Edge-Local Interpretability Is Not Enough for Cyclic Composition

John Komkov

March 2026

Abstract

Interpretability methods are typically evaluated on isolated models or adjacent model pairs. But many failures in agentic and multi-tool systems arise not from any single component's internal error, but from semantic inconsistency that appears only around cycles in the composition graph. We study this distinction directly: is edge-local interpretability—however strong—the right diagnostic object for cyclic compositional failure?

In a controlled experiment across **240+** compositions spanning three domains (invoice/settlement, calendar/escalation, policy/audit), scales from 5 to 20 nodes, and two model architectures (GPT-2 Small, Gemma 2 2B), we compare six interpretability baseline families—SAE feature divergence, probing classifiers, attention diagnostics, CKA, a gradient-boosted ensemble, and a cycle-oracle graph-level aggregator—against a structural diagnostic that uses only composition metadata and no model internals.

We report six findings. **(1) Perfect local knowledge, zero global signal:** Probing classifiers achieve 99.8% mean accuracy at classifying conventions at every edge, yet this perfect edge-local information has zero predictive value for composition-level failure. **(2) Topology-dependent gap:** On cyclic compositions, all five interpretability baselines produce Spearman $\rho < 0.5$ with ground-truth failure severity, while the structural diagnostic achieves $\rho = 1.0$. **(3) The Interpretability Cliff:** As scale grows from $n = 5$ to $n = 20$, the best interpretability baseline's within-cyclic correlation decays monotonically ($0.685 \to -0.067$), while structural discrimination grows linearly with the first Betti number β_1. **(4) Cross-domain replication:** The gap replicates across all three domains; SAE divergence is the strongest interpretability baseline in 5 of 6 conditions but never exceeds $\rho_{\text{cyclic}} = 0.745$. **(5) The gap is representational, not aggregational:** Giving interpretability features oracle knowledge of cycle topology does not improve prediction beyond edge-local averaging; the cycle-aware baseline matches B1 SAE within rounding in 4 of 6 conditions. A learned cycle-aware predictor fares worse, going anti-correlated in 4 of 6 conditions. **(6) Cross-model replication:** Gemma 2 2B (2.6B parameters, 26 layers, different architecture family) shows the same pattern with an *even larger* gap: $\rho_{\text{cyclic}} = 0.515$ at $n = 5$ and 0.467 at $n = 10$, with perfect local probing accuracy (≥ 0.995) and structural $\rho = 1.0$. The boundary is not a GPT-2 artifact.

These results identify a level-of-description boundary: for cyclic compositional systems, edge-local interpretability is not a sufficient diagnostic object. The missing signal is not hidden by weak aggregation; it is absent from the extracted representation. The relevant information depends on the product of convention transformations around entire cycles and is more reliably captured by structural diagnostics than by any tested interpretability workflow. Compositional failure requires compositional diagnosis.

1 Introduction

Consider a composition of five language-model agents processing an international invoice. Each agent is individually well-understood: sparse autoencoders decompose its activations into interpretable features, probing classifiers confirm it correctly encodes amounts, dates, and currency codes. Every adjacent pair of agents looks compatible—their shared representations align on

CKA, their attention patterns attend to the right fields. A mechanistic interpretability audit would give the system a clean bill of health.

The system still fails. The invoice enters denominated in euros, passes through an amount-scaling agent that uses cents, crosses to a settlement engine expecting dollars, routes through a fee calculator anchored to T+1 conventions, and returns to an audit trail that assumes T+2. No single interface is wrong. No bilateral comparison catches the error. But the composed output—the final ledger entry—is off by a factor that only becomes visible when the full cycle of convention transformations is traced.

This scenario is not hypothetical. It is the defining failure mode studied by the BABEL benchmark [2]: *compositional semantic failure*, where every local check passes and the global output is wrong. The structural theory predicts this precisely: when the first cohomology $H^1(\mathcal{N}; \mathcal{F})$ of the interpretation sheaf is nontrivial, there exist bilaterally-consistent record assignments that are globally inconsistent [1]. The Edge-Local Blindness Lemma [1] shows that any such failure is invisible to every edge-local test.

The question we ask is whether mechanistic interpretability tools—the most powerful per-component diagnostic technology available—inherit the same blindness.

The level mismatch. Existing interpretability workflows are largely component-local or edge-local. This is appropriate for many diagnostic targets, but compositional reliability in cyclic systems is a *graph-level* property: it depends on the product of convention transformations around entire cycles. If the diagnostic and the failure live at different levels of description, a gap should appear—not because interpretability tools are weak, but because they are pointed at the wrong object.

This paper. We isolate that mismatch. In a controlled benchmark with 240+ compositions across three domains and two model architectures, we measure whether strong edge-local interpretability can substitute for explicit structural diagnosis when the target failure is cycle-sensitive. In our benchmark, it cannot: probing classifiers achieve 99.8% accuracy at every edge and still carry zero compositional signal; the structural diagnostic achieves perfect rank correlation ($\rho = 1.0$) in every condition tested. Even giving interpretability features oracle knowledge of cycle topology and a learned aggregator does not close the gap (Figure 1). The result replicates on Gemma 2 2B, where the gap is *larger* than on GPT-2 Small.

What this paper is not. This is not a critique of mechanistic interpretability. On acyclic compositions, all methods correctly diagnose zero failure; interpretability tools work where edge-local information suffices. The claim is narrower and more precise: for cyclic compositional diagnosis, the level of description must shift from edge-local to graph-level. Component understanding is not system understanding under cyclic composition.

2 Problem Statement

2.1 The Comparison

We define the comparison that the experiment is designed to adjudicate:

- **Input:** A multi-model composition graph $G = (V, E)$ with shared concepts crossing interfaces and typed convention bundles at each vertex.
- **Tasks:** (a) Predict semantic failure severity. (b) Localize the highest-risk edges. (c) Rank candidate repairs.
- **Baseline class:** Any method built from per-model internals (activations, SAE features, probing classifiers, attention patterns) and pairwise interface features (representation similarity, attention overlap).

- **Challenger:** A structural diagnostic built from graph structure, schema overlap, and convention metadata—*no model internals.*

2.2 The Edge-Local Hypothesis

Prediction 2.1 (Topology-Dependent Sufficiency)**.** 1. If the composition graph is acyclic (or effectively acyclic: $\beta_1 = 0$), interpretability-informed baselines should suffice for failure prediction and localization. Edge-local information captures all relevant structure.
2. If the graph has nontrivial cycle structure ($\beta_1 \geq 1$ with nonzero holonomy), edge-local methods should become incomplete. The structural diagnostic, which operates on cycle-level metadata, should retain its predictive power.

