JOINT-EMBEDDING PREDICTIVE LEARNING CLOSES THE UNIMODAL–MULTIMODAL GAP IN ECG PRETRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

In time series representation learning, pairing signals with free-text descriptions typically outperforms signal-only pretraining; in ECG, where most recordings carry a written interpretive report, the gap is read as evidence that language supplies information the waveform lacks. Whether it survives a controlled comparison is unknown: published comparisons vary the pretraining corpus, the evaluation protocol and whether baselines are retrained at all. We fix all three, retraining six published methods from their official code on 1.21 million recordings and probing every model with the same frozen linear probe on PTB-XL and SPH. Under these controls, joint-embedding predictive learning (JEPA) largely closes the gap: it comes within 0.005 macro-AUROC of the strongest multimodal model on PTB-XL and scores 0.007 above it on SPH, the two being equivalent under two one-sided tests with a 0.02 margin. Reconstruction, discrete-target and contrastive baselines trail behind. Adding reports gives no gain with all labels, but adds 0.015 and 0.045 with 10% of them. Language buys label efficiency rather than a higher ceiling: signal-only latent prediction reaches the same performance without ever seeing a report, and whether text earns its cost elsewhere in time series needs the same controls applied.

1 INTRODUCTION

Across time series domains, representation learning that anchors a signal to paired free-text descriptions reports consistent gains over signal-only pretraining, and the electrocardiogram (ECG) is where that pairing is most readily available at scale. Clinical ECGs are almost always stored with a short interpretive report, such as *sinus rhythm, left axis deviation, anteroseptal infarct*, and because this text is free at scale it is an attractive training signal: methods built on it, notably MERL (Liu et al., 2024) and TolerantECG (Nguyen et al., 2025), lead every signal-only method in published PTB-XL tables (Wagner et al., 2020). This is widely read as evidence that language provides information the waveform does not, a strong claim given that the reports themselves are predominantly machine-generated interpretations with well-documented limitations (Schläpfer & Wellens, 2017; Smulyan, 2019). We ask whether a signal-only objective, evaluated on equal terms, can reach the downstream performance of report-anchored pretraining.

Existing results cannot answer this question, because method, data and evaluation vary together across papers (Figure 1a). Pretraining corpora differ by more than an order of magnitude, from 1.8×10^5 recordings (Kim, 2024) through 1.0×10^6 (Weimann & Conrad, 2025) to 9.1×10^6 (Coppola et al., 2024). Protocols differ in pooling, cropping, label subset and whether the encoder is frozen; the strongest published signal-only result, 0.945, comes from fine-tuning with attention pooling and crop ensembling on a model whose pretraining pool contains PTB-XL itself (Weimann & Conrad, 2025). Baselines are typically evaluated from released checkpoints rather than retrained, which is not neutral: ST-MEM scores 0.896 from its released weights and 0.871 when its own code is retrained on our corpus under the same probe, a difference larger than most reported gaps between methods. The multimodal advantage has therefore never been measured directly, only inferred from tables in which these factors covary.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

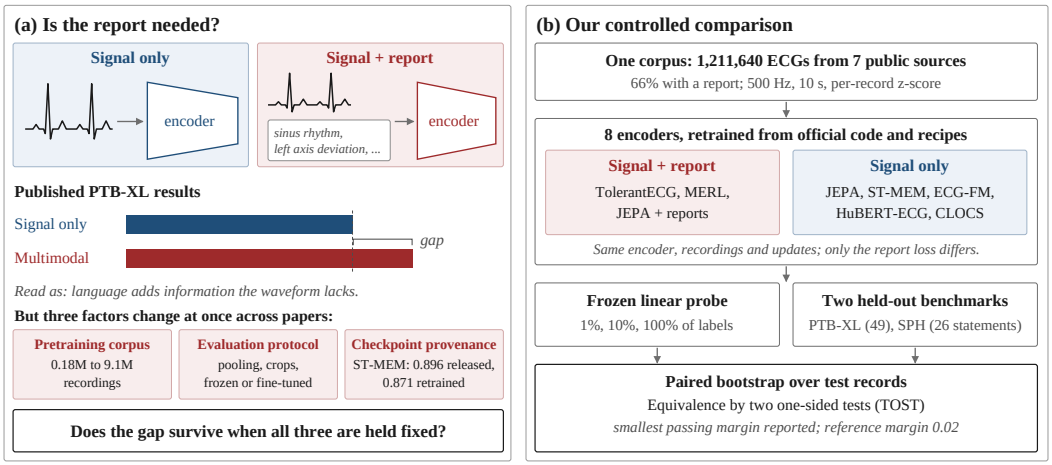

Figure 1: **The question and the protocol. (a)** Multimodal ECG pretraining anchors the signal to its paired report and leads signal-only pretraining in published PTB-XL tables, a gap read as evidence that language adds information the waveform lacks; but corpus, evaluation protocol and checkpoint provenance vary together across papers, so the gap has never been measured directly. **(b)** We hold all three fixed: one corpus, eight encoders retrained from their official code (two of them the same encoder with and without the report loss), one frozen linear probe on two held-out benchmarks, and an equivalence test that reports the smallest margin at which each verdict passes.

Signal-only ECG pretraining has developed along four lines: augmentation-invariant contrastive learning (Chen et al., 2020; Kiyasseh et al., 2021), waveform reconstruction (He et al., 2022; Na et al., 2024), discrete-target prediction (Hsu et al., 2021; Coppola et al., 2024) and continuous latent prediction (Baevski et al., 2022; Assran et al., 2023). The published multimodal advantage was established almost entirely against the first three, and neither MERL nor TolerantECG includes a latent-prediction model among its baselines.

Those three families share a structural difficulty on this signal. Instance-discriminative contrastive learning treats every other recording in a batch as a negative, so the objective is dominated by the largest source of variation between recordings, which in the ECG is patient morphology rather than diagnosis (Chen et al., 2020; Kiyasseh et al., 2021). The small, low-energy variation that separates a normal recording from a non-specific ST change is the variation the objective has least incentive to preserve, and augmentation invariance removes part of it outright. Waveform reconstruction has the opposite problem: its loss is weighted by signal energy, so capacity is spent on the QRS complex, baseline wander and residual noise, and the target is by construction the raw waveform rather than any abstraction of it (He et al., 2022; Assran et al., 2023). Latent prediction instead predicts a teacher's representation of the masked region (Baevski et al., 2022). That target is already invariant to noise the teacher has learned to ignore and already abstracted from amplitude, so the objective can spend capacity on the timing and waveform-shape structure that must be modeled to predict what is hidden. This is a motivation rather than a result; the retrained comparison of §3 is the test.

Our controlled comparison (Figure 1b) evaluates latent prediction under one protocol, adopting the I-JEPA formulation (Assran et al., 2023) as Weimann & Conrad (2025) apply it to ECG, in a 14.5M-parameter 1-D ViT; we claim neither the formulation nor our single deviation from it as a contribution (§3.5). We pretrain on 1,211,640 recordings from seven public sources, principally MIMIC-IV-ECG (Gow et al., 2023) and CODE-15 (Ribeiro et al., 2020), of which 66% carry a machine-generated report, and retrain six published methods (TolerantECG, MERL, ST-MEM, ECG-FM, HuBERT-ECG and CLOCS) on the same corpus from their official implementations. All eight models share one preprocessing pipeline, one frozen linear probe, one label rule and one label-free checkpoint rule (Figure 1b), and are evaluated on PTB-XL (Wagner et al., 2020) and the Shandong Provincial Hospital database (Liu et al., 2022) at 1%, 10% and 100% of the downstream labels (§2). Because the relevant claim against the strongest multimodal baseline is equivalence rather than superiority, we test it with two one-sided tests (Schuirmann, 1987; Lakens, 2017), report

the smallest margin at which each benchmark passes alongside a reference margin of ± 0.02, and never treat a non-significant difference as evidence of equivalence.

Our results separate into two regimes. With all downstream labels the JEPA encoder lands 0.005 below TolerantECG on PTB-XL and 0.007 above it on SPH, neither difference significant; equivalence holds at any margin of 0.011 or more on PTB-XL and 0.017 or more on SPH, the latter failing smaller margins only on JEPA's side. It is the only retrained signal-only method to reach the multimodal baselines, leading every reconstruction, discrete-target and contrastive method by at least 0.032, and adding reports to it changes macro-AUROC by $+0.001$ and -0.003. With 10% of the labels it trails TolerantECG by 0.017 and 0.059 and reports recover most of that gap, $+0.015$ and $+0.045$, significant under bootstrap, permutation and Wilcoxon tests; with 1% the multimodal lead widens to 0.056 and 0.072, but the report effect on the JEPA encoder is no longer reliably detectable. What remains is specific to statement type: the encoder leads on rhythm and conduction and trails on voltage and ST-T statements, exactly where reports help most at reduced budgets.

Our contribution is the measurement rather than the encoder. First, we give the first controlled measurement of the unimodal–multimodal gap in ECG pretraining: eight methods, six of them published, retrained on a single corpus and evaluated under one frozen probe on two benchmarks with different label ontologies, with equivalence tested directly rather than inferred from a non-significant difference. Second, that measurement corrects the reading of published multimodal gains. The gap largely closes with all downstream labels, where a 14.5M-parameter latent-prediction encoder, used here as the instrument, matches the strongest multimodal baseline with 54% of its parameters and gains nothing from reports; the residual benefit of reports is a label-budget effect, measurable at 10% of labels and absent at 100%; and checkpoint provenance alone is worth up to 0.026 macro-AUROC under an otherwise identical protocol. Third, we release the retrained baselines, the SPH evaluation split, per-label probe outputs and the equivalence procedure. The protocol is not specific to the ECG: wherever a time series is paired with text, these three controls are what separate a real multimodal gain from a difference in corpus, protocol or checkpoint provenance.

2 Our Controlled Comparison

2.1 One source pool, one preprocessing

Every model is pretrained from one predefined source pool of $1{,}211{,}640$ recordings from seven public sources, principally MIMIC-IV-ECG ($800{,}035$) and CODE-15 ($345{,}779$), with Ningbo, Georgia, Chapman–Shaoxing and the two CPSC-2018 collections supplying the rest (Appendix Table 4). Each recording is resampled to 500 Hz, band-passed at 0.5–150 Hz with $50/60$ Hz notch filters, per-record z-scored per lead and cropped to a centered 10 s window (12×5000), once and cached, so every recording a model consumes is byte-identical across models; each published method retains its official loader (§2.2). Reports exist only for the MIMIC-IV-ECG subset, so 66.0% of the pool is report-bearing. Neither benchmark enters the pool (verified at the signal level for SPH, Appendix A.3).

