BenchECG and xECG: a benchmark and baseline for ECG foundation models

Riccardo Lunelli[1†], Angus Nicolson[1†], Samuel Martin Pröll[1], Sebastian Johannes Reinstadler[2], Axel Bauer[2], Clemens Dlaska[1*]

[1]Digital Cardiology Lab, University Clinic of Internal Medicine III, Medical University Innsbruck, A-6020 Innsbruck, Austria.
[2]University Clinic of Internal Medicine III, Cardiology and Angiology, Medical University Innsbruck, A-6020 Innsbruck, Austria.

*Corresponding author(s). E-mail(s): clemens.dlaska@i-med.ac.at;
†These authors contributed equally to this work.

Abstract

Electrocardiograms (ECGs) are inexpensive, widely used, and well-suited to deep learning. Recently, interest has grown in developing foundation models for ECGs – models that generalise across diverse downstream tasks. However, consistent evaluation has been lacking: prior work often uses narrow task selections and inconsistent datasets, hindering fair comparison. Here, we introduce BenchECG, a standardised benchmark comprising a comprehensive suite of publicly available ECG datasets and versatile tasks. We also propose xECG, an xLSTM-based recurrent model trained with SimDINOv2 self-supervised learning, which achieves the best BenchECG score compared to publicly available state-of-the-art models. In particular, xECG is the only publicly available model to perform strongly on all datasets and tasks. By standardising evaluation, BenchECG enables rigorous comparison and aims to accelerate progress in ECG representation learning. xECG achieves superior performance over earlier approaches, defining a new baseline for future ECG foundation models.

Keywords: ECG, Foundation Model, Benchmark, Self-supervised learning

Cardiovascular diseases (CVDs) are the leading global cause of death, and early and accurate diagnosis is critical to reducing their burden [1]. Electrocardiograms (ECGs) are inexpensive, non-invasive, and widely available biosignals that integrate complex biological information of the entire organism and are therefore used to detect a broad spectrum of cardiac and non-cardiac conditions. These characteristics make ECGs not only central to routine care but also attractive candidates for automated analysis using machine learning.

While deep learning models have demonstrated strong performance in interpreting ECGs, when trained on large, labelled datasets [4, 5], the emergence of foundation models represents a paradigm shift for medical artificial intelligence (AI) [6]. Foundation models are trained on vast, diverse datasets – often unlabelled – and designed to be general-purpose and adaptable to multiple downstream tasks. Recently, several works have proposed foundation models for ECGs that aim to support a variety of downstream tasks across datasets, tasks, and signals [5, 7–9]. These models typically use self-supervised learning (SSL) approaches and transformer-based architectures [8, 9] that can be trained on vast unlabelled datasets.

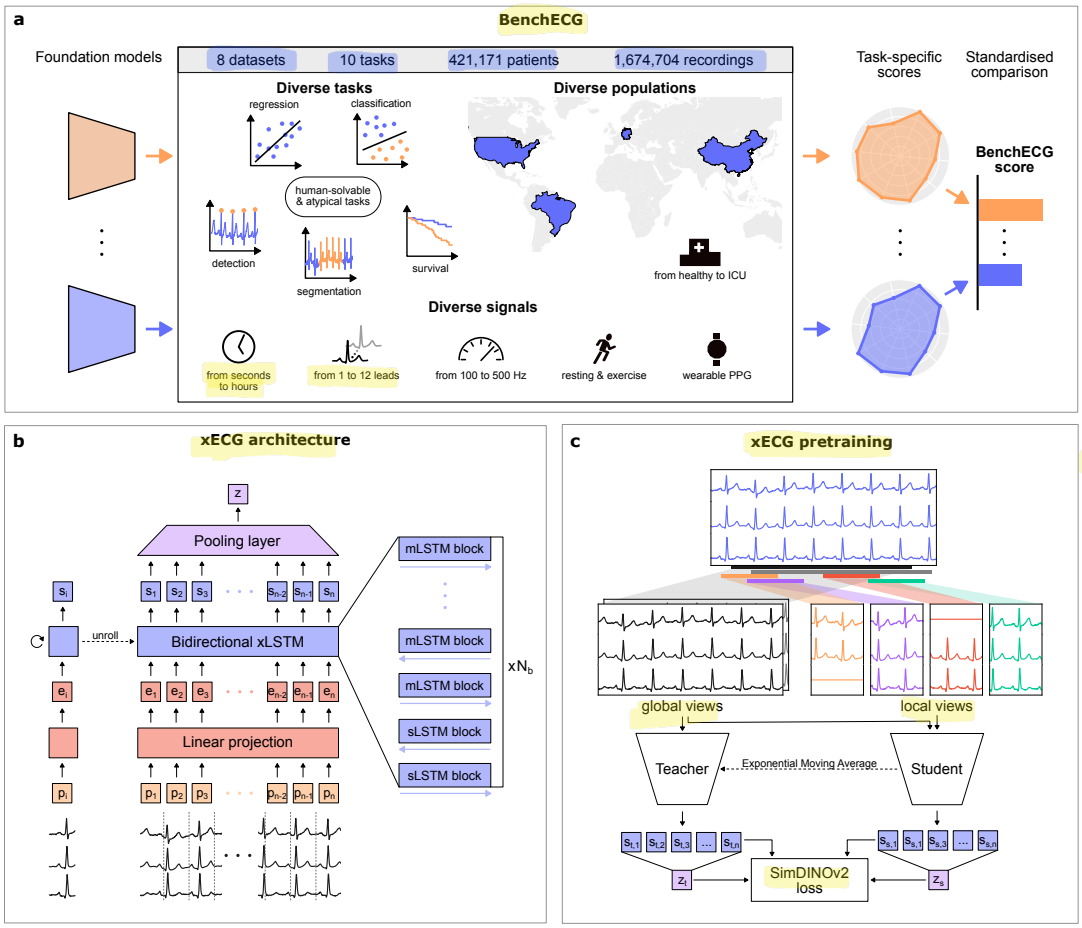

Fig. 1 Overview of BenchECG and xECG. BenchECG (**a**) provides a comprehensive evaluation of ECG foundation models with diverse signals, populations and tasks. xECG (**b**) is a bi-directional recurrent model based on the xLSTM [2] architecture and pretrained via SimDINOv2 [3] self-supervised pretraining (**c**).

Despite promising results, the field faces two major limitations. First, existing evaluation practices are inconsistent, reflecting a broader "reproducibility crisis" in machine learning applied to the sciences [10–14]. Prior work varies significantly in which tasks and datasets are used for validation, making it difficult to compare methods or identify generalisable approaches [7, 15]. Data leakage, improper data preprocessing, and high computational costs often hinder independent replication and verification of results [13]. Given the high-stakes nature of medical diagnostics, there is a critical need for trustworthy and transparent AI evaluation. Second, as we show in this work, the architectural choice of using transformer-based models limits performance in long-context tasks, relevant for real-world settings involving long-term monitoring. While, in principle, powerful for capturing long-range dependencies, standard transformers scale quadratically in time and memory complexity with respect to sequence length, making them computationally expensive and memory-intensive for very long time-series data – a significant limitation given that important clinical applications require analysis of up to hours or days of ECG signals.

A wide range of machine learning fields have addressed similar challenges through the development of standardised, public benchmarks. From well-known benchmarks in vision [16, 17] or language [18, 19], to more specific applications in molecular sequences [20], and federated learning [21]. These benchmarks have enabled robust evaluation, driven architectural innovation, and allowed for rapid progress. In this work, we argue that ECG foundation models would benefit from a similar ecosystem. To this end, we introduce **BenchECG**, a comprehensive benchmark that spans eight publicly available datasets and ten diverse and representative

tasks. BenchECG provides a unified platform for evaluating generalisation, task diversity, and representation quality in ECG foundation models.

We also introduce **xECG**, a novel ECG foundation model based on the extended long short-term memory (xLSTM) architecture [2]. xLSTM models combine the benefits of transformers (parallelisable networks that improve with scale) with the efficiency of recurrent structures, allowing for processing of longer ECG sequences, and have recently been shown to excel at signal forecasting [22, 23]. We pretrain xECG using SimDINOv2 [3], a state-of-the-art SSL method originally developed for computer vision, adapted here to the time-series domain. We show that this combination yields strong, general-purpose ECG representations, providing out-of-the-box flexibility covering all recording scenarios present in cardiac monitoring.

Across BenchECG, xECG achieves the best average rank when finetuning (1.50) and under linear probing (1.20), outperforming prior publicly available state-of-the-art models. xECG is the only model to perform strongly across all BenchECG datasets and task types.

In summary, the contributions of this paper are:

- We introduce BenchECG, the first comprehensive, standardised benchmark for ECG foundation models, enabling rigorous and reproducible evaluation
- We propose xECG, an ECG foundation model that combines the efficiency of xLSTMs with the representational strength of SimDINOv2 self-supervised learning
- We demonstrate that xECG achieves state-of-the-art performance across diverse ECG tasks, establishing a strong and reproducible baseline for future work
- We release code, models, and benchmark tasks to support open research in general-purpose ECG representation learning

A standardised benchmark with clinically relevant tasks

Foundation models, by definition, should be adaptable to many downstream tasks [24]. Hence, we require ECG foundation models to handle variety in three different forms: (1) signal characteristics, (2) dataset characteristics, and (3) task characteristics. An ECG foundation model should be evaluated not only on conventional 12-lead clinical ECGs but also on long-term, few-lead, and related (but non-ECG) physiological signals, across geographically and clinically distinct cohorts, and over a broad set of clinically relevant downstream tasks. To this end, we introduce **BenchECG**, an open benchmark designed to test ECG foundation models made of eight publicly available datasets, comprising 421,171 patients, and a total of 1,674,704 recordings (for an overview see Fig. 1a and Table 1).

BenchECG encompasses signal types used in cardiovascular diagnosis, monitoring and beyond (see Table 1). Besides short 12-lead ECGs (PTB-XL [25], CPSC2018 [26], MIMIC-IV-ECG [27–30], CODE-15% [31]), longer recordings are represented with 30-minute 2-lead signals (MIT-BIH [32]) and single-lead overnight recordings (Apnea-ECG [33]). By sourcing data from many datasets we also encounter different sampling rates (100-500 Hz) as well as different recording conditions ranging from routine ambulatory settings to high-intensity exercise examinations (Exercise-ECG [34]). In addition, wearable photoplethysmography (PPG) waveforms (DeepBeat [35]) are included to test generalisation to other cardiac biosignals.

BenchECG includes cohorts from distinct geographies (Europe, USA, China, Brazil) and populations (healthy individuals to critically ill patients). The datasets also vary in size with the number of participants spanning five orders of magnitude (20 to 233,770 individuals). This ensures the evaluation of models in both low-data and high-data scenarios.

Foundation models should be able to complete a wide variety of downstream tasks, yet key previous works solely evaluate on classification tasks [8, 9]. To provide a more comprehensive picture of model performance, BenchECG includes different types of clinically relevant tasks:

- **Classification**: multilabel classification of diagnostic labels (PTB-XL, CPSC2018), heartbeat-level arrhythmia classification (MIT-BIH), AF classification from PPG signals (DeepBeat), and multilabel classification for simultaneous (ab)normality assessment of various blood test values (MIMIC-IV-ECG).
- **Segmentation**: sleep apnea segmentation in overnight ECG recordings (Apnea-ECG).

- **Detection**: R-peak detection in standard ECGs (MIT-BIH) and exercise ECGs (Exercise-ECG).
- **Regression**: age estimation across populations (finetuned on CODE-15%, evaluated on PTB-XL, CPSC2018 and MIMIC-IV-ECG)
- **Survival analysis**: mortality risk prediction across populations (finetuned on CODE-15%, evaluated on MIMIC-IV-ECG)

By encompassing both *typical* tasks (i.e., human-solvable tasks like arrhythmia classification, R-peak detection) and *atypical* (i.e., machine-learning-enabled tasks and tasks traditionally not ECG-associated like laboratory value classification and age estimation), BenchECG provides a rigorous test bed for general-purpose ECG representation learning. The evaluation of generalisability is further strengthened by specifically testing out-of-distribution (OOD) performance. For instance, the age estimation task is finetuned on a Brazilian population and evaluated on datasets from different countries (China, Germany and the USA). Population and prevalence shifts are also accounted for in the mortality risk prediction task where models are finetuned on a general population cohort (CODE-15%, 5-year mortality 4.7%) and evaluated on ICU patients (MIMIC-IV, 5-year mortality 36%). Furthermore, models are pretrained on resting ECGs, whereas evaluation on the high intensity exercise ECG task assesses generalisation across substantially different signals. Finally, in the AF classification from PPG task, we can test a model's adaptability to a completely new modality.

For a fair comparison, models that wish to be evaluated on BenchECG should not be pretrained on any of the evaluation datasets (Table 1). Note that the CODE-15% dataset is not used in BenchECG evaluation, but is used during finetuning for the age estimation and mortality prediction tasks. This is because it is commonly used in pretraining of self-supervised methods [8, 9] and is a subset of the CODE dataset [36], used in xECG pretraining.

xECG

We introduce xECG, a novel ECG foundation model based on the recently introduced extended long short-term memory (xLSTM) architecture [2]. Designed to overcome key limitations of transformer-based models, xECG enables efficient processing of long ECG sequences, while retaining strong representational capacity through self-supervised pretraining. The model architecture and training procedure are shown in Figure 1b and 1c.

Standard long short-term memory (LSTM) models have limitations in capacity and scalability: limited storage due to scalar cell states, sequential dependencies that prevent parallelism, and rigid gating dynamics that hinder memory flexibility [2]. Transformers [37] address these constraints but introduce quadratic complexity in sequence length, which is impractical for high-resolution ECG signals recorded over long durations. To address this, xLSTM uses both the scalar (sLSTM) and matrix (mLSTM) memory blocks, as well as exponential gating and multi-head memory mechanisms introduced by Beck et al. [2]. These innovations allow xLSTMs to scale linearly in sequence length and, unlike standard LSTMs, support training-time parallelism.

