SPARSE AUTOENCODERS DO NOT GUARANTEE LOCAL FEATURES

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse autoencoders (SAEs) can recover feature coordinates without identifying which inputs determine their activations. A feature associated with one input pattern may still depend on other variables. Objectives based only on representation samples cannot distinguish input maps that induce the same representation distribution. We therefore ask how compactly the activations of a frozen SAE can be explained in terms of declared controlled input variables. The optimal input scope curve measures the least activation variance omitted under a total dependency budget. For every interaction order r, we construct systems with identical representation distributions and the same SAE that require exactly r times the order one scope budget at the same omission tolerance. This separation persists when an exact sparse coding objective uniquely selects the shared dictionary. General lower bounds connect scope requirements to joint input effects and representation directions. Paired interventions provide finite sample control after scope selection. In a controlled factorial experiment with five learned overcomplete SAEs, two input maps increase the mean budget needed to explain 90% of activation variance by 41% and 49%, despite unchanged representation and activation samples.

1 INTRODUCTION

Sparse autoencoders are used to interpret model representations by relating learned features to patterns in the input. A feature associated with one pattern may nevertheless depend on other input variables. Input localization is therefore a relation between a fixed feature system and a declared intervention design. The question is how compactly the controlled variables explain the activations of the recovered coordinates, not only whether those coordinates reconstruct the representation.

Feature recovery theory asks when sparse objectives identify the generating coordinates (Gribonval & Schnass, 2010; Chen et al., 2026; Cui et al., 2026), while causal SAE analyses study which inputs drive their activations (Han et al., 2025). Recovery alone does not settle the latter question. An objective that uses only representation samples sees the same learning problem whenever two input maps induce the same representation distribution. Yet the same feature coordinates may require different inputs under those maps, even if the sparse objective selects the coordinates uniquely. We ask how much their localization requirements can differ while both the representation distribution and the SAE remain fixed.

To answer this question, we introduce the optimal input scope curve for a frozen SAE relative to a specified intervention design. It measures the smallest unexplained activation variance attainable by assigning input subsets to features under a shared dependency budget. Optimizing these assignments distinguishes an inadequate scope choice from dependence that no choice within the budget can capture. This formulation also connects theory to measurement. Paired interventions estimate the omitted variance, and finite sample bounds control the curve after scope selection. Figure 1 previews the corresponding learned comparison: bijective input maps preserve the representation and activation samples while producing different input scope budgets.

Our main contributions are summarized below.

- We prove that order r can require exactly r times the scope budget of order one at the same omission tolerance while the representation distribution and SAE remain fixed. The shared dictionary is uniquely optimal in a specified exact sparse coding problem.

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.

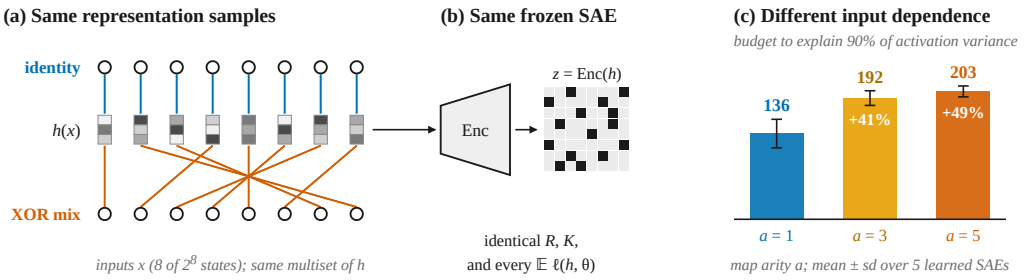

Figure 1: Preview of the central separation. (a) Identity and bijective XOR maps pair the controlled inputs with the same representation samples in different ways. (b) The same frozen SAE therefore receives identical representation and activation samples, leaving reconstruction, sparsity, and every representation only objective unchanged. (c) Across five learned overcomplete SAEs, the mean budget needed to explain 90% of activation variance increases from 136 under the identity map to 192 and 203 for map arities 3 and 5; error bars show one standard deviation. Thus identical representation evidence can support substantially different input dependence.

- We derive compatibility and spectral lower bounds on localization cost. Joint effects constrain which inputs a relevant feature must cover, while their directional spread constrains how many such features are needed.
- We establish stability under aligned encoding perturbations and finite sample control of the scope curve. A controlled factorial experiment with five learned overcomplete SAEs isolates input correspondence by holding representation and activation samples fixed while the localization budget changes.

2 Related Work

Dictionary identification and SAE recovery theory study when sparse objectives recover features under specified generative and training assumptions (Gribonval & Schnass, 2010; Chen et al., 2026; Cui et al., 2026). Our question begins after the representation distribution and feature coordinates have been fixed. We ask whether this information determines how those coordinates depend on controlled inputs. The exact separation shows that it need not, even when the shared dictionary is uniquely selected by an exact sparse coding objective.

Empirical SAE evaluations study feature quality beyond reconstruction and sparsity (Huben et al., 2024; Gao et al., 2025; Karvonen et al., 2025), while input attribution analyses ask which inputs drive activations (Han et al., 2025). Ro et al. (Ro et al., 2026) study visual information scope under spatial perturbations. Our curve instead measures conditional variance relative to declared controlled variables and optimizes feature scopes under a shared dependency budget.

Feature stability work shows that independently trained SAEs can recover different coordinates or reproducible subspaces (Song et al., 2026; Paulo & Belrose, 2026; Gerasimov et al., 2026). Analyses of sparse optima and subspace representations further examine how feature structure depends on the learned dictionary (Dorrell, 2026; Dalili & Mahdavi, 2026). Our comparison begins after the feature coordinates are fixed and isolates differences in their dependence on controlled inputs. This separates uncertainty in the learned basis from uncertainty in the map between inputs and a fixed basis. The resulting scope curve is a property of that feature system under the chosen interventions.

Functional ANOVA isolates effects that require specified input sets (Lengerich et al., 2020). Conditional resampling provides variance estimators (Janon et al., 2014), while effective dimension and sparse approximation give related perspectives on interactions and restricted representations (Owen, 2019; Fletcher et al., 2006). We use these tools for scalar feature functions combined through one fixed decoder. Joint effects must then be carried by features with compatible scopes and decoder directions, which connects input dependence to a total scope budget.

3 INPUT LOCALIZATION FOR A FROZEN SAE

We formalize input localization for a frozen SAE relative to a specified intervention design. The intervention variables, their domains, sampling distribution, and dependency cost define the comparison, so scope curves are directly comparable when this design is held fixed. Pure input effects identify which variables a feature cannot omit, conditional projection measures what a proposed scope leaves unexplained, and a shared dependency budget compares the best scope assignments at the same complexity.

3.1 PURE INPUT EFFECTS UNDER CONTROLLED INTERVENTIONS

To determine which inputs a function cannot omit, we separate effects that require different input sets. Let

$$x = (x_1, \dots, x_n) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_n, \qquad P = \bigotimes_{i=1}^{n} P_i, \tag{1}$$

where each $\mathcal{X}_i$ is finite and each P_i has full support. The variables x_i are executable input choices, such as alternative tokens or controlled concept states. The product distribution specifies the intervention sampling scheme rather than the dependence structure of natural language.

Fix a rule for generating inputs, a model layer, a readout location, and a representation metric, and write $h(x) \in \mathbb{R}^d$. Under P, the functional ANOVA components are

$$h_\varnothing = \mathbb{E}h, \qquad h_T(x_T) = \mathbb{E}[h \mid x_T] - \sum_{U \subsetneq T} h_U(x_U), \tag{2}$$

with $h = \sum_{T \subseteq [n]} h_T$ and orthogonal components in $L^2(P)$. A nonempty component with $|T| = 1$ is a main effect, while $|T| \geq 2$ is a pure interaction. Most importantly, a nonzero pure component on T cannot be represented by a function that omits any variable in T. Thus every exact scope for such a component must contain all of T.

