Electronic Supplementary Material (ESI) for Digital Discovery.


# Data mining crystallization kinetics: Electronic Supporting Information

*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 Kingdon

**Database information**

**Table 1. Search strategies and databases.**

| Database | Search keywords |
|---|---|
| ScienceDirect:<br>https://www.sciencedirect.com/ | (((growth nucleation) OR (kinetic) OR MSMPR) AND (“population balance” crystal) AND (estimation OR determination))) NOT (granulation OR precipitation)<br><br>(growth OR nucleation OR kinetic*) AND (“population balance”) AND (pharmaceutical OR drug OR API) AND crystal*<br><br>(“population balance” AND crystal*) AND (pharma* OR drug ) |
| ACS Publications:<br>https://pubs.acs.org/ | “population balance” crystallization kinetics |
| AIChE:<br>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 |

**Note:** Boolean operators were employed only in ScienceDirect and AIChE websites since those allowed their usage and therefore more complex strategies could be used.

**Table 2. Words used to exclude articles.**

| | | | | |
|---|---|---|---|---|
| Granulation | Dehydration | Wax | Emulsion | Granules |
| Protein | Company | View | Cell | Mills |
| Wet | Fields | Ligand | Edited | Map |
| Review | Decomposition | Graphene | Methane | Biomass |
| Polymerization | Emulsification | Argon | Future | Decracemization |
| Bubble | Atomization | Oxygen | Overview | Hydrogenation |
| Challenges | Diffraction | Zno | Advances | Biological |
| Magnetic | Zeolite | Desulfurization | Emulsions | Electrical |
| Science | Ethylene | Culture | Oil | Catalyzed |
| Milling | Principles | Enzymatic | Granule | Behavior |
| Mill | Ethane | Freezing | Columns | Ball |
| Catalytic | Paper | Scheduling | John | Granular |
| Mcgraw | Monograph | Peptide | Cells | Cavitation |
| Next | Discovery | Rheology | Enzyme | Chromatography |

**Table 3. Information extracted from the final search results.**

| Variable description | Name | Type | Comments |
|---|---|---|---|
| Number of identification | id | Numeric | |
| Article title | title | Alphanumeric | |
| Article journal | journal | Alphanumeric | |
| Article author | author | Alphanumeric | |

| | | | |
|---|---|---|---|
| Solute | | Alphanumeric | |
| Solvent | | Alphanumeric | |
| Antisolvent | | Alphanumeric | When the method is antisolvent |
| Method | | Alphanumeric | |
| Seeding | seeded | Yes (seeded), No (unseeded), both | "both" means the determination of kinetic parameters was based on both seeded and unseeded experiments |
| Exponential term associated with supersaturation in primary nucleation | b | Numeric | |
| Pre-exponential or pre-supersaturation constant associated with primary nucleation | kb | Numeric | |
| Exponential term associated with supersaturation in growth rate | g | Numeric | |
| Pre-exponential or pre-supersaturation constant associated with growth rate | kg | Numeric | |
| Growth rate activation energy | ea.nucleation | Numeric | |
| Nucleation rate activation energy | ea.growth | Numeric | |
| Units kb | kb.units | Alphanumeric | |
| Units kg | kg.units | Alphanumeric | |
| Units Eb | ea.nucleation.units | Alphanumeric | |
| Units Eb | ea.growth.units | Alphanumeric | |
| Growth rate expression | growth.rate | Alphanumeric | |

| | | | |
|---|---|---|---|
| Nucleation rate expression | nucleation.rate | Alphanumeric | |
| Driving force expression | driving.force | Alphanumeric | |
| Driving force units | units.driving.force | Alphanumeric | |
| Others constants | other.constants | Alphanumeric | |
| Comments | - | Alphanumeric | Additional information about experimental conditions, solute characteristics or solvent composition. |

