WHEN TREES ARE NOT ENOUGH: LEARNING MIXED-TOPOLOGY FEATURE GRAPHS WITH ADAPTIVE GRAPH SPARSE AUTOENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse autoencoders (SAEs) expose interpretable features in large language model activations, yet existing structured SAEs impose single-parent trees or forests, while post-hoc graphs permit multiple parents but neither guide feature learning nor ensure reliable relation recovery. We introduce the Adaptive Graph Sparse Autoencoder (AG-SAE), a structure-guided training paradigm that treats each feature's complete parent set as an atomic structural hypothesis and lets evidence select zero, one, or multiple parents. By competing complete parent sets against null, subset, and alternative explanations, AG-SAE identifies jointly necessary multi-parent relations while rejecting redundant or spurious alternatives and verifying that each child contributes beyond its parents. The induced topology over SAE features then defines a differentiable structural loss that guides SAE training, while topology-guided refinement mitigates feature absorption and uses persistent reconstruction gaps exposed by the learned structure to initialize new features. The entire graph is then induced again from the revised dictionary by reassessing every feature's complete parent set, closing the dictionary–graph self-consistency cycle. Experiments demonstrate exact mixed-topology recovery in a controlled toy model, greater relational reliability and semantic validity than structured and post-hoc baselines on real LLM activations, and stronger feature-level causal interventions than conventional SAE features. AG-SAE thereby turns recovered mixed-topology feature structure into an unsupervised training signal that improves the dictionary, enables reliable feature organization beyond the topological limitations of trees, and exhibits stronger causal control beyond reconstruction.

1 INTRODUCTION

Understanding how large language models (LLMs) represent and organize knowledge requires identifying interpretable features within their entangled high-dimensional activations and recovering the relations among those features. Sparse autoencoders (SAEs) provide such features by decomposing activations into sparse combinations from an overcomplete dictionary, but conventional SAEs learn flat dictionaries that reveal which features exist, not how they relate (Bricken et al., 2023; Huben et al., 2024). As SAE capacity increases, broad concepts often split into finer-grained features (Bussmann et al., 2025). Hierarchical SAEs capture these levels of abstraction with nested or multiscale dictionaries, but primarily model hierarchy across dictionary scales rather than relations between individual features (Zaigrajew et al., 2025; Bussmann et al., 2025). Structured SAEs go further by learning feature-level parent–child relations between broad and specialized features (Luo et al., 2026; Cao et al., 2026), while typically imposing single-parent trees or forests that assign each child to only one parent.

However, post-hoc analyses of hierarchical SAEs have found multi-parent associations (Bussmann et al., 2025), and tree-structured SAE studies have acknowledged the limitations of the single-parent constraint (Luo et al., 2026; Cao et al., 2026). Taken together, this evidence indicates that a single-parent tree is too restrictive a prior for feature organization. When a child requires multiple parents to explain its representation, any single-parent assignment omits necessary dependencies and distorts its semantic basis (Figure 1); accumulated across a hierarchy, such omissions can mislead feature navigation and mechanistic analysis. In addition, standard metrics for reconstruction, sparsity, and individual feature quality do not assess relational fidelity (Gao et al., 2025; Karvonen et al., 2025). A structured SAE may score well while recovering an incomplete or misleading structure. Good feature recovery therefore does not guarantee reliable structure recovery.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

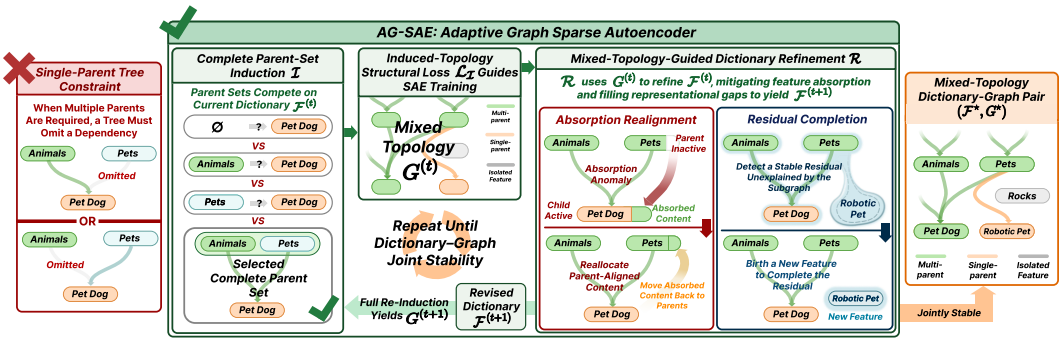

Figure 1: **Overview of AG-SAE.** The induced topology guides SAE training and dictionary refinement; Full Re-Induction closes the cycle, which repeats until the dictionary–graph pair stabilizes. This example is illustrative.

Relaxing the single-parent constraint and adding pairwise edges post hoc merely expands the hypothesis space without ensuring reliable structure recovery. Activation co-occurrence, decoder similarity, local fit gains, redundancy, compositionality, and transitive ancestry can all induce spurious parent–child relations (Grandien et al., 2026). Multi-parent relations must therefore be identified as complete parent sets rather than assembled from independent edges. Yet moving to complete parent sets remains challenging, because a proposed set may still be displaced by a more parsimonious or equally plausible explanation. A child feature may also be a redundant copy of its parent features, carrying no new or more specific information, while a numerically supported association need not correspond to a semantically coherent broad-to-specific relation. Existing methods remain unable to resolve these ambiguities from LLM activations and recover reliable feature structure without imposing a single-parent topology (Bussmann et al., 2025; Luo et al., 2026; Cao et al., 2026).

We therefore introduce the Adaptive Graph Sparse Autoencoder (AG-SAE), a structure-guided SAE training paradigm that treats feature structure as both the object to be recovered and an unsupervised signal for feature formation. AG-SAE couples dictionary and graph learning through a dictionary–graph self-consistency cycle. Given the current dictionary, it assigns each child a parent set by competing complete parent sets against null, subset, and alternative explanations, allowing isolated features, single-parent branches, and multi-parent refinements to emerge without a fixed topology. This distinguishes jointly necessary multi-parent refinements from redundant or spurious relations and verifies that each child retains a nonredundant contribution beyond its parents. The induced topology then guides feature-specific encoder and decoder updates, continuously reallocating parent-aligned content and using persistent reconstruction residuals exposed by the learned structure to seed new features where representational gaps remain, thereby expanding the dictionary. After each update, AG-SAE globally re-induces the feature graph under the revised dictionary, closing the dictionary–graph self-consistency cycle. Figure 1 summarizes this structure-guided refinement cycle. Experiments show that AG-SAE recovers feature structure with greater relational and semantic fidelity than structured SAE and post-hoc baselines (Section 4.3). Beyond these structural gains, we further find that AG-SAE exhibits a non-trivial capability; across both feature addition and feature ablation (Templeton et al., 2024; Marks et al., 2025), features learned under this paradigm support substantially more effective causal interventions than conventionally trained SAE features (Section 4.4). **Taken together, AG-SAE not only learns a reliable mixed-topology feature graph from SAE representations, but also turns that structure into a training signal that improves the dictionary and enables capabilities beyond reconstruction.** Our contributions are as follows:

1. We formulate feature structure recovery as a complete parent-set assignment problem in which each feature may have zero, one, or multiple parents. Comparing each complete parent set against null, subset, and alternative explanations distinguishes jointly necessary multi-parent refinements from redundant or spurious relations and separates inherited parent content from child-specific contributions. The resulting evidence determines each feature's parent count, extending structure recovery beyond single-parent trees.
2. We introduce AG-SAE, a structure-guided SAE training paradigm that uses recovered feature structure as an unsupervised signal for feature formation. Within its dictionary–graph self-consistency cycle, the induced topology guides feature-specific encoder and decoder updates that continuously reallocate parent-aligned content, while persistent graph-

conditioned residuals seed new features that fill representational gaps and expand the dictionary. AG-SAE then globally re-induces all parent sets under the revised dictionary, coupling dictionary optimization to complete structural reassessment.

3. Through experiments, we show that **(a)** AG-SAE achieves reliable mixed-topology recovery, exactly matching all ground-truth complete parent sets in the toy model and producing graphs with greater relational and semantic fidelity than structured SAE and post-hoc graph baselines on real activations. A concrete case further illustrates joint parent support and nonredundant child content; **(b)** AG-SAE's Complete Parent-Set Induction improves multi-parent structural reliability, and its Absorption Realignment substantially reduces feature absorption; and **(c)** in both feature addition and feature ablation, AG-SAE features support substantially more effective causal interventions than conventionally trained SAE features.

2 Related Work and Positioning

Structured SAEs explicitly learn relations between individual features, but restrict each child to a single parent (Muchane et al., 2025; Luo et al., 2026; Cao et al., 2026). Post-hoc analyses reveal multi-parent associations, yet co-activation and redundancy can produce spurious relations (Bussmann et al., 2025; Grandien et al., 2026). The gap is extending SAE training beyond single-parent trees while reliably distinguishing jointly necessary parent sets from redundant or spurious associations. AG-SAE addresses this gap by evaluating complete parent sets against null, subset, and alternative explanations. The induced topology then guides dictionary training and refinement, followed by full re-induction under the revised dictionary. Appendix A provides the full discussion.

3 Adaptive Graph Sparse Autoencoder

The Adaptive Graph Sparse Autoencoder (AG-SAE) casts the learning of feature structure as a self-consistency problem, jointly developing an SAE feature dictionary and a mixed-topology graph over the relations among its features through iterative refinement cycles.

Given an activation domain $\mathcal{X} \subseteq \mathbb{R}^d$ associated with a fixed LLM layer, let $\mathbf{x} \in \mathcal{X}$ denote an activation vector drawn from distribution $\mathcal{D}$ and used as an SAE input during dictionary training. For each feature i, $z_i(\mathbf{x})$ denotes its activation coefficient, and $\mathbf{d}_i \in \mathbb{R}^d$ denotes its unit-norm decoder direction, with $\|\mathbf{d}_i\|_2 = 1$. Let $\mathbf{b} \in \mathbb{R}^d$ denote the reconstruction bias. For an SAE with M features, the reconstruction is written as $\widehat{\mathbf{x}} = \sum_{i=1}^{M} z_i(\mathbf{x})\mathbf{d}_i + \mathbf{b}$. For notational simplicity, we omit the reconstruction bias from the subsequent algorithmic development; see Appendix B.

At refinement cycle t, let M_t denote the number of current features. We use c and p to index child and candidate-parent features, respectively. For a given child c, a candidate parent set $P \subseteq \{1, \ldots, M_t\} \setminus \{c\}$ may satisfy $P = \varnothing$ or contain one or multiple direct parents, with each $p \in P$ denoting one candidate parent. Let $P_c^{(t)}$ denote the parent set ultimately assigned to feature c at refinement cycle t. We then write the global SAE dictionary state and the corresponding mixed-topology graph state as

$$\mathcal{F}^{(t)} \coloneqq \left\{ \left(z_i^{(t)}(\cdot), \mathbf{d}_i^{(t)} \right) \right\}_{i=1}^{M_t}, \qquad G^{(t)} \coloneqq \left(P_c^{(t)} \right)_{c=1}^{M_t}.$$

Thus, $\mathcal{F}^{(t)}$ specifies the current SAE dictionary, while $G^{(t)}$ specifies the relational organization among its features. **As illustrated in Figure 1, the AG-SAE refinement cycle proceeds as follows: (i) Complete Parent-Set Induction $\mathcal{I}$.** Starting from the current dictionary $\mathcal{F}^{(t)}$, $\mathcal{I}$ jointly evaluates complete parent sets rather than independent parent–child relations, allowing zero-, single-, and multi-parent structures to compete and yielding $G^{(t)}$ (Section 3.1). **(ii) Structure-Guided SAE Training.** In addition to the conventional SAE reconstruction objective $\mathcal{L}_{\text{rec}}$ (Bricken et al., 2023), AG-SAE introduces a graph-guided structural loss $\mathcal{L}_{\mathcal{I}}$, the differentiable training counterpart of the relational criteria used by Complete Parent-Set Induction $\mathcal{I}$. Starting from $\mathcal{F}^{(t)}$ and using $G^{(t)}$ as the structural reference for this update, the dictionary is optimized with

$$\mathcal{L}_{\text{rec}}\left(\mathcal{F}^{(t)}\right) + \mathcal{L}_{\mathcal{I}}\left(\mathcal{F}^{(t)}; G^{(t)}\right). \tag{1}$$

The structural loss encourages the learned representations to better support the relations selected by $\mathcal{I}$, thereby feeding the induced structure back into dictionary optimization and driving the dictionary–graph pair toward self-consistency. For conciseness and to avoid redundancy, the underlying relational criteria are detailed in Section 3.1, with the differentiable formulation provided in Appendix B.2. **(iii) Mixed-Topology-Guided Dictionary Refinement $\mathcal{R}$.** Beyond the global

structure-guided training above, $\mathcal{R}$ further refines the dictionary under the guidance of $G^{(t)}$ by realigning absorbed parent-related representation and filling persistent representational gaps, producing $\mathcal{F}^{(t+1)}$ (Section 3.2). **(iv) Full Re-Induction.** Finally, reapplying $\mathcal{I}$ to $\mathcal{F}^{(t+1)}$ yields the next graph $G^{(t+1)}$, and the cycle repeats until the dictionary–graph pair stabilizes (Section 3.3). The overall AG-SAE refinement cycle is therefore summarized as

$$\mathcal{F}^{(t)} \xrightarrow{\mathcal{I}} G^{(t)} \xrightarrow{\mathcal{L}_{\text{rec}}+\mathcal{L}_{\mathcal{I}} \to \mathcal{R}} \mathcal{F}^{(t+1)} \xrightarrow{\mathcal{I}} G^{(t+1)}.$$

3.1 Complete Parent-Set Induction $\mathcal{I}$

Complete Parent-Set Induction $\mathcal{I}$ induces the graph $G^{(t)}$ from the current dictionary $\mathcal{F}^{(t)}$ by jointly evaluating each candidate parent set P, rather than deciding each parent–child relation separately, which empirically improves structural reliability (Section 4.5). We omit the cycle superscript (t) below. The indicator $a_i(\mathbf{x})$ records whether feature i is active, while $a_P(\mathbf{x})$ indicates whether all features in P are jointly active:

$$a_i(\mathbf{x}) \coloneqq \mathbf{1}\{z_i(\mathbf{x}) > 0\}, \qquad a_P(\mathbf{x}) \coloneqq \prod_{p \in P} a_p(\mathbf{x}).$$

Parent-Set Support for the Child. To ensure that the candidate parent set P sufficiently covers the activation regime of child c, AG-SAE requires $\Pr(a_P = 1 \mid a_c = 1) \geq \tau_{\text{cov}}$. Beyond activation coverage, AG-SAE requires the candidate parents to jointly match the child's representation, measured by $\mathcal{S}_{\text{rep}}(c, P)$, which compares the child representation $z_c(\mathbf{x})\mathbf{d}_c$, combining its direction and activation strength, with the joint parent representation $\sum_{p \in P} z_p(\mathbf{x})\mathbf{d}_p$:

$$\mathcal{S}_{\text{rep}}(c, P) \coloneqq 1 - \frac{\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[\left\| z_c(\mathbf{x})\mathbf{d}_c - \sum_{p \in P} z_p(\mathbf{x})\mathbf{d}_p \right\|_2^2 \,\middle|\, a_c(\mathbf{x}) = 1\right]}{\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[\left\| z_c(\mathbf{x})\mathbf{d}_c \right\|_2^2 \,\middle|\, a_c(\mathbf{x}) = 1\right]} \geq \tau_{\text{rep}}. \tag{2}$$

A higher score $\mathcal{S}_{\text{rep}}(c, P)$ indicates that the candidate parent set more closely matches the child's representation. Since $\mathcal{S}_{\text{rep}}$ evaluates the candidate set P jointly, parent sets with different cardinalities can compete under the same criterion. A candidate P is retained only if its score exceeds those of the competing sets by a margin. This enables AG-SAE to extend beyond the single-parent restriction while recovering more reliable structure than independently assembled parent–child relations (Section 4.5). Details are provided in Appendix B.1.

Child Contribution Beyond Parents. We further require c to provide reconstruction contribution beyond its complete parent set. Let $\mathcal{E}_{c,P}(S)$ denote the reconstruction error from feature set S, evaluated on samples where c and all features in P are jointly active:

$$\begin{aligned} \mathcal{E}_{c,P}(S) &\coloneqq \min_{\boldsymbol{\beta} \geq \mathbf{0}} \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[\left\| \mathbf{x} - \sum_{i \in S} \beta_i z_i(\mathbf{x})\mathbf{d}_i \right\|_2^2 \,\middle|\, a_c(\mathbf{x})a_P(\mathbf{x}) = 1\right], \\ \mathcal{S}_{\text{inn}}(c, P) &\coloneqq \frac{\mathcal{E}_{c,P}(P) - \mathcal{E}_{c,P}(P \cup \{c\})}{\mathcal{E}_{c,P}(\varnothing)} \geq \tau_{\text{inn}}. \end{aligned} \tag{3}$$

The score $\mathcal{S}_{\text{inn}}(c, P)$ measures the reduction in reconstruction error when the child c is added to its parent set, comparing reconstruction with P alone against $P \cup \{c\}$. The denominator $\mathcal{E}_{c,P}(\varnothing)$ normalizes this gain by the zero-reconstruction error. A higher score therefore indicates that the child provides additional representational value beyond P. Together with parent-set support, this helps reject redundant or spurious parent–child relations. Fitting details are provided in Appendix B.1. Together, these criteria allow $\mathcal{I}$ to select one parent set per child under global acyclicity, yielding $G = \mathcal{I}(\mathcal{F})$. The resulting graph, together with these criteria, defines the structural signal $\mathcal{L}_{\mathcal{I}}$ that guides dictionary optimization (Appendix B.2).

3.2 Mixed-Topology-Guided Dictionary Refinement $\mathcal{R}$

The operator $\mathcal{R}$ refines the feature dictionary using induced topology G. Learned parent–child relations guide Absorption Realignment to redistribute misallocated parent-related representation, while persistent graph-conditioned residuals guide Residual Completion to fill representational gaps.

