CERTIFIED CROSS-DICTIONARY FEATURE MATCHING VIA CO-ACTIVATION GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Feature matching across dictionaries is central to tracking concepts across layers, models, and fine-tunes, but common matchers return correspondences without a certificate. For non-negative codes evaluated on shared samples, the cross co-activation relation $K_{12}(a, b) = \max_i \min(Z_1[i, a], Z_2[i, b])$ defines a filtered relation whose two feature-side Dowker complexes have identical persistent homology for every pair of codes, even when widths and ambient activation spaces differ. Restricting the induced maps to mutual matches yields a certified partial correspondence with zero round-trip error; its metric distortion δ_{mut} is reported together with coverage. Across twelve synthetic width–sparsity configurations, every committed match is correct (precision 1.000) and $\delta_{\mathrm{mut}} = 0$; additional samples increase coverage with an approximately inverse-linear law in samples per feature. Released GPT-2 dictionaries occupy a different regime: max–min scores saturate with sample count, so coverage must be calibrated empirically rather than extrapolated from the synthetic law. At 8192 tokens, adjacent-layer coverage falls from 0.563 at layers $0 \to 1$ to a minimum of 0.134 at $9 \to 10$, and agreement with correlation and decoder-cosine matchers falls with it from $1 \to 2$ onward; the embedding transition $0 \to 1$ is the one place where the relational matcher commits most and the baselines agree with it least. The method converts feature matching from an unqualified assignment into an auditable correspondence whose precision, coverage, and geometric distortion are explicit.

1 INTRODUCTION

Comparing learned feature sets is now a routine part of mechanistic analysis: features are tracked across depth, aligned across models, and followed before and after fine-tuning. Decoder cosine, activation correlation, crosscoders, and distributional matching all provide useful assignments, but the assignment itself rarely comes with a quantitative certificate in the intrinsic geometry of both dictionaries.

For non-negative codes evaluated on the same samples, cross co-activation defines a weighted relation directly between the two feature sets. Filtered Dowker duality gives this relation identical persistent homology on both sides for every pair of codes, without requiring a shared activation basis, equal dictionary widths, or a margin assumption. The relation also induces maps in both directions. Restricting to mutual matches removes round-trip error exactly and turns the remaining distortion into a certificate for a partial feature correspondence.

On controlled transformations, every committed match is correct across the full width–sparsity sweep, while unresolved correspondence appears explicitly as reduced coverage. On released GPT-2 dictionaries, the same diagnostic reveals a distinct operating regime in which max–min similarity saturates with sample count and adjacent-layer correspondence becomes scarcer with depth across ten of eleven pairs. Agreement with independent correlation and decoder-cosine matchers declines in parallel, providing an external signal of the same representational change.

Contributions.

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.

- **Cross-dictionary duality.** A max–min co-activation relation gives exact filtered Dowker duality for every pair of non-negative codes evaluated on shared samples, including codes with different widths and ambient spaces.
- **Certified partial matching.** Mutual matches have exact round trips; their metric distortion δ_{mut} and coverage provide explicit precision and commitment measures for the returned correspondence.
- **Precision-one recovery on controlled transformations.** Across twelve width–sparsity settings, every committed permutation match is correct and $\delta_{\mathrm{mut}} = 0$, with coverage controlled by the sample budget.
- **Layerwise evidence, and its model dependence.** In GPT-2, real dictionaries show sample-count saturation and a 4.2-fold decline in adjacent-layer coverage from the embedding transition to its minimum in the deep half, with agreement from independent matchers falling over the same span. On Gemma-2-2B the profile is instead U-shaped with an interior minimum at 42% of depth, so the location of the minimum, not a monotone decline, is what transfers.

2 THE CROSS RELATION

Definition 1. For non-negative codes $Z_1 \in \mathbb{R}_+^{m\times n_1}$, $Z_2 \in \mathbb{R}_+^{m\times n_2}$ on shared samples, the *cross co-activation*

$$K_{12}(a,b) = \max_{i\le m} \min\big(Z_1[i,a],\, Z_2[i,b]\big) \tag{1}$$

defines a filtered relation on $J_1 \times J_2$ with cost $-K_{12}$.

