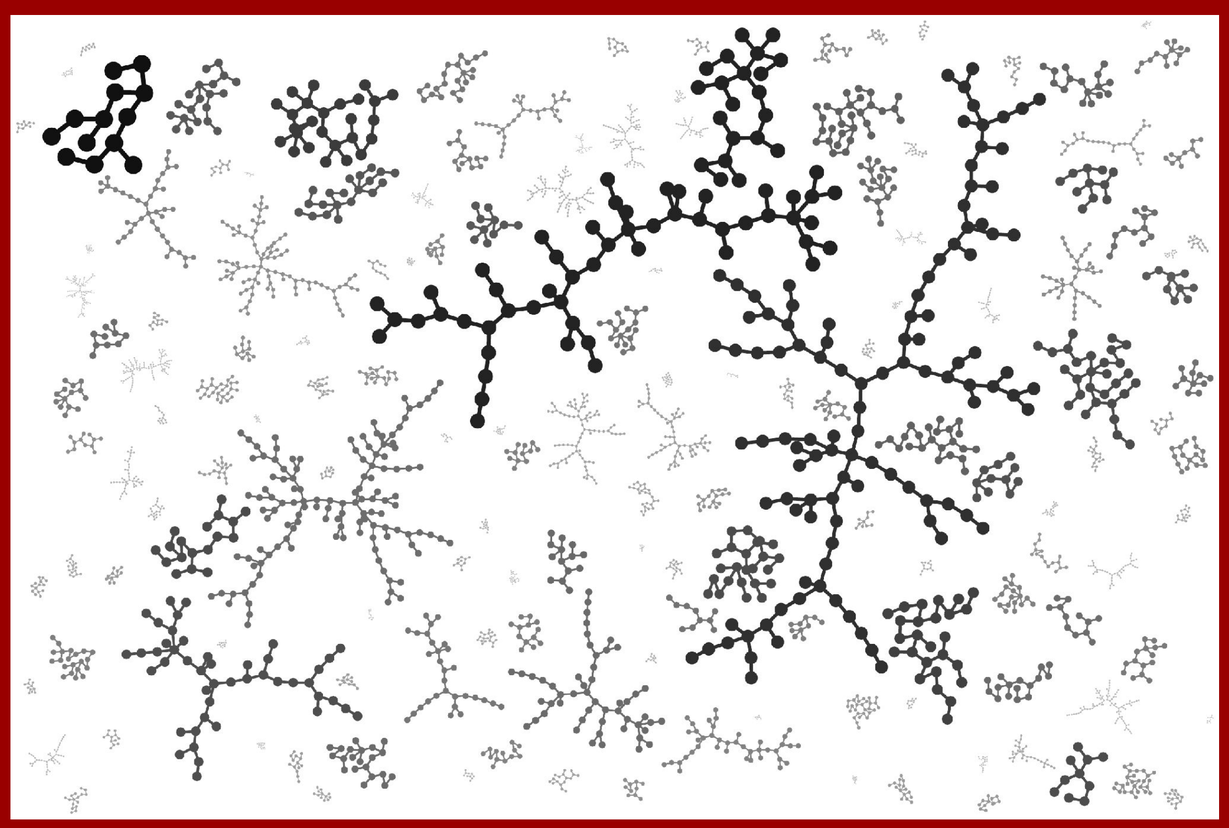


A principled testbed for interpretability

Sparse, self-similar, power-law, and analytically tractable synthetic data with ground-truth latents

Inputs form sparse, self-similar fractal clusters with power-law sizes.

Critical Percolation as a Synthetic Data Model for Interpretability

Ari Brill & Tom Ingebretsen Carlson (Principles of Intelligence)

