EP-Prior: Interpretable ECG Representations via Electrophysiology Constraints

Anonymous Author(s)
Anonymous Institution
anonymous@example.com

Abstract

Accurate ECG classification with limited labeled data requires representations that are both sample-efficient and clinically interpretable. We present EP-Prior, a self-supervised framework that aligns learned ECG representations with cardiac electrophysiology. The encoder produces a prescribed, wave-factorized latent space $(z_P, z_{QRS}, z_T, z_{HRV})$, and an EP-constrained Gaussian-wave decoder reconstructs signals under soft priors on P/QRS/T ordering, refractory periods, and physiologic duration bounds. This structure makes each component inspectable and supports factor-level probes and interventions. We provide a PAC-Bayes analysis that interprets EP constraints as an electrophysiology-informed prior, reducing the complexity term and predicting the largest gains in few-shot regimes. On PTB-XL, EP-Prior improves 10-shot AUROC from 0.627 to 0.699 (+0.072) over a capacity-matched generic SSL baseline, with gains across all five diagnostic superclasses. Ablations show the constraints are necessary: removing the EP losses drops 10-shot AUROC to 0.519, below the baseline. EP-Prior illustrates how embedding domain structure as model priors can jointly improve explainability and label efficiency.

1 Introduction

ECG-based cardiac diagnosis is critical for early detection of arrhythmias and conduction abnormalities [Goldberger *et al.*, 2000]. While deep learning has achieved strong performance on large datasets [Wagner *et al.*, 2020], significant challenges remain in low-data regimes, particularly for wearable and portable devices:

- Many clinically important arrhythmias appear in $<1\%$ of records
- Personalizing models to individual patients requires adaptation from only a handful of examples
- Deploying to new ECG hardware often means starting over with limited labeled data

Equally important is the need for **interpretability**. Black-box models that achieve high accuracy but provide no insight into *what* they have learned face barriers to clinical adoption. Regulatory frameworks increasingly require explainable AI for medical devices.

Cardiac electrophysiology provides rich mathematical structure: P-QRS-T wave morphology, conduction dynamics, refractory constraints. This knowledge is well-understood clinically but rarely exploited in representation learning. Prior work uses EP knowledge in ECGI (inverse problems) but not for learning interpretable representations.

We propose **EP-Prior**, which injects this EP knowledge as architectural priors in a self-supervised framework:

1. A **structured latent space** where encoder outputs decompose into $(z_P, z_{QRS}, z_T, z_{HRV})$
2. An **EP-constrained decoder** using a Gaussian wave model that reconstructs ECG signals
3. **Soft constraint losses** enforcing temporal ordering, refractory periods, and duration bounds

Our contributions are threefold. First, we achieve **interpretability** through a structured latent space with physiologically meaningful components, validated via intervention tests and concept predictability. Second, we provide a **theoretical** PAC-Bayes-motivated analysis explaining *why* EP constraints help in low-data regimes. Third, we demonstrate **empirically** competitive few-shot classification on PTB-XL with inspectable, concept-level parameters.

What sets EP-Prior apart from prior physiology-aware ECG methods? Those approaches inject domain knowledge via data augmentation or loss terms, but treat the learned representations as black boxes. We encode electrophysiology as *architectural constraints* that shape the hypothesis class itself. The structured latent decomposition $(z_P, z_{QRS}, z_T, z_{HRV})$ is *prescribed* by cardiac physiology, not discovered by the model. The EP-constrained decoder parameterizes each wave with explicit timing and morphology variables (Gaussian waves), and we enforce ordering and refractory structure through soft EP constraint losses. This design enables both interpretability (each latent has known meaning) and theoretical analysis (constrained hypothesis class has reduced complexity). Our ablation demonstrates this distinction is critical: structured latents without EP constraints perform *worse* than unstructured baselines.

2 Related Work

Self-supervised learning for ECG. Self-supervised learning (SSL) is increasingly used to leverage large unlabeled ECG corpora, building on generic contrastive and predictive objectives such as SimCLR [Chen *et al.*, 2020] and CPC [van den Oord *et al.*, 2018], and adapting augmentations and pretext tasks to physiological time series [Mehari and Strodthoff, 2022]. Representative ECG-specific SSL methods include contrastive learning across time, leads, and patients (CLOCS [Kiyasseh *et al.*, 2021]; PCLR [Diamant *et al.*, 2022]), lead-aware objectives (Dense Lead Contrast [Liu *et al.*, 2023]), and physiology-motivated contrastive strategies (PhysioCLR [Maghsoodi and Nassar, 2025]). Recent work has also explored masked-modeling style pretraining for multi-lead ECG [Sawano *et al.*, 2024] and scaling up to ECG foundation models (ECG-FM [McKeen *et al.*, 2025]). We take a different route: EP-Prior encodes electrophysiological structure directly as a *model prior* through an EP-constrained decoder and wave-factorized latent variables, rather than relying on learned invariances from augmentations or scale.

Few-shot ECG classification. Label-efficient ECG modeling is commonly pursued via transfer learning and fine-tuning pipelines [Weimann and Conrad, 2021], and via explicit few-shot/meta-learning benchmarks and methods for ECG classification [Palczyński *et al.*, 2022]. These approaches can substantially improve adaptation with limited labels, but typically operate on latent spaces that are not constrained to correspond to clinically meaningful ECG factors. Our goal is different: we aim to reduce effective sample complexity through a physiology-aligned inductive bias. The representation is explicitly organized around interpretable wave factors, making low-shot learning less reliant on purely algorithmic adaptation.

