



# Quantifying the inherent uncertainty associated with nucleation rates estimated from induction time data measured in small volumes

DOI:

[10.1021/acs.cgd.7b00372](https://doi.org/10.1021/acs.cgd.7b00372)

## Document Version

Accepted author manuscript

[Link to publication record in Manchester Research Explorer](#)

## Citation for published version (APA):

Tang, S. K., Hao, H., Davey, R., Vetter, T., & Xiao, Y. (2017). Quantifying the inherent uncertainty associated with nucleation rates estimated from induction time data measured in small volumes. *Crystal Growth and Design*, 17(5), 2852–2863. <https://doi.org/10.1021/acs.cgd.7b00372>

## Published in:

Crystal Growth and Design

## Citing this paper

Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

## General rights

Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

## Takedown policy

If you believe that this document breaches copyright please refer to the University of Manchester's Takedown Procedures [<http://man.ac.uk/04Y6Bo>] or contact [openresearch@manchester.ac.uk](mailto:openresearch@manchester.ac.uk) providing relevant details, so we can investigate your claim.



## Article

## Quantifying the inherent uncertainty associated with nucleation rates estimated from induction time data measured in small volumes

Yan Xiao, Sin Kim Tang, Hongxun Hao, Roger J. Davey, and Thomas Vetter

*Cryst. Growth Des.*, **Just Accepted Manuscript** • DOI: 10.1021/acs.cgd.7b00372 • Publication Date (Web): 14 Apr 2017

Downloaded from <http://pubs.acs.org> on April 19, 2017

### Just Accepted

“Just Accepted” manuscripts have been peer-reviewed and accepted for publication. They are posted online prior to technical editing, formatting for publication and author proofing. The American Chemical Society provides “Just Accepted” as a free service to the research community to expedite the dissemination of scientific material as soon as possible after acceptance. “Just Accepted” manuscripts appear in full in PDF format accompanied by an HTML abstract. “Just Accepted” manuscripts have been fully peer reviewed, but should not be considered the official version of record. They are accessible to all readers and citable by the Digital Object Identifier (DOI®). “Just Accepted” is an optional service offered to authors. Therefore, the “Just Accepted” Web site may not include all articles that will be published in the journal. After a manuscript is technically edited and formatted, it will be removed from the “Just Accepted” Web site and published as an ASAP article. Note that technical editing may introduce minor changes to the manuscript text and/or graphics which could affect content, and all legal disclaimers and ethical guidelines that apply to the journal pertain. ACS cannot be held responsible for errors or consequences arising from the use of information contained in these “Just Accepted” manuscripts.



# Quantifying the inherent uncertainty associated with nucleation rates estimated from induction time data measured in small volumes

Yan Xiao,<sup>†,‡</sup> Sin Kim Tang,<sup>¶</sup> Hongxun Hao,<sup>†,‡</sup> Roger J. Davey,<sup>¶</sup> and Thomas Vetter\*,<sup>¶</sup>

<sup>†</sup>*State Key Laboratory of Chemical Engineering, School of Chemical Engineering and Technology, Tianjin University, Tianjin 300072, P. R. China*

<sup>‡</sup>*Collaborative Innovation Center of Chemical Science and Chemical Engineering, Tianjin 300072, P.R. China*

<sup>¶</sup>*University of Manchester, School of Chemical Engineering and Analytical Science, M13 9PL Manchester, United Kingdom*

E-mail: thomas.vetter@manchester.ac.uk

## Abstract

Unravelling the molecular complexities of crystal nucleation from solutions is predicated on our ability to measure and interpret high quality kinetic data. This allows us to link nucleation rates to supersaturation, as well as to kinetic rate expressions and their parameters, arising from mechanistic considerations. In this context it is vital to be able to assess the reliability of measured nucleation rate data. Accordingly this contribution details a statistical approach that aims at quantifying the inherent uncertainty associated with nucleation rates obtained from induction time measurements carried out in small volumes. We investigate how uncertainties attached to nucleation rates propagate to mechanistic parameters derived from them, and make recommendations for experimental protocols, as well as data analysis strategies that minimize said uncertainty. The approach is applied to induction time measurements obtained for benzoic acid/toluene solutions in a wide range of supersaturations.

## 1 Introduction

The nucleation of organic crystals from solution is a well-studied<sup>1-4</sup>, but by no means fully understood, phenomenon. It remains a challenging subject to study because nuclei are so small that their direct observation is not possible even with current state of the art measurement techniques.

1  
2  
3  
4 5 In recent years, studying nucleation has seen a renaissance. This has been driven by a number of  
5 6 key developments – an increasing awareness of the importance of nucleation with respect to poly-  
6 7 morphic outcome of crystallization processes<sup>5</sup>, the availability of increasing computational power  
8 9 and improved simulation methods enabling the behaviour of larger ensembles of molecules to be  
10 11 studied<sup>6–9</sup>, the availability of off-the-shelf medium-throughput reactor devices for performing mul-  
12 13 tiple repeats of identical experiments and, vitally, a range of easily accessible in situ microscopic  
14 15 and spectroscopic techniques. The latter has enabled a comparison to be made between the extent  
16 17 of self-assembly or solvation of molecules in solution and the building blocks that constitute their  
18 19 crystal structures<sup>10,11</sup>.

20 21 The impact of medium throughput methods for determining nucleation rates can be seen, for  
22 23 example, in the Crystal16 methodology of Jiang and ter Horst<sup>12</sup>. They performed multiple experi-  
24 25 mental repeats of induction time measurements at the same experimental conditions (supersatura-  
26 27 tion, temperature, etc.). Performing such experiments in small solution volumes, which is beneficial  
28 29 from a standpoint of material consumption, accentuates the stochasticity of nucleation. This is  
30 31 seen in the fact that a series of experimental repeats shows a wide distribution of induction times.  
32 33 Jiang and ter Horst<sup>12</sup> developed an approach to derive nucleation rates (with further modifications  
34 35 and applications published later<sup>13,14</sup>) from such stochastic data. To the same end, microfluidic  
36 37 devices have been used to perform multiple repeats in flowing droplet systems; also resulting in  
38 39 nucleation rates<sup>15</sup>. One crucial outcome of this work has been the increasing availability of kinetic  
40 41 data relating nucleation rates to supersaturation and to solvent choice. Such methodologies have  
42 43 been used to investigate the nucleation kinetics of a variety of molecules including carboxylic acids  
44 45 acids<sup>11,12,15–17</sup>, amino acids<sup>12</sup>, amides<sup>13,14</sup> and ketones<sup>18</sup>.

46 47 Surprisingly, the reliability of nucleation rates derived from such inherently stochastic induction  
48 49 time measurements has only been tested in a cursory manner in the past<sup>12,14</sup>, i.e., a level of un-  
50 51 certainty was estimated for a specific case, but it was not quantified how such uncertainty would  
52 53 change under different conditions (for example the number of experimental repeats, etc.). Further-  
54 55 more, the statistical assumptions used in these works remained largely untested. Other authors did  
56 57 not at all quantify the uncertainty attached to either the nucleation rates<sup>15,19–21</sup> or the quantities  
58 59 derived from them<sup>11,18,22</sup>. However, without knowing how reliable nucleation rate datapoints are, it  
60 61 is for example impossible to say whether differences observed between two cases have any statistical

1  
2  
3  
4 35 significance or whether they are coincidental. The same is true for correlations between parameters  
5  
6 36 derived from such estimates of nucleation rates as the uncertainty propagates from rates to param-  
7  
8 37 eters in rate equations. This remains true regardless of what theory has been used to derive the  
9  
10 38 nucleation rate expressions.

11 39 In this article we conduct a statistical analysis of the induction time methodology<sup>12</sup> with the  
12  
13 40 ultimate goal of assessing the confidence levels which may be placed on nucleation rates obtained  
14  
15 41 in this way. To do this, we first revisit the critical assumptions made in developing this technique  
16  
17 42 and recall the cases where these assumptions might not be applicable. Second, we generate large  
18  
19 43 amounts of synthetic induction time data and proceed to establish an analysis approach that allows  
20  
21 44 us to quantify the uncertainty attached to estimates of nucleation rates. Furthermore, we extensively  
22  
23 45 quantify how the level of uncertainty changes with the number of measurements taken (as well as  
24  
25 46 with other factors). This allows us to give recommendations for designing experimental campaigns  
26  
27 47 and for data analysis procedures that yield optimum estimates of nucleation rates (i.e., as reliable as  
28  
29 48 they can be given a certain dataset and measurement procedure). We then show a way to connect  
30  
31 49 the uncertainty quantification obtained from such synthetic data to experimental studies. Finally,  
32  
33 50 we present an updated experimental methodology to obtain induction time data using a Crystal16,  
34  
35 51 which we have used to gather nucleation rate data for the monomorphic compound benzoic acid  
36  
37 52 (CSD refcode BENZAC01) crystallized from toluene.

## 38 39 53 **2 Background**

### 40 41 42 54 **2.1 Nucleation rates from induction times**

43  
44 55 The use of induction time data for measuring nucleation kinetics has been well documented<sup>12,14,22</sup>.  
45  
46 56 Measuring induction times in small volumes has the advantage of lower material consumption and  
47  
48 57 better uniformity of the vessel (or droplet) content, i.e., mixing is typically stronger and tempera-  
49  
50 58 ture control easier to achieve than in larger crystallizers. However, working at small volumes, the  
51  
52 59 stochastic nature of nucleation becomes more apparent than in larger vessels. Indeed, performing  
53  
54 60 many equivalent, independent experiments (same supersaturation, temperature etc.) at solution  
55  
56 61 volumes in the order of 1 mL leads to a wide distribution of observed induction times. Sorting the  
57  
58 62 individual induction times,  $t_{D,i}$  in ascending order, a cumulative probability,  $P_i$  can be obtained  
59  
60

$$P_i = \frac{i}{N_{\text{exp}}} \quad \text{for} \quad i = 1, \dots, N_{\text{nuc}} \quad (1)$$

where  $N_{\text{exp}}$  represents the total number of induction time measurements made and  $N_{\text{nuc}}$  is the number of experiments where nucleation was detected (note  $N_{\text{nuc}} \leq N_{\text{exp}}$ ). In order to characterize nucleation rates from such cumulative probability distributions, a link between the stochasticity exhibited in the series of experiments and the random processes occurring in each individual experiment must be made. Usually this is done by assuming that the formation of each nucleus occurs independently from the others within a given experiment (at least until the nuclei can be observed). If one assumes that supersaturation is generated quickly and kept constant, the probability that at least one nucleus has formed in time  $t_N$  can be expressed in the form of a Poisson distribution<sup>12,14</sup>:

$$P(t_N) = 1 - \exp(-JVt_N) \quad (2)$$

where  $J$  is the nucleation rate and  $V$  is the volume of the solution. This equation assumes that nuclei are detected immediately upon their formation and also before their formation changes the solution properties (supersaturation, temperature, etc.). However, since the size of nuclei falls well below the detection limit of available instruments, changes in some proxy property of the solution (e.g. turbidity, concentration) are used in practice to indicate the onset of nucleation. In the case of detection based on microscopy or turbidimetry, crystal sizes of around a micron and/or sufficient volume fractions are required to detect particles. Clearly, if an instrument can only detect nuclei after they have grown to a considerable size or when present in large numbers, the assumption of a constant nucleation rate (i.e., constant supersaturation) in Eq. (2) is violated and the decrease of supersaturation caused by the formation of nuclei and their subsequent growth should be accounted for, as shown by Maggioni and Mazzotti<sup>23</sup>. While this consideration remains valid regardless of the nucleation mechanism, there are limiting cases in which a modified form of Eq. (2) represents a good approximation and its comparable simplicity is retained. In earlier work of Dugua and Simon on the nucleation of sodium perborate<sup>24</sup> and Kondepudi et al.<sup>25</sup> on sodium chlorate, the nucleation behaviour was rationalized on the basis of an initial low (primary) nucleation rate followed by extensive secondary nucleation. This phenomenon is referred to as the “single nucleus mechanism”<sup>12,14,26</sup> where an initial nucleus forms through primary nucleation and