Absorption Realignment. Feature absorption can cause a child to carry representation aligned with its broader parents (Chanin et al., 2025b). Using the parent set P_c recovered by $\mathcal{I}$, we decompose the decoder direction of each child c with $P_c \neq \varnothing$ as $\mathbf{d}_c = \sum_{p \in P_c} \delta_{cp}\mathbf{d}_p + \mathbf{r}_c$, where each $\delta_{cp} \geq 0$ measures how strongly the child direction $\mathbf{d}_c$ aligns with parent direction $\mathbf{d}_p$, while

$\mathbf{r}_c$ represents the child-specific direction retained after separating the parent-aligned components. Collecting these parent-alignment coefficients as $\boldsymbol{\delta}_c = (\delta_{cp})_{p \in P_c}$, we estimate them jointly through nonnegative ridge fitting, where $\lambda_\delta > 0$ controls the ridge regularization strength:

$$\boldsymbol{\delta}_c \coloneqq \arg\min_{\boldsymbol{\delta}_c \geq \mathbf{0}} \left[\left\| \mathbf{d}_c - \sum_{p \in P_c} \delta_{cp} \mathbf{d}_p \right\|_2^2 + \lambda_\delta \left\| \boldsymbol{\delta}_c \right\|_2^2 \right], \qquad \mathbf{r}_c \coloneqq \mathbf{d}_c - \sum_{p \in P_c} \delta_{cp} \mathbf{d}_p. \tag{4}$$

The decomposition above identifies which parts of the child direction align with its parents and which remain specific to the child. Absorption Realignment uses this structure to redistribute representation within the local parent–child subgraph during training: parent-aligned content is routed along the corresponding parent directions, while child-specific content remains with the child. The same local contribution can be written without changing the reconstruction:

$$z_c(\mathbf{x})\mathbf{d}_c + \sum_{p \in P_c} z_p(\mathbf{x})\mathbf{d}_p = \|\mathbf{r}_c\|_2 z_c(\mathbf{x}) \frac{\mathbf{r}_c}{\|\mathbf{r}_c\|_2} + \sum_{p \in P_c} \left(z_p(\mathbf{x}) + \delta_{cp} z_c(\mathbf{x}) \right) \mathbf{d}_p.$$

Under this decomposition, Absorption Realignment routes parent-aligned content out of the child and back along the corresponding parent directions, while continuously adapting the child–parent allocation during training. This directly mitigates local feature absorption (Section 4.5). Implementation details are provided in Appendix B.3.

Residual Completion. Absorption Realignment mitigates absorption among related features, whereas Residual Completion fills persistent representational gaps. Let $\mathbf{e}(\mathbf{x}) \coloneqq \mathbf{x} - \widehat{\mathbf{x}}$ denote the reconstruction residual of the current SAE. When child c and all features in P_c are active, we define $\mathbf{e}_c(\mathbf{x}) \coloneqq a_c(\mathbf{x}) a_{P_c}(\mathbf{x}) \mathbf{e}(\mathbf{x})$ to retain the residual associated with this parent–child relation. A stable direction in $\mathbf{e}_c$ indicates a persistent representational gap. Residual Completion reuses a matching existing feature when possible; otherwise, it learns a new sparse feature through:

$$\min_{\substack{z \geq 0 \\ \|\mathbf{d}\|_2 = 1}} \mathbb{E}_{\mathbf{x} \sim \mathcal{D}} \left[\|\mathbf{e}_c(\mathbf{x}) - z(\mathbf{x})\mathbf{d}\|_2^2 + \lambda_{\text{sp}} z(\mathbf{x}) \right], \tag{5}$$

where $\lambda_{\text{sp}} > 0$ controls activation sparsity. The new feature captures the persistent residual component exposed by the learned relation, thereby expanding the dictionary to better fill a previously unexplained representational gap. Implementation details are provided in Appendix B.3.

3.3 Full Re-Induction and Joint Stability

After $\mathcal{R}$ uses the induced graph $G^{(t)}$ to refine the dictionary into $\mathcal{F}^{(t+1)}$, Full Re-Induction closes the refinement cycle by reapplying $\mathcal{I}$ to the updated dictionary. Each feature therefore reassesses its parent set under the updated representations, yielding

$$G^{(t+1)} = \mathcal{I}\left(\mathcal{F}^{(t+1)}\right) = \left(P_c^{(t+1)}\right)_{c=1}^{M_{t+1}}.$$

AG-SAE terminates when both the graph and feature contributions stabilize across a refinement cycle. We measure the changes between successive cycles by d_G for the graph and d_F for the feature contributions on held-out activations. For tolerances $\gamma_G, \gamma_F > 0$, AG-SAE stops when

$$d_G\left(G^{(t+1)}, G^{(t)}\right) \leq \gamma_G, \qquad d_F\left(\mathcal{F}^{(t+1)}, \mathcal{F}^{(t)}\right) \leq \gamma_F.$$

Appendix B.4 gives estimator details, while Figure 5 shows the trajectories of d_F and d_G across refinement cycles. We also provide a theoretical analysis of dictionary–graph stability in Appendix C. When both conditions hold, AG-SAE returns the final dictionary–graph pair

$$(\mathcal{F}^\star, G^\star) = (\mathcal{F}^{(t+1)}, G^{(t+1)}).$$

4 Experimental Evaluation

We evaluate AG-SAE by **(a)** measuring its ability to recover structure on a mixed-topology toy model with known ground-truth structure; **(b)** assessing out-of-sample relational reliability and blinded semantic validity on real LLM activations, complemented by a feature-level case study; **(c)** further demonstrating that features learned under the AG-SAE paradigm exhibit stronger causal intervention capabilities; and **(d)** ablating its core operators to isolate their contributions.

4.1 Experimental Setup

Models, data, and baselines. SAE baselines use BatchTopK activations (Bussmann et al., 2024). Our main setting studies layer-13 residual-stream features of Gemma-2-2B on MiniPile (Gemma Team et al., 2024; Kaddour, 2023). We compare AG-SAE with Vanilla SAE Post-hoc, a strong post-hoc baseline using the same mixed-topology search space, Matryoshka SAE with scaled-MCS DAG

Table 1: **AG-SAE structure recovery in toy and real LLM settings. A.** Dictionary quality and structure recovery on the ground-truth mixed-topology toy model. In A, † denotes N/A for methods without a native parent graph; Vanilla SAE's eight no-parent matches reflect the default empty parent set. **B.** Final AG-SAE feature graph on Gemma-2-2B layer-13 activations from MiniPile. Panels C–D use the same setting. Category percentages use the 30,718 active features (99.99% of 30,720); Multi / parented is the fraction of parented features with multiple parents. **C.** Dictionary and graph statistics across refinement cycles. **D.** Representation quality and relation validation across methods (Abs., Feature Absorption; AutoInt., AutoInterp). PSV, NR, and SV denote Predictive Structural Validity, Non-Redundancy, and Semantic Validity; Joint denotes their intersection. † marks Tree SAE's lower L_0 from its native sparsification, which also gives its Abs. score a sparsity advantage.

A. Toy-model evaluation

	Dictionary quality		Structure recovery	
Method	**Test** $R^2 \uparrow$	**Feat. /24** ↑	**Exact PS /24** ↑ **0/1/2-Parent**	**HN Rejected /32** ↑
Vanilla SAE	1.00	**24**	8, (8/0/0)	—†
Matryoshka SAE	0.92	19	8, (8/0/0)	—†
Tree SAE	0.91	10	8, (6/2/0)	28
HSAE	0.99	23	10, (8/2/0)	25
AG-SAE (ours)	**1.00**	**24**	**24**, (**8**/**8**/**8**)	**32**

B. Final mixed-topology graph

Metric	**Result**
Total features	30,720
Isolated	19,912 (**64.82%**)
Roots	3,715 (**12.09%**)
Single-parent	5,421 (**17.65%**)
Multi-parent	1,670 (**5.44%**)
Multi / parented	**23.55%**
Longest path	8 edges

C. Refinement across cycles

Metric	$G^{(0)}$	$G^{(2)}$	$G^{(5)}$
Single-parent	2,343	5,142	5,421
Multi-parent	782	1,637	1,670
Absorption realignment	—	5,973	7,068
Residual completion	—	161	12
Relation gain ↑	0.0456	0.0875	0.0912
Graph change d_G (%) ↓	—	**18.20%**	**3.06%**
Feature change d_F (%) ↓	—	**68.45%**	**9.25%**

D. Real LLM validation

	Representation quality				Recovered-structure reliability						
	Global dictionary				Single-parent (%) ↑			Multi-parent (%) ↑			Joint (%) ↑
Method	**EV** ↑	L_0	**Abs.** ↓	**AutoInt.** ↑	**PSV**	**NR**	**SV**	**PSV**	**NR**	**SV**	**PSV∩NR∩SV**
AG-SAE	0.670	50.13	9.16	**86.44**	**99.75**	**97.00**	**76.88**	**97.00**	**91.88**	**75.00**	**69.56**
Vanilla SAE Post-hoc	0.668	50.00	51.39	80.03	71.50	75.50	69.13	32.63	65.75	58.25	40.63
Matryoshka SAE (DAG)	**0.680**	50.00	12.71	81.02	60.38	66.25	30.38	35.38	57.00	35.00	25.44
Matryoshka SAE (Tree)	**0.680**	50.00	12.71	81.02	48.25	54.50	24.25		N/A		22.38
HSAE	0.665	50.00	10.24	79.90	92.88	92.13	56.88		N/A		51.13
Tree SAE	0.663	40.84†	**5.11**	75.14	24.63	24.13	23.25		N/A		18.75

and best-MCS tree projections, Tree SAE, and HSAE (Huben et al., 2024; Bussmann et al., 2025; Cao et al., 2026; Luo et al., 2026). Toy-model experiments follow SynthSAEBench and Matryoshka SAE (Chanin & Garriga-Alonso, 2026; Bussmann et al., 2025), with cross-model and cross-domain evaluations on Qwen3.5-2B-Base and PubMed (Qwen Team, 2026; National Library of Medicine, 2025). Feature labeling and blinded relation evaluation use Qwen3-30B-A3B under SAEBench AutoInterp and HSAE-style protocols (Qwen Team, 2025; Karvonen et al., 2025; Luo et al., 2026).

Training and evaluation protocol. Hierarchical methods and AG-SAE use matched four-scale dictionaries, with matched training steps and data budgets and comparable sparsity operating points. Structure induction and threshold validation use disjoint data; the resulting dictionary, graph, and thresholds are then frozen for a separate report-only evaluation that cannot affect structure construction, model selection, or dictionary adaptation. Causal interventions evaluate bidirectional feature addition and ablation on 416 held-out prompts spanning 26 first-letter concepts. Complete dictionary widths, layer choices, training specifications, baseline implementations, cross-setting results, and experiment-specific hyperparameters are provided in Appendices D and E.

4.2 Exact Mixed-Topology Recovery in a Controlled Toy Model

Since no reliable ground-truth structure is available for real LLM activations, we evaluate structure recovery in a mixed-topology toy model constructed following the generation procedure of SynthSAEBench (Chanin & Garriga-Alonso, 2026). Observations are sparse linear superpositions of 24 predefined feature directions with positive activation magnitudes. Following the parent-conditioned activation mechanism of the Matryoshka SAE toy model (Bussmann et al., 2025), a child can activate only when all its parents are active. The graph contains eight features each with zero, one, and two parents (Appendix Figure 8). We also construct 32 hard negatives spanning correlation-only candidates, transitive ancestors, and incomplete subsets of true multi-parent sets (Appendix E.1).

At matched sparsity, AG-SAE attains $R^2 = 1.00$, recovers all $24/24$ features and $24/24$ complete parent sets, and rejects all $32/32$ hard negatives (Table 1, Panel A). Vanilla SAE also recovers all $24/24$ features with $R^2 = 1.00$; its lack of a native parent graph leaves only the eight default no-parent matches. Tree SAE and HSAE each recover only two of the eight single-parent sets and none of the eight two-parent sets. These comparisons establish that accurate reconstruction and feature matching do not determine feature organization. In this controlled setting, AG-SAE recovers the full mixed topology and rejects all tested spurious structural explanations.

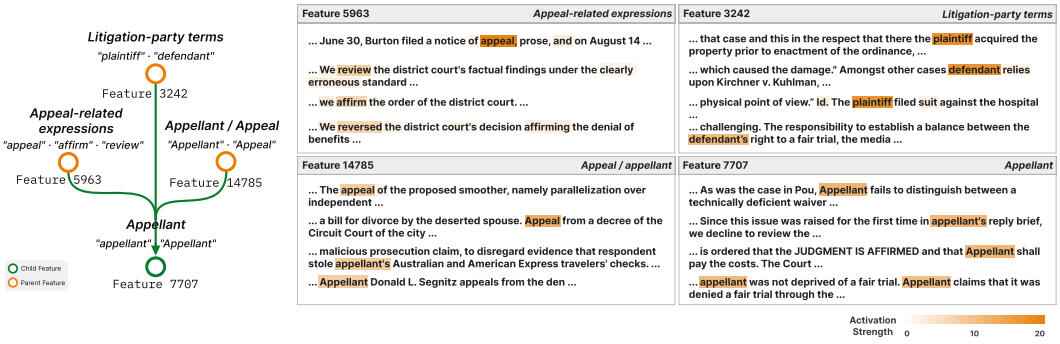

Figure 2: **Learned three-parent relation for appellant references. Left.** Feature 7707 and its complete parent set in final AG-SAE trained on MiniPile Gemma-2-2B layer-13 activations. Feature IDs are global across dictionary banks. **Right.** Activation examples selected per feature from profiling data and report contexts. Orange shading shows per-token activation on the same raw scale.

4.3 Reliable Structure Learning from Real LLM Activations

Learned Mixed-Topology Feature Graph. On layer-13 Gemma-2-2B activations from MiniPile, AG-SAE learns a feature graph containing isolated features, roots, single-parent relations, and multi-parent relations. Panel B of Table 1 shows that multi-parent features comprise **23.55%** of all parented features, establishing a substantial non-tree component in the learned topology. Tree SAE and HSAE inherit graph depth from four predefined dictionary levels, while AG-SAE's longest directed path spans eight edges, or nine feature levels, demonstrating that its joint dictionary–graph training learns relational depth beyond a prescribed dictionary hierarchy. Across refinement cycles, held-out relation gain rises from 4.56% to 9.12%, while d_F and d_G decline after $G^{(2)}$ to 9.25% and 3.06%, respectively (Table 1, Panel C), supporting progressive joint stabilization of the dictionary–graph pair as relation gain approaches a plateau. Appendix D.1 provides full refinement trajectories and analysis (Figure 5), while Appendix D.3 reports cross-model and cross-domain results.

Relation Reliability Validation. With the dictionary, graph, and thresholds frozen, Panel D of Table 1 combines two activation-based structural diagnostics with blinded semantic validation on a disjoint report-only split. Predictive Structural Validity (PSV) measures whether recovered parent–child dependencies remain stable and predictively reliable on unseen activations. Non-Redundancy (NR) measures the child's additional held-out reconstruction contribution beyond its parents. Both diagnostics are method-agnostic and use the same held-out criteria for all approaches. Semantic Validity (SV) tests relational coherence in human-interpretable natural language using frozen semantic labels generated following the SAEBench protocol (Karvonen et al., 2025) and blinded judgments, neither of which enters induction. Joint ($\mathrm{PSV} \cap \mathrm{NR} \cap \mathrm{SV}$) reports the fraction passing all three criteria on the same fixed, arity-matched relation samples.

AG-SAE leads all six single- and multi-parent reliability measures. Its single-parent PSV, NR, and SV reach 99.75%, 97.00%, and 76.88%, exceeding the best competing value per metric by 6.87, 4.87, and 7.75 percentage points. The gap is decisive for multi-parent relations. AG-SAE reaches 97.00% PSV, 91.88% NR, and 75.00% SV, leading the strongest baselines reporting each metric by 61.62, 26.13, and 16.75 points. Its 69.56% Joint rate exceeds the full-space Vanilla SAE Post-hoc comparator by 28.93 points. These establish predictive reliability on unseen activations, nonredundant child content, and the strongest blinded semantic coherence among evaluated methods.

At comparable sparsity, AG-SAE records a 0.01% dead-feature rate, the highest AutoInterp score of 86.44%, and the lowest absorption among methods near $L_0 = 50$, with EV within 0.010 of the best reported value. Tree SAE's lower raw absorption at reduced L_0 accompanies weak relation-validation results, showing that absorption alone does not determine structural reliability. AG-SAE thus achieves the strongest relation reliability with competitive representation quality.

Case Study of a Learned Multi-Parent Relation. To understand how the aggregate reliability results translate into individual feature organization, we study a learned multi-parent relation chosen for its readily interpretable legal vocabulary (Figure 2). In the Gemma-2-2B/MiniPile setting, AG-SAE assigns child 7707 the complete parent set $\{5963, 14785, 3242\}$. The parents activate on appellate expressions, appeal/appellant lexical forms, and litigation-party terms, respectively; the

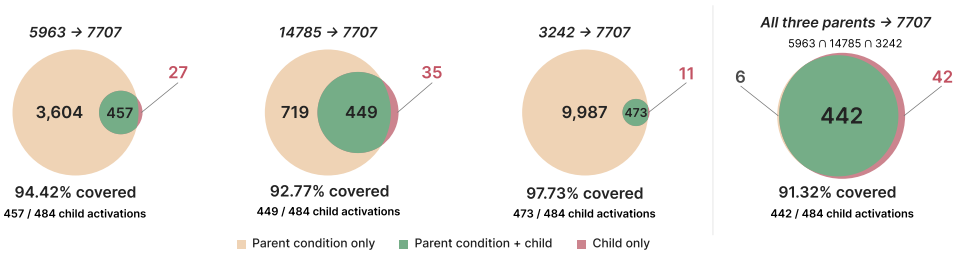

Figure 3: **Individual and joint parent coverage of feature 7707.** Counts use $4{,}910{,}199$ report token positions with frozen activation thresholds. The first three panels show individual parents; the rightmost requires all three parents to be active. Orange denotes parent-only activation, green coactivation with the child, and pink child activation with the parent condition unmet. Coverage uses the 484 child activations as its denominator. Panels use separate visual scales.

child's displayed activations focus on appellant references in litigation. These parental activation scopes overlap and extend beyond the child's, including nonlegal uses of *appeal* for feature 14785.

