# Phenotype-Aware Supervised Contrastive Learning for ECG-CMR Representation Alignment

Daniel Uyterlinde[1], Laura Alvarez-Florez[12], Samuel Ruipérez-Campillo[34], Lukas P.A. Arts[2], Maarten Z.H. Kolk[2], Erik Bekkers[1], Fleur V.Y. Tjong[2]

[1] Informatics Institute, University of Amsterdam, The Netherlands
[2] Department of Cardiology, Amsterdam University Medical Center, Amsterdam, Netherlands
[3] Department of Computer Science, ETH Zurich, Switzerland
[4] Institute of Medical Engineering and Science, Massachusetts Institute of Technology, MA, USA

## Abstract

*Cardiovascular disease remains a leading cause of morbidity and mortality worldwide, driving demand for scalable tools that can extract rich cardiac information from widely available diagnostic tests. Cross-modal contrastive learning between electrocardiograms (ECGs) and cardiac magnetic resonance (CMR) imaging enables transfer of structural cardiac information into the accessible ECG modality. However, standard contrastive alignment objectives can impose overly rigid separation between physiologically similar patients, distorting the continuous structure that downstream clinical tasks rely on. We propose Phenotype-Aware Supervised Contrastive Learning (Pheno-SupCon), which instead uses a soft similarity structure derived from continuous cardiac phenotypes and categorical diagnoses to better preserve clinically meaningful relationships. We align ECG representations to a CMR target space constructed from seven clinical phenotypes from the UK Biobank (∼33K paired subjects). Pheno-SupCon achieves statistically significant improvements over standard contrastive alignment in four of five regression targets, while maintaining comparable classification performance. Full finetuning experiments at reduced fractions of training data further demonstrate that phenotype-aware representations yield consistent gains in low-data regimes for structural phenotype regression.*

## 1. Introduction

Cardiac Magnetic Resonance (CMR) imaging is the gold standard for non-invasive cardiac assessment, providing volumetric data for precise quantification of phenotypes such as Left Ventricular End-Diastolic Volume (LVEDV), ejection fraction (EF), and myocardial mass. However, CMR requires specialized equipment and expertise, limiting its accessibility compared to the widely available 12-lead ECG. Because electrical activity and mechanical function are tightly coupled, recent work has explored *cross-modal contrastive learning* to transfer CMR-derived structural information into ECG representations [1, 2], enabling richer cardiac phenotyping from ECG alone. These methods apply InfoNCE-based objectives [3] to patient-matched ECG-CMR pairs, treating each patient as a unique class: the matched pair is the only positive, and all other batch elements are pushed apart as negatives. While effective, this strict instance discrimination carries a fundamental limitation when applied to clinical data: it treats *physiologically similar patients as negatives*. Two patients with near-identical ejection fractions or ventricular volumes are pushed apart simply because they are different individuals. This *false negative* problem distorts the learned embedding space, penalizing the very continuous functional gradients that downstream clinical tasks depend on.

In this work, we propose **Pheno-SupCon**, a Phenotype-Aware Supervised Contrastive Learning framework that replaces binary instance discrimination with a soft, clinically derived similarity structure. Rather than pushing all non-identical patients apart, the model learns to respect the physiological continuum: patients with similar cardiac function are allowed to remain close in latent space.

## 2. Methodology

Pheno-SupCon aligns ECG representations to a CMR target space via a phenotype-aware contrastive objective. We first describe the encoder infrastructure and CMR target construction shared with our InfoNCE baseline.

### 2.1. Encoders and CMR Target Space.

For ECG encoding we use **ECG-FM** [4], a hybrid CNN-Transformer trained on ∼1.4M 12-lead ECGs. For CMR, we use **CineMA** [5], a multi-view masked autoencoder

pre-trained on the full UK Biobank CMR corpus, which we retrained on a dedicated 70/10/20 split for 100 epochs (5.32M images, ∼2 days) to prevent data leakage.

