

Understanding Ice Crystal Habit Diversity with Self-Supervised Learning

NeurIPS 2025 Workshop:
Tackling Climate Change with
Machine Learning

Joseph Ko¹, Hariprasath Govindarajan^{2,3}, Fredrik Lindsten³, Vanessa Przybylo⁴, Kara Sulia⁴, Marcus van Lier-Walqui¹, Kara D. Lamb¹

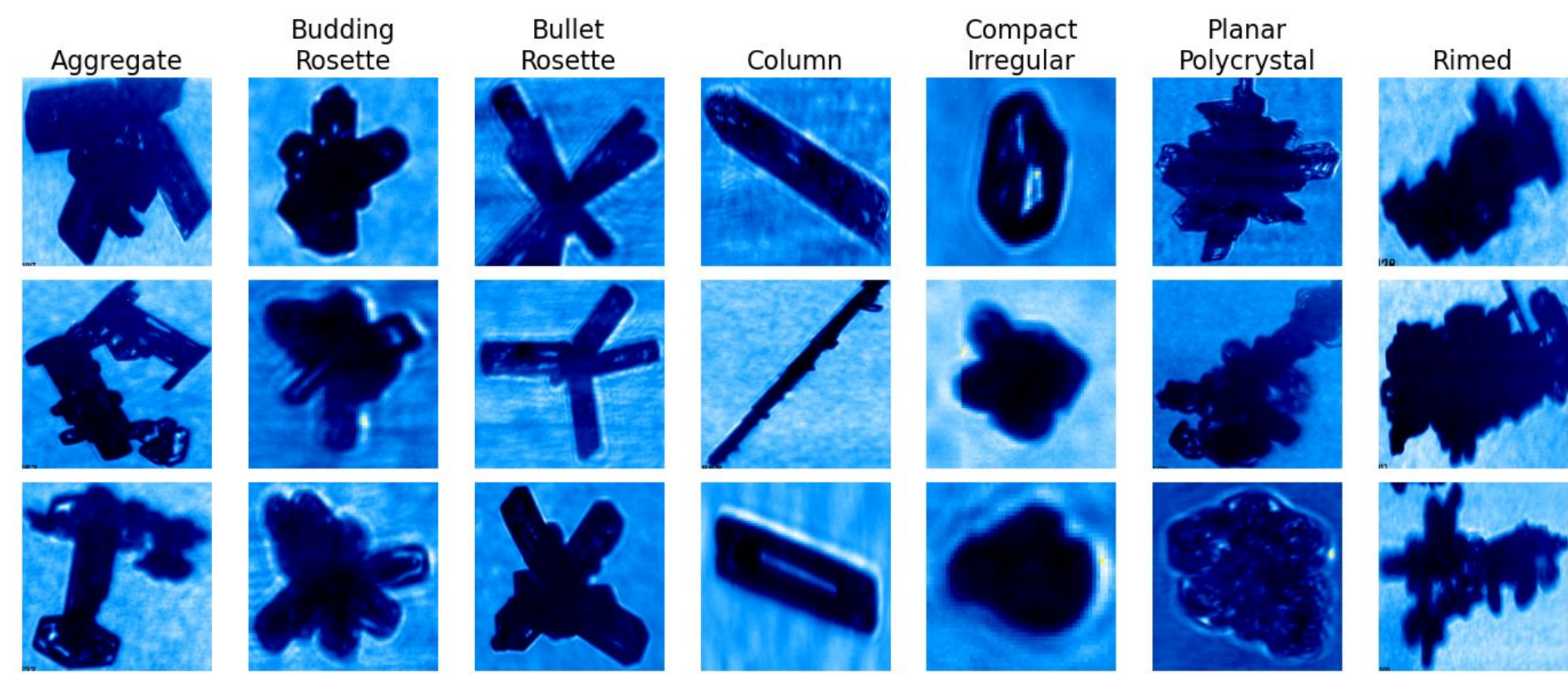
¹Columbia University, ²Qualcomm Auto Ltd Sweden Filial, ³Linköping University, ⁴University at Albany



