REPS: LEARNING REPRESENTATIONS FROM KNOWN RESPONSE LAWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing augmentation-based self-supervised learning uses transformations to create new views, but often leaves their predictable effects on measurable signal properties unused. Our key insight is that when these effects follow known response laws, augmentation can generate not only new views but also new supervision: an existing measurement can be analytically transformed into the expected measurement of each augmented view. We introduce RePs (Response-law Privileged Supervision), which learns measurement-relevant representations by predicting these targets from augmented signals alone, without additional annotation. We test RePs across three distinct response structures and signal domains: temporal scaling in physiological signals, pitch translation in music, and spatial rotation in multichannel audio. RePs outperforms measurement-free baselines in frozen-feature response prediction across all three domains, reducing spatial angular error from 69.0° to 20.2°. With only 10% downstream labels, its ECG representation leads four released encoders on four of five clinical tasks. Controlled ablations further show that direct response supervision yields lower nonlinear readout error than separately predicting source measurements and transformations, supporting known response laws as a source of supervision for representation learning.

1 INTRODUCTION

Many augmentation-based self-supervised methods learn representations by encouraging agreement between transformed views (Chen et al., 2020). For signal-learning tasks that depend on numerical measurements, useful features must retain the variation needed to recover those measurements (Yang et al., 2023). Augmented views of the same source can have different measurement values. Transformation-aware objectives already seek to preserve information about applied actions or their effects in feature space (Dangovski et al., 2022; Devillers & Lefort, 2023). We use known responses of existing measurements to supervise transformed signal views.

In this setting, a training signal is paired with a measurement, and a known response law maps that measurement to a target for a transformed view. Clinical ECG archives and acoustic datasets already provide measurements or annotations of this kind (Gow et al., 2023; Engel et al., 2017; Politis et al., 2022). Digital time dilation scales cardiac intervals, pitch transposition adds a semitone offset, and a change in spatial-audio coordinates rotates a direction vector. These examples share the same learning opportunity: one source measurement can supervise multiple transformed views without annotating each view separately.

We introduce RePs (Response-law Privileged Supervision), which pairs each transformed signal with its response target. Given a signal x, source measurement m, and action g, RePs constructs $v = T_g(x)$ and $r = \mathcal{A}_g(m)$. An encoder and response head predict the standardized target from v alone; m and g construct supervision but do not enter the predictor. Predicting the measurement associated with each transformed view encourages the features to retain measurement-relevant variation. The response loss can train the encoder alone or accompany a contrastive objective through a separate projector (Figure 1).

We evaluate RePs through response prediction across three signal domains, clinical transfer, and ablation studies.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

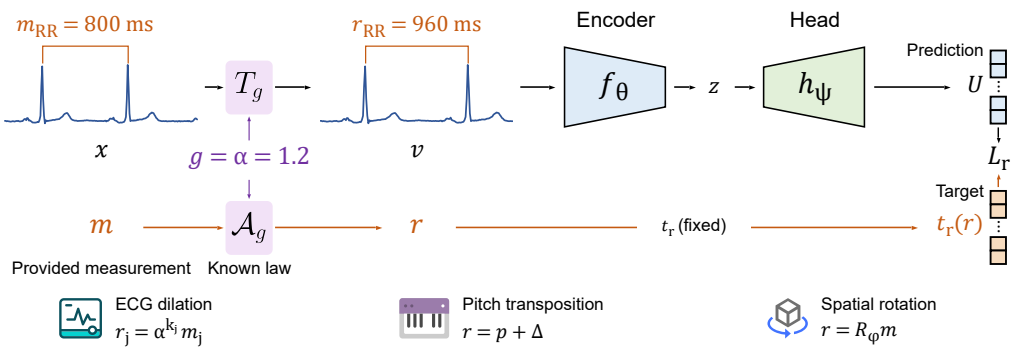


Figure 1: **Learning from known response laws.** The same action g transforms the signal and its provided measurement. Only the transformed view enters the predictor; f_θ and h_ψ produce the standardized prediction U, which is trained against $t_r(r)$ using fixed training statistics. The ECG trace is time-normalized to an illustrative RR of 800 ms; digital dilation by $\alpha = 1.2$ gives 960 ms. The response loss can also be combined with a contrastive objective through a separate projector (Section 3).