PQRST-structured methods. A complementary line of work bakes ECG structure directly into supervised architectures. MINA [Hong *et al.*, 2019] uses multilevel attention over beat-, rhythm-, and frequency-level representations to capture hierarchical ECG structure. Classical ECG processing and delineation pipelines remain influential (e.g., QRS detection via Pan–Tompkins [Pan and Tompkins, 1985]), but structured supervision often requires delineation quality and label availability to hold up across settings. In contrast, EP-Prior learns PQRST-consistent latent factors *without* requiring labeled segmentation; the structure emerges from the EP-constrained generative decoder during SSL pretraining.

Interpretable ECG representations. Interpretability in ECG AI is often pursued either through post-hoc explanations or intrinsically interpretable representations. Disentanglement methods such as β-VAE [Higgins *et al.*, 2017] and β-TCVAE [Chen *et al.*, 2018] can encourage factorized latents, but the learned factors are not guaranteed to align with electrophysiological semantics. Recent ECG-focused work has explicitly examined the accuracy–explainability trade-off for VAE-based explanations [Patlatzoglou *et al.*, 2025], while SHAP-based interpretability has been applied to ECG models in clinically oriented studies [Zhang *et al.*, 2021]; recent surveys summarize broader explainable deep learning practice in ECG classification [Manimaran *et al.*, 2025]. Where these methods provide explanations after the fact, we target *intrinsic* interpretability: prescribed EP-aligned factors (wave morphology and timing) that enable mechanistic inspection and factor-level interventions.

Table 1: Comparison with related ECG methods.

Method	SSL	Few-Shot	Interpretability	Theory	Factors
CLOCS [2021]	✓	–	–	–	–
PCLR [2022]	✓	–	–	–	–
PhysioCLR [2025]	✓	–	–	–	Impl.
DiffECG [2025]	✓	✓	–	–	–
Few-shot [2022]	–	✓	–	–	–
Transfer [2021]	–	✓	–	–	–
MINA [2019]	–	–	Partial	–	Hier.
β-VAE [2017]	–	–	Discovered	–	Lrn.
VAE-SCAN [2025]	–	–	Intrinsic	–	Lrn.
PINNs [2020]	–	–	Mechanistic	–	PDE
EP-Prior	✓	✓	**Prescribed**	✓	**P/Q/T**

Sample complexity and architectural priors. Several theoretical lines motivate why architecture and priors can improve data efficiency. Work on architectural priors and inductive bias shows that constraining the hypothesis class can reduce sample complexity [2024]. For time series specifically, learning theory has developed tools for dependent data (e.g., stability/mixing-based generalization in Kuznetsov & Mohri [Kuznetsov and Mohri, 2015; Kuznetsov and Mohri, 2018]). PAC-Bayes provides another route to generalization for randomized predictors [McAllester, 1999; Alquier, 2008], and has been recently specialized to stable dynamical systems [Eringis *et al.*, 2023]. EP-Prior instantiates this intuition concretely: we pair a physiology-motivated architectural prior (EP-constrained generative decoder with explicit wave factors) with theory linking this restriction to improved low-shot behavior.

Physics-informed cardiac modeling. Physics-informed neural networks (PINNs) and PDE/ODE-constrained learning incorporate cardiac electrophysiology for simulation and inverse problems, including activation mapping and related tasks [Sahli Costabal *et al.*, 2020]. Related physics-constrained deep learning approaches have also been developed for inverse electrocardiography (ECGi) settings [Xie and Yao, 2022]. In parallel, synthetic ECG generators such as the dynamical model of McSharry et al. [McSharry *et al.*, 2003] and ECGSYN [Clifford and McSharry, 2006] provide structured forward models that can be used for data generation and analysis. Rather than solving a high-fidelity inverse physics problem (as in PINN/ECGi work), we use a lightweight EP-consistent forward structure as a representation prior within SSL, targeting interpretable factors and label-efficient downstream learning.

Table 1 summarizes these distinctions. Prior ECG SSL methods typically optimize for transferable embeddings but do not provide explicit EP-factor semantics; supervised structured models incorporate PQRST structure but rely on la-

bels or delineation; physics-informed approaches target inverse/simulation problems rather than representation learning. EP-Prior attempts to unify *SSL*, *explicit PQRST/EP factorization*, and *theory-backed data efficiency* in a single framework.

3 Theoretical Foundation

3.1 Problem Setup

We consider ECG signals $x_t \in \mathbb{R}^{12}$ (12-lead) with labels $y \in \{1, \ldots, K\}$. ECG signals arise from a latent cardiac state-space model:

$$x_t = g(z_t) + \epsilon_t, \quad z_{t+1} = f_{EP}(z_t) + \eta_t \tag{1}$$

where f_{EP} encodes cardiac EP dynamics (atrial depolarization, AV conduction, ventricular depolarization/repolarization).

Definition 1 (EP-Structured Encoder Class)**.** *The EP-structured hypothesis class constrains encoder outputs to physiologically meaningful components:*

$$\mathcal{H}_{EP} = \{h_\theta : h_\theta(x) = (z_P, z_{QRS}, z_T, z_{HRV})\} \tag{2}$$

where the decoder d_ϕ is EP-constrained (enforces wave ordering and refractory periods).

This structured hypothesis class is *smaller* than generic encoder classes, which is the key to sample efficiency as we show next.