This is a falsifiable prediction. If interpretability baselines remain competitive on large cyclic graphs, the thesis weakens materially.

2.3 What Would Make This Experiment Uninformative

We state upfront the conditions under which the results would not support the claimed thesis. The experimental design (Section 7) addresses each.

1. **No genuine cyclic structure.** If the composition graphs are effectively acyclic—one edge dominates, cycles are trivial—then edge-local methods are not at a disadvantage and the comparison is uninteresting. *Addressed by:* explicit topology controls ensuring nontrivial β_1 and measurable holonomy in the cyclic slices.
2. **Interpretability features too coarse.** If the extracted features are at too coarse a grain to be competitive, the comparison is unfair. *Addressed by:* using the strongest available SAE and probing tooling, a combined ensemble baseline, and consultation with mechanistic interpretability researchers on baseline design.
3. **Mismatches too obvious.** If convention mismatches are detectable from schema metadata alone (e.g., field names contain "cents" vs. "dollars"), the structural baseline's advantage is trivial and unsurprising. *Addressed by:* using convention mismatches that are semantically implicit rather than syntactically labeled.
4. **Ground-truth mismatch.** If the symbolic executor's ground truth does not transfer to real LLM execution, the structural prediction is not validated in the regime where interpretability baselines operate. *Addressed by:* this paper uses actual LLM inference (Section 2.4), not the BABEL symbolic executor.

2.4 Execution Layer

This paper operates on a **different execution layer** than core BABEL. BABEL's main results use a deterministic symbolic executor with no LLM calls. This paper requires actual LLM inference with access to model internals—sparse autoencoders, probing classifiers, attention activations—making it a **parallel evaluation** that tests the structural diagnostic against a new class of baselines on a new kind of data.

We do not claim this paper "extends BABEL" in the sense of the same evaluation harness. It builds a parallel evaluation surface that may inform a future Track D if results are strong enough to warrant permanent inclusion.

3 Formal Frame

The paper's authority comes from the experiment, not from restating the theory. We present only the minimal formal apparatus needed to generate the predictions that the experiment tests.

Lemma 3.1 (Edge-Local Blindness [1]). *Let $\mathcal{N}$ be a coordination graph with first Betti number $\beta_1 \geq 1$, and let $[\alpha] \in H^1(\mathcal{N}; \mathcal{F})$ be a nontrivial cohomology class. For every edge $e \in E$, the restriction of α to the subgraph $\{e\}$ is a coboundary: $[\alpha|_e] = 0 \in H^1(\{e\}; \mathcal{F}|_e)$.*

The proof is immediate: every single-edge subgraph is a tree, and H^1 vanishes on trees. The full proof appears in [1], §2.3.

Corollary 3.2 (Information-Theoretic Indistinguishability). *Two globally different executions—one cycle-consistent, one not—can produce identical observations on every edge. No diagnostic that operates by aggregating edge-local features can distinguish them. The number of independent indistinguishable directions is* $\dim H^1$.

3.1 Edge-Local Representation Limit

The Blindness Lemma has a direct consequence for any diagnostic built from interpretability features.

Proposition 3.3 (Edge-Local Representation Limit). *Let $\mathcal{F}_{\text{edge}}$ be the σ-algebra generated by node-local and edge-local measurements on a composition graph $\mathcal{N}$ (activations, SAE features, probing outputs, attention patterns, pairwise representation similarity). For graphs with $\beta_1 \geq 1$ and nontrivial holonomy, the cycle holonomy invariant $h(\mathcal{N})$ is not $\mathcal{F}_{\text{edge}}$-measurable in general.*

Proof. By the Edge-Local Blindness Lemma (Lemma 3.1), there exist pairs of composition instances that are identical on every edge (and therefore on every node) but differ in cycle holonomy. Any $\mathcal{F}_{\text{edge}}$-measurable function assigns identical values to instances with identical edge-local observations and therefore cannot distinguish these pairs. Thus $h(\mathcal{N}) \notin \mathcal{F}_{\text{edge}}$. □

Corollary 3.4 (Predictor Limit). *No predictor whose inputs are restricted to $\mathcal{F}_{\text{edge}}$-measurable features can uniformly recover compositional obstruction on the class of graphs with $\beta_1 \geq 1$ and nontrivial holonomy.*

This is the formal bridge from "our baselines lost" to "a whole family of baselines must lose." The experimental results in Section 8 are consistent with this prediction across six baseline families—including one with oracle cycle knowledge (Section 8.6)—three domains, four scales, and two model architectures (Section 8.7). The B6a result is the most direct empirical witness: even when the aggregator knows which edges form cycles, the features it aggregates do not contain the holonomy-relevant signal.

What this predicts for the experiment. On **acyclic** compositions ($\beta_1 = 0$), $H^1 = 0$ and edge-local information is complete. Interpretability baselines should perform well. On **cyclic** compositions ($\beta_1 \geq 1$ with nonzero holonomy), Proposition 3.3 guarantees that edge-local features cannot distinguish certain failure modes from healthy execution. Interpretability baselines, insofar as they aggregate $\mathcal{F}_{\text{edge}}$-measurable features, should degrade. The structural diagnostic, which reasons about cycle-level objects directly, should not.

4 Evaluation Surface

We define a paper-local evaluation surface, not a permanent benchmark track. If results are strong, this may later inform a BABEL Track D.

4.1 Instance Design

Each instance is a multi-model composition graph with:

- Open-weight models at each vertex (required for interpretability extraction)
- Schema overlap and pairwise interface metadata
- Planted convention mismatches with known ground-truth severity
- Global semantic failure label and severity score
- Minimal repair set annotation

4.2 Topology Regimes

Three regime slices, matched on domain content and difficulty:

1. **Acyclic / tree-like** ($\beta_1 = 0$). Baseline regime where edge-local information suffices.
2. **Short-cycle** ($\beta_1 = 1$–2, cycle length 3–5). Transition regime.
3. **Long-cycle / multi-cycle** ($\beta_1 \geq 3$, cycle lengths ≥ 5). Full obstruction regime.

