# Faithful Is Not Interpretable: Sparse Features, Circuits, and Robustness in Frozen Audio Encoders

**Arnesh Batra** [1 2] **Aniket Khandelwal** [1 2 *] **Arush Gumber** [1 2 *] **Krish Thukral** [3 2 *]

## Abstract

Audio ML systems increasingly reuse frozen encoders for event classification, music, speech affect, ASR-related transfer, and audio-text retrieval. A linear probe can show that a representation works, but not whether its evidence is robust under acoustic shift, concentrated in interpretable features, faithfully routed across layers, or safe to edit. We audit seven frozen audio encoders with sparse autoencoders, sparse transcoders, controlled corruptions, steering, and feature ablation. Two separations dominate. First, faithful inter-layer prediction is not interpretability: transcoders explain up to $R^2 = .998$ of the next-layer representation while Whisper-S has zero class-monosemantic routing, and the correlation between explained variance and monosemanticity across stored transitions is only .123. Second, global stability is not feature stability: Whisper-S keeps directionally stable embeddings under corruption, yet its classifier-relevant sparse features stop firing, producing the largest score drop. By contrast, CLAP-HTSAT concentrates label evidence into about eight SAE features on average, while speech encoders distribute evidence over broader, more polysemantic sets. These diagnostics help audio practitioners distinguish compact acoustic detectors from robust distributed codes and faithful dense routes before deploying or editing frozen audio models.

## 1. Introduction

Frozen audio encoders are now reused across the kinds of problems represented in modern audio workshops: environmental and urban acoustic-event classification, music understanding, speech affect, transcription-adjacent transfer, and multimodal audio-text retrieval (Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2022b;c; Baevski et al., 2022; Radford et al., 2023; Wu et al., 2023; Chen et al., 2022a). For practitioners, good clean accuracy is only the first question. The same encoder may be used inside an acoustic monitoring pipeline, a speech-affect system, a retrieval model, or an audio editing interface; in each case we need to know whether its evidence is stable under noise, localizable to a few features, and editable without moving retain classes. Linear probes reveal that many tasks are encoded somewhere in a representation hierarchy, but they leave open the mechanistic question: is the evidence carried by stable class-aligned features, or by broad polysemantic routes that are hard to intervene on safely? Clean transfer alone cannot answer this.

We study this distinction on Whisper-S, CLAP-HTSAT, HuBERT-B, WavLM-B, UniSpeech-SAT, wav2vec2-B, and Data2Vec-Audio across ESC-50, UrbanSound8K, GTZAN, and CREMA-D (Piczak, 2015; Salamon et al., 2014; Tzanetakis & Cook, 2002; Cao et al., 2014). This panel deliberately spans environmental sound, urban events, music genre, and acted speech affect, while the encoders span ASR supervision, audio-text alignment, masked prediction, denoising SSL, contrastive SSL, speaker-aware SSL, and self-distillation. For each encoder-task pair we take the cached readout with highest clean validation score as a fixed input; deciding that readout depth is outside this paper. Our focus is what the chosen representation contains. We train Top-$k$ sparse autoencoders (SAEs), sparse transcoders, controlled-corruption probes, latent steering maps, and targeted feature ablations on the same frozen activation records.

The main finding is a pair of decouplings. **Faithful is not interpretable**: sparse transcoders can predict the next layer almost perfectly while their latents remain class-polysemantic. **Stable is not stable enough**: Whisper-S has low embedding cosine drift under corruption but high alive-feature drift, so the sparse features used by the classifier change even when the embedding direction barely moves. These decouplings matter for audio mechanistic interpretability because audio categories often share physical attributes. A siren, alarm, and human shout can share spectral structure; a speech-

[1]Indraprastha Institute of Information Technology Delhi (IIIT-Delhi) [2]StarkVision Research [3]Manipal University Jaipur. Correspondence to: Arnesh Batra <arnesh23129@iiitd.ac.in>.

*Proceedings of the $43^{rd}$ International Conference on Machine Learning*, Seoul, South Korea. PMLR 306, 2026. 

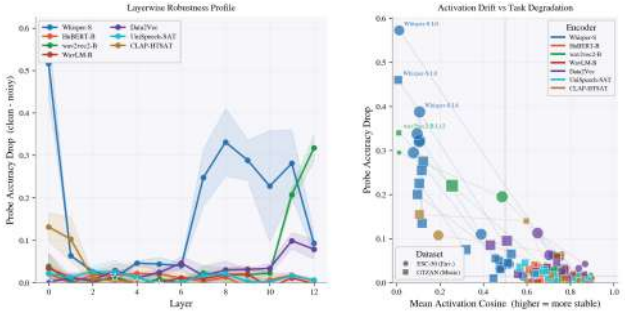


*Figure 1.* Distribution-shift fingerprints under acoustic corruption. The plot jointly summarizes clean-to-corrupted score drop, embedding drift, and alive-feature drift. Whisper-S keeps globally similar embeddings while its classifier-relevant sparse features change, producing the largest performance drop.

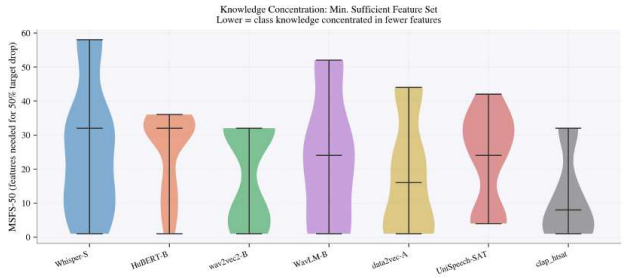


*Figure 2.* MSFS at 50% target drop; smaller values indicate compact evidence.

emotion label can depend on prosody, speaker, and phonetic content. Feature-level audits therefore need robustness and retain-set checks, not only clean probe scores.

This is also an ethical and deployment question, not only a visualization question. Sparse steering or ablation could make monitoring, moderation, or assistive audio systems easier to tune, but the same interventions can suppress or amplify sensitive acoustic attributes. We therefore frame interventions as audits: evidence for where a model stores task information and how much collateral movement a proposed edit would cause, rather than as claims of safe control.

**Contributions.** We (i) give controlled-corruption robustness fingerprints that separate embedding stability from sparse-feature stability, (ii) quantify SAE concentration through minimum sufficient feature sets, (iii) show that transcoder faithfulness and class-monosemanticity are separate axes, and (iv) use steering and feature ablation as diagnostics for when sparse audio features expose actionable handles. We build on sparse interpretability, monosemanticity, superposition, and transcoders (Elhage et al., 2021; 2022; Cunningham et al., 2023; Bricken et al., 2023; Templeton et al., 2024; Dunefsky et al., 2024) and extend audio-native interpretability beyond prior NMF and AudioSAE studies (Parekh et al., 2022; Aparin et al., 2026).

## 2. Methods

**Representations and SAEs.** All encoders are frozen. Speech-style encoders expose layers 0–12; CLAP-HTSAT exposes a stem plus four HTS-AT stages, so its indices are stage-wise. We mean-pool hidden states over time or time-frequency positions before fitting probes and interpretable models. For a selected representation $h^{(\ell)} \in \mathbb{R}^{768}$, the Top-$k$ SAE reconstructs

$$\hat{h} = W_{\text{dec}}\,\text{TopK}(W_{\text{enc}}h + b_{\text{enc}}, k) + b_{\text{dec}}, \qquad (1)$$

with a $6\times$ dictionary (4608 features), $k$ annealed from 130 to 75, and nonnegative Top-$k$ activations. A feature is class-monosemantic when more than 70% of its activation mass falls on one label. We also measure the minimum sufficient feature set (MSFS): the smallest set of class-associated SAE features whose ablation causes a 50% target-class drop.

**Transcoders and faithfulness.** For adjacent readouts, a sparse transcoder predicts the next representation:

$$T_\psi(h^{(\ell)}) = W_{\text{out}}\,\text{TopK}(W_{\text{in}}h^{(\ell)} + b_{\text{in}}, k_t) + b_{\text{out}}, \quad (2)$$

with latent dimension 3840 and $k_t = 16$ for 12-layer encoders. Faithfulness is explained variance,

$$R^2 = 1 - \frac{\sum_i \|h_i^{(\ell+1)} - T_\psi(h_i^{(\ell)})\|_2^2}{\sum_i \|h_i^{(\ell+1)} - \bar{h}^{(\ell+1)}\|_2^2}. \qquad (3)$$

Interpretability is measured separately by the class-monosemantic fraction of active transcoder latents.

**Robustness and interventions.** Controlled perturbations include additive noise, band-pass filtering, and frequency shift. We report clean-to-corrupted score drop, embedding cosine drift, and alive-feature drift. For clean and corrupted Top-$k$ active sets $A_i, A_i'$, alive-feature drift is

$$F = \frac{1}{N}\sum_i \left(1 - \frac{|A_i \cap A_i'|}{|A_i \cup A_i'|}\right). \qquad (4)$$

For steering, we add a scaled empirical class-mean difference vector, $\tilde{z} = z + \alpha v^{(\ell)}$, and measure target gain minus off-target movement. Feature ablation zeros ranked SAE activations before reconstruction and reports target suppression and retain-set movement.

