THE UNIVERSITY *of* EDINBURGH

A Data-Driven Approach to Predicting Heterogeneous Nucleation in Phase Change Materials

Zixuan Wang

A thesis submitted in accordance with the requirements of the School of Chemistry at the University of Edinburgh for the Degree of Doctor of Philosophy

September 2025

Abstract

There is a current requirement for technologies that store heat for both domestic and industrial applications. Phase-change materials (PCMs) are an important class of substances with strong potential for heat storage. For practical use, storage systems must withstand repeated melt/freeze cycles while maintaining a stable melting-crystallisation point and consistent heat output. Salt hydrates are attractive candidates on account of their high energy densities, but there are issues associated with its strong tendency to subcool well below its normal freezing point. While the nucleation problem can be readily solved by the addition of seed crystals of another material, there are a lot of problems that can be encountered that result in nucleator deactivation. Therefore, the problem of identifying suitable heterogeneous nucleating crystallites (NUCs) for PCMs under variable temperature conditions remains a challenging task. In this regard, *in silico* screening methods offers a practical solution to both problems. Through a data driven approach, a workflow is generated by learning from existing experimental reports of working PCM/NUC pairs, in the light of searching for other NUC candidates that may offer improved properties over the additives that are currently used. The focus of this research is therefore to demonstrate the feasibility of a data-driven approach to establish a high-throughput NUC prediction model that could be applied to any given liquid/solid PCM.

In Chapter 2, a workflow generation process is described. The workflow is based on a data-driven approach, and a high-throughput workflow is created based on geometric matching under five related features that returns a binary decision of working/non-working NUC for a given PCM.

In Chapter 3, the trained model is applied with a most extensively studied PCM, ice-I_h. The model is firstly utilised to evaluate the degree of nucleation effectiveness then compared with already existing experimental reports. Bulk water immersion experiments on a set of ten known nucleators sets a delineating temperature to distinguish between good and poor nucleation behaviour. The algorithm is then used to screen 3,500 simple metal oxides and halides taken from the Inorganic Chemistry

Structural Database (ICSD), and show that just 7% of the former and 3% of the latter were predicted to nucleate ice on the basis of geometric slab matching. Subsequent experimental testing of 22 compounds suggests a 64% correct prediction rate, and identifies four new ice nucleators. Inspired by the ice-nucleating efficiency of copper oxides, the copper tubing with local tap water is also tested, and subcooling suppression is observed, most likely due to copper oxide buildup.

In Chapter 4, the model is further trained and tested with working/non-working nucleators from readily existing reliable experimental reports and then the trained model is applied in a high-throughput application for sodium acetate trihydrate (SAT), where over 14,000 candidate NUC structures are screened, from which a list of 521 compounds is identified as potential NUCs for SAT. The result reinforces the success of the current industry-standard NUC for SAT, disodium hydrogen-phosphate hydrates (DSP), which is shown to geometrically match slabs of SAT regardless of the level of hydration present. Other PCMs are sought after, i.e. $Mg(NO_3)_2 \cdot 6H_2O$, $MgCl_2 \cdot 6H_2O$, $CaCl_2 \cdot 6H_2O$, and $LiNO_3 \cdot 3H_2O$. The distribution of prediction from working to non-working NUCs for the four PCMs demonstrates mostly the same trend as the confidence range. This meant this model could be readily used for nucleator mining for other PCM materials.

In Chapter 5, a supervised machine learning workflow is set up with the goal of predicting effective nucleators for any PCM material based on geometric compatibility between their crystallographic slabs. The algorithm is established by learning from highly granular geometric data generated from ice nucleation in Chapter 3 and salt hydrates in Chapter 4, and this approach avoids manually tuning thresholds and instead lets the model discover which geometric criteria (and value ranges) are statistically associated with successful nucleation. The results show prominent prediction power, i.e. success rate on both ice and salt hydrates, and further data analysis showed equal contributions as well as independence of the five features, proving the comprehensiveness of the algorithm.

The impact of this research as well as future works are discussed in Chapter 6.

Lay summary

Heating and cooling account for a large part of energy use in homes and industries. One promising way to store and reuse heat is through special substances called phase change materials (PCMs), which absorb or release heat when they melt or freeze. However, many PCMs, especially a type called salt hydrates, tend to remain liquid even when cooled below their freezing point, which limits their usefulness. To fix this, scientists add additives called nucleators, which help trigger freezing. But finding additives that nucleate specific PCMs is challenging, especially because some nucleators stop working over time or under certain conditions, as well as there is an urgent need to discover new nucleators to expand operating temperatures of PCMs.

This research developed a new computer-based method to help find better nucleators more quickly. Using a combination of geometry-based modelling and machine learning, a predictive tool was created that can scan thousands of materials and suggest which ones are likely to work as nucleators for different PCMs. The method first focused on predicting nucleators for water (ice), and experiments showed that the model was accurate in identifying both known and new ice-nucleating materials. Then the method was expanded to work with salt hydrates, especially sodium acetate trihydrate (SAT), a common material used in heat storage packs. The results suggested several promising new nucleators, and confirmed why some industry-standard ones work so well.

Overall, this research offers a powerful new way to screen materials for heat storage technologies, saving time, cost, and trial-and-error effort in the lab. It opens the door for designing more efficient and reliable thermal energy systems for everyday use.

Acknowledgements

I would like to express my deepest gratitude to my supervisors, Professor Carole Morrison and Professor Colin Pulham, for showing me what scientific research truly is at its core: how to ask the right questions and how to pursue the answers. Their encouragement carried me through moments of imposter syndrome, and their unwavering support during my cancer treatment meant more than I can ever put into words.

I am grateful to Professor Caroline Kirk and Dr. Rebecca Rae for their support with XRD usage, and special thanks to Dr. Hannah Logan for her generous help with the Polar Bear equipment and for the many enlightening discussions we shared. I would also like to thank my colleagues at Sunamp Ltd, Dr. Andrew Bissell, Dr. David E. Oliver, Dr. Emily Goddard, and Dr. Gylen Odling, for their guidance and support.

A very special thank you goes to my partner, Dr. Yong Li. Without him, I would not have been able to complete this PhD. He has stood by me not only as my family but also as an equally sharp mind who challenged and enlightened me as a researcher, and this achievement is as much his as it is mine. I also want to thank my cats Frodo, Gigi (may you frolic happily in cat heaven) and my dog Dobby, whose unconditional love carried me through the ups and downs of this journey, and I love them to my very bones. To my parents Jianhong Wang and Hongxia Guo, I owe endless gratitude for giving everything they could and raising me into who I am today, and this thesis is a testament to that.

Finally, I want to acknowledge myself. Completing a PhD while undergoing cancer treatment is the hardest thing I have ever done, and I am profoundly proud that I never gave up. This thesis is not only a scientific achievement but also a mark of survival, resilience, and determination.

I find strength in the words of Marcus Aurelius: “You have power over your mind, not outside events. Realise this, and you will find strength.” These words constantly remind me that with focus and resolve, one can endure and succeed anything.

Contents

Chapter 1
Introduction

1.1. Challenges and opportunities for thermal energy storage

Thermal energy storage (TES) stands at the heart of the global transition to sustainable energy systems[1,2]. As societies shift toward intermittent renewable sources such as solar[3] and wind[4], the need to store excess energy and deliver it when needed becomes increasingly urgent[5,6]. TES systems offer an elegant solution by storing thermal energy during surplus generation and releasing it on demand[7-10], improving both energy efficiency[11, 12], reliability[13, 14] and sustainability[6,15,16].

Among various TES strategies, phase-change materials (PCMs) are particularly promising[17]. These materials store and release large quantities of latent heat during solid-liquid transitions, enabling compact, efficient thermal storage at near-constant temperatures[18-21]. Unlike sensible heat storage materials, which rely on temperature changes to store energy[22-24], PCMs can store latent heat without significant changes in temperature, making them highly efficient and stable[25,26]. From domestic heating[27,28] to industrial waste heat recovery[29,30], PCMs have the potential to revolutionise how energy is stored, transported, and reused[31-35].

However, the widespread adoption of PCMs is constrained by a practical challenge, namely reliable and repeatable crystallisation[36-38]. Many PCMs suffer from subcooling and sluggish nucleation kinetics[39-41]. Subcooling is a consequence of the failure to form crystallisation sites in the liquid state, and is a common phenomenon displayed by many salt hydrates. Whilst subcooling can be overcome in devices such as disc-activated hand-warmers, where crystallites of the PCM are stored in and released from the rough grooves of the disc, this is a major disadvantage for other, larger scale applications which require reliable and reproducible freezing of the PCM. There is therefore a need for external triggers, i.e. nucleating agents (NUCs), to initiate the crystallisation process[40,42,43]. Without effective NUCs, crystallisation at the material's melting point becomes unpredictable, meaning latent heat energy cannot be predictably released, and thereby undermining the entire purpose of TES device[44,45].

This weakness becomes even more pronounced when extending the operating range of PCMs to higher-temperature systems, where salt hydrates offer immense promise due to their higher sensible heat storage potential compared with organic PCMs such as paraffins.[46,47] At high temperatures the NUC may fail to operate, potentially due to undergoing a phase transition, or dehydration if it too is a hydrate material. Thus, if the full potential of TES systems are to be realised, there is a need to source NUCs that operate under high temperature conditions.

Herein lies the problem: currently, the search for suitable NUCs often relies on trial-and-error screening, i.e. what may be called the "Edison method," guided more by laboratory shelf availability than fundamental understanding. This *ad hoc* approach is time-consuming, inefficient, and unsustainable at scale. To fully unlock the potential of PCMs, a paradigm shift in how to identify and design nucleating agents is urgently needed, to unlock the ability to quickly screen compounds for NUC ability for a given PCM. This is what this thesis aims to address.

This research focuses on two representative categories of inorganic PCMs, namely ice and salt hydrates, to develop and validate a data-driven, generalisable framework for predicting effective heterogeneous nucleators for phase change materials, which results in the potential to screen large structural databases, such as the Inorganic Chemistry Structural Database (ICSD release 2025.1)[48], for potential NUCs for a given PCM.

Ice has been used extensively for millennia as a cold store for food products, but it is only recently that its use in temperature control for buildings has become widespread[49]. It provides a tractable, well-studied system with a wealth of both experimental and computational data that are readily available. Water in many ways is an ideal PCM: it is a one-component system that is cheap, non-toxic, has well-characterised properties, a high latent heat of fusion (334 kJ/kg)[50], and is earth-abundant. Much of the existing literature on heterogeneous nucleation has used ice as a model system, making it a natural choice for method validation[51-57].

Salt hydrates are attractive candidates for mid-temperature TES due to their high energy densities[58,59] and low material costs. Yet, predicting suitable nucleators for

these materials remains particularly challenging due to limited in-situ experimental insight and the complexity of their ionic, hydrated structures[60]. This can be exemplified by sodium acetate trihydrate (SAT), which has a high energy density (values in the range 226-260 kJ kg^{-1} have been reported), a melting point of 58°C (ideal for delivery of domestic hot water), is non-toxic, and has a very low flammability[61-63]. However, SAT suffers from two significant disadvantages that have until now restricted its widespread use as a reliable PCM. These are (i) incongruent melting, whereby instead of melting to form a homogenous solution (melt) of the trihydrate, phase segregation occurs and anhydrous sodium acetate precipitates from solution[63], and (ii) its strong tendency to subcool well below its normal freezing point.

The decision to focus on ice and salt hydrates in this thesis stems from their complementary characteristics: while ice provides a tractable system for methodological development and validation, salt hydrates pose more complex challenges that reflect the diversity of industrial PCM applications. By targeting these inorganic PCMs, this research remains grounded in systems with practical relevance while pushing toward generalisable insights into structure-nucleation relationships.

1.2. Significance of research

To make PCMs viable at scale, especially for mid-temperature storage where the energy density payoff is highest, crystallisation must be controlled. Yet currently, the identification of NUCs is essentially trial-and-error. Materials are selected based on availability, intuition, or historical precedent rather than physical compatibility with the PCM structure. This results in wasted time, resources, and inconsistent performance, even in industrial systems. As TES is integrated into more demanding applications, the reliability of PCMs must match that of conventional technologies.

This research proposes a groundbreaking shift from empirical guesswork to structure-based prediction. By identifying geometric and crystallographic features that underpin effective NUC/PCM interactions, and embedding them into a data-driven workflow, this work aims to build a predictive model for heterogeneous nucleation. By focusing on ice and salt hydrates, materials at the intersection of methodological tractability and industrial relevance, this research not only offers immediate practical insights, but also

lays the groundwork for a generalised NUC prediction framework. This would drastically reduce development time, lower costs, and increase reliability for TES applications. The remainder of this chapter begins with an overview of nucleation theory (Section 1.3), followed by a review of experimental and computational studies on heterogeneous nucleation in ice and salt hydrates (Section 1.4). Section 1.5 then discusses the opportunities and limitations of data-driven approaches for predicting nucleating agents. Finally, Section 1.6 outlines the specific aims and research objectives of this thesis.

1.3. Theoretical framework

The need for nucleators in PCM systems arises from the thermodynamic nature of phase transitions. Without a nucleator, most PCMs experience excessive subcooling tendencies[64,65], which delay crystallisation and make heat release unpredictable[66,67]. By introducing heterogeneous surfaces that lower the nucleation energy barrier, NUCs provide a reliable trigger for the liquid/solid phase change, thereby stabilising PCM performance over repeated thermal cycles.

Nucleation is the initial step in processes such as vapour condensation, crystal formation, melting, and boiling[68]. It involves the localised emergence of a distinct thermodynamic phase at the nanoscale, which expands as growth units attach[69,70]. These phase transitions result from atomistic events driven by thermal fluctuations[71]. Since these events occur on length scales around 10^{-10} m and time scales near 10^{-13} s,[72] *i.e.* comparable to atomic vibrational frequencies, nucleation present a complex phenomenon to study. Despite extensive research on its fundamental aspects, the understanding of nucleation is still incomplete.

Subsequently, Classical Nucleation Theory (CNT)[73] first developed 150 years ago, still offers a foundation for understanding these processes, but it simplifies key aspects such as surface interactions and structural matching. Extensions toward heterogeneous nucleation show that the choice of nucleator surface, including its crystallographic orientation, chemical composition, and structural compatibility, can drastically influence nucleation rates and crystallisation onset temperatures, as outlined in the following sub-section.

1.3.1. Classical nucleation theory (CNT)

Stemming from the work of Volmer and Weber[74], Becker[75], and Frenkel[76], CNT describes the process in which a metastable system gradually transforms into a more stable phase through the formation of nuclei. In the steady-state limit, nucleation occurs at a characteristic rate, often expressed using an Arrhenius-type equation:

$$\mathcal{J} = \omega \exp\left[-\frac{\Delta F^*}{k_B T}\right] \quad (1)$$

where $\mathcal{J}$ is the nucleation rate, ω is a kinetic factor accounting for the rate at which individual particles are incorporated into a nucleus[77], and ΔF^* represents the nucleation free energy barrier that must be overcome for the process to proceed. For homogeneous nucleation, ΔF^* arises from the competition between two opposing contributions, namely:

(i) A bulk (volume) term, which favors nucleation and is proportional to the volume V of the nucleus with radius R, i.e.:

$$V = 4/3\pi R^3 \quad (2)$$

(ii) A surface term, which acts as an energy penalty due to the formation of an interface between the nucleus and the surrounding phase. This term is proportional to the surface area A of the nucleus, given by:

$$A = 4\pi R^2 \quad (3)$$

The formation free energy $\Delta F(R)$ of nucleation can then be written as

$$\Delta F(R) = \Delta pV + \gamma A = -\Delta\mu(\rho_c - \rho_l)\left(\frac{4\pi R^3}{3}\right) + \gamma 4\pi R^2 \quad (4)$$

where γ is the interfacial tension between the nucleus and the surrounding liquid phase. In the standard capillarity approximation, the surface energy γ is assumed to be equivalent to the surface tension of an infinite planar interface between the two coexisting bulk phases—namely, the crystalline phase (with density ρ_c) and the liquid phase (with density ρ_l). At equilibrium, these phases must satisfy the condition of zero chemical potential difference, $\Delta\mu = 0$. The bulk term ΔpV represents the thermodynamic potential gain of the stable phase relative to the metastable phase.

To determine the critical nucleus size, it is possible to define the stationary point of the nucleation free energy by solving:

$$\frac{\partial(\Delta F(R))}{\partial R} = 0 \tag{5}$$

This minimisation yields the critical radius R^* and the free energy barrier ΔF^*_{hom} that must be overcome for homogeneous nucleation to occur, such that:

$$R^* = \frac{2\gamma}{\Delta\mu(\rho_c - \rho_l)}, \Delta F^*_{hom} = \frac{4\pi}{3} R^{*^2}\gamma \tag{6}$$

Substituting the critical nucleation barrier ΔF^*_{hom} at a given supersaturation in Equation 1 gives the final expression for the nucleation rate $\mathcal{J}$, as

$$\mathcal{J} = \omega\, exp[-\frac{16\pi\gamma^3 v^2}{3K_B^3 T^3 (lnS)^2}] \tag{7}$$

Where v is the molecular volume of the solid phase and S is the supersaturation ratio. Finally, the number of molecules in the critical nucleus is approximately:

$$n^* = \frac{4/3\pi R^{*3}}{v} \tag{8}$$

The formation free energy of the cluster, its size, and the rate are the key parameters in CNT research. The size of the critical nucleus has been approximated by various authors using experiments and simulations studies. It has been found that the number of molecules constituting critical nuclei usually falls in the range of tens to thousands[78]. For salt hydrates, the number could be even larger; for example in a study by Adamski, 10−15 g of barium salts form critical nuclei which comprise around a million species[79].

Several key conclusions can be drawn from the preceding analysis. First, nucleation requires overcoming an energy barrier before initiation. Second, larger clusters have a higher probability of serving as nucleation centers due to their increased stability. Finally, a reduction in interfacial energy enhances the likelihood of nucleation. This reduction can be facilitated by interactions with external factors such as heterogeneous molecules, particles, or surfaces, leading to heterogeneous nucleation.

1.3.2. Extension of CNT towards heterogeneous nucleation

Homogeneous nucleation refers to the spontaneous formation of a nanoscopic domain of a new phase within a metastable system due to statistical fluctuations[80-82]. For the

newly formed phase to be stable under the given thermodynamic conditions, it must overcome a significant free energy barrier. In many cases, this is on the order of 50 k_BT (where k_B is Boltzmann's constant and T is the absolute temperature), even when the nucleus consists of only around 100 particles. Given the rarity of such events, heterogeneous nucleation, where pre-existing heterogeneities in the system lower the free energy barrier, is often the more experimentally relevant process.

Consider the extension of Equation 6 to heterogeneous nucleation at a planar wall, assuming partial wetting of the crystal at liquid-solid coexistence, so that $\Delta\gamma = \gamma_{wl} - \gamma_{ws} < \gamma$, where γ is solid-liquid interfacial tension, γ_{wl} is wall-liquid interfacial tension and γ_{ws} is wall-solid interfacial tension. Here, the wall corresponds to the surface of a nucleating agent upon which the PCM nucleates, the solid is the initial PCM nucleus, and the liquid is the PCM melt. Figure 1.1 demonstrates the resulting contact angle and interfacial tensions.

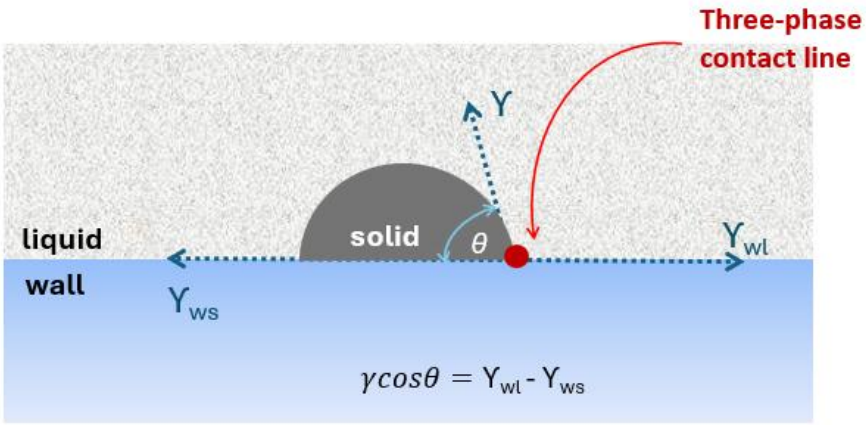


Figure 1.1. Schematic representation of the contact angle (θ) formed at the interface, illustrating the balance of interfacial tensions described by Young's equation.

If the dependence of γ on the orientation of the interface (relative to the crystal axes) is neglected, it is possible to use Young's equation for the contact angle θ,[83] such that:

$$\gamma cos\theta = \Delta\gamma \qquad (9)$$

The shape of the critical nucleus can be approximated as a spherical cap, resting on the wall surface.[84] Equation 4 still applies, but the barrier ΔF^*_{het} is reduced by a factor $f(\theta)$, such that:

$$\Delta F^*_{het} = \Delta F^*_{hom} f(\theta), \qquad f(\theta) = (1 - cos\theta)^2(2 + cos\theta)/4 \tag{11}$$

Where ΔF^*_{hom} is the free energy barrier that must be overcome for homogeneous nucleation, while ΔF^*_{het} is the free energy barrier for heterogeneous nucleation. An illustration of CNT energy profiles for homogeneous and heterogeneous nucleation is seen in Figure 1.2. In the homogeneous case (blue), the system must overcome a large free energy barrier (ΔF^*_{hom}) to reach a critical nucleus size ΔR^*_{hom}. Heterogeneous nucleation (green) lowers the barrier (ΔF^*_{het}) due to favourable interactions with the nucleating agent, as described above.

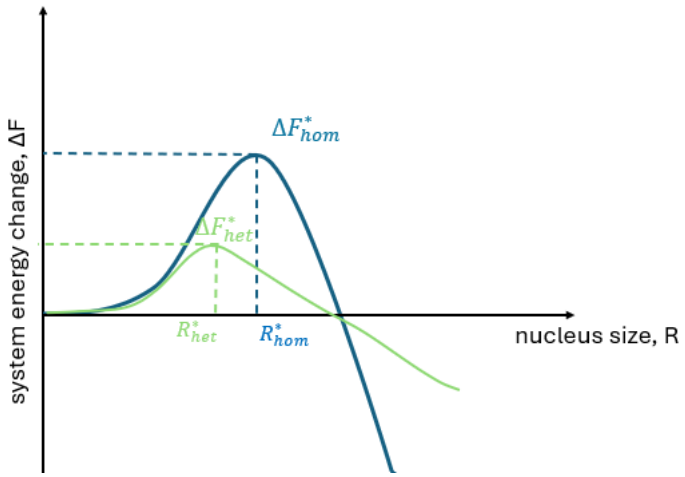


Figure 1.2. CNT energy profiles for homogeneous and heterogeneous nucleation.

Consequently, the presence of a three-phase contact line (Figure 1.1) introduces additional corrections that must be considered[85], factors that have also been highlighted from simulations of wall-attached droplets in lattice gas models[86]. Although new theories such as diffuse interface theory[87] attempts to improve the droplet model by taking into account the interface between solid-liquid, liquid-vapour, it still shows limited success, which is likely due to the omission of the evolutionary stages leading to the critical nucleus not being properly accounted for. Given these

complexities, it remains a non-trivial question whether Equation 11 remains valid in the context of heterogeneous nucleation.

In conclusion, CNT, expressed for both homogeneous and heterogeneous scenarios, relies on several simplifying assumptions:

(i) the nucleus possesses the same macroscopic properties, such as density, structure, and composition, as the stable phase;
(ii) the nucleus is spherical, with a sharply defined interface separating it from the surrounding solution; and
(iii) the vapour-liquid interface is approximated as planar, regardless of the critical cluster size.

Despite numerous extensions and refinements[88-91], CNT remains an important foundational model for describing nucleation, most likely due to the fact that it is based on experimentally accessible parameters.

1.3.3. Two-Step Nucleation Theory

Many excellent studies and reviews have already covered this topic in detail[92,93], so here the focus is on the core concepts only. In the original formulation of CNT, the system must overcome a single free energy barrier associated with forming a crystalline nucleus of a critical size (see Figure 1.3). However, the situation can be quite different for crystal nucleation from solutions. In a supersaturated solution, a significant fluctuation in solute concentration may be necessary just to bring enough solute molecules together to form a certain number of connected entities to form a cluster. It is counterintuitive to assume that these species would simultaneously organise into a crystalline arrangement on the same time scale. In fact, crystal formation in solutions often proceeds via a two-step nucleation mechanism that is absent from the original classical theory. In the typical scenario shown in Figure 1.3, the system must first overcome a precursor free energy barrier $\Delta F^*_{n_{p,two-step}}$ through a density fluctuation that creates a larger number of connected clusters of size $n_{p,two\text{-}step}$. This cluster initially lacks any crystalline order and may be either unstable or metastable relative to the surrounding supersaturated solution. Next, the system must

surmount a second free energy barrier, $\Delta F^*_{n_{two-step}}$, at which stage the number of connected clusters reaches $n_{two\text{-}step}$ to rearrange the molecules within the dense cluster into a crystalline structure.

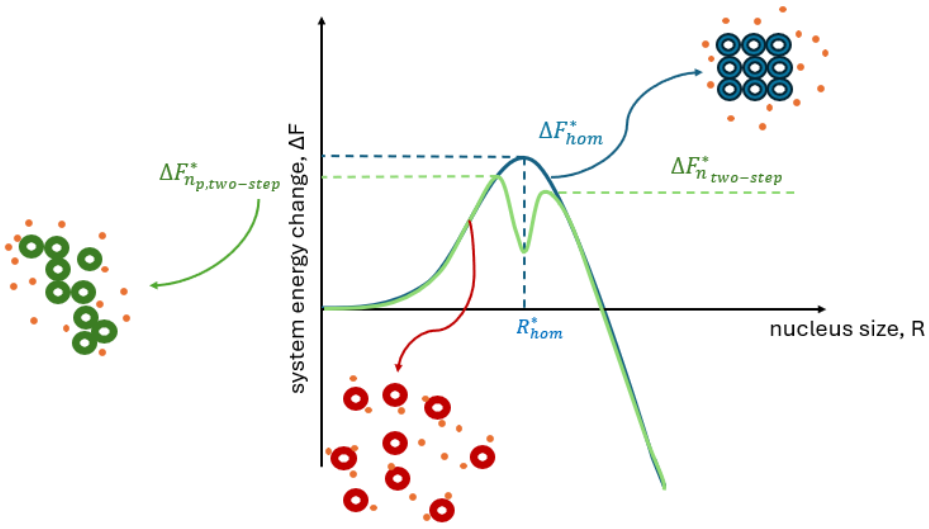


Figure 1.3. Schematic comparison of classical (blue curve) and two-step (green curve) nucleation mechanisms based on free energy profiles as a function of nucleus size. In the classical model, the system crosses a single energy ΔF^*_{hom} at the critical radius ΔR^*_{hom}, leading to direct crystal formation. In contrast, the two-step mechanism involves an initial fluctuation to a metastable intermediate, followed by a secondary transition to the crystalline phase, lowering the overall nucleation barrier $\Delta F^*_{two-step}$.

While some studies suggest a two-step nucleation process arises in certain systems, the structural dependence of nucleation in PCMs, such as sodium acetate trihydrate (see **Chapter 4**), suggests the existence of a strong correlation between the density fluctuations and order formation, which suggests a closer alignment with CNT. Even in models where density fluctuations precede order formation, the emergence of crystallinity is not random. The alignment and matching of structural motifs remain essential for the transition from a disordered phase to a well-definedcrystalline structure, which supports the approach taken in this thesis which is based on NUC/PCM slab matching theory (**Chapter 2**). Due to the limited in-situ nucleation studies on sodium acetate trihydrate, a definitive nucleation pathway has yet to be

established. However, structural considerations suggest that slab matching provides a compelling framework for understanding heterogeneous nucleation.

The slab matching model is strongly grounded in CNT, which assumes that density fluctuations and structural ordering occur simultaneously as a nucleus forms. This implies that nuclei already possess molecular or ionic arrangements that resemble their final crystal structure, making pre-formed slab-like motifs a feasible structural feature of early-stage nucleation. In contrast, the two-step nucleation model suggests that density fluctuations precede order formation, meaning that initial clusters are liquid-like and only develop crystallinity in a secondary step. If nucleation in ice and in salt hydrates followed a purely two-step pathway, slab matching would be a mechanism for crystal growth rather than nucleation formation. Therefore, since the prediction model in this thesis assumes that structural slabs form and integrate early in the nucleation process, it aligns naturally with CNT, where molecular/ionic packing within the nucleus reflects crystalline structure selection from the start.

Despite the emergence of alternative theories, CNT remains widely used due to its simplicity and its ability to qualitatively capture nucleation behaviour across diverse systems. For PCMs, this foundational framework clarifies the rationale behind introducing NUCs, that while deep subcooling can in theory promote homogeneous nucleation by reducing the free energy barrier, it is thermally inefficient and difficult to control. In contrast, heterogeneous nucleation provides a repeatable, interface-guided route with lower energy thresholds and predictable onset temperatures. Within the slab matching framework adopted in this thesis, a structurally compatible NUC not only facilitates nucleation but may also enhance phase purity and long-term cycling stability, particularly in salt hydrates like sodium acetate trihydrate, where repeated melting can otherwise lead to incongruent melting[60].

1.3.4. Fundamental crystallographic concepts relevant to heterogeneous nucleation

The predictive framework developed in this thesis is grounded in crystallographic descriptions of both phase-change materials (PCMs) and heterogeneous nucleators. To ensure clarity and accessibility, this section briefly introduces the core crystallographic

concepts required to interpret the geometric slab-matching methodology employed throughout the thesis. This section is not to provide a comprehensive treatment of crystallography, but to establish a shared vocabulary for describing crystal structure, surface orientation, and lattice compatibility.

Crystalline solids are characterised by the periodic repetition of a unit cell in three-dimensional space[94]. Any crystalline unit cell can be fully described by six lattice parameters: three edge lengths a, b, and c, and three interaxial angles α, β, and γ, where α is the angle between b and c, β between a and c, and γ between a and b[95]. These parameters define the metric geometry of the lattice and determine both the symmetry of the crystal system and the range of possible surface orientations. Based on symmetry and lattice parameters, crystal structures are commonly classified into seven crystal systems: cubic ($a=b=c$, $\alpha=\beta=\gamma=90°$), tetragonal ($a=b\neq c$, $\alpha=\beta=\gamma=90°$), orthorhombic ($a\neq b\neq c$, $\alpha=\beta=\gamma=90°$), hexagonal ($a=b\neq c$, $\alpha=\beta=90°$, $\gamma=120°$), trigonal ($a=b=c$, $\alpha=\beta=\gamma\neq 90°$), monoclinic ($a\neq b\neq c$, $\alpha=\gamma=90°$, $\beta\neq 90°$), and triclinic ($a\neq b\neq c$, $\alpha\neq\gamma\neq\beta\neq 90°$)[96]. These systems describe the constraints on unit cell lengths and interaxial angles and provide a first-order framework for comparing structural symmetry between different materials[97]. Many inorganic PCMs and nucleators considered in this thesis fall within the cubic, hexagonal, or monoclinic systems, each of which presents distinct surface symmetries and cleavage behaviour. Other exceptions of crystal systems are discussed in **Chapter 2**.

More generally, periodic lattices can be categorised into fourteen Bravais lattices[98], which enumerate all unique three-dimensional translational symmetries. The Bravais lattice defines the underlying lattice geometry independent of the atomic basis, and therefore plays a central role in determining how two crystal structures may geometrically align at an interface. In the context of heterogeneous nucleation, compatibility between Bravais lattices influences whether commensurate or near-commensurate interfacial supercells can be constructed.

Crystal surfaces are commonly described using Miller indices (hkl), which specify the orientation of a crystallographic plane relative to the unit cell axes[99]. Miller indices

provide a concise and systematic way to enumerate distinct surface terminations, each of which may exhibit different atomic arrangements, surface symmetries, and interfacial matching behaviour. For hexagonal systems, the four-index Miller-Bravais notation (hkil) is often used to preserve symmetry equivalence within the basal plane. Throughout this thesis, crystallographic slabs are generated by cleaving bulk structures along selected low-index planes, reflecting the fact that such surfaces are more likely to be expressed experimentally and to contribute to nucleation.

When two crystalline surfaces interact during heterogeneous nucleation, the relevant geometric comparison is inherently two-dimensional. Each surface can be represented by an in-plane lattice defined by two basis vectors and an interaxial angle. The degree of lattice registry between a nucleator surface and a PCM surface depends on how closely these two-dimensional lattices can be matched, either directly or through the construction of larger supercells. This concept underpins epitaxial growth theory and motivates the slab-matching descriptors used in later chapters, including lattice vector mismatch, angular deviation, and interfacial area compatibility.

By grounding nucleation prediction in these crystallographic concepts, the framework developed in this thesis explicitly links atomic-scale structure to macroscopic crystallisation behaviour. The crystallographic descriptors introduced here provide the foundation for the geometric screening methodology formalised in Chapter 2 and applied across multiple PCM systems in subsequent chapters. The discussion about skewed slab generation for monoclinic and triclinic crystal systems is in Section 2.3.

1.4. Review on heterogeneous nucleation studies of ice and salt hydrates

Over the years, numerous studies have investigated the mechanisms and factors influencing the heterogeneous nucleation of ice and salt hydrates, employing both experimental techniques and computational simulations. This short review provides an overview of key findings and advancements in this area.

1.4.1. Experimental nucleation studies

Advancements in research methods in recent years have significantly improved the characterisation and observation of heterogeneous nucleation, largely driven by high-resolution solution analytical techniques. Various experimental approaches have been employed to investigate the thermodynamics and kinetics of crystal nucleation in liquids, including salt hydrates. Table 1.1 summarises the key experimental methods used to characterise nucleation, focusing on their spatial resolving power, temporal resolution, and their notable applications.

Table 1.1 Key experimental methods used to characterise heterogeneous nucleation of PCM materials.

Method	Temporal resolution	Spatial resolution	Examples
Confocal scanning microscopy	s	0.3-0.8 μm	CaO on ZrO_2[114], MnS on SiO_2-Al_2O_3-MnO[115]
Atomic force microscopy (AFM)	ms	5-15 nm	brookite TiO_2 on Au[116], calcite on ice[117], Fe(III) oxides on quartz[118], brushite on calcium phosphate[119], $CdCO_3$ on calcite[120], hydroxyapatite ($Ca_{10}(PO_4)_6(OH)_2$) on citrate[121], $BaSO_4$ on barrite[122]
High resolution transmission electron microscopy (HR-TEM)	sub-ps	~0.2 nm	gypsum ($CaSO_4 \cdot 2H_2O$) on bassanite ($CaSO_4 \cdot 0.5H_2O$)[123], $CaCO_3$ on ferrous iron (Fe^{2+})[124]
Cryo-TEM	ms to μs	3-4 Å	$CaCO_3$ on stearic acid[125], calcium phosphate on composite collagen material[126]
Powdered X-ray diffraction (PXRD)	ns to ms	mm to cm	calcite on NX illite[127], calcium phosphates on anatase and rutile[128]
Fourier transform infrared spectroscopy (FTIS)	ns to ms	10-20 μm	$NaNO_3$ on ZnSe[129], $(NH_4)_2SO_4$ on TiO_2, Al_2O_3, or ZrO_2[130], calcium phosphates on anatase and rutile[128]
Optical microscopy	ms	200-500 nm	ice on solid surfaces[105]
Ambient pressure X-ray photoelectron spectroscopy (APXPS)	ms to μs	~10 μm	ice on TiO_2[131]
Differential scanning calorimetry (DSC)	s	>a few cm	$LiNO_3 \cdot 3H_2O$ on $Cu_3(OH)_5(NO_3) \cdot 2H_2O$[57],

			$CaCl_2 \cdot 6H_2O$ on graphene oxide (GO)/$SrCl_2 \cdot 6H_2O$[132]
SEM	ms to s	0.5-4 nm	calcium phosphate on titanium[133], $CaCl_2 \cdot 6H_2O$ on $BaCO_3$[134], calcium phosphates on anatase and rutile[128]
Cloud chamber	ms	a few hundred µm	Ice on AgI[135]
2D solid-state Nuclear Magnetic Resonance (NMR)	ms to s	~10 µm	aragonite on corals[136]
Wide angle X-ray diffraction (WAXS)	ns to ms	1-100 nm	calcite on poly(aspartic acid)[137]
***In situ* liquid-phase TEM**	ms	sub-nm	calcite on Mg^{2+} compounds[138]

As previously discussed, nucleation is a dynamic process that typically occurs on extremely small time and length scales (nanoseconds and nanometres, respectively) posing significant technical challenges in achieving both high spatial and temporal resolution. True microscopic insight remains rare, although in selected cases where nucleation occurs on second-long timescales, dynamical details have been captured. This is the case for salt hydrates, where for example, small-angle X-ray scattering (SAXS) studies on calcium carbonate ($CaCO_3$) nucleation on mica and quartz have demonstrated that hydrophilic NUCs promote nucleation by providing favourable wetting conditions, while hydrophobic surfaces tend to inhibit it[100-102]. Atomic force microscopy (AFM) has enabled the visualisation of molecular-scale interactions during nucleation on a larger (microscopic) scale. For example, AFM studies on the epitaxial growth of brushite ($CaHPO_4 \cdot 2H_2O$) on gypsum ($CaSO_4 \cdot 2H_2O$) revealed two-step growth on specific planes with certain Miller indices, shedding light on the importance of crystallography for nucleator behaviour[103].

However, in most cases, crystal nucleation in liquids occurs within time frames too short for high-resolution instruments to capture sequential snapshots. Under such circumstances, chamber experiments can characterise nucleation as a function of time or temperature, identifying freezing events within a system using techniques such as femtosecond X-ray scattering[51], optical microscopy[104,105], and X-ray powder diffraction[106, 107]. From these data, nucleation rates are reconstructed by measuring induction times[108], which provides connections to theoretical frameworks such as CNT.

Another approach focuses on studying large macroscopic systems, where freezing is detected using techniques such as differential scanning calorimetry (DSC[109]) and Fourier-transform infrared spectroscopy (FT-IR[110]). A notable report used DSC to determine melting and crystallisation temperatures in $LiNO_3 \cdot 3H_2O$, an important TES[111] which will be discussed further in **Chapter 4**. In essence, this work demonstrated a close correlation between the subcooling temperature ΔT with the lattice mismatch between planes of closely packed coordination polyhedra in a number of potential NUCs, including SiO_2, $ZrSiO_4$ and $FeSb_2$. This work went so far as to propose other potential NUCs, such as $Cu_3(OH)_5(NO_3) \cdot 2H_2O$, due to a combination of close lattice matching and low energy chemical bonding interactions.

Cloud chamber experiments, another macroscopic method, have also been employed to determine the frozen fraction of system and nucleation temperatures, especially for ice. Although these methods lack the resolution to capture individual nucleation events, they have provided valuable insights into factors such as solvent effects and the influences of impurities, by analysing the quantity and structure of the crystalline phase. Manglik *et al.*[112] discovered that $Zn(NO_3)_2.6H_2O$ is actually an effective nucleating agent for lithium nitrate hexahydrate. The lattice parameters for the *b*- and *c*-axes varied by 1.43% and 5.40%, respectively, which suggests that the two compounds are closely lattice-matched.

Despite the availability of numerous powerful experimental techniques, and the emergence of new methods such as ultrafast X-ray imaging[113], obtaining microscopic-level insight into nucleation, remains a formidable challenge. In the next section, the impact of machine learning (ML) techniques on this field, which offer a powerful complement to experimental approaches, will be outlined.

1.4.2. Data-driven and Machine Learning approaches

Michaelides *et al*.[148] and Davies *et al*.[149] have laid important groundwork in applying data-driven methods to the prediction of heterogeneous ice nucleation, offering key conceptual parallels to the approach taken in this research. In the initial study,[148] a ML framework was developed to predict the nucleation ability of various crystalline surfaces. Using a random forest classifier trained on features derived from atomistic simulation data, including hydroxyl site geometry, surface patterning, and lattice mismatch, *etc*., the study demonstrated that physically meaningful, low-dimensional features could reliably distinguish between effective and ineffective ice nucleators. This marked an important shift away from purely simulation-heavy or heuristic methods, allowing for efficient, scalable screening of candidate surfaces without the need for costly MD simulations for every new system.

Building upon this, the follow-up paper by Davies *et al*.[149] extended the predictive model to experimental systems. The authors evaluated the nucleation performance of real materials using freezing droplet assays and compared them directly with predictions made by their trained model. Importantly, the study validated that crystal orientation and atomic-scale interfacial structure were just as critical as surface chemistry, reinforcing the need to capture anisotropy and detailed lattice characteristics. An ice nucleating agent prediction model was set up (called "*IcePic*"); the predictive power of which was shown to generalise across a range of material classes, demonstrating the robustness and broader applicability of this approach.

These two papers provide both methodological inspiration and a point of contrast to the present work. Like Davies *et al*.[149], this research uses physically interpretable features (such as angle mismatch and lattice alignment) to predict the performance of nucleating substrates. By contrast, the materials under investigation in this thesis, namely, salt hydrates and other PCMs which differ significantly in both crystal structure and phase behaviour compared to ice. Rather than targeting hydrogen-bonded nucleation under aqueous conditions, this work addresses solid-solid or solid-liquid transitions, where ionic bonding and thermochemical stability are dominant factors. Accordingly, this thesis emphasises geometric compatibility between slab pairs and

interface matching rules, rather than adsorption energetics or site-specific hydrogen bonding.

Furthermore, while the Michaelides-Davies studies[148, 149] rely heavily on molecular simulations for training data, this work takes a high-throughput structural screening approach, aiming to bridge crystallographic features and practical nucleation efficacy in a way that is more computationally tractable. The ability to rapidly scan through large candidate libraries of PCMs and potential nucleators makes this method particularly suitable for industrial applications, such as in TES systems, where identifying generalist nucleators could meaningfully impact energy efficiency.

In data-driven nucleation studies such as those of Michaelides et al.[148] and Davies et al.[149], model performance cannot be judged solely by qualitative agreement with experimental trends. Instead, predictions are evaluated using statistical performance metrics that quantify how reliably a model distinguishes between effective and ineffective nucleators. This is especially important in heterogeneous nucleation, where experimental labels are often sparse, system-dependent, and subject to variability arising from measurement protocols and material preparation.

From a methodological perspective, many data-driven approaches to heterogeneous nucleation can be viewed through the lens of binary classification[176]. In this formulation, candidate nucleators are assigned to one of two categories, working or non-working, based on experimental observation, while computational descriptors are used to infer the likelihood of belonging to either class. This framing reflects the practical reality of nucleation studies, where experimental outcomes are often qualitative or threshold-based rather than continuous, and where the precise microscopic pathway of nucleation remains inaccessible. Treating nucleation prediction as a classification problem allows uncertainty, imbalance, and incomplete labelling to be handled explicitly, and provides a natural framework for evaluating model performance using statistical metrics. This perspective underpins the interpretation of the machine-learning-assisted analyses presented later in this thesis.

Within this framework, accuracy (see Equation 12) provides a measure of the overall proportion of correct predictions[177]. Note that in this thesis, ‘True Positives’ (TP) are

nucleators predicted as working that are also experimentally validated as working. 'False Positives' (FP) are nucleators predicted as working but experimentally are non-working. 'True Negatives' (TN) are nucleators correctly predicted as non-working, while 'False Negatives' (FN) are nucleators predicted as non-working, but which experiments show to be working.

However, accuracy alone can be misleading in datasets where working and non-working nucleators are unevenly represented, a situation frequently encountered in experimental nucleation studies. To address this, complementary metrics such as precision and recall (see Equations 13 & 14) are often used. Precision quantifies the fraction of predicted effective nucleators that are truly effective, and therefore reflects the reliability of positive predictions[178]. This metric is particularly relevant for screening applications, where false positives may lead to unnecessary experimental validation efforts. Recall, by contrast, measures the fraction of truly effective nucleators that are successfully identified by the model, capturing the risk of overlooking promising candidates[179].

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN) \quad \textbf{(12)}$$

$$\text{Recall} = TP/(TP+FN) \quad \textbf{(13)}$$

$$\text{Precision} = TP/(TP+FP) \quad \textbf{(14)}$$

The relative importance of these metrics depends on the scientific or practical objective of the model. In exploratory studies, higher recall may be prioritised to ensure broad coverage of potential nucleators, whereas in resource-constrained industrial contexts, higher precision may be favoured to minimise experimental cost[180]. Consequently, predictive performance is typically assessed using multiple metrics in parallel rather than a single scalar measure. The statistical evaluation metrics provide the necessary framework for interpreting data-driven nucleation predictions. They allow models to be assessed not only in terms of correctness, but also in terms of reliability, selectivity, and practical usefulness. These considerations are central to the interpretation of the machine-learning-assisted analyses presented in the later chapters of this thesis. Further details on the formulation, implementation, and evaluation of this classification framework are provided in **Chapter 5**.

In summary, while operating in distinct material domains, both research streams share a core premise: that structurally derived, interpretable features can yield predictive insight into nucleation phenomena. The progression from data-driven screening to experimentally validated predictions in the Davies *et al*. work provides a valuable precedent as well as a conceptual framework for ongoing efforts to apply similar reasoning in broader crystallisation contexts.

1.4.3. Heterogeneous nucleation of PCM: Key insights

A solid salt hydrate PCM can nucleate from its liquid melt, just as ice does. However, homogeneous nucleation occurs only under highly constrained conditions, making heterogeneous nucleation the dominant pathway in most applications. Despite extensive research efforts (as discussed in Section 1.3), a comprehensive theory and a standardised measurement technique for heterogeneous ice nucleation still do not exist[181]. This gap in understanding is particularly evident in the study of nucleating agents, where:

1. The underlying mechanisms by which different nucleators facilitate heterogeneous nucleation remain unclear;
2. The key properties that define an effective heterogeneous nucleator are not well established;
3. Rational design principles for creating efficient nucleators are still lacking.

A reliable, standard heterogeneous nucleator would be invaluable for establishing a consistent baseline against which all other nucleation processes could be measured and compared. However, efforts to identify such a material have proven challenging. Variability in sample purity, grain size, and surface defects introduces inconsistencies, causing the so-called “baseline” to shift and making standardisation difficult[182]. One source of additional complexity lies in the classification of nucleation modes themselves. Before examining the progress made in identifying influencing factors, it is essential to distinguish between the modes of nucleation, whether immersion, deposition, or contact. This distinction is particularly critical for ice, as each mechanism is governed by different parameters and may lead to different interpretations if not properly classified.

In immersion nucleation, which occurs within subcooled liquid droplets below the surface, the dominant influences are the surface free energy of the NUC and water activity. In contrast, deposition nucleation, where ice forms directly from vapour onto the nucleating agent surface, is highly dependent on temperature and supersaturation levels[183]. This dependence results in different nucleation ability standards between the cloud chamber method and the immersion freezing method (see **Chapter 3**). Finally, contact nucleation refers to the enhanced nucleation of ice at the surface of liquid water when a foreign solid is present at the three-phase contact line (solid-liquid-vapour). This process exhibits a higher nucleation rate than immersion nucleation for the same solid particle[184-187], and is often dictated by surface roughness and defects of the NUC. A schematic representation of the three different nucleation mechanism with respect to T(temperature)-S_i (saturation ratio with respect to ice) is seen below in Figure 1.4.

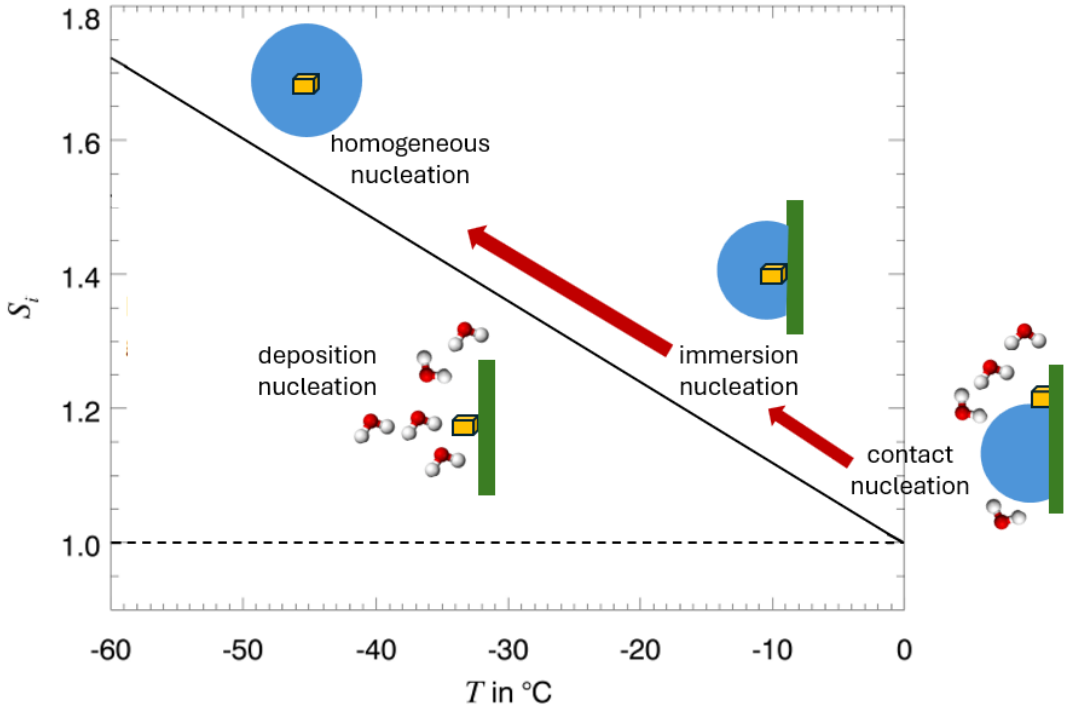


Figure 1.4. Schematic representation of the different modes of ice nucleation plotted as a function of temperature (T) and ice saturation ratio S_i. The temperature versus saturation ratio part of figure is adapted from Hoose *et al*[188].

Above Si = 1 (indicated by the dashed horizontal line in Figure 1.4), ice is the thermodynamically stable phase. Subcooled liquid water is in equilibrium with the

vapour phase along the solid diagonal line, which represents the ice saturation ratio at liquid water saturation.

Ice crystals in the atmosphere have important impacts on various fields, and thus their formation has been studied both in the field and under controlled conditions in laboratory experiments in cloud chambers since many years[56,189] . It is known that water droplets in the atmosphere do not freeze instantaneously at 0 ◦C. Their freezing can either be triggered by NUCs or occur homogeneously at about -38 ◦C[190].The goal of many laboratory studies was and is to assess the ice nucleation ability with presence of a NUC with a specific composition[191, 192].

Contact nucleation, which involves a foreign solid, is a special case of surface nucleation[193]. In both cases, the nucleation rate of ice near a (liquid-water)-(water-vapour) interface is higher than in the bulk liquid phase. Unlike a solid-water interface, surface nucleation and contact nucleation involve a three-phase system where liquid water, solid, and vapour meet. Contact nucleation has been found to be highly significant in heterogeneous ice nucleation. For example, AgI has been shown to be substantially more effective in the contact mode than in the immersion mode[52]. It has also been observed that many heterogeneous nucleators exhibit different nucleation thresholds depending on whether they act in contact or immersion mode[52]. This suggests that the same solid nucleator may behave differently depending on whether it is fully immersed in liquid water or protrudes into the vapour phase.

The reason of contact nucleation is discussed here is, even though the experimental validation method used in this research (see **Chapter 2**) is specifically designed to reduce the (liquid-water)-(water-vapour) interface appearance by applying a layer of silicone oil over the top in each vial, nucleators may not always remain fully immersed in liquid water, and contact nucleation could contribute to apparent discrepancies in nucleation efficiencies. Furthermore, since the three-phase contact line has been found to exhibit significantly enhanced nucleation rates, understanding how and why this occurs provides valuable context for interpreting heterogeneous nucleation processes more broadly. These considerations reinforce the importance of carefully

distinguishing between immersion and contact nucleation modes when analysing experimental results.

These distinctions in nucleation mode, particularly between immersion and contact nucleation, highlight the complexity of interpreting experimental outcomes. This complexity becomes even more pronounced when considering the wide range of physical and structural factors that influence heterogeneous nucleation. In what follows, how properties such as roughness[194], surface morphology[195], and lattice matching[196] impact the nucleation behaviour of salt hydrates and ice are discussed, drawing on both experimental studies and computational models.

Roughness and defects of NUCs

Scanning electron microscopy (SEM) measurements provide insight into nucleation at a micron level, thus shedding light on the dependence of ice nucleation ability on surface topology. Through such experiments it was revealed that edges and cracks at NUC surfaces can play a crucial role in nucleating both salt hydrates and ice, and narrow wedges have been reported as preferred condensation sites[194, 197]. Contact at such a site is effectively one-dimensional and does not require any surmounting of an activation barrier[82]. As such, condensation proceeds without nucleation of supersaturation. After vapour has condensed to such a wedge, however, freezing of the condensed liquid in the narrow wedge may require surmounting a large activation barrier[198], as the angle of the wedge requires distortion of the lattices that costs extra free energy. Instead, nucleation of the solid phase occurs some distance away from the very tip of the narrow wedge and leaves the trapped liquid next to the very tip unfrozen at small subcooling[199]. An illustration of nucleation dynamics in a surface wedge is shown in Figure 1.5. In short, the free energy reduction due to the wetting by the liquid is sufficient to balance out the free energy cost of keeping the small amount of liquid unfrozen below the melting point when the area-to-volume ratio of the pore geometry is sufficiently large.

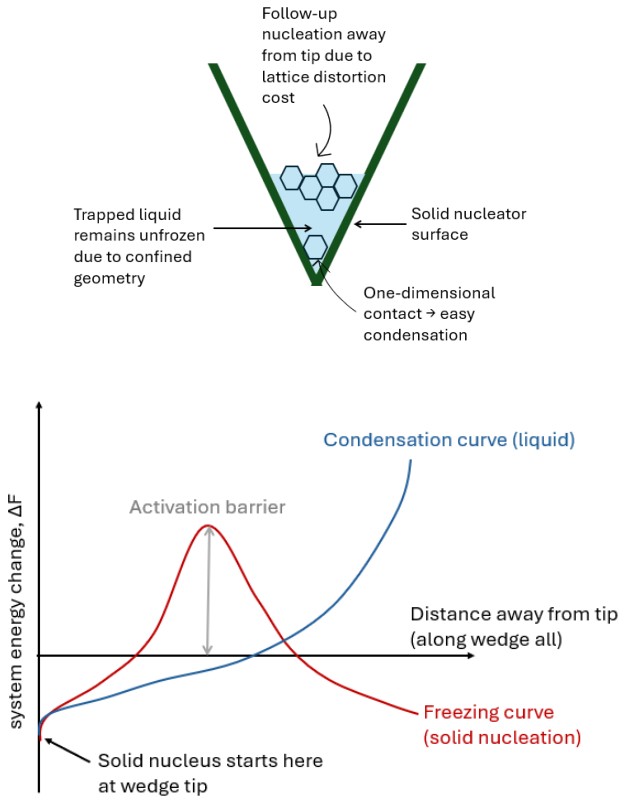


Figure 1.5. Schematic representation of nucleation dynamics within a narrow surface wedge.

Surface morphology of NUCs

Using the coarse-grained mW potential[200], Lupi *et al.*[195] investigated ice nucleation on graphite surfaces, specifically examining how surface hydrophilicity influences nucleation behaviour. They modified the hydrophilicity of the surface in two ways: (i) by uniformly altering the water–surface interaction strength and (ii) by introducing hydrophilic species at the surface. Their findings revealed qualitatively different outcomes.

Cox *et al.*[140] further explored the layering mechanism by investigating ice nucleation rates over a broader range of hydrophilicities. They examined nucleation on two different surfaces: (i) the (111) surface of a face-centred cubic (fcc) crystal, which provides distinct adsorption sites for water molecules, and (ii) a graphite surface, similar to that used in Lupi *et al.*'s study. The study concluded that layering can promote ice nucleation, but only when the surface presents a relatively smooth

potential energy landscape. This highlighted the possibility that heterogeneous nucleation mechanisms can vary significantly across different types of surface morphologies.

Fitzner *et al*.[201] performed brute-force molecular dynamics (MD) simulations of heterogeneous ice nucleation. Their results revealed a complex relationship between surface hydrophobicity and morphology, demonstrating that neither the layering mechanism nor lattice matching alone could fully explain the observed nucleation behaviours. Instead, the study identified three additional microscopic factors that significantly influence heterogeneous ice nucleation: (i) in-plane templating of the first water overlayer on the crystalline surface; (ii) a buckled, ice-like structure in the first overlayer; and (iii) enhanced nucleation beyond the first two overlayers, possibly driven by dynamic effects or extended structural templating from the substrate. Moreover, their findings showed that varying the lattice parameter could induce the nucleation and growth of up to three different ice faces: basal, prismatic, and secondary prismatic (1,1,2,0), on the same surface, further complicating the nucleation landscape.

Lattice matching between NUCs and PCMs

One key finding that challenges the notion of lattice matching as the sole criterion for identifying effective ice nucleating agents comes from research by Sadtchenko *et al*[202]. Silver iodide (AgI) has lattice constants that closely match those of ice, with a deviation of only a few percent. Due to this near-perfect lattice matching, AgI has long been considered an excellent promoter of ice nucleation, making it a prime candidate for cloud seeding. However, despite decades of commercial use, conclusive evidence that AgI significantly enhances rainfall remains elusive. Sadtchenko *et al*. discovered an unexpected phenomenon: rather than forming a uniform epitaxial layer, ice grows exclusively in its hexagonal form on β-AgI. This contradicts the expectation that a lattice-matching substrate should facilitate the epitaxial growth of a uniform crystalline film. Interestingly, this contradiction is not unique to AgI. Barium fluoride (BaF_2)[196], despite its good lattice matching with ice, is also an ineffective ice nucleator. Nutt and co-workers investigated the adsorption structures of water at a model hexagonal surface and at BaF_2 (111) using interaction potentials derived from *ab initio* calculations. Although the surfaces under investigation had structures that matched the

basal face of ice well, they found disordered structures of water to be more favourable than ice-I_h.

1.4.4. Comparative insights and challenges

Both experimental and computational approaches have significantly advanced the understanding of heterogeneous nucleation. While experiments provide empirical evidence and direct observations, computational simulations offer detailed mechanistic insights and the ability to explore parameter spaces inaccessible to experiments. However, challenges remain in bridging the gap between these approaches. For instance, discrepancies often arise due to idealised conditions in simulations compared to the complexities of experimental systems.

Another challenge is the lack of universal models that can accurately predict nucleation behaviour across different salts and substrates. Current efforts focus on integrating experimental data with computational models to develop comprehensive frameworks for heterogeneous nucleation.

Past studies on the heterogeneous nucleation of salt hydrates have revealed intricate dependencies on NUC properties, environmental conditions, and molecular interactions. Experimental advancements have enabled direct observations of nucleation events, while computational simulations have provided atomistic and energetic insights into the process.

In the context of this thesis, the iterative interplay between computational data-driven prediction models and experimental verification serves a dual purpose. On one hand, expanding the database through experimental validation significantly enhances the training pool, thereby improving the accuracy of the predictive model. In turn, a more refined model facilitates the efficient identification of effective NUCs for any given phase-change material PCM in a cost- and time-effective manner. This continuous feedback loop not only streamlines experimental verification but also strengthens the potential for industrial applications by providing a more reliable and scalable approach to nucleation prediction.

1.5. Advantages and challenges of predicting NUCs *in silico*

A persistent challenge in optimising PCM performance is the identification of effective heterogeneous nucleators (NUCs), which promote reliable and repeatable crystallisation. Nucleator discovery has previously relied heavily on empirical trial-and-error approaches, while *in silico* prediction offers an attractive alternative by enabling systematic screening of candidate nucleators prior to experimental testing. Computational approaches can rapidly evaluate large numbers of potential NUCs, reducing experimental workload and narrowing attention to the most promising candidates. This capability is particularly valuable for PCM systems, where experimental nucleation measurements are often costly, difficult to reproduce, or sensitive to uncontrolled variables.

Beyond efficiency, computational methods provide access to mechanistic insights that are difficult to obtain experimentally. By representing crystallographic structure explicitly, *in silico* approaches can interrogate how surface geometry, lattice compatibility, and structural registry influence nucleation behaviour. Such insight supports rational nucleator design, shifting discovery away from empirical heuristics toward physically motivated criteria.

Despite these advantages, *in silico* prediction of heterogeneous nucleation also faces important limitations. Nucleation is inherently stochastic and sensitive to kinetic, chemical, and interfacial effects that may not be fully captured by simplified models. In particular, descriptors that are used in this thesis often rely on idealised structural representations and may neglect factors such as surface chemistry, hydration layers, or dynamic rearrangements at the interface. As a result, computational predictions must be interpreted as probabilistic indicators rather than definitive predictors of nucleation activity.

As computational resources and data-driven methods continue to advance, *in silico* prediction is expected to play an increasingly important role in PCM development. The challenge lies not in replacing experiments, but in constructing predictive frameworks that balance physical interpretability, computational efficiency, and robustness under limited experimental data.

1.6. Aims and research objectives

The overarching aim of this research is to establish a generalisable, data-driven framework for identifying effective heterogeneous NUCs for PCMs, using geometric and crystallographic insights. This aim is pursued through four interconnected chapters, each addressing a different level of understanding and methodological development.

Chapter 2 explains establishment of the data-driven geometric slab-matching framework used throughout the thesis. It formalises the theory of geometric matching grounded in epitaxial growth, developed algorithms for identifying and de-duplicating unique slab terminations, and defined tolerance-based features for high-throughput interface screening. Crucially, it bridges crystallographic features with experimentally validated nucleation outcomes, laying the foundation for the predictive models developed in subsequent chapters.

Chapter 3 establishes ice-I_h as a model PCM system for understanding heterogeneous nucleation. A high-throughput screening workflow is developed that evaluates geometric compatibility between nucleator surfaces and ice slabs cleaved along Miller index planes up to (3,3,3). Experimental validation using immersion freezing confirms the predictive value of geometric matching. This chapter sets the methodological foundation and demonstrates that crystal morphology plays a significant role in nucleation efficiency, with results supporting that a threshold number of well-matched interface models correlates with effective nucleation behaviour.

Building upon the insights from ice nucleation, Chapter 4 extends the geometric slab-matching workflow to salt hydrate PCMs. Unlike ice, the nucleation of salt hydrates remains poorly understood due to the lack of clear experimental data and the complexity of phase transformations. By applying the same interface-matching model, this chapter explores whether geometric indicators used successfully for ice-I_h can generalise to more industrially relevant PCMs. Key challenges addressed include the unknown surface chemistry of salt hydrates and the absence of reliable experimental nucleation metrics. The findings highlight the limitations of relying solely on lattice match without deeper knowledge of surface structure and hydration states, prompting the need for a more adaptable framework.

The final chapter introduces a machine learning framework to bridge the gap between detailed atomistic interface models and limited experimental labelling. While only the overall nucleation effectiveness of each NUC is experimentally known, the geometric compatibility is computed for all possible slab pairings (4,000 per NUC). The model is then trained to identify which geometric features and value ranges are statistically associated with nucleation success. This approach allows the model to infer patterns from highly granular data without needing to manually tune thresholds, offering a scalable method for predictive screening across diverse PCMs.

Together, these chapters form a cohesive strategy that combines atomistic interface modelling, high-throughput screening, and supervised machine learning to understand and predict heterogeneous nucleation. The thesis aims at enabling the rational and transferrable prediction of nucleators for phase changing material nucleation.