2.2 Eight encoders: six published methods, retrained

The eight encoders of Figure 1b are six published methods and the two arms of the JEPA encoder (§2.3, §2.4). Each published method is retrained from scratch on the pool with its official implementation and released configuration, including architecture, batch size, optimizer and training length (Appendix Table 5), so that the retrained model is the method its authors published. What is held fixed is therefore the corpus, the preprocessing, the splits, the probe and the checkpoint rule; training length is part of each method (Appendix A.14 matches budgets). Each recipe was run once (§2.6).

The two multimodal baselines consume only report-bearing rows, as their pipelines require: TolerantECG trains on MIMIC-IV-ECG ($800{,}035$ recordings) and MERL on the MIMIC rows with a usable report, subsampled to $392{,}000$ by its own loader. The report-anchored JEPA arm sees the same $800{,}035$ reports and the remaining rows through its latent loss. The published latent-prediction ECG encoders (Weimann & Conrad, 2025; Kim, 2024) were pretrained on pools that include PTB-XL, so the encoder of §2.3, which follows their formulation, stands in for the family.

Faithfulness is checked against MERL, whose released checkpoint is pretrained on MIMIC-IV-ECG: our retrained copy reproduces it to within 0.006 macro-AUROC on PTB-XL. Where a retrained baseline falls further short, as ST-MEM does at 0.026, we report the gap rather than adjust for it (Appendix Table 16); TolerantECG and CLOCS release no weights.

2.3 EIGHT ENCODERS: THE SIGNAL-ONLY ARM

The encoder is a 1-D vision transformer over non-overlapping 50-sample patches (100 tokens per recording), width 384, depth 8, 8 heads, 14.47M parameters, trained with the I-JEPA formulation (Assran et al., 2023) as Weimann & Conrad (2025) apply it to ECG. The student encodes *only* the visible patches, and a two-block transformer predictor estimates the representations at block-masked positions from the visible context and learned positional queries. Targets come from an exponential-moving-average teacher (momentum $0.999 \rightarrow 0.9999$). Masking uses blocks of 10 patches at a nominal ratio of 0.46; block starts are drawn with replacement, so blocks overlap and the realized masked fraction is 0.34. Writing z_i for the teacher target at masked position i, $\hat{z}_i$ for the predictor's estimate and σ_d for the per-dimension standard deviation of the student's features across the batch, the objective is

$$\mathcal{L} \;=\; \frac{1}{|\mathcal{M}|}\sum_{i\in\mathcal{M}} \mathrm{smooth}_{L_1}\big(\hat{z}_i,\, \mathrm{sg}[z_i]\big) \;+\; \lambda\,\frac{1}{D}\sum_{d=1}^{D}\max\big(0,\, 1-\sigma_d\big), \qquad \lambda = 1. \tag{1}$$

where $\mathrm{sg}[\cdot]$ is the stop-gradient onto the teacher and the second term is a variance hinge against collapse. Training runs $100{,}000$ steps at batch 256 with AdamW, learning rate 8×10^{-4} and weight decay 0.05. Our single departure from the published formulation is the prediction target, the instance-normalized mean of the teacher's eight block outputs rather than its final block alone (Baevski et al., 2022); both are compared in §3.5.

2.4 EIGHT ENCODERS: THE REPORT-ANCHORED ARM

The second arm adds report anchoring to the encoder of §2.3 and changes nothing else. Report embeddings $t \in \mathbb{R}^{768}$ are computed offline with a frozen MedCPT text encoder (Jin et al., 2023). The pooled ECG representation z (a token mean from a clean full-context pass, as at evaluation) and t are each projected by a two-layer MLP to 256 dimensions and ℓ_2-normalized, giving $e = g_\theta(z)$ and $u = h_\psi(t)$, and aligned by a symmetric InfoNCE over the report-bearing rows of the batch,

$$\mathcal{L}_{\text{txt}} \;=\; \tfrac{1}{2}\Big[\mathrm{CE}\big(s\,eu^\top,\, \mathbf{I}\big) + \mathrm{CE}\big(s\,ue^\top,\, \mathbf{I}\big)\Big]. \tag{2}$$

with a learnable temperature s (initialized at $1/0.07$, clamped at 100), added to Equation 1 at weight 1.0. Recordings without a report keep their latent-prediction loss, so both arms see the same recordings in the same order for the same number of updates, and the projection heads are discarded after pretraining, so the evaluated checkpoints are architecturally identical. A *shuffled*-pairing control exists on an earlier backbone (Appendix A.6).

2.5 ONE FROZEN PROBE, TWO HELD-OUT BENCHMARKS

PTB-XL (Wagner et al., 2020) uses the official stratified folds ($17{,}418/2{,}183/2{,}198$) and the SCP-ECG statement set. SPH (Liu et al., 2022) comprises $25{,}770$ recordings from $24{,}666$ patients with AHA/ACC/HRS statements; we split it *by patient* into $15{,}442/5{,}169/5{,}159$, stratified on each patient's first primary statement (Appendix A.3), and label on the 44 primary statements with modifiers stripped. A statement is scored if it has at least ten positive test examples, which retains 49 of 71 on PTB-XL and 26 of 44 on SPH (Appendix A.1); macro-AUROC is the unweighted mean over scored statements, with the scored set held fixed across label budgets.

Encoders are frozen, embeddings are mean-pooled over tokens from a single centered crop, and one linear layer is fit with class-weighted binary cross-entropy under one recipe shared by every model (Appendix A.5). The probe's training split is subsampled to 1%, 10% and 100% with one fixed seed, so every encoder sees the identical labeled subset at each budget, and each model is evaluated at the final checkpoint of its own training configuration, a label-free rule applied identically. The validation split is not subsampled (at 1% the probe is fit on 174 PTB-XL records while early stopping observes

Table 1: **Full-label comparison.** Frozen-probe macro-AUROC with 100% of labels, paired differences against TolerantECG, encoder parameters and FLOPs per 10 s record. Best per column bold, second underlined. †: equivalent at the ± 0.02 reference margin (§2.6; smallest passing margins in the text); all other differences exclude zero.

Model	Inputs	Params (M)	GFLOPs	PTB-XL		SPH	
				AUROC	Δ vs. Tol.	AUROC	Δ vs. Tol.
TolerantECG	signal + report	26.8	13.1	**0.926**	–	0.960	–
MERL	signal + report	3.9	3.4	0.896	-0.031	0.927	-0.033
JEPA	signal	14.5	3.0	0.921	$-0.005^{\dagger}$	**0.967**	$+0.007^{\dagger}$
JEPA + reports	signal + report	14.5	3.0	<u>0.923</u>	$-0.004^{\dagger}$	<u>0.963</u>	$+0.003^{\dagger}$
ST-MEM	signal	85.2	70.7	0.871	-0.056	0.935	-0.025
ECG-FM	signal	90.9	29.2	0.876	-0.050	0.918	-0.042
HuBERT-ECG	signal	30.5	7.3	0.809	-0.117	0.884	-0.075
CLOCS	signal	0.1	0.01	0.608	-0.318	0.671	-0.289

2,183), so every reduced-budget comparison is repeated with the validation split subsampled at the same rate as a control. This deliberately conservative protocol places our absolute numbers below published fine-tuning results; what is compared is models under one protocol.

2.6 Paired bootstrap and equivalence test

All intervals come from a paired bootstrap over test records ($B = 2000$), with the same resampled indices applied to both models. Because both test folds contain patients with more than one recording (23.7% of PTB-XL test records, 8.6% of SPH), every headline comparison is repeated with a bootstrap over patients, which widens intervals by at most 0.0011 and changes no verdict. Each pretraining recipe was run once, so all intervals quantify evaluation uncertainty conditional on the trained checkpoints (§3.5).

Against the strongest multimodal baseline the relevant claim is equivalence, not superiority, so we test it directly with two one-sided tests (Schuirmann, 1987; Lakens, 2017): a difference is equivalent at margin δ when the 90% interval of the paired difference lies within $[-\delta, +\delta]$. Because any verdict depends on δ, the primary quantity reported is the smallest margin at which equivalence holds (the larger absolute bound of that interval), so a reader may apply their own; the verdict at a fixed reference $\delta = 0.02$, used unchanged on both benchmarks, is given for orientation only. Refitting the probe from five initializations on identical features moves the paired difference against TolerantECG with SD 0.0008 on PTB-XL and 0.0052 on SPH, so the reference is four to twenty-five times the evaluation's own noise and below the spacing between consecutive published methods. A non-significant difference is never reported as equivalence. The report effect is assessed with a record-level bootstrap, a within-record permutation test and a Wilcoxon signed-rank test over per-statement AUROCs, paired over different axes so that their agreement is informative (Appendix A.5).

3 Results: Does the Gap Survive?

Figure 1a asked whether the gap survives when corpus, protocol and provenance are held fixed. We report the answer at 100% labels (§3.1), what reports buy at reduced budgets (§3.2), why latent prediction reaches the multimodal band (§3.3), where it does and does not (§3.4) and the checks on alternative explanations (§3.5), every number from the protocol of §2 on frozen features. Each arm is trained once, so the intervals bound evaluation uncertainty for those checkpoints, not variation across training runs.

3.1 Latent prediction reaches the multimodal band with all labels

Table 1 gives the comparison with all downstream labels. On PTB-XL the paired difference between JEPA and TolerantECG is -0.005 with a 90% interval of $[-0.011, +0.001]$, so equivalence holds at any margin of 0.011 or more, and superiority does not ($p = 0.16$). On SPH JEPA leads by $+0.007$ with an interval of $[-0.002, +0.017]$: equivalence holds at any margin of 0.017 or more,

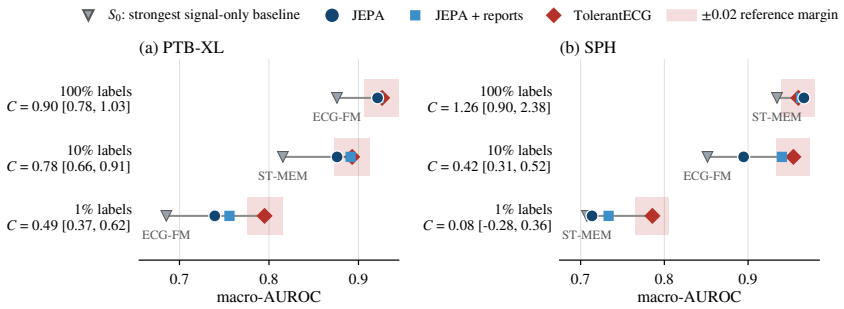

Figure 2: **Latent prediction closes most of the unimodal–multimodal gap at 100% labels, and all of it on SPH.** For each benchmark and label budget: the strongest retrained signal-only baseline S_0 (named beneath each marker), the JEPA encoder, its report-anchored arm and TolerantECG, with the ± 0.02 reference band shaded and the closure C with 95% paired-bootstrap intervals. At 100% JEPA is the only signal-only encoder inside the band.