## 3. Results

### 3.1. Global Stability Is Not Feature Stability

Table 1 and Figure 1 show that clean transfer does not predict robustness. WavLM-B and masked SSL encoders are most stable under the corruption suite, consistent with denoising and masked-prediction objectives that train corruption-tolerant acoustic subspaces. Whisper-S is the

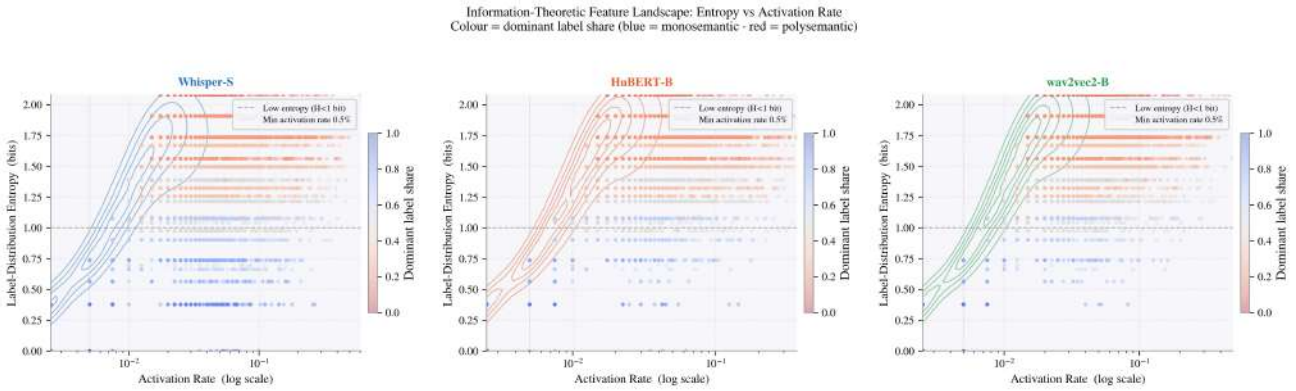


*Figure 3.* Feature entropy versus activation rate on ESC-50, separating rare detectors from high-rate polysemantic features.

*Table 1.* Controlled-perturbation robustness by pretraining paradigm. Lower drops/drifts are better; relative robustness is normalized so the strongest paradigm is 1.00.

| Paradigm | Acc. drop | Act. drift | Feat. drift | Rel. rob. |
|---|---|---|---|---|
| SSL denoising | .044 | .103 | .117 | 1.00 |
| SSL masked pred. | .065 | .195 | .168 | .89 |
| SSL contrastive | .083 | .460 | .138 | .87 |
| Audio-text contr. | .101 | .377 | .207 | .68 |
| Supervised ASR | .280 | .074 | .407 | .00 |

counterexample: its embeddings have the lowest activation drift (.074), yet alive-feature drift is highest (.407) and the score drop is largest (.280). The embedding direction remains similar while the sparse features used by the classifier stop firing. CLAP-HTSAT shows the complementary risk: strong clean transfer can rely on compact category features that shift under acoustic corruption.

### 3.2. Sparse Concentration Predicts Editability

Figure 2 shows that CLAP-HTSAT reaches a 50% target drop after ablating about eight SAE features, while Whisper-S and WavLM-B require broader sets. Figure 3 gives the complementary view: rare, low-entropy features behave like class-selective detectors, while frequent high-entropy features are reused across labels and require retain-set audits. This is why monosemanticity alone is insufficient: a feature must also align with the probe direction and avoid retain classes.

### 3.3. Faithful Routing Is Not Interpretable Routing

Transcoders add a circuit-level view. Across stored transitions, explained variance is high ($R^2 \approx .86$ to .998), but faithfulness and interpretability separate. Whisper-S has class-monosemantic ratio 0 under our criterion even when $R^2$ reaches .998. UrbanSound8K elicits the highest monosemantic transcoder ratios for every encoder, with CLAP-HTSAT reaching .774 at $R^2 = .975$. Across all stored transitions, the correlation between $R^2$ and monosemantic ratio is only .123. High $R^2$ therefore means that the next layer is predictable; it does not mean that the route decomposes into class-readable pieces.

*Table 2.* Sparse-feature and transcoder signatures. MSFS is the number of SAE features whose ablation causes a 50% target-class drop. Larger $R^2$ means faithful next-layer prediction; larger monosemantic ratios mean easier class-level interpretation.

| Family | Sparse features | Transcoder route | Interpretation |
|---|---|---|---|
| CLAP-HTSAT | MSFS $\approx 8$; ESC-50 SAE mono .258 | US8K transcoder mono .774; $R^2 = .975$ | Compact label-aligned event features; perturbation audit needed. |
| Whisper-S | Broad sets; SAE mono $\leq .137$ | Transcoder mono 0 despite $R^2$ up to .998 | Predictable but dense decoder-facing routing. |
| Masked/den. SSL | Moderate selectivity; lowest feature drift .117–.168 | US8K transcoder mono .296–.471 | Partly disentangled routes; stable under corruption. |
| D2V/w2v2 | Evidence often front-loaded in cached sweeps | US8K transcoder mono .190/.152 | Later objective-specialized states can hide useful signal. |

### 3.4. Interventions Are Diagnostics

Steering and ablation ask whether sparse structure is actionable. In Figure 4, class-difference vectors are most selective near the readouts that also support strong probes: Whisper-S steers late, while HuBERT-B and wav2vec2-B steer earlier. Feature ablation gives the sparse counterpart: compact monosemantic feature sets suppress targets with less retain-set movement, whereas distributed encodings require larger ablation sets and incur more collateral movement. We treat both as diagnostic interventions, not as guarantees of safe control or training-data deletion (Ravfogel et al., 2020; Belrose et al., 2023; Cao & Yang, 2015).

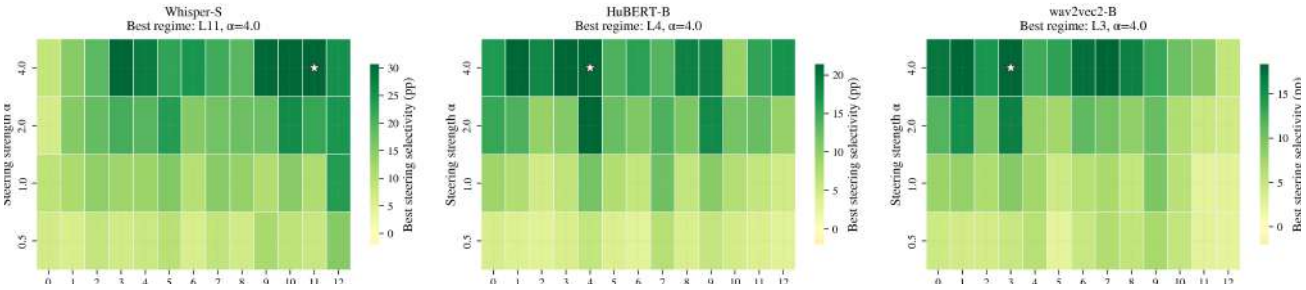


*Figure 4.* Latent steering phase map. Selectivity is target-class gain minus off-target movement as layer and steering strength vary. High-selectivity regions identify readouts with approximately linear intervention handles.

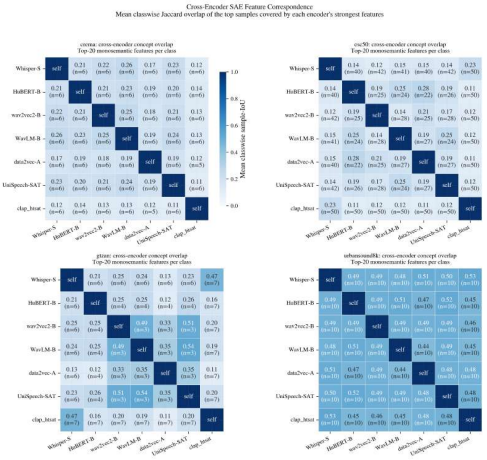


*Figure 5.* Cross-encoder feature overlap measured by activation-profile IoU. High overlap means two encoders activate sparse features on similar examples; low overlap means they represent the same dataset through different internal concepts.

The ablation phase map is the intervention counterpart to MSFS. Low-MSFS encoders are not automatically safe to edit: a small feature set can still sit on a shared acoustic factor and move retain labels. Conversely, distributed encodings can be faithful and robust but require broad feature sets before the target class changes. Reporting target suppression without retain-set movement therefore overstates what sparse audio features can do.

### 3.5. Feature Handles Are Encoder-Specific

If sparse features were merely rediscovering a universal audio vocabulary, then features from different encoders would fire on the same clips. Figure 5 shows a weaker and more useful pattern: feature overlap is structured by encoder family. Related speech SSL encoders share more activation-profile support, while CLAP-HTSAT and Whisper-S carve the same datasets into different sparse concepts. This explains why intervention handles should be audited per encoder even when two probes solve the same task.

This cross-encoder result also qualifies feature-level explanations. A monosemantic feature in one encoder is not automatically the same mechanism as a similarly named feature in another encoder; at most it is a local handle whose activation profile, robustness, and retain-set behavior have been measured. The interpretability unit in this paper is therefore the encoder-specific sparse feature and its route through a faithful transcoder, not a universal acoustic concept dictionary.

## 4. Discussion

Audio interpretability audits should evaluate more than clean score: robustness of the sparse active set, concentration of label evidence, and monosemanticity of faithful routes. This study is observational, since objective, data, architecture, and supervision remain entangled, and SAE claims depend on dictionary size and sparsity. Still, the separations are practically clear: Whisper-S is globally stable but feature-fragile; CLAP-HTSAT is compact yet shift-sensitive; and high-$R^2$ transcoders can remain polysemantic. Faithful audio circuits are therefore not automatically interpretable circuits. The practical implication is to treat sparse audio interpretability as a multi-axis audit. Report clean score, perturbation score drop, alive-feature drift, MSFS, transcoder $R^2$, class-monosemantic ratio, target suppression, and retain-set movement together. Agreement across these axes indicates a compact, stable, and class-readable route; disagreement is itself informative, revealing whether the model uses a distributed code, fragile sparse detector, or faithful but dense polysemantic path.

**Conclusion.** The broader message for machine learning for audio is that faithful reconstruction and clean transfer are only partial tests of interpretability. Sparse features, faithful transfer, and robust intervention behavior must be evaluated jointly. Future work should connect these diagnostics to activation patching through transcoder paths, counterfactual audio edits tied to individual SAE features, human listening studies for semantic validity, and sensitivity analyses over dictionary size, sparsity, and model scale. More broadly, this motivates an audio circuit audit that is faithful, sparse, robust, intervention-aware, and semantically grounded rather than optimized for any single proxy.

