Data mining crystallization kinetics

*Diego A. Maldonado, Antony Vassileiou, Blair Johnston, Alastair J. Florence, Cameron J. Brown**

EPSRC Future Manufacturing Research Hub for Continuous Manufacturing and Advanced Crystallisation (CMAC), University of Strathclyde, Technology and Innovation Centre, 99 George Street, Glasgow G1 1RD, United Kingdom

Crystallization kinetics, population balance, data mining, machine learning, random forest

ABSTRACT

Population balance model is a valuable modelling tool which facilitates the optimization and understanding of crystallization processes. However, in order to use this tool, it is necessary to have previous knowledge of the crystallization kinetics, specifically crystal growth and nucleation. The majority of approaches to achieve proper estimations of kinetic parameters required experimental data. Across time, a vast literature about the estimation of kinetic parameters and population balances have been published. Considering the availability of data, this work built a database with information on solute, solvent, kinetic expression, parameters, crystallization method and seeding. Correlations were assessed and clusters structures identified by hierarchical clustering analysis. The final database contains 336 data of kinetic parameters from 185 different sources. The data were analysed using kinetic parameters of the most common expressions.

Subsequently, clusters were identified for each kinetic model. With these clusters, classification random forest models were made using solute descriptors, seeding, solvent, and crystallization methods as classifiers. Random forest models had an overall classification accuracy higher than 70% whereby they were useful to provide rough estimates of kinetic parameters, although these methods have some limitations.

1. INTRODUCTION

Year by year the challenges that the pharmaceutical sector has to face does not cease to increase. Regulatory requirements, patients' needs, and market competition are becoming more challenging, which has led the industry to rethink the model of business and seek alternatives to improve its productivity. Historically, the business model has been based on the discovery of new molecules and patents protection to a certain extent. However, the costs of new drugs development increase across time and patent expiry time remains the same.[1, 2] In addition, the pharmaceutical industry has been characterized by problems of innovation, flexibility, and efficacy in their processes, which increase costs and hinder to response to customer's demands as required.[2] As a result, the industry is seeking to optimize resources and improve its procedures to satisfy its needs and produce better medicines.

In this way, various initiatives have been introduced in the industry in the last decades. Some include the use of process analytical technology (PAT), the concept of quality by design (QbD), and the development of continuous pharmaceutical manufacturing (CPM), which has come along with technological and scientific advances.[3, 4] Consequently, many methodologies that optimize resources and create more efficient processes have been adopted. In particular, modelling techniques are of great interest given their ability to predict and provide information in an efficient manner.[1, 5]

Modelling techniques aim to depict a material property or a process through a mathematical expression which can be founded on either a physical or empirical relationship.[1, 5, 6] These representations enable the simulation of a process and assess different scenarios in which a condition or property changes.[1, 3] Likewise, modelling techniques facilitate the evaluation and analysis of the effect of factors on processes performance or product quality.[4] In light of these potential usages, the advantages that these models offer are numerous; an adequate model may enable to reduce the number of experiments necessary to obtain certain information,[3] or it may help with quality improvement as modelling provides a valuable insight into the design of a process, which would allow to select conditions or establish specifications systematically with a scientific base.[4] As a result, these tools are being used more frequently in recent years.

For crystallization, a critical unit operation in control and delivery of API with desired specifications, the most common form of modelling is through a population balance model (PBM), typically combined with momentum, mass and energy balances.[7] The main attraction of a PBM is the ability to predict the crystal size distribution (CSD). To fully resolve a PBM, expressions representing the various crystallization phenomenon, such as growth, primary nucleation, secondary nucleation, breakage, agglomeration, are required. For each phenomenon a range of expressions are available, ranging from fundamental fully mechanistic to semi-empirical.[7] Therefore, the selection of the most appropriate kinetic expression and the determination of the respective parameters are crucial in order to obtain accurate predictions. Currently, these activities require the collection of data through an experimental approach, with subsequent application of optimization algorithms that enables proper estimation. Nonetheless, there exists a vast amount of literature tackling PBM and the calculation of kinetic parameters, considering numerous factors such as solute, solvent, operational conditions, etc.

Theoretically, crystallization sub-processes are strongly affected by interactions between solute-solvent and process conditions. In this regard, it could be observed that some kinetic parameters include terms that describe directly any property related to solute and solvent, e.g. surface tension and molar volume. In the same way, it might be expected that kinetic parameters employed in nucleation and growth models, which do not have an explicit relation with physical or chemical properties of the components involved, follow a distribution or correlates to some variables associated with solute, solvent or process. The finding of these relations could potentially be helpful to provide a reasonable range of values within kinetic parameters could be or an approximate estimation of these which may be used in PBM.

This work aims to: 1) build a database containing information on kinetic parameters of primary nucleation and crystal growth of different crystallization processes that includes: solute, solvent, crystallization technique, seeding, and kinetic expression. 2) establish the feasibility of a model that enables estimation of kinetic parameters of growth and primary nucleation by analysis for patterns and correlations with some molecular and process descriptors.

2. METHODS

2.1. Data collection

Initially, a sample frame of potential articles containing the information of interest was built by web-scraping search results from different scientific databases. This procedure was conducted similar to that described by Kwartler.[8] To obtain these results several search strategies were implemented in the following databases: ScienceDirect, ACS Publications, AIChE, and Scientific Research. The combinations of keywords, inclusion and exclusion criteria are detailed in Table 1. To remark, Boolean operators were employed only in ScienceDirect and AIChE websites since those allowed their usage and therefore more complex strategies could be used. All the searches

were performed between June 4 and 6, 2019. The searches were limited to research articles in English – avoiding, for instance, reviews or book chapters - as the main objective was to obtain experimental data.

Table 1. Search strategies and databases.

Database	Search keywords
ScienceDirect: https://www.sciencedirect.com/	(((growth nucleation) OR (kinetic) OR MSMPR) AND ("population balance" crystal) AND (estimation OR determination))) NOT (granulation OR precipitation)
	(growth OR nucleation OR kinetic*) AND ("population balance") AND (pharmaceutical OR drug OR API) AND crystal*
	("population balance" AND crystal*) AND (pharma* OR drug)
ACS Publications: https://pubs.acs.org/	"population balance" crystallization kinetics
AIChE: https://aiche.onlinelibrary.wiley.com/	((growth nucleation) OR (kinetic) OR MSMPR) AND ("population balance" crystal) NOT (granulation)" anywhere published in "AIChE Journal
	(growth OR nucleation OR kinetic AND "population balance") AND (pharmaceutical OR drug OR API)" anywhere and "(crystal*)
Scientific Research: https://www.scirp.org/	population balance crystal kinetic

Subsequently, the following information on all the search results was extracted from the respective websites: title, journal, and authors. The data was next stored in a spreadsheet and pre-processed. Pre-processing consisted of text cleaning, duplicates removal and filtering. Text

cleaning involved stripping extra white spaces and fixing corrupted characters to then remove duplicates, which resulted in a list of 1938 articles. All these tasks were carried out using the R statistical program version 3.5.1 and Microsoft Excel (2016).

This list was later filtered by journal and title. Firstly, it was noticed all the results were published in a total of 125 journals where around 85 % of these papers corresponded to solely 15 journals. Therefore, journals with a number of results lower than 16 were discarded since the remaining 15 % did not reach this number of papers. To verify that important data was not omitted, articles in the discarded journals went through a non-exhaustive review and the most search results turned out to contain non-relevant information. Thus, with the remaining articles, a word frequency analysis of the titles was carried out. Further information on text mining and the frequency analysis can be consulted in Kwartler.[8] Words with a frequency higher than 3 and identified as non-relevant can be seen in the ESI. The article titles containing these words were excluded to finally obtain a list of 1187 articles.

The remaining 1187 articles were then reviewed manually in their totality and data were collected. During the review, various documents were found to have incomplete information or have taken data from another source; therefore, more results were discarded. Likewise, articles that initially were not included in the list were added by considering the source stated on the reviewed papers. The criteria used to select the articles in this stage is illustrated in Figure 1. Information regarding the extracted variable description, name, data type and comments were recorded and can be seen in the ESI.

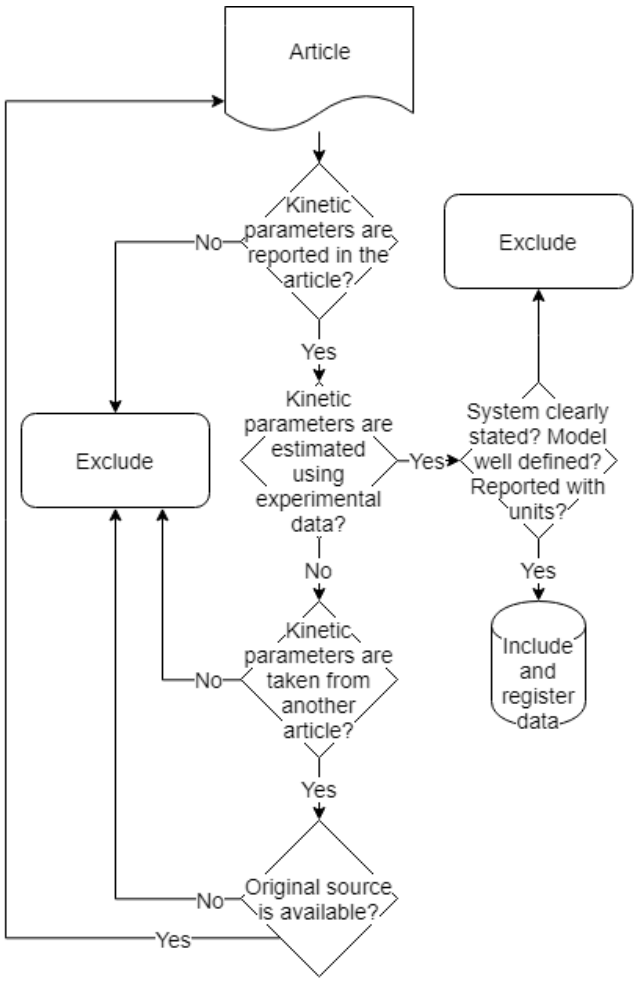

Figure 1. Exclusion/inclusion criteria for the final list of articles.

