ECG2TOK: ECG Pre-Training with Self-Distillation Semantic Tokenizers

Xiaoyan Yuan[1,2] , **Wei Wang**[1,2,*] , **Han Liu**[3] , **Jian Chen**[1] , **Xiping Hu**[1,2]
[1]Artificial Intelligence Research Institute, Shenzhen MSU-BIT University
[2]School of Medical Technology, Beijing Institute of Technology
[3]School of Software, Dalian University of Technology
{3120235281, huxp}@bit.edu.cn, ehomewang@ieee.org,
liu.han.dut@gmail.com, chenj589@mail2.sysu.edu.cn

Abstract

Self-supervised learning (SSL) has garnered increasing attention in electrocardiogram (ECG) analysis for its effectiveness in resource-limited settings. Existing state-of-the-art SSL methods rely on time-frequency detail reconstruction, but due to the inherent redundancy of ECG signals and individual variability, these approaches often yield suboptimal performance. In contrast, discrete label prediction becomes a superior pre-training objective by encouraging models to efficiently abstract ECG high-level semantics. However, the continuity and significant variability of ECG signals pose a challenge in generating semantically discrete labels. To address this issue, we propose an ECG pre-training framework with a self-distillation semantic tokenizer (ECG2TOK), which maps continuous ECG signals into discrete labels for self-supervised training. Specifically, the tokenizer extracts semantically aware embeddings of ECG by self-distillation and performs online clustering to generate semantically rich discrete labels. Subsequently, the SSL model is trained in conjunction with masking strategies and discrete label prediction to facilitate the abstraction of high-level semantic representations. We evaluate ECG2TOK in six downstream tasks, demonstrating that ECG2TOK efficiently achieves state-of-the-art performance and up to a 30.73% AUC increase in low-resource scenarios. Moreover, visualization experiments demonstrate that the discrete labels generated by ECG2TOK exhibit consistent semantics closely associated with clinical features. Our code is available on https://github.com/YXYanova/ECG2TOK.

1 Introduction

Cardiovascular diseases (CVDs), responsible for 17.9 million annual deaths [Kyu *et al.*, 2018], are central to the UN's "Good Health and Well-being" Sustainable Development Goal. While electrocardiogram (ECG) analysis is critical for early CVD detection, its efficacy remains constrained in resource-limited settings due to inadequate healthcare infrastructure. Self-supervised learning (SSL) offers a solution by minimizing labeled data dependency, reducing costs, and improving diagnostic accessibility to advance healthcare equity [Krishnan *et al.*, 2022].

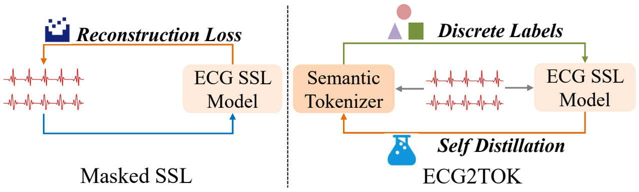

Figure 1: Comparison of ECG2TOK with other SSL methods.

Existing state-of-the-art ECG SSL methods rely on masked reconstruction in the time-frequency domain to learn semantic representations, which has notable limitations in the context of ECG. The inherent characteristics of ECG signals include the periodic repetition of specific waveforms (such as P waves, QRS complexes, and T waves), as well as significant individual differences in amplitude, duration, and morphology [Clifford, 2002; Yuan *et al.*, 2024; Yuan *et al.*, 2025]. However, reconstruction loss tends to emphasize low-level time-frequency features, leading to overfitting of repetitive waveform parts and individual-specific characteristics. Discrete label prediction may serve as a better pre-training objective compared to reconstruction, based on applications in other domains [Ramesh *et al.*, 2021; Bao *et al.*, 2021; Hsu *et al.*, 2021]. This is because discrete label prediction reduces the impact of information redundancy caused by periodic waveforms by mapping ECG signals with similar physiological meanings to the same label space. Additionally, discrete label prediction avoids over-reliance on individual-specific low-level time-frequency features (such as amplitude, duration, and morphology), allowing the model to better learn global semantic features related to heart health.

However, applying discrete label prediction to ECG SSL faces great challenges. On the one hand, ECG signals are continuous and lack discrete semantic units similar to words in natural language processing, which makes label prediction difficult. On the other hand, ECG data exhibits a high degree of variability, and characteristic waves show obvious temporal persistence and morphological differences. For example,

---

[*] Corresponding author.

semantically meaningful waves such as QRS complexes, T waves, and P waves vary greatly in duration and shape, increasing the complexity of extracting semantic token from ECG signals. These challenges make it difficult to effectively apply commonly used tokenizers (such as Vector Quantization [Peng *et al.*, 2022] and Clustering [Hsu *et al.*, 2021; Bao *et al.*, 2021]) to discrete label generation for ECG.

To address these challenges, we propose a novel framework, ECG2TOK, which integrates a self-distilled semantic tokenizer with a SSL model, employing an iterative optimization mechanism for ECG pretraining, as illustrated in Figure 1. Specifically, ECG2TOK utilizes a target encoder to extract context embeddings enriched with knowledge distillation and performs online clustering to refine the high-dimensional embedding space into a semantic codebook, generating corresponding discrete semantic labels. Discrete semantic label prediction and masking strategies are then used to guide SSL model training and enhance the ability of the context encoder to learn universal representations. The introduction of knowledge distillation strengthens the expression of semantics and improves the consistency of the discrete label generation process. ECG2TOK outperforms state-of-the-art ECG SSL methods across six downstream ECG classification tasks. Additionally, visualization experiments reveal that the discrete labels generated by the tokenizer exhibit consistent semantics closely related to clinical features. This demonstrates that ECG2TOK effectively generates discrete labels containing high-level semantics. By predicting these labels, the model gains the ability to overcome signal redundancy and individual variability, achieving efficient and robust ECG modeling and understanding.

