HuBERT-ECG as a self-supervised foundation model for broad and scalable cardiac applications

Edoardo Coppola[1], Mattia Savardi[2], Mauro Massussi[2,3], Marianna Adamo[2,3], Marco Metra[2,3], Alberto Signoroni[2]

[1] Department of Information Engineering–University of Brescia, Brescia, Italy
[2] Dept. of Medical and Surgical specialties, Radiological Sciences and Public Health–Univ. of Brescia, Brescia, Italy
[3] Cardiac Catheterization Laboratory and Cardiology, ASST Spedali Civili di Brescia, Brescia, Italy

Abstract

Deep learning models have shown remarkable performance in electrocardiogram (ECG) analysis, but the limited availability and size of ECG datasets have constrained their success, resulting in systems that are more task specialists than versatile generalists. To counter this, we introduce HuBERT-ECG, a novel self-supervised foundation ECG model pre-trained on a large and diverse dataset of 9.1 million 12-lead ECGs encompassing 164 cardiovascular conditions. By simply adding a proper output layer, HuBERT-ECG can be fine-tuned for a wide array of downstream tasks, from diagnosing diseases to predicting future cardiovascular events. Across diverse real-world scenarios, HuBERT-ECG achieves AUROCs from 0.843 on small datasets to 0.99 on larger sources. When fine-tuned to detect 164 overlapping conditions simultaneously, our model delivers AUROCs above 0.9 and 0.95 for up to 140 and 97 diseases, respectively. HuBERT-ECG can also predict death events within a 2-year follow-up with AUROCs up to 0.91. We release pre-trained models and code as building baselines.

Introduction

The electrocardiogram (ECG) has long been a cornerstone of cardiovascular diagnostics, serving as a non-invasive, cost-effective and widely available tool for assessing cardiac and noncardiac diseases[1]. Through its recurring waveforms, the ECG captures a unique "language of the heart", encoding both physiological and pathological vital information through distinctive electrophysiological "fingerprints"[2]. This language, although seemingly simple in its structure, reveals complex and meaningful patterns that allow the identification of deviations from expected cardiac functions, providing valuable insights into heart health.

Integrating deep learning (DL) into ECG analysis represents a transformational opportunity to significantly improve its clinical utility[3]. In fact, DL algorithms can transform raw ECG traces into powerful digital biomarkers, enabling early detection, risk stratification, and intervention across a wide range of cardiovascular conditions. Previous studies have demonstrated substantial impact of DL in specific areas of ECG interpretation. For example, DL models have consistently shown solid performance in detecting conditions where the ECG is the gold standard for diagnosis (e.g. atrial fibrillation, tachycardia, and bradycardia), leading to validated algorithms[4]. Conversely, DL applications in the diagnosis of morphological conditions, such as heart failure, pulmonary thromboembolism and aortic stenosis, are less widespread but growing[5]. Although these conditions rely primarily on imaging modalities, previous works demonstrate that DL can successfully use ECG data to predict cardiovascular conditions beyond those traditionally associated with ECG patterns,

including non-cardiac conditions such as liver disease or diabetes[6], providing a valuable complement to conventional imaging approaches[7]. Finally, the use of DL to predict future cardiovascular events (CVEs), such as myocardial infarction and stroke, represents the most ambitious application of this technology. While early results are promising, this area presents greater challenges due to the complexity of predicting outcomes from ECG patterns[8].

From a machine learning perspective, current limitations in ECG dataset size, availability, range of conditions and patient demographics have led to developing ECG models that are more task specialists[9] rather than competent and versatile generalists, traits that typically characterise foundation models[10]. In fact, specialised models often struggle to adapt to new domains or even different data distributions within the same task[11]. For example, an ECG model trained solely to detect atrial fibrillation on a dataset where each recording is labelled as either positive or negative for this condition will be unable to identify other cardiac abnormalities unless retrained on a new dataset. Similarly, if a model is trained exclusively on ECGs from individuals over 65, where atrial fibrillation is prevalent, it might fail to perform well on younger patients. In stark contrast, by pre-training in a self-supervised manner on large and diverse unlabelled datasets, foundation models can learn robust and generalised data representations that are transferable to a wide variety of downstream tasks, requiring only lightweight fine-tuning for specific domains or distributions[12].

Unimodal foundation models have achieved significant success across traditional domains—including natural language processing (e.g., the GPT series[13–16], BERT[17], T5[18]), computer vision (e.g., DINO[19], MAE[20]), and audio processing (e.g., Wav2Vec2[21], HuBERT[22])—by leveraging large-scale self-supervised pre-training on extensive unannotated datasets. More recently, these advancements have also facilitated multimodal learning, where models integrate multiple data modalities to achieve a more holistic understanding of information within data (e.g., Google Gemini[a], GPT-4o[b]). In the medical domain, progress has been slower due to the limited availability of large medical datasets[12,23]. However, this trend has begun to shift, leading to the emergence of foundation medical models such as CONCH[24], Prov-GigaPath[25] and Virchow[26] for computational pathology; KAD[27] and CheXzero[28] for radiology; MedSAM[29], for medical image segmentation; EchoCLIP[30] for echocardiography; and BiomedGPT[31] for various biomedical tasks. Previous works in electrocardiography explored self-supervised learning through contrastive approaches[32–36], but none combined a large-scale, diverse pre-training *with* a through evaluation across many datasets and tasks—both crucial requirements for truly foundational ECG models. Among more recent efforts, HeartBEiT[37] and KED[38] stand out. HeartBEiT is notable for its large dataset, but it lacks diversity across sources and a comprehensive generalizability evaluation, testing only three tasks. Conversely, KED, through text-guided pre-training, exhibits zero-shot capabilities and a richer evaluation but does not incorporate a large-scale, diverse pre-training, and evaluations are performed only on rhythm and form conditions. A truly general-purpose ECG foundation model should therefore be (1) versatile and adaptable to various use cases, especially where data is scarce; (2) efficient, requiring minimal fine-tuning to deliver strong performance; (3) designed to support research in underexplored areas.

In this work, inspired by HuBERT[22] architecture, we present HuBERT-ECG, a new ECG foundation model designed for broad transferability across a wide range of tasks, from diagnosing diseases to predicting future CVEs. Unlike most existing ECG models, HuBERT-ECG is pre-trained through self-supervision on an extensive dataset of 9.1 million ECGs from a large and diverse population across four countries. Noteworthy, over 2.4 million ECGs in our dataset are annotated with one or more conditions from a comprehensive set of 164 diagnoses, allowing HuBERT-ECG to learn and generalise over a wide spectrum of pathologies. These diagnostic labels are derived exclusively from

[a] https://gemini.google.com/
[b] https://openai.com/index/hello-gpt-4o/

the ECG signal—without integration of clinical, biomarker, or imaging data—and reflect how ECGs are interpreted in current practice by automated systems or ECG-reading physicians. When fine-tuned and evaluated on 16 datasets and their aggregation, simulating real-world scenarios and use cases of varying complexity, HuBERT-ECG demonstrates both efficiency and accuracy, showing the potential to serve as a strong foundation for addressing three crucial types of tasks: (1) detecting conditions where the ECG is the primary diagnostic tool (e.g., atrial fibrillation, bundle branch blocks); (2) supporting the identification of conditions where the ECG plays a non-primary but potentially informative role (e.g., heart failure, myocardial infarction); and (3) predicting future cardiovascular events (e.g., death, hospitalization).

Finally, to support ongoing research on cardiovascular diseases and accelerate innovation, we release HuBERT-ECG models and code to the community. We make available HuBERT-ECG in three model sizes (or configurations)—SMALL (30 millions parameters), BASE (93 millions parameters) and LARGE (188 million parameters)—to both meet the computational constraints of future users and address a broad range of applications where some model sizes may perform better than others. By sharing insights from comprehensive evaluations on many different downstream datasets, we aim to facilitate HuBERT-ECG successful adoption and customisation for research purposes.

Results

ECG Representation Learning through Self-supervised Pre-training

To develop a robust and generalisable ECG foundation model, we propose HuBERT-ECG, a self-supervised foundation model inspired by HuBERT[22], originally designed for speech signals. HuBERT-ECG learns meaningful ECG representations through an iterative process combining (i) feature extraction, (ii) clustering and (iii) masked modelling. Specifically, descriptive ECG features are extracted from ECG segments and then clustered by a clustering model. Subsequently, during pre-training, embedded versions of ECG segments are masked and HuBERT-ECG is trained to predict which cluster they belong to. By repeating this process across multiple pre-training iterations, and progressively refining the feature extraction, HuBERT-ECG improves its ability to capture the underlying structure of ECG signals, ultimately achieving generalisable ECG representations.

We pre-train HuBERT-ECG on an extensive and diverse dataset of 9.1 million ECGs from four countries (Fig. 1a-d, Fig. 2a, "Methods" – *Data and Preprocessing*). In the first iteration, prior to masked modelling, we extract and cluster Mel Frequency Cepstral Coefficients[39] (MFCCs) from raw ECG segments. MFCCs, known for capturing key frequency components of the signal, highlight the essential structure of the ECG. In the second iteration, we extract semantically higher features from the model's 8th layer (latent representations) and refine the clusters. This iterative refinement allows HuBERT-ECG to learn a deeper understanding of ECG patterns and structure through masked modelling. HuBERT-ECG architecture and pre-training specifics are detailed in "Methods" - *HuBERT-ECG Architecture and Theoretical Framework*. Feature extraction and clustering are illustrated in "Methods" – *Unsupervised label discovery*. Specifications and implementation details are provided in "Methods" – *Implementations* and "Methods" – *Self-supervised pre-training*.

To further evaluate the adaptability of our approach and offer models capable of addressing diverse needs after fine-tuning, we also pre-train SMALL and LARGE model sizes in a single iteration using more refined latent representations.

Supervised fine-tuning on downstream data and evaluation

To comprehensively assess HuBERT-ECG's performance and adaptability across various clinical scenarios, we compile 16 downstream datasets for fine-tuning and evaluation. These datasets include: the labelled partition of Ribeiro[40] (also known as CODE); CPSC and CPSC-Extra[41]; PTB[42]; PTB-XL[43] which, with 6 different sets of conditions, gives rise to 6 different datasets; the publicly available partition of Georgia[44]; Chapman-Shaoxing[45] (Chapman); Ningbo First Hospital[46] (Ningbo); part of the dataset used in Tianchi Arrhythmia Competition[c] (Hefei); Shandong Provincial Hospital[47] (SPH); and SaMi-Trop[48]. In addition, to create a hard benchmark that mirrors real-world cardiological practice—managing a wide range of cardiovascular conditions in patients with diverse health profiles—we harmonise labels from each dataset and merge all these sources into a new, comprehensive dataset named *Cardio-Learning*. This dataset includes 2.4 million ECGs from four countries, each labelled with one or more conditions from a comprehensive, highly imbalanced set of 164 conditions. After fine-tuning, which requires only a small fraction of pre-training time, HuBERT-ECG is evaluated on hold-out test sets of 17 datasets (16 individual + Cardio-Learning). The obtained performance is then compared with that achieved by training the same model from scratch, i.e., using randomly initialised weights. This comparison is common practice in self-supervised learning literature and reveals how self-supervised pre-training enhances both learning efficiency and generalisation. Where applicable, we also benchmark against existing state-of-the-art methods. To ensure consistent evaluation, address label distribution disparities, and employ a robust, threshold-independent metric, we report model performance using the Area Under the Receiver Operating Characteristic Curve (AUROC). The fine-tuning procedure is detailed in "Methods" - *Supervised Fine-tuning*. The following results, presented across diverse application scenarios, are meant to demonstrate the robustness and transferability of HuBERT-ECG's self-supervised ECG representations.

HuBERT-ECG is efficient on small-sampled datasets. In diagnostic scenarios with small datasets, such as PTB (515 instances, 17 conditions), CPSC (6,878 ECGs, 9 conditions) and CPSC-Extra (3,453 ECGs, 52 conditions), HuBERT-ECG consistently shows solid performance and exhibits distinct macro-trends. On PTB (Fig. 3a), the SMALL and BASE configurations perform comparably well when fine-tuned, achieving AUROCs of 0.843±0.029 and 0.848±0.03, respectively. On CPSC (Fig. 3b), the SMALL configuration achieves an AUROC of 0.945±0.002 and is slightly surpassed by the BASE (0.949±0.002) and LARGE (0.949±0.006) models. On this dataset we outperform contrastive learning works (e.g., Liu et al.[34,35] and Tian et al.[38] achieved AUROCs of 0.932, 0.915, and 0.910, respectively) while remaining competitive with models specifically pre-trained and optimised for arrhythmia detection only (e.g., Na et al.[49] with 0.98). On CPSC-Extra (Fig. 3c), a more challenging dataset, HuBERT-ECG SMALL performs best, reaching an AUROC of 0.894±0.076. This result illustrates the advantages of smaller architectures in such intricate, low-data contexts, where larger model configurations might overfit and perform less effectively (BASE 0.756±0.02; LARGE 0.774±0.023). Noteworthy, fine-tuned models also provide 0.03-0.058 AUROC improvements over their randomly initialised counterparts, with the less relevant exception of the BASE configuration on CPSC-Extra, demonstrating the clear benefit of pre-training even on smaller datasets.

HuBERT-ECG remains efficient in increasingly difficult contexts. Diagnostic complexity extends beyond simple metrics like dataset size and the number of classes. It is significantly influenced by the inherent subtlety and difficulty of identifying certain conditions, low per-class sample sizes, and highly skewed label distributions. All these factors can exacerbate the learning challenge and are well-exemplified by the Georgia and Chapman datasets, each featuring over 10,000 ECGs labelled with 62 and 54 conditions unevenly distributed. In these complex diagnostic scenarios, fine-tuning

[c] https://tianchi.aliyun.com/competition/entrance/231754/introduction

HuBERT-ECG consistently leads to better performance than training from scratch, yielding improvements of 0.065-0.078 on Georgia and 0.039-0.061 on Chapman. Specifically, HuBERT-ECG BASE excels on Georgia with an AUROC of 0.832±0.007 (Fig. 3d). For comparison, Tian et al.[38] greatly facilitated the learning problem by considering only 20 classes and reached an AUROC of 0.9. When evaluated on the same classes, HuBERT-ECG BASE reaches strong results (0.853±0.003), despite being fine-tuned to detect all 62 conditions simultaneously. On Chapman (Fig. 3e), all model configurations perform similarly, with the SMALL one topping (0.859±0.008). To address the label distribution skewness, Chapman's authors suggested grouping conditions into four macro-categories, highly simplifying the diagnostic task. Following this approach, Liu et al.[34] and Kiyasseh et al.[32] reported AUROCs of 0.997 and 0.932. When evaluated on these suggested four macro-categories—despite being fine-tuned to recognise the original 54 conditions—HuBERT-ECG achieves strong AUROC (SMALL 0.975±0.002; BASE 0.975±0.009; LARGE 0.984±0.001), showcasing robustness across varying levels of diagnostic granularity and without simplifying the underlying problem during training.

HuBERT-ECG enables accurate diagnostics in large-scale, complicated scenarios. Trained on extensive datasets, diagnostic foundation models are expected to perform robustly in large-scale, complicated scenarios, where they should identify a disease among many others, even accounting for significant inter-patient variations that can substantially affect diagnostic accuracy. To explore HuBERT-ECG's transferability to these scenarios, we leverage large datasets with few to tens of possible conditions including Ribeiro, PTB-XL, Hefei, SPH and Ningbo.
On Ribeiro, a large-scale dataset (2.3 million ECGs, 6 conditions), every HuBERT-ECG configuration performs excellently regardless of weights initialization. The BASE model performs best when fine-tuned (0.9989) and surpasses the trained-from-scratch counterpart (0.9984) whereas the SMALL and LARGE configurations perform similarly (Fig. 3f). Notably, as shown in Table 2, HuBERT-ECG BASE outperforms the model from Ribeiro et al.[40] in terms of sensitivity, specificity, AUROC and area under precision and recall curve (AUPRC). The PTB-XL[43] dataset comprises 21,837 ECGs annotated with 71 possible labels covering diagnostic, form and rhythm statements. The authors state that 44 diagnostic conditions can be aggregated into 23 diagnostic subclasses and 5 broader diagnostic superclasses, while there are 12 possible rhythm statements and 19 different form abnormalities. Therefore, one can evaluate a model on 6 different datasets, referred to as PTB-XL *All*, *Form*, *Rhythm*, *Diagnostic* (*Diag.*), *Diag. Subclass*, *Diag. Superclass*, that differ from each other in the reported labels. On these sources, fine-tuned models surpass randomly initialised counterparts for almost each model configuration (Fig. 3g-l). Particularly, HuBERT-ECG BASE achieves an AUROC greater than 0.90 on 5 out of 6 datasets. Noteworthy, these results are better or comparable to those reported in state-of-the-art works from the literature (Table 3), where models are specifically optimised for PTB-XL benchmark. This robust performance is significant given that PTB-XL is a minimal fraction of our pre-training dataset, underscoring HuBERT-ECG's broad transferability across diverse clinical scenarios after fine-tuning. On Hefei (20,036 samples, 29 conditions), all fine-tuned models surpass the respective randomly initialised counterparts, with the BASE configuration performing best with an AUROC of 0.9661±0.0025 (Fig. 3m). On SPH[47], a collection of 25,770 ECGs annotated with 44 diagnostic statements, all our models achieve comparable results, with the LARGE model size reaching the highest AUROC (0.943) (Fig. 3n). To the best of our knowledge, we are the first to address this dataset in its full complexity. Finally, on Ningbo[46] (34,905 ECGs, 76 co-existing conditions), fine-tuned models generally outperform the trained-from-scratch counterparts, with the BASE size achieving the highest AUROC of 0.9439±0.0053 (Fig. 3o).