Background

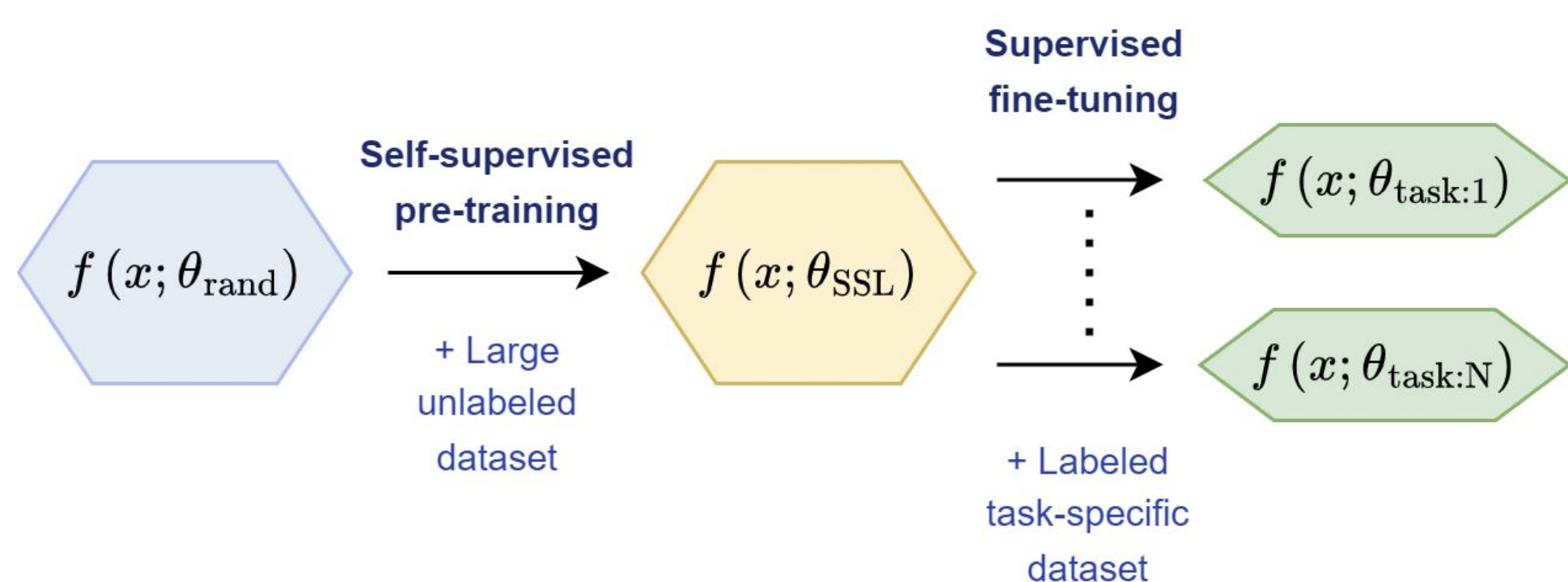


Clouds are one of the largest sources of uncertainty in climate models [1]. Cirrus clouds (see above) are especially hard to model because they are composed of ice crystals that come in many different shapes (see below). Microphysical properties of ice crystals ultimately impact global radiative forcing, precipitation, and the spatiotemporal distribution of clouds. However, the true distributions of ice crystal diversity and their relationship to thermodynamic conditions is not well understood. This work provides a roadmap for learning meaningful latent representations of ice crystal morphology and subsequent downstream applications to improve our fundamental understanding of ice-containing clouds.

