Evaluating Sparse Autoencoders in a Structured Continuous Model: Concept Recovery in Implied-Volatility Forecasting

Anonymous Authors[1]

Abstract

Standard SAE concept-recovery protocols rely on correlation alignment, top-K stability, and patching point estimates without explicit nulls. We propose matched-random baselines (leave-one-out random latents for single-latent patching, and matched-random latent pairs for interaction patching, both with bootstrap confidence intervals) and apply them in a structured continuous testbed where ground-truth descriptors are available: a compact MLP trained to forecast SPY implied volatility from tick-level NBBO data over 45 trading sessions. The standard pipeline reports five cross-seed-stable concept families (moneyness geometry, lagged skew, spot level, rates, lagged curvature), and ridge probes from the SAE codes recover four non-geometric descriptors (lagged skew, lagged curvature, spot, lagged ATM level) at positive held-out R^2 (0.27–0.49). Under matched-random patching, however, only one family survives: moneyness geometry produces single-latent effects at 4.49× its leave-one-out baseline ($p < 0.001$), and the moneyness × lagged skew interaction clears the matched-random null ($p = 0.030$, $N = 300$ random pairs). Every other non-geometric family's mean $|\Delta|$ sits at or below its baseline. The probe-patching gap is the paper's central methodological observation: concepts that the SAE *encodes* (probes) and concepts that the model *uses* (patching) are distinct populations. We argue matched-random baselines should be standard for SAE-interp validation, and we present implied-volatility forecasting as a controlled testbed where this gap is visible because the descriptors are pre-specified.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

Post-hoc attribution methods such as SHAP (Lundberg & Lee, 2017) and LIME (Ribeiro et al., 2016) can identify which inputs matter to a prediction, but they do not directly describe how a model organizes its hidden state. Sparse autoencoders (SAEs) have become a prominent tool for decomposing neural representations in language models (Bricken et al., 2023; Cunningham et al., 2024; Gao et al., 2024; Templeton et al., 2024) and related domain models such as chemistry language models (Cohen et al., 2025). SAEs have been combined with activation patching to identify causal feature circuits (Marks et al., 2025; Conmy et al., 2023), and a related body of work uses causal mediation and patching directly on hidden states to localize mechanisms (Vig et al., 2020; Geva et al., 2023). Yet SAE behavior in structured continuous models (the regime where many scientific and financial models live) is much less understood. This gap matters because in that regime datasets are modest, candidate concepts are correlated, and explicit coordinate structure can dominate learned representations.

We study SAE-based concept recovery in implied-volatility forecasting, where neural network architectures have been used to price and predict options since at least Hutchinson et al. (1994), motivated by the Black–Scholes framework (Black & Scholes, 1973) and its stochastic-volatility extensions (Heston, 1993; Gatheral & Jacquier, 2014). This domain is a useful interpretability testbed because it combines strong geometric coordinates with economically meaningful descriptors such as skew, curvature, spot, rates, and liquidity. It is also a difficult setting. Geometry is explicit in the inputs, market descriptors are correlated, and the predictive model is small enough that strong mechanistic claims would be easy to overstate.

Our primary experiment uses a compact two-hidden-layer MLP that predicts individual cells of a fixed SPY volatility-surface grid. We then train TopK SAEs on hidden activations and evaluate the resulting dictionaries along four axes: predictive substrate quality, reconstruction fidelity, concept alignment and cross-seed stability (using linear CKA (Kornblith et al., 2019) as one stability statistic), and intervention sensitivity under activation patching. **We propose matched-random baselines as a standard for SAE-**

interp validation: a leave-one-out random null for single-latent patching (the family's mean absolute effect compared to the mean across all live latents *not* in that family) and a matched-random-pair null for interaction patching (real concept-family pairs compared to $N = 300$ pairs sampled from live latents outside both families), both reported with bootstrap confidence intervals and one-sided p-values. We argue this comparison should accompany correlation-based concept labels and top-K stability claims, because in our experiments those standard-pipeline claims and the matched-random patching results disagree systematically.

This paper makes three contributions. First, it proposes matched-random baselines for SAE patching validation and demonstrates their value in a structured continuous testbed. Second, it documents a *probe-patching gap*: concepts the SAE encodes well according to held-out ridge probes (lagged skew, lagged curvature, lagged ATM level) do not produce single-latent patching effects distinguishable from typical live latents. Third, it presents implied-volatility forecasting as a controlled testbed for SAE concept-recovery methodology, where pre-specified market descriptors give the comparison ground truth that LLM SAE work typically lacks.

The resulting claim is deliberately limited. We do not argue that the underlying model has been mechanistically decomposed. We argue that the standard SAE-interp pipeline (correlation labels, top-K stability, point-estimate patching) can produce a five-concept dictionary that looks coherent on its own terms but where four of the five concepts fail a matched-random patching null, and that this gap is what concept-recovery papers should be measuring.

2. Experimental Setting

2.1. Data and primary prediction task

We use tick-level National Best Bid and Offer (NBBO) SPY options data covering nine preselected weekly windows from July 2024 through April 2025, for a total of 45 trading sessions. The exact windows and their analysis labels are listed in Appendix A. Quotes are aggregated into 5-minute buckets and converted into volatility-surface cross-sections on a 4×3 grid of log-moneyness and tenor, with the tenor binning $[0,7), [7,21), [21,63)$ days and the log-moneyness binning $[-0.20,-0.10), [-0.10,0), [0,0.10), [0.10,0.20)$. Each timestamp can therefore have up to 12 implied-volatility cells.