2.2. Data analysis

Before conducting the analysis, the collected data went through several cleaning steps. Firstly, the units of k_g and k_b were converted into international system units (SI). Most units of k_g were in either ms^{-1} or $\mathrm{ms}^{-1}(\mathrm{g/g\ solvent})^{-1}$, while k_b was mostly in $\#\ \mathrm{m}^{-3}\mathrm{s}^{-1}$. However, the units of k_b and k_g depended on the factors considered in the kinetic models. Therefore, it was not possible to transform all the units into the same and ensure all the data were comparable in this aspect. Additionally, there were also a few articles in which an equation was given to calculate the

constants and other articles estimating bidimensional growth rate. These cases were recorded but not considered during the analysis. Another adjustment was the scale where logarithm transformation was applied to k_g and k_b, given the order of magnitude that these constants presented. On the other hand, kinetic equations nomenclature was harmonized since several models could be considered equivalent but were expressed in different terms according to the author. Finally, analysis and visualizations were carried out using the R statistical environment software version 3.5.1.

2.2.1. Journal bias by crystallization method

Two analyses were carried out in order to establish the dependency of the reported crystallization method in the journal. A first approach was to employ a Chi-square test of independence having as inputs the entries per journal.[9] In this analysis, it was only considered journals whose number of entries were greater than 10. The second approach was utilizing an analogous analysis but considering the number of articles with a particular method instead of the entries. The reason behind this alternative approach was that an article may have multiple data points but the common pattern was a specific article focuses just on one crystallization method. Therefore, by performing the analysis in this manner, it is possible to avoid bias by excluding journals which may have various data points but very few articles. In the latter approach, the journals with more than 8 papers were used in the evaluation.

The journals used for the analysis were selected based on the number of journals which represent more than 90% of either the entries or artciles, according to the case. Tables in the ESI summarize the number of entries and papers for each journal found in the database.

2.2.2. Molecular descriptors

433 molecular descriptors were initially calculated for all the solutes identified in this revision using Molecular Operating Environment (MOE) software. Afterwards, various descriptors were discarded by considering the following criteria: the same values for all the solutes ($\text{Variance} = 0$), more than one non-determined value (NA), and a high correlation between descriptors (Pearson correlation absolute values greater than 0.9); this resulted in a final list of 110 descriptors. The association of these descriptors with k_b, k_g, b, and g was eventually evaluated.

2.2.3. Hierarchical clustering analysis

Hierarchical clustering (HC) is a methodology of unsupervised classification in which groups or clusters are made based on the similarity (agglomerative) or dissimilarity (divisive) of data.[10] In agglomerative hierarchical clustering (AHC), similar observations form clusters which in turn merge forming larger groups, until a group is obtained containing all the data.[10] In this work, AHC was applied to identify patterns or homogeneous groups in kinetic models that had more than 50 observations. The similarity was measured as Euclidean distances between pairs of (k_g, g) or (k_b, b), according to the case, as can be seen in the equation (1) below.

$$d_{ij} = \sqrt{(x_{1i} - x_{1j})^2 + (x_{2i} - x_{2j})^2} \quad (1)$$

Where d_{ij} represents the distance between the observation i and j, (x_1, x_2) denote the standardized values of either (k_b, b) or (k_g, g). The standardization consisted of subtracting the mean and dividing by the standard deviation. More details regarding the implementation and theoretical aspects can be found elsewhere.[10, 11] As to the selection of the appropriate number of groups, silhouette index was used as a criterion.[11]

2.2.4. Random forest

Random forest (RF) is a technique employed in supervised classification and regression problems.[12] This algorithm generates numerous decision trees using randomly chosen subsets of variables or classifiers.[12] When it is used in classification, each of these trees assigns the problem sample to a determined cluster, by which the same sample may be classified in several groups.[12] As a result, the definite classification is decided by majority votes of decision trees.[12] For the purpose of this study, the main objective of building a RF model was to identify relevant variables that have a certain association with the kinetic parameters.

Thus, a model of classification was first built and the model parameters were tuned. The groups were created by clustering analysis and the classifiers, or potential predictors, corresponded to the molecular descriptors, solvent, method, and seeding. Subsequently, the importance of the predictors was estimated as the mean decrease in accuracy (MDA). RF implementation was performed as detailed elsewhere.[13] The top 15 of most important variables were analysed in detail. To conclude, the selected classifiers were analysed in detail with respect to kinetic constants to assess how they are related.

3. RESULTS AND DISCUSSION

3.1. Data description

The database contains 336 data of kinetic parameters obtained from 185 articles, of which, 21 were not included in the initial sample frame, which means around 1 in 10 revised articles had relevant information. Most of the excluded papers contained incomplete data - for example, the solute identity was stated generically or not provided - or consisted of reviews wherein the primary focus was on mathematical or theoretical aspects of crystallization kinetics. In this manner, if this

approach was to be used in future works, the search strategies ought to be refined to reduce the content which was unrelated and increase search efficacy by including additional keywords, limiting the search to certain journals or considering other filters.

In the recorded data, 297 corresponded to growth rate and 145 related to primary nucleation rate. The data are distributed over 87 solutes and 27 solvents. In particular, solutes are mostly of low molecular weight (< 500 Da) and diverse chemical structure, being 25 inorganics and 62 organic molecules. Another important aspect to highlight is the large predominance of data related to crystallization in aqueous systems. As stated previously, there was a total of 27 solvents where 12 corresponded to aqueous – organic mixtures that, along with water, represented 72.6% of the collected data. Moreover, when antisolvent technique was applied, water was frequently used as an antisolvent (74.5%), which reinforced aqueous systems preponderance. As a consequence, the analysis of this study concerning the effect of solvent on kinetic parameters may be limited due to scarce information on other solvents apart from water. A breakdown of the information related to solute, solvent, method, seeding, and kinetic expressions can be seen in Table 2.

Table 2. Breakdown of information in database. N = 336.

Solute	
Paracetamol	8.93%
Glutamic acid	6.85%
Felodipine	3.87%
Solvent	
Water	65.2%
Ethanol	9.8%
Methanol	8.3%
Method	

Cooling	62.2%
Precipitation	18.2%
Antisolvent	15.2%
Evaporative	1.2%
Combinations	3.2%
Seeding	
Seeded	50.0%
Unseeded	47.6%
Combination of seeded and unseeded	2.4%
Growth rate expression	
$\boldsymbol{G = k_g \Delta C^g}$	31.6%
$\boldsymbol{G = k_g (S-1)^g}$	25.3%
$\boldsymbol{G = k_g (S-1)^g e^{(-E_g/RT)}}$	12.1%
Nucleation rate expression	
$\boldsymbol{B = k_b \Delta C^b}$	42.8%
$\boldsymbol{B = k_b e^{(-B/ln^2 S)}}$	19.3%
$\boldsymbol{B = k_b (S-1)^b e^{(-E_b/RT)}}$	5.5%

Regarding kinetic equations, the expressions used to model growth rate were more diverse than primary nucleation rate. In total, 38 different expressions for growth and 22 different expressions for nucleation rate were found. However, the majority of the crystal growth expressions were derived from the first two shown in Table 2. In these cases, the models included multiplicative terms related to stirring rate, crystal size, or temperature adjustment by Arrhenius; the latter being the most frequent. More complex equations like birth & spread model were also found, but they were isolated cases. For nucleation rate, while there were various ways of modelling, a clear

tendency to use empirical nucleation rate and, to a lesser extent, equations derived from CNT was observed. As can be seen, the power-law models are predominant in both crystal growth and primary nucleation modelling. During the revision, a specific reason to use one or another expression was not found. However, the power-law expressions have long been used in crystallization kinetics modelling since experimental data generally fit well to these equations.[14]

Figure 2 illustrates the sampling distribution of different kinetic parameters. It can be observed the most frequent values were in the order of 10^{8} and 10^{-6}, in international units, for nucleation and growth rate constants, respectively. Likewise, the most common estimations of b and g corresponded approximately to 2.0 and 1.0. All the distributions were right-skewed to a certain extent. However, this behaviour was more notable for the exponents. In this particular case, it was more frequent to find low values of b and g. This fact was emphasized by seeing that 50% of the data were contained within the intervals between 1.0 and 2.0 for g, and between 1.5 and 5.9 for b, which may be considered relatively narrow compared to all the possible values. Returning to kinetic constants, $\log k_b$ values lower than 0 or higher than 30 were not common since they only represented around 13% of the data, while the majority of $\log k_g$ values were lower than 0 with about 75%. Nonetheless, although similar distributions for kinetic parameters values can be seen when separating by kinetic models, some differences between models were observed.