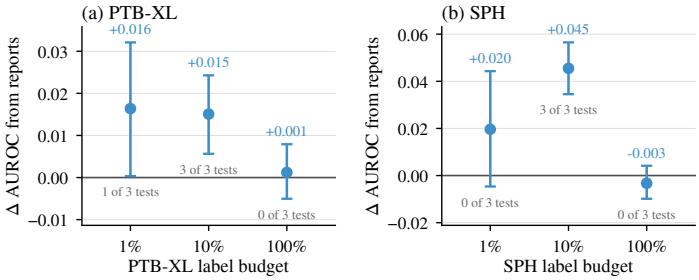

Figure 3: **Reports buy label efficiency, not a higher ceiling.** Paired difference in frozen-probe macro-AUROC between the JEPA encoder with and without the report objective on **(a)** PTB-XL and **(b)** SPH at 1%, 10% and 100% of downstream labels (95% paired bootstrap; beneath each point, how many of the three tests of §2.6 are significant). The full label-budget curves of every contender are in Appendix Figure 8.

and below that the interval fails only on JEPA's side, so non-inferiority holds down to 0.002. At the reference margin of 0.02 both benchmarks are equivalent. The report-anchored arm, with intervals $[-0.009, +0.002]$ and $[-0.003, +0.012]$, is equivalent to TolerantECG at any margin of 0.010 and 0.012 or more. Against every other signal-only method the result is one-sided: JEPA leads all four on both benchmarks, by 0.032 (ST-MEM on SPH) to 0.313 (CLOCS on PTB-XL), every interval excluding zero, and MERL trails it by 0.025 and 0.040. Even from released weights no signal-only checkpoint reaches the band: the strongest, ST-MEM at 0.896 on PTB-XL, remains 0.025 below JEPA (Appendix Table 16), which has 54% of TolerantECG's parameters and 23% of its FLOPs.

How much of the gap closes? With M TolerantECG, S_0 the strongest retrained signal-only baseline in a cell, $G_0 = M - S_0$, $G_J = M - J$ and $C = 1 - G_J/G_0$ the fraction closed, all within the same bootstrap replicates (Figure 2): at 100% labels G_0 is 0.050 on PTB-XL and 0.025 on SPH, G_J is $+0.005$ $[-0.002, +0.012]$ and -0.007 $[-0.018, +0.003]$, and C is 0.90 $[0.78, 1.03]$ and 1.26 $[0.90, 2.38]$. Neither interval excludes complete closure; we say *most* because on PTB-XL the point estimate leaves 0.005 of 0.050, inside the margin and comparable to probe re-initialization noise, but not zero. The two JEPA arms agree: adding reports with recordings and optimizer updates held fixed (§2.4) changes macro-AUROC by $+0.001$ and -0.003; with all labels there is nothing left for the reports to add.

3.2 Reports improve label efficiency; correct ECG–report pairing matters

At reduced budgets the ordering changes (Figure 2; every contender's curve in Appendix Figure 8). With 10% of the labels JEPA trails TolerantECG by 0.017 and 0.059, with 1% by 0.056 and 0.072, and C falls to 0.78 and 0.42: the multimodal advantage is real in this regime. The two JEPA arms

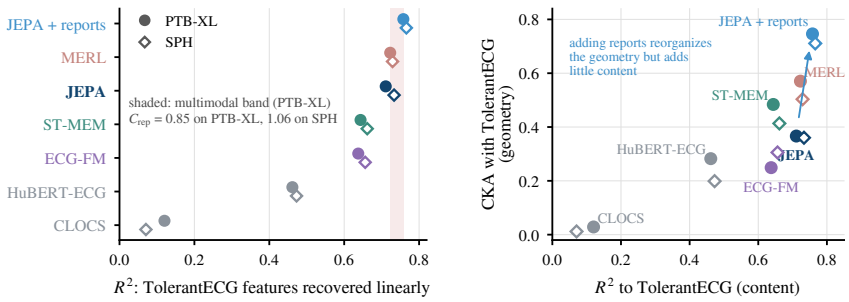

Figure 4: **Much of the multimodal representation is linearly recoverable from the signal-only one, despite a different geometry. (a)**, left: R^2 of a ridge map from each encoder's frozen features to TolerantECG's (circles PTB-XL, diamonds SPH; bootstrap intervals are narrower than the markers), multimodal band shaded. **(b)**, right: the same R^2 against linear CKA with TolerantECG; reports raise CKA from 0.37 to 0.75 and R^2 by 0.05.

change its interpretation (Figure 3). Adding reports to JEPA is worth $+0.015$ $[+0.006, +0.024]$ and $+0.045$ $[+0.035, +0.057]$ at 10%, significant under all three tests on both benchmarks; at 1% the point estimates are $+0.016$ and $+0.020$, but only 1 of 3 and 0 of 3 tests pass and the effect inverts under the validation-subsampling control of §2.5 while the 10% effect survives (Appendix Table 11): only the 10% effect is attributed to text, and the 1% column is treated as exploratory. A seven-budget sweep peaks at 2% and shows no clean threshold (Appendix Figure 9).

A second loss term could help for reasons unrelated to what the reports say; on the earlier `v16` backbone, real reports gain $+0.084$ and $+0.100$ at 10% while reports shuffled across patients *lose* -0.023 and -0.025 (Appendix A.6), so on that independently trained experiment the benefit requires correct ECG–report pairing, and breaking it is worse than no text at all; the control was not rerun on the final encoder.

3.3 What changes when reports are added: linearly recoverable content versus geometry

Three label-free analyses explain the result. First, the frozen signal predicts the report: a ridge regression from frozen features to the MedCPT embedding of the paired report (Appendix A.12) reaches $R^2 = 0.289$ from the signal-only encoder against 0.391 from its report-anchored arm and -0.002 from a shuffled-target control: never having seen a report, the encoder recovers 74% of the report variance that anchoring does.

Second, much of the multimodal representation is linearly recoverable from the signal-only one. Figure 4a asks how much of TolerantECG's embedding a linear map recovers from each encoder's, fit on the training split and scored on the test split with no label involved: signal-only baselines 0.64–0.66, MERL 0.72, JEPA 0.71 on PTB-XL and 0.73 on SPH, a closure of 0.85 and 1.06 that matches the downstream C. Task information agrees: concatenating any other encoder's features to JEPA's gains a probe nothing, whereas TolerantECG still gains from ST-MEM and ECG-FM, an asymmetry that a dimensionality artifact would not produce (Figure 5). These analyses characterize linear recoverability and linear-probe complementarity; they do not establish information-theoretic containment.

Third, what reports change most is the geometry. Linear CKA (Kornblith et al., 2019) between JEPA and TolerantECG is 0.37, below ST-MEM's 0.48 (Figure 4b); the fraction of a record's ten nearest neighbours sharing its superdiagnostic class is 0.53 for JEPA against 0.67 for TolerantECG (chance 0.37; Appendix Table 15). One reading consistent with these numbers is that contrastive alignment clusters same-diagnosis recordings while latent prediction leaves the same content decodable but not clustered, which a probe would need when labels are scarce and not when they are plentiful. This matches the hypothesis of §1, but the experiments test its predictions, not its mechanism: we do not isolate objective choice from architecture or training recipe.

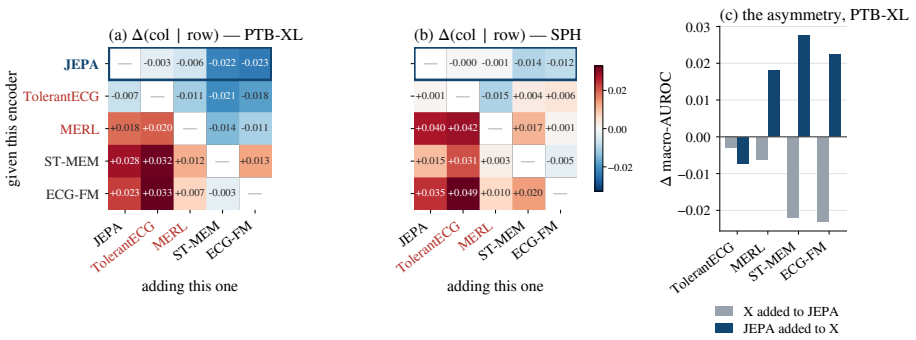

Figure 5: **TolerantECG contributes little additional linearly decodable task information beyond JEPA.** For frozen encoders A, B we write $\Delta(B \mid A) = R([\phi_A, \phi_B]) - R(\phi_A)$ for the macro-AUROC a linear probe gains when B's features are concatenated to A's: the task information B carries that A lacks. **(a,b)** Δ for every ordered pair, on both benchmarks; multimodal encoders are labeled in red. *Only the JEPA row is non-positive throughout on both benchmarks*: no other encoder, multimodal or otherwise, supplies linearly decodable diagnostic information that the signal-only representation is missing. TolerantECG, by contrast, gains from ST-MEM and ECG-FM on SPH. **(c)** The asymmetry that excludes a dimensionality artifact: concatenation is identical in both directions, yet adding JEPA to ST-MEM is worth $+0.028$ while adding ST-MEM to JEPA is worth -0.022. The JEPA–TolerantECG pair is null in both directions, which is the equivalence result restated at the level of linearly decodable information rather than accuracy; this is a statement about linear-probe complementarity, not information-theoretic containment.