Theorem 2. *The Dowker filtrations of* (1) *on J_1 and on J_2 have isomorphic persistent homology in every degree, for every pair of codes. In particular the two feature spaces have the same Dowker persistence diagram, so any discrepancy observed between them is an artefact of the chosen point-cloud metric, not of the dictionaries.*

This is filtered Dowker duality applied to the cross relation; no property of Z_1, Z_2 is used. It is the cross-layer analogue of the observation that a single dictionary's row and column Dowker topologies always agree.

3 THE INDUCED CORRESPONDENCE AND ITS CERTIFICATE

The relation gives a map in each direction: $\mu(a) = \arg\max_b K_{12}(a,b)$ and $\nu(b) = \arg\max_a K_{12}(a,b)$. Let d_1, d_2 be the mean-normalised Dowker metrics on the two feature sets, i.e. from the co-activation kernels of $Z_1^\top$ and $Z_2^\top$. A single map cannot exhibit an interleaving: four conditions must hold, two on the maps and two on the round trips.

Proposition 3 (exact certificate for the total correspondence)**.** *With the convention* $\mathrm{VR}_t = \{\mathrm{diam} \le 2t\}$*, the pair (μ,ν) exhibits a δ-interleaving of* $\mathrm{VR}(J_1,d_1)$ *and* $\mathrm{VR}(J_2,d_2)$ *for*

$$\begin{aligned}\delta_{\mathrm{cross}} = \tfrac{1}{2}\max\Big\{ &\max_{a,a'}\big[d_2(\mu a,\mu a') - d_1(a,a')\big]_+,\ \max_{b,b'}\big[d_1(\nu b,\nu b') - d_2(b,b')\big]_+,\\ &\max_a d_1(a,\nu\mu a),\ \max_b d_2(b,\mu\nu b)\Big\},\end{aligned} \tag{2}$$

and this is the least such δ for this pair of maps. No hypothesis on Z_1, Z_2 is used.

Proof. If $\mathrm{diam}_1(\sigma) \le 2t$ then $d_2(\mu a,\mu a') \le d_1(a,a') + 2\delta \le 2(t+\delta)$ for $a,a' \in \sigma$, so μ maps $\mathrm{VR}_t(J_1)$ into $\mathrm{VR}_{t+\delta}(J_2)$; symmetrically for ν. For the round trip, with $d_1(a,\nu\mu a) \le 2\delta$ and $d_1(a,a') \le 2t$ we get $d_1(a,\nu\mu a') \le d_1(a,a') + d_1(a',\nu\mu a') \le 2t + 2\delta \le 2(t+2\delta)$, so $\nu\mu$ is contiguous to the inclusion $\mathrm{VR}_t(J_1) \hookrightarrow \mathrm{VR}_{t+2\delta}(J_1)$; symmetrically for $\mu\nu$. Each of the four terms is the exact slack of its condition, so no smaller δ satisfies all four. □

Table 1: The cross certificate on the transformation benchmark, ten seeds, $n = 64$, $k = 6$, transformation fraction 0.25, codes max-normalised. The total certificate (2) is near the diameter in every case, including exact permutations; the mutual certificate (3) is exactly zero on the structure-preserving maps and graded on the rest. "recovery" is accuracy on the mutual set where ground truth exists.

	total	mutual set				
case	δ_{cross}	δ_{mut}	q_{95}	mean	coverage	recovery
permute	1.094	**0.000**	0.000	0.0000	0.948	1.000
scale $\times 3.7$	1.627	**0.000**	0.000	0.0000	0.936	1.000
noise 0.05	1.397	1.114	0.437	0.1407	0.527	0.826
split	1.483	0.921	0.457	0.2289	0.722	—
merge	1.587	1.385	0.283	0.0720	0.959	—
unrelated	1.883	1.668	0.735	0.2952	0.447	—

Mutual matching converts the total bound into an informative partial certificate. Proposition 3 is exact but it is a maximum over all pairs, so one mismatched feature drives it to roughly the diameter. We verified both halves on 30 code pairs (permutation, rescaling, splitting, merging, unrelated; $k \in \{4, 8\}$, three seeds): all four conditions hold at δ_{cross} in $30/30$ and at least one fails at $0.98\,\delta_{\mathrm{cross}}$ in $30/30$. On an exact permutation of a k=6 dictionary the total value is 1.094 against a diameter of about 1.4, showing that a maximum over a total map is too coarse for correspondence. Restricting the certificate to mutually supported matches isolates the part of the relation on which the map actually commits.

