Referee #1:

We are very grateful to Dr. Borduas-Dedekind for her careful reading of the manuscript and for providing useful comments and suggestions that we have used to improve the manuscript. Below we give replies to all her comments and describe the modifications implemented in the revision (highlighted in red in this response and in the annotated manuscript for review).

Reviewer: General comments:

The authors present open access Python code to estimate the subpopulations of potential ice nucleating substances from data obtained by drop freezing assays. They present codes that have the potential to be quite important in further discussing the ice-nucleating ability of ambient samples from mineral dust to organic aerosols. I command the others for this important detailed work and for their clear writing. I'd like to raise a few discussion points and point out a few minor issues to be addressed prior to publication.

I'd first like to highlight what I thought were the most important contributions within this paper.

1. Clearly articulated problem to be addressed when using frozen fraction data (for example lines 15-16, 57-58, 75-80)

2. The dilution discussion (section 3.1) is particularly valuable, and the authors can make specific recommendations for the community to move forward in their data analysis.

3. The use of the HUB-forward code to estimate the presence of subpopulations.

Authors: We thank the reviewer for the positive assessment on our work and its presentation.

Reviewer: Here are my recommendations for improvement:

I struggled a little with the chosen terminology of the code. Why use the term "HUB"? What does the "underlying-based" mean in atmospheric science and/or in statistics? The forward/backward terminology was also not intuitive to me, and it's not clear why these terms pointing to a direction were used. Could there be better terms to be used such as "subpopulation determination" for HUB-forward? For example, the term could focus on the outcome of the code?

Authors: We chose the name "HUB" for the method because is short and points to the purpose of the code, which is to connect experimental data with theoretical interpretations. The "underlying-based" words in the "HUB" acronym refer to the outcome of the code, which is the underlying distribution of heterogeneous freezing temperatures. We coined this "underlying distribution" term to distinguish it from the differential spectrum, which approximates it only when the sampling is complete. We now strive to make the notation clearer in the manuscript.

Regarding the terminology "forward" and "backward", we understand that this may not be intuitive but hopefully is well understood after reading the manuscript and/or the manual for the code. In our code, "forward" refers to the determination of the fraction of frozen droplets and the cumulative freezing spectrum from the underlying distribution of heterogeneous freezing temperatures. "Backward" refers to the determination of the differential spectrum, the approximant of the underlying distribution of heterogeneous freezing temperatures, from experimental data.

Reviewer: I'd like to challenge an assumption made in the manuscript (for example on lines 138-139) about the role of dilutions. I think the presented data analysis method is best applied to ice nucleating substances that are intact. For example, mineral dust and P. Syringae proteins. However, there is literature on organic matter and dilution series where dilutions can potentially change the shape, form and composition of ice-nucleating sites. For example, (Bogler and Borduas-Dedekind, 2020) showed that dilutions of the macromolecule lignin influence the mass-normalized ice nucleating ability of the material. I would recommend that the authors expand on the idea that this dilution method is for intact ice-nucleating ability. Alternatively, the authors could also use the open access lignin data and see how their code performs (that would be cool actually!).

Authors: We concur with the reviewer and appreciate her suggestion. We now clarify this in the introduction of the method, page 6, lines 162-167:

"In our model, we assume that the underlying distribution of ice nucleating temperatures $P_u(T)$ does not change with the concentration of INs. This last condition is violated when IN are involved in chemical, aggregation, or solubility equilibria that alter the proportionality between their concentration and the dilution factor of the sample, resulting in a lack of overlap of the pieces of the cumulative spectra $N_m(T)$ obtained from different dilutions (Bogler and Borduas-Dedekind, 2020)."

We refer again to that the lignin data to section 3.3, lines 491-492:

"Supp. Section S5 illustrates this approach for the analysis of droplet freezing data for a sample of lignin (Bogler and Borduas-Dedekind, 2020) in which the IN participate in aggregation equilibria."