HuBERT-ECG can be a good predictor of future cardiovascular events. Predicting future CVEs is an exceptionally challenging task, especially given the scarcity of publicly available datasets. In this context, the SaMi-Trop[48] dataset offers an opportunity to assess HuBERT-ECG's predictive capabilities. It includes 1632 ECGs from patients with Chagas disease monitored during a 2-year

follow-up period, 268 of which are normal recordings and 104 are marked as "death events". On this dataset, the BASE and LARGE model configurations benefit more than the SMALL one from pre-trained weights. In fact, although competitive, the fine-tuned SMALL version reaches an AUROC of 0.8705±0.0279 while the randomly initialised counterpart achieves a score of 0.8883±0.0308. In contrast, the BASE (0.8627±0.0422) and LARGE (0.8622±0.0451) configurations perform marginally better than their trained-from-scratch counterparts. We attribute this behaviour to the low number of events within SaMi-Trop, where a single misclassification can result in significant changes. For comparison purposes, we benchmark against the state-of-the-art work by Ferreira et al.[50]. While we use only the raw ECG traces, they used a combination of handcrafted ECG features, sociodemographic variables and self-reported symptoms to fit a Random Forest estimator, and evaluate their model in terms of G_{mean} score, where $G_{mean} = \sqrt{sensitivity \cdot specificity}$. Fixing the same classification threshold, our best model surpasses theirs, reaching a G_{mean} of 0.78 against one of 0.77, while, in terms of AUROC, no comparison is possible due to the lack of predictions and code.

HuBERT-ECG navigates extremely complex scenarios. In real-world clinical practice, cardiologists are expected to diagnose a wide spectrum of heart-related conditions, adapting to the nuances of each patient's health profile, regardless of how frequently a condition appears. To emulate this complexity, we employ the newly assembled, comprehensive *Cardio-Learning* dataset, which contains more than 2.4 million ECGs and 164 overlapping conditions, where the ECG is either the primary diagnostic tool, a non-primary supportive diagnostic tool, or used to estimate the risk of future CVEs. The fine-tuned HuBERT-ECG BASE provides a macro-averaged AUROC of 0.8517±0.0193, while HuBERT-ECG SMALL and LARGE follow with scores of 0.8453±0.0058 and 0.8403±0.0378, respectively (Fig. 4a-c). When the conditions are grouped according to the ECG diagnostic role—primary diagnostic tool or informative support for diagnosis—the BASE configuration achieves averaged AUROCs of 0.84079±0.019 and 0.8777±0.0197, respectively. Most interestingly, without any additional recordings and by leveraging inter-relationship among conditions, HuBERT-ECG raises AUROC values in predicting future deaths for SaMi-Trop patients to 0.9096±0.0782, significantly surpassing the best model presented for this task in the previous paragraph.

Benchmarking HuBERT-ECG against reported human performance. To contextualize the clinical relevance of HuBERT-ECG, we retrospectively compare the performance of its fine-tuned version on the Ribeiro dataset with that of human readers reported in Ribeiro et al.[40]. That study involved two 4th-year cardiology residents, two 3rd-year emergency residents, and two 5th-year medical students, whose sensitivity and specificity values were averaged across the six diagnostic categories. Using the same dataset and label set, and the balanced accuracy as aggregation metric, we find that always at least two fine-tuned HuBERT-ECG configurations outperform medical residents/students on all the six diagnostic categories, and all configurations always surpass at least two groups of residents/students (see Table 4). However, it is important to note that this is not a direct head-to-head comparison, and no statistical testing can be performed due to the unavailability of individual predictions from the human annotators. Nonetheless, the results suggest that HuBERT-ECG, when fine-tuned, is capable of matching or surpassing human-level performance on key diagnostic tasks, highlighting its potential in downstream clinical applications.

Fine-tuning HuBERT-ECG requires less compute time than training from scratch. While acquiring sufficient labelled data is already a known bottleneck, compute time represents an equally critical constraint in deep learning workflows, particularly in domains where large datasets and complex models necessitate resource-intensive experimentation, hyperparameter optimization, and ablation studies. As demonstrated in Table 5, fine-tuning the pre-trained HuBERT-ECG is markedly more efficient across most scenarios. The aforementioned steps, not included in the presented results, are already performed during the self-supervised pre-training of HuBERT-ECG and can therefore be

skipped by future users, offering a clear advantage over workflows that require developing models entirely from the ground up.

Discussion.

In this paper, we introduce HuBERT-ECG, a novel foundation model for comprehensive ECG analysis available in three scalable configurations (SMALL, BASE, and LARGE) to address diverse clinical needs through a targeted fine-tuning. Pre-trained via self-supervision on an extensive dataset of 9.1 million ECGs from four countries and diverse populations, HuBERT-ECG demonstrates strong performance in identifying 164 cardiovascular conditions. This is rigorously validated across 16 individual datasets and their aggregation (Cardio-Learning), encompassing small to very large ECG collections, annotated with few to many overlapping conditions. Across all evaluations, the three model configurations consistently achieve AUROCs above 0.9 for an average of 136, 138, and 135 conditions and surpass 0.95 AUROC for an average of 95, 94, and 94 conditions, respectively (Fig. 5a). These results underscore the generalisability of HuBERT-ECG's self-supervised representations to highly different clinical contexts. Looking ahead, our performance analysis aims not only to shed light on the data requirements for effective fine-tuning but also to guide performance optimisation on new ECG-based datasets.

In diagnostic contexts with small-sampled datasets, which, requiring limited collection efforts, are common, HuBERT-ECG demonstrates both efficiency and accuracy, even when the number of conditions outpace the number of examples, often resulting in highly imbalanced label distributions. In these contexts, the SMALL and BASE configurations are preferable, demonstrating superior generalisation capabilities, whereas the LARGE configuration is less recommended. As the number of patients, ECGs and conditions increases, simulating more realistic diagnostic environments, HuBERT-ECG delivers remarkable results regardless of diagnostic granularity, albeit requiring adequate data collection efforts for fine-tuning. Furthermore, scaling up model size proves advantageous on larger sources, although smaller configurations remain competitive, thus providing desirable flexibility for complicated diagnostic fine-tuning. Notably, on Ribeiro, the largest dataset we use, HuBERT-ECG showcases the capacity to achieve excellent precision-sensitivity and specificity-sensitivity trade-offs (Table 2), matching or surpassing reported human performance (Table 4). Although Fig. 3f shows that the benefits of pre-training are marginal at the extremely high performance levels obtained from such large datasets with only a few conditions, this does not diminish the value of our approach. In fact, such extensive data collection and labelling efforts—combined with only six conditions—are uncommon in real-world settings. Nonetheless, when benchmarking our performance against prior studies, we demonstrate our straightforward fine-tuning consistently leads to high performance, frequently approaching or surpassing that of models specifically developed for individual datasets or condition groups (e.g., arrhythmias). In addition, aiming at learning efficiency, we show that fine-tuning the pre-trained HuBERT-ECG generally requires less compute time than training from scratch, which also demands costly hyper-parameter tuning and ablation studies. This underscores the generalisability of HuBERT-ECG's representations, especially when data is overall scarce, certain classes are underrepresented, or when compute time is a considerable constraint.

Furthermore, HuBERT-ECG exhibits robust performance in highly complex clinical scenarios exemplified by Cardio-Learning, which aggregates ECGs from diverse sources and countries without task-specific separation and creates a realistic representation of heterogeneous real-world data. In this challenging context, HuBERT-ECG effectively identifies conditions for which the ECG serves as either a primary or informative support signal for diagnosis. Crucially, fine-tuning on Cardio-Learning enables HuBERT-ECG to leverage inter-relationships between co-occurring conditions,

leading to substantial improvements in AUROC for label prediction—even for conditions that appear only in single subsets. For example, when fine-tuning the BASE model on PTB, the AUROCs for Coronary Heart Disease (CHD) and Heart Failure (HF) are 0.7656 and 0.7514, respectively. Remarkably, when fine-tuned on Cardio-Learning, these scores increase significantly to 0.9488±0.0399 for CHD and 0.9674±0.0566 for HF (Fig. 5b).
Similarly, in the highly challenging task of predicting future mortality events, HuBERT-ECG achieves mean AUROCs of 0.8784, 0.9096, and 0.8857 for the SMALL, BASE, and LARGE configurations, respectively, with average improvements of 1.31, 4.69, and 2.35 percentage points (Fig. 5c). These findings strongly suggest that training a model to recognize a broad spectrum of ECG-derived diagnostic labels can significant enhance its performance in complex prediction tasks, including those where the ECG plays a non-primary but potentially informative role. Importantly, although the model is trained on labels from ECG readings, its rich representations can be further fine-tuned using high-quality clinical endpoints (e.g. electronic health records, imaging and/or genomic data), potentially unlocking even greater performance on tasks beyond traditional ECG applications.

Another relevant finding from HuBERT-ECG's label-wise performance on Cardio-Learning is its low prediction variance (Fig. 4). This suggests that merged labels within Cardio-Learning refer to consistent concepts, thus (i) excluding major label inconsistencies, and (ii) supporting both our label harmonisation process and HuBERT-ECG's robustness to label noise. Furthermore, a visual inspection of the fine-tuned HuBERT-ECG attention mechanism confirms its ability to attend to clinically relevant ECG segments rather than arbitrary noise (see Supplementary Information – Sec. 3).

While HuBERT-ECG has been rigorously tested across diverse cardiovascular conditions, our findings suggest its considerable potential to extend its application toward more personalised clinical insights. These include predicting patient response to therapies or identifying non-cardiac conditions through ECG analysis, paving the way for truly personalised cardiovascular care. For example, in the management of heart failure, HuBERT-ECG could be fine-tuned to predict patient responsiveness to resynchronisation therapy, thereby optimising patient selection criteria[51]. This capability would facilitate a more tailored therapeutic approach, aimed not only at predicting adverse events but also at enhancing outcomes, reducing hospitalisations, and improving quality of life. Such advancements could fundamentally transform cardiovascular precision medicine, with the use of foundation ECG models for helping clinicians tailor therapies with greater precision[52].

Despite the highly encouraging results, our approach has certain limitations. First, although the data collection spans more countries than in previous studies, it does not include large and highly populated regions (e.g., Africa and India), which we aim to incorporate in future versions. Second, the scarcity of accessible ECG datasets for predicting future CVEs partially limits our ability to fully assess the HuBERT-ECG's potential in this still underexplored domain. Broader access to suitable datasets for testing the model in this challenging field is desirable to accelerate research. Third, direct and fair comparisons with previous systems, both deep-learning-based and rule-based, remain challenging. Many existing works focus on limited subsets of available datasets (often reducing the actual number of conditions), do not open-source their implementations or evaluation datasets, or do not follow standard practices in metric computation, hampering reproducibility and equitable benchmarking. Lastly, we observe that while some conditions are better classified in Cardio-Learning than in their original datasets, others show lower performance. This discrepancy may stem from erroneous labels or incomplete assignments. For example, in PTB-XL, a major label source, key terms in reports are extracted to retrospectively infer the physician's diagnostic confidence, with labels assigned even at low-confidence levels. Conversely, other datasets may rely on more robust labelling protocols, potentially resulting in identical ECGs from different sources receiving different

diagnoses. Additionally, less experienced physicians may overlook certain conditions, while overly confident ones might exclude others. These inaccuracies can reduce the data efficiency, requiring more data to compensate for their effects.

Methods

Data and Preprocessing

While many ECG-focused studies addressing general clinical questions primarily rely on datasets from the Physionet Challenges[53,54], overlooking other substantial and valid sources, HuBERT-ECG self-supervised pre-training leverages both public and access-on-demand 12-lead ECG datasets. These include the labelled and unlabelled partitions from Ribeiro[40] (except the predefined, publicly available test-set), CPSC and CPSC-Extra[41], PTB[42] and PTB-XL[43], the publicly available partition from Georgia[44], Chapman-Shaoxing[45] and Ningbo First Hospital[46], the partition from the Tianchi Arrhythmia Competition[d], Shandong Provincial Hospital[47], and MIMIC-IV ECG[55]. For supervised fine-tuning and evaluation we reuse all the aforementioned sources and include the new SaMi-Trop[48] dataset. We instead exclude the unlabelled partition of Ribeiro and MIMIC-IV for this phase. Although some datasets appear during both self-supervised pre-training and supervised fine-tuning, this a widely accepted practice in developing foundation models that undergo a self-supervised pre-training. Notable examples include those by Caron et al.[19,56] in computer vision, Hsu et al.[22] and Baevski et al.[21] in speech representation learning, Azizi et al.[57], Pai et al.[58] and Haghighi et al.[59] in various types of medical image analyses, Ritcher et al.[60] in cell genomics, and ultimately Mehary & Strodthoff[36] and Tang et al.[61] in the ECG domain. In addition, a recent survey[62] on the impact of self-supervised pre-training for diagnostic tasks involving many different types of medical data, points out that pre-training with self-supervision and fine-tuning a model on the same data brings no tangible benefit. This is due to the lack of any information leakage, as no clinical/downstream labels or task-specific signal is exposed during the self-supervised pre-training. In our case, pre-training target variables are also pseudo-labels, discovered in an unsupervised manner using feature descriptors (MFCCs and latent representations) that do not carry any disease-specific, clinical, or dataset-specific information. For self-supervised pre-training, all labels and clinical annotations are removed from the pre-training datasets, resulting in an unlabelled dataset of 9.1 million ECGs (Fig. 2a–Pretrain). The effectiveness of the data selection in capturing the diversity of ECG signals is illustrated through a UMAP[63] projection of the ECG embeddings (Fig. 1e). To assess HuBERT-ECG's downstream utility on clinical tasks and datasets of varying complexity, we fine-tune it on each collected dataset[e] and test it on hold-out partitions of such data sources. Additionally, we create a challenging new dataset, which we name *Cardio-Learning*, by merging all the above sources into a new one comprising over 2.4 million ECGs with 164 potentially co-occurring conditions. These conditions can be grouped into three categories based on the diagnostic role of the ECG, as shown in Fig. 1f. During the creation of Cardio-Learning, to ensure label consistency across different sources, diagnostic labels are harmonised independently by three expert cardiologists. First, diagnostic labels that appear under slightly varying nomenclature are consolidated using unique terms. For example, all instances of atrial fibrillation are uniformly labelled as "AF" instead of "AFIB" or "AF_" (see Supplementary Table 1). Second, diagnostic concepts that appear under multiple synonyms are standardised to a single diagnostic category. For instance, following established practices[53,54], ECGs labelled as premature atrial contraction (PAC) or supraventricular premature beats (SVPB) are standardised to PAC|SVPB, as both conditions refer to the same concept. Third, hierarchically broader diagnostic categories are assigned to ECGs already labelled with more specific ones (e.g., Left Bundle Branch Block ECGs also receives the Bundle Branch Block label). However, we avoid two actions: (i) we do not infer

[d] https://tianchi.aliyun.com/competition/entrance/231754/introduction
[e] PTB-XL generates 6 different datasets that differ from each other in the presented conditions.

more specific conditions from more general labels unless explicitly confirmed in the original annotation (e.g., Bundle Branch Block ECGs do not receive the Left Bundle Branch Block label); (ii) if a constituent dataset includes both a more specific and a more general condition, but explicitly does not assign the general condition to ECGs labelled with the specific one, we respect this distinction and do not add it.

Remarkably, as shown in Figs. 1b and 1d, the ECGs used in this study were obtained from patients exhibiting a broad age distribution and diverse geographical origins spanning four countries. To the best of our knowledge, this constitutes one of the largest and most heterogenous dataset ever assembled to date in terms of the number of conditions, demographics, and geographic origin of the individuals.

For pre-processing, consistent with the approach of Natarajan et al.[64], a finite impulse response bandpass filter is applied to exclude frequencies outside the [0.05, 47] Hz range, which is reported to encompass the dominant components of P waves, T waves and QRS complexes[65]. Subsequently, we investigate the impact of the sampling rate, which has no standard value and regulates the degree of dilution of the information content, on both the upstream and downstream performance. Our experimental findings indicate that resampling the ECGs at 100 Hz provides a suitable trade-off between downstream performance and training time, while preserving all the relevant physiological information content according to the Nyquist-Shannon theorem (Supplementary Information–Sec. 1.2). Following this, the ECG signal is rescaled to the [-1, 1] range, mirroring the procedure employed by Natarajan et al.[64]. Finally, unlike other works, we use 5-second 12-lead ECGs instead of 10-second recordings, hence halving memory occupation and accelerating both training and inference. In addition, the selected temporal and spectral parameters align with those derived in Mehari & Strodthoff[66].

HuBERT-ECG Architecture and Theoretical Framework

A schematic representation of the HuBERT-ECG architecture, its pre-training phases and its fine-tuning are shown in Fig. 2b.

Discovering fine-grained pseudo-labels for ECG fragments. In order to produce ground-truth pseudo-labels for the self-supervised pre-training, standard 12-lead ECGs are first flattened into one-dimensional signals and then split into non-overlapping fragments. This fragmentation is necessary to frame short portions of the ECG signal from which feature descriptors can be extracted to fit a clustering model. Its purpose is to discover and provide the fragments with cluster assignments (i.e., pseudo-labels) that are finer than the coarse and noisy ECG feature-based ones used as vocabulary for ECGBERT[67]. Conceptually, let $X = [x_1, x_2, \dots, x_N]$ be a flattened 12-lead ECG composed of N non-overlapping fragments x_i, where $i = 1, 2, \dots N$. Let also $F = [F_1, F_2, \dots, F_N] = [[f_1, f_2, \dots f_d]_1, [f_1, f_2, \dots f_d]_2, \dots, [f_1, f_2, \dots f_d]_N]$ be the sequence of d-dimensional features F_i extracted from each fragment, and K a clustering model fitted on F that finds C different clusters. Then, for each index $i = 1, 2, \dots, N$, under the hypothesis that F_i is a good descriptor of x_i, $K(F_i) = z_i \in \{1, 2, \dots, C\}$ is the pseudo-label assigned to x_i and $Z = [z_1, z_2, \dots, z_N]$ is the sequence of these assignments for each ECG fragment. For the sake of both terminological flexibility and clarity, the terms *pseudo-labels* and *cluster assignments* will be used interchangeably.

Representation learning by predicting pseudo-labels for masked embeddings. HuBERT-ECG operates as a masked prediction model, henceforth denoted as h. When the flattened 12-lead ECG is fed into h, the signal is initially processed by a convolutional waveform embedder. This component captures local contextual information and generates $E = [e_1, e_2, \dots, e_N]$, a sequence of continuous ECG embeddings. Positional encodings are then integrated into the embeddings to provide the model with information about the original sequential position of each lead segment within the flattened ECG

signal. This mechanism allows the model to account for inter-lead relationships that might be less evident after the 12-lead ECG is transformed into a one-dimensional sequence. Subsequently, a set of random indices $M = \{j \mid j \in \{1, 2, \dots, N\}\}$, where $|M| < N$, is generated to determine which embeddings are to be masked, i.e., replaced by a special learnable embedding e_{MASK}. After masking, a Transformer encoder[68], which learns global contextual information, consumes the resulting masked sequence of embeddings E' and produces the output sequence of encodings $O = [o_1, o_2, \dots, o_N]$. Eventually, a probability distribution over the C possible pseudo-labels is constructed for all the N indices: $p_h(\cdot \mid E', i)\ \forall i \in \{1, 2, \dots, N\}$. This distribution is then used in a standard cross-entropy loss that, however, considers only the indices in M: $L(h; E, M, Z) = \sum_{j \in M} log(p_h(z_j \mid E', j))$ [f]. By training the model h to predict the cluster assignments of the masked embeddings $E'_j, \forall j \in M$, which correspond to fragments of the input ECG not seen by the encoder, the model is forced to learn the most informative representations from the visible segments.

