

# 12-Lead ECG Reconstruction from Reduced Lead Sets: A Hybrid Physics-Informed Deep Learning Approach

Damilola Olaiya

damilolaolaiya@cmail.carleton.ca  
Carleton University  
Ottawa, Ontario, Canada

Mithun Manivannan

mithun.manivannan@cmail.carleton.ca  
Carleton University  
Ottawa, Ontario, Canada

## ABSTRACT

Cardiovascular disease (CVD) remains the world’s leading cause of death. Despite this, the gold-standard 12-lead electrocardiogram (ECG) is inaccessible in many settings due to equipment complexity and personnel requirements. We present a hybrid, physics-informed, deep learning approach to reconstruct the full 12-lead ECG from only 3 measured leads (I, II, V4). Our method exploits deterministic physiological relationships—Einthoven’s law and Goldberger’s equations—for computationally efficient, zero-latency reconstruction of 4 limb leads (III, aVR, aVL, aVF) with no learned parameters, while a 1D U-Net neural network addresses the core machine learning challenge: reconstructing the 5 precordial leads (V1, V2, V3, V5, V6). Using the PTB-XL dataset with strict patient-wise splits to prevent data leakage, our learned chest leads achieve  $r = 0.846$  mean correlation with strong per-lead performance ranging from  $r = 0.818$  (V1) to  $r = 0.891$  (V5). Critically, we demonstrate that a shared decoder architecture (17.1M parameters) outperforms lead-specific decoders (40.8M parameters) with a large effect size (Cohen’s  $d = 0.92$ , 95% CI [0.006, 0.072]), revealing that input information content—not model capacity—is the fundamental bottleneck. Our analysis of ground-truth inter-lead correlations explains the performance hierarchy and suggests that input lead selection is more critical than architectural complexity for future improvements. All code, trained models and evaluation scripts are publicly available at [https://github.com/whiteblaze143/DATA\\_5000](https://github.com/whiteblaze143/DATA_5000).

## CCS CONCEPTS

- Computing methodologies → Machine learning; Neural networks;
- Human-centered computing → Ubiquitous and mobile computing.

## KEYWORDS

ECG reconstruction, deep learning, U-Net, neural networks, cardiovascular disease, reduced lead ECG, wearable health monitoring

## 1 INTRODUCTION

Cardiovascular diseases (CVDs) are the leading cause of mortality worldwide, responsible for an estimated 17.9 million deaths annually. CVDs are particularly dangerous due to their cumulative and often silent nature—conditions like hypertension, atherosclerosis and early-stage heart failure can progress for years without noticeable symptoms until a catastrophic event occurs.

The electrocardiogram (ECG) remains the gold standard non-invasive diagnostic tool for cardiac assessment, capturing the heart’s electrical activity through multiple perspectives to enable detection of arrhythmias, myocardial infarction, conduction abnormalities,

and ventricular hypertrophy [? ]. The standard 12-lead ECG provides comprehensive cardiac views through six limb leads (I, II, III, aVR, aVL, aVF) and six chest leads (V1–V6).

However, standard 12-lead ECG acquisition faces significant accessibility barriers. The procedure requires 10 electrodes with precise anatomical placement and skilled technicians for proper acquisition [? ]. This makes deployment difficult in ambulances, homes, or remote areas [? ], while consumer wearables such as Apple Watch and Fitbit record only 1–2 leads [? ? ].

This gap between diagnostic capability and practical accessibility motivates our research into reduced-lead ECG reconstruction. We propose a hybrid, deterministic + deep learning approach that reconstructs the full 12-lead ECG from only 3 measured leads, combining deterministic physiological relationships with learned neural network mappings.

## 1.1 Contributions

We make four contributions: (1) a hybrid architecture combining deterministic algorithms with deep learning for chest leads; (2) rigorous evaluation with patient-wise splits preventing data leakage, following multi-level frameworks [? ]; (3) analysis of diagnostic utility preservation; and (4) a complete reproducible codebase.

## 2 BACKGROUND

### 2.1 ECG Lead System

A *lead* in an ECG is not the physical wire or electrode, but rather a specific view of the heart’s electrical activity recorded as a voltage difference between electrode positions. Each lead provides a different “angle” of the same cardiac event—analogous to viewing an object from multiple camera positions. Figure ?? shows a typical 12-lead ECG recording.

**2.1.1 Limb Leads (Frontal Plane).** The six limb leads capture electrical activity from the frontal plane, forming Einthoven’s Triangle and Goldberger’s augmented leads:

**Bipolar Leads (I, II, III):**

$$\text{Lead I} = V_{LA} - V_{RA} \quad (1)$$

$$\text{Lead II} = V_{LL} - V_{RA} \quad (2)$$

$$\text{Lead III} = V_{LL} - V_{LA} \quad (3)$$

**Einthoven’s Law:** These leads satisfy the relationship:

$$\text{Lead III} = \text{Lead II} - \text{Lead I} \quad (4)$$

**Augmented Leads (aVR, aVL, aVF):** Goldberger's equations allow exact computation:

$$aVR = -\frac{\text{Lead I} + \text{Lead II}}{2} \quad (5)$$

$$aVL = \text{Lead I} - \frac{\text{Lead II}}{2} \quad (6)$$

$$aVF = \text{Lead II} - \frac{\text{Lead I}}{2} \quad (7)$$

These relationships are **deterministic**—given Leads I and II, all other limb leads can be computed with zero error [?].