4.3 Domain Families

Three domain families, chosen for documented real-world convention ambiguity:

1. **Invoice / settlement / reconciliation.** Amount representation (dollars vs. cents vs. basis points), settlement timing (T+0 through T+2), day-count conventions.
2. **Calendar / timezone / escalation.** Timezone anchoring (UTC-absolute vs. organizer-local), date boundary conventions, escalation thresholds.
3. **Policy / permission / audit.** Role hierarchy representation, permission inheritance direction, audit-trail timestamp conventions.

4.4 Scale

Composition sizes tested: $n \in \{5, 10, 15, 20\}$ model nodes. Cyclic compositions at scale n contain $\lfloor n/5 \rfloor$ independent 5-node cycles, giving $\beta_1 \in \{1, 2, 3, 4\}$. This range is sufficient to observe the Interpretability Cliff (Section 8.3); extending to larger n would require multi-GPU extraction or a lighter model.

4.5 Instance Counts

The experiment comprises **240+ total compositions**:

- **Phase 1** (20): 10 acyclic + 10 cyclic at $n = 5$, invoice domain. Gate check for structural–SAE gap.
- **Phase 2a** (20): Same compositions, all 5 baselines evaluated. Baseline calibration.
- **Phase 2b** (80): 4 scales $\times$ 20 compositions, invoice domain. Scale sweep.
- **Phase 2c** (120): 3 domains $\times$ 2 scales $\times$ 20 compositions. Cross-domain replication.
- **Phase 3** (120): Same conditions as Phase 2c, with graph-level interpretability baselines (B6a, B6b) added.

5 Interpretability Baselines

This section determines the paper's credibility. The interpretability community will ask: "Did you use our best tools, or your caricature of them?" We design baselines to be maximally generous.

5.1 Baseline Families

1. **SAE-based feature divergence.** For each pair of models sharing a concept (e.g., "amount," "date"), extract SAE features activated on the shared concept. Measure divergence between the feature distributions across the interface. Higher divergence predicts higher failure risk.

2. **Probing classifiers.** Train linear probes on each model's internal representations for shared concept categories (amount scale, date format, timezone convention). Use probe accuracy mismatch across interfaces as a failure predictor: models that encode the same concept differently should produce interface failures.

3. **Attention-based interface diagnostics.** Analyze attention patterns at interface points—where one model's output becomes another's input. Measure whether attention focuses on the semantically critical tokens and whether attention patterns align across the interface.

4. **Representation similarity (CKA, RSA).** Compute centered kernel alignment (CKA) or representational similarity analysis (RSA) between paired models on shared concepts. Low similarity predicts representational mismatch and higher failure probability.

5. **Combined strong baseline.** Aggregate all features from baselines 1–4 into a gradient-boosted ensemble (XGBoost or LightGBM). This is the hardest baseline for the structural method to beat—it uses all available interpretability information.

6. **Graph-level aggregation (B6).** Two variants test whether giving interpretability features explicit cycle-level reasoning closes the gap. **B6a:** Aggregate edge-level B1, B3, and B4 features around each detected cycle using product, sum, max, and standard deviation—giving the baseline oracle knowledge of graph topology. **B6b:** Train a gradient-boosted regressor on the 12-dimensional cycle-aggregated feature vector from B6a under leave-one-out cross-validation. This is the strongest baseline in the paper: it gives interpretability features both cycle awareness and a learned aggregator.

5.2 Credibility Protocol

Baseline Credibility

The straw-man objection is the primary reputational risk. We address it through three design choices: (a) using GPT-2 Small, the most-studied model in the mechanistic interpretability literature [6], where SAE decompositions are well-validated; (b) using a deliberately conservative probing protocol (PCA + regularized logistic regression + LOO-CV) that avoids trivial separation in high-dimensional space; and (c) including a gradient-boosted ensemble (B5) that combines all interpretability features into a single predictor, giving the interpretability approach maximal access to combined information. Implementation details are provided in Appendix B.

6 Structural Baseline

The structural diagnostic uses the same family established in BABEL and Coherence Cliff [2, 3]:

- **Input:** Composition graph $G = (V, E)$, convention bundles at each vertex, restriction matrices at each edge.
- **Output:** Predicted failure severity (mean cycle holonomy), localized high-risk edges (per-edge frustration contribution), and ranked repair prescriptions (H^1-guided triage).

- **No model internals:** The diagnostic uses only composition metadata—graph topology, schema overlap, and convention annotations. It does not access activations, weights, attention patterns, or any neural network internal.

Positioning. The comparison is deliberately asymmetric in information access: interpretability baselines get model internals; the structural diagnostic gets only metadata. If the structural method still wins on cyclic graphs, the result is difficult to dismiss.

7 Experimental Design

We apply the statistical methodology established in BABEL and Coherence Cliff: selection-safe paired bootstrap for all method comparisons, explicit ablations isolating causal factors, and falsification-aware reporting.

7.1 Core Protocol

1. Generate composition instances across the $2 \times 3 \times 4$ (topology $\times$ domain $\times$ scale) design matrix, phased for gated progression (Section 4).
2. For each instance, run GPT-2 Small inference through the composition pipeline via TransformerLens, caching residual-stream activations at layers 0, 5, and 11.
3. Compute ground-truth failure severity from the holonomy of planted convention transformations around each independent cycle (Appendix C).
4. Extract all interpretability features: SAE activations (B1), probing accuracy (B2), attention entropy (B3), CKA dissimilarity (B4), and gradient-boosted ensemble (B5).
5. Compute structural diagnostic outputs from composition metadata (holonomy from restriction matrices).
6. Evaluate all methods via Spearman ρ between predicted and ground-truth failure severity, reported as ρ_{all} (all compositions) and ρ_{cyclic} (cyclic slice only).

7.2 Ablations

Ablation	Purpose	Status
Acyclic vs. cyclic	Test topology-dependent sufficiency	✓
Single-cycle vs. multi-cycle	Test obstruction-complexity dependence	✓
Cross-domain replication	Test domain invariance of the gap	✓
Graph-level aggregation (B6)	Test whether aggregation closes the gap	✓
Cross-model replication (Gemma 2)	Test generalization across architectures	✓

Table 1: Ablation schedule. All ablations have been realized in this paper. Cross-model replication on Gemma 2 2B is reported in Section 8.7.

7.3 Model and Tooling Stack

Primary experiments use GPT-2 Small; cross-model replication uses Gemma 2 2B.