The xLSTM architecture has demonstrated strong performance in various sequence modelling tasks [22, 23], including bio-signal generation [38] and ECG analysis [39]. This extends to computer vision, where the Vision-LSTM (ViL) [40] architecture processes image patches bidirectionally with mLSTM blocks, achieving a better performance-to-cost trade-off than transformer-based approaches.

We designed xECG with a stack of alternating sLSTM and mLSTM layers arranged bidirectionally. This contrasts with mLSTM-only designs such as ViL [40], as recent evidence suggests a mixed block architecture improves ECG modelling [39]. Specifically, at each layer, one block processes the sequence forward and the other in reverse – see Figure 1b. This design allows the architecture to aggregate information efficiently both forwards and backwards in time.

Each input ECG is divided into non-overlapping temporal patches and projected into an embedding space before being passed to the xLSTM encoder. These patch-level representations can then be flexibly adapted for downstream tasks: pooled for signal-level classification and regression tasks, or directly used for beat-level classification, detection and segmentation tasks.

Dataset	Leads	Participants	Recordings	Length	Task(s)
Apnea-ECG [33]	1	70	70	7-10 h	Segmentation of signal into apnea vs. non-apnea
CPSC2018 [26]	12	6,877	6,877	6-60 s	Multilabel classification of diagnostic classes; age regression (OOD)
DeepBeat [35]	1	169	500,000	25 s	Photoplethysmography AF classification (OOD)
Exercise-ECG [34]	1	20	100	20 s	R-peak detection under exercise conditions (OOD)
MIMIC-IV-ECG [27, 28]	12	161,352	800,035	10 s	Multilabel classification of diagnostic classes, age regression (OOD); Blood test multilabel (a)bnormality classification; survival analysis (OOD)
MIT-BIH [32]	2	44	44	30 min	Multilabel heartbeat-level arrhythmia classification; R-peak detection
PTB-XL [25]	12	18,869	21,799	10 s	Multilabel classification of diagnostic classes; age estimation
CODE-15%	12	233,770	345,779	7-10 s	Training set for age regression and survival analysis.

Table 1 BenchECG datasets and tasks. BenchECG includes a diverse suite of publicly available datasets varying in signal length, number of leads, and task type. Tasks explicitly addressing out-of-distribution aspects are highlighted as OOD. The number of PPGs provided in the DeepBeat dataset include augmentations. CODE-15% is exclusively used for training as it is included in pretraining datasets.

Supervised models are limited by the availability and quality of labelled ECG data. In contrast, self-supervised learning (SSL) enables models to scale to much larger unlabelled datasets by generating supervisory signals directly from the data [24]. Different SSL methods use different strategies: SimCLR [41] contrasts augmented versions of the same signal against others in the batch; Masked Data Modeling (MDM) methods, like Masked Autoencoders [8, 42], reconstruct masked portions of the input; and Joint Embedding Predictive Architectures (JEPA) [9, 43] mask features and predict representations in latent space. Features learnt by SSL tend to be more robust across tasks, particularly in the low-data regime, when compared to supervised methods as they do not overfit on the specific labels used in training [44]. These properties make SSL an attractive pretraining strategy for ECG foundation models designed to generalise across diverse signals, patient populations and tasks.

To pretrain xECG, we adopt an SSL approach using SimDINOv2 [3], a recent variant of the self-distillation with no labels (DINO) framework [45, 46] designed for training stability and reduced hyperparameter sensitivity. Previous work in the medical domain using SSL has typically relied on contrastive methods [41, 47, 48], which require large batch sizes and careful augmentation design. By contrast, SimDINOv2 is based on a non-contrastive teacher-student architecture, in which the student network learns to match representations produced by a slowly evolving teacher model.

In this work, we adapt SimDINOv2 to ECG signals and apply a similar multi-view strategy: each training sample is augmented into multiple global and local views, reflecting different temporal crops and physiologically realistic perturbations. However, by including pretraining datasets that have multiple ECGs per patient, we can use global and local views across different signals from the same patient. The student is trained to produce consistent embeddings across these views, while maintaining high feature diversity via a coding rate regulariser.

The result is a scalable, general-purpose ECG encoder capable of producing rich representations across a wide range of time scales, patient populations, and downstream tasks. When evaluated on BenchECG, xECG achieves the highest overall performance, setting a new baseline for ECG foundation models.

Results

In addition to xECG, we evaluate state-of-the-art ECG foundation models with publicly available pretrained weights on BenchECG: ST-MEM [8], ECG-JEPA [9], and ECGFounder [5]. Further ablations are conducted using a transformer pretrained via our xECG pretraining strategy (SimDINOv2 Transformer) and a supervised xLSTM. (see Methods for an overview of each model).

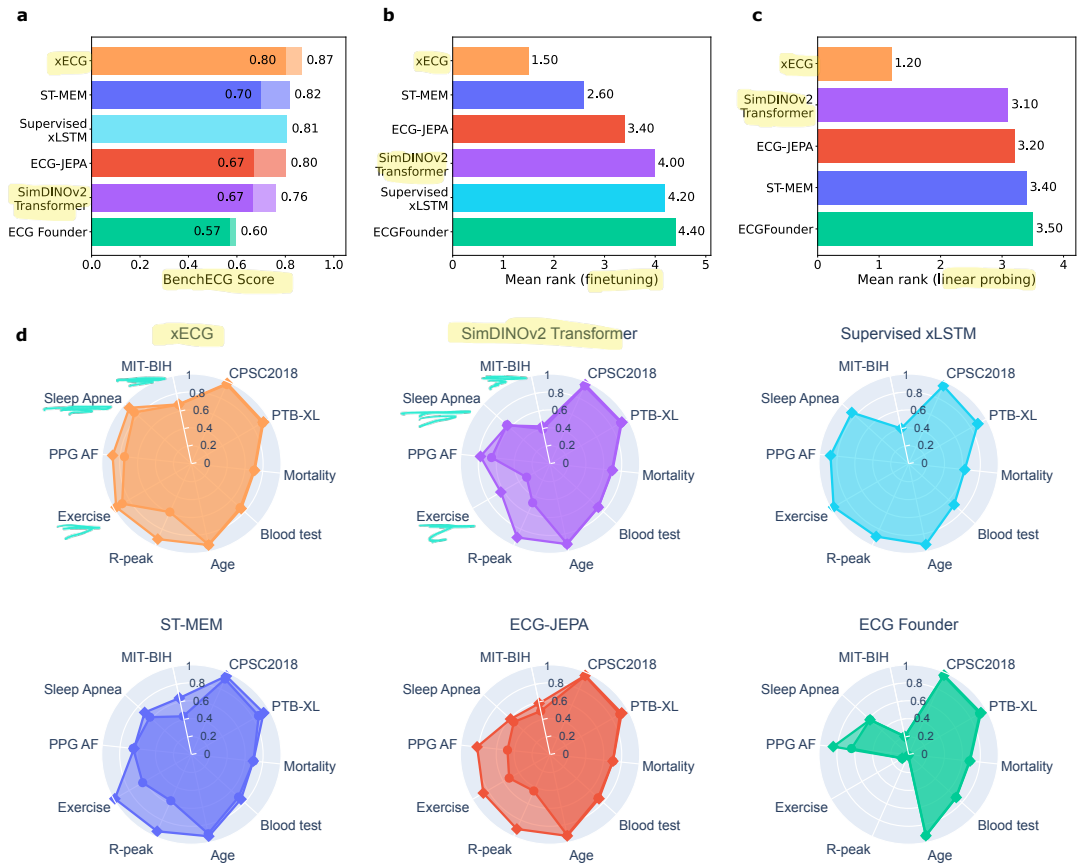

Fig. 2 Performance of state-of-the-art ECG foundation models. a, BenchECG score for state-of-the-art ECG foundation models (and supervised xLSTM baseline) evaluated using finetuning (light colour) and linear probing (full colour). **b,c,** Mean rank of evaluated models across BenchECG tasks for finetuning (**b**) and linear probing (**c**). **d**, Radar plots showing performance of models on individual tasks, both for fine-tuning (diamonds) and linear probing (circles).

For each method, we train the models five times (via both finetuning and linear probing for pretrained models) with different random seeds for batch collection and linear head initialisation, and we report the mean and standard deviation across runs. For CODE-15% and Sleep-Apnea-ECG, which have no published validation split, we used a different random train/-validation split for each run, but for the other datasets we used the validation split consistent with the literature (see Methods for full details).

Comparing foundation models - BenchECG Score

To compare different foundation models, we propose the **BenchECG score**: the mean performance of a model across all tasks in the BenchECG benchmark. For each task we select a metric normalised to the range of 0 to 1. We use area under the receiver operator characteristic curve (AUROC) for classification[1] and segmentation tasks, the symmetric mean absolute percentage error (SMAPE) for regression, F1 score for detection tasks, and the concordance index (C-index) for the survival analysis task (for a detailed justification of metric choices see Methods).

Figure 2a shows the BenchECG score for the different foundation models. xECG achieves the highest BenchECG score of 0.868 ± 0.0030, corresponding to an average rank of 1.50. Results for individual tasks are available in Figure 2d and detailed in Supplementary information. For ranking purposes, pairwise differences between models on the per-task evaluation metrics were

[1] Apart from for the MIT-BIH task where we follow previous works and use F1 score [49–53].

assessed using two-sided Welch's t-tests across independent runs, with significance defined as $p < 0.05$.

xLSTMs outperform in long-context tasks

Given the recurrent nature of xLSTMs and their ability to efficiently process longer signals, they are suited for tasks requiring extended temporal context than transformer- or CNN-based models, which here were limited to fixed 10s inputs. This is apparent in the sleep apnea task, where understanding the variability in a patient's ECG over long time periods is crucial. The supervised xLSTM achieved a higher AUROC than the next best foundation model, ST-MEM (0.853 ± 0.022 vs. 0.702 ± 0.020, $p = 0.000004$), while xECG achieved the highest AUROC overall (0.932 ± 0.014), significantly outperforming ST-MEM ($p = 0.0000001$) and the supervised xLSTM ($p = 0.0003$).

The MIT-BIH arrhythmia classification task also involves long-context inputs with 30-minute ambulatory ECGs. In terms of F1 score, ST-MEM outperforms prior methods reported in the literature [49–53], but is itself surpassed by xECG (0.644 ± 0.007 vs. 0.677 ± 0.025, $p = 0.040$, comparing ST-MEM to xECG). This highlights the benefit of large-scale self-supervised pretraining. However, the gap is larger under linear probing, where xECG achieves an F1 score of 0.674 ± 0.013 compared to 0.436 ± 0.036 for ST-MEM ($p = 0.000032$), indicating that the pretrained xLSTM features capture more relevant information for long-context tasks. Supplementary Table 4 reports the scores for all methods.

Foundation models solve standard evaluation tasks

On shorter 12-lead clinical recordings, which are the focus of most prior work [5, 8, 9], performance differences between models were narrower. On PTB-XL, all foundation models were within 0.008 AUROC of each other (between 0.923 and 0.931) after finetuning, while on CPSC2018 the spread was wider at 0.023 (between 0.958 and 0.981), with xECG achieving the highest score (AUROC 0.981). There were substantially larger gaps between models when linear probing, with AUROCs varying between 0.867 and 0.917 for PTB-XL and 0.922 and 0.968 for CPSC2018 – a difference of 0.050 and 0.046, respectively. For the complete set of scores see Supplementary Tables 2 and 3.

Generalisation across populations

To test cross-population generalisation, BenchECG includes tasks where models were finetuned and evaluated on different datasets. For the age estimation task, each model was trained on CODE-15% and tested on MIMIC-IV-ECG, PTB-XL and CPSC2018. These datasets are collected from Brazil, the USA, Germany, and China, respectively. ST-MEM has the lowest mean absolute error (MAE) across all test sets (mean MAE 8.546 ± 0.070 years), although xECG has the smallest generalisation gap (difference between validation performance on CODE-15% and mean test performance) at 1.77 years – see Figure 3a for prediction distributions.

For the survival analysis task, models were finetuned on CODE-15%, a general patient population in Brazil (5-year mortality of 4.7%) and evaluated on MIMIC-IV-ECG, an ICU population in the USA (5-year mortality of 36%). This tests each model's ability to generalise across different patient populations beyond country of origin. In all cases, the C-index decreased for MIMIC-IV-ECG compared to CODE-15%, but remained above 0.6, suggesting the models retain some predictive value across populations.

To demonstrate the mortality task's clinical relevance, we performed an additional analysis on risk predictions of the xECG model – see Figure 3c. Patients in the MIMIC-IV-ECG dataset were split into high risk, baseline, and low risk groups using the 25th and 75th percentile of the model's risk scores. Adjusting for the age and sex, we fit a cox proportional hazards model and found a hazard ratio of 0.389 (0.383-0.395 95%CI) and 2.10 (2.09-2.12 95%CI) for the low and high risk groups, respectively. Meaning the high risk group is at twice as much risk as the baseline.

