WAVELET-DRIVEN MASKED MULTISCALE RECONSTRUCTION FOR PPG FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Wearable foundation models have the potential to transform digital health by learning transferable representations from large-scale biosignals collected in everyday settings. While recent progress has been made in large-scale pretraining, most approaches overlook the spectral structure of photoplethysmography (PPG) signals, wherein physiological rhythms unfold across multiple frequency bands. Motivated by the insight that many downstream health-related tasks depend on multi-resolution features spanning fine-grained waveform morphology to global rhythmic dynamics, we introduce Masked Multiscale Reconstruction (MMR) for PPG representation learning – a self-supervised pretraining framework that explicitly learns from hierarchical time–frequency scales of PPG data. The pretraining task is designed to reconstruct randomly masked out coefficients obtained from a wavelet-based multiresolution decomposition of PPG signals, forcing the transformer encoder to integrate information across temporal and spectral scales. We pretrain our model with MMR using $\sim$17 million unlabeled 10-second PPG segments collected from over $\sim$32000 smartwatch users largely in naturalistic field settings, ensuring high variability and ecological validity. On 11 of 13 diverse health-related tasks, MMR trained on large-scale wearable PPG data outperforms or matches state-of-the-art open-source PPG foundation models, time-series foundation models and other self-supervised baselines. Extensive analysis of our learned embeddings and systematic ablations underscore the value of wavelet-based representations, showing that they capture robust and physiologically-grounded features. Together, these results highlight the potential of MMR as a step toward generalizable PPG foundation models.

1 INTRODUCTION

Foundation models for biosignals remain in their infancy, despite early promise demonstrating their potential to transform health monitoring and biomarker discovery (Abbaspourazad et al., 2024b; Narayanswamy et al., 2024; Erturk et al., 2025; Xu et al., 2025). Among these signals, photoplethysmography (PPG) is uniquely well-suited for self-supervised learning: it is embedded in virtually every consumer wearable, already underpins multiple deployed machine learning models for applications such as blood pressure, arrhythmia, and stress detection (Song et al., 2019; Bashar et al., 2019; Namvari et al., 2022; Apple Inc., 2025), and offers large-scale data for continuous cardiovascular monitoring (Charlton et al., 2022a; Lee & Akamatsu, 2025). Recent PPG foundation models (Abbaspourazad et al., 2024a; Pillai et al., 2024; Saha et al., 2025) have demonstrated that self-supervised pre-training can outperform traditional ML approaches, underscoring the promise of large-scale pretraining paradigms. More broadly, recent time-series self-supervised models have shown that explicitly modeling spectral-domain information improves robustness and transferability of learnt representations (Kara et al., 2024; Fu & Hu, 2025; Zhang et al., 2022a; Liu et al., 2024). However, existing PPG foundation models either focus solely on time-domain data or use frequency-based methods that leverage fixed-window Fourier transform and late fusion, limiting their ability to capture the adaptive, hierarchical time-frequency features of non-stationary physiological signals such as PPG (Chen et al., 2025; Masserano et al., 2025). To address these limitations, we introduce wavelet-based PPG foundation model, which explicitly learns cross-scale interactions in the time–frequency domain and enables broad generalization across diverse downstream tasks.

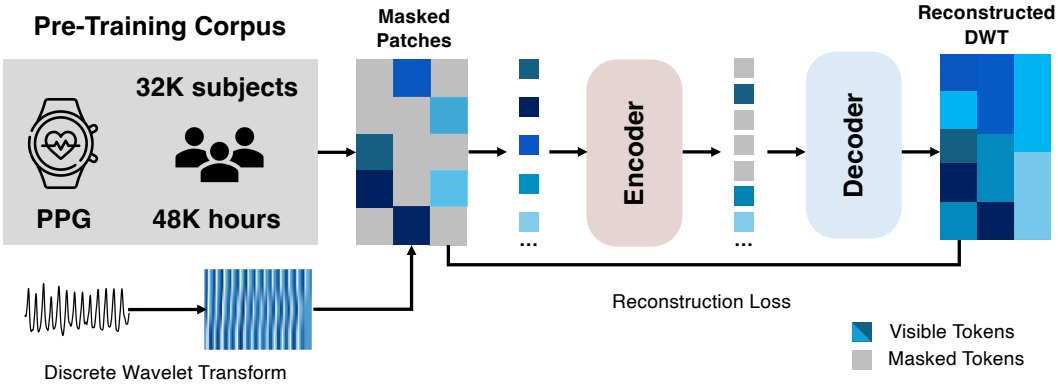


Figure 1: Masked Multiscale Reconstruction for Photoplethysmography (PPG) signals.

Physiological signals such as PPG are inherently multi-scale: local waveform morphology encodes vascular health, and long-term trends capture heart rate variability and global rhythm dynamics (Charlton et al., 2022b; Namvari et al., 2022; Sherebrin & Sherebrin, 2002). Prior approaches for PPG foundation models—temporal embeddings, patient-aware contrastive learning (Abbaspourazad et al., 2024b), morphology-based objectives (Pillai et al., 2024), generally overlook this hierarchical multi-scale structure of PPG. Even hybrid time–frequency models (Zhang et al., 2022a) often treat these domains separately, limiting multi-scale representation learning—a capability crucial for downstream health-related tasks that require features spanning granularities from beat-level morphology to global temporal semantics.

Wavelet decomposition offers a natural, physiologically aligned approach for analyzing PPG in the time-frequency domain. By adaptively trading off time and frequency resolution, wavelets can capture high-frequency transients, subtle waveform changes, and low-frequency rhythms such as respiration and circadian trends in a unified, multi-resolution representation. Motivated by these insights, we establish the Masked Multiscale Reconstruction (MMR) framework: raw PPG signals are decomposed into multiple wavelet bands using the Discrete Wavelet Transform (Daubechies, 1992; Mallat, 2002), and the model is trained to reconstruct masked coefficients across scales. This approach encourages the learning of rich, cross-resolution embeddings.

To summarize, our contributions are: *(i)* We pretrain a large-scale *wavelet-based PPG foundation model* on ∼48K hours of real-world wearable PPG data with a masked multiscale reconstruction objective, enabling the model to capture rich time–frequency information across multiple scales. *(ii)* We demonstrate strong generalization across 13 diverse downstream tasks and provide detailed ablations that examine the impact of design choices such as wavelet family, decomposition scales, and patch size, further underscoring the promise of wavelets for building generalizable PPG foundation models.

2 Related Work

Self-supervised pre-training has become the dominant paradigm for large-scale biosignal modeling. Prior PPG foundation models have leveraged large datasets to learn general representations, using contrastive learning or waveform-based objectives to improve generalization (Abbaspourazad et al., 2024b; Pillai et al., 2024; Saha et al., 2025). Multimodal foundation models transfer representations across PPG, ECG, and other signals, and masked reconstruction has shown promise for multivariate health time series (Abbaspourazad et al., 2024a; Yang et al., 2023; Narayanswamy et al., 2024; Xu et al., 2025). These works mark a shift from traditional task-specific models to general-purpose biosignal foundation models.

Prior frequency-aware models rely on fixed-window spectral features, limiting their ability to capture the adaptive, hierarchical, multi-scale rhythms inherent across time-frequency in physiological signals (Zhang et al., 2022b; Liu et al., 2023; Kara et al., 2024; Cheng et al., 2025; Fu & Hu, 2025; Duan et al., 2024). Early works have applied wavelets to PPG for denoising, feature extraction, or

task-specific pipelines (Alafeef & Fraiwan, 2020; Singh et al., 2023; Shao et al., 2021), while more recent deep learning work has applied end-to-end wavelets for time-series tokenization (Masserano et al., 2025) and modeling biosignals like ECG and EMG (Chen et al., 2025). These methods, however, do not explicitly model cross-resolution interactions at scale for large, real-world PPG datasets.

In contrast, our MMR framework leverages multi-resolution wavelet decomposition to pre-train a large-scale PPG foundation model. By reconstructing masked coefficients across scales, MMR explicitly captures hierarchical, physiologically grounded structure—from local waveform morphology to intermediate oscillations and long-term rhythms. This approach goes beyond time-only and fixed-resolution spectral methods, producing robust, transferable embeddings that generalize across diverse cardiovascular and health monitoring tasks. See Appendix A.1 for an extended related works.

3 METHOD

To build a Masked Multiscale Reconstruction model, we transform PPG signals into multiple time–frequency scales using the Discrete Wavelet Transform (DWT) The resulting wavelet coefficients are interpolated and stacked to form a 2D coefficient map, which is partitioned into patches, many of which are masked, and then processed by a Vision Transformer (ViT) (Dosovitskiy et al., 2020). The model is trained with a multi-scale reconstruction objective, aiming to recover the masked coefficients and thereby learn robust signal representations (Fig. 1).

