Merging Data-Driven And Computational Methods to Understand Ice Nucleation

Martin Fitzner

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

Doctor of Philosophy

of

University College London.

Department of Physics and Astronomy
University College London

November 2, 2018
I, Martin Fitzner, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

November 2, 2018
Abstract

Heterogeneous ice nucleation (IN) is one of the most ubiquitous phase transitions on earth and impacts a plethora of fields in industry (e.g. air transport, food freezing and harsh-weather operations) and science (e.g. freeze avoidance of animals, cryobiology, cloud research). Still to date, we are lacking reliable answers to the question: What is it at the molecular scale that causes an impurity to facilitate the freezing process of supercooled liquid water? In this thesis we make headway towards identifying such microscopic principles by performing computational studies combined with data-driven approaches. In chapter 3 we screen a range of model substrates to disentangle the contributions of lattice match and hydrophobicity and find that there is a complex interplay and an enormous sensitivity to the atomistic details of the interface. In chapter 4 we show that the heterogeneous setting can alter the polymorph of ice that forms and introduce the concept of pre-critical fluctuations, yielding new ideas to design polymorph-targeting substrates. Chapter 5 deals with the liquid dynamics before and during the nucleation event, an aspect of nucleation that mostly goes unrecognized. We show that the homogeneous nucleation event happens in relatively immobile regions of the supercooled liquid, a finding that opens new avenues to understand and influence heterogeneous nucleation by targeting dynamics rather than structure. Finally, Chapter 6 builds on the large amount of data created during this project in that we combine all previously simulated systems and devise a machine-learning approach to find the most important descriptors for their ice nucleation activity (INA). With this we identify new microscopic guidelines and demonstrate that the quantitative prediction of heterogeneous INA is in reach. The unveiling of a computational artifact that potentially
Abstract

affects many computational interface studies is also part of this thesis.
Impact Statement

The insight obtained in this thesis moves forward the field of heterogeneous ice nucleation in several ways. Our examination of the common text book rules revealed that they are inadequate and we propose new guidelines and principles that describe the nucleation enhancement of an impurity. While these still have to be refined further (possibly with data-driven approaches as devised herein) and experimentally tested, this contributes to improved understanding of aerosol behavior in cloud sciences, enhancing our ability to predict the weather. Eventually, a complete collection of microscopic principles to design efficient ice nucleators could also be used to create more ice-phobic surfaces or coatings (by specifically avoiding these principles) which will be of great interest in the aviation, space and oil industries. Forming macroscopic quantities of pure cubic ice and facilitating environments where cubic ice is stable could be achieved with some of the surface design principles we devised. Besides this being an iconic achievement, this could be of use in cryobiology where cubic ice crystals due to their different shape are believed to be benign to tissue as opposed to hexagonal ice crystals. Our findings on the role of liquid dynamics is of fundamental nature, and as such has wide ranging implications for future academic work dealing with the widest range of questions. For instance, studying the role of dynamics at surfaces, in confinement or in solution could explain previously unsolved riddles, uncover new phenomena or open up ways to influence these systems. Another direct impact can be foreseen for the use of potential truncation in computational simulations of interfaces. Since we showed that the commonly applied standard values for cutoffs can drastically change the physical phenomenon, and are also very hard to detect, we anticipate a change in
Impact Statement

best practices for these kind of simulations. Lastly, we also note that, whilst the focus of this thesis was on ice, there are many other materials that could benefit from the improved theoretical understanding (regarding e.g. polymorphism and the role of dynamics) presented here, for instance when dealing with the formation of metallic alloys or pharmaceutical crystals.
Publications

The work presented in this thesis has led to the following publications:


Publications not included in this thesis:


Publications in preparation:


Acknowledgements

I consider myself very lucky having ended up in this city, in this research group, on this research project. Here, I want to thank everyone that made my time in London a success and pleasure.

Foremost, I wish to thank my supervisor Angelos Michaelides for giving me the opportunity of performing research in his group. I am most grateful to him for the many things I learned, professionally and personally, from his direct and indirect teachings as well as providing me with motivation and confidence in all the right moments. I am also indebted to Gabriele C. Sosso (Gabriele#2) for enabling and enriching my work by teaching me countless things and for our many fruitful discussions.

I thank Laurent Joly, Fabio Pietrucci, Silvio Pipolo and Ming Ma for their help and contributions to making the most of our mutual work. I also want to thank Francesco L. Gervasio, Ben Slater and Christoph G. Salzmann for the insightful discussions we had about my research.

The atmosphere in the ICE group is far from icy and not only productive but also always fun and light-hearted. For this I thank all the people who I have met during my time in the group: Andrea, Chiara, Christopher, Conor, Fabian, Gabriele#1, Gabriella, Gerit, Ji, Johan, Michael, Patrick, Philipp, Piero, Sam, Steve, Tai, Yasmine and Wei. Wei is also eternally thanked for bequeathing me his awe-inspiring office chair, which I hereby declare property of whoever is the most senior PhD student in the ICE group (if two students happen to have the same level of seniority the possession should go to the one who can bake the cake that contains the most cherries).
Lastly, my gratitude goes to my parents Jutta and Knut Fitzner who’s unconditional love and support have been a major pillar in my life I am ineffably grateful for.
# Contents

1 Introduction .......................... 23

2 Theory and Computational Methods ........................................... 29
   2.1 Fundamentals of Statistical Mechanics ............................... 29
      2.1.1 Temperature, Entropy and the Microcanonical Ensemble .... 30
      2.1.2 Canonical Ensemble ........................................ 31
      2.1.3 Classical Limit ........................................... 32
   2.2 Molecular Dynamics Simulations ...................................... 33
      2.2.1 Algorithm .................................................. 33
      2.2.2 Realizing the Canonical Ensemble ............................ 34
      2.2.3 Force-Fields for Water .................................... 36
   2.3 Classical Nucleation Theory .......................................... 37
      2.3.1 Homogeneous Nucleation .................................... 37
      2.3.2 Heterogeneous Nucleation .................................. 39
   2.4 Approaches for Studying the Nucleation Event .................... 41
      2.4.1 Brute-Force Simulations .................................... 41
      2.4.2 Free Energy Reconstruction from Unbiased MD ............. 42
      2.4.3 Metadynamics .............................................. 43

3 Interplay Between Surface Morphology and Hydrophobicity .......... 47
   3.1 Introduction ..................................................... 48
   3.2 Methods ......................................................... 51
      3.2.1 System and Computational Methods .......................... 51
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.2.2</td>
<td>Machine-Learning Approaches and Importance Rating</td>
<td>107</td>
</tr>
<tr>
<td>6.3</td>
<td>Results and Discussion</td>
<td>110</td>
</tr>
<tr>
<td>6.3.1</td>
<td>Selected Features and Prediction Performance</td>
<td>110</td>
</tr>
<tr>
<td>6.3.2</td>
<td>Feature Interpretation</td>
<td>114</td>
</tr>
<tr>
<td>6.4</td>
<td>Conclusions</td>
<td>117</td>
</tr>
<tr>
<td>7</td>
<td>Truncated Potentials Can Lead to Unphysical Wetting States</td>
<td>119</td>
</tr>
<tr>
<td>7.1</td>
<td>Introduction</td>
<td>119</td>
</tr>
<tr>
<td>7.2</td>
<td>Computational Settings</td>
<td>121</td>
</tr>
<tr>
<td>7.3</td>
<td>Results and Discussion</td>
<td>122</td>
</tr>
<tr>
<td>8</td>
<td>Summary and Outlook</td>
<td>131</td>
</tr>
<tr>
<td>Appendices</td>
<td>136</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>Supplementary Material for Chapter 3</td>
<td>137</td>
</tr>
<tr>
<td>A.1</td>
<td>Distribution of the $\bar{q}_3(i)$ Order Parameter</td>
<td>137</td>
</tr>
<tr>
<td>A.2</td>
<td>Compressed Exponential Fit</td>
<td>138</td>
</tr>
<tr>
<td>A.3</td>
<td>Critical Nucleus Size on the (111) surface</td>
<td>139</td>
</tr>
<tr>
<td>A.4</td>
<td>Snapshots of Classified Regions</td>
<td>140</td>
</tr>
<tr>
<td>A.5</td>
<td>Distribution of Pre-Critical Nuclei</td>
<td>142</td>
</tr>
<tr>
<td>A.6</td>
<td>Notes on the Water Model</td>
<td>145</td>
</tr>
<tr>
<td>A.7</td>
<td>Higher Temperatures</td>
<td>148</td>
</tr>
<tr>
<td>B</td>
<td>Supplementary Material for Chapter 4</td>
<td>151</td>
</tr>
<tr>
<td>B.1</td>
<td>Computational Details</td>
<td>151</td>
</tr>
<tr>
<td>B.2</td>
<td>Metadynamics Simulations</td>
<td>152</td>
</tr>
<tr>
<td>B.3</td>
<td>Physical Origin of the Nucleation Enhancement</td>
<td>156</td>
</tr>
<tr>
<td>B.4</td>
<td>Cluster Asphericity</td>
<td>157</td>
</tr>
<tr>
<td>B.5</td>
<td>Different Order Parameters</td>
<td>158</td>
</tr>
<tr>
<td>B.6</td>
<td>Higher Temperature with Brute-Force Simulations</td>
<td>160</td>
</tr>
<tr>
<td>B.7</td>
<td>Extension to hetCNT</td>
<td>161</td>
</tr>
</tbody>
</table>
List of Figures

1.1 Summary of different ice nucleators in the atmosphere. 25

2.1 CNT Illustration 38
2.2 Traditional view of free energy profiles in heterogeneous classical nucleation theory. 40
2.3 Metadynamics Illustration 45

3.1 Simulation boxes and model surfaces. 51
3.2 Monitoring the nucleation event via the potential energy. 53
3.3 Influence of the number of simulations on the calculated nucleation rate. 54
3.4 Nucleation rate as a function of surface area. 55
3.5 Heat map results for the nucleation rates. 57
3.6 Liquid density and interfacial water structure for selected systems. 60
3.7 Images of different ice faces that nucleated on very similar substrates. 62
3.8 Heat maps representing the total layering and the layering excluding the contact layer. 67

4.1 Cluster formation and free energy of ice formation for the two model systems. 79
4.2 Metadynamics results for heterogeneous ice nucleation on the two model systems. 82
4.3 Schematic illustration of the connection between free energy profiles, pre-critical fluctuations and polymorph. 84
List of Figures

5.1 Dynamical heterogeneity in supercooled water. ..................... 92
5.2 Connection between dynamical heterogeneity and pre-critical cluster formation. ......................... 95
5.3 Ice nucleation occurs in relatively immobile domains of supercooled water. .......................... 98
5.4 Dynamical propensity of liquid water molecules surrounding a sizeable ice cluster. ....................... 99

6.1 Summary of nucleation temperatures for all model systems. ........ 106
6.2 Illustration of decision trees and random forests. ..................... 108
6.3 Model performance for regression of $T_n$ and classification. ........ 112
6.4 Features that have been selected by the three different methods. ... 114
6.5 Performance of the lattice match as a single descriptor. .......... 115

7.1 Wetting states and wetting transition temperature. .................. 122
7.2 Free energy profiles of wetting for different cutoffs. ................. 124
7.3 Free energy profiles of wetting at the transition temperature with uncorrected setup and for different correction schemes. .......... 127

A.1 Distribution of the $\bar{q}_3(i)$ order parameter. ....................... 137
A.2 Compressed exponential fitting results for two dissimilar nucleation events. ................................. 138
A.3 Committor probability for the 111 surface. ............................. 139
A.4 Snapshots of classified regions for the 111 surface. .................. 140
A.5 Snapshots of classified regions for the 211 surface. .................. 140
A.6 Snapshots of classified regions for the 100 surface. .................. 141
A.7 Snapshots of classified regions for the 110 surface. .................. 141
A.8 Distribution of pre-critical clusters for promoting and inhibiting examples. .............................. 142
A.9 Water density above the substrate for different interaction energies. 144
A.10 Water density comparison for mW and TIP4P/2005. .................... 145
List of Figures

A.11 Probability density distribution of the size of the biggest hexagonal patch of water molecules in the first overlayer. . . . . . . . . . 147
A.12 Temperature dependence of nucleation rates on the 110 substrate. . . 148
B.1 Overview of simulation systems and nucleation behavior. . . . . . . . . 152
B.2 One-dimensional free energy profiles of nucleation obtained with two different methods. . . . . . . . . . . . . . . . . . . . . . . . . 153
B.3 Committor probabilities for the two substrates. . . . . . . . . . . . 155
B.4 Structural properties of water on the substrates. . . . . . . . . . . . 156
B.5 Difference in cluster asphericity for the substrates. . . . . . . . . . . 157
B.6 Consistency of findings with other order parameters 1. . . . . . . . 158
B.7 Consistency of findings with other order parameters 2. . . . . . . . 159
B.8 Kinetic reconstruction of the free energy profiles for higher temperatures. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 160
C.1 Role of statistics in calculating the dynamical propensity. . . . . . 163
C.2 Influence of the DP threshold choice. . . . . . . . . . . . . . . . . . . 165
C.3 DP domain characterization with different order parameters. . . . . 166
C.4 Dynamical Heterogeneity in the mW model. . . . . . . . . . . . . . 167
C.5 Potential energy versus time for the trajectories harvested with TPS. 168
C.6 Indicative snapshots from trajectories that differ by 100 TPS moves. . 169
C.7 Verification of the coarse-grained representation of ice-like space. . . 169
D.1 Overview of initial features computed. . . . . . . . . . . . . . . . . . 172
List of Tables

4.1 Summary of parameters for the simulations examining pre-critical fluctuations. ........................................ 76
4.2 Computed nucleation rates for the two model systems. ......... 78
5.1 Overview of length and time scales used to characterize dynamical heterogeneity. ........................................ 90
D.1 Explanation of acronyms used for the systems studied. ........... 171
Chapter 1

Introduction

The world of ice and of eternal snow, as unfolded to us on the summits of the neighbouring Alpine chain, so stern, so solitary, so dangerous, it may be, has yet its own peculiar charm.

Hermann von Helmholtz, 1865

It is a peculiar spectacle to watch a few of the university’s new first year students during London’s (rare) snowfall perform an oddly excited dance, catching snowflakes and taking pictures outside with an enthusiasm (out in the freezing cold) that can only mean that they have never seen snow before. Subsequently, one could assume that snow and ice have, up to this point, never played any role in their life.

This could not be further from the truth, as for instance climate even in snow-free regions is influenced by clouds and their properties [1, 2]. The latter in turn are strongly affected [3] by airborne particles (such as soot [4, 5], dust [6–8] or organic matter [9, 10]) triggering water droplets to freeze, or being seeds for condensation and subsequent freezing [11]. Ice nucleation also plays a role in the transport industry, the formation of ice on airplane wings or at exposed equipment being a most unfavorable event. In some cases such occurrences have lead to catastrophic consequences, such as the emergency landing of an inbound plane from Beijing at Heathrow airport [12] or the wreckful explosion of one of SpaceX’s landing-rocket prototypes [13]. In agriculture, overnight frost damage in crops can occur at very mild temperatures of only 2 degrees below 0°C due to the presence of an ice nucleating protein in the cell membrane of pseudomonas syringae, a bacteria commonly
found in crops [14]. Genetically modified variants of this bacterium that were lacking the DNA for this particular protein were actually the first genetically modified organisms (GMOs) to be released into the environment [15]. In addition, technological applications such as the cryopreservation of tissue, sperm or organs are hampered by our limited capability to control the ice crystal formation upon cooling so that the material gets damaged.

This non-exhaustive set of examples is rather diverse, but what is a common ingredient in all of them is that liquid water is in contact with some sort of impurity, which then triggers the freezing process. This step is called heterogeneous ice nucleation and is opposed to homogeneous nucleation (i.e. the freezing of water in the absence of any impurity) the dominant starting point for ice freezing in nature. This is due to the fact that liquid water can have a metastable existence below 0°C, indeed micro-droplets in clouds can stay liquid down to −37.5°C [17] and in the lab large volumes of water at −20°C can be kept liquid for more than 100 days [18]. The interface with an alien material reduces the lifetime of the metastable supercooled water and thus enhances the rate of nucleation (we will provide a more quantitative description of nucleation in chapter 2). This opens up a complex question: What is it that makes a certain material good or bad at enhancing the ice nucleation process? The number of possible interfaces is as astronomical as the number of possible materials one can think of (diverse examples of different atmospherically relevant ice nucleating materials are given in Fig. 1.1). Hence, to date there is still no conclusive answer to this, not to speak of a predictive set of rules about the heterogeneous ice nucleation activity (INA) of a substrate. The fundamental questions of how ice is forming was actually named as one of the ten things we still need to know about ice and snow [19].

A significant amount of excellent experimental work was dedicated to understand the formation of ice. These contributions mainly come from two fields. First, there is atmospheric chemistry, where real field samples are investigated regarding their contents and INA. This provides us with the information which materials are

---

1It is also worth noting that the release of genetically modified *pseudomonas syringae* onto a strawberry field created the first public outrage about genetically modified organisms [16]
Figure 1.1: Summary of ice active site density \( n_s \) for a range of materials found in the atmosphere as a function of temperature \( T \). This includes soot, different types of dust, bacteria, pollen, fungal spores and algae. Figure taken from Ref. 11.

capable and responsible for enhanced heterogeneous ice nucleation. For instance, atmospheric measurements revealed that there is ice-nucleating biogenetic material in the sea surface microlayer [9] and even glassy aerosols promote heterogeneous ice nucleation [20]. Yet, this approach is lacking the ability to investigate what is going on at the microscopic scale. Even if bulk structures of the corresponding materials are known, the actual interface with water might not look like what one would expect from simply truncating the crystal\(^2\). For instance we know many complicated surface reconstructions even for simple elemental materials [21] and the interaction with water can also change the surface structure [22]. The second field that contributed significantly to our understanding of ice nucleation is surface science. In there, the interface between a material and water is studied with great resolution and purity, therefore actually establishing the interfacial structure of water and subsequently revealing its potential catalytic effect on ice nucleation. The downside, however, is that even with the current state of the art such measurements are often only possible under artificial conditions such as ultra high vacuum. Of course we note that both of these fields are in continuous development and improvements such as femtosecond X-ray scattering [23, 24] are promising approaches to further our

\(^2\)“God made the bulk; surfaces were invented by the devil.” Wolfgang Pauli
understanding of heterogeneous ice nucleation.

A principle link between the fields of atmospheric chemistry and surface science are computer simulations [25]. With these, we can simulate systems of up to several thousand molecules and investigate their evolution in time. Applied to supercooled water (on structures that are verified by surface science) we can probe the predictions of atmospheric chemistry and understand the microscopic reason behind nucleation. An example of the synergy between experiment and simulation is the feldspar mineral, where in a combined experimental-computational study the odd orientation of ice crystals forming on the surface was explained via metastable surfaces existing in cracks [26]. The simulation of any nucleation event is however far from trivial or routine [27, 28], simply because the evolution times necessary to study ice nucleation fall in a sweet spot (μs to s) that is too slow for atomistic simulations and too fast for microscopic experiments. We will discuss in the methods chapter advanced approaches that still enable the computational study of ice nucleation thanks to recent developments in the field [27, 29–31].

Let us put the contents and aim of this thesis into perspective. The typical aim of studies on heterogeneous ice nucleation is an answer to the question “Why is material X good or bad at heterogeneous ice nucleation?”. For the reasons outlined before all too often this is too complex of a task to be achieved by any single method, and often even in combined efforts. For this reason we propose to split up this question into two parts: 1. “What is the real interface of material X with water?” and 2. “If a surface has interfacial structure or property X, what will it mean for its INA?”. The first part is well suited to be studied with either surface science methods or high-level computational methods, such as density functional theory (DFT) [32, 33]. Question 2 then needs to be answered by a method that can follow the time evolution of the system on the microscopic scale, as we are interested in a rate of formation. Classical molecular dynamics (MD) methods, as mainly applied in this thesis, provide a tool to do exactly this. Hence, it is generally not our aim to study any particular material (except of course the liquid water), but more to answer fundamental questions about the impact on the INA of certain surface and interface
properties. The answers of questions 1 and 2 can then be combined to provide a comprehensive understanding of heterogeneous ice nucleation. It is hoped that this work has gone some way in approaching this goal.

The structure of this thesis is as follows: The next chapter will introduce theoretical concepts and computational methods that are relevant to all subsequent chapters. In chapter 3 we present a screening study of well-controlled model surfaces aimed at investigating if two of the most prevalent “rules” for heterogeneous ice nucleation are indeed working or not and clarify the sensitivity of the nucleation rate to small substrate variations. In chapter 4 we introduce the concept of pre-critical fluctuations and their relevance in heterogeneous nucleation. In there we also investigate the possibility of surface-induced formation of metastable polymorphs. Chapter 5 deals with homogeneous ice nucleation, in that we establish the role and connection of dynamical heterogeneity and liquid dynamics to the nucleation event. We take stock of all data created during this project in chapter 6 and utilize data-driven machine learning approaches to identify the most important descriptors to predicting the INA of a substrate. Chapter 7 highlights a hard-to-detect, potentially wide-spread computational artifact that was encountered during our studies, and how to avoid it. Lastly, we summarize our findings and provide an outlook for the field in chapter 8.
Chapter 2

Theory and Computational Methods

Hence, our trust on Molecular Dynamics simulation as a tool to study the time evolution of many-body systems is based largely on belief. To conclude, let us say there is clearly still a corpse in the closet. We believe this corpse will not haunt us, and we quickly close the closet.

D. Frenkel and B. Smit, Understanding Molecular Simulation

2.1 Fundamentals of Statistical Mechanics

To describe the transition between two phases (liquid water and ice in our case) we need to marry the microscopic details of molecular interactions with the macroscopic\(^1\) behavior of the whole system. The physical theory that achieves this with astounding success\(^2\) is statistical mechanics and in this thesis we will make plenty use of its language and concepts.

While classical and quantum mechanics aim at describing the evolution of a single state, statistical mechanics instead considers an ensemble of systems, i.e. a collection of many states together with a weight proportional to their probability. We consider a quantum-mechanical description first, as the fundamentals are easier derived at this level.

\(^1\)While microscopic interactions are usually considered between few bodies, the macroscopic picture entails a collective view on many more entities at once, typically on the order of the Avogadro number \(N_A \approx 6.0221 \times 10^{23}\)

\(^2\)“It is the only physical theory of general contents of which I am convinced, that regarding the applicability of its fundamentals, will never be overturned.” Albert Einstein (1949)
2.1.1 Temperature, Entropy and the Microcanonical Ensemble

Any given system is described by a quantum state $|i\rangle$ and obeys the Schrödinger equation $\hat{H}|i\rangle = E_i|i\rangle$ with the Hamilton operator $\hat{H}$ and the system's energy $E_i$. We shall denote such state as a *microstate*. Since in a macroscopic system there is an astronomically large amount of degrees of freedom, many of the microstates lead to the same energy. We call $\Omega(E)$ the degeneracy, i.e. the number of microstates with energy $E$. The fundamental assumption of statistical mechanics, termed the *postulate of equal a priori probabilities*, states that in an ensemble of systems with fixed volume $V$ and fixed number of particles $N$, the probability of observing microstate $|i\rangle$ with energy $E_i$ is the same for all states and thus depends only on how many states with this energy exist at these conditions:

$$P(i) = \frac{1}{\Omega(E_i)} \quad (2.1)$$

Much of what follows can be derived from this assumption. However, there is no mathematical proof\(^3\) for its correctness and as with any other axiom, it has been accepted by the community only throughout the course of time due to the success of the resulting framework\(^4\).

We now consider a system that can be split into two weakly interacting subsystems, i.e. there can be energy exchange between them, but the total energy is always $E = E_1 + E_2$. Hence, the number of states is:

$$\Omega(E_1, E_2) = \Omega_1(E_1) \cdot \Omega_2(E_2) \quad (2.2)$$

It can be shown that for the most probable distribution of energies $E_1$ and $E_2$ among the subsystems, the following condition is true

$$\left( \frac{\partial \ln \Omega_1}{\partial E_1} \right)_{N_1, V_1} = \left( \frac{\partial \ln \Omega_2}{\partial E_2} \right)_{N_2, V_2} \quad (2.3)$$

\(^3\)With additional assumptions it can be shown to follow from Schrödinger's equation.

\(^4\)At the time of its proposition Ludwig Boltzmann was one of the biggest proponents of this assumption. He was however harshly criticized by his peers and the inability to deliver proof eventually contributed to his depression that lead to his suicide in 1906 [34].
With Boltzmann’s definition of entropy $S$ \(^5\)

\[
S \equiv k_B \cdot \ln \Omega(N,V,E) \tag{2.4}
\]

where $k_B \approx 1.3806 \times 10^{-23}$ m\(^2\) kg s\(^{-2}\) K\(^{-1}\) is the Boltzmann constant, we see that:

\[
\left( \frac{\partial \ln \Omega}{\partial E} \right)_{N,V} = \frac{1}{k_B} \frac{\partial S}{\partial E} = \frac{1}{k_B T} \tag{2.5}
\]

where the last identity is known from classical thermodynamics. It follows that for the most likely distribution of energies in the two subsystems, their temperatures are equal:

\[
T_1 = T_2 = T \tag{2.6}
\]

and they are said to be in thermal equilibrium. The ensemble where $N$, $V$ and $E$ are fixed is called the microcanonical or NVE ensemble.

### 2.1.2 Canonical Ensemble

Suppose now we want to consider a system that is not at constant energy, but at constant temperature $T$. Once again we imagine a larger system where the subsystem of interest is coupled to another, much larger system we call bath (with $E_{\text{bath}} \gg E_{\text{sys}}$ and $E = E_{\text{sys}} + E_{\text{bath}}$). Since all states of equal energy are equally likely and $\Omega(E) = \Omega_{\text{sys}}(E_{\text{sys}}) \cdot \Omega_{\text{bath}}(E - E_{\text{sys}})$, the probability of observing the system of interest with an energy $E_{\text{sys}}$ is:

\[
P(E_{\text{sys}}) = \frac{1}{\Omega_{\text{sys}}(E_{\text{sys}})} = \frac{\Omega_{\text{bath}}(E - E_{\text{sys}})}{\Omega(E)} \tag{2.7}
\]

\(^5\)Even though much of the theory is based on this equation, it is still somewhat under debate today and alternatives have been proposed [35, 36].
Taking the logarithm and a Taylor approximation of $\Omega_{\text{bath}}$ around $E$ we find

\[
\ln P(E_{\text{sys}}) = \ln \Omega_{\text{bath}}(E - E_{\text{sys}}) - \ln \Omega(E) \\
= \ln \Omega_{\text{bath}}(E) - \frac{\partial \ln \Omega_{\text{bath}}(E)}{\partial E}
\cdot E_{\text{sys}} - \ln \Omega(E) + O(1/T) \tag{2.8}
\]

This means that

\[
P(E_{\text{sys}}) = \frac{\Omega_{\text{bath}}(E)}{\Omega(E)} \exp \left[ -\frac{E_{\text{sys}}}{k_B T} \right] \equiv \frac{1}{Z} \exp \left[ -\beta E_{\text{sys}} \right] \tag{2.9}
\]

where $\beta = \frac{1}{k_B T}$ and $Z = \frac{\Omega(E)}{\Omega_{\text{bath}}(E)} = \sum_i \exp[-\beta E_i]$ is called the \textit{canonical partition function}, where the latter identity has to follow from the constraint $\sum_i P_i = 1$. Equation 2.9 is known as the Boltzmann distribution and it describes the likelihood of observing states of different energies at a fixed temperature. The ensemble where $N$, $V$ and $T$ are fixed is called the \textit{canonical} or $NVT$ ensemble. We know that the Helmholtz free energy $F(N,V,T) = E - TS$ is related to $Z(N,V,T)$ via

\[
\beta \cdot F(N,V,T) = -\ln Z(N,V,T) \tag{2.10}
\]

### 2.1.3 Classical Limit

To obtain the classical limit of the partition function we make use of the quantum mechanical formulation again. $Z$ becomes

\[
Z = \sum_i \exp[-\beta E_i] = \sum_i \langle i | \exp[-\beta \hat{H}] | i \rangle \\
= \text{Tr}\{ \exp[-\beta \hat{H}] \} = \text{Tr}\{ \exp[-\beta (\hat{K} + \hat{U})] \} \tag{2.11}
\]

where $\hat{K}$ and $\hat{U}$ are the operators of kinetic and potential energy. Since the identity $e^{\hat{A} + \hat{B} + [\hat{A}, \hat{B}]} = e^{\hat{A}} e^{\hat{B}}$ holds for Hermitian operators and the commutator $[\hat{K}, \hat{U}] \propto \hbar \rightarrow 0$
in the classical limit we find

\[
Z = \text{Tr}\{\exp[-\beta \hat{K}] \exp[-\beta \hat{U}] \}\n\]

\[
= \sum_r \sum_p \langle r | p \rangle \langle p | \exp[-\beta \hat{K}] | p \rangle \langle p | r \rangle \langle r | \exp[-\beta \hat{U}] | r \rangle
\]

\[
\rightarrow \frac{1}{\Lambda^{3N} N!} \int d\mathbf{r}^N \exp[-\beta U(\mathbf{r}^N)]
\]

(2.12)

where we did the trace over \( \hat{K} \) analytically, leading to the prefactor with \( \Lambda = h / \sqrt{2\pi mk_B T} \). \( N! \) was inserted manually\(^6\). Equation 2.12 essentially connects the microscopics encoded in \( U(\mathbf{r}^N) \) with the macroscopics encoded in \( Z \).

