Benchmarking Cross-Seed Feature Correspondence in Sparse Autoencoders

Anonymous authors
Paper under double-blind review

Abstract

Sparse autoencoders (SAEs) trained on the same model learn seed-dependent dictionaries, raising the question of whether features found by one run correspond to those found by another. We introduce a benchmark that evaluates cross-seed matching methods on *functional* grounds, beyond geometric similarity, using two complementary tests: per-feature ablation fingerprints for scalable screening, and a substitution test that directly measures functional interchangeability by swapping one SAE's feature contribution for another's. Both tests are validated against hard negative controls and stratified by feature activity level.

Evaluating eight matching methods on BatchTopK and ReLU SAEs (five seeds, Pythia-410M layers 4, 8, and 12, with replication on GPT-2 Small), we find that cross-seed correspondence exhibits a quality/coverage tradeoff analogous to precision/recall. At the top of the ranking, greedy cosine and Sinkhorn optimal transport perform equally well ($R = 0.86$ at top-100); in the tail, Sinkhorn with uniform marginals retains higher quality ($R = 0.60$ vs. 0.52 at top-2000), achieving the highest overall AUSQC (area under the substitution-quality curve). Results are validated on a held-out corpus with seed-level bootstrap confidence intervals. All claims are restricted to the fingerprinted feature subset ($\sim$42%), and we show that effect sizes attenuate for low-activity features. The benchmark protocol is designed so that future consistency methods can be evaluated on the same footing, providing a shared standard for measuring progress on feature reproducibility.

1 Introduction

Sparse autoencoders (SAEs) have emerged as a promising tool for mechanistic interpretability, decomposing dense neural network activations into sparse, potentially interpretable features (Bricken et al., 2023; Cunningham et al., 2023; Templeton et al., 2024). Recent work has shown that SAEs trained on the same data can learn substantially different dictionaries (Paulo & Belrose, 2026) and do not find canonical units of analysis (Leask et al., 2025), while Song et al. (2025) argue that feature consistency should be measured explicitly rather than assumed. These results challenge naïve canonicality claims. The unresolved question is not whether features are canonical (they are not) but rather: *despite non-canonicity, can we still recover a practically useful set of cross-run correspondences with causal support?*

This is a challenging matching problem. Given two SAEs with $d_{\text{sae}} = 16{,}384$ features each, features may split (one concept encoded as two features), merge (two concepts encoded as one), or disappear entirely across training runs. Classical geometric alignment methods (orthogonal Procrustes, SVCCA) assume rigid transformations and cannot handle these phenomena. OT-style soft matching has been proposed for unit-level correspondence (Khosla & Williams, 2024), but has not been evaluated against simple baselines with causal validation.

We do not propose a new matcher. We propose a causal benchmark for explicit feature-level correspondence and use it to characterize when post-hoc matching is meaningful. To our knowledge, no prior work evaluates cross-seed SAE feature matching methods via intervention-based ranking curves that measure functional interchangeability at varying coverage levels. As one candidate matcher, we apply dustbin-augmented Sinkhorn

matching, an engineering adaptation of existing soft matching ideas (Khosla & Williams, 2024), and compare it against seven other methods including vanilla Sinkhorn with uniform marginals and thresholded baselines.

Our contributions are:

- A **causal benchmark** for cross-seed feature correspondence using ablation fingerprints (screening) and cross-feature substitution (direct functional test), validated against hard negative controls on a held-out corpus with seed-level bootstrap CIs (effective df ≈ 4).
- **Substitution ranking curves** evaluating methods across coverage levels (top-10 to top-2,000), revealing a quality/coverage tradeoff where greedy cosine matches OT at the top and Sinkhorn (uniform) maintains quality in the tail.
- An **empirical comparison of eight methods** on BatchTopK and ReLU SAEs across Pythia-410M (layers 4, 8, 12) and GPT-2 Small, stratified by activity level. Method rankings invert across architectures, confirming the benchmark's value as an architecture-agnostic protocol.

2 Related work

SAEs for feature discovery. SAEs applied to language model activations learn interpretable features via dictionary learning (Bricken et al., 2023; Cunningham et al., 2023; Templeton et al., 2024). BatchTopK (Bussmann et al., 2025) improved the reconstruction/sparsity tradeoff via exact sparsity constraints; we use this architecture trained via SAELens (Bloom & Chanin, 2024). Recent benchmarks (Karvonen et al., 2025; Chanin et al., 2026; Makelov et al., 2024) have emphasized principled, task-grounded evaluation. Chanin et al. (2024) study feature splitting and absorption, showing that split features undermine naive counting of interpretable units.

Consistency, canonicality, and universality. SAE features are not canonical: Paulo & Belrose (2026) report only ∼30% overlap across independently trained SAEs; Leask et al. (2025) show dictionaries are not uniquely determined; and Song et al. (2025) propose PW-MCC as an explicit consistency metric. PW-MCC computes absolute cosine similarity between decoder weight vectors with Hungarian assignment, producing identical correspondences to greedy cosine, so we do not include it separately. Recent work has further questioned SAE validity: Korznikov et al. (2026) show random baselines match SAEs in synthetic settings; Huang et al. (2025) find simple baselines outperform SAEs for steering; Gould et al. (2025) show interpretability metrics fail to distinguish trained from random transformers. **Our paper does not argue that SAE features are canonical**; instead, we ask: *given* non-canonicity, can we recover a subset of causally meaningful correspondences?

On the universality side, Gurnee et al. (2024) identify 1 to 5% of neurons as universal across seeds; Lan et al. (2024) demonstrate cross-model SAE feature similarity. Our work differs in recovering individual correspondences validated causally rather than measuring aggregate similarity.

Representation alignment and soft matching. Classical tools (SVCCA (Raghu et al., 2017), CKA (Kornblith et al., 2019), Procrustes (Ding et al., 2021)) operate at the subspace level and are intentionally invariant to unit permutations, which can miss unit-level correspondence (Li et al., 2016). On the OT side, Cuturi (2013) introduced entropic regularization; Singh & Jaggi (2020) applied OT for model fusion; and Khosla & Williams (2024) proposed Soft Matching Distance for unit-level correspondence via OT, extended to partial matching by Kapoor et al. (2026) (conceptually equivalent to the dustbin mechanism used here). Balagansky et al. (2025) align SAE features across layers. Training-time approaches (feature-aligned SAEs (Chen et al., 2024), ordered SAEs (Wang, 2025), faithful SAEs (Cho & Kim, 2025), weight regularization (Jedryszek & Crook, 2026), Matryoshka distillation (Martin-Linares & Ling, 2025), and direct consistency optimization (Song et al., 2025)) may reduce the need for post-hoc matching. **OT for soft unit matching is not new**; our contribution is causal validation of correspondences rather than a new matching algorithm.

Causal validation of internal representations. Meng et al. (2022) established causal tracing for identifying editable structure; Geiger et al. (2024) frame alignment as finding correspondences between causal

variables and distributed representations; Wang et al. (2023) demonstrated circuit-level analysis for indirect object identification; and Zhang & Nanda (2024); Heimersheim & Janiak (2024) established best practices for activation patching. Marks et al. (2024) showed SAE features participate in causally meaningful sparse circuits. We adapt per-feature ablations to construct causal fingerprint vectors as a screening metric, supplemented by direct cross-feature substitution tests. **This causal evaluation of cross-seed correspondence quality is the main contribution.**

Positioning relative to existing benchmarks. Several recent works measure SAE consistency or similarity, but none center on intervention-based evaluation of feature-level correspondence. PW-MCC (Song et al., 2025) computes absolute cosine similarity between decoder weight vectors with optimal assignment, producing an aggregate consistency score; the underlying correspondences are identical to greedy cosine matching, and no functional validation is performed. Soft Matching Distance (Khosla & Williams, 2024) measures *representation*-level similarity via OT, producing a scalar distance rather than ranked feature-level correspondences; the partial extension (Kapoor et al., 2026) adds mass discarding but does not evaluate whether individual matched features are interchangeable. SAEBench (Karvonen et al., 2025) evaluates SAE quality (absorption, faithfulness, spurious correlations) rather than cross-seed matching: it answers "is this SAE good?" rather than "do these two SAEs share features?" Our benchmark is complementary: it takes any matcher that produces ranked feature pairs and evaluates whether those pairs are *functionally interchangeable* under causal intervention, at varying coverage levels.

3 Methods

3.1 Optimal transport matching

Given two SAEs with decoder weight matrices $W_A, W_B \in \mathbb{R}^{d_{\text{sae}} \times d_{\text{model}}}$ and sparse feature activations a_A, a_B, we define a composite cost matrix:

$$\mathbf{C}_{ij} = w_{\text{dir}} \cdot C_{\text{dir}}(i,j) + w_{\text{act}} \cdot C_{\text{act}}(i,j) + w_{\text{sp}} \cdot C_{\text{sp}}(i,j) \tag{1}$$

where C_{dir} is angular distance (a true metric on the projective sphere), C_{act} is activation correlation distance (Pearson correlation restricted to positions where either feature is active), and C_{sp} is Jaccard distance on feature supports. We use default weights $(0.4, 0.4, 0.2)$, chosen to balance geometric and behavioral signals without tuning on the causal evaluation metric.

We solve the entropy-regularized optimal transport problem (Cuturi, 2013):

$$\mathbf{T}^* = \arg\min_{\mathbf{T} \geq 0} \langle \mathbf{C}, \mathbf{T} \rangle + \varepsilon H(\mathbf{T}) \quad \text{s.t.} \quad \mathbf{T}\mathbf{1} = \mu_A,\ \mathbf{T}^\top \mathbf{1} = \mu_B \tag{2}$$

where μ_A, μ_B are activation-mass-weighted marginals, $\varepsilon = 0.05$ is the regularization strength, and $H(\mathbf{T}) = -\sum_{ij} T_{ij} \log T_{ij}$ is the entropic regularizer.

Dustbin features. Following the dustbin mechanism of Sarlin et al. (2020), we augment $\mathbf{C}$ with an extra row and column at cost $c_{\text{dustbin}} = 0.5$ (the midpoint of the normalized cost range). Mass flowing to the dustbin represents unmatched features, preventing forced bad matches. This is equivalent to the partial transport formulation of Kapoor et al. (2026).