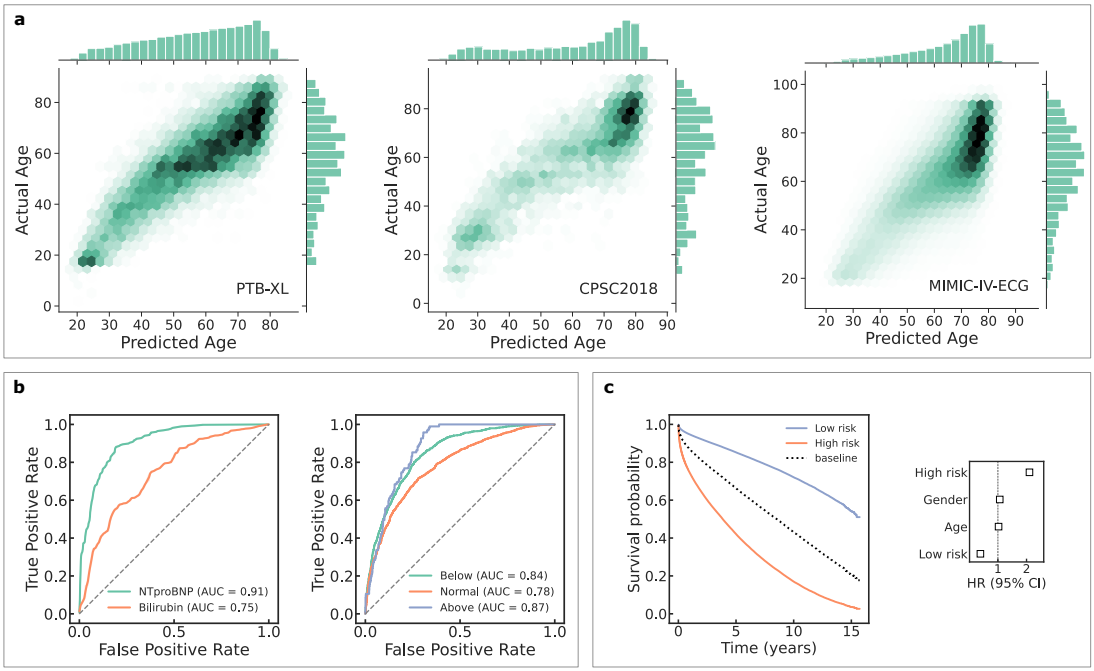

Fig. 3 Additional results for xECG. **a,** The distribution of age predictions on the three test sets. **b,** Performance for NT-proBNP and Bilirubin normal/abnormal classification (left) alongside Albumin abnormally low/normal/abnormally high classification (right) on MIMIC-IV-ECG. **c,** Kaplan Meier curves and hazard ratios for the age/sex adjusted mortality risk on MIMIC-IV-ECG.

Generalisation across tasks

Most of the foundation models adapted well to PPG, an ECG-related modality they were not pretrained on. However, there were large differences in behaviour across models. For example, ST-MEM performed well under linear probing, ranking 2nd, but saw no improvement upon finetuning and had the lowest AUROC, significantly lower than the second-worst model, SimDINOv2 Transformer (AUROC 0.643 ± 0.049 vs 0.780 ± 0.014, $p = 0.0025$). In contrast, ECG-JEPA's frozen features were equivalent to random predictions at AF detection when linear probing (AUROC of 0.477 ± 0.006), but substantially improved after finetuning (AUROC of 0.818 ± 0.021). Supplementary Table 8 summarises the scores for all methods.

For the blood test prediction task performance was limited, with a highest mean AUROC of 0.747 (for xECG). The only blood test result exceeding an AUROC of 0.8 was the prediction of normal/abnormal NT-proBNP, a biomarker used in diagnosing heart failure. This leaves room for improvement for future foundation models. Supplementary Table 11 summarises the scores for all methods.

R-peak detection was finetuned and evaluated on two different datasets: MIT-BIH and Exercise-ECG. If we report F1 score for whether the R-peak predictions are within 150ms of the ground truth, as in previous work [54], the transformer and xLSTM-based models are near-perfect with F1 scores greater than 0.99. Hence, we propose using a more challenging metric, requiring the models to be more precise with their estimates, by reducing the window for a correct R-peak to 20ms. ECGFounder, as a CNN-based foundation model performed poorly (F1 score of 0.011 ± 0.001 and 0.083 ± 0.002 for MIT-BIH and Exercise-ECG, respectively). In contrast, ST-MEM and xECG performed exceptionally well with an F1 greater than 0.9 on both datasets, especially considering the training set for Exercise-ECG consists of only eight patients.

xECG features require less tuning

Linear probing experiments consistently demonstrated that xECG features generalise better across tasks compared to the other foundation models, with xECG ranking first in all tasks

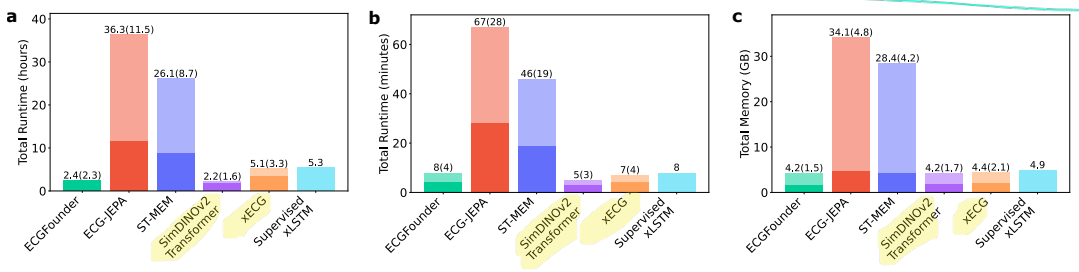

Fig. 4 Runtime and memory consumption for task-specific adaptation of foundation models. Total runtime required for finetuning (shaded color) and linear probing (full color) across all BenchECG tasks (**a**) and PTB-XL task (**b**) on a single Nvidia L40S GPU. (**c**) Corresponding memory consumption of different models for the PTB-XL task.

($p < 0.05$), except mortality risk prediction and PTB-XL, where it ranked second. This led to a mean rank of 1.2 across the benchmark. This advantage is particularly clear when comparing the difference in BenchECG score for finetuning and linear probing for each method (Figure 2a).

Computational Cost

Finally, we compared computational efficiency. Transformer-based models scale quadratically with input length, making them costly to apply to long ECGs. In contrast, xLSTMs scale linearly, allowing them to process longer recordings with substantially reduced memory and time requirements. On the same hardware[2], xECG required $\sim 5\times$ less finetuning time across the benchmark than ST-MEM (5.1 hours vs. 26.1 hours), while achieving higher performance (see Fig. 4). To have a more standardised comparison, we finetuned all the models on PTB-XL for the same number of training steps (30 epochs and 96 batch size to give 5430 training steps) and found that xECG required $\sim 10\times$ less time (7 min vs. 67 min) and $\sim 7\times$ less memory (4.2 GB vs. 28.4 GB) compared to the second-best method ST-MEM. Note that as ECGFounder is a relatively small CNN (30.7M parameters) compared to the other foundation models (ST-MEM 85.2M, JEPA 85.4M, xECG 57.0M, SimDINOv2 Transformer 85.3M) its runtime and memory usage is similar to xECG. Additionally, although the SimDINOv2 Transformer has a similar parameter count to the other transformer models, its temporal patching process produces less tokens from the same length input signal and so it uses less memory and trains faster.

Discussion

BenchECG enables systematic and rigorous comparison of ECG foundation models across a wide range of tasks. This provides a reliable way to measure progress and encourage reproducibility and higher quality research in the field of cardiac signal analysis. Notably, the benchmark can also serve as a practical decision-making tool for model selection. For a novel cardiac application, the optimal foundation model can be selected by consulting its performance on a similar task within BenchECG. Using this benchmark, xECG achieved the highest overall score, with strong and consistent performance across diverse signals, patient populations, and tasks.

A key result is that xLSTM-based architectures performed particularly well on long-context tasks, such as sleep-apnea classification in overnight recordings and arrhythmia detection in 30-minute ambulatory ECGs. These findings support the view that recurrent architectures are advantageous when modelling extended physiological signals, where dependencies span minutes rather than seconds. The superiority of xECG in linear probing on these tasks suggests that its pretrained features are inherently better aligned with long-term temporal dependencies, rather than requiring task-specific finetuning.

On short 12-lead ECGs, which are the most widely studied setting [7], differences between models after finetuning were small. However, larger gaps emerged under linear probing, indicating that some representations were more transferable without extensive adaptation. Here,

[2] A single L40S Nvidia GPU

finetuning can mask representational differences by overwriting pretrained features, whereas linear probing provides a measure of the intrinsic quality of learnt representations. Prior work evaluating models only on PTB-XL and similar tasks [8, 9] may not be able to adequately measure the differences between architectures, highlighting the importance of evaluating a broader range of tasks.

Cross-population tasks exposed additional distinctions. Although ST-MEM achieved the lowest error in age estimation, xECG exhibited smaller generalisation gaps across datasets from Brazil, the USA, Germany, and China. Survival modelling also demonstrated that while concordance indices dropped across populations, when adjusting xECG risk predictions for age and sex, hazard ratios remained predictive, suggesting the prognostic features learnt during training are at least partially transferable across populations. This supports the translation potential of ECG foundation models, but also highlights the need for evaluation across heterogeneous patient cohorts.

Computational comparisons underscore the importance not just of architecture choice (e.g. CNN, transformer, or xLSTM) but also architectural details. Overall, the CNN and xLSTM based models are more computationally efficient than their transformer counterparts. However, when the patching is performed purely temporally (SimDINOv2 Transformer) rather than across leads and time (ECG-JEPA and ST-MEM), the number of tokens, and hence the computational cost, is greatly reduced. Even so, an advantage of recurrent backbones is they provide linear scaling with input length, enabling efficient training and evaluation on long recordings, where transformer-based methods become prohibitive. This computational efficiency is likely to be important for deploying foundation models in real-world clinical settings, where resources can be constrained and continuous Holter monitoring and ICU measurements generate extremely long signals.

Different architectures impose different priors. When designing a foundation model for varied tasks these priors should not be too strongly linked to one task in particular or the model may not generalise well. This is the case with the CNN-based ECGFounder model. The model performs well on classification tasks, but the CNN-based architecture aggregates contextual information at the expense of temporal resolution, making it unsuitable for detection tasks. This is evident in the R-peak detection tasks where ECGFounder performs poorly. ECGFounder's results on BenchECG suggest that while CNN-based foundation models can perform well on classification and regression tasks using 12-lead 10s ECGs, they do not adapt well to long-context tasks or tasks requiring fine temporal information.

Despite its breadth, BenchECG has several limitations. Not all models can be evaluated. If a model has been pretrained on any of the evaluation datasets, it cannot be fairly compared to other methods. Relatedly, like any fixed benchmark, BenchECG risks becoming a target for overfitting as future methods are tuned specifically to its constituent datasets. Although it covers diverse tasks, it remains constrained by the availability of public datasets, which may not capture the full variability of real-world ECG practice. Additionally, there are substantial differences in the pretraining datasets of the models compared in this paper. ECGFounder used HEEDB [55] ($\sim$ 10M ECGs), xECG used CODE, INCART and Chapman & Ningbo ($\sim$ 8M ECGs), whereas ST-MEM and ECG-JEPA used CODE-15%, Chapman & Ningbo ($\sim 400,000$ ECGs, but these works discarded signals shorter than 10s, giving $\sim 200,000$ ECGs). This limits claims on which pretraining methodology is most suitable for ECG feature learning and future work should explore direct comparisons of pretraining methods using the same pretraining data. Regardless, scaling self-supervised training further and pretraining on a wider variety of ECG signals is likely to be essential, as seen in NLP and vision foundation models [56, 57]. Finally, the evaluation focuses on technical metrics and does not directly assess clinical utility or safety, which would require prospective validation in patient care settings. Benchmarking is an important step for model development and comparisons, but should not be seen as the end of the evaluation process.

Overall, the results show that while several foundation models achieve strong performance on conventional ECG benchmarks, xECG provides superior generalisation, flexibility, and computational efficiency.

Methods

BenchECG datasets and tasks

BenchECG includes a total of eight public datasets (7 for evaluation and 1 for training) and 10 tasks. In the following, we describe each task, including the data splits used for training, validation and testing, and how each task is incorporated into the BenchECG score.

PTB-XL. The PTB-XL dataset (Version: 1.0.3) [25] comprises a collection of 21,799, 10 second, 500 Hz, clinical 12-lead ECGs from 18,869 patients in Germany. The dataset includes a broad range of diagnostic classes including a large number of healthy records. For our evaluation, we focused on the multilabel classification of five diagnostic superclasses: Normal ECG (NORM), Myocardial Infarction (MI), ST/T Change (STTC), Conduction Disturbance (CD) and Hypertrophy (HYP). The diagnostic annotations are inherently multilabel, as patients can present with multiple concurrent conditions. We emphasise that we assess performance solely in the multilabel setting as opposed to many other works [8, 9], which used this dataset in a simplified fashion, where ECGs with more than one label were not considered. We adopt the official stratified splits [25], using fold 10 for held-out testing and fold 9 for validation.

CPSC2018. The CPSC2018 dataset [26] comprises 9,831 12-lead-ECGs from 9,458 patients, recorded in eleven distinct hospitals in China, with lengths ranging from 7 to 60 seconds and sampled at 500 Hz. We included this dataset as it complements the PTB-XL dataset in several aspects: (1) different set of diagnostic multilabels, (2) different patient population, and (3) varying signal length. The BenchECG task using this dataset is a multilabel classification problem including nine label categories: normal, atrial fibrillation (AF), left bundle branch block (LBBB), right bundle branch block (RBBB), first-degree atrioventricular block (I-AVB); premature atrial contraction (PAC), premature ventricular contraction (PVC), ST-segment elevation (STE), and ST-segment depression (STD). The test set for CPSC2018 is private, so we use fold 6 as validation and fold 7 as test.

MIT-BIH. To assess performance on a fine-grained (heartbeat-level), multilabel classification task, we used the MIT-BIH Arrhythmia Database (Version: 1.0.0) [32], from the USA. This dataset consists of 48, half-hour, two-lead ambulatory ECG recordings from 47 subjects, digitised at 360 Hz. Each heartbeat is annotated into one of five arrhythmia classes according to the AAMI standard [citation]: Normal (N), Supraventricular Ectopic (SVE), Ventricular Ectopic (VE), Fusion (F), and Unknown (Q). This dataset is challenging due to its limited size, highly imbalanced classes, and the need for beat-level classification. Following standard practice [58], we use the intra-patient DS1/DS2 split for training and testing, respectively. The specific patient IDs for the split DS1 are: 101, 106, 108, 109, 112, 114, 115, 116, 118, 119, 122, 124, 201, 203, 205, 207, 208, 209, 215, 220, 223, 230: and in the DS2 split: 100, 103, 105, 111, 113, 117, 121, 123, 200, 202, 210, 212, 213, 214, 219, 221, 222, 228, 231, 232, 233, 234. As in previous work, we excude four patients with pacemakers. Split DS1 is further divided in train and validation sets keeping patients 114 and 203 for validation.