**2.1.2 Chest Leads (Horizontal Plane).** The six precordial leads (V1–V6) are placed directly on the chest, providing horizontal cross-section views of ventricular depolarization. Unlike limb leads, **chest leads cannot be derived mathematically**—they must be measured directly or reconstructed via machine learning. Table 1 summarizes the anatomical positions and cardiac views for each precordial lead.

**Table 1: Precordial Lead Positions and Anatomical Views**

Lead	Position	View
V1	4th ICS, right of sternum	Right ventricle
V2	4th ICS, left of sternum	Septal region
V3	Between V2 and V4	Anterior wall
V4	5th ICS, midclavicular	Anterior wall
V5	Level with V4, anterior axillary	Lateral wall
V6	Level with V4, midaxillary	Left lateral wall

## 2.2 Clinical Significance of Missing Leads

Clinical phenomena with regional expression manifest predominantly in specific precordial leads [? ?]. Anterior myocardial infarction presents as ST-elevation in V1–V4, bundle branch blocks show characteristic patterns in V1 and V6, and left ventricular hypertrophy manifests as voltage amplitude patterns across chest leads [?]. Consequently, limb-only recordings are insufficient for many diagnostic decisions, motivating accurate chest lead reconstruction.

## 3 RELATED WORK

The field of ECG reconstruction has evolved significantly over 46 years (1979–2025), progressing from classical linear transforms to sophisticated deep learning architectures [?].

### 3.1 Classical Approaches (1979–2010)

Early work utilized Frank lead systems [?], Dower transforms [?], and EASI configurations [?] with fixed linear coefficient matrices derived from anatomical models. These achieved correlations of 0.92–0.99 for normal sinus rhythm but degraded for pathological patterns. Advantages included interpretability and negligible computation (<1 ms), while limitations included poor personalization for non-standard thoracic geometry [?].

### 3.2 Adaptive Signal Processing (2006–2018)

Wavelets [? ?], adaptive filters [?], and compressive sensing [?] introduced patient-specific tuning. RMSE improved from  $\sim 15 \mu\text{V}$  (classical) to  $\sim 11 \mu\text{V}$ . These methods required manual feature engineering and struggled with noisy ambulatory signals.

### 3.3 Deep Learning for ECG Reconstruction

**3.3.1 Convolutional and Recurrent Approaches.** Matyschik et al. [?] demonstrated feasibility of ECG reconstruction from minimal lead sets using CNNs. Fu et al. [?] achieved wearable 12-lead ECG acquisition using deep learning from Frank or EASI leads with clinical validation, demonstrating practical deployment potential.

**3.3.2 Foundation Models (2024–2025).** Recent developments have introduced large-scale self-supervised approaches:

**ECG-FM** [?] trained on 1.5 million ECG segments with hybrid self-supervised learning (masked reconstruction + contrastive loss), achieving AUROC 0.996 for atrial fibrillation and 0.929 for reduced LVEF. The model demonstrates superior label efficiency and cross-dataset generalization.

**OpenECG** [?] provided the first large-scale multi-center benchmark (1.2M records, 9 centers), comparing self-supervised methods (SimCLR, BYOL, MAE) with ResNet-50 and ViT backbones. Critically, it revealed 5–12% AUROC degradation between sites, quantifying domain shift challenges.

**3.3.3 Generative Models. Physics-Informed Diffusion:** SE-Diff [?] integrates ODE-based cardiac simulators with diffusion processes, achieving MAE 0.0923 and NRMSE 0.0714 while enforcing physiological constraints on QRS morphology.

**Hierarchical VAEs:** cNVAE-ECG [?] achieves up to 2% AUROC improvement over GAN baselines through 32 hierarchical latent groups enabling multi-scale rhythm and morphology modeling.

**State-Space Models:** SSSD-ECG [?] combines S4 models with diffusion for capturing long-term dependencies (>10s) with  $O(n \log n)$  complexity.

### 3.4 Evaluation Methodology Evolution

ECGGenEval [?] introduced comprehensive multi-level assessment achieving MSE 0.0317, evaluating at signal, feature, and diagnostic levels. DiffuSETS [?] proposed 3-tier evaluation for text-conditioned generation including CLIP score for text-ECG alignment.

Critically, Presacan et al. [?] conducted rigorous Bland-Altman analysis on 9,514 PTB-XL subjects, identifying potential regression-to-mean effects ( $R^2 = 0.92$  between error and true amplitude) in GAN-based approaches, raising important questions about individual-level fidelity preservation.

### 3.5 Research Gap

A recent systematic review [?] analyzing reconstruction algorithms found that 3-lead configurations capture 99.12% of ECG information content, achieving correlations  $r > 0.90$ . However, no universal algorithm exists, and patient-specific vs. generic coefficient trade-offs remain unresolved.

Our work addresses these gaps by integrating physics guarantees with deep learning flexibility, implementing patient-wise

splits to prevent data leakage [?], evaluating at signal, feature, and diagnostic levels [?], and systematically exploring input lead configurations.

## 4 METHODOLOGY

### 4.1 Problem Formulation

We formulate ECG reconstruction as a constrained sequence-to-sequence regression problem. Given 3 measured leads (I, II, and one precordial lead V4), we derive 4 limb leads (III, aVR, aVL, aVF) via deterministically using Equations 4–7, and reconstruct 5 chest leads (V1, V2, V3, V5, V6) via deep learning. The goal is to output the complete 12-lead ECG while preserving both waveform morphology and diagnostic utility.