Correspondence score. We summarize correspondence quality with $\text{CS}(A, B) = \max(0,\ 1 - \text{OT}(\mu_A, \mu_B, \mathbf{C}))$, where OT denotes the regularized transport cost, a task-specific similarity score in $[0, 1]$. This is a practical summary statistic rather than a metric in the mathematical sense; our primary evaluation relies on the causal validation described below, not on the correspondence score itself.

The pipeline (Figure 6 in Appendix D) combines decoder weights and activations into the composite cost matrix, solved via dustbin-augmented Sinkhorn OT.

3.2 Causal validation

We validate feature correspondences using *causal fingerprints*, per-feature ablation effect vectors that summarize each feature's downstream causal role. This adapts the activation patching framework (Wang et al., 2023; Zhang & Nanda, 2024; Heimersheim & Janiak, 2024) to individual SAE features: for each feature f_i, we compute the mean absolute change in top-k logits when f_i is ablated (zeroed) across a reference corpus:

$$\phi_i = \mathbb{E}_x \left[|\text{logits}(x) - \text{logits}(x \mid f_i = 0)|_{\text{top-}k} \right] \tag{3}$$

If a matching method correctly identifies corresponding features, matched pairs (i, j) should have similar causal fingerprints. We measure causal divergence as cosine distance: $d(\phi_i^A, \phi_j^B) = 1 - |\cos(\phi_i^A, \phi_j^B)|$. For correctly matched pairs, this should be smaller than for random pairs.

3.3 Substitution test

Causal fingerprint similarity shows matched features have similar *sensitivity* to ablation, but does not show they play the same functional role. We strengthen this with a *substitution test*, inspired by interchange interventions (Geiger et al., 2024), but swapping a feature's contribution between independently trained SAEs on the same input rather than across inputs. For a matched pair (A_i, B_j), we ablate A_i's contribution from the residual stream and inject B_j's decoder direction with a scaling factor α fitted via least squares to minimize logit error. The *recovery score* R measures the fraction of A_i's causal effect recovered by the substitution:

$$R = \frac{\|\boldsymbol{\delta}_{\text{sub}}\|^2}{\|\boldsymbol{\delta}_{\text{abl}}\|^2}$$

where $\boldsymbol{\delta}_{\text{abl}} = \texttt{logits}_{\text{orig}} - \texttt{logits}_{\text{ablated}}$ is the logit change from ablating A_i (the causal effect), and $\boldsymbol{\delta}_{\text{sub}} = \texttt{logits}_{\text{substituted}} - \texttt{logits}_{\text{ablated}}$ is the logit change contributed by injecting B_j (how much the substitution restores). Because α is the least-squares projection coefficient, R equals the coefficient of determination (R^2) of regressing $\boldsymbol{\delta}_{\text{abl}}$ onto the substitution direction, i.e., the squared cosine between the two logit-effect vectors (we write R for brevity). This differs from logit-difference recovery in activation patching (Zhang & Nanda, 2024), which restores a scalar output difference; ours measures explained variance across the full logit vector. $R = 1$ indicates perfect recovery; $R = 0$ indicates no effect; values slightly above 1 can occur due to finite-data fitting and LayerNorm nonlinearity. The fitted α accounts for scale differences between SAEs.

4 Experimental setup

Model and SAE architecture. We train SAEs on Pythia-410M (Biderman et al., 2023) (24 layers, $d_{\text{model}} = 1024$) using TransformerLens and SAELens (Bloom & Chanin, 2024). SAEs use the BatchTopK architecture (Bussmann et al., 2025) with $k = 100$ and $d_{\text{sae}} = 16{,}384$ (16× expansion), trained on 100M tokens from the Pile (streamed) with Adam (lr $= 5 \times 10^{-5}$, cosine annealing, 1,000 warmup steps, batch size 4,096 tokens × 256 context length). For architecture comparison (Section 5.9), we additionally train 5 standard ReLU SAEs at layer 8 with $L_1 = 5.0$ (calibrated to produce $L_0 \approx 69$, comparable to BatchTopK $k = 100$), same seeds and all other hyperparameters. Decoder columns are ℓ_2-normalized after each update. Quality metrics for all 28 SAEs are reported in Appendix A.

Seed stability experiment. Five SAEs are trained on the layer-8 residual stream post-activations (`hook_resid_post`) with seeds $\{42, 123, 456, 789, 1024\}$, yielding $\binom{5}{2} = 10$ pairs for matching and $10 \times 2 = 20$ Sinkhorn runs for bidirectional transport.

Matching methods and hyperparameters. We compare eight methods: greedy cosine, sparse Hungarian, Sinkhorn (composite cost), Sinkhorn (cosine-only), Sinkhorn with uniform marginals (inspired by Soft Matching Distance Khosla & Williams, 2024, though using angular cost on decoder weights rather than L_2^2 on activations), mutual nearest neighbors (MNN, $k = 5$), orthogonal Procrustes, and SVCCA. PW-MCC

(Song et al., 2025) is not included as a separate method because it defines the same optimization (absolute cosine similarity with optimal assignment), producing identical correspondences to greedy cosine and sparse Hungarian. Composite cost weights $(w_{\text{dir}}, w_{\text{act}}, w_{\text{sp}}) = (0.4, 0.4, 0.2)$ and Sinkhorn regularization $\varepsilon = 0.05$ follow Section 3.1; dustbin cost is $c_{\text{dustbin}} = 0.5$. Marginals are weighted by activation mass; the sparse top-k candidate set uses $k = 256$ nearest neighbors per feature. All methods are evaluated via fingerprint screening and substitution ranking curves.

Evaluation protocol. For each of the 10 seed pairs, every matching method produces a set of feature correspondences. We compute the causal divergence (cosine distance between causal fingerprint vectors) for each matched pair and compare to a size-matched random baseline. We report Cohen's d and AUROC, averaged across all 10 pairs, as screening metrics; the substitution ranking curves (Section 5.4) provide the primary evaluation.

Causal fingerprints. For each SAE, we compute causal fingerprints by ablating active features at the residual stream after SAE reconstruction across 256 sequences (sampled sequentially from the Pile training split), tracking mean absolute changes in the top-50 logit positions. Logit positions are sampled from the top 5,000 by importance (mean absolute logit magnitude over 16 probe sequences) to ensure diversity across the fingerprint dimensions. Ablation subtracts the single-feature contribution $a_i \cdot \mathbf{w}_i^{\text{dec}}$ from the residual stream at each active position; ablated residuals are batched (32 per forward pass) for efficiency. This yields a fingerprint vector $\phi_i \in \mathbb{R}^{50}$ per feature for those active on the evaluation corpus ($\sim$42% coverage per SAE). We treat fingerprints as a screening metric for functional similarity, not as ground truth; the substitution test (Section 5.4) provides the more direct causal evidence. Activations are drawn from the same distribution used for SAE training; because our goal is to evaluate cross-run feature correspondence rather than generalization, this does not constitute data leakage. SAE training prepends a BOS token whereas evaluation activations do not; because matching aggregates across all 256 token positions, this affects $< 0.4\%$ of tokens and does not materially influence results.

5 Results

5.1 Fingerprint screening separates methods (but is not the primary metric)

Table 1 and Figure 1 show fingerprint screening results across all 10 seed pairs. Sinkhorn composite achieves $d = 1.85 \pm 0.03$, AUROC $= 0.905$; Sinkhorn (uniform, inspired by Khosla & Williams, 2024) achieves $d = 1.70$; greedy cosine $d = 1.34$. Procrustes ($d = 0.00$) and SVCCA ($d = 0.28$) fail as expected: they assume rigid transformations incompatible with feature reordering and splits. The random baseline is weak; the informative comparisons are among geometry-based methods. As the substitution curves show (Section 5.4), OT's fingerprint advantage reflects its smaller, more selective match set rather than better top correspondences: MNN and greedy match or exceed OT at high selectivity.

5.2 Geometry already identifies many top correspondences

Sinkhorn with cosine-only cost uses *the same input* as Procrustes (decoder weights alone) yet achieves $d = 1.76$ vs. $d = 0.00$ (Figure 1), confirming that the gap reflects algorithm flexibility (soft permutations vs. rigid rotation), not additional data. Cosine-only OT nearly matches the composite ($d = 1.76$ vs. 1.85); activation statistics contribute marginally. Transport plan visualization (Figure 7 in Appendix D) reveals near-diagonal structure with off-diagonal entries for splits and merges, consistent with overcomplete dictionaries (Chanin et al., 2024).

Cost component analysis. A cost ablation (Figure 8 in Appendix D) reveals that the composite cost achieves higher fingerprint d (1.85) than cosine-only (1.76), with activation correlation contributing the most to OT transport quality and angular distance acting as a regularizer that trades breadth for precision.

Table 1: Causal validation of matching methods (mean ± std across all 10 seed pairs). Cohen's $d = (\bar{d}_{\text{random}} - \bar{d}_{\text{matched}})/s_{\text{pooled}}$; higher means matched pairs have lower causal divergence than random pairs. "Matched" counts only pairs where both features have nonzero fingerprints (∼42% coverage); methods matching fewer pairs are more selective, so part of the performance gap may reflect selectivity rather than match quality. Note: greedy cosine produces ∼16,000 total matches, of which ∼4,560 have fingerprinted features on both sides (65.9% of fingerprinted features; Table 9); at threshold = 0.5 (used here), only 1,494 fingerprinted matches survive, which is why $d = 1.34$ here vs. $d = 0.54$ at full coverage.