Discrete wavelet transform. The discrete wavelet transform (DWT) decomposes a signal into an approximation A_J and detail bands ${D_j}_{j=1}^{J}$ using paired low- and high-pass filters, with each level downsampled by half for joint time–frequency localization (Mallat, 2002; Daubechies, 1992). Unlike the Fourier transform (Bracewell, 1989), which represents global frequency content and has zero time resolution, or the Short-Time Fourier Transform (STFT) (Durak & Arikan, 2003), which uses fixed time windows, the DWT adapts across scales, offering wide temporal support for low-frequency components and sharp temporal localization for high-frequency transients (Masserano et al., 2025; Stephane, 1999). This makes wavelets well-suited for nonstationary signals with localized bursts or discontinuities. In the DWT, the approximation and detail coefficients are obtained by inner products with scaling and wavelet functions. The scaling function $(\phi_{J,k})$, which acts as a low-pass filter, is used to find the approximation coefficients (a_k^J), while the wavelet function $(\psi_{j,k})$, which acts as a high-pass filter, is used for the detail coefficients (d_k^j). The approximation coefficients capture the low-frequency (coarse) structure of the signal, while the detail coefficients capture the high-frequency (fine) variations, providing a multi-resolution representation (Daubechies, 1992).

To obtain DWT coefficients from PPG signals, we conducted an empirical ablation (Section 5.4) over multiple wavelet families and decomposition levels, and found the Haar wavelet family (Haar, 1909) to perform reliably. We employ a level-4 Haar DWT using `PyWavelets` (Lee et al., 2019), which produces one approximation band and four detail bands. For the Haar wavelet, approximation coefficients are calculated as local averages, whereas detail coefficients capture signed differences, thereby summarizing coarse trends while isolating localized variations. We interpolate the detail and approximation coefficients obtained from the DWT at each subband level, stretching the coefficients at level j to match the original signal length. The resulting subbands are then concatenated in decreasing order of frequency, with high-frequency detail coefficients stacked at the top and the low-frequency approximation band at the bottom, forming a 2-D representation of shape $[n_{\text{subbands}}, \text{time}]$.

Masked Multiscale Reconstruction – MMR. We adopt a Vision Transformer (ViT) encoder–decoder within the masked autoencoder framework (He et al., 2022). The 2-D wavelet coefficient map is divided into non-overlapping patches of size $(1, 25)$ along the temporal axis, yielding a token sequence $\{x_p\}_{p\in\mathcal{P}}$ across subbands. Fixed 2-D sine–cosine positional embeddings encode both temporal order and subband index. During pretraining, a random subset $\mathcal{M} \subset \mathcal{P}$ of 75% patches is masked, and the decoder reconstructs the missing coefficients from the visible tokens $\mathcal{P} \setminus \mathcal{M}$. Our MMR objective exploits the hierarchical structure of the DWT: coarse approximation bands A_J encode global trends at scale 2^J, while detail bands $\{D_j\}$ refine these trends at progressively higher frequencies. The reconstruction task therefore encourages the encoder to capture both *top-*

down (coarse → fine) and *bottom-up* (fine → coarse) dependencies, rather than treating each band independently. This approach encourages cross-scale feature sharing reminiscent of classical multiresolution analysis (Mallat, 2002), but learned directly by the Transformer. Formally, let $X_p \in \mathbb{R}^d$ denote the original coefficient vector in patch p and $\hat{X}_p$ the reconstruction. The MMR loss is the mean-squared error over masked patches:

$$\mathcal{L}_{\text{MMR}} = \frac{1}{|\mathcal{M}|} \sum_{p \in \mathcal{M}} \left\| \hat{X}_p - X_p \right\|_2^2 .$$

Equivalently, writing $X = \{A_J, D_1, \ldots, D_J\}$ for the multiscale decomposition,

$$\mathcal{L}_{\text{MMR}} = \frac{1}{|\mathcal{M}|} \left[\sum_{j=1}^{J} \sum_{p \in \mathcal{M} \cap D_j} \left\| \hat{X}_p^{(j)} - X_p^{(j)} \right\|_2^2 + \sum_{p \in \mathcal{M} \cap A_J} \left\| \hat{X}_p^{(A_J)} - X_p^{(A_J)} \right\|_2^2 \right],$$

which makes explicit that errors are penalized across all scales. This multiscale masking compels the encoder to jointly model spectral and temporal dependencies, yielding richer latent features for downstream tasks.

4 Experimental Setting

Here, we describe the datasets, pretraining setup, and evaluation protocols used to evaluate the MMR approach.

Dataset We pretrain our encoder on large-scale PPG segments collected from diverse [REDACTED] smartwatches. Each input is a 10-second segment, chosen as a practical trade-off between signal quality and model input length: shorter windows reduce the likelihood of motion artifacts or device dropout, while still capturing multiple cardiac cycles necessary for reliable waveform representation. The majority of segments were sampled at lower frequency range of 25 Hz-100Hz, reflecting the low-power, battery-constrained acquisition typical of free-living wearable devices. This setting is deliberately more challenging than those used in much of the prior work, which often relies on clean clinical signals or generic time-series datasets (Pillai et al., 2024). By contrast, our pretraining data reflect the noise, variability, and resource constraints of real-world wearables, aligning model development directly with deployment conditions.

Preprocessing Before feeding each PPG segment as input to MMR, each segment is bandpass filtered (Liang et al., 2018; Lapitan et al., 2024), and subsequently z-score normalized to standardize amplitude across devices and users. These steps reduce trivial sources of variability so that the encoder focuses on physiologically meaningful structure. Because our signals come from real-world smartwatch deployments, they inevitably contain motion artifacts, poor skin contact, and environmental noise. Rather than eliminating this variability altogether, we apply a lightweight signal quality index (SQI)–based filter to discard only the most corrupted segments. We combine entropy (to capture waveform regularity) and autocorrelation (to assess periodicity), two metrics commonly used to identify usable cardiac waveforms (Elgendi, 2016; Karlen et al., 2012; Pradhan et al., 2017). This approach strikes a balance: the retained segments still reflect the variability of naturalistic conditions, but avoid extreme outliers that would otherwise overwhelm representation learning.

Downstream Tasks We evaluate the learned representations on 13 health-related downstream tasks spanning both classification and regression. Classification tasks include hypertension detection, evaluated in both controlled laboratory protocols and free-living settings. We also assess premature ventricular contraction (PVC) detection, which provides clinically relevant information for arrhythmia monitoring and supports atrial fibrillation detection. In addition, we evaluate classification of laboratory biomarker abnormalities, including electrolytes (sodium, potassium, carbon dioxide), hematological indices (hemoglobin, white blood cells), and metabolic markers (creatinine; see Appendix A.4). Regression tasks focus on systolic and diastolic blood pressure prediction, capturing continuous cardiovascular physiology. Collectively, this suite of tasks assesses the extent to which MMR embeddings encode clinically relevant cardiovascular information from wearable PPG data under diverse real-world conditions.

Table 1: Comparison with Self-Supervised Baselines. Best downstream evaluation scores are shown in **bold**, second-best are underlined, and the [min–max] range across the five cross-validation test splits is reported in gray brackets.

Classification - AUROC (↑)	**SimCLR (5M)**	**MSN (2.5M)**	**TF-C (10M)**	**MMR (7M)**	**MMR-Light (2M)**
Hypertension - Lab	69.58 [61.2 – 76.6]	66.90 [53.6 – 76.6]	**73.10** [63.0 – 90.9]	71.18 [60.9 – 92.6]	71.50 [58.5 – 89.1]
Hypertension - Free Living	60.93 [59.1 – 62.8]	55.16 [53.6 – 56.3]	62.92 [61.3 – 65.0]	**68.12** [67.0 – 69.9]	67.08 [65.9 – 69.3]
PVC	74.43 [67.9 – 82.4]	73.60 [65.7 – 79.1]	70.25 [64.9 – 74.8]	**82.04** [74.5 – 88.9]	81.30 [74.7 – 88.0]
Carbon Dioxide	**65.07** [53.5 – 83.7]	62.90 [51.9 – 78.3]	63.02 [53.1 – 77.2]	63.10 [45.7 – 73.7]	63.95 [46.3 – 74.6]
Creatinine	54.18 [38.8 – 71.4]	56.45 [45.6 – 70.4]	53.08 [47.7 – 63.8]	54.82 [38.9 – 69.9]	**57.28** [39.3 – 72.9]
Hemoglobin	**56.10** [29.1 – 69.8]	52.04 [35.4 – 74.2]	51.55 [40.2 – 64.3]	52.63 [35.6 – 79.1]	53.19 [34.8 – 87.7]
Potassium	71.63 [62.3 – 85.1]	69.47 [58.8 – 80.8]	73.76 [59.9 – 81.2]	**74.27** [41.9 – 64.2]	73.23 [50.3 – 81.0]
Sodium	60.37 [53.5 – 65.7]	60.39 [49.6 – 72.7]	55.18 [41.7 – 71.6]	**60.54** [50.2 – 66.9]	57.08 [41.5 – 64.2]
White Blood Cells	**79.51** [55.9 – 86.0]	76.93 [49.9 – 88.3]	77.37 [48.2 – 86.3]	75.71 [42.4 – 91.6]	75.53 [41.8 – 91.4]
Average	65.76 ± 8.12	63.76 ± 8.09	64.47 ± 9.12	**66.93** ± 9.34	66.68 ± 9.00
Regression - MAE (↓)					
Sys. BP (Lab)	11.02 [9.0 – 12.0]	11.02 [8.9 – 11.8]	**10.93** [9.1 – 11.6]	11.12 [9.4 – 11.7]	11.07 [9.4 – 11.8]
Dias. BP (Lab)	7.95 [6.8 – 10.1]	8.07 [6.7 – 10.2]	7.95 [7.2 – 10.2]	7.90 [6.9 – 10.1]	**7.88** [6.9 – 10.1]
Sys. BP	11.75 [11.6 – 11.8]	11.82 [11.7 – 11.9]	11.75 [11.6 – 11.8]	**11.63** [11.4– 11.7]	11.66 [11.5 – 11.8]
Dias. BP	9.47 [9.2 – 9.7]	9.52 [9.2 – 9.8]	9.45 [9.2 – 9.7]	**9.36** [09.1 – 9.5]	9.37 [9.1 – 9.6]
Average	10.04 ± 1.46	10.11 ± 1.43	10.02 ± 1.45	10.00 ± 1.47	**9.99** ± 1.48

