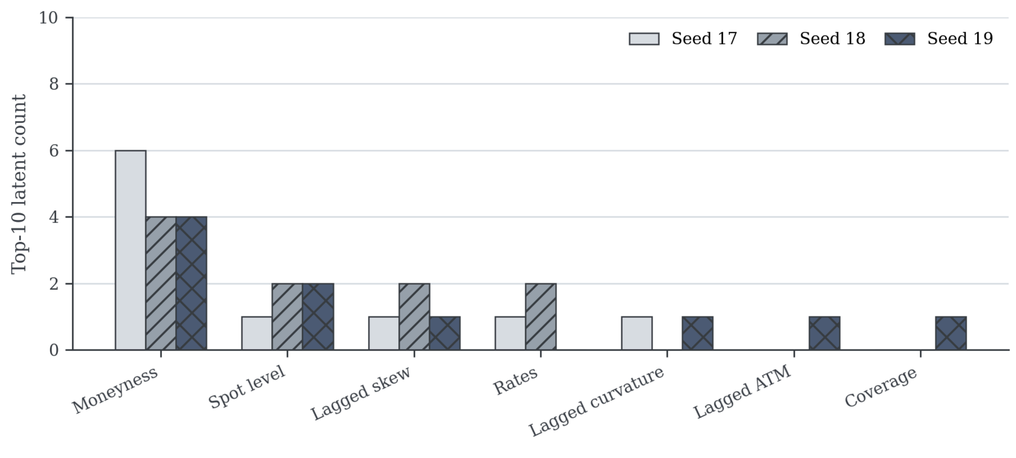
SAE features can look interpretable and still be causally inert.

Sparse Autoencoders for Concept Recovery in a Neural Implied-Volatility Model

Mert Guney

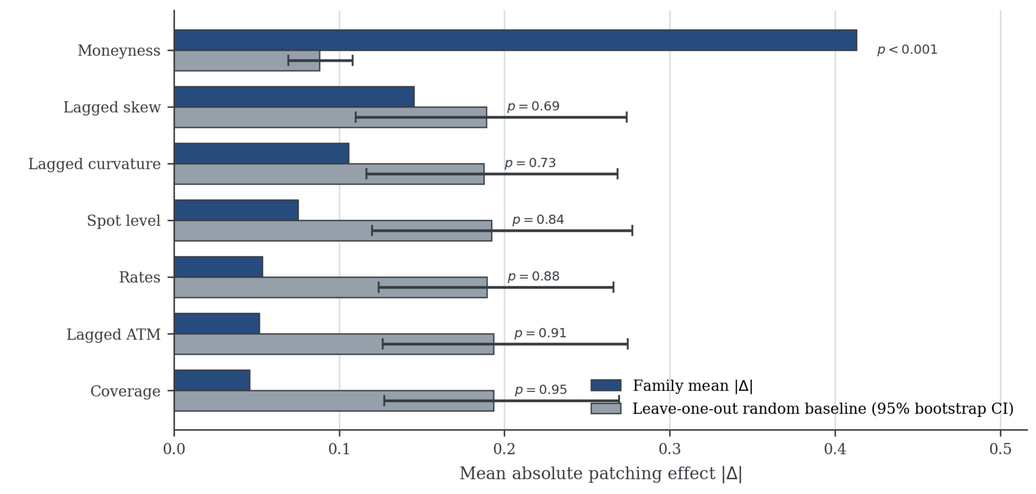
Background. A small neural net (an MLP) predicts the SPY volatility surface: a grid of option-implied volatilities across strike and expiry, built from tick-level options quotes. Of the market concepts a sparse autoencoder (SAE) appears to find inside the net, which does the model actually *use*?

Finding 1. The SAE looks like a clear success: 5 stable, interpretable concepts.



Across 3 seeds, the top-10 SAE latents track five market signals: moneyness (where the strike sits vs the price), lagged skew (vol tilt across strikes), spot (the underlying price), rates, and lagged curvature (smile bend). Ridge probes confirm several are linearly encoded (R^2 up to 0.49).