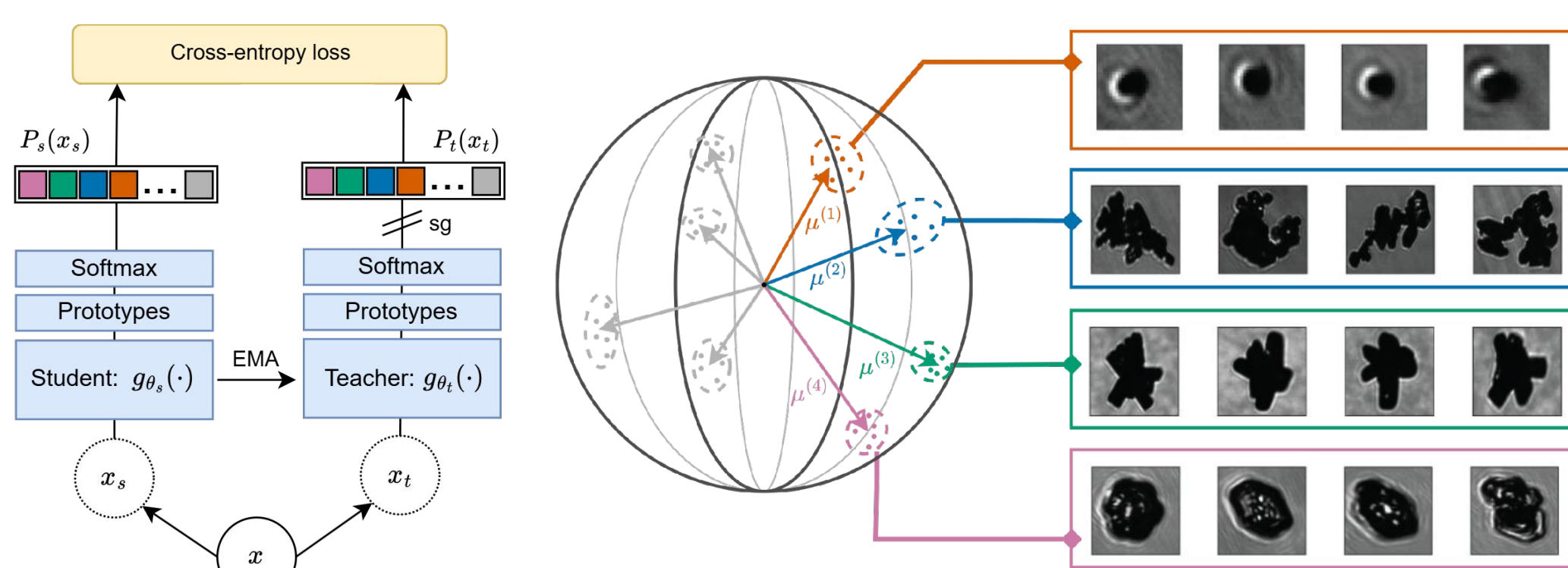


Above: Sample images from cloud particle imager (CPI) probes. Measurements were taken on various airborne field campaigns. A training set of ~3 million unlabelled images were used for this study.

Methods



Self-supervised learning (SSL) utilizes large unlabeled datasets to pre-train a model that can then be utilized for various downstream tasks.



Above: Overview of DINO-vMF method used for pre-training [2, 3]. A small vision transformer architecture is used. The learned representations (i.e., backbone embeddings) have been shown to constitute a mixture model of vMF distributions on a hypersphere.

Right: Description of the different datasets utilized in this study for pre-training, validation, and analysis.

