# ECGDUO: COMPLEMENTARY DISTILLATION FROM ECG AND TIME-SERIES FOUNDATION MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Foundation models provide strong representations for electrocardiogram (ECG) analysis, but full ECG foundation models remain expensive for clinical settings that require compact, fast inference and efficient adaptation. A common remedy is to distill a large ECG teacher into a small encoder. ECG-only distillation, however, suffers from a *single-proxy transfer bottleneck*: its supervision is confined to one ECG-specific representation space, limiting the diversity of temporal knowledge transferred to the student. We therefore propose **ECGDuo**, a compact ECG encoder distilled from two complementary teachers: *(i)* an ECG-specific teacher that transfers cardiac knowledge, and *(ii)* a general time-series teacher that injects domain-agnostic temporal knowledge. ECGDuo uses symmetric contrastive distillation to transfer both forms of knowledge into a lightweight raw-ECG encoder. Each teacher first encodes the same raw ECG signal into either an ECG-specific representation or a domain-agnostic temporal representation. Branch-specific adapters then map the shared student representation into the corresponding teacher spaces, where matched student–teacher pairs are aligned. Across five downstream ECG datasets, ECGDuo consistently outperforms baselines trained from scratch and remains competitive with full-scale foundation models, using only **0.1% of Sundial's parameters**. Code is available in the ***Anonymous Repository***.

## 1 INTRODUCTION

Foundation models (FMs) provide strong transferable representations for electrocardiogram (ECG) analysis (McKeen et al., 2025; Li et al., 2025; Nolin-Lapalme et al., 2025), but full FMs are costly for clinical workflows that require fast inference, small memory footprint, and label-efficient adaptation. Knowledge distillation is a natural compression remedy: a compact student is trained to reproduce the outputs or intermediate representations of a large pretrained teacher, transferring its knowledge without retaining the full model at deployment (Hinton et al., 2015; Romero et al., 2015; Park et al., 2019; Tian et al., 2020; Gou et al., 2021).

Yet current ECG-only distillation suffers from a fundamental *single-proxy transfer bottleneck*: the distillation objective directly supervises only the representation supplied by the ECG teacher. Electrocardiogram foundation models (ECGFMs) organize raw waveforms around ECG-domain structure, including cardiac morphology, rhythm, cross-lead relationships, and diagnostic semantics (McKeen et al., 2025; Li et al., 2025; Nolin-Lapalme et al., 2025). The temporal structures they emphasize are shaped by ECG-specific pretraining data and objectives (McKeen et al., 2025; Li et al., 2025). A compact student distilled from this single space is therefore explicitly constrained to preserve an ECG-specific view of the waveform. This single-view supervision leaves cross-domain temporal structure outside the transfer objective, motivating a second teacher that captures temporal dependencies learned across heterogeneous time-series domains.

We address this bottleneck by introducing a general time-series foundation model (TSFM) as a second source of supervision. Unlike ECGFMs, a TSFM is pretrained across heterogeneous time-series domains and learns temporal dependencies that recur across applications rather than being tied to ECG-specific morphology or diagnostic semantics (Zhou et al., 2023; Goswami et al., 2024; Ansari et al., 2024; Das et al., 2023; Liu et al., 2025). Such cross-domain temporal knowledge provides an independent constraint on information that is under-expressed in the ECGFM space. By aligning with both teachers, the student preserves ECG-specific cardiac knowledge while retaining complementary,



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.

domain-agnostic ECG temporal dynamics. Operationally, the frozen ECGFM and TSFM teachers generate complementary targets for each unlabeled 12-lead ECG segment, which are cached before student training. A lightweight CNN-based student then encodes the same raw segment, and two branch-specific adapters project its shared representation into the corresponding teacher spaces. Within each branch, symmetric contrastive distillation aligns the projected student representation with its matched teacher target, while the two branch losses jointly optimize the shared encoder. For downstream tasks, the encoder remains frozen, and only a linear task head is trained.

Across five downstream ECG benchmarks, the proposed distillation framework, denoted as **ECGDuo**, outperforms baselines trained from scratch and remains competitive with full pretrained ECGFMs, using only 0.1% of Sundial's parameters. Ablations show consistent gains over ECG-only distillation, yielding an overall F1 improvement of 2.06%. Frozen-encoder probing further demonstrates improved label efficiency across matched downstream budgets and evaluation folds.

Our contributions are:

- We identify a *single-proxy transfer bottleneck* in current ECG-only distillation: a single ECGFM provides no direct supervision for structure outside its representation.
- We introduce **ECGDuo**, which combines ECG-specific supervision with a general TSFM constraint into one compact raw-ECG encoder.
- The proposed ECGDuo achieves strong downstream performance on five ECG benchmarks using approximately **0.1% of Sundial's parameters**.

## 2 Related Work

**ECG representation learning and foundation models.** Self-supervised ECG representation learning exploits temporal, lead-wise, and patient-level structure to reduce dependence on labeled data (Kiyasseh et al., 2021; Mehari & Strodthoff, 2022). More recently, ECG foundation models have scaled pretraining to substantially larger ECG corpora and demonstrated broad transfer across clinical tasks and cohorts (McKeen et al., 2025; Li et al., 2025; Nolin-Lapalme et al., 2025). These models provide strong ECG-specific representations, yet a compact student distilled from a single ECG teacher is supervised only by the knowledge exposed through that teacher. ECGDuo investigates whether such compact transfer can benefit from an additional, complementary source of temporal knowledge beyond the ECG-specific teacher.

**General time-series foundation models.** General time-series representation learning aims to capture temporal structure transferable across datasets and domains (Yue et al., 2022; Cheng et al., 2023). Recent time-series foundation models substantially scale this paradigm through heterogeneous pretraining, including MOMENT, Chronos, TimesFM, and Sundial (Goswami et al., 2024; Ansari et al., 2024; Das et al., 2023; Liu et al., 2025). Such broad pretraining can encode temporal regularities that need not coincide with those emphasized by ECG-specific objectives. Rather than adapting a general TSFM as an ECG predictor or retaining it in deployment, ECGDuo uses its representation of the same waveform as an offline source of cross-domain temporal supervision that complements the ECG-specific teacher during pretraining.

**Representation and multi-teacher distillation.** Knowledge distillation transfers predictive outputs or intermediate representations from large teachers to compact students (Hinton et al., 2015; Romero et al., 2015; Park et al., 2019; Tian et al., 2020; Gou et al., 2021). Multi-teacher distillation further combines knowledge from multiple teacher networks (You et al., 2017), with recent work extending this principle to task-agnostic representation learning from diverse pretrained embedders (Formont et al., 2025). ECGDuo addresses a different specialization of this setting: the teachers are deliberately heterogeneous but observe the same physiological waveform; an ECG-specific foundation model provides domain-specialized knowledge, while a general TSFM provides cross-domain temporal knowledge. The key question is therefore not whether multiple teachers can be combined, but whether intentionally complementary teacher spaces improve the representational coverage retained by a compact student. Their distinct representation spaces are aligned through separate branches and distilled into a single compact ECG encoder, with both foundation models discarded after pretraining. Matched student architectures isolate the added supervision at fixed student deployment capacity.

# 3 METHOD

ECGDuo distills two pretrained views of the same ECG waveform into one compact representation. The ECG foundation model supplies cardiac-specific structure, while the general time-series foundation model supplies cross-domain temporal structure. Rather than merging them, ECGDuo keeps the two teacher spaces separate while transferring both into a shared student. Figure 1 summarizes the training and downstream-transfer pipeline. We characterize when predictive complementarity survives compression, then derive the corresponding symmetric contrastive objective for offline teacher targets.

## 3.1 PREDICTIVE COMPLEMENTARITY UNDER COMPRESSION

Let $X$ denote a raw ECG segment, $Y$ a downstream label, and

$$Z_E = g_E(X), \qquad Z_T = g_T(X) \tag{1}$$

be representations from the frozen ECGFM and TSFM. Following the predictive multi-view perspective of Tosh et al. (2021), we distinguish teacher diversity from *task-relevant complementarity*. For the TSFM relative to the ECGFM, define

$$\Delta_{T|E}(Y) = I(Y; Z_T \mid Z_E), \qquad \Delta_{E|T}(Y) = I(Y; Z_E \mid Z_T). \tag{2}$$

These quantities measure predictive information contributed by one teacher after the other is known. In particular, geometrically different teachers can still have zero complementarity for a task.

Let $Z_{ET} = (Z_E, Z_T)$ and let $H = f_\theta(X)$ denote the compact student. We quantify the task-relevant teacher information not yet captured by the student as

$$\rho_Y(H) = I(Y; Z_{ET} \mid H). \tag{3}$$

The following result separates the gain supplied by teacher complementarity from the information lost during compression.