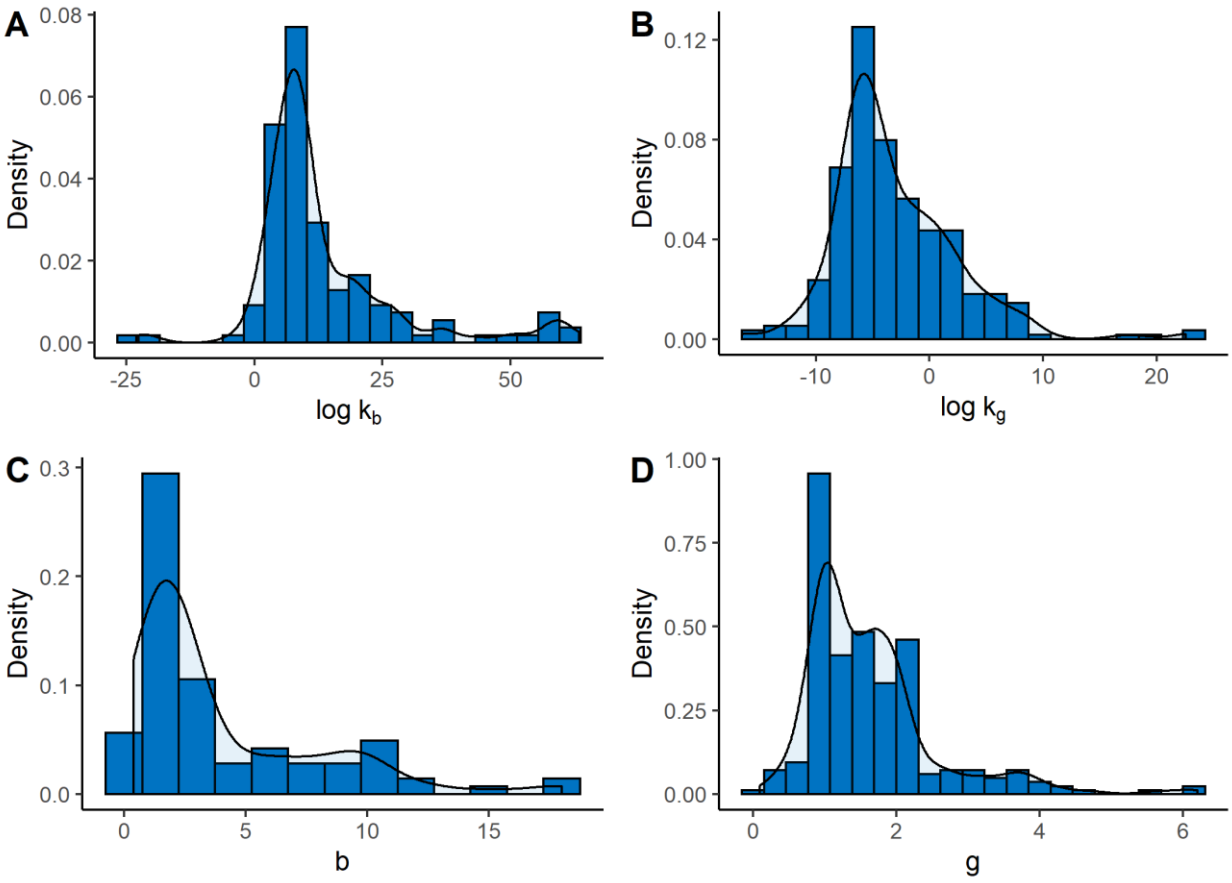

Figure 2. Histograms of kinetic parameters: A) Primary nucleation rate constants; B) Growth rate constants; C) Exponential term associated with supersaturation in primary nucleation rate; D) Exponential term associated with supersaturation in growth rate.

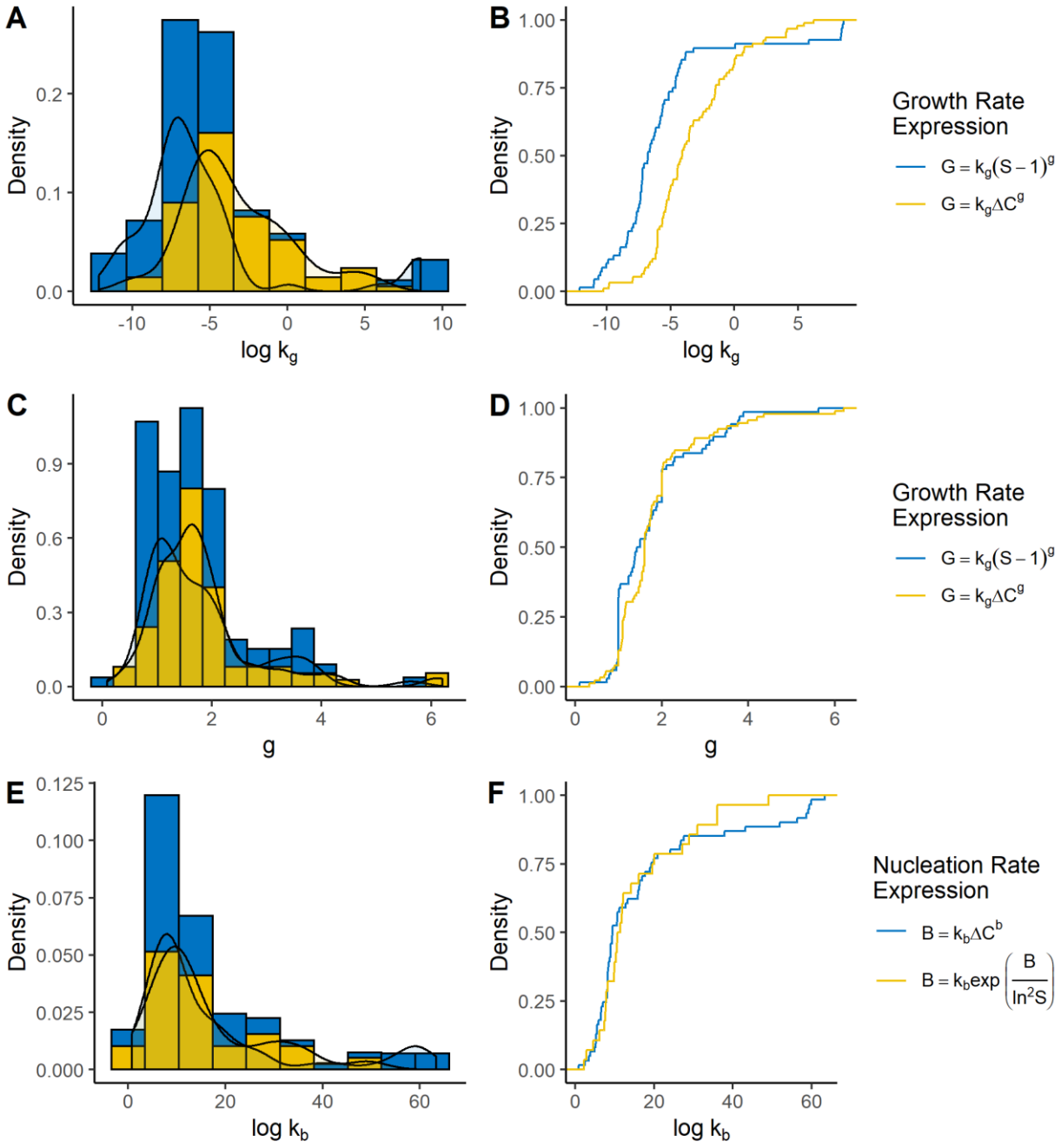

Figure 3. Histograms and empirical cumulative distribution of kinetic parameter by kinetic expression. A-D depict the distributions for the growth rate models with supersaturation ratio ($\log k_g$ median = -6.77, g median = 1.46, n =68) and absolute supersaturation ($\log k_g$ median = -4.11, g median = 1.60, n =92). E-F are the distribution of $\log k_b$ for the empirical nucleation rate model (median = 9.43, n = 61) and CNT (median = 11.00, n = 28).

The comparison of the most common kinetic models and medians are displayed in Figure 3. All the distributions were right-skewed and had a similar shape compared to the discussed previously. By contrasting cumulative distributions, it was possible to notice that k_g values were lower when growth was a function of supersaturation ratio instead of absolute supersaturation (Mann-Whitney U = 1945.5, p-value < 0.05). This difference was around two orders of magnitude. On the other hand, there seems to have been no significant difference in g between growth models (Mann-Whitney U = 2828.5, p-value = 0.301). In the same way, when k_b values from the empirical model were contrasted with CNT model, a high level of coincidence was observed, by which it could be said that the available evidence does not allow to detect significant differences (Mann-Whitney U = 821, p-value = 0.774). Thus, the only constant significantly affected by the model was k_g.

The distributions of temperature-dependent growth equations are shown in Figure 4. With the presented data, it was not possible to establish that there was a significant difference in the activation energy (Mann-Whitney U = 396, p-value = 1.000), the growth rate constant (Mann-Whitney U = 446, p-value = 0.428) and the rate order g (Mann-Whitney U = 499.5, p-value = 0.09) due to the kinetic expression. The rate order g preserved the same behaviour as the overall where the majority of data tended to be between 1.0 and 2.0. Conversely, the median of $\log k_g$ was -0.98 and most data were concentrated within ±4.00. Finally, the activation energies were mostly around 29.79 kJ/mol, with 25% and 75% of the data being lower than 15.95 kJ/mol and 45.78 kJ/mol, respectively. Considering what has been mentioned, the values of kinetic parameters were consistent in their majority with the reported in the literature.