References

1. G. Li and X. Zheng, *Renewable and Sustainable Energy Reviews*, 2016, **62**, 736-757.
2. S. Saher, S. Johnston, R. Esther-Kelvin, J. M. Pringle, D. R. MacFarlane and K. Matuszek, *Nature*, 2024, **636**, 622-626.
3. R. Foster, M. Ghassemi and A. Cota, *Solar Energy: Renewable Energy and the Environment*, CRC press, 2009.
4. A. Sayigh, D. Milborrow and J. Kaldellis, *Innovative Renewable Energy, Cham: Springer International Publishing*, 2020.
5. J. Xu, R. Wang and Y. Li, *Solar Energy*, 2014, **103**, 610-638.
6. S. Hasnain, *Energy Conversion and Management*, 1998, **39**, 1127-1138.
7. T. Yang, W. P. King and N. Miljkovic, *Cell Reports Physical Science*, 2021, **2**.
8. B. Zalba, J. M. Marín, L. F. Cabeza and H. Mehling, *Applied Thermal Engineering*, 2003, **23**, 251-283.
9. M. Abdullah, M. Obayedullah and S. A. Musfika, *International Journal of Energy Research*, 2025, **2025**, 6668430.
10. B. Buonomo, M. R. Golia, O. Manca and S. Nardini, *Thermal Science and Engineering Progress*, 2024, **53**, 102732.
11. A. Elkhatat and S. A. Al-Muhtaseb, *Energies*, 2023, **16**, 4471.
12. I. Sarbu and C. Sebarchievici, *Sustainability*, 2018, **10**, 191.
13. M. M. Islam, T. Yu, G. Giannoccaro, Y. Mi, M. La Scala, M. N. Rajabi and J. Wang, *IEEE Access*, 2024.
14. F. Mohamad, J. Teh, C.-M. Lai and L.-R. Chen, *Energies*, 2018, **11**, 2278.
15. S. Kalaiselvam and R. Parameshwaran, *Thermal Energy Storage Technologies for Sustainability: Systems Design, Assessment and Applications*, Elsevier, 2014.
16. R. Parameshwaran, S. Kalaiselvam, S. Harikrishnan and A. Elayaperumal, *Renewable and Sustainable Energy Reviews*, 2012, **16**, 2394-2433.
17. L. F. Cabeza, A. Castell, C. d. Barreneche, A. De Gracia and A. Fernández, *Renewable and Sustainable Energy Reviews*, 2011, **15**, 1675-1695.
18. B. He, V. Martin and F. Setterwall, *Energy*, 2004, **29**, 1785-1804.
19. J. P. Da Cunha and P. Eames, *Applied Energy*, 2016, **177**, 227-238.
20. C. Wyman, J. Castle and F. Kreith, *Solar Energy*, 1980, **24**, 517-540.
21. J. B. Goodenough, *Energy Storage Materials*, 2015, **1**, 158-161.
22. S. Khare, M. Dell'Amico, C. Knight and S. McGarry, *Solar Energy Materials and Solar Cells*, 2013, **115**, 114-122.
23. G. Li, *Renewable and Sustainable Energy Reviews*, 2016, **53**, 897-923.
24. S. Khare, M. Dell'Amico, C. Knight and S. McGarry, *Solar Energy Materials and Solar Cells*, 2012, **107**, 20-27.
25. G. G. D. Han, H. Li and J. C. Grossman, *Nature Communications*, 2017, **8**, 1446.
26. H. Togun, H. S. Sultan, H. I. Mohammed, A. M. Sadeq, N. Biswas, H. A. Hasan, R. Z. Homod, A. H. Abdulkadhim, Z. M. Yaseen and P. Talebizadehsardari, *Journal of Energy Storage*, 2024, **79**, 109840.
27. U. Masood, M. Haggag, A. Hassan and M. Laghari, *Buildings*, 2023, **13**, 1595.
28. T. Yang, W. P. King and N. Miljkovic, *Cell Reports Physical Science*, 2021, **2**, 100540.

29. D. Jayathunga, H. Karunathilake, M. Narayana and S. Witharana, *Renewable and Sustainable Energy Reviews*, 2024, **189**, 113904.
30. S. A. Ali, K. Habib, M. Younas, S. Rahman, L. Das, F. Rubbi, W. U. Mulk and M. Rezakazemi, *Energy & Fuels*, 2024, **38**, 19336-19392.
31. M. Ahangari and M. Maerefat, *Sustainable Cities and Society*, 2019, **44**, 120-129.
32. G. Gholamibozanjani and M. Farid, *Energies*, 2021, **14**, 1929.
33. L. Wang, L. Guo, J. Ren and X. Kong, *Applied Energy*, 2022, **321**, 119345.
34. R. Aridi and A. Yehya, *Energy Conversion and Management: X*, 2022, **15**, 100237.
35. J. Jeon, J.-H. Lee, J. Seo, S.-G. Jeong and S. Kim, *Journal of Thermal Analysis and Calorimetry*, 2013, **111**, 279-288.
36. H. Yang, Y. Zou and H. Cui, *National Science Open*, 2024, **3**, 20230056.
37. M. Alam, S. Devapriya and J. Sanjayan, *Energy and Buildings*, 2022, **268**, 112226.
38. J. Zhang, Z. Cao, S. Huang, X. Huang, Y. Han, C. Wen, J. Honoré Walther and Y. Yang, *Applied Energy*, 2023, **342**, 121158.
39. W. Cui, Y. Yuan, L. Sun, X. Cao and X. Yang, *Renewable Energy*, 2016, **99**, 1029-1037.
40. M. H. Zahir, S. A. Mohamed, R. Saidur and F. A. Al-Sulaiman, *Applied Energy*, 2019, **240**, 793-817.
41. Y. Song, D. Lilley, D. Chalise, S. Kaur and R. S. Prasher, *Cell Reports Physical Science*, 2023, **4**, 101462.
42. A. Safari, R. Saidur, F. A. Sulaiman, Y. Xu and J. Dong, *Renewable and Sustainable Energy Reviews*, 2017, **70**, 905-919.
43. F. Bruno, M. Belusko, M. Liu and N. H. S. Tay, in *Advances in Thermal Energy Storage Systems*, ed. L. F. Cabeza, Woodhead Publishing, 2015, pp. 201-246.
44. P. Mehta, V. Patel, S. Kumar, V. Sharma, G. G. Tejani and A. J. Santhosh, *Scientific Reports*, 2025, **15**, 13876.
45. Z. Ge, Y. Li, D. Li, Z. Sun, Y. Jin, C. Liu, C. Li, G. Leng and Y. Ding, *Particuology*, 2014, **15**, 2-8.
46. K. Gawron and J. Schröder, *International Journal of Energy Research*, 1977, **1**, 351-363.
47. R. Naumann and H.-H. Emons, *Journal of Thermal Analysis*, 1989, **35**, 1009-1031.
48. D. Zagorac, H. Müller, S. Ruehl, J. Zagorac and S. Rehme, *Applied Crystallography*, 2019, **52**, 918-925.
49. Wood Mackenzie, " Discover global power and renewables insights", https://www.greentechmedia.com/articles/read/ice-energy-finds-profits-in-thermal-energy-storage. Accessed 10 August 2025.
50. H. Kumano, T. Asaoka, A. Saito and S. Okawa, *International Journal of Refrigeration*, 2007, **30**, 267-273.
51. H. Laksmono, T. A. McQueen, J. A. Sellberg, N. D. Loh, C. Huang, D. Schlesinger, R. G. Sierra, C. Y. Hampton, D. Nordlund and M. Beye, *The Journal of Pphysical Chemistry Letters*, 2015, **6**, 2826-2832.
52. C. Marcolli, B. Nagare, A. Welti and U. Lohmann, *Atmospheric Chemistry and Physics*, 2016, **16**, 8915-8937.

53. B. Murray, D. O'sullivan, J. Atkinson and M. Webb, *Chemical Society Reviews*, 2012, **41**, 6519-6554.
54. G. Vali and E. Stansbury, *Canadian Journal of Physics*, 1966, **44**, 477-502.
55. G. C. Sosso, J. Chen, S. J. Cox, M. Fitzner, P. Pedevilla, A. Zen and A. Michaelides, *Chemical Reviews,* 2016, **116**, 7078-7116.
56. V. J. Schaefer, *Chemical Reviews*, 1949, **44**, 291-320.
57. P. J. Shamberger and M. J. O'Malley, *Acta Materialia*, 2015, **84**, 265-274.
58. R. Du, M. Wu, S. Wang, S. Wu, R. Wang and T. Li, *Applied Energy*, 2022, **325**, 119870.
59. N. Xie, Z. Huang, Z. Luo, X. Gao, Y. Fang and Z. Zhang, *Applied Sciences*, 2017, **7**, 1317.
60. D. E. Oliver, A. J. Bissell, X. Liu, C. C. Tang and C. R. Pulham, *CrystEngComm*, 2021, **23**, 700-706.
61. A. Sharma, V. V. Tyagi, C. R. Chen and D. Buddhi, *Renewable and Sustainable energy reviews*, 2009, **13**, 318-345.
62. K. K. Meisingset and F. Grønvold, *The Journal of Chemical Thermodynamics*, 1984, **16**, 523-536.
63. G. A. Lane and G. Lane, *Solar Heat Storage:Llatent Heat Materials*, CRC press Boca Raton, FL, USA:, 1983.
64. A. Safari, R. Saidur, F. Sulaiman, Y. Xu and J. Dong, *Renewable and Sustainable Energy Reviews*, 2017, **70**, 905-919.
65. N. Beaupere, U. Soupremanien and L. Zalewski, *Thermochimica Acta*, 2018, **670**, 184-201.
66. M. Yuan, D. Cao, C. Liu, C. Xu and Z. Liao, *Chemical Engineering Journal*, 2023, **469**, 143743.
67. C. Kutlu, Y. Su, Q. Lyu and S. Riffat, *Renewable Energy*, 2023, **206**, 848-857.
68. S. Karthika, T. K. Radhakrishnan and P. Kalaichelvi, *Crystal Growth & Design*, 2016, **16**, 6663-6681.
69. M. Sleutel, J. Lutsko, A. E. S. Van Driessche, M. A. Durán-Olivencia and D. Maes, *Nature Communications*, 2014, **5**, 5598.
70. L. G. Benning and G. A. Waychunas, in *Kinetics of Water-Rock Interaction*, Springer, 2008, pp. 259-333.
71. K. Binder and P. Virnau, *Soft Materials*, 2021, **19**, 267-285.
72. J. Diemand, R. Angélil, K. K. Tanaka and H. Tanaka, *The Journal of Chemical Physics*, 2013, **139**.
73. V. I. Kalikmanov, in *Nucleation Theory*, Springer, 2012, pp. 17-41.
74. M. Volmer and A. Weber, *Zeitschrift für physikalische Chemie*, 1926, **119**, 277-301.
75. R. Becker and W. Döring, in *Kinetic Treatment of the Nucleation in Supersaturated Vapors*, National Advisory Commitee for Aeronautics, 1954.
76. J. Frenkel, *The Journal of Chemical Physics*, 1939, **7**, 538-547.
77. K. Binder and D. Stauffer, *Advances in Physics*, 1976, **25**, 343-396.
78. N. J. Vitti, P. Majumdar and H. S. White, *Langmuir*, 2023, **39**, 1173-1180.
79. T. Adamski, *Nature*, 1963, **197**, 894-894.
80. P. G. Vekilov, *Crystal Growth & Design*, 2010, **10**, 5007-5019.
81. K. Binder, *Reports on Progress in Physics*, 1987, **50**, 783.
82. D. Kashchiev, *Nucleation*, Elsevier, 2000.

83. J. S. Rowlinson and B. Widom, *Molecular Theory of Capillarity*, Courier Corporation, 2013.
84. D. Turnbull, *Journal of Applied Physics*, 1950, **21**, 1022-1028.
85. B. Block, D. Deb, F. Schmitz, A. Statt, A. Tröster, A. Winkler, T. Zykova-Timan, P. Virnau and K. Binder, *The European Physical Journal Special Topics*, 2014, **223**, 347-361.
86. D. Winter, P. Virnau and K. Binder, *Physical Review Letters*, 2009, **103**, 225703.
87. L. Gránásy, *Journal of Non-Crystalline Solids*, 1993, **162**, 301-303.
88. A. Greer, *The Journal of Chemical Physics*, 2016, **145**.
89. Z. Fan and H. Men, *Metals*, 2022, **12**, 1547.
90. H. R. Pruppacher and J. D. Klett, in *Microphysics of Clouds and Precipitation*, Springer, 2010, pp. 225-281.
91. F. Binsbergen, *Progress in Solid State Chemistry*, 1973, **8**, 189-238.
92. P. G. Vekilov, *Nanoscale*, 2010, **2**, 2346-2357.
93. J. De Yoreo, *Nature Materials*, 2013, **12**, 284-285.
94. S.-Y. Chung, Y.-M. Kim, J.-G. Kim and Y.-J. Kim, *Nature Physics*, 2009, **5**, 68-73.
95. E. M. Pouget, P. H. Bomans, J. A. Goos, P. M. Frederik, G. de With and N. A. Sommerdijk, *Science*, 2009, **323**, 1455-1458.
96. J. Baumgartner, A. Dey, P. H. Bomans, C. Le Coadou, P. Fratzl, N. A. Sommerdijk and D. Faivre, *Nature Materials*, 2013, **12**, 310-314.
97. A. J. Pinto, E. Ruiz-Agudo, C. V. Putnis, A. Putnis, A. Jiménez and M. Prieto, *American Mineralogist*, 2010, **95**, 1747-1757.
98. J. M. Campbell, F. C. Meldrum and H. K. Christenson, *The Journal of Physical Chemistry C*, 2015, **119**, 1164-1169.
99. K. Li, S. Xu, W. Shi, M. He, H. Li, S. Li, X. Zhou, J. Wang and Y. Song, *Langmuir*, 2012, **28**, 10749-10754.
100. M. Wendt, L. K. Mahnke, N. Heidenreich and W. Bensch, *European Journal of Inorganic Chemistry*, 2016, **2016**, 5393-5398.
101. M. A. Levenstein, C. Anduix-Canto, Y. Y. Kim, M. A. Holden, C. González Niño, D. C. Green, S. E. Foster, A. N. Kulak, L. Govada and N. E. Chayen, *Advanced Functional Materials*, 2019, **29**, 1808172.
102. K. Sangwal, *Journal of Crystal Growth*, 2011, **318**, 103-109.
103. S. Krüger and J. Deubener, *Journal of Non-Crystalline Solids*, 2015, **417**, 45-51.
104. H.-J. Tong, J. P. Reid, J.-L. Dong and Y.-H. Zhang, *The Journal of Physical Chemistry A*, 2010, **114**, 12237-12243.
105. P. J. Shamberger and M. J. O'Malley, *Acta Materialia*, 2015, **84**, 265-274.
106. S. Kannan, M. Jog and R. M. Manglik, 2019.
107. J. A. Sellberg, C. Huang, T. A. McQueen, N. Loh, H. Laksmono, D. Schlesinger, R. Sierra, D. Nordlund, C. Hampton and D. Starodub, *Nature*, 2014, **510**, 381-384.
108. I. Sohn and R. Dippenaar, *Metallurgical and Materials Transactions B*, 2016, **47**, 2083-2094.
109. P. Song, Y. Li, T. Zhu and L. Zhang, *Steel Research International*, 2023, **94**, 2200684.

110. H. Asakawa, E. Holmstrom, A. S. Foster, S. Kamimura, T. Ohno and T. Fukuma, *The Journal of Physical Chemistry C*, 2018, **122**, 24085-24093.
111. K. Miyata, Y. Kawagoe, J. Tracey, K. Miyazawa, A. S. Foster and T. Fukuma, *The Journal of Physical Chemistry C*, 2019, **123**, 19786-19793.
112. Y. Hu, B. Lee, C. Bell and Y.-S. Jun, *Langmuir*, 2012, **28**, 7737-7746.
113. P. Ngankam, P. Schaaf, J. Voegel and F. Cuisinier, *Journal of Crystal Growth*, 1999, **197**, 927-938.
114. S. L. Riechers, K. M. Rosso and S. N. Kerisit, *The Journal of Physical Chemistry C*, 2017, **121**, 5012-5019.
115. M. Li, L. Wang, W. Zhang, C. V. Putnis and A. Putnis, *Crystal Growth & Design*, 2016, **16**, 4509-4518.
116. Y. Kuwahara, W. Liu, M. Makio and K. Otsuka, *Minerals*, 2016, **6**, 117.
117. A. Van Driessche, L. G. Benning, J. Rodriguez-Blanco, M. Ossorio, P. Bots and J. García-Ruiz, *Science*, 2012, **336**, 69-72.
118. F. Di Lorenzo, A. Burgos-Cara, E. Ruiz-Agudo, C. V. Putnis and M. Prieto, *CrystEngComm*, 2017, **19**, 447-460.
119. E. M. Pouget, P. H. Bomans, J. A. Goos, P. M. Frederik, G. de With and N. A. Sommerdijk, *Science*, 2009, **323**, 1455-1458.
120. J. P. Patterson, Y. Xu, M.-A. Moradi, N. A. J. M. Sommerdijk and H. Friedrich, *Accounts of Chemical Research*, 2017, **50**, 1495-1501.
121. M. A. Levenstein, C. Anduix-Canto, Y.-Y. Kim, M. A. Holden, C. González Niño, D. C. Green, S. E. Foster, A. N. Kulak, L. Govada, N. E. Chayen, S. J. Day, C. C. Tang, B. Weinhausen, M. Burghammer, N. Kapur and F. C. Meldrum, *Advanced Functional Materials*, 2019, **29**, 1808172.
122. W. Wu and G. H. Nancollas, *Journal of Colloid and Interface Science*, 1998, **199**, 206-211.
123. Q.-N. Zhang, Y. Zhang, C. Cai, Y.-C. Guo, J. P. Reid and Y.-H. Zhang, *The Journal of Physical Chemistry A*, 2014, **118**, 2728-2737.
124. J. H. Han and S. T. Martin, *Journal of Geophysical Research: Atmospheres*, 1999, **104**, 3543-3553.
125. G. Ketteler, S. Yamamoto, H. Bluhm, K. Andersson, D. E. Starr, D. F. Ogletree, H. Ogasawara, A. Nilsson and M. Salmeron, *The Journal of Physical Chemistry C*, 2007, **111**, 8278-8282.
126. Z. Jin, Y. Tian, X. Xu, H. Cui, W. Tang, Y. Yun and G. Sun, *Materials*, 2018, **11**, 1507.
127. F. Barrere, M. M. Snel, C. A. Van Blitterswijk, K. de Groot and P. Layrolle, *Biomaterials*, 2004, **25**, 2901-2910.
128. D. Ibbotson, S. Ahmed and P. J. Shamberger, *The Journal of Physical Chemistry C*, 2024, **128**, 17282-17290.
129. P. Konstantinov, T. Agopian and I. Tchokova, *Bulg. J. Meteorol. Hydrol*, 2000, **2000**, 13-16.
130. S. Von Euw, Q. Zhang, V. Manichev, N. Murali, J. Gross, L. C. Feldman, T. Gustafsson, C. Flach, R. Mendelsohn and P. G. Falkowski, *Science*, 2017, **356**, 933-938.
131. Z. Zou, X. Yang, M. Albéric, T. Heil, Q. Wang, B. Pokroy, Y. Politi and L. Bertinetti, *Advanced Functional Materials*, 2020, **30**, 2000003.

132. Z. Liu, Z. Zhang, Z. Wang, B. Jin, D. Li, J. Tao, R. Tang and J. J. De Yoreo, *Proceedings of the National Academy of Sciences*, 2020, **117**, 3397-3404.
133. Z. Fan and H. Men, *Materials Research Express*, 2020, **7**, 126501.
134. S. J. Cox, S. M. Kathmann, B. Slater and A. Michaelides, *The Journal of Chemical Physics*, 2015, **142**.
135. E. Rosky, W. Cantrell, T. Li, I. Nakamura and R. A. Shaw, *Atmospheric Chemistry and Physics*, 2023, **23**, 10625-10642.
136. H. Dashtian, H. Wang and M. Sahimi, *The Journal of Physical Chemistry Letters*, 2017, **8**, 3166-3172.
137. H. Jiang, A. Haji-Akbari, P. G. Debenedetti and A. Z. Panagiotopoulos, *The Journal of Chemical Physics*, 2018, **148**.
138. R. S. DeFever and S. Sarupria, *The Journal of Chemical Thermodynamics*, 2018, **117**, 205-213.
139. Z. Kargozarfard, A. Haghtalab, S. Ayatollahi and M. H. Badizad, *Industrial & Engineering Chemistry Research*, 2020, **59**, 22258-22271.
140. P. Pedevilla, M. Fitzner, G. C. Sosso and A. Michaelides, *The Journal of Chemical Physics*, 2018, **149**.
141. A. Soni and G.N. Patey, *The Journal of Physical Chemistry C*, 2022, **126**, 6716-6723.
142. M. Fitzner, P. Pedevilla and A. Michaelides, *Nature Communications*, 2020, **11**, 4777.
143. M. B. Davies, M. Fitzner and A. Michaelides, *Proceedings of the National Academy of Sciences*, 2022, **119**, e2205347119.
144. S. Toxvaerd, *The Journal of Chemical Physics*, 2002, **117**, 10303-10310.
145. H. Men, *The Journal of Chemical Physics,* 2024, **160**, 094702.
146. W. Gispen and M. Dijkstra, *The Journal of Chemical Physics*, 2023, **159**.
147. J. R. Espinosa, C. Vega, C. Valeriani, D. Frenkel and E. Sanz, *Soft Matter*, 2019, **15**, 9625-9631.
148. G. C. Sosso, J. Chen, S. J. Cox, M. Fitzner, P. Pedevilla, A. Zen and A. Michaelides, *Chemical Reviews*, 2016, **116**, 7078-7116.
149. H. Fu, M. Zhou, C. Chipot and W. Cai, *The Journal of Physical Chemistry B*, 2024, **128**, 9706-9713.
150. Z. Zou and P. Tiwary, *The Journal of Physical Chemistry B*, 2024, **128**, 3037-3045.
151. Z. Zou, PhD thesis, University of Maryland, 2024.
152. A. R. Finney and M. Salvalaglio, *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2024, **14**, e1697.
153. Y. Jin, G. R. Perez-Lemus, P. F. Zubieta Rico and J. J. de Pablo, *The Journal of Physical Chemistry A*, 2024, **128**, 7257-7268.
154. S. Vandenhaute, S. M. Rogge and V. Van Speybroeck, *Frontiers in Chemistry*, 2021, **9**, 718920.
155. B. Scheifele, I. Saika-Voivod, R. K. Bowles and P. H. Poole, *Physical Review E*, 2013, **87**, 042407.
156. Y. Shibuta, S. Sakane, E. Miyoshi, S. Okita, T. Takaki and M. Ohno, *Naure Communications*, 2017, **8**, 10.
157. Q. Liao, *Progress in Molecular Biology and Translational Science*, 2020, **170**, 177-213.

158. D. M. York, *ACS Physical Chemistry Au*, 2023, **3**, 478-491.
159. D. Ray, *Journal of Chemical Physics*, 2024, **161**.
160. V. Verma, H. Mitchell, M. Guo, B. K. Hodnett and J. Y. Y. Heng, *Faraday Discussions*, 2022, **235**, 199-218.
161. G. M. Torrie and J. P. Valleau, *Chemical Physics Letters*, 1974, **28**, 578-581.
162. A. Laio and F. L. Gervasio, *Reports on Progress in Physics*, 2008, **71**, 126601.
163. P. G. Bolhuis and C. Dellago, *The European Physical Journal Special Topics*, 2015, **224**, 2409-2427.
164. G. Lazzeri, H. Jung, P. G. Bolhuis and R. Covino, *Journal of Chemical Theory and Computation*, 2023, **19**, 9060-9076.
165. J. Behler and M. Parrinello, *Physical Review Letters*, 2007, **98**, 146401.
166. G. C. Sosso, M. Salvalaglio, J. r. Behler, M. Bernasconi and M. Parrinello, *The Journal of Physical Chemistry C*, 2015, **119**, 6428-6434.
167. J. Hegedüs and S. Elliott, *Nature Materials*, 2008, **7**, 399-405.
168. T. Lee and S. Elliott, *Physical Review Letters*, 2011, **107**, 145702.
169. J. Kalikka, J. Akola and R. Jones, *Physical Review B*, 2014, **90**, 184109.
170. W. Cantrell and A. Heymsfield, *Bulletin of the American Meteorological Society*, 2005, **86**, 795-808.
171. M. Polen, E. Lawlis and R. C. Sullivan, *Journal of Geophysical Research: Atmospheres*, 2016, **121**, 11,666-611,678.
172. G. Roudsari, M. Lbadaoui-Darvas, A. Welti, A. Nenes and A. Laaksonen, *Environmental Science: Atmospheres*, 2024, **4**, 243-251.
173. R. P. Sear, *Journal of Physics: Condensed Matter*, 2007, **19**, 466106.
174. T. F. Whale, M. A. Holden, T. W. Wilson, D. O'Sullivan and B. J. Murray, *Chemical Science*, 2018, **9**, 4142-4151.
175. A. Tabazadeh, Y. S. Djikaev and H. Reiss, *Proceedings of the National Academy of Sciences*, 2002, **99**, 15873-15878.
176. E. Mendez-Villuendas and R. K. Bowles, *Physical Review Letters*, 2007, **98**, 185503.
177. C. Hoose and O. Möhler, *Atmos. Chem. Phys.*, 2012, **12**, 9817-9854.
178. P. J. DeMott, O. Möhler, O. Stetzer, G. Vali, Z. Levin, M. D. Petters, M. Murakami, T. Leisner, U. Bundke and H. Klein, *Bulletin of the American Meteorological Society*, 2011, **92**, 1623-1635.
179. H. R. Pruppacher, J. D. Klett and P. K. Wang, *Microphysics of Clouds and Precipitation*, 1998, **28**, 381-382.
180. H. R. Pruppacher and R. Sänger, *Zeitschrift FürAangewandte Mathematik und Physik ZAMP*, 1955, **6**, 485-493.
181. K. Isono, M. Komabayasi and A. Ono, *Journal of the Meteorological Society of Japan. Ser. II*, 1959, **37**, 211-233.
182. A. J. Durant and R. A. Shaw, *Geophysical Research Letters*, 2005, **32**.
183. A. Kiselev, F. Bachmann, P. Pedevilla, S. J. Cox, A. Michaelides, D. Gerthsen and T. Leisner, *Science*, 2017, **355**, 367-371.
184. L. Lupi, A. Hudait and V. Molinero, *Journal of the American Chemical Society*, 2014, **136**, 3156-3164.
185. V. Sadtchenko, G. E. Ewing, D. R. Nutt and A. J. Stone, *Langmuir*, 2002, **18**, 4632-4636.

186. B. Wang, D. A. Knopf, S. China, B. W. Arey, T. H. Harder, M. K. Gilles and A. Laskin, *Physical Chemistry Chemical Physics*, 2016, **18**, 29721-29731.
187. T. Kovács and H. K. Christenson, *Faraday Discussions*, 2012, **159**, 123-138.
188. H. Christenson, *Physical review letters*, 1995, **74**, 4675.
189. V. Molinero and E. B. Moore, *The Journal of Physical Chemistry B*, 2009, **113**, 4008-4016.
190. M. Fitzner, G. C. Sosso, S. J. Cox and A. Michaelides, *Journal of the American Chemical Society*, 2015, **137**, 13658-13669.
191. P. Conrad, G. E. Ewing, R. L. Karlinsey and V. Sadtchenko, *The Journal of Chemical Physics*, 2005, **122**.

Chapter 2
Establishment of a data-driven nucleator prediction model

Building on the theoretical foundations outlined in Chapter 1, this chapter describes the methodology developed to construct and evaluate a data-driven model for predicting the efficacy of nucleators (NUCs) in phase change materials (PCMs). The focus is on combining crystallographic features with carefully curated experimental datasets to provide a robust framework for identifying working and non-working NUCs. The chapter is structured into four sections.

Section 2.1 addresses the reliability of data sources, both crystallographic and experimental, emphasising why confidence in the underlying datasets is essential for developing a trustworthy prediction model.

Section 2.2 introduces geometrical matching as the principal screening criterion. It explains the rationale for prioritising translational symmetry features over other factors, grounds this choice in epitaxial growth theory, and formalises the definition of geometrical matching. The section also develops the constructive algorithm for identifying super-cell matches and discusses the practical threshold values that constrain the search space.

Section 2.3 considers the challenge of identifying equivalent slabs and symmetry reduction. Here, the treatment of symmetry-related Miller indices, basis transformations, and slab construction parameters is detailed to ensure that the dataset consists only of unique and physically meaningful slab terminations.

Finally, Section 2.4 outlines the overall prediction model training workflow. This includes the generation of slab libraries, the interface-matching process, and the definition of geometrical features. It also describes the optimisation strategy for selecting threshold values that best discriminate between working and non-working NUCs, thus linking the crystallographic basis of the methodology to the data-driven framework developed in later chapters.

2.1. Reliability of data sources

The reliability of the data sources used in this study is of central importance, as the robustness of any computational or machine learning framework ultimately depends on the quality of the input data on which it is trained and validated. In the present work, two distinct but complementary considerations arise: (i) the quality of crystallographic information underpinning the geometric analysis of slabs, and (ii) the reliability of experimental datasets used to define working and non-working nucleators for training and testing purposes.

2.1.1. Crystallographic data from database

Crystallographic structures of both PCMs and potential NUCs were primarily sourced from the Inorganic Crystal Structure Database (ICSD release 2025.1)[1], which is the most widely used and authoritative repository of experimentally determined inorganic structures. The ICSD is curated and peer-reviewed, providing atomic coordinates, symmetry information, and crystallographic parameters derived from X-ray diffraction single crystal and powder patterns (PXRD)[2] and neutron diffraction studies[3]. While no crystallographic database is entirely free of error, the ICSD is generally regarded as highly reliable for structural applications.

In the context of this thesis, the accuracy of crystallographic data is particularly critical because the methodology involves slicing slabs according to Miller indices from (001) to $hkl \leq 3$ for cubic unit cells and $hkil \leq 3$ for hexagonal unit cells, and subsequently evaluating their geometric compatibility. Any inaccuracies in lattice parameters or atomic positions would propagate into slab geometries and may distort the calculated interfacial matches. For this reason, only well-determined structures from the ICSD were used, and duplicates or incomplete entries were excluded. As such, although the approach necessarily inherits the limitations of the source database, the reliance on the ICSD ensures that the crystallographic component of the methodology rests upon the most consistent and reliable information available.

2.1.2. Reliability of experimental data

In parallel, the experimental datasets used to either train or validate the transferability of the prediction model require careful consideration of their reliability. For ice

nucleation, a set of ten compounds was selected based on past peer-reviewed and reputable studies (see Chapter 3). These compounds were chosen on the basis of consistent and reproducible reports of their nucleation activity, thereby providing a trusted baseline dataset for training the initial model.

For salt hydrates, the case of sodium acetate trihydrate (SAT) was considered in detail (see Chapter 4). A training set of eighteen nucleators was adopted from the work of David Oliver[4], a predecessor researcher who critically assessed the ability of these compounds to promote or inhibit nucleation of SAT. This dataset is particularly reliable as it was the outcome of a systematic and careful evaluation of NUC performance, providing a balanced set of both working and non-working examples. The confidence in the training data for SAT therefore derives directly from the rigour of this earlier study, as well as its continuity within the research programme.

For the transferability tests, additional datasets were collated for four further hydrated salts: calcium chloride hexahydrate ($CaCl_2 \cdot 6H_2O$), magnesium nitrate hexahydrate ($Mg(NO_3)_2 \cdot 6H_2O$), lithium nitrate trihydrate ($LiNO_3 \cdot 3H_2O$), and magnesium chloride hexahydrate ($MgCl_2 \cdot 6H_2O$). NUC data for these systems were obtained through a targeted literature survey, with results summarised in Chapter 4. Importantly, for each PCM, examples of both effective and ineffective nucleators were identified. This duality is essential, as it provides a more rigorous basis for evaluating whether the geometric slab-matching framework is capable of distinguishing not only successful nucleators but also false candidates. Moreover, many of the relevant studies originated from the same research groups, which helps minimise discrepancies arising from variations in experimental setup that might otherwise influence observed degrees of subcooling.

Together, these two tiers of reliability assessment, i.e. crystallographic integrity on the one hand, and experimental rigour on the other, provide the foundation for the methodology developed in this thesis. By relying on curated crystallographic databases and carefully validated experimental datasets, this study ensures that the subsequent modelling and prediction exercises are grounded in the most robust sources of

information currently available, while also acknowledging the limitations that inevitably accompany them.

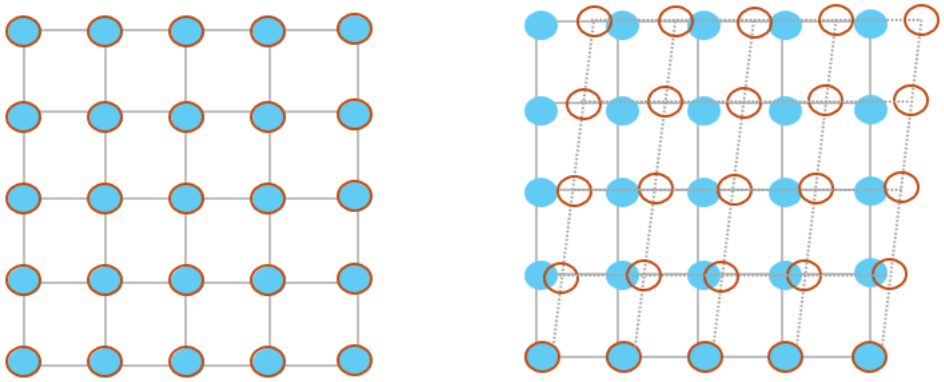
2.2. Geometrical matching as screening criterion

Amongst the wide range of factors known to influence heterogeneous nucleation that were discussed in Section 1.4, including surface chemistry, interfacial interactions, and lattice mismatch, this thesis deliberately prioritises the translational symmetry at the interface relative to the bulk structures on either side, i.e. the degree of geometric matching as the principal feature for assessing NUC potential. This decision is motivated by two principals. First, geometrical features enable a high-throughput and systematic screening of thousands of possible interfaces, which would be impractical if relying on more computationally demanding interaction models. Second, the choice aligns with the established mechanism of epitaxial growth theory in heterogeneous nucleation, wherein crystallographic registry between NUC and PCM surfaces is recognised as a decisive factor in determining nucleation efficacy.

2.2.1. Epitaxial growth and its relevance to geometric matching

Compatibilities between two crystallographic slabs with specific Miller indices are defined by assessing whether the interface's translational symmetry aligns with that of the adjacent bulk structures within an acceptable precision. This concept is referred to as "geometric match".

In recent years, there has been a growing interest in the study of epitaxial growth, where one material is deposited onto another in a controlled manner. The concept of epitaxial growth originates in crystallography, where it describes the ordered growth of a crystalline layer on the surface of another crystal. In epitaxial systems, the overgrown material adopts a crystallographic orientation that is aligned, either fully (see Figure 2.1(a)) or partially (see Figure 2.1(b)), with the underlying substrate. This process depends critically on the extent to which the atomic lattices of the two phases can be brought into registry. A small degree of mismatch between lattice parameters as shown in Figure 2.1(b) may be tolerated through strain accommodation, but beyond a certain threshold, epitaxial growth becomes unfavourable, and defects proliferate.



(a) (b)

Figure 2.1. (a) Illustration of overview of fully matching heterogeneous epitaxial growth, representing a good registry, where the lattice constant of the epitaxial film (blue dot, corresponding to the PCM) is matched to the lattice constant of the substrate (orange hollow dot, corresponding to the NUC); **(b)** Illustration of partial epitaxial growth, representing a strained registry, arising as the epitaxial film has a different lattice constant than the substrate.

The principle is directly relevant to heterogeneous nucleation. In this context, the NUC plays the role of the substrate, while the PCM corresponds to the overgrown crystal. The ability of a NUC to promote nucleation depends on how effectively the crystal structure of the NUC provides a template for the emerging lattice of the PCM. Thus, epitaxial theory provides a clear mechanistic justification for using geometric compatibility between surfaces as a proxy for nucleation propensity.

To quantify this, crystallographic slabs can be constructed along relevant Miller indices for both the PCM and the NUC. Geometric matching features, such as vector mismatches, angle mismatch, and interfacial area registry, are then computed to evaluate the degree of epitaxial fit. In this way, the geometrical registry derived from epitaxial principles offers a tractable and physically grounded criterion for high-throughput screening.

By grounding the feature selection in epitaxial growth theory, this study ensures that the chosen feature is not only computationally efficient but also theoretically motivated. The slab-matching approach employed here can therefore be viewed as a

crystallographic generalisation of epitaxy, adapted to the problem of predicting nucleation activity in PCMs.

2.2.2. Definition of geometrical matching

The formalisation of the geometrical lattice matching theory, as presented in this Chapter (see below), offers a comprehensive and generalised approach applicable to virtually any pair of crystal structures and any interfacial orientation. Under this framework, two lattices are considered to exhibit a geometric match if the two-dimensional lattices formed by their crystal translations parallel to the interface share a common supercell. However, due to inherent structural differences, an exact match is generally unattainable. Instead, a pair of nearly identical supercells (one for each side of the interface) are defined, while establishing rigorous tolerance criteria for deviations. These criteria include vector mismatches between the two slabs, angular deviations, and discrepancies in the unit cell area. To facilitate systematic assessment, a methodology is proposed for deriving a standard slab representation for any given crystal structure, defining it as a parallelogram characterised by side lengths $\vec{a}$ and $\vec{b}$ and an acute interfacial angle α. The degree of mismatch is then quantified through direct comparisons of these parameters, along with additional derived criteria, see Figure 2.2.

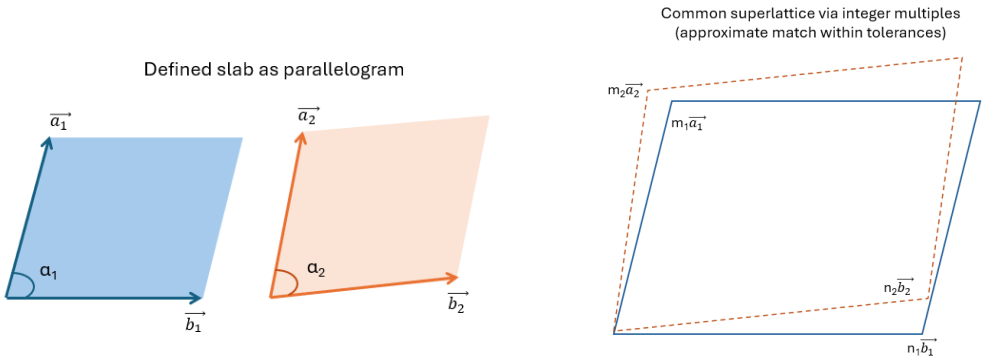


Figure 2.2. Illustration of geometrical lattice matching between two crystallographic slabs. **(a)** Two lattice cells, defined by side vectors a and b and an interfacial angle α, typically show small differences in parameters between the nucleator and PCM. **(b)** By expanding the cells to integer multiples, a common supercell can be constructed.

A key contribution of this study is the development of a constructive algorithm for identifying all possible slab matches within a given mismatch threshold and a unit cell area below a prescribed limit.

A central problem in formalising geometrical matching is to establish a rigorous criterion for when two two-dimensional slabs can be considered to form a match. Under the framework developed in this thesis, two lattices are regarded as matching if each possesses a supercell such that the two supercells are identical up to a rotation, reflection, or choice of primitive translations.

At first glance, exhaustively enumerating and comparing all possible supercells of two given slabs may appear computationally prohibitive. However, the problem simplifies when considering the relationship between the unit-cell areas of the lattices. By definition, the area of any supercell is an integer multiple of the unit-cell area of the original lattice. Given two initial lattices with unit-cell areas A_1 and A_2, their respective supercells will have areas r_1A_1 and r_2A_2, where r_1 and r_2 are integers. For the supercells to match, the condition

$$r_1 A_1 = r_2 A_2 \tag{1}$$

must hold.

In practice, exact equality is rarely achieved because of inherent structural differences between lattices. Instead, approximate solutions are sought by introducing a numerical tolerance that constrains the allowable deviation between r_1A_1 and r_2A_2. This greatly reduces the search space, since only values of r_1 and r_2 lying within the tolerance band need to be explored. Furthermore, for any given lattice, the number of possible supercells of area nA is finite, bounded above by the sum of the divisors of n. This makes systematic enumeration of candidate matches computationally tractable.

The high-throughput screening methodology strongly suggests that sustained epitaxial growth with periodic interfacial structure, which is a prerequisite for effective heterogeneous nucleation, can only occur when a high degree of geometric match is present. The combination of geometric matching principles with data-driven analysis techniques enables the efficient extraction and analysis of slab characteristics for both PCMs and NUCs. By incorporating predefined mismatch tolerance criteria, this

approach offers a systematic and scalable solution to the broader scientific question addressed in this thesis. Ultimately, the methodological advantages in terms of approachability, computational efficiency, and compatibility with high-throughput screening make this the most viable strategy for addressing the nucleation problem in ice and in inorganic salt PCMs.

The geometric matching strategy implemented in this thesis is conceptually rooted in the epitaxial lattice-matching framework introduced by Zur and McGill[8], who developed a systematic algorithm for identifying commensurate supercells between crystalline interfaces. Their method enumerates reduced two-dimensional lattice vectors and evaluates candidate matches based on lattice mismatch, angular deviation, and supercell area, providing a purely geometric criterion for interface compatibility. In this respect, the present workflow shares the same foundational philosophy: interface feasibility is first evaluated through crystallographic commensurability and strain minimisation in two dimensions.

However, several important distinctions arise due to differences in material systems and research objectives. Zur and McGill focused on rigid ionic thin-film systems (*e.g.*, CdTe/GaAs, CdTe/sapphire), where interface chemistry was explicitly excluded from consideration. In contrast, the systems studied here involve hydrated molecular crystals, such as sodium acetate trihydrate, characterised by flexible coordination environments, hydrogen bonding, and surface-dependent structural variability. Consequently, the geometric filtering criteria employed in this work incorporate additional physically motivated constraints, including vector mismatch thresholds, angular tolerances, area consistency conditions, and higher-order descriptors reflecting surface strain and structural feasibility. These constraints are adapted to account for the structural softness and reconstruction behaviour typical of phase-change materials.

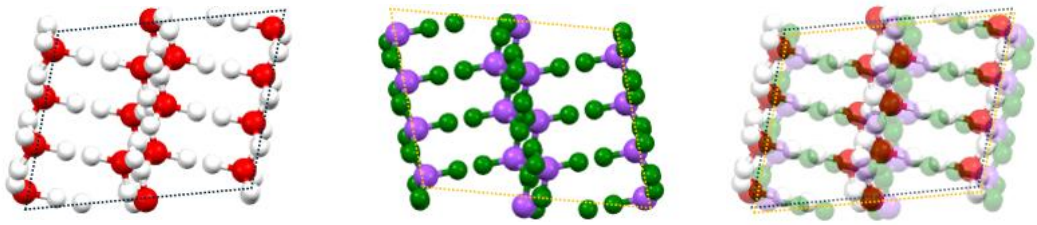
Furthermore, although Zur and McGill qualitatively noted that excessively large supercells may weaken chemical reinforcement at the interface, their framework does not formalise a quantitative upper bound. In the present work, the maximum area overlap criterion is explicitly defined and physically rationalised in terms of interfacial strain distribution, defect accommodation, and the requirement for periodic structural

reconstruction. This provides a systematic constraint on supercell size tailored to heterogeneous nucleation in PCMs.

Finally, the scope of this thesis extends beyond geometric matching. Whereas the Zur-McGill framework terminates upon identifying supercells that satisfy prescribed tolerances, the current research uses the resulting geometric matches as structured descriptors within a data-driven classification model. By linking crystallographic compatibility to experimentally measured nucleation behaviour, the methodology advances from deterministic geometric screening to predictive modelling, enabling probabilistic assessment and feature-importance analysis. In this way, the present work builds upon and generalises the Zur–McGill approach for application to complex molecular systems and heterogeneous nucleation phenomena.

2.3. Identifying equivalent slabs and symmetry reduction

A key methodological challenge in geometric slab matching is determining when two nominally distinct slabs are in fact crystallographic equivalent or nearly identical. If symmetry-related slabs are incorrectly treated as unique, the dataset becomes artificially inflated, and geometric descriptors such as lattice mismatch metrics and correlation matrices may become dominated by redundant entries, leading to biased statistical interpretation. For example, Miller indices such as (121) and $(1\bar{2}1)$ generate surfaces that are geometrically equivalent, and if both were included independently the correlation structure of the dataset would collapse toward zero, see Figure 2.3. It is therefore essential to rigorously identify and eliminate such redundancies to ensure that each slab considered represents a genuinely distinct surface termination.



(a) (b) (c)

Figure 2.3. Symmetry-equivalent slabs in ice-I_h. (a) A Miller index of (121) termination viewed along the a-axis; (b) A Miller index of $(1\bar{2}1)$ termination, viewed along the a-axis; (C) Overlay of the two slabs, showing that they are geometrically equivalent after symmetrical operation. Recognition of such equivalences prevents redundant slab generation and ensures that geometric features represent only distinct surface terminations.

Two principal sources of ambiguity must be addressed. First, slabs may appear distinct due to rotation, inversion, or mirroring, when in fact they represent the same crystallographic surface. Second, lattice descriptions are inherently non-unique; any lattice may be generated by multiple sets of primitive translation vectors, leading to alternative but equivalent representations of the same structure. Consequently, two geometrically identical slabs may appear distinct if expressed using different lattice bases, potentially leading to false mismatches during interfacial comparison.

The first issue was resolved by employing the `SpacegroupAnalyzer` class in `pymatgen.symmetry.analyzer`, which identifies space group equivalences, in combination with the `StructureMatcher` class from `pymatgen.analysis.structure_matcher`. The `StructureMatcher` class searches for transformations (rotation + translation) that map one structure onto another, incorporating inversion symmetry and applying tolerance thresholds to account for small numerical variations. Together, these tools allow recognition of equivalent slabs even when their orientations differ, ensuring that redundant slab pairs are consolidated. These checks were incorporated directly into the slab-processing pipeline by subtracting coordinates of PCM and NUC slabs prior to interface mismatch evaluation (`get_uniq_layercoords`).

The second issue was addressed by applying a Niggli lattice reduction scheme[5] to standardise the representation of slabs. This approach transforms any arbitrary lattice basis into a unique reduced form, preventing false mismatches that would arise if two equivalent lattices were described by different basis vectors. Within Pymatgen, the `SpacegroupAnalyzer` class provides a built-in Niggli reduction implementation, which were employed to ensure equivalence identification. In practice, slabs were generated using `generate_all_slabs()` with conventional cells to remain comparable with experimental crystallography. Subsequently, a 2D reduction heuristic was applied once per slab at the start of the matching stage. This replaces the lattice basis with its reduced equivalent prior to supercell construction, ensuring that structurally equivalent slabs are identified consistently. Figure 2.4 illustrates the effect of Niggli reduction on a two-dimensional lattice. The lattice points (grey) remain unchanged, while the basis vectors spanning the unit cell are replaced. In general, a lattice basis [6] may be transformed into an equivalent basis $\{a', b'\}$ through an integer unimodular transformation:

$$\begin{pmatrix} \vec{a}' \\ b' \end{pmatrix} = U \begin{pmatrix} \vec{a} \\ \vec{b} \end{pmatrix}, U = \begin{pmatrix} p & q \\ r & s \end{pmatrix}, \det(U) = \pm 1$$

Such transformations preserve lattice periodicity and cell area while altering the orientation and skewness of the chosen basis vectors. Niggli reduction selects a canonical basis satisfying metric minimality conditions, ensuring a unique lattice representation without modifying the underlying geometric structure. Although the lattice itself is invariant, the change of basis may affect the numerical transparency of subsequent supercell matching.

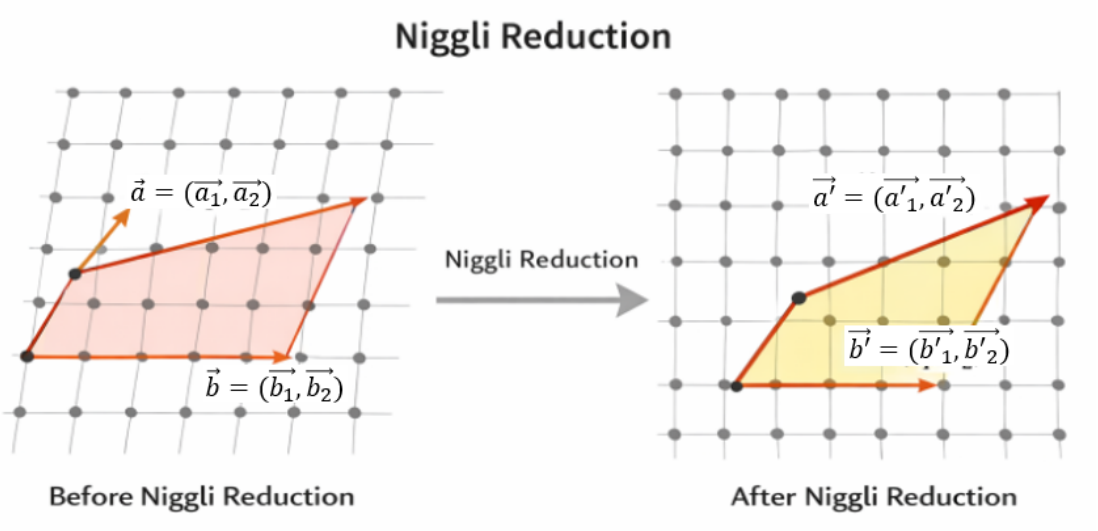


Figure 2.4. Schematic illustration of Niggli lattice reduction in two dimensions.

Although Niggli reduction preserves the metric geometry of the lattice, it may alter the Cartesian orientation and skewness of the chosen basis vectors. Consequently, while the underlying surface lattice remains unchanged, the reduced basis is not always the most geometrically intuitive representation for supercell construction. In low-symmetry systems such as monoclinic and triclinic crystals, rational commensurability relationships that are visually transparent in a conventional setting may become less numerically straightforward after reduction. Supercell construction may therefore require larger integer combinations, and near-commensurate overlaps may appear "irrational" within finite numerical tolerance. These effects reflect a trade-off between lattice uniqueness and basis convenience: reduction improves robustness and prevents duplicate slab representations, but may increase the apparent complexity of interface matching in geometrically skewed systems.

In terms of slab construction, thickness and vacuum spacing were also considered. Slab thickness must correspond to integer multiples of the crystallographic repeat units; for example, a unit cell of 5 Å cannot yield a slab of 7 Å without violating symmetry. `Pymatgen` implements this constraint by allowing users to set `min_slab_size` (minimum slab thickness) and `min_vacuum_size` (minimum vacuum size) as lower bounds, with the algorithm selecting the closest valid configuration. Vacuum thickness, by contrast, is an external parameter that does not affect crystallographic periodicity

and is more easily adjusted to ensure electronic isolation of surfaces. In practice, vacuum convergence was performed first to guarantee slab isolation, followed by convergence of slab thickness.

Another challenge arises when generating slabs from non-orthogonal crystal systems, particularly monoclinic and triclinic structures. In these systems, the crystallographic axes are not mutually perpendicular, and Miller indices (hkl) must be interpreted through reciprocal lattice geometry rather than direct Cartesian directions. While in orthogonal systems the plane normal associated with (hkl) aligns intuitively with the corresponding lattice directions, this simplification does not hold in skewed cells. The true surface normal is defined by a linear combination of reciprocal lattice vectors, and therefore the geometric interpretation of negative Miller indices (*e.g.*, −h, −k, −l) becomes less transparent.

This distinction is illustrated schematically in Figure 2.5. In a tetragonal lattice (left), the planes (101) and (-101) are related by symmetry and therefore correspond to geometrically equivalent surface orientations when viewed along the b-axis. In contrast, in a monoclinic lattice (right), the skewed geometry removes this symmetry relation, and the same sign change produces a distinct plane orientation. The difference arises not from the negative index itself, but from the reduced crystallographic symmetry of the lattice.

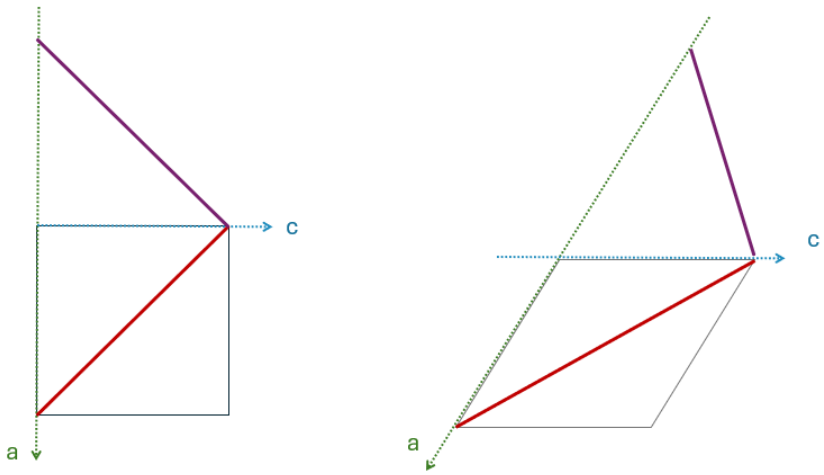


Figure 2.5. Schematic comparison of Miller index interpretation in high- and low-symmetry systems, shown as a projection along the b -axis (displaying only the a–c plane). The red line represents cleavage along the (101) plane, while the purple line represents cleavage along the (-101) plane. In the tetragonal lattice (left), the cleaved planes are symmetry-related and correspond to equivalent surface orientations. In the monoclinic lattice (right), the skewed geometry removes this symmetry relation, and the same sign change results in distinct plane orientations.

Care must therefore be taken to ensure that slab construction uses reciprocal-space definitions to correctly determine plane orientation and to avoid inadvertent duplication or omission of symmetry-related surfaces. These effects are particularly pronounced in triclinic and monoclinic systems, where the absence of symmetry constraints increases the likelihood of generating distinct but closely related surface terminations. Consequently, surface enumeration and matching in low-symmetry crystals are inherently more sensitive to indexing conventions and numerical tolerances.

By combining symmetry recognition (space group analysis, structure matching) with lattice reduction (Niggli scheme), this workflow ensures that only unique slabs are retained for geometric matching. This treatment is essential for preserving both the physical validity and statistical integrity of the subsequent slab-matching analysis. A computational pipeline for unique-slab generation and geometric matching is listed in Figure 2.6.

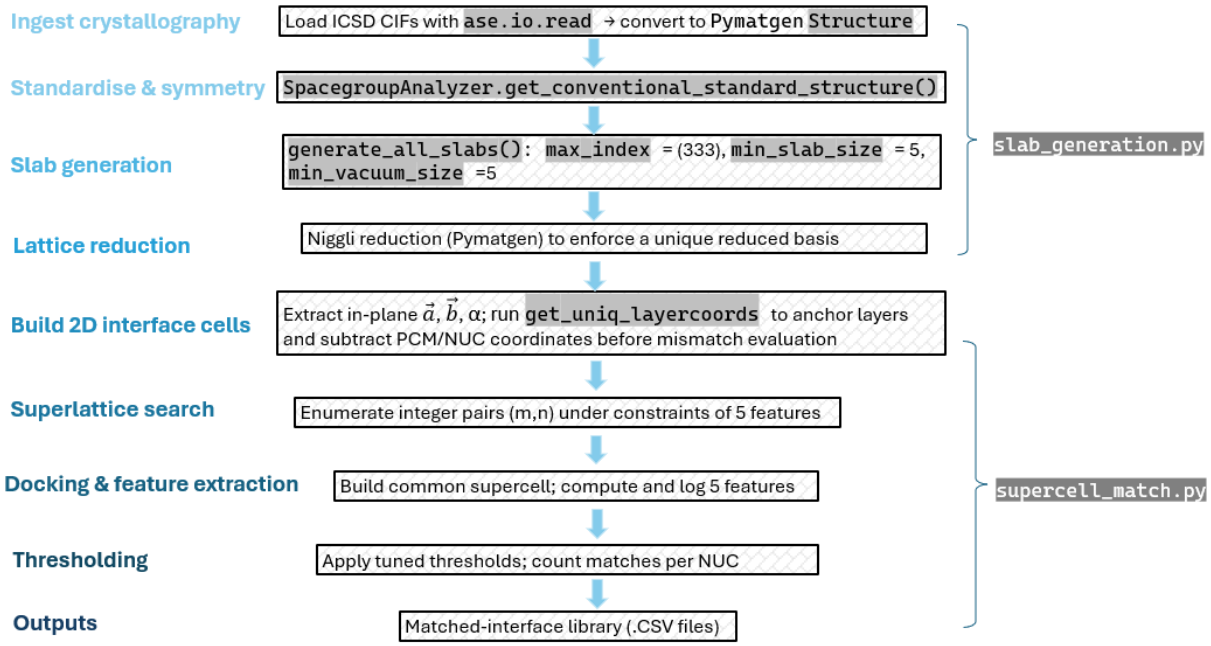


Figure 2.6. Computational flow diagram for unique-slab generation and geometric matching. The workflow integrates multiple Python modules.

2.4. Prediction model training workflow

With an experimental data set thus defined, the crystallographic information files (CIFs) for the 10 compounds as potential nucleator candidates of ice and 18 compounds as potential nucleator candidates of SAT were used as input models for the interface-matching training process. An overview of the general workflow, which was constructed in `Python 3`, underpinned by **ASE**[7] and `Pymatgen`[8] is presented in Figure 2.7.

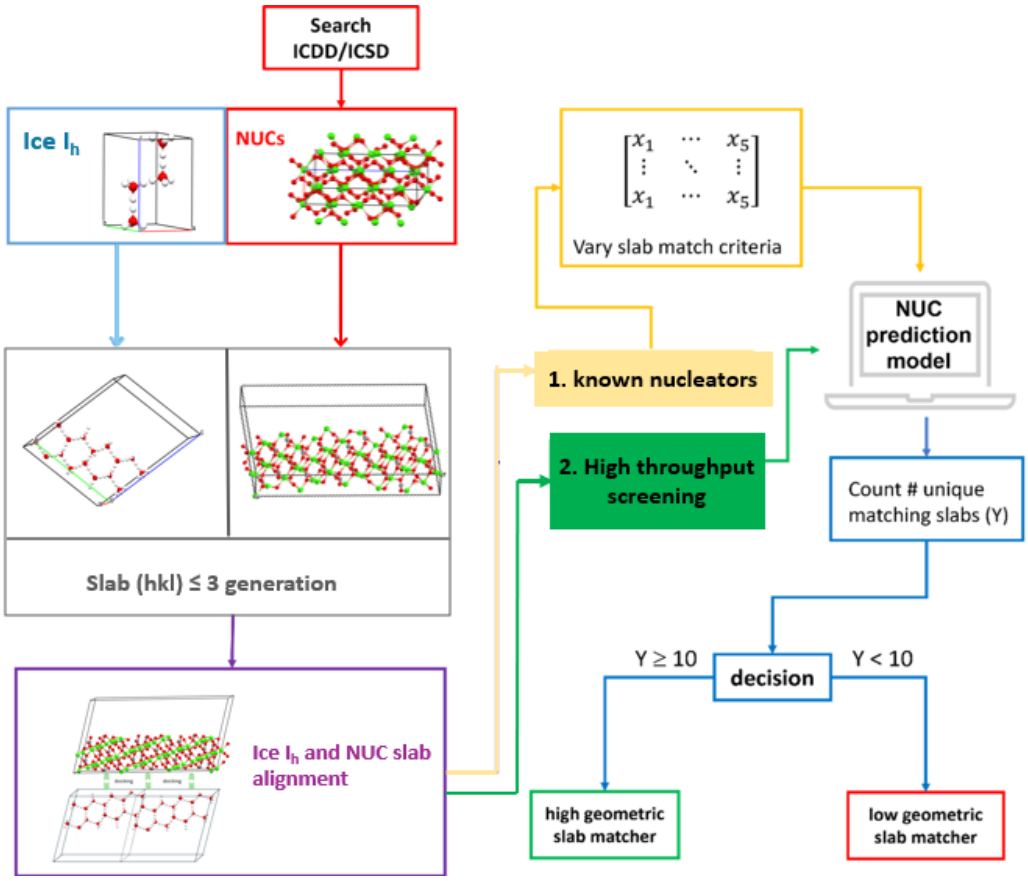


Figure 2.7. Workflow of nucleator prediction model.

Throughout this work, crystal surfaces are referred to using Miller indices (hk(i)l), which define specific cut planes in the unit cell. Interface-matching refers to the geometric alignment of these cut planes, evaluated according to a set of criteria based on area overlap, angle and unit cell mismatch.

In the first instance, sets of surfaces were created by cleaving the corresponding bulk crystal lattices along the Miller index planes hkil ≤ 3 for PCMs and for hk(i)l ≤ 3 for the nucleators to create a pool of 64 non-duplicated surfaces for each crystal lattice. Allowing all PCM surfaces to dock on all NUC surfaces generated a total of 2,401 non-duplicated interface models per NUC. A pseudocode is demonstrated in Algorithm 2.1:

Algorithm 2.1. Slab generation from bulk structures.

1. Input: bulk crystal structure S, set of Miller indices {(hkl)}, minimum slab thickness, vacuum thickness;
2. For each (hkl) in {(hkl)}:
   a. Generate slab S_{hkl} by cleaving S along (hkl)
   b. Export slab structure and record data (crystal ID, (hkl), filepath)
3. Output: collection of slab structures and metadata table.

The geometric matching of two docked surfaces was assessed by searching through integral multiples of the vectors of each surface to find the supercell models that present the smallest unit-cell mismatch. The geometry of each surface is described by two vectors parallel to each slab edge, $\vec{a}$ and $\vec{b}$, expressed as (m, n) supercells, such that $m \cdot \vec{a} \approx n \cdot \vec{b}$. These define the new vectors $\vec{u}$ and $\vec{v}$, respectively (see Figure 2.5). A reduction scheme was then used to express the vectors in the slab frame of reference, to negate the effects of translation, rotation or reflection of the individual surfaces. The two slabs were aligned by minimizing $\left|\frac{||\vec{u_1}||}{||\vec{u_2}||} - 1\right|$ and $\left|\frac{||\vec{u_2}||}{||\vec{u_2}||} - 1\right|$ where the subscripts 1 and 2 denote PCM and NUC supercell surfaces, respectively. Following translational alignment, the slabs were then rotated by transformational matrices $R(\theta)_i$ to lie parallel to each other. The surface generation, alignment and subsequent docking procedures are summarized in Figure 2.8.

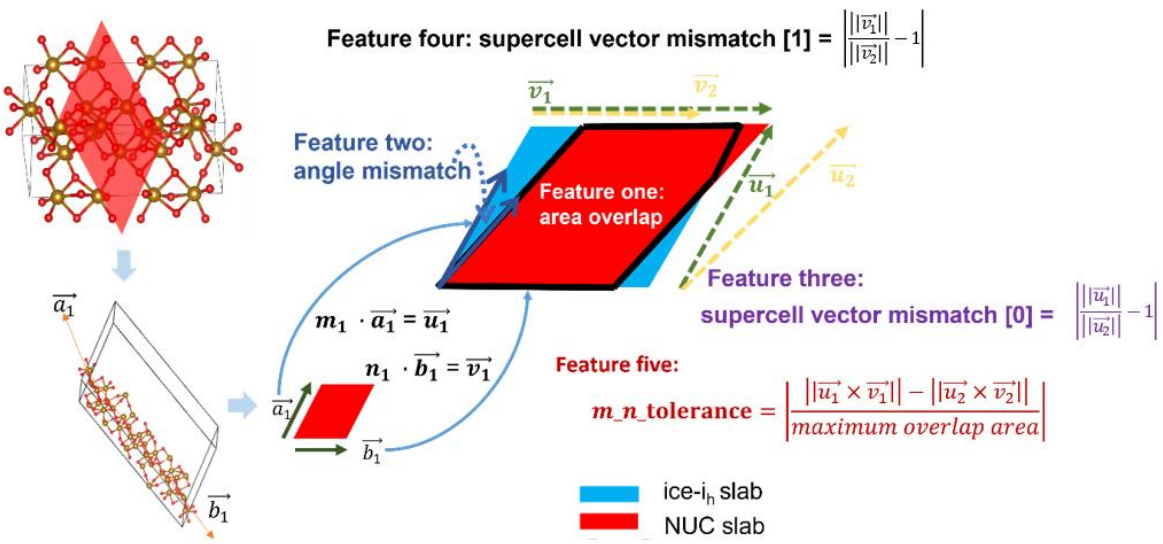


Figure 2.8. The four supercell matching features used to assess PCM/NUC docking.

