NEURAL INFORMATION PROCESSING SYSTEMS

Climate Change AI

NeurIPS 2025 Workshop: Tackling Climate Change with Machine Learning

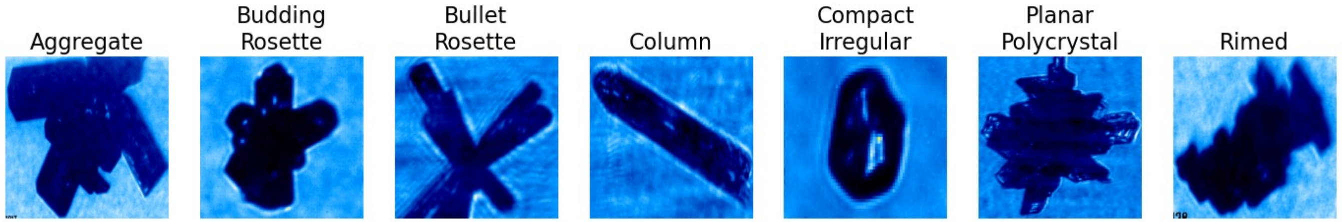
Understanding Ice Crystal Habit Diversity with Self-Supervised Learning

Joseph Ko[1], Hariprasath Govindarajan[2,3], Fredrik Lindsten[3], Vanessa Przybylo[4], Kara Sulia[4], Marcus van Lier-Walqui[1], Kara D. Lamb[1]

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

Clouds are a leading source of climate uncertainty

Below: Cirrus clouds have a wispy, feathery look.



Above: Examples of different ice crystal habits. These images are taken with an in situ imaging probe (CPI) mounted on aircrafts

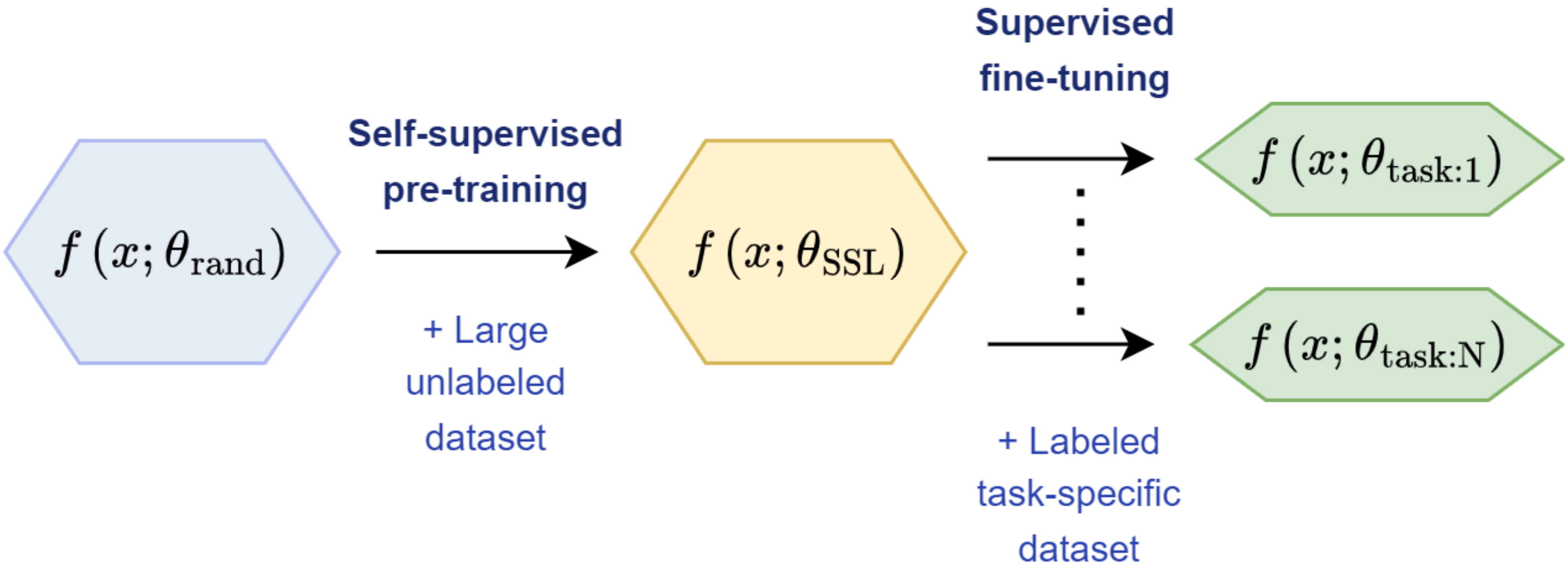
- Cirrus clouds are high-altitude clouds composed of ice crystals
- Ice crystals come in many different habits (i.e., shapes) that are difficult to represent realistically in climate models
- Ice habit impacts microphysical processes at the particle level, impacting the evolution and distribution of cirrus clouds

