Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

CardioJEPA: A Framework for Holistic Cardiac Representations Through Predictive Multimodal Alignment with ECG-Preservation

Daniel Uyterlinde

d.j.uijterlinde@amsterdamumc.nl

University of Amsterdam

Laura Alvarez-Florez

l.alvarezflorez@amsterdamumc.nl

Amsterdam University Medical Center

Samuel Ruipérez-Campillo

ETH Zurich

Lukas P. A. Arts

Amsterdam University Medical Center

Folkert W. Asselbergs

Amsterdam University Medical Centers

Erik J. Bekkers

University of Amsterdam

Ivana Išgum

Mayo Clinic

Fleur V. Y. Tjong

Amsterdam University Medical Center

Research Article

Keywords: Cardiac magnetic resonance imaging, cross-modal alignment, electrocardiogram, foundation models, joint-embedding predictive architecture

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11597188/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

CardioJEPA: A Framework for Holistic Cardiac Representations Through Predictive Multimodal Alignment with ECG-Preservation

Daniel Uyterlinde, Laura Alvarez-Florez, Samuel Ruipérez-Campillo, *Member, IEEE*, Lukas P. A. Arts, Folkert W. Asselbergs, Erik J. Bekkers, Ivana Išgum and Fleur V. Y. Tjong

***Abstract*— The electrocardiogram (ECG) is the most widely used cardiac diagnostic tool, but it reflects cardiac structure and function only indirectly. Aligning ECG and cardiac magnetic resonance (CMR) representations offers a way to transfer imaging-derived structural knowledge into the ECG encoder without compromising the electrophysiological information it already carries. Contrastive learning has become the predominant approach for ECG–CMR alignment, yet the quality of the learned representation is jointly determined by the CMR target representation, the alignment objective, and the preservation of electrophysiological information unique to the ECG. We propose CardioJEPA, a predictive alignment framework built on frozen cardiac foundation models that addresses all three: a target that separates static anatomy from cardiac motion, a predictive objective that maps ECG representations onto a fixed CMR target without instance discrimination, and a preservation term that retains nearly all pre-alignment electrophysiological information. On 63,193 paired ECG–CMR recordings from the UK Biobank, CardioJEPA outperforms previous ECG–CMR alignment methods across structural, electrical, and diagnostic tasks, retains 98.6% of the ECG-unique information present before alignment, substantially reduces the modality gap relative to matched contrastive alignment, and transfers to external cohorts for structural heart disease and Chagas cardiomyopathy without task-specific training, achieving state-of-the-art performance.**

***Index Terms*— Cardiac magnetic resonance imaging, cross-modal alignment, electrocardiogram, foundation models, joint-embedding predictive architecture.**

Manuscript submitted for review on October 29, 2026. This research has been conducted using the UK Biobank Resource under Application Number 24711. This work was supported by grants awarded to F.V.Y.T. by the Dutch Heart Foundation (Dekker 03-005-2025-0129), the Dutch Research Council (NWO Rubicon 452019308 and NWO AINed XS NGF.1609.243.045), Amsterdam Cardiovascular Sciences, and Health Holland, as well as by the Dutch Research Council (NWO) through the Vidi project SIGN (VI.Vidi.233.220), awarded to E.J.B.

D. Uyterlinde and E. J. Bekkers are with the Informatics Institute, University of Amsterdam, The Netherlands. (e-mail: d.j.uijterlinde@amsterdamumc.nl; e.j.bekkers@uva.nl).

L. Alvarez-Florez is with the Department of Biomedical Engineering and Physics, Amsterdam University Medical Center, The Netherlands, and with the Quantitative Healthcare Analysis Group, Informatics Institute, University of Amsterdam, The Netherlands. (e-mail: l.alvarezflorez@amsterdamumc.nl)

S. Ruipérez-Campillo is with the Department of Computer Science, ETH Zurich, Zurich, Switzerland. (e-mail: samuel.ruiperezcampillo@inf.ethz.ch)

L.P.A. Arts, F.W. Asselbergs, and F.V.Y. Tjong are with the Heart Center, Department of Clinical and Experimental Cardiology, Amsterdam University Medical Centers, Amsterdam Cardiovascular Sciences, University of Amsterdam, Amsterdam, The Netherlands. (e-mail: l.p.a.arts@amsterdamumc.nl, f.w.asselbergs@amsterdamumc.nl, f.v.tjong@amsterdamumc.nl)

I. Išgum is with the Department of Radiology, Mayo Clinic, Rochester, MN, USA, the Department of Biomedical Engineering and Physics, Amsterdam University Medical Center, The Netherlands, and with the Department of Radiology and Nuclear Medicine, Amsterdam University Medical Center, The Netherlands. (e-mail: Isgum.Ivana@mayo.edu)

D. Uyterlinde and L. Alvarez-Florez contributed equally to this work as co-first authors. I. Išgum and F. V. Y. Tjong supervised this work jointly.

Parts of this work appeared in the M.S. thesis of D. Uyterlinde [1].

I. Introduction

THE electrocardiogram (ECG) and cardiac magnetic resonance (CMR) imaging capture complementary aspects of cardiac health. The ECG records electrical activity over time, whereas CMR provides detailed anatomical and functional information. However, while ECG is inexpensive and routinely acquired, CMR is costly and requires specialized infrastructure and expertise [2], [3]. Consequently, detailed imaging-derived information is available for only a subset of patients during clinical decision-making. Because cardiac structure and function influence the electrical activity recorded at the body surface, some of this information may nevertheless be reflected in the ECG. This motivates learning ECG representations that incorporate anatomical and functional knowledge derived from CMR, enabling cardiac characteristics that ordinarily require imaging to be inferred from ECG alone.

Large-scale self-supervised pretraining has made this goal increasingly feasible by providing strong unimodal representations for both modalities. ECG foundation models pretrained with masked modeling and related objectives transfer well across downstream tasks and require little labeled data [4]–[6], and CMR encoders can likewise be learned without manual annotations [7], [8]. Paired ECG and CMR acquisitions therefore offer an opportunity to connect these pretrained representations and transfer imaging-derived cardiac information into the ECG latent space.

Early ECG–CMR approaches learned shared representations through cross-modal reconstruction [9], whereas more recent methods have predominantly adopted contrastive learning. ECCL [10] and MMCL [11] use CLIP-style symmetric InfoNCE objectives to align patient-matched ECG and CMR representations, PTACL [12] augments patient-level alignment

with local temporal contrastive learning, and Alvarez-Florez et al. [13] align ECGs with end-diastolic and end-systolic CMR representations to incorporate phase-specific structural information. These studies demonstrate that CMR supervision can enrich ECG representations with structural information. Despite their differences, however, they all build alignment around the information shared between the two modalities using InfoNCE-based objectives, with limited consideration of the modality-specific information that may be lost in the process.

This shared-information formulation is restrictive for two main reasons. First, contrastive instance discrimination treats recordings from different patients as negatives, even when their cardiac phenotypes may be similar, potentially separating physiologically related representations; debiased sampling and similarity-based soft weighting only partially mitigate this [14]–[16]. Second, maximizing the mutual-information bound drives the ECG representation toward a minimal sufficient statistic of the CMR representation [17], [18]. Because the two modalities provide complementary rather than equivalent measurements, this pressure can suppress clinically relevant electrophysiological information, such as rhythm and conduction characteristics, that has no direct counterpart in CMR [19]. Effective ECG–CMR alignment should therefore not only transfer imaging-derived information, but also preserve the information already encoded by the ECG.