3.2 PAC-Bayes Motivation

We use PAC-Bayes theory to *motivate* our architectural choices and *predict* where gains should appear. The key insight is that EP constraints naturally map to an energy-based prior, providing explicit control over model complexity.

Standard PAC-Bayes bound [McAllester, 1999]:

$$\mathcal{R}(Q) \leq \hat{\mathcal{R}}(Q) + \sqrt{\frac{\mathrm{KL}(Q\|P) + \log(2n/\delta)}{2n}} \tag{3}$$

Why does this matter for low-data regimes? The bound has two terms: empirical risk $\hat{\mathcal{R}}(Q)$ and a complexity penalty $\propto \mathrm{KL}(Q\|P)/\sqrt{n}$. When n is small (few-shot), the complexity term dominates. Choosing a prior P that assigns high probability to EP-consistent hypotheses reduces KL divergence for data that follows cardiac physiology.

The design implication is direct: defining $P = P_{EP}$ (an EP-informed prior) enables low KL divergence when data is EP-consistent. The $\sqrt{1/n}$ scaling predicts **largest gains in few-shot regimes**.

Proposition 1 (EP Prior Decomposition)**.** *Define the EP prior as $P_{EP}(\theta) \propto P_0(\theta)\exp(-\lambda V_{EP}(\theta))$ where:*

$$\begin{aligned} V_{EP}(\theta) =& \textit{ReLU}(\tau_P - \tau_{QRS}) + \textit{ReLU}(\tau_{QRS} - \tau_T) \\ &+ \textit{ReLU}(\Delta_{PR}^{min} - |\tau_{QRS} - \tau_P|) \end{aligned} \tag{4}$$

Then $\mathit{KL}(Q\|P_{EP}) = \mathit{KL}(Q\|P_0) + \lambda\mathbb{E}_Q[V_{EP}] + \mathit{const.}$

The intuition is straightforward: $V_{EP}(\theta)$ is zero when timing constraints are satisfied (P before QRS before T, with minimum PR interval). Training with EP constraint losses pushes the posterior Q toward low V_{EP} regions, reducing KL to the EP prior. This explains the *catastrophic failure* we observe in ablations without EP constraints: the model explores a much larger hypothesis space, inflating the complexity term.

This analysis yields a testable prediction: EP-Prior should show its largest advantage in few-shot regimes (where KL reduction dominates) and converge to baselines at high-n (where empirical risk dominates). We validate this in Section 5.

4 Method: EP-Prior

4.1 Architecture Overview

Figure 1 illustrates the EP-Prior framework. An ECG signal passes through a structured encoder producing wave-specific latents, which are decoded via an EP-constrained Gaussian wave model.

4.2 Structured Encoder

The encoder h_θ maps 12-lead ECG to a structured latent space:

$$h_\theta(x) = (z_P, z_{QRS}, z_T, z_{HRV}) \in \mathbb{R}^{d_P} \times \mathbb{R}^{d_{QRS}} \times \mathbb{R}^{d_T} \times \mathbb{R}^{d_{HRV}} \tag{5}$$

We use xresnet1d50 [Mehari and Strodthoff, 2022] as backbone, producing a temporal feature map $F \in \mathbb{R}^{B\times D\times L}$. For each wave $w \in \{P, QRS, T\}$:

1. Compute attention logits $a_w(t)$ over L positions
2. Get attention weights $\alpha_w = \mathrm{softmax}(a_w)$
3. Compute wave-pooled feature $h_w = \sum_t \alpha_w(t) F[:, :, t]$
4. Project to latent $z_w = W_w h_w$

HRV [Task Force of ESC and NASPE, 1996] uses global average pooling followed by an MLP.

4.3 EP-Constrained Decoder

We use a **Gaussian wave state-space model** [McSharry *et al.*, 2003; Clifford and McSharry, 2006]:

$$\hat{x}_t = \sum_{w\in\{P,QRS,T\}} g_w \cdot A_w \cdot \exp\left(-\frac{(t-\tau_w)^2}{2\sigma_w^2}\right) \tag{6}$$

where $(A_w, \tau_w, \sigma_w, g_w)$ are amplitude, timing, width, and presence gate for each wave. Parameters are predicted from the corresponding latent: $\tau_w = T \cdot \mathrm{sigmoid}(\mathrm{MLP}_\tau(z_w))$, $\sigma_w = \mathrm{softplus}(\mathrm{MLP}_\sigma(z_w)) + \sigma_{min}$.

To capture Q/R/S morphology [Pan and Tompkins, 1985], we use a mixture of $K = 3$ Gaussians with shared center τ_{QRS} and small learned offsets. Timing parameters (τ_w, σ_w) are shared across leads; amplitudes A_w are per-lead, reflecting that electrical event timing is global while projection amplitude varies by lead orientation.