Given **millions** of crystal images, can we use **self-supervised learning** to learn robust **representations** of crystal morphology **without labels**?

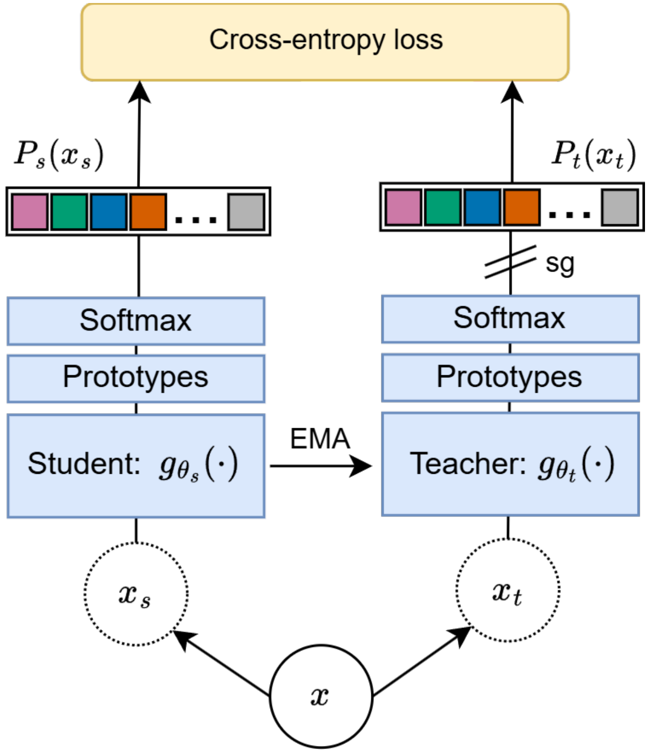
Self-supervised Learning?

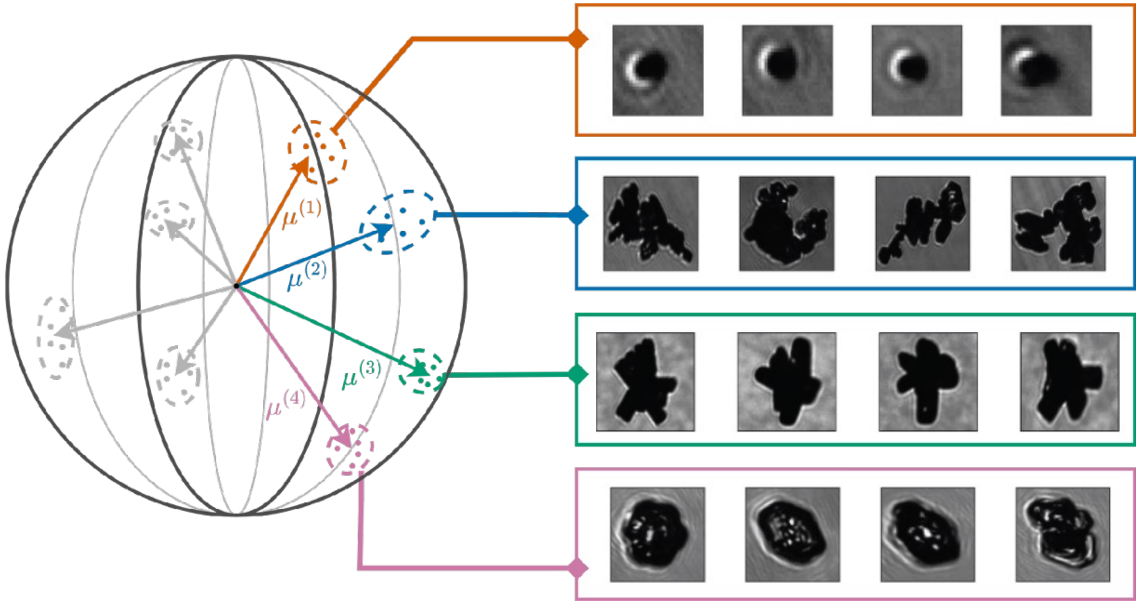
Useful downstream tasks?