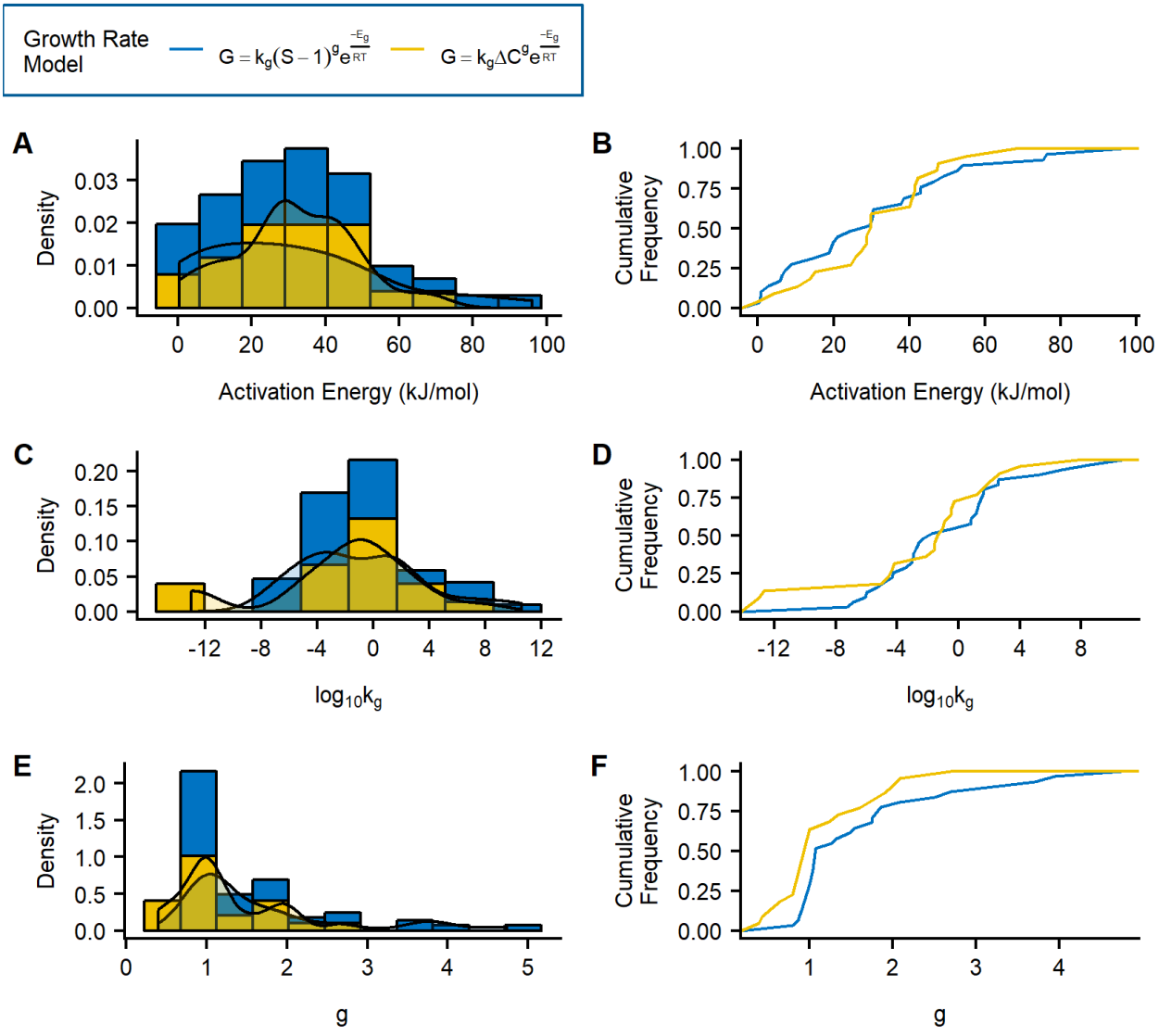

Figure 4. Histograms and cumulative frequency of temperature-dependent growth rate models. Kinetic parameters for supersaturation ratio-based model (E_g median =30.23 kJ/mol, $\log k_g$ = 0.228, g = 1.08, n = 36) and absolute supersaturation-based model (E_g median =29.70 kJ/mol, $\log k_g$ = -1.21, g = 1.00, n = 22). A-B do not include the following database entries due to being possible outliers and hinder a proper view of the majority: 102 and 262.

As for crystal growth, g depends - among other factors - on the growth mechanism which in turn depends on the supersaturation degree.[15] In this manner, it has been reported that g generally is between 1.0 and 2.0, which coincides with the results found in this work, although many data were outside this range.[14, 15] Additionally, g seems not to be affected by the way as supersaturation is expressed. However, k_g showed different values caused by the kinetic model. In line with this, these differences in the magnitude of k_g are expected. Having as a reference the models $G = k_g(S-1)^g$ and $G = k_g \Delta C^g$, it could be said that $k_g^{\{\Delta C\}} = k_g^{\{S\}}/C^{*g}$, which explain the difference. Although this was not seen clearly in temperature-dependent models possibly due to the sample size in these models. Finally, the tendency shows that different may be between 2 and 3 orders of magnitude, where the values of $k_g^{\{\Delta C\}}$ and $k_g^{\{S\}}$ are around $10^{-4.11}\ ms^{-1}(g/g\ solvent)^{-1}$ and $10^{-6.77}\ ms^{-1}$, in that respective order. On the other hand, reference values of k_g were not found for either model. However, according to the literature, growth rates may be in the order of $10^{-7}\ ms^{-1}$ and $10^{-9} - 10^{-8}\ ms^{-1}$ at supersaturation $(S-1)$ of 0.01 and 10 to 100, respectively.[15-17] Assuming g = 1 due to being the most common, k_g might take values in the order of 10^{-11} to $10^{-5}\ ms^{-1}$ for the model using supersaturation ratio. Consequently, it can be noted that most of the recorded data are within the interval previously described, indicating certain agreement with what would be expected. Activation energies in temperature dependent models did not show major divergences considering what was anticipated. Typical E_a are in the order of 40 – 60 $kJ\ mol^{-1}$ and 10 to 20 $kJ\ mol^{-1}$ for crystal growth mediated by surface integration and volume diffusion.[15] Based on the findings, activation energies tended to be around 30.00 $kJ\ mol^{-1}$, which is the middle of both growth mechanisms. Nonetheless, a large proportion of the data falls within these intervals, suggesting that the values are reliable.

Concerning primary nucleation, neither reference ranges of k_b nor b were found for the power-law empirical model. Thus, the pre-exponential terms in CNT was compared to rate constant k_b. In terms of magnitude, no large differences were observed in the models. Therefore, this suggests that the expected values and interpretation of both constants might be similar. In CNT model, the pre-exponential term is expected to be around 10^{30} # $m^3 s^{-1}$ or 10^{10} – 10^{20} # $m^3 s^{-1}$, depending on whether nucleation is homogeneous or heterogeneous.[15] As a result, it can be seen that a big portion of the constants fitted in either CNT or the power-law model is within these intervals, indicating a certain level of concordance compared to previous revisions.

To conclude this part, a database of kinetic parameters was built and, considering all the points exposed for growth and primary nucleation, it can be said, there are no major deviations between the collected data and the information available in other works. This fact provides a certain level of reliability on the data. Additionally, since the source of data is varied in terms of methods and solutes, it is possible to establish approximate intervals in which some kinetic parameters would be expected to belong. However, in this scenario, some constraints are: the limited variety of solvents and most data are concentrated in a few models, by which the studied kinetic parameters in the next sections were limited to the most common models and there might be bias towards aqueous systems.

3.2. Journal bias caused by crystallization method

Detailed results and discussion for the presence of any journal bias to specific crystallization methods is provided in the ESI. In summary, based on the entries, Organic Process Research & Development tends to have more data points related to methods such as precipitation, antisolvent, and evaporative compared to the other journals, which may suggest this journal has a bias towards non-cooling techniques. On the other hand, even though the other journals display differences in

the proportion of crystallization techniques, the available data did not allow to conclude whether these differences are caused by bias or they are of random nature. Based on the papers, journal and crystallization method seem to be independent by which the observed differences may be present by chance.

3.3. Association between kinetic parameters and descriptors

3.3.1. Molecular descriptors

First of all, the evaluation of associations and other analysis were carried out using the following models since they have the most data: $G = k_g \Delta C^g$, $G = k_g (S-1)^g$, and $B = k_b \Delta C^b$. Then, molecular descriptors were used to seek associations between kinetic parameters of the models above mentioned and solute properties. An initial approach to finding out correlations was through Pearson's coefficients (r). A list of moderate and strong correlation is shown in the ESI. The majority of variables presented weak linear correlations ($|r| < 0.3$) for all the kinetic models. In the particular instance of growth rate models, some moderate correlations ($|r|$ between 0.3 and 0.7) could be identified. Specifically, the number of moderate correlations was greater for $G = k_g (S-1)^g$ for either $\log k_g$ and g. In the same way, the variables correlated to g did not match between models, and for $\log k_g$, some overlap such as b_max1len, PEOE_VSA+4, and vsurf_DW13 which were lower in $G = k_g \Delta C^g$. Generally speaking, a similar behaviour was seen in nucleation rate constants compared to growth models. $\log k_b$ and b also showed mainly weak to moderate correlations. Nonetheless, $\log k_b$, in contrast to the other model kinetic parameters, had a strong correlation ($|r| > 0.7$) with two descriptors a_nCl (number of chlorine atoms) and vsurf_DW12 (contact distance between lowest hydrophilic energy) with around 0.78 for both descriptors. However, by analysing these correlations thoroughly, some extreme values were observed which might have caused an overestimation of these relationships. To conclude, overall,

strong correlations between solute descriptors and kinetic parameters could not be identified, except for $\log k_b$, which suggested that linear relationships between the assessed solute properties and the kinetic parameters are poor. These results indicate that these molecular descriptors may not be appropriate predictors or classifiers using linear models, by which, to discard definitely these variables, non-linear associations should be assessed.

3.3.2. Solvent

The effect of solvent on kinetic parameters was diverse. The values of growth kinetic parameters grouped by solvent are displayed in Figure 5. Starting with the model $G = k_g \Delta C^g$, the values of $\log k_g$ associated with MEK and ACN were significantly higher than the rest of solvents and these values exceed 0.0. On the other hand, the rate order g was similar among the distinct solvents, being lower than 2.0. In line with this, ACN values were the lowest with respect to the other solvents. In relation to the model $G = k_g(S-1)^g$, the values of $\log k_g$ and g were comparable to the majority of solvents, solely seeing a large difference of $\log k_g$ in aqueous mixtures and g in EtOH. The results of k_b and b for each solvent are shown in Figure 5. To point out, while MEK and the aqueous mixtures had the highest values of b, they presented the lowest values of $\log k_b$. In contrast, even though EA also possessed a high b, $\log k_b$ was comparable to water and MeOH. As for MeOH, the data were very scattered for both $\log k_b$ and b, thereby hindering the determination of a difference with respect to the other solvents. Thus, for primary nucleation as well as growth models, it was difficult to find significant variations of kinetic parameters with relation to solvent given the majority of solvents showed the tendency to be around the same range and the number of data and solutes per each solvent was rather unbalanced. Nonetheless, there are two cases to highlight: MEK in growth and MeOH in nucleation. It has been documented that numerous solvent properties such as viscosity, polarity, and chemical nature can affect crystal

growth as well as primary nucleation processes.[15, 18] Thus, significative differences among solvents were expected to be observed. However, despite the fact that there were some solvents of different nature, kinetic parameters were rather similar. By observing the particular cases of MEK in growth and MeOH in nucleation, it can be seen that these two have a wide scattering of their parameters compared to water, which is the most frequently employed solvent. MeOH data comprised two solutes – paracetamol and felodipine - crystallized by precipitation and antisolvent; every system exhibited substantial differences in its nucleation parameters. On the other hand, MEK had a wide dispersion of g which is explained by changes in cooling rate in co-crystallization of agomelatine-citric acid. Considering water, there were many more possible combination of methods, solutes and process conditions but such scattering was not exhibited. These facts indicate there might be interactions between solvent and several other factors, such as process conditions, and solvent effect may not be evaluated in isolation. In future studies, a better approach might be to analyse interactions with other factors or use solvent descriptors like viscosity, in order to identify potential associations in a clearer way.