Definition 4 (mutual correspondence)**.** Let $S = \{a : \nu\mu a = a\}$ be the mutual-match set. On S, μ is a bijection onto $\mu(S)$, both round-trip terms of (2) vanish identically, and the certificate reduces to the distortion of a bijection,

$$\delta_{\mathrm{mut}} = \tfrac{1}{2} \max_{a,a' \in S} \big| d_1(a, a') - d_2(\mu a, \mu a') \big|, \qquad \mathrm{cov} = \frac{|S|}{\min(n_1, n_2)}. \tag{3}$$

A partial correspondence with a small certificate and a stated coverage is a more useful object than a total one with a trivial certificate, and this is the form we recommend reporting.

Per-code normalization removes scale ambiguity. If $Z_2 = cZ_1$ with $c > 1$ then $\min(Z_1[i, a], cZ_1[i, b]) = Z_1[i, a]$ for many b, the arg-max ties, and μ is arbitrary: on a rescaled permutation the mutual fraction collapses to 0.17 and δ_{cross} rises to 1.39. Dividing each code by its own maximum before forming K_{12} removes this exactly — the rescaled case then certifies at $\delta_{\mathrm{mut}} = 0.000$ with 100% recovery (Table 1). We normalise throughout. This is a property of the max–min relation and not of the duality theorem, which is invariant to any common monotone reparametrisation of the filtration.

Table 1 separates three properties of the partial correspondence.

Mean distortion resolves sparse local failures. On merging the maximum distortion is 1.385 but the mean is 0.072 and the 95th percentile 0.283: a handful of pairs carry the maximum. Reporting only a max-based certificate would rank merging with the unrelated pair, which is wrong by every other measure.

Coverage measures how strongly the relation commits. It separates related from unrelated dictionaries on its own — 0.94–0.96 for permutation, rescaling and merging against 0.45 for an independently drawn dictionary of the same sparsity — without reference to any metric, and it is the cheapest of these quantities to compute.

Dense noise reduces commitment through max–min saturation. Additive Gaussian noise at $\sigma = 0.05$ costs more coverage (0.53) than merging a quarter of the features (0.96), because $\min(Z_1, Z_2)$ picks up the noise floor at every sample and the $\max_i$ then selects a spurious sample. The construction is a statement about sparse non-negative codes; on dense ones the relation degrades before the matching does.

Table 2: Recovery of a permuted dictionary on the mutual set, five seeds, $m = 1000$ samples. Recovery is exactly 1 and δ_{mut} exactly 0 at every width and sparsity; the cost of scale appears entirely as lost coverage, and is independent of k.

n	k	coverage	recovery	δ_{mut}	s / pair
64	4	0.969	1.000	0	0.65
64	32	0.963	1.000	0	0.67
128	4	0.936	1.000	0	0.74
128	32	0.906	1.000	0	0.52
256	4	0.852	1.000	0	1.36
256	32	0.865	1.000	0	1.64

Table 3: Exact recovery of the intended correspondence, ten seeds, $n = 64$, $k = 6$, transformation fraction 0.25. "centre" uses the generating bump centres and is an oracle for any purely geometric matcher.

transformation	Dowker (ours)	correlation	sliced Wasserstein	centre (oracle)
permute	0.986	**1.000**	**1.000**	1.000
split	0.872	0.847	**0.997**	1.000
merge	**0.983**	0.903	0.717	0.644
absorb	**0.875**	0.772	0.836	1.000
noise	0.986	**1.000**	**1.000**	1.000

3.1 Precision-one matching with sample-controlled coverage

A partial matcher separates two errors: incorrect commitments and missing commitments. In the controlled sweep, only the second occurs.