We add a new section S5 to the supporting section, which presents an analysis of different concentrations of the lignin paper, The analysis suggests that all concentrations correspond to the same IN, in concentrations that are not proportional to the dilution factor, supporting the interpretation of Bogler and Borduas-Dedekind 2020 of an aggregation equilibria that sequesters active IN as the amount of lignin material in solution is increased. We copy here the text and figures added to the SI:

"S5 Modeling the differential spectrum of systems with chemical or phase equilibria using the HUB-backward code

In the derivation of the HUB method, we have assumed that the average number l of IN per droplet is proportional to the dilution of the sample. The IN, however, can be involved in chemical or phase equilibria that would impact the proportionality between l and dilution. The result is a mismatch between the actual concentration of IN in solution and the total mass concentration of the sample. Ice nucleation of Lignin provides such an example (Bogler and Borduas-Dedekind, 2020). This data set combines two challenges. First, that the nucleation from the background water used to prepare the samples produces fraction of ice signal that is highly overlapped with that of the samples themselves (**Fig. S2-A**). This is a characteristic shared by all poor ice nucleants. Second, that the processing of the fraction of ice curves using Vali's equation (**Eq. 1.a** of the manuscript) result in cumulative spectra that do not overlap (**Fig. S2-B**). This implies that the concentration of IN is not proportional to the dilution (in this case, given by the total mass of organic carbon in the sample). In this example, the $N_m(T)$ curves seem to be parallel, suggesting that the nature of the IN is preserved across concentrations, but there is an aggregation equilibrium that makes its concentration increase in a sublinear manner with sample concentration. **Fig. S2-C** shows the differential spectrum n_m obtained with HUB-backward from the fraction of frozen droplets of each concentration; these fits are shown with colored lines in **Fig. S2-A** and faithfully represent each of the $f_{ice}(T)$ curves. As expected from extreme value statistics with incomplete sampling,

the n_m(T) depends on the concentration (see also **Fig 2C** of the main text). The way the peaks move seems to be consistent with the analysis of extreme value statistics, but it is necessary to remove the background in order to do a better analysis.
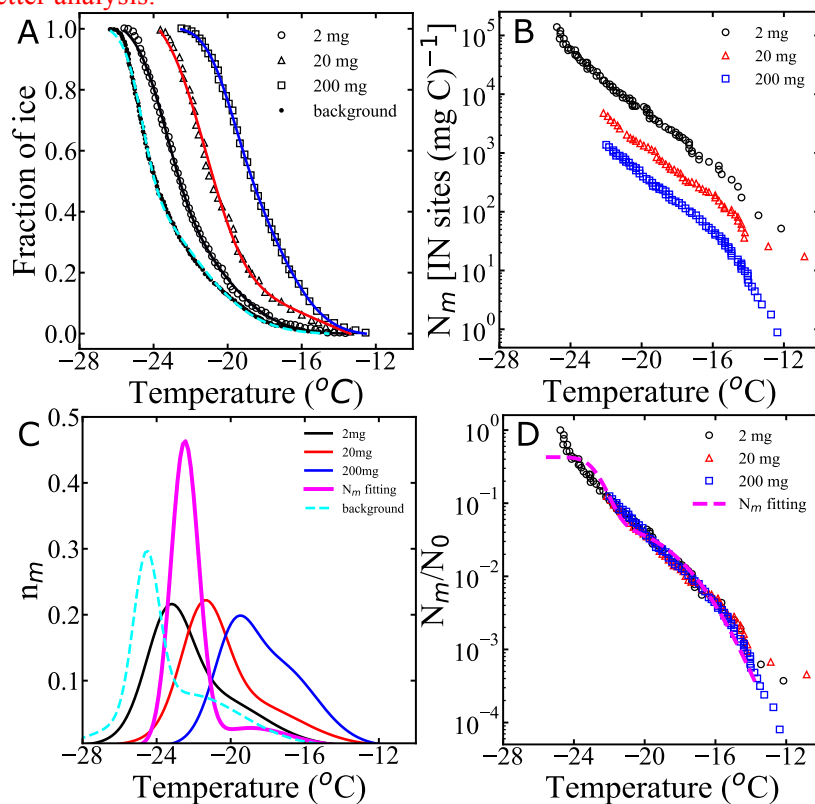
Figure S3: A) Fraction of ice for different concentrations of Lignin (Bogler and Borduas-Dedekind, 2020). Continuous lines represent the fitting of the fraction of frozen droplets obtained with the HUB-backward code using two subpopulations and Gaussian distributions as working basis. B) Cumulative freezing spectrum N_m obtained from Ref. (Bogler and Borduas-Dedekind, 2020). C) The differential freezing spectrum n_m obtained from the fitting shown by continuous lines in A). D) Points represent scaled cumulative freezing spectrum, and magenta dashed line is the fitting of the cumulative spectrum using HUB-backward.
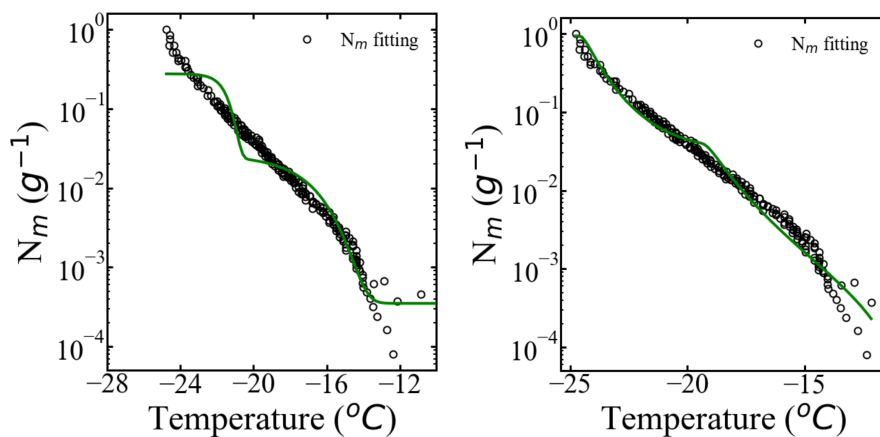
Figure S4: Fitting of Nm using HUB-backward with two subpopulations with a left-tail Gumbel (left), and log-normal (right).