The complete parent configuration identifies the child's activation regime with both broad coverage and high specificity (Figure 3). Using native SAE activations with the dictionary and thresholds frozen, all three parents are active in $442/484 = 91.32\%$ of child-active positions. Conversely, the child is active in $442/448 = 98.66\%$ of the parents' joint activations, whereas it is active in at most 38.44% of any individual parent's activations. Their joint activation therefore identifies the child's more specific response pattern within the parents' broader scopes.

The complete parent set also provides stronger predictive support than its proper subsets and matched alternatives. Its held-out fit to the child's decoder contribution reaches 0.5497, compared with 0.2452 for the best single parent, 0.4543 for the best two-parent subset, and 0.4627 for the strongest evaluated alternative of the same cardinality. Removing any parent and refitting the remaining set reduces the score by at least 0.0954. The child additionally contributes a normalized innovation gain of $\mathcal{S}_{\text{inn}}(c, P) = 0.06177$ on the 442 positions where it and all parents are active. All reconstruction coefficients are fitted on a disjoint split and frozen for report evaluation. Full comparisons and scoring details appear in Appendix D.2. **Together, the aggregate validation and this case study show that AG-SAE learns reliable multi-parent structure with interpretable joint parent support and distinct child contributions.**

4.4 AG-SAE Features Enable Stronger Causal Interventions

We further find that AG-SAE's structure-guided training yields features with greater causal intervention capacity concentrated in their leading features. Building on SAEBench's first-letter probing setup (Karvonen et al., 2025), we use probes to identify the leading feature for each of the 26 first-letter concepts and evaluate these features on 416 disjoint held-out prompts. Following prior SAE intervention studies (Templeton et al., 2024), feature addition tests concept writing, whereas feature ablation tests concept erasure. Our experiments (Figure 4) show that **AG-SAE's leading feature exerts a stronger average causal effect than its Vanilla SAE counterpart, with the improvement appearing simultaneously in both concept writing and erasure for most tested concepts.** Crucially, this advantage is concentrated in the Top-1 feature, demonstrating that AG-SAE consolidates causal control into a single leading feature. Matched training counterfactuals further establish this causal advantage as a functional consequence specific to the AG-SAE paradigm; neither conventional SAE continuation nor decoder scaling reproduces it (Appendix D.4).

AG-SAE features provide stronger bidirectional causal control. The left panel of Figure 4 shows that individual AG-SAE features exert substantially stronger native-scale causal effects in both intervention directions. Mean ablation effect increases from 3.23 to 3.91 logits, a **21.2%** improvement over Vanilla SAE, while mean addition effect increases from 4.22 to 4.79 logits, a **13.6%** improvement. The center panel of Figure 4 shows that this advantage is consistent across concepts. AG-SAE improves addition for $24/26$ concepts and ablation for $21/26$, with **21/26** improving simultaneously in both directions. For AG-SAE, L2-matched random directions and unrelated SAE features produce mean target effects of at most 0.14 logits in magnitude across both intervention directions, establishing the target specificity of the measured effects. Its leading features therefore function as stronger causal control units, more effectively writing their associated concepts when added and erasing them when ablated.

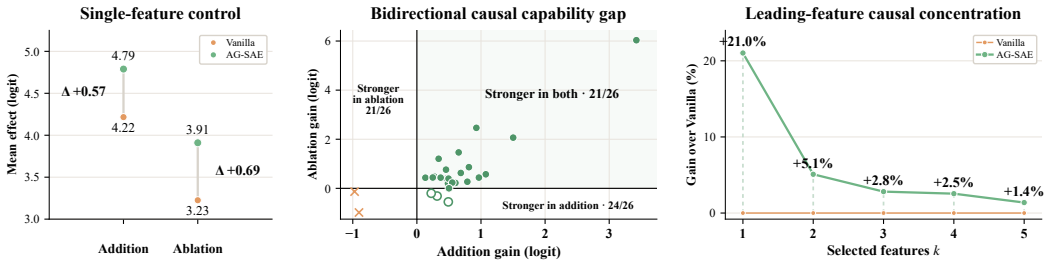

Figure 4: **AG-SAE consolidates bidirectional causal control into its leading features. Left.** Mean causal effects of adding and ablating the leading feature. **Center.** AG-SAE improves both intervention directions for $21/26$ concepts. **Right.** The gain in causal yield per active feature is concentrated at Top-1 and narrows as additional features are included. Together, the three panels establish stronger magnitude, cross-concept consistency, and leading-feature concentration.

Causal control capability is concentrated in the leading feature. The right panel of Figure 4 localizes where this increased causal capability resides. AG-SAE achieves **21.0%** greater causal yield per active feature at Top-1. The advantage contracts to 5.1% at Top-2 and below 3% thereafter, showing that the gain is strongly front-loaded into the leading feature. Moreover, for $20/26$ concepts, Vanilla SAE requires its two leading selected features to exceed the causal effect of a single AG-SAE feature. AG-SAE thus exposes target-relevant causal control through fewer feature units and provides a more compact intervention interface to the model.

These results reveal an unexpected consequence of AG-SAE's dictionary–graph self-consistency cycle. Training to induce faithful feature relations also yields stronger bidirectional causal control, with gains concentrated in each concept's leading feature. This finding motivates further research on enhancing SAE features' causal intervention capabilities beyond reconstruction quality.

4.5 Core Component Ablations

On the same fixed SAE dictionary and candidate space, replacing the assembly of independently selected parent–child edges with Complete Parent-Set Induction raises multi-parent PSV from 34.80% to 66.13% and PSV $\cap$ NR from 31.68% to 64.25%, with identical EV, L_0, and dictionary capacity. Under the SAEBench Feature Absorption protocol (Karvonen et al., 2025; Chanin et al., 2025b), starting from a fixed, fully trained SAE, we perform Complete Parent-Set Induction once and apply a single Absorption Realignment update using the resulting graph. Absorption Fraction decreases from 42.08% to 28.11% and Full Absorption from 38.92% to 27.56%. Evaluation settings and the inference threshold remain fixed; EV and mean L_0 change by only -0.000040 and $+0.0258$, respectively, with no change in the dead-feature fraction. These isolated ablations show that complete parent sets improve structural reliability and that realignment translates recovered structure into reduced feature absorption. Full protocols and analysis are provided in Appendix D.5.

5 Conclusion

AG-SAE extends structured SAE learning from single-parent trees to mixed topologies allowing features with zero, one, or multiple parents. It evaluates each candidate parent set as a complete hypothesis against null, subset, and alternative explanations, distinguishing jointly necessary multi-parent relations from spurious or redundant associations. The resulting graph guides SAE training through a structural loss and further guides dictionary refinement through Absorption Realignment and Residual Completion. Full Re-Induction then reselects every parent set under the revised dictionary, completing the dictionary–graph self-consistency cycle. In the controlled toy model, AG-SAE recovers all 24 ground-truth parent sets and rejects all 32 hard negatives. On real LLM activations, it achieves the strongest single- and multi-parent reliability among the evaluated methods at comparable representation quality. Its 69.56% overall joint validity exceeds the full-space post-hoc baseline by 28.93 percentage points. A concrete case further illustrates complete parent-set support and a nonredundant child contribution within a semantically coherent feature relation. Beyond these structural results, AG-SAE features support stronger concept writing and concept erasure for 21 of 26 concepts, with the gains concentrated in each concept's leading feature. Taken together, AG-SAE not only learns reliable mixed-topology feature structure, but turns that structure from a post-hoc description into an unsupervised training signal that refines the dictionary and yields functional benefits beyond reconstruction quality.

AI USE STATEMENT

Generative AI tools were used to assist with literature search and summarization, code editing and debugging, feedback on experimental design and result interpretation, and language and manuscript polishing. The core research idea, methodological novelty, and scientific contributions of this work were conceived by the authors.

REFERENCES

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Dan Braun, Jordan Taylor, Nicholas Goldowsky-Dill, and Lee Sharkey. Identifying functionally important features with end-to-end sparse dictionary learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3408. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c212c1b88395ab68d5e1671c17883ec6-Abstract-Conference.html.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Bart Bussmann, Patrick Leask, and Neel Nanda. BatchTopK sparse autoencoders. *arXiv preprint arXiv:2412.06410*, 2024. doi: 10.48550/arXiv.2412.06410. URL https://arxiv.org/abs/2412.06410.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with matryoshka sparse autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 6077–6101. PMLR, 2025. URL https://proceedings.mlr.press/v267/bussmann25a.html.

Tue M. Cao, Hoang X. Nhat, Raed Alharbi, Phi Le Nguyen, and My T. Thai. Tree SAE: Learning hierarchical feature structures in sparse autoencoders. *arXiv preprint arXiv:2605.07922*, 2026. doi: 10.48550/arXiv.2605.07922. URL https://arxiv.org/abs/2605.07922.

David Chanin and Adrià Garriga-Alonso. SynthSAEBench: Evaluating sparse autoencoders on scalable realistic synthetic data. In *Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=kALCcAhJa1.

David Chanin, Tomáš Dulka, and Adrià Garriga-Alonso. Feature hedging: Correlated features break narrow sparse autoencoders. *arXiv preprint arXiv:2505.11756*, 2025a. doi: 10.48550/arXiv.2505.11756. URL https://arxiv.org/abs/2505.11756.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. In *Advances in Neural Information Processing Systems*, volume 38, pp. 82318–82355, 2025b. doi: 10.52202/085713-2755. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/764ff7477b8e24dbe01531f6791e8bdf-Abstract-Conference.html.

Valérie Costa, Thomas Fel, Ekdeep Singh Lubana, Bahareh Tolooshams, and Demba Ba. From flat to hierarchical: Extracting sparse representations with matching pursuit. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/311bd32bcd952f93cf820bacd6955f0a-Abstract-Conference.html.

Seyed Arshan Dalili and Mehrdad Mahdavi. Subspace-aware sparse autoencoders for effective mechanistic interpretability. *arXiv preprint arXiv:2606.06333*, 2026. doi: 10.48550/arXiv.2606.06333. URL https://arxiv.org/abs/2606.06333.

Asen L. Dontchev and R. Tyrrell Rockafellar. *Implicit Functions and Solution Mappings: A View from Variational Analysis*. Springer, 2 edition, 2014. doi: 10.1007/978-1-4939-1037-3.

Javier Ferrando, Oscar Obeso, Senthooran Rajamanoharan, and Neel Nanda. Do i know this entity? knowledge awareness and hallucinations in language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=WCRQFlji2q.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tcsZt9ZNKD.

Gemma Team, Morgane Riviere, Shreya Pathak, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. doi: 10.48550/arXiv.2408.00118. URL https://arxiv.org/abs/2408.00118.

Nils Grandien, David Steinmann, Felix Friedrich, and Kristian Kersting. Do sparse autoencoders learn meaningful concept hierarchies? *arXiv preprint arXiv:2606.22994*, 2026. doi: 10.48550/arXiv.2606.22994. URL https://arxiv.org/abs/2606.22994.

Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, Junxuan Wang, Yunhua Zhou, Frances Liu, Qipeng Guo, Xuanjing Huang, Zuxuan Wu, Yu-Gang Jiang, and Xipeng Qiu. Llama scope: Extracting millions of features from Llama-3.1-8B with sparse autoencoders. *arXiv preprint arXiv:2410.20526*, 2024. doi: 10.48550/arXiv.2410.20526. URL https://arxiv.org/abs/2410.20526.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

Rob J. Hyndman and Yanan Fan. Sample quantiles in statistical packages. *The American Statistician*, 50(4):361–365, 1996. doi: 10.1080/00031305.1996.10473566.

Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, and Francis Bach. Proximal methods for hierarchical sparse coding. *Journal of Machine Learning Research*, 12(67):2297–2334, 2011. URL https://www.jmlr.org/papers/v12/jenatton11a.html.

Haoran Jin, Xiting Wang, Shijie Ren, Hong Xie, and Defu Lian. C^2R: Cross-sample consistency regularization mitigates feature splitting and absorption in sparse autoencoders. *arXiv preprint arXiv:2606.30609*, 2026. doi: 10.48550/arXiv.2606.30609. URL https://arxiv.org/abs/2606.30609. Accepted at the 43rd International Conference on Machine Learning (ICML 2026); camera-ready version states PMLR volume 306.

Jean Kaddour. The MiniPile challenge for data-efficient language models. *arXiv preprint arXiv:2304.08442*, 2023. doi: 10.48550/arXiv.2304.08442. URL https://arxiv.org/abs/2304.08442.

Adam Karvonen, Benjamin Wright, Can Rager, Rico Angell, Jannik Brinkmann, Logan Smith, Claudio Mayrink Verdun, David Bau, and Samuel Marks. Measuring progress in dictionary learning for language model interpretability with board game models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 83091–83118, 2024. doi: 10.52202/079017-2644. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/9736acf007760cc2b47948ae3cf06274-Abstract-Conference.html.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Isaac Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum Stuart Mcdougall, Kola Ayonrinde, Demian Till, Matthew Wearden, Arthur Conmy, Samuel Marks, and Neel Nanda. SAEBench: A comprehensive benchmark for sparse autoencoders in language model interpretability. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29223–29264. PMLR, 2025. URL https://proceedings.mlr.press/v267/karvonen25a.html.

Dmitrii Kharlapenko, Stepan Shabalin, Arthur Conmy, and Neel Nanda. Scaling sparse feature circuits for studying in-context learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29819–29851. PMLR, 2025. URL https://proceedings.mlr.press/v267/kharlapenko25a.html.

Anton Korznikov, Andrey Galichin, Alexey Dontsov, Oleg Rogov, Elena Tutubalina, and Ivan Oseledets. OrtSAE: Orthogonal sparse autoencoders uncover atomic features. *arXiv preprint arXiv:2509.22033*, 2025. doi: 10.48550/arXiv.2509.22033. URL https://arxiv.org/abs/2509.22033.

Patrick Leask, Bart Bussmann, Michael Pearce, Joseph Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. Sparse autoencoders do not find canonical units of analysis. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=9ca9eHNrdH.

Yuxiao Li, Eric J. Michaud, David D. Baek, Joshua Engels, Xiaoqing Sun, and Max Tegmark. The geometry of concepts: Sparse autoencoder feature structure. *Entropy*, 27(4):344, 2025. doi: 10.3390/e27040344. URL https://doi.org/10.3390/e27040344.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, Janos Kramar, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 278–300, Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.19. URL https://aclanthology.org/2024.blackboxnlp-1.19/.

Yifan Luo, Yang Zhan, Jiedong Jiang, Tianyang Liu, Mingrui Wu, Zhennan Zhou, and Bin Dong. From atoms to trees: Building a structured feature forest with hierarchical sparse autoencoders. *arXiv preprint arXiv:2602.11881*, 2026. doi: 10.48550/arXiv.2602.11881. URL https://arxiv.org/abs/2602.11881.

Zachary Maas. Conditional dependence structure in sparse autoencoder features. In *Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=wvzMNkaw7m.

Aleksandar Makelov, Georg Lange, and Neel Nanda. Towards principled evaluations of sparse autoencoders for interpretability and control. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/53356aebeea8ffd40a8ac3bb66243162-Abstract-Conference.html.

Alireza Makhzani and Brendan Frey. k-sparse autoencoders. In *The Second International Conference on Learning Representations*, 2014. URL https://arxiv.org/abs/1312.5663.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=I4e82CIDxv.

Mark Muchane, Sean Richardson, Kiho Park, and Victor Veitch. Incorporating hierarchical semantics in sparse autoencoder architectures. *arXiv preprint arXiv:2506.01197*, 2025. doi: 10.48550/arXiv.2506.01197. URL https://arxiv.org/abs/2506.01197.

Anish Mudide, Joshua Engels, Eric J. Michaud, Max Tegmark, and Christian Schroeder de Witt. Efficient dictionary learning with switch sparse autoencoders. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=k2ZVAzVeMP.

National Library of Medicine. MEDLINE/PubMed citations, 2025. URL https://datadiscovery.nlm.nih.gov/Literature/MEDLINE-PubMed-Citations/bg8v-b9q4. Accessed 2026-08-21.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. *arXiv preprint arXiv:2406.01506*, 2024. doi: 10.48550/arXiv.2406.01506. URL https://arxiv.org/abs/2406.01506.

Gonçalo Paulo and Nora Belrose. Sparse autoencoders trained on the same data learn different features. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://iclr.cc/virtual/2026/poster/10010650.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. doi: 10.48550/arXiv.2505.09388. URL https://arxiv.org/abs/2505.09388.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János Kramár, Rohin Shah, and Neel Nanda. Improving sparse decomposition of language model activations with gated sparse autoencoders. In *Advances in Neural Information Processing Systems*, volume 37, pp. 775–818, 2024a. doi: 10.52202/079017-0024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/01772a8b0420baec00c4d59fe2fbace6-Abstract-Conference.html.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with JumpReLU sparse autoencoders. *arXiv preprint arXiv:2407.14435*, 2024b. doi: 10.48550/arXiv.2407.14435. URL https://arxiv.org/abs/2407.14435.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L. Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Christopher Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Christine Ye, Charles O'Neill, John F. Wu, and Kartheik G. Iyer. Sparse autoencoders for dense text embeddings reveal hierarchical feature sub-structure. In *NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning*, 2024. URL https://openreview.net/forum?id=hjROYyfxfO.

Vladimir Zaigrajew, Hubert Baniecki, and Przemyslaw Biecek. Interpreting CLIP with hierarchical sparse autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 73918–73956. PMLR, 2025. URL https://proceedings.mlr.press/v267/zaigrajew25a.html.