3.2 FEATURE SCOPES AND OMITTED DEPENDENCE

Fix a trained encoder, decoder, and bias and write

$$z(x) = \mathrm{Enc}(h(x)), \qquad \widehat{h}(x) = b + Dz(x), \qquad R = \mathbb{E}\|h - \widehat{h}\|_2^2, \tag{3}$$

where $D = [d_1, \dots, d_m]$. We remove zero decoder columns and normalize the remaining columns, rescaling the corresponding activations. A scope assignment $A = (A_1, \dots, A_m)$ specifies the input set $A_j \subseteq [n]$ available to feature j. It is exact when

$$z_j(x) = g_j(x_{A_j}). \tag{4}$$

Let $J_i = \{j : i \in A_j\}$. The total scope size and expected activation sparsity are

$$B(A) = \sum_{j=1}^{m} |A_j| = \sum_{i=1}^{n} |J_i|, \qquad K = \mathbb{E}\|z(x)\|_0. \tag{5}$$

The two quantities count different objects. K counts active features on an example, whereas $B(A)$ counts input–feature dependencies across interventions; repeated use of the same input by different features is charged repeatedly.

For approximate localization, the best L^2 approximation to z_j using only x_{A_j} is its conditional mean. Define

$$\bar{z}_j = \mathbb{E}[z_j \mid x_{A_j}], \qquad \Gamma_z(A) = \mathbb{E}\|z - \bar{z}\|_2^2, \qquad \Lambda_z(A) = \mathbb{E}\|D(z - \bar{z})\|_2^2. \tag{6}$$

Here Γ_z measures dependence omitted from the feature activations, while Λ_z measures the same omission after decoding and satisfies

$$\Lambda_z(A) \leq \|D\|_{\mathrm{op}}^2 \Gamma_z(A). \tag{7}$$

We use Γ_z for the localization curve because decoder columns can cancel omitted feature effects. Appendix F gives an exact ReLU example in which decoded omission vanishes despite nonzero feature-level omission.

3.3 THE OPTIMAL INPUT SCOPE CURVE

A large error for one proposed scope can reflect either broad dependence or simply the wrong choice of inputs. We therefore compare the best assignments under the same total dependency budget.

Definition 1 (Optimal input scope curve)**.** For an integer budget $B \geq 0$, define

$$\boxed{\mathcal{L}_z(B) = \min_{B(A) \leq B} \Gamma_z(A).} \tag{8}$$

Its inverse form is

$$\boxed{\mathcal{B}_z(\tau) = \min\{B : \mathcal{L}_z(B) \leq \tau\}}, \qquad \tau \geq 0. \tag{9}$$

Thus $\mathcal{L}_z(B)$ is the least feature dependence that must remain unexplained at budget B, while $\mathcal{B}_z(\tau)$ is the smallest budget achieving tolerance τ. The SAE is fixed throughout; only the scope assignment changes. Let $W_z = \sum_j \mathrm{Var}(z_j)$. Then $\mathcal{L}_z(0) = W_z$, $\mathcal{L}_z(nm) = 0$, and $\mathcal{L}_z(B)$ is nonincreasing. We use $\mathcal{L}_z(B)/W_z$ when comparing different activation scales.

4 EXACT SEPARATION WITH THE SAME FROZEN SAE

The scope curve distinguishes sparsity from localization by definition. We now ask whether localization can differ substantially when both the representation distribution and the SAE are unchanged.

Lemma 1 (Same representation distribution)**.** *If two input maps $h_1(x)$ and $h_2(x)$ induce the same representation distribution, then every population objective of the form $\mathcal{J}(\theta) = \mathbb{E}\,\ell(h(x), \theta)$ is identical for the two maps, including its local and global optima.*

Thus an objective based only on representation samples cannot distinguish such input maps. Objectives using input labels, intervention pairs, or relations between samples fall outside this comparison.

An invertible input mixing lets us preserve the representation distribution while changing which inputs determine each coordinate. Choose a prime $p > n + 1$ and sample x uniformly from $\mathbb{F}_p^n$. For each $1 \leq r \leq n$, choose an invertible matrix $M_r \in \mathbb{F}_p^{n \times n}$ with exactly r nonzero entries in each row and define

$$t = M_r x \pmod{p}, \qquad q_j^{(r)}(x) = \mathbf{1}[t_j = 0], \qquad \pi = \frac{1}{p}, \qquad v = \pi(1 - \pi), \tag{10}$$

with

$$h^{(r)}(x) = q^{(r)}(x) - \pi \mathbf{1}. \tag{11}$$

Use the same SAE for every r:

$$D = I_n, \qquad b = -\pi \mathbf{1}, \qquad \mathrm{Enc}(h) = \mathrm{ReLU}(h + \pi \mathbf{1}). \tag{12}$$

Theorem 1 (Exact separation for arbitrary interaction order)**.** *For every $1 \leq r \leq n$, the construction in equation 10–equation 12 has the following properties.*

1. *The representation distributions are identical for all r.*
2. *Every nonconstant component of $h^{(r)}$ is a pure effect of order r.*
3. *The same frozen SAE reconstructs exactly and has $z = q^{(r)}$, $R = 0$, $K = n\pi$, and $W_z = nv$ for every r.*
4. *For every integer budget $B \geq 0$,*

$$\boxed{\mathcal{L}_z^{(r)}(B) = v \left[n - \min\left\{ n, \left\lfloor \frac{B}{r} \right\rfloor \right\} \right].} \tag{13}$$

Consequently, for $0 \leq \tau \leq nv$,

$$\boxed{\mathcal{B}_z^{(r)}(\tau) = r\ \mathrm{ceil}\Big(n - \frac{\tau}{v}\Big) = r\, \mathcal{B}_z^{(1)}(\tau).} \tag{14}$$

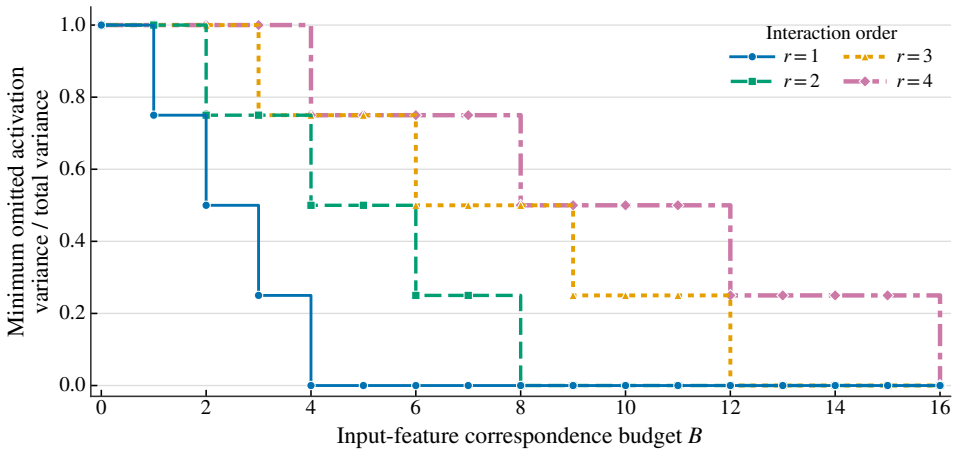

Figure 2: Optimal input scope curves for the finite construction with $n = 4$ and $p = 7$, computed over all $7^4 = 2401$ inputs and every integer budget. The vertical axis shows $\mathcal{L}_z(B)/W_z$.

