ECG-NEST-FM: A FREQUENCY-FOCUSED ECG FOUNDATION MODEL WITH NESTED EMBEDDINGS

Abhishek Sharma *
SEAS, Harvard University
Allston, Massachusetts, USA
abhisheksharma@g.harvard.edu

Lin Yang
Google Research
Cambridge, Massachusetts, USA
linyan@google.com

Cory Y. McLean
Google Research
Cambridge, Massachusetts, USA
cym@google.com

Justin Khasentino †
Google Research
Cambridge, Massachusetts, USA
jtcosentino@google.com

Farhad Hormozdiari †
Google Research
Cambridge, Massachusetts, USA
fhormoz@google.com

ABSTRACT

Electrocardiograms (ECGs) are fundamental to cardiac diagnostics, providing noninvasive insights into cardiovascular conditions. Recent advancements in deep learning have led to foundation models (FMs) capable of learning powerful representations of ECG signals. However, these models often fail to fully exploit the periodic nature and diagnostic frequency bands of ECGs, leading to inefficiencies in computational cost and interpretability. We propose a novel ECG foundation model that learns nested embeddings, where each subset of dimensions encodes progressively higher-frequency information. By explicitly modeling frequency structures and applying a correlation penalty, the method achieves compact, high-rank representations that reduce model size without sacrificing performance. We evaluate our approach on two large-scale datasets for embedding redundancy and prediction performance on downstream clinical tasks such as arrhythmia classification and cardiac condition detection. We observe similar prediction performance AUROC scores and lower embedding redundancy, offering a computationally efficient and interpretable framework for ECG analysis. Finally, we demonstrate that running genome-wide association studies (GWAS) on representations obtained from our model in UK Biobank data captures known cardiovascular variants and detects novel loci, which can be applied to drug discovery.

1 INTRODUCTION

Electrocardiograms (ECGs) are a key tool for non-invasive cardiac diagnostics, providing insights into a patient's heart health at the initial point of clinical contact. Accurate interpretation of ECG signals is essential for effective patient care, yet traditional methods are subject to variability and inaccuracies. Recent advancements in artificial intelligence (AI) have expanded the potential applications of ECGs in diagnostic and predictive medicine (Sau et al., 2024). In particular, foundation models that leverage self-supervised learning (SSL) to capture representations of ECG signals have emerged as a promising solution to scale ECG analysis to large datasets without the need for manual annotation. These models compress high-dimensional ECG signals into low-dimensional representations or features, which transfer to a variety of clinically meaningful tasks such as detecting

*The author contributed to this work while being student researcher at Google research.
†Corresponding authors.

arrhythmias, heart failure, and stratifying patient risk (Abbaspourazad et al., 2024; McKeen et al., 2024; Song et al., 2024; Han et al., 2024).

Given their applications in the clinical setting, it is important that the models deployed are transparent and interpretable (Kiseleva et al., 2022). Such interpretability is also important due to legal and regulatory requirements in healthcare applications (Ennab & Mcheick, 2022). This motivates the development of foundation models which, in addition to learning useful embeddings for prediction, can also allow inspection into what the learned representations encode. ECG signals are periodic and have been shown to have frequency-specific characteristics that can be used for diagnosis and predictive applications (Zhang et al., 2022; Zyout et al., 2023). Also, in cases where ECG machines provide on-device diagnosis, small foundation models which allow efficient and local inference are desirable (Abbaspourazad et al., 2024).

Existing SSL approaches have attempted to address some of these modeling goals, but not all. Abbaspourazad et al. (2024) proposed a PPG and ECG foundation model with 256-dimensional representations. This is considerably smaller than the ECG foundation models proposed in the literature, where the representations are typically chosen to be 1024-dimensional (Song et al., 2024; McKeen et al., 2024). However, they do not exploit frequency-domain information in the ECG signal, and do not focus on the interpretability of the learned representations. Zhang et al. (2022) use frequency-domain information to learn representations using a contrastive loss (Chen et al., 2020a). However, their focus is not on learning a foundation model for ECG signals, and their learned representations do not provide insight into which features of the ECG signal are being captured. McKeen et al. (2024) utilize saliency maps extracted from the final attention layer in the encoder. However, these saliency maps can only be obtained in transformer architectures, and do not provide interpretable attributions for the learned representations.

To address these challenges, we propose a novel ECG foundation model with nested representations (ECG-Nest-FM) that combines a transformer-based encoder with a Matryoshka-inspired decoder (Kusupati et al., 2022). Our model learns nested representations that learn increasingly high frequency components of the ECG signal. By comparing the predictions of the representations from different levels of the hierarchy, we can identify which range of frequencies contribute to a specific task. Our model employs a VICReg loss to minimize cross-correlation between dimensions and incorporates frequency- and time-domain reconstruction losses to enforce interpretability (Bardes et al., 2022). Our approach ensures high effective rank in the representations, giving us diverse and informative representations of the ECG signal. We evaluate our model on two large-scale datasets, MIMIC-IV-ECG and CODE-15%, across multiple clinically relevant prediction tasks like atrial fibrillation. We show how our model's predictions can used to infer the frequency bands that are most important for a specific prediction task. We observed that the ECG-Nest-FM representations derived from UK Biobank Sudlow et al. (2015) data, which includes genomic information, capture genetic signals and replicate known genetic variants previously associated with cardiovascular disease. Furthermore, this interpretability and feature diversity does not come at the cost of downstream performance. Our results show that ECG-Nest-FM achieves comparable performance (w.r.t. AUROC and AUPRC) to standard foundation modeling architectures of similar representation size.

2 Related Work

Self-supervised learning In recent years, self-supervised learning has become increasingly popular across many areas of deep learning. Our work is most related to a large body of self-supervised learning research using masked reconstruction to learn representations without the need for labels (He et al., 2022; Dosovitskiy, 2020; Nie et al., 2023). This is in contrast to contrastive learning, which requires identification of positive and negative samples in the batch to learn representations (Chopra et al., 2005; Chen et al., 2020b). Self-supervised methods are susceptible to representation collapse, and several architectures have been proposed to avoid this, some of which do not require positive and negative samples (Caron et al., 2021; Bardes et al., 2022; He et al., 2020). Self-supervised pre-trained models are promising because they have been shown to encode significant amount of information about downstream targets without seeing any labels during pre-training (Baevski et al., 2020; Cheng et al., 2020; Chen et al., 2021; Gopal et al., 2021; Kiyasseh et al., 2021; Mehari & Strodthoff, 2022; Baevski et al., 2022). There are several works looking at ways to build representations appropriate for time series data (such as ECG), (Zeng et al., 2023; Salinas

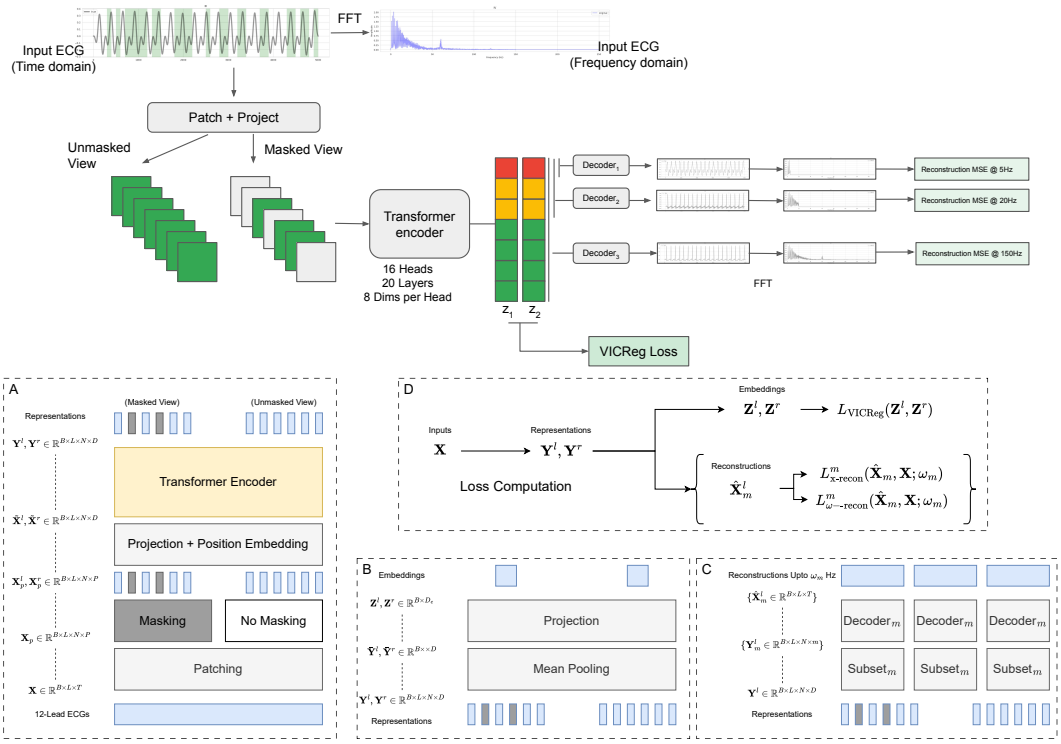


Figure 1: **Top**: Overall pipeline of ECG-Nest-FM for ECG representation learning. **Bottom**: Individual aspects of the architecture and loss computation. A: 12-lead ECGs are patched, optionally masked and encoded using a transformer encoder to generate masked and unmasked views of ECG representations. B: Both views are pooled and projected to create the embeddings. C: Masked representations are decoded using a Matryoshka-decoder, producing reconstructions with progressively increasing frequency information. D: Loss computation leverages embeddings to enforce decorrelation of representation dimensions and ensure encoding of increasing frequency information within the ECG signal representation.

et al., 2020; Sharma et al., 2024; Ruan et al., 2019). However, transformers have been shown to have competitive performance (Nie et al., 2023), leading us to choose a transformer encoder.

SSL in ECG analysis SSL has proven to be effective in ECG analysis, with new methods developed to leverage the special structure of ECG signals. These works use 12-lead ECG signals to learn ECG representations via a self-supervised approach like contrastive learning or masked autoencoder (Abbaspourazad et al., 2024; McKeen et al., 2024; Song et al., 2024; Han et al., 2024; Kiyasseh et al., 2021; Oh et al., 2022; Friedman et al., 2025). However, to the best of our knowledge, none of these works aim to develop nested representations focused at interpreting ECG signals.

Using Frequency information in ECG analysis Both supervised and unsupervised/self-supervised learning methods have been developed to focus on the frequency information in ECG signals (Hu et al., 2020; Aziz et al., 2021; Zyout et al., 2023; Pradhan et al., 2023; Perkins et al., 2020; Zhang et al., 2022). However, these methods do not focus on learning a foundation model whose representations can transfer to predicting several cardiac outcomes.

3 METHODS

3.1 OVERALL APPROACH

Our method aims to create an ECG foundation model with representations that are increasingly complex in the dimension index (i.e., first 64 dimensions are more complex than first 32 dimensions).

To achieve this, we combine a transformer-based encoder with a Matryoshka-style decoder (inspired by Kusupati et al. (2022)). We used a pre-training loss that incorporates masked reconstruction in both frequency- and time-domains, and a correlation penalty to encourage disentanglement of representation dimensions. We present the full architecture of ECG-Nest-FM in Figure 1.