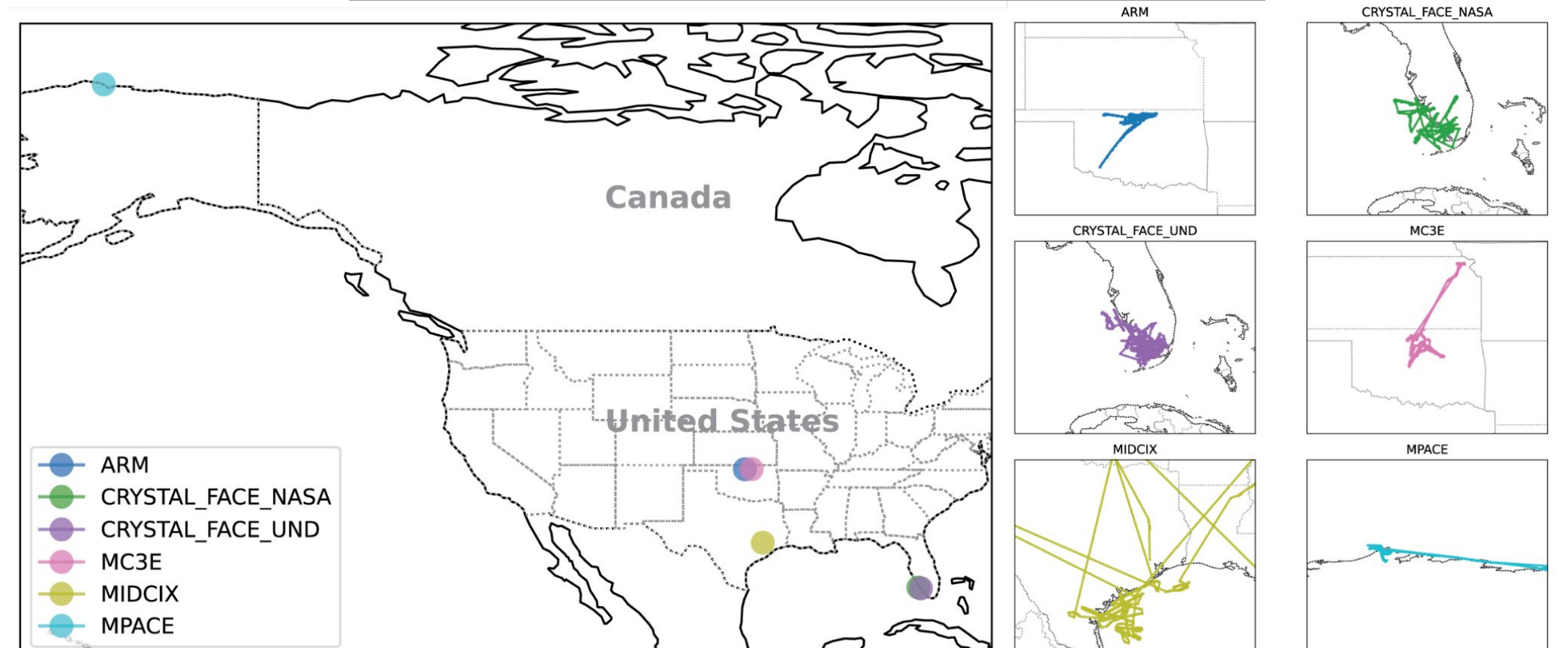
Dataset Name	Description
CPI-3M	~3 million unlabeled CPI images
CPI-21K	Hand-labeled subset of ~21K CPI images
CPI-ENV-500K	~500K CPI images w/ corresponding environmental data
CPI-H-1M	A subset of 1 million CPI images after curation (using hierarchical clustering)

Results

Table below: Model validation results. The top-1 accuracy from kNN and Logistic Regressions are used as baseline metrics to evaluate learned representations. The best model (bolded) used ImageNet pre-trained weights and subsequent pre-training on curated CPI data (CPI-H-1M).