### 2.2 Molecular Dynamics Simulations

#### 2.2.1 Algorithm

The partition function \( Z \) contains all thermodynamic information of the system and we can compute expectation values of physical observables \( X \) according to

\[
\langle X \rangle = \sum_i P_i X_i
\]

(2.13)

The practical evaluation of equation 2.12 however requires approaches to compute these high dimensional integrals. One such approach that is known from mathematics is integration with random (and more sophisticated) sampling, called Monte Carlo. Another approach is to explicitly evolve the system in time according to Newtons equations that follow from the systems interactions in \( U(\mathbf{r}^N) \):

\[
m_i \mathbf{a}_i = m_i \frac{d^2 \mathbf{r}_i}{dt^2} = \mathbf{F}_i = -\nabla_i U(\mathbf{r}^N)
\]

(2.14)

Based on the ergodic hypothesis we can then replace ensemble averages by time averages

\[
\langle X \rangle = \frac{1}{t} \int_0^t d\tau X(\tau)
\]

(2.15)

\(^6\)In quantum statistical mechanics the factor \( N! \) is usually justified by the individual particles being indistinguishable. In classical statistical mechanics it can however be shown that \( N! \) is also required to make the resulting free energy extensive, even if the particles (e.g. colloids) can in principle be distinguished by internal degrees of freedom (e.g. their amorphous internal structure) \cite{37}.
In general, however, systems divided by large free energy barriers are not ergodic to warrant this. In practice this is usually not a problem since we are often interested in averages of certain sub-parts of phase space that are ergodic, i.e. $\langle X \rangle_{\text{liq}}$ is the expected value of the observable $X$ when the system is in liquid state.

For the numerical approximation of the motion (termed trajectory) we need to discretize equation 2.14. We here follow the velocity Verlet scheme based on Taylor approximations:

$$r(t + \Delta t) = r(t) + \Delta t \frac{dr(t)}{dt} + \frac{(\Delta t)^2}{2} \frac{d^2r(t)}{dt^2} + O((\Delta t)^3)$$

$$\approx x(t) + \Delta t \cdot v(t) + (\Delta t)^2 \frac{F(t)}{2m}$$  (2.16)

$$v(t + \Delta t) \approx v(t) + \Delta t \frac{a(t) + a(t + \Delta t)}{2}$$  (2.17)

This allows us to assemble an iterative scheme to integrate the equations of motion:

1. Initialize the system with positions $r^N$ and velocities $v^N$

2. $r_i(t + \Delta t) = r_i(t) + \Delta t \cdot v_i(t) + \frac{a_i(t)}{2} (\Delta t)^2$

3. $a_i(t + \Delta t) = -\frac{1}{m_i} \nabla_i \cdot U(r^N(t + \Delta t))$

4. $v_i(t + \Delta t) = v_i(t) + \frac{a_i(t) + a_i(t + \Delta t)}{2}$

5. Calculate properties of interest and save data

6. Repeat points 2 to 5 until enough time has been simulated

### 2.2.2 Realizing the Canonical Ensemble

The scheme that was just devised would realize the $NVE$ ensemble, provided the timestep $\Delta t$ is chosen small enough, otherwise the energy of the system would drift substantially over the course of the simulation rather than stay constant. The appropriate timestep differs for each system and should be checked individually.

In practice, however, we are mostly interested in systems at constant temperature, i.e. the $NVT$ ensemble. An approach that is directly usable with the velocity
2.2. Molecular Dynamics Simulations

Verlet approach is the *canonical rescaling* of velocities [38], where the temperature is evolved according to a stochastic equation and periodically all velocities are rescaled to this value \(^7\).

Another approach that was used extensively in this thesis is the addition of degrees of freedom to the Hamiltonian, according to Nosé-Hoover [40–42]:

\[ H(p^N, r^N, s, p_s) = K(p^N) + U(r^N) + \frac{p_s^2}{2Q} + \frac{3N+1}{\beta} \ln(s) \]  

(2.18)

where \(s\) is the additional degree of freedom, \(p_s\) is its momentum and \(Q\) its fictitious mass, that is most often converted into a thermostat-relaxation time \(\tau\) via \(\tau = \sqrt{\frac{\beta Q}{N_{df}}}\) with \(N_{df}\) being the number of degrees of freedom. It can be shown that this samples the canonical ensemble, i.e.

\[ \langle X(p^N/s, r^N) \rangle_{NH} = \langle X(\hat{p}^N, r^N) \rangle_{NVT} \]  

(2.19)

We note that this requires a more sophisticated way of integrating the timestep that can be achieved via generating time-reversible integrators with a Liouville operator approach [42, 43].

Lastly, the reader is owed an explanation regarding the epigraph of this chapter. Due to the error introduced by the discretization of the time integration in equation 2.16 the trajectory of a single particle would only approximate the true one. But due to the presence of many other particles and the resulting interactions any computed trajectory starting with given conditions will diverge exponentially fast from the true one. This is termed the Lyapunov instability [44] and is referred to as the “corpse” by Frenkel and Smit. However, as opposed to ordinary mechanics, where the precise evolution of the system given its starting conditions is the central aim, in MD we are interested in statistical predictions of many trajectories with different starting conditions. There is good evidence [45–47] that MD trajectories are still representative of true trajectories in phase space (similar to so called shadow-orbits

\(^7\)Rescaling the velocities every timestep to the target temperature should be avoided, as this does not yield the \(NVT\) ensemble but rather the flying ice cube effect, see [39].
in classical mechanics), albeit no proof exists. For details see Ref. 48.

2.2.3 Force-Fields for Water

The last ingredient to our computational studies is the specific interaction between particles, described by $U(r^N)$. We have used different potentials for the interaction between water and foreign particles and describe the details in the corresponding chapter. Here we briefly describe the two potentials we used for the water-water interaction.

In principle, the forces needed to integrate the equations of motion could be calculated with a higher level of theory, e.g. density functional theory [32, 33, 49], quantum Monte Carlo [50, 51] or machine-learning based potentials [52–54]. However, this will increase the computational cost of the force-calculation step significantly, typically at least 4 orders of magnitude. As a result we would not be able to simulate a sufficient number of timesteps and hence, we choose what is termed classical interaction potentials. The latter are closed analytical formulas, termed force-fields, that are simply evaluated at each timestep. Also within the class of force-fields there are many choices for water, ranging from the highly accurate but expensive MB-POL [55] to coarse-grained water models [56]. In this thesis we have mainly used two variants.

First, we employ the TIP4P/Ice model [57]. It consists of 4 sites, i.e. two charged hydrogens, one uncharged oxygen and a virtual site carrying the rest of the charge. The angles and bonds between the particles are constrained with the SHAKE algorithm [58] which enables a larger timestep. It must be noted that as a result of this the bond vibrations and dissociation cannot be captured by this model, but for our target of simulating nucleation we do not anticipate any major influence coming from these effects. TIP4P/Ice has been parametrized to yield excellent density values for the ices, has a melting point in good agreement with the experimental value and exhibits good liquid densities at ambient temperatures and pressures.

The second potential we employed is the coarse-grained mW potential [29],

---

8Even though these force-fields are often parametrized to give good values for these quantities it is still worth noting that even higher-level theory approaches such as DFT can have a hard time to accurately predict bulk liquid water properties in silico [59]
which unites all atoms of the water molecule in a single site. By noticing the similarity (i.e. regarding the tetrahedrality of bonding patterns or the density anomaly) between water and group-IV materials such as silicon, it was realized that water can be modeled with the Stillinger-Weber potential [60] that was initially devised for those atomic materials. It consists of a two- and three-body term which induces tetrahedral order in the material:

$$ U(i, j, k) = \sum_{i} \sum_{j > i} \phi_2(r_{ij}) + \sum_{i} \sum_{j \neq k > j} \phi_3(r_{ij}, r_{jk}, \theta_{ijk}) $$

$$ \phi_2(r) = A \varepsilon \left[ B \left( \frac{\sigma}{r} \right)^p - \left( \frac{\sigma}{r} \right)^q \right] \exp \left( \frac{\sigma}{r - a\sigma} \right) $$

$$ \phi_3(r, s, \theta) = \lambda \varepsilon [\cos(\theta) - \cos(\theta_0)]^2 \exp \left( \frac{\gamma \sigma}{r - a\sigma} \right) \exp \left( \frac{\gamma \sigma}{s - a\sigma} \right) $$

The mW model yields excellent values for liquid and ice densities, the melting point while decreasing the computational cost by approximately 2 orders of magnitude. While this is a major advantage over fully-atomistic potentials, the obvious shortcomings of the mW model are (by construction) inability to model dissociation, nuclear quantum effects or rotational effects e.g. on the hydrogen-bond network. While we do not expect major implications for nucleation of the two former points, the molecule orientations and their influence on interfacial structures might affect the nucleation process. Being aware of this we can see the mW model as a compromise that with the current computational capabilities of typical high-performance clusters enables a wide variety of nucleation investigations.

### 2.3 Classical Nucleation Theory

#### 2.3.1 Homogeneous Nucleation

Our basic understanding of the nucleation phenomenon was developed almost 100 years ago [61–66]. The qualitative understanding gained therein makes up the foundation of our understanding of phase transitions in liquids and vapors and has been termed classical nucleation theory (CNT). A large body of works has dealt with reviewing, criticizing and extending CNT [67–69]. In here, we will briefly describe
the very basics required to understand the work in this thesis. In particular, in chapter 4 we make use of CNT to understand the meaning of pre-critical fluctuations and the role of different polymorphs.

![Figure 2.1: a) Sketch of the resulting free energy profile for cluster formation as a function of the cluster radius \( r \). b) Illustration of the equilibrium shape for heterogeneous nucleation on a flat substrate.](image)

The reason why nucleation is an activated process rather than a simple relaxation is due to competing effects of bulk energy gain and interfacial energy cost. We consider the free energy \( F \) of a metastable liquid containing a single crystalline cluster with \( n \) molecules:\(^9\)

\[
F(n) = -\frac{n}{\rho_{\text{crys}}} \Delta \mu + A \sigma
\]  

(2.21)

Here \( \Delta \mu = \mu_{\text{liq}} - \mu_{\text{crys}} \) is the chemical potential difference between the two phases\(^{10}\) (positive if the liquid is metastable), \( \rho_{\text{crys}} \) is the density of the crystalline phase, \( A \) is the surface area of the cluster and \( \sigma \) the free energy cost per unit area between the crystalline and the liquid phase. If we assume a spherical shape of the nucleus we obtain:

\[
F(r) = -\frac{4\pi}{3} r^3 \Delta \mu + 4\pi r^2 \sigma
\]  

(2.22)

where \( r \) is the radius of the crystalline nucleus. Equation 2.22 describes a free energy profile with a barrier (see Figure 2.1a). The free energy barrier height \( \Delta F \), the critical radius \( r_{\text{crit}} \) and the critical nucleus size \( n_{\text{crit}} \) can be obtained from this to

\(^9\)Since we are only interested in relative free energies we can without loss of generality assume the free energy of the liquid containing no crystalline cluster to be 0.

\(^{10}\)Typically, \( \mu \) is assumed to be per particle, however here we assume it to be normalized per volume, which is why we need to normalize by \( \rho_{\text{crys}} \).
2.3. Classical Nucleation Theory

be:

\[
\Delta F = \frac{16\pi}{3} \left( \frac{\sigma^3}{(\Delta \mu)^2} \right)
\]

(2.23)

\[
r_{\text{crit}} = \frac{2\sigma}{\Delta \mu}
\]

(2.24)

\[
n_{\text{crit}} = \frac{32\pi}{3} \left( \frac{\sigma^3 \rho_{\text{cys}}}{(\Delta \mu)^3} \right)
\]

(2.25)

The free energy profile can be conveniently expressed as:

\[
F(r) = \Delta F \left[ 3 \left( \frac{r}{r_{\text{crit}}} \right)^2 - 2 \left( \frac{r}{r_{\text{crit}}} \right)^3 \right]
\]

(2.26)

\[
F(n) = \Delta F \left[ 3 \left( \frac{n}{n_{\text{crit}}} \right)^{2/3} - 2 \left( \frac{n}{n_{\text{crit}}} \right) \right]
\]

(2.27)

Based on this the classical nucleation rate can obtained via a kinetic-state model [70] as:

\[
J = \rho_{\text{liq}} Z f_{\text{crit}} \cdot \exp \left[ -\frac{\Delta F}{k_B T} \right]
\]

(2.28)

with the liquid density \(\rho_{\text{liq}}\), the particle attachment frequency to the critical cluster \(f_{\text{crit}}\) and the Zeldovich factor \(Z\) which approximately accounts for barrier recrossings.

2.3.2 Heterogeneous Nucleation

So far we described the homogeneous case. The equilibrium shape in the heterogeneous scenario is no longer purely spherical. If we assume (in good approximation with the systems investigated in this thesis) a planar interface, then the shape of the nucleus is a spherical cap with a contact angle \(\theta\) given by Young’s equation (see also Figure 2.1b):

\[
\cos(\theta) = \frac{\gamma_{\text{sl}} - \gamma_{\text{sc}}}{\gamma_{\text{cl}}}
\]

(2.29)

with the \(\gamma\)’s being the interfacial tensions for substrate-liquid, substrate-cluster and cluster-liquid. In this geometry the resulting quantities that describe the heteroge-
The traditional view of free energy profiles in heterogeneous classical nucleation theory: Shown are free energy profiles for a crystalline cluster containing $n$ molecules for different contact angles. The inset shows the definition of the contact angle and representative spherical caps corresponding to different values of $f_V(\theta)$.

Homogeneous nucleation can be expressed via the homogeneous quantities [70]:

$$\Delta F_{\text{crit}}^{\text{het}} = f(\theta) \cdot \Delta F_{\text{crit}}^{\text{hom}}$$
$$n_{\text{crit}}^{\text{het}} = f(\theta) \cdot n_{\text{crit}}^{\text{hom}}$$

The factor $f(\theta) = (2 + \cos(\theta))(1 - \cos(\theta))^2/4 \in [0, 1]$ is also often referred to as volumetric enhancement factor, because it describes the enhancement of the nucleation rate but also coincides with the fraction of volume that the spherical cap with contact angle $\theta$ occupies compared to a full sphere with the same radius. Illustrative free energy profiles and corresponding spherical caps are displayed in figure 2.2.

The shortcomings of CNT are manifold. For instance the assumption of spherical clusters is often violated [71, 72] (in particular for small clusters), the assumption of a sharp interface is ambiguous [25, 73] and multi-step nucleation processes cannot be described (by construction) [68, 74]. These aspects, however, apply mostly to the assumptions going into the homogeneous case. Indeed, the validity of the heterogeneous extension to CNT (which we term hetCNT) is less well understood. We shall see in chapter 4 that the heterogeneous extension is breaking down should the substrate facilitate the formation of a polymorph that is not the
2.4 Approaches for Studying the Nucleation Event

2.4.1 Brute-Force Simulations

Due to the previously introduced barrier to nucleation, simulating the actual transformation process is not straightforward. The height of the barrier is typically on the order of ten to thousand times the thermal energy $k_B T$ [25]. This means that, even though the transformation process can be rather short once it occurs, the event will usually not be observed on timescales accessible to observation method (MD in our case). Events of this character are often referred to as rare events and occur in a variety of different fields, e.g. pore translocation, earthquakes, protein conformational changes and switching in biochemical networks [75].

Due to the computational efficacy of the mW model however, it is possible to observe the nucleation event directly for certain conditions. This is typically the case when the liquid is strongly supercooled or in contact with a substrate that enhances the heterogeneous nucleation substantially. Studying nucleation in this manner is often referred to as an unbiased or brute-force approach an we made use of it in several parts of this thesis (e.g. chapters 3 and 4).

For studies of heterogeneous nucleation we can directly calculate the nucleation rate $J$ and compare it to the homogeneous one via calculating the survival probability $P_{\text{sur}}(t)$ of the liquid. To this end we measure an ensemble of induction times $\{t_i\}$, i.e. the times it takes for a liquid system to nucleate into ice, from $N$ independent simulations. $P_{\text{sur}}(t)$ is then approximated via

$$P_{\text{sur}}(t) = \frac{1}{N} \sum_{i=1}^{N} \Theta(t_i - t)$$

and fitted according to

$$P_{\text{sur}}(t) = \exp \left[ - (J \cdot t)^\gamma \right]$$

where $\Theta$ is the Heaviside step-function, $J$ is the nucleation rate and $\gamma$ is an additional fit parameter accounting for possible non-exponential kinetics [76]. Strictly
speaking, $\gamma$ should always be 1 for a nucleation event, however in practice we find that $\gamma$ is slightly larger than 1 due to the initial (short compared to the induction time) relaxation of a system that has been quenched to the a temperature. Moreover $\gamma \gg 1$ if the event is a relaxation rather than a nucleation, i.e. when the barrier is on the order of $1k_B T$. Thus, $\gamma$ can serve as an indication for the character of the transformation process, and we include it in any fit we perform.

There is another possible approach we utilize in chapter 6. Instead of calculating the nucleation rate we perform cooling ramps, starting from a temperature above the freezing point down to the homogeneous nucleation temperature. In this manner we often encompass a temperature at which the system will freeze. Hence, if the cooling rate is slow enough (we typically use less than 1 K/ns) we can use the respective nucleation temperature as an indicator for how good the enhancement by the substrate is. This is suitable to compare the absolute INA of several substrates, while the rate approach is more suitable for comparing substrate’s INA at a given temperature and elucidate the nucleation mechanism.

### 2.4.2 Free Energy Reconstruction from Unbiased MD

From direct simulations of the nucleation event it is also possible to reconstruct the free energy profile [77, 78]. To describe the dynamics of activated processes we apply a Fokker-Planck equation [79]:

$$\frac{\partial P(x,t)}{\partial t} = \frac{\partial}{\partial x} \left[ D(x)e^{-\beta F(x)} \frac{\partial}{\partial x} \left( P(x,t)e^{\beta F(x)} \right) \right] = -\frac{\partial J(x,t)}{\partial x}$$  \hspace{0.5cm} (2.34)

where $P(x,t)$ is the probability of the system at time $t$ being in the state described by the order parameter $x$, $J$ the nucleation rate, $D$ a generalized diffusion coefficient and $F(x)$ is the free energy. If we assume the steady state (time derivatives are zero, denoted by the index “st”) we can obtain from this after rearranging and integrating:

$$\beta F(x) = -\ln P_{st}(x) - J_{st} \int \frac{dx'}{D(x')P_{st}(x')} + C$$  \hspace{0.5cm} (2.35)
2.4. Approaches for Studying the Nucleation Event

Note that \( J_{\text{st}} \) is independent of \( x \) because \( \frac{\partial J_{\text{st}}(x)}{\partial x} = -\frac{\partial P_{\text{st}}(x)}{\partial t} = 0 \). We now introduce the mean first passage time

\[
\tau(x, x_0, a) = \int_{x_0}^{x} \frac{dy}{D(y)} e^{\beta F(y)} \int_{a}^{y} e^{-\beta F(z)} dz
\]

which describes the average time that the system needs to reach \( x \) if it started at \( x_0 \) (we also assume lower / upper boundary conditions for \( x \) as \( a \) and \( b \)). This can be arranged to

\[
\frac{\partial \tau}{\partial x} = A = e^{\beta F(x)} \int_{a}^{x} e^{-\beta F(z)} dz
\]

\[
\ln \left[ D(x) A(x) \right] = \beta F(x) + \ln \int_{a}^{x} e^{-\beta F(z)} dz
\]

\[
\frac{\partial \ln \left[ D(x) A(x) \right]}{\partial x} = \frac{\partial \left[ \beta F(x) \right]}{\partial x} + \frac{e^{-\beta F(x)}}{\int_{a}^{x} e^{-\beta F(z)} dz} = \frac{\partial \left[ \beta F(x) \right]}{\partial x} + \frac{1}{D(x) A(x)}
\]

where \( A(x) = \frac{\partial \tau(x)}{\partial x} \) is a substitute. After rearranging and integrating this yields the practical result for obtaining the free energy from a steady state distribution \( P_{\text{st}}(x) \):

\[
\beta F(x) = C + \ln \left[ B(x) \right] - \int_{a}^{x} \frac{dx'}{B(x')}
\]

where \( B(x) \) is a help quantity that can be obtained by combining equations 2.35 and 2.37 as

\[
B(x) = A(x) D(x) = \frac{1}{P_{\text{st}}(x)} \left[ \int_{a}^{x} P_{\text{st}}(x') dx' - \frac{\tau(x)}{\tau(b)} \right]
\]

The choice of the boundary conditions as well as the numerical integration method did in our experience not meaningfully alter any of the results.

2.4.3 Metadynamics

To study nucleation at higher temperatures or with atomistic models we generally require approaches that somehow facilitate the rare event. Such approaches are termed enhanced sampling methods and can briefly be distinguished into path-based (e.g. transition path sampling [80] or forward-flux sampling [30, 75]) and free-energy based (e.g. umbrella sampling [81] or metadynamics [82–84]) methods. In
this thesis we have made use of metadynamics and thus here briefly describe the basic idea behind it.

The fundamental principle of metadynamics is to deposit artificial potential energy in the part of the phase space the system currently occupies and thus “fill up” the free energy basin corresponding to the liquid to enable the system to escape it \[82, 83\]. This is achieved by adding a bias potential \( V_{\text{bias}} \) to the Hamiltonian:

\[
H(p^N, r^N, t) = K(p^N) + U(r^N) + V_{\text{bias}}(s(r^N), t) \quad (2.40)
\]

\( V_{\text{bias}} \) is time-dependent because it is incrementally constructed (typically of Gaussian hills) during the simulation:

\[
V_{\text{bias}}(s, t) = \sum_{t' = \tau_G, 2\tau_G, \ldots, t' < t} h \cdot \exp \left[ \frac{(s(t) - s(t'))^2}{2\delta^2} \right] \quad (2.41)
\]

Here, \( h \) is the height, \( \delta \) is the width and \( \tau_G \) is the deposition stride of the Gaussians. Most importantly, we introduced \( s(t) = s(r^N(t)) \) which is often referred to as collective variable that condenses the information of all particle positions into a single number (although the approach is easily extended to multiple collective variables). To facilitate the nucleation process, \( s \) should be able to distinguish the liquid and solid state of our simulation system, i.e. \( s \) should take very different numbers for both states. In this thesis we used a variety of different order parameters both for metadynamics and the identification of ice and the details are indicated in the corresponding chapters.

In figure 2.3 we show an illustration of the metadynamics algorithm. As can be seen with increasing time and thus increasing amount of deposited Gaussians, the free energy basins fill up, and the system can explore a wider phase space. Eventually, the bias potential will converge to the negative free energy plus an arbitrary constant \( C \):

\[
\lim_{t \to \infty} V_{\text{bias}}(s, t) \approx -F(s) + C \quad (2.42)
\]

We further made use of other modifications of the metadynamics approach, namely
2.4. Approaches for Studying the Nucleation Event

Figure 2.3: Illustration of the metadynamics approach. a) Time evolution of the generic order parameter $s$. b) Bias potential after different number of deposited Gaussians. c) Free energy of the biased system at different times, the free energy of the unbiased system corresponds to the bold black line. This figure is adapted from [83].

well-tempered metadynamics [84], which improves the smoothness of the convergence by decreasing the Gaussian height $h$ with time, and multiple-walker metadynamics [85] which accelerates the filling of the basins via running parallel simulations that periodically exchange their bias potential.
Chapter 3

Interplay Between Surface
Morphology and Hydrophobicity

What makes a material a good ice nucleating agent? Despite the importance of heterogeneous ice nucleation to a variety of fields, major gaps in our understanding of this ubiquitous process still prevent us from answering this question. In this chapter, we have examined the ability of generic crystalline substrates to promote ice nucleation as a function of the hydrophobicity and the morphology of the surface. Nucleation rates have been obtained by brute-force molecular dynamics simulations of coarse-grained water on top of different surfaces of a model fcc crystal, varying the water-surface interaction and the surface lattice parameter. It turns out that the lattice match of the surface with respect to ice, customarily regarded as the most important requirement for a good ice nucleating agent, is at most desirable but not a requirement. On the other hand, the balance between the morphology of the surface and its hydrophobicity can significantly alter the ice nucleation rate and can also lead to the formation of up to three different faces of ice on the same substrate. We have pinpointed three circumstances where heterogeneous ice nucleation can be promoted by the crystalline surface: i) the formation of a water overlayer that acts as an in-plane template; ii) the emergence of a contact layer buckled in an ice-like manner; and iii) nucleation on compact surfaces with very high interaction strength. We hope that this extensive systematic study will foster future experimental work aimed at testing the physiochemical understanding presented here.
3.1 Introduction

Recently, a number of excellent experimental works have succeeded in determining which materials can effectively promote heterogeneous ice nucleation, mostly by measuring ice nucleation temperatures or rates, see e.g. Refs. 20, 86–96. By doing so, the ice nucleating abilities of a large variety of materials has been characterized [11, 97]. This knowledge can for instance be used to decipher and explain the different contributions to ice nucleation in the atmosphere [6, 98–101]. However, experiments currently do not provide information into the molecular details of individual ice nucleation events. Because of the length scale involved (nm), insights into the nucleation process can be obtained instead from computer simulations. And indeed, in the last few years a handful of computational studies have been successful in simulating heterogeneous ice nucleation [31, 102–113]. This indicates that the time is now ripe for furthering our understanding of the microscopic factors that make a material a good ice nucleator. Nevertheless, even being able to explore heterogeneous ice nucleation with simulation approaches may not be enough to understand a priori whether and why a material will be a good ice nucleator or not. This is because many different ingredients like the morphology of the surface [114–117], its hydrophobicity [87, 108, 113, 118], local electric fields [90, 103, 119–121], preferential nucleation sites or surface roughness [107, 109, 122] can simultaneously impact on both the molecular mechanism and the resulting nucleation rate.

The two most discussed “requirements” for an effective ice nucleator are perhaps the crystallographic match with respect to bulk ice and the strength of the water-surface interaction. The former was introduced by Turnbull and Vonneugut [123] in order to characterize the catalytic potential of a surface regarding heterogeneous nucleation. If the atomic arrangements in the contact region are similar, a disregistry or lattice mismatch $\delta$ between the corresponding surface unit cells can be defined in a simplified manner as:

$$\delta = \frac{a_s - a_i}{a_i}$$

(3.1)
where $a_s$ and $a_i$ are the lattice parameters of the surface unit cells of the substrate and a certain face of ice. The idea of a small lattice mismatch $\delta$ being at the heart of the INA dates back to the 1940s, when the ice nucleating capabilities of AgI, featuring only $\delta \approx 2\%$ for the basal face, came to light [97, 124]. Even though both experiments [86, 125] and recent simulations [31, 110] have seriously challenged the validity of this concept and most importantly its generality with respect to other materials [115, 117], a small lattice mismatch is still considered as the primary attribute of an efficient ice nucleator. In the case of bacterial ice nucleating proteins ice-matching patterns have even been used as an *a priori* assumption to infer the three dimensional structure of the residues from the DNA sequence describing the protein [126, 127].

Concerning the water-surface interaction, or the hydrophobicity / hydrophilicity of a surface, in the last two decades a number of experimental studies investigating ice formation on soot [4, 11, 99, 128–131] have prompted a debate about whether a correlation exists between the hydrophilicity of carbonaceous surfaces and their efficacy as ice nucleators. This is a challenging issue, because in most cases the role of the hydrophobicity cannot be disentangled from the influence of the lattice mismatch and surface morphology. As an example, the oxidation of soot taking place in atmospheric aerosols modifies both the hydrophilicity and the morphology of the particles at the nanoscale [132]. Furthermore, Lupi and Molinero [108] found that an increase in hydrophilicity showed adverse effects when it was accomplished by adding OH groups as opposed to just increasing the water-surface interaction strength. And indeed, recent experiments by Whale et al. [96] provide some tentative support for this hypothesis. Cox et al. recently investigated the dependence of the ice nucleation rate as a function of hydrophilicity in the case of model nano-particles [112]. They found a similar interaction range for both a fcc and a graphene-like particle where nucleation is enhanced, leading to a rule-of-thumb for an optimal adsorption strength. They also showed [113] how a simple modification of the surface morphology could lead to a significant change of nucleation rates, demonstrating the potential of atomic-scale control of nucleation.
As far as we know, the interplay between the hydrophobicity and morphology of the surface has not been systematically studied at the molecular scale. In this chapter, we fill this shortfall by investigating ice formation on top of a generic fcc crystal as a function of both the strength of the water-surface interaction and the morphology, taking into account the (111), (100), (110) and (211) surfaces. Strikingly different nucleation scenarios emerge according to the balance between the morphology of the surface and its hydrophobicity, thus demonstrating that the lattice match alone cannot be deemed as the key player in promoting nucleation on crystalline surfaces. In addition, we have found that up to three different faces of ice can nucleate on top of the same surface, and that the microscopic motivation at the heart of the heterogeneous nucleation process is not unique, but actually changes according to both the water-surface interaction and the morphology of the surface. We propose three microscopic factors that lead to enhancement of the nucleation rates: i) the formation of a water overlayer that acts as an in-plane template; ii) the emergence of a contact layer buckled in an ice-like manner; and iii) enhanced nucleation on compact surfaces with very high adsorption energy.

