ECG-InterpBench: Benchmarking the Interpretability of ECG Foundation Models with Matched-Scale Sparse Autoencoders

Yixuan Duan
Department of Electrical and Computer Engineering
Rice University
Houston, Texas, USA
yd68@rice.edu

Wei Qiu*
Department of Electrical and Computer Engineering
Rice University
Houston, Texas, USA
wq8@rice.edu

Abstract

Existing benchmarks for electrocardiogram foundation models primarily evaluate downstream predictive performance, providing limited insight into whether their internal representations can be faithfully decomposed, clinically interpreted, or reproduced across independent analyses. We introduce ECG-InterpBench, a benchmark designed to systematically evaluate the interpretability of ECG foundation-model representations. ECG-InterpBench uses sparse autoencoders as standardized measurement instruments and matches their capacity across models to enable controlled comparisons. We evaluate six frozen ECG foundation models across five standardized encoder depths, five matched dictionary widths, and three random seeds, producing a 450-cell interpretability atlas comprising 75 exactly matched six-model comparison blocks. The benchmark evaluates complementary dimensions of representation interpretability, including sparse reconstruction fidelity, single-feature accessibility and coverage of 49 clinically meaningful ECG measurements, and cross-seed feature reproducibility. The evaluation further quantifies patient-sampling uncertainty, depth- and seed-dependent variation, and sensitivity to the sparsity parameterization. The benchmark reveals that ECG foundation models exhibit distinct interpretability profiles. A matched replication on MIMIC-IV-ECG confirms that reconstruction fidelity and clinical accessibility identify different leading models. The benchmark is accompanied by executable evaluation code, standardized manifests, cell-level metrics, and reproducibility audits. ECG-InterpBench complements performance-centered ECG benchmarks by providing a capacity-controlled and reproducible framework for comparing ECG foundation models across distinct dimensions of representation interpretability.

CCS Concepts

• **Computing methodologies** → **Representation of mathematical objects**; *Unsupervised learning*; • **Applied computing** → *Health informatics*; • **Information systems** → Data analytics.

Keywords

Benchmarking, Sparse Autoencoders, Mechanistic Interpretability, Electrocardiogram, Foundation Models, Representation Analysis

1 Introduction

ECG foundation models (ECG FMs) learn general-purpose representations that support diverse rhythm, morphology, and cardiovascular risk-prediction tasks [12, 25, 28]. Existing evaluations, however, focus primarily on downstream accuracy, transfer performance, or label efficiency [18, 26, 37]. These outcomes do not reveal how the hidden representation is organized: whether it can be faithfully decomposed, whether clinically relevant measurements are localized to identifiable features, or whether the same structure is recovered across independent decompositions. This gap matters in clinical settings because a shared representation can propagate acquisition artifacts, demographic confounding, or cohort-specific shortcuts to every downstream model built on it [10, 31, 36].

Existing interpretability tools address related but different questions. Output attribution identifies waveform regions that influence a prediction, but it does not audit the organization of a task-general hidden representation and may fail parameter-randomization or other sanity checks [1, 33]. Linear probes establish whether a clinical variable is decodable, but not whether it is localized to a small number of features, preserved by a faithful sparse decomposition, or recovered across independent SAE trainings. Representation-level interpretability is therefore not equivalent to predictive accuracy, linear decodability, or visually plausible saliency; it requires explicit, complementary measurements [27].

Sparse autoencoders (SAEs) provide a scalable instrument for this audit by mapping dense activations into sparse codes and reconstructing the original representation from learned decoder directions [6, 8, 9]. Their behavior depends on dictionary width and sparsity, making capacity matching central to controlled cross-model measurement [22, 24, 35]. ECG-InterpBench therefore evaluates every FM with the same family of dictionary scales and sparsity settings and summarizes all models over common scale support. This design separates differences among FM representations from differences in SAE measurement capacity.

ECG-InterpBench uses a matched family of SAEs as the standardized measurement instrument and treats the frozen FM representation as the object being evaluated. We operationalize representation interpretability along three separately reported axes: (1) sparse reconstruction fidelity and dictionary utilization; (2) train-selected single-feature accessibility and coverage of clinical ECG concepts; and (3) cross-seed reproducibility of decoder features and selected feature subspaces. Together, these axes provide a structured profile of how faithfully, locally, and reproducibly each representation can be interpreted.

We evaluate six frozen ECG FMs at five standardized relative depths, five matched SAE scales, and three random seeds. The resulting 450-cell atlas contains 75 comparison blocks in which all six models share the same target relative depth, dictionary capacity, sparsity protocol, and seed. Comparisons are made only within

*Corresponding author.

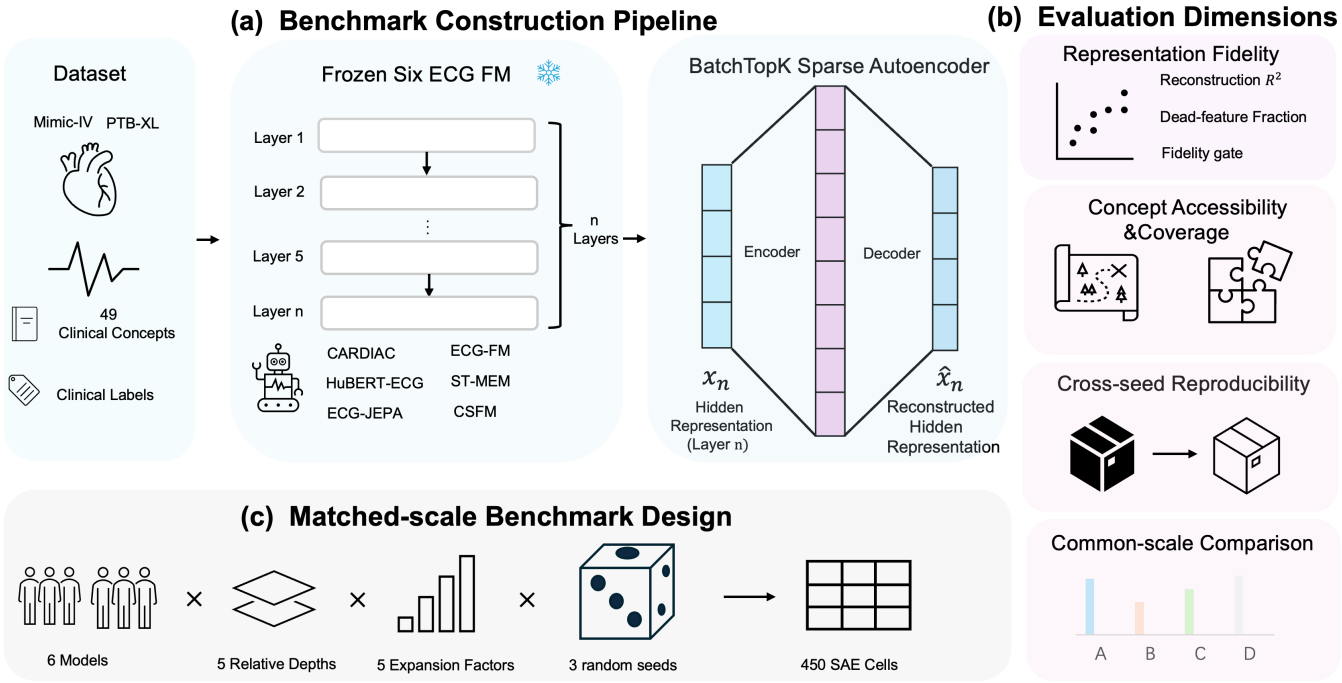

Figure 1: Overview of ECG-InterpBench. (a) PTB-XL and credentialed MIMIC-IV ECGs are passed through six frozen ECG FMs, and layer representations are decomposed with BatchTopK SAEs. (b) Each representation is evaluated separately for reconstruction fidelity, single-feature clinical concept accessibility and coverage, and cross-seed reproducibility. (c) The matched-scale design crosses six models, five relative depths, five expansion factors, and three seeds, producing 450 SAE cells for common-scale model comparison.

these blocks and summarized over the same scale support for every model. We additionally quantify patient-sampling and design uncertainty and test sensitivity to an alternative sparsity parameterization. Figure 1 summarizes this capacity-controlled design.

The benchmark reveals distinct, metric-specific interpretability profiles. Reconstruction fidelity, single-feature clinical accessibility, and decoder reproducibility expose complementary strengths across ECG FMs. A matched MIMIC-IV-ECG replication preserves the separation between reconstruction and accessibility leaders. The resulting profile supports direct model comparison across all measured representation properties.

The benchmark implementation, experiment protocols, and manuscript artifacts are publicly available in our GitHub repository.

Our contributions are:

- We introduce a capacity-controlled benchmark for comparing representation-level interpretability across six ECG FMs under a matched SAE protocol.
- We construct a 450-cell matched-scale interpretability atlas that maps how representation interpretability changes across encoder depth and SAE scale.
- We define leakage-controlled interpretability metrics spanning sparse fidelity, single-feature clinical accessibility and coverage, and cross-seed feature and subspace reproducibility.
- We provide robust interpretability comparisons supported by patient and design uncertainty, sparsity sensitivity, and external-cohort evaluation.

2 Related Work

ECG foundation models. Supervised deep ECG systems established high diagnostic performance for arrhythmia, multi-label 12-lead interpretation, and ventricular-dysfunction screening [2, 13, 30]. PTB-XL subsequently enabled reproducible architecture benchmarking and transfer analysis [34], while contrastive, predictive, and masked-reconstruction objectives showed that useful ECG representations can be learned without task labels [18, 26, 37].

The evaluated encoders use multimodal, contrastive, predictive, clustering, or spatio-temporal objectives. They are CSFM, CARDIAC-FM, ECG-FM, ECG-JEPA, HuBERT-ECG, and ST-MEM [7, 12, 17, 21, 25, 28]. Their released channel selection, sample rate, duration, tokenization, and pooling APIs differ. We preserve those interfaces and standardize only the comparison coordinates—relative depth and SAE capacity.