3.2 DATASETS

We use two large-scale datasets to train and evaluate our model: MIMIC-IV-ECG (Gow et al., 2023) and CODE-15% (Ribeiro et al., 2021). We use 60%-20%-20% splits for training, validation, and held-out evaluation on MIMIC-IV-ECG, and reserve all of CODE-15% as a held-out dataset to test our model's generalization. The splitting is done on the patient IDs to avoid any leakage of same-patient ECGs into validation/test sets. The MIMIC-IV-ECG dataset contains 800,035 ECG recordings from 161,352 subjects. Each ECG is sampled at 500 Hz for 10 seconds (5,000 time steps). Within the CODE-15% dataset, ECG records have lengths of either 10 seconds or 7 seconds (sampled at 400 Hz), and all records are zero-padded to 4,096 timesteps. The CODE-15% dataset contains 345,779 ECGs from 233,770 patients.

Data Preprocessing Both datasets are band-pass filtered from 0.67 Hz to 150 Hz to align with American Heart Association (AHA) recommendations (Kligfield et al., 2007). For the CODE-15% dataset, we used the 10-second ECG records. We first removed the zero-padding, and then upsampled the 400 Hz signal to 500 Hz to match the ECG sampling rate in MIMIC-IV-ECG dataset. This ensured consistent input dimensionality across all samples. We were left with 105,192 ECGs from 73,658 patients in the CODE-15% dataset after this filtering step.

3.3 MODEL ARCHITECTURE

Our model adopts a standard transformer encoder as the backbone. We distinguish representations from embeddings following Bardes et al. (2022): a token vector $\boldsymbol{x} \in \mathbb{R}^D$ is encoded by the encoder f_θ into its *representation* $\boldsymbol{y} = f_\theta(\boldsymbol{x})$, which is then transformed by the projector h_ϕ onto the *embeddings* $\boldsymbol{z} = h_\phi(\boldsymbol{y})$. The representations are used for downstream evaluations and the embeddings are used for computing the VICReg loss. We define f_θ and h_ϕ in the following setup.

Patching Let $\boldsymbol{X} \in \mathbb{R}^{L \times T}$ be the L-lead ECG matrix ($L = 12$ for 12-lead ECGs). Note that the actual data also includes a leading mini-batch dimension but we omit it for clarity. Each lead—i.e., each row of $\boldsymbol{X}$—is a univariate time series with T timesteps. To avoid $O(T^2)$ processing time by a transformer, we reshape $\boldsymbol{X}$ on the T dimension to get N non-overlapping patches of length P (such that $T = NP$) to get $\boldsymbol{X}_p \in \mathbb{R}^{L \times N \times P}$. Such patching is a standard way to 'tokenize' a time series (Nie et al., 2023). For our 5000-timestep-ECGs, we patch the inputs to 10 patches of 500 samples each.

Masking Many SSL methods build two *views* from input by applying two transformations on it (Balestriero et al., 2023). We create the first view $\boldsymbol{X}_p^l$ (superscript l denotes 'left' view) to be a masked version of $\boldsymbol{X}$ where the patches are multiplied with a Bernoulli mask $\boldsymbol{B} \in \{0, 1\}^{L \times N}$ which zeros-out each patch with probability p. We used the second view $\boldsymbol{X}_p^r$ to be the same as unmasked input $\boldsymbol{X}_p$ (superscript r denotes 'right' view). The patches are independently projected to the representation space of dimension D by a 1-layer MLP to get tensors $\tilde{\boldsymbol{X}}^l, \tilde{\boldsymbol{X}}^r \in \mathbb{R}^{L \times N \times D}$. In our setup below, we drop the superscripts l and r whenever a transformation is applied to both views.

Encoder and projector The encoder transformer function $f_\theta : \mathbb{R}^{N \times D} \to \mathbb{R}^{N \times D}$ maps each patch $\tilde{\boldsymbol{X}}^l, \tilde{\boldsymbol{X}}^r$ to representations $\boldsymbol{Y}^l, \boldsymbol{Y}^r \in \mathbb{R}^{L \times N \times D}$ (all L leads are processed independently by the transformer encoder). We use a transformer encoder with 20 layers and 16 self-attention heads per layer. Each attention head has a dimension of 8, leading to a final representation of size 128. We mean-pool the representations over the leads (L) and patches (N), followed by using a projector $h_\phi : \mathbb{R}^D \to \mathbb{R}^{D_e}$ that maps the representations to get the ECG embeddings $\boldsymbol{z}^l, \boldsymbol{z}^r \in \mathbb{R}^{D_e}$, as computing the losses on the embeddings leads to better performance in practice (Balestriero et al., 2023; Bardes et al., 2022).

Matryoshka Decoder Following Kusupati et al. (2022), we choose a set of representation sizes, $\mathcal{M} \in \{1, \cdots, D\}$. Given a size $m \in \mathcal{M}$ and a representations tensor $\boldsymbol{Y}$, we get *nested representations* $\boldsymbol{Y}_m := \boldsymbol{Y}_{:,:,1:m} \in \mathbb{R}^{L \times N \times m}$ by indexing on the final dimension of $\boldsymbol{Y}$. These nested representations are the first m dimensions of the $D-$dimensional representations for each lead and patch. For each $m \in \mathcal{M}$, we also choose frequency thresholds ω_m such that $\boldsymbol{Y}^m$ is constrained to only contain information up to ω_m Hz of the signal. Each nested representation $\boldsymbol{Y}_m$ is mapped to time-domain reconstruction patches $\hat{\boldsymbol{X}}_m \in \mathbb{R}^{L \times T}$ by using a decoder network $g^m : \mathbb{R}^m \to \mathbb{R}^P$ followed by reshaping the patches to ECG's length T (the decoder and reshape operations are applied to each lead independently). Because of its resemblance to the Matryoshka-doll style indexing of the representations, the decoder $\{g^m\}_{m=1}^{|\mathcal{M}|}$ is called a Matryoshka decoder (Kusupati et al., 2022). We use the representation sizes $m \in \{32, 64, 128\}$ and frequency thresholds $\omega_{32} = 5$Hz, $\omega_{64} = 20$Hz, $\omega_{128} = 150$Hz. These thresholds were decided based on the fact that the typical ranges for the T-wave, the P-wave, and the QRS complex are 0–10 Hz, 5-30 Hz, and 8-50 Hz, respectively (Zyout et al., 2023; Tereshchenko & Josephson, 2015). Note that our usage of a Matryoshka decoder is different from Kusupati et al. (2022) because their reconstructions $\{\hat{\boldsymbol{X}}_m\}_m$ would be aim to recover the full ECG $\boldsymbol{X}$, whereas our reconstructions aim to recover only the signal up to ω_m Hz, as we discuss next.

3.4 Pre-training objectives

We train our model using a combination of masked patch reconstruction loss (Nie et al., 2023) and a correlation penalty (Bardes et al., 2022) to encourage the learned representations to capture informative frequency-domain features while ensuring that representations from different views $(\boldsymbol{Y}^l, \boldsymbol{Y}^r)$ remain consistent with each other and diverse in the representation dimensions.

Masked reconstruction loss To encourage the model to learn frequency-domain features, we propose two reconstruction losses. In the first loss, we match the reconstruction of the masked inputs $(\hat{\boldsymbol{X}}_m^l)$ in the frequency domain:

$$L^m_{\omega\text{-RECON}}(\hat{\boldsymbol{X}}_m, \boldsymbol{X}; \omega_m) = \frac{1}{L}\sum_{j=1}^{L} ||\text{mask}_{\omega_m}(\text{FFT}(\hat{\boldsymbol{X}}_m[j])) - \text{mask}_{\omega_m}(\text{FFT}(\boldsymbol{X}[j]))||_2^2 \quad (1)$$

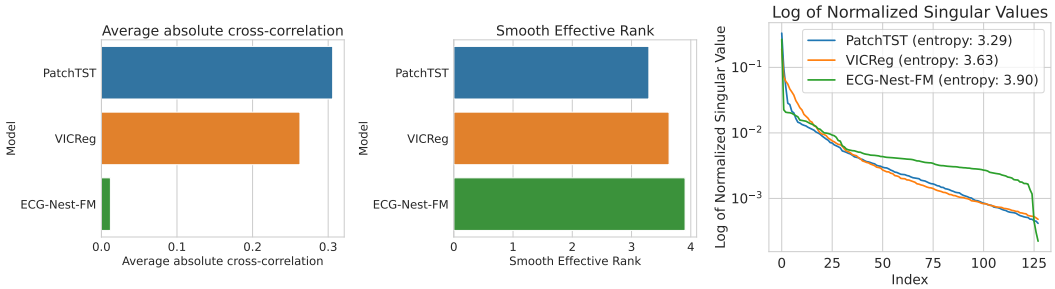
where $\boldsymbol{X}[j] \in \mathbb{R}^T$ indexes the lead j from the ECG signal. FFT denotes the forward Fourier transform, which converts a time domain signal into its frequency-domain representation. By using the mask_{ω_m}, only frequency components below ω_m are considered when comparing the reconstructed and original signals.

The second loss reconstructs the masked components in the time domain:

$$L^m_{\text{X-RECON}}(\hat{\boldsymbol{X}}_m, \boldsymbol{X}; \omega_m) = \frac{1}{L}\sum_{j=1}^{L} ||\text{FFT}^{-1}\left(\text{mask}_{\omega_m}\left(\text{FFT}(\hat{\boldsymbol{X}}_m[j])\right)\right) - \text{FFT}^{-1}\text{mask}_{\omega_m}(\text{FFT}(\boldsymbol{X}[j]))||_2^2 \quad (2)$$

where FFT^{-1} is the inverse Fourier transform that takes a masked frequency-domain signal back to the time domain.

VICReg Correlation Penalty To enforce consistency between the two representation views $\boldsymbol{Y}^l$ and $\boldsymbol{Y}^r$ while preventing representation collapse, we incorporate a loss introduced on the embeddings $\boldsymbol{z}^l, \boldsymbol{z}^r$ (Bardes et al., 2022). We denote $\boldsymbol{Z}^l = [\boldsymbol{z}_1^l, \ldots, \boldsymbol{z}_B^l]$ and $\boldsymbol{Z}^r = [\boldsymbol{z}_1^r, \ldots, \boldsymbol{z}_B^r]$ be the two batches composed of B embedding vectors of dimension D_e: $L_{\text{VICReg}}(\boldsymbol{Z}^l, \boldsymbol{Z}^r) = \lambda s(\boldsymbol{Z}^l, \boldsymbol{Z}^r) + \mu\left(v(\boldsymbol{Z}^l) + v(\boldsymbol{Z}^r)\right) + \eta\left(c(\boldsymbol{Z}^l) + c(\boldsymbol{Z}^r)\right)$ where the similarity term $s(\boldsymbol{Z}^l, \boldsymbol{Z}^r)$, encourages similarity between representations of two views of the same input, the variance term $v(\boldsymbol{Z})$ helps maintain sufficient variation in each latent dimension to avoid representation collapse (i.e., trivial or constant representations), and the covariance term, $c(\boldsymbol{Z})$, acts as a decorrelation penalty so that different dimensions in the representation are not redundant. We discuss the loss terms and implementation details in Appendix A.2.



(a) (b) (c)

Figure 2: Representation diversity on the CODE-15% dataset. (a) Average pairwise cross-correlation of representations. Lower values indicate less redundancy in the learned features thus possibly better embeddings. (b) Smooth Effective Rank (SER) of representations. Lower cross-correlations and higher SER values indicate a greater number of linearly independent features and suggest that the representations possess a higher effective rank and greater feature diversity (Balestriero et al., 2023), both of which are associated with improved downstream task performance. Higher values indicate possibly better embeddings. (c) Log of normalized singular values for the representations matrix provide further evidence that ECG-Nest-FM representations have a more uniform spectrum, suggesting that representations the higher-rank subspace.

