PRETRAINING STRATEGIES AND SCALING FOR ECG FOUNDATION MODELS: A SYSTEMATIC STUDY

Anonymous authors
Paper under double-blind review

ABSTRACT

Specialized foundation models are beginning to emerge in various medical subdomains, but pretraining methodologies and parametric scaling with the size of the pretraining dataset are rarely assessed systematically and in a like-for-like manner. This work focuses on foundation models for electrocardiography (ECG) data, one of the most widely captured physiological time series worldwide. We present a comprehensive assessment of pretraining methodologies, covering five different contrastive and non-contrastive self-supervised learning objectives for ECG foundation models, and investigate their scaling behavior with pretraining dataset sizes up to 11M input samples, exclusively from publicly available sources. Pretraining strategy has a meaningful and consistent impact on downstream performance, with contrastive predictive coding (slightly ahead of JEPA) yielding the strongest overall downstream performance across diverse clinical tasks. Scaling pretraining data continues to yield meaningful improvements up to 11M samples for most objectives and translates into improved downstream performance. We also compare model architectures across all pretraining methodologies and find evidence for a clear superiority of structured state space sequence (S4) models compared to Transformers and CNN models. We hypothesize that the strong inductive biases of S4 models, rather than pretraining scale alone, are the primary driver of effective ECG representation learning, with important implications for future foundation model development in this and potentially other physiological signal domains.

1 INTRODUCTION

Motivation Foundation models (FMs), defined as "models trained on broad data (generally using self-supervision at scale) that can be adapted to a wide range of downstream tasks" (Bommasani et al., 2021), are increasingly gaining traction, including in specialized medical domains (Chen et al., 2024; Zhou et al., 2023), offering specific advantages over training from scratch: (i) higher predictive performance, (ii) improved label efficiency, and (iii) the ability to serve as frozen feature extractors for downstream tasks. In this paper, we shed light on the training of FMs for electrocardiogram (ECG) data. The ECG represents a widely used non-invasive tool for a first-line assessment of the cardiac state and systemic physiology (Siontis et al., 2021). For example, the ECG can be used to detect cardiovascular conditions such as myocardial infarction (Strodthoff et al., 2020), to evaluate cardiovascular risk (Bhatia & Dorian, 2018), and to guide clinical decisions (Rokos et al., 2010).

Research gap The development of ECG FMs has flourished recently (Li et al., 2025; Kim, 2024; Na et al., 2024; Liu et al., 2024; Tian et al., 2024; Coppola et al., 2024; McKeen et al., 2025; Hung et al., 2025; Lunelli et al., 2026; Nolin-Lapalme et al., 2026; Yu et al., 2024; Chen et al., 2025). While comprehensive benchmarking of these models has been a major limitation for a long time, it has been addressed recently by different benchmarking studies (Lunelli et al., 2026; Al-Masud et al., 2026; Wan et al., 2025; Taleshinosrati et al., 2026). This resolved the question of measuring FM performance in a comprehensive manner, but failed to answer central questions on pretraining methodology and the effect of the scale of pretraining data. This is because different FMs differ simultaneously in pretraining methodology, backbone architecture and pretraining data, which makes it impossible to attribute performance differences to any single factor. This paper aims to close this research gap by means of a dedicated study on the pretraining of ECG FMs. We believe this study provides practical guidance for future FM development in the ECG domain and potentially

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

beyond, helping to avoid suboptimal design choices. We focus on FM training purely based on self-supervision. We acknowledge the value of direct/weak supervision, which we envision to be relevant for later training stages of the FM development, but which are not part of this investigation.

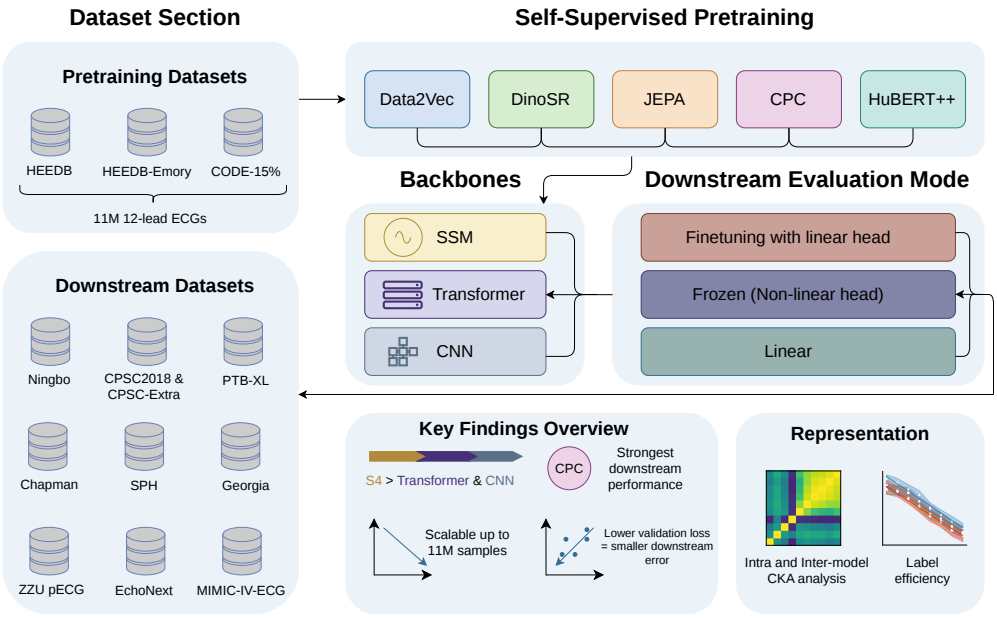


Figure 1: Schematic overview of the design of the study: This work provides a comprehensive assessment of the pretraining of ECG foundation models through self-supervision alone covering pretraining methodologies, model architectures, scaling behavior, label efficiency and representation similarity.

Research questions More specifically, we address the following research questions: **(RQ1)** How does the choice of the model architecture impact the performance of ECG FMs? Here we are interested in investigating the broad classes CNN vs. Transformer vs. state space models. **(RQ2)** Furthermore, what is the impact of the pretraining methodology on the performance and also on the internal structure of ECG FMs? **(RQ3)** Finally, what is the effect of scale, in terms of the size of the pretraining datasets, on the performance of ECG FMs?

Contributions In this paper, we put forward the following technical contributions, which we also summarize in Figure 1: **(1)** We provide the most comprehensive pretraining study on ECG FMs to date, covering five different pretraining objectives trained on the largest publicly available pretraining dataset, comprising over 11M samples, evaluated through a comprehensive evaluation methodology following Al-Masud et al. (2026) in a first like-for-like comparison (see Section 3). **(2)** We provide insights on model architecture comparisons CNN vs. Transformer vs. state space models, confirming an instantiation of the latter (S4) as the superior architecture choice across all pretraining paradigms, in line with comparisons in a supervised setting (see "Backbone analysis" in Section 4). **(3)** We provide comparative insights on five pretraining methodologies popularized in the general self-supervised learning domain, or for the domain of time series/audio data (see Figure 3). Pretraining strategy has a meaningful and consistent impact on downstream performance, with CPC showing the strongest overall downstream performance across diverse clinical tasks, while data2vec consistently lags behind across all evaluation modes and scaling regimes. **(4)** We provide a scaling analysis across all five pretraining methodologies and identify scaling behavior, most clearly for CPC and JEPA. Furthermore, we provide evidence that lower pretraining loss correlates with lower residual errors in downstream tasks (see Figure 4). We further extend the scaling analysis for CPC and JEPA to two architectures per family, S4 and Mamba-2 for SSMs and Transformers with RoPE and sinusoidal positional embeddings, to distinguish scaling behavior specific to a particular model from behavior that holds across an architecture family. **(5)** We complement the analysis

with a representational similarity analysis, which provides additional hints for the superiority of the CPC approach, a cross-dataset transfer evaluation, label/model efficiency analyses and improved finetuning procedures.

2 BACKGROUND

ECG FMs ECG FMs have proliferated rapidly, spanning masked prediction (Na et al., 2024; Coppola et al., 2024), joint-embedding architectures (Kim, 2024), and contrastive and multimodal learning (Liu et al., 2024; Yu et al., 2024; McKeen et al., 2025), with recent models pretrained on around ten million recordings (Li et al., 2025; Coppola et al., 2024). The SSL objectives underlying these models are largely inherited from speech (CPC (Oord et al., 2018), HuBERT (Hsu et al., 2021)) and vision (I-JEPA (Assran et al., 2023)), and have been adapted to physiological time series without systematic evaluation of which objectives or backbone architectures are best suited for ECG data. Backbone choices have similarly followed trends from other domains, with Transformers (Vaswani et al., 2017) dominating despite structured state space models (SSMs) (Gu et al., 2022) showing superior performance on long sequences in supervised ECG settings (Mehari & Strodthoff, 2023; Strodthoff et al., 2024).

Scaling laws Scaling laws relating model performance to pretraining dataset size have been studied extensively in language (Hoffmann et al., 2022) and vision (Nezhurina et al., 2025), typically revealing power-law improvements with increasing data. In the ECG domain, however, whether large pretraining corpora are necessary for strong FM performance or whether carefully designed objectives and architectures can compensate for limited data remains poorly understood, with direct practical implications given the cost of curating large-scale medical datasets. Concurrently, Li et al. (2026) study loss-to-loss scaling laws relating pretraining ID loss to OOD loss across model sizes up to 200M parameters on up to 2.3M records, without systematically relating loss to downstream performance. We instead scale pretraining data to 11M samples across five SSL objectives and backbones including SSMs, and directly quantify the loss–downstream-performance correlation across three clinical tasks.

ECG FM benchmarking Benchmarking of ECG FMs has gained increasing attention recently (Lunelli et al., 2026; Al-Masud et al., 2026; Wan et al., 2025; Taleshinosrati et al., 2026), establishing standardized evaluation protocols and highlighting the importance of strong supervised baselines. However, existing benchmarks evaluate FMs that differ simultaneously in pretraining objective, backbone architecture, and dataset scale, making it impossible to attribute performance differences to any single factor. This paper directly addresses this gap by conducting the first comprehensive, like-for-like comparison of ECG FMs in terms of pretraining objectives and scaling behavior along with consequences for downstream performance, holding architecture and training configuration constant across all conditions.

3 METHODS

Backbone architectures All models share a common encoder comprising a lightweight CNN stem followed by a sequential backbone, with the encoder fixed across all pretraining objectives to enable controlled comparison. The CNN stem consists of four convolutional layers with batch normalization. For the backbone, we consider three families: SSMs (S4 (Gu et al., 2022) and Mamba-2 (Dao & Gu, 2024)), Transformers (with rotary (Su et al., 2024) and sinusoidal (Vaswani et al., 2017) positional embeddings), and CNNs (Net1D (Li et al., 2025)). For all three architectural families, we performed ablations in terms of model type, dimension/depth, learning rate, input size (App. A.9) in supervised manner while keeping a balanced hyperparameter search budget across all of them.

Self-supervised pretraining objectives We investigate five self-supervised pretraining objectives spanning contrastive, predictive, and clustering-based paradigms within a unified architecture and training framework, enabling controlled comparison of their effectiveness for ECG representation learning. **data2vec** (Baevski et al., 2022) trains the model to predict the EMA teacher's averaged top-k contextualized layer representations at masked positions, providing continuous rather than

discrete supervision, with the teacher averaging the outputs of the top 2 S4 layers. **DinoSR** (Liu et al., 2023) trains a student encoder to predict EMA teacher-derived discrete cluster assignments at masked positions, combining online k-means quantization with a masked prediction objective using two codebooks of size 256 each. **JEPA** (Assran et al., 2023) trains a context encoder to predict the EMA teacher's latent representations of masked regions entirely in embedding space via a multi-block masking strategy, avoiding low-level reconstruction. **CPC** (Oord et al., 2018) learns representations by predicting future latent states from causal context via a contrastive objective, discriminating true future steps from within-sequence negatives. **HuBERT++** extends the speech HuBERT (Hsu et al., 2021) approach by replacing offline hard k-means targets with EMA teacher-derived soft cluster assignments computed via Sinkhorn-Knopp optimal transport (Cuturi, 2013), encouraging balanced token participation and preventing representation collapse through online prototype updates (App. A.3). We additionally evaluate **ECGFounder** (Li et al., 2025) and **ECG-JEPA** (Kim, 2024), two of the best-performing external FMs in the benchmark of Al-Masud et al. (2026), as external FM reference points, and a supervised S4 model trained from scratch as a strong task-specific baseline. The overall best-performing ECG-CPC model from Al-Masud et al. (2026) is omitted for comparison as the architecture and training setup largely coincides with CPC model studied in this work (App. A.5.2).

Pretraining datasets We pretrain all models on HEEDB, HEEDB-Emory (Koscova et al., 2026b;a), and CODE-15% (Ribeiro et al., 2021) at 240 Hz, spanning three scales: a small HEEDB subset (106K), a medium HEEDB subset (753K), and the full combined dataset (11M samples). We additionally compare HEEDB against MIMIC-IV-ECG (Gow et al., 2023) as an alternative pretraining source at matched scale, i.e., 753K vs. 759K samples (App. A.9.4) with largely similar downstream performance.

Downstream datasets and tasks We follow the evaluation protocol from Al-Masud et al. (2026), which covers 26 clinically relevant tasks using 10 public datasets comprising 1,616 regression and classification targets. The tasks were assigned to 7 task categories to simplify the interpretation of the results. The categories cover adult/pediatric ECG interpretation, cardiac structure, cardiac and non-cardiac outcomes, acute care and patient characteristics, see Table 2 and App. A.1 for details. Deviating from Al-Masud et al. (2026), we removed two evaluation datasets: CODE-15% as it was used for pretraining in this work and PTB due to its small size.

Evaluation Pretrained encoders are augmented with a linear prediction head and finetuned on each downstream task using AdamW with layer-dependent learning rates, binary cross-entropy for classification, and MAE for regression with z-normalized targets. For linear evaluation, the encoder is kept frozen and only the linear prediction head is trained. For frozen evaluation, the encoder is likewise kept frozen, but the linear head is replaced with a learnable query-attention head (Bardes et al., 2024). We report macro-averaged AUROC for classification and MAE for regression, with statistical significance assessed via pairwise empirical bootstrapping on the test set with 1,000 iterations at a 95% confidence level. Rankings account for confidence interval overlaps, with ties indicating no statistically significant difference. (Macro-)AUROC serves as the primary classification metric, as it is the most widely used metric to characterize the overall discriminative power of a model, even in the presence of label imbalance (McDermott et al., 2024).

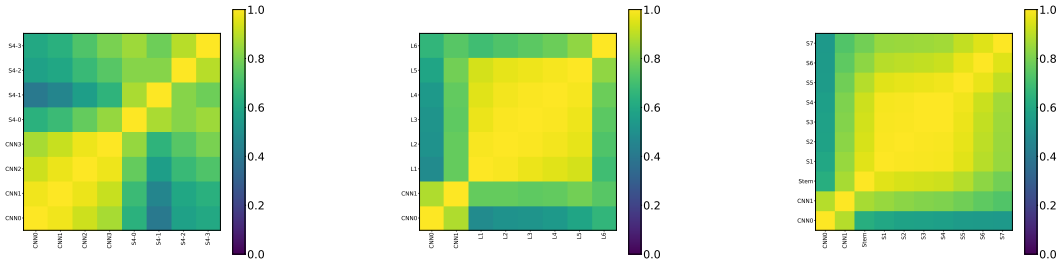
We provide an extensive experimental setup describing details of the model architectures, pretraining and evaluation procedures in App. A.2.

4 Results

Comprehensive evaluation results across finetuning, frozen, and linear evaluation modes are compiled in Tables 4, 5, and 6 in the appendix, where boldface entries denote models not performing statistically significantly worse than the best method per task. Statistical rankings aggregated by task category are visualized in Figure 3 (see Table 8), with per-task ranked lists provided in Table 7.

Backbone analysis: performance Across all five pretraining objectives, the 4-layer S4 backbone consistently outperforms both the Transformer and Net1D backbones on the large majority of downstream tasks and performs on par on the remaining tasks (App. A.6). The performance gap over the

Transformer is most pronounced for JEPA, CPC, and HuBERT++, and is particularly evident on challenging tasks such as pediatric ECG interpretation and cardiac structure prediction. The CNN backbone Net1D performs competitively under data2vec but falls consistently below S4 for all other pretraining objectives. This is further corroborated by the underperformance of the CNN-based FMs MERL (Liu et al., 2024) and ECGFM-KED (Tian et al., 2024) relative to our S4-based models (App. A.5.3), consistent with recent evidence questioning the suitability of CNNs as backbone architectures for physiological time series in supervised settings (Mehari & Strodthoff, 2023; Strodthoff et al., 2024; Al-Masud et al., 2026).



(a) S4 (b) Transformer (c) Net1D

Figure 2: Intra-model layer-wise representation similarity across architectures pretrained using JEPA, measured by CKA (Kornblith et al., 2019) with a Gaussian RBF kernel ($\sigma = 1.0$) on 2,500 PTB-XL samples. Warmer colors indicate greater similarity between layer pairs.

Backbone analysis: representational similarity The intra-model CKA analysis (Kornblith et al., 2019) (Figure 2 and App. A.6.2) provides further insight into the representational structure learned by each backbone. For the S4 backbone, layers develop progressively distinct representations across all pretraining objectives, indicating that each layer captures complementary patterns, a desirable property for hierarchical feature learning. In contrast, Transformer backbones exhibit high inter-layer similarity across intermediate blocks (L1–L5) for most pretraining objectives, with only the final block (L6) showing moderately distinct representations under JEPA and CPC. HuBERT++ with a Transformer backbone shows virtually no layer differentiation across the entire network, suggesting representational collapse. Net1D backbones display uniform similarity in both early CNN layers and later stages across all objectives, pointing to a possible representational bottleneck. The early CNN stem layers show high mutual similarity across all backbone types and pretraining objectives, which is expected given their shared architecture. Overall, the S4 backbone is the only architecture that consistently develops distinct layer-wise representations, with CPC exhibiting the most pronounced differentiation, further supporting S4 as the preferred backbone for ECG FM pretraining.

Finetuning Under full finetuning (App. A.4.1), the external FMs ECG-JEPA and ECGFounder consistently rank among the top performers for adult ECG interpretation, frequently matching or surpassing the supervised S4 baseline. ECGFounder was pretrained using a supervised loss on diagnostic statement prediction, which aligns very closely with the adult ECG interpretation task. At this point, we stress that task-specific finetuning or alignment of ECG data with other modalities could still be applied on top of the FMs trained using self-supervision only but is beyond the scope of this paper. ECG-JEPA profits from the fact that its pretraining datasets serve as test datasets in this study. As shown below, FMs profit from continual pretraining on the target dataset before finetuning, see below, which explains performance differences in ECG-JEPAs pretraining datasets in the category of adult ECG interpretation. Among models pretrained for this paper, CPC and JEPA are the strongest competitors, often falling within the top statistical equivalence group across ECG interpretation and outcome prediction tasks. data2vec lag considerably behind across most categories, while HuBERT++ and DinoSR occupy a middle ground. In categories beyond ECG interpretation, covering cardiac structure, clinical outcomes, and patient characteristics, CPC tends to lead among all evaluated approaches, suggesting its pretraining objective yields more broadly applicable representations across diverse clinical tasks. We also pretrained on MIMIC-IV-ECG and compared against external ECG FMs (MERL (Liu et al., 2024) and ECGFM-KED (Tian et al., 2024)). Our models outperform these external FMs on matched training data, see App. A.5.3, with further gains on larger datasets.

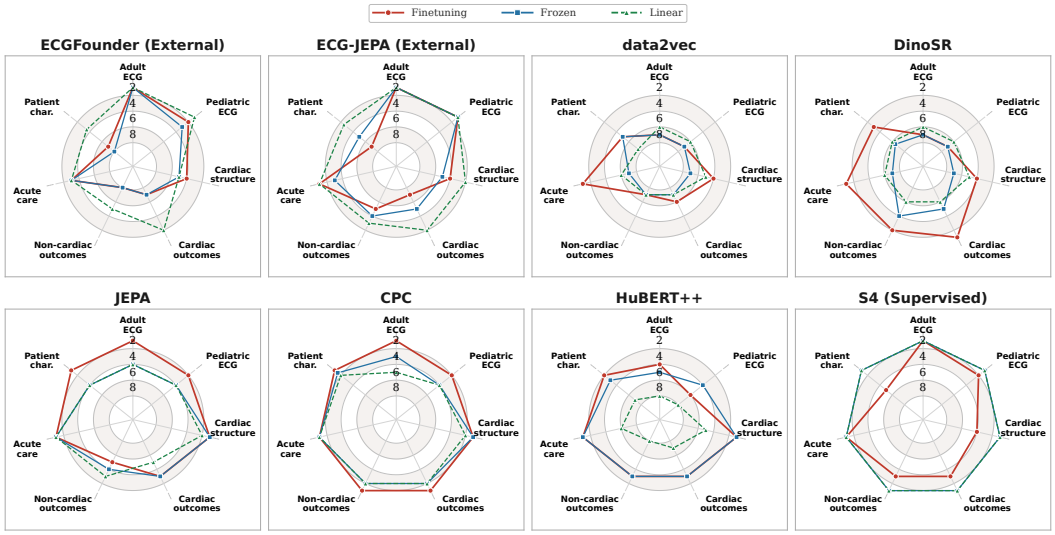
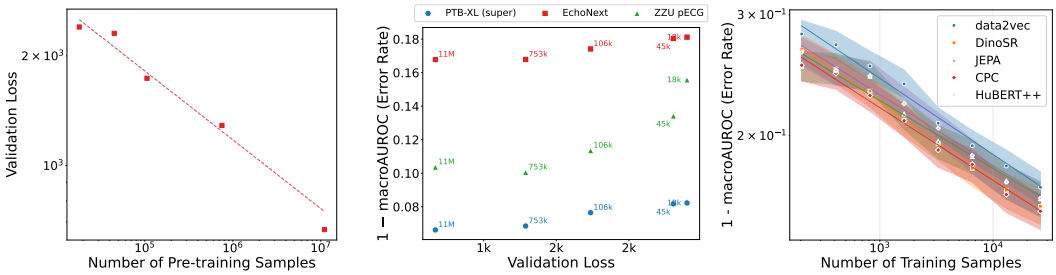


Figure 3: Visualization of performance rankings across seven downstream task categories for the five self-supervised pretraining objectives, where lower rank denotes statistically superior performance. We refer to App. A.4 for detailed classification/regression metrics on individual datasets/tasks.

Frozen and linear evaluation Rankings under frozen evaluation (App. A.4.2) largely mirror finetuning, with ECG-JEPA and ECGFounder retaining strong performance on adult ECG tasks, achieving supervised-level results as frozen feature extractors. CPC maintains its advantage in non-ECG-interpretation categories under frozen evaluation, while data2vec and DinoSR fall further behind relative to finetuning, indicating their representations require task-specific adaptation to be competitive. Under linear evaluation (App. A.4.3), the supervised S4 baseline proves difficult to surpass across most categories, with CPC showing a notable relative drop compared to its frozen evaluation performance, consistent with its token-level pretraining objective not encouraging globally discriminative pooled representations.

