Deep Learning Approaches for Cloud Property Retrieval: Comparing Fine-tuning with Domain-Specific Architectures

Danielle Murphy[1], Kevin Zhang[2], Caleb Parten[3],
Autumn Sterling[4], Haoxiang Zhang[5]
Xingyan Li[6], Jordan A. Caraballo-Vega[7], Jie Gong[7],
Mark Carroll[7], Jianwu Wang[6]

[1]Department of Mathematics, University of California, Berkeley
[2]Department of Computer Science, University of Maryland, College Park
[3]Department of Mathematical Sciences, Eastern New Mexico University
[4]Department of Computer Science, George Mason University
[5]Fairfax Christian School, Herndon, VA
[6]Department of Information Systems, UMBC
[7]NASA Goddard Spaceflight Center

Acknowledgments: NSF (Big Data REU Site), NASA, HPCF, NIH, CIRC, UMBC

Overview

- Accurate cloud property retrieval is critical for near real-time weather forecasting.
- Vital to understanding Earth's climate, energy balance, and hydrological cycle.
- Solution: Use various machine learning models to retrieve these properties
- Accurate retrieval algorithms for cloud properties reduce need for manual labeling of data

Remote Sensing: Satellites and Imagers

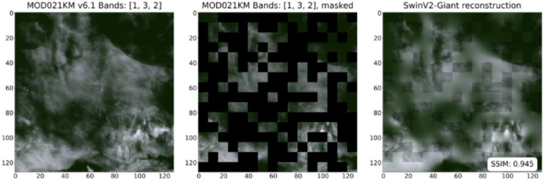
- GOES-R satellites (NOAA) use the **Advanced Baseline Imager (ABI)**.
- ABI provides:
  - 16 spectral bands
  - Higher temporal resolution than MODIS
- Moderate Resolution Imaging Spectroradiometer (MODIS) equipped on NASA's Terra and Aqua satellites
  - Total spectral bands: 36
  - 14 select bands used to train SatVision-TOA

Foundation Models in Remote Sensing

- A foundation model (FM) is a large pre-trained model that serves as a basis for downstream tasks
  - Powerful tool for remote sensing and geospatial tasks.
- Transformers: capture spatial patterns and long-range dependencies.
- Fine-tuning:
  - FM used as encoder
  - Downstream tasks use pre-trained encoder as a starting point
  - Model pipeline may look like:
    (preprocessor) $\rightarrow$ encoder $\rightarrow$ decoder $\rightarrow$ task head

SatVision-TOA

- SatVision-TOA: a foundation model pretrained on 14 MODIS bands.
- Swin-V2 architecture, trained with Masked Image Modeling
- Goal: Fine-tune SatVision-TOA using ABI's enhanced data for cloud property retrieval tasks



Why This Study?

- Most FMs are trained on high-res data (like ABI) $\rightarrow$ less frequent.
  - MODIS data is lower-res but more frequent.
- Channel mismatch: SatVision expects 14 channels
  - ABI has 16 bands
  - We explore methods of handling this mismatch
- Many studies look into segmentation
  - Benchmarking with segmentation *and* regression will help us make stronger conclusions about whether the FM's knowledge can be generalized and used for varying tasks

Cloud Properties

Cloud Mask: Cloudy or not cloudy

Cloud Phase: Clear, Liquid, Supercooled, Mixed, Ice

Cloud Optical Depth (COD):
Measure of cloud opacity (higher = more opaque)

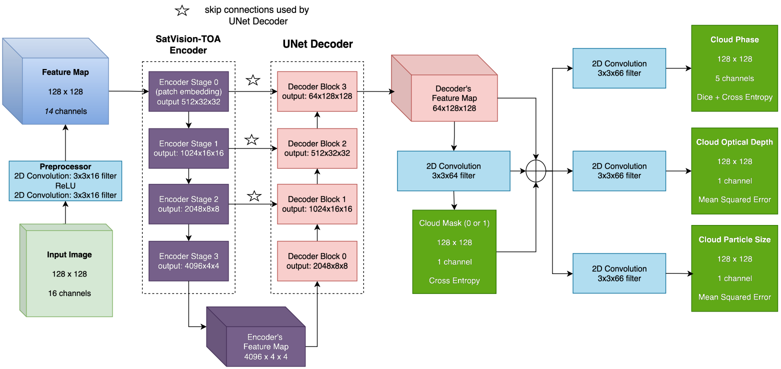
Cloud Particle Size (CPS):
Measure of average cloud particle radius