A critical design choice in our framework is the use of a *frozen supervised CMR target* $\mathbf{z}_{\text{pheno}}$ rather than raw CineMA embeddings. The raw CineMA embedding space is nearly degenerate with respect to functional cardiac variation: the cosine similarity between end-diastole and end-systole representations exceeds 0.996, meaning the encoder assigns near-identical representations to the fully relaxed and fully contracted heart. Despite this similarity, most of the functional information regarding phenotypes like ejection fraction are contained in this small difference vector. We therefore train a lightweight supervised MLP $h_\omega$ ($1536 \to 512 \to 256$) that maps the concatenated end-diastolic and end-systolic CineMA representations $[\mathbf{z}_{\text{ED}}; \mathbf{z}_{\text{ES}}] \in \mathbb{R}^{1536}$ to a compact 256-dimensional target by regressing seven clinical phenotypes (LVEF, LVEDV, LV mass, RVEDV, LVGLS, AFib, MI)[1]:

$$\mathbf{z}_{\text{pheno}}^{(\mathbf{i})} = \mathbf{h}_\omega\left([\mathbf{z}_{\text{ED}}^{(\mathbf{i})}; \mathbf{z}_{\text{ES}}^{(\mathbf{i})}]\right). \tag{1}$$

Once trained, $h_\omega$ is frozen for all subsequent experiments. This concentrates representational capacity on physiologically meaningful and ECG-predictable dimensions of the CMR space, while stabilizing training.

## 2.2. Baselines

Our first baseline is the ECG-FM [4] foundation model, trained purely on ECGs without any cross-modal alignment, serving as a unimodal baseline.

The contrastive baseline is similar to [1, 2]. ECG-FM features (pooled at layer 9 to $\mathbb{R}^{768}$) are projected to the shared 256-dimensional space via an MLP $g_\phi$. The objective is an asymmetric InfoNCE loss; since the CMR target is frozen, gradients flow only through the ECG encoder and its projection layer:

$$\mathcal{L}_{\text{CL}} = -\frac{1}{\mathcal{B}} \sum_{\mathbf{i}=1}^{\mathcal{B}} \log \frac{\exp(\mathbf{z}_{\text{ecg}}^{(\mathbf{i})} \cdot \mathbf{z}_{\text{pheno}}^{(\mathbf{i})}/\tau)}{\sum_{\mathbf{k}=1}^{\mathcal{B}} \exp(\mathbf{z}_{\text{ecg}}^{(\mathbf{i})} \cdot \mathbf{z}_{\text{pheno}}^{(\mathbf{k})}/\tau)}, \tag{2}$$

with temperature $\tau = 0.1$. This loss treats each patient as the only positive and uniformly repels all others, regardless of physiological similarity.

## 2.3. Phenotype-Aware Supervised Contrastive Learning

Pheno-SupCon replaces the binary positive/negative assignment with a soft similarity target derived from clinical phenotypes. Our key contribution is generalizing the supervised contrastive framework of [6] from categorical labels to *continuous phenotypic similarity* through Radial Basis Function kernels. Our method most closely resembles that of [7], who propose SupReMix, a supervised contrastive framework for medical imaging regression that uses embedding-level mixup and label-distance weighting. However, SupReMix operates in the unimodal setting, constructing synthetic contrastive pairs within a single modality. Our work differs in two respects: first, we operate in a cross-modal alignment setting where ECG representations are trained against a frozen CMR target space, meaning the contrastive pairs are real patient-matched recordings rather than synthetic interpolations; second, rather than deriving similarity from a single scalar target, we construct a multivariate similarity matrix over multiple continuous cardiac phenotypes diagnoses, capturing the joint physiological structure of the patient population.

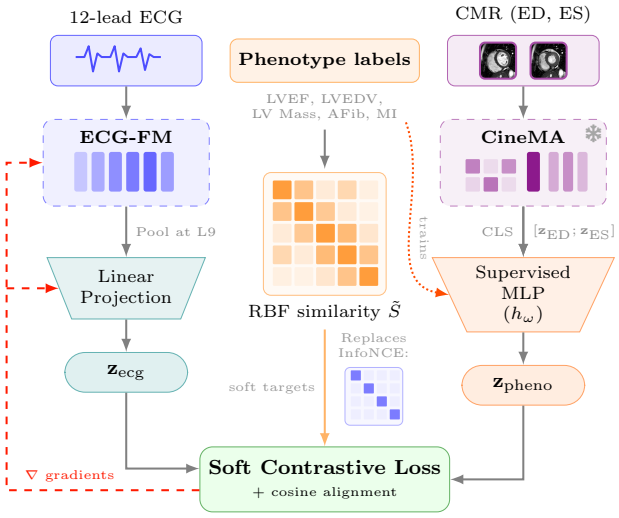


Figure 1. Method overview.