In practice, observed cells in this dataset concentrate in the shortest tenor bin (< 7 days), reflecting the dominance of short-dated SPY options in the NBBO data over this period. Coverage ratios across buckets are 0.25 or 0.33 (3 or 4 cells out of 12), all in tenor bin 0. Two consequences follow. First, the `tenor_bin_index` input is constant across all surviving examples. Second, the *lagged term slope* statistic, defined as $\mathrm{ATM}_{\mathrm{long}} - \mathrm{ATM}_{\mathrm{short}}$ across observed tenors, is identically zero because there is only one observed tenor; it carries no information in this dataset. We retain both as model inputs to keep the feature list compatible with the joint-surface contrast experiment (Section 3.6), where multi-tenor coverage is intentionally simulated.

For each bucket we compute four surface-summary statistics from observed cells: lagged ATM level, lagged skew, lagged term slope, and lagged curvature. These summaries are shifted by one bucket, and the first bucket of every session is dropped to avoid imputation. The resulting lagged per-cell dataset contains 11,583 examples. We apply a temporal 60/20/20 split by bucket timestamp, producing 6,890 training examples, 2,345 validation examples, and 2,348 test examples. The split is contiguous in bucket-timestamp order: it is not stratified, and there is no embargo gap between adjacent partitions. Because the nine session windows themselves are non-contiguous, the practical effect is a $\approx$ six-month gap between the last training bucket (2024-09-17) and the first test bucket (2025-03-31), placing the test set in a different rate and short-volatility regime. This gap strengthens out-of-sample validity but introduces distribution shift in slowly-varying features (notably the risk-free rate; see Appendix F).

Each example contains 11 inputs: tenor bin index, moneyness bin index, tenor center in years, log-moneyness center, spot price, coverage ratio, risk-free rate, and the four lagged surface summaries. The target is the implied volatility observed at the corresponding grid cell.

2.2. Predictive models

Our primary predictive model is a two-hidden-layer MLP with architecture $11 \to 64 \to 32 \to 1$ and ReLU activations. We train with learning rate 0.01, no dropout, no weight decay, and early stopping with patience 60. As a linear reference, we use ridge regression with regularization parameter $\lambda = 10^{-6}$. We train the MLP across seeds 17, 18, and 19. The ridge model is deterministic under the fixed split, so its reported score is identical across seeds.

This architecture is intentionally modest. The goal is not to maximize forecasting performance across model classes, but to ask whether a compact nonlinear forecaster learns hidden structure that can be decomposed in a useful way.

2.3. Sparse autoencoder and evaluation protocol

For each seed, we train a TopK SAE on the 64-dimensional activations of the first hidden layer (`hidden_0`). Unless otherwise stated, the SAE uses a 32-dimensional dictionary with $k = 5$, batch size 16, 200 epochs, and learning rate

0.001. Separate SAEs are trained for each predictive-model seed.

We evaluate each learned dictionary along four axes.

Predictive substrate quality. We first ask whether the base nonlinear model clearly outperforms a simple linear baseline. This guards against interpreting hidden states from a model that does not earn its complexity.

Reconstruction fidelity. We report reconstruction mean squared error (MSE), mean active features, and live/dead feature rates on held-out data. A latent is called *live* if it is nonzero on at least one test example and *dead* otherwise.

Concept alignment and cross-seed stability. We score each latent by Pearson correlation against model inputs, time of day, and regime indicators derived from the preselected weekly windows. Each latent is assigned to its single best-correlated descriptor's concept family (e.g., moneyness geometry, lagged skew, spot level, rates, lagged curvature, liquidity coverage). A concept family is called *cross-seed stable* if it appears among the top-10 latents (by absolute correlation) in at least two of the three seeds. This rule does not claim one-to-one latent matching, but it distinguishes repeated families from seed-specific noise.

Intervention sensitivity. For activation patching, we compute each live latent's train-set mean μ and standard deviation σ, then patch every test example to $\mu + 1.5\sigma$ and $\mu - 1.5\sigma$. We summarize the resulting symmetric output shift and the fraction of examples for which amplification and suppression move the prediction in opposite directions. We compare each concept family's mean absolute output shift to a leave-one-out random baseline (the mean over all live latents not in that family), and we report a one-sided p-value from a 1,000-sample bootstrap of the mean difference. For pairwise interactions, we compare the mean absolute symmetric finite-difference statistic across real cross-family latent pairs to a null distribution of $N = 100$ matched-random pairs per (seed, family-pair) drawn from live latents outside both families ($N = 300$ pooled across seeds), and we report bootstrap CIs and one-sided p-values.

We also run two secondary analyses: matched SAEs on the second hidden layer (`hidden_1`), and a descriptive dictionary-width sweep on seed 18 over latent widths $\{16, 32, 64, 128, 256\}$. In this sweep, k increases modestly with width (4, 5, 5, 8, 10), so rising dead-feature rates cannot be attributed solely to a fixed sparsity budget.

Table 1. Train and test RMSE in the primary lagged per-cell condition. Ridge is deterministic under the fixed split, so its values do not vary across seeds. The MLP train RMSE is reported alongside the test RMSE so the train–test gap is visible: a ratio of $\approx 1.5\times$ indicates moderate overfitting consistent with the small training set ($n = 6{,}890$) and absence of dropout or weight decay.