Table S6: Mean relative error (MRE) and parameters of the differential freezing spectrum obtained using the HUB-backward code from lignin data at various concentrations (Bogler and Borduas-Dedekind, 2020).

	MSE	$T_{mode,1}(^o C)$	s_1	$T_{mode,2}(^o C)$	s_2	c_2
2 mg	2×10^{-4}	-21.05	2.68	-23.33	1.20	0.49
20 mg	2×10^{-4}	-21.46	1.20	-19.24	2.85	0.48
200 mg	2×10^{-4}	-19.84	1.11	-17.51	2.09	0.65
background	2×10^{-4}	-24.55	0.70	-22.40	2.95	0.59
N_m fitting (Gaussian)	0.010	-22.48	0.74	-18.93	2.10	0.14
N_m fitting (log-normal)	0.015	-24.64	0.79	-19.42	0.77	0.05

Reviewer: I also wonder about the choice of Gaussian distributions (Eq3) for the freezing temperatures of populations of IN. Why not log normal? Lines 123-124 mention that other types of normalized distributions could be used, so it would be important to justify this choice. From my own understanding, ambient samples/datasets are typically log normal. See also (Andersson, 2021).

Authors: We selected the Gaussian distribution (or normal distribution) because it is a widely used and well-understood distribution that describes many real-world phenomena. Additionally, it has desirable mathematical properties, such as being symmetric and having a defined mean and variance, which can simplify the analysis of the data. Additionally, we had previously tested the data with left-tailed Gumbel distribution and found it inadequate to represent the differential spectrum. Both Gumbel with left tail and log-normal are asymmetric distributions, with the tail pointing in opposite directions. These distributions have a lower decay towards lower temperatures and higher temperatures, respectively.

Following the suggestion of the reviewer we now add the (right-tailed) log-normal, left-tailed Gumbel distributions to the HUB-forward and HUB-backward codes. We add the following additional text to the manuscript, after introducing the normal (Gaussian) form of the populations, page 6, lines 160-163

"We also provide in the HUB code the option for the user to use the log-normal distribution, which has a tail towards higher temperatures, or the left-tailed Gumbel distribution, which has a tail towards lower temperatures. In our model, we assume that the underlying distribution of ice nucleating temperatures $P_u(T)$ does not change with the concentration of INs."

We updated the code in GitHub with these new options, that are offered to the user in the interactive version of the code.