**Similarity matrix construction.** Let $\mathbf{y_i}$ represent the clinical phenotype vector for patient $i$, partitioned into continuous variables $\mathcal{C}$ (LVEF, LVEDV, LV mass, RVEDV, LVGLS) and binary diagnoses $\mathcal{B}$ (Atrial Fibrillation, Myocardial Infarction).

For continuous phenotypes, we compute a standardized squared Euclidean distance between valid phenotypic intersections (accounting for missing data) and transform it via an RBF kernel:

$$S_{\text{cont}}^{(i,j)} = \exp\left(-\frac{d_{\text{cont}}(\mathbf{y_i}, \mathbf{y_j})^2}{2\sigma}\right), \tag{3}$$

where $\sigma$ is an adaptive bandwidth set to the median pairwise distance within each batch. For binary diagnoses, we apply an exponential penalty over the fractional mismatch:

$$S_{\text{cat}}^{(i,j)} = \exp(-\beta \cdot \text{mismatch}(\mathbf{y_i}, \mathbf{y_j})). \tag{4}$$

[1]LVEF: Left Ventricular Ejection Fraction, LVEDV: Left Ventricular End-Diastolic Volume, LV mass: Left Ventricular mass, RVEDV: Right Ventricular End-Diastolic Volume, LVGLS: Left Ventricular Global Longitudinal Strain, AFib: Atrial Fibrillation, MI: Myocardial Infarction

The two components are combined as $S_{\text{total}}^{(i,j)} = \left(\alpha\, S_{\text{cont}}^{(i,j)} + (1-\alpha)\, S_{\text{cat}}^{(i,j)}\right)^{\gamma}$, where $\alpha \in [0,1]$ controls the relative weight of continuous versus categorical similarity and $\gamma$ sharpens the distribution. Top-$K$ sparsification then retains only the $K$ most similar neighbors per anchor before $L_1$-normalization.

This matrix encodes the clinical insight that a patient with LVEF 58% is more similar to one with LVEF 55% than to one with LVEF 30%, and that shared diagnoses carry additional discriminative weight. Notably, this framework is extensible to additional/different phenotypes with minimal modification.

**Optimization objective.** The ECG projection $\mathbf{z}_{\text{ecg}}^{(\mathbf{i})} = \mathbf{g}_\phi(\mathbf{f}_\theta(\mathbf{x}_{\text{ecg}}^{(\mathbf{i})}))$ is trained against the frozen CMR target via a soft cross-entropy over cosine similarities:

$$\mathcal{L}_{\text{SupCon}} = -\frac{1}{\mathcal{B}} \sum_{\text{i}=1}^{\mathcal{B}} \sum_{\text{j}=1}^{\mathcal{B}} \tilde{\mathcal{S}}_{\text{total}}^{(\text{i,j})} \log \frac{\exp(\mathbf{z}_{\text{ecg}}^{(\mathbf{i})} \cdot \mathbf{z}_{\text{cmr}}^{(\mathbf{j})}/\tau)}{\sum_{\text{k}=1}^{\mathcal{B}} \exp(\mathbf{z}_{\text{ecg}}^{(\mathbf{i})} \cdot \mathbf{z}_{\text{cmr}}^{(\mathbf{k})}/\tau)}. \tag{5}$$

This is combined with a direct patient-wise cosine alignment term to anchor each ECG representation to its own CMR target:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{SupCon}} + \lambda_{\text{align}} \frac{1}{\mathcal{B}} \sum_{\text{i}=1}^{\mathcal{B}} \left(1 - \cos(\mathbf{z}_{\text{ecg}}^{(\mathbf{i})}, \mathbf{z}_{\text{cmr}}^{(\mathbf{i})})\right). \tag{6}$$

## 3. Experimental Setup

**Data.** We use the UK Biobank, selecting ∼33K subjects with paired 12-lead ECG recordings and CMR cine sequences, split 70/10/20 for train/validation/test. Patients without labels are added to the training split. Nine clinical targets are evaluated: LVEF, LVEDV, LV mass, RVEDV, and LVGLS as continuous regression tasks ($R^2$), and $\text{LVEF}_{\leq 40}$, $\text{LVEF}_{\leq 50}$, Myocardial Infarction and Atrial Fibrillation as binary classification tasks (AUROC).