$\ln(1 + x)$ was trained and predicted for both regression tasks instead of the raw value

Tasks and Model Types

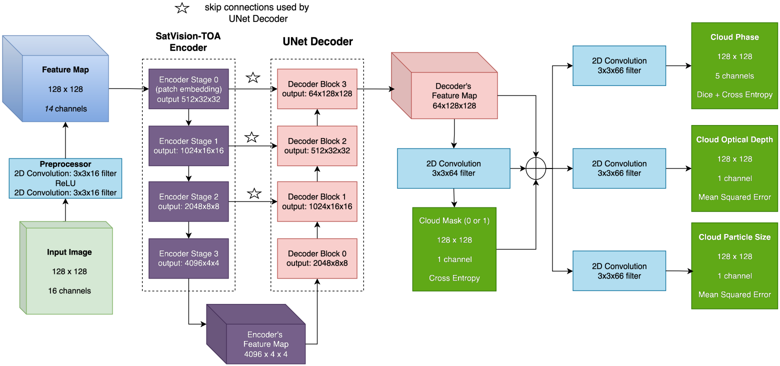
- Goal: Retrieve Level-2 cloud properties from ABI data.
- Architectures: U-Nets, DeepLab, CNNs, hierarchical classifiers
- Tasks:
  - Segmentation: Cloud mask, Cloud phase
  - Regression: Cloud optical depth, Cloud particle size
- Compare two strategies:
  1. Fine-tune foundation model (SatVision-TOA)
  2. Train models from scratch

Multi-Task Fine Tuned Model



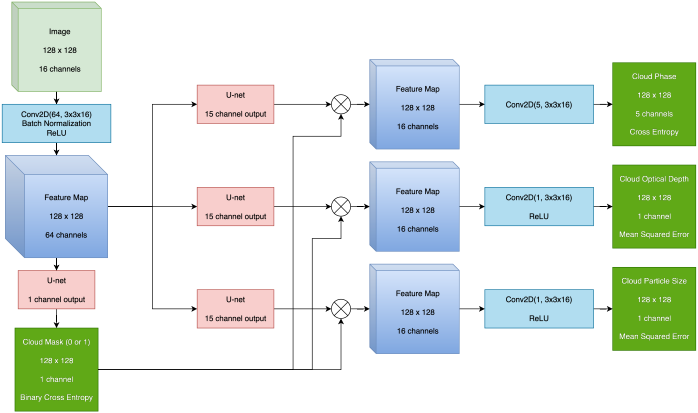
- UNet Decoder: Input downsampled as it goes through encoder stages and upsampled as it goes through the decoder stages
- Skip connections recover spatial detail lost during downsampling

Multi-Task Fine Tuned Model



- Cloud Mask prediction appended to the input of other task heads
- $\text{loss} = 2 \cdot CE_{Mask} + 1 \cdot CE_{Phase} + \frac{1}{100}(MSE_{COD} + MSE_{CPS})$

Multi-Task Model Architecture



- Cloud Mask prediction appended to the **output** of U-nets
- loss $= 1 \cdot CE_{Mask} + 1 \cdot CE_{Phase} + 2(MSE_{COD} + MSE_{CPS})$

Comparing All Models

Table: Performance of Multitask and Single-task Models on Cloud Attribute Prediction.

Model	Task	mIOU	Task	r^2	Train Time
Multitask Models					
Fine Tuned MT	Mask	0.881	COD	0.527	1:56:27
	Phase	0.627	CPS	0.605	
From Scratch MT	Mask	**0.909**	COD	**0.775**	45:59
	Phase	**0.700**	CPS	**0.786**	
Individual Models: Classification					
Fine Tuned	Mask	0.816			1:11:56
Fine Tuned	Phase	**0.713**			1:57:28
Scratch U-net	Mask	**0.896**			19:47
Scratch U-net	Phase	0.664			20:18
Individual Models: Regression					
Fine Tuned			COD	**0.754**	1:51:52
Fine Tuned			CPS	0.680	1:41:11
Scratch U-net			COD	0.717	17:07
Scratch U-net			CPS	**0.738**	17:00

- From-scratch MT model outperforms fine-tuned
- Training time is significantly shorter for from-scratch runs.