Model	Split	Seed 17	Seed 18	Seed 19	Mean
Ridge	Train	1.139	1.139	1.139	1.139
Ridge	Test	1.103	1.103	1.103	1.103
MLP	Train	0.320	0.318	0.329	0.322
MLP	Test	0.557	0.438	0.465	0.487

Table 2. Held-out SAE fidelity by layer in the primary condition. Approximate live features are computed from the mean dead-feature rate for a 32-latent dictionary.

Layer	Recon. MSE	Avg. active	Dead rate	Live / 32
hidden_0	0.0773	4.9999	0.3333	21.3 / 32
hidden_1	0.0068	4.9909	0.6354	11.7 / 32

3. Results

3.1. Predictive performance in the primary condition

The primary lagged per-cell MLP substantially outperforms the ridge baseline on all three seeds, establishing that the hidden activations analyzed by the SAE come from a predictively useful nonlinear model rather than an unstable substrate.

The magnitude of the linear/nonlinear gap matters. If the nonlinear model did not clearly outperform the linear reference, it would be difficult to argue that SAE analysis is decomposing meaningful nonlinear structure. Here, the gap is large enough to justify hidden-state analysis. The MLP shows mild overfitting (test RMSE roughly $1.5\times$ train RMSE), consistent with the small training set and the deliberate decision to avoid regularization. The absolute test RMSE of 0.487 is modest (this is a small forecaster trained on 6,890 examples in a single asset, not a state-of-the-art IV model), so we rely on the linear/MLP gap rather than absolute predictive quality to justify the substrate. We retain the unregularized setup so that the SAE analyzes hidden states from a model that is fit to the training distribution rather than smoothed by an external prior; the train–test gap is small enough that the test-set patching and probing experiments remain interpretable.

3.2. SAE fidelity in the primary condition

Concept recovery is only interpretable when the SAE remains a reasonably faithful summary of the activation space. Table 2 provides that missing layer-level health check.

Two points are important. First, the TopK constraint is

Table 3. Top-10 `hidden_0` concept distribution across seeds in the primary condition.

Concept	S17	S18	S19	Stable?
Moneyness geometry	6	4	4	Yes
Lagged skew	1	2	1	Yes
Spot level	1	2	2	Yes
Rates	1	2	0	Yes
Lagged curvature	1	0	1	Yes
Lagged ATM level	0	0	1	No
Liquidity / coverage	0	0	1	No

active as intended: both layers use about five active features per example on held-out data. Second, the first hidden layer retains a meaningfully live dictionary on test data, with about two thirds of features active at least once. The deeper layer is much sparser in utilization, with only about one third of features live on average.

The raw reconstruction errors are lower in `hidden_1` than in `hidden_0`, but these numbers should be compared cautiously because the layers differ in dimensionality and activation distribution. For the purposes of this paper, the key fact is that the `hidden_0` dictionaries analyzed below are neither collapsed nor trivially dense: they reconstruct the layer with moderate error while preserving a substantial live subdictionary.

3.3. Concept recovery in the first hidden layer

Across the top-10 `hidden_0` latents per seed, moneyness geometry remains the largest concept family, occupying 14 of 30 slots across the three runs. The taxonomy is not purely geometric, however. Lagged skew appears in all three seeds, spot level in all three, rates in two, and lagged curvature in two.

This yields five cross-seed-stable concept families: moneyness geometry, lagged skew, spot level, rates, and lagged curvature. Representative aligned latents reach correlations of approximately $|r| = 0.88$ for moneyness, $|r| = 0.78$ for rates, $|r| = 0.77$ for lagged skew, $|r| = 0.70$ for lagged ATM level, $|r| = 0.66$ for liquidity coverage, and $|r| = 0.65$ for spot. Ridge probes from `hidden_0` SAE codes onto the engineered descriptors confirm that the codes encode several non-geometric concepts at positive held-out R^2 (lagged skew 0.49, lagged curvature 0.37, spot 0.33, lagged ATM level 0.27; full table in Appendix F). The lagged term slope does not appear among the stable families, but this is structurally inevitable: as noted in Section 2.1, the feature is identically zero under single-tenor coverage and so cannot be recovered.

If we stopped here (five cross-seed-stable concept families, six descriptors with $|r| > 0.6$, three non-geometric concepts probe-recoverable at $R^2 > 0.25$), the standard SAE-interp pipeline would conclude that the dictionary recovers a useful mixed taxonomy of coordinate and market-state structure. The next two subsections show that under matched-random patching baselines this conclusion is overstated: only one of the five families produces single-latent patching effects above the random null, and only one cross-family interaction does.

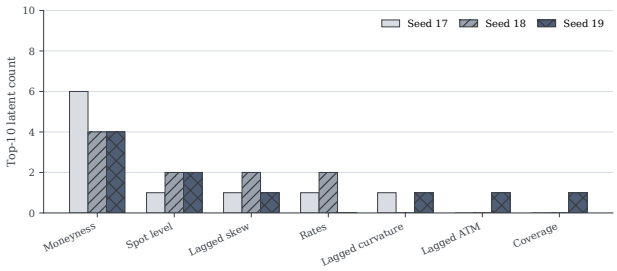

Figure 1. Concept-family counts in the top-10 `hidden_0` latents for each seed. Moneyness remains dominant, but lagged and market-state concepts are consistently present.