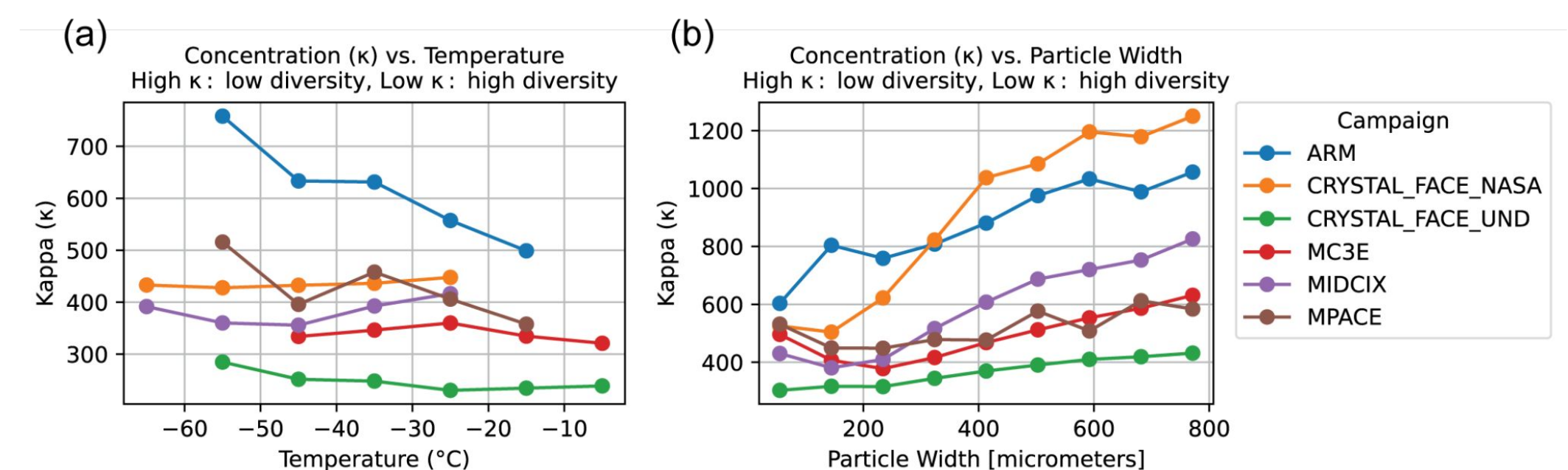
SSL Method	Pre-training dataset	Pre-training epochs	Weight initialization	Top-1 Accuracy (%)	
				kNN	Logistic
DINOv3	LVD-1689M	1000	✗	74.83	81.83
iBOT	ImageNet	800	✗	78.33	82.00
iBOT-vMF	CPI-3M	100	✗	75.05	81.00
iBOT-vMF	CPI-H-1M	100	✗	77.67	83.17
iBOT-vMF	CPI-H-1M	10	✓	81.56	84.39