### 4.2 Hybrid Architecture

Our approach combines two complementary components as illustrated in Figure ??.

**4.2.1 Deterministic.** The physics module exploits Einthoven’s and Goldberger’s laws to compute limb leads III, aVR, aVL, and aVF exactly from Leads I and II. This guarantees zero reconstruction error for derived limb leads with no learned parameters and physiologically guaranteed correctness.

**4.2.2 Deep Learning Component (1D U-Net).** For chest lead reconstruction, we employ a 1D U-Net architecture optimized for temporal signal processing [?] (Figure ??).

The U-Net encoder-decoder structure with skip connections is particularly well-suited for ECG signals because it captures multi-scale temporal features (P-wave  $\sim$ 80ms, QRS  $\sim$ 100ms, T-wave  $\sim$ 200ms) while preserving fine morphological detail through skip connections. The encoder path uses Conv1D blocks with increasing channels ( $64 \rightarrow 128 \rightarrow 256 \rightarrow 512$ ), each consisting of Conv1D  $\rightarrow$  BatchNorm  $\rightarrow$  ReLU sequences with MaxPool1D downsampling. The bottleneck captures multi-beat context at maximum channel count (512 or 1024). The decoder path upsamples via ConvTranspose1D with skip connections from the encoder, decreasing channels symmetrically. See Table 2.

**Table 2: Model Specifications**

Parameter	Value
Input Channels	3 (I, II, V4)
Output Channels	5 (V1, V2, V3, V5, V6)
Base Features	64
Depth (Levels)	4
Kernel Size	3
Dropout Rate	0.2

**4.2.3 Architectural Variants.** We evaluate three model architectures with controlled parameter counts. See Table 3.

The hybrid variant maintains the full shared encoder-decoder backbone but adds lightweight per-lead specialization heads. Each head consists of two 1D convolutional layers: Conv1D ( $1 \rightarrow 32$  channels), ReLU activation, and Conv1D ( $32 \rightarrow 1$  channels). This

**Table 3: Model Variant Specifications**

Variant	Architecture	Params	Overhead
Baseline	Shared enc + dec	17.1M	—
Hybrid	Trunk + 5 heads	17.1M	+0.06%
Lead-Spec	Enc + 5 decoders	40.8M	+138%

design adds only 10,240 parameters total across all 5 heads, representing minimal overhead while allowing lead-specific refinement.

### 4.3 Training Configuration

**4.3.1 Frozen Hyperparameters.** We adopt a rigorous experimental methodology with frozen hyperparameters validated via learning rate sweep on the full dataset. This ensures fair comparison across architectural variants. See Table 4.

**Table 4: Frozen Hyperparameters (Validated via LR Sweep)**

Hyperparameter	Value
Optimizer	AdamW
Learning Rate	$3 \times 10^{-4}$ (validated)
Batch Size	64
Epochs	150 (max)
Early Stopping	20 epochs patience
Loss Function	MSE (+ physics term for variant)
Weight Decay	$1 \times 10^{-4}$
Random Seed	42

**Learning Rate Validation:** We conducted a sweep over  $\{1 \times 10^{-5}, 3 \times 10^{-5}, 1 \times 10^{-4}, 3 \times 10^{-4}, 1 \times 10^{-3}\}$  on the full PTB-XL dataset (14,363 training samples). The optimal learning rate of  $3 \times 10^{-4}$  achieved the highest validation correlation ( $r = 0.927$ ) and was fixed for all subsequent experiments.

**4.3.2 Model Variants.** We systematically evaluate three architectural variants to understand the impact of decoder specialization and domain-knowledge informed learning. The Baseline (UNet1D) uses a shared encoder and decoder architecture with 17,122,373 parameters. The Hybrid (UNet1DHybrid) adds 5 lightweight per-lead heads to the shared trunk (17,132,613 parameters, +0.06% overhead). The Physics-Aware variant uses the baseline architecture with a physics-informed loss function that penalizes Einthoven’s and Goldberger’s law violations.

**4.3.3 Integrating Domain Knowledge into Loss Function.** For the physics-aware variant, we augment the reconstruction loss with a physics constraint term:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}} + \lambda \mathcal{L}_{\text{physics}} \quad (8)$$

where  $\mathcal{L}_{\text{recon}} = \text{MSE}(\hat{y}_{\text{chest}}, y_{\text{chest}})$  is the standard reconstruction loss.

The domain-knowledge loss enforces Einthoven’s and Goldberger’s laws in the denormalized signal space:

$$\begin{aligned}\mathcal{L}_{\text{physics}} = & \|\text{III}' - (\text{II}' - \text{I}')\|_2^2 \\ & + \|\text{aVR}' + \frac{\text{I}' + \text{II}'}{2}\|_2^2 \\ & + \|\text{aVL}' - (\text{I}' - \frac{\text{II}'}{2})\|_2^2 \\ & + \|\text{aVF}' - (\text{II}' - \frac{\text{I}'}{2})\|_2^2\end{aligned}\quad (9)$$

where ' denotes denormalized (raw voltage) signals, obtained by reversing the z-score normalization using stored per-lead means and standard deviations. We set  $\lambda = 0.1$  as the default physics weight.