Method	Matched	Cohen's d	AUROC
Sinkhorn composite	491 ± 8	1.85 ± 0.03	0.905 ± 0.009
Sinkhorn cosine	721 ± 18	1.76 ± 0.02	0.897 ± 0.004
Sinkhorn (uniform)	767 ± 18	1.70 ± 0.03	0.887 ± 0.005
Greedy cosine	$1{,}494 \pm 39$	1.34 ± 0.05	0.825 ± 0.009
MNN (k=5)	$3{,}980 \pm 49$	0.63 ± 0.02	0.665 ± 0.004
Sparse Hungarian	$3{,}312 \pm 43$	0.60 ± 0.02	0.655 ± 0.004
SVCCA	$3{,}631 \pm 58$	0.28 ± 0.01	0.566 ± 0.004
Orthogonal Procrustes	$3{,}047 \pm 45$	0.00 ± 0.02	0.501 ± 0.005

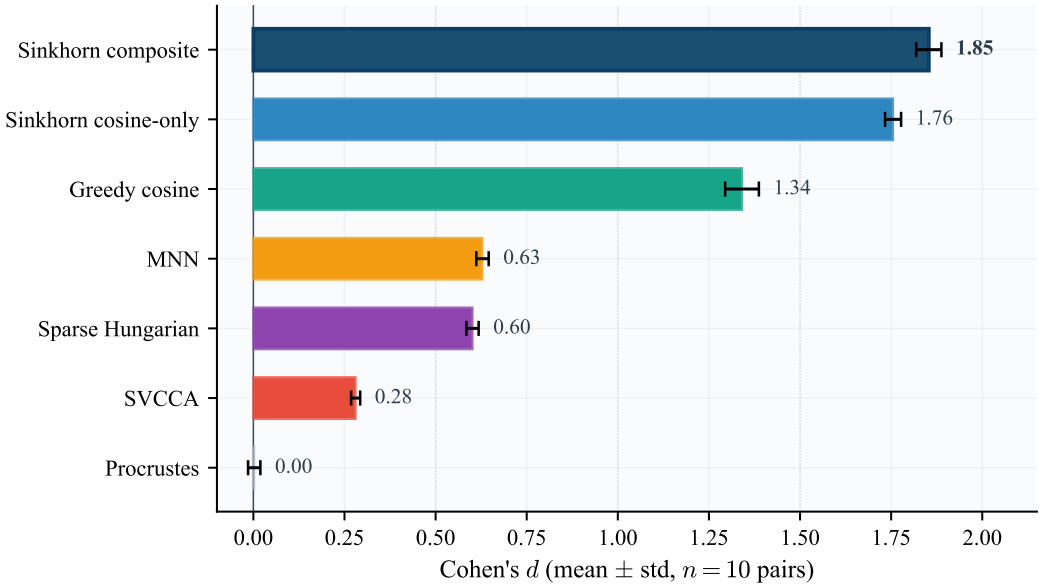

Figure 1: Cohen's d for each matching method (mean ± std across 10 seed pairs, effective df ≈ 4). Classical alignment baselines (Procrustes, SVCCA) fail as expected; the more informative comparison is among geometry-based matchers at matched coverage levels.

5.3 OT design ablations and coverage-controlled comparison

OT component ablations (Table 7 in Appendix C) show the dustbin is the most impactful component ($\Delta d = 0.063$); removing Jaccard has negligible effect. Sinkhorn with uniform marginals (no dustbin, no activation-mass weighting), inspired by Khosla & Williams (2024), achieves $d = 1.70$.

Quality/coverage analysis (Tables 8 and 9 in Appendix C) confirms that thresholded greedy at $t = 0.9$ achieves $d = 2.65$ with only ∼61 pairs, revealing a two-regime structure: at the top, geometry-based matchers recover elite correspondences regardless of algorithm; at broader coverage (top-2000+), OT's dustbin prevents quality degradation (Table 2). Note that PW-MCC (Song et al., 2025) uses the same decoder-weight cosine

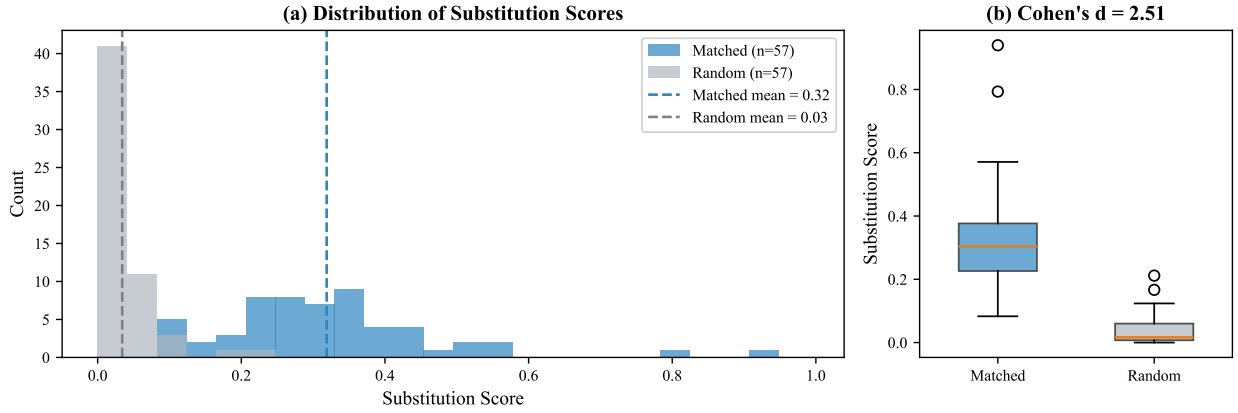

Figure 2: **(a)** Distribution of recovery scores R for top-100 matched and random feature pairs across all 10 seed pairs. Top-ranked matched features recover a large fraction of the ablated feature's causal effect, while random features cluster near zero. **(b)** Box plot showing clear separation ($R = 0.86$ for greedy/Sinkhorn (uniform), $R = 0.60$ for Sinkhorn composite at top-100).

similarity as greedy matching; its aggregate score (mean |cos| of optimally matched pairs) is a useful summary statistic, but the underlying correspondences are identical to greedy cosine.

5.4 Substitution ranking curves: the main result

The substitution test (Section 3.3) is the primary evaluation in this benchmark, as it directly measures whether matched features can recover each other's causal effects. Unlike fingerprint-based screening, the substitution test directly measures functional interchangeability. We evaluate substitution recovery across the *full ranking curve* for five methods that produce distinct ranked feature pairs, at coverage levels top-10, top-50, top-100, top-500, top-1000, and top-2000 (Table 2). Orthogonal Procrustes and SVCCA produce subspace alignments rather than ranked feature-pair lists; sparse Hungarian and PW-MCC produce ranked pairs but yield assignments identical to greedy cosine ($\Delta R < 0.004$; see Section 4), so we report five distinct curves.

Top-pair regime. At top-10 to top-100, greedy cosine, MNN, and Sinkhorn (uniform) achieve $R = 0.95$ and $R = 0.86$ respectively, outperforming Sinkhorn composite ($R = 0.47/0.60$; Figure 2). The strongest correspondences are apparent in decoder geometry alone; Sinkhorn composite's lower top-pair R reflects mass-weighted ranking that conflates feature importance with match quality.

Long-tail regime. Greedy degrades from $R = 0.69$ (top-500) to $R = 0.52$ (top-2000), while Sinkhorn composite holds at $R \approx 0.59$ via dustbin filtering. Sinkhorn (uniform) bridges both regimes: matching greedy through top-1000, then maintaining $R = 0.60$ at top-2000 because its ∼1,100 assignments cap coverage before quality degrades. We summarize with the area under the substitution-quality curve (AUSQC):

$$\text{AUSQC} = \int_{\log_{10} 10}^{\log_{10} K_{\max}} R(k)\, d(\log_{10} k)$$

where $K_{\max} = 5{,}000$ (theoretical max ≈ 2.70). Log-scale integration gives equal weight to each decade of coverage. Sinkhorn (uniform) achieves the highest AUSQC (2.06 vs. greedy's 1.99; Table 2).

Hard negative controls. All comparisons above use random negatives (a weak baseline). Against five types of confound-matched hard negatives (Section 5.6; cosine-neighborhood, mass-matched, Jaccard, permuted top-k, cross-seed decoys), absolute effect sizes decrease but positive separation persists across all types (Table 13).

Table 2: Substitution recovery R across ranking curve (mean ± std across 10 Pythia seed pairs). Random control $R \approx 0.002$ at all levels. AUSQC = area under the substitution-quality curve (mean R integrated over log-coverage). Sinkhorn (uniform) matches greedy at high selectivity and maintains quality in the tail, achieving the highest AUSQC.

Method	top-10	top-50	top-100	top-500	top-1000	top-2000	AUSQC
Greedy cosine	$.95_{\pm.01}$	$.90_{\pm.01}$	$.86_{\pm.01}$	$.69_{\pm.01}$	$.61_{\pm.01}$	$.52_{\pm.01}$	1.99
MNN (k=5)	$.95_{\pm.01}$	$.90_{\pm.01}$	$.86_{\pm.01}$	$.69_{\pm.01}$	$.61_{\pm.01}$	$.52_{\pm.01}$	1.99
Sinkhorn composite	$.47_{\pm.07}$	$.59_{\pm.04}$	$.60_{\pm.04}$	$.59_{\pm.02}$	$.59_{\pm.02}$	$.59_{\pm.02}$	1.55
Sinkhorn cosine	$.48_{\pm.07}$	$.58_{\pm.04}$	$.60_{\pm.03}$	$.56_{\pm.02}$	$.53_{\pm.01}$	$.53_{\pm.01}$	1.49
Sinkhorn (uniform)	$\mathbf{.95}_{\pm.01}$	$\mathbf{.90}_{\pm.01}$	$\mathbf{.86}_{\pm.01}$	$\mathbf{.69}_{\pm.01}$	$\mathbf{.61}_{\pm.01}$	$\mathbf{.60}_{\pm.01}$	**2.06**

5.5 Held-out validation

We recompute all evaluations on a disjoint held-out corpus shard (256 sequences). Matched-vs-random separation persists and method rankings are preserved (Table 11 in Appendix C). Absolute d values differ between shards due to coverage variation, but ordinal rankings and substitution R are stable (CV $< 5\%$).

5.6 Hard negative controls

Beating random pairs is a low bar. We evaluate against five types of confound-matched hard negatives: cosine-neighborhood, activation-mass-matched, Jaccard-neighborhood, permuted top-k, and cross-seed decoys (details and full results in Table 13). Against hard negatives, absolute effect sizes decrease substantially ($d_{\text{hard}} \ll d_{\text{random}}$), but both methods remain above zero separation against all five types. No method-difference Δd reaches statistical significance (seed-level bootstrap 95% CI includes zero for all types). Cosine-neighborhood negatives are the hardest, reducing d by $\sim$20%, indicating that much of the random-baseline effect reflects geometric structure rather than functional correspondence itself.