The remainder of this chapter is organized as follows: Section 3.2 describes the computational setup (3.2.1) together with an assessment of finite-size effects (3.2.2). In section 3.3 we present the nucleation rates for all the different surfaces as a function of adsorption energy and lattice constant. From this data we shall extract and discuss the general trends that emerge (3.3.1). The following subsection 3.3.2 presents the three different scenarios we propose as driving forces behind the nucleation promotion. We then discuss further insight and future perspectives for improved heterogeneous ice nucleation simulations and experiments that could test the suggestions made here in subsection 3.3.3. Finally, the key results and observations are summarized in section 3.4.
3.2 Methods

3.2.1 System and Computational Methods

We considered slab models of crystalline surfaces covered in a water film (see Fig. 3.1a) including 4000 water molecules represented by the coarse-grained mW model [29]. This specific water model has excellent structural properties and a melting point close to experiment [29] but since it is monoatomic it exhibits faster dynamics which in turn allows for brute-force simulations of nucleation [108, 109, 112, 113, 133–135]. The water film is \( \sim 35 \, \text{Å} \) thick, which is enough so that the density is converged to the bulk homogeneous value at \( \sim 12 \, \text{Å} \) above the interface.

![Figure 3.1: a) Example of a simulation box used in a heterogeneous ice nucleation run. The coarse-grained water molecules are depicted as blue spheres while surface atoms are gray. The average box dimensions were \( 60 \times 60 \times 70 \, \text{Å}^3 \). b) Top and side view of the four crystalline surfaces considered. Atoms are colored according to their z-coordinate. Red boxes highlight the symmetry of the surface unit cells.](image)

We remark that in this study we do not aim at investigating specific systems like e.g. metallic surfaces, but instead we intend to extract general insight and useful trends from idealized model substrates. To this end we have taken into account four different crystallographic planes of a generic fcc crystal, namely the (111), (100), (110) and the (211) surfaces, which exhibit significant differences in terms of atomic roughness and the symmetry of the outer crystalline layer (see Fig. 3.1b).
For each of the above mentioned surfaces, we have built a dataset of ten different slabs varying the fcc lattice parameter \( a_{\text{fcc}} \) from 3.52 to 4.66 Å\(^1\). This range encompasses several fcc metals from Ni to Ag, including the well characterized Pt structure [138]. The interaction of the water with the substrate is given by a truncated Lennard-Jones potential:

\[
U(r) = \begin{cases} 
4\varepsilon \left[ \left( \frac{\sigma}{r} \right)^{12} - \left( \frac{\sigma}{r} \right)^6 \right] & r < r_c \\
0 & r \geq r_c
\end{cases}
\] (3.2)

where \( r \) is the distance between a water oxygen and a surface atom. The cutoff distance was set to \( r_c = 7.53 \) Å.

To measure the interaction strength of water with the surface the adsorption energy \( E_{\text{ads}} \) of a single water molecule was computed. In order to vary this quantity \( \varepsilon \) and \( \sigma \) were changed accordingly. \( E_{\text{ads}} \) was computed by minimizing the potential energy of a single water molecule on top of the surface. In this manner well defined adsorption energies can be determined for the (111), (100) and (110) surfaces since only one adsorption site is found by the minimization algorithm. However, for the (211) geometry multiple adsorption sites with considerable energy differences were found \(^2\). For this reason we have chosen to assign every \( (a_{\text{fcc}}, E_{\text{ads}}) \) combination for the (211) orientation the same \( (\varepsilon, \sigma) \) pair as for the (111) surface. This is also motivated by the (111) terrace exhibited by the (211) surface. The final adsorption energy for the (211) geometry as reported in Fig. 3.5 is the arithmetic average of the different adsorption energies found on this particular surface. The averaged results deviate by ca. 5 % from those for the (111) surface, e.g. the highest \( E_{\text{ads}} \) on (111) is around 12.76 kcal/mol while the average value for the (211) surface with the same \( (\varepsilon, \sigma) \) parameters is 13.18 kcal/mol. We note that the range of \( E_{\text{ads}} \) (from 0 to about 13 kcal/mol) we have considered is consistent with typical values obtained by DFT calculations of water monomers on metals [140].

---

\(^1\)All systems were built with the python.ASE [136] and VMD [137] software packages.

\(^2\)This result was not dependent and could not be avoided by different minimization algorithms (steepest descent, conjugate gradient and Hessian-free truncated Newton as implemented in LAMMPS [139]).
3.2. Methods

Figure 3.2: An illustration of how the nucleation induction time \( t_n \) is established by monitoring the change in the potential energy \( E_{\text{pot}} \) in blue. The green data shows the number of water molecules within the biggest ice-like cluster \( N_{\text{cls}} \) and that the jump in \( N_{\text{cls}} \) coincides with nucleation. The data refers to the (111) surface for \( E_{\text{ads}}=1.04 \) kcal/mol and \( a_{\text{fcc}} = 4.16 \text{ Å} \).

3.2.2 Obtaining Nucleation Rates

Heterogeneous ice nucleation events have been simulated by means of brute-force molecular dynamics (MD) simulations, employing the LAMMPS simulation package [139]. We follow a similar protocol to the one of Cox et al. [113]. A time step of 10 fs has been used with periodic boundary conditions in the xy-plane while sampling the NVT canonical ensemble with a chain of 10 Nosé-Hoover thermostats [42, 141] with a relaxation time of 0.5 ps. The positions of the surface atoms were fixed throughout the simulations. Every point of the \((a_{\text{fcc}}, E_{\text{ads}})\) grid corresponds to a specific configuration which has been equilibrated at 290 K for 170 ns. Then 15 uncorrelated (separated by at least 10 ns) snapshots have been selected from the resulting trajectories as starting points for production runs, after having instantaneously quenched the system from 290 to 205 K. Nucleation simulations were terminated 10 ns after a significant drop of the potential energy \((> 0.53 \text{ kcal/mol per water})\) was registered or if the simulation time exceeded 500 ns. In total, we report results from 6000 nucleation and 400 equilibration simulations.

The induction time \( t_n \) of a nucleation event has been detected by monitoring the drop in the potential energy \( E_{\text{pot}} \) of the system associated with the formation of a critical ice nucleus, as shown in Fig. 3.2. We have calculated \( t_n \) by fitting the
potential energy to:

\[ E_{\text{pot}}(t) = a + \frac{b}{1 + \exp[c(t - t_n)]} \]  \hspace{1cm} (3.3)

where \( t_n, a, b \) and \( c \) are fitting parameters. Due to the smoothness of the potential energy surface characterizing the mW model, crystal growth at the supercooling considered here (\( \sim 70 \) K) is extremely fast, resulting in a very sharp potential energy drop that takes place within - at most - 1 ns for all values of \( E_{\text{ads}} \) and \( a_{\text{fcc}} \) considered. Thus, the resulting value of \( t_n \) does not depend on the specific shape of the fitting-function. We thereby estimate the error associated with the calculation of \( t_n \) as \( \pm 1 \) ns. We also verified that no substantial discrepancy with respect to \( t_n \) can be observed by using other order parameters like e.g. the number \( N_{\text{cls}} \) of mW molecules in the biggest ice-like cluster \(^3\), as reported in Fig. 3.2.

From the \( t_n \) dataset, we obtain the nucleation rate via the survival probability \( P_{\text{liq}}(t) \) as outlined in section 2.4. Examples of \( P_{\text{liq}}(t) \) for two very different nucleation events can be found in the Appendix A.2. It is difficult to quantify the error in the nucleation rates from the fitting previously described. Instead, we have em-

\(^3\)In the definition of \( N_{\text{cls}} \) we label the \( i \)-th mW molecule as ice-like according to the order parameter \( \bar{q}_3(i) \) as described by Lechner and Dellago [142], using a radial cutoff of 3.2 \( \text{Å} \). Two ice-like molecules are considered connected up to the same cutoff. The distribution of the \( \bar{q}_3(i) \) parameter can be found in Appendix A.1 (Fig. A.1).
3.2. Methods

Figure 3.4: Nucleation rate constant $J/V$ as a function of surface area (or number of water molecules, see x-axis, top). The legend refers to bulk homogeneous nucleation (Homo), a free-standing slab (Homo$_{\text{VAC}}$) with two vacuum interfaces, nucleation on top of the (111) surface (high$_E$, $E_{\text{ads}}=12.7$ kcal/mol, $d_{\text{fcc}}=3.9$ Å), same as high$_E$ but with a water slab two times thicker (high$_{E,2x}$), and same as high$_E$ but for $E_{\text{ads}}=3.2$ kcal/mol where we see a hexagonal overlayer (Hex).

Employed the Jackknife resampling technique [143] to quantify the error associated with the finite number of simulations, and thus of induction time $t_n$, that we have taken into account to compute each nucleation rate. Jackknife resampling is particularly suitable with respect to e.g. the conventional bootstrap approach when dealing with small sets of data. Results are reported in Fig. 3.3. The number of simulations we have chosen allows for a fairly well converged value of the nucleation rate, although an error bar accounting for about 35% of the value has to be considered. We have chosen to estimate the error bars with respect to $J$ in the worst case scenario, namely for very mild enhancement of $J$ with respect to the homogeneous system for which very long tails in $P_{\text{liq}}(t)$ can be observed. It must be noted that the finite size of our $t_n$ dataset is the major source of error affecting the numerical accuracy of our nucleation rates. In fact, while the calculation of both $t_n$ and $P_{\text{liq}}(t)$ is basically error-free and finite size effects introduce a small systematic error, the long time tails of $P_{\text{liq}}(t)$ can seriously suffer from a small $t_n$ dataset because of the stochastic nature of the nucleation events.
Finite size effects must be thoroughly addressed when dealing with nucleation events. At first, we have calculated the homogeneous nucleation rate $J$ as a function of volume for different models containing 1000, 4000 and 9000 mW molecules. We have considered bulk liquid models as well as free-standing water slabs, in order to take into account the influence of the vacuum-water interface that we have in our slab models. The results are summarized in Fig. 3.4 and led us to choose 4000 mW molecules for our heterogeneous models. Given the fact that the heterogeneous ice nucleation rates reported in this chapter span three orders of magnitude according to the interplay between hydrophobicity and surface morphology, we can safely state that finite size effects have little impact on our results. For instance, we have verified that doubling the area of the (111) crystalline surface (and the number of water molecules as well) only introduces a discrepancy of about a factor two in the nucleation rate constants for $E_{\text{ads}}=3.21$ or $12.76$ kcal/mol ($a_{\text{fcc}}=3.90 \text{ Å}$). This is somehow expected because the strong supercooling, which should guarantee a relatively small critical nucleus size. Indeed, we have obtained an estimate of the critical nucleus for a specific case ((111) surface, $E_{\text{ads}}=1.04$ kcal/mol, $a_{\text{fcc}}=3.90 \text{ Å}$) from a committor analysis [80] based on the number $N_{\text{cls}}$ of mW molecules in the biggest ice-like cluster. This suggests a critical nucleus size of about only 50 mW molecules (see Appendix A.3). This number lies consistently in the range of literature estimates, e.g. 10 molecules at 180 K [134] and 85 [144] to 265 [145] molecules at 220 K.

3.3 Results

3.3.1 No Simple Trend for Nucleation Rates

The nucleation rates on the four surfaces considered are shown as bi-dimensional heat maps as a function of the lattice constant and adsorption energy in Fig. 3.5a. Regions in the 2D plots for which a strong enhancement of the nucleation rates is observed are sketched in Fig. 3.5b and snapshots of representative trajectories for all the classified regions can be found in Appendix A.4. Before even considering any

---

4The $10 \times 10$ mesh of obtained values for each surface was first interpolated from the non-regular to a regular grid via a 2D-spline interpolation. Then this picture was resized to an image with 20 times finer resolution with the lanczos3 algorithm as implemented in MATLAB [146, 147].
3.3. Results

Figure 3.5: a) Heat maps representing the values of ice nucleation rates on top of the four different surfaces considered, plotted as a function of the adsorption energy $E_{ads}$ and the lattice parameter $a_{fcc}$. The lattice mismatch $\delta$ on (111) is indicated below the graph. The values of the nucleation rate $J$ are reported as $\log_{10}(J/J_0)$, where $J_0$ refers to the homogeneous nucleation rate at the same temperature. b) Sketches of the different regions (white areas) in the $(E_{ads}, a_{fcc})$ space in which we observe a significant enhancement of the nucleation rate. We label each region according to the face of $I_h$ nucleating and growing on top of the surface (basal, prism or 2nd prism), together with an indication of what it is that enhances the nucleation. “temp”, “buck”, and “high $E$” refer to the in-plane template of the first overlayer, the ice-like buckling of the contact layer, and the nucleation for high adsorption energies on compact surfaces, as explained in section 3.3.2.

microscopic details of the water structure or nucleation processes, several general observations about the data shown in Fig. 3.5 can be made:

1. The substrates mostly do promote nucleation compared to homogeneous nucleation. On some surfaces enhancements of up to two orders of magnitude are seen for certain values of $a_{fcc}$ and $E_{ads}$. The measured induction times for these events correspond to the transient time rather than the actual nucleation time, as e.g. discussed by Aga et al. [148] and Peng et al. [149]. Therefore nucleation rates at the high end of the values reported should be seen as a lower bound rather than the actual rate.

2. Both $a_{fcc}$ and $E_{ads}$ do not influence nucleation on top of each surface in the same manner. Indeed, the interplay of these two parameters is different for each surface. For instance, variation in $E_{ads}$ for the (211) surface generally has little influence on the nucleation rate. However, on the (111) surface at certain
values of $a_{fcc}$ variation in $E_{ads}$ can have a very big impact on the nucleation rate.

3. The (111) and (110) surfaces promote ice nucleation over a much broader range than the (211) and (100) surfaces. It is worth noticing that surface symmetry alone is definitely not enough to account for such a difference. In fact, the (111) and (110) surfaces possess different symmetry (hexagonal and rectangular respectively, see Fig. 3.1 b), while the (110) and (100) surfaces, although showing completely different ice nucleation capabilities (Fig. 3.5a), have quite comparable surface symmetry (rectangular and square, Fig. 3.1 b). Further evidence for the non-unique role of surface symmetry is given by the fact that simple trends are hard to find even within the very same surface. For instance, the interplay between $a_{fcc}$ and $E_{ads}$ in the case of the (110) surface results in two different regions where nucleation is significantly boosted (see Fig. 3.5b).

4. There is no optimal value for $E_{ads}$. In fact it is surprising how insensitive the nucleation rate is to changes in $E_{ads}$ for some substrates such as the (110) and (211). A notable exception is the (111) surface for $a_{fcc} > 3.9$ Å. Our results here are consistent with the recent work of Cox et al. [112, 113] where an optimal value of $E_{ads}$ around 3 to 6 kcal/mol is found for a fcc(111) and a graphene nanoparticle. The broader range of $a_{fcc}$ and the results for other substrates however reveal that this trend does not hold for the different morphologies.

5. A common feature on all substrates is that the nucleation rate is inhibited for the lowest value of $E_{ads}$. For this adsorption energy the molecules face a hard wall which in turn could even hinder nucleation compared to the homogeneous case [31, 150–152]. By analyzing the distribution of pre-critical nuclei (see Appendix A.5) we find that these avoid the neighborhood of the surface for mentioned $E_{ads}$ range. The effect therefore can be rationalized in a smaller volume available for the nuclei to appear. This volume can be estimated by
the area affected by significant density perturbations due to the presence of the surface (see Fig. A.9). This kind of inhibition is unlikely to be visible in simulations or experiments where the ratio of water volume to contact area is much higher than in our case.

6. The lattice mismatch $\delta$ cannot be regarded as a requirement for a high INA. This issue specifically concerns the (111) substrate, because of its compatibility with the basal face of hexagonal ice $I_h$. We calculated $\delta$ according to equation 3.1, the lattice constant of ice $a_i \approx 4.51$ Å [97] and $a_s = \sqrt{3/2} \cdot a_{fcc}$. Therefore a value of $a_{fcc} = 3.68$ Å corresponds to a zero mismatch ($\delta = 0$) which is indicated in Fig. 3.5a. If $\delta \approx 0$ is the main requirement for enhanced ice nucleation, we would expect a distinct peak around the corresponding value of $a_{fcc}$. The results for nucleation rates however clearly show that this is not the case. Certainly, a small value of $\delta$ does promote nucleation for a wider range of $E_{ads}$, but for adsorption energies between 2 and 6 kcal/mol enhanced nucleation is observed for mismatches even beyond +10 %. We note that for $\delta < 0$ the drop of nucleation rates seems to start sooner than for $\delta > 0$, although the corresponding lattice constants lie somewhat outside of our considered range. This is consistent with all atom simulations which show that a mismatch $\delta$ slightly larger than 0 is more favorable [115]. Furthermore, Mithen and Sear [153] computed nucleation rates of a Lennard-Jones liquid in contact with a substrate and found the maximum close to, but larger than $\delta = 0$. Overall, our results suggest that an as-small-as-possible lattice mismatch is helpful to nucleation, but cannot be regarded as the most important requirement for high INA. For the other surfaces, the definition of disregistry $\delta$ is not as straightforward, because the substrates do not provide a clear template. In fact a strict definition of what can be regarded as similar or not similar is not part of the lattice mismatch theory. We have therefore restricted our discussion of the lattice mismatch to the (111) substrate.
Figure 3.6: Analysis of certain factors important to nucleation. Each row represents data obtained from a representative trajectory for events classified as “temp”, “buck” and both combined (“temp + buck”) mechanisms (see section 3.3.2). The first column depicts the density of water molecules above the surface after freezing (filled curves) and during equilibration before freezing (dashed black line). The second column shows side views and the third column snapshots viewed from above. In all cases the contact layer is colored red while higher layers are colored blue. For ease of visualization in the top view only part of the second layer is shown.

3.3.2 Microscopic Factors for Nucleation

It is unexpected that a simplistic model like the one used here can foster such diverse behavior. However, when we examine the water structures and nucleation processes in detail, general trends do emerge. We now discuss the key features important to nucleation.

In-Plane Template of the First Overlayer

The in-plane structure of the first water overlayer plays an important role in nucleation, because it can act as a template to higher layers. This is particularly evident on the (111) surface, which possesses a hexagonal symmetry compatible with the in-plane symmetry of the basal face of ice (honeycomb). Where nucleation is sig-
3.3. Results

significantly enhanced, we find that a hexagonal overlayer (HOL) of water molecules forms on top of the surface (Fig. 3.6, “temp”), rapidly inducing nucleation. The promotion-effect persists even when a significant number of defects, such as 4-, 5- or 7-membered rings appear within the HOL, as well as in the case of larger lattice mismatches $\delta > 0$ where the HOL is severely stretched. This indicates that the template does not have to be perfect to promote nucleation. The HOL rules the majority of nucleation processes on top of the (111) surface, where only the basal face has been observed to nucleate and grow (see Fig. 3.5b). However, contrary to the idealized bilayer structure of the basal face the overlayers observed here are mostly flat. Reduced buckling in the contact layer has been suggested in a number of studies on metals [154, 155]. The flat hexagonal structures identified here which precede nucleation indicate that a good template needs: (i) the right symmetry and (ii) the right intermolecular distances in the plane, but not necessarily the correct water molecule heights.

We have labeled nucleation events induced by this contact layer as “temp” rather than “hex” to stress that it is not exclusively the bi-layer template of the basal face, typically associated with the term HOL, but rather any possible overlayer compatible with a face of ice. An example of a different overlayer is found on the (110) surface (see Fig. A.7) compatible with the prism face of ice.

**Buckling of the First Overlayer**

Our results concerning the (110) and (211) surfaces suggest that different heights of atoms in the contact layer, termed buckling, is an important factor to enhanced nucleation. The difference between a flat and a buckled overlayer can be seen in the water density and the side views of selected trajectories, depicted in Fig. 3.6. The density for an event characterized by the “temp” mechanism has only a single spike representing the flat hexagonal contact layer. In contrast on the (110) substrate at large lattice constants (Fig. 3.6, “buck”), the first water overlayer is not ice-like but exhibits a pronounced buckling of the contact layer. The fact that we find this combination of a symmetrically unfavorable (and therefore non-templating) but buckled contact layer for many of the enhanced nucleation trajectories leads us to conclude
that the buckling in this case is the microscopic cause for the nucleation enhancement (labeled as “buck”).

As with the (110), the (211) geometry also leads to a significant enhancement of nucleation rates in specific regions. In addition, and quite surprisingly on this surface, nucleation and growth of three different faces of ice are observed. The three regimes roughly correspond to different values of $a_{\text{fcc}}$ (Fig. 3.7a). The (211) substrate has a rectangular in-plane symmetry, but it features (111) micro-facets (see Fig. 3.1 a). For small values of $a_{\text{fcc}}$ (Fig. 3.7a), the spacing between the steps allows for rows of hexagons to form on top of these terraces. This template has a symmetry consistent with the basal face of $I_h$ which in fact exclusively nucleates in this first regime. As an aside we note that the growth direction of the basal face is not exactly parallel to the surface normal of the (111) terraces, leading to the small angle mismatch shown in Fig. 3.7a. As we move on to larger lattice constants, the spacing between the steps becomes too large to accommodate a hexagonal overlayer. Rather a rectangular overlayer appears on top of the surface, wiping out the templating

Figure 3.7: (a) Representative snapshots of the three different faces (basal, prism and 2nd prism face) of hexagonal ice growing on top of the (211) surface (side view). Surface atoms are depicted as balls (gray), while the bonding network of water molecules is represented by sticks (blue). The $\theta$ angle in the top left panel illustrates that the basal face and the normal of the (111) terrace deviate. b) Nucleation rates (circles) and spline interpolation (line) on the (211) surface as a function of the step distance $d$. The red lines indicate the measured characteristic distances $d_1$ and $d_2$ as well as their standard deviation (red shaded area). The meaning of $d$, $d_1$ and $d_2$ is illustrated in the top panels.
effect of the hexagons. These overlayers are buckled in a manner that follows the
corrugation of the surface. This results in the nucleation and growth of the prism
and 2nd prism faces for $a_{\text{fcc}}$ values of 4.16 Å and 4.66 Å respectively (see Fig. 3.7a).
The contact layers, despite being significantly buckled, do not show a favorable in-
plane template (pictures of the overlayers can be found in Appendix A.4).

In the specific case of the (211) substrate $a_{\text{fcc}}$ is a much more sensitive pa-
rameter for the nucleation rate than $E_{\text{ads}}$ which only leads to changes for vanishing
interaction. This suggests that nucleation enhancement by the buckling of the over-
layer is a more geometrical phenomenon. Indeed, we find that the buckling of the
contact layer in these cases coincides with a characteristic periodic length of one
of the ice faces. To support this interpretation, in Fig. 3.7b the nucleation rates for
$E_{\text{ads}} = 6.38$ kcal/mol are displayed as a function of the step distance $d$. The charac-
teristic lengths for prism and 2nd prism face ($d_1$ and $d_2$ respectively) were obtained
from measuring and averaging the corresponding distances in representative trajec-
tories where we see freezing of that particular ice face. The values correspond well
with the step periodicity $d$ at which nucleation is enhanced the most through forma-
tion of the respective face. A similar conclusion was drawn by Zhang et al. [106]
for trenches promoting nucleation the most when they resemble a characteristic
spacing. These effects seem to fade when the roughness is on a larger than atomic
scale [156] or if the surface is amorphous [109].

The results shown in Fig. 3.7b are reminiscent of the predictions of Turnbull
and Vonnegut [123] regarding a small lattice mismatch. Indeed, if one neglects
the fact that the atomic arrangements of the substrate and respective ice face at
the interface are dissimilar, the buckling can be interpreted as a lattice mismatch.
However, this concept is unlikely to be helpful in general as it does not clearly
distinguish the two ingredients that form the buckling: (i) the different heights of
atoms that are adsorbed onto the surface and; (ii) the periodicity that describes the
variation of atomic heights. Contrary to the lattice mismatch, a compatible in-plane
template is not required for the buckling.

We also observed nucleation events in which the overlayer possessed both
atomic scale buckling and favorable in-plane template. In two specific $E_{\text{ads}}$ and $a_{\text{fcc}}$ intervals a buckled first overlayer displaying an in-plane template consistent with the basal or prism face respectively (labeled as “temp + buck”) forms on the (110) surface. The third row in Fig. 3.6 depicts the density and water overlayers in the case of basal face growth. Here, the structured water during equilibration exhibits an appearance that is already close to the double-peak of frozen (basal face) ice. The resulting overlayer consists of hexagonal arrangements, comparable to the basal face of ice - not only as an in-plane template, but also in the buckling. The importance of ice-like structuring along the z direction has been observed and discussed in the case of AgI [105]. Notably a HOL is not enough in this case, suggesting that ice-like buckling could be more effective than in-plane templating.

On the other hand, when neither the in-plane template or the favorable buckling are present, no sizable enhancement of the nucleation rate has been observed. This is what happens for the majority of the $(E_{\text{ads}}, a_{\text{fcc}})$ points on the (100) surface (see Fig. 3.5), which has a square symmetry and being perfectly flat does not cause the contact layer to buckle.

**High Adsorption-Energy Nucleation on Compact Surfaces**

We have also observed the promotion of nucleation in two regions where neither the ice-like in-plane template or buckling of the contact layer was present. The two regions can be found for the (111) and (100) surfaces (see Fig. 3.5) and have been labeled high$_E$ to emphasize that they occur only for the higher adsorption energies. It is also apparent that we find this kind of enhancement on the two compact surfaces rather than the more open ones, which suggests that it is the combination of strong interaction and surface denseness that facilitates the nucleation. The overlayers in these cases were very dense (a disordered overlayer for (111) and perfect squares for (100), see the Fig. A.4 and A.6). It is clear that these structures should be anything but advantageous for nucleation. The analysis of the distribution of precritical nuclei for a representative point (see Appendix A.5) reveals that nucleation happened on top of the first 2~3 water layers. It is therefore clearly a heterogeneous event which the increased rates already suggested. While the actual reason for this
kind of nucleation enhancement is not immediately obvious and potentially interesting, it must be noted that values of $E_{ads}$ in the upper third of the considered range are abstract, as water will probably dissociate on top of the surface rather than being adsorbed. Thus, we have not made further investigations concerning this specific enhancement, however we suggest two possible effects that could be the driving force behind it. First, a layering mechanism similar to the one discussed by Cox et al. [112] could influence higher layers when the coverage of the underlying layers is saturated. This is also supported by the values of layering we have calculated, as discussed later (3.3.3). A second reason for the facilitation could also be dynamical effects, which have been shown to significantly influence molecules and atoms near the interface [157]. The strong adsorption causes the first 1~2 layers to be nearly immobile, effectively extending the surface height and possibly shifting the dynamical effects to layers above $\sim 10$ Å. Lastly, we note that the effects of high nucleation could be shifted towards more realistic interactions for all atom-models of water, since in our tests with the TIP4P/2005 model we observed a slightly more pronounced structuring and layering (see Appendix A.6).

3.3.3 Further Insight and Future Perspective

Having examined heterogeneous ice nucleation on the four crystalline substrates and identified some of the key factors responsible for the enhanced nucleation observed, we now discuss a number of open issues and ways this work could be taken forward in the future.

Layering

Lupi et al. [108, 109] found that the layering $L$ of water at graphitic interfaces correlates with their freezing efficiency. For high adsorption energies Cox et al. [112] also found a correlation between the nucleation and the layering, but only if the layering associated with the contact layer was excluded ($L^*$). These two forms of layering are defined as follows:

$$L = \int_0^{\xi_{bulk}} \left| \frac{\rho(z)}{\rho_0} - 1 \right|^2 \, dz$$  \hfill (3.4)
where $\rho_0$ is the density of bulk water, $z_{\text{bulk}}$ is the height above the surface at which the water density $\rho$ converges to the bulk value and $z_0$ is a height so that the layering contributions of the contact layer are excluded. In all cases $z_{\text{bulk}} = 18$ Å was used.

The results are shown in Fig. 3.8.

Generally, a comparison between the layering plots and the nucleation rates in Fig. 3.5 shows that $L$ and $L^*$ do not correlate very well with the nucleation ability of the surface. We find that both $L$ and $L^*$ monotonically increase with $E_{\text{ads}}$. However, there seems to be a non-trivial dependency on $a_{\text{fcc}}$ where for 3 of the 4 surfaces the layering first increases towards medium values of the lattice constant and then decreases again. This trend reflects a change of the adsorption structure of the water molecules, which was also partially the cause of the different mechanisms influencing nucleation rates. However, the trends of the nucleation maps are not reproduced.

For instance, no region that has been classified as “temp” can be distinguished from its surrounding in the layering plots. If one assumes an optimal value or threshold for $L$ or $L^*$, the corresponding iso-surfaces in the plot would not reproduce any plot of the nucleation rates. While this conclusion has already started to emerge from the work of Cox et al. [112] we can now base the argumentation on a much larger parameter space and additional surfaces.

Interestingly, we find some exceptions. Specifically the areas classified as high $E_{nucleation}$ seem to be similar to areas of strong layering. This could also explain why we do not see this mechanism on the (110) and (211) surfaces, because the layering is still too weak even for the highest adsorption energies. Also the $L^*$ plot for the (211) substrate seems to indicate the regions that have been classified as “buck”. However no quantitative agreement is found, as for instance $L^* \approx 2$ on the (110) surface shapes a region where nucleation rates differ by an order of magnitude.