**Evaluation protocols.** We report two evaluation settings. In *frozen probing*, a linear probe is trained on frozen ECG representations. In *label-efficiency fine-tuning*, the full ECG encoder is fine-tuned using 10%, 50%, and 100% of the labeled training data. The latter directly measures the practical value of learned representations for clinical deployment, where labeled data is scarce.

## 4. Results

**Linear Probing.** Table 1 reports linear probing performance on frozen 9th-layer ECG representations. Pheno-SupCon achieves the highest $R^2$ across all five regression targets, with statistically significant improvements over the contrastive baseline in four (LVEF, LVEDV, LV Mass, RVEDV; $p \leq 0.05$ via bootstrap $\Delta R^2$). Interestingly, our method generalizes well to phenotypes that are not included in the phenotype similarity matrix (RVEDV, LVGLS).

Table 1. Linear probing $9^{th}$ layer features on held-out test set. †Not used in similarity target matrix. *Statistically significant ($p$<0.05) vs. Contrastive (bootstrap $\Delta R^2$; DeLong for AUROC). **Bold**: best per row.

| Phenotype | Baseline | Contrastive | Pheno-SupCon |
|---|---|---|---|
| | *Regression ($R^2$)* | | |
| LVEF | 0.274 | 0.306 | **0.318*** |
| LVEDV | 0.538 | 0.586 | **0.612*** |
| LV Mass | 0.630 | 0.683 | **0.700*** |
| RVEDV† | 0.545 | 0.593 | **0.613*** |
| LVGLS† | 0.207 | 0.223 | **0.233** |
| | *Classification (AUROC)* | | |
| LVEF≤40% | 0.863 | **0.900** | 0.885 |
| LVEF≤50% | 0.816 | 0.829 | **0.836** |
| AFib | 0.725 | 0.731 | **0.734** |
| MI | 0.725 | **0.754** | 0.737 |

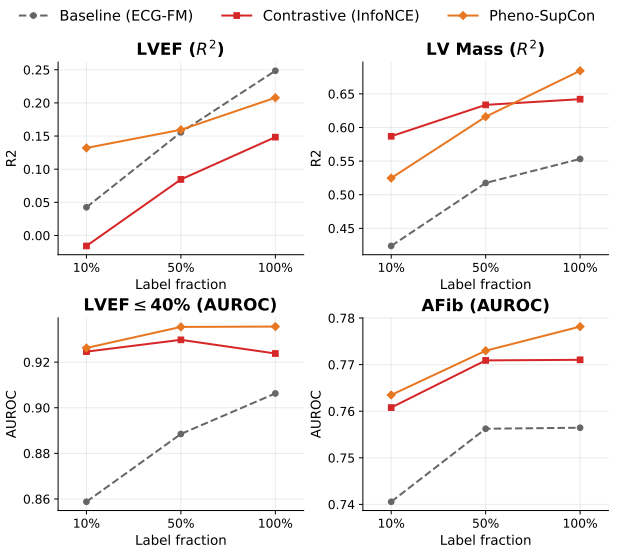


Figure 2. Label-efficiency fine-tuning results for $9^{th}$ layer features across four representative targets. The full ECG encoder is fine-tuned using 10%, 50%, and 100% of labeled training data.

For binary classification, results are mixed; no method is significantly better than the other. Notably, both alignment methods improve substantially over the ECG-FM baseline across all targets, confirming the value of cross-modal alignment.

**Label-Efficiency Fine-Tuning.** Figure 2 shows fine-tuning performance as a function of labeled data availability. Pheno-SupCon establishes a powerful prior for label-

scarce regimes (10%), and scales highly efficiently on LV Mass. However, the baseline's crossover on LVEF ($R^2$) at 100% data indicates that aligning to a frozen cross-modal target may compress task-specific features when fully supervised training data is abundant.

## 5. Discussion

Pheno-SupCon demonstrates that the contrastive objective need not be limited to discrete instance discrimination. By replacing binary positive/negative assignments with a continuous similarity structure derived from clinical phenotypes, the learned manifold can reflect genuine physiological gradients rather than arbitrary patient boundaries. This reinterpretation of contrastive learning over continuous distance functions rather than discrete labels, is broadly applicable beyond the ECG-CMR setting.

