Modelling and Data Analysis of Stochastic Nucleation in Crystallisation

GIOVANNI MARIA MAGGIONI

MODELLING AND DATA ANALYSIS OF STOCHASTIC NUCLEATION IN CRYSTALLISATION

DISS. ETH NO. 24795

MODELLING AND DATA ANALYSIS OF STOCHASTIC NUCLEATION IN CRYSTALLISATION

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by

GIOVANNI MARIA MAGGIONI
Ms. sc., Politecnico di Milano

born on 29th May 1988
Citizen of Italy

Accepted on the recommendation of

Prof. Dr. Marco Mazzotti (ETH Zurich), examiner
Prof. Dr. Paolo Arosio (ETH Zurich), co-examiner
Prof. Dr. Richard D. Braatz (MIT Boston), co-examiner

2018

Giovanni Maria Maggioni: *Modelling and Data Analysis of Stochastic Nucleation in Crystallisation,* © 2018

DOI: 10.3929/ethz-a-

ACKNOWLEDGEMENTS

First, I would like to thank prof. Marco Mazzotti, who took me as one of his doctoral students and who asked me to work on the topic of this thesis. His decision has given me the opportunity to study new subjects, do research rather freely, and spend these interesting and nice years in Zurich. I also want to thank prof. Paolo Arosio and prof. Richard Braatz for accepting to be co-examiners of this work and for their valuable comments.

I want to also acknowledge the students and the colleagues who have collaborated with me over the years, Luca, Léo, Moritz, Maria, Matteo, Francesca, Elena, and Zoran, as well as the technicians in our lab, Markus and Daniel.

I would like to thank several people with whom I have spent many enjoyable moments, from coffee breaks to dinners and more, in particular the colleagues and friends from ETH (Stefano (Bennati), Thomas, Ashwin (among many other things, for a very interesting and good time in India) and Federico, Ian, Stefano (Zanco), Pietro, Davide, Giovanni, Aurélien), the Italian speaking friends (Vasco and Giulia, Matteo, Elena, Paolo and Jacopo), the "Swiss Friends" (Amina and Yassin, Alex and Nadia), as well as some very far away friends (Jazmin, Renato, Tatsu, and Yotam). Finally, in this group I want to specially mention Amir and Alberto, who are the people I have longest known in Zurich.

Some more people deserve a special acknowledgement.

First, I want to thank "Il Sig. Antonio" and his family, always kind and helpful since the moment I arrived in Zurich.

Stefano (Duca) must be thanked, among other things, for all the great coffee breaks and lunches, which always led to very interesting conversations and discussions. We have rather often very different ideas on the details (and, at times, on the essential things, as well),

but, precisely because of this, our chats have been so valuable and fun.

I also want to thank Fabio (Salvatori), with whom I have shared the office for four years and who has had to stand (fairly patiently) for quite a lot of wanderings into mathematics, particularly probability theory and stochastic processes. Moreover, the time in Zurich, not just the long office-hours, would have not been so enjoyable without his reasoned opinions, colourful comments, and sharp remarks.

A special appreciation goes to some old and very good friends, with whom I have shared many nice and enjoyed experiences and moments. Beppe (the "initial" connection), Albi and Edo (with whom I shared several much needed and relaxing holidays over the years and whom I am always happy to meet every time we can), and all the others from "Citterio-Re di Sasso".

Eleonora, a very good and old friend, and her husband Davide, himself an old and good friend now, must be thanked for all pleasant times (before moving to, during my staying, and, I hope, after moving from Zurich) and important experiences I have shared with them.

Together with old friends, I a special acknowledgement is due to two new ones: Leo and Poppy. From the first chat we have had on the way from Barcelona to Villemagne-l'Argentiére, his deep insight in so many (and so different) topics and his good spirit have made and make every conversation with him not only an interesting time, but also a great pleasure. Nevertheless, no conversation would have started if Poppy had not invited me to Villemagne-l'Argentière: I cannot, but thank her. In spite of having been only a short week-end, the timing could have not been more appropriate, the company more enjoyable, and our host a kinder, gentler, and nicer person.

I am also very grateful to Roberto (Mini), whom I have now known for a solid decade. Much has changed in our lives since the time we could have lunch breaks nearby the Scala or we could spend

our Fridays/Saturdays with the other friends “da Bianca”/” in Colonne”, ”sui Navigli”, and in several other places. Notwithstanding his share of heavy burdens, Robi has always been and still is a good, sensible person and a very dear friend, whom I know I can count upon.

Little more than one year ago, I had the luck of meeting Caleb, who has since become a very dear friend. I want to thank him for so many (and always interesting) conversations on several, diverse, assorted, and variegated topics. His sharp mind, insightful observations, open disposition, never trivial questions, and good nature have been bringing great fun, valued considerations, and cherished memories.

Giulia, literally a life-time and the oldest friend, must also be mentioned and thanked. Although we have taken rather different paths since we were children, and even though we can rarely meet in person, her good humour, solar character, kindness, and insight have always turned our reunions into very much appreciated and valued times.

To my family goes all my deepest gratefulness; specially to my grandparents, for their continuous affection and support.
In extremo optimum. To my parents: thank you, for everything and, most of all, simply for always being there.
Vorrei esprimere tutta la mia gratidutine alla mia famiglia, in particolare ai miei nonni, per il loro affetto ed il loro sostegno.
In extremo optimum. Ai miei genitori: grazie, di tutto e, più di tutto, semplicemente d’esserci sempre.

ABSTRACT

Nucleation is an important and challenging scientific problem in its own right, which can be investigated from different perspectives and looking at very different scales, from the level of molecular interactions to that of macroscopic, industrial units. In particular, crystal nucleation is a crucial step in several technologies constituting the backbone of important industries, e.g. pharmaceutical purifications or semi-conductor material production, as well as in several biological processes, e.g insulin metabolism or hemozoin crystallisation in malaria treatments or amyloid proteins aggregation. The models and the methods of data analysis developed in this thesis to rationalise and describe the random behaviour of crystallisation processes are based on the three foundations: the results of nucleation theories, the approach of population balance equations, and the mathematics of stochastic processes.

This work focuses first on the physical premises and on the mathematical construction of the models, elucidating the difference between nucleation time and detection time and discussing the mathematical and practical consequences of this difference.

The models are validated against the experimental data collected in several systems at different operating conditions. The statistical premises necessary to perform a meaningful analysis of the data and the numerical methods used to estimate the model parameters are then illustrated. To examine the quality of the agreement between the models and the data, the possible limitations due to the design of the experiments and to the detection techniques selected for the measurements are discussed.

Since a meaningful characterisation of stochastic processes requires a large amount of data of high quality, the third part of the work focuses on this problem. This part shows the experimental

protocol and the criteria adopted to guarantee well controlled and reproducible operating conditions, to which detection time experiments are known to be very sensitive. A careful analysis of the statistical assumptions allows to determine when one can combine different series of experiments carried out at nominally identical conditions. Furthermore, this analysis also yields a parameter-free method to compute the confidence intervals of nucleation rates estimated from empirical distributions. The method is developed assuming that nucleation follows an exponential distribution, but it can be generalised to other statistics.

Finally, given the importance of polymorphs in a lot of different applications, the competitive nucleation of several polymorphs is investigated, based on the crystallisation experiments of isonicotinamide from ethanol. The stochastic nucleation model previously developed is extended to account for the formation of an arbitrary number of polymorphs, highlighting the main advantages and limitations of the model. The behaviour of the system predicted by the model is then compared with the experiments, commenting on the similarities and differences between them.

The results of this thesis show that a deep understanding of nucleation, and therefore a correct description of this phenomenon, requires data of increasing quantity and quality as well as an accurate analysis of such data, which is based on the mathematical and statistical tools illustrated and developed in this work.

SOMMARIO

La nucleazione è un problema scientifico importante e complesso, il cui studio può essere affrontato da diverse angolazioni ed analizzato a differenti scale d'osservazione, dalle interazioni a livello molecolare sino al comportamento delle apparecchiature macroscopiche d'uso industriale. In particolare, la nucleazione di cristalli é un elemento essenziale in svariate tecnologie, che costituiscono il fondamento sia di importanti processi industriali, come la purificazione di prodotti farmaceutici o la produzione di materiali semi-conduttori, sia di numerosi processi biologici, dal metabolismo dell'insulina alla cristallizzazione dell'emozoina per i trattamenti antimalarici all'aggregazione delle proteine amiloidi. Tanto i modelli matematici quanto i metodi di analisi dei dati sperimentali sviluppati in questa tesi si prefiggono lo scopo di razionalizzare e descrivere il comportamento aleatorio dei processi di cristallizzazione, basandosi su tre elementi principali: i risultati ottenuti dalle teorie della nucleazione, la descrizione delle particelle solide tramite equazioni di bilancio di popolazione e la matematica dei processi stocastici.

Questo lavoro tratta, nella prima parte, della comprensione fisica della nucleazione e la conseguente costruzione matematica dei modelli, evidenziando la differenza tra tempo di nucleazione e tempo di rilevamento, per discutere quindi le implicazioni matematiche e sperimentali legate a tale differenza.

Nella seconda parte di questa tesi, i modelli sono convalidati usando i dati sperimentali ottenuti da diversi sistemi ed a diverse condizioni di processo. Di seguito, si illustrano i presupposti statistici necessari per effettuare una analisi affidabile e coerente dei dati, nonché i metodi numerici utilizzati per stimare i parametri dei modelli. Al fine di esaminare l'accordo tra i modelli proposti

ed i dati sperimentali, si sono discusse sia le possibili limitazioni legate alla progettazione degli esperimenti sia la tecnica scelta per effettuare il rilevamento dei cristalli.

Dato che una caratterizzazione appropriata dei processi stocastici richiede un considerevole numero di punti sperimentali di alta qualità, la terza parte di questo lavoro é dedicata a questo tema. In primo luogo, si discute il protocollo sperimentale ed i criteri adottati per assicurare di effettuare esperimenti in condizioni operative adeguatamente controllate e riproducibili. Una attenta analisi delle ipotesi statistiche consente di determinare quando sia possibile raggruppare serie di esperimenti, misurate indipendentemente, ma alle medesime condizioni nominali. Inoltre, è anche possibile dedurre da tale analisi un metodo, che non necessita di alcun parametro specifico del sistema, per calcolare gli intervalli di confidenza delle velocità di nucleazione stimate dalle distribuzioni empiriche.

Infine, questo lavoro tratta, sulla base di esperimenti di cristallizzazione di isonicotinamide in etanolo, la nucleazione competitiva dei polimorfi, considerando l'importanza di questa proprietà dei cristalli in svariate applicazioni. Un modello stocastico di nucleazione sviluppato in precedenza é generalizzato per considerare la formazione di un numero arbitrario di polimorfi, sottolineandone i principali vantaggi e le maggiori limitazioni. Il comportamento del sistema, predetto dal modello, é quindi confrontato con gli esperimenti, per valutarne le differenze e commentarne l'accuratezza.

I risultati di questa tesi mostrano che una profonda comprensione della nucleazione, pertanto una sua corretta descrizione, necessita non solo dati sperimentali di crescente quantità e qualità, ma anche una analisi accurata ed affidabile di tali dati, basata sui metodi sperimentali e sugli strumenti matematici discussi e sviluppati in questo lavoro.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Symbol	Meaning
μ	Chemical potential
a	Activity
k	Boltzmann constant
T	Temperature
β	Thermodynamic beta
S	Supersaturation
c	Concentration
c_0	Initial concentration
c_s	Solubility
p	System pressure
p_0	Initial pressure
p_e	System vapour pressure
v_l	Molecular volume, liquid phase
γ_a	Activity coefficient
γ	Surface energy
k_a	Surface shape factor
k_v	Volume shape factor
ΔG	Gibbs free energy
ΔF	Helmholtz free energy
J	Nucleation rate
A	Kinetic factor

Symbol	Meaning
B	Thermodynamic factor
G	Growth rate
W	System size
V	System volume
t_{N}	Nucleation time
t_{G}	Growth time
t_{D}	Detection time
t_{D}^{*}	Experimental detection time
Λ	Intensity measure of exponential process
λ	Intensity of exponential process
τ_{N}	Average nucleation time
σ_{N}^{2}	Nucleation time variance
P	Nucleation time probability
P^{*}	Nucleation time empirical probability
F^{*}	Detection time empirical probability
L	Crystal characteristic size
L_{c}	Nucleus size / Critical cluster size
f	Continuous crystal population
ϕ_{j}	j-th order moment, continuous crystal population
$\tilde{f}$	Discrete crystal population
$\tilde{\phi}_{\mathrm{j}}$	j-th order moment, discrete crystal population
ρ_{c}	Crystal density
D	Chebyschev Distance
α	Significance level
C	Confidence level
$\tilde{\sigma}$	Relative sensitivity

1

INTRODUCTION

Science is built up of facts, as a house is with stones. But a collection of facts is no more a science than a heap of stones is a house.

— Henri Poincaré

A theory with mathematical beauty is more likely to be correct than an ugly one that fits some experimental data.

— Paul Dirac

1.1 BACKGROUND

In thermodynamics, a "phase" is a portion of a system characterised by homogeneous properties. The process through which an existing parent phase evolves into a new phase is the so-called "phase transition" [1].

Being a spontaneous thermodynamic transformation, a phase transition brings the system from a state at a higher energy level to a state at a lower energy level. One distinguishes between two situations. If the parent phase is thermodynamically unstable, the phase transition is limited only by diffusion and occurs simultaneously across the system; a process with these characteristics is called "spinodal decomposition". If the parent phase is thermodynamically metastable, the phase transition occurs through the formation of nuclei of the new phase, i.e. nucleation, and their subsequent growth; such a process is called "binodal decomposition". In general, thermodynamic metastability entails that the system is initially at a local minimum of the energy and, through the phase transition, reaches a different local minimum at a lower energy level by overcoming an energy barrier, which separates the two minima.

Because of the necessity to overcome such an energy barrier, the transformations of metastable systems are said to be "activated".

The nuclei constitute the first step of a binodal decomposition and are localised, spatially disjoint, microscopic portions of the new phase. They appear within the parent phase due to the thermal fluctuations, which triggers the binodal decomposition.

The description of nucleation provided by Gibbs, and refined by many others later, has allowed not only the clear understanding of nucleation and of its essential thermodynamic features, but also the recognition that thermodynamic metastability gives nucleation an intrinsically dynamic and stochastic nature. The rich dynamics of binodal decomposition is due to the fact that, differently from the spinodal decomposition, binodal decomposition is not limited only by transport phenomena, but depends also on the occurrence of random fluctuations, which are themselves the reason of the process stochasticity [2]. Additionally, it is important to note that, by conjecturing that nucleation is intrinsically dynamic, it is not meant that nuclei cannot form at a constant rate, but rather that during nucleation and nuclei growth the system evolves over time.

The dynamics of nucleation was first investigated in a deterministic framework by Becker and Döring in 1935 [3], whose work still constitutes the backbone of the modern kinetic theory of nucleation [4]. Interestingly, Eyring, Evans, and Polany developed, also in 1935, the transition state theory for explaining the kinetics of chemical reactions, using arguments akin to those of Becker and Döring. Indeed, since nucleation and chemical reactions are phenomena occurring in metastable systems, both processes can be rationalised and understood using a very similar physical framework and, therefore, a similar mathematical treatment. Effectively, nucleation kinetics have also been described in terms of chemical reactivity, with the addition of each new unit to the sub-critical cluster interpreted as a reaction, much alike a reversible polymerisation. The development of nucleation dynamics ultimately determined the functional dependency of the nucleation rate on the

system state variables (e.g., temperature and composition) as well as on the system-specific parameters (e.g., the interface energy and the crystal molecular volume). The results obtained by the thermodynamic and dynamic theories of nucleation, in many ways so insightful and fundamental, are collectively known as "classical nucleation theory".

While the dynamic features of nucleation have been extensively investigated beyond the work of Becker and Döring, the stochastic nature of nucleation remained initially unexplored, if not neglected, in most engineering applications. Possibly, such an approach was influenced by the experimental behaviour seen with chemical reactions. In fact, the stochasticity characterising the occurrence of an individual reaction at the molecular level is typically not perceived at the macroscopic scale of observations. Since a reaction requires only one or two molecules to interact, and usually in the system there is a gargantuan number of molecules, so many interactions occur simultaneously that, to a macroscopic observer, the system samples at any time almost the entire space of possible events. In other words, assuming the system ergodic, in the vast majority of experiments, the behaviour of all molecules together at the macroscopic level may be consider a fair approximation of the process expectation. Because of the analogy between nucleation and chemical reactions, one may think that a similar argument could hold true also for nucleation. However, nucleation is actually much more complex, since tens, at times even hundreds, of molecules interacting in the proper way are needed to form one nucleus. Therefore, the formation of a nucleus is, by itself, a much rarer (thus, more stochastic) event than a uni- or bi-molecular reaction and the system, again assuming it ergodic, needs a longer time, comparable to that characteristic of macroscopic observation, to sample the space of possible events.

Few, but important studies explicitly modelling the stochastic character of nucleation were conducted in the 1970s, by Kaischew, Toschev, and Kashchiev. In these seminal works, they explicitly

stated that "nucleation is a random phenomenon" [5], and they used stochastic models to describe the experiments. However, their observations and results seemed to have remained largely overlooked, and the poor reproducibility of nucleation experiments, often reported, was generally interpreted as the consequence either of poor mixing conditions or of a high sensitivity to uncontrolled (and almost uncontrollable) factors, such as impurities and surface roughness of the reactors in which nucleation occurred [6].

In the 1990s, the development of new experimental techniques and the resultant observations yielded (and is still yielding) renovated the interest in modelling nucleation as a stochastic phenomenon beyond the molecular level, which physicists have been long investigating [7]. These new techniques, e.g. microfluidics, allow researchers to investigate ever smaller and spatially homogeneous systems, where mass and heat transfer limitations practically do not occur. Furthermore, they perform at one time multiple replicas of the same experiment, i.e. of the same system and at the same conditions. Since the introduction of these new devices, the measured data have been providing abundant evidence that the times at which crystallisation occurs are distributed: indeed, the smaller the investigated system is, the more broadly the measured crystallisation times are distributed. As a consequence, the description of nucleation and the understanding of its dynamics was refined to account for stochasticity, which needed to be modelled within a more appropriate mathematical framework, namely that of stochastic processes, different from the deterministic one adopted to develop the classical approach.

The theory of stochastic processes, within the broader context of probability theory, allows for mathematically clear and tractable modelling, albeit at times not intuitively, of the fundamental features and of the dynamics of a collection of random variables, such as that represented by nucleation times. The theory of probability, with the first descriptions of stochastic processes, dates back to the seventeenth century, with mathematicians such as Fermat, Euler,

Pascal, and Jacob Bernoulli, and this theory has been fundamental in the development of statistical physics by Maxwell, Gibbs, and Boltzmann in the nineteenth century.

However, the modern mathematical theory of stochastic processes was developed only much later, in the 1930s, with the seminal and pioneering works of Kolmogorov, Lévy, Wiener, and later Itô, who initiated the development of stochastic calculus.

Grounded in the theory of measure and in the axiomatic approach of Kolmogorov, the modern theory of stochastic processes extends and generalises most of the usual mathematical objects necessary for modelling (such as the idea of function, continuity, integral, derivative, differential equations, strong and weak solutions,...) and allows one to perform on them, at a formal level, the majority of the standard mathematical manipulations used for the objects' deterministic counterparts. Based on these results, one can mathematically formalise and construct stochastic models used in physics and engineering, as well as social sciences and economics, through which one not only analyses, rationalises, and interprets the empirical observations, but also simulates and studies in detail systems and phenomena of increasing complexity, such as nucleation.

1.2 AIM OF THIS THESIS

Nucleation is clearly a challenging, important, and intriguing scientific problem in its own right, which can be investigated from different perspectives and approaches and looking at very different scales, from the scale of molecular interactions to that of macroscopic, industrial units. However, nucleation, particularly crystal nucleation, is a crucial step not only in several technologies constituting the backbone of important industries, e.g. pharmaceutical purifications or semi-conductor material production, but also in several biological processes, e.g insulin metabolism or hemozoin crystallisation in malaria treatments or amyloid proteins aggregation.

This thesis, focused particularly on engineering and process applications, is a contribution to improve the physical understanding and the description of nucleation, by developing models based on integro-differential population balance equations. Coherently with the engineering scope of the thesis, the behaviour of the units forming a crystal at the molecular level, i.e. how a nucleus forms and grows, is only marginally discussed. Most of the work presented and discussed here aims at developing crystallisation models about when, and with which statistics, the nuclei appear in the system and then grow, also considering the possibility that multiple polymorphs form stochastically in the system. This thesis builds on the results of nucleation theories, on the approach of population balance equations, and on the mathematical theory of stochastic processes for the development of crystallisation models, which account for the random behaviour of nucleation. These models can be used, after being validated against the data, to analyse experiments and to improve the design of new tests and, ultimately, the design and control of crystallisation processes themselves.

The thesis is divided into several, but interconnected chapters, each dealing with different aspects and problems regarding modelling, and analysing the data measured from nucleation experiments.

Chapter 2 discusses nucleation models. First, the classical nucleation theory and the two-step nucleation theory are revised and discussed, highlighting the theories' main results. Then, the probabilistic framework of nucleation is drafted, describing the two stochastic processes adopted to describe the formation of the nuclei. We show how to test some of the fundamental hypotheses of classical nucleation theory by simulating Argon condensation with enhanced molecular dynamics, pointing out the potential and the limitations of such methods to investigate more complex systems. Finally, by applying a novel methodology to analyse the simulations results, we estimate the condensation rate of Argon at conditions comparable to the experimental ones. Argon condensation is

the only case in this work where a type of nucleation, other than crystal nucleation from solution, is considered.

Since in actual experiments one detects the presence of crystals, rather than the formation of nuclei, Chapter 3 deals with crystallisation models, i.e. nucleation and the other phenomena leading to the formation of a new detectable phase, such as growth. After elucidating the difference between nucleation time and detection time, and its mathematical and practical consequences, two different models connecting the formation of the nuclei with the detection of crystals are discussed.

The models are then validated in Chapter 4 against the experimental data, i.e. the measured detection times, collected in several systems and at different operating conditions. To this aim, the statistical assumptions necessary to extract nucleation kinetics from the data are investigated and the problem of collecting a sufficient number of experiments to reconstruct a reliable estimate of the statistic underlying nucleation is discussed from a mathematical perspective. Then, the numerical methods used to estimate the model parameters are discussed. The quality of the agreement between the models and the data is then examined, considering also the possible limitations related to the design of the experiments and to the detection techniques selected for performing the measurements.

Chapter 5 focuses on how to gather and analyse experimental points in large quantity as well as of high quality, through the process of repeating series of experiments performed at nominally identical conditions. First, we illustrate the protocol and the criteria adopted to guarantee well controlled and reproducible operating conditions, to which detection time experiments are known to be very sensitive. Second, we consider the quantitative statistical criteria, which allow one to determine when different series of experiments carried out at nominally identical conditions can be combined and considered together for the estimation of nucleation rates. Then, we develop a method to compute the confidence in-

tervals of nucleation rates estimated from empirical distributions, assuming that nucleation is a Poisson process.

In light of the fact that several organic and inorganic compounds of interest crystallise in different polymorphic forms, Chapter 6 considers the problem of modelling the competitive nucleation of polymorphs in the stochastic framework. Inspired by the analysis of detection time experiments performed with isonicotinamide in ethanol, we extend the results of the first nucleation model discussed in Chapter 2 to the nucleation of an arbitrary number of polymorphs, and we compare the behaviour predicted by such a model with the experimental data.

Finally, we summarise in Chapter 7 the results and conclusions of the thesis and reflect on the possible directions along which the work performed here may be continued and extended.

2

THEORIES OF NUCLEATION

2.1 INTRODUCTION

This Chapter is organised in three sections. The first one reviews the two most important theories of crystal nucleation, presenting and discussing their differences, main results, and shortcomings (Section 2.2).

In the second section, the idea of nucleation as a stochastic process is introduced, and the statistic which is associated to nucleation is illustrated. Based on this statistic, we construct and study two alternative stochastic processes, each of them related to a different model for the formation of primary nuclei (Section 2.3).

Finally, we show and discuss the results obtained simulating the nucleation of liquid argon from its vapour using enhanced molecular dynamics. This relatively simple system allows to illustrate the validity of the basic assumptions and results of classical nucleation theory, to understand when and why such theory may fail to describe more complex systems, and, more interestingly, to emphasise some of the main challenges, achievement, and potential benefits of molecular modelling.

2.2 THERMODYNAMICS AND KINETICS OF NUCLEATION*

Thermodynamically, a phase transition can be envisaged as a process in which the system moves from the state with higher energy (the parent phase) to the state with lower energy (the new phase):

*The material presented in this Section has been reported in: M. Mazzotti, T. Vetter, D. R. Ochsenbein, G. M. Maggioni, C. Lindenberg in *Polymorph. Pharm. Ind. Solid Form Drug Dev.* Springer-Verlag, **2018**.

the energy difference between the two states is the driving force of the phase transition, hence of nucleation and of the subsequent growth of nuclei [4]. First, we describe briefly how such a driving force can be expressed in terms of state variable, then we discuss nucleation theories.

In the following, we consider the phase transition of a system at constant temperature and pressure, thus describing most naturally the process in the Gibbs ensemble with the Gibbs free energy. The same description can also be developed equally validly in different ensembles, such as the canonical ensemble, where the Helmholtz free energy is used. Actually, one can show that for the nucleation of liquid droplets from a vapour the Helmholtz free energy and the Gibbs free energy have the same value [9]. The driving force of the phase transition in the Gibbs ensemble can be expressed as the difference of chemical potentials, $\Delta\mu$, between the system at its current state, μ_ℓ, and the system at its equilibrium state, μ_ℓ^*. $\Delta\mu$ is the logarithm of the ratio of the activities associated the two states, a_ℓ and a_ℓ^*, respectively, where the activity is given by the ratio between the fugacity computed at the system temperature, pressure, and composition and the fugacity computed at some arbitrary reference state [10]. Thus, the driving force can be written as:

$$\Delta\mu = \mu_\ell - \mu_\ell^* = kT \ln\left(\frac{a_\ell}{a_\ell^*}\right) = kT \ln S \tag{2.1}$$

where k is the Boltzmann constant and T the temperature and S is the supersaturation. Considering the case of crystal nucleation from a clear solution, $\Delta\mu$ would be the difference between the chemical potential of the solute in the solution at its current composition (since temperature and pressure are constant) and at its equilibrium composition, i.e. with crystals in equilibrium with the solution. For a system crystallising, $a_l = x\gamma_a$, where x is the molar fraction of the solute (or, more often the concentration of solute) and γ_a its activity coefficient, and $a_l^* = x^*\gamma_a^*$. Quite often, one assumes that $\gamma_a/\gamma_a^* \approx 1$, so that $S = x/x^*$. Eq. (2.1) can be easily

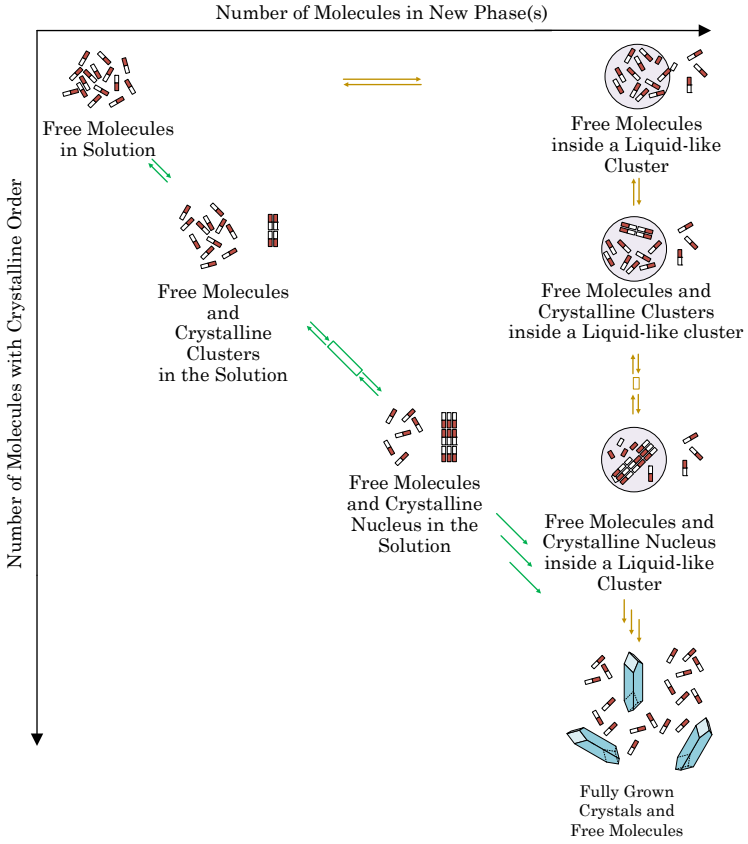


Figure 2.1: Conceptual picture of the mechanisms of nucleation according to CNT (green arrows) and to 2-SNT (yellow arrows), starting from a clear, supersaturated solution. According to both theories, clusters form/decompose through attachment/detachment of building units and, upon reaching a critical cluster size, i.e. becoming a nucleus, the attachment of further building units is energetically favoured, ultimately leading to the formation of a macroscopic crystal. In the CNT, each cluster exhibits the final crystal structure as soon as building units attach, also prior to reaching the critical size. On the contrary, in the 2-SNT, first a disordered, liquid-like cluster forms, which then evolves into a structured cluster in a second step.

adapted to any type of nucleation by considering the activities of the specific phase considered: for liquid condensation from its vapour, $S = p/p^*$, with p being the actual pressure of the vapour and p^* its vapour pressure.

Let us now consider nucleation theories, which aim at deriving expressions of the nucleation rate by describing both the thermodynamic and kinetic aspects of nucleation. An overview of the two most important theoretical frameworks and the related applications shall be given in the following, illustrating the complexity of the phenomenon and the challenges its detailed description entails. The first one is the oldest, and probably best known, theory of nucleation: the so-called classical nucleation theory (CNT). Originally developed for the nucleation of droplets and bubbles, the CNT has also been (and still is) extensively applied to crystals, which, in stark contrast to droplets/bubbles, exhibit a strong supramolecular structure and long-range order. The second theory, initially developed to describe the formation of crystals from solution, will be discussed more qualitatively. It can be interpreted as a refinement and a specialisation of the CNT, since it recognises the difference (neglected by CNT) between short-range ordered phases, such as liquids, and long-range ordered phases, such as crystals. Often referred to as two-step nucleation theory (2-SNT) in the recent literature [11–14], it has been developed mainly during the last three decades, fuelled by considerable progress in measurement devices, which allow for the observation of ever smaller entities in solution [15–17]. A conceptual representation of the hypothesised pathways, of the fundamental differences, and of the similarities in the two theories is provided in Figure 2.1.

2.2.1 *Classical Nucleation Theory*

The thermodynamics of classical nucleation theory has been developed in the work of as early pioneers as Gibbs, while its kinetic foundations dates back to the work of Becker and Döring, Volmer,

and many others [2, 3, 18–21]. The interested reader can find the full derivation, historical details, and various technical aspects of CNT in the classical works of Kashchiev and Debenedetti [4, 22]. In order to describe the thermodynamics of CNT, we derive the relation describing the energy difference between a solution and a solution containing a cluster of n building units (typically, atoms or molecules).

The total change in the system Gibbs free energy to form a cluster of n molecules, ΔG_n, which also corresponds to the work performed by the system, is given by the sum of two terms: a reduction of the energy due to formation of a crystalline bulk phase and an increment of energy caused by the formation of an interface between the new and the parent phase. We may then write:

$$\Delta G_n = -n\Delta\mu + \gamma_n \mathcal{A}_n \tag{2.2}$$

where the crystallisation driving force represents the bulk contribution, $\Delta\mu$ (2.1), and the second term the surface contribution, expressed through a surface energy γ_n and the number of molecules forming the area of the cluster, $\mathcal{A}_n$. Note that such an expression is derived by assuming that a cluster of n molecules can be characterised in terms of a well-defined surface and that it behaves as a whole similarly to a macroscopic crystal. This assumption, known as capillary approximation, has been at the origin of several controversies and it is often formulated in this terms: a crystalline cluster, and thus a nucleus, possesses the same physicochemical properties of a fully-grown, macroscopic crystal, with a similar surface energy, crystalline structure, density, and all other related properties. If the cluster geometry is well described by one characteristic length L, one can derive from Eq. (2.2), function of the discrete variable n, an equivalent formulation function of the continuous variable L. The volume and the area of the cluster, functions of L, are expressed as $\mathcal{V} = k_v L^3$ and $\mathcal{A} = k_a L^2$, respectively, where k_v and k_s are the volume and the surface shape factors (for a cube of side L, $k_v = 1$ and $k_a = 6$; for a sphere of diameter L, $k_v = \pi/6$ and

$k_a = \pi$). Thus, the equivalence of volume is attained by equating the volume of the cluster, $\mathcal{V}$, in the two descriptions:

$$n v_c = \mathcal{V} = k_v L^3 \tag{2.3}$$

$$\gamma_n \mathcal{A}_n = \gamma \mathcal{A} = \gamma k_a L^2 \tag{2.4}$$

where v_c is the molecular volume. The Gibbs free energy change for a cluster of size L, at given temperature T and supersaturation S, can then be expressed as:

$$\Delta G(L) = -\frac{k_v L^3}{v_c} kT \ln S + \gamma k_a L^2 \tag{2.5}$$

where v_c is the molecular volume. The behaviour of ΔG given by Equation (2.5) is shown in Figure 2.2 for several supersaturations: ΔG is zero for $L = 0$, goes through a maximum, and then decreases monotonically. The maximum on each curve in Figure 2.2 represents a point of unstable equilibrium: a cluster with such a size, L_c, is said to be "critical" and corresponds to a nucleus. L_c may be computed from 2.5 by setting $\mathrm{d}\Delta G/\mathrm{d}L = 0$, thus yielding:

$$L_c = \frac{2k_a}{3k_v} \frac{\gamma v_c}{kT \ln S} \tag{2.6}$$

The value of the corresponding free energy change is:

$$\Delta G^* = \left(\frac{4k_a^3}{27 k_v^2} \right) \frac{\gamma^3 v_c^2}{k^2 T^2 \ln^2 S} \tag{2.7}$$

where the term in brackets in Eq. (2.7) is a constant dependent on k_v and k_s, i.e. the geometry of nucleus; for a sphere, it is $16\pi/3$, while for a cube, 32. Eqs. (2.7) and (2.6) indicate that higher supersaturations lead to lower energy barriers for nucleation and smaller nuclei sizes, as one could see already in Fig. 2.2.

The cluster size L can be interpreted also as the coordinate describing the evolution of the system energy, ΔG, during the formation and evolution of a cluster, in analogy to the idea of "reaction

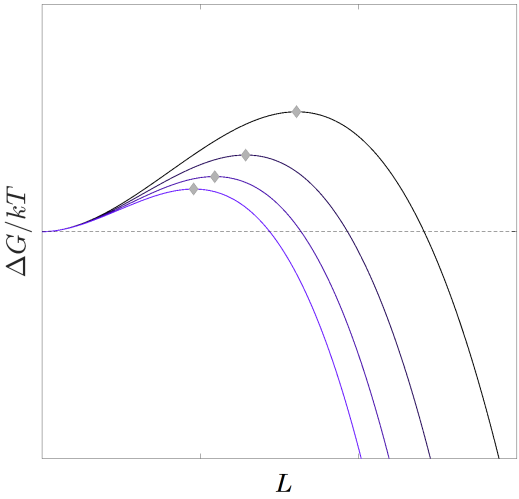


Figure 2.2: Free energy, $\Delta G/kT$, required according to classical nucleation theory, as a function of the reaction coordinate, which here corresponds to the cluster size L. Each curve is drawn at constant values of supersaturation and temperature. The grey symbol on each curve denotes the point associated to the maximum of the Gibbs' free energy and its corresponding critical size L_c.

coordinate" used for the description of chemical reactions. Within this interpretation, the nucleus is the analogous of a transition state, i.e. a (hypothetical) state which is in equilibrium with the reactants, represented here by the building units of the crystal. Similarly, ΔG_{c} represents the activation barrier of nucleation: since the nucleus correspond to the maximum of the free energy difference, it can further grow into a crystal or dissolve back into the solution with the same energy gain, as much as a transition state can evolve into the products of involve into the reactants. Following a reasoning analogous to that of classical transition state theory,

the frequency of clusters passing through the transition state is proportional to the exponential of the height of the free energy barrier that needs crossing [3, 23–26]. Therefore, the nucleation rate J, i.e. the number of nuclei formed in a solution per unit time and unit size of the system (e.g., the system volume) is:

$$J = J_0 \exp\left(-\frac{\Delta G^*}{kT}\right) \tag{2.8}$$

The pre-exponential factor J_0 is typically interpreted as a product of the number of nucleation sites and of the net attachment frequency of building units to the cluster. For the formation of a condensed phase, $J_0 = A(T)S$, thus:

$$J = A(T)S \exp\left(-\frac{\Delta G^*}{kT}\right) = A(T)S \exp\left(-\frac{B(T)}{\ln^2 S}\right) \tag{2.9}$$

where $A(T)$ is often described with an Arrhenius dependence on the temperature:

$$A = A_0 \exp\left(-\frac{A_1}{T}\right) \tag{2.10}$$

and where B is obtained by using Eq. (2.7) as:

$$B = \left(\frac{4k_a^3}{27k_v^2}\right)\frac{\gamma^3 v_c^2}{k^3 T^3} \tag{2.11}$$

Within the literature discussing the CNT, several other cases have been considered when deriving the pre-exponential factor J_0 [4, 27], depending on the phase which is nucleating, and have slightly different functional dependencies on the temperature and concentration.

It is important to note that Eq. (2.9) has been derived for the case of homogeneous nucleation, i.e. a system where no contaminants and impurities are present. In reality, the presence of small (sub-micron- or micro-sized) particles or other impurities in solutions that act as further nucleation sites is experimentally unavoidable. As a consequence, the actual number or nucleation sites is hard

to predict. This issue is believed to be one (if not the) most important cause of discrepancies between experimentally measured data and theoretical predictions. Typically, the predicted nucleation rate are several orders of magnitude smaller different from the ones estimated from the experiment, even though other explanations exist as well (see 2-SNT below). While the presence of "dust" particles and other heterogeneous surfaces (stirrers, crystalliser walls) likely affects the kinetic parameter A more, it is known to also modify the thermodynamic parameter B in Eq. (2.9), which is then usually written as:

$$B = \left(\frac{4k_a^3}{27k_v^2}\right)\frac{\gamma_e^3 v_c^2}{k^3 T^3} \tag{2.12}$$

where $\gamma_e = \chi_e \gamma$ represents the "effective" surface energy, with $0 < \chi_e \leq 1$ an efficiency factor (the smaller, the more effectively impurities induce nucleation). Note that some modifications of CNT require only $0 < \chi_e$, i.e. they allow for impurities to hinder nucleation and make it slower than the homogeneous case. As a consequence, experimentalist and practitioners often gather experimental data on which A and B are fitted as parameters.

In spite of its limitations and simplifications, classical nucleation theory has delivered not only a functional form of the nucleation rate, which proved to be of great value for process design, but also important mechanistic insights. For instance, the observation of a metastable zone, i.e. a zone in the phase diagram where primary nucleation is not observed within a specified time-frame, is qualitatively consistent with the nucleation rate law outlined in Eq. (2.9), even though the metastable zone is likely a kinetic phenomenon. Indeed, the nucleation rate increases rapidly once a certain threshold supersaturation is reached, whereas before this threshold substantial time can elapse before nucleation is observed.

2.2.2 *Two-Step Nucleation Theory*

Differently from CNT, according to 2-SNT the system evolves from an initial state (the clear solution) to the final state (the crystals and the solution) passing through an intermediate metastable state, as shown in Figure 2.1. Such a difference between the reaction path of CNT and that of 2-SNT has also important consequences on the energetics of the two processes, as one can readily see in the qualitative representation of the Gibbs free energy in Figure 2.3. This figure shows the evolution of ΔG along a reaction coordinate, η, which incorporates the information about both the cluster's size and its structural properties. From the initial state ($\Delta G = 0$), i.e. the clear solution, the system must overcome two energy barriers of heights ΔG_1^* and ΔG_2^*, respectively; after the second barrier, it reaches the final state (with a crystal). After the first energy barrier, a metastable state is located, which represents the liquid-like cluster and whose energy level is not necessarily higher than that of the solution, as illustrated in 2.3; in fact, it could also be lower, indicating simply that the liquid-like cluster is more stable than the initial state.

The two-step nucleation theory was originally developed to describe the crystallisation of proteins in aqueous solutions, such as lyzozyme in water [28], for which CNT did not agree well with experimental observations. Specifically, for these systems liquid-like clusters were observed in supersaturated solutions before the formation of crystals. The qualitative picture of the Gibbs free energy is different from the one obtained from CNT, which suggests that the mathematical form of the nucleation rate J may be different as well. For crystal nucleation of proteins in aqueous solutions, a phenomenological model has been derived under system specific assumption, and the corresponding nucleation rate has been estimated [28, 29]. Just as for CNT, also for this phenomenological model accurate ab initio calculations of the energy barriers are difficult, hence the authors took them as fitting parameters [29].

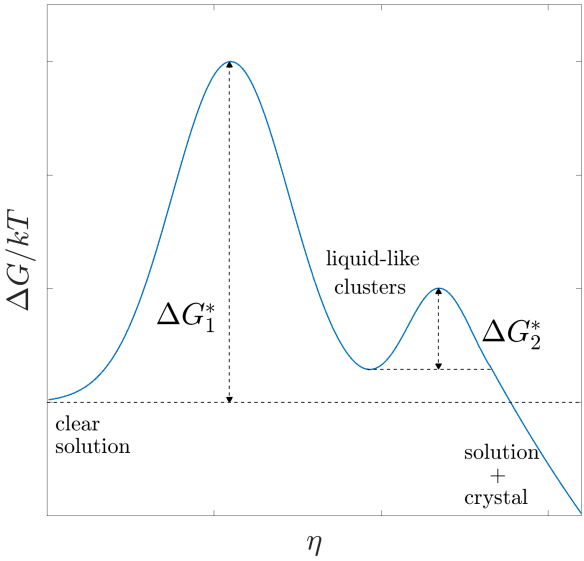


Figure 2.3: A qualitative representation of the energy profile, $\Delta G/kT$, of a cluster as a function of a nucleation coordinate, η, in 2-SNT. The nucleation coordinate encompasses both the cluster size and its degree of crystallinity.

Although the theory was initially conceived to describe nucleation of proteins, i.e. very large macromolecules, in solution [14, 28, 30–32], recent evidence suggests that 2-step nucleation could also describe the crystallisation pathways of amino-acids, small organic molecules, and even inorganic compounds, some of which exhibit features of a 2-step process [33–35]. On the one hand, this might be related to the formal generality of 2-SNT; on the other hand, amino-acids are the building units of proteins, hence it is not surprising that they may share a similar behaviour. As a final remark

about this theory, one should note that the nature and the properties (e.g. density, concentration, structure, etc.) of the intermediate species is still debated. In fact, a consistent, general formulation of the 2-SNT, applicable to all systems, is also still matter of debate.

2.3 STOCHASTIC FORMATION OF PRIMARY NUCLEI*

The theories of nucleation outlined above, although extremely insightful, cannot account by themselves for the poor reproducibility observed in nucleation experiments. In fact, researchers have reported in the literature investigations of several systems, of organic [37–42] as well as of inorganic [5, 43–45] substances, where the variance of measured detection times about the associated mean value is rather large. Two interpretations have been developed to explain such poor reproducibility.

The first one maintains that nucleation, growth, and the other relevant phenomena are deterministic processes, at least when the system must no longer be described looking at the individual molecular interactions of its components. In this case, the macroscopic variability of nucleation experiments is attributed to a high sensitivity to experimental conditions, which are difficult to reproduce exactly in repeated experiments, not dissimilarly from the behaviour of a fluid described by the Navier-Stokes equations. Thus, the equations reported above, either of CNT or of 2-SNT, should still suffice to describe nucleation. Even though it is known that nucleation is affected by mixing and by heat and mass transfer limitations, whose characteristics are not trivial to control, particularly when scaling up, the poor reproducibility observed in typical nucleation experiments has been difficult to understand, rationalise, and control. This difficulty is well known to experimentalists and is truly an issue to be taken in serious consideration. However, it has been

*The material presented in this Section has been reported in: G. M. Maggioni, M. Mazzotti, "Modelling the stochastic behaviour of primary nucleation", *Faraday Discuss.* **2015**, *179*, 359–382 .

observed that the distribution of detection times, from which nucleation properties are inferred, broadens when operating in smaller and smaller systems, where the experimental variables can be finely tuned and the limitations due to transport phenomena are negligible.

The second interpretation, on the contrary, relies on considering the formation of critical nuclei as an intrinsic stochastic phenomenon directly responsible of the variability observed in the outcome of nucleation experiments. An inherent stochasticity of nucleation is consistent with its nature of activated phenomenon and its grounded in statistical mechanics, where random thermal fluctuations are responsible of triggering the formation of a nucleus. The present work develops from and around this second interpretation.

There is currently not only theoretical consensus, but also abundant experimental evidence that nucleation should be indeed treated as stochastic process: we have reported few references in Table 3.1 in Section 3.2. While it is generally accepted that nucleation is stochastic, no final consensus has been reached yet concerning the statistic(s) describing the data. Various distributions and different types of stochastic processes have been used in the literature, the ones most often encountered being the exponential distribution (also improperly called "Poisson distribution", since it can be derived from a Poisson process) [36, 46–48], the Weibull distribution [49, 50], and the Log-Normal distribution [51–53]. A comprehensive review of the alternatives used to describe experiments at constant supersaturation and temperature has been written by Sear [49].

Regardless of the specific statistic chosen for nucleation, a stochastic process can be described according to two different, rather general frameworks. The first one, which has been adopted in most studies and includes also the Poisson process as a special case, considers nucleation as a Markov process, i.e. a process where the probability of future events depends only on the present state of the system and not on its history. The first model presented in

this work (Nucleation Mode 1) develops on this assumption and is discussed in more detail in Section2.3.2.

On the contrary, the second framework assumes history dependence (also termed sometimes "hysteresis"), which has been reportedly observed [54, 55] in some systems, and which can also be hypothesised because of several theoretical considerations, e.g. by considering that the presence of previously formed nuclei and crystals does deplete supersaturation. The case of history dependence is the basis of the second model (Nucleation Mode 2) of this work, presented in Section2.3.3.

2.3.1 *Exponential Stochastic Processes*

In this work, we have adopted the exponential distribution to describe the formation of a single nucleus, based on the following motives. Nucleation is a rare event, resulting from random fluctuations (see Chapter 2) and representing the evolution of the system from an energy minimum to another energy minimum. The exponential distribution is known to be the statistic describing the time between two rare events, for Markov and non-Markov processes alike, the so-called "law of rare events" [25, 46, 56, 57]. According to the well-known Boltzmann theory, it is also the statistic describing the probability that specific configurations are observed in a system at equilibrium. The most general cumulative distribution function P of an exponential distribution is defined as:

$$P(t) = \begin{cases} 1 - \exp\left(-\left(\Lambda(t)\right)\right) & \forall\ t \geq t_0 \\ 0 & \forall\ t < t_0 \end{cases} \tag{2.13}$$

where the so-called intensity measure $\Lambda(t)$ can be written as:

$$\Lambda(t) = \int_{t_0}^{t} \lambda(x)dx \qquad \forall\ t \geq t_0 \tag{2.14}$$

where λ is a time-dependent function. A less general, but often encountered, case is obtained when λ is constant and P reads:

$$P(t) = \begin{cases} 1 - \exp\left(-\left((t - t_0)\lambda\right)\right) & \forall\ t \geq t_0 \\ 0 & \forall\ t < t_0 \end{cases} \tag{2.15}$$

and

$$\Lambda(t) = (t - t_0)\lambda \qquad \forall\ t \geq t_0 \tag{2.16}$$

In order to use the nomenclature most commonly found in crystallisation, we will refer in this work to the exponential distribution as "Poisson distribution", using the nomenclature "Poisson discrete distribution" when referring to the actual Poisson distribution. From Eq. (2.13), one can also compute the average time of nucleation, τ_N, and its variance, σ_N^2:

$$\tau_N = \int_0^{+\infty} t\lambda(t) \exp\left(-\int_0^t \lambda(x)dx\right) dt \tag{2.17}$$

$$\sigma_N^2 = \int_0^{+\infty} (t - \tau_N)^2 \lambda(t) \exp\left(-\int_0^t \lambda(x)dx\right) dt \tag{2.18}$$

The intensity λ is dimensionally a frequency and the (primary) nucleation rate has dimensions of [number of nuclei/unit time/unit size]. Based on a trivial dimensional analysis, one sees that product of the primary nucleation rate, J, and of the size of the system, W, is thus also a frequency. For this reason, the intensity is defined as:

$$\lambda = JW \tag{2.19}$$

This is the typical definition adopted in the literature [46, 58, 59], where either the volume of the solution, V, or the mass of solvent, M, are typically chosen to represent W.

2.3.2 *Nucleation Mode 1: Single Stochastic Event*

Nucleation Mode 1 is the simplest possible stochastic model and is based on the following assumptions:

(M1-H.1) The first nucleation event is stochastic. It follows the statistic of Eq. (2.13), with intensity defined by Eqs. (2.14) and (2.19). No crystal is present in the system prior to the formation of the first nucleus;

(M1-H.2) The formation of the first nucleus triggers the start of a deterministic process, through which new nuclei are formed, while the existing ones grow.