Multi-task learning via cluster ensembles. To increase the granularity of the representations being learned during the pre-training, one can use pseudo-labels generated by an ensemble of clustering models characterised by an increasing number of clusters. This approach constitutes a multi-task learning framework, where a new task is introduced with each additional clustering model. Within this framework, each embedding is associated with multiple pseudo-labels, one from each clustering model in the ensemble. The rationale behind this design is that, while a single clustering model may yield imprecise or coarse cluster assignments, an ensemble of models with an increasing number of clusters can mitigate the introduction of noisy targets and provide valuable complementary information to the model. Denoting the number of clustering models composing the ensemble as Γ, which corresponds to the number of tasks being addressed, the loss function can be reformulated as $L(h;\ E, M, \{Z^\gamma\}) = \sum_{\gamma \in \Gamma} \sum_{j \in M} log(p_h^\gamma(z_j^\gamma \mid E', j))$, where Z^γ represents the sequence of cluster assignments generated by the γ-th clustering model and $\{Z^\gamma\}$ is the set of cardinality Γ comprising all such sequences.

Refining cluster assignments. Analogous to the work of Hsu et al.[22], the quality of the learned representations can be further enhanced by generating more refined cluster assignments for the ECG fragments that are to be used in subsequent pre-training iterations. Importantly, this "finer" mapping does not refer to a temporal refinement, but rather to a more nuanced clustering of the features extracted from the ECG fragments. To achieve it, latent representations are extracted from intermediate layers of the partially pre-trained model h and then clustered. Consequently, while the cluster assignments remain aligned with the original fragments in the input X, their improved quality following this refinement can exert a positive impact on downstream task performance.

Implementation

While the design of HuBERT-ECG mostly aligns with that proposed by Hsu et al.[22], we implement two modifications related to the initial convolutional embedder and the masking strategy. First, since ECGs sampling rate is lower than that of audio signals, there is no need of high-stride convolutions with long filters. A shallower convolutional block with shorter filters proves equally effective. In our model, the convolutional embedder, architecturally tuned to balance embedding information granularity and computational efficiency, generates embeddings at a temporal resolution 0.64 seconds from a flattened 5-second 12-lead ECG sampled at 100 Hz, resulting in 93 embeddings. Second, instead of randomly selecting p% of the embeddings as starting points for masked spans, we opt to mask only the selected embeddings without spanning over adjacent ones. This approach avoids the imprecision introduced by overlapping spans, which complicated determining the exact number of

[f] We express cross-entropy loss as a simple sum of log probabilities, instead of the traditional sum over $p(\cdot)log(p(\cdot))$, because, in masked cross-entropy, $p(\cdot)$ terms are implicitly equal to 1, given the one-hot encoding of labels.

masked embeddings. Our method is equivalent to constructing singleton spans and provides more accurate control over the masking process, enabling us to tune the optimal value of p to use during the self-supervised pre-training (Supplementary Information–Sec. 2.1). Additionally, masking individual embeddings rather than spans has also demonstrated efficacy in the work of Hu et al.[69]

Following the masking operation, the Transformer encoder consumes the masked sequence of embeddings and produces O, the output sequence of encodings. The probability distribution over the cluster assignments from a generic clustering model of the ensemble is parameterised by a look-up embedding matrix A^{γ} with shape $C^{\gamma} \times W$ according to the following formula:

$$p_h^{\gamma}(k \mid E', i) = \frac{exp(cos_sim(B^{\gamma} \cdot o_i,\ A_k^{\gamma})/\tau)}{\sum_{c'=1}^{C^{\gamma}} exp(cos_sim(B^{\gamma} \cdot o_{c'}, A_{c'}^{\gamma})/\tau)}$$

where C^{γ} denotes the number of clusters found by the γ-th clustering model, B^{γ} is a projection matrix used to align the dimensionality of O with the embedding dimension W, A_k^{γ} is the k-th row of the look-up embedding matrix A^{γ} that corresponds to the k-th cluster assignment, $cos_sim(\cdot,\cdot)$ is the cosine similarity between two vectors, and τ is the temperature that scales the logits. The superscript γ denotes the γ-th task being addressed when using an ensemble of Γ clustering models. Specifically, there are as many projection and look-up embedding matrices as there are tasks in the ensemble.

Given the extensive and heterogenous nature of our dataset, we propose HuBERT-ECG in three model sizes: SMALL, BASE and LARGE. As we scale the size, we maintain the same convolutional embedder but increase the encoder depth and width, as well as the label embedding dimension W. Table 1 provides a detailed summary of the architectural specifications for the three model configurations. Following pre-training, to fine-tune HuBERT-ECG for specific downstream tasks, we delete the look-up embedding and projection matrices and attach a randomly initialised linear layer atop the encoder to map the pooled output sequence into logits. Specifically, we exploit the Pytorch implementation provided by Hugging Face and adapt its source code to our specific requirements. All models are trained on a computational node equipped with NVIDIA A100 GPUs.

Unsupervised pseudo-label discovery

HuBERT-ECG BASE undergoes two consecutive pre-training iterations. To generate target pseudo-labels for the first one, that do not include any diagnostic or clinical information, we perform a series of k-means clustering experiments (see Supplementary Information-Sec. 1.1) and find that 100 clusters of feature descriptors consisting of 39-dimensional vectors of Mel Frequency Cepstral Coefficients[39] (MFCCs) (13 base coefficients augmented with their first and second-order derivatives) to be a good and efficient starting point for pre-training. Furthermore, although we obtain only marginal benefit from using a cluster ensemble (see Supplementary Information–Sec. 2.2), when using multiple clusterings for pseudo-label discovery, the additional pseudo-labels are generated by two MFCC-based k-means models with 200 and 300 clusters, respectively.

For the second iteration, to produce enhanced and more refined pseudo-labels, we run the k-means algorithm again, this time increasing the number of clusters. Specifically, we use 500 clusters of latent representations extracted from the 8th encoding layer of HuBERT-ECG BASE after the first iteration. Upon completion of the second iteration, we pre-train the SMALL and LARGE model configurations for one iteration using pseudo-labels generated by clustering representations extracted from 9th encoding layer into 500 and 1000 clusters, respectively (Supplementary Information–Sec. 1.4). Due to memory constraints, we cannot load the entire dataset into memory and, therefore, we opt for a batched implementation of k-means wherein batches of MFCC descriptors, or latent representations, are yielded and the centroid positions are updated incrementally. We set a batch size of 9300, derived from yielding 93 descriptors from 100 ECGs, and use k-means++[70] with 20 random restarts for a better centroid initialisation.

Self-supervised pre-training

For each pre-training iteration we reserve 90% of the unlabelled dataset for training and the remaining 10% for internal validation. We set a batch size of 448 instances and find empirically the optimal masking percentage $p = 33\%$ (see Supplementary Information-Sec. 2.1) The first iteration consists of 80k training steps, while the second iteration counts 770k steps. The third iteration, instead, consists of 362.5k and 422.5k steps for the SMALL and LARGE model sizes, respectively. We use Adam optimiser with β = (0.9,0.98), an initial weight decay of 0.01 and a dropout probability of 0.1 along with a learning rate scheduler that ramps up for the first 8% of the training steps and then decays linearly to zero. Peak learning rates are set to 5e-5/5e-5/2.5e-5 for the BASE, SMALL and LARGE model configurations, respectively. In addition, we find benefits from exploiting a *dynamic regularisation* during pre-training (Supplementary Information–Sec. 1.3). This technique "penalises" the model by incrementally increasing its dropout probability and weight decay if performance does not improve on the internal validation set for *penalty-count* consecutive times. Otherwise, if performance improves on the internal validation set, with respect to the best validation loss or accuracy, the model is "rewarded" by reversing the effects of the last penalty, i.e., decreasing its dropout probability and weight decay. It is important to note that the initial weight decay and dropout probability are the minimum achievable values, while *penalty-count* emerges as a significant hyperparameter to be tuned according to the frequency with which validations are performed. During the first and third iterations, with randomly initialised models, we set the *penalty-count* to 4 and perform an internal validation every 2500 steps. For the second iteration, the internal validation is performed every 5000 steps, while maintaining the same value of *penalty-count*.

Supervised fine-tuning

In order to assess the robustness of HuBERT-ECG's self-supervised representations on clinically relevant datasets simulating real-world scenarios, we fine-tune the model on each labelled dataset considered in this study. These datasets differ greatly from each other in the number of instances (Fig. 1a), patients' age distribution (Fig. 1d), and spectrum of possible conditions, each belonging to one of the three classes outlined in Fig. 1f.

When training-validation-test splits are predefined, or when at least the test set is known and defined a-priori, as in the case of PTB-XL, SPH or Ribeiro, we adhere to the original split to allow fair comparisons with previous works. However, for datasets where such a split is not available or applicable (i.e., Hefei, Ningbo, Chapman, CPSC, Georgia), we first extract a fixed hold-out test set, containing 10% of the dataset instances, using a stratified sampling to ensure that all the dataset diagnostic classes are represented. We then perform a 4-fold cross-validation on the remaining instances to tune the hyperparameters of the four models, selecting the best candidate from each fold for the evaluation on the hold-out test set. Finally, we average performance metrics that the four best candidates obtained during inference. Instead, for datasets with an extremely low cardinality compared to the model size, such as PTB, CPSC-Extra and SaMi-Trop, we split the dataset four times in a stratified manner into four <*training, test*> folds. Then, for each fold, lacking a validation set, we skip any hyperparameter tuning and train four models until either a near-zero training error is reached or a minimum improvement delta is not achieved. We then run inference directly on the corresponding hold-out test set with the last model checkpoint. When fine-tuning on PTB, CPSC-Extra and SaMi-Trop, to prevent overfitting on these small datasets, we set a minimal early stopping's patience of 2 to allow for transitory fluctuations in the training loss descent and set a the minimum improvement delta to 1×10^{-3} for both the loss function and evaluation metrics. Lastly, to construct the Cardio-Learning dataset, we merge the training, validation, and test sets from each constituent dataset to form the corresponding overall training, validation, and test partitions, ensuring the prevention of any data leakage between these partitions. We perform a 4-fold cross-validation even on Cardio-Learning.

Prior to any fine-tuning, we analyse the label distribution of each dataset and exclude the instances labelled with conditions that occur only once. We do so because these conditions are either unlearnable or untestable, as their solitary instances cannot be included in both the training and test sets. However, if a condition occurs twice in the dataset, we assign one instance to the training set and one to the test set, allowing the model to attempt to learn the condition while providing a way to assess its generalisability within the limits of this setup. Eventually, when a condition occurs three times, we place one example per split. During each fine-tuning procedure, we optimise a binary cross-entropy loss function but refer to the macro-averaged multi-label AUROC on the validation set for checkpointing and model selection. The model checkpoint that yields the highest AUROC is used for inference.

To fine-tune HuBERT-ECG we follow a straightforward protocol: we attach atop the pre-trained model a randomly initialised linear layer and fine-tune all the weights of the resulting architecture, except those of the convolutional embedder. In contrast to Hsu et al.[22], we do not use the *freezing-steps* hyperparameter, as our experiments do not demonstrate any efficacy in maintaining the encoder's parameters fixed while training only the final linear layer. Drawing inspiration from Devlin et al.[17], we reduce the batch size to 64 instances and the learning rate to 1e-5. Also, to gain finer control over the identification of a good model candidate, we validate our models every 50 or 500 steps, contingent on the dataset size, hence more frequently than during pre-training. Due to the high number of trainable parameters and the limited size of most datasets, HuBERT-ECG shows a tendency to overfitting but we observe no benefits from using either a strong dropout (up to 0.5) or a high weight decay (up to 0.1), nor from freezing some encoding layers. However, experiments with LayerDrop[71] ([0.0, 0.1, 0.15, 0.2]) indicate its potential to mitigate validation loss divergence and metrics degradation. In addition, we note improvements when using a time-aligned random crop as data augmentation technique, a strategy that we also replicate at test time when we take the most confident prediction among those made on multiple crops of the same instance. In summary, our findings suggest that the best performance is obtained when fine-tuning the entire Transformer encoder, zeroing the dropout probability, maintaining the weight decay at 0.01, and sweeping over the LayerDrop probability. We track all experimental runs using Weights and Biases[g].

Evaluation Metrics

To evaluate the downstream performance of HuBERT-ECG across the diverse datasets considered in this study, we primarily rely on the multi-label AUROC metric, which is widely used in the literature to address the challenges of multi-label classification and class imbalance[36,49,69,72]. The multi-label AUROC generalizes the concept of AUROC, originally developed for binary and multi-class problems, to multi-label scenarios where individual instances can be labelled with more than one class. In this context, the AUROC is computed independently for each condition, treating the prediction of each label as a separate binary classification task. The resulting AUROC scores for individual labels are then averaged to obtain the macro-averaged AUROC, providing a holistic perspective on the model's performance across all diagnostic labels. The AUROC represents the area under the Receiver Operating Characteristic (ROC) curve, which graphically plots the true positive rate (TPR, or sensitivity) against the false positive rate (FPR, or $1 - \text{specificity}$) at varying classification thresholds. Mathematically, TPR and FPR are defined as follows:

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = 1 - \frac{TN}{TN + FP}$$

[g] https://wandb.ai/site

where TP, FN, TN, and FP denote true positives, false negatives, true negatives, and false positives, respectively. The AUROC value quantifies the model's ability to discriminate between instances with or without a specific condition. A ideal classifier achieves an AUROC of 1, with a corresponding ROC curve that reaches the top-left corner of the FPR/TPR plane (coordinate (0,1)). AUROC values closer to 1.0 indicate superior model performance, while a value of 0.5 suggests no predictive capabilities no better than random chance.

On Ribeiro, the largest dataset used in our study, we are able to compute AUPRC as additional metric. This metric is based on the concept of Precision, which quantifies the proportion of correctly predicted positive instances among all instances predicted as positive for a specific diagnostic category and classification threshold. Precision is mathematically defined as follows:

$$Precision = \frac{TP}{TP + FP}$$

High precision values indicate that the model minimises overpredicting positive conditions, thereby reducing the occurrence of false positive alarms. The AUPRC is the area under the precision recall (i.e., sensitivity) curve and evaluates the trade-off between precision and sensitivity across varying classification thresholds. This metric reflects the model's ability to identify true positive samples without being tricked by confounding factors into predicting false positive instances.

Data availability

The PTB, CPSC, CPSC-Extra, Georgia, Chapman, PTB-XL, and Ningbo datasets are publicly accessible through PhysioNet at the following URL: https://physionet.org/content/challenge-2021/1.0.3/training/ningbo/#files-panel.
The Hefei dataset used in this study can be downloaded from the Tianchi platform at: https://tianchi.aliyun.com/competition/entrance/231754/information/.
The SPH dataset is available through Figshare at the following Springer Nature repository: https://springernature.figshare.com/collections/A_large-scale_multi-label_12-lead_electrocardiogram_database_with_standardized_diagnostic_statements/5779802/1.
The SaMi-Trop dataset can be downloaded from Zenodo at: https://zenodo.org/records/4905618.
The Ribeiro dataset is accessible for scientific research upon request to the respective owner. However, its test set is publicly available from Zenodo at: https://zenodo.org/records/3765780.

Code availability

The full pipeline utilised in this study is available at the following GitHub repository: https://github.com/Edoar-do/HuBERT-ECG. This repository includes: (1) code for data preprocessing, from raw data management to the creation of the train-validation-test partitions used in our research; (2) scripts for replicating the pre-training, fine-tuning and inference of every model developed across all datasets; and (3) code for reproducing our performance validation. Eventually, to both facilitate reproducibility and enable rapid implementation, we provide (4) pre-trained weights for all model configurations on Hugging Face[h]. (5) Access to model weights fine-tuned on Cardio-Learning is on-demand. This model is intended as a more disease-oriented baseline for further fine-tuning.

References

1. Smilowitz, N. R. & Berger, J. S. Perioperative Cardiovascular Risk Assessment and Management for Noncardiac Surgery. JAMA 324, 279 (2020).

[h] https://huggingface.co/Edoardo-BS

2. van de Vegte, Y. J. et al. Genetic insights into resting heart rate and its role in cardiovascular disease. Nat Commun 14, 4646 (2023).
3. Friedman, P. A. The Electrocardiogram at 100 Years: History and Future. Circulation 149, 411–413 (2024).
4. Ahsan, M. M. & Siddique, Z. Machine learning-based heart disease diagnosis: A systematic literature review. Artif Intell Med 128, 102289 (2022).
5. Kalmady, S. V. et al. Development and validation of machine learning algorithms based on electrocardiograms for cardiovascular diagnoses at the population level. NPJ Digit Med 7, 133 (2024).
6. Sun, W. et al. ECG for high-throughput screening of multiple diseases: Proof-of-concept using multi-diagnosis deep learning from population-based datasets. Preprint at https://arxiv.org/pdf/2210.06291. (2022).
7. Goto, S. et al. Multinational Federated Learning Approach to Train ECG and Echocardiogram Models for Hypertrophic Cardiomyopathy Detection. Circulation 146, 755–769 (2022).
8. Kim, M. et al. Deep learning for predicting rehospitalization in acute heart failure: Model foundation and external validation. ESC Heart Fail (2024).
9. Musa, N. et al. A systematic review and Meta-data analysis on the applications of Deep Learning in Electrocardiogram. J Ambient Intell Humaniz Comput 14, 9677–9750 (2023).
10. Moor, M. et al. Foundation models for generalist medical artificial intelligence. Nature 616, 259–265 (2023).
11. Finlayson, S. G. et al. The Clinician and Dataset Shift in Artificial Intelligence. New England Journal of Medicine 385, 283–286 (2021).
12. Bommasani, R. et al. On the Opportunities and Risks of Foundation Models. Preprint at https://arxiv.org/abs/2108.07258. (2021).
13. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. Improving Language Understanding by Generative Pre-Training. (2018).
14. Radford, A. et al. Language Models are Unsupervised Multitask Learners. OpenAI blog 1, 9 (2019).
15. Brown, T. B. et al. Language Models are Few-Shot Learners. Adv Neural Inf Process Syst (2020).
16. OpenAI et al. GPT-4 Technical Report. Preprint at https://arxiv.org/abs/2303.08774. (2023).
17. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. in Proceedings of the 2019 Conference of the North 4171–4186 (Association for Computational Linguistics, Stroudsburg, PA, USA, 2019).
18. Raffel, C. et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research (2020).
19. Caron, M. et al. Emerging Properties in Self-Supervised Vision Transformers. in Proceedings of the IEEE/CVF international conference on computer vision (2021).
20. He, K. et al. Masked Autoencoders Are Scalable Vision Learners. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 16000–16009 (2022).
21. Baevski, A., Zhou, Y., Mohamed, A. & Auli, M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. Advances in neural information systems 33, 12449–12460 (2020).
22. Hsu, W.-N. et al. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. IEEE/ACM Trans Audio Speech Lang Process 29, 3451–3460 (2021).
23. Schäfer, R. et al. Overcoming data scarcity in biomedical imaging with a foundational multi-task model. Nat Comput Sci 4, 495–509 (2024).
24. Lu, M. Y. et al. A visual-language foundation model for computational pathology. Nat Med 30, 863–874 (2024).
25. Xu, H. et al. A whole-slide foundation model for digital pathology from real-world data. Nature 630, 181–188 (2024).