5.7 Results stratified by feature activity

Causal fingerprints exist only for features active on the reference corpus ($\sim$42%), creating a potential selection bias. Stratifying by activation mass tercile (Table 14), fingerprint d attenuates 2.2× from high to low ($d = 0.78$ to 0.36). The substitution test is less affected: recovery attenuates only 1.3× ($R = 0.49$ high, $R = 0.38$ low), all far above random ($R \approx 0.002$). Activity bias inflates the secondary screening metric more than the primary evaluation.

5.8 OT sensitivity analysis

A broad hyperparameter sweep (Table 10 in Appendix C) shows that at fixed $\varepsilon = 0.05$, d varies by less than 21% across composite weights, dustbin cost, and marginal mode. The most sensitive parameter is ε itself ($d \in [1.20, 2.86]$), but the practical impact on method ranking is negligible.

5.9 Multi-layer and cross-model scope

All primary results are from layer 8 of Pythia-410M (5 seeds, 10 pairs). We replicate on GPT-2 Small (layer 6, 5 seeds) and extend to Pythia layers 4 and 12 (5 seeds each, 10 pairs per layer) below.

At layer 8 (3.4% dead features), fingerprint screening gives Sinkhorn composite $d = 1.85$, greedy $d = 1.34$, MNN $d = 0.63$; however, the substitution ranking curves (Section 5.4) show that greedy matches or exceeds OT at high selectivity, with OT's benefit limited to broader coverage.

Multi-layer evaluation. To test whether the substitution ranking patterns generalize across network depth, we train additional SAEs at layers 4 and 12 of Pythia-410M (4 new seeds per layer, complementing

the existing seed-42 SAEs from the layer-correspondence experiment, for 5 seeds and 10 pairs per layer). Table 3 shows substitution recovery across three decoder-weight-only methods at layers 4, 8, and 12.

Table 3: Substitution recovery R across network depth (mean ± std, 10 seed pairs per layer). All methods use decoder weights only; no activation files required. AUSQC integrates R over log-coverage (max ≈ 2.70).

Layer	Method	top-10	top-100	top-1000	top-2000	AUSQC
4	Greedy cosine	$0.94_{\pm.01}$	$0.83_{\pm.01}$	$0.61_{\pm.01}$	$0.51_{\pm.01}$	$1.96_{\pm.02}$
	MNN	$0.94_{\pm.01}$	$0.83_{\pm.01}$	$0.61_{\pm.01}$	$0.51_{\pm.01}$	$1.96_{\pm.02}$
	Sinkhorn (uniform)	$0.94_{\pm.01}$	$0.83_{\pm.01}$	$0.61_{\pm.01}$	$\mathbf{0.60}_{\pm.01}$	$\mathbf{2.04}_{\pm.02}$
8	Greedy cosine	$0.95_{\pm.01}$	$0.86_{\pm.01}$	$0.61_{\pm.01}$	$0.52_{\pm.01}$	$1.99_{\pm.02}$
	MNN	$0.95_{\pm.01}$	$0.86_{\pm.01}$	$0.61_{\pm.01}$	$0.52_{\pm.01}$	$1.99_{\pm.02}$
	Sinkhorn (uniform)	$0.95_{\pm.01}$	$0.86_{\pm.01}$	$0.61_{\pm.01}$	$\mathbf{0.60}_{\pm.01}$	$\mathbf{2.06}_{\pm.02}$
12	Greedy cosine	$0.96_{\pm.00}$	$0.85_{\pm.02}$	$0.58_{\pm.01}$	$0.48_{\pm.00}$	$1.96_{\pm.02}$
	MNN	$0.96_{\pm.00}$	$0.85_{\pm.02}$	$0.58_{\pm.01}$	$0.48_{\pm.00}$	$1.96_{\pm.02}$
	Sinkhorn (uniform)	$0.96_{\pm.00}$	$0.85_{\pm.02}$	$0.60_{\pm.01}$	$\mathbf{0.60}_{\pm.01}$	$\mathbf{2.07}_{\pm.03}$

The core pattern replicates at all three layers: methods are within $\Delta R \leq 0.02$ through top-1000, with Sinkhorn's advantage emerging in the tail. Layer 12 shows the sharpest tail divergence ($\Delta = 0.12$, vs. 0.09 at layer 4), consistent with later layers having more specialized features. Standard deviations are ≤ 0.03 across all layers.

GPT-2 Small replication. We replicate on GPT-2 Small (layer 6, 5 seeds, BatchTopK with $k = 80$, $d_{\text{sae}} = 12{,}288$). The ordinal fingerprint-screening ranking (OT > greedy > MNN > SVCCA $\gg$ Procrustes ≈ 0) is consistent across models: Sinkhorn composite $d = 1.88 \pm 0.59$ (Pythia: 1.85 ± 0.03), greedy $d = 1.13 \pm 0.03$ (Pythia: 1.34 ± 0.05). Substitution ranking curves (Table 12) confirm the same pattern: greedy/MNN/Sinkhorn (uniform) achieve $R = 0.79$ at top-10 and $R = 0.64$ at top-100, while Sinkhorn composite trails at $R = 0.31/0.30$. Sinkhorn (uniform) again achieves the highest AUSQC (1.65 vs. greedy's 1.46), maintaining $R = 0.52$ at top-2000 where greedy degrades to $R = 0.30$. The GPT-2 composite result exhibits higher variance in fingerprint d, likely due to lower fingerprint coverage (∼17% vs. 42%); the ordinal ranking is robust but point estimates should be interpreted cautiously.

Architecture comparison: ReLU SAEs. To test whether the benchmark generalizes beyond BatchTopK, we train 5 standard ReLU SAEs at layer 8 of Pythia-410M (same seeds, $d_{\text{sae}} = 16{,}384$, $L_1 = 5.0$), yielding $L_0 \approx 69$ and EV = 0.982. Table 4 shows substitution recovery for the same three decoder-weight methods.

Table 4: Substitution recovery R for ReLU SAEs (mean ± std, 10 seed pairs, layer 8), same columns as Table 2. ReLU SAEs use L_1 regularization ($L_0 \approx 69$) instead of BatchTopK exact sparsity ($k = 100$). Higher R reflects the sparser, higher-variance per-feature activations in ReLU SAEs.

Method	top-10	top-50	top-100	top-500	top-1000	top-2000	AUSQC
Greedy cosine	$\mathbf{1.00}_{\pm.00}$	$\mathbf{1.00}_{\pm.00}$	$\mathbf{1.00}_{\pm.00}$	$\mathbf{1.00}_{\pm.00}$	$\mathbf{1.00}_{\pm.00}$	$\mathbf{1.00}_{\pm.00}$	$\mathbf{2.69}_{\pm.00}$
MNN	$\mathbf{1.00}_{\pm.00}$	$\mathbf{1.00}_{\pm.00}$	$\mathbf{1.00}_{\pm.00}$	$\mathbf{1.00}_{\pm.00}$	$\mathbf{1.00}_{\pm.00}$	$\mathbf{1.00}_{\pm.00}$	$\mathbf{2.69}_{\pm.00}$
Sinkhorn (uniform)	$0.82_{\pm.05}$	$0.88_{\pm.02}$	$0.91_{\pm.01}$	$0.93_{\pm.00}$	$0.93_{\pm.00}$	$0.92_{\pm.00}$	$2.41_{\pm.02}$

Two patterns emerge (Figure 3). First, the benchmark's core validity is architecture-independent: matched features achieve $R \approx 1.0$ vs. $R \approx 0.002$ for random. Second, the *method ranking inverts*: greedy and MNN dominate at every coverage level ($R > 0.99$), while Sinkhorn (uniform) trails ($R = 0.82$ to 0.93; Figure 5). With ReLU's sparser activations ($L_0 \approx 69$), each feature explains more per-token variance, so greedy's exhaustive ranking incurs no quality penalty.

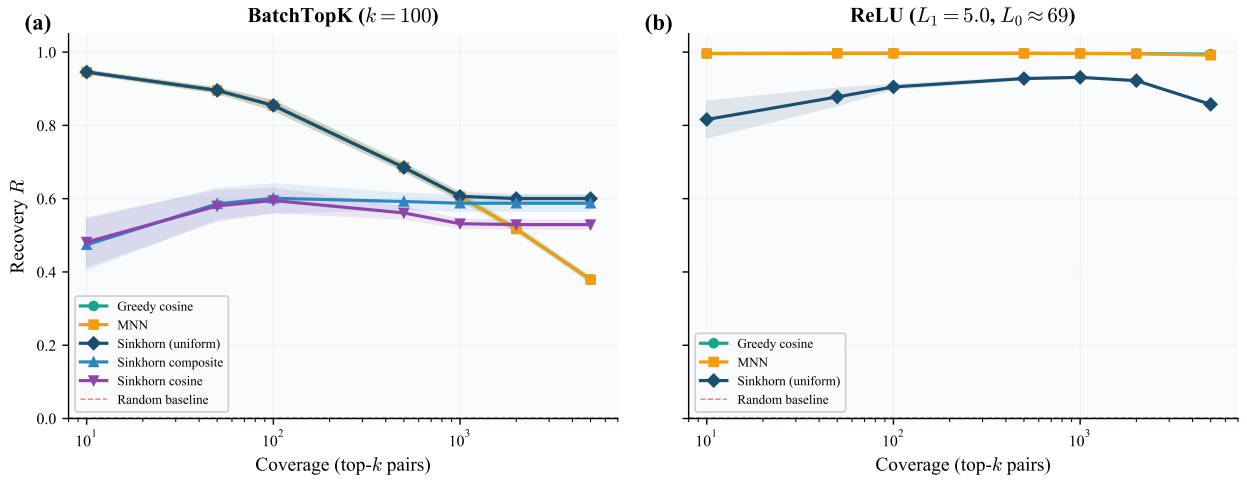

Figure 3: Substitution ranking curves for **(a)** BatchTopK and **(b)** ReLU SAEs. The method ranking inverts across architectures: on BatchTopK, Sinkhorn (uniform) maintains quality in the tail while greedy degrades; on ReLU, greedy and MNN achieve near-perfect recovery ($R > 0.99$) across all coverage levels, while Sinkhorn trails. Shaded regions show ± 1 std across 10 seed pairs.