**Theorem 3.1** (Complementarity survives compression)**.** *For any discrete downstream task, let $\mathcal{R}^\star_{\log}(S) = H(Y \mid S)$ be the Bayes log-loss risk using representation $S$. Then*

$$\boxed{\mathcal{R}^\star_{\log}(H) - \mathcal{R}^\star_{\log}(Z_E) \le \rho_Y(H) - \Delta_{T|E}(Y).} \tag{4}$$

*The symmetric inequality holds relative to $Z_T$. Consequently,*

$$\rho_Y(H) < \Delta_{T|E}(Y) \quad \Longrightarrow \quad \mathcal{R}^\star_{\log}(H) < \mathcal{R}^\star_{\log}(Z_E). \tag{5}$$

**Interpretation.** The theorem separates two competing effects: the *complementarity margin* $\Delta_{T|E}$ contributed by the TSFM and the *compression loss* $\rho_Y(H)$ remaining outside the student. When the former exceeds the latter, the compact student strictly improves over the ECG teacher in optimal prediction. If $H$ fully retains the joint teacher view $Z_{ET}$, then $\rho_Y(H) = 0$, and the full complementarity margin is inherited by the student. Appendix A provides the proof and compares it with its linear-probe counterpart.

## 3.2 OFFLINE COMPLEMENTARY TEACHER TARGETS

For an unlabeled corpus $\mathcal{D}_u = \{x_i\}_{i=1}^N$ with $x_i \in \mathbb{R}^{12\times T}$, the two frozen teachers encode each ECG segment once:

$$z_i^E = g_E(x_i) \in \mathbb{R}^{d_E}, \qquad z_i^T = g_T(x_i) \in \mathbb{R}^{d_T}. \tag{6}$$

The ECG branch follows the native ECGFM preprocessing and representation extraction. For the TSFM branch, each lead is split into fixed-length chunks, encoded independently, aggregated within the lead, and pooled across 12 leads. The two targets therefore retain their native teacher spaces. They are cached before student optimization, so foundation-model inference occurs only once offline.

Each cached pair $(z_i^E, z_i^T)$ shares the same recording and window identity, so both branches supervise one physiological segment. The ECG-specific target preserves cardiac representation geometry, while the TSFM target captures its lead-wise temporal view. Separate caches retain these native coordinates throughout optimization. Both targets remain fixed as gradients from the two branches update the shared student, establishing a stable correspondence between each waveform and its complementary supervision. Teacher inference is completed offline; the resulting compact encoder performs downstream prediction directly from ECG waveforms with a task-specific linear head.

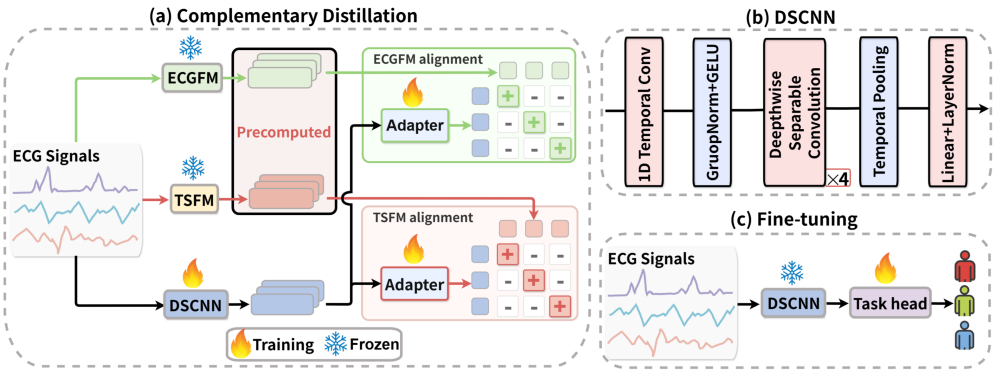


Figure 1: **ECGDuo overview.** (a) Frozen ECGFM and TSFM provide offline targets to branch-specific adapters of a shared DSCNN. (b) Depthwise-separable temporal blocks form the compact student. (c) Teachers and adapters are discarded after pretraining; downstream transfer uses only the frozen encoder and a linear head.

### 3.3 Shared Student with Teacher-Specific Adapters

As shown in Figure 1(a,b), a lightweight DSCNN maps the raw segment to $h_i = f_\theta(x_i) \in \mathbb{R}^d$ using a temporal stem, residual depthwise-separable blocks, adaptive pooling, and a compact projection layer (Howard et al., 2017; Chollet, 2017). Because the teacher embeddings differ in dimensionality and geometry, separate linear adapters map the same $h_i$ into each teacher space,

$$u_i^k = p_k(h_i) = W_k h_i + b_k, \qquad p_k : \mathbb{R}^d \to \mathbb{R}^{d_k}, \quad k \in \{E, T\}. \tag{7}$$

Branch-specific adapters absorb coordinate differences, while the shared backbone must satisfy both supervision sources. Their constraints therefore accumulate in one deployable representation without merging the teacher spaces.

### 3.4 Symmetric Contrastive Distillation

For branch $k \in \{E, T\}$, $(u_i^k, z_i^k)$ is the matched student–teacher pair and the remaining samples in the mini-batch act as negatives. With $\ell_2$-normalized features and $s_{ij}^k = (\bar{u}_i^k)^\top \bar{z}_j^k / \tau_k$, we optimize

$$\mathcal{L}_k = -\frac{1}{2B} \sum_{i=1}^{B} \left[ \log \frac{e^{s_{ii}^k}}{\sum_j e^{s_{ij}^k}} + \log \frac{e^{s_{ii}^k}}{\sum_j e^{s_{ji}^k}} \right], \qquad \mathcal{L}_{\text{ECGDuo}} = \frac{\mathcal{L}_E + \mathcal{L}_T}{2}. \tag{8}$$

The two directions retrieve matched teacher and student representations symmetrically. Both branch losses update the shared compact encoder, while their adapters preserve teacher-specific geometry.

**Proposition 3.2.** *With standard InfoNCE sampling and finite information terms, we obtain*

$$\boxed{I(H; Z_E, Z_T) \geq 2 \log B - \mathcal{L}_E - \mathcal{L}_T - I(Z_E; Z_T).} \tag{9}$$

**Retention view.** The bound rewards information retained from both teachers while discounting their overlap. ECGDuo jointly retains *recoverable* information from both teachers, and their *task-relevant complementarity* determines the value of the joint teacher view. If the student retains all task-relevant information in the joint teacher view, $\rho_Y(H)$ in Theorem 3.1 vanishes. Appendix A proves the bound and relates it to multi-teacher information retention (Formont et al., 2025).

### 3.5 Linear Downstream Transfer

As shown in Figure 1(c), both foundation models and both adapters are discarded after pretraining. For each downstream dataset, the compact encoder remains frozen and only a linear classifier is optimized,

$$h_i = f_\theta(x_i), \qquad \hat{y}_i = W_c h_i + b_c. \tag{10}$$

This head-only protocol evaluates the distilled representation directly and leaves deployment with only the compact raw-ECG encoder and a task-specific linear head.

# 4 EXPERIMENTS

We evaluate the proposed ECGDuo (**the default version is distilled using ECGFounder (Li et al., 2025) and Sundial (Liu et al., 2025) as teachers**) on five downstream ECG classification benchmarks. The distilled encoder remains frozen, and only a linear classifier is fitted for each downstream dataset.

## 4.1 EXPERIMENTAL SETUP

**Datasets.** Distillation uses unlabeled waveforms from MIMIC-IV-ECG (Gow et al., 2023) and CODE-15 (Ribeiro et al., 2021); labels from these corpora are not used. Downstream evaluation is conducted on CPSC2018 (Liu et al., 2018), CSN (Zheng et al., 2022), G12EC (Perez Alday et al., 2020), PTB-XL (Wagner et al., 2020), and PTB (Bousseljot et al., 1995). Table 1 summarizes subject and 3-s window counts, classes, and sampling rates, and separates the unlabeled distillation corpora from the downstream benchmarks.

Table 1: **Dataset statistics.**

| | *Unlabeled distillation* | | *Downstream evaluation* | | | | |
|---|---|---|---|---|---|---|---|
| | MIMIC-IV ECG | CODE-15 | CPSC2018 | CSN | G12EC | PTB-XL | PTB |
| Subjects | 58,792 | 160,661 | 7,131 | 20,757 | 6,081 | 13,273 | 200 |
| Windows | 284,592 | 386,881 | 35,213 | 62,271 | 18,167 | 42,726 | 16,712 |
| Classes | – | – | 5 | 5 | 4 | 4 | 2 |
| Sampling (Hz) | 500 | 400 | 500 | 500 | 500 | 500 | 1000 |

**Evaluation protocol.** We use five-fold subject-independent cross-validation and keep all recordings from the same subject in a single fold. Performance is reported as mean±std across folds for Accuracy, macro Precision, Recall, F1-score, AUROC, and AUPRC. Their arithmetic mean (Avg) is used as the summary score. Table 2 reports this aggregate comparison; the complete metric-wise results are provided in Appendix Tables 4-5.

