ICML
International Conference
On Machine Learning

# Position: Evaluation of ECG Representations Must Be Fixed

Zachary Berger[*1,2], Daniel Prakah-Asante[*1,2], John Guttag[1], Collin Stultz[1,2]

[1]MIT, [2]MGH
[*] Equal contribution

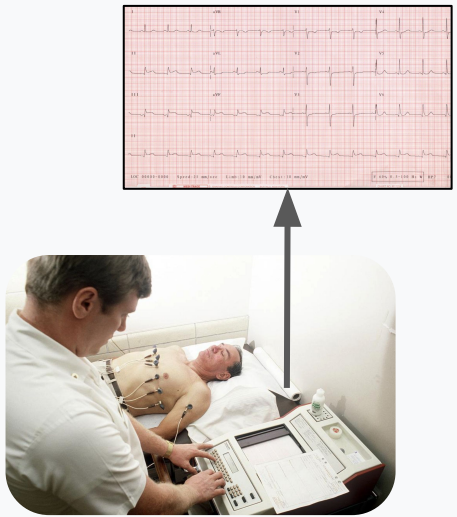
MGH 1811 MASSACHUSETTS GENERAL HOSPITAL

MIT CSAIL

HARVARD MEDICAL SCHOOL

# Background

## 12-lead ECG

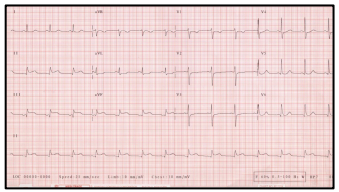


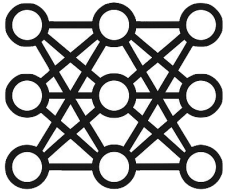
**Diagnose**
Arrhythmias, waveform abnormalities

**Infer patient state**
e.g., structural heart disease, hemodynamics

Images, top (left to right): Glenlarson / Wikimedia Commons; CardioNetworks & Drj / Wikimedia Commons.
Images, bottom (left to right): Theodore Moore / U.S. National Archives; Fruehaufsteher2 / Wikimedia Commons; Collin M. Stultz.

# Background





Pre-trained Encoder

Embedding

# Position

Current 12-lead ECG representation evaluations are inadequate…

**Narrow task set**

- Arrhythmia
- Waveform abnormality

**Unreliable protocol**

- Macro-AUROC
- Point estimates
- Evaluate tasks with insufficient support

# Position

…we must fix them!

**Broader task set**

- Arrhythmia
- Waveform abnormality
- Structural disease
- Hemodynamics
- Patient forecasting

**Revised protocol**

- Task-specific AUCs
- Quantify uncertainty
- Standard baseline
- Exclude tasks with low test-support

# Experiments

We evaluate 6 encoders on 6 datasets.

## Encoders

MERL[1], D-BETA[2], KED[3], HeartLang[4], CLOCS[5],
Randomly initialized ResNet[6]

[1] Liu et al., ICML 2024. [2] Pham et al., ICML 2025. [3] Tian et al., Cell Reports Medicine 2024.
[4] Jin et al., ICLR 2025. [5] Kiyasseh et al., ICML 2021. [6] He et al., CVPR 2016

## Datasets

PTB-XL[7], CPSC2018[8], CSN[9], EchoNext[10], 1yr-HF[11], Hemodynamics[12]

[7] Wagner et al., PhysioNet 2022. [8] Liu et al., 2018. [9] Zheng et al., PhysioNet 2022. [10] Poterucha et al., Nature 2025.
[11] Bergamaschi et al., eClinicalMedicine 2025. [12] Schlesinger et al., JACC: Advances 2022.

**Broader task set + revised protocol ⇒ new conclusions about best methods!**

# Dataset-level Results

No method clearly dominates.

Randomly initialized encoder is surprisingly performant.

| Dataset | Random | CLOCS | KED | HeartLang | MERL | D-BETA |
|---|---|---|---|---|---|---|
| PTB-XL Form | 0.756 0.734–0.780 | 0.715 0.687–0.741 | 0.851 0.827–0.875 | 0.707 0.679–0.736 | 0.853 0.839–0.866 | 0.845 0.821–0.868 |

Table 1. Macro-AUROC on PTB-XL Form with 95% CIs.

