INTERFERENCE BEYOND GEOMETRY IN CONCEPT EXTRACTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Interference is commonly treated as geometric overlap between learned features. We introduce *effective interference*, which combines feature geometry and code statistics to capture realized interactions, distinguishing constructive from destructive interference and frequent weak interactions from rare strong ones. Under local fixed-support assumptions, we characterize how architectural constraints shape interference through four mechanisms: feature orthogonalization, bias compensation, gain adaptation, and encoder-decoder separation. Experiments with sparse autoencoders show that constrained architectures selectively reduce overlap among co-active features, while bias, gain, and encoder freedom allow constructive cross-contributions to remain. Together, these results show that interference in learned representations depends not only on feature geometry, but also on how features are used and on the architecture that produces their codes.

1 INTRODUCTION

Mechanistic interpretability (Sharkey et al., 2025; Bereska & Gavves, 2024; Conmy et al., 2023) seeks to understand the internal mechanisms of large neural networks. Central to this goal is understanding how structure in data is encoded within these models. Two complementary hypotheses have emerged as a foundation for this effort: the *linear representation hypothesis*, in which concepts are represented as directions in activation space (Mikolov et al., 2013; Park et al., 2024; Jiang et al., 2024; Costa et al., 2025), and the *superposition hypothesis*, which proposes that sparse representations allow neural networks to encode more concepts than the dimensionality of their hidden representations (Elhage et al., 2022; Arora et al., 2018; Klindt et al., 2026).

Under superposition, an overcomplete collection of concept directions cannot be mutually orthogonal and must overlap. This overlap is commonly called *interference*: an inevitable cost of gaining representational capacity at the expense of feature separation (Elhage et al., 2022; Bereska & Gavves, 2024; Stevinson et al., 2026; Scherlis et al., 2022). Existing measures treat interference geometrically through pairwise cosine similarity (Kong et al., 2026; Gong et al., 2026) neuron capacity (Scherlis et al., 2022), or mutual coherence (Tropp, 2004), motivating methods that promote orthogonality among extracted concepts (Korznikov et al., 2025; Miller et al., 2026; Till, 2024; Costa et al., 2025).

However, recent observations challenge this view: features with sparse or weakly correlated codes may overlap with little interference (Scherlis et al., 2022; Gurnee et al., 2023), whereas correlated concepts may exploit overlap constructively (Prieto et al., 2026). Overlap may also encode meaningful rather than incidental relationships (Ivanov et al., 2026; Kantamneni & Tegmark, 2025; Gurnee & Tegmark, 2024), including cyclic structures such as days of the week and months of the year (Engels et al., 2025; Stevinson et al., 2026). Together, these observations suggest that a geometric measure of interference is incomplete without accounting for how concepts are used across the data.

Codes provide precisely this missing data-dependent information: unlike the concept dictionary, which is shared across inputs, a code specifies for each input which concepts are active and at what magnitude. Their *support* captures when concepts co-activate and therefore which geometric overlaps are realized. In contrast, a pairwise geometric measure assigns the same weight to an overlap regardless of whether the corresponding concepts frequently co-activate or never do. Their *magnitude*, in turn, captures how much a realized overlap contributes. Even when two concepts co-activate, the strength of their interaction depends on their activation magnitudes.

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.

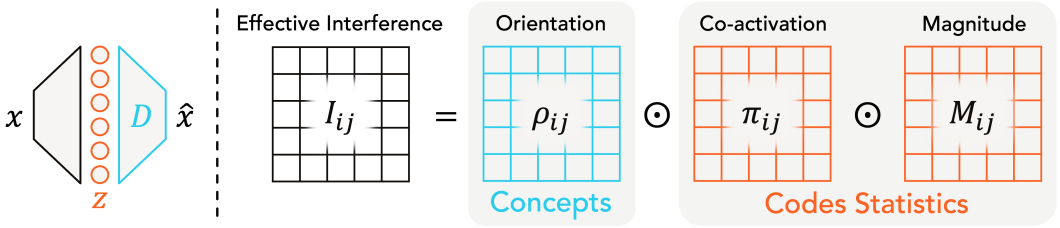


Figure 1: **Effective interference is a joint property of concept geometry and code statistics.** It factors as $I_{ij} = \rho_{ij}\pi_{ij}M_{ij}$, where $\rho_{ij} = \boldsymbol{d}_i^\top \boldsymbol{d}_j$ captures the geometric overlap between dictionary elements $\boldsymbol{d}_i, \boldsymbol{d}_j$, while $\pi_{ij}M_{ij} = \mathbb{E}[z_i z_j]$ captures how their corresponding codes z_i, z_j are used across the data, through co-activation frequency π_{ij} and conditional magnitude M_{ij}.

Interference is therefore fundamentally a joint property of concept geometry and code statistics (Figure 1). This distinction is particularly important for sparse autoencoders (SAEs) used in mechanistic interpretability (Huben et al., 2024; Gao et al., 2025; Bussmann et al., 2024; Rajamanoharan et al., 2024b; Zhu et al., 2026): codes are not observed directly, but inferred from model activations by an encoder. How the encoder is parameterized, through weight tying, bias, or relaxed norm constraints, therefore affects the interference it realizes. We summarize our main contributions below:

- **Effective interference.** We define *effective interference* from pairwise cross-contributions in the reconstruction, combining feature geometry with how the corresponding codes are used across the data. Its factorization into orientation, co-activation frequency, and conditional magnitude separates the geometric and statistical factors underlying realized interactions.
- **Encoder-dependent interference regimes.** Under local fixed-support conditions, we derive an encoder-consistency balance among self-response mismatch, directional incoming cross-talk, bias compensation, and residual coupling, and show how architectural degrees of freedom accommodate this balance through four non-exclusive routes: local orthogonalization, conditional bias compensation, encoder-gain adaptation, and encoder-decoder duality.
- **Controlled validation in SAEs.** We validate our theoretical predictions across SAE families trained on Pythia-160M activations (Biderman et al., 2023), showing that architectural constraints give rise to distinct and consistent interference regimes through orthogonalization, bias, gain, and encoder-decoder freedom. Similar patterns appear in SAEs trained on MNIST and in independently trained language-model SAEs, illustrating how effective interference captures interactions among SAE features that geometric measures alone do not.

2 Related Work

Measuring interference. Under superposition (Elhage et al., 2022), concept directions overlap, yet interference is measured almost exclusively through geometry, depending only on the dictionary. This geometric view treats interference as a cost, motivating global overlap penalties (Korznikov et al., 2025), conditional orthogonalization of co-active concepts (Costa et al., 2025), and geometric design and evaluation (Lee et al., 2025). Interference patterns also transfer across models and can be exploited by adversarial examples, linking them to network vulnerabilities (Stevinson et al., 2026; Gorton & Lewis, 2025; Gong et al., 2026). Few works consider the accompanying code statistics. Exceptions formulate superposition information-theoretically (Bereska et al., 2025), study concept co-occurrence (Li et al., 2025), use co-activation to identify causally relevant concepts (Deng et al., 2026), or assess monosemanticity via latent coherence (Filus et al., 2026). Tolooshams et al. (2025) show that Fourier-neural-operator SAEs admit broader geometric correlations because their functional codes specify where and how concepts are expressed rather than only their magnitude; in our framework, this may keep effective interference low despite geometric correlation. Ivanov et al. (2026) distinguish *incidental* overlap between unrelated concepts (Lecomte et al., 2023) from *structural* interference reflecting relationships in the data distribution (Kantamneni & Tegmark, 2025; Gurnee & Tegmark, 2024). Conditional orthogonalization (Costa et al., 2025) recognizes co-activation but modifies inference rather than measuring interference. We instead measure interference jointly via concept geometry and code statistics.

Superposition. Prior work studies superposition through generative models with prescribed sparse features and code distributions, asking how they are represented in lower-dimensional spaces (Chowdhury & Weiner, 2026). Tied bottleneck autoencoders show that sparsity permits more features than hidden dimensions at the cost of interference (Elhage et al., 2022), while subsequent theory characterizes the capacity of finite-dimensional representations to encode and compute with features (Garg et al., 2026; Hänni et al., 2024; Adler & Shavit, 2024; Borobia et al., 2026). Prieto et al. (2026) further show that co-activation statistics shape interference patterns to reflect data structure. We study the complementary inverse question: without positing a generative model, we start from model activations and ask what interference structure an extractor recovers.

Concept extraction and sparse autoencoders. Concept-based interpretability extracts directions from neural activations (Doshi-Velez & Kim, 2017; Kim et al., 2018). ACE (Ghorbani et al., 2019), ICE (Zhang et al., 2021), and CRAFT (Fel et al., 2023b) can be viewed as dictionary learning (Mairal et al., 2009; Agarwal et al., 2014; Fel et al., 2023a), while newer methods extract features locally (Hindupur et al., 2025; Shafran et al., 2026). Building on sparsity as an interpretability prior (Mairal et al., 2014; Ribeiro et al., 2016; Lipton, 2018), SAEs likewise learn concept dictionaries with sparse codes. They are widely used in mechanistic interpretability (Huben et al., 2024; Gao et al., 2025; Bussmann et al., 2024), with applications beyond language in vision and science (Pach et al., 2026; Fel et al., 2026; Adams et al., 2025; Gujral et al., 2025; Simon & Zou, 2025; Nair et al., 2026). This broad use has focused attention on how architecture and optimization shape recovered concepts. Hindupur et al. (2025) show that SAE projections encode distinct data assumptions, emphasizing alignment between the encoder and target structure. This has motivated architectures for temporal structure (Lubana et al., 2026; Bhalla et al., 2026) and concept hierarchies (Park et al., 2024), as well as objectives such as Matryoshka (Bussmann et al., 2025) for feature absorption (Chanin et al., 2024).

Theory of sparse autoencoders. Sparse autoencoders are closely related to sparse coding, or dictionary learning (Olshausen & Field, 1997). Classical recovery theory gives conditions under which gradient-based methods recover a ground-truth dictionary in shallow SAEs (Agarwal et al., 2014; Chatterji & Bartlett, 2017; Papyan et al., 2017; Rangamani et al., 2018; Nguyen et al., 2019) and deep unrolled architectures (Tolooshams & Ba, 2022), typically assuming independent sparse codes and dictionary incoherence (Donoho & Huo, 2001). Recent SAE analyses identify limits to concept recovery (Klindt et al., 2026), including insufficient sparsity (Cui et al., 2026), non-identifiability and spurious partial minima (Tang et al., 2025), and instability across runs yielding different dictionaries and codes (Nelson et al., 2026; Fel et al., 2025; Paulo & Belrose, 2026). Our closest connection is Dorrell (2026), who avoids generative ground truth and derive optimality constraints for ℓ_1-regularized dictionary learning that explain splitting and absorption (Chanin et al., 2024). We define interactions within a learned additive representation and study how an amortized encoder accommodates them across supports, yielding a frequency-magnitude interference profile and directional encoder-decoder cross-talk not captured by prior optimality constraints.

3 Effective Interference in Additive Representations

Notation. Lowercase a, bold lowercase $\boldsymbol{a}$, and uppercase $\boldsymbol{A}$ denote scalars, vectors, and matrices, respectively; $\boldsymbol{a}_i$ is the i-th column of $\boldsymbol{A}$. We use $\odot$ for the Hadamard product, $[a]_+ := \max\{a, 0\}$, and $[a]_- := \min\{a, 0\}$. For a vector $\boldsymbol{z}$, $\operatorname{supp}(\boldsymbol{z}) = \{i : z_i \neq 0\}$, and $\boldsymbol{z}$ is k-sparse if $|\operatorname{supp}(\boldsymbol{z})| = k$. Given a support S, $\boldsymbol{z}_S$ and $\boldsymbol{D}_S$ restrict $\boldsymbol{z}$ to entries in S and $\boldsymbol{D}$ to columns in S. We call $\boldsymbol{x}$ a *model activation* and $\boldsymbol{z}(\boldsymbol{x})$ its *code*, and use "directions", "features" and "concepts" interchangeably.

Consider an additive representation

$$\widehat{\boldsymbol{x}} = \boldsymbol{D}\boldsymbol{z}(\boldsymbol{x}) \;=\; \sum_{i=1}^{p} \boldsymbol{d}_i z_i(\boldsymbol{x}) \tag{3.1}$$

where $\boldsymbol{x} \sim P$, $\boldsymbol{D} \in \mathbb{R}^{m \times p}$ contains p nonzero concepts, and $\boldsymbol{z}(\boldsymbol{x}) \in \mathbb{R}^p$ their input-dependent codes. We distinguish *potential* interactions, determined by dictionary geometry, from interactions realized by the codes across the data. We assume $\mathbb{E}\|\boldsymbol{d}_i z_i(\boldsymbol{x})\|_2^2 < \infty$ for all i and, unless stated otherwise, nonnegative active codes and normalized concepts $\|\boldsymbol{d}_i\|_2 = 1$. Beyond these conditions, we assume nothing about how the codes are obtained, their sparsity, or the data-generating distribution P.

Definition 3.1 (Effective interference)**.** *For two distinct features $i \neq j$, define their sample-wise interaction as $\iota_{ij}(\boldsymbol{x}) := (\boldsymbol{d}_i^\top \boldsymbol{d}_j) z_i(\boldsymbol{x}) z_j(\boldsymbol{x})$. Their effective interference is the expected interaction*

$$I_{ij} := \mathbb{E}[\iota_{ij}(\boldsymbol{x})] = (\boldsymbol{d}_i^\top \boldsymbol{d}_j)\mathbb{E}[z_i(\boldsymbol{x}) z_j(\boldsymbol{x})]. \tag{3.2}$$

Effective interference captures the realized, code-weighted interaction between learned features, and is symmetric (i.e., $I_{ij} = I_{ji}$). Importantly, it is an expected component cross-term within a chosen decomposition. These pairwise interactions can be collected into a matrix (Figure 1), whose off-diagonal entries measure pairwise feature interactions and whose diagonal entries capture individual feature contributions. This distinction follows from the reconstruction-energy decomposition

$$\mathbb{E}\|\widehat{\boldsymbol{x}}\|_2^2 = \sum_{i=1}^{p} E_i + \sum_{i=1}^{p}\sum_{j\neq i} I_{ij} = \sum_{i=1}^{p} E_i + 2\sum_{i<j} I_{ij}, \tag{3.3}$$

where $E_i := \mathbb{E}\|\boldsymbol{d}_i \boldsymbol{z}_i(\boldsymbol{x})\|_2^2$ and I_{ij} denote individual and pairwise contributions, respectively. Thus, $I_{ij} > 0$ increases reconstruction energy relative to the individual contributions and is *constructive*, whereas $I_{ij} < 0$ is *destructive*. These terms do not by themselves imply benefit or harm to reconstruction error or downstream behaviour. Moreover, this terminology describes how components combine within $\widehat{\boldsymbol{x}}$, not their effect on reconstruction error: the sign of I_{ij} alone does not determine the change in $\mathbb{E}\|\widehat{\boldsymbol{x}} - \boldsymbol{x}\|_2^2 = \mathbb{E}\|\widehat{\boldsymbol{x}}\|_2^2 - 2\mathbb{E}[\boldsymbol{x}^\top \widehat{\boldsymbol{x}}] + \mathbb{E}\|\boldsymbol{x}\|_2^2$, since the alignment term $-2\mathbb{E}[\boldsymbol{x}^\top \widehat{\boldsymbol{x}}]$ also contributes.

As a joint geometric–statistical quantity, effective interference has three useful properties on support awareness, reparameterization invariance, and sign under nonnegative coding (see Proposition D.1). Within this formulation, geometry determines interaction orientation, whereas codes determine whether and how strongly it is realized. Hence, unlike cosine similarity, effective interference is support-aware and scale-aware, offered by the code statistics.

Interference profile. The second moment $\mathbb{E}[z_i z_j]$ combines how often two features co-activate with how large their components are when they do. To separate these effects, define

$$a_i := \mathbf{1}\{z_i \neq 0\}, \qquad \pi_{ij} := \Pr(a_i = 1, a_j = 1), \qquad \rho_{ij} := \boldsymbol{d}_i^\top \boldsymbol{d}_j, \tag{3.4}$$

where a_i is the activation indicator, π_{ij} is the pair's co-activation frequency, and ρ_{ij} is its geometric orientation. For $\pi_{ij} > 0$, let $M_{ij} := \mathbb{E}[z_i z_j \mid a_i = 1, a_j = 1]$. Hence, effective interference can be factorized. For every pair with $\pi_{ij} > 0$, the effective interference can be factorized as follows:

$$\boxed{I_{ij} = \rho_{ij}\, \pi_{ij}\, M_{ij}} \tag{3.5}$$

If $\pi_{ij} = 0$, then $I_{ij} = 0$ and the conditional magnitude M_{ij} need not be defined. We call $\Pi_{ij} := (\rho_{ij}, \pi_{ij}, M_{ij})$ the pair's *effective-interference profile*. I_{ij} measures its average contribution, while Π_{ij} distinguishes geometric orientation, interaction exposure, and conditional magnitude. Consequently, equal values of I_{ij} may arise from frequent weak co-activation or rare strong co-activation.

To compare interference across representations, we separate positive and negative pairwise contributions and define the normalized signed interferences

$$\mathcal{I}_+ := \frac{\mathbb{E}\left[2\sum_{i<j}[\iota_{ij}(\boldsymbol{x})]_+\right]}{\mathbb{E}[\|\widehat{\boldsymbol{x}}\|_2^2]}, \qquad \mathcal{I}_- := \frac{\mathbb{E}\left[2\sum_{i<j}[\iota_{ij}(\boldsymbol{x})]_-\right]}{\mathbb{E}[\|\widehat{\boldsymbol{x}}\|_2^2]}, \qquad \mathcal{I}_{\text{net}} := \mathcal{I}_+ + \mathcal{I}_-. \tag{3.6}$$

The signed terms distinguish constructive from destructive interference, while $\mathcal{I}_{\text{net}}$ captures their net effect and may vanish through cancellation even when both are large.