Cross-dataset transfer To assess robustness to distribution shift, we finetune each model on PTB-XL and apply it without further training to six external datasets, using a label mapping restricted to clinically equivalent classes (App. A.8.5). DinoSR and CPC transfer most robustly, with DinoSR in the top statistical group on all six datasets and CPC on five, while JEPA and data2vec transfer less well.

Scaling with dataset size To investigate the scaling behavior we train models on four subsets of the HEEDB dataset: 18K, 45K, 106K, and 753K, in addition to the full dataset comprising 11M samples, see App. A.2 for the specific dataset composition. For each of the pretraining methods, we first consider the conventional loss scaling with dataset size (Figure 4a and App. A.7.1). The two best-performing methods, CPC and JEPA, show the most consistent power-law scaling according to fit quality with scaling exponents 0.189 and 0.062, respectively. As a lower pretraining loss not necessarily has to align with a better downstream performance, we also investigate the scaling behavior of the residual error defined as $1 -$ macro AUROC for a given specific downstream dataset/task. We also show scaling curves for the downstream performance in App. A.7.1, the most consistent scaling behavior on PTB-XL (super) and mixed results across different other downstream datasets. Finally, we directly compare the pretraining loss to the residual error on three representative datasets (Figure 4b and App. A.7.1). CPC, JEPA and DinoSR show strong positive rank correlations between validation loss and downstream residual error on all three datasets, whereas the correlation is weaker and less consistent for HuBERT++ and data2vec. We also investigated a second member of each of the SSM and Transformer families to elucidate model-specific vs. architecture-specific scaling behavior with results hinting for the former. Interestingly, the Transformer models show a higher scaling exponents in terms of loss, however, this does not translate to better downstream performance, see App. A.7.2.



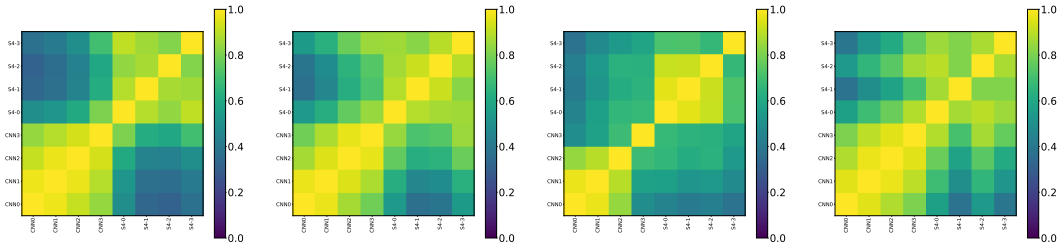
(a) Loss scaling (b) Correlation Loss/error rate (c) EchoNext label efficiency

Figure 4: Scaling Analyses for the CPC model investigating the scaling of the pretraining loss with dataset size and the relation between pretraining loss and downstream performance on selected datasets (representative for adult ECG interpretation, pediatric ECG interpretation and cardiac structure evaluation), see App. A.7 for further results. Figure 4c shows EchoNext label efficiency plot tracing downstream performance in dependence of the number of training samples.

Label efficiency In Figure 4c, we investigate a different scaling behavior, namely the dependence of the downstream performance on the size of the downstream dataset for a given pretrained FM (on the full pretraining dataset), to detect different adaptation behaviors in the small dataset regime. Scaling curves for cardiac structure prediction on EchoNext are shown in Figure 4c, with fit parameters provided in App. A.8.2. Training subsets are subsampled in powers of 2 and we fit scaling curves as above. CPC achieves the lowest error rate consistently across the entire range of training set sizes, establishing it as the most label-efficient model for this task. DinoSR starts with a relatively high error rate in the low-data regime but improves steeply with additional data, eventually surpassing JEPA and approaching CPC at the large-data end, consistent with its highest scaling exponent among all evaluated models. JEPA, by contrast, starts favorably in the low-data regime but scales more slowly, eventually converging toward HuBERT++ at larger dataset sizes. Data2vec lags behind all other models throughout the entire scaling range, indicating that its pretraining objective transfers poorly to cardiac structure prediction regardless of the amount of available labeled data. Overall, the results highlight that the choice of pretraining strategy has a pronounced effect on label efficiency, with differences being most consequential in the low-data regime.

Representation similarity analysis Intra-model CKA similarity matrices for four pretraining objectives are shown in Figure 5. All models exhibit a two-block structure reflecting the hybrid CNN-S4 architecture, with high within-block similarity and lower cross-block similarity between CNN and S4 layers. However, the sharpness of this boundary and the degree of within-block differentiation differ considerably across pretraining strategies. DinoSR shows the sharpest CNN-to-S4 transition with the most distinct separation between the two architectural components, while within each block representations remain relatively homogeneous. JEPA and HuBERT++ exhibit the highest overall representational redundancy, with a notably smoother CNN-to-S4 boundary, suggesting these pretraining objectives do not strongly encourage functional specialization between the two components. CPC stands out with the lowest overall off-diagonal similarity and the most progressive within-block differentiation, where each layer learns increasingly distinct representations forming a clear gradient from local CNN features to higher-level sequential S4 abstractions. This structured representational evolution, largely absent in other pretraining strategies, may underpin CPC's stronger and more consistent downstream performance across diverse clinical tasks. Inter-model CKA analysis across all five pretraining objectives at early, mid, and late network stages is provided in App. A.8.1, revealing the degree to which different pretraining strategies converge to or diverge from similar representational spaces across network depth.

Model efficiency Detailed computational efficiency metrics are provided in App. A.8.3. All five pretraining strategies share the same lightweight SSM backbone with only 3M parameters, operating on 600 timesteps, making them substantially more compact than the reference FMs ECGFounder (33.8M) and ECG-JEPA (87.2M). Consequently, all proposed models are considerably more efficient in terms of FLOPS, GPU memory, and inference latency than ECG-JEPA, and remain competitive with ECGFounder despite the latter's CNN-based architectural advantage in raw throughput.



(a) DinoSR (b) JEPA (c) CPC (d) HuBERT++

Figure 5: Layer-wise representation similarity within each pretraining objective. The CKA matrix for data2vec was omitted due to place constraints and is shown in App. A.8.1.

Across the five pretraining objectives, computational cost is virtually identical since the backbone is shared, meaning that performance differences observed in downstream tasks are attributable entirely to the pretraining objective rather than model capacity or computational budget. This makes our benchmark a controlled and fair comparison of pretraining strategies, and highlights that strong downstream performance can be achieved at minimal computational cost.

Continual pretraining for improved finetuning In line with positive results from other fields (Howard & Ruder, 2018), we propose to use continual pretraining on the target dataset as first finetuning phase before the actual classifier training. Domain-adaptive pretraining followed by a linear head consistently outperforms finetuning both with a standard linear and a non-linear query-attention head (Bardes et al., 2024) across nearly all tasks (App. A.8.4), in line with results of Do et al. (2026), suggesting that closing the distribution gap between pretraining and target domain is more beneficial than increasing prediction head capacity. These findings advocate for domain-adaptive pretraining as a practical strategy when downstream domain data is available. For reasons of comparability, the finetuning results underlying Figure 3 refer to standard finetuning, but we strongly advise to use continual pretraining for any downstream tasks.

HuBERT-ECG vs. HuBERT++ HuBERT++ consistently outperforms HuBERT-ECG (Coppola et al., 2024) across all evaluated tasks (App. A.5.1), with particularly pronounced improvements on adult ECG interpretation and cardiac structure prediction. We attribute this to the combination of Sinkhorn-Knopp cluster assignment preventing prototype collapse, EMA-based target generation providing more stable training signals, and the choice of S4 over Transformer as the backbone, demonstrating that careful refinement of existing objectives can yield meaningful gains without wholesale architectural innovation.

5 Discussion

S4 as preferred backbone The consistent superiority of the S4 backbone over Transformer and CNN alternatives across all five pretraining objectives provides strong evidence that architectural inductive biases are the dominant factor in ECG representation learning. The performance gap is most pronounced for JEPA, CPC, and HuBERT++, and is particularly evident on challenging tasks such as pediatric ECG interpretation and cardiac structure prediction, suggesting that the stable long-range memory and spectral filtering properties of SSMs align naturally with the temporal structure of ECG signals. This is further corroborated by the underperformance of CNN-based external FMs relative to our lightweight S4-based models despite their substantially larger parameter counts, and is mechanistically supported by the CKA analysis showing that S4 is the only backbone that consistently develops progressive and distinct layer-wise representations across all pretraining objectives. These results suggest that backbone architecture should be treated as a first-order design decision in ECG FM development.

Pretraining strategies Pretraining strategies show meaningful and consistent differences in downstream performance. CPC consistently achieves the strongest results across most task categories and evaluation modes, with particular advantages in patient characterization and outcome predic-

tion. JEPA occupies a strong second position, frequently matching CPC under finetuning but falling behind under frozen and linear evaluation. We hypothesize that the sequential nature of CPC's prediction task aligns better with the time series modality than arbitrary block masks used by other approaches. DinoSR and HuBERT++ occupy a middle ground in the performance, with selected strengths but also significant weakness in the core domain of adult and pediatric ECG interpretation. CPC and JEPA perform on par in adult/pediatric ECG interpretation, cardiac structure & function, and acute care prediction, while CPC pulls ahead specifically in cardiac and non-cardiac outcomes, and patient characteristics. Interestingly, DinoSR and CPC seem to transfer most robustly when a finetuned model is evaluated on external datasets without retraining. data2vec lags most consistently across all evaluation modes. Notably, performance gaps widen under frozen and linear evaluation, indicating that representation quality differs substantially across pretraining strategies beyond mere finetuning adaptability. The overall conclusion regarding the relative placement of SSL objectives holds broadly, but the margin and which objective wins outright varies by task category and evaluation mode. Nevertheless, the consistent superiority of the S4 backbone across all five objectives suggests that architectural choice remains the dominant factor, with pretraining strategy playing an important but secondary role.

Pretraining configuration Careful hyperparameter configuration proves critical regardless of the chosen pretraining objective. A backbone model dimension of 512 consistently outperforms both smaller and larger alternatives, and a 4-layer S4 backbone outperforms its 6-layer counterpart across most tasks, indicating that neither capacity nor depth should be maximized indiscriminately. The masking configuration is particularly consequential for JEPA, where the original hyperparameters transfer well to ECG data and yield strong downstream performance.

Scaling The consistent scaling of the pretraining loss, most notably in the cases of the two best-performing methods CPC and JEPA, is a positive sign for the further development of ECG foundation models, and the first systematic finding of this matter in this specific domain. Downstream residual errors correlate positively with pretraining loss for most objectives, most clearly for CPC and JEPA, which represents the second hopeful signal for foundation model development in the field. In spite of an inferior loss scaling compared to Transformer models, S4 models remain the preferable choice due to a consistent relation between pretraining loss and downstream residual errors.

Limitations This work is subject to a number of limitations. First, the paper focuses deliberately on foundation model pretraining through pure self-supervision, without leveraging (weak) supervision through text or other signals. Second, this work does not provided deeper insights into the learned representations apart from linear probing. Third, the initial hyperparameter tuning was done based on supervised performance (across several datasets), which does not necessarily have to align with pretraining performance.

Conclusion We present the first comprehensive, like-for-like study on pretraining strategies and scaling behavior for ECG FMs, covering five self-supervised objectives trained on a unified pretraining dataset of up to 11M samples and evaluated across a diverse set of clinical downstream tasks. Our results establish three key findings. First, backbone architecture is a dominant factor in ECG representation learning, with the S4-based SSM consistently outperforming Transformer and CNN alternatives across all pretraining objectives, a finding mechanistically supported by CKA analysis revealing superior layer-wise representational differentiation in S4 models. Second, pretraining strategy matters: CPC yields the most transferable representations across diverse clinical tasks, with performance gaps that widen under frozen and linear evaluation, while data2vec consistently underperforms regardless of scale. Third, we show firm evidence for scaling behavior in the domain of ECG data, most clearly seen for the two best-performing approaches CPC and JEPA, and provide evidence for correlation of pretraining and downstream performance. Taken together, these findings provide actionable guidance for future ECG FM development, advocating for SSM-based backbones (S4), carefully configured pretraining objectives, and domain-adaptive finetuning as the most impactful levers for improving downstream performance.

AI USE STATEMENT

In this work, we used generative AI tools for language refinement, editing of the manuscript and to draft the label mappings between the external datasets and PTB-XL used for the cross-dataset transfer evaluation (App. A.8.5). We have not used generative AI tools to help develop theoretical models or conceptual frameworks, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, clean and reformat dataset, support qualitative and thematic data analysis, interpret results and generation of synthetic data sets, formulation of mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, assisting with translation are not applicable to this work. Additionally, we used generative AI tools for improving readability of the paper. We have reviewed all AI-assisted work. All mappings were reviewed and verified manually by the authors against the clinical label descriptions and, where available, SNOMED-CT codes. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. All experiments use retrospective electrocardiogram (ECG) recordings from human subjects. No prospective data collection, interventions, or direct contact with participants were performed.

This works builds entirely on publicly available datasets, see App. A.1. For each dataset, use and access were explicitly cleared by the responsible institutional review board (IRB) or ethics committee prior to analysis. Where required, data-use agreements and credentialed-access procedures were followed. All ECG signals and associated metadata were handled in de-identified form in accordance with the approved protocols; no personally identifiable information (PII/PHI) was accessed, stored, or released by this project, and no re-identification attempts were made.

The research is conducted solely for scientific and methodological purposes. The methods and models presented are not intended for clinical decision-making and are not marketed or deployed as medical devices. Any potential clinical application would require separate validation, regulatory review, and oversight.

We report no conflicts of interest and no sensitive sponsorships related to this work. All authors have read and agree to comply with the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

All datasets used for pretraining and evaluation are publicly available, and their sources, licenses and task definitions are listed in App. A.1. The exact HEEDB subsets used for the scaling experiments are specified in App. A.2, together with the full model architectures, the configurations of all five pretraining objectives, the training and evaluation protocols, and the computational resources required. Design details and pseudo-code for HuBERT++ are provided in App. A.3, and the label mappings for the cross-dataset transfer evaluation are listed in App. A.8.5. Code and model weights are available at `https://anonymous.4open.science/r/ecg-pretraining-strategies-4DE3`.

REFERENCES

M A Al-Masud, Juan Miguel Lopez Alcaraz, and Nils Strodthoff. Benchmarking ECG FMs: A reality check across clinical tasks. In *International Conference on Learning Representations*, volume 2026, pp. 146689–146715, 2026.

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

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International conference on machine learning*, pp. 1298–1312. PMLR, 2022.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. V-JEPA: Latent video prediction for visual representation learning, 2024. URL https://openreview.net/forum?id=WFYbBOEOtv.

R Sacha Bhatia and Paul Dorian. Screening for cardiovascular disease risk with electrocardiography. *JAMA Internal Medicine*, 178(9):1163–1164, 2018.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258*, 2021.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervè Jègou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a General-Purpose Foundation Model for Computational Pathology. *Nature medicine*, 30(3):850–862, 2024.

Yanlong Chen, Mattia Orlandi, Pierangelo Maria Rapa, Simone Benatti, Luca Benini, and Yawei Li. PhysioWave: A Multi-Scale Wavelet-Transformer for Physiological Signal Representation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Edoardo Coppola, Mattia Savardi, Mauro Massussi, Marianna Adamo, Marco Metra, and Alberto Signoroni. HuBERT-ECG as a self-supervised foundation model for broad and scalable cardiac applications. *medRxiv*, pp. 2024–11, 2024.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality. In *International Conference on Machine Learning (ICML)*, 2024.

Timothée Darcet, Federico Baldassarre, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Cluster and Predict Latents Patches for Improved Masked Image Modeling. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=Ycmz7qJxUQ.

Duc N. Do, Minh N. Do, Dang Nguyen, Khanh T. Q. Le, Khoa D. Pham, Hung N. Huynh, Phi Pham-Van-Hoang, Quan K. Huynh, Ramez M. Odat, Perisa Ashar, Ethan Philip Lowder, Minh H. N. Le, Hoang Le, Phat V. H. Nguyen, Quan Le, Jacques Kpodonu, and Phat K. Huynh. Domain-Adapted Fine-Tuning of ECG Foundation Models for Multi-Label Structural Heart Disease Screening. *arXiv preprint arXiv:2604.23385*, 2026. URL https://arxiv.org/abs/2604.23385.

Pierre Elias and Joshua Finer. EchoNext: A Dataset for Detecting Echocardiogram-Confirmed Structural Heart Disease from ECGs. PhysioNet, 2025. URL https://doi.org/10.13026/r9pp-3y42.

Brian Gow, Tom Pollard, Larry A Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Jonathan W Waks, Parastou Eslami, Tanner Carbonati, Ashish Chaudhari, Elizabeth Herbst, Dana Moukheiber, Seth Berkowitz, Roger Mark, and Steven Horng. MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset. *PhysioNet*, September 2023. Version 1.0.

Albert Gu, Karan Goel, and Christopher Re. Efficiently Modeling Long Sequences with Structured State Spaces. In *International Conference on Learning Representations*, 2022.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pp. 30016–30030, 2022.

Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 328–339, 2018.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

John Weston Hughes, Linyuan Jing, Joshua Finer, Dustin Hartzel, Christopher Kelsey, Aaron Long, Daniel Rocha, Jeffrey Ruhl, Timothy Poterucha, and Pierre Elias. EchoNext-Mini: A Dataset and Baseline AI Model for Detecting Structural Heart Disease from Electrocardiograms. *NEJM AI*, 3 (5), April 2026. ISSN 2836-9386. doi: 10.1056/aidbp2500516. URL http://dx.doi.org/10.1056/aidbp2500516.

Manh Pham Hung, Aaqib Saeed, and Dong Ma. Boosting Masked ECG-Text Auto-Encoders as Discriminative Learners. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Tan Jian, Haoyi Fan, Jiawei Luo, Yanjie Zhou, Ning Wang, Xizheng Wang, Guizhi Liu, Chengyu Liu, and Zongmin Wang. A pediatric ECG database with disease diagnosis covering 11643 children, 5 2025. URL https://doi.org/10.6084/m9.figshare.27078763.v1.

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

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Zuzana Koscova, Qiao Li, Chad Robichaux, Valdery Moura Junior, Manohar Ghanta, Aditya Gupta, Jonathan Rosand, Aaron D Aguirre, Erik Reinertsen, Steven Song, et al. The harvard-emory ecg database. *Scientific Data*, 2026a.

Zuzana Koscova, Valdery Moura Junior, Matthew Reyna, Shenda Hong, Aditya Gupta, Manohar Ghanta, Reza Sameni, Aaron Aguirre, Qiao Li, Sahar Zafar, Gari Clifford, and M Brandon Westover. Harvard-Emory ECG Database (version 5.0). Brain Data Science Platform, 2026b.

Jiawei Li, Fabio Bonassi, Ming Jin, Stefan Gustafsson, Johan Sundström, Thomas B. Schön, and Antônio H. Ribeiro. How Do Electrocardiogram Models Scale? *arXiv preprint 2605.17276*, 2026. URL https://arxiv.org/abs/2605.17276.

Jun Li, Aaron D Aguirre, Valdery Moura Junior, Jiarui Jin, Che Liu, Lanhai Zhong, Chenxi Sun, Gari Clifford, M Brandon Westover, and Shenda Hong. An Electrocardiogram Foundation Model Built on over 10 Million Recordings. *NEJM AI*, 2(7):AIoa2401033, 2025.

Alexander H Liu, Heng-Jui Chang, Michael Auli, Wei-Ning Hsu, and Jim Glass. Dinosr: Self-distillation and online clustering for self-supervised speech representation learning. *Advances in Neural Information Processing Systems*, 36:58346–58362, 2023.

Che Liu, Zhongwei Wan, Cheng Ouyang, Anand Shah, Wenjia Bai, and Rossella Arcucci. Zero-shot ECG classification with multimodal learning and test-time clinical knowledge enhancement. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 31949–31963, 2024.

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

Hui Liu, Yinglong Wang, Da Chen, Xiyu Zhang, Huijie Li, Lipan Bian, Minglei Shu, and Dan Chen. A large-scale multi-label 12-lead electrocardiogram database with standardized diagnostic statements, 2022b.

Riccardo Lunelli, Angus Nicolson, Samuel Martin Pröll, Sebastian Johannes Reinstadler, Axel Bauer, and Clemens Dlaska. BenchECG and xECG: a benchmark and baseline for ECG foundation models. *npj Digital Medicine*, 2026.

Matthew B McDermott, Haoran Zhang, Lasse H Hansen, Giovanni Angelotti, and Jack Gallifant. A closer look at AUROC and AUPRC under class imbalance. *Advances in Neural Information Processing Systems*, 37:44102–44163, 2024.

Kaden McKeen, Sameer Masood, Augustin Toma, Barry Rubin, and Bo Wang. Ecg-fm: An open electrocardiogram foundation model. *Jamia Open*, 8(5):ooaf122, 2025.

Temesgen Mehari and Nils Strodthoff. Towards quantitative precision for ECG analysis: Leveraging state space models, self-supervision and patient metadata. *IEEE journal of biomedical and health informatics*, 27(11):5326–5334, 2023.

Yeongyeon Na, Minje Park, Yunwon Tae, and Sunghoon Joo. Guiding Masked Representation Learning to Capture Spatio-Temporal Relationship of Electrocardiogram. In *The Twelfth International Conference on Learning Representations*, 2024.

Marianna Nezhurina, Tomer Porian, Giovanni Puccetti, Tommie Kerssies, Romain Beaumont, Mehdi Cherti, and Jenia Jitsev. Scaling Laws for Robust Comparison of Open Foundation Language-Vision Models and Datasets. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Alexis Nolin-Lapalme, Achille Sowa, Jacques Delfrate, Olivier Tastet, Denis Corbin, Merve Kulbay, Derman Ozdemir, Marie-Jeanne Noël, François-Christophe Marois-Blanchet, François Harvey, et al. Foundation models for electrocardiogram interpretation: clinical implications. *European Heart Journal*, pp. ehaf1119, 2026.

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

M. A. Reyna, N. Sadr, E. A. Perez Alday, A. Gu, A. J. Shah, C. Robichaux, A. B. Rad, A. Elola, S. Seyedi, S. Ansari, H. Ghanbari, Q. Li, A. Sharma, and G. D. Clifford. Will Two Do? Varying Dimensions in Electrocardiography: The PhysioNet/Computing in Cardiology Challenge 2021. In *2021 Computing in Cardiology (CinC)*, pp. 1–4, Brno, Czech Republic, 2021. doi: 10.23919/CinC53138.2021.9662687.

M. A. Reyna, N. Sadr, A. Gu, E. A. Perez Alday, C. Liu, S. Seyedi, A. Shah, and G. D. Clifford. Will Two Do? Varying Dimensions in Electrocardiography: The PhysioNet/Computing in Cardiology Challenge 2021 (version 1.0.3). https://doi.org/10.13026/34va-7q14, 2022. PhysioNet. RRID:SCR_007345.

Antônio H. Ribeiro, Gabriela M.M. Paixao, Emilly M. Lima, Manoel Horta Ribeiro, Marcelo M. Pinto Filho, Paulo R. Gomes, Derick M. Oliveira, Wagner Meira Jr, Thömas B Schon, and Antonio Luiz P. Ribeiro. CODE-15%: a large scale annotated dataset of 12-lead ECGs , June 2021.