Appendix

Table of Contents

A Related Work **14**

B Additional Method Details **15**

B.1 Complete Parent-Set Induction $\mathcal{I}$ 15

B.2 Structural Training Loss $\mathcal{L}_{\mathcal{I}}$ 16

B.3 Mixed-Topology-Guided Dictionary Refinement $\mathcal{R}$ 18

B.4 Full Re-Induction and Joint Stability 19

C Theoretical Analysis of Graph Stability in Full Re-Induction **21**

C.1 Setup and Assumptions. 21

C.2 Local Stability of Full Re-Induction 22

C.3 Eventual Graph Stability and Joint Stopping 23

D Additional Experimental Results **25**

D.1 Additional Analysis of the Refinement Process and Learned Mixed Topology 25

D.2 Additional Analysis of the Learned Multi-Parent Relation 26

D.3 Results Across Models and Datasets 28

D.4 Additional Causal Intervention Results 29

D.5 Additional Component Ablations 30

E Experimental Setup and Hyperparameters **32**

E.1 Toy Model Construction and Recovery Evaluation 32

E.2 LLM Activations and Baseline Comparisons 32

E.3 Training Configuration 33

E.4 AG-SAE Hyperparameters 34

E.5 Structural and Semantic Evaluation 34

E.6 Causal Intervention Protocol 36

A Related Work

Sparse Autoencoders and Flat Dictionaries. Sparse autoencoders (SAEs) decompose language-model activations into sparse combinations of latent directions drawn from an overcomplete dictionary. Compared with individual neurons, these directions often correspond to more localized and interpretable features (Bricken et al., 2023; Huben et al., 2024; Marks et al., 2025). Gated and thresholded objectives reduce activation shrinkage and make feature selection more precise. TopK, BatchTopK, and Switch variants further improve the tradeoff between reconstruction and sparsity, distribute sparse activations more effectively, and make dictionary learning easier to scale (Rajamanoharan et al., 2024a;b; Gao et al., 2025; Bussmann et al., 2024; Mudide et al., 2025; Makhzani & Frey, 2014). Large-scale projects have consequently released dictionaries with millions of latents across different LLM activation sites (Templeton et al., 2024; Gao et al., 2025; Lieberum et al., 2024; He et al., 2024). Standardized benchmarks evaluate interpretability, disentanglement, feature recovery, and downstream utility in addition to reconstruction and sparsity (Karvonen et al., 2025; Makelov et al., 2025; Karvonen et al., 2024; Chanin & Garriga-Alonso, 2026). Beyond describing model representations, SAE features also support causal circuit discovery and targeted interventions (Marks et al., 2025; Braun et al., 2024; Ferrando et al., 2025; Kharlapenko et al., 2025). Together, these advances have improved how flat feature dictionaries are learned, evaluated, and used.

Feature Pathologies and Hierarchical Organization. Sparse reconstruction alone does not determine how semantic variation should be distributed among latents. This ambiguity gives rise to feature splitting, absorption, composition, correlation-induced hedging, and different noncanonical solutions across training runs (Bricken et al., 2023; Chanin et al., 2025b; Bussmann et al., 2025; Dalili & Mahdavi, 2026; Jin et al., 2026; Korznikov et al., 2025; Leask et al., 2025; Cao et al., 2026; Chanin et al., 2025a; Paulo & Belrose, 2026). Prior work reduces feature instability and redundancy through cross-sample consistency, decoder orthogonalization, and subspace-based representations (Jin et al., 2026; Korznikov et al., 2025; Dalili & Mahdavi, 2026). Other approaches instead focus on organizing SAE features across semantic scales, using post-hoc coactivation analysis or hierarchical sparse coding (Ye et al., 2024; Li et al., 2025; Maas, 2026; Jenatton et al., 2011; Park et al., 2024). Matryoshka SAE learns nested dictionary prefixes that move from broad features to finer ones. This organization reduces splitting and absorption and lowers a decoder-similarity proxy for composition (Bussmann et al., 2025). Matching-Pursuit SAE extracts conditionally orthogonal hierarchical features from successive residuals (Costa et al., 2025). Together, these approaches organize SAE features at multiple semantic scales, but leave persistent relations between individual features unspecified.

Structured Feature Relations. Structured SAEs extend hierarchical feature organization by explicitly modeling relations between individual features. H-SAE links high-level atoms to expert SAEs, HSAE jointly learns multiple levels while maintaining a partial feature forest, and Tree SAE combines parent-gated activation coverage with cumulative reconstruction across levels (Muchane et al., 2025; Luo et al., 2026; Cao et al., 2026). Their evaluations cover reconstruction, interpretability, absorption, splitting, composition, and relation reliability (Muchane et al., 2025; Luo et al., 2026; Cao et al., 2026). Across the three approaches, explicit coarse-to-fine structure guides representation learning under a single-parent tree or forest prior, which allows each child to have at most one parent. However, post-hoc analyses find both single- and multi-parent associations, showing that feature organization does not uniformly follow a single tree (Bussmann et al., 2025; Grandien et al., 2026). A post-hoc DAG constructed from pairwise relations removes this topological restriction, but coactivation, redundancy, compositional overlap, and transitive ancestry can introduce spurious relations (Grandien et al., 2026). To address both the topological restriction and the risk of spurious relations, AG-SAE formulates each complete parent set as the unit of structural inference. It evaluates each candidate set against null, subset, and alternative explanations, thereby learning a mixed topology that contains features with zero, one, or multiple parents. The induced structure is then fed back into dictionary training through a structural loss and further guides dictionary refinement. The complete structure is subsequently re-induced from the updated dictionary. Together, this feedback loop couples validated structural inference with representation learning under a mixed topology.

B Additional Method Details

This section details Complete Parent-Set Induction $\mathcal{I}$, the structural training loss $\mathcal{L}_{\mathcal{I}}$, Mixed-Topology-Guided Dictionary Refinement $\mathcal{R}$, and the re-induction and joint-stability criteria that close the AG-SAE dictionary–graph self-consistency cycle.

B.1 Complete Parent-Set Induction $\mathcal{I}$

Conditional fitting and evaluation. Complete Parent-Set Induction evaluates the current dictionary $\mathcal{F}$ using native activations and feature contributions $z_i(\mathbf{x})\mathbf{d}_i$. All main-text expectations are estimated by sample averages over their corresponding conditioning events: representational support and complete-set comparisons use $a_c = 1$, coverage measures the fraction of these samples with $a_P = 1$, and the reconstruction errors defining innovation use $a_c a_P = 1$.

For each fixed (c, P), we fit $\boldsymbol{\beta}$ separately for $S = P$ and $S = P \cup \{c\}$ by nonnegative least squares, with one coefficient per feature contribution shared across samples. The fitted coefficients are held fixed when evaluating reconstruction errors on separate comparison samples. Within each data role, both reconstructions and the zero-reconstruction baseline use the same jointly active samples. Their errors define $\mathcal{S}_{\mathrm{inn}}$ in Eq. (3), while $\mathcal{S}_{\mathrm{rep}}$ in Eq. (2) uses the direct joint parent representation $\sum_{p\in P} z_p(\mathbf{x})\mathbf{d}_p$. Candidates require sufficient conditioning-event counts and positive normalizing

energies. Sample requirements and data allocations are given in Appendix E, with separate roles for structure selection, threshold validation, and report-only evaluation.

The population reconstruction errors compare nested function classes on the same conditional distribution. Since the additional child coefficient may be zero, whenever the zero-reconstruction error is positive,

$$0 \le \mathcal{E}_{c,P}(P \cup \{c\}) \le \mathcal{E}_{c,P}(P) \le \mathcal{E}_{c,P}(\varnothing), \qquad 0 \le \mathcal{S}_{\mathrm{inn}}(c, P) \le 1.$$

If the child contribution is already a fixed nonnegative combination of the parent contributions on this distribution, adding c leaves the reconstruction class unchanged and the innovation score is zero.

Complete-set comparisons. For each nonempty candidate parent set P of child c, let Q denote a competing parent set for the same child. We define

$$\mathcal{C}_c(P) = \{Q : Q \subsetneq P\} \cup \{Q : |Q| = |P|,\ Q \neq P\}, \tag{6}$$

combining all proper subsets of P, including the empty set, with equal-cardinality alternatives. All alternatives are drawn from the same candidate search space as P. Every $Q \in \mathcal{C}_c(P)$ is evaluated on the same active-child samples as P, with each nonempty proper subset receiving its own support score. For the empty set, the joint parent representation is zero, and we use

$$a_{\varnothing} = 1, \qquad \mathcal{S}_{\mathrm{rep}}(c, \varnothing) = 0.$$

Subset comparisons measure the support gained by retaining the complete set, while equal-cardinality alternatives test its member identities. A nonempty candidate is retained only if it satisfies the representational-support, activation-coverage, and child-innovation requirements and

$$\mathcal{S}_{\mathrm{rep}}(c, P) \ge \max_{Q \in \mathcal{C}_c(P)} \mathcal{S}_{\mathrm{rep}}(c, Q) + \tau_{\mathrm{cmp}}, \tag{7}$$

where $\tau_{\mathrm{cmp}} > 0$ is the required complete-set comparison margin.

Acyclic selection. Selection starts from an empty graph with $P_c = \varnothing$ for every child. At each step, $\mathcal{I}$ selects, among unassigned children, the retained feasible candidate (c, P) with the largest $\mathcal{S}_{\mathrm{rep}}(c, P)$. A candidate is feasible exactly when the current graph contains no directed path from c to any $p \in P$, so adding all edges $p \to c$ preserves acyclicity. Smaller cardinality breaks score ties, followed by a fixed ordering.

Upon selection, all edges $p \to c$, $p \in P$, are added jointly and $P_c = P$ is assigned. Selection stops when no feasible retained candidate remains; unassigned children keep $P_c = \varnothing$. This preserves the complete parent set as the unit of acceptance and yields the acyclic graph G used by subsequent structure-guided SAE training and dictionary refinement.

B.2 Structural Training Loss $\mathcal{L}_{\mathcal{I}}$

The structural loss $\mathcal{L}_{\mathcal{I}}$ provides differentiable training surrogates for the representational-support, complete-set-preference, activation-coverage, and child-innovation criteria used by Complete Parent-Set Induction $\mathcal{I}$. During each fixed-graph training phase, the induced graph G and comparison sets $\mathcal{C}_c(P_c)$ remain fixed while these surrogates guide feature-parameter updates; Full Re-Induction subsequently reassesses the updated dictionary using the original induction criteria.

The fixed-graph update uses the same graph-conditioned coordinates as Absorption Realignment. We retain $(z_i, \mathbf{d}_i)$ as the canonical unit-decoder representation and use $(z_i^G, \mathbf{d}_i^G)$ only as temporary training coordinates; their construction and conversion back to the unit-decoder representation are detailed in Appendix B.3. Define the contribution of feature i as

$$\mathbf{v}_i(\mathbf{x}) \coloneqq z_i(\mathbf{x})\mathbf{d}_i = z_i^G(\mathbf{x})\mathbf{d}_i^G.$$

For the multi-bank LLM setting, let V_b denote the features in reconstruction bank b and $\mathbf{b}_b$ its reconstruction bias, with $\mathbf{b}_b = \mathbf{0}$ for a bank without a bias. The graph and complete parent sets are global over the feature collection, so selected parents may span banks, whereas reconstruction remains bank-specific:

$$\widehat{\mathbf{x}}^{G,b}(\mathbf{x}) = \mathbf{b}_b + \sum_{i \in V_b} z_i^G(\mathbf{x})\mathbf{d}_i^G.$$

Sparse activations and gradient evaluation. Conditioning events use the native sparse activity indicators and remain fixed within each gradient update. Let $\widetilde{z}_i^G(\mathbf{x}) \geq 0$ denote the nonnegative encoder response before sparsification in the graph-conditioned coordinates, and $\widetilde{a}_i(\mathbf{x}) \in (0,1)$ its smooth presence gate, obtained by applying a sigmoid to the temperature-scaled difference between the encoder preactivation and calibrated activation threshold. The gate temperature is specified in Appendix E.4. These smooth gates provide activation-coverage gradients, including for parents with zero native sparse activation, without replacing the native activity indicators used by Complete Parent-Set Induction.

Representational support and complete-set preference. For a child $c \in V_b$, define the support target

$$\mathbf{y}_c(\mathbf{x}) \coloneqq \mathbf{x} - \widehat{\mathbf{x}}^{G,b}(\mathbf{x}) + \mathbf{v}_c(\mathbf{x}).$$

This is the reconstruction error after removing the child's contribution from bank b. During each gradient update, $\mathbf{y}_c$ is treated as a fixed target and gradients flow through the parent prediction.

At the beginning of each fixed-graph phase, nonnegative prediction coefficients $\boldsymbol{\alpha}^{c,P} = (\alpha_p^{c,P})_{p\in P}$ are fitted separately for each candidate (c,P) on samples with $a_c(\mathbf{x}) = 1$. Each coefficient is shared across samples and held fixed throughout the phase. The representational-support surrogate is

$$\mathcal{L}_{\mathrm{rep}}(c,P) = \frac{\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[\left\|\mathbf{y}_c(\mathbf{x}) - \sum_{p\in P}\alpha_p^{c,P}\widetilde{z}_p^G(\mathbf{x})\mathbf{d}_p^G\right\|_2^2 \,\middle|\, a_c(\mathbf{x}) = 1\right]}{\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[\|\mathbf{y}_c(\mathbf{x})\|_2^2 \,\middle|\, a_c(\mathbf{x}) = 1\right]}.$$

The prediction uses the complete set P, including parents from reconstruction banks different from that of the child.

Complete-set preference compares the selected set P_c with every $Q \in \mathcal{C}_c(P_c)$ from Eq. (6) using the same support target and child-active samples:

$$\mathcal{L}_{\mathrm{cmp}}(c) = \operatorname*{mean}_{Q\in\mathcal{C}_c(P_c)} T\log\left(1 + \exp\left[\frac{m + \mathcal{L}_{\mathrm{rep}}(c,P_c) - \mathcal{L}_{\mathrm{rep}}(c,Q)}{T}\right]\right),$$

where $m > 0$ is the training margin and $T > 0$ controls comparison smoothness. Each competing set uses its own fitted nonnegative prediction coefficients; for $Q = \varnothing$, the parent prediction is zero.

Activation coverage. The coverage criterion used by $\mathcal{I}$ requires $\Pr(a_{P_c} = 1 \mid a_c = 1) \geq \tau_{\mathrm{cov}}$. Its differentiable surrogate encourages every selected parent to be present on child-active samples:

$$\mathcal{L}_{\mathrm{cov}}(c) = -\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[\sum_{p\in P_c}\log\widetilde{a}_p(\mathbf{x}) \,\middle|\, a_c(\mathbf{x}) = 1\right].$$

During each gradient update, the child activity indicator is a fixed conditioning mask, while the smooth parent gates carry encoder gradients. Full Re-Induction evaluates the original coverage criterion using native activity indicators.

Child innovation. The innovation surrogate prevents the child-specific decoder remainder from collapsing during graph-conditioned training. Let $\mathbf{r}_c$ denote this remainder and $\mathbf{d}_c^G$ its corresponding composite decoder in the graph-conditioned parameterization of Appendix B.3. We use

$$\mathcal{L}_{\mathrm{inn}}(c) = \left[\max\left\{0,\ \eta - \frac{\|\mathbf{r}_c\|_2}{\|\mathbf{d}_c^G\|_2}\right\}\right]^2,$$

where $\eta > 0$ is the minimum target for the relative remainder norm. This discourages the child-specific component from vanishing into its parent-aligned allocation. It is only a training surrogate: the reconstruction gain $\mathcal{S}_{\mathrm{inn}}$ in Eq. (3) remains the child-innovation criterion used by Complete Parent-Set Induction and Full Re-Induction.

Combined structural loss. Combining the four components gives

$$\mathcal{L}_{\mathcal{I}}(\mathcal{F}; G) = \lambda_{\mathcal{I}} \operatorname*{mean}_{c: P_c \neq \varnothing} \Big[\lambda_{\mathrm{rep}} \mathcal{L}_{\mathrm{rep}}(c, P_c) + \lambda_{\mathrm{cmp}} \mathcal{L}_{\mathrm{cmp}}(c) + \lambda_{\mathrm{cov}} \mathcal{L}_{\mathrm{cov}}(c) + \lambda_{\mathrm{inn}} \mathcal{L}_{\mathrm{inn}}(c) \Big].$$

The nonnegative component weights control the four surrogate contributions, and the overall structural coefficient $\lambda_{\mathcal{I}}$ is absorbed into $\mathcal{L}_{\mathcal{I}}$ to match Eq. (1). Conditional terms are evaluated only for relations with sufficiently observed conditioning events and positive normalizing energies (Appendix E.4); the mean is over evaluable selected relations, with $\mathcal{L}_{\mathcal{I}} = 0$ if none are available.

As in Eq. (1), $\mathcal{L}_{\mathcal{I}}$ is combined with the conventional SAE reconstruction objective. Within each fixed-graph phase, the selected graph, comparison sets, and fitted prediction coefficients remain fixed, while conditioning masks are recomputed as the dictionary evolves. These surrogates guide parameter updates but do not certify relational validity; Full Re-Induction reassesses the updated dictionary using the original representational-support, complete-set comparison, activation-coverage, and child-innovation criteria of $\mathcal{I}$.

B.3 Mixed-Topology-Guided Dictionary Refinement $\mathcal{R}$

The refinement operator $\mathcal{R}$ uses the induced topology G for Absorption Realignment and Residual Completion. The structural loss for the fixed-graph update is specified in Appendix B.2. We retain the canonical unit-decoder representation $(z_i, \mathbf{d}_i)$ and introduce temporary graph-conditioned coordinates only during this refinement.

Absorption Realignment. With G fixed, let P_i denote the selected parent set of feature i. Each feature is represented by a trainable decoder remainder $\mathbf{r}_i$ and nonnegative parent-alignment coefficients δ_{ip} for $p \in P_i$. Let $\mathbf{d}_i^G$ denote its composite decoder direction in graph-conditioned coordinates. Evaluating features in parent-before-child order gives

$$\mathbf{d}_i^G = \mathbf{r}_i + \sum_{p \in P_i} \delta_{ip} \mathbf{d}_p^G, \qquad \delta_{ip} \geq 0.$$

If $P_i = \varnothing$, the sum is empty and $\mathbf{r}_i = \mathbf{d}_i$ at initialization. Otherwise, the parent-alignment coefficients and $\mathbf{r}_i$ are initialized by the nonnegative ridge decomposition in Eq. (4). This initialization is repeated at the beginning of every refinement cycle using the current native decoder directions. Applied in parent-before-child order, this gives $\mathbf{d}_i^G = \mathbf{d}_i$ for every feature at the start of the fixed-graph update. This graph-conditioned factorization operationalizes the main-text realignment during the fixed-graph update. Because $\mathbf{d}_i^G$ depends on the shared parent coordinates $\{\mathbf{d}_p^G \mid p \in P_i\}$, the parent-aligned portion of feature i's contribution is expressed through those coordinates, and its reconstruction gradients propagate through them. The remainder $\mathbf{r}_i$ parameterizes the child-specific component. The main-text identity gives an equivalent reconstruction-preserving activation view of this computation without requiring an explicit overwrite of native parent activations.