(M1-H.3) perfect mixing, no secondary nucleation, no agglomeration, and no breakage are assumed.

Nucleation Mode 1 bears obvious similarities to other models proposed in the literature, such as that of Jiang et al. [47, 59].

The process described by Nucleation Mode 1 goes as follows. The cooling crystallisation process starts at $t = t_0$, with the temperature following a given cooling profile and the solute concentration, c, constant at its initial value c_0; no crystals are present. Note that the initial time t_0 in Eq. (2.15) is often set equal to zero; using also Eq. (2.19), the cumulative distribution P defining the statistic of nucleation is:

$$P(t) = 1 - \exp\left(-\int_0^t W(x)J(x)dx\right) \tag{2.20}$$

Note that the intensity λ can account also for a changing W, hence, Nucleation Mode 1 is able to describe not only cooling, but also evaporative and anti-solvent crystallisation, even though they are not considered here. At time t_N, the first nucleus appears, hence one crystal of negligible size is present. For any time $t \geq t_N$, the system evolves in a deterministic manner, discussed below in Section 3.3, whose initial conditions are assigned at $t = t_N$. Once the first nucleus has formed, one simulates the growth of the first nucleus and the deterministic formation and growth of the other nuclei. The growth process and the detection conditions are discussed in Section 3.3.

Note also that, according to Nucleation Mode 1, the average nucleation time τ_N and its variance σ_N^2 are given by Eqs. (2.17)-(2.18), showing that not only the average nucleation time decreases with W, but also the variance - hence, the expected variability in measurements of nucleation times. From this interpretation a seemingly reassuring picture seems to emerge: the stochastic nature of nucleation fades away as W increases and the variability in measured nucleation times vanishes accordingly. This behaviour (stochasticity in small systems/slowly nucleating substances and determinism in large systems/fast nucleating substances) is consistent with the mono-nuclear and the poly-nuclear mechanisms developed and used by Kashchiev, Verdoes, and van Rosmalen to describe nucleation and to derive expressions for the experimental induction time in their seminal paper, wherein the appearance of the first nucleus triggers a change of regime, i.e. a transition from a metastable state to a deterministic evolution [60].

Unfortunately, there are several pieces of evidence that challenge the picture of nucleation above. First, it is rather clear that Nucleation Mode 1, assuming only one stochastic event, might be realistic in a very small volume of solution, but looses credibility when the system size increases. It is in fact counter-intuitive that the formation of a critical nucleus, much smaller than the system size, in a specific position in the solution volume impacts immediately the course of nucleation in the whole volume and renders the formation of further nuclei fully deterministic. Therefore, the sharp separation between a stochastic nucleation mechanism for the first crystal and a deterministic growth mechanism (as assumed here and in the model of Jiang et al. [59]) seems to be non-physical. Second, experimental evidence indicates that also in large vessels, i.e. where the standard deviation of nucleation times should be so small to be effectively negligible, nucleation is poorly reproducible [41, 47, 61, 62]. A model proposed by Goh et al. [46] assumes that the formation of several primary nuclei is stochastic and describes nucleation as a Poisson process. However, this model considers all nucleation

events not dependent on the history of the system: this is realistic for the formation of the few first nuclei, but not for that of the subsequent ones, when crystal growth contributes to the depletion of supersaturation, particularly in relatively small systems. These observations motivated the development of an alternative, more general description, where stochasticity and determinism coexist at least during the early stages of the crystallisation process.

2.3.3 *Nucleation Mode 2: Chain of Stochastic Events*

Nucleation Mode 2 is a generalisation of Nucleation Mode 1, but also of the model of Goh et al. [46]. It is based on the following assumptions:

(M2-H.1) All nuclei, not only the first one as in Nucleation Mode 1, are formed stochastically. No crystal is present in the system prior to the formation of the first nucleus;

(M2-H.2) the statistics of time interval elapsed between the formation of any two consecutive nucleation events is exponentially distributed and follows Eq. (2.13), with the intensity given by Eqs. (2.14) and (2.19);

(M2-H.3) after formation, each nucleus grows deterministically, thus giving its deterministic contribution to the depletion of the supersaturation in solution;

(M2-H.4) perfect mixing, no secondary nucleation, no agglomeration, and no breakage are assumed.

In such a framework, the time when older nuclei were stochastically born influences the conditions at which future nuclei are stochastically born.

Such effect of the past on the future leads to a different type of stochastic process, since it means that Nucleation Mode 2 is not a Markov process, hence also not a Poisson process. In fact,

the hypotheses above lead to interpret nucleation as a *self-exciting point process* [56], which can be viewed as a generalisation of a inhomogeneous Poisson process (see Appendix A) where the intensity $\lambda(t)$ is itself stochastic, as it depends on a sequence of past stochastic events; since $\lambda(t)$ is a stochastic quantity, the self-exciting point process considered here is also a Cox process [63]. Equivalently, one can envisage the self-exciting point process as a step-by-step inhomogeneous Poisson process. Note that also the total number of primary nuclei formed during the process up to time t, $N_p(t)$, is a stochastic variable; its formal definition is discussed in Appendix A. According to hypothesis (M2-H.2), the probability P_1 that at least one new nucleus forms in the time interval $\Delta t = t_k - t_{k-1}$ is:

$$P_1(t_{k-1}; t_k) = 1 - \exp\left(- \int_{t_{k-1}}^{t_k} W(x) J(x) dx \right) \tag{2.21}$$

In this case, the formation of the nuclei according to the self-exciting point process defines the chain of stochastic events.

Modelling primary nucleation as a self-exciting point process has two important consequences. The first one concerns the physical description: Nucleation Mode 2 exhibits a higher degree of variability than Nucleation Mode 1; since in the latter the stochasticity is limited to the first event only, while the former to each new primary crystal. However, the process variability still reduces with increasing system size W, since the probability of formation of new primary particles is governed by Eq. (2.21), where the system size W appears in the exponent with a negative sign. This behaviour agrees with experimental evidence. The second consequence is of numerical nature. Some stochastic processes require Monte Carlo based algorithms and rejection methods to compute the average process properties: the efficiency of Monte Carlo algorithms depends on the integration step, Δt, which determines the number of time the rejection process is called. Too fine or too coarse choices can over- or under-estimate the number of nuclei that form, and a large number of simulations is needed simply to validate the

correct time step, seriously limiting the applicability of this specific method for regressing nucleation parameters. This method has been used, for instance, by Biscans et al. [64]. On the contrary, self-exciting processes can be simulated exactly by applying an inversion algorithm, when the intensity rate $\lambda(t)$ is known. This is indeed the case considered in this work, where J has a closed form, as discussed later in Sections 2.4 and 3.3. The algorithmic procedure works starting from the initial condition, $N_p(t_0) = 0$. Then, one generates a uniformly distributed random number and equates it to P_1, finding the corresponding time t_1 drawn from the exponential distribution given in Eq. (2.21). With t_1, one can compute the time interval $\Delta t = t_{N,1} - t_0$ elapsing before the first nucleation event occurs: the system, which now counts one particle, is updated. The next time interval $\Delta t = t_{N,2} - t_{N,1}$, elapsing before the second particle nucleates, can now be computed: this procedure is iterated for the formation of each new nucleus, until the detection condition is met. Note that the system update requires to account for the growth of the primary particles (as well as for any other deterministic phenomenon), i.e. it requires to specify the growth model.

2.4 MOLECULAR DYNAMIC SIMULATIONS*

Condensation of a liquid droplet from a supersaturated vapour phase is the prototypical case of nucleation and was also one of the cases originally considered by Gibbs. Although simpler than crystal nucleation, because the old and the new phase are both not ordered, the formation of a liquid droplet exhibits most of the essential features of nucleation also shared by crystallisation. In

*The results presented in Section 2.4 have been reported in: M. Salvalaglio, P. Tiwary, G. M. Maggioni, M. Mazzotti, M. Parrinello, "Overcoming time scale and finite size limitations to compute nucleation rates from small scale well tempered metadynamics simulations", *J. Chem. Phys.* **2016**, *145*, 211925 .

fact, droplet condensation must overcome an energy barrier, is a stochastic process, and requires the definition of an appropriate collective variable discriminating between the initial metastable state and the final stable state, in order to properly describe the system evolution. Hence, simulations of Argon condensation allow to illustrate well the level of details and understanding which can be gained from molecular dynamics, and the potential benefits of its application to more complex systems, such as crystallisation of organic molecules, salts, and proteins.

At realistic supersaturation conditions, condensation occurs on time scales that far exceed what can be reached with conventional molecular dynamics methods, hence it is challenging to compute its rate from atomistic molecular dynamics (MD) simulations. Another known problem in this context is the distortion of the free energy profile associated to nucleation due to the small, finite size of typical simulation boxes.

Recently, the problem of time scale has been addressed within the framework of enhanced sampling methods: Reguera et al. and Salvalaglio et al. [66, 67] have developed corrections for finite size effects, while Tiwary et al. [68] have proposed a method to directly estimate nucleation rates from enhanced simulations. Combining these approaches, we study the condensation of Argon, a paradigmatic substance, in the canonical NVT-ensemble for a system with a total N_a number of Argon molecules. We also show that characteristic nucleation times in the order of hours can be reliably calculated.

The choice of Argon has a twofold aim: first, to analyse a simple and yet significant system, which has been the subject of both small and large scale nucleation studies and allows to benchmark our results against the existing literature; second, to point out that even in such a simple system, at realistic values of supersaturation nucleation time scales in small volumes rapidly grow out of reach of standard MD. To overcome this last difficulty, we have used Well Tempered Metadynamics (WTmetaD), performing enhanced simu-

lations. The choice of the canonical ensemble also plays a key role in demonstrating the generality of our approach. At constant volume, the finite size effect induced by the coupling between nucleus size and pressure of the vapour phase is analogous to the effect due to the coupling between nucleus size and chemical potential observed in nucleation from multicomponent liquid phases.

The approach developed here for the calculation of transition times and the correction for finite size effects is thus general, and can be exported to systems of increased complexity and density, for which large scale approaches are impractically expensive. Nucleation rates spanning a range of ten orders of magnitude are computed at moderate supersaturation levels, thus bridging the gap between what standard molecular dynamics simulations can do and real physical systems. The results of the simulations, which are well fitted by an exponential distribution, provide further support to its choice as nucleation statistic, discussed in Section 2.3.1. Moreover, they illustrate the difficulties and challenges, and the implemented solutions to overcome them, of simulating nucleation at the molecular scale even for simple systems, e.g. the definition of correct energy potentials and the necessity of defining appropriate collective variables. At the same time, they also clearly highlight the potential benefits of applying computational methods to more complex systems, such as crystallisation from solution, particularly to better understand fundamental features of nucleation, which one cannot typically observe and measure, e.g. the shape of the nuclei at formation.

2.4.1 *Collective Variable, Nucleation Time, and Simulations Details*

First, the essential details of the application of WTmetaD to the calculation of nucleation rates are reported, and a brief analysis of finite size effects is carried out. WTmetaD and the finite size

correction are then combined to outline a systematic strategy for the calculation of nucleation rates.

The acceleration effect associated to WTmetaD has been exploited in order to substantially reduce the simulation time required to observe a nucleation event while simultaneously maintaining the system size small, hence significantly diminishing the overall computational cost. WTmetaD is conventionally used to compute free energy surfaces in a variety of contexts [69, 70]. In WTmetaD the simulated system evolves in a transformed time coordinate, t_{WT}, due to the application of the history-dependent bias potential $V_B(\zeta, t)$ constructed as a function for the collective variable ζ. As discussed in detail in Tiwary and Parrinello [68], rate calculations via WTmetaD do not require a converged estimate of free energy profiles, being instead based on the systematic evaluation of the so-called acceleration factor, which represents the ratio between the physical time and the metadynamics time.

In the context of a nucleation problem, the specific transition time associated to a nucleation event, i.e. the nucleation time t_N, can be computed from the corresponding WTmetaD simulation time $t_{N,WT}$ as:

$$t_N = t_{M,WT} \left\langle \exp\left(\hat{\beta} V_B(\zeta, t)\right) \right\rangle_{WT} \tag{2.22}$$

where the term $\left\langle \exp\left(\hat{\beta} V_B(\zeta, t)\right) \right\rangle_{WT}$ is the acceleration factor, called α_f in the following; note that $\hat{\beta} = 1/kT$, with k the Boltzmann constant. Crucial to this procedure is the hypothesis of negligible bias deposition at the transition state. In the case of rare events, the time which the system takes to cross the transition-state region is rather short compared to the total time of the simulation. Therefore, it is not likely that the crossing of the transition state and the Gaussian deposition are simultaneous, if one increases the time lag between two consecutive depositions. To comply with such hypothesis, the bias potential is constructed through the infrequent deposition of potential Gaussian, in a properly defined space of collective variables; the fulfilment of this condition can be checked

either a priori or a posteriori. In the former case, it is possible to control if the deposition rate is "infrequent enough" by performing some simulations with an increasingly slower deposition frequency: when the transition times converge within the chosen accuracy, the correct rate has been found. A more complete discussion of this method can be found in Tiwary and Parrinello [68]. In the latter case, since the method is based on the assumption that the transition across the energy maximum is a rare event, the statistics of transition times must follow the so-called law of rare events, which leads to the exponential distribution of in Eq. (2.13), as shown also by Salvalaglio et al. [57].

In WTmetaD the bias potential $V_B(\zeta, t)$ is constructed as a function of a collective variable (CV) ζ [69, 70]; here, we have chosen as collective variable the total number of liquid argon atoms in the system, n. n is a global coordinate defining the state of the entire simulation box and does not strictly correspond to the size of the largest cluster, which is the typical reaction coordinate used also in classical nucleation theory. Choosing a global definition of n has two advantages. The first is computational efficiency: a definition of the largest cluster as a CV would in fact require a computationally expensive clustering and sorting procedure performed at each time step. The second reason is due to the system size: here, we are specifically targeting small volumes, in which it is extremely unlikely to observe multiple nuclei simultaneously undergoing an irreversible transition to a stable droplet. Therefore, albeit n is defined globally, when a nucleation event unfolds in small volumes it tends to coincide with the size of the largest cluster. Note that for the case of droplet condensation, the capillary approximation necessary to develop the CNT is restricted to the existence of a surface and its related energy, whereas no ordering of the molecules in the cluster must be assumed: the main difference between the argon in the vapour and liquid phase is in fact the density.

To efficiently compute n at each time step, it is defined according to the ten Wolde-Frenkel definition, in which atoms are consi-

dered liquid when they possess a coordination number larger than a threshold value c_ℓ. For Argon, a value $c_\ell = 5$ has been found to appropriately describe the system [71, 72]. The coordination number of each molecule in the system is defined in a continuous and differentiable form through the expression:

$$c_i = \sum_{j \neq i} \frac{1 - (r_{ij}/r_c)^6}{1 - (r_{ij}/r_c)^1 2} \tag{2.23}$$

where r_{ij} is the Cartesian distance between atoms i and j. The number of molecules possessing $c_i \geq c_\ell$ is thus calculated as:

$$n = \sum_{i=1}^{N_a} \frac{1 - (c_\ell/c_i)^6}{1 - (c_\ell/c_i)^1 2} \tag{2.24}$$

The biased transition time $t_{M,WT}$ is the simulation time associated with the occurrence of a nucleation event in a WTmetaD simulation. Hence to compute $t_{M,WT}$, it is necessary to reliably detect nucleation events. In our case, we have applied the approach described in Tiwary and Parrinello [68], which is based on the fact that a clear time scale separation exists between the residence time in the supersaturated vapour state and the time necessary for a supercritical nucleus to grow in size.

Transition times were computed from simulations in the canonical ensemble (NVT-simulations) of systems consisting of $N_a = 512$ Argon atoms. Argon, being a noble gas, can be well-described as a system of almost ideal, neutral hard sphere, interacting through short range potentials (Lennard-Jones spheres). A Lennard-Jones potential with $\epsilon/k = 119.8$ K, and $\sigma_{LJ} = 0.3405$ nm was adopted to describe the interactions between the atoms. The time step for the integration of the equations of motion was set to 5 fs, following Tiwary et al. [68, 73].

2.4.2 *Free Energy of Nucleation in a Confined system*

We define confinement as the impossibility of exchanging atoms between the system and the surrounding environment. Under this definition an NVT simulation box represents a prototypical confined system, as its total number of atoms N_a is by definition constant. As highlighted in several reference works [67, 74, 75], the free energy change associated with a nucleation process is affected by confinement: because of the small volume of the system, the number of molecules in solution cannot be treated as constant during the formation of the nucleus, hence the supersaturation is depleted while the pre-critical cluster grows into a nucleus. The expression of the free energy derived in the classical nucleation theory must be then modified to account for such an additional effect. The Helmholtz free energy change, ΔF_{N_a}, for the transition from the initial to the final state, i.e. for the formation at constant temperature T and volume V of a n-molecule droplet from a vapour consisting initially of N_a molecules, is given by [74]:

$$\begin{aligned}\Delta F_{N_a}(N_a, n, V, T) = -\frac{n}{\tilde{\beta}} \ln \frac{p}{p_e} + \gamma(an)^{2/3} + \\ + \frac{N_a}{\tilde{\beta}} \ln \frac{p_0}{p_e} + \frac{n}{\tilde{\beta}}(1 - \tilde{\beta} v_\ell p_e)\end{aligned} \tag{2.25}$$

where p is the pressure in the vapour phase, p_0 the pressure in the vapour phase when no atom is in the liquid state, p_e the equilibrium vapour pressure, v_ℓ the molecular volume in the liquid phase, and $a = 6\pi^{1/2} v_\ell$, since the embryo with n atoms is assumed spherical. The two last terms on the r.h.s. of Eq. (2.25) are those related to confinement. Note that we treat the argon vapour as an ideal gas; hence the pressure of the vapour accounting for the formation of droplet with n atoms is:

$$p(N_a, n, V, T) = \frac{(N_a - n)kT}{V_n v_\ell} \tag{2.26}$$

The pressure at the initial state of the system is $p_0 = p(N_a, 0, V, T)$. A typical $\Delta F_{N_a}(n)$ profile is reported in Figure 2.4 (blue line), where it is also compared with the corresponding expression from the CNT (black line). The minimum at n^*_{min} is due to the aforementioned confinement effect, which corresponds to what one observes in a macroscopic system when nuclei grow into crystals in a closed system. In a confined system, even a single nucleus changes the bulk composition by growing, hence depleting the supersaturation: eventually, the system reaches a state, i.e. n^*_{min}, where the addition of any further molecule is energetically unfavourable. The corresponding free energy difference derived in the CNT is:

$$\Delta F(n, V, T) = -\frac{n}{\hat{\beta}} \ln \frac{p}{p_e} + \gamma(an)^{2/3} \tag{2.27}$$

The nucleation rate in the confined system, J_{N_a}, is obtained applying the transition state approached already used in the CNT (Eq. (2.9)), but using the maximum of the Helmholtz free energy $\Delta F^*_{N_a}$:

$$J_{N_a} = A_{N_a} \exp\left(-\hat{\beta} \Delta F^*_{N_a}\right) \tag{2.28}$$

The local maximum $\Delta F^*_{N_a}$ represents the free energy barrier to nucleation in a finite size system. The numerically computable $\Delta F^*_{N_a}$ is associated with a critical nucleus size $n^*_{N_a}$, which represents the liquid embryo in unstable equilibrium with the surrounding vapour. Wedekind et al. [74] and in Salvalaglio et al. [35, 67] have shown that below a minimum value of the initial supersaturation $S_0 = p_0/p_e > 1$, ΔF_{N_a} has not maximum, but increases monotonically. At N_a and T fixed, this condition defines an upper bound for the volume for which nucleation rates can be computed. It is also worth noting that, when a confined system is operated at the same values of T and S_0 as an infinitely large system, $\Delta F^*_{N_a} > \Delta F^*$.

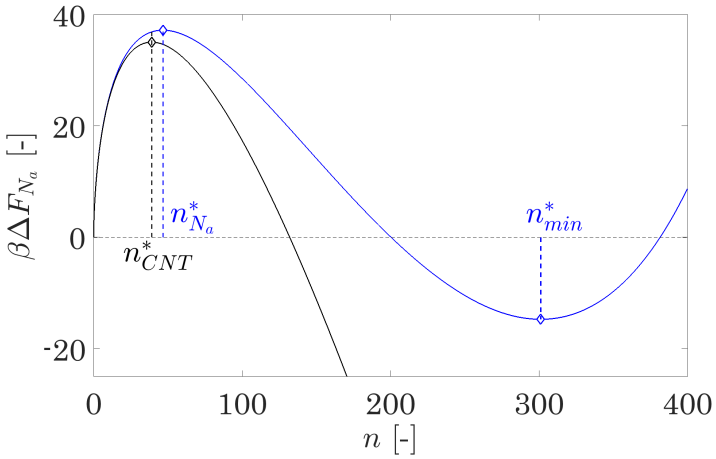


Figure 2.4: Nucleation of a liquid argon droplet from supersaturated vapour. The black line represents the free energy difference predicted by CNT in an infinitely large system, at constant supersaturation; the blue line the free energy difference for a finite size, confined system. Both free energy profiles refer to a system of 512 argon atoms, in a volume of 2197 nm3, at supersaturation S = 6.

2.4.3 *Simulation Results*

Sets of 50–100 simulations were carried out at $T = 80.7$ K at five different supersaturation, 11.4, 8.68, 6.76, 6.01, 5.36, labelled from S_1 to S_5, respectively; other sets with similar numbers of simulations were also performed, but at $T = 72.0$ K at four different supersaturation, 16.9, 15.7, 14.0, 12.0, labelled from S_6 to S_9, respectively.

In the simulations, nucleation is observed to occur randomly, as discussed in Section 2.3. From each set of simulations, one can construct the empirical cumulative probability of nucleation (see

Section 4.2.1 for more details), P^*, and its complement, P_0^*, i.e. the survival nucleation probability, which is plotted in Figure 2.5.

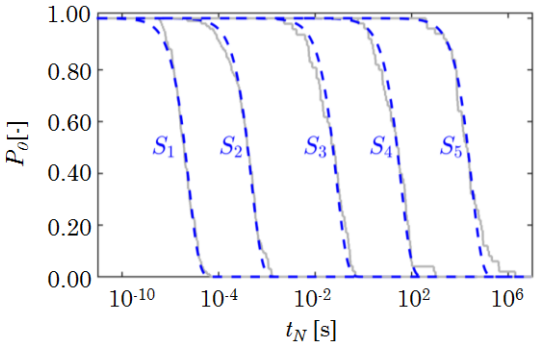


Figure 2.5: Nucleation of a liquid argon droplet from supersaturated vapour. Survival nucleation probability distributions for five different supersaturation levels $S_1 = 11.4$, $S_2 = 8.7$, $S_3 = 6.8$, $S_4 = 6.0$, $S_5 = 5.4$, obtained at $T = 80.7$ K.

The simulations also support the hypothesis of Section 2.3.1 that the exponential distribution is appropriate to describe the statistic of nucleation time: the empirical nucleation probabilities could be fitted very well by Eq. (2.15), with constant intensity $\lambda = 1/\tau_N$, where τ_N is the average nucleation time at a given supersaturation level. The values of τ_N extracted from the simulations are reported in Figure 2.6.

The simulations were stopped when the first nucleus formed in the system; hence, they can be fully described using either Nucleation Mode 1 or Nucleation Mode 2 in Section 2.3.2. The nucleation rate from the simulation results, J_{N_a}, has been thus estimated using Eq. (2.19) as:

$$J_{N_a} = \frac{1}{\tau_N V} \tag{2.29}$$

where the volume has been used for representing the system size. Since the system supersaturation in a finite size before a supercriti-

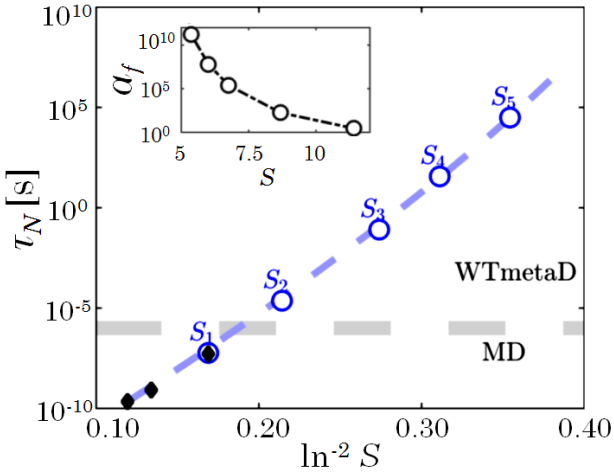


Figure 2.6: Average nucleation times in finite size systems τ_N. In the inset, the average acceleration factor (α_f) is displayed as a function of supersaturation.

cal nucleus forms is essentially the same as in a macroscopic system, i.e. $S = S_0$, the pre-exponential factors are almost the same in the two cases. Therefore, the ratio ϑ between the nucleation rate in macroscopic conditions, J, and in a finite sized confined system, J_{N_a}, reduces to [74]:

$$\vartheta = \frac{J}{J_{N_a}} \approx \exp\left(-\tilde{\beta}(-\Delta F^*_{N_a} - \Delta F^*)\right) \tag{2.30}$$

From Eq. (2.30), nucleation rates in macroscopic conditions are thus computed as:

$$J = J_{N_a} \exp\left(-\tilde{\beta}(-\Delta F^*_{N_a} - \Delta F^*)\right) \tag{2.31}$$

Note that J_{N_a} values systematically underestimate rates obtained from large scale simulations, justifying the finite size correction

ϑ: hence, rate calculations from small scale simulations require a systematic assessment of finite size effects.

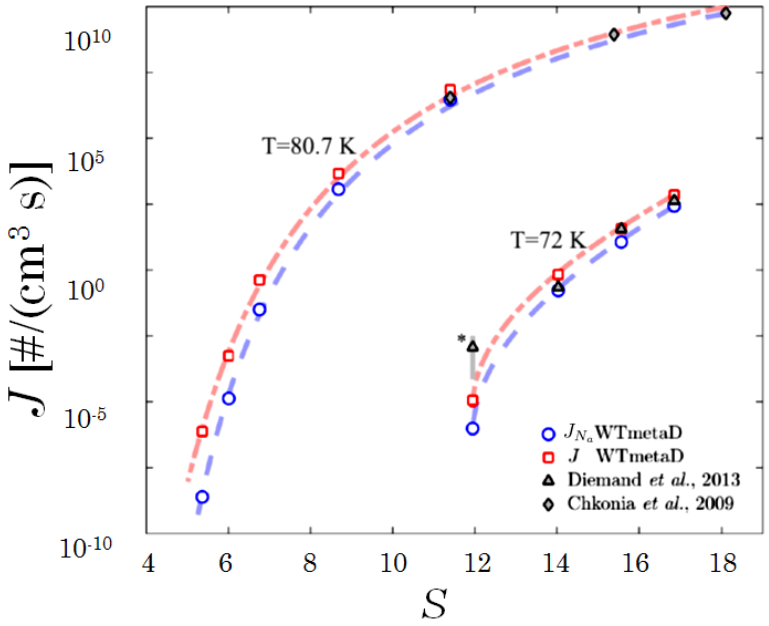


Figure 2.7: Nucleation rates calculated from finite sized WTmetaD simulations (J_{N_a}) and rescaled to the macroscopic limit J. The blue dashed line represents the result of the fitting of the J_{N_a} data with Eq. (2.9). The data-point from Diemand et al. [76], marked with $*$, has been computed on the basis of a single nucleation event.

The estimates of J and of J_{N_a} are reported in Figure 2.7. The finite size correction is substantially captured by the exponential approximation in Eq. (2.30) and it becomes negligible at high supersaturation ($S \geq 11.4$), where the energy barriers of nucleation in the unconfined and confined case are almost indistinguishable. Indeed, in the limit of high supersaturations (and small acceleration factors), the values in Figure 2.7 clearly indicate that the

estimates obtained with our method coincide with those obtained by Chkonia et al. [72] with the survival probability approach based on large-size standard MD simulations; their estimates of the nucleation rate are reported in Figure 2.7 as black symbols. In contrast with the method used by Chkonia and co-workers, however, we significantly extend the accessible nucleation time scale, without significant sacrifices in accuracy. For large scale simulations corresponding to low supersaturation values such as $S_9 = 12.0$, et al. based their conclusions on the observation of a single event. In contrast with the lack of reliable nucleation statistics obtained with large scale simulations, our approach allows to record tens of nucleation events, thus providing meaningful statistics for the estimate of nucleation rates in conditions proven to be challenging even for large scale simulations. To conclude, being able to simultaneously address both time scale and finite size limitations allows to significantly extend the range of nucleation conditions that can be directly investigated with computationally efficient, small-scale molecular simulations: these results unlock this possibility for processes up to ten orders of magnitude slower than those typically accessible through standard molecular dynamic simulations.

2.5 SUMMARY AND CONCLUSIONS

In this Chapter, we have first reviewed the two main theories of nucleation. The first theory, the so-called classical nucleation theory, has been developed over the course of few decades from a complementary thermodynamic and kinetic perspective. It applies not only to the formation of crystals, but also to droplet condensation from a vapour and bubble formation from a liquid; in fact, these two cases were those for which the theory was originally conceived. In the classical nucleation theory, the clusters forming from the solution are envisaged to have all relevant properties of a crystal and the formation of a nucleus is fully described using only one reaction coordinate, i.e. the number of molecules which compose

it. The second theory, the so-called two-step nucleation theory, on the contrary, has been developed specifically for crystal nucleation and accounts for the formation of a new nucleus along two distinct reaction coordinates: a density coordinate, i.e. the number of units gathering together, and an order coordinate, i.e. the orientations of the molecules within the clusters with respect to one another. Both theories are based on the idea of nucleation as an activated process.

The activated nature of nucleation means that, ultimately, nucleation is a stochastic process: this theoretical consideration has been confirmed both computationally and experimentally and it is currently accepted as the standard theory. In light of this, we have discussed how the formation of a new nucleus can be described accounting for its inherent stochastic behaviour. We have done so by modelling the distribution of the time at which a cluster becomes a nucleus, i.e. when an entity identifiable with the new phase appears from the old phase in a system. We have discussed two different models aimed at describing a nucleation. The first one holds that only the very first nucleus appearing in the system forms stochastically, while the subsequent nuclei form deterministically. The second model, on the contrary, assumes that each new nucleus, not only the first one, form stochastically; by hypothesising this mechanism, the second model introduces a dependence of the entire process on the history of the system. These two models will constitute the starting points of more comprehensive descriptions of crystallisation, which include also the growth of the nuclei into fully grown crystals, discussed in Chapter 3 in the following.

Finally, we have investigated nucleation at the molecular level, simulating the nucleation process in a confined system via molecular dynamics. These simulations reproduced the condensation of liquid Argon from its vapour, highlighting few important aspects. First, the exponential statistics indeed represents well the distribution of nucleation times in an ideal system, where conditions can be perfectly set and controlled. Second, in absence of contamination

due to pollutants, the results could be explained and interpreted in terms of classical nucleation theory. The theory needed minor modifications to account for the finite size of the system, because the original formulation considered an infinitely large system. The simulation results suggest that the CNT is at the very least a good starting point for interpreting the behaviour of nucleation. However, more sophisticated theories, such as 2-SNT, may be needed to describe appropriately more complex systems, such as multicomponent systems, or systems constituted by complex molecules, with strongly non-ideal behaviours, such as amino-acids, proteins, or polymers. Third, the simulations illustrated the challenges and the limitations of simulating even a simple model compound such as Argon, i.e. defining appropriate collective variables and performing simulations at experimentally accessible conditions. Applying a method recently proposed to recover the dynamics of the system from metadynamic simulations, we have also shown that improving the efficiency of computations is a crucial field of research, to be able to investigate increasingly more complex, and thus more realistic, systems. The method of Tiwary and Parrinello, based on physical as well as mathematical arguments, can overcome the issue of simulating only experimentally unrealistic conditions, thus yielding computational results possibly directly comparable with experiments. Such a direct comparison would certainly advance the understanding of nucleation, particularly the influence of impurities. Moreover, the computational techniques are a valuable tool to advance the fundamental understanding of nucleation (and related phenomena) at a molecular, not directly observable level.

3

MODELLING CRYSTALLISATION*

3.1 INTRODUCTION

The importance of describing and predicting the course of primary nucleation in crystallisation processes cannot be overestimated. However, these tasks are challenging because the correlations and the rules used to scale up information and models obtained in the lab to industrial processes are often not accurate enough. Furthermore, the characterisation of nucleation events is experimentally always difficult: since monitoring directly nuclei is typically not possible, one uses indirect information, by monitoring when crystals become observable for the first time in the system, i.e. the detection time.

In this Chapter, we first introduce the distinction between nucleation times and detection times and present the general mathematical formalism to describe them. Then, based on the general framework, two specific models are proposed, which differ in how the nuclei form and develop into crystals. Both models account for the stochasticity of primary nucleation and for the possibility of operating at different operating conditions, such as concentration, temperature, temperature profile, and system size.

In the following, we consider a system of size W undergoing cooling crystallisation. The process starts from an initial temperature, T_0, where the solution is saturated and clear (i.e., $c_0 = c_s(T_0)$),

*The results presented in this Chapter have been reported in: G. M. Maggioni, M. Mazzotti, "Modelling the stochastic behaviour of primary nucleation", *Faraday Discuss.* **2015**, *179*, 359–382 . and G. M. Maggioni, M. Mazzotti, "Stochasticity in Primary Nucleation: Measuring and Modeling Detection Times", *Cryst. Growth Des.* **2017**, *17*, 3625–3635 .

and ends at a final temperature, T_f; note that $T_f < T_0$. The temperature evolution of the system is supposed to follow exactly the desired profile thanks to an ideal thermostat, and the system is assumed homogeneous, i.e. well-mixed. During the process, an arbitrary total number N of particles can form, whose growth depletes supersaturation. The models can be easily generalised to different reactors, e.g. tubular crystallisers, and to different crystallisation methods, e.g. evaporative or anti-solvent crystallisation.

The two models presented in this Chapter have been applied and validated against experimental data obtained in different systems, from which the necessary model parameters have been estimated. These aspects are discussed in Chapter 4, which is a natural extension of what is discussed in the present Chapter. Therefore, conclusions concerning Chapters 3 and 4 will be summarised together at the end of Chapter 4.

3.2 NUCLEATION TIME VERSUS DETECTION TIME

Nuclei are nano-scale objects, which normally cannot be observed when they form, but only after growing to a size large enough either to be directly observable with an optical technique or to modify a measurable property of the suspension, such as the light absorbance. Thus, a complete physical and mathematical model of the entire crystallisation, and not only of nucleation, must be provided. In general, the detection time, t_D, is the time elapsed from the establishment of supersaturation to the moment when some measurable property of the system reaches and/or overcomes a threshold value. The nucleation time, t_N, is defined as the time elapsing between the establishment of supersaturation and the formation of the first nucleus, whereas the growth time, t_G, is simply the time elapsing between the nucleation time and the detection time. Hence, t_G depends not only on the growth rate, i.e. the speed at which newly formed nuclei grow while consuming the solute in the mother liquor, but also on the nucleation rate and on

other phenomena possibly occurring during the process, e.g. secondary nucleation, agglomeration, and breakage. With the definitions above, the three times are related as:

$$t_D = t_N + t_G \tag{3.1}$$

The difference between the nucleation time and the detection time as well as the models of t_N and t_G are an crucial aspect of crystallisation, because the interpretation of the experimental data rests ultimately on the ability to retrieve the information about t_N from t_D. Consequently, it is worth discussing such a difference and its implications for modelling in more detail.

To precisely measure detection times, one must be able to monitor, possibly on-line, a specific property of the system. In general, crystallisation experiments rely on measuring turbidity (through turbidimetry and nephelometry), absorbance (typically through infra-red absorbance, ATR-FTIR, more rarely UV-absorbance), or reflectance/scattering of light in the suspension (through Focused Beam Reflectance Measurement, FBRM, and light scattering). One can estimate nucleation kinetics by using a model which links these measured quantities to one or more properties of the crystal distribution (typically moments of the distribution, see Section 3.3) and/or of the solution, for example:

- the number of crystals (estimated from FBRM/light scattering measurements);
- the crystal average size(FBRM/Light Scattering);
- the crystal volume fraction (from turbidimetry, FBRM, and light scattering measurements);
- the total crystal surface per unit volume of the suspension (FBRM, turbidimetry, light scattering)
- the solute concentration (measured by ATR-FTIR).

In order to develop a model which describes how the system evolves from nucleation to detection and to interpret the detection time measurements, further assumptions are necessary, beyond those related to nucleation. In this work, two assumptions have been specifically used:

Det-H.1 The detection time is formed by the additive contribution of nucleation and growth time, i.e. Eq. (3.1);

Det-H.2 the observed stochastic behaviour of detection time is entirely due to primary nucleation, whereas crystal growth is treated as a purely deterministic phenomenon.

Given the hypotheses above, one can mathematically envisage the nucleation time, the growth time, and the detection time as the (model) outputs of three mappings $\mathscr{H}_N$, $\mathscr{H}_G$, and $\mathscr{H}_D$, respectively. These mappings determine the evolution of the crystallising system and use as inputs the initial concentration, c_0, the system size W (which can be the mass of solvent, or the volume of the solution), the temperature evolution, $T(t)$, and the detection condition, Ψ, hence they return the relevant characteristic times:

$$t_N = \mathscr{H}_N[c_0, T(t), W] \tag{3.2}$$

$$t_G = \mathscr{H}_G[c_0, T(t), W, \Psi] \tag{3.3}$$

$$t_D = \mathscr{H}_D[c_0, T(t), W, \Psi] \tag{3.4}$$

The functional forms of these mappings is discussed in Section 3.3 below. Note that, by Det-H.1, it is sufficient to define separately $\mathscr{H}_N$ and $\mathscr{H}_G$, whereas H_D can be obtained by simply adding them together.

At this point, it is worth observing that the exponential distribution discussed in Section 2.3 models the statistic P of t_N, not that of t_D, labelled as F. In general, the distribution functions P and F are different, even though they are related through the mapping $\mathscr{H}_G$. Particularly, we briefly consider two cases of interest for the application in experiments:

C.1 the phenomena between nucleation and detection are deterministic, and the system operating conditions are at a steady-state. Then, the output of $\mathscr{H}_G$ is a constant value of t_G for each t_N, thus t_D is an affine transformation of t_N [77]. Moreover, $P(t_N) = F(t_D)$ and $dP(t_N)/dtN = dF(t_D)/dt_D$;

C.2 the phenomena between the first nucleation event and detection are deterministic, but the value of t_G may depend on the specific t_N associated to it, e.g. if the system is not at steady-state. Thus, $\mathscr{H}_G$ is deterministic and $P(t_N) = F(t_D)$, but $dP(t_N)/dt_N \neq dF(t_D)/dt_D$; t_D is in this case a similarity transformation of t_N.

In the two cases just considered, the mapping between P and F is particularly simple and reduces, geometrically, to a similarity transformation [77]. Note also that case C.1 implies a stronger condition than case C.2: if one knows the value of t_G for one point, one knows the value of t_G for any point of F.

Given the distinction between nucleation time and growth time, it is worth commenting on the conclusions reached by some researchers and found in their works, in which they state that the exponential distribution is not appropriate to described best the data. Rather, these studies have used either a normal, log-normal, or a Weibull distribution to fit the experiments. We note that their analyses were performed by assuming either that the importance of t_G negligible [52] (although at times without any explicit investigation of the actual value of t_G), or by neglecting to verify the actual temperature of the system and the evaporation of the solvent, thus possibly causing variations in supersaturations, or in same cases facing uncontrolled temperature changes during the experiment [50]. Furthermore, only a partial analysis of experimental reproducibility was systematically performed, which could have biased the conclusions about the behaviour of the data, as discussed in Chapters 4 and 5. Under such uncertain conditions, a clear discrimination about which statistics best suits the data is rather

speculative and no absolute conclusion should be reached. Moreover, the distributions used have no physical justification and, in the case of the normal distribution, can hardly be reconciled with the fact that they are properly defined also for negative values of the nucleation times/detection times.

In Section 3.3 below, we build a physically consistent and meaningful, though possibly complex, model, trying not only to minimise the number of parameters required for its implementation, but also to assign them physical meaning. Before introducing our models, we recall the approaches used in some of the most relevant literature and the detection techniques they exploited, which are summarised also in Table 3.1, where, for each paper, the system volume and the detection technique applied are reported. First, in small volumes, and particularly in microfluidic systems, one can often assume that growth is much faster than nucleation, so that the time between particles formation and detection is negligible, i.e. $t_N \approx t_D$ [37, 46, 64, 78–80]. Alternatively, one can decouple nucleation and growth, analysing them independently, as for example by applying the so-called double-pulse technique [11, 81–84], again a technique typically used in microfluidic devices. The most commonly used detection techniques for crystallisation can be divided into two main categories. The first one relies on direct observation of crystals, mainly through optical microscopy, where the detection of crystals depends on the resolution of the optical instrument [11, 37, 46, 64, 78–84]. The second category comprises all experiments where a macroscopic system property, which is related to the volume fraction of the crystals such as transmissivity or turbidity, is measured [38, 39, 44, 47, 51, 55, 59, 61, 62, 85, 86]: crystals are detected, hence the detection time is recorded, when the volume fraction exceeds a specific threshold value. We highlight that this latter technique, albeit very popular, suffers from two weaknesses: first, the quantitative relationship between solid volume fraction and, for example, turbidity is difficult to calibrate; second, the threshold value is difficult to estimate and is dependent on a

series of factors, some related to the device itself, e.g. its sensitivity and accuracy, and some to intrinsic properties of the system, e.g. the fluid-dynamics and the refractive index. Observation using optical microscopy is typically utilised in very small volumes, whereas light absorbance (transmissivity, turbidity, or IR measurements) is exploited in larger volumes.

Table 3.1: Overview of relevant literature on stochastic nucleation from solution and the key features of each study. The category "Stochastic process" applies to the cases where stochastic nucleation is observed but no further specification about its nature is provided.

Literature Overview			
Reference	*Volume*	*Det. Technique*	*Interpretation*
[64, 78, 82–84, 87–90]	10 pL - 100 nL	Optical microscopy	Poisson process
[46, 91, 92]	1-10 µL	Optical microscopy	Poisson process
[47, 85, 93]	1-2 mL	Transmissivity	Poisson process
[38, 52, 94–96]	5-30 mL	Turbidity	Stochastic process
[40, 41]	50-1000 mL	Turbidity	Deterministic process

Finally, we conclude this section reporting as an example of detection time measurements the raw data from the study of Kadam et al. [47]. The data are discussed in further detail in Chapter 4, since they have also been used to validate the models which are discussed in Section 3.3. Kadam and co-workers carried out a large number of experiments of cooling crystallisation of paracetamol from aqueous solutions in vessels of different volumes, to determine the so-called Metastable Zone Width (MZW), using a 1 mL crystalliser and exploring four different initial paracetamol concentrations. All these experiments started by applying a linear cooling profile at the saturation temperature corresponding to the selected concentration. Systematic experiments were also carried out in larger crystallisers, namely of volumes of 500, 700, 900 and 1000 mL, using the same cooling profile. In each experiment, the detection time was measured (using a transmissivity sensor for the 1 mL reactor and by eye inspection in the larger volumes) and cu-

mulative distribution curves of detection times, F^*, were obtained at each initial concentration and for each crystalliser volume (see Chapter 4 for more details). The two sets of data at the lowest and highest explored concentration, namely 15 g/L and 47 g/L (the others being 22 g/L and 32 g/L) are reported in Figure 3.1, where the cumulative probability is plotted as a function of the temperature change between time zero and the detection time, ΔT.

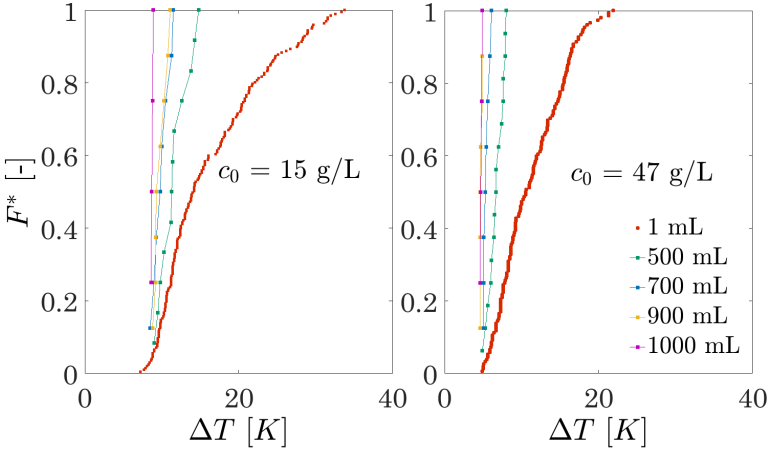


Figure 3.1: The experimental detection times, expressed as empirical cumulative nucleation probability, defined in Section 4.2 and labelled as F^*, vs. the extent of cooling, ΔT, for the cooling crystallisation of paracetamol from aqueous solution, measured by Kadam et al. [47]. The experiments shown here represent two different initial concentrations, 15 g/L (left) and 47 g/L (right), for crystallisers of five different volumes. The lines connecting the data at larger volumes (from 500 to 1000 mL) are drawn as guide for the eye.

3.3 POPULATION BALANCE EQUATION AND SINGLE CRYSTALS MODELS

This section focuses on describing the evolution of the crystallising system from t_N onwards. In a crystallising system, typically multiple crystals are present, forming what can be thought of as an ensemble. The characterisation of how such an ensemble evolves during the course of crystallisation can follow two different approaches:

(i) the crystals collectively form a population and they can be of any size. In this case, one neglect the individual identity of the crystals and describes them in terms of a continuous particle size distribution (PSD, called also equivalently "crystal size distribution"), i.e. a continuous distribution in the continuous variable represented by the crystal size(s);

(ii) the crystals are treated individually, hence considering each crystal a distinct entity, whose evolution is described separately, but not independently from its other siblings. In this case, the crystals form a discrete population of the continuous variable represented by the crystal size(s).

The first approach, introduced by Hulburt and Katz [97] and Randolph [98], clearly does not describe the reality of the crystals. However, it represents a useful approximation, successfully applied to several different systems. In fact, this representation describes many different phenomena, occurring during the particles formation and evolution, so well, that it has become the standard modelling approach in particle process and engineering. This approach is used to define the mapping $\mathscr{H}_G$ for Model 1, as discussed in Section 3.3.1. The second approach represents more close the actual system, but suffers from major limitations: the typical number of crystals in a process can be as low as several thousands (in small systems) to as high as several billions (in the standard lab-scale).

By describing each crystal individually, then, one should deal simultaneously with a corresponding number of equations. This becomes computationally first very expensive and, eventually, not feasible. However, in small systems, or in systems with a very slow rate of formation of particles, the second approach can still be used, as we have done to define the operator $\mathscr{H}_G$ for Model 2 in Section 3.3.2.

In the following, we further assume that the crystals are geometrical objects with a single characteristic dimension, the size L. The volume shape factor, k_v, and the area shape factor, k_s, relate the volume and the area of a crystal to L, as $k_v L^3$ and $k_s L^2$, respectively, as in Eqs. (2.3) and (2.4) in Section 2.2.1. The nuclei are considered as geometrical points, i.e. massless objects. Crystal growth is assumed to be size-independent.

3.3.1 *Model 1: Population Balance Equation*

Since the crystals are fully characterised by L, the (continuous) particle size distribution, $f(t, L)$, is one-dimensional (1D-PSD) [99, 100]. The quantity $f(t, L)dL$ is the number of particle per unit of W (size of the system) of particles of size between L and $L + dL$. Given the population f, one can define the j^{th}-order moment of the 1D-PSD, $\phi_j(t)$:

$$\phi_j(t) = \int_0^{+\infty} L^j f(t, L) dL \tag{3.5}$$

The evolution of a 1D-PSD depends on the production/disappearance and growth/shrinkage of the particle, i.e. on the kinetics of the system, and it is described by a dimensional Population Balance Equation (1D-PBE) or, alternatively but equivalently, by its associated Equations of Moments (EoM). In the case considered here, neither agglomeration nor breakage occur in the system, as for hypothesis (M1-H.3), hence particles can only form and grow. Crystal nucleation is described by the Nucleation Mode 1 nucleation

in Section 2.3.2, before t_N no crystals are present. With these assumptions, the 1D-PBE describing the evolution of f is:

$$\frac{\partial f}{\partial t} + G\frac{\partial f}{\partial L} = J\frac{d\delta_0^*(L)}{dL} \tag{3.6}$$

where G is the growth rate, J is the primary nucleation rate, and $d\delta_0^*(x)/dx$ is the formal derivative of $\delta_0^*(x)$, which is a Dirac measure sitting at $x = 0$ (see Appendix A.3 for further comments about the definition of $\delta_0^*(x)$ and of $d\delta_0^*(x)/dx$). The nuclei (either the first or the subsequent ones) have size $L = 0$. Due to hypothesis (M1-H.1), Eq. (3.6) is valid only for $t \geq t_N$. The initial conditions, assigned at the stochastic value $t = t_N$, and the boundary conditions of Eq. (3.6) are:

$$\begin{cases} f(t = t_N, L) = \dfrac{1}{W}\dfrac{d\delta_0^*(L)}{dL} \\ f(t, L = 0) = 0 \end{cases} \tag{3.7}$$

These conditions are consistent with nuclei represented as massless points. The EoM associated to Eq. (3.6) are:

$$\frac{d\phi_j}{dt} = J\delta_{j0} + jG\phi_{j-1} \qquad j = 0, 1, 2, ... \tag{3.8}$$

where δ_{j0} is the Kronecker-delta (it is 1, if $j = 0$, and 0 otherwise), and $\phi_{-1}(t) = 0$. The initial conditions of Eqs. (3.8), consistent with those in Eq. (3.7), are:

$$\begin{cases} \phi_0(t_N) = \dfrac{1}{W} \\ \phi_j(t_N) = 0 \qquad j = 1, 2, 3, ... \end{cases} \tag{3.9}$$

Note that Eqs. (3.6)-(3.9) can account also for the more general scenario with primary and secondary nucleation; in this case, the primary nucleation rate J in Eqs. (3.6) and (3.8) is replaced by the total nucleation rate J_t, i.e. the sum $J_t = J + J_s$. However, it

is customary to consider only primary nucleation when describing detection time experiments, and we have followed this approach in this work.