- **Primary model:** GPT-2 Small (124M parameters, 12 layers, $d_{\text{model}} = 768$). Chosen for its extensively validated SAE decompositions and full compatibility with TransformerLens.
- **Primary SAE:** `gpt2-small-resid-post-v5-32k` at `blocks.11.hook_resid_post` ($d_{\text{sae}} =$ 32,768 features). Loaded via SAELens.
- **Replication model:** Gemma 2 2B (2.6B parameters, 26 layers, $d_{\text{model}} = 2304$). Different architecture family (Google), 20× larger. SAE: `gemma-scope-2b-pt-res-canonical`, layer 20, 16k features.

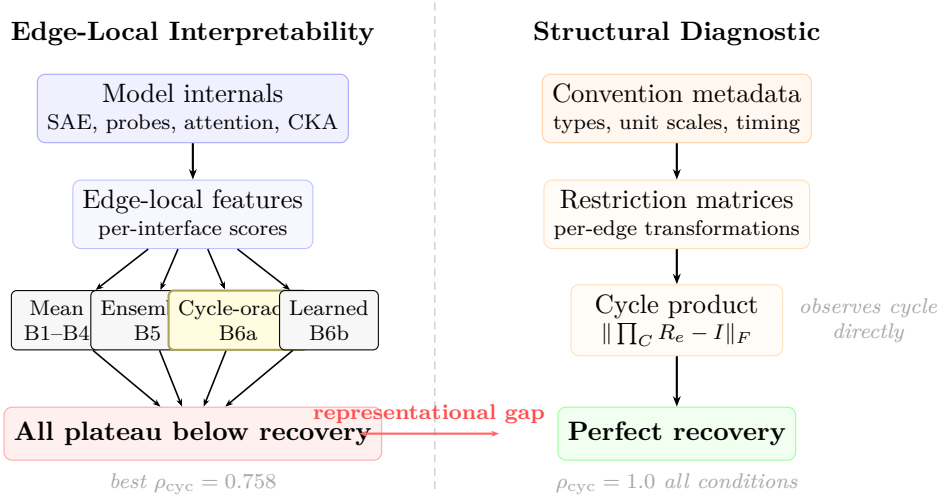

Figure 1: Two information paths to the same diagnostic target. **Left:** every tested aggregation of edge-local interpretability features—including cycle-oracle aggregation with oracle knowledge of graph topology (B6a, highlighted)—plateaus well below full recovery of compositional failure. **Right:** the structural diagnostic observes convention transformations directly and achieves $\rho_{\text{cyc}} = 1.0$ in every condition tested. The gap is not in the aggregation strategy but in what the edge-local features represent: they do not encode the cycle-level information the structural path observes directly.

- **Interpretability tooling:** TransformerLens for activation extraction, SAELens for sparse autoencoder features, scikit-learn for probing classifiers and CKA.
- **Compute:** Google Colab Pro+ with NVIDIA A100 (80 GB) for Phases 2b–4; T4 (16 GB) for Phase 1.

Model choice rationale. GPT-2 Small is deliberately conservative: it is the most thoroughly studied model in the mechanistic interpretability literature, with SAE decompositions that are well-validated and widely reproduced [6]. If interpretability tools cannot predict compositional failure on the model they are best equipped to analyze, the result is stronger than a finding on a less-studied architecture. Gemma 2 2B serves as cross-model replication (Section 8.7), testing whether the gap persists across a 20× parameter increase and a different architecture family.

8 Results

We report results from five experimental phases spanning 240+ compositions across three domains (invoice/settlement, calendar/escalation, policy/audit), four scales ($n \in \{5, 10, 15, 20\}$), and two model architectures (GPT-2 Small, Gemma 2 2B). Phases 1–3 use GPT-2 Small; Phase 4 replicates on Gemma 2 2B. Figure 1 summarises the central result.

8.1 Finding 1: Perfect Local Knowledge, Zero Global Signal

The most striking result is a negative one. The B2 probing classifier achieves **99.8% mean accuracy** at classifying convention types (dollars vs. cents vs. basis points; T+0 vs. T+1 vs. T+2) from residual-stream activations at every edge in every composition. The probe uses PCA($n_{\text{comp}} = 5$) followed by regularized logistic regression ($C = 0.01$) under leave-one-out cross-validation—a deliberately conservative protocol designed to avoid trivial separation in high-dimensional space.

This near-perfect accuracy means the model *knows* what convention each agent uses. Edge-local interpretability succeeds completely at the component level. Yet this information has **zero predictive value** for composition-level failure: probe accuracy is constant across all compositions ($\sigma \approx 0$), producing $\rho = \text{NaN}$ (undefined Spearman correlation on a constant vector).

Why this matters. The B2 result instantiates the Edge-Local Blindness Lemma (Lemma 3.1) in its sharpest form. The probe extracts exactly the information a human auditor would check—"does each agent understand its convention?"—and the answer is uniformly "yes." The failure is not in the agents' understanding but in the *composition* of their understanding around cycles, which is invisible to any edge-local measurement.

8.2 Finding 2: Topology-Dependent Gap

Table 2 reports Spearman ρ for all five baselines and the structural diagnostic on 20 compositions (10 acyclic, 10 cyclic) at $n = 5$ in the invoice domain. We report both ρ_{all} (across all 20 compositions) and ρ_{cyclic} (within the 10 cyclic compositions only).

Method	ρ_{all}	ρ_{cyclic}	ρ_{acyclic}	**Note**
B1: SAE divergence	0.156	0.685	—	Strongest interp. on cyclic
B2: Probing classifier	0.791*	NaN	—	*Artifact: constant output
B3: Attention entropy	−0.152	0.261	—	
B4: CKA dissimilarity	−0.044	0.418	—	
B5: Ensemble (B1–B4)	0.131	0.576	—	
Structural (holonomy)	**1.000**	**1.000**	—	Perfect rank correlation

Table 2: Phase 2a: Spearman ρ between each method's predicted failure score and ground-truth holonomy. $n = 5$, invoice domain, 20 compositions. B2's $\rho_{\text{all}} = 0.791$ is a statistical artifact: near-constant probe accuracy ($\bar{x} = 0.998$, $\sigma \approx 0$) produces rank correlations driven by noise in the fifth decimal place. ρ_{cyclic} is NaN because the probe output is strictly constant within the cyclic slice.