26. Vorontsov, E. et al. A foundation model for clinical-grade computational pathology and rare cancers detection. Nat Med 30, 2924–2935 (2024).
27. Zhang, X., Wu, C., Zhang, Y., Xie, W. & Wang, Y. Knowledge-enhanced visual-language pre-training on chest radiology images. Nat Commun 14, 4542 (2023).
28. Tiu, E. et al. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. Nat Biomed Eng 6, 1399–1406 (2022).
29. Ma, J. et al. Segment anything in medical images. Nat Commun 15, 654 (2024).
30. Christensen, M., Vukadinovic, M., Yuan, N. & Ouyang, D. Vision–language foundation model for echocardiogram interpretation. Nat Med 30, 1481–1488 (2024).
31. Zhang, K. et al. A generalist vision–language foundation model for diverse biomedical tasks. Nat Med (2024).
32. Kiyasseh, D., Tingting Zhu & David A. Clifton. Clocs: Contrastive learning of cardiac signals across space, time, and patients. in Proceedings of the International Conference on Machine Learning 5606–5615 (2021).
33. Oh, J., Chung, H., Kwon, J., Hong, D. & Choi, E. Lead-agnostic self-supervised learning for local and global representations of electrocardiograms. in Conference on Health, Inference and Learning 338–353 (2022).
34. Liu, W. et al. Dense lead contrast for self-supervised representation learning of multilead electrocardiograms. Inf Sci (N Y) 634, 189–205 (2023).
35. Liu, W. et al. A joint cross-dimensional contrastive learning framework for 12-lead ECGs and its heterogeneous deployment on SoC. Comput Biol Med 152, 106390 (2023).
36. Mehari, T. & Strodthoff, N. Self-supervised representation learning from 12-lead ECG data. Comput Biol Med 141, 105114 (2022).
37. Vaid, A. et al. A foundational vision transformer improves diagnostic performance for electrocardiograms. NPJ Digit Med 6, 108 (2023).
38. Tian, Y. et al. Foundation model of ECG diagnosis: Diagnostics and explanations of any form and rhythm on ECG. Cell Rep Med 5, 101875 (2024).
39. Xu, M. et al. HMM-Based Audio Keyword Generation. in 566–574 (2004).
40. Ribeiro, A. H. et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. Nat Commun 11, 1760 (2020).
41. Liu, F. et al. An Open Access Database for Evaluating the Algorithms of Electrocardiogram Rhythm and Morphology Abnormality Detection. J Med Imaging Health Inform 8, 1368–1373 (2018).
42. Bousseljot, R., Kreiseler, D. & Schnabel, A. Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet. Biomedizinische Technik/Biomedical Engineering 317–318 (2009).
43. Wagner, P. et al. PTB-XL, a large publicly available electrocardiography dataset. Sci Data 7, 154 (2020).
44. Goldberger, A. L. et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 101, (2000).
45. Zheng, J. et al. A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. Sci Data 7, 48 (2020).
46. Zheng, J. et al. Optimal Multi-Stage Arrhythmia Classification Approach. Sci Rep 10, 2898 (2020).
47. Liu, H. et al. A large-scale multi-label 12-lead electrocardiogram database with standardized diagnostic statements. Sci Data 9, 272 (2022).
48. Cardoso, C. S. et al. Longitudinal study of patients with chronic Chagas cardiomyopathy in Brazil (SaMi-Trop project): a cohort profile. BMJ Open 6, e011181 (2016).
49. Na, Y., Park, M., Tae, Y. & Joo, S. Guiding Masked Representation Learning to Capture Spatio-Temporal Relationship of Electrocardiogram. in Proceedings of the International Conference on Learning Representations (2024).

50. Ferreira, A. M. et al. Two-year death prediction models among patients with Chagas Disease using machine learning-based methods. PLoS Negl Trop Dis 16, e0010356 (2022).
51. Wouters, P. C. et al. Electrocardiogram-based deep learning improves outcome prediction following cardiac resynchronization therapy. Eur Heart J 44, 680–692 (2023).
52. Leclercq, C. et al. Wearables, telemedicine, and artificial intelligence in arrhythmias and heart failure: Proceedings of the European Society of Cardiology Cardiovascular Round Table. EP Europace 24, 1372–1383 (2022).
53. Perez Alday, E. A. et al. Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. Physiol Meas 41, 124003 (2020).
54. Reyna, M. A. et al. Will Two Do? Varying Dimensions in Electrocardiography: The PhysioNet/Computing in Cardiology Challenge 2021. in 2021 Computing in Cardiology (CinC) 1–4 (IEEE, 2021).
55. Gow, B. et al. MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset. (2023).
56. Caron, M., Bojanowski, P., Joulin, A. & Douze, M. Deep Clustering for Unsupervised Learning of Visual Features. in Proceedings of the European Conference on Computer Vision (ECCV) (2018).
57. Azizi, S. et al. Big Self-Supervised Models Advance Medical Image Classifications. in Proceedings of the IEE/CVF international conference on computer vision 3478–3488 (2021).
58. Pai, S. et al. Foundation model for cancer imaging biomarkers. Nat Mach Intell 6, 354–367 (2024).
59. Haghighi, F., Taher, M. R. H., Gotway, M. B. & Liang, Ji. DiRA: Discriminative, Restorative, and Adversarial Learning for Self-supervised Medical Image Analysis. in Proceedings of the IEE/CVF international conference on computer vision 20824–20834 (2022).
60. Richter, T., Bahrami, M., Xia, Y., Fischer, D. S. & Theis, F. J. Delineating the effective use of self-supervised learning in single-cell genomics. Nat Mach Intell 7, 68–78 (2024).
61. Tang, J., Xia, T., Lu, Y., Mascolo, C. & Saeed, A. Electrocardiogram Report Generation and Question Answering via Retrieval-Augmented Self-Supervised Modeling. in ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1–5 (IEEE, 2025).
62. VanBerlo, B., Hoey, J. & Wong, A. A survey of the impact of self-supervised pretraining for diagnostic tasks in medical X-ray, CT, MRI, and ultrasound. BMC Med Imaging 24, 79 (2024).
63. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. J Open Source Softw 3, 861 (2018).
64. Natarajan, A. et al. A Wide and Deep Transformer Neural Network for 12-Lead ECG Classification. in Computing in Cardiology (2020).
65. Hong, S., Zhang, W., Sun, C., Zhou, Y. & Li, H. Practical Lessons on 12-Lead ECG Classification: Meta-Analysis of Methods From PhysioNet/Computing in Cardiology Challenge 2020. Front Physiol 12, (2022).
66. Mehari, T. & Strodthoff, N. Towards Quantitative Precision for ECG Analysis: Leveraging State Space Models, Self-Supervision and Patient Metadata. IEEE J Biomed Health Inform 27, 5326–5334 (2023).
67. Choi, S. et al. ECGBERT: Understanding Hidden Language of ECGs with Self-Supervised Representation Learning. Preprint at https://arxiv.org/abs/2306.06340. (2023).
68. Vaswani, A. et al. Attention Is All You Need. Adv Neural Inf Process Syst (2017).
69. Hu, R., Chen, J. & Zhou, L. Spatiotemporal self-supervised representation learning from multi-lead ECG signals. Biomed Signal Process Control 84, 104772 (2023).
70. Arthur, D. & Vassilvitskii, S. k-means++: the advantages of careful seeding. in Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms 1027–1035 (ACM-SIAM Symposium on Discrete Algorithms, 2007).
71. Fan, A., Grave, E. & Joulin, A. Reducing Transformer Depth on Demand with Structured Dropout. in Proceedings of the International Conference on Learning Representations (2019).
72. Strodthoff, N., Wagner, P., Schaeffter, T. & Samek, W. Deep Learning for ECG Analysis: Benchmarks and Insights from PTB-XL. IEEE J Biomed Health Inform 25, 1519–1528 (2021).

Acknowledgements

This work has been partly funded by 1) Regione Lombardia, Italy, through the initiative "Programme of measures for economic recovery: development of new cooperation agreements with universities for research, innovation and technology transfer"-DGR n. XI/4445/2021; 2) European Union-Next Generation EU, through the Italian Ministry of Research PRIN 2022, project n. 2022A49KR3 "QT-SEED Quality-of-life Technological and Societal Exploitation of ECG Diagnostics".
We also thank the authors of the CODE study[40] for giving us access to the Ribeiro dataset.

Author Contribution

Edoardo Coppola: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing-original draft, Writing-review & editing. **Mattia Savardi:** Conceptualization, Investigation, Methodology, Project administration, Supervision, Visualization, Writing-original draft, Writing-review & editing. **Mauro Massussi:** Conceptualization, Investigation, Writing-original draft, Writing-review & editing. **Marianna Adamo:** Investigation, Writing-review & editing. **Marco Metra:** Investigation, Writing-review & editing. **Alberto Signoroni:** Conceptualization, Funding acquisition, Investigation, Project administration, Resources, Supervision, Writing-original draft, Writing-review & editing.

Competing Interest

The authors declare no competing financial or non-financial interests.

Ethics statement

For the Ribeiro dataset the CODE Study was approved by the Research Ethics Committee of the Universidade Federal de Minas Gerais, protocol 49368496317.7.0000.5149.

		SMALL	BASE	LARGE
Convolutional embedder	strides	4, 2, 2, 2, 2		
	kernels width	10, 3, 3, 2, 2		
	channels	512		
Transformer encoder	layers	8	12	16
	internal dimension	512	768	960
	feed-forward dimension	2048	3072	3840
	attention heads	8	8	12
Embedding dimension W		256	256	512
Number of parameters		30M	93M	188M

Table 1 | HuBERT-ECG architecture summary.

Condition	Models	Sensitivity	Specificity	AUROC	AUPRC
AF	Ribeiro et al.	0.769	1.000	0.885	0.773
	HuBERT-ECG SMALL	1.0000 (CI:0.9999–1.0000)	0.9754 (CI:0.9656-0.9853)	1.0000 (CI:0.9986-1.0000)	0.9774 (CI:0.9278-1.0000)
	HuBERT-ECG BASE	1.0000 (CI:0.999–1.000)	1.0000 (CI:0.9999–1.0000)	**1.0000** (CI:0.9999–1.0000)	**1.0000** (CI:0.9999–1.0000)
	HuBERT-ECG LARGE	1.0000 (CI:0.9999–1.000)	0.9748 (CI:0.9644–0.9853)	0.9976 (CI:0.9954–0.9999)	0.9240 (CI:0.8535–0.9945)
IAVB	Ribeiro et al.	0.929	0.995	0.962	0.807
	HuBERT-ECG SMALL	0.9290 (CI:0.8580-1.0000)	0.9531 (CI:0.9387-0.9675)	0.9890 (CI:0.9817-0.9963)	0.8309 (CI:0.7366-0.9252)
	HuBERT-ECG BASE	0.7500 (CI:0.6071-0.8929)	0.9981 (CI:0.9962-1.0000)	**0.9977** (CI:0.9961-0.9994)	**0.9560** (CI:0.9264-0.9856)
	HuBERT-ECG LARGE	1.0000 (CI:0.9990–1.0000)	0.9485 (CI:0.9337–0.9637)	0.9960 (CI:0.9935–0.9985)	0.8750 (CI:0.7929–0.9571)
CLBBB|LBBB	Ribeiro et al.	1.000	1.000	1.000	1.000
	HuBERT-ECG SMALL	1.0000 (CI:0.9999–1.0000)	0.9940 (CI:0.9893-0.9987)	1.0000 (CI:0.9999–1.0000)	0.9950 (CI:0.9837-1.0000)
	HuBERT-ECG BASE	0.9000 (CI:0.8333-0.9667)	1.0000 (CI:0.9999–1.0000)	**1.0000** (CI:0.9999–1.0000)	**1.0000** (CI:0.9999–1.0000)
	HuBERT-ECG LARGE	0.9330 (CI:0.8993–0.9667)	0.9950 (CI:0.9913–0.9987)	0.9983 (CI:0.9971–0.9995)	0.9760 (CI:0.9609–0.9911)
CRBBB|RBBB	Ribeiro et al.	1.000	0.995	0.997	0.895
	HuBERT-ECG SMALL	1.0000 (CI:0.9999–1.000)	0.9850 (CI:0.9776-0.9924)	0.9987 (CI:0.9976-0.9998)	0.9800 (CI:0.9616-0.9984)
	HuBERT-ECG BASE	0.9400 (CI:0.8529-1.0000)	0.9943 (CI:0.9899-0.9987)	**0.9986** (CI:0.9974-0.9998)	**0.9740** (CI:0.9513-0.9967)
	HuBERT-ECG LARGE	0.9710 (CI:0.9420–1.0000)	0.9874 (CI:0.9798–0.9950)	0.9977 (CI:0.9957–0.9998)	0.9690 (CI:0.9424–0.9956)
SBRAD	Ribeiro et al.	0.938	0.996	0.967	0.782
	HuBERT-ECG SMALL	1.0000 (CI:0.9999 –1.0000)	0.9808 (CI:0.9716-0.9901)	0.9973 (CI:0.9948-0.9998)	0.8801 (CI:0.7719-0.9883)
	HuBERT-ECG BASE	1.0000 (CI:0.9999–1.0000)	0.9919 (CI:0.9864-0.9975)	**0.9980** (CI:0.9954-1.0000)	**0.8860** (CI:0.7504-1.0000)
	HuBERT-ECG LARGE	1.0000 (CI:0.9999–1.0000)	0.9796 (CI:0.9704–0.9889)	0.9980 (CI:0.9962–0.9998)	0.8950 (CI:0.8014–0.9886)
STACH	Ribeiro et al.	0.937	0.997	0.985	0.923
	HuBERT-ECG SMALL	0.9730 (CI:0.9729–0.9731)	0.9816 (CI:0.9722-0.9911)	0.9969 (CI:0.9946-0.9992)	0.9350 (CI:0.8807-0.9893)
	HuBERT-ECG BASE	0.9730 (CI:0.9729–0.9731)	0.9950 (CI:0.9913-0.9987)	**0.9990** (CI:0.9983-0.9997)	**0.9640** (CI:0.9342-0.9938)
	HuBERT-ECG LARGE	0.9730 (CI:0.9728–0.9732)	0.9829 (CI:0.9734–0.9924)	0.9957 (CI:0.9939–0.9975)	0.9480 (CI:0.9149–0.9811)
MACRO-AVERAGED	Ribeiro et al.	0.935	0.997	0.966	0.863
	HuBERT-ECG SMALL	0.9806 (CI:0.9657–0.9955)	0.9784 (CI:0.9740–0.9828)	0.9974 (CI:0.9964–0.9984)	0.9330 (CI:0.9045–0.9615)
	HuBERT-ECG BASE	0.9255 (CI:0.8943-0.9567)	0.9969 (CI:0.9954-0.9983)	**0.9989** (CI:0.9983-0.9995)	**0.9630** (CI:0.9403-0.9857)
	HuBERT-ECG LARGE	0.9770 (CI:0.9641–0.9899)	0.9782 (CI:0.9740–0.9825)	0.9978 (CI:0.9973–0.9983)	0.9310 (CI:0.9013–0.9607)

Table 2 | Fine-tuned HuBERT-ECG performance on Ribeiro benchmarked against that from Ribeiro et al.[40] according to multiple metrics. Threshold-based metrics are obtained setting the threshold to 0.5, as done by Ribeiro et al., for a fair comparison. Performance uncertainty is reported using 95% confidence intervals obtained via empirical bootstrapping using 1000 bootstrap iterations. Label abbreviations and corresponding diagnosis can be found in Supplementary Table 1.

Models	PTB-XL All	PTB-XL Form	PTB-XL Rhythm	PTB-XL Diag.	PTB-XL Diag. Subclass	PTB-XL Diag. Superclass
	Macro-averaged AUROC					
HuBERT-ECG SMALL	0.900±0.006	0.838±0.012	0.941±0.013	0.919±0.005	0.914±0.005	0.907±0.007
HuBERT-ECG BASE	0.903±0.002	0.855±0.012	0.953±0.010	0.917±0.005	0.917±0.006	0.911±0.008
HuBERT-ECG LARGE	0.896±0.003	0.828±0.018	0.935±0.016	0.905±0.002	0.919±0.006	0.903±0.007
Hu et al.[69(*)]	**0.947**	~ 0.895	**~ 0.980**	**~ 0.950**	**~ 0.940**	**~ 0.938**
Liu et al.[34(*)]	N.A.	N.A.	N.A.	N.A.	N.A.	0.912
Liu et al.[35(*)]	N.A.	N.A.	N.A.	N.A.	N.A.	0.877
Mehari & Strodthoff [36(**)]	0.942	N.A.	N.A.	N.A.	N.A.	N.A.
Strodthoff et al.[72(***)]	0.925	**0.896**	0.957	0.937	0.929	0.928
Na et al.[49(***)]	N.A.	N.A.	N.A.	N.A.	N.A.	0.933

Table 3 | Fine-tuned HuBERT-ECG performance on PTB-XL datasets against the state-of-the-art. Performance uncertainty is reported using 95% confidence intervals obtained through empirical bootstrapping using 1000 bootstrap iterations. "N.A." stands for "Not Available". (*) Code not available. Results preceded by '~' symbol are estimated by looking at paper graphs. (**) Based on the available code, the AUROC is calculated on individual batches and then averaged—introducing batch-size dependency—with custom handling for any NaN values that arise due to class imbalances within batches. (***) Based on the available code, the AUROC is computed after aggregating predictions and corresponding ground-truth labels across all batches—a standard approach we also adopt.