grows to a certain size. At this point, it becomes large enough to undergo abrasion or a complete shattering, triggering rapid secondary nucleation, which is detected via a rapid increase in, for example, turbidity. The resulting detection time may then be formulated as

$$t_D = t_N + t_g + t_e \quad (3)$$

where  $t_D$  is the detection time,  $t_N$  is the time at which the single nucleus has formed,  $t_g$  is the time for the nucleus to grow to the size at which it triggers secondary nucleation and  $t_e$  is the time elapsed between secondary nucleation and the detection of particles by the measurement device. While any or all of the terms on the right hand side of Eq. (3) could be stochastic (e.g.,  $t_g$  could be stochastic if growth rate dispersion<sup>27,28</sup> occurs) the prevalent assumptions made in the literature<sup>12</sup> are that  $t_e$  is negligible and  $t_g$  is deterministic. Using these assumptions, Eq. (2) can be modified using Eq. (3) to obtain:

$$P(t_D) = 1 - \exp(-JV(t_D - t_g)) \quad (4)$$

While there is experimental evidence supporting the single nucleus mechanism<sup>24,25</sup> and hence allowing the use of the simple approximation, Eq. (4), we believe that it is vital to bear in mind the underlying assumptions enabling this simplification. In particular the assumption of constant supersaturation up to the point of nuclei detection might be violated for systems with a high detection threshold (instrument specific) and low solubility (solute/solvent/temperature specific). However, we deem an extensive exploration of this conundrum to be outside the scope of the present work. The experimental cumulative probability distributions (Eq. (1)) can then be fitted with the nonlinear Eq. (4) by minimizing the sum of squared differences between model and experiment, defined as:

$$\Phi_{J,NL} = \sum_{i=1}^{N_{nuc}} (P_i - P(t_{D,i}))^2 \quad (5)$$

where  $P_i$  are the points on the experimental cumulative probability curve (Eq. (1)) and  $P(t_{D,i})$  are the corresponding points calculated with Eq. (4). Alternatively, the nucleation rate can be obtained by fitting a linearized version of Eq. (4), i.e.:

$$\ln(1 - P(t_D)) = -JV(t_D - t_g) \quad (6)$$

to the data using

$$\Phi_{J,L} = \sum_{i=1}^{N_{\text{nuc}}} (\ln(1 - P(t_{D,i})) - \ln(1 - P_i))^2 \quad (7)$$

In these two alternatives, the data points in Eq. (1) are weighted differently, which in general results in different estimates of the nucleation rate  $J$ . Several approaches to arrive at  $t_g$  values have been introduced in the literature. It can be taken as the shortest measured induction time within a dataset<sup>11,12</sup>. Here one makes the implicit assumption that a sufficient number of induction times has been measured, so that one of them represents an instance where  $t_N = 0$ . Alternatively  $t_g$  can be estimated as a parameter when fitting Eq. (4) or Eq. (6) to the data (Eq. (1)), but this is often found to give meaningless, negative values for  $t_g$ . Finally it can, of course, be calculated if growth kinetics are available and the size at which secondary nucleation is triggered is known<sup>12-14</sup>. The first of these approaches seems most appealing, as it requires the least amount of prior, independent information. It appears that an assessment of the most successful of these for determining the true nucleation rate has not been carried out in the literature.

## 2.2 Nucleation rate expressions and parameters according to classical nucleation theory

Nucleation rates  $J$  obtained from experiments carried out at different supersaturations  $S$  can be fitted by nucleation rate expressions that are derived from mechanistic assumptions. In this contribution rate expressions stemming from classical nucleation theory (CNT) will be used, but the statistical approach presented in this paper is applicable to any rate expression derived from other theories, e.g., two step nucleation theory<sup>2</sup>. A formal derivation of CNT rate expressions and their underlying assumptions can be found elsewhere<sup>1,10</sup>. Depending on the rate-limiting step assumed in CNT, the nucleation rate  $J$  can be written as

$$J = AS \exp\left(-\frac{B}{\ln^2 S}\right) \quad (8)$$

1  
2  
3  
4 131 OR

$$J = AS \ln S \exp\left(-\frac{B}{\ln^2 S}\right) \quad (9)$$

5  
6  
7  
8 132 for interface-transfer or volume-diffusion control, respectively. The supersaturation is defined as  $S =$   
9  
10 133  $x/x_s$  where  $x$  and  $x_s$  are the mole fractions of the solute in the supersaturated and the equilibrium  
11  
12 134 state, respectively. The two parameters  $A$  and  $B$  are typically referred to as the “thermodynamic”  
13  
14 135 parameter  $B$ :

$$B = \frac{16\pi v_0^2 \gamma^3}{3k_b^3 T^3} \quad (10)$$

15  
16  
17  
18 136 and the “kinetic” parameter  $A$ :

$$A = \frac{f_0 C_0}{\sqrt{12\pi B}} \quad (11)$$

19  
20  
21  
22  
23 137 In Eqs. (10) and (11),  $v_0$  is the molecule volume,  $\gamma$  the interfacial energy of the cluster/solution  
24  
25 138 interface,  $k_b$  the Boltzmann constant,  $T$  the absolute temperature,  $f_0$  the supersaturation indepen-  
26  
27 139 dent part of the attachment frequency of building units to a cluster and  $C_0$  the concentration of  
28  
29 140 nucleation sites.

30  
31 141  $A$  and  $B$  (for interface-transfer or volume-diffusion control, respectively) may be derived either  
32  
33 142 by fitting the nonlinear Eqs. (8) and (9) to the data or via the equations’ linearized forms, Eqs. (12)  
34  
35 143 and (13).

$$\ln\left(\frac{J}{S}\right) = \ln A - \frac{B}{\ln^2 S} \quad (12)$$

$$\ln\left(\frac{J}{S \ln S}\right) = \ln A - \frac{B}{\ln^2 S} \quad (13)$$

36  
37  
38  
39  
40  
41  
42  
43  
44 144 In both cases, the fitting is typically conducted with a least squares approach. Additionally, we  
45  
46 145 note that Dunning and Notley<sup>29</sup> derived a further linearized form of the CNT equation in 1957:

$$\ln J + 3 \ln(\ln S) = \ln wn + \ln\left(\frac{2av_0^2 \sigma^3}{k_b^3 T^3}\right) - \frac{\pi q^2 O}{k_b^2 T^2 \ln S} - \frac{av_0^2 \gamma^3}{k_b^3 T^3 \ln^2(S)} \quad (14)$$

47  
48  
49  
50  
51  
52 146 where  $w$  is a frequency factor,  $n$  is the number of critical nuclei,  $a$  is a geometrical factor,  $q$  is  
53  
54 147 the free energy per unit length of the edge of the surface nucleus, and  $O$  is the area occupied by  
55  
56 148 a molecule on the cluster surface. This form of the equation is related to the interface transfer  
57  
58 149 controlled Eq. (8), but differs in that it specifically assumes nucleus growth to take place by a  
59  
60

1  
2  
3  
4 150 surface nucleation mechanism. This introduces additional supersaturation dependent terms in the  
5  
6 151 equation, but plots of  $\ln J + 3 \ln(\ln S)$  vs  $\ln^{-2} S$  for the nucleation of cyclonite and sucrose showed  
7  
8 152 a good linear relationship suggesting that the last term in Eq. (14) is dominant<sup>29,30</sup>.

### 11 3 Statistical analysis approach

12  
13  
14 154 The methodology employed here, i.e., using measured induction probabilities to access nucleation  
15  
16 155 rates, has found considerable use, but only limited assessment has been made<sup>12</sup> of the reliability of  
17  
18 156 both the derived nucleation rates and the fitted values of  $A$  and  $B$  in the CNT equations. Here we  
19  
20 157 set out to explore more fully the impact of various data fitting methods and the relationship between  
21  
22 158 the number of measured data points and the confidence that we may place on the calculated value  
23  
24 159 of the nucleation rate (and by extension the parameters  $A$  and  $B$ ).

#### 26 3.1 Generation of synthetic induction times

27  
28  
29 161 The following procedure was adopted: firstly, a range of supersaturations was chosen and at each  
30  
31 162 supersaturation the “true” nucleation rate was calculated using Eq. (8) together with the values of  
32  
33 163  $A$  and  $B$  given in Table 1. Random numbers were then generated between 0 and 1 to represent  
34  
35 164 “synthetic” probabilities of induction,  $P$ . The random numbers were generated using the imple-  
36  
37 165 mentation of the Mersenne twister<sup>31</sup> as built into MatLAB2015a. For each of these probabilities a  
38  
39 166 “synthetic” induction time was calculated by using Eq. (4) together with the “true” nucleation rate  
40  
41 167 and the growth time,  $t_g$  calculated using Eqs. (15) and (16):

$$44 \quad G = k(S - 1)^g \quad (15)$$

$$48 \quad t_g = \frac{r_a}{G} \quad (16)$$

51  
52 168 where  $G$  is the growth rate,  $g$  and  $k$  are the parameters of crystal growth (Table 1), and  $r_a$  is  
53  
54 169 the size of the nucleus before rapid secondary nucleation occurs, taken to be 10  $\mu\text{m}$  (without loss  
55  
56 170 of generality). Nucleation rates were then obtained by fitting these data to Eq. (4) or Eq. (6)  
57  
58 171 and by using  $t_g$  estimated in one of three ways: firstly by making it a parameter to be identified

1  
2  
3  
4 172 during the fitting, secondly taking it to be the shortest induction time and finally using Eqs. (15)  
5  
6 173 and (16). In this way, it was possible to create sets of synthetic data corresponding to any number  
7  
8 174 of experimental observations ( $N_{\text{exp}}$ ) at any supersaturation ( $S$ ) and to repeat the “experimental  
9  
10 175 campaign” as many times as desired. Thus, for the five supersaturations in Table 1, experiments  
11  
12 176 were simulated in which  $N_{\text{exp}} = 20, 30, 40, \dots, 400$ . By repeating each of these experimental  
13  
14 177 campaigns 10,000 times large numbers of  $J$  values were generated which enabled their probability  
15  
16 178 density distribution and confidence intervals to be explored as functions of both  $N_{\text{exp}}$  and  $S$ .