Our ablations compare direct response prediction with separate prediction of the source measurement and transformation using the same measurements. A shifted note can reveal its resulting pitch without uniquely specifying its original pitch and applied shift. We evaluate both newly fitted readers and predictors retained from encoder training. The main contributions of this work are summarized as follows:

- We introduce RePs, which turns existing measurements into supervision for transformed signals through known response laws. The response objective trains signal-only representations without annotating the transformed views, either on its own or alongside a contrastive loss.
- Across ECG timing, musical pitch, and spatial audio, RePs with a contrastive objective yields lower frozen-feature response error than the evaluated measurement-free objectives. A separately pretrained ECG encoder achieves the highest mean on four of five clinical tasks against four released encoders using 10% of downstream training labels.
- We show that the preferred supervision target can depend on how the trained model is used. At fixed measurement access, direct response supervision yields lower full-support nonlinear readout error than separate factor prediction across three domains, while their ranking reverses on musical pitch when trained predictors are reused.

2 Related Work

Transformations and numerical targets. Transforming a signal and its label consistently is an established form of supervised data augmentation. For direction-of-arrival estimation, Mazzon et al. (2019) rotate first-order Ambisonics signals and their direction labels together; Lie point symmetries generate transformed training solutions for differential equations (Brandstetter et al., 2022). Without absolute measurements, SimPer learns periodicity from relative playback speeds, and PESTO learns pitch through transposition consistency with synthetic calibration (Yang et al., 2023; Riou et al., 2023). RePs makes known measurement responses a training objective for signal representations. A single source measurement generates view-specific targets, which can train the encoder alone or alongside contrastive learning.

Transformation-aware representations. Rotation prediction, AugSelf, and E-SSL supervise applied transformations to preserve information that view alignment may discard (Gidaris et al., 2018; Lee et al., 2021; Dangovski et al., 2022). For ECG signals, temporal–spatial reverse detection likewise uses transformation prediction as a self-supervised task (Zhang et al., 2023). EquiMod and NFT learn how actions operate in feature space (Devillers & Lefort, 2023; Koyama et al., 2024). CARE, SIE, and SER constrain representation geometry or separate where invariance and equivariance are encouraged (Gupta et al., 2024; Garrido et al., 2023; Lee et al., 2026). These approaches preserve transformation information without the source measurements used by RePs.

Supervision and representation evaluation. LUPI formalizes training-only information, while TRAM uses it to model label noise (Vapnik & Vashist, 2009; Collier et al., 2022). In RePs, measurements construct targets rather than enter the encoder. AuxiLearn learns auxiliary tasks or their loss combinations (Navon et al., 2021), and task-geometry studies examine how supervised objectives shape representations (Alleman et al., 2024).

Readout model switching studies reader complexity and data efficiency, while risk decomposition separates representation usability from probe and encoder generalization (Li et al., 2023; Dubois et al., 2023). Other work examines label efficiency, cross-task transfer, and encoder versus projector features (Shi et al., 2023; Ouyang et al., 2025). We vary supervision at fixed measurement access and evaluate both newly fitted readers and jointly trained predictors.

Physiological and time-series learning. ECG pretraining objectives use subject relationships, temporal–spectral structure, and heartbeat or rhythm structure (Lan et al., 2022; Yang & Hong, 2022; Jin et al., 2025); other studies refine positive-pair selection or use physiological priors (Lan et al., 2024; Maghsoodi et al., 2026). ECG-language alignment supports zero-shot classification (Li et al., 2024), while ECGFounder studies large-scale pretraining with diagnostic labels (Li et al., 2025). Released encoders learn from reports, diagnostic descriptions, or waveform-only pretraining (Liu et al., 2024; Tian et al., 2024; Coppola et al., 2024; McKeen et al., 2025). Related physiological-sensing applications include human-body digital twins and wearable sleep monitoring (Tang et al., 2024; 2025; 2026). Within physiological signal learning, RePs addresses how existing measurements can supervise representations through known signal transformations.

3 Learning from known response laws

3.1 Constructing a response target

Each training example contains a signal x_i and source measurement m_i. The same sampled action g_i determines both the transformed view and its target: T_{g_i} acts on the signal, and the known response law $\mathcal{A}_{g_i}$ acts on the measurement. For example, shifting a note by Δ semitones changes its pitch target from p to $p + \Delta$. An encoder f_θ and response head h_ψ predict the target from the view alone. The fixed map t_r standardizes targets using encoder-training statistics; t_r^{-1} restores their units:

$$\begin{aligned} v_i &= T_{g_i}(x_i), & r_i &= \mathcal{A}_{g_i}(m_i), \\ z_i &= f_\theta(v_i), & \widehat{r}_i &= t_r^{-1}(h_\psi(z_i)) \,. \end{aligned} \tag{1}$$

Only v_i enters the network: m_i and g_i define supervision, not prediction inputs.