R-peak. To evaluate R-peak detection performance we also used the MIT-BIH dataset, using the R-peak annotation locations as ground truth for this task. The same train/val/test splitting is used.

Sleep Apnea. To evaluate a model's ability to handle long-duration signals and tasks, we include the PhysioNet Apnea-ECG Database (Version: 1.0.0) [59]. This dataset contains 70 single-lead ECG recordings collected in Germany, each approximately 7 to 10 hours in duration, with sleep apnea annotations (presence/absence of sleep apnea) at a one-minute resolution. Each participant has a single ECG. The benchmark task is to segment the signal into presence or absence of sleep apnea at a one-minute resolution. The test set is comprised of the filenames beginning with 'x'. The rest of the patients are divided in a random split where 20% of subjects are chosen to be in the validation set and the rest is used for training.

PPG AF. In order to evaluate foundation models on an atypical and out of distribution scenario we include the DeepBeat PPG Dataset [35]. It provides photoplethysmography (PPG) signals annotated with atrial fibrillation (AF) labels. PPG signals are recorded from wearable devices rather than traditional ECG equipment, presenting unique challenges related to noise artifacts and signal quality that are characteristic of wrist-worn sensors. The dataset comprises

over 500,000 labelled PPG signals from more than 100 individuals, collected from three distinct sources to ensure diversity and robustness:

- Cardioversion Cohort: 107 patients undergoing elective cardioversion procedures at Stanford University, all presenting with atrial fibrillation.
- Exercise Stress Test Cohort: 41 participants undergoing elective exercise stress tests, representing normal sinus rhythm cases.
- IEEE Signal Processing Cup 2015 Dataset: Publicly available data supplementing the Stanford collections to provide out-of-institution examples and enhance generalisability.

We used the official train, validation and test splits [35].

Exercise. The ECG in High Intensity Exercise Dataset (Exercise-ECG) consist of ECGs extracted in different times of a maximal exercise test. In particular, for each patient the first segment was extracted 30s before a heavy intensity effort and the second 60s after. The third 30s before VO_2 max (highly severe intensity up to exhaustion). The fourth at the moment of exhaustion (centred on the VO2 max measurement); the fifth 60s post-exercise, i.e. during the recovery after exhaustion. This is a very small dataset with only 20 participants and 5 recordings each, for a total of 100 ECGs, each 20s long. The dataset includes R-peak annotations which we use to train for R-peak detection. We used patients 10-13 for validation and 13-20 for testing.

Age. A regression task where models are trained to predict the age of a patient from their ECG. We used the CODE-15% dataset for training and validation while the entire PTB-XL, CPSC2018 and MIMIC-IV-ECG (Version: 1.0) [27] datasets for testing in order to assess the performance of the models across different populations (Europe, China and USA, respectively). Ages range from 17-100 in the training set and 1-101 in the testing sets. For the PTB-XL dataset, patients older than 90 years have a recorded age of 300 years for privacy reasons, hence we excluded these patients. In the CPSC2018 dataset we did not considered patients with invalid age (e.g., -1 or not a numbers).

Mortality. In order to evaluate foundation models on survival analysis, we again used the CODE-15% dataset for training and validation and the MIMIC-IV-ECG dataset for testing. CODE-15% includes follow-up time in years for all ECGs and a binary label indicating mortality. For MIMIC-IV-ECG, we extracted the mortality data from the admission file of the original MIMIC-IV dataset (Version: 3.1) [30]. The models are trained to minimize the Cox proportional hazards partial likelihood function [60]:

$$\mathcal{L}_{\text{cox}} = \sum_{i=1}^{n} \delta_i \Big\{ \phi(x_i) - \ln \sum_{j \in R(t_i)} e^{\phi(x_j)} \Big\} \tag{1}$$

where δ_i is a boolean indicator of the subject i's status (mortality) and $\phi(x_i)$ is the scalar output of the model, representing the survival prediction (specifically, the natural logarithm of the hazard ratio) and $R(t_i)$ is the subset of patients not censored at the time subject i died or became censored ($\{j : t_j > t_i\}$).

Blood test. To assess the capability of foundation models to extract patient information normally collected separately from an ECG, we predict (ab)normal blood test results for patients in the MIMIC-IV-ECG dataset. We extract the labels from the original MIMIC-IV dataset [30]. Previous work analysing blood marker prediction from MIMIC-IV-ECG [61] found many markers are too challenging to predict from ECG signals. In the benchmark we evaluate only those markers which achieved an AUROC above 0.7 in this prior work [61]. These markers are: Albumin; Hemoglobin; NTproBNP; Acetaminophen; Hematocrit; PT; Red Blood Cells; 25-OH Vitamin D; RDW-SD; INR(PT); Urea Nitrogen; Monocytes; Acetaminophen; Absolute Basophil Count; Urea Nitrogen; C-Reactive Protein; Cholesterol, HDL; Bilirubin, Direct; Creatinine; Sedimentation Rate; pO2; Osmolality, Measured; Bicarbonate. We treat the task as a multi-task, multi-class setting. Where the model predicts whether the result is below, inside or above the reference values. Reference values considered for the classification tasks are chosen to be the median values of the reference values, following the motivation of the work cited before. For the test split we used patients in fold 19, for the validation, fold 18, and the rest for

training. These folds come from the diagnostic labels in the MIMIC-IV-ECG-Ext-ICD external dataset (Version: 1.0.1) [62].

BenchECG Metrics

In order to define a BenchECG score, each task needs to be evaluated using a metric normalised between 0-1, with 1 indicating perfect performance. In this section we justify our choice of metric for each task.

For classification tasks we use area under the receiver operator characteristic curve (AUROC). AUROC is a suitable metric as it is not affected by prevalence and it represents a model's predictive power across all operating points. An AUROC of 1 indicates perfect classification and an AUROC of 0.5 is equivalent to a random model. The one exception to using AUROC is for MIT-BIH, where we use F1 score. This is because F1 score is typically reported in the literature [49–53], unlike AUROC, and so allows for better comparisons to previous work.

For R-peak detection, we also use F1. However, more detail is required, as how close does a prediction need to be for it to count as correct? A recent review on R-peak detection [54] suggests using a tolerance window of 150ms, but, as can be seen from our results, this leads to saturated performance with many methods being near-perfect. Instead, we use a stricter 20ms total tolerance window to push the boundaries of this task and have a more useful comparison across methods.

For age regression, we use the symmetric mean absolute percentage error (SMAPE) wich is bounded between 0 and 1:

$$\text{SMAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|x_i - y_i|}{|x_i| + |y_i|} \tag{2}$$

where x_i is the model prediction and y_i is the ground truth. An SMAPE of 0 indicates no error so, for our BenchECG score, we use $1-$ SMAPE.

For the survival analysis task, we use concordance index. Concordance index represents the probability that two randomly selected subjects will maintain the same relative order of their observed event (death in this case) as predicted by the model. A higher C-index, closer to 1, indicates better predictive performance, while a value of 0.5 suggests the model is no better than random guessing, similar to AUROC.

Model selection and data preprocessing

Our primary baselines are ST-MEM [8], ECG-JEPA [9] and ECG founder [5]. To ensure a fair and direct comparison, we use the official pretrained weights provided by the authors and re-evaluate them within our standardised evaluation framework. For each model we applied the preprocessing strategies as detailed in their respective original works [5, 8, 9]. Below, we briefly describe each model.

ST-MEM. A foundation model based on the vision transformer (ViT-Base) architecture [63]. It is pretrained in a self-supervised manner following a masked auto-encoder paradigm (MAE) [42]. The authors introduce spatio-temporal patchifying, where the signals are tokenised across both time and leads.

ECG-JEPA. A foundation model based on the vision transformer (ViT-Base) architecture [63]. It is pretrained in a self-supervised manner following JEPA [43]. The authors introduce a custom cross pattern attention [9] which replaces standard attention. As in ST-MEM, signals are tokenised across both time and leads.

ECGFounder. This foundation model is based on convolutional neural network (CNN) and was trained with supervision on a large scale labelled dataset of over 10 million ECGs [5]. Here the model accept the full signal in input and outputs a single feature map.

SimDINOv2 Transformer. This is an ablation study for xECG, where we use a vision transformer (ViT-Base) architecture [63] in place of the xLSTM, but apply the same simDINOv2 pretraining and the same patching strategy.

xLSTM Supervised. A further ablation of our xECG model where we maintain the same exact architecture but we do not pre-train it.

Together, these models span different architectural approaches currently used for ECG analysis, including transformer-based, convolutional, and recurrent networks.

xECG Architecture

The core of our xECG is an encoder composed of a stack of nine alternating sLSTM and mLSTM blocks (s, s, m, m, s, s, m, m, s) processing sequences of ECG patches bidirectionally. A high-level overview of our architecture is shown in Figure 1b.

Let an input ECG signal be denoted by $s \in \mathbb{R}^{L \times C}$, where L is the signal length and C is the number of leads (channels). Given a model sampling frequency $f_m \in \mathbb{R}_{\geq 0}$ and an input signal frequency $f_s \in \mathbb{R}_{\geq 0}$, the signal is first resampled to f_m. We intentionally omit any other kind of pre-processing steps, e.g., normalisation or extensive filtering. Thus, the model directly learns from the raw signal, allowing it to capture any potentially relevant information present in the signal.

The resampled signal is then divided along only the temporal dimension into a sequence of N non-overlapping patches $P = (p_e, p_1, p_2, ..., p_N)$, where each patch has a size of P_s. The signal is truncated to ensure its length is a multiple of P_s. Each patch is flattened and transformed via a linear projection into the model's embedding space, resulting in a sequence of patch embeddings $s_e = (e_1, e_2, ..., e_N)$, where each $e_i \in \mathbb{R}^E$ and E is the model's embedding dimension.

The sequence of patch embeddings s_e is processed by the xLSTM encoder. The encoder consists of stacked pairs of sLSTM and mLSTM blocks. To achieve bidirectionality, we adopt a layer-wise alternating strategy. Each pair of blocks processes the sequence in opposing directions: one block processes the sequence from start to finish, while the second processes a reversed version.

After passing through the final encoder layer, we obtain a sequence of rich patch representations s_r. These patches can be used in different ways: pooled to produce a single, fixed-size vector for downstream classification or as they are for segmentation or detection tasks. During pretraining these representations are aggregated using an attention pooling layer [64].

Self-supervised pretraining

The adopted SimDINOv2 [3] pretraining strategy is based on a teacher-student framework where the student network learns by matching the output of the teacher network. The student parameters θ_t are updated via standard backpropagation. The teacher parameters θ_t are not trained directly but are instead an exponential moving average (EMA) of the student's weights. At each training iteration, the teacher is updated using the following rule:

$$\theta_t \leftarrow \lambda_t \theta_t + (1 - \lambda_t)\theta_s \tag{3}$$

where λ_t is the momentum parameter scheduled to increase during training. This schedule encourages the teacher to follow the student's progress more closely at the beginning and then stabilize as training progresses. The schedule is defined as:

$$\lambda_t \leftarrow \lambda_{\text{base}} + (t/N)(1 - \lambda_{\text{base}}) \tag{4}$$

Here, λ_{base} is the initial momentum value, t is the current training iteration and N is the total number of training iterations.

A key component of the DINO family of algorithms is the use of multi-crop augmentations, where the model learns to associate different views of the same sample. We adapt this concept from the image domain to time-series ECG data. For each ECG signal in a batch, we generate a set of augmented "views" by creating random sub-sequences of varying lengths:

- **Global Views**: Two long, overlapping sub-sequences (e.g., 6-10 seconds long).
- **Local Views**: Four shorter sub-sequences (e.g., 1-3 seconds long).

In Fig. 1c an example of different views of the same sample is shown.

During training, the teacher network processes only the global views. The student network, instead, process all views (both global and local). The training objective will force the student's

output for every view to match the teacher's output for the corresponding global view. Because in the pretraining we might have multiple ECGs for each patient, we consider different samples of the same patient as a single source where to get different views. This means that if a patient has two ECGs, a global (or local) view is selected to be a slice of one of the two samples, selected randomly. At loss level we treat then these two views coming from two different samples (but same patient) as if they belong to the same original signal. This loss objective of SimDINOv2 consists of three distinct components:

- **Patch-level objective ($\mathcal{L}_{\text{patch}}$)**: For each global view provided to the student network, a random subset of its patch embeddings are replaced with a shared, learnable [MASK] token. The teacher network receives the same global view but without any masking. The student is then asked to reconstruct the original, unmasked patch representations from the teacher. The loss is computed as the Mean Squared Error (MSE) between the student's normalised output for the masked positions and the teacher's corresponding outputs. Formally, let $\hat{s}_t$ be the sequence of l_2 normalised patch representations from the teacher for an unmasked view, and let $\hat{s}_s$ be the l_2 normalised student's output for the masked view. If M is the set of indices for the masked patches, the loss is:

$$\mathcal{L}_{\text{patch}} = \frac{1}{|M|} \sum_{i \in M} \|\hat{s}_{s,i} - \hat{s}_{t,i}\|_2^2 , \tag{5}$$

with

$$\hat{s}_{s,i} = \frac{s_{s,i}}{\|s_{s,i}\|_2}, \; \hat{s}_{t,i} = \frac{s_{t,i}}{\|s_{t,i}\|_2} \tag{6}$$

- **Sample-level objective ($\mathcal{L}_{\text{view}}$):** The student network is trained to produce representations for all views (both global and local) of the same sample to be closer to the respective teacher's global views representations. Specifically, the loss is calculated between the student's representation for one view and the teacher's representation for a different global view from the same original signal. This is performed across all pairs of student-teacher views. Let K be the total amount of global and local views and $Z_t = \{z_{t,1}, ..., z_{t,G}\}$ be the set of pooled representations of the global views processed by the teacher, where $G < K$ is the number of global views. Let $Z_s = \{z_{s,1}, ..., z_{s,K}\}$ be the set of representations of all the K views presented to the student, the loss is defined as:

$$\mathcal{L}_{\text{view}} = \sum_{i=1}^{G} \sum_{\substack{j=1 \\ j \neq i}}^{K} \|\hat{z}_{t,i} - \hat{z}_{s,j}\|_2^2 , \tag{7}$$

with

$$\hat{z}_{s,i} = \frac{z_{s,i}}{\|z_{s,i}\|_2}, \; \hat{z}_{t,i} = \frac{z_{t,i}}{\|z_{t,i}\|_2} \tag{8}$$

- **Coding rate regularizer ($\mathcal{L}_{\text{cr}}$)**: This regularizer operates on the covariance matrix of the l_2 normalised feature vectors $\hat{z}_s$ produced by the student network. Formally, the regularization term to be minimised is:

$$\mathcal{L}_{\text{cr}} = -\gamma R_\epsilon(Cov[\hat{z}_s]) \tag{9}$$

where γ is a scalar hyperparameter that controls the strength of the regularization defined by:

$$\gamma = \epsilon \sqrt{B/(E \min\{E, B\})} \tag{10}$$

being ϵ an hyperparameter, E the embedding size and B the batch size. $Cov[\hat{z}_s]$ is the sample covariance matrix of the student's normalised feature vectors within a batch, and R_ϵ is the coding rate function defined as:

$$R_\epsilon(\Gamma) = \frac{1}{2} \log\det\left(\mathbf{I} + \frac{E}{\epsilon}\Gamma\right) \tag{11}$$

Here, Γ is the covariance matrix and I is the identity matrix.