**Data analysis: correlation molecular descriptors vs kinetic parameters**

**Table 4. Moderate and strong correlations between molecular descriptors and kinetic parameters**

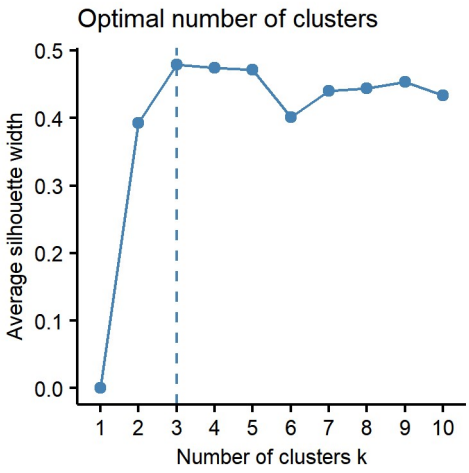
| **Kinetic parameter** | **Pearson correlation absolute value ($\|r\|$)** | **Descriptor** |
|---|---|---|
| | | $G = k_g \Delta C^g$ |
| $\log k_g$ | 0.4 – 0.5 (1) | vsurf_IW8 |
| | 0.3 – 0.4 (13) | a_ICM, b_max1len, lip_don, PEOE_RPC-, PEOE_VSA+4, SMR_VSA4, vsurf_CW2, vsurf_DW12, vsurf_DW13,vsurf_ID7, vsurf_ID8, vsurf_IW7, vsurf_Wp6 |
| $g$ | 0.3 – 0.4 (1) | MNDO_dipole |
| | | $G = k_g (S - 1)^g$ |
| $\log k_g$ | 0.5 – 0.6 (3) | b_max1len, PEOE_VSA+4, SMR_VSA1 |
| | 0.4 – 0.5 (6) | balabanJ, GCUT_PEOE_0, h_pKb, h_pstrain, SMR_VSA6, vsurf_DW13 |
| | 0.3- 0.4 (9) | GCUT_SMR_0, lip_don, logP(o/w), rsynth, SlogP_VSA3, SMR_VSA0, vsa_other, vsurf_ID8, vsurf_IW1 |
| $g$ | 0.5 – 0.6 (2) | PEOE_VSA-1, pmiZ |
| | 0.4 – 0.5 (3) | E_rsol, h_pstates, logP(o/w), opr_brigid, PEOE_VSA+5, vsurf_Wp6 |

| | | |
|---|---|---|
| | 0.3 – 0.4 (13) | a_nS, GCUT_PEOE_1, h_pavgQ, h_pKa, npr1, PEOE_VSA-2, PEOE_VSA-6, PEOE_VSA_FPNEG, SlogP, SMR_VSA1, std_dim2, vsa_other, vsurf_R |
| | | $B = k_b \Delta C^b$ |
| $\log k_b$ | > 0.7 (2) | a_nCl, vsurf_DW12 |
| | 0.6 – 0.7 (3) | E_ang, SlogP_VSA6, vsurf_DW13 |
| | 0.5 – 0.6 (5) | BCUT_PEOE_1, E_str, PEOE_VSA+3, vsurf_CP, vsurf_CW1 |
| | 0.4 – 0.5 (4) | BCUT_PEOE_2, npr2, PEOE_VSA-1, vsurf_IW7 |
| | 0.3 – 0.4 (12) | GCUT_PEOE_2, KierA1, logP(o/w), npr1, PEOE_VSA+2, PEOE_VSA+5, PM3_dipole, SMR_VSA6, std_dim2, vsa_acc, vsa_other, vsurf_CW2 |
| $b$ | 0.5 – 0.6 (3) | E_rsol, E_str, PEOE_VSA+3 |
| | 0.4 – 0.5 (5) | h_pKa, PEOE_VSA+4, rsynth, vsurf_CP, vsurf_IW8 |
| | 0.3 – 0.4 (12) | BCUT_PEOE_1, BCUT_SLOGP_1, E_ang, h_pavgQ, h_pstates, npr2, PEOE_VSA-4, PEOE_VSA-6, SMR_VSA1, SMR_VSA6, std_dim3, vsurf_DW12 |

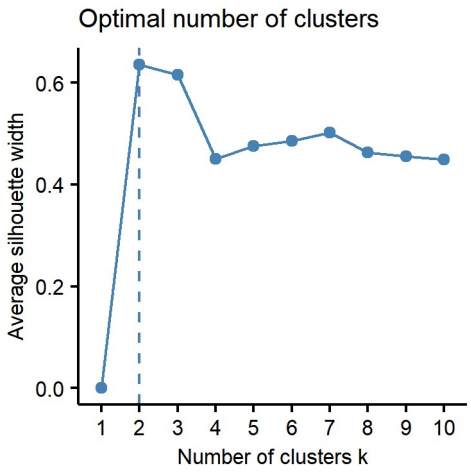
**Cluster analysis and silhouette index plots**

The results for the selection of the optimal number of cluster can be found below. The optimal number of clusters corresponds to the one that provides the highest index.