Baselines We compare MMR against three groups of baselines. First, self-supervised learning methods trained on the same wearable PPG data, including SimCLR (Chen et al., 2020), Masked Siamese Network (MSN) (Assran et al., 2022), and TF-C (Zhang et al., 2022a), representing contrastive, masked patch/data matching, and multi-view frequency–time approaches, respectively (see Appendix A.3.1 for details). These baselines allow us to assess the benefit of our proposed architecture and pretraining strategy when applied to the same real-world PPG signals. Second, we evaluate open-source pretrained models, divided into two categories: (i) general-purpose time-series foundation models such as Chronos (Ansari et al., 2024), which are trained on diverse multivariate time-series but not specifically on physiological signals, and (ii) domain-specific models such as PaPaGei (Pillai et al., 2024), which leverage high-quality fingertip PPG with stratified binning and explicit morphological feature learning to capture clinical waveform properties at high sampling rates (125–500 Hz); we use the PaPaGei-S variant in this work. These models provide a point of comparison to evaluate whether pretraining on clean or general-purpose data transfers effectively to noisy, wearable PPG. Finally, we include classical statistical feature–based models without pretraining to provide a non-learned baseline. Together, these baselines span the current state of both general-purpose and domain-specialized representation learning for PPG signals, highlighting the advantages of MMR in capturing real-world cardiovascular physiology.

5 Downstream Evaluation

We evaluate the quality and generalization of the learned representations by training downstream classifiers on top of frozen encoders. We consider both our full MMR model (7M params)and a smaller variant, MMR-Light (2M params), which has fewer parameters to study efficiency-performance trade-offs. For all experiments, encoder representations serve as input features to random forest models for classification and regression tasks, with performance measured on held-out test data using 5-fold cross-validation. Binary classification tasks are evaluated via AUROC, and regression tasks via mean absolute error (MAE). Our evaluation encompasses several aspects: (i) performance across 13 downstream health-related tasks compared to self-supervised and pretrained baselines, (ii) analysis of the learned embedding space to assess patient discriminability and physiological structure, (iii) the impact of pretraining data size and model parameter scaling, and (iv) ablation studies examining key architectural choices such as wavelet family, decomposition level, and patch size. This comprehensive evaluation allows us to assess both the predictive power and the interpretability of the representations learned by MMR and MMR-Light.

Table 2: Comparison with Open-source Pretrained Models and Statistical Features. Best downstream evaluation scores are shown in **bold**, second-best are underlined, and the minimum–maximum range across the five cross-validation splits is reported in gray brackets.

Classification - AUROC (↑)	**Stat. Feat.**	**Chronos (200M)**	**PaPaGei (5M)**	**MMR (7M)**	**MMR-Light (2M)**
Hypertension - Lab	69.28 [59.7–83.1]	67.26 [55.8–79.8]	65.85 [58.6–75.4]	71.18 [60.9–92.6]	**71.50** [58.5–89.1]
Hypertension - Free Living	56.59 [54.7–57.7]	59.95 [58.3–61.8]	59.94 [59.2–60.2]	**68.12** [67.0–69.9]	67.08 [65.9–69.3]
PVC	70.41 [64.2–77.2]	65.73 [65.3–75.3]	72.20 [67.0–79.9]	**82.04** [74.5–88.9]	81.30 [74.7–88.0]
Carbon Dioxide	57.20 [50.3–69.2]	61.14 [54.0–75.1]	61.23 [52.3–75.3]	63.10 [45.7–73.7]	**63.95** [46.3–74.6]
Creatinine	49.32 [40.2–56.7]	**61.14** [49.2–74.0]	51.86 [43.4–70.0]	54.82 [38.9–69.9]	57.28 [39.3–72.9]
Hemoglobin	**56.13** [35.6–65.1]	53.65 [43.7–59.3]	49.28 [39.5–59.2]	52.63 [35.6–79.1]	53.19 [34.8–87.7]
Potassium	47.87 [41.6–55.3]	63.49 [55.4–85.1]	68.51 [56.1–77.2]	**74.27** [41.9–64.2]	73.23 [50.3–81.0]
Sodium	52.72 [47.5–56.7]	**63.33** [53.6–74.2]	58.44 [51.1–70.3]	60.54 [50.2–66.9]	57.08 [41.5–64.2]
White Blood Cells	55.52 [52.7–58.3]	**77.51** [55.4–88.1]	73.64 [44.1–87.1]	75.71 [42.4–91.6]	75.53 [41.8–91.4]
Average	57.12 ± 7.60	63.47 ± 7.73	62.77 ± 8.61	**66.93** ± 9.34	66.68 ± 9.00
Regression - MAE (↓)					
Sys. BP (Lab)	11.08 [9.3–11.9]	**10.79** [9.2–11.9]	11.04 [9.6–12.2]	11.12 [9.4–11.7]	11.07 [9.4–11.8]
Dias. BP (Lab)	8.12 [6.7–10.3]	7.93 [7.0–8.9]	8.09 [6.9–10.1]	7.90 [6.9–10.1]	**7.88** [6.9–10.1]
Sys. BP	11.75 [11.6–11.8]	11.82 [11.7–11.9]	11.75 [11.6–11.8]	**11.63** [11.4–11.7]	11.66 [11.5–11.8]
Dias. BP	9.47 [9.2–9.7]	9.52 [9.2–9.8]	9.45 [9.2–9.7]	**9.36** [9.1–9.5]	9.37 [9.1–9.6]
Average	10.04 ± 1.46	10.11 ± 1.43	10.02 ± 1.45	10.00 ± 1.47	**9.99** ± 1.48

5.1 Evaluating Transferrability of learned features across Dievsre Downstream tasks

Detailed results for all baselines are reported in Tables 1 and 2. Across 13 downstream health-related tasks, MMR consistently achieves superior or competitive performance relative to state-of-the-art baselines. On average, MMR reaches 66.9% AUROC across 9 binary classification tasks, with notable strengths in PVC detection and free-living hypertension classification. Its lightweight variant, MMR-Light, achieves comparable performance, demonstrating that parameter efficiency can be maintained without substantial loss in predictive quality. For regression tasks, MMR attains the lowest error on diastolic blood pressure in the lab setting and on both hypertension-related regression tasks in free-living conditions. These results hold even when compared to self-supervised baselines trained on identical wearable PPG data (SimCLR, MSN, TF-C). While TF-C, leveraging both time- and frequency-domain augmentations, performs well on hypertension and systolic BP regression, SimCLR and MSN underperform on free-living hypertension and PVC detection. This highlights the value of multi-scale representation learning: MMR's design captures richer temporal and spectral structure and consistently outperforming competing approaches.

When compared to pretrained PPG models, MMR outperforms PaPaGei by approximately 4% AUROC on average across classification tasks. PaPaGei, trained on high-quality fingertip PPG at high sampling rates, relies on fiducial point extraction, which can limit adaptability to noisy, low-sampling-rate wearable data. Similarly, MMR improves over the large, general-purpose time-series model Chronos on 5 of 9 tasks (average +3.5% AUROC), showing that domain-specific pretraining on large-scale wearable PPG is critical for capturing cardiovascular signal characteristics that general multivariate models may miss.

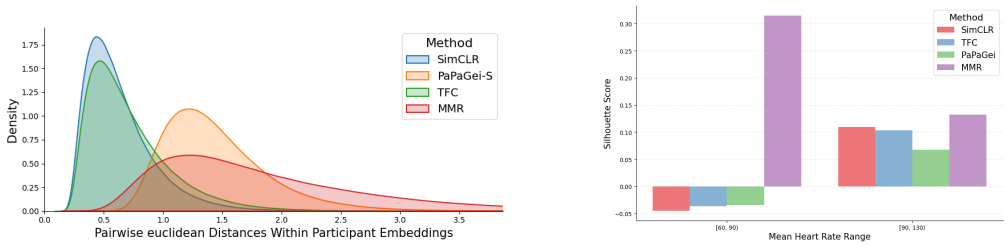
Taken together, these results indicate that MMR not only excels across a wide range of tasks but also produces representations that are robust to real-world variability in device, user, and signal quality. Unlike models trained on clinical or high-fidelity data, MMR demonstrates strong generalization under noisy, resource-constrained conditions, making it particularly suitable for practical, real-world deployment.

5.2 Learned Feature Analysis

To further interpret how MMR captures clinically relevant information, we examine the structure of the learned embeddings on an unseen downstream task: the Hypertension- Free Living dataset.