Ivan C Rokos, William J French, Amal Mattu, Graham Nichol, Michael E Farkouh, James Reiffel, and Gregg W Stone. Appropriate cardiac cath lab activation: optimizing electrocardiogram interpretation and clinical decision-making for acute ST-elevation myocardial infarction. *American heart journal*, 160(6):995–1003, 2010.

Konstantinos C Siontis, Peter A Noseworthy, Zachi I Attia, and Paul A Friedman. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nature Reviews Cardiology*, 18(7):465–478, 2021.

Nils Strodthoff, Patrick Wagner, Tobias Schaeffter, and Wojciech Samek. Deep learning for ECG analysis: Benchmarks and insights from PTB-XL. *IEEE journal of biomedical and health informatics*, 25(5):1519–1528, 2020.

Nils Strodthoff, Juan Miguel Lopez Alcaraz, and Wilhelm Haverkamp. Prospects for artificial intelligence-enhanced electrocardiogram as a unified screening tool for cardiac and non-cardiac conditions: an explorative study in emergency care. *European Heart Journal - Digital Health*, 5(4):454–460, 07 2024. ISSN 2634-3916. doi: 10.1093/ehjdh/ztae039. URL https://doi.org/10.1093/ehjdh/ztae039.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Amirhossein Taleshinosrati, Yangyang Wang, Atitaya Phoemsuk, Vahid Abolghasemi, Naser Hossein Motlagh, Sadasivan Puthusserypady, Daniel Teichmann, and Abdolrahman Peimankar. FOUND-AF: Benchmarking ECG Foundation Models for Atrial Fibrillation Detection. *arXiv preprint arXiv:2608.03597*, 2026.

Jian Tan, Haoyi Fan, Jiawei Luo, Yanjie Zhou, Ning Wang, Xizheng Wang, Guizhi Liu, Chengyu Liu, and Zongmin Wang. A pediatric ECG database with disease diagnosis covering 11643 children. *Scientific Data*, 12(1):867, 2025. doi: 10.1038/s41597-025-05225-z.

Yuanyuan Tian, Zhiyuan Li, Yanrui Jin, Mengxiao Wang, Xiaoyang Wei, Liqun Zhao, Yunqing Liu, Jinlei Liu, and Chengliang Liu. Foundation model of ECG diagnosis: Diagnostics and explanations of any form and rhythm on ECG. *Cell Reports Medicine*, 5(12), 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I Lunze, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset. *Scientific data*, 7(1):1–15, 2020. doi: 10.1038/s41597-020-0495-6.

Philipp Wagner, Nils Strodthoff, Ralf Bousseljot, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset (version 1.0.3). https://doi.org/10.13026/kfzx-aw45, 2022. PhysioNet. RRID:SCR_007345.

Zhijiang Wan, Qianhao Yu, Jia Mao, Wenfeng Duan, and Cheng Ding. OpenECG: Benchmarking ECG Foundation Models with Public 1.2 Million Records. *arXiv preprint arXiv:2503.00711*, 2025. URL https://arxiv.org/abs/2503.00711.

Han Yu, Peikun Guo, and Akane Sano. ECG semantic integrator (ESI): A foundation ECG model pretrained with LLM-enhanced cardiological text. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Jianwei Zheng, Jianming Zhang, Sidy Danioko, Hai Yao, Hangyuan Guo, and Cyril Rakovski. A 12-lead electrocardiogram database for arrhythmia research covering more than 10, 000 patients. *Scientific Data*, 7(1), February 2020. ISSN 2052-4463. doi: 10.1038/s41597-020-0386-x. URL http://dx.doi.org/10.1038/s41597-020-0386-x.

Yukun Zhou, Mark A Chia, Siegfried K Wagner, Murat S Ayhan, Dominic J Williamson, Robbert R Struyven, Timing Liu, Moucheng Xu, Mateo G Lozano, Peter Woodward-Court, et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163, 2023.

A Appendix

A.1 Datasets

This section describes all datasets used in this study. We first summarize the data sources and their licenses (App. A.1.1), and then detail the downstream tasks, their grouping into task categories, and the effective sample sizes available for each task (App. A.1.2).

A.1.1 Data Sources and Licenses

All datasets used in this study are publicly available, which makes the full pretraining and evaluation pipeline reproducible. HEEDB and HEEDB-Emory, both collected at US hospitals, are by a wide margin the largest publicly available ECG datasets. We complement them with CODE-15% from Brazil to broaden the geographic and population diversity of the pretraining dataset. The downstream datasets additionally cover China and Germany, so a large part of the evaluation tests transfer to populations and acquisition settings not seen during pretraining.

Table 1: Datasets used for pretraining and downstream evaluation, with their region of origin and license.

Dataset	Region	License
Pretraining		
HEEDB (Koscova et al., 2026b;a)	USA	BDSP Credentialed Health Data License 1.5.0
HEEDB-Emory (Koscova et al., 2026b;a)	USA	BDSP Credentialed Health Data License 1.5.0
CODE-15% (Ribeiro et al., 2021)	Brazil	Creative Commons Attribution 4.0 International
Downstream evaluation		
Ningbo (Reyna et al., 2022)	China	Open Data Commons Open Database License v1.0
CPSC2018 (Reyna et al., 2021; 2022)	China	Open Data Commons Open Database License v1.0
CPSC-Extra (Reyna et al., 2021; 2022)	China	Open Data Commons Open Database License v1.0
Georgia (Reyna et al., 2021; 2022)	USA	Open Data Commons Open Database License v1.0
Chapman (Zheng et al., 2020)	China	Creative Commons Attribution 4.0 International
SPH (Liu et al., 2022a;b)	China	Creative Commons Zero 1.0 Universal
PTB-XL (Wagner et al., 2022; 2020)	Germany	Creative Commons Attribution 4.0 International
ZZU pECG (Tan et al., 2025; Jian et al., 2025)	China	Creative Commons Attribution 4.0 International
EchoNext (Elias & Finer, 2025; Hughes et al., 2026)	USA	PhysioNet Restricted Health Data License 1.5.0
MIMIC-IV-ECG (Gow et al., 2023)	USA	Open Data Commons Open Database License v1.0

A.1.2 Task Categories and Dataset Characteristics

Table 2 summarizes the 26 downstream tasks, grouped into seven task categories that move progressively further away from classical ECG interpretation, from adult and pediatric ECG statements to echocardiographic findings, clinical outcomes and patient characteristics.

Table 2: This table, adapted from (Al-Masud et al., 2026), provides an overview of the datasets and prediction tasks included in this study, covering routine cardiac diagnostics, clinical outcomes, and patient metadata, along with their sample sizes, patient counts, and label structures. For all tasks, we report effective sample sizes rather than the total number of available ECGs. In classification tasks, the effective sample size corresponds to the number of ECGs and patients with at least one positive label, since samples containing only negative labels do not contribute information about the presence of a condition. In regression tasks, it corresponds to the number of ECGs and patients with at least one valid numeric target value, as missing targets cannot be used for training or evaluation. Consequently, effective sample sizes may be substantially smaller than the total dataset size. We adopt this definition to provide a realistic estimate of usable data per task, ensure comparability across tasks, and avoid overstating the available training signal. †: evaluation only

Task	Dataset	Type	Samples	Patients	Outputs
Pretraining					
Pretraining	HEEDB	-	10,445,276	1,802,503	-
Pretraining	HEEDB-Emory	-	998,843	349,556	-
Pretraining	CODE-15%	-	345,109	233,480	-
Downstream evaluation: adult ECG interpretation					
ECG interpretation	Ningbo	Multi-label	34,808	unknown	68
ECG interpretation	CPSC2018	Multi-label	6,867	unknown	9
ECG interpretation	CPSC-Extra	Multi-label	3,441	unknown	33
ECG interpretation	Georgia	Multi-label	10,286	unknown	50
ECG interpretation	Chapman	Multi-label	10,646	10,646	42
ECG interpretation	Chapman (rhythm)†	Multi-label	10,646	10,646	9
ECG interpretation	SPH	Multi-label	25,770	24,666	35
ECG interpretation	PTB-XL(all/sub/super)	Multi-label	21,799	18,869	71/23/5
ECG interpretation†	PTB-XL(diag/form/rhythm)	Multi-label	21,799	18,869	44/19/12
Downstream evaluation: pediatric ECG interpretation					
ECG interpretation	ZZU pECG	Multi-label	12,328	10,350	58
Downstream evaluation: cardiac structure & function					
Echocardiogram findings	EchoNext	Multi-label	82,543	36,286	11
Downstream evaluation: cardiac outcomes					
Cardiac discharge diagnoses	MIMIC-IV-ECG	Multi-label	114,355	49,400	158
Downstream evaluation: non-cardiac outcomes					
Non-cardiac discharge diagnoses	MIMIC-IV-ECG	Multi-label	178,163	81,930	918
Downstream evaluation: acute care predictions					
Clinical deterioration	MIMIC-IV-ECG	Multi-label	5,577	4,595	6
Mortality	MIMIC-IV-ECG	Multi-label	17,639	10,220	7
ICU admission	MIMIC-IV-ECG	Multi-label	18,690	13,868	2
Downstream evaluation: patient characteristics					
Sex	MIMIC-IV-ECG	Binary	182,076	83,736	1
Age	MIMIC-IV-ECG	Regression	182,076	83,736	1
Biometrics	MIMIC-IV-ECG	Regression	119,214	53,702	3
ECG features	MIMIC-IV-ECG	Regression	181,989	83,721	7
Lab values	MIMIC-IV-ECG	Regression	88,448	53,293	18
Vital signs	MIMIC-IV-ECG	Regression	131,602	66,605	6

A.2 Experimental Setup

Input preprocessing All models operate on standard 12-lead ECG signals sampled at 240 Hz. Raw waveforms are segmented into non-overlapping windows of 2.5 s (600 timesteps) and passed directly to the encoder. No data augmentation is applied during pretraining beyond the masking or context selection strategies intrinsic to each objective. Pretraining data are partitioned into 10 folds, with 9 folds used for training and the remaining fold held out for validation; no held-out test set is used during pretraining.

Shared encoder architecture All five pretraining objectives share an identical encoder that is kept fixed across every comparison, ensuring that any observed performance differences are attributable solely to the pretraining loss and its associated components. The encoder consists of a CNN stem followed by an S4-based sequential backbone. The CNN stem comprises four convolutional layers with output feature dimensions $[512, 512, 512, 512]$, kernel sizes $[3, 1, 1, 1]$, strides $[2, 1, 1, 1]$, and dilations $[1, 1, 1, 1]$; batch normalization is applied after each convolutional layer (no layer normalization in the stem). The S4 backbone uses 4 layers with model dimension 512, state dimension 8, dropout rate 0.2, and no prenormalization or batch normalization inside the S4 blocks. The backbone operates in non-causal mode for all objectives except CPC, which requires a causal backbone consistent with its autoregressive prediction objective.

Shared training configuration All models are optimized with AdamW, learning rate 3e-3, weight decay 1e-3, and batch size 64. Training is performed in full (fp32) precision on a single NVIDIA RTX PRO 6000 Blackwell Server Edition GPU. All main results are based on a 4-layer S4 backbone pretrained for 10 epochs, while the hyperparameter ablations were conducted with a 6-layer S4 backbone pretrained for 20 epochs unless stated otherwise in the respective table captions.

The pretraining dataset at each scale corresponds to specific HEEDB subsets: the 18K-sample run uses subset S0001-1987, the 45K-sample run uses subset S0001-1990, the 106K-sample run uses subset S0001-2019, and the 753K-sample run uses subset S0001-2007. The full 11M-sample dataset combines all available HEEDB cohorts, HEEDB-Emory, and CODE-15%.

data2vec configurations The EMA teacher averages the contextualized representations produced by the top 2 S4 backbone layers (layers 3 and 4). Layer normalization is applied to the averaged teacher targets before the regression loss is computed. The SSL prediction head is a single-layer non-causal S4 module with the same model dimension as the backbone (d=512).

DinoSR configurations The EMA teacher generates discrete pseudo-labels through online k-means quantization with two codebooks of size 256 each, updated with an EMA momentum of 0.9. Teacher features are layer-normalized before quantization, and the student is trained with a cross-entropy loss at softmax temperature 1.0. The SSL prediction head is a single-layer non-causal S4 module. Masking is applied in the latent space using span masking with a mask probability of 0.065 and a span length of 10 tokens.

JEPA configurations A multi-block masking strategy is employed following Assran et al. (2023). The context encoder receives an unmasked region uniformly sampled from $[85\%, 100\%]$ of the input (1 context block), and the predictor targets 4 non-overlapping prediction blocks, each spanning $[15\%, 20\%]$ of the input, with a minimum of 64 tokens retained and overlap between prediction blocks disallowed. The EMA teacher produces latent representations from the top backbone layer, to which layer normalization is applied twice: once at the individual layer level and once to the final aggregated targets. The prediction loss is the smooth-ℓ_1 loss with β=1.0. The SSL prediction head is a single-layer non-causal S4 module.

CPC configurations No masking or EMA teacher is used. The model employs a causal S4 backbone and learns to predict 14 future latent steps from the causal context representation via a contrastive objective, drawing all negative samples exclusively from within the same input sequence. The SSL head is a linear projection without bias term.

HuBERT++ configurations The EMA teacher generates soft cluster assignments by applying Sinkhorn–Knopp optimal transport online to two codebooks of sizes 128 and 256. The student

is trained with a KL divergence loss between its predicted log-probabilities and the teacher's soft assignments. The loss is computed on both masked and unmasked tokens, weighted by α=0.75 and $1-\alpha$, respectively. Masking follows the same span masking as DinoSR, with a mask probability of 0.065 and a span length of 10 tokens. The SSL prediction head is a multi-layer perceptron (MLP). Full design details and pseudo-code are provided in App. A.3.

Downstream evaluation Downstream evaluation follows the protocol of (Al-Masud et al., 2026) across 26 clinically relevant tasks from 10 public datasets covering 7 task categories. All downstream runs use AdamW, learning rate 1e-3, weight decay 1e-3, batch size 64, and 100 epochs. Three evaluation modes are considered. *Finetuning (linear head):* the pretrained encoder is optimized end-to-end together with a linear prediction head using layer-dependent learning rates: the head is updated at the base rate (1e-3), the predictor module at 1e-4, and the CNN stem at 1e-5 (discriminative factor 0.1 applied per group). Sequence-level predictions are obtained by mean-pooling the encoder output before the linear head. *Frozen:* encoder weights are kept fixed and a learnable query-attention pooling head (Bardes et al., 2024) (16 attention heads, no bias) is trained at the base learning rate, operating directly on the full token sequence. *Linear:* encoder weights are kept fixed and a single linear layer is trained on mean-pooled representations at the base learning rate. Binary cross-entropy is used for classification tasks and MAE for regression tasks with z-normalized targets. Macro-averaged AUROC serves as the primary metric for classification; standardized MAE is used for regression. Statistical significance is assessed via pairwise empirical bootstrapping with 1,000 iterations on the held-out test set, with rankings determined by statistical equivalence groups (ties indicate no statistically significant difference at the 95% confidence level).

Computational resources All pretraining and downstream evaluation experiments were conducted on NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs for pretraining and NVIDIA L40 GPUs for downstream evaluation, with each pretraining run using a single GPU. To accelerate data loading, all pretraining datasets were copied to `$TMPDIR` (a per-job local NVMe-backed file system on the HPC cluster) prior to each run; without this step, I/O bottlenecks would significantly delay training. Approximate wall-clock training times per objective at each pretraining scale are summarized in Table 3.

Table 3: Approximate wall-clock pretraining time per run (single NVIDIA RTX PRO 6000 Blackwell Server Edition). 'm' stands for minute, 'h' for hour.

Objective	Small (106K, 10 ep.)	Medium (753K, 10 ep.)	Full (11M, 10 ep.)
data2vec	33.9 m	3.9 h	31.2 h
DinoSR	37.7 m	4.3 h	38.6 h
JEPA	35.2 m	4.0 h	34.8 h
CPC	42.4 m	4.8 h	45.1 h
HuBERT++	27.6 m	3.0 h	26.3 h

A.3 HuBERT++

A.3.1 Design decisions

We briefly describe the design decisions underlying HuBERT++. We draw on strong non-contrastive SSL methods such as I-JEPA (Assran et al., 2023), HuBERT (Hsu et al., 2021), DINO (Caron et al., 2021), DinoSR (Liu et al., 2023) and CAPI (Darcet et al., 2025) from the vision and speech domain. Its design is most closely related to DinoSR and CAPI. We follow the categorization of design choices put forward by Darcet et al. (2025):

1. We use exponential moving averages of a student network as prediction target (CAPI Figure 3) like I-JEPA, HuBERT, CAPI
2. We use clustering as loss (CAPI Figure 4) like CAPI. To avoid backpropagation through the clustering, we therefore need another non-SGD cluster update.
3. We use Sinkhorn-Knopp like DINO and unlike DinoSR, which encourages equiparticipation and prevents collapse.
4. We use granular, soft prediction targets (unlike DinoSR), which would also allow us to use different temperatures (as in DINO).
5. We use cluster assignments using Sinkhorn-Knopp optimal transport (unlike CAPI, which uses a quite ad-hoc procedure). This has the nice side effect that prediction target computation and cluster updates happen consistently.

A.3.2 Pseudo-code

We provide pseudo-code for the three most crucial components of the algorithm.

Computing EMA targets

```
@torch.no_grad()
def get_ema_targets(self, x):
    """
    Get soft targets from EMA path using Sinkhorn-Knopp.

    Args:
        x: input images [B, C, T]

    Returns:
        targets: [B, K] soft assignment probabilities
    """
    # Extract and normalize EMA features
    z_ema = self.ema_encoder(x)
    z_ema = self.ema_projector(z_ema)
    z_ema = F.normalize(z_ema, dim=1)

    # Compute similarity to prototypes
    logits = z_ema @ self.ema_prototypes.T / self.temperature  # [B,
    ↪  K]

    # Apply Sinkhorn-Knopp for balanced soft assignments
    targets = sinkhorn_knopp(
        logits,
        num_iters=self.sinkhorn_iters,
        epsilon=self.sinkhorn_epsilon
    )

    return targets, z_ema
```

Update cluster prototypes

```
@torch.no_grad()
def update_prototypes(self, ema_features, assignments):
```

```
        """
        Update prototypes using soft assignments from Sinkhorn-Knopp.

        Args:
            ema_features: [B, D] normalized features from EMA encoder
            assignments: [B, K] soft assignment matrix from
            ↪  Sinkhorn-Knopp
        """
        if self.step_count < self.freeze_prototypes_steps:
            return

        # Compute new prototype positions as weighted average of features
        new_prototypes = assignments.T @ ema_features  # [K, D]
        new_prototypes = F.normalize(new_prototypes, dim=1)

        # EMA update of prototypes
        self.ema_prototypes.data = (
            self.prototype_momentum * self.ema_prototypes.data +
            (1 - self.prototype_momentum) * new_prototypes
        )
        self.ema_prototypes.data = F.normalize(self.ema_prototypes.data,
        ↪  dim=1)
```

Forward method

```
    def forward(self, x):
        """
        Forward pass with single augmented view.

        Args:
            x: augmented view [B, C, T]

        Returns:
            loss: scalar loss value
        """
        # Get targets from EMA path (no gradients)
        with torch.no_grad():
            targets, z_ema = self.get_ema_targets(x)

        # Get predictions from online path (with gradients)
        z_online = self.online_encoder(x)
        z_online = self.mask(z_online) # optional masking in latent space
        z_online = self.online_projector(z_online)
        z_online = self.online_predictor(z_online)
        z_online = F.normalize(z_online, dim=1)

        # Compute logits for online features against prototypes
        logits_online = z_online @ self.ema_prototypes.T /
        ↪  self.temperature

        # Cross-entropy loss: predict EMA targets with online features
        loss = -torch.sum(targets * F.log_softmax(logits_online, dim=1),
        ↪  dim=1).mean()

        # Update prototypes (no gradient flow)
        with torch.no_grad():
            self.update_prototypes(z_ema, targets)

        self.step_count += 1

        return loss
```

A.4 Main Results on the Full Pretraining Dataset

This section reports the complete downstream results for all models pretrained on the full 11M-sample dataset, alongside the external FMs ECGFounder and ECG-JEPA and the supervised S4 baseline. Results are shown separately for finetuning, frozen and linear evaluation, and form the basis of the rankings in Fig. 3 and App. A.4.4.

A.4.1 Finetuning Evaluation

Table 4: Comparison of aggregated macro-AUROC for classification and MAE for regression under finetuning with a linear prediction head. We highlight with ↑ tasks where higher AUROC is better and ↓ tasks where lower standardized MAE values are better. The best-performing result is highlighted in boldface and underlined, while models that do not perform statistically significantly worse are also highlighted in boldface. † signifies evaluation of a model trained on the parent dataset (listed above).

	FMs (Finetuned)		**Pretrained by Us (Finetuned)**					**Supervised**
	ECGFounder	**ECG-JEPA**	**data2vec**	**DinoSR**	**JEPA**	**CPC**	**HuBERT++**	**S4**
Adult ECG interpretation								
Ningbo ↑	**0.974**	**0.973**	0.965	0.965	0.970	**0.974**	0.969	**0.972**
CPSC2018 ↑	0.966	**0.974**	0.953	0.953	**0.971**	**0.967**	0.962	0.962
CPSC-Extra ↑	**0.906**	**0.897**	0.873	0.873	**0.888**	**0.897**	**0.892**	0.852
Georgia ↑	**0.920**	**0.918**	0.894	0.894	**0.921**	**0.920**	0.911	0.903
Chapman ↑	**0.968**	**0.972**	0.945	0.945	**0.964**	**0.963**	0.962	**0.963**
-Chapman (rhythm)† ↑	**0.991**	**0.989**	0.977	0.977	**0.989**	**0.990**	0.987	0.986
SPH ↑	**0.983**	0.980	0.974	0.974	0.974	**0.982**	**0.984**	**0.981**
PTB-XL (all) ↑	0.934	**0.940**	0.929	0.929	0.938	**0.945**	0.939	**0.941**
-PTB-XL (diag) † ↑	**0.950**	**0.946**	0.941	0.941	0.935	**0.946**	**0.946**	**0.943**
-PTB-XL (form)† ↑	0.875	0.912	0.904	0.904	**0.925**	**0.926**	0.906	**0.919**
-PTB-XL (rhythm)† ↑	**0.965**	**0.956**	0.921	0.921	**0.961**	**0.963**	0.952	**0.956**
PTB-XL (sub) ↑	**0.943**	**0.935**	**0.936**	**0.936**	**0.939**	**0.941**	**0.942**	**0.938**
PTB-XL (super) ↑	**0.935**	0.921	0.929	0.929	**0.934**	**0.934**	**0.936**	0.932
Pediatric ECG interpretation								
ZZU pECG ↑	0.898	**0.911**	0.875	0.875	0.895	0.896	0.892	0.897
Cardiac structure & function								
EchoNext ↑	0.817	0.817	0.824	0.824	**0.830**	**0.832**	**0.831**	0.819
Cardiac outcomes								
MIMIC (Cardiac) ↑	0.768	0.772	0.778	**0.784**	0.780	**0.786**	0.782	0.780
Non-cardiac outcomes								
MIMIC (Non-cardiac) ↑	0.701	0.711	0.708	0.719	0.713	**0.722**	0.716	0.714
Acute care predictions								
MIMIC (Deterioration) ↑	0.717	**0.747**	**0.754**	**0.762**	**0.764**	**0.763**	**0.757**	**0.756**
MIMIC (Mortality) ↑	**0.810**	**0.792**	**0.799**	**0.794**	**0.801**	**0.808**	**0.800**	**0.793**
MIMIC (ICU) ↑	0.748	0.742	0.749	0.752	0.752	**0.758**	**0.753**	0.745
Patient characteristics								
MIMIC (Sex) ↑	0.913	0.904	0.922	0.931	0.932	**0.939**	0.932	0.919
MIMIC (Age) ↓	0.461	0.463	0.462	0.450	**0.435**	0.445	0.445	0.455
MIMIC (Biometrics) ↓	0.637	0.640	0.623	0.610	0.606	**0.596**	0.609	0.626
MIMIC (ECG Features) ↓	0.458	0.460	0.454	0.454	**0.448**	0.455	0.450	0.452
MIMIC (Lab Values) ↓	0.679	0.677	0.678	**0.670**	**0.670**	**0.669**	**0.670**	0.675
MIMIC (Vital Signs) ↓	0.704	0.703	0.702	0.701	**0.698**	**0.699**	**0.699**	0.701

A.4.2 FROZEN EVALUATION

Table 5: Comparison of aggregated macro-AUROC for classification and MAE for regression under frozen evaluation. We highlight with ↑ tasks where higher AUROC is better and ↓ tasks where lower standardized MAE values are better. The best-performing result is highlighted in boldface and underlined, while models that do not perform statistically significantly worse are also highlighted in boldface. † signifies evaluation of a model trained on the parent dataset (listed above).