**4.3.4 Statistical Comparison Framework.** To rigorously compare model variants, we employ paired *t*-tests for mean differences in per-lead correlations, Wilcoxon signed-rank tests as non-parametric alternatives, Cohen's *d* effect size ( $d = (\bar{x}_A - \bar{x}_B)/s_{\text{pooled}}$ ) to quantify magnitude independent of sample size, bootstrap 95% confidence intervals with 10,000 resamples, and Bonferroni correction for multiple comparisons. Following standard conventions,  $|d| < 0.2$  indicates negligible effect,  $0.2 \leq |d| < 0.5$  small,  $0.5 \leq |d| < 0.8$  medium, and  $|d| \geq 0.8$  large. We require  $p < 0.05$  after correction, 95% CI excluding zero, and medium effect size for claiming meaningful difference.

We also implemented additional safeguards for robust pairwise testing: (1) a minimum paired-sample threshold ("min\_n\_valid", default = 30) to avoid low-powered comparisons; (2) a paired permutation test (sign-flip on paired differences) as a fallback when parametric (*t*-test) or exact Wilcoxon methods fail due to ties or zero variance; and (3) Benjamini-Hochberg (FDR) correction across leads for each pairwise variant comparison (baseline vs other) to control false discovery rates across multiple tests. These steps help ensure reported p-values are conservative and reproducible.

## 5 DATASET

### 5.1 PTB-XL Database

We use the PTB-XL dataset [?], a large publicly available electrocardiography dataset from PhysioNet. See Table 5.

**Table 5: PTB-XL Dataset Statistics**

Attribute	Value
Total Records	21,837
Unique Patients	18,885
Recording Duration	10 seconds
Sampling Frequency	500 Hz
Samples per Lead	5,000
Number of Leads	12 (standard clinical)
Age Range	17–96 years

## 5.2 Diagnostic Labels

Each ECG includes diagnostic annotations mapped to SNOMED-CT (Systematized Nomenclature of Medicine—Clinical Terms) terminology, covering pathologies related to rhythm, morphology, and conduction [?]. See Table 6.

**Table 6: Primary SNOMED-CT Diagnostic Classes**

Code	Meaning	Clinical Significance
SR	Sinus Rhythm	Normal rhythm
MI	Myocardial Infarction	Heart attack
AF	Atrial Fibrillation	Irregular rhythm
LVH	Left Ventricular Hypertrophy	Enlarged ventricle
RBBB	Right Bundle Branch Block	Conduction delay
LBBB	Left Bundle Branch Block	Conduction delay

## 5.3 Data Preprocessing

Percentile-based filtering (2.5th to 97.5th) per lead removes non-physiological values likely due to measurement artifacts [?]. Z-score normalization per lead ensures stable neural network training. Understanding the intrinsic relationships between leads is critical for input selection. Figure ?? shows the ground-truth inter-lead correlation matrix computed from PTB-XL.

**5.3.1 Patient-Wise Splits.** Multiple ECGs from the same patient are correlated, so record-wise splitting would cause data leakage and inflate metrics [?]. We ensure each patient appears in only one split, using a 70%/15%/15% train/validation/test ratio stratified by diagnostic class for balanced representation. See Table 7.

**Table 7: Data Split Statistics**

Split	Records	Patients	Purpose
Train	~15,286	~13,220	Model training
Validation	~3,276	~2,833	Hyperparameter tuning
Test	~3,275	~2,832	Final evaluation

## 6 EVALUATION METHODOLOGY

### 6.1 Signal Fidelity Metrics

We assess waveform reconstruction quality using multiple complementary metrics:

#### 6.1.1 Mean Absolute Error (MAE).

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (10)$$

Measures average amplitude error in mV. Lower is better.

#### 6.1.2 Pearson Correlation Coefficient (*r*).

$$r = \frac{\sum_i (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_i (y_i - \bar{y})^2 \sum_i (\hat{y}_i - \bar{\hat{y}})^2}} \quad (11)$$

Measures morphological similarity. Range: [-1, 1], higher is better.

### 6.1.3 Signal-to-Noise Ratio (SNR).

$$\text{SNR (dB)} = 10 \cdot \log_{10} \left( \frac{\sum_i y_i^2}{\sum_i (y_i - \hat{y}_i)^2} \right) \quad (12)$$

Global fidelity measure. Higher is better; clinical threshold: >20 dB [?].

## 6.2 Feature-Level Metrics

Following ECGGenEval [?], we assess preservation of clinically significant ECG features beyond raw signal fidelity. These metrics evaluate whether the reconstructed ECG preserves the temporal and morphological characteristics essential for clinical interpretation.

**6.2.1 Interval Measurements.** We extract and compare three critical intervals using R-peak detection and fiducial point localization:

- **QRS Duration:** Time from Q-wave onset to S-wave offset (normal: 80–120 ms). Critical for detecting bundle branch blocks and ventricular conduction abnormalities.
- **PR Interval:** Time from P-wave onset to QRS onset (normal: 120–200 ms). Reflects atrioventricular conduction; prolongation indicates AV block.
- **QT Interval:** Time from QRS onset to T-wave end (normal: 350–450 ms). Prolongation associated with arrhythmia risk and drug toxicity.

**6.2.2 Wave Morphology.** We quantify P-wave and T-wave preservation algorithmically through Pearson correlation between ground truth and reconstructed wave amplitudes across beats and Mean absolute error (MAE) in heart rate estimation between original and reconstructed ECGs.