Condition	Predictor	Sensitivity	Specificity	Balanced Accuracy
AF	HuBERT-ECG SMALL	1.0000 (CI:0.9999–1.0000)	0.9754 (CI:0.9656-0.9853)	*0.9877 (CI:0.9827-0.9926)*
	HuBERT-ECG BASE	1.0000 (CI:0.999–1.000)	1.0000 (CI:0.9999–1.0000)	**1.0000 (CI:0.9999–1.0000)**
	HuBERT-ECG LARGE	1.0000 (CI:0.9999–1.000)	0.9748 (CI:0.9644–0.9853)	0.9874 (CI:0.9821-0.9926)
	Cardiology residents	0.7690	0.9960	0.8825
	Emergency residents	0.6150	0.9980	0.8065
	Medical students	0.9230	0.9890	0.9560
IAVB	HuBERT-ECG SMALL	0.9290 (CI:0.8580-1.0000)	0.9531 (CI:0.9387-0.9675)	0.9411 (CI:0.8983-0.9837)
	HuBERT-ECG BASE	0.7500 (CI:0.6071-0.8929)	0.9981 (CI:0.9962-1.0000)	0.8741 (CI:0.8016-0.9464)
	HuBERT-ECG LARGE	1.0000 (CI:0.9990–1.0000)	0.9485 (CI:0.9337–0.9637)	**0.97425 (CI:0.9663-0.9818)**
	Cardiology residents	0.6790	0.9970	0.838
	Emergency residents	0.8210	0.9840	0.9025
	Medical students	0.929	0.9790	*0.9540*
CLBBB|LBBB	HuBERT-ECG SMALL	1.0000 (CI:0.9999–1.0000)	0.9940 (CI:0.9893-0.9987)	**0.997 (CI:0.9946-0.9993)**
	HuBERT-ECG BASE	0.9000 (CI:0.8333-0.9667)	1.0000 (CI:0.9999–1.0000)	0.9500 (CI:0.9166-0.9833)
	HuBERT-ECG LARGE	0.9330 (CI:0.8993–0.9667)	0.9950 (CI:0.9913–0.9987)	*0.9640 (CI:0.9453-0.9827)*
	Cardiology residents	0.9000	1.0000	0.9500
	Emergency residents	0.8670	0.9990	0.9330
	Medical students	0.9000	0.9970	0.9485
CRBBB|RBBB	HuBERT-ECG SMALL	1.0000 (CI:0.9999–1.000)	0.9850 (CI:0.9776-0.9924)	0.9672 (CI:0.9887-0.9962)
	HuBERT-ECG BASE	0.9400 (CI:0.8529-1.0000)	0.9943 (CI:0.9899-0.9987)	0.9792 (CI:0.9214-0.9993)
	HuBERT-ECG LARGE	0.9710 (CI:0.9420–1.0000)	0.9874 (CI:0.9798–0.9950)	*0.9825 (CI:0.9609-0.9975)*
	Cardiology residents	0.9710	0.9940	0.8820
	Emergency residents	0.7650	0.9990	0.9685
	Medical students	0.9410	0.9960	**0.9904**
SBRAD	HuBERT-ECG SMALL	1.0000 (CI:0.9999 –1.0000)	0.9808 (CI:0.9716-0.9901)	*0.9960 (CI:0.9857-0.9950)*
	HuBERT-ECG BASE	1.0000 (CI:0.9999–1.0000)	0.9919 (CI:0.9864-0.9975)	**0.9990 (CI:0.9931-0.9987)**
	HuBERT-ECG LARGE	1.0000 (CI:0.9999–1.0000)	0.9796 (CI:0.9704–0.9889)	0.9670 (CI:0.9851-0.9944)
	Cardiology residents	0.9380	0.9960	0.9355
	Emergency residents	0.8750	0.9960	0.8725
	Medical students	0.7500	0.9950	0.9773
STACH	HuBERT-ECG SMALL	0.9730 (CI:0.9729–0.9731)	0.9816 (CI:0.9722-0.9911)	0.9773 (CI:0.9725-0.9821)
	HuBERT-ECG BASE	0.9730 (CI:0.9729–0.9731)	0.9950 (CI:0.9913-0.9987)	**0.9840 (CI:0.9821-0.9859)**
	HuBERT-ECG LARGE	0.9730 (CI:0.9728–0.9732)	0.9829 (CI:0.9734–0.9924)	*0.9790 (CI:0.9731-0.9828)*
	Cardiology residents	0.8110	0.9990	0.9050
	Emergency residents	0.9460	0.9970	0.9715
	Medical students	0.8380	0.9960	0.9170

Table 4 | Fine-tuned HuBERT-ECG performance on Ribeiro benchmarked against two 4th year cardiology residents (Cardiology residents), two 3rd year emergency residents (Emergency residents), two 5th year medical students (Medical students) involved in the study by Ribeiro et al.[40]. Both residents' and students' performance are averages, as reported in the original study. Label abbreviations and corresponding diagnoses can be found in Supplementary Table 1.

Dataset	Model size		
	HuBERT-ECG SMALL (FT / TfS)	HuBERT-ECG BASE (FT / TfS)	HuBERT-ECG LARGE (FT / TfS)
PTB	**3.3k** / 3.5k	2.3k / 2.15k	1.8k / 1.5k
Cpsc	**5.8k** / 6.7k	4.8k / 4.4k	3.2k / 3.1k
Cpsc-Extra	9.2k / 8.15k	6.7k / 3.9k	5.3k / 2.8k
Georgia	**5.2k** / 8.13k	**5.4k** / 6.3k	**2.8k** / 4.8k
Chapman	**5k** / 9.22k	**4.4k** / 5.3k	**3.3k** / 3.5k
Ribeiro	**55k** / 60k	**45k** / 55k	**50k** / 40k
PTB-XL All	**12.5k** / 27k	**9k** / 19k	**8.5k** / 19k
PTB-XL Diag.	**11k** / 21k	**9.5k** / 24.5k	**7.5k** / 20k
PTB-XL Diag. Superclass	**7k** / 18.5k	**6k** / 15.5k	**4.5k** / 16.5k
PTB-XL Diag. Subclass	**8k** / 12.5k	**6.5k** / 24.5k	8k / 8k
PTB-XL Rhythm	**7.5k** / 19.5k	**8.5k** / 25.5k	**5.5k** / 20k
PTB-XL Form	**6.5k** / 17k	**6.5k** / 9k	**5.5k** / 9k
Hefei	**10.1k** / 18.2k	**5.8k** / 15.2k	**5.2k** / 13.7k
SPH	**10.5k** / 27k	**5.5k** / 14.5k	**7.5k** / 21k
Ningbo	**15.8k** / 35.5k	**12.6k** / 21.7k	**10.1k** / 20k
SaMi-Trop	0.9k / 0.8k	1.1k / 0.6k	0.8k / 0.5k
Cardio-Learning	**30k** / 30.1k	**27.8k** / 42.1k	32.5k / 27.4k

Table 5 | Comparison between fine-tuning (FT) and training-from-scratch time (TfS), on average, across datasets and model sizes. Time is expressed as number of training steps.

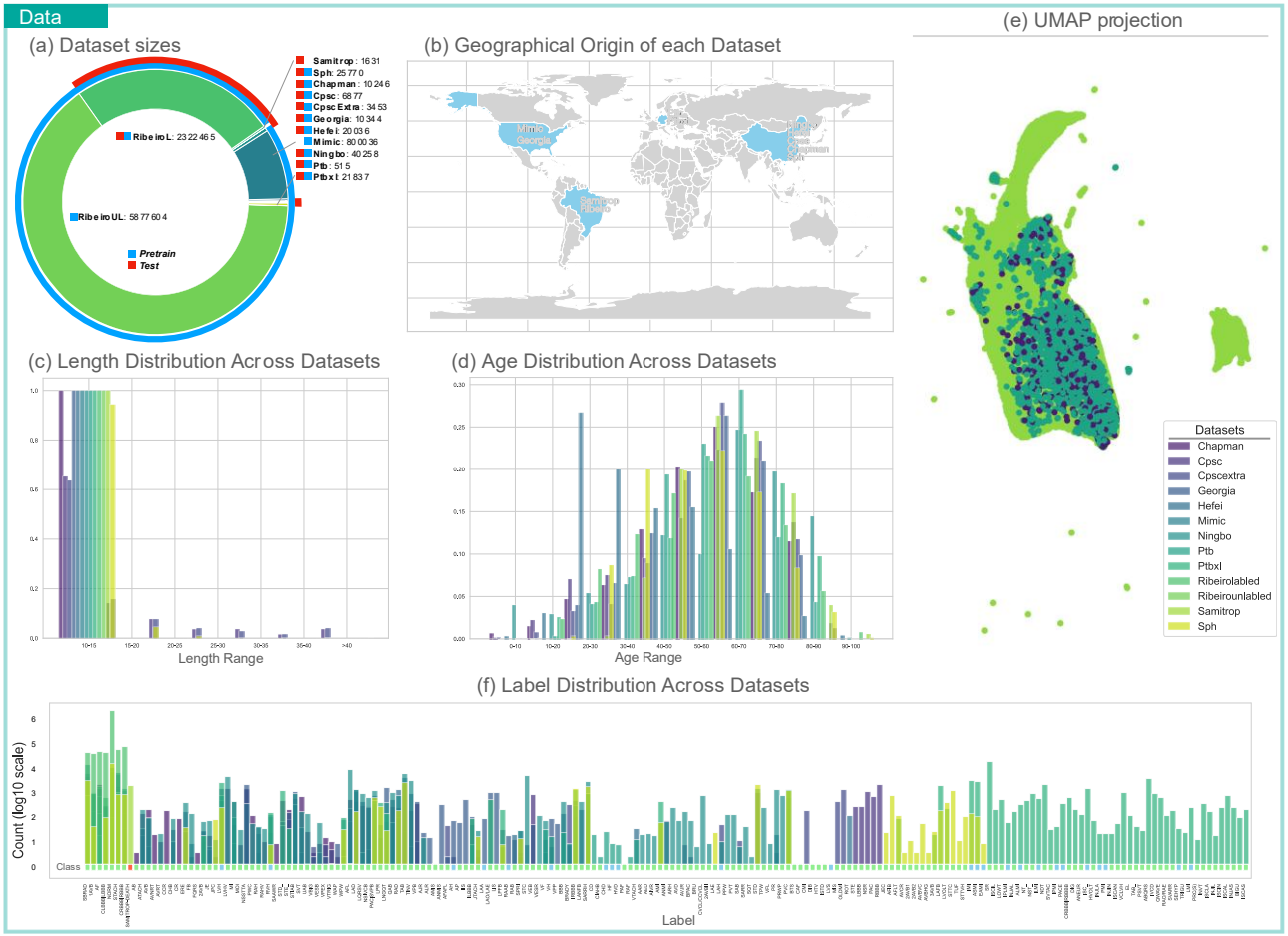


Fig 1. (a) Used dataset and their sizes. The blue bar highlights the dataset used for the pre-training while the red one is the testing one. **(b)** Geographical origin of the considered datasets. **(c)** ECG length in seconds. **(d)** Age distribution. **(e)** UMAP projection (uniform sample of 30% of all the dataset for the sake of visualisation). **(f)** Label distribution and grouping according to 3 classes (see the coloured bar below the distribution): Class 0: Green colour for labels for which the ECG the primary diagnosis tool; Class 1: Blue colour for labels for which the ECG is not the primary diagnostic tool but a potentially informative support; Class 2: Red colour for labels for which the ECG is used to predict future CVEs. Label abbreviations and corresponding diagnosis are reported in Supplementary Table 1.

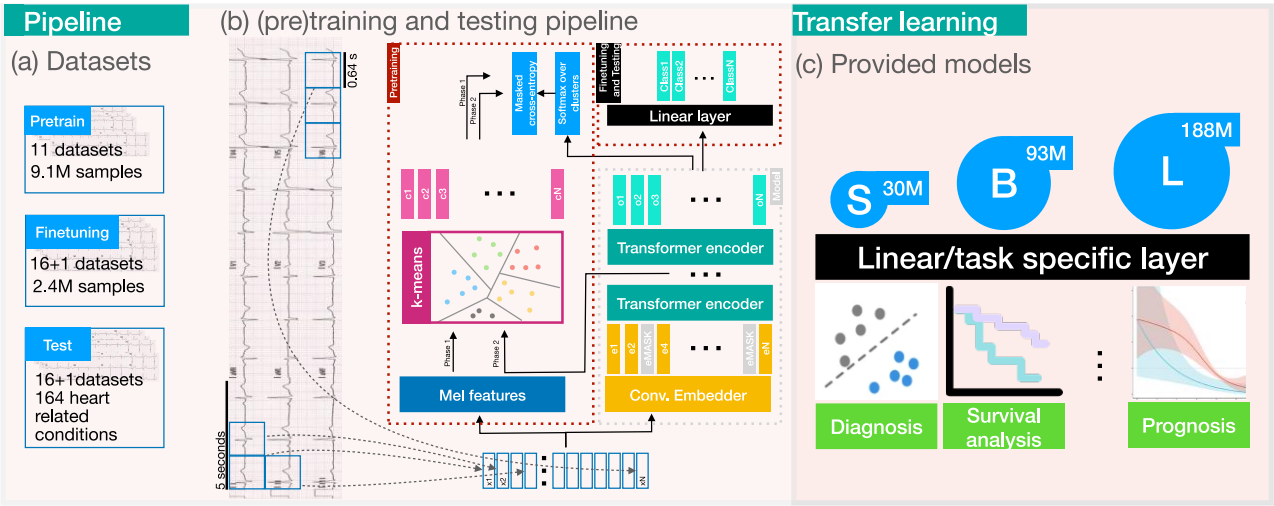


Fig 2. An overview of the proposed deep learning pipeline. (a) **Datasets**: The model is pre-trained on 11 datasets comprising over 9.1 million samples. Then, the model is fine-tuned and tested on 16 datasets and their aggregation covering 164 heart-related conditions. (b) **Pre-training and Testing Pipeline**: The pipeline begins with Mel feature extraction, followed by k-means clustering, and includes two training phases with masked cross-entropy loss and softmax classification. A convolutional embedder and a Transformer encoder process the data, culminating in a linear output layer for classification. (c) **Transfer Learning and Provided Models**: Three model variants (**S**MALL, **B**ASE, **L**ARGE) with increasing parameter counts (30M, 93M, and 188M) are pre-trained, tested and, eventually, provided to enable transfer learning for many downstream tasks such as diagnosis, survival analysis, and prognosis. Fine-tuning can be performed by simply adding a task-specific linear layer.

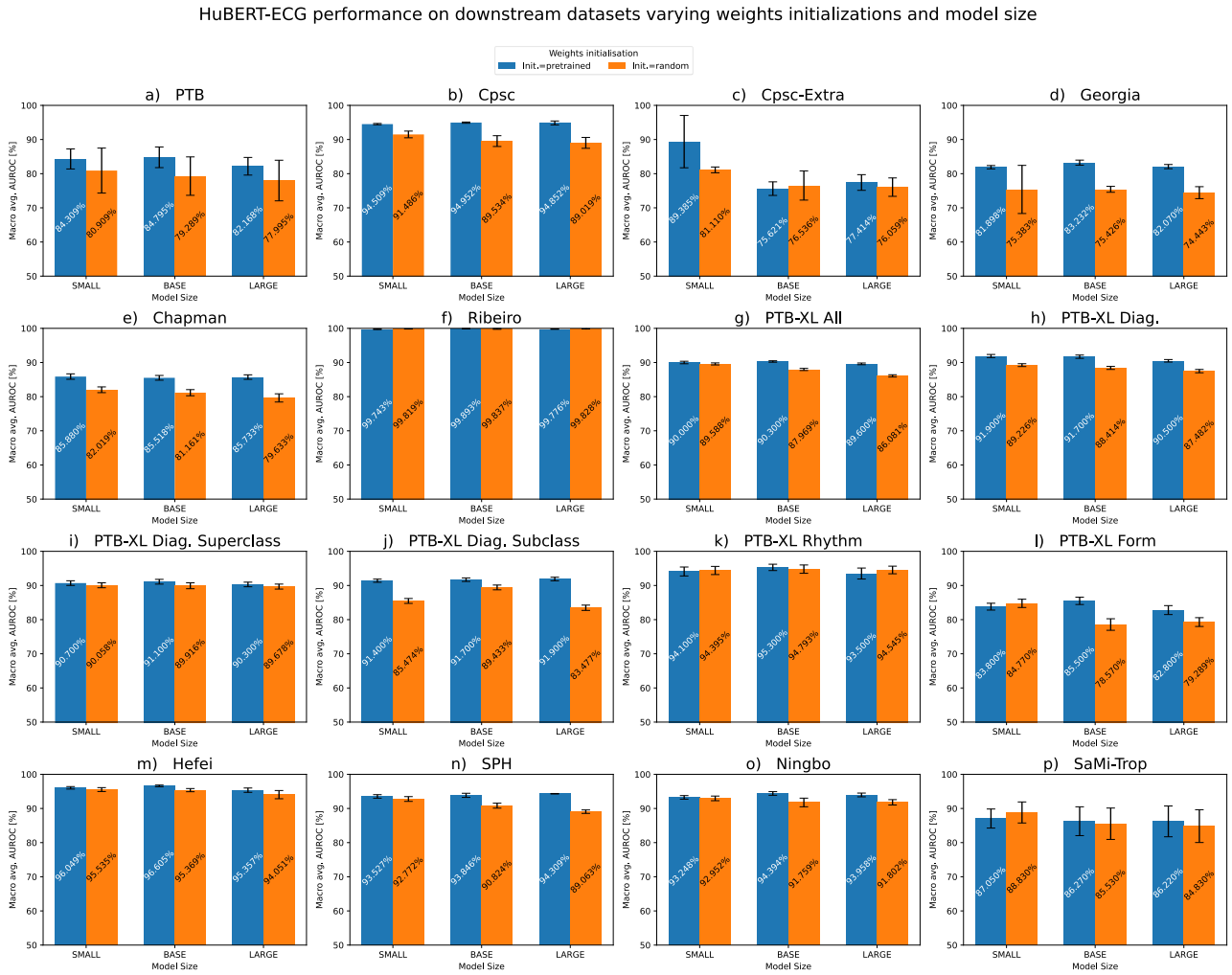


Fig. 3. HuBERT-ECG macro-averaged performance across 16 downstream datasets varying weights initialization and model size. Performance uncertainty on datasets with no predefined test set is reported through standard deviations across fold performance, while on datasets with predefined test sets uncertainty is assessed using 95% confidence intervals obtained via empirical bootstrapping with 1000 bootstrap iterations.