Predictive latent learning offers an alternative to contrastive learning. JEPA [20], [21] performs prediction directly in representation space without relying on negative samples, and CrossJEPA [22] extended this principle to cross-modal knowledge transfer using a fixed target encoder. Because such predictive objectives involve no instance-level repulsion, the source representation is not forced to discard information beyond what is needed to predict the target. However, changing the alignment objective alone does not determine what information can be transferred or guarantee that source-specific information is preserved. Successful transfer therefore still depends on both the quality of the CMR target and the explicit preservation of ECG-specific information.

To address these considerations, we propose CardioJEPA, a predictive framework for ECG–CMR representation alignment built on pretrained cardiac foundation models. CardioJEPA makes three methodological contributions. First, we construct a CMR target that separates static anatomy from cardiac motion by applying principal component analysis independently to each component. This prevents dominant anatomical variation from overwhelming the functional signal and increases the functional information available for transfer from CMR to ECG. Second, we replace contrastive instance discrimination with predictive latent alignment against a frozen CMR target so that patients with similar phenotypes are no longer pushed apart. Third, we incorporate a dedicated preservation objective that anchors the ECG encoder to its pretrained representation, explicitly retaining the electrophysiological information encoded by the original model. Together, these components transfer imaging-derived structural and functional information into the ECG representation without sacrificing the electrophysiological information that gives the ECG its independent diagnostic value.

The remainder of the manuscript is structured as follows. Section II formalizes the problem and introduces CardioJEPA. In Section III, we evaluate the framework across structural, electrophysiological, and diagnostic tasks, investigating how CMR target construction, alignment strategy, and ECG-information preservation determine the quality of the resulting representation. We then compare CardioJEPA with previous ECG–CMR alignment approaches and assess its generalization to external cohorts for structural heart disease and Chagas cardiomyopathy detection. Finally, Sections IV and V discuss the implications and limitations of the findings and present the main conclusions.

II. Methodology

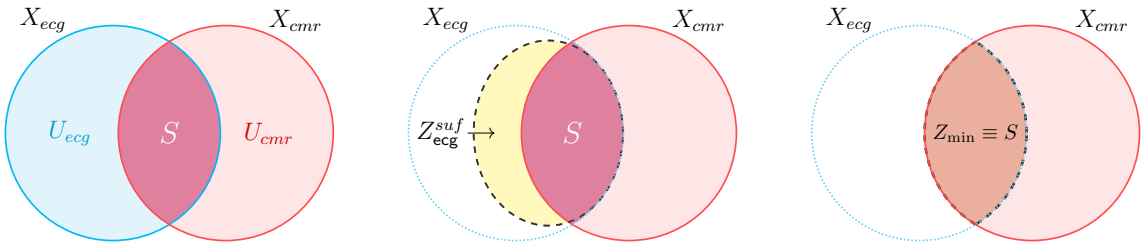
This section presents **CardioJEPA**, our framework for predictive ECG–CMR representation alignment. At deployment, only the ECG is available, so the quantity to maximize is the task-relevant information carried by the ECG representation alone. We first decompose this quantity to identify which of its components alignment can transfer and which it must preserve. This decomposition organizes the four components of the framework: pretrained unimodal encoders (Section II-A), a CMR target that isolates function from anatomy (Section II-B), a predictive alignment objective (Section II-C), and a distillation term that preserves ECG-unique information (Section II-D).

Let Z_{ecg} and Z_{cmr} denote the ECG and CMR encoder outputs and Y a set of downstream clinical phenotypes. Following [19], the task-relevant information decomposes as:

$$I(Z_{\text{ecg}}, Z_{\text{cmr}}; Y) = \underbrace{I(Z_{\text{ecg}}; Z_{\text{cmr}}; Y)}_{S\text{: Shared}} + \underbrace{I(Z_{\text{cmr}}; Y \mid Z_{\text{ecg}})}_{U_{\text{cmr}}\text{: CMR-unique}} + \underbrace{I(Z_{\text{ecg}}; Y \mid Z_{\text{cmr}})}_{U_{\text{ecg}}\text{: ECG-unique}}. \tag{1}$$

Here, $I(A;B)$ denotes mutual information, $I(A,B;C)$ the information the joint variable (A,B) carries about C, $I(A;B \mid C)$ conditional mutual information, and $I(A;B;C)$ the interaction information, i.e., the information that A and B share about C. Following Liang *et al.* [19], we identify this quantity as the shared component S. The three terms are nonnegative and partition the total task-relevant information by the chain rule (Fig. 1a).

By construction, U_{cmr} is inaccessible at deployment. Therefore, the objective reduces to maximizing $I(Z_{\text{ecg}}; Y) = S + U_{\text{ecg}}$. The shared component S comprises attributes observable in both modalities, such as ventricular volumes, ejection fraction, and myocardial mass; U_{ecg} captures electrophysiological characteristics absent from CMR, including conduction delays and rhythm abnormalities. An effective alignment framework must therefore maximize the shared information transferred through the CMR target, which upper-bounds what can be transferred at all (Section II-B), while preserving U_{ecg} during alignment (Sections II-C and II-D; Fig. 1a).



(a) Raw Information $I(X_{ecg}; X_{cmr}) = S$

(b) Sufficient Representation Contains at least all S

(c) Minimal Sufficient Contains only S (contrastive convergence)

Fig. 1. Information-theoretic view of ECG and CMR representations. (a) The two modalities share information $\boldsymbol{S}$, while each also carries modality-unique information ($\boldsymbol{U_{ecg}}$ and $\boldsymbol{U_{cmr}}$). (b) Predictive alignment targets a sufficient representation that contains at least $\boldsymbol{S}$ and leaves ECG-unique information unconstrained. (c) Contrastive learning encourages a minimal sufficient representation that contains only $\boldsymbol{S}$, discarding $\boldsymbol{U_{ecg}}$.

A. Pretrained Encoders and Data Preprocessing

We initialize the ECG and CMR encoders from the frozen unimodal foundational models ECG-FM [23] and CineMA [8], respectively. Crucially, both encoders natively encode substantial information regarding cardiac phenotypes prior to any cross-modal alignment.

ECG-FM is a CNN-Transformer hybrid pretrained on 1.5 million 12-lead ECGs using a combination of masked reconstruction and contrastive learning. Its robust baseline performance across downstream tasks makes it an optimal starting point, allowing our framework to focus entirely on structural knowledge transfer and electrophysiological preservation. During preprocessing, ECG signals are split into 5-second windows, with one window drawn uniformly at random per forward pass to serve as a temporal augmentation, and independently z-scored per lead. For the ECG encoder, all representations are extracted at layer 9 (either before or after the predictive CMR alignment). Through layer-wise linear probing, we established this as the optimal extraction point for maximizing both structural and electrical task performance.