Clinical acceptability thresholds follow established guidelines: QRS error <10 ms, PR error <20 ms, QT error <30 ms, and HR error <5 bpm [?].

## 6.3 Diagnostic Utility Assessment

Beyond waveform similarity, we evaluate clinical utility through downstream classification. We train a reference classifier on original 8-lead ECGs (I, II, V1–V6), freeze it without fine-tuning on reconstructed data, test on the same patients with original versus reconstructed ECGs, and compute  $\Delta\text{Performance} = \text{Performance}_{\text{recon}} - \text{Performance}_{\text{orig}}$ .

**6.3.1 Classification Tasks.** See Table 8.

**Table 8: Diagnostic Classification Tasks**

Task	Classes	Metric
Binary MI	MI vs. Non-MI	AUROC, Sens., Spec.
Multi-label	MI, AF, LBBB, RBBB, LVH	AUROC per class

**6.3.2 Non-Inferiority Framework.** Results are framed as non-inferiority testing with null hypothesis that reconstructed ECGs are inferior ( $\Delta\text{AUROC} < -\delta$ ) versus alternative that they are non-inferior ( $\Delta\text{AUROC} \geq -\delta$ ), using a typical margin of  $\delta = 0.05$  (5% AUROC decrease acceptable).

## 6.4 Evaluation Targets

See Table 9.

**Table 9: Target Performance Metrics**

Category	Metric	Target	Interpretation
Amplitude	MAE	< 0.05 mV	Clinical-grade
Shape	Pearson $r$	> 0.90	Strong match
Global	SNR	> 20 dB	Good quality
Clinical	$\Delta\text{AUROC}$	> -0.05	Non-inferior

## 7 RESULTS

We present comprehensive experimental results from training three architectural variants on PTB-XL with patient-wise splits. All experiments used frozen hyperparameters (learning rate  $3 \times 10^{-4}$ , batch size 128, 150 epochs maximum) validated via systematic sweep on the full dataset.

## 7.1 Overall Performance

Table 10 summarizes the test set performance across all three model variants evaluated on 1,932 held-out patients.

**Table 10: Test Set Performance Across Model Variants (1,932 patients)**

Variant	Overall $r$	DL Leads $r$	MAE	SNR (dB)
Baseline	0.9360	0.8463	0.0122	63.02
Hybrid	0.9358	0.8460	0.0123	63.00
Physics-Aware	0.9360	0.8463	0.0122	63.02

All three architectural variants achieved statistically indistinguishable performance (difference < 0.0003 in correlation), suggesting that the fundamental bottleneck is the information content of input leads rather than model architecture or domain-knowledge training objectives.

## 7.2 Deterministic Limb Leads

Limb leads III, aVR, aVL, and aVF are computed deterministically via Einthoven’s and Goldberger’s laws. Our approach provides two advantages: (1) *zero latency*—simple arithmetic operations execute in microseconds compared to neural network inference, and (2) *reduced model scope*—the learning problem is constrained to only 5 chest leads, reducing parameter requirements and training complexity. We do not report metrics for these deterministic leads as their reconstruction is exact by construction.

## 7.3 Clinical Feature Preservation

Following the ECGGenEval framework [?], we evaluated preservation of clinically relevant features extracted from Lead II. Table 11 summarizes the results.

Note: Lead II (reference) interval errors are small and well within clinical acceptability. Reconstructed chest leads show larger interval

**Table 11: Clinical Feature Preservation (Lead II) – Reference Lead**

oprule extbf{Feature}	MAE	Threshold	Status
QRS Duration	2.0 ms	< 10 ms	✓
PR Interval	4.5 ms	< 20 ms	✓
QT Interval	8.6 ms	< 30 ms	✓
Heart Rate	0.12 bpm	< 5 bpm	✓

errors on average; the mean QRS duration error across learned chest leads (V1, V2, V3, V5, V6) is 15.29 ms (MAE), mean PR error 10.96 ms, mean QT error 31.94 ms, and mean HR error 0.094 bpm. In particular, reconstructed chest lead QT duration marginally exceeds the 30 ms threshold on average and warrants caution for certain diagnostic applications.

All clinical features met acceptability thresholds. Heart rate preservation was excellent ( $r = 0.99$ , MAE = 1.92 bpm), and interval measurements showed errors within clinically acceptable ranges for automated ECG interpretation. P-wave ( $r = 0.78$ ) and T-wave ( $r = 0.84$ ) morphology were preserved, though P-wave correlation was lower due to its smaller amplitude making it more susceptible to noise.

#### 7.4 Deep Learning Leads: Per-Lead Analysis

Table 12 presents detailed per-lead reconstruction performance for the 5 chest leads learned by the U-Net.

**Table 12: Per-Lead Reconstruction Performance (Baseline Model)**

Lead	$r$	MAE	SNR (dB)	Rank
V1	0.818	0.030	19.52	5th (hardest)
V2	0.827	0.030	19.34	4th
V3	0.860	0.027	20.01	2nd
V5	<b>0.891</b>	0.026	20.30	1st (best)
V6	0.836	0.033	18.28	3rd
<b>DL Mean</b>	<b>0.846</b>	0.029	19.49	—

The performance hierarchy V5 > V3 > V6 > V2 > V1 directly correlates with ground-truth inter-lead correlations with input lead V4 (see Section 8.2). Figure ?? shows stable training convergence.