Across $n \in \{64, 128, 256\}$ and $k \in \{4, 8, 16, 32\}$ — twelve configurations, five seeds each — recovery on the mutual set is 1.000 in every one, and δ_{mut} is exactly 0 (Table 2). Every match the relation commits to is the intended one. What degrades with width is coverage, from 0.97 at $n = 64$ to 0.85 at $n = 256$, and it degrades not because the dictionary is wide but because the sample budget is fixed: at $n = 256$, $k = 8$,

m/n	1.0	2.0	3.9	7.8	15.6	31.2
coverage	0.562	0.727	0.839	0.914	0.967	0.983
recovery	1.000	1.000	1.000	1.000	1.000	1.000

and a log–log fit gives

$$1 - \mathrm{cov} \;=\; 0.51\,(m/n)^{-0.96}, \qquad R^2 = 0.985, \tag{4}$$

i.e. inverse-linear in samples per feature to within 4% in the exponent. Two ties in the arg-max are what break mutuality, and the probability that a feature's true partner is beaten by a spurious one falls like the reciprocal of the number of samples per feature.

The practical form of (4) is a sample budget: $m \gtrsim 5n$ for 90% coverage, $10n$ for 95%, $51n$ for 99%. This is a *precision-one* matcher whose recall is a purchasable quantity, which is the opposite of the usual trade and is what makes it safe to use for auditing: a reported correspondence never needs to be second-guessed, only supplemented.

Cost. Forming K_{12} is $O(mn_1n_2)$ in the dense form and $O(m\,k_1k_2)$ on k-sparse codes, since only co-active pairs contribute; the two Dowker metrics are $O(mn_i^2)$ and $O(mk_i^2)$ respectively. Measured end-to-end on the dense path, a pair of $n = 256$ dictionaries over $m = 1000$ samples certifies in 1.4–1.6 s on one CPU core, and the sparse path is what makes width-2^{15} dictionaries feasible.

The matchers have complementary failure modes. The relational matcher dominates on **merging** — 0.983 against 0.717 for the distributional matcher and 0.644 for the geometric oracle — because a merged feature has neither a single centre nor a coherent distribution, but it does have a well-defined

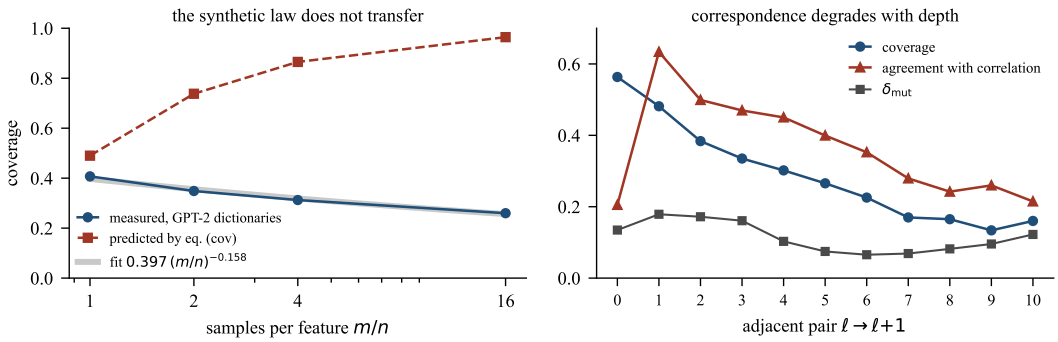

Figure 1: Left: the coverage law fitted on synthetic codes predicts coverage rising with sample count; on released GPT-2 dictionaries it falls, because the max–min relation saturates on codes whose features co-activate heavily. The law is fitted on synthetic codes and inverts on real ones. Right: adjacent-layer correspondence in GPT-2. Coverage falls by a factor of four from the embedding to the top of the network, and the relational and correlation matchers stop agreeing with each other in the same place.

maximal overlap with each of its sources. It also wins on absorption. The distributional matcher dominates on **splitting** (0.997 against 0.872), which is equally natural: a split preserves the shape of the support, so optimal transport tracks it, while max-overlap is ambiguous between the two halves. On the trivial transformations the relational matcher is marginally worse.