Fingerprint screening on ReLU reveals an asymmetry: Sinkhorn cosine achieves $d = 2.60 \pm 0.04$ vs. greedy's $d = 2.40 \pm 0.04$, even though greedy dominates on substitution, confirming that the two evaluation axes measure complementary properties.

6 Discussion

A benchmark, not a matcher. We do not claim OT is the best matcher or that SAE features are canonical. The benchmark provides functional tests that separate methods appearing similar under geometric metrics. Recent work questions SAE feature validity (Korznikov et al., 2026; Huang et al., 2025; Gould et al., 2025); our substitution test provides complementary evidence. If features were arbitrary directions, cosine-matched pairs across independent runs should not be functionally interchangeable, yet $R = 0.86$ for matched vs. $R = 0.002$ for random at top-100 persists across both models, multiple layers, and two SAE architectures (rising to $R \approx 1.0$ on ReLU). This does not prove features are "real," but provides evidence that top-ranked correspondences capture shared computational structure.

Simple methods work well. Greedy cosine matching, requiring only decoder weights, achieves the same substitution recovery as OT through top-1000 on BatchTopK and *dominates at all coverage levels* on ReLU ($R > 0.99$; Figure 3). Methods diverge only in the tail, where OT's dustbin prevents quality degradation on BatchTopK (highest AUSQC on both Pythia and GPT-2). OT's tail advantage does not transfer to ReLU, where sparser activations virtually eliminate tail degradation. Hard negatives reduce absolute effect sizes but both methods maintain positive separation under all five types (Table 13).

Activity bias and scope. Top-ranked matches are dominated by high-activity features: 95% of the top-100 pairs fall in the high-activity tercile (Section 5.7). However, substitution attenuation from high to low terciles is only 1.3× ($R = 0.49$ vs. 0.38), compared to 2.2× for fingerprint screening. All claims are restricted to the fingerprinted subset (∼42%) with CIs reflecting 5 seeds (effective df ≈ 4).

Limitations. This benchmark covers two decoder-only families at three Pythia layers and one GPT-2 layer; replication on architecturally distinct models and additional SAE architectures (Gated, Matryoshka) remains important. Fingerprint coverage (∼42%) correlates with activation mass. Causal fingerprints project onto $\mathbb{R}^{50}$; methods with fewer matches are inherently more selective. Hard negatives reduce but do not eliminate the concern that method-level differences reflect cost matrix properties rather than functional

correspondence. Training-time consistency methods (Chen et al., 2024; Wang, 2025; Song et al., 2025) can be evaluated on the same protocol.

7 Conclusion

We propose a causal benchmark for cross-seed SAE feature correspondence, centering on substitution tests validated against hard negative controls with seed-level bootstrap CIs. Cross-seed correspondence is a quality/coverage tradeoff: simple cosine baselines match or exceed optimal transport at the top of the ranking. Which method wins overall depends on SAE architecture: on BatchTopK, Sinkhorn with uniform marginals achieves the highest AUSQC by maintaining tail quality; on ReLU, greedy cosine dominates at all coverage levels ($R > 0.99$). This architecture dependence, consistent across three Pythia layers and GPT-2 Small and restricted to the fingerprinted subset (∼42%), reinforces the need for architecture-aware evaluation when selecting a matching method. Future work should compare post-hoc matching against training-time consistency methods (Song et al., 2025; Chen et al., 2024; Wang, 2025) and extend to more model families.

Ethics Statement

This work studies the internal representations of language models and does not involve human subjects, private data, or deployment of AI systems. Our experiments use publicly available models (Pythia-410M, GPT-2 Small) and open datasets (the Pile, OpenWebText). We see no direct negative societal impacts from the methodology presented here, though improved understanding of neural network internals could in principle be applied to both beneficial (safety, interpretability) and harmful (adversarial manipulation) ends.

Reproducibility Statement

All code, configurations, and analysis scripts are available in the supplementary material. SAEs are trained using SAELens with fixed random seeds; all hyperparameters are specified in configuration files. Causal fingerprints are computed deterministically given a fixed reference corpus. We report results across five random seeds and include standard deviations for all aggregate metrics. The threshold sweep (Appendix B) and held-out validation (Table 11) demonstrate robustness of the main findings.

References

Nikita Balagansky, Basim Mustafa, Andreas Steiner, and Luca Bauer. Mechanistic permutability: Match features across layers. *International Conference on Learning Representations*, 2025.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. *International Conference on Machine Learning*, 2023.

Joseph Bloom and David Chanin. SAELens. https://github.com/jbloomAus/SAELens, 2024. Software library for training and analyzing sparse autoencoders.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Bart Bussmann, Patrick Leask, and Neel Nanda. BatchTopK sparse autoencoders. *International Conference on Learning Representations*, 2025.

David Chanin, James Wilken-Smith, Tomáš Dulka, Joseph Bloom, and Neel Nanda. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. *arXiv preprint arXiv:2409.14507*, 2024.

David Chanin, Eoin Farrell, and Joseph Bloom. SynthSAEBench: A synthetic benchmark for evaluating sparse autoencoders. *arXiv preprint arXiv:2602.14687*, 2026.

Trenton Chen, Shivam Prakash, and Tarun Guha. Enhancing neural network interpretability with feature-aligned sparse autoencoders. *arXiv preprint arXiv:2411.01220*, 2024.

Sangjun Cho and Peter Y Kim. Faithful sparse autoencoders: Training without out-of-distribution fake features. *arXiv preprint arXiv:2506.17673*, 2025.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 2013.

Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding representation similarity through statistical testing. *Advances in Neural Information Processing Systems*, 2021.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D Goodman. Finding alignments between interpretable causal variables and distributed neural representations. *Conference on Causal Learning and Reasoning (CLeaR), PMLR*, 236:160–187, 2024.

Stephen Gould et al. Automated interpretability metrics do not distinguish trained and random transformers. *arXiv preprint arXiv:2501.17727*, 2025.

Wes Gurnee, Theo Horsley, Neel Nanda, Ziming Lim, and Dimitris Bertsimas. Universal neurons in GPT-2 language models. *arXiv preprint arXiv:2401.12181*, 2024.

Stefan Heimersheim and Jett Janiak. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024.

Zhengfu Huang, Alessandro Stolfo, Wentao Xu, Mrinmaya Sachan, Yibo Cao, and Samyak Arora. AxBench: Steering LLMs? even simple baselines outperform sparse autoencoders. In *International Conference on Machine Learning (ICML)*, 2025.

Krzysztof Jedryszek and Paul Crook. Stable and steerable sparse autoencoders with weight regularization. *arXiv preprint arXiv:2603.04198*, 2026.

Surbhi Kapoor, Alex H Williams, and Meenakshi Khosla. Partial soft-matching distance for neural representational similarity. *arXiv preprint arXiv:2602.19331*, 2026.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum 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. *International Conference on Machine Learning*, 2025.

Meenakshi Khosla and Alex H Williams. Soft matching distance: A metric on neural representations that captures single-neuron tuning. *Proceedings of the First Workshop on Unifying Representations in Neural Models (UniReps), PMLR*, 243:326–341, 2024.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. *International Conference on Machine Learning*, 2019.

Anton Korznikov, Andrey Galichin, Alexey Dontsov, Oleg Rogov, Ivan Oseledets, and Elena Tutubalina. Sanity checks for sparse autoencoders: Do SAEs beat random baselines? *arXiv preprint arXiv:2602.14111*, 2026.

Michael Lan, Philip Torr, Austin Meek, Ashkan Khakzar, David Krueger, and Fazl Barez. Sparse autoencoders reveal universal feature spaces across large language models. *arXiv preprint arXiv:2410.06981*, 2024.

Patrick Leask, Bart Bussmann, Michael Pearce, Joseph Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. Sparse autoencoders do not find canonical units of analysis. *International Conference on Learning Representations*, 2025.

Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. Convergent learning: Do different neural networks learn the same representations? *International Conference on Learning Representations*, 2016.

Aleksandar Makelov, George Lange, and Neel Nanda. Towards principled evaluations of sparse autoencoders for interpretability and control. *arXiv preprint arXiv:2405.08366*, 2024.

Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. *arXiv preprint arXiv:2403.19647*, 2024.

Alejandro Martin-Linares and Charles Ling. Attribution-guided distillation of matryoshka sparse autoencoders. *arXiv preprint arXiv:2512.24975*, 2025.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 2022.

Goncalo Paulo and Nora Belrose. Sparse autoencoders trained on the same data learn different features. *International Conference on Learning Representations*, 2026.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in Neural Information Processing Systems*, 2017.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Sidak Pal Singh and Martin Jaggi. Model fusion via optimal transport. *Advances in Neural Information Processing Systems*, 2020.

Xiangchen Song, Aashiq Muhamed, Yujia Zheng, Lingjing Kong, Zeyu Tang, Mona T Diab, Virginia Smith, and Kun Zhang. Position: Mechanistic interpretability should prioritize feature consistency in SAEs. *International Conference on Machine Learning*, 2025.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. *International Conference on Learning Representations*, 2023.

Michael Wang. Enforcing orderedness in sparse autoencoders to improve feature consistency. *Advances in Neural Information Processing Systems*, 2025.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. *International Conference on Learning Representations*, 2024.

A SAE training details

We train 28 SAEs across five experimental conditions: seed stability (5 BatchTopK SAEs at layer 8), layer correspondence (5 SAEs at layers 4/8/12/16/20 with seed 42), checkpoint tracking (5 SAEs), multi-layer seed expansion (4 additional seeds at layers 4 and 12 each), and architecture comparison (5 standard ReLU SAEs at layer 8 with $L_1 = 5.0$, yielding $L_0 \approx 69$). Table 5 reports quality metrics for all SAEs, and Figure 4 visualizes explained variance across conditions. We report reconstruction quality as $1 - \|\hat{x} - x\|^2/\|x\|^2$, the explained variance metric commonly used in prior SAE work (Bricken et al., 2023; Cunningham et al., 2023). The five seed-stability SAEs (used for all main-text analyses) achieve consistent quality: explained variance 0.821 to 0.824, cosine similarity 0.906 to 0.908, and $< 4\%$ dead features.