Overall, the interference profile distinguishes several regimes that a single geometric or aggregate statistic conflates. Large $|\rho_{ij}|$ together with $\pi_{ij} \approx 0$ describes *unrealized overlap*: the directions admit interaction, but their supports prevent it. Large π_{ij} and small M_{ij} describe frequent weak interactions, whereas small π_{ij} and large M_{ij} describe rare strong interactions.

4 How Sparse Autoencoders Handle Interference

The effective-interference profile characterizes how features interact in a learned representation. The architecture and optimization of the extractor determine which profiles arise. Consider a typical amortized SAE architecture (Huben et al., 2024; Bricken et al., 2023; Gao et al., 2025),

$$\boldsymbol{z} = \sigma(\boldsymbol{W}^\top \boldsymbol{x} + \boldsymbol{b}), \qquad \widehat{\boldsymbol{x}} = \boldsymbol{D}\boldsymbol{z}, \tag{4.1}$$

where $\boldsymbol{W}$ and $\boldsymbol{D}$ are the encoder and decoder weights, $\boldsymbol{b}$ is the encoder bias, and σ is a sparsifying nonlinearity. Unlike normalized concepts in $\boldsymbol{D}$, encoder weights $\boldsymbol{W}$ need not be normalized. An

SAE shapes interference at two levels: encoder scores together with σ determine the support and hence π_{ij}, while the encoder and decoder govern interactions among co-active features and hence ρ_{ij} and M_{ij}. We study the latter on a fixed support, so our conclusions are local to an active region and, for gradient dynamics, to updates that preserve the support. Within a fixed support S, the encoder is locally affine:

$$\boldsymbol{z}_S = \boldsymbol{W}_S^\top \boldsymbol{x} + \boldsymbol{b}_S, \qquad \boldsymbol{z}_{S^c} = 0. \tag{4.2}$$

Rather than optimizing codes separately for each input, an SAE predicts them with a shared encoder. The map in (4.2) thus amortizes fixed-support least squares across inputs. Proposition 4.1 sets code-optimality condition for each input example. We treat this condition as an ideal target, and develops this section's analyses for when the code-optimality gap is small. This code-optimality gap relates to amortization gap, and can be small for in-distribution data (Margossian & Blei, 2024) (Section 5 further investigates this in SAEs).

Proposition 4.1 (Code-optimality condition)**.** *For a fixed decoder $\boldsymbol{D}$ and support S, any interior least-squares optimum satisfies*

$$\boldsymbol{D}_S^\top \boldsymbol{r} = 0, \qquad \boldsymbol{r} := \boldsymbol{x} - \boldsymbol{D}_S \boldsymbol{z}_S. \tag{4.3}$$

For nonnegative codes, interior means $z_i > 0 \; \forall i \in S$.

Code optimality is a decoder condition: it constrains $\boldsymbol{D}$ given the codes, but not the encoder $\boldsymbol{W}$. We therefore require an encoder consistency identity (Proposition 4.2). Let $\alpha_i = \|\boldsymbol{w}_i\|_2$ and $\gamma_{ij} := \boldsymbol{w}_i^\top \boldsymbol{d}_j / \|\boldsymbol{w}_i\|_2$ be the cosine similarity between $\boldsymbol{w}_i$ and $\boldsymbol{d}_j$, so that $\boldsymbol{w}_i^\top \boldsymbol{d}_j = \alpha_i \gamma_{ij}$. Unlike $\boldsymbol{D}^\top \boldsymbol{D}$, which captures interactions among reconstruction components, $\boldsymbol{W}^\top \boldsymbol{D}$ couples encoder and decoder directions and is asymmetric.

Proposition 4.2 (Encoder consistency identity)**.** *Suppose the encoder is locally affine as in (4.2). For every active feature $i \in S$,*

$$0 = (\alpha_i \gamma_{ii} - 1) z_i + \sum_{j \neq i} \alpha_i \gamma_{ij} z_j + b_i + \boldsymbol{w}_i^\top \boldsymbol{r}. \tag{4.4}$$

Consequently, for every i with $\Pr(a_i = 1) > 0$,

$$0 = \underbrace{(\alpha_i \gamma_{ii} - 1)\mathbb{E}[z_i \mid a_i = 1]}_{\textit{self-response mismatch}} + \underbrace{\sum_{j \neq i} \alpha_i \gamma_{ij} \mathbb{E}[z_j \mid a_i = 1]}_{\textit{incoming cross-talk}} + \underbrace{b_i}_{\substack{\textit{bias} \\ \textit{compensation}}} + \underbrace{\mathbb{E}[\boldsymbol{w}_i^\top \boldsymbol{r} \mid a_i = 1]}_{\textit{residual coupling}}, \tag{4.5}$$

where γ_{ii} and γ_{ij} are encoder-decoder self-alignment and cross-alignment, respectively. The encoder gain α_i rescales the entire response of encoder i, including its self-response, incoming cross-talk, and residual coupling. With the code-optimality condition, Proposition 4.2 exposes which architectural terms can accommodate interactions. Its incoming cross-talk is directional: the effect of feature j on the encoder score of feature i depends on $\boldsymbol{w}_i^\top \boldsymbol{d}_j$. This contrasts with the symmetric decoder interference, $I_{ij} = (\boldsymbol{d}_i^\top \boldsymbol{d}_j)\mathbb{E}[z_i z_j]$. Moreover, for tied weights, (4.3) removes the residual coupling term in (4.5), but not generally for an untied encoder, since $\boldsymbol{D}_S^\top \boldsymbol{r} = 0$ does not imply $\boldsymbol{W}_S^\top \boldsymbol{r} = 0$.

To isolate the role of each term in the encoder-consistency balance, we begin with a tied, unit-norm, bias-free baseline and relax one constraint at a time by introducing bias, encoder gain, and independent encoder directions. Holding the remaining degrees of freedom fixed allows each change in interference to be attributed to a specific accommodation mechanism. These settings serve as analytical controls rather than assumptions about practical SAEs.

Route I: constrained local orthogonalization. Consider the tied and normalized $\alpha_i = 1$, $\gamma_{ii} = 1$, $\gamma_{ij} = \rho_{ij}$, and bias-free $\boldsymbol{b} = 0$ setting. Under the fixed-support code optimality condition per input (4.3), the encoder-consistency condition (4.4) gives

$$(\boldsymbol{D}_S^\top \boldsymbol{D}_S - \boldsymbol{I})\boldsymbol{z}_S = 0. \tag{4.6}$$

For a single input, this only constrains $\boldsymbol{D}_S^\top \boldsymbol{D}_S - \boldsymbol{I}$ along the observed code vector $\boldsymbol{z}_S$. However, if the code vectors observed on support S span $\mathbb{R}^{|S|}$, the condition must hold in every direction and therefore implies $\boldsymbol{D}_S^\top \boldsymbol{D}_S = \boldsymbol{I}$. Thus, tied normalized SAEs are driven toward orthogonality among concepts that co-activate, rather than across the entire overcomplete dictionary.

The same local orthogonalizing tendency can be seen directly from the training dynamics. For two active features $S = \{i, j\}$, define $\Delta(\boldsymbol{D}_S) := \|\boldsymbol{D}_S^\top \boldsymbol{D}_S - \boldsymbol{I}\|_F^2 = 2\rho_{ij}^2$, where the second equality follows from unit-norm columns.

Proposition 4.3 (Local orthogonalization for two active features)**.** *Consider the tied, normalized, bias-free model with fixed support $S = \{i, j\}$ and loss $\frac{1}{2}\|\widehat{\boldsymbol{x}} - \boldsymbol{x}\|_2^2$. Let $\boldsymbol{D}_S^+$ denote the active dictionary after one gradient step of size η, followed by column normalization. Then*

$$\Delta(\boldsymbol{D}_S^+) - \Delta(\boldsymbol{D}_S) = -8\eta\rho_{ij}^2\|z_i\boldsymbol{d}_i - z_j\boldsymbol{d}_j\|_2^2 + O(\eta^2). \tag{4.7}$$

Hence, if $\rho_{ij} \neq 0$ and $z_i\boldsymbol{d}_i \neq z_j\boldsymbol{d}_j$, a gradient step cannot increase $\Delta(\boldsymbol{D}_S)$ to first order.

Equation (4.7) identifies an orthogonalizing contribution on examples where the pair is jointly active. Thus, π_{ij} controls the frequency of this direct two-feature contribution, although other support states can also change the pairwise angle. The dynamics therefore predict selective orthogonalization of frequently co-active directions, not global dictionary orthogonalization. Proposition 4.3 concerns one-example gradient descent with a support-preserving step. Section 5 further shows experiments with higher k, batch updates, and Adam optimizer where this predicted tendency still remains.

Route II: conditional bias compensation. We next retain tied, unit-norm encoder-decoder weights but allow a learned encoder bias. Tying implies $\alpha_i = 1$, $\gamma_{ii} = 1$, and $\gamma_{ij} = \rho_{ij}$ for $i \neq j$. Hence, the self-response mismatch vanishes, while fixed-support code optimality removes residual coupling. We obtain the following population-level condition:

$$\hat{b}_i = -\sum_{j \neq i} \rho_{ij}\mathbb{E}[z_j \mid a_i = 1]. \tag{4.8}$$

Thus, a shared bias can compensate for the *mean incoming cross-talk* on examples where feature i is active. It cannot generally cancel the sample-wise cross-talk separately for every input.

Route III: encoder-gain adaptation. We next keep the encoder direction tied to its corresponding decoder direction, but allow the encoder norm to vary: $\boldsymbol{w}_i = \alpha_i\boldsymbol{d}_i$, with $\alpha_i > 0$. Consequently, $\gamma_{ii} = 1$ and $\gamma_{ij} = \rho_{ij}$. In this setting, fixed-support code optimality also gives $\boldsymbol{w}_i^\top \boldsymbol{r} = 0$. The conditional balance (Proposition 4.2) therefore gives the following estimate on the encoder gain

$$\hat{\alpha}_i = \left(1 + \frac{\sum_{j \neq i} \rho_{ij}\mathbb{E}[z_j \mid a_i = 1]}{\mathbb{E}[z_i \mid a_i = 1]}\right)^{-1}. \tag{4.9}$$

Thus, positive mean incoming cross-talk predicts $\alpha_i < 1$, whereas negative incoming cross-talk predicts $\alpha_i > 1$, provided the denominator is positive. Encoder gain rescales both the self-response and incoming cross-talk, providing an alternative to decoder angular orthogonalization. This behaviour manifests itself in the unnormalized learning dynamics (Proposition F.1): the two-feature angular dynamics depend on both the pair's cosine similarity and its weight norm.

Route IV: encoder-decoder duality. Untying separates decoder interaction from inference cross-talk. For each pair, $(\rho_{ij}, \pi_{ij}M_{ij})$ describes decoder interaction, while $\boldsymbol{w}_i^\top \boldsymbol{d}_j = \alpha_i\gamma_{ij}$ is the directed cross-gain through which decoder feature j enters encoder score i. Fixed-support least squares admits a biorthogonal encoder (D.9), with targets $\alpha_i\gamma_{ii} = 1$ and $\alpha_i\gamma_{ij} = 0$ for $j \neq i$. Because different supports induce different duals, a shared encoder can only approximate these targets jointly. We therefore predict small $|\alpha_i\gamma_{ij}|$ for pairs with large $\pi_{ij}M_{ij}$ even when $\rho_{ij} > 0$, particularly when residual coupling is small. When α_i is bounded away from zero, this also predicts γ_{ij} near zero. Thus, untying can preserve constructive effective interference while suppressing incoming cross-talk, reflecting the roles of encoders as feature detectors and decoders as reconstruction components (Bricken et al., 2024; Rajamanoharan et al., 2024a; Nanda, 2023).

The four routes are controlled limits of the same balance and are not mutually exclusive. A practical encoder may distribute accommodation across decoder geometry, encoder gain, encoder-decoder alignment, bias, and residual coupling. We now show how these signatures recur across SAE families.

5 Results

We evaluate whether the architectural degrees of freedom identified in Section 4 predict the interference structure learned by SAEs. Our experiments use TopK (Gao et al., 2025; Makhzani & Frey, 2013), JumpReLU (Rajamanoharan et al., 2024b), BatchTopK (Bussmann et al., 2024), and Matryoshka (Bussmann et al., 2025) SAEs trained on Pythia-160M-deduped residual-stream activations at layer 8 with dictionary width $p = 16{,}384$ and sparsity $k = 40$. For additional results and experimental details, see Sections I and J, respectively.

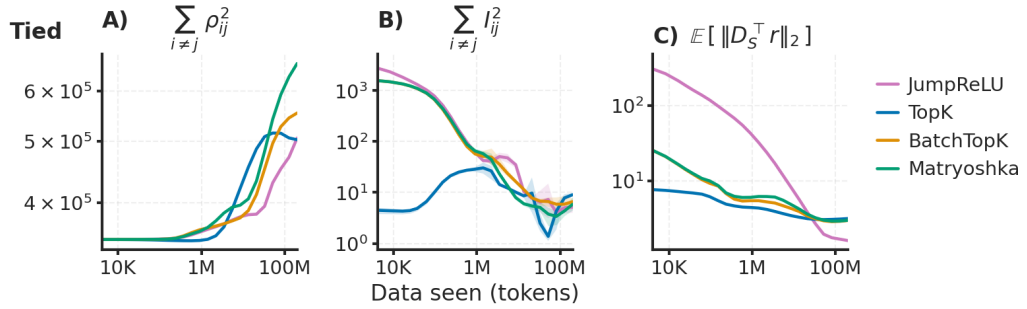


Figure 2: **Geometric and effective interference have distinct training dynamics. A)** Global geometric overlap rises, whereas **B)** effective interference falls from its peak. Consistent with Proposition 4.3, this reflects selective orthogonalization of co-active pairs, reducing realized interactions without globally orthogonalizing the dictionary. **C)** Residual coupling declines but remains nonzero, so fixed-support code optimality (Proposition 4.1) is approached only approximately (see Figure 7 for joint distribution of decoder orientation ρ_{ij} and co-activation-weighted magnitude $\pi_{ij} M_{ij}$).

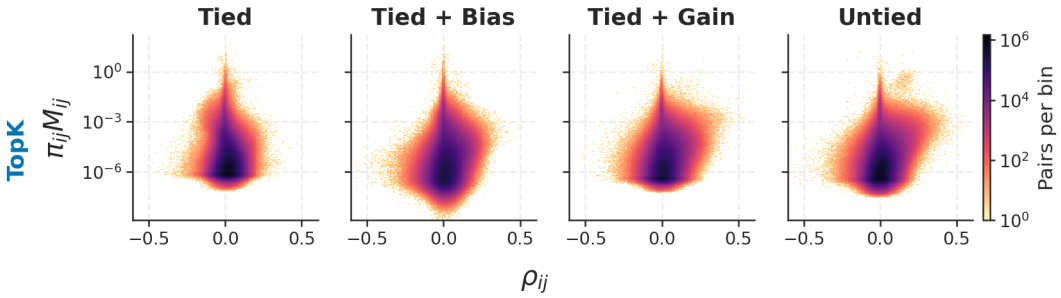


Figure 3: **Architectural freedom shifts co-active feature pairs toward constructive decoder interactions.** Joint distributions of orientation ρ_{ij} and co-activation-weighted magnitude $\pi_{ij} M_{ij}$. In the tied, normalized, bias-free model, pairs with large $\pi_{ij} M_{ij}$ concentrate near $\rho_{ij} = 0$, consistent with local orthogonalizing pressure. With bias, encoder-gain freedom, or an untied encoder, some of these pairs extend toward positive ρ_{ij}. This pattern recurs across other SAEs (Figure 8).

How to interpret the results. We do not rank SAE families by absolute interference. Instead, each family tests the same mechanisms by measuring how untying, encoder bias, and gain adaptation alter interference and encoder-decoder cross-talk. The evidence lies in the direction and recurrence of these within-family changes, not absolute differences across sparsifiers.

We employ the code statistics on activation frequency and magnitude, $\pi_{ij} M_{ij}$, together with the orientation entry ρ_{ij} to show how pairwise interaction is allocated between the decoder and codes. For when encoder gain is enabled, we report α_i and encoder-decoder cross-alignment γ_{ij}.

Constrained SAEs suppress realized interactions without globally orthogonalizing. Proposition 4.3 predicts support-local rather than global orthogonalizing pressure. Across SAE families, global squared geometric overlap increases during training, whereas squared effective interference falls from its peak (Figure 2A-B). Thus, the dictionary need not become globally orthogonal for realized interactions to decrease; code statistics determine which overlaps matter. This behaviour is consistent with selective pressure on co-active pairs and refines the global orthogonality picture suggested by the linear representation hypothesis (Costa et al., 2025). The code-optimality gap $\mathbb{E}[\|\boldsymbol{D}_S^\top \boldsymbol{r}\|_2]$ also decreases but remains nonzero (Figure 2C), showing that fixed-support optimality is an increasingly accurate, but still approximate, description of the learned codes. This behaviour is aligned with prior investigations of amortization gap in SAEs (O'Neill et al., 2025).

Architectural freedom enables constructive decoder interactions. The joint interference profiles in Figure 3 reveal a common effect of relaxing architectural constraints. In the tied, normalized, bias-free model, pairs with large $\pi_{ij} M_{ij}$ concentrate near $\rho_{ij} = 0$ or shift toward negative. Adding bias or encoder gain, or untying the encoder, allows such pairs to extend toward positive ρ_{ij} and therefore retain constructive effective interference. These profiles establish the common decoder behaviour but do not identify how each architecture accommodates the resulting interactions; the encoder-consistency balance (4.5) provides this distinction.