3.5 Evaluation Metrics

Embedding Cross-Correlation To verify that the correlation penalty in 3 effectively reduces redundancy across embedding dimensions, we compute the average absolute cross-correlation of the representations: $\frac{1}{|U|}\sum_{i<j}|R_{ij}|$, where $R \in \mathbb{R}^{D\times D}$ is the correlation matrix of the representation features (with U denoting the set of off-diagonal index pairs) computed over the test data. Lower average cross-correlation indicates greater decorrelation among features, which has been shown to improve performance on downstream tasks (Bardes et al., 2022).

Smooth effective rank Smooth effective rank (SER) is the entropy of the normalized singular value distribution of a matrix (Roy & Vetterli, 2007). SER serves as a metric to quantify the diversity and rank collapse of learned representations, helping assess the effectiveness of representation learning methods without the use of any labels (Garrido et al., 2023). Also, without requiring the labels, average cross-correlation and SER help in model selection since they correlate with downstream "usefulness" (Balestriero et al., 2023).

Downstream Classification Performance on CODE-15% To ensure clinical validity of the learned representations on a held-out dataset, we perform non-linear probing of representations to predict the outcomes in the CODE-15% dataset. These outcomes include the patient's sex, and the following cardiac conditions: 1st degree AV block (1dAVb), left/right bundle branch block (LBBB, RBBB), sinus bradycardia (SB), sinus tachycardia (ST), and atrial fibrillation (AFib). We use a non-linear probe to better quantify the diagnostic information captured about each frequency band in the representations (Pimentel et al., 2020b), and also to be sure that low-frequency representations indeed have no information about higher frequencies when they have lower AUROC/AUPRCs. We also share the linear probing results in the Appendix A.3 to accompany the non-linear probing results to show sensitivity of performance to probe complexity (Pimentel et al., 2020a). We present both AUROC and AUPRC values for these classification tasks in Figure 6. While AUROC quantifies the overall discriminative quality, AUPRC is particularly useful for rare outcomes, like in our dataset, as it quantifies how the model correctly classifies the outcome label. We train a Histogram-based Gradient Boosting Classification Tree on each subset of the learned representations (32, 64, or 128 dimensions) as a non-linear probe. Detailed descriptions of how the targets are created are provided in Appendix A.1. We also present analogous results for the MIMIC-IV-ECG dataset in Appendix A.3. On MIMIC-IV-ECG, the outcomes include sinus rhythm, atrial fibrillation, sinus arrhythmia, left and right ventricular hypertrophy, left-axis deviation, sinus tachycardia, and sinus bradycardia.

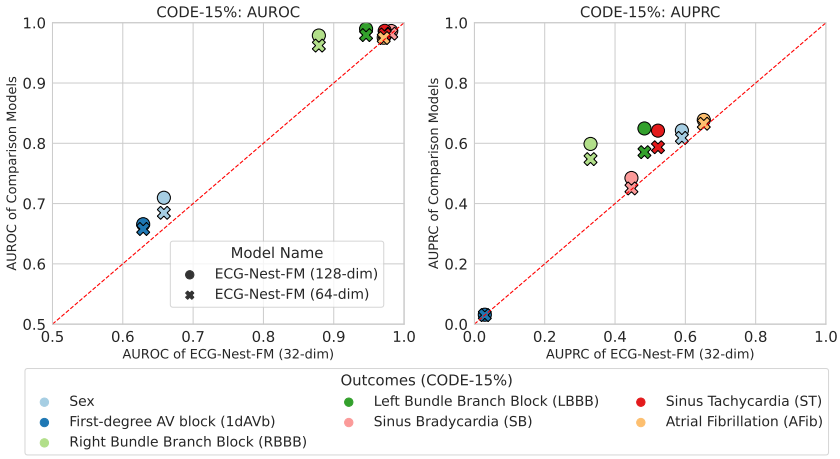


Figure 3: Comparison of AUROC (left) and AUPRC (right) for predicting clinical outcomes using 32-dim, 64-dim, and 128-dim representations in CODE-15% dataset. The x-axis shows AUROC with 32-dim representations (information up to 5 Hz), while the y-axis shows AUROC with higher dimensions——crosses for 64-dim (up to 20 Hz) and circles for 128-dim (full signal up to 150 Hz).

4 RESULTS

ECG-Nest-FM helps identify outcomes with low frequency components While most outcomes perform best with full representations, Figure 3 shows that 32-dimensional representations already perform comparably in classifying atrial fibrillation, sinus tachycardia, and sinus bradycardia. Fibrillatory wave frequencies are typically identified in the range of 4–9 Hz (Husser et al., 2007). Thus, it is clinically relevant that even representations limited to up to 5 Hz of information perform well in classification. Sinus tachycardia is characterized by a heart rate exceeding 100 beats per minute (bpm), often due to physiological or pathological causes (Mayuga et al., 2022). This corresponds to a frequency range of approximately 1.67–3.33 Hz (corresponding to 100–200 bpm, where 100 bpm is the diagnostic lower bound and 200 bpm is the typical upper bound). Sinus bradycardia is defined by a heart rate below 60 bpm, which typically translates to a frequency range of 0.5–1 Hz for heart rates between 30 and 60 bpm (Hafeez & Grossman). Note that this analysis was only enabled by our nested representations—unlike standard representation learning methods for ECG, which do not disentangle the effects of different frequency ranges. A natural question arises if this comes at a cost of downstream performance of the representations.

ECG-Nest-FM representations encode patient health information We evaluated the health information encoded in the ECG-Nest-FM representations using non-linear probing on MIMIC-IV-ECG and CODE-15%, and present the AUROC and AUPRC values in Figure 6. We compared ECG-Nest-FM to alternate architectures: PatchTST (i.e., a masked autoencoder without the VICReg loss and the matryoshka decoder) and VICReg (i.e., ECG-Nest-FM with the matryoshka decoder). ECG-Nest-FM achieved AUROCs and AUPRCs similar to those of PatchTST and VICReg (Figure 6). This indicates that introducing nested representations does not compromise downstream performance. Combined with the previous result, *we observe that the improved interpretability does not come at a cost of predictive performance*. By disentangling frequency components, our model helps the practitioner understand which frequencies contribute to which outcomes. Notably, our approach effectively predicted left and right bundle branch blocks, which aligns with previous work demonstrating that the frequency spectrum of the QRS complex (typically 8–50 Hz) is informative for these diagnoses (Niebauer et al., 2014; Alventosa-Zaidin et al., 2019; Zyout et al., 2023). By disentangling frequency components, our representations are able to capture these clinically relevant features. However, performance was weaker for predicting left and right ventricular hypertrophy, likely due to the more complex and less standardized frequency-based diagnostic criteria for these conditions. We see from Figure 2 that ECG-Nest-FM achieves high effective rank compared to the baseline models. This suggests that the learned features are decorrelated and diverse, and that the representations can capture more information.

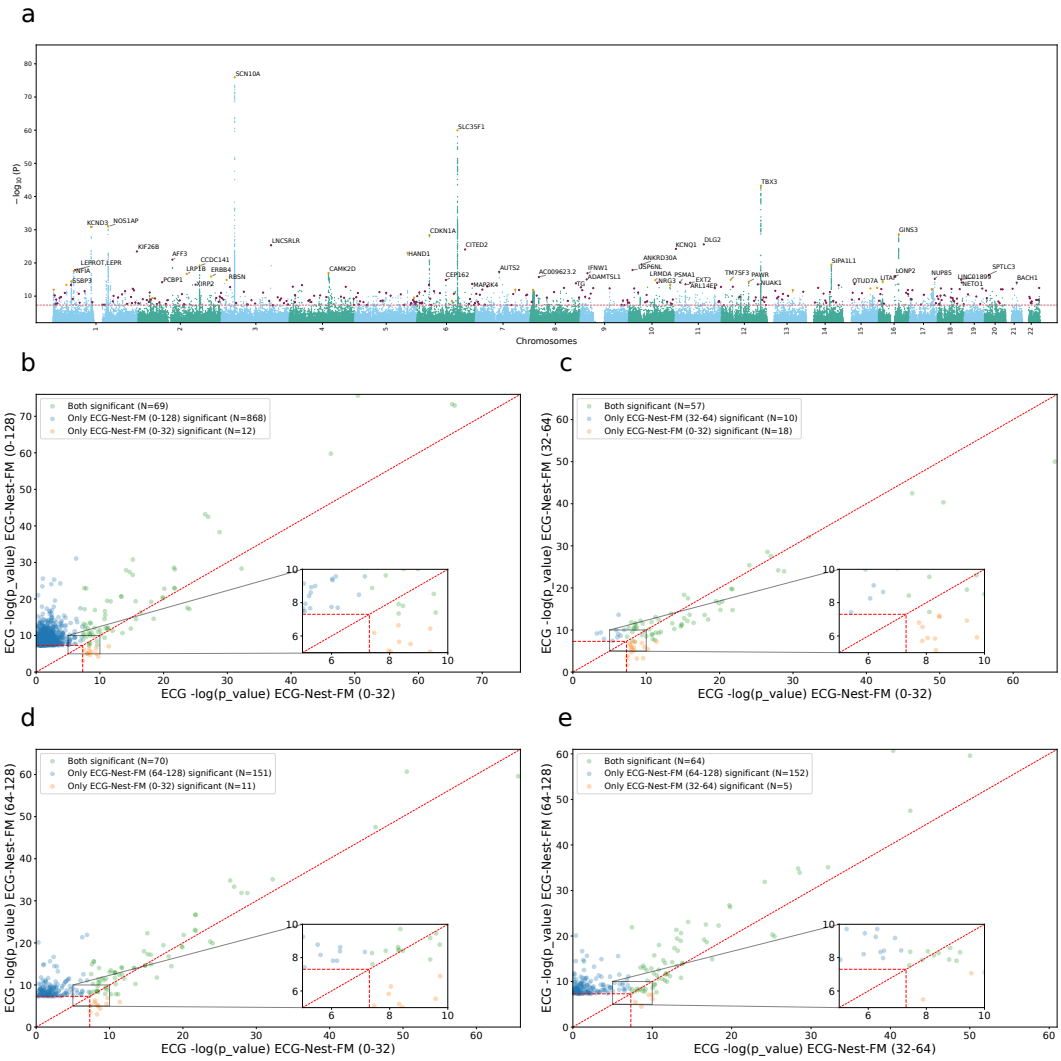


Figure 4: ECG-Nest-FM representations enhance genetic discovery. a) Manhattan plot of 128 ECG-Nest-FM representation GWAS. GWAS p-values for all 22 autosomal chromosomes. Black gene names indicate the closest gene for each locus with $-\log_{10} p > 20$. Purple dots denote the GWS loci uniquely detected by all 128 ECG-Nest-FM. Orange dots indicate loci also identified in first 32 ECG-Nest-FM, b) Power comparison between ECG-Nest-FM 0-128 (all) and 0-32 (first 32) representations, and c) Power comparison between ECG-Nest-FM 32-64 and 0-32 representations, d) Power comparison between ECG-Nest-FM 64-128 and 0-32 representations, and e) Power comparison between ECG-Nest-FM 64-128 and 32-64 representations.