Table 4. Patching summary across all three seeds in the primary condition. Mean $|\Delta|$ is the mean absolute symmetric output shift; opp.-dir. is the fraction of examples where amplify and suppress effects have opposite signs. The column n is the number of latents in the family pooled across seeds. "Baseline $|\Delta|$" is the mean $|\Delta|$ across all live latents *not* in that concept family (leave-one-out null), with 95% bootstrap CI in brackets. "Ratio" is the family mean $|\Delta|$ divided by its baseline. The p-value is from a one-sided bootstrap test (1,000 resamples) that the family's mean exceeds the baseline's mean.

Concept	$\|\Delta\|$	Opp.-dir.	n	Baseline $\|\Delta\|$ [95% CI]	Ratio	p
Moneyness geometry	0.395	41.5%	19	0.088 [0.063, 0.117]	**4.49×**	**< 0.001**
Lagged skew	0.136	29.6%	12	0.189 [0.116, 0.273]	0.72×	0.69
Lagged curvature	0.100	19.1%	6	0.187 [0.122, 0.262]	0.53×	0.73
Spot level	0.088	19.9%	8	0.192 [0.124, 0.272]	0.46×	0.84
Lagged ATM level	0.063	16.9%	7	0.193 [0.125, 0.275]	0.33×	0.91
Rates	0.059	7.3%	5	0.189 [0.121, 0.271]	0.31×	0.88
Liquidity / coverage	0.041	10.4%	6	0.194 [0.123, 0.276]	0.21×	0.96

3.4. Intervention sensitivity via activation patching

Activation patching results, evaluated against a leave-one-out random baseline, show that only moneyness geometry produces patching effects clearly above what other live latents produce on average.

Moneyness latents produce mean $|\Delta|$ of 0.395, which is 4.49× the leave-one-out baseline of 0.088 ($p < 0.001$ by one-sided bootstrap with 1,000 resamples). No other concept family reaches the $p < 0.05$ threshold against its leave-one-out baseline. Lagged skew is the strongest non-geometric concept, but its mean $|\Delta|$ of 0.136 sits *below* its baseline of 0.189, with $p = 0.69$. The same pattern holds for lagged curvature, spot, lagged ATM level, rates, and liquidity coverage: each family's mean $|\Delta|$ is below the mean $|\Delta|$ of latents outside that family.

Read alongside Table 3, this is a meaningful disconnect: non-geometric descriptors are recovered as cross-

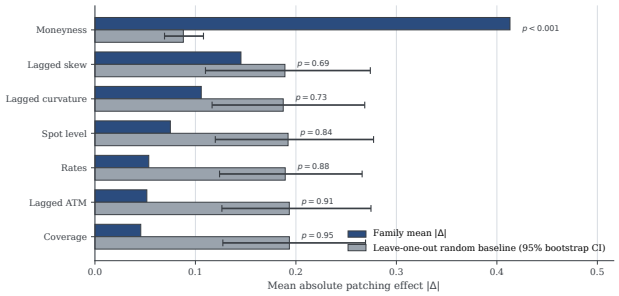

Figure 2. Mean absolute patching effect by concept family (dark) and matched leave-one-out random baseline (light, with 95% bootstrap CI). One-sided p-values are annotated on each row. Only moneyness exceeds its baseline at $p < 0.05$; every non-geometric family's mean $|\Delta|$ sits at or below its leave-one-out baseline. Full numerical breakdown is in Table 4.

seed-stable correlation labels, but the patching evidence does not support the claim that they exert mechanistic influence on predictions *distinguishable from typical live latents*. Direction consistency tells a slightly more favorable story for lagged skew: at 29.6% it is 1.19× its leave-one-out baseline (24.9%), the only non-geometric family whose direction consistency exceeds its baseline. For all other non-geometric families, direction consistency is also below baseline.

We therefore restrict causal language in this paper. The patching evidence supports a single clear claim: moneyness geometry is mechanistically encoded in first-layer SAE latents in this setting. Non-geometric descriptors are recovered as labels that the model could in principle use, but their interventional signature is not separable from the population of live latents.

Probe-patching gap. Combining Section 3.3's probe results with Table 4 highlights the central methodological observation of this paper. Ridge probes from the SAE codes recover lagged skew ($R^2 = 0.49$), lagged curvature ($R^2 = 0.37$), spot ($R^2 = 0.33$), and lagged ATM level ($R^2 = 0.27$) on the held-out test split: by the standard concept-recovery criterion, the SAE *encodes* these descriptors. Yet under matched-random patching, none of these families produces output shifts distinguishable from a typical live latent ($p \geq 0.69$ in each case). The two evaluations are not measuring the same property. Probes ask whether a linear readout from the codes can reconstruct a target; patching asks whether perturbing a code along the direction the SAE has aligned with that target moves the prediction more than perturbing an arbitrary live direction. In our experiments, "concept-encoded by the SAE" and "concept-mechanistically-used by the model" are distinct populations. We argue this gap should be measured, not assumed away, in any concept-recovery claim.