With bias, the predicted $\hat{b}_i$ closely matches the learned b_i across SAE families (Figure 4A), supporting compensation of mean incoming cross-talk. TopK and BatchTopK lie closest to equality, whereas Matryoshka and JumpReLU deviate more and correlate less. Because $\hat{b}_i - b_i$ equals the residual-coupling term in this controlled setting, these deviations indicate a less exact fixed-support approximation: its mean magnitude is $0.11 - 0.14$ for Matryoshka and JumpReLU against $0.04 - 0.05$ for TopK and BatchTopK. With encoder gain, $\hat{\alpha}_i$ similarly tracks the learned α_i (Figure 4B), with the same ordering ($0.11 - 0.15$ versus $0.03 - 0.06$). Positive mean incoming cross-talk predicts $\alpha_i < 1$, allowing the encoder to attenuate its self-response and incoming cross-talk without changing decoder normalization. This mechanism parallels Prieto et al. (2026), who observe increased constructive interference on toy models trained with weight decay. Similarly, by training an unnormalized decoder, a smaller decoder norms relax the pressure toward local orthogonality (Proposition F.1), allowing positively aligned directions to constructively reconstruct.

Untying shifts decoder pairs toward positive alignment ρ_{ij} and increases positive effective interference I_{ij} (Figure 5 left), showing that decoder overlap need not be reduced to control inference cross-talk. The pairwise profiles reveal the mechanism: pairs with large $\pi_{ij}M_{ij}$ exhibit a pronounced tilt toward positive ρ_{ij}, while their encoder-decoder cross-gains $\alpha_i\gamma_{ij} = \boldsymbol{w}_i^\top \boldsymbol{d}_j$ remain thin and concentrated near zero (Figure 5 right). Untying therefore separates reconstruction from inference: the decoder retains constructive interactions, while the encoder limits how decoded feature j enters the score of feature i. This pattern is consistent with approximate support-conditional biorthogonalization, although these pairwise distributions do not establish exact duality on every support.

Architectural freedom reallocates interference handling. We decompose the normalized off-diagonal contribution into constructive $\mathcal{I}_+$ and destructive $\mathcal{I}_-$ terms, with $\mathcal{I}_{\text{net}} = \mathcal{I}_+ + \mathcal{I}_-$ (Figure 6). This separation is essential: in the tied baseline, sizable positive and negative contributions largely cancel, particularly for JumpReLU, so a small net value does not imply weak feature interactions. Within every SAE family, introducing bias, encoder gain, or independent encoder directions shifts the balance toward positive net interference relative to the tied baseline, although the intermediate variants are not strictly ordered because the accommodation mechanisms interact rather than combine additively. Untying produces the largest constructive contribution across all four families: $\mathcal{I}_+ + |\mathcal{I}_-|$ accounts for roughly 40–50% of reconstruction energy, while the net contribution remains approximately 15–25%. Effective interference is therefore not a small residual effect, nor something SAEs generally minimize; architectural constraints

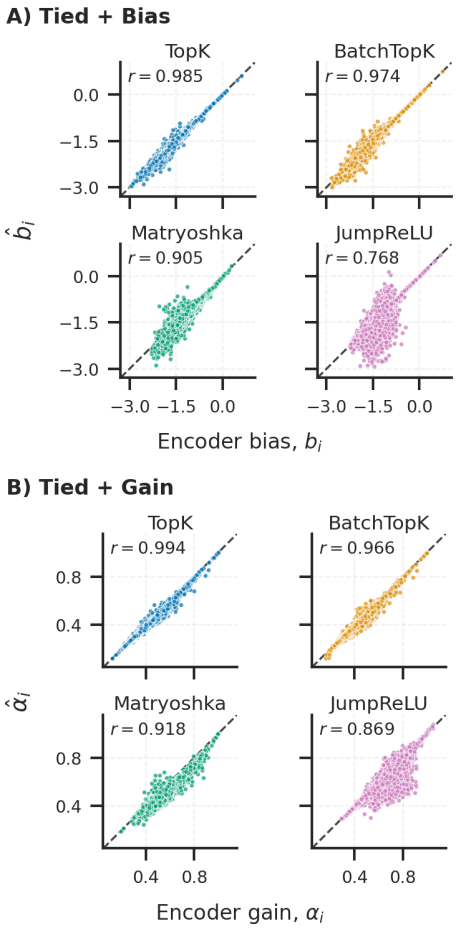


Figure 4: **Bias and encoder gain compensate mean incoming cross-talk.** Each point is a feature; horizontal axes show learned parameters and vertical axes their encoder-balance predictions. **A)** In tied + bias models, $\hat{b}_i$ from (4.8) closely matches b_i. **B)** In tied + gain models, $\hat{\alpha}_i$ from (4.9) tracks α_i. Dashed lines indicate equality and r is Pearson correlation. This behaviour appears across other SAEs (Figure 9).

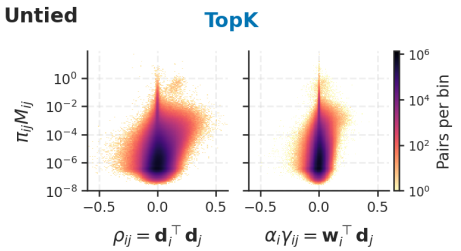


Figure 5: **Untying separates decoder interactions from encoder cross-talk.** Joint distributions of co-activation-weighted magnitude $\pi_{ij}M_{ij}$ against decoder orientation ρ_{ij} and cross-gain $\alpha_i\gamma_{ij} = \boldsymbol{w}_i^\top \boldsymbol{d}_j$. Pairs with large $\pi_{ij}M_{ij}$ tilt toward positive decoder alignment, while their cross-gains concentrate near zero.

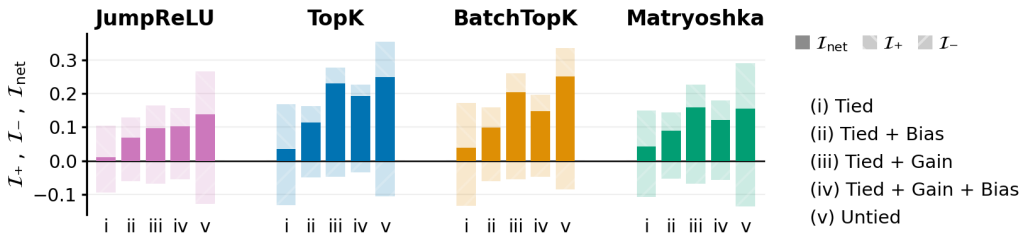


Figure 6: **SAE reconstruction contains substantial signed effective interference.** Constructive $\mathcal{I}_+$, destructive $\mathcal{I}_-$, and net $\mathcal{I}_{\text{net}}$ contributions across architectural variants and SAE families. Sizable positive and negative terms can cancel, so small net interference does not imply weak interactions. Relative to the tied case, architectural freedom generally shifts the balance toward higher interference.

determine whether cross-feature interactions are canceled or retained constructively. This refines a purely geometric view of the LRH (Park et al., 2024): interference depends not only on dictionary overlap, but also on which features are jointly used. Finally, we demonstrate the generality of these interference signatures beyond our main setting using raw MNIST images (Section J.3.1) and pre-trained language-model SAEs (Section J.3.2).

Illustrating the interference profile across support size k and dictionary width p. As a diagnostic example, we vary k and p in TopK SAEs (Figure 10). Let $\tilde{p}$ denote the number of utilized features, accounting for features that never activate. The mean co-activation frequency $\overline{\pi}_{ij}$ closely follows the reference $k(k-1)/\tilde{p}(\tilde{p}-1)$ (Section D.5). At fixed p, increasing k raises $\overline{\pi}_{ij}$ while reducing the mean conditional magnitude $\overline{M}_{ij}$. Mean absolute decoder overlap $\overline{|\rho_{ij}|}$ also generally decreases in the biased, gain-adaptive, and untied variants. Net interference $\mathcal{I}_{\text{net}}$, however, has no universal monotonic trend, indicating that increased co-activation can be offset by weaker magnitudes and signed geometric contributions. At fixed k, increasing p spreads co-activation across more pairs, reducing $\overline{\pi}_{ij}$, while $\overline{M}_{ij}$ and $\overline{|\rho_{ij}|}$ change less in most settings. The largest tied configurations also show elevated FVU and are therefore interpreted cautiously. This illustrative experiment shows how the interference profile separates compensating statistical and geometric changes without asserting a universal scaling law.

6 Conclusion

Discussion and limitations. Effective interference measures realized pairwise cross-contributions within a learned additive decomposition, allowing us to study how SAE architectures allocate reconstruction energy and which interaction regimes they favor. It does not characterize interactions among latent ground-truth concepts, as studied by Prieto et al. (2026), nor is it invariant to phenomena such as splitting or absorption; identifying these requires evidence beyond I_{ij} and aligned decompositions across extractors. More generally, effective interference diagnoses energy allocation, not representation quality: constructive and destructive describe cross-term signs without implying reconstruction quality, semantic interpretability, causal relevance, or downstream behaviour.

Our theory is local to fixed supports and approximates a shared amortized encoder by per-input code optimality. This approximation can be accurate in-distribution (Margossian & Blei, 2024) but degrade under distribution shift (O'Neill et al., 2025), consistent with our small but nonzero measured gaps. Similarly, our local orthogonalization proposition considers two active features and a small, support-preserving gradient step, and thus does not establish global convergence or cover larger supports, minibatch training, adaptive optimizers, or support transitions. Overall, our experiments support the proposed compensation and orthogonalization mechanisms without uniquely identifying a single mechanism.

Conclusion. Interference is not a property of feature geometry alone, but of geometry together with how features co-activate. We formalized this interaction as *effective interference* and showed that SAE architectures realize distinct interference regimes: constrained encoders tend to orthogonalize co-active features, whereas bias, gain, and encoder freedom allow constructive cross-contributions to remain. These results provide a framework for characterizing how architectural choices shape the interactions realized by learned representations, without treating any single interference regime as universally preferable.

REFERENCES

Etowah Adams, Liam Bai, Minji Lee, Yiyang Yu, and Mohammed AlQuraishi. From mechanistic interpretability to mechanistic biology: Training, evaluating, and interpreting sparse autoencoders on protein language models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=zdOGBRQEbz.

Micah Adler and Nir Shavit. On the complexity of neural computation in superposition. *arXiv preprint arXiv:2409.15318*, 2024.

Alekh Agarwal, Animashree Anandkumar, Prateek Jain, Praneeth Netrapalli, and Rashish Tandon. Learning sparsely used overcomplete dictionaries. In Maria Florina Balcan, Vitaly Feldman, and Csaba Szepesvári (eds.), *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pp. 123–137, Barcelona, Spain, 13–15 Jun 2014. PMLR. URL https://proceedings.mlr.press/v35/agarwal14a.html.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495, 2018.

Leonard Bereska and Stratis Gavves. Mechanistic interpretability for AI safety - a review. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=ePUVetPKu6.

Leonard Bereska, Zoe Tzifa-Kratira, Reza Samavi, and Stratis Gavves. Superposition as lossy compression — measure with sparse autoencoders and connect to adversarial vulnerability. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=qaNP6o5qvJ.

Usha Bhalla, Alex Oesterling, Claudio Mayrink Verdun, Himabindu Lakkaraju, and Flavio Calmon. Temporal sparse autoencoders: Leveraging the sequential nature of language for interpretability. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=bojVI4l9Kn.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling, 2023. URL https://arxiv.org/abs/2304.01373.

Joseph Bloom, Curt Tigges, Anthony Duong, and David Chanin. Saelens. https://github.com/decoderesearch/SAELens, 2024.

Hector Borobia, Elies Seguí-Mas, and Guillermina Tormo-Carbó. Linear-readout floors and threshold recovery in computation in superposition. *arXiv preprint arXiv:2605.01192*, 2026.

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. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Trenton Bricken, Jonathan Marcus, Siddharth Mishra-Sharma, Meg Tong, Ethan Perez, Mrinank Sharma, Kelley Rivoire, Thomas Henighan, and Adam Jermyn. Using dictionary learning features as classifiers. 2024. URL https://transformer-circuits.pub/2024/features-as-classifiers/index.html.

Bart Bussmann, Patrick Leask, and Neel Nanda. Batchtopk sparse autoencoders. In *NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning*, 2024. URL https://openreview.net/forum?id=d4dpOCqybL.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with matryoshka sparse autoencoders. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 6077–6101. PMLR, 13–19 Jul 2025.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, and Joseph Isaac Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. In *Interpretable AI: Past, Present and Future*, 2024. URL https://openreview.net/forum?id=Wzav8fesTL.

Niladri S Chatterji and Peter L Bartlett. Alternating minimization for dictionary learning: Local convergence guarantees. *arXiv preprint arXiv:1711.03634*, 2017.

Mriganka Basu Roy Chowdhury and Eric McLaughlin Weiner. Effects of sparsity and superposition on loss in simple autoencoders, 2026. URL https://arxiv.org/abs/2606.18538.

Tom Conerly, Adly Templeton, Trenton Bricken, Jonathan Marcus, and Tom Henighan. Update on how we train SAEs. Transformer Circuits Thread, April 2024.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=89ia77nZ8u.

Valérie Costa, Thomas Fel, Ekdeep Singh Lubana, Bahareh Tolooshams, and Demba E. Ba. From flat to hierarchical: Extracting sparse representations with matching pursuit. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=Ll5miDx8KB.

Jingyi Cui, Qi Zhang, Yifei Wang, and Yisen Wang. On the limits of sparse autoencoders: A theoretical framework and reweighted remedy. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=DSOTgzeH3w.

Ruixuan Deng, Xiaoyang Hu, Miles Gilberti, Shane Storks, Aman Taxali, Mike Angstadt, Chandra Sripada, and Joyce Chai. Sparse feature coactivation reveals causal semantic modules in large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3419–3451, 2026.

D.L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47(7):2845–2862, 2001. doi: 10.1109/18.959265.

William Dorrell. How optimality structures sparse dictionaries: A theory for understanding sae representations. *arXiv preprint arXiv:2606.02385*, 2026.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Scott C Douglas, Shun-ichi Amari, and S-Y Kung. On gradient adaptation with unit-norm constraints. *IEEE Transactions on Signal processing*, 48(6):1843–1847, 2002.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition, 2022. URL https://arxiv.org/abs/2209.10652.

Josh Engels, Eric Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *International Conference on Learning Representations*, volume 2025, pp. 84591–84622, 2025.

Thomas Fel, Victor Boutin, Louis Béthune, Rémi Cadène, Mazda Moayeri, Léo Andéol, Mathieu Chalvidal, and Thomas Serre. A holistic approach to unifying automatic concept extraction and concept importance estimation. *Advances in Neural Information Processing Systems*, 36: 54805–54818, 2023a.

Thomas Fel, Agustin Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. Craft: Concept recursive activation factorization for explainability. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2711–2721, 2023b.

Thomas Fel, Ekdeep Singh Lubana, Jacob S. Prince, Matthew Kowal, Victor Boutin, Isabel Papadimitriou, Binxu Wang, Martin Wattenberg, Demba E. Ba, and Talia Konkle. Archetypal SAE: Adaptive and stable dictionary learning for concept extraction in large vision models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=9v1eW8HgMU.

Thomas Fel, Binxu Wang, Michael Lepori, Matthew Kowal, Andrew Lee, Randall Balestriero, Sonia Joseph, Ekdeep Singh Lubana, Talia Konkle, Demba Ba, et al. Into the rabbit hull: From task-relevant concepts in dino to minkowski geometry. In *International Conference on Learning Representations*, volume 2026, pp. 154538–154584, 2026.

Katarzyna Filus et al. Measuring monosemanticity in sparse autoencoders via latent activation coherence. *arXiv preprint arXiv:2607.17770*, 2026.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling, 2020. URL https://arxiv.org/abs/2101.00027.

Leo Gao, Tom Dupre 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.

Nikhil Garg, Jon Kleinberg, and Kenny Peng. How many features can a language model store under the linear representation hypothesis? In Steve Hanneke and Tor Lattimore (eds.), *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 5358–5376. PMLR, 29 Jun–03 Jul 2026.

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems*, pp. 9273–9282, 2019.

Bofan Gong, Shiyang Lai, James Evans, and Dawn Song. Signal in the noise: Polysemantic interference transfers and predicts cross-model influence. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=tYeuz2LwVU.

Liv Gorton and Owen Lewis. Adversarial examples are not bugs, they are superposition. In *Mechanistic Interpretability Workshop at NeurIPS 2025*, 2025. URL https://openreview.net/forum?id=z0YTS0FExr.

Onkar Gujral, Mihir Bafna, Eric Alm, and Bonnie Berger. Sparse autoencoders uncover biologically interpretable features in protein language model representations. *Proceedings of the National Academy of Sciences*, 122(34):e2506316122, 2025.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jE8xbmvFin.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. Finding neurons in a haystack: Case studies with sparse probing. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=JYs1R9IMJr.

Kaarel Hänni, Jake Mendel, Dmitry Vaintrob, and Lawrence Chan. Mathematical models of computation in superposition. *arXiv preprint arXiv:2408.05451*, 2024.

Sai Sumedh R. Hindupur, Ekdeep Singh Lubana, Thomas Fel, and Demba E. Ba. Projecting assumptions: The duality between sparse autoencoders and concept geometry. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=SoA8rMxDaF.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, volume 2024, pp. 7827–7845, 2024.

Georgi Ivanov, Narmeen Oozeer, Shivam Raval, Tasana Pejovic, Shriyash Upadhyay, and Amir Abdullah. Spectral superposition: A theory of feature geometry. *arXiv preprint arXiv:2602.02224*, 2026.

Yibo Jiang, Goutham Rajendran, Pradeep Kumar Ravikumar, Bryon Aragam, and Victor Veitch. On the origins of linear representations in large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 21879–21911. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/jiang24d.html.

Subhash Kantamneni and Max Tegmark. Language models use trigonometry to do addition. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*, 2025. URL https://openreview.net/forum?id=CqViN4dQJk.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum McDougall, Kola Ayonrinde, et al. Saebench: A comprehensive benchmark for sparse autoencoders in language model interpretability. *arXiv preprint arXiv:2503.09532*, 2025.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

David Klindt, Charles O'Neill, Patrik Reizinger, Harald Maurer, and Nina Miolane. A unifying framework from neural superposition to sparse interpretable codes. *Nature Machine Intelligence*, pp. 1–13, 2026.