We further tested the log-normal as a basis in the HUB-backward code to fit the fungi and bacteria data (magenta line below shows the analysis with log-normal). In these examples, the log-normal distribution does not provide a better fit to the data. Likewise, the reviewer can see in the SI introduced in response to the previous question that the log-normal distribution performs slightly worse than the Gaussian in representing the data.
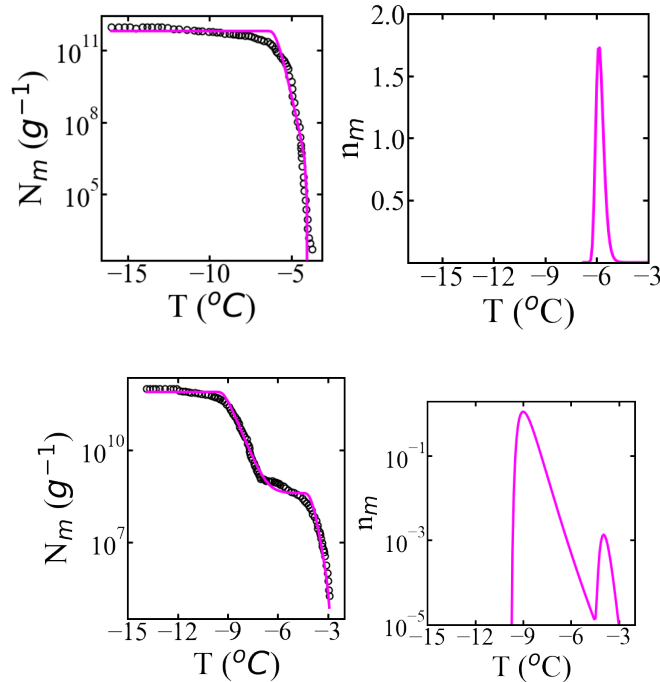
Reviewer: The manuscript is well written and well-motivated. The flow could be improved with more subsections to be able to find the information rapidly for the future reader. For instance, after reading the paragraph at lines 178-186 – I would have been interested to see this code applied in the following section. There could also be a Method section for the details of the math and then a Results and Discussion section with subsections for categories related to recommendations like dilutions series, subpopulations, etc. Subsections within pages 9-10-11 would also help the flow.

Authors: We hope that the last paragraph of the introduction, which details the contents of the sections, serves as the guide that the reviewer is asking for. We now clarify there that Section 2 is the methods section. That section provides already all the mathematical formulations used in the paper. We have included the methods section 2 –examples to help readers better understand the methodology, which we believe is a more effective pedagogical approach to present the algorithms. Section 3 is the results section, where we applied HUB-forward and HUB-backward to address multiple scenarios that illustrate their power and potential.

Reviewer: There are additional references that I would encourage the authors to consider, and I've added them throughout my specific comments below.

Authors: Thank you for making us aware of these studies.

Specific comments:

Reviewer: Title: The title might be improved by specifying the types of ice nuclei as well as either defining HUB or removing the acronym.

Authors: We prefer to keep the title as initially proposed: the name of the method (HUB), and then the description of what it does ("a method to model and extract the distribution of ice nucleating temperatures from drop freezing experiments"). Keeping HUB in the title makes it easier for the readers to identify the manuscript with the code, which we hope will be of interest to many in the ACP community. The

BINARY (Budke and Koop, 2015) and CHILL+ (Nguyen and Molinero, 2015) are well-known examples of using backronyms in the title of a paper to identify the code or technique.

Reviewer: Lines 32-36 has a rather random assortment of references of some drop freezing assays. I can refer the authors to a < 2021 comprehensive table of reported techniques: Table 1 in (Miller et al., 2021)

Authors: We thank the reviewer for pointing out the comprehensive list of reported techniques in the literature. We replace the long list by (page 2, line 34):

"A comprehensive report of various drop freezing techniques can be found in (Miller et al., 2021)."

Reviewer: Line 36: I would also comment that many drop freezing techniques are also used for ambient measurements with unknown concentrations and unknown surface area like sea surface samples and ambient aerosols. How would the authors use their code on these types of samples?