Our contributions are as follows:

- We propose a novel ECG pretraining framework with self-distillation semantic tokenizers, which utilizes discrete label prediction loss and outperforms traditional reconstruction loss methods, opening new avenues for ECG pretraining.
- We provide an effective semantic tokenizer to quantize continuous ECG features into semantically compact discrete labels, facilitating future ECG pretraining and time series pretraining work.
- Our method achieves state-of-the-art results on six ECG classification benchmark tasks with minimal pretraining cost, and up to a 30.73% AUC improvement under low-resource conditions.

2 Related Work

2.1 ECG-based SSL

Recent advances in self-supervised learning for ECG have shown strong generalization capabilities, greatly enhancing downstream task performance [Lai *et al.*, 2023]. Among these, contrastive learning has gained significant attention. CLOCS [Kiyasseh *et al.*, 2021] promotes representation consistency across spatial, temporal, and patient dimensions, while ASTCL [Wang *et al.*, 2023] introduces adversarial spatiotemporal contrastive learning to improve robustness and semantic invariance. Other approaches [Lan *et al.*, 2022; Yang and Hong, 2022] explore inter- and intra-subject representations, as well as temporal and frequency domain alignment. However, contrastive learning methods often rely on complex pair construction and prior knowledge [Zhang *et al.*, 2022], limiting their applicability. In contrast, masked reconstruction methods explicitly encourage the model to capture global contextual information by predicting masked content. For example, CRT [Zhang *et al.*, 2023] proposes reconstructing both temporal and frequency domain data to discover cross-domain correlations and enhance representation learning. ST-MEM [Na *et al.*, 2024] adopts a similar masked modeling strategy. However, traditional reconstruction methods tend to focus on low-level time-frequency detail recovery, neglecting high-level semantic information, which limits their generalization capability. This highlights the need for incorporating high-level semantic modeling to further enhance the representation learning of ECG models. HeartLang [Jin *et al.*, 2025] treats ECG as a language with words (QRS waves) and sentences (rhythms), facilitating representation learning by predicting the vocabulary index of each word.

2.2 Discrete Labels-based SSL

Discrete label-based self-supervised learning (SSL) has made significant strides in natural language processing (NLP) [Lan, 2019; Kenton and Toutanova, 2019a; Liu, 2019], computer vision (CV) [Bao *et al.*, 2021; Peng *et al.*, 2022], and speech processing [Hsu *et al.*, 2021; Liu *et al.*, 2023]. In NLP, BERT [Kenton and Toutanova, 2019b] uses masked prediction to effectively learn representations from discrete input sequences. In CV, Beit [Peng *et al.*, 2022] introduce vector-quantized (VQ) visual tokenizers, enabling the generation of discrete tokens for masked image patches and the prediction of their original tokens. Similarly, HuBERT [Hsu *et al.*, 2021] employs iterative hidden state clustering to generate discrete labels for speech SSL tasks, showcasing the versatility of this approach across domains.

However, applying discrete label-based self-supervised learning (SSL) methods to ECG signals presents two major challenges. First, ECG signals are continuous-valued sequences, lacking discrete semantic units like phonemes in speech or words in NLP, making it difficult to directly apply predictive loss. Second, due to significant variations in waveform and rhythm across individuals, ECG signals contain excessively larger data variations, making it challenging to effectively capture unified semantic patterns using methods like VQ tokenizers or clustering. To address this, our approach incorporates self-distillation combined with clustering as a discrete tokenizer, enhancing the semantic representation capability before discretization.

3 Methodology

3.1 Overview

ECG2TOK pre-training framework integrates a semantic tokenizer and an ECG masked SSL model, leveraging discrete label prediction to optimize the pre-training process. As shown in Figure 2, the semantic tokenizer generates discrete semantic labels from unlabeled ECG data, which are then used to train the ECG SSL model through a combination of