Discriminability of Patient-Wise Embeddings. To evaluate the quality and interpretability of the learned representations, we examine the separation of patient embeddings by computing inter-patient distances. Well-separated embeddings indicate that the model separates diverse patient co-

horts, which can improve robustness and accuracy in downstream tasks Pillai et al. (2024). Figure 2a shows the average pairwise distances across patients in the Hypertension- Free Living dataset. MMR achieves the largest separation among all methods, demonstrating its ability to disentangle individual patient identities in the embedding space. In contrast, contrastive baselines such as SimCLR and TF-C show weaker separation, with many embeddings collapsing together, while PaPaGei performs moderately well but still lags behind MMR.



(a) Comparison of participant-wise distance distributions of learned embeddings across different baselines. Wider curves mean stronger separation and discrimination.

(b) Silhouette scores of heart rate clusters in the average patient embeddings learned by different baselines. A higher score indicates better-defined clusters and stronger separation in the embedding space.

Figure 2: Analyzing the learned embeddings on the unseen downstream dataset of the Hypertension – Free Living.

Physiological Structure in the Embedding Space. Beyond patient-wise discriminability, the learned embeddings capture meaningful physiological variation. Visualizing participant-level mean embeddings with t-SNE (van der Maaten & Hinton, 2008), colored by heart rate ranges (Figure 11), reveals a smooth gradient from low to high heart rate. This observation indicates that MMR encodes underlying cardiovascular dynamics rather than producing arbitrary or randomly aligned representations. To quantify this structure, we compute silhouette scores (Rousseeuw, 1987) for elevated (90-130 bpm) vs. normal (60-90 bpm) heart rate groups. MMR consistently achieves the highest scores, demonstrating that its embeddings cluster more coherently by physiological state than those of competing baselines. These findings suggest that MMR produces *physiologically aligned representations*, enabling downstream models to leverage subtle but clinically relevant variations in cardiovascular signals.

5.3 Evaluating MMR at Varied Data and Parameter Settings

We next examine how pretraining dataset size and model capacity influence downstream performance. By comparing MMR across different amounts of pretraining data and varying parameter scales, we gain insight into the trade-offs between model complexity, data efficiency, and predictive accuracy.

Pretraining Data Scaling We pretrain MMR on datasets of increasing size—1M, 5M, and 17M segments—to examine how the amount of pretraining data affects downstream classification performance. As shown in Table 3, larger datasets consistently improve results on hypertension and PVC detection, highlighting the benefit of data volume. The smallest dataset (1M segments) underperforms relative to larger splits, achieving 68.0% and 66.1% AUROC on lab and free-living hypertension tasks, compared to 69.1% and 67.5% for 5M segments. Pretraining on the full 17M segments yields the strongest overall performance across tasks. Some lab biomarkers, such as abnormal WBC and creatinine, show little improvement with additional data (see Appendix A.6), suggesting that certain endpoints are less sensitive to segment-level diversity. Overall, these results indicate that diverse and large-scale pretraining data—spanning multiple users, devices, and real-world contexts—is critical for learning robust, generalizable representations.

Model Scaling We next consider the effect of model capacity on downstream performance. MMR models with varying parameter counts were pretrained to assess how size influences predictive accuracy. As shown in Table 6 (full numbers in Appendix A.6, Table 5), the mid-sized 7M-parameter

Classification AUROC (↑)	Data Scaling (Segments)			Model Scaling (Params)		
	1M	5M	17M	2M	7M	27M
Hypertension-Lab	67.99	69.06	**71.18**	71.50	71.18	**80.54**
Hypertension	66.08	67.50	**68.12**	67.08	**68.12**	66.65
PVC Detection	76.62	79.84	**82.04**	81.30	**82.04**	81.41

Table 3: MMR Scaling Results: Mean AUROC on 5-fold cross-validation for selected predictive tasks under both model scaling (parameters) and data scaling (pretraining segments).

model performs strongly across tasks, ranking first or second on 7 of 9 classification tasks. The smaller 2M model also achieves competitive results, indicating that MMR-Light provides an effective, parameter-efficient alternative suitable for deployment on resource-constrained devices. Larger models offer notable improvements for select endpoints, such as +10% AUROC on hypertension-lab and +5% on sodium prediction, but these gains do not generalize uniformly across all tasks. Overall, scaling model capacity can enhance performance for specific tasks, but even compact models retain substantial predictive power.

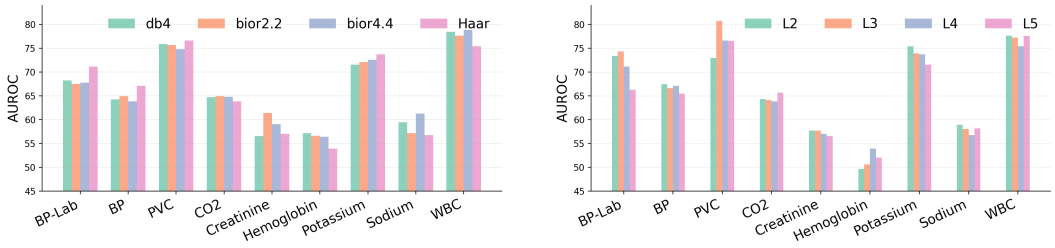
5.4 Ablation Studies

To understand the contribution of individual components within MMR, we conducted ablation experiments using a 2M-parameter model (MMR-Light) pretrained on a 1M-segment subset. We systematically varied the wavelet family, decomposition level, masking strategy, and patch size, and evaluated the resulting impact across all classification tasks. In Figure 3, the Hypertension-Free Living dataset is labeled as BP, and the Hypertension Lab as BP Lab.

Wavelet family design. We evaluated several wavelet bases (db4, bior2.2, bior4.4, Haar) (Daubechies, 1988; Karoui & Vaillancourt, 1994; Haar, 1909) under identical training conditions (2M parameters, 1M segments). Across hypertension classification tasks, Haar consistently achieved the highest AUROC, and it also outperformed most other wavelets on PVC detection. This advantage likely stems from Haar's compact and sharply changing basis functions, which are well-suited to capturing abrupt waveform changes that signal irregular heartbeats (Yang et al., 2019). In contrast, smoother wavelets like db4 and bior4.4 can miss these fine transients, reducing performance on tasks that depend on detecting sharp signal features. For other biomarkers and vitals, differences across wavelets were smaller, with average AUROC generally between 64–68. Notably, bior2.2 improved Creatinine prediction, while bior4.4 gave modest gains for WBC and Sodium lab outcomes. These results suggest that smoother wavelets may be preferable for tasks dominated by slower, more global signal dynamics, whereas Haar supports tasks requiring the detection of rapid, localized changes.

Effect of decomposition depth. We evaluated the effect of wavelet decomposition depth (levels 2–5 using Haar) in our PPG $\rightarrow$ DWT transformation. Decomposition depth determines the number of multi-resolution sub-bands used to represent the signal. Moderate depths (levels 3) consistently yielded the best performance on hypertension classification, while deeper decompositions (level 5) led to a 7–8 % AUROC drop, likely because excessive decomposition fragments the signal and discards fine-grained, predictive features. A similar pattern was observed for PVC detection: level 3 achieved over 80% AUROC with only 1M pretraining samples, outperforming level 2 (73%) and level 5 (75%). Some lab biomarkers, including WBC, Sodium, and carbon dioxide, benefited from deeper decomposition, indicating that tasks dominated by slower, global signal trends may require more sub-bands. Overall, these results show that decomposition depth has a task-dependent impact and highlight the potential value of adaptive decomposition strategies rather than using a fixed hierarchy.

Patch size ablation Building on the previous analysis of decomposition depth, we next examine how patch size—the temporal resolution of input segments—affects downstream performance (refer Fig. 5b in Appendix). Patch size determines how the encoder captures local transients versus broader signal dynamics. Smaller patches, such as $(1, 25)$, consistently achieve the strongest hypertension classification, highlighting the importance of preserving fine-grained temporal details.



(a) Comparison across different wavelet families (db4, bior2.2, bior4.4, and Haar) shows that Haar provides consistently strong performance across tasks.

(b) Evaluation of decomposition levels (L2–L5) indicates that moderate depths, particularly L4, achieve the best balance.

Figure 3: Ablation of various wavelet families and decomposition levels for PPG signal analysis.

Larger patches, like $(1, 100)$, tend to smooth out fine-grained events, leading to a 6–7% drop in hypertension lab AUROC, but provide modest gains for lab biomarkers such as WBC and potassium. The intermediate patch size $(1, 50)$ offers a compromise, performing reasonably on lab outcomes but remaining weaker for hypertension. The across-band scheme $(2, 25)$ underperforms across nearly all tasks, suggesting that mixing sub-bands can reduce discriminative power.

Overall, these results reveal that different tasks benefit from distinct temporal and frequency scales and suggest that multi-scale PPG representations provide complementary cues, highlighting wavelet-based encodings as an effective pretraining strategy for robust and adaptive physiological monitoring.

6 Discussion and Future Work

Foundation models for biosignals hold strong promise for generalizable digital health applications, yet most existing approaches overlook the spectral structure underlying physiological rhythms. In this work, we introduced Masked Multiscale Reconstruction (MMR), a pretraining framework for PPG that leverages wavelet-based time–frequency representations, and demonstrated robust performance across 13 diverse health tasks, matching or surpassing state-of-the-art baselines. Our analyses show that MMR embeddings capture physiologically meaningful information and enable subject-level discrimination, while ablations highlight the importance of explicitly modeling spectral hierarchies for representation learning. A key limitation of our current design is the reliance on fixed decomposition levels, which leads to uneven performance across tasks and motivates exploration of adaptive multiscale strategies. More broadly, our findings suggest that learning directly in the joint time–frequency domain is a powerful paradigm for PPG foundation models, opening paths toward multimodal integration, longitudinal modeling, and clinically meaningful health prediction.

