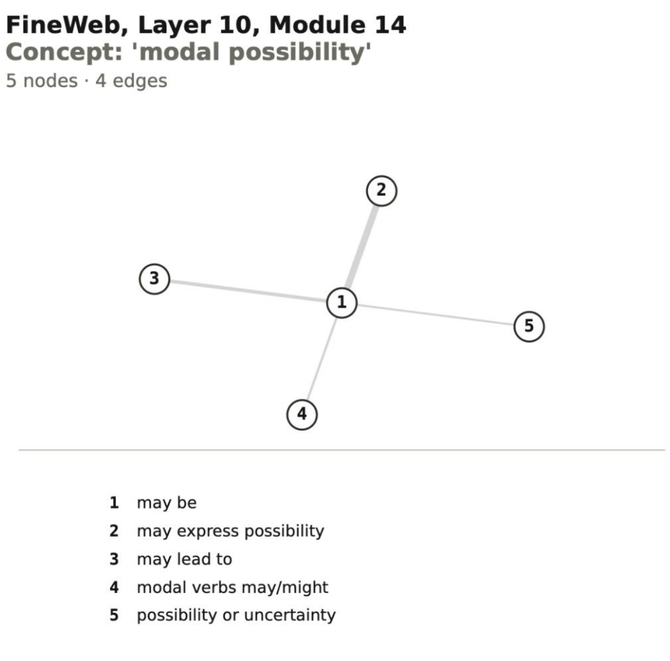
SAE features organize into stable, recoverable dependence graphs

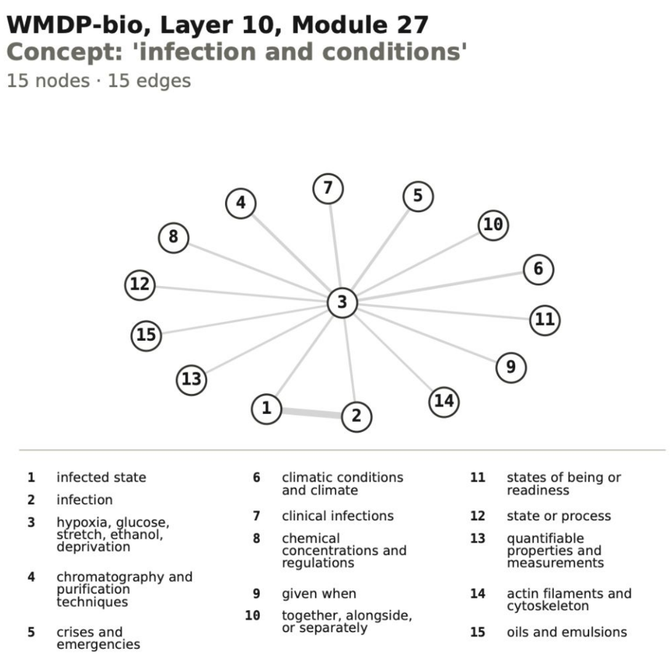
Conditional Dependence Structure in Sparse Autoencoder Features

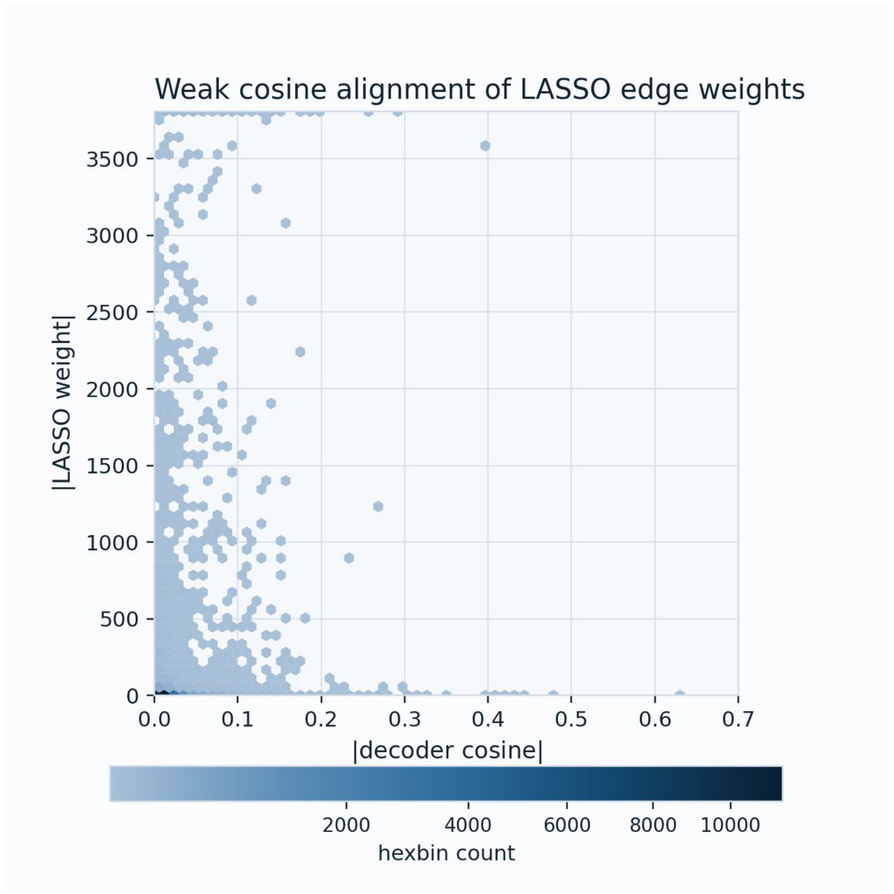
Background: SAEs learn overcomplete dictionaries with thousands of interpretable features, but how these features are organized remains poorly understood.