**Baselines.** We use two baseline groups. The supervised group contains XResNet1D (Strodthoff et al., 2021), Inception1D (Ismail Fawaz et al., 2020), MedGNN (Fan et al., 2025), iTransformer (Liu et al., 2024), Medformer (Wang et al., 2024), and PatchTST (Nie et al., 2023), all trained end-to-end on each downstream dataset. The frozen foundation-model group contains *Sundial* (Liu et al., 2025), *ECG-FM* (McKeen et al., 2025), and *ECGFounder* (Li et al., 2025); their encoders are frozen and only a linear classifier is fitted. ECGDuo follows the same frozen-encoder protocol. Both teachers, together with the teacher-specific adapters, are discarded after distillation and are not used downstream. The two settings are reported separately in Table 2.

Table 2: **Downstream transfer under two comparison regimes.** Entries are Avg (mean±std over five subject-independent folds); the strongest non-ECGDuo result in each column is underlined.

| **Model** | **CPSC2018** | **CSN** | **G12EC** | **PTB-XL** | **PTB** | **Mean** |
|---|---|---|---|---|---|---|
| **A. Supervised task-specific models** *(end-to-end training)* | | | | | | |
| XResNet1D | 57.24±7.20 | 53.65±7.92 | 63.81±4.03 | 58.21±6.43 | 72.56±5.78 | 61.09 |
| Inception1D | 61.54±5.37 | 61.48±4.74 | 59.67±7.61 | 69.56±5.81 | 72.51±7.52 | 64.95 |
| MedGNN | 60.02±3.47 | 54.77±4.33 | 57.94±2.81 | 70.29±2.83 | 70.57±8.65 | 62.72 |
| iTransformer | 45.42±0.70 | 76.14±1.12 | 63.08±0.69 | 56.85±1.39 | 76.22±4.51 | 63.54 |
| Medformer | 53.07±0.84 | 63.25±0.97 | 51.48±1.06 | 59.34±1.49 | 70.80±5.00 | 59.59 |
| PatchTST | 62.68±1.46 | 73.03±1.12 | 70.19±1.63 | 69.69±0.67 | 80.30±4.20 | 71.18 |
| **ECGDuo (Ours)** | **78.27±1.10** | **89.50±0.42** | **78.56±1.25** | **81.06±1.01** | **91.26±1.78** | **83.73** |
| **B. Frozen foundation models** *(linear probe)* | | | | | | |
| Sundial | 75.53±1.78 | 87.97±0.37 | 77.82±1.63 | 76.49±0.87 | 84.94±2.77 | 80.55 |
| ECG-FM | 76.86±1.05 | 87.61±2.36 | 78.05±3.07 | 76.75±3.65 | 87.05±3.79 | 81.26 |
| ECGFounder | 78.59±1.63 | 90.37±0.76 | 82.44±1.26 | 81.26±1.22 | 87.33±5.23 | 84.00 |
| **ECGDuo (Ours)** | **78.27±1.10** | **89.50±0.42** | **78.56±1.25** | **81.06±1.01** | **91.26±1.78** | **83.73** |

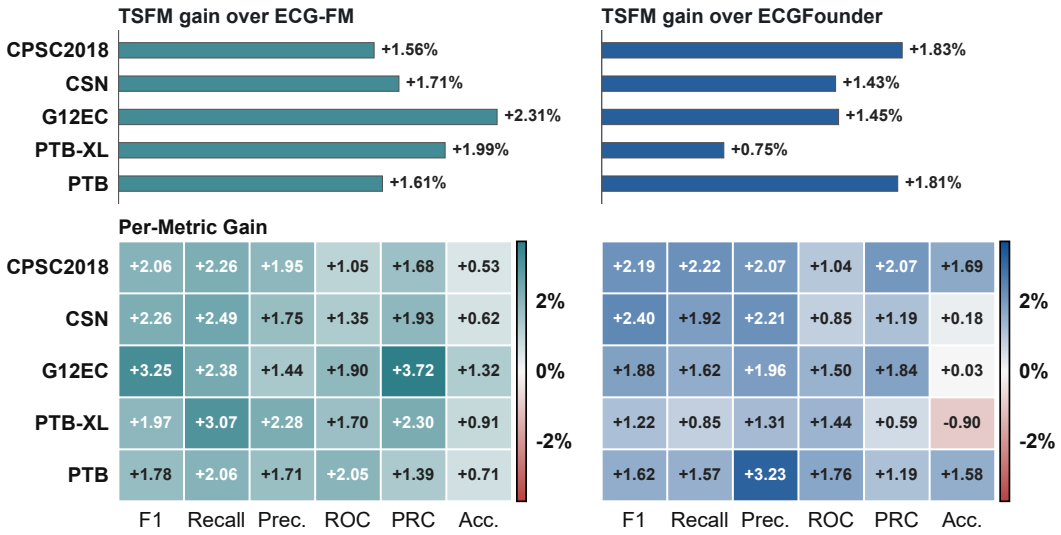


Figure 2: Each distilled dual-teacher student is compared with its corresponding ECG-only student. **Top**: relative improvement across all metrics. **Bottom**: metric-wise relative improvement.

### 4.2 Main Results

Table 2 evaluates the value of transferring ECG-specific (from ECGFM) and general time-series knowledge (from TSFM) into a compact student from two complementary perspectives. ***Setting A*** compares ECGDuo with dataset-specific models trained end-to-end from scratch, testing whether the distilled representation offers an advantage over learning solely from the labeled target data. ***Setting B*** compares frozen pretrained encoders under the same linear-probe protocol, testing how much of the full foundation models' capability remains after being compressed into a compact student.

In *Setting A*, ECGDuo ranks first on all five benchmarks and improves the strongest supervised result by $11.92\%$ to $24.87\%$ in overall relative Avg. The leading baseline differs across datasets: PatchTST on CPSC2018, G12EC, and PTB; iTransformer on CSN; and MedGNN on PTB-XL, indicating distinct characteristics of these heterogeneous datasets. Despite this variation, ECGDuo consistently outperforms the strongest dataset-specific alternative. Its advantage over convolutional, graph-based, and Transformer architectures with different inductive biases indicates that the distilled representation transfers robustly across downstream settings.

In *Setting B*, under the same frozen-encoder linear-probe protocol, ECGDuo outperforms *Sundial* and *ECG-FM* on every benchmark while remaining highly competitive with *ECGFounder*: it nearly matches *ECGFounder* on CPSC2018 and PTB-XL and exceeds it by 3.93 points on PTB, resulting in only a $0.27$-point margin in the five-dataset mean ($83.73$ *vs.* $84.00$). Crucially, ECGDuo achieves this near-teacher performance with only **0.137M** parameters, approximately $0.45\%$ of *ECGFounder* and $0.1\%$ of *Sundial*. This result establishes a favorable accuracy-efficiency trade-off: ECGDuo retains performance very close to a full-scale ECG foundation model while being far more efficient. Its reduced memory footprint and computational cost make the distilled encoder better suited to high-throughput ECG analysis and deployment in resource-constrained clinical environments.

### 4.3 Ablation Study

Figure 2 evaluates the contribution of TSFM (*i.e.*, *Sundial*) supervision. Adding TSFM supervision to *ECG-FM* consistently improves Avg by $1.56\%$ to $2.31\%$ across all datasets, with an overall gain of $1.84\%$. When paired with the stronger *ECGFounder*, it still raises Avg by $0.75\%$ to $1.83\%$, averaging $1.45\%$; the only metric-level decrease is PTB-XL Accuracy ($-0.90\%$), while all other metrics improve. The smaller but consistently positive improvements with *ECGFounder* indicate that the TSFM supervision provides complementary general temporal knowledge even when the ECG-specific teacher is already strong. These results show that ECGDuo benefits from combining complementary teacher views rather than merely compensating for limitations of a particular ECGFM.

