
Understanding Ice Crystal Habit Diversity with Self-Supervised Learning

Joseph Ko

Columbia University
New York, New York
jk473@columbia.edu

Hariprasath Govindarajan

Qualcomm Auto Ltd Sweden Filial &
Linköping University
Linköping, Sweden
hargov@qti.qualcomm.com

Fredrik Lindsten

Linköping University
Linköping, Sweden
fredrik.lindsten@liu.se

Vanessa Przybylo

University at Albany
Albany, New York
Vanessa.Przybylo@nationalgrid.com

Kara Sulia

University at Albany
Albany, New York
ksulia@albany.edu

Marcus van Lier-Walqui

Columbia University
New York, New York
mv2525@columbia.edu

Kara D. Lamb

Columbia University
New York, New York
k13231@columbia.edu

Abstract

Ice-containing clouds strongly impact climate, but they are hard to model due to ice crystal habit (i.e., shape) diversity. We use self-supervised learning (SSL) to learn latent representations of crystals from ice crystal imagery. By pre-training a vision transformer with many cloud particle images, we learn robust representations of crystal morphology, which can be used for various science-driven tasks. Our key contributions include (1) validating that our SSL approach can be used to learn meaningful representations, and (2) presenting a relevant application where we quantify ice crystal diversity with these latent representations. Our results demonstrate the power of SSL-driven representations to improve the characterization of ice crystals and subsequently constrain their role in Earth’s climate system.

1 Introduction

Clouds are one of the largest sources of uncertainty in climate models [1, 2]. They are notoriously difficult to represent accurately in models, and ice-containing clouds are especially challenging due to highly diverse properties such as crystal morphology [3]. Ice microphysical properties alter particle-radiation interactions and aerodynamics at the single-particle scale; and influence global radiative forcing, precipitation, and spatiotemporal distributions of clouds through a cascade of multiscale interactions [4, 5]. Improving our understanding of clouds is crucial, since uncertainties in future cloud behavior largely drive the overall uncertainty of future climate projections [6–8].

One important way to constrain ice microphysical properties is to take in situ measurements. For example, to understand the distribution of ice crystal habit (i.e., shape), millions of images of cloud particles have been taken on numerous airborne campaigns using cloud particle imagers (CPI) [9]. Historically, image processing techniques have been used to extract microphysically-relevant properties from CPI images [10, 11], and more recently, supervised ML has been used to improve

predictions of particle properties [12]. However, unsupervised ML has largely been underutilized in the context of analyzing CPI data and in situ microphysical observations more broadly.

To our knowledge, this is the first application of self-supervised learning to explore patterns of latent ice crystal representations. We pre-train a state-of-the-art vision transformer on a large CPI dataset to learn robust crystal representations that can support downstream science-oriented tasks. We also demonstrate an efficient pre-training pipeline that leverages existing pre-trained models and data curation. Model validation using a smaller, labeled test set confirms that the representations are encoding physically meaningful features. This work highlights the benefits of learning robust crystal representations and paves the way for more accurate, data-driven ice microphysical models.

2 Data and methods

2.1 Dataset description

The main data in this study are CPI images that come from various federally-funded airborne field campaigns. In brief, a CPI is an optical imager that takes single-channel images of cloud particles with a charge-coupled device (CCD) camera. The native CPI resolution is $2.3\mu\text{m}$, but each image was resized to a resolution of 224×224 pixels. In total, ~ 3.2 million unlabeled CPI images from across 13 field campaigns were used as the available pre-training dataset for our model (hereafter CPI-3M). To validate learned representations, we used a smaller, hand-labeled subset of $\sim 21,000$ CPI images (hereafter CPI-21K). In addition, the CPI-3M also contained habit classification labels that were predicted using a fine-tuned VGG16 convolutional neural network from Przybylo et al. [12]. Examples of CPI images are shown in Figure 1. A subset of CPI-3M had corresponding environmental data that we used for downstream analysis (see Section 3). The environmental data include measurements such as pressure, temperature, and ice water content. $\sim 524,000$ CPI images had corresponding environmental measurements (hereafter CPI-ENV-500K).

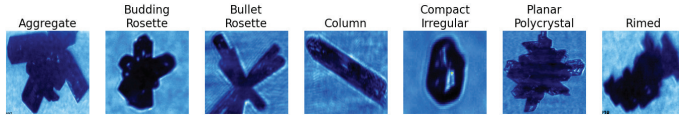


Figure 1: Examples of CPI images grouped by habit (i.e., shape).