Sparse representation analysis. SAEs are motivated by sparse coding [23, 29] and have been used to recover interpretable directions from language, biological, and other foundation models [5, 32].

Large-scale dictionary releases and scaling studies show that dictionary width, sparsity, reconstruction, and feature quality interact [8, 9, 22, 35]; BatchTopK makes the batch-level activation budget explicit [6]. Evaluations of SAE fidelity, sparse probing, and intervention further caution that no single metric establishes interpretability [15, 24]. These results motivate a capacity sweep, but they also imply that a cross-model study must match the sweep rather than select a separate favorable capacity for each encoder.

Prior work applies TopK SAEs to EEG transformers across layers and scales, evaluating dictionary health, clinical semantics, steering, and spectral decoding [20]. ECG-InterpBench focuses instead on controlled interpretability comparisons across ECG foundation models. It restricts primary comparisons to common scales, integrates identical scale support for every model, pairs patient resamples across models, freezes concept selection before test evaluation, and evaluates seed-level feature identity against random matching.

Concept and benchmark evaluation. Interpretability depends on the intended audience and use, and medical explanations require evaluation against clinical needs rather than visual plausibility alone [10, 27, 31, 36]. ECG attribution studies can recover diagnostically relevant waveform regions [33], but output attribution answers a different question from auditing the organization of a task-general hidden representation. Concept activation vectors define supervised directions [16], concept bottlenecks expose human-specified intermediate variables [19], and LEACE removes linearly decodable subspaces in closed form [3]. These methods likewise answer different questions from an unsupervised sparse dictionary.

Explanation methods can pass qualitative inspection while failing sanity, leakage, or control tests [1, 14]. SAE studies accordingly report heterogeneous criteria, including one-dimensional binary probe loss, automated explanation, TCAV, intervention selectivity, and reconstruction fidelity [9, 15, 20]. Those values are not numerically commensurate with held-out Pearson correlation for continuous ECG measurements, so we do not assign our correlation scale a literature-wide "strong alignment" threshold. Our benchmark instead uses it as a leakage-controlled relative comparison under matched ECG-FM depth–scale blocks, while retaining every failed cell in a fixed denominator.

3 Benchmark Design

3.1 Source cohorts, frozen models, and representations

PTB-XL supplies 21,799 twelve-lead ECGs and 49 waveform-derived measurement and morphology concepts [38]. We use the existing patient-disjoint split: 15,149 records for training, 3,325 for validation, and 3,325 for test. The test set contains 2,862 unique patients. Every model receives the same ordered record manifest, and preprocessing statistics are estimated on training records only. PTB-XL is the source cohort for the primary benchmark. For external validation, we repeat the complete 450-cell multiscale SAE audit on a credentialed, patient-partitioned MIMIC-IV-ECG manifest of 100,000 ECGs from 48,491 patients (Appendix E).

All six pretrained encoders remain frozen throughout the benchmark. We update no FM parameters and perform no task-specific fine-tuning; only the SAEs are trained on activations extracted from the fixed checkpoints. Each encoder is called through its released preprocessing path (Appendix C, Table S3). We audit pre-projection transformer-layer states for CARDIAC-FM; its released 768-to-512 pooled projection is used only in the downstream stage and is not the dimension of the source layer atlas. Consequently, all six source layer atlases have $d = 768$.

3.2 Standardized depth and exact scale matching

For an encoder exposing L audited states, target relative depth $q \in \{0, .25, .5, .75, 1\}$ maps to

$$\ell(q) = \left\lfloor q(L-1) + \tfrac{1}{2} \right\rfloor. \tag{1}$$

We record both q and the realized depth $\ell/(L-1)$. CSFM maps to layers $(0, 1, 3, 4, 5)$; each 12-state encoder maps to $(0, 3, 6, 8, 11)$. Relative depth provides a shared coordinate for aligning early, intermediate, and final representation stages across encoders with different layer counts. This alignment enables systematic depth-wise comparison of where reconstruction fidelity, clinical accessibility, and reproducible sparse structure emerge across models.

For a batch of standardized activations $X \in \mathbb{R}^{B\times d}$, BatchTopK computes positive pre-activations and retains the largest Bk values over the complete batch:

$$P = \mathrm{ReLU}\big((X - b_d)W_e^\top + b_e\big), \tag{2}$$

$$Z = \mathrm{BatchTopK}_{Bk}(P), \qquad \widehat{X} = ZW_d^\top + b_d. \tag{3}$$

Decoder columns are constrained to unit norm. Thus k specifies the batch-average active-feature budget, while the reported realized mean L_0 summarizes per-record activation sparsity.

The primary grid fixes expansion $E = N/d \in \{1, 4, 8, 16, 32\}$ and $k/d = 1/8$. Because $d = 768$ for every source layer, all models use exactly the same widths

$$N \in \{768, 3072, 6144, 12288, 24576\} \tag{4}$$

and the same $k = 96$ at every scale. Each SAE is trained for 8,000 steps with batch size 256, Adam learning rate 3×10^{-4}, and seed 4311, 4312, or 4313. The complete grid contains

$$6 \text{ models} \times 5 \text{ depths} \times 5 \text{ scales} \times 3 \text{ seeds} = 450 \tag{5}$$

cells, arranged as 75 blocks in which all six FMs share the same target depth, N, k, and seed. Primary contrasts pair all models at the same prespecified E and summarize performance over the shared scale grid, with scale choices fixed before test evaluation.

3.3 Leakage-controlled single-coordinate clinical accessibility

Concept values are aligned by ECG identifier and standardized with training-split means and standard deviations. A missing or non-finite value is replaced by the training mean, which is zero in standardized coordinates. Coordinate selection uses the same deterministic 4,096-record subset of the training split in every cell. For concept c and SAE code coordinate j, let r_{jc}^{train} denote their correlation on that subset. We select

$$j_c^* = \arg\max_j |r_{jc}^{\mathrm{train}}| \tag{6}$$

once and evaluate the same coordinate on validation and test records. The test single-coordinate accessibility score is the mean

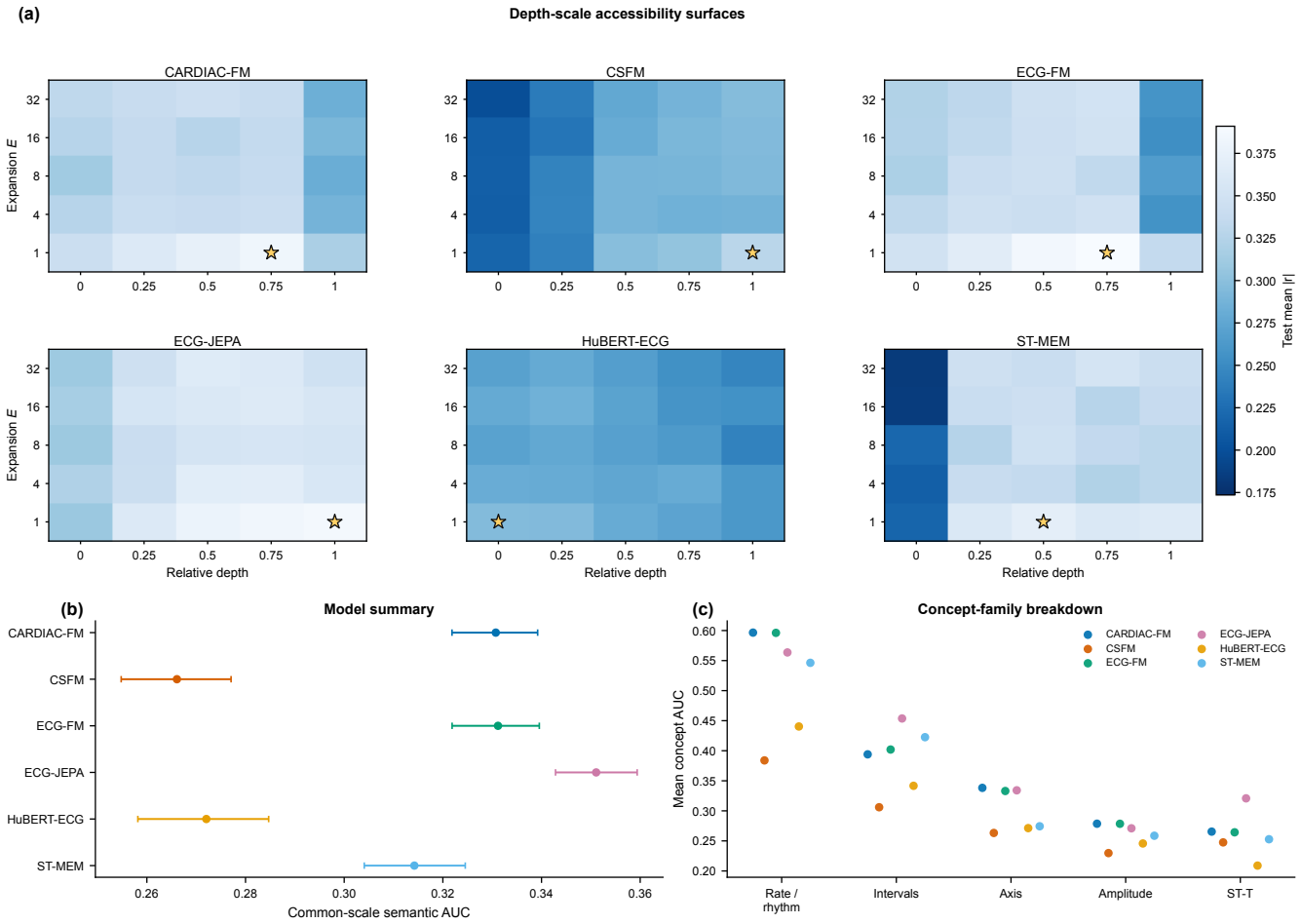

Figure 2: Single-coordinate clinical accessibility across the matched SAE atlas. (a) Test mean absolute concept correlation over five relative depths and five exactly matched expansion factors; circles mark each model's peak cell. (b) Common-scale semantic AUC with paired patient-cluster 95% intervals. (c) Mean concept-level common-scale AUC by ECG concept family. Features are selected on the training subset and frozen before held-out evaluation.