Table 5: Quality metrics for all 28 trained SAEs. EV = explained variance ($1-\|\hat{x}-x\|^2/\|x\|^2$), Cos = cosine similarity, L0 = mean L0 sparsity, Dead = percentage of dead features. Multi-layer seed SAEs are trained at layers 4 and 12 with seeds {123, 456, 789, 1024} to complement existing seed-42 SAEs. ReLU SAEs use standard L_1 regularization (higher EV but lower cosine similarity than BatchTopK at comparable L_0).

Experiment	SAE	EV	Cos	L0	Dead %
Seed Stability	Seed 42	0.821	0.906	105.2	3.4
	Seed 123	0.824	0.908	105.3	3.8
	Seed 456	0.824	0.907	105.3	3.4
	Seed 789	0.822	0.907	105.1	3.9
	Seed 1024	0.822	0.907	105.1	3.8
Layer Corr.	L4 s42	0.887	0.941	106.6	1.0
	L8 s42	0.823	0.907	105.1	4.7
	L12 s42	0.820	0.905	103.7	9.0
	L16 s42	0.877	0.935	103.0	40.6
	L20 s42	0.890	0.942	105.3	49.1
Multi-layer Seeds	L4 s123	0.886	0.941	106.2	0.9
	L4 s456	0.886	0.941	106.2	0.9
	L4 s789	0.887	0.941	106.7	1.1
	L4 s1024	0.886	0.941	106.4	0.8
	L12 s123	0.821	0.906	104.0	8.1
	L12 s456	0.822	0.907	104.1	8.6
	L12 s789	0.821	0.906	103.9	8.0
	L12 s1024	0.821	0.906	104.0	8.5
ReLU (L8, L_1=5)	Seed 42	0.982	0.828	68.4	26.0
	Seed 123	0.982	0.828	68.9	25.3
	Seed 456	0.982	0.829	69.0	25.7
	Seed 789	0.982	0.828	68.7	25.5
	Seed 1024	0.982	0.828	68.8	25.4
Checkpoint	1K steps	0.339	0.590	100.9	45.5
	10K steps	0.741	0.860	100.6	34.0
	50K steps	0.855	0.924	105.8	0.4
	100K steps	0.869	0.932	105.4	0.3
	143K steps	0.872	0.934	102.1	0.3

B Causal validation by pair

Table 6 reports Cohen's d for the Sinkhorn composite method on each of the 10 seed pairs. Results are highly consistent (CV = 1.8%).

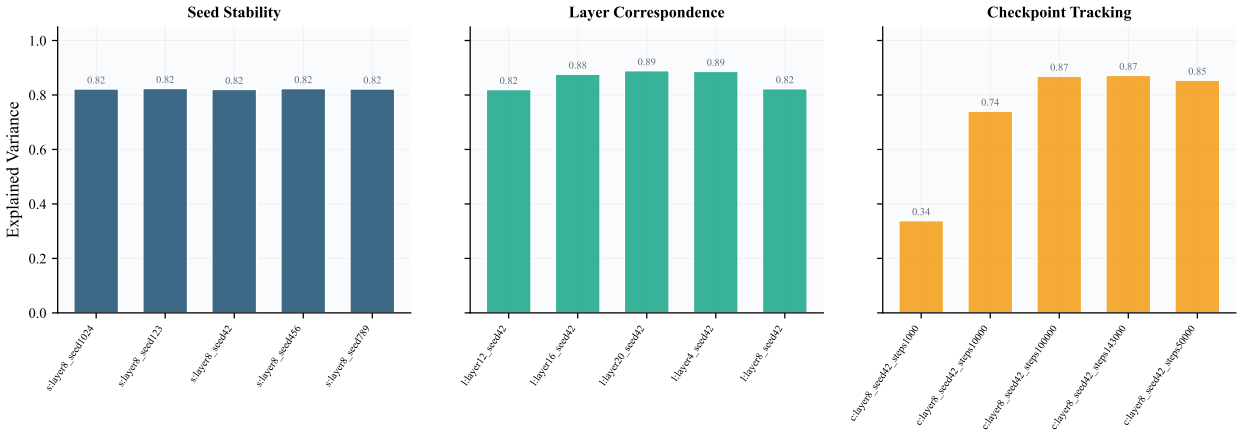

Figure 4: Explained variance for all 28 SAEs across five experimental conditions. Seed-stability and multi-layer-seed SAEs show consistent quality within each layer; checkpoint SAEs show progressive improvement with training steps; ReLU SAEs achieve higher EV than BatchTopK at comparable sparsity.

Table 6: Cohen's d (Sinkhorn composite) for each seed pair.

Seed Pair	Cohen's d
1024 vs 123	1.815
1024 vs 42	1.791
1024 vs 456	1.884
1024 vs 789	1.896
123 vs 42	1.890
123 vs 456	1.852
123 vs 789	1.831
42 vs 456	1.838
42 vs 789	1.862
456 vs 789	1.875
Mean ± std	$\mathbf{1.853 \pm 0.033}$

C Extended ablation results

C.1 OT design ablations

Table 7 reports the effect of removing individual OT components.

Table 7: OT design ablations (mean ± std across 10 seed pairs). "Uniform marginals" uses the full OT pipeline (with dustbin) but replaces activation-mass marginals with uniform ones. "Sinkhorn (uniform)" removes both dustbin and activation-mass weighting, a separate method inspired by Khosla & Williams (2024). Note: "Full OT (default)" is the same method as Sinkhorn composite in Table 1; the slight std difference (0.047 vs. 0.03) reflects a different null-distribution sample size (5,000 vs. 10,000 random draws) and sample vs. population std.

Variant	Matched	Cohen's d	AUROC
Full OT (default)	491	$\mathbf{1.858 \pm 0.047}$	$\mathbf{0.904 \pm 0.010}$
No Jaccard	436	1.864 ± 0.064	0.906 ± 0.013
Uniform marginals	453	1.820 ± 0.041	0.898 ± 0.010
No dustbin	526	1.795 ± 0.057	0.897 ± 0.012
Sinkhorn (uniform)	767	1.702 ± 0.027	0.887 ± 0.005

C.2 Threshold sweep

To evaluate thresholded baselines, we vary the cosine distance threshold $t \in \{0.3, \ldots, 0.9\}$ for greedy cosine and MNN, retaining only matches below threshold (Table 8).

Table 8: Threshold sweep for greedy cosine and MNN (mean across 10 seed pairs).

Method	Matched	Cohen's d	AUROC
Greedy $t = 0.3$	2,663	0.95	0.748
Greedy $t = 0.5$	1,494	1.33	0.825
Greedy $t = 0.7$	490	1.94	0.921
Greedy $t = 0.9$	61	2.65	0.996
MNN $t = 0.3$	2,659	0.96	0.751
MNN $t = 0.5$	1,494	1.33	0.825
MNN $t = 0.7$	490	1.92	0.920
MNN $t = 0.9$	61	2.65	0.996

At high thresholds ($t \geq 0.7$), thresholded greedy and MNN exceed Sinkhorn's full-coverage d (1.85), confirming that selectivity drives part of the gap. However, these operating points retain only ∼3% of features, while Sinkhorn composite retains 7.1% at $d = 1.85$. At matched coverage (∼7%, corresponding roughly to $t \approx 0.6$ to 0.7), Sinkhorn achieves higher fingerprint d than thresholded baselines, though the substitution ranking curves (Section 5.4) show a more nuanced picture.

C.3 Precision-coverage curves

Table 9 reports the full quality/coverage summary across all methods and 10 seed pairs.

Table 9: Quality/coverage summary (mean across 10 pairs). "Full coverage" evaluates all matches; "best d" reports the peak Cohen's d achievable by retaining only the most confident matches.

Method	Coverage	d (full)	Best d	Best cov.
Sinkhorn composite	7.1%	1.85	2.78	0.3%
Sinkhorn cosine	13.6%	1.78	2.71	0.7%
Sinkhorn (uniform)	11.1%	1.71	2.75	0.5%
Greedy cosine	65.9%	0.54	2.26	3.3%
MNN (k=5)	57.5%	0.62	2.33	2.9%

C.4 OT sensitivity analysis

Table 10 summarizes how fingerprint d varies across OT hyperparameter choices.

Table 10: OT sensitivity analysis (representative pair, seed 42 vs. 123; validated on 2 additional pairs). At fixed ε, d varies by $< 21\%$ across all other parameter choices.

Parameter	Values tested	d range
Composite weights	8 combinations	[1.54, 1.89]
ε	{0.01, 0.02, 0.05, 0.1, 0.2}	[1.20, 2.86]
Dustbin cost	{auto, 0.3, 0.5, 0.7, 0.9}	[1.77, 2.05]
Marginal mode	{mass, uniform, activity}	[1.77, 1.89]

C.5 Held-out validation

Table 11 reports the held-out causal validation results, comparing matches computed on the original corpus against evaluation on a disjoint held-out shard.

Table 11: Held-out causal validation (mean across 10 seed pairs). Matches are computed on the original corpus; causal fingerprints and substitution tests use a disjoint held-out shard (256 sequences). Held-out fingerprints use the same evaluation operator as the main experiment (full per-feature ablation with forward passes), ensuring directly comparable effect sizes.

Method	d	AUROC	Sub. recovery R		
			top-50	top-100	top-200
Sinkhorn composite	3.55	0.976	0.88	0.83	0.78
Sinkhorn cosine	3.21	0.969	0.90	0.86	0.79
Greedy cosine	1.00	0.743	0.90	0.86	0.79

C.6 GPT-2 Small substitution ranking

Table 12 reports the full substitution ranking curve for GPT-2 Small (layer 6, 5 seeds, 10 pairs), complementing the prose summary in Section 5.9.

Table 12: Substitution recovery R on GPT-2 Small (layer 6, mean across 10 seed pairs). The same pattern holds: Sinkhorn (uniform) matches greedy at the top and maintains quality in the tail.