2.2 Efficient self-supervised pre-training

Self-supervised learning (SSL) is an effective approach to learn informative representations of unlabeled data [13, 14]. Such representations enable various downstream analyses as well as efficient predictions with minimum labeling efforts [15]. CPI images exhibit natural clusters characterized by ice habits. Hence, we consider state-of-the-art, clustering-based SSL methods from the DINO family [16, 17], which are trained to assign each image to a cluster such that its augmented views, obtained by applying data augmentations, are also assigned to the same cluster. Govindarajan et al. [18] showed that the representations constitute a mixture model of von Mises-Fisher (vMF) distributions. We use the iBOT-vMF method [18] to pre-train our models and use the small Vision Transformer model architecture with a patch size of 16 (more details in A.5.1). We evaluate the learned representations with the downstream task of classifying the CPI-21K dataset.

SSL models pre-trained on ImageNet [19] are publicly available. These models transfer well to ImageNet-related domains, but their performance in entirely new domains is unclear [20]. First, we evaluate a CPI-3M pre-trained model and compare it with the best ImageNet pre-trained model (see Table 1). We also consider the recent DINOv3 [21] model that is pre-trained on a larger private dataset consisting of 1.7B naturalistic images. We observe that the ImageNet pre-trained model performs well, showing that features learned from ImageNet can transfer well to CPI images. Pre-training the DINO family models on imbalanced data is a known challenge [22, 23]. We address this limitation by following a data curation strategy proposed by Vo et al. [23]. Specifically, we curate 1.2 million images from CPI-3M through hierarchical sampling of data in the learned latent space (see A.5.2 for more details). This produces images that are more uniformly distributed in the latent space and

hence, less imbalanced. We call this new dataset CPI-H-1M. Pre-training on this $\sim 3\times$ smaller dataset results in an improved model, demonstrating the importance of data curation.

The above experiments showed that data curation is important and that ImageNet pre-trained models work reasonably well on CPI data. Also, pre-training on well-curated datasets in the target domain demonstrated potential for improved performance with better evaluation results. With this motivation, we investigated if we could pre-train models for CPI data more efficiently. Specifically, we pre-train a model on the curated CPI-H-1M dataset using iBOT-vMF for only 10 epochs and initialize the model with the ImageNet pre-trained model weights. Given a well-curated dataset like CPI-H-1M, this approach is $\sim 30\times$ more compute efficient than directly pre-training on the large CPI-3M dataset and resulted in the best performance based on validation of learned representations (see Table 1).

Table 1: Comparison of self-supervised learning models on the task of classifying the CPI-21K dataset using kNN and logistic regression classifiers. We use ViT-Small model architecture for all results and report the Top-1 accuracy metric.

SSL Method	Pre-training dataset	Pre-training epochs	Weight initialization	Top-1 Accuracy (%)	
				kNN	Logistic
DINOv3	LVD-1689M	1000	$\times$	74.83	81.83
iBOT	ImageNet	800	$\times$	78.33	82.00
iBOT-vMF	CPI-3M	100	$\times$	75.05	81.00
iBOT-vMF	CPI-H-1M	100	$\times$	77.67	83.17
iBOT-vMF	CPI-H-1M	10	$\checkmark$	81.56	84.39

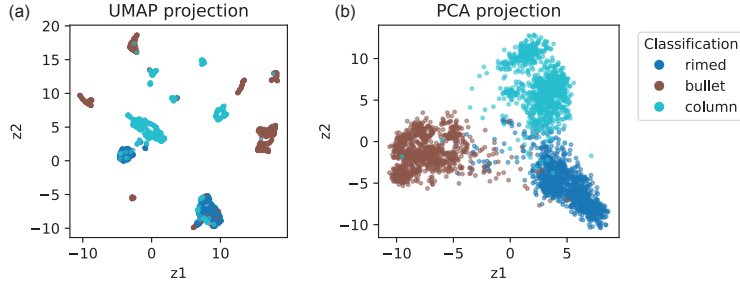


Figure 2: 2D projections of the 384-dimensional latent embeddings. A subset of 3000 samples is shown here. (a) Non-linear dimensionality reduction with UMAP. (b) Linear projection with PCA.