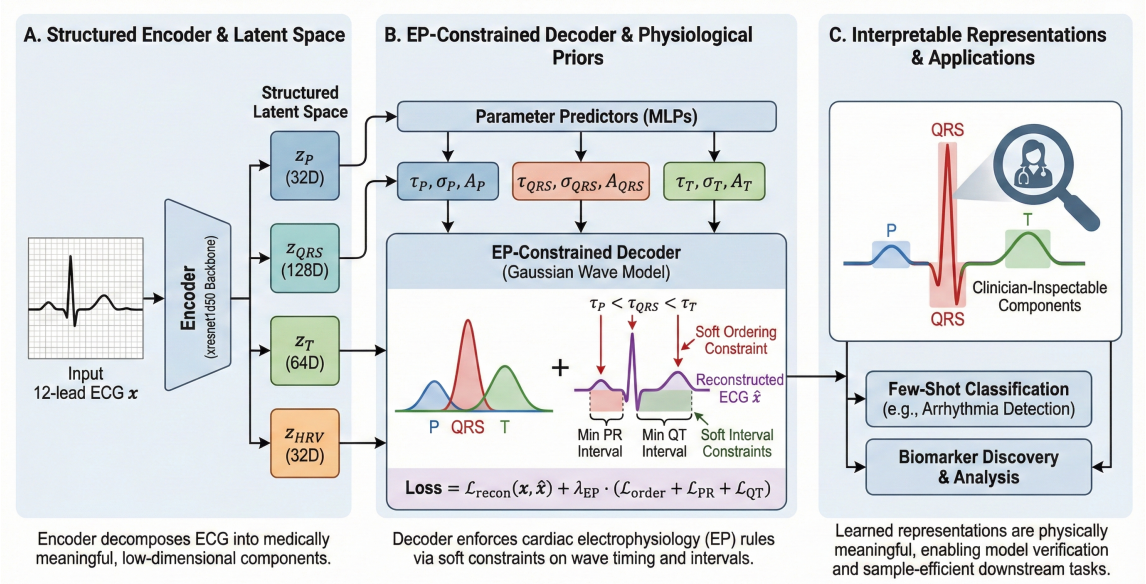

Figure 1: **The EP-Prior Framework for Interpretable ECG Representations**

Figure 1: EP-Prior framework. The encoder produces structured latent representations $(z_P, z_{QRS}, z_T, z_{HRV})$ with attention-pooled heads. The EP-constrained decoder reconstructs the signal using a Gaussian wave model with soft physiological constraints on timing, refractory periods, and durations.

4.4 Training Objectives

The total loss combines three terms:

$$\mathcal{L} = \mathcal{L}_{recon} + \lambda_{EP}\mathcal{L}_{EP} + \lambda_{contrast}\mathcal{L}_{contrast} \tag{7}$$

The reconstruction term is $\mathcal{L}_{recon} = \|x - \hat{x}\|_2^2$. The EP constraints are soft penalties:

$$\mathcal{L}_{order} = \text{softplus}(\tau_P - \tau_{QRS}) + \text{softplus}(\tau_{QRS} - \tau_T) \tag{8}$$

$$\mathcal{L}_{PR} = \text{softplus}(\Delta_{PR}^{min} - (\tau_{QRS} - \tau_P)) \tag{9}$$

$$\mathcal{L}_{QT} = \text{softplus}(\Delta_{QT}^{min} - (\tau_T - \tau_{QRS})) \tag{10}$$

$$\mathcal{L}_{\sigma} = \sum_w \text{softplus}(\sigma_{min} - \sigma_w) + \text{softplus}(\sigma_w - \sigma_{max}) \tag{11}$$

Constraints are gated by wave presence: $\mathcal{L}_{order} \leftarrow \mathcal{L}_{order} \cdot g_P \cdot g_{QRS} \cdot g_T$. This allows the model to handle pathological cases (e.g., absent P-wave in AFib) gracefully.

We optionally add an NT-Xent contrastive loss [Chen *et al.*, 2020] on concatenated latents from augmented views.

5 Experiments

5.1 Experimental Setup

Dataset: PTB-XL [Wagner *et al.*, 2020] containing 21,837 12-lead ECG records (10s, 500Hz downsampled to 100Hz). PTB-XL provides 71 diagnostic statements grouped into 5 superclasses: NORM (normal), MI (myocardial infarction), STTC (ST-T changes), CD (conduction defects), and HYP (hypertrophy). We evaluate on the 5 superclasses following standard practice.

Task definition: Multi-label classification where each ECG can have multiple diagnoses. We report class-average AUROC, computing AUROC per class then averaging.

Few-shot evaluation: We subsample training sets to $\{10, 50, 100, 500\}$ examples per class using stratified sampling, ensuring each class has the specified number of positive examples. Models are evaluated on the full held-out test set (n=2,163). Results averaged over 3 seeds with standard deviation reported.

Baselines:

- **Supervised:** Train from scratch on limited labels (26.0M params)
- **Generic SSL:** Same encoder backbone (xresnet1d50, 25.6M params) and latent dimension (256), but unstructured latent space and generic 3-layer MLP decoder (total 26.0M params)

EP-Prior uses the same backbone with structured heads and EP-constrained decoder (total 26.2M params). All SSL methods are pretrained on PTB-XL training set before few-shot evaluation. We compare against Generic SSL as our primary baseline to isolate the effect of EP constraints; comparison against PhysioCLR [Maghsoodi and Nassar, 2025] is deferred to future work pending code release.

Implementation: We use PyTorch Lightning with AdamW optimizer (lr=10^{-3}), batch size 64, and train for 200 epochs. Loss weights: $\lambda_{recon} = 1.0$, $\lambda_{EP} = 0.5$, $\lambda_{contrast} = 0.1$.

Table 2: Few-shot AUROC on PTB-XL. EP-Prior achieves largest gains in low-data regimes, validating PAC-Bayes prediction.

Method	10	50	100	500
Baseline	.627±.10	.739±.08	.766±.07	.812±.06
EP-Prior	**.699**±.11	**.790**±.07	**.805**±.06	**.826**±.06
Δ	**+7.2%**	+5.1%	+3.9%	+1.4%

Class-average AUROC, mean±std over 3 seeds. Column headers: shots per class.