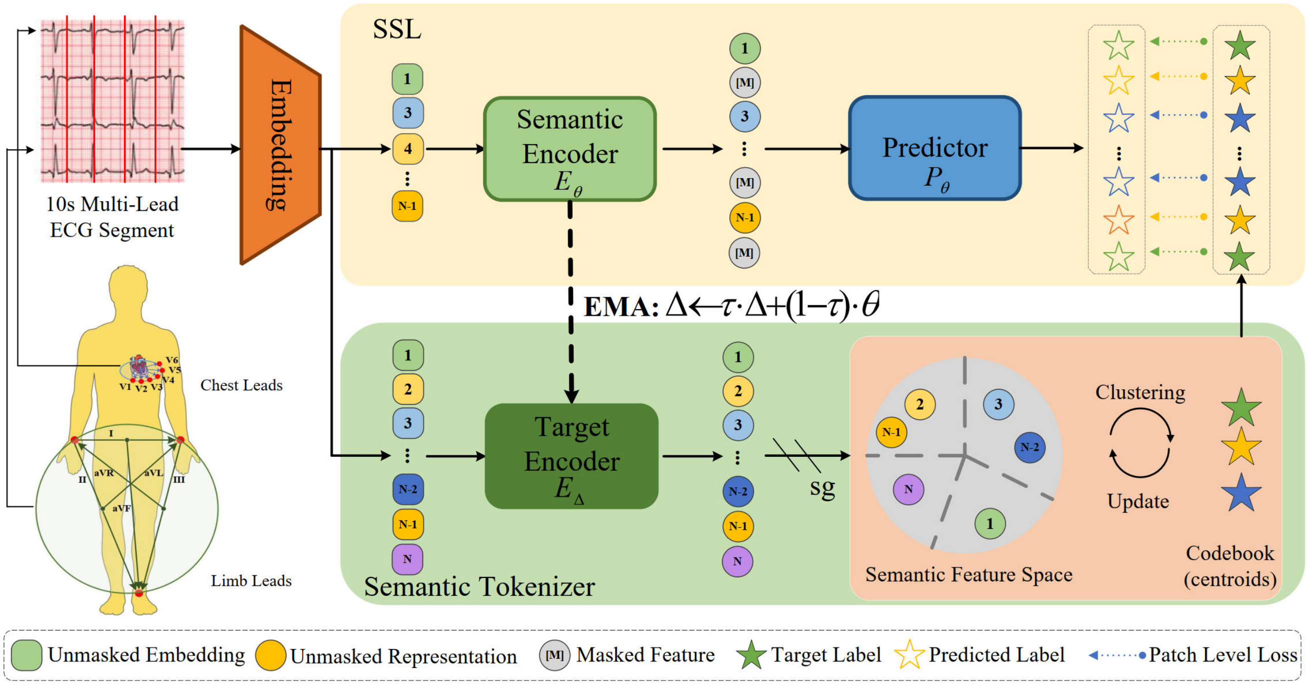

Figure 2: Overview of ECG2TOK: The target encoder, maintained as an exponential moving average of the context encoder, extracts target features from unmasked ECG signals. High-dimensional outputs from the target encoder are processed through online clustering, while the SSL learns to predict clusters index for each ecg patch. Both the target encoder and the clustering module (shaded areas) operate without gradient computation.

masked prediction and discrete label prediction tasks. During iterative training, ECG2TOK uses knowledge distillation to train a new semantic tokenizer, replacing the previous tokenizer in the next training iteration.

This process enables mutual enhancement between the semantic tokenizer and the SSL model: the tokenizer learns semantically rich representations from the SSL model, while the SSL model benefits from the discrete semantic labels generated by the tokenizer. Together, they produce robust and semantically aligned ECG representations.

3.2 Masked ECG SSL Model

We propose a masked modeling task for training ECG pretraining models, as illustrated in the upper part of Figure 2. Unlike traditional methods that rely on reconstructing the time-frequency details of input ECG signals for model optimization, our approach introduces a Transformer-based label predictor to replace the conventional decoder, predicting patch-level discrete labels. These predicted labels are combined with pseudo-labels generated by the semantic tokenizer (detailed in Section 3.3) to compute the loss, providing an innovative and efficient solution for ECG model pre-training.

Input Embedding. Inspired by ST-MEM [Na *et al.*, 2024], the raw ECG signal $\mathbf{I} \in R^{L\times 12}$ is divided into a set of non-overlapping patches $\mathbf{I}' \in R^{T\times(12\cdot p)}$, where $T = L/p$ and p represents the length of each patch. Subsequently, a learnable linear projection is applied to each patch, resulting in the embeddings:

$$\mathbf{E}_p = \text{LinearProj}(I') \in R^{T\times d_{model}}, \tag{1}$$

where E_p denotes the embedding for each patch, and d_{model} is the projected feature dimension.

To construct the final input, the patches embedding $\mathbf{E}_p$ is combined with a temporal positional embedding $\mathbf{E}_t \in R^{T\times 1\times d_{model}}$ and a spatial positional embedding $\mathbf{E}_s \in R^{1\times 12\times d_{model}}$:

$$\mathbf{X} = \mathbf{E}_p + \mathbf{E}_t + \mathbf{E}_s, \tag{2}$$

where X represents the final input patch embedding.

Masked Prediction. Given the input patch embedding $\mathbf{X} = \{x_t\}_{t=1}^{T}$, we apply random masking to 75% of the patches embedding. The masked positions in the patches are denoted by $\mathcal{M} = \{1,\ldots,T\}^{0.75T}$. Next, the unmasked patch $\mathbf{X}^U = \{x_t : t \in \mathcal{M}\}_{t=1}^{T}$ is fed into the ViT encoder E_θ, producing the encoded representations $\mathbf{R}^U = \{r_t : t \in \mathcal{M}\}_{t=1}^{T}$. The encoded non-masked patch features are then combined with the masked patch features, and passed to the label predictor P_θ. The predictor outputs the predicted discrete labels $\mathbf{Z} = \{z_t\}_{t=1}^{T}$. The calculation formula is as follows:

$$\mathbf{R}^U = E_\theta(\mathbf{X}^U), \tag{3}$$

$$\mathbf{Z} = \{z_t\}_{t=1}^{T} = P_\theta(\mathbf{R}^U, \{0 : t \notin \mathcal{M}\}_{t=1}^{T}), \tag{4}$$

where $\mathbf{X}^U$ is the unmasked patches embedding, $\mathbf{R}^U$ is the unmasked patches representation and the zero vector represents the masked patch features.

3.3 Self-Distilled Semantic Tokenizer

We propose a self-distilled semantic tokenizer that combines self-distillation with online clustering to quantize continuous

ECG features into discrete semantic labels. The framework leverages a target encoder to distill knowledge from the context encoder, generating high-dimensional semantic aware embeddings. Online clustering is then applied to these high-dimensional embeddings from target encoder to construct a codebook. The pseudo-labels are obtained through the codebook indices and are used to guide the self-supervised training of the semantic network. Unlike traditional offline clustering or vector quantization methods, our approach integrates online clustering directly into the training process, enabling dynamic adaptation between semantic representations and discrete labels. This significantly enhances the model's optimization efficiency and semantic consistency.

Target Encoder Parameterization. As illustrated in Figure 2, our method is inspired by data2vec [Baevski *et al.*, 2023] and adopts the same overall framework. The objective is to train the context encoder E_θ under the guidance of the target encoder E_Δ, where both models share the same architecture. In our work, this architecture is a 12-layer Transformer encoder [Vaswani, 2017].

The weights of the context encoder, denoted as θ, are updated using backpropagation based on the gradient of the loss function. The target encoder weights, represented by Δ, are initialized to be identical to the context encoder weights at the beginning of training. During training, the target encoder weights are updated through an **Exponentially Moving Average (EMA)** of the context encoder weights, expressed as:

$$\Delta \leftarrow \tau\Delta + (1-\tau)\theta, \tag{5}$$

where τ is a hyperparameter that governs the update frequency of the target encoder weights. The value of τ increases linearly throughout training, starting from an initial value τ_0 and gradually approaching 1. This scheduling allows the target encoder to adapt quickly during the early stages of training while stabilizing in later stages.