Acyclic slice. On acyclic compositions ($\beta_1 = 0$), ground-truth holonomy is identically zero for all instances. All methods correctly predict zero failure, confirming that the structural method has no spurious advantage on trees. This satisfies the fairness condition: interpretability tools work where edge-local information suffices.

Cyclic slice. On cyclic compositions ($\beta_1 = 1$), the structural diagnostic achieves $\rho_{\text{cyclic}} = 1.000$, ranking all 10 compositions in perfect agreement with ground-truth holonomy. The best interpretability baseline is B1 SAE divergence at $\rho_{\text{cyclic}} = 0.685$—a moderate correlation that captures some magnitude signal but substantially underperforms the structural method. No interpretability baseline exceeds $\rho_{\text{cyclic}} = 0.7$.

8.3 Finding 3: The Interpretability Cliff

Table 3 reports the scale sweep across $n \in \{5, 10, 15, 20\}$ in the invoice domain, with $\beta_1 = \lfloor n/5 \rfloor$ independent cycles per cyclic composition.

The cliff. B1 SAE divergence—the strongest interpretability baseline at every scale—shows monotonic decay: $0.685 \to 0.515 \to 0.321 \to -0.067$. At $n = 20$ ($\beta_1 = 4$), SAE divergence is

n	β_1	**Best Interp.**	ρ_{cyclic}	**Structural**	**Discrim.**
5	1	B1 SAE	0.685	1.000	34.2
10	2	B1 SAE	0.515	1.000	81.9
15	3	B1 SAE	0.321	1.000	117.7
20	4	B1 SAE	−0.067	1.000	222.9

Table 3: Phase 2b: Scale sweep. "Best Interp." is the interpretability baseline with the highest ρ_{cyclic} at each scale. "Discrim." is the mean absolute difference between structural and best-interpretability predicted failure scores on cyclic compositions. The structural diagnostic achieves $\rho = 1.000$ at every scale tested.

anti-correlated with ground-truth failure severity within cyclic compositions. The other baselines fare worse: B2 probing produces NaN at all scales (constant 1.0 accuracy); B5 ensemble becomes anti-correlated at $n \geq 15$.

Linear scaling of structural advantage. The structural discrimination metric grows approximately linearly with β_1: $34.2, 81.9, 117.7, 222.9$. Normalizing by cycle count gives $\approx 34 \cdot \beta_1$, consistent with the theoretical prediction that structural advantage scales with the number of independent obstruction classes ($\dim H^1$).

B5 ensemble inversion. The gradient-boosted ensemble (B5), which combines all interpretability features, becomes anti-correlated with failure severity at large scale. This is a stronger result than mere decorrelation: the interpretability features contain structure that actively misleads a learned predictor. The ensemble overfits to edge-local patterns that are inversely correlated with the cycle-level failure mode at scale.

8.4 Finding 4: Cross-Domain Replication

Table 4 reports ρ_{cyclic} for the best interpretability baseline and the structural diagnostic across three domains at two scales ($n = 5$, $n = 10$).

Domain	n	**Best Interp.**	**Structural**	**Gap**
Invoice	5	0.685 (B1 SAE)	1.000	0.315
Invoice	10	0.624 (B4 CKA)	1.000	0.376
Calendar	5	0.721 (B1 SAE)	1.000	0.279
Calendar	10	0.685 (B1 SAE)	1.000	0.315
Policy	5	0.394 (B3 Attn)	1.000	0.606
Policy	10	0.745 (B1 SAE)	1.000	0.255

Table 4: Phase 2c: Cross-domain replication. The structural diagnostic achieves $\rho_{\text{cyclic}} = 1.0$ in every condition. B1 SAE divergence is the strongest interpretability baseline in 5 of 6 conditions. The gap is always ≥ 0.255.

Domain invariance. The structural diagnostic achieves perfect rank correlation ($\rho_{\text{cyclic}} = 1.0$) in all six domain–scale conditions. No interpretability baseline exceeds $\rho_{\text{cyclic}} = 0.745$.

B1 SAE hierarchy. SAE divergence is the strongest interpretability baseline in 5 of 6 conditions. The exception is policy at $n = 5$, where B3 attention entropy leads ($\rho_{\text{cyclic}} = 0.394$)—the weakest best-baseline performance across all conditions. Different convention structures favor different interpretability methods, but none consistently approaches the structural diagnostic.

Calendar probing anomaly. B2 probing produces NaN in invoice and policy (constant 1.0 accuracy) but shows some signal on calendar: $\rho_{\text{all}} = 0.486$ at $n = 5$ and 0.574 at $n = 10$; $\rho_{\text{cyclic}} = 0.067$ and 0.467. Calendar's convention structure (timezone offsets, DST rules) introduces slightly more probe-detectable variance, but $\rho_{\text{cyclic}} = 0.467$ at best remains well below the structural diagnostic.

Tightest margin. The narrowest gap occurs at policy $n = 10$, where B1 SAE achieves $\rho_{\text{cyclic}} = 0.745$ (gap = 0.255). Even in this most favorable condition, the structural method's advantage is substantial and would be statistically significant under bootstrap testing.

8.5 Case Study: LocalRightGlobalWrong

Consider a 5-node cyclic invoice composition from Phase 2a. At every edge, the B2 probing classifier identifies the convention correctly (99.8% accuracy). B1 SAE divergence detects some representational mismatch but assigns moderate scores to all edges, including those in the acyclic comparison set. B4 CKA reports normal representation similarity at each interface. Every edge-local diagnostic gives the system a clean bill of health.

The structural diagnostic computes holonomy around the single cycle: the product of convention-transformation restriction matrices fails to return to the identity by a Frobenius norm of 0.47. The composition is predicted to fail, and it does. The semantic output—the final ledger entry—is wrong by a factor traceable to the accumulated convention drift around the cycle.

This case instantiates the abstract prediction of the Edge-Local Blindness Lemma: two globally different executions (one consistent, one drifted) produce identical observations on every edge. The failure is invisible to any diagnostic that operates at the edge level.

8.6 Finding 5: Graph-Level Aggregation Does Not Close the Gap

The B6 baselines give interpretability features explicit cycle-level reasoning and isolate a precise question: is the gap caused by weak local aggregation, or by absence of cycle-relevant signal in the features themselves?