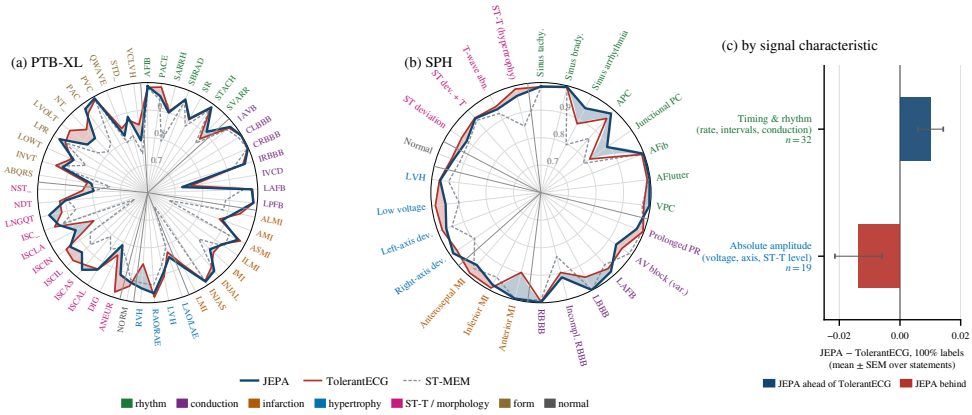

Figure 6: **Where the signal-only encoder wins and where it does not. (a,b)** Per-statement AUROC at 100% labels for every scored statement on PTB-XL (49) and SPH (26) for JEPA, TolerantECG and ST-MEM, radial axis 0.6–1.0; the band between the JEPA and TolerantECG curves is blue where JEPA is ahead and red where it is behind (intervals in Appendix Table 13). **(c)** The same statements, both benchmarks pooled, grouped by the property of the signal they depend on: timing (rate, rhythm, intervals, conduction; $n = 32$) and absolute amplitude (voltage, axis and diffuse ST-T level; 19); infarction and Q/T-morphology statements, on which the two are level, and the normal-ECG statements are omitted. Bars are the paired difference JEPA−TolerantECG at 100% labels, mean ± SEM over statements: JEPA leads on timing and trails on absolute amplitude.

3.4 Where the encoder wins, and where it does not

Figure 6 shows what the macro average averages over. With all labels JEPA matches or exceeds TolerantECG on 24 of 49 PTB-XL statements and 14 of 26 SPH statements, and exceeds ST-MEM on 44 and 22. Its largest gains, with intervals excluding zero, are rhythm and ectopy (sinus arrhythmia $+0.063$, premature complexes, atrial flutter) and right-axis deviation; its largest deficits, 10 on PTB-XL and 8 on SPH with intervals excluding zero before any multiplicity correction, are the same

kind of statement on both ontologies: *low voltage* is the largest on each (-0.080 and -0.036), then the voltage criterion for hypertrophy, the diffuse ST-T statements and infarction localizations (two on PTB-XL, one on SPH). Grouped by the signal property a statement depends on (Figure 6c), JEPA leads on timing statements by $+0.010$ and trails on absolute-amplitude statements by 0.014. The deficit tracks statement type, not prevalence (Spearman $\rho = -0.17$ and -0.06, $p > 0.2$).

The pattern is consistent with the hypothesis of §3.3. Rhythm, ectopy and conduction are timing and waveform-shape structure a masked-latent objective must model to predict what it cannot see; global QRS amplitude and diffuse, low-energy repolarization shifts contribute little to that loss, and a report names them outright, so contrastive alignment turns each name into a readout. Three observations are consistent with a readout deficit rather than missing information: the 10% report effect concentrates on exactly these statements; the deficit shrinks from -0.056 to -0.033 and from -0.121 to -0.022 between 10% and 100% labels; and on them JEPA still leads ST-MEM by $+0.026$ and $+0.024$, so the shortfall is relative to text anchoring, not to signal-only pretraining (Appendix Table 12). The effect is two-sided: on rhythm statements the report objective *costs* accuracy at 100% labels, -0.028 on PTB-XL and -0.021 on SPH (both intervals exclude zero), and -0.121 on SPH sinus arrhythmia at 10% (Appendix Table 13), as the readout account would predict.

3.5 Prediction target and robustness

Our one departure from the I-JEPA recipe, averaging the teacher's normalized upper eight blocks instead of its final block (§2.3), is worth $+0.005$ on PTB-XL and $+0.001$ on SPH at 100% labels, neither significant, so it is not what closes the gap; on a mask-token encoder the same change is worth $+0.028$ to $+0.058$ (Appendix Figure 7), so both modifications address the same deficiency. Released checkpoints score 0.006 to 0.026 above our retrained ones; released and retrained checkpoints are not interchangeable, and corpus, schedule and stochasticity cannot be separated after the fact (Appendix Tables 16–18). Matching passes controls exposure, not compute, since architectures, batch sizes and per-step FLOPs differ. Appendix A.14 compares each baseline with the JEPA checkpoint nearest its own budget; the ordering against every signal-only method is unchanged at matched passes, and the JEPA encoder's own trajectory plateaus by 13 passes (Appendix Figure 10), so the schedule is not what separates it from the baselines. The final five checkpoints vary by SD ≤ 0.0015, an MLP probe puts the smallest passing margins at 0.014 on PTB-XL and 0.008 on SPH. The report effect, though not the equivalence result, replicates on an independently trained encoder, the earlier `v16` backbone: $+0.084$ and $+0.100$ at 10% labels (Appendix Table 10).

Assumptions. Frozen linear probing is the evaluation of interest, so fine-tuning, zero-shot classification and report retrieval, where paired text defines the task, are not studied and the equivalence claim does not extend to them. Each pretraining recipe was run once. The intervals therefore quantify evaluation uncertainty conditional on the trained checkpoints and do not include optimization-seed variability; the equivalence claim is restricted to the trained checkpoints under the stated protocol, and the two benchmarks, differing in population, ontology and statement set, are a robustness check on that claim rather than a substitute for independent pretraining runs. The final checkpoint is used, a label-free rule that could favor objectives still improving there. The 1% column is indicative only, because early stopping observes more labeled records than the probe is fit on. Both benchmarks are 10 s 12-lead resting recordings and reports cover 66% of the pool; we claim nothing beyond that setting.

4 Conclusion

Report-anchored ECG pretraining is usually read as evidence that language adds information the waveform lacks. Under a controlled comparison we find the opposite: most of what the report contributes is already in the signal. A latent-prediction encoder that never sees a report is statistically equivalent to the strongest multimodal model on two benchmarks with all labels, and anchoring it to reports changes nothing. What reports add is arrangement rather than content: they cluster recordings by the findings they name, which helps a linear probe at 10% of labels and not at 100%. Paired text is therefore best understood as a label-efficiency tool, not a source of missing information. The deficit that remains sits on voltage and diffuse ST-T statements, and the open question is what signal-only objective would make such low-energy, global features as readable as timing and rhythm already are.

Ethics Statement

This work uses only de-identified, publicly documented electrocardiogram databases released for research use: MIMIC-IV-ECG, CODE-15, Ningbo, Georgia, Chapman–Shaoxing and the two CPSC-2018 collections for pretraining, and PTB-XL and the Shandong Provincial Hospital database for evaluation. MIMIC-IV-ECG is obtained through PhysioNet credentialed access and used under its data use agreement; the remaining sources are openly distributed by their maintainers. No new data were collected, no human subjects were recruited, and no attempt was made to re-identify any record. All reported numbers are frozen-probe macro-AUROC under the protocol of §2, which characterizes representation quality rather than clinical performance; none of the encoders studied here is validated for diagnostic use, and the paired reports are machine-generated interpretations with documented limitations. Readers should not treat the comparisons as evidence of clinical readiness for any of the eight models.

Reproducibility Statement

Code for pretraining, baseline retraining, probing and the equivalence procedure, together with the pretrained encoder weights, the retrained baseline checkpoints, the SPH split and the per-label probe outputs, will be released publicly upon acceptance.

AI Use Statement

We used generative AI tools solely to edit and polish the manuscript for readability, clarity, grammar, and style. We did not use generative AI tools for literature retrieval, research ideation or execution, drafting substantive sections of the paper, generating synthetic datasets, developing theoretical or conceptual frameworks, formulating or proving mathematical claims, designing methodology or experiments, or interpreting results. All AI-assisted edits were reviewed and revised by the authors, who take full responsibility for the final content of the paper.

References

Erick A. Perez Alday, Annie Gu, Amit J. Shah, Chad Robichaux, An-Kwok Ian Wong, Chengyu Liu, Feifei Liu, Ali Bahrami Rad, Andoni Elola, Salman Seyedi, Qiao Li, Ashish Sharma, Gari D. Clifford, and Matthew A. Reyna. Classification of 12-lead ECGs: The PhysioNet/computing in cardiology challenge 2020. *Physiological Measurement*, 41(12):124003, 2020. doi: 10.1088/1361-6579/abc960.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629, 2023.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pp. 1298–1312, 2022.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 1597–1607, 2020.

Edoardo Coppola, Mattia Savardi, Mauro Massussi, Marianna Adamo, Marco Metra, and Alberto Signoroni. HuBERT-ECG: A self-supervised foundation model for broad and scalable cardiac applications. *medRxiv*, 2024. doi: 10.1101/2024.11.14.24317328. URL https://www.medrxiv.org/content/10.1101/2024.11.14.24317328.

Brian Gow, Tom Pollard, Larry A. Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Jonathan W. Waks, Parastou Eslami, Tanner Carbonati, Ashish Chaudhari, Elizabeth Herbst, Dana Moukheiber, Seth Berkowitz, Roger Mark, and Steven Horng. MIMIC-IV-ECG: Diagnostic electrocardiogram matched subset (version 1.0). PhysioNet, 2023.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, 2022.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 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.

Qiao Jin, Won Kim, Qingyu Chen, Donald C. Comeau, Lana Yeganova, W. John Wilbur, and Zhiyong Lu. MedCPT: Contrastive pre-trained transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651, 2023. doi: 10.1093/bioinformatics/btad651.

Sehun Kim. Learning general representation of 12-lead electrocardiogram with a joint-embedding predictive architecture. *arXiv preprint arXiv:2410.08559*, 2024. URL https://arxiv.org/abs/2410.08559.

Dani Kiyasseh, Tingting Zhu, and David A. Clifton. CLOCS: Contrastive learning of cardiac signals across space, time, and patients. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 5606–5615, 2021.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, 2019.

Daniël Lakens. Equivalence tests: A practical primer for t-tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362, 2017. doi: 10.1177/1948550617697177.

Che Liu, Zhongwei Wan, Cheng Ouyang, Anand Shah, Wenjia Bai, and Rossella Arcucci. Zero-shot ECG classification with multimodal learning and test-time clinical knowledge enhancement. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2403.06659. arXiv:2403.06659.

Feifei Liu, Chengyu Liu, Lina Zhao, Xiangyu Zhang, Xiaoling Wu, Xiaoyan Xu, Yulin Liu, Caiyun Ma, Shoushui Wei, Zhiqiang He, Jianqing Li, and Eddie Ng Yin Kwee. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. *Journal of Medical Imaging and Health Informatics*, 8(7):1368–1373, 2018.