The three domains instantiate scaling, translation, and rotation responses. For ECG dilation, $T_\alpha x(t) = x(t/\alpha)$ and $(\mathcal{A}_\alpha(m))_j = \alpha^{k_j} m_j$, where k_j specifies how measurement coordinate j scales. For pitch transposition, $\mathcal{A}_\Delta(p) = p + \Delta$ in semitones. For spatial audio, $\mathcal{A}_\phi(m) = R_\phi m$, where m is a three-dimensional direction and R_ϕ rotates its azimuth. The signal operations, action grids, and ECG exponents are specified in Appendix A.2.

The ECG law describes digital resampling rather than physiological adaptation to heart rate. Provider measurements describe the full record, while the network sees an eight-second view. Measurements recomputed on a cropped view can differ from the nominal target. Appendix C.4 examines how remeasured targets affect the readout comparison.

3.2 Training the representation

The response loss trains the features to predict the transformed measurement. It can be used alone or combined with a contrastive loss on a separate projection of the same features. In the three-domain response-learning experiments, the response loss is standardized squared error: we average observed entries within each coordinate, then average coordinates. Let $q_{ij} \in \{0, 1\}$ indicate target availability. The set $J = \{j : \sum_i q_{ij} > 0\}$ contains the coordinates observed in a minibatch, and U_{ij} denotes the standardized prediction. Let R and Q collect the minibatch targets and masks. With $t_{r,j}(r) = (r_j - \mu_{r,j})/\sigma_{r,j}$, the loss is

$$\mathcal{D}_r(U, R, Q) = \frac{1}{C_r |J|} \sum_{j \in J} \frac{\sum_i q_{ij} \big(U_{ij} - t_{r,j}(r_i)\big)^2}{\sum_i q_{ij}}. \tag{2}$$

The target statistics and constant-predictor scale C_r are fixed from encoder-training data (Appendix A.3). We write this response loss as L_r and set it to zero for an empty J.

We call training with L_r alone *response-only RePs*. For the combined objective, we generate two views per source and apply the same encoder and response head to both. A separate projector q_ϕ supplies features to NT-Xent:

$$L_{\text{shared}} = \tfrac{1}{2}\left(L_r^{(1)} + L_r^{(2)}\right) + \lambda_{\text{base}} L_{\text{NT-Xent}}. \tag{3}$$

Both terms update the encoder, but they act on different heads: h_ψ predicts the numerical target and q_ϕ supports view alignment (Chen et al., 2020). We call this configuration *shared-base RePs*.

The initial response-to-NT-Xent encoder-gradient norm ratio on training batches sets λ_{base}, which remains fixed during training. Appendix A.3 gives the calibration procedure and coefficients.

3.3 Training configurations and model use

Table 1 distinguishes the three separately trained configurations. The comparison with measurement-free objectives uses shared-base RePs. The supervision ablations use both shared-base and response-only training. The separate clinical encoder, RePs-SSL, uses NT-Xent on two views and Huber response loss on a third, with unit coefficients. Its Huber threshold is 1 in standardized target space, and its loss averages observed cells rather than the coordinate-wise means in Eq. 2. Appendix C.1 gives that training recipe.

Table 1: **Three separately trained RePs configurations.** Each row uses its own checkpoints. L_r is normalized masked MSE; the clinical encoder instead uses masked Huber loss. Appendices A and C give the respective training protocols.

Configuration	Objective	Training data	Schedule
Shared-base	Response + contrastive (Eq. 3)	ECG, NSynth, STARSS22	20 pairs/source 40 forwards
Aligned response-only	Response only (L_r)	ECG, NSynth, STARSS22	40 single views per source
Fullscale RePs-SSL	Contrastive (views 1, 2) Huber response (view 3)	360,455 ECG records	10 epochs 14,080 steps

We evaluate the trained model in two ways. A *refitted reader* is a new predictor trained on frozen encoder features. We fit Ridge, MLP-1, MLP-2, and bilinear readers to the same response targets, using MLP-2 for the principal comparisons. Appendix A.5 specifies their architectures, fitting data, and selection rules.

A *retained predictor* reuses the response head or factor-composition path learned during encoder training. Both principal evaluations use the signal alone. Appendix B.4 additionally evaluates readers supplied with the true action.