Codebook Assignment. After distilling high-dimensional semantic aware embeddings from the target encoder, we leverage online clustering to construct a codebook (set of centroids) $\mathbf{E} = \{\mathbf{e}_1, \ldots, \mathbf{e}_V\}$, where there are V codewords (centroids) $\mathbf{e}_i \in R^D$. To align the target encoder outputs with the codebook, each codebook entry v is associated with a set $\tilde{\mathbf{Z}}_v$, which contains the representations closest to the current centroid e_v. Formally, this is defined as:

$$\tilde{\mathbf{Z}}_v = \left\{ \tilde{\mathbf{z}}_t \;\middle|\; v = \arg\min_{i\in\{1,\ldots,V\}} \|\tilde{\mathbf{z}}_t - \mathbf{e}_i\|_2 \right\}, \tag{6}$$

where the set index v will be used as a pseudo label to train the semantic network.

Codebook Update. To ensure the codebook dynamically adapts to the target encoder outputs, each codeword $\mathbf{e}_v$ is updated using an EMA. The update rules are given by:

$$s_v \leftarrow \sigma s_v + (1-\sigma)\sum \tilde{\mathbf{Z}}_v, \tag{7}$$

$$n_v \leftarrow n_v + (1-\sigma)\left|\tilde{\mathbf{Z}}_v\right|, \tag{8}$$

$$\mathbf{e}_v \leftarrow \frac{s_v}{n_v}, \tag{9}$$

where s_v represents the accumulated sum of all neighboring target representations (i.e., $\tilde{\mathbf{Z}}_v$ as defined in Eq. 6), while n_v denotes the count of neighbors. Both terms are updated using an EMA mechanism with a decay rate σ, enabling the computation of the codeword $\mathbf{e}_v$ as the weighted average of its neighboring representations. During initialization, s_v, n_v, and $\mathbf{e}_v$ are all set to 1.

By defining codewords based on their corresponding neighboring representations, these codewords can be interpreted as semantic units derived from the target model in an unsupervised manner. Subsequently, these units are used to train the SSL model. The clustering approach produces discrete labels for ECG based on their context, which have consistent semantics. As shown in the section 5.4, the semantics of these codewords are closely consistent with those of the clinical patterns.

$$\mathcal{L}_{\text{p}} = \sum_{t=1}^{T} \log p_\psi\left(v|\mathbf{z}_t\right). \tag{10}$$

where ψ is the softmax activation over the codebook, v is the codeword index of the corresponding patch from the target network (i.e., $\tilde{\mathbf{z}}_t \in \tilde{\mathbf{Z}}_v$), and $\mathbf{z}_t$ is the corresponding output feature of the context network.

4 Experiments

4.1 Pre-training Configuration

MIMIC-IV-ECG. We use the MIMIC-ECG dataset [Gow *et al.*, 2023] to pre-train the proposed ECG2TOK framework. This dataset comprises 800,035 ECG recordings collected from 161,352 unique subjects, with each recording lasting 10 seconds and sampled at 500 Hz. To handle missing or invalid values (e.g., ”NaN” and ”Inf”) in the ECG data, we replaced them with the average of six neighboring points. After this preprocessing step, we obtained a final pre-training dataset containing 771,693 samples.

Implementation. MIMIC-IV-ECG dataset is split into training and validation sets in a 9:1 ratio, with the validation set used for semantic token analysis. During pre-training, a randomly initialized transformer [Vaswani, 2017] is employed as the target encoder, and input data is processed with a 75% random masking rate. The number of discrete semantic tokens generated through online clustering is set to 128, with a cluster center dimension of 768. The learning rate is configured to 1.5×10^{-4}, and the model is trained for a total of 7 epochs using the AdamW optimizer with momentum parameters [0.9, 0.95]. All experiments are conducted on a single NVIDIA A800 GPU with a batch size of 256. To ensure reproducibility, the random seed is fixed at 0.

4.2 Downstream Tasks Configuration

We evaluate our method on six tasks across three publicly available ECG datasets, covering over 100 cardiac conditions. Data splitting details are provided in the appendix.

PTB-XL. The PTB-XL dataset [Wagner *et al.*, 2020] contains 21,837 12-lead ECG recordings from 18,885 patients, each lasting 10 seconds and sampled at 500 Hz. The dataset supports classification into **Superclass** (5 classes), **Subclass**

(23 classes), **Form** (19 classes), and **Rhythm** (12 classes). We adopted the official split [Wagner *et al.*, 2020] for training, validation, and testing.

CPSC2018. The CPSC2018 dataset [Liu *et al.*, 2018] includes 6,877 12-lead ECG recordings with durations between 6 and 60 seconds. It contains 9 distinct labels and is divided into 70% for training, 10% for validation, and 20% for testing.

Chapman-Shaoxing-Ningbo (CSN). The CSN dataset [Zheng *et al.*, 2020; Zheng *et al.*, 2022] contains 45,152 12-lead ECG recordings, each lasting 10 seconds and sampled at 500 Hz. After excluding samples with "unknown" annotations, the refined dataset includes 23,026 recordings across 38 labels. We used a 70%:10%:20% split for training, validation, and testing.

Implementation. For linear probing, the target encoder remains frozen, and only the linear classifier with randomly initialized parameters is trained. To assess the method's performance under low-resource conditions, linear probing is conducted on each task using 1%, 10%, and 100% of the training dataset. The training process spans 100 epochs, including a 5-epoch warmup phase, with a base learning rate of 5.0×10^{-3}, a weight decay of 0.05, and a batch size of 16. The AdamW optimizer is utilized, with gradient clipping and distributed evaluation omitted. Binary cross-entropy (BCE) is employed as the loss function, suitable for multi-label classification. Macro AUC is used as the evaluation metric for all downstream tasks. To ensure reproducibility, the random seed is set to 0, and `torch.manual_seed` is fixed at 42. Test results are derived from the best-performing validation model.

