

SAE features organize into stable, recoverable dependence graphs

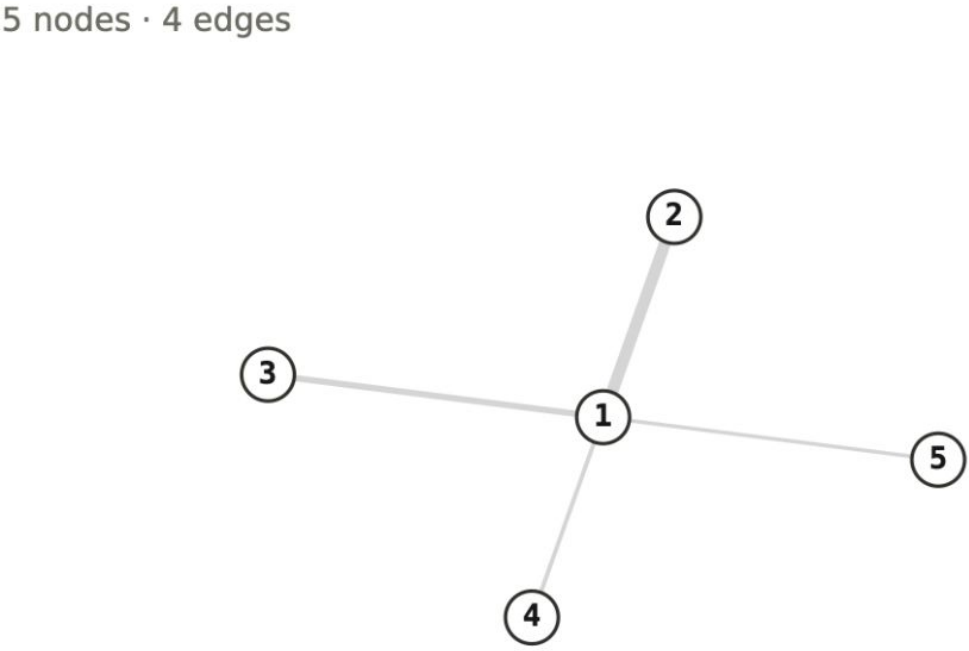
Conditional Dependence Structure in Sparse Autoencoder Features

Background: SAEs learn overcomplete dictionaries with thousands of interpretable features, but how these features are organized remains poorly understood.

Result 1: Features organize into semantically coherent sparse dependence graphs that are stable under resampling

Result 2: This structure isn't well explained by similarity-based baselines. Graph edges are not recovered by cosine similarity or correlation

FineWeb, Layer 10, Module 14
Concept: 'modal possibility'