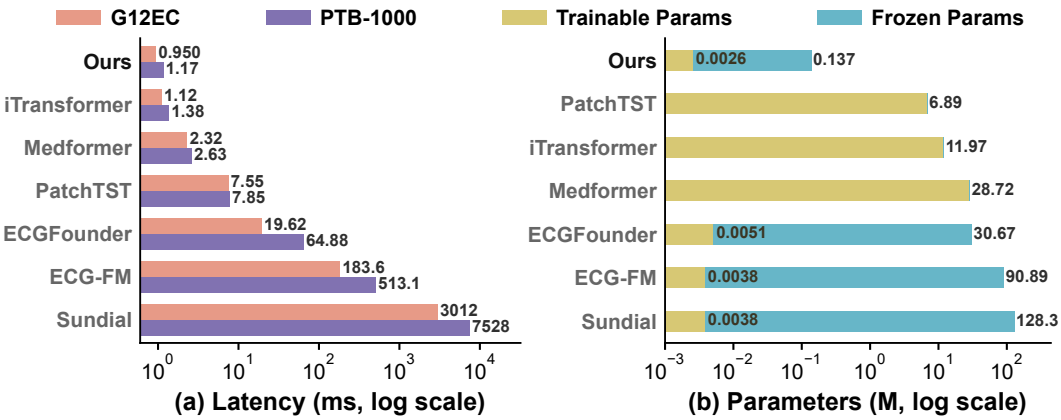


Figure 3: **Deployment efficiency. (a)**: inference latency on G12EC and PTB. **(b)**: total and trainable parameters comparison. We use a batch size of $bs = 128$ and a model dimension of $d_{\text{model}} = 512$.

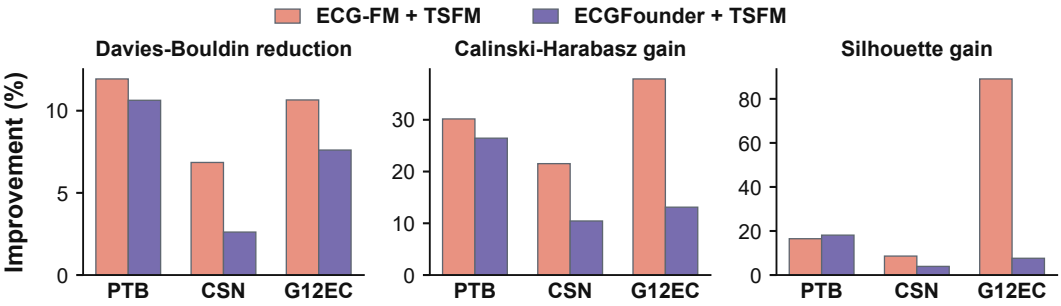


Figure 4: **Change in feature cluster quality after adding Sundial.** Bars report relative changes from the corresponding ECG-only student on PTB, CSN, and G12EC.

### 4.4 Efficiency Analysis

This section evaluates the deployment cost of ECGDuo against representative dataset-specific models and its full-scale foundation model counterparts, considering both inference latency and parameter footprint. Figure 3**(a)** reports inference latency on G12EC ($500Hz$) and PTB ($1000Hz$), where ECGDuo is consistently the fastest model despite the different input resolutions. Figure 3**(b)** shows that ECGDuo contains only $0.137M$ parameters, approximately $50\times$ *fewer than PatchTST and over* $900\times$ *fewer than Sundial*, of which merely $0.0026M$ ($1.87\%$) are optimized during linear probing. In comparison, frozen foundation models require few trainable parameters but **still retain their complete** $30.67$ **to** $128.33M$**-parameter encoders** at inference. By distilling the ECG-specific and general time-series knowledge of large foundation models into a compact ECGDuo encoder, our method **preserves their complementary knowledge without inheriting their deployment cost, achieving both low-latency inference and a small model footprint.**

### 4.5 Representation Visualization

This section examines the effect of TSFM supervision on the representation learned by ECGDuo. Figure 4 quantifies this change using three complementary clustering metrics: the Davies-Bouldin index decreases by $2.62\%$ to $11.93\%$, while the Calinski-Harabasz index and Silhouette coefficient increase by $10.45\%$ to $37.89\%$ and $3.93\%$ to $89.05\%$, respectively. The improvement is consistent across both *ECG-FM* and *ECGFounder*. Figure 5 further compares the frozen feature space of ECG-only and ECG+TSFM students on PTB, CSN, and G12EC. For both *ECG-FM* and *ECGFounder* branches, adding TSFM (*i.e.*, Sundial) produces more compact class regions with reduced overlap, indicating that complementary temporal knowledge is well-organised in the student representation. These results show that TSFM supervision complements both ECG teachers, with larger benefits when the ECG-specific representation leaves more class-relevant temporal structure uncaptured.

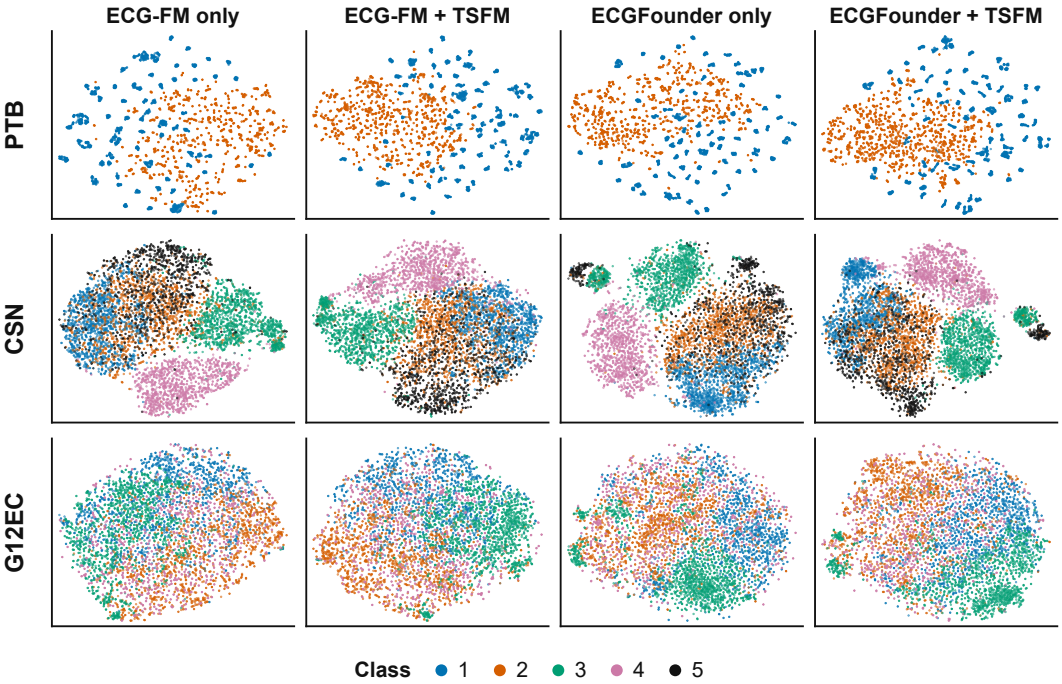


Figure 5: **Pretrained representation geometry before linear probing.** Each row projects an identical set of 6,000 ECG windows across teacher configurations, so within-row changes isolate TSFM supervision from sampling variation. Colors denote diagnostic classes.

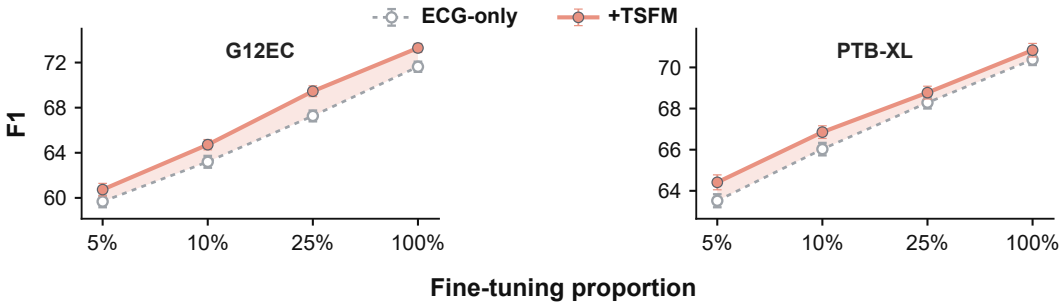


Figure 6: **Frozen-encoder transfer under label scarcity.** Only the class-stratified subjects used to fit the linear head vary; encoder features and evaluation folds are fixed. Each point averages 45 matched fold/seed runs; error bars are 95% CIs and shading denotes the paired mean gap.