Let $z_i^G(\mathbf{x})$ denote the sparse activation of feature i in graph-conditioned coordinates, initialized as $z_i^G(\mathbf{x}) = z_i(\mathbf{x})$. For $\|\mathbf{d}_i^G\|_2 > 0$, the corresponding unit-decoder representation is

$$\mathbf{d}_i = \frac{\mathbf{d}_i^G}{\|\mathbf{d}_i^G\|_2}, \qquad z_i(\mathbf{x}) = \|\mathbf{d}_i^G\|_2 z_i^G(\mathbf{x}). \tag{8}$$

Thus, the two coordinate systems preserve the same feature contribution,

$$z_i(\mathbf{x}) \mathbf{d}_i = z_i^G(\mathbf{x}) \mathbf{d}_i^G,$$

and, because the rescaling factor is positive, the activity indicator $a_i(\mathbf{x})$ is unchanged.

Using the reconstruction-bank notation introduced in Appendix B.2, let V_b denote the features in bank b and $\mathbf{b}_b$ its reconstruction bias. The graph-conditioned reconstruction of bank b is

$$\widehat{\mathbf{x}}^{G,b}(\mathbf{x}) = \mathbf{b}_b + \sum_{i \in V_b} z_i^G(\mathbf{x}) \mathbf{d}_i^G.$$

The graph and complete parent sets are global, so a selected parent of a feature in V_b may belong to a different reconstruction bank, while reconstruction remains bank-specific. Since $z_i^G = z_i$ and $\mathbf{d}_i^G = \mathbf{d}_i$ at initialization, the graph-conditioned parameterization preserves the current reconstruction of every bank.

During the fixed-graph update, encoder parameters, decoder remainders, and parent-alignment coefficients are optimized in graph-conditioned coordinates under the reconstruction loss and the structural loss of Appendix B.2, as summarized in Eq. (1). The main-text notation suppresses the temporary graph-conditioned coordinates; $\mathcal{F}$ denotes the corresponding unit-decoder feature representation. After the fixed-graph update, the temporary coordinates are converted back to the unit-decoder representation using Eq. (8), preserving each feature contribution. The resulting native sparse activations and unit-norm decoder directions are then used by Residual Completion.

Residual Completion. For a child $c \in V_b$ with selected parent set P_c, Residual Completion uses the native reconstruction of the child's bank b with graph routing disabled:

$$\widehat{\mathbf{x}}^{\,b}(\mathbf{x}) = \mathbf{b}_b + \sum_{i \in V_b} z_i(\mathbf{x})\mathbf{d}_i, \qquad \mathbf{e}_b(\mathbf{x}) = \mathbf{x} - \widehat{\mathbf{x}}^{\,b}(\mathbf{x}).$$

Here, $\mathbf{e}_b(\mathbf{x})$ is the native reconstruction residual of bank b. The parent–child-conditioned residual is

$$\mathbf{e}_c(\mathbf{x}) = a_c(\mathbf{x}) a_{P_c}(\mathbf{x}) \mathbf{e}_b(\mathbf{x}).$$

Thus, the residual comes from the child's reconstruction bank, while the conditioning event requires the child and all selected parents to be active, including parents from other banks.

Leading directions of the uncentered residual second moment

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[\mathbf{e}_c(\mathbf{x})\mathbf{e}_c(\mathbf{x})^\top\right]$$

provide candidates for the completion objective in Eq. (5), oriented by the reconstruction energy captured under nonnegative projection. For a fixed unit direction $\mathbf{d}$, let $u \geq 0$ denote its nonnegative amplitude on a residual sample. The optimal amplitude for the sparse rank-one completion objective is

$$\arg\min_{u \geq 0} \left\{ \|\mathbf{e}_c(\mathbf{x}) - u\mathbf{d}\|_2^2 + \lambda_{\text{sp}} u \right\} = \max\left\{ 0,\ \mathbf{d}^\top \mathbf{e}_c(\mathbf{x}) - \frac{\lambda_{\text{sp}}}{2} \right\},$$

where $\lambda_{\text{sp}} > 0$ is the sparsity coefficient from Eq. (5). This gives the nonnegative activation target for a fixed residual direction; candidate validation and encoder initialization construct a learned approximation to it.

A candidate residual component must have sufficient contextual observations and retain reconstruction gain on separate validation samples. The observation, gain, and fit-to-validation consistency thresholds are specified in Appendix E.4.

Residual Completion first tests whether an existing feature in the same reconstruction bank V_b can capture the accepted residual direction. A directionally matched feature is reused only if its native activation and decoder contribution also satisfy the conditional reconstruction-gain criterion on validation samples. An accepted reused feature retains its existing identity and participates in subsequent dictionary refinement. If no existing feature in V_b passes the reuse criterion, an unused entry in bank b is initialized with the accepted residual direction and an activation threshold calibrated to the prescribed sparsity. The initialized entry receives a new feature identity, becomes part of $\mathcal{F}^{(t+1)}$, and has no assigned parent set under the current fixed graph G. Its encoder is further optimized in subsequent fixed-graph phases using Eq. (1). Its parent set is not inherited from the relation that produced the residual; instead, Full Re-Induction determines its complete parent set together with those of all other features.

B.4 Full Re-Induction and Joint Stability

Full Re-Induction. After Mixed-Topology-Guided Dictionary Refinement $\mathcal{R}$ produces $\mathcal{F}^{(t+1)}$, the temporary graph-conditioned coordinates are converted back through Eq. (8) to the native unit-decoder representation $(z_i, \mathbf{d}_i)$, preserving each feature contribution.

Full Re-Induction then reapplies $\mathcal{I}$ directly to $\mathcal{F}^{(t+1)}$ using the updated native sparse activations $z_i(\mathbf{x})$, activity indicators $a_i(\mathbf{x})$ recomputed from them, and feature contributions $z_i(\mathbf{x})\mathbf{d}_i$. Every feature, including newly initialized ones, is reassessed under the candidate search rules of Appendix B.1; candidate coefficients and scores are recomputed from the updated feature contributions, followed by complete-set comparisons and acyclic selection:

$$G^{(t+1)} = \mathcal{I}\Big(\mathcal{F}^{(t+1)}\Big) = \Big(P_c^{(t+1)}\Big)_{c=1}^{M_{t+1}}.$$

Existing parent sets compete under the same criteria as all other candidates, so each feature's assignment may change with its updated representation.

Feature identity and stability. Persistent features retain their identities across refinement cycles; reuse preserves an existing identity, while each newly initialized feature receives a new one. Let $\mathcal{V}^{(t)}$ denote the feature identities in $\mathcal{F}^{(t)}$, with $|\mathcal{V}^{(t)}| = M_t$, and define

$$\mathcal{U} \coloneqq \mathcal{V}^{(t)} \cup \mathcal{V}^{(t+1)}$$

as the identities present in either successive dictionary state. Parent-set members are compared across cycles using these persistent identities.

Structural change is the fraction of identities whose membership or complete parent-set assignment changes:

$$d_G\Big(G^{(t+1)}, G^{(t)}\Big) = 1 - \frac{\Big|\Big\{c \in \mathcal{V}^{(t)} \cap \mathcal{V}^{(t+1)} : P_c^{(t+1)} = P_c^{(t)}\Big\}\Big|}{|\mathcal{U}|}.$$

The numerator counts persistent features with unchanged complete parent sets; newly introduced features therefore contribute to structural change even when assigned an empty parent set. Hence $d_G = 0$ exactly when feature membership and all parent-set assignments are unchanged.

Feature change compares individual feature contributions on the same fixed validation samples. For any vector-valued function $\boldsymbol{\phi}(\mathbf{x})$, define

$$\|\boldsymbol{\phi}\|_{\mathrm{val}}^2 \coloneqq \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\mathrm{val}}}\big[\|\boldsymbol{\phi}(\mathbf{x})\|_2^2\big],$$

where $\mathcal{D}_{\mathrm{val}}$ is the validation distribution used for stability evaluation and the expectation is estimated by sample average. The normalized feature change is

$$d_F\Big(\mathcal{F}^{(t+1)}, \mathcal{F}^{(t)}\Big) = \left(\frac{\displaystyle\sum_{i\in\mathcal{U}} \Big\| z_i^{(t+1)}(\cdot)\mathbf{d}_i^{(t+1)} - z_i^{(t)}(\cdot)\mathbf{d}_i^{(t)} \Big\|_{\mathrm{val}}^2}{\displaystyle\sum_{i\in\mathcal{U}} \left(\Big\| z_i^{(t+1)}(\cdot)\mathbf{d}_i^{(t+1)} \Big\|_{\mathrm{val}}^2 + \Big\| z_i^{(t)}(\cdot)\mathbf{d}_i^{(t)} \Big\|_{\mathrm{val}}^2 \right)}\right)^{1/2}.$$

For an identity absent from one dictionary state, its contribution is defined as zero in that state; we also set $d_F = 0$ when the denominator is zero. Contributions are compared per feature before aggregation, preventing cancellation across feature updates. Since d_F depends on $z_i(\mathbf{x})\mathbf{d}_i$, it is invariant to contribution-preserving activation–decoder rescaling.

Stopping and returned state. Joint stability is evaluated after both dictionary refinement and Full Re-Induction have completed for the current cycle. For $\gamma_G, \gamma_F > 0$, AG-SAE terminates when

$$d_G\Big(G^{(t+1)}, G^{(t)}\Big) \le \gamma_G, \qquad d_F\Big(\mathcal{F}^{(t+1)}, \mathcal{F}^{(t)}\Big) \le \gamma_F.$$

When both conditions hold, AG-SAE returns

$$(\mathcal{F}^\star, G^\star) = (\mathcal{F}^{(t+1)}, G^{(t+1)}), \qquad G^\star = \mathcal{I}(\mathcal{F}^\star).$$

Thus, the returned graph is induced from the returned dictionary, and the completed cycle satisfies both structural and feature-contribution stability tolerances. Stopping uses validation samples separate from report-only evaluation; their allocation and the tolerances are specified in Appendix E, while Appendix C provides sufficient conditions for eventual graph stability and joint stopping.

C THEORETICAL ANALYSIS OF GRAPH STABILITY IN FULL RE-INDUCTION

We analyze the stability of the graph produced by Full Re-Induction as the dictionary evolves across refinement cycles. Small dictionary changes could in principle alter the discrete graph through changes in candidate acceptance and selection. We show that around a nondegenerate stabilized state, sufficiently small changes preserve all complete parent sets, preventing repeated structural changes from blocking the joint stopping rule.

C.1 SETUP AND ASSUMPTIONS

Let $\mathcal{V}$ be a finite, nonempty set of feature identities, with $M = |\mathcal{V}|$, and let $\theta \in \Theta \subseteq \mathbb{R}^q$ denote the continuous state determining the native activations, decoder directions, and induction thresholds. Write

$$\mathcal{F}_\theta = \{(z_i(\cdot;\theta), \mathbf{d}_i(\theta))\}_{i\in\mathcal{V}}, \qquad \mathbf{v}_i(\mathbf{x};\theta) = z_i(\mathbf{x};\theta)\mathbf{d}_i(\theta),$$

where $z_i \geq 0$ and $\|\mathbf{d}_i\|_2 = 1$. For multiple banks, $\mathcal{V}$ contains the features from all banks, and parent sets may cross banks. Reconstruction follows the bankwise objectives in Appendices B.2 and B.3. Each refinement step returns to this native representation through the contribution-preserving conversion in Eq. (8).

Fix a finite empirical induction configuration Ξ, comprising the retrieval, fitting, independent comparison, and control-calibration data, batch contexts, external randomness, and deterministic tie rules. For BatchTopK, occurrences of an input in different batch contexts are distinct sample positions. We write $\mathcal{I}_\Xi(\mathcal{F}_\theta)$ for the induction operator with this configuration made explicit. It recomputes sparse supports, candidate families, fitted coefficients, calibrated thresholds, and the graph, including dictionary-dependent control retrieval.

The empirical scores are those in Appendix B.1. The support score $\mathcal{S}_{\mathrm{rep}}$ compares $\mathbf{v}_c$ with the direct sum $\sum_{p\in P}\mathbf{v}_p$. The innovation score $\mathcal{S}_{\mathrm{inn}}$ uses separate nonnegative least-squares (NNLS) fits for P and $P\cup\{c\}$, evaluated on independent comparison samples. Let $\mathrm{cov}(c,P;\theta)$ denote the fraction of active-child samples on which all parents in P are active. For each nonempty candidate passing the sample-count and energy requirements, define its acceptance margin by

$$\begin{aligned} m_{c,P}(\theta) = \min\big\{\ & \mathrm{cov}(c,P;\theta) - \tau_{\mathrm{cov}}, \\ & \mathcal{S}_{\mathrm{rep}}(c,P;\theta) - \tau_{\mathrm{rep}}(c,P;\theta), \\ & \mathcal{S}_{\mathrm{inn}}(c,P;\theta) - \tau_{\mathrm{inn}}, \\ & \mathcal{S}_{\mathrm{rep}}(c,P;\theta) - \max_{Q\in\mathcal{C}_c(P)} \mathcal{S}_{\mathrm{rep}}(c,Q;\theta) - \tau_{\mathrm{cmp}}\big\}. \end{aligned} \tag{9}$$

Here $\mathcal{C}_c(P)$ is the comparison family in Eq. (6), including the empty set with $\mathcal{S}_{\mathrm{rep}}(c,\varnothing) = 0$. The threshold $\tau_{\mathrm{rep}}(c,P;\theta)$ is calibrated as in Appendix E.4; its arguments record the dictionary dependence and calibration group of (c,P). Candidates in the same model, dictionary-width, and parent-set-cardinality group share the threshold. A candidate is retained exactly when $m_{c,P}(\theta) \geq 0$. Starting from the empty graph, acyclic selection repeatedly chooses the feasible retained candidate with the largest support score among unassigned children and adds its complete parent set.

Condition C.1 (Nondegenerate reference state)**.** Let $\bar\theta \in \Theta$ be a valid native state. The reference execution of $\mathcal{I}_\Xi$ at $\bar\theta$ satisfies:

(i) **Discrete comparisons and branch continuity.** The comparisons determining native supports, eligible samples, preliminary screening, candidate pools, search spaces, and control identities form a finite collection. After substituting earlier reference choices, each comparison is represented by a scalar function $h_j(\theta)$ continuous in a relative neighborhood of $\bar\theta$. Either $h_j(\bar\theta) \neq 0$, or h_j is identically zero in that neighborhood and uses a fixed tie rule. Activation amplitudes, decoder directions, and preprocessing quantities are continuous on the reference branch. Sparse supports are described by the comparisons generating them, allowing inactive outputs to remain identically zero.

(ii) **Well-posed fitting and calibration.** Every nonempty NNLS problem solved on the reference data branch has a fitting matrix of full column rank. All normalizing energies used in scoring and calibration are positive, and each calibration group has a finite, nonempty list of valid control scores.

(iii) **Strict acceptance margins.** Every candidate passing preliminary screening has $m_{c,P}(\bar{\theta}) \neq 0$.

(iv) **Strict selection gaps.** At each step of acyclic selection, the selected candidate has a strictly larger support score than every other feasible retained candidate. This condition is vacuous when only one candidate is feasible.

C.2 Local Stability of Full Re-Induction

To show that small dictionary changes preserve the induced graph, we first control how refitting changes the relation scores. The following lemma establishes that the NNLS coefficients vary continuously with their inputs, so refitting does not introduce abrupt changes in the reconstruction errors used to assess child innovation. This sensitivity property follows from the theory of solution mappings for strongly monotone variational inequalities (Dontchev & Rockafellar, 2014, Section 2F). We give a direct argument with an explicit perturbation bound.

Lemma C.2 (Continuity of NNLS solutions)**.** *Suppose $A(\theta)$ and $\mathbf{y}(\theta)$ are continuous in a neighborhood of $\bar{\theta}$, and $A(\bar{\theta})$ has full column rank. Then, in a sufficiently small neighborhood, the solution*

$$\boldsymbol{\beta}(\theta) = \operatorname*{arg\,min}_{\boldsymbol{\beta} \geq \mathbf{0}} \tfrac{1}{2}\|A(\theta)\boldsymbol{\beta} - \mathbf{y}(\theta)\|_2^2$$

exists, is unique, and is continuous in θ.

Proof. Write $H = A^\top A$ and $\mathbf{w} = A^\top \mathbf{y}$. Full column rank and continuity give $H \succeq \mu I$ for some $\mu > 0$ throughout a sufficiently small neighborhood. The objective is strongly convex and coercive on the closed nonnegative orthant, so its minimizer exists and is unique.

For two nearby instances, distinguish their data and solutions by primes. The first-order optimality conditions (Boyd & Vandenberghe, 2004, Section 4.2.3) give

$$\langle H\boldsymbol{\beta} - \mathbf{w}, \boldsymbol{\beta}' - \boldsymbol{\beta} \rangle \geq 0,$$
$$\langle H'\boldsymbol{\beta}' - \mathbf{w}', \boldsymbol{\beta} - \boldsymbol{\beta}' \rangle \geq 0.$$

Adding the inequalities and setting $\boldsymbol{\Delta} = \boldsymbol{\beta}' - \boldsymbol{\beta}$ yields

$$\mu\|\boldsymbol{\Delta}\|_2^2 \leq \langle \boldsymbol{\Delta}, \mathbf{w}' - \mathbf{w} - (H' - H)\boldsymbol{\beta} \rangle.$$

By the Cauchy–Schwarz inequality,

$$\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2 \leq \frac{\|\mathbf{w}' - \mathbf{w}\|_2 + \|H' - H\|_{\mathrm{op}}\|\boldsymbol{\beta}\|_2}{\mu}. \tag{10}$$

The bound also holds when $\boldsymbol{\Delta} = \mathbf{0}$. Fixing the unprimed instance and letting the primed instance approach it proves continuity. □

We now show how continuous score changes lead to unchanged parent-set decisions. Under the nondegeneracy conditions, sufficiently small dictionary perturbations preserve the comparisons governing candidate acceptance and acyclic selection. Full Re-Induction therefore recovers the same mixed topology throughout a neighborhood of the reference state.