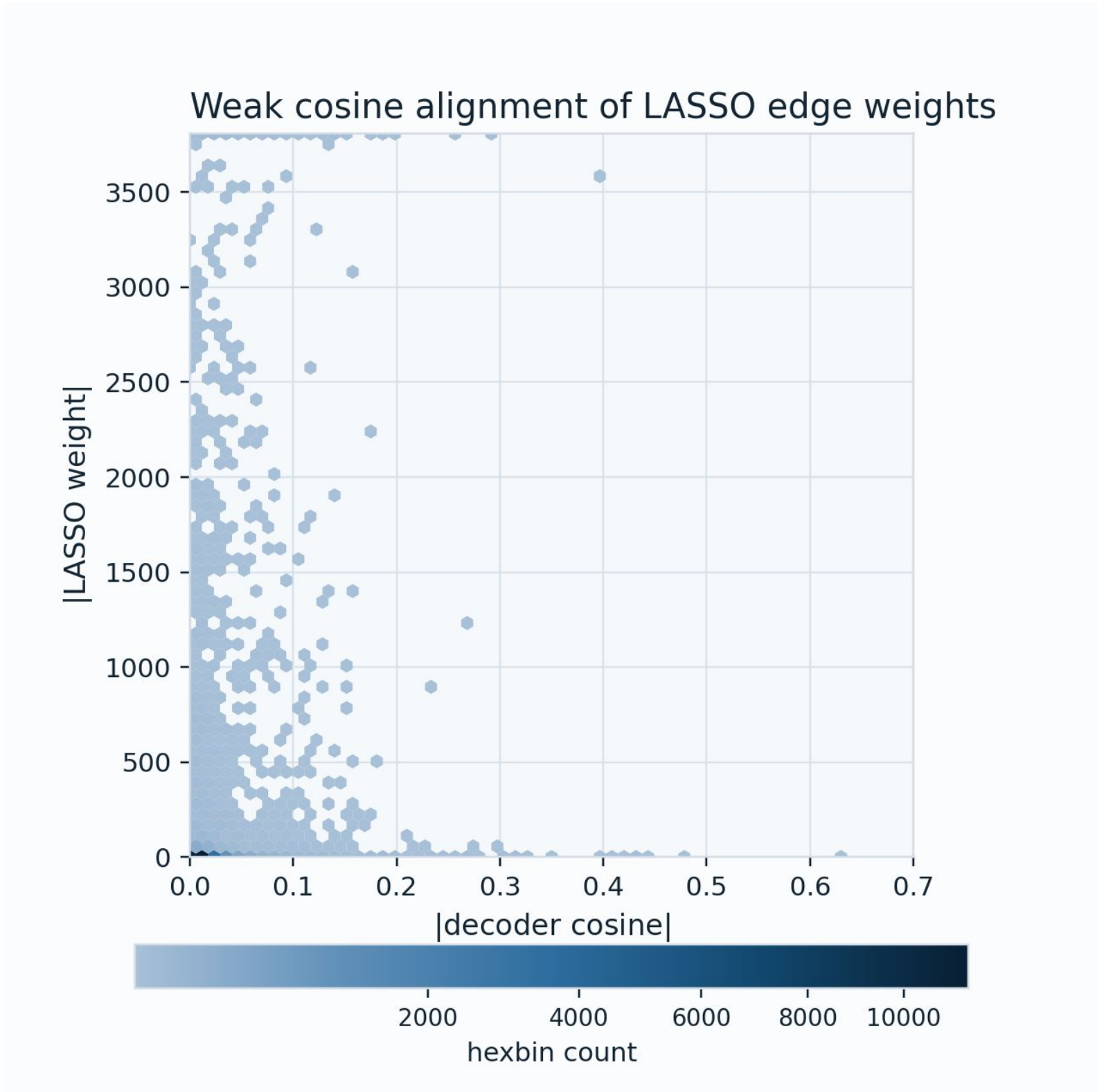


- 1 may be
- 2 may express possibility
- 3 may lead to
- 4 modal verbs may/might
- 5 possibility or uncertainty

WMDP-bio, Layer 10, Module 27
Concept: 'infection and conditions'