Table 5. Cross-seed interaction patching summary. Real $|I|$ is the mean absolute interaction across cross-family latent pairs in the named concept-family pair, with 95% bootstrap CI in brackets. Null $|I|$ is the mean absolute interaction over matched random pairs drawn from live latents outside both families. The n_{real} count is the number of distinct latent-index pairs (a,b) available across all three seeds where latent a belongs to family A and latent b to family B; the $n_{\text{null}} = 300$ count is 100 random pairs per seed pooled across seeds. "p" is the one-sided fraction of null pairs whose absolute interaction equals or exceeds the family mean.

Family A	Family B	Real $\|I\|$ [95% CI]	Null $\|I\|$ (p95)	n_{real} / n_{null}	p
Moneyness geom.	Lagged skew	0.077 [0.057, 0.098]	0.014 (0.058)	78 / 300	**0.030**
Moneyness geom.	Lagged curvature	0.055 [0.037, 0.077]	0.019 (0.064)	36 / 300	0.060
Moneyness geom.	Rates	0.055 [0.023, 0.099]	0.015 (0.058)	33 / 300	0.060
Spot level	Rates	0.010 [0.004, 0.019]	0.047 (0.206)	15 / 300	0.673

3.5. Pairwise interaction patching

Single-latent patching does not reveal whether two concepts interact in shaping the prediction, beyond what either does in isolation. We probe interactions through a symmetric finite-difference estimator: for each pair of latents (a,b), we patch their codes to all four combinations of $\mu \pm 1.5\sigma$ and form the interaction $f_{++} - f_{+-} - f_{-+} + f_{--}$ on the held-out test set. We aggregate the per-example interaction into mean signed and mean absolute statistics for every cross-family pair of latents available in each seed, with bootstrap confidence intervals on the means and a matched-random null in which both latents are drawn from live latents *outside* the two concept families involved. The null uses $N = 100$ random pairs per (seed, family-pair), and the cross-seed null pools to $N = 300$.

Two patterns emerge. First, interactions involving moneyness geometry produce mean absolute interactions in the 0.05–0.08 range, with confidence intervals that exclude zero. Among them, only moneyness × lagged skew clears the $p < 0.05$ threshold against the matched-random null; moneyness × lagged curvature and moneyness × rates are borderline ($p = 0.06$ each). Second, the spot × rates interaction is small in absolute magnitude (0.010) and sits below its random null (0.047, $p = 0.67$), consistent with these two families having no detectable joint effect beyond what random pairs of live latents produce.

This pattern is consistent with the single-latent patching results: moneyness geometry is the only family that produces a clear effect on predictions both in isolation and in interaction. Lagged skew, while not significant on its own against the leave-one-out baseline, produces a significant interaction with moneyness, suggesting that the model uses lagged skew *conditional on* the moneyness coordinate rather than as a standalone driver. We treat the borderline interactions ($p = 0.06$) as suggestive but not conclusive.

3.6. Contrast experiments across task formulations

The primary lagged per-cell experiment is most informative when viewed alongside two adjacent task formulations. The first omits lagged surface summaries and predicts cells independently; the second predicts the surface jointly. These contrasts reveal that concept richness is not meaningful in isolation. It must be interpreted together with the predictive credibility of the substrate being analyzed.

Three lessons follow. First, the original per-cell task is predictively viable but overwhelmingly geometry-heavy. Second, the joint-surface task yields a richer taxonomy, including liquidity, smile curvature, term structure, and timing concepts, but the predictive model no longer beats the linear baseline and the SAE reconstruction error is much worse. Third, the lagged per-cell condition (the substrate analyzed in Sections 3.3–3.5) improves prediction substantially over both alternatives and surfaces non-geometric descriptors as correlation labels, even though Section 3.4 shows that those non-geometric labels do not survive matched-random patching.

The contrast strengthens the paper's overall framing. The richer concept list in the joint-surface task arises in a setting where the predictive model itself is compromised; without a credible substrate, a richer taxonomy is not stronger evidence. Together with the probe-patching gap in Section 3.4, this argues that concept-recovery claims should be stress-tested along multiple axes (substrate quality, fidelity, probe encoding, and matched-random patching) rather than judged on the concept list alone.

3.7. Depth and dictionary width

The two hidden layers support different concept mixtures. In `hidden_0`, the stable concept families are moneyness geometry, lagged skew, lagged curvature, spot level, and rates. In `hidden_1`, they shift toward moneyness geometry, lagged ATM level, intraday timing, liquidity / coverage, and rates.

This change is suggestive of depth-dependent organization, but it comes with much higher sparsity in the deeper layer. Test dead-feature rates for `hidden_1` range from 0.531 to 0.750, compared with 0.250 to 0.375 for `hidden_0`. Because the deeper layer also differs in dimension and raw activation scale, we treat the layer comparison as descriptive rather than definitive evidence of abstraction.

The width sweep points in the same general direction. Even with larger k at wider dictionaries, dead-feature rates rise sharply while concept diversity does not improve reliably.

The 128-wide dictionary attains the best held-out reconstruction error in this sweep, but it does not expose more distinct top-10 concept families than width 16. The 256-wide dictionary has many live latents but an even higher dead-feature rate and a narrower top-10 taxonomy. In other words, larger dictionaries are not a free interpretability gain in this setting.