$|r^{\text{test}}_{j^*_c c}|$ over 49 concepts; coverage is the fraction with $|r^{\text{test}}_{j^*_c c}| \geq .20$. Training-only selection fixes every SAE coordinate, layer, and scale before held-out evaluation. This estimand quantifies how strongly one sparse coordinate locally exposes each concept, and the .20 cutoff provides a consistent descriptive operating point across models.

3.4 Fixed-scale accessibility calibration

At the common $E = 8$ cells ($N = 6144, k = 96$), we compare seven readouts: fixed-$\alpha = 10$ ridge from the dense state, full SAE code, or 16/four train-ranked SAE coordinates; and fit-free train-selected native, SAE, or random coordinates. Dense coordinates are evaluated once per model–depth, whereas SAE metrics average three seeds. The random control averages 20 shared seeds, each projecting the normalized state onto N Gaussian unit directions before the same ReLU, BatchTopK, batch order, and training-only selection. We report held-out mean $|r|$, coverage at $|r| \geq .20$, paired differences, and random-seed percentile intervals. Coordinates and readout settings are fixed from the training data before validation and test outcomes are used for held-out evaluation.

3.5 Metric-specific FM profiles

Measurement fidelity. We report normalized reconstruction $R^2 = 1 - \text{SSE}/\text{SST}$, dead-feature fraction, realized mean L_0, and the fraction of layer–scale cells satisfying the preregistered validation fidelity gate ($R^2 \geq .90$ and dead fraction $< .20$). Reconstruction quantifies how faithfully the sparse instrument preserves a representation and is interpreted alongside clinical accessibility.

Single-coordinate clinical accessibility. Train-selected test correlation and concept coverage quantify how locally the fixed concept panel is exposed by individual sparse-code coordinates. Their primary role is relative model comparison under an identical selection and capacity protocol. Joint interpretation with reconstruction fidelity connects local concept exposure to the quality of the underlying sparse decomposition.

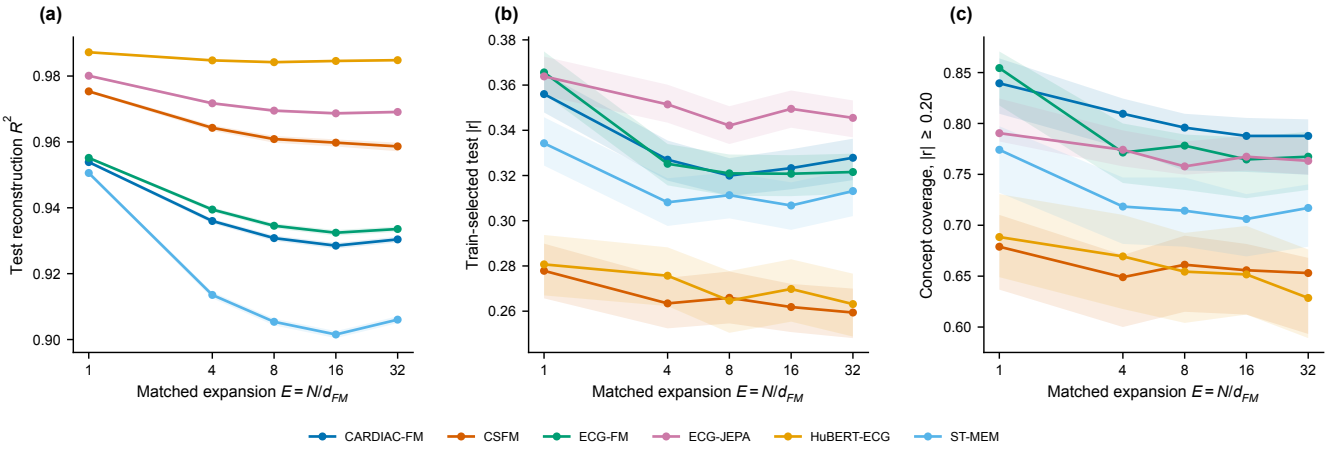

Figure 3: Patient-bootstrap FM profiles at each common SAE scale: (a) reconstruction R^2, (b) train-selected single-coordinate clinical accessibility, and (c) concept coverage at $|r| \geq 0.20$. Lines compare models at identical E and therefore identical absolute N; bands are paired 95% patient-cluster intervals.

Reproducibility. Within each model, depth, and scale, we retain up to the 256 most active non-dead decoder directions per seed according to validation firing rate. Hungarian matching maximizes signed decoder cosine for each seed pair. We report matched cosine above a 100-permutation random-pairing floor and normalized overlap between the selected decoder subspaces.

Common-scale summary. For metric curve $m(E)$, the model profile is the normalized log-scale area

$$\mathrm{AUC}_E(m) = \frac{1}{\log 32 - \log 1} \int_{\log 1}^{\log 32} m(e^u)\, du, \tag{7}$$

Here, E is the SAE expansion factor, evaluated at $\{1, 4, 8, 16, 32\}$, and $m(E)$ may denote reconstruction R^2, single-coordinate mean $|r|$, concept coverage, matched decoder cosine, or selected-subspace overlap. We approximate the integral by trapezoidal integration over these five points in $\log E$. Division by $\log 32 - \log 1$ makes AUC_E a normalized average across SAE scales. The resulting scale summary is then averaged over the five depths and three seeds as specified by each inferential analysis. Fidelity, single-coordinate accessibility, coverage, and stability jointly define the model's multiscale interpretability profile.

3.6 Paired uncertainty and multiplicity

The primary test inference resamples PTB-XL patients with replacement 2,000 times and retains all records belonging to a sampled patient. The sorted patient set and bootstrap seed are identical across all 450 cells, enabling paired FM differences. BatchTopK codes are computed once in the frozen original test order with batch size 256; bootstrap weights are then applied to patient-aggregated sufficient statistics, preserving a common encoding across all resamples. Reconstruction uses the frozen full-test reference mean in the SST term for every bootstrap draw. These intervals quantify held-out patient sampling uncertainty conditional on the frozen reference, FM and SAE weights, and train-selected coordinates. The three SAE seeds are averaged within each draw.

A second 10,000-sample crossed bootstrap resamples the five depth units and three SAE seeds to quantify design variation after integrating the frozen scale curve. Pairwise model contrasts are adjusted by Benjamini–Hochberg within each metric family [4]. Scale sensitivity is summarized by Kendall rank correlation between every pair of common E values. Together, patient and design intervals characterize complementary sources of uncertainty: held-out patient sampling and depth–seed design variation.

3.7 Sparsity sensitivity

The primary fixed-k/d arm holds the absolute active budget constant while dictionary width grows. A preregistered mid-depth sensitivity arm instead fixes $k/N = 1/64$, giving $k \in \{12, 48, 96, 192, 384\}$ over the same five exact widths, six models, and three seeds. At relative depth $q = 0.5$, this arm retains the primary training protocol and uses $N = \{768, 3072, 6144, 12288, 24576\}$ with $k = N/64 = \{12, 48, 96, 192, 384\}$. It evaluates the stability of metric-specific FM ordering under an alternative sparsity parameterization, with both arms and their comparison preregistered before test evaluation.

4 Results

4.1 ECG FMs have distinct depth–scale decomposition surfaces

Figure 2 summarizes single-coordinate clinical accessibility at three resolutions: test depth–scale surfaces for every FM, paired patient-bootstrap common-scale semantic AUCs, and concept-family profiles. The corresponding reconstruction and dead-feature atlases are included in Appendix D. These surfaces are the primary result object: a model can localize clinical measurements more strongly at one depth while requiring a wider dictionary or exhibiting less stable atom identity elsewhere. We therefore retain all 25 depth–scale cells per model as the prespecified multiscale evaluation surface.

Table 1: Final-layer input-harmonization audit. Δ values are changes in test mean $|r|$ from the native protocol. Coverage is the joint-protocol fraction of 49 concepts reaching $|r| \geq 0.20$; CKA compares joint and native test representations.

Model	Native mean $\|r\|$	Δ Lead	Δ Temporal	Δ Joint	Joint coverage	Joint CKA
CARDIAC-FM	0.2922	+0.0000	−0.0058	−0.0058	0.857	0.991
CSFM	0.3321	−0.0004	+0.0000	−0.0004	0.837	**1.000**
ECG-FM	0.3166	−0.0000	−0.0048	−0.0048	**0.959**	0.997
ECG-JEPA	**0.4380**	+0.0000	+0.0000	+0.0000	**0.959**	**1.000**
HuBERT-ECG	0.2923	**+0.0006**	**+0.0005**	**+0.0006**	0.735	**1.000**
ST-MEM	0.3891	−0.0000	−0.0131	−0.0131	0.898	0.770

Table 2: MIMIC-IV-ECG held-out multiscale SAE profiles. Reconstruction, semantic alignment, and coverage are normalized expansion-curve AUCs averaged over five relative depths. Stability is matched decoder-feature cosine across SAE seeds; subspace stability is normalized overlap of the selected decoder subspaces. Higher is better except for dead-feature fraction.

Model	Recon.	Semantic	Coverage	Dead fraction	Feature stability	Subspace stability
CARDIAC-FM	0.958	0.295	0.674	0.166	0.366	0.718
CSFM	0.979	0.261	0.582	0.380	0.293	**0.778**
ECG-FM	0.955	0.282	0.625	**0.099**	**0.378**	0.702
ECG-JEPA	0.985	**0.324**	**0.781**	0.447	0.340	0.748
HuBERT-ECG	**0.990**	0.279	0.610	0.546	0.323	0.762
ST-MEM	0.960	0.274	0.600	0.211	0.348	0.760

Interpretability is consequently represented as a structured profile of sparse reconstructability, single-coordinate clinical accessibility, coverage, and feature reproducibility. Common-scale AUC provides a capacity-robust comparison within each axis, while the underlying depth–scale surfaces support operating-point selection for specific analysis goals.