3 Results and discussion

3.1 Quality of learned representations

To supplement the validation results in Table 1, here we (1) confirm if clusters match expert habit labels, and (2) compare to a feature-extraction based baseline to demonstrate the benefit of SSL representations. For (1), we used dimensionality reduction to inspect clusters in 2D space. For clarity, we used a balanced subset of 3000 CPI-3M samples, and three of the seven classes to reduce visual clutter (see Appendix A.1). UMAP and PCA were used to project the 384-dimensional SSL embeddings into 2D. Figure 2 visualizes the projections, with points colored by labels predicted by a CNN from Przybylo et al. [12]. PCA reveals three distinct clusters, while UMAP forms more fragmented groupings. This aligns with the strong performance of the logistic regression on CPI-21K and suggests our model is learning approximately linearly separable morphological features without explicit guidance. For (2), we trained a baseline classifier using the CPI-21K dataset, using 13 extracted geometric features as predictors. These geometric features are derived using traditional image processing techniques, and include features such as aspect ratio, laplacian blur, and circularity, among others. Further details about these features can be found in Przybylo et al. [12]. The feature-based logistic regression performed with a top-1 accuracy of 65%, which is much lower than the 84% accuracy from our logistic regression validation using our best model (see Table 1).

3.2 Application: quantifying ice crystal diversity

After validating the learned representations, we applied them to a downstream science task: quantifying ice habit diversity in real-world clouds. Existing methods to quantify habit diversity rely on pre-designated classes and assumptions about certain morphological features, whereas SSL-driven embeddings enable a purely data-driven approach without any prior assumptions. Since our representations follow vMF distributions, the most appropriate metric to characterize diversity is the κ (i.e., “concentration”) metric [24] (details in Appendix A.3). Using CPI-ENV-500K, we analyzed how κ varies as a function of air temperature, particle size, and campaign. Figure 3a generally shows increasing habit diversity with increasing temperatures, and Figure 3b shows decreasing diversity with increasing particle size. Additionally, we see a wide spread between campaigns, highlighting the variability in crystal diversity between different cloud systems (see Appendix A.2).

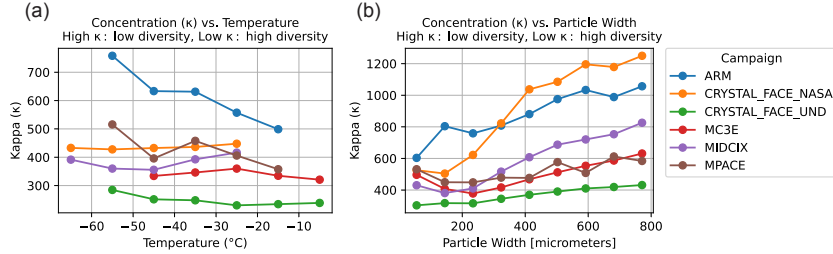


Figure 3: Crystal diversity (κ) using CPI-ENV-500K. (a) κ as a function of air temperature and stratified by campaign. (b) κ as a function of particle size (width) and stratified by campaign.

We also quantified both intra- and inter-cluster similarity. Assuming clusters loosely follow expert labels, we computed the mean cosine similarity for each predicted class, with respect to the centroid of all other predicted classes. Figure 4 shows the heatmap of the mean cosine similarity, where the diagonal describes the intra-cluster diversity, and the off-diagonal values describe the inter-cluster similarities. In other words, the heatmap shows us which habit classes show the most variability within that cluster, and also which clusters are most similar or dissimilar to each other.

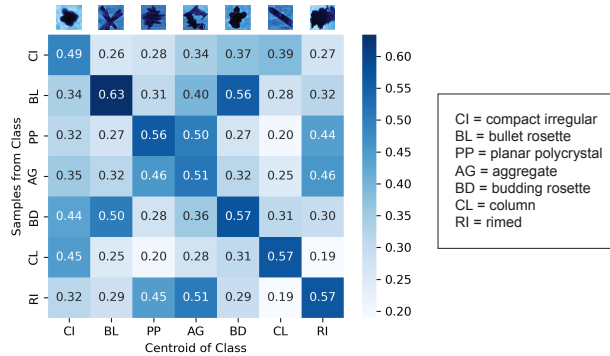


Figure 4: Cosine similarity heatmap. Intra-(diagonal) and inter-(row-wise) class similarity. Representative CPI images from each class are shown at the top.

4 Conclusion

We used the iBOT-vMF SSL vision transformer [18] to learn ice crystal representations from a large CPI dataset. Standard SSL pre-training can be computationally expensive. Through data curation and by leveraging ImageNet pre-trained weights, we outline an efficient pre-training pipeline. This resulted in robust embeddings with strong linear predictive power for downstream habit classification, validating our learned representations. These latent representations enable fully data-driven pipelines to characterize ice crystal morphology, reducing dependence on pre-defined classes and expert-driven assumptions. As a case study, we revealed systematic ice crystal diversity variations with temperature, particle size, and campaign, and quantified intra- and inter-class similarity across habit types. For future work, we will explore how learned embeddings can support anomaly detection, identifying mislabeled or rare habits, in addition to further linking microphysical properties to thermodynamic histories. This work demonstrates that SSL can effectively capture ice crystal morphology as latent representations, providing a scalable framework to reduce microphysical uncertainties and improve the representation of ice-containing clouds in climate models.