The final composite loss $\mathcal{L}$ is then the sum of these terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{patch}} + \mathcal{L}_{\text{view}} + \mathcal{L}_{\text{cr}} \tag{12}$$

Adapting the multi-crop strategy from the image domain to ECG signals requires a set of carefully designed augmentations:

- **Random lead dropout**[65]: Every lead is zeroed out with a probability p_{drop}. This do not apply to lead II to prevent all leads are zeroed out. This forces the model to learn redundant information present across the 12 leads and reconstruct a complete cardiac picture from partial data.
- **Low-frequency component swap:** To teach invariance to baseline wander caused by patient movement or respiration, we isolate and swap low-frequency components across signals in a batch. First, we apply a low-pass filter (a Butterworth filter with a cutoff at 0.5 Hz) to extract the low-frequency baseline from each signal. These baseline signals are then randomly shuffled and added back to the corresponding high-frequency components of different signals in the batch.
- **Multiplicative gaussian jitter:** To simulate realistic sensor noise that often scales with signal magnitude, we apply a multiplicative form of Gaussian jitter. With a probability p_{jitter}, the augmented signal $s'(t)$ is generated from the original signal $s(t)$ according to the formula:

$$s'(t) = s(t) \cdot [1 + A \cdot n(t)] \tag{13}$$

  where A is a scalar "amplitude" hyperparameter, and $n(t)$ is noise sampled from a standard normal distribution $\mathcal{N}(0, \sigma^2)$.
- **Random amplitude scaling:** To account for variations in overall signal strength due to factors like electrode-skin impedance, the entire signal is scaled by a single random scalar α. With a probability p_{scale}, the scalar α is drawn from a uniform distribution:

$$\alpha \sim \mathcal{U}\left(1 - \frac{R}{2}, 1 + \frac{R}{2}\right) \tag{14}$$

  where R is the amplitude range hyperparameter. The augmented signal is then $s'(t) = \alpha \cdot s(t)$.

This pretraining strategy used a large-scale corpus of ECG data aggregated from several publicly available sources:

- **CODE** [36]: This extensive dataset comprises 2,322,513 12-lead ECG recordings from 1,676,384 patients. The signals have durations ranging from 7 to 10 seconds and were recorded at various sampling rates between 300 and 600 Hz.
- **Chapman** [66] and **Ningbo** [67]: Together, these datasets contribute 45,152 10-second, 12-lead ECGs sampled at 500 Hz.
- **Incart** [28]: This dataset consists of 75 12-lead Holter recordings, each 30 minutes in duration and sampled at 257 Hz, with each recording from a different patient. These long-duration signals provide a source for extracting multiple, diverse segments from a single continuous recording.

To ensure a minimum standard of quality and consistency across the aggregated datasets, a simple yet effective preprocessing pipeline was applied to each ECG signal. We filter out corrupted or invalid recordings. We excluded any signals that met one of the following criteria:

1. The signal contained missing (NaN) values.
2. The signal was completely composed by zeroes.
3. The signal exhibited excessive noise or artifacts, identified by a variance greater than 10 combined with an absolute amplitude exceeding 15 mV.

Then, all signals were uniformly resampled to a target frequency of 100 Hz. This step serves a dual purpose: it standardises the temporal resolution and acts as a low-pass filter, removing high-frequency noise while preserving the essential morphological features of the ECG

waveform. Lower sampling rate reduces computational complexity without compromising downstream task performance [68].

Data availability

All datasets used in BenchECG are publicly available:

- PTB-XL: freely available at https://physionet.org/content/ptb-xl/1.0.3/.
- CPSC2018: freely available at https://physionet.org/content/challenge-2020/1.0.2/training/cpsc_2018/.
- MIT-BIH Arrhythmia Database: freely available at https://www.physionet.org/content/mitdb/1.0.0/.
- ECG in High Intensity Exercise Dataset: freely available at https://zenodo.org/records/5727800.
- Sleep Apnea-ECG Database: freely available at https://www.physionet.org/content/apnea-ecg/1.0.0/.
- DeepBeat PPG: available at https://www.synapse.org/Synapse:syn21985690/wiki/.
- MIMIC-IV-ECG (v1.0): available at https://physionet.org/content/mimic-iv-ecg/1.0/.
- MIMIC-IV-ECG-Ext-ICD (v1.0.1): available at https://physionet.org/content/mimic-iv-ecg-ext-icd-labels/1.0.1/ (requires PhysioNet credential access and training completion).
- MIMIC-IV (v3.1): available at https://physionet.org/content/mimiciv/3.1/ (requires PhysioNet credential access and training completion).
- CODE-15: freely available at https://zenodo.org/records/4916206.

For pre-training we use the CODE dataset. For access, contact the authors of [36]. Additionally we used INCART (available at urlhttps://physionet.org/content/incartdb/1.0.0/) and Chapman & Ningbo (available at https://physionet.org/content/ecg-arrhythmia/1.0.0/).

Code availability

We release xECG weights and the BenchECG code in the following repository: https://github.com/dlaskalab/bench-xecg/.

References

[1] Lindstrom, M. *et al.* Global Burden of Cardiovascular Diseases and Risks Collaboration, 1990-2021. *J. Am. Coll. Cardiol.* **80**, 2372–2425 (2022).

[2] Beck, M. *et al.* xLSTM: Extended long short-term memory. *Adv. Neural Inf. Process. Syst.* **37**, 107547–107603 (2024).

[3] Wu, Z. *et al.* Simplifying DINO via Coding Rate Regularization. *arXiv preprint arXiv:2502.10385* (2025).

[4] Ribeiro, A. H. *et al.* Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nat Commun* **11**, 1760 (2020).

[5] Li, J. *et al.* An Electrocardiogram Foundation Model Built on over 10 Million Recordings. *NEJM AI* **2**, AIoa2401033 (2025).

[6] Moor, M. *et al.* Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).

[7] Han, Y., Liu, X., Zhang, X. & Ding, C. Foundation Models in Electrocardiogram: A Review. *arXiv preprint arXiv:2410.19877* (2024).

[8] Na, Y., Park, M., Tae, Y. & Joo, S. Guiding Masked Representation Learning to Capture Spatio-Temporal Relationship of Electrocardiogram. *arXiv preprint arXiv:2402.09450* (2024).

[9] Kim, S. Learning General Representation of 12-Lead Electrocardiogram with a Joint-Embedding Predictive Architecture. *arXiv preprint arXiv:2410.08559* (2024).

[10] Henderson, P. *et al.* Deep Reinforcement Learning That Matters. *Proc. AAAI Conf. Artif. Intell.* **32** (2018).

[11] Pineau, J. *et al.* Improving Reproducibility in Machine Learning Research(A Report from the NeurIPS 2019 Reproducibility Program). *J. Mach. Learn. Res.* **22**, 1–20 (2021).

[12] Roberts, M. *et al.* Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat Mach Intell* **3**, 199–217 (2021).

[13] Kapoor, S. & Narayanan, A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* **4**, 100804 (2023).

[14] Ball, P. Is AI leading to a reproducibility crisis in science? *Nature* **624**, 22–25 (2023).

[15] Bernardini, A., Brunello, A., Gigli, G. L., Montanari, A. & Saccomanno, N. AIOSA: An approach to the automatic identification of obstructive sleep apnea events based on deep learning. *Artif. Intell. Med.* **118**, 102133 (2021).

[16] Deng, J. *et al.* ImageNet: A large-scale hierarchical image database. *CVPR* 248–255 (2009).

[17] Lin, T.-Y. *et al.* Fleet, D., Pajdla, T., Schiele, B. & Tuytelaars, T. (eds) *Microsoft COCO: Common Objects in Context.* (eds Fleet, D., Pajdla, T., Schiele, B. & Tuytelaars, T.) *Comput. Vis. – ECCV 2014*, 740–755.

[18] Wang, Y. *et al.* MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. *NeurIPS* (2024).

[19] Bosma, J. S. *et al.* The DRAGON benchmark for clinical NLP. *npj Digit. Med.* **8**, 289 (2025).

[20] Ektefaie, Y. *et al.* Evaluating generalizability of artificial intelligence models for molecular datasets. *Nat Mach Intell* **6**, 1512–1524.

[21] Karargyris, A. *et al.* Federated benchmarking of medical artificial intelligence with MedPerf. *Nat Mach Intell* **5**, 799–810.

[22] Auer, A. *et al.* TiRex: Zero-Shot Forecasting Across Long and Short Horizons with Enhanced In-Context Learning. *arXiv preprint arXiv:2505.23719* (2025).

[23] Kong, Y. *et al.* Unlocking the Power of LSTM for Long Term Time Series Forecasting. *Proc. AAAI Conf. Artif. Intell.* **39**, 11968–11976 (2025).

[24] Bommasani, R. *et al.* On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258* (2021).

[25] Wagner, P. *et al.* PTB-XL, a large publicly available electrocardiography dataset. *Sci. Data* **7**, 1–15 (2020).

[26] Liu, F. *et al.* An Open Access Database for Evaluating the Algorithms of Electrocardiogram Rhythm and Morphology Abnormality Detection. *J. Med. Imaging Health Inform.* **8**, 1368–1373 (2018).

[27] Gow, B. *et al.* MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset. *PhysioNet* (2023).

[28] Goldberger, A. L. *et al.* PhysioBank, PhysioToolkit, and PhysioNet. *Circulation* **101**, e215–e220 (2000).

[29] Johnson, A. *et al.* MIMIC-IV. *PhysioNet* (2023).

[30] Johnson, A. E. W. *et al.* MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data* **10**, 1 (2023).

[31] Ribeiro, A. H. *et al.* CODE-15%: A large scale annotated dataset of 12-lead ECGs (2021). URL https://doi.org/10.5281/zenodo.4916206.

[32] Moody, G. B. & Mark, R. G. The impact of the MIT-BIH Arrhythmia Database. *IEEE Eng. Med. Biol. Mag.* **20**, 45–50 (2001).

[33] Chang, H.-Y., Yeh, C.-Y., Lee, C.-T. & Lin, C.-C. A Sleep Apnea Detection System Based on a One-Dimensional Deep Convolution Neural Network Model Using Single-Lead Electrocardiogram. *Sensors* **20**, 4157 (2020).

[34] De Giovanni, E., Teijeiro, T., Meier, D., Millet, G. & Atienza, D. ECG in high intensity exercise dataset (2021). URL https://doi.org/10.5281/zenodo.5727800.

[35] Torres-Soto, J. & Ashley, E. A. Multi-task deep learning for cardiac rhythm detection in wearable devices. *npj Digit. Med.* **3**, 116 (2020).

[36] Ribeiro, A. L. P. *et al.* Tele-electrocardiography and bigdata: The CODE (Clinical Outcomes in Digital Electrocardiography) study. *J. Electrocardiol.* **57**, S75–S78 (2019).

[37] Vaswani, A. *et al.* Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30** (2017).

[38] Schmidinger, N. *et al.* Bio-xLSTM: Generative modeling, representation and in-context learning of biological and chemical sequences. *arXiv preprint arXiv:2411.04165* (2025).

[39] Kang, L., Fu, X., Vazquez-Corral, J., Valveny, E. & Karatzas, D. xLSTM-ECG: Multi-label ECG Classification via Feature Fusion with xLSTM. *arXiv preprint arXiv:2504.16101* (2025).

[40] Alkin, B., Beck, M., Pöppel, K., Hochreiter, S. & Brandstetter, J. Vision-LSTM: xLSTM as generic vision backbone. *ICLR* (2025).

[41] Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. *ICML* (2020).

[42] He, K. *et al.* Masked Autoencoders Are Scalable Vision Learners. *CVPR* 16000–16009 (2022).

[43] Assran, M. *et al.* Self-Supervised Learning From Images With a Joint-Embedding Predictive Architecture. *CVPR* 15619–15629 (2023).

[44] Ericsson, L., Gouk, H. & Hospedales, T. M. How Well Do Self-Supervised Models Transfer? *CVPR* (2021).

[45] Caron, M. *et al.* Emerging Properties in Self-Supervised Vision Transformers. *ICCV* 9650–9660 (2021).

[46] Oquab, M. *et al.* DINOv2: Learning Robust Visual Features without Supervision. *TMLR* (2024).

[47] Huang, S.-C. *et al.* Self-supervised learning for medical image classification: A systematic review and implementation guidelines. *npj Digit. Med.* **6**, 74 (2023).