4 Experiments

We evaluate response prediction across three signal domains and clinical transfer with a separately pretrained ECG encoder. We then ablate the supervision targets and losses, and analyze predictor reuse.

4.1 Evaluation protocol

The response-learning experiments use MIMIC-IV-ECG, NSynth, and STARSS22 (Gow et al., 2023; Engel et al., 2017; Politis et al., 2022). Patients, instruments, and recording clips define disjoint groups for encoder training, reader fitting, reader development, and evaluation. The principal comparisons use a common ResNet1D backbone, 128-dimensional features, five encoder seeds, and final-step checkpoints. Appendix A specifies the splits and training recipes.

Primary evaluation tests unseen actions on held-out source groups. Readers are fitted and selected on the action grids used for encoder training. We report patient-equal standardized mean absolute error (MAE) across six ECG timing coordinates, pitch MAE in semitones for NSynth, and clip-macro angular error in degrees for STARSS22. Lower error indicates more accurate response prediction.

Reader-fitting budgets count distinct sources, not augmented views. The full reader-fitting set contains 84,735 ECG records, 12,349 NSynth notes spanning 32 training-defined pitch strata, or 871 STARSS22 windows; we call this *full support*. Smaller budgets use three source draws paired across methods; NSynth retains the same pitch support. Feature normalization uses fitting data only, and development data select regularization or update counts.

Uncertainty is computed across five encoder initializations on a fixed group split. We average reader starts and source draws within each encoder, then report means and pointwise $t(4)$ intervals. We use seed-paired differences for the ablations, with separate Holm corrections for the 12 principal nonlinear-reader contrasts and four spatial-factorial tests. Appendix A.6 defines the other statistical families.

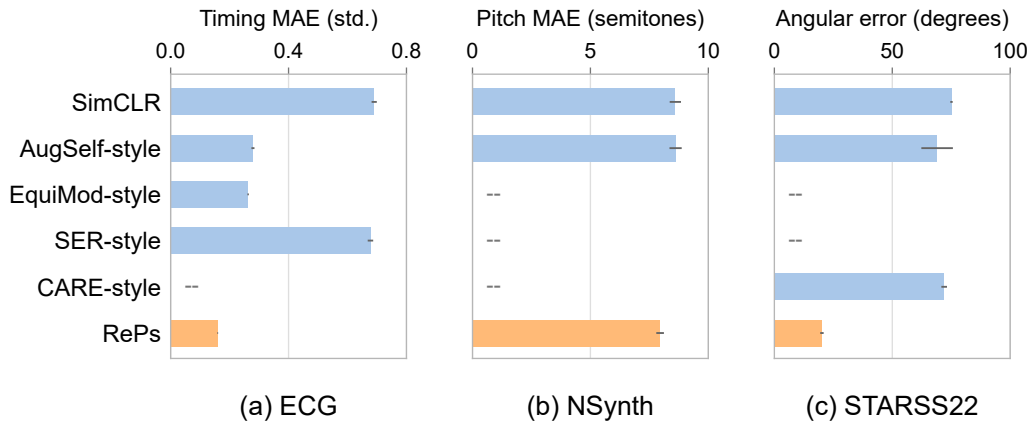


Figure 2: **Response readout after pretraining with or without measurement supervision.** Full-support MLP-2 error from frozen 128-dimensional features; lower is better. RePs uses measurement targets with a contrastive base, while the other objectives are measurement-free adaptations. All use 40 encoder forwards per source. Bars are five-encoder means with pointwise 95% intervals. Dashes mark unevaluated combinations. Error axes start at zero.

4.2 Learning numerical responses from measurements

We compare shared-base RePs with SimCLR and AugSelf-style in all three domains, EquiMod-style and SER-style in ECG, and CARE-style in spatial audio (Chen et al., 2020; Lee et al., 2021; Devillers & Lefort, 2023; Lee et al., 2026; Gupta et al., 2024). We adapt these measurement-free objectives to a common signal backbone; Appendix A.4 specifies the changes to their losses and signal operations.

All methods use the same signals and 40 encoder forwards per source; shared-base RePs additionally uses measurement-derived targets. Section 4.4 examines target choice through ablations using the same measurements.

RePs has the lowest mean MLP-2 error among these objectives in each domain (Figure 2). RePs reduces mean ECG error from 0.263 (EquiMod-style) to 0.160, pitch error from 8.60 (SimCLR) to 7.96 semitones, and spatial error from 69.01° (AugSelf-style) to 20.17°. RePs also gives lower mean error with Ridge, MLP-1, and bilinear readers (Appendix B.1).