	FMs (Frozen)		**Pretrained by Us (Frozen)**					**Supervised**
	ECGFounder	**ECG-JEPA**	**data2vec**	**DinoSR**	**JEPA**	**CPC**	**HuBERT++**	**S4**
Adult ECG interpretation								
Ningbo ↑	0.961	**0.971**	0.895	0.895	0.941	0.955	0.941	**0.972**
CPSC2018 ↑	0.966	**0.975**	0.915	0.915	0.939	0.961	0.937	0.962
CPSC-Extra ↑	**0.907**	**0.902**	0.823	0.823	0.856	0.872	0.863	0.852
Georgia ↑	**0.924**	**0.910**	0.846	0.846	0.888	0.904	0.878	0.903
Chapman ↑	**0.967**	**0.964**	0.914	0.914	0.941	0.941	0.940	**0.963**
-Chapman (rhythm)† ↑	**0.983**	**0.988**	0.941	0.941	0.966	**0.984**	0.959	**0.986**
SPH ↑	0.966	**0.980**	0.923	0.923	0.964	0.960	0.964	**0.981**
PTB-XL (all) ↑	0.927	0.934	0.876	0.876	0.913	0.928	0.918	**0.941**
-PTB-XL (diag) † ↑	**0.940**	**0.943**	0.888	0.888	0.922	0.932	0.933	**0.943**
-PTB-XL (form) † ↑	0.876	0.889	0.853	0.853	0.888	**0.903**	0.886	**0.919**
-PTB-XL (rhythm) † ↑	**0.958**	**0.966**	0.862	0.862	0.915	0.946	0.908	**0.956**
PTB-XL (sub) ↑	**0.939**	**0.934**	0.905	0.905	0.926	**0.932**	0.925	**0.938**
PTB-XL (super) ↑	0.928	0.917	0.913	0.913	0.923	0.907	0.918	**0.932**
Pediatric ECG interpretation								
ZZU pECG ↑	0.891	**0.905**	0.815	0.815	0.875	0.872	0.869	**0.897**
Cardiac structure & function								
EchoNext ↑	0.803	0.811	0.799	0.799	**0.815**	**0.820**	**0.819**	**0.819**
Cardiac outcomes								
MIMIC (Cardiac) ↑	0.745	0.757	0.752	0.758	0.764	0.774	0.766	**0.780**
Non-cardiac outcomes								
MIMIC (Non-cardiac) ↑	0.671	0.688	0.676	0.688	0.691	0.706	0.695	**0.714**
Acute care predictions								
MIMIC (Deterioration) ↑	0.697	0.702	0.705	**0.727**	**0.745**	**0.741**	**0.726**	**0.756**
MIMIC (Mortality) ↑	**0.769**	**0.788**	0.740	0.761	**0.773**	**0.783**	**0.785**	**0.793**
MIMIC (ICU) ↑	0.731	0.734	0.720	0.727	0.736	**0.750**	0.737	**0.745**
Patient characteristics								
MIMIC (Sex) ↑	0.872	0.894	0.893	0.885	0.905	**0.920**	0.914	**0.919**
MIMIC (Age) ↓	0.515	0.484	0.504	0.511	0.486	0.465	0.473	**0.455**
MIMIC (Biometrics) ↓	0.702	0.700	0.666	0.671	0.655	0.637	0.646	**0.626**
MIMIC (ECG Features) ↓	0.489	0.477	0.487	0.487	0.468	0.468	0.463	**0.452**
MIMIC (Lab Values) ↓	0.703	0.694	0.693	0.698	0.684	**0.676**	0.680	**0.675**
MIMIC (Vital Signs) ↓	0.719	0.716	0.712	0.714	0.709	**0.702**	0.704	**0.701**

A.4.3 LINEAR EVALUATION

Table 6: Comparison of aggregated macro-AUROC for classification and MAE for regression under linear evaluation. We highlight with ↑ tasks where higher AUROC is better and ↓ tasks where lower standardized MAE values are better. The best-performing result is highlighted in boldface and underlined, while models that do not perform statistically significantly worse are also highlighted in boldface. † signifies evaluation of a model trained on the parent dataset (listed above).

	FMs (Linear)		**Pretrained by Us (Linear)**					**Supervised**
	ECGFounder	**ECG-JEPA**	**data2vec**	**DinoSR**	**JEPA**	**CPC**	**HuBERT++**	**S4**
Adult ECG interpretation								
Ningbo ↑	**0.970**	**0.970**	0.814	0.848	0.911	0.899	0.793	**0.972**
CPSC2018 ↑	0.964	**0.975**	0.814	0.809	0.895	0.896	0.781	0.962
CPSC-Extra ↑	**0.910**	**0.902**	0.780	0.774	0.857	0.779	0.711	0.852
Georgia ↑	**0.923**	**0.920**	0.771	0.818	0.855	0.845	0.735	0.903
Chapman ↑	**0.968**	**0.962**	0.823	0.850	0.894	0.854	0.775	**0.963**
-Chapman (rhythm)†	**0.987**	**0.989**	0.830	0.892	0.954	0.940	0.814	**0.986**
SPH ↑	0.975	0.967	0.869	0.867	0.940	0.922	0.840	**0.981**
PTB-XL (all) ↑	0.931	0.928	0.799	0.805	0.878	0.866	0.753	**0.941**
-PTB-XL (diag)† ↑	**0.947**	0.925	0.807	0.800	0.886	0.882	0.760	**0.943**
-PTB-XL (form)† ↑	0.874	**0.908**	0.761	0.767	0.842	0.796	0.701	**0.919**
-PTB-XL (rhythm)† ↑	**0.961**	**0.969**	0.811	0.863	0.903	0.916	0.799	0.956
PTB-XL (sub) ↑	**0.945**	0.916	0.823	0.838	0.896	0.881	0.813	**0.938**
PTB-XL (super) ↑	0.924	0.911	0.826	0.822	0.880	0.860	0.809	**0.932**
Pediatric ECG interpretation								
ZZU pECG ↑	**0.900**	**0.891**	0.763	0.757	0.844	0.846	0.726	**0.897**
Cardiac structure & function								
EchoNext ↑	0.795	0.806	0.790	0.795	0.805	0.799	0.785	**0.819**
Cardiac outcomes								
MIMIC (Cardiac) ↑	0.751	0.751	0.704	0.722	0.746	0.752	0.701	**0.780**
Non-cardiac outcomes								
MIMIC (Non-cardiac) ↑	0.671	0.675	0.638	0.652	0.673	0.682	0.633	**0.714**
Acute care predictions								
MIMIC (Deterioration) ↑	0.713	0.720	0.670	0.681	**0.731**	**0.728**	0.684	**0.756**
MIMIC (Mortality) ↑	**0.774**	**0.782**	0.698	0.742	**0.762**	**0.761**	0.720	**0.793**
MIMIC (ICU) ↑	0.730	**0.736**	0.686	0.697	0.725	0.732	0.701	**0.745**
Patient characteristics								
MIMIC (Sex) ↑	0.872	0.883	0.817	0.826	0.870	0.882	0.809	**0.919**
MIMIC (Age) ↓	0.511	0.489	0.632	0.580	0.552	0.534	0.608	**0.455**
MIMIC (Biometrics) ↓	0.700	0.685	0.733	0.713	0.689	0.680	0.732	**0.626**
MIMIC (ECG Features) ↓	0.488	0.490	0.572	0.536	0.492	0.501	0.549	**0.452**
MIMIC (Lab Values) ↓	0.693	0.695	0.709	0.711	0.692	0.684	0.704	**0.675**
MIMIC (Vital Signs) ↓	0.716	0.714	0.724	0.722	0.710	0.707	0.719	**0.701**

A.4.4 STATISTICAL RANKINGS

To summarize performance across tasks, we rank all models per task and evaluation mode based on statistical equivalence groups. Models within the same equivalence group share the same rank, with the top group consisting of all models that do not perform significantly worse than the best model. Ranks follow standard competition ranking, so that if three models share rank 1, the next model is assigned rank 4. Tab. 7 reports these per-task rankings, and Tab. 8 aggregates them as the median rank within each task category, which underlies Fig. 3.

Table 7: Per-task statistical rankings of all models across evaluation modes, reported as Finetuned / Frozen / Linear. Lower ranks indicate better performance. † evaluated using the model trained on the parent label set (listed above).

	FMs		Pretrained by Us					Supervised
	ECGFounder	**ECG-JEPA**	**data2vec**	**DinoSR**	**JEPA**	**CPC**	**HuBERT++**	**S4**
Adult ECG interpretation								
Ningbo	1/3/1	1/1/1	7/7/7	7/7/6	5/5/4	1/3/4	5/5/7	1/1/1
CPSC2018	4/2/2	1/1/1	7/7/6	7/7/6	1/5/4	1/2/4	4/5/8	4/2/2
CPSC-Extra	1/1/1	1/1/1	6/7/5	6/7/5	1/3/3	1/3/5	1/3/8	6/3/3
Georgia	1/1/1	1/1/1	7/7/7	7/7/6	1/3/4	1/3/4	5/6/8	5/3/3
Chapman	1/1/1	1/1/1	7/7/7	7/7/5	1/4/4	1/4/5	6/4/8	1/1/1
-Chapman (rhythm)†	1/1/1	1/1/1	7/7/7	7/7/6	1/5/4	1/1/5	5/5/7	5/1/1
SPH	1/3/2	5/1/2	6/7/6	6/7/6	6/3/4	1/3/5	1/3/8	1/1/1
PTB-XL (all)	4/2/2	1/2/2	7/7/6	7/7/6	4/5/4	1/2/4	4/5/8	1/1/1
-PTB-XL (diag)†	1/1/1	1/1/3	6/7/6	6/7/6	6/4/4	1/4/4	1/4/8	1/1/1
-PTB-XL (form)†	8/3/3	4/3/1	4/7/5	4/7/5	1/3/3	1/1/5	4/3/8	1/1/1
-PTB-XL (rhythm)†	1/1/1	1/1/1	7/7/7	7/7/6	1/5/4	1/4/4	6/5/7	1/1/3
PTB-XL (sub)	1/1/1	1/1/3	1/7/6	1/7/6	1/5/4	1/1/5	1/5/8	1/1/1
PTB-XL (super)	1/2/2	8/4/3	5/6/6	5/6/6	1/3/4	1/8/5	1/4/8	5/1/1
Pediatric ECG interpretation								
ZZU pECG	2/3/1	1/1/1	7/7/6	7/7/6	2/4/4	2/4/4	6/4/8	2/1/1
Cardiac structure & function								
EchoNext	4/5/5	4/5/2	4/7/5	4/7/5	1/1/2	1/1/2	1/1/5	4/1/1
Cardiac outcomes								
MIMIC (Cardiac)	7/7/2	7/5/2	6/7/7	1/5/6	3/3/5	1/2/2	3/3/7	3/1/1
Non-cardiac outcomes								
MIMIC (Non-cardiac)	8/8/5	5/4/3	7/7/7	2/4/6	5/4/3	1/2/2	3/3/8	3/1/1
Acute care predictions								
MIMIC (Deterioration)	8/6/5	1/6/1	1/6/5	1/1/5	1/1/1	1/1/1	1/1/5	1/1/1
MIMIC (Mortality)	1/1/1	1/1/1	1/8/6	1/7/6	1/1/1	1/1/1	1/1/6	1/1/1
MIMIC (ICU)	3/3/3	7/3/1	3/7/8	3/7/6	3/3/3	1/1/3	1/3/6	7/1/1
Patient characteristics								
MIMIC (Sex)	7/8/4	8/5/2	5/5/7	2/7/6	2/4/4	1/1/2	2/3/8	6/1/1
MIMIC (Age)	6/7/3	6/4/2	6/6/8	4/7/6	1/4/5	2/2/4	2/3/7	5/1/1
MIMIC (Biometrics)	7/7/5	7/7/2	5/5/7	2/6/6	2/4/4	1/2/2	2/3/7	5/1/1
MIMIC (ECG Features)	7/8/2	8/5/3	4/6/8	4/6/6	1/3/4	4/3/5	2/2/7	3/1/1
MIMIC (Lab Values)	7/8/3	5/5/3	7/5/7	1/7/7	1/4/3	1/1/2	1/3/6	5/1/1
MIMIC (Vital Signs)	7/8/4	7/6/4	4/5/8	4/6/7	1/4/3	1/1/2	1/3/6	4/1/1

Table 8: Median statistical rankings of all models per task category, reported as Finetuned / Frozen / Linear. Lower ranks indicate better performance.

	FMs		Pretrained by Us					Supervised
	ECGFounder	**ECG-JEPA**	**data2vec**	**DinoSR**	**JEPA**	**CPC**	**HuBERT++**	**S4**
Adult ECG interpretation	1/1/1	1/1/1	7/7/6	7/7/6	1/4/4	1/3/5	4/5/8	1/1/1
Pediatric ECG interpretation	2/3/1	1/1/1	7/7/6	7/7/6	2/4/4	2/4/4	6/4/8	2/1/1
Cardiac structure & function	4/5/5	4/5/2	4/7/5	4/7/5	1/1/2	1/1/2	1/1/5	4/1/1
Cardiac outcomes	7/7/2	7/5/2	6/7/7	1/5/6	3/3/5	1/2/2	3/3/7	3/1/1
Non-cardiac outcomes	8/8/5	5/4/3	7/7/7	2/4/6	5/4/3	1/2/2	3/3/8	3/1/1
Acute care predictions	3/3/3	1/3/1	1/7/6	1/7/6	1/1/1	1/1/1	1/1/6	1/1/1
Patient characteristics	7/8/3.5	7/5/2.5	5/5/7.5	3/6.5/6	1/4/4	1/1.5/2	2/3/7	5/1/1

A.5 Comparison with Existing ECG FMs

This section compares models from this study with existing ECG FMs that share either the pretraining objective or the pretraining data, complementing the comparison with ECGFounder and ECG-JEPA in App. A.4.

A.5.1 Comparison between HuBERT++ and HuBERT-ECG

Tab. 9 compares HuBERT++ with HuBERT-ECG (BASE, 93M parameters) (Coppola et al., 2024), which uses a Transformer encoder trained to predict hard cluster assignments obtained by offline k-means clustering. HuBERT++ instead derives soft targets online from an EMA teacher via Sinkhorn-Knopp assignments and uses an S4 backbone with only 3M parameters. Although HuBERT-ECG was pretrained on 9.1M ECGs that include most of our downstream datasets, HuBERT++ outperforms it on all tasks.

Table 9: Finetuning performance (macro-AUROC) of HuBERT-ECG and HuBERT++. The best-performing result is highlighted in boldface and underlined, while models that do not perform statistically significantly worse are also highlighted in boldface.

	FMs (Finetuned)	
	HuBERT-ECG	**HuBERT++**
Adult ECG interpretation		
Ningbo	0.958	**<u>0.969</u>**
CPSC2018	**0.956**	**<u>0.962</u>**
CPSC-Extra	0.876	**<u>0.892</u>**
Georgia	0.883	**<u>0.911</u>**
Chapman	0.941	**<u>0.962</u>**
SPH	0.953	**<u>0.984</u>**
PTB-XL (all)	0.915	**<u>0.939</u>**
PTB-XL (sub)	0.918	**<u>0.942</u>**
PTB-XL (super)	0.908	**<u>0.936</u>**
Pediatric ECG interpretation		
ZZU pECG	**0.883**	**<u>0.892</u>**
Cardiac structure & function		
EchoNext	0.792	**<u>0.831</u>**

A.5.2 Comparison between CPC and external CPC

Tab. 10 provides a direct comparison to the ECG-CPC model from (Al-Masud et al., 2026). The model shows a relatively similar level of performance with slight advantages for the CPC model trained as part of this work, most likely due to a slightly larger training dataset and longer pretraining.

Table 10: Comparison of aggregated macro-AUROC between CPC and external CPC (Al-Masud et al., 2026) for classification and MAE for regression under finetuning with a linear prediction head. We highlight with ↑ tasks where higher AUROC is better and ↓ tasks where lower standardized MAE values are better. The best-performing result is highlighted in boldface and underlined, while models that do not perform statistically significantly worse are also highlighted in boldface. † signifies evaluation of a model trained on the parent dataset (listed above). Both CPC and CPC (external) use 4 layers. CPC is pretrained on HEEDB, HEEDB-Emory, and CODE-15% for 10 epochs with a learning rate of 3e-3, while CPC (external) is pretrained on HEEDB for 2 epochs with a learning rate of 1e-3.

	CPC	CPC (external)
Adult ECG interpretation		
Ningbo ↑	**0.974**	**0.973**
CPSC2018 ↑	**0.967**	**0.969**
CPSC-Extra ↑	**0.897**	**0.898**
Georgia ↑	**0.920**	**0.913**
Chapman ↑	**0.963**	**0.962**
-Chapman (rhythm)† ↑	**0.990**	0.987
SPH ↑	**0.982**	**0.981**
PTB-XL (all) ↑	0.945	**0.949**
-PTB-XL (diag)† ↑	0.946	**0.951**
-PTB-XL (form)† ↑	**0.926**	**0.934**
-PTB-XL (rhythm)† ↑	**0.963**	**0.959**
PTB-XL (sub) ↑	**0.941**	**0.940**
PTB-XL (super) ↑	**0.934**	**0.934**
Pediatric ECG interpretation		
ZZU pECG ↑	**0.896**	**0.892**
Cardiac structure & function		
EchoNext ↑	**0.832**	**0.831**
Cardiac outcomes		
MIMIC (Cardiac) ↑	**0.786**	0.781
Non-cardiac outcomes		
MIMIC (Non-cardiac) ↑	**0.722**	0.719
Acute care predictions		
MIMIC (Deterioration) ↑	**0.763**	**0.764**
MIMIC (Mortality) ↑	**0.808**	**0.803**
MIMIC (ICU) ↑	**0.758**	0.753
Patient characteristics		
MIMIC (Sex) ↑	**0.939**	0.933
MIMIC (Age) ↓	0.445	**0.437**
MIMIC (Biometrics) ↓	**0.596**	0.604
MIMIC (ECG Features) ↓	0.455	**0.451**
MIMIC (Lab Values) ↓	**0.669**	0.673
MIMIC (Vital Signs) ↓	**0.699**	**0.700**

A.5.3 COMPARISON WITH FMS PRETRAINED ON MIMIC-IV-ECG

To compare with MERL (Liu et al., 2024) and ECGFM-KED (Tian et al., 2024) under matched pretraining data, we additionally pretrain all five objectives on MIMIC-IV-ECG, the dataset used by both external FMs. Tab. 11 reports the finetuning results.

Table 11: Finetuning performance (macro-AUROC) of models pretrained on MIMIC-IV-ECG (759K samples). Our models use a 6-layer S4 backbone pretrained at 500 Hz for 20 epochs with learning rate 3e-3. MERL and ECGFM-KED were pretrained for 50 and 4 epochs with learning rates 2e-4 and 5e-5, respectively. The best-performing result is highlighted in boldface and underlined, while models that do not perform statistically significantly worse are also highlighted in boldface.

	FMs (Finetuned)		**Pretrained by Us (Finetuned)**					**Supervised**
	MERL	**KED**	**data2vec**	**DinoSR**	**JEPA**	**CPC**	**HuBERT++**	**S4**
Adult ECG interpretation								
Ningbo	0.955	0.940	0.957	**0.966**	0.962	**0.970**	**0.972**	**<u>0.972</u>**
CPSC2018	0.936	0.930	0.957	0.965	**0.966**	**<u>0.971</u>**	**0.971**	0.962
CPSC-Extra	0.873	0.824	0.835	0.864	0.866	**<u>0.908</u>**	**0.895**	0.852
Georgia	**0.912**	0.877	0.873	**0.913**	0.889	**0.911**	**<u>0.917</u>**	**0.903**
Chapman	0.946	0.917	0.932	0.950	0.937	0.956	**0.957**	**<u>0.963</u>**
SPH	0.944	0.932	0.967	0.969	0.968	**<u>0.982</u>**	**0.980**	**0.981**
PTB-XL (all)	0.925	0.889	0.915	0.934	0.924	**0.940**	**0.937**	**<u>0.941</u>**
PTB-XL (sub)	**0.937**	0.908	0.896	**0.934**	0.905	0.931	0.932	**<u>0.938</u>**
PTB-XL (super)	**0.930**	0.905	0.909	0.920	0.910	0.920	0.918	**<u>0.932</u>**
Pediatric ECG interpretation								
ZZU pECG	**0.886**	0.861	0.846	0.878	0.870	**0.886**	**0.887**	**<u>0.897</u>**
Cardiac structure & function								
EchoNext	**0.822**	0.806	0.817	**0.825**	0.817	**<u>0.828</u>**	**0.826**	0.819

A.6 Backbone Architecture Comparison

To isolate the effect of the backbone architecture, we pretrain each of the five objectives with an S4, a Transformer and a Net1D backbone, keeping the pretraining data and training configuration identical. We compare the resulting models in terms of downstream performance, layer-wise representational similarity and computational efficiency. Notably, the S4 backbone achieves the best results with only 3M parameters, compared to 19.2M for the Transformer and 10M for Net1D.