Five features were then defined to quantify the quality of fit for the resulting bank of ice-I_h/nucleator interfaces (Figure 2.9). These were the (i) maximum area overlap, (ii) angle mismatch, (iii) supercell vector mismatch [0] for vector $\vec{u}$, and the (iv) supercell vector mismatch [1] for vector $\vec{v}$, according to Equations (2-5). Finally, to temper the (m, n) supercell generation to sensible outcomes compared to the maximum area overlap, a maximum value of tolerance for variables m and n is set, according to Equation (6).

$$\text{Maximum area overlap} = \left|\left|\vec{u_1} \times \vec{v_1}\right|\right| \approx \left|\left|\vec{u_2} \times \vec{v_2}\right|\right| \tag{2}$$

$$\text{Angle mismatch} = \angle(\vec{u_1}, \vec{u_2}) - \angle(\vec{v_1}, \vec{v_2}) \tag{3}$$

$$\text{Supercell vector mismatch [0]} = \left|\frac{||\vec{u_1}||}{||\vec{u_2}||} - 1\right| \tag{4}$$

$$\text{Supercell vector mismatch [1]} = \left|\frac{||\vec{v_1}||}{||\vec{v_2}||} - 1\right| \tag{5}$$

$$m\_n\_\text{tolerance} = \left|\frac{\left|\left|\vec{u_1} \times \vec{v_1}\right|\right| - \left|\left|\vec{u_2} \times \vec{v_2}\right|\right|}{maximum\ area\ overlap}\right| \tag{6}$$

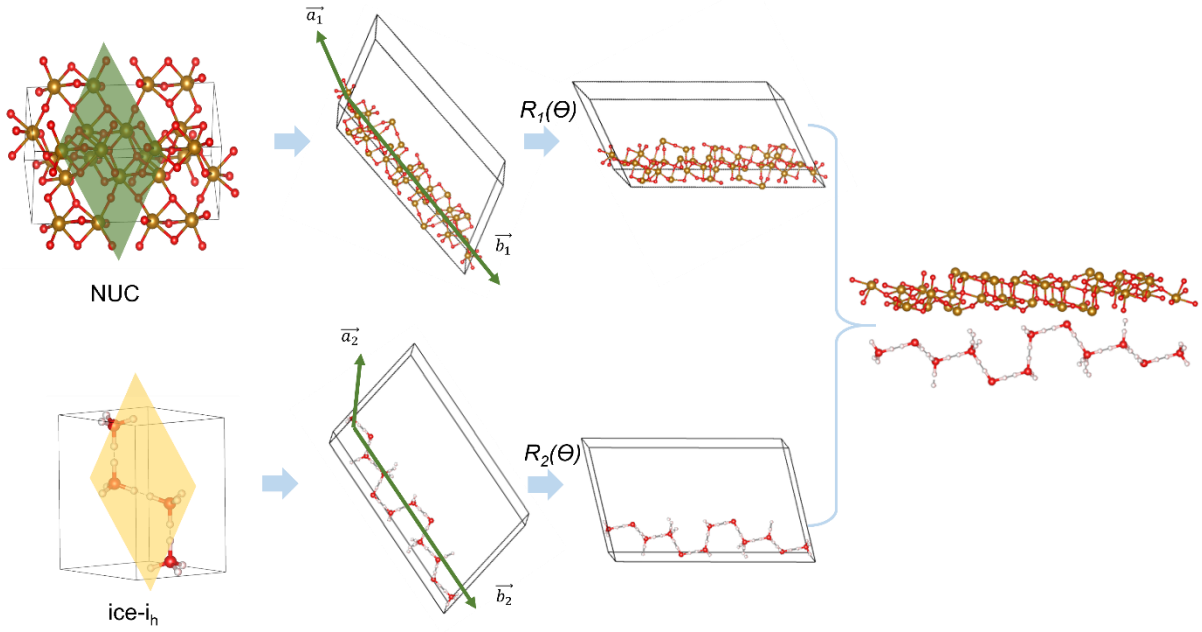


Figure 2.9. NUC and PCM slab generation and docking.

A pseudocode demonstrating the match of two slabs and computing of descriptors is as follows:

Algorithm 2.2. Identification of best-matching supercells between PCM and NUC slabs.

1. Input: PCM slab, NUC slab, thresholds (A_{max}, maximum angle mismatch ε_θ, maximum vector mismatches $\varepsilon_u = \varepsilon_v$, m_n_tolerance ε_r)
2. Compute primitive surface areas A_{PCM}, A_{NUC} from $\left|\left|\vec{a} \times \vec{b}\right|\right|$
3. Construct candidate multiplier pairs (r_1, r_2) such that $\left|\frac{|||\vec{u_1}\times\vec{v_1}||-||\vec{u_2}\times\vec{v_2}||}{maximum\ area\ overlap}\right| \leq \varepsilon_r$ and both supercell areas $\leqslant A_{max}$
4. For each candidate (r_1, r_2):
   a. Enumerate PCM supercells with multiplier r_1
   b. Enumerate NUC supercells using HNF with multiplier r_2
   c. For each pair of candidate supercells $(\vec{u_1}, \vec{v_1})$, $(\vec{u_2}, \vec{v_2})$:
      i. Compute length mismatches $\delta_u = \left|\frac{||\vec{u_1}||}{||\vec{u_2}||} - 1\right|$, $\delta_v = \left|\frac{||\vec{v_1}||}{||\vec{v_2}||} - 1\right|$
      ii. Compute angle mismatch $\Delta\theta = \angle(\vec{u_1}, \vec{u_2}) - \angle(\vec{v_1}, \vec{v_2})$

iii. Compute areas $A_1 = \left||\overrightarrow{u_1} \times \overrightarrow{v_1}|\right|$, $A_2 = \left||\overrightarrow{u_2} \times \overrightarrow{v_2}|\right|$, $A\max = \max(A1, A2)$

iv. Accept match if $|\delta_u| < \varepsilon_u$, $|\delta_v| < \varepsilon_v$, and $\Delta\theta < \varepsilon_\theta$

5. Select best accepted match according to rule
6. Output: best match vectors and descriptor tuple (A_{max}, $\Delta\ \theta$, $|\delta_u|$, $|\delta_v|, m_n$_tolerance)

The final step in the slab-matching workflow is to determine the most effective set of numerical thresholds for distinguishing between working and non-working nucleators. Rather than fixing tolerance values a priori, a systematic optimisation strategy was employed, whereby thresholds for each criterion were iteratively optimised. A sensible value for the maximum-area overlap feature (Equation 2) can be deduced based on the maximum surface area of possible surfaces generated from the bulk PCM lattice. This defines the largest possible area of vector overlap achieved during slab docking. For example for ice-I_h, the upper limit value of $\vec{a}$ (4.5193 Å) × $\vec{b}$ (7.3595 Å) × 10 ≈ 330 Å^2 (where the multiplier by 10 ensures ample tolerance of surface size differences) was chosen for Equation 2, and was held fixed while the criteria for Equations (3)-(6), which by definition are assumed to adopt values close to zero, were allowed to vary. Tightening the parameters refers to applying stricter numerical thresholds for lattice vector length mismatch and angular deviation. The 'loose' and 'medium' thresholds were chosen empirically based on literature tolerance ranges observed in epitaxial lattice matching.[9]

This process is deemed as a classifier model, as the ultimate decision of whether a compound is a working or non-working NUC for a given PCM is recorded as a binary y = 0 (negative result, representing a non-working NUC) or y = 1 (positive result, representing a working NUC). To reach this overall decision for a given training set of pre-defined working or non-working NUCs, binary decisions were recorded for each slab docking scenario as defined by the set of four feature values at a given numerical tolerance limit. The solution then takes the form of a decision tree (shown in Figure 2.10), to determine the numerical values for each of the four features that correctly capture all working NUCs while at the same time excluding all non-working NUCs,

by adjusting the upper limit (maximum value) for each feature independently. Values for the features were then systematically tightened to essentially train the model to make binary decisions, as the number of interface models that fit the increasingly stringent conditions fall for the non-working NUCs. Finally, NUCs that remained inconsistent with the optimised thresholds were flagged as outliers. These cases highlight the limits of purely geometrical features, highlighting that additional factors, such as surface chemistry, structure rearrangements, defects *etc.*, that are omitted in this workflow can play significant roles.

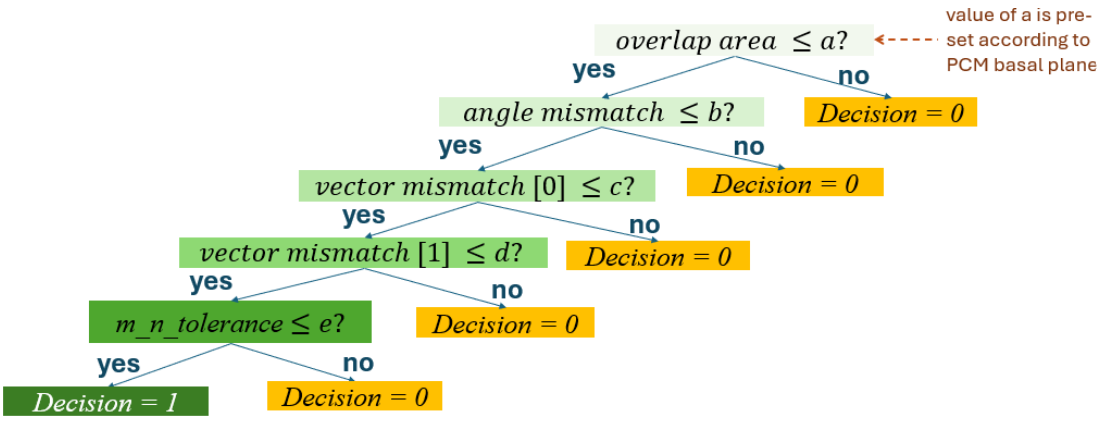


Figure 2.10. Binary decision making for prediction model training.

This optimisation process ensures that the chosen thresholds maximise the discriminatory power of the geometric features while remaining grounded in the experimental classifications of working and non-working NUCs. It also provides a bridge between the crystallographic basis of slab matching and the subsequent machine-learning framework developed in later Chapter 5.

One further note is warranted on the different functions undertaken by the maximum area overlap feature (Equation 2) and the *m_n*_tolerance (Equation 6). The former limits the absolute size of the interface unit cell that can be considered a valid match, while the latter defines the permissible ratio between the supercell scaling factors of the two slabs.The inclusion of both features mitigates a key limitation inherent in the purely geometric matching strategy,, whereby the potential to allow matches with increasingly large supercell cells increases the likelihood of finding a close geometric registry. However, as the interface unit cell grows, the chemical interactions that

reinforce epitaxial growth become less effective, undermining the physical relevance of the match. Thus, by constraining the tolerance on the scaling factors in Equation 6, the supercell generation process is restricted to manageable integer multiples that preserve physical plausibility while maintaining computational tractability.

References

1. D. Zagorac, H. Müller, S. Ruehl, J. Zagorac and S. Rehme, *Applied Crystallography*, 2019, **52**, 918-925.
2. H. E. Swanson, *Standard X-ray diffraction powder patterns*, US Department of Commerce, National Bureau of Standards, 1953.
3. G. E. Bacon, 1975.
4. D. E. Oliver, A. J. Bissell, X. Liu, C. C. Tang and C. R. Pulham, *CrystEngComm*, 2021, **23**, 700-706.
5. P. Niggli, *Krystallographische und strukturtheoretische grundbegriffe*, Akademische verlagsgesellschaft m. b. h., 1928.
6. J. M. Campbell, F. C. Meldrum and H. K. Christenson, *The Journal of Physical Chemistry C*, 2015, **119**, 1164-1169.
7. A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer and C. Hargus, *Journal of Physics: Condensed Matter*, 2017, **29**, 273002.
8. S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson and G. Ceder, *Computational Materials Science*, 2013, **68**, 314-319.
9. A. Zur and T. McGill, *Journal of applied physics*, 1984, **55**, 378-386.

Chapter 3
Data-Driven Prediction of Heterogeneous Ice Nucleators

3.1. Introduction

Super-cooling is a well-known phenomenon whereby a liquid exists in a metastable state below its freezing point. It can affect phase change materials, such as salt hydrates and ice banks, that are used for thermal energy storage[1,2]. For ice formation, which is the subject of this study, the control of crystal nucleation also has applications in cloud seeding and the production of artificial snow[3-5]. The pathway to mitigate against super-cooling is through the addition of heterogenous nucleating agents; these are insoluble materials that present solid templating surfaces to facilitate the nucleating process. This differentiates from freezing point depression, which is a thermodynamic effect driven through the addition of solutes.[6] Herein the focus is on the former: the onset of freezing induced by surface-facilitated nucleation. Particles that are known to be effective heterogeneous nucleators for ice span a broad range of materials,[5,7] from ionic salts,[8-10] to minerals,[11] carbonaceous materials[12] and organic matter.[13]

The available literature testifies to the complex nature of heterogeneous ice nucleation, with papers citing the importance of crystallographic similarities,[14,15] surface chemistries,[7,14,16-18] topologies,[19] water structuring effects and adsorption strengths,[20-22] and suspended solid or liquid particle sizes,[23-25] with experimental conditions that range from macroscopic observations on ice formation in the atmosphere[18,25], to those performed in ultra-clean materials chemistry labs[26]. Moreover, it is known that different nucleation pathways exist, depending on whether the nucleating particles are immersed in liquid water or suspended in a supersaturated vapour.[27, 28] The broadness of the field, combined with the variations in experiments conducted, results in variable reporting of 'good' or 'poor' ice-nucleating ability.

From a theoretical perspective, early reports[8, 29] often attributed ice-nucleating ability with a zero-lattice mismatch registry, i.e. a close similarity between the unit-cell dimensions of the nucleator and a particular face of the hexagonal phase of ice (ice-I_h), typically defined as the $\overrightarrow{ab}$ basal plane (0001).[30] While this has proven effective to

account for the well-known ice-nucleating properties of simple compounds such as AgI,[8] it is now widely accepted as an over-simplification,[31] not least because it does not take into account the chemistry of the nucleator/ice-forming interface. Computer simulation has made significant inroads into providing insights at the atomic level for heterogeneous[32] and homogeneous[33] ice nucleation. In particular, work by Michaelides *et al*. has highlighted the importance of understanding surface hydrophobicity, morphology and the variation in the adsorption energy landscape,[14, 17] as well as considering how ordered water molecule layers build up on a nucleating substrate,[5, 17] and how the density of the liquid water near the surface reduces.[22] More recently, machine-learning techniques trained on images of water-contact layers and the resulting prediction model (*IcePic*) have demonstrated success at accurately and rapidly predicting heterogeneous ice-nucleating behaviour.[24]

Herein a different approach has been taken that looks to take advantage of the wealth of potential heterogenous nucleators available through databases such as the International Centre for Diffraction Data (ICDD)[34] and the Inorganic Chemistry Structural Database (ICSD).[35] It is sought to generate a geometric docking model that assesses the quality of fit between ice-I_h/nucleator docked slabs cleaved along Miller index planes from the respective bulk crystal lattices. While this has similarities to the zero-lattice mismatch approach, it goes beyond the low-index planes to consider the docking of all interfaces (both nucleator and ice- I_h) described by the Miller indices up to (3,3,3). In this way, some of the structural complexity of the nucleation process by considering crystal morphology are addressed, where ice crystallites could seed on the faces, edges, corners, defects or other surface features of the nucleating crystal that could be described by these higher Miller-index planes. While this study focuses on ice nucleation in bulk water, the overarching goal is to build a generalisable high-throughput framework for predicting heterogenous nucleation agents for any given phase change material.

Given the variation in the literature regarding experimental set up, the study begins with establishing its own experimental benchmarking, *via* bulk water immersion experiments, on a set of ten widely known effective or poor nucleators for ice that could be readily sourced. A data-driven approach is then derived that is capable of

identifying new heterogeneous crystal nucleators using geometric interface matching, where the quality of fit between ice-I_h/nucleator docked slabs cleaved along Miller index planes from the respective bulk crystal lattices are assessed and ranked. By tightening a set of geometric criteria that describe the fit of the docked nucleator and ice-I_h cut planes, the number of matching slab interfaces that remain can act as a guide to the likely classification of a good or poor nucleator for ice-I_h. On this basis, the ICSD is then screened for several thousand simple metal oxide and halide structures. Testing the predicted outcomes for 22 compounds showed a 64% success rate. The procedure has also led to the discovery of four compounds, along with standard copper tubing, that can act as ice nucleators under immersion conditions.

3.2. Experimental benchmarking of known nucleators

3.2.1. Motivation and approach

To establish a physically meaningful and reproducible benchmark for classifying heterogeneous ice nucleation behaviour, a series of immersion freezing experiments using a well-characterised set of nucleators are conducted. This step is crucial not only for validating the slab-matching prediction model developed in Chapter 2 but also for defining a decision threshold that would enable binary classification of working versus non-working nucleators in a reproducible and chemically interpretable way.

The benchmarking set consisted of ten inorganic compounds, selected based on prior reports in the literature regarding their performance as ice nucleators, as well as their availability and stability under experimental conditions. The compounds chosen for their known effective behavior are MnO,[36] FeO (Wüstite),[36] AgI,[30, 37] Cu_2O,[38] AgCl[37], CuO[39] and SiO_2 (quartz).[27] For the poor nucleating agents, BaF_2,[40] $CaCO_3$ (calcite),[41] and $Al(OH)_3$ (gibbsite).[42] are chosen. The selected materials span a broad spectrum of reported nucleation efficiencies and include examples with both strong and weak templating capabilities. This diversity allows the experimental benchmark to capture a wide behavioural range and supported the development of a generalisable classification boundary.

3.2.2. Sample preparation and characterisation

All compounds were used in powder form with no surface functionalisation or post-treatment, to reflect the native crystallographic features relevant to immersion nucleation. To confirm the crystallographic identity and purity of nucleator samples, powder X-ray diffraction (PXRD) was performed on all purchased or synthesised compounds prior to use in nucleation experiments. PXRD patterns were collected at ambient temperature using a Bruker D2-Phaser instrument operating at 30 kV and 10 mA, using Cu-Kα radiation (λ = 1.5418 Å), with a step size of 0.03° and a step time of 0.5 s, over an angular range of 5-60° 2θ. Where needed, samples were lightly ground in a mortar and pestle and loaded without any further treatment onto a flat-plate Si-wafer sample holder.

The experimental diffraction patterns were compared against reference data obtained from the International Centre for Diffraction Data (ICDD) and the Inorganic Crystal Structure Database (ICSD). As an example, Figure 3.1 shows the PXRD pattern of NiO collected in this study (red), overlaid with the ICDD reference pattern for NiO (card no. 00-044-1159, blue). The excellent agreement in peak positions and relative intensities confirms that the laboratory sample is phase-pure NiO and crystallographically consistent with the reference structure used in the computational slab-matching model.

This protocol was repeated for all nucleator compounds employed in this thesis, ensuring that the experimental materials tested correspond to the correct polymorphs and unit cell parameters of the crystallographic entries used in the computational workflow. The 10 compounds with nucleating abilities known used for benchmarking, their polymorph/mineral names, and associated ICDD reference codes are found in Table 3.1.

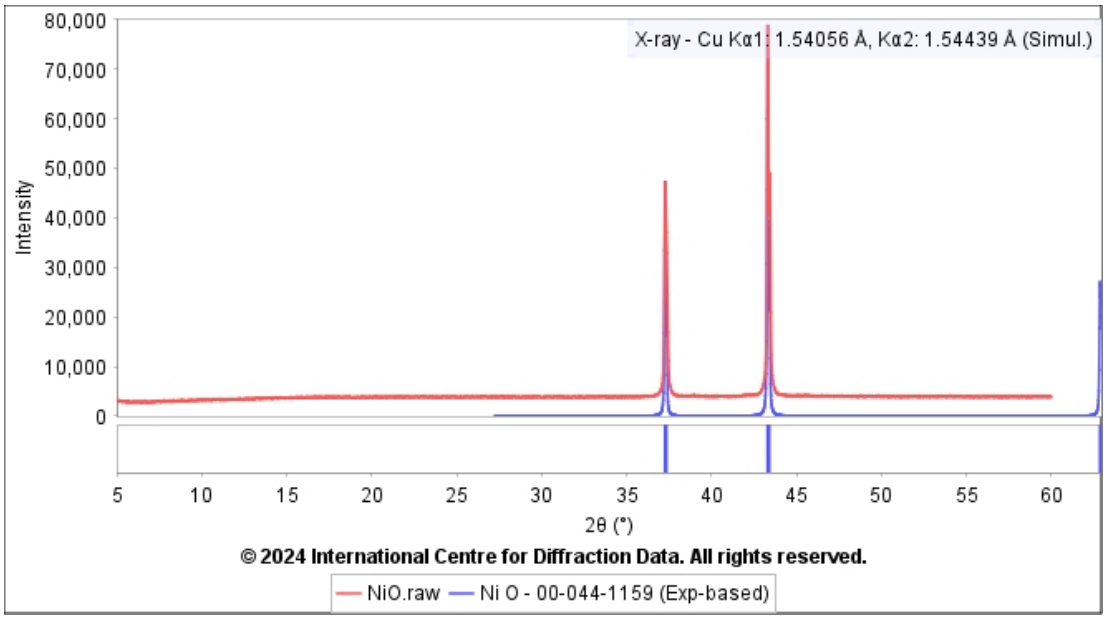


Figure 3.1. PXRD pattern of NiO.

Table 3.1. 10 compounds with nucleating abilities known used for benchmarking, their polymorph/mineral names, and associated ICDD reference codes.

Compound	ICDD code	Polymorph/mineral name
FeO	00-006-0615	Würstite
AgI	00-009-0374	β-/Iodargyrite
Cu_2O	01-078-2076	Cuprite
AgCl	00-006-0480	Form I/Chlorargyrite
CuO	00-041-0254	Tenorite
SiO_2	00-033-1161	Quartz
BaF_2	00-004-0452	Frankdicksonite
$CaCO_3$	00-005-0586	Calcite
MnO	01-075-6876	Manganosite
$Al(OH)_3$	01-070-2038	Gibbsite

3.2.3. Immersion freezing protocol and instruments

For immersion experiments, nucleation samples were prepared in screw-top borosilicate glass vials. Prior to use, all vials were rigorously cleaned to minimise contamination, first by washing in hot soapy water and then by repeatedly rinsing with ultra-pure water (18.2 MΩ·cm) passed through a Millex (33 mm) sterile filter unit fitted with a Millipore Express PES membrane. This ensured that no residual particulates or surfactants remained that could interfere with nucleation measurements.

For each experiment, 1 wt% of lightly ground nucleator powder was weighed into the bottom of the vial, onto which 10 mL of ultra-pure filtered water was added. To form a protective barrier against both airborne contamination and evaporation, a 1 mL overlayer of silicone oil was carefully pipetted onto the water surface. An RS PRO Type K thermocouple (exposed junction, 0.2 mm diameter), also pre-washed in ultra-pure water, was then positioned through the oil layer such that the junction was immersed in the aqueous phase (see Figure 3.2). This arrangement minimised the possibility of spurious nucleation events arising from the thermocouple tip itself while ensuring accurate temperature monitoring.

The vials were subjected to thermal cycling between +20 °C and -20 °C in a Polar Bear Plus instrument, using a controlled cooling and heating rate of 0.5 °C min^{-1}. Each sample underwent four freeze–thaw cycles to account for cycle-to-cycle variability. The efficacy of nucleation was quantified by recording the freezing onset temperature, defined as the point of initial exothermic release. A higher onset temperature (i.e. reduced subcooling) was interpreted as indicative of more effective nucleation. In the absence of any nucleating agent, the ultra-pure water reproducibly froze at -12 ± 3 °C under these conditions. All experiments were performed in triplicate, providing a basis for reproducibility.

It is important to note that this protocol captures the behaviour of the most active nucleation site within the sample, rather than an average across all particles present. Variability in particle dispersion, surface exposure, or local impurities can therefore influence the exact freezing onset recorded. Despite this statistical limitation, when experimental parameters such as particle loading, cooling rate, and sample handling are tightly controlled, the method provides a reproducible and meaningful comparative measure of nucleator efficacy across different materials.

Figure 3.2. Image of screw-top borosilicate glass vial, containing 1%wt nucleating agent (on the bottom), 10 ml of ultra-pure, filtered water, and 1 ml layer of silicone oil on the top, into which sits the thermocouple.

3.2.4. Results and establishment of decision threshold

Figure 3.3 shows the observed freezing onset temperatures for all ten compounds. All measurements are in line with expectations of effective or poor ice nucleation capability according to literature precedents.[10,19] Effective nucleators such as AgI, MnO, FeO (Wüstite), Cu_2O, CuO, and SiO_2 consistently triggered freezing events above -4 °C, while poor nucleators such as BaF_2, $Al(OH)_3$ (gibbsite), and $CaCO_3$ (calcite) exhibited freezing below -5 °C in most cases.

Ice nucleation onset temperatures are known to vary significantly, even for well-documented compounds like AgI,[10, 43-45] owing to the inherent stochastic nature of the nucleation process and the variability of different experimental protocols. The set-up affords sufficient reliability to differentiate between effective nucleators (e.g. AgI, Cu_2O) and weak or inactive ones (e.g. $Al(OH)_3$, BaF_2). This is sufficient for the means of this method, as the aim is merely to define a boundary temperature to delineate between the two behaviours. According to the accuracy limitations afforded by the Polar Bear apparatus and the sample preparations (1 wt% solid loading in 10 mL ultra-pure water), the boundary temperature is set to -4 °C; this temperature distinction will

be used to experimentally classify all further compounds as either a good or poor ice nucleator under these immersion conditions. In the absence of a nucleator, the experimental set-up achieves reliable sub-cooling to -12 ± 3 °C, as shown in Figure 3.4 of a sample polar bear temperature cycling data.

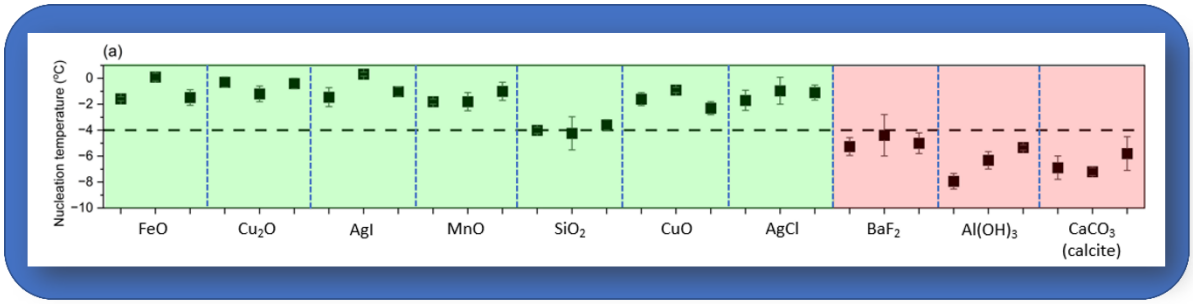


Figure 3.3. Freezing onset temperatures for the 10 benchmark nucleators. Error bars represent standard deviation across three independent trials. Green bars indicate materials classified as effective; red bars indicate poor nucleators (green: ≥10 matches; red: <10 matches).

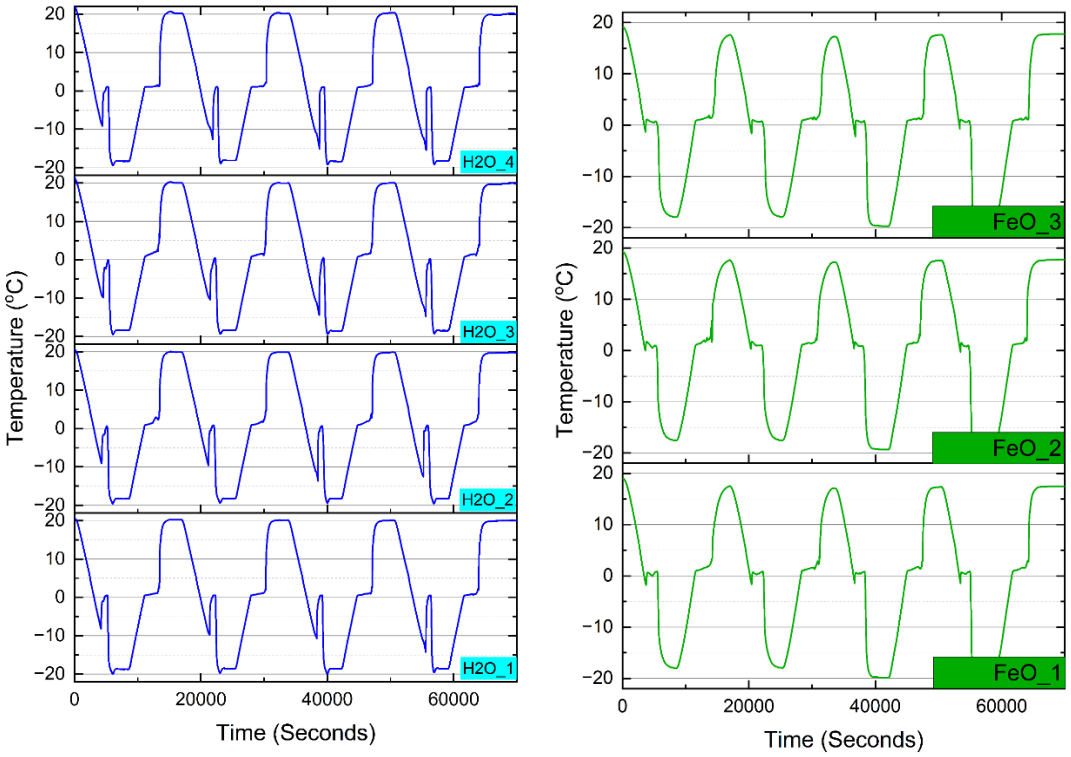


Figure 3.4. Temperature cycling data obtained for **(a)** ultra-pure, filtered water (four independent runs), showing reliable subcooling to -12 ± 3 °C; **(b)** FeO (three independent runs), in presence of 1% FeO.

It is fully acknowledged that e.g. nucleation chamber experiments[46] and drop-freezing assays[47] would yield far more reliable nucleation temperatures than what we have achieved here. This boundary is admittedly conservative: for example, some materials classified as 'poor' may show moderate activity under different conditions (e.g., droplet freezing or cloud chamber experiments). However, for the purposes of this thesis which focuses on interface geometry rather than broader physico-chemical influences, it provides a consistent framework for subsequent validation.

3.3. Establishing Geometric Matching Criteria

The prediction framework described in Chapter 2 relies on five geometric features to quantify the degree of crystallographic compatibility between the surface slabs of a phase-change material (PCM) and those of a potential nucleator (NUC). These features, i.e. maximum area overlap, angle mismatch, supercell vector mismatch ($\vec{u}$ and

$\vec{v}$ directions), and $m_n_$tolerance, were used to evaluate how well the two surfaces could geometrically dock in a lattice-matched configuration.

While the detailed methodology for calculating these features and aligning slabs has been described previously, this section outlines the practical criteria developed to distinguish between working and non-working nucleators based on their resulting match statistics. The goal was to determine a set of tolerance thresholds that could robustly differentiate between materials that consistently nucleate ice under experimental conditions and those that do not.

3.3.1. Descriptor overview

As a recap. each ice-NUC interface model generated in the workflow was evaluated using the five following geometric features:

Maximum area overlap: the total interfacial area shared between aligned slabs

Angle mismatch: angular deviation between slab lattice vectors

Supercell vector mismatch ($\vec{u}$ and $\vec{v}$): percent deviation in lattice vector lengths between the ice and NUC surfaces along two directions

***m_n_*tolerance**: a penalising factor to avoid excessive surface scaling via large supercell multipliers

The specific equations and reduction scheme are provided in Chapter 2, Section 2.3.

3.3.2. Threshold tuning: Loose to tight criteria

An overview of the general workflow, which was constructed in Python 3, underpinned by ASE[48] and Pymatgen[49] was presented in Chapter 2. With an experimental data set thus defined, the crystallographic information files (CIFs) for the ten compounds discussed in the previous section were used as input models for the interface-matching process. Throughout this work, crystal surfaces were referred by using Miller indices (hk(i)l), which define specific cut planes in the unit cell.

To define a working classification boundary, the numerical thresholds for these features were progressively tightened and the number of matching interface models

each known nucleator returned at each threshold level were evaluated. The three tolerance regimes tested are summarised in Table 3.2. Note that a sensible value for the maximum-area overlap descriptor can be deduced based on the maximum surface area of possible surfaces generated from the bulk ice-I_h lattice. This is the largest possible area of vector overlap achieved during slab docking. For this, the upper limit value of $\vec{a}$ (4.5193 Å) × $\vec{b}$ (7.3595 Å) × 10 ≈ 330 Å^2 (where the multiplier by 10 ensures ample tolerance of surface size differences) was chosen, and was held fixed while the thresholds for other features which by definition are assumed to adopt values close to zero, were allowed to vary. Tightening the parameters refers to applying stricter numerical thresholds for lattice vector length mismatch and angular deviation. The 'loose' and 'medium' thresholds were chosen empirically based on literature tolerance ranges observed in epitaxial lattice matching.[50]

Table 3.2. Descriptor thresholds under loose, medium, and tight matching criteria.

maximum-area overlap (Å^2)	angle mismatch	vector mismatch $\vec{u}$ direction	vector mismatch $\vec{v}$ direction	m_n_tolerance
330	0.2	0.2	0.2	0.1
330	0.1	0.01	0.01	0.1
330	0.1	0.01	0.01	0.01

At each level, the number of valid interface models per NUC was counted, and the separation between known good and poor nucleators was assessed. Figure 3.5 shows this comparison, with results grouped according to experimental nucleation performance (as described in Chapter 3, Section 3.2).

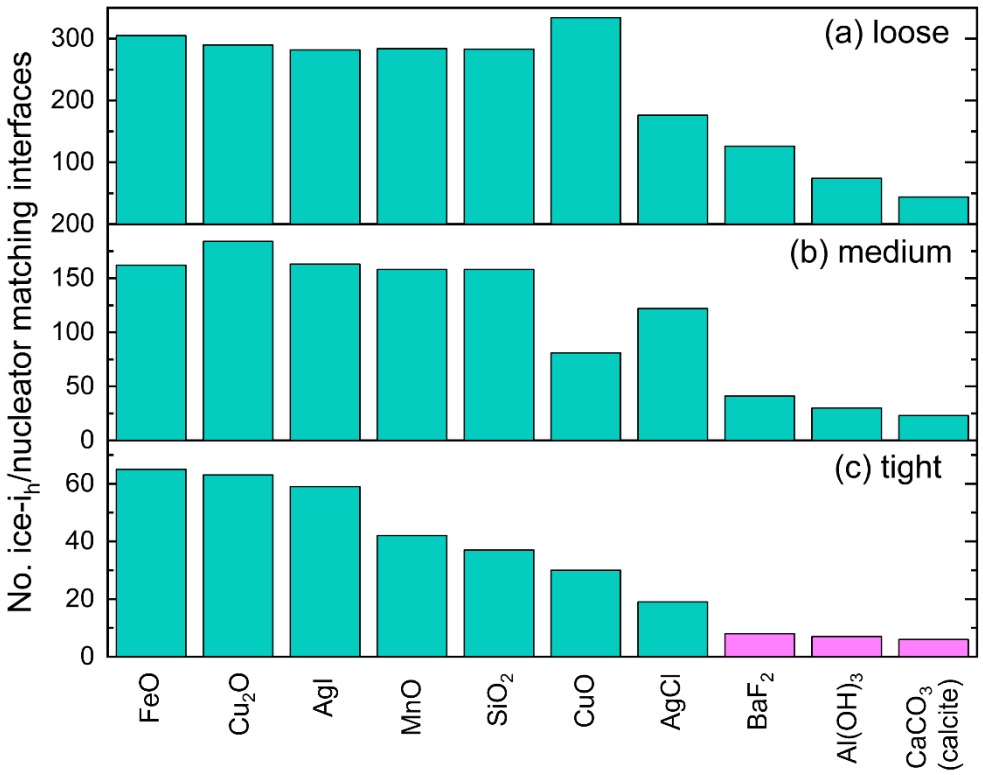


Figure 3.5. Interface-matching results for ice-I_h/nucleator pairs based on progressively stricter geometric thresholds: **(a)** loose, **(b)** medium, and **(c)** tight tolerance levels. Compounds shown in pink return less than 10 slab matching interfaces.

The results clearly shows that under tight tolerance values, effective nucleators retained $\geqslant$10 matching interface models, while poor nucleators dropped below this number. Tightening the parameters any further results in a substantial and

comprehensive loss in matching interface models; thus these fit criteria represent a probability boundary to differentiate between predicted good and poor ice nucleation behaviour. The Miller indices of the matching interface models are shown in Table 3.3, along with in-plane vector mismatches (along $\vec{u}$ *and* $\vec{v}$ directions)

The results clearly shows that under tight tolerance values, effective nucleators retained $\geqslant 10$ matching interface models, while poor nucleators dropped below this number. Tightening the parameters any further results in a substantial and comprehensive loss in matching interface models; thus these fit criteria represent a probability boundary to differentiate between predicted good and poor ice nucleation behaviour.

Table 3.3. Results from geometric slab matching using the tight geometry constraints. Basal (0, 0, 0, 1), primary (0, 1, $\bar{1}$, 0), (1, 0, $\bar{1}$, 0) and secondary (1, 1, $\bar{2}$, 0) prism ice faces are highlighted.

Nucleator (NUC)	Vector mismatch $\vec{u}$	Vector mismatch $\vec{v}$	(hk(i)l) of NUC slab	(hkil) of ice-I_h slab
MnO	0.008606	0.016237	(0, 0, 1)	(0, 0, 0, 1)
	0.008606	0.016237	(0, 0, 1)	(0, 1, $\bar{1}$, 0)
	0.005214	0.016237	(0, 0, 1)	(0, 1, $\bar{1}$, 2)
	0.008606	0.016237	(0, 0, 1)	(0, 2, $\bar{2}$, 1)
	0.004264	0.016237	(0, 0, 1)	(0, 2, $\bar{2}$, 3)
	0.008606	0.036978	(0, 0, 1)	(1, 3, $\bar{4}$, 0)
	0.008606	0.036978	(0, 0, 1)	(3, 1, $\bar{4}$, 0)
	0.003364	0.016237	(0, 1, 1)	(0, 0, 0, 1)
	0.004320	0.016237	(0, 1, 1)	(0, 1, $\bar{1}$, 0)
	0.007246	0.072499	(0, 1, 1)	(0, 1, $\bar{1}$, 1)
	0.004409	0.016237	(0, 1, 1)	(0, 1, $\bar{1}$, 2)
	0.003281	0.016237	(0, 1, 1)	(0, 1, $\bar{1}$, 3)
	0.006389	0.016237	(0, 1, 1)	(0, 2, $\bar{2}$, 1)
	0.006387	0.016237	(0, 1, 1)	(0, 2, $\bar{2}$, 3)
	0.000378	0.016237	(0, 1, 2)	(0, 0, 0, 1)
	0.000378	0.016237	(0, 1, 2)	(0, 1, $\bar{1}$, 0)
	0.007267	0.016237	(0, 1, 3)	(0, 0, 0, 1)
	0.001799	0.016237	(0, 1, 3)	(0, 1, $\bar{1}$, 3)
	0.000378	0.016237	(0, 2, 1)	(0, 0, 0, 1)
	0.002674	0.016237	(0, 2, 1)	(0, 1, $\bar{1}$, 0)
	0.006267	0.016237	(0, 2, 3)	(0, 0, 0, 1)
	0.009483	0.016237	(0, 2, 3)	(0, 1, $\bar{1}$, 2)

	0.001799	0.016237	(0, 3, 1)	(0, 0, 0, 1)
	0.007262	0.016237	(0, 3, 1)	(0, 1, $\overline{1}$, 3)
	0.006267	0.016237	(0, 3, 2)	(0, 0, 0, 1)
	0.009483	0.016237	(0, 3, 2)	(0, 1, $\overline{1}$, 2)
	0.008920	0.046361	(1, 1, 2)	(0, 1, $\overline{1}$, 0)
	0.002510	0.074700	(1, 1, 2)	(0, 1, $\overline{1}$, 2)
	0.007483	0.046360	(1, 1, 2)	(1, 1, $\overline{2}$, 0)
	0.003157	0.016237	(1, 1, 2)	(1, 1, $\overline{2}$, 1)
	0.008920	0.046361	(1, 1, 2)	(1, 3, $\overline{4}$, 0)
	0.008920	0.046361	(1, 1, 2)	(3, 1, $\overline{4}$, 0)
	0.005329	0.051687	(1, 1, 3)	(1, 2, $\overline{3}$, 0)
	0.005329	0.051687	(1, 1, 3)	(2, 1, $\overline{3}$, 0)
	0.005329	0.051687	(1, 1, 3)	(2, 3, $\overline{5}$, 0)
	0.005329	0.051687	(1, 1, 3)	(3, 2, $\overline{5}$, 0)
	0.008920	0.046360	(1, 2, 1)	(0, 1, $\overline{1}$, 0)
	0.003157	0.016237	(1, 2, 1)	(1, 3, $\overline{4}$, 0)
	0.005329	0.051687	(1, 3, 1)	(1, 2, $\overline{3}$, 0)
	0.005329	0.051687	(1, 3, 1)	(2, 1, $\overline{3}$, 0)
	0.005329	0.051687	(1, 3, 1)	(2, 3, $\overline{5}$, 0)
	0.005329	0.051687	(1, 3, 1)	(3, 2, $\overline{5}$, 0)
FeO	0.046993	0.039302	(0, 0, 1)	(0, 0, 0, 1)
	0.046993	0.024151	(0, 0, 1)	(0, 1, $\overline{1}$, 0)
	0.046993	0.033408	(0, 0, 1)	(0, 1, $\overline{1}$, 1)
	0.046993	0.068985	(0, 0, 1)	(0, 1, $\overline{1}$, 2)
	0.046993	0.060454	(0, 0, 1)	(0, 2, $\overline{2}$, 1)
	0.046993	0.060454	(0, 0, 1)	(0, 2, $\overline{2}$, 3)
	0.008566	0.008079	(0, 0, 1)	(1, 1, $\overline{2}$, 1)
	0.046993	0.037502	(0, 1, 1)	(0, 0, 0, 1)
	0.046993	0.034549	(0, 1, 1)	(0, 1, $\overline{1}$, 0)
	0.053589	0.059878	(0, 1, 1)	(0, 1, $\overline{1}$, 2)
	0.046993	0.054372	(0, 1, 1)	(1, 1, $\overline{2}$, 3)
	0.018807	0.024618	(0, 1, 1)	(1, 2, $\overline{3}$, 0)
	0.018807	0.024618	(0, 1, 1)	(2, 1, $\overline{3}$, 0)
	0.018807	0.024618	(0, 1, 1)	(2, 3, $\overline{5}$, 0)
	0.018807	0.024618	(0, 1, 1)	(3, 2, $\overline{5}$, 0)
	0.080608	0.069961	(0, 1, 2)	(0, 0, 0, 1)
	0.046993	0.046890	(0, 1, 2)	(0, 1, $\overline{1}$, 0)
	0.046993	0.045766	(0, 1, 2)	(0, 1, $\overline{1}$, 2)
	0.046993	0.045766	(0, 1, 2)	(0, 1, $\overline{1}$, 3)
	0.046993	0.043967	(0, 1, 3)	(0, 0, 0, 1)
	0.046993	0.057518	(0, 1, 3)	(0, 1, $\overline{1}$, 0)
	0.004557	0.002173	(0, 1, 3)	(0, 1, $\overline{1}$, 2)
	0.080608	0.074671	(0, 1, 3)	(1, 2, $\overline{3}$, 0)

	0.080608	0.074671	(0, 1, 3)	(2, 1, $\bar{3}$, 0)
	0.080608	0.074671	(0, 1, 3)	(2, 3, $\bar{5}$, 0)
	0.080608	0.074671	(0, 1, 3)	(3, 2, $\bar{5}$, 0)
	0.080608	0.069961	(0, 2, 1)	(0, 0, 0, 1)
	0.046993	0.046890	(0, 2, 1)	(0, 1, $\bar{1}$, 0)
	0.046993	0.045766	(0, 2, 1)	(0, 1, $\bar{1}$, 2)
	0.046993	0.042478	(0, 2, 1)	(0, 1, $\bar{1}$, 3)
	0.046993	0.055037	(0, 2, 3)	(0, 1, $\bar{1}$, 0)
	0.002592	0.008079	(0, 2, 3)	(1, 1, $\bar{2}$, 2)
	0.046993	0.055037	(0, 2, 3)	(1, 3, $\bar{4}$, 0)
	0.046993	0.055037	(0, 2, 3)	(3, 1, $\bar{4}$, 0)
	0.046993	0.043967	(0, 3, 1)	(0, 1, $\bar{1}$, 0)
	0.046993	0.055037	(0, 3, 2)	(0, 1, $\bar{1}$, 0)
	0.002592	0.008079	(0, 3, 2)	(1, 1, $\bar{2}$, 2)
	0.046993	0.055037	(0, 3, 2)	(1, 3, $\bar{4}$, 0)
	0.046993	0.055037	(0, 3, 2)	(3, 1, $\bar{4}$, 0)
	0.046993	0.037502	(1, 1, 0)	(0, 0, 0, 1)
	0.053589	0.059878	(1, 1, 0)	(0, 1, $\bar{1}$, 2)
	0.046993	0.054372	(1, 1, 0)	(1, 1, $\bar{2}$, 3)
	0.018807	0.024618	(1, 1, 0)	(1, 2, $\bar{3}$, 0)
	0.018807	0.024618	(1, 1, 0)	(2, 1, $\bar{3}$, 0)
	0.018807	0.024618	(1, 1, 0)	(2, 3, $\bar{5}$, 0)
	0.018807	0.024618	(1, 1, 0)	(3, 2, $\bar{5}$, 0)
	0.018075	0.010816	(1, 1, 1)	(0, 1, $\bar{1}$, 2)
	0.046993	0.048254	(1, 1, 2)	(0, 0, 0, 1)
	0.013647	0.010816	(1, 1, 2)	(0, 1, $\bar{1}$, 0)
	0.013647	0.027341	(1, 1, 2)	(1, 1, $\bar{2}$, 0)
	0.046993	0.057911	(1, 1, 2)	(1, 1, $\bar{2}$, 2)
	0.013647	0.018807	(1, 1, 2)	(1, 2, $\bar{3}$, 0)
	0.013647	0.018807	(1, 1, 2)	(2, 1, $\bar{3}$, 0)
	0.013647	0.018807	(1, 1, 2)	(2, 3, $\bar{5}$, 0)
	0.013647	0.018807	(1, 1, 2)	(3, 2, $\bar{5}$, 0)
	0.018807	0.024566	(1, 1, 3)	(0, 0, 0, 1)
	0.046993	0.048254	(1, 2, 1)	(0, 0, 0, 1)
	0.013647	0.027341	(1, 2, 1)	(1, 1, $\bar{2}$, 0)
	0.065494	0.051501	(1, 2, 2)	(0, 2, $\bar{2}$, 1)
	0.065494	0.051501	(1, 2, 2)	(0, 2, $\bar{2}$, 3)
	0.018807	0.024566	(1, 3, 1)	(0, 0, 0, 1)
	0.065494	0.051501	(2, 1, 2)	(0, 2, $\bar{2}$, 1)
	0.065494	0.051501	(2, 1, 2)	(0, 2, $\bar{2}$, 3)
	0.065494	0.051501	(2, 2, 1)	(0, 2, $\bar{2}$, 1)
	0.065494	0.051501	(2, 2, 1)	(0, 2, $\bar{2}$, 3)
β-AgI	0.016157	0.016157	(0, 0, 0, 1)	(0, 0, 0, 1)

	0.016157	0.080815	(0, 0, 0, 1)	(0, 1, $\overline{1}$, 0)
	0.016157	0.045739	(0, 0, 0, 1)	(0, 1, $\overline{1}$, 1)
	0.016157	0.012887	(0, 0, 0, 1)	(0, 1, $\overline{1}$, 2)
	0.016157	0.045739	(0, 1, $\overline{1}$, 0)	(0, 0, 0, 1)
	0.016157	0.045739	(0, 1, $\overline{1}$, 0)	(0, 1, $\overline{1}$, 0)
	0.016157	0.021369	(0, 1, $\overline{1}$, 0)	(0, 1, $\overline{1}$, 2)
	0.016157	0.013768	(0, 1, $\overline{1}$, 0)	(0, 2, $\overline{2}$, 1)
	0.016157	0.013768	(0, 1, $\overline{1}$, 0)	(0, 2, $\overline{2}$, 3)
	0.020503	0.022202	(0, 1, $\overline{1}$, 0)	(1, 1, $\overline{2}$, 0)
	0.040547	0.043916	(0, 1, $\overline{1}$, 0)	(1, 1, $\overline{2}$, 1)
	0.040547	0.0635	(0, 1, $\overline{1}$, 0)	(1, 1, $\overline{2}$, 2)
	0.040547	0.03268	(0, 1, $\overline{1}$, 0)	(1, 1, $\overline{2}$, 3)
	0.016157	0.010207	(0, 1, $\overline{1}$, 1)	(0, 1, $\overline{1}$, 0)
	0.016157	0.019546	(0, 1, $\overline{1}$, 1)	(0, 1, $\overline{1}$, 1)
	0.085676	0.082062	(0, 1, $\overline{1}$, 1)	(1, 1, $\overline{2}$, 3)
	0.016157	0.009021	(0, 1, $\overline{1}$, 2)	(0, 1, $\overline{1}$, 0)
	0.016157	0.015688	(0, 1, $\overline{1}$, 2)	(0, 1, $\overline{1}$, 1)
	0.016157	0.018198	(0, 1, $\overline{1}$, 2)	(0, 1, $\overline{1}$, 2)
	0.016157	0.018198	(0, 1, $\overline{1}$, 2)	(0, 1, $\overline{1}$, 3)
	0.0635	0.068305	(0, 1, $\overline{1}$, 2)	(1, 1, $\overline{2}$, 2)
	0.016157	0.009021	(0, 1, $\overline{1}$, 2)	(1, 3, $\overline{4}$, 0)
	0.016157	0.007485	(0, 2, $\overline{2}$, 1)	(0, 0, 0, 1)
	0.016157	0.035853	(0, 2, $\overline{2}$, 1)	(0, 1, $\overline{1}$, 2)
	0.016157	0.007485	(0, 2, $\overline{2}$, 3)	(0, 0, 0, 1)
	0.016157	0.035853	(0, 2, $\overline{2}$, 3)	(0, 1, $\overline{1}$, 2)
	0.040547	0.056021	(1, 1, $\overline{2}$, 0)	(0, 0, 0, 1)
	0.020503	0.022202	(1, 1, $\overline{2}$, 0)	(0, 1, $\overline{1}$, 0)
	0.020503	0.016157	(1, 1, $\overline{2}$, 0)	(1, 1, $\overline{2}$, 0)
	0.040547	0.033209	(1, 1, $\overline{2}$, 0)	(1, 1, $\overline{2}$, 1)
	0.020503	0.002154	(1, 1, $\overline{2}$, 0)	(1, 2, $\overline{3}$, 0)
	0.020503	0.022202	(1, 1, $\overline{2}$, 0)	(1, 3, $\overline{4}$, 0)
	0.020503	0.002154	(1, 1, $\overline{2}$, 0)	(2, 1, $\overline{3}$, 0)
	0.020503	0.002154	(1, 1, $\overline{2}$, 0)	(2, 3, $\overline{5}$, 0)
	0.020503	0.022202	(1, 1, $\overline{2}$, 0)	(3, 1, $\overline{4}$, 0)
	0.020503	0.002154	(1, 1, $\overline{2}$, 0)	(3, 2, $\overline{5}$, 0)
	0.016157	0.006997	(1, 1, $\overline{2}$, 1)	(0, 0, 0, 1)
	0.016157	0.026062	(1, 1, $\overline{2}$, 2)	(0, 0, 0, 1)
	0.016157	0.019315	(1, 1, $\overline{2}$, 2)	(1, 1, $\overline{2}$, 2)
	0.020503	0.016157	(1, 2, $\overline{3}$, 0)	(1, 2, $\overline{3}$, 0)
	0.020503	0.016157	(1, 2, $\overline{3}$, 0)	(2, 1, $\overline{3}$, 0)
	0.020503	0.016157	(1, 2, $\overline{3}$, 0)	(2, 3, $\overline{5}$, 0)
	0.020503	0.016157	(1, 2, $\overline{3}$, 0)	(3, 2, $\overline{5}$, 0)
	0.040547	0.016157	(1, 3, $\overline{4}$, 0)	(0, 0, 0, 1)

	0.040547	0.03268	(1, 3, $\overline{4}$, 0)	(1, 1, $\overline{2}$, 3)
	0.020503	0.016157	(2, 1, $\overline{3}$, 0)	(1, 2, $\overline{3}$, 0)
	0.020503	0.016157	(2, 1, $\overline{3}$, 0)	(2, 1, $\overline{3}$, 0)
	0.020503	0.016157	(2, 1, $\overline{3}$, 0)	(2, 3, $\overline{5}$, 0)
	0.020503	0.016157	(2, 1, $\overline{3}$, 0)	(3, 2, $\overline{5}$, 0)
	0.020503	0.016157	(2, 3, $\overline{5}$, 0)	(1, 2, $\overline{3}$, 0)
	0.020503	0.016157	(2, 3, $\overline{5}$, 0)	(2, 1, $\overline{3}$, 0)
	0.020503	0.016157	(2, 3, $\overline{5}$, 0)	(2, 3, $\overline{5}$, 0)
	0.020503	0.016157	(2, 3, $\overline{5}$, 0)	(3, 2, $\overline{5}$, 0)
	0.040547	0.016157	(3, 1, $\overline{4}$, 0)	(0, 0, 0, 1)
	0.040547	0.03268	(3, 1, $\overline{4}$, 0)	(1, 1, $\overline{2}$, 3)
	0.020503	0.016157	(3, 2, $\overline{4}$, 0)	(1, 2, $\overline{3}$, 0)
	0.020503	0.016157	(3, 2, $\overline{4}$, 0)	(2, 1, $\overline{3}$, 0)
	0.020503	0.016157	(3, 2, $\overline{4}$, 0)	(2, 3, $\overline{5}$, 0)
	0.020503	0.016157	(3, 2, $\overline{4}$, 0)	(3, 2, $\overline{5}$, 0)
Cu_2O	0.055512	0.090601	(0, 0, 1)	(0, 0, 0, 1)
	0.055512	0.043996	(0, 0, 1)	(0, 1, $\overline{1}$, 0)
	0.055512	0.024171	(0, 0, 1)	(0, 1, $\overline{1}$, 1)
	0.055512	0.059429	(0, 0, 1)	(0, 1, $\overline{1}$, 2)
	0.055512	0.0781	(0, 0, 1)	(0, 1, $\overline{1}$, 3)
	0.055512	0.050975	(0, 0, 1)	(0, 2, $\overline{2}$, 1)
	0.055512	0.050975	(0, 0, 1)	(0, 2, $\overline{2}$, 3)
	0.055512	0.028228	(0, 1, 1)	(0, 0, 0, 1)
	0.055512	0.093654	(0, 1, 1)	(0, 1, $\overline{1}$, 0)
	0.055512	0.086297	(0, 1, 1)	(0, 1, $\overline{1}$, 1)
	0.004586	0.001781	(0, 1, 1)	(1, 1, $\overline{2}$, 0)
	0.055512	0.044947	(0, 1, 1)	(1, 1, $\overline{2}$, 3)
	0.070949	0.078274	(0, 1, 2)	(0, 0, 0, 1)
	0.026852	0.034635	(0, 1, 2)	(0, 1, $\overline{1}$, 2)
	0.055512	0.033159	(0, 1, 2)	(0, 1, $\overline{1}$, 3)
	0.055512	0.048065	(0, 1, 3)	(0, 1, $\overline{1}$, 0)
	0.004422	0.006785	(0, 1, 3)	(0, 1, $\overline{1}$, 2)
	0.090601	0.082943	(0, 1, 3)	(1, 1, $\overline{2}$, 0)
	0.070949	0.082943	(0, 1, 3)	(1, 2, $\overline{3}$, 0)
	0.070949	0.082943	(0, 1, 3)	(2, 1, $\overline{3}$, 0)
	0.070949	0.082943	(0, 1, 3)	(2, 3, $\overline{5}$, 0)
	0.070949	0.082943	(0, 1, 3)	(3, 2, $\overline{5}$, 0)
	0.070949	0.078274	(0, 2, 1)	(0, 0, 0, 1)
	0.026852	0.034635	(0, 2, 1)	(0, 1, $\overline{1}$, 2)
	0.055512	0.074326	(0, 2, 3)	(0, 1, $\overline{1}$, 2)
	0.055512	0.074326	(0, 2, 3)	(1, 3, $\overline{4}$, 0)
	0.055512	0.045606	(0, 2, 3)	(3, 1, $\overline{4}$, 0)
	0.070949	0.082943	(0, 3, 1)	(1, 2, $\overline{3}$, 0)

	0.070949	0.082943	(0, 3, 1)	(2, 1, $\bar{3}$, 0)
	0.070949	0.082943	(0, 3, 1)	(2, 3, $\bar{5}$, 0)
	0.070949	0.082943	(0, 3, 1)	(3, 2, $\bar{5}$, 0)
	0.055512	0.045606	(0, 3, 2)	(0, 1, $\bar{1}$, 0)
	0.055512	0.074326	(0, 3, 2)	(0, 1, $\bar{1}$, 2)
	0.055512	0.045606	(0, 3, 2)	(1, 3, $\bar{4}$, 0)
	0.055512	0.045606	(0, 3, 2)	(3, 1, $\bar{4}$, 0)
	0.055512	0.028228	(1, 1, 0)	(0, 0, 0, 1)
	0.055512	0.093654	(1, 1, 0)	(0, 1, $\bar{1}$, 0)
	0.055512	0.086297	(1, 1, 0)	(0, 1, $\bar{1}$, 1)
	0.004586	0.001781	(1, 1, 0)	(1, 1, $\bar{2}$, 0)
	0.055512	0.044947	(1, 1, 0)	(1, 1, $\bar{2}$, 3)
	0.0097	0.0097	(1, 1, 1)	(0, 0, 0, 1)
	0.055512	0.068566	(1, 1, 2)	(0, 0, 0, 1)
	0.004586	0.001781	(1, 1, 2)	(0, 1, $\bar{1}$, 0)
	0.055512	0.048454	(1, 1, 2)	(1, 1, $\bar{2}$, 2)
	0.020164	0.022541	(1, 1, 2)	(1, 1, $\bar{2}$, 3)
	0.004586	0.0097	(1, 1, 2)	(1, 2, $\bar{3}$, 0)
	0.004586	0.0097	(1, 1, 2)	(2, 1, $\bar{3}$, 0)
	0.004586	0.0097	(1, 1, 2)	(2, 3, $\bar{5}$, 0)
	0.004586	0.0097	(1, 1, 2)	(3, 2, $\bar{5}$, 0)
	0.020164	0.022541	(1, 2, 1)	(1, 1, $\bar{2}$, 3)
	0.05597	0.059979	(1, 2, 2)	(0, 2, $\bar{2}$, 1)
	0.05597	0.059979	(1, 2, 2)	(0, 2, $\bar{2}$, 3)
	0.055512	0.067693	(1, 2, 3)	(1, 1, $\bar{2}$, 2)
	0.055512	0.067693	(1, 3, 2)	(1, 1, $\bar{2}$, 2)
	0.0097	0.004586	(2, 1, 1)	(1, 2, $\bar{3}$, 0)
	0.0097	0.004586	(2, 1, 1)	(2, 1, $\bar{3}$, 0)
	0.0097	0.004586	(2, 1, 1)	(2, 3, $\bar{5}$, 0)
	0.0097	0.004586	(2, 1, 1)	(3, 2, $\bar{5}$, 0)
	0.05597	0.059979	(2, 1, 2)	(0, 2, $\bar{2}$, 1)
	0.05597	0.059979	(2, 1, 2)	(0, 2, $\bar{2}$, 3)
	0.05597	0.059979	(2, 2, 1)	(0, 2, $\bar{2}$, 1)
	0.05597	0.059979	(2, 2, 1)	(0, 2, $\bar{2}$, 3)
AgCl	0.002333	0.002333	(1, 1, 1)	(0, 0, 0, 1)
	0.002333	0.019152	(0, 0, 1)	(1, 1, $\bar{2}$, 1)
	0.066111	0.079298	(0, 1, 1)	(0, 1, $\bar{1}$, 0)
	0.066111	0.07202	(0, 1, 1)	(1, 2, $\bar{3}$, 0)
	0.066111	0.07202	(0, 1, 1)	(2, 3, $\bar{5}$, 0)
	0.063135	0.058721	(1, 1, 2)	(0, 1, $\bar{1}$, 1)
	0.066111	0.079298	(0, 1, 1)	(1, 1, $\bar{2}$, 0)
	0.002333	0.017918	(0, 1, 1)	(0, 0, 0, 1)
	0.07202	0.071539	(0, 0, 1)	(0, 0, 0, 1)

	0.066111	0.054992	(1, 1, 2)	(0, 1, $\bar{1}$, 0)
	0.058721	0.063135	(2, 1, 1)	(0, 1, $\bar{1}$, 1)
CuO	0.036359	0.069884	(0, 0, 1)	(0, 1, $\bar{1}$, 0)
	0.036359	0.016217	(0, 0, 1)	(0, 0, 0, 1)
	0.036359	0.041924	(0, 0, 1)	(0, 1, $\bar{1}$, 1)
	0.036359	0.044336	(0, 0, 1)	(0, 1, $\bar{1}$, 2)
	0.036359	0.012057	(0, 0, 1)	(0, 1, $\bar{1}$, 3)
	0.036359	0.044939	(0, 0, 1)	(0, 2, $\bar{2}$, 1)
	0.036359	0.044939	(0, 0, 1)	(0, 2, $\bar{2}$, 3)
	0.069884	0.077016	(0, 0, 1)	(1, 1, $\bar{2}$, 0)
	0.087426	0.093092	(0, 0, 1)	(1, 1, $\bar{2}$, 1)
	0.018674	0.009755	(1, 0, 1)	(0, 1, $\bar{1}$, 0)
	0.042266	0.060243	(1, 0, 1)	(0, 0, 0, 1)
	0.018674	0.009755	(1, 0, 1)	(1, 1, $\bar{2}$, 0)
	0.042266	0.0568	(1, 0, 1)	(1, 1, $\bar{2}$, 2)
	0.042266	0.026174	(1, 0, 1)	(1, 1, $\bar{2}$, 3)
	0.018674	0.009755	(1, 0, 1)	(1, 3, $\bar{4}$, 0)
	0.018674	0.009755	(1, 0, 1)	(3, 1, $\bar{4}$, 0)
	0.069884	0.081133	(1, 0, 2)	(0, 1, $\bar{1}$, 0)
	0.061115	0.053355	(1, 0, 2)	(0, 1, $\bar{1}$, 2)
	0.069101	0.053355	(1, 0, 3)	(0, 1, $\bar{1}$, 0)
	0.009755	0.005144	(1, 0, 3)	(0, 0, 0, 1)
	0.005144	0.00925	(1, 0, 3)	(1, 1, $\bar{2}$, 2)
	0.069884	0.058696	(2, 0, 1)	(0, 1, $\bar{1}$, 0)
	0.000374	0.00844	(2, 0, 1)	(0, 0, 0, 1)
	0.044336	0.058696	(2, 0, 1)	(0, 1, $\bar{1}$, 2)
	0.069884	0.062706	(3, 0, 1)	(0, 1, $\bar{1}$, 0)
	0.069884	0.073719	(3, 0, 1)	(1, 1, $\bar{2}$, 0)
	0.062542	0.053355	(3, 0, 2)	(0, 0, 0, 1)
	0.009755	0.002179	(3, 0, 2)	(0, 1, $\bar{1}$, 1)
	0.069884	0.062542	(3, 0, 2)	(1, 1, $\bar{2}$, 0)
SiO_2	0.087377	0.087377	(0, 0, 0, 1)	(0, 0, 0, 1)
	0.087377	0.074747	(0, 0, 0, 1)	(0, 1, $\bar{1}$, 0)
	0.041024	0.048013	(0, 0, 0, 1)	(0, 1, $\bar{1}$, 2)
	0.087377	0.078646	(0, 0, 0, 1)	(0, 1, $\bar{1}$, 3)
	0.087377	0.065147	(0, 1, $\bar{1}$, 1)	(0, 1, $\bar{1}$, 0)
	0.087377	0.096542	(0, 1, $\bar{1}$, 1)	(0, 2, $\bar{2}$, 1)
	0.087377	0.096542	(0, 1, $\bar{1}$, 1)	(0, 2, $\bar{2}$, 3)
	0.087377	0.0799	(0, 1, $\bar{1}$, 2)	(0, 0, 0, 1)
	0.087377	0.086597	(0, 1, $\bar{1}$, 2)	(0, 1, $\bar{1}$, 0)
	0.087377	0.092742	(0, 1, $\bar{1}$, 2)	(0, 1, $\bar{1}$, 1)
	0.087377	0.061508	(0, 1, $\bar{1}$, 2)	(0, 1, $\bar{1}$, 2)
	0.087377	0.090447	(0, 1, $\bar{1}$, 2)	(0, 1, $\bar{1}$, 3)

	0.087377	0.086597	(0, 1, $\bar{1}$, 2)	(1, 3, $\bar{4}$, 0)
	0.087377	0.058028	(0, 1, $\bar{1}$, 3)	(0, 1, $\bar{1}$, 0)
	0.087377	0.089662	(0, 1, $\bar{1}$, 3)	(0, 2, $\bar{2}$, 1)
	0.087377	0.089662	(0, 1, $\bar{1}$, 3)	(0, 2, $\bar{2}$, 3)
	0.087377	0.070717	(0, 2, $\bar{2}$, 1)	(0, 1, $\bar{1}$, 1)
	0.087377	0.090683	(0, 2, $\bar{2}$, 3)	(0, 1, $\bar{1}$, 0)
	0.087377	0.088715	(0, 3, $\bar{3}$, 1)	(0, 1, $\bar{1}$, 0)
	0.058304	0.064798	(1, 1, $\bar{2}$, 0)	(0, 2, $\bar{2}$, 1)
	0.058304	0.064798	(1, 1, $\bar{2}$, 0)	(0, 2, $\bar{2}$, 3)
	0.087377	0.081834	(1, 1, $\bar{2}$, 0)	(1, 1, $\bar{2}$, 0)
	0.087377	0.0611	(1, 1, $\bar{2}$, 0)	(1, 1, $\bar{2}$, 2)
	0.058304	0.075801	(1, 1, $\bar{2}$, 1)	(0, 1, $\bar{1}$, 0)
	0.058304	0.068564	(1, 1, $\bar{2}$, 1)	(0, 1, $\bar{1}$, 1)
	0.087377	0.083254	(1, 1, $\bar{2}$, 1)	(1, 1, $\bar{2}$, 2)
	0.066702	0.076224	(1, 1, $\bar{2}$, 2)	(0, 0, 0, 1)
	0.058304	0.0544	(1, 1, $\bar{2}$, 2)	(0, 1, $\bar{1}$, 3)
	0.087377	0.086427	(1, 1, $\bar{2}$, 3)	(1, 1, $\bar{2}$, 3)
	0.041024	0.035885	(1, 2, $\bar{3}$, 0)	(0, 0, 0, 1)
	0.087377	0.079213	(1, 3, $\bar{4}$, 0)	(0, 0, 0, 1)
	0.042502	0.035885	(1, 3, $\bar{4}$, 0)	(1, 1, $\bar{2}$, 1)
	0.041024	0.035885	(2, 1, $\bar{3}$, 0)	(0, 0, 0, 1)
	0.041024	0.035885	(2, 3, $\bar{5}$, 0)	(0, 0, 0, 1)
	0.087377	0.079213	(3, 1, $\bar{4}$, 0)	(1, 1, $\bar{2}$, 1)
	0.041024	0.035885	(3, 2, $\bar{5}$, 0)	(0, 0, 0, 1)
BaF_2	0.029942	0.051898	(0, 0, 1)	(0, 1, $\bar{1}$, 1)
	0.015254	0.009938	(0, 1, 1)	(1, 1, $\bar{2}$, 2)
	0.029942	0.031783	(1, 1, 1)	(0, 1, $\bar{1}$, 0)
	0.015254	0.029942	(1, 1, 1)	(1, 1, $\bar{2}$, 2)
	0.029942	0.028902	(1, 1, 2)	(0, 0, 0, 1)
	0.029942	0.000506	(1, 1, 2)	(0, 1, $\bar{1}$, 2)
	0.029942	0.028902	(1, 2, 1)	(0, 0, 0, 1)
	0.029942	0.000506	(1, 2, 1)	(0, 1, $\bar{1}$, 2)
$Al(OH)_3$	0.076796	0.055188	(1, 0, 0)	(0, 1, $\bar{1}$, 2)
	0.039724	0.034498	(0, 0, 1)	(0, 1, $\bar{1}$, 0)
	0.005015	0.027389	(0, 0, 1)	(1, 1, $\bar{2}$, 2)
	0.076796	0.086632	(1, 0, 0)	(0, 1, $\bar{1}$, 1)
	0.055188	0.062748	(1, 0, 2)	(0, 1, $\bar{1}$, 2)
	0.076796	0.080446	(1, 0, 0)	(0, 1, $\bar{1}$, 0)
	0.077086	0.080679	(1, 0, 1)	(0, 0, 0, 1)
$CaCO_3$-calcite	0.004348	0.004348	(0, 0, 0, 1)	(0, 0, 0, 1)
	0.025855	0.03613	(0, 0, 0, 1)	(0, 1, $\bar{1}$, 0)
	0.043409	0.048044	(0, 0, 0, 1)	(0, 1, $\bar{1}$, 1)
	0.025855	0.03613	(0, 0, 0, 1)	(1, 3, $\bar{4}$, 0)

	0.025855	0.023871	(2, 3, $\bar{5}$, 0)	(0, 1, $\bar{1}$, 1)
	0.025855	0.023871	(3, 2, $\bar{5}$, 0)	(0, 1, $\bar{1}$, 1)

The geometric matches span a range of Miller indices for the nucleators, demonstrating that even moderately high-index planes can contribute to favourable interface docking, especially when surface scaling penalties are limited by the *m_n_*tolerance filter. Additionally, the recurrent pairing of ice faces such as (0,0,0,1) and (0,1,$\bar{1}$,0) with several nucleator surfaces underscores the role of ice basal and prismatic planes as common docking targets. This reinforces the idea that heterogeneous nucleation is not governed by a single ideal lattice match, but by a distribution of interfacial compatibilities across different surface combinations, which the current geometric model captures effectively.

