Automatic Classification of Sedimentary Particles for Insights into Past Climate & Environment

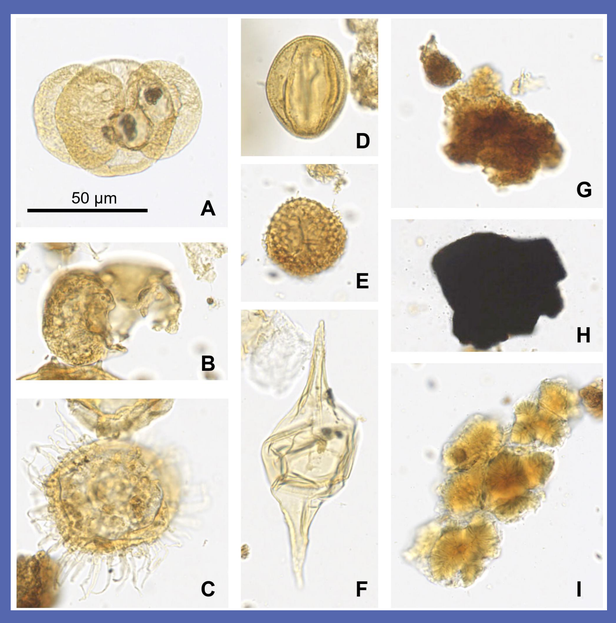
Nikolai Andrianov and Kasia K. Sliwinska
Geological Survey of Denmark and Greenland (GEUS)

AI in Science Summit 2025
EU2025DK

We explore an automated approach to classifying sedimentary particles in microscope images from samples representing different geological time periods. Our goal is to enable rapid, objective analysis of past environmental conditions using machine learning. The method combines automated particle segmentation—via thresholding and the pre-trained SAM2 model—with unsupervised clustering using the DINOv2 model. Early results show effective segmentation, though particle groupings remain mixed, underscoring the challenge of label-free classification. Future work will integrate limited labeled data to enhance model accuracy and interpretability.

Understanding the Past to Inform the Future

- Reconstructing past environments supports climate studies, resource mapping, and sustainable energy
- Approach: identify and classify microscopic organic particles (palynofacies) preserved in rocks
- Examples of particle categories are shown below



A) Well preserved bisaccate pollen; B) Large fragment of a bisaccate pollen; C) Dinocyst; D) Non saccate pollen; E) Spore; F) Dinocyst; G) Amorphous organic matter; H) Phytoclast (fragment of black wood); I) Freshwater algae (Botryococcus). Note that particles C and F belong to the same category.

Limitations of the Traditional Workflow

- Manual counting of 10,000s of particles are time-consuming and hard to reproduce
- Analysis area often undefined → limited reproducibility
- Subjective identification → analyst bias and poor comparability
- Verification or correction requires full re-analysis

Our Digital Approach

- Introduces a standardized, automated workflow for sedimentary organic particle analysis
- Goals: transparency, reproducibility, efficiency, scalability
- Enables extraction of high-quality environmental data from microscopy images

Related Work & Motivation

- Automatic particle recognition common in biology, materials science, and life sciences
- Most methods rely on supervised models that lack generalization to new sample types
- Self-supervised learning learns features from unlabeled images, reducing manual effort
- Applied to palynofacies, this enables scalable, reproducible classification and reconstruction