Linghao Kong, Inimai Subramanian, Yonadav G Shavit, Micah Adler, Dan Alistarh, and Nir N Shavit. Expand neurons, not parameters. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=cYxavyv2C7.

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.

Victor Lecomte, Kushal Thaman, Rylan Schaeffer, Naomi Bashkansky, Trevor Chow, and Sanmi Koyejo. What causes polysemanticity? an alternative origin story of mixed selectivity from incidental causes. *arXiv preprint arXiv:2312.03096*, 2023.

Sewoong Lee, Adam Davies, Marc E. Canby, and Julia Hockenmaier. Evaluating and designing sparse autoencoders by approximating quasi-orthogonality. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=XhdNFeMclS.

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, March 2025. ISSN 1099-4300. doi: 10.3390/e27040344. URL http://dx.doi.org/10.3390/e27040344.

Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

Ekdeep Singh Lubana, Can Rager, Sai Sumedh R. Hindupur, Valérie Costa, Oam Patel, Sonia Krishna Murthy, Thomas Fel, Greta Tuckute, Daniel Wurgaft, Eric Bigelow, Demba E. Ba, Melanie Weber, and Aaron Mueller. Priors in time: Missing inductive biases for language model interpretability. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=4J2e3nWiC8.

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th annual international conference on machine learning*, pp. 689–696, 2009.

Julien Mairal, Francis Bach, and Jean Ponce. Sparse modeling for image and vision processing. *Foundations and Trends in Computer Graphics and Vision*, 8(2-3):85–283, 2014.

Alireza Makhzani and Brendan Frey. K-sparse autoencoders. *arXiv preprint arXiv:1312.5663*, 2013.

Charles C. Margossian and David M. Blei. Amortized variational inference: When and why? In Negar Kiyavash and Joris M. Mooij (eds.), *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pp. 2434–2449. PMLR, 15–19 Jul 2024. URL https://proceedings.mlr.press/v244/margossian24a.html.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff (eds.), *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 746–751, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/N13-1090/.

Moritz Miller, Florent Draye, and Bernhard Schölkopf. Superposition without interference? towards isolated interventions via almost orthogonal features in language models, 2026. URL https://arxiv.org/abs/2602.04718.

Akira A Nair, Jaehyun Joo, Jonghyun Lee, Lina Takemaru, Yidi Huang, Manu Shivakumar, Matthew Eric Lee, Jaesik Kim, Sokratis Apostolidis, and Dokyoon Kim. Interpreting genomic language models using sparse autoencoders. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=boBP35BB2U.

Neel Nanda. Open source replication & commentary on anthropic's dictionary learning paper. AI Alignment Forum, October 23th 2023. URL https://www.alignmentforum.org/posts/fKuugaxt2XLTkASkk/open-source-replication-and-commentary-on-anthropic-s.

Walter Nelson, Theofanis Karaletsos, and Francesco Locatello. Toward identifiable sparse autoencoders. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=miLK9YcxtA.

Thanh V Nguyen, Raymond KW Wong, and Chinmay Hegde. On the dynamics of gradient descent for autoencoders. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2858–2867. PMLR, 2019.

Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

Charles O'Neill, Alim Gumran, and David Klindt. Compute optimal inference and provable amortisation gap in sparse autoencoders. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 46877–46896. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/o-neill25a.html.

Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=DaNnkQJSQf.

Vardan Papyan, Jeremias Sulam, and Michael Elad. Working locally thinking globally: Theoretical guarantees for convolutional sparse coding. *IEEE Transactions on Signal Processing*, 65(21): 5687–5701, 2017.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=UGpGkLzwpP.

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://openreview.net/forum?id=EjInprGpk9.

Lucas Prieto, Edward Stevinson, Melih Barsbey, Tolga Birdal, and Pedro A. M. Mediano. From data statistics to feature geometry: How correlations shape superposition. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=7akSRQS5Xh.

Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János Kramár, Rohin Shah, and Neel Nanda. Improving dictionary learning with gated sparse autoencoders. *arXiv preprint arXiv:2404.16014*, 2024a.

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.

Akshay Rangamani, Anirbit Mukherjee, Amitabh Basu, Ashish Arora, Tejaswini Ganapathi, Sang Chin, and Trac D. Tran. Sparse coding and autoencoders. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 36–40, 2018. doi: 10.1109/ISIT.2018.8437533.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Adam Scherlis, Kshitij Sachan, Adam S. Jermyn, Joe Benton, and Buck Shlegeris. Polysemanticity and capacity in neural networks. *CoRR*, abs/2210.01892, 2022. URL https://doi.org/10.48550/arXiv.2210.01892.

Or David Shafran, Shaked Ronen, Omri Fahn, Shauli Ravfogel, Atticus Geiger, and Mor Geva. From directions to regions: Decomposing activations in language models via local geometry. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Tz7n7pX9cO.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeffrey Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Isaac Bloom, Stella Biderman, Adrià Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Mary Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, William Saunders, Eric J Michaud, Stephen Casper, Max Tegmark, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Thomas McGrath. Open problems in mechanistic interpretability. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=91H76m9Z94. Survey Certification.

Elana Simon and James Zou. Interplm: discovering interpretable features in protein language models via sparse autoencoders. *Nature methods*, 22(10):2107–2117, 2025.

Xiangchen Song, Aashiq Muhamed, Yujia Zheng, Lingjing Kong, Zeyu Tang, Mona T. Diab, Virginia Smith, and Kun Zhang. Mechanistic interpretability should prioritize feature consistency in sparse autoencoders. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2172–2210, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.99. URL https://aclanthology.org/2026.acl-long.99/.

Edward Stevinson, Lucas Prieto, Melih Barsbey, and Tolga Birdal. Adversarial vulnerability from interference between features in superposition. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=fD6c4bD7sR.

Yiming Tang, Harshvardhan Saini, Zhaoqian Yao, Yizhen Liao, Qianxiao Li, Mengnan Du, and Dianbo Liu. On the theoretical foundation of sparse dictionary learning in mechanistic interpretability. *arXiv preprint arXiv:2512.05534*, 2025.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Demian Till. Do sparse autoencoders find "true features", 2024. URL https://www.lesswrong.com/posts/QoR8noAB3Mp2KBA4B/do-sparse-autoencoders-find-true-features.

Bahareh Tolooshams and Demba E. Ba. Stable and interpretable unrolled dictionary learning. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=e3S0Bl2RO8.

Bahareh Tolooshams, Ailsa Shen, and Anima Anandkumar. Mechanistic interpretability with sparse autoencoder neural operators. *arXiv preprint arXiv:2509.03738*, 2025.

Ivana Tošić and Pascal Frossard. Dictionary learning. *IEEE Signal Processing Magazine*, 28(2): 27–38, 2011.

Joel A Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information theory*, 50(10):2231–2242, 2004.

Ruihan Zhang, Prashan Madumal, Tim Miller, Krista A Ehinger, and Benjamin IP Rubinstein. Invertible concept-based explanations for cnn models with non-negative concept activation vectors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11682–11690, 2021.

Xudong Zhu, Mohammad Mahdi Khalili, and Zhihui Zhu. Abstopk: Rethinking sparse autoencoders for bidirectional features. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=EEs6I4cO7S.

A Appendix - AI Use Statement

Authors used generative AI tools to assist with polishing and improving the writing of the manuscript, including clarity and readability.

We used Claude Code to adapt our existing SAE implementation, training procedure, and evaluation metrics, originally developed and validated on MNIST, to run at scale on cached Pythia activations. The underlying architectures, training objective, and evaluation quantities were fixed beforehand.

We also used GPT-5.6 to obtain feedback on the manuscript, including suggestions on the organization of the results and overall write-up. We did not use generative AI tools for research formulation, experimental or methodological design, derivations, or theoretical development. We reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

B Appendix - Reproducibility Statement

Appendix I provides the complete experimental setup, including the model and data, training budget, architectures, initialization, optimization, and constraints. It also reports holdout reconstruction, sparsity, and dead-feature statistics for all runs (Table 2). Complete proofs of all formal results are provided in the appendix.

The SAEBench experiments (Section J.3.2) use publicly released checkpoints and SAEBench's own evaluation pipeline. We identify the checkpoints by model, layer, and sparsity target, allowing these results to be reproduced directly from public artifacts.

C Appendix - Ethics Statement

Interpretability is an important component of developing and auditing reliable AI systems, as it can help researchers understand model representations, identify failure modes, and evaluate unintended behaviour. Sparse autoencoders are used for this purpose by extracting human-interpretable features from neural network activations. Our work highlights a consideration for this approach: the extent to which SAE features are separable depends on architectural choices that may otherwise appear incidental. We therefore view our results as supporting more cautious interpretation of individual SAE features and interventions based on them.

Our experiments use publicly released pretrained models, datasets, and SAE checkpoints. No new data were collected and no human subjects were involved. The language corpora are web-derived and may contain personal or biased content, but our analysis is restricted to aggregate geometric statistics and does not inspect or attempt to recover individual examples or sensitive attributes. As with interpretability research more broadly, the methods studied here could in principle contribute to model analysis or reverse engineering; however, our work does not introduce new capabilities for eliciting model behaviour or recovering training data, and we do not identify additional direct risks beyond those associated with existing interpretability methods.

D Appendix - Effective Interference

The effective interference has the following properties.

Proposition D.1 (Properties of effective interference)**.** *For distinct features $i \neq j$:*

1. ***Support awareness.*** *If $\Pr(z_i \neq 0, z_j \neq 0) = 0$, then $I_{ij} = 0$, irrespective of $\boldsymbol{d}_i^\top \boldsymbol{d}_j$.*
2. ***Reparameterization invariance.*** *For positive scalars β_i, the transformations $\boldsymbol{d}_i \mapsto \beta_i \boldsymbol{d}_i$ and $z_i \mapsto z_i/\beta_i$ leave every component $\boldsymbol{d}_i z_i$ and hence every I_{ij} unchanged.*
3. ***Sign under nonnegative coding.*** *If $z_i, z_j \geq 0$ and the pair co-activates with nonzero probability, then I_{ij} has the sign of $\boldsymbol{d}_i^\top \boldsymbol{d}_j$.*

D.1 Relation to geometric measures

Effective interference recovers geometric measures only under restrictive code statistics. Suppose that $\mathbb{E}[z_i z_j] = s$ for every $i \neq j$. Then

$$I_{ij} = s(\boldsymbol{d}_i^\top \boldsymbol{d}_j), \tag{D.1}$$

so the off-diagonal effective-interference matrix is proportional to the off-diagonal Gram matrix. Purely geometric interference therefore implicitly assumes that the code second moments contain no pair-specific information. In the binary, unit-normalized case, $M_{ij} = 1$, and effective interference instead reduces to cosine similarity weighted by the probability of co-activation. This makes explicit the role of latent statistics in models of superposition (Elhage et al., 2022; Prieto et al., 2026).

Mutual coherence plays a different role. For normalized directions, let

$$\mu := \max_{k \neq \ell} |\boldsymbol{d}_k^\top \boldsymbol{d}_\ell|. \tag{D.2}$$

Under nonnegative coding,

$$|I_{ij}| \leq \mu \, \pi_{ij} M_{ij}. \tag{D.3}$$

Coherence therefore gives a support-agnostic worst-case geometric bound, rather than a measure of the interaction realized by a particular representation (Donoho & Huo, 2001; Tropp, 2004). The bound is loose when co-activation concentrates on pairs whose geometric overlap is far below the global maximum, mirroring the distinction between worst-case geometric recovery conditions and data-dependent behaviour in sparse coding and dictionary learning (Mairal et al., 2009; Tošić & Frossard, 2011).

D.2 Nonnegative Codes

Our formulation and results consider nonnegative codes. For signed codes, one can replace the nonnegative profile by

$$M_{ij}^{\text{signed}} := \mathbb{E}[z_i z_j \mid a_i = a_j = 1], \qquad A_{ij} := \mathbb{E}[|z_i z_j| \mid a_i = a_j = 1], \tag{D.4}$$

where now A_{ij} is the conditional magnitude. When $A_{ij} > 0$, we define the signed code alignment as

$$\kappa_{ij} := \frac{M_{ij}^{\text{s}}}{A_{ij}} \in [-1, 1], \qquad I_{ij} = \rho_{ij} \pi_{ij} \kappa_{ij} A_{ij}. \tag{D.5}$$

Under nonnegative coding, $\kappa_{ij} = 1$ and the original profile is recovered. We do not explore this setting in the paper and leave for future work.

D.3 Optimality and consistency

Proposition 4.1 (Code-optimality condition)**.** *For a fixed decoder $\boldsymbol{D}$ and support S, any interior least-squares optimum satisfies*

$$\boldsymbol{D}_S^\top \boldsymbol{r} = 0, \qquad \boldsymbol{r} := \boldsymbol{x} - \boldsymbol{D}_S \boldsymbol{z}_S. \tag{4.3}$$

For nonnegative codes, interior means $z_i > 0 \; \forall i \in S$.

Proof. For fixed support S, the least-squares objective is

$$\mathcal{L}(\boldsymbol{z}_S) = \frac{1}{2} \|\boldsymbol{x} - \boldsymbol{D}_S \boldsymbol{z}_S\|_2^2 . \tag{D.6}$$

At an interior optimum, the first-order optimality condition requires

$$\nabla_{\boldsymbol{z}_S} \mathcal{L} = -\boldsymbol{D}_S^\top (\boldsymbol{x} - \boldsymbol{D}_S \boldsymbol{z}_S) = -\boldsymbol{D}_S^\top \boldsymbol{r} = 0. \tag{D.7}$$

Hence,

$$\boldsymbol{D}_S^\top \boldsymbol{r} = 0. \tag{D.8}$$

For nonnegative codes, interiority ensures that no inequality constraint is active, i.e., $z_i > 0$ for all $i \in S$, so the unconstrained first-order condition applies[1]. ■

Proposition 4.2 (Encoder consistency identity)**.** *Suppose the encoder is locally affine as in (4.2). For every active feature $i \in S$,*

$$0 = (\alpha_i \gamma_{ii} - 1) z_i + \sum_{j \neq i} \alpha_i \gamma_{ij} z_j + b_i + \boldsymbol{w}_i^\top \boldsymbol{r}. \tag{4.4}$$

Consequently, for every i with $\Pr(a_i = 1) > 0$,

$$0 = \underbrace{(\alpha_i \gamma_{ii} - 1)\mathbb{E}[z_i \mid a_i = 1]}_{\textit{self-response mismatch}} + \underbrace{\sum_{j \neq i} \alpha_i \gamma_{ij} \mathbb{E}[z_j \mid a_i = 1]}_{\textit{incoming cross-talk}} + \underbrace{b_i}_{\substack{\textit{bias} \\ \textit{compensation}}} + \underbrace{\mathbb{E}[\boldsymbol{w}_i^\top \boldsymbol{r} \mid a_i = 1]}_{\textit{residual coupling}}, \tag{4.5}$$

Proof. Substitute $\boldsymbol{x} = \boldsymbol{D}\boldsymbol{z} + \boldsymbol{r}$ into $z_i = \boldsymbol{w}_i^\top \boldsymbol{x} + b_i$ on the active region, separate the $j = i$ term, and take the conditional expectation. ■

D.4 Encoder-decoder duality

Consider an untied encoder. The decoder Gram $\boldsymbol{D}^\top \boldsymbol{D}$ continues to determine decoder geometry, whereas the cross-Gram $\boldsymbol{W}^\top \boldsymbol{D}$ determines inference cross-talk. These two matrices need not have the same off-diagonal structure.

When $\boldsymbol{D}_S$ has full column rank, the least-squares code and its support-specific linear encoder are

$$\boldsymbol{z}_S^* = (\boldsymbol{D}_S^\top \boldsymbol{D}_S)^{-1} \boldsymbol{D}_S^\top \boldsymbol{x}, \qquad \boldsymbol{W}_S^* = \boldsymbol{D}_S (\boldsymbol{D}_S^\top \boldsymbol{D}_S)^{-1}, \qquad (\boldsymbol{W}_S^*)^\top \boldsymbol{D}_S = \boldsymbol{I}, \tag{D.9}$$

where $\boldsymbol{W}_S^*$ is the ideal encoder conditioned on current active support.

D.5 Average co-activation probability pairs under uniformly distributed support

Here we derive the average co-activation probability pairs when the support is uniformly chosen across data examples. For a k-sparse autoencoder. Assume support is uniformly distributed across all features. For activation indicator a_i, a k-sparse code satisfy

$$\sum_{i \neq j} a_i a_j = \left(\sum_i a_i \right)^2 - \sum_i a_i = k(k-1), \tag{D.10}$$

resulting in sum co-activation pair probability of

$$\sum \pi_{ij} = \mathbb{E}[a_i a_j] = k(k-1). \tag{D.11}$$

The average co-activation probability pairs from $\tilde{p}$ total used features (i.e., $\tilde{p} = p -$ `# dead neurons`) would be

$$\overline{\pi_{ij}} = \frac{1}{\tilde{p}(\tilde{p}-1)} \sum \pi_{ij} = \frac{k(k-1)}{\tilde{p}(\tilde{p}-1)} \tag{D.12}$$

where $\tilde{p}(\tilde{p}-1)$ are number of ordered pairs from $\tilde{p}$ features.

[1]The affine identity assumes unit slope on active coordinates. ReLU- and threshold-based encoders are affine on their active coordinates, while TopK-like encoders are affine within regions where the selected indices do not change (Huben et al., 2024; Bricken et al., 2023; Gao et al., 2025; Rajamanoharan et al., 2024b). For BatchTopK, this holds conditional on the batch selection mask, hence, not point-wise sparsifier.