4.3 TRANSFER TO CLINICAL ECG TASKS

We assess clinical transfer with frozen-encoder linear probes and 10% of the training labels. RePs-SSL is pretrained separately for ten epochs on 360,455 MIMIC-IV-ECG records from 87,431 patients with the three-view NT-Xent–Huber recipe in Section 3. We compare it with four released encoders: MERL, HuBERT-ECG, ECG-FM, and ECGFM-KED (Liu et al., 2024; Coppola et al., 2024; McKeen et al., 2025; Tian et al., 2024). A matched NT-Xent-only SimCLR control shares the pretraining corpus, backbone, and schedule, with the response-loss weight set to zero. The tasks are PTB-XL Rhythm-12, CPSC2018, Chapman rhythm, MIMIC cardiac labels, and CODE-15 (Wagner et al., 2020; Liu et al., 2018; Zheng et al., 2020; Ribeiro et al., 2020). All systems share downstream splits and the probe recipe; input adapters follow each encoder's requirements (Appendix C). All four released encoders report MIMIC pretraining, and HuBERT-ECG reports sources covering all five tasks; record-level overlaps with the evaluation sets are unquantified (Appendix C.2).

Table 2: **Clinical transfer with 10% training labels.** The upper four rows use released encoders; the lower two use matched fullscale pretraining, with response supervision disabled for SimCLR. Values are mean ± SD over five repeats with frozen-encoder linear probes; higher is better. The SD includes encoder and reader variation for SimCLR and RePs-SSL, but only reader variation for released encoders. Bold and underline mark the highest and second-highest means. Appendix C gives the protocols.

Encoder	PTB-XL AUROC	CPSC AUPRC	Chapman F1	MIMIC AUPRC	CODE-15 AUPRC
MERL	$\underline{0.783 \pm 0.024}$	$\underline{0.585 \pm 0.006}$	0.820 ± 0.001	$\underline{0.067 \pm 0.003}$	0.304 ± 0.016
HuBERT-ECG	0.749 ± 0.061	0.467 ± 0.003	$\underline{0.827 \pm 0.002}$	0.058 ± 0.002	$\mathbf{0.476 \pm 0.009}$
ECG-FM	0.702 ± 0.038	0.481 ± 0.003	0.653 ± 0.002	0.058 ± 0.001	0.312 ± 0.006
ECGFM-KED	0.644 ± 0.075	0.290 ± 0.010	0.610 ± 0.002	0.049 ± 0.001	0.055 ± 0.015
SimCLR	0.662 ± 0.030	0.423 ± 0.012	0.621 ± 0.017	0.055 ± 0.001	0.164 ± 0.005
RePs-SSL	$\mathbf{0.836 \pm 0.019}$	$\mathbf{0.636 \pm 0.008}$	$\mathbf{0.873 \pm 0.004}$	$\mathbf{0.070 \pm 0.001}$	$\underline{0.376 \pm 0.023}$

Under matched pretraining, RePs-SSL achieves higher mean scores than SimCLR on all five tasks (Table 2). PTB-XL AUROC rises from 0.662 to 0.836, and Chapman F1 from 0.621 to 0.873. Appendix C.5 reports the full five-objective clinical comparison.

Against the four released encoders, the 0.219M-parameter RePs-SSL encoder achieves the highest mean on PTB-XL, CPSC, Chapman, and MIMIC. It ranks second on CODE-15 at 0.376 AUPRC, versus 0.476 for HuBERT-ECG. The released-encoder comparison evaluates complete pretrained systems.

4.4 ABLATION STUDIES AND READOUT ANALYSIS

We ablate supervision targets, prediction paths, and auxiliary losses in the three signal domains, using fixed source signals, measurements, and action schedules.

Direct response prediction targets the transformed measurement; Factorized instead targets its source measurement and applied action. We evaluate each objective with newly fitted readers and retained predictors.