**A** $G = k_g \Delta C^g$




**B** $G = k_g (S-1)^g$




**C** $B = k_b \Delta C^b$

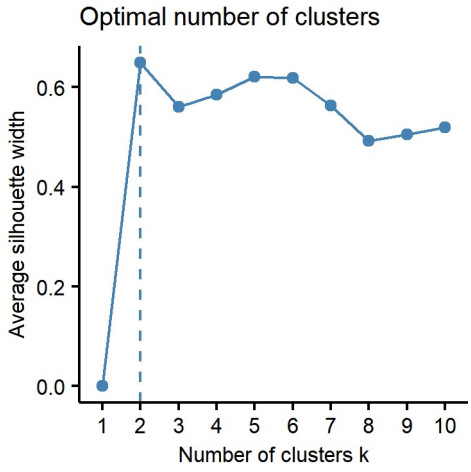


**Table 5. Summary statistics of cluster obtained for the model $G = k_g \Delta C^g$ (G1).**

| Cluster | Mean | Median | Min | Max | Standard deviation |
|---|---|---|---|---|---|
| | | $\log k_g$ | | | |
| 1 | -1.55 | -3.19 | -5.09 | 5.45 | 3.43 |
| 2 | 0.22 | -0.40 | -2.42 | 6.24 | 6.24 |
| 3 | -5.62 | -5.55 | -10.29 | -2.74 | -2.74 |
| | | $g$ | | | |
| 1 | 3.82 | 3.50 | 2.67 | 6.20 | 1.14 |
| 2 | 1.48 | 1.60 | 0.33 | 2.63 | 0.52 |
| 3 | 1.49 | 1.57 | 0.45 | 2.29 | 0.43 |

**Table 6. Summary statistics of cluster obtained for the model $G = k_g (S-1)^g$ (G2).**

| Cluster | Mean | Median | Min | Max | Standard deviation |
|---|---|---|---|---|---|
| | | $log\ k_g$ | | | |
| 1 | -6.99 | -6.83 | -12.15 | 0.06 | 2.34 |
| 2 | -6.53 | -7.19 | -8.05 | -3.19 | 1.40 |
| 3 | 8.02 | 8.42 | 5.83 | 8.58 | 1.07 |
| | | $g$ | | | |
| 1 | 1.43 | 1.38 | 0.10 | 2.50 | 0.50 |
| 2 | 3.62 | 3.50 | 2.93 | 5.62 | 0.74 |
| 3 | 1.05 | 1.00 | 1.00 | 1.29 | 0.12 |

**Table 7. Summary statistics of cluster obtained for the model $B = k_b \Delta C^b$ (B1)**

| Cluster | Mean | Median | Min | Max | Standard deviation |
|---|---|---|---|---|---|
| | | $\log k_b$ | | | |
| 1 | 9.88 | 8.89 | 3.46 | 19.22 | 4.07 |
| 2 | 6.85 | 5.69 | 4.93 | 12.76 | 2.66 |
| 3 | 1.58 | 1.58 | 0.78 | 2.38 | 1.13 |
| 4 | 24.01 | 24.11 | 16.24 | 37.85 | 6.73 |
| 5 | 56.45 | 58.88 | 43.17 | 63.36 | 6.28 |
| | | $b$ | | | |
| 1 | 1.92 | 1.85 | 0.38 | 5.60 | 1.24 |
| 2 | 9.15 | 8.80 | 7.10 | 12.40 | 1.60 |
| 3 | 17.80 | 17.80 | 17.60 | 18.00 | 0.28 |
| 4 | 5.88 | 6.23 | 3.00 | 7.63 | 1.51 |
| 5 | 10.90 | 10.00 | 10.00 | 15.00 | 1.83 |

**Journal bias evaluation**

- **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 articles, according to the case. Tables 8 to 10 summarize the number of entries and papers for each journal found in the database.