Acknowledgements

We acknowledge funding from NSF through the Learning the Earth with Artificial Intelligence and Physics (LEAP) Science and Technology Center (STC) (Award #2019625). This research was also financially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP), and the Excellence Center at Linköping–Lund in Information Technology (ELLIIT). The collaboration was initiated during the ELLIIT Focus Period on Machine Learning for Climate Science in Linköping, 2024, and we thank ELLIIT for the support during the program. Computations were enabled by the Berzelius resource at the National Supercomputer Centre, provided by the Knut and Alice Wallenberg Foundation.

References

- [1] Hugh Morrison, Marcus van Lier-Walqui, Ann M. Fridlind, Wojciech W. Grabowski, Jerry Y. Harrington, Corinna Hoose, Alexei Korolev, Matthew R. Kumjian, Jason A. Milbrandt, Hanna Pawlowska, Derek J. Posselt, Olivier P. Prat, Karly J. Reimel, Shin Ichiro Shima, Bastiaan van Diedenhoven, and Lulin Xue. Confronting the Challenge of Modeling Cloud and Precipitation Microphysics. *Journal of Advances in Modeling Earth Systems*, 12(8), 2020. ISSN 19422466. doi: 10.1029/2019MS001689.
- [2] Kara Diane Lamb, Clare E. Singer, Kaitlyn Loftus, Hugh Morrison, Margaret Powell, Joseph Ko, Jatan Buch, Arthur Z. Hu, Marcus van Lier Walqui, and Pierre Gentine. Perspectives on Systematic Cloud Microphysics Scheme Development with Machine Learning, July 2025.
- [3] Emma Järvinen, Olivier Jourdan, David Neubauer, Bin Yao, Chao Liu, Meinrat O. Andreae, Ulrike Lohmann, Manfred Wendisch, Greg M. McFarquhar, Thomas Leisner, and Martin Schnaiter. Additional global climate cooling by clouds due to ice crystal complexity. *Atmospheric Chemistry and Physics*, 18(21):15767–15781, November 2018. ISSN 16807324. doi: 10.5194/ACP-18-15767-2018.
- [4] Shin Ichiro Shima, Yousuke Sato, Akihiro Hashimoto, and Ryohei Misumi. Predicting the morphology of ice particles in deep convection using the super-droplet method: Development and evaluation of SCALE-SDM 0.2.5-2.2.0, -2.2.1, and -2.2.2. *Geoscientific Model Development*, 13(9), 2020. ISSN 19919603. doi: 10.5194/gmd-13-4107-2020.
- [5] Kamal Kant Chandrakar, Hugh Morrison, Jerry Y. Harrington, Gwenore Pokrifka, and Nathan Magee. What Controls Crystal Diversity and Microphysical Variability in Cirrus Clouds? *Geophysical Research Letters*, 51(11):e2024GL108493, 2024. ISSN 1944-8007. doi: 10.1029/2024GL108493. URL <https://onlinelibrary.wiley.com/doi/abs/10.1029/2024GL108493>. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1029/2024GL108493>.
- [6] Sandrine Bony, Robert Colman, Vladimir M Kattsov, Richard P Allan, Christopher S Bretherton, Jean-Louis Dufresne, Alex Hall, Stephane Hallegatte, Marika M Holland, William Ingram, et al. How well do we understand and evaluate climate change feedback processes? *Journal of Climate*, 19(15):3445–3482, 2006.
- [7] S. C. Sherwood, M. J. Webb, J. D. Annan, K. C. Armour, P. M. Forster, J. C. Hargreaves, G. Hegerl, S. A. Klein, K. D. Marvel, E. J. Rohling, M. Watanabe, T. Andrews, P. Braconnot, C. S. Bretherton, G. L. Foster, Z. Hausfather, A. S. von der Heydt, R. Knutti, T. Mauritsen, J. R. Norris, C. Proistosescu, M. Rugenstein, G. A. Schmidt, K. B. Tokarska, and M. D. Zelinka. An Assessment of Earth’s Climate Sensitivity Using Multiple Lines of Evidence. *Reviews of Geophysics*, 58(4):e2019RG000678, 2020. ISSN 1944-9208. doi: 10.1029/2019RG000678.
- [8] Ivy Tan, Chen Zhou, Aubert Lamy, and Catherine L. Stauffer. Moderate climate sensitivity due to opposing mixed-phase cloud feedbacks. *npj Climate and Atmospheric Science*, 8(1):86, March 2025. ISSN 2397-3722. doi: 10.1038/s41612-025-00948-7.
- [9] Brad Baker and R. Paul Lawson. Improvement in determination of ice water content from two-dimensional particle imagery. Part I: Image-to-mass relationships. *Journal of Applied Meteorology and Climatology*, 45(9), 2006. ISSN 15588424. doi: 10.1175/JAM2398.1.