4.2 Matched-scale curves produce metric-specific FM orderings

Figure 3 compares all FMs only at the same E. Shaded regions are paired patient-cluster intervals after averaging the three fixed SAE seeds and five target depths. For reconstruction, the highest common-scale point estimate is HuBERT-ECG (AUC 0.985, 95% CI [0.985, 0.985], bootstrap rank-1 probability 1.000). The complete reconstruction order is HuBERT-ECG > ECG-JEPA > CSFM > ECG-FM > CARDIAC-FM > ST-MEM, with 15 of 15 pairwise contrasts surviving BH correction.

Single-coordinate clinical accessibility has a separate ordering: ECG-JEPA > ECG-FM > CARDIAC-FM > ST-MEM > HuBERT-ECG > CSFM, with 13 of 15 FM contrasts surviving BH correction. The matched model-level profile is complemented by substantial concept-level resolution: for ECG-JEPA, the 49 concept-specific common-scale AUCs have median 0.406 and interquartile range 0.250–0.442; 26 concepts reach at least 0.40, eight reach at least 0.50, and four reach at least 0.60. Ventricular rate, QRS duration, P-wave detection, and mean RR are among the most locally accessible concepts. Concept coverage similarly orders the models as CARDIAC-FM > ECG-FM > ECG-JEPA > ST-MEM > HuBERT-ECG > CSFM, with 13 BH-significant pairwise contrasts. Exact common-scale AUC estimates and paired confidence intervals are provided in Appendix Table S1.

Appendix E reports the MIMIC-IV-ECG replication, preserving metric-specific leaders. Table 1 summarizes the final-layer input-harmonization audit detailed in Appendix F; the six-model accessibility order is preserved under the harmonized protocols.

4.3 External validation preserves metric-specific leaders

We repeated the complete 450-cell multiscale SAE audit on a credentialed, patient-partitioned MIMIC-IV-ECG cohort; cohort construction and sensitivity analyses are detailed in Appendix E. Table 2 reports the held-out model profiles after integrating the same five-point expansion curve and averaging over the same five relative depths.

The replication reveals complementary leaders across interpretability dimensions. HuBERT-ECG has the highest reconstruction profile (0.990), whereas ECG-JEPA has the highest single-feature semantic alignment (0.324) and concept coverage (0.781). ECG-FM has the lowest dead-feature fraction (0.099) and the highest feature stability (0.378), while CSFM has the highest subspace stability (0.778). Together, reconstruction fidelity, local clinical accessibility, dictionary utilization, and cross-seed reproducibility provide a multidimensional comparison of the six models.

4.4 A calibration ladder separates availability from localization

Detailed calibration results are provided in Appendix Table S2. The full SAE retains 79.9–87.4% of dense-ridge mean $|r|$. Single SAE coordinates outperform the mean of 20 matched random dictionaries by 0.034–0.088 across every model's five depths, with 78.8–87.8% concept–depth wins. These held-out results demonstrate consistent

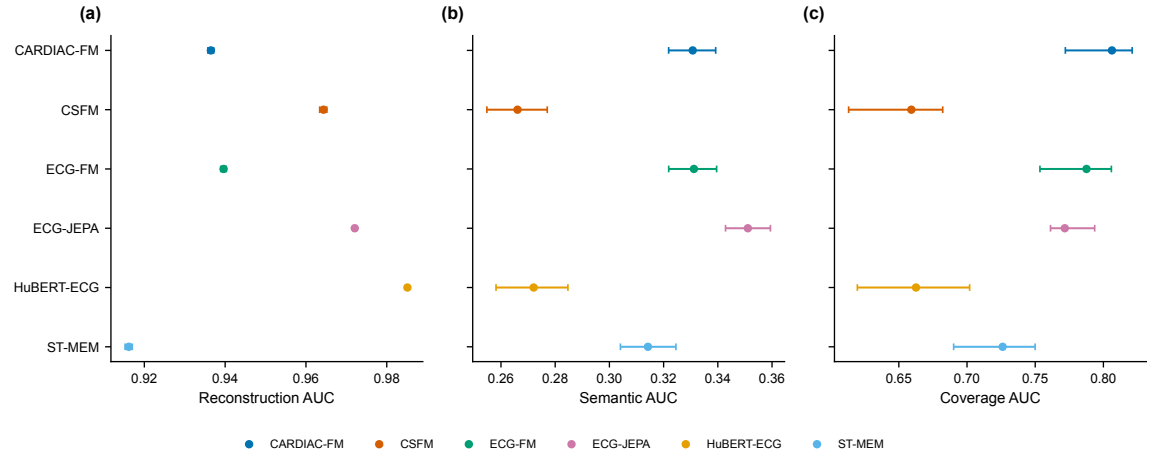

Figure 4: Metric-specific common-scale AUC estimates and paired patient-cluster intervals for (a) reconstruction, (b) single-coordinate clinical accessibility, and (c) concept coverage. A higher value is favorable within each panel, and the panels jointly characterize complementary dimensions of representation interpretability.

recovery of clinically aligned sparse structure by the learned SAE dictionaries across all six ECG FMs.

This calibration connects distributed concept availability with sparse localization. Full-code readouts quantify preserved clinical information, and single-coordinate readouts identify aligned sparse features across all six FMs. Together, these readouts establish the SAE profile as a common sparse instrument for comparing FMs.

The single-coordinate profile quantifies how strongly each clinical measurement is localized to one train-selected SAE coordinate and retained under held-out evaluation. Its 49-concept macro-average supports standardized model-level comparison, while the concept-level distribution reveals a substantial high-accessibility subset and resolves differences across waveform measurements. Together, these views provide both a common summary and concept-specific resolution of sparse clinical accessibility.

4.5 Rank stability is itself a benchmark result

Rank stability across SAE capacities is quantified over all ten pairs of common E values. Reconstruction rank agreement has median Kendall $\tau = 1.00$ and minimum $\tau = 1.00$; single-coordinate accessibility has median $\tau = 0.73$ and minimum $\tau = 0.73$; coverage has median $\tau = 0.73$ and minimum $\tau = 0.60$. Figure 4 presents common-scale AUC as a capacity-robust summary, complemented by full curves for decisions tied to a particular SAE budget.

Rank agreement across E directly tests whether an FM conclusion survives measurement capacity. The shared expansion grid provides a capacity-aware assessment of FM ordering under a common evaluation protocol. The current atlas also matches absolute width because every audited layer has $d = 768$.

4.6 Feature identity and sparse fidelity are different axes

Cross-seed matching asks whether a favorable fidelity or accessibility profile is implemented by reproducible decoder directions. The highest above-random matched-cosine AUC belongs to ECG-FM (0.371), whereas the highest selected-subspace overlap AUC belongs to CSFM (0.757). Figure 5(a)–(b) shows how both quantities change over common scale. Joint reporting of stability, reconstruction, and single-coordinate clinical accessibility captures complementary representation properties.

Feature-level matching and selected-subspace overlap resolve two complementary forms of reproducibility across SAE seeds.

4.7 Model ordering is tested under a second sparsity parameterization

To isolate the effect of the active-code budget, we compare the primary fixed-k/d arm with a preregistered fixed-k/N arm at relative depth $q = 0.5$. Because $d = 768$, the primary arm uses $k = 96$ at every E, whereas the sensitivity arm uses $k = \{12, 48, 96, 192, 384\}$ for $E = \{1, 4, 8, 16, 32\}$; dictionary widths, models, seeds, and training settings remain matched. Reconstruction ordering is highly stable across the two arms (median/minimum Kendall $\tau = 1.00/0.87$), while accessibility (0.73/0.60) and coverage (0.87/0.47) respond more to the active budget, with the largest coverage reordering at $E = 16$. Relative to fixed-k/d, fixed-k/N changes common-scale AUC by -0.012 to -0.042 for reconstruction, -0.004 to -0.014 for accessibility, and -0.027 to $+0.016$ for coverage. At the shared $E = 8$ anchor, both arms use $N = 6144$, $k = 96$, and independently trained cells differ by at most 0.0272. Figure 5(c)–(e) displays the corresponding model-level profiles.

5 Limitations and Claim Boundary

First, the current benchmark covers six representative ECG foundation models spanning a diverse set of encoder architectures and pretraining objectives. Although this collection enables controlled comparisons across these models, future versions could incorporate newly released ECG foundation models and additional model scales to further broaden the benchmark.

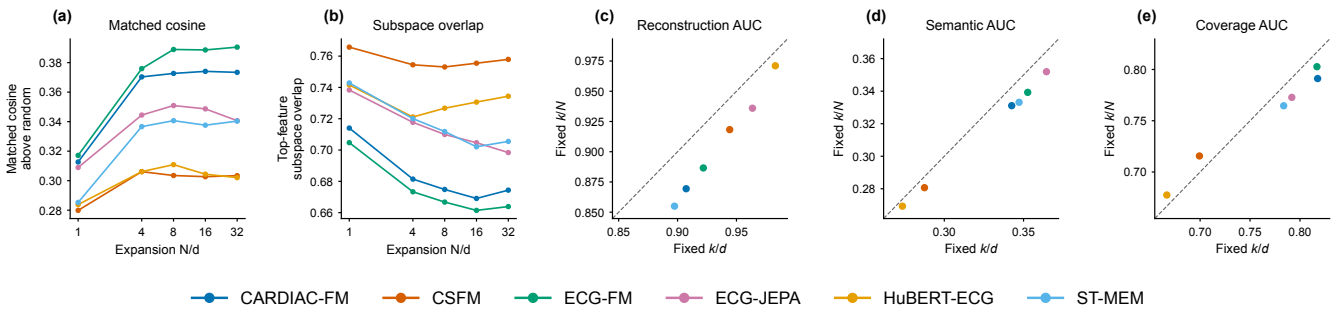

Figure 5: Robustness axes. (a) Cross-seed decoder matching above random and (b) selected-subspace overlap over common scale. Panels (c)–(e) show mid-depth common-scale AUC for reconstruction, clinical accessibility, and concept coverage under fixed k/d and fixed k/N; the diagonal denotes unchanged profiles.