- **Journal bias caused by crystallization method**

Detailed results and discussion for the presence of any journal bias to specific crystallization methods is provided below. 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.

**Table 8. Number of entries and papers for each journal included in the database. The journals employed for both journal bias analyses are in bold.**

| Journal | Entries | Papers | Entries/paper |
|---|---|---|---|
| **Crystal Growth & Design** | 67 | 38 | 1.76 |
| **Industrial & Engineering Chemistry Research** | 65 | 36 | 1.81 |
| **Journal of Crystal Growth** | 48 | 24 | 2.00 |
| **AIChE Journal** | 35 | 19 | 1.84 |
| **Chemical Engineering Science** | 28 | 19 | 1.47 |
| **Chemical Engineering Research and Design** | 22 | 14 | 1.57 |
| **Organic Process Research & Development** | 15 | 9 | 1.67 |
| **Chemical Engineering and Processing: Process Intensification** | 13 | 8 | 1.62 |
| **Chemical Engineering Journal** | 10 | 5 | 2.00 |
| Powder Technology | 3 | 3 | 1.00 |

| | | | |
|---|---|---|---|
| CrystEngComm | 14 | 2 | 7.00 |
| Chirality | 1 | 1 | 1.00 |
| Computers & Chemical Engineering | 1 | 1 | 1.00 |
| Crystal Research and Technology | 2 | 1 | 2.00 |
| Industrial & Engineering Chemistry Fundamentals | 3 | 1 | 3.00 |
| International Journal of Modern Physics B | 4 | 1 | 4.00 |
| Journal of Crystallization Process and Technology | 2 | 1 | 2.00 |
| Journal of Process Control | 1 | 1 | 1.00 |
| The Canadian Journal of Chemical Engineering | 2 | 1 | 2.00 |

**Table 9. Contingency table for Journal and Crystallization Method based on entries.**

| Journal | Abb | Cooling | Others |
|---|---|---|---|
| AIChE Journal | AICJ | 29 | 6 |
| Chemical Engineering and Processing: Process Intensification | CEaPPI | 9 | 4 |
| Chemical Engineering Journal | ChEJ | 7 | 3 |
| Chemical Engineering Research and Design | CERaD | 13 | 9 |
| Chemical Engineering Science | ChES | 19 | 9 |
| Crystal Growth & Design | CG&D | 41 | 26 |
| Industrial & Engineering Chemistry Research | I&ECR | 28 | 37 |
| Journal of Crystal Growth | JoCG | 35 | 13 |
| Organic Process Research & Development | OPR&D | 10 | 5 |

**Table 10. Contingency table for Journal and Crystallization Method based on papers.**

| Journal | Abb | Cooling | Others |
|---|---|---|---|
| AIChE Journal | AICJ | 14 | 5 |
| Chemical Engineering and Processing: Process Intensification | CEaPPI | 5 | 3 |
| Chemical Engineering Research and Design | CERaD | 8 | 6 |

| | | | |
|---|---|---|---|
| Chemical Engineering Science | ChES | 12 | 7 |
| Crystal Growth & Design | CG&D | 26 | 12 |
| Industrial & Engineering Chemistry Research | I&ECR | 19 | 17 |
| Journal of Crystal Growth | JoCG | 18 | 6 |
| Organic Process Research & Development | OPR&D | 7 | 2 |

**Journal bias caused by crystallization method**

- **Analysis by number of entries**

Due to the number of available data points for non-cooling crystallization not being enough to get a reliable conclusion regarding the association of the variables, the analysis was carried out combining the methods different from cooling crystallization into one category. This analysis was done for both entries and papers. Additionally, CrystEngComm was excluded of this analysis since the number of papers was small and had many entries. Number of entries or papers per journal are summarized in the ESI.