5 Results and Discussion

5.1 Main Results

Table 1 presents the linear probing performance of our proposed ECG2TOK model across six tasks and three resource configurations, compared to the current state-of-the-art SSL methods. For each dataset, the **bold** numbers represent the best performance, while the underlined numbers indicate the second-best performance (this formatting is applied consistently throughout all tables in this paper). The results demonstrate that ECG2TOK achieves significant performance improvements across multiple tasks, obtaining the highest AUC in almost all resource settings (highlighted in bold in the table). Moreover, the AUC performance of ECG2TOK is particularly outstanding on the PTBXL-Rhythm, CPSC2018, and CSN datasets, with average AUC improvements of 16.4%, 9.43%, and 15.6%, respectively, over the next-best models. These results indicate that, compared to other SOTA methods, ECG2TOK exhibits superior representation capability, effectively extracting high-level semantic features and adapting to various downstream tasks.

Notably, under **low-resource** conditions (using only 1% and 10% of the training samples), ECG2TOK continues to exhibit superior performance compared to other models, highlighting its strong generalization ability in resource-constrained environments. For instance, with only 1% of the training data, ECG2TOK achieves an AUC of 81.16 on the PTBXL-Rhythm dataset, representing an improvement of approximately 30.73% over the next-best model, HeartLang. Similarly, on the CSN dataset, ECG2TOK achieves an improvement of around 19.66% over the next-best model, BarlowTwins. In some classification tasks, such as CPSC2018 and CSN, ECG2TOK outperforms other SSL methods trained on 100% of the data, even when using only 10% of the training data. These significant improvements suggest that ECG2TOK learns more advanced, semantically rich, and general representations, outperforming other self-supervised learning methods and further validating its robustness and effectiveness under low-resource conditions.

In summary, ECG2TOK consistently achieves superior performance across various downstream tasks, demonstrating its strong representation ability and broad adaptability in self-supervised learning. Its remarkable performance in low-resource scenarios underscores its robustness and effectiveness, making it highly suitable not only for large-scale datasets but also for practical applications in resource-constrained fields such as healthcare and health monitoring.

5.2 Pre-training Efficiency

The experimental results demonstrate that ECG2TOK significantly outperforms existing ECG self-supervised learning methods in both pre-training efficiency and performance. As shown in Table 2 and Figure 3, ECG2TOK achieves exceptional results with only 7 pre-training epochs, reducing pre-training time by 28.5 times compared to HeartLang and 9.12 times compared to MERL. Notably, ECG2TOK surpasses HeartLang (pre-trained for 200 epochs) and MERL (pre-trained for 50 epochs) after just 4 epochs, highlighting its ability to quickly learn high-quality representations. After 7 total pre-training epochs, ECG2TOK achieves an average AUC of 81.18 across six downstream tasks, significantly outperforming both models despite their longer pre-training periods.

The efficiency improvement of ECG2TOK can be attributed to three key factors. First, ECG2TOK employs a 75% high masking rate during pre-training. This large masking means that a significant portion of the ECG data is excluded before being input to the context encoder, enhancing batch processing capabilities and leveraging parallel computing advantages to improve efficiency. Second, the reconstruction objective in the proposed self-supervised framework differs from traditional ECG input space reconstruction objectives. It provides semantically rich discretized tokens (cluster center) as pre-training targets and encourages the model to discard redundant details, allowing the model to capture key features faster, thereby accelerating the convergence of the pre-training process. Lastly, the incorporation of self-distillation further speeds up the learning by refining the model's feature representations, allowing for faster and more efficient training.

5.3 Ablation Study

Training Target. Table 3 illustrates the importance of each component in the ECG2TOK framework. Removing the semantic tokenizer (**w/o semantic tokenizer**), where the model

	PTBXL-Super			PTBXL-Sub			PTBXL-Form			PTBXL-Rhythm			CPSC2018			CSN		
Method	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%
random init	70.45	77.09	81.61	55.82	67.60	77.91	55.82	62.54	73.00	46.26	62.36	79.29	54.96	71.47	78.33	47.22	63.17	73.13
SimCLR[Chen *et al.*, 2020]	63.41	69.77	73.53	60.84	68.27	73.39	54.98	56.97	62.52	51.41	69.44	77.73	59.78	68.52	76.54	59.02	67.26	73.20
BYOL[Grill *et al.*, 2020]	71.70	73.83	76.45	57.16	67.44	71.64	48.73	61.63	70.82	41.99	74.40	77.17	60.88	74.42	78.75	54.20	71.92	74.69
BarlowTwins[Zbontar *et al.*, 2021]	71.87	75.96	78.41	62.57	70.84	74.34	52.12	60.39	66.14	50.12	73.54	77.62	55.12	72.75	78.39	60.72	71.64	77.43
MoCo-v3[Chen *et al.*, 2021]	73.19	76.65	78.26	55.88	69.21	76.69	50.32	63.71	71.31	51.38	71.66	74.33	62.13	76.74	75.29	54.61	74.26	77.68
SimSiam[Chen and He, 2021]	73.15	72.70	75.63	62.52	69.31	76.38	55.16	62.91	71.31	49.30	69.47	75.92	58.35	72.89	75.31	58.25	68.61	77.41
TS-TCC[Eldele *et al.*, 2021]	70.73	75.88	78.91	53.54	66.98	77.87	48.04	61.79	71.18	43.34	69.48	78.23	57.07	73.62	78.72	55.26	68.48	76.79
CLOCS[Kiyasseh *et al.*, 2021]	68.94	73.36	76.31	57.94	72.55	76.24	51.97	57.96	72.65	52.38	71.88	76.31	59.59	77.78	77.49	54.38	71.93	76.13
ASTCL[Wang *et al.*, 2023]	72.51	77.31	81.02	61.86	68.77	76.51	44.14	60.93	66.99	52.38	71.98	76.05	57.90	77.01	79.51	56.40	70.87	75.79
CRT[Zhang *et al.*, 2023]	69.68	78.24	77.24	61.98	70.82	78.67	46.41	59.49	68.73	47.44	73.52	74.41	58.01	76.43	82.03	56.21	73.70	78.80
ST-MEM[Na *et al.*, 2024]	61.12	66.87	71.36	54.12	57.86	63.59	55.71	59.99	66.07	51.12	65.44	74.85	59.69	63.32	70.39	59.77	66.87	71.36
HeartLang[Jin *et al.*, 2025]	78.94	85.59	87.52	64.68	79.34	**88.91**	**58.70**	63.99	80.23	62.08	76.22	90.34	60.44	66.26	77.87	57.94	68.93	82.49
ECG2TOK(Ours)	**81.23**	**85.68**	**87.99**	**72.31**	**79.42**	84.52	55.88	**73.78**	**85.28**	**81.16**	**89.41**	**91.41**	**68.42**	**83.41**	**91.04**	**72.66**	**85.14**	**92.84**