17 Table 1: Parameters used in the induction times generation process

Parameter	Value
$A$	$8 \times 10^5 \text{ m}^{-3} \text{ s}^{-1}$
$B$	3.5
$k$	$1 \times 10^{-8} \text{ m s}^{-1}$
$g$	2
$r_a$	$1 \times 10^{-5} \text{ m}$
$V$	$1.5 \times 10^{-6} \text{ m}^3$
$N_{\text{exp}}$	20, 30, $\dots$ , 400
$S$	1.6, 1.7, $\dots$ , 2.0
Experimental repeats	10,000

### 179 3.2 Derivation and reliability of nucleation rate values

180 These simulations provide, at each  $S$  and  $N_{\text{exp}}$ , 10,000 values of  $J$  for each fitting method and for  
181 each way of dealing with  $t_g$ . For each fitting method and set of  $J$  values both the complete probability  
182 density distribution ( $p$ ) of  $J$  and the 95% confidence interval as a function of  $N_{\text{exp}}$  were calculated.  
183 These computations were designed to aid an assessment of the best method for calculating  $J$  from  
184 experimental data and the reliability of the obtained  $J$  values (assuming the single nucleus model  
185 applies and that the supersaturation stays constant up to nuclei detection; see Section 2.1). We  
186 also highlight that for low  $S$  the available experimental time (typically 5-8 hours) is insufficient  
187 for all the vials to nucleate and hence the number of measured induction times,  $N_{\text{nuc}} < N_{\text{exp}}$ .  
188 Consequently the largest value of  $P$  inferred from the detection times is less than one. To assess  
189 the reliability of the nucleation rates in such cases, we defined a threshold value of  $P = P_{\text{cutoff}}$ ,  
190 being the fraction of data points used in calculating  $J$ .  $P_{\text{cutoff}}$  was assigned values between 0.1 to  
191 1 in the data presented here. In practice, low values of  $P_{\text{cutoff}}$  are the result of slow nucleation

1  
2  
3  
4 192 at low supersaturations while larger values of  $P_{\text{cutoff}}$  result at high supersaturations (at constant  
5  
6 193 maximum experiment/observation time). Allowing more time for observation in experiments yields  
7  
8 194 higher  $P_{\text{cutoff}}$  values. The choice of  $P_{\text{cutoff}}$  as a representative number (rather than a maximum  
9  
10 195 observation time) is beneficial, because unlike the observation time, the probability is dimensionless  
11  
12 196 and hence transferrable between solute/solvent systems (that exhibit different kinetics).

### 14 197 3.3 Reliability of CNT Parameters and derived properties

16  
17 198 The above procedure allows quantitative assessment of the uncertainty of nucleation rate values  
18  
19 199 obtained from a set of induction times. This uncertainty will propagate to the CNT parameters  $A$   
20  
21 200 and  $B$  as well as any subsequently derived parameters. To investigate this, we sample the synthetic  
22  
23 201  $J$  distributions (one nucleation rate for each supersaturation; for  $n$  samples) and proceed to fit the  
24  
25 202 obtained data sets in a least squares sense with the linear or nonlinear version of the CNT rate  
26  
27 203 equations, Eqs. (8) and (12) (or their equivalent for other mechanisms), to obtain the parameters  
28  
29 204  $A$  and  $B$ . We minimize the sum of squares by varying  $A$  and  $B$ , for instance in Eq. (12), to achieve  
30  
31 205 this:

$$32 \quad \Phi_{\text{CNT}} = \sum_{k=1}^K \left( \ln \left( \frac{J_k}{S_k} \right) - \ln \left( \frac{J_{\text{exp},k}}{S_k} \right) \right)^2 \quad (17)$$

33  
34  
35 206 where  $K$  is the number of supersaturations for which induction time data was measured,  $S_k$  the  
36  
37 207 supersaturations,  $J_k$  the calculated nucleation rates at the present  $A$  and  $B$  values, and  $J_{\text{exp},k}$  the  
38  
39 208 nucleation rates estimated from the experimental cumulative probability distributions. Repeating  
40  
41 209 this fitting for many samples drawn from the underlying probability distributions of  $J$  leads in  
42  
43 210 turn to probability distributions of the CNT parameters from which confidence intervals etc. can  
44  
45 211 be calculated. We then compare these results against the “real” parameter values (see Table 1) so  
46  
47 212 as to decide which fitting method leads to better parameter estimates.

48  
49 213 When estimating the CNT parameters  $A$  and  $B$  from real experimental data, the  $J$  values at  
50  
51 214 different supersaturations are often, out of necessity, determined from cumulative probability curves  
52  
53 215 containing a varying number of data points  $N_{\text{nuc}}$  at each supersaturation (see, e.g., Table 5). As we  
54  
55 216 will show below, this translates into different levels of uncertainty (different widths of confidence  
56  
57 217 intervals) in the estimated nucleation rates. We will quantify in the results section whether there is  
58  
59 218 a benefit in accounting for this variability in the uncertainty attached to the data points at different  
60

1  
2  
3  
4 219 supersaturations. To this end, we define the weighted sum of squares, to be minimized by varying  
5  
6 220  $A$  and  $B$ , as:

$$\Phi_{\text{CNT}} = \sum_{k=1}^K w_k \left( \ln \left( \frac{J_k}{S_k} \right) - \ln \left( \frac{J_{\text{exp},k}}{S_k} \right) \right)^2 \quad (18)$$

7  
8  
9  
10 221 The weights  $w_k$  should be chosen according to<sup>32</sup>

$$w_k = \frac{1}{\sigma_k^2} \quad (19)$$

11  
12  
13  
14  
15  
16  
17 222 where  $\sigma_k^2$  is the variance of  $\ln(J_{\text{exp},k}/S_k)$ , which can be obtained from the synthetic data as briefly  
18  
19 223 mentioned above and further detailed in the results section. Using such weights the values of  $J$  are  
20  
21 224 more heavily weighted the more data points are available (since these  $j$  values carry less uncertainty).  
22  
23 225 While this will always decrease the uncertainty around parameter estimates, we will quantify this  
24  
25 226 effect in the results section.

## 27 28 227 **4 Experimental methods**

29  
30  
31 228 Benzoic acid (BA, Acros Organics,  $\geq 99\%$ ) and toluene (Sigma-Aldrich,  $\geq 99.9\%$ ) were used as-  
32  
33 229 received. Solubility data of BA in toluene were measured gravimetrically and are reported in Table  
34  
35 230 S1 in the supporting information.

36  
37 231 Induction times were measured in toluene using the Crystal16 multiple reactor setup (Technobis  
38  
39 232 Crystallization Systems). Each of its 16 wells can hold one standard HPLC glass screw topped vial  
40  
41 233 of maximum volume approximately 1.8 mL; stirring is by small variable speed magnetic stirrers;  
42  
43 234 the presence of crystals is detected turbidimetrically. The difference between the time at which the  
44  
45 235 turbidity starts to increase and time zero at which supersaturation was established was taken as the  
46  
47 236 detection time (Eq. (3)). To minimize the difference between the set and actual temperature in the  
48  
49 237 wells, a temperature recalibration of the Crystal16 was performed. In designing an experimental  
50  
51 238 protocol for these measurements, a number of issues were considered. It is known from previous  
52  
53 239 work<sup>11,12,14</sup> that the optimum combination of solubility, stirring and solution volume is essential  
54  
55 240 in avoiding two major problems – crystal “crowning” around the liquid meniscus and ineffective  
56  
57 241 suspension of the crystals. The former leads to inconsistency in solution composition for repeat  
58  
59 242 experiments while stirring not only suspends the crystal for detection, but also—through secondary  
60

1  
2  
3  
4 243 processes—impacts the detection time in a complex fashion. To avoid introducing further variability  
5  
6 244 it is therefore vital to compare data sets obtained at the same stirring rate. In the literature stirring  
7  
8 245 speeds of 700 and 900 rpm have been used with solution volumes in the range 1 – 1.8 mL. While  
9  
10 246 crowning may be overcome by using increased solution volumes to eliminate the meniscus, volumes  
11  
12 247 in excess of 1.6 mL are to be used with care since some of the solution then lies outside the  
13  
14 248 heating block of the Crystal16 (when standard HPLC vials are used). This obviously jeopardizes  
15  
16 249 temperature uniformity in the vial.

17 250 Overall there is no “correct” methodology – compromise is inevitable. For example, experiments  
18  
19 251 carried out in our laboratory on *p*-nitrobenzoic acid revealed a dependency of growth times and  
20  
21 252 nucleation rates on stirring in the range (900 rpm vs. 1400 rpm). We therefore chose 900 rpm in all  
22  
23 253 experiments presented here, also in order to link to previous work<sup>11,12</sup>. We prepared stock solutions  
24  
25 254 from 100 mL solvent stirred with the requisite mass of solute for one hour in a sealed, jacketed  
26  
27 255 vessel held at 40°C. 1.5 mL of that solution was delivered to each vial using a micropipette. Proper  
28  
29 256 sealing of the vials against evaporative losses is essential and here we found a combination of a  
30  
31 257 screw top and cling film to be highly effective in preventing weight loss. The total weight of all vials  
32  
33 258 including solution was recorded before placing them in the Crystal16 wells and then again at the  
34  
35 259 end of each experiment. Induction times from vials with a mass variation bigger than 0.05% were  
36  
37 260 rejected. Crowning was checked frequently in each cycle visually or by shaking the vials to observe  
38  
39 261 any solid material falling down into the solution.

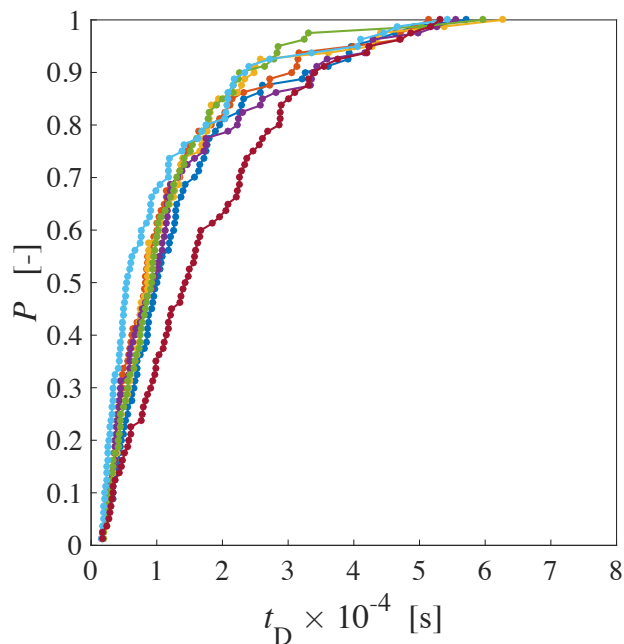
40 262 Upon placement in the instrument, the vials were heated back to the hold temperature (40°C)  
41  
42 263 with a rate of 1°C/min and held there for one hour. The vials were then crash cooled at 5°C/min to  
43  
44 264 20°C and held at this temperature for 8 h in order to allow time for nucleation to occur. Five cycles  
45  
46 265 were carried out, resulting in 80 data points per run. It is useful to perform a screening test to  
47  
48 266 choose the minimum and maximum supersaturations ( $S_{\min}$  and  $S_{\max}$ ) for a particular solute/solvent  
49  
50 267 system. We define  $S_{\min}$  as the value below which less than 50% of vials nucleate in an 8 h hold  
51  
52 268 time and  $S_{\max}$  is that above which any detection times lower than 120 s are measured<sup>33</sup>. Such  
53  
54 269 short detection times could result from cases where substantial nucleation already occurred during  
55  
56 270 the generation of supersaturation (i.e., during cooling). This would lead to complex effects of  
57  
58 271 temperature on the results and would invalidate the constant supersaturation assumption made in  
59  
60 272 Eq. (4). Note that we present a way to account for the added uncertainty in nucleation rate when

1  
2  
3  
4 273 working at supersaturations smaller than  $S_{\min}$  in the results section (cf. Section 5.2), so that  $S_{\min}$   
5  
6 274 should be understood as a soft boundary, while  $S_{\max}$  represents a hard boundary in the context  
7  
8 275 considered here. Within the workable supersaturation range, typically at least five supersaturations  
9  
10 276 including  $S_{\min}$  and  $S_{\max}$  should be chosen for data collection. In the benzoic acid in toluene data  
11  
12 277 presented here, we carried out experiments at nine different supersaturations.