1 TL;DR

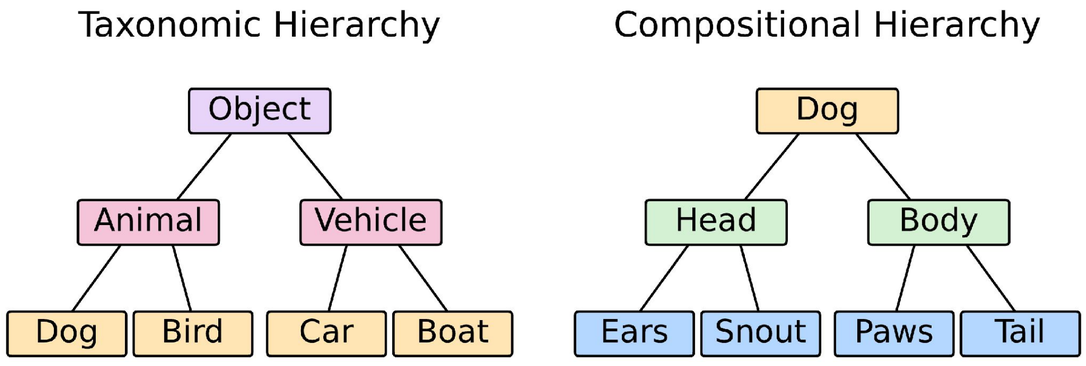
- We propose a synthetic dataset based on critical mean-field percolation: sparse, self-similar fractal clusters with power-law sizes and a built-in taxonomic hierarchy.
- An almost-linear-time cyclic coalescent algorithm samples a random tree and its latent hierarchy together, scaling to millions of points.
- Ground-truth hierarchical latents are linearly recoverable from a trained network's activations using linear probes.

2 Background & Motivation

- Synthetic datasets give interpretability methods ground truth, but most lack the multi-scale structure of real data.
- Real data is sparse, hierarchical, low-dimensional, and power-law distributed; fractality yields all four at once.
- Can a tractable fractal model be a realistic interpretability testbed?

3 The Percolation Model

- Sites occupied at random on a high-dimensional lattice modeling the data space form clusters; above 6 dimensions they are cycle-free trees.
- At criticality the model is exactly solvable: power-law sizes (exponent 5/2) and fractal dimension D = 4, with no hyperparameter tuning.
- Targets come from latent variables in a binary tree, giving an explicit ground-truth taxonomic hierarchy.



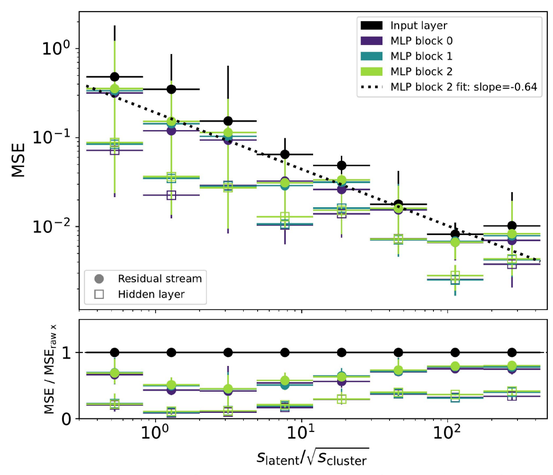
Targets follow a taxonomic hierarchy, distinct from compositional structure.

4 Method & Results

Sampling: the cyclic coalescent

- Critical clusters are uniform random labeled trees.
- We propose an efficient variant of Pitman's additive coalescent that merges blocks of nodes in a random cyclic order.
- Returns a tree and its latent hierarchy in effectively linear O(n α(n)) time.

Cyclic coalescent: arrange nodes in a random cycle, then each block iteratively merges with its successor.



Per-latent probe MSE beats the raw-input baseline and trends as a power law.

Probing: are latents represented linearly?

- We train a residual MLP on the regression task and fit linear probes to recover ground-truth hierarchical latents.
- Latents are linearly decodable: probe error beats the raw-input baseline and follows a power law.
- MLP matches a 1-NN best case ($R^2 \approx 0.94$ / 0.88); probe quality is flat across layers.

5 Discussion & Limitations

- One hyperparameter-free model unites sparsity, self-similarity, power laws, and analytical tractability.
- We show a proof of concept that theory-grounded synthetic datasets can serve interpretability research.
- Future work could test for analogs of sparse autoencoder pathologies such as feature splitting and absorption, and investigate the relationship between data structure and neural activations.
- Directions to extend the model include supporting tokenized data and classification tasks, and modeling compositional concepts using supercritical percolation allowing for clusters with cycles.

PAPER

CODE

Mechanistic Interpretability Workshop, ICML 2026

aryeh.brill@gmail.com