Authors: Our code (and Vali's formulation, in general) can be used even when the absolute concentrations or areas of the IN are not known, provided that the user knows the relative concentration of the dilution series of the parent sample. The foundation of this is apparent in eq. 1a, where the normalizing factor X could have any arbitrary units (e.g. some pollen data is presented in units of grains of pollen, and then the cumulative spectrum would be in units of IN per grain of pollen). We now clarify this in the introduction, lines 70-75:

"For soluble INs, the normalization factor is commonly defined by the mass of the ice nucleating material $X = \rho \, (V_{drop}/d)$, where ρ is the density of the initial solution, V_{drop} is the droplet volume and d is the dilution factor (Kunert et al., 2018). The IN surface area per drop, $X = A_{drop}$, is sometimes used as normalization factor for insoluble INs (e.g., dust, crystals), although it is challenging to measure the total IN surface area accurately (Knopf et al., 2020). We note that Eq. 1a can be used even when the absolute concentrations or areas of the IN are unknown, provided that the user knows the relative concentration of the dilution series derived from a parent sample."

Reviewer: Lines 42-46 discuss the role of cooling rate which is important in data evaluation. I would encourage the authors to comment and reference (Wright et al., 2013). Also relevant to the discussion on lines 440-447.

Authors: We now add to the introduction a discussion of stochastic modeling of ice nucleation data that references the work of Wright and Petters and others, and also extend the discussion of section 3.3 to highlight future research opportunities:

"Our analysis of the freezing data of cholesterol monohydrate shows that even a three-fold change in the cooling rate can have significant impact on the differential spectrum (**Fig. 11B**). As expected, the modes of the three populations move towards warmer temperatures upon decreasing the cooling rate. We note, however, that the shift of the peaks is not uniform; the middle one seems to be more sensitive to the cooling rate. Different sensitivity of the freezing rate of subpopulations to temperature has been also reported in simulations of nucleation data of minerals using the stochastic and modified singular frameworks (Herbert et al., 2014; Murray et al., 2011) The modified singular model proposes an empirical correction the relation between $f_{ice}(T)$ and $N_m(T)$ to account for the effect of the cooling rate on the shift of these quantities (Vali, 1994). That analysis could be extended to the analysis of the subpopulations of IN obtained with HUB-backward. Moreover, it would be interesting in future studies to use the rate dependence of the mode of the subpopulations to extract the steepness of the nucleation barrier with temperature using nucleation theory (Budke and Koop, 2015), and to investigate the

relationship between the cooling rate dependence of the differential spectrum obtained in the singular approximation with the interpretation of the same data modelled with the stochastic framework, such as in (Wright et al., 2013; Herbert et al., 2014).”

Reviewer: Eq1b and differential freezing spectra have been discussed previously in (Creamean et al., 2019) and so this reference should be added and discussed.”

Authors: Eq1b is the formulation presented by Vali in 1971 (we now add the citation). We now mention Creamean et al., 2019 in the paragraph that discusses the differential spectra in terms of populations, lines 110-113:

“While several studies have broadly defined populations by the range of nucleation temperatures they encompass(Turner et al., 1990; Creamean et al., 2019) or the origin of the sample (Steinke et al., 2020), there is currently no simple procedure to identify and quantify subpopulations or classes from cumulative freezing spectra $N_m(T)$.”

Reviewer: Scheme 1: “I_u” is not defined. I also think this scheme could be improved by using graphics instead of terms. In other words, the authors could show a frozen fraction graph and show the type of graphs that may be generated based on their code. (especially since different research groups use different terms, a graphical visualization would be helpful – and could also serve as a TOC graphic)

Authors: Thanks for pointing this out, the definition of “I_u” appeared only later in section 2.2. We now refer it for the first time in caption of Figure 1, and point to section 2.2. where the equation is derived:

“The intrinsic cumulative spectrum $I_u(T)$ is proportional to $\int_{T_m}^{T} P_u(T')dT'$ (section 2.2).”

