WHEN ARE FOUNDATION MODELS GOOD DISTILLATION TEACHERS? A CONTROLLED STUDY OF TEACHER SELECTION IN ECG

Anonymous authors
Paper under double-blind review

ABSTRACT

Large pretrained models are increasingly used as knowledge-distillation teachers, yet selecting a teacher is often reduced to choosing the largest or strongest available pretrained model. This assumes that model-to-task transferability also predicts teacher-to-student transfer, an assumption that need not hold when supervision must pass through a compact student. We study this distinction in a controlled setting using multi-label ECG classification. Holding the student architecture, distillation objective, optimization protocol, and evaluation fixed, we compare two public foundation models with task-specific supervised teachers; on PTB-XL, our audit spans six distinct teacher models, and we repeat the central foundation-teacher comparison on PhysioNet/CinC 2021. The two foundation teachers exhibit sharply different transfer behavior: ECG-FM provides no robust student improvement on either dataset, whereas ECGFounder improves PTB-XL AUROC, mAP, and F1@0.5 and outperforms ECG-FM on those metrics on PhysioNet. We test calibration, downstream fine-tuning, teacher scale, and architecture family, but none individually accounts for the transfer gap. Across the tested teacher pool, downstream target-task fit separates ineffective KD candidates more clearly than foundation-model status or parameter scale, although it does not perfectly rank the strongest teachers. These results motivate treating distillation-teacher selection as a task-conditioned transfer problem distinct from selecting a model to fine-tune.

1 INTRODUCTION

Foundation models are increasingly used not only as predictors, but as teachers for smaller deployable models. Knowledge distillation (KD) (Hinton et al., 2015) provides a natural mechanism for transferring knowledge from their predictions into compact students, but it introduces a model-selection problem that standard foundation-model evaluation does not answer: which pretrained model is actually a useful teacher for a particular student and target task? A natural heuristic is to prefer the largest or independently strongest teacher. Yet teacher strength alone does not determine distillation performance. Increasing teacher accuracy or capacity can fail to improve a smaller student, motivating teacher assistants, student-compatible training protocols, and a distinction between teacher imitation and student generalization (Cho & Hariharan, 2019; Mirzadeh et al., 2020; Stanton et al., 2021; Beyer et al., 2022). Intermediate checkpoints can even outperform more accurate converged teachers (Wang et al., 2022). As ecosystems of pretrained models grow, selecting which model to distill from therefore becomes a distinct machine-learning problem.

Recent work increasingly treats teacher quality as relational rather than intrinsic. SoTeacher trains teachers with downstream student performance in view (Dong et al., 2024); progressive distillation finds that intermediate teachers can provide an implicit curriculum (Panigrahi et al., 2025); and GRACE uses student-gradient information to select language-model teachers for particular student–task pairs (Panigrahi et al., 2026). Related transferability methods such as LogME (You et al., 2021) and TEMPLATE (Zhang et al., 2025) ask which pretrained model is likely to adapt successfully to a downstream task. But model-to-task transferability and teacher-to-student transferability are different problems. A model that adapts well to a target task need not provide soft targets that a substantially smaller student can exploit. Thus the best model to fine-tune need not be the best model to distill from.

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.

We study this distinction using ECG foundation models as a controlled testbed. ECG is particularly suitable because a rapidly growing ecosystem of pretrained models is now being systematically compared across downstream tasks. An ICLR 2026 benchmark evaluated eight ECG foundation models across 26 clinically relevant tasks and 12 public datasets (Al-Masud et al., 2026), while the recent BenchECG benchmark similarly evaluates foundation-model transfer across diverse ECG tasks and datasets (Lunelli et al., 2026). These efforts establish both the promise and heterogeneity of ECG foundation models, but they primarily ask which pretrained model transfers well to the task itself. They do not establish which of those models transfers knowledge effectively through a compact student. That distinction is practically consequential: the two foundation teachers considered here contain 30.7M and 90.4M parameters, whereas our deployed student contains 1.22M and runs approximately 7–32$\times$ faster in our CPU profile.

We therefore treat foundation-model distillation as a controlled teacher-selection problem. We use ECG-FM (McKeen et al., 2025) and ECGFounder (Li et al., 2025) as focal public foundation teachers rather than as an exhaustive FM benchmark. On PTB-XL, the study spans six distinct teacher models: the two foundation models, three same-family supervised ResNet teachers spanning 3.2M–17.5M parameters, and a cross-family Transformer teacher. We repeat the central foundation-teacher comparison on PhysioNet/CinC 2021. Throughout the headline comparisons, we fix the 1.22M-parameter student, KD objective, optimization protocol, and evaluation procedure, holding the student-side pipeline fixed while changing the teacher. Two additional compact architectures provide auxiliary student-side sanity checks.

The resulting teacher outcomes are sharply different. ECG-FM provides no robust student improvement over no-KD on either target dataset, whereas ECGFounder improves PTB-XL AUROC, mAP, and F1@0.5 and outperforms ECG-FM on those metrics on PhysioNet. Calibration correction, additional downstream tuning, a controlled teacher-size ladder, and a cross-family comparison do not individually account for the discrepancy. Across the tested teacher pool, measured downstream target-task fit separates ineffective KD candidates more clearly than foundation-model status or parameter scale, although it does not perfectly rank the strongest teachers. Together, these results expose a distinction not captured by conventional foundation-model benchmarks: choosing a model to distill from is a separate model-selection problem from choosing one to fine-tune.

Contributions.

1. **A controlled audit of foundation-model teacher selection.** We connect recent ECG foundation-model benchmarking to the distinct problem of distillation, evaluating two public foundation teachers within a six-teacher PTB-XL audit while holding the compact student and student-side KD pipeline fixed across the central comparisons.
2. **A systematic test of common teacher-quality explanations.** We examine calibration correction, stronger downstream fine-tuning, a same-family teacher-capacity ladder, and a cross-family teacher. None of these factors alone explains the large difference in distillation utility between the foundation teachers under our protocol.
3. **Evidence for task-conditioned teacher screening.** Across the tested teacher pool, downstream target-task performance distinguishes ineffective KD candidates more clearly than foundation-model status or parameter scale, suggesting that foundation teachers should be screened for the target task before committing to student distillation.

2 Setup and evaluation protocol

We study how teacher choice affects logit-based knowledge distillation into a fixed compact ECG classifier. The student, distillation objective, training protocol, and evaluation protocol are held fixed in the headline KD comparisons; the teacher varies. The ECG-FM recipe ablation and exploratory composed condition are separate interventions described in §4.

Datasets. **PTB-XL** (Wagner et al., 2020) at 500 Hz, filtered to the five diagnostic superclasses (NORM, MI, STTC, CD, HYP) as a multi-label task. We use the standard fold split: folds 1–8 for train (17,073 records), fold 9 for validation (2,145), fold 10 for test (2,157). Per-lead z-score normalisation is computed on the training split only. **PhysioNet/CinC 2021** (Reyna et al., 2021)

at 500 Hz, restricted to the five sub-databases that satisfy the 12-lead/500 Hz filter (CPSC 2018, CPSC 2018 Extra, Georgia, Chapman–Shaoxing, Ningbo), yielding 60,211 recordings annotated with 30 SNOMED-CT labels. PTB-XL is excluded from this set to avoid train/test overlap with the PTB-XL evaluation. We treat PhysioNet/CinC 2021 as a second target task with a different label space, not as a pretraining-disjoint generalization benchmark: ECG-FM's pretraining corpus draws on publicly available ECG sources that may overlap with PhysioNet 2021 constituent databases (see §6). We use an 80/10/10 stratified split on the primary label (6,022 test records).

Compact student. A 1.22 M-parameter 1D ResNet (COMPACTRESNET1D): stem convolution (k=15, s=2), four residual blocks with stride-2 downsampling and channel widths $[64, 128, 128, 256]$, global average pooling, and a linear head. Full details are in Appendix D. We additionally evaluate two smaller student architectures (a plain CNN and a ECGTRANSFORMER) in Appendix D; the main paper uses the ResNet student throughout.

Teachers. **Foundation teacher A: ECG-FM** (90.4 M parameters) (McKeen et al., 2025), a transformer-based ECG encoder pretrained with hybrid self-supervision on public ECG sources including PhysioNet 2021 and MIMIC-IV-ECG (Johnson et al., 2023). We use the authors' public MIMIC-IV-ECG-finetuned checkpoint, replace its head, and fine-tune on each downstream dataset. We re-implement ECG-FM in pure PyTorch to keep a uniform training stack (Appendix E). We also evaluate a post-hoc temperature-scaled (Guo et al., 2017) variant and a best-of-sweep fine-tuned variant found by a four-configuration learning-rate / weight-decay / warmup sweep. **Foundation teacher B: ECGFounder** (30.7 M) (Li et al., 2025), a Net1D (RegNet-1D-style) CNN backbone supervised-pretrained on a $\sim$10 M-record cohort for 150-class ECG condition prediction. We load the public 12-lead checkpoint, strip the pretraining head, and attach a fresh linear classification head under the same fine-tuning protocol (Appendix E). **Supervised ResNet teacher** (7.8 M), a same-family ResNet1D trained from scratch on the target dataset. **Teacher ladder** (§4.3): three additional same-family teachers at 3.2 M, 7.8 M, and 17.5 M parameters, plus a cross-family matched-capacity Transformer teacher at 4.8 M, trained on the target dataset under the same protocol.