Here $\mathrm{ceil}(u) = \min\{k \in \mathbb{Z} : k \geq u\}$. The separation has a simple mechanism. Invertibility makes $t = M_r x$ uniform, so the representation and activation distributions are unchanged across r. But each coordinate depends on exactly r inputs, and omitting any one of them returns its conditional mean to π. Retaining one feature's nonconstant variation therefore costs r units of scope, producing the staircase in Figure 2 and the exact factor of r in equation 14.

For example, with $n = 10$ and $p = 101$, every order has $K \approx 0.099$. At budget $B = 10$, order one has zero omission while order five retains at least 80% relative omission. Appendix B constructs the required matrices and proves the result. The construction fixes a dictionary. We next ask whether the same dictionary is uniquely selected in a corresponding exact sparse coding problem.

Proposition 1 (Unique optimum for exact sparse coding at fixed width)**.** *Let $h = q - \pi\mathbf{1}$ with independent $q_i \sim \mathrm{Bernoulli}(\pi)$ and $0 < \pi < 1/2$. Among representations of width n with decoder columns of unit norm, nonnegative codes, free bias, and exact reconstruction, the minimum of $\mathbb{E}\|z\|_1$ is $n\pi$. Up to a column permutation, the unique optimal decoder is I_n, the code equals q on the data support, and the bias is $-\pi\mathbf{1}$.*

Thus the separation does not rely on an arbitrary or suboptimal dictionary in this exact fixed-width problem. The sparse objective selects the same coordinates uniquely up to permutation, while their input scopes still depend on the intervention map. This isolates the exact separation from dictionary ambiguity. Section 7 asks whether the same distinction persists beyond this idealized regime in trained overcomplete SAEs. Appendix C gives the proof and a near-optimal stability result; Appendix B also treats representation redesign under a scope budget. The exact family shows how large the separation can be, but its all-or-nothing feature costs do not explain what can force broad scopes in a general representation. We turn to that question next.

5 General Lower Constraints on Feature Scopes

For a general representation, a joint input effect can be carried only by features whose scopes contain the required variables and whose decoder directions can represent the effect. We first keep the identity of each interacting set, then aggregate these requirements into a total scope lower bound.

5.1 Joint Effects Require Shared Features

A pure effect involving every variable in T cannot be assembled from scalar feature functions that each omit some variable in T. For nonempty T, define the compatible features and their decoder span by

$$I_T = \{j : T \subseteq A_j\} = \bigcap_{i \in T} J_i, \qquad V_T = \mathrm{span}\{d_j : j \in I_T\}. \tag{15}$$

The localized reconstruction can represent its pure component on T only inside V_T. Let Π_T be the orthogonal projector onto V_T, with $\Pi_T = 0$ when I_T is empty, and define

$$Q(D, A) = \sum_{T \neq \varnothing} \mathbb{E}\|(I - \Pi_T)h_T\|_2^2. \tag{16}$$

Thus $Q(D, A)$ is the interaction energy unreachable by features with compatible scopes and decoder directions. Decoder alignment alone is insufficient: two scalar features that separately depend on inputs 1 and 2 cannot produce a pure interaction between them.

Theorem 2 (Compatibility of joint effects and feature scopes)**.** *For any input function, frozen SAE, and scope assignment in Section 3,*

$$\boxed{\sqrt{R} + \sqrt{\Lambda_z(A)} \geq \sqrt{Q(D, A)}.} \tag{17}$$

Hence

$$\Lambda_z(A) \geq [\sqrt{Q(D, A)} - \sqrt{R}]_+^2, \qquad \Gamma_z(A) \geq \frac{[\sqrt{Q(D, A)} - \sqrt{R}]_+^2}{\|D\|_{\mathrm{op}}^2}. \tag{18}$$

If the scope declaration is exact, then $R \geq Q(D, A)$.

The bound follows because the localized reconstruction has its T component in V_T, while

$$h - \bar{h} = (h - \widehat{h}) + D(z - \bar{z}).$$

Thus incompatible interaction energy must appear as reconstruction error, decoded scope omission, or both. Appendix D gives the proof, tightness of the square-root form, and a relaxed construction attaining $Q(D, A)$.

Because $Q(D, A)$ retains the identity of every interacting set, it can be difficult to turn into a direct total-budget statement. We next give up that identity and aggregate by interaction order.

5.2 Interaction order and representation directions

Interaction order alone does not determine total scope cost. A function $h(x) = d\, g(x)$ can contain arbitrarily high-order interactions while all effects remain collinear, so one sufficiently broad feature may suffice. The relevant constraints are therefore both how broad a feature must be and how many decoder directions are needed.

For each threshold r, collect the effects of order at least r:

$$h_{\geq r} = \sum_{|T| \geq r} h_T, \qquad C_r = \mathbb{E}[h_{\geq r} h_{\geq r}^{\mathsf{T}}], \tag{19}$$

with eigenvalues $\lambda_{r,1} \geq \cdots \geq \lambda_{r,d} \geq 0$. Let

$$m_r(A) = \#\{j : |A_j| \geq r\} \tag{20}$$

be the number of features whose scopes are broad enough to carry such effects. These features span at most $m_r(A)$ decoder directions, so any remaining spectral mass must appear as approximation error.

Theorem 3 (Spectral budget bound)**.** *For a scope assignment A, write $\Lambda = \Lambda_z(A)$. For every $1 \leq r \leq n$,*

$$\boxed{(\sqrt{R} + \sqrt{\Lambda})^2 \geq \sum_{k > m_r(A)} \lambda_{r,k} \geq \sum_{k > \lfloor B(A)/r \rfloor} \lambda_{r,k}.} \tag{21}$$

Define

$$N_r(t) = \min \left\{ q \in \{0, \ldots, d\} : \sum_{k > q} \lambda_{r,k} \leq t \right\}. \tag{22}$$

Then every scope assignment satisfies

$$B(A) \geq \sum_{r=1}^{n} N_r\big((\sqrt{R} + \sqrt{\Lambda})^2\big). \tag{23}$$

Consequently,

$$\boxed{\mathcal{B}_z(\tau) \geq \sum_{r=1}^{n} N_r\Big((\sqrt{R} + \|D\|_{\mathrm{op}}\sqrt{\tau})^2\Big)\,.} \tag{24}$$

For an assignment attaining $\mathcal{B}_z(\tau)$, $\Gamma_z(A) \leq \tau$, so equation 7 gives $\Lambda \leq \|D\|_{\mathrm{op}}^2\tau$, yielding equation 24.

The aggregation uses the exact counting identity

$$B(A) = \sum_{r=1}^{n} m_r(A),$$

because a feature with scope size s is counted once at every threshold $r \leq s$. The bound therefore combines a breadth requirement from interaction order with a multiplicity requirement from directional spread. Appendix E gives the proof and a corresponding bound for a specified input set.

When $R = 0$ and exact localization is required,

$$\mathcal{B}_z(0) \geq \sum_{r=1}^{n} \mathrm{rank}(C_r). \tag{25}$$

For the exact order-r family in Section 4, $C_s = vI_n$ for $s \leq r$ and $C_s = 0$ for $s > r$, so the bound is tight: $\sum_s \mathrm{rank}(C_s) = nr = \mathcal{B}_z^{(r)}(0)$. In general, aggregating by order discards the identities of interacting sets and can be loose. Section 7 measures this gap in learned SAEs, while Appendix H shows that the interaction covariance is stable under small reconstruction error.

6 Stability and finite-sample measurement

The exact separation would be less informative if small changes to aligned feature coordinates could erase it, and the scope curve would be difficult to use if only its population definition were available. We address these two questions in turn.