A.6.1 Downstream Performance

Tables 12–16 report finetuning results for each pretraining objective with three backbones, a 4-layer S4, a 6-block Transformer and a 7-stage Net1D. All models were pretrained on the 753K-sample HEEDB subset for 20 epochs with a learning rate of 3e-3. The best-performing result is highlighted in boldface and underlined, while models that do not perform statistically significantly worse are also highlighted in boldface.

Table 12: Finetuning performance (macro-AUROC) of data2vec with different backbones.

	data2vec (Finetuned)		
	S4	**Transformer**	**Net1D**
Adult ECG interpretation			
Ningbo	**<u>0.965</u>**	**0.962**	**0.964**
CPSC2018	**<u>0.960</u>**	0.950	**0.959**
CPSC-Extra	**0.839**	**<u>0.858</u>**	**0.858**
Georgia	**<u>0.894</u>**	0.867	**0.892**
Chapman	**<u>0.946</u>**	0.935	**0.937**
SPH	**<u>0.976</u>**	0.964	**0.975**
PTB-XL (all)	**0.924**	0.905	**<u>0.925</u>**
PTB-XL (sub)	**<u>0.934</u>**	0.915	**0.930**
PTB-XL (super)	**<u>0.926</u>**	0.911	0.917
Pediatric ECG interpretation			
ZZU pECG	**0.868**	0.850	**<u>0.873</u>**
Cardiac structure & function			
EchoNext	**0.816**	**0.821**	**<u>0.823</u>**

Table 13: Finetuning performance (macro-AUROC) of DinoSR with different backbones.

DinoSR (Finetuned)	S4	Transformer	Net1D
Adult ECG interpretation			
Ningbo	**0.964**	0.956	**0.967**
CPSC2018	**0.968**	0.942	0.962
CPSC-Extra	**0.883**	0.843	**0.884**
Georgia	**0.906**	0.849	0.889
Chapman	**0.960**	0.925	0.952
SPH	**0.975**	0.956	0.968
PTB-XL (all)	**0.935**	0.907	0.920
PTB-XL (sub)	**0.938**	0.921	0.925
PTB-XL (super)	**0.932**	0.905	0.913
Pediatric ECG interpretation			
ZZU pECG	**0.888**	0.861	**0.882**
Cardiac structure & function			
EchoNext	**0.829**	0.813	0.817

Table 14: Finetuning performance (macro-AUROC) of JEPA with different backbones.

JEPA (Finetuned)	S4	Transformer	Net1D
Adult ECG interpretation			
Ningbo	**0.971**	0.952	0.936
CPSC2018	**0.970**	0.949	0.945
CPSC-Extra	**0.888**	0.853	0.829
Georgia	**0.913**	0.868	0.844
Chapman	**0.957**	0.934	0.927
SPH	**0.976**	0.961	0.921
PTB-XL (all)	**0.943**	0.907	0.896
PTB-XL (sub)	**0.924**	0.901	0.875
PTB-XL (super)	**0.932**	0.896	0.898
Pediatric ECG interpretation			
ZZU pECG	**0.891**	0.847	0.841
Cardiac structure & function			
EchoNext	**0.826**	0.802	0.801

Table 15: Finetuning performance (macro-AUROC) of CPC with different backbones.

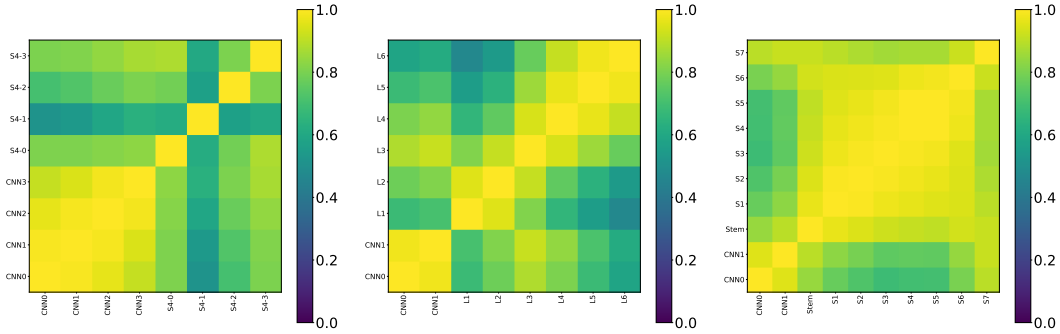
	CPC (Finetuned)		
	S4	**Transformer**	**Net1D**
Adult ECG interpretation			
Ningbo	**0.974**	0.962	0.945
CPSC2018	**0.972**	0.950	0.942
CPSC-Extra	**0.906**	0.814	0.840
Georgia	**0.912**	0.875	0.878
Chapman	**0.959**	0.932	0.929
SPH	**0.982**	0.965	0.943
PTB-XL (all)	**0.943**	0.907	0.902
PTB-XL (sub)	**0.937**	0.911	0.921
PTB-XL (super)	**0.932**	0.914	0.907
Pediatric ECG interpretation			
ZZU pECG	**0.899**	0.864	0.852
Cardiac structure & function			
EchoNext	**0.832**	0.812	0.816

Table 16: Finetuning performance (macro-AUROC) of HuBERT++ with different backbones.

	HuBERT++ (Finetuned)		
	S4	**Transformer**	**Net1D**
Adult ECG interpretation			
Ningbo	**0.970**	0.940	0.952
CPSC2018	**0.970**	0.926	0.941
CPSC-Extra	**0.891**	0.774	0.847
Georgia	**0.915**	0.833	0.861
Chapman	**0.958**	0.906	0.916
SPH	**0.982**	0.948	0.951
PTB-XL (all)	**0.939**	0.904	0.905
PTB-XL (sub)	**0.940**	0.888	0.906
PTB-XL (super)	**0.930**	0.888	0.903
Pediatric ECG interpretation			
ZZU pECG	**0.896**	0.839	0.834
Cardiac structure & function			
EchoNext	**0.830**	0.805	0.808

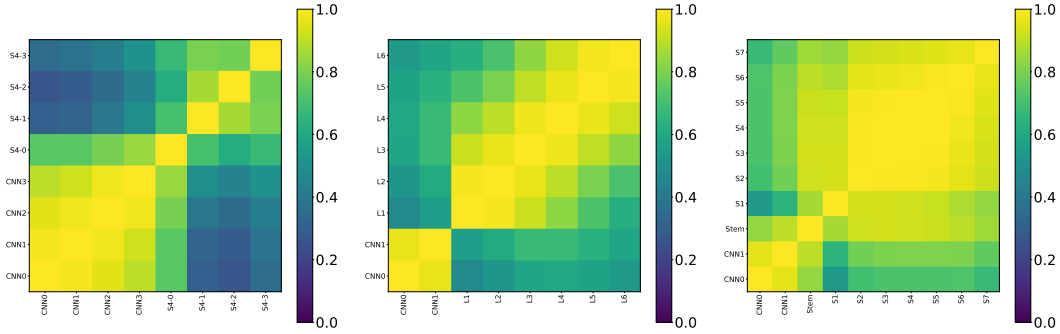
A.6.2 LAYER-WISE REPRESENTATIONAL SIMILARITY

Figures 6–10 show intra-model layer-wise representational similarity for all five pretraining objectives and the three backbones, using the same models as in the downstream comparison above. Similarity is measured by CKA with a Gaussian RBF kernel ($\sigma = 1.0$) on 2,500 PTB-XL samples, where warmer colors indicate greater similarity between layer pairs.



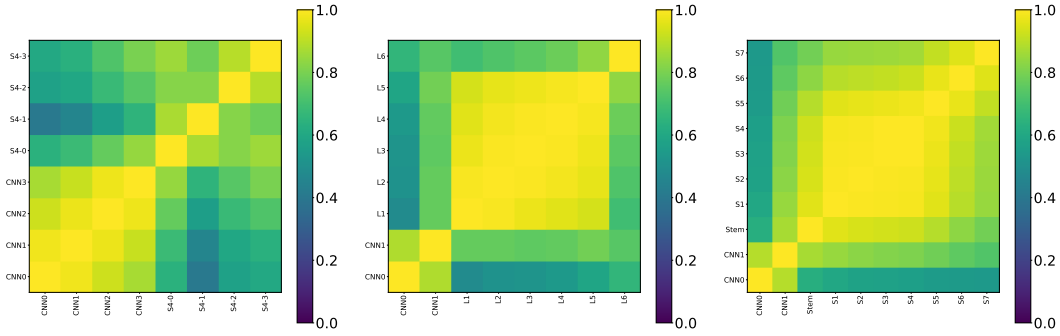
(a) S4 (b) Transformer (c) Net1D

Figure 6: Intra-model layer-wise representational similarity for data2vec.



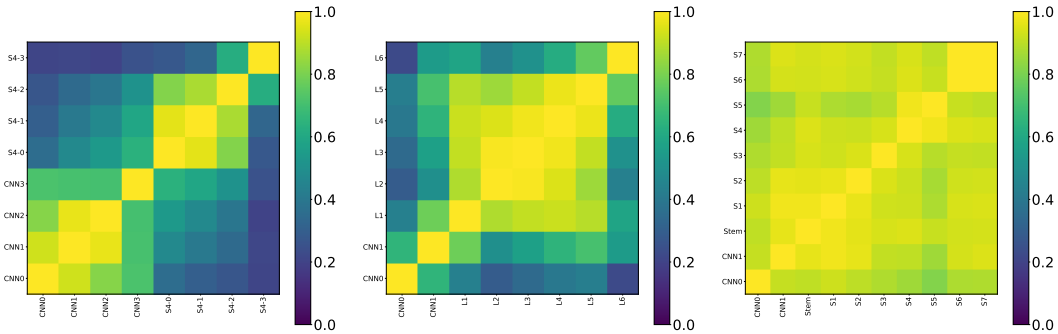
(a) S4 (b) Transformer (c) Net1D

Figure 7: Intra-model layer-wise representational similarity for DinoSR.



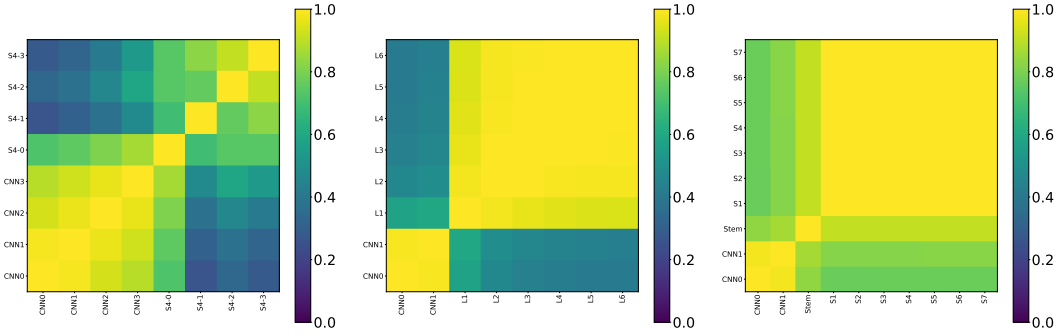
(a) S4 (b) Transformer (c) Net1D

Figure 8: Intra-model layer-wise representational similarity for JEPA.



(a) S4 (b) Transformer (c) Net1D

Figure 9: Intra-model layer-wise representational similarity for CPC.



(a) S4 (b) Transformer (c) Net1D

Figure 10: Intra-model layer-wise representational similarity for HuBERT++.

A.6.3 COMPUTATIONAL EFFICIENCY

Tab. 17 compares the computational cost of the three backbones under each pretraining objective. Parameters, FLOPs and memory depend only on the backbone and are therefore identical across objectives, while throughput and latency vary slightly between runs. The S4 backbone is the most efficient on all metrics, requiring about 16 times fewer forward FLOPs and 4 times less GPU memory than the Transformer and achieving roughly twice its throughput, while also achieving the best downstream performance.

Table 17: Computational cost of the S4, Transformer and Net1D backbones under each pretraining objective. GFLOPs for forward (F) and backward (B) passes are measured with batch size 1, while peak GPU memory, throughput (samples/s) and latency (ms/sample) are measured on PTB-XL (all) with batch size 64, all on an NVIDIA L40. The best value per objective is highlighted in boldface and underlined.

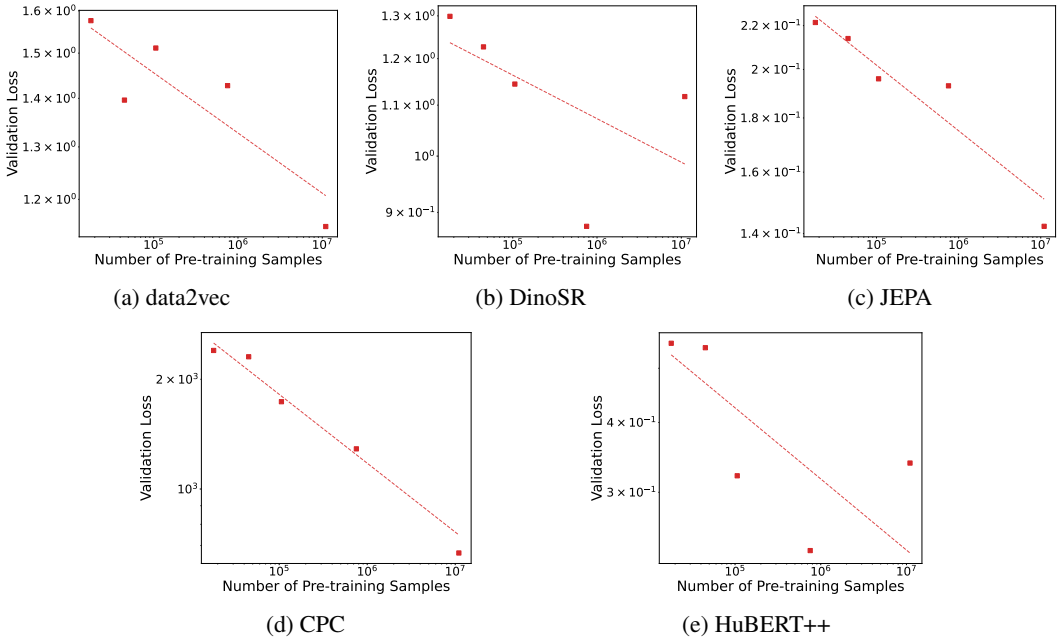
Model	Backbone	Parameters ↓	GFLOP (F/B) ↓	GPU Mem (MB) ↓	Thr↑/ Lat↓
data2vec	S4	**3M**	**1.741** / **5.213**	**482.59**	**441.10** / **2.267**
	Transformer	19.2M	27.410 / 82.207	1952.84	239.12 / 4.182
	Net1D	10M	8.845 / 73.817	582.76	103.92 / 9.623
DinoSR	S4	**3M**	**1.741** / **5.213**	**482.59**	**436.51** / **2.291**
	Transformer	19.2M	27.410 / 82.207	1952.84	215.89 / 4.632
	Net1D	10M	8.845 / 73.817	582.76	192.12 / 5.205
JEPA	S4	**3M**	**1.741** / **5.213**	**482.59**	**409.39** / **2.443**
	Transformer	19.2M	27.410 / 82.207	1952.84	223.23 / 4.480
	Net1D	10M	8.845 / 73.817	582.76	171.55 / 5.829
CPC	S4	**3M**	**1.741** / **5.213**	**482.59**	**401.14** / **2.493**
	Transformer	19.2M	27.410 / 82.207	1953.18	229.64 / 4.355
	Net1D	10M	8.845 / 73.817	582.76	186.24 / 5.369
HuBERT++	S4	**3M**	**1.741** / **5.213**	**482.59**	**418.21** / **2.391**
	Transformer	19.2M	27.410 / 82.207	1952.84	226.81 / 4.409
	Net1D	10M	8.845 / 73.817	582.76	164.67 / 6.073

A.7 Scaling Analysis

We study scaling with pretraining dataset size in two settings. First, we compare the five pretraining objectives with a fixed S4 backbone across pretraining datasets from 18K to 11M samples. Second, we compare four backbone architectures for CPC and JEPA across pretraining datasets from 18K to 753K samples.

A.7.1 Scaling Across Pretraining Objectives

Validation loss scaling Figure 11 shows the pretraining validation loss as a function of pretraining dataset size and Table 18 reports the corresponding power-law fit parameters.



(a) data2vec (b) DinoSR (c) JEPA

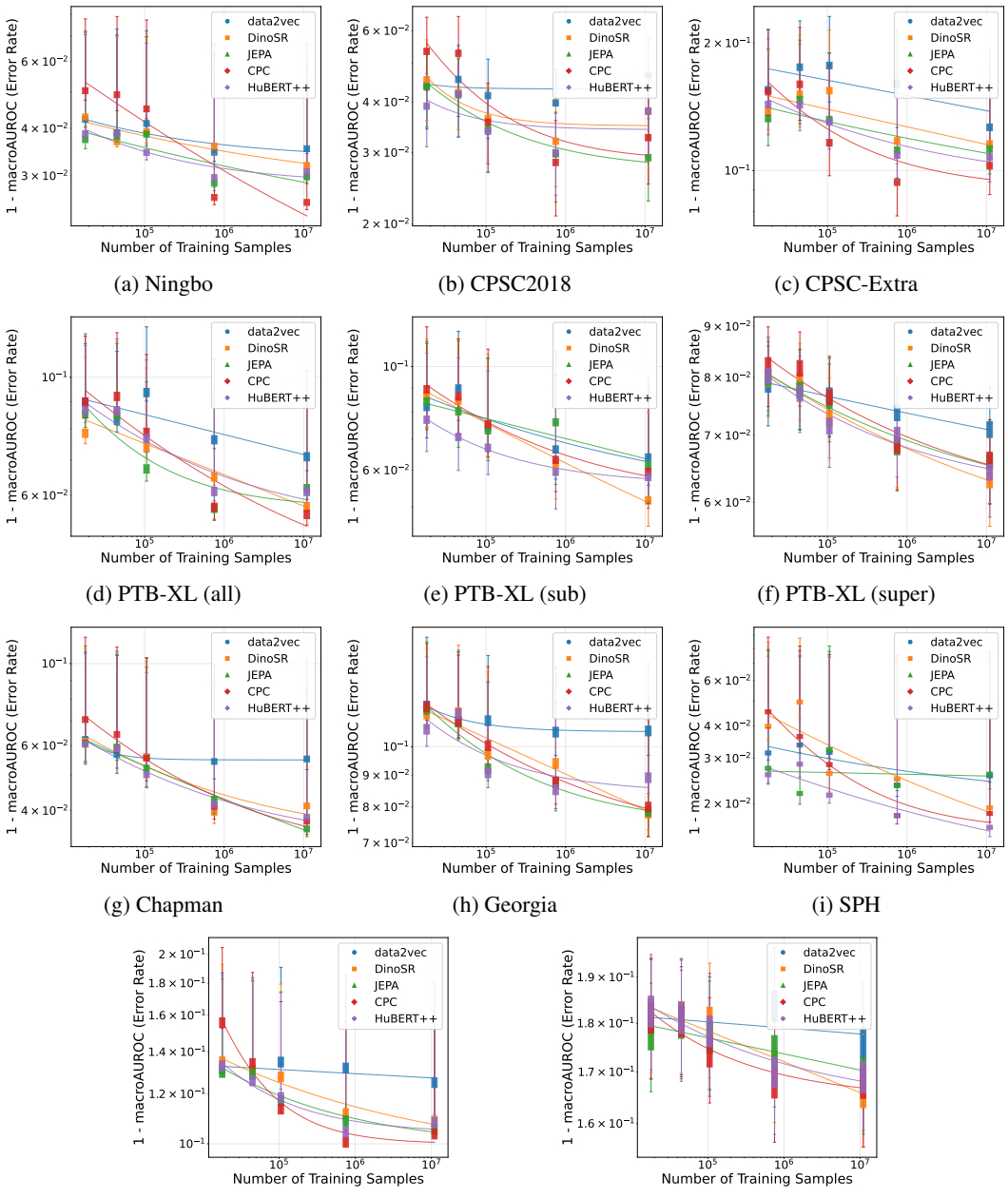
(d) CPC (e) HuBERT++

Figure 11: Validation loss as a function of pretraining dataset size for each model. Dashed lines show power-law fits of the form $f(n) = C \cdot N^{-\alpha}$.

Table 18: Power-law fit parameters $f(n) = C \cdot N^{-\alpha}$ for validation loss across models.

Model	C	α	R^2
data2vec	2.304	0.0399	0.7592
DinoSR	1.749	0.03538	0.3616
JEPA	0.4119	0.06194	0.9127
CPC	1.601×10^4	0.1889	0.9702
HuBERT++	1.838	0.1272	0.5606

Downstream performance scaling Figure 12 shows the downstream error, defined as $1 -$ macro-AUROC, as a function of pretraining dataset size N. We fit curves of the form $C \cdot N^{-\alpha} + L_0$, where α governs the rate of improvement and L_0 captures the residual error floor. Fit parameters for each dataset are given in Tables 19–29, and Table 30 summarizes the scaling exponents across datasets, weighted by fit quality. Not all combinations of objective and dataset show consistent scaling. The clearest scaling behavior is observed on PTB-XL (super), a comparatively coarse task with only 5 classes.



(a) Ningbo (b) CPSC2018 (c) CPSC-Extra

(d) PTB-XL (all) (e) PTB-XL (sub) (f) PTB-XL (super)

(g) Chapman (h) Georgia (i) SPH

(j) ZZU pECG (k) EchoNext

Figure 12: Downstream error ($1 -$ macro-AUROC) as a function of pretraining dataset size across adult ECG interpretation, pediatric ECG interpretation (ZZU pECG) and cardiac structure prediction (EchoNext).

Table 19: Fit parameters for the scaling analysis of Ningbo dataset.

Model	C	α	L_0	R^2
data2vec	0.1352	0.2761	0.0332	0.757
DinoSR	0.0631	0.1436	0.0261	0.827
JEPA	0.0581	0.1014	0.0175	0.708
CPC	0.2428	0.1752	0.0093	0.894
HuBERT++	0.3262	0.3438	0.0285	0.861

Table 20: Fit parameters for the scaling analysis of CPSC2018 dataset.

Model	C	α	L_0	R^2
data2vec	48.6179	1.0801	0.0431	0.033
DinoSR	38.3061	0.8317	0.0349	0.753
JEPA	0.9634	0.4093	0.0271	0.939
CPC	4.4522	0.5198	0.0286	0.818
HuBERT++	5.6429	0.6937	0.0342	0.294

Table 21: Fit parameters for the scaling analysis of CPSC-Extra dataset.

Model	C	α	L_0	R^2
data2vec	0.2475	0.0360	0.0000	0.518
DinoSR	0.2257	0.0417	0.0000	0.595
JEPA	0.1709	0.0869	0.0678	0.676
CPC	4.0753	0.4125	0.0902	0.789
HuBERT++	0.5334	0.2324	0.0924	0.912

Table 22: Fit parameters for the scaling analysis of PTB-XL (all) dataset.

Model	C	α	L_0	R^2
data2vec	0.1311	0.0376	0.0000	0.695
DinoSR	0.1479	0.0588	0.0000	0.847
JEPA	3.2329	0.4721	0.0566	0.827
CPC	0.4017	0.1964	0.0359	0.893
HuBERT++	0.4598	0.2535	0.0513	0.903

Table 23: Fit parameters for the scaling analysis of PTB-XL (sub) dataset.