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



Teacher-student self-distillation method used for pre-training





- iBOT-vMF method[1] is used for pre-training
- Small vision transformer used as backbone
- Student and teacher models output probability distributions over *K* pseudo-classes

The von Mises-Fischer (vMF) modifications[1] encourages the model to learn distributions of representations on a unit hypersphere.

[1]Details in Govindarajan et al. 2023 (ICLR): https://doi.org/10.48550/arXiv.2405.10939

Representations were validated with a smaller subset of labeled data (CPI-21K)

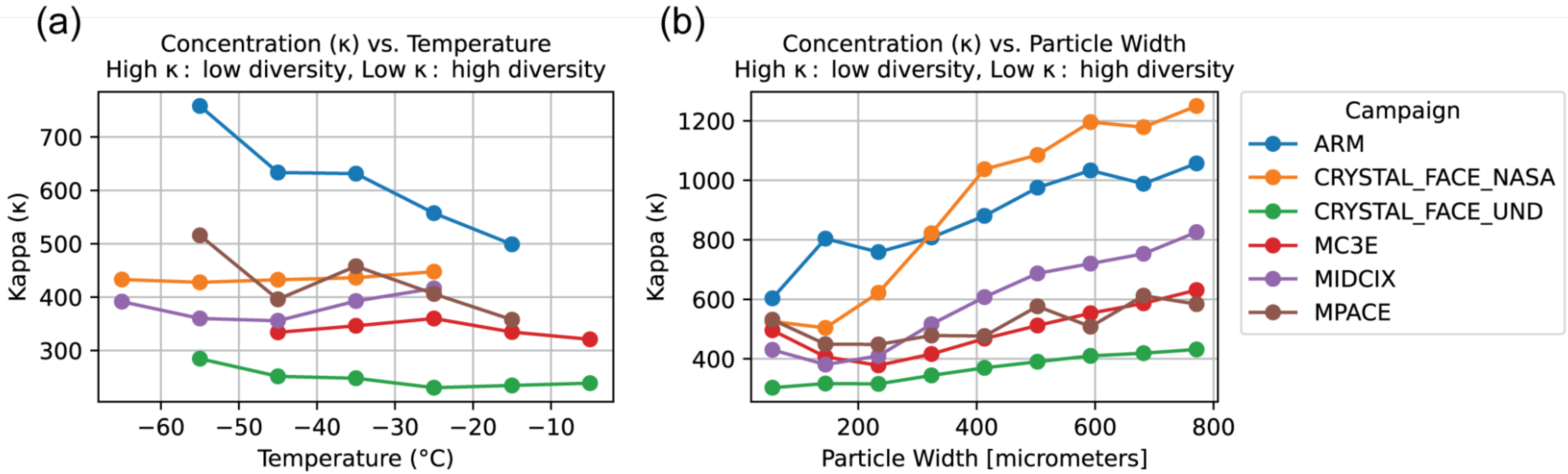
Dataset Name	Description
CPI-3M	~3 million unlabeled CPI images
CPI-21K	Hand-labeled subset of ~21K CPI images
CPI-ENV-500K	~500K CPI images w/ corresponding environmental data
CPI-H-1M	A subset of 1 million CPI images after curation (using hierarchical clustering)

Dataset description: details of various datasets used in our study

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**

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).