If one is interested only in computing some average properties, e.g. the average size or the crystal volume fraction, Eqs. (3.8) are solved only for $j = 0, 1, 2, 3, 4$, since most physical properties of interest of a 1D-PSD are related to the first four moments. The zero-th order moment, $\phi_0(t)$, is the total number of particles per unit size; therefore, the total number of particles in the systems is:

$$N_p(t) = W\phi_0(t) \tag{3.10}$$

The third order moment, $\phi_3(t)$ multiplied by k_v, represents the volume of solids in the suspension per unit of W. Through the mass balance, it links the actual solute concentration $c(t)$ to the initial concentration, c_0:

$$c(t) = c_0 - k_v\rho_c\phi_3(t) \tag{3.11}$$

where ρ_c is the crystal density (kg/m^3).

The integration of the EoM is performed until the detection condition is fulfilled, and the time associated to it is the (model) detection time t_D. Such condition is typically given in terms of a function of one or more moments of the distribution reaching a specific value:

$$\Psi(\phi_0(t_D), \phi_1(t_D), \phi_2(t_D), \phi_3(t_D), ...) = \psi \tag{3.12}$$

where the function Ψ depends on the physical principles behind the measurement and ψ is system- and instrument-specific threshold value. A common detection condition in the literature [41, 47] (used for techniques based on transmissivity, IR and Raman spectroscopy, and turbidity) is defined in terms of the crystal volume fraction in system. Such a volume fraction changes during the course of crystallisation due to the formation and growth of new crystals and it is proportional to the third-order moment, i.e. $\dfrac{V_c(t)}{V} \propto \phi_3(t)$.

Consequently, the detection time t_D is defined in these cases as the time when the third moment reaches a threshold value:

$$\Psi = k_v \phi_3(t_D) = \psi_V \tag{3.13}$$

where V_c is the total crystal volume in the suspension, V is the volume of the suspension, and ψ_V the threshold value. In the following, the threshold value is set to $\psi_V = 10^{-4}$. A second type of detection condition is based on the largest crystal reaching a threshold size and, thus, triggering detection (e.g., by promoting secondary nucleation [47, 59, 85], or by becoming visible in microfluidic experiments monitored with optical microscopy [64, 78, 88]). In this case, the detection condition is given by a threshold value, ψ_L, on the average number weighted crystal size, $L_{10} = \phi_1 / \phi_0$:

$$\Psi = L_{10}(t_D) = \frac{\phi_1(t_D)}{\phi_0(t_D)} = \psi_L \tag{3.14}$$

In the following, the threshold value is set to $\psi_L = 10$ µm.

3.3.2 *Model 2: Discrete Population*

The discrete population, associated in this work to the Nucleation Mode 2 (Section 2.3.3), accounts for the fact that all crystals form stochastically; hence, at time t, their number is $N_p(t)$, which is a monotonically increasing staircase function, with $N_p(0) = 0$ and all the step increments equal to 1. In this approach, crystals are considered individually and each new crystal, labelled k, is born at the corresponding nucleation time $t_{N,k}$. It then grows at deterministic rate according to the following equation:

$$\frac{dL_k}{dt} = G(S,T) \qquad t \geq t_{N,k} \tag{3.15}$$

where G is the growth rate, dependent on the supersaturation S and on the temperature T. Since the nuclei are assumed to be

massless, hence having zero size, and since for $t \leq t_{N,k}$, the k-th crystal does not exist, the initial condition associated to Eq. (3.15):

$$L_k(t \leq t_{N,k}) = 0 \tag{3.16}$$

For $t = t_{N,k}$, Eq. (3.16) is the initial condition associated to Eq. (3.15). Considering the hypothesis (M2-H.4), one can define the j^{th}-order moment of the discrete population:

$$\tilde{\phi}_j(t) = \sum_{k=1}^{N_p(t)} L_k^j(t) \qquad j = 1, 2, 3.... \tag{3.17}$$

while for the zero-order moment, it holds:

$$\tilde{\phi}_0(t) = \frac{1}{W} \sum_{k=1}^{N_p(t)} \delta_0^*(t - t_{N,k}) \tag{3.18}$$

Using Eqs. (3.15) and (3.17), one can show that:

$$\frac{d\tilde{\phi}_j}{dt} = jG\phi_{j-1} \qquad j = 1, 2, 3, ... \tag{3.19}$$

Note that the previous equation is not valid for the zero-th order moment, which is not differentiable in the classical sense, because it is a step function with discontinuities. Nevertheless, by interpreting the differentiation and the integration in the sense of stochastic calculus [101], though, one can formally write:

$$\frac{d\tilde{\phi}_0}{dt} = \sum_{k=1}^{N_p(t)} \frac{d\delta_0^*(t - t_{N,k})}{dt} \tag{3.20}$$

where the formal derivative of δ_0^* has the same meaning as in Eq. (3.7). The initial conditions of the moments of the discrete population are given at $t = t_{N,1}$ and they are formally identical to those of the moments of the continuous population:

$$\begin{cases} \tilde{\phi}_0(t_{N,1}) & = \dfrac{1}{W} \\ \tilde{\phi}_j(t_{N,1}) & = 0 \qquad j = 1, 2, ... \end{cases} \tag{3.21}$$

Similarly to the PBE description in Section 3.3.1, the mass balance and the detection condition can be written in terms of the third moment. The mass balance is:

$$c(t) = c_0 - k_v \rho_c \tilde{\phi}_3(t) = c_0 - k_v \rho_c \sum_{k=1}^{N_p(t)} L_k^3(t) \tag{3.22}$$

If crystal detection is based on the crystal volume fraction, the detection condition Ψ is a function of $\tilde{\phi}_3$ as:

$$\Psi = k_v \tilde{\phi}_3(t_D) = \frac{k_v \sum_{k=1}^{N_p(t_D)} L_k^3(t_D)}{V} = \psi_V \tag{3.23}$$

where ψ_V has the same value as in Eq. (3.13). Analogous results can be obtained with the condition given by Eq. (3.14).

3.3.3 *Constitutive Equations*

The driving force of crystallisation is expressed in terms of supersaturation, rigorously defined as the ratio of the activity at the actual conditions and at the equilibrium conditions. However, as discussed in Section 2.2, when the ratio of the activity coefficients $\gamma_a(T, c)/\gamma_a(T, c_s) \approx 1$, the supersaturation S is expressed simply as the ratio between the actual solute concentration, c, and the temperature dependent solubility, $c_s(T)$:

$$S = \frac{c}{c_s(T)} \tag{3.24}$$

where T is the actual temperature of the system. The functional form of $c_s(T)$ depends on the specific solute-solvent system: as an exponential law based on the Van't Hoff ideal solubility, if one adopts an approach based on physical consideration, or as a polynomial, if it is considered a mere fitting in the temperature range of interest. For antisolvent crystallisation, c_s would be not only a function of T, but also of the antisolvent concentration.

The constitutive equations necessary to complete Model 1 and Model 2 are those of the primary nucleation rate, J, and of the growth rate, G. In this work, without loss of generality, they have been chosen based on the expressions provided by the classical nucleation theory, discussed in Chapter 2 and given by Eq. (2.9), and on the (size independent) birth and spread growth model, for J and G, respectively [4, 41, 102]. They are:

$$J(S,T) = A_0 \exp\left(-\frac{A_1}{T}\right) S \exp\left(-\frac{B}{T^3 \ln^2 S}\right) \tag{3.25}$$

$$G(S,T) = K_0 \exp\left(-\frac{K_1}{T}\right) (S-1)^{2/3} (\ln S)^{1/6} \exp\left(-\frac{K_2}{T^2 \ln S}\right) \tag{3.26}$$

where S is the supersaturation and T the temperature; A_0, A_1, and B are the parameters for the nucleation rate, while K_0, K_1, and K_2 those for the growth rate. This choice of the kinetic rates is justified, for the solutes and solvents discussed in Chapters 4-5, by several pieces of evidence reported in the literature [40, 41, 103–106], where Eqs. (3.25) and (3.26) has been shown to describe the data well.

The set of model parameters can thus be represented, for both models, as the vector $\boldsymbol{\omega} = [A_0; A_1; B; K_0; K_1; K_2]^T$, and has to be assigned to solve the model equations.

4

VALIDATING THE MODELS AGAINST EXPERIMENTAL DATA*

4.1 INTRODUCTION

The development of new experimental devices has helped producing a wealth of data from which one could formulate new theories and models of nucleation. At the same time, though, these results have yielded some new and important challenges for both experimentalists and modellers [13, 36, 37, 46, 49, 58, 59, 80, 85, 106–113].

First of all, each set of analysed experiments should be truly representative of the stochastic process. On the one hand, it should be possible to reproduce experiments in a reliable way, so that the empirical statistics extracted from different repetitions of the same experiment effectively belong to the same underlying distribution. On the other hand, enough measurements should be collected to properly reconstruct the actual probability distribution function from the experiments, as it is the case for every statistical phenomenon. Were the data not representative of the parent statistics, the analysis, hence the results of such analysis, would be meaningless.

A second issue concerns how to estimate nucleation kinetics from detection time measurements characterised by an inherent stochas-

*The results presented in this Chapter have been reported in: G. M. Maggioni, M. Mazzotti, "Modelling the stochastic behaviour of primary nucleation", *Faraday Discuss.* **2015**, *179*, 359–382 . and G. M. Maggioni, M. Mazzotti, "Stochasticity in Primary Nucleation: Measuring and Modeling Detection Times", *Cryst. Growth Des.* **2017**, *17*, 3625–3635 .

ticity. We have discussed in Chapter 3 how additional assumptions about detection and growth have to be made, how nucleation and growth kinetics have to be estimated at the same time, and how the stochasticity and the variability of nucleation rate measurements propagate to the model parameters belonging to the nucleation rate expression: all these problems are well documented also in the literature [36, 47, 59, 93, 106, 114].

A third issue, strictly correlated to experiments reproducibility, concerns the sensitivity of the system to small variations of initial and boundary conditions (always occurring within different repetitions of a theoretically identical set of experiments). A system should not be sensitive to such variations in a strong way, otherwise it would not be possible to discriminate between variations due to the stochastic behaviour and those due to experimental uncertainties. The latter, even if fundamentally different in nature, can produce results that appear to be stochastic. Despite the similarity, though, these types of variabilities have a different physical nature and interpretation.

Additionally, it is mandatory to properly understand the detection technique used to measure the experimental data, in order to provide an appropriate description of them with a model. In fact, the definition of the detection conditions provides, ultimately, the link between the nucleation time and the detection time.

Exploiting the mathematical models discussed in Chapter 3, we now discuss the issues listed above. First, we provide the theoretical background on which our statistical analysis is based (Section 4.2) and we discuss how the vector of model parameters, $\boldsymbol{\omega}$, can be estimated by fitting experimental results (Section 4.3). Then, we report and discuss the results of applying the models to two different types of cooling crystallisation experiments (Section 4.4). Third, in order to consider all possible reasons for discrepancies between the model and the data, we comment on the sensitivity of the system to operating parameters (Section 4.5) and we inspect in more detail the detection technique used to measure the data,

discussing whether a better definition for the detection condition could be implemented in the model.

Lastly, we summarise the results and draw some conclusions (Section 4.7).

4.2 FUNDAMENTAL HYPOTHESES AND STATISTICAL PRELIMINARIES

Detection time experiments have been repeatedly observed and reported to be the result of a stochastic process [37, 46, 47, 59, 85, 115]. A proper analysis and a quantitative interpretation of such experiments rest on three fundamental assumptions:

(H.1) the experiments are carried out under identical and reproducible experimental conditions;

(H.2) the detection times are stochastic events, which sample one and the same cumulative probability function, which depends on the chemico-physical properties of the system;

(H.3) the experiments are in a sufficient number to form a representative sample of the parent statistic.

To provide a precise understanding of these hypotheses and of the implications they have, we introduce in this section some necessary basic definitions and concepts.

4.2.1 *Empirical Distribution Function*

Let us consider a set of Q acceptable experiments and q detection times, t_D, where q is the number of experiments in which crystals have indeed been detected (hence, $Q - q$ is the number of acceptable experiments in which no crystal has been detected before the end of the experiment). A more detailed explanation of when an experiment is "acceptable" is discussed in Section 5.2.2.3. Since

nucleation is an activated phenomenon [3], the nucleation process is stochastic, hence the measured detection time t_D (i.e. the sum of the nucleation time, t_N, and the growth time, t_G, as discussed in Section 3.2) is itself a stochastic variable distributed according to an intrinsic underlying cumulative probability function, $F(t_D)$:

$$t_D \sim F \tag{4.1}$$

This expression should be read as "t_D is distributed according to the cumulative distribution F". Let us define:

$$\mathbf{F} = [F_1, F_2, ..., F_q]^T \ : F_j = F(t_{D,j}),\ j = 1, 2, ..., q \tag{4.2}$$

The q measured detection times form the ordered vector $\mathbf{t^*_D}$:

$$\mathbf{t^*_D} = \left[t^*_{D,1}, t^*_{D,2}, ..., t^*_{D,q}\right]^T \ : 0 \leq t^*_{D,1} \leq t^*_{D,2} \leq ... \leq t^*_{D,q} < +\infty \tag{4.3}$$

through which the following empirical cumulative distribution function (eCDF, also simply "empirical distribution function") $F^*_Q(t_D)$ is defined as:

$$F^*_Q(t_D) = \frac{1}{Q} \sum_{j=1}^{q} \mathbb{1}_{[t^*_{D,j}, +\infty)}(t_D) \tag{4.4}$$

where $\mathbb{1}_{[a,+\infty)}(x)$ is the indicator function on the interval $[a, +\infty)$, i.e. equal to 1 if $a \leq x < \infty$, and 0 otherwise. The vector $\mathbf{t^*_D}$ constitutes a discrete, finite collection of samples of the stochastic variable t_D, whereas $F^*_Q(t_D)$ is an estimate of $F(t_D)$, obtained using a number of experiments $Q \in \mathbb{N}_{>0}$. It is also convenient to define:

$$\mathbf{F}^*_Q = \left[F^*_{Q,1}, F^*_{Q,2}, ..., F^*_{Q,q}\right]^T \ : F^*_{Q,j} = F^*_Q(t^*_{D,j}),\ j = 1, 2, ..., q \tag{4.5}$$

In the following, t_D, F, t^*_D, and F^*_Q without index will denote generic quantities, while the indexed quantities will correspond to specific experiments.

To better illustrate the quantities and concepts above, let us assume that nucleation follows Nucleation Mode 1, with constant nucleation rate J (see Section 2.3.1, Eqs. (2.15) and (2.16), and Section 2.3.2). The nucleation time $t_N =$ (see Eq. (3.1)) is then distributed according to the following cumulative probability function of t_D, as discussed in Section 3.2, case C.2:

$$P(t_N) = F(t_D) = 1 - \exp(-(t_D - t_G)JV) \tag{4.6}$$

where the size of the system is given by the volume, i.e. $W = V$. We have plotted in Figure 4.1 the function $F(t_D)$ (black line), and the two eCDFs $F^*_{Q_1}(t_D)$ and $F^*_{Q_2}(t_D)$, computed according to Equation 4.6 from two randomly sampled sets of detection times, with $Q_1 = 20$ (blue) and $Q_2 = 80$ (green) points, respectively (the two sets have been generated according to the inverse transform sampling algorithm, using the uniform random number generator implemented in Matlab$^{\text{TM}}$). It is evident that the quality of the estimation of $F(t_D)$ improves with an increasing number of experimental measurements, in agreement with both intuition and theory [116].

Note that, although the cumulative probability function and its empirical estimates used above are functions of t_D, the definitions in Eqs. (4.3)-(4.5), as well as the properties discussed in Section 4.2.2 below, are valid in general for any continuous stochastic variable [117], such as t_N. The choice of the detection time as reference variable is due to the fact that t_D is the quantity measured in experiments as well as the variable used to build the objective function for the estimation of the model parameters.

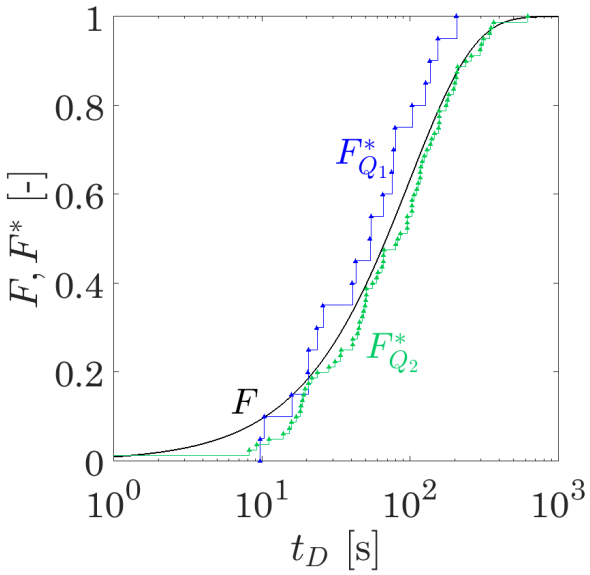


Figure 4.1: Plots of the cumulative probability function, $F(t_D)$, of Eq. (4.6) (with $J = 10,000$ #/(m^3 s), $t_G = 0$ s, and $V = 1$ mL), black, and of two eCDFs sampled from it, i.e. $F^*_{Q_1}$ with $Q_1 = 20$, blue, and $F^*_{Q_2}$ with $Q_2 = 80$, green. Note that in this case $q_1 = Q_1$ and $q_2 = Q_2$.

4.2.2 *Convergence Properties of the Empirical Distribution Function and Sample Representativity*

$\mathbf{F}^*_Q$ is an approximation of $\mathbf{F}$ and the distance between $\mathbf{F}^*_Q$ and $\mathbf{F}$ is measured by means of the Chebyschev distance, D_Q, also used in the Kolmogorov-Smirnov test:

$$D_Q = \left\|\mathbf{F}^*_Q - \mathbf{F}\right\|_\infty = \sup_{1\le j\le q} |F^*_j - F_j| \tag{4.7}$$

Using the Chebyschev distance, the Glivenko-Cantelli theorem [118] proves the uniform convergence of $\mathbf{F}^*_Q$ to $\mathbf{F}$, known in probability theory as almost sure convergence [117]:

$$\lim_{q\to\infty} D_Q \overset{a.s.}{=} 0 \tag{4.8}$$

Eq. (4.8) means that the sequence $\mathbf{F}^*_Q$ converges to $\mathbf{F}$ as $q \to \infty$ almost surely, i.e. with probability 1. Note that the Glivenko-Cantelli theorem in its general mathematical statement uses the limit $Q \to \infty$, since in theory one can assume that the experiments run as long as needed in order to let the system produce an observation. Here, though, to be consistent with the physical possibility that $q \leq Q$, it has been chosen to use q, since Q approaches infinity if q does, while the opposite is not true. Eq. (4.8) assures that, given a sufficient number of successful experiments, i.e. here experiments where crystals were detected, the probability built from experiments using Eqs. (4.3) and (4.4) represents a reliable approximation of the real probability, whatever this is. One can define the Chebyschev distance also between two empirical cumulative distributions, $F^*_{Q_1}$ and $F^*_{Q_2}$, with Q_1 and Q_2 points, respectively:

$$D_{Q_1,Q_2} \quad = \quad \left\| F^*_{Q_1}(t_D) - F^*_{Q_2}(t_D) \right\|_\infty \tag{4.9}$$

The Glivenko-Cantelli theorem, Eq. (4.8), guarantees almost sure convergence of $\mathbf{F}^*_{\mathbf{Q}}$ to $\mathbf{F}$, but does not provide any estimation of the rate of convergence. Thus, the theorem cannot be used to test (H.3), i.e. if and when the number of experiments Q is large enough to constitute a statistically representative sample of the parent population: the validity of hypothesis (H.3) it is essential to model the stochastic behaviour based on experimental evidence.

A minimum rate of convergence of Eq. (4.8) is estimated by the *Dvoretzky-Kiefer-Wolfowitz-Massart inequality* [119, 120] (DKWM), which places probabilistic bounds (DKWM-bounds) on the maximum value of D_Q, defined in Eq. (4.7), and reads:

$$\Pr\left(D_Q > \varepsilon\right) \leq 2e^{-2Q\varepsilon^2} \tag{4.10}$$

where ε (> 0) represents the upper bound of D_Q, Q is the number of experiments, and the left hand side of the equation represents the probability that D_Q is larger than ε. From Eq. (4.10), one can define the confidence level (also known simply as confidence), C:

$$C = \Pr\left(D_Q \leq \varepsilon\right) \geq 1 - 2e^{-2Q\varepsilon^2} \tag{4.11}$$

The meaning of the confidence level C, and its use to analyse the experiments, are discussed in more detail in Chapter 5. Equation (4.11) is used to compute a lower bound on the number of points Q to achieve a given accuracy ε, with a given confidence C, as:

$$Q \geq \frac{1}{2\varepsilon^2} \ln\left(\frac{2}{1-C}\right) \tag{4.12}$$

and a lower bound on the maximum value of D_Q for a given number of points Q, with a given a given confidence C, as:

$$\varepsilon \geq \sqrt{\frac{1}{2Q} \ln\left(\frac{2}{1-C}\right)} \tag{4.13}$$

Quite significantly, these results are independent of the underlying statistics and are applicable to any stochastic process with a continuous cumulative distribution probability, e.g. the Poisson process. Eq. (4.12) is useful for the design of experiments, namely to choose the number of measurements, Q. Equations (4.11) and (4.13) are useful to analyse the quality of a given set of Q measurements. One should also note the strong non-linear dependence of Q on C and ε, which leads to a very rapid increment of Q if high accuracy (small ε) and high confidence (large C) of $\mathbf{F}^*$ are required.

4.3 ESTIMATION OF MODEL PARAMETERS

The necessary input values to use the models discussed in Chapter 3 consist, for each experimental series in a specific solvent, of the vectors representing the sets of measured detection times and

corresponding eCDFs. Let us consider that N_s series have been performed in the same solvent at different supersaturations. For each series k, Q_k is the total number of experiments and q_k the number of experiments where crystals have been observed. To simplify the notation, in the following the vectors containing the experimental data associated to the sequence Q_k will be labelled simply as $\mathbf{t}^*_{\mathbf{D},k}$ and $\mathbf{F}^*_k$, with the index $k = 1, 2, ..., N_s$ related to the series of experiments from which the data have been measured; the j^{th} element of the vectors will be indicated as $t^*_{D,jk}$ and F^*_{jk}, respectively. The actual cumulative distribution function corresponding to the operating conditions at which $\mathbf{t}^*_{\mathbf{D},k}$ and $\mathbf{F}^*_k$ have been measured is indicated as F_k, with $\mathbf{F}_k = F(\mathbf{t}^*_{\mathbf{D},k})$. Clearly, the Chebyschev distance between $F^*_{Q_k}$ and F (Eq. (4.7)) will be similarly denoted as D_k, while that between, for instance, $F^*_{Q_1}$ and $F^*_{Q_2}$, as D_{12}.

To estimate the model parameters one minimises the distance between the experimental data and the corresponding simulated output of the model, i.e. the total error. The definition of the total error depends on the metric defining the distance. In this work, two different metrics will be used and a specific method is associated to each metric. The first metric, and its associated method, are discussed in Section 4.3.1, whilst the second one in Section 4.3.2. The two approaches lead in principle to different parameters and in the following it is discussed which method is best and why. All optimisations have been performed using the Matlab$^{\text{TM}}$ routine *fminsearch*, a derivative-free method.

4.3.1 *Euclidean Metric and Non-linear Least Squares Method*

The first metric uses, for each series k, the distance e_{jk} between each detection time forecast by the model and its experimental

counterpart; such a distance is defined as the difference between experimental and model values, relative to the geometric mean:

$$e_{jk} = \frac{t_{D,jk} - t^*_{D,jk}}{\sqrt{t_{D,jk} t^*_{D,jk}}} \tag{4.14}$$

The use of the weight, namely the reciprocal of the geometric mean in this case, guarantees that the fitting procedure is not biased in favour of longer detection times. The total error for each individual series k is thus computed as the euclidean norm E_k of the vector $\mathbf{e}_k$, consisting of the e_{jk} associated to the k^{th} series:

$$E_k = \|\mathbf{e}_k\|_2 = \sqrt{\sum_{j=1}^{q_k} e_{jk}^2} \tag{4.15}$$

Finally, the total error, E_s, is given by the euclidean norm of the vector $\mathbf{e}_s$, constituted of the errors E_k associated to the different series of experiments, divided by the number of series N_s:

$$E_s = \frac{1}{N_s} \|\mathbf{E}\|_2 = \frac{1}{N_s} \sqrt{\sum_{k=1}^{N_s} E_k^2} = \frac{1}{N_s} \sqrt{\sum_{k=1}^{N_s} \sum_{j=1}^{q_k} e_{jk}^2} \tag{4.16}$$

Minimising E defined by Eq. (4.16) implies solving a non-linear least squares problem, which entails establishing an iterative procedure, at each n^{th} step of which a set of parameters, $\boldsymbol{\omega}^{(n)}$, is computed as current guess, yielding the values $t_{D,jk}$ needed in Eq. (4.14). Thus, the method associated to the first metric is a non-linear least squares method (NLSM).

4.3.2 *Chebyschev Metric and p-value Maximisation*

The second metric is based on cumulative probabilities instead of detection times, and corresponds to the Chebyschev distance introduced in Eq. (4.7):

$$D_k = \|\mathbf{F}^*_k - \mathbf{F}_k\|_\infty \tag{4.17}$$

In this case, D_k represents the analogous of E_k for the Euclidean metric used in NLSM.

The definition of the total error to minimise, based on the DKWM inequality and its underlying statistical model, exploits the p-value concept, with which one measures the probability that a certain assumption - in this case the consistency between model and measurements - is not fulfilled. This new approach towards parameter estimation is based on the following statistical analysis.

Let us consider first, for the case with N_s experimental series at different supersaturation levels, all measured detection times. A two-step statistical analysis is carried out, where one checks first the ability of the model to describe each individual experimental series k, and then the ability of the model to describe the whole set of experimental series belonging to that solvent.

For a given set of estimated model parameters one can compare, for every experimental series k, the empirical cumulative distribution function and the corresponding distribution obtained with the model by calculating the distance, D_k using Eq.(4.17).

The first step of the statistical analysis is that of testing, at the level of the individual experimental series, a first hypothesis (first level null hypothesis), which states that the empirical cumulative distribution function is sampled from the corresponding distribution obtained with the model. To this aim one determines a p-value, p_k, for each series (each series consisting of Q_k experimental points). The p-value measures, in this case, the probability, under the statistical model associated to the DKWM inequality, that, by repeating the experiments, the Chebyschev distance associated to the repetition would be equal to or larger than its observed value, D_k [121]. Thus, this p-value corresponds to the probability defined by Eq. (4.10) and is given by:

$$p_k = 2\exp\left(-2Q_k D_k^2\right) \tag{4.18}$$

Note that when $D_k < \left(\ln 2/(2Q_k)\right)^{1/2}$ we set $p_k = 1$, to avoid a p-value larger than one. The smaller the p-value, the larger the sta-

tistical incompatibility of the empirical distribution $\mathbf{F}_k^*$, obtained from the experimental series k, and the corresponding distribution F, obtained with the model, assuming that all the underlying assumptions (including the physical assumptions incorporated in the model) used to calculate the p-value are fulfilled.

The second step can be viewed as a meta-analysis of the results obtained in the first step, namely of the set of N_s p-values, where N_s is the number of experimental series for the solvent under consideration. This is performed by applying the Fisher's method [122, 123] to combine the independent p_k values belonging to each solvent into a new p-value, p_s, characterising the whole series of experiments. The method relies on the fact that p_k is a stochastic variable with a uniform distribution [124], since the nucleation times (hence the detection times) are stochastic variables. As a consequence, the sum of the log-transform of the p_k values defines a new variable X_s^2 for each solvent, which follows a χ^2-distribution with $2N_s$ degrees of freedom. The new variable is defined as:

$$X_s^2 = -2 \sum_{k=1}^{N_s} \ln p_k \tag{4.19}$$

and the corresponding p-value is computed as:

$$p_s = 1 - \int_0^{X_s^2} \chi_{2N_s}^2(y) dy \tag{4.20}$$

Again, the smaller the value of p_s, the larger the statistical incompatibility of the whole set of experimental series in that solvent and the corresponding model, assuming that all the underlying physical and statistical assumptions behind the model and used to calculate the p-value are fulfilled. It is important to underline that, differently from E_s (Eq. (4.16), p_s belongs to the interval 0 to 1 and its numerical value provides a guidance to assess the goodness of description not only in relative terms, i.e. if the performance of the model in a system is better than in other systems, but also in

absolute terms, i.e. how much the model is incompatible with the data.

4.4 RESULTS OF THE VALIDATION

In this Section, the two models presented in Chapter 3 are validated against the experimental data reported in the literature for the cooling crystallisation from solution of two different model compounds, each following a different cooling profile.

The first model compound (Section 4.4.1) is paracetamol and the experimental data have been measured and reported by Kadam et al. [47]. Paracetamol has been crystallised from water using a linear cooling and each experiment has been run until crystal detection occurred. Model 1 is used to estimate the model parameters from the experiments using the non-linear least squares method. First, the quality of the fitting (which followed the non-linear least square method above), its predictive capabilities, and the correlation among the estimated parameters for the expressions of J and G are discussed; then, the estimated parameters are used in Model 2, analysing its similarities and differences with respect to Model 1. Simulations illustrating the effect of changing the cooling rate are shown and discussed, as well. All results are presented so that the reader is informed not only about average properties of the crystallisation process, but also about the statistic distribution.

The second model compound (Section 4.4.2) is p-aminobenzoic acid (p-ABA) crystallised from three different organic solvents, namely acetonitrile (ACN), 2-propanol (IPA), and ethyl-acetate (EtOAc), using a crash cooling profile. These experiments have been performed and reported by Sullivan et al. [85]. Note that because of the crash cooling profile, the experiments are actually isothermal: ideally, the system passes instantaneously from the saturation temperature at the given concentration to the final temperature, at which it remains isothermally. The experiments have been left running for 9 hours before being stopped. Crystals may

or may not have been detected during this time window, hence not all empirical cumulative distributions necessarily attain the value $F_q^* = 1$. The data of p-ABA, in all solvents, are described with Model 1 and the model parameters estimated through both the non-linear least squares method and the p-value maximisation method. The results obtained with the two estimation methods are compared and discussed.

The operating conditions, i.e. the temperature $T(t)$, the system size W, and the initial solute concentration c_0 are the input parameters.

4.4.1 *Case 1: Paracetamol in Water*

Of the four different initial concentrations used in the work of Kadam et al. [47], we selected two, namely 15 mg/mL and 47 mg/mL; for each of these two values, Kadam and co-workers reported experiments at four different volumes, 1, 500, 700, 900, and 1000 mL. The volume V, constant during crystallisation, was chosen to represent the system size, thus $W = V$. The investigations at different volumes aimed at understanding the behaviour of the system when scaling the process. Assuming an ideal thermostat, we modelled the linear cooling as:

$$T(t) = T_0 - \beta t \tag{4.21}$$

where the cooling rate β was set to the value reported in the original work, i.e. 0.5 K/min. We used an empirical temperature dependent solubility of paracetamol in water, given by the following polynomial expression [125]:

$$c_s(T) = \sum_{i=0}^{4} \Gamma_i (T - 273.15)^i \tag{4.22}$$

The parameters Γ_i were fitted on experimental data from the literature [47] and are reported in Table 4.1. The volume shape factor was $k_v = \pi/6$ and the crystal density $\rho_c = 1260$ kg/m^3.

Table 4.1: The coefficients of the solubility of paracetamol in water.

Parameter	Value
Γ_0 [g L^{-1}]	7.147
Γ_1 [g L^{-1} K^{-1}]	1.986 10^{-1}
Γ_2 [g L^{-1} K^{-2}]	5.048 10^{-3}
Γ_3 [g L^{-1} K^{-3}]	-1.273 10^{-4}
Γ_4 [g L^{-1} K^{-4}]	3.0188 10^{-6}

4.4.1.1 *Model 1: Parameter Estimation*

Model 1 has been applied to estimate the six parameters in the nucleation and growth rate expressions of Eqs. (3.25) and (3.26) by fitting the experimental detection times measured in the 1 mL crystalliser with the ones calculated with the model. Figure 4.2 shows the experimental cumulative distributions of detection times, the corresponding calculated distributions as well as the calculated distributions of nucleation times. The values of the six estimated parameters are reported as *Reference kinetics* in the second column of Table 4.2.

A few remarks are worth making. The model predicts a detection time which is between 0 s and about 400 s larger than the nucleation time. This indicates the importance of growth in these crystallisation experiments on paracetamol from aqueous solution. Note that the slower the nucleation, the lower the temperature, since T decreases in time due to cooling, and the higher the supersaturation at which it occurs. Higher supersaturation (that accelerate growth) prevails on lower temperature (that reduces it), thus leading to a decrease in growth times for larger values of t_N (from about 400 s, when nucleation is the fastest, to essentially 0 s, i.e., instantaneous growth to the detection size, when $t_N \approx 4000$ s).

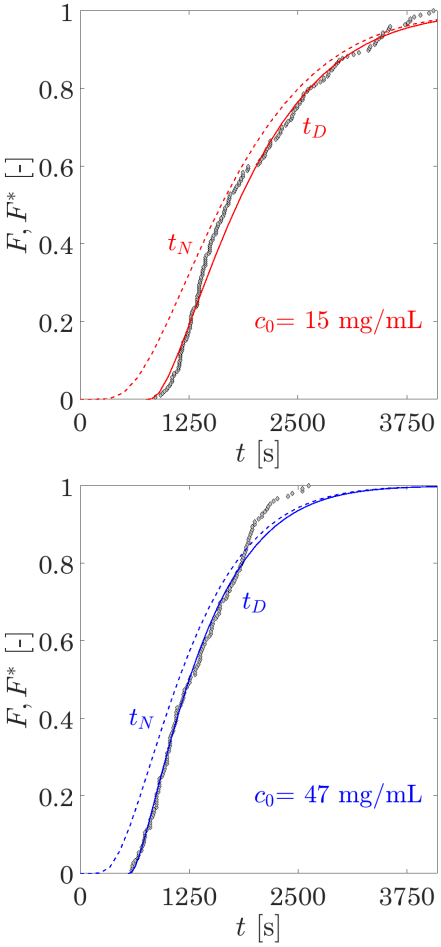


Figure 4.2: Results of Model 1 applied to the experimental data of Kadam et al. [47] in 1 mL vessel. Symbols: experimental data; solid lines: detection time probability curves; dashed lines: nucleation time probability curves (given by Eq. (2.20)).

Table 4.2: Kinetic parameters obtained from fitting the 1 mL-crystalliser data from Kadam et al. [47].

Parameters, $\boldsymbol{\omega}$	*Reference kinetics*	Faster growth	Slower growth
A_0 [#/(m^3 s)]	7.55×10^3	8.410×10^3	5.5×10^3
A_1 [K]	660	570	630
B [K^3]	7.35×10^5	2.15×10^6	1.3×10^5
K_0 [m s^{-1}]	1.67×10^{-4}	1.67×10^{-3}	8.35×10^{-5}
K_1 [K]	1.0×10^3	1.0×10^3	1.0×10^3
K_2 [K^2]	3.22×10^3	3.22×10^3	3.22×10^3

The agreement between measured and simulated detection times is rather good, not only at the initial concentrations of 15 mg/mL and 47 mg/mL, but also at 22 mg/mL and at 32 mg/mL (data not shown here). This is an important result, because such a good fitting has been obtained with temperature dependent rate models for nucleation and growth, where only six parameters have been estimated. The resulting model can reliably be used to predict crystallisation at different initial concentrations and for different cooling rates, i.e. something not possible with the original model used by of Kadam and co-workers and its twelve estimated parameters.

When looking more in detail, at $c_0 = 47$ mg/mL a deviation between experiments and simulations can be observed at values of the detection time beyond about 1800 s. In fact ,the experimental data for $t_D > 1800$ s seem to follow a different trend with respect to those for $t_D < 1800$ s. Considering that upon cooling supersaturation increases, a change of nucleation mechanism might explain the deviation, particularly considering that the supersaturation is as high as 1.8, observed when nucleation leading to $t_D = 1800$ s occurs. However, more experimental data would be needed to demonstrate this property thoroughly.

4.4.1.2 *Model 1: Correlation among Parameters*

The set of parameters indicated as *Reference kinetics* in Table 4.2 has been obtained by minimising the deviation between experiments and simulations using the non-linear least squares method. By analysing the optimisation results, one realises that the objective function of this problem is rather flat and that the minimum rather shallow. In fact, the estimated model parameters are correlated and very similar distributions of detection times, as the ones plotted in Figure 4.2, can be obtained by increasing the growth rate and reducing the nucleation rate, or vice versa.

To demonstrate this, we have selected two different sets of values for the growth rate parameters, and estimated the nucleation rate parameters to minimise the error between experiments and simulations. As reported in Table 4.2, the growth rate constant K_0 has been set to a value ten times larger (column 3, *Faster growth*) and to a value two times smaller (column 4, *Slower growth*) than that of the *Reference kinetics*, whereas K_1 and K_2 have been kept unchanged. After fitting, the nucleation rate parameters A_0, A_1, and B have attained the new values, also reported in Table 4.2.

Figure 4.3 illustrates the results, by showing the experimental detection time distributions at the two concentrations considered in the 1 mL crystalliser, together with the same distributions calculated with the three different sets of parameters; it is rather clear that the fitting accuracy is quite comparable. In the insets, the distributions of nucleation times for the two initial concentration levels (on the l.h.s. and on the r.h.s. the larger and the smaller concentration, respectively), as calculated with the three sets of parameters are also plotted. One can readily observe that indeed nucleation times are the longest for the set of *Faster growth* parameters, whereas the *Reference kinetics* yields intermediate values of t_N.

Thus, estimating parameters of nucleation and growth rate from statistics of detection times only is not possible, because the two groups of parameters are highly correlated, which is a mathema-

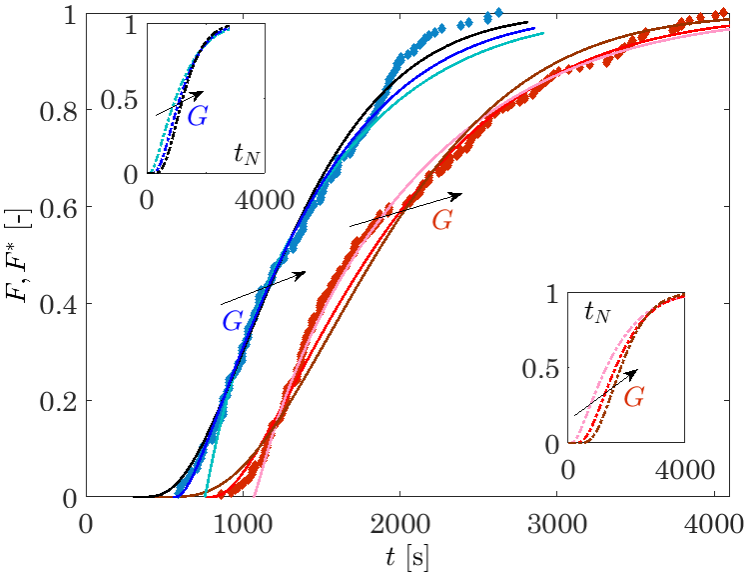


Figure 4.3: Results of Model 1 with three different sets of kinetic parameters (cfr. Table 4.2) and experimental data of Kadam et al. [47] in 1 mL. In the main figure, symbols: experimental data; solid lines: detection time probabilities. In the figure insets, dashed lines: nucleation time probability (given by Eq. (2.20)). For $c_0 = 15$ mg/mL and 47 mg/mL (red and blue shades, respectively), G increases from the lightest to the darkest colour.

tical property reflecting the physical reality that both nucleation and growth contribute to determine detection times. To overcome this problem, either independent growth measurements or additional, different pieces of information (e.g., the number of particles or the average particle size at some time t) are required, as indicated also by Nagy et al. [114].

Interestingly, though, the simulation results indicate that, if in 1 mL the different sets of kinetic parameters yield very similar dis-

tributions of detection times, in larger volumes the corresponding distributions are well distinct. This is illustrated in Fig. 4.4 for the case of $c_0 = 47$ mg/mL, where the cumulative distributions of nucleation and detection times obtained for the three sets of kinetic parameters are plotted and compared with the experimental distributions.

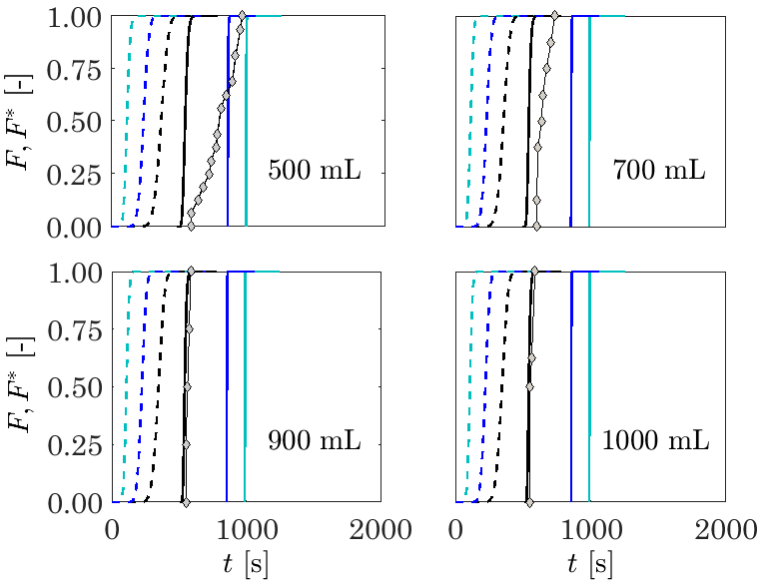


Figure 4.4: Results of Model 1 with 3 different sets of kinetic parameters (cfr. Table4.2) and experimental data, in 4 different volumes, for $c_0 = 47$ mg/mL. Symbols: experimental data of Kadam et al. [47]; solid lines: detection time probabilities (given by Model 1); dashed lines: nucleation time probability (given by Eq. (2.20)). G increases with colour from light blue to black.

It is worth noting that the set of parameters labelled as *Faster growth* performs better than the two other sets. Although one might be tempted to argue that the volume effect could be used to identify

the best among different sets of parameters, in reality one should be cautious and rule out first the possible causes of differences in measured detection times when scaling up the system volume, for instance differences in the detection technique or device used, which could exhibit different accuracy and precision.

4.4.1.3 *Model 1: Effect of the System Volume*

Figure 4.5 illustrates results for larger crystalliser volumes, namely 500, 700, 900, and 1,000 mL; each diagram refers to a specific volume as indicated, and shows the experimental cumulative distributions of detection times (for the two initial concentrations, namely 15 mg/mL and 47 mg/mL associated to larger and smaller detection times, respectively), and the corresponding distributions of both detection times and nucleation times calculated using Model 1 with the *Reference kinetics* parameters.

It is rather clear that the model is in general not able to describe the experiments, as it happens in the case of the model used by Kadam et al. to analyse the same data [47]. The model is unable to consistently predict the average detection times, and underestimates the variability. The experiments exhibit a much larger variability than predicted by the model. This behaviour might be due to a model mismatch, e.g. by not accounting for secondary nucleation or to the same reasons spelled out above with reference to the attempts made by Kadam and co-workers, hence making the extrapolation to larger volumes infeasible. We also tried to specifically fit the measurements in larger volumes with a set of *ad hoc* parameters different from the *Reference kinetics* in Table 4.2, without achieving satisfactory results though.

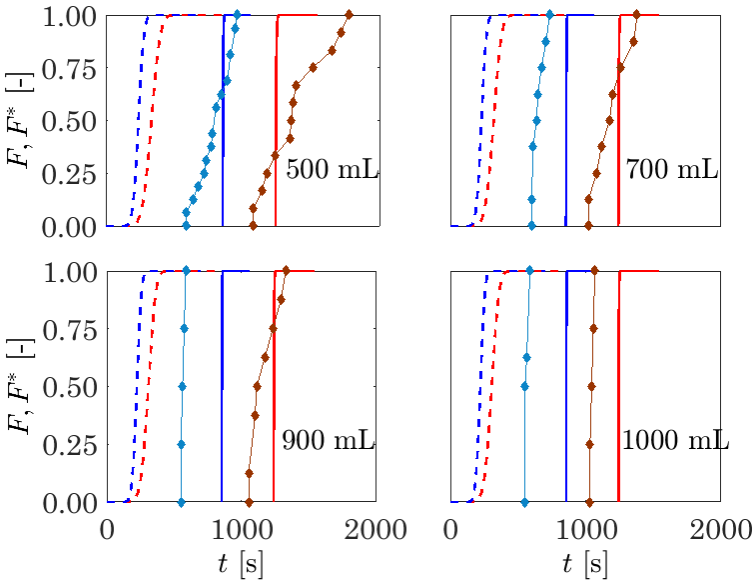


Figure 4.5: Results of Model 1 and experimental data of Kadam et al. [47], in 4 different volumes. Initial concentrations are 15 (red) and 47 (blue) mg/mL. Symbols: experimental data; solid lines: detection time probability curves; dashed lines: nucleation time probability curves (given by Eq. (2.20)).

4.4.1.4 *Model 2: Comparison with Model 1*

Let us now consider Model 2; for the sake of comparison with Model 1, rather than estimating a new set of kinetic parameters, the *Reference kinetics* set of parameters in Table 4.2 has been used to carry out simulations of nucleation and growth with both models, starting from the two reference initial concentration values, and for a number of system volumes between 1 mL and 1000 mL. Each simulation using Model 1 or Model 2 has been performed independently and, for each volume V at each initial concentration c_0, at

least one thousand simulations have been carried out in order to build reliable statistics.

The key differences between the results obtained with the two models can be appreciated by examining Figure 4.6, which plots the total number of particles present at the detection time, i.e. $N(t_D)$, for simulations carried out at different initial concentration, c_0, and different system volume, V. Let us focus on the results for $c_0 = 15$ mg/mL and $V = 100$ mL; similar considerations can be made for all the other cases. There are two sets of points: the aligned brown (darker) symbols on the l.h.s. obtained through Model 1, and the red (lighter) symbols, forming a cloud on the r.h.s. and calculated using Model 2.

In the case of Model 1, the stochasticity of nucleation yields a range of detection times and of number of particles as discussed above; however, for each stochastic value of the nucleation time, there are only one value of the detection time and one number of particles, because after the birth of the first nucleus the evolution of the system is deterministic.

In the case of Model 2, on the contrary, the same detection time may correspond to different numbers of particles, and vice versa; this is indeed the consequence of the stochastic nature of $\mathscr{H}_G$ due to the co-existence of stochastic nucleation and of deterministic growth after the formation of the first nucleus.

It is also worth observing that, as a consequence of the different structure of the two models, the average detection time and number of particles obtained with Model 2 are larger than those calculated using Model 1. On the one hand, the variability of the detection times decreases with increasing volume for both models, i.e. the self-exciting point process of Model 2 behaves in that respect like the inhomogeneous Poisson process of Model 1. On the other hand, the variability (or the standard deviation) of the Model 2 statistics of detection times is larger than that of Model 1.

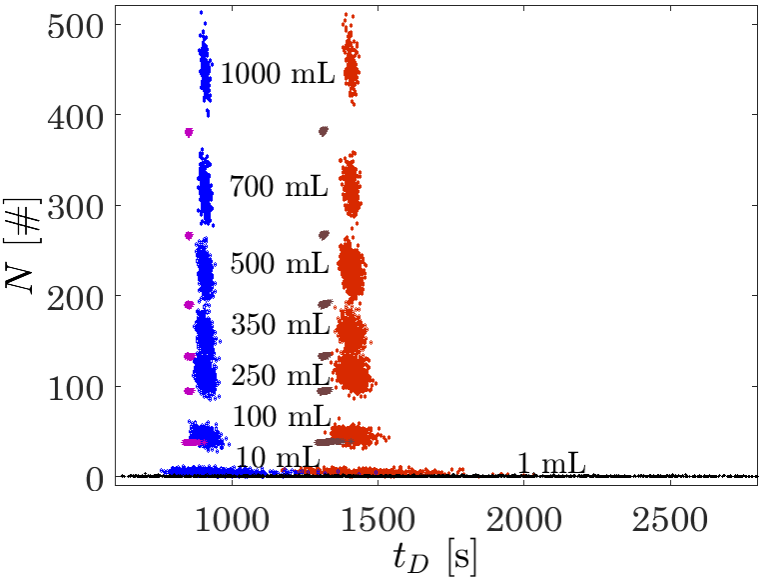


Figure 4.6: Results of simulations of Model 2 given as number of crystals, N, formed at the detection time, t_D, for different volumes. Two different initial concentrations, 15 and 47 mg/mL, are represented in red and blue, respectively, for volumes from 10 to 1000 mL. Results of simulations in 1 mL volume, largely overlapping for both initial concentrations, are all represented in black. Results of simulations of Model 1 for 100, 250, 350, 500, 700, and 1000 mL are indicated as violet and brown points (for c_0 = 47 mg/mL and c_0 = 15 mg/mL, respectively).