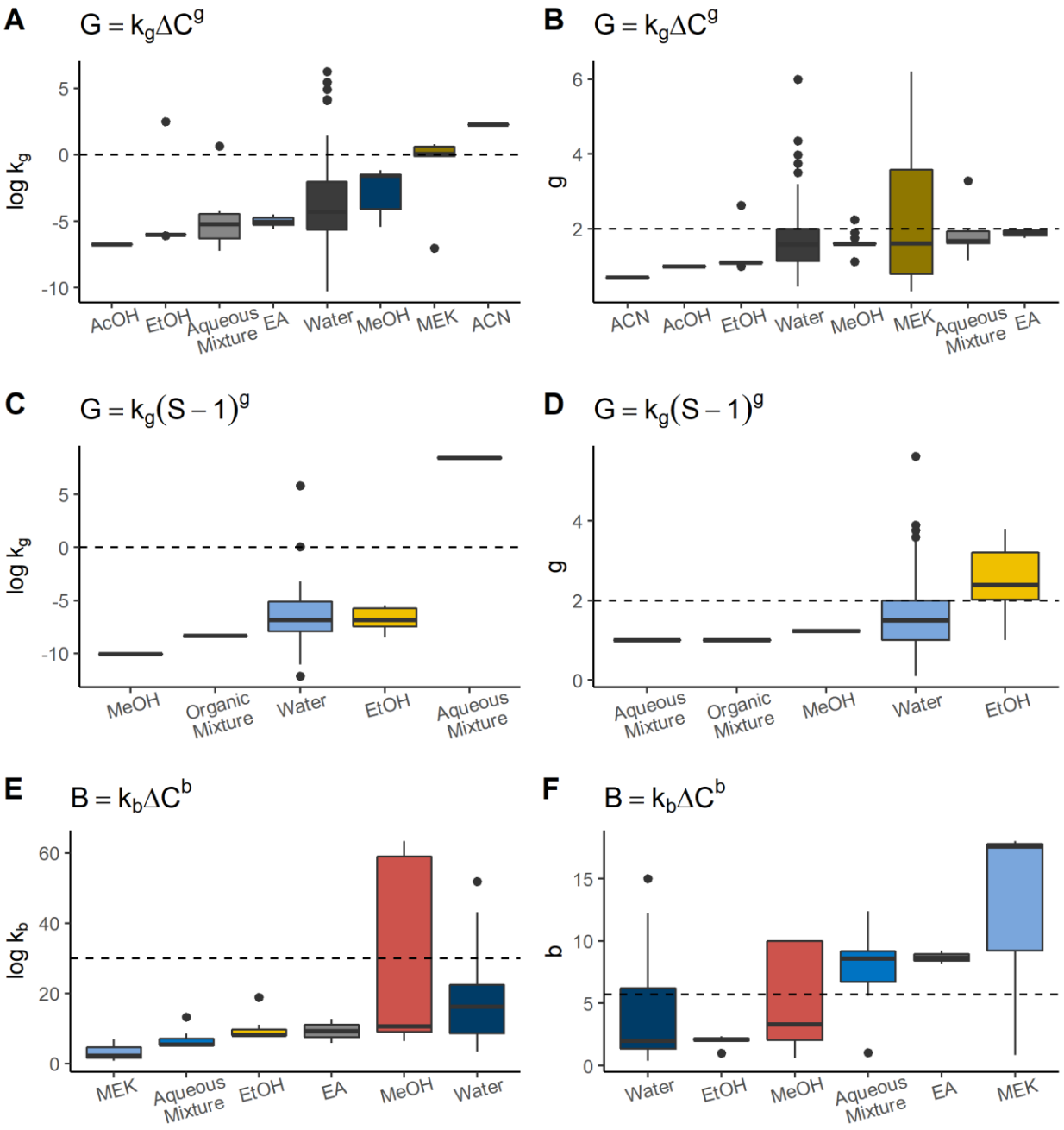

Figure 5. A-D Association growth kinetic parameters and solvent. E-F Association between primary nucleation kinetic parameters and solvent. AcOH, acetic acid; EtOH, ethanol; EA, ethyl acetate; MeOH, methanol; MEK, methyl ethyl ketone; Aqueous mixture, mixture of water + an organic solvent; organic mixture, mixture of several organic solvents. From left to right, solvents are placed in ascending order of medians.

3.3.3. Crystallization technique

Figure 6 shows boxplots of kinetic parameters separated by crystallization technique for the modes $G = k_g \Delta C^g$ and $G = k_g (S-1)^g$. Cooling and reactive crystallization presented the highest values of $\log k_g$ and g in the model $G = k_g \Delta C^g$ followed by evaporative and antisolvent. In cooling crystallization, it was observed that the data exhibited the highest scattering in both parameters by which, despite having the highest values of both kinetic parameters, these were not notably different to the other techniques. These results contrasted with the model $G = k_g (S-1)^g$ since the same patterns were not seen. In this model, for example, precipitation and cooling had the lowest values of $\log k_g$ and g. The values of g in both models tended to be high or moderately higher than 1.0 for precipitation and antisolvent. These techniques are characterised for reaching a very high level of supersaturation.[15-17] In these conditions, g is generally higher than 2.0 given the low solubility in the system.[15] In this manner, the results are consistent. Conversely, k_g does not exhibit the same behaviour, suggesting that k_g may not necessarily show a pattern related to the technique. As for cooling crystallization, the dispersion is generally wider than the other techniques. A reason might be that cooling crystallization was the most frequent and more variations of process conditions can be found. In this manner, all of these changes may lead to have a larger variance in growth constants.

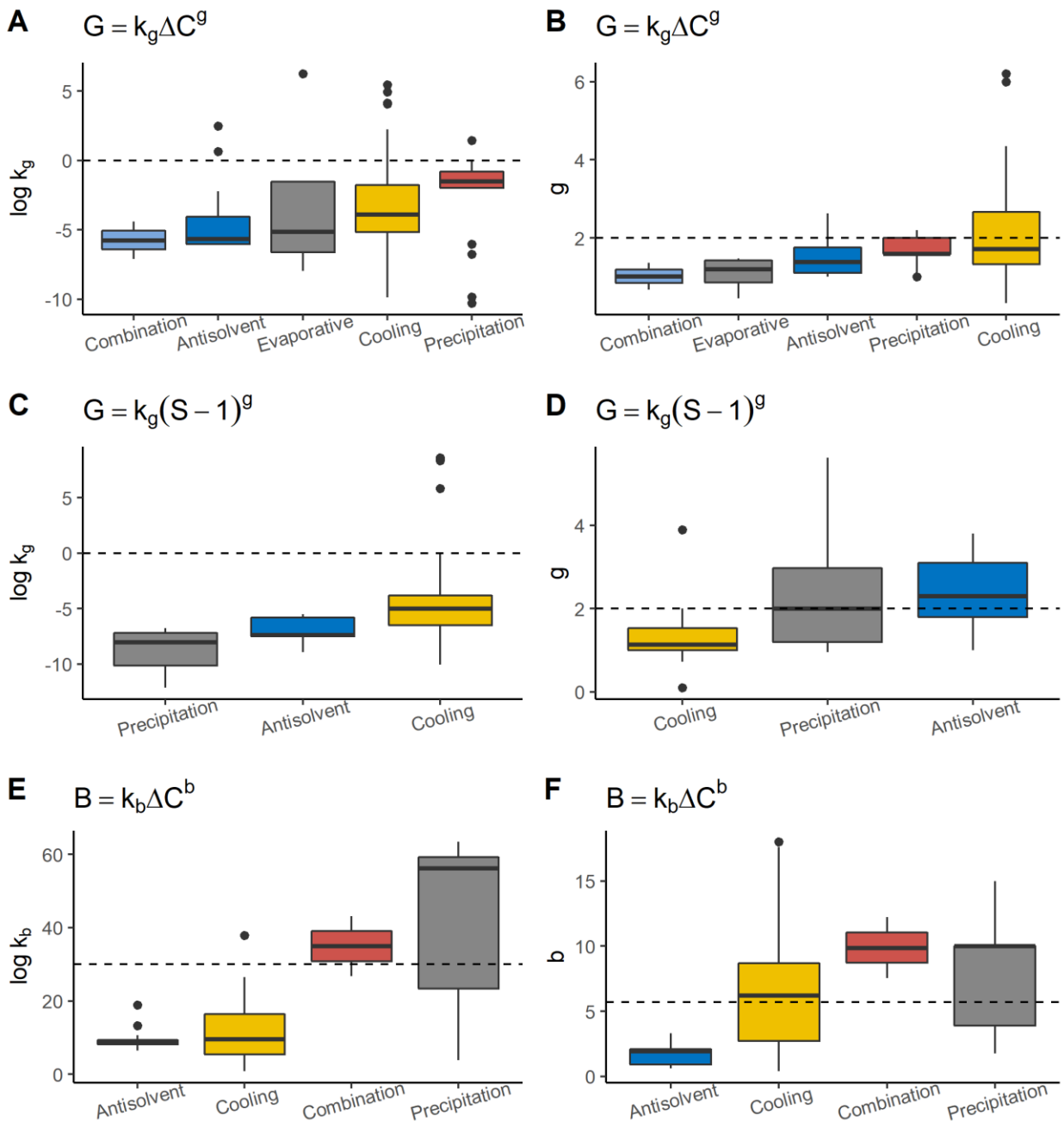

Figure 6. A-D Association growth kinetic parameters and crystallization technique. E-F Association primary nucleation parameters and crystallization technique. From left to right, techniques are placed in ascending order of medians

Nucleation rate data showed that $\log k_b$ and b have the same pattern i.e., a technique with high b, it has high $\log k_b$. Although the scattering was the highest, precipitation exhibited the largest

k_b and b preceded by cooling crystallization. It could also be observed that the majority of methods displayed values of b higher than 5.9. In opposition, antisolvent technique shows the lowest values for both nucleation parameters. Precipitation and antisolvent are characterised by large nucleation rates.[15] In this way, their parameters are expected to show the same tendency. This trend was seen for precipitation but not antisolvent. A possible reason is that the solutes crystalized by antisolvent show a moderate solubility in the system solvent – antisolvent.[15] The results are portrayed in **Figure 6**. Finally, the data indicate that there may be patterns such as the case of precipitation where, especially for primary nucleation, higher values of all the parameters compared to the others were observed.