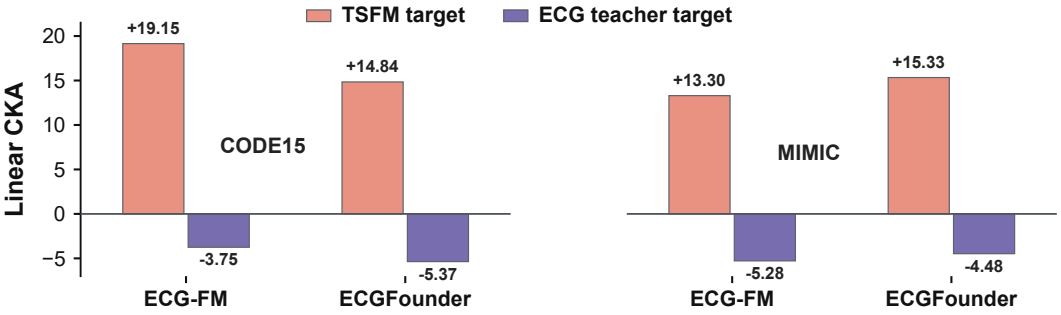


Figure 7: **Student–teacher representation alignment.** Bars show $100 \times \Delta$ Linear CKA on 5,000 held-out windows per dataset, using TSFM and the matched ECG teacher as separate targets. Signs indicate increased or decreased alignment after TSFM supervision.

**Label-efficient transfer.** Figure 6 evaluates whether the benefit of TSFM supervision persists when downstream labels are scarce. Under the same frozen-encoder protocol, the dual-teacher student outperforms its ECG-only counterpart across all eight *dataset-budget* combinations. Relative to ECG-only distillation, TSFM supervision yields an average relative F1 gain of $2.44\%$ on G12EC, peaking at $3.26\%$ with only $25\%$ of the labels. On PTB-XL, the advantage is most pronounced in the most label-constrained regimes, reaching $1.41\%$ and $1.25\%$ with $5\%$ and $10\%$ of the labels, respectively, and remains consistently positive as label coverage increases. Moreover, every paired $95\%$ confidence interval excludes zero across the 45 matched fold/seed runs, showing that the improvement is not driven by a particular split or label budget.

Because only the linear head is optimized, this systematic upward shift cannot be attributed to downstream adaptation of the encoder. Instead, ECGDuo makes task-relevant temporal structure more readily accessible to a linear classifier, consistently strengthening frozen-encoder transfer from scarce-label to fully supervised settings.

**Complementary knowledge retention.** Figure 7 uses Linear CKA (Kornblith et al., 2019) to compare student alignment with each teacher before and after adding TSFM supervision. Relative to the corresponding ECG-only student, adding TSFM supervision increases Linear CKA with the TSFM target by $13.30$ to $19.15$ points across CODE-15 and MIMIC-IV-ECG, whereas alignment with the matched ECG teacher decreases by only $3.75$ to $5.37$ points. In every *teacher-dataset* pairing, the increase toward the TSFM target is at least $2.5\times$ larger than the corresponding loss toward the ECG teacher. This asymmetric retention profile is reproduced with both *ECG-FM* and *ECGFounder*, demonstrating that the effect is not tied to a particular ECG teacher.

Rather than trading one teacher view for the other, ECGDuo substantially broadens the student's representational coverage: ***it acquires strong cross-domain temporal alignment while preserving most of the ECG-specific structure inherited from the ECG teacher***. The concurrent downstream improvements in Figure 2 further show that this additional structure is functionally useful rather than merely geometric. Together, the alignment analysis and matched downstream gains demonstrate that ECGDuo integrates complementary teacher knowledge into a compact encoder, expanding temporal coverage while retaining ECG-specific structure.

## 5 Conclusion

This work addresses the single-view supervision of ECG-only distillation: one ECG teacher transfers cardiac knowledge, while cross-domain temporal structure remains outside its transfer targets. ECGDuo overcomes this *single-proxy transfer bottleneck* by jointly distilling an ECG-specific foundation model and a general time-series foundation model into a shared lightweight encoder. Teacher-specific adapters preserve the distinct supervision provided by the two teacher spaces during training, while both teachers and adapters are removed after distillation, leaving the compact encoder for downstream deployment. Across five subject-independent benchmarks, ECGDuo consistently outperforms task-specific models trained from scratch and retains performance competitive with full-scale ECGFMs using only $0.137$M parameters (**0.1% of Sundial**). Matched teacher ablations demonstrate that complementary supervision improves transfer at fixed student architecture and capacity. TSFM supervision improves nearly every downstream metric across two different ECG teachers, produces more discriminative feature geometry, and consistently strengthens frozen-encoder transfer across label budgets. Moreover, Linear CKA reveals that the substantial increase in TSFM alignment is accompanied by only a limited reduction in ECG-teacher alignment. ECGDuo therefore expands the representational coverage of the compact student without discarding the ECG-specific structure already provided by the ECG teacher. Taken together, these results establish the proposed complementary distillation framework as an effective way to extend the supervision available to a compact student beyond a single teacher space. ECGDuo decouples the capacity required to acquire diverse pretrained knowledge from the capacity required to deploy it: high-capacity teachers provide complementary supervision during training, whereas only the compact shared encoder remains at inference. This principle offers a practical foundation for building efficient physiological encoders when no single pretrained model provides a sufficiently complete transfer target.

## REFERENCES


Abdul Fatir Ansari et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

R. Bousseljot, D. Kreiseler, and A. Schnabel. Nutzung der EKG-signaldatenbank CARDIODAT der PTB über das internet. *Biomedizinische Technik*, 40(s1):317–318, 1995. doi: 10.1515/bmte.1995.40.s1.317.

Mingyue Cheng, Qi Liu, Zhiding Liu, Hao Zhang, Rujiao Zhang, and Enhong Chen. TimeMAE: Self-supervised representations of time series with decoupled masked autoencoders. *arXiv preprint arXiv:2303.00320*, 2023. URL https://arxiv.org/abs/2303.00320v3.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv:2310.10688*, 2023.

Wei Fan et al. MedGNN: Towards multi-resolution spatiotemporal graph learning for medical time series classification. *arXiv preprint arXiv:2502.04515*, 2025.

Philippe Formont, Maxime Darrin, Banafsheh Karimian, Jackie CK Cheung, Eric Granger, Ismail Ben Ayed, Mohammadhadi Shateri, and Pablo Piantanida. Learning task-agnostic representations through multi-teacher distillation. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3664.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *International Conference on Machine Learning*, 2024.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819, 2021.

Brian Gow et al. MIMIC-IV-ECG: Diagnostic electrocardiogram matched subset. *PhysioNet*, 2023. doi: 10.13026/4nqg-sb35.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Andrew G. Howard et al. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Hassan Ismail Fawaz et al. Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery*, 34:1936–1962, 2020.

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

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 3519–3529, 2019.

Jun Li et al. An electrocardiogram foundation model built on over 10 million recordings. *NEJM AI*, 2(7):AIoa2401033, 2025.

Feifei Liu et al. 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.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*, 2024.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. In *International Conference on Machine Learning*, 2025.

Kaden McKeen et al. ECG-FM: An open electrocardiogram foundation model. *JAMIA Open*, 8(5): ooaf122, 2025.

Temesgen Mehari and Nils Strodthoff. Self-supervised representation learning from 12-lead ECG data. *Computers in Biology and Medicine*, 141:105114, 2022.

Yuqi Nie et al. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

Alexis Nolin-Lapalme et al. Foundation models for generalizable electrocardiogram interpretation: Comparison of supervised and self-supervised electrocardiogram foundation models. *medRxiv*, 2025. doi: 10.1101/2025.03.02.25322575.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Erick A. Perez Alday et al. Classification of 12-lead ECGs: The physionet/computing in cardiology challenge 2020. *Physiological Measurement*, 41(12):124003, 2020. doi: 10.1088/1361-6579/abc960.

Antônio H. Ribeiro et al. CODE-15%: A large scale annotated dataset of 12-lead ECGs, 2021.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015.

Nils Strodthoff et al. Deep learning for ECG analysis: Benchmarks and insights from PTB-XL. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1519–1528, 2021.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2020.

Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive learning, multi-view redundancy, and linear models. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pp. 1179–1206. PMLR, 2021.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Patrick Wagner et al. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific data*, 7 (1):154, 2020.

Yihe Wang, Nan Huang, Taida Li, Yujun Yan, and Xiang Zhang. Medformer: A multi-granularity patching transformer for medical time-series classification. In *NeurIPS*, 2024. URL https://openreview.net/forum?id=jfkid2HwNr.

Shan You, Chang Xu, Chao Xu, and Dacheng Tao. Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1285–1294, 2017. doi: 10.1145/3097983.3098135.

Zhihan Yue et al. TS2Vec: Towards universal representation of time series. In *AAAI Conference on Artificial Intelligence*, 2022.

Jianwei Zheng, Hangyuan Guo, and Huimin Chu. A large scale 12-lead electrocardiogram database for arrhythmia study. *PhysioNet*, 2022. doi: 10.13026/wgex-er52.

Tian Zhou, Peisong Niu, Xue Wang, Liang Sun, and Rong Jin. One fits all: Power general time series analysis by pretrained lm. In *Advances in Neural Information Processing Systems*, 2023.