- [10] G M McFarquhar, D Baumgardner, A Bansemer, S J Abel, J Crosier, J French, P Rosenberg, A Korolev, A Schwarzenboeck, D Leroy, J Um, W Wu, A J Heymsfield, C Twohy, A Detwiler, P Field, A Neumann, R Cotton, D Axisa, and J Y Dong. Processing of Ice Cloud In Situ Data Collected by Bulk Water, Scattering, and Imaging Probes: Fundamentals, Uncertainties, and Efforts toward Consistency. *Meteorological Monographs*, 58, 2017. doi: 10.1175/AMSMONOGRAPHS-D-16-0007.1.
- [11] R. P. Lawson, S. Woods, E. Jensen, E. Erfani, C. Gurganus, M. Gallagher, P. Connolly, J. White-way, A. J. Baran, P. May, A. Heymsfield, C. G. Schmitt, G. McFarquhar, J. Um, A. Protat, M. Bailey, S. Lance, A. Muehlbauer, J. Stith, A. Korolev, O. B. Toon, and M. Krämer. A Review of Ice Particle Shapes in Cirrus formed In Situ and in Anvils. *Journal of Geophysical Research: Atmospheres*, 124(17-18):10049–10090, September 2019. ISSN 21698996. doi: 10.1029/2018JD030122.
- [12] Vanessa M. Przybylo, Kara J. Sulia, Carl G. Schmitt, and Zachary J. Lebo. Classification of Cloud Particle Imagery from Aircraft Platforms Using Convolutional Neural Networks. *Journal of Atmospheric and Oceanic Technology*, 39(4):405–424, April 2022. ISSN 0739-0572. doi: 10.1175/JTECH-D-21-0094.1.
- [13] Linus Ericsson, Henry Gouk, Chen Change Loy, and Timothy M. Hospedales. Self-Supervised Representation Learning: Introduction, Advances and Challenges. *IEEE Signal Processing Magazine*, 39(3):42–62, May 2022. ISSN 1053-5888, 1558-0792. doi: 10.1109/MSP.2021.3134634.
- [14] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1920–1929, 2019. doi: 10.1109/CVPR.2019.00202.
- [15] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 456–473, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19821-2.
- [16] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021. doi: 10.1109/ICCV48922.2021.00951.
- [17] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations*, 2022.
- [18] Hariprasath Govindarajan, Per Sidén, Jacob Roll, and Fredrik Lindsten. DINO as a von mises-fisher mixture model. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=cMJo1FTwBTQ>.
- [19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.
- [20] Hasan Abed Al Kader Hammoud, Tuhin Das, Fabio Pizzati, Philip H. S. Torr, Adel Bibi, and Bernard Ghanem. On pretraining data diversity for self-supervised learning. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 54–71, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72992-8.
- [21] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

- [22] Hariprasath Govindarajan, Per Sidén, Jacob Roll, and Fredrik Lindsten. On partial prototype collapse in the dino family of self-supervised methods. In *35th British Machine Vision Conference 2024, BMVC 2024, Glasgow, UK, November 25-28, 2024*. BMVA, 2024. URL <https://papers.bmvc2024.org/0949.pdf>.
- [23] Huy V. Vo, Vasil Khalidov, Timothée Darcet, Théo Moutakanni, Nikita Smetanin, Marc Szafraniec, Hugo Touvron, Camille Couprie, Maxime Oquab, Armand Joulin, Herve Jegou, Patrick Labatut, and Piotr Bojanowski. Automatic data curation for self-supervised learning: A clustering-based approach. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.
- [24] Suvrit Sra. A short note on parameter approximation for von Mises-Fisher distributions: And a fast implementation of $I_s(\mathbf{x})$. *Computational Statistics*, 27(1):177–190, March 2012. ISSN 1613-9658. doi: 10.1007/s00180-011-0232-x.