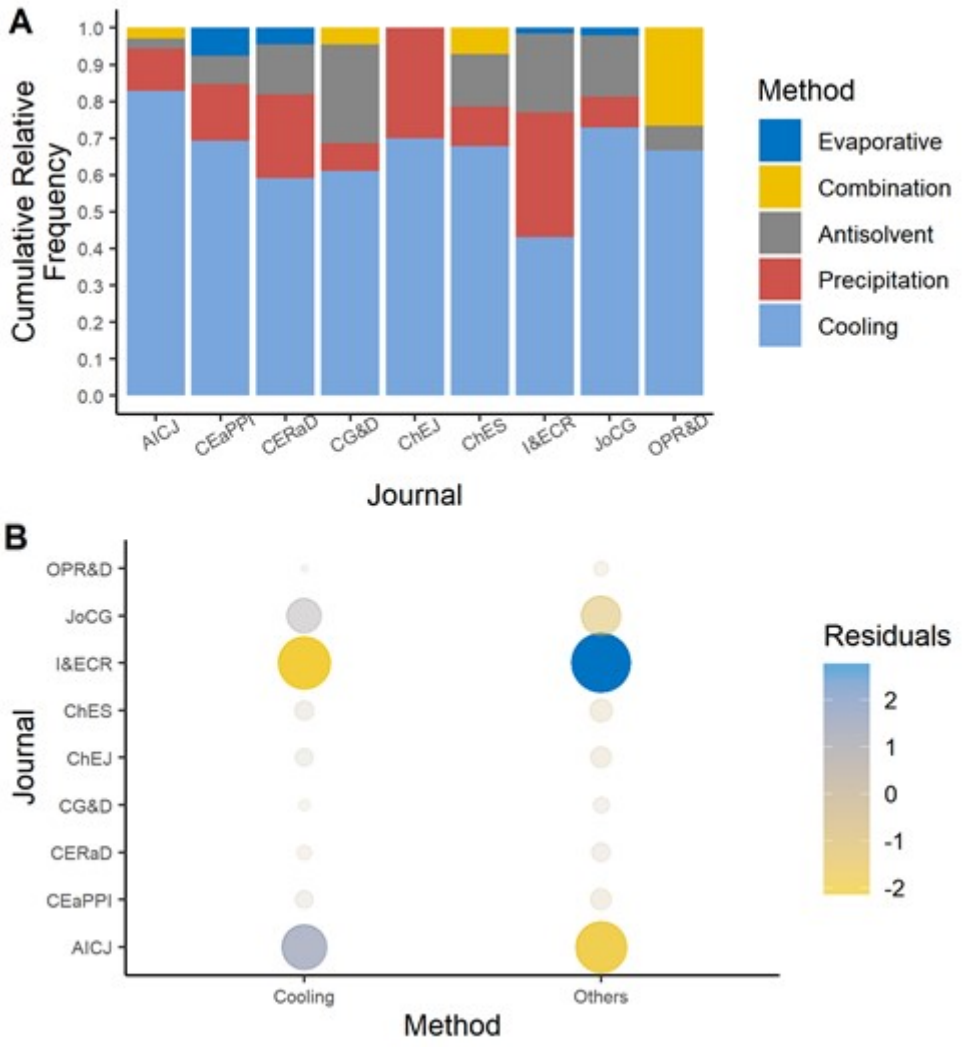



**Figure 1.** A Proportion of entries for each journal. B Residuals of Chi-square test of Journal – Method (entries).

As can be seen in **Figure 1** there are differences in each group which suggests that the journals may have a tendency to have more or fewer entries of particular crystallization techniques. By performing the independency test, the association between journal and technique is confirmed (Chi-square test, p-value = 0.01, df = 8). However, upon revising the residuals closely in **Figure 1B**, it is possible to observe there is just a journal - I&ECR - which contributes significantly to the dependency of the crystallization technique – residual higher than 2.[2] This journal is the only one in which the majority of data points corresponds to non-cooling methods, whereas cooling is predominant in the other journals. On the other hand,

given the residuals in the others journal are rather low, the observed differences may be random rather than a bias of the journals towards a specific method. Thus, the variations might have happened by chance excluding the Industrial & Engineering Chemistry Research Journal, which favour the obtention of data related to alternative techniques to cooling crystallization.

- **Analysis by number of papers**

While the analysis by journal yields similar results to the previous one when comparing figure 1 and 3, the test of independence shows that both variables are independent (Chi-square test, p-value = 0.62, df = 7) in this case. Overall, the residuals in this analysis are smaller than those in the evaluation by entries comparing **Figure 1B** and **Figure 2B**. This fact leads to conclude that all the differences in all the journals occurs by a random variation and it is not possible to establish the bias caused by a journal based on the available data.