## References


Aparin, G., Sadekova, T., Rukhovich, A., Yermekova, A., Kushnareva, L., Popov, V., Kuznetsov, K., and Piontkovskaya, I. AudioSAE: Towards understanding of audio-processing models with sparse AutoEncoders. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3221–3254, Rabat, Morocco, March 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.eacl-long.149. URL https://aclanthology.org/2026.eacl-long.149/. arXiv preprint arXiv:2602.05027.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12449–12460, 2020. URL https://arxiv.org/abs/2006.11477.

Baevski, A., Hsu, W.-N., Xu, Q., Babu, A., Gu, J., and Auli, M. data2vec: A general framework for self-supervised learning in speech, vision and language. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1298–1312. PMLR, 2022. URL https://arxiv.org/abs/2202.03555.

Belrose, N., Schneider-Joseph, D., Ravfogel, S., Cotterell, R., Raff, E., and Biderman, S. LEACE: Perfect linear concept erasure in closed form. *Advances in Neural Information Processing Systems*, 36, 2023. doi: 10.52202/075280-2884.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N. L., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., and Verma, R. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014. doi: 10.1109/TAFFC.2014.2336244.

Cao, Y. and Yang, J. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pp. 463–480. IEEE, 2015. doi: 10.1109/SP.2015.35.

Chen, K., Du, X., Zhu, B., Ma, Z., Berg-Kirkpatrick, T., and Dubnov, S. HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 646–650. IEEE, 2022a. doi: 10.1109/ICASSP43922.2022.9746312. URL https://arxiv.org/abs/2202.00874.

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Qian, Y., Seltzer, M. L., Wang, S., Chen, L., Meng, H., Yu, D., and Wei, F. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022b. doi: 10.1109/JSTSP.2022.3188113. URL https://arxiv.org/abs/2110.13900.

Chen, S., Wu, Y., Wang, C., Chen, Z., Chen, Z., Liu, S., Wu, J., Qian, Y., Wei, F., Li, J., Zeng, X., and Yu, D. UniSpeech-SAT: Universal speech representation learning with speaker aware pre-training. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6152–6156. IEEE, 2022c. doi: 10.1109/ICASSP43922.2022.9747077. URL https://arxiv.org/abs/2110.05752.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models, 2023. URL https://arxiv.org/abs/2309.08600.

Dunefsky, J., Chlenski, P., and Nanda, N. Transcoders find interpretable LLM feature circuits. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0768.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., and Olah, C. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. HuBERT: Self-supervised speech representation learning by masked

prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. doi: 10.1109/TASLP.2021.3122291. URL https://arxiv.org/abs/2106.07447.

Parekh, J., Parekh, S., Mozharovskyi, P., d'Alché Buc, F., and Richard, G. Listen to interpret: Post-hoc interpretability for audio networks with NMF. In *Advances in Neural Information Processing Systems*, volume 35, pp. 35201–35214, 2022. doi: 10.52202/068431-2556.

Piczak, K. J. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pp. 1015–1018. ACM, 2015. doi: 10.1145/2733373.2806390.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. Robust speech recognition via large-scale weak supervision, 2023. URL https://arxiv.org/abs/2212.04356.

Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., and Goldberg, Y. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.647.

Salamon, J., Jacoby, C., and Bello, J. P. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM International Conference on Multimedia*, pp. 1041–1044. ACM, 2014. doi: 10.1145/2647868.2655045.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., and Henighan, T. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Tzanetakis, G. and Cook, P. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002. doi: 10.1109/TSA.2002.800560.

Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., and Dubnov, S. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2023. doi: 10.1109/ICASSP49357.2023.10095969. URL https://arxiv.org/abs/2211.06687.

# A. Datasets, Encoders, and Implementation Details

This appendix contains the experimental details and additional diagnostics for the robustness, sparse-feature, transcoder, steering, ablation, and genealogy claims in the main text.

## A.1. Datasets and Splits

Table 3: Dataset summary used in the frozen-representation experiments.

| Dataset | Task | Clips | Classes | Train | Test/held-out |
|---|---|---|---|---|---|
| ESC-50 | Environmental sound | 2000 | 50 | 1600 | 400 |
| GTZAN | Music genre | 999 | 10 | 799 | 200 |
| CREMA-D | Speech emotion | 7442 | 6 | 6136 | 1306 |
| UrbanSound8K | Urban sound | 8732 | 10 | 7079 | 1653 |

The four datasets stress complementary acoustic abstractions. ESC-50 tests broad environmental semantics; UrbanSound8K tests narrow urban events; GTZAN combines instrumentation, rhythm, texture, and recording style; and CREMA-D isolates acted speech affect with speaker variation.

## A.2. Model Catalog

Table 4: Studied encoders and analyzed readout depths.

| Encoder | Objective family | Parameters | Width | Depths analyzed |
|---|---|---|---|---|
| Whisper-S | Weakly supervised ASR | 244M | 768 | 0–12 |
| CLAP-HTSAT | Audio-text contrastive | 86M | 768 | 0–4 (stem+4 stages) |
| HuBERT-B | Offline-unit masked prediction | 95M | 768 | 0–12 |
| WavLM-B | Masked prediction + denoising | 95M | 768 | 0–12 |
| UniSpeech-SAT | Speaker-aware masked prediction | 94M | 768 | 0–12 |
| wav2vec2-B | Contrastive masked prediction | 95M | 768 | 0–12 |
| Data2Vec-Audio | Self-distillation | 94M | 768 | 0–12 |

## A.3. Implementation Hyperparameters

Table 5: Implementation hyperparameters for Paper B diagnostics.

| Component | Settings |
|---|---|
| Frozen extraction | Audio is loaded as mono, resampled to the checkpoint processor rate, and clipped to 5 s for ESC-50, UrbanSound8K, and CREMA-D, and 30 s for GTZAN. Hidden states are extracted with `output_hidden_states=True` and mean-pooled over time or time-frequency positions. |
| SAE training | Top-$k$ SAE with hidden multiplier 6, dictionary size 4608 for width-768 encoders, $k_0 = 130$ annealed to $k = 75$ over the first 50% of steps, sparsity weight $5 \times 10^{-5}$ with 10% warmup, 600 epochs, batch size 128, Adam learning rate $10^{-4}$, cosine schedule, decoder normalization, dead feature revival window 75, and seed 13. |
| Transcoder training | Sparse transcoder with hidden multiplier 5, Top-$k = 16$, 20 epochs, batch size 512, Adam learning rate $9 \times 10^{-4}$, sparsity weight $3 \times 10^{-4}$, and seed 17. Metrics report reconstruction MSE, explained variance, active fraction, mean latent $\ell_1$, and class-monosemantic ratio. |
| Perturbations and interventions | Robustness sweeps use controlled additive noise, band-pass filtering, and frequency shift. Steering uses empirical class-mean differences with strengths $\{0.5, 1, 2, 4\}$. Feature ablation ranks class-associated SAE features by activation mass and reports target suppression, retain score, and off-target drift. |

# B. SAE, Transcoder, and Intervention Objectives

The Top-$k$ SAE in Equation 1 is trained by reconstruction MSE with an $\ell_1$ sparsity penalty on normalized activations. Its reconstruction coefficient is

$$R^2_{\mathrm{SAE}} = 1 - \frac{\sum_i \|h_i - \hat{h}_i\|_2^2}{\sum_i \|h_i - \bar{h}\|_2^2}. \tag{5}$$

Across stored SAE artifacts, reconstruction MSE has median $1.9 \times 10^{-5}$ and mean $4.7 \times 10^{-5}$; median reconstructed-probe change is $-0.018$ score points and the dead-feature fraction is below 2%.

For each SAE feature we measure activation rate and label entropy,

$$H^{(f)} = -\sum_{c=1}^{C} p(y = c \mid f \text{ active}) \log_2 p(y = c \mid f \text{ active}), \tag{6}$$

which separates rare class-selective detectors from frequent polysemantic features. MSFS at threshold $\tau$ is

$$\text{MSFS-}\tau = \min\{|S| : \text{Acc}_{\text{ablate}(S)} \leq (1 - \tau/100)\,\text{Acc}\}. \tag{7}$$

For a downstream logit at layer $\ell + 1$, $g_c(h) = w_c^\top h + b_c$, transcoder error bounds probe-logit error by Cauchy-Schwarz:

$$\left| g_c(h^{(\ell+1)}) - g_c(T_\psi(h^{(\ell)})) \right| \leq \|w_c\|_2 \, \|h^{(\ell+1)} - T_\psi(h^{(\ell)})\|_2. \tag{8}$$

High explained variance therefore bounds a residual that can perturb linear readouts, but does not certify monosemantic routing.

For feature-level ablation, let $\hat{z} = b + \sum_f a_f d_f$ and let $g_c(z) = w_c^\top z + b_c$. Ablating a set $S$ changes a class margin by

$$\Delta M_{y,c}(S) = -\sum_{f \in S} a_f (w_y - w_c)^\top d_f. \tag{9}$$

If features in $S$ are target-aligned for class $y$, $(w_y - w_c)^\top d_f \geq \beta_f > 0$, and retain-neutral for non-target contrasts, $|(w_r - w_{c'})^\top d_f| \leq \kappa_f$, then target margin reduction is at least

$$\sum_{f \in S} a_f \beta_f, \tag{10}$$

while retain margin changes are at most

$$\sum_{f \in S} a_f \kappa_f. \tag{11}$$

This is why MSFS, monosemanticity, and retain-set movement must be reported together.