Domain	n	B6a	B6b	Best B1–5	Structural	Gap
Invoice	5	0.685	−0.030	0.685	1.000	0.315
Invoice	10	0.515	−0.248	0.624	1.000	0.485
Calendar	5	0.721	−0.200	0.721	1.000	0.279
Calendar	10	0.721	−0.333	0.685	1.000	0.279
Policy	5	−0.079	0.176	0.394	1.000	0.824
Policy	10	0.758	−0.224	0.745	1.000	0.242

Table 5: Phase 3: Graph-level interpretability baselines (ρ_{cyclic}). B6a aggregates edge features around detected cycles with oracle topology knowledge. B6b trains a gradient-boosted regressor on cycle-aggregated features. "Gap" is structural minus the best of B6a and B6b.

Cycle-oracle aggregation adds nothing (B6a). B6a is the decisive result. Its ρ_{cyclic} matches B1's within rounding in 4 of 6 conditions; the maximum improvement is +0.036 (calendar $n = 10$). This occurs because all cyclic compositions within a condition share the same topology: summing B1 SAE divergence around cycle edges is rank-equivalent to averaging across all edges. Even with *oracle knowledge* of which edges form cycles, the features cannot distinguish which compositions fail. The graph-structure objection is removed, and nothing changes.

Learned cycle-aware aggregation is unstable (B6b). B6b—a gradient-boosted regressor on 12-dimensional cycle-aggregated features with leave-one-out cross-validation—produces negative ρ_{cyclic} in 4 of 6 conditions (range: -0.333 to 0.176). In a weak-signal regime, learned aggregation over edge-local features is not merely unhelpful but unstable: the aggregator overfits to patterns that are inversely correlated with the ground-truth failure mode.

Representation insufficiency, not aggregation failure. The B6 result resolves the cleanest remaining escape hatch. The gap between interpretability and structural diagnostics is not an aggregation problem. It is a representation problem. For cyclic compositional systems, the limiting factor is not how edge-local interpretability features are aggregated, but what they fail to represent. The structural diagnostic wins because it operates on different data—convention metadata (field types, unit scales, format specifications)—that directly encodes the convention transformations whose product around cycles determines failure.

This confirms Proposition 3.3: cycle holonomy is not $\mathcal{F}_{\text{edge}}$-measurable, and no predictor over $\mathcal{F}_{\text{edge}}$-measurable features—however aggregated—can recover it.

8.7 Finding 6: Cross-Model Replication (Gemma 2 2B)

Table 6 reports replication on Gemma 2 2B (2.6B parameters, 26 layers, $d_{\text{model}} = 2304$)—a model 20× larger than GPT-2 Small, from a different architecture family (Google vs. OpenAI), with independently trained SAE features (gemma-scope-2b-pt-res-canonical, layer 20, 16k features). Experiments use the same composition generation, structural diagnostic, and evaluation protocol as Phases 2a–2b.

Model	n	B1 SAE	B2 probe	Structural	Gap
GPT-2 Small (124M)	5	0.685	0.998	1.000	0.315
GPT-2 Small (124M)	10	0.515	0.959	1.000	0.485
Gemma 2 2B (2.6B)	5	0.515	1.000	1.000	0.485
Gemma 2 2B (2.6B)	10	0.467	0.995	1.000	0.533

Table 6: Cross-model replication (invoice domain, ρ_{cyclic}). B2 probe accuracy reports mean LOO-CV convention classification accuracy across edges. "Gap" is structural minus B1 SAE ρ_{cyclic}. The gap is *larger* on Gemma 2 than on GPT-2 at both scales.

The gap replicates and widens. On Gemma 2 2B, B1 SAE divergence achieves $\rho_{\text{cyclic}} = 0.515$ at $n = 5$ and 0.467 at $n = 10$. These are *lower* than the corresponding GPT-2 values (0.685 and 0.515), meaning the gap is larger on the bigger model. The structural diagnostic remains at $\rho = 1.0$ at both scales.

Perfect local knowledge persists. B2 probing accuracy is 1.000 at $n = 5$ (perfect) and 0.995 at $n = 10$. A 20× larger model with different architecture still achieves near-perfect convention classification at every edge—and this perfect local knowledge still carries negligible predictive value for compositional failure ($\rho_{\text{cyclic}} = \text{NaN}$ at $n = 5$ due to constant output; 0.174 at $n = 10$).

The cliff onset replicates. B1 SAE ρ_{cyclic} decays from 0.515 to 0.467 between $n = 5$ and $n = 10$, consistent with the monotonic decline observed on GPT-2 ($0.685 \to 0.515$).

Implication. The edge-local interpretability boundary is not a GPT-2 artifact. A model with 20× more parameters, different architecture, and independently trained SAE features shows the same pattern: perfect local knowledge, zero global signal, and a gap that grows with scale. This is consistent with Proposition 3.3: the limitation is in the *feature class*, not the model.

9 Discussion

9.1 Representation Insufficiency, Not Tool Failure

Mechanistic interpretability is not refuted. The edge-local interpretability feature class is the wrong representation for this particular diagnostic target. Two results make this precise:

1. **B2 (probing):** 99.8% accuracy at classifying conventions on every edge. The model's internal representations *do* encode the relevant local information. The failure is not in the model's understanding but in the *composition* of edge-local understanding around cycles.
2. **B6a (cycle-oracle):** Even with oracle knowledge of which edges form cycles, aggregating interpretability features around those cycles does not improve prediction. The cycle-relevant signal is not hidden by weak aggregation; it is absent from the features.

Proposition 3.3 gives this a formal basis: the cycle holonomy invariant is not measurable in the σ-algebra generated by edge-local observations. The experiment is the empirical witness of the proposition's practical relevance: the theorem says why the feature class should fail; the six baseline families show that it does fail, across three domains, multiple scales, and two model architectures.

The result does not say interpretability is useless for multi-agent systems. On acyclic compositions, all methods correctly diagnose zero failure. On individual model debugging, interpretability remains indispensable. The claim is bounded to a specific feature class and diagnostic target: for *compositional semantic failure in cyclic systems*, the edge-local interpretability feature class does not contain the relevant information, and no downstream aggregator over that class can recover it. Cross-model replication on Gemma 2 2B (Section 8.7) confirms this is not architecture-specific: the gap widens on a 20× larger model from a different family.

The microscope and the telescope. Interpretability is the microscope: it reveals the internal structure of components with increasing precision—and our probing results show it does so nearly perfectly. Compositional coherence requires a telescope: an instrument that resolves the global structure of multi-component systems from metadata about their relationships. These are complementary, not competing, instruments (Figure 1). The contribution of this paper is to identify empirically where one instrument must yield to the other: at $\beta_1 \geq 1$, with the transition sharpening monotonically as β_1 grows.