Factorized predicts the source measurement and action separately with $\frac{1}{2}L_m + \frac{1}{2}L_g$. Direct+Aux adds $\frac{1}{4}L_m + \frac{1}{4}L_g$ to L_r and retains a direct response head. F+Comp adds supervision of the response obtained by composing those predictions. For a minibatch of B examples, its loss is

$$\begin{aligned} L_{\mathrm{F+Comp}} &= \tfrac{1}{4}L_m + \tfrac{1}{4}L_g + \tfrac{1}{2}L_c, \\ \widetilde{r}_i &= \mathcal{A}_{\widehat{g}_i}(\widehat{m}_i), \qquad \widetilde{R} = (\widetilde{r}_i)_{i=1}^{B}, \qquad L_c = \mathcal{D}_r\Big(t_r(\widetilde{R}), R, Q\Big). \end{aligned} \tag{4}$$

L_m and L_g follow the coordinate-standardized loss convention in Eq. 2, with separate fixed scales; state coordinates use availability masks and actions are fully observed. The head outputs are inverse-

standardized before composition. The composed response is then standardized, so L_c supplies standardized predictions and raw targets to $\mathcal{D}_r$.

Appendix A.3 gives the branch definitions and raw-space parameterizations.

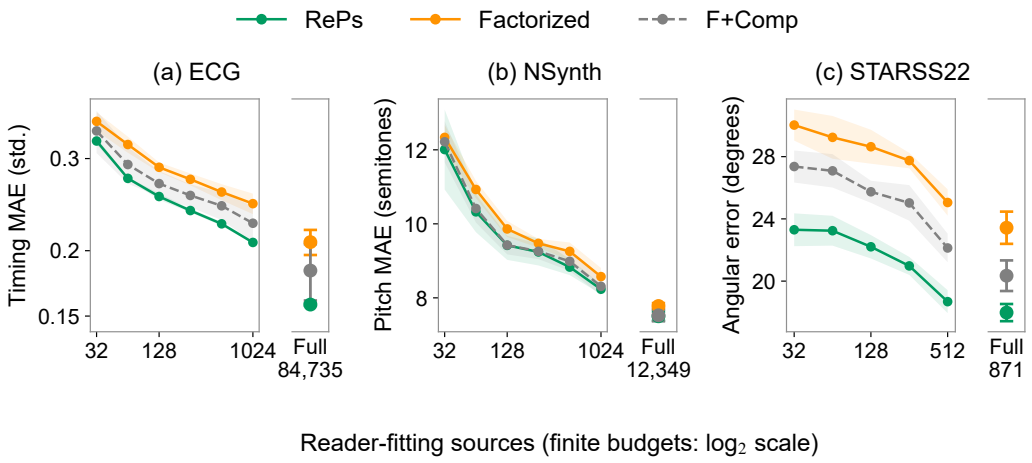


Figure 3: **Response readout across fitting budgets.** A new MLP-2 reader is fitted on frozen response-only features at each budget. Curves are five-encoder means with pointwise 95% intervals; lower is better. Finite budgets use a $\log_2$ axis; Full is a separate condition labeled by source count. NSynth retains 32-pitch support, STARSS22 has no 1,024-source condition, and ECG uses a logarithmic error axis.

Supervision with a shared contrastive base. Direct response supervision gives lower mean MLP-2 error than Factorized in all three domains when the contrastive coefficient is fixed within each domain (Table 3). A composition loss narrows the gap. Auxiliary factor targets help in the NSynth shared-base setting: Direct+Aux has the lowest mean, 7.912 semitones versus 7.955 for RePs.

Table 3: **Supervision ablations with a fixed contrastive base.** Full-support MLP-2 means over five encoders, using the same source measurements and domain-specific λ_{base} for every objective. Lower is better. Bold identifies the lowest mean in each column.

Objective	ECG Timing MAE (std.) ↓	NSynth Semitones ↓	STARSS22 Degrees ↓
RePs	**0.160**	7.955	**20.174**
Factorized	0.201	8.080	24.916
F+Comp	0.188	8.013	22.975
Direct+Aux	0.169	**7.912**	21.767

Response-only supervision. Without the contrastive term, RePs has lower full-support MLP-2 and bilinear error than Factorized in all three domains (Table 4). All six paired contrasts pass the principal 12-test Holm correction. Adding composition closes much of the gap: only the two STARSS22 RePs–F+Comp contrasts remain significant in that family. Figure 3 reports the MLP-2 comparison at smaller fitting budgets. Response-only RePs also has lower mean MLP-2 error than shared-base RePs in each domain (Tables 3 and 4).

Reusing the trained predictors reverses the ordering on NSynth. With refitted MLP-2, RePs gives 7.50 semitones and Factorized gives 7.78; with retained prediction, the errors are 7.95 and 6.78, respectively. The retained F+Comp and Direct+Aux paths reach 6.48 and 6.24 semitones. Direct+Aux has the lowest retained mean here, combining a direct response head with auxiliary factor targets. Retained predictors are trained jointly with their encoders; refitted readers are trained on the separate fitting set.