## 15 278 5 Results

### 18 279 5.1 Nucleation rates from induction times obtained with unlimited observation 19 20 280 time

21  
22 281 In this section “synthetic data” is used to help address key questions in the estimation and reliability  
23  
24 282 of  $J$  values obtained from experiments by considering the impact of  $N_{\text{exp}}$ , the evaluation method  
25  
26 283 of  $t_g$  and the fitting technique used to transform the experimental probability data into nucleation  
27  
28 284 rates.



30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52 Figure 1: Synthetic data for seven experimental campaigns with  $N_{\text{exp}} = 80$ ,  $S = 1.8$ , and nucleation  
53 rate specified as in Table 1.

54  
55  
56 285 Figure 1 shows seven typical synthetic data sets taken from the 10,000 repeats of an experimental  
57  
58 286 campaign with  $N_{\text{exp}} = 80$  at  $S = 1.8$ . There is considerable variation in the probability curves

1  
2  
3  
4 287 obtained, even though they were generated in the same Poisson process (same nucleation rate,  
5  
6 288 volume, etc.). At each  $S$  at which simulations were performed, we therefore have 10,000 cumulative  
7  
8 289 probability curves,  $P$ , available. When fitting any of these probability curves using Eq. (4) or  
9  
10 290 Eq. (6), a different value for the nucleation rate  $J$  will be obtained. For each fitting method and  
11  
12 291 means of dealing with  $t_g$ , this results in 10,000 calculated  $J$  values. The corresponding probability  
13  
14 292 density distribution of these nucleation rate values,  $p(J)$  is defined such that

$$\int_0^{\infty} p(J) \, dJ = 1 \quad (20)$$

15  
16  
17  
18  
19  
20 293 The confidence intervals can be obtained using the cumulative distribution,  $P(J)$ , where the values  
21  
22 294 of  $J$  at 2.5% and 97.5%, were used—signified by  $J_{P=0.025}$  and  $J_{P=0.975}$ —to calculate 95% confidence  
23  
24 295 intervals. Since the integral in Eq. (20) represents a dimensionless probability and  $J$  has dimen-  
25  
26 296 sions of  $[\text{m}^{-3} \text{ s}^{-1}]$ , it follows that  $p(J)$  has dimensions of  $[\text{m}^3 \text{ s}]$ . Throughout this contribution, a  
27  
28 297 probability density distribution (of any property) will be indicated with small  $p$ , while cumulative  
29  
30 298 distributions are indicated with a capital  $P$ .

31  
32 299 Figure 2(a) shows, for different  $N_{\text{exp}}$ , and  $S = 1.8$ , the frequency density distribution of  $J$  values  
33  
34 300 estimated using the nonlinear fitting method (Eq. (5)) and taking  $t_g$  as the shortest induction time.  
35  
36 301 Figure 2(b) shows, for the same data, the effect of dealing with  $t_g$  by each of the above methods on  
37  
38 302 the 95% confidence interval of the  $J$  values. The results of the linear fitting (Eq. (7)) and frequency  
39  
40 303 density distributions of  $J$  with  $t_g$  again treated in each of the three ways are shown in Figure S1  
41  
42 304 in the Supporting Information and are similar to the results shown in Figure 2. Some specific  
43  
44 305 confidence interval values for  $J$  are given in Table 2.

45  
46 306 A number of features captured in Figure 2 are worth mentioning since they have important  
47  
48 307 implications for the experimental protocol. Firstly, for all six methods, the peak values of the  
49  
50 308  $J$  density distribution are around the “true”  $J$  values with the distribution narrowing as  $N_{\text{exp}}$   
51  
52 309 increases. This means that the most likely estimate is close to the real value and that more measured  
53  
54 310 data is beneficial in increasing the certainty of this estimate. Secondly, regarding the growth time  
55  
56 311  $t_g$ , it is evident that for low numbers of experimental points when  $t_g$  is taken as the shortest  
57  
58 312 induction time, the confidence intervals of  $J$  are slightly different from those when  $t_g$  is regarded  
59  
60 313 as a parameter or calculated from the crystal growth rate. For increasing numbers of points this

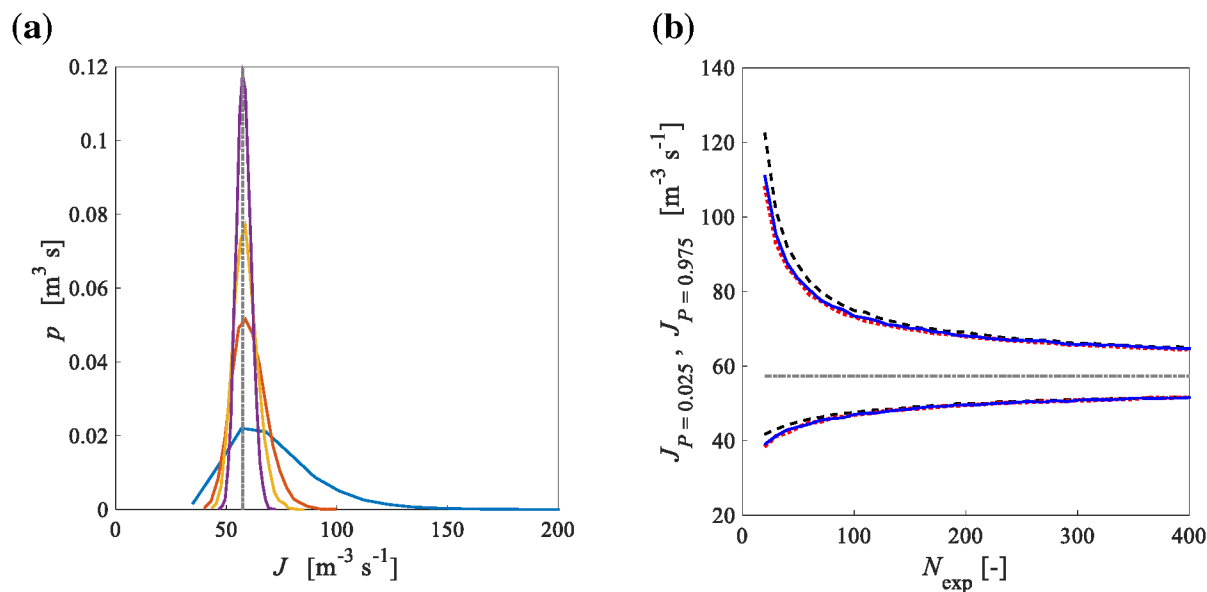


Figure 2: (a) Frequency density distribution of  $J$  values calculated using Eq. (4) and taking  $t_g$  as shortest induction time with  $S = 1.8$ ,  $N_{\text{exp}} = 20$  (blue), 80 (orange), 160 (yellow), 400 (purple); (b) the 95% confidence interval of  $J$  as a function of  $N_{\text{exp}}$  for  $S = 1.8$  and dealing with  $t_g$  using three different methods (black dash line:  $t_g$  as shortest induction time, red dot line:  $t_g$  as a parameter, blue solid line:  $t_g$  calculated from growth rate). The grey dash dot line is the real value of  $J$ .

314 difference disappears. It is noteworthy that—by definition— $t_g$  is a deterministic quantity and should  
 315 monotonically decrease with increasing supersaturation. However, when selecting  $t_g$  as the smallest  
 316 induction time contained within a limited dataset, some of the stochasticity of the nucleation time  
 317 is shifted onto  $t_g$ , such that the growth time also becomes a probabilistic quantity with its own  
 318 probability density distribution. These distributions might overlap when they stem from similar  
 319 supersaturations (see Fig S2 in the Supporting Information for  $S = 1.7$  and  $S = 1.75$ ). This is  
 320 important to keep in mind when considering real experimental data, where  $t_g$  values determined  
 321 using this approach do not necessarily decrease monotonically with increasing supersaturation.  
 322 While this is an undesired (and counter-intuitive) effect, Figure 2(b) indicates that it has little  
 323 impact on the estimated nucleation rates for datasets containing a reasonable number of detection  
 324 times. Overall it can be seen from both Figure 2 and Table 2 that the six methods show a consistent  
 325 estimation of the nucleation rate and the confidence intervals of  $J$  when 80 experiment points or more  
 326 are used. A method in which approximately 80 detection times are measured per supersaturation,  $t_g$   
 327 is taken to be the shortest induction time, and the nucleation rates are estimated by fitting Eq. (4)  
 328 to the experimental cumulative probability distribution, therefore constitutes a reasonable option

for an experimental protocol applied to real experimental data.

Table 2: Some specific 95% confidence intervals of the nucleation rate  $J$  calculated at  $S = 1.8$ . Parameters in Table 1 result in a  $J_{\text{real}}$  of  $57 \text{ m}^{-3}\text{s}^{-1}$ ; confidence intervals given in the same units.

$N_{\text{exp}}$	$t_g$ calculated from growth rate		$t_g$ as shortest induction time		$t_g$ as parameter	
	linear	nonlinear	linear	nonlinear	linear	nonlinear
80	47 – 75	46 – 76	47 – 76	47 – 78	45 – 79	46 – 76
160	50 – 69	49 – 70	50 – 69	49 – 70	48 – 72	49 – 70
400	52 – 64	52 – 65	52 – 65	52 – 65	51 – 66	52 – 65

Furthermore, we would like to highlight here that the relative error density distribution of  $J$  for 10,000 synthetic experiments does not change with  $S$ , as shown in Figure 3. This means that these comparisons and the conclusions derived from them are independent of supersaturation and that  $S$  is not a key factor influencing the outcome of the different methods, as long as  $S_{\text{min}}$  and  $S_{\text{max}}$  are chosen appropriately (see Section 4). In this context, it is worth comparing our values to the uncertainty assessment performed in the original work by Jiang and ter Horst<sup>12</sup>. They investigated the effect of the inherent stochasticity (at  $N_{\text{exp}}$ ), and concluded that “there is about an 80% chance that we measured a nucleation rate within 20% of the actual nucleation rate.”. Figure 3 supports that assessment. However, we present here a more extended analysis of the underlying probability distributions, extend it to other values of  $N_{\text{exp}}$ , and investigate the effect of further parameters (such as limited observation times) as detailed below.

## 5.2 Nucleation rates from induction time data with limited observation time

In experiments performed at low  $S$  the  $P_{\text{cutoff}}$  values typically reduce from 1 down to 0.1 (in extreme cases), because insufficient time is provided for every vial to nucleate. Qualitatively speaking, at a constant  $N_{\text{exp}}$  and decreasing  $P_{\text{cutoff}}$  the number of experiments where nucleation has been detected,  $N_{\text{nuc}}$ , decreases as well, therefore the nucleation rates estimated from such datasets should exhibit a higher degree of uncertainty and are therefore less reliable. Figure 4(a) shows (for  $S = 1.8$  and  $N_{\text{exp}} = 80, 200, 400$ ) how the relative confidence interval width (defined as the confidence interval width divided by the real  $J$  value) widens as  $P_{\text{cutoff}}$  decreases.