Considering the statistics of the number of particles, with increasing volume the variability of the number of particles calculated with Model 2 increases a lot. It can also be noted that for $V = 1$ mL, i.e., the volume on which the quantitative study of Kadam et al. [47] is based, only a very few number of particles are formed, typically less than five, which correspond to a very broad range of detection times (see Figure 4.2 for the cumulative probability

distribution of the detection times measured and calculated with Model 1 under the same conditions); from $V = 250$ mL on, the number of particles exceeds about one hundred. This result gives a quantitative confirmation and a visual illustration of how at different system volumes a mono-nuclear or a poly-nuclear mechanism may be the better way to describe nucleation [60]. It shows also that the mono-nuclear mechanism implicit in the Kadam et al. model and in our Model 1 is inadequate to describe volumes above 10 mL, in this case.

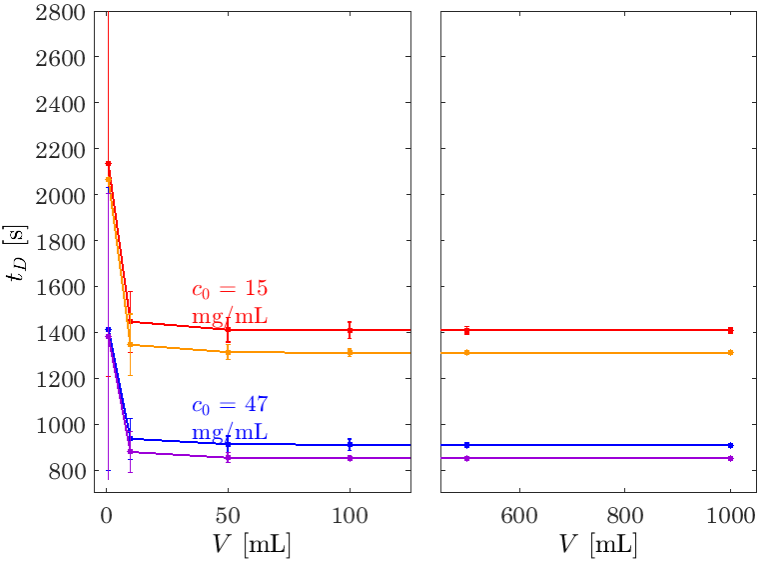


Figure 4.7: Results of simulations of Model 2, $c_0 = 47$ mg/mL (blue), and $c_0 = 15$ mg/mL (red), and Model 1, $c_0 = 47$ mg/mL (violet), and $c_0 = 15$ mg/mL (orange), given as average values of detection times, t_D. In the left figure, the volume ranges from 0 to 120 mL; in the right one, from 450 to 1100 mL. The error bars indicate the standard deviation of the corresponding average t_D value.

The same results are illustrated in a different manner for the same two initial concentration values in Figure 4.7, which shows as a function of the volume, V, the mean detection time, t_D, together with error bars, that are calculated by adding and subtracting from the mean t_D the value of the relevant standard deviation of the t_D distribution. It is worth noting that the two models behave very similarly when $V = 1$ mL, as expected since at that volume a few particles only are formed; since the first nucleus follows the same statistics for both models, with such a small number of particles to be formed, the differences between the results obtained with the two models can only be very small.

4.4.1.5 *Model 2: Sensitivity to Kinetic Parameters*

Simulations using Model 2 have been repeated for the *Faster growth* parameters in Table 4.2, i.e. the parameters that when using Model 1 exhibited the best accuracy in describing measurements at higher system volumes (see Fig.4.4). Figure 4.8 shows as a function of the volume, V, the mean detection time, t_D, together with error bars, calculated as in Figure 4.7. For each initial concentration value, two curves obtained using the two different sets of kinetic parameters are plotted. Moreover, the average detection times observed in the experiments of Kadam et al. [47] in crystallisers with $V = 500$ mL, 700 mL, 900 mL and 1000 mL are plotted.

As already observed using Model 1, the two sets of parameters yield very similar results when $V = 1$ mL, but diverge progressively as the volume increases. At both concentrations the *Faster growth* parameters yield lower values of the average detection time (about 500 s less and 300 s less at 15 mg/mL and at 47 mg/mL, respectively) and larger values of the standard deviation (which is typical for slower nucleation rates, as discussed in the Introduction when summarizing the main features of a homogeneous Poisson process) than the *Reference kinetics* parameters.

When comparing the simulation results and the experimental data in Figure 4.8, one notes that the experiments fall in the re-

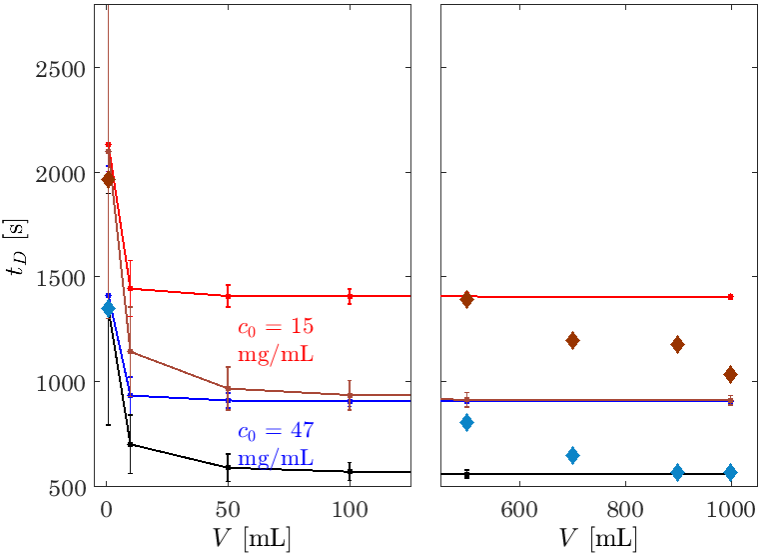


Figure 4.8: Comparison of simulations of Model 2, in 1 mL crystalliser, given as average detection times, t_D, and the associated standard deviation (error bars). Two different sets of parameters, *Reference kinetics* (blue and red, $c_0 = 47$ mg/mL and $c_0 = 15$ mg/mL, respectively) and Faster growth (black and brown, $c_0 = 47$ mg/mL and $c_0 = 15$ mg/mL, respectively) have been used (see Table 4.2). The volume ranges from 0 to 120 mL (left box) and from 450 to 1100 mL (right box). The diamond symbols indicate the experimental data of Kadam and co-workers [47] at the corresponding volumes. The observed overlapping of the brown and blue line is a coincidence of this particular system, not a specific feature of the model.

gion delimited by the two curves calculated with the *Reference kinetics* and with the *Faster growth* parameters; the former set works better in the 500 mL crystalliser, whereas the latter in the 900 mL and 1000 mL crystallisers. Even though the standard de-

viation predicted by Model 2 is larger than that by Model 1, it is still much smaller than the experimental one.

To summarise, though Model 2 is a clear improvement with respect to Model 1 in terms of physical consistency, based on the experimental results of Kadam et al. in large crystallisers only, it is not possible to demonstrate that Model 2 can capture completely the effects due to the change from very small volumes to larger ones. Clearly, to this aim more measurements, following a well designed experimental protocol, must to be performed.

4.4.1.6 *Model 2: Effect of Cooling Rate*

Being intrinsically temperature dependent, both Model 1 and Model 2 can be used to study the effect of the cooling rate on the nucleation stochasticity and on the metastable zone width (MZW), $\Delta T(t_D)$, i.e. difference between the temperature at time zero, when the solution is saturated, and at the detection time.

We have investigated such an effect using Model 2 and illustrated it in Figure 4.9, where the average values of ΔT with its variability (error bars) are plotted vs. the system volume. These have been calculated using the parameters corresponding to the *Reference kinetics* for the initial concentration of 47 mg/mL and three different cooling rates, namely $\beta = 0.25$ K/min, 0.5 K/min (the value used in all simulations reported so far), and 1 K/min; therefore, the data corresponding to $\beta = 0.5$ K/min have already been reported in terms of average detection time both in Figure 4.7 and in Figure 4.8. Such average detection times at large volumes are reported also in Figure 4.9 beside the corresponding $\Delta T(t_D)$ vs. V curve.

As expected based on physical intuition, larger cooling rates lead to larger values of the MZW (and of its variability) and correspondingly smaller values of t_D. This is due to the fact that the MZW is determined by the competition between a supersaturation generating mechanism, i.e. cooling, and a supersaturation depleting process, namely particle formation and growth. Changing the cool-

ing rate affects the former but not the latter, thus yielding larger values of MZW, the reduction of the detection time being a mere consequence of the fact that the same temperature change requires less time when the cooling rate is larger.

The time-dependent model can also describe the effect of nucleation stochasticity on the MZW in very small volumes, namely that the MZW and its variability are larger in very small volumes than in large volumes. In volumes beyond a system specific threshold value, the MZW reaches an asymptotic value for each cooling rate, while its variability in practice vanishes.

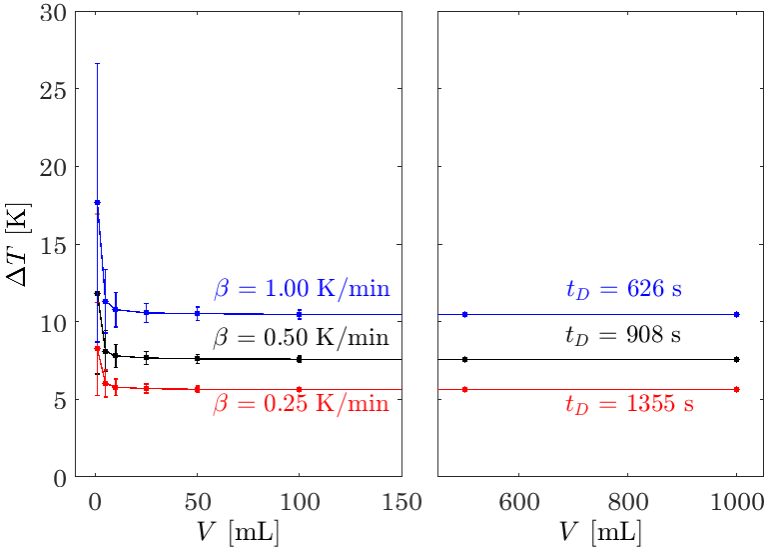


Figure 4.9: Average Metastable Zone Width, ΔT, simulated with Model 2 at different volumes, for $c_0 = 47$ mg/mL, with three different cooling rates, β. In the left figure, the volume ranges from 0 to 120 mL; in the right one, from 450 to 1100 mL (where the asymptotic values of the average t_D for each cooling rate are indicated). The error bars represent the standard deviation at the corresponding average ΔT value.

4.4.2 *Case 2: p-Aminobenzoic Acid in Organic Solvents*

For the experiments with p-ABA performed by Sullivan et al. [85] in ACN, IPA, and EtOAc, we have assumed an ideal thermostat, capable of regulating instantaneously the system temperature from the initial temperature to the set-point temperature, T_{set}:

$$T(t) = \begin{cases} T(t) = T_0 & \forall\ t \leq t_0 \\ T(t) = T_{\text{set}} & \forall\ t > t_0 \end{cases} \tag{4.23}$$

where we have assumed $t_0 = 0$. Note that, differently from the experiments of paracetamol in water by Kadam et al. [47] discussed above, these sets of experiments ran cycles with an isothermal waiting time of 9 hours; when the waiting time was over, a new cycle started, hence terminating the previous experiment. During each experiment, thus, crystals may or may not have been formed and detected; for this reason, not all eCDF obtained by Sullivan et al. [85] attain the value $F^* = 1$ at the largest detection time measured in each series of experiments. Since the system was operated at isothermal conditions, the process intensity is described by Eq. (2.16), i.e. $\Lambda(t) = WJt$, K_0 and K_1 in Eq. (3.26) can be lumped into a single parameter, i.e., $K = K_0 \exp(-K_1/T)$, while A_0 and A_1 in Eq. (3.25) can be combined into $A = A_0 \exp(-A_1/T)$. Thus, the vector of model parameters is $\boldsymbol{\omega} = [A, B, K, K_2]$. We express the system size as mass of solvent, $W = M$, thus the units of J are [# /(kg s)]; Sullivan et al. [85] carried out the original experiments with a volume of 1.5 (in ACN and IPA) or 1.8 mL (EtOAc), at a stirring rate of 700 rpm.

The solubility of p-ABA $c_s(T)$ could be expressed in terms of T through:

$$c_s(T) = \Gamma_0 \exp\left(-\frac{\Gamma_1}{T}\right) \tag{4.24}$$

with Γ_0, Γ_1 being two constants; here c_s and Γ_0 are in gram of solute per gram of solvent [g/g], while T and Γ_1 in [K]. The values of the

solubility parameters, $(\Gamma_0; \Gamma_1)$, have been obtained from fitting Eq. (4.24) to the data found in Svard et al. [86]; they are $(328; 2510)$, $(224; 2380)$, and $(3.1; 1078)$, for ACN, IPA, and EtOAc, respectively.

Different supersaturations were investigated, thus producing 6 series of experiments p-ABA crystallised from ACN, 8 from IPA, and 7 from EtOAc. Since the system is isothermal and Model 1 is used to describe it, $t_N = t_D - t_G$, with t_G costant for each supersaturation, (see Chapter 3, Section 3.2, case C.2), t_N is distributed as t_D, i.e.

$$P(t_N) = F(t_D) = F(t_N + t_G) \tag{4.25}$$

hence, the eCDF of detection times is representative also of the eCDF of nucleation times.

The non-linear least squares method is discussed first and the p-value maximisation method afterwards. Before proceeding with discussing the parameter estimation procedure, it is worth making three points. First, p-ABA exhibits rather fast growth kinetics, which implies very short growth times and detection occurring very soon after nucleation. Under these conditions and for isothermal experiments (not linear cooling experiments, as analysed in our previous contribution [36]) the nucleation rate and the growth rate parameters are not correlated, as we could demonstrate by estimating the nucleation rate parameters repeatedly for different assigned values of the growth rate parameters; by doing so, we have always obtained very similar nucleation rate parameters. Second, it is worth keeping in mind that the nucleation rate parameters are strongly dependent on the threshold value used in the detection condition of Eq. (3.14). We have chosen the value for such threshold in accordance with earlier papers [85, 106], which represents an arbitrary, but sensible choice. Finally, we have estimated confidence intervals for the parameters according to a heuristic procedure, i.e. we have repeated the estimation of the parameters from randomised initial guesses and looked for the minimum of the errors, using a genetic algorithm. The minimum found by the opti-

miser, i.e. the genetic algorithm, was assumed to be also the global minimum; the parameters from the local minima whose associated objective functions had values differing not more than 1% from the objective function from the global minimum were used to estimate the standard deviation, which was taken as confidence interval for the parameters.

4.4.2.1 *NLS Method*

The iterative procedure for the NLS method at the n^{th} step uses the current guess of the parameters, $\boldsymbol{\omega}^{(n)}$; it is also assumed that the eCDF exhibits the exact values of the cumulative probability of the detection times, i.e.$F^*_{jk} = F_{jk}$. First, the experimental F^*_{jk} value associated to the measured $t^*_{D,jk}$ value has been used to compute $t_{N,jk}$, since with Model 1 $F_{jk} = P_{jk}$, with the isothermal version of Eq. (2.16), thus obtaining $t_{N,jk} = -(\ln(1 - F^*_{jk})/(WJ_{p,k}))$, where $J_{p,k}$ is the nucleation rate at the conditions (supersaturation and temperature) of the series k. Then, Eqs. (3.8) and (3.12) are integrated from $t_{N,jk}$ to when the detection condition (Eq. (3.12)) is fulfilled: this establishes the time corresponding to the requested value $t_{D,jk}$.

All experiments in the same solvent have been fitted together, thus obtaining three sets of parameters, reported in Table 4.3. The corresponding errors for each series of experiments, at the corresponding value of supersaturation, are reported for all twenty-one series in Table 4.4.

The comparison between experimental data and estimated results is illustrated in Figure 4.10, where the cumulative distribution functions obtained in each series of experiments, i.e. the points $(\mathbf{t}^*_{\mathbf{D}}, \mathbf{F}^*)$, are compared with the curves calculated using the model, i.e. $(\mathbf{t}_{\mathbf{D}}, \mathbf{F})$. In each figure, the supersaturation level increases from lower right to upper left, i.e. going from longer detection times to shorter ones.

Table 4.3: Non-linear least square methods. Estimated values of the model parameters for p-ABA in the three solvent investigated.

Parameters, $\boldsymbol{\omega}$	ACN	IPA	EtOAc
A [# /(kg s)]	0.94 ± 0.1	0.44 ± 0.1	0.42 ± 0.05
$B \times 10^{-5}$ [K^3]	2.81 ± 0.1	4.45 ± 0.2	4.11 ± 0.2
$K \times 10^{7}$ [m s^{-1}]	1.78 ± 0.5	9.67 ± 2	3.37 ± 0.7
$K_2 \times 10^{-3}$ [K^2]	8.61 ± 0.2	1.58 ± 0.1	0.92 ± 0.2

Let us consider the experimental points first. It is apparent that some of the experimental curves exhibit a rather regular behaviour, e.g. those in ACN in general and a few in IPA and in EtOAc. On the contrary, a few curves have a shape rather different from the exponential Poisson distribution, such as in IPA at the highest supersaturation. In other cases the experimental behaviour appears to be counter-intuitive, e.g. at the fourth and fifth supersaturation level in IPA the two curves intersect with each other, or in EtOAc the third, fourth, fifth and sixth curves are almost indistinguishable in certain detection time ranges. Thus summarising, the experimental data themselves seem to exhibit some intrinsic inaccuracy.

Let us now evaluate the quality of the fitting of the model description of the experimental curves. First of all, the comparison between experiments and simulations highlights the inconsistency of some of the measurements, e.g. those at the lowest supersaturations in IPA and EtOAc. It is also striking that the error values reported in Table 4.4 are the worst for the experimental data that appear more problematic on visual inspection (see the discussion in the previous paragraph). It is worth underlining that when the calculated cumulative distribution and the measured one differ, this might be due to the wrong choice of the statistics (see for instance

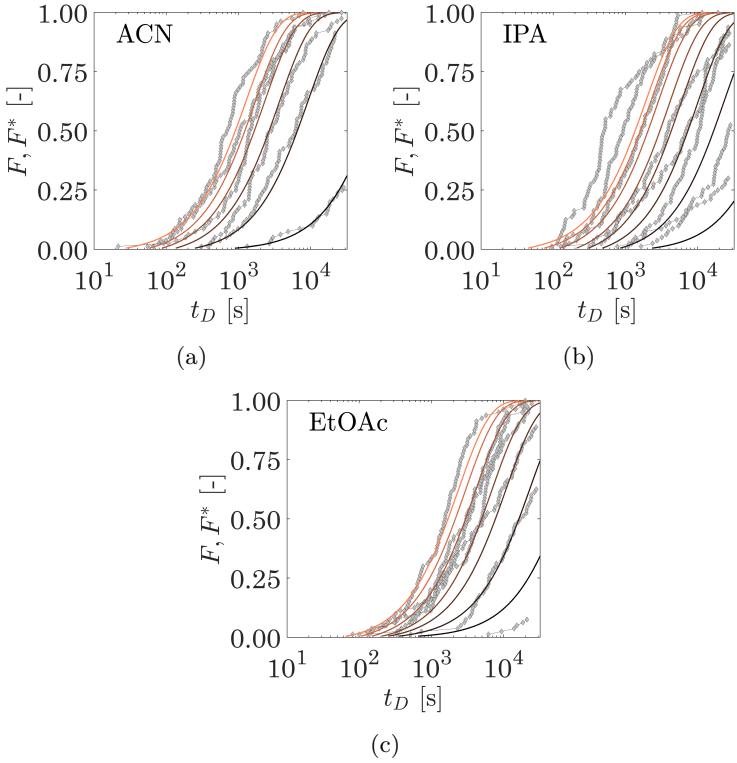


Figure 4.10: Comparison between experimental measurements and simulation results calculated using the parameters estimated with the non-linear least squares method. The experimental points of Sullivan et al. [85] are represented by grey diamonds, while the calculated distributions are plotted as continuous lines, from black to orange, indicating increasing supersaturation. The abscissa is in logarithmic scale.

the case of the highest supersaturation in IPA). However, when the distance between experimental curves obtained at different supersaturation is completely different from what obtained with the

Table 4.4: Non-linear least squares method. Operating conditions, total error for each individual series (from Eq.(4.15)) and total error for each solvent (from Eq.(4.16)).

Series	ACN		IPA		EtOAc	
k	S	E_k	S	E_k	S	E_k
1	1.05	0.83	1.07	4.78	1.07	3.52
2	1.07	1.21	1.09	2.95	1.08	1.27
3	1.09	2.08	1.10	2.69	1.10	4.59
4	1.11	0.96	1.12	3.81	1.11	1.57
5	1.14	2.04	1.14	3.87	1.14	1.45
6	1.17	1.51	1.18	1.41	1.17	2.55
7	-	-	1.22	3.19	1.18	1.51
8	-	-	1.26	4.62	-	-
E_s		0.62		1.26		1.00

model, this might be due to a wrong dependence on supersaturation. While the models are able to describe the supersaturation dependence of the measurements in ACN, they are not able to do so in all cases for the experiments in the other two solvents. In each solvent, the quality of the fitting differs at different supersaturation, without exhibiting any clear trend, e.g. that lower supersaturation leads to larger errors. The quality of the data in ACN appears, both on visual inspection and upon analysis of the errors, to be better than the one of those in EtOAc, which is in turn better than the quality of the data in IPA. Nevertheless, it is hard to judge whether the difficulty we experience in interpreting the data in IPA and EtOAc is due to the data or to the model.

4.4.2.2 *Statistical Analysis of NLSM*

Before discussing the second method, it is useful to evaluate the NLSM results in light of the representativity issue outlined in Section 4.2.2. The least-square estimation is based on an error minimisation where experimental measurements are assumed to be affected only by Gaussian error (if the error were negligible, the measurements would be accurate), whereas the measured empirical cumulative distribution function intrinsically deviates from the underlying distribution, as the DKWM inequality clearly states. For each solvent, Figure 4.11 illustrates two series of experiments at two different supersaturation levels, by plotting the eCDF, F^* (symbols), and its corresponding distribution calculated with the model, F (solid black line); the two chosen series correspond to the smallest (upper part of the figure) and to the largest (lower part of the figure) model error (see Table 4.4). In each diagram of Figure 4.11, three uncertainty regions around the experimental points are shaded, namely those obtained by applying the DKWM inequality for the values of C equal to 0.10 (the most colourful and smallest region), 0.70 (the intermediate region), and 0.99 (the largest grey region). One sees immediately that the distance between the experimental and the calculated distributions differs in the six diagrams, and that the calculated distributions are contained in different uncertainty regions. In the diagrams on the left, the calculated solid line is contained in the smallest uncertainty region (C=0.10), while, in the diagrams on the right, F is (barely) contained in the smallest uncertainty region only in the case of ACN, whereas it reaches the largest uncertainty region (C=0.99) in the case of EtOAc, and it goes even beyond it in the case of IPA. Despite intuitively suggesting that the model describes the case of ACN much better than the other two, could the deviation between experiments and the model in the case of IPA simply be due to the unavoidable stochastic effects? The p-value method looks at this question from a different perspective, as discussed in Section 4.3.

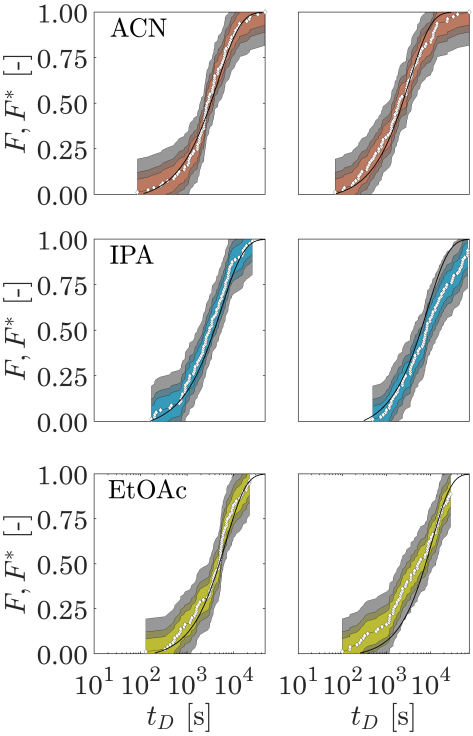


Figure 4.11: Comparison between the model (solid black lines) and the experimental points (grey symbols), accounting for the uncertainty (shaded areas) due to the estimation of $\mathbf{F}^*$. The selected datasets (4 - 5, 6 - 5, and 4 - 3 for ACN, IPA, and EtOAc, respectively) represent, for each solvent, the best (l.h.s. figure) and the worst (r.h.s. figure) fittings. The uncertainties are computed with Eq. (4.13) for $C = 0.10, 0.70, 0.99$ (from red, blue, and yellow to grey colour, for the three solvents). The model fitting lies within the boundaries for ACN, but not for IPA, for which part of the model curves still lies outside even for $C = 0.99$.

4.4.2.3 *p-Value Method*

The maximisation of the p-value of each solvent, p_s, from Eq. (4.20), yields alternative estimates of the model parameters, whose values are reported in Table 4.5. An iterative procedure has been set up for this method as well. At each n^{th} step a set of model parameters, $\boldsymbol{\omega}^{(n)}$, was available as current guess; with these, first the nucleation cumulative probability was calculated using $\Lambda(t) = WJt$, thus obtaining $P_k(t_{N,j}) = 1 - \exp(-WJ_k t_{N,j})$, where J_k is the nucleation rate at the values of T and S prevailing in series k, and $t_{N,j}$ represents the j^{th} nucleation time of the series. Then, from $t_{N,jk}$, the corresponding detection times, i.e. $t_{D,jk}$, are calculated by integrating the model equations (3.8)-(3.12); to those values the probability, $F_k(t)$, calculated above is associated. Finally, one uses the Chebyschev distance to determine the difference, D_k, between the calculated probabilities F_{jk} and the measured values F^*_{jk}, and thus p_k from Eq. (4.18). The p_k values calculated for each series in each solvent, on which the estimation of these parameters has been based, are reported in Table 4.6, together with the three p_s values, obtained with Eqs. (4.19) and (4.20). The estimated parameters in Table 4.5 can be compared with those in Table 4.3. The nucleation rate parameters A and B change only slightly for the two estimation methods in case of ACN and EtOAc, but significantly for the case of IPA. Thus, by using a different optimisation criterion based on the value of p_s, a modest quantitative change in the model parameters is observed for the systems already described well by the NLSM, i.e. nucleation in ACN, but no qualitative change in the accuracy of the model description. On the contrary, a quantitative change is observed for the system captured worst by the model with the parameter estimated by the NLSM„ i.e. IPA, but, also in this case, no qualitative change. With the parameters obtained with the p-value method, one can draw a new set of plots similar to those reported in Figure 4.10, to which the new plots (not reported here) are very similar.

Table 4.5: p-value maximisation method. Estimated values of the model parameters for the three solvents.

Parameters, $\boldsymbol{\omega}$	ACN	IPA	EtOAc
A [# /(kg s)]	1.05 ± 0.1	0.76 ± 0.1	0.40 ± 0.1
$B \times 10^{-5}$ [K^3]	2.79 ± 0.1	6.60 ± 0.1	4.47 ± 0.05
$K \times 10^{6}$ [m/s]	1.41 ± 1	2.33 ± 0.7	3.59 ± 1
$K_2 \times 10^{-3}$ [K^2]	6.90 ± 3	17.8 ± 5	14.7 ± 4

Based on the calculated p_k and p_s values in Table 4.6, several observations can be made, and a few conclusions can be drawn. All p_k values for ACN have the same order of magnitude and are larger than 0.2, thus above the reference value of 0.05 which is sometimes used as a lower threshold (see a recent paper for an in depth discussion about the meaning of the p-value and of the controversy about the 0.05 threshold [121]). On the contrary, three of the p_k values for EtOAc and four for IPA are below such reference threshold; in the case of IPA, the p_k values span a broad range of orders of magnitude. As a consequence, the calculated p_s value is rather large for ACN, namely 0.61, smaller for EtOAc, namely 0.009, i.e. below the conventional threshold, and very small for IPA, namely $4.5{\times}10^{-4}$, i.e. well below the conventional threshold. Independently of the comparison with any specific threshold value and consistently with the results illustrated in Figure 4.10, it is fair to conclude that, according to the p_s values analysis, the model represents best the system in the case of the nucleation of p-ABA in ACN, poorly in that of EtOAc, and very poorly in the case of IPA.

There can be several reasons for such difference in the capability of the same model (with different parameters of course) to describe the nucleation of the same substance in different solvents:

Table 4.6: p-value maximisation method. Operating conditions, p_k-value for each individual series (from Eq. (4.18)) and p_s-value associated to each solvent (from Eq. (4.20)).

Series	ACN		IPA		EtOAc	
k	S	p_k	S	p_k	S	p_k
1	1.05	1.00	1.07	8.96×10^{-5}	1.07	0.02
2	1.07	0.53	1.09	0.88	1.08	0.76
3	1.09	0.37	1.10	0.90	1.10	1.30×10^{-2}
4	1.11	0.75	1.12	2.92×10^{-4}	1.11	0.39
5	1.14	0.22	1.14	1.85×10^{-3}	1.14	0.40
6	1.17	0.22	1.18	0.87	1.17	2.87×10^{-2}
7		-	1.22	0.64	1.18	0.42
8		-	1.26	3.68×10^{-3}		-
p_s		0.61		4.50×10^{-4}		0.009

not having run the experiments, we can only list possible reasons and indicate the need for further investigations. First, the physical model used to describe the experiments might be inaccurate in the case of IPA and EtOAc, possibly not in how nucleation occurs, but in how crystals evolve from nucleation to detection, e.g. undergoing a polymorphic transformation. Secondly, the statistical model might be inaccurate, i.e. another model rather the Poisson process should be used (see Section 2.3.1). However, this hypothesis seems unlikely, as we discussed briefly in Section 3.2. Assuming no history dependence of the data analysed, we have fitted the experiments using for the nucleation time statistics a Weibull distribution, which reads:

$$P(t) = 1 - \exp\left(-\left(JWt\right)^{\eta}\right) \tag{4.26}$$

where the parameter η is strictly positive (if $\eta = 1$ the Weibull distribution becomes identical to the exponential distribution of Eq. (2.15) with constant λ). Despite the additional parameter, the fitting did not differ significantly from those obtained assuming a Poisson statistics; in fact, δ was estimated to be close to 1 from the fitting. That the exponential distribution does not provide a worse description of the data would suggest that the chosen statistic is not the reason for the discrepancy between the data and the model. Finally, the experiments might have been carried out in a region of high sensitivity due to the choice of the operating conditions, thus leading to an intrinsic lack of reproducibility, which is discussed in more detail in Section 4.5.

4.5 SENSITIVITY ANALYSIS

The question of reproducibility and reliability of experiments is further explored in this Section, under the assumption that the statistical model describing the process and the system is correct. To guarantee experimental reproducibility, the operating conditions, as well as the process variables, such as system size, temperature, and concentration (or supersaturation) must be accurately controlled.

The non-linear character of the mechanisms involved in crystallisation rises the question of how much the output of a crystallisation process changes in response to small, experimentally non-measurable variations of the operating conditions, i.e. what the sensitivity of the system output is with respect to a specific input variable [126]. Mathematically, the first order (local) sensitivity, s_x^y, and the relative sensitivity, $\tilde{\sigma}_x^y$, of a function y of m variables $(x_1, x_2, ..., x_m)$ are defined as:

$$s_x^y = \frac{\partial y}{\partial x} \tag{4.27}$$

$$\tilde{\sigma}_x^y = \frac{\partial \ln y}{\partial \ln x} = \frac{x}{y} s_x^y \tag{4.28}$$

where for the sake of simplicity x indicates any of the generic independent variables, x_j. For small deviations from a reference state, i.e. from y_r and x_r, keeping everything else constant, one obtains:

$$dy = s_x^y dx \tag{4.29}$$

$$\frac{dy}{y_r} = \tilde{\sigma}_x^y \frac{dx}{x_r} \tag{4.30}$$

A system output y is insensitive with respect to a variable x if it is not affected at all by its changes, i.e. $s_x^y = 0$; it is linearly sensitive if its sensitivity is independent of x, i.e. $s_x^y = \text{const}$; it is non-linearly sensitive in any other case, i.e. when $s_x^y = s_x^y(x)$.

For both Model 1 and Model 2 described in Chapter 3, the output consists in the nucleation probability, i.e. $y = P$, and in the detection probability, i.e. $y = F$. The input x can be any among temperature, T, concentration, c, system size, W, or detection time, t_D. The method can be generally applied to either Model 1 or Model 2: the main difference in the sensitivity analysis between the two models consists in the fact that, in the latter, the nucleation probability and the detection probability are not related by a one-to-one mapping, since the mapping $\mathscr{H}_G$ is stochastic, while in the former one assumes $F = P$, given the relation between nucleation time and detection time (Eq. (3.1)). For the sake of simplicity, the sensitivity analysis discussed in this Section is applied to isothermal crystallisation described by Model 1, and the results obtained from the analysis of p-ABA data will be used; the results from the data of paracetamol crystallised from water could also be equivalently used. All qualitative conclusions and most results, though, are directly applicable also to Model 2. Considering isothermal conditions, λ is constant (see Eqs. (2.19) and (3.25)); for the analysis, it is convenient to use Eq. (3.1) and write the nucleation probability P as:

$$P(z) = 1 - \exp\left(-zJW\right) \tag{4.31}$$

where $z = t_N = t_D - t_G$. Applying Eqs. (4.27) and (4.28) to Eq.(4.31) gives:

$$s_x^P = \frac{dP}{d\Lambda}\frac{\partial \Lambda}{\partial x} = e^{-\Lambda} s_x^\Lambda \tag{4.32}$$

$$\tilde{\sigma}_x^P = \frac{\Lambda}{e^\Lambda - 1}\tilde{\sigma}_x^\Lambda \tag{4.33}$$

with $\Lambda = zJW$, with J depending on T and c, but neither on W nor on z. Hence, one obtains:

$$\tilde{\sigma}_x^\Lambda = \tilde{\sigma}_x^J + \tilde{\sigma}_x^w + \tilde{\sigma}_x^z \tag{4.34}$$

and from Eq. (4.33):

$$\tilde{\sigma}_x^P = \frac{\Lambda}{e^\Lambda - 1}\left(\tilde{\sigma}_x^J + \tilde{\sigma}_x^w + \tilde{\sigma}_x^z\right) \tag{4.35}$$

Note that the factor $\Lambda/(e^\Lambda - 1)$ is monotonically decreasing from 1, when $\Lambda \to 0$, to 0, when $\Lambda \to \infty$. Therefore, the relative sensitive of P with respect to any x is at most as large as the largest among the corresponding relative sensitivities of J, W and z. Note also that if $x = W$, then $\sigma^W_{x=W} = 1$, whereas $\tilde{\sigma}^W_{x\neq W} = 0$ otherwise; consequently, variations of the system size are not responsible for any non-linear sensitivity of P.

As far as the remaining two terms are concerned, i.e. $\tilde{\sigma}_x^\tau$ and $\tilde{\sigma}_x^J$, we shall consider two cases. In the first, simple case, we assume that the growth time is negligible compared to the nucleation time, i.e. $t_D \approx t_N = z$, hence $\tilde{\sigma}^z_{t_D} = 1$ and $\tilde{\sigma}^z_{x\neq t_D} = 0$. Therefore, any non-linear sensitivity of P stems from that of the nucleation rate, J, with respect to T or c, i.e. from $\tilde{\sigma}_c^J$ and $\tilde{\sigma}_T^J$. In the second, more general case, the growth time is not negligible, and $z = t_N = t_D - t_G$; thus, the relative sensitivity of z depends also on the sensitivity of the growth rate G, $\tilde{\sigma}_x^G$, based on the detection condition expressed by Eq. (3.12). As a consequence, the relative sensitivity of P may in this case be non-linear and large if and only if the relative sensitivities of J or G (obviously with respect only to temperature,

T, and concentration, c, as neither system size nor detection time affect J and G) are themselves non-linear and large.

Deriving explicitly $\tilde{\sigma}_x^J$ and $\tilde{\sigma}_x^G$ from Eqs.(3.26) and (3.25), with x equal to either c or T, one obtains:

$$\tilde{\sigma}_c^J = 1 + \frac{2B}{T^3 \ln^3 \frac{c}{c_s(T)}} \tag{4.36}$$

$$\tilde{\sigma}_T^J = \frac{A_1}{T} + \frac{3B}{T^3 \ln^2 \frac{c}{c_s(T)}} + \tilde{\sigma}_c^J \frac{q_1}{T} \tag{4.37}$$

$$\tilde{\sigma}_c^G = \frac{2}{3} \frac{\frac{c}{c_s(T)}}{\frac{c}{c_s(T)} - 1} + \frac{1}{6 \ln \frac{c}{c_s(T)}} + \frac{K_2}{T^2 \ln^2 \frac{c}{c_s(T)}} \tag{4.38}$$

$$\tilde{\sigma}_T^G = \frac{K_1}{T} + \frac{2K_2}{T^2 \ln \frac{c}{c_s(T)}} + \tilde{\sigma}_c^G \frac{q_1}{T} \tag{4.39}$$

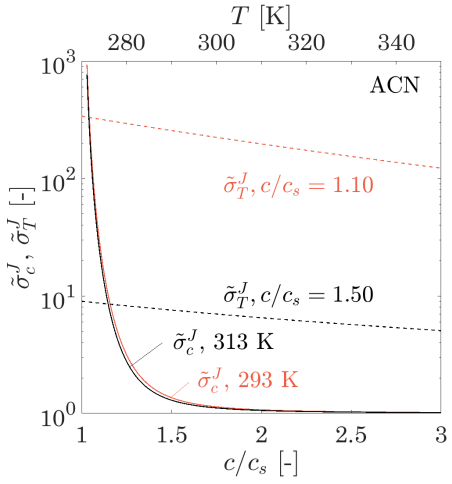
The values of the parameters needed in these equations are reported for p-ABA in ACN, IPA, and EtOAC in Section 4.4.2, above.

Figure 4.12 reports the values of $\tilde{\sigma}_c^J$ and $\tilde{\sigma}_T^J$ (left hand side) and of $\tilde{\sigma}_c^G$ and $\tilde{\sigma}_T^G$ (right hand side) computed with the parameters of p-ABA at two different constant levels of T and c, respectively; these are plotted as function of supersaturation, i.e. $c/c_s(T)$, and temperature. Note that A_1 and K_1, which account for the temperature dependence of J and G, respectively, could not be estimated from the experiments of Sullivan et al. [85] with p-ABA, since all experiments were carried out at one constant temperature, we have used two physically reasonable values, namely $A_1 = 1200$ K and $K_1 = 1000$ K. As clearly observed in Figure 4.12, J and G are almost linearly sensitive with respect to T alone, even though the absolute value of $\tilde{\sigma}_T^J$ and $\tilde{\sigma}_T^G$ depend strongly on the value of c at which it is computed. On the contrary, J and G are very non-linearly sensitive with respect to c, particularly at low supersaturation levels, i.e. when $c \sim c_s$, where small changes in concentration

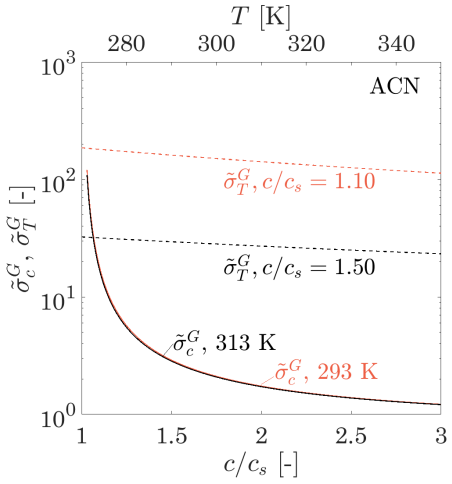
and temperature yield much larger changes in the rates of nucleation and growth.

Though the curves in Figure 4.12 have been calculated with the model parameters obtained from the experiments in ACN, we have verified, on the one hand, that the same trends are observed for the experiments of p-ABA in IPA and EtOAc as well and, on the other hand, that there are no major differences in the absolute values of the sensitivities among the three solvents, which might justify the observed differences in model accuracy, discussed in Section 4.4.2.

To summarise, the detection probability is extremely sensitive with respect to both solute concentration and solution temperature, whereas it is very little or not at all sensitive with respect to system size and to detection time, i.e. to its measured value. Being aware of such sensitivity is essential both in planning and in evaluating experimental results. Underestimating the importance of such a sensitivity might lead to interpreting what is the outcome of poor experimental reproducibility as a manifestation of stochasticity, or to misjudging the quality of a model due to its inability to describe poorly reproducible measurements. Nevertheless, it should be noticed that the sensitivities plotted in Figure 4.12 are low only at high supersaturation; operating under conditions of low sensitivity might lead to bigger relative uncertainties in defining very short detection times. Such situation indicates that there is a need to find a trade-off between reducing the sensitivity while working with large enough detection times.



(a)



(b)

Figure 4.12: (a) The relative sensitivity of J, with respect to c (solid line, as a function of c/c_s at $T = 293$ K, red, and $T = 313$ K, black) and T (dashed lines, σ_T^J has been plotted for $c/c_s = 1.10$, green, and $c/c_s = 1.50$, black). (b) The relative sensitivity of G, at the same conditions used for $\tilde{\sigma}_c^J$ in Figure 4.12a.

4.6 CRYSTAL DETECTION VIA LIGHT TRANSMISSIVITY

One last source of uncertainty, which can be responsible for the mismatch between the experimental data and the model results, is related to how the detection process itself is modelled. The experimental detection times measured and analysed in this Chapter (and in Chapters 5 and 6) have been (mostly) obtained through the measurement of transmissivity, which is among the most common techniques used to perform nucleation experiments. Given its important role in the model capability of describing the data, it is worth investigating such a technique in more detail. First, we briefly recall the basic principles behind this technique and, then, we discuss some experimental results concerning the characterisation of the Crystal16.

4.6.1 *Physical Principle of Transmissivity*

In the context of nucleation experiments, transmissivity, $\mathcal{T}$, is a measure of the total amount of the light intensity travelling unscattered through the solution, from the light source (also known as the emitter) to the sensor. The light source emits a constant number of photons at the same energy level, i.e. at a constant intensity, I_0. In a nucleation experiment, it is convenient to take as reference value the intensity of the clear solution, i.e. without particles: this value, corresponding to the initial state of a nucleation experiments, corresponds also to the maximum transmissivity, i.e. $I(t_0) = I_0 = \max(I)$; we assume in the following $I_0 = 100$, thus $0 \leq I \leq 100$ for any I. Note that the transmissivity, similarly to the absorbance, is given in arbitrary units. The transmissivity is then defined as:

$$\mathcal{T} = \frac{I_0 - I}{I_0} = 1 - \frac{I}{I_0} \tag{4.40}$$

Figure 4.13 illustrates a schematic representation of the emitter-sensor system in the Crystal16, which was used to measure the

data presented above and in Chapter 5 below, as well as most of the data discussed in Chapter 6. The light source (a red laser emitter) and the detector (a light-sensor with an area of about 1 mm^2) are aligned and positioned outside each vial, whose geometry can described as a cylinder of diameter $d = 10$ mm and height $h = 32$ mm.

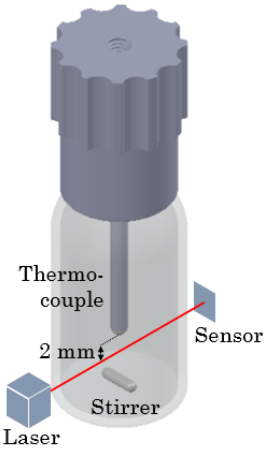


Figure 4.13: Schematics of the transmissivity sensor in the Crystal16.

Let us now consider how the particles in suspension can affect the transmissivity. Intuitively, the intensity measured by the sensor reduces proportionally to the fraction of its area not reached by the light, i.e. obscured by the particles; hence, the transmissivity $\mathcal{T}$ depends on the crystal population. A more quantitative link between the crystal population and $\mathcal{T}$ can be established by exploiting the results derived to analyse turbidity [127] and based on the notion of absorbance. The absorbance, $\mathcal{A}$, is defined as the logarithm of

the ratio between the incident, I_0, and the transmitted, I, light intensity [127]:

$$\mathcal{A} = \ln\left(\frac{I_0}{I}\right) \tag{4.41}$$

Thus, the transmissivity is related to $\mathcal{A}$ by:

$$\mathcal{T} = 1 - \exp(-\mathcal{A}) \tag{4.42}$$

Turbidity, $\mathscr{T}$, is defined as:

$$\mathscr{T} = \frac{1}{l}\mathcal{A} \tag{4.43}$$

where l is the length of the optical path, i.e. the actual geometric length crossed by the light times the refractive index of the suspension. For a suspension of spherical, non-absorbing, and isotropic particles, in the absence of multiple scattering, $\mathscr{T}$ can be expressed as a function of the particle size distribution and of a scattering coefficient, K_s:

$$\mathscr{T} = \int_0^{+\infty} \frac{\pi L^2}{4} K_s(L) f(L) dL \tag{4.44}$$

where K_s depends on the size L of the particles, on the wavelength of the light beam, and on the ratio of the refractive index of the particles with respect to that of the solution. Thus, using Eqs. (4.42), (4.43), and (4.44):

$$\mathcal{T} = 1 - \exp\left(-l \int_0^{+\infty} \frac{\pi L^2}{4} K_s(L) f(L) dL\right) \tag{4.45}$$

Note that the integral in Eq. (4.45) corresponds to the value of K_s averaged on the projected surface area of the particles, which, for spherical particles, is directly related to the second moment ϕ_2 of the PSD. This fact also suggests that two PSDs, f_a and f_b, different, but having the same second moment, $\phi_{a,2} = \phi_{b,2}$, would have a similar effect on the turbidity. Interestingly, in spite of these

considerations, in the literature turbidity is often calibrated with respect to the third moment, ϕ_3, since this is more easily quantified than ϕ_2 [127–129], even though some works accounting explicitly for the second moment have also been published [130]. We shall now investigate whether ϕ_2 should actually be the preferred choice.

4.6.2 *Characterisation Experiments*

To assess the correctness of the hypothesis about the dependence of $\mathcal{T}$ on the second moment, we have performed two types of experiments with glass beads (silica) in water. Glass beads are a convenient choice to study the property of the measurement device itself for two reasons. First, they can be considered geometrically perfect spheres, thus facilitating the characterisation of the PSD, since several techniques can be used to accurately measure the PSD of spherical objects. Second, they neither dissolve in water nor break at the typical operating conditions applied in the Crystal16. All experiments were performed using three different populations of glass beads, labelled as GB_1, GB_3, and GB_2, with a volume average size, L_{43}, of 60, 120, and 80 µm, respectively; they were characterised via Laser Diffraction using a Helos Sympatec^{TM}.

In the experiments, a known mass of particles was added to a clear aqueous solution; the transmissivity was then measured for several minutes to be sure that the signal was stable. Several different values of masses were tested, with each measurement repeated thrice to roughly estimate the standard deviation of the intensity measurements, which was taken as an indication of the measurement error. Knowing the PSD, in each experiment the first four moments could be computed from the mass and associated to the measured intensity. The results are reported in Figure 4.14, which shows the intensity as a function of the different moments; the populations GB_1, GB_3, and GB_2 are associated to the red, blue, and black symbols, respectively. By looking at the figure, it is clear that ϕ_0 and ϕ_1 are different between population GB_1, on the one

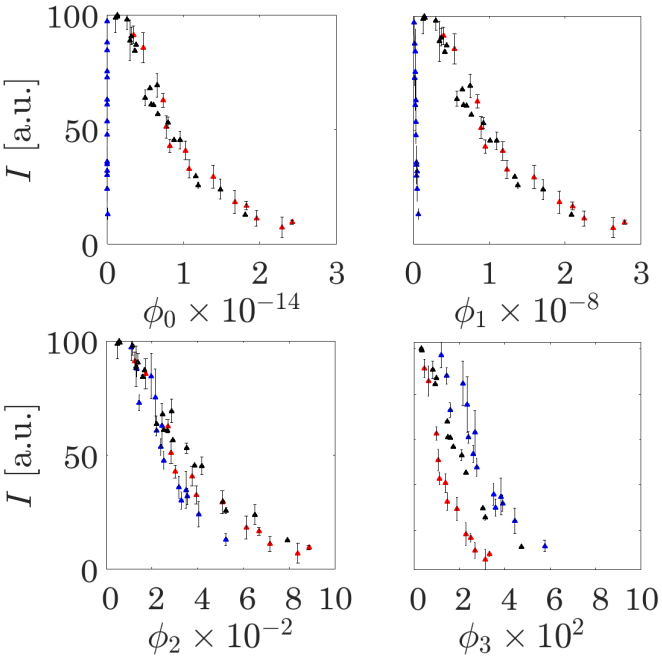


Figure 4.14: The intensity measured in characterisation experiments with three different silica glass beads populations, GB_1 (red), GB_3(blue), and GB_2(black), with a number average size of 60, 120, and 80 µm, respectively

hand, and GB_2-GB_3, on the other hand. The values of ϕ_3, though more similar among the three populations, are however still clearly distinguishable, and each population can be well identified. On the contrary, the measurements of ϕ_2 largely overlap for all populations, consistently with what one expects based on the hypothesis made in Section 4.6.1.