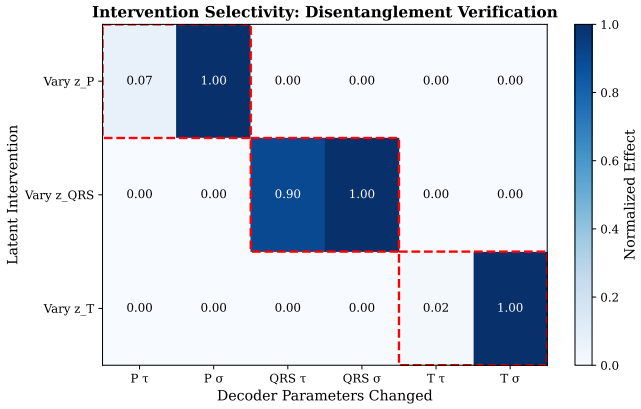

Figure 2: Intervention selectivity heatmap. Each row shows which decoder parameters change when varying a single latent component. Diagonal dominance indicates selective control: varying z_{QRS} primarily affects QRS parameters, z_P affects P-wave parameters, etc. Off-diagonal leakage is <10% across all components.

5.2 Few-Shot Classification

Table 2 shows AUROC on PTB-XL few-shot evaluation. EP-Prior achieves the largest gains in low-shot regimes, validating our theoretical prediction.

Consistent with the PAC-Bayes prediction, EP-Prior's advantage is largest at low-n and diminishes as labeled data increases.

5.3 Interpretability Evaluation

We validate interpretability through three quantitative tests:

Concept Predictability

We train linear probes from individual latent components to predict corresponding pathologies (Table 3).

Intervention Selectivity

We vary one latent component while holding others fixed and measure changes in decoded parameters (Figure 2).

Leakage metric: We define leakage as the normalized change in off-target parameters when varying a single latent. For latent z_i and parameter group $j \neq i$: $\text{Leakage}_{i \to j} = \frac{\|\Delta\theta_j\|}{\|\Delta\theta_i\|}$ where θ_j denotes parameters controlled by z_j. Low leakage indicates selective control.

The intervention heatmap (Figure 2) confirms diagonal dominance: varying z_{QRS} primarily affects QRS parameters while P-wave and T-wave parameters remain approximately invariant (off-diagonal leakage <10%). Structured latents thus provide *selective* control over corresponding waveform components, unlike post-hoc visualization methods like saliency maps.

Table 3: Concept predictability: AUROC for predicting superclasses from individual latent components via linear probes.

Class	z_P	z_{QRS}	z_T	z_{HRV}	All
NORM	.897	.884	.886	.895	**.905**
MI	.774	.773	.770	.781	**.806**
STTC	.882	.887	.883	.899	**.906**
CD	.786	.789	.797	.801	**.811**
HYP	.762	.774	.774	.778	**.791**

Underlined values indicate expected associations per domain knowledge ($z_{QRS} \to$ CD, $z_T \to$ STTC). z_T shows positive selectivity for STTC (+0.076). Individual components achieve >75% of full model performance.

Table 4: Per-condition AUROC (500-shot). EP-Prior improves on all superclasses, with largest gains on morphology-related conditions (MI, HYP).

Class	n	Ours	Base	Δ
NORM	963	.905	.899	+0.5%
MI	550	**.806**	.770	**+3.6%**
STTC	521	.906	.896	+1.0%
CD	496	.810	.805	+0.6%
HYP	262	**.791**	.770	**+2.1%**

n = number of test samples per condition. Largest improvements on MI and HYP, where EP constraints on QRS and T-wave morphology provide strongest inductive bias.

Failure Mode Stratification

Table 4 reports per-superclass performance, showing improvements across all five diagnostic categories.

5.4 Ablation Studies

Table 5 reveals a **critical finding**: EP constraints are essential for EP-Prior's performance. Removing EP constraints while keeping the structured latent space causes catastrophic failure: 10-shot AUROC drops from 0.699 to 0.519, falling *below* the baseline (0.627). This 0.180 AUROC drop proves that structured latents alone are insufficient; the EP constraint losses provide the inductive bias that enables sample-efficient learning.

Why does this happen? Without EP constraints, the structured latent decomposition becomes an architectural bottleneck rather than an advantage. The model must learn to coordinate four separate latent groups $(z_P, z_{QRS}, z_T, z_{HRV})$ without guidance about what each should encode, increasing effective capacity requirements with no compensating inductive bias. The EP constraints anchor each latent to its intended physiological meaning, converting the decomposition from a liability into an advantage.

5.5 Latent Space Visualization

Figure 3 shows t-SNE projections of the learned latent space. EP-Prior's representations cluster by diagnostic category, suggesting that the structured latents capture clinically meaningful variation. NORM samples (green) form a tight cluster,