Effective nucleators such as MnO, FeO, β-AgI, Cu_2O, AgCl, CuO, and SiO_2 consistently return a large and diverse set of matching interface models with ice-I_h, many of which involve low lattice mismatch values ($|\delta| < 0.02$) and frequent pairings with the basal and primary prismatic faces. These materials typically show multiple crystallographic orientations capable of meeting the tight geometric criteria, suggesting a variety of potential docking configurations that could seed ice formation. In contrast, the poor nucleators, i.e. BaF_2, $Al(OH)_3$ and $CaCO_3$ (calcite), exhibit a sparse and limited set of matches, often with larger mismatch values, fewer ice faces represented, and an absence of repeated low-mismatch pairings. This disparity implies that effective nucleators not only meet the strict geometric tolerance more frequently, but also do so across multiple surface combinations, increasing the probability of nucleation under variable growth conditions. Conversely, the geometric scarcity observed for poor nucleators provides a structural rationale for their inability to trigger freezing in the immersion experiments, supporting the $\geqslant$10-match threshold as a robust discriminant between working and non-working candidates.

The threshold of $\geqslant$10 matching interfaces under tight criteria was selected as the classification rule to differentiate predicted working versus non-working nucleators for further screening tasks. This value was used in the remainder of the study, including

the high-throughput screening of 3,500 compounds from the ICSD and the subsequent experimental testing of 22 selected candidates (see Section 3.4).

While this approach does not incorporate surface chemistry or solubility factors, it provided a computationally efficient and statistically grounded baseline for identifying promising nucleating candidates based purely on geometric compatibility. The effectiveness of this threshold will be evaluated in Section 3.4 against experimental outcomes.

3.4. High-Throughput Screening and Validation

3.4.1. Computational Screening Results

Following the establishment of the $\geqslant 10$ match threshold under tight geometric criteria (Section 3.3), the interface-matching workflow was applied in a high-throughput mode to identify new potential ice nucleators from the Inorganic Crystal Structure Database (ICSD). The search targeted simple binary metal halides (1,257 entries) and metal oxides (2,267 entries), chosen to maximise the likelihood of identifying materials with comparable lattice spacings to ice-I_h while avoiding excessive chemical complexity. Restricting the search to binary systems also ensured that selected candidates were more likely to be commercially available in high purity and low aqueous solubility, both essential for immersion-freezing testing.

The slab-matching process evaluated up to Miller indices (3,3,3) for both nucleator and ice-I_h surfaces, producing millions of unique docking configurations. Applying the tight criteria rapidly reduced this to a manageable subset, with ca. 7% of halides and ca. 3% of oxides exceeding the $\geqslant 10$-match classification threshold (Figure 3.6).

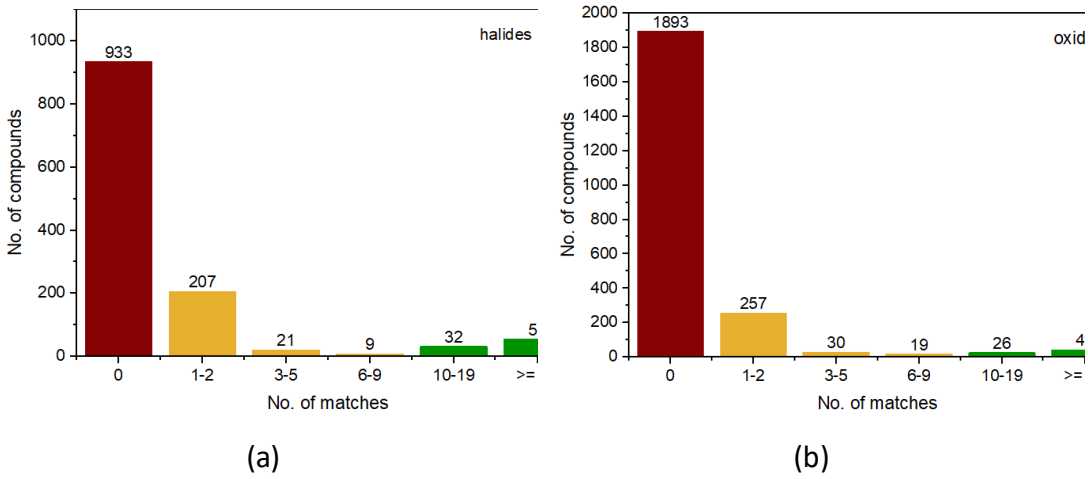


Figure 3.6. Distribution of the number of matching ice-I_h/nucleator interface models for **(a)** binary metal halides (n = 1,257) and **(b)** binary metal oxides (n = 2,267) identified in the ICSD, evaluated under the tight geometric criteria described in Section 3.3. The green bars indicate the ⩾10-match classification threshold used to predict effective nucleators. Compounds to the right of this threshold formed the candidate pool for experimental validation.

Figure 3.6 reveals that the vast majority of screened compounds returned very few or no matching interfaces with ice-I_h under tight tolerance conditions, with the modal bin for both oxides and halides being zero matches (1,893 and 933 compounds respectively). Only a small fraction exceeded the ⩾10-match classification threshold derived from the benchmarking study (Section 3.3): 84 oxides (3% of the set) and 87 halides (7% of the set). These above-threshold compounds occupy the long right-hand tail of each distribution, indicating rare but significant geometric compatibility across multiple crystallographic faces. A comprehensive list of working halides and oxides are available in Table 3.4.

Table 3.4. High throughput ICSD predictions on metal halides and oxides as ice nucleators.

Halides					
ICSD id	**compound**	**ICSD id**	**compound**	**ICSD id**	**compound**
22039	$EuCl_2$	65441	Hg_2Cl_2	419661	CrF_5
759787	MgF_2	62239	InBr	425449	InCl
414242	LiI	68819	PbI_2	48206	$NdCl_2$
71833	CuF_2	56539	AgCl	73165	PdF_2
140153	NF_3	56547	AgBr	78894	SnF_4
79678	γ-AgI	165593	KCl	51283	CaF_2
424210	GeF_2	68735	MnF_2	72354	Hg_2F_2
181148	NaCl	90997	TlF	130021	AlF_3
56552	β-AgI	73255	γ-CuCl	154994	PbF_2
1165	$OsCl_4$	2110	α-AgI	33752	$MnCl_2$
4070	UBr_3	135637	BiI_3	38074	$SrCl_2$
63535	TeI_4	202130	$PbCl_2$	56763	$CaBr_2$
68442	Ag_2F	202558	GeF_4	65479	$LaBr_3$
72424	$ThBr_4$	281551	MgI_2	78895	PbF_4
32705	ZrI_3	16142	SbF_3	124816	AgF_2
67500	$MnBr_2$	148399	YbF_2	130655	$ZrCl_4$
250363	$SbCl_5$	251213	$YbBr_2$	192942	$CrCl_2$
167473	TbF_3	86440	$CdCl_2$	23892	CrI_2
200460	U_2F_9	409450	$CuBr_2$	113757	Xe_2F_6
257262	CsCl	30708	YCl	138290	IrI3
14135	Cr_2F_5	41120	FeF_3	186512	$BaCl_2$
16112	TlF	80220	γ-CuI	1558	BiI
23903	$CrBr_2$	124813	BrF_5	14194	SnF_2
74729	LaF_3	9536	$CmCl_3$	14213	B_2Cl_4
2459	$ZnCl_2$	15101	SrI_2	20364	ZnF_2
4060	$PuCl_3$	25828	$CrCl_3$	65702	FeF_2
23147	$NdCl_3$	61348	RbI_3	252834	NbF_4
30052	$ZrCl_2$	70164	$SmCl_3$	260162	$TcBr_3$
56815	SmF_2				
Oxides					
ICSD id	**compound**	**ICSD id**	**compound**	**ICSD id**	**compound**
24729	BaO_2	176176	EuO	8491	PbO_2
40180	PbO	162039	MnO	27919	Cs_2O
4415	PtO_2	82233	FeO	37534	SnO_2
202407	PtO_2	128532	CoO	15028	V_6O_{13}
9863	MgO	23803	OsO_4	30361	BiO
15070	OsO_2	166362	MoO_3	38979	ZnO_2
1504	VO_2	134060	Na_2O_2	154021	ReO_2
167953	TiO_2	172174	Cu_2O	87942	CoO_2
163625	SrO	15620	BeO	27431	Ga_2O_3

166122	NiO	75198	NbO_2	140215	Tc_2O_7
99464	SeO_2	156423	Bi_2O_4	163628	CaO
15516	SnO	14124	HgO	1462	Ti_2O_3
80829	WO_2	167237	CrO_2	9206	Rh_2O_3
87184	KO_2	84574	GeO_2	66939	RuO_2
167268	Cr_2O_3	164007	Fe_2O_3	69094	CuO
171866	MnO_2	165720	CeO_2	142790	HfO_2
23722	MoO_2	172161	ZrO_2	161691	TeO_2
30005	Mn_3O_4	176081	Nb_2O_5	7840	Tb_3O_5
162843	ZnO	281041	Ag_2O	15798	V_2O_5

3.4.2. Experimental validation

The selection of candidates for experimental testing was drawn from this $\geqslant 10$ bin, ensuring that all tested materials satisfied the geometric criterion most predictive of successful nucleation. Within this subset, additional filters, such as low aqueous solubility, commercial availability, and diversity of match counts, were applied to produce a manageable set for laboratory validation. Including candidates from both the lower and upper ends of the $\geqslant 10$ range also allowed the robustness of the threshold rule to be evaluated across a gradient of predicted compatibility.

The 22 shortlisted candidates identified through this distribution were then subjected to immersion freezing experiments under the same conditions as the benchmarking set to assess the predictive accuracy of the geometric model. The selection strategy balanced three considerations:

- **Predicted classification diversity**: Seventeen predicted as effective nucleators, five as poor nucleators, allowing assessment of both true positive and true negative performance;
- **Chemical practicality**: Low solubility in water to ensure surfaces remained intact during testing; non-toxic or readily handled materials preferred;
- **Availability and representativeness**: Commercial availability from reliable suppliers and coverage of a broad range of match scores within each prediction class.

Experimental testing used the same immersion freezing protocol as for the benchmarking set (Section 3.2), with 1 wt% solid loading in 10 mL ultra-pure water,

a 1 mL silicone oil layer, and controlled cooling in the Polar Bear Plus apparatus. Freezing onset temperatures were measured over three independent runs per sample, and the -4 °C decision threshold (Section 3.2.4) was applied to classify the outcome. Table 3.5 summarises the 22 compounds used with their mineral names as well as their ICDD reference codes.

Table 3.5. 22 compounds with nucleating abilities known used for benchmarking, their polymorph/mineral names, and associated ICDD reference codes.

Compound	ICDD code	Polymorph/mineral name
NiO	00-044-1159	–
CuI	00-006-0246	γ-/Marshite
MgO	00-004-0829	Periclase
TiO_2	00-021-1276	Rutile
CoO	00-048-1719	–
PbO_2	00-041-1492	Plattnerite
SnO_2	00-041-1445	Cassiterite
$CaCO_3$	00-041-1475	Aragonite
Ti_2O_3	00-010-0063	Tistarite
CeO_2	00-004-0593	Cerianite
ZnO	04-004-4531	Zincite
MnO_2	01-090-9047	β-/Pyrolusite
Ag_2O	00-041-1104	–
Fe_2O_3	01-080-5405	Hematite
WO_3	01-083-0950	γ-
Al_2O_3	01-084-9871	γ-
Fe_3O_4	01-080-7683	Magnetite
Mn_2O_3	04-007-0856	α-/Bixyite
$PbBr_2$	04-005-4710	–
Bi_2O_3	00-041-1449	Bismite
Co_3O_4	00-042-1467	Guite
TiO_2	00-021-1272	Anatase

Figure 3.7 summarises the experimental results, with compounds ordered by descending model predicted match count, as shown in Figure 3.8. 14 of the 22 compounds were correctly categorised as working or non-working nucleators, yielding a 64% prediction success rate, while five were wrongly assigned (CoO, $CaCO_3$ (aragonite), TiO_2 (anatase), Mn_2O_3, and Fe_2O_3), and three were ambiguous (MgO, Fe_3O_4 and Co_3O_4). This level of agreement is comparable to or better than many

structure-activity screening approaches in materials science, especially given the simplicity of the geometric-only model.

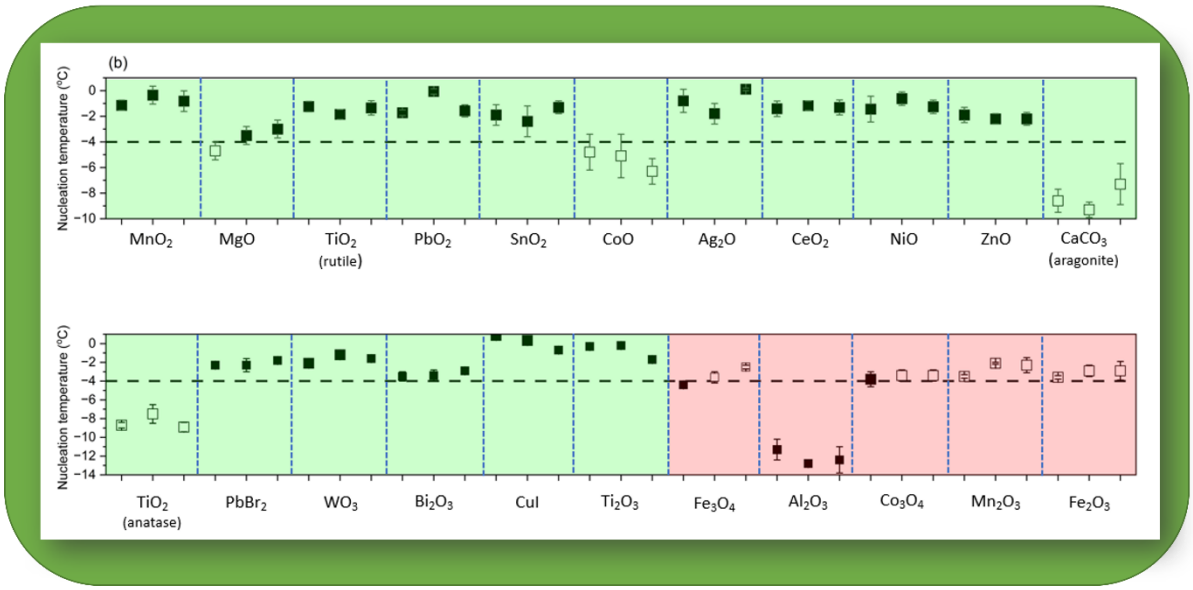


Figure 3.7. Experimental results for 22 test compounds predicted via the geometric interface-matching workflow. Compounds are ordered by descending number of predicted matching interfaces with ice-I_h (green: ≥10 matches; red: <10 matches). Filled symbols indicate agreement between prediction and experiment; empty symbols indicate disagreement.

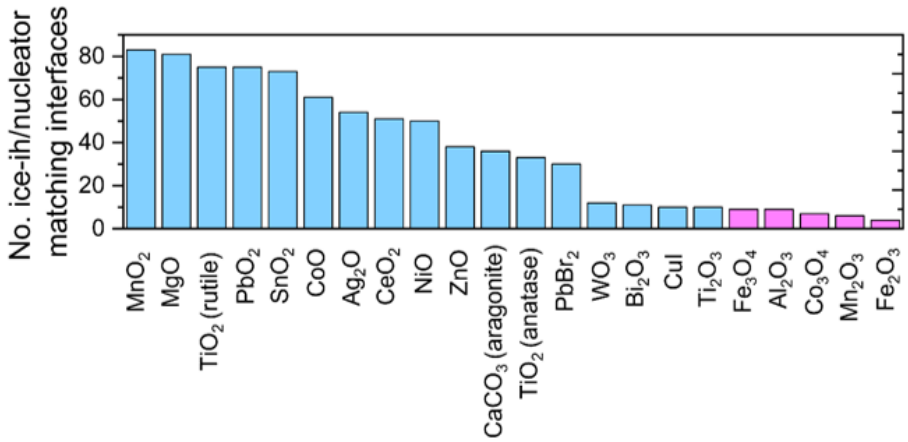


Figure 3.8. Interface-matching results for the 22 test compounds according to the tight tolerance criteria. Compounds shown in pink return less than 10 slab matching interfaces.

Of the correct predictions, some have previously been reported as ice nucleators in cloud-chamber experiments. The very early work by Fukuta[51] is highlighted here, who investigated the behavior of many metal salts under vapor-deposition conditions, including MnO_2, MgO, CoO, Ag_2O, NiO, ZnO, and $PbBr_2$. NiO was trialed for artificial snow production as far back as 1956,[52] while more recently lead oxide was highlighted as an anthropogenic climate modifier.[53] Early reports by Vonnegut[54] suggested that solid solutions of CuI with AgI improve the nucleating ability of the latter under immersion conditions, which was attributed to improvements in the lattice mismatch with the crystal structure of ice-I_h. MgO, TiO_2, AgI, Al_2O_3, and SiO_2 have attracted attention from materials scientists and computational modelers, who have studied the ice-nucleating abilities of individual faces. For instance, an experimental study on pristine MgO (100) and TiO_2 (100) (rutile) suggested the absence of a templating effect,[55] which matches the observation that these surfaces do not interface match with ice-I_h (see Table 3.6). The same paper reports that the (110) face of TiO_2 (rutile) supports the growth of cubic ice.[41] While the model specifically matched against ice-I_h, the modelling suggests that the (110) TiO_2 (rutile) surface only matches against one cleaved surface from ice-I_h; instead, the (001), (010), and (011) feature more heavily, geometrically matching with the basal and primary ice-I_h faces (see Table 3.6).

Molecular dynamics simulations performed on the (0001) and $(10\bar{1}0)$ faces of the beta-polymorph of AgI (as studied here) concluded that ice-I_h nucleates on the former, but not the latter.[56] This also matches the outcome of this study, with the (0001) face pairing with both the basal and primary ice-I_h faces, whereas none of the ice surface models interface with the $(10\bar{1}0)$ face of AgI.

For Al_2O_3, the gamma polymorph (as studied here, although with fairly low crystallinity) is known to be the oxide that forms on aluminum surfaces when exposed to the atmosphere.[57] Immersion studies have previously shown that α-Al_2O_3 is a more effective ice nucleator, and that any effect by γ-Al_2O_3 is weak.[58] This is borne out in this work, where γ-Al_2O_3 is classed as a weak nucleator, which was substantiated by the low number of matching interfaces it presents with ice-I_h (Figure 3.8); rerunning

the slab-matching process with the α-polymorph resulted in considerably more matching interfaces, which may be indicative of a higher ice nucleating ability (see Table 3.6).

SiO_2 $(10\,\bar{1}\,0)$ has been observed to template ice-I_h under immersion freezing conditions.[27] While this face was initially paired with multiple ice-I_h faces under the loose geometric criteria (Figure 3(a)), upon tightening the criteria other Miller planes of SiO_2 (notably (0001)) were found to match more closely with those of ice- I_h (see Table 3.6).

Thin films of the nanocomposite SnO_2 (cassiterite) /TiO_2 (anatase), spin coated with a Krytox grease lubricant have been shown to display anti-icing properties,[59] suggesting that ice templates poorly on this substrate. While geometric slab matching for SnO_2 and TiO_2 (anatase) returns high numbers of matching interfaces (Table 3.6), suggesting that both should template for ice- I_h, the latter was one of the five wrong assignments, as the bulk water immersion experiments showed that TiO_2 (anatase) does not nucleate ice (Figure 3.7). While there are likely to be many reasons why the SnO_2/TiO_2 nanocomposite inhibits ice growth, it would be of interest to explore if this could be attributed to TiO_2 (anatase) dominating the suppression of ice formation.

To the best of knowledge no prior reports have attributed ice nucleation properties to CeO_2, WO_3, Bi_2O_3 or Ti_2O_3, suggesting these could be new ice-I_h nucleators under immersion conditions.

While these findings are generally encouraging for a high throughput screening approach for identification of heterogeneous nucleating agents based purely on interface matching, it is important to note that predictions will miss any potential nucleators that do not fulfil the matching criterion, as indeed illustrated by Co_3O_4, Mn_2O_3 and Fe_2O_3. It also does not consider any surface chemistry effects (such as surface polarity), allow for any surface reactions or reconstructions which may promote nucleation, or rank the relative stabilities of the surface models. The data set highlights a number of false negatives (CoO, $CaCO_3$ (aragonite) and TiO_2 (anatase)). MgO displays ambiguous behaviour, despite presenting with one of the highest number of matching interface models with ice-I_h. One possible explanation is that the solid is

undergoing a surface reaction in water to form $Mg(OH)_2$ (brucite) or a hydrate. Repeating this slab matching approach with $Mg(OH)_2$ (ICSD code 34401), and $Mg(OH)_2{\cdot}2H_2O$ (ICSD code 118781) return a total of 45 and 4 slab matching interfaces, respectively, with ice-I_h down considerably from the 81 predicted interfaces with MgO. Without an in-depth experimental validation that explores ice nucleation onto defined nucleator surfaces, it remains unknown whether nucleation actually proceeds *via* these geometrically matching interfaces. The model's reliance on idealised interface matching introduces potential for false negatives, especially for materials where nucleation arises from less ordered or transient interface states. Nevertheless, it is noted that, reassuringly, the Miller index of the basal face of ice-I_h features heavily on the paired slabs list, as do the primary and secondary prismatic faces (see Table 3.6).

Finally, the data presented in Figures 3.7 and 3.8 illustrates an important point that a higher number of matching interface models does not correlate with a greater extent of suppression of subcooling; rather this demonstrates that the given crystal nucleator morphology (the edges, corners and potential defect sites captured by the Miller index planes up to hkl = (333)) are more likely to geometrically match with a corresponding Miller index place for ice-I_h.

In an attempt to show whether interface matching offers new information beyond the zero-lattice mismatch approach,[30] the mismatch registry parameter for the unit-cell parameters and for each of the matching interface models is also calculated (see Table 3.6). The analysis shows that, based on similarities of unit-cell dimensions between the nucleator and ice-I_h basal face, just seven of the 32 compounds that were explored experimentally would be correctly predicted as an effective or poor nucleator. For the matched interface models, while some of the effective nucleators do return low lattice mismatch values, the majority of the interface pairings do not. Overall, this suggests that interface matching offers a broader search criterion for potential nucleation behaviour than the lattice mismatch approach.

Table 3.6. Results of validated 22 compounds from geometric slab matching using the tight geometry constraints. Basal (0, 0, 0, 1), primary (0, 1, $\bar{1}$, 0), (1, 0, $\bar{1}$, 0) and secondary (1, 1, $\bar{2}$, 0) prism ice faces are highlighted.

Nucleator (NUC)	Vector mismatch $\vec{u}$	Vector mismatch $\vec{v}$	(hk(i)l) of NUC slabs	(hkil) indices of ice-ih slabs
Fe_2O_3	0.004068	0.004068	(0, 0, 0, 1)	(0, 0, 0, 1)
	0.062895	0.06379	(2, 3, $\bar{5}$, 0)	(0, 1, $\bar{1}$, 0)
	0.035322	0.02606	(2, 3, $\bar{5}$, 0)	(2, 3, $\bar{5}$, 3)
	0.062895	0.06379	(3, 2, $\bar{5}$, 0)	(0, 1, $\bar{1}$, 0)
Fe_3O_4	0.071134	0.06833	(0, 1, 0)	(0, 1, $\bar{1}$, 1)
MgO	0.003428	0.003428	(1, 1, 1)	(0, 0, 0, 1)
	0.008476	0.011245	(1, 1, 2)	(0, 1, $\bar{1}$, 0)
	0.067792	0.063731	(2, 2, 3)	(0, 1, $\bar{1}$, 0)
	0.067792	0.054672	(1, 2, 1)	(0, 1, $\bar{1}$, 0)
	0.008476	0.014859	(1, 3, 3)	(1, 1, $\bar{2}$, 2)
	0.003428	0.008476	(2, 1, 1)	(0, 1, $\bar{1}$, 0)
	0.039505	0.054672	(1, 1, 1)	(0, 2, $\bar{2}$, 1)
	0.008476	0.003428	(1, 1, 2)	(1, 2, $\bar{3}$, 0)
	0.008476	0.006899	(1, 3, 2)	(0, 1, $\bar{1}$, 0)
	0.067792	0.074885	(0, 1, 1)	(0, 1, $\bar{1}$, 0)
	0.067792	0.072173	(1, 0, 1)	(1, 1, $\bar{2}$, 0)
	0.057024	0.045906	(0, 0, 1)	(2, 1, $\bar{3}$, 3)
	0.008476	0.003428	(1, 1, 2)	(2, 1, $\bar{3}$, 3)
	0.04224	0.039961	(1, 2, 2)	(2, 1, $\bar{3}$, 3)
	0.067792	0.063134	(0, 0, 1)	(0, 1, $\bar{1}$, 0)
	0.04224	0.039961	(1, 2, 2)	(0, 1, $\bar{1}$, 0)
	0.067792	0.079434	(1, 1, 0)	(2, 1, $\bar{3}$, 1)
	0.008476	0.006899	(1, 3, 2)	(2, 3, $\bar{5}$, 3)
	0.067792	0.032011	(3, 2, 3)	(0, 1, $\bar{1}$, 0)
	0.067792	0.074885	(0, 1, 1)	(0, 1, $\bar{1}$, 1)
	0.008476	0.003428	(1, 1, 2)	(1, 1, $\bar{2}$, 0)
	0.057024	0.045906	(0, 0, 1)	(1, 2, $\bar{3}$, 0)
	0.008476	0.003428	(1, 1, 2)	(0, 1, $\bar{1}$, 2)
	0.008476	0.006899	(1, 2, 3)	(0, 1, $\bar{1}$, 0)
	0.04224	0.039961	(1, 2, 2)	(0, 0, 0, 1)
	0.04224	0.039961	(1, 2, 2)	(0, 3, $\bar{3}$, 2)
	0.04224	0.039961	(2, 1, 2)	(0, 0, 0, 1)
	0.067792	0.032011	(3, 2, 3)	(0, 0, 0, 1)

	0.067792	0.096172	(3, 2, 3)	(0, 3, $\bar{3}$, 2)
	0.067792	0.076421	(0, 0, 1)	(0, 0, 0, 1)
	0.067792	0.064082	(0, 0, 1)	(0, 3, $\bar{3}$, 2)
	0.017367	0.010854	(3, 1, 0)	(1, 0, $\bar{2}$, 1)
	0.017367	0.010854	(0, 1, 3)	(2, 3, $\bar{5}$, 1)
	0.017367	0.010854	(0, 1, 3)	(3, 2, $\bar{5}$, 1)
	0.003428	0.003428	(1, 1, 1)	(1, 1, $\bar{2}$, 0)
	0.008476	0.011245	(1, 1, 2)	(1, 3, $\bar{4}$, 0)
	0.067792	0.065602	(0, 1, 1)	(0, 0, 0, 1)
	0.067792	0.075784	(0, 1, 1)	(1, 2, $\bar{3}$, 0)
	0.067792	0.075784	(0, 1, 1)	(2, 1, $\bar{3}$, 3)
	0.008476	0.006899	(1, 3, 1)	(2, 1, $\bar{3}$, 3)
	0.008476	0.006899	(1, 1, 3)	(2, 1, $\bar{3}$, 3)
	0.067792	0.032011	(2, 2, 3)	(2, 1, $\bar{3}$, 3)
	0.067792	0.065602	(0, 1, 1)	(0, 1, $\bar{1}$, 0)
	0.008476	0.006899	(1, 1, 3)	(0, 1, $\bar{1}$, 0)
	0.008476	0.006899	(1, 3, 1)	(0, 1, $\bar{1}$, 0)
	0.016157	0.017525	(0, 2, 1)	(0, 1, $\bar{1}$, 0)
	0.016157	0.014554	(0, 1, 1)	(0, 1, $\bar{1}$, 2)
	0.016157	0.004652	(0, 1, 0)	(0, 1, $\bar{1}$, 0)
	0.016157	0.014554	(0, 1, 1)	(0, 1, $\bar{1}$, 1)
	0.016157	0.014554	(0, 1, 1)	(0, 1, $\bar{1}$, 0)
	0.016157	0.020039	(0, 0, 1)	(0, 3, $\bar{3}$, 2)
	0.016157	0.055449	(0, 1, 0)	(0, 0, 0, 1)
	0.016157	0.029086	(0, 1, 0)	(0, 3, $\bar{3}$, 2)
	0.016157	0.039598	(0, 1, 1)	(0, 1, $\bar{1}$, 3)
	0.016157	0.016659	(0, 1, 2)	(0, 1, $\bar{2}$, 0)
	0.016157	0.029086	(0, 1, 0)	(2, 1, $\bar{3}$, 3)
TiO_2	0.016157	0.016659	(0, 2, 1)	(0, 2, $\bar{2}$, 1)
	0.016157	0.070139	(0, 3, 1)	(0, 2, $\bar{2}$, 1)
	0.016157	0.008447	(0, 2, 3)	(0, 2, $\bar{2}$, 1)
	0.016157	0.002657	(0, 0, 1)	(1, 0, $\bar{1}$, 2)
	0.016157	0.063987	(0, 0, 1)	(0, 1, $\bar{1}$, 0)
	0.016157	0.016659	(0, 2, 1)	(0, 2, $\bar{2}$, 3)
	0.016157	0.016659	(0, 1, 2)	(0, 0, 0, 1)
	0.016157	0.070139	(0, 1, 3)	(0, 1, $\bar{1}$, 2)
	0.016157	0.039598	(0, 1, 0)	(0, 3, $\bar{3}$, 1)
	0.016157	0.01373	(0, 1, 1)	(2, 3, $\bar{5}$, 1)
	0.016157	0.01373	(0, 1, 1)	(3, 1, $\bar{4}$, 2)
	0.016157	0.008447	(0, 3, 2)	(0, 1, $\bar{1}$, 0)

	0.016157	0.016659	(0, 2, 1)	(1, 3, $\bar{4}$, 0)
	0.016157	0.036569	(0, 1, 0)	(0, 1, $\bar{1}$, 2)
	0.016157	0.016659	(0, 1, 2)	(2, 1, $\bar{3}$, 3)
	0.016157	0.020039	(0, 0, 1)	(0, 2, $\bar{2}$, 1)
	0.016157	0.016659	(1, 2, 0)	(0, 2, $\bar{2}$, 1)
	0.016157	0.01373	(0, 1, 1)	(0, 2, $\bar{2}$, 1)
	0.090673	0.093279	(1, 1, 0)	(2, 3, $\bar{5}$, 0)
	0.090673	0.093279	(1, 1, 0)	(1, 0, $\bar{1}$, 2)
	0.016157	0.008447	(0, 3, 2)	(0, 1, $\bar{1}$, 2)
	0.016157	0.014554	(0, 1, 1)	(2, 3, $\bar{5}$, 3)
	0.016157	0.01373	(0, 1, 1)	(2, 1, $\bar{3}$, 1)
	0.016157	0.070139	(0, 3, 1)	(0, 0, 0, 1)
NiO	0.088848	0.072758	(1, 1, 1)	(0, 0, 0, 1)
	0.00197	0.0028	(1, 2, 1)	(0, 1, $\bar{1}$, 0)
	0.00197	0.023003	(1, 1, 2)	(1, 2, $\bar{3}$, 0)
	0.00197	0.023003	(1, 2, 1)	(1, 2, $\bar{3}$, 0)
	0.017326	0.00126	(1, 2, 3)	(0, 1, $\bar{1}$, 0)
	0.017326	0.00126	(1, 3, 2)	(0, 1, $\bar{1}$, 0)
	0.019257	0.008116	(0, 0, 1)	(2, 1, $\bar{3}$, 3)
	0.017326	0.00126	(2, 1, 2)	(2, 1, $\bar{3}$, 3)
	0.019257	0.042379	(1, 0, 0)	(2, 1, $\bar{3}$, 3)
	0.017326	0.02079	(3, 2, 3)	(2, 1, $\bar{3}$, 3)
	0.019257	0.043148	(0, 0, 1)	(0, 1, $\bar{1}$, 0)
	0.017326	0.00126	(2, 1, 2)	(0, 1, $\bar{1}$, 0)
	0.019257	0.042379	(1, 0, 0)	(0, 1, $\bar{1}$, 0)
	0.017326	0.02079	(3, 2, 3)	(0, 1, $\bar{1}$, 0)
	0.019799	0.006635	(0, 1, 2)	(1, 1, $\bar{2}$, 2)
	0.00197	0.00126	(0, 1, 1)	(0, 1, $\bar{1}$, 1)
	0.00197	0.0028	(1, 1, 2)	(1, 1, $\bar{2}$, 0)
	0.00197	0.0028	(1, 2, 1)	(1, 1, $\bar{2}$, 0)
	0.019257	0.042379	(0, 0, 1)	(1, 2, $\bar{3}$, 0)
	0.061535	0.05628	(1, 2, 1)	(0, 1, $\bar{1}$, 2)
	0.019257	0.008116	(0, 0, 1)	(0, 3, $\bar{3}$, 2)
	0.017326	0.00126	(2, 1, 2)	(0, 3, $\bar{3}$, 2)
	0.017326	0.02079	(3, 2, 3)	(0, 3, $\bar{3}$, 2)
	0.00126	0.00126	(0, 0, 1)	(0, 0, 0, 1)
	0.061535	0.05628	(2, 1, 2)	(0, 0, 0, 1)
	0.017326	0.02079	(3, 2, 3)	(0, 0, 0, 1)
	0.00197	0.00126	(0, 1, 1)	(0, 0, 0, 1)
	0.088848	0.072758	(1, 1, 1)	(0, 2, $\bar{2}$, 1)

	0.061535	0.05628	(1, 1, 3)	(2, 3, $\bar{5}$, 0)
	0.088848	0.072758	(2, 2, 3)	(2, 3, $\bar{5}$, 0)
CoO	0.055667	0.068391	(1, 1, 2)	(0, 1, $\bar{1}$, 0)
	0.055667	0.048282	(1, 1, 2)	(1, 2, $\bar{3}$, 0)
	0.055667	0.067518	(1, 3, 2)	(0, 1, $\bar{1}$, 0)
	0.055667	0.02806	(0, 1, 1)	(1, 1, $\bar{2}$, 0)
	0.009535	0.009535	(1, 1, 1)	(0, 0, 0, 1)
	0.019996	0.022701	(1, 1, 2)	(2, 1, $\bar{3}$, 3)
	0.070773	0.078425	(0, 1, 2)	(0, 0, 0, 1)
	0.055667	0.093475	(0, 1, 1)	(0, 1, $\bar{1}$, 2)
	0.055667	0.02806	(0, 1, 1)	(0, 1, $\bar{1}$, 0)
	0.070773	0.083093	(1, 3, 1)	(1, 1, $\bar{2}$, 2)
	0.055667	0.086119	(0, 0, 1)	(0, 2, $\bar{2}$, 1)
	0.055667	0.048282	(1, 1, 2)	(1, 1, $\bar{2}$, 0)
	0.004421	0.009535	(1, 1, 2)	(2, 3, $\bar{5}$, 0)
	0.070773	0.083093	(1, 3, 1)	(0, 1, $\bar{1}$, 0)
	0.055667	0.077923	(0, 0, 1)	(0, 1, $\bar{1}$, 1)
	0.055796	0.060134	(1, 2, 2)	(0, 1, $\bar{1}$, 1)
	0.055667	0.07415	(3, 2, 3)	(0, 1, $\bar{1}$, 1)
	0.055667	0.07415	(3, 2, 3)	(2, 1, $\bar{3}$, 1)
	0.055667	0.024003	(1, 0, 0)	(2, 1, $\bar{3}$, 1)
	0.055667	0.067518	(2, 1, 2)	(2, 1, $\bar{3}$, 1)
	0.055667	0.067518	(2, 1, 2)	(0, 2, $\bar{2}$, 1)
	0.055667	0.045435	(3, 2, 3)	(0, 2, $\bar{2}$, 1)
PbO_2	0.097009	0.068525	(0, 1, 1)	(1, 2, $\bar{3}$, 0)
	0.097009	0.049962	(0, 1, 0)	(0, 1, $\bar{1}$, 0)
	0.000512	0.001428	(1, 3, 0)	(0, 0, 0, 1)
	0.047302	0.046295	(0, 2, 3)	(1, 3, $\bar{4}$, 0)
	0.097009	0.079473	(0, 1, 1)	(0, 1, $\bar{1}$, 0)
	0.097009	0.068525	(0, 1, 1)	(2, 1, $\bar{3}$, 3)
	0.005217	0.03427	(1, 1, 2)	(1, 1, $\bar{2}$, 2)
	0.000512	0.004232	(1, 2, 0)	(0, 1, $\bar{1}$, 0)
	0.000512	0.004232	(1, 2, 0)	(2, 3, $\bar{5}$, 3)
	0.047302	0.078721	(2, 3, 0)	(0, 0, 0, 1)
SnO_2	0.026869	0.017879	(1, 3, 0)	(0, 1, $\bar{1}$, 2)
	0.017879	0.010939	(2, 3, 1)	(0, 1, $\bar{1}$, 0)
	0.048418	0.030106	(0, 1, 1)	(0, 1, $\bar{1}$, 0)
	0.023185	0.032894	(1, 1, 1)	(1, 1, $\bar{2}$, 2)
	0.048418	0.025618	(0, 1, 0)	(1, 3, $\bar{4}$, 0)

	0.011542	0.013654	(2, 1, 3)	(1, 3, $\bar{4}$, 0)
	0.089501	0.062253	(2, 3, 3)	(1, 3, $\bar{4}$, 0)
	0.048418	0.049979	(1, 0, 0)	(1, 3, $\bar{4}$, 0)
	0.048418	0.027337	(1, 0, 1)	(1, 0, $\bar{1}$, 0)
	0.048418	0.043327	(0, 1, 1)	(0, 1, $\bar{1}$, 2)
CuI	0.09384	0.071846	(1, 2, 1)	(0, 1, $\bar{1}$, 1)
	0.09384	0.092361	(0, 1, 1)	(0, 1, $\bar{1}$, 0)
	0.052707	0.062576	(0, 0, 1)	(1, 1, $\bar{2}$, 1)
	0.052707	0.058826	(2, 0, 3)	(1, 2, $\bar{3}$, 0)
	0.052707	0.058826	(0, 2, 3)	(1, 2, $\bar{3}$, 0)
	0.052707	0.068689	(1, 1, 2)	(1, 1, $\bar{2}$, 2)
	0.09384	0.071846	(1, 1, 1)	(1, 1, $\bar{2}$, 0)
	0.052707	0.058826	(0, 2, 3)	(1, 0, $\bar{1}$, 2)
	0.052707	0.068689	(1, 2, 1)	(1, 1, $\bar{2}$, 2)
	0.052707	0.068689	(2, 1, 1)	(1, 1, $\bar{2}$, 2)
	0.09384	0.092361	(1, 0, 1)	(0, 1, $\bar{1}$, 0)
	0.052707	0.058826	(2, 1, 1)	(0, 1, $\bar{1}$, 1)
$CaCO_3$ (aragonite)	0.031247	0.020374	(1, 0, 1)	(1, 1, $\bar{2}$, 1)
	0.018057	0.00256	(1, 0, 0)	(1, 1, $\bar{2}$, 2)
	0.031247	0.018057	(3, 0, 1)	(0, 0, 0, 1)
	0.018057	0.00822	(0, 0, 1)	(1, 1, $\bar{2}$, 3)
	0.082836	0.075509	(3, 0, 2)	(1, 1, $\bar{2}$, 3)
	0.082836	0.083838	(2, 3, 1)	(1, 1, $\bar{2}$, 1)
	0.031247	0.018057	(1, 0, 3)	(2, 2, $\bar{4}$, 3)
	0.097901	0.085753	(0, 1, 1)	(2, 3, $\bar{5}$, 0)
	0.018057	0.011665	(1, 0, 0)	(0, 0, 0, 1)
Ti_2O_3	0.003357	0.003357	(0, 1, 2)	(0, 0, 0, 1)
	0.014723	0.003866	(1, 1, 1)	(0, 0, 0, 1)
	0.014723	0.012777	(1, 1, 1)	(0, 2, $\bar{2}$, 1)
	0.014723	0.003866	(1, 1, 2)	(1, 1, $\bar{2}$, 0)
	0.014723	0.003866	(1, 2, 1)	(1, 1, $\bar{2}$, 0)
	0.014723	0.003866	(1, 2, 2)	(1, 1, $\bar{2}$, 2)
	0.01275	0.014723	(2, 1, 2)	(1, 1, $\bar{2}$, 2)
ZnO	0.099597	0.057488	(0, 0, 0, 1)	(0, 0, 0, 1)
	0.079783	0.079537	(0, 1, $\bar{1}$, 2)	(1, 1, $\bar{2}$, 1)
	0.0875	0.088319	(1, 1, $\bar{2}$, 1)	(0, 1, $\bar{1}$, 2)
	0.099597	0.08926	(1, 2, $\bar{3}$, 0)	(1, 1, $\bar{2}$, 2)
	0.099597	0.08926	(2, 1, $\bar{3}$, 0)	(1, 1, $\bar{2}$, 2)
	0.042534	0.040193	(1, 0, $\bar{1}$, 2)	(1, 1, $\bar{2}$, 1)

MnO_2	0.026791	0.023417	(0, 0, 1)	(1, 1, $\bar{2}$, 2)
	0.026791	0.001684	(0, 1, 3)	(0, 1, $\bar{1}$, 0)
	0.071499	0.063529	(0, 1, 3)	(2, 3, $\bar{5}$, 3)
	0.026791	0.034409	(0, 2, 1)	(1, 1, $\bar{2}$, 0)
	0.071499	0.063529	(1, 0, 3)	(2, 3, $\bar{5}$, 3)
CeO_2	0.018872	0.01292	(0, 1, 2)	(1, 2, $\bar{3}$, 0)
	0.015676	0.004621	(0, 1, 1)	(0, 1, $\bar{1}$, 2)
	0.046956	0.033413	(0, 0, 1)	(0, 1, $\bar{1}$, 2)
	0.046956	0.033413	(0, 0, 1)	(1, 2, $\bar{3}$, 0)
Ag_2O	0.044391	0.040666	(0, 1, 1)	(0, 0, 0, 1)
	0.044391	0.068085	(0, 1, 1)	(0, 3, $\bar{3}$, 2)
	0.092998	0.044391	(1, 1, 3)	(0, 1, $\bar{1}$, 0)
	0.044391	0.033745	(3, 2, 3)	(0, 0, 0, 1)
WO_3	0.006863	0.008648	(0, 0, 1)	(0, 1, $\bar{1}$, 0)
Al_2O_3	0.053463	0.053463	(1, 0, $\bar{1}$, 1)	(0, 1, $\bar{1}$, 1)
	0.053463	0.010705	(1, 0, $\bar{1}$, 1)	(2, 1, $\bar{3}$, 1)

3.4.3. Baseline comparison

The manually tuned geometric matching model was evaluated on 22 experimentally characterised compounds. The resulting confusion matrix is shown in Figure 3.9. Two evaluation terms, accuracy and recall, are defined as follows:

Accuracy = (TP+TN)/(TP+TN+FP+FN)

(i.e. the overall proportion of correctly classed nucleators)

Recall = TP/(TP+ FN)

(i.e. the fraction of working nucleators successfully identified)

Of the 22 compounds, 14 were correctly classified, yielding an overall accuracy of 64%. The model demonstrated strong sensitivity (recall for working class) toward identifying working nucleators (11/13 correctly identified; 84.6%), but considerably lower specificity (recall for non-working class) for non-working compounds (3/9 correctly identified; 33.3%). This asymmetry indicates a tendency to over-predict nucleation activity, leading to a higher false positive rate.

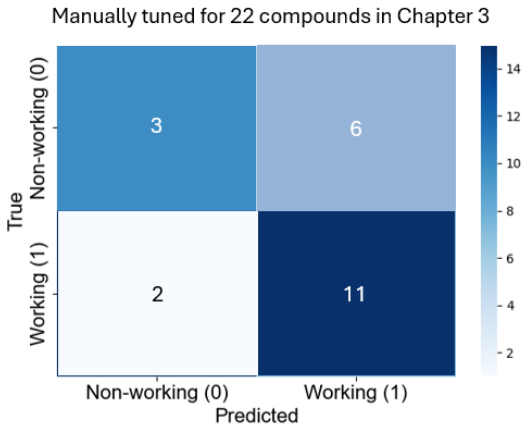


Figure 3.9. Random Forest (RF) predictions for manually tuned prediction model for the 22 compounds.

To account for class imbalance, balanced accuracy was also calculated as the mean of sensitivity and specificity, yielding 58.9%. This provides a more conservative estimate of overall discriminative performance.

To contextualise the predictive performance of the geometric matching model, two simple baseline classifiers were evaluated. For binary classification, random guessing yields an expected accuracy of 50%. The majority-class baseline was calculated by assigning all compounds to the most frequent experimental class (working nucleator), which comprised 15 of the 22 validation compounds, corresponding to an accuracy of 68.2%. The geometric interface-matching model achieved an accuracy of 64%, outperforming random guessing but slightly underperforming the majority-class baseline. This indicates that while geometric compatibility captures meaningful structural signal, class imbalance in the validation set limits the discriminative performance of the geometry-only model. Figure 3.10 shows the comparison of predictive performance for the geometric interface-matching model for ice against baseline classifiers.

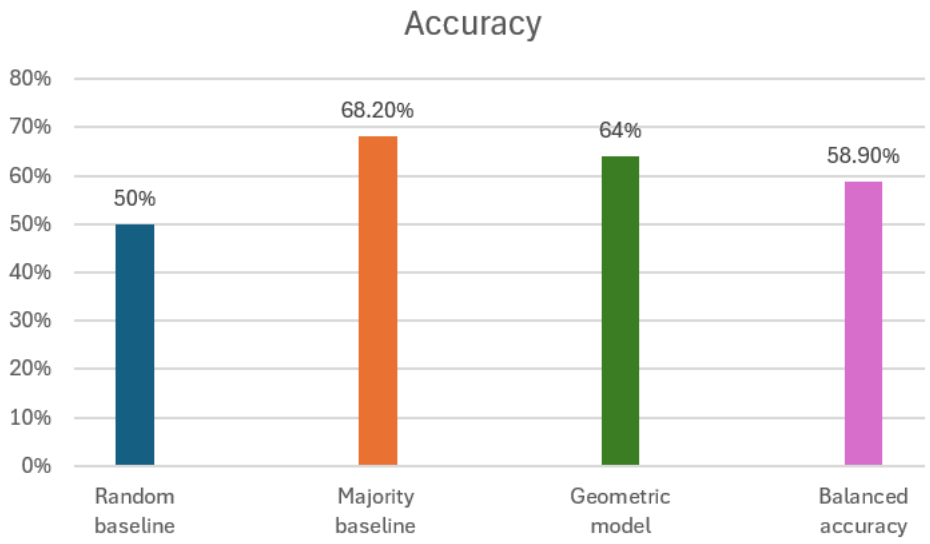


Figure 3.10. comparison of predictive performance for the geometric interface-matching model for ice against baseline classifiers. Random guessing yields 50% expected accuracy for binary classification. Majority-class guessing achieves 68.2% due to class imbalance (15/22 working nucleators). The geometric model achieves 64% overall accuracy and 58.9% balanced accuracy.

3.4.4. Copper tubing test

Given the high performance of copper oxides in the validation list, combined with the recent report that copper oxide nanoparticles act as ice nucleation sites,[59, 60] 'off-the-shelf' copper tubing for ice nucleation is tested. This follows given that a copper surface will readily oxidize in contact with air and water, and thus could promote the nucleation of ice in bulk water. This is particularly relevant given that copper is widely used in plumbing, HVAC (heating ventilation and cooling) applications, and electrical transmission power lines, for which the formation of ice contributes to burst water pipes and power outages.[61, 62] Sections of copper pipe (BS EN 1057 standard, 10 mm external diameter) were cut into 1 cm lengths, rinsed, and placed into sealed vials containing 20 mL of tap water (East Lothian supply, $CaCO_3$ concentration = 74.23 ppm). Samples were thermally cycled between –20 °C and +20 °C in the Polar Bear Plus Crystal apparatus, using a heating/cooling rate of 1 °C min^{-1}.

Each copper-containing sample was subjected to 25 freeze-thaw cycles, while blank samples containing only tap water (no tubing) were run in parallel as controls. Measurements were performed in triplicate. The large number of cycles was chosen to assess not only the nucleating potential of copper but also its reproducibility across repeated cycling, reflecting real-world conditions where nucleation sites may undergo repeated activation and deactivation.

The resulting data show that the copper tubing did indeed induce ice nucleation at -2.3 ± 0.2° C (compared to -10.3 ± 0.9°C in its absence). Moreover, the nucleation temperature increased over multiple cycles, potentially correlating with increased oxidation of the metal surface (see Figures 3.11 and 3.12).

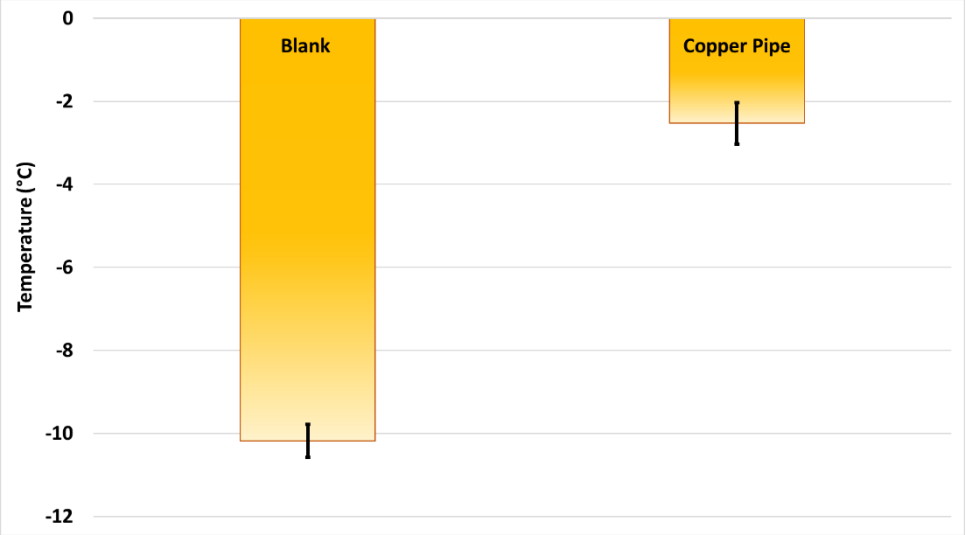


Figure 3.11. Ice nucleation data for tap water (74.23 ppm hardness as $CaCO_3$) in presence of BS EN 1057 copper tubing.

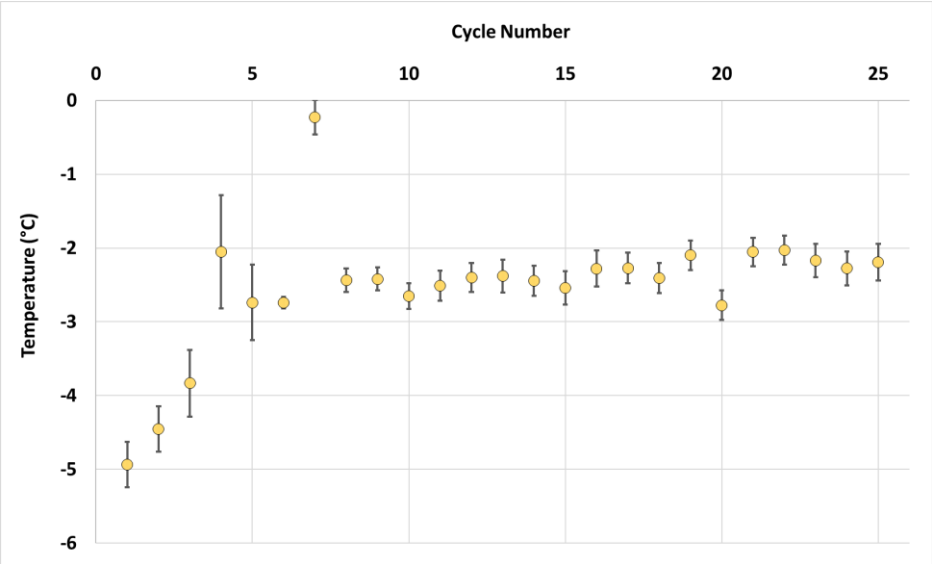


Figure 3.12. Evolution of ice nucleation temperature in the presence of copper pipe over repeated thermal cycles.

3.4.54. Summary

This combined computational-experimental workflow successfully reduced a pool of over 3,500 potential candidates to a manageable set for laboratory validation, resulting in the discovery of four new ice nucleators and confirming several others. The approach demonstrates the feasibility of geometric slab matching as a high-throughput screening tool, while also revealing its limitations and opportunities for integration with other descriptor sets.

3.5. Limitations and future aspects

3.5.1. Geometry-only assumptions

At the core of the method lies the assumption that crystallographic lattice matching between a PCM and a candidate nucleator is a primary driver of heterogeneous nucleation efficiency. This is supported by the strong separation observed between working and non-working benchmark nucleators when assessed under tight geometric criteria (Section 3.3), and by the correct prediction of several new ice nucleators.

However, lattice geometry represents only one facet of the nucleation problem. In reality, the ice-nucleator interface is influenced by a complex interplay of factors

beyond surface registry, including interfacial free energy, hydration layer structure, chemical bonding interactions, and defect-mediated nucleation pathways. A good geometric fit can therefore be a necessary but not sufficient condition for effective nucleation.

3.5.2. Missing surface chemistry

The geometric model treats nucleator surfaces as ideal, clean, and chemically inert lattices, an assumption that is rarely valid under experimental conditions. Many oxides and halides develop hydrated, hydroxylated, or otherwise chemically modified surface layers upon exposure to water. These surface layers can disrupt or enhance hydrogen-bond templating to the ice lattice, either suppressing nucleation despite geometric compatibility (false positives) or enabling it in the absence of strong lattice matches (false negatives).

For example, TiO_2 (anatase), predicted to be a strong nucleator based on its high match count, consistently failed to trigger freezing above -4 °C in immersion tests. The discrepancy may be attributed to the formation of hydrated surface layers that are structurally or energetically incompatible with ice nucleation. Conversely, Fe_3O_4 displayed nucleation activity near the decision threshold despite a relatively modest match count, suggesting that favourable surface chemistry, possibly involving hydroxyl-mediated hydrogen bonding, can partially compensate for suboptimal lattice alignment.

3.5.3. Role of polymorphs

The importance of accurate polymorph identification was underscored by cases such as $CaCO_3$, where different phases yield significantly different surface spacings and thus different predicted match counts. Experimental samples may also contain mixed polymorphs or undergo phase transformations during preparation or storage, further complicating predictive accuracy. This polymorph sensitivity reinforces the need for reliable structural characterisation prior to computational screening and suggests that future models could benefit from probabilistic weighting of polymorph likelihood under given synthesis or environmental conditions.

3.5.4. Real-world surface complexity

In nature and in application, nucleator surfaces are rarely atomically flat single-crystal planes. Instead, they present a landscape of terraces, steps, edges, defects, and microstructural heterogeneities. Surface roughness can expose multiple crystallographic orientations within a single particle, potentially increasing the number of effective docking configurations beyond that predicted for an ideal slab. In contrast, surface oxidation, contamination, or amorphisation can obscure or distort otherwise compatible lattice sites.

For instance, the observation that copper tubing in contact with tap water exhibited strong nucleation activity is likely linked to the in-situ growth of copper oxide/hydroxide surface layers with geometries favourable to ice nucleation. This dynamic surface evolution is not captured in the static geometric model, yet it is highly relevant for real-world performance.