In Figure 4(b) the relative confidence interval width is plotted against the number of nucleated points ( $N_{\text{nuc}}$ ) for all values of  $S$  (1.6 – 2),  $N_{\text{exp}}$  (20 – 400) and different  $P_{\text{cutoff}}$  (0.1 – 1) conditions. Clearly, the relative confidence interval width decreases rapidly with increasing  $N_{\text{nuc}}$  (without ever

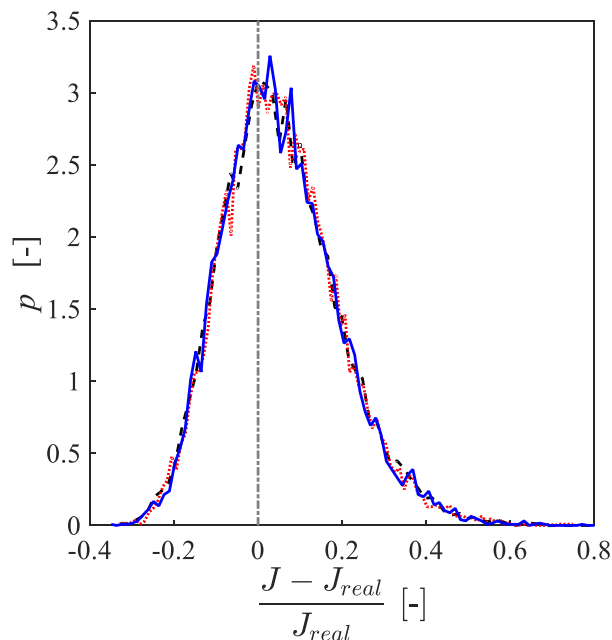


Figure 3: The relative error density distribution of  $J$  for 10,000 modelling experiments for  $S = 1.6$  (black dashed line), 1.8 (red dotted line) and 2.0 (blue solid line), with  $N_{\text{exp}} = 80$ . The vertical dotted line corresponds to the real nucleation rate (calculated with the values in Table 1).

decreasing to zero). In addition it can be seen that, apart from minor noise caused by random sampling, all the points can be collapsed onto a single line, indicating that it is the absolute number ( $N_{\text{nuc}}$ ) of nucleated points which determines the confidence intervals of  $J$ , regardless of the underlying values of  $S$ ,  $N_{\text{exp}}$  and  $P_{\text{cutoff}}$ . It is again noteworthy that this observation is independent of the value of supersaturation and hence of the absolute value of the nucleation rate. The confidence interval width is only one property of the underlying probability density distribution and different probability density distributions could result in the same confidence intervals. However, in Figure 4(c) we show that the probability density distributions for ( $N_{\text{exp}} = 80$ ,  $P_{\text{cutoff}} = 1$ ), ( $N_{\text{exp}} = 160$ ,  $P_{\text{cutoff}} = 0.5$ ) and ( $N_{\text{exp}} = 400$ ,  $P_{\text{cutoff}} = 0.2$ ), all having  $N_{\text{nuc}} = 80$ , overlap with each other (blue curves). The same is true for the two additional curves generated with ( $N_{\text{exp}} = 40$ ,  $P_{\text{cutoff}} = 1$ ) and ( $N_{\text{exp}} = 80$ ,  $P_{\text{cutoff}} = 0.5$ ), thus both having  $N_{\text{nuc}} = 40$ . Our analysis in Figure 4(c) shows that for a certain value of  $N_{\text{nuc}}$  the distribution of relative errors in  $J$  is indeed the same, regardless of the underlying  $P_{\text{cutoff}}$  and  $N_{\text{exp}}$  value. We also note that the probability density distributions are not normal distributions, but instead closer to lognormal distributions (i.e., heavier tail at larger  $J$  values). In fact, the probability density distribution becomes further away from a

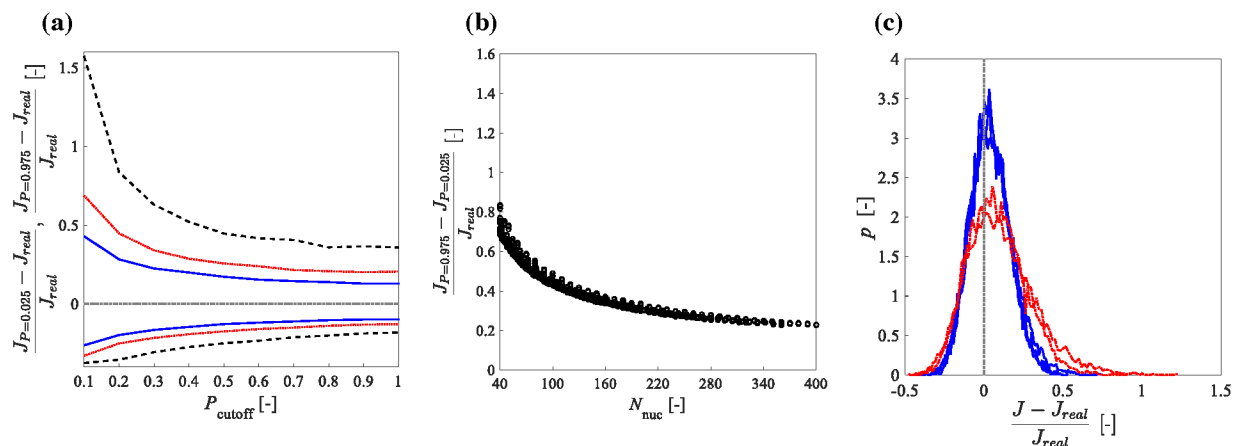


Figure 4: (a) Confidence intervals of  $J$  when reducing  $P_{\text{cutoff}}$  values from 1 down to 0.1 for  $S = 1.8$ ,  $N_{\text{exp}} = 80$  (black dash line), 200 (red dotted line) and 400 (blue solid line). (b) Relative confidence interval width (defined as 95% confidence interval width divided by the real  $J$  values) vs. the number of nucleated vials ( $N_{\text{nuc}}$ ) for all values of  $S$  (1.6 – 2),  $N_{\text{exp}}$  (20 – 400) and different  $P_{\text{cutoff}}$  (0.1 – 1) conditions. (c) The relative error density distribution of  $J$  for 10,000 modelling experiments for  $S = 1.8$ , blue lines:  $N_{\text{exp}} = 80$ ,  $P_{\text{cutoff}} = 1$ ,  $N_{\text{exp}} = 160$ ,  $P_{\text{cutoff}} = 0.5$ ,  $N_{\text{exp}} = 400$ ,  $P_{\text{cutoff}} = 0.2$ , red lines:  $N_{\text{exp}} = 40$ ,  $P_{\text{cutoff}} = 1$ ,  $N_{\text{exp}} = 80$ ,  $P_{\text{cutoff}} = 0.5$ .

normal distribution with decreasing  $N_{\text{nuc}}$ . We believe that this behaviour has not previously been discussed in the literature and symmetric confidence intervals reported around nucleation rates<sup>14</sup> should therefore be treated with caution – especially at low values of  $N_{\text{nuc}}$ . Crucially, these findings indicate that the reliability (or uncertainty) of nucleation rates estimated from a set of detection times can be assessed quantitatively regardless of the underlying nucleation rate without introducing further assumptions (apart from those already outlined in Section 2.1).

In real experiments only one  $P(t_D)$  curve is measured for each supersaturation and temperature value, rather than the 10,000 simulated here. Therefore, it is not possible to generate equivalent probability density distributions  $p(J)$  directly from experimental data. However, this analysis shows that the shape of the synthetic  $p(J)$  distribution, due to its independence from the absolute value of the nucleation rate, can be used to generate the  $p(J)$  distribution for experimental data as long as  $N_{\text{nuc}}$  is known. To accomplish this, one must realize that the nucleation rate estimated from one set of detection times corresponds to one position (not necessarily the peak – although it represents the most likely position) on the distribution shown in Figure 4(c). Since we have performed the simulations for a variety of  $N_{\text{nuc}}$  values, c.f. Figure 4(b), we know all these distributions. For each position of the experimental  $J$  value on this curve (which is associated with a known probability),

1  
2  
3  
4 383 a re-scaled probability distribution can be calculated. By weighting and summing these re-scaled  
5  
6 384 distributions, one obtains an approximation of the sought after probability density distribution for  
7  
8 385 the experimental  $J$  values. This procedure would result in Student's t-distribution if the underlying  
9  
10 386 probability distributions were normal<sup>34</sup>. However, as shown in Figure 4(c), they are not normal  
11  
12 387 distributions when  $N_{\text{nuc}}$  is small. The approach detailed here nonetheless gives consistent estimates  
13  
14 388 of probability density distributions and hence of confidence intervals. In summary: knowing the  
15  
16 389 value of  $N_{\text{nuc}}$  of each real experiment, the distributions of  $J$  values or the 95% confidence intervals  
17  
18 390 can be calculated using the distributions of relative error of  $J$  (examples shown in Figure 4(c))  
19  
20 391 obtained from the modelling results.

### 21 392 **5.3 Estimation of CNT parameters**

22  
23  
24 393 In this section, we investigate the influence of the chosen fitting method on the reliability of the  
25  
26 394 estimated CNT parameters. Figure 5 shows the synthetic data plotted in nonlinear and linearized  
27  
28 395 fashions according to Eq. (8) and Eq. (12). It is apparent that while the width of the confidence  
29  
30 396 interval around  $J$  increases with  $S$  in the non-linear plot (Figure 5(a)), the confidence intervals in the  
31  
32 397 linearized version (Figure 5(b)), show similar widths at each  $S$ . This indicates that the data exhibit  
33  
34 398 considerable heteroscedasticity (changes in the variance of the nucleation rate with the independent  
35  
36 399 variable (supersaturation)). The peak points of the distribution for both  $J$  and  $\ln(J/S)$  are around  
37  
38 400 the "true" values, as shown for one supersaturation in the insets in the respective figure. From the  
39  
40 401 insets, it is also apparent that the probability density distribution appears lognormal in Figure 5(a)  
41  
42 402 and (approximately) normal in Figure 5(b).

43 403 We can explore the effect of such heteroscedasticity on the estimated values for  $A$  and  $B$  by  
44  
45 404 randomly selecting values of  $J$  for each of the five  $S$  (at each supersaturation, there are 10,000  
46  
47 405 values to choose from) and then minimizing the sum of squared residuals between the set of selected  
48  
49 406  $J$  values and Eq. (8) or Eq. (12). In this way 10,000 sets of  $A$  and  $B$  were obtained for both  
50  
51 407 fitting methods. This provided distributions of the parameters, shown in Figure 6. The distribution  
52  
53 408 density of  $A$ ,  $p(A)$  (similarly defined as  $p(J)$ , see Eq. (20)), from nonlinear fitting is much wider  
54  
55 409 than from linear fitting, does not show a normal distribution and the peak is not at the "true"  $A$   
56  
57 410 value (Table 1).  $A$  values derived by fitting the linearized Eq. (12), on the other hand, are normally  
58  
59 411 distributed and show a narrow distribution centred on the "true" value. The distribution of  $B$   
60