Finding 2. Causal patching shows it leans mainly on one: moneyness, essentially a raw input coordinate.



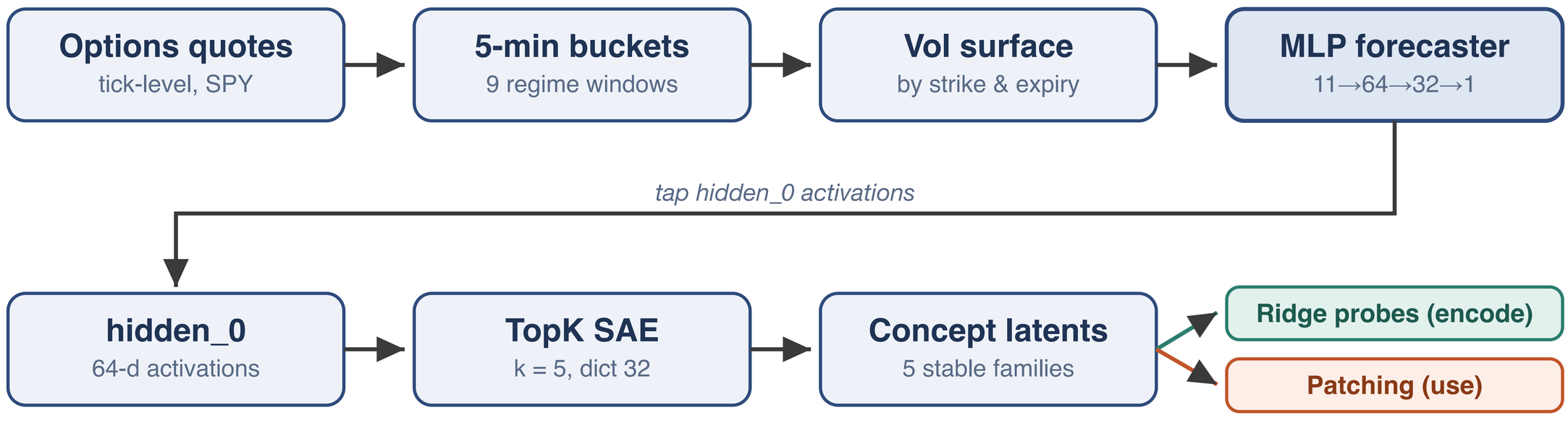
Patch each concept against a matched-random baseline (the model's other live latents). On its own, only moneyness beats the null: |Δ| = **0.395**, **4.49×** baseline ($p < 0.001$). The other four fail as standalone drivers; lagged skew (|Δ| = 0.136 vs 0.189, $p = 0.69$) returns only as a **moneyness-conditional interaction** ($p = 0.030$).

The probe-patching gap

What an SAE **encodes** (correlation, probes) is a different population from what the model **uses** (causal patching). The fix: make **matched-random baselines** a standard validation step for SAE interpretability.

Methods

Tick-level SPY options · TopK SAE (k=5, dict 32) · Activation patching · Matched-random baseline



9 regime-diverse weeks (**45** sessions), Jul 2024 to Apr 2025 · **11,583** records (6,890 / 2,345 / 2,348) · MLP **0.487** vs linear **1.103** RMSE (2.26×) · Moneyness fills **14 of 30** top slots · Only **1 of 5** drives predictions on its own · 3 seeds, cross-seed stable

Limitation. A deliberately limited single-asset pilot: small MLP, correlation-based labels. The deeper layer (hidden_1) is weak (53 to 75% dead features), and patching is limited intervention evidence, not a clean causal decomposition. Wider SAE dictionaries mostly add dead features (up to 73.8%). Takeaway: a standard SAE pipeline can yield a coherent 5-concept dictionary where 4 of 5 fail their causal test.

Mert Guney

Mechanistic Interpretability Workshop, NeurIPS 2026