Running relational and distributional matchers together makes disagreement informative: merging and absorption favor relational overlap, whereas splitting favors distributional transport. The certificate quantifies the geometry of the relational matches that are returned.

Complementary behavior persists across sparsity levels. Repeating the sweep at $k = 3$ and $k = 12$ changes almost nothing about the two methods that carry a guarantee, and quite a lot about the one that does not:

	merge		split		absorb	
k	Dowker	sliced W	Dowker	sliced W	Dowker	correlation
3	0.983	0.722	0.883	1.000	0.875	0.886
6	0.983	0.717	0.872	0.997	0.875	0.772
12	0.983	0.716	0.867	0.997	0.875	0.759

The 26-point margin on merging and the corresponding margin on splitting are both invariant to k, while activation correlation — a second-moment statistic — degrades from 0.886 to 0.759 on absorption as the code gets denser.

The relational matcher reaches 0.986 on the trivial transformations while remaining invariant to feature position. On absorption, the distributional matcher attains lower distortion (0.035 versus 0.060) but recovers fewer of the intended pairs, showing that metric consistency and recovery accuracy are distinct evaluation axes. We therefore pair the distortion certificate with coverage and recovery statistics.

4 Sample-count saturation and layerwise feature correspondence in GPT-2

We evaluate released GPT-2 dictionaries (jbloom, width 2^{15}, the 1024 most active alive features per layer) on a fixed token stream across all eleven adjacent layer pairs and a sweep of token budgets. Real dictionaries enter a saturation regime that reverses the synthetic sample-count trend.

Real dictionaries show sample-count saturation. Equation (4), fitted on synthetic TopK codes, predicts coverage rising to 0.96 by $m/n = 16$. On real dictionaries it *falls*:

Table 4: Adjacent-layer correspondence in GPT-2, $m = 8192$, the 1024 most active features per layer. "agreement" is the fraction of features on which the relational matcher and the baseline return the same partner. Coverage falls by a factor of 4.2 from the embedding to its minimum at $9 \to 10$ and by 3.5 to the final pair; agreement falls with it from $1 \to 2$ onward. The $0 \to 1$ pair is an outlier in the opposite direction: the relational matcher commits on more features there than anywhere else while the two baselines agree with it least.

pair	coverage	δ_{mut}	mean distortion	agree (cosine)	agree (corr.)
$0 \to 1$	**0.563**	0.135	0.0015	0.076	0.205
$1 \to 2$	0.481	0.179	0.0081	**0.540**	**0.634**
$2 \to 3$	0.384	0.172	0.0082	0.438	0.499
$3 \to 4$	0.335	0.161	0.0061	0.409	0.470
$4 \to 5$	0.302	0.103	0.0026	0.397	0.450
$5 \to 6$	0.266	0.075	0.0034	0.354	0.399
$6 \to 7$	0.226	**0.065**	0.0039	0.302	0.353
$7 \to 8$	0.170	0.069	0.0074	0.247	0.279
$8 \to 9$	0.165	0.082	0.0070	0.223	0.242
$9 \to 10$	**0.134**	0.096	0.0093	0.224	0.260
$10 \to 11$	0.160	0.122	0.0097	0.188	0.215

m/n	1	2	4	16
coverage, measured	0.407	0.349	0.312	**0.260**
coverage, predicted by (4)	0.490	0.738	0.865	0.964

The measured trend is $\mathrm{cov} = 0.397\,(m/n)^{-0.158}$ ($R^2 = 0.984$), so the real-dictionary exponent has the opposite sign from the positive synthetic sample-budget law.

The mechanism follows from the max–min construction: K_{12} is a *maximum* over samples, so adding samples can only raise it, for every pair. On synthetic codes with near-disjoint random supports that saturation is slow and extra samples mostly break ties in favour of the true partner. Real dictionary features co-activate heavily, so K_{12} saturates toward its supremum for many pairs at once, the arg-max becomes less decisive, and mutual matches are lost. Coverage therefore has an empirical optimum in sample count whose location depends on code co-activation density. A sample-count sweep directly calibrates this operating point for real dictionaries.