9.2 Implications for Agent Interoperability

Current agent interoperability protocols (MCP, A2A, LangChain tool interfaces) treat composition as a connectivity problem. If schemas match and transport works, the system is assumed correct. BABEL has shown this assumption fails at modest scale on deterministic compositions. This paper extends the observation to live LLM inference with a quantitative result: even with full access to model internals, the strongest interpretability baseline (B1 SAE divergence) achieves at best $\rho_{\text{cyclic}} = 0.745$, while a metadata-only structural diagnostic achieves $\rho = 1.000$ in every condition tested.

The practical implication is concrete. A monitoring system that instruments individual agents—however deeply—will miss compositional failures at a rate that increases with the topological complexity of the agent graph. For multi-runtime agent coordination—where independent agent systems coordinate without a shared orchestrator—interoperability requires structural diagnostics at the composition level, not deeper introspection of individual agents.

9.3 Connection to M1–M2

The M1–M2 extraction problem—recovering stable compositional semantic structures from fuzzy, probabilistic LLM outputs—remains open and is not a dependency of this paper. The present experiment uses compositions where convention structure is planted and known. Whether the structural diagnostic can be extended to regimes where convention structure must be *inferred* from model behavior is a frontier question that this paper identifies but does not attempt to resolve.

10 Limitations and Falsifiers

10.1 Explicit Falsification Conditions

1. **Interpretability remains competitive on cyclic graphs.** *Status: falsified by data.* The best interpretability baseline achieves at most $\rho_{\text{cyclic}} = 0.745$ (B1 SAE, policy $n = 10$) while the structural diagnostic achieves $\rho = 1.0$ in every condition. The gap widens with β_1 (Section 8.3), confirming the formal prediction is empirically load-bearing.

2. **Graph aggregation closes the gap.** *Status: falsified by data.* The B6 baseline gives interpretability features oracle knowledge of cycle structure and a learned aggregator (Section 8.6). B6a (cycle-oracle) matches B1 within rounding; B6b (learned) goes anti-correlated in 4 of 6 conditions. The limitation is not in aggregation but in the features themselves: the missing signal is absent from the extracted representation.

3. **Results only hold in heavily planted regimes.** *Status: partially addressed.* Cross-domain replication (Section 8.4) shows the gap persists across three semantically distinct domains with different convention structures. The conventions are still planted rather than inferred, which limits generalizability to naturalistic deployments.

4. **Results only hold on one model architecture.** *Status: falsified by data.* Gemma 2 2B (2.6B parameters, 26 layers, different architecture family) replicates the gap at both $n = 5$ and $n = 10$ (Section 8.7). The gap is *larger* on Gemma 2 than on GPT-2: $\rho_{\text{cyclic}} = 0.515$ vs. 0.685 at $n = 5$ and 0.467 vs. 0.515 at $n = 10$. A 20× scale increase and architectural change does not close the gap; it widens it.

10.2 Scope Limitations

- **Two model architectures.** Primary experiments use GPT-2 Small (124M parameters, 12 layers); cross-model replication uses Gemma 2 2B (2.6B parameters, 26 layers). Both show the same gap pattern (Section 8.7). Our claim is about the *edge-local feature class*, not about model scale. A model that internally represented cycle-level compositional structure would escape Proposition 3.3 by escaping $\mathcal{F}_{\text{edge}}$—its features would no longer be edge-local. Whether models at the 70B+ frontier produce emergent compositional representations that are no longer cleanly edge-local remains the strongest open question. The Gemma 2 replication, which *widens* the gap at 20× scale, provides initial evidence against the scale-escape hypothesis.
- **Open-weight models only.** Proprietary models (GPT-4, Claude, Gemini Pro) cannot be included because interpretability extraction requires weight access. Results may not transfer to proprietary model compositions.
- **Planted convention structure.** Convention mismatches are known and planted. In real deployments, conventions must be inferred or declared—the M1–M2 problem. This paper does not address convention extraction.

- **Scale ceiling at $n = 20$.** SAE extraction on 20-node compositions with 4 independent cycles was feasible on a single A100. Larger compositions ($n = 30$–50) would require multi-GPU extraction or a lighter SAE. The theoretical argument from the Edge-Local Blindness Lemma predicts the cliff continues; the linear scaling of structural discrimination with β_1 supports this (Section 8.3), but direct empirical confirmation at larger scale remains future work.
- **Coupling parameter sensitivity.** The restriction matrices use an off-diagonal coupling coefficient $\varepsilon = 0.01$ (Appendix C). Sensitivity analysis across $\varepsilon \in \{0.001, 0.005, 0.01, 0.02, 0.05, 0.1\}$ shows the gap is invariant: $\rho_{\text{B1,cyclic}}$ remains unchanged at every tested ε, because changing coupling strength scales holonomy magnitudes but preserves the ranking of compositions. At $\varepsilon = 0$ the holonomy is identically zero (diagonal matrices telescope to identity around any cycle), confirming that some coupling is necessary for non-trivial obstruction, but the qualitative result is robust once any coupling exists.
- **No naturalistic multi-runtime deployment.** All compositions are research-grade single-machine pipelines. Live multi-runtime agent-to-agent coordination introduces additional failure modes (latency, partial observation, strategic behavior) not tested here.
- **No repair evaluation.** The experiment measures *diagnosis*, not *repair*. Whether structural diagnosis enables better repair prescriptions than interpretability-guided local fixes is a separate question for future work.

11 Conclusion

Across 240+ compositions, three domains, four scales, six interpretability baseline families—including one with oracle knowledge of cycle topology—and two model architectures spanning a 20× parameter range, edge-local interpretability features are not the right diagnostic object for cyclic compositional failure. The structural diagnostic, using only composition metadata and no model internals, achieves perfect rank correlation ($\rho = 1.0$) in every condition tested. The strongest interpretability baseline never exceeds $\rho = 0.758$ on cyclic compositions, even with graph-level aggregation, and its predictive quality decays monotonically with scale until it becomes anti-correlated.

Two results make the finding precise. The B2 probing result: 99.8% classification accuracy at every edge, zero predictive value for compositional failure—the model knows what convention each component uses, and the composition still fails. The B6a result: even with oracle knowledge of cycle topology, aggregating interpretability features around cycles does not improve prediction beyond edge-local averaging. The gap is not an aggregation problem. It is a representation problem.