3.5.5. The “false negative” dilemma

False negatives, materials predicted to be poor nucleators yet found experimentally to be effective, represent a particular challenge for model refinement. They expose gaps in the feature set that allow genuinely active surfaces to be overlooked. In this study, the occurrence of false negatives suggests that the current model’s strict reliance on basal/prismatic lattice registry may miss alternative nucleation mechanisms. For example, certain substrates may promote ice formation by templating less commonly considered crystallographic faces of ice Ih, such as secondary prism or pyramidal planes, rather than the basal or primary prism faces explicitly evaluated in this framework. Other mechanisms could include epitaxial growth mediated by adsorbed ions or molecules, or defect-driven pathways in which local surface geometry deviates significantly from the ideal bulk structure.

From a practical perspective, false negatives are more damaging than false positives in the context of discovery, as they eliminate promising candidates from further consideration. This motivates a more cautious approach to excluding low-match-count materials, and supports the development of models that integrate geometric, chemical, and energetic features.

3.5.6. Future directions

Several avenues could improve the predictive power of the prediction model, first and foremost by integration with machine learning, as explored in Chapter 5, geometric features can be further regulated in a supervised ML framework. Feature attribution tools such as SHAP could help identify the relative contribution of geometry vs chemistry to nucleation efficiency.

3.6. Conclusions

Herein, a high throughput data-driven workflow is presented for identifying potential heterogeneous nucleating agents, like ice, from structural chemistry databases, such as the ICSD. The model is based on assessing the quality of fit between ice-I_h and nucleator docked slabs, formed from cleaving along Miller index planes from the respective bulk crystal lattices. While this has similarities to the zero-lattice mismatch approach, it goes beyond the low-index planes to consider the docking of all interfaces described by the Miller indices up to (3,3,3). In this way, some of the structural complexity of the nucleation process is addressed by considering crystal morphology, where ice crystallites could seed on the faces, edges, corners, defects or other surface features of the nucleating crystal that could be described by these higher Miller-index planes.

Numerical tolerance limits for the docking model were derived from a training set of ten compounds. The bulk water freezing experiments were sufficiently reliable to classify each compound correctly as an effective or poor ice nucleator, based on expectations from the literature. Tightening the geometric matching criteria resulted in a fall in the number of matching interface models until differentiation between the two classes was obtained.

The trained model screened approximately 3,500 simple metal oxides and halides from the ICSD for predicted nucleation behaviour. Subsequent experimental measurements of 22 compounds showed a 64% prediction success rate, as defined by the freezing temperature boundary obtained from the experimental training set data. The workflow also identified four previously unreported ice nucleating agents (CeO_2, WO_3, Bi_2O_3, Ti_2O_3).

Given the high ice nucleating ability demonstrated for copper oxides, the nucleating ability of standard copper tubing is also tested immersed in samples of local tap water. This was also found to suppress sub-cooling, likely due to the build-up of copper oxides over the timescale of the experiment.

While the approach that has been taken here is undoubtedly simplistic, and does not account for many important aspects, such as surface chemistry effects, reactions and reconstructions, it nevertheless demonstrates an acceptable level of success to form the basis for a high throughput computational screening approach to locate potential heterogeneous nucleating agents.

References

1. S. Kiyabu, P. Girard and D. J. Siegel, *Journal of the American Chemical Society*, 2022, **144**, 21617-21627.
2. S. Hasnain, *Energy Conversion and Management*, 1998, **39**, 1139-1153.
3. O. Boucher, D. Randall, P. Artaxo, C. Bretherton, G. Feingold, P. Forster, V.-M. Kerminen, Y. Kondo, H. Liao and U. Lohmann, in *Climate Change 2013,* Cambridge University Press, 2013, pp. 571-657.
4. W. K. Tao, J. P. Chen, Z. Li, C. Wang and C. Zhang, *Reviews of Geophysics*, 2012, **50**.
5. B. Murray, D. O'sullivan, J. Atkinson and M. Webb, *Chemical Society Reviews*, 2012, **41**, 6519-6554.
6. G. Vali, P. J. DeMott, O. Möhler and T. F. Whale, *Atmos. Chem. Phys.*, 2015, **15**, 10263-10270.
7. Z. A. Kanji, L. A. Ladino, H. Wex, Y. Boose, M. Burkert-Kohn, D. J. Cziczo and M. Krämer, *Meteorological Monographs*, 2017, **58**, 1.1-1.33.
8. B. Vonnegut, *Journal of Applied Physics*, 1947, **18**, 593-595.
9. B. Vonnegut and H. Chessin, *Science*, 1971, **174**, 945-946.
10. C. Marcolli, B. Nagare, A. Welti and U. Lohmann, *Atmospheric Chemistry and Physics*, 2016, **16**, 8915-8937.
11. A. Kiselev, F. Bachmann, P. Pedevilla, S. J. Cox, A. Michaelides, D. Gerthsen and T. Leisner, *Science*, 2017, **355**, 367-371.
12. B. Friedman, G. Kulkarni, J. Beránek, A. Zelenyuk, J. A. Thornton and D. J. Cziczo, *Journal of Geophysical Research: Atmospheres*, 2011, **116**.
13. T. W. Wilson, L. A. Ladino, P. A. Alpert, M. N. Breckels, I. M. Brooks, J. Browse, S. M. Burrows, K. S. Carslaw, J. A. Huffman and C. Judd, *Nature*, 2015, **525**, 234-238.
14. M. Fitzner, G. C. Sosso, S. J. Cox and A. Michaelides, *Journal of the American Chemical Society*, 2015, **137**, 13658-13669.
15. C. Hoose and O. Möhler, *Atmospheric Chemistry and Physics*, 2012, **12**, 9817-9854.
16. P. Pedevilla, M. Fitzner and A. Michaelides, *Physical Review B*, 2017, **96**, 115441.
17. J. Carrasco, A. Hodgson and A. Michaelides, *Nature Materials*, 2012, **11**, 667-674.
18. T. Koop, B. Luo, A. Tsias and T. Peter, *Nature*, 2000, **406**, 611-614.
19. T. F. Whale, M. A. Holden, A. N. Kulak, Y.-Y. Kim, F. C. Meldrum, H. K. Christenson and B. J. Murray, *Physical Chemistry Chemical Physics*, 2017, **19**, 31186-31193.
20. G. C. Sosso, T. F. Whale, M. A. Holden, P. Pedevilla, B. J. Murray and A. Michaelides, *Chemical Science*, 2018, **9**, 8077-8088.
21. H. R. Pruppacher, J. D. Klett and P. K. Wang, *Microphysics of Clouds and Precipitation*, Taylor & Francis, 1998.
22. M. Fitzner, P. Pedevilla and A. Michaelides, *Nature Communications*, 2020, **11**, 4777.
23. J. Chen, Z. Wu, J. Chen, N. Reicher, X. Fang, Y. Rudich and M. Hu, *Atmospheric Chemistry and Physics*, 2021, **21**, 3491-3506.

24. M. B. Davies, M. Fitzner and A. Michaelides, *Proceedings of the National Academy of Sciences*, 2022, **119**, e2205347119.
25. G. Vali, *Journal of Atmospheric Sciences*, 1971, **28**, 402-409.
26. G. Vali, *Atmospheric Chemistry and Physics*, 2008, **8**, 5017-5031.
27. M. A. Holden, J. M. Campbell, F. C. Meldrum, B. J. Murray and H. K. Christenson, *Proceedings of the National Academy of Sciences*, 2021, **118**, e2022859118.
28. C. Marcolli, *Atmospheric Chemistry and Physics*, 2014, **14**, 2071-2104.
29. B. Mason, *Advances in Physics*, 1958, **7**, 221-234.
30. D. Turnbull and B. Vonnegut, *Industrial & Engineering Chemistry*, 1952, **44**, 1292-1298.
31. G. C. Sosso, J. Chen, S. J. Cox, M. Fitzner, P. Pedevilla, A. Zen and A. Michaelides, *Chemical Reviews*, 2016, **116**, 7078-7116.
32. M. Fitzner, G. C. Sosso, S. J. Cox and A. Michaelides, *Proceedings of the National Academy of Sciences*, 2019, **116**, 2009-2014.
33. T. Li, D. Donadio and G. Galli, *Nature Communications*, 2013, **4**, 1887.
34. S. N. Kabekkodu, A. Dosen and T. N. Blanton, *Powder Diffraction*, 2024, 1-13.
35. D. Zagorac, H. Müller, S. Ruehl, J. Zagorac and S. Rehme, *Journal of Applied Crystallography*, 2019, **52**, 918-925.
36. E. Chong, K. E. Marak, Y. Li and M. A. Freedman, *Physical Chemistry Chemical Physics*, 2021, **23**, 3565-3573.
37. W. G. Finnegan and S. K. Chai, *Journal of the Atmospheric Sciences*, 2003, **60**, 1723-1731.
38. C. X. Kronawitter, C. Riplinger, X. He, P. Zahl, E. A. Carter, P. Sutter and B. E. Koel, *Journal of the American Chemical Society*, 2014, **136**, 13283-13288.
39. X. ZHANG, X. MENG, Q. ZHAO and C. LI, *CIESC Journal*, 2014, **65**, 4321.
40. P. Conrad, G. E. Ewing, R. L. Karlinsey and V. Sadtchenko, *The Journal of Chemical Physics*, 2005, **122**.
41. L. Kaufmann, C. Marcolli, J. Hofer, V. Pinti, C. R. Hoyle and T. Peter, *Atmospheric Chemistry and Physics*, 2016, **16**, 11177-11206.
42. A. Kumar, C. Marcolli and T. Peter, *Atmospheric Chemistry and Physics*, 2019, **19**, 6059-6084.
43. A. J. Miller, C. Fuchs, F. Ramelli, H. Zhang, N. Omanovic, R. Spirig, C. Marcolli, Z. A. Kanji, U. Lohmann and J. Henneberger, *Atmospheric Chemistry and Physics*, 2025, **25**, 5387-5407.
44. A. Soni and G. Patey, *The Journal of Physical Chemistry C*, 2022, **126**, 6716-6723.
45. Y. Yu, M. Chen, Y. Lei and H. Niu, *Nature Communications Physics*, 2025, **8**, 7.
46. B. Nagare, C. Marcolli, A. Welti, O. Stetzer and U. Lohmann, *Atmospheric Chemistry and Physics*, 2016, **16**, 8899-8914.
47. J. C. Lee, T. Hansen and P. L. Davies, *Cryobiology*, 2023, **113**, 104584.
48. A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer and C. Hargus, *Journal of Physics: Condensed Matter*, 2017, **29**, 273002.

49. S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson and G. Ceder, *Computational Materials Science*, 2013, **68**, 314-319.
50. A. Zur and T. McGill, *Journal of Applied Physics*, 1984, **55**, 378-386.
51. Z. Lin, C. Du, B. Yan and G. Yang, *Journal of Catalysis*, 2019, **372**, 299-310.
52. N. Fukuta, *Journal of Atmospheric Sciences*, 1958, **15**, 17-26.
53. I. Sano, Y. Fujitani, K. Ito and S. Kitani, *Journal of the Meteorological Society of Japan. Ser. II*, 1956, **34**, 185-189.
54. D. J. Cziczo, O. Stetzer, A. Worringen, M. Ebert, S. Weinbruch, M. Kamphus, S. J. Gallavardin, J. Curtius, S. Borrmann and K. D. Froyd, *Nature Geoscience*, 2009, **2**, 333-336.
55. R. E. Passarelli Jr, H. Chessin and B. Vonnegut, *Science*, 1973, **181**, 549-551.
56. R. Souda, T. Aizawa, N. Sugiyama and M. Takeguchi, *Physical Chemistry Chemical Physics*, 2020, **22**, 20515-20523.
57. S. A. Zielke, A. K. Bertram and G. N. Patey, *The Journal of Physical Chemistry B*, 2015, **119**, 9049-9055.
58. A. Raveh, Z. Tsameret and E. Grossman, *Surface and Coatings Technology*, 1997, **88**, 103-111.
59. E. Chong, M. King, K. E. Marak and M. A. Freedman, *The Journal of Physical Chemistry A*, 2019, **123**, 2447-2456.
60. F. L. Heale, I. P. Parkin and C. J. Carmalt, *ACS Applied Materials & Interfaces*, 2019, **11**, 41804-41812.
61. Z. R. Schiffman, M. S. Fernanders, R. D. Davis and M. A. Tolbert, *ACS Earth and Space Chemistry*, 2023, **7**, 812-822.
62. J. C. Pohlman and P. Landers, *IEEE Transactions on Power Apparatus and Systems*, 1982, 2443-2450.
63. M. Farzaneh, C. Volat and A. Leblond, in *Atmospheric Icing of Power Networks*, Springer, 2008, pp. 229-268.

Chapter 4

Transferability of prediction model: Case studies for nucleator prediction of salt hydrates

In the previous chapter, the heterogeneous nucleation of ice was investigated, identifying key geometric descriptors that govern compatibility likelihood for various candidate nucleators. While ice is one of the most extensively studied phase change materials (PCMs), its relevance is primarily confined to environmental and cryopreservation contexts[1-3]. To test the transferability and generalisability of the nucleator prediction model, this thesis now turns to sodium acetate trihydrate [$Na(CH_3COO)\cdot 3H_2O$], hereafter abbreviated to SAT, a PCM of direct industrial significance in thermal energy storage systems[4,5].

In the UK, over half of domestic sector energy consumption is spent on heating[6]. As society seeks to decarbonise the energy grid, heat storage batteries, which are PCMs that store renewable energy in the form of latent and sensible heat, offer a practical means of balancing fluctuations in energy supply and demand[7-10]. Among all PCM candidates, salt hydrates such as SAT are particularly attractive due to their low cost, non-flammability, and high energy storage density[11,12]. SAT, with a melting temperature of 58 °C, is ideally suited for domestic water heating systems[13]. In addition to latent heat storage, it also possesses a sensible heat capacity approximately four times greater than that of an equivalent mass of water[14].

Despite these advantages, SAT suffers from a pronounced tendency to subcool, remaining in a metastable liquid state well below its melting point due to the absence of spontaneous nucleation events[15]. While this issue is easily circumvented in small-scale applications like hand warmers where mechanical flexing of a metal disc releases seed crystals stored on its rough surface, i.e. homogenous nucleation[16], such solutions are not viable in large-scale, sealed heat battery systems.

A more scalable approach involves heterogeneous nucleation: introducing solid nucleator particles to trigger crystallisation. Several such additives have been

identified, including tetrasodium pyrophosphate decahydrate ($Na_4(P_2O_7)\cdot 10H_2O$) and disodium hydrogen phosphate hydrates ($Na_2(HPO_4)\cdot 7H_2O$ and $Na_2(HPO_4)\cdot 2H_2O$)[13]. A critical requirement is that the nucleator must remain stable and solid throughout operation, i.e., it must possess a melting point above that of SAT. Even when this condition is met, further challenges remain, such as dehydration to a non-nucleating anhydrous form[17] or dissolution into the PCM melt at elevated temperatures[18].

Traditionally, the search for suitable nucleators has relied heavily on trial-and-error experimentation, often dictated by the availability of reagents rather than the underlying scientific rationale[19-21]. This approach is resource-intensive and rarely provides mechanistic insights into why a particular additive succeeds or fails to nucleate the PCM.

In contrast, in silico screening offers a promising and rational alternative. Computer simulations have provided valuable theoretical insights into the mechanisms of heterogeneous nucleation[22-25], yet they have not yielded generalisable descriptors or reliable prediction frameworks. In particular, for inorganic salt systems like SAT, the structural and energetic properties that govern nucleator performance remain poorly defined[26].

This challenge is well-suited to a supervised data-driven approach, particularly when a curated dataset of both successful and unsuccessful nucleators is available. For SAT, such a dataset exists, comprising three known effective nucleators and fifteen documented failures[27] that are very well experimented. By training a classification model on this dataset, geometric features and constraints that are statistically predictive of nucleation success can be uncovered. This, in turn, enables us to search virtually for new nucleator candidates that might outperform existing materials—for instance, by offering a reduced a subcooling effect, thereby improving the efficiency and robustness of SAT-based heat batteries.

This chapter therefore builds upon the methodology reported in the previous chapter for ice, herein extending its application to salt hydrates including SAT to demonstrate

that data-driven nucleator prediction is transferable across materially and functionally distinct PCMs.

4.1. Model modification: Adding a new slab proportion feature

4.1.1. Considerations of size comparability of PCM-NUC interface pairings

In the previous work on ice nucleation, the necessity of constraining slab sizes was less pronounced due to the relatively small and symmetry-consistent unit cells involved. Ice-I_h has modest lattice parameters, and many of its known nucleators, such as simple halides and oxides that were explored in Chapter 3, possess comparably sized and isotropic unit cells. As a result, surface slabs generated from these materials often fell within a naturally compatible geometric range, and excessive scaling was rarely required to achieve good registry during interface matching.

However, when extending the model to accommodate larger, low-symmetry PCMs such as sodium acetate trihydrate (SAT), the problem of scale mismatch becomes significantly more pronounced. SAT features a substantially larger unit cell with anisotropic dimensions, and its sliced surfaces can span several times the area of typical nucleator slabs. Without a constraint on the relative sizes of the nucleator and PCM slabs, the matching algorithm can produce unrealistic supercells where small nucleators are tiled repeatedly against large PCM surfaces (or vice versa), leading to physically unrealistic interfaces. These extreme mismatches are geometrically valid in the mathematical sense, but they are unlikely to yield coherent or energetically favourable nucleation sites in practice.

This disparity necessitates a screening feature to assess size proportionality between candidate slabs, either through bulk volumetric comparison or, more appropriately, via an in-plane slab area ratio. The introduction of such a constraint helps ensure that the docked slabs share commensurate spatial scales, allowing the prediction model to remain physically grounded as it is applied to a broader class of PCMs.

4.1.2. From Telke's rule to surface area proportion

Volumetric constraints have historically been used as a heuristic in nucleator selection, most notably in the context of Telkes' rule[28], which posits that effective nucleators tend to possess a similar unit cell volume to the phase change material (PCM) they induce, and that nucleation may occur when the crystallographic data of the nucleation catalyst and the salt to be crystallised agree within 15%. This empirical guideline, while useful in coarse-grained screening, is inherently bulk-focused and was never designed with surface-based geometric matching in mind. Volume is a bulk property that includes the out-of-plane dimension, which may be irrelevant or misleading for surface compatibility. Moreover, in the context of interface-driven heterogeneous nucleation, the use of the volumetric comparison may result in promising slab combinations being inadvertently discarded due to differences in lattice parameters perpendicular to the interface plane or due to crystallographic anisotropy, i.e. features that are irrelevant to the formation of two-dimensional interfaces.

Thus, in order to extend the interface matching model to allow study of potential PCM/nucleator combinations where large differences in unit cell parameters arise, a 6th feature, the slab area ratio tolerance criterion, is introduce, which considers only the unit cell areas of the interfaced slabs. This criterion better reflects the conditions at the interface and ensures that screening is performed in a parallel and modular fashion, rather than prematurely eliminating candidates before the slabs can be evaluated. In practice this feature makes sure that the unit cell areas of the two slabs, $<\overrightarrow{a_1}, \overrightarrow{b_1}>$ and $<\overrightarrow{a_2}, \overrightarrow{b_2}>$, lie within a specified relative range defined by the user, such that:

$$\frac{\left|\overrightarrow{a_1}, \overrightarrow{b_1}\right|}{\left|\overrightarrow{a_2}, \overrightarrow{b_2}\right|} - 1 \leq \epsilon \qquad (1)$$

where the subscripts 1 and 2 refer to the PCM and nucleator, respectively, and ϵ is slab area ratio tolerance. Unlike *m_n_*tolerance, which is described by geometric slab feature 5 (as described in the previous chapter), and which is responsible for evaluating how well the scaled supercells align, this new feature controls the relative sizes of the supercell combinations that are permitted to grow, and thus permits the user to exclude

any slab combinations with fundamentally mismatched unit cell sizes. In the context of heterogeneous nucleation, if the native (unit cell-level) areas of NUC slab and PCM slab are very different, then forcing them to match via large supercells might still be mathematically possible, i.e. meeting features 1-5, but physically unrealistic. Such mismatches would typically require large or strained supercells which increases energy penalties, resulting in high interfacial energy and poor coherence. By enforcing a bounded area ratio, the search is biased toward slab pairs that are more likely to form low-strain, energetically favourable interfaces in realistic conditions. In effect it acts as a sanity check to avoid attempting forced commensuration on badly mismatched pairs.

In terms of considerations on whether feature 5, *i.e.* *m_n_*tolerance is redundant from slab area ratio tolerance, while the maximum overlapping area provides a measure of the absolute size of a common supercell that can be formed by repeating the two slabs, it does not directly account for how proportionally well the individual unit cells align within that overlap. The *m_n_*tolerance parameter was defined as the normalised difference between the in-plane supercell area formed by repeating slab 1 m times along its lattice vectors a_1 and b_1, and that formed by repeating slab 2 n times along its lattice vectors a_2 and b_2, scaled by the overlapping area. In other words, The *m_n_*tolerance parameter was defined as the normalised difference between the areas of the two in-plane supercells, $m<\overrightarrow{a_1} \times \overrightarrow{b_1}>$ and $n<\overrightarrow{a_2} \times \overrightarrow{b_2}>$ scaled by the overlapping area. This value captures the relative area mismatch between the repeated unit cells of the two slabs, rather than the total area they can occupy together. In cases where the maximum overlap is large but achieved through mismatched repetition (*e.g.*, over- or under-stretching one slab), the *m_n_*tolerance remains sensitive to this disproportion, thereby serving as a strain-aware refinement of the matching condition. It ensures that the interface is not only geometrically possible but also minimally distorted, which is critical when screening for physically realistic or low strain heterostructures.

A modified workflow targeting SAT nucleator prediction as well as other salt hydrates is therefore demonstrated in Figure 4.1.

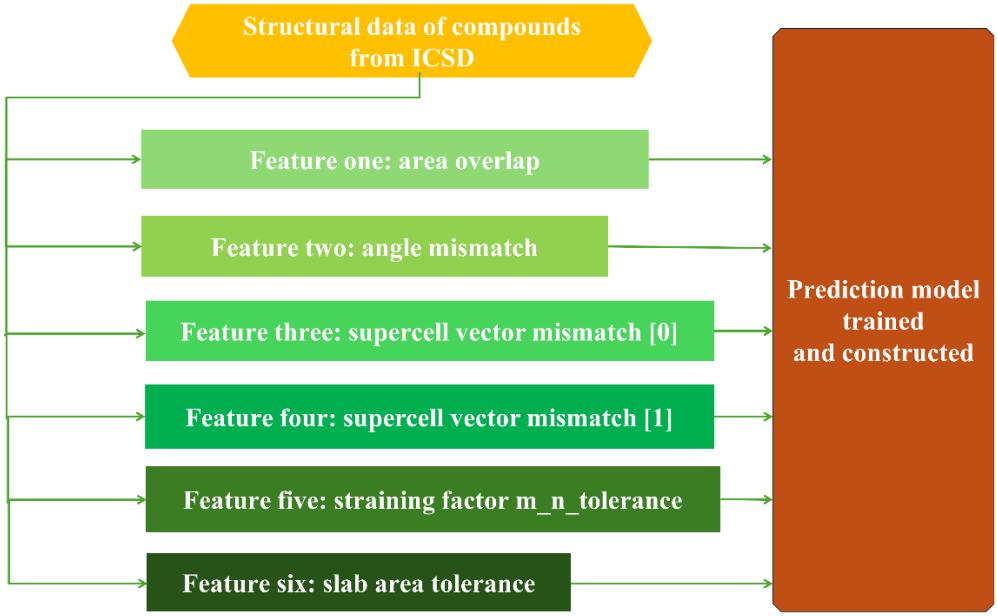


Figure 4.1 Schematic workflow of the geometric interface-matching model used to train and test the nucleation prediction algorithm regarding salt hydrate PCMs other than ice. Structural data from the ICSD is used to generate surface slabs, from which six geometric features are extracted: (1) area overlap, (2) angle mismatch, (3-4) supercell vector mismatches in orthogonal directions, (5) straining factor (*m_n_*tolerance), and (6) slab area tolerance. These features form the input space for training a prediction model to classify potential nucleators.

4.2. Training of modified prediction model and high-throughput prediction for nucleators of SAT

4.2.1. Training of modified prediction model

After the bank of 276,768 interface models $(h,k,l \leq 3)$ cleaved from 3 working nucleators and 15 non-working nucleators were put into prediction model as training data, values for the six features were extracted into a data frame matrix. As this is a classifier model, the ultimate decision is whether a compound is assigned as a working or non-working NUC for SAT [i.e. recorded as a binary y = 0 (negative result, representing a non-working NUC) or y = 1 (positive result, representing a working NUC)]. To reach this overall decision for the training set binary decisions were recorded for each slab docking scenario as defined by the set of 6 feature values, which

were initially set to relatively loose tolerance limits learned from Chapter 3, with maximum area overlap set to approximately 10 times the size of basal plane of SAT unit cell, giving the interface matches sensible supercell combinations, slab area ratio value set as 1.5 for comparable slab sizes to be matched, and maximum angle mismatch, *m_n_*tolerance, vector mismatch [u] and [v] set close to 0. Values for the four features (maximum angle mismatch, *m_n_*tolerance, vector mismatch [u] and [v]) were then systematically tightened to essentially train the model to make binary decisions, as the number of interface models that fit the increasingly stringent conditions will drop to zero for the non-working NUCs. The results for the training process are shown in Figure 4.3, where the number of matching slab models for PCM/NUC slab combinations that fit a less strict upper limit setting is shown, alongside the results obtained when the binary decision making process was reached, i.e. the point at which the number of interface models for the non-working NUCs that fulfilled the feature criteria fell to zero. Miller indices of matching pairs for the three known working nucleators under the final values is also shown in Table 4.1. The final values for the features were thus set as follows:

$$\text{Maximum area overlap} = 1100\ \text{Å}^2 \tag{2}$$

$$\text{Maximum angle mismatch} = 0.01° \tag{3}$$

$$\text{Maximum supercell vector mismatch [0]} = 0.01 \tag{4}$$

$$\text{Maximum supercell vector mismatch [1]} = 0.01 \tag{5}$$

$$m\_n\_\text{tolerance} = 0.02 \tag{6}$$

$$\varepsilon = 1.5 \tag{7}$$

Figures 4.3(a) and (c) show the number of matching slab pairs (“hits”) obtained when all crystallographic duplicates, i.e. symmetry-equivalent cleaves such as (111), (222), and (333) are included. In contrast, Figures 4.3(b) and (d) present results with duplicate surfaces removed, such that each unique cleaving direction is only counted once. This filtering is critical for accurately estimating model precision, as multiple hits from symmetry-equivalent surfaces can artificially inflate the apparent predictive power of certain nucleators.

In the looser constraint case (Figures 43(a) and (b)), three working nucleators, i.e. $Na_2(HPO_4)\cdot 2H_2O$, $Na_2(HPO_4)\cdot 7H_2O$ and $Na_4P_2O_7\cdot 10H_2O$ yield high numbers of hits,

consistent with their experimentally verified activity. However, several of the remaining nucleators, previously reported as ineffective, also produce hits, indicating false positives under the relaxed screening regime. These are particularly prominent when duplicates are included (Figure 4.3(a)), suggesting that high feature tolerance may allow superficially compatible but geometrically redundant surfaces to enter the pool.

In contrast, the tighter constraint case (Figures 4.3(c)and (d)) produces a more selective outcome. The same three known nucleators are correctly predicted with clear hit counts, but the number of false positives is markedly reduced. Only two experimentally unsuccessful nucleators, sodium tungstate dihydrate and sodium molybdate dihydrate, produce any hits at all, and these hits are fewer and less consistent than those from the confirmed active group. Of course, there are factors beyond simple geometric matching that account for the success or failure of a particular nucleating agent. Nevertheless, this reflects an overall increase in prediction precision and a reduction in false discovery rate, validating the tightened constraints as more effective filters. Moreover, the general decline in total hit count suggests that the stricter thresholds improve the model's discriminative sharpness, preserving only those interfaces with strong geometric plausibility. Sodium pyrophosphate ($Na_4P_2O_7$) is also worthy of note, as some literature reports cite this as a working NUC for SAT[27]. Other reports point out that upon heating and cooling it will likely convert to the decahydrate form in-situ, i.e. the anhydrous form is not the active form for this compound[13]. Reassuringly, the decahydrate form is correctly predicted to be a working NUC[29]. Also, the workflow generates no false negatives, i.e. false reporting of known working NUC behaviour, which is extremely encouraging. The overall prediction success rate based purely on geometric screening is therefore 16/18 (89%).

In terms of the role of duplicate filtering in hit distribution, comparing the panels with (a and c) and without (b and d) symmetry-equivalent surfaces reveal a subtle but important distinction. While the inclusion of duplicates allows for a denser hit map, this does not necessarily improve prediction quality. For example, in Figures 4.3(a) and (c),nucleators whose crystal structures generate multiple symmetry-equivalent slab terminations contribute disproportionately to hit counts.. This can give a

misleading impression of nucleation compatibility. When these duplicates are removed (Figures 4.3(b) and (d)), the model is forced to rely on unique surface matches, resulting in a more balanced and physically interpretable output. Crucially, even in the tightly constrained, duplicate-filtered case (Figure 4.3(d)), the model successfully retains all true positives (the three experimentally verified nucleators), while reducing the number of false positives to a minimal and clearly identifiable subset. This strongly supports the idea that the final constraint set used in this study offers an effective balance between sensitivity and specificity.

These four panels demonstrate the importance of both constraint refinement and symmetry-aware filtering in improving the interpretability and reliability of computational nucleator screening. Looser geometric thresholds may overestimate compatibility by including strained or geometrically redundant interfaces, while duplicate surface inclusion can exaggerate the perceived success of symmetric structures. The results shown here justify the implementation of a tight, symmetry-filtered constraint set as the final configuration for predictive screening. This not only increases consistency with experimental observations but also enhances the physical meaningfulness of the model by aligning hit frequency with realistic interface matching conditions.

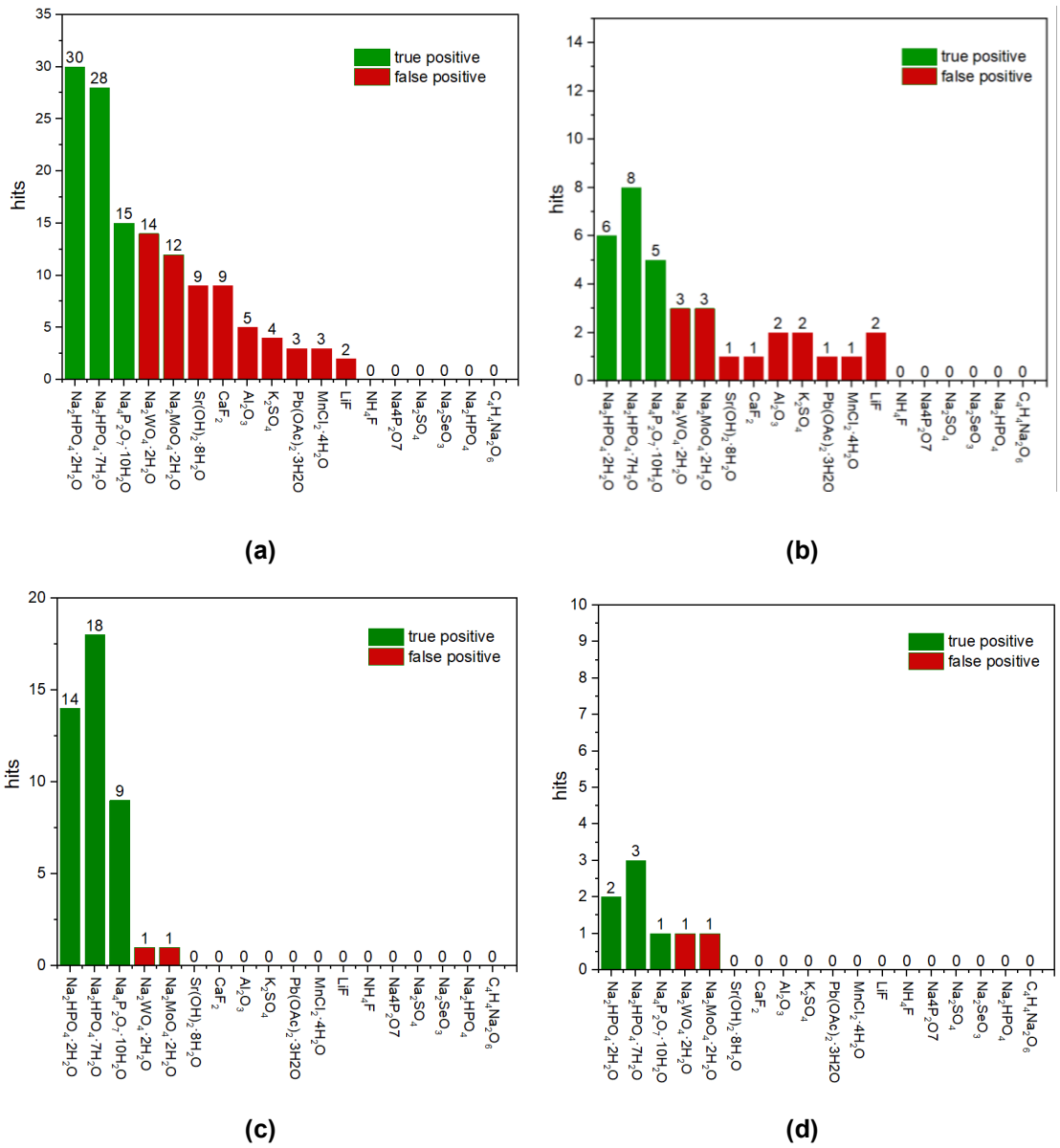


Figure 4.3. Number of matching slab models for SAT/NUC slab combinations for 18 NUCs that **(a)** fit a less strict feature limit (maximum area overlap = 1100 Å^2, maximum angle mismatch = 0.1°, maximum supercell vector mismatch [0] = 0.1, maximum supercell vector mismatch [1] = 0.1), symmetrical duplicates included, **(b)** symmetrical duplicates excluded, and **(c)** fit the more strict feature limit (maximum area overlap = 1100 Å^2, maximum angle mismatch = 0.01°, maximum supercell vector mismatch [0] = 0.01, maximum supercell vector mismatch [1] = 0.01), symmetrical duplicates included, **(d)** symmetrical duplicates excluded. True positives are shown in green and false positives in red.

Table 4.1. Miller indices of matching pairs for the three known working nucleators under criteria maximum area overlap = 1100 Å^2, maximum angle mismatch = 0.01°, maximum supercell vector mismatch [0] = 0.01, maximum supercell vector mismatch [1] = 0.01, *m_n_*tolerance = 0.02, slab area tolerance = 1.5.

NUC	Miller indices of matching surface of NUC	Miller indices of matching surface of SAT
$Na_2HPO_4 \cdot 2H_2O$	(101)	(010)
$Na_2HPO_4 \cdot 2H_2O$	(100)	(001)
$Na_2HPO_4 \cdot 7H_2O$	(100)	(010)
$Na_2HPO_4 \cdot 7H_2O$	(001)	(001)
$Na_2HPO_4 \cdot 7H_2O$	(101)	(010)
$Na_2P_2O_7 \cdot 10H_2O$	(001)	(100)

4.2.2. Redundancy validation

To assess the informational independence of the geometric features used in the nucleation prediction model, a correlation matrix was constructed based on pairwise Pearson correlation coefficients[30-32]. This statistical tool quantifies the degree of linear dependence between pairs of features, returning values between -1 (perfect negative correlation) and +1 (perfect positive correlation). Values near zero indicate little to no linear correlation[33]. In the context of feature engineering for data-driven training, this type of analysis is critical for identifying redundant or collinear variables, which may lead to overfitting, inflated feature importance, or instability during training.

Here, the correlation matrix serves two main purposes: first, to validate the inclusion of the newly introduced feature 6 (slab area ratio tolerance), and second, to confirm that all six geometric features contribute distinct information about slab-to-slab compatibility. The matrix was computed using all interface pairings evaluated during model training (described below), and a colour-coded scheme was applied to highlight high ($r > 0.7$), moderate ($r = 0.3$-0.7), and low ($r < 0.3$) correlations.

As shown in Figure 4.4, the results reveal that all pair-wise correlations are below 0.15, indicating that the features are effectively independent and non-redundant. The highest correlation observed is between vector mismatch [1] and *m_n_*tolerance ($r = 0.13$), which is both expected and modest. This weak association likely arises because both features involve supercell construction, albeit different aspects: vector mismatch

targets unit cell vector discrepancies, while *m_n_*tolerance evaluates how well the scaled supercells align.

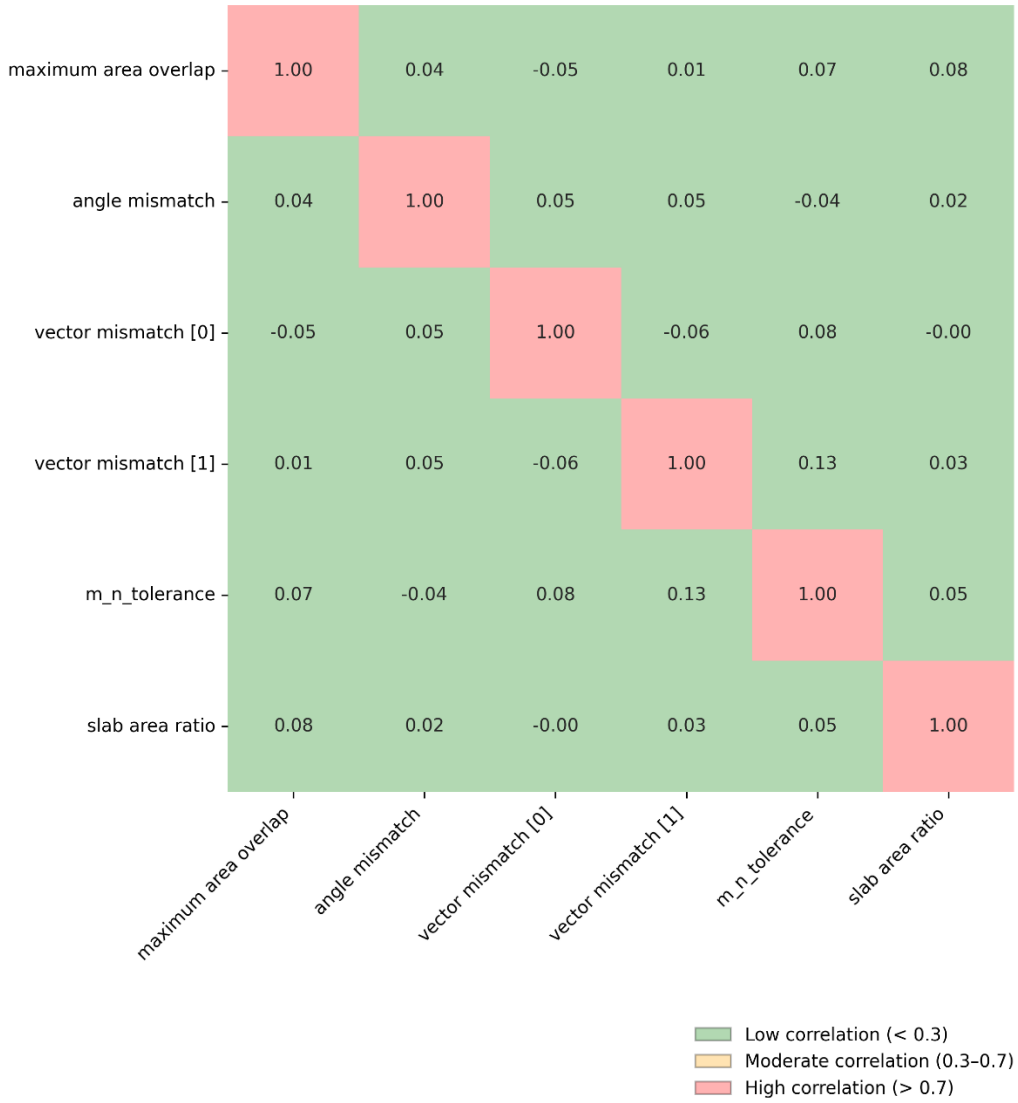


Figure 4.4. Correlation matrix of the six geometric features used in the slab matching model. Each cell represents the Pearson correlation coefficient between a pair of features, coloured according to strength of correlation (green = low < 0.3, pink = high > 0.7). All features exhibit low pairwise correlation, indicating that they provide independent information. In particular, the newly added slab area ratio (Feature 6) is uncorrelated with all other descriptors, supporting its role as a physically meaningful and non-redundant addition to the predictive feature set.

Most notably, the slab area ratio (feature 6) shows negligible correlation with all other features ($r \leqslant 0.08$), supporting its role as a structurally orthogonal addition to the descriptor set. This justifies its inclusion as a meaningful new constraint that adds discriminatory power without introducing collinearity. Conceptually, this feature assesses global size compatibility between surface unit cells, whereas features like area overlap and vector mismatch describe more localised geometric alignments. The lack of correlation reinforces the idea that area proportion provides unique, physically interpretable filtering that cannot be inferred from other metrics.

Interestingly, even features derived from similar constructs, such as vector mismatch [0] and vector mismatch [1], remain uncorrelated ($r = -0.06$), reflecting the structural anisotropy inherent in the dataset and the independent treatment of the two lattice vectors in interface geometry. This uncorrelated behaviour may stem from the presence of low-symmetry unit cells (*e.g.* monoclinic or triclinic), where the lattice vectors are not constrained to orthogonal directions, allowing mismatch in one vector without necessarily affecting the other; In contrast, for high-symmetry systems like orthorhombic cells, once one vector is matched, the other may follow automatically due to 90° inter-axial angles and uniform dimensions. The observed lack of correlation thus suggests that many slab pairings involve cells where anisotropy in edge alignment plays a critical role. This independence may be further modulated by the γ angle between lattice vectors, which can decouple the directional mismatches in non-orthogonal systems.

These findings confirm that the six features form a non-redundant, complementary descriptor set for quantifying slab match quality. This diversity in geometric criteria improves the model's ability to distinguish viable nucleator/PCM pairs, while also ensuring interpretability and robustness in downstream machine learning applications.

4.2.3. Baseline comparison

The manually tuned geometric matching model was evaluated on 18 candidate nucleators for SAT. The resulting confusion matrix is shown in Figure 4.5. Of the 18 compounds, 16 were correctly classified, yielding an overall accuracy of 88.9%.

The model achieved perfect sensitivity toward working nucleators (3/3 correctly identified) and strong specificity toward non-working compounds (13/15 correctly identified, 86.7%). The balanced accuracy, calculated as the mean of sensitivity and specificity, was 93.3%, indicating strong discriminative performance across both classes.

To contextualise predictive performance, two baseline classifiers were evaluated. For binary classification, random guessing yields an expected accuracy of 50%. The majority-class baseline was calculated by assigning all compounds to the most frequent experimental class (non-working nucleator), which comprised 15 of the 18 candidates, corresponding to an accuracy of 83.3%.

The geometric matching model achieved an accuracy of 88.9%, exceeding both the random and majority-class baselines. This indicates that, for SAT, geometric compatibility provides strong discriminatory power and is not merely reflecting class imbalance.

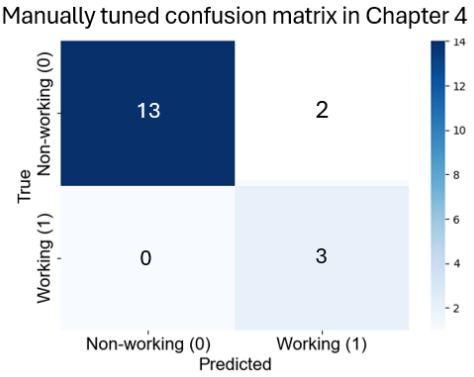


Figure 4.5. Confusion matrix for the manually tuned geometric matching model applied to 18 candidate nucleators for SAT.

4.2.4. High-throughput screening of database for potential nucleators of SAT

Successful validation of the prediction model now permits application of a high-throughput screening of many NUC candidates, in order to identify potential new additives with enhanced properties for SAT-based thermal heat storage applications. This could include, for example, having a higher melting point in order to extend the

temperature operating conditions of the device. Other criteria can be considered and screened for at this stage, including compound availability, toxicity and price.

In principle, while the entire Inorganic Crystallographic Structure Database (ICSD)[34] can be screened, it is sensible to restrict the search space to six categories, namely nitrates, sulfates, phosphates, pyrophosphates, hydrates and halides, on the likely basis of stability, high melting points and unit cell size matching. This initial input database of structures contained 14101 entries, comprising 779 nitrates, 2031 sulfates, 5580 phosphates, 832 pyrophosphates, 2741 hydrates, and 2138 halides. Taking these candidates through the NUC predictor algorithm trained to generate binary predictions resulted in 30 nitrates, 91 sulfates, 214 phosphates, 67 pyrophosphates, 78 hydrates, and 41 halides that fulfilled the stringent geometric matching conditions (see Table 4.2). The purchasable compounds according to Chemical Availability Search (ChASe)[35] are underlined green in the table for future experimental validation purposes.

Table 4.2. High throughput predictions of new crystal nucleators for SAT.

ICSD	Compound	ICSD	Compound
Nitrates: 30/779			
65949	$Ag(NH_3)_3NO_3$	31353	$Cu_2(OH)_3NO_3$
6298	$Co(NO_3)_2(NH_3)_4NO_3$	98714	$ErHPO_4NO_3{\cdot}3H_2O$
78922	$Cu(NH_3)(NO_3)_2$	112306	$GdTeO_3NO_3$
415719	$Hg(OH)NO_3{\cdot}H_2O$	100711	$KAg(NO_3)_2$
35735	$Ag(NH3)_2NO_3$	37180	$LiNO_2{\cdot}H_2O$
36557	$Cd(OH)NO_3{\cdot}H_2O$	87623	$Lu(NO_3)_3{\cdot}4H_2O$
4426	$CdCu_3(OH)_6(NO_3)_2{\cdot}H_2O$	22367	$Mn(OH2)_6(NO_3)_2$
38152	$Cu_2(NO_3)(OH)_3$	421223	$MoO_2(NO_3)_2$
64618	$CdMg(NO_2)_4{\cdot}2H_2O$	128436	$SrNH_2SO_3NO_3{\cdot}H_2O$
50327	$CsNO_2$	14356	$Th(NO_3)_4{\cdot}5H_2O$
260072	$NdONO_3$	250750	$Pb_3(B_3O_7)NO_3$
28327	$Ni(NO_3)_2$	51474	$Pb_3O_2(OH)NO_3$
139091	$Ni_3(B_7O_{13})NO_3$	56434	$CsCa(NO_2)_3$
128935	$Pb_2(PO_4)NO_3{\cdot}H_2O$	411533	$PrCl_2NO_3{\cdot}5H_2O$
66709	$RbNO_3$	243986	$Rb_2Na(NO_3)_3$

Sulfates: 91/2031			
78195	$(Ag(NH_3)_2)_2SO_4$	96348	$Cu_3(OH)_4SO_4$
100856	$(NH_4)Al(SO_4)_2{\cdot}12H_2O$	280578	$FeSO_4{\cdot}7H_2O$
59891	$(NH_4)Cr(SO_4)_2{\cdot}12H_2O$	252116	$GdNa(SO_4)_2{\cdot}H_2O$
174961	$(NH_4)Sb(SO_4)_2$	280548	$KAl(SO_4)_2{\cdot}12H_2O$
108844	$(NH_4)_2Cr(SO_4)_2{\cdot}6H_2O$	170708	$KCe(SO_4)_2{\cdot}H_2O$
62991	$(NH_4)_2Cu(SO_4)_2{\cdot}6H_2O$	95459	$KCu_3O(SO_4)_2Cl$
40982	$(NH_4)_2(Ni(SO_4)_2{\cdot}6H_2O$	236129	$KFe(SO_4)_2{\cdot}4H_2O$
90099	$(NH_4)_2(Zn(SO_4)_2{\cdot}6H_2O$	91951	$KLiSO_4$
201523	$(NH_4)_2Cu(SO_4)_2{\cdot}6H_2O$	97874	$KMnH(SO_4)_2{\cdot}2H_2O$
23824	$(NH_4)_2SO_3{\cdot}H_2O$	430240	$KSb(SO_4)_2$
266693	$(NH_4)_2SO_4$	144213	$KTb(SO_4)_2$
391376	$NiNa(SO_4)_2{\cdot}6H_2O$	50582	$K_2Co(SO_4)_2{\cdot}6H_2O$
97313	$VOSO_4{\cdot}3H_2O$	171289	$K_2Cu(SO_4)_2{\cdot}6H_2O$
62368	$BaSO_4$	162314	$K_2Fe(SO_4)_2{\cdot}6H_2O$
262106	$CaSO_4{\cdot}0.5H_2O$	162313	$K_2Mg(SO_4)_2{\cdot}6H_2O$
62641	$CdSO_3$	60762	K_2SO_3
133732	$CdSO_4{\cdot}H_2O$	47341	$K_2Be(SO_4)_2$
43984	$CdZr(SO_4)_3$	137924	$K_2Ca_2(SO_4)_3$
415601	$Cd_2(OH)_2SO_4$	40279	$K_2Cd_2(SO_4)_3$
280407	$Cd_2Tl_2(SO_4)_3$	81082	$K_2Co_2(SO_4)_3$
33736	$CoSO_4$	100420	$K_2Mg_2(SO_4)_3$
65979	$CrSO_4{\cdot}3H_2O$	239300	$K_2Ni(SO_4)_2{\cdot}6H_2O$
14316	$CsAl(SO_4)_2{\cdot}12H_2O$	67827	$K_2Ru(SO_4)_2{\cdot}6H_2O$
69435	$CsFe(SO_4)_2{\cdot}12H_2O$	161155	$K_2Zn(SO_4)_2{\cdot}6H_2O$
81085	$CsLiSO_4$	82872	$K_2Zn_2\ (SO_4)_3$
175547	$CsSb(SO_4)_2$	200547	$K3Yb(SO_4)_3$
409738	$Cs_2Mn(SO_4)_2{\cdot}6H_2O$	25055	$Li(NH_4)SO_4$
409741	$Cs_2Ni(SO_4)_2{\cdot}6H_2O$	63179	$LiCsSO_4$
93171	γ-$CuSO_4$	88458	$Li_2V_2(SO_4)_3$
93999	Cu_2OSO_4	174059	$LiRbSO_4$
759711	$MgSO_4{\cdot}H_2O$	414448	$NaGd(SO_4)_2{\cdot}H_2O$
26772	$MgK_2(SO_4)_2{\cdot}6H_2O$	419316	$Na_2Co(SO_4)_2{\cdot}6H_2O$
119912	$MgSO_4{\cdot}5H_2O$	194362	$Na_2Fe(SO_4)_2{\cdot}2H_2O$
66174	$Mg_3(OH)2(SO_4)_2$	170139	$Na_2Zn(SO_4)_2$
74810	$MnSO_3{\cdot}H_2O$	249552	$Na_3H(SO_4)_2$

423300	$MnSO_4 \cdot 3H_2O$	43988	$NiZr(SO_4)_3$
43986	$MnZr(SO_4)_3$	100625	$PbSO_4$
425664	$Mn_2(OH)_2SO_3$	79126	$Pb_4(PO_4)_2SO_4$
14310	$Na(NH_4)SO_4 \cdot 2H_2O$	59388	$RbCe(SO_4)_2 \cdot H_2O$
15368	$NaAl(SO_4)_2 \cdot 12H_2O$	280321	$RbCr(SO_4)_2 \cdot 12H_2O$
143396	$NaBi(SO_4)_2 \cdot H_2O$	119019	$Rb_2Cu(SO_4)_2 \cdot 6H_2O$
81224	$NaCe(SO_4)_2 \cdot H_2O$	203159	Rb_2SO_4
160770	$Rb_2Be(SO_4)_2 \cdot 2H_2O$	252117	$TbNa(SO_4)_2 \cdot H_2O$
186951	$SnSO_4$	83659	$Tl2Cd2(SO_4)_3$
148779	$Rb_2Ca_2\ (SO_4)_3$	44051	$ZnZr(SO_4)_3$
96631	$Rb_2Cd_2(SO_4)_3$		
Phosphates: 214/5580			
80044	$(NH_4)CoPO_4 \cdot H_2O$	26816	$Al_3(OH)_3(PO_4)_2 \cdot 5H_2O$
46787	$(NH_4)H_2PO_4$	28311	$Al_3(PO_4)_2(OH)_3 \cdot 6H_2O$
259131	$(NH_4)PO_3$	88133	$Ba((VO)(\ PO_4))_2 \cdot 4H_2O$
90844	$(NH_4)SnPO_4$	66699	$Ba(VO_2)PO_4$
78004	$(NH_4)(VO)PO_4$	50608	$BaCu_2(PO_4)_2 \cdot H_2O$
91845	$(NH_4)(VO_2)PO_4 \cdot 3H_2O$	64136	$BaNi_2Fe(PO_4)_3$
200085	$(NH_4)CdPO_4 \cdot H_2O$	81457	$Ba_2Cu(PO_4)_2$
710018	$(NH_4)Fe_4(PO_4)_3$	91803	$Ba_3Bi(PO_4)_3$
249131	$(NH_4)Gd(PO_3)_4$	200376	$Ba3Y(PO_4)_3$
239340	$(NH_4)MnPO_4 \cdot H_2O$	54163	$BiPO_4$
79749	$PrPO_4$	73894	$BiMg_2O_2PO_4$
23510	$(NH_4)_2Mn(PO_3)_4$	10406	$CaMgAl(OH)(PO_4)_2 \cdot 4H_2O$
170053	$(NH_4)_7(HPO_4)_2PO_4$	162030	$CaTh(PO_4)_2$
93766	$NbOPO_4$	23667	$CaZn_2(PO_4)_2 \cdot 2H_2O$
50671	$AgMg(PO_3)_3$	91524	$Ca_{10}Na(PO_4)_7$
261617	$Ag_2Co_3(PO_4)_2(HPO_4)$	126134	$Ca_2Sr(PO_4)_2$
93735	$Ag_2FeMn_2(PO4)_3$	97500	$Ca_3(PO_4)_2$
261616	$Ag_2Ni_3(HPO_4)(PO_4)_2$	177722	$Ca_3Co_3(PO_4)_4$
245002	$Ag_3In_2(PO_4)_3$	177721	$Ca_3Ni_3(PO_4)_4$
14000	Ag_3PO_4	56311	$Ca_5(OH)(PO_4)_3$
35652	$AlPO_4$	125897	$Ca_6(PO_3)_4$
66872	$CsNd(PO_3)_4$	432393	$KTb(PO_3)_4$
108840	$Na_2(HPO_4) \cdot 12H_2O$	151428	$SrFe_2(PO_4)_2$
10127	$CsPr(PO_3)_4$	182657	$KTiOPO_4$

20908	$Cs_2Co(PO3)_4$	420128	$SrSn(PO_4)_2$
161678	Cs_3PO_4	88956	$KZnPO_4$
202613	$Cs_3Fe_4(PO_4)_5$	23484	$K_2Cu(PO_3)_4$
250189	$Cu(NH_4)PO_4·H_2O$	158487	$K_2LuZr(PO_4)_3$
15833	$Cu_2(OH)PO_4$	40307	$K_2Ti_2(PO_4)_3$
188025	$Cu_2ZnPO_4(OH)_3·H_2O$	80167	$K_3Fe_3(PO_4)_4·H_2O$
417732	$Cu_3(PO_4)_2·H_2O$	172528	$K_3Gd_5(PO_4)_6$
100019	$Cu_5(OH)_4(PO_4)_2$	427805	$LiCr_4(PO_4)_3$
88011	$Fe(NH_3)_2PO_4$	153518	$LiEr(PO_3)_4$
281079	$FePO_4$	428811	$LiMg_3PO_4P_2O_7$
241128	$Fe_2(OH)PO_4$	25834	$LiMnPO_4$
66405	$Fe_2Na_3(PO_4)_3$	248138	$LiSm(PO_3)_4$
280884	$Fe_7(PO_4)_2(HPO_4)_4$	240703	$LiTb(PO_3)_4$
30881	$GaPO_4$	185595	$LiVOPO_4·2H_2O$
86632	$GaPO_4·2H_2O$	162784	$LiY(PO_3)_4$
416879	$GdLi(PO_3)_4$	79351	$LiZnPO_4·H_2O$
15227	$Ge(OH)PO_4$	91112	$LiZr_2(PO_4)_3$
420532	$HgPd_2(PO_4)_2$	51333	$Li_2FeZr(PO_4)_3$
80297	$InPO_4·2H_2O$	84703	$Li_2Na(MoO)_2(PO_4)_3$
173806	KH_2PO4	151920	$Li_2TiCr(PO_4)_3$
20970	$K(TiO)PO_4$	151919	$Li_2TiFe(PO_4)_3$
79651	$K(VO)PO_4$	20208	Li_3PO_4
291383	$K(VO_2)_2PO_4·3H_2O$	86461	$Li3Cr_2(PO_4)_3$
63552	$KBaFe_2(PO_4)_3$	62244	$Li3Fe_2(PO_4)_3$
96263	$KBi(PO_3)_4$	50420	$Li3Sc_2(PO_4)_3$
59295	$KCoAl(PO_4)_2$	70808	$Li3V_2(PO_4)_3$
135859	$KCoPO_4$	23017	$Li_4Zn(PO_4)_2$
20806	$KFe_4(PO_4)_3$	245847	$Li_5(VO)(PO_4)_2$
59836	$KGd(PO_3)_4$	159331	$Mg_2(OH)PO_4$
5972	$KMnPO_4·H_2O$	20796	$Mg_2(OH)PO_4·3H_2O$
20254	$KNd(PO_3)_4$	100784	$Mg_2Mn(PO_4)_2$
20034	$KNiPO_4$	238697	$PuPO_4$
83598	$KSrPO_4$	418084	$Na_4Mg_3(PO_4)_2(P_2O_7)$
87806	$Na4Ni_5(PO_4)_2(P_2O_7)_2$	74738	$RbPO_3$
201760	$Na_4Ni_7(PO_4)_6$	71906	$Rb(TiO)PO_4$
400310	$Na_5Ca_2Al(PO_4)_4$	421738	$RbMgPO_4$
12494	$Na_5Ni_2(PO_4)_3·H_2O$	89458	$RbMn_4(PO_4)_3$

62162	$NdPO_4$	10019	$RbNd(PO_3)_4$
4253	$NdLi(PO_3)_4$	66873	$RbSm(PO_3)_4$
155452	$PbCu_2(PO_4)_2$	86893	$Rb_2Mn(PO_3)_4$
129315	$PbMn_2Fe_2(PO_4)_3(OH)_3$	161677	Rb_3PO_4
177732	$PbNi_3(PO_4)_3$	74726	$RhPO_4$
50945	$PbSn(PO4)_2$	126200	$ScPO_4{\cdot}2H_2O$
128935	$Pb_2PO_4NO_3{\cdot}H_2O$	133688	$SmPO_4$
255465	$Pb_2Mg(PO_4)_2$	27567	$Sn_2(OH)PO_4$
290268	$Pb_2Ni(PO_4)_2$	67661	$Sr(VO)_2(PO_4)_2{\cdot}4H_2O$
8094	$Pb_3(PO_4)_2$	152395	$SrCu_2(PO_4)_2$
79126	$Pb_4(PO_4)_2SO_4$	17644	$Ce_2O(PO_4)_2$
125063	$Ca_8MnTb(PO_4)_7$	40523	$Co_2Fe(PO_4)_2$
89003	$Ca_9CoK(PO_4)_7$	100318	$Co_2Mg(PO_4)_2$
85109	$Ca_9MgK(PO_4)_7$	6208	$Co_3(PO_4)_2$
94539	$Ca_9MnNa(PO_4)_7$	193725	$MnHPO_4{\cdot}3H_2O$
82090	$Cd(MoO_2)PO_4$	62220	$MnPO_4{\cdot}H_2O$
15861	$Cd(PO_3)_2$	24066	$Mo(OH)_3PO_4$
160403	$Cd_2Cu_2(PO_4)_2SO_4{\cdot}5H_2O$	5148	$(NH_4)MgPO_4{\cdot}H_2O$
63548	$Cd_3(PO_4)_2$	9271	$NaBePO_4$
33598	$CePO_4$	66676	$NaCdIn_2(PO_4)_3$
100319	$Co_3Mg_3(PO_4)_4$	131813	$NaCdPO_4$
60836	$CrPO_4$	280175	$NaCoPO_4$
406829	$Cr_3(PO_4)_2$	251862	$NaCoCr_2(PO_4)_3$
710069	$CsMnGa_2(PO_4)_3{\cdot}2H_2O$	88719	$SrFe_3(PO_4)_3$
170792	$CsGd(PO_3)_4$	56292	$NaFePO_4$
200318	CsH_2PO_4	61696	$NaFe_3(PO_4)_3$
24550	$CsMgPO_4{\cdot}6H_2O$	250074	$NaGa_3(OH)_4(PO_4)_2$
142121	$CsNa_2Gd_2(BO_3)(PO_4)_2$	69967	$NaLi_2PO_4$
51490	$Mg_3(PO_4)_2{\cdot}4H_2O$	36249	$NaMnPO_4$
173888	$Zr_2(MoO_4)(PO_4)_2$	59357	$NaNi(PO_3)_3$
23642	$Mg_3Ca_3(PO_4)_4$	280116	$NaZnPO_4{\cdot}H_2O$
100783	$Mg_3Fe_3(PO_4)_4$	408373	$Na_2CaMg(PO_4)_2$
92226	$Na_2FeTi(PO_4)_3$	253558	$Na_2CaMnFe(PO_4)_3$
243841	$Na_2Fe_3(PO_4)_3$	20064	$Na_2Fe(OH)PO_4$
254401	$Na_2Ni_2Al(PO_4)_3$	14090	Na_3PO_4
239687	$Na_2Ni_2Cr(PO_4)_3$	93790	$Na_3Fe(PO_4)_2$
250437	$Na_2Tb(MoO_4)PO_4$	95532	$Na_3Fe_3(PO_4)_4$

416014	$Zr_2O(PO_4)_2$	56865	$Na_3Sc_2(PO_4)_3$
420127	$SrTi(PO_4)_2$	82116	$Na_4Co_3(PO_4)_2(P_2O_7)$
81370	$Th_4(PO_4)_4P_2O_7$	176183	$Na_4Fe_3(PO_4)_2(P_2O_7)$
36520	$TiPO_4$	82713	$Na_4Ni_3(PO_4)_2(P_2O_7)$
80450	$Tl(TiO)PO_4$	75446	$Zr_2(Na(PO_4))_4 \cdot 6H_2O$
79321	TlBe PO_4	80390	$Zr_2(WO_4)(PO_4)_2$
20108	$TlNd(PO_3)_4$	82286	VPO_4
132784	$TlPO_4 \cdot 2H_2O$	85565	$VPO_4 \cdot H_2O$
250163	$PrK_2Zr(PO_4)_3$	415924	$VOPO_4$
Pyrophosphates: 67/832			
44770	$(NH_4)TiP_2O_7$	133969	$Cs_3BaBi(P_2O_7)_2$
50920	$(NH_4)_2(VO)_3(P_2O_7)_2$	157108	$Cu_2P_2O_7$
27556	$(NH_4)_4P_2O_7$	65656	$KMoP_2O_7$
67501	$Ag_2H_2P_2O_7$	255046	$KNaCuP_2O_7$
90917	$Ag_2ZnP_2O_7$	415324	$KYbP_2O_7 \cdot 2H_2O$
434124	$Ag_6(VO)_2(PO_4)_2P_2O_7$	202671	$K_2(VO)_3(P_2O_7)_2$
83649	$BaFe_2(P_2O_7)_2$	97977	$K_2PdP_2O_7$
39398	$BaMgP_2O_7$	143564	$K_3SrBi(P_2O_7)_2$
69103	$BaTi_2(P_2O_7)_2$	195300	$K_4Pd_4(P_2O_7)_3$
66538	$BaV_2(P_2O_7)_2$	106101	$LiCsPbP_2O_7$
57336	$Ca_2P_2O_7$	428811	$LiMg_3PO_4P_2O_7$
72673	$CdBaP_2O_7$	41433	$Li_2BaP_2O_7$
72672	$CdSrP_2O_7$	247985	$Li_2MnP_2O_7$
162885	$CrTi_2(P_2O_7)_2$	260280	$Li_2Zn_3(P_2O_7)_2$
49919	$CsMoP_2O_7$	39814	$Li_4P_2O_7$
242169	CsLiCd P_2O_7	13430	$Li_6Zn(P_2O_7)_2$
265036	$CsLiZnP_2O_7$	154075	$NaCeP_2O_7$
177784	$CsNaZnP_2O_7$	759983	$NaCsZnP_2O_7$
252184	$Cs_2BaP_2O_7$	202751	$NaTiP_2O_7$
430123	$Cs_2CuP_2O_7$	80417	$Na_2CuP_2O_7$
82713	$Na_4Ni_3(PO_4)_2P_2O_7$	82116	$Na_4Co_3(PO_4)_2P_2O_7$
101930	$RbNaMgP_2O_7$	265035	$RbLiZnP_2O_7$
417072	$RbWOP_2O_7$	75599	$Tl(NbO)P_2O_7$
34177	$Rb_2(TeO)P_2O_7$	98627	$Tl_2Ni_4(PO_4)_2P_2O_7$
203091	$Rb_2(VO)P_2O_7$	195302	Tl2Pd3$(P_2O_7)_2$
68977	$Rb_2(VO)_3(P_2O_7)_2$	64634	V2$(VO)(P_2O_7)_2$

115543	$Rb_2CaP_2O_7$	51095	$Zn_2P_2O_7$
208899	$Rb_2CuP_2O_7$	166875	$Pd_2P_2O_7$
425511	$Rb_2Cu_3(P_2O_7)_2$	75598	$Rb(NbO)P_2O_7$
14836	$Rb_2MgP_2O_7$	177785	$RbLiMgP_2O_7$
131707	$Rb_2PbP_2O_7$	31004	$Sr_2P_2O_7$
19047	SiP_2O_7	44811	TiP_2O_7
418084	$Na_4Mg_3(PO4)2P_2O_7$	236316	$Na_4Fe_3(PO_4)_2P_2O_7$
65673	$RbMoP_2O_7$		
Hydrates: 78/2741			
59837	$SnCl_4·3H_2O$	201842	$Cu_3(PO_4)_2·H_2O$
402119	$AgClO_4·H_2O$	82459	$Hf(HPO_4)_2·H_2O$
63173	$Ba(VO_3)_2·H_2O$	15404	$HgSO_4·H_2O$
20332	$BaBr_2·H_2O$	192268	$Ho(NO_3)_3·H_2O$
100847	$CaCO_3·H_2O$	79714	$LiBrO_4·H_2O$
151960	$Ca6(SiO_3)_6·H_2O$	37180	$LiNO_2·H_2O$
408100	$CoSeO_3·H_2O$	6302	$MgSO_4·H_2O$
23500	$CdSO_4·H_2O$	68563	$CoHPO_3·H_2O$
59347	$CdSeO_3·H_2O$	62220	$MnPO_4·H_2O$
260000	$Cd_4(SO_3)_4·6H_2O$	100781	$CuHPO_4·H_2O$
960	$CaCl_2·2H_2O$	9924	$HfF_4·3H_2O$
409114	$Ni(ClO_4)_2·8H_2O$	136862	$Ho(ReO_4)_3·2H_2O$
363	$(VO_2)IO_3·2H_2O$	14060	$HoIO_5·4H_2O$
39800	$Ba(ReO_4)_2·4H_2O$	280068	$InPO_4·2H_2O$
36587	$Be_2(OH)(PO_4)·4H_2O$	20020	$KI·2H_2O$
181191	$Be_3(AsO_4)_2·2H_2O$	236189	$K_4P_2O_6·4H_2O$
151682	$Ca_2CrO_5·3H_2O$	62001	$K_4V_2O_7·2H_2O$
280410	$Ca_2P_2O_7·4H_2O$	46784	$LuPO_4·2H_2O$
132781	$CdSeO_4·2H_2O$	24250	$Mg(ClO_4)_2·6H_2O$
11510	$Co_2(As_2O_7)·2H_2O$	88518	$Mg(HSeO_3)_2·3H_2O$
67374	$CsNa_2(OH)_3·6H_2O$	120226	$MgCO_3·6H_2O$
261584	$Fe_2(SO_4)_3·3H_2O$	119912	$MgSO_4·5H_2O$
20760	$TlCl_3·4H_2O$	433807	$MnC_2O_4·3H_2O$
142131	$Y(NO_3)_3·5H_2O$	193725	$MnHPO_4·3H_2O$
32265	$Zn_4(OH)_6SO_4·4H_2O$	423300	$MnSO_4·3H_2O$
290924	$ZrF_4·3H_2O$	81423	$MnV_2O_6·2H_2O$
85565	$VPO_4·H_2O$	10204	$Ni(IO_3)_2·2H_2O$

79356	$YbI_2 \cdot H_2O$	410916	$Pr(ClO_3)_3 \cdot 2H_2O$
125343	$Zr(HPO_4)_2 \cdot H_2O$	61171	$Rb_3VO_4 \cdot 4H_2O$
20494	$Zr(SO_4)_2 \cdot H_2O$	421778	$Rh_2(SO_4)_3 \cdot 2H_2O$
74810	$MnSO_3 \cdot H_2O$	126200	$ScPO_4 \cdot 2H_2O$
66746	$MnSeO_4 \cdot H_2O$	14059	$SmIO_5 \cdot 4H_2O$
95893	$Na_3HP_2O_7 \cdot H_2O$	59816	$Sr_S2O_3 \cdot 5H_2O$
74049	$NiHPO_3 \cdot H_2O$	187024	$SrV_3O_7 \cdot 4H_2O$
37083	$SrBr_2 \cdot H_2O$	136861	$Tb(ReO_4)_3 \cdot 2H_2O$
37082	$SrCl_2 \cdot H_2O$	132532	$TlAsO_4 \cdot 2H_2O$
132783	$MnSeO_4 \cdot 2H_2O$	26744	$MgCl_2 \cdot 6H_2O$
9300	$Mn_2P_2O_7 \cdot 2H_2O$	16566	$NaBr \cdot 2H_2O$
419040	$NaBH_4 \cdot 2H_2O$	425696	$NaClO_4 \cdot 2H_2O$
Halides: 41/2138			
56553	AgI	22401	HgI_2
15707	BaI_2	71062	In_7Cl_9
31696	$BeCl_2$	23126	InBr
135621	BiI_3	425449	InCl
56768	$CaBr_2$	60402	KCl
31582	$CeBr_3$	113572	Mn_2F_5
81674	CeF_3	12167	MnF_2
59971	CrF_3	242167	Mo_6I_{16}
52171	FeF_3	28535	Nb_6Cl_{14}
241170	Hg_2I_4	72190	NdI_2
144380	$HgBr_2$	10510	PrI_2
14146	ZnF_2	22073	$PtCl_4$
23163	$ZrCl_3$	14217	$SbBr_3$
15917	$ZnCl_2$	425148	WCl_6
412110	$SbCl_5$	52159	TiF_3
85526	$SiCl_2$	173785	TiI_3
63865	$SnBr_2$	109143	TlCl
31093	SnI_4	30268	TlF
15972	$SrBr_2$	18029	TlF_3
203137	SrI_2	26056	$TeCl_4$
167471	TbF_3		

The first observation of the table is that the screening yielded a total of 521 unique hits, reflecting a significant narrowing down from the initial 14,101 structural entries, and

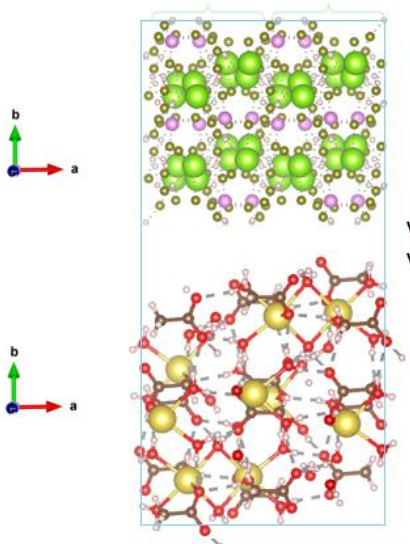
thus underscoring the strong selectivity imposed by the geometric interface-matching model. Among the six chemical groups, phosphates emerged as the most successful category, with 214 hits, representing over 40% of the total. This high hit rate is likely attributed to their inherent structural diversity, high symmetry, and frequent occurrence of orthophosphate building blocks that result in a wide range of cleavable surface geometries. The sulfates and pyrophosphates also contributed substantially, producing 91 and 67 viable nucleator candidates, respectively. Both groups are chemically and structurally related to phosphates, sharing similar backbone anions and coordination geometries, which may explain their success under the current matching algorithm. Hydrates and halides, while well represented in the initial database, returned fewer matches (78 and 41 respectively), likely due to mismatches in cleaved surface areas or angle criteria. In contrast, nitrates, although comprising a modest initial pool (779 entries), produced 30 hits, suggesting that while they are not the most abundant in the source database, certain nitrate structures are nonetheless geometrically well-suited to SAT interfaces.