Let us now focus on the behaviour of the intensity I as a function of ϕ_2. Initially, in the range $0.1 \leq \phi_2 \leq 0.4$, I decreases rather steeply with increasing values of ϕ_2. Beyond $\phi_2 \approx 0.4$, the decrease

becomes less pronounced and appears to flatten for $\phi_2 \geq 0.7$, more evidently so for the data belonging to GB_1 (red symbols). By plotting the logarithm of the intensity, the data seem to exhibit a linear behaviour, suggesting that the functional form of Eq. (4.45) is correct, as one can observe in Figure 4.15, where the data have been reported with the associated error bars.

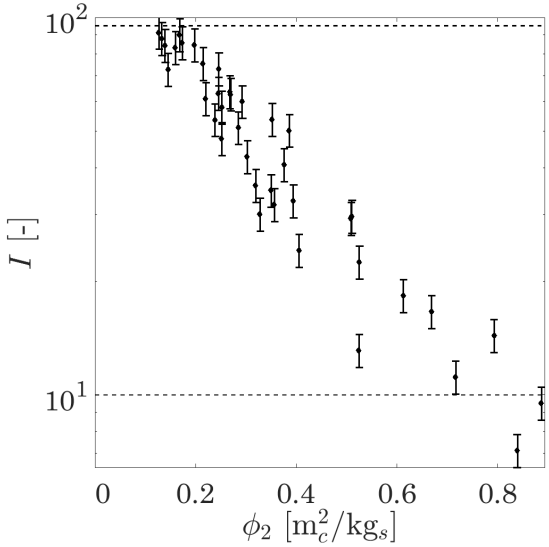


Figure 4.15: The Intensity I measured during the characterisation experiments is plotted as a function of the second moment, for all populations together. The vertical axis is on a logarithmic scale.

In light of these results, and based on Eq. (4.45), the data have been fitted with an exponential function of the form:

$$I(\phi_2) = \begin{cases} 100 \exp\left(\theta_0 - \theta_1 \phi_2\right) & \forall\ \phi_2 \geq \theta_2 \\ 0 & \forall\ \phi_2 < \theta_2 \end{cases} \tag{4.46}$$

where θ_0, θ_1, are the parameters to fit, while θ_2 is obtained from the condition $I(\phi_2 = \theta_2) = 100$. Note that, using Eq. (4.46), one can define a threshold value for detection, ψ_A, in analogy to the value α_V in Eq. (3.13), e.g. by setting the detection condition at $\psi_A = 2\theta_2$. Clearly, the kinetic parameters obtained with this detection condition would be in general different from those obtained in Section 4.4 above. Further investigations would be necessary, on the one hand, to determine a sensible detection threshold and, on the other hand, to quantify the extent of change in the model parameters on the same system for different detection conditions. The latter issue is particularly relevant to fairly compare the kinetics estimated from data measured using different detection techniques.

By fitting the logarithm of I with a straight line, using a simple least-square methods, θ_0 and θ_1 have been estimated as $\theta_0 = 0.268$, $\theta_1 = -3.24$, thus $\theta_2 = 0.08$; the associated R^2 is 0.91. Eq. (4.46) with the chosen set of parameters is illustrated as a blue solid line in Figure 4.16. We repeated the fitting procedure using the third order moment, ϕ_3, rather than ϕ_2; the fit is rather poor and yields a R^2 of 0.45, when all data were fitted together. By fitting each population individually, on the contrary, the fit improved significantly $R^2 \approx 0.90$; however, in this case, the threshold value identified, which should be related to the parameter θ_2, is different for each different population. Consequently, using the same detection criterion for the third moment with every different population would introduce further inaccuracies in the model and discrepancies in the description of the data.

Finally, we have also investigated the influence of the stirring rate on detection. The interest in this aspect of detection is twofold. On the one hand, it is in general important to understand if by changing the operating conditions the outcome of a process is changing in an unexpected way, since this is an indication that an important feature of the process is not captured by the model. On the other hand, several previous works in the literature have reported a change in detection time measurements when changing

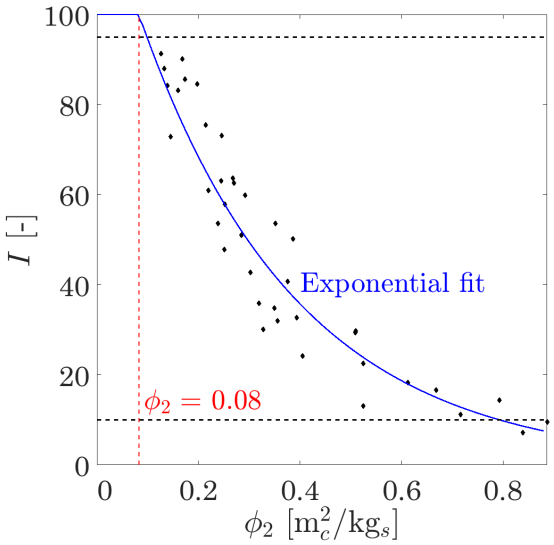


Figure 4.16: The Intensity I, plotted as a function of the second moment, for all measurements (black symbols) and the fitting using Eq. (4.46) (blue solid line).

the stirring rate [53, 95, 131–133]. The framework of classical nucleation theory, which was used to interpret those results, did not offer an explanation for such an effect; at the same time, it was never verified in the aforementioned works whether the observed variations were due to spurious effects, i.e. due to the influence on the measurements device itself. A similar situation arises also when PSD characterisations are performed using the so-called equivalent sphere approximation, i.e. by interpreting the data generated from particles of different shapes, assuming that they are actually spheres with an equivalent volume. This assumption is known to cause a false interpretation of the data, to which a pseudo-size dependent growth is associated, even though the actual growth rate is size-independent [134].

Here, we aimed at testing the absence of spurious effects of the stirring rate on the detection. Eq. (4.45) suggests that the light intensity I should not be influenced by the stirring rate, i.e. by the fluid-dynamic conditions, as long as the system is well-fluidised, i.e. the crystals in suspension neither settle nor gather in a dead mixing zone of the reactor. In fact, no explicit dependence on the fluid dynamic conditions appears explicitly in Eq. (4.45). To test this hypothesis, we conducted experiments at constant value of ϕ_2 and measured the intensity in the Crystal16 at different values of the stirring rate. The glass beads populations GB_1 and GB_2 were used and eight different stirring rates were tested, from a minimum of 500 rpm (rounds per minute), below which the system was no longer well fluidised, and a maximum of 1250 rpm, the maximum value attainable by the device. Figure (4.17) shows the results obtained in these experiments: one can observe that no evident variation has been recorded at the different stirring rates, thus confirming that the stirring does not influence the detection mechanism itself. Consequently, the differences observed in previous experiments with a Crystal16 (or similar devices) should be due to an actual effect of the fluid dynamics on the processes of nucleation and growth.

Thus summarising, the results discussed in this Section allow to highlight two important conclusions for the interpretation and the modelling of the data. First, although different detection conditions have been used in the literature to model the system, the most appropriate one, independent of the specific population of crystals forming in the volume, seems to be a condition related to the second order moment. Given that parameters such as the crystal habit, or the population mean size, change from system to system, neglecting them, while using at the same time detection conditions related to the third order moment or to a minimum detection size, may contribute to introduce further discrepancies between the data and the model. Moreover, since the nucleation kinetics estimated from the experiments are in general heavily dependent on the detection condition, neglecting the differences among unrela-

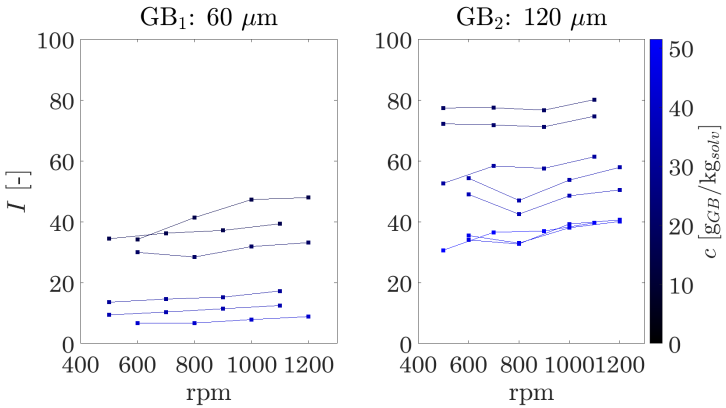


Figure 4.17: The light intensity I, measured at the same value of ϕ_2, but for different stirring rate. As one can see, the intensity does not change in the range of operating conditions investigated.

ted detection techniques (or even different detection devices) while comparing kinetics obtained from experiments conducted with different detection techniques may actually lead to incorrect, biased conclusions. Second, the theoretical framework discussed above and the performed experiments allow to discard the hypothesis of spurious effects on the intensity measured by the device due to the stirring rate, hence also on the measured detection times. If such an influence is observed, it should be related to an effect of the fluid dynamics on the phenomena leading to crystal detection.

4.7 SUMMARY AND CONCLUSIONS

In Chapters 3 and 4 we have developed two different models of crystallisation from solution, with the process initiated by primary nucleation and we have tested these models on several different systems, analysing the numerical methods to estimate the model

parameters and assessing the extent and type of discrepancies between the models and the experimental data.

In Chapter 3, we have discussed hot two different stochastic process of primary nucleation can be implemented in the framework of crystal populations and how the undetectable nucleation events can be connected to observable quantities. To do so, we have assumed that variability of detection times observed in cooling crystallisation experiments is the consequence of an intrinsic stochastic nature of nucleation, which is mathematically defined as a stochastic process, consistently with its physical interpretation. On the contrary, subsequent crystal growth has been assumed to be well characterised as a deterministic process. Model 1 (Sections 2.3.2 and 3.3.1) assumes that the stochastic nature of nucleation, modelled as a Poisson process, influences only the very beginning of the crystallisation, i.e. the formation of the first particle, whereas the subsequent nuclei formation is perfectly deterministic and described by a Population Balance Equation or, equivalently, by the method of moments. The effect of nucleation stochasticity determines only the time *when* the system shifts to the deterministic regime. Model 2 (Sections 2.3.3 and 3.3.2), on the contrary, considers that the formation of all nuclei, not only the first one, is stochastic, whereas the growth is deterministic. Due to the different stochastic process used to describe nucleation, the second model exhibits always a larger detection time variability. However, the magnitude of detection time variance reduces with increasing system volume in both models, consistently with physical considerations and experimental evidence. The models not only account for dynamic profiles of temperature and concentration, but also can be easily modified to describe systems with a non-constant size; furthermore, additional phenomena, such as secondary nucleation, could be also accounted for. Hence, both models can be used to analyse isothermal as well as non-isothermal nucleation experiments, and crystallisation where the driving force is not due to cooling, but, e.g., solvent evaporation or anti-solvent addition. The

models can also be used to simulate the behaviour of the system at different initial concentration, with different temperature profiles, and, by assuming scale-independent kinetics, at different volumes. This last aspect is important to gain a better understanding of the effects of scaling-up and to assess the feasibility of extrapolating kinetic information obtained in small, controlled volumes to larger values, by directly comparing experimental data and simulation results at different scales.

In the present Chapter 4, we have tested and verified the developed models on several different systems. First, we have focused on the necessary statistical preliminaries which allow the use of empirical data to estimate parameters of stochastic processes, by reviewing the notion of empirical cumulative distribution function, and of convergence and rate of convergence of such function to the real, but unknown, cumulative distribution.

Second, we have discussed two different methodologies to estimate from the data the relevant model parameters. The first one based on the classical non-linear least square approach, applied to an objective function based on the Euclidean metric; the second one based on the statistical idea of p-value maximisation of Fisher's meta-analysis, applied to an objective function based on the Chebyschev metric, which arises in the context of almost sure convergence related to empirical distribution functions.

Then, we have validated Model 1 as well as Model 2 against the two rich sets of data from the literature. The first set comprises detection time measurements of cooling crystallisation of paracetamol from water, obtained by Kadam et al. [47] in well-mixed reactors of different volumes (from 1 to 1000 mL). In these experiments, supersaturation has been created by cooling the system at a constant rate, from the saturation temperature until crystallisation was detected; the parameters for this system were estimated using Model 1 and the approach based on the non-linear least square method and the parameters were then used to investigate also the system behaviour with Model 2. We have discussed in detail the

results, particularly focusing on the correlation between nucleation and growth kinetics, for which we have shown that, at the experimental conditions chosen by Kadam et al., independent growth measurements are necessary to avoid parameters correlation. We have also illustrated and discussed the differences between the models in describing the data, and on the effect of the volume and of the cooling rate on the distribution of detection times. Although Model 1 and Model 2 show that different sets of kinetic parameters yield very similar detection times at the specific volume they have been obtained, they produce clearly different results when the system volume changes. The correct kinetics may thus be discriminated by repeating detection time experiments at different system sizes. The second set of literature data, produced by Sullivan et al. [85], comprises detection times experiments obtained from the crystallisation of p-ABA from ACN, IPA, and EtOAc, with supersaturation induced by crash cooling, thus performing isothermal experiments. The data have been interpreted using Model 1, whose nucleation and growth parameters have been estimated. Two different methods have been proposed to estimate the parameters of the model: the one based on the standard non-linear least square approach, the other based on a statistical approach and Fisher's meta-analysis.

Fourth, the discrepancies observed between the models and the experiments in the different systems have also led to study of the system sensitivity on the operating conditions, since nucleation and growth are known to be strongly non-linear functions of the system supersaturation and temperature. The sensitivity analysis has shown that indeed the nucleation experiments should be carefully planned to minimise sensitivity on the initial and operating conditions and that this feature should be accounted for when analysing the results. However, sensitivity alone does not justify why the models describe specific systems and operating conditions better than others.

Finally, we have investigated in more detail the detection technique used to measure the data. In spite of the common practice found in the literature, where often the third moment of the PSD is used to determine detection with transmissivity and turbidity (see for instance the works of Lindenberg et al. [41] and of Kadam et al. [47]), a more correct modelling would be obtained by defining the detection condition on the second moment. Identifying the correct property to measure is important for selecting which detection condition to use in the model (Eq. (3.12)) and how to characterise the threshold value, which ultimately defines the detection time. Inaccuracies in modelling detection can indeed lead to a mismatch between a model and the data, as well as between to apparent incoherences among sets of data measured in different conditions and/or with different techniques.

Thus summarising, three main conclusions can be drawn. First, the presented models are a valid contribution towards the development of sound and reliable protocols for designing, analysing, and interpreting nucleation experiments. The underlying physics, i.e. the crystal formation and growth mechanisms, the statistical model, and the detection method must be well understood and mathematically properly described, lest the model description of the experiments is meaningless. Second, the stochastic nature of primary nucleation and its influence on subsequent processes need accounting for when analysing the experimental data. To this aim, we have explored the criteria for fulling the critical hypotheses behind the statistical analysis of the experimental measurements, namely (1) the statistical representativity of the experimental sample, and (2) the methods to compare the experimental data and a statistical model, via a meta-analysis of the results. The former issue can be dealt with by applying a number of statistical properties, whose cornerstone is the DKWM inequality of Eq. (4.10), which provides a rate of convergence of the measurements towards the underlying statistics, as a function of the number of measured points. Finally, experimental operating conditions have to be controlled very well

and a correct description of the system cannot be limited to the physical and mathematical modelling of the phenomena constituting crystallisation, but must also entail the correct modelling of the detection technique itself, lest even small experimental errors spoil completely the accuracy and usability of the measurements The analysis performed in this Chapter shows that general mathematical descriptions are necessary to understand better the physics of nucleation and to obtain reliable kinetic parameters for process design, optimisation, and control.

5

EXPERIMENTAL METHODS AND DATA ANALYSIS FOR ACCURATE DETECTION TIME MEASUREMENTS

5.1 INTRODUCTION

When measuring nucleation times and characterising nucleation rates, the stochastic nature of nucleation poses a number of challenges, which have been the subject of many papers published in the past [38, 46], as well as recently [51, 93, 95, 135, 136]. It is fair to say that characterising nucleation rates accurately and reliably is still an unsolved scientific problem [49]. Issues in measuring nucleation rates can be categorised as follows.

First, since the nucleation rate is strongly sensitive to operating conditions [48], namely temperature and solute concentration, experimental operating conditions have to be tightly controlled in order to guarantee consistency from one experiment to the next. As a consequence each experiment used to build a representative statistics has to be double-checked before acceptance, as thoroughly discussed recently in the literature [50, 106].

Secondly, as each and every nucleation experiment exhibits a different detection time, it is clear that a large number of experiments are needed to build a representative statistics. It is therefore necessary to establish quantitative statistical criteria that allow to decide when different series of experiments carried out at nominally

The results presented in this Chapter have been reported in: G. M. Maggioni, L. Bosetti, E. dos Santos, M. Mazzotti, "Statistical Analysis of Series of Detection Time Measurements for the Estimation of Nucleation Rates", *Cryst. Growth Des.* **2017**, *17*, 5488–5498 .

identical conditions can be combined and considered together to estimate nucleation rates. We believe that the best approach involves the use of Kolmogorov-Smirnov statistical tests (KS-test), as proposed by Little, Sear, and Keddie [50].

Thirdly, care has to be taken in defining a protocol to extract nucleation rates from distributed nucleation times, and attention has to be given to how to determine the confidence limits of such estimates [38, 46, 53, 133]. A recent paper has exploited the results of a numerical analysis to quantify the generic inherent uncertainty that has to be associated to experimental estimates of growth rates obtained in small volume experiments [106].

Finally, as discussed in some detail in Chapters 2, 3, and 4 above, nucleation is detected after crystals have grown enough, hence nucleation rate and growth rate are necessarily correlated. The estimate of one depends on that of the other, unless one can assume that crystals grow from nucleation to detection in a negligible time as compared to the nucleation time [47, 50, 61, 106, 115].

This Chapter aims at addressing the first three issues above, by devoting to each one a section. As to the fourth issue, we will assume that the growth time is negligible in all the experimental series presented.

Novel aspects with respect to the literature can be summarised as follows. First, we have carried out nucleation experiments with isonicotinamide in ethanol in a Crystal16 set-up (Technobis Crystallization Systems) where it has been possible to crash-cool the solution in less than two minutes. This device has been custom-modified to allow for on-line temperature monitoring, which provides a vital piece of information for determining whether experiments were conducted exactly at the same conditions. Secondly, we have applied the KS-tests through the whole range of significance levels, and not only at 0.05 as typically done [50], thus providing a much richer insight on the reproducibility of the experimental series and on the accuracy of the nucleation rate estimate. Based on the KS-test, we have also developed a statistical method that al-

lows to decide if one can combine an arbitrary number of repeated series – not just two of them – potentially obtaining much larger pools of data. Thirdly, instead of using a numerical analysis to establish confidence intervals on the estimated nucleation rates [106], we have derived and carried out an exact sensitivity analysis for a Poisson model of nucleation, independent of the specific substance used and of its nucleation rate parameters.

5.2 MATERIALS AND METHODS

In this section, we describe first the equipment, the experimental protocol, and the material used in the experiments, then we summarise the background theory, with focus on statistical analysis of data and estimation of nucleation rates.

5.2.1 *Equipment*

All nucleation experiments were performed in a customised version of the Crystal16, engineered and built by Technobis Crystallization Systems^{TM}, according to our specifications. The instrument consists of 16 independent reactors, each of them a cylindrical glass vial of about 1.8 mL capacity. The presence of solids in solution was determined by transmissivity measurements. Transmissivity (see Section 4.6) is a measure of the light, which travels unscattered through the solution from a source to a sensor. In the Crystal16, transmissivity is monitored independently in each reactor; the light source is a red laser. The reference state, i.e. 100 % transmissivity, is taken as the unsaturated, clear solution at the dissolution temperature. Particles in suspension are detected when transmissivity steadily drops below 90%, and remains steadily below such values, so as to avoid experimental errors due to signal noise.

The modified version of the instrument, named Crystal16 Temperature Measurement System, is equipped with an independent thermocouple in each reactor. The system has been designed to

avoid that the thermocouple interfere with the turbidity measurement and the magnetic bar stirrer (Figure 5.1). The thermocouples are k-type probes (Iconel 600) 45 mm long and 1.5 mm in diameter, purchased from and certified by Picolog^{TM}. Since the temperature inside each vial is measured, the supersaturation attained in each experiment is estimated more accurately. It is worth noting that the thermocouples provide evidence that crystallisation has occurred in the system, independently of transmissivity, as shown in Figure 5.1. The drop in transmissivity occurs at the same time

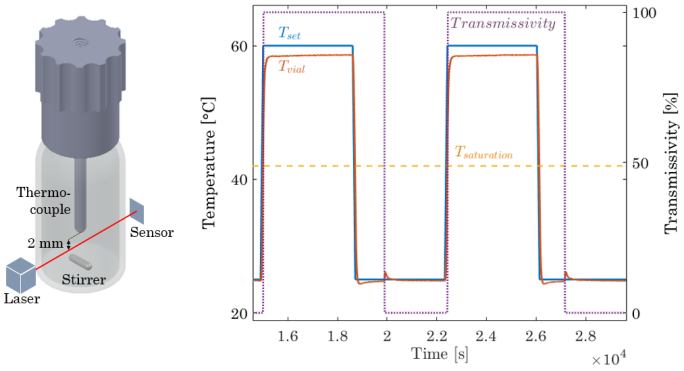


Figure 5.1: Experimental set-up and protocol. On the left hand side: schematic representation of one crystalliser ($\sim$ 1.8 mL) in the C16-TMS, with the thermocouple inserted through the cap and the laser sensor. On the right hand side: typical experimental cycles with INA-EtOH performed at $S = 1.72$, with dissolution at 333 K and crystallisation at 298 K. The detection event corresponds to the transmissivity drop and to the tiny temperature peak. The saturation temperature corresponds to $S = 1.72$.

when the temperature changes. Due to the isothermal conditions imposed by the thermostat, such change must be connected to a release of heat due to crystallisation, which occurs so suddenly that is not immediately compensated by the thermostat.

5.2.2 *Experimental protocol*

5.2.2.1 *Experimental procedure*

For each series of experiments, a stock solution (60 g) was prepared by dissolving the desired amount of solute in the solvent at a temperature above saturation (333 K) in a closed vessel, and used immediately thereafter. For each series of experiments, a new set of 16 glass vials was used as crystallisers and, before each new experimental series, all caps and stirrers were first washed with water and ethanol to dissolve any solids from previous experiments and then dried with compressed air to remove mechanically any possible residual fine particle. Each position in the Crystal16 was assigned to a specific stirrer: given that only the glass vials changed from one series to another, this protocol allowed us to track possible biases due to a specific stirrer, or a specific position, when more repetitions at the same conditions (i.e. same supersaturation) were performed. The caps were not kept in the same position: after accurate tests, their assignment to a specific position was proved neither to influence the results nor to improve the procedure. Before filling, the empty vials were weighed, each together with its stirrer and cap. The following precautions were taken during the loading of the vials with clear, undersaturated solution, typically at a temperature above room temperature:

- the vials are filled as fast as possible with the stock solution at 333 K, to avoid cooling below the saturation temperature;
- the syringe used to fill the vials must be clean at any moment; in the case crystalline formations are observed on its outer or inner part, the syringe is substituted with a new one;
- after loading, the vials are immediately closed to prevent solvent evaporation.

The filling protocol was designed to have conditions as homogeneous as possible inside each crystalliser. The filled and closed vials

were then weighed again to determine the actual mass loaded in each position. The coefficient of variation of the loaded masses was less than 2% in all experiments. The vials were placed in their positions in the Crystal16 and accurately cleaned thermocouples were immersed into the solutions.

A typical cycle consisted of two main steps. First, the solution was kept at the dissolution temperature, T_{dis}, 333 K in this work, for a certain amount of time (60 min), to ensure complete crystal dissolution. The dissolution temperature was selected to be sufficiently higher than the saturation temperature, but also much lower than the boiling point of the solvent. Then, the solution was cooled (at a rate of 20 K/min in this work) to the crystallisation temperature, T_{cry}, i.e. 298 K in all experiments reported in this paper (see Section 5.2.2.3 for further details about temperature accuracy). The duration of the isothermal step of each cycle varies according to the supersaturation (30 to 240 minutes from the highest to the lowest value of S in this work). Finally, after the crystallisation step, the system was heated at T_{dis} to dissolve the particles before starting a new cycle. For the entire duration of the experiments, the vials were stirred at 700 rpm with a standard magnetic bar. Once the whole series of cycles were completed, the vials were weighed for a third time to quantify the amount of solvent evaporated during the experiment.

Note that, while for each cycle only up to sixteen nucleation events and the corresponding detection times can be measured, the number of cycles in a complete series is limited by the extent of evaporation (see below, Section 5.2.2.3). Therefore, in order to obtain a large set of data points, typically more than one series at the same nominal conditions were performed.

5.2.2.2 *Filtration of the Solution*

The clear, dissolved parent solution was filtered at 333 K before being loaded in the reactors. Filtration was performed during the loading of the vials: once the syringe was full with parent solution,

a 0.22 µm PTFE filter was attached to the tip, then the vial was filled with the solution passing through the filter. The filter was then removed from the tip of the syringe, which was accurately washed with the solvent and inserted again in the reactor for refilling. All the other steps described for the loading of vials were kept unchanged, but additional care was payed when filters had to be substituted. After loading four vials the filter was usually replaced with a new one. The change could be necessary before, if crystals were observed in the filter, or if the pressure needed to empty the syringe increased.

5.2.2.3 *Corroboration of Experimental Measurements*

The actual crystallisation temperature do not always match exactly the set temperature and evaporation could occur during the cycles. Temperature changes and solvent evaporation affect the kinetics by modifying the actual supersaturation inside the vials. Thus, we have defined criteria to validate each experimental point, by verifying that it is indeed measured at the nominal operating conditions or very close to them [50, 59]. Such validation procedure guarantees that measurements are homogeneous and can be considered as sampled from the same statistical distribution. Note that three numbers characterise each series of experiments: the total number of experiments, Q_{tot}, given simply by the number of cycles times 16; the number of experiments complying with the criteria for reproducibility discussed below, i.e. the acceptable experiments (see also Section 4.2), Q; and the number of experiments where crystal detection has been possible before the cooling period is completed and the vials are heated up again (see Figure 5.1), q.

Temperature Criterion.
Since the thermocouples measure the temperature inside each vial for the entire duration of the experiment, i.e. between establishment of the nominal supersaturation and detection, it is possible to verify that during this whole time interval the temperature remains within a 0.5 K range around the set value, i.e. it is of 298 ± 0.5

K. Only experiments fulfilling this criterion yield a valid detection time measurement.

Concentration Criterion.
In this work, we assume a moderate change of the solute's activity coefficient in a range between the actual solute concentration, c, and its temperature-dependent solubility, $c_s(T)$; i.e. we have assume that $\gamma(T, c) \simeq \gamma^*(T, c_s)$. Therefore the supersaturation is defined as in Eq. (3.24) in Section 3.3.3. Note that because of the temperature criterion above, supersaturation has always been computed using the value of solubility at 298 K, i.e. 94 g/kg. Solvent evaporation has been observed during the experiments, thus it must be considered. It has been assumed that the same amount of ethanol evaporates during each cycle, progressively increasing the concentration (hence the supersaturation) in the individual vials. Only the detection times, whose corresponding supersaturation is less than a value of 0.02 larger that the set value, have been considered valid measurements.

5.2.3 *Materials*

Isonicotinamide (INA), purity $\geq 99\%$, was purchased from Sigma-Aldrich, while ethanol (EtOH) from two suppliers, Fluka (Ethanol Laboratory Reagent, absolute, $\geq 99.5\%$) and EMD Millipore (Ethanol Absolute For Analysis Emsure Acs, Iso, Reag. Ph. Eur.). For the experiments where the solution was filtered, 13 mm syringe filters (PTFE Hydrophobic, Nonsterile, Pore Size 0.22 μm) were used. It is known that INA has different polymorphs [137, 138]. The supersaturation in this work is calculated for the form stable at 298 K, named Form 1 by Hansen and co-workers [137]. We measured the solubility of INA in ethanol at 298 K gravimetrically and found it to be 0.093 g/g (gram of substance per gram of solvent); this is the average value of six independent measurements.

5.2.4 *Estimation of Nucleation Rates*

The definition of the empirical distribution function (eCDF) F^* has been already discussed in Section 4.2.1, to which we forward the reader for the definition of the eCDF (Eq. (4.4)) and its related properties. Here, we recall only that each series of $Q_{tot,k}$ measurements yields Q_k acceptable experiments, i.e. fulfilling the criteria discussed above in Section 5.2.2.3, and q_k values of the detection time, where q_k is the number of experiments in which crystals have indeed been detected, while $Q_k - q_k$ the number of acceptable experiments in which no crystal has been detected before the end of the experiment.

The ultimate goal of detection time experiments is to measure the nucleation rate, J. In this Chapter, J is estimated, for each supersaturation value S, using the non-linear least square method, based on two hypotheses. First, nucleation is proceeds according to Model 1, i.e. the first events follows a Poisson process, with the cumulative distribution function defined by Equation (2.20) in Section 2.3.2. Second, following the model proposed by Jiang et al. [59] and Xiao et al. [106], the growth time t_G is defined, dependent on the experimental conditions (supersaturation and temperature), is estimated, for each series of experiments as supersaturation S_k, as the shortest detection time measured at such conditions:

$$t_{G,k} = \min_{S_k}(\mathbf{t}^*_{\mathbf{D},\mathbf{k}}) \tag{5.1}$$

The value of J at each supersaturation, for a given system volume V, can be estimated by using the linearised form of Equation 4.6, minimising the following sum of the squares of the residues between experiments and model:

$$\Psi_k = \sum_{j=1}^{q_k} (\ln(1 - F^*_Q(t^*_{D,jk})) + JV(t^*_{D,jk} - t_{G,k}))^2 \tag{5.2}$$

5.2.5 *Repeated Experiments*

In a deterministic process, reproducibility is the possibility to replicate the experimental results for the same set of conditions and operating parameters within a certain accuracy, which is controlled by the background noise affecting the measured signals. Thus, in an ideal, zero-noise deterministic system, the values of the variables measured in different repetitions of the same experiment do not change. On the contrary, the outcome of a stochastic process such as nucleation is by definition not reproducible in this sense. With reference to the experiments reported here, in principle each series samples the same underlying distribution, but its measurements are expected to be different from the measurements obtained in a different series performed under the same operating conditions.

This point is illustrated in Figure 5.2, where four series of experiments carried out at the same temperature (298 K) and supersaturation ($S = 1.41$) are shown in the form of the corresponding eCDF (blue, red, yellow and violet for series 1, 2, 3 and 4; the number of cycles and of retained measurements for each series in reported in the inset of the same figure). When confronted with this set of data or similar ones, the question is whether to consider them self-consistent or reproducible, i.e. whether to consider them as estimates of the same underlying cumulative probability function. If this were the case, then one could combine the four sets of data and obtain a single set of data, with a much larger overall number of data points. The latter could be the basis for a more accurate estimate of the nucleation rate corresponding to the selected supersaturation level.

In this section we want to explore when and how multiple series of nucleation experiments can be combined, by first providing the statistical basis for such investigation and then applying it to the four series shown in Figure 5.2.

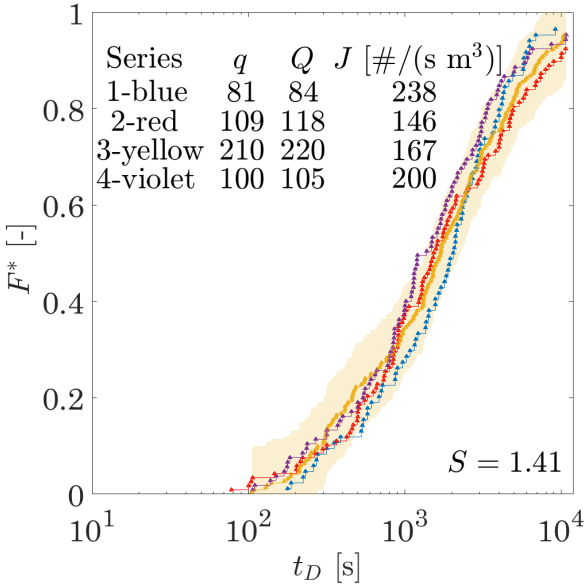


Figure 5.2: Nucleation experiments of INA in EtOH. Cumulative distribution functions of 4 experimental series at the same nominal conditions, namely $S = 1.41$. For series 3, the corresponding uncertainty area has been drawn: its boundaries have been computed with $\tilde{C} = 0.95$ according to the procedure detailed in Section 5.2.7. Note that all series are mostly within the uncertainty area.

5.2.6 *Kolmogorov Statistic*

As introduced in Section 4.2, one proper measure of the distance in the space of distribution functions between either an eCDF $F^*_Q(t_D)$ and a cumulative distribution function $F(t_D)$, or a pair of independent eCDFs $F^*_{Q_1}(t_D)$ and $F^*_{Q_2}(t_D)$, is the Chebyschev distance, defined as the maximum of the absolute values of the difference between the two distributions, D_Q given by Eqs. (4.7) and D_{Q_2,Q_1} by

Eq. (4.9), respectively: More specifically, let us define the following two norms:

$$D_Q = \left\| F^*_Q(t_D) - F(t_D) \right\|_\infty \tag{5.3}$$

$$D_{Q_2,Q_1} = \left\| F^*_{Q_2}(t_D) - F^*_{Q_1}(t_D) \right\|_\infty \tag{5.4}$$

Such norms are in turn random variables, whose distribution has been studied first by Kolmogorov [139] and Smirnov [140], and then by many others. To simplify the notation, let us call both of them D_Q , where $Q = Q_1$ in the former case, and $Q = Q_1 Q_2/(Q_1 + Q_2)$ in the latter case, as shown by Kolmogorov [139]. Let us also define the null hypothesis that $F^*_Q(t_D)$ is sampled from the continuous cumulative distribution function $F(t_D)$ in the first case, and that $F^*_{Q_1}(t_D)$ and $F^*_{Q_2}(t_D)$ are both sampled independently from the same continuous cumulative distribution function in the second case. Under the null hypothesis, D_Q is a random variable and the quantity $\sqrt{Q}D_Q$ (called for brevity K-number in the first case and KS-number in the second) exhibits a limiting cumulative distribution function $\Phi(\kappa)$ given by:

$$\Phi(\kappa) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}\exp(-2i^2\kappa^2) \tag{5.5}$$

This equation should be interpreted as follows. Under the null hypothesis, the probabilities that the product $\sqrt{Q}D_Q$ is smaller than κ (the confidence level, C) or larger than κ (the significance level, α) are:

$$C = \Pr\{\sqrt{Q}D_Q \leq \kappa\} \xrightarrow{Q\to+\infty} \Phi(\kappa) \tag{5.6}$$

$$\alpha = \Pr\{\sqrt{Q}D_Q > \kappa\} \xrightarrow{Q\to+\infty} 1 - \Phi(\kappa) \tag{5.7}$$

The probabilities on the l.h.s. of these equations are in general functions of Q and κ, but for $Q \geq 35$ (a condition that is always fulfilled in the experiments reported here) they are indistinguishable from the asymptotic expression on the r.h.s. [141].

The asymptotic relationships $C = \Phi(\kappa) = 1 - \alpha$ are illustrated in Figure 5.3. Note that the definitions of confidence and significance levels provided by Eqs. (5.6) and (5.7) are mathematically rigorous and should be referred to whenever interpreting data using C and α. For data yielding values of the K- or KS-number corresponding to point A in Figure 5.3, the null hypothesis is fulfilled with a confidence level as low as 0.2 and with a significance as high as 0.8. For data corresponding to point B in the figure, the null hypothesis is fulfilled at a higher confidence level but at a lower significance level. Point R in the figure corresponds to the value which is used as a standard in statistics, namely a confidence level of 0.95 at a significance level of 0.05, i.e. corresponding to $\kappa = 1.358$.

It is also worth noting that the following inequality on α (DKWM inequality, from Dvoretzky, Kiefer, Wolfowitz [119], and Massart [120]) holds true for every Q and κ:

$$\alpha = \Pr\{\sqrt{Q}D_Q > \kappa\} \leq \Phi_{DKWM}(\kappa) = 2\exp(-2\kappa^2) \qquad (5.8)$$

Using $\Phi_{DKWM}(\kappa)$ instead of $\Phi(\kappa)$ yields an error of less than 0.1% on α for α for $\kappa \geq 1$ and $\alpha \leq 0.271$. The function $\Phi_{DKWM}(\kappa)$ can be easily inverted, thus yielding:

$$\kappa = \sqrt{\frac{1}{2}\ln\left(\frac{2}{\alpha}\right)} \qquad (5.9)$$

which is a fully acceptable approximation when $\alpha \leq 0.271$.

5.2.7 *Statistical Tests*

Let us consider again the four experimental series illustrated in Figure 5.2, having a number of data points between 84 and 220, with the objective of merging them into a single series with a larger number of data points by combining as many of them as possible, i.e. by combining those that can be considered well reproduced under the chosen experimental conditions. Such discussion shall

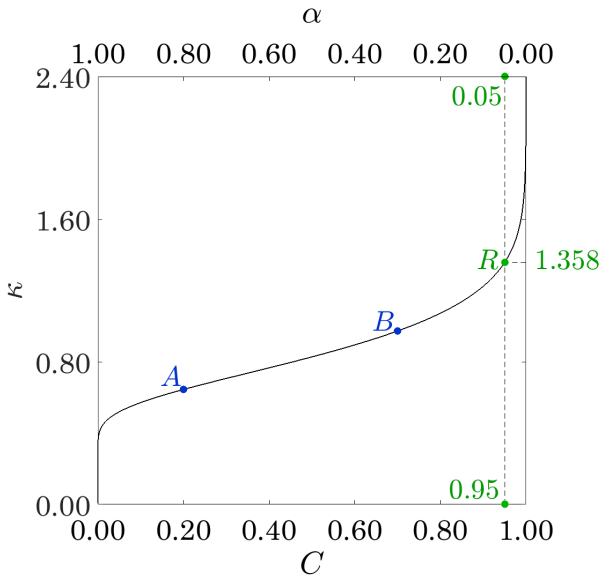


Figure 5.3: Kolmogorov distribution $\Phi(\kappa)$ (Equation (5.5)). Both the confidence level C (lower x-axis) and the significance level α (upper x-axis) are indicated. A and B represent two arbitrary points with $C = 0.20$ and $C = 0.70$, while R represents the standard value $\tilde{C} = 0.95$, corresponding to $\tilde{\alpha} = 0.05$ and $\kappa = 1.358$.

be based on two statistical tests, which are in turn based on the two norms defined in the previous section.

The first test considers a single eCDF $F^*_Q(t_D)$, e.g. the yellow series in Figure 5.2 having $Q = 220$, and evaluates the probabilistic distance of the eCDF from the continuous cumulative distribution function $F(t_D)$ given by Eq. (4.7), under the null hypothesis that the former is indeed sampled from the latter [141]. To this aim, one must (i) select a significance level α, e.g. $\alpha = 0.05$, (ii) calculate the corresponding κ value either from Eq. (5.9) if $\alpha \leq 0.271$ or by inverting Eq. (5.7), i.e. as $\kappa = \Phi^{-1}(1 - \alpha)$, otherwise, e.g. $\kappa =$

1.358 in the case $\alpha = 0.05$, (iii) determine the distance between $F^*_Q(t_D)$ and $F(t_D)$ as $D_Q = \kappa/\sqrt{Q}$ (obtained by substituting the inequality $\sqrt{Q}D_Q > \kappa$ in Eq.(5.7) with the equality, and using $Q = Q_1$), e.g. $D_Q = 0.0916$ in the case $\alpha = 0.05$, and (iv) use it to draw a band around the eCDF, which is defined as $F^*_Q(t_D) \pm D_Q$ and is shown in Figure 5.2 in light yellow colour.

The interpretation of this first test is that under the null hypothesis the continuous cumulative distribution function $F(t_D)$ will lie within the light yellow band of Figure 5.2 with a probability $C = 1 - \alpha$. It is worth making a couple of additional remarks. Firstly, the amplitude of such uncertainty band depends only on the number of data points, Q, and on the significance level, α. Secondly, in the example of Figure 5.2 all empirical distribution functions appear to lie within the yellow band, which is defined for $\alpha = 0.05$ in this case.

Could this feature be considered a proof of good reproducibility? In principle yes, but with the caveat that a different (larger) value of α would yield a different (narrower) band, hence a different assessment of reproducibility. To deal with this conundrum the second statistical test is useful.

5.3 COMBINATION OF EXPERIMENTAL SERIES

5.3.1 *Combining Series*

The second test, a 2-KS test, analyses a pair of independent eCDFs $F^*_{Q_1}(t_D)$ and $F^*_{Q_2}(t_D)$ and checks the null hypothesis that they are sampled independently from the same continuous cumulative distribution function [50]. To this aim, one must (i) define $\bar{Q} = Q_1 Q_2/(Q_1 + Q_2)$, (ii) calculate D_{Q_1,Q_2} through Eq. (4.9), (iii) compute the corresponding KS-number $\kappa = \sqrt{\bar{Q}} D_{Q_1,Q_2}$, and (iv) determine the corresponding C and α values through equations (5.6) and (5.7) (or by reading the values in Figure 5.3).

The KS-numbers for all six possible combinations of two series from the four in Figure 5.2 span the range 0.69 (pair 2-3) to 1.19 (pair 1-4) and are used to plot the crosses in Figure 5.4 on the background of the function $\Phi(\kappa)$ shown in Figure 5.3. Their coordinates are the KS-number itself (κ coordinate) and for the sake of comparison and clarity the value of C associated to the smallest KS-number among them (yellow cross), which is $C = 0.271$ for the pair 2-3.

Since uncertainty, e.g. the uncertainty band in Figure 5.2, scales with $1/\sqrt{Q}$, i.e. the reciprocal of the square root of the number of data points, combining experimental series reduces uncertainty.

Combining the data points of series 2 and 3 into one, thus forming what we call the pair 2-3, with a total number of $N_{23} = 338$ data points, constitutes the safest, most conservative choice. This is also the one that imposes the strictest constraint on reproducibility between series, namely a confidence level of 0.271 and a significance level of 0.739, i.e. rather far away from the typical values of 0.95 and 0.05, respectively, used in statistical tests. The yellow-shaded region in Figure 5.4 highlights the KS-number and the corresponding confidence level (0.271) associated to the pair 2-3.

The next step in increasing the number of data points and decreasing the uncertainty is that of forming a triple, e.g. in this case the triple 1-2-3. To this aim, one considers the KS-numbers of all pairs involved, i.e. 1-2, 2-3 and 1-3, and those of the 2KS tests between these three pairs and the third single series, i.e. (1-2)-3, (2-3)-1, and (1-3)-2. All such ternary 2KS tests and the corresponding KS-numbers are used to plot the blue triangle symbols in Figure 5.4, while the corresponding 2KS tests on pairs above are plotted as the two blue crosses and as the yellow cross in the same figure (the black triangles and the black crosses correspond to the other possible combinations involving also series 4).

Let us consider the four possible triples that can be obtained from the four series in Figure 5.2, i.e. 1-2-3, 1-2-4, 1-3-4, and 2-3-4,

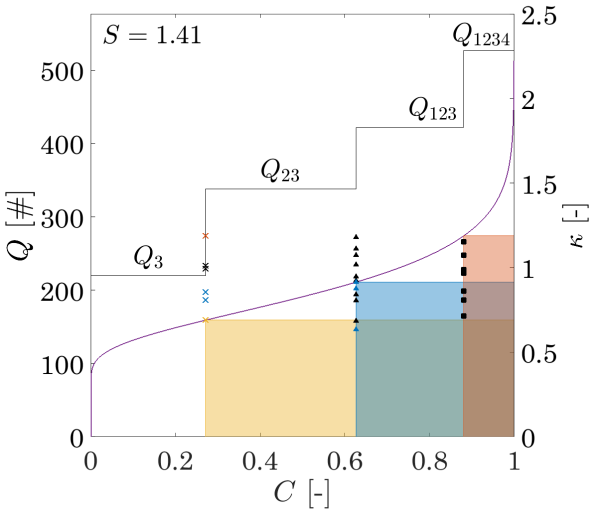


Figure 5.4: Statistical analysis and combination of experimental series (at $S = 1.41$) with different number of points Q (see Section 5.2.4). Results of the KS-tests illustrated with respect to the Kolmogorov distribution. Crosses, triangles, and boxes have vertical coordinate corresponding to the KS-number for pairs, triples, and quadruples, respectively, and horizontal coordinate as discussed in the main text. The boundaries of the three shaded areas are defined by the KS-number, κ (vertical coordinate), and the confidence level, P (horizontal coordinate), associated to the three combined series, yellow for the pair, blue for the triple, and red for the quadruple. The staircase line gives for the series 3, pair 2-3, triple 1-2-3, and quadruple 1-2-3-4 the corresponding number of data points.

as well as the corresponding six KS-numbers, which are associated to each triple as described above for the triple 1-2-3; among these, let us use the largest KS-number for each triple. In the case of the data in Figure 5.2, the triple 1-2-3 exhibits the smallest largest KS-number (0.91) among the four possible triples. As a consequence,

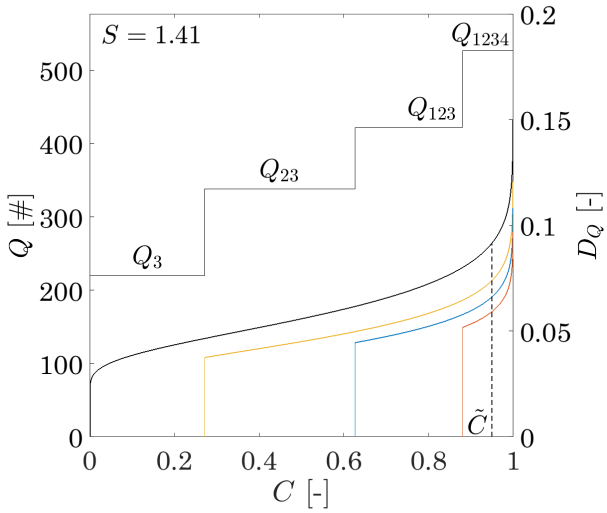


Figure 5.5: Statistical analysis and combination of experimental series (at $S = 1.41$) with different number of points Q (see Section 5.2.4). The four curves plot the function $E_j(P)$ defined in the interval $[C_j, 1)$, as defined in the main text; colors are black, yellow, blue, and red, for the series 3, 2-3, 1-2-3, and 1-2-3-4, respectively.

the triple 1-2-3, having a total of $N_{123} = 443$ data points, is the safest, most conservative choice for a triple, and it is thus chosen as new combination. The blue-shaded region in Figure 5.4 highlights its KS-number and the corresponding confidence level (0.626). The triple 1-2-3 would fulfil the null hypothesis for all possible 2KS tests made with its three component series with a confidence level of 0.626 or larger (hence with a significance level of 0.374 or smaller).

Finally, all four series in Figure 5.2 can be combined, after evaluating the KS-numbers of all pairs, of all merged pairs with singles series, of merged pairs with other merged pairs (tests involving four original series overall), and of all triples with single series (in-

volving four series as well). All KS-numbers obtained by including four series corresponds to the box symbols in Figure 5.4. In this particular case, it turns out that the largest KS number of all is that of the 1-4 pair (1.19, corresponding to the red cross in Figure 5.4). Such value determines the red-shaded region in Figure 5.4 corresponding to the quadruple 1-2-3-4 (defined by a total of $N_{1234} = 527$ data points), bounded by $\kappa = 1.19$ and $C = 0.881$.

With reference to the quadruple 1-2-3-4 (red region), the triple 1-2-3 (blue region), and the pair 2-3 (yellow region), Figure 5.4 illustrates effectively where each of these combinations of single experimental series is feasible, i.e. for decreasing values of KS-number, κ, and of confidence level, P, in going from the quadruple to the pair. It is obvious that in choosing how many individual series to combine, one has to trade off inclusiveness (in going from left to right in the diagram more series are included whilst the confidence level increases and the significance level decreases) and strictness (in going from right to left the κ value decreases, and the requirements for experimental series to be considered as sampling the same continuous cumulative distribution function become more and more stringent).

Contrary to what typically done in the literature, where fixed values of P and α are considered (typically 0.95 and 0.05, respectively) and based on these values the null hypothesis is tested, we argue that Figure 5.4, where the whole range of values of C and α are explored, provides a more complete picture of the statistical features and the interrelation of the different experimental series. Figure 5.4 shows also the number of data points, on which the three combinations above, as well as the single series with the largest number of data points, are based.