Overall, we find that the layering does not generally correlate with the nucleation ability which is likely due to the fact that this quantity averages over lateral
3.3. Results

Figure 3.8: Heat maps representing the total layering $L$ (top) and the layering excluding the contact layer $L^*$ (bottom) calculated from the equilibration runs. The dashed lines indicate the regions where nucleation was enhanced through high $E_{\text{nucleation}}$ (black) and buckling (red). Note that the color range for $L$ and $L^*$ is different.

Structure effects such as the in-plane symmetry and template. The exceptions are such cases where the potential energy surface is smooth, as for instance high $E_{\text{nucleation}}$ events where nucleation happens further away from the surface or nucleation on graphene-like surfaces [108, 112].

Notes on the Water Model

The fast dynamics associated with the coarse-grained mW model has made the current systematic study possible. However it is important to consider, at the very least, how the absence of explicit hydrogens affects the results of this study. To this end we have performed test calculations with the all-atom TIP4P/2005 model [158] which provides a reasonable description of water [159].

Firstly, we compared the water densities for one ($a_{\text{fcc}},E_{\text{ads}}$) point on each surface (see Fig. A.10). The densities obtained are very similar for both water models and we conclude that the structuring they show is nearly equivalent.

Secondly, a set of simulations at supercooling with the all-atom model was carried out on the (111) surface. Here, the almost instantaneous formation of a hexagonal overlayer was the driving force behind the nucleation enhancement for mW. With TIP4P/2005 we did not observe the complete formation of such an
overlayer within 100 ns. However, an analysis of the hexagonal cluster distribution (see Fig. A.11) shows that the largest patches of hexagons can be found for $E_{\text{ads}} \approx 3.2$ kcal/mol. This is precisely the value for which we observe the fastest formation in the case of mW as well. This trend confirms that while - especially heterogeneous - nucleation processes modeled by mW water are for certain non-physically fast, they can still capture part of the underlying physics.

Higher Temperatures

It is interesting to understand how the trends observed in this study could depend on temperature, especially because our simulations were performed in the deeply supercooled regime. It is currently beyond reach to carry out such an extensive set of simulations at a significantly higher temperature with the brute-force approach. However, to estimate the effect of the strong supercooling we performed further calculations at 210 and 215 K for 3 adsorption energies on the (110) surface (the results can be found in Appendix A.7). We find that the nucleation rates in regions where no specific mechanism has been attributed heavily decline, but otherwise no significant changes can be observed. That includes the trends of the nucleation rate as well as the adsorption structures, which are the basis for the mechanisms we propose. This indicates that our conclusions are also valid for higher temperatures.

Future Perspective and Experimental Verification

Before concluding we discuss some aspects that should be addressed in future studies as well as making some suggestions about how the insight presented here could be tested experimentally.

A first step will be to use all-atom models of water [159, 160] such as the TIP4P/2005 discussed above or its cousin TIP4P/Ice [57] specifically designed for the study of ice. Recently all-atom simulations of homogeneous ice nucleation have been performed [161] with the help of the forward-flux sampling technique [30, 75]. The latter seems like a promising approach for nucleation sim-

---

5To gauge interaction energies for different water models the heat of vaporization is often used. The latter for both models is nearly the same [29, 158] which means we can compare the adsorption energies directly.
3.3. Results

Simulations [111, 144, 150, 161–164] although there are of course many other free energy and enhanced sampling techniques [80–83, 165–169] that could be used. Improvement in the water-surface interaction potential is of equal importance if one wishes to investigate heterogeneous ice nucleation. For instance, an extension of the present study to realistic clean metal surfaces needs to account for the orientational dependence of the water molecules on the surface and polarization and charge effects [170, 171]. Fitting water-surface interaction potentials to density functional theory or higher-level electronic structure theories is one way to take such effects into account and work in our group in this direction is ongoing [140, 172–174]. In the attempt to take another step towards the modeling of realistic surfaces, the effect of surface vibrations, structural relaxation and surface reconstruction, together with the influence of surface defects must be taken into account, and will be the focus of future computational efforts. Furthermore, it has been shown that dissociation of water molecules occurs at reactive metal surfaces so that the overlayers can be comprised of water-hydroxyl mixtures [175–178]. Taking this issue into account will require a suitable and accurate dissociable model of water. Lastly, it will be important that nucleation studies approach experiments more closely. Especially the supercooling in computational studies is a major concern since it is too strong to directly allow for comparison with e.g. atmospheric or laboratory measurements.

Our results could be most directly probed by measurements that can reliably characterize surface structures with molecular level of resolution. This would require ultra-high vacuum prepared levels of cleanliness. A most promising candidate for an experimental study would be gold surfaces because of their resistance to oxidation and golds fcc crystal structure. With $a_{\text{fcc}} \approx 4.08 \, \text{Å}$ [179] and $E_{\text{ads}} \approx 3-7 \, \text{kcal/mol}$ [140] our simulations indicate that the nucleation rates on the (111) and (100) gold substrates should differ by 2-3 orders of magnitude. This has been estimated from the data in Fig. 3.5 in the region of the Au lattice constant and $E_{\text{ads}}$.

$^6$The actual difference in nucleation rates for Au(111) and Au(100) could be even larger, since the highest nucleation rates computed in our study are only lower bounds. Therefore a significant difference between these two substrates should also be detectable if one measures freezing temperatures.
particles that expose different facets. By adding molecules that are inactive for ice nucleation but selectively bind to the promoting facets of the particle the nucleation rate could be controlled. Indications of freezing in a well defined surface-science-style study could be obtained with e.g. ambient pressure x-ray photoelectron spectroscopy [180, 181] or surface x-ray diffraction [182]. Another class of interesting materials are halogenated graphene [183, 184] and graphane [185]. The functionalization of graphene with different atoms such as H, F, Cl, Br or I should alter the underlying geometry of the 2D material only slightly [186, 187], but the water-surface interaction could greatly vary [188]. This could be exploited to verify our predictions for different interaction strengths by examining ice nucleation on these compounds. This idea can even be extended to other quasi-2D honeycomb materials such as silicene [189], germanene [190] and stanene [191] that have different lattice constants [188, 192] if grown on appropriate supports and if they remain stable in an aqueous environment. In such a manner the interplay between morphology and hydrophobicity could be examined experimentally, possibly yielding a similar nucleation map to Fig. 3.5a. Moreover, self assembled monolayers [193–195] provide the possibility to create specific morphologies. For instance different headgroups for aliphatic chains can alter the hydrophobicity of the resulting surface, while functional groups in the chain can change the spacing between them. Additionally, different chain lengths could be used to design a buckled surface. In combination with non hydrogen-bonding headgroups this could enable the design of interfaces useful for testing the buckling mechanism. Finally, we note that the exciting capabilities of femtosecond x-ray scattering [23, 24] techniques that have recently been used to explore homogeneous ice formation in water droplets could possibly be extended to heterogeneous systems.

3.4 Conclusions

In summary, we have examined the interplay between surface morphology and hydrophobicity on the ability of a generic crystalline surface to promote ice nucleation. We have calculated the nucleation rates of a coarse grained model of water on top
of four different crystalline surfaces of an ideal fcc crystal by means of brute-force molecular dynamics simulations, sweeping a comprehensive range of adsorption energies and lattice parameters.

Strikingly different nucleation scenarios have emerged on the various crystalline surfaces considered. Even for a specific surface the balance between lattice constant and hydrophobicity fosters non trivial trends. Most surprisingly the nucleation and growth of up to three different faces of hexagonal ice on top of the same surface could be induced by altering the lattice parameter alone.

We have demonstrated that on the (111) surface a small lattice mismatch with respect to ice is certainly not a requirement for promoting ice nucleation. This implies that in the search for understanding of the nucleation performance of known materials or the design of new ones one should not exclusively focus on the lattice mismatch issue. Nonetheless, our results show that it is important which surface is present, as nucleation rates can vary from inhibition to promotion for different faces of the same material. This means that experiments have to carefully characterize the atomic structure of ice nucleators, because the sheer morphological difference in samples could account for varying nucleation rates. Additionally, this provides exciting possibilities to change the ice nucleation behavior of materials through e.g. growth-habit control [196] to strengthen the inhibition effect or to turn nucleating nano particles into inhibitors and vice versa.

In most cases nucleation is promoted in a wide range of $E_{\text{ads}}$ without changing the molecular mechanism. Therefore, optimal interaction strengths are rare exceptions and only found for some specific $a_{\text{fcc}}$ ranges.

Finally, we have pinpointed three different scenarios that facilitate the nucleation process.

1. The ability of the surface to create a first water overlayer that provides an in-plane template consistent with one of the faces of ice. Such an overlayer is typically, but not exclusively found on top of a surface that already displays a compatible symmetry.

2. The ability of the surface to structure the first two water overlayers in such a
way that they resemble either the density profile perpendicular to the surface or a characteristic buckling distance in the surface plane of one of the faces of ice. This typically requires a certain roughness at the atomic scale.

3. Even in the case of a first overlayer lacking both an in-plane template and structuring, nucleation can be promoted within the higher water layers. This kind of enhancement requires a compact surface with high adsorption energy.

Whether or not one of these scenarios could take place on top of a given surface, depends in a non trivial manner on both the morphology and hydrophobicity of the surface. Such a large body of findings will hopefully encourage and guide future work addressing heterogeneous ice nucleation on top of realistic surfaces, in the hope of furthering our understanding of what makes a material a good ice nucleator.
Chapter 4

Role of Pre-critical Fluctuations and Metastable Polymorphs in Heterogeneous Ice Nucleation

Polymorphism plays a great role in nucleation and its application, e.g. for pharmaceuticals. The role of polymorphs in heterogeneous nucleation is generally less well understood than e.g. for the homogeneous case. It is unclear whether and when there can be differences in the nucleation mechanism, e.g. if they are already in the earliest occurrences of crystalline clusters. Here we investigate through molecular dynamics simulations how the formation of pre-critical crystalline clusters is connected to the kinetics of nucleation. Considering heterogeneous water freezing as a prototype of relevance, we find that pre-critical fluctuations connote which crystalline polymorph will form. The emergence of metastable phases can thus be promoted by templating crystal faces characteristic of specific polymorphs. As a consequence heterogeneous classical nucleation theory cannot describe our simulation results, because the different substrates lead to the formation of different ice polytypes. We discuss how the issue of polymorphism needs to be incorporated into analysis and comparison of heterogeneous and homogeneous nucleation. Our results will help to interpret and analyze the growing number of experiments and simulations dealing with crystal polymorph selection.
4.1 Introduction

Despite its flaws [25, 68], classical nucleation theory [61, 64] (CNT) provides a qualitative understanding of nucleation and due to its simplicity is still the most widely used theoretical framework to interpret experiments and simulations. In Fig. 2.2 we showed the free energy profiles predicted by hetCNT. If we arrange the CNT equations discussed in section 2.3 (and line-tension effects are assumed negligible) we can obtain an expression for the enhancement factor \( f_V \) which is independent of the nucleus size \( n \) [69] and we can express \( f_V \) in terms of the heterogeneous and homogeneous free energies:

\[
(f_V)^{1/3} = \frac{F_{\text{het}}(n)}{F_{\text{hom}}(n)} \left[ 1 - \frac{2}{3} \left( \frac{n}{n_c} \right)^{1/3} \right] + \frac{2}{3} \left( \frac{n}{n_c} \right)^{1/3} \]  

(4.1)

Interestingly, equation 4.1 requires only knowledge of the homogeneous \( n_c \), as the free energies \( F(n) \) can be obtained for arbitrary cluster sizes \( n \). Knowing the value of \( f_V \) for a given substrate is fundamental as it encodes all information about the nucleation enhancement which is reflected in the fact that all the curves in Fig. 2.2 retain the same functional shape and the steepness ratio

\[
\chi(f_V) = \frac{\Delta F(f_V)}{n_c(f_V)} = \frac{f_V \cdot \Delta F_{\text{hom}}}{f_V \cdot n_{c,\text{hom}}} = \frac{\Delta F_{\text{hom}}}{n_{c,\text{hom}}} = \chi
\]

(4.2)

is independent of the enhancement. This has been key in several nucleation studies [153, 197, 198], on evaluating the performance of CNT [199, 200], or as a bridge between atmospheric cloud models and the microscopic description of ice nucleation [201, 202]. It is however tremendously challenging to measure \( f_V \), since knowledge of difficult to obtain quantities like \( \Delta F \) is needed. In principle one has easier access to pre-critical quantities, such as the probability \( P(n) \) of finding a cluster of size \( n < n_c \) in the liquid state. Since \( P_{\text{het}}(n)/P_{\text{hom}}(n) \propto F_{\text{het}}(n)/F_{\text{hom}}(n) \), equation 4.1 could be evaluated from the statistics of pre-critical clusters (termed pre-critical fluctuations) and thus without having to observe the rare nucleation event itself. Although many aspects of nucleation have been studied in great detail, the role of pre-critical fluctuations in heterogeneous nucleation is less well understood.
However, deeper understanding could potentially be exploited to gain insight into fundamental aspects of heterogeneous crystal nucleation.

In this chapter we aim to understand pre-critical fluctuations and their connection to nucleation kinetics by comparing cluster fluctuations on two model substrates that enhance the nucleation of ice to the same extent. To this end we perform an extensive set of molecular dynamics simulations. From this it emerges that the traditional hetCNT picture can break down because a substrate can facilitate the formation of different polymorphs. As a consequence, when using hetCNT one must choose a different bulk-reference to describe the nucleation process correctly. Although here, we illustrate the potential role of pre-critical fluctuations in the context of CNT, in principle they could be used with any theory that provides a free energy profile for nucleation. We hope that the new insight obtained furthers the theoretical understanding of heterogeneous nucleation, polymorph selection and the role of pre-critical fluctuations.

4.2 Methods

4.2.1 Unbiased Molecular Dynamics

We performed molecular dynamics simulations of heterogeneous ice nucleation with 18000 water molecules, represented by the coarse-grained mW model [29], a model that is widely used to study phenomena involving water [109, 111, 162, 200, 203–207]. The water molecules are placed in a film geometry on top of two pristine, rigid fcc surfaces (termed s1 and s2). See additional details about the structures and computational methods in the Appendix B.1. These surfaces have proven useful in disentangling the contributions of lattice match and hydrophobicity to heterogeneous ice nucleation [208] (see chapter 3). The two examples presented here have been selected because of their striking difference in pre-critical fluctuations. While not aiming at representing any specific material, they are most similar to metal-water interfaces [178, 209]. The substrate-water interaction is given by a Lennard-Jones potential tuned to achieve the same absolute nucleation rate (see Tab. 4.1 for interaction parameters). Following established protocols [113, 208] we first equili-
brane each structure for 10 ns at 300 K. Then, production runs are quenched to the
target temperature and coupled to a 10-fold Nosé-Hoover chain [42] to sample the
NVT ensemble, integrating the equations of motion with a timestep of 10 fs. The
relaxation time after the quench is on the order of 10 ps and can thus be considered
non-disturbing to the nucleation. The nucleation events themselves are detected
by a sudden drop in the potential energy, upon which we terminate the computa-
tion and collect the current time as induction time. 100 simulations for each of
the two substrates at 218 K and 50 simulations at 221 K have been performed with
LAMMPS [139]. From the collection of induction times we fit the survival probabil-
ity, as was discussed in section 2.4. We also made use of the kinetic reconstruction
of the free energy as discussed in there.

### 4.2.2 Identification of Ice

Ice-like molecules were detected using an order parameter according to Li et
al. [144] as implemented in PLUMED2 [210, 211]. First we compute for each
molecule $i$ the quantity:

$$l_{qlm}(i) = \frac{1}{N_b(i)} \sum_{k=1}^{N_b(i)} Y_{lm}(\theta_{ik}, \phi_{ik})$$

(4.3)

where the sum goes over the $N_b(i)$ neighbors of molecule $i$, $Y_{lm}$ are spherical har-
monics and $\theta_{ik}$ and $\phi_{ik}$ are the relative orientational angles between the molecule $i$
and $k$. For a given $l$ we then compute the quantity for all possible values of $m$ and
store them in a vector $\vec{l}_{ql}(i)$ containing $2l + 1$ components. Finally we calculate
values $l_{ql}$ according to:

$$l_{ql}(i) = \frac{1}{N_b(i)} \sum_{k=1}^{N_b(i)} \frac{\vec{l}_{ql}(i) \cdot \vec{l}_{ql}(k)}{|\vec{l}_{ql}(i)| \cdot |\vec{l}_{ql}(k)|}$$

(4.4)
For the particular choice of $l = 3$ the values of $l q_3$ can distinguish both between solid and liquid molecules as well as between cubic and hexagonal ice. For values of $l q_3 < -0.69$ we classify a molecule as ice-like. Additionally if $l q_3 < -0.85$ the molecule belongs to cubic ice and otherwise to hexagonal ice (the distribution of $l q_3$ for different water phases can be found elsewhere [144]).

4.2.3 Metadynamics Simulations
We performed well-tempered metadynamics simulations [82, 84] with 20 walkers [85] at 235 K. We employed smaller simulation boxes with 8000 molecules in a $60 \times 60 \times 78$ Å cell with 3D periodic boundary conditions. Nucleation was facilitated by biasing the path variables [212] $s$ and $z$ constructed by measuring the generalized distances of the systems permutation invariant vector [213, 214] (PIV) to two reference states (liquid and frozen simulation cells obtained from brute force MD at 205 K). In essence, the PIV is the vector of irreducible adjacency matrix entries ordered by magnitude, where entries decay smoothly from one to zero for intermolecular distances beyond 3.4 Å. The metadynamics parameters were Gaussian height $\delta = 0.2$ kcal/mol, Gaussian width $(\sigma_s, \sigma_z) = (0.022, 0.38)$, deposition stride 2 ps and a biasfactor of 50. We employ a repulsive wall at $s = 1.5$ to restrain our simulation to cluster sizes relevant to nucleation. The resulting free energy profiles were checked for convergence by reweighting [215] to the one-dimensional free energy profile $F(s)$ (see Appendix B.2). The critical cluster sizes were obtained via a committor analysis [80] seeded from the metadynamics trajectories (see also Appendix B.2).

4.3 Results
4.3.1 Identical Rates Despite Different Pre-Critical Fluctuations
The results for the ice nucleation rates of supercooled water in contact with two model substrates (termed s1 and s2) are summarized in Tab. 4.2. Because nucleation rates usually differ by many orders of magnitude we can label the resulting rates from our systems as essentially identical. The presence of the substrates compared to the homogeneous case increases the nucleation rate by many orders of mag-
Chapter 4. Pre-Critical Fluctuations in Heterogeneous Ice Nucleation

Table 4.2: Computed nucleation rates $J$ for the two systems $s_1$ and $s_2$ at two temperatures:
For the sake of comparison we have normalized the rates by the water-substrate contact area. The error in the rate has been estimated by Jackknife resampling.

<table>
<thead>
<tr>
<th>Temperature</th>
<th>System</th>
<th>$s_1$</th>
<th>$s_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$218$ $K$</td>
<td>$J_{218K}$ $[ns^{-1}Å^{-2}] \times 10^{-6}$</td>
<td>$(7.18 \pm 1.30)$</td>
<td>$(4.23 \pm 0.71)$</td>
</tr>
<tr>
<td></td>
<td>$J_{221K}$ $[ns^{-1}Å^{-2}] \times 10^{-7}$</td>
<td>$(2.82 \pm 1.02)$</td>
<td>$(2.88 \pm 0.77)$</td>
</tr>
</tbody>
</table>

Pre-critical fluctuations in $s_1$ and $s_2$ and the homogeneous case. In Fig. 4.1a we plot the size distribution of the biggest ice-like cluster that can be found in each snapshot from the trajectories. The probability densities for the homogeneous system and $s_2$ line up almost exactly, while the probability to spontaneously form large clusters is clearly enhanced in $s_1$. Furthermore, we plot the spatial probability distribution of these pre-critical clusters in Fig. 4.1b. For $s_1$ there is clearly a large peak near the surface. In contrast the same density for $s_2$ is nearly flat, which means - contrary to the expectations of hetCNT - fluctuations are not enhanced near the substrate. We checked that no permanent structures are detected as pre-critical clusters, confirming for example that the peak for $s_1$ stems from small ice-like patches that are continuously fluctuating rather than from a permanent ice-like overlayer. The difference between the two systems is also reflected in the asphericity of the nuclei formed where $s_1$ tends to form flatter clusters than $s_2$ (see Appendix B.4). We verified that these findings hold for the same order parameter but with the first hydration shell included in the cluster-definition, as well as for a different choice of underlying spherical harmonics (details in Appendix B.5). The difference in pre-critical fluctuations is surprising since the nucleation rate in both systems is effectively the same. In the framework of hetCNT this would imply that both substrates should have the same free energy profile and steepness ratio $\chi$ (equation 4.2), which is incompatible with the differences in pre-critical cluster sizes we observe. This shows that pre-critical fluctuations and hetCNT as commonly applied...
do not adequately describe the behavior observed on the two surfaces studied.

Figure 4.1: Cluster formation and free energy of ice formation for the two substrates at 218 K: a) Probability distribution of the size of the biggest ice-like cluster $N_{\text{cls}}$. The inset shows the average composition of pre-critical clusters within 5 Å of the surface. b) Probability distribution (blue) for the z component of the center of mass of ∼100000 pre-critical clusters. Positions of critical clusters are indicated in red. The water density is also indicated by dashed black lines (arbitrary scale). c) Free energy profiles as a function of $N_{\text{cls}}$.

4.3.2 Different Polymorphs Yield Different Free Energy Profiles

To understand why the fluctuations are so different on the two substrates we now look at the free energy profiles for nucleation. These are shown in Fig. 4.1c where it
can be seen that the curve for s2 follows the homogeneous one more closely than s1 does. The fact that the curves for s2 and hom. are similar is consistent with the fact that we find essentially the same cluster size distribution in the two. The free energy profile for s1 is flatter and therefore the formation of larger clusters is less costly compared to s2 and hom. It is also clear that in disagreement with hetCNT the ratio $\chi_{s1} \neq \chi_{s2}$ (see equation 4.2). The results for 221 K can be found in Appendix B.6 and show essentially the same features. Furthermore, a reconstruction of $F$ as a function of the CNT coordinate $n$ reveals similar differences, particularly in the steepness of the free energy paths. If hetCNT was adequate for our observation the functional shape of s1 and s2 should be identical (as illustrated in fig. 2.2). But for instance $\Delta F$ and $n_c$ are not being scaled by a single factor which suggests that at least one additional degree of freedom might enter in heterogeneous nucleation. In what follows we show that the polymorph, which can be influenced by the substrate, can account for this.

We now try to understand why we see different functional forms for the free energy curves. In line with other studies [109, 111, 200] we find that for s1 stacking-disordered ice $I_{sd}$ has formed, where the basal face of ice is in contact with the substrate. This is also the type of ice that forms in homogeneous simulations [134, 144, 161, 162]. The shape of the free energy profile for s1 and the enhanced pre-critical fluctuations compared to the homogeneous case can thus be explained by traditional hetCNT. In contrast, s2 forms a crystal face (primary prism) in contact with the surface that is only found in hexagonal ice $I_h$ but not in $I_c$. Thus $I_c$ layers cannot grow on top of that and subsequently the stacking-disorder, which usually leads to the formation of $I_{sd}$, is strongly disfavored at the surface (illustrated in Fig. 4.2d). The traditional hetCNT does not work for that because the nucleus on s2 is purely hexagonal and therefore not related to the homogeneous one that traditional hetCNT chooses as reference (stacking-disordered). Hence, it is no surprise that the shape of the free energy profile for s2 is different from s1, despite their identical nucleation rates. The pre-critical fluctuations in s2 appear unaltered compared to the homogeneous case, not because the substrate has no impact but
rather because they are fluctuations of a different polytype and this comparison is ill-defined. This can be seen in the inset in Fig. 4.1a where we show that an average pre-critical cluster within 5 Å of the surface in s1 is stacking disordered with a cubicity (fraction of Ic) of \( \sim 60\% \), while s2 forms 90% pure \( I_h \) clusters. The apparent \( \sim 10\% \) of \( I_c \)-like molecules in s2 are due to uncertainties in classifying interfacial molecules at the edge of the cluster. We have visually verified that, in contrast to s1, in s2 we never observe clusters near the substrate that are in their core stacking disordered (see also Fig. 4.2c). Although the statistics we have for critical clusters are worse, we note that on each surface their composition was nearly identical to the one of the respective pre-critical clusters. This further suggests that there is a causal connection between critical and pre-critical clusters.

### 4.3.3 The Polymorphic Differences Remain At Higher Temperature

To understand if our findings hold at higher temperatures we performed metadynamics simulations in our two systems at an elevated temperature of 235 K. This is around the highest temperature we can aim to study with our system size (as we expect the homogeneous critical cluster size to be \( \approx 600 \) molecules [111]). In Fig. 4.2a we show the free energy profiles obtained, where we note that the variable \( s \) describes the path from a liquid (\( s \approx 1.1 \)) to a frozen (\( s \approx 1.9 \)) simulation cell. We have employed an artificial soft wall at \( s = 1.5 \) to aid convergence for the region describing cluster sizes relevant to nucleation rather than growth. From these simulations we obtain (details in Appendix B.2) a free energy barrier on s1 of \( 204 \pm 5 \) k\( _B \)T and on s2 of \( 227 \pm 5 \) k\( _B \)T and critical cluster size on s1 of \( 211 \pm 11 \) and on s2 of \( 104 \pm 3 \). Finding that \( \Delta F_{s1} < \Delta F_{s2} \) and \( n_{c,s1} > n_{c,s2} \) is entirely consistent with the trends obtained at lower temperatures. In addition it can be seen from Fig. 4.2b that the polytype of ice formed in s1 and s2 is not the same, the former being \( \approx 55\% \) stacking-disordered and the latter being almost purely hexagonal. The deviations for smaller clusters are once again artifacts of the local order parameter employed at the cluster interface, where the classification is ambiguous. To illustrate the difference in the cluster cores, Fig. 4.2c shows representative snapshots for the critical
clusters in $s_1$ and $s_2$. We highlight in there hexagonal and double-diamond cages, the building blocks of $I_h$ and $I_c$ [161] that are a stronger topological feature than the local order parameter. In panel d) of Fig. 4.2 we illustrate that the substrate in $s_2$ avoids the stacking-disorder by stacking ice double-layers perpendicular to the surface which is a result of the crystal face (prism) in contact with the surface. We note in passing that this could be a general recipe for water and other tetrahedral liquids (e.g. group-IV elements or silica) and could also be exploited to design surfaces
that nucleate pure cubic ice. Overall, the findings for the higher temperature agree
with the simulations at lower temperature, suggesting that our reasoning also holds
for situations where pre-critical and critical clusters are separated by more than one
order of magnitude in size.

4.4 Discussion

We now try to place the results of this chapter in a broader context and discuss
some of the implications of our findings. The first consequence drawn from the
possible occurrence of different polymorphs is that the fundamental result of het-
CNT that reads $n_{c, \text{het}} = f_V n_{c, \text{hom}}$ and $\Delta F_{\text{het}} = f_V \Delta F_{\text{hom}}$ is not true for cases where
the substrate promotes the formation of a polymorph different than the one that is
formed homogeneously. This is because the enhancement factor $f_V$ is only properly
defined if the heterogeneous quantity it describes refers to the homogeneous refer-
ence of that polymorph. In general when thinking about heterogeneous nucleation
there are three possible ways to account for the enhancement factor: (i) An expres-
sion in terms of a shape factor $f_V = \frac{V_{\text{het}}}{V_{\text{hom}}}$; (ii) An expression in terms of a nucleus
factor $f_N = \frac{n_{c, \text{het}}}{n_{c, \text{hom}}}$; and (iii) an expression in terms of a potency factor $f_P = \frac{\Delta F_{\text{het}}}{\Delta F_{\text{hom}}}$. These three definitions are equivalent under the assumption that they describe events
where the same polymorph has been formed. However, if different polymorphs are
compared the concept of the enhancement factor becomes ill-defined. We derive in
the Appendix B.7 correction factors in the framework of hetCNT that account for
this change. The fact that increasing the temperature accentuated the difference in
the free energy profiles observed on the two substrates is an indication that effects
like line tension [216] and cluster asphericity [25] are not the main reason for our
observation (as those likely decrease with increasing temperature / increasing clus-
ter sizes), but rather it is caused by the different polymorphs. Hence, we believe
that the polymorph is a separate issue that should be taken into account in a com-
prehensive (heterogeneous) nucleation theory, in addition to known shortcomings
of CNT or its heterogeneous extension. We speculate that for the same reason the
colour could even be the most relevant deviation from hetCNT at high temper-
Another implication of this work is that pre-critical fluctuations are comparable for different substrates only if compared to the correct homogeneous fluctuations of their corresponding polymorph. In our study, the comparison of the pre-critical fluctuations of s1 and s2 with the homogeneous case would have resulted in the conclusion that s1 enhances the nucleation and s2 does not (from Fig. 4.1a), while they actually lead to nearly identical enhancement. We have illustrated this in Fig. 4.3a-b where we draw homogeneous and heterogeneous nucleation profiles for two different polymorphs. The gray shaded area and how far it stretches on the x-axis illustrates what cluster sizes can be reached through thermal fluctuations. This
ultimately determines the extent of pre-critical fluctuations and is very different for the two polymorphs as a result of their different homogeneous free energy profiles. Upon comparing to the homogeneous nucleation of a single (homogeneously dominant) polymorph (which without loss of generality we assume to be hom,1 in Fig. 4.3c) the apparent discrepancy becomes clear.