6.1 Stability under changes to the encoding

Proposition 2 (Stability of the scope curve)**.** *Let z and $\widetilde{z}$ be aligned encodings on the same input distribution, with the same number and normalization of feature coordinates. If $\nu = \|z - \widetilde{z}\|_{L^2}$, then for every budget B,*

$$\boxed{\Big|\sqrt{\mathcal{L}_z(B)} - \sqrt{\mathcal{L}_{\widetilde{z}}(B)}\Big| \leq \nu.} \tag{26}$$

Conditional expectation is an orthogonal projection for each coordinate, so its complement is a contraction and the bound survives minimization over scope assignments. Applied to Theorem 1,

$$\mathcal{L}_{\widetilde{z}}(B) \geq \Big[\sqrt{v\,(n - \min\{n, \lfloor B/r \rfloor\})} - \nu\Big]_+^2. \tag{27}$$

Thus the separation cannot collapse under an arbitrarily small change in aligned feature functions. The proposition does not assert invariance to feature splitting, rotations, or arbitrary reparameterizations. Appendix C connects the bound to near-optimal solutions of the exact sparse objective.

6.2 Measuring localization from controlled interventions

For a candidate scope A of feature j, independently sample $x, x' \sim P$ and form $y = (x_A, x'_{-A})$. Then

$$\boxed{\mathbb{E}\left[(z_j - \mathbb{E}[z_j \mid x_A])^2\right] = \frac{1}{2}\mathbb{E}\left[(z_j(x) - z_j(y))^2\right].} \tag{28}$$

Conditioned on x_A, the two activations on the right are independent draws from the same conditional distribution. Repeated paired interventions therefore estimate the omission of a fixed candidate scope without fitting the conditional mean.

Once candidate errors are available, the remaining problem is scope selection. If $z_{j,T}$ is the ANOVA component of feature j and $w_{j,T} = \mathbb{E}[z_{j,T}^2]$, then

$$e_j(A) = \sum_{T \not\subseteq A} w_{j,T}, \qquad \Gamma_z(A_1, \ldots, A_m) = \sum_j e_j(A_j). \tag{29}$$

For small n, enumerating the 2^n candidate scopes for each feature and allocating the total budget by dynamic programming gives the exact curve. For larger designs, the same estimation and selection machinery applies to a prespecified candidate family and yields its restricted curve, with simultaneous control determined by the family size. Appendix G gives the recurrence and certified optimization bounds for restricted candidate families.

Estimating a fixed scope does not by itself control the curve selected from many candidates. Let $\widehat{e}_j(A)$ estimate every candidate error and define

$$\widehat{\mathcal{L}}_z(B) = \min_{B(A) \le B} \sum_j \widehat{e}_j(A_j). \tag{30}$$

We therefore require simultaneous control before minimizing.

Proposition 3 (Finite-sample control of the selected curve)**.** *If $|\widehat{e}_j(A) - e_j(A)| \le \epsilon_j$ simultaneously for every feature j and every candidate scope A, then*

$$\sup_B \left| \widehat{\mathcal{L}}_z(B) - \mathcal{L}_z(B) \right| \le \sum_j \epsilon_j. \tag{31}$$

If a deterministic activation range width bound ρ_j is available for feature j and N independent paired samples are used per candidate scope, then with probability at least $1 - \delta$ the bound holds with

$$\epsilon_j = \rho_j^2 \sqrt{\frac{\log(2m2^n/\delta)}{8N}}. \tag{32}$$

The bound is intentionally conservative but remains valid after scope selection. Appendix G gives the proof and the corresponding statement for any fixed restricted candidate family. We next test whether the separation and the finite-sample measurement remain visible when the representation and SAE are learned rather than stipulated.

7 Controlled validation with learned SAEs

The experiments provide a controlled learned realization in which representation evidence is exactly matched across intervention maps. This isolates input correspondence and lets us test three consequences of the theory: whether localization can differ under identical learned representation samples, whether paired interventions recover the optimized scope curve from finite data, and how informative the general spectral certificate is outside the exact construction.

Localization differs under identical learned representation samples. Each seed uses one trained representation network and one trained SAE, while three bijective input maps permute the same complete $2^8 = 256$ state factorial grid. The representation and SAE activation samples are therefore identical within a seed; only their correspondence to the controlled inputs changes.

For each of five seeds, an $8 \to 24 \to 8$ ReLU network is trained to reconstruct eight binary factors, and its 24-dimensional hidden state defines h. An overcomplete $24 \to 48 \to 24$ ReLU SAE is then trained on 192 states, with 64 states held out from both training procedures. After freezing both models, we compare the identity map with cyclic XOR maps involving three or five neighboring factors. Scope information is never used during training, and the learned features need not be pure interactions of the map arity. Appendix J gives the full protocol.

We enumerate all 256 candidate scopes for each frozen feature and compute the exact population curve. The budget required to reduce omitted activation variance below $0.1W_z$ rises from 136.0 ± 22.6 for the identity map to 192.4 ± 11.6 and 202.8 ± 8.6 for the three- and five-factor maps. Both mixed maps require a larger budget in every seed. Across the five learned SAEs, relative reconstruction

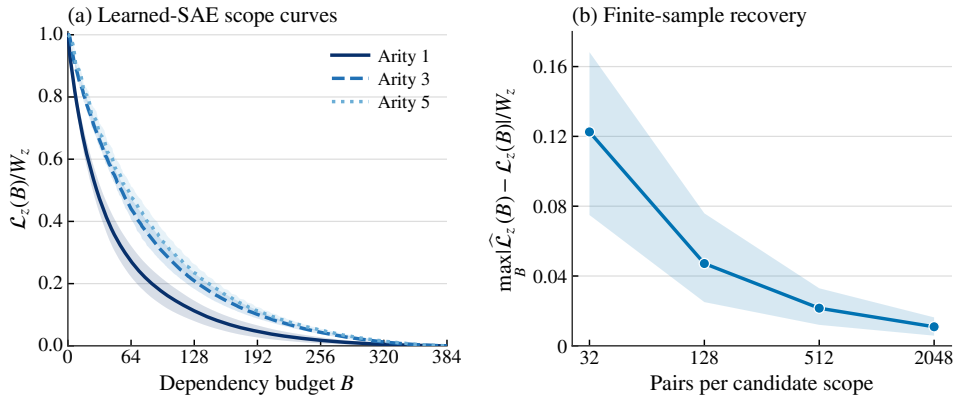

Figure 3: Controlled validation with learned overcomplete SAEs. (a) Mean optimal scope curves across five independently trained representation–SAE pairs; shading shows one standard deviation. Within each seed, the three bijective input maps induce exactly the same representation and activation samples. (b) Finite paired resampling approaches the exact population curve. Points average the maximum normalized curve error over five seeds, three maps, and ten repetitions; shading spans the 10th–90th percentiles. The horizontal sample count is the number of pairs per candidate scope.

error is $0.246\% \pm 0.062\%$ on the full intervention grid and $0.639\% \pm 0.170\%$ on held-out states, with 14.41 ± 1.69 active features per state. Figure 3a shows the population curves.

Finite paired samples recover the optimized curve. The second experiment asks whether the paired estimator remains accurate after candidate scopes are reselected from noisy estimates. For every seed, map, and candidate scope, we estimate omission from N paired interventions and re-optimize the full curve. Averaged over five seeds, three maps, and ten repetitions, the maximum normalized curve error falls from 0.123 at $N = 32$ to 0.047, 0.0216, and 0.0109 at $N = 128, 512, 2048$. The mean absolute error in $\mathcal{B}_z(0.1W_z)$ falls from 27.6 to 6.0, 1.83, and 0.65 budget units. Figure 3b shows the curve error; Appendix J reports the full sampling audit.