Learned representations can be leveraged to quantify **crystal diversity** as a function of key variables

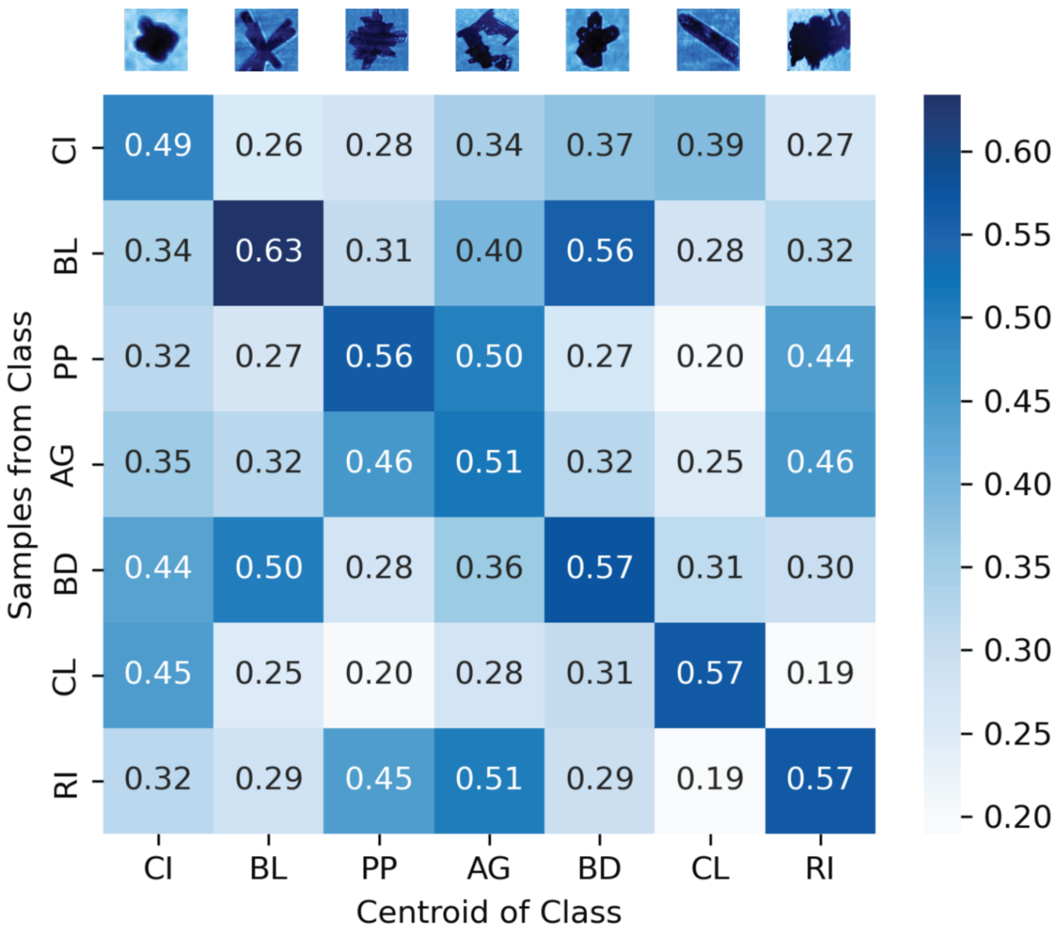


Crystal diversity is quantified with κ, a scalar concentration parameter for vMF distributions.
(a) Diversity generally *increases* as temperature increases.
(b) Diversity generally *decreases* as ice particle size increases.
The line colors correspond to different campaigns.

Representations can be used to explore similarities **within** and *between* classes

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

Conclusion & Future Work

Contact: Joseph Ko
jk4730@columbia.edu

Main points:

- Robust crystal representations were efficiently learned using a self-supervised vision transformer
- Trends in crystal diversity were explored using learned representations
- Intra-cluster and inter-cluster similarities were quantified using a hand-labeled subset of data

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)

Link to paper: https://doi.org/10.48550/arXiv.2509.07688

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).