In summary, we have presented a comparative study of molecular dynamics simulations of heterogeneous ice nucleation on two distinct model-substrates. It was shown that, in disagreement with hetCNT as traditionally applied, their pre-critical fluctuations can differ substantially, and yet identical nucleation rates are obtained, which we attribute to the formation of different polytypes. From this we draw the following conclusions:

1. Substrates can promote the formation of metastable phases by templating crystal faces that are unique to the respective polymorph. This is an extension of the rationale applied in experimental studies where iso-structural templates are used [217, 218] since the substrate does not require the same structure but rather any structure that nucleates the right crystal face. In particular, for materials with different stackings (e.g. ice, group-IV elements or silicates), the templating of faces so that the stacking is perpendicular to the surface normal seems most promising and can avoid stacking-disorder.

2. Traditionally applied hetCNT can break down when a polymorph nucleates that is not the dominant homogeneous polymorph. This should be corrected for by choosing the right bulk reference, an aspect that is largely disregarded in the nucleation literature.

3. We hypothesize that the extent of pre-critical fluctuations carries information about the enhancement of nucleation and can serve as an early indicator of which polymorph will form. This provides a possible route to efficiently rank the nucleation ability of substrates and will be the subject of future investigation. Since pre-critical fluctuations are less sensitive to finite size effects than approaches in which the full nucleation path is examined, they could prove
particularly useful for studying nucleation at very low levels of supercooling, provided one pays attention to the comparability of systems (same contact area, temperature etc.).

While these implications have arisen from simulations of heterogeneous ice nucleation, it is clear that they are general to the phenomenon of heterogeneous nucleation and particularly relevant to the description of materials that display polymorphism, such as e.g. alumina [219], silicon [218], xenon [220], n-alkane [221] or the epilepsy drug carbamazepine [217]. A quantitative treatment of nucleation in these systems such as comparing the enhancement of nucleation to the homogeneous case requires information about the polymorph that formed. If the polymorph is disregarded this could lead to false inferences about nucleation rates, mechanisms or the accuracy of CNT and might also cause widely used CNT-based models that use a single enhancement factor $f_V$ as a parameter to appear inadequate. Further studies aimed at understanding and potentially exploiting pre-critical fluctuations are needed.
Chapter 5

Influence of Liquid Dynamics on the Homogeneous Ice Nucleation Process

Generally, when an ice crystal is born from liquid water two key changes occur: (i) the molecules order; and (ii) the mobility of the molecules drops as they adopt their lattice positions. Most research on ice nucleation (and crystallization in general) has focused on understanding the former with less attention paid to the latter. However, supercooled water exhibits fascinating and complex dynamical behavior, most notably dynamical heterogeneity (DH), a phenomenon where spatially separated domains of relatively mobile and immobile particles coexist [222]. The structural origins of these distinct dynamical regions has been studied [223–227], but the microscopic connection between DH and the nucleation event itself has yet to be examined directly for water or, indeed, for any liquid. Here we address this issue on the basis of computer simulations which reveal that: (i) ice nucleation occurs in low-mobility regions of the liquid; (ii) there is a dynamical incubation period in which the mobility of the molecules drops prior to any ice-like ordering; and (iii) ice-like clusters cause arrested dynamics in surrounding water molecules. With this we establish a clear connection between dynamics and homogeneous nucleation and anticipate that our findings are a starting point for the examination of the role of dynamical heterogeneities in heterogeneous and solution-based nucleation.
5.1 Introduction

Considerable effort has gone into understanding supercooled water and the microscopic details of ice nucleation [23, 228–230]. In particular computer simulations (either atomistic or coarse-grained) have proved to be of great importance in understanding ice nucleation recently. They have for instance shown that the nucleation of hexagonal ice I\textsubscript{h} proceeds through a polytype of ice known as stacking-disordered ice I\textsubscript{sd} [144, 161, 162, 203] and that this polytype is entropically stabilized at cluster sizes relevant to nucleation [231]. Simulations have also looked at dynamical heterogeneities in supercooled liquid water. It was suggested that there is more tetrahedrality in immobile regions [223–225] and that 5-membered rings [226] act as locally favored structures. While these works have not focused on the evolution of molecular mobility during and before the nucleation event the pre-ordering in less mobile domains could be seen as \textit{a priori} argument for nucleation preferring those regions [232]. On the other hand Mazza \textit{et al.} [233] found that rotational and translational heterogeneities correlate in water, and thus, the transformative motions of the rearrangements necessary for crystallization could be hindered in strongly immobile regions. Indeed, it has been argued for metallic liquids [234] that the enhanced mobility near the surface can explain surface-induced crystallization rather than the traditional view of a heterogeneous reduction of the nucleation barrier. Mobile molecules would have the potential for a more effective exploration of phase-space and therefore the ability to more readily perform collective self-assembly. Overall, one can find arguments for both immobile and mobile regions being preferential domains for nucleation. But whether there is any preference at all is not established to date for water, or indeed any other supercooled liquid, as the liquid particle mobility immediately before and during crystallization has not been examined on the molecular scale.
5.2 Methods

5.2.1 Molecular Dynamics Simulations

We mainly study the DH of a system containing 10,000 water molecules, represented by the TIP4P/Ice model [57]. All our MD simulations are performed with the LAMMPS code [139], integrating the equations of motion with a 2 fs time-step and using a 10-fold Nosé-Hoover chain [42] with a relaxation time of 200 fs to control temperature. We employ a cubic simulation box with 3-dimensional periodic boundary conditions and approximate volume of \(68 \times 68 \times 68 \, \text{Å}^3\). Static bonds and angles have been constrained with the SHAKE algorithm [58]. To avoid quenching effects upon generating starting configurations at different temperatures we performed (after 10 ns equilibration at melting temperature) a 0.5 K/ns cooling ramp in the NPT ensemble (ten-fold Nosé-Hoover chain barostat with relaxation time of 2 ps). At 273, 260, 250, 240, 230, 220 and 210 K we save configurations. Those are propagated for 10 ns in the NVT ensemble at equilibrium volume to calculate dynamical properties as well as drawing 5 snapshots for each temperature that are apart at least 1 ns to use for an iso-configurational analysis (ISOCA) [235, 236].

5.2.2 Work Flow to Characterize the Liquid Dynamics

In order to characterize the liquid dynamics, appropriate length and time scales have to be chosen, which is achieved by the procedure involving the following steps:

1. From the NVT simulation of the system with \(N = 10,000\) molecules at the target temperature we obtain the oxygen-oxygen radial distribution function

\[
g_{oo}(r) = \frac{1}{2\pi r^2 \rho} \sum_{i=1}^{N} \sum_{j>i}^{N} \delta \left( r - \| \mathbf{r}_i - \mathbf{r}_j \| \right)
\]

where the sums consider \(N\) oxygen atoms and their positions \(\mathbf{r}_{i/j}\), \(\rho\) is the liquid density and the average is over all trajectory frames.

2. We calculate the isotropic structure factor

\[
S(q) = 1 + \frac{4\pi\rho}{q} \int_0^{\infty} dr \, r \sin(qr) \left[ g_{oo}(r) - 1 \right]
\]

and define \(q_0\) as the value where \(S(q)\) has its first peak, with \(q\) being a reciprocal length.

3. For this \(q_0\) we calculate the quantity

\[
\Phi(q, t) = \frac{1}{N} \sum_{j=1}^{N} \exp \left( iq \cdot [\mathbf{r}_j(0) - \mathbf{r}_j(t)] \right)
\]
4. Via $\Phi(q, t)$ we obtain the self-intermediate scattering function $F(q, t) = \langle \Phi(q, t) \rangle$ and the dynamical susceptibility $\chi_4(q, t) = N \left[ \langle |\Phi(q, t)|^2 \rangle - \langle \Phi(q, t) \rangle^2 \right]$. Isotropic averages taken over 200 independent directions according to $F(q_0, t) = \langle F(q, t) \rangle_{\|q\|=q_0}$ are evaluated.

5. The time of maximum heterogeneity $t_0$ is taken as the time where $\chi_4(q_0, t)$ has its maximum, i.e. where the movements at the nearest-neighbor range are most heterogeneous.

The resulting values for $q_0$ and $t_0$ for all temperatures can be found in Table 5.1 together with the structural relaxation time $\tau_{\text{liq}}$ that was obtained as $\alpha$-relaxation value from $F(q, t)$.

<table>
<thead>
<tr>
<th>Temperature ($K$)</th>
<th>$q_0$ ($\text{Å}^{-1}$)</th>
<th>$t_0$ ($\text{ps}$)</th>
<th>$\tau_{\text{liq}}$ ($\text{ps}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>273 K</td>
<td>2.01</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>260 K</td>
<td>1.96</td>
<td>11</td>
<td>6</td>
</tr>
<tr>
<td>250 K</td>
<td>1.91</td>
<td>27</td>
<td>14</td>
</tr>
<tr>
<td>240 K</td>
<td>1.84</td>
<td>115</td>
<td>68</td>
</tr>
<tr>
<td>230 K</td>
<td>1.80</td>
<td>620</td>
<td>356</td>
</tr>
<tr>
<td>220 K</td>
<td>1.77</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>210 K</td>
<td>1.76</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 5.1: Overview of length ($q_0$) and time ($t_0$) scales used to characterize dynamical heterogeneity and structural relaxation time $\tau_{\text{liq}}$ at different temperatures. Because of the computational cost we did not consider 220 K and 210 K for the rest of the study.

### 5.2.3 Transition Path Sampling Simulations

We will study the nucleation event with transition path sampling [80] (TPS), which was performed by Stephen J. Cox using an in-house code that interfaces with LAMMPS [139]. The system comprised of 4000 mW molecules and a pressure of 1 atm was maintained using a barostat with a damping constant of 5 ps and a timestep of 10 fs. Langevin dynamics was used to maintain a temperature of 235 K with a damping constant of 10 ps.

To define whether or not a trajectory was reactive or not, we used the size of the largest ice-like cluster $N_{\text{cls}}$ as defined by Li et al. [144]. The system was considered to be liquid if $N_{\text{cls}} < 50$ and ice if $N_{\text{cls}} > 800$. An initial reactive trajectory was generated by unbiased simulation at 205 K, with configurations and velocities stored every 2 ps. A 2.4 ns segment of this trajectory was extracted, and the velocities were rescaled by a factor $\sqrt{\frac{235}{205}}$. An equilibration of 1000 TPS moves was then performed.
with a 2:1 ratio of shooting to shifting moves. As we used stochastic dynamics, we made use of the one-way shooting algorithm [80]. The maximum length of a shifting move was 80 ps. After this equilibration period, the length trajectory was further reduced to 1.6 ns, and a further 7500 TPS moves were performed as a production run, with every 100 trajectories saved. A discussion about the statistical independence of the TPS trajectories is included in Appendix C.5.

To classify a molecule $i$ as either mobile or immobile we employed the enduring displacement formalism [237]:

$$m_i(t) = h(|\bar{r}_i(t + \Delta t) - \bar{r}_i(t)| - a), \quad (5.1)$$

where $a = 1$ Å, $\Delta t = 2$ ps, and $\bar{r}_i(t)$ is the inherent structure position of molecule $i$ at time $t$ and the Heaviside step function $h(x)$. To find the inherent structure positions we perform energy minimization on trajectory snapshots from molecular dynamics, which lets molecules move into the nearest basin of potential energy. Minimization was done with the FIRE algorithm [238]. To calculate overlaps between immobile and crystalline regions of space we first define the coarse grained immobility ($\mathcal{I}$) and crystallinity ($\mathcal{Q}$) fields:

$$\mathcal{I}(r) = \sum_{i=1}^{N} (-1)^{m_i}g(|r - r_i|), \quad (5.2)$$

$$\mathcal{Q}(r) = \sum_{i=1}^{N} (-1)^{q_i}g(|r - r_i|) \quad (5.3)$$

where the sum runs over all molecules, $g(r) = (2\pi\xi^2)^{-\frac{3}{2}}\exp\left(-r^2/2\xi^2\right)$ is a normalized Gaussian with $\xi = 2.8$ Å. $m_i = 1$ and $q_i = 1$ if the molecule identified with the position vector $r_i$ is mobile / ice-like respectively.

### 5.2.4 Results

Dynamical Heterogeneity in Supercooled Water

We begin by uncovering DH in the liquid together with a brief structural analysis of the resulting dynamically distinct domains. To this end we perform molecular
Figure 5.1: Dynamical heterogeneity in supercooled liquid water. a) Spatial distribution of the dynamical propensity (DP) at 230 and 273 K. Relatively immobile and mobile molecules (only oxygens shown) are colored according to the scale on the left. b) Probability density distribution of the DP at 230 K. Blue and red shaded regions highlight the 5% of water molecules labeled as most immobile (MI) and most mobile (MM). c) Number of $n$-membered primitive rings within the respective domain at 230 K. The dashed portions of the bars represent the fraction of those rings that is fully connected by hydrogen bonds. The top insets show an example for a fully and non-fully hydrogen-bonded 5-membered ring, where solid lines between oxygens are a guide to the eye and do not imply hydrogen bonds.
5.2. Methods

dynamics (MD) simulations with the TIP4P/Ice [57] model in a homogeneous water system in the temperature range 273 K to 230 K. In addition we utilize ISOCA, a technique which allows us to obtain spatially resolved maps of DH. We quantify the tendency of each molecule to move with a dynamical propensity (DP):

$$DP_i = \left\langle \frac{\|\mathbf{r}_i(t_0) - \mathbf{r}_i(0)\|^2}{\text{MSD}} \right\rangle_{\text{ISO}}$$  \hspace{1cm} (5.4)

where $\mathbf{r}_i(t)$ is the position vector of molecule $i$ at time $t$, $t_0$ is the time of maximum heterogeneity (see definition in Methods) and MSD is the mean square displacement of all oxygen atoms. In this approach we average the outcome over multiple trajectories that start from the same initial configuration, indicated by the notation $\langle \ldots \rangle_{\text{ISO}}$. For all snapshots (5 for each of the 5 different temperatures) we calculate the dynamical propensity (DP, see Eq. 5.4) by performing 200 independent MD runs (we assess the role of the number of trajectories in Appendix C.1) in the NVT ensemble. As length of these production runs we choose the time $t_0$. For our purpose it is sufficient to consider oxygens only when calculating the DP. In Fig. 5.1a we show two snapshots of initial configurations used for the ISOCA at 230 and 273 K, where each oxygen atom is colored according to its DP. This choice of temperatures is to illustrate the maximum difference in the extent of DH as well as the relevance of strong supercoolings like 230 K in homogeneous nucleation of ice [161, 231, 239]. It can be seen that spatially localized domains of relatively immobile (blue) and mobile (red) particles emerge and that the spatial aggregation of the domains differs drastically. As reported in Fig. 5.1b, the probability density distribution of the DP at 230 K is rather broad with a factor of 30 between the mobility of particles at opposite tails. To obtain structural insight into the different domains we label the top and bottom 5% of the DP distribution (see Fig. 5.1b) as most mobile (MM) and most immobile (MI). A detailed analysis of the structural differences in these domains has been performed, where in agreement with the literature [223, 225] we find the MI domains to be in a more tetrahedral environment (see Appendix C.3). The structural properties of these regions are then established by calculating the number of primitive $n$-membered rings within them (utilizing
the R.I.N.G.S. code [240]), as well as computing distributions of common order parameters such as $q_{\text{tetra}}$ or topological patterns such as cages. We did not investigate the two lowest temperatures as the computational expense for establishing the DP with statistical significance becomes too costly. All results reported in this text are averages over the results for the 5 snapshots per temperature. As can be seen from Fig. 5.1c the MI regions have a rings distribution strongly peaked around $6 \pm 1$ members while the distribution for the MM domains is very broad. Moreover, the amount of entirely hydrogen-bonded rings is substantially higher in the MI region. In particular, an abundance of $6 \pm 1$-membered hydrogen bonded rings can be regarded as the key structural characteristic of the MI domains in the liquid. We show that the choice of threshold for labeling MI or MM regions has a negligible influence in Appendix C.2. While the majority of the 6-membered rings are ice-like if regarded in isolation, when also considering their surroundings we find that only a negligible fraction of the molecules involved can be labeled as ice-like (details in Appendix C.3). This indicates that it is the relative orientation of the rings that is not yet ice-like thus suggesting that the MI regions could indeed be the birthplace of ice clusters.

Dynamics of Pre-Critical Fluctuations

As a first step to understand the connection between DH and ice nucleation we focus on pre-critical clusters, i.e. the ice-like clusters that form via frequent thermal fluctuations and thus are readily probed by unbiased MD [241]. Ice-like molecules were detected using an order parameter ($l_{q_6}$) according to Li et al. [144] as implemented in PLUMED2 [210, 211]. For values of $l_{q_6} > 0.5$ we classify a molecule as ice-like, otherwise as liquid. The key results of our analysis are reported in Fig. 5.2, where we find a strong tendency for the pre-critical ice nuclei to form within MI domains, rather than the MM domains. To quantify their preference to form in the immobile regions we have split the whole range of (sorted) DP values into 20 equal sections (i.e. each corresponding to 5% of the whole range). This means that the molecules in the first/last DP section are the same molecules as the molecules in the MI/MM regions. With this choice the expected overlap in the absence of any corre-
5.2. Methods

Figure 5.2: Connection between dynamical heterogeneity (DH) and pre-critical cluster formation. a) Representative snapshot of a spontaneously formed cluster (red bonds and spheres) immersed in the MI region (transparent blue surface representation). b) Average overlap between the molecules in the largest ice-like cluster and molecules in the relative DP range. Each bar corresponds to a 5% fraction of (sorted) DP values, i.e. the first / last bar corresponds to the MI / MM regions. The expected overlap if clusters were uncorrelated with the DP would be 5% (indicated by the dashed line). c) Average evolution of the mobility (DP) and crystallinity ($l_6$) for molecules in a cluster before its first time of assembly (taken to be $t = 0$). Dashed lines indicate the mean values of DP and $l_6$ of the liquid. $\tau_{\text{liq}}$ is the structural relaxation time. Shaded regions indicate 95% confidence intervals. All data was obtained with the TIP4P/Ice water model at 240 K.
ration with the DP would be 5%. The results reported in Fig. 5.2b,c were calculated from a TIP4P/Ice trajectory at 240 K. We define as the time of first assembly of a cluster the frame that for the first time has the biggest ice-like cluster comprised of molecules that have not been part of the biggest ice-like cluster in the 75 ps before that frame. This is slightly larger than the structural relaxation time of 68 ps at that temperature and larger or smaller choices did not qualitatively alter the findings. In Fig. 5.2b we plot the average overlap of molecules in the biggest ice-like cluster with these DP sections. We clearly see that there is a strong preference for pre-critical ice clusters to belong to the lower DP sections (i.e. more immobile molecules). In addition, we find that the formation of pre-critical clusters is suppressed in the MM domains as the overlap values there are below the base-line of 5%. Considering the fact that nucleation is stochastic in nature, this is strong evidence for the connection between immobile and ice-forming regions.

Having shown that pre-critical ice clusters strongly overlap with the immobile regions from their early stages we now study the temporal connection between immobility and clusters before their first time of assembly. In Fig. 5.2c we show the average value of the DP and the crystallinity parameter $l_q$ [144] of molecules that belong to a cluster at its first time of assembly (taken to be at $t = 0$). The 95% confidence intervals for the DP evolution result from point-wise estimates from $10^6$-fold bootstrap resampling of all relevant clusters, the corresponding intervals for the $l_q$ curve are smaller than the line width plotted. It can be seen that the mobility drops at $\sim -1000$ ps, which is significant compared to the structural relaxation time of the liquid: $\tau_{\text{liq}} \approx 68$ ps at 240 K (see Methods). In addition this drop occurs earlier and is much less abrupt than the change in structure, which can be associated with the rise of the $l_q$ order parameter at about 400 ps before the assembly. This finding is crucial and it confirms that immobility on average precedes ice-cluster formation by a significant timespan. This can be thought of as a dynamical incubation period in which the dynamics of the molecules changes prior to the structural change towards ice. While this mechanism is not necessarily orthogonal to the commonly applied reasoning of purely structural ordering, our results suggest that arguing in terms of
5.2. Methods

a process that involves distinct dynamical and subsequent structural steps is a viable route for a better description of nucleation.

Dynamics During Nucleation

The results reported above point strongly towards an interplay between structural motifs pertinent to nucleation and the dynamics of the system. Unbiased MD simulations however, cannot directly sample nucleation events, except under extreme conditions close to the homogeneous nucleation temperature [203]. In this section, we report results from transition path sampling (TPS) simulations that allow us to sample many nucleation events at reasonably high temperature. Specifically, we study a system of water molecules represented by the mW model [29] at 235 K. In the case of the brute force simulations, we employed an ISOCA to quantify mobility. However, in the case of TPS we harvest many more (7500) reactive trajectories, making ISOCA for each frame of each trajectory impractical. To fully exploit the statistical sampling provided by TPS, we therefore employ the *enduring displacement* formalism, which permits on-the-fly calculation of each particle’s mobility. Using this approach has the added benefit of allowing us to validate the robustness of our previous results.

In order to identify regions of space as either ice-like/liquid-like and immobile/mobile (and the boundaries separating them), we introduce the coarse-grained crystallinity field \( Q(r) \), and immobility field \( I(r) \). In brief, \( Q(r) \) is obtained by smearing each ice-like molecule’s position with a positive, normalized Gaussian, and each liquid-like molecule with a negative Gaussian. In regions that are predominantly ice-like, \( Q(r) \) will take values close to the density of ice, whereas in regions that are predominantly liquid-like, it will take values close to the negative density of liquid water. Boundaries separating ice-like and liquid-like regions are defined by surfaces with \( Q(r) = 0 \). In Appendix C.5 we show that \( Q(r) \) performs well at identifying regions as ice-like that fit with our intuitive understanding, while simultaneously neglecting small fluctuations. In a similar fashion, \( I(r) \) is obtained by smearing immobile particles with a positive Gaussian, and mobile particles with a negative Gaussian. In Fig. 5.3a we show a snapshot of \( Q(r) \) and \( I(r) \) prior to
Figure 5.3: Ice nucleation occurs in relatively immobile domains of supercooled water. Time evolution of the coarse grained immobility $I(r)$ (translucent silver) and crystallinity $Q(r)$ (opaque blue) fields, from a trajectory harvested by TPS. (a) Prior to nucleation we see large immobile domains and an absence of crystalline order. During nucleation, the ice nucleus forms (b) and grows (c) within the immobile domain. The ice cluster in snapshots b) and c) comprised of 83 and 296 molecules respectively. The diameter of the ice-like region in c) is approximately 3.4 nm. The TPS simulations in this picture were created by Stephen J. Cox.

nucleation (only regions $Q(r) > 0$ and $I(r) > 0$ are shown). While there are no regions identified as ice-like, we do see large immobile domains in the supercooled liquid. Figs. 5.3b,c show similar snapshots after the onset of nucleation. Crucially, and consistent with our findings for the pre-critical nuclei, it is clear that the ice nucleus forms within an immobile domain. Overall, the results obtained from TPS support and extend our observations from the unbiased simulations and strengthen our conclusion since we now also have insight: (i) for larger clusters; (ii) for a large ensemble of nucleation trajectories; and (iii) with a different method of classifying immobility. In addition, as can be seen in Fig. 5.3b,c the surrounding immobile region is larger than the ice-like region, which suggests that ice-clusters further slow down their surroundings. We will explore this aspect next with the atomistic model.

Dynamics Near Ice Clusters

So far we have shown that ice nucleation occurs in relatively immobile regions. To complete the emerging picture we further studied the effect of a sizeable ice-cluster on the dynamics of the surrounding water with the TIP4P/Ice model. We considered an ice cluster from an earlier study [242] that consists of $\sim$ 300 water molecules, which is roughly the critical cluster size at 230 K [161]. The ice cluster is solvated with 100,000 water molecules in a large cubic box (side length $\sim$16 nm) to
Figure 5.4: Dynamical heterogeneity (DH) in presence of an ice-cluster at 230 K. a) DH of water molecules (spheres colored according to the color scale) surrounding a critical ice-cluster (orange bonds and spheres). The presence of the cluster causes a “cloud” of immobility around it. b) Average DP of a water molecule as a function of its minimum distance to the cluster $r_{\text{min}}$. The slowdown of dynamics around the cluster extends for $\sim 10$ Å, which roughly corresponds to 2.5 hydration layers.

avoid finite-size effects in the DP and cluster self-interaction. After equilibrating the simulation box at 300 K, whilst keeping the cluster fixed, we perform a quench to 230 K. Then we run several molecular dynamics steps (for roughly two times the relaxation time of TIP4P/Ice at 230 K each), alternating between keeping the cluster and the water molecules fixed. In this manner we equilibrate the interface between the two, avoiding that the cluster grows or shrinks significantly. After equilibrating the interface, we performed an ISOCA on the resulting system, incorporating only liquid molecules into the calculation of the DP. In Fig. 5.4a we show a slice through the ice-cluster (orange) which we find to be surrounded by a “cloud” of immobile molecules (that are not part of the ice but rather the liquid). The resulting MI regions are of comparable size to the cluster itself. This indicates that ice-clusters not only prefer to appear in immobile regions, but once a large cluster has formed it fur-
ther slows down the surrounding water. As shown in the previous section the TPS results also show a surrounding shell of immobile particles around sizeable ice-like regions. This is relevant to crystal growth as the liquid molecules in direct vicinity are significantly less mobile than those in the bulk. We have quantified the range of the cluster influence in Fig. 5.4b where we see that it takes approximately 10 Å (∼ 2.5 hydration layers) for the dynamics to return to the average. This is roughly the same distance that was found to mark the change from confined/interfacial to bulk behavior of water in nanopores [243] and at solid surfaces [244], indicating an interesting correspondence between the change in dynamics via confinement and vicinity to a crystal.

5.3 Discussion and Conclusions

The results of our study suggest that the complex nature of dynamics should not be overlooked in theoretical descriptions of nucleation. We have shown that liquid molecules in the vicinity of a nucleus are slowed down significantly, which implies that their diffusivity (connected to the attachment rate) is reduced compared to bulk molecules, an aspect neglected by classical nucleation theory. A new theoretical approach is also suggested by our findings and their connection to the literature [161]: The population of rings seems to smoothly describe the transition from liquid to crystal and could thus be a promising ansatz in terms of thermodynamic models, similar to the proposed two-state model of Russo and Tanaka [226].

We verified that our results on the rings distribution also hold for the coarse-grained mW model [29] even though for mW the extent of DH is much smaller (see Appendix C.4). This suggests that the characteristic features identified in our study are not sensitive to the specific hydrogen bond parametrization, but rather caused by the tetrahedral order inherent in the modeled material [227]. Thus, our findings may be of relevance not just to water but to a much broader range of materials, evidently ones with tetrahedral order (such as group IV elements and silica). More broadly, it remains to be seen if the immobility of pre-crystalline structures is connected to nucleation in non-tetrahedral liquids in the same manner since the population of
e.g. rings will be material-specific. However, based on our findings we can suggest that it is the correspondence between immobile and crystalline topological features (such as rings) that connects immobile regions with nucleation.

The connection of DH and nucleation in water could be probed experimentally by investigating heavier water molecules as their mobility might be different. It is for instance established that liquid D$_2$O has a higher nucleation rate than H$_2$O [245]. However this cannot be taken as direct evidence in support of our observation as the change in the hydrogen bonding induced by nuclear quantum effects [246] potentially influences the nucleation rate too. A more rigorous experimental validation of our findings would be the nucleation rate comparison for H$_2^{18}$O and ordinary water, which to the best of our knowledge has not been achieved. If one of the two liquids is more/less mobile (diffusive) our results suggest a decreased/enhanced nucleation rate. Materials other than water could be studied in greater detail if their crystallization kinetics are slower, e.g. the correlation between structural ordering and dynamics could be probed in colloidal systems with real-space imaging [247] by studying the mobility of particles in crystalline clusters before their time of first assembly.

Our findings are likely of relevance to heterogeneous nucleation and nucleation from solution. Generally, previous work has focused on investigating the structural or templating role of nucleating agents [25, 109, 208, 248]. However, an impurity or substrate is bound to impact the dynamics of the supercooled liquid in its vicinity, possibly leading to a novel mechanism of heterogeneous nucleation. For the example of water freezing, ice nucleation on hydrophobic surfaces (basically incapable of structuring the water network to a major extent) has been reported [109, 208] as well as intriguing alternating hydrophilic-hydrophobic patterns in the prominent ice nucleating bacteria *Pseudomonas syringae* [249] and a nucleation enhancement by soluble molecules [250] that could be connected to the liquid dynamics. Moreover, for nucleation from solution it is well known that different solutes change the nucleation mechanisms, i.e. in the case of urea [251]. Understanding how solutes change the dynamics and impact the formation of amorphous precursors could shed
light on this issue. As such, we hope that this work will push the community to take into account the role of dynamics, and particularly of DH in connection with crystal nucleation and growth, with emphasis on the cases of heterogeneous nucleation and nucleation from solution.