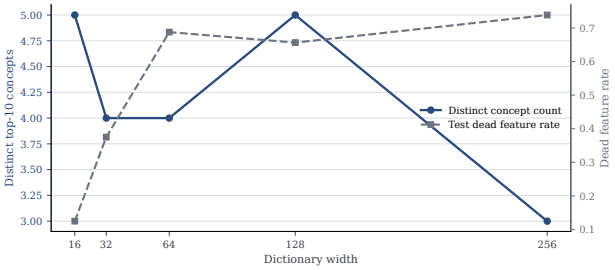

Figure 3. Dictionary-width sweep on seed 18. Concept diversity does not improve reliably with width, while dead-feature rates rise sharply.

4. Discussion

The central finding is the probe-patching gap. The standard SAE-interp pipeline (correlation labels, top-K stability, point-estimate patching) reports five cross-seed-stable concept families in our model, three of which probes recover at positive held-out R^2. Under matched-random patching baselines, four of the five concepts fail. Only moneyness geometry and the moneyness $\times$ lagged skew interaction survive the comparison. The other "recovered" concepts are real as labels but not as mechanism: probing finds them, patching cannot distinguish them from typical live latents.

We take three lessons from this.

First, matched-random baselines change the conclusion. Without them, our setup looks like a successful concept-recovery study with five repeated families and substantial mean $|\Delta|$ values. With them, only one family clears the null. The bar for what counts as a recovered concept moves substantially when the comparison is made explicit. We argue this should be the default comparison in SAE-interp, not an optional supplement.

Second, "encoded" and "used" are different properties. Probes ask whether a linear readout can reconstruct a target from the codes; patching asks whether perturbing the code along the SAE's chosen direction moves predictions more than perturbing an arbitrary live direction. These are not redundant criteria. In structured continuous models with correlated descriptors, an SAE can align a latent with a feature that the downstream model is not actually using, and a probe can recover the alignment without that alignment being mechanistically load-bearing. Treating probe success as evidence of mechanism conflates the two.

Table 6. Three task formulations compared in the study.

Condition	Lin.	MLP	Dead	Recon.	Dominant top-10 concept mix
Original per-cell	1.140	0.854	0.604	0.089	22 moneyness, 3 rates, 2 spot, 2 liquidity, 1 volatility regime
Lagged per-cell	1.103	0.487	0.333	0.077	14 moneyness, 5 spot, 4 lagged skew, 3 rates, 2 lagged curvature, 1 lagged ATM, 1 liquidity
Joint surface	0.563	0.712	0.573	0.218	11 liquidity, 6 rates, 3 spot, 3 smile curvature, 3 term structure, 2 intraday timing, 2 time regime

Table 7. Dictionary-width sweep for seed 18 in the primary condition.

Width	k	Dead	Recon.	Live	Top-10 concepts
16	4	0.125	0.157	14	5
32	5	0.375	0.116	20	4
64	5	0.688	0.144	20	4
128	8	0.656	0.099	44	5
256	10	0.738	0.124	67	3

Third, richer concept lists are not interpretability gains in isolation. Our contrast experiments (Section 3.6) show that the joint-surface task produces a richer concept taxonomy than the per-cell tasks but on a substrate that no longer beats the linear baseline; without a credible predictive model, a richer taxonomy is not stronger evidence of what the model is doing. Combined with the probe-patching gap, this suggests that concept-recovery evaluations should report (i) substrate quality, (ii) reconstruction fidelity, (iii) probe encoding R^2, and (iv) matched-random patching effects together, not the concept list alone.

The IV-forecasting substrate makes the gap visible because the descriptors are pre-specified rather than discovered post hoc. A natural next step is to apply matched-random patching to published SAE-interp results in domains where the same test can be run (small models with public SAE checkpoints) to see whether the gap is general or domain-specific. Our prior is that it is general: most current SAE-interp claims rely on correlation alignment and point-estimate patching without explicit nulls.

5. Limitations and Conclusion

We proposed matched-random baselines for SAE concept-recovery validation (leave-one-out random latents for single-latent patching, matched-random pairs for interaction patching, both with bootstrap CIs and one-sided p-values) and applied them in a structured continuous testbed where ground-truth descriptors exist. The standard SAE-interp pipeline reports five cross-seed-stable concept families and ridge probes recover four non-geometric descriptors (lagged skew, lagged curvature, spot, lagged ATM level) at positive held-out R^2. Under matched-random patching, only one family (moneyness geometry, 4.49× baseline, $p < 0.001$) and one interaction (moneyness × lagged skew, $p = 0.030$) survive. We argue that the gap between probe-encodable and patching-distinguishable concepts is what concept-recovery evaluations should be measuring, and that matched-random nulls should be the default comparison rather than an optional supplement.

Limitations. The study is intentionally controlled rather than comprehensive. It focuses on one asset, one primary nonlinear architecture (a compact two-layer MLP), and one predictive-baseline family (ridge). The dataset has effectively single-tenor coverage in this period, which simplifies the volatility surface and makes `tenor_bin_index` and `lagged_term_slope` constant inputs (Section 2.1); applying the same methodology in a multi-tenor regime is a natural follow-up. The cross-seed stability criterion is threshold-based on top-10 appearance rather than explicit latent alignment via, e.g., the Hungarian assignment over correlation. The deeper-layer (`hidden_1`) and width-sweep experiments are reported descriptively and were not run through matched-random patching; we expect they would also fail the null pattern but have not verified this. Finally, the regime indicators are analysis labels derived from preselected windows, not learned signals.