3.3.4. Seeding

Growth kinetic constants are compared in Figure 7. Although seeded and unseeded crystallization did not seem to differ markedly, it was still possible to see small differences between groups. In general, unseeded processes showed values slightly higher than seeded crystallization. This tendency was especially more notable in g for both models. However, $\log k_g$ in $G = k_g(S-1)^g$ model exhibited the opposite, where seeded processes have greater values of k_g. In this manner, kinetic parameters were different depending on seeding but this difference did not appear substantial overall by which this parameter may not be useful to characterise growth rate parameters.

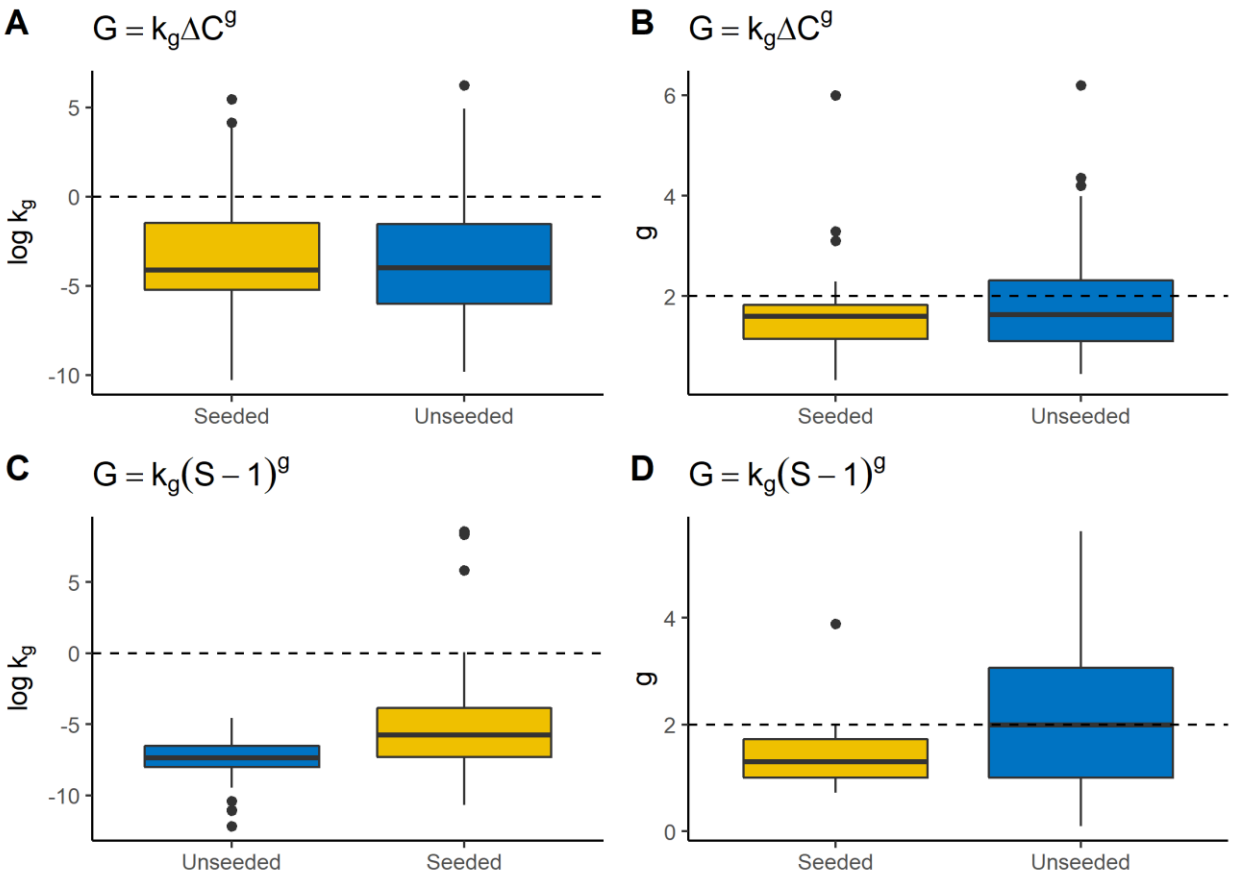

Figure 7. Association growth kinetic parameters and seeding. From left to right, seed and unseeded processes are placed in ascending order of medians.

3.4. Cluster analysis

In response to previous results in which no clear associations could be established between certain properties and kinetic parameter, ACH was performed in the observation of several models. The objective was to first identify whether the data have a cluster structure and form homogeneous groups founded on the kinetic parameters. Secondly, through a complementary methodology, it was aimed to find characteristics that enable to classify the crystallization of a solute under certain conditions in a group and provide a rough estimation of possible values for kinetic parameters.

Thus, AHC was carried out over the next models since they have more than 50 observations: $G = k_g \Delta C^g$ (G1), $G = k_g (S-1)^g$ (G2), and $B = k_b \Delta C^b$ (B1).

Initially, the optimal number of clusters was 3 in the model G1, while the optimal was 2 for the others based on the maximum Silhouette index (see ESI). Nonetheless, in the models G2 and B1, 2 clusters did not provide a good differentiation between groups in relation to the rate constant and the supersaturation rate order together. Therefore, the chosen number of clusters for these cases was the second optimal number according to the index. In this manner, the final number of clusters of 3, 3, and 5 were reached for the models G1, G2, and B1, respectively. The results for the model G1 are shown in Figure 8 and summary statistics of the cluster are listed in the ESI. Cluster 2 and 3 could be clearly discriminated by $\log k_g$ values since k_g is higher for the former. However, both groups had similar values of g as seen by comparing the means. Consequently, the growth rate of these observations is limited by the rate constant rather than supersaturation, at the same supersaturation levels. Instead, Cluster 1 showed larger values of g with respect to the other groups, but the range of $\log k_g$ was approximately the same as Cluster 2. Therefore, the growth rate of the observations that belong to Cluster 1 is more strongly dependent on supersaturation.

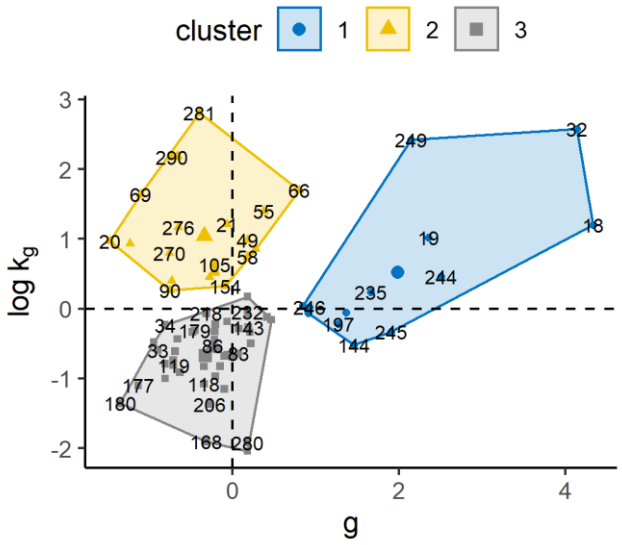

Figure 8. Scatter plot of standardised $log\ k_g$ and g for the model $G = k_g \Delta C^g$ (G1). The labels represent the identification number of the observations. Cluster observations are distributed as follows: cluster 1 (n = 13), cluster 2 (n = 27), and cluster 3 (n = 52).

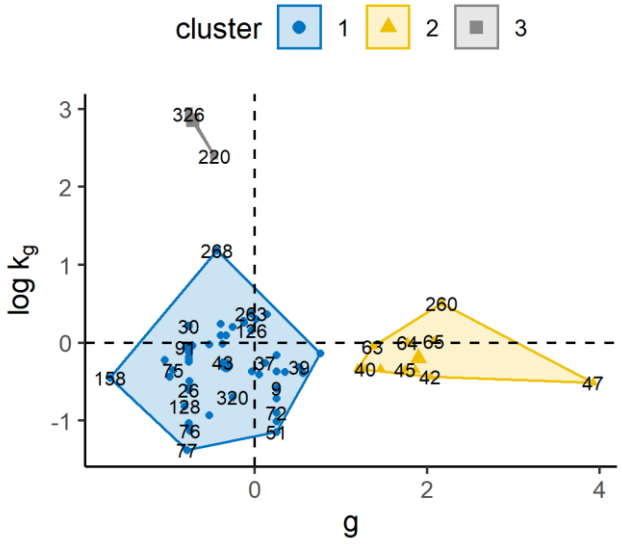

Figure 9. Scatter plot of standardised $logk_g$ and g for the model $G = k_g(S-1)^g$ (G2). The labels represent the identification number of the observations. Cluster observations are distributed as follows: cluster 1 (n = 51), cluster 2 (n = 11), and cluster 3 (n = 6).

In a similar way, 3 clusters were identified based on kinetic parameters of the model G2. In particular, Cluster 1 and 2 were comparable in terms of k_g but differ in g, where Cluster 2 presented larger rate orders. In opposition, Cluster 3 is characterized mainly by having high values of $\log k_g$ and low values of g. These results can be observed in Figure 9 and the ESI.

Regarding primary nucleation, 5 clusters were identified. The scatter plot and summary statistics can be found in **Figure 10** and ESI, respectively. Although all the groups presented different means for all the kinetic parameter, they still had some values that could overlay. In relation to this, Cluster 1 to 3 showed similar values of $\log k_b$, whereas the order b was distinct between groups. Similarly, Cluster 4 showed observations similar to those of group 1 and 2. Conversely, Cluster 5 was distinct in $\log k_b$ as well as b values. To note, Cluster 3 was composed by 2 observations only which belonged to the same solute. These observations corresponded to an experiment related to co-crystallization of agomelatine/citric acid. Given the characteristics of the solutes, this group was not included in the later analysis since molecular descriptors were not appropriate.