Certificate strength reflects available information. We finally compare the exact budgets with the spectral lower bound in equation 24. At tolerance $0.1W_z$, the mean lower bounds are 4.2 ± 0.4, 12.6 ± 1.3, and 20.6 ± 1.9 for map arities one, three, and five, far below the exact budgets. The representation-level certificate is coarse by construction because it uses order-wise interaction covariance while discarding interacting-set identities and feature-level ANOVA structure. It is nonzero and increases with the imposed input mixing, but remains a weak constraint on the exact budgets. Appendix J reports per-seed compatibility diagnostics and a more information-rich feature-order relaxation. Its tighter values illustrate the tradeoff between the information used by a certificate and the localization constraint it can provide. These experiments validate the distinction in a controlled factorial setting; they do not estimate the prevalence of broad scopes in current language-model SAEs.

8 Discussion and Conclusion

Sparse features are often interpreted by connecting individual activations to recognizable inputs, but reconstruction and sparsity alone do not determine that connection. Our results show that input localization should instead be treated as a relation between a fixed feature system and a declared intervention design. This distinction matters for SAE evaluation because a feature can appear sparse and stable while still requiring broad input context to explain its activation.

The practical implication is that claims about what drives an SAE feature require evidence at the input–feature level, not only representation-level quality metrics. Controlled interventions provide one way to supply that evidence and to compare feature systems under a common experimental design. The scope curve formalizes this comparison, complementing existing evaluations of reconstruction, sparsity, feature quality, and downstream behavior.

AI USE STATEMENT

In this work, we used generative AI tools to improve the readability of the manuscript, to check the correctness and internal consistency of mathematical derivations and proofs, and to assist with experimental plotting.

REPRODUCIBILITY STATEMENT

The mathematical assumptions, constructions, and proofs are given in the main text and appendices. The controlled validation specifies the intervention grid, model and SAE architectures, training and held out splits, optimization settings, random seeds, candidate scopes, paired sample sizes, repetition counts, and numerical checks used in the reported experiments.

REFERENCES

Siyu Chen, Heejune Sheen, Xuyuan Xiong, Tianhao Wang, and Zhuoran Yang. Taming polysemanticity in LLMs: Theory-grounded feature recovery via sparse autoencoders. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/ac4543f52e9de25f19e925a9f3e4fea6-Abstract-Conference.html.

Jingyi Cui, Qi Zhang, Yifei Wang, and Yisen Wang. On the limits of sparse autoencoders: A theoretical framework and reweighted remedy. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/fc6bd0eef19459655d5b097af783661d-Abstract-Conference.html.

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

William Dorrell. How optimality structures sparse dictionaries: A theory for understanding SAE representations. *arXiv preprint arXiv:2606.02385*, 2026. URL https://arxiv.org/abs/2606.02385.

Alyson K. Fletcher, Sundeep Rangan, Vivek K. Goyal, and Kannan Ramchandran. Denoising by sparse approximation: Error bounds based on rate-distortion theory. *EURASIP Journal on Advances in Signal Processing*, 2006:26318, 2006. doi: 10.1155/ASP/2006/26318. URL https://doi.org/10.1155/ASP/2006/26318.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/42ef3308c230942d223c411adf182c88-Abstract-Conference.html.

Gleb Gerasimov, Timofei Rusalev, Nikita Balagansky, Daniil Laptev, Vadim Kurochkin, and Daniil Gavrilov. Unstable features, reproducible subspaces: Understanding seed dependence in sparse autoencoders. *arXiv preprint arXiv:2606.12138*, 2026. URL https://arxiv.org/abs/2606.12138.

Rémi Gribonval and Karin Schnass. Dictionary identification—sparse matrix-factorization via ℓ_1-minimization. *IEEE Transactions on Information Theory*, 56(7):3523–3539, 2010. doi: 10.1109/TIT.2010.2048466.

Sangyu Han, Yearim Kim, and Nojun Kwak. Causal interpretation of sparse autoencoder features in vision. *arXiv preprint arXiv:2509.00749*, 2025. URL https://arxiv.org/abs/2509.00749.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_fi

les/paper/2024/hash/1fa1ab11f4bd5f94b2ec20e794dbfa3b-Abstract-Conference.html.

Alexandre Janon, Thierry Klein, Agnès Lagnoux, Maëlle Nodet, and Clémentine Prieur. Asymptotic normality and efficiency of two Sobol index estimators. *ESAIM: Probability and Statistics*, 18: 342–364, 2014. doi: 10.1051/ps/2013040. URL https://www.numdam.org/articles/10.1051/ps/2013040/.

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

Benjamin Lengerich, Sarah Tan, Chun-Hao Chang, Giles Hooker, and Rich Caruana. Purifying interaction effects with the functional ANOVA: An efficient algorithm for recovering identifiable additive models. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2402–2412. PMLR, 2020. URL https://proceedings.mlr.press/v108/lengerich20a.html.

Art B. Owen. Effective dimension of some weighted pre-sobolev spaces with dominating mixed partial derivatives. *SIAM Journal on Numerical Analysis*, 57(2):547–562, 2019. doi: 10.1137/17M1158975. URL https://doi.org/10.1137/17M1158975.

Gonçalo Paulo and Nora Belrose. Sparse autoencoders trained on the same data learn different features. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/3c1fe56b043848b211030c202764c6a7-Abstract-Conference.html.

Yusung Ro, Jaehyun Choi, and Junmo Kim. Beyond semantics: Disentangling information scope in sparse autoencoders for CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 3252–3261, 2026. URL https://arxiv.org/abs/2604.05724.

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 *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2172–2210, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.99. URL https://aclanthology.org/2026.acl-long.99/.

A Functional ANOVA and conditional projection

For any square integrable scalar or vector valued function, equation 2 is equivalent to

$$h_T = \sum_{U \subseteq T} (-1)^{|T|-|U|} \mathbb{E}[h \mid x_U]. \tag{33}$$

Because P is a product distribution, expectation operators over distinct input variables commute. If $i \in T$, averaging equation 33 over x_i cancels terms in pairs, so $\mathbb{E}_{x_i} h_T = 0$. If $T \neq U$, averaging first over a variable in their symmetric difference gives

$$\mathbb{E}[h_T h_U^{\mathsf{T}}] = 0. \tag{34}$$

Hence

$$\|h\|_{L^2}^2 = \sum_T \|h_T\|_{L^2}^2. \tag{35}$$

If a scalar function f depends only on x_A, then $f_T = 0$ whenever $T \nsubseteq A$. Conversely,

$$\mathbb{E}[f \mid x_A] = \sum_{T \subseteq A} f_T, \tag{36}$$

and therefore

$$\mathbb{E}\left[(f - \mathbb{E}[f \mid x_A])^2\right] = \sum_{T \nsubseteq A} \mathbb{E}[f_T^2]. \tag{37}$$

Taking conditional expectations coordinate by coordinate is therefore an orthogonal projection on $L^2(P; \mathbb{R}^m)$, a fact used in Proposition 2.

B Proof of the exact separation for arbitrary interaction order

B.1 Existence of the matrices

For $r = 1$, take $M_1 = I_n$. For $r > 1$, choose a binary matrix C with zero diagonal and exactly $r - 1$ nonzero entries in each row, for example by selecting the next $r - 1$ columns cyclically. The polynomial $\det(I + tC)$ has constant term one and degree at most n. It therefore has at most n roots in $\mathbb{F}_p$. Since $p - 1 > n$, there exists a nonzero $t \in \mathbb{F}_p$ that is not a root. Then

$$M_r = I + tC \tag{38}$$

is invertible and has exactly r nonzero entries in every row.

B.2 Identical representation distributions

An invertible linear map is a bijection on $\mathbb{F}_p^n$, so $t = M_r x$ is uniform whenever x is uniform. Its coordinates are therefore independent and uniform on $\mathbb{F}_p$. Consequently, each $q_j^{(r)} = \mathbf{1}[t_j = 0]$ is independent Bernoulli$(1/p)$, independent of r. Thus $q^{(r)}$ and $h^{(r)}$ have the same joint distributions for all interaction orders.