Second, the clinical concept inventory emphasizes waveform measurements and diagnostic concepts that can be evaluated consistently across models and cohorts. Extending the inventory to finer-grained morphological patterns, continuous physiological attributes, longitudinal changes, and expert-annotated concepts could provide a more comprehensive view of representation interpretability.

Third, ECG-InterpBench uses a standardized SAE capacity and sparsity grid to ensure matched comparisons across encoders. Future work could evaluate finer-grained capacity ranges, alternative sparsity objectives, and additional dictionary-learning methods while preserving the same capacity-controlled evaluation protocol. These extensions would also facilitate adaptation of the benchmark to foundation models for EEG, medical imaging, electronic health records, and other clinical modalities.

Throughout, ECG-InterpBench characterizes the structure of frozen model representations under a matched SAE protocol; it does not establish physiological mechanisms, clinical utility, or prospective patient outcomes. The reported model orderings are therefore specific to the evaluated representation properties and should not be interpreted as rankings of diagnostic performance or overall clinical value.

6 Conclusion

ECG-InterpBench uses a calibrated multi-scale SAE family to audit and compare ECG foundation-model representations. Its core unit is an exact six-model matched block. The resulting depth–scale profiles, paired patient inference, seed stability, sparsity sensitivity, and dense-to-single calibration distinguish faithful sparse reconstruction, distributed concept availability, single-coordinate accessibility, and reproducible feature identity. This design turns ECG-FM interpretability into a systematic and reproducible model-comparison problem. More broadly, ECG-InterpBench provides a general framework for identifying where, at what capacity, and how consistently clinically meaningful structure emerges within foundation-model representations. By connecting representation geometry with clinical concepts across models and scales, it establishes a foundation for the principled development and selection of interpretable ECG foundation models. The same capacity-controlled auditing framework can be extended to other clinical foundation models, including those for medical imaging, electroencephalography, electronic health records, physiological time series, and multimodal patient data. With modality-specific clinical concepts and matched representation interfaces, this framework can support systematic interpretability comparisons across a broader range of medical AI systems.

7 Ethics and Responsible Use

All analyses use de-identified secondary ECG datasets under their respective licenses and data-access agreements. No new participants were recruited, and no clinical intervention was performed. PTB-XL provides the directly accessible public reproduction path for the benchmark, while MIMIC-IV-ECG is a publicly available dataset that requires credentialed access through an application and approval process. Released artifacts exclude restricted waveforms, protected metadata, and record-level MIMIC identifiers.

ECG-InterpBench is a research benchmark for evaluating the representation-level interpretability of ECG foundation models. Its outputs are not intended for direct use in diagnosis, treatment, patient management, or other clinical decision-making settings.

The accompanying artifact includes executable evaluation code, standardized manifests, aggregate and configuration-level metrics, fixed random and bootstrap seeds, audit outputs, and figure-generation scripts. The public PTB-XL workflow supports end-to-end reproducible evaluation, while the MIMIC-IV-ECG analysis provides an additional reproduction path for approved credentialed users. Model weights and source datasets remain governed by their original repositories, licenses, and access requirements.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity Checks for Saliency Maps. In *Advances in Neural Information Processing Systems*, Vol. 31. https://arxiv.org/abs/1810.03292

[2] Zachi I. Attia, Suraj Kapa, Francisco Lopez-Jimenez, Paul M. McKie, Dorothy J. Ladewig, Gaurav Satam, Patricia A. Pellikka, Maurice Enriquez-Sarano, Peter A. Noseworthy, Thomas M. Munger, Samuel J. Asirvatham, Christopher G. Scott, Rickey E. Carter, and Paul A. Friedman. 2019. Screening for Cardiac Contractile Dysfunction Using an Artificial Intelligence-Enabled Electrocardiogram. *Nature Medicine* 25 (2019), 70–74. doi:10.1038/s41591-018-0240-2

[3] Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. LEACE: Perfect Linear Concept Erasure in Closed Form. In *Advances in Neural Information Processing Systems*, Vol. 36. https://arxiv.org/abs/2306.03819

[4] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57, 1 (1995), 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x

[5] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. 2023. Towards Monosemanticity: Decomposing Language Models with Dictionary Learning. *Transformer Circuits Thread* (2023). https://transformer-circuits.pub/2023/monosemantic-features/index.html

[6] Bart Bussmann, Patrick Leask, and Neel Nanda. 2024. BatchTopK Sparse Autoencoders. *arXiv preprint arXiv:2412.06410* (2024). https://arxiv.org/abs/2412.06410

[7] Edoardo Coppola, Mattia Savardi, Mauro Massussi, Marianna Adamo, Marco Metra, and Alberto Signoroni. 2024. HuBERT-ECG as a Self-Supervised Foundation Model for Broad and Scalable Cardiac Applications. *medRxiv* (2024). doi:10.1101/2024.11.14.24317328

[8] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2024. Sparse Autoencoders Find Highly Interpretable Features in Language Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=F76bwRSLeK

[9] Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2025. Scaling and Evaluating Sparse Autoencoders. In *International Conference on Learning Representations*. https://openreview.net/forum?id=tcsZt9ZNKD

[10] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L. Beam. 2021. The False Hope of Current Approaches to Explainable Artificial Intelligence in Health Care. *The Lancet Digital Health* 3, 11 (2021), e745–e750. doi:10.1016/S2589-7500(21)00208-9

[11] Brian Gow, Tom Pollard, Larry A. Nathanson, Alistair Johnson, Benjamin Moody, Catarina Fernandes, Nathaniel Greenbaum, Jonathan W. Waks, Parastou Eslami, Timothy Carbonati, et al. 2023. *MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset.* doi:10.13026/4nqg-sb35

[12] Xiao Gu, Wei Tang, Jinpei Han, Veer Sangha, Fenglin Liu, Shreyank N. Gowda, Antonio H. Ribeiro, Patrick Schwab, Kim Branson, Lei Clifton, et al. 2026. Cardiac Health Assessment across Scenarios and Devices Using a Multimodal Foundation Model Pretrained on Data from 1.7 Million Individuals. *Nature Machine Intelligence* 8, 2 (2026), 220–233. doi:10.1038/s42256-026-01180-5

[13] Awni Y. Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H. Tison, Codie Bourn, Mintu P. Turakhia, and Andrew Y. Ng. 2019. Cardiologist-Level Arrhythmia Detection and Classification in Ambulatory Electrocardiograms Using a Deep Neural Network. *Nature Medicine* 25 (2019), 65–69. doi:10.1038/s41591-018-0268-3

[14] Neil Jethani, Mukund Sudarshan, Yindalon Aphinyanaphongs, and Rajesh Ranganath. 2023. Have We Learned to Explain?: How Interpretability Methods Can Learn to Encode Predictions in Their Interpretations. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 206)*. 1459–1491. https://proceedings.mlr.press/v206/jethani23a.html

[15] Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum McDougall, Kola Ayonrinde, Matthew Wearden, Arthur Conmy, Samuel Marks, and Neel Nanda. 2025. SAEBench: A Comprehensive Benchmark for Sparse Autoencoders in Language Model Interpretability. *arXiv preprint arXiv:2503.09532* (2025). https://arxiv.org/abs/2503.09532

[16] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viégas, and Rory Sayres. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*. 2668–2677. https://proceedings.mlr.press/v80/kim18d.html

[17] Sehun Kim. 2024. Learning General Representation of 12-Lead Electrocardiogram with a Joint-Embedding Predictive Architecture. *arXiv preprint arXiv:2410.08559* (2024). https://arxiv.org/abs/2410.08559

[18] Dani Kiyasseh, Tingting Zhu, and David A. Clifton. 2021. CLOCS: Contrastive Learning of Cardiac Signals Across Space, Time, and Patients. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*. 5606–5615. https://proceedings.mlr.press/v139/kiyasseh21a.html

[19] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept Bottleneck Models. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*. 5338–5348. https://proceedings.mlr.press/v119/koh20a.html

[20] William Lehn-Schiøler, Magnus Ruud Kjær, Rahul Thapa, Magnus Guldberg Pedersen, Anton Storgaard Mosquera, Nick Williams, Radu Gatej, Tue Lehn-Schiøler, Andreas Brink-Kjær, Sadasivan Puthusserypady, Sándor Beniczky, James Zou, and Lars Kai Hansen. 2026. Mechanistic Interpretability of EEG Foundation Models via Sparse Autoencoders. *arXiv preprint arXiv:2605.13930* (2026). https://arxiv.org/abs/2605.13930

[21] Fumin Li, Siting Li, Yuhan Qian, Bojun Chen, Jennifer A. Brody, Vidhushei Yogeswaran, Kerri L. Wiggins, Colleen M. Sitlani, Joshua C. Bis, Ali Shojaie, et al. 2026. CARDIAC-FM: A Multimodal Foundation Model for Cardiovascular Risk Prediction Using ECG and Cardiac MRI. *medRxiv* (2026). doi:10.64898/2026.03.16.26348526

[22] Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma Scope: Open Sparse Autoencoders Everywhere All at Once on Gemma 2. *arXiv preprint arXiv:2408.05147* (2024). https://arxiv.org/abs/2408.05147

[23] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. 2009. Online Dictionary Learning for Sparse Coding. In *Proceedings of the 26th International Conference on Machine Learning*. 689–696. doi:10.1145/1553374.1553463