E APPENDIX - PROOF OF PROPOSITION 4.3

Proposition 4.3 (Local orthogonalization for two active features)**.** *Consider the tied, normalized, bias-free model with fixed support $S = \{i, j\}$ and loss $\frac{1}{2}\|\widehat{\boldsymbol{x}} - \boldsymbol{x}\|_2^2$. Let $\boldsymbol{D}_S^+$ denote the active dictionary after one gradient step of size η, followed by column normalization. Then*

$$\Delta(\boldsymbol{D}_S^+) - \Delta(\boldsymbol{D}_S) = -8\eta\rho_{ij}^2\|z_i\boldsymbol{d}_i - z_j\boldsymbol{d}_j\|_2^2 + O(\eta^2). \tag{4.7}$$

Hence, if $\rho_{ij} \neq 0$ and $z_i\boldsymbol{d}_i \neq z_j\boldsymbol{d}_j$, a gradient step cannot increase $\Delta(\boldsymbol{D}_S)$ to first order.

Proof. Since $S = \{i, j\}$ and the concept directions are unit norm, the quantity $\Delta(\boldsymbol{D}_S)$ reduces to

$$\Delta(\boldsymbol{D}_S) = 2(\boldsymbol{d}_i^\top \boldsymbol{d}_j)^2 = 2\rho_{ij}^2. \tag{E.1}$$

It is therefore sufficient to characterize the first-order change in ρ_{ij} induced by one gradient step followed by column normalization.

In the tied, normalized and bias-free setting ($b_i = 0$), the active codes satisfy

$$z_i = \boldsymbol{d}_i^\top \boldsymbol{x}, \qquad z_j = \boldsymbol{d}_j^\top \boldsymbol{x}, \tag{E.2}$$

and the reconstruction is

$$\hat{\boldsymbol{x}} = \boldsymbol{d}_i z_i + \boldsymbol{d}_j z_j. \tag{E.3}$$

We first derive the normalized update of each active feature, then use it to obtain the corresponding dynamics of ρ_{ij} and, consequently, of $\Delta(\boldsymbol{D}_S)$.

Normalized update. Let $\boldsymbol{g}_k := \nabla_{\boldsymbol{d}_k}\mathcal{L}$ denote the gradient with respect to an active feature $\boldsymbol{d}_k$. A gradient step of size η gives $\boldsymbol{d}_k - \eta\boldsymbol{g}_k$, which is subsequently projected back onto the unit sphere by column normalization. Expanding this normalized update to first order in η yields

$$\boldsymbol{d}_k^+ = \boldsymbol{d}_k - \eta(\boldsymbol{I} - \boldsymbol{d}_k\boldsymbol{d}_k^\top)\boldsymbol{g}_k + O(\eta^2). \tag{E.4}$$

Thus, normalization removes the radial component of the gradient, and only its component tangent to the unit sphere contributes to the first-order evolution of the feature direction. We refer to Appendix G for the derivation of this expansion.

Projected gradients. The next step is to replace the gradients in the normalized update by their expressions for the tied, bias-free model. As derived in Appendix H, for each active feature $\boldsymbol{d}_k$, we have

$$\boldsymbol{g}_k = (\hat{\boldsymbol{x}} - \boldsymbol{x})z_k + \alpha_k\boldsymbol{x}\boldsymbol{d}_k^\top(\hat{\boldsymbol{x}} - \boldsymbol{x}) \tag{E.5}$$

Using Equation (E.2), Equation (E.3) and the fact that $\alpha_k = 1$, we get

$$\begin{aligned}
\boldsymbol{g}_i &= (\boldsymbol{d}_i z_i + \boldsymbol{d}_j z_j - \boldsymbol{x})z_i + \boldsymbol{x}\,\boldsymbol{d}_i^\top(\boldsymbol{d}_i z_i + \boldsymbol{d}_j z_j - \boldsymbol{x}) \\
&= \boldsymbol{d}_i z_i^2 + \boldsymbol{d}_j z_i z_j - \boldsymbol{x}z_i + \boldsymbol{x}\,(z_i + \rho_{ij}z_j - z_i) \\
&= \boldsymbol{d}_i z_i^2 + \boldsymbol{d}_j z_i z_j + \boldsymbol{x}\,(\rho_{ij}z_j - z_i)\,,
\end{aligned} \tag{E.6}$$

where we used $\|\boldsymbol{d}_i\|_2 = 1$. Projecting the gradient onto the tangent space of the unit sphere and using

$$(\boldsymbol{I} - \boldsymbol{d}_i\boldsymbol{d}_i^\top)\boldsymbol{d}_i = 0, \qquad (\boldsymbol{I} - \boldsymbol{d}_i\boldsymbol{d}_i^\top)\boldsymbol{d}_j = \boldsymbol{d}_j - \rho_{ij}\boldsymbol{d}_i, \tag{E.7}$$

we obtain

$$\begin{aligned}
(\boldsymbol{I} - \boldsymbol{d}_i\boldsymbol{d}_i^\top)\boldsymbol{g}_i &= (\boldsymbol{d}_j - \rho_{ij}\boldsymbol{d}_i)z_i z_j + \big(\boldsymbol{x} - \boldsymbol{d}_i(\boldsymbol{d}_i^\top\boldsymbol{x})\big)\,(\rho_{ij}z_j - z_i) && \text{(E.8)} \\
&= (\boldsymbol{d}_j - \rho_{ij}\boldsymbol{d}_i)z_i z_j + (\boldsymbol{x} - \boldsymbol{d}_i z_i)\,(\rho_{ij}z_j - z_i). && \text{(E.9)}
\end{aligned}$$

By symmetry,

$$(\boldsymbol{I} - \boldsymbol{d}_j\boldsymbol{d}_j^\top)\boldsymbol{g}_j = (\boldsymbol{d}_i - \rho_{ij}\boldsymbol{d}_j)z_i z_j + (\boldsymbol{x} - z_j\boldsymbol{d}_j)(\rho_{ij}z_i - z_j). \tag{E.10}$$

Correlation dynamics. We now use the projected gradients to characterize how the correlation between the two active directions evolves after one normalized gradient step. Recall that $\rho_{ij} = \boldsymbol{d}_i^\top \boldsymbol{d}_j$. Therefore, using the first-order normalized update in Equation (E.4), the correlation after the update is

$$(\boldsymbol{d}_i^+)^\top \boldsymbol{d}_j^+ = \boldsymbol{d}_i^\top \boldsymbol{d}_j - \eta \boldsymbol{d}_i^\top (\boldsymbol{I} - \boldsymbol{d}_j \boldsymbol{d}_j^\top)\boldsymbol{g}_j - \eta \boldsymbol{d}_j^\top (\boldsymbol{I} - \boldsymbol{d}_i \boldsymbol{d}_i^\top)\boldsymbol{g}_i + O(\eta^2). \tag{E.11}$$

Substituting Equation (E.9) and Equation (E.10) yields the two symmetric contributions

$$\boldsymbol{d}_i^\top (\boldsymbol{I} - \boldsymbol{d}_j \boldsymbol{d}_j^\top)\boldsymbol{g}_j = (1 - \rho_{ij}^2) z_i z_j + (z_i - z_j \rho_{ij})(\rho_{ij} z_i - z_j), \tag{E.12}$$

$$\boldsymbol{d}_j^\top (\boldsymbol{I} - \boldsymbol{d}_i \boldsymbol{d}_i^\top)\boldsymbol{g}_i = (1 - \rho_{ij}^2) z_i z_j + (z_j - z_i \rho_{ij})(\rho_{ij} z_j - z_i). \tag{E.13}$$

Noticing that these two contributions are equal, summing them yields

$$(\boldsymbol{d}_i^+)^\top \boldsymbol{d}_j^+ = \rho_{ij} - 2\eta \left((1 - \rho_{ij}^2) z_i z_j + (z_i - z_j \rho_{ij})(\rho_{ij} z_i - z_j) \right) + O(\eta^2) \tag{E.14}$$

$$= \rho_{ij} - 2\eta \left(z_i z_j - \rho_{ij}^2 z_i z_j + \rho_{ij} z_i^2 - z_i z_j - \rho_{ij}^2 z_i z_j + \rho_{ij} z_j^2 \right) + O(\eta^2) \tag{E.15}$$

$$= \rho_{ij} - 2\eta \left(\rho_{ij} z_i^2 + \rho_{ij} z_j^2 - 2\rho_{ij}^2 z_i z_j \right) + O(\eta^2) \tag{E.16}$$

$$= \rho_{ij} - 2\eta \rho_{ij} \left(z_i^2 + z_j^2 - 2\rho_{ij} z_i z_j \right) + O(\eta^2) \tag{E.17}$$

$$= \rho_{ij} - 2\eta \rho_{ij} \left(z_i^2 \|\boldsymbol{d}_i\|_2^2 + z_j^2 \|\boldsymbol{d}_j\|_2^2 - 2 z_i z_j \boldsymbol{d}_i^\top \boldsymbol{d}_j \right) + O(\eta^2) \tag{E.18}$$

$$= \rho_{ij} - 2\eta \rho_{ij} \left\| z_i \boldsymbol{d}_i - z_j \boldsymbol{d}_j \right\|_2^2 + O(\eta^2). \tag{E.19}$$

Finally, substituting the correlation dynamics into $\Delta(\boldsymbol{D}_S) = 2\rho_{ij}^2$ gives

$$\Delta(\boldsymbol{D}_S^+) - \Delta(\boldsymbol{D}_S) = 2\big((\boldsymbol{d}_i^+)^\top \boldsymbol{d}_j^+\big)^2 - 2\rho_{ij}^2 \tag{E.20}$$

$$= 2 \left(\rho_{ij} - 2\eta \rho_{ij} \left\| z_i \boldsymbol{d}_i - z_j \boldsymbol{d}_j \right\|_2^2 + O(\eta^2) \right)^2 - 2\rho_{ij}^2 \tag{E.21}$$

$$= -8\eta \rho_{ij}^2 \left\| z_i \boldsymbol{d}_i - z_j \boldsymbol{d}_j \right\|_2^2 + O(\eta^2). \tag{E.22}$$

The first-order term is of order $\eta \rho_{ij}^2$, while the remainder is of order η^2. Writing the remainder as $R(\eta) = O(\eta^2)$, there exist constants $C > 0$ and $\eta_0 > 0$ such that

$$|R(\eta)| \le C\eta^2, \qquad 0 < \eta < \eta_0. \tag{E.23}$$

Hence,

$$\Delta(\boldsymbol{D}_S^+) - \Delta(\boldsymbol{D}_S) \le -8\eta \rho_{ij}^2 \left\| z_i \boldsymbol{d}_i - z_j \boldsymbol{d}_j \right\|_2^2 + C\eta^2. \tag{E.24}$$

Therefore, the difference is non-positive whenever

$$C\eta^2 \le 8\eta \rho_{ij}^2 \left\| z_i \boldsymbol{d}_i - z_j \boldsymbol{d}_j \right\|_2^2, \tag{E.25}$$

or equivalently,

$$\eta \le \frac{8}{C} \rho_{ij}^2 \left\| z_i \boldsymbol{d}_i - z_j \boldsymbol{d}_j \right\|_2^2. \tag{E.26}$$

Provided $\|z_i \boldsymbol{d}_i - z_j \boldsymbol{d}_j\|_2^2 > 0$, this quantity is a strictly positive constant for fixed $z_i, z_j, \boldsymbol{d}_i, \boldsymbol{d}_j$. Thus, under the assumption $\eta \ll \rho_{ij}^2$ and for sufficiently small $\eta < \eta_0$, the above bound is satisfied. Consequently, the negative first-order term dominates the $O(\eta^2)$ remainder, and

$$\Delta(\boldsymbol{D}_S^+) - \Delta(\boldsymbol{D}_S) \le 0. \tag{E.27}$$

This proves the result. ■

F APPENDIX - PROPOSITION F.1

Proposition F.1 (Local orthogonalization without normalization)**.** *Consider the tied, bias-free setting with active support $S = \{i, j\}$ and $z_i z_j > 0$. Let*

$$n_k = \|\boldsymbol{d}_k\|_2, \qquad \tilde{\boldsymbol{d}}_k = \frac{\boldsymbol{d}_k}{n_k}, \qquad \tilde{z}_k = \tilde{\boldsymbol{d}}_k^\top \boldsymbol{x} = \frac{z_k}{n_k}, \qquad \rho_{ij} = \tilde{\boldsymbol{d}}_i^\top \tilde{\boldsymbol{d}}_j,$$

and let $\tilde{\boldsymbol{D}}$ denote the column-normalized version of $\boldsymbol{D}$. After one gradient step of size η applied to the unnormalized dictionary $\boldsymbol{D}$,

$$\Delta(\tilde{\boldsymbol{D}}_S^+) - \Delta(\tilde{\boldsymbol{D}}_S) = -8\eta\rho_{ij}\left[\rho_{ij}(\tilde{z}_i^2 + \tilde{z}_j^2) + \tilde{z}_i\tilde{z}_j\left((n_i^2 + n_j^2)(1 - \rho_{ij}^2) - 2\right)\right] + O(\eta^2), \quad \text{(F.1)}$$

where $\boldsymbol{D}^+$ is the updated, unnormalized dictionary and $\tilde{\boldsymbol{D}}^+$ its column-normalized version.

Consequently, for sufficiently small η, $\Delta(\tilde{\boldsymbol{D}}_S)$ is guaranteed to be non-increasing to first order whenever

$$\begin{cases} n_i^2 + n_j^2 > \dfrac{2}{1+\rho_{ij}}, & 0 < \rho_{ij} < 1, \\ n_i^2 + n_j^2 < \dfrac{2}{1+\rho_{ij}}, & -1 < \rho_{ij} < 0. \end{cases} \quad \text{(F.2)}$$

At $\rho_{ij} = 0$, the first-order change vanishes; at $\rho_{ij} = 1$, the two directions remain positively collinear. The case $\rho_{ij} = -1$ is incompatible with $z_i z_j > 0$ for nonzero tied codes.

Importantly, the gradient update itself is performed on the unnormalized dictionary $\boldsymbol{D}$. We evaluate the resulting change through its column-normalized version $\tilde{\boldsymbol{D}}$ because we are interested in the evolution of the angle between feature directions, independently of changes in their column norms.

Proof. Since we are interested in the angular dynamics, we track the normalized feature directions. After one gradient step on the unnormalized dictionary,

$$\tilde{\boldsymbol{d}}_i^+ = \frac{\boldsymbol{d}_i^+}{\|\boldsymbol{d}_i^+\|_2} = \frac{\boldsymbol{d}_i - \eta\nabla_{\boldsymbol{d}_i}\mathcal{L}}{\|\boldsymbol{d}_i - \eta\nabla_{\boldsymbol{d}_i}\mathcal{L}\|_2}. \quad \text{(F.3)}$$

Dividing the numerator and denominator by n_i and using Equation (G.6) gives

$$\tilde{\boldsymbol{d}}_i^+ = \frac{\tilde{\boldsymbol{d}}_i - \frac{\eta}{n_i}\nabla_{\boldsymbol{d}_i}\mathcal{L}}{\|\tilde{\boldsymbol{d}}_i - \frac{\eta}{n_i}\nabla_{\boldsymbol{d}_i}\mathcal{L}\|_2} \quad \text{(F.4)}$$

$$= \left(\tilde{\boldsymbol{d}}_i - \frac{\eta}{n_i}\nabla_{\boldsymbol{d}_i}\mathcal{L}\right)\left(1 + \frac{\eta}{n_i}\tilde{\boldsymbol{d}}_i^\top\nabla_{\boldsymbol{d}_i}\mathcal{L} + O(\eta^2)\right) \quad \text{(F.5)}$$

$$= \tilde{\boldsymbol{d}}_i - \frac{\eta}{n_i}\left(\boldsymbol{I} - \tilde{\boldsymbol{d}}_i\tilde{\boldsymbol{d}}_i^\top\right)\nabla_{\boldsymbol{d}_i}\mathcal{L} + O(\eta^2). \quad \text{(F.6)}$$

Thus, to first order, only the component of the gradient orthogonal to $\tilde{\boldsymbol{d}}_i$ changes the feature direction, while the radial component changes only its norm.