In conclusion, we have established the structural hallmarks of DH in water and shown that ice nuclei originate within immobile regions of the supercooled liquid and that there is a dynamical incubation period in which the mobility of particles drops prior to any structural change. Additionally, the presence of an ice crystal-lite causes arrested dynamics in water molecules that surround it. This connection between dynamics and structure provides a novel perspective on the physics of nucleation.
Chapter 6

Data-Driven Approach to Identify Descriptors for Heterogeneous Ice Nucleation

There are many aspects of an impurity that influence whether it will be good or bad at heterogeneous ice nucleation. On the microscopic level we are lacking a predictive set of rules, guidelines or descriptors to predict \textit{a priori} whether a substrate will have a good or bad ice nucleating ability (INA). The most prominent rule around is still the lattice match of the substrate to ice, which has however been shown to not be necessary for a good INA. In this chapter, we compute the INA of a large set of structurally diverse model substrates as well as many potential physical descriptors that can be obtained from simple simulations. We then devise a data-driven approach based on machine-learning methods to identify which of these features are most important to predicting the INA. A physical interpretation of our results is that no single feature is sufficient for a prediction of INA and the four most important contributions for the INA are: i) 2D-lattice match to ice; ii) local density reduction of the liquid; iii) the uniformity of the adsorption environment of ice structures; and iv) local tetrahedral ordering induced in the liquid. With this we take a first step at identifying a new and predictive set of microscopic descriptors for the INA of substrates and show that the quantitative predictability of heterogeneous ice nucleation is in reach.
6.1 Introduction

We have seen in chapter 3 that the screening of well-defined model systems has the advantage of being able to disentangle various contributions to the INA as well as exposing trends that would go unnoticed when considering only a small sample of systems. Hence, we believe that this approach can be complementary to traditionally performed studies where a few systems are prepared to match a particular material. Thus, in the data-driven approach of this chapter we are not aiming to represent a certain material with high accuracy, but rather study a large variety of different systems that are well-defined with controllable characteristics such as lattice spacing, interaction energy or symmetry. We then aim at finding descriptors, i.e. computable and ideally experimentally accessible quantities, that enable the prediction of the substrate’s INA.

Such a microscopic understanding would potentially benefit the macroscopic comprehension of IN. A simpler case where something similar was achieved is heterogeneous catalysis, where it was for instance found that for the methanation reaction \( \text{CO} + 3\text{H}_2 \rightarrow \text{CH}_4 + \text{H}_2\text{O} \) the activation and formation barriers scale linearly with the CO dissociation energy on the surface [252, 253]. This enables the computational screening (at the DFT level) of potential catalysts. Recently, the identification of suitable predictive parameters (also termed descriptors) via machine-learning methods poses a promising ansatz in the field of materials science, as was demonstrated e.g. by Ghiringhelli et al. [254] for the band gap of zinc blende and wurtzite structures. For heterogeneous ice nucleation the “gold standard” for the unknown INA of a given impurity still is the set of requirements outlined by Pruppacher and Klett in the seventies [97]. While some of these deal with more macroscopic properties from an atmospheric chemistry viewpoint (insolubility and aerosol size requirement) the others (chemical bond, active site and crystallographic requirement) deal with in principle microscopic characteristics. But the specific definition of e.g. what will be an active site for IN or quantitative statements are missing. Moreover, there are many known exceptions to these rules [20, 255, 256], some of which we have also added in chapter 3.
In this chapter, we will combine the model systems investigated previously to characterize their ice nucleation ability (INA). We will then compute a large number of possible physical observables that describe the substrate or the substrate-water interface at coexistence temperature and from this we aim to predict the INA. To this end we will utilize state-of-the-art machine learning models that enable the assessment of predictive power of a given set of descriptors to find and interpret the most important ones. This is a first step towards a comprehensive and predictive set of descriptors and rules to quantitatively estimate the INA in computational screenings and contributes towards the understanding of what makes a certain material good or bad at IN.

6.2 Methods

6.2.1 Simulations and Systems

We performed simulations of water, represented by the coarse-grained mW model [29], in contact with a variety of model substrates. The latter comprise of the 400 Lennard-Jones substrates of chapter 3, the OH group patterns of Ref. 257, graphitic surfaces with modulated water-substrate interaction strength similar to the ones used in Ref. 112, graphene oxide modelled as in Ref.108 and several additional OH group patterns with symmetries not yet found in Ref. 257. In total we have 900 substrates.

On these systems we perform 5 cooling ramps each (at 1K/ns) from 273 K to 200 K. We define as the nucleation temperature $T_n$ the temperature at which a significant drop in the potential energy is seen, which is (similar to what was shown in chapter 3) sufficient to identify the onset of nucleation as the drop is much more pronounced than the slope caused by the cooling. With this approach we have a measure so that we can rank the substrates by their INA as well as a measure for the uncertainty of the latter. In Fig. 6.1 we show $T_n$ for all substrates grouped by the class of systems they belong to. Interestingly we find a strong imbalance between
Figure 6.1: a) Nucleation temperature $T_n$ for all substrates considered. The acronyms below are identifiers for different classes of systems and are explained in appendix D. b) Example snapshots for some models systems, where gray / red / light blue spheres represent Lennard-Jones / oxygen / carbon atoms.

good and bad ¹ nucleators. For instance, if we split the whole range for $T_n$ into three equal sections we find that we have 751 bad nucleators ($T_n < 222$ K), 55 excellent ice nucleators ($T_n > 245$ K) and 94 mediocre ice nucleators ($T_n$ between the two other thresholds). This is somewhat surprising since many of the systems probed were comprised of hydroxyl groups. Indeed, often the presence of hydroxyl patterns in residues of biological molecules alone is seen as sufficient for explaining their ice

¹The choices to label nucleators as good or bad are in principle arbitrary. In this work we have subsequently checked in several places the influence of temperature thresholds for this classification, often yielding very similar results.
6.2. Methods

nucleation abilities [126, 127].

The features we compute for each of the systems and the method used to obtain them are listed in detail in appendix D. We have generally taken into account forces, displacements, densities, generalized lattice matches, velocities, adsorption energies of many different ice structures and local structuring measures, and also distinguish different liquid water layers regarding their vicinity to the surface. In total we have about 1500 initial features that will be assessed.

6.2.2 Machine-Learning Approaches and Importance Rating

We make use of tree-based machine-learning methods that have been developed to produce predictive models. A single decision tree is a binary tree that implements decisions about features $f_i$. If $f_i$ is larger than a certain value $\delta_i$ then the observation would go into one side of the node, if not it will go into the other branch of the node. This is illustrated in Fig. 6.2. This process can be repeated for $k$ features and creates on the order of $2^k$ tree nodes. If a node does not split up further it is called a terminal node or leaf. Leafs enable the desired prediction, as each observation that falls into a certain leaf will be labelled according to the associated class (for classification) or the associated range of the continuous quantity to be predicted (for regression).

Decision trees have the advantage that they are in principle interpretable, i.e. one can follow the step-wise reasoning between inputting an observation at the first node (called root) up to the step where it is classified. To train such a decision tree when creating a new node it is checked which property $f_i$ yields the best improvement over the current model. Improvement in the prediction error is usually measured with either the gini-impurity for a set of $J$ classes ($E = \sum_i P_i(1 - P_i)$) or the information gain ($E = -\sum_i P_i \log_2 P_i$) computed at the node for classification and the mean square error for regression of a continuous quantity.

Another advantage of these models is that an inherent measure $I(f_j)$ for the importance of feature $f_j$ can be constructed. Once the tree is fit we permute the values of the $j$'th feature for all observations from $f_j$ into $f_{\pi_j}$ and assess the tree
accuracy before and after that:

\[ I(X_j) \propto E(f_j) - E(f_{\pi_j}) \]  \hspace{1cm} (6.1)

where \( E \) is the desired measure of the model error (as mentioned above). It is commonplace to use this approach (in the form of a random forest, see below) with genetic arrays to identify gene expressions that correlate with certain diseases \cite{258, 259}. We will train a random forest on the complete set of initial features and assess their importance score. It is critical to realize that correlations amongst the features generally can lead to biases of the importance rating, i.e. smaller groups of correlated features are favored over larger groups, even if they are intrinsically equally important. This means we cannot just take the top \( n \) features with the high-
6.2. Methods

est importance ranking and declare them as the best set of descriptors. There is an ongoing area of research and several methods like unbiased trees [260, 261] or partial permutations [262, 263] have been proposed to deal with correlated features. We will take care of correlations later on and devise schemes to select a subset of important features out of the large set of initial features taking into account importance score as well as correlation with other features.

We are using more powerful models that are based on single decision trees to obtain our predictive accuracy and the importance measure. This has the advantage that non-linear correlations can be captured (better than e.g. in linear regression) since we expect the INA of the substrates to be a highly non trivial function of the features. We will train a random forests (RF) [264], which are a collection of many decision trees, that are however not trained on the same data each, but rather bootstrap samples of the original training set and only use a randomly selected subset of all \( k \) features. In this manner the variance of the models is significantly reduced and the predictive power greatly increased [265, 266]. The feature importance of a random forest is simply the mean (and in some sense normalized) feature importance over all trees that included this feature. A further model we utilize to test predictive power are boosted trees (BT) [265, 267]. In this approach a single first decision tree is fit. Then the residual error of the prediction based on that tree is taken as target value for another decision tree that is then added to the previous model, and this is repeated in an iterative fashion. In this manner we obtain a prediction for the \( k' \)th iteration:

\[
\hat{y}_k = \hat{y}_{k-1} + \eta \hat{f}_k = \sum_{i}^{k} \eta \hat{f}_i
\]

(6.2)

where \( \eta \) is the so-called learning rate and each new model \( \hat{f}_k \) is fit on the residuals \( e_{k-1} = y - \hat{y}_{k-1} \) of the accumulated previous models where \( y \) denotes the true value and \( \hat{y} \) the prediction. In general, several hyperparameters (e.g. learning rate, number of trees, maximum tree depth, fraction of features to be used etc.) have to be set in all of the models considered in here. To obtain a reasonable estimate of the latter
we perform a grid search in the space of hyperparameters to obtain something that is close to the optimal value. We note that we also used a model that is not based on trees (support-vector-machine [268, 269]), but we omit the results here since they were always equal or inferior in terms of prediction accuracy compared to the BT.

### 6.3 Results and Discussion

We now perform the importance rating of the initial features and construct a framework that picks a minimally correlated subset of the most important ones. We then test these subsets of features to obtain which are the important ones and how many features are necessary to predict INA with reasonable accuracy and then discuss the most important groups.

#### 6.3.1 Selected Features and Prediction Performance

Our estimate of the importance rating of the various features is based on a RF trained to classify nucleators as good ($T_n > 225$ K) or bad ($T_n \leq 225$ K) on all $\sim 1500$ initial features. The results obtained for a threshold choice of 240 K are almost identical. After doing so each initial feature has an importance rating. It would however be wrong to simply pick the best $n$ features regarding this metric since the RF does not take care of strong correlations, i.e. the “best” 10 features might be highly correlated. Any model or set of descriptors would benefit from a minimum correlation between each of the selected features and the total number needed to produce reasonable predictions will be reduced drastically.

Cross-Correlation and Feature Selection

First we compute the correlations between all features according to the maximum information coefficient (MIC) [270]. This has the strong advantage over taking just the linear correlation in that we can better identify non-linear correlations between
features. The MIC is defined as the mutual information based on entropy $S$:

$$\text{MIC}(X, Y) = \sup_{G_x, G_y} \left\{ S_{G_x}(X) - S_{G_x, G_y}(X|Y) \right\}$$

$$= \sup_{G_x, G_y} \left\{ S_{G_y}(Y) - S_{G_y, G_x}(Y|X) \right\}$$

with $$S_{G_x}(X) = -\sum_{x \in G_x} P(x) \log P(x)$$

and $$S_{G_y, G_x}(Y|X) = -\sum_{x \in G_x} \sum_{y \in G_y} P(x, y) \log \left[ \frac{P(x, y)}{P(x)} \right]$$ (6.3)

where the sums are done over grids $G_x$ and $G_y$ that are optimized to yield the maximum value of the MIC [270]. The MIC is also normalized so that it falls in the range $[0, 1]$, where 1 means maximum possible correlation.

To select reasonable sets of $n$ features to be used for training the models we then use feature-selection based on three different methods:

1. The MIC can be used as a distance metric $d(X, Y) = 1 - \text{MIC}(X, Y)$. This means we can simply apply hierarchical clustering and group the features into sets of correlated ones. We then select the most distant $n$ clusters and take the most important feature in each cluster. We applied this to the 200 most important features with average linkage.

2. The clustering described above gives higher priority to decorrelated features rather than their importance. Since it is not clear a priori if this will yield the best results we also devise an approach that gives priority to the importance rating. We take the list of best features and select the most important one. We then search the list of features in order of descending importance and select a feature if the mean MIC with the already selected features is smaller than a threshold. This is done till $n$ features have been selected.

3. The same approach as 2 but taking as criterion that the maximum rather than the mean of MICs between already selected and to-be selected features is smaller than a threshold. We tested different thresholds and find a value of
0.4 to be a reasonable compromise between avoiding true correlations while not excluding features that are correlated only by random noise.

Regression Performance

Now we have everything at hand to select $n$ features and train the RF and BT models to predict $T_n$. By assessing their quality we can gain insight i) how many features are generally needed for a good prediction and ii) do the selected features contain enough information to predict $T_n$? The latter question is most important to establish the physically relevant features, even if we cannot directly interpret the resulting models. Also it is naturally clear that the fewer features a model needs to achieve a certain accuracy, the more “efficient” it can be considered compared to similarly accurate models that however need to include more features.

To assess the quality of the models we performed 10-fold cross validation according to commonplace practices. This means we randomly split all 900 observations into 10 groups roughly containing the same number of elements. Then 10 models are fit, each omitting one of these groups for training but rather using that group as test data to establish a mean square error. The 10-fold cross-validated mean-square error (10-CV MSE) is the average of the MSE of the 10 models.

In Fig. 6.3 we now show the performance in terms of the 10-fold cross validated root mean square error (10-CV RMSE) for the BT models and for the three
different feature-selection methods devised previously. The following main observations can be made:

- The performance between the different feature selection methods is similar. The most gains in terms of accuracy are made within the first 5 features and after 10 to 12 there is no significant improvement. For classification the use of more than 5 features is disadvantageous.

- Selection method 2 yields the most efficient improvements in terms of how many features need to be included, the most efficient model selecting $n = 4$ features.

- The best 10-CV RMSE we get is about 5.5K. The natural standard deviation of the $T_n$ we calculated was around 2.5 K. This means that we can predict $T_n$ in a 11 K window, while the natural uncertainty in $T_n$ is about 5 K.

The last point can be considered a decent prediction of the INA of a given system. The total window of $T_n$’s is about 70 K, which means with a prediction window of 11 K we can distinguish between 6 and 7 different classes of nucleators. To the best of our knowledge we are not aware of any previous work that was able predict the INA across a diversity of many different substrates.

**Classification Performance**

In addition to estimating the descriptor performance by regressing $T_n$ we also trained classification models based on simple decision trees, i.e. models that can distinguish between good ($T_n > 225$ K) and bad ($T_n \leq 225$ K) ice nucleators. We also repeated this with 240 K as the threshold, yielding models that were only marginally worse.

To assess the performance of a classification model we calculate the area under the curve (AUC) of the receiver-operator curve (ROC). ROC’s are commonplace in machine learning and plot the true-positive rate versus the false-positive rate of a model regarding its prediction of a certain class, in our case being a good ice nucleator. The AUC of this always lies in the interval $[0.5,1]$ and depending on the value we can label the model as $([0.5,0.6]$ - useless), $([0.6,0.7]$ - bad), $([0.7,0.8]$ - fair), $([0.8,0.9]$ - good), $([0.9,1]$ - excellent).
Fig. 6.3 shows the results for the classification AUC, where the latter has been averaged over 3000 models, each randomly splitting the whole dataset in 2/3 for training and 1/3 for calculating the ROC and AUC. Generally, we find a very good performance of the resulting models, most of them being at the border of good and excellent. Considering we are using fairly simplistic decision trees this means that the selected features carry enough information to distinguish what is a good and bad ice nucleator.

6.3.2 Feature Interpretation

Now that we have established that with the initial features and devised protocols for feature selection we can build predictive models for the INA. In this section we will now attempt to interpret the commonalities between all these models and shape them into physical interpretations of the descriptors.

In Fig. 6.4 we show the first 12 features that get selected by the three different selection methods. They are displayed in order, i.e. if we were to ask for the $n = 3$ feature model of selection method 1 we would take the top three features in the left column. We are not discussing in detail all of the features (the precise meanings and method of computation can be found in the appendix D), but rather highlight the
Lattice Matches

The first and single most important feature in all selection methods is the (2 dimensional, i.e. in-plane) lattice match with the basal face of ice. This is on the one hand in agreement with our results of chapter 3, in that a good lattice match resulted in a good ice nucleation rate. On the other hand we find many exceptions. In Fig. 6.5 we see that there are many substrates that have a good match, but bad INA (blue) and a bad match but a good INA (red). This reinforces our reasoning that a lattice match is beneficial but neither sufficient nor a requirement. In terms of other ice faces we find that the match to the prism face is second most important, albeit not being selected as a feature due to its correlation with other features.
Liquid Density

The average number of nearest neighbors is another feature that always appears on place 2. We find that substrates that are able to decrease the number of nearest neighbors due to structuring of the liquid are on average better nucleators. This has a straightforward physical interpretation: If the average number of nearest neighbors is higher in the liquid than in the crystal (which is true for water), then upon nucleating the emerging crystal has to “push” out the excess neighbors, likely resulting in an energetic cost. If a surface is able to reduce the number of neighbors on average in the liquid in its proximity, this will therefore benefit this region’s ability to nucleate ice. We find both substrates that decrease and increase the average amount of neighbors in the close-by liquid, which means that this principle could also be used to argue for nucleation-inhibiting surfaces. We are not aware of this microscopic density-principle being established for heterogeneous nucleation.

Adsorption Uniformity

Many of the features appearing in Fig. 6.4 are results of random structure searches to estimate statistics of adsorption energies of diverse ice-like structures, i.e. ranging from water monomers to hexamers to ice cages, etc. Interestingly, the statistics that appear are always measures of the spread of values (such as variance or entropy) and never a measure of absolute values (such as mean or median). We interpret this as the uniformity of the adsorption environment with a more uniform environment (smaller variance) being more beneficial to ice nucleation.

Another unexpected finding is that not the most ice-similar structures are the most relevant ones for this measure. For instance, we see that the adsorption uniformity for tetra- and pentamers is most important, followed by mono- and dimers. The hexamers or cages investigated were less relevant and also received a lower importance rating. This perhaps indicates that interfacial structures that are not necessarily ice-like play an important role and a good ice nucleator does not require the interfacial water layers to be precisely ice-like.
Local Ordering

The fourth class of features relates to the local ordering of the liquid by the substrate. In the computation of these we made sure to use slightly above coexistence temperature to avoid accidentally having pre-formed ice at the surface. The best measure of local ordering was the $l_{q_3}$ measure (see chapter 4) closely followed by the $q_3$, $l_{q_6}$ and $l_{q_{12}}$. This is an interesting finding since with this we have established that the liquid ordering at coexistence temperature is indicative of the nucleation temperature and thus the INA of the surface. This can easily be probed by simulations, but also experiments might be able to measure a quantity that relates to the tetrahedral ordering probed by the $l_{q_3}$.

6.4 Conclusions

In this chapter we started by establishing the INA of a large number of diverse substrates in contact with supercooled water as well as computing a vast number of potentially descriptive features. We then devised a data-driven approach based on machine-learning methods to select features descriptive of the INA and established that the resulting models are indeed predictive. Our main findings can be summarized as:

- There is no single descriptor that can reasonably well predict the INA. For each of the devised important descriptors there are exceptions of the INA if used in isolation.

- Using these four microscopic principles we can predict the INA of a given substrate:
  
  - Lattice match of the substrate to ice.
  
  - Density reduction of the liquid near the surface.
  
  - As uniform as possible adsorption environment (for tetra- and pentamers rather than hexamers).
  
  - Local tetrahedral ordering of the liquid near the surface, ideally measured with the $l_{q_3}$.
The experimental verification of these guidelines could be achieved if methods are designed to specifically probe the above descriptors. The lattice match can already be probed by crystallographic methods. Density and local ordering could eventually be probed by developments of surface x-ray [271, 272] or sum-frequency generation spectroscopy [249, 273]. The adsorption uniformity is related to how pristine and extended a perfect patch of substrate is and could thus be studied with TEM [274, 275], AFM [276] or STM [87, 277, 278] methods. Furthermore, the computational evaluation of these descriptors is straight-forward, even for more complex water and interface models.

With this work we performed a first step at identifying a new and predictive set of microscopic descriptors for the INA of substrates and show that the quantitative predictability of heterogeneous ice nucleation is in reach. Further work is needed to refine the descriptors we found into simple, experimentally accessible quantities.
Chapter 7

Truncated Interaction Potentials Can Lead to Unphysical Wetting States in Simulations of Liquid Interfaces

Non-bonded potentials are included in most force fields and therefore widely used in classical molecular dynamics (MD) simulations of materials and interfacial phenomena. It is commonplace to truncate these potentials for computational efficiency based on the assumption that errors are negligible for reasonable cutoffs or compensated for by adjusting other interaction parameters. Arising from a metadynamics study of a wetting transition of water on a solid substrate we find that the influence of the cutoff is unexpectedly strong and can change the character of the wetting transition from continuous to first order by creating artificial metastable wetting states. Common cutoff corrections such as the use of a force switching function, a shifted potential or a shifted force do not avoid this. Such a qualitative difference urges caution and suggests that using truncated non-bonded potentials can induce unphysical behavior that cannot be fully accounted for by adjusting other interaction parameters.

7.1 Introduction

Short- to medium-range potentials such as the Lennard-Jones [279] or the Buckingham [280] potential are the backbone of classical MD simulations. They represent
Pauli repulsion as well as non-directional dispersion attraction and there exist multiple flavors implemented in most MD codes under the term of non-bonded interactions. In practice there is a need to truncate these potentials since the number of neighbors that have to be considered for each entity grows enormously, drastically increasing the computational cost for the force calculation. Truncating between \( r_c = 2.5\sigma \) and \( 3.5\sigma \), where \( \sigma \) is the characteristic interaction range, is a very common practice in MD studies [48] and has become the minimum standard, assuming that errors arising from this are small enough. Several studies have reported that with these settings significant problems can arise. For instance the truncation can alter the phase diagram of the Lennard-Jones system [281, 282] or yield different values for interfacial free energies [283–287]. These effects are quantitative in nature, meaning that they can in certain circumstances be analytically corrected for [288–290] or compensated for by other interaction parameters such as interaction strength or interaction range. The latter is important for the development of force fields where non-bonded potentials are often included and the cutoff can be seen as another fitting parameter. Naturally, a parametrization with a small cutoff would be preferred to another one if they deliver equal accuracy. This however is only true in the assumption that the underlying physical characteristics that are created by truncated and longer ranging potentials are the same.

In this chapter we investigated the influence of the cutoff for the interfacial phenomenon of water-wetting on a solid substrate. We found that the effect of the cutoff of the water-substrate interaction was not only unexpectedly strong, but also changed the fundamental physics of the wetting transition in an unprecedented way by creating metastable wetting states that have also never been seen in experiments. We show that proposed cutoff corrections such as the use of a force switching function, a shifted potential or a shifted force did not fix this and could even worsen the effect. This finding shows that atomistic simulations of interfaces need to be treated with great care since unphysical behavior could occur and easily remain undetected. This is particularly relevant since a large number of MD studies using truncated potentials are reported each year. Our results suggest the use of much larger-than-
common cutoffs or long-range versions of non-bonded potentials in MD studies of wetting and interfacial phenomena. We explicitly note that the type of wetting transition studied here, which was the one between droplet and film at constant particle number, should not be confused with other wetting transitions involving the vapor phase at varying particle number [291].

7.2 Computational Settings

We investigated two droplets comprised of 3000 and 18000 water molecules which were represented by the coarse-grained mW model [29], on top of a rigid, pristine fcc(100) surface (lattice parameter 4.15 Å). Whilst this substrate does not aim at representing any particular material, similar systems have been used to study ice nucleation [31, 112, 113, 145, 208, 292, 293] or water-metal interfaces [294, 295]. The simulation cell had dimensions $17 \times 17 \times 11$ nm which is enough to avoid interaction of the water molecules with their periodic images for all wetting states. Even though the liquid is rather non-volatile even at the highest temperature considered, we employed a reflective wall at the top of the cell to avoid evaporation and mimic experimental conditions. Our simulations were performed with the LAMMPS code [139], integrating the equations of motion with a timestep of 10 fs. This rather large timestep is commonly used in combination with the mW model and is acceptable for our system since during NVE simulations the total energy drift was found to be only about $2 \times 10^{-9}$ eV per water molecule per ps. In addition, we verified that we obtain the same results using standard protocols for updating the neighbor lists compared with unconditionally updating them every timestep. All production simulations were performed in the NVT ensemble with constant temperature maintained by a ten-fold Nosé-Hoover chain [42] with a relaxation time of 1 ps. The substrate-water interaction was given by a distance ($r$) dependent Lennard-Jones potential

$$U_{LJ}(r) = 4\varepsilon \left[ \left( \frac{\sigma}{r} \right)^{12} - \left( \frac{\sigma}{r} \right)^{6} \right]$$  

(7.1)
with $\varepsilon = 29.5$ meV, $\sigma = 2.5$ Å truncated at a cutoff $r_c$. This resulted in a maximum interaction energy of 154 meV for an adsorbed water monomer (weakly depending on the cutoff). Additionally we performed well-tempered metadynamics simulations [82, 84] for the smaller droplet with the PLUMED2 code [210]. In these simulations the Gaussian height, width, bias-factor and deposition stride were 2.16 meV, 0.15 Å, 20 and 20 ps respectively. Metadynamics is usually applied to drive rare events such as nucleation [25, 211, 242, 296] or protein folding [83, 297]. In our systems, this method helped to uncover the underlying free energy profile of wetting.

### 7.3 Results and Discussion

**Figure 7.1:** a) Side view of the two wetting states for the small droplet. Water is blue and surface atoms are gray. b) Temperature of the wetting transition $T_w$ (points) versus cutoff radius $r_c$ and fit (red line). The $T_w$ were obtained from the free energy profiles (see text) and we estimate errors to be $\pm 3$ K. $T_0$ is the converged wetting temperature.

We studied the wetting behavior of the larger droplet by performing standard MD runs at different temperatures first. As starting configurations we chose either
7.3. Results and Discussion

A flat water film in direct contact or a spherical droplet placed above the substrate. Within at most 5 ns the simulation was equilibrated and a seemingly stable configuration was reached, where the water is either wetting (contact angle $\theta = 0^\circ$) or partially wetting ($0^\circ < \theta < 180^\circ$). An illustration of the two wetting states can be found in Fig. 7.1a. Initially we employed a radial cutoff at $r_c = 3.0\sigma$ for the water-substrate interaction. With this setting we found that interestingly a wetting transition happened at finite angle $\theta_0 \approx 23^\circ$, i.e. a smaller non-zero contact angle was not possible. This behavior cannot be explained by the standard Young’s equation.

However, upon increasing the cutoff we found that the wetting behavior drastically changed. First, the wetting temperature $T_w$ at which the wetting transition took place increased as we increased the cutoff (Fig. 7.1b). Whilst $T_w$ shows a clear convergence behavior with $r_c$, it is unexpectedly slow. A reasonably converged wetting temperature $T_0$ is only reached for $r_c > 7\sigma$. Second, we noticed that for an increasing cutoff the minimum possible contact angle $\theta_0$ got smaller and eventually vanished. Most importantly, we also found that for temperatures around $T_w$ the stable configuration that was reached after the 5 ns could depend on the starting configuration for smaller cutoffs, while for larger $r_c$ it always reached the same state. This suggests that for small $r_c$ we actually found metastable wetting states that are absent for large $r_c$. This also means that $T_w$ cannot naively be defined through visual analysis of trajectories at different temperatures but needs to be defined by the free energy of wetting. For a first order phase transition we define $T_w$ to be the temperature where the two basins (corresponding to wetting and partial wetting) have the same free energy. For a continuous phase transition $T_w$ is the temperature where the single basin represents a contact angle of $\theta = 0^\circ$ for $T < T_w$ and $\theta > 0^\circ$ for $T > T_w$.

Understanding the character of these wetting states with standard MD can prove difficult as the dependence on the starting configuration always leaves doubt on the outcome of the equilibrated configuration obtained from it. To clarify, we show the results from the metadynamics simulations in Fig. 7.2. As a collective vari-
Figure 7.2: Free energy profiles of wetting for different cutoffs in a small temperature range around the respective transition temperature $T_w$ (generally at or near the central column for each system). As collective variable we chose the center of mass of the water droplet (COM$_z$, substrate at $z = 0$). We note that for the largest cutoff of 8σ the temperature range is slightly larger to highlight the shape of the free energy profile for complete and partial wetting.