# A Theory for Complementary Distillation

## A.1 Proof of Theorem 3.1

**Log-loss proof.** Let $Z_{ET} = (Z_E, Z_T)$. By the chain rule,

$$\Delta_{T|E}(Y) = I(Y; Z_T \mid Z_E) = H(Y \mid Z_E) - H(Y \mid Z_{ET}). \tag{A.1}$$

Using Eq. (A.1), the student-to-ECG risk gap becomes

$$\mathcal{R}^\star_{\log}(H) - \mathcal{R}^\star_{\log}(Z_E) = H(Y \mid H) - H(Y \mid Z_E) \tag{A.2}$$

$$= H(Y \mid H) - H(Y \mid Z_{ET}) - \Delta_{T|E}(Y). \tag{A.3}$$

Conditioning cannot increase entropy, so $H(Y \mid Z_{ET}, H) \le H(Y \mid Z_{ET})$. Therefore

$$H(Y \mid H) - H(Y \mid Z_{ET}) \le H(Y \mid H) - H(Y \mid Z_{ET}, H) = I(Y; Z_{ET} \mid H) = \rho_Y(H). \tag{A.4}$$

Substituting Eq. (A.4) into Eq. (A.3) proves

$$\mathcal{R}^\star_{\log}(H) - \mathcal{R}^\star_{\log}(Z_E) \le \rho_Y(H) - \Delta_{T|E}(Y). \tag{A.5}$$

Thus $\rho_Y(H) < \Delta_{T|E}(Y)$ implies a strict log-loss advantage over the ECG teacher. If $Z_{ET}$ is fully recoverable from $H$, then $I(Y; Z_{ET} \mid H) = 0$, so the complete task-relevant complementarity survives compression. Exchanging $E$ and $T$ gives the symmetric statement.

**Squared-loss and linear-probe corollary.** For vector-valued $Y$, let $m_E = \mathbb{E}[Y \mid Z_E]$ and $m_{ET} = \mathbb{E}[Y \mid Z_{ET}]$. The $L_2$ projection identity (Tosh et al., 2021) yields the exact relation

$$\mathcal{R}^\star_{\text{sq}}(Z_E) - \mathcal{R}^\star_{\text{sq}}(Z_{ET}) = \mathbb{E}\|m_{ET} - m_E\|_2^2. \tag{A.6}$$

Thus complementarity also appears as an exact squared-risk gain. Moreover, if the alignment maps recover both targets affinely, $Z_E = P_E(H)$ and $Z_T = P_T(H)$, every linear predictor on $Z_{ET}$ pulls back to a linear predictor on $H$. Consequently,

$$\mathcal{R}^\star_{\text{lin}}(H) \le \mathcal{R}^\star_{\text{lin}}(Z_{ET}) \le \mathcal{R}^\star_{\text{lin}}(Z_E), \tag{A.7}$$

with strict improvement whenever the joint teacher view has a strict linear predictive margin over $Z_E$. This linear-probe analogue of Theorem 3.1 matches our frozen-encoder evaluation.

## A.2 Proof of Proposition 3.2

**Joint-retention proof.** Let $U_k = p_k(H)$ be the student projection for branch $k \in \{E, T\}$. Standard InfoNCE with one positive and $B - 1$ marginal negatives gives (van den Oord et al., 2018)

$$I(U_k; Z_k) \ge \log B - \mathcal{L}_{S\to k}, \qquad I(U_k; Z_k) \ge \log B - \mathcal{L}_{k\to S}. \tag{A.8}$$

Because $\mathcal{L}_k$ averages the two directional losses, these inequalities imply the symmetric lower bound $I(U_k; Z_k) \ge \log B - \mathcal{L}_k$. Since $U_k$ is a deterministic function of $H$, data processing yields

$$I(H; Z_k) \ge I(U_k; Z_k) \ge \log B - \mathcal{L}_k. \tag{A.9}$$

The interaction identity gives

$$I(H; Z_E, Z_T) = I(H; Z_E) + I(H; Z_T) - I(Z_E; Z_T) + I(Z_E; Z_T \mid H). \tag{A.10}$$

Using $I(Z_E; Z_T \mid H) \ge 0$ and substituting Eq. (A.9) for the two branches yields

$$I(H; Z_E, Z_T) \ge 2\log B - \mathcal{L}_E - \mathcal{L}_T - I(Z_E; Z_T), \tag{A.11}$$

which proves Proposition 3.2.

**Interpretation.** The bound separates *retention* from *redundancy*: each branch bounds recoverable teacher information, and their overlap is discounted once. Frozen teachers make $I(Z_E; Z_T)$ constant, so minimizing the branch losses raises the lower bound on joint information retained by $H$. This connects the contrastive objective to task-agnostic multi-teacher retention (Formont et al., 2025). Separate adapters preserve native teacher coordinates while imposing both alignment constraints on the shared encoder. Theorem 3.1 then relates retention to downstream prediction: the student benefits when the teachers' task-relevant complementarity exceeds the student's residual compression loss.

# B Experimental Details

## B.1 Data, task construction, and representation acquisition

**Code and reproducibility.** All code required to reproduce ECGDuo, including training and evaluation scripts, data-preprocessing pipelines, configuration files, and execution details, is available in the ***Anonymous GitHub Repository***.

**Corpus roles and targets.** MIMIC-IV-ECG and CODE-15 supply waveform windows for unlabeled representation acquisition; their diagnostic labels never enter the distillation loss or model selection. CPSC2018, CSN, G12EC, PTB-XL, and PTB are reserved for downstream transfer. Eligibility is resolved before any fold is drawn: CPSC2018 and G12EC require one retained target group, CSN requires one original diagnosis among five rhythms, PTB-XL requires one non-HYP superclass shared by every patient record, and PTB requires concordant normal/control or myocardial-infarction records. Table 3 combines corpus scale, prediction unit, and the resulting endpoint contract.

**Waveform contract.** Signals remain at their native sampling rates and are reordered to the canonical leads I, II, III, aVR, aVL, aVF, and V1–V6. Records with an invalid shape, a missing or duplicated canonical lead, or a non-finite value are rejected. Each accepted lead is standardized using complete-record statistics with $10^{-8}$ stabilization, then divided into contiguous non-overlapping 3-s windows; the incomplete tail is discarded. We neither resample nor augment, pad, or normalize individual windows. CODE-15 trailing zeros are removed before statistics, and PTB Frank leads are excluded.

Table 3: Dataset scale, learning role, and endpoint contract. Source labels are excluded from distillation; transfer targets are fixed before subject-level partitioning.

| Dataset | Role / unit | Subjects | Windows | Hz | Target space and eligibility |
|---|---|---|---|---|---|
| ***Representation-learning sources with labels withheld*** | | | | | |
| MIMIC-IV-ECG | Source / window | 58,792 | 284,592 | 500 | Valid canonical signals; stable subject and record identifiers; all source labels remain withheld. |
| CODE-15 | Source / window | 160,661 | 386,881 | 400 | Valid exams; strip trailing zero padding before complete-record statistics; labels withheld. |
| ***Downstream benchmarks with subject-independent partitions*** | | | | | |
| CPSC2018 | Transfer / record | 7,131 | 35,213 | 500 | NORM, AF, CD, PB, or STA; exactly one group; NORM has no additional diagnosis. |
| CSN | Transfer / record | 20,757 | 62,271 | 500 | One of five retained SNOMED rhythm diagnoses. |
| G12EC | Transfer / record | 6,081 | 18,167 | 500 | NSR, STTA, SB, or LQT; exactly one group; NSR has no additional diagnosis in any retained record. |
| PTB-XL | Transfer / patient | 13,273 | 42,726 | 500 | NORM, MI, STTC, or CD; records agree within patient; the target space excludes HYP. |
| PTB | Transfer / subject | 200 | 16,712 | 1000 | Normal/control versus MI; all records are valid, in target, and concordant within subject. |

**Offline targets and shared bottleneck.** Frozen ECGFounder and Sundial form the default teacher pair; ECG-FM replaces ECGFounder only in the matched ECG-teacher ablation. Native transformations are applied once. The ECG teacher follows its official twelve-lead pathway, while Sundial embeddings are pooled across chunks within each lead and then across leads. Both detached targets are cached under the source-window key. A 0.137M-parameter DSCNN maps the same waveform through a temporal stem, four residual depthwise-separable stages of widths 64/96/128/128, global average pooling, and a 512-dimensional projection with group normalization, GELU, and dropout. Separate linear adapters preserve the two teacher geometries during acquisition; neither adapter nor teacher is exported.