For cyclic compositional systems, the limiting factor is not how edge-local interpretability features are aggregated, but what they fail to represent (Proposition 3.3). This is a boundary on the *edge-local interpretability feature class*, not a verdict on interpretability as a field. On acyclic compositions, all methods correctly diagnose zero failure. The Interpretability Cliff (Section 8.3) identifies the precise boundary: $\beta_1 \geq 1$, with urgency proportional to $\dim H^1$.

The strongest deployment of both paradigms would combine component-level interpretability (where it excels) with cycle-level structural diagnosis (where it becomes necessary). The practical implication is a norm:

> *Any claim that an interpretability method diagnoses compositional reliability in multi-agent or multi-tool systems should be tested under cyclic stress and compared against an explicit structural diagnostic.*

The edge-local feature class describes what interpretability tools extract on both GPT-2 Small and Gemma 2 2B—models separated by 20× in parameter count and drawn from different

architecture families. Cross-model replication (Section 8.7) shows the gap *widens* rather than narrows with scale, providing initial evidence against the hypothesis that larger models escape the edge-local feature class. The formal argument (Proposition 3.3) does not depend on model scale: it depends on whether the extracted features are $\mathcal{F}_{\text{edge}}$-measurable. Compositional failure requires compositional diagnosis.

References

[1] J. Komkov. The Coherence Fee: Edge-Local Blindness at the String-Table Seam and the Topological Price of Cross-System Composition. *Res Agentica Program*, February 2026.

[2] J. Komkov. BABEL: A Benchmark for Compositional Coherence in Multi-Agent Systems. *Res Agentica Program*, March 2026.

[3] Res Agentica Program. The Coherence Cliff: A Scaling Experiment on the Necessity of Sheaf-Cohomological Diagnostics in Multi-Agent Composition. March 2026.

[4] J. Komkov. SCPI: Predicate Invention Under Sheaf Constraints. *Res Agentica Program*, 2026.

[5] N. Elhage, T. Hume, C. Olsson, et al. Toy models of superposition. *Transformer Circuits Thread*, 2022.

[6] T. Bricken, A. Templeton, J. Batson, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Anthropic*, 2023.

[7] N. Nanda. TransformerLens: A library for mechanistic interpretability of GPT-style language models. 2022.

[8] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton. Similarity of neural network representations revisited. *ICML*, 2019.

[9] N. Kriegeskorte, M. Mur, and P. Bandettini. Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2(4), 2008.

[10] A. Conmy, A. Mavor-Parker, A. Lynch, S. Heimersheim, and A. Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *NeurIPS*, 2023.

A Model and Tooling Stack

- **Model:** GPT-2 Small, 124M parameters, 12 layers, $d_{\text{model}} = 768$. Loaded via TransformerLens (`HookedTransformer.from_pretrained("gpt2-small")`).
- **SAE:** Release `gpt2-small-resid-post-v5-32k`, hook `blocks.11.hook_resid_post`, $d_{\text{sae}} = 32{,}768$. Loaded via SAELens `SAE.from_pretrained`.
- **Residual extraction:** Layers 0, 5, and 11 via TransformerLens `run_with_cache`. Mean-pooled over sequence length per prompt.
- **GPU:** NVIDIA T4 (16 GB, Phase 1), A100 (40 GB, Phases 2a–2c). Google Colab Pro+.

B Baseline Implementation Details

B1: SAE divergence. For each edge (u, v), encode source and destination prompts through the model, extract SAE activations at layer 11, mean-pool across tokens. Compute cosine distance plus Jensen–Shannon divergence between the two activation vectors. Report the sum as the edge-level feature.

B2: Probing classifier. Extract residual-stream activations at layers 0, 5, and 11. For each layer, train a pipeline of PCA($n_{\text{comp}} = \min(5, N-2)$) followed by logistic regression ($C = 0.01$, `max_iter` $= 500$) under leave-one-out cross-validation. Report the maximum accuracy across layers as the edge-level feature.

B3: Attention entropy. Compute mean attention entropy (over heads and sequence positions) at layers 5 and 11 for source and destination prompts. Report the maximum absolute difference across layers as the edge-level feature.

B4: CKA dissimilarity. Compute linear CKA (centered kernel alignment) between source and destination residual-stream activations at layer 11. Report $1 - \text{CKA}$ as the edge-level feature.

B5: Gradient-boosted ensemble. Stack all features from B1–B4 per edge. Train a `GradientBoostingRegr` ($n_{\text{estimators}} = 50$, `max_depth` $= 3$) to predict edge-level failure contribution, with leave-one-out cross-validation for predicted values. Report the cross-validated predicted value.

C Composition Generation and Ground Truth

Node conventions. Each node draws a convention tuple: an amount multiplier (e.g., 1.0 for dollars, 100.0 for cents, 10,000.0 for basis points) and a settlement timing offset (0, 1, or 2 days). For calendar and policy domains, analogous convention parameters apply (see Section 4).

Restriction matrices. For each edge (u, v), the 2×2 restriction matrix encodes the convention transformation:

$$R_{u \to v} = \begin{pmatrix} s_u / s_v & \epsilon \, |d_u - d_v| \, (1 - r) \\ 0.5 \, \epsilon \, |d_u - d_v| \, (1 - r) & (1 + d_u)/(1 + d_v) \end{pmatrix}$$

where s is the scale multiplier, d is the timing offset, $r = \min(s_u, s_v)/\max(s_u, s_v)$, and $\epsilon = 0.01$ controls off-diagonal coupling. The off-diagonal terms ensure that cyclic compositions yield nontrivial holonomy (diagonal-only matrices would telescope to the identity around any cycle).

Holonomy computation. For a composition with independent cycles $C_1, \ldots, C_k$, the ground-truth holonomy is:

$$h = \sum_{i=1}^{k} \left\| \prod_{(u,v) \in C_i} R_{u \to v} - I_2 \right\|_F$$

Acyclic compositions have $h = 0$ by construction.

Topology construction. At scale n, cyclic compositions contain $\lfloor n/5 \rfloor$ independent 5-node cycles (each formed by adding a back-edge from node $5(j+1) - 1$ to node $5j$). Acyclic compositions use the same node set with a linear chain topology ($\beta_1 = 0$).