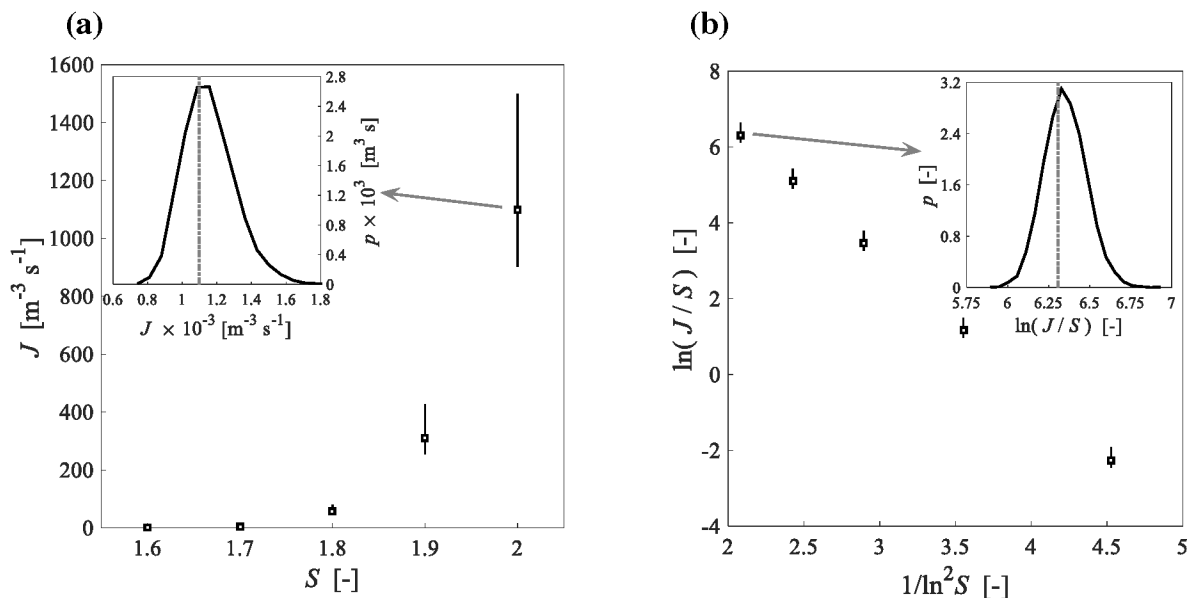


Figure 5: Synthetic nucleation rate data for different supersaturations plotted in nonlinear and linearized form: (a)  $J(S)$  vs.  $S$  and (b)  $\ln(J/S)$  vs.  $1/\ln^2(S)$ . The squares in both plots represent the peak values with 95% confidence intervals around them. The probability density distribution of  $J$  and  $\ln(J/S)$  for  $S = 2$ , respectively, are shown in the inserted figures, in which the dash dot lines show the corresponding “true” values.

values is likewise much wider in the nonlinear fitting than that from the linear fitting. Therefore, as expected, one would have to account for the presence of heteroscedasticity when using a least squares fitting approach. From Figure 6 we conclude that an estimation of  $A$  and  $B$  is best done by fitting Eq. (12) (or its equivalent for other mechanisms) to the data in a least squares sense (in contrast to using Eq. (8)). While this is common practice in the literature, we have substantiated the reason for this choice here.

We now focus on the case when  $J$  values at different supersaturations stem from cumulative probability curves containing a varying number of data points  $N_{\text{muc}}$  at each supersaturation (see, e.g., our dataset on benzoic acid shown in Table 5). We have shown that such variations translate into different levels of uncertainty (different widths of confidence intervals, cf. Figure 4(b)) in the estimated nucleation rates. In short such datasets exhibit heteroscedasticity so that applying an unweighted least squares method to estimate  $A$  and  $B$  is ill-advised. Admittedly, we<sup>11</sup> and others<sup>12,13</sup> have done so in the past, however, this is no reason not to make improvements. We therefore use the weighted least squares approach introduced in Eq. (18) on the synthetic datasets listed in

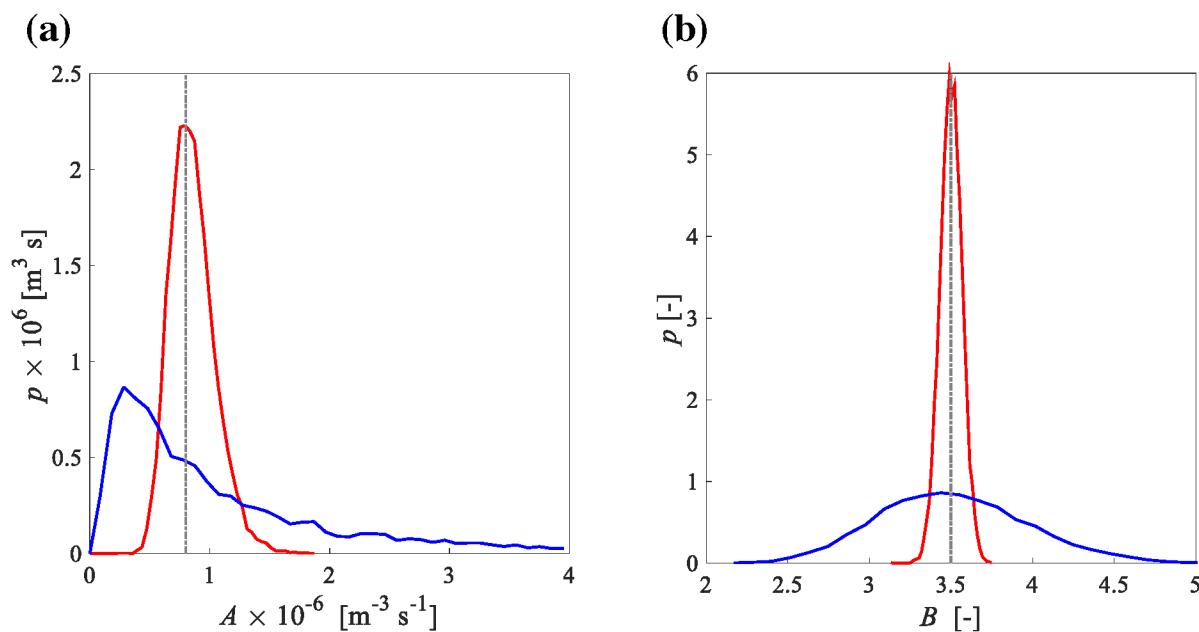


Figure 6: Probability density distributions of parameters (a)  $A$  and (b)  $B$  estimated from synthetic data either by fitting Eq. (8) (blue lines) or Eq. (12) (red lines). The dash dot lines show the corresponding real values.

Table 3. In this table, case 1 represents the base case with equal  $N_{\text{nuc}}$  measured for all supersaturations. Case 2 with low  $P_{\text{cutoff}}$  values at low supersaturations might stem from an experimental study where a maximum observation time has been set rather low, so that not all vials nucleated at lower supersaturations due to the inherently slower kinetics. Case 3 represents the reverse and case 4 closely mirrors the situation of our experimental dataset on benzoic acid (c.f. Table 5).

Figure 7 shows the probability density distributions for parameters  $A$  and  $B$  when the uncertainty on the nucleation rates at each supersaturation is propagated using the methodology detailed in Section 3.3. In both subfigures ( $A$  shown in Figure 7(a),  $B$  shown in Figure 7(b)) the dashed lines represent the weighted sum of squares approach, while the solid lines represent the regular sum of squares approach. The weighted least squares approach performs slightly better than the regular least squares approach for in cases 2, 3 and 4, as shown by the narrower probability density distributions that are better centered around the true parameter values (shown as the vertical dotted lines). For a more quantitative comparison, we also report the 95% confidence intervals that can be obtained from these probability density distributions in Table 4. While in all cases the true parameter values (c.f. Table 1) are contained within the reported confidence intervals, the intervals

Table 3: Cases analyzed to evaluate CNT parameter estimation methodologies

Case	$S$	$N_{\text{exp}}$	$P_{\text{cutoff}}$
1	1.60	80	1.0
	1.70	80	1.0
	1.80	80	1.0
	1.90	80	1.0
	2.00	80	1.0
2	1.60	80	0.2
	1.70	80	0.4
	1.80	80	0.6
	1.90	80	0.8
	2.00	80	1.0
3	1.60	80	1.0
	1.70	80	0.8
	1.80	80	0.6
	1.90	80	0.4
	2.00	80	0.2
4	1.60	80	0.7
	1.65	80	0.6
	1.70	80	0.7
	1.75	240	0.7
	1.80	80	0.9
	1.85	80	0.9
	1.90	220	0.9
	1.95	140	1.0
2.00	110	1.0	

obtained from the weighted least squares approach are tighter and thus preferable. It is also worth highlighting that the statistical analysis presented here confirms our intuition that adding more datapoints (both in terms of the number of measured supersaturations, as well as in terms of overall measured nucleation points) improves the parameter estimation considerably, as one can see with the comparison of case 1 vs. case 4.

This approach can also be used in the estimation of probability density distributions for  $A$  and  $B$  values estimated from experimental  $J$  values. This originates from the fact that an estimation of the uncertainty attached to these  $J$  values (and therefore the variance around them) can be performed as detailed in Section 5.2. It is therefore possible to use the weighted sum of squares approach for such data without any further modification.

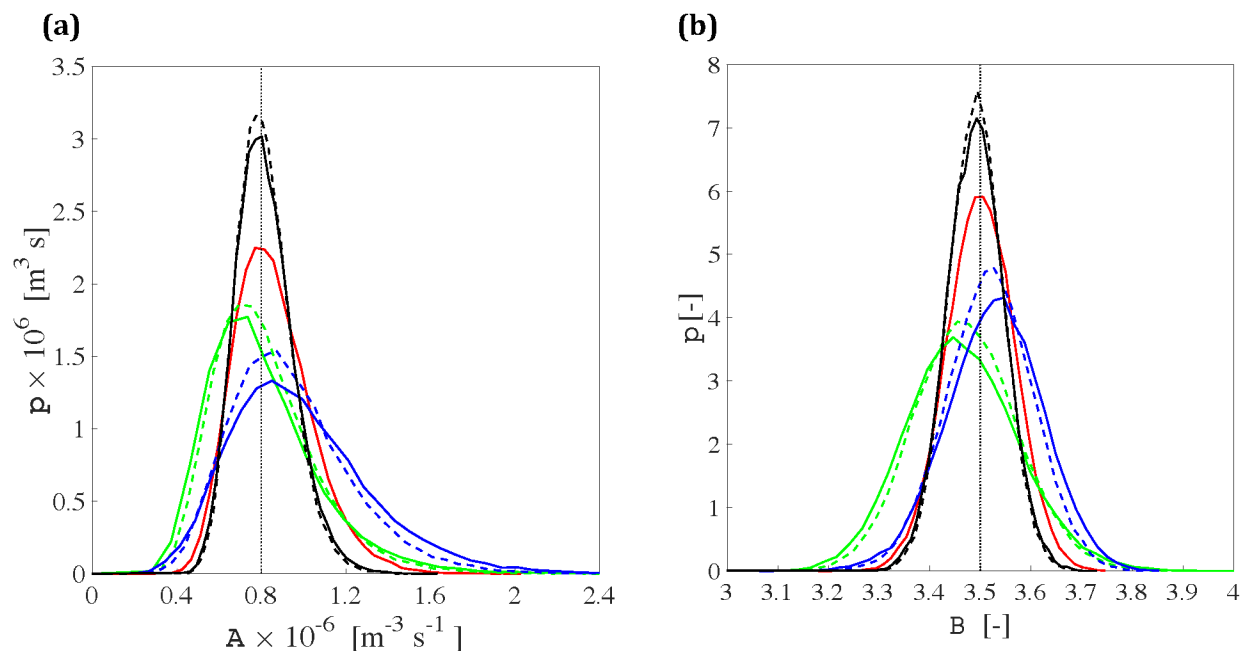


Figure 7: Comparison between regular sum of squares (Eq. (17); solid lines) and weighted sum of squares (Eq. (18); dashed lines) used to estimate CNT parameters: (a) results for parameters  $A$ , (b) results for parameters  $B$ . Red lines (overlapping): case 1; green lines: case 2; blue lines: case 3; black lines: case 4. Cases detailed in Table 3.

## 5.4 Application to experimental data

In this section detection time data gathered on BA/toluene solutions is analysed in detail using the methodology described in the sections above.