Future work. The methodological contribution is portable. The most informative next experiment is to apply matched-random patching to publicly released SAE checkpoints in adjacent settings (for example, the small-model SAEs released alongside sparse-feature-circuits work (Marks et al., 2025) or the public Pythia SAEs from EleutherAI) to test whether the probe-patching gap is general. A finding of similar magnitudes in those settings would convert this paper's testbed-specific claim into a field-wide methodological observation. Additional directions include broader IV experiments with multi-tenor coverage, supervised-probe and PCA/ICA decomposition baselines for the underlying activations, and donor-based or counterfactual-localization patching to replace the current population-statistic clamps.

Code and data availability. An anonymized code repository containing the patching, probing, and stability pipelines, together with model checkpoints, SAE checkpoints, per-bucket SAE codes, and the raw CSVs that back Tables 4 and 5, is provided to reviewers at `https://anonymous.4open.science/r/`

iv-sae-mechinterp and will be moved to a public, non-anonymous repository upon acceptance. Reproducing the results from the released artifacts requires only the SAE codes and metadata; no further access to the raw NBBO quote data (which is licensed and cannot be redistributed) is necessary.

A. Weekly windows and analysis labels

Table 8. Preselected weekly windows used in the study. These labels are analysis metadata, not learned regimes.

Label	Start	End
low_vol	2024-07-08	2024-07-12
low_vol	2024-07-15	2024-07-19
low_vol	2024-08-19	2024-08-23
event	2024-07-29	2024-08-02
event	2024-09-16	2024-09-20
high_vol	2024-08-05	2024-08-09
high_vol	2025-03-10	2025-03-14
high_vol	2025-03-31	2025-04-04
high_vol	2025-04-07	2025-04-11

B. Per-seed SAE fidelity by layer in the primary condition

C. Raw top-10 `hidden_0` latent alignments in the primary condition

D. Stable concept families by layer

E. Width-sweep detail for seed 18

F. Ridge probe validation of SAE codes

We validate concept recovery by training ridge regression probes from the SAE codes onto each engineered descriptor. Probes are tuned by 5-fold blocked CV over contiguous timestamp blocks within the training split. Test R^2 is reported on the held-out test split.

For `hidden_0`, lagged skew, lagged curvature, and lagged ATM level achieve substantial CV R^2 (≥ 0.80) and positive test R^2, validating that the SAE codes encode these descriptors and the encoding generalizes. Spot, coverage ratio, and time of day show high CV R^2 but degrade on test, consistent with mild distribution shift. The probe trends weaken at `hidden_1`, where most test R^2 values turn negative; combined with the higher dead-feature rate of `hidden_1` (Table 9), this is consistent with the picture that the second hidden layer is less amenable to a 32-dimensional sparse decomposition in this setting.

G. SHAP versus SAE in seed 18

References

Black, F. and Scholes, M. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3): 637–654, 1973.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Cohen, J., Swanepoel, S., Cass, M. E., Hein, K., and Wigh, D. S. Unveiling latent knowledge in chemistry language models through sparse autoencoders. In *AI4Mat Workshop, NeurIPS*, 2025.

Conmy, A., Mavor-Parker, A. N., Lynch, A., Heimersheim, S., and Garriga-Alonso, A. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, 2023.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, 2024.

Gao, L., la Tour, T. D., Tillman, H., Goh, G., Tow, R., Radford, A., Sutskever, I., Leike, J., and Wu, J. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

Gatheral, J. and Jacquier, A. Arbitrage-free svi volatility surfaces. *Quantitative Finance*, 14(1):59–71, 2014.

Geva, M., Bastings, J., Filippova, K., and Globerson, A. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Heston, S. L. A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The Review of Financial Studies*, 6(2): 327–343, 1993.

Hutchinson, J. M., Lo, A. W., and Poggio, T. A nonparametric approach to pricing and hedging derivative securities via learning networks. *The Journal of Finance*, 49 (3):851–889, 1994.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pp. 3519–3529, 2019.

Table 9. Per-seed SAE fidelity metrics for `hidden_0` and `hidden_1` in the primary lagged per-cell condition.

Seed	Layer	Tr. recon	Te. recon	Tr. active	Te. active	Tr. dead	Te. dead
17	hidden_0	0.002384	0.054039	4.9997	5.0000	0.3125	0.3750
18	hidden_0	0.005035	0.115876	4.9991	5.0000	0.2813	0.3750
19	hidden_0	0.003477	0.062094	4.9975	4.9996	0.1563	0.2500
17	hidden_1	0.000252	0.004913	4.9984	5.0000	0.6875	0.7500
18	hidden_1	0.001037	0.009999	4.9971	5.0000	0.5313	0.6250
19	hidden_1	0.001062	0.005492	4.9906	4.9727	0.5313	0.5313

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Marks, S., Rager, C., Michaud, E. J., Belinkov, Y., Bau, D., and Mueller, A. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *International Conference on Learning Representations*, 2025.

Ribeiro, M. T., Singh, S., and Guestrin, C. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jermyn, A., et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024.

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., and Shieber, S. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, 2020.

Table 10. Raw top-10 `hidden_0` latent alignments in the primary lagged per-cell condition.