Knowledge distillation objective. We use logit-based KD with sigmoid soft targets, which respects the multi-label structure:

$$\mathcal{L} \;=\; \alpha \cdot \mathcal{L}_{\text{BCE}}(z_s, y) + (1-\alpha) \cdot T^2 \cdot \mathcal{L}_{\text{BCE}}\Big(\frac{z_s}{T},\, \sigma\Big(\frac{z_t}{T}\Big)\Big)\,, \tag{1}$$

where z_s and z_t are student and teacher logits, σ is the sigmoid, and T^2 compensates for the temperature-induced gradient scale (Hinton et al., 2015). Teacher logits are detached. Our default recipe is T=4, α=0.5. An exploratory ECG-FM stress test (§4.2) instead uses the recipe selected post hoc by the highest single-seed *test* F1 in a $T \times \alpha$ grid with the default ECG-FM teacher.

Training protocol. AdamW (lr $= 10^{-3}$, weight decay 10^{-4}), cosine annealing, batch size 64, up to 60 epochs with patience 15 on validation AUROC, BCE-with-logits loss for no-KD and the objective in Eq. 1 for KD. The same optimization and early-stopping rule apply, although the selected epoch and realized training duration can differ. All experiments run on a single NVIDIA RTX 5090.

Metrics and hierarchy. We report four macro-averaged metrics and one calibration diagnostic, organised in an explicit hierarchy. **Primary comparative metrics (threshold-free): macro-AUROC** (ranking) and **macro-AP (mAP)** (precision-recall); our headline claims are anchored on these, and we use *mAP* as the short name in the rest of the paper. **Deployment-facing metric: macro-F1 at the fixed 0.5 threshold** (*F1@0.5*; standard in ECG multi-label work). **Supplementary metric: macro-F1 at per-class validation-tuned thresholds** (*F1@vtuned*; per-class thresholds chosen on the validation split and applied to the test split); we report it for completeness but not as a headline metric, because re-tuning per-class thresholds inside each bootstrap resample is rare-class-unstable and we do not put paired CIs on it. **Diagnostic metric:** expected calibration error (**macro-ECE**, 15 bins), for teacher and student; ECE is reported to characterise failure modes (e.g. the overconfident-teacher hypothesis of §4.1) rather than to order conditions. mAP is particularly important on PhysioNet, where class imbalance makes F1 at a fixed threshold unstable.

Statistical protocol. Unless otherwise stated, each condition is run for three training seeds ($\{0, 1, 2\}$). For headline comparisons we report multi-seed mean $\pm$ std, 95% paired bootstrap

confidence intervals on test-set metric differences (1000 resamples; logits seed-averaged before bootstrapping so paired CIs reflect test-sample variance rather than training-seed variance), and the *bootstrap win fraction* — the fraction of resamples in which one condition exceeds the other. Two conditions are considered to differ when the paired Δ CI excludes zero. We compute paired-bootstrap CIs for AUROC, mAP, ECE, and F1@0.5; F1@vtuned is reported as a seed-averaged difference only, because re-tuning per-class thresholds inside every resample is unstable when a class has few positives in a given resample (the standard rare-class instability noted in (Reyna et al., 2021)). The supplementary material reports per-seed F1@vtuned for every condition.

Reporting convention. Tables report two distinct estimators. Point rows are the mean$\pm$std of metrics computed separately for each training seed. For paired Δ rows, we first average logits across the three seeds within each condition, then bootstrap paired test examples and compute the metric difference. A Δ is therefore not generally the difference between the displayed point-row means; nonlinear metrics such as F1 can even differ in direction. For example, PTB-XL ECG-FM has a higher per-seed mean F1 than no-KD but a slightly negative paired ΔF1. Primary comparisons should be read from the Δ rows; e.g. the "Δ(ECG-FM $-$ baseline) $= +0.0425^{\dagger}$" ECE entry in Table 1 means ECG-FM raises student ECE by 0.0425 under this paired estimator, with † marking that the 95% paired CI excludes zero. These CIs capture test-example uncertainty conditional on seed-averaged logits, not full training-seed uncertainty. Per-seed values are in the supplementary material.

3 Contrasting student outcomes from two foundation teachers

We evaluate four distillation conditions on two public ECG benchmarks, holding the student-side KD and evaluation pipeline fixed while changing the teacher: (i) *no teacher* — the ResNet1D student trained with BCE against the multi-label targets; (ii) *KD from ECG-FM* (90.4 M foundation model, fine-tuned); (iii) *KD from ECGFounder* (30.7 M foundation model, fine-tuned); and (iv) *KD from a supervised ResNet teacher* (7.8 M, trained from scratch). All four conditions are run with three seeds under the statistical protocol of §2.

3.1 KD from ECG-FM does not improve the student in this setting

On PTB-XL, the compact ResNet1D student without a teacher already reaches 0.928 macro-AUROC and 0.733 F1@0.5 (Table 1), exceeding the 90.4 M ECG-FM teacher's 0.892 test AUROC at roughly 1/75th the parameter budget. KD from ECG-FM does not improve ranking (ΔAUROC CI spans zero) and worsens calibration: every bootstrap resample has the KD student worse-calibrated than the baseline (ΔECE $+0.0425$, CI excludes zero). On PhysioNet, the failure mode is ranking instead of calibration: ΔAUROC -0.0039 (CI excludes zero, bootstrap win fraction for KD $= 0.000$), with threshold-robust metrics directionally consistent. Across both datasets ECG-FM never produces a robust advantage over the no-KD baseline.

3.2 A second foundation teacher and the supervised teacher both help

Swapping the foundation teacher for ECGFounder (Net1D/CNN backbone, $\sim 3\times$ fewer parameters than ECG-FM) reverses the direction on PTB-XL AUROC, mAP, and F1@0.5, with all three paired Δ CIs excluding zero (Table 1) and no ECE degradation. On PhysioNet the absolute AUROC gain over no-KD is small for any teacher (strong baseline), but the paired contrast between the two foundation teachers is sharp: ECGFounder beats ECG-FM on AUROC, mAP, and F1@0.5 with all three CIs excluding zero. KD from the 7.8 M supervised ResNet, which serves as the high-fit reference for the pipeline, also helps on F1@0.5 with CIs excluding zero on both datasets. On PTB-XL, ECGFounder has the highest AUROC and mAP point estimates among the KD teachers tested.

3.3 Second target task: PhysioNet/CinC 2021

Moving from PTB-XL (17 K records, 5 superclasses) to PhysioNet/CinC 2021 (60 K records, 30 classes) does not change the pattern (Table 2). ECG-FM KD again hurts: paired ΔAUROC CI excludes zero and is negative. ECGFounder is positive on every metric except ECE, with high bootstrap win fractions but CIs that span zero against the strong baseline. The cleanest PhysioNet contrast is again between the two foundation teachers: ECGFounder beats ECG-FM on AUROC,

Table 1: PTB-XL main result (3 seeds). ResNet1D student (1.22 M), four teacher conditions; ECE lower is better. Point rows report mean±std across three seeds; Δ rows are paired-bootstrap estimates over seed-averaged logits (§2; F1@vtuned is a per-seed mean difference). † marks 95% CIs excluding zero; primary comparisons should be read from the Δ rows.

	macro-AUROC	mAP	F1@0.5	F1@vtuned	ECE
No-KD baseline	$0.9278_{\pm.0011}$	$0.8224_{\pm.0036}$	$0.7325_{\pm.0113}$	$0.7512_{\pm.0014}$	$0.0291_{\pm.0067}$
KD from ECG-FM	$0.9263_{\pm.0031}$	$0.8208_{\pm.0073}$	$0.7392_{\pm.0078}$	$0.7469_{\pm.0058}$	$0.0776_{\pm.0028}$
KD from ECGFounder	$\mathbf{0.9341}_{\pm.0002}$	$\mathbf{0.8347}_{\pm.0009}$	$\mathbf{0.7539}_{\pm.0059}$	$\mathbf{0.7559}_{\pm.0006}$	$\mathbf{0.0278}_{\pm.0042}$
KD from supervised	$0.9298_{\pm.0010}$	$0.8257_{\pm.0024}$	$0.7488_{\pm.0059}$	$0.7517_{\pm.0024}$	$0.0317_{\pm.0039}$
Δ(ECG-FM − baseline)	-0.0012	-0.0009	-0.0014	-0.0043	$+0.0425^{\dagger}$
Δ(ECGFounder − baseline)	$+0.0045^{\dagger}$	$+0.0094^{\dagger}$	$+0.0145^{\dagger}$	$+0.0047$	-0.0017
Δ(supervised − baseline)	-0.0004	-0.0016	$+0.0115^{\dagger}$	$+0.0005$	$+0.0019$
Δ(ECGFounder − ECG-FM)	$+0.0057^{\dagger}$	$+0.0103^{\dagger}$	$+0.0159^{\dagger}$	$+0.0090$	$-0.0442^{\dagger}$