Substituting the gradient from Equation (H.10) yields

$$\nabla_{\boldsymbol{d}_i}\mathcal{L} = (\hat{\boldsymbol{x}} - \boldsymbol{x})z_i + \boldsymbol{x}\boldsymbol{d}_i^\top(\hat{\boldsymbol{x}} - \boldsymbol{x}) \quad \text{(F.7)}$$

$$= (z_i\boldsymbol{d}_i + z_j\boldsymbol{d}_j - \boldsymbol{x})\,z_i + \boldsymbol{x}\boldsymbol{d}_i^\top(z_i\boldsymbol{d}_i + z_j\boldsymbol{d}_j - \boldsymbol{x}) \quad \text{(F.8)}$$

$$= z_i^2\boldsymbol{d}_i + z_i z_j\boldsymbol{d}_j - z_i\boldsymbol{x} + \boldsymbol{x}\left(z_i n_i^2 + z_j n_i n_j\rho_{ij} - z_i\right) \quad \text{(F.9)}$$

$$= z_i^2\boldsymbol{d}_i + z_i z_j\boldsymbol{d}_j + \left[(n_i^2 - 2)z_i + n_i n_j\rho_{ij}z_j\right]\boldsymbol{x}. \quad \text{(F.10)}$$

We now project each term separately. First, since $\boldsymbol{d}_i = n_i\tilde{\boldsymbol{d}}_i$,

$$\left(\boldsymbol{I} - \tilde{\boldsymbol{d}}_i\tilde{\boldsymbol{d}}_i^\top\right)z_i^2\boldsymbol{d}_i = z_i^2\left(\boldsymbol{I} - \tilde{\boldsymbol{d}}_i\tilde{\boldsymbol{d}}_i^\top\right)n_i\tilde{\boldsymbol{d}}_i = 0. \quad \text{(F.11)}$$

Second,

$$\left(\boldsymbol{I}-\tilde{\boldsymbol{d}}_i\tilde{\boldsymbol{d}}_i^\top\right) z_i z_j \boldsymbol{d}_j = z_i z_j \left(\boldsymbol{I}-\tilde{\boldsymbol{d}}_i\tilde{\boldsymbol{d}}_i^\top\right) n_j \tilde{\boldsymbol{d}}_j \tag{F.12}$$

$$= n_j z_i z_j \left(\tilde{\boldsymbol{d}}_j - \tilde{\boldsymbol{d}}_i\tilde{\boldsymbol{d}}_i^\top \tilde{\boldsymbol{d}}_j\right) = n_j z_i z_j \left(\tilde{\boldsymbol{d}}_j - \rho_{ij}\tilde{\boldsymbol{d}}_i\right). \tag{F.13}$$

Finally,

$$\left(\boldsymbol{I}-\tilde{\boldsymbol{d}}_i\tilde{\boldsymbol{d}}_i^\top\right)\boldsymbol{x}\left[(n_i^2-2)z_i + n_i n_j \rho_{ij} z_j\right] = \left[\boldsymbol{x} - \tilde{\boldsymbol{d}}_i\left(\tilde{\boldsymbol{d}}_i^\top \boldsymbol{x}\right)\right]\left[(n_i^2-2)z_i + n_i n_j \rho_{ij} z_j\right] \tag{F.14}$$

$$= \left(\boldsymbol{x} - \tilde{z}_i\tilde{\boldsymbol{d}}_i\right)\left[(n_i^2-2)z_i + n_i n_j \rho_{ij} z_j\right]. \tag{F.15}$$

Therefore,

$$\left(\boldsymbol{I}-\tilde{\boldsymbol{d}}_i\tilde{\boldsymbol{d}}_i^\top\right)\nabla_{\boldsymbol{d}_i}\mathcal{L} = n_j z_i z_j \left(\tilde{\boldsymbol{d}}_j - \rho_{ij}\tilde{\boldsymbol{d}}_i\right) + \left[(n_i^2-2)z_i + n_i n_j \rho_{ij} z_j\right]\left(\boldsymbol{x} - \tilde{z}_i\tilde{\boldsymbol{d}}_i\right). \tag{F.16}$$

Substituting this expression into the first-order update of the normalized direction gives

$$\tilde{\boldsymbol{d}}_i^+ = \tilde{\boldsymbol{d}}_i - \frac{\eta}{n_i}\left[n_j z_i z_j \left(\tilde{\boldsymbol{d}}_j - \rho_{ij}\tilde{\boldsymbol{d}}_i\right) + \left[(n_i^2-2)z_i + n_i n_j \rho_{ij} z_j\right]\left(\boldsymbol{x} - \tilde{z}_i\tilde{\boldsymbol{d}}_i\right)\right] + O(\eta^2). \tag{F.17}$$

By symmetry, the corresponding update for atom j is

$$\tilde{\boldsymbol{d}}_j^+ = \tilde{\boldsymbol{d}}_j - \frac{\eta}{n_j}\left[n_i z_i z_j \left(\tilde{\boldsymbol{d}}_i - \rho_{ij}\tilde{\boldsymbol{d}}_j\right) + \left[(n_j^2-2)z_j + n_i n_j \rho_{ij} z_i\right]\left(\boldsymbol{x} - \tilde{z}_j\tilde{\boldsymbol{d}}_j\right)\right] + O(\eta^2). \tag{F.18}$$

Given the required inner products,

$$\tilde{\boldsymbol{d}}_j^\top\left(\tilde{\boldsymbol{d}}_j - \rho_{ij}\tilde{\boldsymbol{d}}_i\right) = 1-\rho_{ij}^2, \qquad \tilde{\boldsymbol{d}}_j^\top\left(\boldsymbol{x} - \tilde{z}_i\tilde{\boldsymbol{d}}_i\right) = \tilde{z}_j - \rho_{ij}\tilde{z}_i, \tag{F.19}$$

$$\tilde{\boldsymbol{d}}_i^\top\left(\tilde{\boldsymbol{d}}_i - \rho_{ij}\tilde{\boldsymbol{d}}_j\right) = 1-\rho_{ij}^2, \qquad \tilde{\boldsymbol{d}}_i^\top\left(\boldsymbol{x} - \tilde{z}_j\tilde{\boldsymbol{d}}_j\right) = \tilde{z}_i - \rho_{ij}\tilde{z}_j. \tag{F.20}$$

We can now compute the first-order evolution of the normalized correlation $\rho_{ij} = \tilde{\boldsymbol{d}}_i^\top\tilde{\boldsymbol{d}}_j$. Using

$$\rho_{ij}^+ = (\tilde{\boldsymbol{d}}_i^+)^\top\tilde{\boldsymbol{d}}_j^+, \tag{F.21}$$

and keeping only first-order terms in η, we get

$$\begin{aligned}
\rho_{ij}^+ &= \rho_{ij} - \frac{\eta}{n_i}\left[n_j z_i z_j \tilde{\boldsymbol{d}}_j^\top\left(\tilde{\boldsymbol{d}}_j - \rho_{ij}\tilde{\boldsymbol{d}}_i\right) + \left[(n_i^2-2)z_i + n_i n_j \rho_{ij} z_j\right]\tilde{\boldsymbol{d}}_j^\top\left(\boldsymbol{x} - \tilde{z}_i\tilde{\boldsymbol{d}}_i\right)\right] \\
&\quad - \frac{\eta}{n_j}\left[n_i z_i z_j \tilde{\boldsymbol{d}}_i^\top\left(\tilde{\boldsymbol{d}}_i - \rho_{ij}\tilde{\boldsymbol{d}}_j\right) + \left[(n_j^2-2)z_j + n_i n_j \rho_{ij} z_i\right]\tilde{\boldsymbol{d}}_i^\top\left(\boldsymbol{x} - \tilde{z}_j\tilde{\boldsymbol{d}}_j\right)\right] + O(\eta^2) \\
&= \rho_{ij} - \eta z_i z_j\left(\frac{n_j}{n_i} + \frac{n_i}{n_j}\right)(1-\rho_{ij}^2) \\
&\quad - \frac{\eta}{n_i}\left[(n_i^2-2)z_i + n_i n_j \rho_{ij} z_j\right](\tilde{z}_j - \rho_{ij}\tilde{z}_i) \\
&\quad - \frac{\eta}{n_j}\left[(n_j^2-2)z_j + n_i n_j \rho_{ij} z_i\right](\tilde{z}_i - \rho_{ij}\tilde{z}_j) + O(\eta^2). \\
&= \rho_{ij} - \eta\tilde{z}_i\tilde{z}_j\left(n_i^2+n_j^2\right)\left(1-\rho_{ij}^2\right) \\
&\quad - \eta\left[(n_i^2-2)\tilde{z}_i + n_j^2\rho_{ij}\tilde{z}_j\right](\tilde{z}_j - \rho_{ij}\tilde{z}_i) \\
&\quad - \eta\left[(n_j^2-2)\tilde{z}_j + n_i^2\rho_{ij}\tilde{z}_i\right](\tilde{z}_i - \rho_{ij}\tilde{z}_j) + O(\eta^2) \\
&= \rho_{ij} - \eta\tilde{z}_i\tilde{z}_j(n_i^2+n_j^2)(1-\rho_{ij}^2) \\
&\quad - \eta\Big[(n_i^2-2)\tilde{z}_i\tilde{z}_j - (n_i^2-2)\rho_{ij}\tilde{z}_i^2 + n_j^2\rho_{ij}\tilde{z}_j^2 - n_j^2\rho_{ij}^2\tilde{z}_i\tilde{z}_j\Big]
\end{aligned} \tag{F.22}$$

$$- \eta\Big[(n_j^2-2)\tilde{z}_i\tilde{z}_j - (n_j^2-2)\rho_{ij}\tilde{z}_j^2 + n_i^2\rho_{ij}\tilde{z}_i^2 - n_i^2\rho_{ij}^2\tilde{z}_i\tilde{z}_j\Big] + O(\eta^2) \tag{F.23}$$

$$= \rho_{ij} - \eta\tilde{z}_i\tilde{z}_j(n_i^2+n_j^2)(1-\rho_{ij}^2)$$

$$- \eta\Big[(n_i^2+n_j^2-4)\tilde{z}_i\tilde{z}_j - (n_i^2+n_j^2)\rho_{ij}^2\tilde{z}_i\tilde{z}_j + 2\rho_{ij}(\tilde{z}_i^2+\tilde{z}_j^2)\Big] + O(\eta^2) \tag{F.24}$$

$$= \rho_{ij} - \eta\Big[(n_i^2+n_j^2)(1-\rho_{ij}^2)\tilde{z}_i\tilde{z}_j + (n_i^2+n_j^2-4)\tilde{z}_i\tilde{z}_j$$

$$- (n_i^2+n_j^2)\rho_{ij}^2\tilde{z}_i\tilde{z}_j + 2\rho_{ij}(\tilde{z}_i^2+\tilde{z}_j^2)\Big] + O(\eta^2) \tag{F.25}$$

$$= \rho_{ij} - 2\eta\Big[\rho_{ij}(\tilde{z}_i^2+\tilde{z}_j^2) + \tilde{z}_i\tilde{z}_j\big((n_i^2+n_j^2)(1-\rho_{ij}^2)-2\big)\Big] + O(\eta^2). \tag{F.26}$$

Setting $n_i = n_j = 1$ recovers the normalized case in Equation (E.17).

Since

$$\Delta(\tilde{\boldsymbol{D}}_S) = 2\rho_{ij}^2, \tag{F.27}$$

the first-order change in interference measure is

$$\Delta(\tilde{\boldsymbol{D}}_S^+) - \Delta(\tilde{\boldsymbol{D}}_S) = -8\eta\rho_{ij}\Big[\rho_{ij}(\tilde{z}_i^2+\tilde{z}_j^2)$$

$$+ \tilde{z}_i\tilde{z}_j\big((n_i^2+n_j^2)(1-\rho_{ij}^2)-2\big)\Big] + O(\eta^2). \tag{F.28}$$

Hence, the interference measure decreases to first order whenever

$$\rho_{ij}\Big[\rho_{ij}(\tilde{z}_i^2+\tilde{z}_j^2) + \tilde{z}_i\tilde{z}_j\big((n_i^2+n_j^2)(1-\rho_{ij}^2)-2\big)\Big] > 0. \tag{F.29}$$

Using

$$\rho_{ij}\left(\tilde{z}_i^2+\tilde{z}_j^2\right) + \tilde{z}_i\tilde{z}_j\left[(n_i^2+n_j^2)(1-\rho_{ij}^2)-2\right] \tag{F.30}$$

$$= \rho_{ij}\left(\tilde{z}_i^2+\tilde{z}_j^2-2\tilde{z}_i\tilde{z}_j\right) + \tilde{z}_i\tilde{z}_j\left[(n_i^2+n_j^2)(1-\rho_{ij}^2)-2+2\rho_{ij}\right] \tag{F.31}$$

$$= \rho_{ij}\left(\tilde{z}_i-\tilde{z}_j\right)^2 + \tilde{z}_i\tilde{z}_j\left[(n_i^2+n_j^2)(1-\rho_{ij}^2)-2(1-\rho_{ij})\right] \tag{F.32}$$

$$= \rho_{ij}\left(\tilde{z}_i-\tilde{z}_j\right)^2 + (1-\rho_{ij})\left[(n_i^2+n_j^2)(1+\rho_{ij})-2\right]\tilde{z}_i\tilde{z}_j. \tag{F.33}$$

and assuming $\tilde{z}_i\tilde{z}_j > 0$, the exact threshold is

$$\begin{cases} n_i^2+n_j^2 > \dfrac{2}{1+\rho_{ij}} - \dfrac{\rho_{ij}}{1-\rho_{ij}^2}\dfrac{(\tilde{z}_i-\tilde{z}_j)^2}{\tilde{z}_i\tilde{z}_j}, & 0<\rho_{ij}<1, \\ n_i^2+n_j^2 < \dfrac{2}{1+\rho_{ij}} - \dfrac{\rho_{ij}}{1-\rho_{ij}^2}\dfrac{(\tilde{z}_i-\tilde{z}_j)^2}{\tilde{z}_i\tilde{z}_j}, & -1<\rho_{ij}<0. \end{cases} \tag{F.34}$$

The correction term always shifts the threshold in the direction that makes the inequality easier to satisfy. Therefore, a $\tilde{z}_i, \tilde{z}_j$-independent sufficient condition is

$$\begin{cases} n_i^2+n_j^2 > \dfrac{2}{1+\rho_{ij}}, & 0<\rho_{ij}<1, \\ n_i^2+n_j^2 < \dfrac{2}{1+\rho_{ij}}, & -1<\rho_{ij}<0. \end{cases} \tag{F.35}$$

This guarantees a first-order decrease for any $\tilde{z}_i\tilde{z}_j > 0$.

The boundary cases are immediate. At $\rho_{ij} = 0$, the first-order change vanishes. At $\rho_{ij} = 1$, the normalized feature directions remain positively collinear, so $\rho_{ij}^+ = 1$ exactly. The case $\rho_{ij} = -1$ is incompatible with $\tilde{z}_i\tilde{z}_j > 0$ for nonzero tied codes. ■

When $n_i = n_j = 1$, the sufficient condition holds for all $-1 < \rho_{ij} < 1$, $\rho_{ij} \neq 0$, recovering the normalized result. Without normalization, however, the angular dynamics also depend on $n_i^2 + n_j^2$. Thus, gradient descent need not enforce local orthogonalization: feature norms provide an additional degree of freedom through which interference can be reduced.

G Appendix - Column Normalization

The mean squared error objective does not uniquely determine the scale of the decoder directions, since scaling a decoder feature $\boldsymbol{d}_i$ by a constant and inversely scaling its activation z_i leaves the reconstruction unchanged. To remove this scaling ambiguity, sparse dictionary learning commonly constrain each decoder feature to have unit norm. In practice, SAE implementations typically enforce this constraint through two operations applied sequentially at each training step (Huben et al., 2024; Bricken et al., 2023; Gao et al., 2025). First, the component of the gradient parallel to each decoder column is projected out, so that the optimizer update is tangent to the unit sphere. After the optimizer step, each decoder column is explicitly renormalized to unit norm.

The gradient projection alone does not enforce the constraint exactly: even a tangent update can cause the feature norms to drift over long training horizons (Douglas et al., 2002). It nevertheless ensures that the gradient passed to adaptive optimizers such as Adam (Kingma & Ba, 2014) is consistent with the unit-norm constraint, rather than relying on a subsequent renormalization to correct an unconstrained update. This has been observed empirically to yield small but consistent improvements in the optimization objective (Bricken et al., 2023). The subsequent explicit renormalization then corrects the remaining norm drift and restores unit norm exactly after every optimizer step.

In the following, we make the relationship between these two sequential operations explicit by deriving the first-order expansion of a renormalized gradient update. We show that, to the first order, explicit renormalization itself induces a projection of the gradient onto the tangent space of the unit sphere, with the discrepancy appearing only at higher order.

Consider a dictionary feature $\boldsymbol{d}_i$ satisfying $\|\boldsymbol{d}_i\|_2 = 1$, and denote its gradient by

$$\boldsymbol{g}_i := \nabla_{\boldsymbol{d}_i} \mathcal{L}. \tag{G.1}$$

After a gradient step with step size $\eta > 0$ followed by renormalization, the updated feature is

$$\boldsymbol{d}_i^+ = \frac{\boldsymbol{d}_i - \eta \boldsymbol{g}_i}{\|\boldsymbol{d}_i - \eta \boldsymbol{g}_i\|_2}. \tag{G.2}$$

The normalization factor satisfies

$$\|\boldsymbol{d}_i - \eta \boldsymbol{g}_i\|_2 = \sqrt{(\boldsymbol{d}_i - \eta \boldsymbol{g}_i)^\top (\boldsymbol{d}_i - \eta \boldsymbol{g}_i)} \tag{G.3}$$

$$= \sqrt{1 - 2\eta \boldsymbol{g}_i^\top \boldsymbol{d}_i + \eta^2 \boldsymbol{g}_i^\top \boldsymbol{g}_i}. \tag{G.4}$$

Define

$$\alpha_i := 2\eta \boldsymbol{g}_i^\top \boldsymbol{d}_i - \eta^2 \boldsymbol{g}_i^\top \boldsymbol{g}_i. \tag{G.5}$$

Then

$$\frac{1}{\|\boldsymbol{d}_i - \eta \boldsymbol{g}_i\|_2} = (1 - \alpha_i)^{-1/2}. \tag{G.6}$$

Using Taylor's theorem with Lagrange remainder,

$$(1 - \zeta_i)^{-1/2} = 1 + \frac{\zeta_i}{2} + \frac{3\zeta_i^2}{8}(1 - \xi)^{-5/2}, \tag{G.7}$$

for some ξ between 0 and ζ_i. Substituting the expression for ζ_i, we get

$$\boldsymbol{d}_i^+ = (\boldsymbol{d}_i - \eta \boldsymbol{g}_i) \Bigg[1 + \eta \boldsymbol{g}_i^\top \boldsymbol{d}_i - \frac{\eta^2}{2} \boldsymbol{g}_i^\top \boldsymbol{g}_i + \frac{3}{8} \left(2\eta \boldsymbol{g}_i^\top \boldsymbol{d}_i - \eta^2 \boldsymbol{g}_i^\top \boldsymbol{g}_i \right)^2 (1 - \xi)^{-5/2} \Bigg]. \tag{G.8}$$

Keeping the first-order terms in η yields

$$\boldsymbol{d}_i^+ = \boldsymbol{d}_i - \eta \boldsymbol{g}_i + \eta \boldsymbol{d}_i \boldsymbol{d}_i^\top \boldsymbol{g}_i + R_i \tag{G.9}$$

$$= \boldsymbol{d}_i - \eta \left(\boldsymbol{I} - \boldsymbol{d}_i \boldsymbol{d}_i^\top \right) \boldsymbol{g}_i + R_i, \tag{G.10}$$

where R_i collects the second- and higher-order terms. More explicitly,

$$R_i = -\eta^2 \boldsymbol{g}_i(\boldsymbol{g}_i^\top \boldsymbol{d}_i) - \frac{\eta^2}{2}\boldsymbol{d}_i(\boldsymbol{g}_i^\top \boldsymbol{g}_i) + \frac{3}{8}(\boldsymbol{d}_i - \eta \boldsymbol{g}_i)\left(2\eta \boldsymbol{g}_i^\top \boldsymbol{d}_i - \eta^2 \boldsymbol{g}_i^\top \boldsymbol{g}_i\right)^2 (1-\xi)^{-5/2}, \tag{G.11}$$

for some ξ between 0 and $2\eta \boldsymbol{g}_i^\top \boldsymbol{d}_i - \eta^2 \boldsymbol{g}_i^\top \boldsymbol{g}_i$.