[24] Aleksandar Makelov, Georg Lange, and Neel Nanda. 2025. Towards Principled Evaluations of Sparse Autoencoders for Interpretability and Control. In *International Conference on Learning Representations*. https://openreview.net/forum?id=1Njl73JKjB

[25] Kaden McKeen et al. 2025. ECG-FM: An Open Electrocardiogram Foundation Model. *JAMIA Open* 8, 5 (2025), ooaf122. doi:10.1093/jamiaopen/ooaf122

[26] Temesgen Mehari and Nils Strodthoff. 2022. Self-Supervised Representation Learning from 12-Lead ECG Data. *Computers in Biology and Medicine* 141 (2022), 105114. doi:10.1016/j.compbiomed.2021.105114

[27] W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. Definitions, Methods, and Applications in Interpretable Machine Learning. *Proceedings of the National Academy of Sciences* 116, 44 (2019), 22071–22080. doi:10.1073/pnas.1900654116

[28] Yeongyeon Na, Minje Park, Yunwon Tae, and Sunghoon Joo. 2024. Guiding Masked Representation Learning to Capture Spatio-Temporal Relationship of Electrocardiogram. In *International Conference on Learning Representations*. https://openreview.net/forum?id=WcOohbsF4H

[29] Bruno A. Olshausen and David J. Field. 1997. Sparse Coding with an Overcomplete Basis Set: A Strategy Employed by V1? *Vision Research* 37, 23 (1997), 3311–3325. doi:10.1016/S0042-6989(97)00169-7

[30] Antônio H. Ribeiro, Manoel Horta Ribeiro, Gabriela M. M. Paixão, Derick M. Oliveira, Paulo R. Gomes, Jéssica A. Canazart, Milton P. S. Ferreira, Carl R. Andersson, Peter W. Macfarlane, Wagner Meira, Jr., Thomas B. Schön, and Antonio Luiz P. Ribeiro. 2020. Automatic Diagnosis of the 12-Lead ECG Using a Deep Neural Network. *Nature Communications* 11 (2020), 1760. doi:10.1038/s41467-020-15432-4

[31] Cynthia Rudin. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence* 1 (2019), 206–215. doi:10.1038/s42256-019-0048-x

[32] Elana Simon and James Zou. 2025. InterPLM: Discovering Interpretable Features in Protein Language Models via Sparse Autoencoders. *Nature Methods* (2025). doi:10.1038/s41592-025-02836-7

[33] Nils Strodthoff and Claas Strodthoff. 2019. Detecting and Interpreting Myocardial Infarction Using Fully Convolutional Neural Networks. *Physiological Measurement* 40, 1 (2019), 015001. doi:10.1088/1361-6579/aaf34d

[34] Nils Strodthoff, Patrick Wagner, Tobias Schaeffter, and Wojciech Samek. 2021. Deep Learning for ECG Analysis: Benchmarks and Insights from PTB-XL. *IEEE Journal of Biomedical and Health Informatics* 25, 5 (2021), 1519–1528. doi:10.1109/JBHI.2020.3022989

[35] Adly Templeton et al. 2024. Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. *Transformer Circuits Thread* (2024). https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html

[36] Sana Tonekaboni, Shalmali Joshi, Melissa D. McCradden, and Anna Goldenberg. 2019. What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. In *Proceedings of the 4th Machine Learning for Healthcare Conference (Proceedings of Machine Learning Research, Vol. 106)*. 359–380. https://proceedings.mlr.press/v106/tonekaboni19a.html

[37] Akhil Vaid et al. 2023. A Foundational Vision Transformer Improves Diagnostic Performance for Electrocardiograms. *npj Digital Medicine* 6 (2023), 108. doi:10.

1038/s41746-023-00840-9
[38] Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I. Lunze, Wojciech Samek, and Tobias Schaeffter. 2020. PTB-XL, a Large Publicly Available Electrocardiography Dataset. *Scientific Data* 7 (2020), 154. doi:10.1038/s41597-020-0495-6

Table S1: Common-five-scale model profiles with paired patient-cluster 95% intervals. The semantic column reports the train-selected single-coordinate $|r|$ AUC defined in Section 3. Each AUC integrates the same $E \in \{1, 4, 8, 16, 32\}$ support and averages the same five depth targets and three SAE seeds.

Model	Recon AUC	Semantic AUC	Coverage AUC
CARDIAC-FM	0.936 [0.936, 0.937]	0.331 [0.322, 0.339]	**0.806** [0.772, 0.821]
CSFMe	0.964 [0.963, 0.965]	0.266 [0.255, 0.277]	0.659 [0.613, 0.682]
ECG-FM	0.940 [0.939, 0.940]	0.331 [0.322, 0.340]	0.788 [0.753, 0.806]
ECG-JEPA	0.972 [0.972, 0.972]	**0.351** [0.343, 0.359]	0.772 [0.761, 0.794]
HuBERT-ECG	**0.985** [0.985, 0.985]	0.272 [0.258, 0.285]	0.663 [0.620, 0.702]
ST-MEM	0.916 [0.915, 0.917]	0.314 [0.304, 0.325]	0.726 [0.690, 0.750]

A Common-Scale Model Profiles

Table S1 reports the exact model-level common-scale AUC estimates and paired patient-cluster confidence intervals underlying the visual comparison in Figure 4.

B Fixed-Scale Accessibility Calibration

Table S2 reports the complete held-out calibration that compares distributed ridge readouts with increasingly localized SAE, native-coordinate, and random-coordinate readouts at the common $E = 8$ scale.

Table S2: Held-out $E = 8$ accessibility calibration over five depths and 49 concepts. Ridge columns report mean $|r|$. One-dimensional entries report mean $|r|$/coverage at $|r| \geq .20$; SAE values average three seeds and random values average 20 dictionaries. All coordinates are selected on training data only.

Model	Dense ridge	Dense 1D	Full SAE	Top-16	Top-4	SAE 1D	Random 1D
CARDIAC-FM	0.764	0.377/0.927	**0.677**	0.514	0.423	0.320/**0.796**	0.232/0.591
CSFM	0.651	0.301/0.739	0.536	0.410	0.341	0.266/0.661	0.190/0.480
ECG-FM	**0.766**	0.385/**0.943**	0.675	0.513	0.417	0.321/0.778	0.243/0.634
ECG-JEPA	0.726	**0.421**/0.910	0.645	**0.535**	**0.430**	**0.342**/0.758	**0.290**/**0.697**
HuBERT-ECG	0.665	0.325/0.776	0.565	0.412	0.340	0.265/0.654	0.231/0.562
ST-MEM	0.725	0.379/0.886	0.615	0.474	0.391	0.311/0.714	0.245/0.572

Table S3: Frozen source-layer interfaces. The benchmark preserves each released ECG input path while matching the analyzed layer width. "Layers" counts available audited states, including the indexed boundary states used by the released extraction code.

Model	Pretraining family	Input used	Audited layer state	Layers	d
CSFM	multimodal sensing	12×2500	released encoder/classification stream	6	**768**
CARDIAC-FM	ECG+MRI risk	12×5000	transformer state before 768-to-512 pool projection	**12**	**768**
ECG-FM	wav2vec and contrastive	12×5000	token states	**12**	**768**
ECG-JEPA	joint-embedding prediction	8×2500	encoder states after released lead selection	**12**	**768**
HuBERT-ECG	masked clustering	flattened 12×2500	token states	**12**	**768**
ST-MEM	spatio-temporal masking	12×2250	`forward_encoding` states	**12**	**768**

C Frozen Source-Layer Interfaces

Table S3 defines the representation-extraction contract used throughout the benchmark. Each released checkpoint is frozen and evaluated through its native lead selection, temporal resolution, and tokenization path. The models consequently receive inputs ranging from eight to twelve leads and from 2,250 to 5,000 samples, while the audited states are taken before any model-specific downstream projection that would change the source width. Preserving these interfaces measures the representations available under realistic model use rather than introducing an untrained common tokenizer.

The shared audited width $d = 768$ permits exact capacity matching: the five expansion factors correspond to absolute dictionary widths $N \in \{768, 3072, 6144, 12288, 24576\}$, with the primary active budget fixed at $k = 96$. Relative depth maps the six available CSFM states and the twelve states exposed by each other encoder onto the same five index positions, providing a common coordinate over each encoder trajectory. Table S3 specifies the resulting representation interfaces, and Appendix F tests how heterogeneous waveform interfaces affect the final-layer model ordering.

D Additional Depth–Scale Atlases

Figures A1 and A2 expose the two validation fidelity diagnostics for all 150 model–depth–scale configurations underlying the 450 trained SAE cells. Each heatmap cell averages the three fixed SAE seeds, and each metric uses one shared color scale across all six models. These complete surfaces complement the clinical-accessibility atlas in Figure 2 and prevent a favorable depth or dictionary width from being selected in isolation.

Reconstruction remains high overall but does not improve monotonically with dictionary width under the fixed $k = 96$ budget. Averaged descriptively over all models, depths, and seeds, validation R^2 is 0.968, 0.953, 0.949, 0.947, and 0.948 for $E = 1, 4, 8, 16$, and 32, respectively. The model-averaged profiles range from 0.918 for ST-MEM to 0.986 for HuBERT-ECG, while the depth average falls from 0.978 at the initial state to approximately 0.937 at relative depths 0.5–0.75 before reaching 0.962 at the final state. Thus additional dictionary coordinates do not substitute for model- and depth-specific reconstruction structure when the active-code budget is held constant.

Dictionary utilization changes much more strongly with scale. The corresponding mean dead-feature fractions are 0.019, 0.042, 0.127, 0.338, and 0.604 from $E = 1$ through $E = 32$. Model-level averages range from 0.137 for ECG-FM to 0.337 for HuBERT-ECG, and the initial-state average (0.388) is substantially higher than the averages near relative depths 0.5 and 0.75 (0.156 and 0.158). HuBERT-ECG therefore combines the highest descriptive reconstruction average with the highest dead-feature average, directly illustrating that sparse fidelity and dictionary utilization are distinct properties. Reconstruction, dead-feature fraction, and the preregistered validation fidelity gate jointly characterize operating points across the complete dictionary-width curve.