The committed matches remain metrically tight — mean distortion 0.001–0.004 across the sweep, and 0.0061 averaged over all eleven layer pairs, against a diameter of order one. Coverage directly records the loss of mutual commitments as saturation increases.

The depth profile. At a fixed budget of 8192 tokens, adjacent-pair coverage falls with depth over ten of the eleven pairs, from 0.563 to a minimum of 0.134 at $9 \to 10$, and recovers slightly to 0.160 at the final pair $10 \to 11$. The three matchers stop agreeing with each other over the same span.

Averaging, layers 0–3 give coverage 0.441 and agreement 0.452; layers 8–10 give 0.153 and 0.239. So in GPT-2 feature identity between adjacent layers becomes harder to establish with depth by all three methods simultaneously, and they corroborate each other least exactly where each is individually least willing to commit. That is the correlation an auditor most needs to know about: in the deep half of GPT-2, a reported cross-layer feature correspondence is both scarcer and less independently supported.

The certificate is smallest in the middle of the network ($\delta_{\mathrm{mut}} = 0.065$ at $6 \to 7$) and larger at both ends (0.135 at $0 \to 1$, 0.122 at $10 \to 11$), so the features that *do* correspond correspond most rigidly in the middle layers.

Coverage tracks layer distance. Coverage should fall as the two layers get further apart, and it does: over five long-range pairs it averages 0.127 against 0.290 for adjacent pairs, and δ_{mut} rises correspondingly (0.136–0.280 against 0.065–0.179). The extreme case is the whole network, $0 \to 11$: coverage 0.089, and the relational matcher agrees with decoder-cosine on 0.7% of features. Agreement falls from roughly half the features near the bottom of the network to almost none across

the full depth, so cross-matcher disagreement tracks representational separation rather than behaving as unstructured noise.

The profile covers one model, one dictionary family and one token stream, at the most active 1024 features per layer. The inversion of (4) is a sign change rather than a magnitude one, so it does not depend on those choices; the depth profile is a measurement of these dictionaries, and the same three numbers — coverage, δ_{mut}, cross-matcher agreement — are what to compute on any other pair.

The depth profile is not monotone across models. The GPT-2 profile above declines over ten of its eleven pairs. That is not a law. We repeat the measurement on Gemma Scope (Lieberum et al., 2024b), a JumpReLU suite for Gemma-2-2B (26 layers, width 2^{14}), under the same protocol as the GPT-2 table: for each layer the released sparsity level whose realised mean support is closest to 32, the 1024 most active alive features, and $m = 8192$ tokens with the initial token excluded. Six consecutive pairs spanning the network give coverage 0.231 ($2 \to 3$), 0.156 ($6 \to 7$), 0.055 ($10 \to 11$), 0.128 ($14 \to 15$), 0.165 ($18 \to 19$) and 0.296 ($22 \to 23$).

The profile is U-shaped. It falls to a minimum at 42% of depth and then rises, ending 28% above where it started, so coverage in Gemma-2-2B does not decline with depth. Sparsity does not explain the shape: four of the six pairs use an identical L_0 on both sides, all use 1024 features on $m \approx 8000$ shared tokens, and pairs at nearly the same L_0 span the whole range of coverage ($L_0 \approx 39$–41 gives 0.055, $L_0 = 40$ gives 0.165, $L_0 = 38$ gives 0.296).

What survives is weaker and more useful. Adjacent-layer correspondence has an interior minimum whose location is model-specific: GPT-2 reaches its minimum at 82% of depth with one upward pair after it, Gemma-2-2B reaches its at 42% with a sustained rise. The distortion agrees on the qualitative point in both models — δ_{mut} is smallest exactly at the coverage minimum in Gemma-2-2B (0.043 at $10 \to 11$) and smallest in the middle in GPT-2 — so the features that do correspond correspond most rigidly near the minimum. An auditor should locate that minimum for the model at hand rather than assume that deeper means scarcer.