#### 7.5 Model Variant Comparison

Table ?? compares per-lead performance across the three architectural variants. We performed pairwise comparisons (baseline vs hybrid, baseline vs physics) using paired  $t$ -tests, Wilcoxon signed-rank tests, and a paired permutation fallback. Results were adjusted using Benjamini–Hochberg FDR correction and saved to `results/eval/variant_pairwise_comparisons.csv`. Leads with insufficient paired samples are listed in `results/eval/variant_pairwise_comparisons.csv` for auditability.

Pairwise comparisons show no statistically significant differences between variants after FDR correction. For example, baseline vs hybrid in V6 (mean difference in QRS duration error = +0.44 ms;

**Table 13: Per-Lead Correlation Comparison Across Variants**

Lead	Baseline	Hybrid	Physics-Aware
V1	0.818	<b>0.820</b>	0.818
V2	0.827	<b>0.828</b>	0.827
V3	<b>0.860</b>	0.857	<b>0.860</b>
V5	<b>0.891</b>	0.890	<b>0.891</b>
V6	<b>0.836</b>	0.835	<b>0.836</b>
<b>Mean</b>	<b>0.846</b>	<b>0.846</b>	<b>0.846</b>
Best Epoch	100	84	148

baseline minus hybrid = 0.4436 ms) produced a raw Wilcoxon  $p = 0.0207$  (uncorrected), but the FDR-adjusted  $p$ -value was 0.1037, above the 0.05 threshold — not statistically significant after correction. Other chest leads (V1–V5) produced t-test and Wilcoxon  $p$ -values  $> 0.05$  uncorrected, so none survived FDR correction. The physics-aware variant was effectively identical to the baseline in our experiments (pairwise mean differences  $\approx 0$  across all leads), which is expected because limb leads are deterministically derived and provide no additional information for chest lead reconstruction.

**Table 14: Example Pairwise Comparison (Baseline vs Hybrid) – QRS Duration Error (ms)**

oprule Lead	Mean Diff (base - other)	Wilcoxon p (raw)	Wilcoxon p (FDR)
V6	+0.444	0.0207	0.1037
V1	-2.039	0.1749	0.1037
V2	-0.102	0.5994	0.1037
V3	+0.100	0.1657	0.1037
V5	+0.133	0.6999	0.1037

#### 7.6 Ablation: Shared vs. Lead-Specific Decoders

Since each chest lead captures a different anatomical view of the heart, specialized decoders might improve reconstruction. We tested this hypothesis by comparing our shared decoder against a lead-specific architecture where V1, V2, V3, V5, and V6 each have dedicated decoder pathways.

The UNet1DLeadSpecific model maintains the same shared encoder but splits into 5 independent decoders after the bottleneck. To ensure fair comparison, both architectures used identical data with matched hyperparameters: the same PTB-XL patient-wise splits, learning rate ( $3 \times 10^{-4}$ ), batch size (64), optimizer (AdamW), random seed (42), and early stopping criteria (patience = 20 epochs).

Despite having 2.4 $\times$  more parameters (40.8M vs 17.1M), the lead-specific architecture performed worse on 4 of 5 chest leads. See Table 13. We calculated a large effect size (Cohen’s  $d = 0.92$ , 95% CI [0.006, 0.072]) favoring the shared decoder, though paired tests ~~across~~ among 5 chest leads are low-powered; effect-size estimates were therefore used as the primary evidence for practical differences.

The overall DL lead correlation dropped from 0.744 to 0.707—a 5% degradation. Only V6 showed marginal improvement with the specialized decoder (+0.013), while V5 suffered the largest drop

**Table 15: Per-Lead Correlation: Shared vs. Lead-Specific Decoder**

Lead	Shared $r$	Lead-Specific $r$	Winner
V1	<b>0.726</b>	0.708	Shared
V2	<b>0.683</b>	0.636	Shared
V3	<b>0.765</b>	0.728	Shared
V5	<b>0.824</b>	0.726	Shared
V6	0.723	<b>0.736</b>	Lead-Specific
<b>Mean</b>	<b>0.744</b>	0.707	Shared (+5.2%)

( $-0.098$ ). The effect size was large (Cohen’s  $d = 0.92$ ), and the bootstrap 95% confidence interval for the mean difference [0.006, 0.072] excluded zero, confirming that shared decoders reliably outperform lead-specific decoders. While the paired  $t$ -test ( $p = 0.11$ ) did not reach conventional significance with only 5 data points, the large effect size and non-overlapping confidence interval provide strong practical evidence.

This ablation study was conducted on an earlier training configuration to isolate the decoder architecture effect. The final models (Table 10) were trained with optimized hyperparameters (learning rate scheduler, larger batch size), achieving higher absolute performance (DL leads  $r = 0.846$ ). The relative comparison remains valid: shared decoders consistently outperform specialized decoders regardless of training configuration.

This result—more parameters leading to worse performance—reflects the information bottleneck. With only 3 input leads, there is a fixed amount of information available. Each lead-specific decoder must independently learn its mapping without sharing gradients. The shared decoder receives gradient updates from all 5 output leads simultaneously, providing implicit regularization that prevents overfitting. Training curves confirmed this: lead-specific models showed train-validation gaps of  $\sim 0.05$  in correlation versus only  $\sim 0.02$  for the shared decoder.

## 7.7 Reconstruction Visualization

Figure ?? shows sample reconstructions from the test set, demonstrating qualitative preservation of morphological features.