Table 2: PhysioNet/CinC 2021 main result (3 seeds). Same student and KD pipeline, 30-class multi-label; conventions as in Table 1.

	macro-AUROC	mAP	F1@0.5	F1@vtuned	ECE
No-KD baseline	$0.9574_{\pm.0014}$	$0.5401_{\pm.0011}$	$0.4899_{\pm.0081}$	$0.5389_{\pm.0070}$	$0.0078_{\pm.0005}$
KD from ECG-FM	$0.9554_{\pm.0004}$	$0.5343_{\pm.0016}$	$0.4775_{\pm.0038}$	$0.5310_{\pm.0070}$	$0.0092_{\pm.0001}$
KD from ECGFounder	$\mathbf{0.9609}_{\pm.0016}$	$0.5488_{\pm.0101}$	$0.4973_{\pm.0184}$	$0.5421_{\pm.0032}$	$0.0096_{\pm.0003}$
KD from supervised	$0.9604_{\pm.0005}$	$\mathbf{0.5532}_{\pm.0010}$	$\mathbf{0.5094}_{\pm.0011}$	$\mathbf{0.5464}_{\pm.0049}$	$\mathbf{0.0076}_{\pm.0003}$
Δ(ECG-FM − baseline)	$-0.0039^{\dagger}$	-0.0093	-0.0135	-0.0079	$+0.0028^{\dagger}$
Δ(ECGFounder − baseline)	$+0.0018$	$+0.0059$	$+0.0117$	$+0.0032$	$+0.0030^{\dagger}$
Δ(supervised − baseline)	-0.0007	-0.0016	$+0.0212^{\dagger}$	$+0.0075$	$+0.0014^{\dagger}$
Δ(ECGFounder − ECG-FM)	$+0.0056^{\dagger}$	$+0.0152^{\dagger}$	$+0.0251^{\dagger}$	$+0.0111$	$+0.0001$

mAP, and F1@0.5 with all three CIs excluding zero. Both ECGFounder and the supervised teacher show a small PhysioNet ECE penalty versus no-KD, with paired-bootstrap CIs excluding zero; we return to this in §6.

On the magnitude of these effects. The PTB-XL ECGFounder gains have paired-bootstrap CIs excluding zero but are numerically modest, against a strong 0.928-AUROC baseline on which F1 is the most movable metric. ECG-FM's Δs sit at-or-below zero on every primary metric across both datasets, while ECGFounder's sit at-or-above zero. At deployment, the cost of the wrong foundation teacher is the lost ECGFounder gain plus a $\sim$0.05 PTB-XL student-ECE penalty — a shift large enough to change operating-point behaviour.

4 Testing explanations for the ECG-FM result

§3 establishes different student outcomes but does not identify their cause. We test whether ECG-FM's shortfall persists after soft-target calibration and stronger downstream fine-tuning, then compare teacher size and architecture family under the same student pipeline. Table 3 summarizes the tested interventions. Calibration correction and retuning do not rescue ECG-FM, while the size and family controls do not explain the gap. These tests do not establish what causes the difference between the two foundation teachers.

4.1 Calibration is not sufficient to explain the failure

One natural hypothesis, following Guo et al. (2017), is that ECG-FM's logits are over-confident and the shortfall is therefore a calibration artifact. The default-fine-tuned ECG-FM's soft targets are indeed nearly hard (Table 7): mean max-probability 0.987; 10.7% of (sample, class) pairs are confident-wrong; only 1.6% sit in $[0.3, 0.7]$; per-class binary entropy is numerically undefined because the sigmoid saturates — though the "tuned" row of Table 7 reaches $H(p)$=0.22, so saturation

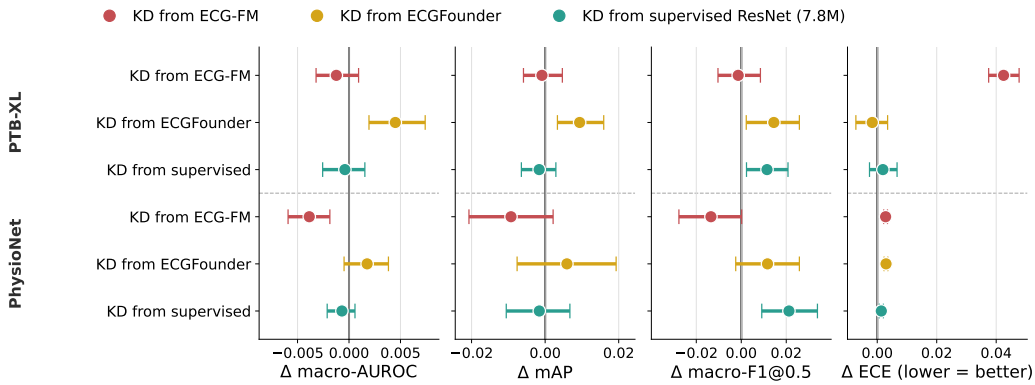


Figure 1: **Core result: paired-bootstrap Δ (KD − no-KD)** across the four metrics on PTB-XL (top) and PhysioNet/CinC 2021 (bottom). Points are paired-bootstrap means over seed-averaged logits; bars are 95% CIs. Red: ECG-FM. Amber: ECGFounder. Teal: supervised ResNet (7.8 M). ECG-FM sits at or below zero on AUROC/mAP/F1 and above zero on ECE; ECGFounder sits above zero on AUROC/mAP/F1 (PTB-XL CIs exclude zero); supervised KD gains F1 and is near zero elsewhere.

Table 3: Tests of candidate explanations for ECG-FM's shortfall. The composed recipe is exploratory because its single-seed grid cell was selected by test F1. Numerical evidence is in Table 4 and §4.1–§4.3.

Possible explanation	Intervention	Result	Verdict
Miscalibrated soft targets	Post-hoc temperature scaling of ECG-FM	Student ECE drops vs default KD; no clear AUROC/F1 gain vs no-KD	Not sufficient
Under-tuned ECG-FM	LR/WD/warmup sweep; exploratory composed KD recipe	Tuned-teacher AUROC CI spans zero vs no-KD; composed means show no clear gain	No rescue shown
Teacher size	Same-family supervised ResNet ladder (3–17.5 M)	AUROC point estimates are close; F1 is non-monotone	No monotonic effect observed
Architecture family	Cross-family matched-capacity ECGTransformer (4.8 M)	Negative student AUROC and F1 Δ CIs exclude zero	Not sufficient in this control
Native target-task fit	Compare native validation AUROC with student outcomes	Higher-fit teachers group with more favorable KD outcomes here	Descriptive association

is recipe-specific, not architectural. Distilling at the default temperature therefore trains the student on effectively hard labels plus a high-confidence noise channel, consistent with student ECE more than doubling.

We tested H1 by post-hoc temperature-scaling ECG-FM's logits on the validation split before distillation ("ECG-FM, T-scaled" in Table 4). Softening the teacher reduces student ECE (-0.027 vs default ECG-FM, Table 4) but does not recover ranking or F1 (ΔAUROC -0.0018, ΔF1 $+0.0057$; both CIs span zero). Figure 3 in the appendix shows the pattern across all ECG-FM variants: successive interventions drive student ECE back toward the no-KD baseline without a consistent F1 gain.

4.2 Teacher retuning does not rescue the result

We next tested whether stronger ECG-FM fine-tuning or a better KD recipe rescues the result. We swept learning rate, weight decay, and warmup, picked the best-of-sweep teacher by validation AUROC ("ECG-FM, tuned": native PTB-XL val AUROC 0.919 vs 0.896 default, Table 8), and

Table 4: Teacher-choice evidence (PTB-XL, 3 seeds per row). Same student and test split throughout. Default KD recipe (T=4, α=0.5) on every row except "ECG-FM, best-case," which uses a recipe selected post hoc by single-seed test F1 (T=4, α=0.9). "T-scaled" and "tuned" modify only the teacher (post-hoc temperature scaling and LR/WD/warmup sweep, respectively). Point AUROC/F1 rows are per-seed mean±std; ECE is a per-seed mean. Except for the best-case row, Δ columns are paired-bootstrap means vs no-KD over seed-averaged logits (§2); † marks 95% CIs excluding zero. Best-case Δs are differences of per-seed means and have no paired-bootstrap CI.