A Appendix

A.1 Dimensionality reduction

Figure 5 shows the 2D UMAP and PCA projections including all classes, as referenced in Section 3.1. 1000 samples per class are shown here, analogous to Figure 2. Note the high degree of overlap in 2D space, which is not surprising given that 384-dimensional embeddings are being reduced to 2D. Three distinct classes were chosen out of the seven classes shown here for the main text to reduce visual clutter and for the sake of demonstration.

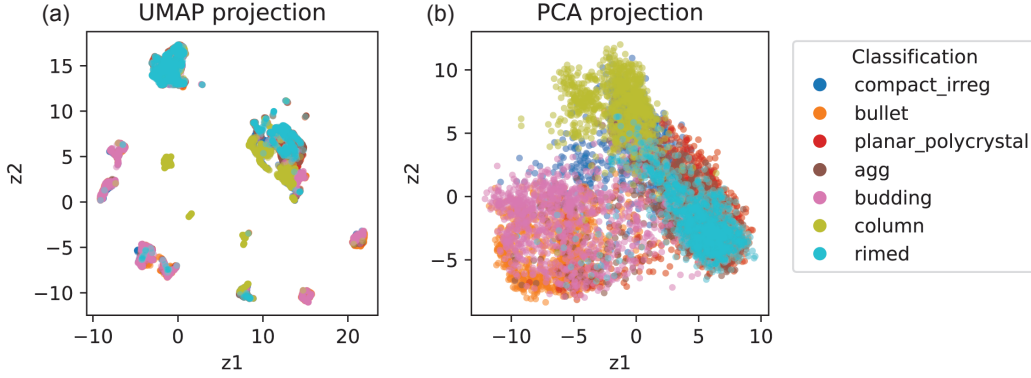


Figure 5: 2D projections of the 384-dimensional latent embeddings. 7000 samples (1000 samples per class) are shown here. (a) Non-linear dimensionality reduction with UMAP. (b) Linear projection with PCA.

A.2 Campaign details

Additional details regarding the campaigns are described here. Figure 6 shows a map indicating where the various campaigns were conducted, as well as displaying the flight tracks of the individual campaigns in more detail. Figure 7 shows the variability in conditions (e.g., temperature, ice water content, and altitude) between the different field campaigns. As mentioned in Section 3.2, we observed a wide inter-campaign range of crystal diversity. The large differences in environmental conditions for the different campaigns are highlighted here.

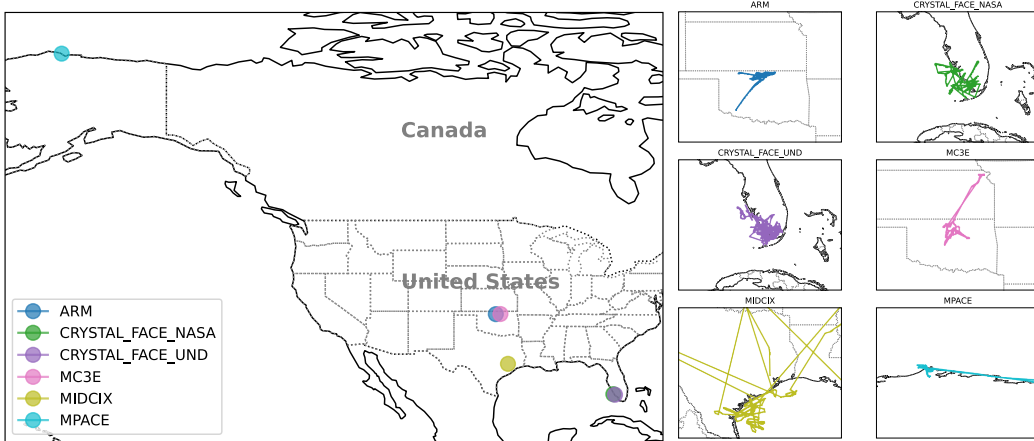


Figure 6: Maps showing the general locations (left) of the different field campaigns mentioned in Section 3.2 as well as the respective zoomed-in views (right) of the flight tracks.