CineMA is a multi-view masked autoencoder that jointly reconstructs masked frames across short- and long-axis views, yielding embeddings that capture 3D cardiac geometry without segmentation supervision. The publicly available CineMA weights were trained on the entire UK Biobank CMR cohort [24]. To avoid data leakage during downstream evaluation, we retrained the CineMA architecture from scratch on a strictly isolated 70/10/20 subject split (100 epochs, with a total of 5.32M images). We extract the end-diastolic (ED) and end-systolic (ES) frames using the pretrained CineMA segmentation model. We keep this encoder frozen throughout alignment. Predicting against a fixed target rules out the degenerate solution in which predictor and target encoder jointly collapse Z_{cmr} to a constant, a failure mode of joint-embedding objectives that do not rely on negative pairs [21]. Beyond this, large-scale self-supervised pre-training has already given the encoder a strong representation of cardiac anatomy and motion, and fine-tuning it on the alignment objective alone risks degrading representations that are already well suited to the task. Prior ECG–CMR alignment work adopts the same asymmetry [10], [13].

B. Creating a CMR Target Representation

The CMR target representations must capture both static anatomy and dynamic function. We find that CineMA ED and ES embeddings exhibit a mean cosine similarity exceeding 0.996 on the training split. Because the raw ED and ES embeddings are nearly identical, the alignment loss is largely explained by their shared anatomical component, leaving only a weak training signal for cardiac function. To address this imbalance, we construct an unsupervised, PCA-reduced target, which we call the *isolated-PCA* target, that explicitly separates function from anatomy to prevent static anatomical variance from overshadowing the subtle functional signal. We apply PCA independently to the ED embedding (static anatomy) and to the motion vector $\boldsymbol{\Delta} = \mathbf{z}_{\text{cmr}}^{\text{ES}} - \mathbf{z}_{\text{cmr}}^{\text{ED}}$ (function), z-score normalize each stream, and concatenate:

$$\mathbf{z}_{\text{cmr}} = \left[\, \text{PCA}_{256}(\mathbf{z}_{\text{cmr}}^{\text{ED}}) \,\middle\|\, \text{PCA}_{256}(\boldsymbol{\Delta}) \,\right] \in \mathbb{R}^{512}, \quad (2)$$

where $\|$ denotes concatenation and $\text{PCA}_{256}(\cdot)$ the projection onto the top 256 principal components, fitted on the training split. Projecting the two streams separately eliminates the anatomy–motion cross-covariance blocks, preventing anatomical variance from masking the functional signal. Under a Gaussian assumption, the retained variance increases monotonically with $I(Z_{\text{cmr}}^{\text{raw}}; Z_{\text{target}})$, so among all rank-$k$ linear projections, PCA is the information-maximizing choice, and the resulting target is the most information-rich linear compression of the CMR state that gives anatomy and function equal weight. This departs from the standard JEPA target, produced by an EMA teacher [20] or, in the cross-modal case, a frozen encoder of the second modality [22]. We retain the frozen-encoder asymmetry but add a fixed linear projection on top of it, a deliberate choice given the near-degeneracy of the raw ED and ES embeddings.

C. The Alignment Objective

CardioJEPA aligns the ECG encoder to the frozen CMR target (Section II-B) through a predictive rather than a contrastive objective. Deployment requires carrying U_{ecg} through alignment, but InfoNCE instead maximizes a lower bound on the mutual information between the two representations, $I(Z_{\text{ecg}}; Z_{\text{cmr}})$ [17]. This places a ceiling on the ECG-unique

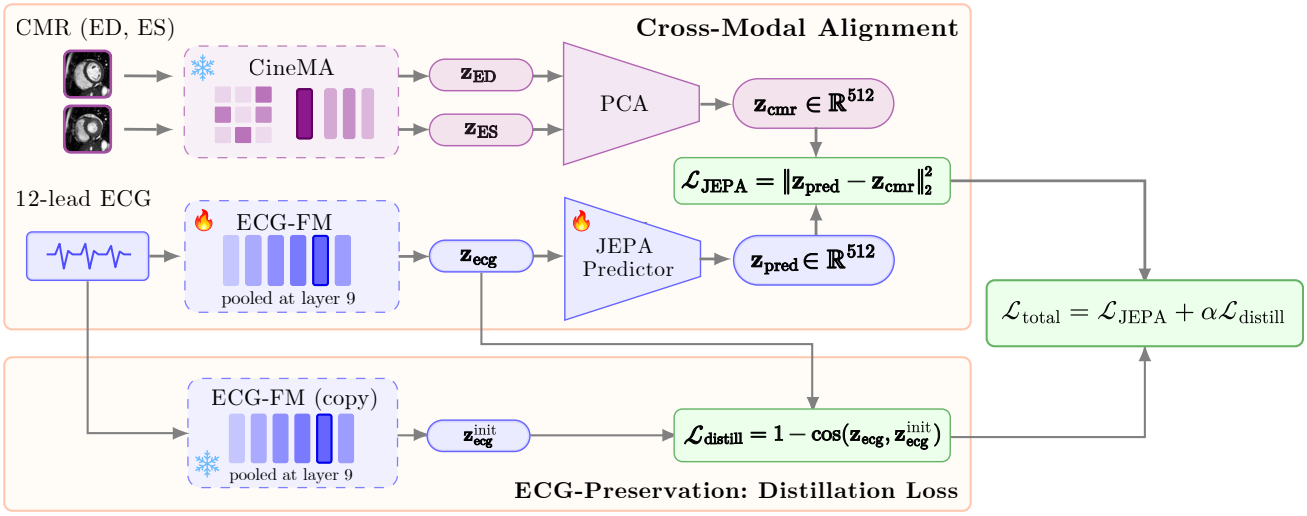


Fig. 2. **CardioJEPA architecture.** The framework aligns an ECG encoder to a frozen, PCA-reduced CMR target ($\mathbf{z}_{cmr}$) using a predictive mean-squared error loss ($\mathcal{L}_{JEPA}$). A distillation loss ($\mathcal{L}_{distill}$) anchors the trainable ECG representation ($\mathbf{z}_{ecg}$) to its frozen initialization ($\mathbf{z}_{ecg}^{init}$), preserving ECG-unique information.

information that can be retained:

$$U_{ecg} = I(Z_{ecg}; Y \mid Z_{cmr}) \le h(Z_{ecg} \mid Z_{cmr}) = h(Z_{ecg}) - I(Z_{ecg}; Z_{cmr}), \quad (3)$$

where $h(\cdot)$ denotes differential entropy, so $h(Z_{ecg} \mid Z_{cmr})$ quantifies the variability in the ECG representation that the CMR representation does not account for. Asymptotically, InfoNCE drives the encoder toward the minimal sufficient point at which $h(Z_{ecg})$ collapses to $I(Z_{ecg}; Z_{cmr})$, so $h(Z_{ecg} \mid Z_{cmr}) \to 0$ and U_{ecg} vanishes with it, leaving S as the only task-relevant information the representation retains (Fig. 1c) [18], [25]. A second limitation is geometric: the InfoNCE gradient weights each negative by its softmax similarity, so repulsion is strongest between the most phenotypically similar patients, precisely where similarity is clinically meaningful (Fig. 3).