Proposition C.3 (Local constancy of Full Re-Induction)**.** *If Condition C.1 holds at $\bar{\theta}$, there exists $r_{\mathcal{I}} > 0$ such that*

$$\theta \in \Theta, \quad \|\theta - \bar{\theta}\|_2 < r_{\mathcal{I}} \quad \Longrightarrow \quad \mathcal{I}_\Xi(\mathcal{F}_\theta) = \mathcal{I}_\Xi(\mathcal{F}_{\bar{\theta}}). \tag{11}$$

Thus every complete parent set and the global DAG remain unchanged, including features with zero, one, or multiple parents.

Proof. We preserve the discrete execution branch, establish continuity of the scores and thresholds, and then use the strict margins to preserve candidate acceptance and acyclic selection.

Step 1: Discrete execution branch. Consider the comparisons in Condition C.1(i). For each $h_j(\bar{\theta}) \neq 0$, continuity gives a neighborhood on which

$$|h_j(\theta) - h_j(\bar{\theta})| < \tfrac{1}{2}|h_j(\bar{\theta})|.$$

Intersect these finitely many neighborhoods with those preserving the identically zero comparisons. If an execution first departed from the reference choices, all preceding choices would agree, so the departing comparison would use the same reference expression h_j. Its sign or fixed tie decision is unchanged, a contradiction. Hence the supports, conditional samples, preliminary screening, candidate families, and control identities agree with the reference execution. Sample counts and coverage fractions are therefore constant.

Step 2: Refitted scores. On this data branch, the feature contributions are continuous, so the NNLS matrices and targets are continuous. Condition C.1(ii) and Lemma C.2 give continuous fitted coefficients. For each refitted reconstruction, let $n_{\mathrm{cmp}} > 0$ be the number of comparison samples. Let $A_{\mathrm{cmp}}(\theta)$ and $\mathbf{y}_{\mathrm{cmp}}(\theta)$ denote the corresponding design matrix and target vector, respectively, obtained by stacking feature contributions and reconstruction targets across these samples. The coefficient vector $\boldsymbol{\beta}(\theta)$ is estimated on the separate fitting samples. Each error evaluated on the fixed comparison samples has the continuous form

$$\frac{1}{n_{\mathrm{cmp}}}\|A_{\mathrm{cmp}}(\theta)\boldsymbol{\beta}(\theta) - \mathbf{y}_{\mathrm{cmp}}(\theta)\|_2^2, \qquad n_{\mathrm{cmp}} > 0.$$

The innovation score is a difference of these errors divided by a positive energy, and is therefore continuous. The direct support score is likewise continuous, since it consists of finite averages of continuous contributions with a positive denominator.

Step 3: Calibration and acceptance. Let $s_1(\theta), \ldots, s_K(\theta)$ be the continuous scores of a fixed control list. Their kth order statistic satisfies

$$|s_{(k)}(\theta) - s_{(k)}(\bar{\theta})| \le \max_{1 \le j \le K} |s_j(\theta) - s_j(\bar{\theta})|.$$

Indeed, perturbing every entry by at most δ changes each ordered entry by at most δ. Standard sample quantiles select or interpolate order statistics (Hyndman & Fan, 1996). For a fixed list length, quantile level, and rule, the interpolation weights are fixed, so the quantile is continuous. Adding the calibration margin and taking the maximum with the floor preserve continuity; hence τ_{rep} is continuous. Finite maxima and minima then give continuity of the acceptance margin in Eq. (9). By Condition C.1(iii), a smaller neighborhood preserves the signs of all acceptance margins and thus the retained candidate set.

Step 4: Acyclic selection. Both executions start from the empty graph. If their first k choices agree, their current graphs and unassigned children agree. Reachability from each child to its candidate parents is identical, so the feasible candidates are the same. The reference winner's score gap over each competitor is continuous and strictly positive by Condition C.1(iv). It therefore remains the winner in a smaller neighborhood. Induction over at most M selections gives the same complete parent sets and the same termination point; all remaining children keep empty parent sets. Only finitely many neighborhood restrictions are required, yielding a common positive radius $r_{\mathcal{I}}$. □

Decision margins and the stability radius. Let $\mathcal{H}$ collect the nonzero reference comparisons, acceptance margins, and selection score gaps whose signs are preserved in the proof. Suppose these functions have Lipschitz constants L_h on a relative ball of radius $r_0 > 0$ that also preserves full rank, positive energies, and persistent ties. Then one may take

$$r_{\mathcal{I}} = \min\left\{ r_0,\ \min_{h \in \mathcal{H}:\, L_h > 0} \frac{|h(\bar{\theta})|}{2L_h} \right\}, \tag{12}$$

where the inner minimum is $+\infty$ when its index set is empty. The admissible perturbation is thus controlled by the decision margins relative to their sensitivity to the dictionary state.

C.3 Eventual Graph Stability and Joint Stopping

Local stability provides the graph-side guarantee needed for joint stopping. We next consider a refinement sequence whose dictionary state stabilizes to a nondegenerate limit. Once the dictionary states remain within the corresponding stability neighborhood, Full Re-Induction returns the same graph. Under the continuity conditions below, feature-contribution changes also vanish, ensuring that the joint stopping rule in Appendix B.4 is eventually satisfied. Fix nonempty validation samples

X_{val} with fixed batch contexts, separate from report-only evaluation. The sample-average validation norm is

$$\|\mathbf{u}\|_{\mathrm{val}}^2 \coloneqq \frac{1}{|X_{\mathrm{val}}|} \sum_{\mathbf{x} \in X_{\mathrm{val}}} \|\mathbf{u}(\mathbf{x})\|_2^2.$$

Empirically, the refinement trajectories in Figure 5 exhibit a stabilization trend consistent with the regime considered below. After $G^{(2)}$, both d_F and d_G decrease monotonically.

Corollary C.4 (Eventual stability and finite joint stopping)**.** *Consider a continued sequence of the stated refinement updates, with $\mathcal{F}^{(t)} = \mathcal{F}_{\theta_t}$ and $G^{(t)} = \mathcal{I}_\Xi(\mathcal{F}_{\theta_t})$ under the same configuration Ξ at every cycle. Suppose that:*

(i) After some finite t_0, the feature identity set is the same finite, nonempty $\mathcal{V}$. New identities may be introduced before t_0.

(ii) The states satisfy $\theta_t \to \theta_\infty \in \Theta$, and Condition C.1 holds at θ_∞.

(iii) Every validation contribution $\mathbf{v}_i(\mathbf{x}; \theta)$ is continuous at θ_∞, and

$$E_{\mathrm{val}}^\infty \coloneqq \sum_{i \in \mathcal{V}} \|\mathbf{v}_i(\cdot; \theta_\infty)\|_{\mathrm{val}}^2 > 0.$$

Let $G^\infty = \mathcal{I}_\Xi(\mathcal{F}_{\theta_\infty})$. There exists a finite $T_G \geq t_0$ such that

$$G^{(t)} = G^\infty, \qquad d_G(G^{(t+1)}, G^{(t)}) = 0 \quad (t \geq T_G),$$
$$d_F(\mathcal{F}_{\theta_{t+1}}, \mathcal{F}_{\theta_t}) \longrightarrow 0.$$

Consequently, for any fixed $\gamma_G, \gamma_F > 0$, there exists a finite T_γ such that both joint-stability inequalities hold for every $t \geq T_\gamma$.

Proof. Apply Proposition C.3 at θ_∞. Since $\theta_t \to \theta_\infty$, all sufficiently late states lie in its stability neighborhood. Thus $G^{(t)} = G^\infty$ for every $t \geq T_G$, for some finite $T_G \geq t_0$.

Let $W(\theta)$ concatenate all validation contributions in a fixed feature and sample order, scaled by $|X_{\mathrm{val}}|^{-1/2}$. On the fixed-identity tail, the metrics in Appendix B.4 are exactly

$$\begin{aligned} d_G(G^{(t+1)}, G^{(t)}) &= \frac{1}{M} \sum_{c \in \mathcal{V}} \mathbf{1}\{P_c^{(t+1)} \neq P_c^{(t)}\}, \\ d_F(\mathcal{F}_{\theta_{t+1}}, \mathcal{F}_{\theta_t}) &= \frac{\|W(\theta_{t+1}) - W(\theta_t)\|_2}{\left(\|W(\theta_{t+1})\|_2^2 + \|W(\theta_t)\|_2^2\right)^{1/2}}. \end{aligned} \tag{13}$$

As in the original definition, $d_F = 0$ when its denominator is zero. Eventual equality of the graphs gives $d_G = 0$. Continuity of the finitely many validation contributions gives $W(\theta_t) \to W(\theta_\infty)$, and hence

$$\begin{aligned} \|W(\theta_{t+1}) - W(\theta_t)\|_2 \leq{}& \|W(\theta_{t+1}) - W(\theta_\infty)\|_2 \\ &+ \|W(\theta_t) - W(\theta_\infty)\|_2 \longrightarrow 0. \end{aligned}$$

The squared denominator in Eq. (13) tends to $2E_{\mathrm{val}}^\infty > 0$, proving $d_F \to 0$. Choose T_F such that $d_F \leq \gamma_F$ for all $t \geq T_F$, and set $T_\gamma = \max\{T_G, T_F\}$. Both stopping inequalities then hold for every $t \geq T_\gamma$. Checking them after each completed cycle therefore triggers joint stopping no later than cycle T_γ along this continuation. □

Corollary C.4 gives sufficient conditions for reaching a stopping cycle. At any stopping cycle t_s, the returned dictionary–graph pair satisfies

$$(\mathcal{F}^\star, G^\star) = (\mathcal{F}_{\theta_{t_s+1}}, G^{(t_s+1)}), \qquad G^\star = \mathcal{I}_\Xi(\mathcal{F}^\star).$$

Thus the returned graph is induced from the returned dictionary. Positive tolerances allow the joint stopping criterion to be met before the graph becomes exactly constant.

Together, these theoretical results show that, under the stated conditions, once the dictionary stabilizes around a nondegenerate state, Full Re-Induction eventually preserves all complete parent sets. The graph therefore also stabilizes rather than repeatedly changing structure, allowing the joint stopping criterion to be satisfied.

D Additional Experimental Results

D.1 Additional Analysis of the Refinement Process and Learned Mixed Topology

On Gemma-2-2B layer-13 activations from MiniPile, we analyze how the dictionary–graph pair evolves during refinement and how the resulting feature relations are organized across dictionary scales and parent-set arities.

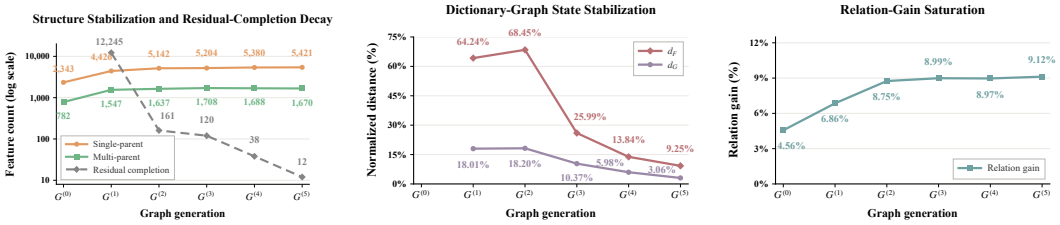

Figure 5: **AG-SAE refinement dynamics and joint stability.** Results use Gemma-2-2B layer-13 activations from MiniPile. **Left.** Counts of single- and multi-parent features across graph generations, together with accepted Residual Completion updates between successive refinement cycles. **Middle.** Normalized feature-contribution change d_F on held-out activations and graph change d_G in feature membership and complete parent-set assignments between successive refinement cycles. Both decrease after $G^{(2)}$, reaching 9.25% and 3.06%, respectively, in the final transition. **Right.** Relation gain is the held-out fraction of residual representation explained by parents with the child withheld, averaged over all selected relations, increasing from 4.56% at $G^{(0)}$ to 9.12% at $G^{(5)}$.

Refinement Across Cycles. The state snapshots in Panel C of Table 1 summarize the refinement statistics at representative graph generations, with Absorption Realignment counting the relations undergoing realignment during the corresponding refinement cycle. The complete trajectories in Figure 5 show that AG-SAE first forms most of its mixed topology and then progressively consolidates the dictionary–graph pair. By $G^{(2)}$, the single- and multi-parent populations are already close to their final levels. The multi-parent population peaks at $1{,}708$ in $G^{(3)}$ and then declines slightly to $1{,}670$ by $G^{(5)}$, showing that AG-SAE consolidates its learned topology rather than relying on unrestricted structural expansion. The left panel of Figure 5 shows that accepted Residual Completion updates fall from $12{,}245$ in the first transition to 12 in the final transition. This decline indicates that progressively fewer completion updates are accepted as refinement proceeds, consistent with Residual Completion addressing representational gaps exposed by the induced structure and the remaining demand for such corrections diminishing. The middle panel shows that feature-contribution change d_F increases from 64.24% to 68.45% across the first two transitions, coinciding with the period of greatest topology growth and Residual Completion activity. After $G^{(2)}$, d_F decreases monotonically to 9.25%. This non-monotonic trajectory is consistent with an early dictionary-reorganization phase followed by the later consolidation of feature contributions. The same panel shows that the exact parent-set change fraction d_G falls from approximately 18% in the first two transitions to 3.06% in the last, indicating that Full Re-Induction increasingly retains the parent sets selected under the preceding dictionary. Meanwhile, the right panel shows that mean held-out relation gain doubles from 4.56% at $G^{(0)}$ to 9.12% at $G^{(5)}$, remaining near 9% after $G^{(2)}$. Thus, the diminishing demand for completion accompanies stronger relational explanation, while reduced feature-contribution change, reduced parent-set reassignment, and the plateau in relation gain indicate that the dictionary–graph pair is approaching a stable configuration over the observed cycles. The final d_F and d_G values fall below their prescribed tolerances, establishing operational joint stability. Together, these trajectories support AG-SAE's central training premise that induced structure guides targeted dictionary refinement and the revised dictionary yields increasingly stable and better-supported relations.

Final Topology Across Dictionary Scales and Parent-Set Arity. The resulting graph contains multi-parent relations across all four dictionary banks, comprising 23.55% of parented features overall (Table 1). Graph roles separate systematically by scale, while relational depth extends beyond the four bank levels. Joint validity differs by only 0.75 percentage points between two- and three-

Table 2: **Multi-parent topology spans all four dictionary banks, and Joint validity changes by only 0.75 percentage points between two- and three-parent sets. A.** Cov. % is the number of parented features divided by bank width, and Multi % is the number of multi-parent features divided by the number of parented features. **B.** Structures denotes the total number of recovered relations in each arity cohort. Panel B applies the frozen report-only protocols of Table 1 to fixed arity-specific relation cohorts. PSV, NR, and SV are evaluated on the same relations, while Joint reports the percentage satisfying all three criteria. All rates are reported as percentages.

A. Coverage by child width					
Width	Parented	Cov. %	Single	Multi	Multi %
2,048	422	20.61	218	204	48.34
4,096	1,197	29.22	847	350	29.24
8,192	2,118	25.85	1,636	482	22.76
16,384	3,354	20.47	2,720	634	18.90
B. Validity by multi-parent arity					
Arity	Structures	PSV	NR	SV	Joint
2	1,239	98.75	95.25	70.75	65.50
3	431	95.25	88.50	79.25	64.75

parent sets. Panel A of Table 2 decomposes the 23.55% multi-parent share reported in Table 1 across the four dictionary banks. Among the $7{,}091$ parented features, $1{,}670$ (23.55%) have multiple parents. Every bank contains between 204 and 634 multi-parent children, which constitute 18.90% to 48.34% of its parented features. Parented features likewise cover 20.47% to 29.22% of each bank across the eightfold range of dictionary widths. The within-bank multi-parent share increases monotonically as dictionary width decreases, reaching 48.34% in the $2{,}048$-feature bank.

Graph roles separate systematically with dictionary scale. The left panel of Figure 6 shows that roots fall monotonically from 27.29% of the $2{,}048$-feature bank to 5.94% of the $16{,}384$-feature bank, while leaves rise monotonically from 6.49% to 18.85%. Roots outnumber leaves by 4.20 to 1 in the narrowest bank, whereas leaves outnumber roots by 3.17 to 1 in the widest. The fraction of features that serve as a parent for at least one child follows the same ordering, decreasing from 41.41% in the narrowest bank to 7.56% in the widest. The induction procedure permits edges in either scale direction and within a bank, making this monotonic role separation an empirical property of the recovered relations. The resulting gradient is consistent with the increase in feature granularity with dictionary width documented in prior work (Bricken et al., 2023; Bussmann et al., 2025). Narrow banks disproportionately contain explanatory roots, whereas wide banks disproportionately contain terminal refinements. Graph depth exceeds the limit of a strict four-level scale hierarchy. If every edge followed increasing dictionary width, any directed path beginning at a root could contain at most three edges. The right panel of Figure 6 places 394 features beyond this bound and reaches a maximum depth of eight. The graph also contains $2{,}381$ interior features with both a parent set and at least one child. Directed edges move from a narrower to a wider bank in 71.80% of cases, from a wider to a narrower bank in 27.87%, and within the same bank in the remaining 0.33%. These edge directions and paths extend the recovered organization beyond a four-level progression through dictionary scale.

The final graph contains $1{,}239$ two-parent sets and 431 three-parent sets. Panel B of Table 2 reports arity-specific validity under the frozen report-only protocols of Table 1. Two-parent sets reach 98.75% PSV, 95.25% NR, and 70.75% SV, with a Joint rate of 65.50%. Three-parent sets reach 95.25%, 88.50%, and 79.25% on the same metrics, with a Joint rate of 64.75%. The two Joint rates remain nearly identical, differing by only 0.75 percentage points. The pooled Joint rate in Table 1 is higher because it also includes the single-parent cohort. At both evaluated multi-parent arities, approximately two thirds of the sampled relations satisfy all three validity criteria.

D.2 Additional Analysis of the Learned Multi-Parent Relation

We provide additional activation statistics and reconstruction comparisons for the relation illustrated in Figures 2 and 3. The final AG-SAE trained on Gemma-2-2B layer-13 activations from MiniPile assigns child $c = 7707$ the complete parent set $P = \{5963, 14785, 3242\}$.