Hui Liu, Dan Chen, Da Chen, Xiyu Zhang, Huijie Li, Lipan Bian, Minglei Shu, and Yinglong Wang. A large-scale multi-label 12-lead electrocardiogram database with standardized diagnostic statements. *Scientific Data*, 9(1):272, 2022. doi: 10.1038/s41597-022-01403-5.

Yeongyeon Na, Minje Park, Yunwon Tae, and Sunghoon Joo. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=WcOohbsF4H.

Huynh Dang Nguyen, Trong-Thang Pham, Ngan Le, and Van Nguyen. TolerantECG: A foundation model for imperfect electrocardiogram. In *Proceedings of the 33rd ACM International Conference on Multimedia (ACM MM)*, 2025. URL https://arxiv.org/abs/2507.09887. arXiv:2507.09887.

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. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature Communications*, 11(1):1760, 2020. doi: 10.1038/s41467-020-15432-4.

Jürg Schläpfer and Hein J. Wellens. Computer-interpreted electrocardiograms: Benefits and limitations. *Journal of the American College of Cardiology*, 70(9):1183–1192, 2017. doi: 10.1016/j.jacc.2017.07.723.

Donald J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, 1987. doi: 10.1007/BF01068419.

Harold Smulyan. The computerized ECG: Friend and foe. *The American Journal of Medicine*, 132 (2):153–160, 2019. doi: 10.1016/j.amjmed.2018.08.025.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I. Lunze, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7(1):154, 2020. doi: 10.1038/s41597-020-0495-6.

Kuba Weimann and Tim O. F. Conrad. Self-supervised pre-training with joint-embedding predictive architecture boosts ECG classification performance. *Computers in Biology and Medicine*, 196: 110809, 2025. doi: 10.1016/j.compbiomed.2025.110809. arXiv:2410.13867.

Jianwei Zheng, Jianming Zhang, Sidy Danioko, Hai Yao, Hangyuan Guo, and Cyril Rakovski. A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Scientific Data*, 7(1):48, 2020. doi: 10.1038/s41597-020-0386-x.

A Appendix

A.1 Retained statement lists

Statements are scored when they have at least ten positive test records (§2.5); n_+ is that count. PTB-XL retains 49 of 71 SCP-ECG statements and SPH 26 of 44 AHA primary statements.

Table 2: **Scored PTB-XL statements** (49 of 71).

Code	Statement	n_+	Code	Statement	n_+
1AVB	First degree AV block	79	LMI	Lateral myocardial infarction	20
ABQRS	Abnormal QRS	322	LNGQT	Long QT-interval	11
AFIB	Atrial fibrillation	152	LOWT	Low amplitude T-waves	44
ALMI	Anterolateral myocardial infarct…	27	LPFB	Left posterior fascicular block	18
AMI	Anterior myocardial infarction	35	LPR	Prolonged PR interval	34
ANEUR	ST-T changes compatible with ven…	10	LVH	Left ventricular hypertrophy	214
ASMI	Anteroseptal myocardial infarction	234	LVOLT	Low QRS voltages in the frontal…	18
CLBBB	Complete left bundle branch block	54	NDT	Non-diagnostic T abnormalities	182
CRBBB	Complete right bundle branch block	54	NORM	Normal ECG	963
DIG	Digitalis-effect	18	NST_	Non-specific ST changes	77
ILMI	Inferolateral myocardial infarct…	48	NT_	Non-specific T-wave changes	42
IMI	Inferior myocardial infarction	267	PAC	Atrial premature complex	40
INJAL	Subendocardial injury in anterol…	14	PACE	Normal functioning artificial pa…	28
INJAS	Subendocardial injury in anteros…	22	PVC	Ventricular premature complex	114
INVT	Inverted T-waves	29	QWAVE	Q waves present	55
IRBBB	Incomplete right bundle branch b…	112	RAO/RAE	Right atrial overload/enlargement	10
ISCAL	Ischemic in anterolateral leads	66	RVH	Right ventricular hypertrophy	12
ISCAS	Ischemic in anteroseptal leads	17	SARRH	Sinus arrhythmia	77
ISCIL	Ischemic in inferolateral leads	18	SBRAD	Sinus bradycardia	64
ISCIN	Ischemic in inferior leads	22	SR	Sinus rhythm	1674
ISCLA	Ischemic in lateral leads	13	STACH	Sinus tachycardia	82
ISC_	Non-specific ischemic	128	STD_	Non-specific ST depression	101
IVCD	Non-specific intraventricular co…	79	SVARR	Supraventricular arrhythmia	14
LAFB	Left anterior fascicular block	162	VCLVH	Voltage criteria (QRS) for left…	87
LAO/LAE	Left atrial overload/enlargement	42			

Table 3: **Scored SPH statements** (26 of 44).

Code	Statement	n_+	Code	Statement	n_+
1	Normal ECG	2777	105	Incomplete right bundle-branch b…	248
21	Sinus tachycardia	145	106	Right bundle-branch block	147
22	Sinus bradycardia	547	120	Right-axis deviation	34
23	Sinus arrhythmia	310	121	Left-axis deviation	27
30	Atrial premature complexes	117	125	Low voltage	62
36	Junctional premature complexes	13	142	Left ventricular hypertrophy	38
50	Atrial fibrillation	134	145	ST deviation	365
51	Atrial flutter	20	146	ST deviation with T-wave change	206
60	Ventricular premature complexes	205	147	T-wave abnormality	434
82	Prolonged PR interval	52	153	ST-T change of ventricular hyper…	23
86	AV block, varying conduction	10	160	Anterior MI	10
101	Left anterior fascicular block	31	161	Inferior MI	24
104	Left bundle-branch block	17	165	Anteroseptal MI	23

A.2 Pretraining corpus

Table 4 lists the seven sources of the pretraining pool of §2.1, their native sampling rates and durations, and whether they carry reports. All are unified by the one preprocessing pipeline described there.

Table 4: Pretraining corpus. All sources are unified to 500 Hz, 10 s, per-record z-scored. Patient identifiers are not released uniformly across sources and are therefore not tabulated.

Source	Records	Native rate	Native duration	Reports
MIMIC-IV-ECG (Gow et al., 2023)	800,035	500 Hz	10 s	yes
CODE-15 (Ribeiro et al., 2020)	345,779	400 Hz	7–10 s	no
Ningbo (Zheng et al., 2020)	34,905	500 Hz	10 s	no
Georgia (Alday et al., 2020)	10,344	500 Hz	10 s	no
Chapman–Shaoxing (Zheng et al., 2020)	10,247	500 Hz	10 s	no
CPSC-2018 (Liu et al., 2018)	6,877	500 Hz	variable	no
CPSC-2018-Extra (Liu et al., 2018)	3,453	500 Hz	variable	no

A.3 SPH evaluation split and held-out verification

SPH (Liu et al., 2022) comprises 25,770 recordings from 24,666 patients. Because some patients contribute several recordings, the split is made by patient, $15{,}442/5{,}169/5{,}159$ recordings for training, validation and test, stratified on each patient's first primary AHA statement, with zero patient overlap between splits verified after the fact. SPH contributes no records to the pretraining corpus of Table 4. We verified this twice. First, against the shard index of the pretraining pool, which records the source of every one of the $1{,}211{,}640$ pretraining recordings: none is drawn from SPH. Second, at the signal level, because a recording could enter two public collections under different identifiers: for a random sample of SPH test recordings we computed the maximum normalized cross-correlation against every pretraining recording from the Chinese-population sources in the pool, lead by lead, and compared it with the same statistic between pool recordings themselves, excluding self-matches. The largest SPH-to-pool correlation lies below the pool-to-pool distribution and no recording matches on any of the 12 leads, whereas the same procedure does recover the 746 duplicate pairs (1.3%) that exist within the pool itself, chiefly between CPSC-2018 and its companion collection. SPH is therefore held out from every encoder in the comparison.

A.4 Baseline retraining

Each baseline is retrained from its official implementation on the corpus of Table 4 with its own released configuration, including its architecture, input length, sampling rate and optimizer schedule, so that the comparison is between methods as published rather than between objectives on a common backbone (§2.2). Training length is part of each released configuration, so passes over the recordings a method saw differ, from 4.0 for ST-MEM to 21.1 for the JEPA encoder (Table 5); ECG-FM's fairseq recipe accumulates gradients over eight batches per update, so its $40{,}000$ updates at batch

64 are 16.9 passes, not 2.1. Three baselines do not consume the whole corpus: TolerantECG's pipeline trains on MIMIC-IV-ECG only, MERL's loader subsamples the MIMIC rows with a usable report to 392,000, and HuBERT-ECG's feature-dump stage is run on a 196,000-record subsample, all following the released code. Table 17 places every baseline against the JEPA checkpoint nearest its own budget. Every model is evaluated at the final checkpoint of its run. For the four methods that ship public weights, Table 16 reports the released checkpoint under the same probe. TolerantECG and CLOCS release no weights, so those rows exist only in retrained form. Frozen features for every model and benchmark were extracted once and cached, and each cached feature set was verified against a re-extraction of 128 records from the named checkpoint before use.

Table 5: Models compared. All are pretrained from the corpus of Table 4 and evaluated with the identical frozen probe. Parameter counts are for the encoder used at evaluation; "Reports" indicates whether the method consumes paired report text during pretraining. "Recordings seen" is the number of distinct corpus recordings the released pipeline trains on, and "Passes" the number of passes over those recordings under the released training length; each baseline follows its own configuration, so neither is equalized (§2.2, Appendix A.4).

Model	Objective family	Params	Reports	Recordings	Passes
JEPA	latent prediction	14.5M	no	1,211,640	21.1
JEPA + reports	latent prediction + CLIP	14.5M	yes	1,211,640	21.1
TolerantECG	distillation + CLIP	26.8M	yes	800,035[a]	5.0
MERL	CLIP	3.9M	yes	392,000[b]	13.0
ST-MEM	waveform reconstruction	85.2M	no	1,211,640	4.0
ECG-FM	wav2vec2 + CMSC	90.9M	no	1,211,640	16.9
HuBERT-ECG	discrete-target prediction	30.5M	no	196,000[c]	13.1
CLOCS	augmentation contrastive	0.1M	no	1,211,640	8.5

[a] MIMIC-IV-ECG only; its pipeline requires paired text. [b] MIMIC rows with a usable report, subsampled by its loader. [c] Random subsample at its feature-dump stage.

A.5 Probe protocol, noise anchoring and bootstrap variants

The three tests of §2.6. The bootstrap is paired over resampled test records. The permutation test exchanges the two models' predictions within each record and rebuilds the null for the macro gap, conditioning on the observed test set rather than resampling it. The Wilcoxon signed-rank test over the per-statement AUROCs is paired over the label axis and distinguishes a gain spread across statements from one carried by a few, the failure mode a macro average hides. Per-statement DeLong tests with Benjamini–Hochberg correction are reported alongside them.