Thus, to first order in the learning rate, explicit renormalization is equivalent to replacing the gradient $\boldsymbol{g}_i$ by its projection onto the tangent space of the unit sphere at $\boldsymbol{d}_i$:

$$\boldsymbol{g}_i \longmapsto \left(\boldsymbol{I} - \boldsymbol{d}_i \boldsymbol{d}_i^\top\right) \boldsymbol{g}_i. \tag{G.12}$$

Therefore, explicit normalization and tangent-space gradient projection induce the same leading-order dynamics on the unit sphere. Their difference is captured by the higher-order remainder $R_i = O(\eta^2)$.

ℓ_1 regularization. The unit-norm constraint is particularly important for SAEs trained with ℓ_1 regularization, as otherwise the sparsity penalty can be reduced by increasing the decoder norms while proportionally decreasing the codes, without changing the reconstruction. Recent work has proposed removing the unit-norm constraint by instead using the reparameterization-invariant regularizer $\lambda \sum_{i \in S} \|\boldsymbol{d}_i\|_2 z_i$ (Rajamanoharan et al., 2024b; Conerly et al., 2024).

H APPENDIX - GRADIENT COMPUTATIONS

In this appendix, we derive the gradient expressions used throughout the proofs. We consider a single input $\boldsymbol{x}$ and define its active support as

$$S(\boldsymbol{x}) := \{i : z_i(\boldsymbol{x}) \neq 0\}. \tag{H.1}$$

We treat the active support as locally fixed throughout the derivation, such that inactive features receive zero gradient. The reconstruction is then

$$\hat{\boldsymbol{x}} = \sum_{i \in S(\boldsymbol{x})} \boldsymbol{d}_i z_i, \tag{H.2}$$

and we consider the reconstruction loss

$$\mathcal{L}(\boldsymbol{x}) = \frac{1}{2}\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2^2. \tag{H.3}$$

In the tied setting, the encoder and decoder weights are shared, such that $\boldsymbol{W} = \boldsymbol{D}$. For an active feature $i \in S(\boldsymbol{x})$, the activation is

$$z_i = \boldsymbol{d}_i^\top \boldsymbol{x} + b_i, \tag{H.4}$$

and the reconstruction is therefore

$$\hat{\boldsymbol{x}} = \sum_{i \in S(\boldsymbol{x})} \boldsymbol{d}_i z_i = \sum_{i \in S(\boldsymbol{x})} \boldsymbol{d}_i \left(\boldsymbol{d}_i^\top \boldsymbol{x} + b_i\right). \tag{H.5}$$

The reconstruction loss is

$$\mathcal{L}(\boldsymbol{x}) = \frac{1}{2}\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2^2 = \frac{1}{2}\Big\| \sum_{i \in S(\boldsymbol{x})} \boldsymbol{d}_i \left(\boldsymbol{d}_i^\top \boldsymbol{x} + b_i\right) - \boldsymbol{x}\Big\|_2^2. \tag{H.6}$$

Gradient with respect to a dictionary feature. Because the weights are tied, $\boldsymbol{d}_i$ affects the reconstruction in two ways: directly through the decoder and indirectly through the activation z_i. For an active feature $i \in S(\boldsymbol{x})$, we have

$$\frac{\partial z_i}{\partial \boldsymbol{d}_i} = \boldsymbol{x}^\top. \tag{H.7}$$

Using the product rule, the derivative of the reconstruction with respect to $\boldsymbol{d}_i$ is therefore

$$\frac{\partial \hat{\boldsymbol{x}}}{\partial \boldsymbol{d}_i} = \frac{\partial}{\partial \boldsymbol{d}_i}\left(\boldsymbol{d}_i z_i\right) = \left(\frac{\partial \boldsymbol{d}_i}{\partial \boldsymbol{d}_i}\right) z_i + \boldsymbol{d}_i \left(\frac{\partial z_i}{\partial \boldsymbol{d}_i}\right) = \boldsymbol{I} z_i + \boldsymbol{d}_i \boldsymbol{x}^\top. \tag{H.8}$$

Since the gradient of the reconstruction loss with respect to $\hat{\boldsymbol{x}}$ is $\hat{\boldsymbol{x}} - \boldsymbol{x}$, applying the chain rule yields

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{d}_i} = (\hat{\boldsymbol{x}} - \boldsymbol{x})^\top \left(\frac{\partial \hat{\boldsymbol{x}}}{\partial \boldsymbol{d}_i}\right) = (\hat{\boldsymbol{x}} - \boldsymbol{x})^\top \left(\boldsymbol{I} z_i + \boldsymbol{d}_i \boldsymbol{x}^\top\right) = (\hat{\boldsymbol{x}} - \boldsymbol{x})^\top z_i + \left((\hat{\boldsymbol{x}} - \boldsymbol{x})^\top \boldsymbol{d}_i\right) \boldsymbol{x}^\top. \tag{H.9}$$

Hence,

$$\nabla_{\boldsymbol{d}_i} \mathcal{L} = \begin{cases} (\hat{\boldsymbol{x}} - \boldsymbol{x}) z_i + \boldsymbol{x} \boldsymbol{d}_i^\top (\hat{\boldsymbol{x}} - \boldsymbol{x}) & i \in S(\boldsymbol{x}), \\ 0, & i \notin S(\boldsymbol{x}). \end{cases} \tag{H.10}$$

The first term corresponds to the contribution of $\boldsymbol{d}_i$ through the decoder, while the second arises from its contribution through the tied encoder.

I Appendix - Experimental Details

Model and data. We use Pythia-160M-deduped (Biderman et al., 2023), taking residual-stream activations after layer 8 (`blocks.8.hook_resid_post`, $m = 768$). Activations come from the uncopyrighted Pile (Gao et al., 2020), using 1024-token contexts with a prepended BOS token, and are cached with SAELens (Bloom et al., 2024). We center activations by the training-set geometric median as in (Gao et al., 2025), but compute $\boldsymbol{b}_{\text{pre}}$ once and keep it fixed. Throughout, $\boldsymbol{x} = \boldsymbol{x}_{\text{raw}} - \boldsymbol{b}_{\text{pre}}$; no other normalization is applied.

Training. Each SAE is trained for 200M tokens (48,828 steps of 2^{14} tokens) on disjoint activations and evaluated on a fixed, disjoint 2M-token holdout set.[2] We use Adam (Kingma & Ba, 2014) with learning rate 2×10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, no weight decay, schedule, or warmup.

Architecture. All models use

$$\boldsymbol{z} = \sigma\big(\mathrm{ReLU}\big(\boldsymbol{W}^\top \boldsymbol{x} + \boldsymbol{b}\big)\big), \qquad \hat{\boldsymbol{x}} = \boldsymbol{D}\boldsymbol{z}, \tag{I.1}$$

with $\boldsymbol{D} \in \mathbb{R}^{m \times p}$ unit-norm columnwise, $p = 16{,}384$, and $k = 40$. We compare TopK (Gao et al., 2025), BatchTopK (Bussmann et al., 2024), JumpReLU (Rajamanoharan et al., 2024b), and Matryoshka (Bussmann et al., 2025). For JumpReLU, we enforce sparsity through L0 target regularization. For Matryoshka, we use the SAEBench block boundaries at 512, 1536, 3584, and 7680 features (Karvonen et al., 2025).

Initialization and constraints. $\boldsymbol{D}$ is Gaussian-initialized and column-normalized, with $\boldsymbol{W} = \boldsymbol{D}$ and $\boldsymbol{b} = \boldsymbol{0}$ at initialization. Decoder columns remain unit norm by projecting out the radial gradient component and renormalizing after each update (Section G).

We parameterize the encoder weight as $\boldsymbol{W} = \boldsymbol{\alpha} \odot \boldsymbol{D}$ or $\boldsymbol{W} = \boldsymbol{D}$, where $\boldsymbol{\alpha} \in \mathbb{R}^p$ is a separately learnable, per-feature gain (initialized as $\boldsymbol{\alpha} = \boldsymbol{1}$) that rescales each column of $\boldsymbol{D}$ independently of its unit-norm constraint. We consider five settings:

(i) *tied*: $\boldsymbol{W} = \boldsymbol{D}$, $\boldsymbol{\alpha} = \boldsymbol{1}$, $\boldsymbol{b} = \boldsymbol{0}$;
(ii) *tied + bias*: $\boldsymbol{W} = \boldsymbol{D}$, $\boldsymbol{\alpha} = \boldsymbol{1}$, with $\boldsymbol{b}$ learned;
(iii) *tied + gain*: $\boldsymbol{W} = \boldsymbol{\alpha} \odot \boldsymbol{D}$, with $\boldsymbol{\alpha}$ learned and $\boldsymbol{b} = \boldsymbol{0}$;
(iv) *tied + gain + bias*: $\boldsymbol{W} = \boldsymbol{\alpha} \odot \boldsymbol{D}$, with both $\boldsymbol{\alpha}$ and $\boldsymbol{b}$ learned;
(v) *untied*: $\boldsymbol{W}$ is learned independently, with $\boldsymbol{\alpha} = \boldsymbol{1}$ and $\boldsymbol{b} = \boldsymbol{0}$.

In the *tied + gain* settings, this lets us isolate the effect of decoupling the encoder's per-feature scale from the decoder's direction, without also decoupling the direction itself. In the untied setting, $\boldsymbol{W}$ already carries per-feature scale, so we do not additionally learn $\boldsymbol{\alpha}$.

Seeds. Figure 2 averages over three seeds, varying both initialization and token order; all other experiments use a single seed. Although the aggregate geometric and effective interference patterns in this figure are stable across seeds, the learned dictionaries differ substantially, consistent with previous observations (Fel et al., 2025; Nelson et al., 2026; Song et al., 2026; Paulo & Belrose, 2026). For the runs in Figure 2, Table 1 quantifies this using mean maximum cosine similarity,

$$\mathrm{MMCS} = \frac{1}{p} \sum_i \max_j \cos(\boldsymbol{d}_i^A, \boldsymbol{d}_j^B), \tag{I.2}$$

and its absolute-value counterpart.

	MMCS	median	$\mathrm{MMCS}_{\lvert\cdot\rvert}$	>0.9	>0.7
JumpReLU	0.711	0.785	0.712	24.8%	61.1%
TopK	0.548	0.533	0.552	8.9%	31.2%
BatchTopK	0.436	0.335	0.444	6.0%	18.5%
Matryoshka	0.425	0.332	0.433	3.1%	15.8%

Table 1: Cross-seed dictionary agreement for the runs in Figure 2, averaged over seed pairs $(0, 1)$, $(0, 2)$, and $(1, 2)$.

[2] (Karvonen et al., 2025) report that many SAE evaluation metrics achieve most of their performance by 50M training tokens.

Evaluation. We report fraction of variance unexplained (FVU), average sparsity $\mathbb{E}\|\boldsymbol{z}\|_0$, and the fraction of features that never activate on the holdout set. Sparsity is closely matched across settings, while FVU generally decreases as constraints are relaxed.

	TopK			BatchTopK			Matryoshka			JumpReLU		
Constraint	FVU	L0	Dead	FVU	L0	Dead	FVU	L0	Dead	FVU	L0	Dead
(i) Tied	.118	40.0	0.00	.124	40.9	0.00	.126	40.7	0.00	.100	45.6	0.02
(ii) Tied + Bias	.086	40.0	0.00	.085	40.1	0.00	.091	40.1	0.00	.086	45.5	0.48
(iii) Tied + Gain	.091	40.0	0.00	.092	40.0	0.12	.096	40.2	0.01	.087	45.4	0.62
(iv) Tied + Gain + Bias	.085	40.0	0.01	.084	40.0	0.02	.090	40.1	0.01	.085	45.4	0.90
(v) Untied	.083	40.0	0.54	.084	40.8	5.41	.088	40.6	6.05	.083	45.0	17.62

Table 2: Holdout evaluation after 200M training tokens. "Dead" is the percentage of features that never activate on the holdout set.

J Appendix - Additional Results

J.1 Extended results

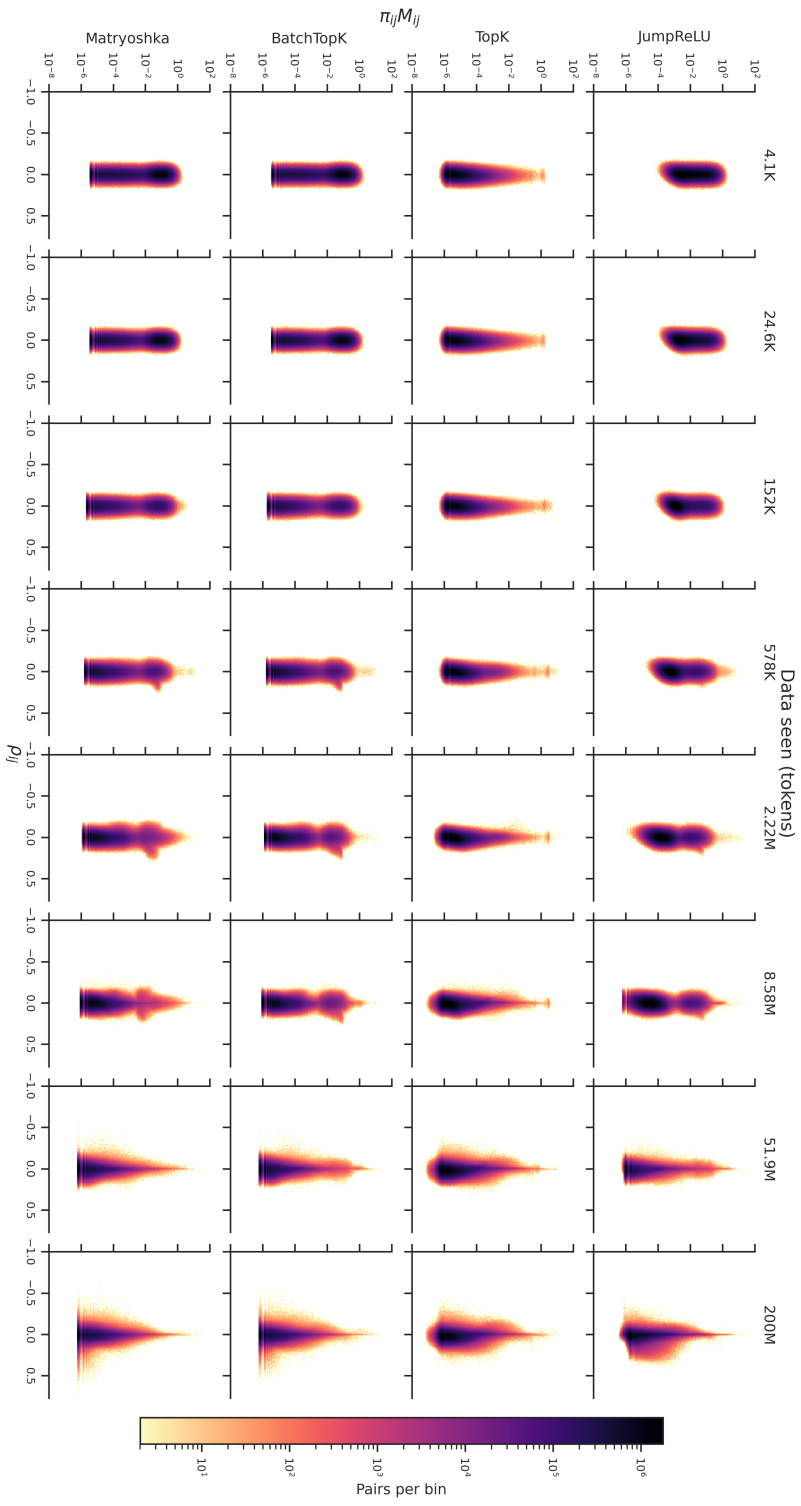


Figure 7: **Transition of interference factors as training progress.** This figure complements Figure 2.