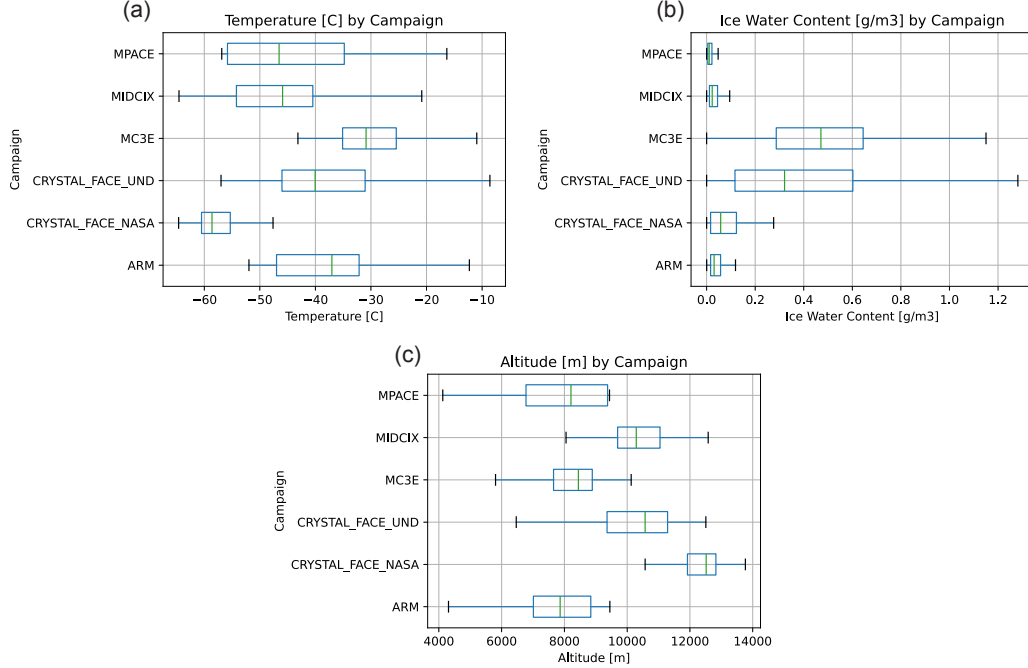


Figure 7: Distributions of (a) air temperature, (b) ice water content, and (c) altitude for the different field campaigns represented as box-and-whisker plots.

A.3 Metrics details

For a random p -dimensional unit vector $\mathbf{x}_i$, the von Mises-Fisher (vMF) probability distribution is given by $f(\mathbf{x}_i; \boldsymbol{\mu}, \kappa) = C_p(\kappa) \exp(\kappa \boldsymbol{\mu}^T \mathbf{x}_i)$, where $\boldsymbol{\mu}$ is a mean vector with $\|\boldsymbol{\mu}\| = 1$, κ is a scalar concentration parameter that measures isotropic precision, and $C_p(\kappa)$ is a normalizing constant. A higher value of the parameter κ denotes a higher concentration of samples around the mean vector $\boldsymbol{\mu}$ and lower variance or diversity. On the other hand, a lower value of the parameter κ denotes a lower concentration and consequently a higher variance or diversity. We used the following equations to estimate κ for a set of embedding vectors [24]:

$$\hat{\kappa} = \frac{\bar{R}(p - \bar{R}^2)}{1 - \bar{R}^2} \quad \bar{R} = \frac{\left\| \sum_{i=1}^N \mathbf{x}_i \right\|}{N}$$

where $\mathbf{x}_i \in \mathbb{R}^p$ is the i -th normalized embedding vector, N is the number of samples, and p is the embedding dimensionality.

We computed cosine similarity between two embedding vectors $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^p$ as:

$$\text{cosine_similarity}(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{\|\mathbf{x}_i\| \|\mathbf{x}_j\|}$$

A.4 Datasets

The CPI-3M dataset is a raw uncurated dataset obtained by combining images from all airborne field campaigns. The CPI-H-1M is curated in an unsupervised manner from the CPI-3M dataset, as detailed in the Section 2.2 and Appendix A.5.2. A subset of data from CPI-3M dataset contained environmental data and this subset is denoted as CPI-ENV-500K. CPI-21K is a smaller dataset where the ice habit classes are hand-labeled. This dataset is further split into training and test splits consisting of 19,737 and 1,800 images respectively. The test split is balanced and contains equal number of images per class, that is, 200 images in each of the nine classes.

Table 2: CPI Image Datasets. Pseudo-labeled means labels were predicted using a supervised CNN described in Przybylo et al. [12].