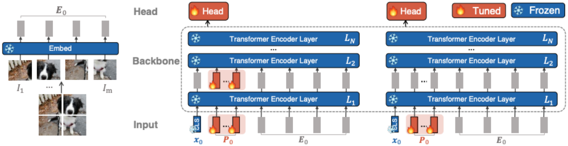
Fine-Tuning Experiments

Fine-Tuning Models and Experiments

Fine-Tuning Experiments

- Significant computation is required for fine-tuning and train times are long.
- Parameter-Efficient Fine-Tuning (PEFT) strategies aim to reduce this cost.

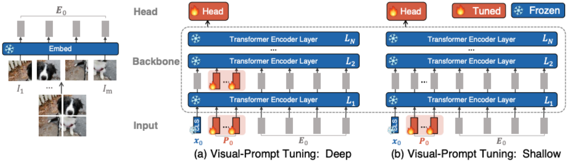
Visual Prompt Tuning



(a) Visual-Prompt Tuning: Deep (b) Visual-Prompt Tuning: Shallow

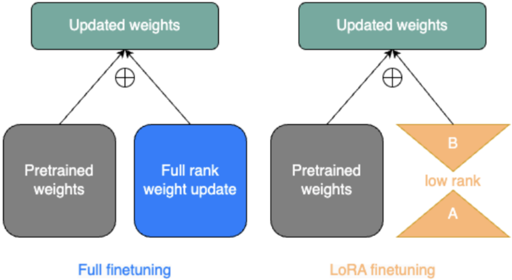
- In VPT, the inputs to the model are wrapped in learnable prompts
- During training, the entire encoder is frozen but prompts are trainable
- This allows the model to still learn but we only train a small amount of parameters

Visual Prompt Tuning



- We implemented VPT Shallow; where prompts are injected just to the first layer of the transformer
- Prompts are added element-wise to patches.
  1 prompt = 1 patch

Low Rank Adaptation



- Weight updates of the encoder approximated with low rank matrices A and B: $W = W_{frozen} + AB$
- If W is $n \times n$, A is $n \times r$ and B is $r \times n$
  - n^2 trainable parameters turns into $2 \cdot r \cdot n$

Comparing Fine Tuning Strategies

Table: Best Individual Task Performance for each Fine Tuning Strategy

Task	Hyperparams	Time to Train	mIOU/r^2
Mask			
FFT		1:45:50	0.749
LoRA	rk 32	1:11:56	0.816
VPT	300 prompts	1:01:32	0.675
Phase			
FFT		1:51:42	0.649
LoRA	rk 64	1:12:57	0.614
VPT	300 prompts	1:02:33	0.512
Optical Depth			
FFT		1:51:52	0.754
LoRA	rk 16	1:09:37	0.645
VPT	200 prompts	0:58:40	0.586
Particle Size			
FFT		1:41:11	0.680
LoRA	rk 32	1:08:10	0.664
VPT	100 prompts	0:58:55	0.574

- VPT provides the best improvement in training time
- LoRA is more balanced: decreased training time, producing competitive results

Visual Prompt Tuning: Best # of Prompts

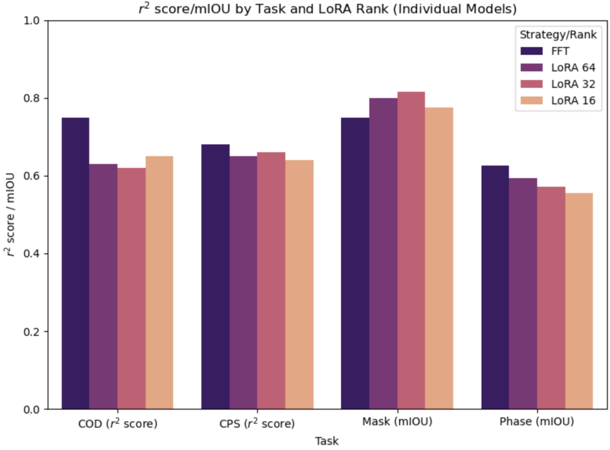
Table: Performance of fine-tuned individual models with Visual Prompt Tuning (VPT).