REFERENCES

Salar Abbaspourazad, Anshuman Mishra, Joseph Futoma, Andrew C Miller, and Ian Shapiro. Wearable accelerometer foundation models for health via knowledge distillation. *arXiv preprint arXiv:2412.11276*, 2024a.

Salar Abbaspourazad et al. Large-scale training of foundation models for wearable biosignals. In *ICLR 2024 Workshop or Poster*, 2024b. URL https://openreview.net/forum?id=pC3WJHf51j. Self-supervised learning on Apple Heart Movement Study (PPG/ECG).

Maha Alafeef and Mohammad Fraiwan. Smartphone-based respiratory rate estimation using photoplethysmographic imaging and discrete wavelet transform. *Journal of Ambient Intelligence and Humanized Computing*, 11(2):693–703, 2020.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

Apple Inc. Hypertension notifications validation paper. https://www.apple.com/health/pdf/Hypertension_Notifications_Validation_Paper_September_2025.pdf, September 2025. Accessed: 2025-09-23.

Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *European conference on computer vision*, pp. 456–473. Springer, 2022.

Syed Khairul Bashar, Dong Han, Shirin Hajeb-Mohammadalipour, Eric Ding, Cody Whitcomb, David D McManus, and Ki H Chon. Atrial fibrillation detection from wrist photoplethysmography signals using smartwatches. *Scientific reports*, 9(1):15054, 2019.

Ronald N Bracewell. The fourier transform. *Scientific American*, 260(6):86–95, 1989.

Yong-Mei Cha, Glenn K Lee, Kyle W Klarich, and Martha Grogan. Premature ventricular contraction-induced cardiomyopathy: a treatable condition. *Circulation: Arrhythmia and Electrophysiology*, 5(1):229–236, 2012.

Peter H Charlton, Panicos A Kyriacou, Jonathan Mant, Vaidotas Marozas, Phil Chowienczyk, and Jordi Alastruey. Wearable photoplethysmography for cardiovascular monitoring. *Proceedings of the IEEE*, 110(3):355–381, 2022a.

Peter H Charlton, Birutė Paliakaitė, Kristjan Pilt, Martin Bachler, Serena Zanelli, Daniel Kulin, John Allen, Magid Hallab, Elisabetta Bianchini, Christopher C Mayer, et al. Assessing hemodynamics from the photoplethysmogram to gain insights into vascular age: a review from vascagenet. *American Journal of Physiology-Heart and Circulatory Physiology*, 322(4):H493–H522, 2022b.

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, 2020.

Xiaoli Chen, Rui Wang, Phyllis Zee, Pamela L Lutsey, Sogol Javaheri, Carmela Alcántara, Chandra L Jackson, Michelle A Williams, and Susan Redline. Racial/ethnic differences in sleep disturbances: the multi-ethnic study of atherosclerosis (mesa). *Sleep*, 38(6):877–888, 2015.

Yanlong Chen, Mattia Orlandi, Pierangelo Maria Rapa, Simone Benatti, Luca Benini, and Yawei Li. Physiowave: A multi-scale wavelet-transformer for physiological signal representation. *arXiv preprint arXiv:2506.10351*, 2025.

Rui Cheng, Xiangfei Jia, Qing Li, Rong Xing, Jiwen Huang, Yu Zheng, and Zhilong Xie. Fat: Frequency-aware pretraining for enhanced time-series representation learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 310–321, 2025.

Ingrid Daubechies. Orthonormal bases of compactly supported wavelets. *Communications on pure and applied mathematics*, 41(7):909–996, 1988.

Ingrid Daubechies. *Ten lectures on wavelets*. SIAM, 1992.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Jufang Duan, Wei Zheng, Yangzhou Du, Wenfa Wu, Haipeng Jiang, and Hongsheng Qi. Multi-frequency contrastive learning representation for time series. In *ICML*, 2024.

Lutfiye Durak and Orhan Arikan. Short-time fourier transform: two fundamental properties and an optimal implementation. *IEEE Transactions on Signal Processing*, 51(5):1231–1242, 2003.

Mohamed Elgendi. Optimal signal quality index for photoplethysmogram signals. *Bioengineering*, 3(4):21, 2016.

Eray Erturk, Fahad Kamran, Salar Abbaspourazad, Sean Jewell, Harsh Sharma, Yujie Li, Sinead Williamson, Nicholas J Foti, and Joseph Futoma. Beyond sensor data: Foundation models of behavioral data from wearables improve health predictions. *arXiv preprint arXiv:2507.00191*, 2025.

En Fu and Yanyan Hu. Frequency-masked embedding inference: A non-contrastive approach for time series representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 16639–16647, 2025.

Alfred Haar. *Zur theorie der orthogonalen funktionensysteme*. Georg-August-Universitat, Gottingen., 1909.

Dong Han, Syed Khairul Bashar, Fahimeh Mohagheghian, Eric Ding, Cody Whitcomb, David D McManus, and Ki H Chon. Premature atrial and ventricular contraction detection using photoplethysmographic data from a smartwatch. *Sensors*, 20(19):5683, 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.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Denizhan Kara, Tomoyoshi Kimura, Shengzhong Liu, Jinyang Li, Dongxin Liu, Tianshi Wang, Ruijie Wang, Yizhuo Chen, Yigong Hu, and Tarek Abdelzaher. Freqmae: Frequency-aware masked autoencoder for multi-modal iot sensing. In *Proceedings of the ACM Web Conference 2024*, pp. 2795–2806, 2024.

Walter Karlen, K Kobayashi, J Mark Ansermino, and Guy Albert Dumont. Photoplethysmogram signal quality estimation using repeated gaussian filters and cross-correlation. *Physiological measurement*, 33(10):1617, 2012.

A Karoui and R Vaillancourt. Families of biorthogonal wavelets. *Computers & Mathematics with Applications*, 28(4):25–39, 1994.

Denis G Lapitan, Dmitry A Rogatkin, Elizaveta A Molchanova, and Andrey P Tarasov. Estimation of phase distortions of the photoplethysmographic signal in digital iir filtering. *Scientific Reports*, 14(1):6546, 2024.

Gregory Lee, Ralf Gommers, Filip Waselewski, Kai Wohlfahrt, and Aaron O'Leary. Pywavelets: A python package for wavelet analysis. *Journal of Open Source Software*, 4(36):1237, 2019.

Simon A Lee and Kai Akamatsu. Foundation models for physiological signals: Opportunities and challenges. 2025.

Yongbo Liang, Mohamed Elgendi, Zhencheng Chen, and Rabab Ward. An optimal filter for short photoplethysmogram signals. *Scientific data*, 5(1):1–12, 2018.

Ran Liu, Ellen L Zippi, Hadi Pouransari, Chris Sandino, Jingping Nie, Hanlin Goh, Erdrin Azemi, and Ali Moin. Frequency-aware masked autoencoders for multimodal pretraining on biosignals. *arXiv preprint arXiv:2309.05927*, 2023.

Ran Liu, Ellen L. Zippi, Hadi Pouransari, Chris Sandino, Jingping Nie, Hanlin Goh, Erdrin Azemi, and Ali Moin. Frequency-aware masked autoencoders for multimodal pretraining on biosignals. In *ICLR Workshop on Learning from Time Series for Health*, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Stephane G Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence*, 11(7):674–693, 2002.

Luca Masserano, Abdul Fatir Ansari, Boran Han, Xiyuan Zhang, Christos Faloutsos, Michael W. Mahoney, Andrew Gordon Wilson, Youngsuk Park, Syama Sundar Rangapuram, Danielle C. Maddix, and Bernie Wang. Enhancing foundation models for time series forecasting via wavelet-based tokenization. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=B6WalMoQJW.

Mina Namvari, Jessica Lipoth, Sheida Knight, Ali Akbar Jamali, Mojtaba Hedayati, Raymond J Spiteri, and Shabbir Syed-Abdul. Photoplethysmography enabled wearable devices and stress detection: a scoping review. *Journal of Personalized Medicine*, 12(11):1792, 2022.

Girish Narayanswamy, Xin Liu, Kumar Ayush, Yuzhe Yang, Xuhai Xu, Shun Liao, Jake Garrison, Shyam Tailor, Jake Sunshine, Yun Liu, et al. Scaling wearable foundation models. *arXiv preprint arXiv:2410.13638*, 2024.

National Library of Medicine (US). Medlineplus. https://medlineplus.gov/, 2020. [Internet]. Bethesda (MD): National Library of Medicine (US); [updated 2020-06-24; cited 2025-08-31].

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

Arvind Pillai et al. Papagei: Open foundation models for optical physiological signals. *arXiv preprint arXiv:2410.20542*, 2024. URL https://arxiv.org/abs/2410.20542.

Nikhilesh Pradhan, Sreeraman Rajan, Andy Adler, and Calum Redpath. Classification of the quality of wristband-based photoplethysmography signals. In *2017 IEEE international symposium on medical measurements and applications (MeMeA)*, pp. 269–274. IEEE, 2017.

Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

Mithun Saha et al. Pulse-ppg: An open-source field-trained ppg foundation model for wearable applications. *arXiv preprint arXiv:2502.01108*, 2025. URL https://arxiv.org/abs/2502.01108.

Shiliang Shao, Ting Wang, Lebing Wang, Sinan Li, and Chen Yao. A photoplethysmograph signal preprocess method based on wavelet transform. In *2021 36th Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, pp. 771–775. IEEE, 2021.

MH Sherebrin and RZ Sherebrin. Frequency analysis of the peripheral pulse wave detected in the finger with a photoplethysmograph. *IEEE Transactions on biomedical engineering*, 37(3):313–317, 2002.

Bikesh Kumar Singh, Neelamshobha Nirala, et al. Expert diagnostic system for detection of hypertension and diabetes mellitus using discrete wavelet decomposition of photoplethysmogram signal and machine learning technique. *Medicine in Novel Technology and Devices*, 19:100251, 2023.

Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016.

Andrius Solosenko and Vaidotas Marozas. Automatic premature ventricular contraction detection in photoplethysmographic signals. In *2014 IEEE Biomedical Circuits and Systems Conference (BioCAS) Proceedings*, pp. 49–52. IEEE, 2014.

Kwangsub Song, Ku-young Chung, and Joon-Hyuk Chang. Cuffless deep learning-based blood pressure estimation for smart wristwatches. *IEEE Transactions on Instrumentation and Measurement*, 69(7):4292–4302, 2019.

Mallat Stephane. A wavelet tour of signal processing, 1999.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

Maxwell A Xu, Girish Narayanswamy, Kumar Ayush, Dimitris Spathis, Shun Liao, Shyam A Tailor, Ahmed Metwally, A Ali Heydari, Yuwei Zhang, Jake Garrison, et al. Lsm-2: Learning from incomplete wearable sensor data. *arXiv preprint arXiv:2506.05321*, 2025.

Chaoqi Yang, M. Brandon Westover, and Jimeng Sun. Biot: Biosignal transformer for cross-data learning in the wild. In *NeurIPS 2023*, 2023. URL https://openreview.net/forum?id=c2LZyTyddi.

Chengming Yang, Cesar Veiga, Juan J Rodriguez-Andina, Jose Farina, Andres Iniguez, and Shen Yin. Using ppg signals and wearable devices for atrial fibrillation screening. *IEEE Transactions on Industrial Electronics*, 66(11):8832–8842, 2019.

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, 2022a.

Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency. In *NeurIPS*, 2022b.

A APPENDIX

A.1 EXTENDED RELATED WORK

Self-supervised pretraining has emerged as the dominant paradigm for large-scale biosignal modeling. For example, (Abbaspourazad et al., 2024b) trained foundation models on PPG and ECG from ∼141K Apple Watch users, demonstrating the value of contrastive learning at scale. In parallel, (Pillai et al., 2024) introduced PaPaGei , an open-source PPG foundation model trained on 20M unlabeled fingertip PPG segments that explicitly leverages waveform morphology, while (Saha et al., 2025) developed Pulse-PPG using 100 days of field data from 120 participants, showing improved efficiency and generalizability. Beyond single-modality PPG, multimodal biosignal foundations transfer representations across ECG, PPG, and other signals either via knowledge distillation (Abbaspourazad et al., 2024a) or unified embeddings (Yang et al., 2023). Related work has also applied masked reconstruction on multivariate health time series, yielding strong generative and discriminative performance on tasks such as activity classification (Narayanswamy et al., 2024; Xu et al., 2025). Together, these advances reflect a shift from task-specific models to general-purpose foundation models for biosignals.

While these foundation models highlight the value of large-scale self-supervision, most treat signals purely in the time domain. A growing body of work shows that explicitly incorporating spectral information provides a powerful inductive bias for robust and transferable representations. For instance, Time-Frequency Consistency (Zhang et al., 2022b) proposed aligning time- and frequency-domain views via contrastive loss, while bioFAME (Liu et al., 2023) introduced a frequency-aware transformer encoder with multi-head spectral filters. Similarly, FreqMAE (Kara et al., 2024) leveraged temporal-shifting encoders to model spectral content in multimodal IoT data. More recent approaches, such as FAT (Cheng et al., 2025), FEI (Fu & Hu, 2025), and MF-CLR (Duan et al., 2024), further illustrate how spectral modeling can enhance time-series representation learning. These findings suggest that frequency-aware pretraining can serve as a complementary approach to large-scale training for physiological signals such as PPG. However, most rely on a fixed-size Fourier transform and fail to capture multi-scale representations of PPG.

Wavelet analysis provides a natural way to capture information at different temporal scales by decomposing signals into multi-resolution frequency bands. Earlier PPG studies applied discrete wavelet transforms (DWT) for denoising and handcrafted features, for example, in respiratory rate estimation (Alafeef & Fraiwan, 2020), hypertension and diabetes detection (Singh et al., 2023), and peak stabilization pipelines (Shao et al., 2021). More recently, deep learning models have incorporated wavelets end-to-end, such as wavelet-based tokenization for time-series foundation models (Masserano et al., 2025) and PhysioWave (Chen et al., 2025), which couples learned wavelet decompositions, frequency guided masking with Transformers for physiological signals such as ECG and EMG. Building on this line of work, we introduce a multi-resolution masked pretraining framework for large-scale PPG data collected from smartwatches in real-world settings. By leveraging the fact that health tasks rely on information at multiple signal granularities, our approach provides more physiologically grounded and transferable representations.

A.2 MMR ABLATIONS

MMR is downstream data efficient In this evaluation, the frozen MMR model representations were evaluated with varying proportions $\{25\%, 50\%, 75\%\}$ of labeled downstream data. MMR demonstrates clear efficiency in limited labeled settings, outperforming both TFC and PaPaGei-S. This early advantage persists as data availability increases, showing that MMR's representations remain robust and transferable across scales. TFC follows closely, improving steadily with additional supervision but never closing the gap with MMR. PaPaGei exhibits the steepest relative improvement as labeled supervision increases, making its performance at full data more competitive. Overall, these patterns highlight MMR's performance consistency across downstream data regimes.

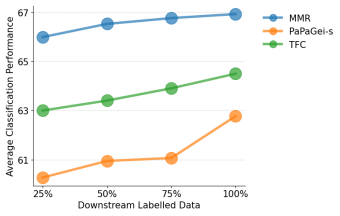
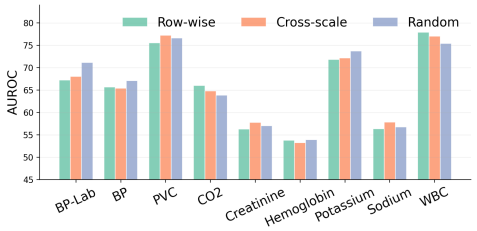
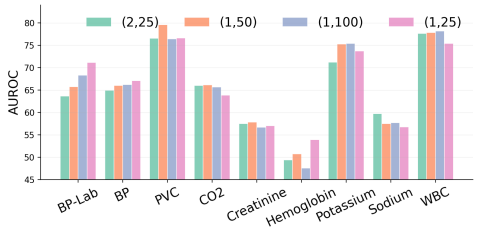


Figure 4: Average performance of MMR, PaPaGei-S, and TFC models across increasing percentages of labeled downstream data. Performance improves with more labeled data, with MMR consistently leading, followed by TFC and PaPaGei-S.

Masking strategy ablation. We evaluate three masking strategies for our masked autoencoder: Random, Row-wise: masking entire rows of approximation or detail coefficients, and Cross-scale: masking columns across the hierarchical scales. Random masking achieves the strongest performance on hypertension classification, outperforming Row-wise by 5–6% and Cross-scale by 2–3%. This suggests that eliminating entire subband coefficients or masking across scales impedes the model's ability to predict low-frequency coefficients without access to high-frequency components, and vice versa. Across most tasks, random masking either leads or performs competitively. However, task-specific patterns emerge: cross-scale proves advantageous for Creatinine and Sodium prediction. For WBC classification, both row-wise and cross-scale masking achieve 78%, representing a +3 point improvement over random. Overall, random masking serves as a strong default strategy, while structured masking (row-wise, cross-scale) offers selective benefits for particular biomarkers.



(a) Masking strategy: We ablate across three masking strategies—row-wise, cross-scale, and random—and find that random masking performs most consistently, providing a robust choice across downstream tasks.



(b) Patch size: We ablate across multiple patch size configurations, varying the number of subbands and temporal span per patch, and observe that $(1, 25)$ achieves the most consistent performance across the downstream task.

Figure 5: Ablation studies analyzing the effect of masking strategy and patch size in the proposed wavelet masked modeling framework, MMR.