## C. Robustness and Steering Details

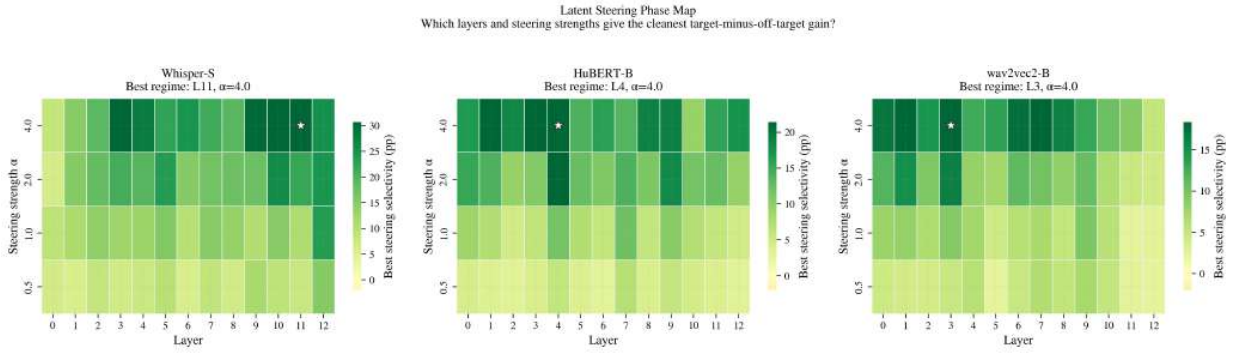


*Figure 6.* Latent steering phase maps. The map sweeps readout depth and steering strength for class-difference vectors, measuring target-class gain after subtracting off-target movement. Selective steering is strongest near readouts that also maximize clean probe utility.

For a linear probe with class weights $w_c$ and biases $b_c$, define the clean margin of example $i$ with true class $y_i$ at layer $\ell$ as

$$\Gamma_i^{(\ell)} = \min_{c \neq y_i} \left[ (w_{y_i} - w_c)^\top z_i^{(\ell)} + b_{y_i} - b_c \right]. \tag{12}$$

If $\Gamma_i^{(\ell)} > 0$ and corruption-induced drift $\tilde{\delta}_i^{(\ell)}$ satisfies

$$\max_{c \neq y_i} \left| (w_{y_i} - w_c)^\top \tilde{\delta}_i^{(\ell)} \right| < \Gamma_i^{(\ell)}, \tag{13}$$

then the linear probe prediction for example $i$ is unchanged. A model can therefore have small global cosine drift while predictions change if drift lands in the probe or sparse-feature subspace.

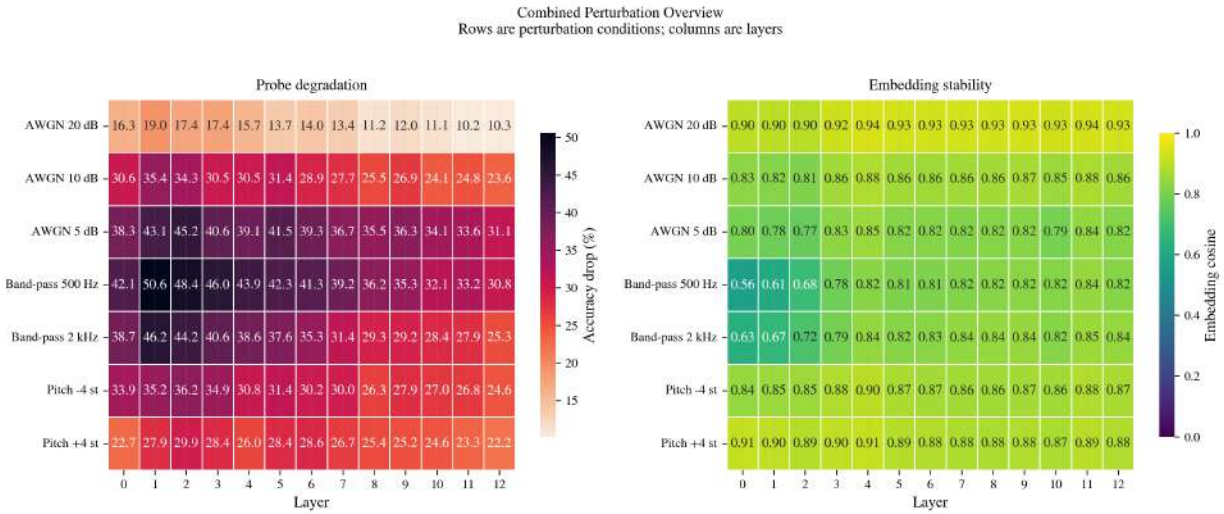


*Figure 7.* Controlled perturbation analysis across additive noise, band-pass filtering, and frequency shift. The panels separate score drop, representation drift, and sparse-feature drift, revealing structure that clean accuracy alone cannot see.

## D. Sparse Feature and Transcoder Tables

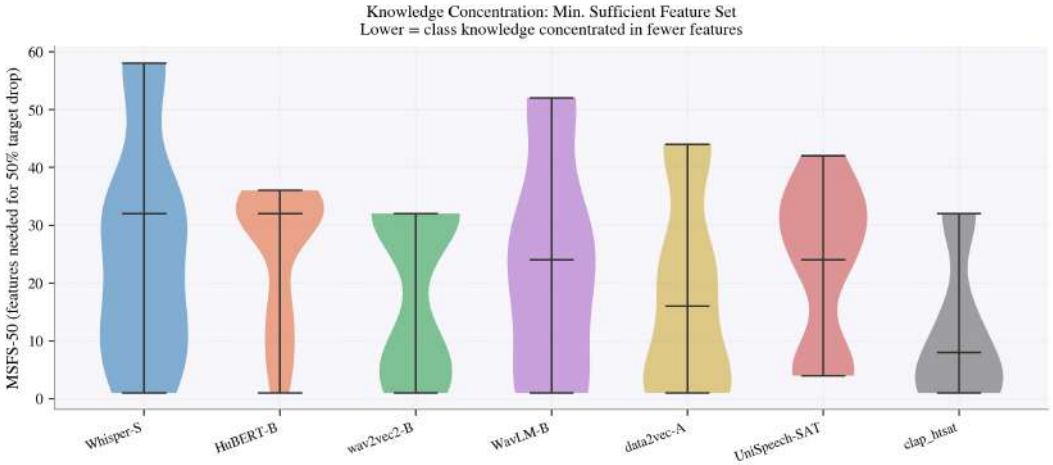


*Figure 8.* Minimum sufficient feature set at 50% target drop. Smaller values mean the classifier depends on a compact set of SAE features.

For a traditional descriptor family $\mathcal{T}$ and SAE feature $f$, the feature-level bridge is

$$B_f(\mathcal{T}) = \max_{g \in \mathcal{T}} |\mathrm{Corr}(a_f(x_i), g(x_i))| \,. \tag{14}$$

$B_f$ asks whether an individual sparse feature has a classical acoustic interpretation.

## E. Editing and Additional Figures

## F. Sparse-Feature Genealogy

Genealogy plots trace SAE features through adjacent depths by activation-profile overlap. They are feature-lineage diagrams rather than causal explanations: branches indicate that later features inherit activation mass from earlier features, while merges indicate that a later feature combines earlier acoustic or semantic factors.

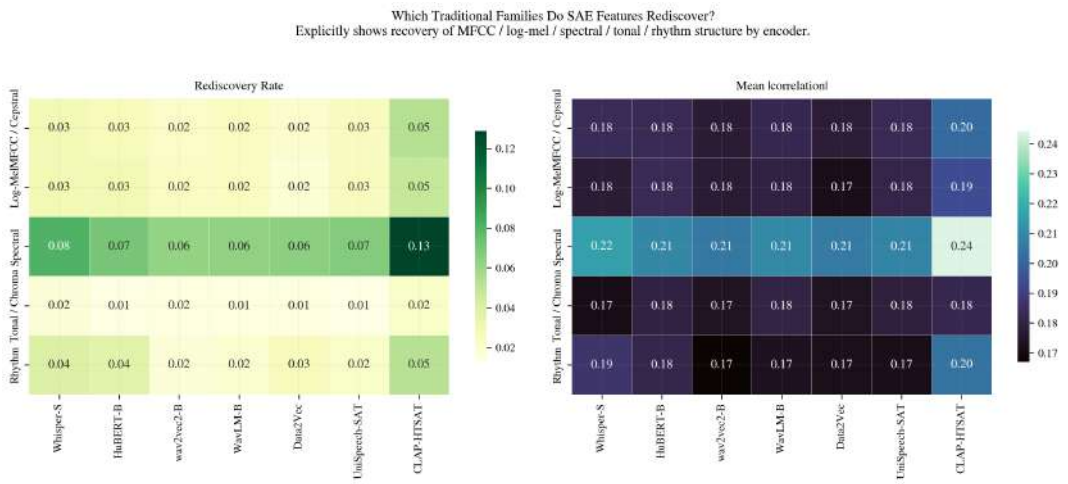


*Figure 9.* Alignment between SAE features and classical audio descriptor families. The rediscovery score compares SAE activations with MFCC/cepstral, mel, spectral, tonal/chroma, rhythm, and combined descriptors.