Teacher condition (ResNet1D student)	AUROC	F1	ECE	ΔAUROC	ΔF1
No-KD baseline	$0.9278_{\pm.0011}$	$0.7325_{\pm.0113}$	0.0291	–	–
ECG-FM, default	$0.9263_{\pm.0031}$	$0.7392_{\pm.0078}$	0.0776	−0.0012	−0.0014
ECG-FM, T-scaled	$0.9257_{\pm.0030}$	$0.7398_{\pm.0034}$	0.0508	−0.0018	+0.0057
ECG-FM, tuned	$0.9271_{\pm.0008}$	$0.7254_{\pm.0106}$	0.0331	−0.0014	$-0.0147^{\dagger}$
ECG-FM, best-case	$0.9287_{\pm.0034}$	$0.7373_{\pm.0052}$	0.0283	+0.0009	+0.0048
ECGFounder (pretrained)	$\mathbf{0.9341}_{\pm.0002}$	$\mathbf{0.7539}_{\pm.0059}$	**0.0278**	$+0.0045^{\dagger}$	$+0.0145^{\dagger}$
Supervised ResNet, 3 M	$0.9307_{\pm.0003}$	$0.7460_{\pm.0031}$	0.0302	+0.0005	+0.0069
Supervised ResNet, 7.8 M	$0.9298_{\pm.0010}$	$0.7488_{\pm.0059}$	0.0317	−0.0004	$+0.0115^{\dagger}$
Supervised ResNet, 17.5 M	$0.9311_{\pm.0004}$	$0.7362_{\pm.0087}$	0.0355	+0.0005	−0.0022
ECGTransformer, 5 M	$0.9236_{\pm.0003}$	$0.7020_{\pm.0102}$	0.0462	$-0.0065^{\dagger}$	$-0.0382^{\dagger}$

distilled with the default KD recipe. The tuned teacher improves native AUROC, but the student's F1 drops (ΔF1 $-0.0147^{\dagger}$), and AUROC is indistinguishable from no-KD (ΔAUROC -0.0014, CI spans zero).

As an exploratory stress test, we composed the sweep-best ECG-FM teacher with T=4, α=0.9, the recipe selected *post hoc* for the highest single-seed *test* F1 in the default-teacher grid (Appendix F). Three-seed per-seed means put its AUROC at 0.9287 versus 0.9278 for no-KD, and F1 at 0.7373 versus 0.7325; the corresponding differences of +0.0009 and +0.0048 do not demonstrate a clear improvement. Student ECE is similar to no-KD (0.0283 versus 0.0291). No paired-bootstrap comparison was saved for this condition, and test-set recipe selection prevents it from serving as independent confirmation of the teacher-fit pattern. The default-recipe tuned-teacher comparison remains the controlled retuning test; the supervised teacher also provides a reference without an analogous rescue sweep.

4.3 Teacher Size and Architecture Family Do Not Individually Explain the Gap

We test whether the gap is a size or family artifact. Holding the student, pipeline, and KD recipe fixed on PTB-XL, we distil the same ResNet1D student from a same-family supervised ladder (3.2 M / 7.8 M / 17.5 M ResNet) and from a matched-capacity cross-family 4.8 M ECGTransformer teacher. The three supervised-ResNet student AUROCs are close across the 6× teacher-size range; the saved paired comparisons of the 3.2 M and 17.5 M students with no-KD have AUROC CIs spanning zero. F1 is non-monotone in teacher size, and the cross-family Transformer teacher *degrades* the student (ΔAUROC $-0.0065^{\dagger}$, ΔF1 $-0.0382^{\dagger}$). In this teacher pool, measured native PTB-XL validation AUROC also groups the weaker candidates (ECG-FM default and tuned; ECGTransformer) apart from ECGFounder and the supervised ResNets (Table 8, Figure 2). The composed ECG-FM condition shares the tuned teacher's native score but uses a different, test-selected KD recipe. Native fit groups the weaker candidates in this pool but does not exactly order the stronger teachers. ECGFounder at 0.929 yields the highest per-seed mean student AUROC in this PTB-XL set.

5 Native Target-Task Fit as a Teacher-Selection Signal

In the PTB-XL teacher pool, ECG-FM's default and tuned checkpoints and the ECGTransformer have lower measured native validation AUROC than ECGFounder and the three supervised ResNets (Table 8). Their student KD outcomes are also less favorable under the tested pipeline (Table 4). The two public foundation teachers make the point sharply: foundation-model status does not separate

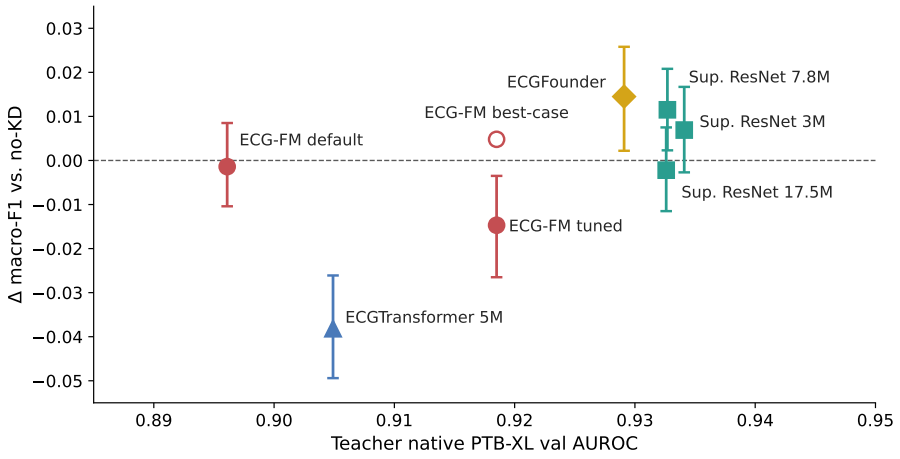


Figure 2: **Measured native target-task fit and student KD outcome in the tested PTB-XL pool.** Each point is a selected condition from Table 4: x is teacher validation AUROC and y is student macro-F1 difference from no-KD. Filled markers show paired bootstrap means with 95% CIs over test examples conditional on seed-averaged logits. The open best-case marker shows only the difference of per-seed mean F1 values and has no paired CI; its recipe was selected using test F1. The T-scaled condition is omitted because it changes soft-target calibration, not the underlying teacher's native AUROC measurement. The higher-fit teachers group with more favorable KD outcomes, while native AUROC does not exactly order the stronger teachers.

the positive and negative student-transfer outcomes, whereas their measured downstream fit differs substantially. The supervised size ladder supplies no monotone student gain with teacher scale.

Native target-task fit thus separates the weakest candidates in these comparisons, but it does not resolve differences among the higher-fit teachers. The tuned and exploratory composed ECG-FM conditions share a teacher checkpoint but use different KD recipes. The retained records lack three-seed no-KD validation metrics and complete grid validation results, so they do not support a prospective numerical accept/reject cutoff.

These results suggest evaluating candidate teachers on the downstream task before committing to distillation. Student-side evaluation should then compare that KD recipe directly with a no-KD baseline on an independent test set.

6 Discussion and Limitations

ECG-FM provides no robust student gain on either dataset; ECGFounder yields PTB-XL gains and outperforms ECG-FM on PhysioNet under the same student-side KD pipeline. This study covers one logit-based KD family, two multi-label ECG datasets, and one compact student. In this pool, native target-task fit aligns with student outcomes more clearly than foundation-model status or size. Calibration correction, ECG-FM retuning, a supervised size ladder, and a cross-family teacher do not individually explain the gap. The composed ECG-FM recipe, selected using single-seed test F1, is exploratory. We did not vary no-KD recipes such as label smoothing, focal loss, or MixUp-style augmentation, which could change the ordering. Native fit combines effects of labels, representations, calibration, optimization, and preprocessing; the tested interventions do not isolate a causal mechanism. Generality across more teachers, KD objectives, and task families remains untested. ECGFounder has a small PhysioNet paired-estimator ECE penalty versus no-KD (Table 2); deployment calibration remains a separate consideration.

ECGFounder implementation caveats. We use the per-sample z-score renormalisation expected by ECGFounder's upstream code. ECG-FM received an LR/WD/warmup sweep and a $T \times \alpha$ KD-recipe ablation; ECGFounder did not. The sweep tested H2 (ECG-FM was under-tuned), which applies to the teacher that fails under the default recipe. This creates asymmetric tuning: more

ECG-FM configurations were explored, while ECGFounder used the default recipe. A matched ECGFounder sweep could change its measured KD outcome; our comparison supports only the configurations tested.

Cross-teacher input normalization. ECGFounder uses its upstream per-sample z-score; ECG-FM and the supervised teachers use the dataloader's per-lead z-score. Each pretrained teacher thus sees its expected input distribution, but without a normalization-swap ablation we cannot quantify the KD effect. A common-normalization comparison remains untested.

Pretraining-corpus overlap. PhysioNet/CinC 2021 is a second target task and label-space stress test, not a pretraining-disjoint benchmark: ECG-FM's public pretraining sources may overlap PhysioNet 2021 constituent databases. We are not aware of PTB-XL or PhysioNet overlap in ECG-Founder's pretraining cohort. Such overlap would *advantage* ECG-FM, so it cannot explain its PhysioNet underperformance. Independent dataset-overlap audits would strengthen future evaluations.

Seed count and uncertainty. We use three training seeds per condition. Mean$\pm$std rows in Tables 1, 2, and 4 report training-seed variance; paired-bootstrap CIs condition on seed-averaged logits and quantify test-example uncertainty for matched contrasts, not training-seed uncertainty. The exploratory best-case row has no paired CI. More seeds and a nested bootstrap are natural follow-ups.