### 5.4.1 Induction time probability distributions and nucleation rates of BA in toluene

Figure 8 shows five representative induction time probability distribution curves of BA in toluene ( $S = 1.38, 1.41, 1.50, 1.57, 1.63$ ). It is apparent that the higher  $S$ , the faster the cumulative probability,  $P$ , increases, which means a higher  $J$ . The  $J$  values as well as  $t_g$  and  $N_{\text{nuc}}$  at different  $S$  obtained from fitting results to all supersaturations measured are listed in Table 5. The calculated  $J$  values do not always increase monotonically with  $S$ . This can occur despite efforts to control experimental variables, due to the inevitable uncertainty in estimating  $J$  using the probability distribution method, as highlighted in Figure 1. This leads to the fact that the probability distributions of  $J$  for different supersaturations might overlap. Hence, observing a non-monotonic behaviour in  $J$  is entirely possible and should not (necessarily) be interpreted as erroneous. Furthermore, for reasons

Table 4: 95% confidence intervals of CNT parameters for cases detailed in Table 3<sup>‡</sup>

Case	$A \times 10^{-6} \text{ m}^3 \text{ s}^{-1}$		$B$	
	unweighted	weighted	unweighted	weighted
1 <sup>†</sup>	0.55 – 1.28 (90.8%)		3.37 – 3.63 (7.50%)	
2	0.43 – 1.46 (129.2%)	0.45 – 1.39 (117.1%)	3.26 – 3.70 (12.5%)	3.28 – 3.68 (11.4%)
3	0.47 – 1.76 (161.3%)	0.49 – 1.61 (139.3%)	3.34 – 3.71 (10.6%)	3.35 – 3.68 (9.6%)
4	0.58 – 1.12 (66.9%)	0.59 – 1.10 (63.5%)	3.38 – 3.60 (6.3%)	3.39 – 3.60 (6.0%)

<sup>‡</sup> Percentage numbers are widths of confidence intervals divided by “real” parameter value (see Table 1).

<sup>†</sup> For case 1 weighted and unweighted approach yields the same confidence intervals.

discussed in Section 5.1 the expected monotonic decrease of  $t_g$  with supersaturation can also be absent, as shown in the reported data. However, as Section 5.1 also shows, this does not affect the estimated nucleation rates. The probability density distributions of  $J$  calculated for synthetic data can be used to estimate the probability density distribution of the nucleation rate values estimated from experimental data as detailed in Section 5.2. The resulting 95% confidence intervals are also reported in Table 5.

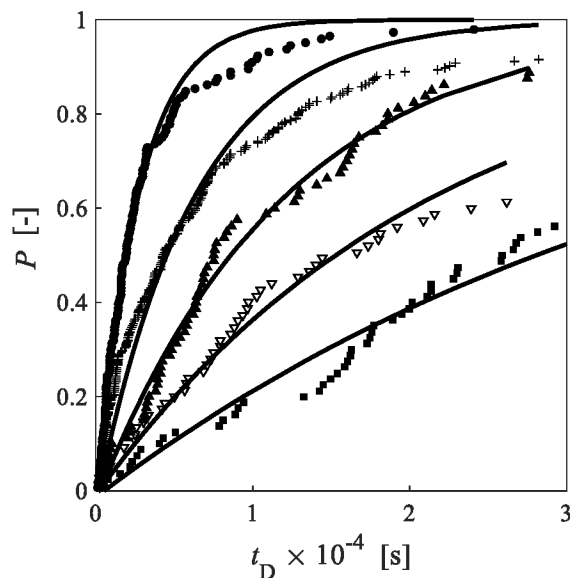


Figure 8: The measured cumulative induction time probabilities  $P$  (symbols) and fitted curves  $P(t_D)$  (Eqs. (4) and (5)) for BA in toluene at  $S = 1.38$  (squares), 1.41 (white triangles), 1.50 (black triangles), 1.57 (pluses), 1.63 (circles).

Table 5: Values of  $N_{\text{nuc}}$ ,  $J(S)$  and  $t_g$  for BA in toluene

$S$	$N_{\text{nuc}}$	$J^\dagger / \text{m}^{-3} \text{s}^{-1}$			$t_g / \text{s}$
1.38	57	17	11 – 25	523	
1.41	46	31	19 – 50	195	
1.44	53	31	20 – 50	237	
1.46	170	44	33 – 56	130	
1.50	71	56	37 – 87	274	
1.55	72	134	91 – 197	160	
1.57	206	107	84 – 133	211	
1.63	141	254	187 – 337	122	
1.69	109	587	441 – 788	122	

<sup>†</sup> Single values represent the most likely nucleation rate; ranges are 95% confidence intervals.

#### 5.4.2 Kinetic and Thermodynamic parameters of BA in toluene

With the different  $J$  values and the corresponding number of nucleated vials,  $N_{\text{nuc}}$ ,  $A$  and  $B$  and their confidence intervals can be estimated using the weighted least square method introduced in Section 5.3. The results for BA in toluene are reported in Table 6.

Table 6: Nucleation parameters of BA in toluene obtained using Eq. (12). Results from volume diffusion control are shown in Table S2 in the Supporting Information<sup>a</sup>

$A \times 10^{-2} / \text{m}^{-3} \text{s}^{-1}$		$B \times 10$		$f_0 C_0 / M^b / \text{mol}^{-1} \text{s}^{-1}$		$\gamma / \text{mJ m}^{-2}$	
13	8.7 – 18	5.2	4.6 – 5.8	10	6.5 – 15	4.4	4.2 – 4.6

<sup>a</sup> Single values represent the most likely parameter values; ranges are 95% confidence intervals.

<sup>b</sup>  $M$  represents the molarity of the solution.

For the sake of completeness, we performed linear fitting to all three forms of the CNT nucleation equations (Eqs. (12) to (14)) using the  $J(S)$  data of BA in toluene. The fitting results are shown in Figure 9 where it can be seen that all three linearized fits (Figure 9(a)) are similarly satisfactory. In Figure 9(b) the same fits and data are shown in the non-linearized form. It is clear from this plot that Eq. (14) gives a flatter trend with increasing  $S$  than Eq. (12) or Eq. (13) indicating the latter two to be better choices. However, all three equations significantly underestimate the nucleation rate at high supersaturations, primarily due to the fact that we provided more data at low nucleation rates (low supersaturations), hence the fits are understandably biased to that region. The fact that Eqs. (12) and (13) provide similar fits prevents us from using such data to discriminate between diffusion or interface transfer as rate limiting steps.

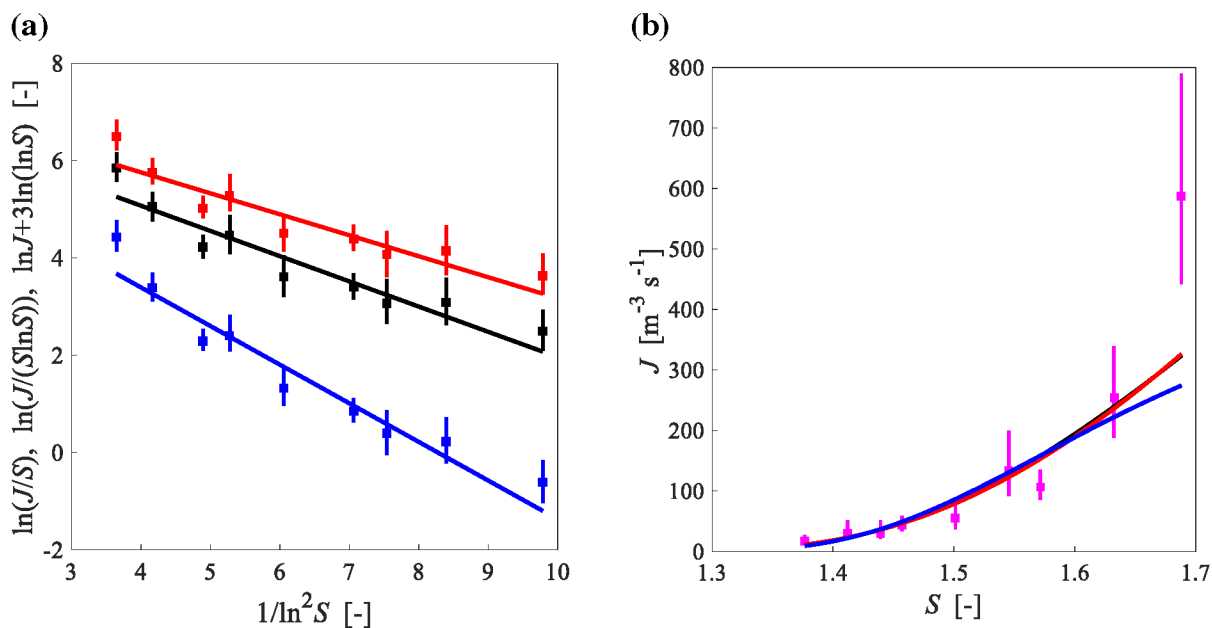


Figure 9: Linear fitting results of the  $J(S)$  data for BA in toluene to all three forms of the CNT nucleation equations (Eqs. (12) to (14)). (a) linear fits, (b) the curves are calculated from the parameters obtained from the linear fitting results in (a). Black: Eq. (12), red: Eq. (13), blue: Eq. (14), purple: real experiment points.

## 6 Discussion and conclusions

An approach to quantify the reliability of nucleation rates obtained from induction time experiments performed in small volumes of solution was presented. The approach utilized millions of induction times generated in silico to achieve this. Our approach assumed that a single nucleus mechanism is valid and assumed that nucleation is a stochastic event following a Poisson distribution. Under these circumstances, our data modelling showed that 80 or more induction times should be measured in the real experiments, that setting growth time,  $t_g$ , as the shortest induction time is acceptable and requires no additional information about the growth rate, and that using a nonlinear fitting method to the Poisson equation is the best option for estimating the nucleation rate. The kinetic and thermodynamic parameters,  $A$  and  $B$ , within CNT rate expressions are best obtained by linear data fitting<sup>35</sup>. Surprisingly, the relative error of  $J$  depends only on the number of nucleated vials,  $N_{\text{nuc}}$ , in the experiments and the uncertainty in  $J$  can be estimated using the modelling results.

A few points are worth making regarding the above conclusions and the wider applicability of the approach presented here. First of all, this stochastic methodology could be adapted to cases where

1  
2  
3  
4 498 the single nucleus mechanism does not apply, where the supersaturation does not stay constant—  
5  
6 499 either through cooling (i.e., during metastable zone width measurements) or because crystals are  
7  
8 500 only detected in a measurement device after a considerable amount of supersaturation has been  
9  
10 501 consumed by the nucleating/growing crystals. In the latter instance, a model similar to the one  
11  
12 502 presented by Maggioni and Mazzotti<sup>23</sup> could be employed together with the approach to generate  
13  
14 503 synthetic induction time data and evaluating the uncertainty attached to the derived nucleation  
15  
16 504 rates. However, in such a case, the resulting model loses the intriguing simplicity of the limiting case  
17  
18 505 analyzed here. Second, our approach can be used whether the stochastic process of nucleation follows  
19  
20 506 a Poisson distribution (which results from assuming that each nucleation event is independent, as  
21  
22 507 long as the experimental conditions ( $S$ ,  $T$ , etc.) remain constant) or one selected because it fits  
23  
24 508 experimental detection time data better (which is preferred by some authors<sup>18,21</sup>; although some  
25  
26 509 of the selected distributions lack a theoretical basis). Finally, it should be noted that the levels  
27  
28 510 of uncertainty calculated here must be seen as optimistic, because the inherent stochasticity of  
29  
30 511 the nucleation process was assumed to be the sole contributor to uncertainty. Naturally, there  
31  
32 512 are many other sources of uncertainty that could be important, especially fluctuating experimental  
33  
34 513 conditions (temperature fluctuations, evaporating solvent through slightly untight vials, deviations  
35  
36 514 in mixing/secondary nucleation caused by stirrer) and model-experiment mismatches. However, the  
37  
38 515 effect of experimental issues can be reduced by tighter control of temperature, ensuring that vials  
39  
40 516 are tight and so on, while the inherent stochasticity of nucleation cannot be reduced other than by  
41  
42 517 measuring a more extensive data set. Even then, we have shown that the uncertainty stemming from  
43  
44 518 the stochastic process results in confidence intervals of at least 20% relative width (cf. Figure 4(b)).