ECG-Nest-FM representations enhance genetic discovery To utilize ECG-Nest-FM representations for genetic discovery, we performed genome-wide association study (GWAS). We considered four main GWAS: GWAS on learned representations for the first 32 representations (ECG-Nest-FM-0-32), GWAS on learned representations for 32 to 64 representations (ECG-Nest-FM-32-64), GWAS on learned representations for 64 to 128 representations (ECG-Nest-FM-64-128), and finally GWAS on all learned representations (ECG-Nest-FM-0-12)). To be able to combine the GWAS for a set of desired learned representations, we performed PCA to make ECG-Nest-FM representations uncorrelated Aschard et al. (2014); Zhou et al. (2024) and then combined the GWAS statistics. For more details on GWAS process, see Appendix A.4. We observed that all four GWAS had reasonable genomics inflation (Figure 10, Figure 11, Figure 12, and Figure 13), obtained 55, 48, 194, and, 911 genome-wide significant loci from the ECG-Nest-FM-0-32, ECG-Nest-FM-32-64, ECG-Nest-FM-

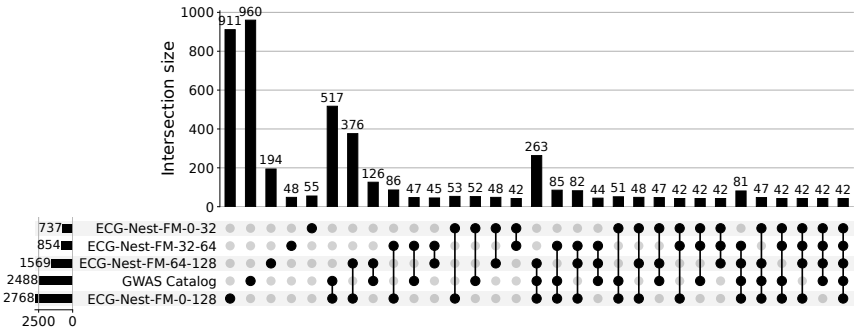


Figure 5: Genetic loci identified using ECG-Nest-FM representations overlap with known loci from the GWAS Catalog related to cardiovascular diseases and traits.

64-128, and ECG-Nest-FM-0-128, respectively (Figure 4). In addition, ECG-Nest-FM-0-128 have non-zero heritability for some learned representations Table 2.

We observed that 52/55, 47/48, 126/194, and 517/911 loci associated with cardiovascular traits were replicated in the GWAS catalog (i.e., have been previously reported in existing literature). Although GWAS on 128 ECG-Nest-FM representations have the highest power compared to 64 and 32, the 32 ECG-Nest-FM representations capture the largest fraction of known cardiovascular disease (Figure 5).

5 Discussion and Conclusion

We proposed a foundation model for learning ECG representations. ECG-Nest-FM disentangles the effects of different frequency bands, thereby providing insights into which frequencies are most useful for a diagnosis. Our model's architecture allowed us to provide empirical evidence that specific frequency bands suffice for specific diagnoses, as the clinical literature suggests. This attribution to frequency bands makes our representations interpretable. Notably, this interpretability does not come at a cost of downstream diagnostic quality of the representations, as the model remained competitive in diagnosing common cardiac conditions. The trade-off between interpretability and predictive performance is a well-known challenge in machine learning (Huysmans et al., 2006; Dziugaite et al., 2020), often hindering the adoption of more interpretable methods. However, it is possible to build performant, interpretable models (Bell et al., 2022), and in the context of ECG foundation modeling, we show that our architectural innovation achieves both (a notion of) interpretability and high diagnostic accuracy .

Limitations and Future Work While ECG-Nest-FM offers improvements in interpretability and maintains competitive diagnostic performance, our approach has limitations that warrant further investigation. One limitation is that while the paper emphasizes interpretability in the frequency domain, certain outcomes have a more natural interpretation in the time domain (Pradhan et al., 2023). Therefore, it would be interesting to learn hybrid representations which can attribute contributions of both frequency- and time-domain features towards a specific diagnosis. A challenge in building such representations is the possible collapse of frequency-domain features if the time-domain features are sufficient to reconstruct the signal (or vice versa) (Havasi et al., 2022). However, this may need rethinking the decorrelation penalty because the frequency- and time-domain features are correlated. Addressing these limitations would improve the model's diagnostic capabilities, as well as provide alternate ways to interpret the representations from the model.

References

Salar Abbaspourazad, Oussama Elachqar, Andrew Miller, Saba Emrani, Udhyakumar Nallasamy, and Ian Shapiro. Large-scale training of foundation models for wearable biosignals. In

The Twelfth International Conference on Learning Representations, 2024. URL https://openreview.net/forum?id=pC3WJHf51j.

M Alventosa-Zaidin, L Guix Font, M Benitez Camps, C Roca Saumell, G Pera, M Teresa Alzamora Sas, R Fores Raurell, Oriol Rebagliato Nadal, Antoni Dalfó-Baqué, and J Brugada Terradellas. Right bundle branch block: prevalence, incidence, and cardiovascular morbidity and mortality in the general population. *European Journal of General Practice*, 25(3):109–115, 2019.

Hugues Aschard, Bjarni J. Vilhjálmsson, Nicolas Greliche, Pierre-Emmanuel Morange, David-Alexandre Trégouët, and Peter Kraft. Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. *The American Journal of Human Genetics*, 94(5):662–676, May 2014. ISSN 0002-9297. doi: 10.1016/j.ajhg.2014.03.016. URL http://dx.doi.org/10.1016/j.ajhg.2014.03.016.

Saira Aziz, Sajid Ahmed, and Mohamed-Slim Alouini. Ecg-based machine-learning algorithms for heartbeat classification. *Scientific reports*, 11(1):18738, 2021.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

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

Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, Avi Schwarzschild, Andrew Gordon Wilson, Jonas Geiping, Quentin Garrido, Pierre Fernandez, Amir Bar, Hamed Pirsiavash, Yann LeCun, and Micah Goldblum. A cookbook of self-supervised learning, 2023. URL https://arxiv.org/abs/2304.12210.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=xm6YD62D1Ub.

Andrew Bell, Ian Solano-Kamaiko, Oded Nov, and Julia Stoyanovich. It's just not that simple: an empirical study of the accuracy-explainability trade-off in machine learning for public policy. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pp. 248–266, 2022.

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

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020a.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020b.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9640–9649, 2021.

Joseph Y Cheng, Hanlin Goh, Kaan Dogrusoz, Oncel Tuzel, and Erdrin Azemi. Subject-aware contrastive learning for biosignals. *arXiv preprint arXiv:2007.04871*, 2020.

Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pp. 539–546. IEEE, 2005.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Gintare Karolina Dziugaite, Shai Ben-David, and Daniel M Roy. Enforcing interpretability and its statistical impacts: Trade-offs between accuracy and interpretability. *arXiv preprint arXiv:2010.13764*, 2020.

Mohammad Ennab and Hamid Mcheick. Designing an interpretability-based model to explain the artificial intelligence algorithms in healthcare. *Diagnostics*, 12(7):1557, 2022.

Sam F Friedman, Shaan Khurshid, Rachael A Venn, Xin Wang, Nate Diamant, Paolo Di Achille, Lu-Chen Weng, Seung Hoan Choi, Christopher Reeder, James P Pirruccello, et al. Unsupervised deep learning of electrocardiograms enables scalable human disease profiling. *npj Digital Medicine*, 8 (1):23, 2025.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann Lecun. Rankme: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *International conference on machine learning*, pp. 10929–10974. PMLR, 2023.

Bryan Gopal, Ryan Han, Gautham Raghupathi, Andrew Ng, Geoff Tison, and Pranav Rajpurkar. 3kg: Contrastive learning of 12-lead electrocardiograms using physiologically-inspired augmentations. In *Machine Learning for Health*, pp. 156–167. PMLR, 2021.

Brian Gow, Tom Pollard, Larry A. Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Jonathan W. Waks, Parastou Eslami, Tanner Carbonati, Ashish Chaudhari, Elizabeth Herbst, Dana Moukheiber, Seth Berkowitz, Roger Mark, and Steven Horng. MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset, 2023. URL `https://doi.org/10.13026/4nqg-sb35`.

Y. Hafeez and S. A. Grossman. Sinus bradycardia. URL `https://www.ncbi.nlm.nih.gov/books/NBK493201/`. [Updated 2023 Aug 7].

Yu Han, Xiaofeng Liu, Xiang Zhang, and Cheng Ding. Foundation models in electrocardiogram: A review. *arXiv preprint arXiv:2410.19877*, 2024.

Marton Havasi, Sonali Parbhoo, and Finale Doshi-Velez. Addressing leakage in concept bottleneck models. *Advances in Neural Information Processing Systems*, 35:23386–23397, 2022.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Yusong Hu, Yantao Zhao, Jihong Liu, Jin Pang, Chen Zhang, and Peizhe Li. An effective frequency-domain feature of atrial fibrillation based on time–frequency analysis. *BMC Medical Informatics and Decision Making*, 20:1–11, 2020.

Daniela Husser, David S Cannom, Anil K Bhandari, Martin Stridh, Leif Sörnmo, S Bertil Olsson, and Andreas Bollmann. Electrocardiographic characteristics of fibrillatory waves in new-onset atrial fibrillation. *Europace*, 9(8):638–642, 2007.

Johan Huysmans, Bart Baesens, and Jan Vanthienen. Using rule extraction to improve the comprehensibility of predictive models. 2006.

Anastasiya Kiseleva, Dimitris Kotzinos, and Paul De Hert. Transparency of ai in healthcare as a multilayered system of accountabilities: between legal requirements and technical limitations. *Frontiers in artificial intelligence*, 5:879603, 2022.

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

Paul Kligfield, Leonard S Gettes, James J Bailey, Rory Childers, Barbara J Deal, E William Hancock, Gerard Van Herpen, Jan A Kors, Peter Macfarlane, David M Mirvis, et al. Recommendations for the standardization and interpretation of the electrocardiogram: part i: the electrocardiogram and its technology: a scientific statement from the american heart association electrocardiography and arrhythmias committee, council on clinical cardiology; the american college of cardiology foundation; and the heart rhythm society endorsed by the international society for computerized electrocardiology. *Circulation*, 115(10):1306–1324, 2007.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, et al. Matryoshka representation learning. *Advances in Neural Information Processing Systems*, 35:30233–30249, 2022.

Kenneth A Mayuga, Artur Fedorowski, Fabrizio Ricci, Rakesh Gopinathannair, Jonathan Walter Dukes, Christopher Gibbons, Peter Hanna, Dan Sorajja, Mina Chung, David Benditt, et al. Sinus tachycardia: a multidisciplinary expert focused review. *Circulation: Arrhythmia and Electrophysiology*, 15(9):e007960, 2022.

Joelle Mbatchou, Leland Barnard, Joshua Backman, Anthony Marcketta, Jack A. Kosmicki, Andrey Ziyatdinov, Christian Benner, Colm O'Dushlaine, Mathew Barber, Boris Boutkov, Lukas Habegger, Manuel Ferreira, Aris Baras, Jeffrey Reid, Goncalo Abecasis, Evan Maxwell, and Jonathan Marchini. Computationally efficient whole-genome regression for quantitative and binary traits. *Nature Genetics*, 53(7):1097–1103, May 2021. ISSN 1546-1718. doi: 10.1038/s41588-021-00870-7. URL http://dx.doi.org/10.1038/s41588-021-00870-7.

Kaden McKeen, Laura Oliva, Sameer Masood, Augustin Toma, Barry Rubin, and Bo Wang. Ecg-fm: An open electrocardiogram foundation model. *arXiv preprint arXiv:2408.05178*, 2024.

Temesgen Mehari and Nils Strodthoff. Self-supervised representation learning from 12-lead ecg data. *Computers in biology and medicine*, 141:105114, 2022.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

Mark J. Niebauer, John Rickard, Landon Polakof, Patrick J. Tchou, and Niraj Varma. Qrs frequency characteristics help predict response to cardiac resynchronization in left bundle branch block less than 150 milliseconds. *Heart Rhythm*, 11(12):2183–2189, 2014. ISSN 1547-5271. doi: https://doi.org/10.1016/j.hrthm.2014.07.034. URL https://www.sciencedirect.com/science/article/pii/S1547527114008078. Focus Issue: Devices.