In addition to satisfying geometric constraints, an important practical consideration is the chemical availability of shortlisted compounds. Compounds marked in green underline in Table 4.2 indicate positive identification in the Chemical Availability Search (ChASe), flagging them as commercially purchasable. These entries are of immediate interest for follow-up experimental screening, as they offer a low barrier to laboratory validation. Notably, most categories retained a healthy proportion of purchasable hits, especially within the phosphate and sulfate groups—implying that geometric compatibility is not confined to exotic or obscure compounds. In many cases, multiple hits are seen within the same chemical family, indicating the potential for chemical substitution strategies or formulation of mixed systems to explore synergistic nucleation effects. For instance, several sodium-based orthophosphates and calcium-based sulfates feature prominently, offering a balance between availability, cost-effectiveness, and predicted compatibility.

These screening results provide a robust foundation for experimental prioritisation. By coupling the prediction model with structural and commercial filters, this work narrows the candidate space to a manageable and chemically meaningful list of

nucleators. Furthermore, the concentration of hits within particular structural classes suggests that future refinements to the model could consider category-specific descriptors or chemistry-informed pre-screening, particularly for compounds that share bonding motifs or coordination environments with previously validated nucleators. The successful outcome of this screening also points to the transferability of the predictor model from ice nucleation to industrially relevant PCMs such as SAT. The flexibility of the framework makes it feasible to extend to other materials of interest, with minimal retraining required once a sufficiently representative dataset of known nucleators is available.

Another significant discovery is that the high throughput process found another nucleator that was already known by SUNAMP Ltd and is used as part of their commercial formula[27]. This is $Na_2(HPO_4)\cdot 12H_2O$, which was not part of training dataset, and therefore acts as a further positive validation point for prediction model. It should be noted that disodium hydrogen-phosphate hydrates exhibit excellent nucleator properties for SAT that appear to be insensitive to the level of hydration, as both the dihydrate and hepta-hydrate forms have also been reported as NUCs for SAT[13]. These compounds were successfully predicted in training data sets, while Table 4.2 also lists the decahydrate form as a predicted successful NUC. In an effort to understand this effect further the geometries of the most successful $Na_2(HPO_4)\cdot nH_2O$/SAT slabs are shown in Figure 4.6, from which it is apparent that the patterns of repeat of the disodium hydrogen-phosphates units are unaffected by the presence of varying amounts of intercalated water molecules. Thus, despite disodium hydrogen-phosphate hydrates readily interconverting through thermal dehydration and condensation[36], the templating properties of the NUC slabs are retained, making this a particularly robust nucleator for SAT.

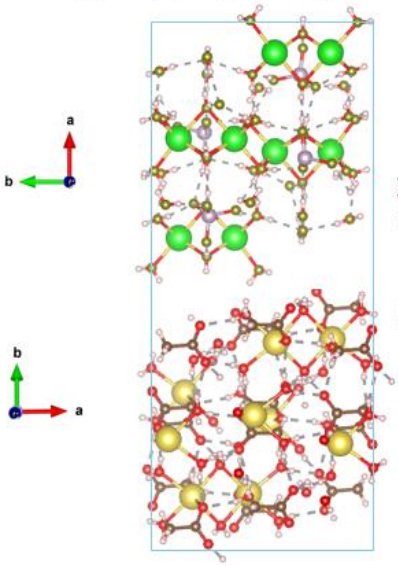


vector_mismatch_u = 0.25%
vector_mismatch_v = 2.07%

SAT(010), viewing along c-axis

(a)

$Na_2(HPO_4)\cdot 7H_2O$ (100), viewing along c-axis

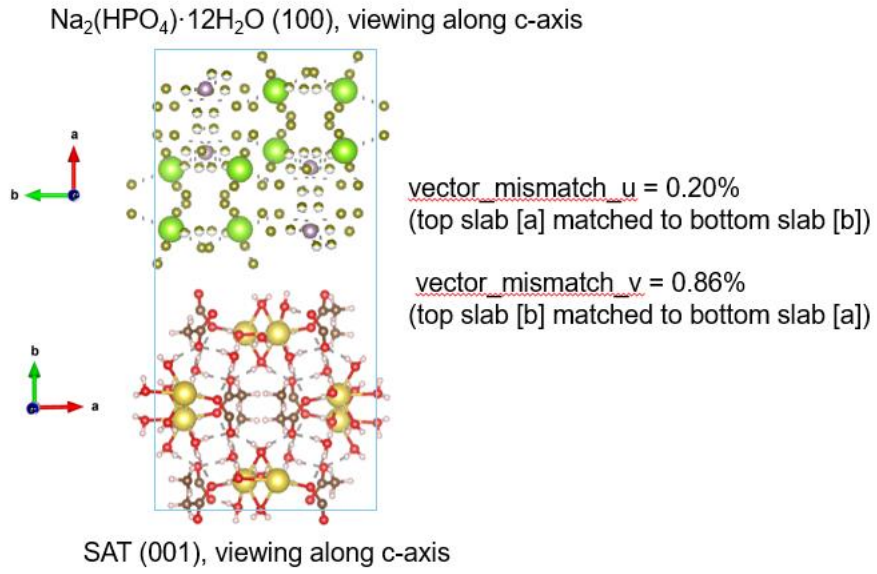


vector_mismatch_u = 1.40%
(top slab [a] matched to bottom slab [b])

vector_mismatch_v = 0.07%
(top slab [b] matched to bottom slab [a])

SAT (010), viewing along c-axis

(b)

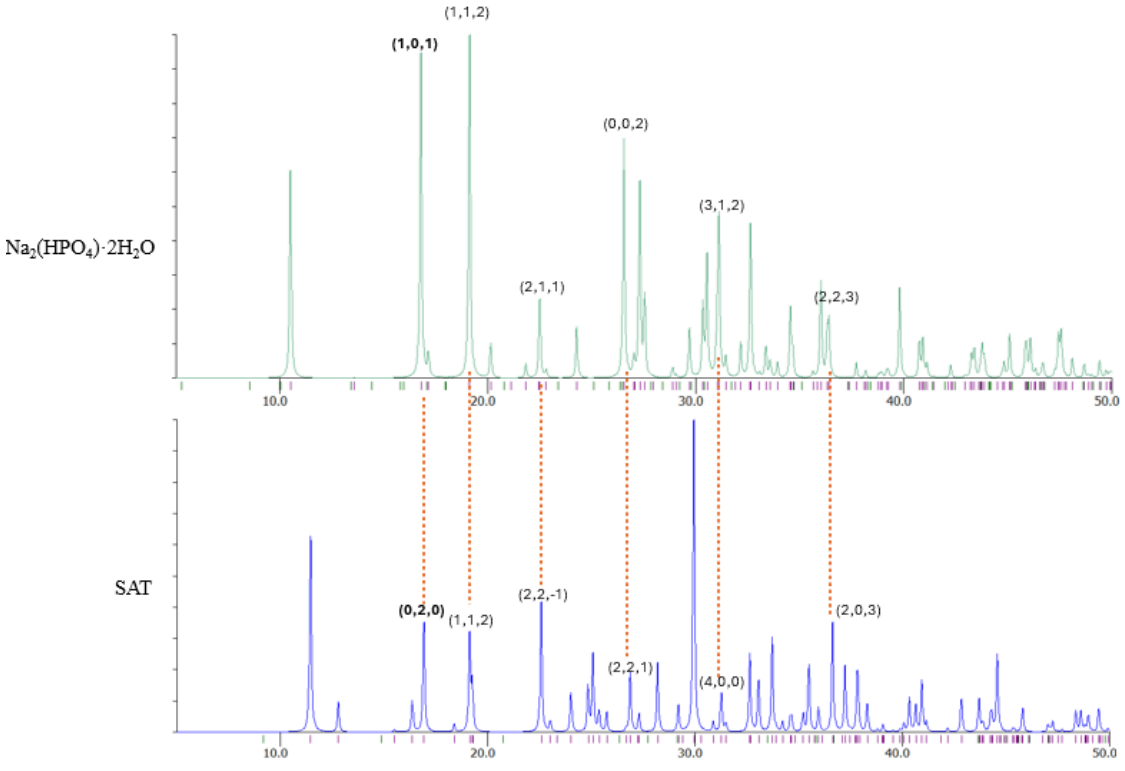


(c)

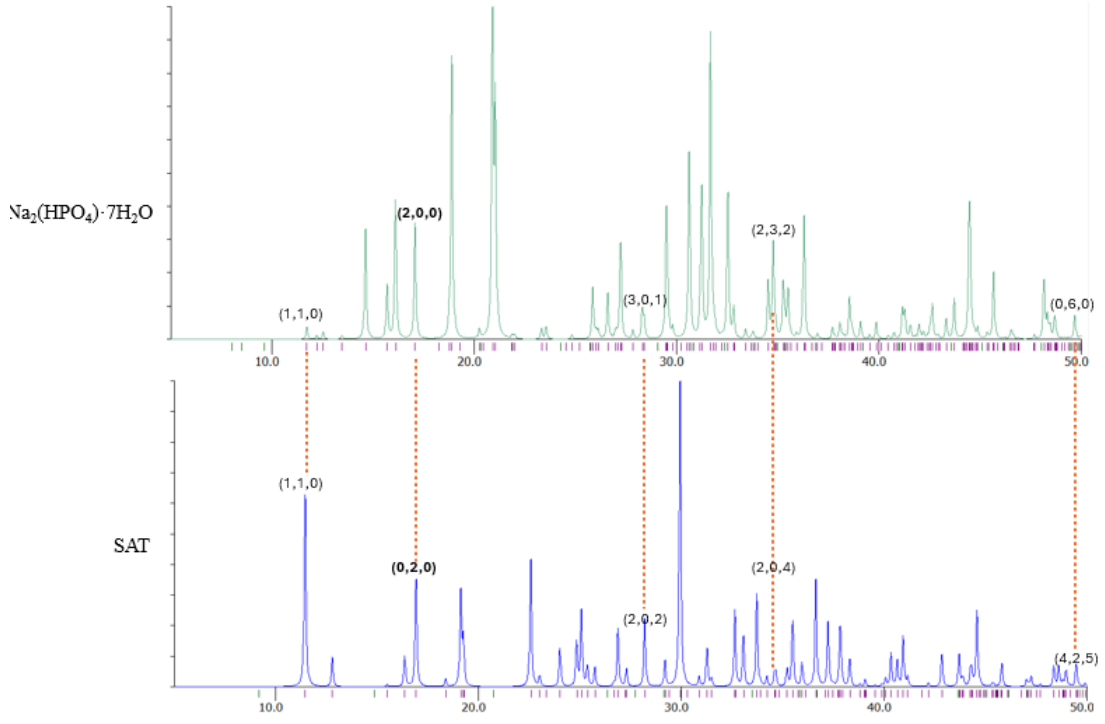
Figure 4.6. Examples of $Na_2(HPO_4)\cdot xH_2O$/SAT supercells that fulfil the strict geometric matching criteria, where **(a)** x = 2, **(b)** x = 7, **(c)** x = 12.

The way geometric matching of oriented slabs from NUC and PCM plays an important part in heterogeneous nucleation of inorganic salts at an atomic level was also worth looking into, at which the process of heterogeneous nucleation on a substrate can be considered to be atom-by-atom building of the initial solid phase on a template. Therefore, the lattice matching across the solid/substrate interface should play an important role in this process. Relevant to this process is epitaxial growth of a thin layer of one material (the PCM for example) on the surface of another material (the substrate, the NUC for example). The scientific basis for epitaxial growth of strained layers on a substrate originates from the theory of Frank and Van der Merwe[37]. In epitaxial systems, a coherent crystalline layer can grow on a substrate if the lattice mismatch between specific crystallographic planes is minimal and the interplanar spacings (d-spacings) are sufficiently similar to allow structural continuity. To further validate the physical relevance of the slab matching results, their consistency with classical epitaxial growth theory was explored. This condition is often detectable through powder X-ray diffraction (XRD), where overlapping diffraction peaks between two materials indicate shared or harmonically related d-spacings, therefore the output of geometric slab matching with simulated XRD patterns of both the phase

change material (PCM) and nucleating agent (NUC) were cross referenced. When a matched pair of surfaces, for example $(hkl)_{PCM}$ and $(h'k'l')_{NUC}$, exhibits not only high geometric compatibility but also overlapping XRD peaks at the corresponding d-spacing, this suggests a strong theoretical basis for epitaxial alignment. Such dual-consistency, in both real space (surface orientation and lattice vector alignment) and reciprocal space (Bragg reflection positions), strengthens the plausibility that these interfaces represent favourable pathways for heterogeneous nucleation. This integrated criterion thus allows us to flag "epitaxially probable" matches, which may be prioritised for further investigation, including experimental validation or inclusion in predictive models.


Na2(HPO4)·2H2O
(1,1,2)
(1,0,1)
(0,0,2)
(3,1,2)
(2,1,1)
(2,2,3)
10.0
20.0
30.0
40.0
50.0
SAT
(0,2,0)
(1,1,2)
(2,2,-1)
(2,0,3)
(2,2,1)
(4,0,0)
10.0
20.0
30.0
40.0
50.0

(a)


Na2(HPO4)·7H2O
(2,0,0)
(2,3,2)
(1,1,0)
(3,0,1)
(0,6,0)
10.0
20.0
30.0
40.0
50.0
SAT
(1,1,0)
(0,2,0)
(2,0,4)
(2,0,2)
(4,2,5)
10.0
20.0
30.0
40.0
50.0

(b)

Figure 4.7. Simulated powder X-ray diffraction (XRD) patterns for **(a)** SAT (bottom) and $Na_2HPO_4{\cdot}2H_2O$ (top), (b) SAT (bottom) and $Na_2HPO_4{\cdot}7H_2O$ (top), shown with peak indexing. Dotted lines indicate overlapping reflections with matching d-spacings. Miller indices for selected peaks are annotated, and d-spacing values may be included to support quantitative comparison. Overlapping peaks with corresponding geometrically matched planes are candidates for epitaxial alignment, in accordance with coherent nucleation theory.

To this end, XRD patterns was compared between PCM and NUC in order to find (hkl) faces with the same d-spacing, which is a prerequisite for epitaxial growth in terms of heterogeneous nucleation process (see examples given in Figure 4.7). Next, the indexed overlapping peaks were cross-referenced with geometric matching screening results to determine if there are coincident matches of (hkl) faces between NUC and PCM. If a match is found then it is likely that nucleation can proceed at the highlighted interface via the epitaxial growth model. As shown in Figure 4.5 and summarised in Table 4.3, six overlapping peaks with very similar 2θ values (to within 0.01°) were identified between SAT and $Na_2HPO_4{\cdot}2H_2O$, along with five overlapping peaks between SAT and $Na_2HPO_4{\cdot}2H_2O$, according to the PDF-2 202032 database cards.

The next step would be to identify which planes were matching planes between PCM and NUC according to purely geometric matching concerns. Miller indices values of slabs comprising the matching supercell of SAT-$Na_2HPO_4{\cdot}7H_2O$ interface system were listed in Table 4.1. Cross-referencing Table 4.1 and Table 4.3, the matchings of (hkl) values of planes that satisfy both requirements: i) identical d-spacings; ii) geometrically matched, shown in column 8 of Table 4.3. The nucleator planes could very much likely be the true planes that SAT crystals grow upon with the preferred orientation of planes respectively, because these plane matchings satisfy two prerequisites according to epitaxial growth theories[38,39]: i) geometry similarities; ii) identical d-spacings. Note the appearance of negative values of (hkl) in XRD pattern while no negative values were shown in geometric matching process is because the negative values were automatically contained in the geometric matching process, in that $(22\bar{1})$ for example is geometrically symmetrical to (221) in geometric concerns, and thus not included in case of an independent result.

Table 4.3. Cross-comparison of simulated XRD patterns and geometric slab matching between SAT and $Na_2HPO_4 \cdot 2H_2O$/ $Na_2HPO_4 \cdot 7H_2O$. Peaks with closely matching 2θ positions are identified, and corresponding d-spacings are listed for both materials. The d-spacing mismatch (Δd) is calculated as a percentage. Matched (hkl) indices are shown alongside a qualitative assessment of their geometric compatibility, leading to a final evaluation of each pair's plausibility as an epitaxial interface candidate.

Nucleator	XRD Peak (2θ)	d-spacing (SAT, Å)	d-spacing (NUC, Å)	Δd(%)	SAT (hkl)	NUC (hkl)	Geometric match?
$Na_2HPO_4 \cdot 2H_2O$	16.90(6)	5.241	5.232	0.180	(020)	(101)	yes
	19.17(7)	4.624	4.609	0.334	(112)	(112)	no
	22.56(2)	3.938	3.915	0.577	(22-1)	(211)	no
	26.86(3)	3.316	3.337	-0.627	(221)	(002)	no
	31.24(3)	2.861	2.882	-0.749	(400)	(312)	no
	36.69(5)	2.447	2.450	-0.129	(203)	(223)	no
$Na_2HPO_4 \cdot 7H_2O$	11.45(1)	7.721	7.735	-0.177	(110)	(110)	no
	16.90(5)	5.241	5.237	0.067	(020)	(200)	yes
	28.17(7)	3.164	3.160	0.142	(202)	(301)	no
	34.62(8)	2.589	2.583	0.205	(204)	(232)	no
	49.44(2)	1.842	1.846	-0.220	(425)	(060)	no

An intriguing ambiguity arises when attempting to reconcile geometric surface matching with crystallographic evidence from X-ray diffraction (XRD). For instance, a geometric match may occur between the (101) surface of $Na_2HPO_4 \cdot 2H_2O$ and the (010) surface of SAT, but the corresponding (010) reflection may be absent in XRD due to systematic absences as a consequence of the space group symmetry; $Na_2HPO_4 \cdot 2H_2O$ belongs to space group P*2_1/n*, for which the (010) reflection is systematically absent due to the 2_1 screw axis along the b-axis. In such scenarios, higher-order reflections like (020), which are parallel to (010) but possess half the d-spacing, may still appear in the pattern, offering indirect evidence of the underlying orientation. This raises a philosophical tension: should (010) and (020) be considered as distinct surfaces, or as manifestations of the same crystallographic orientation viewed through different lenses? In reciprocal space, they are clearly distinct, with unique d-spacings and diffraction intensities. Yet in real space, they represent the same family of planes, differing only by periodicity, implying that epitaxial interaction may still occur on the (010) plane even if only (020) is observed. This geometric matching framework, grounded in real-space analysis, is therefore capable of identifying interfacial relationships that XRD alone may obscure. Rather than resolving this

ambiguity, it is preserved here as a reminder that systematic absences in reciprocal space do not imply the absence of corresponding real-space lattice planes; diffraction visibility reflects symmetry-imposed structure factor conditions rather than the physical non-existence of a crystallographic orientation.

The link between the geometric matching screening process and XRD d-spacing overlap analysis suggests that an epitaxial growth model for heterogenous nucleation of inorganic salts is plausible, and thus may shed light on the nucleation mechanism. The value and impact of this analysis also lie in the fact that very little is known about interfaces of supercells from heterogenous nucleation of inorganic salt crystals, especially the Miller indices of matching planes of PCMs and NUCs. While *e.g.* experimental tomography measurements have been undertaken for ice nucleation[40,41], the same in-depth analysis for salt hydrate nucleation has not been done. While modelling simulations using, *e.g.* long-timescale quantum mechanical molecular dynamics simulations could be undertaken they are time consuming and costly in terms of computational resources required. This utilisation of the high-throughput screening calculation makes it possible to clarify matching planes in real nucleation processes, which is valuable new insight for industrial applications, where the aide of crystal habit modifiers could potentially change the morphology of crystals to one that PCMs can crystallise upon more readily[42-46].

4.3. Transferring the modified model to five other salt hydrates

4.3.1. Literature search on potential nucleators of $CaCl_2·6H_2O$, $Mg(NO_3)_2·6H_2O$, $LiNO_3·3H_2O$ and $MgCl_2·6H_2O$

The workflow documented above was developed for SAT, and thus the next stage is to identify whether it can be transferred across to other salt hydrates. If this can be done without further modification this suggests that the procedure is more widely applicable to salt hydrates, in general, and potentially to other systems for which nucleating agents might be sought, *e.g.* directing polymorphism.

To this end, information is gained toward literature reporting on potential nucleators for four other common PCMs, namely (i) $CaCl_2·6H_2O$, (ii) $Mg(NO_3)_2·6H_2O$, (iii) $LiNO_3·3H_2O$ and (iv) $MgCl_2·6H_2O$. The results from this short literature survey are

given in Table 4.4. If running the prediction model trained for SAT on this new data set shows that the reported nucleators slab match with the corresponding PCM this adds further evidence that the geometric slab matching approach is generally applicable for more materials beyond what it was initially trained for. Crucially, examples of both working and non-working nucleators were found for each salt hydrate PCM, which provides a more rigorous test of the geometric slab matching prediction capability. Also of note is that many of the experimental studies were reported in the same papers, suggesting that variations in experimental set up, which could result in significant variations in the degree of subooling observed, will be minimised.

Table 4.4. Summary of experimentally reported nucleating agents for four salt hydrate PCMs ($CaCl_2 \cdot 6H_2O$, $Mg(NO_3)_2 \cdot 6H_2O$, $LiNO_3 \cdot 3H_2O$, and $MgCl_2 \cdot 6H_2O$), including NUC names, added weight percentage, corresponding subcooling reduction efficacy, and literature source. is NUCs that were experimentally reported to be working and predicted as working, is NUCs that were experimentally reported to be working but predicted as non-working, is NUCs that were experimentally reported to be non-working and predicted as non-working, and is NUCs that were experimentally reported to be non-working but predicted as working.

PCM	NUC	Weight percentage of NUC added	Subcooling reduction (%)	Ref.
$CaCl_2 \cdot 6H_2O$	$SrCl_2 \cdot 6H_2O$	1 wt%	100.0	Lane[47]
	$BaCl_2$	0.5 wt%	100.0	Lane[47]
	K_2CO_3	0.5 wt%	100.0	Sutjahja et al[48]
	$BaCO_3$	0.5 wt%	100.0	Lane[47]
	$Ba(OH)_2$	0.1 wt%	98.3	Lane[47]
	$BaSO_4$	0.1 wt%	98.3	Lane[47]
	BaO	0.1 wt%	97.5	Lane[47]
	γ-Al_2O_3	0.5 wt%	97.0	Li et al[49]
	$SrBr_2$.6H2O	1 wt%	90.6	Lane[47]
	$Ba(OH)_2 \cdot 8H_2O$	1 wt%	90.0	Cao et al[50]
	$BaCl_2 \cdot 2H_2O$	2 wt%	90.0	He et al[51]
	$Sr(OH)_2$	0.1 wt%	79.0	Lane[47]
	$SrCO_3$	1 wt%	65.0	Abhat[52]
	BaF_2	1 wt%	65.0	Abhat[52]
	SrF_2	1 wt%	65.0	Abhat[52]
	KNO_3	3 wt%	45.0	Blen et al[53]

	BaI_2	1 wt%	non-working	Lane[47]
	$CaI_2·6H_2O$	1 wt%	non-working	Lane[47]
$Mg(NO_3)_2·6H_2O$	$Ca(OH)_2$	0.1 wt%	100.0	Lane[47]
	$CaCO_3$	0.1 wt%	100.0	Lane[47]
	$MgCO_3$	0.1 wt%	100.0	Lane[47]
	$SrCO_3$	0.1 wt%	100.0	Lane[47]
	$BaCO_3$	0.1 wt%	99.0	Lane[47]
	MgO	0.1 wt%	98.9	Lane[47]
	$Mg(OH)_2$	0.1 wt%	98.5	Lane[47]
	$CaSO_4·2H_2O$	0.5 wt%	92.2	Lane[47]
	$CuNa_2(SO_4)_2·2H_2O$	0.5 wt%	90.3	Lane[47]
	$Mg_3(PO_4)_2$	0.1 wt%	76.7	Lane[47]
	$ZnSO_4·4H_2O$	0.5 wt%	74.7	Lane[47]
	BaF_2	0.1 wt%	59.1	Lane[47]
	$BaSO_4$	0.1 wt%	54.5	Lane[47]
	$BaHPO_4$	0.1 wt%	50.9	Lane[47]
	TiO_2	0.5 wt%	47.5	Gupta et al[54]
	ZnO	0.5 wt%	45.0	Gupta et al[54]
	$Ba(NO_3)_2$	0.1 wt%	35.1	Lane[47]
	$BaCl_2$	0.1 wt%	30.6	Lane[47]
	Fe_2O_3	0.1 wt%	29.7	Lane[47]
	AlO(OH)	1 wt%	25.7	Honcova et al[55]
	BaO	1 wt%	14.2	Honcova et al[55]
	$CuSO_4.3H_2O$	0.5 wt%	11.1	Lane[47]
	$Ba(OH)_2·8H_2O$	1 wt%	non-working	Honcova et al[55]
	BaO_2	1 wt%	non-working	Honcova et al[55]
	$CaHPO_4·2H_2O$	0.1 wt%	non-working	Lane[47]
	CaO	1 wt%	non-working	Honcova et al[55]
	$KF·4H_2O$	1 wt%	non-working	Lane[47]
	$MgCl_2·2H_2O$	1 wt%	non-working	Lane[47]
	$MgSO_4·4H_2O$	1 wt%	non-working	Lane[47]
	$Sr(OH)_2$	1 wt%	non-working	Honcova et al[55]
$LiNO_3·3H_2O$	$Zn_3(OH)_4(NO_3)_2$	3 wt%	93.8	Kannan et al[56]
	$Cu_3(OH)_5(NO_3)·2H_2O$	5 wt%	91.3	Shamberger et al[57]
	$Zn(NO_3)_2.6H_2O$	3 wt%	90.0	Kannan et al[56]
	$FeSb_2$	3 wt%	80.0	Shamberger et al[57]
	$ZrSiO_4$	3 wt%	75.0	Shamberger et al[57]

	SiO_2	3 wt%	69.9	Shamberger et al[57]
	TiO_2	3 wt%	69.0	Shamberger et al[57]
	$Zn(NO_3)_2 \cdot 2[Zn(OH)_2]$	0.9 wt%	23.5	Shamberger et al[57]
	$Zn(NO_3)_2 \cdot 4[Zn(OH)_2] \cdot 2H_2O$	3 wt%	12.5	Shamberger et al[57]
	$Cu(NO_3)(OH)_2[Cu(OH)_2] \cdot 2H_2O$	3 wt%	non-working	Shamberger et al[57]
	$Cu_2(NO_3)(OH)_3$	3 wt%	non-working	Shamberger et al[57]
$MgCl_2 \cdot 6H_2O$	$CaC_2O_4 \cdot H_2O$	0.5 wt%	100.0	Lane[47]
	$CaK_2(SO_4)_2 \cdot H_2O$	0.5 wt%	51.5	Lane[47]
	Na_3AlF_6	0.5 wt%	91.9	Lane[47]
	$Ba(OH)_2$	0.5 wt%	100.0	Lane[47]
	BaO	0.5 wt%	100	Lane[47]
	$Ca(OH)_2$	0.5 wt%	97.4	Lane[47]
	CaO	0.5 wt%	100.0	Lane[47]
	$LiOH \cdot H_2O$	0.5 wt%	91.4	Lane[47]
	$Mg(OH)_2$	0.5 wt%	100.0	Lane[47]
	$Sr(OH)_2$	0.5 wt%	100.0	Lane[47]
	$SrCO_3$	0.5 wt%	100	Lane[47]
	$MgBr_2 \cdot 6H_2O$	0.5 wt%	non-working	Lane[47]
	$AgNO_3$	0.5 wt%	non-working	Lane[47]
	$(NH_4)_2SO_4 \cdot NH_4NO_3$	0.5 wt%	43.9	Lane[47]
	$Ca(NH_4)_2(SO_4)_2 \cdot H_2O$	0.5 wt%	49.4	Lane[47]
	$(NH_4)_2SO_4$	0.5 wt%	48.5	Lane[47]
	K_2SO_4	0.5 wt%	non-working	Lane[47]
	K_2CrO_4	0.5 wt%	40.4	Lane[47]
	$CaSO_4$	0.5 wt%	15.2	Lane[47]

A closer examination of Table 4.4 reveals several important trends that support the generalisability of the feature set used in this model. Firstly, across different PCMs, many compounds that are highly effective in subcooling reduction, such as $BaCl_2$, $Ba(OH)_2$, $BaCO_3$, and $SrCl_2$ share compositional or structural similarities with the original PCM host. These additives frequently appear across multiple systems, suggesting a potentially transferable mechanistic basis for nucleation promotion. Secondly, certain anionic species, such as carbonate, hydroxide, and sulphate, repeatedly demonstrate efficacy despite varying cation identities, indicating that anionic match or interaction with the host lattice may play a significant role.

The table also distinguishes between working and non-working additives/nucleators, many of which are compositionally or structurally dissimilar from the host. This binary outcome provides clear labelling that can be used for future classification model training or for retrospective model testing using derived features. The variation in weight percentages further allows investigation into concentration sensitivity, although for now the presence or absence of subcooling suppression is treated as a binary indicator.

4.3.2. Results

The potential of the materials working as a nucleator is arranged from left to right (strong to weak) according to subcooling reduction degrees (a large % degree of subcooling reduction specifies a good nucleator) on the y-axis on the left in a dot-and-line fashion on Figure 4.8. These NUCs and PCMs were then put into the geometric matching process to generate matching model hit histograms; the results from this process are also shown on Figure 4.8.

The values of the six features were the same as used for SAT. Therefore, the transferable values are as follows:

$$\text{Maximum area overlap} = \text{<variable, PCM dependent>}\ \text{Å}^2 \quad (8)$$

$$\text{Maximum angle mismatch} = 0.01° \quad (9)$$

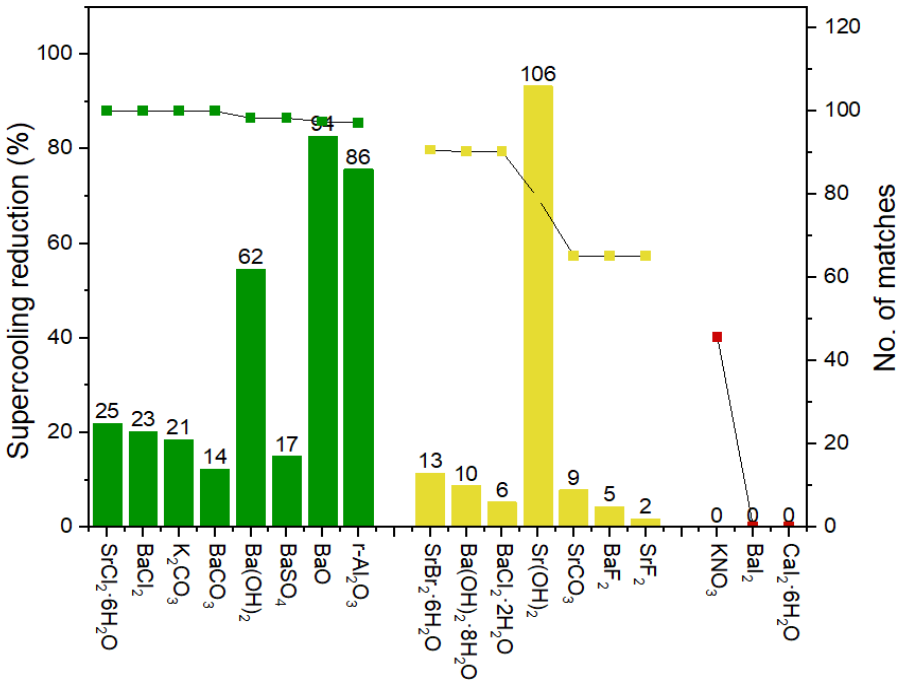
$$\text{Maximum supercell vector mismatch [0]} = 0.01 \quad (10)$$

$$\text{Maximum supercell vector mismatch [1]} = 0.01 \quad (11)$$

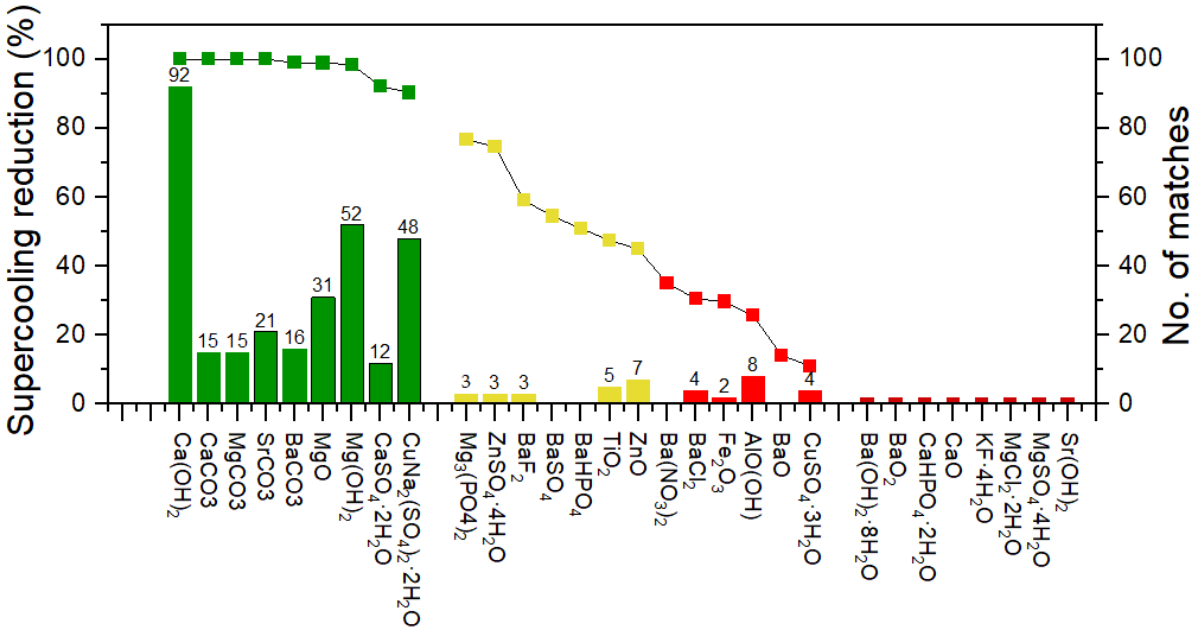
$$m\_n\_\text{tolerance} = 0.02 \quad (12)$$

$$\text{slab area tolerance} = 1.5 \quad (13)$$

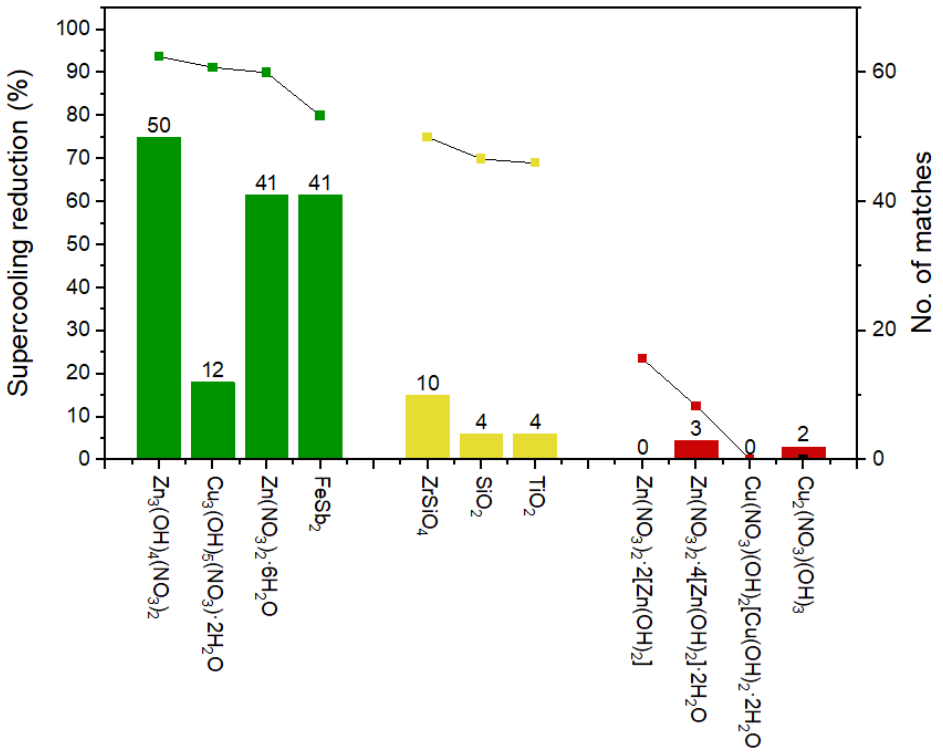
The numerical values adopted for feature 1, i.e. the Maximum area overlap, are PCM dependent, and thus is the only feature that needed to be changed for each slab matching run. For $CaCl_2 \cdot 6H_2O$ a value of 610 Å^2 was chosen (based on the ab plane). Following a similar logic, for $Mg(NO_3)_2 \cdot 6H_2O$ a value of 790 Å^2 was chosen (based on the ab plane), for $LiNO_3 \cdot 3H_2O$ a value of 750 Å^2 was chosen (based on the ab plane), for $MgCl_2 \cdot 6H_2O$ a value of 700 Å^2 was chosen (based on the ab plane).


Supercooling reduction (%)
No. of matches
25
23
21
14
62
17
94
86
13
10
6
106
9
5
2
0
0
0
SrCl2·6H2O
BaCl2
K2CO3
BaCO3
Ba(OH)2
BaSO4
BaO
r-Al2O3
SrBr2·6H2O
Ba(OH)2·8H2O
BaCl2·2H2O
Sr(OH)2
SrCO3
BaF2
SrF2
KNO3
BaI2
CaI2·6H2O

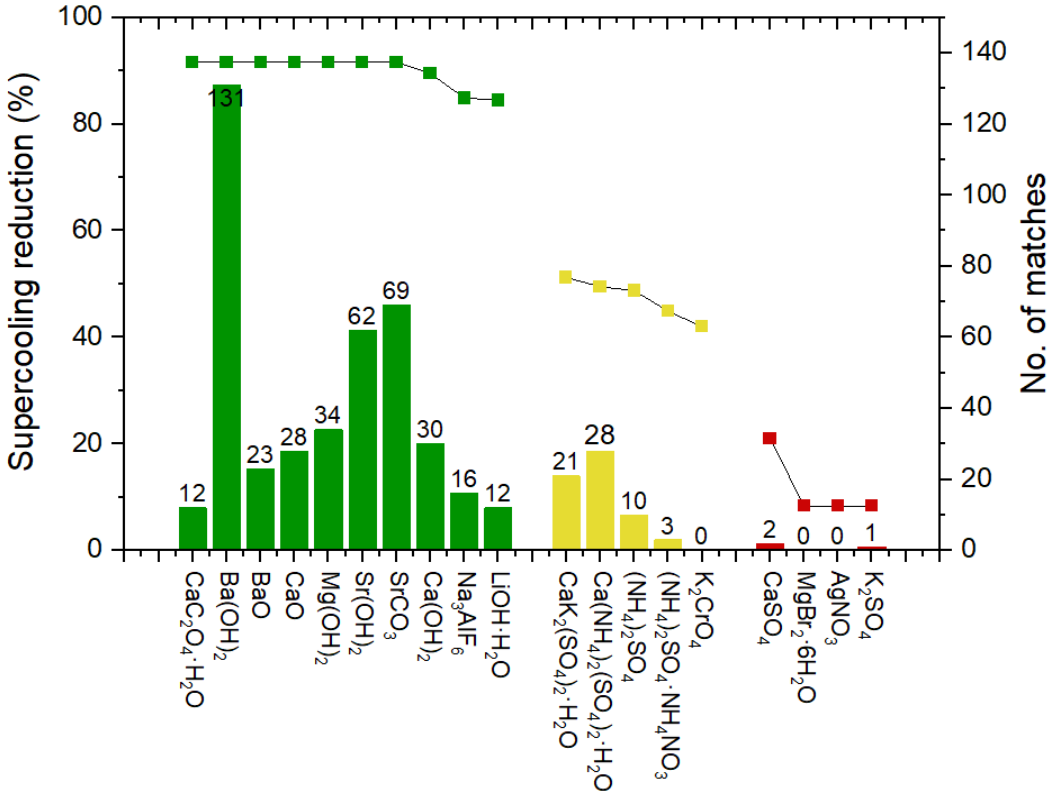
(a)


Supercooling reduction (%)
No. of matches
92
15
15
21
16
31
52
12
48
3
3
3
5
7
4
2
8
4
Ca(OH)2
CaCO3
MgCO3
SrCO3
BaCO3
MgO
Mg(OH)2
CaSO4·2H2O
CuNa2(SO4)2·2H2O
Mg3(PO4)2
ZnSO4·4H2O
BaF2
BaSO4
BaHPO4
TiO2
ZnO
Ba(NO3)2
BaCl2
Fe2O3
AlO(OH)
BaO
CuSO4·3H2O
Ba(OH)2·8H2O
BaO2
CaHPO4·2H2O
CaO
KF·4H2O
MgCl2·2H2O
MgSO4·4H2O
Sr(OH)2

(b)


Supercooling reduction (%)
No. of matches
50
12
41
41
10
4
4
0
3
0
2
$Zn_3(OH)_4(NO_3)_2$
$Cu_3(OH)_5(NO_3)\cdot 2H_2O$
$Zn(NO_3)_2\cdot 6H_2O$
$FeSb_2$
$ZrSiO_4$
SiO_2
TiO_2
$Zn(NO_3)_2\cdot 2[Zn(OH)_2]$
$Zn(NO_3)_2\cdot 4[Zn(OH)_2]\cdot 2H_2O$
$Cu(NO_3)(OH)_2[Cu(OH)_2\cdot 2H_2O$
$Cu_2(NO_3)(OH)_3$

(c)


Supercooling reduction (%)
No. of matches
12
131
23
28
34
62
69
30
16
12
21
28
10
3
0
2
0
0
1
$CaC_2O_4\cdot H_2O$
$Ba(OH)_2$
BaO
CaO
$Mg(OH)_2$
$Sr(OH)_2$
$SrCO_3$
$Ca(OH)_2$
Na_3AlF_6
$LiOH\cdot H_2O$
$CaK_2(SO_4)_2\cdot H_2O$
$Ca(NH_4)_2(SO_4)_2\cdot H_2O$
$(NH_4)_2SO_4$
$(NH_4)_2SO_4\cdot NH_4NO_3$
K_2CrO_4
$CaSO_4$
$MgBr_2\cdot 6H_2O$
$AgNO_3$
K_2SO_4

(d)

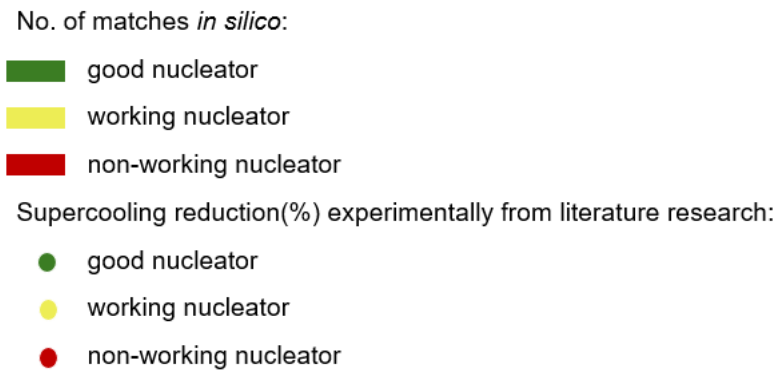


Figure 4.8. Correlation between geometric matching predictions and experimental performance of reported nucleators for four different salt hydrate PCMs: **(a)** $CaCl_2 \cdot 6H_2O$, **(b)** $Mg(NO_3)_2 \cdot 6H_2O$, **(c)** $LiNO_3 \cdot 3H_2O$, and **(d)** $MgCl_2 \cdot 6H_2O$. Bar heights represent the number of geometrically matched configurations as identified by *in silico* screening model. Overlaid markers denote experimentally reported subcooling reduction percentages from literature. Green bars and dots indicate "good nucleators" with strong experimental performance and high in silico match counts; yellow denotes moderate performers; red indicates additives classified as "non-working" in literature. The alignment between high subcooling reduction and a greater number of predicted matches highlights the transferability and cross-material validity of the SAT-derived feature set for nucleator screening.

The corresponding confusion matrices are shown in Figure 4.9, with predictions based on the rule that nucleators with >0 predicted matches are classified as “working” and those with 0 matches as “non-working.” Experimental ground truth was determined from immersion freezing tests, with ⩾40% subcooling reduction taken as evidence of nucleation activity. To contextualise performance, results are compared against two simple baselines: random guessing (50% expected accuracy for binary classification) and majority-class prediction, defined as assigning all candidates to the most frequent experimental class within each PCM system.

Working systematically through the four systems (a-d):

$CaCl_2 \cdot 6H_2O$ (a): The model showed the highest agreement, correctly capturing 15 working nucleators with only a single false negative. Notably, 100% of the experimentally non-working nucleators (KNO_3, BaI_2, $CaI_2 \cdot 6H_2O$) were also predicted

as non-working, yielding a clean separation. The model showed the highest agreement, correctly classifying 17 of 18 nucleators (94.4% accuracy). Balanced accuracy was 96.9%, reflecting both strong sensitivity (93.8%) and perfect specificity (100%). The majority-class baseline for this system was 88.9%, indicating that the model provides genuine improvement beyond class imbalance.

$Mg(NO_3)_2 \cdot 6H_2O$ (b): Performance was weaker, with 14 true positives, 2 false negatives, and 4 false positives, giving an accuracy of 80.0%. Out of 14 experimentally non-working nucleators, 10 (71%) were correctly identified with zero matches, while 4 produced spurious hits. This underscores the greater challenge of distinguishing weaker nucleators in this system, where the geometric criteria alone appear less discriminating. Performance was lower but still strong, with 24 correct predictions out of 30 (80.0% accuracy) and a balanced accuracy of 79.5%. The majority baseline was 53.3%, substantially below the model performance, demonstrating clear discriminative capability.

$LiNO_3 \cdot 3H_2O$ (c): The model performed well overall, correctly classifying all working nucleators (7 true positives, no false negatives). However, 2 nucleators that were experimentally non-working nevertheless gave non-zero matches, producing false positives. The model correctly classified 9 of 11 nucleators (81.8% accuracy), achieving perfect sensitivity toward working compounds but lower specificity (50%), resulting in a balanced accuracy of 75.0%. The majority-class baseline was 63.6%, which the model exceeded.

$MgCl_2 \cdot 6H_2O$ (d): Results were intermediate, with 14 true positives, 1 false negative, and 2 false positives, corresponding to an accuracy of 84.2%. Among the 4 experimentally non-working nucleators, 2 (50%) were correctly identified with zero matches, while 2 were misclassified as working. This partial separation suggests that $MgCl_2 \cdot 6H_2O$ is moderately well described by the model but still prone to false positives. Results were intermediate, with 16 of 19 correctly classified (84.2% accuracy) and a balanced accuracy of 71.7%, reflecting strong sensitivity but moderate specificity. The majority baseline (78.9%) was marginally lower than model performance.

Across all four systems, the geometric matching model consistently exceeded random guessing and majority-class prediction, although specificity varied depending on the PCM. This indicates that while geometric compatibility provides robust predictive signal, the degree of separation between working and non-working nucleators remains systemly dependent.

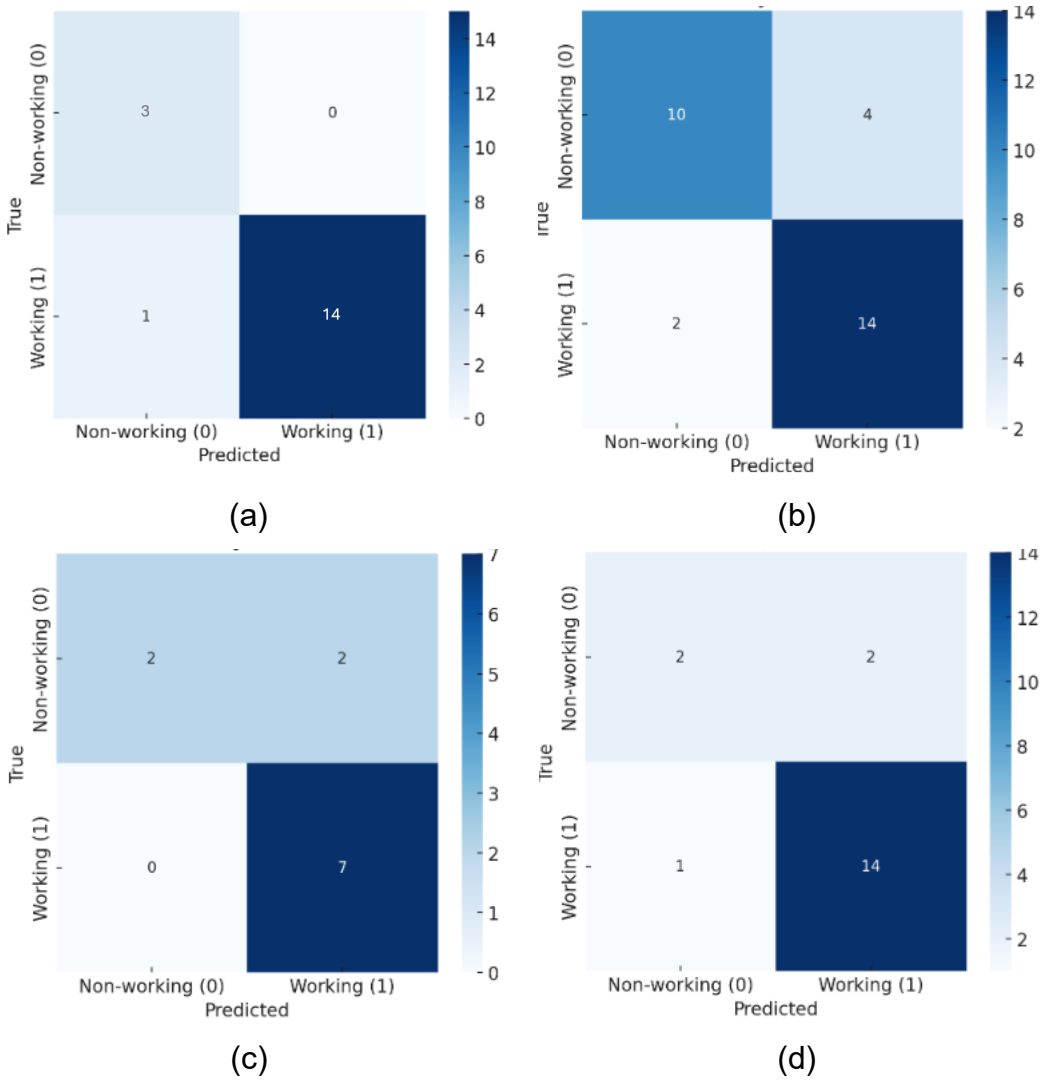


Figure 4.9. Confusion matrices for four representative PCM systems **(a)** $CaCl_2{\cdot}6H_2O$, **(b)** $Mg(NO_3)_2{\cdot}6H_2O$, **(c)** $LiNO_3{\cdot}3H_2O$, and **(d)** $MgCl_2{\cdot}6H_2O$.

This comparative visualisation highlights both consistencies and divergences. Strong nucleators such as $Ca(OH)_2$, $Ba(OH)_2$, and $SrCO_3$ appear on the left of each panel with both high experimental subcooling reduction (green dots) and high modelled match counts (green bars), reinforcing that the feature-based model captures essential

structural compatibilities underpinning nucleation. Conversely, known ineffective additives (red dots) almost always returned zero matches (red bars), giving a robust binary discriminant: across all four PCMs, an average of ~65% of non-working nucleators were correctly predicted with zero matches. This provides a useful semi-quantitative descriptor of model reliability in filtering out false positives.

Conversely, additives known to be ineffective (i.e., "non-working" nucleators from literature) show negligible or zero geometric matches in the model (red bars), supporting the model's discriminatory capacity. Thus, in terms of a binary outcome, geometric matching appears to offer a ‘good enough’ differentiator. Interestingly, some yellow-classified nucleators (i.e. intermediate subcooling nucleators) show moderate hit counts but lower subcooling performance, likely indicating a more complex situation, such as concentration sensitivity, kinetic limitations, or surface reconstruction, all of which are not explicitly captured by the purely geometric model.

Inevitably, there are a few outliers in the four panels, among which $Mg(NO_3)_2 \cdot 6H_2O$ paired with $ZnSO_4 \cdot 4H_2O$ is particularly notable. The data point appears mid-range in subcooling reduction (~74.7%) but yields a relatively low number of geometric matches in the prediction model (yellow bar, low height). Given its substantial experimental performance, this discrepancy suggests that the model may have underestimated its nucleation potential. One plausible explanation is that $ZnSO_4 \cdot 4H_2O$ may promote nucleation via a different crystallographic face than those prioritised in the slab-matching pipeline. If the default cleavage planes or surface terminations chosen for modelling do not correspond to the functional interface during nucleation, the match count would be artificially suppressed. A second possibility is that ionic or hydrogen-bonding effects dominate the nucleation mechanism for this compound, rather than purely geometric epitaxy. Zn^{2+} has a relatively high charge density and can form stable hydrate structures or coordinate with nitrate in solution, which may enhance nucleation through chemical templating or local field effects that the geometric model does not account for. This outlier serves as a reminder that while geometric compatibility is a strong indicator, it is not the sole determinant of nucleation behaviour, pointing to future work integrating surface chemistry descriptors or hydration dynamics into the model.

Importantly, the consistent trend across distinct PCM chemistries, ranging from chlorides to nitrates, suggests that the selected six features reflect geometrically universal descriptors of epitaxial nucleation likelihood. This robustness is particularly noteworthy given that no additional training or re-fitting was done; the same parameter set used for SAT predictions was directly applied to all four cases. Therefore, this test validates the model's practical applicability for high-throughput nucleator discovery across a broader chemical space, minimising the need for system-specific recalibration.

It is important to note that the classification of "working" versus "non-working" nucleators here is contingent on the threshold chosen for experimental subcooling reduction, here set at $\geqslant$ 40%. This value was selected as a pragmatic compromise: stringent enough to exclude marginal effects but lenient enough to capture nucleators with moderate efficacy. However, if a more stringent threshold ($\geqslant$ 50% reduction) were applied, several borderline cases (*e.g.* K_2CrO_4 at 40.4% in $MgCl_2{\cdot}6H_2O$, and $CaK_2(SO_4){\cdot}2H_2O$ at 51.5%) would shift categories. In $MgCl_2{\cdot}6H_2O$, for instance, raising the cut-off to 50% would reclassify K_2CrO_4 as non-working, increasing the number of true negatives but simultaneously introducing an additional false negative, lowering the overall accuracy from 84.2% to ~78%. Conversely, if a more relaxed threshold ($\geqslant$30% reduction) were used, several compounds currently treated as non-working (*e.g.* $Ba(NO_3)_2$ at 35.1% in $Mg(NO_3)_2{\cdot}6H_2O$) would be reclassified as working. In this case, $Mg(NO_3)_2{\cdot}6H_2O$'s accuracy would drop from 75% to 70%, since the model predicts non-zero matches for $Ba(NO_3)_2$, inflating the false positive rate.