We chose the $z$-component of the center of mass of the water droplet (COM$_z$), where $z$ is the surface normal direction. While this choice is not equivalent to the contact angle (as they are related in a non-linear manner) it is clear that significantly different values for COM$_z$ correspond to different contact angles and can therefore distinguish the different wetting states. For the smallest cutoff at $T_w$ and around we found that two basins coexist, one being the flat film (COM$_z \approx 4$ Å) and the other being a droplet with certain contact angle (COM$_z \gtrsim 5$ Å). These two states are separated by a significant barrier larger than $20 k_B T$, which explains why we observed metastable states in the unbiased simulations for small $r_c$. This corresponds to a first-order phase transition between the wetting states. The occurrence of a minimum possible contact angle $\theta_0$ is explained by the existence of the second basin, which does not approach the wetting basin, but rather becomes less stable as temperature changes. However, this character faded as we increased $r_c$. The barrier became smaller and the distance between the basins got smaller. For the largest cutoff investigated (8σ) we clearly see that only a single basin exists that changes its position with temperature. As a result no metastable wetting states exist and
the phase transition is continuous. We note that in this case the estimate of $T_w$ is more difficult than for the first order transitions, however in this work we aim at presenting qualitative results and from Fig. 7.2 it is clear that $T_w$ is higher than for the smaller cutoffs.

Only the results for the largest cutoff are in agreement with the fact that water wetting transitions are generally continuous when probed in experiments [298, 299] and finite-angle wetting transitions have, to the best of our knowledge, never been observed experimentally. Therefore, the correct qualitative wetting behavior in our system is not achieved with standard cutoffs and if undetected could potentially lead to false conclusions. Differences between short and long-ranged interactions have been highlighted for other interfacial phenomena, such as drying [291] or grain boundary melting [300].

We further study the effect of the most commonly used correction schemes to cutoffs:

1. A shifted potential (sp) which ensures that the value of the potential energy $U$ does not jump at the cutoff distance, given by:

$$U_{sp}(r) = U_{LJ}(r) - U_{LJ}(r_c)$$  \hspace{1cm} (7.2)

The corresponding force $F$ remains unaltered:

$$F_{sp}(r) = F_{LJ}(r)$$  \hspace{1cm} (7.3)

2. A switching function (switch) which brings the force to zero between an inner $r_{c,1}$ and an outer cutoff $r_{c,2}$ (we chose $3\sigma$ and $4\sigma$):

$$F_{switch}(r) = F_{LJ}(r) \hspace{1cm} r \leq r_{c,1}$$  \hspace{1cm} (7.4)

$$F_{switch}(r) = \sum_{k=0}^{3} C_k (r - r_{c,1})^k \hspace{1cm} r_{c,1} < r \leq r_{c,2}$$

where $C_k$ are constants determined to ensure a smooth behavior [139].
3. A shifted-force potential (sf), which ensures that force and potential do not jump:

\[
U_{sf}(r) = U_{LJ}(r) - U_{LJ}(r_c) - (r - r_c)F_{LJ}(r_c)
\]

\[
F_{sf}(r) = F_{LJ}(r) - F_{LJ}(r_c)
\]  

The latter approach was found to give good results for a homogeneous system and even allowed for a reduction of the cutoff \[301\]. Our results for these three corrections can be found in Fig. 7.3. By definition and thus unsurprisingly, the shifted potential does not yield any significant difference (where the remaining minor deviations are due to the metadynamics sampling) over the plain cutoff since forces remain unaltered. The smooth cutoff via switching function seems to improve the situation, however the fact that the transition temperature lies between the ones we found for a plain cutoff at 3\(\sigma\) and 4\(\sigma\) suggests that the improvement stems from the effectively increased interaction range rather than the fact that the force vanishes smoothly. Interestingly, the shifted force with the same cutoff performs worst out of all candidates as the barrier increases by a factor of two, which increases the likelihood that simulations are performed in the metastable state without realizing it. The fact that none of the considered correction schemes significantly improved the character of the wetting free energy profile leads us to conclude that it is not the way in which the cutting is done that matters most, but rather the effective cutoff distance as well as the overall interaction strength at that distance.

As an initial attempt to understand the results obtained we looked at the potential energies of the various systems with the different cutoffs considered. This, however, did not reveal any obvious explanation. One possible interpretation for the creation of metastable states in our systems with shorter cutoff can be obtained by considering the droplet state (not assuming anything about the stability relative to the film state). For a transition towards the film state, there needs to be thermal fluctuations of water molecules that are above the contact layer in the downwards direction (the fact that COM\(_z\) has proven a good reaction coordinate supports this
Figure 7.3: Free energy profiles of wetting approximately at the transition temperature with uncorrected setup (cut) and for different correction schemes [shifted potential (sp), force switch (switch) and shifted force (sf)] applied with a cutoff at $3\sigma$. None of the schemes show the correct behavior, which is shown in Fig. 7.2 to be a single basin.

...
The entropic change between these two states will be monotonic for a single water molecule and therefore also for the whole droplet.

Finding a general recipe for how to avoid such unphysical wetting states is difficult. Other aspects like e.g. the substrate density or the liquid-liquid interaction strength will have an influence on how strongly the fluctuations in the droplet state are affected by $r_c$. Generally, cutoffs that are deemed acceptable from the intermolecular perspective do not necessarily mean that the interaction between macroscopic states such as a film/droplet and a substrate is sufficiently captured. This is especially important in an interfacial simulation setting such as a slab, where a cutoff-caused change in interaction from the substrate side is not compensated by an equal change from the vacuum side. Consequently, only employing much larger cutoffs or techniques to calculate the long-range part of the dispersion force [302–304] can ensure that unphysical effects are avoided. A minimal sanity check for future wetting studies could be to start simulations from both a wetting film and a spherical liquid snapshot. If both of them end up in the same configuration the existence of an unphysical metastable wetting state is unlikely.

In light of the vast amount of work that is done in the MD community using similar interactions, our findings urge extreme caution when dealing with truncated non-bonded potentials in simulations of interfacial phenomena [305]. We have seen both quantitative and qualitative differences for the wetting transition studied here. The former could be accounted for by changing other interaction parameters to reproduce the transition at the right temperature $T_0$. This assumption is fundamental to fitting force fields with truncated potentials to obtain quantitative agreement with e.g. experimental values. But it does not hold for the character of the transition because it arises purely from the value of the cutoff itself. If the resulting metastability of states remains undetected, the use of truncated interaction potentials could lead to wrong inferences about physical properties being made. While this conclusion has resulted from a simulation of wetting, similar implications could hold for other interfacial phenomena such as capillary flow [306, 307],
evaporation/condensation [308, 309], mixtures [310–312] or heterogeneous nucleation [111, 200, 313, 314] where it is commonplace to use truncated interactions.

\footnote{We have indeed done a check for a few selected systems from chapter 3, recalculating their nucleation rate with a much larger cutoff for the substrate-water interaction (40 Å). The resulting rates never differed more than a factor of 2 from the previously calculated rates, thus not changing any of our conclusions from that chapter.}
Chapter 8

Summary and Outlook

In this thesis we studied heterogeneous and homogeneous ice nucleation from different perspectives, utilizing computational and data-enabled methods. In chapter 3 we investigated the heterogeneous nucleation rate on top of four model fcc substrates with a coarse-grained water model in a screening approach. Due to the well-controlled character of the systems we were able to disentangle different contributions to the nucleation enhancement, the most important finding being that neither the lattice match nor the interaction strength with water alone can describe the nucleation rate but rather that there can be a non-trivial interplay between the two. This is for instance contrary to the most prominent “rule” that a good ice nucleator requires a good lattice match to ice. Whilst some experiments already showed that materials that do not have an obvious match to ice are good at ice nucleation, the microscopic interface at the surface of impurities could still form ice-like structures due to e.g. surface reconstruction. Our work has for the first time shown counterexamples to the lattice match requirement at the molecular scale. We rationalized this in the occurrence of buckled overlayers that seemed to facilitate ice nucleation when their periodicity coincides with the one of ice, even if the in-plane structure was not ice-like. Additionally, we also found a remarkable sensitivity to the surface properties as small changes could not only drastically alter the nucleation rate but also what type of ice formed. We saw that even on the same surface up to three different crystal faces of ice could be formed. These findings make very clear that to truly understand the ice nucleation ability of a given material, its surface structure under
wet and ambient conditions must be known and reasoning purely based on clean bulk structures might be inadequate.

The findings of chapter 3 lead to the question of the role of different polymorphs in heterogeneous nucleation of ice and in general. This aspect was investigated further in chapter 4 where we studied in great detail the ice nucleation mechanism of two particular model substrates that exhibit the same enhanced heterogeneous nucleation rate. We find that due to the fact that a surface can facilitate a specific face of ice, it can also form a metastable polytype by nucleating a face of that polymorph that is not found in other polymorphs. This leads to different free energy profiles that cannot be described by the heterogeneous extension of classical nucleation theory. This shows that models, like e.g. the ones used in cloud science assuming a single polymorph for all temperatures and aerosol impurities, can be inadequate since surface induced polymorphism alters the nucleation rate in non-trivial ways. We have further introduced the concept of pre-critical fluctuations and studied their role. We found that in principle they can be indicative, if not descriptive of the nucleation enhancement, if the polymorph is properly taken into account. More work is needed to establish a quantitative framework, the weakest part currently being the reliance on classical nucleation theory and its downsides (e.g. not accounting for stacking-disorder). However, the principle of studying pre-critical fluctuations quantitatively can be applied to any nucleation framework and also is not restricted to Boltzmann sampling. It is imaginable that the initial steps of e.g. a metadynamics or forward-flux simulation which go beyond cluster sizes of unbiased sampling could be harnessed to compare pre-critical fluctuations, which would be much cheaper than always studying the total nucleation mechanism.

In chapter 5 we elucidated the role of liquid dynamics in homogeneous ice nucleation, an aspect that often goes unnoticed. This is the first work dealing with the dynamics at onset and during the nucleation process of any liquid at the microscopic level. We found that the dynamical heterogeneity in supercooled water comes with a preference of nucleation happening in immobile regions. We also elucidated the structural differences in terms of the rings network in regions of different liquid mo-
bility. This work provides a first step at investigating liquid dynamics and its role on the nucleation process in heterogeneous scenarios. For instance it is clear that at surfaces, in confinement or in solution the liquid dynamics could be drastically altered; and in here we built the fundamentals to comprehend and further study these changes. Indeed, we have preliminary results that can explain the fact that nucleation in thin free-standing water films happens in the bulk-interior but is orders of magnitude faster than the bulk nucleation rate [315].

Chapter 6 builds on chapter 3 in that we combined the structures used therein with many more model systems that were utilized in other studies. After establishing the ice nucleation activity of all systems in terms of their nucleation temperature we use a data-driven feature engineering approach to distinguish good and bad nucleators and predict the nucleation temperature. This not only demonstrates that the ice nucleation ability of substrates can be quantitatively predicted for a diverse set of systems (which is something that to the best of our knowledge has never been achieved), but also yields a set of four surface and interfacial properties that are most relevant to heterogeneous ice nucleation. Among these is the principle of local density reduction, a novel aspect that could find its way into theoretical modeling of the nucleation process. In ongoing work we are building on this to create an easily applicable and interpretable model utilizing recent advancements in the machine-learning community.

Finally, chapter 7 elucidated a computational artifact that was found during our studies on droplets at surfaces. The setup we used is extremely common in state-of-the-art simulations of interfaces. This urges to great caution when dealing with truncated non-bonded potentials. In fact, the computational tools to treat non-bonded interactions with Ewald methods just as is routinely done for coulombic interactions are there, however most of these are ignored as errors are assumed to be small and in any case quantitative. The fact that we showed the standard treatment of truncated interaction potentials can yield unphysical (and most importantly qualitatively wrong) behavior will hopefully impact the treatment of such interactions.
Chapter 8. Summary and Outlook

The future of the computational field of heterogeneous ice nucleation studies lies mainly in two directions: First, more accurate modeling is needed to add new layers of complexity so that computer simulations eventually approach real materials. This includes all-atom force fields, which however already significantly complicate and prolong the simulation process. With the continued development of faster computers (with the potentially ground breaking changes quantum computers pose on the horizon) it might one day be possible to also account for properties such as polarization and nuclear quantum effects in nucleation simulations. As of today this can only be introduced in expensive studies of single systems, therefore potentially overlooking trends. Second, data-driven methods such as screening or high-throughput offer excellent possibilities to extract from simulations a more reliable set of rules. This approach however needs what is stated in the name: data. The traditional method of a computational study in materials science and physical chemistry considers few systems, if not a single one. For the statistical extraction of knowledge with the goal of predicting what makes a good or bad ice nucleator the reasoning needs to be based on a broader set of systems in general.

There is also a diverse set of topics in heterogeneous ice nucleation that at the time of writing this thesis remain opaque. Under certain conditions, the preferred nucleation at the triple-boundary between water, substrate and air has been observed [316], but no explanation is known. The microscopic mechanisms for shock nucleation or ice-recognition by organic molecules are unknown or at least debated. The formation of cubic ice in macroscopic quantities at ambient conditions or the targeted “blunting” of ice crystals in cryo-frozen matter could be attempted based on the insight presented herein. In contrast to organic materials, for inorganic materials it still seems unclear what actually resembles the “active site” that catalyzes the nucleation. It seems clear that it is not the flat outer surface of such material, but rather some characteristic found within trenches and cracks [11]. It is unknown whether this always is due to a metastable surface being present in such features [26] or whether other effects such as strong electric fields could be the cause of this. The influence of the liquid dynamics is also not well understood, but
could further our understanding of the nucleation in droplets, capillaries, nanotubes or at surfaces.

The rise of machine-learning in natural sciences creates an exciting time and could contribute to solving some of the problems mentioned above. While algorithms for analyzing big data and extracting features are available they will likely continue to improve to deliver more reliable, cheaper and more precise predictions. In addition, the recent developments in machine-learning potentials [52, 54, 317–319] will create a middle ground, in that simulations accurate enough (and also capable of dealing with effects traditional force-fields can have trouble with, like e.g. dissociation) can be done on time scales that allow the simulation of nucleation.
Appendix A

Supplementary Material for Chapter 3

A.1 Distribution of the $\bar{q}_3(i)$ Order Parameter

![Figure A.1: Distribution for the $\bar{q}_3(i)$ parameter for liquid water, hexagonal ice $I_h$ and cubic ice $I_c$. The data was obtained by a short time simulation of the liquid and pristine crystals respectively (205 K, 4096 molecules, NPT, 10 ns). The dashed line indicates the threshold above which particles have been considered as solid.](image)

Nucleation was monitored by following the change in the potential energy. As a separate check, for selected trajectories we also monitored $N_{cls}$, the number of molecules in the largest solid-like cluster. The state of molecules was characterized by a modified version of the local $\bar{q}_3(i)$ parameter [142]. Figure A.1 shows the distribution of $\bar{q}_3(i)$ for different phases. We applied a cutoff of 3.2 Å for both the $\bar{q}_3(i)$ neighbor-list and the cluster algorithm.
A.2 Compressed Exponential Fit

![Graphs showing compressed exponential fit results for two dissimilar nucleation events.](image)

**Figure A.2:** Compressed exponential fitting results for two dissimilar nucleation events. $P_{\text{liq}}(t)$ (red circles) and fit after equation 2.33 (blue lines) for the (110) surface and $a_{\text{fcc}} = 3.9$ Å. a) $E_{\text{ads}} = 11.63$ kcal/mol and b) $E_{\text{ads}} = 5.3$ kcal/mol.

The simulation protocol involves an instantaneous quench from the equilibration temperature to the one at which we study nucleation. Because the system has to relax into quasi-equilibrium first, the nucleation rate increases with time, resulting in a deviation from perfect exponential characteristics. The effect of this non-exponential behavior can be appreciated in Figure A.2, where we show the $t_n$ datasets and the resulting $P_{\text{liq}}(t)$ for two dissimilar nucleation scenarios observed on the (110) surface as a function of the strength of the water-surface interaction. In the case of a) the nucleation typically proceeds on a timescale ranging from 1 to 100 ns, resulting in well behaved exponential decay ($\gamma \sim 1$ in equation 2.33 for the survival probability). On the other hand, the fitting of the data shown in Figure A.2b gave $\gamma \gg 1$, which in turn implies a nucleation rate that increases with time, as the timescale for $t_n$ (0.1-1 ns) is indeed comparable with the relaxation time of the system. This occurrence takes place mainly for those ($E_{\text{ads}}, a_{\text{fcc}}$) values for which we observe the basically instantaneous (10-1000 ps) formation of almost perfect ice-like overlayers on top of the surface.
A.3 Critical Nucleus Size on the (111) surface

To obtain an estimate of the critical nucleus size we performed a committor analysis [80] on the (111) surface. The results are depicted in Figure A.3 and suggest a critical nucleus size of circa 50 mW molecules, which is much smaller than our system size (4000 mW).

**Figure A.3**: Committor probability ($P_B$) with respect to the number $N_{cls}$ of mW molecules in the biggest ice-like cluster for the (111) surface ($E_{ads} = 1.04$ kcal/mol, $a_{fcc} = 3.90$ Å). Three different thresholds $N_B$ for the order parameter have been considered and reported. The analysis has been obtained by shooting 30 statistically independent MD runs (2 ns long) from 40 different starting configurations taken along a nucleation trajectory. The arrow marks the critical nucleus size $\approx 50$. 
A.4 Snapshots of Classified Regions

Figure A.4: Classified regions for the 111 surface. Snapshots are taken from regions indicated by the blue dot.

Figure A.5: Classified regions for the 211 surface. Snapshots are taken from regions indicated by the blue dot.
Figure A.6: Classified regions for the 100 surface. Snapshots are taken from regions indicated by the blue dot.

Figure A.7: Classified regions for the 110 surface. We note that the side view of the last region does not show the typical side perspective of the prism face. However, it results from a 90° rotation around the z axis of the typical view, as seen e.g. in the prism region of the (211) surface. Snapshots are taken from regions indicated by the blue dot.
A.5 Distribution of Pre-Critical Nuclei

The distribution of pre-critical nuclei in Figure A.8 is helpful for discussing what is the only common feature found for all of the four surfaces: inhibition of the nucleation rate for the smallest value of $E_{ads}$. For this interaction strength the molecules essentially face a hard wall which in turn could even hinder nucleation compared to the homogeneous case [31, 150–152]. To understand this we must first mention, what happens for the homogeneous bulk case (which we term Homo) and the case of a free standing water slab (called Homo$_{VAC}$) with two water-vacuum interfaces. The nucleation rate of Homo$_{VAC}$ will be lower than in the bulk case (Homo), which can be roughly rationalized in terms of a smaller volume available for the nuclei to appear than in the bulk case. This effect is visible in Figure A.8, where the distribution for Homo corresponds to a constant line, while the probability for Homo$_{VAC}$ is decreased towards the interface. In fact, the nucleation rate constant for Homo$_{VAC}$ computed by excluding the volume of the system affected by the presence of the water-vacuum surface (which can be estimated by looking at the density profile along the z-coordinate, see SI Figure A.9a) is basically the same as obtained for
Homo. In the case of our models, the presence of the LJ surface could introduce significant density perturbations in the water film for all values of $E_{\text{ads}}$ (see SI Figure A.9b). As a result, when no efficient template can be provided by the surface, pre-critical nuclei tend to strictly avoid the neighborhood of the LJ surface as well, as reported in Figure A.8 (Inh). This effect could be even stronger than the inhibition coming from the water-vacuum interface, and as a result the effective volume available for the nuclei to appear is even less than in the Homo$_{\text{VAC}}$ case, thus causing a net inhibiting effect due to the presence of the surface. It is worth noting that while the promotion of the nucleation rate observed for many ($a_{\text{fcc}}, E_{\text{ads}}$) points can be rather strong, the inhibition effect is usually much weaker, as it basically accounts for the removal of the portion of the system affected by the presence of the LJ surface. This kind of inhibition is therefore unlikely to be visible in simulations or experiments where the ratio of water volume to contact area is much higher than in our case.
Figure A.9: a) z-Density profile of a homogeneous mW model (4000 molecules) $H$ and the same model as a free standing slab ($H_{VAC}$), featuring two water-vacuum interfaces. The x-axis refers to the distance from the center of mass of the mW slab along the z coordinate (only one side of the slab is shown). The shaded region highlights the fraction of the system affected by the presence of the water-vacuum interface because of density oscillations. b) z-Density profile of mW water on top of the (100) surface ($a_{fcc} = 3.9 \, \text{Å}$, obtained at 290 K) as a function of $E_{ads}$. 
A.6 Notes on the Water Model

Firstly, we checked that the water densities observed are not an artifact of the coarse grained water model by comparing them to the results of TIP4P/2005. The test was done for one \( (a_{fcc}, E_{ads}) \) point on each surface. The water-surface interaction was kept identical to the mW case, i.e. the surface atoms interact with the TIP4P/2005 oxygens through the same LJ potential. The resulting densities (see Figure A.10) are very similar. While the latter shows slightly stronger peaks and layering, the peak positions agree with the mW values. We conclude that the structuring exhibited by the mW model is nearly equivalent to the one of TIP4P/2005. Our results concerning the buckling and the structuring perpendicular to the surface should therefore be applicable to all-atom models of water. It appears that potential differences for overlayer patterns obtained from simulations are artifacts of the very different time scales on which both models evolve rather than actual structural differences.

![Figure A.10: Comparison of the density perpendicular to the surface for the mW and TIP4P/2005 models of water. Results are based on at least 75 ns long equilibration trajectories approximately 15 K above the melting point of the corresponding model. All graphs were computed for \( a_{fcc} \approx 3.9 \, \text{Å} \) and \( E_{ads} \approx 3.2 \, \text{kcal/mol} \).](image-url)

Secondly, we performed nucleation simulations with TIP4P/2005 water on the
(111) surface for \( a_{fcc} = 3.90 \, \text{Å} \) and different values of \( E_{\text{ads}} \). Contrary to the mW case, using the all-atom model we did not observe the formation of a complete hexagonal overlayer within 100 ns. This might not seem a surprise, specifically because of the lack of a hydrogen bond network. This deficiency results in a much faster dynamics (we have estimated a mismatch in the self-diffusion coefficient of about three orders of magnitude at the supercooling considered here) of the water molecules with respect to both experiments and basically any full atomistic water model [29, 320]. Besides, the mW model potential energy surface is much smoother than one in which hydrogen bonds would be taken into account. However, the number and the size of hexagonal patches within the first overlayer is consistent with what we have observed in the case of the mW model. In Figure A.11 we report the probability density function of the size of the biggest hexagonal patch of TIP4P/2005 water molecules on top of the (111) surface. The tails of the distributions, corresponding to sizable hexagonal patches can only be observed for \( E_{\text{ads}} = 3.2 \, \text{kcal/mol} \), which is exactly the value for which we observe the fastest formation of the hexagonal overlayer in the case of mW. To gauge interaction energies for different water models the heat of vaporization is often used. The latter for both models is nearly the same [29, 158] which means we can compare the adsorption energies directly. The trend holds for different supercooling as well, and confirms that while - especially heterogeneous - nucleation processes modeled by mW water are for certain non-physically fast, the model could still capture part of the underlying physics of the problem.
Figure A.11: Probability density distribution of the size of the biggest hexagonal patch of TIP4P/2005 water molecules within the first overlayer for different values of $E_{\text{ads}}$ and $a_{\text{fcc}}=3.90$ Å on the (111) surface. Results were taken from 20 ns trajectories sampled with a 2 fs timestep. The inset depicts a fairly large cluster of about 20 molecules.
A.7 Higher Temperatures

Our simulations have been performed in the deeply supercooled regime. It is beyond reach to do such an extensive set of simulations as performed here at significantly higher temperature. Nonetheless it is interesting to understand how the phenomena observed might depend on temperature. To estimate the effect of the strong supercooling on the results and especially the proposed mechanisms we performed calculations at higher temperatures for three adsorption energies on the (110) surface (depicted in Figure A.12). Since the computational cost significantly increases, the nucleation rates become less accessible with the brute force approach and we therefore limit this trial to only a few \((a_{\text{fcc}}, E_{\text{ads}})\) points. As previously mentioned, the values at the top and bottom range of the nucleation rate should be seen as a lower/upper bound to the actual nucleation rate. Conclusively, a missing temperature dependence of these points indicates a nucleation rate out of the limit that can be resolved with simulations of 500 ns length, rather than one that is constant with temperature. We find that the trends seen at 205 K are stable against the

Figure A.12: Temperature dependence of nucleation rates (circles) and spline interpolation (colored lines) for 3 adsorption energies on the (110) surface. All values were normalized by the homogeneous nucleation rate \(J_0\) at 205 K.
temperature increase and only in regions where no specific mechanism has been attributed the rates heavily decline. In fact the increased temperature can help to identify the values of $a_{fcc}$ inducing a certain mechanism in a more precise way because the gaps between the enhanced regions increase. Furthermore, the structures of the adsorption layers did not show any noticeable change so that we can assume that our conclusions are valid also for higher temperatures.
Appendix B

Supplementary Material for Chapter 4

B.1 Computational Details

Brute-force simulations of heterogeneous ice nucleation were performed with 18000 water molecules, represented by the coarse-grained mW model [29], placed in a film geometry on top of two pristine, rigid fcc surfaces (termed s1 and s2). The substrate-water interaction is given by a Lennard-Jones interaction (cut-off at 3\(\sigma\)), tuned to achieve the same absolute nucleation rate. We established that the pitfall of finite-size effects is already avoided for smaller systems containing only 4000 water molecules [208]. Generally, finding two impurities that nucleate different polymorphs at the same rate at the same temperature and study the nucleation event itself on them can be considered a very costly endeavor. Following established protocols [112, 113, 208] we first equilibrate each structure for 10 ns at 300 K. Then production runs are quenched to the target temperature and coupled to a 10-fold Nosé-Hoover chain [42, 141] to sample the NVT ensemble, integrating the equations of motion with a timestep of 10 fs. The relaxation time after the quench is on the order of 10 ps and can thus be considered non-disturbing to the nucleation. The nucleation events themselves are detected by a sudden drop in the potential energy, upon which we terminate the computation and collect the current time as induction time. 100 simulations for each of the two substrates at 218 K and 50 simulations
at 221 K have been performed with LAMMPS [139]. From the collection of induction times we fit the survival probability of the supercooled liquid to obtain the nucleation rate. Supplementary Figure B.1 shows the substrate morphology and the resulting survival probabilities which are almost identical for both scenarios and temperatures. Considering the fact that nucleation rates usually differ by many orders of magnitude we can label the resulting rates from our two systems as identical.

![Figure B.1](image_url)

**Figure B.1**: Overview of simulation systems and nucleation behavior: a) Perspective view of a representative simulation box with approximate dimensions 12 × 12 × 7 nm. Water is blue and surface atoms are gray. b) Top and side views of the two model substrates s1 and s2. The highest, second highest and the lower layers are colored in different shades of gray. c) Survival probability of the liquid, used to determine the nucleation rate.

## B.2 Metadynamics Simulations

The brute-force simulations have the disadvantage that pre-critical and critical clusters are similar in size. To see if our findings also hold for conditions where these are very different in size we perform simulations at a higher temperature of 235 K. At this undercooling we expect the hom. critical cluster size to be ≈ 600 [111] and therefore this temperature is close to the upper limit of temperatures that can be investigated with a system of our size (∼10,000 molecules).

To facilitate the nucleation we start by constructing the permutation invariant vector (PIV [213, 214]) V for the system. This entails computing the adjacency matrix of oxygens and reducing it to the irreducible part (then taken as a vector).
**Figure B.2:** One-dimensional free energy profiles as a function of the variable $s$ which describes the direct path from the liquid ($s \approx 1.1$) to the fully frozen simulation box ($s \approx 1.9$). The position of the repulsive soft-wall is indicated by the dashed grey line at $s = 1.5$. The results obtained via the standard use of the bias-potential (solid blue line) agree well with the estimates obtained by reweighting (dashed red line).

The distance ($r$) dependent criteria for adjacency was softened with a switching function $f_{\text{switch}}(r) = \frac{1-(r/r_0)^n}{1-(r/r_0)^m}$ with $n = 4$, $m = 12$ and $r_0 = 3.4$ Å, i.e. considering only nearest neighbors. The final PIV vector is then sorted, rendering it invariant under permutation of identical atoms. The PIV is also constructed for two reference snapshots ($V_A$ for a liquid and $V_B$ for a fully frozen simulation cell generated at 205 K). From these vectors we compute generalized distances (as squared Euclidean distances) $\mathcal{D}_A = \|V - V_A\|^2$ and $\mathcal{D}_B = \|V - V_B\|^2$. We use these distances to create a path [212] from basin A to basin B, defining the variables

$$s = \frac{1 \cdot e^{-\lambda \mathcal{D}_A} + 2 \cdot e^{-\lambda \mathcal{D}_B}}{e^{-\lambda \mathcal{D}_A} + e^{-\lambda \mathcal{D}_B}}$$

$$z = -\lambda^{-1} \log \left( e^{-\lambda \mathcal{D}_A} + e^{-\lambda \mathcal{D}_B} \right)$$

where $s$ measures the direct progression from A to B, while $z$ is a measure for the...
movement perpendicular to this path. For the parameter $\lambda$ we chose $\lambda = \frac{2.3}{D_{AB}}$ where $D_{AB}$ is the generalized distance between the two references states. This choice leads to the free energy basins for A / B being around $s \approx 1.1 / 1.9$. In addition we add a biasing potential

$$V_{\text{wall}}(s) = k (s - 1.5)^2 \cdot \theta(s - 1.5)$$  \hspace{1cm} (B.2)

with $k = 10,000$ kcal/mol and the Heaviside step function $\theta$. This represents a soft-wall to avoid the system getting trapped for too long in the crystalline basin corresponding to the fully frozen cell. The system for cluster sizes relevant to nucleation ($s < 1.5$) is unaffected by the wall.