[48] Liu, Z., Alavi, A., Li, M. & Zhang, X. Self-Supervised Contrastive Learning for Medical Time Series: A Systematic Review. *Sensors* **23**, 4221 (2023).

[49] Xia, Y., Xiong, Y. & Wang, K. A transformer model blended with CNN and denoising autoencoder for inter-patient ECG arrhythmia classification. *Biomed. Signal Process. Control* **86**, 105271 (2023).

[50] Sellami, A. & Hwang, H. A robust deep convolutional neural network with batch-weighted loss for heartbeat classification. *Expert Syst. Appl.* **122**, 75–84 (2019).

[51] Li, Y., Qian, R. & Li, K. Inter-patient arrhythmia classification with improved deep residual convolutional neural network. *Comput. Methods Programs Biomed.* **214**, 106582 (2022).

[52] Marinho, L. B. *et al.* A novel electrocardiogram feature extraction approach for cardiac arrhythmia classification. *Future Gener. Comput. Syst.* **97**, 564–577 (2019).

[53] Li, F., Xu, Y., Chen, Z. & Liu, Z. Automated Heartbeat Classification Using 3-D Inputs Based on Convolutional Neural Network With Multi-Fields of View. *IEEE Access* **7**, 76295–76304 (2019).

[54] Ali, S. T. A., Kim, S. & Kim, Y.-J. Towards Reliable ECG Analysis: Addressing Validation Gaps in the Electrocardiographic R-Peak Detection. *Appl. Sci.* **14**, 10078.

[55] Koscova, Z. *et al.* The Harvard-Emory ECG database. *medRxiv* (2025).

[56] Kaplan, J. *et al.* Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020).

[57] Zhai, X., Kolesnikov, A., Houlsby, N. & Beyer, L. Scaling Vision Transformers. *CVPR* 12104–12113 (2022).

[58] De Chazal, P., O'Dwyer, M. & Reilly, R. B. Automatic classification of heartbeats using ECG morphology and heartbeat interval features. *IEEE Trans. Biomed. Eng.* **51**, 1196–1206 (2004).

[59] Penzel, T., Moody, G. B., Mark, R. G., Goldberger, A. L. & Peter, J. H. The apnea-ecg database. *Computers in Cardiology 2000. Vol. 27 (Cat. 00CH37163)* 255–258 (2000).

[60] Bello, G. A. *et al.* Deep learning cardiac motion analysis for human survival prediction. *Nat Mach Intell* **1**, 95–104 (2019).

[61] Miguel Lopez Alcaraz, J. & Strodthoff, N. CardioLab: Laboratory Values Estimation from Electrocardiogram Features - An Exploratory Study. *2024 Computing in Cardiology Conference* (2024).

[62] Strodthoff, N., Lopez Alcaraz, J. M. & Haverkamp, W. MIMIC-IV-ECG-Ext-ICD: Diagnostic labels for MIMIC-IV-ECG (version 1.0.1). *PhysioNet* (2024).

[63] Dosovitskiy, A. *et al.* An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR* (2021).

[64] Bolya, D. *et al.* Perception Encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181* (2025).

[65] Oh, J., Chung, H., Kwon, J.-m., Hong, D.-g. & Choi, E. Lead-agnostic Self-supervised Learning for Local and Global Representations of Electrocardiogram. *Proc. Conf. Health Inference Learn.* 338–353 (2022).

[66] Zheng, J. *et al.* A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Sci Data* **7**, 48 (2020).

[67] Zheng, J. *et al.* Optimal Multi-Stage Arrhythmia Classification Approach. *Sci Rep* **10**, 2898 (2020).

[68] Mehari, T. & Strodthoff, N. Advancing the State-of-the-Art for ECG Analysis through Structured State Space Models. *arXiv preprint arXiv:2211.07579* (2022).

[69] Jyotishi, D. & Dandapat, S. An Attentive Spatio-Temporal Learning-Based Network for Cardiovascular Disease Diagnosis. *IEEE Trans. Syst. Man Cybern. Syst.* **53**, 4661–4671 (2023).

[70] Murugesan, B. *et al.* ECGNet: Deep Network for Arrhythmia Classification. *IEEE MeMeA* 1–6 (2018).

[71] Shen, Q., Qin, H., Wei, K. & Liu, G. Multiscale Deep Neural Network for Obstructive Sleep Apnea Detection Using RR Interval From Single-Lead ECG Signal. *IEEE Trans. Instrum. Meas.* **70**, 1–13 (2021).

Supplementary Information

Individual BenchECG Results

In this section, we report the individual results of each model for each task in BenchECG. Each model is trained 5 times and we report the mean± standard deviation (SD). Pairwise differences between models were assessed using two-sided Welch's t-tests across independent runs, with significance defined as $p < 0.05$. Bold results indicate best performance across models and underlined indicates 2nd best.

Methods	PTB-XL			
	Accuracy	F1 Score	AUROC	AUPRC
Supervised				
xLSTM-ECG [39]	0.875	-	0.913	-
ASTLNet [69]	0.629	-	0.913	-
ECGNet [70]	0.873	-	0.901	-
Supervised xLSTM	0.860 ± 0.005	0.667 ± 0.007	0.891 ± 0.004	0.748 ± 0.006
Linear Probing				
ST-MEM	0.847 ± 0.001	0.564 ± 0.002	0.867 ± 0.001	0.693 ± 0.001
ECG-JEPA	0.864 ± 0.001	**0.712 ± 0.002**	0.902 ± 0.000	0.769 ± 0.000
ECGFounder	**0.883 ± 0.001**	0.702 ± 0.002	**0.917 ± 0.000**	**0.799 ± 0.000**
SimDINOv2 Transformer	0.875 ± 0.001	0.703 ± 0.002	0.914 ± 0.000	0.782 ± 0.001
xECG	0.876 ± 0.001	0.690 ± 0.001	0.915 ± 0.000	0.788 ± 0.001
Finetuning				
ST-MEM	**0.892 ± 0.001**	**0.739 ± 0.005**	**0.930 ± 0.001**	**0.822 ± 0.002**
ECG-JEPA	0.884 ± 0.001	0.733 ± 0.003	0.923 ± 0.000	0.809 ± 0.000
ECGFounder	**0.891 ± 0.001**	**0.734 ± 0.002**	**0.930 ± 0.000**	**0.822 ± 0.001**
SimDINOv2 Transformer	0.880 ± 0.002	0.710 ± 0.017	0.925 ± 0.001	0.808 ± 0.001
xECG	0.884 ± 0.002	0.718 ± 0.011	0.928 ± 0.001	0.816 ± 0.003

Table 2 PTB-XL. Individual model performance for the PTB-XL task. Each model is trained 5 times and we report the mean ± standard deviation. Pairwise differences between models were assessed using two-sided Welch's t-tests across independent runs, with significance defined as $p < 0.05$. Bold results indicate best performance across models and underlined indicates 2nd best.

Methods	CPSC2018			
	Accuracy	F1 Score	AUROC	AUPRC
Supervised				
Supervised xLSTM	0.952 ± 0.001	0.719 ± 0.008	0.951 ± 0.002	0.813 ± 0.006
Linear Probing				
ST-MEM	0.935 ± 0.000	0.455 ± 0.005	0.922 ± 0.001	0.722 ± 0.004
ECG-JEPA	0.889 ± 0.002	0.613 ± 0.003	0.961 ± 0.000	0.830 ± 0.001
ECGFounder	0.958 ± 0.000	0.750 ± 0.007	0.962 ± 0.000	0.830 ± 0.001
SimDINOv2 Transformer	0.953 ± 0.001	0.733 ± 0.008	0.951 ± 0.001	0.787 ± 0.008
xECG	**0.961 ± 0.002**	**0.781 ± 0.011**	**0.968 ± 0.001**	**0.861 ± 0.004**
Finetuning				
ST-MEM	0.963 ± 0.001	0.789 ± 0.012	0.958 ± 0.003	0.835 ± 0.011
ECG-JEPA	0.942 ± 0.003	0.749 ± 0.009	0.965 ± 0.001	0.842 ± 0.003
ECGFounder	0.964 ± 0.000	0.794 ± 0.005	0.969 ± 0.001	0.854 ± 0.002
SimDINOv2 Transformer	0.961 ± 0.001	0.796 ± 0.008	0.965 ± 0.003	0.853 ± 0.006
xECG	**0.968 ± 0.001**	**0.822 ± 0.008**	**0.981 ± 0.001**	**0.888 ± 0.003**

Table 3 CPSC2018. Individual model performance for the CPSC2018 task. Each model is trained 5 times and we report the mean± standard deviation. Pairwise differences between models were assessed using two-sided Welch's t-tests across independent runs, with significance defined as $p < 0.05$. Bold results indicate best performance across models and underlined indicates 2nd best.

Methods	MIT-BIH				
	Accuracy	Sensitivity	PPV	Specificity	F1 Score
Supervised					
Xia et al.[49]	0.976	0.510	0.563	0.930	0.529
Sellami et al. [50]	0.953	0.690	0.555	0.960	0.558
Li et al.[51]	0.889	0.521	0.568	0.947	0.533
Marinho et al. [52]	0.943	0.478	-	0.914	-
Li et al. [53]	0.914	0.616	0.489	0.950	0.539
Sup. xLSTM	0.896 ± 0.013	0.448 ± 0.028	0.416 ± 0.020	0.910 ± 0.004	0.402 ± 0.014
Linear Probing					
ST-MEM	0.937 ± 0.003	0.424 ± 0.027	0.507 ± 0.035	0.920 ± 0.003	0.436 ± 0.036
ECG-JEPA	0.946 ± 0.000	0.470 ± 0.003	0.574 ± 0.003	0.933 ± 0.000	0.496 ± 0.003
ECGFounder	0.865 ± 0.012	0.210 ± 0.004	0.209 ± 0.004	0.808 ± 0.003	0.206 ± 0.005
SimDINOv2 Transf.	0.931 ± 0.000	0.381 ± 0.004	0.549 ± 0.012	0.924 ± 0.004	0.395 ± 0.005
xECG	**0.972 ± 0.002**	**0.686 ± 0.021**	**0.669 ± 0.014**	**0.976 ± 0.003**	**0.674 ± 0.013**
Finetuning					
ST-MEM	0.967 ± 0.002	**0.625 ± 0.007**	0.671 ± 0.013	**0.965 ± 0.003**	0.644 ± 0.007
ECG-JEPA	0.954 ± 0.001	0.557 ± 0.006	0.629 ± 0.007	0.951 ± 0.001	0.583 ± 0.007
ECGFounder	0.863 ± 0.008	0.211 ± 0.002	0.207 ± 0.002	0.809 ± 0.002	0.207 ± 0.002
SimDINOv2 Transf.	0.896 ± 0.008	0.475 ± 0.002	0.457 ± 0.005	0.905 ± 0.001	0.425 ± 0.006
xECG	**0.976 ± 0.007**	**0.651 ± 0.035**	**0.721 ± 0.019**	**0.973 ± 0.010**	**0.677 ± 0.025**

Table 4 MIT-BIH. Individual model performance for the MIT-BIH task. Each model is trained 5 times and we report the mean± standard deviation. Pairwise differences between models were assessed using two-sided Welch's t-tests across independent runs, with significance defined as $p < 0.05$. Bold results indicate best performance across models and underlined indicates 2nd best.

Methods	MIT-BIH (R-peak)	
	F1 Score (150ms)	F1 Score (20ms)
Supervised		
Supervised xLSTM	0.980 ± 0.007	0.889 ± 0.012
Linear Probing		
ST-MEM	0.970 ± 0.001	0.564 ± 0.002
ECG-JEPA	**0.976 ± 0.000**	0.444 ± 0.001
ECGFounder	0.073 ± 0.005	0.010 ± 0.001
SimDINOv2 Transformer	0.864 ± 0.001	0.476 ± 0.000
xECG	0.945 ± 0.005	**0.590 ± 0.011**
Finetuning		
ST-MEM	0.994 ± 0.001	**0.937 ± 0.001**
ECG-JEPA	**0.996 ± 0.000**	0.909 ± 0.004
ECGFounder	0.081 ± 0.006	0.011 ± 0.001
SimDINOv2 Transformer	0.983 ± 0.001	0.900 ± 0.001
xECG	0.995 ± 0.000	0.921 ± 0.001

Table 5 R-peak. Individual model performance for the R-peak detection task (MIT-BIH). Each model is trained 5 times and we report the mean± standard deviation. Pairwise differences between models were assessed using two-sided Welch's t-tests across independent runs, with significance defined as $p < 0.05$. Bold results indicate best performance across models and underlined indicates 2nd best.

Methods	Exercise (R-peak) F1 Score (150ms)	F1 Score (20ms)
Supervised		
Supervised xLSTM	0.990 ± 0.002	0.966 ± 0.002
Linear Probing		
ST-MEM	0.822 ± 0.018	0.627 ± 0.008
ECG-JEPA	0.924 ± 0.004	0.526 ± 0.007
ECGFounder	0.421 ± 0.016	0.083 ± 0.002
SimDINOv2 Transformer	0.603 ± 0.003	0.301 ± 0.001
xECG	**0.948 ± 0.004**	**0.892 ± 0.004**
Finetuning		
ST-MEM	0.988 ± 0.002	**0.986 ± 0.001**
ECG-JEPA	0.975 ± 0.002	0.859 ± 0.007
ECGFounder	0.423 ± 0.006	0.082 ± 0.005
SimDINOv2 Transformer	0.853 ± 0.008	0.633 ± 0.013
xECG	**0.996 ± 0.001**	0.968 ± 0.002

Table 6 Exercise. Individual model performance for the Exercise task (R-peak detection in Exercise-ECG). Each model is trained 5 times and we report the mean± standard deviation. Pairwise differences between models were assessed using two-sided Welch's t-tests across independent runs, with significance defined as $p < 0.05$. Bold results indicate best performance across models and underlined indicates 2nd best.