We also follow the suggestion of the reviewer and update Figure 1 with figures that illustrate the type of generated graphs from our analysis:
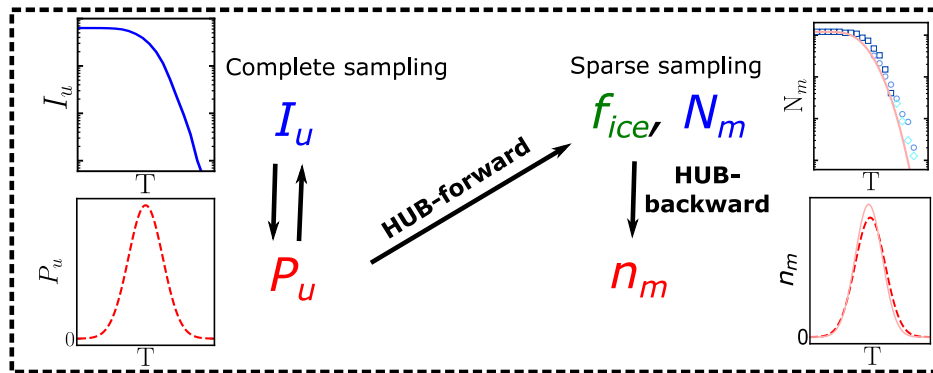
Reviewer: Lines 102-109 could be omitted entirely as these sentences are redundant (more appropriate for a thesis rather than a manuscript)

Authors: We prefer to keep the paragraph about the organization of the manuscript, to allow readers to easily find where the methods (section 2) and applications (section 3) are discussed in this study and access them according to their interests and needs.

Reviewer: The idea of Eq2 and the sum of all parts has been nicely discussed in (Steinke et al., 2020) and the authors should consider mentioning this work.

Authors: We now mention (Steinke et al., 2020) in the paragraph about populations cited above.

Reviewer: Figure 1 – PMF should also be defined in the text. It's also difficult to see the black line in figure 1. Perhaps making it bold would help?

Authors: We update Fig. 2 B-C with a magenta dashed line to enhance the contrast and add the following statement to the manuscript:

"shows the probability mass function (PMF)"

Reviewer: Would it be worth relegating the tables to the SI? Some of the values could be added directly onto the graphs for instance.

Authors: There are only two tables in the manuscript (six more are already in the Supporting Information). These tables provide the uncertainty in the fit of the cumulative spectra and important information of the populations that is relevant to the main arguments of the paper, that should be readily accessible to the reader.

Reviewer: Line 330-331 – there is much value in having code now to support this claim! Well done to the authors.

Authors: Thank you, we appreciate and share your enthusiasm for highlighting the importance of dilutions in the calculation of the freezing spectra.

Reviewer: Lines 335-343 – excellent recommendations

Authors: Thank you!

Reviewer: Figure 6 – specify in the caption the difference between panels A, B, C and D.

Authors: We thank the reviewer for pointing that out. To clarify this, we add the following statement to the caption of Fig. 6:

"Panels A, B, C and D were computed with a different number of dilutions."

Reviewer: Line 368-369: it would be worth describing how the choice of "2 subpopulations" was made. If I understood correctly, it was previously optimized? Or are the authors sourcing this information another way? It would be worth clarifying.

Authors: The choice of "2 subpopulations" was not sourced from any previous optimization but was rather determined through our analysis of the data. In our study, we performed a series of tests to determine the best fit for the data and found that a two-subpopulation model provided the best explanation of the observations. This conclusion was drawn based on a number of criteria, including the mean squared error, the goodness of the fit, and the simplicity of the model. Table 2 shows the mean squared error of each model.

We now add an extra column to Table 2 that indicates explicitly the number of subpopulations for each optimization.

Reviewer: Figure 8 – there's an error on the panel labels in the caption. ABC should be ACD.

Authors: Thank you for bringing this to our attention. We have reviewed the caption of Figure 8 and could not find the error you have described. However, we now add the panel letters and not just left and right, to the caption of figures 7 and 8 to avoid misinterpretations.

Reviewer: Line 386-387 – why were some points omitted from the optimization procedure?

Authors: The differential spectrum obtained from the sparsely sampled black data of the original submission and the total (black plus gray) data points are almost identical. We now indicate this in the manuscript, lines 464-466:

"Section S4 shows that the differential spectrum optimized from the whole data set and its sparse sampling are almost identical, because HUB-forward interpolates and smooths the input data to produce an equispaced data set.