*Table 6.* Sensitivity of SAE class-monosemantic feature ratios to the dominant-label threshold. Ratios are computed on each dataset's selected SAE readout and then pooled by encoder.

| Encoder | Share $\geq 60\%$ | Share $\geq 70\%$ | Share $\geq 80\%$ | Mean share |
|---|---|---|---|---|
| CLAP-HTSAT | .300 | .188 | .100 | .513 |
| Data2Vec-Audio | .255 | .156 | .075 | .485 |
| WavLM-B | .232 | .143 | .072 | .475 |
| Whisper-S | .227 | .136 | .060 | .474 |
| HuBERT-B | .226 | .135 | .065 | .470 |
| UniSpeech-SAT | .208 | .119 | .055 | .463 |
| wav2vec2-B | .193 | .110 | .050 | .456 |

*Table 7.* Best-transcoder monosemantic ratio by dataset and encoder. UrbanSound8K elicits the highest monosemantic routing for every encoder.

| Encoder | CREMA-D | ESC-50 | GTZAN | US8K |
|---|---|---|---|---|
| CLAP-HTSAT | .000 | .026 | .000 | **.774** |
| UniSpeech-SAT | .115 | .057 | .015 | **.471** |
| HuBERT-B | .190 | .057 | .000 | **.321** |
| WavLM-B | .086 | .089 | .011 | **.296** |
| wav2vec2-B | .050 | .093 | .000 | **.152** |
| Data2Vec | .021 | .000 | .014 | **.190** |
| Whisper-S | .000 | .000 | .000 | .000 |

*Table 8.* Best-transcoder transition by encoder and dataset.

| Encoder | CREMA-D | ESC-50 | GTZAN | US8K |
|---|---|---|---|---|
| CLAP-HTSAT | $3 \to 4$ | $0 \to 1$ | $0 \to 1$ | $3 \to 4$ |
| Whisper-S | $10 \to 11$ | $10 \to 11$ | $2 \to 3$ | $10 \to 11$ |
| HuBERT-B | $11 \to 12$ | $0 \to 1$ | $0 \to 1$ | $0 \to 1$ |
| UniSpeech-SAT | $1 \to 2$ | $0 \to 1$ | $0 \to 1$ | $2 \to 3$ |
| WavLM-B | $11 \to 12$ | $0 \to 1$ | $0 \to 1$ | $0 \to 1$ |
| wav2vec2-B | $10 \to 11$ | $10 \to 11$ | $11 \to 12$ | $10 \to 11$ |
| Data2Vec | $11 \to 12$ | $11 \to 12$ | $6 \to 7$ | $11 \to 12$ |

*Table 9.* Full best-transcoder metrics. $R^2$ is explained variance of the next layer from sparse latents; Mono is the fraction of active latents whose activation mass is dominated by one class.

| Dataset | Encoder | Src | Tgt | Recon. MSE | $R^2$ | Mean $\ell_1$ | Mono |
|---|---|---|---|---|---|---|---|
| CREMA-D | CLAP-HTSAT | 3 | 4 | .0746 | .9832 | .0829 | .0000 |
| CREMA-D | Data2Vec | 11 | 12 | .0002 | .9802 | .0085 | .0213 |
| CREMA-D | HuBERT-B | 11 | 12 | .0021 | .9470 | .0078 | .1899 |
| CREMA-D | UniSpeech-SAT | 1 | 2 | .0009 | .9402 | .0057 | .1154 |
| CREMA-D | wav2vec2-B | 10 | 11 | .0090 | .9940 | .0263 | .0500 |
| CREMA-D | WavLM-B | 11 | 12 | .0017 | .9409 | .0047 | .0855 |
| CREMA-D | Whisper-S | 10 | 11 | .0054 | .9978 | .1080 | .0000 |
| ESC-50 | CLAP-HTSAT | 0 | 1 | .0061 | .8667 | .0173 | .0263 |
| ESC-50 | Data2Vec | 11 | 12 | .0009 | .9442 | .0107 | .0000 |
| ESC-50 | HuBERT-B | 0 | 1 | .0018 | .9059 | .0038 | .0571 |
| ESC-50 | UniSpeech-SAT | 0 | 1 | .0020 | .8999 | .0047 | .0565 |
| ESC-50 | wav2vec2-B | 10 | 11 | .0224 | .9589 | .0195 | .0932 |
| ESC-50 | WavLM-B | 0 | 1 | .0017 | .8847 | .0040 | .0886 |
| ESC-50 | Whisper-S | 10 | 11 | .0381 | .9895 | .1265 | .0000 |
| GTZAN | CLAP-HTSAT | 0 | 1 | .0042 | .8896 | .0167 | .0000 |
| GTZAN | Data2Vec | 6 | 7 | .0027 | .8586 | .0054 | .0139 |
| GTZAN | HuBERT-B | 0 | 1 | .0015 | .9079 | .0036 | .0000 |
| GTZAN | UniSpeech-SAT | 0 | 1 | .0020 | .8804 | .0046 | .0154 |
| GTZAN | wav2vec2-B | 11 | 12 | .0015 | .9354 | .0067 | .0000 |
| GTZAN | WavLM-B | 0 | 1 | .0016 | .8651 | .0037 | .0112 |
| GTZAN | Whisper-S | 2 | 3 | .0037 | .9326 | .0138 | .0000 |
| US8K | CLAP-HTSAT | 3 | 4 | .1551 | .9746 | .0798 | .7739 |
| US8K | Data2Vec | 11 | 12 | .0005 | .9830 | .0140 | .1905 |
| US8K | HuBERT-B | 0 | 1 | .0010 | .9568 | .0041 | .3214 |
| US8K | UniSpeech-SAT | 2 | 3 | .0013 | .9502 | .0064 | .4714 |
| US8K | wav2vec2-B | 10 | 11 | .0089 | .9740 | .0141 | .1524 |
| US8K | WavLM-B | 0 | 1 | .0011 | .9410 | .0040 | .2958 |
| US8K | Whisper-S | 10 | 11 | .0172 | .9955 | .1311 | .0000 |

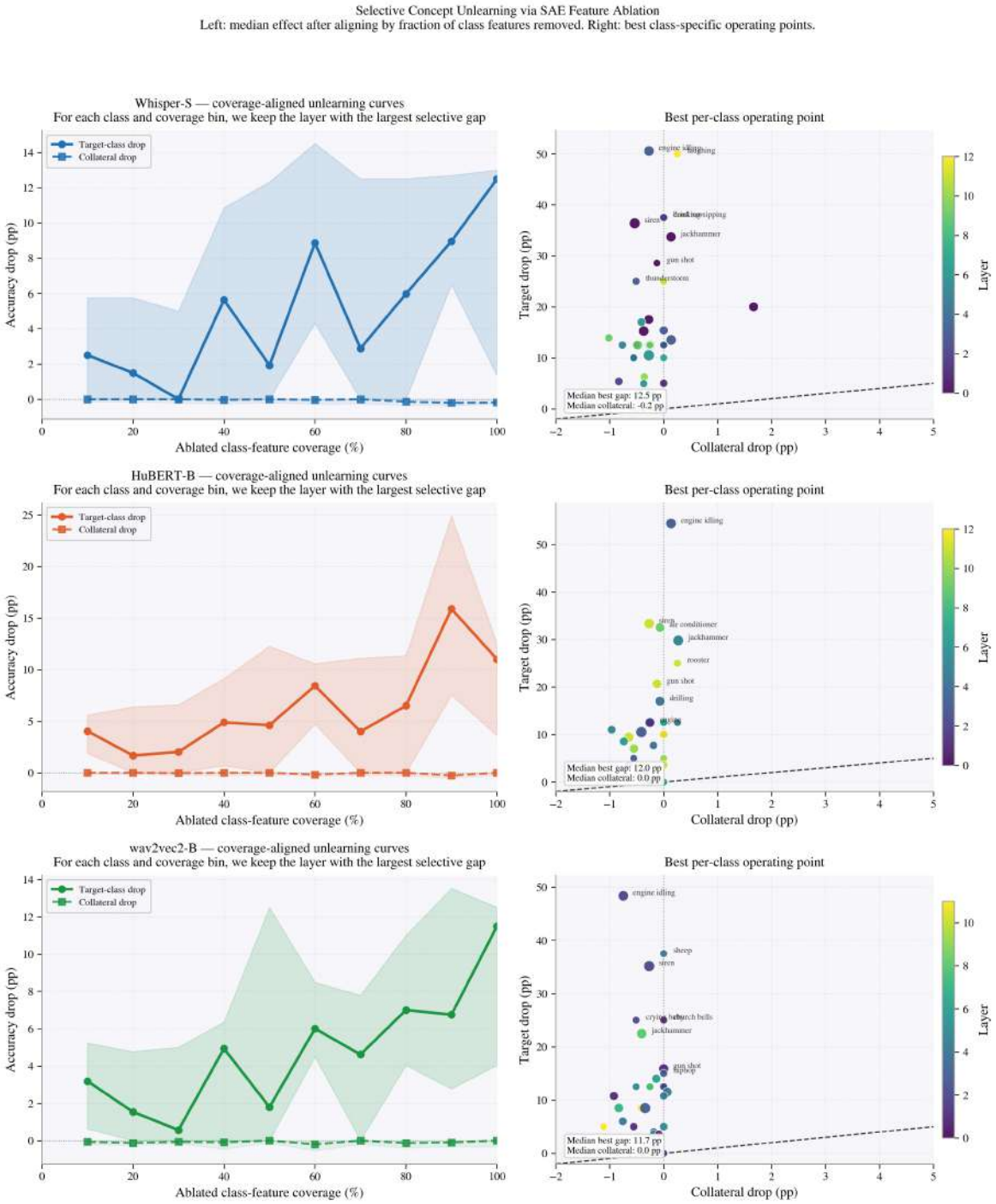


*Figure 10.* Top-$k$ probing and targeted feature-ablation diagnostics. The plot tracks how probe performance changes as increasingly many target-associated sparse features are kept or removed.

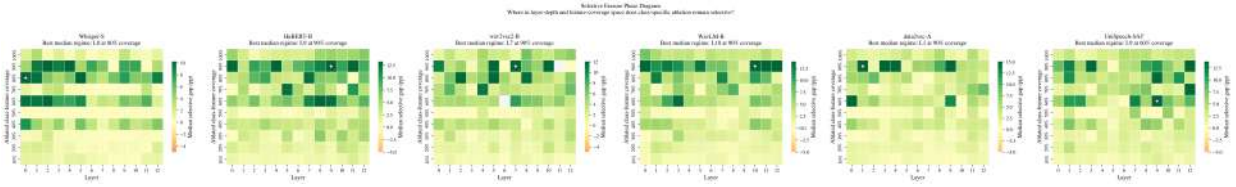

*Figure 11.* Feature-ablation phase diagram over target suppression and retain-set interference. Points in the desirable region suppress the target class while preserving off-target classes.