Our predictive formulation is asymmetric, with gradients flowing only through the ECG branch such that the CMR manifold stays fixed. A linear predictor g_ϕ on top of the ECG encoder predicts the frozen CMR target, $\mathbf{z}_{pred} = g_\phi(\mathbf{z}_{ecg})$:

$$\mathcal{L}_{JEPA} = \frac{1}{B}\sum_{i=1}^{B} \left\|\mathbf{z}_{pred}^{(i)} - \mathbf{z}_{cmr}^{(i)}\right\|_2^2. \quad (4)$$

The predictor gradient $2g_\phi^\top(g_\phi \mathbf{z}_{ecg}^{(i)} - \mathbf{z}_{cmr}^{(i)})$ has no sum over patients, no softmax weights, and no repulsion: phenotypically similar patients exert zero repulsion, so the ECG cluster can mirror the geometry of the frozen target (Fig. 3). Because ECG-FM produces T=156 tokens, we aggregate by token-wise projection with a learned softmax weighting $w^{(t)}$, refining the definition to $\mathbf{z}_{pred} = \sum_t w^{(t)} g_\phi(\mathbf{z}_{ecg}^{(t)})$, which emphasizes the most CMR-informative segments. Fig. 2 gives an overview.

D. Preserving ECG-Unique Information

With finite encoder capacity, increasing S appears to come at the expense of U_{ecg}, but only if the conditional entropy

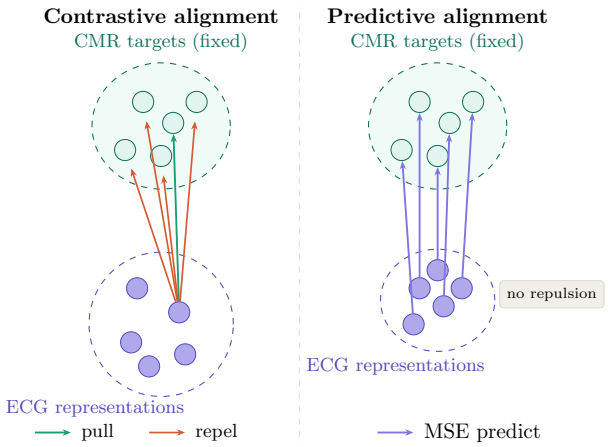


Fig. 3. **Contrastive vs. predictive alignment against a frozen target.** Contrastive alignment (left) introduces batch-wide repulsion that scatters ECG representations. Predictive alignment (right) regresses each ECG representation toward its paired CMR target without repulsion, letting the ECG cluster mirror the geometry of the frozen CMR space.

$h(Z_{ecg} \mid Z_{cmr})$ is left unconstrained. Contrastive alignment drives $h(Z_{ecg} \mid Z_{cmr})$ toward zero (Section II-C); predictive alignment does not penalize variance orthogonal to the target, but offers no mechanism to retain it either (Fig. 1b). Because the ECG foundation model captures substantial electrophysiological information before alignment, $h(Z_{ecg} \mid Z_{cmr})$ starts from a non-trivial baseline: the challenge is not to learn more ECG-unique information, but to prevent its suppression as the encoder specializes toward Z_{cmr}. This motivates maximizing

$$\mathcal{J}(\theta) = I(Z_{ecg}; Z_{cmr}) + \alpha\, h(Z_{ecg} \mid Z_{cmr}), \quad (5)$$

which decouples shared from unique information: α preserves pretrained electrophysiological variance rather than requiring new capacity.

We lower-bound the second term by anchoring the encoder to its pretrained representation, instantiating it as a *distillation loss* that CardioJEPA pairs with the predictive objective, giving

$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{JEPA}} + \alpha\,\mathcal{L}_{\text{distill}}$. It acts as a cosine penalty anchoring the active encoder to its pretrained embedding:

$$\mathcal{L}_{\text{distill}} = \frac{1}{B}\sum_{i=1}^{B}\big(1 - \cos(\mathbf{z}_{\text{ecg}}^{(i)}, \mathbf{z}_{\text{ecg}}^{\text{init},(i)})\big), \tag{6}$$

where $\mathbf{z}_{\text{ecg}}^{\text{init}}$ is the embedding produced by the frozen copy of the pretrained ECG encoder. The loss penalizes angular deviation only, so the encoder may rescale freely while its angular orientation is held close to the pretrained embedding.

III. Experiments & Results

We evaluate CardioJEPA by progressively validating the design principles introduced in Section II. We first establish the dataset and evaluation metrics (Sections III-A and III-B), then assess each design component independently: the CMR target representation (Section III-C), the alignment and preservation objectives (Section III-D), and the resulting latent space geometry (Section III-E). Lastly, we compare CardioJEPA with prior ECG–CMR alignment methods (Section III-F) and evaluate its out-of-distribution generalization (Section III-G).

A. Dataset

Our experimental cohort comprises 63,193 matched ECG–CMR pairs from the UK Biobank, of which 34,044 have available clinical phenotype annotations (Table I). All participants provided written informed consent as described in the original cohort publications [24]. From the available UK Biobank annotations, we selected five CMR-derived structural phenotypes: left ventricular (LV) ejection fraction (LVEF), left ventricular end-diastolic volume (LVEDV), left ventricular mass (LVM), right ventricular end-diastolic volume (RVEDV), and left ventricular global longitudinal strain (LVGLS). We additionally selected four ECG-derived phenotypes: PR interval, QRS duration, corrected QT interval (QTc), and P-wave duration, together with four clinical diagnoses: atrial fibrillation (AF), bundle branch block (BBB), cardiomyopathy (CM), and myocardial infarction (MI).

TABLE I
UK Biobank dataset splits.

Split	N	Purpose
Training	47,683	Alignment; probe fit on labeled subset
Validation	5,299	Alignment loss / early stopping
LP-validation	3,404	In-training probe evaluation
LP-test	6,807	Held-out test
Total	63,193	

B. Evaluation Metrics

To quantify both structural transfer and preservation of ECG-unique information, we probe frozen representations using Ridge regression for continuous phenotypes (R^2), and logistic regression for binary outcomes (AUROC), reported on the held-out test split, LP-test (Table I).

To summarize the trade-off between structural transfer and electrophysiological preservation, we define two composite metrics. $\tilde{S}$ represents the shared information and is computed as the mean R^2 across three phenotypes accessible to both modalities: LVEDV, RVEDV, and LV mass. $\tilde{U}_{\text{ecg}}$ measures ECG-unique information using four electrophysiological phenotypes: PR interval, QRS duration, QTc interval, and P-wave duration. To isolate electrophysiological information beyond cardiac structure, we first remove the component that can be linearly explained by five structural covariates (LVEF, LVEDV, LV mass, RVEDV, and LVGLS) using ordinary least squares (OLS) residualization. The regression coefficients are estimated on the training split and applied to the test split. OLS residuals are orthogonal to these covariates by construction [26], so any remaining decodable variance is inaccessible from the five structural summaries used. A positive residual R^2 therefore provides direct evidence of linearly accessible electrophysiological information beyond cardiac structure.