Table 4: **Refitted response readout versus retained prediction.** Full-support means over five response-only encoders; lower is better. MLP-2 and bilinear readers are newly fitted, whereas Retained reuses the trained path. Stars on Factorized or F+Comp values mark paired differences from RePs that pass the principal 12-test Holm correction. Direct+Aux and retained comparisons are descriptive. Appendix B.2 gives the complete paired comparisons.

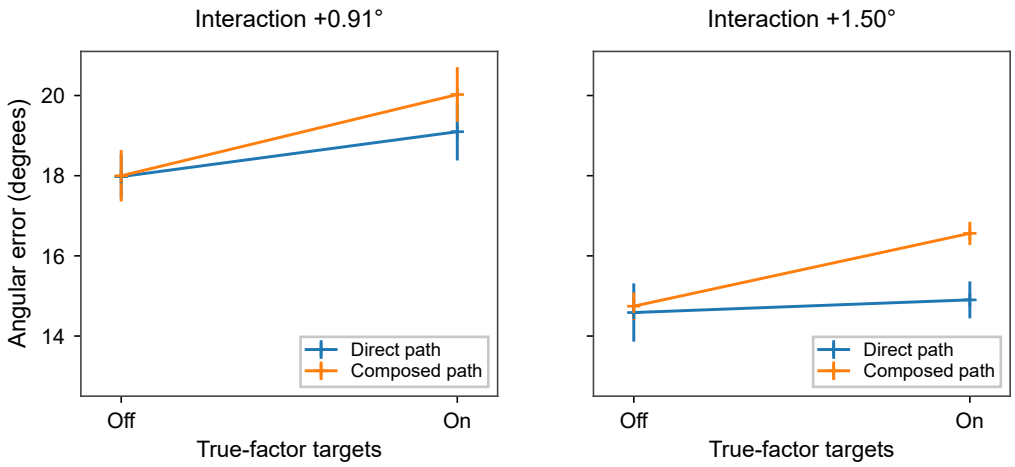
	Refitted readers		Trained path
Objective	MLP-2	Bilinear	Retained
ECG: standardized timing MAE ↓			
RePs	0.158	0.155	0.147
Factorized	0.208*	0.210*	0.240
F+Comp	0.183	0.181	0.187
Direct+Aux	0.168	0.165	0.158
NSynth: pitch MAE (semitones) ↓			
RePs	7.50	8.29	7.95
Factorized	7.78*	8.71*	6.78
F+Comp	7.53	8.64	6.48
Direct+Aux	7.56	8.65	6.24
STARSS22: angular error (degrees) ↓			
RePs	17.98	21.49	14.59
Factorized	23.43*	28.46*	19.67
F+Comp	20.35*	24.88*	17.01
Direct+Aux	19.10	23.68	14.90

Separating prediction path from factor supervision. Factor-based prediction changes both the path used to produce a response and the targets supplied during training. A factorial ablation on STARSS22 crosses direct and composed prediction paths with the presence or absence of true-factor targets. The direct cells are RePs and Direct+Aux; the composed cells are CompOnly and Comp+Aux. Each cell has response coefficient one, with $\frac{1}{4}L_m + \frac{1}{4}L_g$ added in auxiliary conditions. CompOnly learns factor-shaped outputs through response loss without true-factor targets. F+Comp has response coefficient one-half and is not a factorial cell.

Without true-factor targets, RePs and CompOnly give MLP-2 errors of $17.98°$ and $18.00°$, a paired CompOnly-minus-RePs difference of $0.02°$ with 95% interval $[-0.40, 0.44]$. True-factor targets increase error by $1.12°$ on the direct path and $2.03°$ on the composed path. The interaction is $0.91°$ for MLP-2 and $1.50°$ for retained prediction; both pass the four-test Holm correction (Figure 4). At these coefficients, true-factor targets incur a larger error cost on the composed path for MLP-2 readout and retained prediction. Ridge and bilinear interaction intervals include zero (Appendix B.3).

The benefit of correct action–target coupling depends on the domain and reader. For example, predicting the original pitch from transformed NSynth signals (Static) gives full-support MLP-2 error of 7.491 semitones, versus 7.502 for RePs (Appendix B.2).