Dataset name	# images	Labels	Label type	Environmental data	Data curation
CPI-3M	3,200,351	✓	Pseudo-labeled using a CNN	✗	✗
CPI-H-1M	1,200,000	✓	Pseudo-labeled using a CNN	✗	✓
CPI-ENV-500K	524,000	✓	Pseudo-labeled using a CNN	✓	✗
CPI-21K	21,537	✓	Hand-labeled	✗	✗

A.5 Implementation details

A.5.1 Self-supervised Pre-training

Method overview: We use a clustering-based self-supervised pre-training method from the DINO family, known as iBOT [17]. This uses a teacher-student self-distillation setup where both the teacher and the student have the same model architecture and are initialized with random weights. Given images x from an unlabeled image dataset, we obtain randomly augmented views of the image, $x_s = A_s(x)$ and $x_t = A_t(x)$. Here, A_s and A_t are random augmentations specific to the student and teacher models. The student and teacher networks consist of a backbone model (typically a Vision Transformer) and a prediction head. The backbone model is used for different downstream tasks but the prediction head (typically an MLP) is only used during pre-training and then discarded. The student and teacher models output probability distributions over K pseudo-classes or clusters. The student model is trained to match the cluster assignment of the teacher, using a cross-entropy loss between the student and teacher outputs. The student model weights are updated using the loss gradients and the teacher model is updated using an exponential moving average of the student model weights. For a more detailed description and motivation behind this training methodology, we refer the reader to the original papers of DINO [16], iBOT [17] and DINO-vMF [18].

Implementation: We use the public codebase of iBOT ¹ and use the vMF normalized formulation proposed in Govindarajan et al. [18]. We modified the image augmentations used during the pre-training based on their suitability to CPI images. Since, CPI images are monochromatic, we remove the jitter to the image saturation and hue. In addition to the random horizontal flip, we added a random vertical flip with a probability of 0.5, as the crystals can be freely rotated in space. In the existing random resized crop augmentation, we reduced the change in aspect ratio of the crop by setting the new aspect ratio to be in the range of (0.9, 1.1). The ice crystals contain spikes of varying thickness which is an important distinguishing feature of the crystal and we want this information to be preserved in the learned representations. We use this modified set of augmentations for all the pre-training experiments that we conducted. Other hyperparameter settings for both the standard pre-training setup and the shorter and more efficient pre-training setup are provided in Table 3. In the efficient pre-training setup, we initialize the model with the weights from a model pre-trained on ImageNet-1K dataset using the iBOT method [17].

A.5.2 Data curation

In this section, we provide additional details on how we curate the CPI-H-1M dataset from the larger CPI-3M dataset. Firstly, we run a hierarchical KMeans algorithm on the latent representations using a hierarchy as follows: 3.2M images $\rightarrow$ 50K clusters $\rightarrow$ 5K clusters $\rightarrow$ 1K clusters $\rightarrow$ 200 clusters. We use the latent representations obtained from the ViT model pre-trained using iBOT-vMF on CPI-3M dataset. If an existing pre-trained model (such as those trained on ImageNet) would perform reasonably well on the target dataset, then one could also consider those latent representations for this step. Then, we use hierarchical sampling where we compute the number of samples per sub-tree, starting from the coarsest level in the hierarchy. As demonstrated in Vo et al. [23], this produces uniform distribution of samples across different levels in the hierarchy. We used the code available in their public repository ².

¹<https://github.com/bytedance/ibot/>

²<https://github.com/facebookresearch/ssl-data-curation>

Table 3: Hyperparameter settings for iBOT-vMF pre-training using the standard setup and the proposed efficient setup using weights initialized from an ImageNet pre-training.

Hyperparameter	Standard iBOT-vMF	Efficient iBOT-vMF
training epochs	100	10
batch size	1024	1024
learning rate	$4e-4$	$3e-4$
warmup epochs	10	8
freeze last layer epochs	1	1
min. learning rate	$1e-6$	$1e-6$
weight decay	$0.04 \rightarrow 0.4$	$0.04 \rightarrow 0.1$
stochastic depth	0.1	0.1
gradient clip	1.0	1.0
optimizer	adamw	adamw
shared head	✓	✓
fp16	✓	✓
momentum	$0.996 \rightarrow 1.0$	$0.996 \rightarrow 1.0$
global crops	2	2
global crops scale	$[0.32, 1.0]$	$[0.32, 1.0]$
local crops	10	10
local crops scale	$[0.1, 0.32]$	$[0.1, 0.32]$
head mlp layers	3	3
head hidden dim.	1024	1024
head bottleneck dim.	64	64
norm last layer	✗	✗
num. prototypes	2048	2048
vmf normalization	✓	✓
centering	probability	probability
teacher temp.	$0.04 \rightarrow 0.07$	0.04
temp. warmup epochs	30	—
student temp.	0.1	0.1
pred. ratio	$[0.0, 0.3]$	$[0.0, 0.3]$
pred. ratio variance	$[0.0, 0.2]$	$[0.0, 0.2]$
pred. shape	block	block