Probe. The probe of §2.5 is a single linear layer fit with `BCEWithLogitsLoss` and per-statement positive weighting clipped to $[1, 100]$, using Adam at learning rate 10^{-2}, weight decay 10^{-4} and minibatches of 512, for at most 200 passes over the training split (7,000 steps on PTB-XL) with early stopping on validation macro-AUROC, evaluated once per pass with a patience of 20; features are standardized with training-split statistics. The same recipe, and the same fixed seed, are used for every encoder and every label budget. Refitting the probe from five initializations on identical features gives the standard deviations of Table 6; the ± 0.02 reference margin of §2.6 is checked against the paired-gap noise there, 0.0008 on PTB-XL and 0.0052 on SPH. Table 7 repeats every headline comparison with a bootstrap over patients rather than records, and Table 8 replaces the linear probe by a one-hidden-layer MLP. Table 9 evaluates the final five saved checkpoints of the JEPA run and of its report-anchored arm.

Table 6: **Probe re-initialization noise.** Standard deviation of macro-AUROC over five probe fits from different initializations on identical frozen features, and of the paired JEPA−TolerantECG gap across those fits. The ± 0.02 reference margin is four to twenty-five times the gap noise.

Benchmark	JEPA	JEPA + reports	TolerantECG	Paired gap
PTB-XL	0.0018	0.0014	0.0021	0.0008
SPH	0.0013	0.0014	0.0045	0.0052

Table 7: **Record against patient-cluster bootstrap.** Paired differences in macro-AUROC at 100% labels with 95% intervals from resampling test records (the primary procedure) and from resampling patients with all of their records. No verdict changes; no bound moves by more than 0.0011.

Benchmark	Comparison	Record bootstrap	Patient bootstrap
PTB-XL	JEPA - TolerantECG	-0.005 $[-0.012, +0.002]$	-0.005 $[-0.012, +0.002]$
PTB-XL	reports on JEPA	$+0.001$ $[-0.005, +0.008]$	$+0.002$ $[-0.006, +0.008]$
PTB-XL	JEPA - ST-MEM	$+0.051$ $[+0.042, +0.060]$	$+0.051$ $[+0.041, +0.060]$
PTB-XL	JEPA - ECG-FM	$+0.045$ $[+0.037, +0.053]$	$+0.045$ $[+0.037, +0.054]$
PTB-XL	JEPA - HuBERT-ECG	$+0.112$ $[+0.102, +0.123]$	$+0.113$ $[+0.101, +0.123]$
PTB-XL	JEPA - CLOCS	$+0.313$ $[+0.298, +0.328]$	$+0.313$ $[+0.298, +0.329]$
SPH	JEPA - TolerantECG	$+0.007$ $[-0.003, +0.018]$	$+0.007$ $[-0.003, +0.018]$
SPH	reports on JEPA	-0.003 $[-0.010, +0.004]$	-0.003 $[-0.010, +0.004]$
SPH	JEPA - ST-MEM	$+0.032$ $[+0.022, +0.042]$	$+0.032$ $[+0.022, +0.042]$
SPH	JEPA - ECG-FM	$+0.049$ $[+0.040, +0.058]$	$+0.049$ $[+0.039, +0.058]$
SPH	JEPA - HuBERT-ECG	$+0.082$ $[+0.073, +0.092]$	$+0.082$ $[+0.072, +0.092]$
SPH	JEPA - CLOCS	$+0.296$ $[+0.275, +0.316]$	$+0.297$ $[+0.275, +0.317]$

Table 8: **Probe capacity.** The primary linear probe against a one-hidden-layer MLP probe on the same frozen features at 100% labels, with the 90% interval of the paired difference, the smallest margin at which equivalence holds (its larger absolute bound) and the verdict at the ± 0.02 reference. Attention pooling and multi-crop probes are not executable under this protocol: PTB-XL records are exactly 10 s against a 10 s window, and TolerantECG's adapter returns an already-pooled vector.

Benchmark	Probe	JEPA	TolerantECG	Δ [90%]	Min. margin	At 0.02
PTB-XL	linear	0.921	0.926	-0.005 $[-0.011, +0.001]$	0.011	equivalent
PTB-XL	MLP	0.932	0.942	-0.010 $[-0.014, -0.006]$	0.014	equivalent
SPH	linear	0.967	0.960	$+0.007$ $[-0.002, +0.017]$	0.017	equivalent
SPH	MLP	0.973	0.971	$+0.002$ $[-0.003, +0.008]$	0.008	equivalent

Table 9: **Final five saved checkpoints.** Macro-AUROC at 100% labels of the last five checkpoints of the JEPA run and of its report-anchored arm, their standard deviation, and how many satisfy the equivalence test against TolerantECG at the ± 0.02 reference. On SPH the JEPA checkpoints sit 0.005–0.008 *above* TolerantECG, with 90% upper bounds of 0.015–0.018, so the smallest passing margin there is 0.018.

Benchmark	Arm	AUROC, last five	SD	TOST
PTB-XL	JEPA	0.920 0.921 0.921 0.920 0.921	0.0006	5/5
PTB-XL	JEPA + reports	0.920 0.921 0.923 0.923 0.923	0.0011	5/5
SPH	JEPA	0.965 0.968 0.968 0.966 0.967	0.0015	5/5
SPH	JEPA + reports	0.963 0.963 0.959 0.960 0.963	0.0018	5/5

A.6 Report-anchored variant and the shuffled-report control

The report-anchored arm of §2.4 adds a symmetric InfoNCE objective between the pooled ECG embedding and a frozen MedCPT embedding of the paired report, with weight 1.0, to the latent-prediction loss; recordings without a report keep the latent-prediction loss alone, so the two arms see identical recordings and identical optimizer updates. The shuffled-report control assigns every recording the report of another patient drawn at random, destroying the pairing while holding the objective, the text distribution and the training budget fixed. It exists on the earlier `v16` backbone, on which signal-only, real-report and shuffled-report arms were each trained for $100{,}000$ steps; Table 10 gives the three arms under the primary probe. No noise-embedding control was trained, and we do not report one.

Table 10: **The shuffled-report control.** All three arms are the 100,000-step checkpoints of the earlier `v16` backbone (§2.2): signal-only, real reports, and reports shuffled across patients. Differences are paired against the signal-only arm with 95% bootstrap intervals. No noise-embedding control was trained.

Benchmark	Arm	10%	Δ vs signal	100%	Δ vs signal
PTB-XL	signal-only	0.820	–	0.876	–
PTB-XL	+ reports	0.904	+0.084 [+0.073, +0.096]	0.923	+0.047 [+0.038, +0.055]
PTB-XL	+ shuffled reports	0.797	−0.023 [−0.036, −0.010]	0.871	−0.005 [−0.015, +0.004]
SPH	signal-only	0.845	–	0.920	–
SPH	+ reports	0.945	+0.100 [+0.085, +0.116]	0.966	+0.045 [+0.033, +0.056]
SPH	+ shuffled reports	0.820	−0.025 [−0.043, −0.007]	0.922	+0.001 [−0.009, +0.011]

A.7 Recipe ablation

Figure 7 separates the two changes that distinguish the JEPA encoder from the `v16` baseline of §2.2: the switch to a visible-tokens-only student, and the averaging of the teacher's upper eight blocks. It is the evidence behind the claim in §3.5 that the first does nearly all the work and that the second is largely redundant with it once the encoder no longer spends capacity on mask tokens.

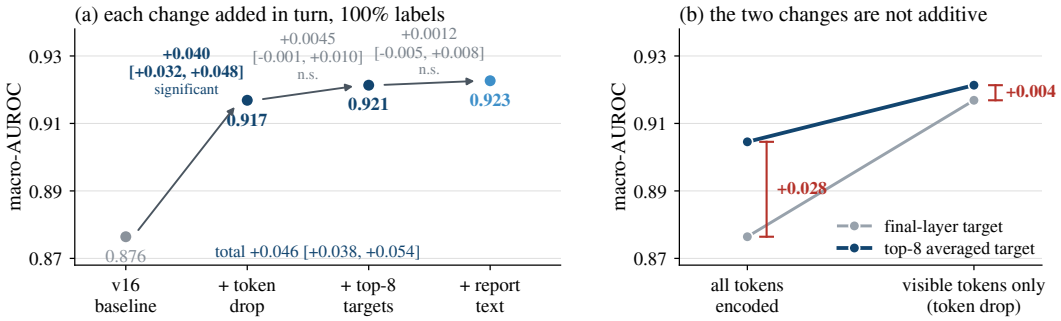

Figure 7: **One of the two changes does nearly all the work, and the second is largely redundant with it.** PTB-XL frozen-probe macro-AUROC at 100% labels (49 labels with ≥10 test positives); intervals are paired bootstrap over the test set (B=2000). **(a)** the recipe built up one change at a time. Switching the encoder to visible-tokens-only (token drop) is worth +0.040 [+0.032, +0.048]; on top of that, averaging the teacher target over the last eight blocks adds +0.0045 [−0.001, +0.010] and a frozen report-embedding objective adds +0.0012 [−0.005, +0.008], neither significant. The recipe as a whole is worth +0.046 over the `v16` baseline. **(b)** why the order in (a) is misleading if read alone: top-8 target averaging is worth +0.028 on an encoder that sees every token, but only +0.004 once the encoder is visible-only. The two changes address the same deficiency, and once the encoder no longer wastes capacity on mask tokens, richer targets have little left to add.

A.8 Label-budget curves and sweep

Figure 8 gives every contender's frozen-probe macro-AUROC at the three label budgets of §3.2, with the ±0.02 reference band around TolerantECG; the report effect on the JEPA encoder at the same budgets is Figure 3 in the main text.