7 Conclusion

Distillation utility depends on the target task and student. ECG-FM provides no robust gain on either dataset; ECGFounder yields PTB-XL AUROC, mAP, and F1 gains with paired CIs excluding zero and outperforms ECG-FM on those metrics on PhysioNet. Calibration correction and further ECG-FM tuning do not rescue the tested configurations; neither the supervised size ladder nor cross-family comparison resolves the gap. Measured downstream fit distinguishes useful from ineffective teachers in this pool more clearly than foundation-model status or scale. Teacher choice should be checked by comparing the KD student directly with a no-KD baseline. More broadly, foundation-model teacher selection should be treated as a task- and student-conditioned transfer problem rather than as a by-product of foundation-model scale or standalone performance.

AI use statement

ChatGPT and OpenAI Codex assisted with literature discovery and synthesis, manuscript drafting and editing, and methodological and interpretive review. Codex inspected repository sources, checked claims against retained results, and edited LaTeX and bibliography files. We verified AI-assisted citations against original publications and interpretations against experimental artifacts. Generative AI was not used to generate or modify experimental data or experimental results. The authors made all final methodological, interpretive, and editorial decisions and take responsibility for the paper.

REFERENCES

M A Al-Masud, Juan Miguel Lopez Alcaraz, and Nils Strodthoff. Benchmarking ECG FMs: A reality check across clinical tasks. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=xXRqWpt3Xr.

Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. Knowledge distillation: A good teacher is patient and consistent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10925–10934, 2022.

Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4794–4802, 2019.

Chengyu Dong, Liyuan Liu, and Jingbo Shang. Toward student-oriented teacher network training for knowledge distillation. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/pdf?id=wsWGcw6qKD.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 1321–1330, 2017.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023. doi: 10.1038/s41597-022-01899-x.

Jun Li, Aaron D. Aguirre, Valdery Moura, Jiarui Jin, Che Liu, Lanhai Zhong, Chenxi Sun, Gari Clifford, M. Brandon Westover, and Shenda Hong. An electrocardiogram foundation model built on over 10 million recordings with external evaluation across multiple domains. *NEJM AI*, 2(7), 2025. doi: 10.1056/aioa2401033. Preprint: arXiv:2410.04133 (2024).

Riccardo Lunelli, Angus Nicolson, Samuel Martin Pröll, Sebastian Johannes Reinstadler, Axel Bauer, and Clemens Dlaska. BenchECG and xECG: a benchmark and baseline for ECG foundation models. *npj Digital Medicine*, 2026. doi: 10.1038/s41746-026-03196-y. URL https://doi.org/10.1038/s41746-026-03196-y.

Kaden McKeen, Laura Oliva, Sameer Masood, Augustin Toma, Barry Rubin, and Bo Wang. ECG-FM: An open electrocardiogram foundation model. *JAMIA Open*, 2025. doi: 10.1093/jamiaopen/ooaf122. Preprint: arXiv:2408.05178 (2024).

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5191–5198, 2020.

Abhishek Panigrahi, Bingbin Liu, Sadhika Malladi, Andrej Risteski, and Surbhi Goel. Progressive distillation induces an implicit curriculum. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/pdf?id=wPMRwmytZe.

Abhishek Panigrahi, Bingbin Liu, Sadhika Malladi, Sham Kakade, and Surbhi Goel. In good GRACES: Principled teacher selection for knowledge distillation. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/pdf?id=m276fke38H.

Matthew A. Reyna, Nadi Sadr, Erick A. Perez Alday, Annie Gu, Amit J. Shah, Chad Robichaux, Ali Bahrami Rad, Andoni Elola, Salman Seyedi, Sardar Ansari, Hamid Ghanbari, Qiao Li, Ashish Sharma, and Gari D. Clifford. Will two do? Varying dimensions in electrocardiography: The PhysioNet/Computing in Cardiology Challenge 2021. *Computing in Cardiology*, 48:1–4, 2021. doi: 10.23919/CinC53138.2021.9662687.

Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A. Alemi, and Andrew G. Wilson. Does knowledge distillation really work? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 6906–6919, 2021.

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.

Chaofei Wang, Qisen Yang, Rui Huang, Shiji Song, and Gao Huang. Efficient knowledge distillation from model checkpoints. In *Advances in Neural Information Processing Systems*, volume 35, pp. 607–619, 2022. doi: 10.52202/068431-0044. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/03e0712bf85ebe7cec4f1a7fc53216c9-Abstract.html.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. LogME: Practical assessment of pre-trained models for transfer learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12133–12143. PMLR, 2021. URL https://proceedings.mlr.press/v139/you21b.html.

Weiyang Zhang, Xinyang Chen, Xiucheng Li, Kehai Chen, Weili Guan, and Liqiang Nie. Unified transferability metrics for time series foundation models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 41155–41178, 2025. doi: 10.52202/085713-1374. URL https://papers.neurips.cc/paper_files/paper/2025/hash/3adfe6e3d8da4a75d174f466e1efc039-Abstract-Conference.html.

A Reproducibility statement

All experiments use only public data: PTB-XL (Wagner et al., 2020) and PhysioNet/CinC 2021 (Reyna et al., 2021). Splits are deterministic (stratified $80/10/10$ for PhysioNet, using the first scored SNOMED Dx code in each record's header as the stratification class so that all 30 classes appear in each split; the standard fold-1–8 / fold-9 / fold-10 partition for PTB-XL). The first-Dx convention is a class-coverage device, not a clinical-priority claim; the stratifier balances overall class representation rather than guaranteeing balanced minority-label distributions across splits. The two foundation-teacher checkpoints are the public MIMIC-IV-ECG-finetuned release of ECG-FM (McKeen et al., 2025) and the public 12-lead release of ECGFounder (Li et al., 2025) on HuggingFace (`PKUDigitalHealth/ECGFounder`, MIT license). An accompanying code release contains preprocessing, model, KD, training, and evaluation code, orchestration scripts, `requirements.txt`, and `docs/REPRODUCIBILITY.md` with environment and run-order instructions. Headline student and KD runs use seeds $\{0, 1, 2\}$. The retained single-seed KD-grid launcher defaults to seed 0, but its invocation log is not retained, so the executed grid seed is unverified. Raw data is downloaded by the supplied scripts from the original dataset hosts; no human-subject data is collected by us.

PTB-XL preprocessing. 12-lead waveforms are loaded with `wfdb` at the native 500 Hz, no resampling. Each record contributes a single 10 s window (12×5000 samples). Diagnostic SCP codes are mapped to the five superclasses (NORM, MI, STTC, CD, HYP) using the `diagnostic_class` column of `scp_statements.csv`; records with no superclass label are dropped. Per-lead z-score normalisation: mean and std are computed on the training fold (folds 1–8) only and applied to validation and test. No beat segmentation or band-pass filtering is applied beyond what PTB-XL already ships. Teacher inputs are taken from the same dataloader, so the student and supervised teachers see identical preprocessing; ECGFounder additionally applies the per-sample renormalisation described in Appendix E (see §6 for discussion).

B Model inventory

Tables 5 and 6 summarise the models used in the controlled audit; the following appendices give the implementation details needed for reproduction (students in Appendix D, teachers in Appendix E).

Table 5: Main student and teacher models (Part 1 of 2). Params are for PTB-XL (C=5); PhysioNet (C=30) head adds a small number of parameters (e.g. the ResNet1D student grows from 1,220,485 to 1,226,910). All teacher models are adapted to the target dataset before their logits are used as soft targets: pretrained teachers are fine-tuned, while supervised teachers are trained from scratch.

Role	Model	Params	Backbone / family	Initialization / pretraining	Use in this study
Compact student	CompactResNet1D	1.22 M	1D ResNet	Trained from scratch on target dataset	Deployed student in every no-KD and KD condition; fixed across all teachers.
Foundation teacher	ECG-FM (McKeen et al., 2025)	90.4 M	Transformer / wav2vec 2.0-style ECG encoder	Public MIMIC-IV-ECG-finetuned checkpoint; head replaced and fine-tuned on target dataset	Main KD teacher (§3); calibration, tuning, and best-case rescue conditions (§4). Low native PTB-XL fit in this study; not rescued by the tested interventions.
Foundation teacher	ECGFounder (Li et al., 2025)	30.7 M	Net1D / RegNet-1D-style CNN	Public 12-lead checkpoint, supervised-pretrained on ∼10 M-record ECG cohort; head replaced and fine-tuned on target dataset	Second foundation-teacher comparison (§3); positive KD on PTB-XL, separates from ECG-FM on PhysioNet.
Supervised teacher	ResNet-large	7.8 M	1D ResNet (wider CompactResNet1D)	Trained from scratch on target dataset	Supervised-teacher reference (§3).

Table 6: Control and auxiliary models (Part 2 of 2). Size- and family-control teachers used in the rescue tests, plus two auxiliary students included only as sanity checks.

Role	Model	Params	Backbone / family	Initialization / pretraining	Use in this study
Size-control teachers	ResNet-small / large / xlarge	3.2 / 7.8 / 17.5 M	1D ResNet (same family as student)	Trained from scratch on target dataset	Same-family capacity ladder (§4.3); probes whether teacher size alone would order student KD gain.
Family-control teacher	ECGTransformer	4.8 M	Transformer (patch-embedded)	Trained from scratch on target dataset	Cross-family matched-capacity teacher (§4.3); probes whether architecture family alone would explain the result.
Auxiliary student	SimpleCNN1D	0.56 M	1D CNN	Trained from scratch on target dataset	Sanity check only (Appendix D); not used in the headline results.
Auxiliary student	ECGTransformer (student)	0.44 M	Transformer (patch-embedded)	Trained from scratch on target dataset	Sanity check only (Appendix D); not used in the headline results.