Thus, while the 40% cut-off provides a useful operational definition, the results should be interpreted as semi-quantitative rather than absolute. Shifting the threshold by even 10% can move borderline nucleators between categories and visibly alter accuracy, precision, and recall. This sensitivity underscores the importance of viewing the geometric model as a relative discriminator of strong versus weak nucleators, rather than a binary pass/fail test tied to a single experimental threshold.

Finally, the steep drop-off observed in match counts towards the right-hand side of each plot also hints at a useful thresholding mechanism: materials below a certain

match number could be deprioritised in experimental screening, offering an efficient triage strategy for narrowing down candidate libraries.

4.4. Conclusion and future work

Utilising sodium acetate trihydrate (SAT) as a liquid/solid phase change material (PCM) in heat storage battery applications is hampered by the need to control the crystallisation process through the addition of chemical nucleating agents (NUCs). Finding new NUCs that remain solid at higher temperatures would allow the energy storage potential of SAT to be extended, to capitalise on both its high sensible heat storage as well as its high latent heat storage potential. Currently, experimentalists approach this task by systematically testing the stock of chemicals on the laboratory shelf, a process that is both time- and resource-consuming, while yielding little gain in overall understanding. Using a data-driven approach, however, a high-throughput workflow has been created based on geometric feature matching that returns a binary decision of working/non-working NUC for a given PCM. Following an improvement of the current prediction model on ice nucleation, a reliable dataset of three working and fifteen non-working NUCs for SAT was constructed to train and test this model. Libraries of slabs cleaved from crystal structures (up to $h\ k\ l \leq 4$) were created, and the resulting supercells created upon NUC/SAT slab docking were assessed by just four geometric features (the lattice overlap, angle mismatch, and two lattice mismatch vectors), which in turn were derived from just two vectors for each NUC and SAT slab model. Following a manual training process, the upper limit values for the six features were systematically tightened until the number of non-working NUC models fell to zero, at which point the model became a binary classifier. Correlation matrices obtained from the working NUC models showed little correlation between the six features.

The trained model was then applied in a high-throughput application, where over 14,000 candidate NUC structures were screened, from which a list of 521compounds were identified as potential NUCs for SAT. This revealed some important statistical information. Firstly, locating 521 hits from a long list of 14,000 suggests a success rate of just 3.7%. This highlights the need to address this problem through a data-driven

approach, where resources can be focused on a pre-screened list known to result in a geometric match. Data also shows that it is statistically more likely that successful NUCs for SAT will be based on sodium phosphate/pyrophosphate salts. The data also reinforces the success of the current industry-standard NUC for SAT, disodium hydrogen-phosphate hydrate, which is shown to geometrically match slabs of SAT regardless of the level of hydration present. The list of potential new NUCs for SAT include some obvious candidates for experimental study.

The result of this high-throughput application combined with XRD pattern analysis demonstrated the possibility of epitaxial growth mechanism during heterogenous nucleation of inorganic salts, proved by overlapping of matching planes between pure geometric screening and XRD analysis.

An important question to raise is whether the features identified in this study will apply to other PCMs beyond SAT. This is exemplified with four other common salt-hydrate based PCMs, and showed that the same geometric feature tuning obtained for SAT could be applied to $CaCl_2 \cdot 6H_2O$, $Mg(NO_3)_2 \cdot 6H_2O$, $LiNO_3 \cdot 3H_2O$, and $MgCl_2 \cdot 6H_2O$. Notably, this revealed that the poor nucleators, i.e. those that showed almost no reduction in subcooling were likely to report no matching geometric interfaces. This confirms the transferability of this model in extending to other salt hydrates.

The prospects of this project lie in three aspects. From the research standpoint, this prediction model allowed us to tune geometric features to build a workflow that could be a fully automated ML process. From a fundamental theory development aspect, the statistical data generated will create new insights on the processes that drive heterogeneous nucleation. And finally from a commercial aspect, the ability to screen many thousands of compounds in a short space of time, to generate a short list of candidates that share features with known successful compounds, is an extremely valuable new angle on what was previously a serendipitous endeavour.

References

1. O. Boucher, D. Randall, P. Artaxo, C. Bretherton, G. Feingold, P. Forster, V.-M. Kerminen, Y. Kondo, H. Liao and U. Lohmann, in *Climate change 2013,* Cambridge University Press, 2013, pp. 571-657.
2. W. K. Tao, J. P. Chen, Z. Li, C. Wang and C. Zhang, *Reviews of Geophysics*, 2012, **50**.
3. B. Murray, D. O'sullivan, J. Atkinson and M. Webb, *Chemical Society Reviews*, 2012, **41**, 6519-6554.
4. T. Wada, Y. Matsuo and R. Yamamoto, *Solar Energy*, 1984, **33,** 3-4.
5. M. Dannemand, J. M. Schultz, J. B. Johansen and S. Furbo, *Applied Thermal Engineering*, 2015, **91**, 671-678.
6. L. Shorrock and J. Utley, *Domestic Energy Fact File 2003*, BRE Bookshop Watford, 2003.
7. B. Zalba, J. M. Marín, L. F. Cabeza and H. Mehling, *Applied Thermal Engineering*, 2003, **23**, 251-283.
8. A. Sharma, V. V. Tyagi, C. R. Chen and D. Buddhi, *Renewable and Sustainable Energy Reviews*, 2009, **13**, 318-345.
9. M. Farid, A. M. Khudhair, S. A. K. Razack and S. Al-Hallaj, *Thermal Energy Storage with Phase Change Materials*, 2021, 4-23.
10. F. Agyenim, N. Hewitt, P. Eames and M. Smyth, *Renewable and Sustainable Energy Reviews*, 2010, **14**, 615-628.
11. J. Jaguemont, N. Omar, P. Van den Bossche and J. Mierlo, *Applied Thermal Engineering*, 2018, **132**, 308-320.
12. L. Jiang, H. Zhang, J. Li and P. Xia, *Energy*, 2019, **188**, 116048.
13. D. E. Oliver, A. J. Bissell, X. Liu, C. C. Tang and C. R. Pulham, *CrystEngComm*, 2021, **23**, 700-706.
14. Z. Ma, H. Bao and A. P. Roskilly, *Solar Energy Materials and Solar Cells*, 2017, **172**, 99-107.
15. Y. He, N. Zhang, Y. Yuan, X. Cao, L. Sun and Y. Song, *Journal of Thermal Analysis and Calorimetry*, 2018, **133**, 859-867.
16. M. A. Rogerson and S. S. Cardoso, *AIChE Journal*, 2003, **49**, 505-515.
17. T. Kimura and N. Koga, *The Journal of Physical Chemistry A*, 2011, **115**, 10491-10501.
18. N. Kumar, D. Banerjee and R. Chavez Jr, *Journal of Energy Storage*, 2018, **20**, 153-162.
19. J. Li, M. A. Parkes and C. G. Salzmann, *Crystal Growth & Design*, 2024, **24**, 8292-8300.
20. H. Kimura and J. Kai, *Solar energy*, 1985, **35**, 527-534.
21. W. Cui, T. Guo, Y. Xinhui, H. Ma, D. Liang and J. Dong, *Solar Energy Materials and Solar Cells*, 2023, **250**, 112098.
22. M. Lauricella, S. Meloni, N. J. English, B. Peters and G. Ciccotti, *The Journal of Physical Chemistry C*, 2014, **118**, 22847-22857.
23. J. Anwar and D. Zahn, *Angewandte Chemie International Edition*, 2011, **50**, 1996-2013.
24. P. Pedevilla, M. Fitzner and A. Michaelides, *Physical Review B*, 2017, **96**, 115441.
25. D. Zahn, *The Journal of Physical Chemistry B*, 2007, **111**, 5249-5253.

26. M. Sun, T. Liu, M. Li, T. Liu, X. Wang, G. Chen and D. Jiang, *Journal of Energy Storage*, 2023, **62**, 106956.
27. D. E. Oliver, Degree of Doctor of Philosophy, University of Edinburgh, 2015.
28. M. Telkes, *Industrial & Engineering Chemistry*, 1952, **44**, 1308-1310.
29. T. Wada and R. Yamamoto, *Bulletin of the Chemical Society of Japan*, 1982, **55**, 3603-3606.
30. G. James, D. Witten, T. Hastie and R. Tibshirani, *An Introduction to Statistical Learning*, Springer, 2013.
31. M. Kuhn and K. Johnson, *Applied Predictive Modeling*, Springer, 2013.
32. I. Guyon and A. Elisseeff, *Journal of Machine Learning Research*, 2003, **3**, 1157-1182.
33. C. F. Dormann, J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, J. R. G. Marquéz, B. Gruber, B. Lafourcade and P. J. Leitão, *Ecography*, 2013, **36**, 27-46.
34. M. Hellenbrandt, *Crystallography Reviews*, 2004, **10**, 17-22.
35. Willoughby, Cerys. "Searching Chemical Availability (ChASe)". PSDI (accessed 10 August 2025).
36. A. Ghule, C. Bhongale and H. Chang, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 2003, **59**, 1529-1539.
37. F. C. Frank and J. H. van der Merwe, *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 1949, **198**, 205-216.
38. L. Stauffer, A. Mharchi, C. Pirri, P. Wetzel, D. Bolmont, G. Gewinner and C. Minot, *Physical Review B*, 1993, **47**, 10555.
39. S. Belyakov and C. Gourlay, *Acta Materialia*, 2014, **71**, 56-68.
40. S. Deville, J. Adrien, E. Maire, M. Scheel and M. Di Michiel, *Acta Materialia*, 2013, **61**, 2077-2086.
41. S. Deville, E. Maire, A. Lasalle, A. Bogner, C. Gauthier, J. Leloup and C. Guizard, *Journal of the American Ceramic Society*, 2010, **93**, 2507-2510.
42. S. Li, X. Xie and Y. Liu, *Scientific Reports*, 2024, **14**, 8051.
43. S. Pu and K. Hadinoto, *Chemical Engineering Research and Design*, 2024, **201**, 45-66.
44. S. Mirza, I. Miroshnyk, J. Heinämäki, O. Antikainen, J. Rantanen, P. Vuorela, H. Vuorela and J. Yliruusi, *AAPS PharmSciTech*, 2009, **10**, 113-119.
45. J. Jennings, M. F. Butler, M. McLeod, E. Csányi, A. J. Ryan and O. O. Mykhaylyk, *Crystal Growth & Design*, 2018, **18**, 7094-7105.
46. S. K. Pramanik and B. Ganguly, *CrystEngComm*, 2025, **27**, 2439-2451.
47. G. A. Lane, *Solar Energy Materials and Solar Cells*, 1992, **27**, 135-160.
48. I. Sutjahja, S. R. AU, N. Kurniati, I. D. Pallitine and D. Kurnia, *Journal of Physics: Conference Series,* 2016, **739**, 012064.
49. X. Li, Y. Zhou, H. Nian, X. Zhang, O. Dong, X. Ren, J. Zeng, C. Hai and Y. Shen, *Energy & Fuels*, 2017, **31**, 6560-6567.
50. S. Cao, X. Luo, X. Han, X. Lu and C. Zou, *Energies*, 2022, **15**, 824.
51. Z. He, H. Ma and S. Lu, *Journal of Energy Storage*, 2024, **90**, 111906.
52. A. Abhat, *Solar Energy*, 1983, **30**, 313-332.
53. K. Blen, T. F. and K. and Kaygusuz, *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, 2008, **30**, 775-787.

54. N. Gupta, A. Kumar, H. Dhasmana, V. Kumar, A. Kumar, P. Shukla, A. Verma, G. V. Nutan, S. K. Dhawan and V. K. Jain, *Journal of Energy Storage*, 2020, **32**, 101773.
55. P. Honcova, R. Pilar, V. Danielik, P. Soska, G. Sadovska and D. Honc, *Journal of Thermal Analysis and Calorimetry*, 2017, **129**, 1573-1581.
56. S. Kannan, N. Kumar, M. A. Jog and R. M. Manglik, *Industrial & Engineering Chemistry Research*, 2022, **61**, 16341-16351.
57. P. J. Shamberger and M. J. O'Malley, *Acta Materialia*, 2015, **84**, 265-274.

Chapter 5

Towards a generalisable prediction framework for heterogeneous nucleation

Heterogeneous nucleation remains one of the most stubbornly elusive problems in materials science. Although it is widely recognised as the controlling step in crystallisation processes, the ability to predict which substrates will act as effective nucleators has lagged far behind our ability to measure them experimentally. The earlier chapters of this thesis have shown that geometric descriptors grounded in epitaxial growth theory can explain why certain nucleators are effective for ice and sodium acetate trihydrate, yet those demonstrations are necessarily system-specific. If the predictive study of nucleation is to advance beyond isolated case studies, what is needed is a framework that unifies descriptors, thresholds, and decision rules into a generalisable, data-driven methodology. This chapter provides that synthesis. By comparing machine-learned thresholds with hand-tuned values across the previously discussed PCM systems, and by codifying a consistent aggregation rule for nucleator-level classification, this chapter seeks not only to validate earlier insights but to demonstrate that nucleation can be studied through a systematic, transferable, and reproducible pipeline.

5.1. Aim and Scope

This chapter presents a generalisable prediction framework that links geometric slab matching to nucleator-level predictions for phase-change materials (PCMs). Whereas Chapter 3 established ice as a benchmark system and Chapter 4 demonstrated sodium acetate trihydrate (SAT) as a case study, here the emphasis shifts from system-specific analyses to the development of a unified prediction framework. The overarching aim is to test whether thresholds inferred automatically from validated datasets reproduce or refine the hand-tuned geometric cut-offs derived in those previous chapters, and whether a consistent aggregation rule yields stable predictions across distinct PCM families.

The inclusion of this chapter is not merely technical, but also conceptual. Without it, the preceding chapters would remain as isolated demonstrations: geometric principles applied to ice, then to SAT, then extended piecemeal to other salt hydrates. Chapter 5 provides the synthesis and generalisation that transforms those individual demonstrations into a broader methodology. By unifying descriptors, thresholds, and decision rules across chemically distinct systems, this chapter addresses the critical question of portability: can the same pipeline be applied consistently to different PCMs, and do the geometric parameters identified in one system transfer to others? The further question of whether a single combined model can be learned across multiple PCMs is considered later in this chapter.

A further justification lies in the tension between expert-driven rules and data-driven inference. In Chapters 3 and 4, thresholds were selected by informed judgement, guided by epitaxial growth theory and empirical performance. Such hand-tuned rules are persuasive in context, but they carry the risk of subjectivity and limited portability. Chapter 5 introduces machine learning not as a replacement, but as a rigorous test of whether the data support or challenge expert intuition. If data-driven thresholds converge on the same values chosen by hand, this strengthens confidence in the physical soundness of the earlier work. Conversely, if divergences arise, they highlight precisely where geometric assumptions may need refinement. In both cases, the analysis presented in this chapter advances the thesis beyond anecdotal validation into the realm of reproducible, quantitative methodology.

Finally, this chapter situates the thesis in the broader landscape of materials informatics. Predicting heterogeneous nucleation remains one of the most challenging and least understood aspects of phase-change material design. By showing how crystallographic descriptors can be consistently transformed into predictive features, validated across multiple systems, and interpreted through model explainability techniques, Chapter 5 demonstrates that heterogeneous nucleation can be studied not only case by case, but also through a generalisable data-driven framework. This contribution, bridging physical insight and machine learning, is essential for moving beyond qualitative reasoning towards scalable prediction tools for new PCMs.

5.2. Datasets

The datasets considered here are derived from the compiled pair-level outputs of the slab-matching pipeline (Chapter 2), filtered to include only those nucleator–PCM pairs for which experimental labels are available. Four independent subsets are defined:

Ice: nucleator slabs matched with ice-I_h cleavage planes. Ground truth working status is taken from immersion freezing experiments (Chapter 3);

Sodium acetate trihydrate (SAT): nucleator slabs matched with SAT surfaces. Working status is derived from laboratory immersion freezing and cycling experiments (Chapter 4);

Four salt hydrates ($CaCl_2{\cdot}6H_2O$, $Mg(NO_3)_2{\cdot}6H_2O$, $LiNO_3{\cdot}3H_2O$ and $MgCl_2{\cdot}6H_2O$): nucleator slabs matched with four additional salt hydrates, for which limited but validated experimental data from literature research are available (Chapter 4).

Cross-PCM pooled dataset: in addition to analysing each PCM individually, a combined dataset was also constructed by pooling all nucleator-PCM pairings across ice, SAT, and the four hydrates. This unified set enables direct testing of whether a single model can learn transferable geometric rules across chemically distinct systems, and whether such a model retains predictive power when applied to each PCM in isolation.

Each subset comprises multiple PCM cleavage planes and multiple slabs per nucleator, generating hundreds of potential pair-level matches. Importantly, all three subsets (ice, SAT, and four additional salt hydrates) were generated using an identical pipeline for slab generation, symmetry reduction, alignment, and descriptor calculation (see Chapter 2). This consistency is critical: this machine learning technique-based prediction framework can only operate meaningfully if the features describing each example are defined in the same way across all systems. If the slab construction rules, tolerance definitions, or feature formulations varied between datasets, then observed differences could not be attributed to nucleation physics, but merely to inconsistencies in data preparation.

A key consideration about the nature of this thesis that adds to the complexity for this prediction framework dataset is that, while a single successful interface match can make a nucleator function experimentally, it remains unknown which specific slab-slab interface drives nucleation in each case. It is not clear which individual slab pairings are successful; only whether the nucleator as a whole has been experimentally validated. This introduces a key challenge: how to learn from highly granular geometric data when labels exist only at the aggregate level. The unified schema i.e. features defined on a slab-pairing level, as well as binary labels defined on a nucleator level, therefore represents a fundamental pillar for the development of the predictive models presented here in Chapter 5.

For each nucleator-PCM pair, the following features were computed in this machine learning pipeline:

Angle mismatch ($\Delta\theta$): the deviation in interfacial angle between the PCM and NUC slabs;

Vector mismatch ($\Delta u = \left|\frac{\overrightarrow{u_1}}{\overrightarrow{u_2}} - 1\right|$ and $\Delta v = \left|\frac{\overrightarrow{v_1}}{\overrightarrow{v_2}} - 1\right|$): the relative length mismatches of the two in-plane lattice vectors of the matched slabs;

Maximum area overlap ($|\overrightarrow{u_1} . \overrightarrow{v_1}| \approx |\overrightarrow{u_2} . \overrightarrow{v_2}|$): the overlap area of the two slabs when expanded to their commensurate supercells, bounded by a dynamic constraint proportional to the basal (001) PCM slab area;

m_n_tolerance ($\left|\frac{|\overrightarrow{u_1} . \overrightarrow{v_1}| - |\overrightarrow{u_2} . \overrightarrow{v_2}|}{maximum\ area\ overlap}\right|$): to temper the ($m$, n) supercell generation to sensible outcomes compared to the maximum area overlap.

The binary labels of working = 1 or non-working = 0 are defined at the nucleator level, based on experimental validation of whether each nucleator promotes or fails to promote crystallisation of the target PCM. Because the geometric descriptors are calculated for all slab pairings, each nucleator is associated with many candidate interfaces. This necessitates an aggregation step, whereby multiple pair-level descriptors are combined to yield a nucleator-level classification.

By structuring the dataset in this way, consistent pair-level features across systems aggregated into nucleator-level labels, and experimentally validated outcomes, the groundwork is laid for the machine learning analysis. This framework allows thresholds learned from one PCM system (*e.g.* ice) to be compared directly with those hand-tuned in earlier chapters (Chapter 3 and 4), and tested for stability and transferability across distinct PCMs. In addition, the pooled dataset provides a direct test of framework generalisability, asking whether a single set of descriptors and thresholds can capture nucleation behaviour across multiple PCMs simultaneously.

5.3 Framework description

The generalisable prediction framework developed in this chapter can be understood as a multi-stage workflow that transforms raw geometric matches into nucleator-level predictions. At its core, the approach links fine-grained slab-pair descriptors to the coarser experimental labels available at the nucleator level, while systematically defining thresholds, applying aggregation rules, and testing model interpretability. Figure 5.1 illustrates the workflow: beginning with geometric descriptors derived from slab matching, proceeding through aggregation and thresholding strategies, followed by classifier training under a strict cross-validation scheme, and culminating in feature attribution via SHAP (SHapley Additive exPlanations) analysis[1]. This structured pipeline makes transparent not only how predictions are generated, but also how expert-derived rules and data-driven inference interact within a reproducible methodology.

Figure 5.1. Workflow and data outputs of the generalisable prediction framework.

5.3.1 From pair-level matches to nucleator-level classification

The fundamental challenge of this problem is that labels exist only at the nucleator level: experimentally, one can determine whether a compound promotes or inhibits nucleation of a given PCM, but not which specific slab-slab interface is responsible. Yet the geometric analysis produces data at the much finer pair level, with dozens or hundreds of candidate interfaces per nucleator. Without aggregation, there is a mismatch between the scale of descriptors and the scale of labels.

To bridge this gap, herein a pair-to-nucleator aggregation strategy has been adopted. For each PCM-NUC slab pair, four geometric descriptors are calculated, namely (i)

angle mismatch, (ii, iii) vector mismatches $\bar{u}$ and $\bar{v}$, and (iv) m_n_tolerance. Maximum area overlap values are fixed by a dynamic constraint proportional to the basal (001) PCM slab area. A pair is deemed to "pass" if all descriptors fall below their threshold values. A nucleator is then classified as working if it achieves at least K passes across its PCM matches. The choice of K is not unique but rather defines the level of stringency: at the most extreme, $K = 0$ would imply that zero matches defines a non-working nucleator (and therefore 1 or more matching interfaces defines a working nucleator), whereas more stringent values such as $K = 10$ require a broader set of consistent matches. Unless otherwise stated, $K = 5$ is used as a balanced criterion, reflecting the expectation that effective nucleators exhibit multiple independent opportunities for favourable registry rather than relying on a single isolated match. This rule is implemented directly in the pipeline:

```
(nuc_grouped["predicted_label"]=(nuc_grouped["num_matches"
] >= 5))
```

from `Pandas` module[2]. Sensitivity tests with K = 0, 5, 10, 15 are performed to confirm robustness. This aggregation step ensures that predictions are aligned with experimental evidence, mitigating noise from uncertain individual slab contributions, and reflecting physical intuition that robust nucleators should support several good registries rather than one isolated coincidence.

5.3.2 Threshold determination strategies

Thresholds define what counts as an acceptable geometric match. In earlier chapters, values were hand-tuned via systematic tightening to mirror experimental binary outcomes. In contrast, in this chapter, thresholds are instead inferred from the data. This serves two purposes: (i) to test whether machine-learned thresholds corroborate the expert-selected values, and (ii) to provide a reproducible way to transfer the approach across PCMs. Thresholds define what counts as a "good" geometric match. For this, three automated data-driven strategies were pursued:

> **Quantile rule (default)**: Thresholds are defined using the quantile rule, where each descriptor cut-off is set to the 30th percentile of its distribution (*e.g.*

`df["vector_mismatch_u"].quantile(0.3)`). In other words, only the best/lowest 30% of values (*i.e.* the best geometric matches) are allowed to pass. This simple rule requires no training, avoids overfitting, and provides a consistent baseline across systems[3];

Receiver operating characteristic (ROC)-optimised thresholds: For each descriptor, the cut-off values were varied and assessed on how well it separated working from non-working nucleators[4]. The quality of separation was measured by balancing two rates: the rate of true working nucleators correctly identified, also called the rate of true positives (TP), and the rate of non-working nucleators incorrectly labelled as working, also called the rate of false positives (FP)[5]. The cut-off that gave the best balance was chosen, with the evaluation done by systematically leaving one nucleator out at a time (i.e. so the test nucleator was never used in training). This is implemented by calling `roc_curve` from `sklearn.metrics` library[6];

Joint optimisation by grid search: In this approach, thresholds for all descriptors and the vote cut-off K were tuned together by exhaustively scanning over a grid of possible values. The best combination was chosen as the one that maximised balanced accuracy[7]; that is, the average of correct classification rates for working and non-working nucleators. To prevent overfitting, this was evaluated using grouped cross-validation (CV, see Section 5.3.4.), where entire nucleators (not individual slabs) were left out of the training step. This was implemented by importing `GridSearchCV` module from `sklearn.model_selection` library[6].

Together, these three strategies permit a comparison of a simple rule (quantile) with more data-driven optimisations, and to test whether the machine-learned cut-offs converge on or refine the expert-selected values used in earlier chapters.

5.3.3 Model training

To complement the threshold-based voting rule, two machine learning classifiers were trained on the aggregated nucleator-level datasets. The intention is not to maximise

predictive accuracy per se, but to provide a systematic, reproducible means of quantifying which geometric descriptors contribute most strongly to nucleation outcomes. Because of the modest dataset sizes and the need for interpretability, simple, well-established models were deliberately selected: (i) penalised logistic regression and (ii) shallow random forest.

Penalised logistic regression (LR)

Logistic regression is a linear classification model that estimates the probability that an input belongs to class y =1 (working nucleator) or class y = 0 (non-working nucleator)[8, 9]. It assumes that the log-odds of the outcome can be expressed as a weighted sum of the input features:

$$\log\left(\frac{P(y=1)}{P(y=0)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \qquad (1)$$

where x_i are the geometric descriptors (*e.g.* angle mismatch, vector mismatches), and β_i are the learned coefficients. The sign and magnitude of each β_i indicates how strongly the corresponding feature contributes to predicting a nucleator as a working nucleator for the given PCM. Because small datasets can lead to unstable estimates, a L2 regularisation[10, 11] (also called ridge penalty) was further used, as outlined in Equation 2:

$$L(\beta) = -\sum_{i=1}^{n}[y_i \log \hat{p_i} + (1-y_i)\log(1-\hat{p_i})] + \lambda \sum_{j=1}^{k} \beta_j^2 \qquad (2)$$

Where $L(\beta)$ is the loss (penalised negative log-likelihood), $\hat{p}_i$ is the predicted probability for sample i, λ is the regularisation hyperparameter, tuning the strength of the L2 penalty applied to the squared coefficients, the first term is negative log-likelihood fit to data, and the second term is L2 penalty, i.e. the squared coefficients. This adds a constraint that discourages excessively large coefficient values by penalising their squared magnitude during training. Regularisation prevents overfitting and improves generalisability, while still allowing coefficients to be directly interpreted as feature weights. Logistic regression therefore serves as a transparent

baseline model, linking descriptors to nucleation outcomes in a mathematically simple and physically interpretable way. The logistic regression was called by importing `LogisticRegression` from `scikit-learn` library[6].

Random Forest (RF)

While logistic regression captures only linear effects of individual features, nucleation behaviour may also depend on nonlinear interactions (*e.g.* a slab being favourable only when both angular mismatch and lattice mismatch are simultaneously small). To capture such dependencies, a Random Forest (RF) classifier was also trained. This is an ensemble of decision trees, each built on a bootstrap re-sample of the data and a random subset of features[12, 13]. Individual decision trees operate by recursively splitting the dataset on feature thresholds (*e.g.* "if $\Delta\theta > 0.2°$, go left; otherwise, go right"), until terminal nodes (leaves) are reached that assign class labels. By averaging predictions across many trees, a RF classifier reduces the variance inherent in any single tree, leading to greater robustness. Figure 5.2 shows the workflow of a bootstrap resampled RF technique. The RF was called by `RandomForestClassifier` module from the `sklearn.ensemble` library[6].

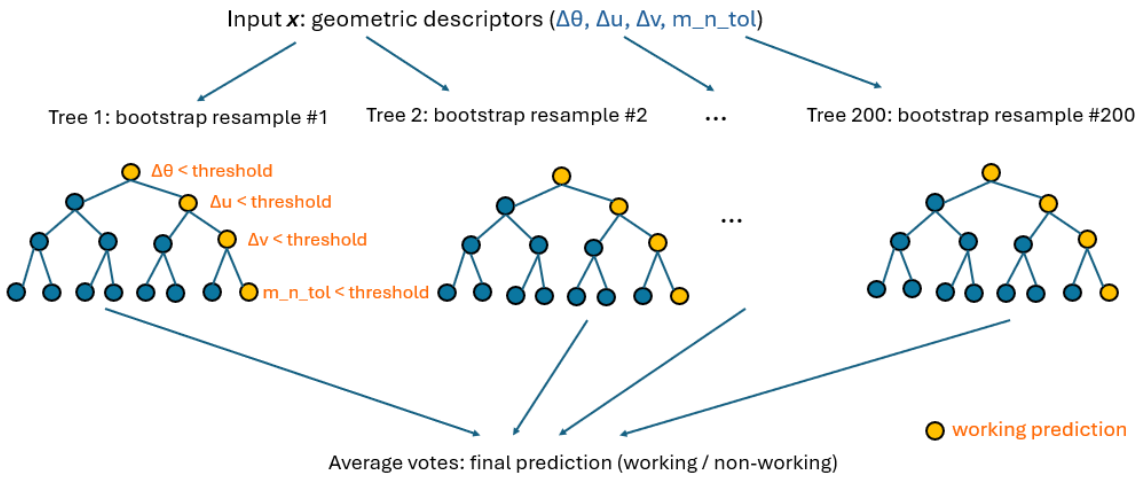


Figure 5.2. Random Forest workflow adapted for heterogeneous nucleation prediction. Geometric descriptors (angle mismatch, vector mismatches, m_n_tolerance) are used

as inputs. Multiple decision trees are trained on bootstrap resamples of the nucleator dataset, each producing working/non-working predictions.

RF classifiers are well suited for the task here for three reasons. First, they can capture non-linear interactions between geometric descriptors (*e.g.* angle tolerance and vector mismatch jointly influencing registry quality), which linear models may miss. Second, they are relatively robust to noise and small sample sizes, especially when constrained to shallow depth and moderate tree numbers[14, 15] (100 estimators were used here). Third, they provide a natural link to feature importance analysis[16], which is extended using SHAP (SHapley Additive exPlanations, see 5.3.5) values to achieve more interpretable, per-feature contribution scores.

It is important to note that correlations among features can bias raw feature importance estimates in tree ensembles[17, 18]. This is partly mitigated here by the use of SHAP analysis, which accounts for feature interactions more fairly than simple permutation measures[19]. Nevertheless, to avoid over-interpretation, importance rankings are discussed in the context of descriptor interdependence and physical meaning, rather than treated as absolute.

5.3.4 Cross-validation protocol

Cross-validation (CV) provides a systematic way to estimate model performance on unseen data[20, 21] . In CV, the dataset is split into separate parts: one part is used to train the model, and the remaining part is used to test it. By repeating this process several times with different splits, an average performance measure that reflects how the model is likely to generalise beyond the available data is obtained[21]. Because the datapoints are groups of slab matches per nucleator, splitting slabs from the same nucleator across train/test would leak information. Therefore, a method specific to this thesis called "Leave-One-Nucleator-Out CV (LONOCV)" is used, called out by `GroupKFold` from `sklearn.model_selection` library. Here, each nucleator is held out in turn as the test set, while all remaining nucleators form the training set. This is illustrated in Figure 5.3. This maximises the amount of training data in each fold, while ensuring that all slab matches associated with a given nucleator are excluded from training when that nucleator is used for testing. LONOCV is

particularly appropriate for the relatively small datasets employed in this work, as it provides the most stringent possible estimate of model generalisability.

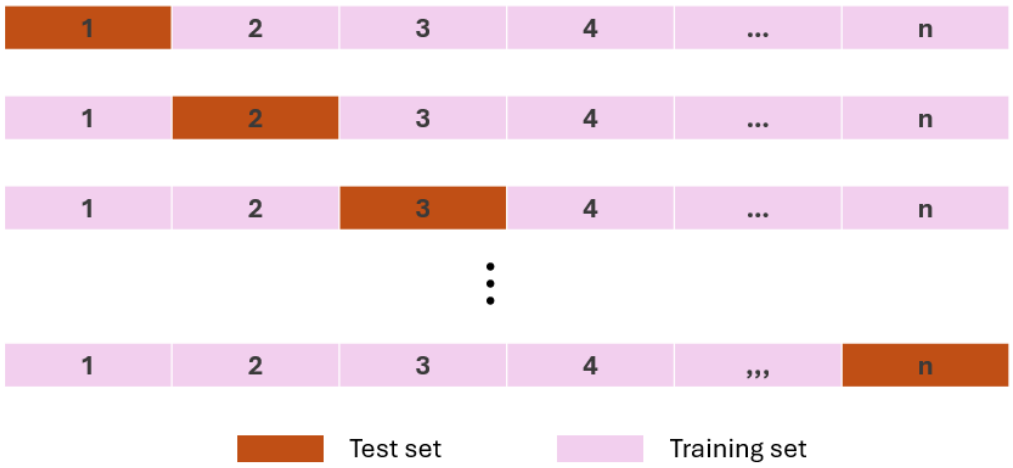


Figure 5.3. Schematic of cross-validation strategy of Leave-One-Nucleator-Out CV (LONOCV), the strictest form of grouped CV, each nucleator is excluded in turn while all others are used for training.

Given the small, validated datasets and the emphasis on methodological transparency, herein raw counts and distributions (confusion matrices, threshold tables) that rely on sensitivity analyses (varying cut-offs and K) are reported to assess robustness. Formal uncertainty quantification (*e.g.*, grouped cross-validation, bootstrapped confidence intervals) is not applied in the present workflow and is left as future work once larger validated cohorts are available.

A key methodological consideration in this work is the granularity at which the data are split into training and testing sets. Although geometric descriptors are computed at the slab-pair level, labels exist only at the nucleator level. If slab matches from the same nucleator were distributed across both training and test folds, information leakage would occur, i.e. the model could indirectly "see" geometric patterns associated with a nucleator during training and thereby artificially inflate performance during testing. Grouped cross-validation at the nucleator level is therefore essential to ensure that predictions for a held-out nucleator are made entirely without access to its own slab data.

The choice of Leave-One-Nucleator-Out CV (LONOCV) represents the strictest possible grouped validation strategy for small datasets. It maximises training data in each fold while guaranteeing independence of the test nucleator. However, this approach also has limitations. Because the number of nucleators per PCM is modest, performance estimates are inherently high-variance and sensitive to individual influential samples. In addition, LONOCV evaluates interpolation within the existing chemical space rather than true extrapolation to entirely novel classes of materials. No fully independent external test cohort is available, which limits claims of universal generalisability.

Finally, in the pooled cross-PCM dataset, training and testing are still conducted at the nucleator level, meaning that a nucleator is never evaluated on its own data. However, the model may still encounter chemically related systems during training, which could inflate apparent transferability. These constraints are intrinsic to the currently available experimentally validated datasets and should be borne in mind when interpreting reported accuracies.

5.3.5 SHAP feature attribution

Interpreting machine learning models is as important as measuring their predictive accuracy, particularly in small scientific datasets where explanatory value is essential. To this end, SHAP (SHapley Additive exPlanations)[1] plots have been utilised to quantify the contribution of each geometric descriptor to model predictions.

SHAP is grounded in cooperative game theory[22]. It treats each feature as a "player" in a game whose "payout" is the model's prediction. The contribution of a feature is defined as the average change in prediction when that feature is added to all possible subsets of the other features. This produces a unique, fair allocation of importance scores across features. In practice, efficient algorithms exist to compute SHAP values for tree-based models such as Random Forests[23].

In this chapter, SHAP is applied to the Random Forest model trained on nucleator-level labels. To make this idea tangible, two standard SHAP visualisations are discussed here as background context (Figures 5.4 and 5.5). In the former, a bar plot

of average absolute SHAP values shows the mean magnitude of contribution of a descriptor across the test set. Longer bars indicate features that consistently exert a stronger influence on the model's output, and in this way a global ranking of feature importance is obtained.

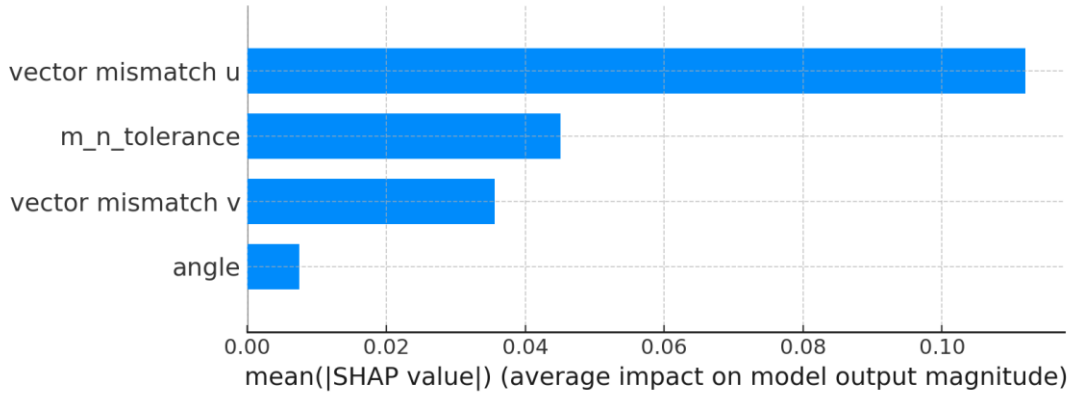


Figure 5.4. Illustrative bar plot of average absolute SHAP values for the Random Forest classifier. Each bar shows the mean magnitude of a descriptor's contribution to model predictions across the test set. Longer bars correspond to features that consistently exert stronger influence on whether a nucleator is classified as working or non-working. This plot provides a global ranking of descriptor importance.

In contrast, the so-called 'violin plot' of SHAP value distributions (Figure 5.5) shows how SHAP values for a feature vary across all nucleators. The horizontal spread indicates whether the feature sometimes drives predictions positive (right) or negative (left). The colour scale encodes the raw feature value (red = high, blue = low), so one can see whether high or low numerical values of a given descriptor are associated with increased likelihood of a nucleator being classified as working or non-working. This plot therefore provides not only global importance but also directionality and consistency of effects across the dataset.

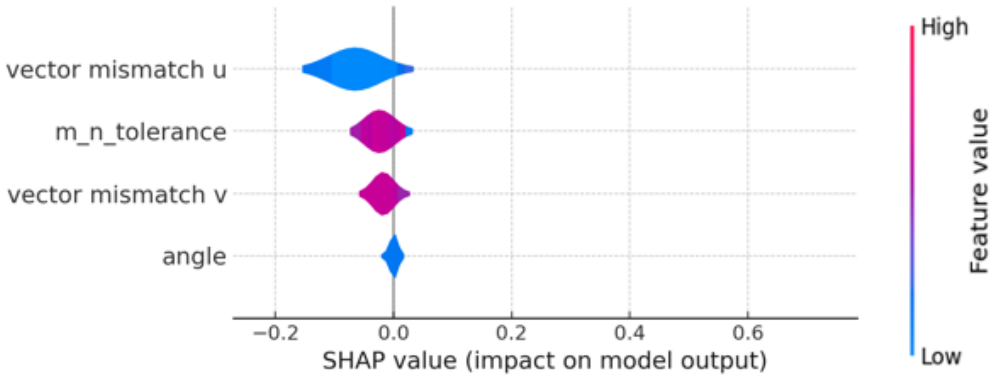


Figure 5.5. Illustrative violin plot of SHAP value distributions. Each "violin" represents the spread of SHAP values for one descriptor across all nucleators. Horizontal spread to the right indicates that the feature contributes positively towards a "working" classification, while spread to the left indicates the opposite. Colour encodes the raw feature value (red = high, blue = low), showing whether high or low values are associated with working nucleators.

Together, these plots allow both an overview of which descriptors matter most, and an intuitive sense of how they influence predictions. They transform the Random Forest from a "black box" into a model that can be directly compared to physical expectations: for instance, whether smaller angular mismatches or smaller vector mismatches indeed shift predictions toward "working" nucleators, as expected.

5.3.6 Bootstrap feature importance distributions

In addition to SHAP analysis, which attributes model predictions to features at both local and global levels, a bootstrap resampling of the RF models was also undertaken to characterise feature importance[24]. In this procedure, the nucleator-level dataset is repeatedly resampled with a replacement, a RF classifier is trained on each resample, and the resulting feature importance scores are collected[25]. Repeating this process many times (200 replicates was employed in this study) produces a distribution of importance values for each descriptor rather than a single estimate. Figure 5.6 depicts the bootstrapping technique that is used here. Bootstrap feature importance distributions were obtained by repeated training of RF classifiers (`RandomForestClassifier`, `scikit-learn`[6]) on bootstrap resamples

generated with `NumPy`[26]. The resulting distributions of feature importances were aggregated with `Pandas`[2] and visualised as boxplots using `Matplotlib`[27].

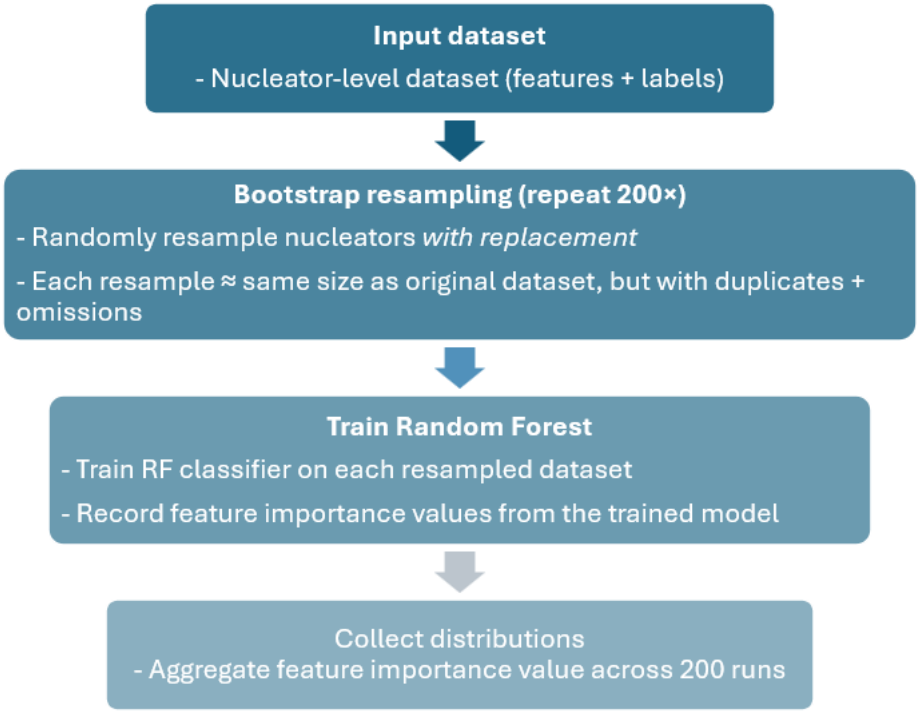


Figure 5.6. Bootstrap resampling procedure for Random Forest feature importance.

This approach complements SHAP by providing an uncertainty-aware perspective. Whereas SHAP offers a rigorous attribution of feature contributions grounded in cooperative game theory, bootstrap resampling quantifies the stability of feature importance rankings under dataset variability. Taken together, the two methods provide a balanced interpretability framework: SHAP highlights how features contribute to individual and overall predictions, while bootstrap distributions convey the robustness and variability of those contributions across resampled datasets.

5.4. Results and Discussion: Ice as a benchmark system

The prediction framework was first applied to the ice dataset, comprising 63 unique ice slabs paired against 32 × 63=2,016 nucleator slabs, yielding a total of 127,008 candidate slab-slab interfaces. This baseline dataset allows us to test whether data-driven thresholds reproduce the manually optimised values derived in Chapter 3. Three

thresholding strategies were compared: ROC-optimised, quantile rule (30th percentile), and joint optimisation by grid search, under both Logistic Regression (LR) and Random Forest (RF) classifiers. In practice, these thresholds serve as the decision boundary that determines whether a nucleator is classified as working or non-working, i.e. whether the number of matched interfaces exceeds the cut-off.

Unless otherwise stated, aggregation is carried out with $K = 5$. However, to illustrate how performance depends on this choice, Tables 5.1-5.5 report multiple K values, spanning more stringent criteria (e.g. K = 10, requiring at least ten valid slab matches for classification as working) and more relaxed criteria (e.g. K = 0, where no minimum hit count is imposed).. In each case, the thresholding strategy is evaluated across this range, and the corresponding classification outcomes are benchmarked. The confusion matrices shown in Fig. 5.7 and related figures adopt the K value that gave the highest classification accuracy for that thresholding method, which for some methods corresponds to $K = 10$.

Across all strategies, Leave-One-Nucleator-Out CV (LONOCV) balanced accuracies fell in the range of 0.64-0.71, indicating moderate predictive power in distinguishing working from non-working nucleators. ROC-optimised thresholds consistently yielded the highest balanced accuracy for LR (≈0.75), with grid search performing comparably and quantile-based cut-offs underperforming (≈0.44-0.56). Importantly, the ROC-derived thresholds converged very closely with the expert-selected geometric cut-offs from Chapter 3 ($\Delta u \approx 0.027$, $\Delta v \approx 0.029$, $\Delta\theta \approx 10^{-8}$, $m_n_$tolerance ≈ 0.031), suggesting that data-driven optimisation corroborates the physical reasoning underpinning those earlier selections (see Table 5.1).

Table 5.1. Comparison of threshold values under hand-tuned, ROC-optimised, and quantile regimes. The highlighted row represents the best performing model.

Method	Δu	Δv	Δθ	*m_n*_tolerance	Accuracy
K=0 ROC RF	0.027	0.029	1.11×10^{-8}	0.031	64.71%
K=5 ROC RF	0.027	0.029	1.11×10^{-8}	0.031	67.64%
K=10 ROC RF	0.027	0.029	1.11×10^{-8}	0.031	71.90%
K=15 ROC RF	0.027	0.029	1.11×10^{-8}	0.031	71.90%
K=0 quantile RF	0.016	0.012	0	0.008	43.82%
K=5 quantile RF	0.016	0.012	0	0.008	43.82%
K=10 quantile RF	0.016	0.012	0	0.008	43.82%
K=15 quantile RF	0.016	0.012	0	0.008	43.82%
grid RF	0.016	0.016	0	0.010	56.34%
K=0 ROC LR	0.027	0.029	1.11×10^{-8}	0.031	64.71%
K=5 ROC LR	0.027	0.029	1.11×10^{-8}	0.031	74.58%
K=10 ROC LR	0.027	0.029	1.11×10^{-8}	0.031	71.90%
K=15 ROC LR	0.027	0.029	1.11×10^{-8}	0.031	67.64%
K=0 quantile LR	0.016	0.011	0	0.008	44.11%
K=5 quantile LR	0.016	0.011	0	0.008	44.11%
K=10 quantile LR	0.016	0.011	0	0.008	41.47%
K=15 quantile LR	0.016	0.011	0	0.008	41.47%
Grid LR	0.016	0.016	0	0.010	52.94%
Manual	0.01	0.01	0.1	0.01	63.64%

Confusion matrices provide a more granular view of predictive behaviour by exposing accuracy, recall and precision values [Figure 5.7 (a-c) relates to the work reported here, while (d) is a reminder of the manually-tuned outcome established in Chapter 3]. In

this context, ‘True Positives’ (TP) are nucleators predicted as working that are also experimentally validated as working. ‘False Positives’ (FP) are nucleators predicted as working but experimentally are non-working. ‘True Negatives’ (TN) are nucleators correctly predicted as non-working, while ‘False Negatives’ (FN) are nucleators predicted as non-working, but which experiments show to be working. The following definitions then apply:

Accuracy = (TP+TN)/(TP+TN+FP+FN)

(i.e. the overall proportion of correctly classed nucleators)

Recall = TP/(TP+ FN)

(i.e. the fraction of working nucleators successfully identified)

Precision = TP/(TP+FP)

(i.e. the fraction of predicted working nucleators that are truly correct)

ROC-optimised thresholds [Figure 5.7(a)] achieved the most balanced performance, with an overall accuracy of 71.9%, high precision (92.9%) and moderate recall (61.9%). This reflects how the ROC procedure explicitly seeks to balance TP and FP rates, yielding thresholds that retain most true working nucleators without inflating FP values. In contrast, the quantile rule [Figure 5.7(b)] performed poorly: recall collapsed to only 14.3%, meaning nearly all true working nucleators were mis-classified as non-working. This arises because a fixed 30% quantile cut-off is overly restrictive, allowing very few slab matches to pass and therefore underestimating working behaviour. Grid optimisation [Figure 5.7(c)] produced intermediate behaviour, capturing more true positives than the quantile rule (recall 38.1%), though at the cost of lower overall accuracy (56.3%) and slightly reduced precision (88.9%). Its improvement over the quantile rule arises from jointly tuning descriptor cut-offs with the vote threshold K, which relaxes the overly strict quantile constraint and admits more candidate matches.

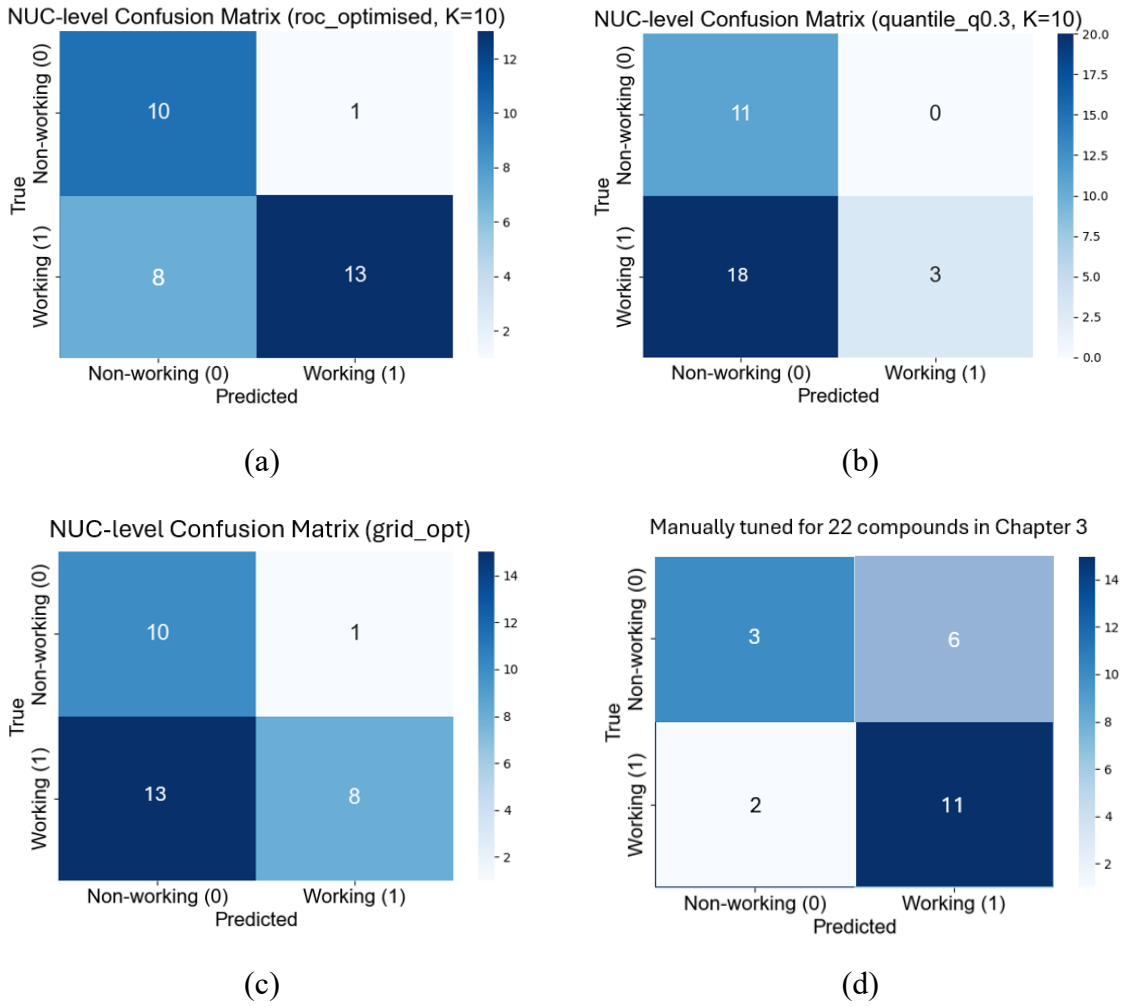


Figure 5.7. Comparison of Random Forest (RF) nucleator-level predictions for ice under three thresholding strategies with K = 10 aggregation. **(a)** ROC-optimised thresholds; **(b)** Quantile rule (30th percentile); **(c)** Grid search optimisation; **(d)** manually tuned prediction model for the 22 compounds from Chapter 3.

Notably, precision values were consistently high across all strategies ($\geqslant$88%). This indicates that when a nucleator was predicted as working, it was very likely to be correct, *i.e.* false positives were rare. The particularly high precision of the quantile rule (100%) simply reflects its extreme conservatism: by classifying almost everything as non-working, the few nucleators predicted as working were always genuine, but at the expense of missing most true working cases (low recall). In practical terms, high precision across all methods means the framework is good at avoiding false claims of

nucleation activity, but only the ROC-optimised thresholds achieve this without sacrificing sensitivity to true working nucleators.

Table 5.2. Comparison of accuracy, recall, and precision for the three thresholding strategies applied to ice nucleators (corresponding to Figure 5.7a-c).

Threshold strategy	Classifier	Accuracy	Recall	Precision
ROC-optimised (K=10)	RF/LR	71.9%	61.9%	92.9%
Quantile rule (q=0.3, K=10)	RF/LR	43.8%	14.3%	100%
Grid optimisation	RF/LR	56.3%	38.1%	88.9%
Manual (Chapter 3)	Heuristic	63.6%	84.6%	64.7%
Random baseline	N/A	50%	N/A	N/A
Majority baseline	N/A	68.2%	N/A	N/A

In Chapter 3, hand-tuned geometric thresholds were benchmarked against a set of 22 nucleators, selected from high-throughput slab-matching calculations and then validated experimentally. These represented the most promising candidates identified by the screening pipeline at the time. The confusion matrix in Figure 5.7(d) shows that this manual scheme achieved an overall accuracy of 63.6% (14/22 correct classifications), with a recall of 84.6% (11/13 working nucleators correctly identified) and a precision of 64.7%. This reflects its strength in capturing most true working nucleators (high recall), but also its weakness in generating a relatively high number of false positives (6 misclassified as working), lowering precision.

By contrast, the ROC-optimised thresholds in Figure 5.7(a), when applied to the full expanded dataset of 32 nucleators (the 22 validated compounds plus the 10 compounds introduced during the Chapter 3 training stage), yielded a more balanced performance profile. ROC achieved lower recall than the manual scheme (61.9% vs 84.6%), but much higher precision (92.9% vs 64.7%), reducing the risk of incorrectly predicting non-working nucleators as working. Overall accuracy was also higher for ROC (71.9% vs 63.6%), and crucially, ROC thresholds were selected in a systematic, data-driven way rather than by manual tightening.

For consistency with Chapters 3 and 4, performance was benchmarked against random guessing (50% expected accuracy) and the majority-class baseline (68.2%). Random guessing yields an expected accuracy of 50%, while majority-class prediction (assigning all compounds to the most frequent experimental class, working nucleator) yields an accuracy of 65.6%. The ROC-optimised thresholds therefore exceed both baselines, achieving a +6.3 percentage point improvement over majority-class prediction. In contrast, the manually tuned scheme in Chapter 3 did not surpass its corresponding majority baseline. This demonstrates that systematic ROC-based optimisation not only improves overall accuracy relative to manual tuning, but also achieves genuine discriminative performance beyond trivial class imbalance effects. This shift from heuristic threshold adjustment to data-driven optimisation therefore results in both improved precision and demonstrable performance gains relative to naïve classifiers.

It is important to clarify why Figure 5.7(d) reports performance on only 22 nucleators, whereas the machine-learning models in this chapter are benchmarked against the full set of 32. In Chapter 3, the aim was to establish a hand-tuned prediction model. The initial 12 compounds were used as a training set for manually tightening geometric thresholds; by construction, these thresholds were adjusted until all 12 were classified correctly, thereby enforcing a training accuracy of 100%. Including them in a confusion matrix would therefore have been circular and uninformative. Instead, validation was carried out on the additional 22 compounds identified by high-throughput screening and subsequently tested experimentally, providing a genuine test set for the manually tuned rules. By contrast, the machine-learning framework developed in this chapter treats the full cohort of 32 nucleators as training data, with performance estimated through strict cross-validation. Leave-One-Nucleator-Out CV ensures that each nucleator is evaluated only on predictions made without its own data contributing to training, thereby avoiding the circularity issue that applied to the hand-tuned scheme. This distinction explains why the Chapter 3 confusion matrix is limited to 22 compounds, while the Chapter 5 matrices span all 32, and underlines the conceptual advance: the former demonstrates the feasibility of geometric thresholds, while the latter embeds them in a reproducible, data-driven framework.

This comparison highlights the trade-off between the two approaches: the hand-tuned thresholds prioritised sensitivity, capturing nearly all true working nucleators but at the cost of false positives, while the ROC optimisation balanced recall with precision, yielding a more conservative but more reliable classifier. Importantly, the ROC-based framework also scales better: it can be retrained on larger datasets, resampled for uncertainty quantification, and interpreted through SHAP and bootstrap analyses (see below), features not accessible in the earlier hand-tuned workflow.

Feature attribution using SHAP provides further insight into which geometric descriptors most strongly influence nucleator classification in ice (Figure 5.8). The bar plot of mean absolute SHAP values shows that vector mismatch u is the dominant feature, exerting the largest average impact on model predictions, followed by *m_n_*tolerance and vector mismatch_*v*, while angular mismatch contributes negligibly. The violin plots refine this picture by showing how different feature values drive the predictions. Low values of vector mismatch_*u* (blue) consistently push the model towards predicting a working nucleator, whereas high values (pink) shift predictions towards non-working. A similar, though weaker, trend is seen for *m_n_*tolerance and vector mismatch_*v*, where tighter tolerances and smaller mismatches are favourable. Angle mismatch shows almost no effect across its range. It is important to note that the absolute position of the SHAP distributions (negative or positive) reflects the model's baseline output and is therefore not directly meaningful when comparing across datasets. What matters physically are the colour-coded trends: in all cases, low mismatch or tolerance values are associated with working nucleators, while high values are associated with non-working ones.

The stronger influence of vector mismatch_*u* can be traced to the way the reduced supercell vectors are constructed in the workflow. As implemented in the code:

```
def get_uv(ab, t_mat):
    u = np.array(ab[0]) * t_mat[0][0] + np.array(ab[1]) *
t_mat[0][1]
    v = np.array(ab[1]) * t_mat[1][1]
    return [u, v]
```

Here, `ab[0]` and `ab[1]` are the two primitive in-plane lattice vectors of the slab. The vector **u** is formed as a linear combination of both primitives, and therefore carries the main degrees of freedom for aligning one slab with another. By contrast, **v** is defined as a scaled version of a single primitive direction, following passively once the orientation is fixed by **u**. In practice, this means that mismatch in **u** governs the registry between the two lattices, while mismatch in **v** reflects only the multiples required to complete the supercell. Consequently, vector **u** emerges as the more discriminating descriptor for nucleator classification. This ranking aligns with physical expectations that lattice registry between substrate and ice is governed more strongly by translational lattice matching than by angular alignment.[28]

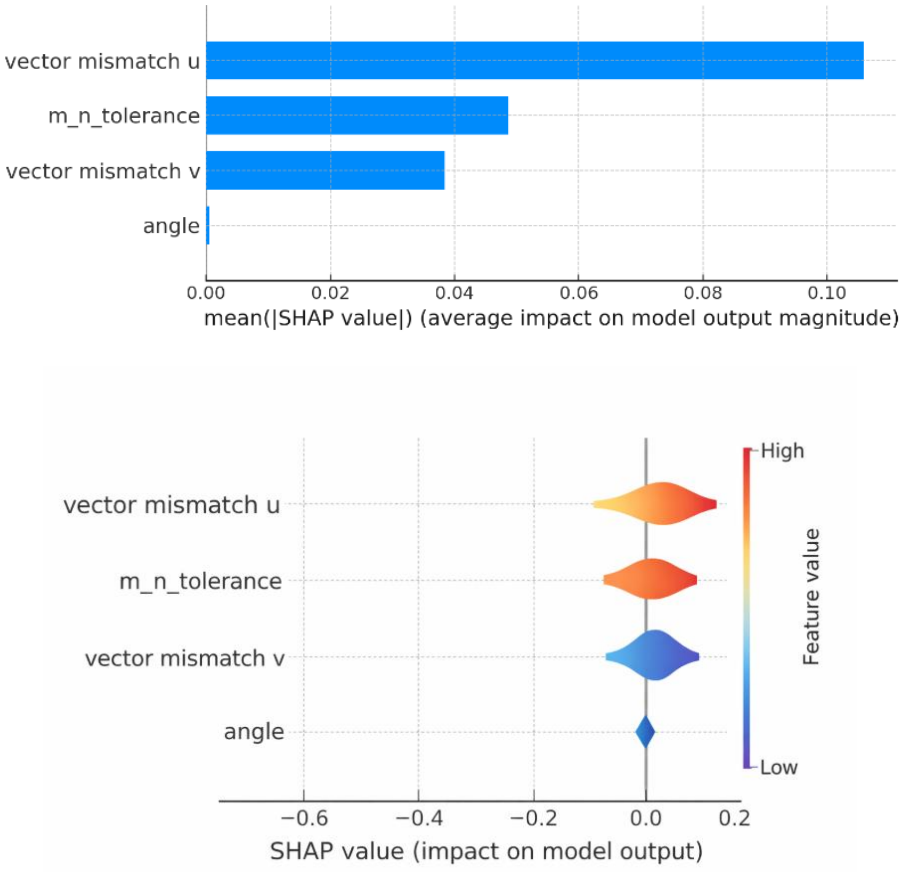


Figure 5.8. Feature attribution for the Random Forest model applied to ice nucleators using SHAP (SHapley Additive exPlanations). Top: Mean absolute SHAP values (bar plot) ranking descriptors by average contribution to predictions. Bottom: SHAP violin plot showing the distribution of per-nucleator contributions, with colour indicating feature value (blue = low, red = high).

Bootstrap re-sampling of RF feature importance offers a more stable picture. Across 200 replicates, the two vector mismatch descriptors consistently emerged as the most influential features, while angular mismatch contributed little and *m_n_*tolerance showed intermediate importance (Figure 5.9). The distributions of feature importance were broad, reflecting uncertainty from the small dataset, but the relative ordering was robust. This dual perspective, *i.e.* SHAP for formal attribution and bootstrap for uncertainty quantification, provides an overall balanced interpretability framework.

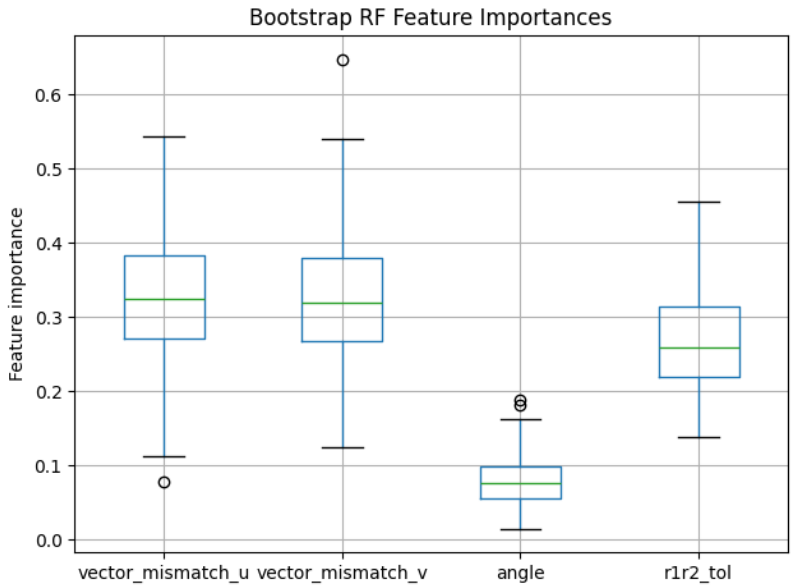


Figure 5.9. Bootstrap resampling of ROC-RF, showing stable central tendency and variance.