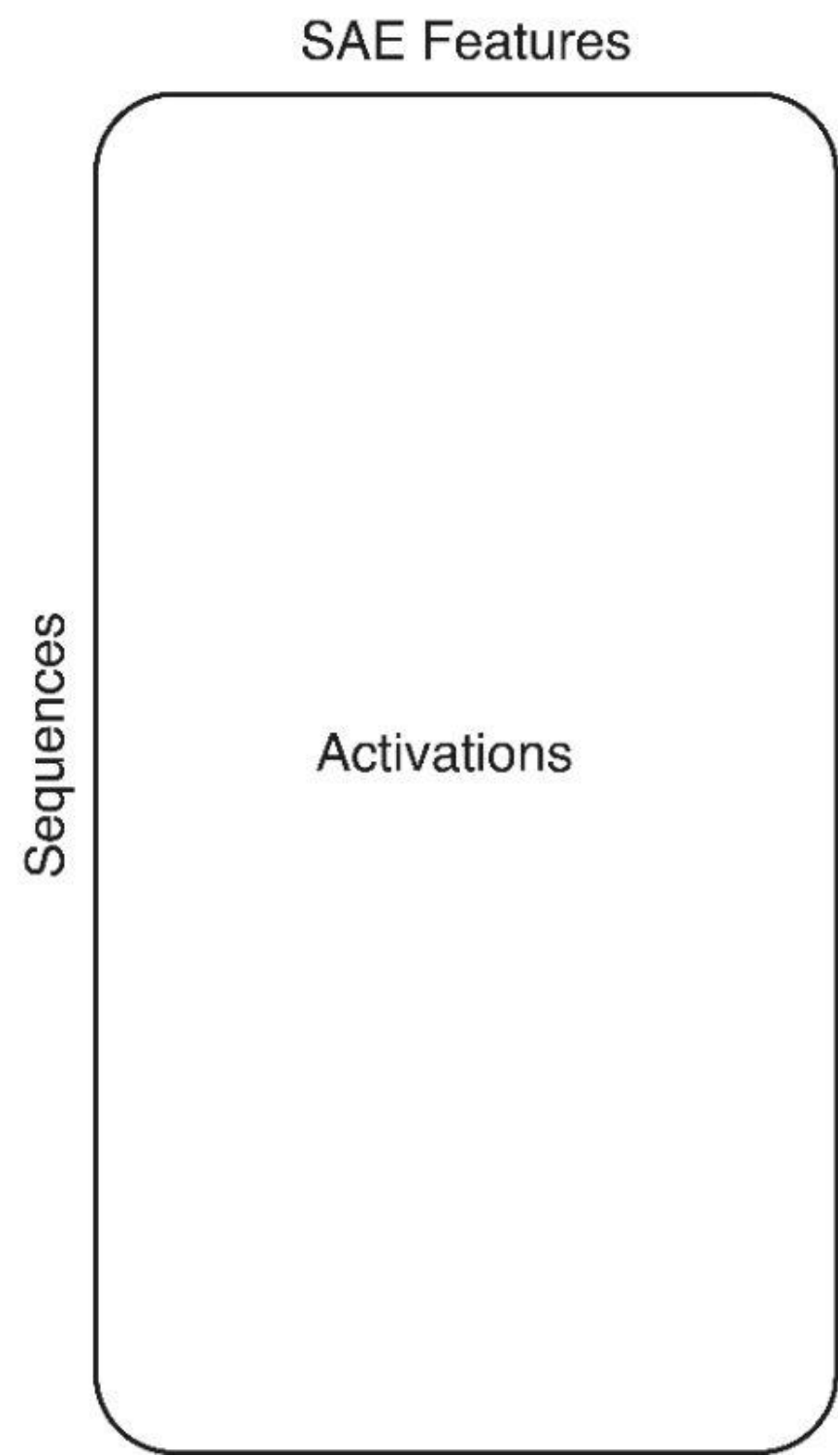


- | | | |
|---|---|---|
| 1 infected state | 6 climatic conditions and climate | 11 states of being or readiness |
| 2 infection | 7 clinical infections | 12 state or process |
| 3 hypoxia, glucose, stretch, ethanol, deprivation | 8 chemical concentrations and regulations | 13 quantifiable properties and measurements |
| 4 chromatography and purification techniques | 9 given when | 14 actin filaments and cytoskeleton |
| 5 crises and emergencies | 10 together, alongside, or separately | 15 oils and emulsions |