C. The Importance of the CMR Alignment Target

To assess the structural and functional information encoded by different CMR target representations, we evaluate how well each target predicts clinically relevant cardiac phenotypes via linear probing. We compare six target representations: a static end-diastolic (ED) frame, a static end-systolic (ES) frame, their difference ($\mathbf{\Delta}$) as a motion representation, their concatenation, a supervised representation learned to predict clinical variables, and the proposed isolated-PCA target. The supervised target is constructed by taking the intermediate representations of an MLP, trained to regress/classify 7 cardiac variables on top of the CMR encoder: LVEF, LVM, LVEDV, RVEDV, LVGLS, AF and MI.

The results in Table II show that static anatomy and cardiac motion provide complementary information. The ED embedding strongly represents ventricular volumes and mass but captures cardiac function poorly, whereas the motion representation ($\mathbf{\Delta}$) substantially improves LVEF and LVGLS prediction. Concatenating the raw CineMA embeddings provides the strongest overall balance across anatomical and functional phenotypes, but its high dimensionality ($d = 1536$) makes it impractical as an alignment target. The proposed *isolated-PCA* target separately compresses the anatomy and motion streams into a compact $d = 512$ representation, matching the supervised target on anatomical phenotypes while providing stronger functional information. Despite being unsupervised, it matches the supervised alternative on anatomy and exceeds it on function (LVEF $R^2 = 0.663$ versus 0.628).

TABLE II
Linear probing R^2 of CMR target representations.

Representation	Dim.	LVEDV	RVEDV	LVM	LVEF	LVGLS
Raw CineMA embeddings						
ED	768	0.954	0.939	0.965	0.247	0.322
ES	768	0.900	0.895	0.958	0.507	0.392
$\mathbf{\Delta}$ (ES−ED)	768	0.932	0.924	0.925	0.651	0.551
Concat (ED, ES)	1536	**0.965**	**0.953**	**0.973**	**0.691**	**0.606**
Constructed targets (ours)						
Supervised	256	0.954	0.942	0.964	0.628	0.565
Isolated PCA	512	0.954	0.939	0.960	0.663	0.570

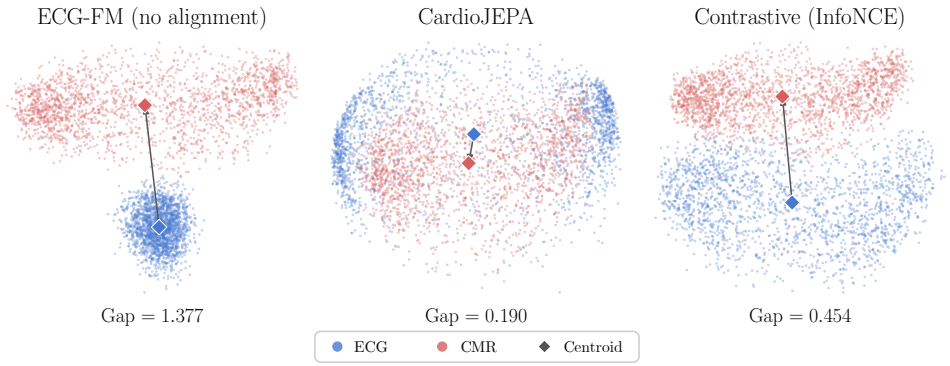


Fig. 4. **Modality gap in the aligned latent space (supervised target).** First two principal components of ℓ_2-normalized ECG and CMR embeddings on the held-out test set; the (modality) gap is the ℓ_2 distance between centroids. CardioJEPA achieves clear overlap.

D. The Role of Alignment and Preservation Objectives in Retaining ECG-Unique Information

To isolate the contribution of the alignment objective itself, we compare CardioJEPA against a matched contrastive baseline that differs only in the alignment loss. Specifically, we replace the predictive JEPA objective with an asymmetric InfoNCE objective [10], [11], while keeping the frozen CMR target, ECG encoder, and optimization protocol identical. For the InfoNCE loss, the embeddings are ℓ_2-normalized, and the temperature is set to 0.1.

First, we disable the preservation term ($\alpha = 0$) to isolate the effect of the alignment objective from the explicit preservation objective. The results show that both paradigms transfer an equivalent amount of shared structural cardiac information ($\tilde{S} = 0.636 \pm 0.002$ for both). In contrast, CardioJEPA retains approximately 2.8% more ECG-unique information than contrastive learning ($\tilde{U}_{\text{ecg}} = 0.660 \pm 0.008$ versus 0.642 ± 0.003; Table III). These findings indicate that predictive alignment preserves ECG-unique information even without an explicit preservation objective, whereas contrastive alignment suppresses it (Section II-C).

We next quantify the cost of alignment, defined as the reduction in ECG-unique information ($\tilde{U}_{\text{ecg}}$) relative to the original unaligned ECG representation. The ECG-FM baseline provides this upper bound, attaining $\tilde{U}_{\text{ecg}} = 0.709$ (Table IV). Without explicit preservation, alignment reduces $\tilde{U}_{\text{ecg}}$ to 0.660 ($\Delta = -0.049$) for CardioJEPA and 0.642 ($\Delta = -0.067$) for contrastive learning, indicating that predictive alignment sacrifices less ECG-unique information than its contrastive counterpart. Introducing the preservation objective recovers nearly all of this loss: CardioJEPA with distillation reaches 0.699 ($\Delta = -0.010$), while contrastive learning with distillation reaches 0.689 ($\Delta = -0.020$). These results show that explicit preservation largely eliminates the cost of alignment, with predictive alignment remaining consistently closer to the unaligned ECG representation.

E. Latent Space Geometry

To evaluate how the alignment objective shapes the geometry of the shared latent space, we compare the aligned ECG and CMR embeddings produced by the predictive and contrastive alignment objectives (Fig. 4). Following Liang et al. [27], we measure the modality gap as the ℓ_2 distance between the centroids of ℓ_2-normalized aligned ECG and CMR embeddings. We report this analysis on the supervised target space.

TABLE III
PARADIGM COMPARISON AT $\boldsymbol{\alpha = 0}$ (NO PRESERVATION TERM). $\tilde{\boldsymbol{S}}$: STRUCTURAL COMPOSITE $\boldsymbol{R^2}$. $\tilde{\boldsymbol{U}}_{\mathbf{ecg}}$: ECG-UNIQUE COMPOSITE $\boldsymbol{R^2}$. MEAN $\pm$ STD. OVER THREE SEEDS.

Method	$\tilde{\boldsymbol{S}}$	$\tilde{\boldsymbol{U}}_{\mathbf{ecg}}$
Predictive (CardioJEPA)	0.636 ± 0.002	$\mathbf{0.660 \pm 0.008}$
Contrastive	0.636 ± 0.002	0.642 ± 0.003

TABLE IV
ALIGNMENT COST: $\tilde{\boldsymbol{U}}_{\mathbf{ecg}}$ DEVIATION FROM THE UNALIGNED ECG-FM BASELINE, BY METHOD AND PRESERVATION SETTING.