Table 1: Linear probing performance comparison between our method and other eSSL approaches. The highest AUC are highlighted in **bold**, while the number underlined represents the second best.

Model	Epoch	Hour × GPU	Speedup	AUC(Mean±SE)
ST-MEM [Na *et al.*, 2024]	800	800	1×	63.31±1.48
HeartLang [Jin *et al.*, 2025]	200	200	4×	73.92±2.53
MERL [Liu *et al.*, 2024]	50	64	12.5×	78.11±2.55
ECG2TOK(Ours)	**7**	**7**	**114×**	**81.18±2.18**

Table 2: Comparison of Pre-training Costs with Other SOTA eSSL Models: All models are uniformly fine-tuned using linear methods on downstream datasets. The results are presented as (average AUC ± standard error) across six downstream tasks in three resource settings.

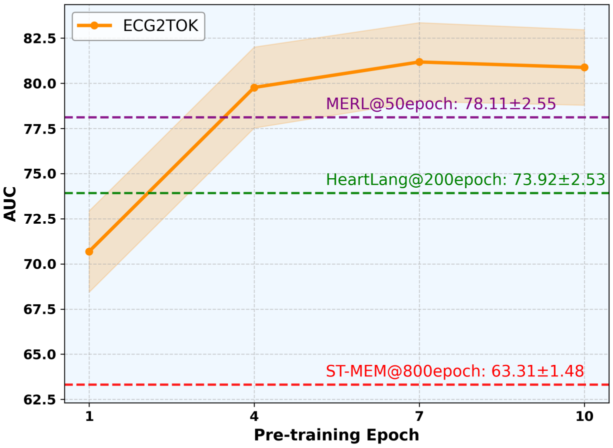

Figure 3: Comparison of our method (ECG2TOK) with three other eSSL approaches in terms of pre-training epochs, with ECG2TOK pre-trained for 10 epochs. All models are fine-tuned uniformly across six downstream tasks, and the results are averaged over the evaluation sets.

reverts to a traditional masked reconstruction of raw signals, results in the most significant performance drop of 10.1%. Excluding self-distillation (**w/o self-distillation**), where ECG's continuous input space is directly discretized into labels without knowledge compression from a target model, leads to a performance decrease of 4.13%. Finally, removing the clustering module (**w/o clustering**), which shifts the training objective to reconstructing the abstract feature rather than clustering-based discretization, results in a drop of 3.79%. Overall, the full ECG2TOK model achieves the best performance (average AUC: 81.18), demonstrating the complementary contributions of these components in effectively capturing and leveraging ECG signal representations.

Model	Training Targets	AUC(Mean±SE)
w/o semantic tokenizer	input reconstruction	72.98±2.25
w/o self-distilltion	discrete label prediction	77.82±2.21
w/o clustering	features reconstruction	78.10±2.39
ECG2TOK(Ours)	discrete label prediction	**81.18±2.18**

Table 3: Ablation Study of EEG2TOK Components: 'Mean' refers to the average AUC calculated from 18 results across six downstream tasks in three resource settings, while 'SE' denotes the standard error.

Masking Ratio. We investigate the impact of random masking rates during pre-training. The masking rate largely determines the number of unmasked patches accessible to the predictor during training, which directly affects the complexity of the pre-training task. Figure 4 illustrates how fine-tuning performance on downstream tasks changes with different masking rates. Notably, when the masking rate is set to 75%, our proposed ECG2TOK model achieves the best fine-tuning performance, indicating that at this moderately high masking rate, the model is able to effectively transfer and optimize downstream tasks while maintaining a higher pre-training task complexity.

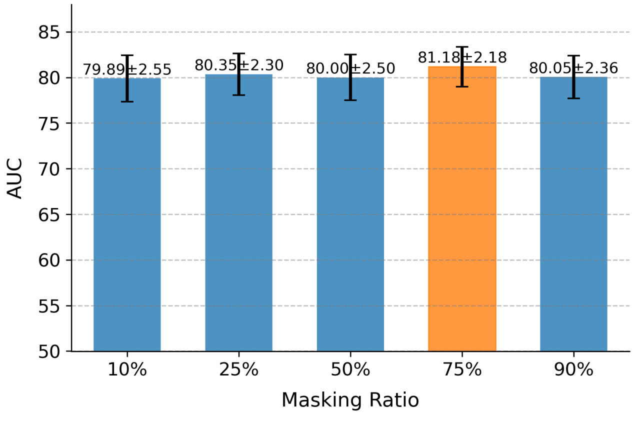

Figure 4: AUC at different masking ratios. The x-axis represents masking ratios, and the y-axis shows the corresponding average AUC values. The error bars represent standard deviations. The orange bar highlights the best-performing ratio.

Cluster size	1%	10%	100%
32	71.29 ± 3.72	82.45 ± 1.88	88.31 ± 1.31
128	**71.94 ± 3.51**	**82.74 ± 2.06**	**88.85 ± 1.29**
512	68.49 ± 3.62	79.08 ± 2.26	86.10 ± 1.55
1024	69.35 ± 3.55	79.68 ± 2.32	86.88 ± 1.41

Table 4: Performance comparison with varying codebook size V.

