Teaching AI Agents to See: Geometric Machine Learning for Single-Cell Data

Brian Di Bassinga Computer Science
Smita Krishnaswamy, Computer Science and Genomics, Yale University

WHY AGENTS NEED TO SEE

Complex biological data needs more intelligent analysis

Embeddings reveal cell relationships:
- ➢ Clusters → Different kinds of cells
- ➢ Trajectories → Cell Development
- ➢ Branching → Cells "choosing" a path

Current agents describe structure in word, losing geometric information

We teach agents to "see" structure through quantitative metrics.

KEY FINDINGS

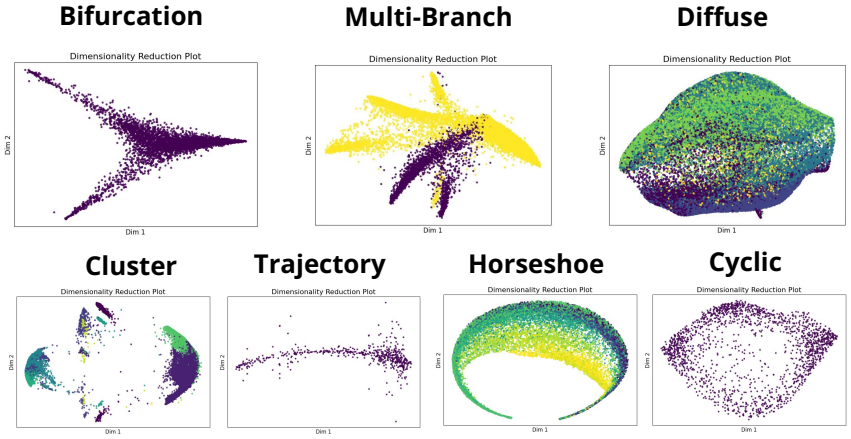
- **> 40.43%** accuracy on 7-class classification
- **> 2.9x** improvement over random baseline
- **> 53** geometric metrics extracted per embedding
- **> Spatial Entropy features are most predictive**
- > Distinctive shapes (horseshoe) classified well
- > Similar structures need more features (i.e., Clusters vs Multi-Branch)

LEARNING TO SEE

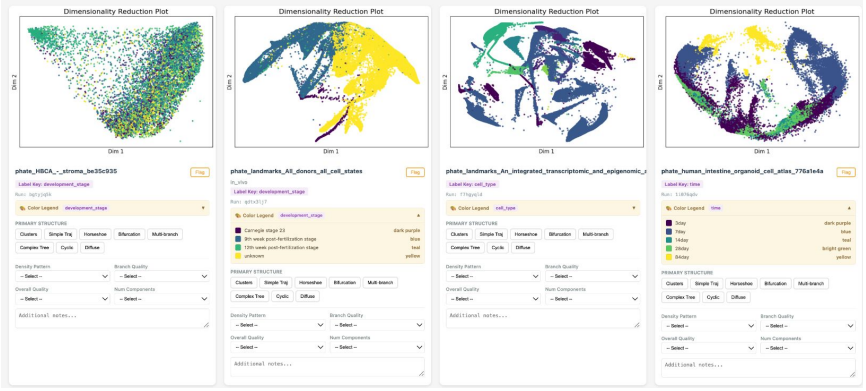
Data Pipeline:
- ➢ 100 datasets from CELLxGENE (Chan Zuckerberg Institute)
- ➢ PHATE embeddings (2D projections)
- ➢ 100 expert-labeled structures

Machine Learning Model:
- ➢ 53 geometric features (density, entropy, topology)
- ➢ ML classifier (SVM with RBF kernel)
- ➢ Leave-one-out validation

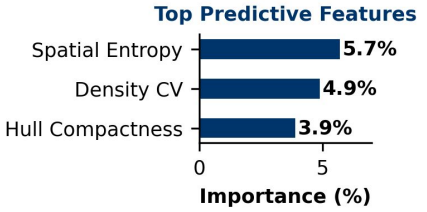
7-CLASS TAXONOMY

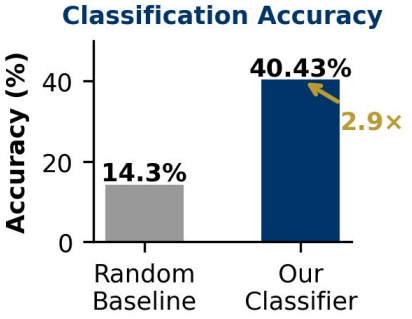


MANUAL CLASSIFICATION WEBSITE



RESULTS





NEXT STEPS

- ➢ Scale labeled data
- ➢ Integrate with multi-agent orchestration