Method	top-10	top-50	top-100	top-500	top-1000	top-2000	AUSQC
Greedy cosine	.79	.68	.64	.48	.37	.30	1.46
MNN (k=5)	.79	.68	.64	.48	.37	.30	1.46
Sinkhorn composite	.31	.30	.30	.33	.33	.33	0.83
Sinkhorn cosine	.33	.34	.34	.35	.35	.35	0.91
Sinkhorn (uniform)	**.79**	**.68**	**.64**	**.52**	**.52**	**.52**	**1.65**

C.7 Architecture comparison: AUSQC summary

Figure 5 shows the AUSQC for each method and SAE architecture. On BatchTopK, Sinkhorn (uniform) achieves the highest AUSQC (2.06) by maintaining tail quality; on ReLU, greedy cosine and MNN achieve the highest AUSQC (2.69), approaching the theoretical maximum (2.70). The ranking inversion is clearly visible: the method advantage on BatchTopK becomes a disadvantage on ReLU.

C.8 Hard negative controls (full results)

Table 13 reports causal validation against five types of confound-matched hard negatives.

The five hard-negative types are: (1) cosine-neighborhood negatives (features sharing geometric similarity but not matched); (2) activation-mass-matched negatives (±10% mass, low cosine); (3) Jaccard-neighborhood negatives (similar support overlap, different directions); (4) permuted top-k negatives (matched tier preserved, correspondence broken); and (5) cross-seed decoys (different seed pair, similar geometric profile). Greedy cosine achieves higher d on random, cosine-neighborhood, mass-matched, and cross-seed decoy controls, while Sinkhorn wins on Jaccard and permuted top-k.

C.9 Activity stratification (full results)

Table 14 reports the fingerprint screening stratification; Table 15 reports the substitution stratification.

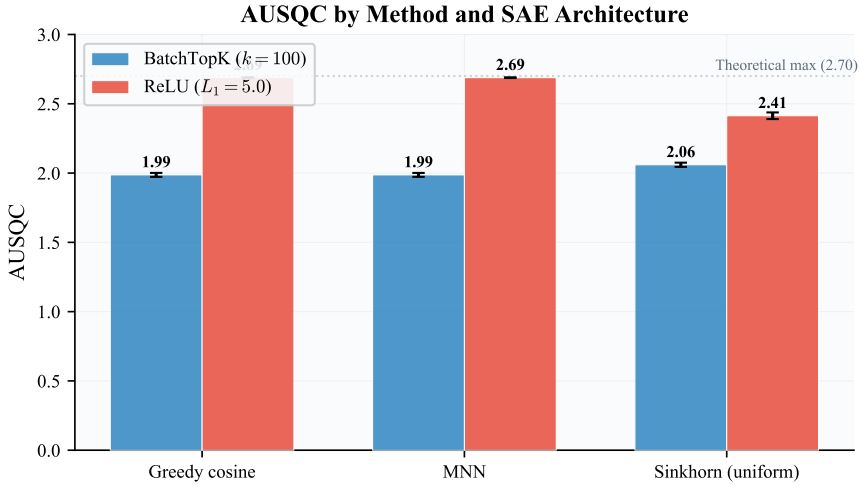

Figure 5: AUSQC by method and SAE architecture. On BatchTopK, Sinkhorn (uniform) achieves the highest AUSQC by maintaining tail quality; on ReLU, greedy/MNN dominate, approaching the theoretical max of 2.70. Error bars: ±1 std across 10 seed pairs.

Table 13: Causal validation against hard negatives (mean across 10 seed pairs, full-coverage operating point). All hard-negative types substantially reduce effect sizes compared to random. Δd = Sinkhorn composite minus greedy cosine. No hard-negative Δd reaches statistical significance (seed-level bootstrap 95% CI includes zero for all types).

Negative type	Sinkhorn d	Greedy d	Δd	Sig?
Random (reference)	2.04	2.25	−0.21	n/a
Cosine neighborhood	1.67	1.78	−0.11	No
Mass-matched	1.38	1.43	−0.05	No
Jaccard neighborhood	1.44	1.39	+0.05	No
Permuted top-k	1.73	1.66	+0.07	No
Cross-seed decoy	2.02	2.20	−0.18	No

Table 14: Fingerprint screening stratified by feature activity level (mean across 10 seed pairs). "High", "Medium", "Low" refer to terciles of total activation mass among fingerprinted features. Greedy cosine shows 2.2× attenuation from high to low; Sinkhorn's low-stratum estimate ($n = 6$) is unreliable.

Method	Stratum	Coverage	d	AUROC	n
Sinkhorn composite	High	34%	1.40	0.87	393
	Medium	33%	1.32	0.80	41
	Low	33%	1.87	0.88	6
Greedy cosine	High	34%	0.78	0.73	1570
	Medium	33%	0.37	0.60	696
	Low	33%	0.36	0.58	594

C.10 Fingerprint robustness

Causal fingerprints are noisy as per-feature signatures (mean cosine similarity ∼0.29 across hyperparameter settings), but the downstream method ranking they induce is stable (Table 16). Causal validation d remains $>$ 3.2 across all $n_{\text{sequences}}$ and n_{logits} settings tested, so the protocol's conclusions about which matcher performs better are not sensitive to these choices, even though individual fingerprint vectors vary substantially.

Table 15: Substitution recovery R stratified by activation mass tercile (greedy cosine top-2000 matched pairs, mean ± std across 10 seed pairs). Attenuation is only 1.3× from high to low, compared to 2.2× for the fingerprint screening metric (Table 14). All terciles remain far above random ($R \approx 0.002$).

Stratum	R (mean ± std)	n pairs	Fraction of top-2000
High	0.49 ± 0.004	1,210	60.5%
Medium	0.41 ± 0.004	582	29.1%
Low	0.38 ± 0.006	197	9.9%
Random	0.002	n/a	n/a

Table 16: Fingerprint robustness to hyperparameters.

Setting	Mean cosine	n_{valid}
$n_{\text{seq}} = 32$	0.297	1,692
$n_{\text{seq}} = 64$	0.292	2,571
$n_{\text{seq}} = 128$	0.290	3,820
$n_{\text{seq}} = 256$ (ref)	0.287	5,060
$n_{\text{logits}} = 10$	0.468	5,060
$n_{\text{logits}} = 25$	0.358	5,060
$n_{\text{logits}} = 50$ (ref)	0.287	5,060
$n_{\text{logits}} = 100$	0.332	5,060

C.11 Seed-level bootstrap confidence intervals

Table 17 reports 95% bootstrap confidence intervals using seed-level resampling (10,000 resamples of 5 seeds with replacement, then computing all $\binom{5}{2}$ pairs from each resample). This correctly accounts for the non-independence of overlapping seed pairs (effective df ≈ 4). Pairs where both resampled seeds are identical are excluded (they would have zero cost).

Table 17: Seed-level bootstrap 95% CIs for causal validation metrics (10,000 resamples).

Method	Cohen's d [95% CI]	AUROC [95% CI]
Sinkhorn composite	1.85 [1.82, 1.89]	0.905 [0.895, 0.915]
Sinkhorn cosine	1.76 [1.73, 1.78]	0.897 [0.892, 0.902]
Greedy cosine	1.34 [1.30, 1.38]	0.825 [0.816, 0.832]
MNN (k=5)	0.63 [0.61, 0.65]	0.665 [0.661, 0.670]
Sparse Hungarian	0.60 [0.59, 0.62]	0.655 [0.651, 0.660]
SVCCA	0.28 [0.27, 0.30]	0.566 [0.563, 0.571]
Procrustes	0.00 [−0.02, 0.02]	0.501 [0.495, 0.506]

Table 18 reports bootstrap CIs on the *difference* in Cohen's d between key method pairs. If the 95% CI excludes zero, the advantage is statistically significant under seed-level resampling.

Table 18: Bootstrap 95% CIs on method-difference Δd (seed-level resampling). Bold = CI excludes zero.

Comparison	Δd	95% CI
Sinkhorn composite − Greedy cosine	+0.51	[0.475, 0.557]
Sinkhorn composite − MNN	+1.23	[1.185, 1.262]
Sinkhorn cosine − Greedy cosine	+0.41	[0.379, 0.454]
Greedy cosine − MNN	+0.71	[0.674, 0.744]

Ranking stability analysis shows that Sinkhorn composite holds the top rank in 100% of bootstrap resamples, followed by Sinkhorn cosine (100% rank 2) and greedy cosine (100% rank 3). The top-3 ranking is perfectly stable across all 10,000 resamples.

D Additional figures and robustness analyses

This section collects supplementary figures and tables referenced from the main text.

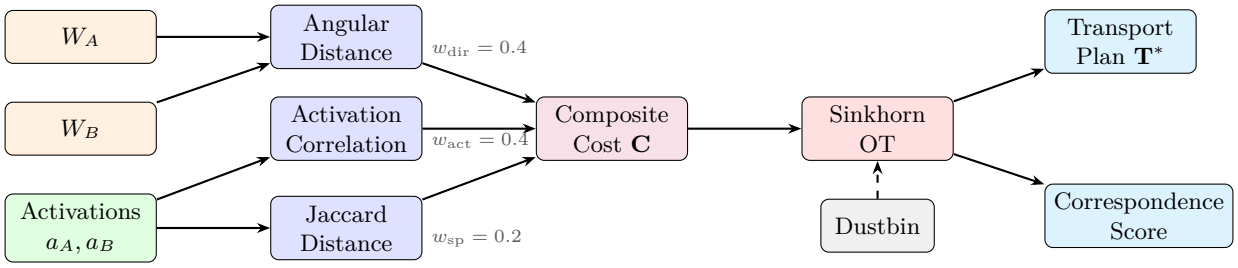

Figure 6: OT matching pipeline overview. Decoder weights and sparse activations are combined into a composite cost matrix, then solved via dustbin-augmented Sinkhorn OT to produce a transport plan and correspondence score.