Finally, the data were segmented into different groups for each considered model. Clusters exhibited particular values of either the rate constant or the supersaturation order. The relationships between clusters and molecular descriptors, solvent, methods and seeding are analysed in the next section through random forest model (RF).

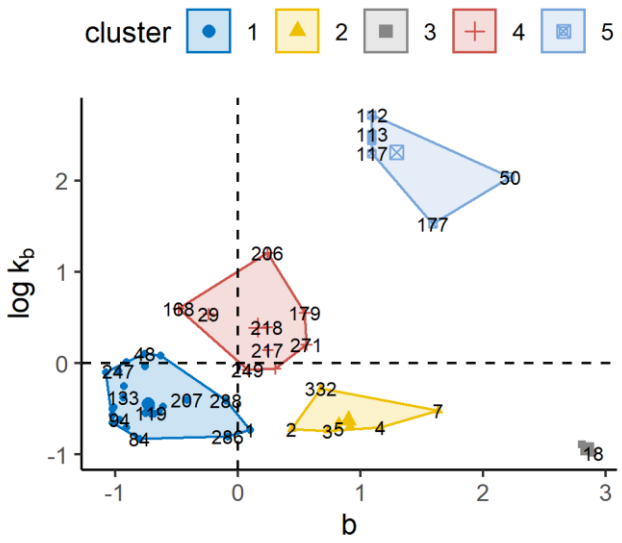

Figure 10. Scatter plot of standardised log k_b and b for the model $B = k_b \Delta C^b$ (B1). The labels represent the identification number of the observations. Cluster observations are distributed as follows: cluster 1 (n =34), cluster 2 (n = 8), and cluster 3 (n = 2), cluster 4 (n = 9), and cluster 5 (n = 8).

3.5. Descriptors importance

RF possesses, among many other advantages, the ability to deal with non-linear relationships and redundant information, and assign importance to the classifiers, which is useful in the selection of variables and search of patterns. With this in mind, RF models were built for the kinetic expressions G1, G2, and B1 with the next parameters: ntree = 15000, mtry = 10, and set.seed = 50. The out-of-bag (OOB) and class prediction error are listed below in Table 3. The high errors within groups were generally associated with the smallest size class. Additionally, the predictability was evaluated via leave-one-out cross-validation. The overall classification accuracy was 74.11%, 85.45%, and 83.05% for the models G1, G2, and B1, respectively. Previous works dealing with application of RF in crystallization phenomenon showed a level of accuracy around 70%.[19] Therefore, the proposed models can be considered acceptable in this aspect.

Table 3. OOB and class error of the RF models

	$G = k_g \Delta C^g$ (G1)	$G = k_g (S-1)^g$ (G)	$B = k_b \Delta C^b$ (G3)
OOB (%)	25.88	14.55	16.95
Class error (%)			
Cluster 1	36.36	10.26	2.94
Cluster 2	30.43	30.00	0.00
Cluster 3	21.56	16.67	-
Cluster 4	-	-	77.78
Cluster 5	-	-	25.00

Figure 11 shows the top 15 of the most important variables for RF classification. All the models included solvent, method, seeding and 110 molecular descriptors as classifiers. For all the three models, among the most common and important classifiers were found mostly descriptors related to partial charges (PEOE), topological indices such as BCUT and GCUT, and volume-surface-shape indices (vsurf). Variables such as seeding and solvent were not as relevant as the other descriptors. Instead, crystallization technique (method) was among the top 15 of the most important variables only in primary nucleation rate model. The **Figure 11** also shows that after the first one or two ranked variables, MDA is reduced slowly which suggests that there are no large differences in the importance after the first one. Thus, this might indicate that the contribution of the majority of variables to the model predictability is similar. As a result, there are no outstanding variables but most of them contribute equally.

By observing Table 4, it is possible to notice that the 3 most important variable were different with respect to mean throughout all clusters. As a result, these classifiers can be potentially useful to distinguish one group from another. However, some clusters had a high standard deviation and

so a high scattering. Therefore, the observations of two clusters may overlap. In this manner, the most important descriptors may not be enough to provide accurate discrimination between groups. This can be seen for instance in the descriptor GCUT_PEOE_3 of model G1. Cluster 1 had a lower value than the others but the descriptor in Cluster 2 and 3 was rather similar, around 2.1. Thus, the best descriptor only can identify Cluster 1 from the rest in this case. Furthermore, Cluster 1 has a wide scattering with respect to its average, which means some observations of this group might overlap with the others, thereby being confused. In light of the mentioned limitations of the descriptors, the high scattering within clusters may provide an explanation for why the MDA is rather similar and low in the models given the descriptors may separate a cluster from another but not all the clusters. Consequently, this suggests that a variable in isolation cannot explain the variability between clusters and the best model requires many variables.

Table 4. Expected values (standard deviation) of the 3 most important classifiers for each cluster.

Cluster	**Descriptors**		
Model G1	GCUT_PEOE_3	vsurf_IW7	BCUT_PEOE_3
1	1.56 (0.66)	1.95 (2.23)	1.80 (0.65)
2	2,12 (0.50)	1.26 (2.12)	2.24 (0.37)
3	2.17 (0.46)	3.53 (2.13)	2.44 (0.34)
Model G2	PEOE_VSA+2	vsurf_R	PEOE_RPC+
1	6.65 (11.13)	1.61 (0.16)	0.51 (0.36)
2	20.23 (16.04)	1.38 (0.15)	0.22 (0.19)
3	24.72 (12.11)	1.23 (0.02)	0.11 (0.04)
Model B1	BCUT_SLOGP_1	GCUT_SLOGP_1	GCUT_PEOE_2
1	-0.68 (0.43)	-0.53 (0.48)	0.14 (0.14)
2	-0.57 (0.06)	-0.43 (0.10)	0.09 (0.00)

4	-0.63 (0.52)	-0.59 (0.54)	0.28 (0.31)
5	-0.21 (0.06)	-0.28 (0.08)	-0.05 (0.04)

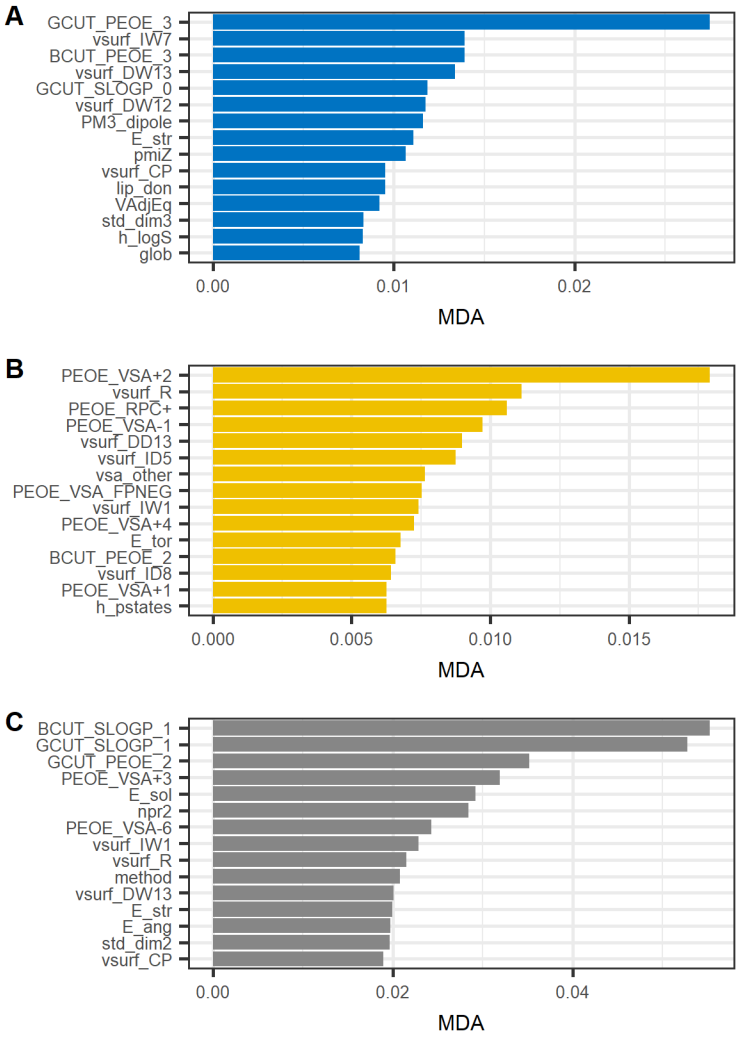

Figure 11. Top 15 of the most important classifiers based on mean decrease in accuracy (MDA). A) model G1, $G = k_g \Delta C^g$, B) model G2, $G = k_g (S - 1)^g$, C) model B1, $B = k_b \Delta C^g$.

By comparing the most important descriptor in the proposed models to previous works in crystallization and solubility, several coincidences can be found. Specifically, MOE descriptors such as BCUT, GCUT and partial charge (PEOE) have been found useful to predict solubility and crystallisability,[19, 20] which match with the findings in this work to a certain extent. From a

conceptual point of view, BCUT and GCUT descriptors are topological indices which are calculated based on molecular graphs.[21] This group of indices have been related to chemical features like branching, size and cyclicity which in turn are related to molecular flexibility and rigidity.[22] These properties have been described to influence on crystallization tendency and kinetics.[23, 24] In this way, descriptors that measure properties like molecular flexibility are expected to be relevant in crystallization models. Similarly, partial charge is important since affect the interactions solute-solvent and solute-solute.[14, 25] These descriptors were primarily relevant in the model G2 and model B1. The difference between model G1 and G2 may be given by the definition of the rate constant in which, as mentioned in previous sections, $k_g^{\{\Delta C\}} = k_g^{\{S\}}/C^{*g}$. As can be seen, k_g in the model G1 is more solubility-dependent whereby differences in important descriptors can arise, even though both models describe the same process. Lastly, vsurf descriptors comprise indices that characterise surface properties which include hydrophobic and hydrophilic interactions, shape, etc.[26] This group of indices is calculated considering molecular conformation which makes them different from partial charge descriptors, for example.[26] These types of interactions are also important in nucleation and crystal growth.[15] In this manner, descriptors that represent interactions between solute-solvent or solute-solute may be of help to describe crystallization kinetics.