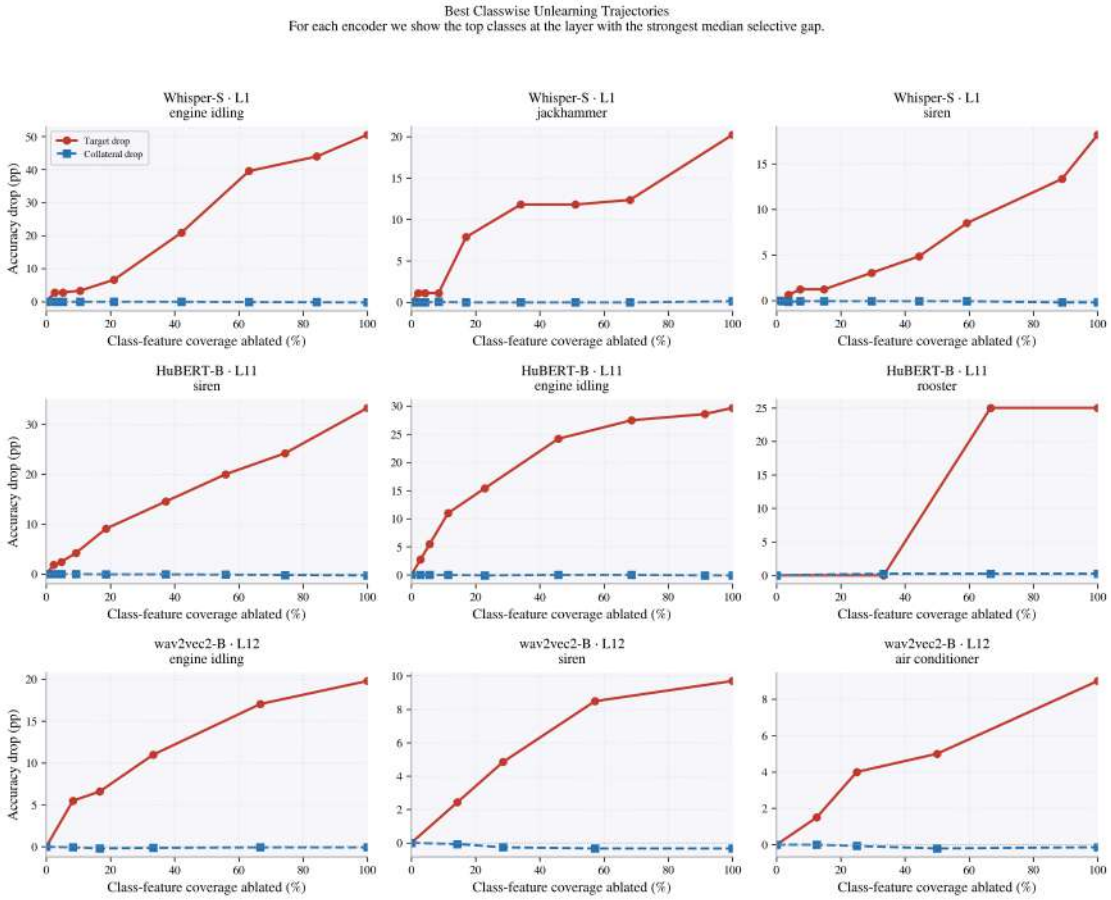


*Figure 12.* Top class-level ablation examples. The grid highlights classes whose SAE features are most removable under the target-suppression criterion and classes where removal spills into acoustically related labels.

*Table 10.* Sparse-feature genealogy event rates for the six common-depth encoders. Rows aggregate available dataset/encoder lineage records.

| Encoder | Persist | Split | Merge | Die | Emerge |
|---|---|---|---|---|---|
| Whisper-S | .001 | .576 | .015 | .191 | .216 |
| wav2vec2-B | .026 | .587 | .039 | .161 | .188 |
| HuBERT-B | .031 | .549 | .046 | .166 | .209 |
| WavLM-B | .031 | .537 | .047 | .184 | .200 |
| Data2Vec-Audio | .029 | .451 | .047 | .223 | .249 |
| UniSpeech-SAT | .041 | .445 | .072 | .208 | .234 |

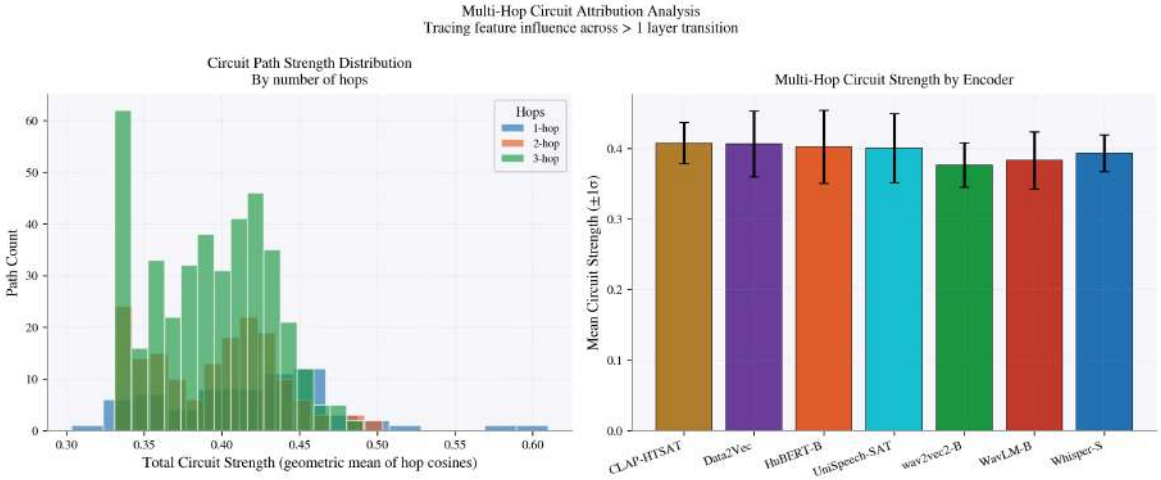


*Figure 13.* Sparse transcoder attribution graph. Nodes are sparse latent features and edges indicate attribution through an inter-layer transcoder.

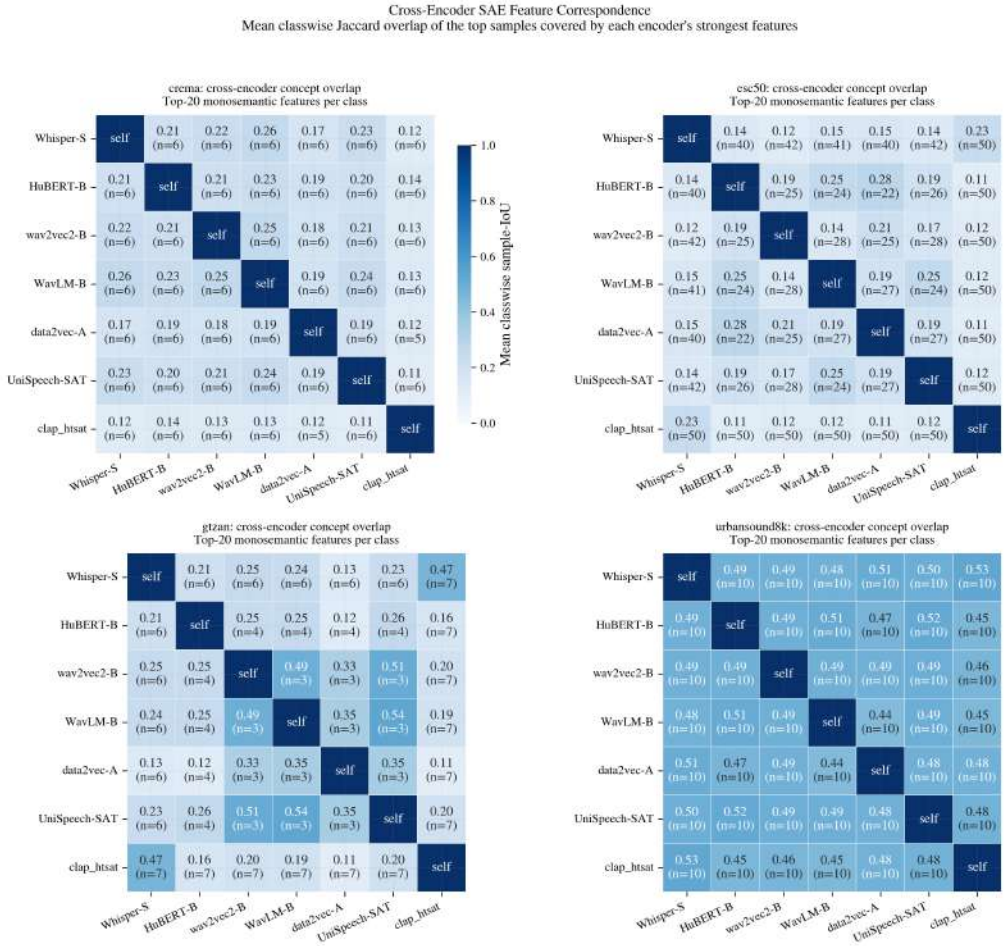


*Figure 14.* Cross-encoder feature overlap measured by activation-profile IoU. High overlap means two encoders activate sparse features on similar examples; low overlap means they carve the same dataset into different internal concepts.