A.3 Training Setup

We pretrain MMR using the AdamW optimizer (Loshchilov & Hutter, 2017) with a base learning rate of 1×10^{-4}, a cosine decay schedule, and a linear warmup applied over the first 10% of steps. Training is carried out with a batch size of 512, a weight decay of 1×10^{-5}, and gradient clipping at 1.0. To the PPG signal, we apply the same augmentations as LSM (Narayanswamy et al., 2024) prior to wavelet decomposition, specifically time-flipping, adding Gaussian noise, and stretching along the temporal axis. The MMR model architecture follows the LSM-Small configuration (approximately 7M parameters), consisting of 8 encoder blocks with hidden size 256, 4 attention heads, and a feedforward size of 1024. For reconstruction during pretraining, we use a lightweight decoder composed of 2 blocks with hidden size 192 and 4 attention heads. The smaller variant, MMR-Light (approximately 2M parameters), follows the LSM-Tiny configuration from (Narayanswamy et al., 2024), with 4 encoder blocks (hidden size 192, 3 heads) and a lightweight decoder of 2 blocks (hidden size 128, 4 heads). Hyperparameter tuning was minimal and limited to a small grid search over candidate learning rates $\in \{10^{-2}, 10^{-3}, 10^{-4}\}$ and weight decay values $\in \{10^{-3}, 10^{-4}, 10^{-5}\}$. These sweeps and ablations were conducted on a subset of the pretraining dataset consisting of approximately 1 million data points. All experiments were performed on four Tesla T4 GPUs (16GB each) using distributed data parallel (DDP) training in PyTorch (Paszke et al., 2019).

A.3.1 Baseline Methods

For the PaPaGei family of models, we use the open-source pretrained weights released by Pillai et al. (2024) for PaPaGei-S, the morphology-aware pretraining model, which we refer to simply as PaPaGei. The model employs a ResNet-style convolutional encoder with 18 blocks, starting with 32 filters that double every four layers, and produces a 512-dimensional embedding through a projection head. The PaPaGei-S variant additionally includes two mixture-of-expert heads for refining morphology-related indices (sVRI, IPA, SQI), resulting in approximately 5M parameters overall. Pretraining is performed on 57,000 hours of PPG data (around 20M segments) from large clinical datasets such as MIMIC-III (Johnson et al., 2016) and MESA (Chen et al., 2015), using

a morphology-aware self-supervised objective with augmentations including cropping, Gaussian noise, flipping, negation, and magnitude scaling.

For SimCLR (Chen et al., 2020), we adopt the same ResNet-18 backbone as PaPaGei (∼5M parameters) and apply the standard NT-Xent loss (Sohn, 2016; Chen et al., 2020) with a temperature of $\tau = 0.2$. The augmentation pipeline includes random cropping (0.5), time flipping (0.2), negation (0.2), scaling (0.4), and Gaussian noise (0.35). For TF-C (Zhang et al., 2022a), we likewise use a ResNet-18 encoder, with a total of ∼10M parameters since two encoders are employed, and train with a time-frequency contrastive loss. Both SimCLR and TF-C are optimized using Adam with a base learning rate of 1×10^{-4}, a weight decay of 1×10^{-5}, a batch size of 128, and cosine learning rate scheduling. For the Masked Siamese Network (MSN) (Assran et al., 2022), we adopt transformer encoders with 4 attention heads and 12 layers. The latent dimension of the attention layers is 128, and the feedforward networks have a hidden size of 512, resulting in approximately 2.5M parameters overall. MSN is trained with a base learning rate of 5×10^{-4}, a weight decay of 1×10^{-4}, and linear warmup. Pretraining is performed on the same smartwatch PPG dataset as MMR, using a patch size of 10 and a masking ratio of 0.75.

As a lightweight baseline, we also implement the Statistical Features approach similar to (Pillai et al., 2024). Each 10s PPG segment is represented by a handcrafted feature vector: mean, standard deviation, 25th percentile, 50th percentile (median), 75th percentile, minimum, and maximum values (i.e., [mean, std, p_{25}, p_{50}, p_{75}, min, max]). These features are computed per segment and directly used as input to a random forest classifier or regressor, depending on the downstream task.

Finally, we include the open-source Chronos-T5 (Base, 200M parameters) (Ansari et al., 2024) as a large-scale time-series foundation model baseline. Chronos is based on the T5 encoder–decoder transformer architecture, adapted for time series by quantizing values into discrete tokens and applying mean-scaling normalization. It is trained autoregressively with cross-entropy loss on a large collection of public datasets and synthetic series, using additional augmentation strategies such as TSMixup. Chronos represents a state-of-the-art zero-shot time-series model that is substantially larger than our other baselines, providing a complementary point of comparison for evaluating scale and domain adaptation effects.

A.4 Dataset details

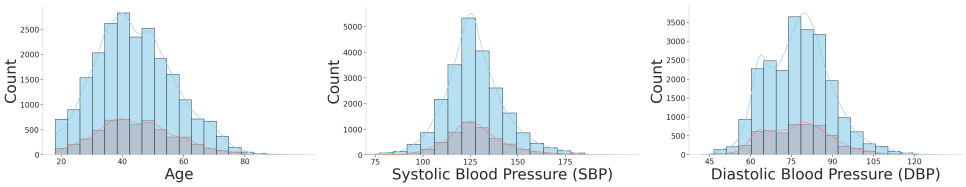


Figure 6: Hypertension Free Living Data Statistics

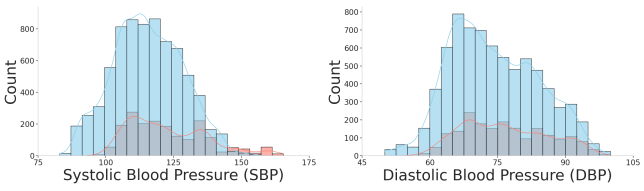


Figure 7: Hypertension Lab Data Statistics

We pretrain on data collected from various types of [REDACTED] smartwatches where PPG is sampled at different sampling frequencies (e.g., 25-100 Hz). These datasets provide diverse signals collected under [REDACTED] distinct studies and user groups. Such data closely reflect real-world conditions, making them highly representative for PPG-based wearable applications. We evaluate on several downstream datasets collected in diverse settings. For probing tasks, we train random forest classifiers and regression models using 5-fold cross-validation. We adopt a grouped and stratified 5-fold cross-validation procedure. This approach ensures that all data from the same subject is

Table 4: Downstream datasets. Counts are (#positive / #negative) segments

Task	Setting	Positive	Negative
Hypertension	Lab (protocol)	646	3,595
	Naturalistic (field)	5,952	8,189
PVC Detection	Naturalistic (field)	40,780	442,484
Laboratory Tests			
Sodium	Clinical reports	9,714	8,721
Potassium	Clinical reports	10,103	9,509
Creatinine	Clinical reports	8,176	8,879
Carbon Dioxide	Clinical reports	7,846	9,141
Hemoglobin	Clinical reports	6,429	6,286
White Blood Cells	Clinical reports	7,384	6,380

contained within a single fold to prevent data leakage, while also balancing the distribution of target classes across folds.

Hypertension Classification We define hypertension as a binary classification task based on clinical guidelines: individuals are labeled as *Hypertensive* (label 1) if their systolic blood pressure is $\geq$ 130 mmHg or diastolic blood pressure is $\geq$ 80 mmHg, and *Normal* (label 0) otherwise. We apply buffer thresholds of ± 8 mmHg around the diagnostic cutoffs.

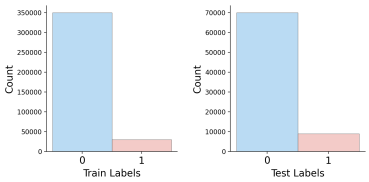


Figure 8: PVC label distributions.

PVC Detection Premature Ventricular Contractions (PVCs) are early heartbeats originating in the ventricles (Cha et al., 2012). They may indicate underlying cardiac conditions or an increased risk of arrhythmias. Prior studies (Han et al., 2020; Solosenko & Marozas, 2014) have investigated the detection of PVCs using PPG data. In our setting, we label high PVC burden as class 1 and low PVC burden as class 0.

Laboratory Tests For various laboratory tests (explained below as per (National Library of Medicine (US), 2020)), we adopt a binary classification scheme where abnormal values are labeled as class 1 and class 0 otherwise.

- Sodium: Elevated sodium (hypernatremia) is linked to dehydration or adrenal gland/kidney dysfunction.
- Potassium: High potassium (hyperkalemia) may cause cardiac arrhythmias; low potassium (hypokalemia) is associated with muscle weakness, fatigue, and rhythm disturbances.
- Creatinine: Elevated creatinine reflects impaired kidney function and may indicate acute kidney injury, chronic kidney disease, or other kidney problems.
- Carbon Dioxide: Abnormal carbon dioxide levels suggest an acid–base imbalance. Low levels may indicate metabolic acidosis or respiratory alkalosis, whereas high levels may reflect metabolic alkalosis or chronic respiratory acidosis.
- Hemoglobin: Low hemoglobin levels may indicate anemia, which can result from iron deficiency, chronic disease, or blood loss.
- White Blood Cells (WBC): Elevated WBC (leukocytosis) may be seen with infection, inflammation, stress, or blood disorders. Low WBC (leukopenia) can indicate bone marrow suppression, viral infection, or autoimmune conditions.

A.5 Example Visualization of Discrete Wavelet Transforms of Wearable PPG Signal

In this section, we present visualizations illustrating the transformation of PPG signals into DWT coefficients (Fig. 9) and their subsequent processing within our masked modeling framework. The patchified coefficient maps are heavily masked, and the MMR encoder–decoder is trained to reconstruct the missing subbands (Fig. 10). These examples demonstrate how the DWT captures multi-resolution structure and how MMR exploits this representation to recover meaningful time–frequency information from incomplete inputs.