Using these numbers the uncertainty associated to each of these four series, i.e. the single series 3, the pair 2-3, the triple 1-2-3 and the quadruple 1-2-3-4, can be calculated for any confidence level larger than the smallest confidence level, at which that series is feasible (left border of the shaded regions in Figure 5.4). Let us

label each of these four eCDFs with the index j, chosen as the number of original series included, i.e. the series 3 has $j = 1$, the pair 2-3 has $j = 2$, the triple 1-2-3 has $j = 3$ and the quadruple has $j = 4$; accordingly the number of data points is called M_j, so as $M_1 = Q_3$, $M_2 = Q_{23}$, $M_3 = Q_{123}$, and $M_4 = Q_{1234}$. Since they have different numbers of data points, the uncertainty, with which the underlying continuous cumulative distribution function is estimated (see the beginning of Section 5.2.7), is obtained by dividing κ by the square root of a different number, namely M_j; we call such uncertainty E_j. Let us now call C_j the minimum confidence level at which the j-th eCDF is feasible, and κ_j the corresponding K-number from equation 9 (upper border of the shaded regions in Figure 5.4); note that $C_1 = \kappa_1 = 0$. The uncertainty associated to the j-th combined eCDF is a function of the confidence level as calculated above, or equivalently of κ; it defines a curve in the open interval $[C_j, 1)$, which is plotted in Figure 5.5, and will be called $E_j(P)$.

It is worth noting that the smallest uncertainty levels for the three combined series, i.e. $E_j(C_j)$, are rather similar, although they correspond to very different confidence levels, whereas at a given confidence level, e.g. $\tilde{C} = 0.95$, where all combinations are feasible, the uncertainty, i.e. $E_j(\tilde{C})$, decreases for increasing number of data points, as expected. This effect highlights once again the importance of considering the whole spectrum of KS-number and confidence levels, when considering how to combine single eCDFs into a larger set of data points as a basis for the estimation of physical properties such as the nucleation rate.

5.4 ESTIMATION OF NUCLEATION RATES

5.4.1 *Nucleation Rates from Combined eCDFs*

The ultimate goal of this work is to determine an estimate of the nucleation rate from the eCDF at a given supersaturation through

the procedure described in Section 5.2.4, and to assess its accuracy using the statistical tools presented in the previous sections. The most important result of the previous section is that such estimation can be based on individual experimental series or on different combinations thereof, each combination being feasible beyond a minimum confidence level and corresponding to a different uncertainty (depending on the number of data points, based on which it is constructed). The goal of this section is to link the statistical features of the eCDF to the uncertainty to be attributed to the corresponding estimate of the nucleation rate.

Let us consider again the four eCDFs measured at $S = 1.41$ and plotted in Figure 5.2, as well as all the eCDFs that can be formed by combining them in all possible ways. In Figure 5.6, the values of the nucleation rate estimated by applying the method described in Section 5.2.4 to the four individual eCDFs (circles), the six possible pairs (crosses), the four triple (triangles), and the quadruple (box symbol) are plotted vs. the number of data points forming that eCDF.

It is immediately apparent how different the estimates obtained from the individual series are, namely from about 140 $\#/(\mathrm{s\ m^3})$ to about 240 $\#/(\mathrm{s\ m^3})$, despite the relative similarity of the four eCDFs in Figure 5.2. As series are combined, the variation between estimates from pairs and from triples becomes smaller and smaller, whereas of course there is only one quadruple, hence only one nucleation rate estimate that is based on all the data points of all the four original eCDFs.

It is thus obvious that if we had to select one of the single series, even that with the largest number of points, we would risk a large error. But this is also true if we decided to use a pair (or a triple) of experimental series, e.g. to retain only the points belonging to the two series, whose KS-number is the smallest, i.e. the pair 2-3 associated to the cross at the rightmost vertex of the yellow-shaded triangle in Figure 5.6 (in fact the difference between the nucleation

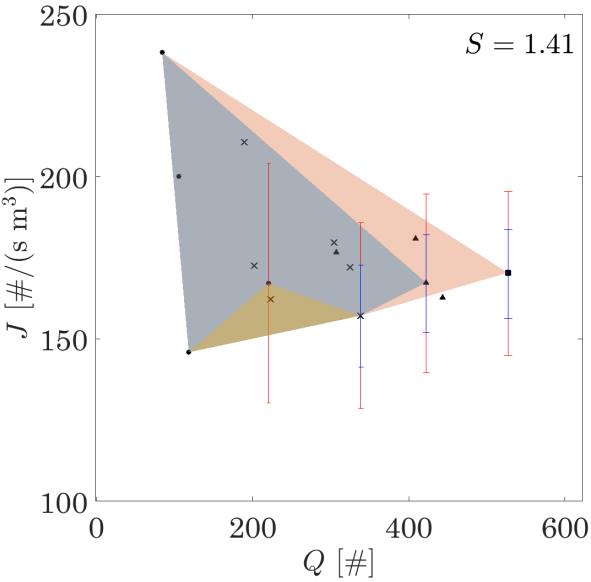


Figure 5.6: Nucleation rates estimated from the four experimental series at $S = 1.41$. Symbols refer to all individual series (●), and all possible combinations of two (×), three (▲), or four (■) series; the horizontal coordinate is the number of data points of the individual series of or the combines ones. Red error bars represent the confidence intervals estimated at $\tilde{C} = 0.95$, whereas the blue error bars represent those at $C = C_j$, using Equation (5.13).

rate estimated from the pair 2-3 and that estimated from the pair 1-4 is about 50% of the former value).

On the other hand, the variability exhibited by the estimates obtained from the single series might reflect an experimental variability that is intrinsic and that for the sake of best experimental practices should be accounted for. Based on this argument, one should combine all data points and estimate the nucleation rate from the quadruple. By doing so, however, one risks to include in

the analysis one or more real outliers caused by an undetected experimental error, e.g. the single series yielding the largest nucleation rate value, and thus to spoil the final estimate of the nucleation rate.

The discussion above demonstrates the existence of a real issue to be addressed by the researcher. It appears natural to exploit the statistical analysis outlined in the previous section to provide a rationale for the choice of the experimental series that have to be included in the combined eCDF, on which the final nucleation rate estimate is based.

5.4.2 *Propagation of Uncertainty*

This section aims at establishing a quantitative relationship between the intrinsic uncertainty of the eCDF, measured by the distance D_Q (see Eq. (4.7)) and related to its number of data points N, and the uncertainty that characterises the nucleation rate estimated from the eCDF via the regression presented in Section 5.2.4. Recently such relationship has reportedly been determined through a large number of numerical calculations yielding a 0.95 confidence interval of J as a function of N (see Figure 2 of Xiao et al. [106]). Assuming a general validity of the result thus obtained, the authors have directly applied such numerical confidence interval to their experimental data.

Here, we apply a different approach, based on the following observations. On the one hand, any uncertainty on the eCDF propagates to a related uncertainty on the value of J extracted from that eCDF. On the other hand, any change in the value of J causes a related change in the cumulative distribution function given for instance by Eq. (4.6). The latter effect is easy to compute, as shown below, and can be used to determine the former.

Let us consider two cumulative Poisson distributions, $F(t_D; J')$ and $F(t_D; J)$, calculated for the same system (same system volume, V, and same growth time t_G), but using two different nucleation

rates, namely J' and J. The relative change in the value of the nucleation rate, η, is defined as:

$$\eta = \frac{J' - J}{J} \tag{5.10}$$

Using Eq. (4.6), the difference between the two distributions (defined for $t_D \geq t_G$) is:

$$\Delta F(t_D) = \exp(-JV(t_D - t_G)) - \exp(-J'V(t_D - t_G)) \tag{5.11}$$

This function is 0 when $t_D = t_G$ and $t_D \to \infty$, and it is positive or negative if $J' > J$ or $J' < J$, respectively. The stationary point of $\Delta F(t_D)$ (maximum if positive, minimum if negative) is where $\mathrm{d}\Delta F(t_D)/\mathrm{d}t_D = 0$, namely where:

$$(J' - J)V(t_D - t_G) = \ln J' - \ln J \tag{5.12}$$

Substituting this into Eq. (5.11) yields the following equation for the stationary value of $\Delta F(t_D)$, denominated ΔF_c:

$$\Delta F_\mathrm{c} = \frac{\eta}{(1+\eta)^{1+1/\eta}} \simeq \frac{\eta}{e} \tag{5.13}$$

where e is the Euler number. The functional relationship between ΔF_c and η is strictly monotonically increasing. The approximated linear relationship holds for small values of η (i.e., $\Delta F^*/\eta$ approaches $1/e$ when $\eta \to 0$), causing an error on the value of ΔF^* of up to 10% for $|\eta| \leq 0.2$; e is the Euler number.

The Kolmogorov statistics establishes the intrinsic uncertainty, D_Q, with which the eCDF $F_N^*(t_D)$ based on N data points estimates the underlying continuous cumulative distribution function for any given confidence level C (see Eq. (5.6)). More precisely, according to the first test presented in Section 5.2.7, such uncertainty is calculated from the K-number associated to C, i.e. κ obtained from Eq. (5.6), as $D_Q = \kappa/\sqrt{Q}$. Such uncertainty on the underlying continuous cumulative distribution function corresponds to

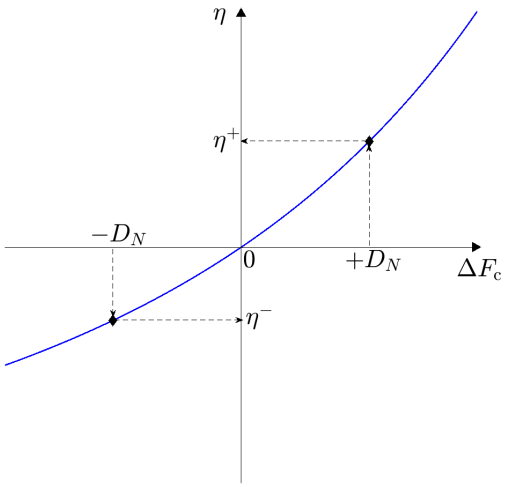


Figure 5.7: Propagation of uncertainty and confidence intervals for nucleation rate estimates. Plot of the function $\Delta F^*(\eta)$ defined by Eq. (5.13). For $\Delta F^* \pm D_Q$, the curve gives upper and lower relative bounds of the confidence interval on the nucleation rate estimates, η^+ and η^- (see main text).

an uncertainty on the associated nucleation rate, and such correspondence must exactly be given by Eq. (5.13), where $\Delta F_c = \pm D_Q$. Kolmogorov statistics applies in fact to a distance calculated using the infinite norm of Eq. (4.7). Note that Eq. (5.13) indicates that the relative uncertainty on the nucleation rate is about 2.7 times larger than that on the eCDF (in terms of absolute values).

The exact formula yields asymmetric confidence intervals for the nucleation rate estimate, with the distance of the upper bound from the estimated value larger than that of the lower bound. Hence for the same absolute value of the left hand side of Eq. (5.13), but

different sign, one obtains two values of η, namely η^+ and η^-, with $\eta^+ > 0 > \eta^-$ and $|\eta^+| > |\eta^-|$. The asymmetric behaviour of η can be seen in Figures 5.8, where η is plotted as a function of ΔF_c, and in Figure 5.7, where η^+ and η^- are reported as function of Q, for two different confidence levels, namely $\tilde{C} = 0.95$ (solid line) and $C = 0.75$ (dashed line). Note that the relative uncertainty on the nucleation rate is still rather large even for a large number of points ($|\eta| \approx 0.2$ for both confidence levels when $N = 500$, as can be seen in Figure 5.7); the shape of the curves in Figure 5.7 suggest that further improving the accuracy of the nucleation rate estimate requires a number of experiments which might be experimentally challenging to obtain. These results are consistent both qualitatively and semi-quantitatively with those reported by Xiao et al. [106] (see Figure 2 of their work).

Thus concluding, C confidence intervals on the value of J estimated from J', obtained from a specific eCDF as to Section 5.2.4, are given by the following inequalities $(1 + \eta^-)J' \leq J \leq (1 + \eta^+)J'$; if and only if the linearised formula is used then an approximate symmetric confidence interval is obtained. Finally, although Eq. (5.13) is valid for a Poisson process only, the logic followed here could be applied to any stochastic process, whatever its underlying statistics.

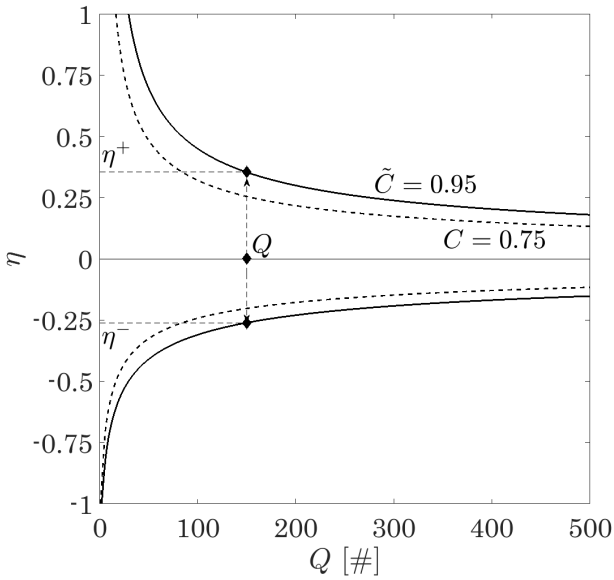


Figure 5.8: Propagation of uncertainty and confidence intervals for nucleation rate estimates. Plot of the (relative) confidence interval, η^+ and η^-, vs. number of data points, N, at two confidence levels, namely $\tilde{C} = 0.95$ (see Eq. (5.14), solid line), and a generic value $C = 0.75$ (dashed line). Note that the confidence interval is always asymmetric, i.e. $\left|\eta^+\right| > \left|\eta^-\right|$, and that they agree qualitatively and semi-quantitatively with those calculated by Xiao et al. [106].

5.4.3 *Estimates and Confidence Intervals*

In this section we aim at considering the different possible estimates of the nucleation rate reported in Figure 5.6, and at attributing to each of them proper confidence intervals. We focus on the four combined eCDFs highlighted in Figure 5.5 and on the corresponding nucleation rate estimates plotted in Figure 5.6 (the rightmost

circle, the rightmost cross, the triangle symbol at the right vertex of the blue shaded region, and the square symbol; see also Table 5.2 for the quantitative values). Consistently with the notation introduced in Section 5.2.7, we call such estimates J_j, with $j = 1, \ldots, 4$.

Let us consider first a specific confidence level, namely $\tilde{C} = 0.95$, because of its broad use in the context of parameter estimation; the associated K-number is $\tilde{\kappa} = 1.358$. For the j-th combined eCDF the left hand side of Eq. (5.13) is calculated as $\Delta F^* = \pm E_j(\tilde{C})$; by solving this equation one obtains the two values $\tilde{\eta}_j^+$ and $\tilde{\eta}_j^-$ hence the following confidence interval:

$$(1 + \tilde{\eta}_j^-)J_j \leq J \leq (1 + \tilde{\eta}_j^+)J_j \quad (j = 1, 2, 3, 4) \tag{5.14}$$

Such confidence intervals are indicated as red error bars in Figure 5.6, while the linear approximation is reported in Table 5.2, last column, as relative error. It is obvious that the amplitude of the confidence interval at a fixed value of the confidence level is controlled by the number of data points, on which the nucleation rate estimate is based. Therefore, such amplitude increases when going from the quadruple 1-2-3-4 to the single series 3, namely from 0.34 to 0.56 in relative terms. In the case of the single series, such amplitude is quantitatively similar to that calculated as the difference between the maximum and minimum nucleation rates estimated from the four single series divided by the average value (about 0.53).

While evaluating the 0.95 confidence intervals is the standard approach, in this work we shall consider also the confidence intervals calculated at the minimum confidence level, at which each combined eCDF is feasible (such approach is meaningless in the case of a single series, where such minimum confidence level is zero). In other words, for the j-th eCDF ($j > 1$) the left hand side of Eq. (5.13) is calculated as $\Delta F^* = \pm E_j(C_j)$ ($j > 1$), where C_j is the mentioned minimum confidence level, which is highlighted in Figure 5.5 and reported in the fifth column of Table 5.2. By solving

Eq. (5.13) one obtains the two values η_j^+ and η_j^-, using which the following confidence interval is obtained:

$$(1+\eta_j^-)J_j \leq J \leq (1+\eta_j^+)J_j \quad (j=2,3,4) \tag{5.15}$$

These confidence intervals are indicated as blue error bars in Figure 5.6, whereas the corresponding linear approximations are reported in Table 5.2, second to last column, as relative errors. It is apparent that the amplitude of these confidence intervals (blue error bars) is smaller than that of the previous ones (red error bars), because the former are calculated at a lower confidence level than the latter. It is also evident that, differently from what observed in the previous approach, the relative amplitude increases when going from the single series 3 to the quadruple 1-2-3-4, namely from 0.22 to 0.30, and such an increase is less pronounced. It is worth noting that in the case of the combined eCDF consisting of two series for instance, i.e. the combination 2-3, the blue error bar is half of the red one.

Before discussing these results any further, it is worth underlining that we are analysing specifically the four experimental series at supersaturation 1.41, but this without any loss of generality. We have in fact observed exactly the same trends with the experimental series measured at all the other supersaturation levels (see Section 5.5 below).

5.4.4 *Choice of an Estimate and of its Confidence Interval*

Let us first underline that the statistical meaning of the confidence interval on an estimated quantity at a given confidence level C is the following: if the same statistical test were repeated many times, i.e. in our case if the series 3 were repeated at the same experimental conditions performing the same number of experiments, and the C confidence intervals were calculated at every repetition, such intervals would bracket the true value of the estimated quantity in a fraction C of the repetitions [142]. In looser terms this definition can in our case be rephrased as follows: the confidence interval of

the nucleation rate estimate obtained in a single repetition brackets its true value with probability C. For the sake of brevity, we will use the latter formulation in the following.

Since the amplitude of the confidence interval increases with increasing confidence level, then reducing both reduces also the range of values attained by the physical properties that depend on the estimated nucleation rate (e.g. the amount of fines in a crystallization process), when the nucleation rate is taken within its confidence intervals. Such a reduction is obtained at the price of a reduced probability that the true value of the nucleation rate belongs to the considered confidence interval and that the corresponding physical property belongs to the associated range of values obtained. In more quantitative terms, since the amplitude of the confidence interval for a given eCDF scales with κ and with the reciprocal of the square root of its number of data points, to halve the 0.95 confidence interval ($\kappa = 1.358$) the confidence level has to be reduced to 0.20 ($\kappa = 0.645$), whatever the underlying number of data points.

Based on the discussion in the previous section, it is clear that when considering a single series, i.e. series 3 in our case, for which $j = 1$, confidence intervals on the estimated nucleation rate for any confidence level C between 0 and 1 can be calculated. The choice of the relevant confidence level C is determined exclusively by the trade off just described, between inclusiveness (high values of C) and strictness in constraining experimental variability (low values of C). The lowest value of C is 0, hence the smallest amplitude is also 0; therefore, only the confidence interval at $\tilde{C} = 0.95$ is plotted in Figure 5.6 (red error bar). It is obvious that larger values of C could be considered, but also that letting C approach 1 would be meaningless considering that in this case the confidence interval would become infinitely large.

In the cases of combined eCDFs instead ($j = 2, 3, 4$), there is a minimum confidence level C_j, below which the j-th combination is not feasible. This yields a minimum amplitude of the confidence

interval (blue error bar in Figure 5.6); between the blue and the red error bars in Figure 5.6 lie all possible confidence intervals for the nucleation rate J_j calculated at confidence levels between C_j and $\tilde{C} = 0.95$. Again the choice of the relevant confidence level C is determined by the trade off just described, but in this case there is a lower bound that cannot be overcome to avoid loss of feasibility of the considered combined eCDF.

5.5 RESULTS OF THE ANALYSIS

5.5.1 *Supersaturation*

The statistical analysis discussed in Sections 5.3 and 5.4 has been applied to all values of supersaturation that we have investigated, namely $S = 1.58, 1.71$, and 1.91, beside $S = 1.41$; either four or five experimental series have been carried out for each supersaturation level yielding results qualitatively comparable to those shown in Figure 5.2. Here, we illustrate and discuss the outcome of the statistical analysis, by plotting the experimental data in terms of diagrams similar to those in Figure 5.5.

As one can see in Figure 5.9, for each supersaturation, our method identifies the minimum confidence level at which two, three, four or five series can be combined; such levels are different for different combinations. Furthermore, such minimum levels vary widely for different supersaturations: they are evenly spaced for $S = 1.41$ and $S = 1.71$, but differ strongly from one another for $S = 1.58$ and $S = 1.91$.

As discussed in Section 5.3.1, the spacing of C_j provides an immediate, qualitative insight on how close and similar the experimental series are with each other. Another important observation concerns the feasibility of combining all series at the standard confidence level $\tilde{C} = 0.95$. Indeed, for the results reported in this work, only $S = 1.41$ could form a combination with all series, since for its quadruple $C_4 < \tilde{C}$; for the remaining three supersaturations, the

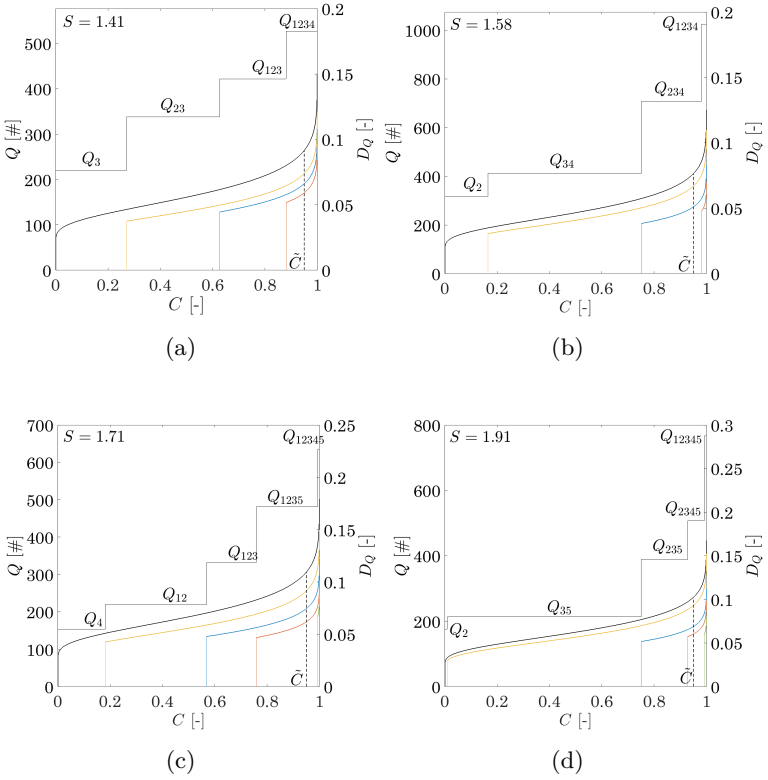


Figure 5.9: Statistical analysis and combination of series at different supersaturation levels. These four plots are obtained and have the same meaning as the plot of Figure 5.5 (which is reproduced here as Figure 5.9a for the sake of comparison with the others).

combination of all distributions can be obtained only for a confidence level larger than the reference value. We recall that the choice $\tilde{C} = 0.95$, albeit commonly used and accepted, bears in itself no specific, privileged statistical meaning. This was originally pointed

out already by Fisher [143], who introduced such concept and recommended to choose the confidence level on a case by case basis, on the base of the knowledge one has of the specific experimental system. Our approach complies with two different requirements: it allows not only to perform a sound statistical analysis of the data, but also to detail clearly and thoroughly all the steps leading to the choice of whether combining different experimental series, thus enhancing the transparency, the portability, and the verifiability of the results.

Finally, we highlight a rather practical point: collecting data at high supersaturation is cumbersome, since often experiments are rejected by the criteria on temperature and concentration discussed in Section 5.2.2 (see Table 5.1). In fact, at high supersaturation crystals can form more likely during the cooling phase, before the system attains the desired crystallisation temperature. Investigating such conditions would thus require several independent repetitions, further stressing the importance of a clear procedure to combine different series of experiments together.

Table 5.1: Effect of supersaturation level on the relationship between the number of experiments and the number of acceptable experiments, as defined in Section 5.2.2.3. All detection times have been measured at 298 ± 0.5 K (see Section 5.2.2). At each supersaturation, the largest combined series feasible at $\tilde{C} = 0.95$ has been considered.

S	Series	Q_{tot}	Q	Q/Q_{tot}
1.41	1-2-3-4	704	527	0.74
1.58	2-3-4	896	728	0.81
1.71	1-2-3-5	800	481	0.60
1.91	2-3-4-5	1072	507	0.47

5.5.2 *Nucleation rates*

After performing the statistical analysis of the experimental series and combining them accordingly, we can estimate for each supersaturation the values of the nucleation rate associated to each series (individual and combined), J_j; these are plotted in Figure 5.10 vs. the confidence level C_j, corresponding to the individual or combined eCDF. Note that the choice of the linear scale for J penalises the visibility of the data at $S = 1.41$, which however are clearly illustrated in Figure 5.6. In light of the results of Section 5.5.1, we can also compute the inherent uncertainty of each nucleation rate, reported in Figure 5.10 as red and blue error bars estimated according to Equations (5.14) and (5.15), respectively (see also Table 5.2 for the exact values).

Even though one can arbitrarily choose a different confidence value $C \geq C_j$, the uncertainties on the nucleation rates will not be smaller than those computed at C_j. No uncertainty is reported for the quadruple at $S = 1.58$ and for the quintuples at $S = 1.71$ and at $S = 1.91$, since these combinations could not be formed at the reference confidence level $\tilde{C} = 0.95$, as highlighted in the previous Section 5.5.1. Note that, for all values of supersaturations investigated, the amplitude of the confidence interval of the nucleation rate decreases by combining more series together, since the number of points M_j increases (see Table 5.2). However, by comparing the uncertainties at different supersaturations, the data appear to be heteroscedastic (i.e. the variance of the data is not constant but depends on the supersaturation) since such amplitude increases with S; this behaviour has been reported also by Xiao et al. [106].

The values of J obtained so far are point-estimates, i.e. each value of S yields one value of J. However, we would rather be able to fit a functional form dependent on supersaturation and temperature, which could also be used for modelling and predicting the behaviour of the system at conditions not yet explored experimentally. Several alternative models have been discussed and repor-

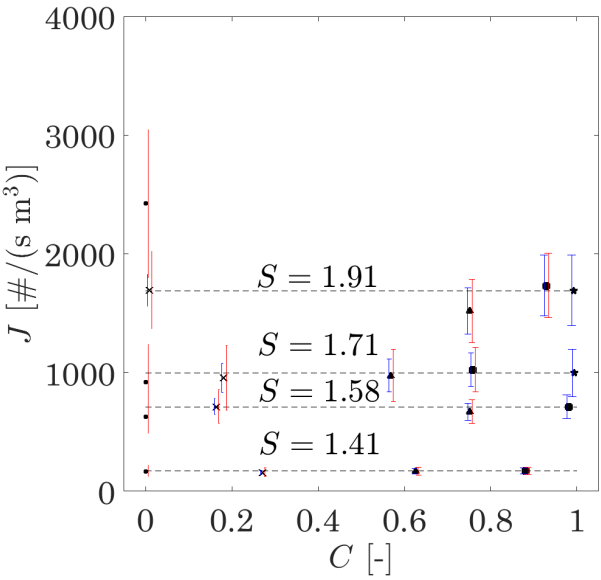


Figure 5.10: Nucleation rate, at different values of S, with the associated asymmetric confidence intervals computed according to Eq. (5.14) and (5.15), represented as blue and red error bars, respectively. The bullets (•), the crosses (×), the triangles (▲), the square (■), and the stars (★) correspond to values of J calculated from the single distributions, the pairs, the triples, the quadruple, and the quintuples, respectively. For each value of S, the dashed line corresponds to the value of J of the distribution combined at the highest value of C. Nucleation rates are plotted vs. the confidence level C_j of the corresponding combination of series. Since the combinations of all series at $S \geq 1.58$ are feasible at a confidence level larger than $\tilde{C} = 0.95$, the red error bars are missing in these cases.

ted in the literature to interpret detection time experiments [59, 85, 106] and we have investigated ourselves this issue in two previous papers [36, 48]. A proper fitting can be based on classical least-squares methods or on criteria derived from the Kolmogorov-

Smirnov statistics [48], but both approaches require nucleation and growth rate models, physical conditions for detection, a proper understanding of secondary nucleation, its role and mechanism. Finally, we are well aware that the system investigated here, likewise others presented in the literature [137, 138], exhibits more polymorphic forms.

While the nucleation rate estimate, in terms of nuclei formed per unit time and volume, has a clear meaning also in the case of polymorphic systems, its interpretation requires to determine which polymorph forms under which conditions, and how stochasticity affects the formation of the different polymorphs. All this goes beyond the scope of this work, and will be the subject of a follow up work.

Table 5.2: Summary of the results obtained by applying the statistical analysis and the procedure to estimate nucleation rates and confidence intervals to the series measured at different supersaturations. The same individual and combined series selected in Figure 5.9 and 5.10 are considered. The corresponding number of points, M_j, and estimated nucleation rate, J_j, are given. Note that the confidence levels are calculated for the sake of simplicity using the linear approximation given by Eq.(5.13) (symmetric intervals); η_j is computed at the minimum feasible confidence level C_j (column 5), and $\tilde{\eta}_j$ at the standard confidence level $\tilde{C} = 0.95$.

S [-]	Series	M_j [#]	J_j [# /(s m^3)]	C_j [-]	η_j [-]	$\tilde{\eta}_j$ [-]
1.41	3	220	167	0	0	0.28
	2-3	338	157	0.27	0.11	0.22
	1-2-3	422	167	0.63	0.13	0.20
	1-2-3-4	527	170	0.88	0.15	0.17
1.58	2	316	628	0	0	0.23
	3-2	412	711	0.16	0.09	0.20
	2-3-4	709	667	0.75	0.11	0.15
	1-2-3-4	1,025	709	0.98	0.14	-
1.71	4	154	920	0	0	0.35
	1-2	261	953	0.18	0.13	0.26
	1-2-3	373	973	0.57	0.14	0.21
	1-2-3-5	481	1,023	0.76	0.14	0.18
	1-2-3-4-5	635	994	0.99	0.20	-
1.91	2	260	2424	0	0	0.26
	3-5	213	1694	0.01	0.08	0.29
	2-3-5	389	1516	0.75	0.13	0.21
	2-3-4-5	649	1744	0.93	0.15	0.16
	1-2-3-4-5	767	1689	0.99	0.18	-

5.6 SUMMARY AND CONCLUSIONS

Estimating the rate of nucleation, i.e. the rate of a stochastic process, from measurements presents challenges foreign to deterministic processes. Even though it is rather obvious that more experiments should result in a better estimate, provided such experiments are consistent with each other, i.e. carried out under exactly the same experimental conditions, it is also clear that any experimental technique will be intrinsically limited in the number of data points that can be collected in a single series of measurements.

Here, we have addressed the issue of when and how multiple experimental series, each yielding a so called eCDF (empirical cumulative distribution function), can be merged to obtain a combined series with a number of data points given by the sum of those of each individual series. We have investigated such problem starting from a well established statistical framework, namely the classical studies by Kolmogorov and Smirnov.

We have then presented few novel contributions to the lively debate and to the growing literature on primary nucleation. First, we have applied an engineered version of a commercial tool to improve its experimental accuracy and to enhance its productivity, since rigorous control of experimental conditions is crucial. Then, we have utilised Kolmogorov and Kolmogorov-Smirnov tests to discuss under which conditions two individual series can be combined and, building on this analysis, we have discussed how an arbitrary number of experimental series can be analysed and possibly combined into one. While avoiding to conduct the whole discussion in terms of standard values of the confidence level and related concepts (e.g. 95% confidence level), we have explored the whole available range of confidence level to enhance our understanding. By doing so, we have introduced a sort of spectrum of the characteristics of the series that we would like to merge and we have shown how informative such tool is. While we have not given any one-fits-all recommendation for the choice of a specific estimate of

the nucleation rate from the several ones that we have presented and discussed, we believe that the tools and the insight provided here will help the researcher, who wants to use them, to make informed and rational choices. In fact, since combining all nucleation time measurement might be misleading, the procedure used for combining individual series must be clearly documented: it is each researcher's choice whether combining the data together and at which confidence level, but the use of the methods illustrated here will enhance transparency and clarity.

We have also discussed how to estimate nucleation rates from the combined series, as well as the trade-off between inclusiveness in using as many original series as possible and strictness in limiting the acceptable experimental variability from one series to the other. Such analysis has been heavily based on an exact relation for the propagation of the uncertainty from the eCDFs to the nucleation rates, which has been derived here. Different functional forms would be obtained of course by applying statistics different from the Poisson distribution, but our methodology would still apply and we believe our main conclusions would not be altered.

Differently from the model validation procedure discussed in Chapter 4, we have not used the nucleation rates obtained at different supersaturation levels to estimate the parameters of a specific nucleation rate expression either. We have made this choice not because we consider this unimportant and in fact, but because we are aware that a corrected determination of nucleation rates of the system studied in this Chapter presents additional challenges, namely the presence of multiple polymorphs. However, we believe that the statistical analysis described and applied here is a prerequisite to estimate nucleation rates at specific operating conditions, independently of the functional form of the nucleation rate on physical properties and on operating conditions. Moreover, estimating the nucleation rate parameters requires making assumptions on or investigating more deeply crystal growth and detection methods and, for the system considered here, its polymorphism. This fea-

ture requires appropriate investigations beyond the determination of the point-wise value of J and it is discussed in further detail in Chapter 6. With respect to all the previous issues, the analysis developed here is, indeed, a necessary prerequisite.

6

STOCHASTIC NUCLEATION OF POLYMORPHS - EXPERIMENTAL EVIDENCE AND MATHEMATICAL MODELLING

6.1 INTRODUCTION

In addition to the difficulties and issues discussed in the Sections above, for substances that exhibit different polymorphs, there is the need of establishing which polymorph has been formed and of attributing the estimated nucleation rate to the appropriate polymorph, which is challenging. The Ostwald's rule of stages, stating that crystals of the least stable polymorph nucleate first and then possibly transform via a solution-mediated mechanism towards more stable forms, was believed to apply [145, 146], while it was understood that in a solution at a specific solute concentration the polymorph that forms first is that with the largest nucleation rate among those for which the solution is supersaturated, i.e. often, but not always, the least stable polymorph [147–149]. In a deterministic framework, both interpretations take for granted that at a given supersaturation and temperature there is one polymorph that always forms first in repeated nucleation experiments.

Considering nucleation in a stochastic framework leads to questioning the conclusion above. If nucleation times are distributed, even if one polymorph has a shorter average nucleation time than another, it might well be that in a specific experiment the nuclea-

The results presented in this Chapter have been reported in: G. M. Maggioni, L. Bezinge, M. Mazzotti, "Stochastic Nucleation of Polymorphs: Experimental Evidence and Mathematical Modeling", *Cryst. Growth Des.* **2017**, *17*, 6703–6711 .

tion time of the former is longer than that of the latter; hence in that specific experiment, the generally slow-nucleating polymorph nucleates before the generally fast-nucleating polymorph does. Considering that the average nucleation time of a polymorph and the variance of the distribution of nucleation times scale with the reciprocal of the nucleation rate and of the system size, such effects are supposed to be more or less observable depending on the experimental conditions (supersaturation and temperature) and on the size of the crystalliser. Furthermore, it could also happen that the polymorph that forms first transforms into a more stable one before being detected (or before being collected at the end of the experiment for off-line analysis).

Experimental evidence concerning the stochastic nature of nucleation of polymorphs has been reported in recent years [150–152]. Among others, Kulkarni et al. [61, 138] and Caridi et al. [153] studied the nucleation of isonicotinamide (INA), of which at least three polymorphs are known, in various solvents and solvent mixtures. They reported that when INA is in ethanol (EtOH), methanol (MeOH), and 2-propanol (IPA) only the most stable form (Form II, in their work) nucleated, whereas in nitrobenzene and nitromethane the Form IV and the Form I nucleated, respectively. Kulkarni and co-workers [154] studied also the crystallisation of INA in two solute mixtures: in the case of ethanol/nitrobenzene, they reported the formation of both Form II and Form I, while in the case of ethanol/nitromethane of both Form II and Form V. These authors observed that only one polymorph formed when crystallisation was performed in small volumes (3 mL). The polymorph forming in each individual experiment appeared to be random, while the occurrence frequency of the polymorphs depended on the initial concentration. On the contrary, in large volume experiments they observed solid mixtures of multiple polymorphs. In the same work [154], Kulkarni et al. reported that also 4-hydroxyacetophenone in ethyl acetate yielded different polymorphs in different experiments. In this system, though, liquid-

liquid phase separation was reported, which may admittedly have affected the nucleation process. Hansen et al. [150] investigated the crystallisation of a non-steroidal drug, piroxicam, in acetone and acetone/water mixtures with several additives. They reported that piroxicam crystallised either Form I, or Form II, or a mono-hydrate crystal, in a random way and with an occurrence frequency depending on the initial solute concentration. To interpret their data, Kulkarni et al. [154], as well as Hansen et al. [150], postulated the stochastic formation from the clear solution of a first single nucleus, dictating the evolution of the system in small volumes. In larger systems, a more conventional poly-nuclear mechanism would act, resulting in mixtures of multiple solid forms. The formation of different polymorphs at the same conditions randomly, rather than in the deterministic way predicted by the classical theory, have also been recently reported for other systems [50, 155, 156], such as p-ABA in water/EtOH mixtures [151, 152], L-glutamic acid in water [156], and m-hydroxibenzoic acid in 1-propanol [95]. Furthermore, Cui et al. [157] have observed that either form α, or form β, or a mixture of the two crystallised on a functionalised substrate via contact secondary nucleation, which is an activated mechanism [158] similar to primary nucleation. It is also worth noting that Hansen et al. [137] in a recent study have further investigated the crystallisation of INA in different solvents (MeOH, acetonitrile, acetone, ethyl acetate, and dichloromethane) and have concluded that the polymorphism of INA depends not only on the solvent chosen, but also on the concentration and on the temperature at which crystallisation is carried out.

Recently, the possibility of modelling these phenomena has been demonstrated using an approach based on the description of nucleation as a Poisson process [159, 160]. By expanding on the results that we have presented above (see in particular Chapters 2), the stochastic nucleation of polymorphs is studied both experimentally (crystallising INA in ethanol) and theoretically. The Chapter is structured as follows: experimental methods and experimental

results are reported in Sections 6.2 and 6.3; the stochastic model is developed, discussed, and utilised in Section 6.4; conclusions are drawn in Section 6.5.

6.2 EXPERIMENTAL

6.2.1 *Materials*

Isonicotinamide (INA), purity ≥99%, was purchased from Sigma-Aldrich; ethanol (EtOH) from Fluka (Ethanol Laboratory Reagent, absolute, ≤ 99.5%) and EMD Millipore (Ethanol Absolute For Analysis Emsure Acs, Iso, Reag. Ph. Eur.). Prior to crystallisation experiments, the solution has been filtered through 13 mm syringe filters (PTFE Hydrophobic, non-sterile, Pore Size 0.22 µm); INA was used as received. Different polymorphs of INA are known and their nomenclature in the literature [137, 138] appears somewhat confusing. In this work, we have labelled the three polymorphs in order of decreasing stability at 298 K in EtOH as α, β, and γ. They correspond to the forms labelled as EHOWIH01, EHOWIH02, and EHOWIH04 in the CSD [137, 138], respectively. The solubilities of the polymorphs have been measured gravimetrically and their values are reported in Table 6.1. We have assumed the ratio of activity coefficients $\gamma_a(T, c)/\gamma_a(T, c_s) \approx 1$, hence the supersaturation S_j for each polymorph has been expressed as:

$$S_j(T, c) = \frac{c}{c_{s,j}(T)} \qquad j = \alpha, \beta, \gamma \tag{6.1}$$

where c is the concentration of INA in EtOH and $c_{s,j}$ the solubility of the polymorph j, expressed in grams of solute per kilogram of solvent [g/kg$_s$]. Unless explicitly indicated otherwise, in the following, with the generic term "solubility" and "supersaturation", we refer to those of the stable polymorph, i.e. $c_{s,\alpha}$ and S_α. The crystallisation experiments have been performed in a range of supersaturations between $S_\alpha = 1.10$ and $S_\alpha = 2.10$.

Table 6.1: The values of the solubilities $c_{s,j}$, of the three polymorphs of INA in EtOH, measured gravimetrically at different temperatures.

T [K]	$c_{s,\alpha}$	$c_{s,\beta}$ [g/kg$_s$]	$c_{s,\gamma}$
293.15	79.02 ± 0.60	80.69 ± 3.00	85.22 ± 2.20
298.15	93.41 ± 1.00	95.75 ± 3.20	100.6 ± 2.20
303.15	-	115.1 ± 1.90	117.1 ± 4.05
308.15	-	134.4 ± 3.00	-
310.15	138.4 ± 0.70	-	-
316.65	170.1 ± 0.30	-	-

6.2.2 *Methods*

Cooling crystallisation experiments of INA in EtOH were performed at a constant temperature of 298.15 K; due to the accuracy of the thermocouples, this temperature as well as the other temperatures reported in the following, are accurate within 0.5 K. The crystallisation experiments were performed with an amount of solvent equal either to 1.40 g (small size experiments) or to 50 g (large size experiments), the former in a Crystal16 (Technobis Crystallization Systems, multiple reactor set-up with 16 wells, each equipped with an independent thermocouple [115]), and the latter in an EasyMax 400 (Mettler Toledo). Both devices were equipped with glass reactors; for the Crystal16, standard HPLC glass screw topped vials were used as vessels. As discussed elsewhere, the reactors were properly sealed during the experiments to minimise solvent evaporation, since this would change the solute concentration, hence the supersaturation [115]. We have also monitored the

evaporation by weighing the reactors (individual vials in Crystal16 and single glass reactor in EasyMax) before and after each experiment: if evaporation resulted in a change of supersaturation larger than 2%, the experiment was discarded and repeated. The vials in the Crystal16 were stirred with magnetic bars at 700 rpm, while the reactor in the EasyMax with a 4-blade impeller at 400 rpm, to guarantee that the system was well-mixed. The stock solution was kept for 60 minutes at 333 K to ensure the complete dissolution of INA and immediately thereafter used to fill the vials. For the large scale experiments, the solution was prepared directly inside the glass reactor.

Supersaturation was induced by cooling from different saturation points, and multiple cycles were performed for both small and large size experiments. A crystallisation cycle consisted of 4 stages: a heating ramp from 298.15 K to 333.15 K (2.0 and 2.5 minutes, for Crystal16 and EasyMax, respectively), a dissolution step at 333.15 K (60 minutes, both devices), a cooling ramp from 333.15 K to 298.15 K (2.0 and 2.5 minutes, for Crystal16 and EasyMax, respectively), and finally an isothermal step at 298.15 K (its duration varied from 10 to 180 minutes, depending on the supersaturation and the system size). The differences in the heating and cooling stage depended on the different configuration and controller of the devices used.

6.2.3 *Measurement protocols*

The polymorphism of the crystals produced in the experiments was assessed off-line, ex situ through XRD measurements (experiments in Crystal16 and EasyMax) and on-line, in situ through Raman spectroscopy (RA 400 Raman spectrometer from Mettler Toledo, for the experiments in the EasyMax). XRD and Raman spectra of the pure polymorphs are shown in Figure 6.1 and 6.2, respectively.

Three characteristic peaks for each polymorph have been highlighted in Figure 6.1 as vertical bands in blue ($2\theta = 17.8, 23.0, 26.9$),

red ($2\theta = 14.5, 23.9, 25.9$), and yellow ($2\theta = 30.9, 32.2, 37.0$), for α, β, and γ, respectively. For the experiments in the Crystal16, we measured the XRD spectra only of the crystals produced during the very last cycle of a series of experiments using the same solution.

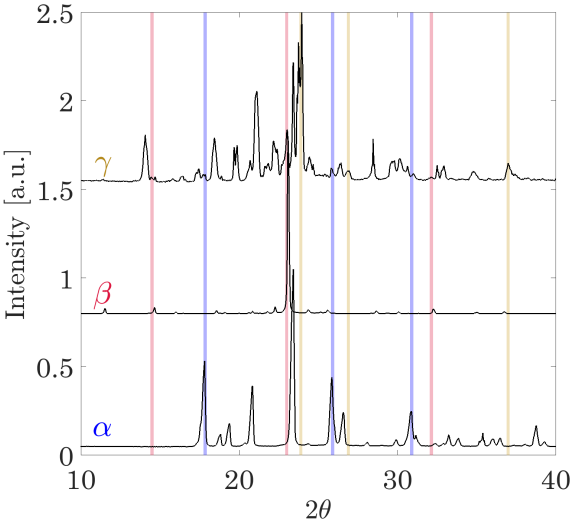


Figure 6.1: Characterisation of the polymorphs. The XRD spectra of the powder of α, β, and γ; the bands in blue, magenta, and yellow, highlight the characteristic peaks of the three polymorphs, respectively.

The Raman spectra of the pure forms in EtOH are reported in Figure 6.2 and are consistent with those given by Kulkarni et al. [138]. The three polymorphs can be clearly distinguished in two intervals of the Raman spectrum, namely $980 - 1010$ cm^{-1} and $1200 - 1230$ cm^{-1}, as highlighted in Figure 6.2. Raman spectra were collected in the range of Raman shift from 100 to 3600 cm^{-1}, with a resolution of 1 cm^{-1} and averaged over 10 scans using an exposure time of 6 s, corresponding to one measurement per minute. The raw data from

Raman measurements have been treated according to standard procedures for spectroscopic data pre- and post-processing: the signal has been smoothed with a Savitzky-Golay filter [161] and a linear baseline correction [162] has been applied to the spectra. In the EasyMax, at least nine cycles for each supersaturation S_α have been performed, thus obtaining a number of repetitions comparable with those (sixteen) obtained from the last cycle of the experiments in the Crystal16. We used Raman spectroscopy also to verify that the system attained unbiased initial conditions during the dissolution step after each crystallisation step. The complete disappearance of the peaks associated to the solid forms of INA within 5 minutes from the beginning of the dissolution stage and the reproducibility of the Raman spectra of the clear solution at high temperature in different cycles indicated that the initial conditions of the system were the same at the beginning of each crystallisation experiment.

Several XRD spectra measured from samples of small scale experiments were markedly different from those of pure polymorphs. Since solvates of INA have been reported for crystallisation from water [163], we have taken care of excluding the presence of solvates or solvent inclusions by means of two different tests. The first test consisted in repeating the XRD measurements on the same samples after 5 days, during which the crystals were kept in a dry environment: all spectra were unchanged. The second test was a thermal gravimetric measurement, with temperature up to 500 K, i.e. well beyond the melting temperature of any possible solvent, of all known polymorphs and of any solvates. The thermal gravimetric analysis showed no loss of material, hence indicating that no ethanol was trapped in the solids and also pointing out that the spectra exhibiting characteristic peaks of two or three polymorphs were originated by a solid mixture of pure crystals (i.e., a mixture of polymorphs).

Detection times, t_D, were also recorded during all experiments (see Figure 6.3). For the Crystal16 experiments, the detection time was defined as the difference in time between the attainment of

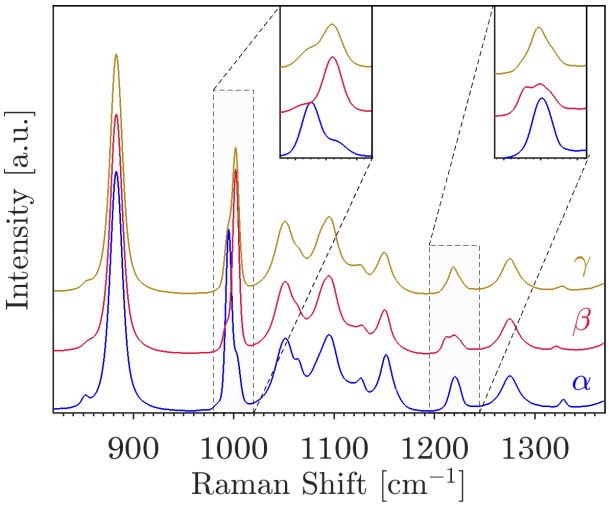


Figure 6.2: Characterisation of the polymorphs. The Raman spectra of pure α, β, and γ in ethanol, in blue, red, and yellow, respectively.

the isothermal temperature and the decrease of light transmissivity below a given threshold. Following the same protocol discussed in detail elsewhere [115], we retained only the detection times in the small scale experiments whose temperature was within a band of ± 0.5 K from the set value of 298.15 K.

For the EasyMax experiments, the detection time was defined as the difference in time between the attainment of the saturation point and the time when a characteristic peak of Ethanol (at wavelength 883 cm^{-1}) decreased below a threshold value with respect to the peak maximum in the dissolved state before cooling (see Fig. 6.3).

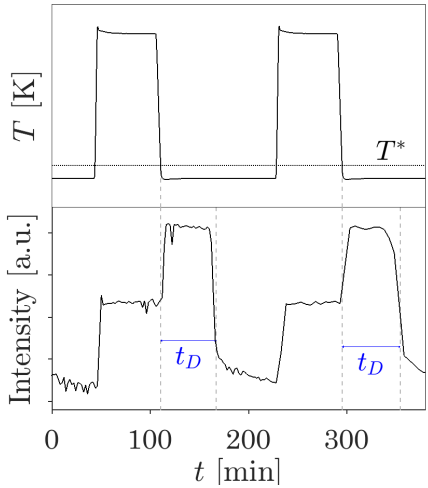


Figure 6.3: An illustration of detection times based on the temperature profile (top) and on the Raman intensity I (bottom), measured for a peak typical of ethanol, for the case with $S = 1.30$. T^* indicates the saturation temperature of the most stable form.