To highlight, seeding, solvent, and methods were not important for growth models, and only the crystallization technique had some relevance in primary nucleation model. These results were expected since no associations between kinetic parameters and these variables were observed, except between the crystallization technique and nucleation parameters, as discussed in previous sections. By revising the results of model B1, a clearer association between crystallization technique and nucleation parameters can be observed given there is a dominant method in every

cluster as follows: Cluster 1: 64.7% antisolvent, Cluster 2: 100% cooling, Cluster 4: 77.8% cooling and Cluster 5: 87.5% precipitation. This might suggest that every cluster may also be associated with a determined crystallization method. Nonetheless, this result did not include evaporative crystallization as there were not data of primary nucleation under this condition. In the end, this indicates that RF models were able to discriminate irrelevant variables and select the most important in the correspondent model.

To summarise, RF classification models with acceptable accuracy were built. These models may yield very rough estimates of kinetic parameters for the models $G = k_g \Delta C^g$, $G = k_g (S-1)^g$, and $B = k_b \Delta C^b$, by providing mostly information on certain molecular descriptors and crystallization technique. Among the main limitations of these models, it can be found that most training data were limited to water. Although solvent was not important, a possible reason is that there was no sufficient variety of solvents to capture the variability and have an appropriate measurement of its effect, whereby it would be recommended to incorporate more solvents and study solvent molecular descriptors. Another constraint was the sample size per cluster. It would have been desirable to have a larger sample with a greater number of solutes to produce better groups and obtain more accurate models. A final limitation was concerning molecular descriptors. Specifically, 3D descriptors such as vsurft are dependent on the molecule conformation. For this work, the optimal conformation was not selected so that in future works, this might be considered to obtain more accurate values.

4. CONCLUSIONS

A database was built containing relevant information on kinetic parameters of different solutes and solvents at several conditions of seeding and crystallization technique. The data were

contrasted to theoretical data and showed to be consistent, thereby being useful to develop other analysis. The most common kinetic models were $G = k_g \Delta C^g$, $G = k_g (S-1)^g$, and $B = k_b \Delta C^b$. The parameters of these models were used to assess association with other variables. In specific, kinetic parameters relationships with 110 solute molecular descriptors, solvent, seeding and crystallization technique were studied. No strong linear correlations were found between molecular descriptors and kinetic parameters. Similarly, a clear association of kinetic parameters with seeding or solvent was not observed. On the other hand, while crystallization technique did not display a tendency in regards growth parameters, a notable association was seen with primary nucleation parameters.

In order to look for patterns, hierarchical clustering analysis was performed in the kinetic parameters of each model. A cluster structure was identified and the observations were assigned to a group. Later, random forests models were built to classify observations in the groups established by clustering analysis, using as classifiers the variables employed during the assessment of the associations. Three random forest models were obtained for each kinetic model. The overall classification accuracy calculated by leave one out-cross-validation was higher than 70% for all the models. The most important variables for classification were topological (BCUT and GCUT), partial charge (PEOE), and vsurf descriptors showing certain association with kinetic parameters. In addition, crystallization technique was relevant to classify observation in primary nucleation, which confirms its relationship with nucleation parameters.

These models may be employed to yield a rough estimate of kinetic parameters of crystal growth and primary nucleation. However, the models are mostly constraint to aqueous systems. In this manner, it was possible to establish that developing a model to predict kinetic constants is feasible.

Future works in this field should focus on providing more accurate estimations. In this scenario, considering the following factors might be useful:

1. Increase the number of solutes for each model.
2. Increase the number and nature of solvents.
3. Model solvent molecular descriptor.
4. Select optimal conformation to calculate solute molecular descriptors.

To aid in points 1 and 2, the authors welcome contributions from researchers to expand the database. Original and updated versions of the database will remain freely available from the University of Strathclyde KnowledgeBase at https://doi.org/10.15129/8f47a175-3ac7-4791-a310-82e6652bd9f5.

ASSOCIATED CONTENT

All data underpinning this publication are openly available from the University of Strathclyde KnowledgeBase at https://doi.org/10.15129/8f47a175-3ac7-4791-a310-82e6652bd9f5**:**

- All the data collected with and without pre-processing, observations whose kinetic parameters were a function of solvent or antisolvent concentration, observations whose growth was measured as volume, data adjusted according to what was explained in the article (dataset_raw.csv and dataset_preprocessed.csv)
- Molecular descriptors employed in random forests of the compounds in the database (moe_descriptors.csv)
- Code employed to perform cluster analysis and random forests in R (script.html)

AUTHOR INFORMATION

Corresponding Author

*E-mail: cameron.brown.100(at)strath.ac.uk

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors would like to thank EPSRC and the Future Continuous Manufacturing and Advanced Crystallisation Research Hub (Grant Ref: EP/P006965/1) for funding this work.

REFERENCES

(1) Rantanen, J.; Khinast, J., The Future of Pharmaceutical Manufacturing Sciences. *Journal of pharmaceutical sciences* **2015,** 104, (11), 3612-3638.
(2) Lee, S. L.; O'Connor, T. F.; Yang, X.; Cruz, C. N.; Chatterjee, S.; Madurawe, R. D.; Moore, C. M. V.; Yu, L. X.; Woodcock, J., Modernizing Pharmaceutical Manufacturing: from Batch to Continuous Production. *Journal of Pharmaceutical Innovation* **2015,** 10, (3), 191-199.
(3) Gernaey, K. V.; Cervera-Padrell, A. E.; Woodley, J. M., A perspective on PSE in pharmaceutical process development and innovation. *Computers & Chemical Engineering* **2012,** 42, 15-29.
(4) Rogers, A.; Ierapetritou, M., Challenges and opportunities in modeling pharmaceutical manufacturing processes. *Computers & Chemical Engineering* **2015,** 81, 32-39.
(5) Pandey, P.; Bharadwaj, R.; Chen, X., 1 - Modeling of drug product manufacturing processes in the pharmaceutical industry. In *Predictive Modeling of Pharmaceutical Unit Operations*, Pandey, P.; Bharadwaj, R., Eds. Woodhead Publishing: 2017; pp 1-13.
(6) Kremer, D. M.; Hancock, B. C., Process Simulation in the Pharmaceutical Industry: A Review of Some Basic Physical Models. *Journal of pharmaceutical sciences* **2006,** 95, (3), 517-529.
(7) Omar, H. M.; Rohani, S., Crystal Population Balance Formulation and Solution Methods: A Review. *Crystal Growth & Design* **2017,** 17, (7), 4028-4041.
(8) Kwartler, T., *Text Mining in Practice with R*. ed.; Wiley: 2017.
(9) Agresti, A., *Categorical Data Analysis*. ed.; Wiley: 2003.
(10) Johnson, R. A.; Wichern, D. W., *Applied Multivariate Statistical Analysis (Classic Version)*. ed.; Pearson: 2018.
(11) Kassambara, A., *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning*. ed.; STHDA: 2017.
(12) Hastie, T.; Tibshirani, R.; Friedman, J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. ed.; Springer New York: 2009.

(13) Williams, G., *Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery*. ed.; Springer New York: 2011.
(14) Myerson, A. S.; Erdemir, D.; Lee, A. Y., *Handbook of Industrial Crystallization*. 3 ed.; Cambridge University Press: 2019.
(15) Lewis, A.; Seckler, M.; Kramer, H.; van Rosmalen, G., *Industrial Crystallization: Fundamentals and Applications*. ed.; Cambridge University Press: 2015.
(16) Rudolph, P., *Handbook of Crystal Growth: Bulk Crystal Growth*. ed.; Elsevier Science: 2014.
(17) Nishinaga, T., *Handbook of Crystal Growth: Fundamentals*. ed.; Elsevier Science: 2014.
(18) Du, W.; Yin, Q.; Gong, J.; Bao, Y.; Zhang, X.; Sun, X.; Ding, S.; Xie, C.; Zhang, M.; Hao, H., Effects of Solvent on Polymorph Formation and Nucleation of Prasugrel Hydrochloride. *Crystal Growth & Design* **2014,** 14, (9), 4519-4525.
(19) Bhardwaj, R. M.; Johnston, A.; Johnston, B. F.; Florence, A. J., A random forest model for predicting the crystallisability of organic molecules. *CrystEngComm* **2015,** 17, (23), 4272-4275.
(20) Palmer, D. S.; O'Boyle, N. M.; Glen, R. C.; Mitchell, J. B. O., Random Forest Models To Predict Aqueous Solubility. *Journal of Chemical Information and Modeling* **2007,** 47, (1), 150-158.
(21) Roy, K.; Kar, S.; Das, R. N., Chapter 2 - Chemical Information and Descriptors. In *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*, Roy, K.; Kar, S.; Das, R. N., Eds. Academic Press: Boston, 2015; pp 47-80.
(22) Hu, Q.-N.; Liang, Y.-Z.; Yin, H.; Peng, X.-L.; Fang, K.-T., Structural Interpretation of the Topological Index. 2. The Molecular Connectivity Index, the Kappa Index, and the Atom-type E-State Index. *Journal of Chemical Information and Computer Sciences* **2004,** 44, (4), 1193-1201.
(23) Bai, J.; Fang, H.; Zhang, Y.; Wang, Z., Studies on crystallization kinetics of bimodal long chain branched polylactides. *CrystEngComm* **2014,** 16, (12), 2452-2461.
(24) Yu, L.; Reutzel-Edens, S. M.; Mitchell, C. A., Crystallization and Polymorphism of Conformationally Flexible Molecules: Problems, Patterns, and Strategies. *Organic Process Research & Development* **2000,** 4, (5), 396-402.
(25) Kowacz, M.; Prieto, M.; Putnis, A., Kinetics of crystal nucleation in ionic solutions: Electrostatics and hydration forces. *Geochimica et Cosmochimica Acta* **2010,** 74, (2), 469-481.
(26) Cruciani, G.; Mannhold, R.; Kubinyi, H.; Folkers, G., *Molecular Interaction Fields: Applications in Drug Discovery and ADME Prediction*. ed.; Wiley: 2006.