Task	100 Prompts	200 Prompts	300 Prompts
Classification mIOU			
Mask	0.610	0.670	0.675
Phase	0.496	0.488	0.512
Regression r^2 Score			
COD	0.512	0.586	0.551
CPS	0.574	0.520	0.508

- Classification tasks may prefer a higher number of prompts than regression

Low Rank Adaptation: Ranks

- We tried different ranks across the single-task models



Low Rank Adaptation: Multitask Training

Table: Multitask Model: Full Fine Tuning vs. LoRA rk. 16

Model	Task	mIOU	Task	r^2	Train Time
FFT	Mask	0.838	COD	0.550	1:53:54
	Phase	0.578	CPS	0.610	
LoRA	Mask	0.755	COD	0.479	1:24:32
	Phase	0.508	CPS	0.504	

- Training time decreased by about 25%
- On average, task performance decreased by 13.075%
- Most drastic change seen in the r^2 score for CPS (-17.4%).

Tuning Losses

$$CE(p_t) = -\log(p_t)$$

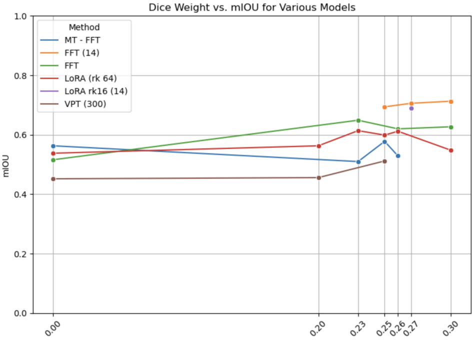
$$FL(p_t) = -(1-p_t)^{\gamma}\log(p_t)$$

$$Dice = 1 - \frac{2 \cdot TP}{(TP+FP)+(TP+FN)}$$

- With CE loss for cloud phase, models had difficulty handling edges and were adversely affected by class imbalance in the dataset
- Focal loss was helpful for cloud mask, but did not work for phase.
- Using *just* dice loss did not work → weighted sum of Dice, CE

Tuning Losses: Dice Weights

$$\text{loss} = \text{dice weight} \cdot \text{dice loss} + (1 - \text{dice weight}) \cdot \text{CE loss}$$



- Improved average recall from 0.719 (FFT, 16 bands, with just CE) to .886 (FFT, 14 bands)

Number of Bands

During much of our work, we were motivated to try to use all 16 bands. We used a preprocessor: two 2D Convolutions to go from 16 channels to 14.

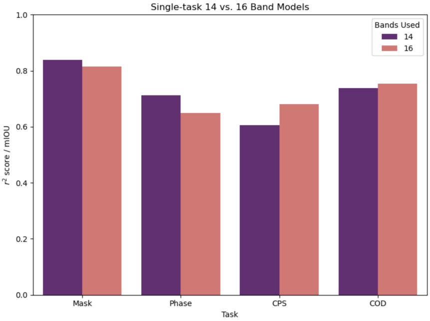
- We started trying 14 band models to see how performance changed
- We got our highest mask mIOU (from fine-tuning) from MT 14-band model

Table: 14-Band and 16-Band Multitask Models

Attribute	**14 Bands**	**16 Bands**
Dice Weight	0.30	0.23
Learning Rate	3e-4	3e-4
Mask mIOU	0.881 (+5.1%)	0.838
Phase mIOU	0.627 (+8.5%)	0.578
COD r^2	0.527 (-4.2%)	0.550
CPS r^2	0.605 (-0.7%)	0.609

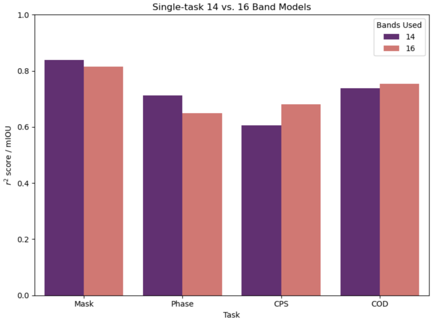
14 Band Single-Task Performance

- We used full fine tuning and LoRA for each individual task, adjusting: learning rates, dice weight, and rank (if training with LoRA)
- Overall, the 14 band models obtain comparable results to 16 band models