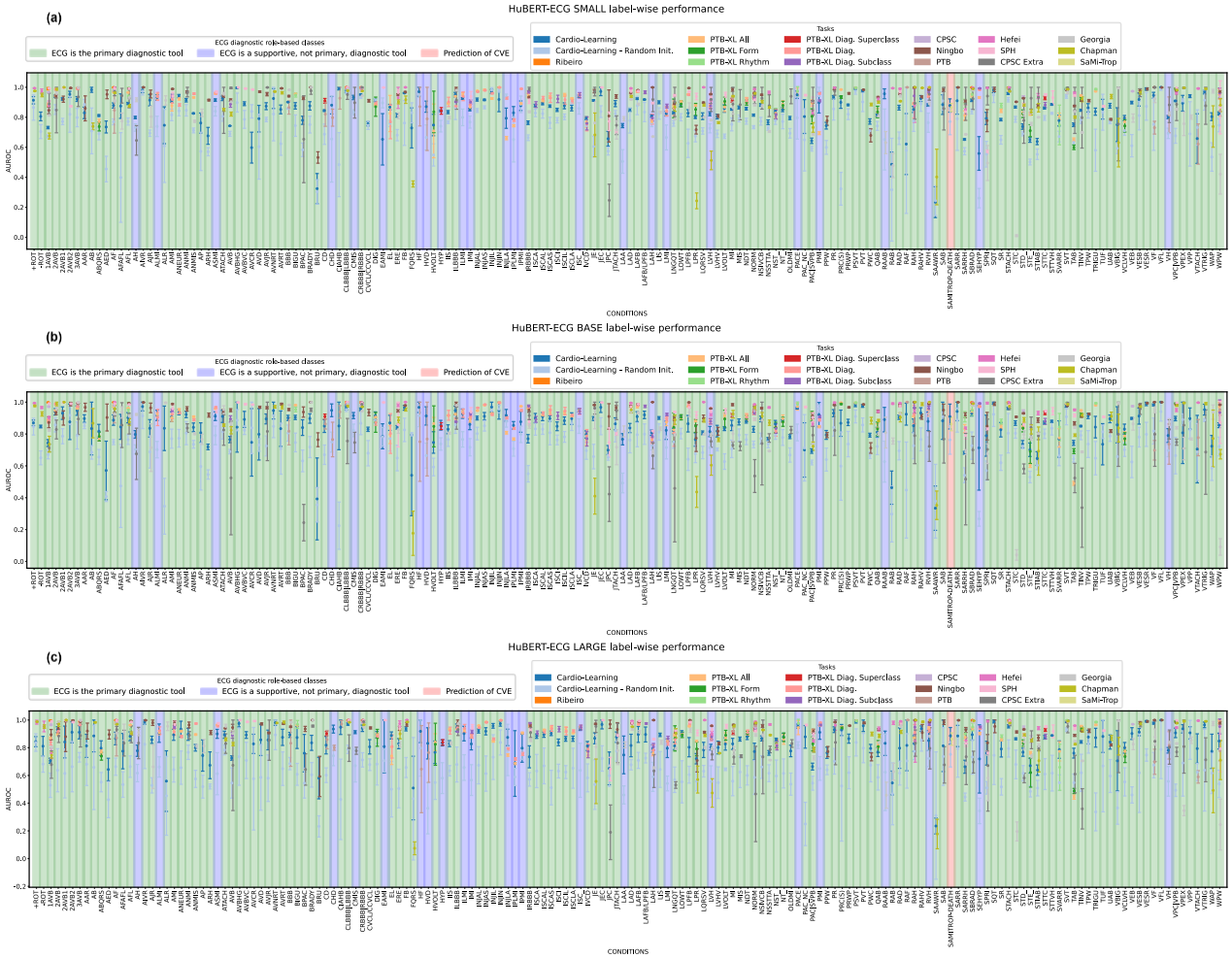


Fig. 4. HuBERT-ECG label-wise performance for **(a)** SMALL, **(b)** BASE, and **(c)** LARGE configurations across all datasets, including Cardio-Learning. The reported AUROC values are averages obtained from cross-validation and are presented with standard deviations if a label comes a dataset with no predefined test set. Otherwise, if a label comes from a dataset with a predefined test set, uncertainty is expressed with 95% confidence intervals obtained through empirical bootstrapping using 1000 bootstrap iterations. Label backgrounds are coloured based on the diagnostic role of the ECG. Supplementary Table 1 provides label abbreviations and their corresponding diagnoses. (Zoom in for better view).

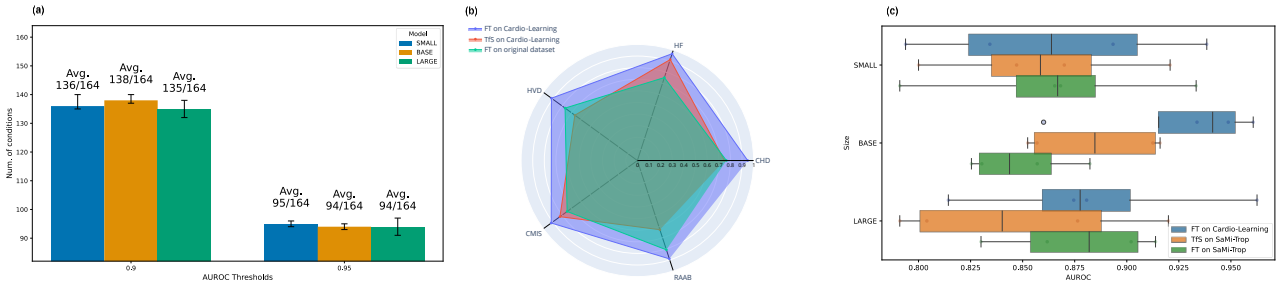


Fig. 5. **(a)** Average number of conditions, with min-max ranges, classified with at least 0.90 and 0.95 AUROC by HuBERT-ECG model configurations in at least one downstream dataset. **(b)** HuBERT-ECG averaged AUROC performance for conditions for which the ECG is an informative support for diagnosis and that appear only in one downstream dataset when the model is (i) fine-tuned on Cardio-Learning, (ii) trained from scratch on Cardio-Learning, and (iii) fine-tuned on the original downstream dataset. "FT" means fine-tuning; "TfS" means training from scratch. Conditions abbreviations: "HF" means heart failure; "CHD" means coronary heart disease; "RAAB" means right atrial abnormality; "CMIS" means chronic myocardial ischemia; "HVD" means heart valve disorder. **(c)** HuBERT-ECG AUROC trend in the prediction of death events for SaMi-Trop patients when it is fine-tuned on Cardio-Learning, trained from scratch on Sami-Trop, and fine-tuned on SaMi-Trop.

HuBERT-ECG as a self-supervised foundation model for broad and scalable cardiac applications

Edoardo Coppola[1], Mattia Savardi[2], Mauro Massussi[2,3], Marianna Adamo[2,3], Marco Metra[2,3], Alberto Signoroni[2]

Supplementary Information

1. Pre-training analyses

This section presents an analysis of how pre-training with different ECG feature descriptors and sampling rates impact downstream performance. We then investigate the interesting effects of using our dynamic regularisation during pre-training. Finally, we examine the clustering quality across both HuBERT-ECG encoding layers and pre-training iterations to measure their downstream impact. To these aims, we pre-train HuBERT-ECG BASE for one iteration with fixed hyperparameters and running configuration that are described in each of the following sections. Then, to evaluate the impact of different choices, we simply perform linear evaluations, i.e., training of a randomly initialised linear layer atop a frozen pre-trained model, on a development set that we extract from the labelled dataset used by Ribeiro et al.[1]. We build such a set, hereafter referred to as *Ribeiro-dev*, by deliberately excluding all normal ECGs, as we are more interested in assessing the ability to detect cardiac abnormalities rather than normal ECGs. Linear evaluation results represent a conservative estimate, as they are derived using fewer trainable parameters and development set that include more than 300k instances.

1.1 Exploring feature descriptors for ECG segments

In order to identify the most appropriate ECG segment descriptor to use in the initial pre-training iteration, we perform k-means clustering on the entire training set with 3 different feature descriptors extracted from the ECG segments. It is important to mention that we choose k-means over other clustering approaches (e.g. DBSCAN, hierarchical clustering, Gaussian Mixture models) for its computational efficiency and scalability, two crucial ingredients when working in large-scale setups as ours. The first descriptor takes into account 16 simple time-frequency features, while the second one is based on 39 Mel Frequency Cepstral coefficients[2] (MFCCs) following Hsu et al.[3]. These descriptors have been successfully used in ECG analysis[4,5]. The third descriptor combines the first 13 MFCCs and all the previous time-frequency features. For each descriptor, we run k-means with 10, 30, 50, 100, 150, 200 and 300 clusters and compute the corresponding sum of squared errors (SSE), or inertia (Supplementary Fig. 1). These clusters correspond to the pseudo-labels used as target variable during the self-supervised pre-training of HuBERT-ECG. We observe that MFCC- and time-frequency based descriptors provide comparable inertia and a global mean squared error in the order of 1e-5, and thus appear as promising candidates. Following the elbow method[6], and driven by the intuition that more clusters capture too fine and specific ECG patterns for the first pre-training iteration, we set C = 100 as the best number of clusters to use irrespective of the descriptor. Given this number of clusters, we compute the Davies-Bouldin[7] index to have an additional selection criterion and obtain extremely low scores. Interestingly, while clustering on MFCCs yields higher inertia than clustering on time-frequency features, clusters on the former are slightly more compact, as suggested by a marginally lower Davies-Bouldin score. Therefore, we experiment with the use of both MFCC- and time-frequency based k-means models during pre-training and, then, linearly evaluate the resulting models on *Ribeiro-dev* until performance plateau or a maximum number of 80k training steps is reached. The linear evaluation results, presented in Supplementary Table 2, show that pre-training with pseudo-labels generated by an MFCC-based k-means model yields better downstream performance.

[1] Department of Information Engineering – University of Brescia, Brescia, Italy
[2] Dept. of Medical and Surgical specialties, Radiological Sciences and Public Health – Univ. of Brescia, Brescia, Italy
[3] Cardiac Catheterization Laboratory and Cardiology, ASST Spedali Civili di Brescia, Brescia, Italy

1.2 How the sampling rate affects downstream performance

The sampling rate of an ECG depends on the specific settings of the machine recording the electrical activity of the heart and there is no standard practice in this regard. For example, ECGs from the Ribeiro et al.[1] dataset are sampled at 400 Hz, while ECGs from PTB-XL[8] and SPH[9] are sampled at 400 and 500 Hz, respectively. We investigate the effect that the ECGs sampling rate has on downstream performance. This parameter is important as it regulates the dimensionality of the data, therefore the computation speed and the memory occupation, and the degree of dilution of the information content. While it is true that training with ECGs sampled at lower frequencies is computationally less expensive, it is also true that sampling ECGs at lower frequencies may not provide enough samples to capture significant features and nuances in the input signal. Increasing the sampling rate may remedy this issue, at the cost of slower training and higher memory occupation. However, an oversampled ECG may contain redundant samples that overly dilute the information content and may be perceived as noise by a deep learning model. For these reasons, after band-pass filtering our ECGs to exclude frequencies outside [0.05, 47] Hz, which is reported to contain the dominant components of P waves, T waves and QRS complexes[10], we investigate the effects that sampling a 12-lead ECG at 50 and 100 Hz on linear evaluation performance over *Ribeiro-dev*. For these experiments, to determine a suitable sampling rate to work with, we linearly evaluate pre-trained models until the performance plateau or 80k training steps are performed. To always have the same number of embeddings being processed by the Transformer encoder, we design waveform convolutional embedders that become increasingly shallow as the sampling rate decreases, while also maintaining the same number of ECG segments. All the other design choices and running configurations are fixed across the experiments, as shown in Supplementary Table 3. Supplementary Fig. 2a illustrates the validation loss curves observed during pre-training at 50 and 100 Hz while Supplementary Fig. 2b shows the corresponding linear evaluation performance on *Ribeiro-dev*. Two discernible trends are evident from the former: as the sampling rate decreases, the validation loss curves start at lower values than those corresponding to higher sampling rates; furthermore, as the sampling rate decreases, the models tend to show earlier overfitting to the training data. We suspect that at lower sampling rates, the information needed to solve the upstream task becomes more readily accessible, thereby facilitating the learning process. However, when clinical downstream labels are introduced, there is a noticeable performance discrepancy between models operating at 50 Hz and those operating at twice this sampling rate (Supplementary Fig. 2b and Supplementary Table 4). We attribute this discrepancy to the insufficient number of samples available to capture condition-related patterns. In summary, sampling at 100 Hz captures all the meaningful physiological information, satisfies the Nyquist-Shannon theorem, and represents a desirable trade-off between computational cost and accurate downstream performance. All results are shown in Supplementary Table 4.

1.3 The effects of the dynamic regularisation

In order to gain a deeper insight into the impact of dynamic regularisation on training time and the time required to tune regularisation terms, including weight decay and dropout probabilities, we consider four different pre-training setups, each followed by a linear evaluation to assess the actual benefit on downstream performance. The first and the second ones consist in performing the entire first pre-training iteration (260k steps with no early stopping) with a weight decay of 0.01 and a dropout probability of 0.1 (referred to as *default*), with and without dynamic regularisation, respectively. The last two setups, instead, consist in performing the same pre-training iteration, with and without dynamic regularisation, but setting the initial weight decay and dropout probability to the maximum values found by the dynamic regularisation. All experiments consider the BASE architecture, ECGs sampled at 100 Hz, a MFCC-based k-means as pseudo-label generator (C = 100), a batch size of 448 instances and a masking percentage $p = 33\%$. The results of our experiments, presented in Supplementary Table 5 and labelled with letters for clarity, reveal interesting insights. When using dynamic regularisation with default weight decay and dropout probability (setup A), HuBERT-ECG converges in 80k steps, resulting in the best macro-averaged AUROC of 0.933 during linear evaluation. In contrast, disabling the dynamic regularisation while maintaining the default regularisation terms (setup B), significantly slows down the convergence and leads to inferior downstream results. Instead, when comparing setups C and D, we observe two opposite behaviours: setup C shows faster convergence but worse upstream performance, while setup D shows slower convergence but better upstream performance. Despite these discordant trends, both setups produce close downstream results, both surpassing those from setup B, but still falling short of those of setup A. This suggests that pre-training with the dynamic regularisation, or the maximum regularisation terms it finds, accelerate pre-training and allows the model to adapt itself to avoid overfitting, resulting in improved linear evaluation when compared to scenarios where the dynamic regularisation is not used at all. Nevertheless, we hypothesise that initiating pre-training with already high regularisation terms, irrespective of whether dynamic regularisation is employed or not (setups C and D), may impair the model's learning capability at the most crucial stage, resulting in inferior, albeit marginal, downstream performance. To investigate into the source of these benefits, it seems plausible that a dynamic dropout rate encourages the model to upweight alternative hidden paths in the Transformer's feed-forward blocks when necessary, without permanently excluding the dropped ones. In such cases, when also the weight decay increases, this complementary form of regularisation prevents significant changes to the weights of those paths, thereby maintaining training stability.

1.4 The downstream impact of clustering quality across encoding layers and iterations

In the work of Hsu et al.[3], prior to the second and third pre-training iterations, the most suitable number of clusters was determined empirically, as well as the layer from which latent features had to be extracted. This was achieved through the analysis of the quality of clusters of hidden representations extracted from each model layer. In particular, an automatic-speech-recognition model was used to produce frame-level phonetic labels that served as targets to measure the correctness of forced-aligned cluster assignments. The same approach is not applicable to our case as, to the best of our knowledge, there is no powerful open-source ECG model that can produce frame-level forced-aligned cluster assignments. Consequently, to determine which layer's latent features should be clustered and how many clusters are to be found, we make use once again of traditional compactness and separability metrics. In addition, to limit the set of layers to explore, we combine such metrics with findings from the NLP domain regarding layers transferability[11]. After sampling 10% of pre-training ECGs, we cluster their segments' latent representations from the 5th – 10th Transformer layer and measure the clustering quality in terms of inertia, Davies-Bouldin index and Calinsky-Harabasz[12] index. We exclude from consideration shallower layers on the grounds that they would produce too coarse representations to be of use in clustering. Similarly, deeper layers are excluded on the grounds that they would generate representations that would lead to too task-specific pseudo-labels for a generic pre-training. Upon completion of the first iteration, we identify 500 and 1000 clusters of latent features from each of these layers, since we consider that setting a higher number of clusters than that used in the first iteration is necessary for two reasons: 1) to try to generate much finer cluster assignments; 2) to avoid measuring, through clustering, the degree of separability of the classes learnt during the previous iteration. The results of this final analysis are presented in Supplementary Fig. 3. As can be seen, each metric shows a monotonic trend, regardless of the representations being clustered. Therefore, they do not provide an unequivocal indication of the best encoding layer to extract features from. To continue the pre-training of the BASE configuration, we then choose to cluster latent representations from the 8th layer into 500 clusters, as this point seems to mark a non-negligible change in the metrics we consider. After completing the second iteration, we re-introduce diagnostic labels and perform a linear evaluation with the same running configurations reported in previous paragraphs to quantify the relative improvement with respect to previous evaluations. As shown in Supplementary Fig. 4, when the first-iteration model saturates, the second-iteration one achieves results that are approximately 5 AUROC points better and still has room for improvement. Once HuBERT-ECG BASE completes the second iteration, we repeat the clustering step described above and report the results in Supplementary Fig. 3. Once again, the metrics we consider show a monotonic trend that does not facilitate the choice of an acceptable number of clusters (500 or 1000) nor which layer the latent representations should be extracted from. For these reasons, to start pre-training the SMALL and LARGE model configurations, we decide to extract latent features from the 9th layer and to cluster them into 500 and 1000 clusters, respectively. We believe that pre-training these third-iteration models requires, for any configuration, to select a deeper layer than those selected for the previous iterations. Furthermore, increasing the quantity and fineness of cluster assignments can be beneficial for a model with higher learning capacity such as HuBERT-ECG LARGE. Conversely, for the SMALL model size, maintaining the same number and granularity of cluster assignments can facilitate the task of mimicking the BASE model without significant loss of performance. The benefits of these choices are also displayed in Supplementary Fig. 4. Upon completion of 362.5k- and 422.5k-step pre-trainings, the SMALL and LARGE configurations perform similarly when linearly evaluated on *Ribeiro-dev* for 80k steps: both achieve a macro-averaged AUROC slightly lower than that of the BASE configuration, but show no sign of strong saturation within the training time of this linear evaluation.