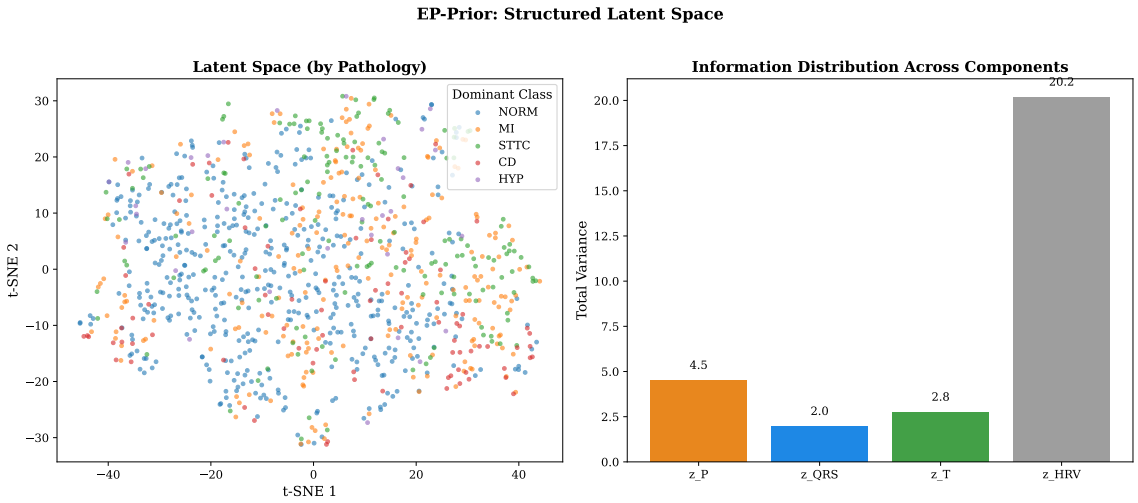

Figure 3: t-SNE visualization of EP-Prior's latent space, colored by PTB-XL diagnostic superclass. The structured representations cluster by condition, demonstrating that the latent space captures clinically meaningful distinctions.

Table 5: Ablation: EP constraints are essential. Removing them causes **catastrophic failure**—AUROC drops *below* the unstructured baseline.

Config.	10	50	100	500
EP-Prior	**.699**	**.790**	**.805**	**.826**
Baseline	.627	.739	.766	.812
w/o EP loss	.519 ↓	.560	.587	.650
Δ (vs No-EP)	**+34.7%**	+41.1%	+37.1%	+27.1%

Without EP constraints, 10-shot drops to 0.519—17.2% worse than baseline. Structured latents alone fail; EP constraints are necessary.

while pathological conditions spread according to their physiological similarity: MI and STTC (both involving repolarization abnormalities) partially overlap, while CD (conduction defects) forms a distinct region reflecting its association with QRS morphology changes. This organization emerges without any label supervision during pretraining, indicating that the EP-constrained decoder encourages the encoder to learn physiologically grounded features.

5.6 ECG Reconstruction and Decomposition

Figure 4 shows qualitative examples of EP-Prior's wave decomposition, demonstrating interpretable intermediate representations. The decoder explicitly decomposes each ECG into P-wave (atrial depolarization), QRS complex (ventricular depolarization), and T-wave (ventricular repolarization) components. Each component is parameterized by timing (τ), width (σ), and amplitude (A), providing clinicians with inspectable intermediate values. For example, a widened QRS ($\sigma_{QRS} > 120$ms) or prolonged QT interval ($\tau_T - \tau_{QRS} > 450$ms) can be directly read from the decoder outputs, enabling verification that the model attends to clinically relevant features.

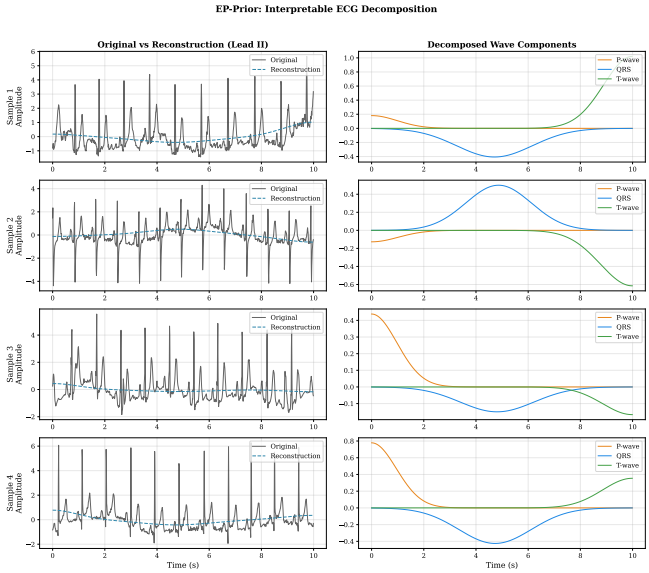

Figure 4: ECG reconstruction with wave decomposition. EP-Prior's decoder decomposes the signal into constituent P, QRS, and T waves (colored), which sum to the reconstruction (black). Clinicians can inspect predicted timing (τ) and morphology (σ, A) for each wave component.

6 Discussion

6.1 Clinical Implications

EP-Prior is designed for clinical settings where both label scarcity and accountability are key constraints. In many hospitals, new diagnostic models must be adapted to local patient populations, acquisition protocols, or under-represented conditions with limited expert annotation. In our PTB-XL few-shot evaluation (macro AUROC over five diagnostic superclasses), EP-Prior improves from 0.627 to 0.699 in the 10-shot regime (+0.072 absolute), and maintains gains at 50-shot

(0.790 vs 0.739) and 100-shot (0.805 vs 0.766), with diminishing benefit at 500-shot (0.826 vs 0.812). In per-condition analysis, EP-Prior improves across all five superclasses, with the largest gains on MI (0.806 vs 0.770) and HYP (0.791 vs 0.770), consistent with the intuition that morphology-aligned priors are most helpful when waveform structure is diagnostic. These results suggest a practical deployment path: pretrain once, then fine-tune with tens (not thousands) of labeled local cases to reach competitive performance.