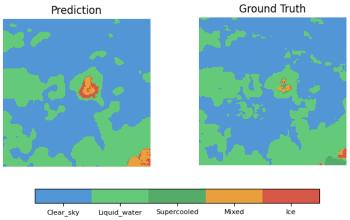
14 Band Single-Task Performance

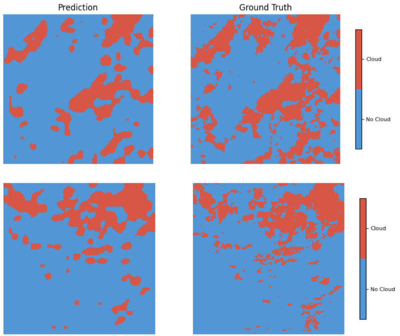
- Using the 14 "matched" to MODIS bands may take better advantage of encoder's pretrained knowledge
- Further exploration of 14 band models may improve our overall fine-tuned performance
- Preprocessor is a viable option, may be useful for future work if bands are not as easily matched?



Multitask Model Visuals

These predictions are from a 16-band multitask model, trained with LoRA.





Overall, we find were successful in our goal: working with SatVision-TOA to fine-tune meaningful cloud prediction models

From-Scratch

From-Scratch Models and Experiments

MLPs

- 3 hidden layers with ReLU activation
- Baseline before trying spatially aware models
- Trained on 160 images, batch size of 2048 pixels

Table: MLPs benchmark evaluation

Model	**Task**	**mIOU**	**Task**	**r^2**
MLP	Mask	0.823	COD	0.724
	Phase	0.578	CPS	0.640

Trees, Linear Regression, Forest, Gradient Boosting

Other algorithmic models used for both pixel-by-pixel classification and regression with Sci-kit Learn.

Table: MLPs benchmark evaluation

Model	**Task**	**mIOU**	**Task**	**r^2**
Decision Tree	Mask	**0.903**		
	Phase	**0.729**		
Linear Regression			COD	0.212
			CPS	0.299
Regression Forest			COD	0.663
			CPS	0.609
Hist Grad Boosting			COD	**0.786**
			CPS	**0.739**

Pixel-by-pixel models

Decision trees and Histogram-based Gradient Boosting outperformed MLP models

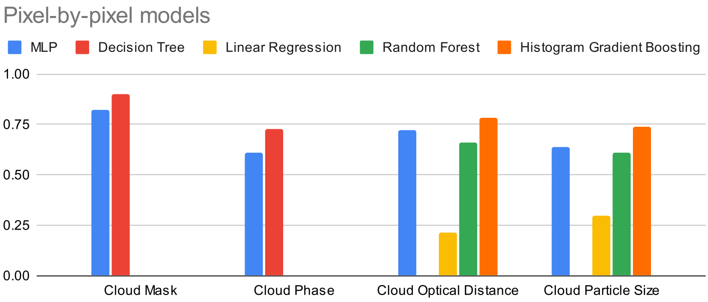


Figure: Comparing pixel-based model evaluations

Individual U-nets

- Type of Convolutional Neural Network
- Uses Resnet-34 encoder
- Skip layers capture multi-level features

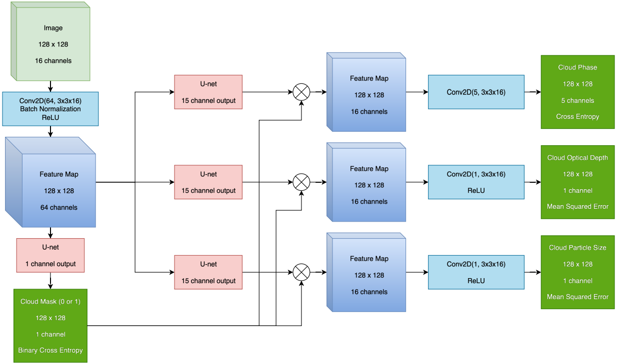
Table: Single U-net evaluation

Model	Task	mIOU	Task	r^2
U-net	Mask	**0.896**	COD	0.717
	Phase	0.664	CPS	0.738

Multi-task

- V1: Cloud **mask** output appended to **input** of other U-nets
- V2: Cloud **mask and phase** output appended to **input** of other U-nets
- V3: Encoder and decoder setup. Cloud mask appended to **output** of other U-nets
- V4: U-nets replaced with DeepLab

Multi-task Diagram



- Cloud Mask prediction appended to the **output** of U-nets
    - Batch Normalization added in encoder

Multi-task cont.