2 Ablation Study

In this section, we present a short sequence of ablation studies to investigate the effects of specific architectural choices and the impact of important hyperparameters. In particular, we study (1) the impact of our masking strategy, and (2) the effects of multi-task learning when experimenting with multiple cluster ensembles. To do so we perform multiple pre-trainings of HuBER-ECG BASE (first iteration only) followed by linear evaluations on *Ribeiro-dev* with fixed hyperparameters. If not otherwise mentioned, the experimental configurations are the same of the previous section.

2.1 Impacts of the Masking Strategy

We consider setting the value of the masking percentage p of crucial importance to make HuBERT-ECG learn high quality representations of 12-lead ECGs. An excessively low value would generate a trivial upstream task, while an exaggeratedly high value would result into a nearly impossible one. For this reason, we experiment with multiple values of this hyperparameter in order to see how it impacts on downstream performance and plot the results of these experiments in Figure 5. For the sake of comparison, we also plot the exact percentage of embeddings that are masked when the masking strategy proposed by Hsu et al.[3] is used. Beyond some statistical and label noise, what emerges clearly is that

setting $p = 33\%$ guarantees the best results, as linear evaluation performance peaks at this masking percentage, while following the masking strategy used to pre-train HuBERT leads to suboptimal performance. We believe this finding is due to the more regular patterns and high information redundancy of ECGs compared to audio signals.

2.2 The Effects Of Multi-Task Learning

To observe the downstream effects of pre-training HuBERT-ECG in a multi-task learning framework, we perform three pre-trainings with an increasing number of tasks to solve. Initially, we pre-train HuBERT-ECG using pseudo-labels generated by a single MFCC-based k-means model with 100 clusters. Then, we pre-train again with an additional clustering model with 200 clusters and, eventually, we consider an ensemble of three k-means models with 100, 200 and 300 clusters. Supplementary Table 6 reports the attained results. Interestingly, adding just a new clustering model does not change pre-training nor affects downstream performance. In contrast, although marginal, we see an improvement when considering an ensemble of three k-means models providing much more granular pseudo-labels. Since solving three tasks at once is harder than solving just one of them, it is not surprising to see that such performance gains follow a much longer pre-training. We also believe that the cardinality of the ensemble is not as relevant as the maximum number of clusters, which is likely more useful to the model to capture label-specific patterns in downstream data. This is analogous to the refinement of the cluster assignments that is performed prior to the second iteration, except for how such refinement is achieved, since both refinements aim to generate finer and more granular pseudo-labels for the ECG segments. Considering both the longer pre-training time and the marginal improvement obtained on the results reached after training with pseudo-labels generated by a single clustering model, we do not proceed in pre-training with cluster ensembles.

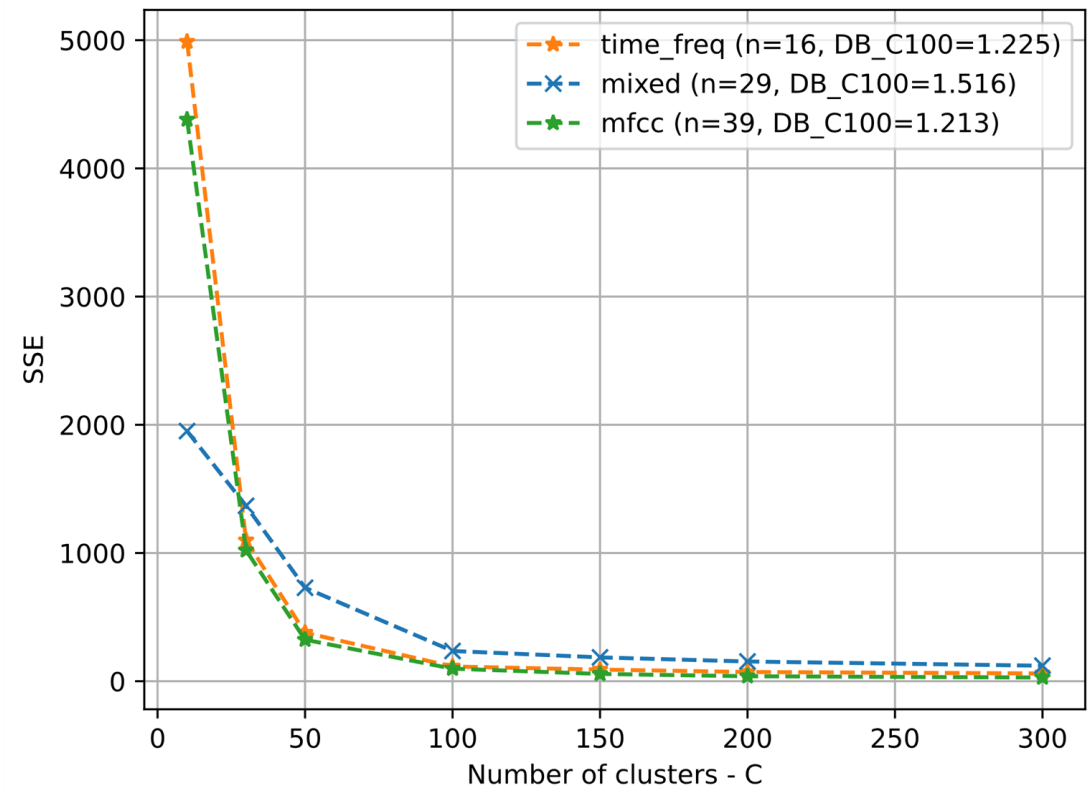
3 Explaining fine-tuned diagnostic models

The self-supervised pre-trained HuBERT-ECG model has no knowledge of diseases or their manifestations in ECG signals. During pretraining, the model learns only the structural and statistical properties of the signal, encoding them into general-purpose representations[13]. This learning process is guided by the pre-training task of predicting pseudo-labels—generated in an unsupervised, disease-unrelated way—for masked ECG embeddings. As such, the representations learned are not directly aligned with clinically useful, human-interpretable ECG characteristics. Because of this, applying explainability techniques directly to the pretrained model would yield explanations that are clinically uninformative: the internal features are not intended to capture pathophysiological phenomena. In contrast, once HuBERT-ECG is fine-tuned on a supervised diagnostic task, its representations become aligned with disease-specific patterns, making explainability techniques more meaningful.

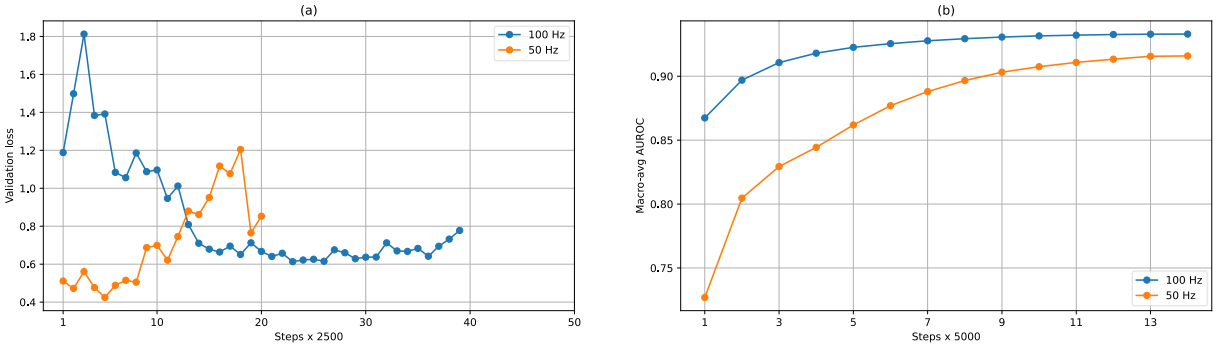
In this section, we apply explainability methods to a supervised fine-tuned version of HuBERT-ECG. Specifically, we use attention maps[14], which are well-suited to Transformer-based architectures, to visualize which parts of the ECG the model focuses on when making predictions. In Supplementary Fig. 6, we present four ECGs from the PTB-XL dataset labelled as Sinus Rhythm and correctly classified by the model. The attention maps show that the model focuses on multiple regions of the signal, with some variability across recordings. This variability is expected and desirable, reflecting inter-patient differences and signal-specific features that HuBERT-ECG learns during fine-tuning and leverages when predicting the output condition. In Supplementary Fig. 7, we show two ECGs labelled as Atrial Fibrillation, also correctly detected. In these cases, the attention maps highlight clinically relevant signal segments, suggesting that the model is focusing on meaningful characteristics rather than spurious patterns or noise.

While these visualizations do not provide a direct one-to-one mapping with classical ECG features (e.g., P-wave morphology or QRS duration), they offer qualitative evidence that the fine-tuned model attends to diagnostic signal regions in a task-consistent manner. The observed variability in attention patterns reflects the model's flexibility in handling heterogeneous ECG presentations, rather than a lack of intelligent feature extraction.

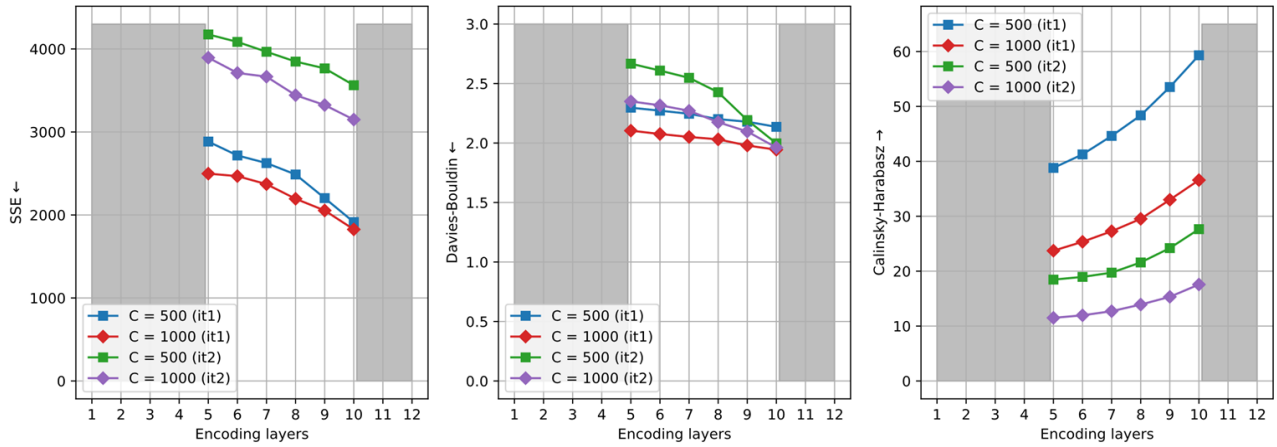
Supplementary Figures



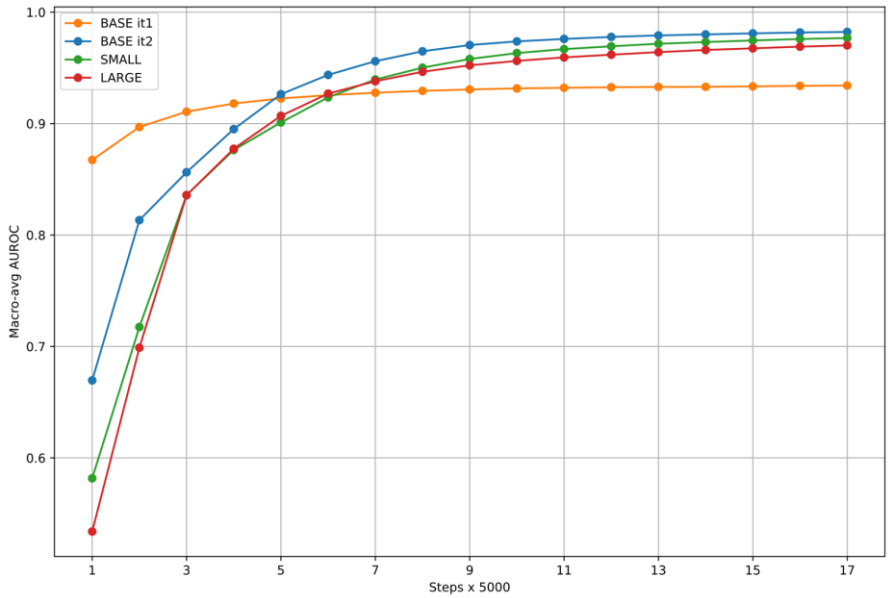
Supplementary Fig. 1 | Performance comparison of multiple k-means clustering runs in terms of Sum of Squared Errors (SSE). Once fixed the optimal number of clusters (C = 100), the Davies-Bouldin index (DB) is computed and reported as "DB_C100". Lower DB values indicate better clustering.



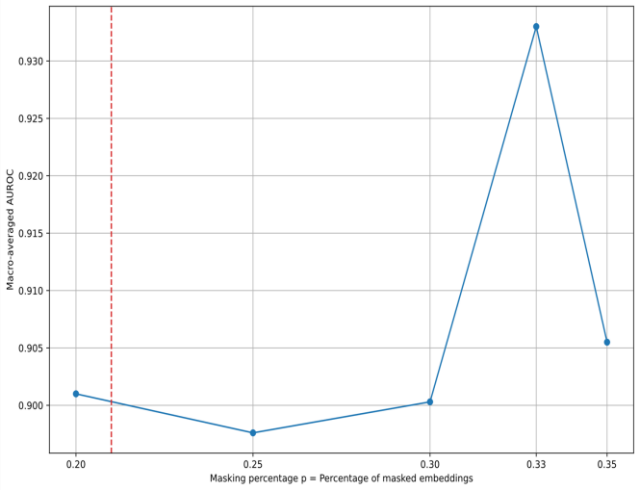
Supplementary Fig. 2 | (a) Validation loss curves obtained when pre-training HuBERT-ECG with ECGs sampled at 50 and 100 Hz. **(b)** Downstream performance obtained when HuBERT-ECG is linearly evaluated on *Ribeiro-dev* after being pre-trained with ECGs sampled at 50 and 100 Hz.



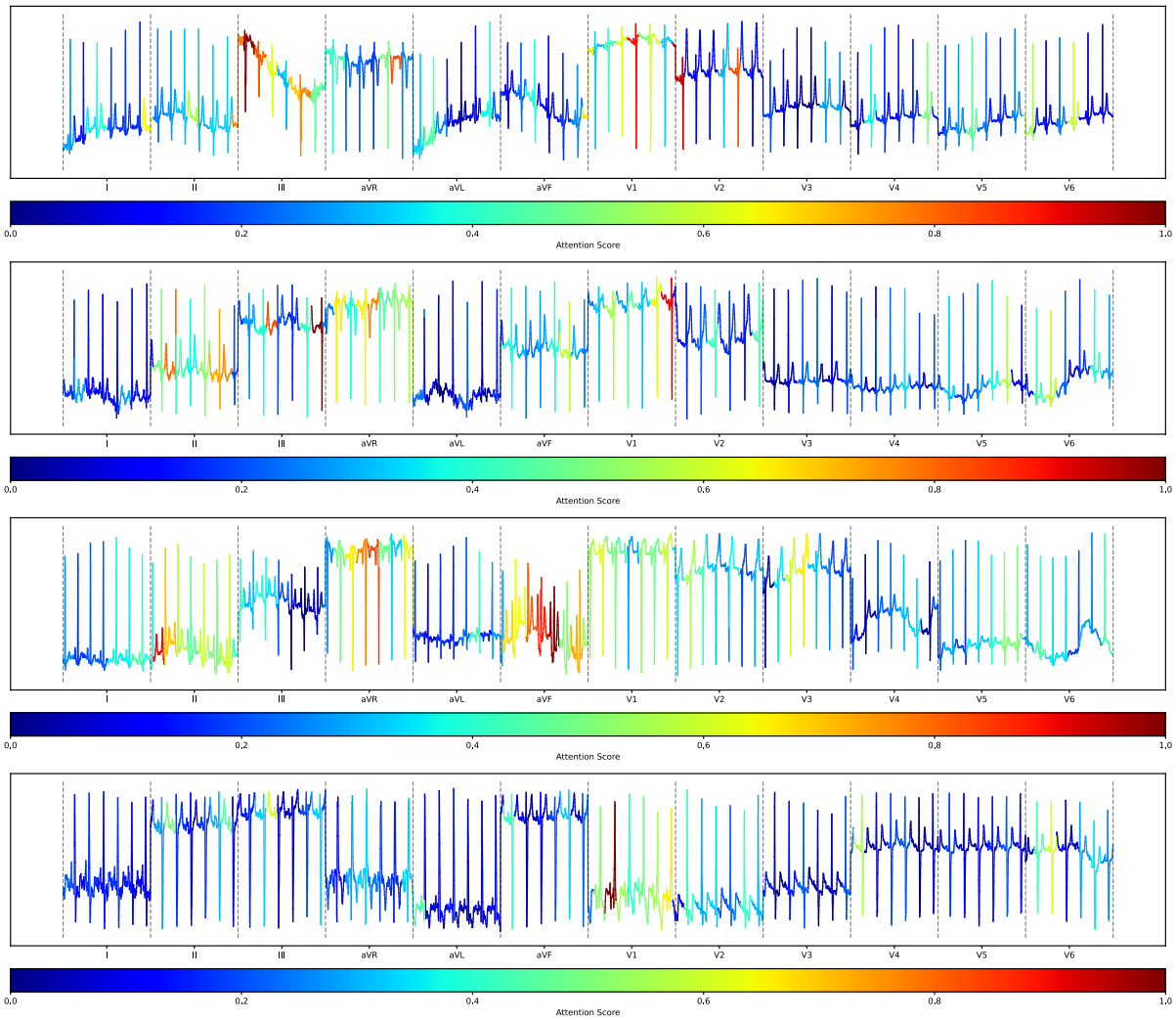
Supplementary Fig. 3 | Clustering quality metrics (inertia, Davies-Bouldin index, Calisky-Harabasz index) across HuBERT-ECG BASE encoding layers after the first (it1) and second (it2) pre-training iterations. Symbols ↑and ↓ indicate whether a metric needs to be maximised or minimised, respectively. Gray-shaded regions refer to encoding layers not considered in this analysis.