Number of Clusters. We investigate the impact of different codebook sizes (Cluster size) on the performance of the ECG2TOK model. The results in Table 4 show that codebook size significantly affects model performance. A small codebook (e.g., 32) leads to lower performance, likely due to insufficient feature representation. Performance improves with larger codebooks, peaking at a size of 128, which balances feature representation and computational complexity. However, further increasing the codebook size to 512 or 1024 results in a slight decline, possibly due to dispersed features and reduced generalization. In conclusion, a codebook size of 128 is found to be the most optimal, as it effectively captures data features and provides the best performance under varying resource conditions.

5.4 Visualization

To validate whether the semantic tokenizer in ECG2TOK captures clinically valuable information, we perform inference on the MIMIC-IV test set, which includes annotated clinical text reports and ECG signals from diverse patients. We cluster the ECG patches and visualize the 10 patches closest to each cluster center from different subjects, displaying their corresponding clinical reports for comparison. Each label is presented as ”P xx(ID xx):...xxx...”, where ”P” denotes the ECG patch index, ”ID” is the subject ID, and ”...xxx...” is the clinical report. By comparing the ECG waveforms with the associated clinical reports, we assess whether the discrete labels accurately reflect clinical features, thus confirming the effectiveness of the semantic tokenizer.

As shown in Figure 5, ECG patches sharing the same label or cluster center exhibit consistent semantic patterns aligned with clinical phenomena. For example, patch ”37” shows ST segment changes, indicative of myocardial ischemia, while ”47” demonstrates QT interval prolongation, a sign of arrhythmias or drug effects. Patch ”67” presents normal waveforms, indicating a healthy heart. These results suggest that the generated labels capture clinically relevant ECG features and contain meaningful, consistent semantic information.

The consistent semantic patterns across different subjects further indicate that the pseudo-labeling process is robust and generalizes well across clinical scenarios. This validates that the semantic tokenizer effectively captures high-level clinical patterns, not just memorizing patient-specific data, confirming the generated discrete labels reflect the clinical semantics of the ECG signals and are useful for ECG analysis.

In conclusion, these findings demonstrate that the semantic tokenizer in ECG2TOK effectively capture consistent and clinically relevant patterns in ECG signals, validating the practicality of ECG2TOK for ECG analysis.

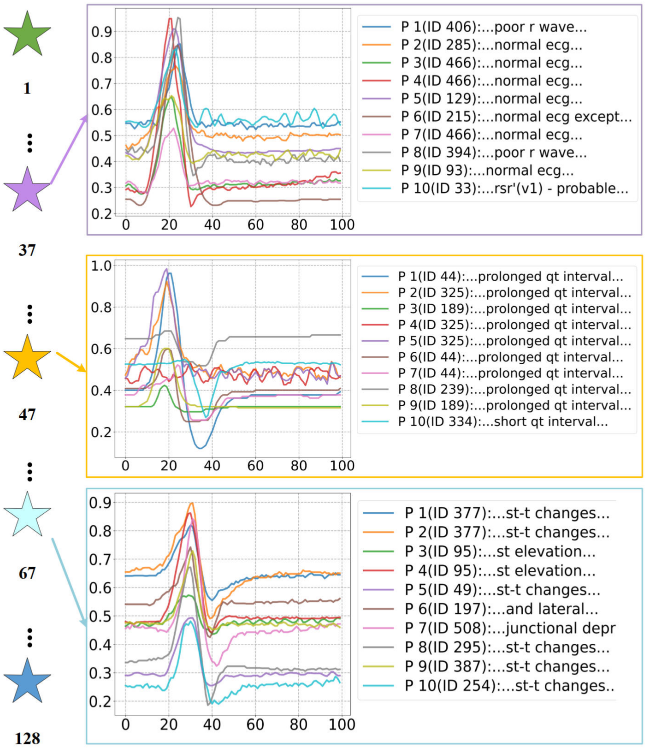

Figure 5: Visualization of the correspondence between discrete labels and input ECG patches. ”P” represents the patch index, ”ID” denotes the subject ID, and the text following the ”:” is the corresponding ECG report. ECG patches with the same label exhibit consistent semantic patterns.

6 Conclusion

In this paper, we present ECG2TOK, a novel ECG pretraining framework that incorporates a self-distillation semantic tokenizer. Unlike conventional approaches that focus on time-frequency signal reconstruction, ECG2TOK leverages discrete label prediction to drive the model towards higher-level semantic abstraction. The key innovation of ECG2TOK lies in the self-distillation semantic tokenizer, which transforms continuous ECG signals into a compact semantic space and generates discrete labels for self-supervised learning. Our experimental results demonstrate that ECG2TOK surpasses state-of-the-art methods across six downstream classification tasks, achieving up to a 30.73% improvement in AUC under low-resource conditions. Furthermore, the self-distillation framework contributes to enhanced pretraining efficiency, accelerating the learning process without sacrificing performance. Visualization analysis further highlights that the semantic tokenizer produces discrete labels with consistent and meaningful semantics. These findings underscore the potential of self-supervised learning with discrete label prediction in overcoming challenges such as individual variability and redundancy in ECG analysis, providing innovative solutions to improve healthcare access and efficiency, particularly in resource-limited settings.

Acknowledgments

This work was supported in part by the Sustainable Development Science and Technology Project of Shenzhen Science and Technology Innovation Commission under Grant KCXFZ20201221173411032.

References

[Baevski *et al.*, 2023] Alexei Baevski, Arun Babu, Wei-Ning Hsu, and Michael Auli. Efficient self-supervised learning with contextualized target representations for vision, speech and language. In *International Conference on Machine Learning*, pages 1416–1429. PMLR, 2023.

[Bao *et al.*, 2021] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[Chen and He, 2021] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.

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

[Chen *et al.*, 2021] 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*, pages 9640–9649, 2021.

[Clifford, 2002] G Clifford. *Signal processing methods for heart rate variability*. PhD thesis, Oxford University, UK, 2002.

[Eldele *et al.*, 2021] Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan. Time-series representation learning via temporal and contextual contrasting. *arXiv preprint arXiv:2106.14112*, 2021.