45  
46 519 Nevertheless, the present approach can be used to characterize the impact of all these effects  
47  
48 520 on the reliability/uncertainty of nucleation rates or any derived quantity, such as parameters in  
49  
50 521 nucleation rate expressions. Only by employing such a quantitative approach can some level of  
51  
52 522 statistical significance be established, e.g., when correlating or comparing parameters obtained from  
53  
54 523 different solvents or solutes. While this method has been used here solely in the context of benzoic  
55  
56 524 acid/toluene, we have successfully applied it to a larger dataset, as reported elsewhere<sup>36</sup> to enable  
57  
58 525 molecular insights into the nucleation of aromatic carboxylic acids from a variety of solvents.  
59  
60

## Acknowledgement

YX and SKT gratefully acknowledge funding from the Chinese Scholarship Council (CSC) and Pfizer UK Ltd., respectively.

## Supporting Information Available

Solubility data, further figures, notation table. Codes and data values are available on request from the corresponding author. This material is available free of charge via the Internet at <http://pubs.acs.org/>.

## References

- (1) Kashchiev, D. *Nucleation: Basic Theory with Applications*; Butterworth-Heinemann: Oxford, 2000.
- (2) Vekilov, P. G. Nucleation. *Cryst. Growth Des.* **2010**, *10*, 5007–5019.
- (3) Anwar, J.; Zahn, D. Uncovering molecular processes in crystal nucleation and growth by using molecular simulation. *Angew. Chem. Int. Ed.* **2011**, *50*, 1996–2013.
- (4) Yi, P.; Rutledge, G. Molecular origins of homogeneous crystal nucleation. *Annu. Rev. Chem. Biomol. Eng.* **2012**, *3*, 157–182.
- (5) Weissbuch, I.; Lava, M.; Leiserowitz, L. Toward Stereochemical Control, Monitoring, and Understanding of Crystal Nucleation. *Cryst. Growth Des.* **2003**, *3*, 125–150.
- (6) Chen, J.; Trout, B. Computational Study of Solvent Effects on the Molecular Self-Assembly of Tetrolic Acid in Solution and Implications for the Polymorph Formed from Crystallization. *J. Phys. Chem. B* **2008**, *112*, 7794–7802.
- (7) Toroz, D.; Hammond, R.; Roberts, K.; Harris, S.; Ridley, T. 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.
- (8) Salvalaglio, M.; Perego, C.; Giberti, F.; Mazzotti, M.; Parrinello, M. Molecular-dynamics simulations of urea nucleation from aqueous solution. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, E6–E14.
- (9) Salvalaglio, M.; Mazzotti, M.; Parrinello, M. Urea homogeneous nucleation mechanism is solvent dependent. *Faraday Discuss.* **2015**, *179*, 291–307.
- (10) Davey, R.; Schroeder, S.; ter Horst, J. Nucleation of Organic Crystals—Molecular Perspective. *Angew. Chem. Int. Ed.* **2013**, *52*, 2166–2179.

- 1  
2  
3  
4 556 (11) Sullivan, R.; Davey, R.; Sadiq, G.; Dent, G.; Back, K.; ter Horst, J.; Toroz, D.; Hammond, R.  
5 557 Revealing the Roles of Desolvation and Molecular Self-Assembly in Crystal Nucleation from  
6 558 Solution: Benzoic and *p*-Aminobenzoic Acids. *Cryst. Growth Des.* **2014**, *14*, 2689–2696.
- 7  
8 559 (12) Jiang, S.; ter Horst, J. Crystal Nucleation Rates from Probability Distributions of Induction  
9 560 Times. *Cryst. Growth Des.* **2011**, *11*, 256–261.
- 10  
11 561 (13) Kulkarni, S.; Kadam, S.; Meekes, H.; Stankiewicz, A.; ter Horst, J. Crystal Nucleation Kinetics  
12 562 from Induction Times and Metastable Zone Widths. *Cryst. Growth Des.* **2013**, *12*, 2435–2440.
- 13  
14 563 (14) Brandel, C.; ter Horst, J. Measuring induction times and crystal nucleation rates. *Faraday*  
15 564 *Discuss.* **2015**, *179*, 199–214.
- 16  
17 565 (15) Rossi, D.; Gavriilidis, A.; Kuhn, S.; Candel, M.; Jones, A.; Price, C.; Mazzei, L. Adipic Acid  
18 566 Primary Nucleation Kinetics from Probability Distributions in Droplet-Based Systems under  
19 567 Stagnant and Flow Conditions. *Cryst. Growth Des.* **2015**, *15*, 1784–1791.
- 20  
21 568 (16) Khamar, D.; Zeglinski, J.; Mealey, D.; Rasmuson, A. Investigating the Role of Solvent?Solute  
22 569 Interaction in Crystal Nucleation of Salicylic Acid from Organic Solvents. *J. Am. Chem. Soc.*  
23 570 **2014**, *136*, 11664–11673.
- 24  
25  
26 571 (17) Davey, R.; Back, K.; Sullivan, R. Crystal nucleation from solutions – transition states, rate  
27 572 determining steps and complexity. *Faraday Discuss.* **2015**, *179*, 9–26.
- 28  
29 573 (18) Mealey, D.; Zeglinski, J.; Khamar, D.; Rasmuson, A. Influence of solvent on crystal nucleation  
30 574 of risperidone. *Faraday Discuss.* **2015**, *179*, 309–328.
- 31  
32 575 (19) Ildfonso, M.; Candoni, N.; Veesler, S. A Cheap, Easy Microfluidic Crystallization Device  
33 576 Ensuring Universal Solvent Compatibility. *Org. Process Res. Dev.* **2012**, *16*, 556–560.
- 34  
35 577 (20) Mealey, D.; Croker, D.; Rasmuson, A. Crystal nucleation of salicylic acid in organic solvents.  
36 578 *CrystEngComm* **2015**, *17*, 3961–3973.
- 37  
38 579 (21) Javid, N.; Kendall, T.; Burns, I.; Sefcik, J. Filtration Suppresses Laser-Induced Nucleation of  
39 580 Glycine in Aqueous Solutions. *Cryst. Growth Des.* **2016**, *16*, 4196–4202.
- 40  
41 581 (22) Yang, H.; Rasmuson, A. Nucleation of Butyl Paraben in Different Solvents. *Cryst. Growth*  
42 582 *Des.* **2013**, *13*, 4226–4238.
- 43  
44 583 (23) Maggioni, G.; Mazzotti, M. Modelling the stochastic behaviour of primary nucleation. *Faraday*  
45 584 *Discuss.* **2015**, *179*, 359–382.
- 46  
47 585 (24) Dugua, J.; Simon, B. Crystallization of sodium perborate from aqueous solutions: I. Nucleation  
48 586 rates in pure solution and in presence of a surfactant. *J. Cryst. Growth* **1978**, *44*, 265–279.
- 49  
50 587 (25) Kondepudi, D.; Kaufman, R.; Singh, N. Chiral symmetry breaking in sodium chlorate crystal-  
51 588 lization. *Science* **1990**, *250*, 975–976.
- 52  
53  
54 589 (26) Kulkarni, S.; Meekes, H.; ter Horst, J. Polymorphism Control through a Single Nucleation  
55 590 Event. *Cryst. Growth Des.* **2014**, *14*, 1493–1499.
- 56  
57 591 (27) Harding, M.; Rule, R.; Oldman, R.; Davey, R. Growth rate dispersion in small crystals and its  
58 592 relation to mosaic spread. *J. Cryst. Growth* **1992**, *123*, 373–384.

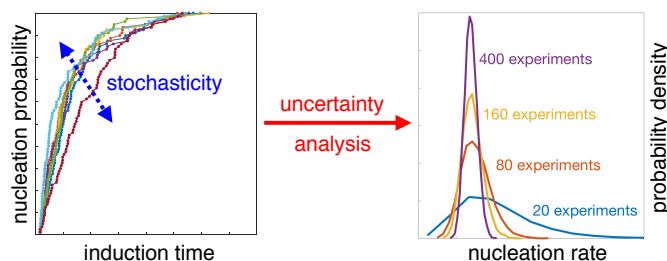
- 1  
2  
3  
4 593 (28) Ochsenbein, D.; Schorsch, S.; Vetter, T.; Morari, M.; Mazzotti, M. Modeling the facet growth  
5 594 rate dispersion of  $\beta$ -L-glutamic acid—Combining single crystal experiments with  $n$ D particle  
6 595 size distribution data. *Chem. Eng. Sci.* **2015**, *133*, 30–43.
- 7  
8 596 (29) Dunning, W.; Notley, N. Kinetics of Crystallization. III. *Z. Phys. Chem.* **1957**, *61*, 55–59.
- 9  
10 597 (30) Dunning, W.; Shipman, A. *Proc. Agric. Industries 10th International Conference* **1957**, 1448–  
11 598 1456.
- 12  
13 599 (31) Matsumoto, M.; Nishimura, T. Mersenne twister: a 623-dimensionally equidistributed uniform  
14 600 pseudo-random number generator. *ACM Trans Model Comput Simul* **1998**, *8*, 3–30.
- 15  
16 601 (32) Strutz, T. *Data Fitting and Uncertainty – A practical introduction to weighted least squares*  
17 602 *and beyond*; Springer: Heidelberg, 2016.
- 18  
19  
20 603 (33) The 120 s represent a, admittedly arbitrary, safety margin here.
- 21  
22 604 (34) Student, The Probable Error of a Mean. *Biometrika* **1908**, *6*, 1–25.
- 23  
24 605 (35) This would also hold true for any other rate expression with an exponential term in it, such  
25 606 as the ones associated with two step nucleation theory<sup>2</sup>.
- 26  
27 607 (36) Cruz-Cabeza, A.; Davey, R.; Sachitanathan, S.; Smith, R.; Tang, S.; Vetter, T.; Xiao, Y.  
28 608 Revealing the role of aromatic stacking in the nucleation of substituted benzoic acids. *in*  
29 609 *preparation*
- 30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

## 610 For Table of Contents Use Only

611 **Title:** Quantifying the inherent uncertainty associated with nucleation rates estimated from in-  
612 duction time data measured in small volumes

613 **Authors:** Yan Xiao, Sin Kim Tang, Hongxun Hao, Roger J. Davey, and Thomas Vetter



614 **Synopsis:** The reliability of nucleation rates obtained from induction time experiments performed  
615 in small volumes of solution is quantified. Utilizing millions of induction times generated in silico, the  
616 propagation of the inherent stochasticity of nucleation to estimates of parameters in rate expressions  
617 of classical nucleation theory is investigated. The approach is applied to experimental data of benzoic  
618 acid nucleated from toluene..