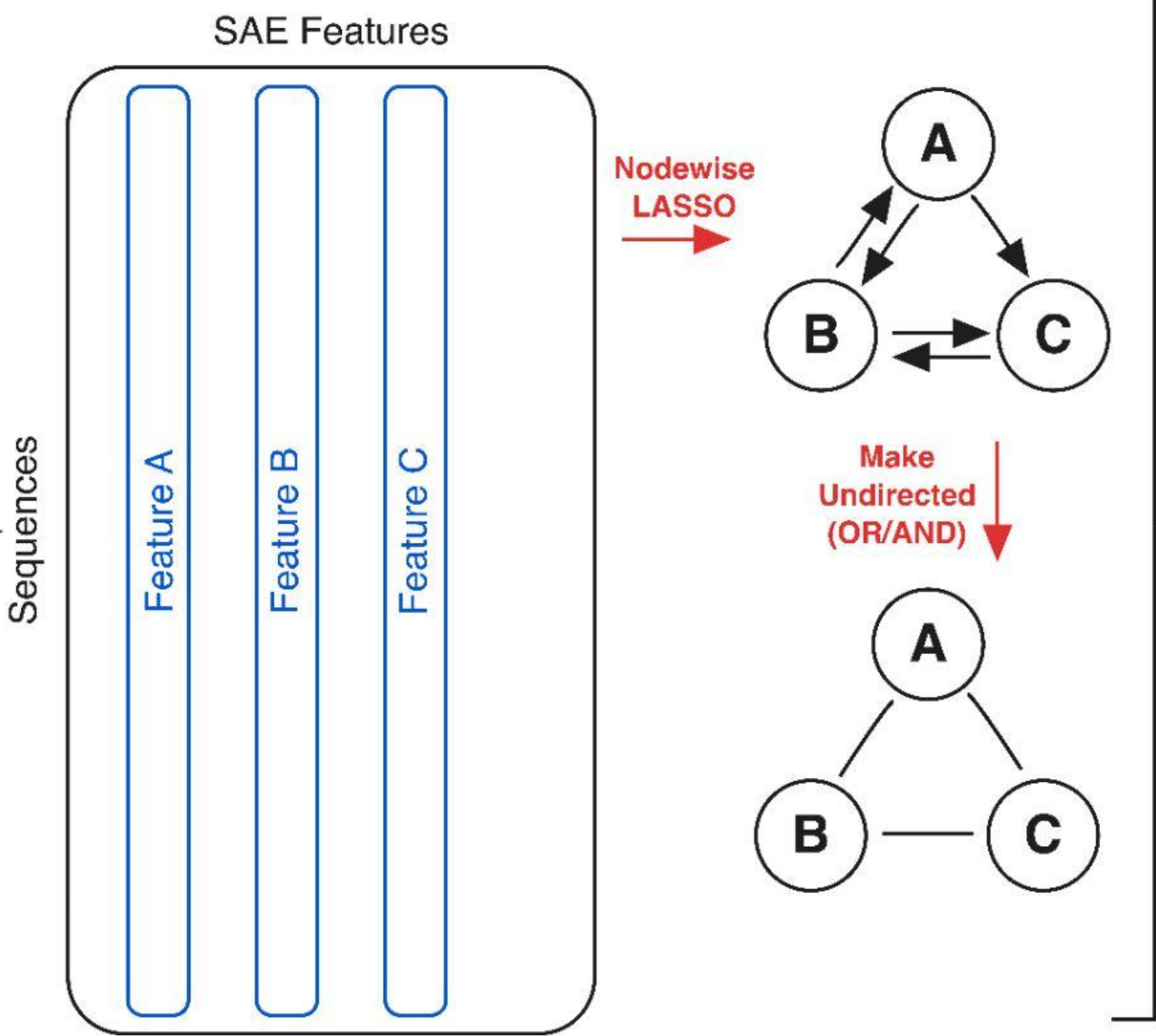


Method

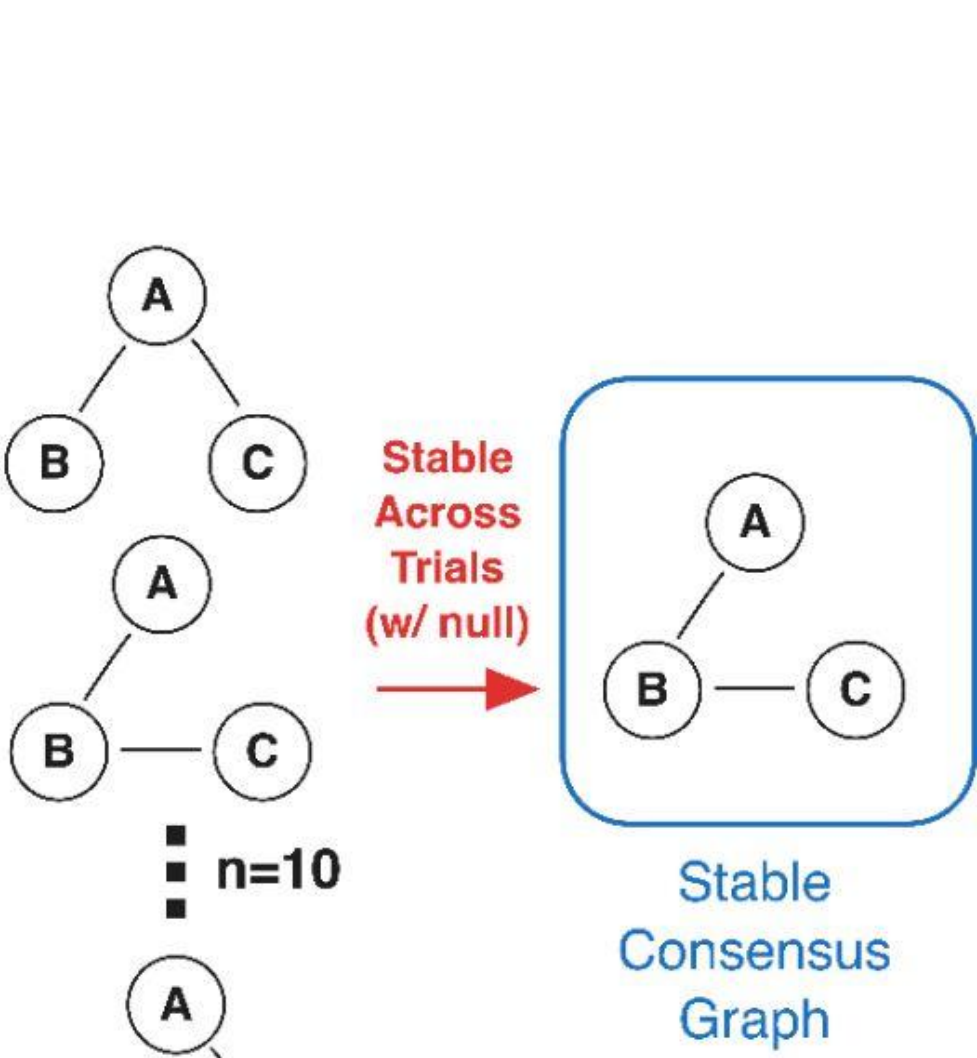
1. Sample SAE Activations



2. Build Dependence Graph



3. Repeat and Merge



Limitation: These graphs estimate statistical dependence, not causal circuits. More work is required to understand how this structure relates to feature splitting and merging



PAPER

Zachary Maas

Mechanistic Interpretability Workshop, NeurIPS 2026