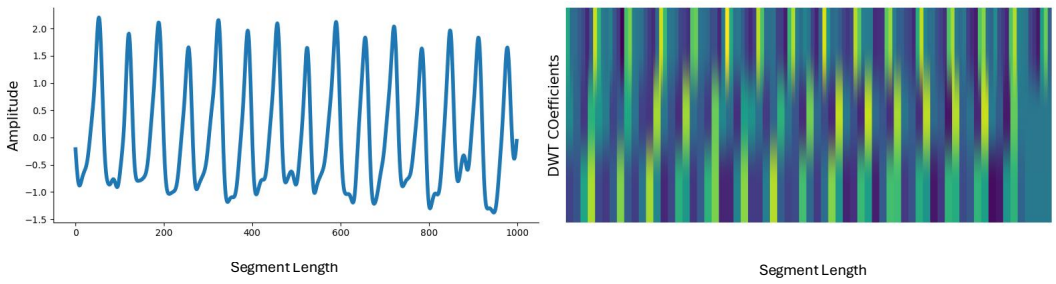


Figure 9: Example PPG signal (left) and its corresponding 2-D representation of discrete wavelet transform (DWT) coefficients (right). The DWT decomposes the signal into multi-resolution subbands, where higher-frequency detail coefficients appear at the top and the low-frequency approximation band at the bottom, providing a time–frequency representation.

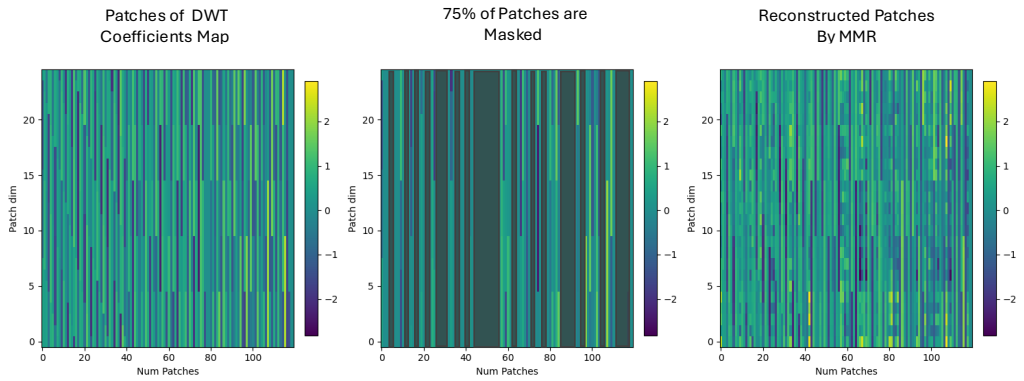


Figure 10: Illustration of the masked modeling framework applied to DWT coefficients. Left: original patchified DWT coefficient map of the PPG signal. Middle: masked input where 75% of patches are randomly removed. Right: reconstructed patches generated by the MMR model, demonstrating its ability to recover missing time–frequency structure from partial observations.

A.6 Additional Classification Results for MMR and Best Performing Baselines

Table 5 and Table 6 report the full set of AUROC scores for model scaling and data scaling experiments. These results show that larger models (27M parameters) achieve the highest performance on several endpoints, including Hypertension-Lab and Sodium prediction, while the mid-sized model remains competitive across most tasks. Similarly, increasing pretraining data size from 1M to 17M segments yields consistent gains for hypertension and PVC detection, with more modest improvements for certain lab biomarkers such as WBC and creatinine.

Classification AUROC (↑)	**2 Million**	**7 Million**	**27 Million**
Hypertension-Lab	71.50	71.18	**80.54**
Hypertension	67.08	**68.12**	66.65
PVC Detection	81.30	**82.04**	81.41
Laboratory Tests			
Carbon Dioxide	**63.95**	63.10	62.57
Creatinine	57.28	54.82	**58.26**
Hemoglobin	53.19	52.63	**54.57**
Potassium	73.23	**74.27**	71.34
Sodium	57.08	60.54	**64.45**
White Blood Cells	75.53	75.71	**76.12**

Table 5: **MMR Model Scaling Results:** This table reports *mean AUROC* on the 5-fold cross-validation performance metrics for MMR models of increasing scale (2 million, 7 million, and 27 million parameters). Each column represents a model size, while rows correspond to the predictive tasks (e.g., hypertension, PVC, WBC).

Classification AUROC (↑)	**1 Million**	**5 Million**	**17 Million**
Hypertension-Lab	67.99	69.06	**71.18**
Hypertension	66.08	67.50	**68.12**
PVC	76.62	79.84	**82.04**
Laboratory Tests			
Carbon Dioxide	**63.80**	63.54	63.10
Creatinine	**59.60**	54.60	54.82
Hemoglobin	52.21	53.89	52.63
Potassium	74.03	74.92	74.27
Sodium	57.69	57.25	**60.54**
White Blood Cells	**76.80**	75.16	75.71

Table 6: **MMR Data Scaling Results:** This table reports *mean AUROC* on the 5-fold cross-validation performance metrics for MMR model with increasing pretraining data scale (1 million, 5 million, and 17 million segments). Each column represents pretraining data size, while rows correspond to the predictive tasks (e.g., hypertension, PVC, WBC).

A.7 Analyzing Learned Representations for Downstream Tasks

We present t-SNE visualizations of patient-level embeddings learned by MMR and PaPaGei for the Hypertension task (Free Living). Figure 11 (left) shows the embeddings colored by age bins. While the different age groups are not perfectly separable, the latent space reveals slight clustering, with patients above 50 years more prominently shifting toward the left side of the embedding. This indicates that the model encodes some age-related demographic information. We found that, beyond patient-level discriminability, the learned embeddings capture clear physiological structure. As shown in Figure 11, t-SNE visualizations of participant-level mean embeddings colored by heart rate reveal a smooth gradient from low to high values. This indicates that the MMR model encodes physiologically meaningful variation rather than random alignment. Such clustering is not observed in t-SNE of PaPaGei embeddings in Fig. 12.

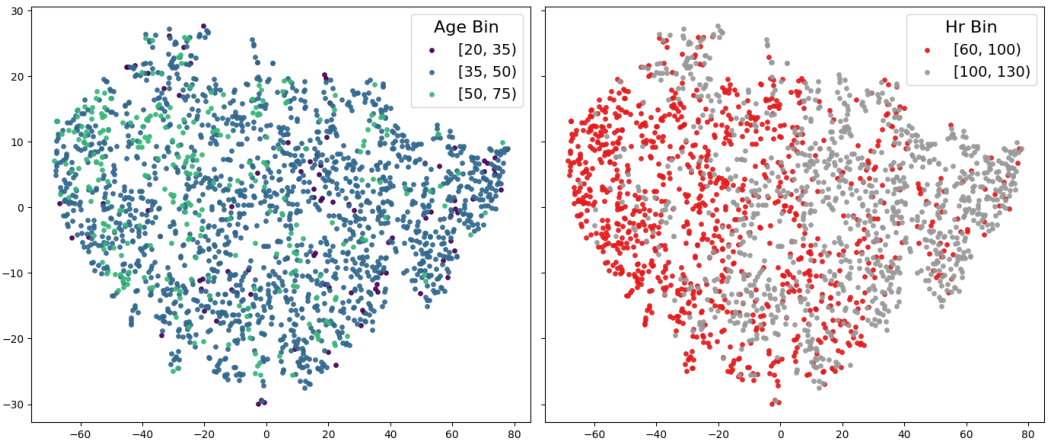


Figure 11: t-SNE visualization of patient-level embeddings (mean of all segments per patient) learned by MMR for the Hypertension Task – Free Living. Left: embeddings colored by age bins show only slight clustering, consistent with prior observations (Narayanswamy et al., 2024). Right: embeddings colored by heart rate bins reveal a gradient, indicating that the representations capture meaningful physiological variability.

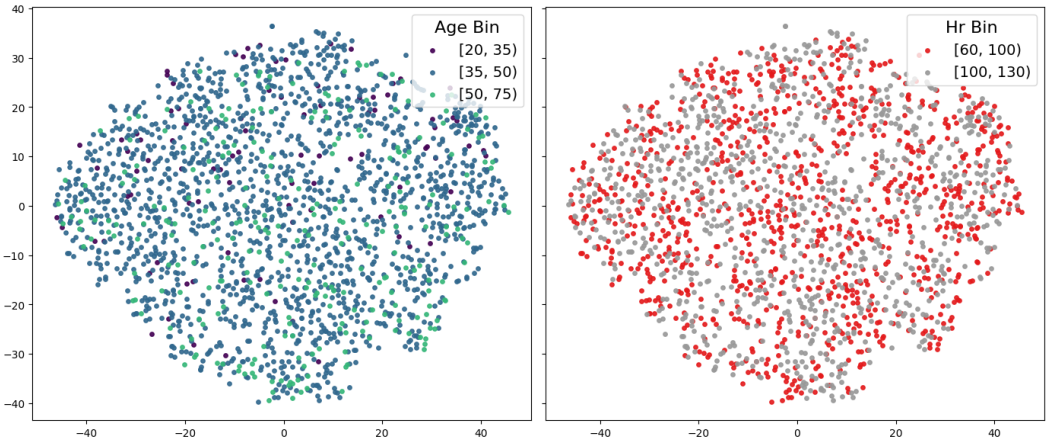


Figure 12: t-SNE visualization of PaPaGei embeddings for Hypertension Task – Free Living.