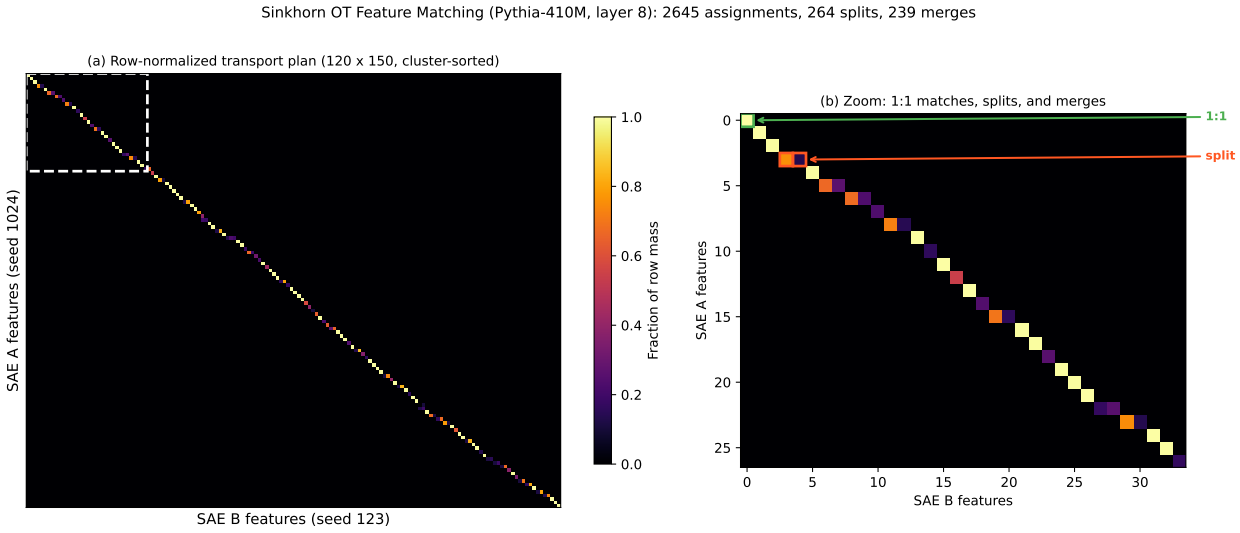

Figure 7: Sinkhorn transport plan for one seed pair (Pythia-410M, seeds 1024 vs. 123). **(a)** Row-normalized transport plan for the top 120 features, sorted by cluster assignment, revealing near-diagonal (permutation-like) structure with off-diagonal entries for splits and merges. **(b)** Zoomed view of the dashed region. Annotations highlight a 1:1 match (green) and a split where one SAE A feature maps to two SAE B features (red).

Figure 9 illustrates a single substitution case study. Figure 10 and Figure 11 show correspondence score heatmaps for the layer-correspondence and seed-stability experiments, respectively.

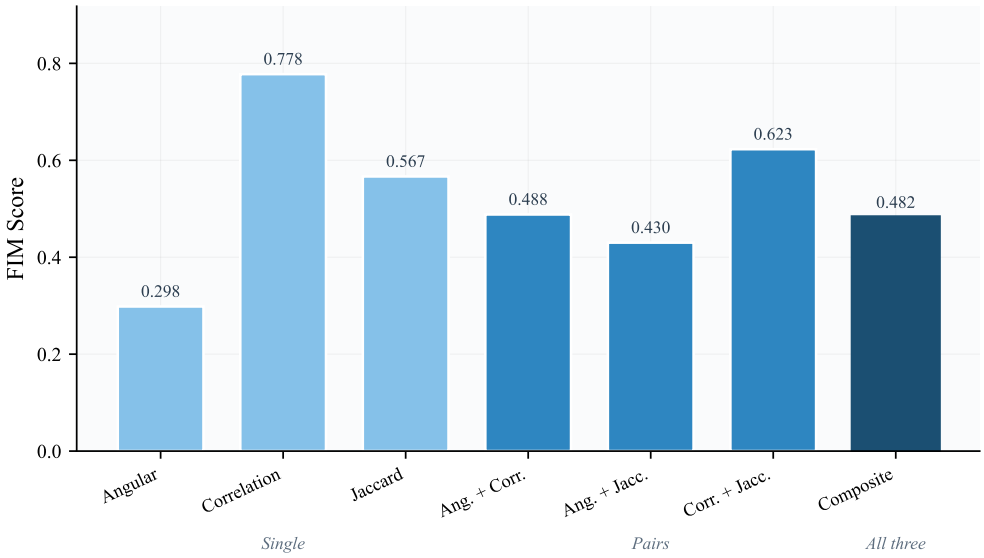

Figure 8: Correspondence score under different cost component combinations. The composite cost balances geometric (angular) and behavioral (correlation) similarity signals.

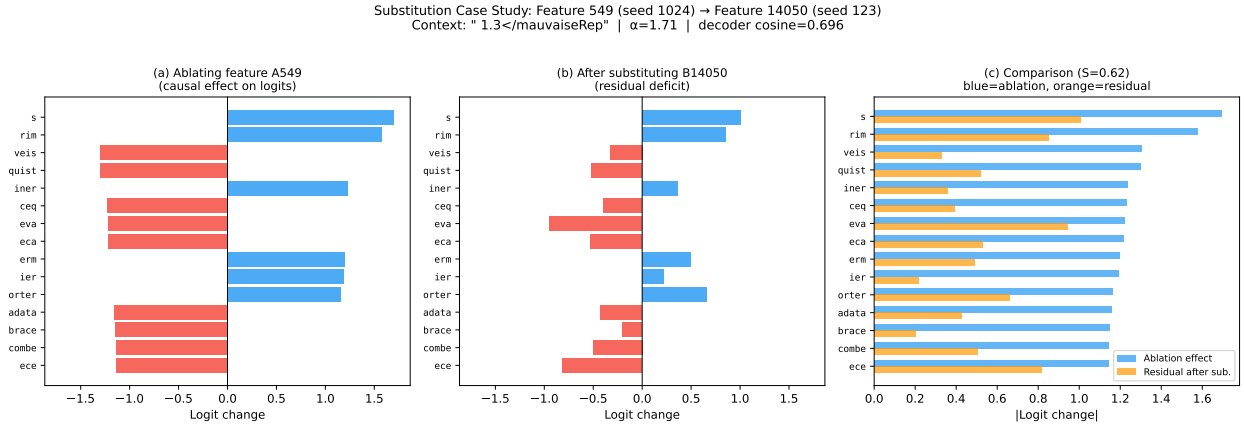

Figure 9: Substitution case study for a single context position (features 549 and 14050, decoder cosine = 0.70). **(a)** Logit changes from ablating feature A (the causal effect). **(b)** Logit changes after substituting feature B with fitted scaling; the effect is substantially recovered. **(c)** Absolute comparison showing the ablation effect (blue) and residual deficit (orange) for the top-15 most affected tokens. Aggregate across all active positions: $R = 0.57$, $\alpha = 0.96$.

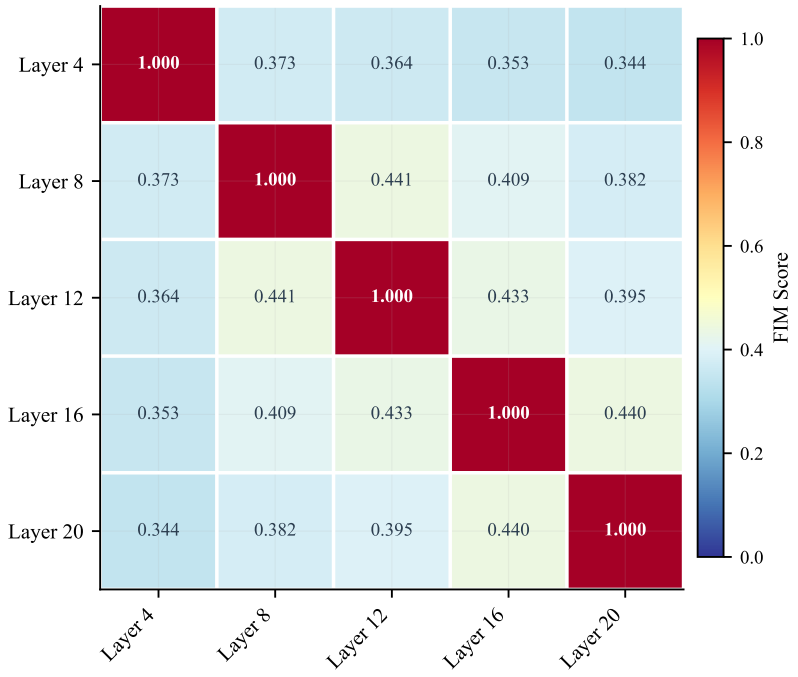

Figure 10: OT correspondence score heatmap for layer correspondence experiment. Adjacent layers show higher feature similarity, as expected, with correspondence decreasing as layer distance increases.

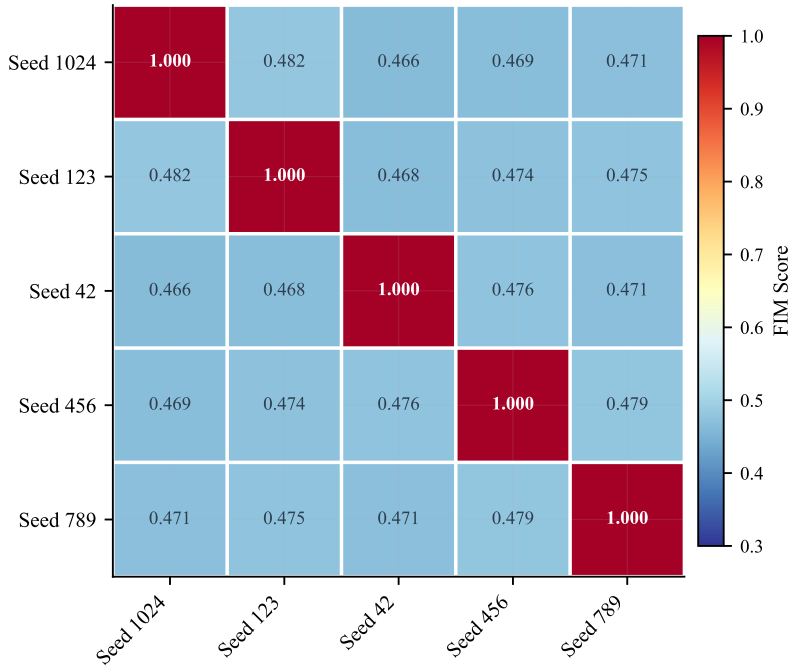

Figure 11: OT correspondence score heatmap for the seed stability experiment. Off-diagonal scores (∼0.48) reflect moderate feature correspondence across seeds.