Jungwoo Oh, Hyunseung Chung, Joon-myoung Kwon, Dong-gyun Hong, and Edward Choi. Lead-agnostic self-supervised learning for local and global representations of electrocardiogram. In *Conference on Health, Inference, and Learning*, pp. 338–353. PMLR, 2022.

Paxml. *Paxml: a Jax-based machine learning framework for training large scale models*, 2022. https://github.com/google/paxml [Accessed: 2025-02-11].

Garrett Perkins, Chase McGlinn, Muhammad Rizwan, and Bradley M Whitaker. Detecting cardiac abnormalities from 12-lead ecg signals using feature selection, feature extraction, and machine learning classification. In *2020 Computing in Cardiology*, pp. 1–4. IEEE, 2020.

Tiago Pimentel, Naomi Saphra, Adina Williams, and Ryan Cotterell. Pareto probing: Trading off accuracy for complexity. *arXiv preprint arXiv:2010.02180*, 2020a.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. Information-theoretic probing for linguistic structure. *arXiv preprint arXiv:2004.03061*, 2020b.

Bikash K. Pradhan, Bala Chakravarty Neelappu, J. Sivaraman, Doman Kim, and Kunal Pal. A review on the applications of time-frequency methods in ecg analysis. *Journal of Healthcare Engineering*, 2023(1):3145483, 2023. doi: https://doi.org/10.1155/2023/3145483. URL https://onlinelibrary.wiley.com/doi/abs/10.1155/2023/3145483.

A. H. Ribeiro, G. M. M. Paixao, E. M. Lima, M. Horta Ribeiro, M. M. Pinto Filho, P. R. Gomes, D. M. Oliveira, W. Meira Jr, T. B. Schon, and A. L. P. Ribeiro. CODE-15%: a large scale annotated dataset of 12-lead ECGs, 2021. URL https://doi.org/10.5281/zenodo.4916206.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pp. 606–610. IEEE, 2007.

Tong Ruan, Liqi Lei, Yangming Zhou, Jie Zhai, Le Zhang, Ping He, and Ju Gao. Representation learning for clinical time series prediction tasks in electronic health records. *BMC medical informatics and decision making*, 19:1–14, 2019.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International journal of forecasting*, 36(3):1181–1191, 2020.

Arunashis Sau, Libor Pastika, Ewa Sieliwonczyk, Konstantinos Patlatzoglou, Antonio H Ribeiro, Kathryn A McGurk, Boroumand Zeidaabadi, Henry Zhang, Krzysztof Macierzanka, Danilo Mandic, et al. Artificial intelligence enabled electrocardiogram for mortality and cardiovascular risk estimation: An actionable, explainable and biologically plausible platform. *medRxiv*, pp. 2024–01, 2024.

Abhishek Sharma, Pilar F Verhaak, Thomas H McCoy, Roy H Perlis, and Finale Doshi-Velez. Identifying data-driven subtypes of major depressive disorder with electronic health records. *Journal of Affective Disorders*, 356:64–70, 2024.

Junho Song, Jong-Hwan Jang, Byeong Tak Lee, DongGyun Hong, Joon-myoung Kwon, and Yong-Yeon Jo. Foundation models for ecg: Leveraging hybrid self-supervised learning for advanced cardiac diagnostics. *arXiv preprint arXiv:2407.07110*, 2024.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.

Larisa G Tereshchenko and Mark E Josephson. Frequency content and characteristics of ventricular conduction. *Journal of electrocardiology*, 48(6):933–937, 2015.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 11121–11128, 2023.

Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency. *Advances in Neural Information Processing Systems*, 35:3988–4003, 2022.

Yuchen Zhou, Justin Cosentino, Taedong Yun, Mahantesh I. Biradar, Jacqueline Shreibati, Dongbing Lai, Tae-Hwi Schwantes-An, Robert Luben, Zachary McCaw, Jorgen Engmann, Rui Providencia, Amand Floriaan Schmidt, Patricia Munroe, Howard Yang, Andrew Carroll, Anthony P. Khawaja, Cory Y. McLean, Babak Behsaz, and Farhad Hormozdiari. Utilizing multimodal ai to improve genetic analyses of cardiovascular traits. March 2024. doi: 10.1101/2024.03.19.24304547. URL http://dx.doi.org/10.1101/2024.03.19.24304547.

Ala'a Zyout, Hiam Alquran, Wan Azani Mustafa, and Ali Mohammad Alqudah. Advanced time-frequency methods for ecg waves recognition. *Diagnostics*, 13(2):308, 2023.

A Appendix

A.1 Dataset details

We describe the labels in the CODE-15% and MIMIC-IV-ECG datasets in Table 1.

Table 1: Labels in ECG dataset

Dataset	Label	Description
CODE-15%	is_male	Indicates whether the patient is male.
CODE-15%	1st degree AV block (1dAVb)	Presence of a first degree atrioventricular block.
CODE-15%	right bundle branch block (RBBB)	Presence of a right bundle branch block.
CODE-15%	left bundle branch block (LBBB)	Presence of a left bundle branch block.
CODE-15%	sinus bradycardia (SB)	Presence of sinus bradycardia.
CODE-15%	sinus tachycardia (ST)	Presence of sinus tachycardia.
CODE-15%	atrial fibrillation (AFib)	Presence of atrial fibrillation.
CODE-15%	normal_ecg	Indicates a normal electrocardiogram.
MIMIC-IV-ECG	abnormal_ecg	Patient has an overall abnormal electrocardiogram.
MIMIC-IV-ECG	atrial_fibrillation	Patient exhibits an irregular, often rapid heart rhythm.
MIMIC-IV-ECG	left_axis_deviation	The electrical axis of the heart is shifted leftward.
MIMIC-IV-ECG	left_ventricular_hypertrophy	Patient's left ventricle is enlarged.
MIMIC-IV-ECG	right_ventricular_hypertrophy	Patient's right ventricle is enlarged.
MIMIC-IV-ECG	sinus_arrhythmia	Patient has a sinus rhythm with irregular heart rate.
MIMIC-IV-ECG	sinus_bradycardia	Patient has a slow sinus rhythm (below 60 bpm).
MIMIC-IV-ECG	sinus_rhythm	Patient exhibits a normal sinus rhythm.
MIMIC-IV-ECG	sinus_tachycardia	Patient has a fast sinus rhythm (above 100 bpm).

A.2 Training details

VICReg Correlation Penalty To enforce consistency between the two representation views $\boldsymbol{Y}^l$ and $\boldsymbol{Y}^r$ while preventing representation collapse, we incorporate a loss introduced on the embeddings $\boldsymbol{z}^l, \boldsymbol{z}^r$ (Bardes et al., 2022). We denote $\boldsymbol{Z}^l = [\boldsymbol{z}_1^l, \ldots, \boldsymbol{z}_B^l]$ and $\boldsymbol{Z}^r = [\boldsymbol{z}_1^r, \ldots, \boldsymbol{z}_B^r]$ be the two batches composed of B embedding vectors of dimension D_e.

$$L_{\text{VICReg}}(\boldsymbol{Z}^l, \boldsymbol{Z}^r) = \lambda s(\boldsymbol{Z}^l, \boldsymbol{Z}^r) + \mu \left(v(\boldsymbol{Z}^l) + v(\boldsymbol{Z}^r)\right) + \eta \left(c(\boldsymbol{Z}^l) + c(\boldsymbol{Z}^r)\right) \tag{3}$$

where the terms are defined as:

$$s(\boldsymbol{Z}^l, \boldsymbol{Z}^r) = \frac{1}{B} \sum_{b=1}^{B} ||\boldsymbol{z}_b^l - \boldsymbol{z}_b^r||^2 \tag{4}$$

$$v(\boldsymbol{Z}) = \frac{1}{D_e} \sum_{k=1}^{D_e} \max\left(0, 1 - \sqrt{[\text{Cov}(\boldsymbol{Z})]_{kk} + \epsilon}\right) \tag{5}$$

$$c(\boldsymbol{Z}) = \frac{1}{D_e} \sum_{i \neq j} [\text{Cov}(\boldsymbol{Z})]_{i,j}^2 \tag{6}$$

where ϵ is a small constant to avoid numerical instabilities and $\text{Cov}(\boldsymbol{Z}) \in \mathbb{R}^{D_e \times D_e}$ is the covariance matrix of embeddings $\boldsymbol{z}$ estimated from the current batch (computed for each view separately):

$$\text{Cov}(\boldsymbol{Z}) = \frac{1}{B-1} \sum_{b=1}^{B} (\boldsymbol{z}_b - \bar{\boldsymbol{z}})(\boldsymbol{z}_b - \bar{\boldsymbol{z}})^\top, \text{ where } \bar{\boldsymbol{z}} = \sum_{b=1}^{B} \boldsymbol{z}_b \tag{7}$$

The similarity term $s(\boldsymbol{Z}^l, \boldsymbol{Z}^r)$, encourages similarity between representations of two views of the same input, the variance term $v(\boldsymbol{Z})$ helps maintain sufficient variation in each latent dimension to avoid representation collapse (i.e., trivial or constant representations), and the covariance term, $c(\boldsymbol{Z})$, acts as a decorrelation penalty so that different dimensions in the representation are not redundant.

Optimizer We use the Adam optimizer, employing an inverse-decay learning rate schedule:

$$\alpha(t) = \frac{\alpha_0}{\sqrt{D}} \min\left(\frac{t+1}{T_{\text{warmup}}^{3/2}}, \frac{1}{\sqrt{t+1}}\right) \tag{8}$$

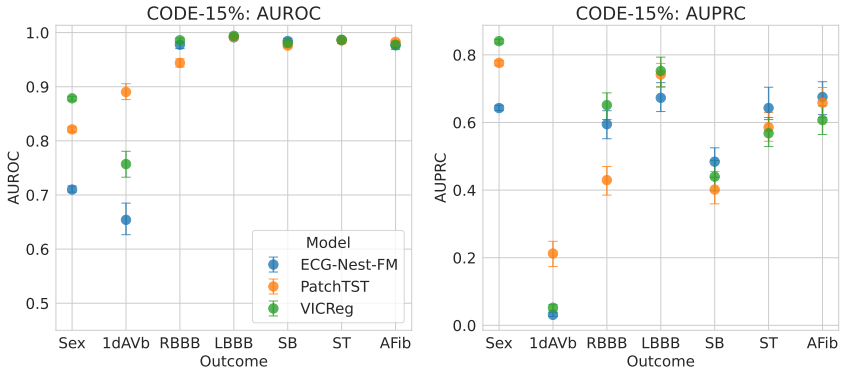


Figure 6: Non-linear probing results (AUROC and AUPRC) of the foundation models for different clinical outcomes predicted on the held-out CODE-15% dataset ($n = 26,264$ ECGs). Error bars indicate the 95% confidence intervals from 100 bootstrap samples. ECG-Nest-FM demonstrates comparable performance across most cardiac outcomes. Prevalence of outcomes: Sex (M): 41.2%, 1dAVb: 1.6%, RBBB: 2.2%, LBBB: 1.8%, SB: 1.6%, ST: 2.1%, AF: 2.2%.

where $\alpha_0 = 0.5$ is the base learning rate, $T_{\text{warmup}} = 4000$ is the number of warmup steps, and $D = 128$ is the transformer encoder dimension. After warmup, the learning rate decays according to the inverse square root of the current step, mitigating overfitting and ensuring stable convergence.

Implementation Details We implement our model in the Pax (Paxml, 2022) framework and train on TPUs for scalable computation. The training runs for a total of 2 million steps with a batch size of 1,024.