Method	$\tilde{\boldsymbol{U}}_{\mathbf{ecg}}$ (R^2)	**Δ from init**
ECG-FM [23]	0.709	—
CardioJEPA + $\mathcal{L}_{\text{distill}}$	0.699	-0.010
Contrastive Learning + $\mathcal{L}_{\text{distill}}$	0.689	-0.020
CardioJEPA ($\alpha = 0$)	0.660	-0.049
Contrastive Learning ($\alpha = 0$)	0.642	-0.067

The unaligned ECG-FM baseline exhibits the largest modality gap (1.377). After alignment, CardioJEPA reduces this gap to 0.190, producing a broad, overlapping latent manifold, whereas contrastive learning retains a substantially larger separation (0.454) by concentrating the two modalities into distinct cones. This is consistent with our theoretical analysis: the predictive objective attracts ECG representations toward their CMR targets without the repulsive forces inherent to contrastive learning, so the ECG manifold can mirror the geometry of the frozen target rather than being compressed into a cone. We report this as a descriptive property of the aligned latent space; whether a smaller modality gap has downstream benefit is not tested here.

F. Comparison with Prior Work

To position our work against existing ECG–CMR alignment methods, we compare our approach with the MMCL [11], PTACL [12], and ECCL [10] frameworks. The ablations in Sections III-D and III-E identify CardioJEPA with distillation as the best-performing configuration, which we use for all subsequent experiments. We compare against the released

TABLE V
LINEAR-PROBING PERFORMANCE (R^2 FOR REGRESSION, AUROC FOR CLASSIFICATION) ON STRUCTURAL, ELECTRICAL, AND DIAGNOSTIC CARDIAC PHENOTYPES, POSITIONED AGAINST PRIOR ECG–CMR ALIGNMENT METHODS.

		Structural (R^2)				**Electrical (R^2)**				**Diagnosis (AUROC)**				**Composites (R^2)**	
Input	**Method**	**LVEF**	**LVM**	**LVEDV**	**RVEDV**	**PR**	**QRS**	**QTc**	**P-wave**	**AFib**	**BBB**	**CM**	**MI**	$\tilde{S}$	$\tilde{U}_{\text{ecg}}$
ECG	*ECG-FM [23]*	*0.286*	*0.649*	*0.554*	*0.559*	*0.912*	*0.796*	*0.703*	*0.506*	*0.724*	*0.826*	*0.795*	*0.712*	*0.587*	*0.709*
ECG	ECCL† [10]	0.176	0.449	0.372	0.397	–	–	–	–	–	–	–	–	0.406	–
	MMCL‡ [11]	0.139	0.401	0.342	0.363	0.494	0.640	0.488	0.252	0.683	0.767	0.629	0.689	0.369	0.474
	PTACL [12]	0.257	0.613	0.505	0.531	0.872	0.770	0.657	0.406	0.719	**0.808**	0.772	0.711	0.550	0.624
	CardioJEPA	**0.323**	**0.702**	**0.604**	**0.606**	**0.908**	**0.791**	**0.699**	**0.493**	**0.741**	0.807	**0.790**	**0.742**	**0.637**	**0.699**
CMR	*CMR Target*	*0.661*	*0.960*	*0.954*	*0.939*	*0.528*	*0.235*	*0.167*	*0.028*	*0.773*	*0.796*	*0.765*	*0.786*	*0.951*	*0.182*

† ECCL released no checkpoint and uses full fine-tuning; scores are quoted from the original publication.
‡ MMCL's original publication reports fine-tuning with attention pooling; the released checkpoint is evaluated here under our linear-probing protocol.

MMCL and PTACL checkpoints under our linear-probing protocol on the held-out split, while reporting ECCL from the original publication because no public checkpoint is available (full fine-tuning, marked †).

CardioJEPA achieves the strongest overall performance among ECG–CMR alignment methods across structural and electrical phenotypes (Table V). Notably, the unaligned ECG-FM initialization already exceeds all previously published alignment methods on the four structural phenotypes. CardioJEPA further increases the structural composite to $\tilde{S} = 0.637$, while retaining an ECG-unique composite close to the pre-alignment baseline ($\tilde{U}_{\text{ecg}} = 0.699$ versus 0.709). Among aligned methods, CardioJEPA preserves the largest amount of the original ECG representation, remaining closest to ECG-FM across all four electrical phenotypes. It also achieves the highest AUROC among aligned methods for atrial fibrillation, cardiomyopathy, and myocardial infarction, while the unaligned ECG-FM remains strongest for bundle branch block. These results demonstrate that CMR-derived structural information can be incorporated into the ECG representation without substantially sacrificing its original electrophysiological content.

G. Out-of-Domain External Validation

To assess whether the learned representations generalize beyond the UK Biobank training distribution, we evaluate CardioJEPA on two external tasks: structural heart disease detection in EchoNext [28] and Chagas disease detection across independent Latin American cohorts.

1) Structural heart disease detection (EchoNext): To evaluate whether the structural cardiac information learned through ECG–CMR alignment transfers across datasets and patient populations, we compare CardioJEPA against the EchoNext models [29], [30] on the EchoNext dataset [28]. EchoNext and EchoNext-mini are state-of-the-art models for detection of structural heart disease (SHD) using ECG waveform, demographics, and ECG-derived tabular features. Following the EchoNext benchmark, we evaluate performance on the primary SHD composite endpoint, which is positive if any of 11 individual structural abnormalities are present. Both EchoNext models are trained end-to-end for this task, whereas our encoder is evaluated in a frozen-representation setting, where only a linear probe is trained on top of the aligned ECG embeddings without any SHD-specific optimization. Despite this asymmetry, CardioJEPA exceeds both end-to-end baselines (Table VI). Relative to EchoNext-mini, AUROC improves from 0.820 to 0.850 and AUPRC from 0.789 to 0.823, with non-overlapping confidence intervals in both cases. The margin also holds against the full EchoNext model, trained on 1.2 million private multicentre pairs, which reaches 0.831 AUROC and 0.750 AUPRC on this split: CardioJEPA improves on it by $+0.019$ AUROC and $+0.073$ AUPRC. CardioJEPA further exceeds EchoNext-mini on 8 of the 11 individual abnormalities, with the largest margin on reduced ejection fraction (LVEF $\leq 45\%$, 0.912 versus 0.852 AUROC). The learned ECG representation therefore captures a broad spectrum of structural cardiac information rather than a single dominant phenotype.

TABLE VI
STRUCTURAL HEART DISEASE DETECTION (ECHONEXT [28]).

Method	**AUROC**	**AUPRC**
EchoNext† [29]	0.831	0.750
EchoNext-mini [30]	0.820 [0.809,0.831]	0.789 [0.772,0.804]
CardioJEPA (ours)	**0.850** [0.839,0.861]	**0.823** [0.809,0.839]

† Full EchoNext model, trained on 1.2M private multicentre pairs, evaluated on this public test split (results quoted from [29]).

2) Chagas disease detection: Chagas disease is a parasitic cardiomyopathy whose early, treatable stage is frequently missed. Detecting it from the 12-lead ECG alone was the task of the 2025 George B. Moody PhysioNet Challenge [31], and was first approached with deep learning by Jidling *et al.* [32], who trained a ResNet ensemble on the CODE and SaMi-Trop cohorts. Chagas cardiomyopathy combines structural myocardial remodeling with conduction and rhythm abnormalities, making it an ideal setting to evaluate both components of the learned representation. Successful transfer requires both the structural cardiac information acquired through ECG–CMR alignment and the electrophysiological information preserved during alignment. We evaluate the frozen representation using five-fold cross validation under two published protocols (Table VII), applying the encoder fully out of distribution: it

TABLE VII
FROZEN-ENCODER TRANSFER TO CHAGAS DISEASE CLASSIFICATION.