5 Related work

Four comparisons, not one. A non-negative code invites several topological comparisons that are easy to conflate, and they are not interchangeable. (i) The row and column point clouds of a single code can be compared under Vietoris–Rips filtrations, where agreement is a property to be certified because it can fail. (ii) The same two sides can be compared in the relational geometry of the code itself, where Dowker duality makes the comparison exact for every matrix and no certificate is needed. (iii) A code can be compared against the data it encodes, which asks whether the co-activation geometry tracks the activation geometry. (iv) Two codes on shared samples can be compared against each other, with no common ambient space. All four share the four-condition structure of a δ-interleaving, the filtrations differ, and none of the four implies another: a code can be perfectly self-dual and encode its data badly, or track its data well and correspond to a second code hardly at all. This paper is about (iv).

Matching features across dictionaries. Sparse crosscoders (Lindsey et al., 2024; Jiralerspong & Bricken, 2026; Chaudhari et al., 2026) train a joint dictionary across layers or models and read the correspondence off the shared latents; they answer a different question — learn a common basis — and require retraining for every pair. Decoder-cosine matching assumes each feature is a single direction in a comparable ambient basis, an assumption the not-all-features-are-linear results of Engels et al. (2024) put under strain. The closest work to ours is the semantic-optimal-transport line (Cao et al., 2026), which also abandons single directions: it represents each feature by an activation-weighted distribution, projects those into a shared reference space, matches with a Wasserstein distance (Peyré & Cuturi, 2019; Bonneel et al., 2015; Cuturi, 2013), and proves invariance to activation rescaling, stability under perturbation and recovery under a finite-sample margin condition.

Our construction operates directly on co-activation and therefore requires no projection into a shared reference space; dictionaries may differ in width and activation basis as long as they are evaluated on shared samples. Theorem 2 gives exact Dowker duality for every pair of codes, without a margin or separation condition. The matching itself is then audited in each dictionary's intrinsic metric by

δ_{mut}, separating the unconditional topological relation from the empirical quality and coverage of the committed correspondences.

Representational similarity. CKA (Kornblith et al., 2019), canonical-correlation methods (Raghu et al., 2017; Morcos et al., 2018), shape metrics (Williams et al., 2021) and representational similarity analysis (Kriegeskorte et al., 2008) compare *sample* geometry between two representations and return a scalar; surveys are given by Klabunde et al. (2023). They are not directly comparable to what we do — we return a partial *feature-to-feature* correspondence with a certificate — but they motivate the question, and the platonic-representation line (Huh et al., 2024) is the strongest version of the claim that such correspondences should exist at all. Gromov–Wasserstein matching (Mémoli, 2011) is the metric-geometry ancestor of the distortion criterion we report.

Dictionaries and their evaluation. The dictionaries we match are sparse autoencoders in the line of Bricken et al. (2023); Cunningham et al. (2024); Templeton et al. (2024), descended from sparse coding (Olshausen & Field, 1996) and motivated by superposition (Elhage et al., 2022); recent architectural work (Gao et al., 2024; Rajamanoharan et al., 2024a;b) and open releases (Lieberum et al., 2024a) are what make a layerwise audit possible. Feature splitting and absorption (Chanin et al., 2024) are the transformations our benchmark is built to separate, and benchmark suites (Karvonen et al., 2025; Makelov et al., 2024) evaluate along axes orthogonal to correspondence.

Dowker duality. Due to Dowker (1952), with proofs collected by Björner (1995) and the functorial filtered form we use established by Chowdhury & Mémoli (2018); Robinson (2022) identifies filtrations for which it fails. Interleaving and stability are Chazal et al. (2009); Cohen-Steiner et al. (2007); Chazal et al. (2016), and persistence is computed as in Bauer (2021).

AI USE STATEMENT

We used GPT solely to polish the writing.