Methods	Physionet Sleep Apnea-ECG Accuracy	F1 Score	AUROC
Supervised			
Chang et al.[33]	0.920	0.879	0.865
Shen et al. [71]	0.894	0.866	0.946
Bernardini et al. [15]	0.936	0.916	0.981
Supervised xLSTM	0.778 ± 0.024	0.668 ± 0.059	0.853 ± 0.022
Linear Probing			
ST-MEM	0.617 ± 0.039	0.609 ± 0.036	0.623 ± 0.041
ECG-JEPA	0.473 ± 0.019	0.456 ± 0.030	0.543 ± 0.006
ECGFounder	0.546 ± 0.043	0.538 ± 0.048	0.580 ± 0.041
SimDINOv2 Transformer	0.683 ± 0.012	0.650 ± 0.029	0.648 ± 0.032
xECG	**0.786 ± 0.021**	**0.700 ± 0.032**	**0.856 ± 0.025**
Finetuning			
ST-MEM	0.620 ± 0.034	0.618 ± 0.036	0.702 ± 0.020
ECG-JEPA	0.580 ± 0.014	0.576 ± 0.015	0.592 ± 0.031
ECGFounder	0.544 ± 0.030	0.537 ± 0.031	0.579 ± 0.026
SimDINOv2 Transformer	0.620 ± 0.033	0.617 ± 0.032	0.638 ± 0.035
xECG	**0.848 ± 0.031**	**0.767 ± 0.075**	**0.932 ± 0.014**

Table 7 Sleep Apnea. Individual model performance for the Sleep Apnea task (sleep apnea segmentation Apnea-ECG). Each model is trained 5 times and we report the mean± standard deviation. Pairwise differences between models were assessed using two-sided Welch's t-tests across independent runs, with significance defined as $p < 0.05$. Bold results indicate best performance across models and underlined indicates 2nd best.

Methods	DeepBeat PPG		
	Accuracy	F1 Score	AUROC
Supervised			
Supervised xLSTM	0.771 ± 0.016	0.733 ± 0.012	0.876 ± 0.006
Linear Probing			
ST-MEM	**0.667 ± 0.048**	0.560 ± 0.018	0.653 ± 0.006
ECG-JEPA	0.524 ± 0.010	0.448 ± 0.007	0.477 ± 0.006
ECGFounder	**0.714 ± 0.064**	0.491 ± 0.072	0.641 ± 0.019
SimDINOv2 Transf.	0.618 ± 0.034	0.562 ± 0.010	0.656 ± 0.003
xECG	**0.701 ± 0.061**	**0.616 ± 0.034**	**0.751 ± 0.013**
Finetuning			
ST-MEM	0.616 ± 0.065	0.555 ± 0.030	0.643 ± 0.049
ECG-JEPA	0.698 ± 0.029	0.660 ± 0.022	0.818 ± 0.021
ECGFounder	**0.722 ± 0.048**	0.686 ± 0.031	0.846 ± 0.012
SimDINOv2 Transf.	0.660 ± 0.034	0.626 ± 0.020	0.780 ± 0.014
xECG	**0.778 ± 0.024**	**0.738 ± 0.018**	**0.887 ± 0.017**

Table 8 PPG AF. Individual model performance for the PPG AF task (AF classification in the PPG dataset DeepBeat). Each model is trained 5 times and we report the mean± standard deviation. Pairwise differences between models were assessed using two-sided Welch's t-tests across independent runs, with significance defined as $p < 0.05$. Bold results indicate best performance across models and underlined indicates 2nd best.

Methods	Age Prediction			
	MAE (MIMIC-IV)	MAE (PTB-XL)	MAE (CPSC)	MAE (mean)
Supervised				
Supervised xLSTM	9.464 ± 0.124	8.229 ± 0.081	8.754 ± 0.121	8.594 ± 0.089
Linear Probing				
ST-MEM	11.566 ± 0.025	9.804 ± 0.004	11.151 ± 0.026	11.208 ± 0.016
ECG-JEPA	9.980 ± 0.018	8.389 ± 0.008	8.899 ± 0.017	8.809 ± 0.004
ECGFounder	**9.298 ± 0.019**	8.098 ± 0.017	8.595 ± 0.014	**8.402 ± 0.009**
SimDINOv2 Transf.	10.284 ± 0.025	9.253 ± 0.025	9.959 ± 0.025	9.735 ± 0.012
xECG	9.334 ± 0.011	**7.989 ± 0.006**	**8.489 ± 0.006**	8.659 ± 0.006
Finetuning				
ST-MEM	**8.247 ± 0.087**	**7.393 ± 0.123**	**7.245 ± 0.133**	**7.225 ± 0.084**
ECG-JEPA	8.849 ± 0.123	7.927 ± 0.195	8.169 ± 0.118	7.829 ± 0.100
ECGFounder	9.061 ± 0.139	7.884 ± 0.223	8.260 ± 0.168	7.970 ± 0.128
SimDINOv2 Transf.	9.866 ± 0.069	8.664 ± 0.068	9.303 ± 0.073	8.932 ± 0.032
xECG	8.818 ± 0.040	7.670 ± 0.036	7.943 ± 0.068	7.790 ± 0.028

Table 9 Age. Individual model mean absolute error (MAE) for the Age task (age regression, trained on CODE-15% and evaluated on MIMIC-IV-ECG, PTB-XL and CPSC2018). Each model is trained 5 times and we report the mean± standard deviation. Pairwise differences between models were assessed using two-sided Welch's t-tests across independent runs, with significance defined as $p < 0.05$. Bold results indicate best performance across models and underlined indicates 2nd best.

Methods	Age Prediction			
	SMAPE (MIMIC-IV)	SMAPE (PTB-XL)	SMAPE (CPSC)	SMAPE (mean)
Supervised				
Sup. xLSTM	0.079 ± 0.001	0.075 ± 0.001	0.082 ± 0.001	0.079 ± 0.001
Linear Probing				
ST-MEM	0.098 ± 0.000	0.089 ± 0.000	0.104 ± 0.000	0.105 ± 0.000
ECG-JEPA	0.083 ± 0.000	0.076 ± 0.000	0.083 ± 0.000	0.081 ± 0.000
ECGFounder	**0.077 ± 0.000**	0.073 ± 0.000	0.081 ± 0.000	**0.078 ± 0.000**
SimDINOv2 Trans.	0.086 ± 0.000	0.085 ± 0.000	0.094 ± 0.000	0.091 ± 0.000
xECG	0.078 ± 0.000	**0.073 ± 0.000**	**0.080 ± 0.000**	0.081 ± 0.000
Finetuning				
ST-MEM	**0.069 ± 0.001**	**0.067 ± 0.001**	**0.068 ± 0.001**	**0.067 ± 0.001**
ECG-JEPA	0.073 ± 0.001	0.070 ± 0.002	0.075 ± 0.001	0.071 ± 0.001
ECGFounder	0.075 ± 0.001	0.072 ± 0.002	0.078 ± 0.001	0.074 ± 0.001
SimDINOv2 Trans.	0.083 ± 0.001	0.079 ± 0.001	0.088 ± 0.001	0.083 ± 0.000
xECG	0.073 ± 0.000	0.070 ± 0.000	0.075 ± 0.001	0.072 ± 0.000

Table 10 Age. Individual model symmetric mean absolute percentage error (SMAPE) for the Age task (age regression, trained on CODE-15% and evaluated on MIMIC-IV-ECG, PTB-XL and CPSC2018). Each model is trained 5 times and we report the mean± standard deviation. Pairwise differences between models were assessed using two-sided Welch's t-tests across independent runs, with significance defined as $p < 0.05$. Bold results indicate best performance across models and underlined indicates 2nd best.

Methods	Blood test		
	Accuracy	F1 Score	AUROC
Supervised			
Supervised xLSTM	0.425 ± 0.004	0.421 ± 0.003	0.683 ± 0.001
Linear Probing			
ST-MEM	0.424 ± 0.000	0.416 ± 0.000	0.708 ± 0.000
ECG-JEPA	0.432 ± 0.000	0.424 ± 0.000	0.722 ± 0.000
ECGFounder	0.430 ± 0.000	0.424 ± 0.000	0.711 ± 0.000
SimDINOv2 Transformer	0.437 ± 0.000	0.431 ± 0.000	0.724 ± 0.001
xECG	**0.439 ± 0.000**	**0.435 ± 0.001**	**0.733 ± 0.000**
Finetuning			
ST-MEM	**0.456 ± 0.002**	**0.457 ± 0.002**	0.744 ± 0.002
ECG-JEPA	0.440 ± 0.000	0.434 ± 0.001	0.735 ± 0.001
ECGFounder	0.440 ± 0.002	0.437 ± 0.003	0.714 ± 0.003
SimDINOv2 Transformer	0.445 ± 0.004	0.443 ± 0.005	0.725 ± 0.001
xECG	0.451 ± 0.001	0.450 ± 0.002	**0.747 ± 0.002**

Table 11 Blood test. Individual model performance for the Blood test task ((ab)normal classification of blood test results from MIMIC-IV-ECG). Each model is trained 5 times and we report the mean± standard deviation. Pairwise differences between models were assessed using two-sided Welch's t-tests across independent runs, with significance defined as $p < 0.05$. Bold results indicate best performance across models and underlined indicates 2nd best.

Methods	Survival Analysis		
	CI (CODE-15, Val)	CI (MIMIC-IV-ECG, Test)	CI (difference)
Supervised			
Supervised xLSTM	0.650 ± 0.018	0.630 ± 0.016	0.020 ± 0.015
Linear Probing			
ST-MEM	0.755 ± 0.005	0.697 ± 0.001	**0.058 ± 0.006**
ECG-JEPA	**0.830 ± 0.004**	**0.706 ± 0.001**	0.124 ± 0.004
ECGFounder	0.817 ± 0.004	0.683 ± 0.001	0.134 ± 0.004
SimDINOv2 Transformer	0.779 ± 0.005	0.695 ± 0.001	0.084 ± 0.005
xECG	0.811 ± 0.005	0.703 ± 0.001	0.108 ± 0.005
Finetuning			
ST-MEM	0.774 ± 0.003	0.694 ± 0.004	**0.080 ± 0.000**
ECG-JEPA	**0.827 ± 0.004**	0.706 ± 0.001	0.122 ± 0.005
ECGFounder	**0.822 ± 0.005**	0.688 ± 0.002	0.134 ± 0.003
SimDINOv2 Transformer	0.785 ± 0.007	0.701 ± 0.004	**0.084 ± 0.008**
xECG	0.817 ± 0.005	**0.710 ± 0.001**	0.107 ± 0.005

Table 12 Mortality. Concordance Index (CI) for individual models for the Mortality task (survival analysis for mortality prediction, trained on CODE-15% and evaluated on MIMIC-IV-ECG). Each model is trained 5 times and we report the mean± standard deviation. Pairwise differences between models were assessed using two-sided Welch's t-tests across independent runs, with significance defined as $p < 0.05$. Bold results indicate best performance across models and underlined indicates 2nd best.

Downstream tasks hyperparameters

This section details the hyperparameters and data processing procedures used for linear probing and fine-tuning on all downstream tasks. It is important to note that ECGFounder, as a convolutional-based model, does not utilize drop path or layer-wise learning rate decay. No data augmentations were applied to any downstream task, with the exception of the DeepBeat PPG dataset, which was only available in a pre-augmented format for the training split. For all downstream tasks, we used the AdamW optimiser with a learning rate scheduler. The first epoch was used for learning rate warm-up, after which the learning rate was progressively reduced to zero following a cosine schedule by the final epoch.

Patch representation is specific to our transformer and xLSTM-based models (xECG, SimDINOv2, and Supervised xLSTM). This hyperparameter determines the pooling method applied to patch embeddings to generate the final signal representation. We employed two pooling strategies: average pooling (avg), which computes the mean of all patch embeddings, and max pooling (max), which takes the maximum value across the sequence for each feature dimension. For all the tasks and all the methods we selected the best model on the validation set on the metric used for the BenchECG score.

PTB-XL. For the PTB-XL dataset, a batch size of 256 was used for linear probing across all models. During fine-tuning, the batch size was reduced for certain models due to memory constraints. Table 13 provides a comprehensive list of all hyperparameters for this task.

Hyperparameter	**PTB-XL**					
	ECGFounder	ECG-JEPA	ST-MEM	SimDINOv2 Transf.	xECG	Sup. xLSTM
Linear Probing						
learning rate head	0.001	0.001	0.001	0.001	0.001	-
patch representation	-	-	-	avg	avg	-
Finetuning						
learning rate head	0.001	0.001	0.0003	0.001	0.001	0.0001
learning rate encorder	0.00001	0.00001	0.0003	0.00001	0.00003	0.0001
layerwise lr decay	-	0.75	0.75	0.75	0.75	-
drop path	-	0.5	0.5	-	0.5	-
batch size	256	96	128	256	256	256
patch representation	-	-	-	avg	avg	avg

Table 13 Hyperparameters for the PTB-XL task.

CPSC2018. In the nine-class multi-label classification task on the CPSC2018 dataset, all models were trained for 30 epochs with a weight decay of 0.1. The learning rates varied depending on the specific model. A batch size of 256 was used for linear probing, while for fine-tuning, it was adjusted for some models to accommodate memory limitations. Because this dataset contain signals with different lenghts, spanning from 7 to 60 seconds we cropped signals to 10 seconds and applied padding if they were smaller than that. The complete hyperparameter settings for this task are presented in Table 14.