A.3 Supplementary results

Here, we present additional analyses to further show the behavior of the learned representations across different embedding dimensionalities and probing methods on both the MIMIC-IV-ECG and CODE-15% datasets. For the MIMIC-IV-ECG dataset, we include two sets of experiments. The first set, illustrated in Figure 8, employs a non-linear probe to compare the diagnostic performance of 32-dimensional sub-embeddings against their 64- and 128-dimensional counterparts using AUROC, AUPRC, cross-correlation, and smooth effective rank (SER) metrics. The second set, shown in Figure 9, presents similar comparisons using a linear probe (using ridge regression). For the CODE-15% dataset, Figure 7 summarizes the linear probing results, analogous to the Figures 3, 2 and 6 in the main paper. Together, these supplementary analyses provide a comprehensive view of how embedding dimensionality and probe selection influence representation quality and downstream diagnostic performance.

A.4 Genomics-wide association study (GWAS)

We generated ECG-Nest-FM representations for the 12-lead ECG using the combined training and validation data. To achieve uncorrelated coordinates, we applied principal component analysis (PCA) to each set of representations. First, genome-wide association study (GWAS) were then conducted on each PCA-ed ECG-Nest-FM representations using REGENIE Mbatchou et al. (2021). We adjusted GWAS for age, sex, body mass index, standing height, genotyping array, and the top 15 genetic principal components.

To obtain an overall result of genetic discovery on the multimodal representations, we combined the PC GWAS. We summed the chi-square statistics for each GWAS and computed the combined p-value using the combined chi-square statistic, with the number of phenotypes as the degrees of freedom. This combined chi-square statistic provided the final GWAS result used in this work Aschard et al. (2014).

We consider hits as independent genome-wide significant variants ($R^2 \leq 0.1$ and $P \leq 5 \times 10^{-8}$) and loci were obtained by merging hits within 250 kb.

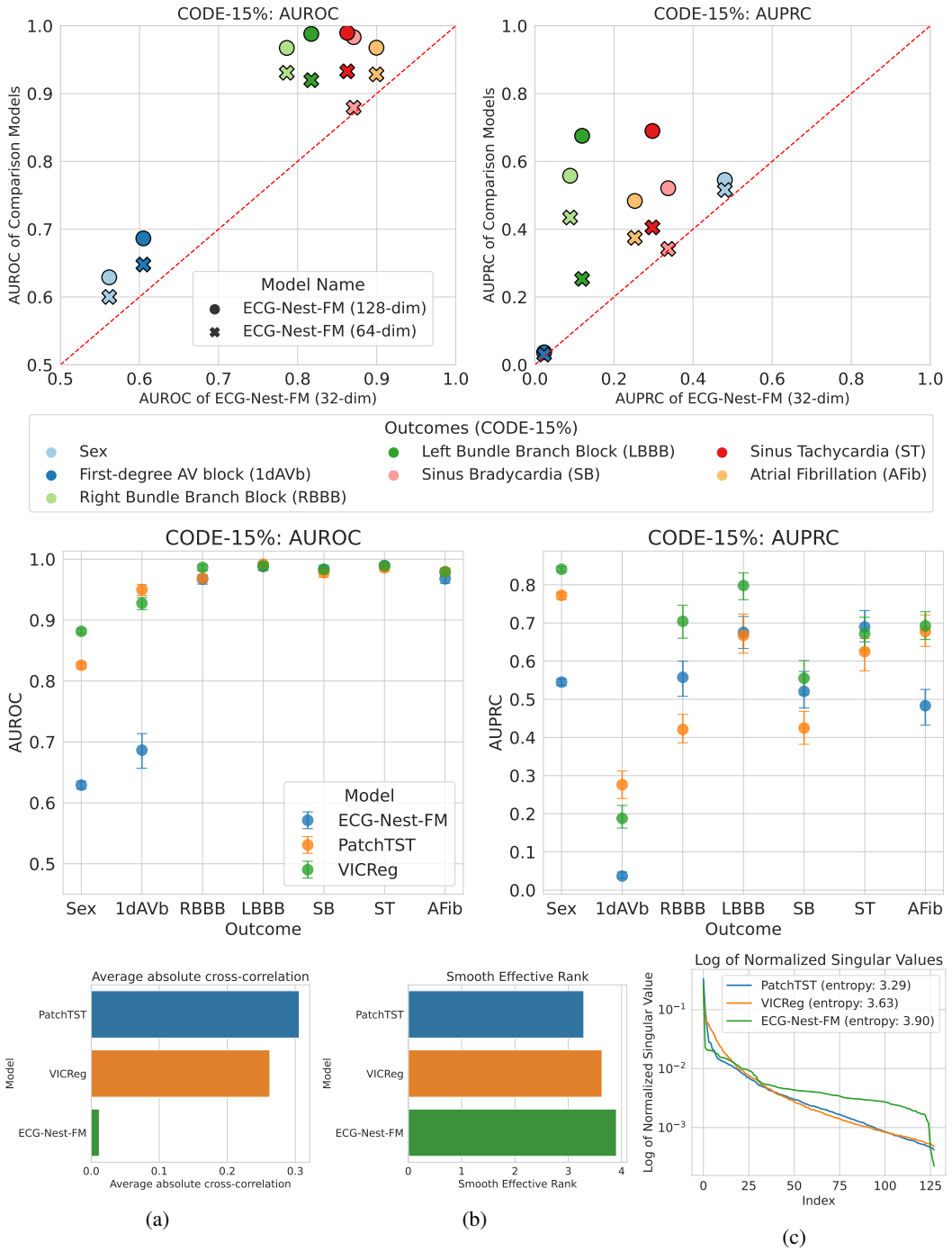


Figure 7: Linear probing results and avg-cross correlations and smooth effective ranks for representations on the CODE-15% dataset.

Representations	S-LDSC Intercept	S-LDSC SNP-heritability
Latent-0	0.9853 (0.0106)	0.1929 (0.0275)
Latent-1	1.0218 (0.0088)	0.0577 (0.0218)
Latent-2	0.9885 (0.0096)	0.005 (0.0257)
Latent-3	0.9987 (0.0117)	0.0041 (0.0298)

Latent-4	1.0059 (0.0083)	0.0131 (0.0205)
Latent-5	1.0027 (0.01)	-0.0028 (0.0238)
Latent-6	0.9992 (0.0094)	0.0326 (0.0229)
Latent-7	0.9943 (0.0091)	0.018 (0.0226)
Latent-8	1.0026 (0.0107)	0.1218 (0.0256)
Latent-9	1.0019 (0.0089)	0.0048 (0.0225)
Latent-10	0.9963 (0.0091)	-0.0016 (0.021)
Latent-11	0.9949 (0.0103)	0.1292 (0.0262)
Latent-12	0.9909 (0.014)	0.0682 (0.0365)
Latent-13	1.0031 (0.0092)	-0.0079 (0.022)
Latent-14	1.0079 (0.0088)	0.0133 (0.0219)
Latent-15	1.0094 (0.0092)	-0.0158 (0.0231)
Latent-16	0.995 (0.009)	0.0183 (0.0214)
Latent-17	1.0037 (0.0093)	0.0043 (0.0213)
Latent-18	1.0032 (0.009)	-0.0046 (0.0229)
Latent-19	0.9963 (0.0092)	0.0162 (0.0228)
Latent-20	0.9986 (0.0126)	0.0247 (0.0335)
Latent-21	1.0032 (0.0093)	0.0247 (0.0227)
Latent-22	0.9948 (0.0152)	0.0702 (0.0435)
Latent-23	1.0122 (0.0089)	-0.0149 (0.0217)
Latent-24	0.9995 (0.0096)	0.0163 (0.0221)
Latent-25	0.9872 (0.0095)	0.0405 (0.023)
Latent-26	0.9938 (0.0102)	0.0836 (0.0245)
Latent-27	1.0043 (0.0098)	0.0004 (0.0268)
Latent-28	0.9948 (0.0083)	0.0223 (0.0217)
Latent-29	0.9958 (0.0089)	0.0172 (0.0214)
Latent-30	1.0028 (0.009)	0.0059 (0.0242)
Latent-31	1.0082 (0.0093)	-0.0121 (0.0208)
Latent-32	1.0111 (0.0091)	0.0098 (0.0211)
Latent-33	1.0082 (0.0093)	-0.0078 (0.0218)
Latent-34	0.9981 (0.0082)	0.0278 (0.0219)
Latent-35	0.9939 (0.0118)	0.0275 (0.0292)
Latent-36	0.9888 (0.0126)	0.0593 (0.0333)
Latent-37	1.0076 (0.0094)	0.0323 (0.0227)
Latent-38	1.003 (0.0092)	0.0577 (0.0239)
Latent-39	0.9796 (0.0176)	0.1229 (0.0501)
Latent-40	0.9826 (0.0096)	0.0515 (0.0211)
Latent-41	0.9965 (0.0087)	0.0364 (0.0231)
Latent-42	0.9828 (0.0101)	0.0613 (0.028)
Latent-43	0.9987 (0.0084)	-0.0026 (0.0218)
Latent-44	1.0094 (0.0102)	-0.0038 (0.0264)
Latent-45	0.996 (0.0088)	0.01 (0.0194)
Latent-46	0.9774 (0.0097)	0.0703 (0.0241)
Latent-47	0.9861 (0.0104)	0.0394 (0.0256)
Latent-48	0.9775 (0.0098)	0.113 (0.0205)
Latent-49	0.9984 (0.0088)	0.022 (0.0202)
Latent-50	1.0053 (0.0091)	0.0142 (0.0218)
Latent-51	0.9905 (0.009)	0.028 (0.0217)
Latent-52	1.0116 (0.0097)	-0.0225 (0.0225)
Latent-53	1.0015 (0.0087)	0.002 (0.0209)
Latent-54	1.0027 (0.0087)	-0.0008 (0.0192)
Latent-55	1.0063 (0.0092)	-0.0132 (0.0196)
Latent-56	1.007 (0.0089)	0.0036 (0.0218)
Latent-57	1.0021 (0.009)	0.0161 (0.0217)
Latent-58	0.9913 (0.009)	0.0286 (0.0224)
Latent-59	0.9945 (0.0088)	0.0433 (0.0226)
Latent-60	0.9938 (0.0086)	0.016 (0.0218)
Latent-61	0.9971 (0.0094)	0.013 (0.0232)
Latent-62	1.0083 (0.0083)	0.0144 (0.0212)