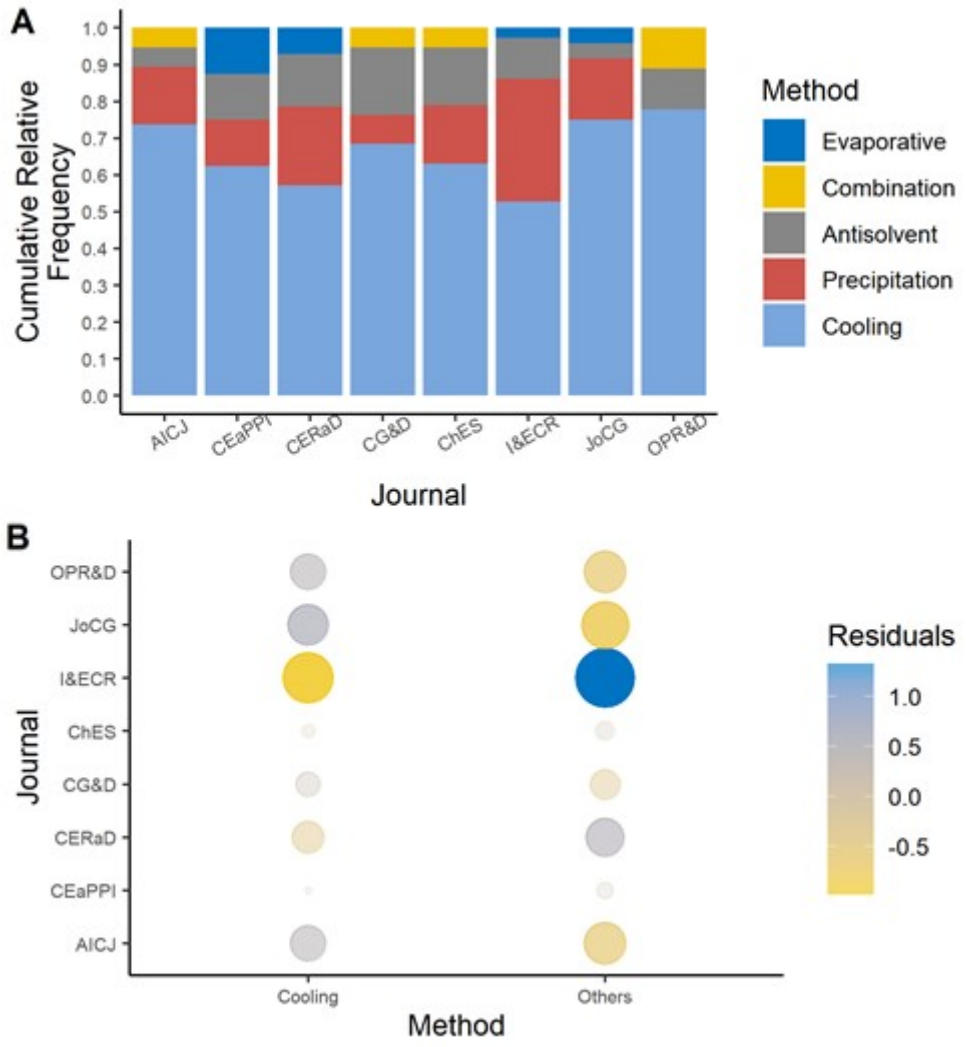



**Figure 2** A Proportion of entries for each journal. B Residuals of Chi-square test of Journal – Method (papers).

Both approaches – by-paper and by-entries – showed different conclusions. This outcome may result from the number of data points that a paper can provide. As seen previously, cooling crystallization is predominant in all the journals and papers. Upon going through the database in more detail, most of the papers that report cooling experiments provide many more data points for each paper, contrary to what happened with the papers associated with alternative methods. This means that the bias observed in the evaluation by entries might be associated to the paper or the author rather than the journal, and by analysing using papers only, this bias might be omitted which make this approach more reliable to evaluate journals bias.