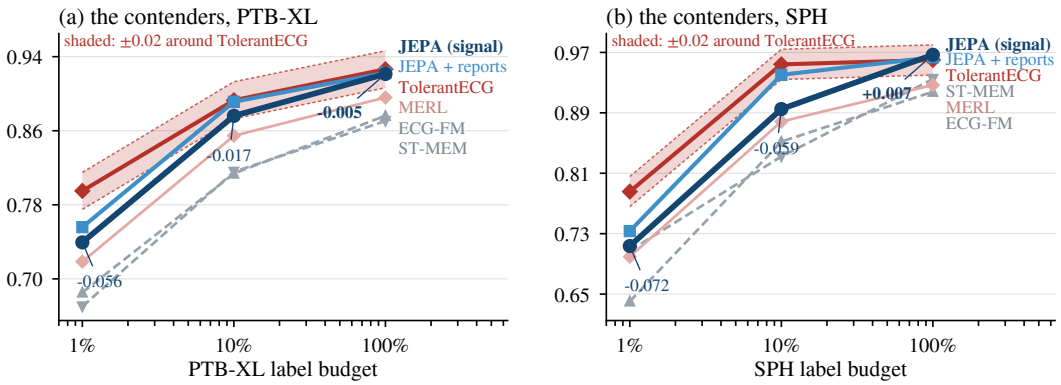

Figure 8: **The gap is a function of the label budget.** Frozen-probe macro-AUROC on **(a)** PTB-XL and **(b)** SPH against the fraction of downstream training labels for JEPA, its report-anchored arm, TolerantECG, MERL, ST-MEM and ECG-FM, with the ± 0.02 reference band around TolerantECG shaded and the JEPA$-$TolerantECG difference annotated at each budget.

The headline comparison uses three label budgets. Figure 9 repeats the report-effect measurement at seven, so that the 10% result of §3.2 can be read against its neighbours rather than against two distant points.

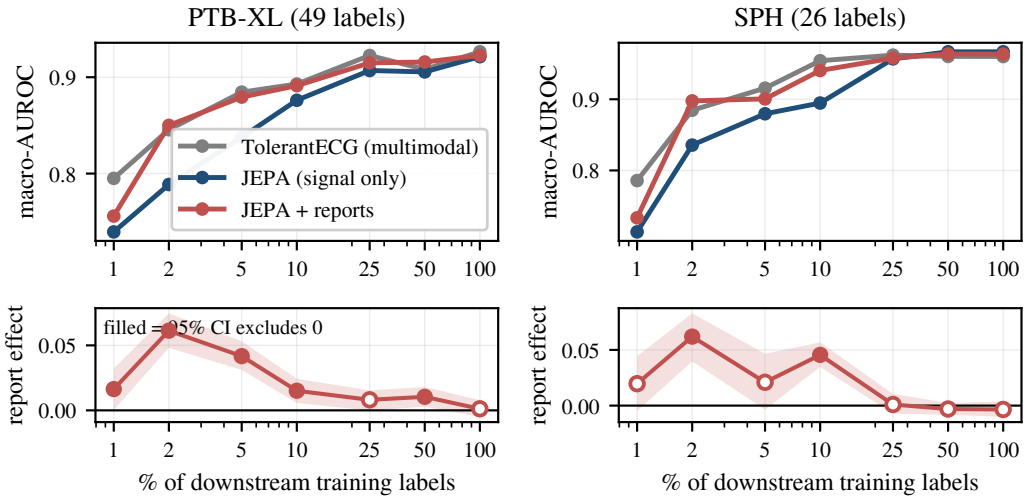

Figure 9: **The report effect across seven label budgets.** Top: frozen-probe macro-AUROC of the JEPA encoder, its report-anchored arm, and TolerantECG at 1, 2, 5, 10, 25, 50, and 100% of downstream labels, with the same subsampling seed at every budget. Bottom: the paired report effect with 95% bootstrap intervals; filled markers exclude zero. The effect peaks at 2%, is large below 10%, and is absent at 100% labels; adjacent budgets disagree by more than their intervals suggest, which is what a single probe fit per cell produces, so no threshold is claimed.

A.9 The report effect by label budget

Table 11 gives the numbers behind §3.2: the paired effect at each budget, how many of the three tests of §2.6 it passes, and what becomes of it when the validation split is subsampled at the training rate. The 10% effect survives that control on both benchmarks and the 1% effect does not, which is why only the former is attributed to text.

Table 11: **The report effect on the same encoder, by label budget.** Paired difference in macro-AUROC between the JEPA encoder with and without the report objective, the number of the three tests (paired bootstrap, permutation, Wilcoxon over per-statement AUROCs) significant at 0.05, and the same quantity with the validation split subsampled at the training rate (§2.5).

Benchmark	Budget	JEPA	+ reports	Δ [95% CI]	Tests	Δ, val. control	Tests
PTB-XL	1%	0.739	0.756	$+0.016$ $[+0.000, +0.032]$	1/3	-0.002	0/3
PTB-XL	10%	0.876	0.891	$+0.015$ $[+0.006, +0.024]$	3/3	$+0.028$	3/3
PTB-XL	100%	0.921	0.923	$+0.001$ $[-0.005, +0.008]$	0/3	$+0.001$	0/3
SPH	1%	0.714	0.733	$+0.020$ $[-0.005, +0.044]$	0/3	-0.023	0/3
SPH	10%	0.895	0.940	$+0.045$ $[+0.035, +0.057]$	3/3	$+0.017$	2/3
SPH	100%	0.967	0.963	-0.003 $[-0.010, +0.004]$	0/3	-0.003	0/3

A.10 Per-category breakdown

Table 12 aggregates the per-statement results of §3.4 by statement group. This is the form in which the main text quotes the rhythm-against-morphology pattern and the concentration of the 10% report effect on the statements a report names explicitly.

Table 12: **JEPA against TolerantECG, and the report effect, by statement group.** PTB-XL groups follow the SCP-ECG statement classes; SPH groups the AHA primary statements into rhythm (categories C–H) and morphology (I–M). n is the number of scored statements in the group; intervals are 95% paired bootstrap.

Benchmark	Group	n	Budget	JEPA $-$ TolerantECG	Report effect
PTB-XL	diagnostic	31	10%	-0.008 $[-0.022, +0.007]$	$+0.017$ $[+0.005, +0.030]$
PTB-XL	diagnostic	31	100%	-0.002 $[-0.011, +0.008]$	$+0.007$ $[-0.002, +0.016]$
PTB-XL	form	15	10%	-0.045 $[-0.067, -0.022]$	$+0.014$ $[-0.005, +0.033]$
PTB-XL	form	15	100%	-0.017 $[-0.029, -0.006]$	$+0.014$ $[+0.002, +0.026]$
PTB-XL	rhythm	7	10%	$+0.006$ $[-0.014, +0.028]$	-0.014 $[-0.032, +0.006]$
PTB-XL	rhythm	7	100%	$+0.002$ $[-0.013, +0.015]$	-0.028 $[-0.048, -0.008]$
SPH	rhythm	10	10%	-0.004 $[-0.010, +0.002]$	-0.014 $[-0.034, +0.006]$
SPH	rhythm	10	100%	$+0.016$ $[-0.002, +0.032]$	-0.021 $[-0.037, -0.004]$
SPH	morphology	15	10%	-0.096 $[-0.111, -0.081]$	$+0.087$ $[+0.074, +0.101]$
SPH	morphology	15	100%	$+0.001$ $[-0.010, +0.019]$	$+0.009$ $[+0.005, +0.014]$

A.11 Per-statement differences

Table 13 lists the individual statements behind Figure 6, restricted to those on which the paired interval excludes zero at 100% labels, with the report effect on the same statements alongside for comparison.

Table 13: **Statements on which the JEPA–TolerantECG interval excludes zero at 100% labels.** Per-statement AUROC at 100% labels for JEPA, TolerantECG and ST-MEM, the paired difference JEPA$-$TolerantECG with its 95% bootstrap interval, and the report effect on the JEPA encoder at 10% labels for the same statement. n_+ is the number of positive test records. Sorted by the difference; every statement not listed has an interval that includes zero. Intervals are not corrected for multiplicity.

Bench	Code	Statement	n_+	JEPA	Tol.	ST-MEM	JEPA$-$Tol. [95%]	Rep. 10%
PTB-XL	LVOLT	Low QRS voltages in the frontal and ho…	18	0.880	0.962	0.832	-0.080 $[-0.147, -0.027]$	$+0.096$
PTB-XL	VCLVH	Voltage criteria (QRS) for left ventri…	87	0.791	0.849	0.744	-0.057 $[-0.111, -0.010]$	-0.000
PTB-XL	ISCIL	Ischemic in inferolateral leads	18	0.923	0.966	0.881	-0.042 $[-0.097, -0.002]$	$+0.065$
PTB-XL	NST_	Non-specific ST changes	77	0.797	0.837	0.799	-0.040 $[-0.073, -0.007]$	$+0.026$
PTB-XL	PACE	Normal functioning artificial pacemaker	28	0.954	0.987	0.941	-0.029 $[-0.093, -0.000]$	$+0.023$
PTB-XL	NDT	Non-diagnostic T abnormalities	182	0.901	0.921	0.892	-0.020 $[-0.035, -0.006]$	$+0.006$
PTB-XL	ISC_	Non-specific ischemic	128	0.942	0.960	0.929	-0.017 $[-0.033, -0.002]$	$+0.030$
PTB-XL	IMI	Inferior myocardial infarction	267	0.906	0.923	0.851	-0.017 $[-0.032, -0.003]$	$+0.024$
PTB-XL	NORM	Normal ECG	963	0.931	0.946	0.922	-0.014 $[-0.021, -0.008]$	$+0.003$
PTB-XL	ASMI	Anteroseptal myocardial infarction	234	0.969	0.979	0.934	-0.010 $[-0.017, -0.004]$	-0.004
PTB-XL	PVC	Ventricular premature complex	114	0.993	0.989	0.986	$+0.005$ $[+0.001, +0.010]$	$+0.000$
SPH	125	Low voltage	62	0.949	0.986	0.900	-0.036 $[-0.064, -0.017]$	$+0.107$
SPH	165	Anteroseptal MI	23	0.951	0.990	0.968	-0.036 $[-0.094, -0.003]$	$+0.230$
SPH	153	ST-T change due to ventricular hypertr…	23	0.962	0.988	0.849	-0.024 $[-0.046, -0.012]$	$+0.153$
SPH	82	Prolonged PR interval	52	0.972	0.997	0.995	-0.023 $[-0.056, -0.006]$	$+0.022$
SPH	145	ST deviation	365	0.928	0.950	0.920	-0.022 $[-0.031, -0.012]$	$+0.061$
SPH	121	Left-axis deviation	27	0.968	0.983	0.946	-0.015 $[-0.031, -0.002]$	$+0.056$
SPH	147	T-wave abnormality	434	0.942	0.953	0.925	-0.011 $[-0.019, -0.002]$	$+0.150$
SPH	1	Normal ECG	2777	0.940	0.947	0.915	-0.007 $[-0.012, -0.003]$	$+0.021$
SPH	22	Sinus bradycardia	547	0.996	0.994	0.994	$+0.002$ $[+0.001, +0.002]$	-0.005
SPH	51	Atrial flutter	20	0.998	0.989	0.995	$+0.008$ $[+0.004, +0.014]$	$+0.024$
SPH	120	Right-axis deviation	34	0.987	0.970	0.893	$+0.017$ $[+0.005, +0.032]$	$+0.098$
SPH	60	Ventricular premature complexes	205	0.997	0.976	0.971	$+0.019$ $[+0.005, +0.038]$	$+0.036$
SPH	30	Atrial premature complexes	117	0.984	0.960	0.864	$+0.023$ $[+0.007, +0.043]$	-0.021
SPH	23	Sinus arrhythmia	310	0.947	0.883	0.797	$+0.063$ $[+0.046, +0.082]$	-0.121
SPH	160	Anterior MI	10	0.996	0.898	0.997	$+0.089$ $[+0.000, +0.323]$	$+0.079$

A.12 Representation analyses

Table 14 gives the report-predictability regression of §3.3 in full, Table 15 the label-free geometry measures behind Figure 4; the residual task-information matrices are Figure 5 in the main text.