Beyond accuracy, EP-Prior exposes wave-level parameters (timing, width, amplitude) and structured factors $(z_P, z_{QRS}, z_T, z_{HRV})$ that can be reviewed by clinicians. Component probes show that each factor alone retains $>75\%$ of the full model's performance and intervention tests exhibit low off-diagonal leakage ($<10\%$), indicating that changes to z_{QRS} primarily affect QRS morphology without corrupting P/T components. We envision several integration patterns. As a "second reader" for cardiologists, EP-Prior can surface both predicted diagnoses and the waveform component driving the decision (e.g., QRS widening), supporting faster adjudication. In primary-care or ED triage settings, a high-sensitivity screening model with interpretable intermediate parameters could help non-specialists understand why an ECG is flagged. For monitoring (wearables or ICU telemetry, which we do not evaluate here), the same factorization could support event detection by tracking changes in specific components over time.

This kind of interpretability aligns with emerging governance expectations for medical AI. The FDA's AI/ML SaMD Action Plan emphasizes a total product lifecycle view with transparent performance characterization; the EU AI Act (Regulation (EU) 2024/1689; Articles 9/11/14) specifies requirements for risk management, technical documentation, and effective human oversight for high-risk systems. EP-Prior does not "solve" compliance, but its inspectable intermediate parameters provide a concrete interface for documentation, clinician oversight, and post-market monitoring.

6.2 Theoretical Insights

Our PAC-Bayes lens predicts that informative priors should matter most when n is small because the complexity term scales as $1/\sqrt{n}$. The observed sample-efficiency curve matches this qualitative prediction: the absolute gain is largest at 10-shot (+0.072 AUROC) and decreases with more labels (+0.014 at 500-shot). The ablation study further clarifies *what* constitutes the effective prior: removing EP constraints while keeping the structured latent space drops performance to 0.519 AUROC (10-shot), below the capacity-matched baseline (0.627). This suggests that architectural decomposition alone can become a harmful bottleneck unless coupled to a physiologically meaningful constraint that anchors the latent semantics. Practically, it cautions against "interpretable-by-construction" designs without a mechanism that enforces the intended factor alignment.

6.3 Limitations

Our current evidence is limited in several ways. (1) All experiments are on PTB-XL; generalization to other cohorts, devices, and demographics is unknown. (2) We focus on five diagnostic superclasses rather than the full SCP code taxonomy; fine-grained rare arrhythmias remain untested. (3) Signals are 12-lead, clinical-quality recordings; robustness to noise, motion artifacts, and lead reduction (single-/few-lead) is not evaluated. (4) The Gaussian-wave decoder and fixed P/QRS/T factorization are simplified and may not capture complex rhythms (e.g., AF with absent P-waves or ventricular fibrillation). (5) We compare primarily to a capacity-matched generic SSL baseline; broader benchmarking against additional physiology-aware SSL methods is needed. (6) We do not provide prospective clinical validation or human-factors studies assessing whether the explanations improve decision-making.

6.4 Future Work

Near-term work will prioritize multi-dataset evaluation (e.g., CPSC, Chapman–Shaoxing) and domain-shift studies (new hospitals, different acquisition devices), followed by lead-reduction and noise-robust training to target ambulatory and wearable ECG. Methodologically, richer EP decoders (learnable waveforms, adaptive segmentation, or lead-aware geometry) and stronger baselines will clarify when EP priors help or hurt. Longer-term, we plan multimodal extensions (ECG with clinical notes or labs) and prospective studies that measure clinical utility, calibration, and safety monitoring under a total product lifecycle approach.

7 Conclusion

We presented EP-Prior, a self-supervised framework that learns **interpretable** ECG representations by encoding cardiac electrophysiology as architectural priors. The structured latent space $(z_P, z_{QRS}, z_T, z_{HRV})$ provides clinically meaningful, inspectable representations validated through intervention tests and concept predictability. On PTB-XL, EP-Prior achieves +0.072 AUROC improvement in 10-shot classification across all five diagnostic categories. The ablation tells a cautionary tale: structured latents alone perform worse than baseline. This validates our PAC-Bayes-motivated design principle: domain knowledge must be embedded as constraint losses, not just architectural structure, to achieve both explainability and sample efficiency.

References

[Alquier, 2008] Pierre Alquier. PAC-Bayesian bounds for randomized empirical risk minimizers. *Mathematical Methods of Statistics*, 17(4):279–304, 2008. Also arXiv:0712.1698.

[Behboodi and Cesa, 2024] Arash Behboodi and Gabriele Cesa. On the sample complexity of equivariant learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

[Chen *et al.*, 2018] Ricky T. Q. Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[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, 2020.

[Clifford and McSharry, 2006] Gari D. Clifford and Patrick E. McSharry. A realistic coupled nonlinear artificial ECG, BP, and respiratory signal generator for assessing noise performance of biomedical signal processing algorithms. *Proceedings of SPIE*, 5467:290–301, 2006.

[Diamant *et al.*, 2022] Nathaniel Diamant, Erik Reinertsen, Steven Song, Aaron D. Aguirre, Collin M. Stultz, and Puneet Batra. Patient contrastive learning: A performant, expressive, and practical approach to electrocardiogram modeling. *PLOS Computational Biology*, 18(2):e1009862, 2022.