We note that nucleation (and detection) can occur also during the cooling ramp, as soon as $S_j > 1$: this can be more likely observed in systems at high supersaturations and/or of large size, since the average nucleation time is inversely proportional to the system size and the nucleation rate, the latter increasing with increasing S_α [36, 47]. Indeed, in the large size experiments detection occurred always during the cooling ramp for all experiments with $S_\alpha > 1.40$. A direct quantitative comparison between the values of the detection times measured in the Crystal16 and those measured in the EasyMax is not possible, since different detection techniques and detection conditions were applied in the two devices.

6.3 EXPERIMENTAL RESULTS

We have performed two sets of nucleation experiments, which provide different, complementary pieces of information. The Crystal16 allows, on the one hand, to gather a large number of detection time data, but cannot monitor which polymorph is associated to such events and the XRD measurements can be obtained only from the samples collected after the final cycle of a series. On the other hand, the EasyMax has a much lower productivity in terms of number of experiments, but it allows to measure online the time-resolved Raman spectra.

6.3.1 *Detection Times*

Figures 6.4 (Crystal16) and 6.5 (EasyMax) illustrate, for each experimental supersaturation (horizontal axis) the detection times measured in tens or hundreds of repeated experiments (vertical axis). The detection times are clearly distributed and their spread reflects the stochastic nature of primary nucleation, already reported for several systems [37, 46, 47, 59, 85, 93]. In Figure 6.4, t_D varies from a minimum of about 1 min to a maximum of about 5 hours ($S_\alpha = 1.20$), and from a minimum of about 1 minute to a maximum of about 30 minutes ($S_\alpha = 2.10$). In Figure 6.5, for the same supersaturation values, the maximum t_D decreases to about 10 and 4 minutes, at $S_\alpha = 1.20$ and $S_\alpha = 2.10$, respectively.

The sharp decrease of the broadness of the distribution observed in the data when increasing the supersaturation, from $S_\alpha = 1.20$ to $S_\alpha = 2.08$, or the system size, from 1.40 to 50 g of solvent, is also consistent with the statistics associated to the measurement of stochastic processes [47, 59].

Since nucleation is stochastic, the detection time t_D is a stochastic variable (as illustrated in Figures 6.4 and 6.5), distributed according to an intrinsic underlying cumulative probability function, $F(t_D)$. The function F can be estimated from the experiments

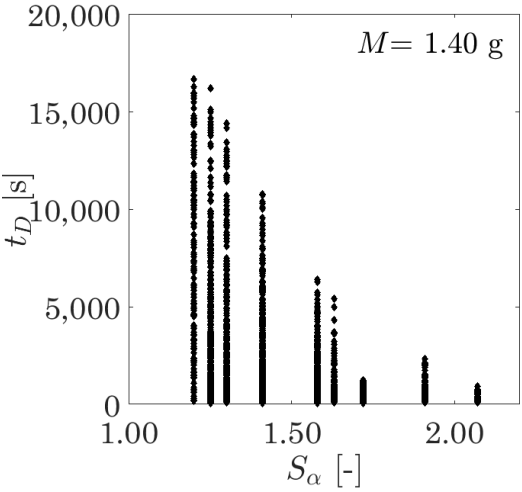


Figure 6.4: The detection times measured from the experiments in the Crystal16 at different supersaturation levels. M is the mass of ethanol used in the experiments.

through the empirical cumulative distribution F^*, as discussed in Section 4.2.

6.3.2 *Polymorphic Form*

Figure 6.6 (Crystal16) and Figure 6.7 (EasyMax) demonstrate that, in repeated experiments carried out at exactly the same conditions starting from identical initial solutions, different polymorphs are obtained first.

This is illustrated in Figure 6.6 through XRD spectra of the final crystals obtained at the same supersaturation. In experiment E1 the β polymorph was obtained first and in experiment E4 polymorph α, whereas in experiments E2 and E3 the β and γ polymorphs and the α and β polymorphs, respectively, are both present.

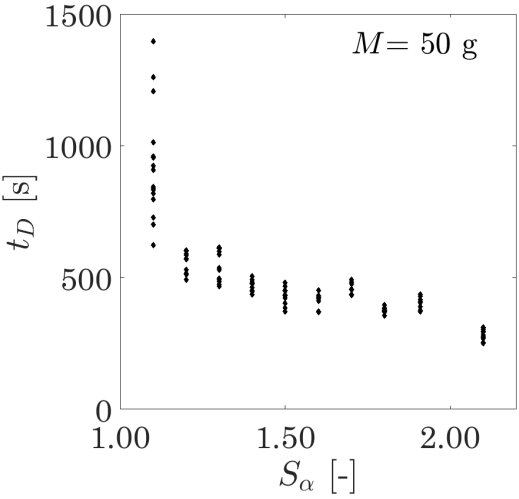


Figure 6.5: The detection times measured from the experiments in the EasyMax at different supersaturation levels. M is the mass of ethanol used in the experiments.

Figure 6.7 shows at each supersaturation the relative occurrence Q of the three polymorphs (i.e., the number of experiments in which a specific polymorph was observed, as a percentage of all experiments performed at that specific supersaturation level), as obtained from the Raman spectra both at detection (large vertical bars) and at the end of the experiment (thin vertical bars).

Two remarks are worth making. First, the distribution of relative occurrences changes markedly from one supersaturation to the next; we will provide a justification thereof in Section 6.4, based on the different supersaturation dependence of the nucleation rate of the different polymorphs. Secondly, the distribution of relative occurrences changes from the time of detection to the end of the experiment, because of polymorph transformation, as discussed in Section 6.3.3.

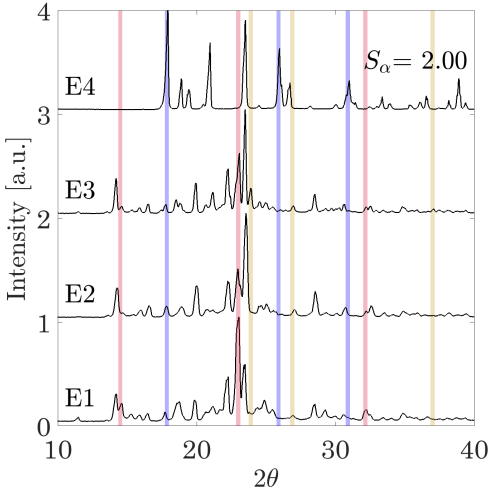


Figure 6.6: XRD spectra at $S_\alpha = 2.00$ obtained from four different vials (labelled as E1, E2, E3, E4) of the same typical nucleation experiment. Depending on the vial analysed, either one of the three forms (or possibly a mixture) can be obtained. The colours of the vertical bands associate to the peaks typical of each polymorph, as in Figure 6.1: blue, red, and yellow refer to α, β, and γ, respectively.

Figure 6.8 illustrates the distribution of detection times as observed in the same experiments, by plotting the corresponding empirical cumulative distribution functions (eCDFs).

Figure 6.8a shows the eCDFs (as black lines) of the data obtained in the Crystal16 at $S_\alpha = 1.30$, 1.58 and 1.91, as computed using Eq. (4.4); for each supersaturation level, the experiments for which the XRD spectra were also measured are represented with a coloured symbol, in blue, red, or yellow to indicate that polymorphs α, β, or γ, respectively, has nucleated first. It is apparent that

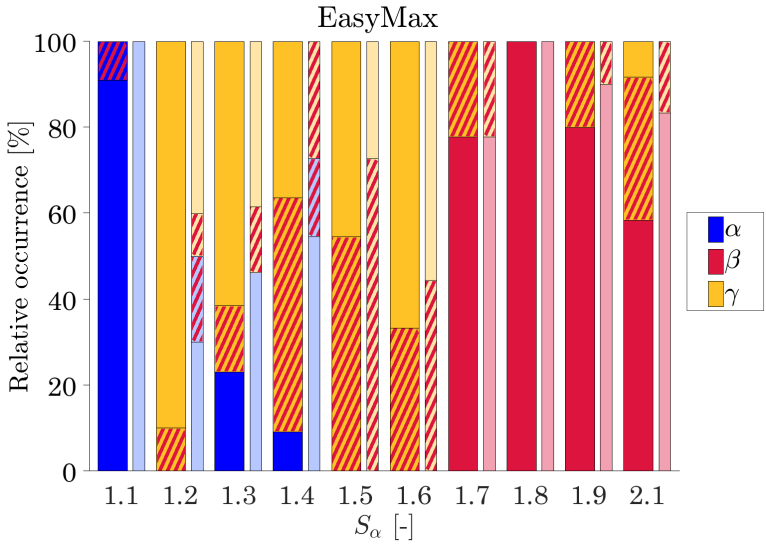


Figure 6.7: The percentage occurrence frequency obtained from identifying the polymorph nucleating with Raman spectroscopy; the larger bars indicate the occurrence frequency (as percentage Q) at detection, while the smaller bars with lighter colours the frequency at the end of the cycle. The colours follow the usual code, with blue, red, and yellow for α, β, and γ, respectively. The striped bars in Figure 6.7 indicate the experiments in which two different polymorphs were observed together.

all polymorphs have distributed detection times, and any among the three can form first.

Figure 6.8b shows the eCDFs of the data measured in the EasyMax experiments at $S_\alpha = 1.30$ and 2.10, after analysing the Raman spectra at detection. The black eCDF is that of the detection times, when ignoring which specific polymorph was detected, i.e. the same type of curve plotted in Figure 6.8a (but with a much smaller number of data points).

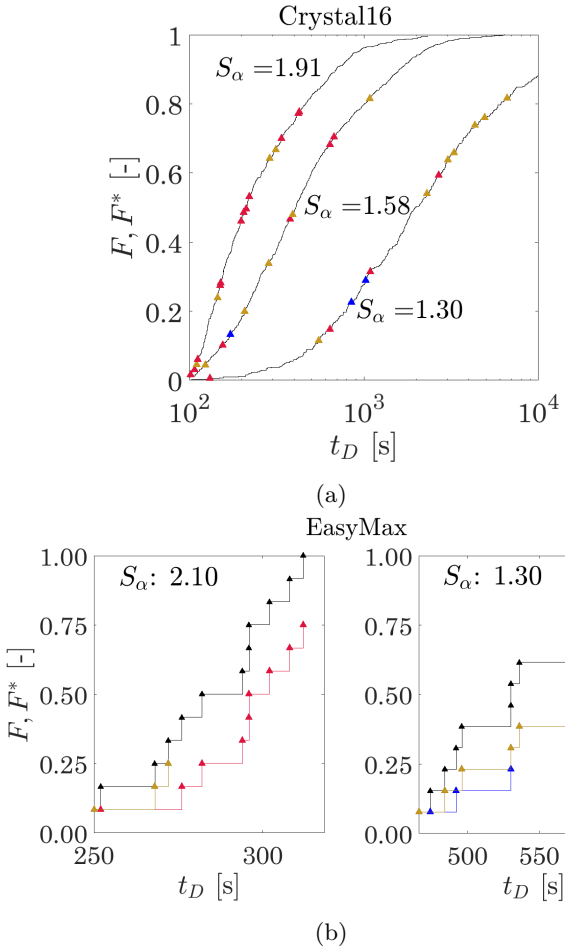


Figure 6.8: (a) The eCDFs for experiments at $S = 1.30$, 1.58, and 1.91 in the Crystal16; the symbols represents the samples for which the XRD spectra have been measured. The blue, red, and yellow symbols refer to α, β, and γ, respectively. (b) The eCDFs for $S = 1.30$ and 2.10 obtained from identifying the polymorph nucleating with time-resolved Raman spectroscopy in the EasyMax, where a much smaller number of points was available.

The coloured eCDFs are those of the individual polymorphs, where the horizontal coordinate, i.e. the detection time, is the same as that of the corresponding point on the black eCDF, whereas the vertical coordinate, corresponding to the cumulative probability, attains a smaller value because it is calculated on the data points of that polymorph only.

6.3.3 *Solid Phase Transformation*

Figure 6.7 provides evidence that new crystals, after forming in a certain polymorphic form, whose relative occurrence depends on supersaturation, may transform into a more stable form, possibly through a solvent mediated transformation that has been well documented by several other studies [138, 151], including a few from our group [164, 165]. While Figure 6.7 provides information about the polymorphic form at two points in time, namely detection and experiment end, and proves that less stable forms may evolve into more stable ones, time resolved Raman spectra, as those shown in Figure 6.9, allow to monitor the dynamics of the whole transformation with great clarity.

More specifically, Figure 6.9a shows the transformation from polymorph γ to α occurring at S_α =1.30, and Figure 6.9b shows that from polymorph γ to β occurring at $S_\alpha = 1.90$; in both cases seven spectra are shown, namely from detection to 30 min after that, with each spectra taken five minutes after the previous.

A few remarks are worth making. First, the transformation in Figure 6.9a is possibly not yet completed, as the characteristic peak of polymorph γ is still visible in the last spectra recorded. Secondly, in the case of Figure 6.9b the transformation to the most stable form α is not visible, but it indeed occurred as we could observe after six hours during which the suspension had been left undisturbed. Finally, as we observe in Figure 6.9b that the transformation

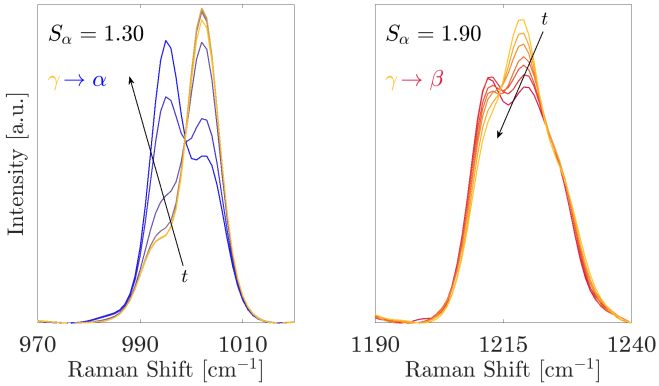


Figure 6.9: The transformation of γ polymorph, monitored over 30 minutes. (a) A direct transformation of γ into α at low supersaturations. (b) A transformation from γ into β, at high supersaturations. Typically, the β form is then stable for several hours, but transforms ultimately into the α form.

to the most stable form may take very long, we also conclude that when the most stable polymorph α is present at detection (see measurements at low supersaturation levels in Figure 6.7), it has most likely formed via primary nucleation and not through solid phase transformation from another metastable polymorph (which would have taken too long to occur in the short time before detection).

6.3.4 *Discussion of Experimental Result*

The experiments in the Crystal16 and in the EasyMax provide complementary information and allow to identify the main features of the nucleation process, particularly the effect of the system volume and of the solute concentration. The experimental data clearly show not only that detection times are statistically distributed (see Section 6.3.1, Figures 6.4 and 6.5), but also that the

type of polymorph nucleating in each experiment is itself a stochastic property (Section 6.3.2, Figures 6.6 and 6.7). Moreover, the data indicate that the relative occurrence of the different polymorphs depends strongly on the solute concentration (see Figure 6.8b). Therefore, a proper understanding of this non-deterministic behaviour requires analysing the experiments in the context of statistical models of nucleation [5, 37, 46, 59].

Stochastic models of nucleation for non-polymorphic systems describe the nucleation time, t_N, i.e. the time when the first primary nucleus forms, with $S > 1$, as a a stochastic quantity. Even though in some cases crystal growth has been found to be itself stochastic [104, 166, 167], primary nucleation is considered to be the major source of stochasticity during crystallisation from a clear solution, whereas growth can be considered to be deterministic, and can be modelled accordingly. Unfortunately, nucleation times t_N cannot be measured directly, because the nuclei are too small for being observed [36, 46–48, 59, 85]. Consequently, one measures t_D and, using a model, retrieves t_N, from which then the nucleation kinetics can be estimated. Since nucleation times are distributed [37, 59, 85], also detection times are distributed. A thorough and comprehensive discussion about the estimation of t_N from t_D and the necessary assumptions behind it can be found elsewhere [36, 48, 59, 85, 106].

6.4 MODELLING THE STOCHASTIC NUCLEATION OF POLYMORPHS

6.4.1 *Mathematical Model*

Primary nucleation has successfully been modelled as a Poisson process [36, 46–48, 59, 61, 85, 93, 96], which is the mathematical formalism describing the occurrence of rare events [56], e.g. activated phenomena such as chemical reactions and nucleation [168]. It is often assumed that only the first nucleus forms stochastically at

the time t_N, whereas the later nuclei – primary and secondary alike – form deterministically [36, 46–48, 59, 106]. For the sake of simplicity, but without loss of generality, we consider here only unseeded isothermal crystallisation [36, 46, 47]. Note that the primary nucleation rate J is constant in an isothermal, closed system with no crystal and no chemical reactions, since the solution composition does not change until the first nucleus forms. By assuming the first nucleation event as the only source of stochasticity, we make use of Model 1 in Chapter 2 as reference nucleation model. Clearly, this is a rather strong hypothesis, whose validity for non-polymorphic systems we discussed previously [36]. This hypothesis, however, allows to build a model conceptually simple, computationally affordable, and semi-quantitatively accurate.

Let us now extend the stochastic model of nucleation to a system with an arbitrary number of polymorphs $(\alpha, ..., \pi)$. Such a system can be thought of as occupying discrete states: an empty state with no nuclei (which is also the initial state of the system) and a state where a nucleus of polymorph j has formed $(j = \alpha, ..., \pi)$; hence, the presence of a nucleus of polymorph j represents a possible state different from the empty state. The (random) appearance of the first nucleus causes the irreversible transition of the system from the original empty state to the new state. The functions $P_0(t)$ and $P_j(t)$ represent the probabilities that, at time t, the system is still in the initial state and that it is in the state j, respectively. Note that the system must be in one of the possible states, therefore:

$$P_0(t) + \sum_{j=\alpha}^{\pi} P_j(t) = 1 \tag{6.2}$$

and that the probability $P(t)$ that a nucleus of any polymorph has formed, irrespectively of which one, is by definition of $P_0(t)$ simply:

$$P(t) = 1 - P_0(t) = \sum_{j=\alpha}^{\pi} P_j(t) \tag{6.3}$$

The evolution of $P_0(t)$ and $P_j(t)$ can be described in terms of the rates of transition between the possible states of the system. The rate of change of the probability $P_j(t)$ (state with one nucleus of polymorph j) is given by the probability $P_0(t)$ (state without nuclei) times the transition rate λ_j, while the rate of change of P_0 is minus the product of $P_0(t)$ times an overall transition rate $\hat{\lambda}$:

$$\begin{cases} \dfrac{dP_0}{dt} = -\hat{\lambda} P_0 & (6.4) \\ \dfrac{dP_j}{dt} = \lambda_j P_0 \qquad \forall j = \alpha, ..., \pi & (6.5) \end{cases}$$

The corresponding initial conditions are:

$$\begin{cases} P_0(0) = 1 & (6.6) \\ P_j(0) = 0 \qquad \forall j = \alpha, ..., \pi & (6.7) \end{cases}$$

By summing term by term all Eqs. (6.4) and (6.5), and employing the conservation of total probability given by Eq. (6.2), one obtains:

$$\hat{\lambda} = \sum_{j=\alpha}^{\pi} \lambda_j \tag{6.8}$$

which is a physically sensible result. Based on physical arguments [46], λ_j is the product of the primary nucleation rate of the polymorph j, J_j, and the system size, V, i.e. $\lambda_j = J_j V$ $(j = \alpha, ..., \pi)$, hence $\hat{\lambda} = V \sum_{j=\alpha}^{\pi} J_j = V\hat{J}$, where $\hat{J} = \sum_{j=\alpha}^{\pi} J_j$ represents an apparent nucleation rate and corresponds to what an observer incapable of distinguishing the different polymorphs would measure. Solving Eqs. (6.4) and (6.5), using also Eq. (6.3), yields:

$$P_0(t) = \exp\left(-\hat{\lambda} t\right) \tag{6.9}$$

$$P(t) = 1 - \exp\left(-\hat{\lambda} t\right) \tag{6.10}$$

$$P_j(t) = \frac{\lambda_j}{\hat{\lambda}} P(t) \qquad \forall j = \alpha, ..., \pi \tag{6.11}$$

Note that $P_j(t)$ is the so-called joint probability, i.e. the probability that at time t a nucleus has formed and it is of polymorph j. Dividing P_j by the probability P that a nucleus of any type has formed, yields the occurrence probability ξ_j:

$$\xi_j = \frac{P_j(t)}{P(t)} = \frac{\lambda_j}{\hat{\lambda}} = \frac{J_j}{\hat{J}} \qquad \forall j = \alpha, ..., \pi \tag{6.12}$$

which is the conditional probability, namely the probability that, at any time, the nucleus which has formed belongs to polymorph j, given that a (first) nucleus of any type has formed. Note also that, for a homogeneous system such as the one considered here, ξ_j is a function only of the nucleation kinetics.

Figure 6.10 illustrates the results obtained above for the case with three polymorphs, namely $j = \alpha, \beta, \gamma$. The black solid line represents the cumulative probability function given by Eq. (6.10), while the blue, magenta, and yellow line, for form α, β, and γ, respectively represent the cumulative probability functions given by Eq. (6.11). The probability P, associated to $\hat{J}$, behaves identically to the cumulative probability of a standard isothermal Poisson process and, as expected, asymptotically approaches 1 for $t \to \infty$. On the contrary, $t \to \infty$ the probabilities $P_j(t)$ do not approach 1, but the asymptotic value ξ_j, thus reflecting the fact that a polymorph can or cannot form. Figure 6.10 also shows clearly that, as long as the value of $\hat{J}$ is constant (see Eq. (6.12)), the occurrence probability of the polymorphs does not depend on the nucleation time itself, as indicated by Eq. (6.12): to change such probability, one must alter the nucleation kinetics of the polymorphs, for instance by operating at a different supersaturation, or at a different temperature.

It is worth highlighting three points. First, according to Eq.(6.10) all polymorphs contribute to the total nucleation probability, $P(t)$, through the individual intensities λ_j, hence ultimately through J_j. Second, the polymorphs associated with higher nucleation rates are more likely to nucleate, but in general they do not necessarily

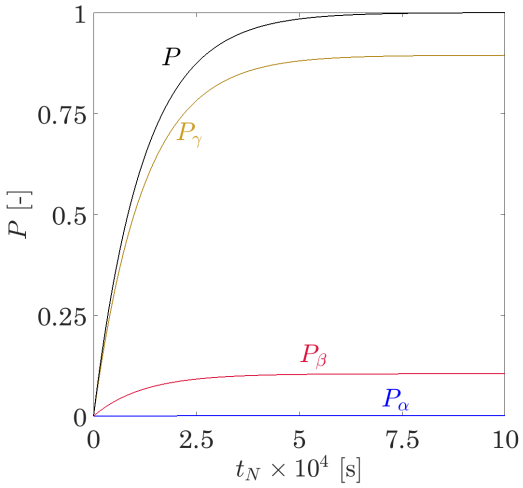


Figure 6.10: The cumulative probabilities for a system with three polymorphs in a system with $V = 1.8$ mL. The black, blue, red, and yellow lines refer to the probability of forming any polymorph, α, β, γ, respectively, and are calculated for the nucleation rates $\hat{J} = 490$, $J_\alpha = 4.30 \times 10^{-4}$, $J_\beta = 61.4 \times 10^{-3}$, $J_\gamma = 0.525$ [#/(kg s)]. The values of the nucleation rates have been computed using Eqs. (6.13) and (6.14) with the parameters reported in Table 6.3, at $T = 298.15$ K and $S_\alpha = 1.50$.

nucleate first in each individual repetition of an experiment: each polymorph has non-zero (albeit possibly very small) probability to nucleate, unless its supersaturation is below 1. Third, the system size V influences the nucleation probability P, i.e. the nucleation time, but does not affect the occurrence probability ξ_j. Thus, which polymorph nucleates depends on the conditions of the system, but not on the system size. Finally, we can use the model to investigate the influence of the operating parameters, namely the solute concentration and the system temperature.

6.4.2 *Comparison between the Model Results and the Experiments*

In this section, we focus on studying the effect of the solute concentration on the system; since experiments were isothermal, we equivalently look at the influence of the supersaturation, S_α. The primary nucleation rate of each individual polymorph is assumed to follow the expression of the Classical Nucleation Theory [2–4], Eq. (3.25) in Section 3.3.3:

$$J_j(c,T) = S_j(c,T) A_j \exp\left(-\frac{B_j}{T^3 \ln^2 S_j(c,T)}\right) \qquad j = \alpha, \beta, \gamma \tag{6.13}$$

At constant temperature T, the pseudo-rate $\hat{J}$ is then a function of the actual solute concentration c (through the supersaturation, Eq. (6.1)) and of all individual nucleation rates of α, β, and γ:

$$\hat{J}(c) = J_\alpha(c) + J_\beta(c) + J_\gamma(c) \tag{6.14}$$

Equations (6.13) and (6.14) show that such model needs six parameters to describe the INA/EtOH isothermal system, namely the pair (A_j, B_j) for each polymorph, with $j = \alpha, \beta, \gamma$. Determining these six parameters is not trivial; two points are worth mentioning. First, when no information about the specific polymorph forming is available, such as during detection time experiments in the Crystal16, the empirical cumulative distribution function obtained from experiments allows to estimate only the value of $\hat{J}$ at the chosen supersaturation level. Second, a proper estimation of the parameters would require the characterisation not only of crystal growth, secondary nucleation, and dissolution, but also of the detection technique itself [48, 114]: such studies go beyond the scope of this contribution. For these reasons, even though we have selected the parameters (see Table 6.3) to run our simulations based on a fitting procedure, we consider such values appropriate only for semi-quantitative studies.

The parameters used to run the simulations are reported in Table6.2. They have been estimated accoriding the following procedure.

From the small scale experiments, detection times at nine different supersaturations, $S_{\alpha,i}$, were collected (see also Figure 6.4). All data were measured in this case at a temperature $T = 298.15$ K (Section 6.2). The values of supersaturations of each experimental series $i = 1, ..., N_s$ are reported in Table 6.2.

Using Model 1 (see Section 2.3.2) for modelling the process of primary nucleation, we have computed the values of $\hat{J}_i$ from the experiments associated to each $S_{\alpha,i}$:

$$t_N = t_D - t_G \tag{6.15}$$

$$P(t_N) = F(t_D) = 1 - \exp\left(-\lambda(t_D - t_G)\right) \tag{6.16}$$

where the growth time t_G, as in Chapter 5 (Section 5.2.4), has been defined using the approximation of Jiang et al. [59, 93, 115]; for each $S_{\alpha,i}$, one assumed then an associated constant value of $t_{G,i}$ given by:

$$t_{G,i} = \min_{S_{\alpha,i}}(\mathbf{t_{D,i}}) \qquad \forall i = 1, ..., N_s \tag{6.17}$$

Note that for the sake of simplicity, we ignore the possibility that t_G at a given S_α depends on the polymorph that has formed first.

All estimates $\hat{J}_i$ are reported in the second column of Table 6.2 and illustrated as green symbols in Figure 6.12. The estimate $\hat{J}$ from an eCDF has always an inherent uncertainty [106, 115] and a formula for estimating the uncertainty of the nucleation rate from that of the eCDF, which we have used here, has been derived in a recent contribution [115]. Using such formula, which yields asymmetric confidence intervals, one can calculate the relative uncertainties (at a confidence level of 0.95) of each J_i according to the following expression:

$$(1 + \eta_i^-)J_i \leq J_i \leq (1 + \eta_i^+)J_i \qquad (i = 1, ..., 9) \tag{6.18}$$

Table 6.2: The estimates $\hat{J}_i$ from the small size experiments at the different supersaturations, with the associated relative uncertainties, η_i^+ and η_i^-; the relative uncertainties η^+ and η^- of each nucleation rate, and the number of points n used in each eCDF.

i	$S_{\alpha,i}$	$\hat{J}_i$ [# /(m^3 s)]	η_i^+	η_i^-	n
1	1.20	0.05	0.30	-0.23	197
2	1.25	0.18	0.19	-0.16	446
3	1.30	0.28	0.19	-0.16	447
4	1.41	0.35	0.17	-0.15	527
5	1.58	1.64	0.12	-0.11	1025
6	1.63	1.50	0.27	-0.22	233
7	1.72	1.89	0.16	-0.14	635
8	1.91	4.40	0.14	-0.12	767
9	2.07	10.40	0.32	-0.24	181

The values of the uncertainties are reported in the third- and second-last columns of Table 6.2; η_i^+ refers to a positive deviation from the estimate $\hat{J}_i$, while η_i^- to a negative deviation.

The last column of Table 6.2 reports the number n of experimental points of each eCDF. The data we have collected, and the associated estimates $\hat{J}_i$ (Table 6.2), are not sufficient to determine a unique set of model parameters. Therefore, we have considered the following additional constraints on the values of the occurrence

probabilities ξ_j, which are based on the data presented in Figure 6.7:

$$\begin{cases} \xi_\alpha > \xi_\gamma & 1.00 \leq S_\alpha \leq 1.15 \\ \xi_\gamma > \xi_\beta & 1.00 \leq S_\alpha \leq 1.50 \\ \xi_\beta > \xi_\gamma & S_\alpha \geq 1.65 \end{cases} \tag{6.19}$$

The values for the pairs (A_j, B_j) reported in Table 6.3 are those that fit best the nucleation rates in Table 6.2, while at the same time fulfilling the constraints of Eq. (6.19). They have been used in Eq. (6.14) to compute the ξ_j values, which are plotted in Figure 6.12 as a function of supersaturation and compared with the occurrence frequency from the experimental measurements.

Figure 6.11 illustrates the comparison between the nucleation rates from the model and and the estimates $\hat{J}_i$ from the experimental data, with the related uncertainties. The nucleation rates from the model have been computed at $T = 298.15$ K and plotted as function of S_α: the blue, red, and yellow curves represent J_α, J_β, and J_γ, respectively; the black line is the apparent nucleation rate $\hat{J}$, which can be compared with the the experimental estimates, reported as green symbols.

Table 6.3: The parameters of the nucleation rate J_j selected and used in the simulations.

Parameters / Form	α	β	γ
A_j [# /(kg s)]	23	17,715	392
B_j [K^3]	475	2.37×10^7	8.85×10^5

Let us now compare the simulations results and the experimental data. Eq. (6.13) indicates that one can modify the occurrence

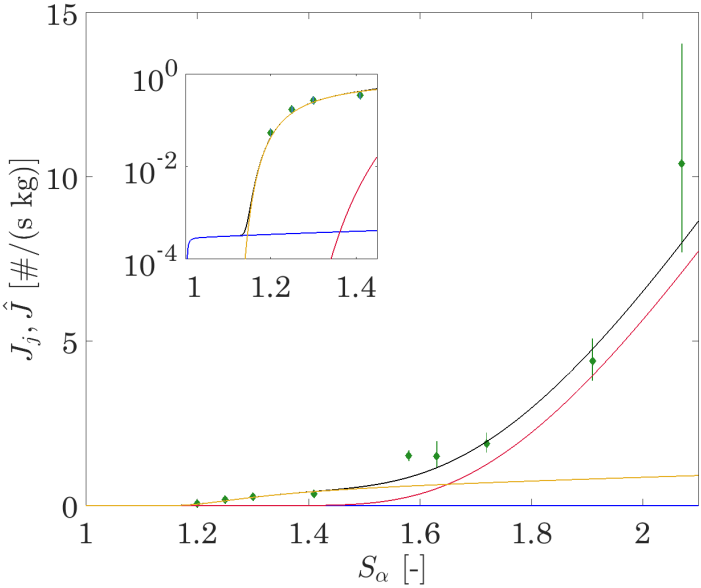


Figure 6.11: The apparent nucleation rates, $\hat{J}_i$ (green symbols), from small scale experiments at $T = 298.15$ K, for different values of $S_{\alpha,i}$. The uncertainty interval (green bars) associated to each $S_{\alpha,i}$ has been computed at a confidence level of 0.95. Using the parameters in Table 6.3, the primary nucleation rates of the individual polymorphs, J_j have been drawn as solid lines in blue, magenta, and yellow for α, β, and γ, respectively, and the associated sum, i.e. $\hat{J}$, as a black solid line.

probability of the polymorphs (Eq. (6.12)) by varying the supersaturation, hence the nucleation rate; using the parameters reported in Table 6.3, one can compute the values of ξ_α, ξ_β, and ξ_γ as a function of S_α. These values are represented in Figure 6.12 as the blue, red, and yellow solid line, for ξ_α, ξ_β, and ξ_γ, respectively.

Figure 6.7, on the other hand, reports the (percentage) occurrence frequency Q of form α, β, and γ, at different values of S_α:

one can interpret the occurrence frequency of each polymorph as an estimate of its occurrence probability.

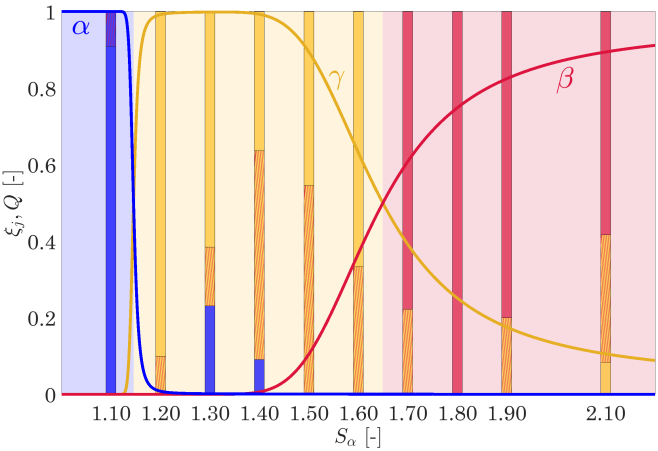


Figure 6.12: The occurrence probabilities ξ_α, ξ_β, and ξ_γ, in blue, red, and yellow, plotted as function of S_α, while the stacked bars represents the rescaled relative occurrences Q measured in the large scale experiments and reported also in Figure 6.7.

By normalising Q from Figure 6.7, one can then compare the values of Q with those of ξ_j, at each S_α, as shown in Figure 6.12. One can readily see that the model results and the experimental measurements exhibit a similar behaviour.

Thus concluding, these results suggest that the model presented in Section 6.4.1 is not only plausible, but also consistent with the data, further indicating that it indeed describes the main, fundamental physical features of the system.

6.5 SUMMARY AND CONCLUSIONS

First, we make a few remarks about the deterministic, kinetic interpretation of Ostwald's rule of stages developed by Davey, Garside, Cardew and others [145–149, 169], mentioned in Section 6.1. Even though the deterministic theory cannot explain the experiments of Section 6.3 of INA in EtOH, a deep connection exists between such theory and the stochastic model. Davey, Garside, and others interpreted the primary nucleation rate as a measure of how much a certain polymorph j is favoured against the others in a deterministic framework. This "bias" towards a specific polymorph would promote the formation of pre-critical clusters of a specific configuration and would eventually lead to the formation of the polymorph with the highest nucleation rate at the given conditions. In the stochastic framework, the "bias" of the deterministic theory can be interpreted as the probability of the system to form a certain polymorph (see Eq.(6.12)). In fact, the intervals of S_α where the stochastic model predicts that a polymorph j is more likely to appear than the others correspond to the intervals where the deterministic theory predicts that the same polymorph j nucleates always first. Thus, the deterministic model can be seen as a special case of the stochastic model, in the sense that its prediction corresponds to the most-likely outcome predicted by the stochastic approach. Second, we note that when J_j at the given conditions is such that $\xi_j \approx 1$, then the system exhibits a pseudo-deterministic behaviour, i.e. the experiments yield almost always the polymorph j only. Finally, it is worth noting that, through a qualitative analysis not reported here, we can show that the region where the α polymorph nucleates with higher probability should become larger when T decreases; this conclusion further supports the conjecture that the difference between our findings and those of Kulkarni et al. [138] can be due to the different temperatures at which these authors have conducted their experiments, namely 298.15 K in our case and 278.15 K in theirs.

Thus summarising, in this work we have presented nucleation experiments of isonicotinamide in ethanol where all polymorphs, and not only the stable one (α, form EHOWIH01), have been crystallised from clear solutions. The production of the metastable β and γ forms (form EHOWIH02 and EHOWIH04, respectively) has been consistently obtained in a broad range of initial concentrations, and has been shown to occur also at two different system sizes. The formation of the polymorphs β and γ is in contrast with one hypothesis made in the literature about crystallisation of INA in EtOH [138, 153], but is consistent with the behaviour of INA observed in other solvents [137]. Even more interestingly, our experiments exhibited a stochastic nature in terms not only of detection times, but also of the polymorph forming first. On the basis of the theoretical results concerning the stochastic nature of primary nucleation, we have developed (Section 6.4) a simple model which naturally extends previous work on stochastic primary nucleation [36, 46–48, 85] to a polymorphic system. Such model describes in a coherent manner the stochasticity of nucleation/detection times, as well as that of the type of polymorph nucleating, as observed in the experiments.

7

CONCLUSIONS AND OUTLOOK

The truth does not change according to our ability to stomach it.
— Flannery O'Connor

Nucleation is the first step in the transition from a metastable parent phase to a new thermodynamic phase. It is not only a scientifically challenging and important problem in its own right, but also a crucial process in several technologies used in many important industrial processes. Notwithstanding a long history of investigations, several features of nucleation have yet to be properly understood and described: above all, its stochastic nature and the practical impossibility of directly observing the formation of nuclei. Thus, one major challenge is to perform experiments which corroborate theory predictions.

First, the experimental characterisation of any phenomenon must comply with a minimal requirement, namely that the data are of a high quality. By this, one means that the system in which they are measured should be affected as little as possible by noise. Second, when one studies stochastic processes a further requirements must be considered, i.e that a sufficient number of data points must be gathered to perform a meaningful analysis and retrieve the statistical properties of the system. Third, in the case of crystallisation, one also needs to link the unobservable formation of nuclei and observable fully grown crystals.

This thesis had, thus, a twofold purpose.

On the one hand, it aimed at investigating primary nucleation, particularly the formation of crystals from a clear solution. The purpose was to develop and implement mathematical models, which account for the intrinsic stochastic nature of nucleation, in a system either with one species of crystals or with multiple polymorphs. An

important criterion of model design in this thesis is its applicability to a broad range of operating conditions.

On the other hand, this thesis addressed how to perform accurate experiments and how to analyse their results in a statistically robust manner. The ultimate goal was to retrieve reliable information about the physical and statistical properties of the system, thereby obtaining nucleation kinetics, which are a vital piece of information for process design, optimisation, and control.

Chapter 2 reviewed the fundamental theories of nucleation, highlighting their thermodynamic foundations, particularly the activated nature of nucleation and its kinetic consequences. Building on the essential results of the deterministic framework, the idea of nucleation as a stochastic process was then introduced and discussed. The key element to a stochastic model of nucleation is the exponential statistic, which describes activated phenomena and, ultimately, nucleation. Using the exponential statistic, two different stochastic models of primary nucleation are proposed. The first one (Nucleation Model 1), following a hypothesis often found in the literature, assumed that only the first primary nucleus originates stochastically, whereas subsequent nuclei form deterministically, e.g. according to the rate given by classical nucleation theory. The second model (Nucleation Mode 2), on the contrary, assumed that each primary nucleus forms stochastically, but still according to an exponential probability law. In this case, though, the probability depends on the current system state, which is determined by the formation and growth of previous nuclei, i.e. by the history of the system. Finally, we illustrated the differing model predictions by simulating the condensation of Argon from its vapour using enhanced molecular dynamics, implementing a novel method to investigate realistic nucleation conditions. On the one hand, Argon condensation elucidated the challenges of performing detailed simulations at a molecular level. On the other hand, this system offered a clear insight into the main features of nucleation and into the validity and limitations of classical nucleation theory. Moreover,

the results highlights the inherently stochastic nature of nucleation even in such a simple, small system. A second contribution of this Chapter is that these simulations demonstrated the benefits (and the necessity) of developing efficient computational techniques to explicate the fundamental mechanisms of nucleation at a level of detail still beyond direct experimental investigation.

Chapter 3 focused on building the connection between nucleation and the observation of fully grown crystals. To this aim, we first distinguished between nucleation time and detection time, and we proposed a mathematical formalisation of this difference. Then, we introduced two alternative approaches to describe crystallisation from the formation of the first nucleus until crystals detection, using the nucleation models developed in Chapter 2. The first approach, based on a continuous population model, was implemented into the framework of population balance equations. The stochastic nature of nucleation manifested in this model as a randomisation of the initial conditions given by the nucleation time. The second approach, based on a discrete population model, used the second nucleation model, according to which each crystal forms stochastically but grows deterministically. The growth of each crystal was described by an ordinary differential equation. Finally, the implementation of the detection condition and the constitutive equations of nucleation and growth were also discussed.

In **Chapter 4**, we investigated the statistical premises and properties of empirical distribution functions of random variables, i.e. the estimate of cumulative distribution functions reconstructed from experimental data. Relying on important results of statistics, we examined the problem of convergence of such an empirical estimate to its underlying distribution function as well as the rate of convergence, highlighting the inherent, unavoidable error present in the former. Second, we discussed how to estimate the model parameters from the data and their empirical distributions through two alternative approaches. The first one was based on the Euclidean metric and exploited a non-linear least squares method to

obtain the values of the model parameters by error minimisation. The second one developed on the Chebyschev metric and performed an estimation of the parameters based on Fisher's meta-analysis by p-value maximisation, relying on statistical tools and properties of stochastic processes. Then, the two models were validated against the experimental observations of several systems reported in the literature: we used the models to describe the experiments, estimate the relevant model parameters, and assess the parameter correlations. By inspecting the fitting results, we not only evaluated the predictive ability of the models, but we also investigated the similarities and the differences between the two models. To better understand why the models described some sets of data better than others, we performed a sensitivity analysis on the operating variables, namely the system size, the temperature, and the concentration (or equivalently, the supersaturation). We focused particularly on the sensitivity of the nucleation probability and of the kinetic laws. The analysis showed that several investigations were performed in conditions suffering from a high sensitivity to operating conditions, but it did not provide a direct explanation why the same model would be better suited for some sets of data than for others. Nevertheless, the sensitivity analysis pointed out that performing experiments under sensitive conditions introduces further uncertainties on the data, biases the interpretation of the experiments, and hinders an appropriate modelling, thus deteriorating the quality of the estimated parameters. Finally, we investigated in more detail the detection technique used to produce most of the data, the definition of the detection condition itself, and its mathematical implementation - an often neglected aspect of crystallisation models. Indeed, theoretical arguments and characterisation experiments confirmed that transmissivity, used in an increasing number of experimental investigations, would be more appropriately described using the second order moment of the crystal size distribution.

In light of these results of Chapter 4, **Chapter 5** focused on the analysis of experimental data and on the estimation of nucleation kinetics from multiple series of detection time measurements which are repeated at the same conditions. By investigating the behaviour of isothermal crystallisation of isonicotinamide from ethanol, we first studied the reproducibility of detection time measurements. We proposed a statistical method to identify the confidence level if several distinct empirical distribution functions belong or not to the same statistic, building upon and extending the results of Kolmogorov and Smirnov. In agreement with the results of Chapter 2, we also presented a method to estimate the error propagation from an empirical distribution to its associated nucleation rate estimate, assuming a Poisson model of nucleation and exponentially distributed nucleation times. This method is not only parameter-free, but it also yields asymmetric confidence intervals which automatically account for the inherent inaccuracies of any empirical distribution and the confidence level one attributes to the data.

Finally, in **Chapter 6**, we investigated the concomitant nucleation of multiple polymorphs. We showed experimental results obtained from hundreds of crystallisation experiments of isonicotinamide from ethanol in two different volumes, where the polymorphism of the crystals was monitored with both XRD and Raman spectroscopy. The results showed that, in such systems, the nucleation time (and hence the detection time) and the type of polymorph forming are stochastic. To rationalise the results, we described the competing stochastic nucleation of polymorphs by expanding on the results of Model 1 in Chapter 2 and showed that a polymorphic system has a much richer and more complex behaviour that a non-polymorphic one. The model equations also indicated that, if one cannot discriminate among the different polymorphs, the nucleation rate estimated from detection time distributions corresponds to an apparent, total nucleation rate. In addition, this rate's dependency on supersaturation and temperature may not be properly described by classical nucleation theory, even though

the rates of the individual polymorphs are individually described by it in the correct manner.

In summary, the following main conclusions can be drawn from the results presented in this work:

1. Molecular dynamics, successfully applied to estimate the rate of Argon condensation from its vapour at realistic conditions, can be used for more challenging systems to assess (or discard) some hypotheses concerning the mechanisms of formation of new nuclei and of their growth.

2. Since detection times are stochastic, their experimental measurements must be analysed using appropriate statistical tools. One must also be aware of the error inherent to any empirical distribution function and its rate of convergence to the actual cumulative distribution, since, ultimately, nucleation kinetics are estimated from empirical distributions.

3. Primary nucleation is the principal cause of detection times stochasticity. However, since nucleation time distributions are not directly observable, this can be inferred from the data by using a mathematical model describing not only nucleation, but also the evolution and the detection of the crystals, which may not be easily formulated.

4. The estimation of the model parameters from experimental data can be structured such that the statistical properties of the model are accounted for during the estimation procedure itself. Given the limited accuracy of detection time experiments, any experimentalist should always consider the problem of sensitivity, that is, selecting the operating conditions as to maximise the distance between two distributions or minimise the consequences of sensitivity.

5. The existence of several polymorphs, a feature common to many substances, leads to a richer stochastic behaviour du-

ring crystallisation, which entails a more detailed mathematical description of the process.

7.1 OUTLOOK OF FUTURE WORK

In order to improve the optimisation and the control over the processes in which nucleation is a determining phenomenon, models must still be improved and refined. Notwithstanding many achievements, much still remains to do. Among the several areas of research which can be explored, only a few are briefly discussed here, namely those which are a natural continuation of the work presented in the previous Chapters.

First, one must determine the domain of applicability of the classical nucleation theory and of the two-step nucleation theory. While the classical nucleation theory has been applied rather successfully to describe the nucleation of many compounds, it has failed in many other circumstances. At the same time, the existence of pre-nucleation clusters with a non-crystalline nature has been confirmed experimentally in several, but not all systems: the two-step nucleation theory is grounded and built upon the existence of such non-crystalline structures. The natural direction of research in this area is to construct a more general theory of nucleation, capable of accounting for the results of the classical nucleation theory as well as of the two-step nucleation theory. Such a theoretical development would also be instrumental in understanding which operating parameters control and/or influence nucleation, thus enabling one to shed light on experimental results which cannot be explained in the current framework. An important example in this context is provided by the influence of fluid-dynamics, particularly of the shear stresses. Experiments have shown that shear can favour the formation of one polymorph over another as well as influence the velocity at which nucleation starts: these observations, specially when small molecules are involved rather than proteins or polymers, cannot be explained by the classical nucleation theory, but

neither by two-step nucleation theory, at least in the present formulation.

Second, a better comprehension of the physical nature and of the characteristics of the pre-nucleation clusters would allow one to describe more appropriately the stochastic process underlying nucleation. Each energy barrier is physically associated to a stochastic event: a system overcoming two barriers should exhibit statistical features different from one evolving according to a classical pathway, particularly if the intermediated species is relatively stable. Additionally, recent works have proposed that the description of the stochastic process of nucleation should not be described using time as appropriate coordinate, but rather other variables. Although still at an embryonic stage, this idea deserves further consideration and investigation, particularly given the thermodynamic nature of nucleation. Furthermore, understanding better the abstract space in which the stochastic process is appropriately described, could improve the definition and the selection of the observables to be monitored during the experiments.

Third, one should inspect theoretical predictions and experimental observations with the aid of molecular dynamic simulations, since they allow one to directly look into the system at a scale and with a resolution otherwise not reachable. However, it is known that the computational burden these simulations entail can still be too demanding for studying systems at real conditions. Moreover, the potentials used to describe the systems in the simulations are at times not accurate enough yet. A concerted effort should thus be directed at developing better numerical algorithms and more accurate potentials need also developing.

Finally, to complement the previously listed areas, one should focus on the design of the experiments: on the one hand, on how to actually perform appropriate measurements and, on the other hand, on how to account for model and measure uncertainties. Concerning the measurements, research has been recently moving towards new applications and novel combinations of techniques to

gain a deeper comprehension of the structure of the solution: since optical microscopy cannot be used to look at the clusters, scattering techniques have been used, such as neutron, X-ray, and light scattering. Special attention has been paid to the system behaviour at conditions close to saturation, where the pre-nucleation clusters are more stable, hence they have longer life-times and are easier to monitor. Few important and enlightening experiments have been performed in this direction in recent years. Further improvements along this line of research should combine and complement the information provided by scattering techniques with that given by spectroscopy and optical microscopy for a comprehensive characterisation of the system. Since the simple existence of pre-nucleation clusters may not provide sufficient information to characterise the nucleation process, one should consider to carefully analyse the data to extract other properties of the clusters (such as the cluster size distribution, or the cluster geometry, compact or fractal) and of the system (such as hysteresis). These pieces of information are often present in the measurements, but they do require lengthier, and often rather complex, analyses of high-quality data, i.e. with few or almost no noise. To gather such data, new protocols must be developed, to ensure the reproducibility, the repeatability, and the reliability of crystallisation experiments. From an experimental perspective, though, some factors cannot be controlled, even though they greatly affect the system, such as the presence and type of impurities. Since impurities severely hinder experimental reproducibility, their impact on the system should be further investigated and methods to analyse the data while accounting for the unavoidable, inherent experimental uncertainties should be further developed. To achieve this goal, crystallisation might draw inspiration from and further build on the results obtained in related fields, such as chemical industry. In several chemical and petro-chemical plants, models (based on polynomial chaos-expansion, machine learning, and Bayesian statistics) have been successfully implemented in the past two decades. These models account not only for the uncertain-

ties in the composition of streams directed to industrial synthesis, but allow one to implement control and optimisation policies, even though the physics of the synthesis is not fully understood.