We here copy the new section S4:

"**S4 Effect of sparse sampling of a cumulative spectrum on the estimation of the differential spectrum with HUB-backward**

Figure 9E-F presents an analysis of a sample of the pollen data of (Dreischmeier, 2019). **Supp. Fig. S2** shows that the analysis of the full data set shown **Fig. 9E** produces an almost identical differential spectrum, because HUB-backward interpolates the input data to produce a smooth and equally spaced data set (**Supp. Fig. S**3).
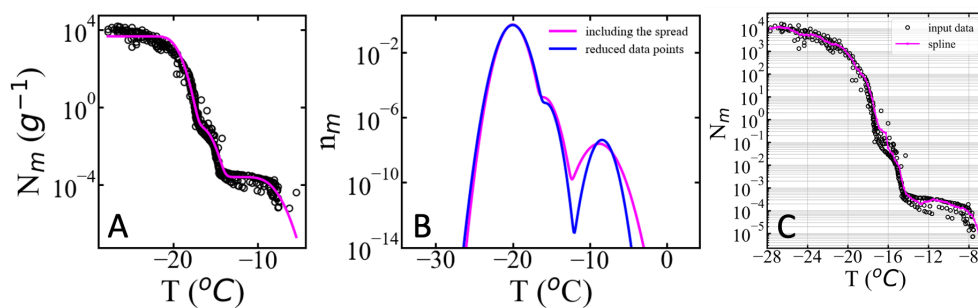
Figure S2: Effect of sparsely sampling a dense data set. A) cumulative spectra (black circles) of pollen from (Dreischmeier, 2019) and its fitting to two populations with HUB-backward fitting (magenta line). B) differential spectrum derived by that analysis from the analysis of all experimental data points (magenta line) is almost indistinguishable from the one obtained by sparsely sampling the data set (blue line, also shown in Fig 9F). C) cumulative spectra (black circles) of pollen from (Dreischmeier, 2019) used as input for HUB-backward and interpolated data using the default parameters of the code (magenta line)."

Reviewer: Figure 9 Panel A is arguably an important graph and would benefit from being highlighted separately (perhaps moving the other panels to the SI?).

Authors: We find all three examples in figure 9 important, as they illustrate different aspects of the optimization (number and relative weights of populations, noisy data…). We appreciate your recognition of the importance of the data for fungi, which is our focus in a separate study, to which we now refer in lines 426-427:

"We refer the reader to (Schwidetzky et al., 2023) for an interpretation of the size of the ice nucleating surface of *F. acuminatum* based on its differential spectrum and nucleation theory."

Reviewer: References:

Andersson, A.: Mechanisms for log normal concentration distributions in the environment, Sci Rep, 11, 16418, https://doi.org/10.1038/s41598-021-96010-6, 2021.

Bogler, S. and Borduas-Dedekind, N.: Lignin's ability to nucleate ice via immersion freezing and its stability towards physicochemical treatments and atmospheric processing, Atmospheric Chemistry and Physics, 20, 14509–14522, https://doi.org/10.5194/acp-20-14509-2020, 2020.

Creamean, J. M., Mignani, C., Bukowiecki, N., and Conen, F.: Using freezing spectra characteristics to identify ice-nucleating particle populations during the winter in the Alps, Atmospheric Chemistry and Physics, 19, 8123–8140, https://doi.org/10.5194/acp-19-8123-2019, 2019.

Miller, A. J., Brennan, K. P., Mignani, C., Wieder, J., David, R. O., and Borduas-Dedekind, N.: Development of the drop Freezing Ice Nuclei Counter (FINC), intercomparison of droplet freezing techniques, and use of soluble lignin as an atmospheric ice nucleation standard, Atmospheric Measurement Techniques, 14, 3131–3151, https://doi.org/10.5194/amt-14-3131-2021, 2021.

(Steinke et al., 2020), I., Hiranuma, N., Funk, R., Höhler, K., Tüllmann, N., Umo, N. S., Weidler, P. G., Möhler, O., and Leisner, T.: Complex plant-derived organic aerosol as ice-nucleating particles – more than the sums of their parts?, Atmospheric Chemistry and Physics, 20, 11387–11397, https://doi.org/10.5194/acp-20-11387-2020, 2020.

Wright, T. P., Petters, M. D., Hader, J. D., Morton, T., and Holder, A. L.: Minimal cooling rate dependence of ice nuclei activity in the immersion mode, Journal of Geophysical Research: Atmospheres, 118, 10,535-10,543, https://doi.org/10.1002/jgrd.50810, 2013.

Citation: https://doi.org/10.5194/egusphere-2022-1242-RC1