# Task-level Results

No method clearly dominates. Randomly initialized encoder is competitive.

| Method | NORM | CLBBB |
|---|---|---|
| Random | 0.891 0.877–0.903<br>/ 0.839 0.813–0.863 | 0.973 0.934–0.999<br>/ 0.880 0.784–0.955 |
| CLOCS | 0.859 0.843–0.872<br>/ 0.788 0.763–0.812 | 0.984 0.974–0.993<br>/ 0.676 0.551–0.794 |
| KED | 0.932 0.921–0.942<br>/ 0.900 0.883–0.916 | 0.998 0.997–1.000<br>/ 0.949 0.906–0.983 |
| HeartLang | 0.875 0.860–0.890<br>/ 0.822 0.798–0.846 | 0.993 0.984–0.998<br>/ 0.867 0.767–0.944 |
| MERL | 0.926 0.915–0.937<br>/ 0.885 0.862–0.904 | 0.999 0.997–1.000<br>/ 0.947 0.889–0.988 |
| D-BETA | 0.929 0.919–0.939<br>/ 0.894 0.875–0.912 | 0.999 0.997–1.000<br>/ 0.970 0.937–0.995 |

Table 2. AUROC (top) and AUPRC (bottom) with 95% CIs on PTB-XL Norm and CLBBB.

# Hemodynamics + Patient Forecasting

No method clearly dominates, performance is marginal beyond random baseline.

| **Method** | mPCWP | mPA | 1YR-HF |
|---|---|---|---|
| Random | 0.70 0.67–0.73<br>/ 0.71 0.68–0.75 | 0.72 0.68–0.76<br>/ 0.86 0.83–0.88 | 0.78 0.78–0.79<br>/ 0.58 0.58–0.59 |
| CLOCS | 0.68 0.64–0.71<br>/ 0.71 0.68–0.74 | 0.66 0.63–0.70<br>/ 0.84 0.81–0.86 | 0.75 0.74–0.75<br>/ 0.53 0.53–0.54 |
| KED | 0.72 0.69–0.75<br>/ 0.75 0.71–0.78 | 0.76 0.73–0.79<br>/ 0.89 0.87–0.91 | 0.83 0.82–0.83<br>/ 0.66 0.65–0.66 |
| HeartLang | 0.68 0.65–0.71<br>/ 0.70 0.66–0.73 | 0.71 0.67–0.74<br>/ 0.86 0.84–0.88 | 0.79 0.78–0.79<br>/ 0.58 0.58–0.59 |
| MERL | 0.74 0.71–0.77<br>/ 0.76 0.73–0.79 | 0.76 0.73–0.80<br>/ 0.88 0.86–0.90 | 0.83 0.83–0.83<br>/ 0.66 0.66–0.67 |
| D-BETA | 0.71 0.68–0.74<br>/ 0.73 0.70–0.77 | 0.74 0.71–0.78<br>/ 0.87 0.85–0.90 | 0.82 0.82–0.82<br>/ 0.64 0.63–0.64 |

Table 3. AUROC (top) and AUPRC (bottom) with 95% CIs on hemodynamic (mPCWP, mPA) and patient forecasting (1yr-HF).

# Key Takeaways

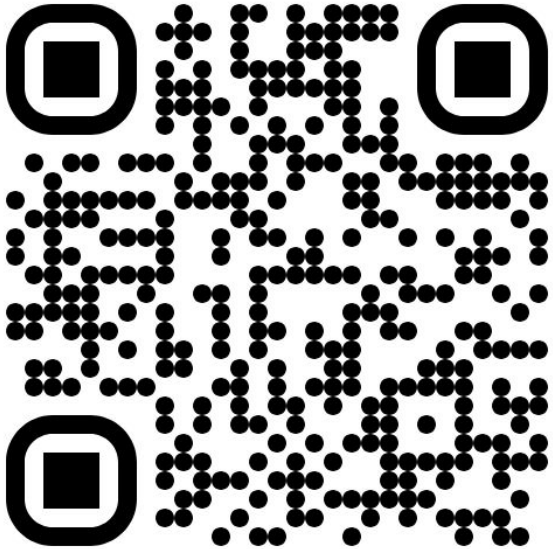
1) Current 12-lead ECG evaluations are inadequate
    - Narrow task set
    - Unreliable protocol

2) We propose a broader task set and revised protocol

3) We find
    - No clear best-performing method
    - Randomly initialized encoder does surprisingly well

# Evaluation of ECG Representations Must Be Fixed

Code and paper:



https://ecgfix.csail.mit.edu/