Model	C	α	L_0	R^2
data2vec	0.1791	0.1625	0.0497	0.756
DinoSR	0.2114	0.1041	0.0120	0.982
JEPA	0.1268	0.0429	0.0000	0.845
CPC	0.5924	0.2752	0.0513	0.963
HuBERT++	1.0934	0.4022	0.0559	0.992

Table 24: Fit parameters for the scaling analysis of PTB-XL (super) dataset.

Model	C	α	L_0	R^2
data2vec	0.0555	0.0366	0.0402	0.994
DinoSR	0.1156	0.1372	0.0506	0.964
JEPA	0.1207	0.1687	0.0575	0.967
CPC	0.1717	0.1942	0.0580	0.948
HuBERT++	0.2071	0.2407	0.0606	0.958

Table 25: Fit parameters for the scaling analysis of Chapman dataset.

Model	C	α	L_0	R^2
data2vec	7209.1044	1.4070	0.0549	0.984
DinoSR	0.3651	0.2530	0.0330	0.881
JEPA	0.1618	0.1190	0.0117	0.979
CPC	0.7315	0.2909	0.0296	0.979
HuBERT++	0.3795	0.2583	0.0318	0.955

Table 26: Fit parameters for the scaling analysis of Georgia dataset.

Model	C	α	L_0	R^2
data2vec	13.1679	0.7297	0.1057	0.935
DinoSR	0.2007	0.0593	0.0018	0.923
JEPA	1.2382	0.3460	0.0744	0.946
CPC	0.4540	0.2273	0.0680	0.992
HuBERT++	1.7855	0.4301	0.0843	0.663

Table 27: Fit parameters for the scaling analysis of SPH dataset.

Model	C	α	L_0	R^2
data2vec	0.0707	0.1692	0.0199	0.661
DinoSR	0.2519	0.2001	0.0087	0.672
JEPA	0.0222	0.0089	0.0064	0.013
CPC	4.1350	0.5014	0.0157	0.981
HuBERT++	0.0954	0.1712	0.0097	0.829

Table 28: Fit parameters for the scaling analysis of ZZU pECG dataset.

Model	C	α	L_0	R^2
data2vec	0.1417	0.0067	0.0000	0.369
DinoSR	0.2564	0.1844	0.0944	0.947
JEPA	0.5077	0.2785	0.0988	0.931
CPC	57.4340	0.7056	0.0999	0.977
HuBERT++	3.0097	0.4690	0.1040	0.933

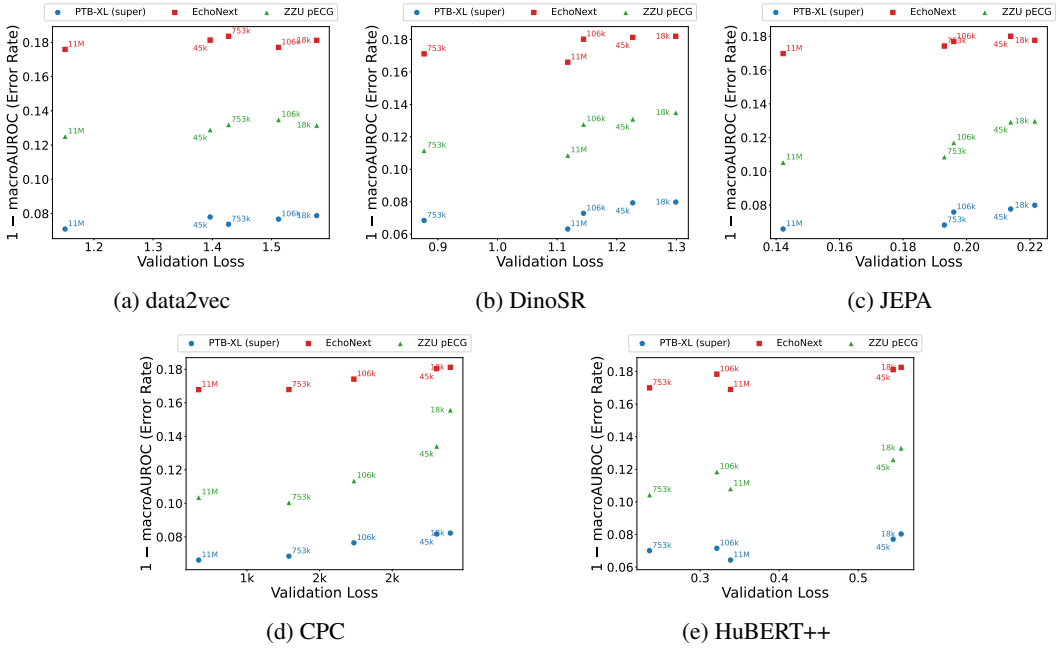
Table 29: Fit parameters for the scaling analysis of EchoNext dataset.

Model	C	α	L_0	R^2
data2vec	0.1796	0.0032	0.0072	0.196
DinoSR	0.2111	0.0160	0.0028	0.958
JEPA	0.1944	0.0082	0.0001	0.880
CPC	0.4532	0.3308	0.1647	0.921
HuBERT++	0.1656	0.2115	0.1627	0.938

Table 30: Weighted average of the scaling exponent $\bar{\alpha}$ across datasets, using R^2-weighted and R^4-weighted ($(R^2)^2$) aggregation.

Model	$\bar{\alpha}$ (R^2 weighted)	$\bar{\alpha}$ ($(R^2)^2$ weighted)
data2vec	0.3813	0.4479
DinoSR	0.1772	0.1692
JEPA	0.2084	0.2131
CPC	0.3474	0.3471
HuBERT++	0.3118	0.3023

Correlation between validation loss and downstream error To test whether a lower pretraining loss translates into better downstream performance, Fig. 13 plots the downstream error against the pretraining validation loss across all pretraining dataset sizes, and Tab. 31 reports the corresponding Spearman rank correlations on three datasets representing adult ECG interpretation, pediatric ECG interpretation and cardiac structure prediction.



(a) data2vec (b) DinoSR (c) JEPA

(d) CPC (e) HuBERT++

Figure 13: Correlation between pretraining validation loss and downstream error (1 − macro-AUROC) on PTB-XL (super), ZZU pECG and EchoNext.

Table 31: Spearman's rank correlation (r) between pretraining validation loss and downstream error rate (1 – macro-AUROC) for five self-supervised models across three datasets.

	PTB-XL (super)		**EchoNext**		**ZZU pECG**	
Model	r	p	r	p	r	p
data2vec	0.700	0.188	0.200	0.747	0.700	0.188
DinoSR	0.900	0.037	0.900	0.037	0.900	0.037
JEPA	1.000	0.000	0.900	0.037	1.000	0.000
CPC	1.000	0.000	1.000	0.000	0.900	0.037
HuBERT++	0.700	0.188	0.700	0.188	0.900	0.037

Downstream results at intermediate scales Tables 32 and 33 report finetuning results for all five objectives pretrained on the 106K- and 753K-sample HEEDB subsets, using the same configuration as the main results (4-layer S4 backbone, 10 epochs), together with the supervised S4 baseline. The best-performing result is highlighted in boldface and underlined, while models that do not perform statistically significantly worse are also highlighted in boldface.

Table 32: Finetuning performance (macro-AUROC) of models pretrained on the 106K-sample HEEDB subset.

	Pretrained Models (Finetuned)					**Supervised**
	data2vec	**DinoSR**	**JEPA**	**CPC**	**HuBERT++**	**S4**
Adult ECG interpretation						
Ningbo	0.959	0.961	0.961	0.955	0.965	**0.972**
CPSC2018	0.959	**0.964**	**0.966**	**0.964**	**0.966**	**0.962**
CPSC-Extra	0.823	0.846	0.868	**0.884**	**0.870**	0.852
Georgia	0.890	**0.903**	**0.908**	0.900	**0.909**	**0.903**
Chapman	0.944	0.944	0.948	0.944	0.950	**0.963**
SPH	0.968	**0.974**	0.968	0.972	**0.978**	**0.981**
PTB-XL (all)	0.906	0.926	0.933	0.921	0.923	**0.941**
PTB-XL (sub)	0.926	0.925	0.927	0.925	**0.933**	**0.938**
PTB-XL (super)	0.923	0.927	0.924	0.924	0.928	**0.932**
Pediatric ECG interpretation						
ZZU pECG	0.865	0.872	0.883	**0.886**	0.881	**0.897**
Cardiac structure & function						
EchoNext	**0.823**	0.820	**0.823**	**0.826**	**0.822**	**0.819**

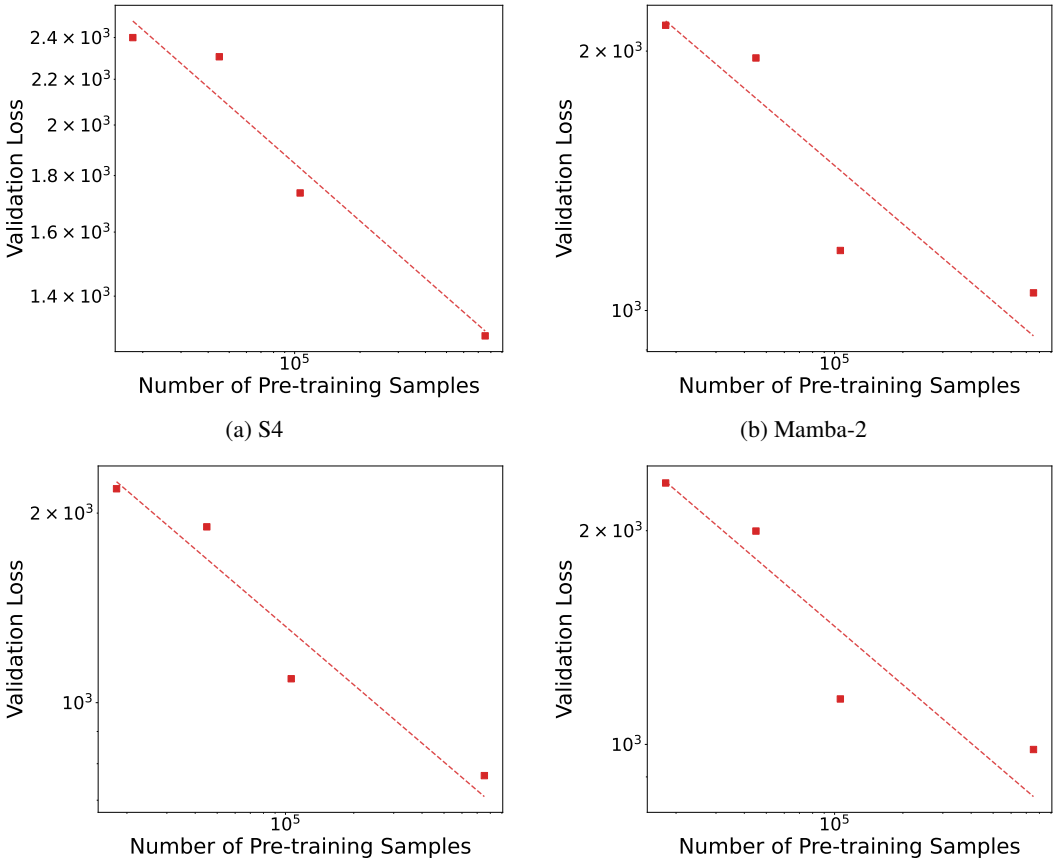
Table 33: Finetuning performance (macro-AUROC) of models pretrained on the 753K-sample HEEDB subset.

	Pretrained Models (Finetuned)					**Supervised**
	data2vec	**DinoSR**	**JEPA**	**CPC**	**HuBERT++**	**S4**
Adult ECG interpretation						
Ningbo	0.965	0.964	0.971	**0.974**	0.970	**0.972**
CPSC2018	0.960	**0.968**	**0.970**	**0.972**	**0.970**	0.962
CPSC-Extra	0.839	0.883	0.888	**0.906**	**0.891**	0.852
Georgia	0.894	**0.906**	**0.913**	**0.912**	**0.915**	**0.903**
Chapman	0.946	**0.960**	0.957	**0.959**	**0.958**	**0.963**
SPH	**0.976**	0.975	**0.976**	**0.982**	**0.982**	**0.981**
PTB-XL (all)	0.924	0.935	**0.943**	**0.943**	**0.939**	**0.941**
PTB-XL (sub)	0.934	**0.938**	**0.924**	**0.937**	**0.940**	**0.938**
PTB-XL (super)	0.926	**0.932**	**0.932**	**0.932**	**0.930**	**0.932**
Pediatric ECG interpretation						
ZZU pECG	0.868	0.888	0.891	**0.899**	**0.896**	**0.897**
Cardiac structure & function						
EchoNext	0.816	**0.829**	0.826	**0.832**	**0.830**	0.819

A.7.2 Scaling Across Backbone Architectures

To test whether the scaling behavior depends on the backbone, we pretrain CPC and JEPA with four backbones, namely S4, Mamba-2, and a Transformer with either rotary (RoPE) or sinusoidal positional encodings, on HEEDB subsets of 18K, 45K, 106K and 753K samples. Due to the computational constraints, we did not extend this comparison to the full 11M-sample pretraining dataset. As a result, the S4 fit for CPC differs slightly from the one reported in App. A.7.1. For each objective, we report the scaling of the pretraining validation loss and its correlation with downstream error on PTB-XL (super), ZZU pECG and EchoNext.

CPC validation loss scaling Fig. 14 shows the CPC validation loss as a function of pretraining dataset size for each backbone, with power-law fit parameters in Tab. 34.



(a) S4 (b) Mamba-2

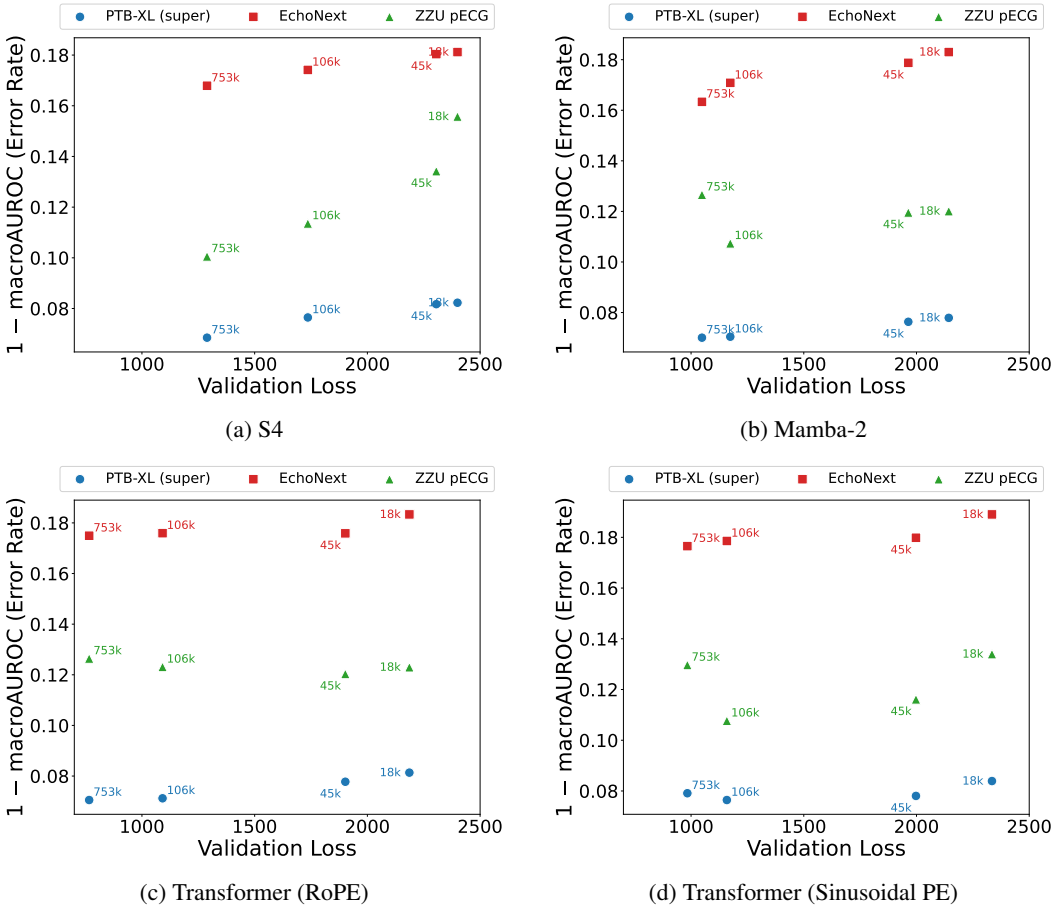
(c) Transformer (RoPE) (d) Transformer (Sinusoidal PE)

Figure 14: CPC validation loss as a function of pretraining dataset size for each backbone. Dashed lines show power-law fits of the form $f(n) = C \cdot N^{-\alpha}$.

Table 34: Power-law fit parameters $f(n) = C \cdot N^{-\alpha}$ for CPC validation loss across backbone architectures.

Backbone architecture	C	α	R^2
S4	1.354×10^4	0.1730	0.9380
Mamba-2	1.972×10^4	0.2253	0.8553
Transformer (RoPE)	4.604×10^4	0.3084	0.9299
Transformer (Sinusoidal PE)	3.419×10^4	0.2735	0.8974

CPC correlation between validation loss and downstream error Fig. 15 plots the downstream error against the CPC validation loss for each backbone, and Tab. 35 reports the corresponding correlations and linear fit slopes.



(a) S4

(b) Mamba-2

(c) Transformer (RoPE)

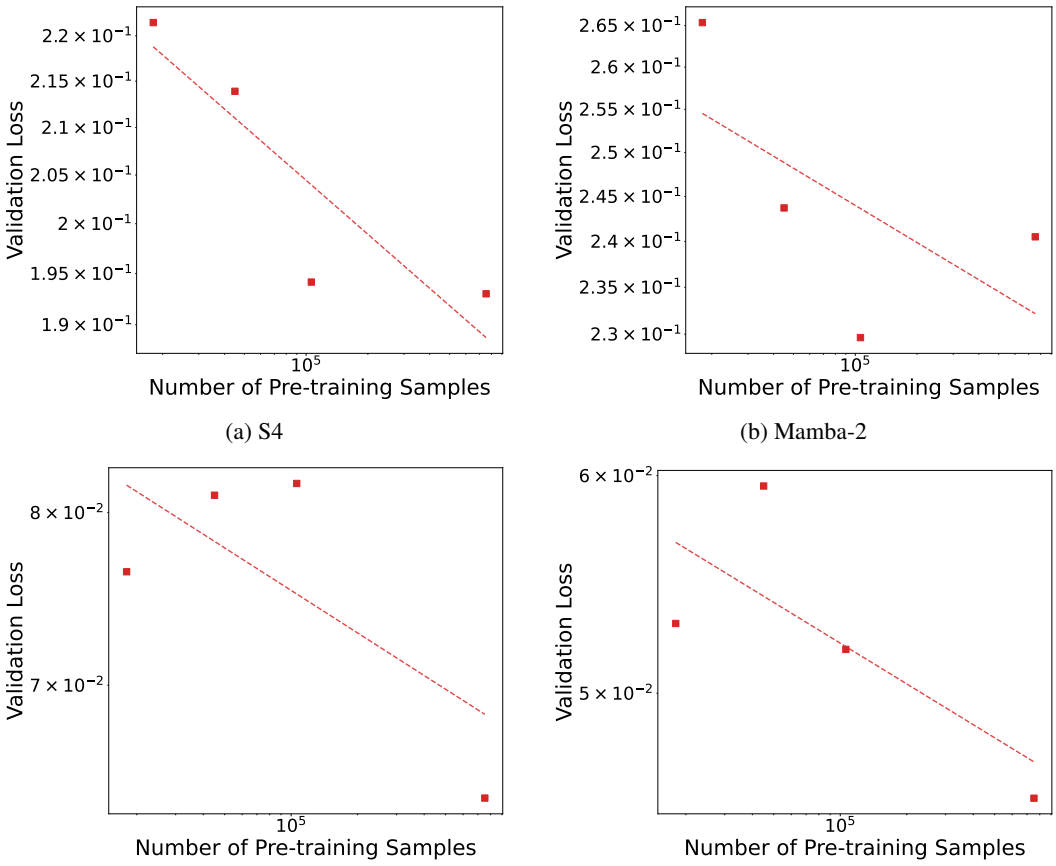
(d) Transformer (Sinusoidal PE)

Figure 15: Correlation between CPC validation loss and downstream classification error (1 – macro-AUROC) on PTB-XL (super), ZZU pECG and EchoNext.

Table 35: Correlation between CPC pretraining validation loss and downstream error (1 − macro-AUROC) for different backbones. The slope β is obtained from a linear fit and reported in units of 10^{-6}. Pearson p-values are given in parentheses.

Backbone		PTB-XL (super)	ZZU pECG	EchoNext
S4	Spearman ρ	0.949	1.000	1.000
	Pearson r	0.985 (0.015)	0.946 (0.054)	0.998 (0.002)
	Slope β	12.57	43.74	11.56
Mamba-2	Spearman ρ	0.949	−0.316	1.000
	Pearson r	0.994 (0.007)	0.114 (0.886)	0.958 (0.042)
	Slope β	7.43	1.72	15.41
Transformer (RoPE)	Spearman ρ	0.949	−0.633	0.949
	Pearson r	0.978 (0.022)	−0.695 (0.305)	0.759 (0.241)
	Slope β	7.41	−2.55	4.20
Transformer (Sinusoidal)	Spearman ρ	0.400	0.400	1.000
	Pearson r	0.689 (0.311)	0.325 (0.675)	0.841 (0.159)
	Slope β	3.60	6.04	7.49

JEPA validation loss scaling Fig. 16 shows the JEPA validation loss as a function of pretraining dataset size for each backbone, with power-law fit parameters in Tab. 36.