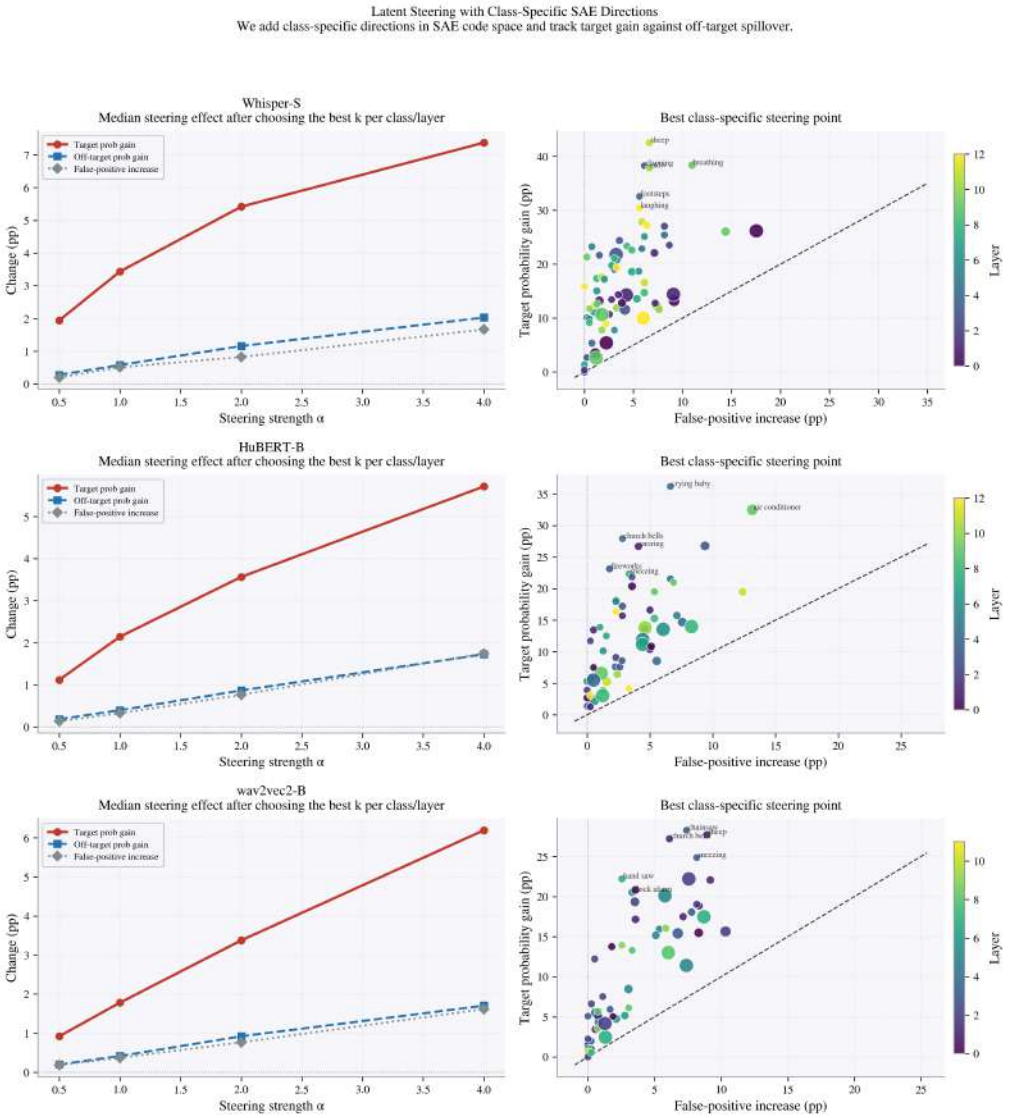


*Figure 15.* Latent steering tradeoff between target gain and off-target movement. The desirable region increases the target class while leaving retain classes unchanged.

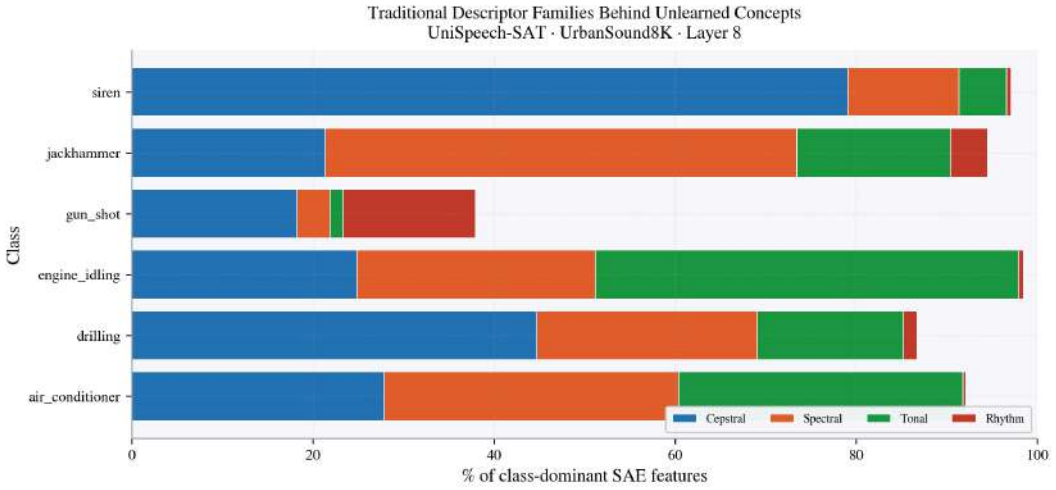


*Figure 16.* Relationship between concept ablation and traditional acoustic feature families. Concepts aligned with spectral or cepstral families can be easier to interpret but may also share evidence with nearby sound classes.

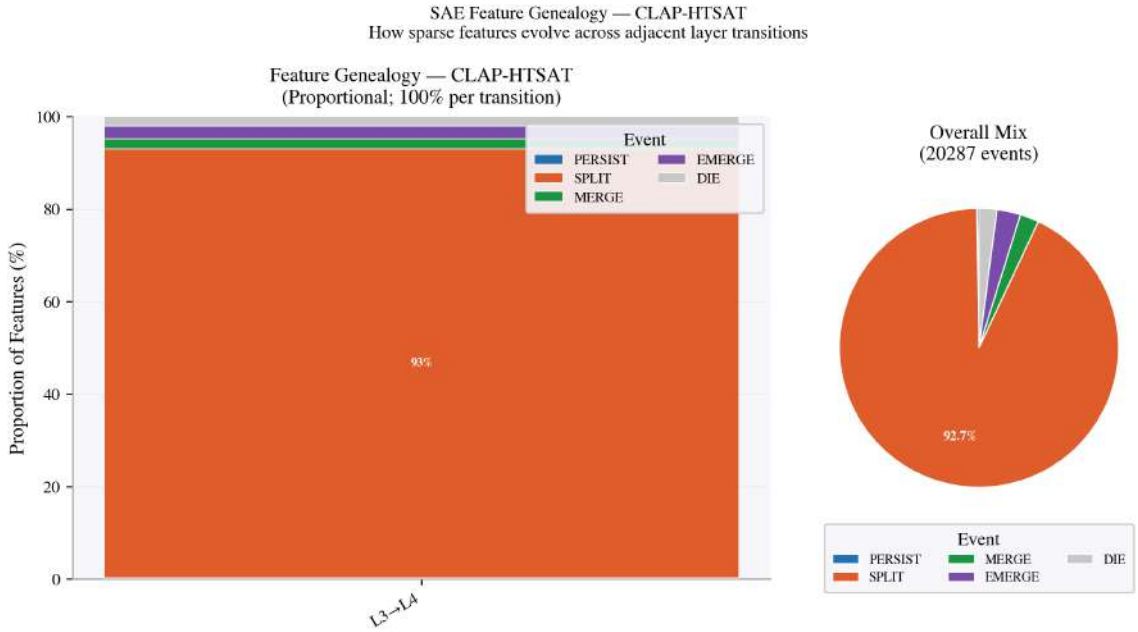


*Figure 17.* Feature genealogy for CLAP-HTSAT. Compact branches indicate that audio-text alignment forms label-aligned features usable for class-level intervention diagnostics.

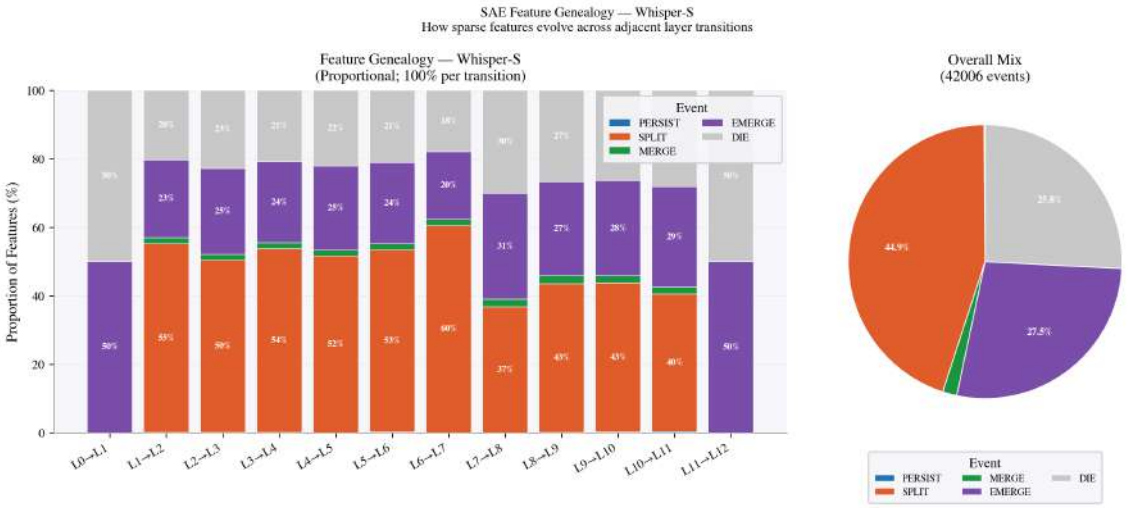


*Figure 18.* Feature genealogy for Whisper-S. Longer chains and repeated merges are consistent with late decoder-aligned, polysemantic routing.