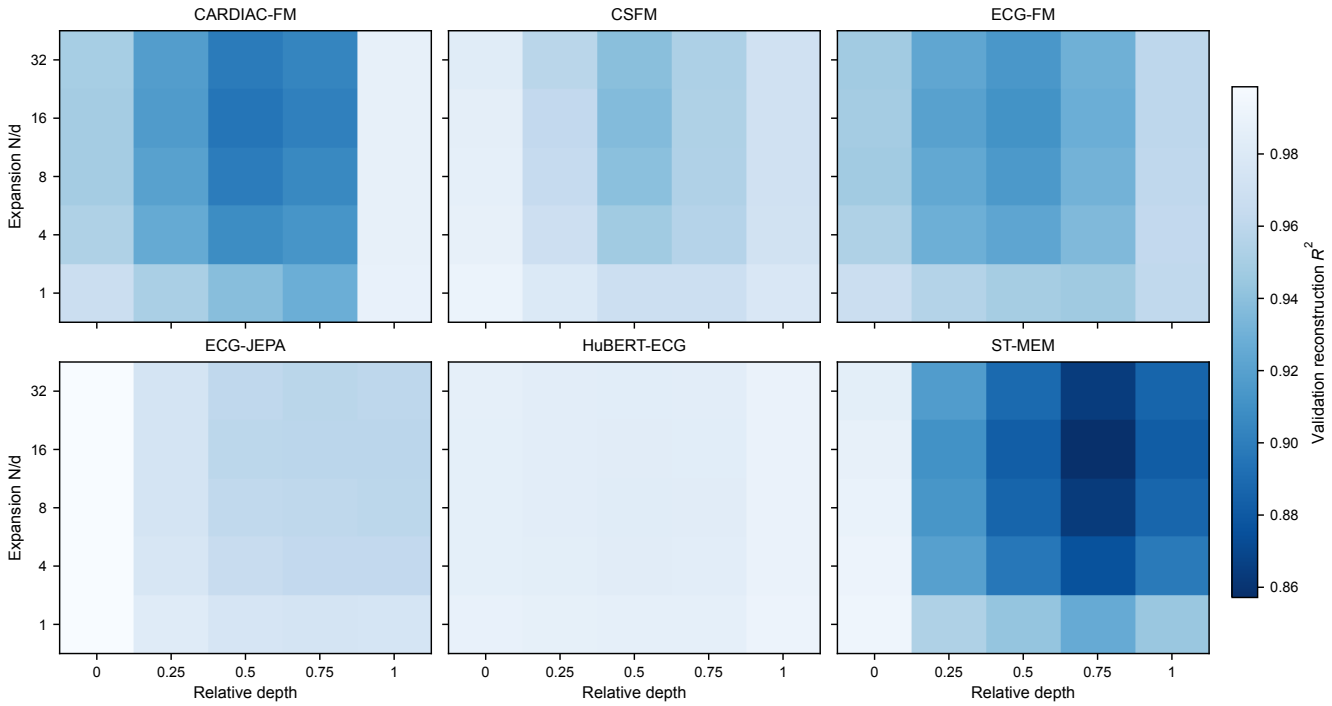

Figure A1: Validation reconstruction R^2 over the complete matched depth–scale atlas.

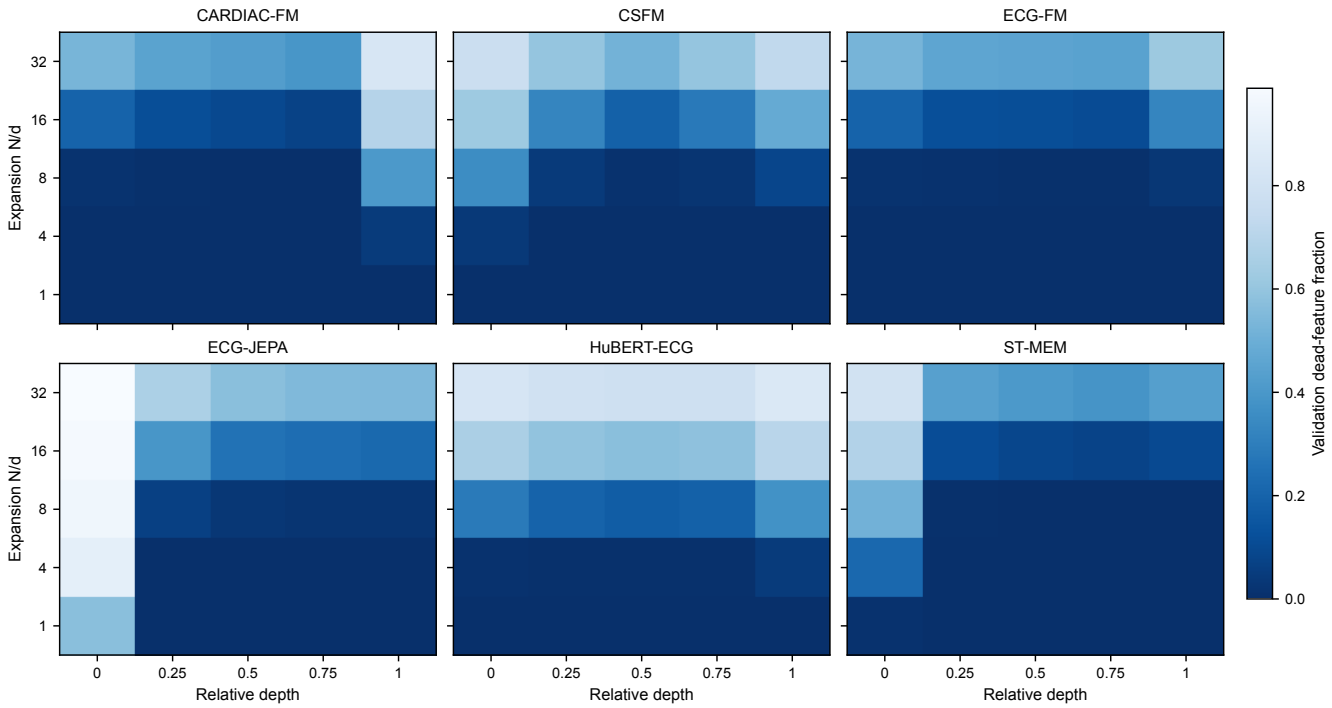

Figure A2: Validation dead-feature fraction over the complete matched depth–scale atlas.

Table S4: MIMIC-IV-ECG scale and depth sensitivity, aggregated across the six models. Expansion rows average over five depths; depth rows average over five expansions. All rows average the three SAE seeds.

Axis	Setting	Recon. R^2	Semantic alignment	Coverage	Dead fraction
Expansion	$E = 1$	0.973	**0.299**	**0.710**	0.032
	$E = 4$	0.971	0.281	0.624	0.134
	$E = 8$	0.971	0.284	0.632	0.277
	$E = 16$	0.971	0.284	0.637	0.461
	$E = 32$	0.970	0.283	0.627	**0.636**
Relative depth	0	**0.986**	0.295	0.640	0.542
	0.25	0.973	0.297	0.662	0.293
	0.50	0.964	0.298	0.689	0.211
	0.75	0.959	0.289	0.692	0.194
	1.00	0.973	0.252	0.546	0.301

E Credentialed MIMIC-IV-ECG Replication

We repeated the multiscale SAE audit on the credentialed MIMIC-IV-ECG cohort [11]. The fixed manifest contains 100,000 ECGs from 48,491 patients, partitioned by patient into 70,260 training, 9,768 validation, and 19,972 test records; 54,591 records have complete waveform-derived measurements. Source waveforms and identifiers remain in the credentialed environment. The replication uses the same six frozen encoders, relative depths $\{0, .25, .5, .75, 1\}$, expansion factors $E \in \{1, 4, 8, 16, 32\}$, $k = 96$, three SAE seeds, and 8,000 training steps as the primary benchmark, for 450 completed SAE cells. For this SAE-only analysis, semantic alignment and coverage use seven continuous waveform concepts: heart rate, QRS duration, PR interval, a QT-like interval, and global R-, ST-, and T-amplitude measurements.

The model-level held-out profiles are reported in Table 2; this appendix provides the cohort, uncertainty, and scale–depth sensitivity details.

Patient-cluster uncertainty uses 2,000 test-set bootstrap draws over 6,609 patients, retaining all ECGs from each sampled patient. The leading profiles are precise: HuBERT-ECG reconstruction is 0.9905 (95% CI 0.9904–0.9906), while ECG-JEPA semantic alignment is 0.3244 (0.3190–0.3299) and coverage is 0.7810 (0.7400–0.8057). The intervals quantify patient-sampling uncertainty conditional on the frozen encoders, trained SAEs, and training-selected coordinates.

The aggregate sensitivity analysis in Table S4 demonstrates the resolution gained from a multiscale, multidepth audit. Across expansion factors, reconstruction remains near 0.97, semantic alignment is highest at $E = 1$, and the dead-feature fraction traces dictionary utilization from 0.032 at $E = 1$ to 0.636 at $E = 32$. Across depth, semantic alignment is highest near relative depth 0.5 and coverage peaks near 0.75. These profiles show how MIMIC representation properties vary jointly with SAE capacity and encoder depth.

Matched decoder features vary across seeds (model-level stability AUC 0.293–0.378), whereas their selected subspaces are more reproducible (overlap AUC 0.702–0.778). This distinction identifies the selected subspace as the more stable unit for cross-seed interpretation and supports model-level conclusions about reproducible representation structure. The release audit passes all requirements with 450/450 SAE cells and no missing matched depth–scale block. Together, these results provide a complete credentialed-cohort replication of the benchmark profile.