For branch $b \in \{E, T\}$ and batch $B$, let $H = f_\theta(B)$, $U^b = p_b(H)$, and $Z^b$ be the cached target; hats denote row-wise $\ell_2$ normalization. ECGDuo minimizes

$$S_b = \widehat{U}^b \widehat{Z}^{b\top}/\tau, \qquad \mathcal{L}_b = \tfrac{1}{2}\left[\mathrm{CE}(S_b, I) + \mathrm{CE}(S_b^\top, I)\right], \qquad \mathcal{L}_{\mathrm{Duo}} = \tfrac{1}{2}(\mathcal{L}_E + \mathcal{L}_T), \quad \tau = 0.1. \tag{B.1}$$

The two directions retrieve matched teacher and student representations symmetrically; averaging active branches keeps the objective scale fixed in single- versus dual-teacher comparisons. Dataset, subject, record, and window identifiers join signals, targets, folds, embeddings, and predictions. Cardinality, uniqueness, finiteness, and checksums are verified before optimization; a failed join aborts rather than imputes a target or reuses a neighboring window.

### B.2 Optimization, subject-independent transfer, and paired analyses

**Algorithm 1** Offline complementary distillation and teacher-free export

**Require:** indexed source windows $\mathcal{D}_U$; frozen teachers $g_E, g_T$; native transforms $T_E, T_T$; seed $s = 2026$.
**Return:** frozen encoder $f_\theta$; no teacher-side state.

```
Validate canonical waveforms and freeze the corpus/subject/record/window index.
For every x_i ∈ D_U, cache both teacher targets under its immutable window key:
    z_i^E ← stopgrad(g_E(T_E(x_i))).
    z_i^T ← stopgrad(pool(g_T(T_T(x_i)))).
Verify a finite one-to-one join of (x_i, z_i^E, z_i^T); abort on missing or duplicate keys.
Initialize f_θ, p_E, p_T and the merged-source sampler with seed s.
For epochs e = 1, ..., 5, traverse aligned mini-batches B of size 1,024:
    h ← f_θ(x); u^b ← norm(p_b(h)) for b ∈ {E, T}.
    Evaluate both retrieval directions in Equation B.1; set L ← (L_E + L_T)/2.
    Update θ, p_E, p_T with AdamW and cosine learning-rate decay.
    Clip the global gradient norm at 16 before the optimizer step.
After five complete passes, hash the checkpoint, configuration, cache, and source manifest.
Discard g_E, g_T, p_E, p_T and all target caches; return only the frozen encoder f_θ.
```

**Pretraining.** The merged source corpora are traversed for five epochs using AdamW with learning rate $10^{-2}$, weight decay $10^{-4}$, cosine annealing, batch size 1,024, temperature 0.1, and gradient-norm limit 16. Seed 2026 fixes initialization and sample order. Single- and dual-teacher students share windows, architecture, sampler, optimizer, and update budget; only the named branch changes. Teacher inference is offline, and gradients reach the encoder and enabled adapters only.

**Subject-independent evaluation.** Five deterministic class-stratified folds are created once at the subject level and reused by every method. In each rotation, one fold is held out for testing, and a class-stratified 25% of the remaining subjects is used as validation data, giving approximately 60/20/20 train/validation/test proportions. Evaluation uses 3-s windows, which inherit their subject partition. Checkpoints are selected by validation macro F1 with patience 10 and a 50-epoch ceiling; the corresponding test fold is evaluated once. Target mappings, folds, stopping rules, and metric code are frozen before testing.

**Downstream models.** All learners use cross-entropy, AdamW, cosine decay, and weight decay $10^{-4}$. ECGDuo freezes its encoder and fits a linear head with batch 128 and learning rate $10^{-4}$. Frozen-model probes use cached features, batch 512, and learning rate $10^{-3}$; end-to-end baselines use batch 128, evaluation batch 256, learning rate $10^{-4}$, and published architectures. Frozen comparisons therefore measure linear accessibility, while supervised baselines retain end-to-end training.

**Metrics and reporting.** We report Accuracy, macro Precision/Recall/F1, AUROC, and AUPRC. Binary ranking uses positive-class scores; multiclass ranking uses macro one-vs-rest aggregation. Avg is the unweighted mean of the six measures. Fold predictions, checkpoint identifiers, and metric inputs are stored before aggregation; results are reported as mean$\pm$std over the five folds. Hashes are checked before paired joins; incomplete or mismatched folds are reported rather than replaced by cached values.

**Paired analyses.** For label efficiency, only the class-stratified fraction of G12EC and PTB-XL training subjects changes (5/10/25/100%); ECG-only and ECG+TSFM students share selected subjects, frozen feature keys, validation/test folds, and head seeds, and confidence intervals use paired fold/seed differences. For retention, 5,000 subjects per source corpus are excluded before dedicated matched retraining. Centered Linear CKA (Kornblith et al., 2019) is evaluated on 5,000 common held-out windows per corpus against fixed Sundial and matched ECG-teacher targets. The reported contrast, $100 \times [\mathrm{CKA}_{\mathrm{dual}} - \mathrm{CKA}_{\mathrm{single}}]$, therefore varies teacher count while holding data, targets, seed, capacity, and optimization fixed.

**Controlled contrasts.** A gain is attributed to complementary supervision only when paired runs share accepted waveform keys, preprocessing, student capacity, source membership, optimizer settings, downstream partitions, and metric implementation. The deployable checkpoint contains neither a teacher, an adapter, nor cached target state, so transfer and efficiency results describe the common student bottleneck rather than hidden teacher-side computation.

# C Complete Downstream Results

Table 4: **Complete results against supervised task-specific models.** Entries are mean±std over five subject-independent folds. Supervised baselines are trained end-to-end; ECGDuo uses the frozen distilled encoder with a linear head.