B.3 Pure interaction order

Let S_j be the support of row j of M_r, with $|S_j| = r$. The function $q_j^{(r)} - \pi$ depends only on x_{S_j}. If any variable x_i with $i \in S_j$ is averaged out while the other inputs are held fixed, the modular sum remains uniform because the coefficient of x_i is nonzero. Hence

$$\mathbb{E}_{x_i}[q_j^{(r)} - \pi] = 0 \tag{39}$$

for every $i \in S_j$. By the ANOVA characterization in Appendix A, the coordinate is a pure component on S_j and has no nonconstant components of lower order.

B.4 Exact SAE and optimal scope curve

On the data support, $h_j + \pi = q_j \in \{0, 1\}$, so equation 12 gives $z = q$ and exact reconstruction. The expected number of nonzero activations is $K = n\pi$ and the total activation variance is $W_z = nv$.

Fix a declared scope A_j. If $S_j \subseteq A_j$, then q_j is measurable with respect to x_{A_j} and

$$\mathbb{E}[q_j \mid x_{A_j}] = q_j. \tag{40}$$

If $S_j \not\subseteq A_j$, at least one nonzero coefficient in the modular sum is still averaged out, so

$$\mathbb{E}[q_j \mid x_{A_j}] = \pi. \tag{41}$$

Therefore

$$\mathbb{E}\left[\left(q_j - \mathbb{E}[q_j \mid x_{A_j}]\right)^2\right] = \begin{cases} 0, & S_j \subseteq A_j, \\ v, & S_j \not\subseteq A_j. \end{cases} \tag{42}$$

Exactly localizing one feature requires a scope containing r inputs. Under budget B, at most $\min\{n, \lfloor B/r \rfloor\}$ features can be exactly localized, and choosing any such set achieves the bound. This proves equation 13; solving for the smallest budget that gives error at most τ proves equation 14.

B.5 Allowing the representation to change

One can instead allow a new representation with exact scopes under budget B and minimize reconstruction error. Denote this optimum by $R_r^\star(B)$. The same construction satisfies

$$R_r^\star(B) = v\left[n - \min\{n, \lfloor B/r \rfloor\}\right]. \tag{43}$$

The lower bound follows from Theorem 3 with $\Lambda = 0$. At most $\lfloor B/r \rfloor$ features can have scope size at least r, while the covariance at order r is vI_n. For achievability, keep any $k = \min\{n, \lfloor B/r \rfloor\}$ coordinates of the identity decoder and predict zero in the rest after bias adjustment. Equation 43 happens to match the frozen curve numerically in this construction, but it optimizes a different object.

C Sparse optimality and stability near the optimum

C.1 Proof of Proposition 1

The support of q contains all 2^n binary vertices with positive probability. An exact decoder at width n must therefore be invertible; otherwise it cannot reconstruct all vertex differences. Write $W = D^{-1}$. On the data support, every exact encoding has the form

$$z = Wq + c, \qquad c = -W(\pi\mathbf{1} + b). \tag{44}$$

Nonnegativity on all vertices implies

$$c_j \geq \sum_i (-W_{ji})_+. \tag{45}$$

Let

$$S = \sum_{j,i} |W_{ji}|, \qquad N = \sum_{j,i} (-W_{ji})_+, \qquad U = \sum_j \left(c_j - \sum_i (-W_{ji})_+\right) \geq 0. \tag{46}$$

A direct expectation over independent Bernoulli coordinates gives

$$\mathbb{E}\|z\|_1 = \pi S + (1 - 2\pi)N + U. \tag{47}$$

Since $e_i = DW_{:i}$ and every decoder column has unit norm,

$$1 = \|e_i\|_2 \leq \sum_j |W_{ji}|, \tag{48}$$

so $S \geq n$. Because $\pi < 1/2$, equation 47 is at least $n\pi$, attained by $D = I_n$, $z = q$, and $b = -\pi\mathbf{1}$.

If equality holds, then $N = U = 0$ and $S = n$. Thus $W \geq 0$, and each column of W has ℓ_1 norm one. Equality in

$$e_i = \sum_j W_{ji} d_j \tag{49}$$

with nonnegative weights and unit decoder columns forces every used decoder column to point along e_i. There are exactly n columns, so each standard basis direction receives exactly one column. Hence D is a column permutation of I_n, $c = 0$, and the code and bias are the corresponding permutations of q and $-\pi\mathbf{1}$.

C.2 Stability under a small objective gap

Suppose the same exact reconstruction problem at fixed width satisfies

$$\mathbb{E}\|z\|_1 \leq n\pi + \varepsilon, \qquad \varepsilon \leq \min\{\pi/16, (1-2\pi)/2\}. \tag{50}$$

Equation 47 implies

$$S - n \leq \frac{\varepsilon}{\pi}, \qquad N \leq \frac{\varepsilon}{1-2\pi} \leq \frac{1}{2}. \tag{51}$$

For column i of W, write

$$s_i = \sum_j |W_{ji}|, \qquad p_i = \sum_j (W_{ji})_+. \tag{52}$$

Then $s_i \geq 1$ and $p_i \geq 1 - N \geq 1/2$. The exact identity

$$\sum_j |W_{ji}| \| \operatorname{sign}(W_{ji}) d_j - e_i\|_2^2 = 2(s_i - 1) \tag{53}$$

implies that some positively weighted column $d_{j(i)}$ satisfies

$$\|d_{j(i)} - e_i\|_2^2 \leq 4(s_i - 1). \tag{54}$$

Under equation 50, each selected column lies within $1/2$ of its standard basis vector. One decoder column cannot lie in the balls of radius $1/2$ around two distinct standard basis vectors, whose mutual distance is $\sqrt{2}$. The assignment $i \mapsto j(i)$ is therefore a permutation P, and

$$\boxed{\|DP - I\|_F \leq 2\sqrt{\varepsilon/\pi}.} \tag{55}$$

Set $D' = DP$, $z' = P^\mathsf{T} z$, and $\delta = 2\sqrt{\varepsilon/\pi} \leq 1/2$. Then

$$\|(D')^{-1} - I\|_{\mathrm{op}} \leq \frac{\delta}{1-\delta}. \tag{56}$$

Using equation 44 and equation 47, the aligned code satisfies

$$\nu = \|z' - q\|_{L^2} \leq \frac{\delta}{1-\delta}\sqrt{n\pi} + \frac{\varepsilon}{1-2\pi} + \varepsilon. \tag{57}$$

Combining equation 57 with Proposition 2 converts a small objective gap in this sparse objective into a lower bound on the scope curve.

D Proof of the compatibility theorem and tightness

For a fixed assignment A, let $\bar{h} = b + D\bar{z}$. Because $\bar{z}_j$ depends only on x_{A_j}, its pure ANOVA component on T is zero unless $T \subseteq A_j$. Therefore $\bar{h}_T \in V_T$. Orthogonal projection gives

$$\mathbb{E}\|h_T - \bar{h}_T\|_2^2 \geq \mathbb{E}\|(I - \Pi_T)h_T\|_2^2. \tag{58}$$

Summing over all nonempty T and using orthogonality yields

$$\|h - \bar{h}\|_{L^2}^2 \geq Q(D, A). \tag{59}$$

On the other hand,

$$h - \bar{h} = (h - \widehat{h}) + D(z - \bar{z}), \tag{60}$$

so

$$\|h - \bar{h}\|_{L^2} \leq \sqrt{R} + \sqrt{\Lambda_z(A)}. \tag{61}$$

Combining this inequality with equation 59 proves equation 17. Rearrangement and equation 7 give equation 18.