MIT-BIH. For the beat-level classification task on the MIT-BIH dataset, the training procedures differed significantly across the various methods, reflecting the R-peak level annotations. For the transformer and xLSTM-based models, which have sufficiently small patch sizes (ECG-JEPA: 20ms, ST-MEM: 300ms, our baselines: 250ms), we adopted a patch-level classification approach. This ensures that each patch contains at most one heartbeat. During training, we utilised overlapping windows centered around the R-peak, with the window length being model-dependent. For validation and testing, non-overlapping windows were used. Our recurrent baselines processed the entire signal as a single sample. ECGFounder, due to its architecture, does not support a fine-grained, beat-level output. Therefore, this model was trained by feeding it single heartbeats, each within a 1-second window centered on the corresponding R-peak. Because this dataset consists of signals recorded with a two leads configuration, the missing leads are set to zero for the whole duration of the ecg. For both linear probing and fine-tuning on this dataset, models were trained for 20 epochs. A consistent learning rate of 0.001 was used for linear probing, while different learning rates were employed for fine-tuning. The

Hyperparameter	CPSC2018					
	ECGFounder	ECG-JEPA	ST-MEM	SimDINOv2 Transf.	xECG	Sup. xLSTM
Linear Probing						
learning rate head	0.001	0.001	0.001	0.001	0.001	-
final normalisation	-	-	-	batch	-	-
patch representation	-	-	-	max	max	-
Finetuning						
learning rate head	0.001	0.001	0.0003	0.001	0.001	0.0001
learning rate encorder	0.00001	0.00001	0.0003	0.00001	0.00003	0.0001
layerwise lr decay	-	0.75	0.75	0.75	0.75	-
drop path	-	0.5	0.5	-	0.5	-
batch size	256	96	128	256	128	128
final normalisation	-	-	-	batch norm	-	-
patch representation	-	-	-	max	max	max

Table 14 Hyperparameters for classification task on CPSC2018.

batch size for linear probing was 256 and was adjusted during fine-tuning based on memory availability. All hyperparameters for the MIT-BIH classification task are detailed in Table 15 .

R-peak. For the R-peak detection task, a similar windowing approach was utilised. However, a key distinction was the use of non-overlapping windows during the training phase for all models. For our recurrent models, this non-overlapping window strategy was also applied during evaluation, instead of processing the entire 30-minute signal. ST-MEM and ECG-JEPA output 12 and 8 patch embeddings, respectively, for the same segment of ECG originally divided into patches. We then use the patch embedding from lead II as the input to the detection head. To enable detection at a signal-pixel level, a detection head was appended to the models. For the transformer and xLSTM-based architectures, this head produces an output with a dimensionality matching the input patch size, thereby allowing for fine-grained temporal localization. In the case of ECGFounder, which lacks inherent spatial correspondence in its output due to its convolutional nature, a prediction head was applied that matched the size of its accepted input window. For both linear probing and fine-tuning on this dataset, models were trained for 30 epochs. All the hyperparameters are shown in Table 16

Hyperparameter	MIT-BIH (classification)					
	ECGFounder	ECG-JEPA	ST-MEM	SimDINOv2 Transf.	xECG	Sup. xLSTM
Linear Probing						
learning rate head	0.001	0.001	0.001	0.001	0.001	-
weight decay	0.1	0.1	0.1	0.1	0.1	-
window len	10s	10s	10s	10s	36s	-
patch representation	-	-	-	avg	avg	-
Finetuning						
learning rate head	0.001	0.001	0.001	0.01	0.001	0.0001
learning rate encorder	0.00001	0.00001	0.001	0.00001	0.000001	0.0001
weight decay	0.1	0.1	0.1	0.1	0.0	0.0
layerwise lr decay	-	0.75	0.75	0.75	0.75	-
drop path	0.0	0.5	0.5	0.5	0.0	0.0
batch size	256	96	96	96	64	64
window len	10s	10s	10s	10s	36s	36s

Table 15 Hyperparameters for each group.

Exercise. On the ECG in High Intensity Dataset we followed the same procedure of the R-peak detection task of MIT-BIH. The only difference is that in this dataset each recording is of 20 seconds so window sizes matches the 20 seconds for our recurrent models and 10 seconds for all the others. Training was done for 30 epochs and a batch size of 8 (due to the very small amount of data). All the other parameters are shown in Table 17. The ECGs in this dataset consist only of lead II recordings, thus we padded to zero the rest of the leads for all the model except for ECGFounder where another version of the model was released accepting one single lead.

Hyperparameter	MIT-BIH (R-peak detection)					
	ECGFounder	ECG-JEPA	ST-MEM	SimDINOv2 Transf.	xECG	Sup. xLSTM
Linear Probing						
learning rate head	0.001	0.01	0.001	0.01	0.01	-
weight decay	0.1	0.1	0.1	0.1	0.1	-
window len	10s	10s	10s	10s	1m12s	-
patch normalisation	-	-	-	layer norm	-	-
Finetuning						
learning rate head	0.001	0.001	0.01	0.01	0.001	0.01
learning rate encorder	0.00001	0.001	0.01	0.001	0.001	0.01
weight decay	0.1	0.1	0.1	0.1	0.0	0.0
layerwise lr decay	-	0.75	0.75	0.75	0.75	0.75
drop path	-	0.5	0.5	0.5	-	0.0
batch size	256	96	96	96	64	64
window len	10s	10s	10s	10s	1m12s	1m12s

Table 16 Hyperparameters for each group.

Hyperparameter	ECG in Intense Exercise Dataset (R-peak detection)					
	ECGFounder	ECG-JEPA	ST-MEM	SimDINOv2 Transf.	xECG	Sup. xLSTM
Linear Probing						
learning rate head	0.001	0.1	0.01	0.1	0.1	-
weight decay	0.1	0.1	0.1	0.1	0.1	-
window len	10s	10s	10s	10s	20s	-
patch normalisation	-	-	-	layer norm	-	-
Finetuning						
learning rate head	0.001	0.01	0.01	0.1	0.01	0.001
learning rate encorder	0.00001	0.001	0.01	0.001	0.001	0.001
weight decay	0.1	0.1	0.1	0.1	0.0	0.0
layerwise lr decay	-	0.75	0.75	0.75	0.75	-
drop path	-	0.5	0.5	0.5	0.5	0.0
window len	10s	10s	10s	10s	20s	20s

Table 17 Hyperparameters for each group.

Sleep Apnea. Training on the Sleep Apnea-ECG dataset required different approaches for recurrent and non-recurrent models due to the minute-level annotations. For non-recurrent models that accept 10-second inputs, each overnight signal was divided into non-overlapping 10-second segments, each inheriting the label of its corresponding minute. A linear head on the final signal representation was used for prediction, and the loss was calculated for each 10-second segment during training. For evaluation, predictions from the six segments within each minute were aggregated and averaged. Conversely, our recurrent models (xECG and Supervised xLSTM) were fed 3-minute signal segments to leverage their ability to process longer sequences. A prediction head was appended to each patch, and the loss was applied at the patch level. During validation and testing, these patch-level predictions were averaged at the minute level to align with the ground-truth labels. All models were trained for 20 epochs with batch sizes adjusted based on memory consumption. The corresponding hyperparameters are shown in Table 18. The ECGs in this dataset consist only of lead II recordings, thus we padded to zero the rest of the leads for all the model except for ECGFounder where the single-lead configuration was utilised.

PPG AF. For the Atrial Fibrillation classification on the DeepBeat PPG dataset, all models were trained for 30 epochs using the maximum batch size that each method could use to accelerate training. As we did not have access to the original, non-augmented data, we adopted a strategy of training on a randomly selected 10% subset of the training set for each epoch. All hyperparameters for this task are detailed in Table 19. The recordings in this dataset are single-channel, which we treated as lead II recordings. For all models except ECGFounder, we zero-padded the remaining leads, while for ECGFounder we used the single-lead configuration.

Age. On the age regression task, all models were trained for 15 epochs. The batch size for linear probing was 512, while for fine-tuning, it was reduced based on memory limits. The hyperparameters are presented in Table 20.

Hyperparameter	Sleep Apnea-ECG					
	ECGFounder	ECG-JEPA	ST-MEM	SimDINOv2 Transf.	xECG	Sup. xLSTM
Linear Probing						
learning rate head	0.001	0.001	0.001	0.0001	0.001	-
weight decay	0.1	0.1	0.1	0.1	0.1	-
batch size	256	256	256	256	3	
window len	10s	10s	10s	10s	3m	-
patch normalisation	-	-	-	layer norm	-	-
patch representation	-	-	-	avg	max	-
Finetuning						
learning rate head	0.001	0.001	0.0003	0.00001	0.001	0.0001
learning rate encorder	0.000001	0.000001	0.0003	0.000001	0.0001	0.0001
weight decay	0.1	0.1	0.1	0.1	0.1	0.0
layerwise lr decay	-	0.75	0.75	0.75	0.9	-
drop path	-	0.5	0.5	0.5	0.2	-
batch size	256	96	128	128	3	3
window len	10s	10s	10s	10s	3m	3m
patch representation	-	-	-	avg	max	avg

Table 18 Hyperparameters for each group.

Hyperparameter	DeepBeat PPG					
	ECGFounder	ECG-JEPA	ST-MEM	SimDINOv2 Transf.	xECG	Sup. xLSTM
Linear Probing						
learning rate head	0.1	0.001	0.0003	0.001	0.001	-
weight decay	0.1	0.1	0.1	0.1	0.1	-
batch size	512	96	128	256	512	
patch normalisation	-	-	-	layer norm	-	-
patch representation	-	-	-	avg	avg	-
Finetuning						
learning rate head	0.01	0.001	0.0003	0.001	0.001	0.0001
learning rate encorder	0.001	0.00001	0.0003	0.0001	0.00001	0.0001
weight decay	0.1	0.1	0.1	0.1	0.1	0.1
layerwise lr decay	-	0.75	0.75	0.75	1.0	-
drop path	-	0.5	0.5	0.5	-	-
batch size	128	96	128	256	128	128
patch normalisation	-	-	-	layer norm	-	-
patch representation	-	-	-	avg	avg	avg

Table 19 Hyperparameters for each group.

Hyperparameter	Age regression					
	ECGFounder	ECG-JEPA	ST-MEM	SimDINOv2 Transf.	xECG	Sup. xLSTM
Linear Probing						
learning rate head	0.1	0.1	0.01	0.01	0.01	-
weight decay	0.01	0.01	0.01	0.01	0.01	-
patch representation	-	-	-	avg	avg	-
Finetuning						
learning rate head	0.1	0.1	0.0003	0.01	0.01	0.001
learning rate encorder	0.01	0.01	0.01	0.01	0.1	0.01
weight decay	0.01	0.01	0.01	0.01	0.1	0.01
layerwise lr decay	-	0.75	0.75	0.75	0.75	0.75
drop path	-	0.5	-	-	0.2	0.5
batch size	512	96	128	512	512	512
patch representation	-	-	-	avg	avg	avg

Table 20 Hyperparameters for each group.

Blood test. For the blood test abnormality detection task, all the models were trained for 20 epochs. Batch size for linear probing was set to 512, while for fine-tuning, it was reduced based on memory limits. Hyperparameters are presented in Table 21

Mortality. For the mortality risk task, all the models were trained for 30 epochs. Batch size for linear probing was set to 512, while for fine-tuning, it was reduced based on memory limits. Hyperparameters are presented in Table 22

Hyperparameter	Blood test					
	ECGFounder	ECG-JEPA	ST-MEM	SimDINOv2 Transf.	xECG	Sup. xLSTM
Linear Probing						
learning rate head	0.001	0.003	0.01	0.001	0.001	-
weight decay	0.1	0.1	0.1	0.1	0.1	-
final normalisation	-	-	-	batch norm	-	-
patch representation	-	-	-	avg	avg	-
Finetuning						
learning rate head	0.001	0.001	0.001	0.001	0.001	0.0001
learning rate encorder	0.0001	0.00001	0.001	0.00001	0.00003	0.0001
weight decay	0.1	0.1	0.1	0.1	0.1	0.1
layerwise lr decay	-	0.75	0.75	0.75	0.75	-
drop path	-	0.5	-	-	0.2	0.5
batch size	512	96	96	512	512	512
patch representation	-	-	-	avg	avg	avg

Table 21 Hyperparameters for each group.

Hyperparameter	Mortality					
	ECGFounder	ECG-JEPA	ST-MEM	SimDINOv2 Transf.	xECG	Sup. xLSTM
Linear Probing						
learning rate head	0.001	0.001	0.01	0.001	0.001	-
weight decay	0.1	0.1	0.1	0.1	0.1	-
final normalisation	-	-	-	batch norm	-	-
patch representation	-	-	-	avg	avg	-
Finetuning						
learning rate head	0.001	0.001	0.0003	0.001	0.001	0.001
learning rate encorder	0.0001	0.000001	0.00003	0.00001	0.000001	0.0001
weight decay	0.1	0.1	0.1	0.1	0.1	0.1
layerwise lr decay	-	0.75	0.75	0.75	0.75	-
drop path	-	0.5	0.0	0.0	0.0	-
batch size	512	96	96	512	512	512
final normalisation	-	-	-	batch norm	-	-
patch representation	-	-	-	avg	avg	avg

Table 22 Hyperparameters for each group.

Pre-training hyperparameters

Our xECG model was pretrained using the SimDINOv2 framework for 100 epochs with a batch size of 512. The learning rate was initialised to 0.0001 and managed by a cosine annealing scheduler, which included a five-epoch linear warm-up phase. We applied a layer-wise learning rate decay of 0.9, and gradient clipping was set to a maximum norm of 3.0. Both weight decay and the teacher model's exponential moving average (EMA) momentum (λ_t) were scheduled. Weight decay increased linearly from 0.04 to 0.4 over the course of training, while the EMA momentum was scheduled to increase from 0.99 to 1.0 by the final epoch.

For the self-distillation process, we generated two global views (random crops of 80% of the signal length) and four local views (random crops of 40%) for each input sample. Patches within the student model's input views were masked with a probability of 0.3. Random lead dropout was used with a probability of 0.2 for all leads except for lead II. Multiplicative Gaussian jitter (amplitude A=0.6) and random amplitude scaling (range $R = 0.2$) were each applied with a probability of 0.1.

The xECG model operates on signals sampled at a frequency (f_m) of 100 Hz. The input signal is divided into patches, where each patch consists of 25 time points, corresponding to a duration of 250 ms. The transformer baseline was pretrained using the same set of hyperparameters described above. The sole architectural difference was the inclusion of learnable positional embeddings, which are necessary for the transformer but not for our recurrent xECG model.