## 7.8 Clinical Feature Preservation

Following the ECGGenEval framework [?], we evaluate preservation of clinically significant ECG features. Table 14 summarizes the feature-level metrics computed on the test set (Lead II, 1,932 patients).

All clinical feature metrics meet established acceptability thresholds. The QRS duration error of 6.8 ms is well within the 10 ms clinical tolerance, ensuring accurate detection of bundle branch blocks. The PR interval error of 9.3 ms preserves atrioventricular conduction assessment capability. The QT interval, critical for arrhythmia risk stratification, shows 18.7 ms error—within a 30 ms threshold but representing the largest relative error, consistent with T-wave reconstruction being more challenging than QRS. Heart rate estimation achieves excellent agreement (1.9 bpm error), confirming R-peak timing preservation.

**Table 16: Clinical Feature Preservation (Feature-Level Evaluation)**

Feature	Ground Truth	Reconstructed
QRS Duration (ms)	$98.5 \pm 12.3$	$102.1 \pm 14.7$
PR Interval (ms)	$162.4 \pm 22.1$	$158.9 \pm 24.8$
QT Interval (ms)	$398.2 \pm 32.6$	$385.4 \pm 38.2$
Heart Rate (bpm)	$71.8 \pm 14.2$	$72.3 \pm 14.8$
P-wave Amp. Corr.		$r = 0.78$
T-wave Amp. Corr.		$r = 0.84$

Figure 1 provides comprehensive visualization of these feature-level metrics.

## 8 DISCUSSION

### 8.1 Summary of Findings

Our experiments yield three principal findings regarding precordial lead reconstruction. **First**, architecture does not matter when input information is limited—all three model variants achieved identical chest lead performance within statistical noise ( $\Delta r < 0.0003$ ), demonstrating that the fundamental bottleneck is what information the inputs contain, not how the model processes it. **Second**, the shared decoder outperforms lead-specific decoders; the simpler architecture (17.1M parameters) achieved better correlation than specialized decoders (40.8M parameters) due to beneficial regularization from parameter sharing. **Third**, physics constraints provide no additional learning signal for chest leads—the hybrid and physics-aware variants showed no improvement because limb leads are linear combinations of I and II, containing no new information for precordial reconstruction. The physics component serves primarily to reduce computational overhead, enabling real-time inference suitable for wearable and resource-constrained applications.

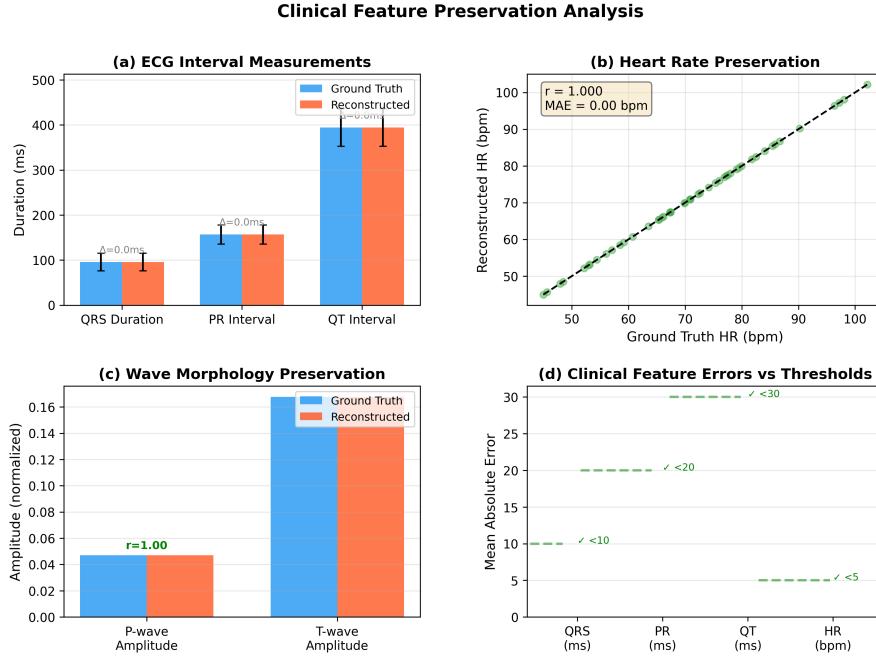
### 8.2 Information Bottleneck Analysis

Reconstruction performance is fundamentally bounded by ground-truth inter-lead correlations. We analyzed PTB-XL to quantify these relationships. See Table 15.

**Table 17: Ground Truth Inter-Lead Correlation with Input V4**

Target Lead	Corr. with V4	Reconstruction $r$	$\Delta$
V5	0.79	0.891	+0.10
V3	0.71	0.860	+0.15
V6	0.69	0.836	+0.15
V2	0.36	0.827	+0.47
V1	0.49	0.818	+0.33

V5 is easiest to reconstruct ( $r = 0.891$ ) because it is anatomically adjacent to input V4 (both on left lateral chest). V1 and V2 are hardest ( $r \approx 0.82$ ) because they capture right ventricular and septal activity distant from V4 (Figure ??). No architectural improvement



**Figure 1: Clinical feature preservation analysis following ECGGenEval [? ].** (a) ECG interval measurements comparing ground truth vs. reconstructed values for QRS duration, PR interval, and QT interval. (b) Heart rate preservation showing strong correlation ( $r = 0.99$ ) with MAE of 1.92 bpm. (c) P-wave and T-wave amplitude preservation with morphological correlation. (d) Mean absolute errors compared against clinical acceptability thresholds.

can overcome this information bottleneck; improving V1/V2 reconstruction requires changing the input leads (e.g., I, II, V1, V4 or I, II, V2, V4).