F Input-Interface Harmonization Audit

The primary benchmark preserves each released model's input interface to reflect realistic use, but this choice could confound encoder differences with lead selection, sampling, or windowing. We therefore repeated final-layer dense-coordinate accessibility under four protocols on the same 21,799 PTB-XL records and patient split. *Native* retains each released interface. *Lead* supplies the common independent leads I, II, and V1–V6 and reconstructs the dependent limb leads for 12-lead models using $\mathrm{III} = \mathrm{II} - \mathrm{I}$, $\mathrm{aVR} = -(\mathrm{I} + \mathrm{II})/2$, $\mathrm{aVL} = \mathrm{I} - \mathrm{II}/2$, and $\mathrm{aVF} = \mathrm{II} - \mathrm{I}/2$. *Temporal* maps the complete 10-second waveform to a common 250-Hz, 2,500-point grid before adapting its length to each checkpoint. *Joint* applies both transformations. All harmonized waveforms are constructed in physical units and standardized per lead. Tokenizers, patch embeddings, and pretrained weights remain model-native; this is therefore a waveform-harmonized, tokenizer-native audit rather than a shared-tokenizer experiment.

For each of 49 waveform concepts, the best dense coordinate and its sign are selected only on a fixed 4,096-record subset of the training split. Mean absolute correlation and coverage at $|r| \geq 0.20$ are then evaluated without reselection on 3,325 patient-disjoint test records. We additionally compare each harmonized representation with its native counterpart using linear CKA. Uncertainty uses 2,000 patient-cluster bootstrap draws, with Benjamini–Hochberg correction applied across the 18 model-level comparisons and, separately, across all concept-level comparisons.

The model-level results are reported in Table 1.

Lead harmonization preserves mean $|r|$ within 0.0006 and achieves lead-to-native CKA of at least 0.9999. Temporal harmonization yields changes of $\Delta = -0.0058$ for CARDIAC-FM (95% CI $[-0.0095, -0.0022]$, $q = 0.016$), $\Delta = -0.0048$ for ECG-FM (95% CI $[-0.0083, -0.0013]$, $q = 0.016$), and $\Delta = -0.0131$ for ST-MEM (95% CI $[-0.0196, -0.0067]$, $q = 0.009$). The nearly identical joint and temporal effects identify time-grid and window handling as the source of the measurable changes. ST-MEM's joint CKA of 0.770 reflects the contrast between its native centered 9-second crop and the harmonized complete 10-second record mapped to 2,250 input points.

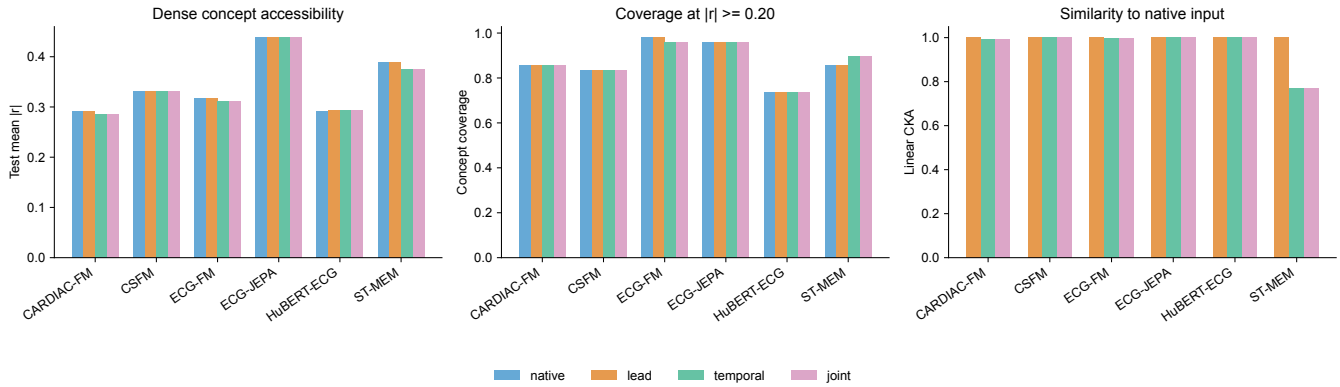

Figure A3: Final-layer accessibility under native, lead-harmonized, temporal-harmonized, and jointly harmonized inputs. Left: test mean absolute correlation after training-only coordinate selection. Center: concept coverage at $|r| \geq 0.20$. Right: linear CKA between each harmonized representation and its native counterpart.

Model-level coverage is stable after FDR correction (all $q = 1.0$), and the six-model ordering by mean accessibility is identical under every protocol (Spearman $\rho = 1.0$; Kendall $\tau = 1.0$). Coverage ordering is also stable (Spearman $\rho = 0.971$ for temporal and joint). ECG-JEPA and HuBERT-ECG mean differences range from 3.5×10^{-5} to 5.7×10^{-4}, further demonstrating the close agreement between native and harmonized accessibility profiles.

Overall, the stable ordering across all four input protocols indicates that the observed cross-model differences primarily reflect encoder-representation structure. Lead harmonization has negligible effect, and temporal harmonization preserves the ordering across models. This audit strengthens the attribution of the benchmark profiles to the learned representations.

G Validation-Matched SAE–Dense Intervention Selectivity

The accessibility analyses ask whether a concept can be read from one coordinate. We additionally test whether a sparse coordinate can change a frozen concept readout with less spillover than a native dense coordinate. This is a final-layer, matched-effect comparison over the same 49 waveform concepts. Both methods receive 768 candidate coordinates: the native FM dimensions for Dense and a fixed, target-independent 768-coordinate subset of the $E = 8$ dictionary for SAE. For each concept, the top five coordinates and the high-concept centroid are selected using training data only. Intervention doses are then calibrated on validation data to a common achievable target-readout change, capped at +0.25 standard deviations, with a minimum feasible change of 0.05 and interpolation coefficient at most one. Features, centroids, and doses are frozen before patient-disjoint test evaluation.

The SAE arm uses three training seeds and 20 fixed candidate subsets per seed. Primary inference requires a concept to pass the validation effect gate for all three seeds; failed concepts remain in the fixed denominator and are not replaced. The two primary endpoints are cross-family off-target root mean square change and the corresponding wrong-behavior index (WBI), defined as off-target RMS divided by the absolute target effect. Lower values indicate a more selective intervention. Confidence intervals and two-sided tests use 2,000 paired patient-cluster bootstrap draws, followed by Benjamini–Hochberg correction over the preregistered metric families.

Figure A4 visualizes the paired bootstrap contrasts. All ten 95% confidence intervals lie entirely below zero, showing consistent SAE reductions in both off-target RMS and WBI across the five models. Off-target RMS differences range from -0.014 for HuBERT-ECG to -0.087 for CSFM, while WBI differences range from -0.176 to -1.189 for the same models. CSFM shows the largest reduction on both endpoints, and the favorable direction remains consistent across eligibility sets ranging from 5 to 15 concepts. Across the resulting model–concept evaluations, SAE has a favorable and FDR-significant difference in all 10 model–metric tests ($q \leq .0011$). At a validation-matched frozen readout effect, the selected SAE coordinates consistently produce less cross-family readout spillover than native dense coordinates.

Table S5: Final-layer matched-effect comparison at $k = 5$ across the five models satisfying the all-seed validation gate. The eligibility count n requires all three SAE seeds to pass the gate. WBI and deltas are measured on the patient-disjoint test split; negative SAE–Dense deltas favor SAE.

Model	n	Dense WBI	SAE WBI	Δ off-target RMS	Δ WBI
CARDIAC-FM	5	1.614	0.971	−0.046	−0.643
CSFM	11	**2.134**	0.944	−0.087	−1.189
ECG-JEPA	**15**	1.436	0.691	−0.057	−0.746
HuBERT-ECG	10	1.315	**1.138**	**−0.014**	**−0.176**
ST-MEM	10	1.678	0.736	−0.063	−0.942

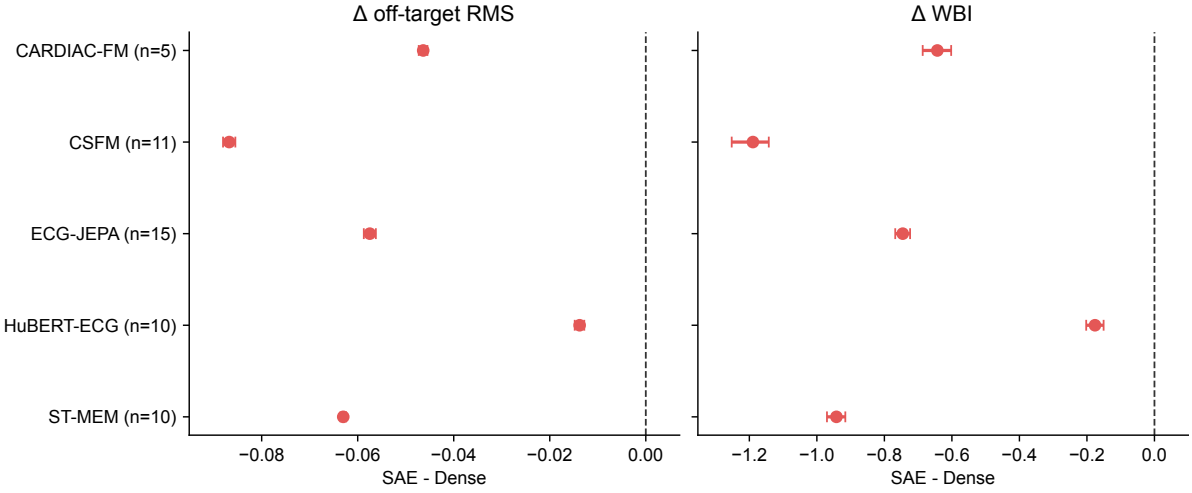

Figure A4: SAE–Dense differences in cross-family off-target RMS and WBI with paired patient-cluster 95% confidence intervals. Negative values favor SAE; model labels give the number of concepts satisfying the strict all-seed gate.