| Dataset | Model | Accuracy | Precision | Recall | F1 | AUROC | AUPRC | Avg |
|---|---|---|---|---|---|---|---|---|
| **CPSC2018** (5 classes) | XResNet1D | 53.83±8.50 | 56.49±5.49 | 48.52±9.30 | 47.25±9.30 | 81.85±4.20 | 55.51±6.42 | 57.24±7.20 |
| | Inception1D | 58.20±6.41 | 60.83±5.50 | 52.10±5.84 | 52.54±5.64 | 84.60±3.38 | 60.98±5.44 | 61.54±5.37 |
| | MedGNN | 56.63±3.51 | 55.16±3.38 | 53.52±3.22 | 53.59±3.54 | 83.71±2.26 | 57.52±4.93 | 60.02±3.47 |
| | iTransformer | 42.93±0.68 | 39.85±0.70 | 39.27±0.75 | 39.38±0.69 | 71.17±0.64 | 39.92±0.74 | 45.42±0.70 |
| | Medformer | 50.37±0.83 | 47.84±0.71 | 47.41±1.16 | 47.49±0.90 | 77.04±0.52 | 48.30±0.92 | 53.07±0.84 |
| | PatchTST | 59.84±1.48 | 57.33±1.16 | 56.87±2.03 | 56.98±1.57 | 85.03±0.83 | 60.01±1.71 | 62.68±1.46 |
| | ECGDuo (Ours) | 75.30±1.09 | 73.79±1.22 | 74.01±1.28 | 73.82±1.17 | 93.34±0.52 | 79.38±1.30 | 78.27±1.10 |
| **CSN** (5 classes) | XResNet1D | 43.34±9.15 | 60.32±9.10 | 35.89±4.69 | 32.92±5.88 | 85.21±7.69 | 64.19±10.99 | 53.65±7.92 |
| | Inception1D | 53.38±8.93 | 69.63±3.81 | 48.72±2.80 | 46.70±3.38 | 85.74±4.16 | 64.72±5.33 | 61.48±4.74 |
| | MedGNN | 47.88±6.34 | 53.94±3.92 | 49.17±1.24 | 41.03±2.37 | 81.14±5.60 | 55.46±6.52 | 54.77±4.33 |
| | iTransformer | 84.25±0.75 | 71.53±1.07 | 68.20±1.57 | 69.07±1.39 | 92.55±0.48 | 71.24±1.44 | 76.14±1.12 |
| | Medformer | 68.97±0.82 | 57.16±1.68 | 54.60±0.62 | 55.50±0.91 | 85.46±0.61 | 57.80±1.18 | 63.25±0.97 |
| | PatchTST | 76.67±1.16 | 67.83±1.14 | 65.51±1.00 | 66.65±1.09 | 90.92±0.92 | 70.60±1.41 | 73.03±1.12 |
| | ECGDuo (Ours) | 91.07±0.30 | 88.64±0.83 | 84.38±0.53 | 85.79±0.51 | 97.45±0.11 | 89.65±0.23 | 89.50±0.42 |
| **G12EC** (4 classes) | XResNet1D | 64.42±4.12 | 61.42±3.04 | 54.66±6.25 | 54.97±6.60 | 84.17±1.02 | 63.23±3.15 | 63.81±4.03 |
| | Inception1D | 62.18±7.18 | 54.71±6.48 | 52.28±9.20 | 52.30±8.87 | 80.07±5.59 | 56.49±8.33 | 59.67±7.61 |
| | MedGNN | 61.60±3.22 | 56.46±3.33 | 48.37±2.61 | 48.19±3.44 | 79.05±1.63 | 53.98±2.63 | 57.94±2.81 |
| | iTransformer | 66.27±1.38 | 57.49±0.69 | 56.93±0.80 | 56.94±0.62 | 80.46±0.19 | 60.41±0.44 | 63.08±0.69 |
| | Medformer | 55.99±1.51 | 45.84±1.04 | 44.77±0.96 | 44.93±0.96 | 71.23±0.92 | 46.09±0.98 | 51.48±1.06 |
| | PatchTST | 72.59±2.02 | 65.02±2.54 | 64.20±1.37 | 64.28±1.71 | 86.98±0.72 | 68.05±1.41 | 70.19±1.63 |
| | ECGDuo (Ours) | 79.72±0.95 | 75.94±2.21 | 71.36±0.78 | 72.56±1.28 | 92.71±0.80 | 79.10±1.49 | 78.56±1.25 |
| **PTB-XL** (4 classes) | XResNet1D | 66.35±1.97 | 64.83±9.45 | 39.23±4.80 | 40.59±5.58 | 80.85±6.85 | 57.43±9.92 | 58.21±6.43 |
| | Inception1D | 73.57±3.43 | 70.55±8.49 | 55.39±5.53 | 59.04±7.52 | 87.91±3.60 | 70.92±6.27 | 69.56±5.81 |
| | MedGNN | 75.21±1.27 | 67.87±2.92 | 60.60±2.95 | 63.03±2.18 | 87.32±3.16 | 67.72±4.51 | 70.29±2.83 |
| | iTransformer | 65.12±1.06 | 51.95±2.13 | 47.68±1.04 | 49.24±1.46 | 77.35±0.94 | 49.74±1.68 | 56.85±1.39 |
| | Medformer | 66.88±0.86 | 55.24±1.82 | 50.63±1.89 | 52.03±1.09 | 78.42±1.79 | 52.86±1.51 | 59.34±1.49 |
| | PatchTST | 74.41±0.47 | 66.13±0.77 | 61.58±0.95 | 63.46±0.77 | 86.11±0.28 | 66.46±0.80 | 69.69±0.67 |
| | ECGDuo (Ours) | 82.57±0.77 | 78.40±1.44 | 73.56±1.20 | 75.69±1.07 | 93.74±0.42 | 82.40±1.16 | 81.06±1.01 |
| **PTB** (2 classes) | XResNet1D | 79.45±3.79 | 70.10±11.05 | 61.94±4.65 | 62.02±5.26 | 71.59±7.02 | 90.27±2.90 | 72.56±5.78 |
| | Inception1D | 78.52±3.83 | 66.21±10.74 | 64.20±6.60 | 63.16±6.56 | 73.37±11.34 | 89.60±6.05 | 72.51±7.52 |
| | MedGNN | 78.60±6.62 | 60.87±13.07 | 61.69±6.70 | 60.18±9.82 | 71.72±10.46 | 90.35±5.20 | 70.57±8.65 |
| | iTransformer | 80.97±4.95 | 69.94±7.87 | 68.00±3.61 | 68.13±5.62 | 77.72±2.98 | 92.58±2.04 | 76.22±4.51 |
| | Medformer | 78.05±2.93 | 63.32±6.55 | 62.57±6.05 | 62.86±6.34 | 68.54±6.73 | 89.45±1.38 | 70.80±5.00 |
| | PatchTST | 82.96±2.58 | 72.46±7.20 | 73.29±4.63 | 72.35±5.33 | 84.83±4.45 | 95.92±1.00 | 80.30±4.20 |
| | ECGDuo (Ours) | 92.12±1.64 | 88.74±2.51 | 84.99±3.05 | 86.37±1.35 | 96.36±1.47 | 98.99±0.66 | 91.26±1.78 |

Table 5: **Complete results against frozen foundation models (linear probe).** Entries are mean±std over five subject-independent folds. All encoders are frozen and only a linear classifier is fitted.

| Dataset | Model | Accuracy | Precision | Recall | F1 | AUROC | AUPRC | Avg |
|---|---|---|---|---|---|---|---|---|
| **CPSC2018** (5 classes) | Sundial | 72.22±1.89 | 70.48±1.91 | 70.96±2.06 | 70.72±1.92 | 91.91±0.93 | 76.88±1.97 | 75.53±1.78 |
| | ECG-FM | 73.66±0.76 | 71.96±0.96 | 73.42±1.45 | 72.11±1.08 | 92.46±0.58 | 77.55±1.47 | 76.86±1.05 |
| | ECGFounder | 75.40±2.17 | 73.93±1.86 | 74.81±1.48 | 73.66±2.32 | 93.11±0.60 | 80.66±1.34 | 78.59±1.63 |
| | ECGDuo (Ours) | 75.30±1.09 | 73.79±1.22 | 74.01±1.28 | 73.82±1.17 | 93.34±0.52 | 79.38±1.30 | 78.27±1.10 |
| **CSN** (5 classes) | Sundial | 90.77±0.31 | 86.68±0.71 | 82.62±0.35 | 83.76±0.36 | 96.91±0.10 | 87.08±0.40 | 87.97±0.37 |
| | ECG-FM | 90.14±1.98 | 86.47±3.54 | 82.41±2.19 | 83.58±2.55 | 96.56±0.86 | 86.52±3.02 | 87.61±2.36 |
| | ECGFounder | 92.33±0.24 | 90.47±1.41 | 84.76±1.61 | 86.01±1.03 | 98.19±0.09 | 90.43±0.19 | 90.37±0.76 |
| | ECGDuo (Ours) | 91.07±0.30 | 88.64±0.83 | 84.38±0.53 | 85.79±0.51 | 97.45±0.11 | 89.65±0.23 | 89.50±0.42 |
| **G12EC** (4 classes) | Sundial | 77.43±1.65 | 75.90±2.10 | 70.95±1.95 | 72.31±1.70 | 92.22±0.67 | 78.12±1.72 | 77.82±1.63 |
| | ECG-FM | 78.88±2.76 | 75.56±3.97 | 71.33±3.29 | 72.42±3.53 | 92.43±1.59 | 77.65±3.31 | 78.05±3.07 |
| | ECGFounder | 82.35±0.87 | 79.64±1.70 | 78.09±1.54 | 77.97±1.03 | 92.50±0.82 | 84.07±1.62 | 82.44±1.26 |
| | ECGDuo (Ours) | 79.72±0.95 | 75.94±2.21 | 71.36±0.78 | 72.56±1.28 | 92.71±0.80 | 79.10±1.49 | 78.56±1.25 |
| **PTB-XL** (4 classes) | Sundial | 79.16±0.52 | 75.14±1.20 | 66.59±1.39 | 69.76±0.93 | 91.28±0.27 | 77.01±0.92 | 76.49±0.87 |
| | ECG-FM | 79.21±3.33 | 75.39±4.58 | 67.14±3.01 | 70.54±3.87 | 91.00±2.02 | 77.24±5.08 | 76.75±3.65 |
| | ECGFounder | 82.34±0.55 | 79.71±1.55 | 73.55±3.00 | 74.61±1.55 | 93.74±0.23 | 83.59±0.43 | 81.26±1.22 |
| | ECGDuo (Ours) | 82.57±0.77 | 78.40±1.44 | 73.56±1.20 | 75.69±1.07 | 93.74±0.42 | 82.40±1.16 | 81.06±1.01 |
| **PTB** (2 classes) | Sundial | 87.55±1.81 | 80.96±4.15 | 78.24±2.83 | 78.61±3.35 | 88.98±2.63 | 95.33±1.83 | 84.94±2.77 |
| | ECG-FM | 87.39±2.10 | 80.12±6.80 | 85.14±3.41 | 81.04±3.89 | 92.23±4.11 | 96.36±2.43 | 87.05±3.79 |
| | ECGFounder | 87.67±5.27 | 81.31±9.14 | 84.77±5.43 | 81.39±7.65 | 92.76±3.13 | 96.06±0.73 | 87.33±5.23 |
| | ECGDuo (Ours) | 92.12±1.64 | 88.74±2.51 | 84.99±3.05 | 86.37±1.35 | 96.36±1.47 | 98.99±0.66 | 91.26±1.78 |

## AI USE STATEMENT

Large language models, including Claude Code, were used only for AI-assisted drafting, editing, and literature checks. They did not influence experimental design, execution, analysis, or interpretation. All scientific decisions, analyses, and results were produced and fully verified solely by the authors.