Table: Multi-task common hyper-parameters

Images	14973
Train/Validation/Test Split	80/10/10
Optimizer	Adam
Batch size	128
Learning rate	.00002
Learning rate scheduler	Patience=3, Factor=.5
Epochs	100
Loss	Unweighted sum of individual losses

Multi-task from Scrach Results

Table: Multitask evaluation

Model	Task	mIOU	Task	r^2	Train Time
V1	Mask	0.819	COD	0.740	40:16
	Phase	0.642	CPS	0.742	
V2	Mask	0.707	COD	0.719	40:48
	Phase	0.471	CPS	0.471	
V3	Mask	0.911	COD	0.767	43:07
	Phase	0.692	CPS	0.776	
V3.1	Mask	**0.915**	COD	**0.769**	44:30
	Phase	**0.696**	CPS	**0.781**	
V4	Mask	0.847	COD	0.697	48:41
	Phase	0.632	CPS	0.700	

Multi-task from Scrach Results

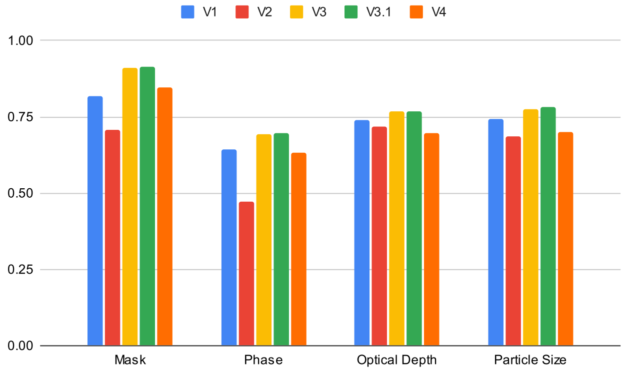


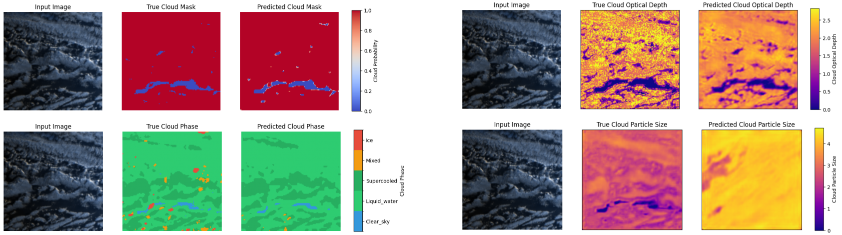
Figure: Comparing multi-task model evaluations

Multi-task Loss Weight Tuning Cont.

Table: Adjusting Loss Weights in MT V3.1

Weights	Task	mIOU	Task	r^2	Train Time
(1, 1, 1, 1)	Mask	**0.915**	COD	0.769	44:30
	Phase	0.696	CPS	0.781	
(1, 1, .5, .5)	Mask	0.866	COD	0.706	39:34
	Phase	0.648	CPS	0.716	
(1, 1, 2, 2)	Mask	0.909	COD	**0.775**	45:59
	Phase	**0.700**	CPS	**0.786**	
(2, 1, 1, 1)	Mask	0.887	COD	0.734	38:40
	Phase	0.654	CPS	0.743	

Multi-task Loss Weight Tuning Cont.



Conclusion

- SatVision-TOA performs well on both segmentation and regression tasks when fine-tuned with ABI data
  - Low rank adaptation is successful in achieving comparable results to full fine tuning while reducing training time
- Multi-task models offer efficiency and improved task results in some cases
- Comparing foundation model adaptation vs. training from scratch reveals:
  - Trade-offs in accuracy vs. training cost
  - Task-specific differences in performance

Key Insights and Products

- Knowledge from foundation models pretrained on MODIS can be transferred to ABI-based tasks despite a different number of spectral bands and resolution differences
- Future research may look further into the band mismatch problem
- Multi-task learning consolidates inference pipelines

Github: https://github.com/asterli6/big-data-reu
https://github.com/big-data-lab-umbc/big-data-reu/tree/main/2025-projects/team-1

[1] D. Murphy, K. Zhang, C. Parten, A. Sterling, H. Zhang, et al., tech. rep. HPCF–2025–4, 2025.

[2] D. Murphy, K. Zhang, C. Parten, A. Sterling, H. Zhang, et al., REU Symposium, *ICDM 2025*, 2025.