REPRODUCIBILITY STATEMENT

The equality of the two feature-side Dowker diagrams and the zero round-trip property of the mutual restriction are proved where they are stated. The synthetic sweep enumerates its twelve width–sparsity configurations explicitly, and every real-dictionary measurement reports the token budget, since the max–min score saturates with sample count and coverage cannot be extrapolated across budgets. Baseline matchers are the correlation and decoder-cosine matchers computed on the same token stream and the same feature subset. All results come from public artifacts and deterministic procedures. The GPT-2 dictionaries are the released `GPT2-Small-SAEs-Reformatted` suite at `blocks.{L}.hook_resid_pre`. The Gemma-2-2B dictionaries are Gemma Scope residual-stream SAEs at width 2^{14}; two conventions there change the numbers by more than an order of magnitude and are worth stating explicitly. The suite's `layer_L` corresponds to the residual stream *after* block L, and the initial token must be excluded: its residual norm is $12.2\times$ the mean of the remaining tokens in our measurements, and retaining it inflates the fraction of unexplained variance from 0.17 to 6.9. Where a single sparsity level is used we take the released level whose realised mean support is closest to 32 and restrict to the 1024 most active alive features, matching the GPT-2 protocol.

REFERENCES

Ulrich Bauer. Ripser: Efficient computation of vietoris–rips persistence barcodes. *Journal of Applied and Computational Topology*, 5(3):391–423, 2021.

Anders Björner. Topological methods. In *Handbook of Combinatorics*, pp. 1819–1872. Elsevier, 1995.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

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.

Tue M. Cao, Nguyen Do, and My T. Thai. Semantic optimal transport for sparse autoencoder feature matching and circuit compression. *arXiv preprint arXiv:2605.28567*, 2026.

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

Marmik Chaudhari, Nishkal Hundia, and Idhant Gulati. Sparse crosscoders for diffing moes and dense models. *arXiv preprint arXiv:2603.05805*, 2026.

Frédéric Chazal, David Cohen-Steiner, Marc Glisse, Leonidas J. Guibas, and Steve Y. Oudot. Proximity of persistence modules and their diagrams. In *Symposium on Computational Geometry*, pp. 237–246, 2009.

Frédéric Chazal, Vin de Silva, Marc Glisse, and Steve Oudot. *The Structure and Stability of Persistence Modules*. Springer, 2016.

Samir Chowdhury and Facundo Mémoli. A functorial dowker theorem and persistent homology of asymmetric networks. *Journal of Applied and Computational Topology*, 2(1):115–175, 2018.

David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete & Computational Geometry*, 37(1):103–120, 2007.

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

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

C. H. Dowker. Homology groups of relations. *Annals of Mathematics*, 56(1):84–95, 1952.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *Transformer Circuits Thread*, 2022.

Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. *arXiv preprint arXiv:2405.14860*, 2024.

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. *arXiv preprint arXiv:2406.04093*, 2024.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. In *International Conference on Machine Learning*, 2024.

Thomas Jiralerspong and Trenton Bricken. Cross-architecture model diffing with crosscoders: Unsupervised discovery of differences between llms. *arXiv preprint arXiv:2602.11729*, 2026.

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

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *arXiv preprint*, 2023.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pp. 3519–3529, 2019.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis — connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. *arXiv preprint*, 2024a.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. *arXiv preprint arXiv:2408.05147*, 2024b.

Jack Lindsey, Adly Templeton, Jonathan Marcus, Tom Conerly, Joshua Batson, and Christopher Olah. Sparse crosscoders for cross-layer features and model diffing. *Transformer Circuits Thread*, 2024.

Aleksandar Makelov, Georg Lange, and Neel Nanda. Towards principled evaluations of sparse autoencoders for interpretability and control. In *International Conference on Learning Representations*, 2024.

Facundo Mémoli. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, 11(4):417–487, 2011.

Ari Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems*, 2018.

Bruno A. Olshausen and David J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

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

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

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders. *arXiv preprint*, 2024b.

Michael Robinson. Cosheaf representations of relations and dowker complexes. *Journal of Applied and Computational Topology*, 6(1):27–63, 2022.

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.

Alex H. Williams, Erin Kunz, Simon Kornblith, and Scott Linderman. Generalized shape metrics on neural representations. In *Advances in Neural Information Processing Systems*, 2021.