We emphasize that this order parameter approach does not enforce or facilitate any specific ice polymorph as path variables do not make any assumption about the transition and the PIV we employ describes only the nearest neighbor environment which is identical for I$_h$ and I$_c$.

Finally, we perform well-tempered metadynamics [82, 84] with 20 walkers [85] on the variables $s$ and $z$ with parameters gaussian height $\delta = 0.2$ kcal/mol, gaussian width $(\sigma_s, \sigma_z) = (0.022, 0.38)$, deposition stride 2 ps and a biasfactor of 50. Simulations were done with a PLUMED [210] patched version of LAMMPS [139].

To check if our simulations are reasonably converged we compare the one-dimensional free energy profiles obtained from integrating out the $z$-degree of freedom, i.e.

$$F(s) = -\beta^{-1} \cdot \ln \left[ \int \exp(-\beta F(s,z)) \, dz \right]$$  \hspace{1cm} (B.3)

with the one obtained via the reweighting algorithm from Tiwary and Parinello [215]:

$$F(s) = -\beta^{-1} \cdot \ln \left\langle \exp[\beta V(\tilde{s}(t),\tilde{z}(t),t) - c(t)] \cdot \delta(\tilde{s}(t) - s) \right\rangle_t$$  \hspace{1cm} (B.4)

where the weight (exponential factor depending on the metadynamics bias $V$ and a time-dependent constant $c(t) = \frac{1}{k_BT}$) is calculated on the fly and the average $\langle ... \rangle_t$ goes over all snapshots generated during the MD simulation. The cubicity reported in the main text figure 3 was also obtained from these trajectories with
the corresponding weights applied to obtain the unbiased distribution. The fact that we get very similar profiles (see Supplementary Figure B.2) indicates that our simulations are reasonably converged and by all means suitable for the qualitative discussion in the main text.

Since the metadynamics simulation in the variables $s$ and $z$ does not directly yield the critical cluster size we performed a committor analysis [80] seeded from these metadynamics trajectories (we have restricted our selection to one starting point for each $n$ since the size of the biggest cluster is known to be a good reaction coordinate [111]). For each point we performed 200 simulations that were 2 ns long. As target values for the basins A and B we chose $n_A = 30$ and $n_B = 1000$. The committor probability $p_B(n)$ is then fitted according to:

$$p_B(n) = \frac{1}{2} \left[ 1 + \tanh \left( \frac{n - n_c}{a} \right) \right] \quad (B.5)$$

where $a$ and $n_c$ are fitting parameters. From the plot in Supplementary Figure B.3 we obtain $n_{c,s1} = 211 \pm 11$, $n_{c,s2} = 104 \pm 3$ and $a_{s1} = 121 \pm 20$, $a_{s2} = 22 \pm 3$. The errors are estimated 95% confidence intervals. Since the width of the committor
curve around $n_c$ is related to the curvature of the free energy $F(n)$ around $n_c$ the values of $a$ confirm again our results for lower temperatures, where $s1$ has a broader profile around the transition state (larger $a$) than $s2$. The differences here are more pronounced than for the lower temperature, suggesting that the influence of the different polymorphs is even more significant for higher temperatures.

### B.3 Physical Origin of the Nucleation Enhancement

![Figure B.4](image)

**Figure B.4:** Structural properties of water on the two substrates: a) Number density of supercooled water in contact with the substrate (at $z = 0$ Å) obtained from trajectories at 218K, sampling only the pre-nucleation part. b) Snapshot of the layers (top-down view of a portion of the simulation cell) corresponding to the peaks with the same color in the density plots.

A look at the density of liquid water on top of the substrates (see Supplementary Figure B.4a) reveals distinct peaks. The corresponding layers show how for $s1$ there is an in-plane templating effect as (stretched) hexagons in the first and to a lesser extent in the second layer are present even before nucleation. As a result in all simulations of $s1$ the basal face of ice has nucleated in contact with the surface.

The situation in $s2$ however is different since the first two layers exhibit a static non ice-like structure that does not change upon nucleation while the third layer is
disordered before freezing. It can be assumed that this is due to the corruga-
tion of the substrate s2. The first overlayer there is effectively comprised of chains
of molecules within the trenches and has little to no space to rearrange. Such an
example of a non-ice like first overlayer facilitating ice-like higher layers has e.g.
also been found from density functional calculations of feldspar [322] and in our
previous work [208]. The ice crystal face in contact with the substrate s2 is always
the prism face of hexagonal ice I_h. We suspect that the structure of the second layer
(blue) is key, because albeit being non ice-like the formation of stripes (see blue
layer for s2 in Supplementary Figure B.4) resembles a sub-structure in the prism
face of hexagonal ice I_h.

B.4 Cluster Asphericity

![Cluster Asphericity Graph](image)

**Figure B.5:** Difference in cluster asphericity between the two substrates: Scatter plot of the
asphericity parameter $\kappa$ versus COM$_z$ of 50000 pre-critical clusters (blue dots). Red circles
indicate the points for nearly-critical clusters.

We examined the morphology of clusters depending on their proximity to the
surface. In Supplementary Figure B.5 we plot of the asphericity parameter

$$\kappa = \sqrt{\frac{3\text{Tr}[(S - \frac{3}{3\text{Tr}S/3})^2]}{2(\text{Tr}S)^2}} \quad (B.6)$$

versus the z component of the center of mass (COM$_z$) of 50000 pre-critical clusters.
Appendix B. Supplementary Material for Chapter 4

(blue dots). $S$ is the gyration tensor of all molecules belonging to the respective cluster. Special values of $\kappa$ are 0 (perfect sphere), 0.5 (rod or pancake shaped) and 1 (chain). While $s_2$ clearly does not exhibit any density or asphericity increase the latter increases in $s_1$ when closer to the substrate, indicating pancake and chain morphologies. This further highlights that both systems, despite having the same nucleation rate, take two very distinct paths.

**B.5 Different Order Parameters**

We consider the influence of a different order parameter on the detection of clusters. For this purpose we calculated the cluster size distributions (see Supplementary Figure B.6) based on a Steinhardt parameter [323] with spherical harmonics of order 6 and spatial coarse-graining as defined by Lechner and Dellago [142]. The order parameter used in the main text used spherical harmonics of order 3 and phase averaging between neighbors [144]. The results qualitatively agree with the ones reported in the main text. The tails for $s_2$ seem so be more pronounced, which indicates that the order parameter is more generous in detecting clusters that are directly at the surface.

![Figure B.6: Validation of findings with other order parameters](image)

Figure B.6: Validation of findings with other order parameters 1: a) Distribution for the alternate order parameter $\overline{q}_6$. The dashed line indicates the cutoff used to distinguish liquid water from ice. b) Probability distribution for the size of the largest ice-like cluster $N_{\text{cls}}$ obtained from all trajectories at 218 K before the nucleation event happened.
The issue of the detection of solid-like molecules at the interface is important and debated. Despite being crystal-like an order parameter might not detect these molecules because of their interfacial environment. A straightforward way to consider them is to include all the neighbors of the detected cluster, i.e. the first hydration shell (molecules within 3.2 Å) in our case. While this adds a degree of arbitrariness the main conclusions should be robust to that. We show the results for the free energy profile and the cluster size distribution in Supplementary Figure B.7, where we denote the order parameter from the main text as \( N_{\text{cls}} \) and the one including the surrounding shell as \( N_{\text{cls}+\text{surf}} \). Additionally we considered a reconstruction as a function of the CNT coordinate \( n_{\text{cls}} \) since strictly speaking CNT deals with the free energy of an isolated cluster \( n_{\text{cls}} \). However, in a real system there will be many clusters of different sizes and the distribution of the biggest cluster \( P(N_{\text{cls}}) \) will only be equal to the distribution of clusters \( P(n_{\text{cls}}) \) for very large values of \( n_{\text{cls}} \). In our systems the amount of molecules that are ice-like before nucleation is still much smaller than their total number and cluster appearances can be seen as independent. Therefore we approximate \( P(n_{\text{cls}}) \) from the histogram of all clus-

Figure B.7: Validation of findings with other order parameters 2: The left column shows the free energy as a function of the number of molecules in the biggest ice-like cluster \( F(N_{\text{cls}}) \), same but additionally including the first water hydration shell \( F(N_{\text{cls}+\text{surf}}) \) and as a function of the CNT coordinate (isolated cluster) \( F(n_{\text{cls}}) \). The right column shows the corresponding cluster size distributions.
ters in all frames in all trajectories. We see that the enhancement of fluctuations is completely overshadowed by the weight of the many more smaller clusters (see \(P(n_{\text{cls}})\) in Supplementary Figure B.7) and therefore \(N_{\text{cls}}\) is a better coordinate to examine fluctuations. However the same trend found in the free energy profiles is also found for \(F\) as a function of \(n_{\text{cls}}\). The profile for system \(s2\) which nucleates a different-than-homogeneous polymorph is steeper and closer to the homogeneous line. Overall, we conclude that our results and conclusions are robust against different choices of order parameters.

### B.6 Higher Temperature with Brute-Force Simulations

![Graphs showing free energy profiles for systems s1 and s2 at 218 K and 221 K](image)

**Figure B.8:** Results for higher temperatures: Free energy profiles obtained by kinetic reconstruction for the two systems in this study at 218 and 221 K.

We consider a range of different parameters and settings that could potentially change our main results. First of all, we examine the effect of temperature. In Supplementary Figure B.8 we plot the resulting free energy profiles for both temperatures considered in this study. While the difference in temperature is only 3 K (the brute-force approach becomes exceedingly costly) the rate changes already by an order of magnitude (see main text). It is apparent that both systems do not change the trend we see for the lower temperatures. In particular the barrier height for \(s2\) is larger than for \(s1\), while the transition state is located at a smaller value of \(N_{\text{cls}}\) for \(s2\) than for \(s1\). We conclude that indeed the curves for each substrate belong to different families of curves, representing the homogeneous path way they are based on (\(s1\) is based on the dominant homogeneous path way, while \(s2\) follows the
B.7 Extension to hetCNT

Having shown that the heterogeneous classical nucleation theory (hetCNT), as traditionally applied, cannot describe the phenomena observed in our simulations, we now lay out an extension of hetCNT which helps to make it more general by accounting for different bulk references. We note that it is not our aim to include corrections for several of the already known possible shortcomings of CNT and hetCNT (e.g. neglect of the line tension [216, 324]) but we rather focus solely on how to account for a change in polymorph induced by the substrate. CNT yields the following expressions for the free energy barrier $\Delta F$ and the number of atoms/molecules in the critical nucleus $n_c$:

$$\Delta F_{\text{hom}} = \frac{16\pi}{3} \frac{\sigma^3}{(\Delta \mu)^3}$$
$$n_{c,\text{hom}} = \frac{32\pi}{3} \frac{\sigma^3}{(\Delta \mu)^3} \cdot \rho_{\text{cry}}$$

(B.7)

where $\sigma$ is the interfacial free energy between solid and liquid, $\Delta \mu = \mu_{\text{liq}} - \mu_{\text{cry}}$ is the chemical potential difference to the liquid and $\rho_{\text{cry}}$ is the number density for the crystalline phase. The assumptions implied in this are: i) the nucleus has spherical shape, ii) thermodynamic properties of small clusters are assumed to be the values of the bulk and iii) a well-defined surface that separates cluster from liquid. For the step towards heterogeneous CNT we denote with the subscript 1 and 2 the two different polymorphs (in our case 1 = stacking-disordered and 2 = hexagonal) occurring in the systems s1 and s2 respectively. From hetCNT we obtain the resulting barrier and critical nucleus size for system 1 where polymorph 1 has
formed:

\[ \Delta F_1 = f_{V,1}(\theta_1) \cdot \Delta F_{\text{hom},1} \]
\[ n_{c,1} = f_{V,1}(\theta_1) \cdot n_{c,1}^{\text{hom},1} \]

(B.8)

where \( f_{V,1} \) is the volumetric factor of polymorph 1. In the case of system 2 the comparison to the hom. formation of polymorph 2 would be straightforward. This however is unlikely to be relevant since kinetics for the hom. pathway of polymorph 2 are unknown and much harder (if not impossible) to measure than those for the dominant polymorph 1. Therefore we introduce correction factors to equations B.7 to compare to polymorph 1:

\[ \Delta F_2 = f_{V,2}(\theta_2) \cdot \Delta F_{\text{hom},1} \cdot \left( \frac{\Delta \mu_1}{\Delta \mu_2} \right)^2 \left( \frac{\sigma_2}{\sigma_1} \right)^3 \]
\[ n_{c,2} = f_{V,2}(\theta_2) \cdot n_{c,1}^{\text{hom},1} \cdot \left( \frac{\Delta \mu_1}{\Delta \mu_2} \right)^3 \left( \frac{\sigma_2}{\sigma_1} \right)^3 \left( \frac{\rho_2}{\rho_1} \right) \]

(B.9)

Note that the scaling with \( \Delta \mu \) is different for the barrier height and the critical nucleus size, which results in a different ratio \( \chi = \frac{\Delta F}{n_c} \). hetCNT predicts that \( \chi \) is independent and the same for all supercoolings since the enhancement factors cancel out. However our comparison of systems s1 and s2 clearly indicates that this ratio is not the same in the two systems. Therefore, this extension may account for why heterogeneous free energy profiles such as our results in the main text figure 2c can have different functional shapes as opposed to the prediction of hetCNT. Note that our extension has not introduced any further assumptions, but we have solely used the tools supplied by CNT to illustrate how a change in polymorph needs to be included in the theory. Since \( \frac{\Delta F_1}{n_{c,2}} \cdot \frac{n_{c,2}}{n_{c,1}} = \left( \frac{\Delta \mu_1}{\Delta \mu_2} \right) \left( \frac{\rho_2}{\rho_1} \right) \), equations B.9 also have the potential to estimate the chemical potential difference of different polymorphs from nucleation data, provided the approximations of CNT are reasonable.
Appendix C

Supplementary Material for Chapter 5

C.1 The Role of Statistics for calculating the DP

\[ \langle \Delta \text{DP}_i \rangle \% \]

Figure C.1: Assessment of the role of statistics on the calculation of dynamical propensity (DP). Shown is the average difference (in \%) of the calculated DP value for each molecule, when using \( n \) instead of 200 runs to calculate the square-displacement. The estimate was done with Jackknife re-sampling [321].

The growing time of maximum heterogeneity \( t_0 \) with increasing supercooling also increases the cost of the iso-configurational analysis. Therefore we aim at choosing a number of shootings \( n \) for each snapshot that is a compromise between a reasonable accuracy of the dynamical propensity and computational saving. In figure C.1 we show an assessment of the influence of choice of \( n \). We estimated the
average difference of all DP values in a given snapshot obtained with \( n \) shootings compared to the estimate with 200 shootings, where we used the latter number to characterize snapshots for the rings-analysis. We see from this that even halving the number of runs to \( n = 100 \) would introduce an average error of only less than 10\% on the DP values which is not enough to qualitatively change any of our conclusions.

For certain parts of this work many frames had to be analyzed regarding their DP distribution. This is a substantial computational effort, especially for the atomistic model which has a time of maximum heterogeneity \( t_0 \) (which we set as length of each run) of more than 100 ps at 240 K and more than 600 ps at 230 K. Thus we reduced the number of runs for the ISOCOA from 200 to 40 for (i) the overlap calculation of regions with ice-like clusters (Fig. 3b in the main text); (ii) the plot of the temporal evolution of the DP before cluster assembly (Fig 3c in the main text); and (iii) the supplementary video. As we can see in figure C.1, large average deviations for the DP values are only obtained for choices of \( n < 20 \). For \( n = 40 \) we obtain an average deviation of 15\%. We can see that doubling this number would marginally increase the accuracy while also doubling the cost for simulation. Furthermore, an average deviation of 15\% is unlikely to change our results regarding the structural analysis of the resulting domains since this is not enough to significantly change the ranking of the top/bottom 5\% of DP values.

### C.2 DP Threshold Choice for Defining MI and MM Domains

To investigate the structural characteristics of MI and MM domains we make a DP threshold choice \( \varepsilon \) for the fraction of molecules we will label as mobile/immobile if they belong to the top/bottom \( \varepsilon \)-fraction of DP values. This choice is in principle arbitrary and since in the main text we reported results for a choice of \( \varepsilon = 5\% \), we here show the results of the rings-analysis for halving/doubling this threshold. In figure C.2 we can clearly see that the characteristic occurrence of 6-membered rings for the MI domain and of smaller / less hydrogen-bonded rings in the MM region is not altered by that choice. This suggests that our findings are truly characteristic
C.3 Domain Characterization With Other Order Parameters

After having identified the MI and MM domains we performed comprehensive structural analysis in terms of rings population and hydrogen bonding. In figure C.3 we also report the distributions for other order parameters for the aforementioned regions as well as in the overall snapshot. The measures for tetrahedrality $q_{\text{tetra}}$ and local structuring $l_{q3}$ and $l_{q6}$ all show minor differences between the regions. The most important result from this is the fact that the overlap with the ice-like region of the respective order parameter (indicated in gray) is marginal, suggesting that a drop in mobility could precede nucleation.

Lastly we note that an analysis of the MI and MM domains in terms of hexagonal and double-diamond cages (see e.g. Refs.161, 199) did not yield any significant occurrence of such ($\sim 1$ per snapshot), further indicating that our configurations
Figure C.3: Characterization of different domains by distributions of customarily used order parameters: the tetrahedrality measure for each molecule $q_{tetra}$ [224] and phase-averaged local Steinhardt parameters [144, 323], using spherical harmonics of order 3 and 6 ($lq3$ and $lq6$). lines for the most immobile (MI) / most mobile (MM) region are blue/red while averages over the whole simulation cell are black. The gray shaded regions show values of the order parameter that indicate an ice-like environment.

even at the strongest supercooling are still liquid.

C.4 Characterization of Dynamics for the mW Model

The entire workflow performed mostly on the all-atom model was repeated for the mW model at coexistence (274 K) and close to the homogeneous freezing temperature (205 K). In figure C.4 we see that the extent of dynamical heterogeneity is much weaker than for the all-atom model. Only at the low temperature (which corresponds to a supercooling of $\approx 70$ K) there is a hint of heterogeneous dynamics comparable to that of the other model for $\approx 10$ K supercooling. This is expected and in agreement with the fact that this liquid has a larger diffusion coefficient [29]. However, the population of the rings network shows similar signatures regarding the type and amount of rings that is seen in the MI and MM regions. We take this
C.5 Transition Path Sampling for the mW Model

When performing a transition path sampling (TPS) study we need to address whether or not we are effectively sampling different trajectories. In Fig. C.5 we show the potential energy vs time for all of the trajectories harvested using TPS (76 trajectories from a total of 7500 TPS moves, saving every 100 trajectories.) The plot clearly illustrates that many types of trajectory are being sampled: Some have long induction times while others look like they start nucleating straightaway; others reach a much lower potential energy state than others too. It also appears that there is a range of crossing times. Perhaps most importantly, the trajectories go back-and-forth between these different types of trajectory, suggesting that we are actually sampling a steady state distribution in trajectory space rather than still
relaxing the trajectories toward some steady state.

![Graph](image)

**Figure C.5:** Potential energy versus time for the trajectories harvested with TPS.

Fig. C.6 shows snapshots from two trajectories that differ by 100 TPS moves. Ice-like molecules, with $l q_6 > 0.5$, are shown in red and blue for the different trajectories. It can clearly be seen that the ice nucleus is forming in different regions of space (compare the snapshot at 0.10 ns and 0.42 ns). Furthermore, visual inspection of the structures at 0.80 ns indicates that there is not a significant degree of overlap between the ice-like molecules between the two trajectories.

In order to quantify the statistical independence of the trajectories, we have considered the following quantity for trajectory $m$:

$$N_m = \sum_{i \in \text{trans}} N_{\text{core}}(t_i) \quad (C.1)$$

The notation ‘$i \in \text{trans}$’ indicates that we are only summing over configurations for which $10 < N_{\text{core}} < 500$. The decay of the correlation function $\langle N_0, N_m \rangle$ will then give an indication of how correlated the trajectories are. The correlation ‘time’ from this analysis is $\tau \approx 75$ trajectories, indicating that every 100th trajectory should be
C.5. Transition Path Sampling for the mW Model

Figure C.6: Indicative snapshots from trajectories that differ by 100 TPS moves.

close to statistically independent. In the main text we introduced the coarse grained

Figure C.7: Verification of the coarse-grained representation of ice-like space. In grey we
show the coarse-grained field for $\mathcal{Q}(r) > 0$, while blue spheres and bonds correspond to
ice-like molecules characterized with the $lq_6$.

fields $\mathcal{I}$ and $\mathcal{Q}$ to analyse the trajectories harvested from TPS. We note that the
framework we have used to compute $\mathcal{I}$ and $\mathcal{Q}$ bears a strong resemblance to that of
Willard and Chandler [325], although the determination of interfaces is simplified in the current context. To demonstrate the method, Fig. C.7 shows the ‘blue trajectory’ from Fig. C.6 at 0.42 ns. The transparent iso-surface is representing regions of space where \( \mathcal{Q}(\mathbf{r}) > 0 \). We can see that this coarse graining procedure does a very good job of encompassing the nascent ice nucleus. Due to the fact that ice-like and liquid-like regions contribute with different signs, this procedure also benefits from the fact that only regions of space that are very ice-like appear to give \( \mathcal{Q}(\mathbf{r}) > 0 \): We can see that small clusters away from the nucleus do not show up in this coarse grained representation.
Appendix D

Supplementary Material for Chapter 6

D.1 Overview of Substrates Studied

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
<th>Interaction</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>LJ_fcc111</td>
<td>Fcc crystal with (111) surface exposed</td>
<td>LJ</td>
<td>208</td>
</tr>
<tr>
<td>LJ_fcc100</td>
<td>Fcc crystal with (100) surface exposed</td>
<td>LJ</td>
<td>208</td>
</tr>
<tr>
<td>LJ_fcc110</td>
<td>Fcc crystal with (110) surface exposed</td>
<td>LJ</td>
<td>208</td>
</tr>
<tr>
<td>LJ_fcc211</td>
<td>Fcc crystal with (211) surface exposed</td>
<td>LJ</td>
<td>208</td>
</tr>
<tr>
<td>OH_HEX</td>
<td>Back wall + hexagonal OH patterns</td>
<td>LJ + mW</td>
<td>257</td>
</tr>
<tr>
<td>OH_SQU</td>
<td>Back wall + square OH patterns</td>
<td>LJ + mW</td>
<td>257</td>
</tr>
<tr>
<td>OH_TRI</td>
<td>Back wall + triangular OH patterns</td>
<td>LJ + mW</td>
<td>257</td>
</tr>
<tr>
<td>OH_SQU-OCT</td>
<td>Back wall + OH squares and octagons</td>
<td>LJ + mW</td>
<td>257</td>
</tr>
<tr>
<td>OH_TRI-HEX</td>
<td>Back wall + OH triangles and hexagons</td>
<td>LJ + mW</td>
<td>257</td>
</tr>
<tr>
<td>OH_TRI-SQU</td>
<td>Back wall + OH triangles and squares</td>
<td>LJ + mW</td>
<td>257</td>
</tr>
<tr>
<td>OH_SQ-HE-OCT</td>
<td>Back wall + OH squares, hexagons and octagons</td>
<td>LJ + mW</td>
<td>257</td>
</tr>
<tr>
<td>OH_TR-SQ-HE</td>
<td>Back wall + triangles, squares and hexagons</td>
<td>LJ + mW</td>
<td>257</td>
</tr>
<tr>
<td>GRA</td>
<td>Graphene structure of varied interaction strength</td>
<td>LJ</td>
<td>112</td>
</tr>
<tr>
<td>GRAO</td>
<td>Graphene structure + OH patterns</td>
<td>LJ + mW</td>
<td>108</td>
</tr>
<tr>
<td>OH_REC</td>
<td>Back wall + rectangular OH patterns</td>
<td>LJ + mW</td>
<td>-</td>
</tr>
<tr>
<td>OH_SHEX</td>
<td>Back wall + stretched hexagonal OH patterns</td>
<td>LJ + mW</td>
<td>-</td>
</tr>
<tr>
<td>OH_SREC</td>
<td>Back wall + stretched rectangular OH patterns</td>
<td>LJ + mW</td>
<td>-</td>
</tr>
<tr>
<td>OH_ANGE</td>
<td>Name of the author as OH pattern</td>
<td>LJ + mW</td>
<td>-</td>
</tr>
<tr>
<td>OH_MART</td>
<td>Name of the author as OH pattern</td>
<td>LJ + mW</td>
<td>-</td>
</tr>
<tr>
<td>OH_PHIL</td>
<td>Name of the author as OH pattern</td>
<td>LJ + mW</td>
<td>-</td>
</tr>
</tbody>
</table>

Table D.1: Acronyms of the systems as displayed in Fig. 6.1 explained. The interaction of water with the LJ substrates, the back wall and the carbon atoms in graphene was modelled with a Lennard-Jones interaction. The interaction of OH groups with water was treated as mW-mW [29] interaction.
D.2 Initial Features

We give a brief overview of the initial features we consider and how they were computed. Fig. D.1 shows an illustration of the different feature classes and how the corresponding acronym is formed. Each graph starting from a blue box can be considered a new class of features which we then pre-process to obtain statistical measures (\textit{stat} in red in Fig. D.1) for the corresponding quantities while also distinguishing different layers perpendicular to the surface (\textit{layers} in green in Fig. D.1, see also the inset on the top right) for some of them.

• \textit{dyn}:

Starting point is a simulation of water interfacing with the substrate. We run two sets, one which is 100 ns long where we save every 1 ps (for Steinhardt
\[ q_i \] [323], local Steinhardt \( lq_i \) [144] and number of nearest neighbors \( nn \) and one which is 100 ps where we save every 10 fs (for forces and velocities).

- **\( \text{disp} \):**
  Displacements in either dimension \((x, y, z, \text{lateral } xy \text{ and total } r)\) after 1, 2, 5, 10, 20, 35, 50, 75, 100 and 150 ps.

- **\( \text{rssA} \):**
  Random structure search approach similar to Ref. 326 where we probe the adsorption energy of hemispherical ice seeds \( (I_h(001), I_h(100), I_h(110), I_c(001) \text{ and } I_c(111)) \) for different sizes (100, 300 and 500 molecules).

- **\( \text{rssB}_{\text{flex}} \):**
  Energies from minimization of \( n \)-mer water clusters positioned in many random positions above the surface.

- **\( \text{rssB}_{\text{rigid}} \):**
  Similar to the flexible approach but keeping the structure of the deposited ice structure rigid. Since energy minimization with rigid bodies is highly not-trivial we performed short MD simulations with rigid constraints [327, 328], slightly pushing the ice-structure downwards while draining out the kinetic energy with a friction term in the equations of motion. In this manner we find that the ice structures have enough room to find a local minima with respect to position and orientation.

- **\( \text{lmatch} \):**
  Generalized lattice match calculated as in Ref. [257]:
  \[
  \zeta = \min_{r_0, \theta} \left( \sqrt{\frac{1}{N_M} \sum_{i=1}^{N_{\infty}} (r_i(r_0, \theta) - r_s)^2} \right) \tag{D.1}
  \]

  Here we place ice layers corresponding to a certain face randomly over the surface and compute the shortest distance to a substrate atom for each ice molecule provided this is shorter than a certain cutoff (yielding \( N_M \) contacts).
We considered 2D projections of the ice lattices as well as their actual 3D structure and different cutoffs for neighbor choices. Ice faces considered were Ih(001), Ih(100), Ih(110) and Ic(001).

- dens:
  Number density of liquid water in different layers obtained from the dyn runs.

Here are four examples of acronyms and their actual meaning:

1. **lmatch2D_Ih001_c2**:
   Generalized lattice match calculated with the second cutoff (c2 = 3.2 Å) of the 2D projected basal face (Ih001)

2. **dyn_nn_all_median**:
   Median number of nearest neighbors in the whole liquid (cutoff 3.4 Å)

3. **rssBr_4mer_Eall_range**:
   Total range of all adsorption energies of the water tetramer obtained with the rigid random structure search approach

4. **dyn_lq3_l12_var**:
   Variance of the lq3 parameter in the water layer 112 (definition in Fig. D.1)
Bibliography


[27] Aleks Reinhardt, Jonathan PK Doye, Eva G Noya, and Carlos Vega. Local


[58] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman JC Berendsen. Numerical integration of the cartesian equations of motion of a system with con-


[96] Thomas F. Whale, Martin Rosillo-Lopez, Benjamin J. Murray, and


[104] Arpa Hudait and Valeria Molinero. Ice crystallization in ultrafine water-salt


[131] H. Bingemer, H. Klein, M. Ebert, W. Haunold, U. Bundke, T. Herrmann, K. Kandler, D. Müller-Ebert, S. Weinbruch, A. Judt, A. Wber, B. Nillius,


[218] Håkon Ikaros T Hauge, Marcel A Verheijen, Sonia Conesa-Boj, Tanja Etzel- storfer, Marc Watzinger, Dominik Kriegner, Ilaria Zardo, Claudia Fasolato,


[313] Yuqing Qiu and Valeria Molinero. Strength of alkane-fluid attraction determines the interfacial orientation of liquid alkanes and their crystalliza-