Method	AUROC	Top5%-TPR
CODE-15 + SaMi-Trop + PTB-XL (2025 PhysioNet Challenge split)		
Van Santvliet *et al*. [33] (frozen)†	–	0.381 ± 0.003
ECG-FM (init)	0.838 ± 0.006	0.386 ± 0.015
CardioJEPA (ours)	$\mathbf{0.859} \pm 0.005$	$\mathbf{0.432} \pm 0.008$
CODE-15 + SaMi-Trop (Jidling et al. split)		
Jidling *et al*. [32] (end-to-end)‡	0.80	–
ECG-FM (init)	0.831 ± 0.008	0.378 ± 0.016
CardioJEPA (ours)	$\mathbf{0.852} \pm 0.005$	$\mathbf{0.425} \pm 0.012$

† Frozen-backbone ablation of the challenge winners, matched to our frozen setting. Their fine-tuned model reached 0.490 top-5% TPR (0.867 AUROC).
‡ In-distribution validation AUROC; their CNN was trained end-to-end on CODE and SaMi-Trop, whereas our encoder is applied out-of-distribution.

has never seen these Latin American cohorts or any Chagas label during pre-training. Under the Challenge protocol, the unaligned ECG-FM initialization already matches the frozen configuration of the challenge-winning system [33] (top-5% TPR 0.386 versus 0.381), confirming that large-scale ECG pretraining alone captures much of the relevant signal. CardioJEPA improves on this foundation rather than simply inheriting it, increasing top-5% TPR to 0.432 and AUROC to 0.859; the same pattern holds under the Jidling *et al*. protocol, where our models reach an AUROC of 0.852, above the 0.80 reported for an end-to-end CNN trained *in distribution* on these cohorts. These results show that an imaging-based, disease-agnostic ECG representation can detect an unseen disease spanning both structural and electrophysiological abnormalities, providing direct evidence that cross-modal alignment does not require sacrificing modality-unique information.

IV. DISCUSSION

This study introduced CardioJEPA, a framework for ECG–CMR alignment that combines an informative CMR target representation with a predictive objective and explicit preservation of ECG-unique information. The resulting cardiac representation consistently outperformed previous ECG–CMR alignment methods across structural, electrical, and diagnostic tasks. Notably, the unaligned ECG foundation model already exceeded all of them on structural phenotype regression (Table V), highlighting the importance of large-scale unimodal pre-training. This observation is consistent with the recent shift toward large-scale self-supervised foundation models [4]–[6]. Building on this initialization, CardioJEPA further improved structural performance while preserving the electrophysiological information that gives the ECG its unique clinical value, indicating that multimodal alignment contributes cardiac information beyond what ECG pre-training alone provides.

The central finding of this work is that incorporating structural information from CMR does not require sacrificing ECG-specific information. As discussed in Section II-C, contrastive objectives are theoretically expected to compress modality-specific information. The empirical margin is nonetheless modest, because the bound in Eq. (3) describes the asymptotic limit of InfoNCE: alignment against a frozen target converges before the ECG encoder reaches minimal sufficiency, so the suppressive effect appears as a small difference rather than the collapse the theory describes in the limit. Our experiments (Table III) show that, when aligning to a frozen CMR target, both predictive and contrastive objectives transfer a similar amount of structural information. Their primary difference lies in how well they preserve ECG-specific information: predictive alignment consistently retains more, and adding the distillation objective lifts both paradigms close to the pre-alignment ceiling. The resulting representation approaches the original ECG foundation model on electrophysiological phenotypes while simultaneously improving structural prediction, demonstrating that preserving ECG-specific information is not incompatible with learning imaging-derived structural representations. These findings support the theoretical motivation for predictive latent alignment [21], [22] in complementary modalities, while also emphasizing that explicit preservation is an essential component of successful ECG–CMR alignment.

Our experiments further demonstrate that the CMR target representation is equally important as it determines the structural information available for transfer. As motivated in Section II, alignment cannot recover information absent from the target representation. The ablations in Table II show this directly. A representation derived from a single end-diastolic frame accurately captures ventricular anatomy but poorly represents cardiac function, whereas separating static anatomy from cardiac motion substantially improves functional phenotypes. The proposed isolated-PCA target preserves this complementary information in a compact latent representation and, despite requiring no supervision, matches the supervised alternative on anatomical phenotypes while exceeding it on cardiac function. Overall, this representation provides a compact but efficient solution to maximize the information that is available for transfer.

Unlike previous ECG–CMR alignment methods, which were evaluated only in-distribution [10]–[12], we showed that CardioJEPA generalized across external datasets, downstream tasks, and disease populations. The magnitude of improvement, however, depended on the nature of the target disease. Purely electrophysiological conditions, such as bundle branch block, derived little benefit from imaging supervision because the relevant information is already contained within the ECG (Table V). In contrast, diseases characterized by structural remodeling consistently benefited from multimodal alignment. Most notably, despite relying only on a frozen representation and a linear probe, CardioJEPA outperformed the end-to-end EchoNext-mini model [30] for structural heart disease detection (Table VI). It likewise improved ECG-based detection of Chagas cardiomyopathy relative to the frozen configuration of the challenge-winning model [33] (Table VII), without disease-specific fine-tuning. The finding that the same frozen representation supports both an echo-derived structural endpoint and a parasitic cardiomyopathy, in cohorts and populations it was never trained on, suggests it encodes a broad cardiac state rather than any single phenotype.

Despite the overall improvements in structural prediction, regression performance remained lower for LVEF than for other phenotypes such as ventricular volumes and mass. Importantly, this difference was already apparent in the CMR

target representation, which achieved an R^2 of 0.661 for LVEF compared with $R^2 > 0.93$ for the anatomical phenotypes. This suggests that the lower ECG performance is not solely a limitation of cross-modal transfer, but also reflects how functional information is represented in the CMR target space. While the proposed representation incorporates differences between end-diastolic and end-systolic states, LVEF depends on dynamic changes throughout the cardiac cycle that may not be fully captured by this representation. Future work could therefore investigate richer temporal CMR representations that explicitly model motion across the complete cardiac cycle, potentially providing a more informative functional target for ECG–CMR alignment.

Several limitations should be acknowledged. First, our evaluation relies on frozen representations and linear probing. Electrophysiological information recoverable only through nonlinear adaptation is therefore not captured, and the representational capacity of CardioJEPA may be underestimated. Second, the framework is limited to cine CMR and can only transfer information present in that modality; extending the target to sequences that capture tissue characteristics, such as late gadolinium enhancement or parametric mapping, may further enrich the resulting ECG representation. Finally, although CardioJEPA generalized across multiple external datasets and clinical tasks, prospective multicenter validation is required to determine whether these improvements directly translate into earlier diagnosis in clinical practice [34], [35].