Latent-63	1.0048 (0.0099)	0.0223 (0.025)
Latent-64	0.9925 (0.0081)	0.0211 (0.0198)
Latent-65	0.9999 (0.0095)	0.0311 (0.0229)
Latent-66	1.0148 (0.0093)	-0.0197 (0.0217)
Latent-67	0.9981 (0.0097)	0.0196 (0.0226)
Latent-68	0.9929 (0.0098)	0.0343 (0.0223)
Latent-69	0.9933 (0.0102)	0.0278 (0.0254)
Latent-70	0.9926 (0.0085)	0.0295 (0.021)
Latent-71	1.0036 (0.0091)	-0.0081 (0.0226)
Latent-72	1.0071 (0.0098)	-0.0187 (0.0234)
Latent-73	1.0002 (0.0086)	0.0257 (0.0218)
Latent-74	1.0064 (0.0087)	0.0014 (0.0229)
Latent-75	0.9922 (0.0091)	0.0411 (0.023)
Latent-76	1.0137 (0.0091)	-0.013 (0.0206)
Latent-77	0.9966 (0.0102)	0.0204 (0.0255)
Latent-78	0.9995 (0.008)	-0.0034 (0.0193)
Latent-79	0.9963 (0.008)	0.0091 (0.0204)
Latent-80	1.013 (0.0084)	-0.017 (0.0204)
Latent-81	0.9996 (0.009)	0.0369 (0.0217)
Latent-82	0.9969 (0.0088)	0.0191 (0.0211)
Latent-83	0.982 (0.0149)	0.0824 (0.0399)
Latent-84	0.9933 (0.0092)	0.0298 (0.0218)
Latent-85	1.012 (0.0088)	-0.0074 (0.0218)
Latent-86	1.0082 (0.0089)	-0.0129 (0.0226)
Latent-87	0.9951 (0.0088)	-0.0021 (0.0219)
Latent-88	0.9945 (0.0093)	0.0366 (0.0225)
Latent-89	0.9993 (0.0141)	0.0264 (0.0382)
Latent-90	1.0039 (0.0085)	-0.01 (0.0212)
Latent-91	1.0116 (0.0089)	-0.0081 (0.0227)
Latent-92	1.0183 (0.0101)	-0.0362 (0.0252)
Latent-93	1.0054 (0.0091)	-0.0073 (0.0221)
Latent-94	1.0032 (0.0089)	0.0083 (0.0218)
Latent-95	1.0039 (0.009)	-0.0164 (0.0207)
Latent-96	0.9981 (0.0092)	0.0206 (0.0245)
Latent-97	1.0097 (0.0093)	-0.016 (0.0209)
Latent-98	0.999 (0.0086)	-0.0024 (0.0219)
Latent-99	0.9962 (0.0085)	0.0178 (0.0231)
Latent-100	1.0112 (0.0082)	-0.0062 (0.0208)
Latent-101	1.0105 (0.0092)	-0.0165 (0.0221)
Latent-102	1.0165 (0.0108)	-0.0162 (0.0256)
Latent-103	0.9952 (0.0086)	0.0047 (0.0222)
Latent-104	0.9935 (0.0094)	0.0114 (0.0224)
Latent-105	1.0019 (0.0091)	-0.0076 (0.0225)
Latent-106	1.009 (0.0093)	-0.0113 (0.0223)
Latent-107	0.9908 (0.0083)	0.0116 (0.0227)
Latent-108	1.0044 (0.0095)	0.0004 (0.0229)
Latent-109	0.9876 (0.008)	0.0131 (0.021)
Latent-110	0.9882 (0.0102)	0.0264 (0.0232)
Latent-111	0.99 (0.0106)	0.0367 (0.0246)
Latent-112	1.0026 (0.0088)	-0.0146 (0.0237)
Latent-113	1.0133 (0.0095)	-0.0327 (0.0233)
Latent-114	1.0041 (0.0084)	0.005 (0.0227)
Latent-115	1.0132 (0.0091)	-0.0098 (0.0225)
Latent-116	0.9883 (0.0088)	0.0235 (0.0205)
Latent-117	1.0006 (0.0093)	0.0094 (0.0224)
Latent-118	1.0013 (0.0085)	0.0099 (0.0214)
Latent-119	1.01 (0.0086)	-0.0309 (0.0222)
Latent-120	0.9875 (0.0087)	0.0095 (0.0211)
Latent-121	1.0017 (0.0113)	-0.0004 (0.028)

Latent-122	0.9997 (0.0098)	0.0018 (0.0232)
Latent-123	0.9977 (0.0086)	0.0184 (0.0222)
Latent-124	1.0019 (0.0087)	0.0201 (0.0206)
Latent-125	1.0062 (0.0097)	-0.0052 (0.023)
Latent-126	0.9956 (0.0089)	0.0187 (0.0213)
Latent-127	0.9885 (0.0087)	0.0291 (0.023)

Table 2: S-LDSC intercept and SNP-heritability for ECG-Nest-FM-0-128 GWAS. In LD score regression (S-LDSC), the intercept is a crucial diagnostic measure for assessing potential biases and confounding in GWAS results. Ideally, we want the S-LDSC intercept to be as close to 1 as possible.

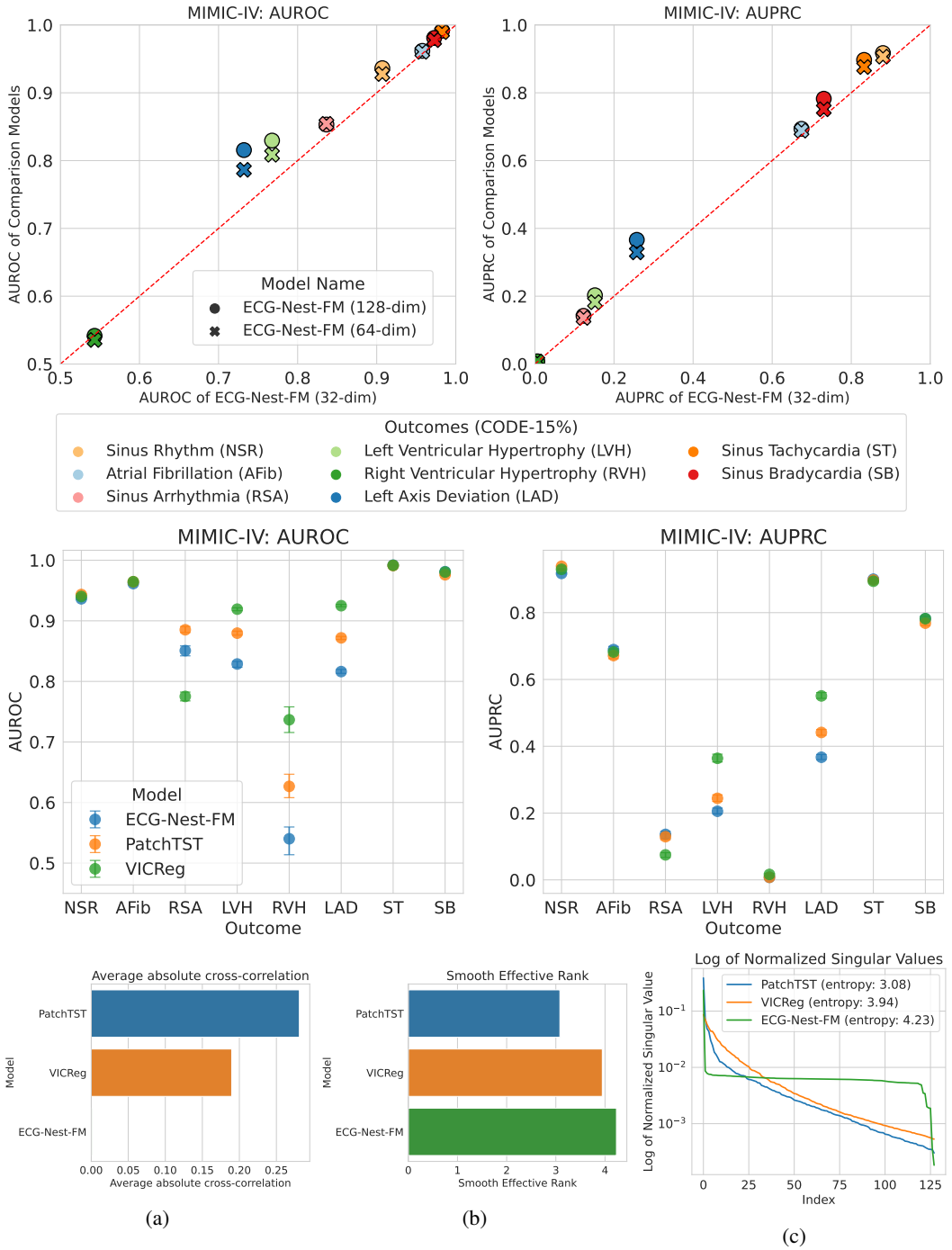


Figure 8: Non-linear probing results and avg-cross correlations and smooth effective ranks for representations on the MIMIC-IV-ECG dataset.

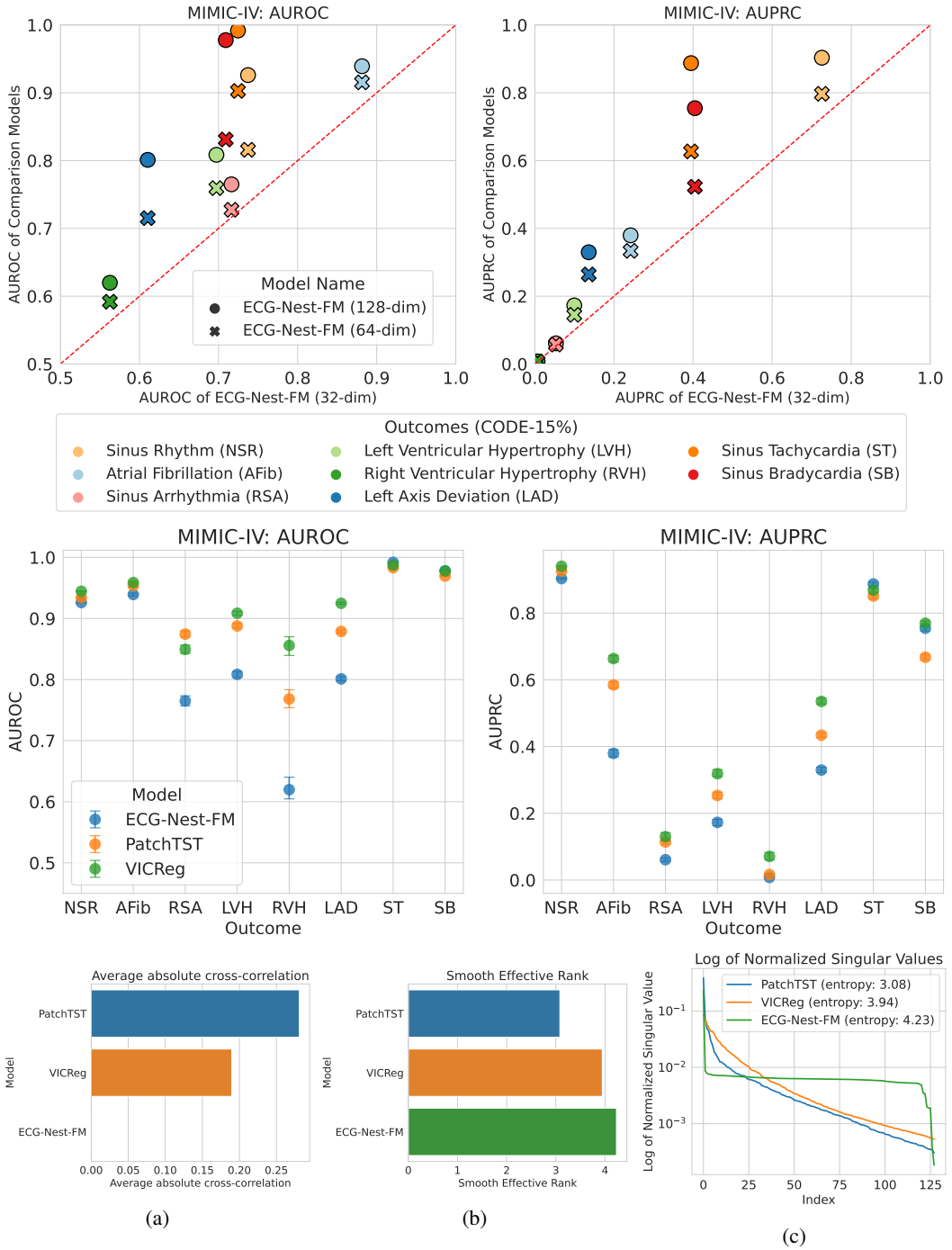


Figure 9: Linear probing results and avg-cross correlations and smooth effective ranks for representations on the MIMIC-IV-ECG dataset.