By complementing the work carried out in this thesis, one could expect not only to gain a more comprehensive knowledge of nucleation, but also to improve and intensify processes of industrial relevance.

BIBLIOGRAPHY

[1] H. B. Callen, *Thermodynamics and an Introduction to Thermostatistics*, 2nd ed., Wiley, **1985** (cit. on p. 1).

[2] J. W. Gibbs, *The collected works of J. Willard Gibbs*, Longmans, Green and Co., **1928** (cit. on pp. 2, 13, 184).

[3] R. Becker, W. Döring, "Kinetische Behandlung der Keimbildung in übersättigten Dämpfen", *Ann. Phys.* **1935**, *416*, 719–752 (cit. on pp. 2, 13, 16, 62, 184).

[4] D. Kashchiev, *Nucleation, Basic theory with applications*, Butterworth Heinemann, **2000** (cit. on pp. 2, 10, 13, 16, 58, 184).

[5] S Toschev, A Milchev, S Stoyanov, "On some probabilistic aspects of the nucleation process", *J. Cryst. Growth* **1972**, *13-14*, 123–127 (cit. on pp. 4, 20, 179).

[6] W. J. Dunning, A. J. Shipman in Proc. Agric. Ind. 10th Int. Conf. **1954**, 1448 (cit. on p. 4).

[7] D. Frenkel, K. J. Schrenk, S. Martiniani, "Monte Carlo sampling for stochastic weight functions", **2016**, *114*, DOI 10.1073/pnas.1620497114 (cit. on p. 4).

[8] M. Mazzotti, T. Vetter, D. R. Ochsenbein, G. M. Maggioni, C. Lindenberg in *Polymorph. Pharm. Ind. Solid Form Drug Dev.* Springer-Verlag, **2018** (cit. on pp. 9, 231).

[9] F. F. Abraham, *Homogeneous nucleation theory*, Advances i, Academic Press, New York, **1974** (cit. on p. 10).

[10] J. M. Prausnitz, R. Lichtenthaler, E. G. Azevedo, *Molecular Thermodynamics in Fluid Phase Equilibria*, 3rd ed., Prentice-Hall, **1999** (cit. on p. 10).

[11] O. Galkin, P. G. Vekilov, "Direct Determination of the Nucleation Rates of Protein Crystals", *J. Phys. Chem. B* **1999**, *103*, 10965–10971 (cit. on pp. 12, 48).

[12] D. Kashchiev, P. G. Vekilov, A. B. Kolomeisky, "Kinetics of two-step nucleation of crystals", *J. Chem. Phys.* **2005**, *122*, DOI 10.1063/1.1943389 (cit. on p. 12).

[13] D. Erdemir, A. Y. Lee, A. S. Myerson, "Nucleation of crystals from solution: Classical and two-step models", *Acc. Chem. Res.* **2009**, *42*, 621–629 (cit. on pp. 12, 59).

[14] P. G. Vekilov, "Nucleation", *Cryst. Growth Des.* **2010**, *10*, 5007–5019 (cit. on pp. 12, 19).

[15] D. Vivarès, E. W. Kaler, A. M. Lenhoff, "Quantitative imaging by confocal scanning fluorescence microscopy of protein crystallization via liquid-liquid phase separation", *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2005**, *61*, 819–825 (cit. on p. 12).

[16] A. M. Beale, A. M. Van Der Eerden, S. D. Jacques, O. Leynaud, M. G. O'Brien, F. Meneau, S. Nikitenko, W. Bras, B. M. Weckhuysen, "A combined SAXS/WAXS/XAFS setup capable of observing concurrent changes across the nano- to-micrometer size range in inorganic solid crystallization processes", *J. Am. Chem. Soc.* **2006**, *128*, 12386–12387 (cit. on p. 12).

[17] N. Pienack, W. Bensch, "In-situ monitoring of the formation of crystalline solids", *Angew. Chemie - Int. Ed.* **2011**, *50*, 2014–2034 (cit. on p. 12).

[18] M. Volmer, I. Estermann, "Über den Mechanismus der Molekülabscheidung an Kristallen", *Zeitschrift fuer Phys.* **1921**, *7*, 13–17 (cit. on p. 13).

[19] J. Frenkel, "A General Theory of Heterophase Fluctuations and Pretransition Phenomena", *J. Chem. Phys.* **1939**, *7*, 538 (cit. on p. 13).

[20] J. W. Cahn, J. E. Hilliard, "Free Energy of a Nonuniform System. I. Interfacial Free Energy", *J. Chem. Phys.* **1958**, *28*, 258 (cit. on p. 13).

[21] J. W. Cahn, "Free Energy of a Nonuniform System. II. Thermodynamic Basis", *J. Chem. Phys.* **1959**, *30*, 1121 (cit. on p. 13).

[22] P. G. Debenedetti, *Metastable Liquids Concepts and Principles*, Princeton University Press, **1996** (cit. on p. 13).

[23] H. H. Kramers, "Brownian motion in a field of force and the diffusion model of chemical reactions", *Physica* **1940**, *7*, 284–304 (cit. on p. 16).

[24] H. M. Rosenstock, M. B. Wallenstein, A. L. Wahrhaftig, H Eyring, "Absolute Rate Theory for Isolated Systems and the Mass Spectra of Polyatomic Molecules.", *Proc. Natl. Acad. Sci. U. S. A.* **1952**, *38*, 667–678 (cit. on p. 16).

[25] J. Langer, "Statistical theory of the decay of metastable states", *Ann. Phys. (N. Y).* **1969**, *54*, 258–275 (cit. on pp. 16, 22).

[26] D. G. Truhlar, B. C. Garrett, S. J. Klippenstein, "Current Status of Transition-State Theory", *J. Phys. Chem.* **1996**, *100*, 12771–12800 (cit. on p. 16).

[27] D. Kashchiev, G. M. van Rosmalen, "Review: Nucleation in solutions revisited", *Cryst. Res. Technol.* **2003**, *38*, 555–574 (cit. on p. 16).

[28] P. G. Vekilov, "Two-step mechanism for the nucleation of crystals from solution", *J. Cryst. Growth* **2005**, *275*, 65–76 (cit. on pp. 18, 19).

[29] W. Pan, A. B. Kolomeisky, P. G. Vekilov, "Nucleation of ordered solid phases of proteins via a disordered high-density state: phenomenological approach.", *J. Chem. Phys.* **2005**, *122*, 174905 (cit. on p. 18).

[30] P. G. Vekilov, "The two-step mechanism of nucleation of crystals in solution", *Nanoscale* **2010**, *2*, 2346 (cit. on p. 19).

[31] P. G. Vekilov in *Adv. Chem. Phys. Vol. 151*, (Eds.: G. Nicolis, D. Maes), **2012**, Chapter The two-st, 79 (cit. on p. 19).

[32] M. A. Vorontsova, D. Maes, P. G. Vekilov, "Recent advances in the understanding of two-step nucleation of protein crystals", *Faraday Discuss.* **2015**, *179*, 27–40 (cit. on p. 19).

[33] B. A. Garetz, J. Matic, A. S. Myerson, "Polarization switching of crystal structure in the nonphotochemical light-induced nucleation of supersaturated aqueous glycine solutions.", *Phys. Rev. Lett.* **2002**, *89*, 175501 (cit. on p. 19).

[34] D. Gebauer, A. Völkel, H. Cölfen, "Stable prenucleation calcium carbonate clusters.", *Science* **2008**, *322*, 1819–22 (cit. on p. 19).

[35] M. Salvalaglio, M. Mazzotti, M. Parrinello, "Urea homogeneous nucleation mechanism is solvent dependent", *Faraday Discuss.* **2015**, *179*, 291–307 (cit. on pp. 19, 35).

[36] G. M. Maggioni, M. Mazzotti, "Modelling the stochastic behaviour of primary nucleation", *Faraday Discuss.* **2015**, *179*, 359–382 (cit. on pp. 20, 21, 43, 59, 60, 89, 155, 170, 179, 180, 191, 231).

[37] A. F. Izmailov, A. S. Myerson, S. Arnold, "A statistical understanding of nucleation", *J. Cryst. Growth* **1999**, *196*, 234–242 (cit. on pp. 20, 48, 59, 61, 171, 179).

[38] O. Pino-García, Å. C. Rasmuson, "Primary Nucleation of Vanillin Explored by a Novel Multicell Device", *Ind. Eng. Chem. Res.* **2003**, *42*, 4899–4909 (cit. on pp. 20, 48, 49, 121, 122).

[39] O. Pino-García, A. Rasmuson, "Influence of Additives on Nucleation of Vanillin: Experiments and Introductory Molecular Simulations", *Cryst. Growth Des.* **2004**, *4*, 1025–1037 (cit. on pp. 20, 48).

[40] J. Schöll, L. Vicum, M. Müller, M. Mazzotti, "Precipitation of L-Glutamic Acid: Determination of Nucleation Kinetics", *Chem. Eng. Technol.* **2006**, *29*, 257–264 (cit. on pp. 20, 49, 58).

[41] C. Lindenberg, M. Mazzotti, "Effect of temperature on the nucleation kinetics of α L-glutamic acid", *J. Cryst. Growth* **2009**, *311*, 1178–1184 (cit. on pp. 20, 25, 49, 54, 58, 118).

[42] N. a. Mitchell, P. J. Frawley, "Nucleation kinetics of paracetamol–ethanol solutions from metastable zone widths", *J. Cryst. Growth* **2010**, *312*, 2740–2746 (cit. on p. 20).

[43] R. Ohmura, M. Ogawa, K. Yasuoka, Y. H. Mori, "Statistical Study of Clathrate-Hydrate Nucleation in a Water/Hydrochlorofluorocarbon System: Search for the Nature of the "Memory Effect"", *J. Phys. Chem. B* **2003**, *107*, 5289–5293 (cit. on p. 20).

[44] D. G. Randall, J. Nathoo, F. Genceli-Güner, H. J. M. Kramer, G. J. Witkamp, A. E. Lewis, "Determination of the metastable ice zone for a sodium sulphate system", *Chem. Eng. Sci.* **2012**, *77*, 184–188 (cit. on pp. 20, 48).

[45] N. Gherras, G. Fevotte, "Comparison between approaches for the experimental determination of metastable zone width: A case study of the batch cooling crystallization of ammonium oxalate in water", *J. Cryst. Growth* **2012**, *342*, 88–98 (cit. on p. 20).

[46] L. Goh, K. Chen, V. Bhamidi, G. He, N. C. S. Kee, P. J. A. Kenis, C. F. Zukoski, R. Braatz, "A Stochastic Model for Nucleation Kinetics Determination in Droplet-Based Microfluidic Systems.", *Cryst. Growth Des.* **2010**, *10*, 2515–2521 (cit. on pp. 21–23, 25, 26, 48, 49, 59, 61, 121, 122, 171, 179–181, 191).

[47] S. S. Kadam, S. A. Kulkarni, R. Coloma Ribera, A. I. Stankiewicz, J. H. ter Horst, H. J. Kramer, "A new view on the metastable zone width during cooling crystallization", *Chem. Eng. Sci.* **2012**, *72*, 10–19 (cit. on pp. 21, 24, 25, 48–50, 54, 55, 60, 61, 71, 72, 74, 75, 77–80, 82, 84, 85, 88, 116, 118, 122, 170, 171, 179, 180, 191).

[48] G. M. Maggioni, M. Mazzotti, "Stochasticity in Primary Nucleation: Measuring and Modeling Detection Times", *Cryst. Growth Des.* **2017**, *17*, 3625–3635 (cit. on pp. 21, 43, 59, 121, 155, 156, 179, 180, 184, 191, 231).

[49] R. P. Sear, "Quantitative studies of crystal nucleation at constant supersaturation: experimental data and models", *CrystEngComm* **2014**, *16*, 6506–6522 (cit. on pp. 21, 59, 121).

[50] L. J. Little, R. P. Sear, J. L. Keddie, "Does the γ Polymorph of Glycine Nucleate Faster? A Quantitative Study of Nucleation from Aqueous Solution", *Cryst. Growth Des.* **2015**, *15*, 5345–5354 (cit. on pp. 21, 47, 121, 122, 127, 135, 163).

[51] H. Yang, Å. C. Rasmuson, "Nucleation of Butyl Paraben in Different Solvents", *Cryst. Growth Des.* **2013**, *13*, 4226–4238 (cit. on pp. 21, 48, 121).

[52] D. Mealey, D. M. Croker, Å. C. Rasmuson, "Crystal nucleation of salicylic acid in organic solvents", *CrystEngComm* **2015**, *17*, 3961–3973 (cit. on pp. 21, 47, 49).

[53] C. Forsyth, I. S. Burns, P. A. Mulheran, J. Sefcik, "Scaling of Glycine Nucleation Kinetics with Shear Rate and Glass-Liquid Interfacial Area", *Cryst. Growth Des.* **2016**, *16*, 136–144 (cit. on pp. 21, 112, 122).

[54] A. J. Durant, R. A. Shaw, "Evaporation freezing by contact nucleation inside-out", *Geophys. Res. Lett.* **2005**, *32*, 1–4 (cit. on p. 22).

[55] M. Kuhs, J. Zeglinski, Å. C. Rasmuson, "Influence of History of Solution in Crystal Nucleation of Fenoxycarb: Kinetics and Mechanisms", *Cryst. Growth Des.* **2014**, *14*, 905–915 (cit. on pp. 22, 48).

[56] D. L. Snyder, M. I. Miller, *Random Point Processes in Time and Space*, Springer-Verlag, New York, NY, **1991** (cit. on pp. 22, 27, 179).

[57] M. Salvalaglio, P. Tiwary, M. Parrinello, "Assessing the Reliability of the Dynamics Reconstructed from Metadynamics", *J. Chem. Theory Comput.* **2014**, *10*, 1420–1425 (cit. on pp. 22, 32).

[58] R. Dombrowski, J. D. Litster, N. J. Wagner, Y. He, "Crystallization of alpha-lactose monohydrate in a drop-based microfluidic crystallizer", *Chem. Eng. Sci.* **2007**, *62*, 4802–4810 (cit. on pp. 23, 59).

[59] S. Jiang, J. H. ter Horst, "Crystal Nucleation Rates from Probability Distributions of Induction Times", *Cryst. Growth Des.* **2011**, *11*, 256–261 (cit. on pp. 23–25, 48, 55, 59–61, 127, 129, 155, 171, 179, 180, 185).

[60] D. Kashchiev, D. Verdoes, G. van Rosmalen, "Induction time and metastability limit in new phase formation", *J. Cryst. Growth* **1991**, *110*, 373–380 (cit. on pp. 25, 83).

[61] S. A. Kulkarni, S. S. Kadam, H. Meekes, A. I. Stankiewicz, J. H. ter Horst, "Crystal Nucleation Kinetics from Induction Times and Metastable Zone Widths", *Cryst. Growth Des.* **2013**, *13*, 2435–2440 (cit. on pp. 25, 48, 122, 162, 179).

[62] F. L. Nordström, M. Svärd, Å. C. Rasmuson, "Primary nucleation of salicylamide: the influence of process conditions and solvent on the metastable zone width", *CrystEngComm* **2013**, *15*, 7285 (cit. on pp. 25, 48).

[63] D. Cox, "Some statistical methods connected with series of events", *J. R. Stat. Soc. Ser. B* **1955**, *17*, 129–164 (cit. on p. 27).

[64] S. Teychené, B. Biscans, "Crystal nucleation in a droplet based microfluidic crystallizer", *Chem. Eng. Sci.* **2012**, *77*, 242–248 (cit. on pp. 28, 48, 49, 55).

[65] M. Salvalaglio, P. Tiwary, G. M. Maggioni, M. Mazzotti, M. Parrinello, "Overcoming time scale and finite size limitations to compute nucleation rates from small scale well tempered metadynamics simulations", *J. Chem. Phys.* **2016**, *145*, 211925 (cit. on pp. 28, 231).

[66] D. Reguera, R. K. Bowles, Y. Djikaev, H. Reiss, "Phase transitions in systems small enough to be clusters", *J. Chem. Phys.* **2003**, *118*, 340 (cit. on p. 29).

[67] M. Salvalaglio, C. Perego, F. Giberti, M. Mazzotti, M. Parrinello, "Molecular-dynamics simulations of urea nucleation from aqueous solution", *Proc. Natl. Acad. Sci.* **2015**, *112*, E6–E14 (cit. on pp. 29, 34, 35).

[68] P. Tiwary, M. Parrinello, "From Metadynamics to Dynamics", *Phys. Rev. Lett.* **2013**, *111*, 230602 (cit. on pp. 29, 31–33).

[69] A. Barducci, M. Bonomi, M. Parrinello, "Metadynamics", *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 826–843 (cit. on pp. 31, 32).

[70] O. Valsson, P. Tiwary, M. Parrinello, "Enhancing Important Fluctuations: Rare Events and Metadynamics from a Conceptual Viewpoint", *Annu. Rev. Phys. Chem.* **2016**, *67*, annurev–physchem–040215–112229 (cit. on pp. 31, 32).

[71] H. Grubmüller, "Predicting slow structural transitions in macromolecular systems: Conformational flooding", *Phys. Rev. E* **1995**, *52*, 2893–2906 (cit. on p. 33).

[72] G. Chkonia, J. Wölk, R. Strey, J. Wedekind, D. Reguera, "Evaluating nucleation rates in direct simulations.", *J. Chem. Phys.* **2009**, *130*, 064505 (cit. on pp. 33, 40).

[73] P. Tiwary, M. Parrinello, "A Time-Independent Free Energy Estimator for Metadynamics", *J. Phys. Chem. B* **2015**, *119*, 736–742 (cit. on p. 33).

[74] J. Wedekind, D. Reguera, R. Strey, "Finite-size effects in simulations of nucleation", *J. Chem. Phys.* **2006**, *125*, DOI 10.1063/1.2402167 (cit. on pp. 34, 35, 38).

[75] R. Grossier, S. Veesler, "Reaching One Single and Stable Critical Cluster through Finite-Sized Systems", *Cryst. Growth Des.* **2009**, *9*, 1917–1922 (cit. on p. 34).

[76] J. Diemand, R. Angélil, K. K. Tanaka, H. Tanaka, "Large scale molecular dynamics simulations of homogeneous nucleation.", *J. Chem. Phys.* **2013**, *139*, 074309 (cit. on p. 39).

[77] J. N. Cederberg, *A Course in Modern Geometries*, Springer New York, New York, NY, **2001** (cit. on p. 47).

[78] S. Teychené, B. Biscans, "Microfluidic device for the crystallization of organic molecules in organic solvents", *Cryst. Growth Des.* **2011**, 4810–4818 (cit. on pp. 48, 49, 55).

[79] K. Chen, L. Goh, G. He, P. J. A. Kenis, C. F. Zukoski, R. D. Braatz, "Identification of nucleation rates in droplet-based microfluidic systems", *Chem. Eng. Sci.* **2012**, *77*, 235–241 (cit. on p. 48).

[80] D. Knezic, J. Zaccaro, A. S. Myerson, "Nucleation Induction Time in Levitated Droplets", *J. Phys. Chem. B* **2004**, *108*, 10672–10677 (cit. on pp. 48, 59).

[81] P. Laval, J.-B. Salmon, M. Joanicot, "A microfluidic device for investigating crystal nucleation kinetics", *J. Cryst. Growth* **2007**, *303*, 622–628 (cit. on p. 48).

[82] M. Ildefonso, N. Candoni, S. Veesler, "A Cheap, Easy Microfluidic Crystallization Device Ensuring Universal Solvent Compatibility", *Org. Process Res. Dev.* **2012**, *16*, 556–560 (cit. on pp. 48, 49).

[83] M. Ildefonso, N. Candoni, S. Veesler, "Heterogeneous Nucleation in Droplet-Based Nucleation Measurements", *Cryst. Growth Des.* **2013**, *13*, 2107–2110 (cit. on pp. 48, 49).

[84] Z. Hammadi, N. Candoni, R. Grossier, M. Ildefonso, R. Morin, S. Veesler, "Small-volume nucleation", *Comptes Rendus Phys.* **2013**, *14*, 192–198 (cit. on pp. 48, 49).

[85] R. A. Sullivan, R. J. Davey, G. Sadiq, G. Dent, K. R. Back, J. H. ter Horst, D. Toroz, R. B. Hammond, "Revealing the Roles of Desolvation and Molecular Self-Assembly in Crystal Nucleation from Solution: Benzoic and p -Aminobenzoic Acids", *Cryst. Growth Des.* **2014**, *14*, 2689–2696 (cit. on pp. 48, 49, 55, 59, 61, 71, 88, 89, 92, 102, 117, 155, 171, 179, 191).

[86] M. Svärd, F. L. Nordström, E.-M. Hoffmann, B. Aziz, Å. C. Rasmuson, "Thermodynamics and nucleation of the enantiotropic compound p-aminobenzoic acid", *CrystEngComm* **2013**, *15*, 5020 (cit. on pp. 48, 89).

[87] R. Grossier, Z. Hammadi, R. Morin, S. Veesler, "Predictive Nucleation of Crystals in Small Volumes and Its Consequences", *Phys. Rev. Lett.* **2011**, *107*, 025504 (cit. on p. 49).

[88] M. Ildefonso, N. Candoni, S. Veesler, "Using Microfluidics for Fast, Accurate Measurement of Lysozyme Nucleation Kinetics", *Cryst. Growth Des.* **2011**, *11*, 1527–1530 (cit. on pp. 49, 55).

[89] M. Ildefonso, E. Revalor, P. Punniam, J. Salmon, N. Candoni, S. Veesler, "Nucleation and polymorphism explored via an easy-to-use microfluidic tool", *J. Cryst. Growth* **2012**, *342*, 9–12 (cit. on p. 49).

[90] S. Teychené, B. Biscans, "Nucleation kinetics of polymorphs: Induction period and interfacial energy measurements", *Cryst. Growth Des.* **2008**, *8*, 1133–1139 (cit. on p. 49).

[91] S. Talreja, P. J. A. Kenis, C. F. Zukoski, "A kinetic model to simulate protein crystal growth in an evaporation-based crystallization platform.", *Langmuir* **2007**, *23*, 4516–22 (cit. on p. 49).

[92] V. Bhamidi, P. J. A. Kenis, C. F. Zukoski, "Probability of Nucleation in a Metastable Zone: Induction Supersaturation and Implications", *Cryst. Growth Des.* **2017**, *17*, 1132–1145 (cit. on p. 49).

[93] C. Brandel, J. H. ter Horst, "Measuring induction times and crystal nucleation rates", *Faraday Discuss.* **2015**, *179*, 199–214 (cit. on pp. 49, 60, 121, 171, 179, 185).

[94] M. Svärd, F. L. Nordström, T. Jasnobulka, Å. C. Rasmuson, "Thermodynamics and nucleation kinetics of m-aminobenzoic acid polymorphs", *Cryst. Growth Des.* **2010**, *10*, 195–204 (cit. on p. 49).

[95] J. Liu, M. Svärd, Å. C. Rasmuson, "Influence of Agitation and Fluid Shear on Nucleation of m -Hydroxybenzoic Acid Polymorphs", *Cryst. Growth Des.* **2014**, *14*, 5521–5531 (cit. on pp. 49, 112, 121, 163).

[96] D. Mealey, J. Zeglinski, D. Khamar, Å. C. Rasmuson, "Influence of solvent on crystal nucleation of risperidone", *Faraday Discuss.* **2015**, *179*, 309–328 (cit. on pp. 49, 179).

[97] H. Hulburt, S. Katz, "Some problems in particle technology", *Chem. Eng. Sci.* **1964**, *19*, 555–574 (cit. on p. 51).

[98] A. D. Randolph, "A population balance for countable entities", *Can. J. Chem. Eng.* **1964**, *42*, 280–281 (cit. on p. 51).

[99] A. Randolph, M. Larson, *Theory of Particulate Processes*, Academic Press, London, **1971** (cit. on p. 52).

[100] D. Ramkrishna, *Population Balances: Theory and Applications to Particulate Systems in Engineering*, Academic Press, **2000**, 355 (cit. on p. 52).

[101] B. Øksendal, A. Sulem, *Applied Stochastic Control of Jump Diffusions*, 2nd ed., Springer Berlin Heidelberg, Berlin, Heidelberg, **2007** (cit. on p. 56).

[102] *Crystallization Technology Handbook*, 2nd ed., (Ed.: A. Mersmann), Marcel Dekker Inc., New York, **2001** (cit. on p. 58).

[103] T. Vetter, M. Mazzotti, J. Brozio, "Slowing the Growth Rate of Ibuprofen Crystals Using the Polymeric Additive Pluronic F127", *Cryst. Growth Des.* **2011**, *11*, 3813–3821 (cit. on p. 58).

[104] D. R. Ochsenbein, S. Schorsch, F. Salvatori, T. Vetter, M. Morari, M. Mazzotti, "Modeling the facet growth rate dispersion of β l-glutamic acid—Combining single crystal experiments with nD particle size distribution data", *Chem. Eng. Sci.* **2015**, *133*, 30–43 (cit. on pp. 58, 179).

[105] D. Toroz, R. B. Hammond, K. J. Roberts, S. Harris, T. Ridley, "Molecular dynamics simulations of organic crystal dissolution: The lifetime and stability of the polymorphic forms of para-amino benzoic acid in aqueous environment", *J. Cryst. Growth* **2014**, *401*, 38–43 (cit. on p. 58).

[106] Y. Xiao, S. K. Tang, H. Hao, R. J. Davey, T. Vetter, "Quantifying the Inherent Uncertainty Associated with Nucleation Rates Estimated from Induction Time Data Measured in Small Volumes", *Cryst. Growth Des.* **2017**, *17*, 2852–2863 (cit. on pp. 58–60, 89, 121–123, 129, 143, 146, 147, 154, 155, 179, 180, 185).

[107] R. P. Sear, "Nucleation via an unstable intermediate phase", *J. Chem. Phys.* **2009**, *131*, 074702 (cit. on p. 59).

[108] B. Peters, "Supersaturation rates and schedules: Nucleation kinetics from isothermal metastable zone widths", *J. Cryst. Growth* **2011**, *317*, 79–83 (cit. on p. 59).

[109] R. P. Sear, "Estimation of the Scaling of the Nucleation Time with Volume When the Nucleation Rate Does Not Exist", *Cryst. Growth Des.* **2013**, *13*, 1329–1333 (cit. on p. 59).

[110] M. R. Thorson, S. Goyal, Y. Gong, G. G. Z. Zhang, P. J. A. Kenis, "Microfluidic approach to polymorph screening through antisolvent crystallization", *CrystEngComm* **2012**, *14*, 2404 (cit. on p. 59).

[111] P. G. Vekilov, "Crystal nucleation: Nucleus in a droplet", *Nat. Mater.* **2012**, *11*, 838–840 (cit. on p. 59).

[112] A. Jawor-Baczynska, J. Sefcik, B. D. Moore, "250 nm glycine-rich nanodroplets are formed on dissolution of glycine crystals but are too small to provide productive nucleation sites", *Cryst. Growth Des.* **2013**, DOI 10.1021/cg300150u (cit. on p. 59).

[113] P. G. Vekilov, M. A. Vorontsova, "Nucleation precursors in protein crystallization", *Acta Crystallogr. Sect. FStructural Biol. Commun.* **2014**, *70*, 271–282 (cit. on p. 59).

[114] Z. K. Nagy, M. Fujiwara, X. Y. Woo, R. D. Braatz, "Determination of the kinetic parameters for the crystallization of paracetamol from water using metastable zone width experiments", *Ind. Eng. Chem. Res.* **2008**, *47*, 1245–1252 (cit. on pp. 60, 77, 184).

[115] G. M. Maggioni, L. Bosetti, E. dos Santos, M. Mazzotti, "Statistical Analysis of Series of Detection Time Measurements for the Estimation of Nucleation Rates", *Cryst. Growth Des.* **2017**, *17*, 5488–5498 (cit. on pp. 61, 121, 122, 165, 169, 185, 231).

[116] M. H. DeGroot, M. J. Schervish, C. Sheet, *Probability and Statistics*, Pearson, **2011**, 911 (cit. on p. 63).

[117] K. Borovkov, *Elements of Stochastic Modelling*, 2nd ed., World Scientific, Singapore, **2014** (cit. on pp. 63, 65).

[118] M. R. Kosorok, *Introduction to Empirical Processes and Semiparametric Inference*, Springer-Verlag, **2008** (cit. on p. 65).

[119] A. Dvoretzky, J. Kiefer, J. Wolfowitz, "Asymptotic Minimax Character of the Sample Distribution Function and of the Classical Multinomial Estimator", *Ann. Math. Stat.* **1956**, *27*, 642–669 (cit. on pp. 65, 133).

[120] P. Massart, "The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz Inequality", *Ann. Probab.* **1990**, *18*, 1269–1283 (cit. on pp. 65, 133).

[121] R. L. Wasserstein, N. A. Lazar, "The ASA's Statement on p -Values: Context, Process, and Purpose", *Am. Stat.* **2016**, *70*, 129–133 (cit. on pp. 69, 97).

[122] R. A. Fisher, "On the Mathematical Foundations of Theoretical Statistics", *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **1922**, *222*, 309–368 (cit. on p. 70).

[123] F. Mosteller, R. A. Fisher, "Questions and Answers", *Am. Stat.* **1948**, *2*, 30 (cit. on p. 70).

[124] H. M. J. Hung, R. T. O'Neill, P. Bauer, K. Kohne, "The Behavior of the P-Value When the Alternative Hypothesis is True", *Biometrics* **1997**, *53*, 11 (cit. on p. 70).

[125] C. Lindenberg, M KraÌˆttli, "Design and optimization of a combined cooling/antisolvent crystallization process", *Cryst. Growth Des.* **2009**, *9*, 1124–1136 (cit. on p. 72).

[126] P. M. Frank, *Introduction to System Sensitivity Theory*, Academic Press, **1978** (cit. on p. 99).

[127] A. Chianese, H. J. M. Kramer, *Industrial Crystallization Process Monitoring and Control*, (Eds.: A. Chianese, H. J. M. Kramer), Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, **2012** (cit. on pp. 106–108).

[128] M. Parisi, A. Chianese, "Investigation of Nucleation Kinetics by Nephelometric Measurements", *Chem. Eng. Technol.* **2013**, *36*, 1335–1340 (cit. on p. 108).

[129] M. Parisi, A. Chianese, "The Use of Nephelometric Measurements in Crystallization Operations", **2013**, *32*, 2065–2070 (cit. on p. 108).

[130] S. H. Chung, D. L. Ma, R. D. Braatz, "Optimal model-based experimental design in batch crystallization", *Chemom. Intell. Lab. Syst.* **2000**, *50*, 83–90 (cit. on p. 108).

[131] J. Liu, Å. C. Rasmuson, "Influence of Agitation and Fluid Shear on Primary Nucleation in Solution", *Cryst. Growth Des.* **2013**, *13*, 4385–4394 (cit. on p. 112).

[132] J. Liu, M. Svärd, Å. C. Rasmuson, "Influence of Agitation on Primary Nucleation in Stirred Tank Crystallizers", *Cryst. Growth Des.* **2015**, *15*, 4177–4184 (cit. on p. 112).

[133] C. Forsyth, P. A. Mulheran, C. Forsyth, M. D. Haw, I. S. Burns, J. Sefcik, "Influence of Controlled Fluid Shear on Nucleation Rates in Glycine Aqueous Solutions", *Cryst. Growth Des.* **2015**, *15*, 94–102 (cit. on pp. 112, 122).

[134] I. de Albuquerque, M. Mazzotti, D. R. Ochsenbein, M. Morari, "Effect of needle-like crystal shape on measured particle size distributions", *AIChE J.* **2016**, *62*, 2974–2985 (cit. on p. 112).

[135] H. Yang, M. Svärd, J. Zeglinski, Å. C. Rasmuson, "Influence of Solvent and Solid-State Structure on Nucleation of Parabens", *Cryst. Growth Des.* **2014**, *14*, 3890–3902 (cit. on p. 121).

[136] C. A. Pons Siepermann, S. Huang, A. S. Myerson, "Nucleation Inhibition of Benzoic Acid through Solution Complexation", *Cryst. Growth Des.* **2017**, *17*, 2646–2653 (cit. on p. 121).

[137] T. B. Hansen, A. Taris, B.-G. Rong, M. Grosso, H. Qu, "Polymorphic behavior of isonicotinamide in cooling crystallization from various solvents", *J. Cryst. Growth* **2016**, *450*, 81–90 (cit. on pp. 128, 156, 163, 164, 191).

[138] S. A. Kulkarni, E. S. McGarrity, H. Meekes, J. H. ter Horst, "Isonicotinamide self-association: the link between solvent and polymorph nucleation", *Chem. Commun.* **2012**, *48*, 4983 (cit. on pp. 128, 156, 162, 164, 167, 177, 190, 191).

[139] A Kolmogorov, "Confidence limits for an unknown distribution function", *Ann. Math. Stat.* **1941**, *12*, 461–463 (cit. on p. 132).

[140] N. Smirnov, "Table for Estimating the Goodness of Fit of Empirical Distributions", *Ann. Math. Stat.* **1948**, *19*, 279–281 (cit. on p. 132).

[141] F. J. Massey, "The Kolmogorov-Smirnov Test for Goodness of Fit", *J. Am. Stat. Assoc.* **1951**, *46*, 68–78 (cit. on pp. 132, 134).

[142] S. M. Ross, *Introduction to probability and statistics for engineers and scientists*, Academic Press, **2004** (cit. on p. 149).

[143] R. A. Fisher, *Statistical methods for research workers*, Fifth, Oliver & Boyd, Edinburgh, **1925** (cit. on p. 153).

[144] G. M. Maggioni, L. Bezinge, M. Mazzotti, "Stochastic Nucleation of Polymorphs: Experimental Evidence and Mathematical Modeling", *Cryst. Growth Des.* **2017**, *17*, 6703–6711 (cit. on pp. 161, 231).

[145] W Ostwald, "Studien über die Bildung und Umwandlung fester Körper", *Zeitschrift für Phys. Chemie* **1897**, *22*, 289–302 (cit. on pp. 161, 190).

[146] J. Nývlt, "The Ostwald Rule of Stages", *Cryst. Res. Technol.* **1995**, *30*, 443–449 (cit. on pp. 161, 190).

[147] P. T. Cardew, R. J. Davey, A. J. Ruddick, "Kinetics of polymorphic solid-state transformations", *J. Chem. Soc. Faraday* **1984**, *80*, 659–668 (cit. on pp. 161, 190).

[148] R. Davey, P. Cardew, D. McEwan, D. Sadler, "Rate controlling processes in solvent-mediated phase transformations", *J. Cryst. Growth* **1986**, *79*, 648–653 (cit. on pp. 161, 190).

[149] R. Davey, J. Garside, *From molecules to crystallizers*, **2000**, 1 (cit. on pp. 161, 190).

[150] T. B. Hansen, H. Qu, "Formation of Piroxicam Polymorphism in Solution Crystallization: Effect and Interplay of Operation Parameters", *Cryst. Growth Des.* **2015**, *15*, 4694–4700 (cit. on pp. 162, 163).

[151] J. F. B. Black, R. J. Davey, R. J. Gowers, A. Yeoh, "Ostwald's rule and enantiotropy: polymorph appearance in the crystallisation of p-aminobenzoic acid", *CrystEngComm* **2015**, *17*, 5139–5142 (cit. on pp. 162, 163, 177).

[152] R. K. Garg, D. Sarkar, "Polymorphism control of p-aminobenzoic acid by isothermal anti-solvent crystallization", *J. Cryst. Growth* **2016**, *454*, 180–185 (cit. on pp. 162, 163).

[153] A. Caridi, S. A. Kulkarni, G. Di Profio, E. Curcio, J. H. Ter Horst, "Template-induced nucleation of isonicotinamide polymorphs", *Cryst. Growth Des.* **2014**, *14*, 1135–1141 (cit. on pp. 162, 191).

[154] S. A. Kulkarni, H. Meekes, J. H. ter Horst, "Polymorphism Control through a Single Nucleation Event", *Cryst. Growth Des.* **2014**, *14*, 1493–1499 (cit. on pp. 162, 163).

[155] V. Bhamidi, S. H. Lee, G. He, P. S. Chow, R. B. H. Tan, C. F. Zukoski, P. J. A. Kenis, "Antisolvent Crystallization and Polymorph Screening of Glycine in Microfluidic Channels Using Hydrodynamic Focusing", *Cryst. Growth Des.* **2015**, *15*, 3299–3306 (cit. on p. 163).

[156] N. Jiang, Z. Wang, L. Dang, H. Wei, "Effect of supersaturation on l-glutamic acid polymorphs under droplet-based microchannels", *J. Cryst. Growth* **2016**, *446*, 68–73 (cit. on p. 163).

[157] Y. Cui, J. Stojakovic, H. Kijima, A. S. Myerson, "Mechanism of Contact-Induced Heterogeneous Nucleation", *Cryst. Growth Des.* **2016**, *16*, 6131–6138 (cit. on p. 163).

[158] S. G. Agrawal, A. H. J. Paterson, "Secondary Nucleation: Mechanisms and Models", *Chem. Eng. Commun.* **2015**, *202*, 698–706 (cit. on p. 163).

[159] G. M. Maggioni, M. Mazzotti, "Stochastic modelling and statistical analysis of detection time experiments", *48th BACG Annu. Conf.* **2017** (cit. on p. 163).

[160] W. Sun, G. Ceder, "Induction time of a polymorphic transformation", *CrystEngComm* **2017**, *19*, 4576–4585 (cit. on p. 163).

[161] A. Savitzky, M. J. E. Golay, "Smoothing and Differentiation of Data by Simplified Least Squares Procedures", *Anal. Chem.* **1964**, *36*, 1627–1639 (cit. on p. 168).

[162] V. Mazet, C. Carteret, D. Brie, J. Idier, B. Humbert, "Background removal from spectra by designing and minimising a non-quadratic cost function", *Chemom. Intell. Lab. Syst.* **2005**, *76*, 121–133 (cit. on p. 168).

[163] N. B. BaÌthori, A. Lemmerer, G. A. Venter, S. A. Bourne, M. R. Caira, "Pharmaceutical Co-crystals with Isonicotinamide—Vitamin B3, Clofibric Acid, and Diclofenac—and

Two Isonicotinamide Hydrates", *Cryst. Growth Des.* **2011**, *11*, 75–87 (cit. on p. 168).

[164] J. Cornel, C. Lindenberg, M. Mazzotti, "Experimental Characterization and Population Balance Modeling of the Polymorph Transformation of L-Glutamic Acid", *Cryst. Growth Des.* **2009**, *9*, 243–252 (cit. on p. 177).

[165] J. Cornel, P. Kidambi, M. Mazzotti, "Precipitation and Transformation of the Three Polymorphs of d-Mannitol", *Ind. Eng. ...* **2010**, 5854–5862 (cit. on p. 177).

[166] R. C. Zumstein, R. W. Rousseau, "Growth rate dispersion in batch crystallization with transient conditions", *AIChE J.* **1987**, *33*, 1921–1925 (cit. on p. 179).

[167] R. C. Zumstein, R. W. Rousseau, "Growth rate dispersion by initial growth rate distributions and growth rate fluctuations", *AIChE J.* **1987**, *33*, 121–129 (cit. on p. 179).

[168] R. J. Davey, S. L. M. Schroeder, J. H. ter Horst, "Nucleation of Organic Crystals-A Molecular Perspective", *Angew. Chemie Int. Ed.* **2013**, *52*, 2166–2179 (cit. on p. 179).

[169] J. Bernstein, R. J. Davey, J.-O. Henck, "Concomitant Polymorphs", *Angew. Chemie Int. Ed.* **1999**, *38*, 3440–3461 (cit. on p. 190).

A

STOCHASTIC PROCESSES

This appendix chapter contains some of the fundamental mathematical notions used throughout the thesis, either explicitly or implicitly.

In particular, Section A.3 is dedicated to a more precise definition of the Population Balance Equation with randomised initial conditions.

A.1 DEFINITION OF STOCHASTIC PROCESS

Let $(E, \mathscr{E})$ be a measurable space, where E is a set and $\mathscr{E}$ a σ-algebra on E. Let T be an arbitrary set, countable or uncountable. For each t in T, let X_t be a random variable taking values in $(E, \mathscr{E})$. Then, the collection $\{X_t : t \in T\}$ is called a *stochastic process* with state space $(E, \mathscr{E})$ and parameter set T.

For each ω in Ω, let $X(\omega)$ denote the function $t \mapsto X_t(\omega)$ from T into E; then, $X(\omega)$ is an element of $\mathscr{E}^T$. Thus, the mapping $X : \omega \mapsto X(\omega)$ from Ω into $\mathscr{E}^T$ is measurable relative to $\mathscr{H}$, i.e. the σ-algebra on Ω, and $\mathscr{E}^T$. $\mathscr{E}^T$ is the space of all the possible X-valued functions of $t \in T$ that map from the set T into the space E.

Consequently, one may regard the stochastic process $\{X_t : t \in T\}$ simply as a random variable X taking values in the product space $(F, \mathscr{F}) = (E^T, \mathscr{E}^T)$.

The distribution of the random variable X, that is, the probability measure $\mathbb{P} \circ X^{-1}$ on $(F, \mathscr{F})$, is called the probability law of the stochastic process $\{X_t : t \in T\}$.

To underline the dependence of X on t, the notation $X(t)$ is also used.

A.2 COUNTING PROCESSES

A counting process is a stochastic process $N_t, t \geq 0$;

(1) $N_t \geq 0$

(2) $N_t \in \mathbb{N}$

(3) $s \leq t \leftrightarrow N_s \leq N_t$

In other words, it is a process with values that are non-negative, integer, and non-decreasing. Note that typically the process is also right-continuous.

A.2.1 *Poisson Process*

Two equivalent definitions can be given of a non-stationary (also "non-homogeneous" or "inhomogeneous") Poisson process.

The first one, more abstract, relies explicitly on measure theory and states that the counting process N is said to be a Poisson process with rate λ with respect to F if it is adapted to $\mathscr{F}$ and $\forall (s,t) \in \mathbb{R}_+$ and for all positive functions f on $\mathbb{N}$:

$$\mathbb{E}_s f(N_{s+t} - N_s) = \sum_{k=0}^{\infty} \frac{e^{-\Lambda(s,t)} \left(\Lambda(s,t)\right)^k}{k!} f(k) \tag{A.1}$$

N has independent increments and its probability law is completely determined by the positive function $\Lambda(s,t) = \int_s^{s+t} \lambda(z)dz$, also known as "parameter measure" of the process; note that $N_0 = 0$. If the increments are also stationary : $\Lambda(s,t) = \lambda t$.

The second definition is less abstract and states that a counting process $N_t, t \geq 0$ is a Poisson process if:

(1) $N_0 = 0$

(2) $\Pr\left(N_{s+t} - N_s = 1\right) = \lambda(t) + o(t)$

(3) $\Pr\left(N_{s+t} - N_s \geq 2\right) = o(t)$

where $o(t) : o(t)/t \to 0$ if $t \to 0$ and:

$$\Pr(N_{s+t} - N_s = n) = \frac{e^{-\Lambda(s,t)} \left(\Lambda(s,t)\right)^n}{n!} \tag{A.2}$$

with $\Lambda(s,t) = \int_s^{s+t} \lambda(z)dz$

A.3 POPULATION BALANCE EQUATION, STOCHASTIC PROCESSES, AND DIRAC MEASURE

While presenting Model 1 in Section 3.3.1, the Dirac measure δ_x^* was introduced. Although this object and the Dirac-delta function δ share several properties, δ_x^* is actually more general. First, we briefly recall the definitions of δ and δ_x^*; then, we show that it is consistent with the PBE written in Chapter 3.

A.3.0.1 *Dirac Measure*

Let us first look again at the definition of the Dirac delta in the theory of generalised functions. For any function $g : \mathbb{R} \mapsto X \subseteq \mathbb{R}$, with $g \in C_c^\infty(\mathbb{R})$, i.e. g is a smooth function with compact support, δ is the linear functional such that:

$$\delta[g] = \int_{\mathbb{R}} g(x)\delta(x)dx = g(0) \tag{A.3}$$

An equivalent, but more abstract and flexible definition of δ can be given in the framework of measure theory. δ is understood as the purely atomic measure on $\mathbb{R}$, such that:

$$\delta(x) = \begin{cases} 1 & x = 0 \\ 0 & x \neq 0 \end{cases} \tag{A.4}$$

and the integral of the δ-measurable function g on $\mathbb{R}$ is:

$$\int_{\mathbb{R}} g(x)\delta(dx) = g(0) \tag{A.5}$$

Let us now define the more general Dirac measure; one should be careful to the notation, which in this context has a slightly different meaning. The Dirac measure $\delta_x^*(X)$ over the set X (it is assumed here $X \subseteq \mathbb{R}$, but the set can be on more general fields) is:

$$\delta_x^*(X) = \begin{cases} 1 & x \in X \\ 0 & x \notin X \end{cases} \tag{A.6}$$

The integral of the δ_x^*-measurable function g of bounded variation on X is:

$$\int_A g(X) d\delta_x^*(X) = \begin{cases} g(X) & x \in X \\ 0 & x \notin X \end{cases} \tag{A.7}$$

Note that when $g(X) = 1$, i.e. the constant unit function:

$$\int_A g(X) d\delta_x^*(X) = \int_A d\delta_x^*(X) = g(x) = 1 \tag{A.8}$$

With an abuse of notation, one can also write:

$$\int_X \frac{d\delta_x^*(X)}{dX} dX = \int_X d\delta_x^*(X) = \delta_x^*(X) \tag{A.9}$$

where the expression $\dfrac{d\delta_x^*(X)}{dx}$ is only *formally* a derivative. Note that the Dirac measure preserves all the properties of interest of the Dirac delta function, but it is defined on more general domain $X \in \mathbb{R}$.

A.3.0.2 *Application to the PBE*

Let us check that the Dirac measure is consistent with the framework of the PBE used in Chapter 3. First, we note that the 1D-PSD f is defined on $\mathbb{R}^+$, on which it is of bounded variation. Thus, one can apply the result of Eq. (A.9) to Eq. (3.7), with $x = 0$, i.e.

the Dirac measure is purely atomic and sits on 0. Integrating over $A \equiv \mathbb{R}^+$, one obtains:

$$\int_{\mathbb{R}^+} f(t_N, L)dL = \phi_0(t_N) = \frac{1}{W}\int_{\mathbb{R}^+} \frac{d\delta_0^*(L)}{dL}dL = \frac{1}{W}\int_{\mathbb{R}^+} d\delta_0^*(L) = \frac{1}{W} \tag{A.10}$$

which corresponds to Eq. (3.9). A similar result also holds for Equation (3.6), from which Eqs. (3.8) are derived.

Note that, since $X \equiv \mathbb{R}^+$, the use of the Dirac measure also solves an inconsistency rising with the use of the more common δ. It has been pointed out that by using δ, properly defined on $\mathbb{R}$, one recovers the unit doublet upon differentiating the PSD f with respect to L. Also the unit doublet is defined, however, on $\mathbb{R}$, with half of its "mass" located on the negative axis, on which f, though, is not defined. On the contrary, the Dirac measure δ_0^* is defined on $\mathbb{R}^+$, and its pseudo-derivative would thus be defined in the same domain.

CURRICULUM VITAE

PERSONAL DATA

Name	Giovanni Maria Maggioni
Date of Birth	29 May 1988
Place of Birth	Milan, Italy
Citizen of	Italy

EDUCATION

2007 - 2012	Politecnico di Milano, Milan, Italy *Final degree:* Ms. sc. Chemical Engineering
2012 - 2017	Politecnico di Milano, Milan, Italy *Final degree:* Bs. sc. Chemical Engineering
2002 - 2007	Liceo Classico L. Dehon (High School) Monza, Italy *Final degree:* Maturità Classica

PUBLICATIONS

Articles in peer-reviewed journals:

5. G. M. Maggioni, M. Mazzotti, "Modelling the stochastic behaviour of primary nucleation", *Faraday Discuss.* **2015**, *179*, 359–382 .

4. M. Salvalaglio, P. Tiwary, G. M. Maggioni, M. Mazzotti, M. Parrinello, "Overcoming time scale and finite size limitations to compute nucleation rates from small scale well tempered metadynamics simulations", *J. Chem. Phys.* **2016**, *145*, 211925 .

3. G. M. Maggioni, M. Mazzotti, "Stochasticity in Primary Nucleation: Measuring and Modeling Detection Times", *Cryst. Growth Des.* **2017**, *17*, 3625–3635 .

2. G. M. Maggioni, L. Bosetti, E. dos Santos, M. Mazzotti, "Statistical Analysis of Series of Detection Time Measurements for the Estimation of Nucleation Rates", *Cryst. Growth Des.* **2017**, *17*, 5488–5498 .

1. G. M. Maggioni, L. Bezinge, M. Mazzotti, "Stochastic Nucleation of Polymorphs: Experimental Evidence and Mathematical Modeling", *Cryst. Growth Des.* **2017**, *17*, 6703–6711 .

Book Chapters:

1. M. Mazzotti, T. Vetter, D. R. Ochsenbein, G. M. Maggioni, C. Lindenberg in *Polymorph. Pharm. Ind. Solid Form Drug Dev.* Springer-Verlag, **2018**.