Workflow Overview

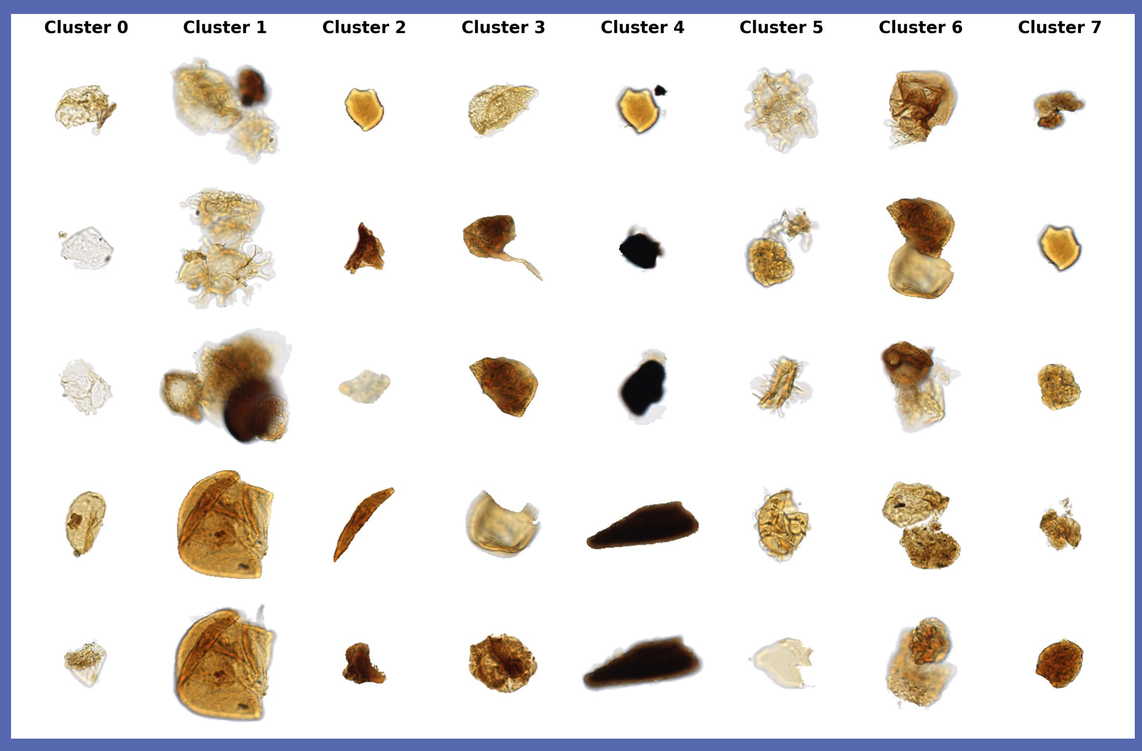
Particle Segmentation & Extraction
- Process microscopy slides containing 10,000s of particles stored in multi-focal NDPI format
- Automatic detection and isolation of individual particles from microscope images
- Grayscale conversion, Gaussian smoothing, contrast enhancement, Otsu thresholding
- Segment Anything Model v2 (SAM2) for overlapping particles
- **Output**: individual particle images ready for feature extraction and classification

Feature Extraction & Unsupervised Classification
- Each segmented particle is represented as a feature embedding using the DINOv2 vision transformer
- Focus on unsupervised learning due to limited labeled data
- Use pre-trained DINOv2 features (frozen backbone) — no fine-tuning or learned head
- k-means clustering applied to the embeddings to form pseudo-labels
- Visually similar particles are grouped together, forming a structured dataset for downstream analysis
- Enables robust classification without manual annotations

Implementation

https://github.com/nikolai-andrianov/palynofacies

Unsupervised Clustering – All Particles



- Many similar particles are clustered correctly, but several particle types are mixed, e.g.:
  - Cluster 4: effectively identifies black wood particles but includes others
  - Cluster 6: groups overlapping particles that should be separated
- Clustering driven mainly by color and shape
- Promising for rapid, unbiased organization of large image datasets

Supervised Learning – Detecting Spores

Spores as Calibration Markers for Particle Quantification
- Spores are routinely added during sample preparation
  - A tablet containing a known quantity of spores is mixed with a weighed sediment sample
  - Counting spores and target particles on a slide yields amount per gram of sediment
  - This allows quantification of absolute particle abundance in each sample
- Dataset Preparation
  - Target class: artificially added spores
  - Labels: small curated set of spore and non-spore particles
  - Augmentation: rotation, flips, color jitter

Model Architecture & Training
- Base model: ResNet-18, pre-trained on ImageNet
- First convolutional and batch normalization layers, and residual blocks 1–2 are frozen
  - General low-level features (edges, corners, basic textures) transfer well to microscopy images
  - Prevent overfitting on small dataset
  - Reduces training time and memory usage
- Layers 3–4 and the final classifier trained
  - Learn domain-specific features
  - Binary classification (spore/non-spore)
- ~5 million trainable parameters
- Loss: binary cross-entropy
- Clear signs of overfitting after epoch 4 → early stopping

Results

Category	Description	Estimate
True Positives	Correctly identified spores	≈ 80%
False Positives	Incorrectly labeled as spores	≈ 10%
Uncertain	Particles out of focus	≈ 10%
True Negatives	Correctly identified as non-spores	Qualitatively high
False Negatives	Missed spores among non-spores	Qualitatively low

Next steps

- Extend the image-reading workflow to access all focal planes in microscopy slides
- Expand labeled dataset with new spore variants
- Use active learning to refine uncertain cases
- Integrate CNN outputs into the broader unsupervised pipeline for hybrid classification