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

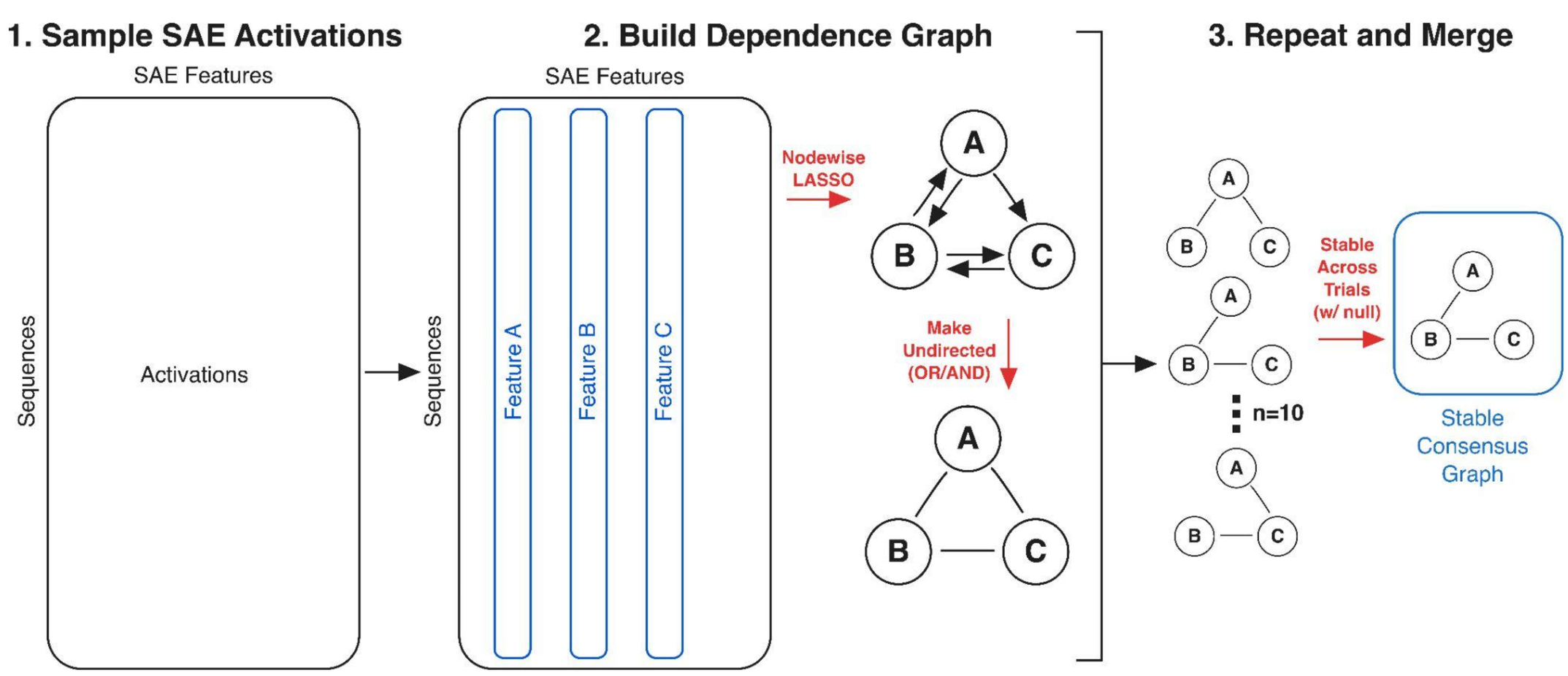
Result 2: This structure isn't well explained by similarity-based baselines. Graph edges are not recovered by cosine similarity or correlation







Method



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