V. Conclusion

We presented CardioJEPA, a predictive ECG–CMR alignment framework built on frozen cardiac foundation models. Comparing target construction, alignment, and preservation under one protocol shows that the CMR target and the preservation term govern downstream performance, while the two alignment objectives, under an identical target and encoder, differ mainly in how much ECG-unique information they retain, and in the geometry of the resulting latent space. CardioJEPA outperforms published ECG–CMR alignment methods, and the predictive objective yields a more coherent shared latent space than a matched contrastive control. The resulting frozen representations detect structural heart disease and Chagas cardiomyopathy in external cohorts without task-specific training, suggesting that ECG models carrying imaging-derived structure could extend cardiac assessment where CMR is unavailable.

References

[1] D. Uyterlinde, "Beyond Contrastive Learning: Preserving Modality-Unique Information Through Predictive ECG–CMR Alignment," M.S. thesis, Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands, 2026. [Online]. Available: https://scripties.uba.uva.nl/search?id=record_58520

[2] D. J. Pennell, "Cardiovascular magnetic resonance," *Circulation*, vol. 121, no. 5, pp. 692–705, 2010.

[3] P. Kligfield *et al.*, "Recommendations for the standardization and interpretation of the electrocardiogram," *Circulation*, vol. 115, no. 10, pp. 1306–1324, 2007.

[4] A. Vaid *et al.*, "A foundational vision transformer improves diagnostic performance for electrocardiograms," *NPJ digital medicine*, vol. 6, no. 1, p. 108, 2023.

[5] E. Coppola, M. Savardi, M. Massussi, M. Adamo, M. Metra, and A. Signoroni, "HuBERT-ECG as a self-supervised foundation model for broad and scalable cardiac applications," *medRxiv*, 2024.

[6] J. Li *et al.*, "An electrocardiogram foundation model built on over 10 million recordings," *NEJM AI*, vol. 2, no. 7, p. AIoa2401033, 2025.

[7] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Contrastive learning of global and local features for medical image segmentation with limited annotations," *Advances in neural information processing systems*, vol. 33, pp. 12 546–12 558, 2020.

[8] Y. Fu *et al.*, "A versatile foundation model for cine cardiac magnetic resonance image analysis tasks," *arXiv preprint arXiv:2506.00679*, 2025.

[9] A. Radhakrishnan *et al.*, "Cross-modal autoencoder framework learns holistic representations of cardiovascular state," *Nature Communications*, vol. 14, no. 1, p. 2436, 2023.

[10] Z. Ding *et al.*, "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, pp. 109–119.

[11] Ö. Turgut *et al.*, "Unlocking the diagnostic potential of electrocardiograms through information transfer from cardiac magnetic resonance imaging," *Medical Image Analysis*, vol. 101, p. 103451, 2025.

[12] A. Selivanov, P. Müller, Ö. Turgut, N. Stolt-Ansó, and D. Rueckert, "Global and Local Contrastive Learning for Joint Representations from Cardiac MRI and ECG," in *Medical Image Computing and Computer Assisted Intervention, MICCAI 2025 - 28th International Conference, 2025, Proceedings*, 2026, pp. 217–227.

[13] L. Alvarez-Florez *et al.*, "Dual-phase cross-modal contrastive learning for cmr-guided ecg representations for cardiovascular disease assessment," in *SPIE Medical Imaging 2026: Image Processing*, 2026.

[14] C.-Y. Chuang, J. Robinson, Y.-C. Lin, A. Torralba, and S. Jegelka, "Debiased Contrastive Learning," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 8765–8775.

[15] P. Khosla *et al.*, "Supervised Contrastive Learning," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 18 661–18 673.

[16] Z. Ding *et al.*, "Generating Cardiac Magnetic Resonance Images from Electrocardiograms — A Multicenter Study," *NEJM AI*, vol. 3, no. 4, p. AIoa2500549, 2026.

[17] A. van den Oord, Y. Li, and O. Vinyals, "Representation Learning with Contrastive Predictive Coding," *arXiv preprint arXiv:1807.03748*, 2018.

[18] H. Wang, X. Guo, Z. Deng, and Y. Lu, "Rethinking Minimal Sufficient Representation in Contrastive Learning," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2022, pp. 16 020–16 029.

[19] P. P. Liang, Z. Deng, M. Q. Ma, J. Zou, L.-P. Morency, and R. Salakhutdinov, "Factorized Contrastive Learning: Going Beyond Multi-view Redundancy," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 32 971–32 998.

[20] Y. LeCun, "A Path Towards Autonomous Machine Intelligence," *OpenReview*, 2022.

[21] M. Assran *et al.*, "Self-Supervised Learning From Images With a Joint-Embedding Predictive Architecture," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 15 619–15 629.

[22] A. Perera *et al.*, "CrossJEPA: Cross-Modal Joint-Embedding Predictive Architecture for Efficient 3D Representation Learning from 2D Images," *arXiv preprint arXiv:2511.18424*, 2025.

[23] K. McKeen, S. Masood, A. Toma, B. Rubin, and B. Wang, "ECG-FM: an open electrocardiogram foundation model," *JAMIA Open*, vol. 8, no. 5, p. ooaf122, 10 2025.

[24] C. Sudlow *et al.*, "UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age," *PLOS Medicine*, vol. 12, no. 3, pp. 1–10, 03 2015.

[25] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola, "What Makes for Good Views for Contrastive Learning?" in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 6827–6839.

[26] R. Frisch and F. V. Waugh, "Partial Time Regressions as Compared with Individual Trends," *Econometrica*, vol. 1, no. 4, pp. 387–401, 1933.

[27] V. W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Zou, "Mind

the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 17 612–17 625.

[28] P. Elias and J. Finer, "EchoNext: A Dataset for Detecting Echocardiogram-Confirmed Structural Heart Disease from ECGs," *PhysioNet*, Apr. 2026, version 1.1.1. [Online]. Available: https://doi.org/10.13026/7cfw-d091

[29] T. J. Poterucha *et al.*, "Detecting structural heart disease from electrocardiograms using AI," *Nature*, vol. 644, no. 8075, pp. 221–230, 2025.

[30] J. W. Hughes *et al.*, "EchoNext-mini: A dataset and baseline AI model for detecting structural heart disease from electrocardiograms," *NEJM AI*, vol. 3, no. 5, p. AIdbp2500516, 2026.

[31] M. A. Reyna *et al.*, "Detection of Chagas Disease from the ECG: The George B. Moody PhysioNet Challenge 2025," in *Computing in Cardiology 2025*, vol. 52, 2025, pp. 1–4.

[32] C. Jidling *et al.*, "Screening for Chagas disease from the electrocardiogram using a deep neural network," *PLoS neglected tropical diseases*, vol. 17, no. 7, p. e0011118, 2023.

[33] L. Van Santvliet, P. X. Nguyen, B. Vandenberk, and M. De Vos, "Detecting Chagas disease using a vision transformer-based ECG foundation model," *Proceedings of CinC 2025*, vol. 52, 2025.

[34] E. H. Kemper *et al.*, "ESR Essentials: how to get to valuable radiology AI: the role of early health technology assessment—practice recommendations by the European Society of Medical Imaging Informatics," *European radiology*, vol. 35, no. 6, pp. 3432–3441, 2025.

[35] O. A. Ramwala *et al.*, "Establishing a validation infrastructure for imaging-based artificial intelligence algorithms before clinical implementation," *Journal of the American College of Radiology*, vol. 21, no. 10, pp. 1569–1574, 2024.