Table 14: **Report predictability from frozen features.** Ridge regression from frozen features to the 768-d MedCPT embedding of the paired report on $20,081$ MIMIC-IV-ECG recordings for fitting and $19,919$ patient-disjoint recordings for evaluation; retrieval is a recording's own report among $1,000$ candidates (chance 0.001). Both encoders pretrained on all MIMIC rows, so this measures representational content under matched exposure; the ridge is out-of-sample.

Encoder	R^2	R@1	R@5	R@10
JEPA (never shown a report)	0.289	0.148	0.275	0.337
JEPA + reports	0.391	0.215	0.381	0.450
Shuffled target (chance)	-0.002	–	–	–

Table 15: **Representation geometry against TolerantECG, label-free.** R^2 of a ridge map from each encoder's frozen features to TolerantECG's (fit on the training split, evaluated on the test split, 95% bootstrap over test records), linear CKA with TolerantECG, mutual 10-NN agreement with TolerantECG, and effective rank (TolerantECG: 40.5 on PTB-XL, 37.6 on SPH). Recoverability places JEPA in the multimodal band; CKA and neighborhood agreement do not, which is the geometry–information distinction of §3.3.

	PTB-XL				SPH			
Encoder	R^2	CKA	10-NN	erank	R^2	CKA	10-NN	erank
JEPA	0.711 [0.706, 0.715]	0.366	0.162	19.4	0.733 [0.730, 0.737]	0.360	0.115	17.9
JEPA + reports	0.758 [0.754, 0.762]	0.746	0.325	46.0	0.766 [0.763, 0.769]	0.711	0.235	49.5
MERL	0.723 [0.718, 0.728]	0.570	0.162	23.7	0.729 [0.725, 0.733]	0.503	0.108	22.3
ST-MEM	0.644 [0.638, 0.649]	0.484	0.107	31.5	0.661 [0.657, 0.666]	0.413	0.065	31.8
ECG-FM	0.638 [0.631, 0.644]	0.249	0.150	53.5	0.656 [0.652, 0.661]	0.305	0.102	48.3
HuBERT-ECG	0.462 [0.453, 0.470]	0.282	0.061	28.3	0.473 [0.467, 0.478]	0.199	0.027	28.1
CLOCS	0.120 [0.112, 0.128]	0.029	0.014	71.1	0.071 [0.067, 0.075]	0.012	0.009	72.1

A.13 CHECKPOINT PROVENANCE

Table 16 quantifies the provenance confound of §1. The four methods that ship public weights are scored twice under the identical probe, once from those weights and once from our retraining of the same code on the corpus of Table 4.

Table 16: **Released against retrained checkpoints under one probe.** Macro-AUROC at 100% labels. Every shift is negative: the released weights score higher than the same method retrained on our corpus.

	PTB-XL			SPH		
Model	released	retrained	shift	released	retrained	shift
MERL	0.902	0.896	−0.006	0.943	0.927	−0.016
ST-MEM	0.896	0.871	−0.026	0.944	0.935	−0.010
ECG-FM	0.883	0.876	−0.007	0.926	0.918	−0.008
HuBERT-ECG	0.834	0.809	−0.025	0.911	0.884	−0.026

A.14 JEPA AT EACH BASELINE'S TRAINING BUDGET

Training length is part of each released configuration (§2.2). Figure 10 shows the JEPA encoder's trajectory over the saved checkpoints, and Table 17 places every baseline against the JEPA checkpoint nearest its own budget. The ordering against every signal-only method is unchanged at matched passes.

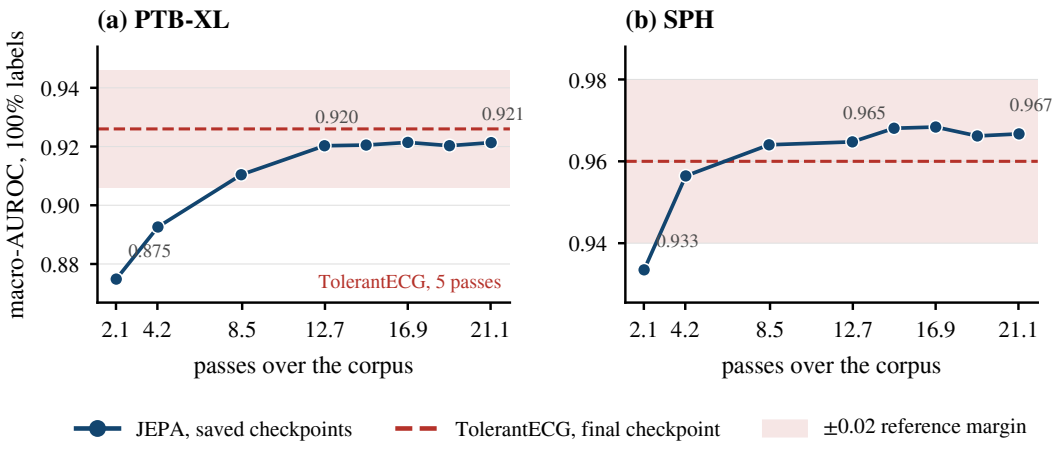

Figure 10: **The encoder's trajectory plateaus by** 13 **passes.** Frozen-probe macro-AUROC of the JEPA encoder at 100% labels at every saved checkpoint, against passes over the corpus (2.1 to 21.1). On PTB-XL the score is 0.920 at 12.7 passes and 0.921 at 21.1; on SPH 0.965 and 0.967. The dashed line is TolerantECG's final checkpoint at its authors' released recipe, five passes over MIMIC-IV-ECG; the band is the ±0.02 reference margin. The final schedule is therefore not what separates the encoder from the baselines, which it leads from 8.5 passes onward.

Table 17: **Matched-budget comparison.** Each baseline at its own released budget, in passes over the recordings it saw (Table 5), against the saved JEPA checkpoint nearest that budget in passes over the corpus, under the primary probe at 100% labels. The ECG-FM row uses its true 16.9 passes (Appendix A.4); against the checkpoint at 2.1 passes it would be a tie (−0.001 on PTB-XL, +0.016 on SPH).

	passes		**PTB-XL**			**SPH**		
Baseline	its	JEPA	baseline	JEPA	lead	baseline	JEPA	lead
ST-MEM	4.0	4.2	0.871	0.893	+0.022	0.935	0.956	+0.022
TolerantECG	5.0	4.2	0.926	0.893	−0.034	0.960	0.956	−0.003
CLOCS	8.5	8.5	0.608	0.910	+0.302	0.671	0.964	+0.293
MERL	13.0	12.7	0.896	0.920	+0.024	0.927	0.965	+0.038
ECG-FM	16.9	16.9	0.876	0.921	+0.045	0.918	0.968	+0.050

TolerantECG at its own 5-pass budget exceeds the JEPA checkpoint at 4.2 passes on PTB-XL; whether it would continue to improve beyond its released schedule is not testable without retraining it, and we do not speculate.

A.15 Alternative explanations

Table 18 collects the remaining threats to the full-label result that can be tested without training anything further, each paired with the evaluation-only check that addresses it.

Table 18: **Alternative explanations, tested without training.** Each row is a threat to the full-label result and the evaluation-only check that addresses it. "4 passes" compares the JEPA checkpoint nearest ST-MEM's budget with ST-MEM.

Threat	PTB-XL	SPH	Outcome
Baselines retrained badly	released−retrained shifts $+0.006$ to $+0.026$	$+0.008$ to $+0.026$	released and retrained checkpoints are not interchangeable; corpus, schedule and stochasticity cannot be separated after the fact
JEPA trained longer	$+0.022$ over ST-MEM at 4 passes	$+0.022$	the lead survives a matched budget
Target averaging does the work	$+0.005$ $[-0.001, +0.010]$	$+0.001$ $[-0.003, +0.006]$	neutral at 100%; not the mechanism
One-checkpoint accident	last-5 SD 0.0006; $5/5$ inside the margin	SD 0.0015; all five 0.005–0.008 ahead	no checkpoint sensitivity
Linear probe favors JEPA	MLP probe: -0.010 $[-0.014, -0.006]$	MLP probe: $+0.002$ $[-0.003, +0.008]$	equivalent under both probes
Test-set patient structure	patient bootstrap moves bounds ≤ 0.0011	≤ 0.0011	0 of 12 verdicts change

A.16 EFFICIENCY

Table 19 gives the cost measurements summarized by the parameter and FLOP columns of Table 1. Each model is measured through its own native input pipeline rather than a common one, so the figures reflect the methods as published.

Table 19: **Efficiency.** Parameters of the evaluated encoder, FLOPs and wall-clock of one forward pass over a single 10 s 12-lead record through each model's native input pipeline (batch size 1, one GPU), embedding dimension, native input shape, and probe training cost per pass over the PTB-XL training split ($17{,}418$ records, 35 steps of 512).

Model	Params (M)	GFLOPs	Fwd (ms)	d	Input	Probe (ms/pass)
JEPA	14.47	3.00	1.01	384	(12, 5000)	17
JEPA + reports	14.47	3.00	1.01	384	(12, 5000)	13
TolerantECG	26.81	13.06	3.08	768	(12, 5000)	14
MERL	3.85	3.43	0.72	512	(12, 5000)	14
ST-MEM	85.15	70.71	7.99	768	(12, 2250)	12
ECG-FM	90.88	29.24	5.28	768	(12, 2500)	12
HuBERT-ECG	30.47	7.34	2.20	512	(6000,)	14
CLOCS	0.09	0.01	1.91	256	(1, 2500, 12)	11