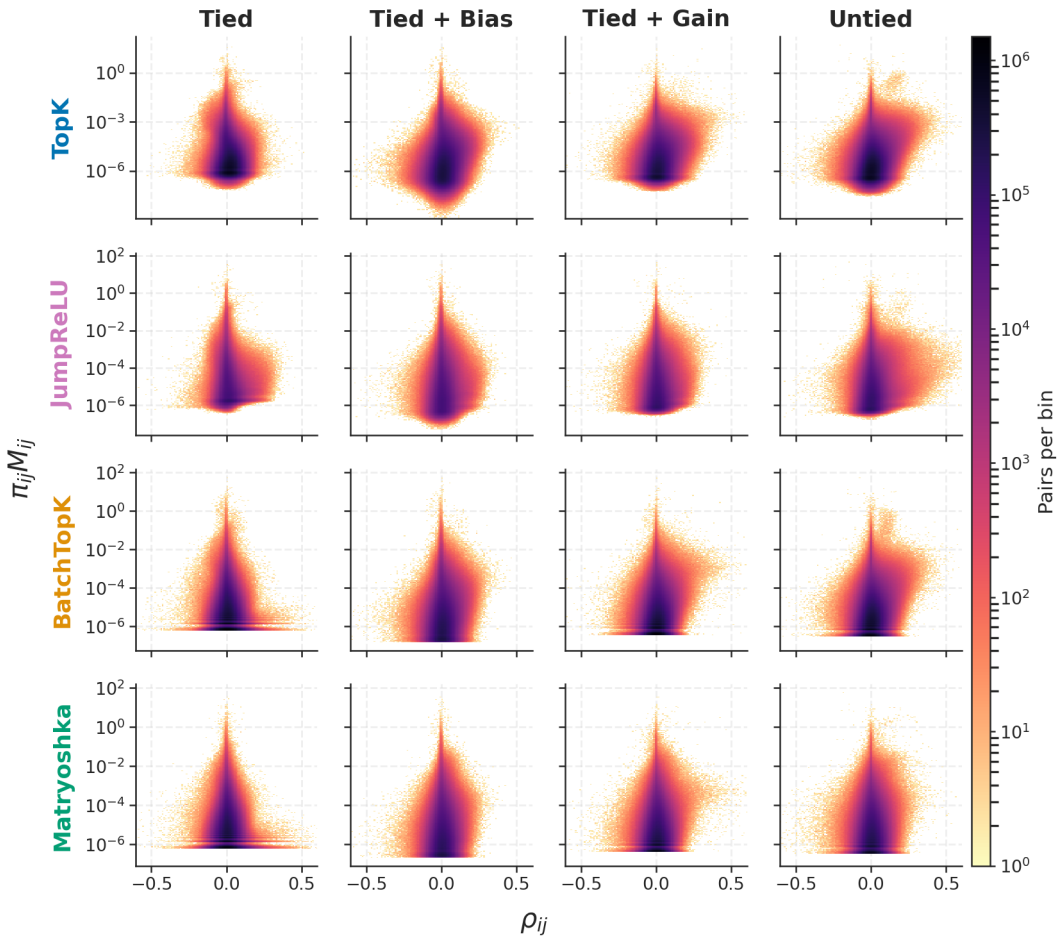


Figure 8: **Consistent trend on architectural freedom shifting co-active feature pairs toward constructive decoder interactions**. This figure complements Figure 3.

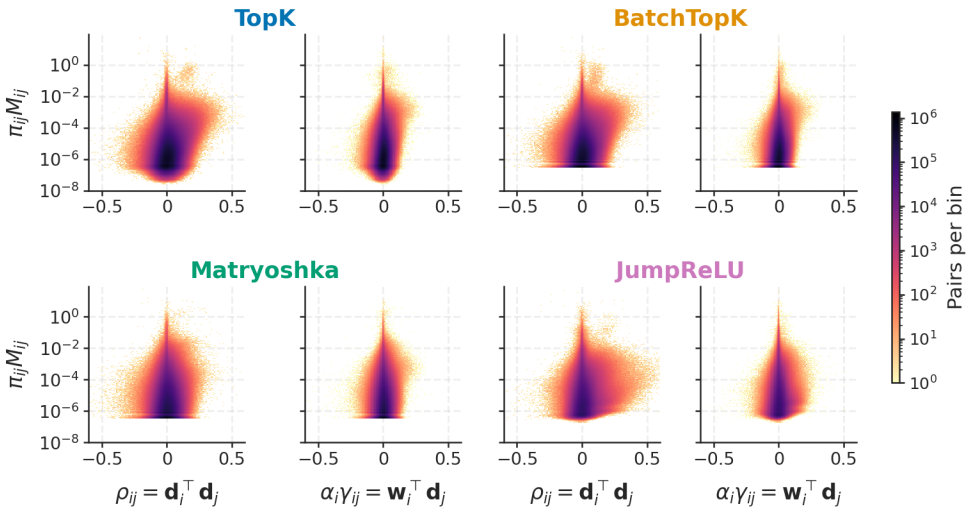


Figure 9: **The untying mechanism to separate constructive decoder interactions from encoder cross-talk is consistent across SAEs, including TopK, BatchTopK, Matryoshka, and JumpReLU.** This figure complements Figure 5.

J.2 Interference under varying sparsity and dictionary width

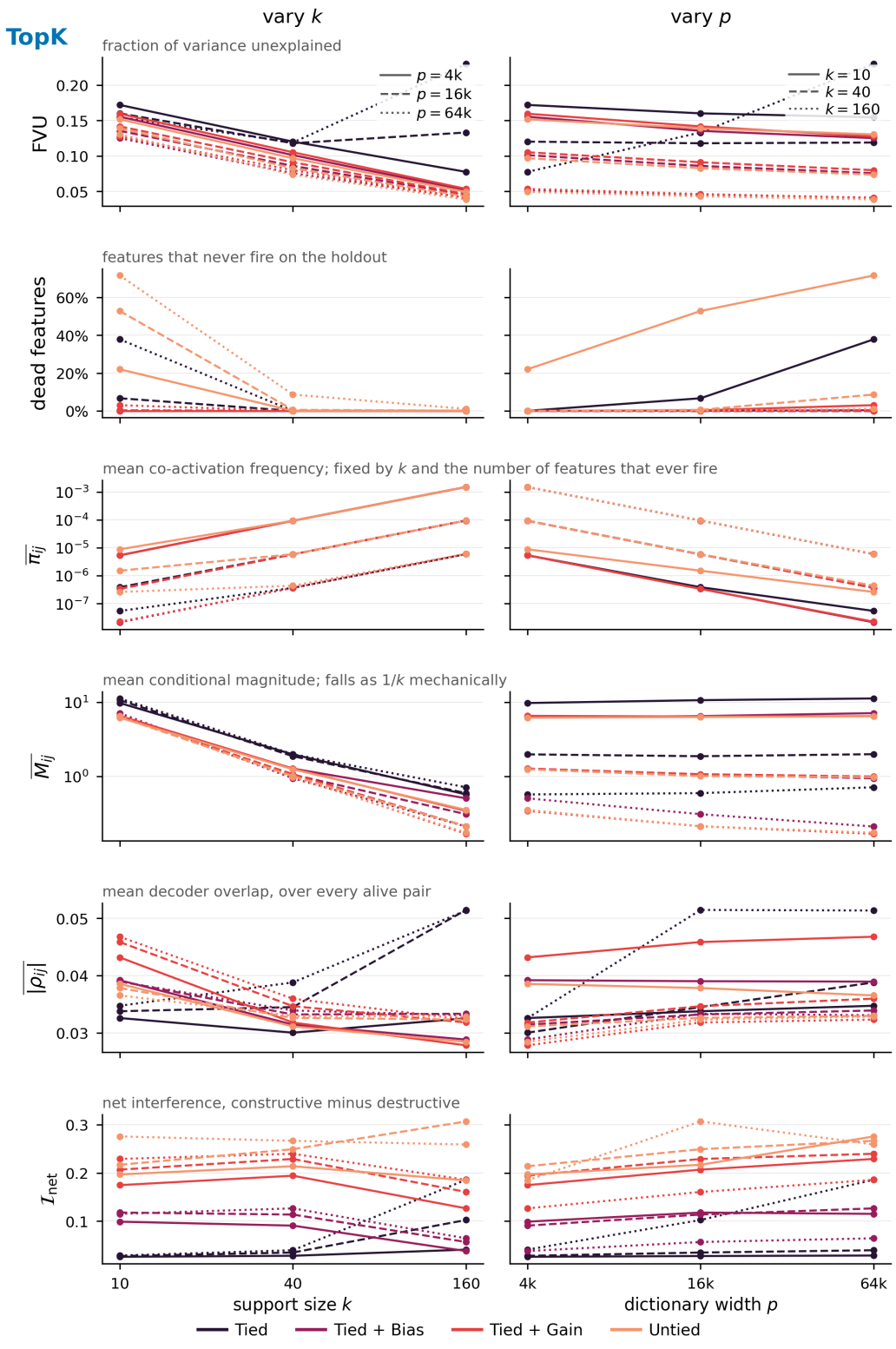


Figure 10: **Interference across sparsity and dictionary width.** TopK SAEs trained on Pythia-160M, varying constraints, support size k, and width p. **Left:** varying k **Right:** varying p.

J.3 Generality of the interference signatures

We test whether the interference signatures observed in the main experiments generalize beyond our training setup. We consider two settings: a different modality (MNIST; Section J.3.1) and independently trained language-model SAEs from SAEBench (Section J.3.2).

J.3.1 A different modality: MNIST

We train SAEs on MNIST with pixels scaled to $[0, 1]$, using 50,000 training and 10,000 validation examples. Models have $p = 1{,}000$ features and $k = 10$, and are trained for 1,000 epochs with Adam at learning rate 10^{-4}. Decoder columns are kept unit norm, and inputs are centered using the geometric median of the training set (Bricken et al., 2023; Gao et al., 2025; Bussmann et al., 2024). JumpReLU uses an ℓ_0-target penalty with coefficient 10 (Rajamanoharan et al., 2024b); Matryoshka uses nested groups of sizes 250, 500, 750, and 1,000 (Bussmann et al., 2025).

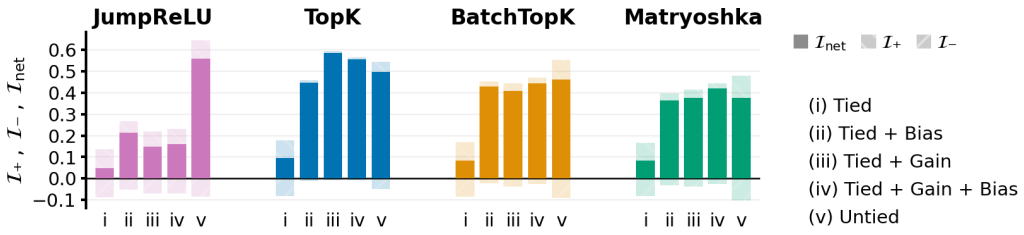


Figure 11: **Effective interference on MNIST.** The same qualitative pattern as in language models appears: interference is lowest in the most constrained tied setting and increases as architectural degrees of freedom are added. This figure complements Figure 6.

J.3.2 Independently trained SAEs: SAEBench

Our main experiments are deliberately controlled: all architectures are trained on the same model, data, width, sparsity, and optimization setup, allowing us to isolate the effect of architectural constraints. To test whether the same interference signatures persist beyond this controlled setting, we also evaluate independently trained SAEs released with SAEBench (Karvonen et al., 2025). We use $p = 16{,}384$ checkpoints targeting $k = 40$ for gemma-2-2b (Team et al., 2024) layer 12 and Pythia-160M-deduped layer 8, evaluated using SAEBench's own OpenWebText pipeline.

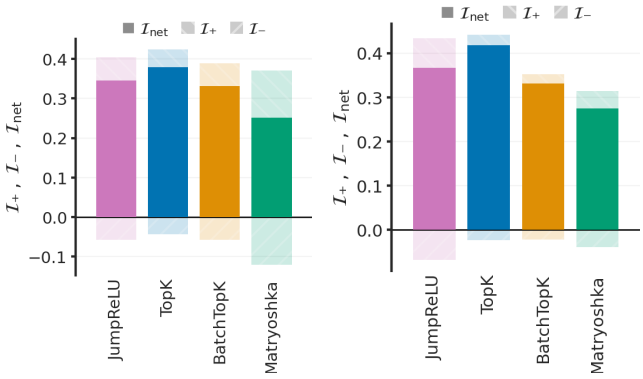


Figure 13: **Effective interference for independently trained SAEBench SAEs.** Gemma-2-2B layer 12 (left) and Pythia-160M layer 8 (right), across the four architectures.

These SAEs differ from ours in several aspects of the training recipe: they are trained for longer, use learning-rate warmup and decay, learn the pre-encoder bias rather than fixing it to the geometric median, and, for the hard-sparsity architectures, include an auxiliary loss designed to revive dead features. They are also available only in the untied setting, so their absolute interference levels are not directly comparable to those in our controlled sweep. Despite these differences, the same qualitative signatures remain visible across both model families (Figure 14). In particular, the SAEBench models exhibit substantial effective interference (Figure 13), generally more than we observe in our own untied runs, suggesting that additional optimization and training choices in these pipelines may themselves contribute to effective interference.

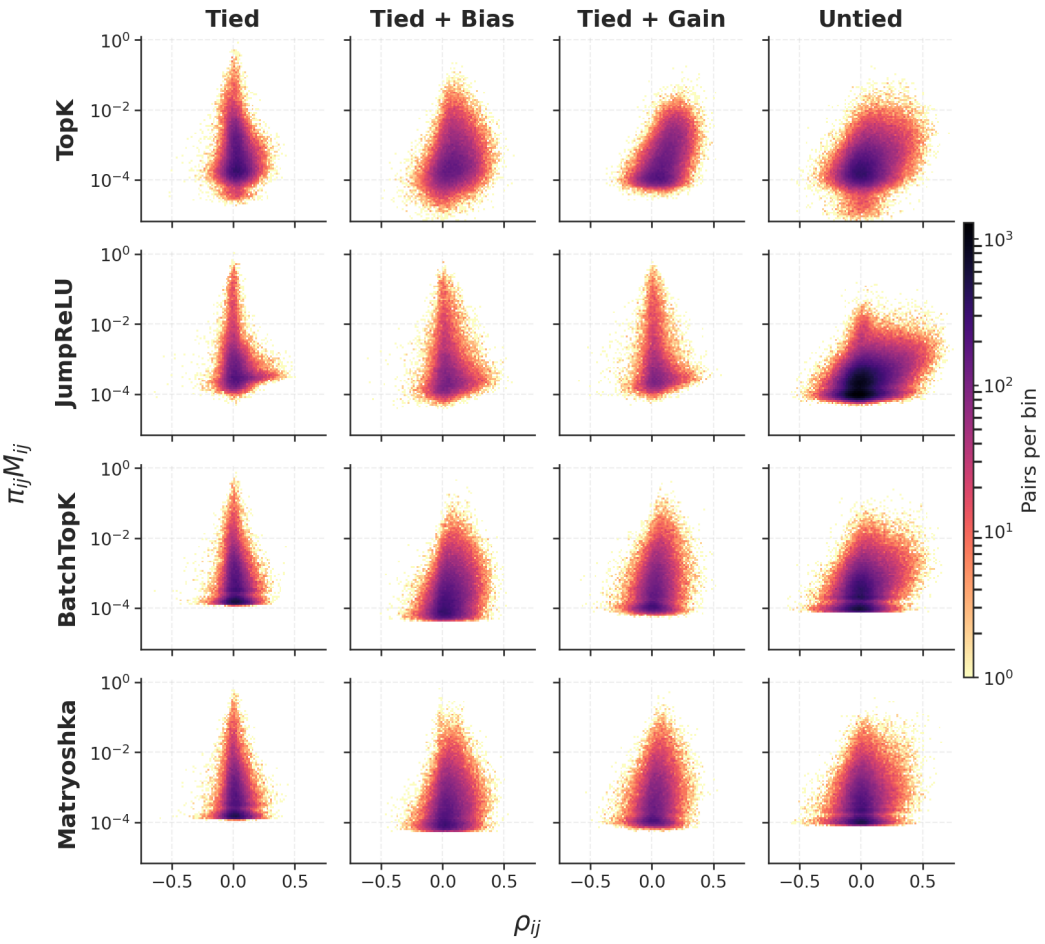


Figure 12: **Decoder orientation and co-activation on MNIST.** Relaxing the constraints broadens the distribution of decoder correlations and increases co-activation among non-orthogonal feature pairs. This figure complements Figure 3.

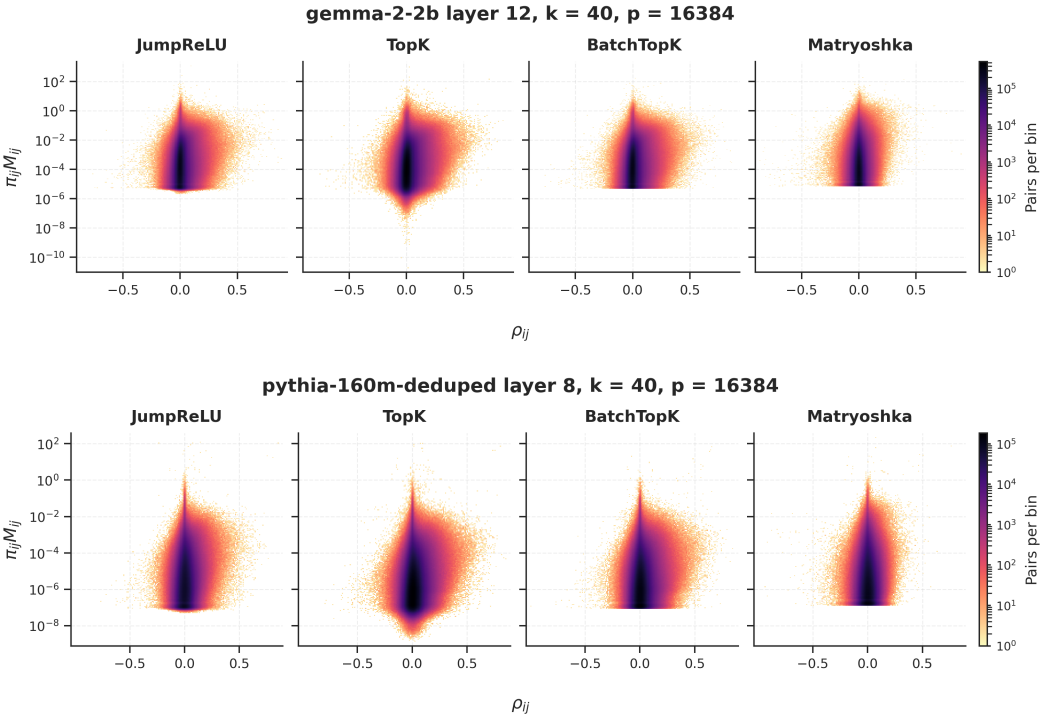


Figure 14: **Decoder orientation and co-activation for independently trained SAEBench SAEs.** Gemma-2-2B layer 12 (top) and Pythia-160M layer 8 (bottom).