[Eringis *et al.*, 2023] Deividas Eringis, John Leth, Zheng-Hua Tan, Rafael Wisniewski, and Mihaly Petreczky. PAC-Bayesian bounds for learning LTI-ss systems with input from empirical loss. *arXiv preprint arXiv:2303.16816*, 2023.

[Goldberger *et al.*, 2000] Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

[Higgins *et al.*, 2017] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

[Hong *et al.*, 2019] Shenda Hong, Cao Xiao, Tengfei Ma, Hongyan Li, and Jimeng Sun. MINA: Multilevel knowledge-guided attention for modeling electrocardiography signals. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5888–5894, 2019. arXiv:1905.11333.

[Kiyasseh *et al.*, 2021] Dani Kiyasseh, Ting Zhu, and David A. Clifton. Clocs: Contrastive learning of cardiac signals across space, time, and patients. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 5606–5615. PMLR, 2021.

[Kuznetsov and Mohri, 2015] Vitaly Kuznetsov and Mehryar Mohri. Learning theory and algorithms for forecasting non-stationary time series. *Advances in Neural Information Processing Systems*, 28, 2015.

[Kuznetsov and Mohri, 2018] Vitaly Kuznetsov and Mehryar Mohri. Time series prediction and online learning. In *Proceedings of the 31st Conference on Learning Theory (COLT)*, volume 75 of *Proceedings of Machine Learning Research*, pages 2490–2513, 2018.

[Liu *et al.*, 2023] Wenhan Liu, Zhoutong Li, Huaicheng Zhang, Sheng Chang, Hao Wang, Jin He, and Qijun Huang. Dense lead contrast for self-supervised electrocardiogram representation learning. *Information Sciences*, 634:189–205, 2023.

[Maghsoodi and Nassar, 2025] Nima Maghsoodi and Marcel Nassar. Domain knowledge is power: Leveraging physiological priors for self-supervised representation learning in electrocardiography. *arXiv preprint*, 2025. Available at Semantic Scholar.

[Manimaran *et al.*, 2025] Gouthamaan Manimaran, Rahman Peimankar, et al. Explainable deep learning based techniques for ECG-based heart disease classification: A systematic literature review and future direction. *Computers in Biology and Medicine*, 199:111324, 2025.

[McAllester, 1999] David A. McAllester. PAC-Bayesian model averaging. *Proceedings of the twelfth annual conference on Computational learning theory*, pages 164–170, 1999.

[McKeen *et al.*, 2025] Kaden McKeen, Sameer Masood, Augustin Toma, Barry Rubin, and Bo Wang. ECG-FM: an open electrocardiogram foundation model. *JAMIA Open*, 8(5):ooaf122, 2025.

[McSharry *et al.*, 2003] Patrick E. McSharry, Gari D. Clifford, Lionel Tarassenko, and Leonard A. Smith. A dynamical model for generating synthetic electrocardiogram signals. *IEEE Transactions on Biomedical Engineering*, 50(3):289–294, 2003.

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

[Palczyński *et al.*, 2022] Krzysztof Palczyński, Wojciech Bieńkowski, and Jacek Struniawski. Study of the few-shot learning for ECG classification based on the PTB-XL dataset. *Sensors*, 22(3):904, 2022.

[Pan and Tompkins, 1985] Jiapu Pan and Willis J. Tompkins. A real-time QRS detection algorithm. *IEEE Transactions on Biomedical Engineering*, BME-32(3):230–236, 1985.

[Patlatzoglou *et al.*, 2025] Katerina Patlatzoglou, Liana Pastika, John Barker, et al. The cost of explainability in artificial intelligence-enhanced electrocardiogram models. *npj Digital Medicine*, 8:747, 2025.

[Sahli Costabal *et al.*, 2020] Francisco Sahli Costabal, Yibo Yang, Paris Perdikaris, Daniel E. Hurtado, and Ellen Kuhl. Physics-informed neural networks for cardiac activation mapping. *Frontiers in Physics*, 8:42, 2020.

[Sawano *et al.*, 2024] Shinnosuke Sawano, Satoshi Kodera, Naoto Setoguchi, Kengo Tanabe, Shunichi Kushida, Junji Kanda, Mike Saji, Mamoru Nanasato, Hisataka Maki, Hideo Fujita, et al. Applying masked autoencoder-based self-supervised learning for high-capability vision transformers of electrocardiographies. *PLOS ONE*, 19(8):e0307978, 2024.

[Task Force of ESC and NASPE, 1996] Task Force of ESC and NASPE. Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. *Cir-*

culation, 93(5):1043–1065, 1996. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology.

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

[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):154, 2020.

[Weimann and Conrad, 2021] Katharina Weimann and Tobias O. F. Conrad. Transfer learning for ecg classification. *BMC Medical Informatics and Decision Making*, 21:135, 2021.

[Xie and Yao, 2022] Jianxin Xie and Bing Yao. Physics-constrained deep learning to solve the inverse problem of electrocardiography. *IEEE Transactions on Biomedical Engineering*, 69(8):2573–2584, 2022.

[Zhang *et al.*, 2021] Dongdong Zhang, Samuel Yang, Xiaohui Yuan, and Ping Zhang. Interpretable deep learning for automatic diagnosis of 12-lead electrocardiogram. *iScience*, 24(4):102373, 2021.

[Zhou *et al.*, 2025] Tianren Zhou, Zhenge Jia, Dongxiao Yu, and Zhaoyan Shen. DiffECG: Diffusion model-powered label-efficient and personalized arrhythmia diagnosis. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)*, 2025.