[Gow *et al.*, 2023] Brian Gow, Tom Pollard, Larry A Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Seth Berkowitz, Dana Moukheiber, Parastou Eslami, et al. Mimic-iv-ecg-diagnostic electrocardiogram matched subset. *Type: dataset*, 2023.

[Grill *et al.*, 2020] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

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

[Jin *et al.*, 2025] Jiarui Jin, Haoyu Wang, Hongyan Li, Jun Li, Jiahui Pan, and Shenda Hong. Reading your heart: Learning ecg words and sentences via pre-training ecg language model. *arXiv preprint arXiv:2502.10707*, 2025.

[Kenton and Toutanova, 2019a] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota, 2019.

[Kenton and Toutanova, 2019b] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota, 2019.

[Kiyasseh *et al.*, 2021] 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*, pages 5606–5615. PMLR, 2021.

[Krishnan *et al.*, 2022] Rayan Krishnan, Pranav Rajpurkar, and Eric J Topol. Self-supervised learning in medicine and healthcare. *Nature Biomedical Engineering*, 6(12):1346–1352, 2022.

[Kyu *et al.*, 2018] Hmwe Hmwe Kyu, Degu Abate, Kalkidan Hassen Abate, Solomon M Abay, Cristiana Abbafati, Nooshin Abbasi, Hedayat Abbastabar, Foad Abd-Allah, Jemal Abdela, Ahmed Abdelalim, et al. Global, regional, and national disability-adjusted life-years (dalys) for 359 diseases and injuries and healthy life expectancy (hale) for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet*, 392(10159):1859–1922, 2018.

[Lai *et al.*, 2023] Jiewei Lai, Huixin Tan, Jinliang Wang, Lei Ji, Jun Guo, Baoshi Han, Yajun Shi, Qianjin Feng, and Wei Yang. Practical intelligent diagnostic algorithm for wearable 12-lead ecg via self-supervised learning on large-scale dataset. *Nature Communications*, 14(1):3741, 2023.

[Lan *et al.*, 2022] Xiang Lan, Dianwen Ng, Shenda Hong, and Mengling Feng. Intra-inter subject self-supervised learning for multivariate cardiac signals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4532–4540, 2022.

[Lan, 2019] Zhenzhong Lan. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[Liu *et al.*, 2018] Feifei Liu, Chengyu Liu, Lina Zhao, Xiangyu Zhang, Xiaoling Wu, Xiaoyan Xu, Yulin Liu, Caiyun Ma, Shoushui Wei, Zhiqiang He, et al. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. *Journal of Medical Imaging and Health Informatics*, 8(7):1368–1373, 2018.

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

[Liu *et al.*, 2024] Che Liu, Zhongwei Wan, Cheng Ouyang, Anand Shah, Wenjia Bai, and Rossella Arcucci. Zero-shot ecg classification with multimodal learning and test-time clinical knowledge enhancement. *arXiv preprint arXiv:2403.06659*, 2024.

[Liu, 2019] Yinhan Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364, 2019.

[Na *et al.*, 2024] Yeongyeon Na, Minje Park, Yunwon Tae, and Sunghoon Joo. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. *arXiv preprint arXiv:2402.09450*, 2024.

[Peng *et al.*, 2022] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022.

[Ramesh *et al.*, 2021] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

[Vaswani, 2017] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[Wagner *et al.*, 2020] Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I Lunze, Wojciech Samek, and Tobias Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific data*, 7(1):1–15, 2020.

[Wang *et al.*, 2023] Ning Wang, Panpan Feng, Zhaoyang Ge, Yanjie Zhou, Bing Zhou, and Zongmin Wang. Adversarial spatiotemporal contrastive learning for electrocardiogram signals. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[Yang and Hong, 2022] Ling Yang and Shenda Hong. Unsupervised time-series representation learning with iterative bilinear temporal-spectral fusion. In *International conference on machine learning*, pages 25038–25054. PMLR, 2022.

[Yuan *et al.*, 2024] Xiaoyan Yuan, Wei Wang, Xiaohe Li, Yuanting Zhang, Xiping Hu, and M Jamal Deen. Catransformer: A cycle-aware transformer for high-fidelity ecg generation from ppg. *IEEE Journal of Biomedical and Health Informatics*, 2024.

[Yuan *et al.*, 2025] Xiaoyan Yuan, Wei Wang, Junxin Chen, Kai Fang, Ali Kashif Bashir, Tapas Mondal, Xiping Hu, and M Jamal Deen. Enhancing multi-label ecg classification via task-guided lead correlations in internet of medical things. *IEEE Internet of Things Journal*, 2025.

[Zbontar *et al.*, 2021] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pages 12310–12320. PMLR, 2021.

[Zhang *et al.*, 2022] Huaicheng Zhang, Wenhan Liu, Jiguang Shi, Sheng Chang, Hao Wang, Jin He, and Qijun Huang. Maefe: Masked autoencoders family of electrocardiogram for self-supervised pretraining and transfer learning. *IEEE Transactions on Instrumentation and Measurement*, 72:1–15, 2022.

[Zhang *et al.*, 2023] Wenrui Zhang, Ling Yang, Shijia Geng, and Shenda Hong. Self-supervised time series representation learning via cross reconstruction transformer. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[Zheng *et al.*, 2020] Jianwei Zheng, Huimin Chu, Daniele Struppa, Jianming Zhang, Sir Magdi Yacoub, Hesham El-Askary, Anthony Chang, Louis Ehwerhemuepha, Islam Abudayyeh, Alexander Barrett, et al. Optimal multi-stage arrhythmia classification approach. *Scientific reports*, 10(1):2898, 2020.

[Zheng *et al.*, 2022] J Zheng, H Guo, and H Chu. A large scale 12-lead electrocardiogram database for arrhythmia study (version 1.0. 0). *PhysioNet 2022Available online httpphysionet orgcontentecg arrhythmia10 0accessed on*, 23, 2022.