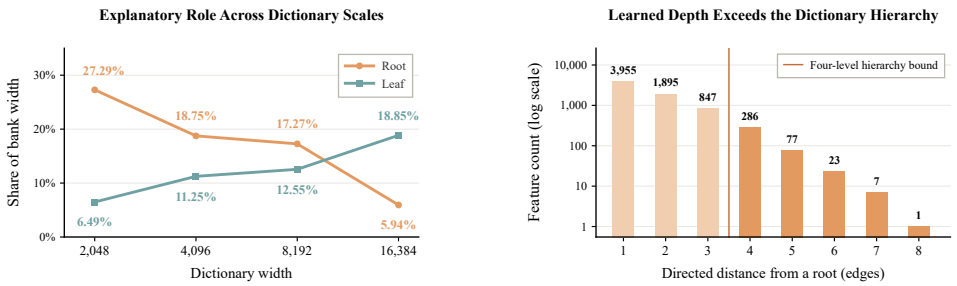

Figure 6: **Dictionary scale organizes graph roles, while graph depth extends beyond the four bank levels. Left.** Root and leaf shares within each dictionary bank. Roots fall from 27.29% to 5.94% of bank width, while leaves rise from 6.49% to 18.85%. **Right.** A feature's graph depth is the maximum number of edges among all directed paths that begin at a root and end at that feature. The dashed line marks the maximum depth of three permitted by a strict progression from narrower to wider banks. The recovered graph places 394 features beyond this bound and reaches depth eight.

Table 3: **Complete parent-set comparison for feature 7707.** The fit score used in PSV is one minus the squared prediction error for the child's decoder contribution, normalized by its total squared norm over the same 484 report positions where the child is active. Nonnegative coefficients are fitted separately for each parent set on a disjoint split and frozen for evaluation. The matched alternative is the strongest replacement set in the evaluated candidate pool. The null hypothesis has score 0.

Comparison	Parent set	Held-out fit ↑
Complete learned set	$\{5963, 14785, 3242\}$	**0.5497**
Remove 5963	$\{14785, 3242\}$	0.4543
Remove 14785	$\{5963, 3242\}$	0.4409
Remove 3242	$\{5963, 14785\}$	0.3460
Best single parent	$\{14785\}$	0.2452
Best matched alternative	$\{5295, 14785, 3242\}$	0.4627

Activation Coverage and Specificity. All activation statistics use the frozen dictionary with graph routing disabled and activation thresholds fixed before report evaluation. On the report split of 4,910,199 token positions, the child activates at 484 positions. Each parent covers 92.77%–97.73% of these positions, and all three parents jointly cover $442/484 = 91.32\%$. Conversely, only 11.25%, 38.44%, and 4.52% of the activations of features 5963, 14785, and 3242, respectively, coincide with child activation. For the complete parent configuration, this proportion reaches $442/448 = 98.66\%$. Thus, each parent individually covers most child activations while responding over a substantially broader set of positions. Their joint activation retains coverage of over nine tenths of child-active positions and identifies the child's activation regime with much greater specificity than any individual parent.

Complete Parent-Set Comparisons. Table 3 compares the learned parent set with its subsets and the strongest evaluated matched alternative. The complete set achieves a held-out fit of 0.5497, compared with 0.2452 for the best single parent and 0.4543 for the best two-parent subset. Removing any parent and refitting the remaining coefficients reduces the report score by at least 0.0954. Feature 14785 is particularly informative because its responses to *appeal* and *appellant* overlap lexically with the child. Removing it reduces the score to 0.4409, demonstrating incremental predictive support beyond the other two parents despite this lexical overlap. At the same parent-set cardinality, replacing feature 5963 with 5295 gives the strongest evaluated matched alternative, with a score of 0.4627. The learned set exceeds this alternative by 0.0870, supporting the value of its specific parent combination. Together, the subset and replacement comparisons establish predictive contributions from every selected parent and favor the learned combination over the evaluated alternatives.

Nonredundant Child Contribution. The child also supplies reconstruction value within the activation regime shared with its complete parent set. On the 442 report positions where the

Table 4: **AG-SAE dictionary quality, graph composition, and relation validity across models and datasets.** Gemma/PubMed uses Gemma-2-2B layer 13 with four dictionary scales; Qwen/MiniPile uses Qwen3.5-2B-Base layer 18 with two scales and 49,152 features in total. **A.** EV and L_0 are report-only measurements with graph routing disabled at each dictionary scale. Cov. % divides parented features by bank width. Single % and Multi % divide the respective feature counts by parented features; Root % and Leaf % divide their counts by bank width. Roots have children and no parents; leaves have parents and no children. **B.** Validation follows the frozen report-only protocols of Table 1, Panel D. Structures denotes the total number of recovered relations in each cohort. PSV, NR, SV, and Joint are evaluated under the same frozen report-only protocol within each cohort, with Joint reporting simultaneous satisfaction of all three criteria.

A. Dictionary Quality and Graph Composition

		Dictionary		Graph					
Setting	Width	EV	L_0	Parented	Cov. %	Single %	Multi %	Root %	Leaf %
Gemma / PubMed	2,048	0.660	50.29	328	16.02	50.61	49.39	34.28	4.74
	4,096	0.686	50.12	1,082	26.42	70.15	29.85	23.07	7.08
	8,192	0.710	49.74	2,349	28.67	74.37	25.63	17.19	14.64
	16,384	0.727	50.36	3,700	22.58	79.35	20.65	4.72	19.46
Qwen / MiniPile	16,384	0.771	49.74	3,873	23.64	54.66	45.34	20.06	12.34
	32,768	0.798	49.71	5,689	17.36	75.73	24.27	13.77	15.32

B. Relation Validity

Setting	Cohort	Structures	PSV	NR	SV	Joint
Gemma / PubMed	Single-parent	5,608	99.63	96.38	73.75	71.75
	Multi-parent	1,851	97.00	91.38	65.50	60.75
Qwen / MiniPile	Single-parent	6,425	99.38	93.13	65.50	63.13
	Multi-parent	3,137	95.13	88.25	74.38	62.25

child and all three parents are active, adding feature 7707 yields a normalized innovation gain of $\mathcal{S}_{\text{inn}}(c, P) = 0.06177$. The parent-only and parent-plus-child reconstructions use the same report positions, with nonnegative coefficients fitted separately on a disjoint split and frozen for evaluation. Following the definition of $\mathcal{S}_{\text{inn}}$, the reduction in reconstruction error is normalized by the zero-reconstruction error on these positions. The parent-set fit measures how well the parents predict the child's decoder contribution, while the innovation gain measures the child's additional contribution to reconstructing the original activations. The two comparisons therefore provide complementary support for the learned relation. The complete parent set supplies joint predictive support, and the child retains a nonredundant representational contribution even where their activations closely coincide.

D.3 Results Across Models and Datasets

AG-SAE recovers validated mixed topology under both an isolated corpus change and a joint change in model family, layer, and dictionary configuration. The PubMed setting holds Gemma-2-2B, layer 13, and the four dictionary scales fixed while replacing MiniPile with a biomedical corpus. The Qwen setting retains MiniPile while changing the model family, layer, number of dictionary scales, and total dictionary capacity.

D.3.1 Validated Mixed Topology Persists under a Corpus Change

AG-SAE reproduces the main mixed-topology pattern on PubMed after four refinement cycles (Table 4, Panel A). Multi-parent relations comprise 24.82% of parented features, compared with 23.55% on MiniPile (Table 1, Panel B), and maximum graph depth reaches 11. The topology spans all four dictionary banks, with multi-parent shares decreasing from 49.39% to 20.65% as width grows, following the same ordering as MiniPile (Table 2, Panel A). Roots become less prevalent and leaves more prevalent with increasing width, reproducing the scale-dependent role inversion.

Parented features cover 16.02%–28.67% of each bank; EV increases with dictionary width while native L_0 remains near 50.

Structural reliability also persists under the corpus change (Table 4, Panel B). Single-parent and multi-parent PSV reach 99.63% and 97.00%, within 0.12 and 0.00 percentage points of their MiniPile counterparts (Table 1, Panel D). NR changes by less than one percentage point in both cohorts. Joint validity reaches 71.75% for single-parent relations and 60.75% for multi-parent relations. Thus, more than three fifths of relations in each cohort simultaneously satisfy predictive, nonredundancy, and semantic criteria. With the model, layer, dictionary scales, and evaluation protocols held fixed, these results establish that broad cross-domain heterogeneity is not required for AG-SAE to recover validated multi-parent topology.

D.3.2 Validated Mixed Topology Persists under a New Model and Dictionary Configuration

AG-SAE recovers mixed topology and the same direction of scale-dependent role separation in Qwen3.5-2B-Base under a larger, two-scale dictionary configuration. Total capacity increases from $30{,}720$ to $49{,}152$ features, while the number of scales decreases from four to two; the main setting uses four scales to match the hierarchical baselines. After eight refinement cycles, $3{,}137$ of $9{,}562$ parented features have multiple parents, a share of 32.81% compared with 23.55% in the main setting (Table 1, Panel B). Within-bank multi-parent shares are 45.34% and 24.27% in the narrower and wider banks, respectively. Roots become less prevalent and leaves more prevalent in the wider bank, preserving the role ordering observed on Gemma with both MiniPile and PubMed. EV increases with width, with native L_0 near 50 at both scales (Table 4, Panel A).

Relational depth extends well beyond the two dictionary scales. A strict progression from the narrower to the wider bank permits a maximum depth of one; the recovered graph reaches depth nine, with $2{,}880$ features exceeding that bound, corresponding to 30.1% of all parented features. This substantial population demonstrates learned relational paths beyond the hierarchy defined by dictionary scale.

The recovered Qwen relations also retain joint predictive, nonredundancy, and semantic validity (Table 4, Panel B). Single-parent and multi-parent Joint rates reach 63.13% and 62.25%, respectively, so more than three fifths of relations in each cohort satisfy all three criteria. Together, the PubMed and Qwen results establish that validated mixed topology, systematic role separation across dictionary scales, and relational depth beyond those scales persist under an isolated corpus change and under a joint change in model family, layer, scale count, and total dictionary capacity.

D.4 Additional Causal Intervention Results

Matched training counterfactuals and paired statistics across concepts establish that the stronger causal control reported in Figure 4 is specific to full AG-SAE training among the evaluated procedures. The gain holds simultaneously in addition and ablation across most concepts and remains concentrated in the leading feature at the same Top-1 activation rate.

Only full AG-SAE training strengthens bidirectional causal control. Starting from the same pre-refinement checkpoint, we compare full AG-SAE training with two conventional SAE continuation procedures. Normalized continuation retains decoder normalization, and free-scale continuation allows decoder scale to vary under the conventional SAE objective. All three branches use the same data order and training budget of $40{,}000$ steps. The evaluation preserves the same 13 feature identities, fixed before the branch outcomes were assessed. For each feature, we subtract the matched random-control effect separately from its addition and ablation effects, compute the change from the shared checkpoint in each intervention direction, and retain the weaker change. A positive change therefore requires causal control to strengthen in both intervention directions. Table 5 reports the mean change from the shared checkpoint in this weaker effect, its bootstrap confidence interval, and the number of features with positive change. Full AG-SAE training increases bidirectional causal control by 0.867 logits, with a 95% confidence interval of $[0.298, 1.506]$, and improves $10/13$ features. Normalized and free-scale continuation produce mean changes of -0.215 and -0.160 logits and improve only $3/13$ and $5/13$ features. Relative to normalized continuation, full AG-SAE achieves a 1.082-logit advantage with a 95% confidence interval of $[0.450, 1.894]$, favoring $11/13$

Table 5: Matched changes in bidirectional causal control under full AG-SAE training and two conventional SAE continuation controls.

Condition	Change	95% CI	Wins
Full AG-SAE	$\mathbf{+0.867}$	$[\mathbf{0.298}, \mathbf{1.506}]$	$\mathbf{10/13}$
Normalized continuation	-0.215	$[-0.597, 0.184]$	$3/13$
Free-scale continuation	-0.160	$[-0.546, 0.235]$	$5/13$

Table 6: Paired AG-SAE gains over Vanilla SAE across the 26 first-letter concepts.

Metric	Gain	95% CI	Wins	p-value
Addition	$\mathbf{+0.573}$	$[\mathbf{0.308}, \mathbf{0.893}]$	$\mathbf{24/26}$	1.05×10^{-5}
Ablation	$\mathbf{+0.685}$	$[\mathbf{0.260}, \mathbf{1.240}]$	$\mathbf{21/26}$	0.0025
Bidirectional	$\mathbf{+0.674}$	$[\mathbf{0.281}, \mathbf{1.210}]$	$\mathbf{21/26}$	0.0025
Top-1 concentration	$\mathbf{+0.681}$	$[\mathbf{0.259}, \mathbf{1.241}]$	$\mathbf{20/26}$	0.0094

features. The advantage also holds separately in both intervention directions, reaching 0.794 logits in addition and 1.055 logits in ablation. Under matched initialization, data order, training budget, and feature identities, only the full AG-SAE branch strengthens addition and ablation jointly. Additional conventional SAE optimization and allowing decoder scale to vary do not reproduce the gain.

The causal advantage spans concepts and remains concentrated at Top-1. Table 6 aggregates the paired AG-SAE gains over Vanilla SAE across the 26 first-letter concepts. Gain denotes the mean difference in logits, and Wins counts concepts favoring AG-SAE. For Bidirectional, we first take the weaker of the addition and ablation effects within each method for each concept, and then compute the paired AG-SAE–Vanilla difference; Top-1 reports the single-feature cumulative ablation comparison. Intervals are 95% paired bootstrap confidence intervals, and p-values use exact two-sided sign tests. AG-SAE improves addition by 0.573 logits across $24/26$ concepts and ablation by 0.685 logits across $21/26$ concepts. Both paired confidence intervals are strictly positive, with exact sign-test p-values of 1.05×10^{-5} and 0.0025. This weaker-direction comparison yields a 0.674-logit bidirectional gain with a 95% confidence interval of $[0.281, 1.210]$ across $21/26$ concepts. The direction-specific results additionally show simultaneous strengthening of concept writing and erasure across most evaluated concepts. At Top-1, AG-SAE retains a 0.681-logit advantage with a 95% confidence interval of $[0.259, 1.241]$ across $20/26$ concepts, despite the identical 85.1% activation rate of the two methods. This locates the increased causal yield in the leading feature rather than in more frequent feature activation. Together, the matched continuation counterfactuals and paired statistics attribute the observed increase in causal control to the full AG-SAE training paradigm among the evaluated procedures.

D.5 Additional Component Ablations

We evaluate how Complete Parent-Set Induction and Absorption Realignment recover reliable relations and translate structural information into dictionary improvements. Figure 7 compares induction procedures on a fixed dictionary, measures the additional gains from dictionary optimization, and quantifies the reach and effectiveness of absorption correction. We then connect these component results to the full refinement trajectories.

Complete Parent-Set Induction improves structural reliability. On the same fixed SAE dictionary and candidate space, Pairwise Assembly constructs parent sets from independently selected parent–child edges, while Complete Parent-Set Induction evaluates each combination as an atomic structural hypothesis. Identical EV, L_0, and dictionary capacity isolate the induction procedure's contribution from changes in representation quality or sparsity. Complete Parent-Set Induction raises multi-parent PSV from 34.80% to 66.13% ($+31.33$ percentage points) and $\mathrm{PSV} \cap \mathrm{NR}$ from 31.68% to 64.25% ($+32.57$ points; Figure 7, left). At the same dictionary operating point, it more than doubles the fraction of multi-parent relations combining predictive parent support with a nonredundant child contribution, directly supporting the complete parent set as the unit of structural inference.

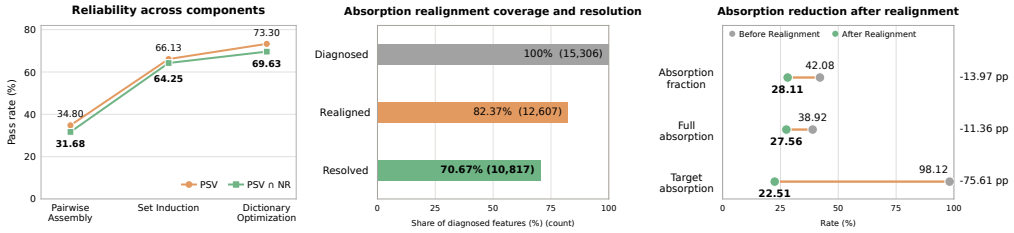

Figure 7: **Complete parent-set induction improves structural reliability, and realignment reduces feature absorption. Left.** Multi-parent PSV and PSV ∩ NR under Pairwise Assembly, Complete Parent-Set Induction, and Dictionary Optimization. Dictionary Optimization denotes the full post-induction dictionary update, including Structure-Guided SAE Training and Mixed-Topology-Guided Dictionary Refinement, followed by one Full Re-Induction on the updated dictionary to obtain the evaluated graph. The first two configurations share a fixed dictionary and candidate space. Set Induction denotes Complete Parent-Set Induction. **Middle.** Realignment coverage and resolution over 15,306 benchmark-diagnosed absorption cases. Of these, 12,607 receive an Absorption Realignment update and 10,817 are subsequently resolved; percentages use all diagnosed cases as the denominator. **Right.** Absorption measurements before and after realignment. Gray and green mark the before and after values, respectively; reductions are in percentage points.

Dictionary Optimization further raises PSV to 73.30% and PSV ∩ NR to 69.63%, gains of 7.17 and 5.38 percentage points over induction alone. These comparisons establish two complementary improvements. Complete Parent-Set Induction extracts more reliable structure from a fixed representation, and adapting the dictionary improves relational reliability beyond that structural selection step.

Absorption Realignment achieves broad coverage and substantial absorption reduction. The middle and right panels of Figure 7 report complementary measurements from a targeted experiment evaluating Absorption Realignment. Starting from a fixed, fully trained SAE, we perform Complete Parent-Set Induction once and apply Absorption Realignment once using the resulting graph. The SAEBench Feature Absorption diagnostic (Karvonen et al., 2025), which builds on the first-letter evaluation of Chanin et al. (2025b), identifies an evaluation cohort of 15,306 absorption cases. The benchmark diagnosis defines this cohort, while induction and realignment use the SAE's representations and model-internal quantities. The middle panel shows that 12,607 cases (82.37%) receive a realignment update and 10,817 (70.67%) are subsequently resolved. Both rates use the full diagnosed population, including cases that receive no update or remain unresolved. The operator thus reaches over four fifths of diagnosed cases and resolves more than seven tenths of the full cohort. The right panel reports the corresponding absorption reductions. Under identical evaluation settings and the same frozen inference threshold, Absorption Fraction falls from 42.08% to 28.11% (−13.97 percentage points), Full Absorption from 38.92% to 27.56% (−11.36 points), and target absorption from 98.12% to 22.51% (−75.61 points). The target correction thus accompanies improvements in both aggregate absorption measurements. EV changes by only −0.000040, mean L_0 by +0.0258, and the dead-feature fraction is unchanged. Together with the frozen threshold, this stable operating point makes the absorption reductions informative about feature organization. These results demonstrate that Absorption Realignment converts the graph recovered by Complete Parent-Set Induction into effective absorption correction, supporting its intended role in reallocating parent-aligned content and AG-SAE's use of learned structure to refine the dictionary.