The ice benchmark demonstrates that the framework recovers physically meaningful descriptors, achieves moderate predictive performance despite small datasets, and highlights the conditions under which data-driven optimisation improves upon heuristic thresholds. These findings provide confidence that the same methodology can be applied to more complex PCMs in the following sections, and lay the groundwork for testing whether a common rule set can emerge across systems. In particular, the best-performing configuration for ice was obtained with ROC-optimised thresholds at $K = 5$ under logistic regression, yielding a balanced accuracy of 74.6%.

5.5. Results and Discussion: Sodium Acetate Trihydrate (SAT)

Sodium acetate trihydrate (SAT) was selected as the first chemically distinct PCM to test the transferability of the framework beyond ice. Unlike the highly symmetric ice lattice, SAT crystallises in a more complex hydrogen-bonded structure with less predictable cleavage behaviour. This makes it a more stringent benchmark: if the framework recovers meaningful thresholds and descriptors in SAT, it demonstrates that the geometric principles identified in ice are not just pertinent to that particular system. The data set size for this system is 63 unique SAT slabs paired against 18×63=1,134 nucleator slabs, yielding a total of 71,442 candidate slab-slab interfaces. The SAT training labels are more heavily skewed towards non-working nucleators, with only three compounds experimentally identified as active nucleators. This imbalance makes the classification task inherently more challenging, but also more representative of real PCM screening, where most candidate additives are ineffective.

Table 5.3 directly compares thresholding strategies, classifiers, and *K*-values for SAT. The results show that quantile thresholds (q = 0.3) consistently outperformed ROC optimisation across both RF and LR models, reaching the highest overall accuracies of 88.9% at $K = 0$ and $K = 5$. The associated thresholds ($\Delta u = 0.010$, $\Delta v = 0.012$, $\Delta\theta = 0$, m_n_tolerance = 0.008) are more conservative than those derived from ROC ($\Delta u = 0.012$, $\Delta v = 0.026$, $\Delta\theta \approx 10^{-9}$, m_n_tolerance = 0.012), reflecting a stricter filter on acceptable mismatches. Grid optimisation also reached 88.9% accuracy, but with less interpretable thresholds ($\Delta u = 0.007$, $\Delta v = 0.010$, m_n_tolerance = 0.005). In contrast, ROC thresholds plateaued at 55.6-66.7% accuracy regardless of classifier or *K*, indicating that they systematically over-predicted working nucleators. These results make clear that the quantile rule (q = 0.3) offers the most reliable and physically meaningful threshold set for SAT, striking a better balance between recall and precision than either ROC or grid-based optimisation.

Table 5.3. Performance of different thresholding strategies for SAT under Logistic Regression (LR) and Random Forest (RF) classifiers at varying K values. The highlighted rows represent the best performing models.

method	Δu	Δv	Δθ	*m_n_*tolerance	accuracy
K=0 ROC RF	0.012	0.026	9.40×10^{-9}	0.012	55.56%
K=5 ROC RF	0.012	0.026	9.40×10^{-9}	0.012	61.11%
K=10 ROC RF	0.012	0.026	9.40×10^{-9}	0.012	66.67%
K=15 ROC RF	0.012	0.026	9.40×10^{-9}	0.012	66.67%
K=0 quantile=30 RF	0.010	0.012	0	0.008	88.89%
K=5 quantile=30 RF	0.010	0.012	0	0.008	88.89%
K=10 quantile=30 RF	0.010	0.012	0	0.008	83.33%
K=15 quantile=30 RF	0.010	0.012	0	0.008	83.33%
grid RF	0.007	0.010	0	0.005	88.89%
K=0 ROC LR	0.012	0.026	9.40×10^{-9}	0.012	55.56%
K=5 ROC LR	0.012	0.026	9.40×10^{-9}	0.012	61.11%
K=10 ROC LR	0.012	0.026	9.40×10^{-9}	0.012	66.67%
K=15 ROC LR	0.012	0.026	9.40×10^{-9}	0.012	66.67%
K=0 quantile=30 LR	0.010	0.012	0	0.008	88.89%
K=5 quantile=30 LR	0.010	0.012	0	0.008	88.89%
K=10 quantile=30 LR	0.010	0.012	0	0.008	83.33%
K=15 quantile=30 LR	0.010	0.012	0	0.008	83.33%
Grid LR	0.007	0.010	0	0.005	88.89%

Across logistic regression (LR) and random forest (RF) classifiers, quantile-based thresholds (30th percentile) yielded the strongest performance, achieving overall accuracies of 88.9%, with recall of 66.7% and precision of 66.7% (Table 5.4). ROC-optimised thresholds performed notably worse, reaching only 61.1-66.7% accuracy with recall of 100% but much lower precision (33.3%), while grid search optimisation matched the quantile rule in accuracy (88.9%) but achieved lower recall (66.7%) and higher precision (66.7%) at less interpretable cut-offs.

Compared against simple baseline classifiers, random guessing yields an expected accuracy of 50%, while majority-class prediction (assigning all candidates to the dominant non-working class) yields an accuracy of 83.3%. The quantile-based and grid optimisation strategies therefore exceed the majority baseline by approximately 5.6 percentage points, demonstrating genuine discriminative improvement beyond class imbalance effects. In contrast, both the ROC-optimised and manually tuned schemes fall below the majority baseline, indicating that their apparent performance is partially explained by class imbalance rather than true separation between working and non-working nucleators.

Table 5.4. Comparison of accuracy, recall, and precision for the three thresholding strategies applied to ice nucleators (corresponding to Figure 5.10 (a-c)).

Threshold strategy	Classifier	Accuracy	Recall	Precision
ROC-optimised (K=15)	RF/LR	66.7%	100%	33.3%
Quantile rule (q=0.3, K=5)	RF/LR	88.9%	66.7%	66.7%
Grid optimisation	RF/LR	88.9%	66.7%	66.7%
Manual (Chapter 4)	Heuristic	72.7%	60%	100%
Random baseline	N/A	50%	N/A	N/A
Majority baseline	N/A	83.3%	N/A	N/A

The superiority of the quantile rule in this case contrasts with the ice benchmark, where ROC thresholds were consistently the top-performing choice. This inversion likely reflects differences in the distribution of geometric match scores between the two systems. For ice, working and non-working nucleators were more evenly separable along ROC-derived cut-offs, making Youden's J statistic a natural choice. In SAT, however, the score distributions overlap more heavily, with working nucleators not always achieving the strongest geometric matches. As a result, ROC optimisation inflates recall (capturing all true working cases) but at the cost of precision, mis-classifying many non-working nucleators as working. By contrast, the quantile cut-off acts more conservatively, filtering out borderline matches and thereby reducing false positives. This behaviour is consistent with the more heterogeneous lattice environments in SAT, where geometric registry alone may not perfectly track

nucleation propensity, and conservative thresholds help stabilise predictive performance.

The confusion matrices (Figure 5.10) reveal these trade-offs. The quantile rule ($q = 0.3$, $K = 5$) correctly classified 16 out of 18 nucleators, with only two errors, balancing recall (66.7%) and precision (66.7%). ROC-optimised thresholds ($K = 15$) identified all true working nucleators (recall = 100%) but generated six false positives, reducing precision to 33.3% and overall accuracy to 66.7%. Grid optimisation achieved the same accuracy as the quantile rule (88.9%), but relied on statistically optimised thresholds ($\Delta u = 0.007$, $\Delta v = 0.010$, m_n_tolerance = 0.005) that were selected to maximise classification performance rather than derived from physically meaningful strain or geometric tolerances.

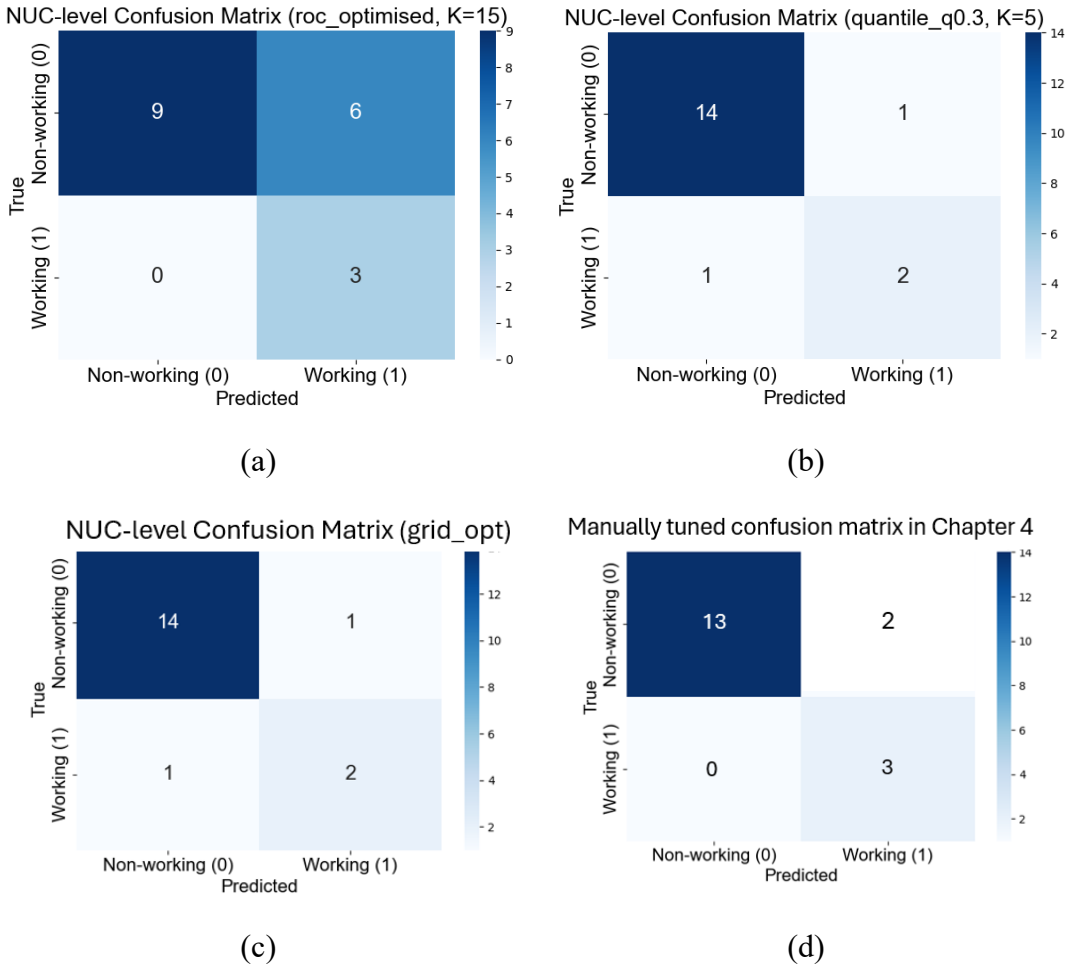


Figure 5.10. Confusion matrices comparing classification performance for SAT nucleators under different thresholding strategies. (a) ROC-optimised thresholds (K = 15) achieved perfect recall but poor precision, misclassifying six non-working nucleators as working; (b) Quantile thresholds (q = 0.3, K = 5) achieved the best overall balance, with only two misclassifications (accuracy 88.9%); (c) Grid optimisation achieved the same overall accuracy but relied on less interpretable thresholds. (d) Manually tuned thresholds from Chapter 4 ($\Delta u = \Delta v = 0.01$, $\Delta\theta = 0.01°$, *m_n*_tolerance = 0.02).

The manually tuned thresholds from Chapter 4 ($\Delta u = \Delta v = 0.01$, $\Delta\theta = 0.01°$, *m_n*_tolerance = 0.02) achieved an overall accuracy of 72.7%, with recall of 60% and precision of 100%. This scheme successfully captured all three working nucleators but produced a relatively high number of false positives. By contrast, the ML-driven quantile rule reduced false positives while maintaining a balanced trade-off, improving

accuracy to 88.9%. This demonstrates how systematic optimisation improves upon hand-tuned cut-offs, particularly in filtering out borderline false positives.

Bootstrap RF and SHAP analyses (Figures 5.11-12) consistently identified the vector mismatch descriptor Δu as the most influential feature for classification, followed by *m_n_*tolerance, while angular mismatch contributed negligibly. This ranking largely mirrors the ice benchmark, where translational registry dominated over angular alignment, but with a stronger role for *m_n_*tolerance in SAT. This difference likely reflects the much larger crystallographic unit cell of SAT compared to ice, that the greater the unit cell size, the more challenging it becomes to achieve commensurate lattice matching, and small deviations in supercell construction tolerance can substantially degrade epitaxial compatibility.

The SHAP violin plots show how each descriptor contributes to the classification of working versus non-working nucleators. The x-axis indicates the direction of impact relative to the model's baseline prediction: values to the right increase the likelihood of being classified as a working nucleator, while values to the left increase the likelihood of a non-working classification. Importantly, the sign of the SHAP values reflects the model's internal baseline and is therefore not physically meaningful across datasets; what matters is the magnitude (feature importance) and the colour distribution (direction of effect). The colour gradient encodes the actual feature values, with blue denoting low values and pink denoting high values. In practice, low values of vector mismatch u, vector mismatch v, and *m_n_*tolerance (blue) are consistently associated with positive SHAP contributions, *i.e.* a greater probability of being predicted as a working nucleator. High values of these descriptors (pink) shift predictions towards non-working. Angular mismatch exerts only a weak influence, though lower values are still marginally more favourable. Together, this confirms the physical intuition that tight translational registry and low supercell tolerance favour epitaxial compatibility.

One thing to note is that although the absolute SHAP values differ in sign between the ice and SAT classifiers, the physical interpretation is consistent across both datasets: low values of vector mismatch_u (blue) shift predictions towards working nucleators,

whereas high values (pink) favour non-working outcomes. The apparent inversion in the ice plots (Figure 5.8) arises from the model's baseline offset, not from a reversal of physical trends. Thus, both systems reinforce the conclusion that translational registry (Δu in particular) is the dominant factor governing heterogeneous nucleation.

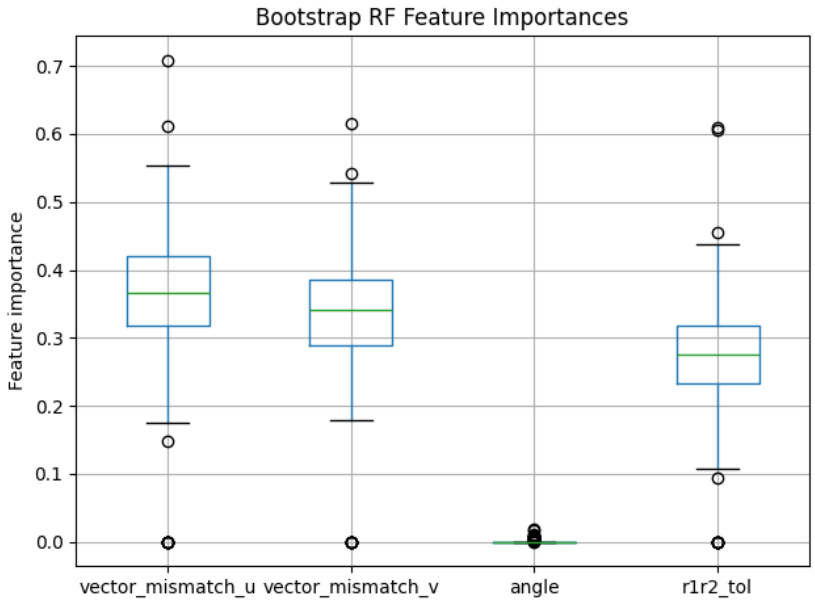


Figure 5.11. Bootstrap random forest feature importances for SAT nucleators.

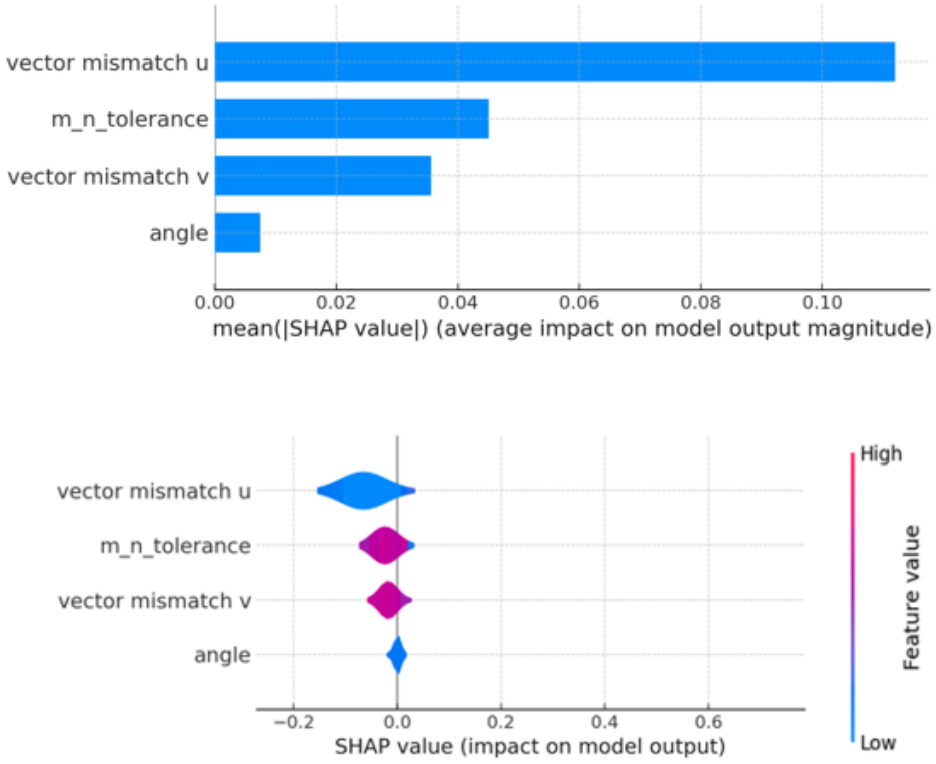


Figure 5.12. SHAP feature attribution for SAT nucleators under a random forest classifier. **(Top)** Bar plot of mean absolute SHAP values showing that vector mismatch u has the greatest influence, followed by *m_n*_tolerance and vector mismatch_v, with angle mismatch negligible. **(Bottom)** Violin plot distributions reveal that low mismatch values (blue) contribute positively to working classifications, whereas higher mismatches (red) push predictions toward non-working.

Together, the SAT results demonstrate that the framework generalises beyond ice, but that the optimal thresholding strategy shifts with crystal chemistry. While ROC thresholds were best for ice, quantile thresholds proved superior for SAT, likely because SAT's heterogeneous lattice makes conservative filtering more robust against false positives. Importantly, feature attribution confirmed that the model recovered physically meaningful descriptors, with vector mismatches consistently emerging as the dominant predictors of nucleation activity. SAT thus serves as a bridge system, showing both continuity (vector mismatch dominance) and divergence (different optimal thresholds) relative to ice, and reinforcing the framework's portability.

Whether such portability extends to a single transferable model across systems is assessed in the pooled analysis.

5.6. Results and Discussion: Salt hydrates as a stress test of framework generalisability

The next evaluation of the geometric machine learning framework was conducted on the set of four salt hydrates: $CaCl_2 \cdot 6H_2O$, $Mg(NO_3)_2 \cdot 6H_2O$, $LiNO_3 \cdot 3H_2O$, and $MgCl_2 \cdot 6H_2O$, previously reported via manual tuning in Chapter 4. These systems represent the most stringent test of the pipeline. Unlike ice (Section 5.2), which offered a large and well-curated dataset for benchmarking, or sodium acetate trihydrate (Section 5.4), where a moderate volume of high-quality data enabled transferability testing, the salt hydrates are characterised by fragmented and imbalanced experimental validation data, comprising 78 nucleators across 309,582 slab pairings in total of 4 salt hydrate PCMs. This inherent data limitation makes them a genuine stress test for robustness, reproducibility, and generalisability of the approach.

In this chapter, the ML classifiers were initially trained and evaluated separately for each PCM, hence the individual performance scores reported in Table 5.5. This was necessary given the imbalance in data volume and quality across the four systems, but it also raises a key conceptual question: is PCM-specific tuning essential, or could a combined model spanning multiple salts (and even SAT) achieve transferable predictive power? The system-specific analyses suggest that while the same descriptors consistently dominate across PCMs, the precise threshold values optimising performance vary from one material to another, especially for hydrates with large or low-symmetry unit cells where registry constraints are more severe. To address this question directly, a pooled dataset comprising all data (ice, SAT and the four other salt hydrates) was also constructed, providing a first test of whether a single cross-PCM model can retain predictive accuracy. The results of this pooled analysis are presented in Section 5.6.

Table 5.5. Performance of different thresholding strategies for $CaCl_2 \cdot 6H_2O$, $Mg(NO_3)_2 \cdot 6H_2O$, $LiNO_3 \cdot 3H_2O$, and $MgCl_2 \cdot 6H_2O$ under Logistic Regression (LR) and Random Forest (RF) classifiers at varying K values. The highlighted rows represent the best performing models.

NUC	method	Δu	Δv	Δθ	*m_n_*tolerance	accuracy
$CaCl_2 \cdot 6H_2O$	*K*=0 ROC RF/LR	0.031	0.005	5.75×10^{-8}	0.011	72.22%
	K=5 ROC RF/LR	0.031	0.005	5.75×10^{-8}	0.011	66.67%
	K=10 ROC RF/LR	0.031	0.005	5.75×10^{-8}	0.011	66.67%
	K=15 ROC RF/LR	0.031	0.005	5.75×10^{-8}	0.011	55.56%
	K=0 quantile=30 RF/LR	0.006	0.007	0	0.009	38.89%
	K=5 quantile=30 RF/LR	0.006	0.007	0	0.009	38.89%
	K=10 quantile=30 RF/LR	0.006	0.007	0	0.009	38.89%
	K=15 quantile=30 RF/LR	0.006	0.007	0	0.009	38.89%
	grid	0.006	0.007	0	0.009	38.89%
	Manual	0.01	0.01	0.01	0.02	89.52%
$Mg(NO_3)_2 \cdot 6H_2O$	*K* =0 ROC RF/LR	0.008	0.011	0.008	0.022	66.67%
	K =5 ROC RF/LR	0.008	0.011	0.008	0.022	63.33%
	K =10 ROC RF/LR	0.008	0.011	0.008	0.022	56.67%
	K =15 ROC RF/LR	0.008	0.011	0.008	0.022	53.33%
	K =0 quantile=30 RF/LR	0.011	0.012	1.42×10^{-14}	0.011	53.33%
	K =5 quantile=30 RF/LR	0.011	0.012	1.42×10^{-14}	0.011	56.67%
	K =10 quantile=30 RF/LR	0.011	0.012	1.42×10^{-14}	0.011	53.33%
	K =15 quantile=30 RF/LR	0.011	0.012	1.42×10^{-14}	0.011	50.00%
	grid	0.007	0.012	6.76×10^{-9}	0.013	60.0%
	Manual	0.01	0.01	0.01	0.02	76.66%
$LiNO_3 \cdot 3H_2O$	*K*=0 ROC RF/LR	0.028	0.011	4.23×10^{-10}	0.028	72.73%
	K =5 ROC RF/LR	0.028	0.011	4.23×10^{-10}	0.028	72.73%
	K=10 ROC RF/LR	0.028	0.011	4.23×10^{-10}	0.028	72.73%
	K=15 ROC RF/LR	0.028	0.011	4.23×10^{-10}	0.028	72.73%
	K=0 quantile=30 RF/LR	0.011	0.006	0	0.015	63.64%
	K=5 quantile=30 RF/LR	0.011	0.006	0	0.015	54.55%
	K=10 quantile=30 RF/LR	0.011	0.006	0	0.015	54.55%
	K=15 quantile=30 RF/LR	0.011	0.006	0	0.015	45.45%
	grid	0.011	0.011	0	0.024	72.73%
	Manual	0.01	0.01	0.01	0.02	81.84%
$MgCl_2 \cdot 6H_2O$	*K*=0 ROC RF/LR	0.016	0.012	0.039	0.007	78.95%
	K=5 ROC RF/LR	0.016	0.012	0.039	0.007	84.27%

	K=10 ROC RF/LR	0.016	0.012	0.039	0.007	78.95%
	K=15 ROC RF/LR	0.016	0.012	0.039	0.007	73.68%
	K=0 quantile=30 RF/LR	0.008	0.012	0	0.010	61.11%
	K=5 quantile=30 RF/LR	0.008	0.012	0	0.010	61.11%
	K=10 quantile=30 RF/LR	0.008	0.012	0	0.010	61.11%
	K=15 quantile=30 RF/LR	0.008	0.012	0	0.010	55.56%
	grid	0.006	0.008	0	0.007	78.95%
	Manual	0.01	0.01	0.01	0.02	84.27%

From Table 5.5, it can be seen that performance of this ML prediction model was inevitably lower and more variable than for ice or SAT. For example, $CaCl_2 \cdot 6H_2O$ retained relatively strong agreement with experiments (~ 90% accuracy), whereas $Mg(NO_3)_2 \cdot 6H_2O$ proved more challenging, with performance falling closer to 75%. $LiNO_3 \cdot 3H_2O$ and $MgCl_2 \cdot 6H_2O$ lay between these extremes. These differences partly reflect the greater experimental uncertainty associated with nucleation measurements in salt hydrates, and partly the small number of tested nucleators per system, which amplifies the influence of individual compounds. In such a setting, a single misclassified nucleator can swing apparent accuracy by 5-10 percentage points, underlining why bootstrap resampling (Figure 5.13) was employed to provide more reliable uncertainty bounds.

A further point is that the best accuracies for the hydrates were generally achieved at lower *K* values (0 or 5), indicating that effective nucleators in these systems may be identifiable even from a small number of matching interfaces. This contrasts with ice and SAT, where higher K values provided more robust predictions, and highlights the impact of data sparsity on aggregation rules. In addition, the numerical values of the four descriptor thresholds (Δu, Δv, $\Delta\theta$, and *m_n*_tolerance) vary substantially across the hydrates and compared with SAT. For instance, $CaCl_2 \cdot 6H_2O$ was optimised with Δu = 0.006-0.01 and *m_n*_tolerance = 0.009-0.02, whereas $Mg(NO_3)_2 \cdot 6H_2O$ required larger Δv values (0.011-0.012) and looser tolerances. These variations emphasise that there is no single universal set of cut-offs, but rather that optimal geometric criteria depend strongly on the crystallography of the PCM in question.

Despite these challenges, the qualitative behaviour of the model remains consistent with that observed for ice and SAT. Across the four salt hydrates, an average of ~ 65%

of non-working nucleators were correctly predicted as having zero matches (see confusion matrices), which provides a semi-quantitative descriptor of model reliability in filtering false positives.

To quantify stability under sparse data, the salt-hydrate datasets were also bootstrapped and performance metrics recomputed (Figure 5.13). The distributions make two points clear. First, the rank-ordering of thresholding strategies seen in Table 5.4 is stable under resampling (quantile/grid $\geqslant$ ROC), indicating the conclusions are not artifacts of a single split. Second, the spread reflects data scarcity and class imbalance: systems with more marginal or few non-working examples exhibit broader intervals (*e.g.*, $MgCl_2 \cdot 6H_2O$, $MgNO_3 \cdot 6H_2O$), whereas systems dominated by clear positives/negatives show tighter distributions (*e.g.*, $CaCl_2 \cdot 6H_2O$). In all cases, the bootstrap medians align with the point estimates reported in Table 5.4, but the interval widths remind us that precision is limited by sample size, *i.e.* a key reason to report uncertainty alongside accuracy in these stress-test systems.

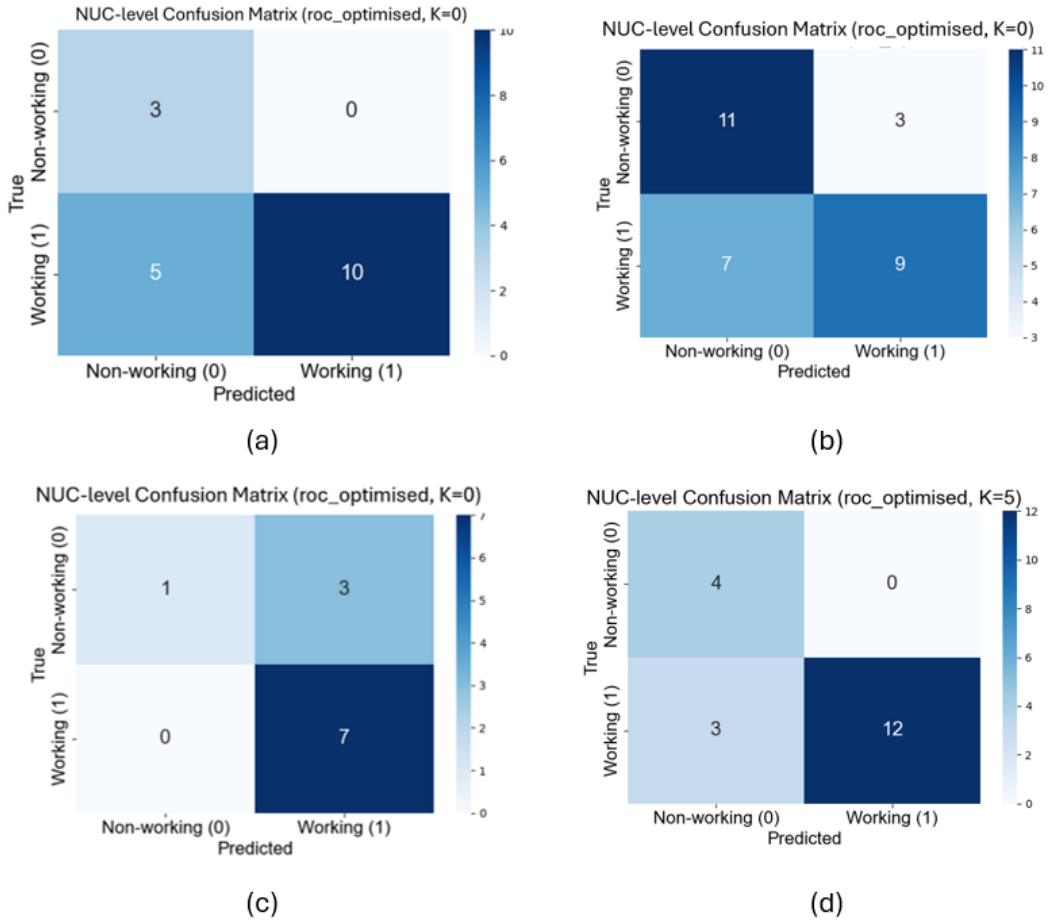


Figure 5.13. Confusion matrices with best classification performance for four representative PCM systems **(a)** $CaCl_2{\cdot}6H_2O$, **(b)** $Mg(NO_3)_2{\cdot}6H_2O$, **(c)** $LiNO_3{\cdot}3H_2O$, and **(d)** $MgCl_2{\cdot}6H_2O$.

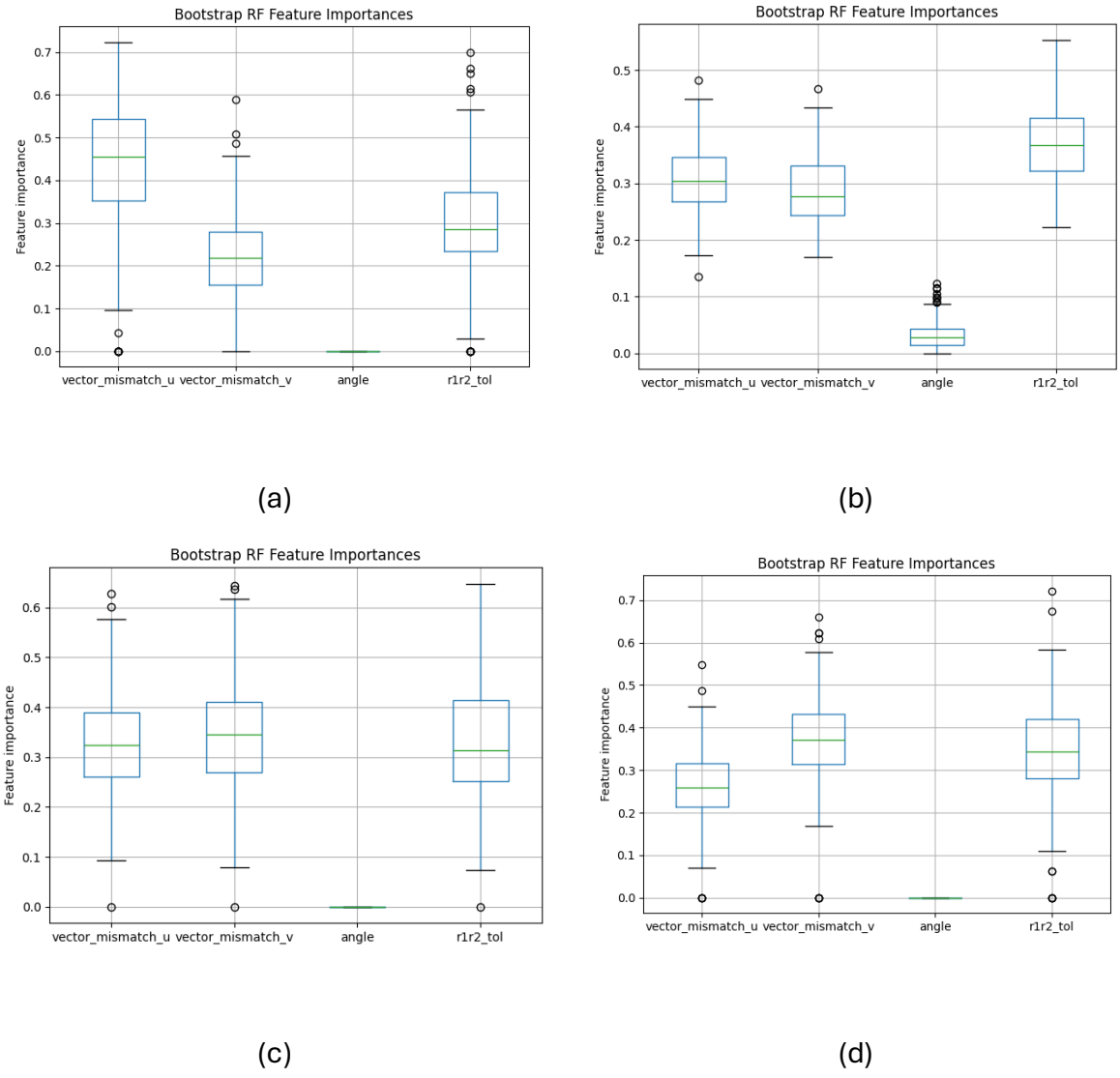
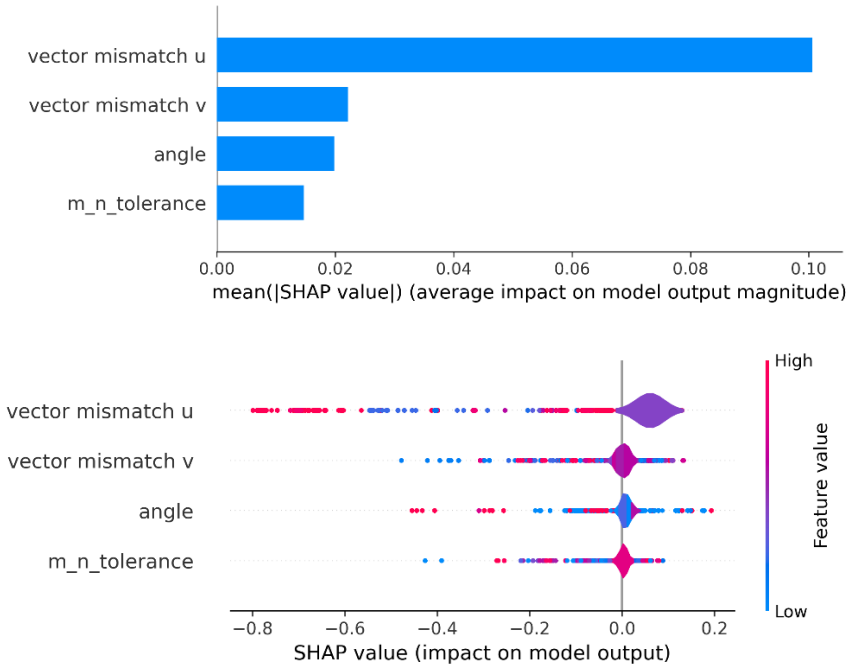


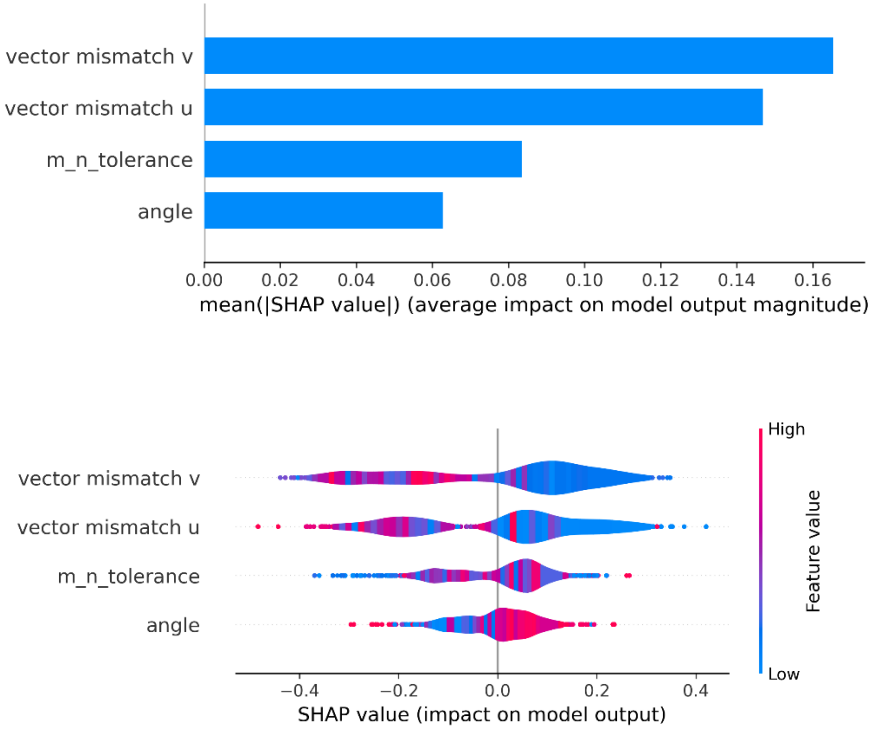
Figure 5.14. Bootstrap random forest feature importances for nucleators of **(a)** $CaCl_2 \cdot 6H_2O$, **(b)** $Mg(NO_3)_2 \cdot 6H_2O$, **(c)** $LiNO_3 \cdot 3H_2O$, and **(d)** $MgCl_2 \cdot 6H_2O$.

SHAP analyses further reinforce continuity with earlier systems. Although absolute SHAP values are noisier under limited data, the ranking of descriptors remained broadly stable: vector mismatches (particularly along vector_u) and *m_n*_tolerance emerged as the most influential, while angular mismatch contributed negligibly (Figure 5.15). This mirrors the findings from both ice (Figure 5.8) and SAT (Figure 5.11), underscoring that the geometric descriptors capture transferable physical principles rather than system-specific artefacts. The noisier distributions observed here are consistent with the limited dataset size, yet importantly, the signal was never lost

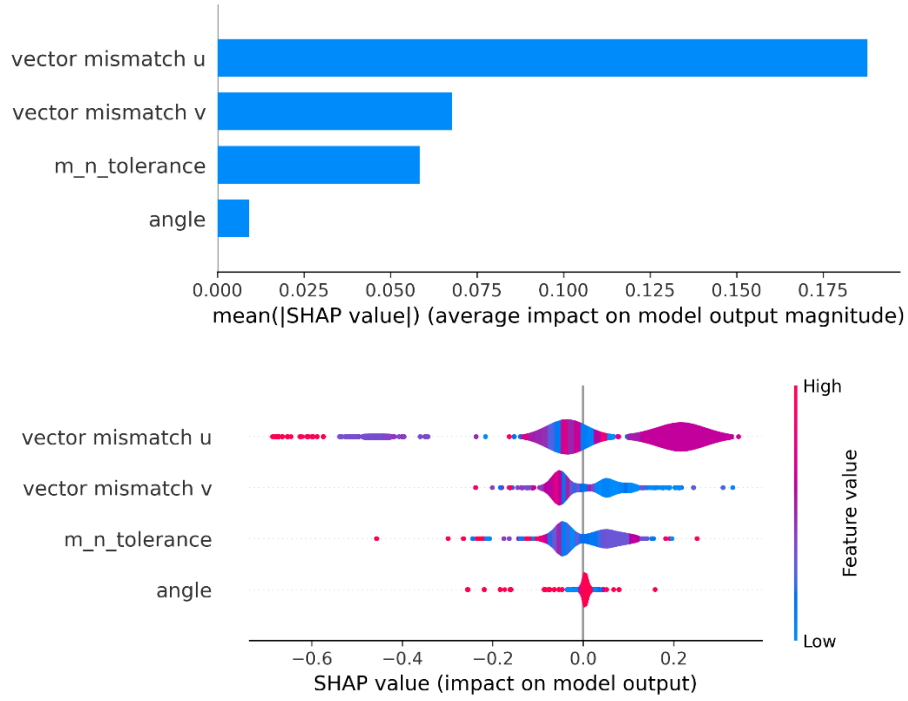
entirely, demonstrating that the workflow continues to extract physically meaningful features under constrained conditions.



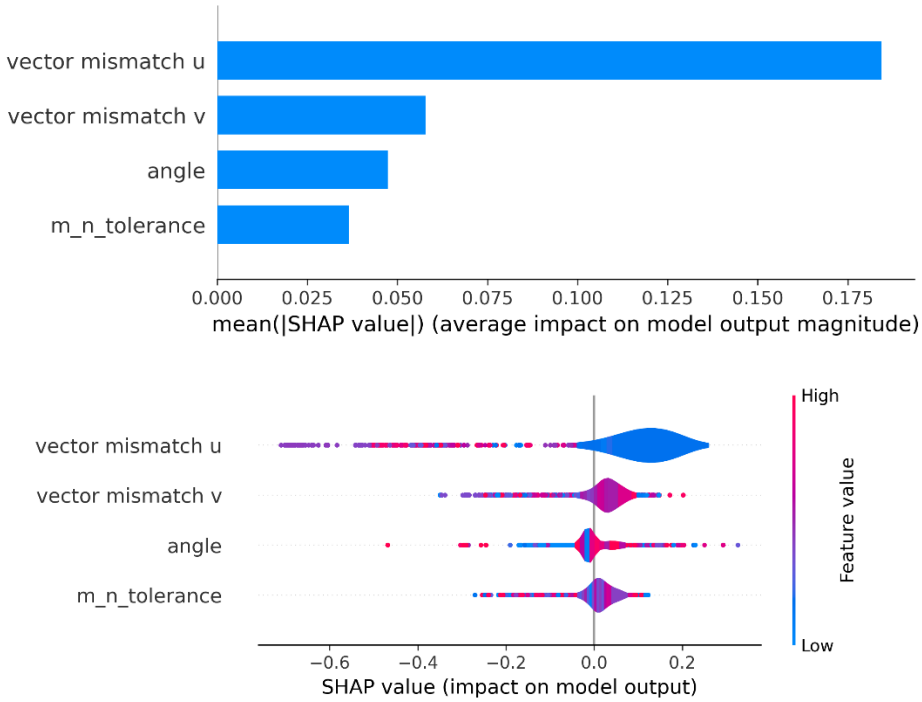
(a)



(b)



(c)



(d)

Figure 5.15. Feature attribution for the Random Forest model applied to nucleators of **(a)** $CaCl_2 \cdot 6H_2O$, **(b)** $Mg(NO_3)_2 \cdot 6H_2O$, **(c)** $LiNO_3 \cdot 3H_2O$, and **(d)** $MgCl_2 \cdot 6H_2O$ using SHAP

(SHapley Additive exPlanations). Top: Mean absolute SHAP values (bar plot) ranking descriptors by average contribution to predictions. Bottom: SHAP violin plot showing the distribution of per-nucleator contributions, with colour indicating feature value (blue = low, red = high).

5.7 Cross-PCM pooled model

Finally, to test whether a single transferable model could be established across multiple systems, all available nucleator-PCM data (ice, SAT, and four salt hydrates) were pooled into a unified dataset. This comprised 130 nucleators, each represented by 63 slabs, giving a total of approximately 508,000 candidate slab-slab interfaces (127,008 for ice, 71,442 for SAT, and 309,582 across the four hydrates).

Table 5.6 shows the performance of ROC-optimised, quantile, and grid search strategies for the pooled dataset. The optimised thresholds converge on intermediate values ($\Delta u = 0.012$, $\Delta v = 0.011$, $\Delta\theta = 2.52\times10^{-8}$, *m_n_*tolerance = 0.026), which lie between those observed in the single-system models. For example, Δu and Δv cut-offs are tighter than the more permissive values found in $Mg(NO_3)_2\cdot6H_2O$ but broader than those for $CaCl_2\cdot6H_2O$; similarly, the pooled *m_n_*tolerance is larger than the strict SAT optimum (0.008) but smaller than hydrate-specific values (0.022-0.028). This averaging behaviour is expected when combining PCMs with distinct crystallographic unit cells and mismatch distributions, and suggests that the pooled model balances different registry constraints by converging on compromise thresholds.

Table 5.6. Performance of different thresholding strategies for the universal PCM dataset (ice, SAT, and four salt hydrates combined) under Logistic Regression (LR) and Random Forest (RF) classifiers at varying K values. The highlighted rows represent the best-performing models.

method	Δu	Δv	Δθ	m_n_tolerance	accuracy
K=0 ROC RF/LR	0.012	0.011	2.52×10^{-8}	0.026	66.15%
K=5 ROC RF/LR	0.012	0.011	2.52×10^{-8}	0.026	68.46%
K=10 ROC RF/LR	0.012	0.011	2.52×10^{-8}	0.026	66.15%
K=15 ROC RF/LR	0.012	0.011	2.52×10^{-8}	0.026	63.85%
K=0 quantile=30 RF/LR	0.008	0.010	0	0.011	66.15%
K=5 quantile=30 RF/LR	0.008	0.010	0	0.011	60.00%

K=10 quantile=30 RF/LR	0.008	0.010	0	0.011	56.92%
K=15 quantile=30 RF/LR	0.008	0.010	0	0.011	55.38%
grid	0.079	0.011	0	0.010	66.15%

The overall accuracy reached 68.46% under ROC optimisation at *K*=5, broadly comparable to the individual PCM models, though slightly below the best single-system results (*e.g.* 74-89% for ice, SAT, and some hydrates). Thus, while the universal model is less accurate than PCM-specific tuning, it does not collapse under the heterogeneity of the data, retaining moderate predictive ability across chemically distinct systems. This suggests that a pooled approach may provide a viable transferable predictor, particularly for screening PCMs where no experimental calibration is available, albeit at the cost of some precision relative to PCM-specific models.

The confusion matrix in Figure 5.16 provides further insight. Out of 130 nucleators, the model correctly identified 52 working nucleators (true positives, TP) and 37 non-working nucleators (true negatives, TN). However, 27 working nucleators were missed (false negatives, FN), and 14 non-working nucleators were incorrectly classified as working (false positives, FP). This pattern reflects a moderately conservative classifier: false positives are relatively rare (only 14 cases), which means that nucleators flagged as working are likely to be genuine. By contrast, the larger number of false negatives indicates reduced sensitivity: the universal model misses a subset of genuine nucleators, underestimating their activity. In practice, this bias may be acceptable in exploratory screening, where avoiding false claims of nucleation activity (FPs) is more important than achieving exhaustive recall.

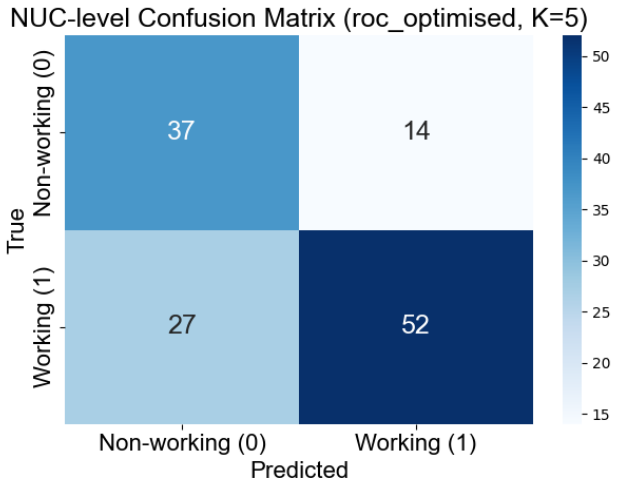


Figure 5.16. Confusion matrix for the pooled PCM dataset using ROC-optimised thresholds at $K = 5$. Out of 130 nucleators, the model correctly identified 52 working nucleators and 37 non-working nucleators, with 27 false negatives and 14 false positives.

The pooled analysis confirms that vector mismatch descriptors and $m_n_$tolerance retain predictive value across diverse systems, and that a single set of thresholds can achieve workable classification accuracy. Nevertheless, as expected, PCM-specific tuning still provides superior performance, particularly for well-characterised systems such as ice and SAT. The universal model should therefore be viewed not as a replacement, but as a complementary strategy: it offers a 'good enough' transferable predictor where no calibration data exist, while PCM-specific optimisation remains preferable when sufficient validation is available.

5.8 Conclusions

The results highlight the hierarchical narrative of framework validation. Ice served as a benchmark proof of concept, showing that physically interpretable descriptors can be linked to nucleation outcomes when abundant data are available. SAT provided a transferability test, demonstrating that the same principles apply to a chemically distinct PCM with a different lattice structure and hydrogen-bonding network. The salt hydrates extended this progression further, acting as a stress test for robustness under

sparse and fragmented data, and showing that even under such conditions the pipeline recovers physically meaningful descriptors while transparently conveying uncertainty.

The most interesting step was the pooled analysis, in which all available PCM systems were combined into a single dataset. This experiment showed that intermediate threshold values emerge when diverse systems are considered together, and that a universal model can retain moderate predictive power (68% accuracy) across more than 500,000 candidate interfaces. Although such pooled models do not yet match the accuracy of PCM-specific tuning, they demonstrate that the same descriptors remain predictive across chemically diverse systems, offering a first step towards a genuinely transferable screening tool. In contexts where no calibration data are available, such universal models could provide a valuable starting point, while PCM-specific optimisation remains the preferred route when sufficient experimental validation exists.

The key outcome of this chapter is therefore twofold. First, the framework is portable: it can be applied reproducibly across chemically distinct PCMs without modification of descriptors or aggregation rules. Second, it is beginning to show signs of transferability, as a pooled model achieves workable classification across multiple systems. These findings point towards a future research agenda that balances system-specific optimisation with the development of universal cross-PCM predictors, ultimately aiming to provide scalable, generalisable tools for heterogeneous nucleation screening.

Thus, the validation ladder now spans four levels: ice as a benchmark, SAT as a transferability test, salt hydrates as a robustness stress test, and finally the pooled dataset as a first demonstration of cross-PCM generalisation.

References

1. S. M. Lundberg and S.-I. Lee, *Advances in Neural Information Processing Systems*, 2017, **30**.
2. W. McKinney, *Scipy*, 2010, **445**, 51-56.
3. A. Naik, Y. Wu, M. Naik and E. Wong, *International Conference on Machine Learning*, 2023, **202**, 25677-25693.
4. C. E. Metz, *Seminars in Nuclear Medicine*, 1978, **8**, 283-298.
5. K. H. Zou, W. J. Hall and D. E. Shapiro, *Statistics in medicine*, 1997, **16**, 2143-2156.
6. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss and V. Dubourg, *the Journal of Machine Learning Research*, 2011, **12**, 2825-2830.
7. Á. B. Jiménez, J. L. Lázaro and J. R. Dorronsoro, in *Innovations in Hybrid Intelligent Systems*, Springer, 2008, pp. 120-127.
8. D. R. Cox, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 1958, **20**, 215-232.
9. D. Maulud and A. M. Abdulazeez, *Journal of Applied Science and Technology trends*, 2020, **1**, 140-147.
10. C. Cortes, M. Mohri and A. Rostamizadeh, *arXiv preprint arXiv:1205.2653*, 2012.
11. R. W. Hoerl, *Technometrics*, 2020, **62**, 420-425.
12. L. Breiman, *Machine Learning*, 2001, **45**, 5-32.
13. A. Cutler, D. R. Cutler and J. R. Stevens, in *Ensemble Machine Learning*, Springer, 2012, pp. 157-175.
14. J. Luan, C. Zhang, B. Xu, Y. Xue and Y. Ren, *Fisheries Research*, 2020, **227**, 105534.
15. A. Nadi and H. Moradi, *Expert Systems with Applications*, 2019, **138**, 112801.
16. B. Gregorutti, B. Michel and P. Saint-Pierre, *Statistics and Computing*, 2017, **27**, 659-678.
17. C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin and A. Zeileis, *BMC Bioinformatics*, 2008, **9**, 307.
18. L. Auret and C. Aldrich, *Chemometrics and Intelligent Laboratory Systems*, 2011, **105**, 157-170.
19. I. Covert, S. M. Lundberg and S.-I. Lee, *Advances in Neural Information Processing Systems*, 2020, **33**, 17212-17223.
20. S. Varma and R. Simon, *BMC Bioinformatics*, 2006, **7**, 91.
21. S. Bates, T. Hastie and R. Tibshirani, *Journal of the American Statistical Association*, 2024, **119**, 1434-1445.
22. L. Merrick and A. Taly, *International Cross-Domain Conference for Machine Learning and Knowledge Extraction,* 2020, **12279**, 17-38.
23. Y. Kim and Y. Kim, *Sustainable Cities and Society*, 2022, **79**, 103677.
24. A. Palczewska, J. Palczewski, R. Marchese Robinson and D. Neagu, in *Integration of Reusable Systems*, Springer, 2014, pp. 193-218.
25. J. I. Gimenez-Nadal, M. Lafuente, J. A. Molina and J. Velilla, *Empirical Economics*, 2019, **56**, 233-267.
26. C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S.

Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke and T. E. Oliphant, *Nature*, 2020, **585**, 357-362.
27. J. D. Hunter, *Computing in Science & Engineering*, 2007, **9**, 90-95.
28. M. Camarillo, J. Oller-Iscar, M. M. Conde, J. Ramírez and E. Sanz, *The Journal of Chemical Physics*, 2024, **160**.

Chapter 6

Conclusion and Future Perspectives

The overarching motivation of this thesis was set out in Chapter 1: to move beyond heuristic or trial-and-error approaches to for heterogenous nucleator discovery, and to establish a framework capable of predicting nucleation behaviour across a variety of phase-change materials. Sections 1.5 and 1.6 laid out the central research questions: can geometric descriptors of slab matching capture the key physical principles underlying heterogeneous nucleation; are these descriptors transferable across chemically distinct PCMs; can they be embedded into a machine learning workflow that is interpretable, systematic, and robust; and to what extent can such a framework provide practical guidance in the face of limited or noisy experimental data. The work presented across Chapters 2 to 5 addresses each of these questions in turn, culminating in a coherent framework for generalisable prediction.

The first question concerned whether geometric compatibility between a nucleator surface and a crystallising phase can be quantified in a way that meaningfully predicts nucleation activity. Chapter 2 introduced a slab-matching algorithm designed precisely for this purpose, producing descriptors that capture translational mismatch, angular alignment, supercell tolerance, and interface area. Chapter 3 then put this approach to the test using ice as a benchmark system. By comparing geometric predictions against a library of twenty-two experimentally characterised nucleators, the analysis demonstrated that materials yielding large numbers of well-matched interfaces tended also to exhibit strong nucleation activity in immersion freezing conditions. Although geometric matching was not a perfect predictor, i.e. some false positives emerged where geometry alone could not account for chemical or kinetic subtleties, the study confirmed that structural registry is indeed a dominant factor. Thus, the first research question is answered affirmatively: geometric slab matching provides a physically grounded, quantifiable, and predictive handle on heterogeneous nucleation.

The second question asked whether the same geometric descriptors identified in ice are transferable to chemically distinct systems. To address this, sodium acetate trihydrate was chosen as a test case in Chapter 4. Unlike ice, SAT possesses a more

complex hydrogen-bonded lattice with different cleavage behaviour, raising the possibility that descriptors tuned on ice might not generalise. The results showed that while threshold values shifted in response to the new lattice environment, the same descriptors retained predictive power. Nucleators that strongly reduced subcooling in SAT were those that exhibited non-zero geometric matches under the slab-matching framework, whereas ineffective additives corresponded to zero matches. Importantly, this was achieved without retraining or modifying the underlying algorithm, apart from a minor refinement to include slab area ratio as an additional descriptor. This framework was shown to be transferrable to three further, data scarce, salt hydrate PCMs. The conclusion is that the framework is not locked into the peculiarities of ice: it can adapt to chemically distinct PCMs, demonstrating genuine transferability.

A third question concerned how best to formalise threshold selection and whether machine learning could add value without sacrificing interpretability. Chapter 5 addressed this by embedding the geometric descriptors into logistic regression and random forest classifiers, and by testing three thresholding strategies: quantile rules, ROC-optimised cut-offs, and joint grid search. On ice, ROC thresholds produced the most balanced performance, recovering the same physically meaningful cut-offs as the manually tuned scheme but with improved precision. In SAT, however, the quantile rule emerged as optimal, reflecting differences in the distribution of geometric scores across systems. The fact that different PCMs favour different thresholding strategies is itself informative: it shows that while the descriptors are portable, their operationalisation requires system-sensitive optimisation. Crucially, the machine learning framework provided not only higher predictive performance but also uncertainty quantification via bootstrap resampling and interpretability via SHAP analyses. These additions directly answer the research question about whether data-driven methods can contribute more than black-box predictions: they can, by making thresholding systematic, reproducible, and transparent.

The final question set in Chapter 1 asked whether such a framework could remain useful under the most difficult conditions: small, sparse, and noisy datasets. Here, three further data-sparse salt hydrates provided the most stringent stress test. With limited experimental validation and pronounced variability, these systems exposed the limits

of predictive accuracy. Yet even here, the workflow continued to operate reproducibly. Bootstrap analyses revealed that although performance metrics fluctuated more widely, the relative ranking of threshold strategies was stable, and SHAP confirmed the continued dominance of vector mismatches and supercell tolerance as key descriptors. Across the four salt hydrates examined, approximately two-thirds of experimentally non-working nucleators were predicted correctly as having zero matches, offering a semi-quantitative measure of reliability. These results do not suggest that high accuracy is achievable in such constrained settings, but they do show that the framework does not collapse under data scarcity. Instead, it provides physically interpretable signals, quantifies uncertainty, and offers guidance on where predictions are more or less reliable. When all datasets were pooled and treated collectively, the final machine learning models confirmed that the same set of geometric descriptors retained predictive value across chemically diverse systems. Training on the full corpus of nucleators provided a more stable basis for threshold selection, smoothing over system-specific fluctuations while consistently highlighting vector mismatches and supercell tolerance as the dominant features. Although absolute accuracy was inevitably diluted by the heterogeneity of the data, this unified model underscores the wider applicability of the framework, that nucleation behaviour across different PCMs can be rationalised within a single descriptor space, without the need to construct bespoke models for each material system.

Beyond answering these core research questions, the work makes broader methodological contributions. It establishes a reproducible algorithm for geometric slab matching, demonstrates its transferability across distinct PCM chemistries, integrates it with interpretable machine learning, and validates it across a ladder of increasing difficulty: ice as a benchmark, SAT as a transferability test, and salt hydrates as a stress test. This progression shows not only that geometric descriptors are predictive, but that they are portable across systems and resilient under challenging data regimes.

Limitations of this research need to be addressed. Experimental labels are defined at the nucleator level, whereas the specific crystallographic faces responsible for nucleation remain uncertain. The framework therefore aggregates across many slab

pairings without being able to identify the singular active interface. Moreover, geometry alone cannot capture the role of chemical bonding, hydration, or kinetic effects, which are known to influence nucleation. In this sense, the framework inherits the assumptions of the theoretical models on which it is grounded: slab matching operationalises the epitaxial growth model, where lattice registry is treated as the dominant control, but it cannot account for the kinetic prefactors and interfacial free energies emphasised in Classical Nucleation Theory. The thresholds themselves also remain sensitive to dataset composition and size, meaning that performance metrics can shift as new nucleators are added. These limitations do not undermine the core findings, but they highlight the scope for extending the framework beyond epitaxy-based descriptors towards models that integrate chemical interactions and dynamical effects.

Future work should pursue several directions. Expanding experimental datasets is paramount, as larger and more diverse libraries of nucleators will reduce uncertainty and enable external validation. Integrating chemical descriptors alongside geometric ones offers another pathway, particularly features capturing surface chemistry, hydration layers, or ion coordination. Linking slab matching with atomistic simulations could provide a multi-scale perspective, connecting the static geometry of crystal faces with the dynamic processes of nucleation. From a methodological standpoint, extending bootstrap analysis into full Bayesian treatments could yield probabilistic predictions with confidence intervals, improving decision-making in data-scarce environments. Finally, there are clear opportunities for industrial application. In thermal energy storage, where the discovery of effective nucleators for PCMs remains largely empirical, a generalisable predictive tool could accelerate material selection, reduce costs, and enable rational formulation design. Similar benefits extend to cryopreservation, pharmaceuticals, and other fields where controlling nucleation is critical.

The conclusion that emerges from this thesis is that heterogeneous nucleation, long regarded as resistant to prediction, can in fact be approached with a generalisable, data-driven methodology. By combining crystallographic intuition with machine learning, it is possible to move from descriptive heuristics to reproducible predictions. The

framework developed here recovers physically meaningful descriptors, adapts to new PCM chemistries, remains interpretable under data scarcity, and highlights uncertainty transparently. It does not offer perfect accuracy, nor does it claim to capture all aspects of nucleation, but it provides a foundation on which more comprehensive models can be built.

In summary, the research questions posed at the outset have been answered. Geometric descriptors are predictive; they are transferable across distinct PCMs; machine learning provides systematic and interpretable thresholding; and the framework remains useful, albeit noisier, under sparse data. Taken together, these findings establish a progressive validation ladder from ice, through SAT, to salt hydrates, illustrating benchmark performance, chemical transferability, and robustness under stress. The work thus provides not only specific results but a methodological foundation: a portable, interpretable, and extensible framework for heterogeneous nucleation prediction. With further refinement and expansion, it has the potential to transform nucleator discovery from a largely empirical exercise into a rational and predictive science.

Appendix A

Publications

[Finding heterogeneous nucleating agents for ice using a data-driven approach]

Accepted for publication in PCCP, on 22 September 2025.

Appendix B

Code availability

All scripts and code used in this thesis, including the slab-matching pipeline and machine learning workflows, are openly available on GitHub at:

https://github.com/mirandawangorange-creator/geometric-matching-NUC-finder

The repository contains fully documented Python scripts, example input/output files, and instructions for reproducing the analyses described in Chapters 2-5.

Appendix C

Statement of experimental contributions

All experimental design, data analysis, computational modelling, and interpretation presented in this thesis were conducted by the author. The only exceptions are the polar bear ice nucleation experiments and the PXRD characterisations, which were performed by Professor Carole Morrison under the author's direction. The resulting data were analysed and interpreted by the author and integrated into the broader framework of this research.