(a) S4 (b) Mamba-2

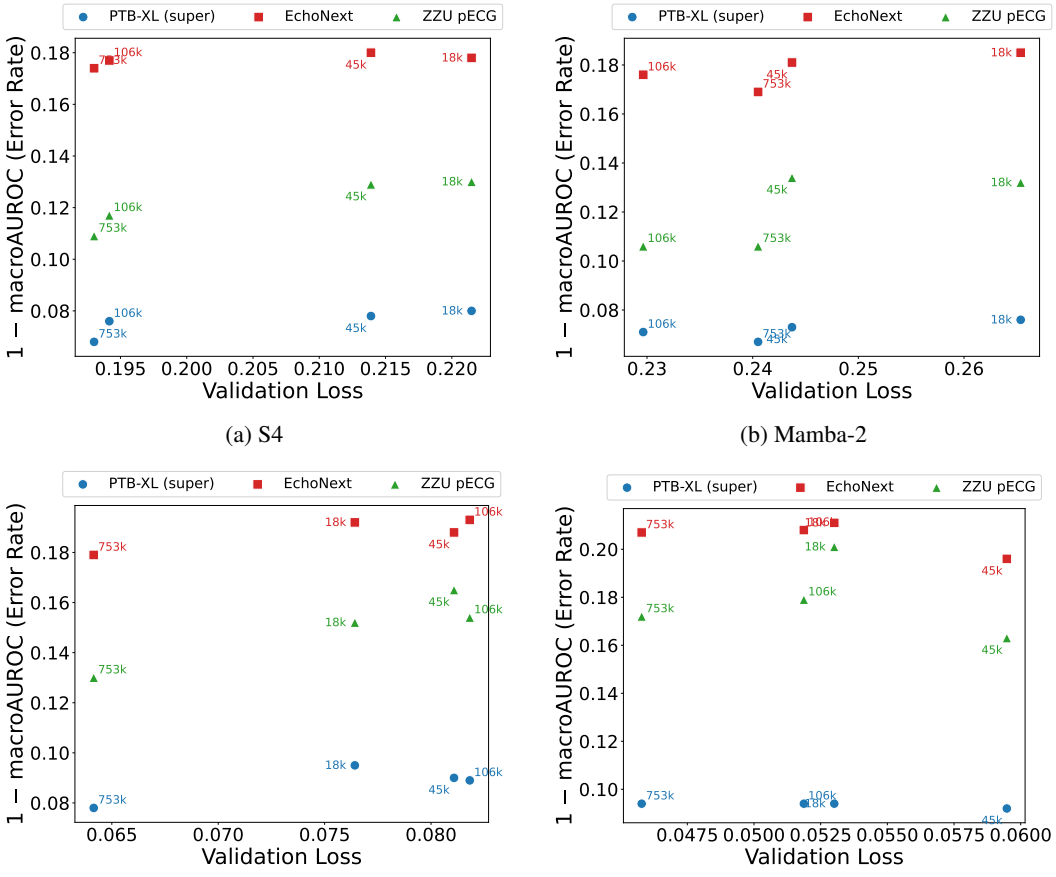
(c) Transformer (RoPE) (d) Transformer (Sinusoidal PE)

Figure 16: JEPA validation loss as a function of pretraining dataset size for each backbone. Dashed lines show power-law fits of the form $f(n) = C \cdot N^{-\alpha}$.

Table 36: Power-law fit parameters $f(n) = C \cdot N^{-\alpha}$ for JEPA validation loss across backbone architectures.

Backbone architecture	C	α	R^2
S4	0.3223	0.0395	0.7870
Mamba-2	0.3240	0.0246	0.3906
Transformer (RoPE)	0.1301	0.0474	0.5032
Transformer (Sinusoidal PE)	0.0919	0.0493	0.5398

JEPA correlation between validation loss and downstream error Fig. 17 plots the downstream error against the JEPA validation loss for each backbone, and Tab. 37 reports the corresponding correlations and linear fit slopes.



(a) S4

(b) Mamba-2

(c) Transformer (RoPE)

(d) Transformer (Sinusoidal PE)

Figure 17: Correlation between JEPA validation loss and downstream classification error (1 – macro-AUROC) on PTB-XL (super), ZZU pECG and EchoNext.

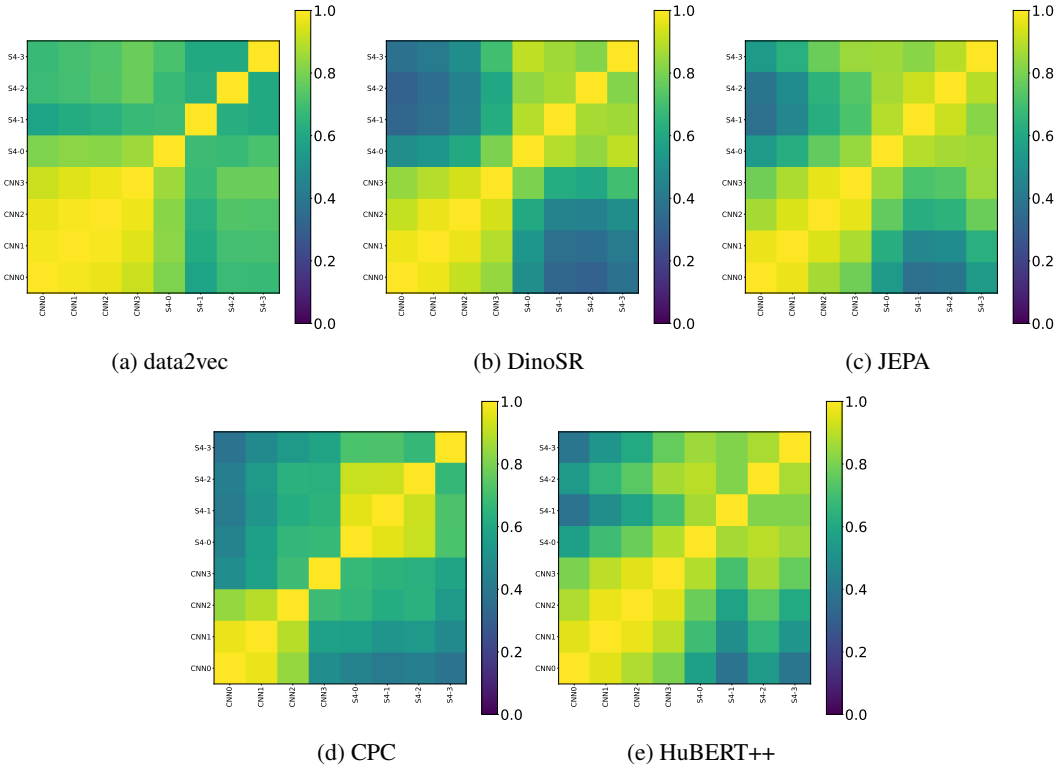
Table 37: Correlation between JEPA pretraining validation loss and downstream error (1 − macro-AUROC) for different backbones ($n = 4$). The slope β is obtained from a linear fit. Pearson p-values are given in parentheses.

Backbone		**PTB-XL (super)**	**ZZU pECG**	**EchoNext**
S4	Spearman ρ	1.000	1.000	0.800
	Pearson r	0.804 (0.196)	0.942 (0.058)	0.734 (0.266)
	Slope β	0.296	0.665	0.129
Mamba-2	Spearman ρ	0.800	0.738	0.800
	Pearson r	0.695 (0.305)	0.719 (0.281)	0.677 (0.323)
	Slope β	0.175	0.750	0.312
Transformer (RoPE)	Spearman ρ	0.200	0.800	0.800
	Pearson r	0.783 (0.217)	0.939 (0.061)	0.879 (0.121)
	Slope β	0.686	1.686	0.686
Transformer (Sinusoidal)	Spearman ρ	−0.775	−0.200	−0.200
	Pearson r	−0.825 (0.175)	−0.196 (0.804)	−0.681 (0.319)
	Slope β	−0.147	−0.566	−0.796

A.8 Further Analyses

A.8.1 Representation Similarity

Intra-model similarity Fig. 18 shows layer-wise CKA similarity within each model for all five pretraining objectives, complementing Fig. 5 in the main text, which omits data2vec.

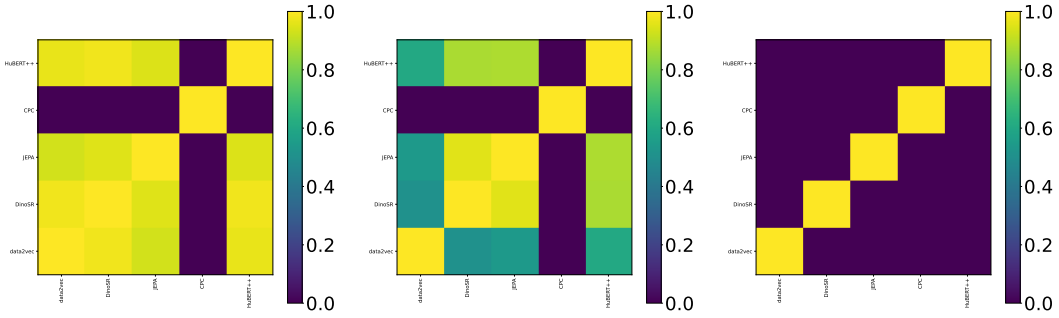


(a) data2vec (b) DinoSR (c) JEPA

(d) CPC (e) HuBERT++

Figure 18: Layer-wise representation similarity within each pretraining objective, measured by CKA with a Gaussian RBF kernel ($\sigma = 1.0$) on 2,500 PTB-XL samples. Warmer colors indicate greater similarity between layer pairs.

Inter-model similarity Fig. 19 compares representations across pretraining objectives at early, mid and late network stages, showing how similar the learned representations are.



(a) Early stage (b) Mid stage (c) Late stage

Figure 19: Inter-model representational similarity across network depths. CKA heatmaps comparing corresponding stages across five FMs (data2vec, DinoSR, JEPA, CPC and HuBERT++). Higher values (yellow) indicate similar representations between layers. CKA computed using Gaussian RBF kernel ($\sigma = 1.0$) on 256 samples of PTB-XL (all) dataset per model.

A.8.2 LABEL EFFICIENCY

Tab. 38 reports the fit parameters for the label efficiency curves on EchoNext shown in Fig. 4c, using the same functional form as in App. A.7.1.

Table 38: Fit parameters for the EchoNext label efficiency.

Model	C	α	L_0	R^2
data2vec	0.5207	0.1107	0.0000	0.984
DinoSR	0.4843	0.1274	0.0231	0.997
JEPA	0.4448	0.0986	0.0000	0.959
CPC	0.4565	0.1059	0.0000	0.988
HuBERT++	0.4684	0.1033	0.0000	0.980

A.8.3 COMPUTATIONAL EFFICIENCY

Tab. 39 compares the computational cost of our models with the external FMs ECGFounder and ECG-JEPA.

Table 39: Unified comparison of computational cost, memory usage, and inference efficiency for all ECG representation learning models. Reported metrics include: (i) GFLOPs for forward (F) and backward (B) passes (lower is better), measured with batch size 1 on an NVIDIA L40; (ii) peak GPU memory during inference (lower is better), measured on PTB-XL (all) with batch size 64; and (iii) throughput (samples/s; higher is better) and latency (ms/sample; lower is better) under the same hardware and batch size. Timesteps correspond to the input sequence length for each model, and parameter counts include all trainable weights. The best model is highlighted in bold face and underlined.

Model	**Timesteps**	**Parameters ↓**	**GFLOP (F/B) ↓**	**GPU Mem (MB) ↓**	**Thr↑/ Lat↓**
ECGFounder (CNN)	1250	33.8M	**0.602 / 5.066**	**187.37**	**2220.67 / 0.450**
ECG-JEPA (Transformer)	2500	87.2M	73.877 / 221.6	2136.79	98.71 / 10.131
data2vec (SSM)	600	**3M**	1.741 / 5.213	482.59	441.10 / 2.267
DinoSR (SSM)	600	**3M**	1.741 / 5.213	482.59	436.51 / 2.291
JEPA (SSM)	600	**3M**	1.741 / 5.213	482.59	409.39 / 2.443
CPC (SSM)	600	**3M**	1.741 / 5.213	480.53	401.14 / 2.493
HuBERT++ (SSM)	600	**3M**	1.741 / 5.213	482.59	418.21 / 2.391

A.8.4 FINETUNING STRATEGIES

Tables 40 and 41 compare three finetuning strategies for CPC and HuBERT++, namely finetuning with a linear head (Linear), finetuning with a learnable query-attention pooling head (Non-linear), and continual pretraining on the downstream data for 20 epochs followed by finetuning with a linear head (Domain Pretrain + Linear). All models were pretrained on the full 11M-sample dataset for 20 epochs with a 6-layer S4 backbone. The best-performing result is highlighted in boldface and underlined, while models that do not perform statistically significantly worse are also highlighted in boldface.

Table 40: Finetuning performance (macro-AUROC) of CPC with different finetuning strategies.

	CPC (Finetuned)		
	Linear	**Non-linear**	**Domain Pretrain + Linear**
Adult ECG interpretation			
Ningbo	**0.971**	0.963	**0.967**
CPSC2018	0.968	0.961	**0.974**
CPSC-Extra	**0.892**	0.867	**0.903**
Georgia	0.899	0.889	**0.909**
Chapman	**0.959**	0.946	**0.963**
SPH	**0.981**	0.973	**0.977**
PTB-XL (all)	**0.934**	0.921	**0.937**
PTB-XL (sub)	0.921	0.919	**0.934**
PTB-XL (super)	0.916	0.914	**0.921**
Pediatric ECG interpretation			
ZZU pECG	0.880	**0.885**	**0.894**
Cardiac structure & function			
EchoNext	0.820	0.813	**0.829**

Table 41: Finetuning performance (macro-AUROC) of HuBERT++ with different finetuning strategies.

	HuBERT++ (Finetuned)		
	Linear	**Non-linear**	**Domain Pretrain + Linear**
Adult ECG interpretation			
Ningbo	**0.973**	0.964	**0.975**
CPSC2018	0.968	0.968	**0.974**
CPSC-Extra	**0.896**	0.863	**0.899**
Georgia	**0.900**	0.884	**0.901**
Chapman	0.959	0.946	**0.964**
SPH	**0.976**	0.975	**0.982**
PTB-XL (all)	**0.937**	0.929	**0.936**
PTB-XL (sub)	**0.931**	**0.935**	**0.931**
PTB-XL (super)	0.918	0.915	**0.922**
Pediatric ECG interpretation			
ZZU pECG	**0.887**	**0.888**	**0.888**
Cardiac structure & function			
EchoNext	**0.822**	0.813	**0.825**

A.8.5 Cross-Dataset Transfer

To assess how well the learned representations generalize across data distributions, we finetune each model on PTB-XL (all) and evaluate it without further training on six external datasets. Predictions are mapped to each target dataset via a label crosswalk restricted to clinically equivalent diagnostic classes. As shown in Tab. 42, DinoSR and CPC transfer most robustly, with DinoSR in the top group on all six datasets and CPC on five.

Label mapping Each source label is mapped to a PTB-XL (all) class only if both describe the same clinical entity, based on the clinical descriptions and, where available, matching SNOMED-CT codes. Labels without an exact counterpart, such as generic labels for which PTB-XL only has more specific classes or concepts absent from PTB-XL, are excluded. A label may map to several PTB-XL classes when these share the same SNOMED-CT code (e.g., NORM and SR for sinus rhythm). The mappings were drafted with the assistance of a large language model and verified manually by the authors. Tables 43–48 list all mappings.

Table 42: Cross-dataset transfer performance (macro-AUROC) of models finetuned on PTB-XL (all) and evaluated without further training on six external datasets. The best-performing result is highlighted in boldface and underlined, while models that do not perform statistically significantly worse are also highlighted in boldface.

	data2vec	**DinoSR**	**JEPA**	**CPC**	**HuBERT++**
Ningbo	0.808	**0.825**	0.793	**0.820**	**0.814**
CPSC2018	0.704	**0.747**	**0.734**	**0.746**	0.717
CPSC-Extra	**0.791**	**0.770**	0.713	**0.746**	0.754
Georgia	0.729	**0.759**	**0.741**	**0.764**	**0.746**
Chapman	0.736	**0.806**	0.744	**0.813**	**0.799**
SPH	0.830	**0.868**	0.797	0.828	0.830

Table 43: Label mapping from CPSC2018 to PTB-XL (all).

Source label	**Description**	**PTB-XL target(s)**
NORMAL	Normal ECG	NORM, SR
AF	Atrial fibrillation	AFIB
I-AVB	First-degree AV block	1AVB
LBBB	Left bundle branch block	CLBBB
PAC	Premature atrial contraction	PAC
PVC	Premature ventricular contraction	PVC, PRC(S)
STD	ST depression	STD_
STE	ST elevation	STE_

Table 44: Label mapping from Georgia to PTB-XL (all).

Source label	**PTB-XL target(s)**
1st degree AV block	1AVB
2nd degree AV block	2AVB
Anterior ischemia	ISCAN
Atrial fibrillation	AFIB
Atrial flutter	AFLT
Complete right bundle branch block	CRBBB
Incomplete left bundle branch block	ILBBB
Incomplete right bundle branch block	IRBBB
Inferior ischaemia	ISCIN
Lateral ischaemia	ISCLA
Left anterior fascicular block	LAFB
Left atrial enlargement	LAO/LAE
Left bundle branch block	CLBBB
Left posterior fascicular block	LPFB
Left ventricular hypertrophy	LVH
Low QRS voltages	LVOLT
Nonspecific intraventricular conduction disorder	IVCD
Nonspecific ST-T abnormality	STTC
Premature atrial contraction	PAC
Premature ventricular complexes	PVC, PRC(S)
Prolonged QT interval	LNGQT
Q wave abnormal	QWAVE
Right ventricular hypertrophy	RVH
Sinus arrhythmia	SARRH
Sinus bradycardia	SBRAD
Sinus rhythm	SR, NORM
Sinus tachycardia	STACH
ST depression	STD_
ST elevation	STE_
Supraventricular tachycardia	SVTAC
T wave abnormal	TAB_
T wave inversion	INVT
Ventricular ectopic beats	PVC

Table 45: Label mapping from CPSC-Extra to PTB-XL (all).

Source label	PTB-XL target(s)
1st degree AV block	1AVB
2nd degree AV block	2AVB
Anterior myocardial infarction	AMI
Atrial fibrillation	AFIB
Atrial flutter	AFLT
Complete heart block	3AVB
Complete right bundle branch block	CRBBB
Incomplete left bundle branch block	ILBBB
Incomplete right bundle branch block	IRBBB
Left bundle branch block	CLBBB
Left ventricular hypertrophy	LVH
Myocardial ischemia	ISC_
Nonspecific ST-T abnormality	STTC
Premature atrial contraction	PAC
Premature ventricular contractions	PVC
Right atrial hypertrophy	RAO/RAE
Right ventricular hypertrophy	RVH
Sinus arrhythmia	SARRH
Sinus bradycardia	SBRAD
Sinus tachycardia	STACH
ST depression	STD_
ST elevation	STE_
Supraventricular premature beats	SVARR
T wave abnormal	TAB_

Table 46: Label mapping from SPH to PTB-XL (all).

Source code	Description	PTB-XL target(s)
1	Normal ECG	NORM
21	Sinus tachycardia	STACH
22	Sinus bradycardia	SBRAD
23	Sinus arrhythmia	SARRH
30	Atrial premature complex(es)	PAC
50	Atrial fibrillation	AFIB
51	Atrial flutter	AFLT
60	Ventricular premature complex(es)	PVC
82	Prolonged PR interval	LPR
85	2:1 AV block	2AVB
88	Complete (third-degree) AV block	3AVB
101	Left anterior fascicular block	LAFB
105	Incomplete right bundle branch block	IRBBB
108	Ventricular preexcitation	WPW
125	Low voltage	LVOLT
140	Left atrial enlargement	LAO/LAE
142	Left ventricular hypertrophy	LVH
146	ST deviation with T-wave change	STTC
147	T-wave abnormality	TAB_
148	Prolonged QT interval	LNGQT
160	Anterior myocardial infarction	AMI
161	Inferior myocardial infarction	IMI
165	Anteroseptal myocardial infarction	ASMI

Table 47: Label mapping from Chapman to PTB-XL (all).

Source label	Description	PTB-XL target(s)
1AVB	First-degree AV block	1AVB
AF	Atrial flutter	AFLT
AFIB	Atrial fibrillation	AFIB
APB	Atrial premature beats	PAC
AQW	Q wave abnormal	QWAVE
IVB	Intraventricular block	IVCD
LFBBB	Left anterior fascicular block	LAFB
LVH	Left ventricular hypertrophy	LVH
MIFW	Myocardial infarction, anterior wall	AMI
MILW	Myocardial infarction, inferior wall	IMI
PRIE	Prolonged PR interval	LPR
QTIE	Prolonged QT interval	LNGQT
SA	Sinus arrhythmia	SARRH
SB	Sinus bradycardia	SBRAD
SR	Sinus rhythm	SR
ST	Sinus tachycardia	STACH
STDD	ST depression	STD_
STTC	ST-T change	STTC
SVT	Supraventricular tachycardia	SVTAC
TWC	T wave change	TAB_
TWO	T wave inversion	INVT
VPB	Ventricular premature beat	PVC
VPE	Ventricular preexcitation	WPW

Table 48: Label mapping from Ningbo to PTB-XL (all).

Source label	Description	PTB-XL target(s)
1AVB	First-degree AV block	1AVB
2AVB2	Second-degree AV block (type II)	2AVB
AF	Atrial flutter	AFLT
AnMI	Acute myocardial infarction, anterior wall	AMI
APB	Atrial premature beats	PAC
CAVB	Complete AV block	3AVB
CLBBB	Complete left bundle branch block	CLBBB
CRBBB	Complete right bundle branch block	CRBBB
IRBBB	Incomplete right bundle branch block	IRBBB
IVB	Intraventricular block	IVCD
LAnFB	Left anterior fascicular block	LAFB
LQRSV	Low QRS voltages	LVOLT
LVH	Left ventricular hypertrophy	LVH
MILW	Myocardial infarction, inferior wall	IMI
PR	Pacemaker rhythm	PACE
PRIE	Prolonged PR interval	LPR
PVC	Premature ventricular contractions	PVC
QAb	Q wave abnormal	QWAVE
RAH	Right atrial hypertrophy	RAO/RAE
RVH	Right ventricular hypertrophy	RVH
SA	Sinus arrhythmia	SARRH
SB	Sinus bradycardia	SBRAD
SR	Sinus rhythm	SR, NORM
ST	Sinus tachycardia	STACH
STC	ST changes	NST_
STDD	ST depression	STD_
SVT	Supraventricular tachycardia	SVTAC
TWO	T wave inversion	INVT
WPW	Wolff-Parkinson-White pattern	WPW

A.9 Ablation Studies

This section reports the ablations underlying the configuration choices in the main text. Unless stated otherwise, SSL models were pretrained on the 753K-sample HEEDB subset for 20 epochs with a 6-layer S4 backbone, and supervised models were trained from scratch for 100 epochs with learning rate 1e-3. In all tables, the best-performing result is highlighted in boldface and underlined, while models that do not perform statistically significantly worse are also highlighted in boldface.

A.9.1 Supervised Backbone Configurations

To select a strong configuration for each backbone family, we train each architecture from scratch in a supervised fashion across several sizes. For S4, we vary the model dimension (Tab. 49), and we additionally compare S4 with Mamba-1 and Mamba-2 (Tab. 50). For Transformers, we vary the model dimension for both RoPE and sinusoidal positional encodings (Tables 51 and 52), and for Net1D, we vary the block depth (Tab. 53).

Table 49: Performance (macro-AUROC) of supervised S4 models (6 layers) with model dimensions 512, 768 and 1024 and corresponding state dimensions 8, 12 and 16.

	S4 (Supervised)		
	512	**768**	**1024**
Adult ECG interpretation			
Ningbo	**0.973**	**0.973**	**0.972**
CPSC2018	**0.967**	**0.964**	**0.967**
CPSC-Extra	**0.888**	0.861	**0.889**
Georgia	**0.917**	0.898	0.902
Chapman	**0.957**	**0.956**	**0.960**
SPH	**0.978**	**0.979**	**0.980**
PTB-XL (all)	**0.940**	0.929	**0.938**
PTB-XL (sub)	**0.935**	**0.933**	**0.932**
PTB-XL (super)	**0.931**	**0.931**	**0.932**
Pediatric ECG interpretation			
ZZU pECG	0.879	**0.893**	0.880
Cardiac structure & function			
EchoNext	**0.823**	**0.823**	**0.822**

Table 50: Performance (macro-AUROC) of supervised S4, Mamba-1 and Mamba-2 models.