D.1 TIGHTNESS OF THE SQUARE ROOT BOUND

Take the exact family with $h = q - \pi\mathbf{1}$, declare every scope empty, and define

$$z_\alpha(h) = \mathrm{ReLU}(\alpha h + \pi\mathbf{1}), \qquad D = I_n, \qquad b = -\pi\mathbf{1}, \qquad 0 \le \alpha \le 1. \tag{62}$$

On the data support, $\widehat{h} = \alpha h$. Hence

$$R = (1-\alpha)^2 nv, \qquad \Lambda = \Gamma = \alpha^2 nv, \qquad Q = nv. \tag{63}$$

Therefore

$$(\sqrt{R} + \sqrt{\Lambda})^2 = Q \tag{64}$$

for every α. The constant and the square root form in Theorem 2 cannot be uniformly improved without additional assumptions.

D.2 RELAXED ACHIEVABILITY OF Q

For fixed decoder and scopes, set $p_T = \Pi_T h_T$. Choose a fixed linear right inverse of the decoder submatrix indexed by I_T and write

$$p_T = \sum_{j \in I_T} d_j a_{j,T}(x_T). \tag{65}$$

Linear transformations preserve the pure component on T. Define

$$f_j = \sum_{\varnothing \neq T \subseteq A_j} a_{j,T}. \tag{66}$$

Then f_j depends only on x_{A_j} and the reconstruction

$$\mathbb{E}h + \sum_j d_j f_j \tag{67}$$

has squared error exactly $Q(D, A)$. On finite input spaces, a constant shift makes every f_j nonnegative and can be absorbed by the bias. This construction allows arbitrary coefficient functions of x_{A_j} and therefore does not address whether a restricted encoder can compute them from h alone when h is not injective.

E PROOF OF THE SPECTRAL BUDGET BOUND

Only features with $|A_j| \ge r$ can contribute to ANOVA components of order at least r. Therefore the part of the localized reconstruction with order at least r, $\bar{h}_{\ge r}$, lies in the span of at most $m_r(A)$ fixed decoder columns. The best approximation to a random vector with covariance C_r by a linear subspace of dimension $m_r(A)$ has mean squared error at least the spectral tail

$$\sum_{k > m_r(A)} \lambda_{r,k}. \tag{68}$$

Projection onto the ANOVA components of order at least r cannot increase $\|h - \bar{h}\|_{L^2}$, while the compatibility proof gives

$$\|h - \bar{h}\|_{L^2} \le \sqrt{R} + \sqrt{\Lambda}. \tag{69}$$

This proves the first inequality in equation 21. Since every feature counted by $m_r(A)$ has scope size at least r,

$$r\, m_r(A) \le B(A), \tag{70}$$

which proves the second.

For a fixed threshold $t = (\sqrt{R} + \sqrt{\Lambda})^2$, the first inequality implies $m_r(A) \ge N_r(t)$. The counting identity

$$B(A) = \sum_j |A_j| = \sum_j \sum_{r=1}^{n} \mathbf{1}[|A_j| \ge r] = \sum_{r=1}^{n} m_r(A) \tag{71}$$

therefore gives equation 23. Any assignment attaining $\mathcal{B}_z(\tau)$ has $\Gamma \le \tau$ and, by equation 7,

$$\Lambda \le \|D\|_{\mathrm{op}}^2 \tau. \tag{72}$$

Substitution into equation 23 proves equation 24.

E.1 A BOUND FOR A SPECIFIED INPUT SET

For nonempty U, define

$$C_U^{\text{joint}} = \sum_{T \supseteq U} \mathbb{E}[h_T h_T^\intercal], \qquad q_U = \left| \bigcap_{i \in U} J_i \right|. \tag{73}$$

Because $V_T \subseteq V_U$ when $T \supseteq U$, the full certificate satisfies

$$(\sqrt{R} + \sqrt{\Lambda})^2 \geq Q(D, A) \geq \sum_{k > q_U} \lambda_k(C_U^{\text{joint}}). \tag{74}$$

Thus effects containing U require enough features whose scopes contain all inputs in U.

F CANCELLATION AND WHY Γ DEFINES THE MAIN CURVE

Let x be uniform on $\{-1, +1\}$ and $h = x$. Use three identical feature activations

$$z_j = \text{ReLU}(1 + h) = 1 + x, \tag{75}$$

with decoder $D = (1, -1, 1)$ and bias $b = -1$. This is an exact SAE with decoder columns of unit norm.

Declare $A_1 = A_2 = \varnothing$ and $A_3 = \{1\}$. The omitted encoding is $(x, x, 0)$, whose decoded contribution cancels, so $(\Lambda, \Gamma) = (0, 2)$. Now enlarge only A_1 to $\{1\}$. The omitted encoding becomes $(0, x, 0)$, giving $(\Lambda, \Gamma) = (1, 1)$. The larger scope reduces omission in the individual feature activations but destroys decoder cancellation. Hence Γ is monotone under scope expansion, while Λ need not be.

G MEASUREMENT AND OPTIMIZATION DETAILS

G.1 EXACT DYNAMIC PROGRAMMING FOR SMALL INPUT SPACES

Equation 37 gives equation 29. Define

$$g_j(k) = \min_{|A| \leq k} e_j(A), \qquad k = 0, \ldots, n. \tag{76}$$

Then

$$\boxed{\mathcal{L}_z(B) = \min_{\sum_j k_j \leq B} \sum_j g_j(k_j).} \tag{77}$$

For one direction, any feasible scope assignment induces $k_j = |A_j|$, so the right side is no larger. For the other, choose a scope attaining each $g_j(k_j)$; its true cost is at most k_j, so it is feasible. After enumerating 2^n scopes per feature, allocating the budget across features costs $O(mnB)$ with a direct dynamic program.

For a pure interaction, adding one input can leave the error unchanged until every required input is included. Greedy allocation is therefore not generally optimal.

G.2 CERTIFIED LOWER AND UPPER BOUNDS

For every $\lambda \geq 0$,

$$\sum_j \min_A [e_j(A) + \lambda |A|] - \lambda B \leq \mathcal{L}_z(B). \tag{78}$$

Indeed, for any feasible assignment the left side is at most

$$\Gamma_z(A) + \lambda(B(A) - B) \leq \Gamma_z(A). \tag{79}$$

Optimizing over λ can tighten the lower bound but does not guarantee zero discrete duality gap. Any feasible scope assignment supplies an upper bound, so their gap reports optimization uncertainty.

G.3 PAIRED CONDITIONAL RESAMPLING

Let $y = (x_A, x'_{-A})$ with independent $x, x' \sim P$. Conditioned on x_A, the variables $z_j(x)$ and $z_j(y)$ are independent and identically distributed. Therefore

$$\frac{1}{2}\mathbb{E}[(z_j(x) - z_j(y))^2 \mid x_A] = \mathrm{Var}(z_j \mid x_A). \tag{80}$$

Taking expectation proves equation 28. The identity also holds for nonuniform product marginals, but the resampling must follow those marginals.

G.4 FINITE-SAMPLE CONTROL AFTER SCOPE SELECTION

For any assignment $A = (A_1, \ldots, A_m)$, simultaneous candidate-scope errors satisfying $|\widehat{e}_j(A_j) - e_j(A_j)| \le \epsilon_j$ imply

$$\left| \sum_j \widehat{e}_j(A_j) - \sum_j e_j(A_j) \right| \le \sum_j \epsilon_j.$$

Apply this inequality first to an optimizer of the population objective and then to an optimizer of the empirical objective. The two directions give equation 31 simultaneously for every budget B.