C ADDITIONAL TEACHER DIAGNOSTICS

Table 7: Teacher soft-target diagnostics (PTB-XL test split). Mean binary entropy $H(p)$ over (sample, class) pairs; mean per-sample max-probability; "confident-wrong" rate (fraction of pairs with $p>0.9$ against a true 0 or $p<0.1$ against a true 1); and "uncertain" rate (fraction of pairs with $p \in [0.3, 0.7]$). $H(p)$ is numerically undefined for ECG-FM default because the sigmoid saturates to floating-point zero on at least one class. Confident-wrong rates are pooled over all $5 \times N$ (PTB-XL) sample-class pairs and so blend across class prevalences; under PTB-XL's mostly-negative-per-record label structure most pairs are true-0, so high-prevalence classes weight the pooled rate more heavily. We use the metric only for relative comparison across teachers under the same labels, not as an absolute calibration score.

Teacher	$H(p)$	max p	conf. wrong	uncertain
ECG-FM, default	– (sat.)	0.9873	0.1067	0.0157
ECG-FM, tuned	0.2199	0.8700	0.0361	0.0975
ECGFounder	0.2299	0.8561	0.0252	0.1048
Supervised ResNet, 7.8 M	0.2237	0.8865	0.0283	0.1038
ECGTransformer, 5 M	0.2092	0.8293	0.0446	0.0973

Table 8: Teacher native PTB-XL performance (validation split). Each row is the best-validation checkpoint of the teacher used to produce soft targets for the KD rows in Table 4. Native AUROC groups weaker and stronger student outcomes descriptively in this pool, but it does not establish a predictive cutoff or an exact ranking within the higher-fit group.

Teacher	Params	Val AUROC	Val F1
ECG-FM, default	90.4 M	0.8961	0.7193
ECG-FM, tuned (low-lr)	90.4 M	0.9185	0.6961
ECGTransformer, 5 M	4.8 M	0.9049	0.6713
ECGFounder (pretrained)	30.7 M	0.9291	0.7284
Supervised ResNet, 3 M	3.2 M	**0.9341**	0.7055
Supervised ResNet, 7.8 M	7.8 M	0.9327	**0.7324**
Supervised ResNet, 17.5 M	17.5 M	0.9326	0.7120

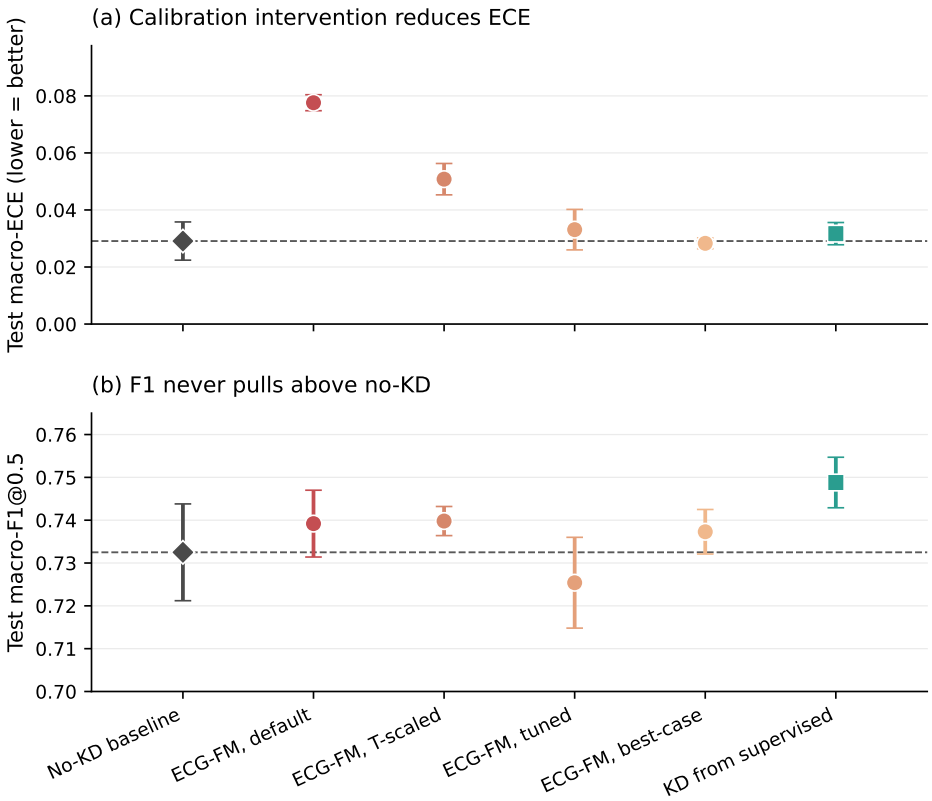


Figure 3: **The calibration gap is recipe-fixable; the F1 gap is not.** PTB-XL test metrics across the no-KD baseline, four ECG-FM KD conditions of increasing effort, and the supervised-teacher KD reference. Error bars are ±1 seed-std; the dashed line is the no-KD baseline. The exploratory composed condition has per-seed mean ECE near no-KD; F1 changes across ECG-FM conditions do not show a clear and consistent gain. Its recipe was selected using test F1, and no paired CI was saved for that condition.

D FULL STUDENT ARCHITECTURE DETAILS

CompactResNet1D (1.22 M, used in main paper). 12-lead ECG input ($B \times 12 \times 5000$). Stem: `Conv1d` (k=15, stride 2, padding 7, 12→64) → BN → ReLU. Body: four `_ResBlock1D` blocks at channel widths $[64, 128, 128, 256]$, kernel size 7, dropout 0.1. Block 1 keeps temporal resolution; blocks 2–4 each apply stride-2 downsampling, so the sequence dimension reduces from 5000 to 5000/16 ≈ 312 tokens. Each residual block is `Conv-BN-ReLU-Dropout-Conv-BN`, with a 1×1 convolutional shortcut whenever the channel count or the stride changes. Head: `AdaptiveAvgPool1d(1)` → `Linear`(256→C). Total trainable parameters: 1,220,485 for C=5 (PTB-XL); 1,226,910 for C=30 (PhysioNet). Weights are initialised with Kaiming-normal for convolutions, ones/zeros for batch-norm γ/β, and Xavier-normal for the linear head.

Auxiliary student – SimpleCNN1D (0.56 M). Same stem as above. Three `Conv-BN-ReLU-MaxPool(2)` stages with kernel sizes $[7, 5, 5]$ and channel widths $[128, 256, 256]$. `AdaptiveAvgPool1d(1)` → `Dropout(0.1)` → `Linear`(256→C). No residual connections; included as a deliberately weaker inductive prior.

Auxiliary student – ECGTransformer (0.44 M). Patch embedding: `Conv1d`(12→128, k=25, stride 25, no overlap), so the 5000-sample input becomes 200 patch tokens of width 128. Fixed sinusoidal positional encoding. Transformer encoder: 3 pre-norm layers, 4 attention heads, FFN width 256, dropout 0.1. `AdaptiveAvgPool1d(1)` over the token dimension → `Linear`(128→C).

The ResNet student is the only one used in the headline experiments (§3 and §4); the SimpleCNN1D and ECGTransformer students above appear only as cross-architecture sanity checks and are not used in the headline tables. The "ECGTransformer, 5 M" row in Table 4 refers to the separate, larger ECGTransformer *teacher* described in Appendix E (used as the cross-family matched-capacity teacher in §4.3), not to the 0.44 M auxiliary Transformer *student* above.

E FULL TEACHER DETAILS

ECG-FM integration. We re-implement ECG-FM (McKeen et al., 2025) as a pure-PyTorch `nn.Module` (no fairseq dependency) in `src/models/teacher.py`, restoring the published wav2vec 2.0-style encoder layer-by-layer: a 7-block convolutional feature extractor that maps the raw 12-lead waveform to a 768-dim token sequence, a grouped 1D convolutional positional encoding (`kernel_size=128, groups=16`), and a 12-layer Transformer encoder with 12 heads and a 3072-dim FFN per layer. We instantiate this encoder, load pretrained weights from the authors' MIMIC-IV-ECG-finetuned checkpoint (`mimic_iv_ecg_finetuned.pt`), discard the original 17-class head, and attach a fresh `Linear(768→`C`)` head for the downstream task. Total: 90.38 M parameters (encoder fully unfrozen during fine-tuning). We use this MIMIC-IV-ECG-finetuned release because it is the public artifact ECG-FM's authors publish for downstream use: a practitioner adopting ECG-FM as a foundation teacher gets this checkpoint by default. Starting from a purely self-supervised pretrained checkpoint and fine-tuning under our protocol is a natural follow-up we did not run.

Default ECG-FM fine-tuning recipe. AdamW (lr 1×10^{-4}, weight decay 10^{-4}), 3-epoch linear-probe warm-up with the encoder frozen, then full fine-tuning with cosine annealing for up to 30 epochs and patience 10 on validation F1. Batch size 32 (the `scripts/train_teacher.py` default, distinct from the batch-64 student/KD-training protocol of §2). The default-recipe checkpoint reaches val AUROC 0.8961 and val F1 0.7193 at epoch 19 on PTB-XL.