Crystal Diversity Analysis



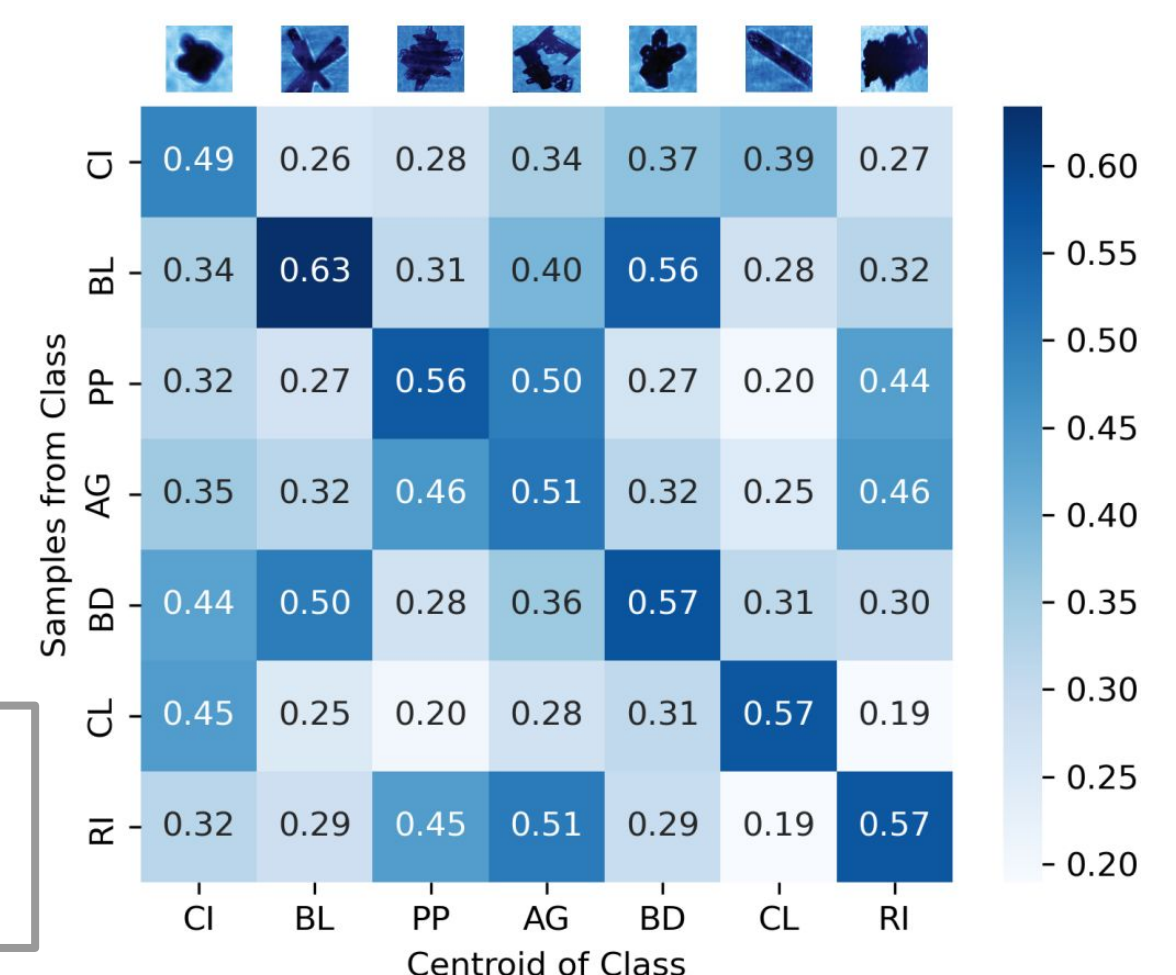
Above: Map of flight tracks for the five campaigns that constitute CPI-ENV-500K, which was used for downstream analysis of crystal diversity.

Below: Crystal diversity is quantified here with κ , which is a scalar concentration parameter for vMF distributions. (a) Diversity generally *increases* as temperature increases. (b) Diversity generally *decreases* as ice particle size increases. The line colors correspond to different campaigns.



Right: Matrix showing mean cosine similarity values for different habit clusters, using the labeled CPI-21K dataset. The matrix expresses *intra-cluster* (within cluster) diversity along the diagonal and *inter-cluster* (between different clusters) similarity on the off-axis cells.

CI = compact irregular BD = budding rosette
BL = bullet rosette CL = column
PP = planar polycrystal RI = rimed
AG = aggregate



Conclusion & Future Work

Main points:

- Robust crystal representations were efficiently learned using a self-supervised vision transformer, leveraging ImageNet pre-trained weights and data curation.
- Trends in crystal diversity were explored using learned representation embeddings. Diversity generally increased with increasing ambient temperatures and decreasing particle size
- Intra-cluster and inter-cluster similarities were also quantified using a hand-labeled subset of data, revealing (i) which classes exhibit the most in-class diversity, and (ii) which classes are most similar to each other.

Future work will include additional downstream scientific applications e.g.,

- Utilize embeddings for anomaly detection to identify rare crystals and potentially uncover new crystal formation mechanisms.
- Dive deeper into the relationship between ice morphology and environmental conditions (both instantaneous and historical).

References

- Morrison et al. Confronting the Challenge of Modeling Cloud and Precipitation Microphysics. Journal of Advances in Modeling Earth Systems, 2020
- 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.
- 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.