	SSMs (Supervised)		
	S4	**Mamba-1**	**Mamba-2**
Adult ECG interpretation			
Ningbo	**0.973**	0.957	0.944
CPSC2018	**0.967**	0.944	0.931
CPSC-Extra	**0.888**	0.840	0.792
Georgia	**0.917**	0.904	0.878
Chapman	**0.957**	0.892	0.938
SPH	**0.978**	0.873	0.949
PTB-XL (all)	**0.940**	0.921	0.917
PTB-XL (sub)	**0.935**	0.891	0.917
PTB-XL (super)	**0.931**	0.911	0.915
Pediatric ECG interpretation			
ZZU pECG	**0.879**	0.820	0.853
Cardiac structure & function			
EchoNext	**0.823**	0.794	0.796

Table 51: Performance (macro-AUROC) of supervised Transformers (6 blocks) with rotary positional encodings (RoPE) and model dimensions 512, 768 and 1024.

Transformer RoPE (Supervised)	512	768	1024
Adult ECG interpretation			
Ningbo	**0.823**	0.757	0.739
CPSC2018	**0.873**	0.833	0.807
CPSC-Extra	**0.849**	0.772	0.755
Georgia	**0.799**	0.732	0.705
Chapman	**0.845**	0.809	0.738
SPH	**0.831**	0.808	0.783
PTB-XL (all)	**0.850**	0.803	0.755
PTB-XL (sub)	**0.878**	0.822	0.767
PTB-XL (super)	**0.875**	0.824	0.786
Pediatric ECG interpretation			
ZZU pECG	**0.789**	0.715	0.694
Cardiac structure & function			
EchoNext	**0.732**	0.717	0.500

Table 52: Performance (macro-AUROC) of supervised Transformers (6 blocks) with sinusoidal positional encodings and model dimensions 512, 768 and 1024.

Transformer Sinusoidal PE (Supervised)	512	768	1024
Adult ECG interpretation			
Ningbo	**0.880**	0.865	0.832
CPSC2018	**0.901**	0.869	0.845
CPSC-Extra	**0.793**	**0.811**	0.791
Georgia	**0.841**	0.817	0.793
Chapman	**0.907**	0.889	0.832
SPH	**0.922**	0.883	0.842
PTB-XL (all)	**0.858**	0.850	0.818
PTB-XL (sub)	**0.892**	0.881	0.844
PTB-XL (super)	**0.875**	0.849	0.843
Pediatric ECG interpretation			
ZZU pECG	0.790	**0.812**	0.778
Cardiac structure & function			
EchoNext	**0.748**	0.729	**0.750**

Table 53: Performance (macro-AUROC) of supervised Net1D models with small $[2, 2, 2, 2, 2, 2, 2]$, base $[2, 2, 2, 3, 3, 4, 4]$ and large $[2, 2, 2, 3, 3, 16, 4]$ block depths across the seven stages.

Net1D (Supervised)	small	base	large
Adult ECG interpretation			
Ningbo	**0.960**	**0.964**	**0.966**
CPSC2018	**0.950**	**0.945**	**0.943**
CPSC-Extra	**0.826**	0.808	**0.846**
Georgia	**0.900**	**0.890**	**0.896**
Chapman	**0.950**	0.944	**0.956**
SPH	**0.953**	**0.973**	**0.956**
PTB-XL (all)	**0.922**	**0.924**	**0.926**
PTB-XL (sub)	0.925	**0.930**	**0.933**
PTB-XL (super)	**0.925**	**0.922**	**0.922**
Pediatric ECG interpretation			
ZZU pECG	0.873	**0.886**	**0.879**
Cardiac structure & function			
EchoNext	**0.806**	**0.808**	**0.810**

A.9.2 Pretraining Hyperparameters

Learning rate Tables 54–58 compare learning rates of 3e-3, 1e-3 and 3e-4 for all five objectives.

Table 54: Finetuning performance (macro-AUROC) of data2vec pretrained with learning rates 3e-3, 1e-3 and 3e-4.

data2vec (Finetuned)	3e-3	1e-3	3e-4
Adult ECG interpretation			
Ningbo	**0.960**	**0.960**	**0.963**
CPSC2018	**0.958**	**0.960**	0.954
CPSC-Extra	**0.860**	**0.853**	0.833
Georgia	**0.879**	**0.883**	**0.877**
Chapman	**0.927**	**0.929**	**0.931**
SPH	**0.965**	**0.965**	**0.969**
PTB-XL (all)	**0.908**	**0.908**	**0.914**
PTB-XL (sub)	**0.907**	**0.897**	0.892
PTB-XL (super)	**0.911**	**0.912**	0.909
Pediatric ECG interpretation			
ZZU pECG	**0.853**	**0.857**	0.849
Cardiac structure & function			
EchoNext	**0.818**	**0.820**	**0.818**

Table 55: Finetuning performance (macro-AUROC) of DinoSR pretrained with learning rates 3e-3, 1e-3 and 3e-4.

	DinoSR (Finetuned)		
	3e-3	**1e-3**	**3e-4**
Adult ECG interpretation			
Ningbo	**0.965**	**0.967**	0.962
CPSC2018	**0.965**	**0.966**	**0.963**
CPSC-Extra	**0.865**	**0.867**	**0.861**
Georgia	**0.909**	0.903	0.886
Chapman	**0.948**	**0.946**	0.934
SPH	**0.971**	**0.977**	**0.975**
PTB-XL (all)	**0.927**	**0.927**	0.921
PTB-XL (sub)	**0.930**	**0.925**	0.915
PTB-XL (super)	**0.920**	0.917	0.915
Pediatric ECG interpretation			
ZZU pECG	**0.878**	**0.870**	**0.870**
Cardiac structure & function			
EchoNext	**0.826**	**0.827**	0.815

Table 56: Finetuning performance (macro-AUROC) of JEPA pretrained with learning rates 3e-3, 1e-3 and 3e-4.

	JEPA (Finetuned)		
	3e-3	**1e-3**	**3e-4**
Adult ECG interpretation			
Ningbo	**0.962**	0.954	0.947
CPSC2018	**0.961**	**0.958**	0.953
CPSC-Extra	**0.846**	**0.837**	**0.849**
Georgia	**0.873**	**0.877**	**0.878**
Chapman	**0.934**	**0.937**	0.924
SPH	**0.965**	**0.960**	0.958
PTB-XL (all)	**0.925**	0.913	0.905
PTB-XL (sub)	**0.905**	0.894	0.884
PTB-XL (super)	**0.912**	**0.912**	0.909
Pediatric ECG interpretation			
ZZU pECG	**0.845**	**0.835**	**0.841**
Cardiac structure & function			
EchoNext	**0.817**	**0.821**	**0.817**

Table 57: Finetuning performance (macro-AUROC) of CPC pretrained with learning rates 3e-3, 1e-3 and 3e-4.

CPC (Finetuned)	3e-3	1e-3	3e-4
Adult ECG interpretation			
Ningbo	**0.971**	**0.968**	0.965
CPSC2018	**0.969**	**0.973**	**0.970**
CPSC-Extra	**0.882**	**0.893**	**0.893**
Georgia	**0.904**	**0.901**	0.896
Chapman	**0.958**	0.952	0.947
SPH	**0.979**	**0.978**	**0.976**
PTB-XL (all)	**0.936**	**0.938**	**0.933**
PTB-XL (sub)	**0.933**	**0.931**	0.915
PTB-XL (super)	**0.921**	**0.920**	0.915
Pediatric ECG interpretation			
ZZU pECG	**0.876**	**0.889**	**0.882**
Cardiac structure & function			
EchoNext	**0.824**	**0.822**	**0.821**

Table 58: Finetuning performance (macro-AUROC) of HuBERT++ pretrained with learning rates 3e-3, 1e-3 and 3e-4.

HuBERT++ (Finetuned)	3e-3	1e-3	3e-4
Adult ECG interpretation			
Ningbo	**0.970**	**0.970**	**0.968**
CPSC2018	**0.972**	**0.970**	0.966
CPSC-Extra	**0.884**	**0.892**	0.875
Georgia	**0.906**	0.896	0.889
Chapman	**0.955**	**0.954**	0.942
SPH	0.979	**0.981**	0.974
PTB-XL (all)	**0.932**	**0.935**	0.926
PTB-XL (sub)	**0.931**	**0.931**	0.923
PTB-XL (super)	**0.919**	0.916	0.912
Pediatric ECG interpretation			
ZZU pECG	**0.898**	0.889	0.880
Cardiac structure & function			
EchoNext	**0.824**	**0.823**	**0.821**

Model dimension Tables 59–63 compare S4 backbones with model dimensions 256 and 512.

Table 59: Finetuning performance (macro-AUROC) of data2vec pretrained with S4 backbone dimensions 256 and 512.

data2vec (Finetuned)	256	512
Adult ECG interpretation		
Ningbo	0.949	**0.960**
CPSC2018	0.938	**0.960**
CPSC-Extra	0.843	**0.853**
Georgia	0.866	**0.883**
Chapman	**0.931**	**0.929**
SPH	**0.959**	**0.965**
PTB-XL (all)	**0.906**	**0.908**
PTB-XL (sub)	**0.887**	**0.897**
PTB-XL (super)	0.906	**0.912**
Pediatric ECG interpretation		
ZZU pECG	0.837	**0.857**
Cardiac structure & function		
EchoNext	**0.817**	**0.820**

Table 60: Finetuning performance (macro-AUROC) of DinoSR pretrained with S4 backbone dimensions 256 and 512.

DinoSR (Finetuned)	256	512
Adult ECG interpretation		
Ningbo	0.959	**0.965**
CPSC2018	**0.956**	**0.965**
CPSC-Extra	**0.866**	**0.865**
Georgia	0.874	**0.909**
Chapman	0.937	**0.948**
SPH	**0.971**	**0.971**
PTB-XL (all)	**0.922**	**0.927**
PTB-XL (sub)	**0.922**	**0.930**
PTB-XL (super)	0.916	**0.920**
Pediatric ECG interpretation		
ZZU pECG	0.864	**0.878**
Cardiac structure & function		
EchoNext	0.819	**0.826**

Table 61: Finetuning performance (macro-AUROC) of JEPA pretrained with S4 backbone dimensions 256 and 512.

JEPA (Finetuned)	256	512
Adult ECG interpretation		
Ningbo	0.942	**0.962**
CPSC2018	0.945	**0.961**
CPSC-Extra	**0.851**	**0.846**
Georgia	**0.882**	**0.873**
Chapman	**0.935**	**0.934**
SPH	0.949	**0.965**
PTB-XL (all)	0.901	**0.925**
PTB-XL (sub)	**0.907**	**0.905**
PTB-XL (super)	**0.913**	**0.912**
Pediatric ECG interpretation		
ZZU pECG	**0.839**	**0.845**
Cardiac structure & function		
EchoNext	**0.814**	**0.817**

Table 62: Finetuning performance (macro-AUROC) of CPC pretrained with S4 backbone dimensions 256 and 512.

CPC (Finetuned)	256	512
Adult ECG interpretation		
Ningbo	0.963	**0.971**
CPSC2018	**0.969**	**0.969**
CPSC-Extra	**0.901**	**0.882**
Georgia	0.892	**0.904**
Chapman	**0.955**	**0.958**
SPH	0.976	**0.979**
PTB-XL (all)	**0.936**	**0.936**
PTB-XL (sub)	**0.929**	**0.933**
PTB-XL (super)	0.918	**0.921**
Pediatric ECG interpretation		
ZZU pECG	**0.883**	**0.876**
Cardiac structure & function		
EchoNext	**0.825**	**0.824**

Table 63: Finetuning performance (macro-AUROC) of HuBERT++ pretrained with S4 backbone dimensions 256 and 512.

HuBERT++ (Finetuned)	256	512
Adult ECG interpretation		
Ningbo	0.961	**0.970**
CPSC2018	0.948	**0.972**
CPSC-Extra	0.859	**0.884**
Georgia	0.875	**0.906**
Chapman	0.933	**0.955**
SPH	**0.962**	**0.979**
PTB-XL (all)	0.918	**0.932**
PTB-XL (sub)	0.902	**0.931**
PTB-XL (super)	0.904	**0.919**
Pediatric ECG interpretation		
ZZU pECG	0.850	**0.898**
Cardiac structure & function		
EchoNext	0.812	**0.824**

Input length Tables 64–68 compare input lengths of 2.5 s and 5 s. These models were pretrained on the 106K-sample HEEDB subset with a 4-layer S4 backbone.

Table 64: Finetuning performance (macro-AUROC) of data2vec pretrained with input lengths of 2.5 s and 5 s.

data2vec (Finetuned)	2.5s	5s
Adult ECG interpretation		
Ningbo	**0.957**	**0.959**
CPSC2018	0.953	**0.959**
CPSC-Extra	**0.839**	0.823
Georgia	**0.894**	**0.890**
Chapman	**0.947**	**0.944**
SPH	**0.962**	**0.968**
PTB-XL (all)	**0.929**	0.906
PTB-XL (sub)	**0.931**	**0.926**
PTB-XL (super)	**0.926**	0.923
Pediatric ECG interpretation		
ZZU pECG	0.847	**0.865**
Cardiac structure & function		
EchoNext	0.817	**0.823**

Table 65: Finetuning performance (macro-AUROC) of DinoSR pretrained with input lengths of 2.5 s and 5 s.

DinoSR (Finetuned)	2.5s	5s
Adult ECG interpretation		
Ningbo	**0.956**	**0.961**
CPSC2018	0.958	**0.964**
CPSC-Extra	**0.866**	0.846
Georgia	**0.904**	**0.903**
Chapman	**0.952**	0.944
SPH	**0.976**	**0.974**
PTB-XL (all)	**0.925**	**0.926**
PTB-XL (sub)	**0.931**	**0.925**
PTB-XL (super)	**0.928**	**0.927**
Pediatric ECG interpretation		
ZZU pECG	**0.869**	**0.872**
Cardiac structure & function		
EchoNext	**0.824**	**0.820**

Table 66: Finetuning performance (macro-AUROC) of JEPA pretrained with input lengths of 2.5 s and 5 s.

JEPA (Finetuned)	2.5s	5s
Adult ECG interpretation		
Ningbo	**0.960**	**0.961**
CPSC2018	**0.959**	**0.966**
CPSC-Extra	**0.868**	**0.868**
Georgia	**0.902**	**0.908**
Chapman	**0.952**	**0.948**
SPH	**0.969**	**0.968**
PTB-XL (all)	0.923	**0.933**
PTB-XL (sub)	**0.938**	0.927
PTB-XL (super)	**0.930**	0.924
Pediatric ECG interpretation		
ZZU pECG	0.864	**0.883**
Cardiac structure & function		
EchoNext	**0.825**	**0.823**

Table 67: Finetuning performance (macro-AUROC) of CPC pretrained with input lengths of 2.5 s and 5 s.

CPC (Finetuned)	2.5s	5s
Adult ECG interpretation		
Ningbo	**0.961**	0.955
CPSC2018	0.958	**0.964**
CPSC-Extra	**0.877**	**0.884**
Georgia	**0.901**	**0.900**
Chapman	**0.948**	**0.944**
SPH	**0.959**	**0.972**
PTB-XL (all)	**0.929**	0.921
PTB-XL (sub)	**0.932**	0.925
PTB-XL (super)	**0.929**	0.924
Pediatric ECG interpretation		
ZZU pECG	**0.888**	**0.886**
Cardiac structure & function		
EchoNext	**0.826**	**0.826**

Table 68: Finetuning performance (macro-AUROC) of HuBERT++ pretrained with input lengths of 2.5 s and 5 s.

HuBERT++ (Finetuned)	2.5s	5s
Adult ECG interpretation		
Ningbo	**0.964**	**0.965**
CPSC2018	**0.960**	**0.966**
CPSC-Extra	**0.866**	**0.870**
Georgia	0.897	**0.909**
Chapman	**0.949**	**0.950**
SPH	**0.974**	**0.978**
PTB-XL (all)	**0.925**	**0.923**
PTB-XL (sub)	**0.937**	0.933
PTB-XL (super)	**0.931**	0.928
Pediatric ECG interpretation		
ZZU pECG	**0.876**	**0.881**
Cardiac structure & function		
EchoNext	**0.826**	**0.822**

A.9.3 Objective-Specific Ablations

HuBERT++ SSL head and codebook size Tab. 69 compares an MLP and an S4 SSL head for HuBERT++, each with codebook sizes [128, 256] and [256, 512].

Table 69: Finetuning performance (macro-AUROC) of HuBERT++ with different SSL heads and codebook sizes. Base uses an MLP head with codebooks of size [128, 256], S4 replaces the MLP with an S4 head, [256, 512] uses an MLP head with larger codebooks, and S4 + [256, 512] combines both changes.

	HuBERT++ (Finetuned)			
	Base	**S4**	**[256, 512]**	**S4 + [256, 512]**
Adult ECG interpretation				
Ningbo	**0.970**	**0.967**	**0.970**	**0.967**
CPSC2018	**0.972**	0.968	**0.971**	0.968
CPSC-Extra	**0.884**	**0.883**	**0.882**	**0.887**
Georgia	**0.906**	**0.903**	**0.910**	0.901
Chapman	**0.955**	**0.954**	**0.953**	0.948
SPH	**0.979**	0.976	**0.981**	**0.978**
PTB-XL (all)	**0.932**	0.932	**0.936**	**0.933**
PTB-XL (sub)	**0.931**	**0.934**	0.927	**0.931**
PTB-XL (super)	**0.919**	**0.918**	**0.918**	**0.918**
Pediatric ECG interpretation				
ZZU pECG	**0.898**	0.887	**0.891**	0.888
Cardiac structure & function				
EchoNext	**0.824**	**0.826**	**0.824**	**0.825**

JEPA SSL head Tab. 70 compares S4 SSL heads with 1, 3 and 6 layers for JEPA. These models were pretrained on the 106K-sample HEEDB subset.

Table 70: Finetuning performance (macro-AUROC) of JEPA with S4 SSL heads of 1, 3 and 6 layers. The 3- and 6-layer heads use model dimension 256, and the 1-layer head uses no additional linear layer.

	JEPA (Finetuned)		
	1 Layer	**3 Layers**	**6 Layers**
Adult ECG interpretation			
Ningbo	**0.956**	0.950	0.952
CPSC2018	**0.955**	**0.956**	0.949
CPSC-Extra	0.839	**0.854**	**0.847**
Georgia	**0.875**	**0.879**	0.866
Chapman	**0.929**	**0.928**	**0.927**
SPH	**0.965**	**0.958**	**0.953**
PTB-XL (all)	**0.908**	**0.910**	**0.904**
PTB-XL (sub)	**0.891**	**0.890**	**0.880**
PTB-XL (super)	**0.910**	0.905	0.907
Pediatric ECG interpretation			
ZZU pECG	0.827	**0.847**	0.837
Cardiac structure & function			
EchoNext	**0.817**	**0.821**	**0.817**

A.9.4 PRETRAINING DATASET

Tables 71–75 compare pretraining on the 753K-sample HEEDB subset with pretraining on MIMIC-IV-ECG (759K samples), both at 240 Hz, to assess whether the choice of pretraining source affects downstream performance at matched scale.

Table 71: Finetuning performance (macro-AUROC) of data2vec pretrained on HEEDB and MIMIC-IV-ECG.

data2vec (Finetuned)	HEEDB	MIMIC
Adult ECG interpretation		
Ningbo	**0.960**	**0.962**
CPSC2018	**0.960**	**0.961**
CPSC-Extra	**0.853**	**0.855**
Georgia	**0.883**	**0.886**
Chapman	**0.929**	**0.932**
SPH	**0.965**	**0.963**
PTB-XL (all)	**0.908**	**0.910**
PTB-XL (sub)	**0.897**	**0.909**
PTB-XL (super)	**0.912**	**0.913**
Pediatric ECG interpretation		
ZZU pECG	**0.857**	0.848
Cardiac structure & function		
EchoNext	**0.820**	**0.821**

Table 72: Finetuning performance (macro-AUROC) of DinoSR pretrained on HEEDB and MIMIC-IV-ECG.

DinoSR (Finetuned)	HEEDB	MIMIC
Adult ECG interpretation		
Ningbo	0.965	**0.968**
CPSC2018	**0.965**	**0.967**
CPSC-Extra	**0.865**	**0.864**
Georgia	**0.909**	**0.913**
Chapman	**0.948**	**0.951**
SPH	**0.971**	**0.978**
PTB-XL (all)	0.927	**0.939**
PTB-XL (sub)	**0.930**	**0.933**
PTB-XL (super)	**0.920**	**0.921**
Pediatric ECG interpretation		
ZZU pECG	**0.878**	**0.879**
Cardiac structure & function		
EchoNext	**0.826**	**0.830**

Table 73: Finetuning performance (macro-AUROC) of JEPA pretrained on HEEDB and MIMIC-IV-ECG.

JEPA (Finetuned)	HEEDB	MIMIC
Adult ECG interpretation		
Ningbo	**0.962**	0.955
CPSC2018	**0.961**	**0.965**
CPSC-Extra	**0.846**	**0.868**
Georgia	**0.873**	**0.881**
Chapman	**0.934**	**0.934**
SPH	**0.965**	**0.974**
PTB-XL (all)	**0.925**	0.920
PTB-XL (sub)	**0.905**	**0.910**
PTB-XL (super)	**0.912**	**0.913**
Pediatric ECG interpretation		
ZZU pECG	0.845	**0.875**
Cardiac structure & function		
EchoNext	0.817	**0.822**

Table 74: Finetuning performance (macro-AUROC) of CPC pretrained on HEEDB and MIMIC-IV-ECG.

CPC (Finetuned)	HEEDB	MIMIC
Adult ECG interpretation		
Ningbo	**0.971**	**0.973**
CPSC2018	**0.969**	**0.973**
CPSC-Extra	0.882	**0.900**
Georgia	**0.904**	**0.904**
Chapman	**0.958**	**0.959**
SPH	**0.979**	**0.979**
PTB-XL (all)	0.936	**0.943**
PTB-XL (sub)	**0.933**	**0.933**
PTB-XL (super)	**0.921**	**0.920**
Pediatric ECG interpretation		
ZZU pECG	0.876	**0.892**
Cardiac structure & function		
EchoNext	**0.824**	**0.829**

Table 75: Finetuning performance (macro-AUROC) of HuBERT++ pretrained on HEEDB and MIMIC-IV-ECG.

	HuBERT++ (Finetuned)	
	HEEDB	**MIMIC**
Adult ECG interpretation		
Ningbo	0.970	**0.973**
CPSC2018	**0.972**	**0.973**
CPSC-Extra	**0.884**	**0.890**
Georgia	**0.906**	**0.900**
Chapman	**0.955**	**0.960**
SPH	**0.979**	**0.977**
PTB-XL (all)	**0.932**	**0.937**
PTB-XL (sub)	**0.931**	**0.929**
PTB-XL (super)	**0.919**	**0.917**
Pediatric ECG interpretation		
ZZU pECG	**0.898**	**0.901**
Cardiac structure & function		
EchoNext	**0.824**	**0.826**