ECG-FM fine-tuning sweep (Phase 3.2). To preempt "the teacher was under-tuned" as an explanation for the KD failure, we sweep four variants of the recipe above. Best-of-sweep on validation AUROC is the "low-lr" variant, which we use throughout §4 as "ECG-FM, tuned."

Table 9: ECG-FM fine-tuning sweep on PTB-XL. Each row is a different choice of learning rate, weight decay, warm-up length and total epochs; all other hyperparameters are held fixed. Best-of-sweep by val AUROC is **low-lr**.

Config	lr	wd	warmup	epochs	val AUROC	val F1
default	1×10^{-4}	1×10^{-4}	3	30	0.8961	0.7193
low-lr	5×10^{-5}	1×10^{-4}	3	30	**0.9185**	0.6961
low-lr + high-wd	5×10^{-5}	1×10^{-3}	3	30	0.9168	0.7054
high-wd	1×10^{-4}	1×10^{-3}	3	30	0.7232	0.1232
long-warmup	1×10^{-4}	1×10^{-4}	8	35	0.7362	0.0000

The two degenerate rows (high-wd, long-warmup) collapse to near-trivial F1; we treat them as failed sweep points and select the best non-degenerate configuration.

ECGFounder integration. We load the pretrained 12-lead ECGFounder (Li et al., 2025) from the public HuggingFace release (`PKUDigitalHealth/ECGFounder`, file `12_lead_ECGFounder.pth`; MIT license). The upstream architecture is `Net1D`, a RegNet-1D-style CNN with Squeeze-and-Excitation blocks and Swish activations, instantiated with the canonical constructor arguments (`base_filters=64, filter_list=[64, 160, 160, 400, 400, 1024, 1024], m_blocks_list=[2, 2, 2, 3, 3, 4, 4], kernel_size=16, stride=2, groups_width=16, use_bn=False, use_do=False`) taken verbatim from the upstream `ft_12lead_ECGFounder` helper. We drop the 150-class pretraining head, attach a fresh `Linear(`$1024 \to C$`)` head for the downstream task, and wrap the backbone in `src/models/teacher_ecgfounder.py` with the same interface as our ECG-FM classifier. Total: 30.67 M parameters (backbone fully unfrozen during fine-tuning). A

forward-pass renormalization step applies the global per-sample z-score expected by ECGFounder's upstream fine-tune pipeline, overriding our dataloader's per-lead normalization so the backbone sees the distribution it was trained on. The default-recipe checkpoint (same recipe as ECG-FM default) reaches val AUROC 0.9291 and val F1 0.7284 at epoch 5 on PTB-XL, and val AUROC 0.9709 at epoch 7 on PhysioNet. We do not run a fine-tuning sweep for ECGFounder in this submission — the default recipe already places it in the high-native-fit band that, in our experiments, contains the useful teachers — but a sweep is a natural follow-up if reviewers request matched treatment to ECG-FM.

Supervised teacher ladder. The supervised teachers are built from the same `CompactResNet1D` class as the student, only with wider channel lists. All three are trained from scratch on the target dataset under the student training protocol of §2 (AdamW lr 10^{-3}, wd 10^{-4}, cosine, batch 64, up to 60 epochs, patience 15 on val AUROC).

Table 10: Supervised teacher ladder used in §4.3. `CompactResNet1D` channel widths in brackets; depth fixed at 4 residual blocks; kernel size 7 throughout.

Tag	Channels	Params
`resnet_small`	$[96, 192, 256, 384]$	3.2 M
`resnet_large`	$[128, 256, 512, 512]$	7.8 M
`resnet_xlarge`	$[192, 384, 768, 768]$	17.5 M

Cross-family teacher (Transformer). The cross-family teacher in §4.3 is built from the same `ECGTransformer` class as the auxiliary student, only wider and deeper: d_{model}=256, h=8 heads, 6 layers, FFN width 1024, patch size 10 (so the token count is $5000/10 = 500$ rather than 200). Total: 4.8 M parameters. Trained under the student protocol on the target dataset.

F Full KD recipe ablation

We ran a $T \times \alpha$ grid using the default ECG-FM teacher (no temperature-scaling, no fine-tuning sweep), the ResNet1D student of §2, and a single training seed. The retained launcher defaults to seed 0, but the invocation log is unavailable, so the executed seed cannot be verified. PTB-XL *test*-set metrics are in Table 11. AUROC ranges from 0.9210 to 0.9315; the grid shows no large positive AUROC gain over the no-KD baseline. ECE generally declines as α increases for $T \in \{2, 4, 8\}$, but not monotonically at T=1. The highest single-seed *test* F1 cell, T=4, α=0.9 (F1 0.7524), was selected post hoc for the three-seed composed ECG-FM stress test in §4.2. Its reuse of the test split makes that comparison exploratory rather than independent confirmation.

G Per-class AUROC

Per-class test AUROC for the four headline conditions on each dataset is computed on seed-averaged logits across the three training seeds. The tables show that the macro-level pattern is not confined to one class, while individual effects vary in size and direction. For example, ECGFounder KD improves four of five PTB-XL classes relative to no-KD but is lower on HYP; the PhysioNet class effects are also heterogeneous.

The PhysioNet per-class pattern is consistent with the macro story: KD from ECGFounder matches or improves the baseline on 21 of 30 classes while ECG-FM KD is below baseline on 20 of 30. The macro pattern is not confined to one class, but the effects are uneven: under ECG-FM KD, SA, IRBBB, LQRSV, BBB, and LBBB have the largest absolute drops vs no-KD, while PVC, IAVB, AFL, and NSIVCB improve. ECGFounder's largest gains include PR, PVC, PAC, RAD, and NSIVCB.

H Training and learning curves

Early stopping (PTB-XL). All student conditions use the same patience-15 rule on validation AUROC, but the rule can select different checkpoints and produce different realized training durations. Thus the teacher comparison holds the optimization and selection *protocol* fixed, not the number of

Table 11: KD recipe ablation on the PTB-XL test set (default ECG-FM teacher, ResNet1D student, single seed). Bold: per-column maximum (AUROC, F1) or minimum (ECE). The highest test-F1 cell (T=4, α=0.9) was selected post hoc for the exploratory composed row in Table 4.

T	α	AUROC	F1	ECE	Note
1	0.3	0.9301	0.7478	0.0332	
1	0.5	0.9305	0.7439	0.0382	
1	0.7	**0.9315**	0.7469	**0.0329**	
1	0.9	0.9306	0.7456	0.0347	
2	0.3	0.9301	0.7451	0.0601	
2	0.5	0.9301	0.7424	0.0596	
2	0.7	0.9304	0.7457	0.0507	
2	0.9	0.9308	0.7425	0.0406	
4	0.0	0.9277	0.7453	0.0793	
4	0.3	0.9286	0.7471	0.0789	
4	0.5	0.9294	0.7482	0.0746	(Tab. 1 default)
4	0.7	0.9284	0.7421	0.0735	
4	0.9	0.9295	**0.7524**	0.0564	(best-case recipe)
8	0.3	0.9210	0.7296	0.0882	
8	0.5	0.9229	0.7277	0.0835	
8	0.7	0.9265	0.7445	0.0813	
8	0.9	0.9298	0.7463	0.0652	

Table 12: PTB-XL per-class test AUROC (3-seed-averaged logits).

Class	No-KD	KD ECG-FM	KD ECGFounder	KD sup.
NORM	0.9510	0.9499	**0.9517**	0.9508
MI	0.9396	0.9370	**0.9435**	0.9407
STTC	0.9348	0.9339	**0.9421**	0.9350
CD	0.9224	0.9246	**0.9354**	0.9269
HYP	**0.9112**	0.9079	0.9088	0.9038
macro	0.9318	0.9307	**0.9363**	0.9315

updates. The repository does not retain the per-epoch histories for all headline conditions, so we do not infer that their learning trajectories overlap or attribute the entire outcome gap solely to soft-target content.

Subset learning curve. Figure 4 plots no-KD test AUROC across three student architectures trained on $\{25, 50, 75, 100\}$ % of the PTB-XL training fold. Slopes are roughly linear and parallel; the ResNet1D dominates the CNN and Transformer at every fraction.

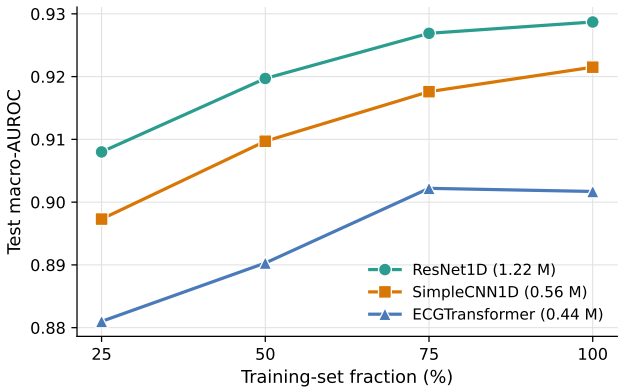


Figure 4: No-KD test macro-AUROC vs. training-set fraction on PTB-XL, for the three student architectures.