The regression improvements under linear probing support this: Pheno-SupCon consistently produces better-organized representations along physiological gradients, with significant gains on four of five regression targets. No significant differences between methods were observed for classification targets, consistent with the threshold-based nature of binary tasks, which require correct cluster placement rather than fine-grained manifold structure.

The label-efficiency results are practically significant. In clinical deployment, labeled ECG-CMR pairs are expensive to acquire, and the ability to achieve competitive performance with 10% of labels directly reduces annotation burden. That Pheno-SupCon's advantage is most pronounced in this regime suggests that its representations encode clinically meaningful structure that is otherwise only recoverable through extensive supervised fine-tuning. The magnitude of the effect may also depend on the choice of phenotypes used to construct the similarity matrix; a richer or better-weighted set of clinical features could strengthen the inductive bias and potentially yield larger gains. This represents a promising direction for future work.

Several limitations should be noted. First, our similarity matrix is constructed from five phenotypes which overlap with evaluation targets. However, improvements on held-out phenotypes (RVEDV, LVGLS) suggest the gains reflect genuine representational structure rather than overfitting to the similarity construction. The choice of phenotypes and their relative weighting remains a design decision, and sensitivity to these choices requires further study. Second, the RBF bandwidth $\sigma$ is set via a batch-level median heuristic, which may be suboptimal for highly skewed phenotype distributions, or in the presence of outliers. Finally, a trainable CMR projection head could allow the shared space to reorient under downstream supervision, potentially recovering task-specific structure that the frozen projection discards. We leave this to future work.

In conclusion, Pheno-SupCon demonstrates that clinically grounded continuous similarity structure can meaningfully improve cross-modal ECG-CMR alignment for phenotype regression. The framework is easily extensible to additional phenotypes and similarity kernels, offering a principled path toward representation spaces that respect the continuous structure of cardiovascular phenotypes.

## Acknowledgments

This research has been conducted using the UK Biobank Resource under application number 24711 and has been supported by grants from the Dutch Heart Foundation (Dekker 03-005-2025-0129) Dutch Research Council (Rubicon 452019308, AINed XS NGF.1609.243.045) and Amsterdam Cardiovascular Sciences, Health Holland. Computational resources have been provided by the University of Amsterdam and SURF.

## References

[1] Ding Z, Hu Y, Li Z, Zhang H, Wu F, Xiang Y, Li T, Liu Z, Chu X, Huang Z. Cross-modality cardiac insight transfer: A contrastive learning approach to enrich ecg with cmr features. In International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2024; 109–119.

[2] Özgün Turgut, Müller P, Hager P, Shit S, Starck S, Menten MJ, Martens E, Rueckert D. Unlocking the diagnostic potential of electrocardiograms through information transfer from cardiac magnetic resonance imaging. Medical Image Analysis 2025;101:103451. ISSN 1361-8415.

[3] van den Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding, 2019.

[4] McKeen K, Masood S, Toma A, Rubin B, Wang B. Ecg-fm: an open electrocardiogram foundation model. JAMIA Open 10 2025;8(5):ooaf122. ISSN 2574-2531.

[5] Fu Y, Bai W, Yi W, Manisty C, Bhuva AN, Treibel TA, Moon JC, Clarkson MJ, Davies RH, Hu Y. A versatile foundation model for cine cardiac magnetic resonance image analysis tasks. arXiv preprint arXiv250600679 2025;.

[6] Khosla P, Teterwak P, Wang C, Sarna A, Tian Y, Isola P, Maschinot A, Liu C, Krishnan D. Supervised contrastive learning. In Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds.), Advances in Neural Information Processing Systems. Curran Associates, Inc.; 18661–18673.

[7] Wu Y, Dong Z, Chen C, Zhou W, Zhou JH. Supremix: Supervised contrastive learning for medical imaging regression with mixup. Medical Image Analysis 2025;103909.

Address for correspondence:

Laura Alvarez-Florez

Department of Clinical and Experimental Cardiology, Amsterdam University Medical Center

Meibergdreef 9, 1105 AZ Amsterdam, NL

l.alvarezflorez@amsterdamumc.nl