For the paired estimator, suppose z_j has range width at most ρ_j. Each summand $Y = (z_j(x) - z_j(y))^2/2$ lies in $[0, \rho_j^2/2]$. Hoeffding's inequality gives

$$\Pr\left(|\widehat{e}_j(A) - e_j(A)| > \epsilon_j\right) \le 2\exp\left(-\frac{8N\epsilon_j^2}{\rho_j^4}\right).$$

There are $m\,2^n$ feature-scope candidates when every subset is considered. A union bound with the choice in equation 32 therefore establishes the simultaneous premise of Proposition 3 with probability at least $1 - \delta$. The same proof replaces $m2^n$ by the total size of any fixed restricted candidate family.

H STABILITY OF THE INTERACTION COVARIANCE

Apply the same ANOVA decomposition to $\widehat{h}$ and define

$$\widehat{C}_r = \mathbb{E}[\widehat{h}_{\ge r}\widehat{h}_{\ge r}^{\mathsf{T}}], \qquad V_r = \mathrm{tr}(C_r). \tag{81}$$

Let $e = \widehat{h}_{\ge r} - h_{\ge r}$. Projection onto ANOVA components of order at least r is a contraction, so $\mathbb{E}\|e\|_2^2 \le R$. Expanding

$$\widehat{C}_r - C_r = \mathbb{E}[h_{\ge r}e^{\mathsf{T}} + eh_{\ge r}^{\mathsf{T}} + ee^{\mathsf{T}}] \tag{82}$$

and applying Cauchy–Schwarz gives

$$\boxed{\|\widehat{C}_r - C_r\|_{\mathrm{op}} \le 2\sqrt{V_r R} + R.} \tag{83}$$

Thus low reconstruction error can preserve the covariance structure of input effects even when compact localization of the feature scopes remains constrained.

I NUMERICAL CHECKS FOR THE EXACT CONSTRUCTION

We numerically checked the closed form identities, the dynamic program, and the finite constructions on small instances. The discrepancies were at floating point precision. To reduce ambiguity in machine-parsed versions of the manuscript, all scalar residual variances are written with explicit parentheses before squaring rather than with absolute-value bars.

Table 1: Per-seed learned-SAE validation. $B_{0.1} = \mathcal{B}_z(0.1W_z)$. Reconstruction and sparsity are identical across map arities within each seed because each map permutes the same representation and activation samples.

Seed	$B_{0.1}, a = 1$	$B_{0.1}, a = 3$	$B_{0.1}, a = 5$	$R/\operatorname{Var}(h)$	K
0	128	206	215	0.00236	13.96
1	147	197	204	0.00220	15.52
2	168	197	205	0.00345	16.17
3	129	186	192	0.00252	14.59
4	108	176	198	0.00178	11.79
Mean	136.0	192.4	202.8	0.00246	14.41

J CONTROLLED LEARNED-SAE VALIDATION DETAILS

Training and intervention grid. The intervention variables are eight binary factors, giving all $2^8 = 256$ states. Inputs are represented as $\{-1, +1\}^8$. For each seed in $\{0, 1, 2, 3, 4\}$, 192 states are selected for training and 64 are held out. A dense $8 \to 24 \to 8$ network with a ReLU hidden layer is trained by Adam for 1500 steps to reconstruct the input factors. The 24-dimensional hidden representation is centered and scaled using only the training states. An overcomplete SAE with 48 ReLU features is then trained on the same 192 representations for 2500 Adam steps. Decoder columns are normalized inside the forward pass and the training loss is reconstruction error plus 0.01 times the sum of feature activations. Scope information is not used in either training objective.

After training, both models are frozen. For arity $a \in \{1, 3, 5\}$ we define a cyclic XOR map G_a by replacing factor i with the XOR of a consecutive factors starting at i, with indices taken modulo eight. Exhaustive enumeration verifies that all three maps are bijections of the 256-state grid. Hence $\{h(G_a(x))\}_x$ and $\{z(G_a(x))\}_x$ are identical multisets for all three arities within each seed. The arity labels describe the imposed input map and do not assert that learned SAE features are pure interactions of that order.

Exact curves. For every feature and every one of the 256 candidate scopes, we compute the population omission using the Boolean Fourier form of functional ANOVA and independently cross-check it against direct conditional means. The largest discrepancy across seeds, maps, features, and scopes is below 4×10^{-16}. Dynamic programming then gives the exact optimal curve on the complete intervention population.

Finite paired sampling. For each seed, map, and candidate scope, we estimate Equation 28 from $N \in \{32, 128, 512, 2048\}$ paired interventions. The same base pairs may be reused across scopes; each candidate remains an average of N independent paired observations. We repeat each seed, map, and N combination ten times and re-optimize the full curve after estimation. Thus each reported mean at a fixed N averages 150 independently seeded curve estimates. The mean maximum normalized curve errors are 0.123, 0.0472, 0.0216, and 0.0109, while the mean absolute errors in $\mathcal{B}_z(0.1W_z)$ are 27.6, 5.99, 1.83, and 0.65 budget units. Conservative simultaneous Hoeffding bands cover all exact curves in these runs but remain wide; their role is coverage rather than tight inference.

Certificate diagnostics. At tolerance $0.1W_z$, Equation 24 gives mean spectral budget lower bounds 4.2, 12.6, and 20.6 for arities one, three, and five, compared with exact budgets 136.0, 192.4, and 202.8. The representation-level certificate uses only order-wise interaction covariance, so it tracks the direction of the imposed dependence change while remaining coarse. Figure 4 compares it with a more information-rich feature-order relaxation.

For the latter, write $w_{j,T} = \mathbb{E}[z_{j,T}^2]$ and define $a_j(k) = \sum_{|T|>k} w_{j,T}$. Every scope with $|A_j| = k$ necessarily omits all components of order greater than k, so

$$e_j(A_j) \geq a_j(|A_j|).$$

Consequently,

$$\underline{\mathcal{L}}_z(B) = \min_{\sum_j k_j \leq B} \sum_j a_j(k_j) \leq \mathcal{L}_z(B). \tag{84}$$

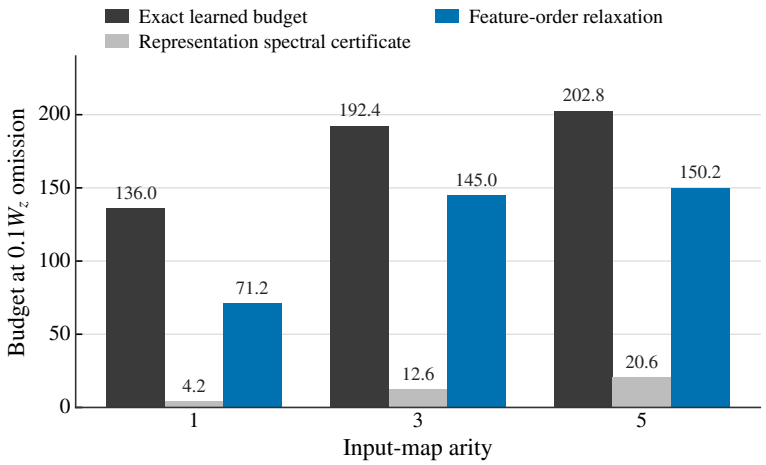

Figure 4: Certificate diagnostics for the learned-SAE validation, averaged over five seeds. The representation-level spectral certificate uses only order-wise covariance and is correspondingly coarse. The feature-order relaxation in equation 84 is tighter because it also uses ANOVA energies of the learned code.

This relaxation uses the learned code's ANOVA energies and therefore retains feature-level structure discarded by the representation-level spectral certificate. Its mean budget lower bounds at $0.1W_z$ are 71.2, 145.0, and 150.2. The tighter values illustrate the tradeoff between the information available to a certificate and the strength of the resulting localization constraint; the relaxation does not replace Theorem 3.