Table 13: PhysioNet/CinC 2021 per-class test AUROC (3-seed-averaged logits). Class abbreviations follow the official 30-class scored set; "macro" is the unweighted mean of the 30 per-class AUROCs. Columns: no-KD baseline; KD from ECG-FM; KD from ECGFounder; KD from supervised.

Class	No-KD	ECG-FM	ECGF.	sup.	Class	No-KD	ECG-FM	ECGF.	sup.
AF	0.9543	0.9525	**0.9557**	0.9533	PR	0.9953	0.9952	**0.9984**	0.9947
AFL	0.9799	**0.9814**	0.9812	0.9795	PRWP	0.9803	0.9735	**0.9818**	0.9803
BBB	**0.9843**	0.9750	0.9831	0.9834	PVC	0.9174	0.9325	**0.9505**	0.9179
Brady	**0.9911**	0.9888	0.9904	0.9888	LPR	**0.9811**	0.9762	0.9724	0.9759
CLBBB	**0.9983**	0.9973	0.9963	0.9976	LQT	0.9583	0.9601	**0.9666**	0.9594
CRBBB	0.9854	0.9863	0.9836	**0.9865**	QAb	0.9720	0.9728	**0.9752**	0.9717
IAVB	0.9878	0.9904	**0.9919**	0.9873	RAD	0.9881	0.9869	**0.9904**	0.9871
IRBBB	**0.9653**	0.9545	**0.9653**	0.9619	RBBB	0.9902	0.9876	0.9898	**0.9905**
LAD	0.9843	0.9819	**0.9863**	0.9832	SA	0.8606	0.8035	0.8534	**0.8622**
LAnFB	0.9886	0.9874	**0.9908**	0.9902	SB	0.9970	**0.9979**	0.9975	0.9972
LBBB	0.9861	0.9816	**0.9880**	0.9861	STach	0.9957	0.9951	**0.9964**	0.9952
LQRSV	**0.9830**	0.9606	0.9681	0.9824	SVPB	**0.9214**	0.9101	0.8978	0.8992
NSIVCB	0.8929	0.8981	**0.9110**	0.9003	TAb	0.9268	0.9223	0.9272	**0.9278**
NSR	0.9831	**0.9840**	0.9836	0.9834	TInv	0.9234	0.9216	**0.9253**	0.9216
PAC	0.8791	0.8685	**0.8880**	0.8859	VPB	0.9037	0.9167	**0.9223**	0.9043

macro: no-KD 0.9618 KD ECG-FM 0.9580 KD ECGFounder **0.9636** KD sup. 0.9612

The retained subset-learning-curve summary above concerns no-KD students at different training-set fractions. It does not compare the per-epoch trajectories of ECGFounder KD and no-KD. Those trajectories would require the archived run histories.

I Latency and deployment profile

Teachers are train-time only; only the student is deployed. The teacher latencies below are reported as context for the parameter-count and compute differences between conditions, not as a deployment constraint. We benchmark CPU latency on a single Intel CPU thread (mean±std over 100 forward passes per model after a 5-pass warm-up; one 12-lead 5,000-sample input per pass; FP32, no batching). MACs are computed with `thop` on the same 12-lead, 5000-sample input, also FP32. The ResNet1D student has $\sim$74$\times$ fewer parameters than the ECG-FM teacher and runs $\sim$32$\times$ faster in single-thread CPU inference; the Transformer student is $\sim$124$\times$ faster. ECGFounder, at 30.7 M parameters, sits between the two foundation baselines in size and latency (FP32 parameter memory $\sim$122.7 MB, $\sim$27 ms/record at 1 thread).

Table 14: Single-sample CPU inference profile. Latency is single-thread FP32, mean±std over 100 runs (20 for ECG-FM due to throughput; 30 for ECGFounder). "FP32 params (MB)" is the raw parameter memory (params $\times$ 4 bytes); it is *not* the on-disk checkpoint size, which for several models additionally serialises Adam optimizer state and is correspondingly larger. GMACs are reported on the same input shape.

Model	Params	FP32 params (MB)	CPU latency (ms)	GMACs
Teacher: ECG-FM	90.4 M	361.5	129.66 ± 0.82	28.36
Teacher: ECGFounder (12-lead)	30.7 M	122.7	26.71 ± 0.12	1.17
Teacher: supervised ResNet (7.8 M)	7.8 M	31.2	17.59 ± 0.20	4.58
Student: ResNet1D (used in main)	1.22 M	4.9	4.02 ± 0.08	0.78
Student: SimpleCNN1D	0.56 M	2.2	2.40 ± 0.08	0.59
Student: ECGTransformer	0.44 M	1.8	1.05 ± 0.04	0.05

The deployment regime our paper targets is therefore single-record, real-time CPU inference at $\sim$4 ms/record for the ResNet1D student, which fits in $\sim$4.9 MB of FP32 parameter memory — compatible with many edge-class deployment settings; additional quantization would be needed for tighter microcontroller budgets. This student-side latency and size advantage holds regardless of which teacher is chosen — only training cost differs across teacher conditions — and at deployment

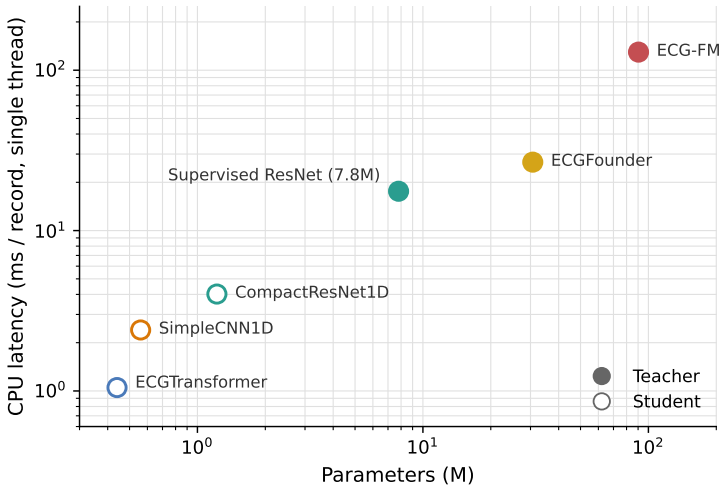


Figure 5: CPU latency vs. parameter count for every model in Table 14 (log–log). Teachers are filled markers, students are open markers; family colours match Figure 2. The student tier sits roughly an order of magnitude below the teacher tier in both axes.

time the student's roughly 32–124× speedup over any of the teachers is what a practitioner actually ships.

J Robustness: lead dropout and Gaussian noise

In this single-seed evaluation, KD from ECG-FM shows no consistent robustness advantage over no-KD. The two students remain close under Gaussian noise, whereas lead-dropout differences change sign and reach about 0.006 AUROC. Lead-dropout augmentation produces much larger gains under missing leads. The headline teacher comparisons do not depend on these perturbation results.

Lead-dropout robustness. We zero out a uniformly chosen subset of $k \in \{1, 3, 6, 9\}$ leads at test time and re-evaluate. "ResNet (robust)" is the same student trained with random lead dropout as augmentation. As expected, the lead-dropout penalty is severe for all unsupervised-augmentation models: the baseline ResNet drops from 0.927 AUROC at k=0 to 0.644 at k=9, while the augmentation-trained student stays above 0.848.

Table 15: PTB-XL test AUROC under random lead-dropout (zeroing k of 12 leads, single seed). At k=9, three leads remain. ECG-FM KD has no consistent advantage over no-KD; the augmentation-trained "robust" student performs better under heavy lead dropout.

Model	k=0	k=1	k=3	k=6	k=9
Teacher (ECG-FM)	0.892	0.888	0.872	0.831	0.741
Student: ResNet (no-KD)	0.927	0.918	0.888	0.800	0.644
Student: ResNet (KD-ECG-FM)	0.927	0.918	0.889	0.806	0.638
Student: ResNet (robust)	0.930	0.928	0.924	0.907	0.848

Gaussian-noise robustness. Additive white Gaussian noise at five SNRs from 40 dB down to 0 dB applied to the test inputs. All ECG models in our setting are remarkably robust to additive noise — the baseline ResNet loses just ∼0.013 AUROC even at 0 dB — so this perturbation is not a useful discriminator between conditions and is reported only for completeness.

Summary. Under the saved noise conditions, ECG-FM KD differs from no-KD by at most 0.0009 AUROC. Under lead dropout the differences range from −0.0053 to +0.0062 AUROC and change sign. This single-seed evaluation does not show a consistent robustness gain from ECG-FM KD.

Table 16: PTB-XL test AUROC under additive white Gaussian noise applied at test time (single seed).

Model	40 dB	20 dB	10 dB	5 dB	0 dB
Teacher (ECG-FM)	0.892	0.892	0.889	0.885	0.872
Student: ResNet (no-KD)	0.927	0.927	0.925	0.922	0.914
Student: ResNet (KD-ECG-FM)	0.927	0.927	0.926	0.923	0.913
Student: ResNet (robust)	0.930	0.929	0.928	0.925	0.915