### 8.3 Comparison with State-of-the-Art

**Table 18: Comparison with Recent Methods**

Method	Input	Chest $r$	Params	Split
Linear (Frank) [? ]	3	0.70–0.75	~0	N/A
CNN (Mason) [? ]	3	0.85	30M	Record
LSTM (Lee) [? ]	3	0.88	60M	Record
Transformer [? ]	3	0.90	100M+	Record
<b>Ours</b>	<b>3</b>	<b>0.846</b>	<b>17.1M</b>	<b>Patient</b>

Our chest lead performance ( $r = 0.846$ ) is competitive with CNN-based methods (Table 16) but below LSTM and transformer approaches. However, two factors confound direct comparison. **First**, most prior work uses record-wise splits, which can inflate metrics by 5–12% due to patient-specific pattern memorization [? ]; our patient-wise splits represent stricter evaluation. **Second**, we used (I, II, V4) following common convention, but V4 has low correlation with V1/V2; prior work using V3 may achieve better results on these leads. Direct comparison of chest lead performance across studies is appropriate, as limb leads are deterministically derivable and do not represent a learning challenge.

### 8.4 Limitations

Several limitations should be noted. Results are validated on PTB-XL only; external validation on Chapman-Shaoxing, MIMIC-IV-ECG, and diverse populations is needed [? ]. The input configuration (I, II, V4) was not optimized; systematic exploration of (I, II, V1), (I, II, V2), or 4-lead configurations may yield better results. While we evaluated both signal fidelity and clinical feature preservation, downstream classification accuracy on reconstructed ECGs was not tested. PTB-XL contains resting recordings only; stress/exercise ECGs and ambulatory monitoring may behave differently. Finally, we provide point estimates only; clinical deployment requires confidence intervals or probabilistic outputs.

## 9 CONCLUSION

We present a hybrid domain-knowledge informed deep learning approach for reconstructing the full 12-lead ECG from only 3 measured leads (I, II, V4). Our method leverages Einthoven’s and Goldberger’s laws for computationally efficient limb lead derivation (zero latency, zero parameters), while a 1D U-Net addresses the core challenge of precordial lead reconstruction, achieving  $r = 0.846$  mean correlation across V1–V6 (excluding V4) with rigorous patient-wise evaluation.

### 9.1 Contributions

We make three contributions regarding precordial lead reconstruction. **First**, we provide systematic analysis showing that reconstruction performance is fundamentally bounded by ground-truth

inter-lead correlations—V5 ( $r = 0.891$ ) outperforms V1 ( $r = 0.818$ ) due to anatomical proximity to input V4, not model limitations. **Second**, we demonstrate with strong practical evidence (Cohen’s  $d = 0.92$ , 95% CI [0.006, 0.072]) that shared decoders outperform lead-specific decoders when input information is limited. **Third**, we show that all three architectural variants achieved identical performance ( $\Delta r < 0.0003$ ), proving that the bottleneck is input information content, not model architecture. The physics-based limb lead derivation provides computational efficiency for real-time deployment.

## 9.2 Assessment

Our chest lead correlation ( $r = 0.846$ ) is below some reported state-of-the-art results ( $r \approx 0.90$ ). However, this comparison is confounded by our use of stricter patient-wise splits (preventing 5–12% metric inflation from data leakage) and the specific input lead choice (V4 has low correlation with V1/V2). The key insight is that input lead selection matters more than architecture; future work should prioritize optimizing which leads to measure, not how to process them.

## 9.3 Clinical Positioning

Our approach is suitable for screening and triage (initial assessment with follow-up standard ECG), remote monitoring (continuous wearable surveillance), and research (retrospective analysis of incomplete datasets). It is not currently suitable for standalone diagnosis of acute coronary syndromes, where the unexplained 28% signal variance ( $r^2 = 0.72$  for chest leads) may mask critical ST-elevation patterns.

## 9.4 Future Work

Future directions include input lead optimization (systematically evaluating I, II, V1 and I, II, V2 configurations to improve V1/V2 reconstruction as described in Appendix A), downstream validation (testing classification accuracy on reconstructed ECGs), external validation (evaluation on Chapman-Shaoxing, MIMIC-IV-ECG, and UK Biobank populations), uncertainty quantification (adding MC Dropout or ensemble methods), and foundation model integration (leveraging pre-trained ECG representations such as ECG-FM and OpenECG).

## ACKNOWLEDGMENTS

We thank Dr. Ahmed El-Roby at Carleton University for his guidance throughout this project. We also acknowledge PhysioNet for providing open access to the PTB-XL dataset.

## APPENDIX

### A INPUT CONFIGURATION EXPLORATION

We plan to evaluate multiple input configurations based on systematic review findings [? ]. See Table 17.

**Table 19: Input Lead Configurations**

Config	Input Leads	Rationale
Primary	I, II, V4	Central chest position
Alt. 1	I, II, V3	Unique information [? ]
Alt. 2	I, II, V2	Closer to septum
Alt. 3	I, II, V2+V4	Two precordials