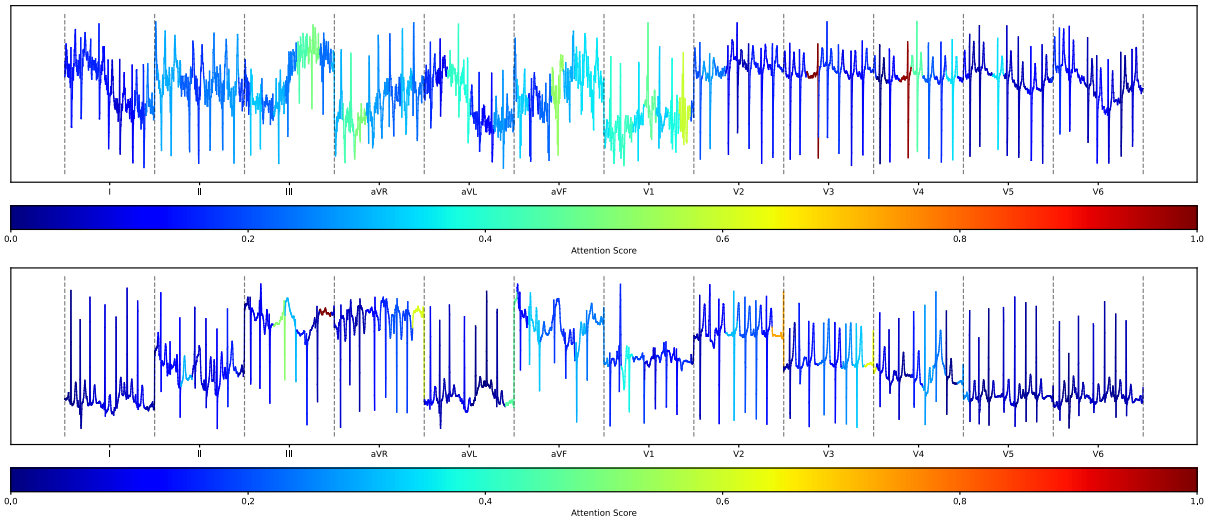
Supplementary Fig. 4 | Linear evaluation performance of HuBERT-ECG BASE, after first and second pre-training iteration, HuBERT-ECG SMALL and HuBERT-ECG LARGE on *Ribeiro-dev*.



Supplementary Fig. 5 | HuBERT-ECG BASE linear evaluation performance after pre-training with different values of the hyper-parameter *p* (i.e. the percentage of ECG embeddings to mask). The red dashed line indicates the percentage of ECG embeddings that would be masked if we followed the masking strategy used to pre-train HuBERT[3].



Supplementary Fig. 6 | *Attention maps on predictions formulated by a fine-tuned HuBERT-ECG for correclty classified Sinus Rhythm ECGs*



Supplementary Fig. 6 | *Attention maps on predictions formulated by a fine-tuned HuBERT-ECG for correclty classified Atrial Fibrillation ECGs*

Supplementary Tables

Supplementary Table 1: This table lists the harmonised diagnoses and the corresponding abbreviations used throughout this study. For conditions that appear with different names across various datasets, the table also includes found synonyms. When generating the Cardio-Learning dataset, condition appearing under various synonyms are standardised to a unique name. The 'Super-conditions' column indicates hierarchically broader diagnostic categories. When creating the Cardio-Learning dataset, if broader and more specific conditions come from different constituent datasets, the broader conditions augment the annotation of an ECG that is labelled with the more specific condition. If an ECG is labelled with a more general condition, its annotation is not augmented with a more specific diagnostic category, unless explicitly confirmed in the original annotation. Importantly, these hierarchical augmentations are performed only when the hierarchy involves diagnostic labels coming from different constituent datasets. This means that if a more specific condition and its broader diagnostic category appear in the same dataset, but an ECG from that source is explicitly labelled only the more specific one, the hierarchical mapping is not performed to respect the distinction made by the original annotators.

Diagnosis	**Abbreviation**	**Found Abbreviated Synonyms**	**Super-conditions**
2nd Degree Av Block	2AVB	IIAVB	AVB
Second Av Block Mobitx Type I	2AVB1	IIAVB I, MOI	AVB, 2AVB
Mobitz Type II Atrioventricular Block	2AVB2	IIAVB II,	AVB, 2AVB
Av Block Varying Conduction	AVBVC		AVB
Av Block Advanced (High Grade)	AVBHG		AVB
Av Block Complete (Third Degree)	3AVB	IIIAVB, CHB	AVB
1st Degree Av Block	1AVB	IAVB	AVB
Anterior Myocardial Infarction	ANMI		MI
Extensive Anterior Myocardial Infarction	EAMI		ANMI, MI
St Elevation	STE_	STE	MI
Inferoposterolateral Myocardial Infarction	IPLMI		MI
Posterior Myocardial Infarction	PMI		MI
Inferolateral Myocardial Infarction	ILMI		MI
Inferoposterior Myocardial Infarction	IPMI		MI
Anterolateral Myocardial Infarction	ALMI		MI
Lateral Myocardial Infarction	LMI		MI
Acute Myocardial Infarction	AMI		MI
Anteroseptal Myocardial Infarction	ASMI		MI
Inferior Myocardial Infarction	IMI		MI
Ischemic In Anterior Leads (Subclass)	ISCA		MIS
Anterior Ischemia	ANMIS	ISCAN	MIS
Inferior Ischaemia	IIS	ISCIN	MIS
Subendocardial Injury In Inferior Leads	INJIN		MIS
Ischemic In Inferior Leads	ISCI		MIS
Subendocardial Injury In Inferolateral Leads	INJIL		MIS
Ischemic In Inferolateral	ISCIL		MIS
Subendocardial Injury In Lateral Leads	INJLA		MIS
Ischemic In Lateral Leads	ISCLA		MIS
Lateral Ischaemia	LIS		MIS

Subendocardial Injury In Anteroseptal Leads	INJAS		MIS
Ischemic In Anteroseptal Leads	ISCAS		MIS
Subendocardial Injury In Anterolateral Leads	INJAL		MIS
Ischemic In Anterolateral Leads	ISCAL		MIS
Left Atrial Abnormality	LAA		
Left Atrial Hypertrophy	LAH	LAO, LAE	AH
Right Atrial Abnormality	RAAB		
Right Atrial Hypertrophy	RAH	RAO, RAE	AH
Left Ventricular Hypertrophy	LVH		VH
Voltage Criteria (QRS) For Left Ventricular Hypertrophy	VCLVH		VH
Septal Hypertrophy	SEHYP		VH
Right Ventricular Hypertrophy	RVH		VH
ST-T Change Due To Ventricular Hypertropy	STTVH		VH
Incomplete Right Bundle Branch Block	IRBBB		BBB
Complete Left Bundle Branch Block | Left Bundle Branch Block	CLBBB|LBBB	CLBBB, LBBB	BBB
Incomplete Left Bundle Branch Block	ILBBB		BBB
Complete Right Bundle Branch Block | Right Bundle Branch Block	CRBBB|RBBB	CRBBB, RBBB	BBB
Transient Ischemic Attack	TIA		
Atrial Hypertrophy	AH		
Myocardial Infarction	MI		
Myocardial Ischemia	MIS		
Ventricular Hypertrophy	VH		
Coronary Heart Disease	CHD		
Chronic Myocardial Ischemia	CMIS	CMI	
Heart Failure	HF		
Heart Valve Disorder	HVD		
Left Ventricular Strain	LVS		
Countercolockwise Rotation	-ROT	CCR	CVCL/CCVCL
Clockwise Rotation	+ROT	CR	CVCL/CCVCL
Accelerated Atrial Escape Rhythm	AAR		
Atrial Bigeminy	AB		
Abnormal QRS	ABQRS		
Atrial Escape Beat	AED		
Atrial Fibrillation	AF	AFIB	
Atrial Fibrillation And Flutter	AFAFL		
Atrial Flutter	AFL	AFLT	
Accelerated Idioventricular Rhythm	AIVR		
Accelerated Junctional Rhythm	AJR		
Suspect Arm Ecg Leads Reversed	ALR		
Atrial Pacing Pattern	AP		
Atrial Rhythm	ARH		
Atrial Tachycardia	ATACH		
Av Block	AVB	_AVB	
Atrioventricular Dissociation	AVD		
Atrioventricular Junctional Rhythm	AVJR		

Atrioventricular Node Reentrant Tachycardia	AVNRT		
Atrioventricular Reentrant Tachycardia	AVRT		
Bundle Branch Block	BBB		
Blocked Premature Atrial Contraction	BPAC		
Bradycardia	BRADY		
Brugada	BRU		
Brady Tachy Syndrome	BTS		
Cardiac Dysrhythmia	CD		
Congenital Incomplete Atrioventricular Heart Block	CIAHB		
Clockwise Or Counterclockwise Vectorcardiographic Loop	CVCL/CCVCL		
Diffuse Intraventricular Block	DIB		
Early Repolarization	ERE		
Fusion Beats	FB		
Fqrs Wave	FQRS		
High T-Voltage	HTV		
Indeterminate Cardiac Axis	ICA		
Idioventricular Rhythm	IR		
Junctional Escape	JE		
Junctional Premature Complex	JPC		
Junctional Tachycardia	JTACH		
Left Axis Deviation	LAD		
Left Posterior Fascicular Block	LPFB		
Prolonged Pr Interval	LPR		
Low Qrs Voltages	LQRSV		
Prolonged Qt Interval	LNGQT	LQT	
Left Ventricular High Voltage	LVHV		
Nonspecific Intraventricular Conduction Disorder	NSIVCB		
Normal Sinus Rhythm	NORM		
Nonspecific St T Abnormality	NSSTTA		
Old Myocardial Infarction	OLDMI		MI
Premature Atrial Contraction | Supraventricular Premature Beats	PAC|SVPB	PAC, SVPB, SVPB|PAC	
Prolonged P Wave	PPW		
Pacing Rhythm	PR		
Poor R Wave Progression	PRWP		
Paroxysmal Supraventricular Tachycardia	PSVT		
Paroxysmal Ventricular Tachycardia	PVT		
P Wave Change	PWC		
Qwave Abnormal	QAB	QAVE	
R Wave Abnormal	RAB		
Right Axis Deviation	RAD		
Rapid Atrial Fibrillation	RAF		
Right Atrial High Voltage	RAHV		
Sinus Arrhythmia	SARRH	SA	
Sinus Atrium To Atrial Wandering Rhythm	SAAWR		
Sinoatrial Block	SAB		

Sinus Arrest	SARR		
Sinus Bradycardia	SBRAD	SB	
Sinus Node Dysfunction	SND		
Shortened Pr Interval	SPRI		
Decreased Qt Interval	SQT		
Sinus Tachycardia	STACH	ST	
S T Changes	STC		
St Depression	STD_	STD	
St Interval Abnormal	STIAB		
Supraventricular Bigeminy	SVB		
Supraventricular Tachycardia	SVT	SVTACH	
T Wave Abnormal	TAB	TAB_	
T Wave Inversion	TINV	INVT	
Tall P Wave	TPW		
U Wave Abnormal	UAB		
Ventricular Bigeminy	VBIG		
Ventricular Ectopics	VEB		
Ventricular Escape Beat	VESB		
Ventricular Escape Rhythm	VESR		
Ventricular Fibrillation	VF		
Ventricular Flutter	VFL		
Premature Ventricular Contractions | Ventricular Premature Beats	VPC|VPB	VPC, VPB, PVC	
Ventricular Pre Excitation	VPEX		
Ventricular Pacing Pattern	VPP		
Paired Ventricular Premature Complexes	VPVC		
Ventricular Tachycardia	VTACH		
Ventricular Trigeminy	VTRIG		
Wandering Atrial Pacemaker	WAP		
Wolff-Parkinson-White Pattern	WPW		
Low Voltage	LVOLT		
TU fusion	TUF		
atrial premature complexes non-conducted	PAC_NC		PAC|SVPB
av conduction ration N:D	AVCR		
left anterior fascicular block	LAFB	LANFB	
junctional escape complex(es)	JEC		
ST deviation with T-wave change	STTC		
left anterior/posterior fascicular block	LAFB/LPFB		
non specific intraventricular conduction disturbance	IVCD		
non specific t wave changes	NT_		
sinus rhythm	SR		
digitalis effect	DIG		
premature complex(es)	PRC(S)		
supraventricular arrhythmia	SVARR		
trigeminal pattern (unknown origin SV or Ventricular)	TRIGU		
low amplitude t waves	LOWT		
electrolytic disturbance or drug (former EDIS)	EL		

bigeminal pattern (unknown origin SV or Ventricular)	BIGU		
normal functionning artificial pacemaker	PACE		
non-diagnostic t abnormalities	NDT		
ST-T changes compatible with ventricular aneurysm	ANEUR		
non specific ST changes	NST_		
non specific ischemic	ISC_		
hypertrophy	HYP		
high qrs voltage	HVOLT		

Supplementary Table 2: Linear evaluation performance of pre-trained HuBERT-ECG BASE models when using pseudo-labels generated by MFCC- and time-frequency based k-means models.

Label generator	Macro-averaged AUROC
MFCC-based k-means – C = 100	0.933
Time-frequency based k-means – C = 100	0.913

Supplementary Table 3: Architecture design and running configuration when experimenting with ECGs sampled at 50 and 100 Hz.

Architectural outline	Sampling rate	
	50 Hz	100 Hz
Convolutional embedder	kernels = (10, 3, 3, 2) strides= (4, 2, 2, 2) channels = 512	kernels = (10, 3, 3, 2, 2) strides= (4, 2, 2, 2, 2) channels = 512
Transformer encoder	BASE	
k-means	MFCC, C = 100	
Pre-training steps Linear evaluation steps Batch size P LayerDrop Dropout Weight decay	130k 80k 448 33% 0.1 0.1 0.01	

Supplementary Table 4: Linear evaluation performance of HuBERT-ECG on *Ribeiro-dev* after being pre-trained at 50 and 100 Hz. Linear evaluation steps necessary to let validation AUROC plateau are also reported.

ECG sampling rate	macro-averaged AUROC	Linear evaluation steps at plateau
50 Hz	0.9176	65k
100 Hz	0.9330	

Supplementary Table 5: Effects of the use of dynamic regularisation when pre-training HuBERT-ECG in terms of pre-training steps and upstream performance under the form of validation loss and validation accuracy. For each setup, linear evaluation performance on *Ribeiro-dev* is also reported.

Setups	Pre-training steps	Pre-training (validation loss, validation accuracy)	macro-averaged AUROC
Dynamic regularisation & Default regularisation Terms	80k	0.6379, 0.8205	0.933
No Dynamic regularisation & Default regularisation Terms	235k	0.6493, 0.8182	0.921
(C) Dynamic regularisation & Maximum regularisation Terms	72.5k	0.6775, 0.8111	0.9263
No Dynamic regularisation & Maximum regularisation Terms	90k	0.5860, 0.8312	0.9244

Supplementary Table 6: Linear evaluation performance of HuBERT-ECG BASE after being pre-trained in a multi-task learning framework in which tasks are represented by multiple k-means models composing an ensemble of label generators. The corresponding number of pre-training steps is reported for every task/cluster ensemble.

K-means ensemble	Macro-averaged AUROC	Pre-training steps
MFCC-based k-means, C = 100	0.933	80k
MFCC-based k-means, C = 100, 200	0.930	75k
MFCC-based k-means, C = 100, 200, 300	0.943	135k

Supplementary References

1. Ribeiro, A. H. et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. Nat Commun 11, 1760 (2020).
2. Xu, M. et al. HMM-based audio keyword generation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 3333, (2004).
3. Hsu, W.-N. et al. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. IEEE/ACM Trans Audio Speech Lang Process 29, 3451–3460 (2021).
4. Singh, A. K. & Krishnan, S. ECG signal feature extraction trends in methods and applications. Biomed Eng Online 22, 22 (2023).
5. Boussaa, M., Atouf, I., Atibi, M. & Bennis, A. ECG signals classification using MFCC coefficients and ANN classifier. in 2016 International Conference on Electrical and Information Technologies (ICEIT) 480–484 (IEEE, 2016).
6. Thorndike, R. L. Who belongs in the family? Psychometrika 18, 267–276 (1953).
7. Davies, D. L. & Bouldin, D. W. A Cluster Separation Measure. IEEE Trans Pattern Anal Mach Intell PAMI-1, 224–227 (1979).
8. Wagner, P. et al. PTB-XL, a large publicly available electrocardiography dataset. Sci Data 7, 154 (2020).
9. Liu, H. et al. A large-scale multi-label 12-lead electrocardiogram database with standardized diagnostic statements. Sci Data 9, 272 (2022).
10. Hong, S., Zhang, W., Sun, C., Zhou, Y. & Li, H. Practical Lessons on 12-Lead ECG Classification: Meta-Analysis of Methods From PhysioNet/Computing in Cardiology Challenge 2020. Front Physiol 12, (2022).
11. Rogers, A., Koleva, O. & Rumshisky, A. A primer in BERTology: What we know about how BERT works. Trans Assoc Comput Linguist 8, 842–866 (2021).
12. Calinski, T. & Harabasz, J. A dendrite method for cluster analysis. Commun Stat Theory Methods 3, 1–27 (1974).
13. Liu, X. et al. Self-Supervised Learning: Generative or Contrastive. IEEE Trans Knowl Data Eng 35, 857–876 (2023).
14. Abnar, S. & Zuidema, W. Quantifying Attention Flow in Transformers. in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics 4190–4197 (Association for Computational Linguistics, Stroudsburg, PA, USA, 2020).