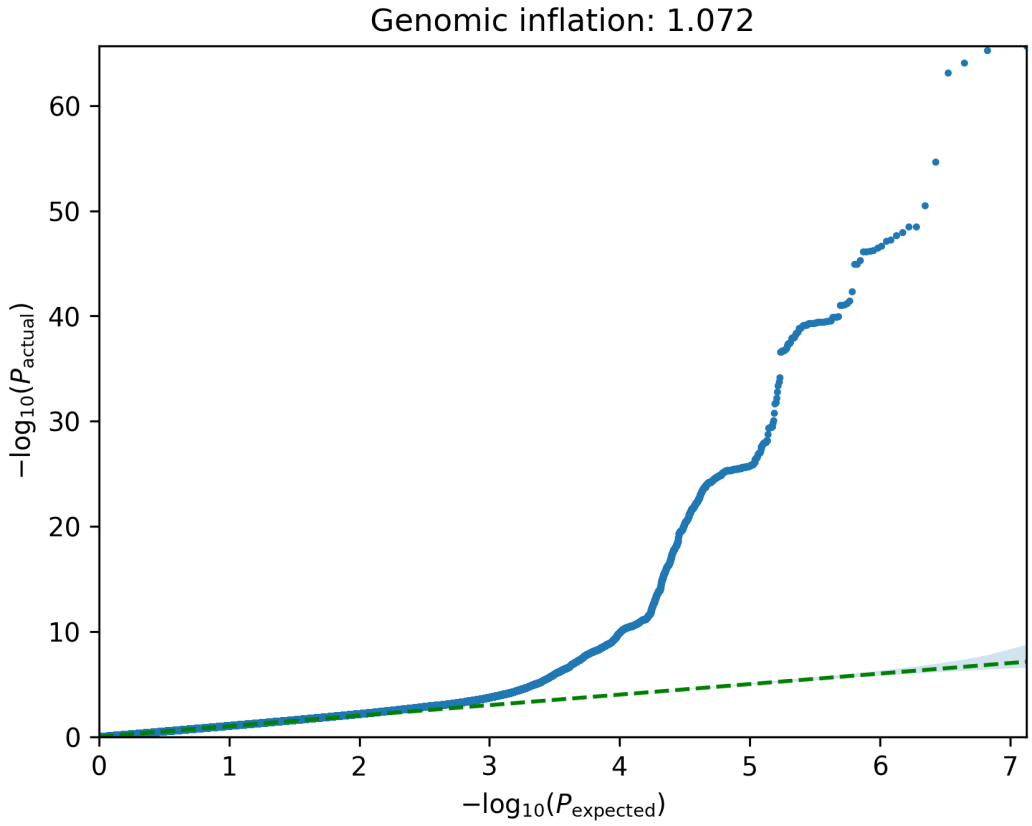


Figure 10: **QQ-plot ECG-Nest-FM combined on all 32 representations obtained from 12 lead ECG.**

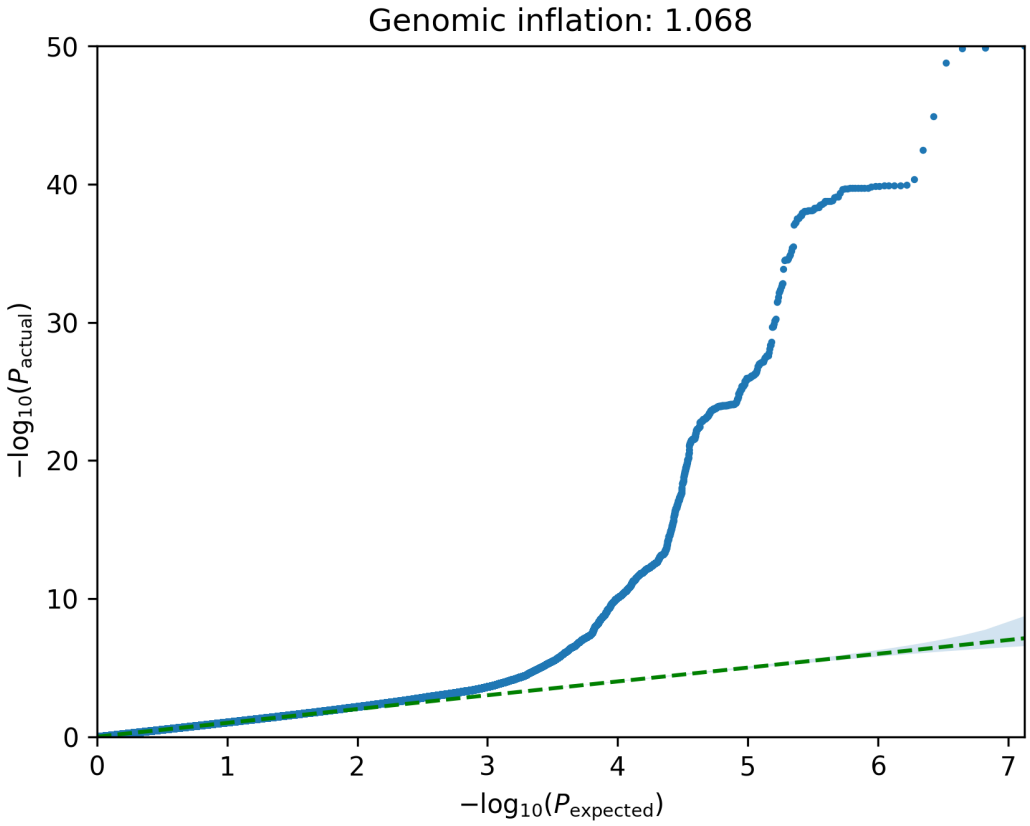


Figure 11: **QQ-plot ECG-Nest-FM combined on all 32 to 64 representations obtained from 12 lead ECG.**

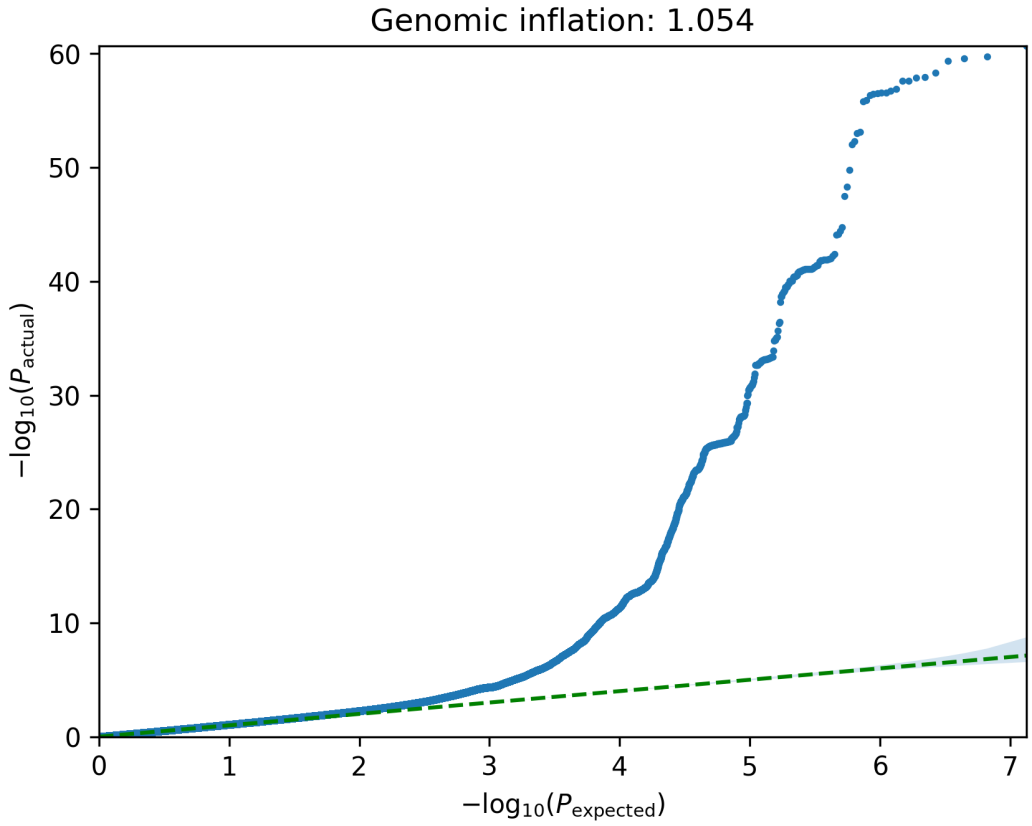


Figure 12: **QQ-plot ECG-Nest-FM combined on all 64 to 128 representations obtained from 12 lead ECG.**

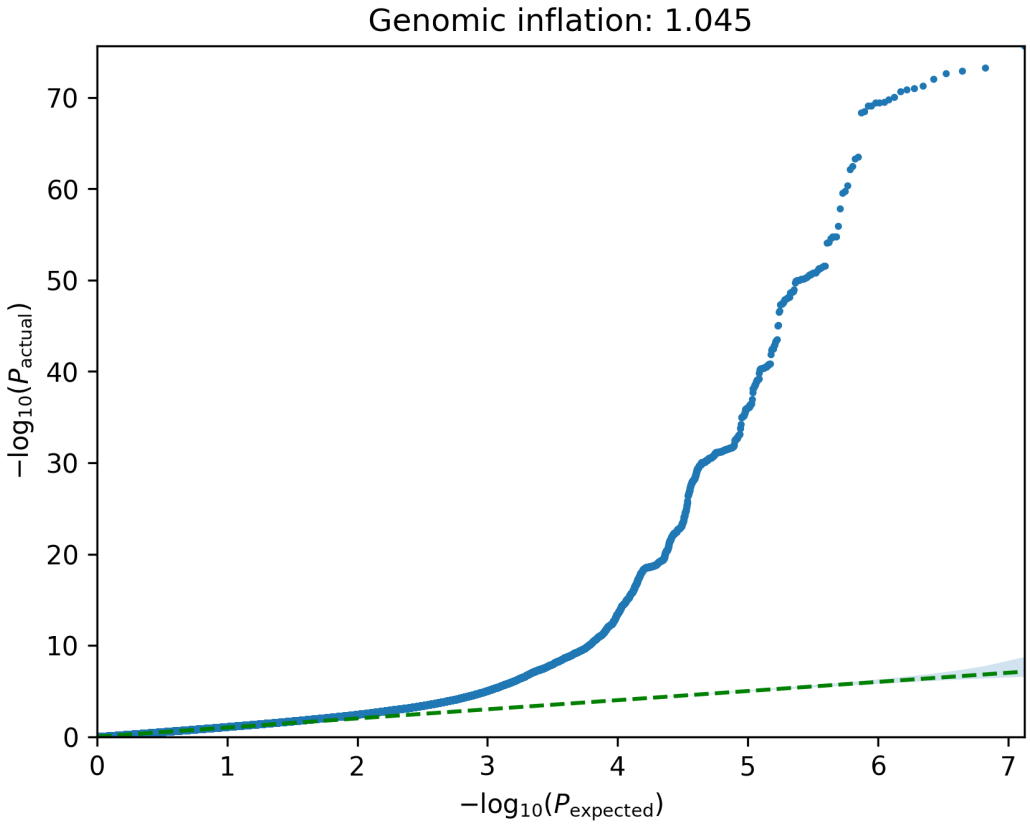


Figure 13: **QQ-plot ECG-Nest-FM combined on all 128 representations obtained from 12 lead ECG.**