Component effects align with the full training dynamics. The refinement trajectories in Figure 5 connect these component effects to the complete learning cycle. Residual Completion updates decline from 12,245 in the first transition to 12 in the last, consistent with diminishing demand for corrections to representational gaps exposed by the recovered structure. Feature-contribution change d_F and graph change d_G fall to 9.25% and 3.06%, respectively, while mean held-out relation gain rises from 4.56% to 9.12%. The multi-parent population settles at 1,670 after peaking at 1,708, showing that stronger relation support accompanies structural consolidation without continued expansion of the multi-parent population. Together, these trajectories support progressive dictionary–graph stabilization as completion demand and parent-set reassignment diminish. The

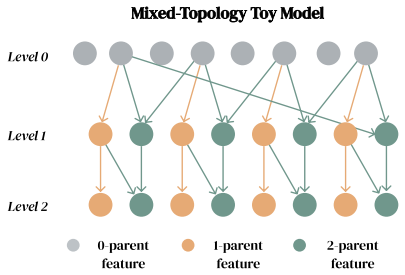

Figure 8: Ground-truth mixed topology of the toy model. Gray, orange, and green nodes have zero, one, and two immediate parents, respectively. Child activation requires all immediate parents to be active.

fixed-dictionary comparison establishes the benefit of complete parent-set inference, the isolated realignment test establishes effective absorption correction, and the dictionary-optimization gains and refinement trajectories support their integration into an alternating process that uses recovered relations to improve the dictionary.

E Experimental Setup and Hyperparameters

We specify the data construction, baseline configurations, and evaluation protocols underlying the reported results, together with the core training hyperparameters.

E.1 Toy Model Construction and Recovery Evaluation

Our toy model follows the SynthSAEBench framework (Chanin & Garriga-Alonso, 2026), generating sparse linear mixtures of 24 unit directions in 24 dimensions, with Gaussian-copula correlations among candidate activations and Gaussian magnitudes (mean 1.0, standard deviation 0.1) clipped at zero, with zero bias and no added noise. Feature directions are generated with graph-local correlations, and related active features share a common latent scale with feature-specific variation. Extending Matryoshka SAE's parent-conditioned toy model (Bussmann et al., 2025), child activation requires all immediate parents to be active. The three-level graph contains four isolated features, four roots, eight single-parent features, and eight two-parent features (Figure 8). Target activation probabilities for isolated features, roots, level-1 features, and level-2 features are 0.08, 0.20, 0.03, and 0.003, respectively, yielding an expected 1.384 active features per observation. Before training, we specify 32 hard negatives comprising four correlation-only edges, 12 non-immediate ancestor edges, and the 16 singleton subsets of the eight true two-parent sets. Each correlation-only pair targets a conditional activation probability of 0.70 for the associated isolated feature given child activation, compared with its marginal probability of 0.08. Methods share the generated instance, 200,000 validation observations, and 1,000,000 independent test observations, with ground-truth directions and parent assignments reserved for evaluation. Feature recovery uses Hungarian decoder matching with an absolute cosine threshold of 0.80. Exact parent-set recovery additionally requires recovery of the child and all true parents and equality of the complete predicted and ground-truth immediate parent sets. For methods producing a graph, hard-negative rejection measures the absence of forbidden edges and the nonselection of incomplete parent sets, complementing reconstruction and feature recovery with evaluation of the generating relations.

E.2 LLM Activations and Baseline Comparisons

Activation data and comparison setting. Following HSAE's choice of model, corpus, and reference dictionary sizes (Luo et al., 2026), we use Gemma-2-2B (Gemma Team et al., 2024) activations on MiniPile (Kaddour, 2023) from transformer block 13 (zero-based). All methods share 100 million training and 5 million validation activations, extracted from sequences of up to 512 tokens with padding and the first eight positions excluded, and standardized per coordinate using training statistics. The reference widths are 2,048, 4,096, 8,192, and 16,384, corresponding to independently reconstructing dictionaries for Vanilla SAE, HSAE, and AG-SAE, and feature groups within a 30,720-feature dictionary for Matryoshka SAE and Tree SAE. BatchTopK (Bussmann et al., 2024)

targets 50 active features per independent dictionary or full cumulative reconstruction; realized sparsity is reported in the main comparison table.

Vanilla SAE Post-hoc. We freeze the four independently trained SAEs and construct edges using the directed scaled masked cosine similarity of Bussmann et al. (2025, Appendix F), with the published threshold of 0.6. The score computes activation cosine similarity on tokens where the child is active, scaled by the ratio of maximum child to maximum parent activation. The candidate search space matches AG-SAE, including same-scale and all cross-scale candidates. Activation-frequency ordering ensures acyclicity, and each child retains up to three highest-scoring qualifying parents.

Matryoshka SAE. Following Bussmann et al. (2025), a single global BatchTopK operation selects activations, and equally weighted reconstruction losses over cumulative prefixes of 2,048, 6,144, 14,336, and 30,720 features update the shared parameters. Graph extraction uses the same activation score and threshold between adjacent feature groups. From the same trained dictionary, the DAG variant retains all qualifying parents, while the tree variant follows the paper's visualization projection and retains only the highest-scoring parent; children without qualifying parents remain unassigned.

Tree SAE. Following Cao et al. (2026), features occupy four privilege layers, with each feature assigned to the imaginary root or one parent in a lower privilege layer. Training combines equally weighted cumulative reconstruction losses with layer-specific auxiliary reconstruction for inactive features, retaining dynamic reallocation of inactive children and root fallback. The per-layer BatchTopK budgets are 34, 7, 5, and 4. Parent gating follows sparsification, retaining a child only when its parent is active; removed activations are not replaced, so fewer than 50 features may remain active.

HSAE. We implement HSAE (Luo et al., 2026) with BatchTopK, initializing it from the four pretrained SAEs. Each level reconstructs the same input. Following the original method, each child selects one parent by maximum encoder cosine similarity in the preceding level; the lowest-scoring 20% remain unassigned. The objective retains the parent–children contribution constraint with the published weight of 0.01, and reconstruction replaces a parent's contribution with the sum of its children's contributions with probability 5%. Sparsification precedes substitution, and the constraint uses unperturbed contributions. Assignments are updated every 5,000 steps and after training; joint optimization processes 20 million cached training activations. Training budgets and optimization schedules are detailed in Section E.3.

Additional model and corpus settings. Additional AG-SAE experiments use PubMed (National Library of Medicine, 2025) with the same Gemma-2-2B layer and dictionary configuration, and Qwen3.5-2B-Base (Qwen Team, 2026) at layer 18 on MiniPile with independent dictionaries of 16,384 and 32,768 features.

E.3 Training Configuration

For the main Gemma-2-2B comparison, Vanilla SAE (Huben et al., 2024) receives 120k, 160k, 220k, and 300k updates for dictionary sizes of 2,048, 4,096, 8,192, and 16,384, respectively. AG-SAE uses the same per-dictionary training budgets, including 25k, 25k, 50k, and 75k initialization updates. AG-SAE's algorithmic stopping criterion is the joint dictionary–graph stability rule defined in Appendix B.4. For the main baseline comparison, however, we additionally impose the matched training budgets to prevent AG-SAE from receiving an advantage from additional optimization. Accordingly, AG-SAE stops either when the joint stability criterion is satisfied or when the matched training budget is exhausted. Each dictionary is frozen upon reaching its update ceiling, and the returned graph is freshly induced from the final dictionaries. Matryoshka SAE (Bussmann et al., 2025) and Tree SAE (Cao et al., 2026) each receive 248k updates, matching the average of the four reference budgets weighted by their initial dictionary widths and corresponding to 507.904M activation presentations at batch size 2,048. HSAE (Luo et al., 2026) initializes from the fully trained Vanilla dictionaries and receives an additional 20k joint updates at batch size 1,024, with each shared batch updating all four dictionaries and providing 20.48M additional activation presentations per dictionary. All methods use the same cache of 100M training activations, with presentation counts including repeated cache passes. LLM baselines use AdamW with zero weight decay, gradient clipping

Table 7: **Model-specific AG-SAE hyperparameters.**

Hyperparameter	Gemma	Qwen
λ_{cov}	0.006	0.012
λ_{inn}	0.0015	0.003
Initial $\lambda_{\mathcal{I}}$	0.05	0.015
LR, cycles 1–5	$(2, 1.5, 1, 0.6, 0.3) \times 10^{-4}$	$(3, 2.2, 1.5, 0.9, 0.6) \times 10^{-4}$
LR, cycles 6–8	3×10^{-5}	6×10^{-5}
LR, cycles 9–16	3×10^{-5}	3×10^{-5}
Updates/cycle, 9–16	40,000	40,000
Min. residual gain	10^{-4}	2×10^{-4}
Min. residual validation/fitting ratio	0.25	0.50

at 1.0, and a peak learning rate of 3×10^{-4}, with 10% linear warmup followed by cosine decay to zero within each training stage; HSAE's joint stage starts a fresh optimizer and learning-rate schedule. AG-SAE uses AdamW with zero weight decay, gradient clipping at 1.0, 500 warmup steps, and a 2,000-step structural-loss ramp per cycle, with learning rates of 2×10^{-4}, 1.5×10^{-4}, 10^{-4}, 6×10^{-5}, and 3×10^{-5} for cycles one through five and 3×10^{-5} thereafter. All toy-model methods receive 40,000 Adam updates with $\beta = (0.5, 0.9375)$ and batch size 256, corresponding to 10.24M training observations per method, using a learning rate of 3×10^{-2} for the first 38,000 updates and 3×10^{-3} for the final 2,000 updates; AG-SAE performs its full fixed-graph update, including Structure-Guided SAE Training and Mixed-Topology-Guided Dictionary Refinement, during these final 2,000 updates.

E.4 AG-SAE Hyperparameters

For both Gemma and Qwen, we restrict parent sets to $|P| \leq 3$, retrieve 24 candidates, and use an exact-search pool of 12, with $\tau_{\text{cmp}} = 10^{-3}$, $\tau_{\text{cov}} = 0.40$, and $\tau_{\text{inn}} = 3 \times 10^{-4}$. The representational-support threshold τ_{rep} is calibrated separately for each model, dictionary width, and parent-set cardinality using six random and six wrong-parent controls, a 0.99 control quantile, a 0.001 margin, and a 0.01 floor. The reconstruction coefficients $\boldsymbol{\beta}$ used for child innovation are fitted by unregularized NNLS, requiring 128 fitting and 64 validation events for both child and joint support. The nonnegative prediction coefficients $\boldsymbol{\alpha}^{c,P}$ used by $\mathcal{L}_{\text{rep}}$ are likewise fitted by unregularized NNLS on child-active samples at the beginning of each fixed-graph training phase. AG-SAE training and refinement use $\lambda_\delta = 10^{-6}$, $\lambda_{\text{sp}} = 0.1$, $\lambda_{\text{rep}} = 0.02$, $\lambda_{\text{cmp}} = 1$, $m = 0.005$, $T = 0.02$, $\eta = 0.15$, and smooth-gate temperature 0.5 for both models; model-specific parameters are summarized in Table 7. The overall structural weight $\lambda_{\mathcal{I}}$ targets a 3% structural gradient contribution with a 2,000-update ramp. Training uses AdamW with batch size 2,048, zero weight decay, gradient clipping at 1.0, and target $L_0 = 50$ per bank. Residual Completion requires 64 fitting and 64 validation events, uses a reuse cosine-similarity threshold of 0.95, calibrates newly initialized feature thresholds at the 0.995 activation quantile, and caps capacity growth at 5%. Joint stability uses $\gamma_G = 0.05$ and $\gamma_F = 0.10$ over 262,144 validation samples. Training is capped at 16 cycles, followed by fresh re-induction of the terminal graph.

E.5 Structural and Semantic Evaluation

Structural evaluation. Recovered relations are evaluated using the same scoring rules and thresholds across methods, with dictionaries, graphs, and thresholds frozen before report evaluation. For the main Gemma-2-2B/MiniPile comparison, we use fixed evaluation cohorts of 800 single-parent, 400 two-parent, and 400 three-parent relations from each frozen output for every arity supported by the method. Each cohort is sampled once and reused for PSV, NR, SV, and Joint. Joint pools these fixed cohorts over all relation arities supported by the method. Thus, methods supporting both single- and multi-parent relations are evaluated over $1{,}600$ sampled relations for Joint, while single-parent-only methods use the 800-relation single-parent cohort. Nonnegative coefficients are fitted on a separate split and held fixed for evaluation. PSV compares the complete parent set with the null, proper subsets, and evaluated matched alternatives of the same cardinality in predicting the child's decoder contribution on held-out child-active samples. The score is one minus squared prediction error divided by the child's contribution energy; passing requires a margin of at least 10^{-3} over the strongest competitor. NR compares reconstruction using the parents alone with reconstruction using

the parents and child on jointly active samples; passing requires an error reduction of at least 10^{-3} after normalization by the zero-reconstruction error.

Semantic evaluation. AG-SAE and all baselines use identical prompts, evidence counts, and sampling procedures within each task, with Qwen3-30B-A3B as the evaluator (Qwen Team, 2025). Following SAEBench (Karvonen et al., 2025), feature descriptions are generated independently from ten highest-activation and five activation-weighted contexts, then frozen before relation evaluation. Single-parent SV adopts HSAE's subset-or-specialization criterion (Luo et al., 2026), using descriptions and representative activating contexts. Multi-parent SV evaluates joint semantic coverage from the descriptions, requiring every parent to be relevant and rejecting mere co-occurrence. Judges receive no method identities, feature identifiers, or quantitative scores, and semantic outputs never enter induction. Within each arity, SV reports the fraction of sampled relations satisfying the semantic criterion, and Joint reports the fraction satisfying PSV, NR, and SV simultaneously.

Single-parent semantic evaluation uses the following prompt.

```
[System]
We are studying neurons in a sparse autoencoder (SAE). Each neuron activates on specific
words, substrings, or concepts in short documents, with the activating text marked by
<<...>>.

Determine whether the Child neuron's activating concept is a subset or a more specific
version of the Parent neuron's activating concept.

Output exactly:
HaveRelationship: Yes or No
Confidence: High, Medium, or Low

Do not include any other text.

[User]
Parent neuron
Frozen semantic label: {parent_label}
Top-activation text 1: {parent_top_1}
Top-activation text 2: {parent_top_2}
Activation-weighted text: {parent_weighted}

Child neuron
Frozen semantic label: {child_label}
Top-activation text 1: {child_top_1}
Top-activation text 2: {child_top_2}
Activation-weighted text: {child_weighted}
```

Multi-parent semantic evaluation uses the following prompt, with parent labels presented in randomized order.

```
[System]
You judge whether a set of Parent SAE feature labels jointly forms a semantic envelope
that covers one Child feature label.

Answer Yes when the combined semantic scope of the Parents covers the stable core meaning
of the Child, so the Child is a natural semantic specialization, intersection, or
combination within the Parent meanings. Different Parents may cover different aspects of
the Child. No individual Parent is required to contain the entire Child.

Before deciding, check every Parent one by one. Every Parent must be directly relevant to
a core semantic aspect of the Child. If even one Parent is irrelevant, answer No. Do not
answer Yes merely because one Parent already covers the Child while ignoring the other
Parents.

Also answer No when the Parents and Child are merely related or commonly co-occur, when
the combined Parent scope misses an essential part of the Child meaning, or when the
Child is not a coherent specialization, intersection, or combination of the Parent
meanings.

Judge only the frozen feature labels. Do not use graph structure, activation strength,
quantitative scores, feature IDs, or method identity. Parent order has no meaning.

Return exactly one JSON object:
{"joint_coverage":"Yes","reason":"brief phrase"}
or
{"joint_coverage":"No","reason":"brief phrase"}

[User]
```

```
Parent feature labels:
{numbered_parent_labels}

Child feature label:
{child_label}
```

E.6 CAUSAL INTERVENTION PROTOCOL

We compare AG-SAE and Vanilla SAE on all 26 first-letter concepts, using the task introduced by Chanin et al. (2025b) and SAEBench's single-feature sparse probing procedure (Karvonen et al., 2025). For each dictionary, probing on the training split selects one leading feature per letter; feature identities, inference thresholds, and addition scales are frozen before evaluation. Both methods use the same 416 held-out prompts, with eight target-letter and eight non-target-letter examples per concept. Following SAE feature-intervention studies (Templeton et al., 2024), we intervene at the queried token's layer-13 residual stream in Gemma-2-2B, expressing decoder vectors in the original residual-stream coordinates. Addition applies the selected decoder vector scaled by its median nonzero activation on target-class training examples to non-target prompts; ablation subtracts its observed thresholded contribution from target prompts. These interventions evaluate causal control at each feature's native contribution scale. Controls use random directions orthogonal to the target decoder and unrelated letter features, each matched to the actual intervention's L2 norm on each prompt. The target margin is the target-letter logit minus the mean of the other 25 letter logits. Effects are margin increases for addition and margin decreases for ablation relative to the same prompt without intervention, including zero-activation ablation cases. To measure concentration, we cumulatively ablate the leading one to five features in the frozen sparse-probe ranking, computing all contributions from the same unmodified activation. Single-feature ablation and the cumulative curve are evaluated in separate bfloat16 forward passes, batching the four intervention controls and the six prefixes $k = 0, \ldots, 5$, respectively; the Top-1 curve value is therefore an independently executed estimate of the same leading-feature intervention. Causal yield per active feature is the mean ablation effect divided by the mean number of active intervened features. We average prompts within each letter and compare methods across letters using paired bootstrap 95% confidence intervals and exact two-sided sign tests.