Seed	Latent index	Top feature	Concept family	Top correlation	Live example count
17	1	moneyness_bin_index	moneyness_geometry	0.881417	3322
17	22	moneyness_bin_index	moneyness_geometry	-0.850284	3312
17	14	risk_free_rate	rates	-0.783815	2155
17	10	moneyness_bin_index	moneyness_geometry	0.755398	2161
17	31	lagged_skew	lagged_skew	-0.661820	1095
17	11	spot	spot_level	-0.646470	1981
17	24	moneyness_bin_index	moneyness_geometry	-0.579353	1760
17	17	moneyness_bin_index	moneyness_geometry	-0.568497	2820
17	25	moneyness_bin_index	moneyness_geometry	-0.555056	1663
17	5	lagged_curvature	lagged_curvature	-0.543596	3485
18	31	lagged_skew	lagged_skew	0.770889	1502
18	15	moneyness_bin_index	moneyness_geometry	-0.763855	2170
18	28	moneyness_bin_index	moneyness_geometry	0.752065	1846
18	11	moneyness_bin_index	moneyness_geometry	0.722576	2613
18	21	lagged_skew	lagged_skew	-0.676135	1083
18	29	risk_free_rate	rates	-0.647269	668
18	6	risk_free_rate	rates	-0.644416	3550
18	9	spot	spot_level	-0.622835	1327
18	5	spot	spot_level	-0.593488	1044
18	3	moneyness_bin_index	moneyness_geometry	-0.579426	2007
19	29	moneyness_bin_index	moneyness_geometry	-0.762801	2474
19	12	moneyness_bin_index	moneyness_geometry	0.746196	1776
19	20	lagged_atm_level	lagged_atm_level	0.704546	1676
19	21	moneyness_bin_index	moneyness_geometry	-0.657506	1483
19	9	coverage_ratio	liquidity_coverage	-0.655553	1150
19	31	lagged_curvature	lagged_curvature	-0.635973	3623
19	19	lagged_skew	lagged_skew	0.618648	1290
19	17	spot	spot_level	-0.613895	2036
19	26	moneyness_bin_index	moneyness_geometry	0.592979	2334
19	13	spot	spot_level	-0.589269	1600

Table 11. Cross-seed-stable concept families by layer.

hidden_0	hidden_1
Moneyness geometry	Moneyness geometry
Lagged skew	Lagged ATM level
Lagged curvature	Intraday timing
Spot level	Liquidity / coverage
Rates	Rates

Table 12. Full width-sweep detail for seed 18 in the primary condition.

Width	k	Train dead	Test dead	Train recon	Test recon	Live test	Distinct top-10	Mean abs. top-5 corr.	Top-10 concept counts
16	4	0.1250	0.1250	0.009162	0.157291	14	5	0.7671	lagged ATM = 2; lagged skew = 2; moneyness = 3; rates = 1; spot = 2
32	5	0.2813	0.3750	0.005035	0.115876	20	4	0.7371	lagged skew = 2; moneyness = 4; rates = 2; spot = 2
64	5	0.3750	0.6875	0.003486	0.143807	20	4	0.7390	lagged skew = 2; moneyness = 5; rates = 2; spot = 1
128	8	0.3906	0.6563	0.001177	0.099212	44	5	0.7198	lagged curvature = 1; lagged skew = 1; moneyness = 6; rates = 1; spot = 1
256	10	0.5820	0.7383	0.000622	0.123617	67	3	0.7063	moneyness = 7; rates = 1; spot = 2

Table 13. Ridge probe test R^2 from SAE codes to engineered descriptors, averaged across the three seeds. CV R^2 is the cross-validated R^2 on the training split (used for α selection); test R^2 is on the held-out test split. Two rows are degenerate by construction. *Lagged term slope* is identically zero in this dataset (single-tenor coverage; see Section 2.1), so the trivial $R^2 = 1.000$ carries no information. *Risk-free rate* test R^2 is dominated by an out-of-distribution shift: train rates span 4.75%–5.23% (July–September 2024) while test rates span 4.17%–4.24% (March–April 2025), with no overlap. The catastrophic test R^2 values reflect ridge extrapolation, not the SAE codes' inability to encode the rate. We report the rate row for transparency but exclude it from concept-recovery interpretation.

Descriptor	`hidden_0` CV	`hidden_0` test	`hidden_1` CV	`hidden_1` test
Lagged skew	0.94	0.49	0.39	-3.29
Lagged curvature	0.89	0.37	0.45	-1.39
Lagged ATM level	0.80	0.27	0.35	-0.15
Spot	-1.67	0.33	-10.38	-0.27
Coverage ratio	0.84	-0.87	0.34	-1.67
Time of day	0.76	-1.17	0.52	-0.19
Lagged term slope†	1.00	1.00	1.00	1.00
Risk-free rate‡	-8.17	-228.46	-29.02	-747.44

† Constant in this dataset (single-tenor coverage); $R^2 = 1$ is trivial.
‡ Out-of-distribution test split; ridge extrapolates catastrophically. Excluded from interpretation.

Table 14. Seed-18 SHAP versus SAE comparison in the primary condition.

Feature	Mean |SHAP|	SAE freq.
Lagged ATM level	0.5157	0
Moneyness bin	0.4947	4
Log moneyness	0.4840	0
Lagged curvature	0.1638	0
Risk-free rate	0.1500	2
Spot	0.1420	2
Lagged skew	0.0897	2
Coverage ratio	0.0199	0
Tenor bin	0.0000	0
Tenor years	0.0000	0
Lagged term slope	0.0000	0