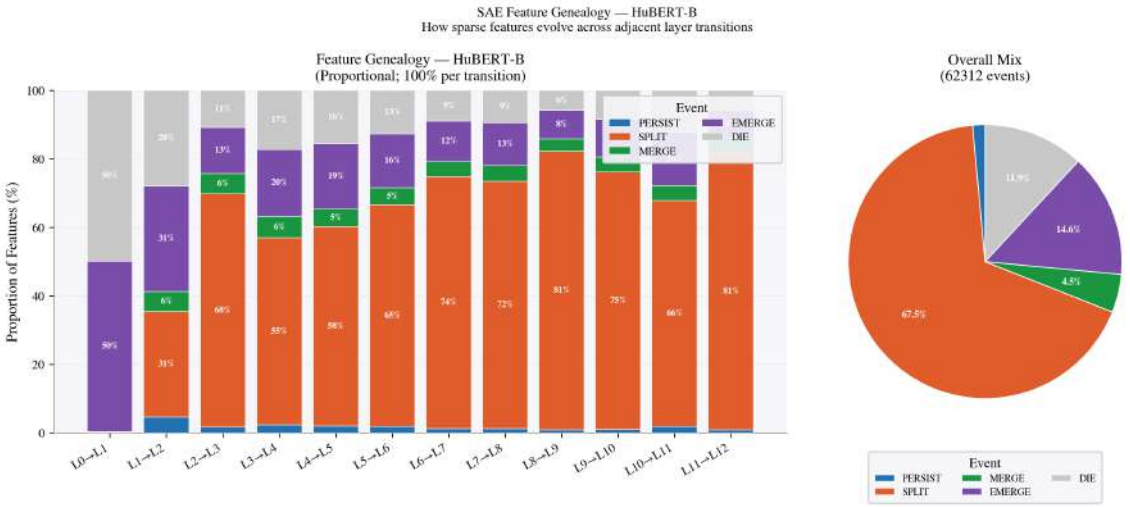


*Figure 19.* Feature genealogy for HuBERT-B. Discrete-unit pretraining produces clustered feature families that specialize early and then recombine.

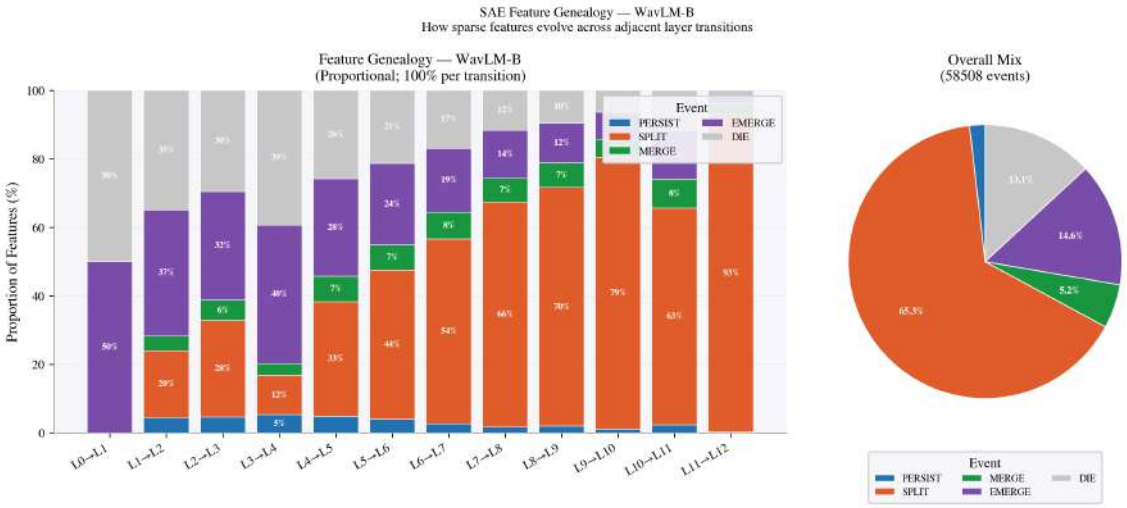


*Figure 20.* Feature genealogy for WavLM-B. Denoising encourages stable acoustic lineages that persist under perturbation more reliably than Whisper-S features.

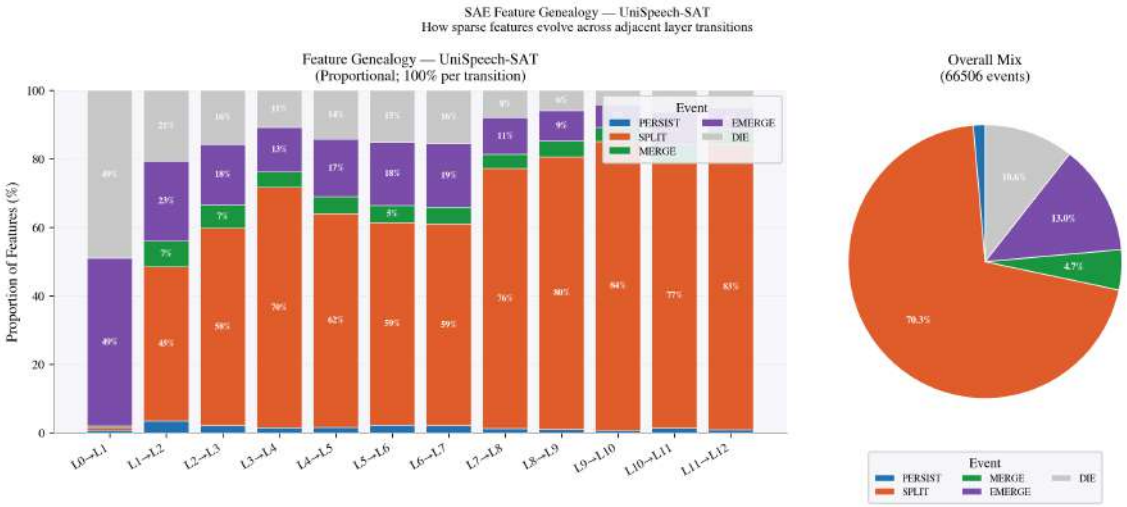


*Figure 21.* Feature genealogy for UniSpeech-SAT. Speaker-aware masked prediction creates intertwined content and speaker feature branches.

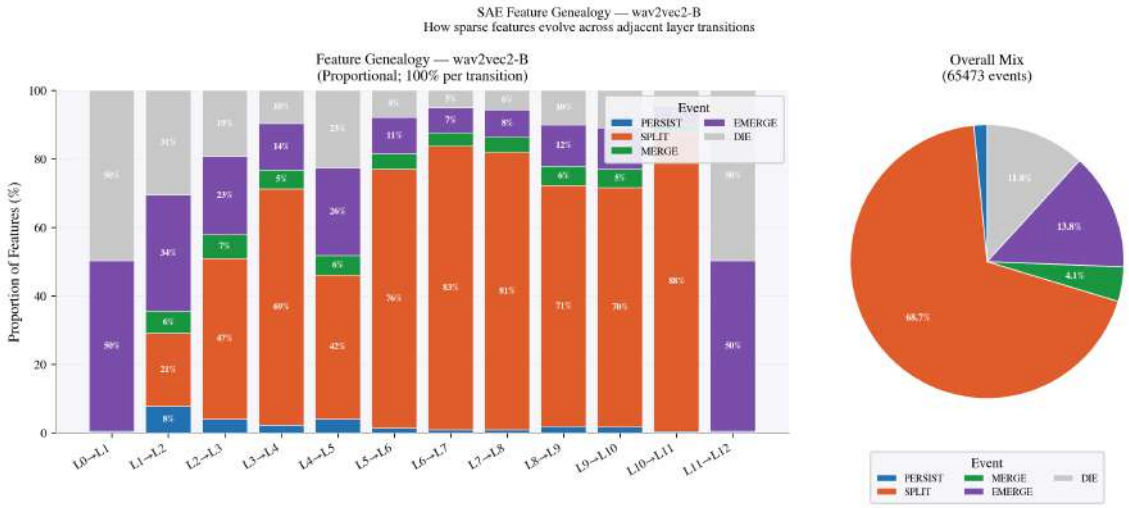


*Figure 22.* Feature genealogy for wav2vec2-B. Contrastive pretraining yields early acoustic branches that become less useful when later layers specialize.

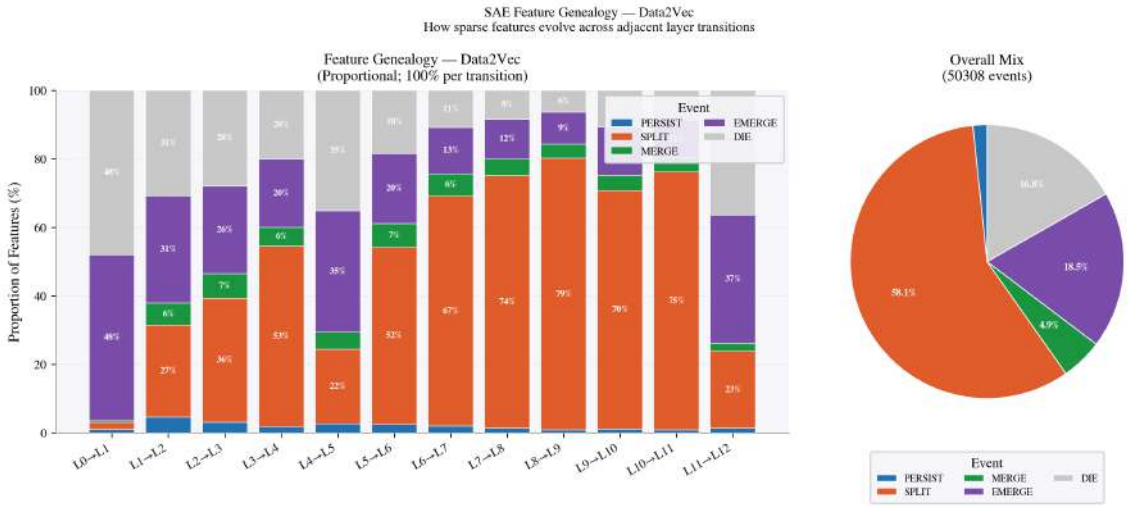


*Figure 23.* Feature genealogy for Data2Vec-Audio. Self-distillation produces a front-loaded genealogy, matching the frequent layer-0 optimum in probe results.