The supervision ranking also changes with the prediction target. Identity, which predicts source measurements from untransformed views, gives lower native ECG timing error, while audio classification rankings depend on the label (Appendix C.3).



(a) Refitted MLP-2 (b) Retained predictor

Figure 4: **Prediction path and true-factor targets at matched coefficients.** Factor targets off/on give RePs/Direct+Aux on the direct path and CompOnly/Comp+Aux on the composed path. All cells use response coefficient one. Points and whiskers are five-encoder means and pointwise 95% intervals. The paired interactions are $+0.91°$ (95% CI $[0.19, 1.62]$, $p_H = 0.049$) for refitted MLP-2 and $+1.50°$ (95% CI $[1.16, 1.84]$, $p_H = 0.001$) for retained prediction; p_H is Holm-adjusted within the four-test family. Appendix B.3 gives all four interfaces.

Training-budget sensitivity. Doubling STARSS22 encoder exposure from 40 to 80 forwards per source preserves the mean MLP-2 ordering of RePs, F+Comp, and Factorized (Appendix D.1). The Factorized–RePs gap changes from $5.45°$ to $5.17°$, and the F+Comp–RePs gap from $2.37°$ to $2.71°$. A separate test keeps the 40-forward encoders fixed and increases the reader update cap from 4,000 to 8,000, with patience extended from five to ten checks. RePs retains lower mean error than Factorized and F+Comp with both MLP-2 and bilinear readers.

Response prediction with observable factors. A synthetic control provides both generating factors in the input, so recovering them does not require resolving an ambiguous signal decomposition (Appendix D.3). The response is either their sum or their product. With a one-dimensional representation, RePs, F+Comp, and Direct+Aux all have lower response error than Factorized under both laws, across retained and refitted prediction. At larger widths, the ordering depends on the law and prediction interface. The narrow-representation result illustrates why factor accessibility alone does not determine the preferred supervision: encoding the resulting response and retaining its two generating factors place different demands on a limited representation.

5 Conclusion

RePs uses known response laws to turn existing measurements into supervision for transformed views without annotating those views individually. The response objective trains signal-only representations on its own or alongside a contrastive loss. Across ECG timing, musical pitch, and spatial audio, the combined objective improves frozen-feature response readout over the evaluated measurement-free adaptations. A separately pretrained ECG encoder achieves the highest mean on four of five clinical tasks against four released encoders using 10% of downstream training labels.

The target ablations show that direct response supervision gives lower full-support nonlinear readout error than separate factor prediction using the same measurements; a composition loss narrows the gap. On NSynth, their ranking reverses when the trained predictors are reused. The preferred target can therefore differ between learning features for a new reader and training a predictor for reuse.

AI USE STATEMENT

Large language models were used to assist with language polishing and readability, literature discovery, research brainstorming, and feedback on experimental design and result interpretation. The authors reviewed all AI-assisted content, independently made the final scientific decisions, and take full responsibility for the manuscript.

REPRODUCIBILITY STATEMENT

An anonymized code repository is available at `https://anonymous.4open.science/r/RePs_Anonymous-6131`. Sections 3 and 4, together with the appendices, provide the data processing, model configurations, training and evaluation protocols, fixed seeds, and statistical procedures used in the experiments.

REFERENCES

Matteo Alleman, Jack Lindsey, and Stefano Fusi. Task structure and nonlinearity jointly determine learned representational geometry. In *International Conference on Learning Representations*, 2024.

Johannes Brandstetter, Max Welling, and Daniel E Worrall. Lie point symmetry data augmentation for neural PDE solvers. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 2241–2256. PMLR, 2022. URL https://proceedings.mlr.press/v162/brandstetter22a.html.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 1597–1607, 2020.

Mark Collier, Rodolphe Jenatton, Effrosyni Kokiopoulou, and Jesse Berent. Transfer and marginalize: Explaining away label noise with privileged information. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 4219–4237, 2022.

Edoardo Coppola, Mattia Savardi, Mauro Massussi, Marianna Adamo, Marco Metra, and Alberto Signoroni. HuBERT-ECG as a self-supervised foundation model for broad and scalable cardiac applications. *medRxiv*, 2024. doi: 10.1101/2024.11.14.24317328.

Rumen Dangovski, Li Jing, Charlotte Loh, Seungwook Han, Akash Srivastava, Brian Cheung, Pulkit Agrawal, and Marin Soljacic. Equivariant contrastive learning. In *International Conference on Learning Representations*, 2022.

Alexandre Devillers and Mathieu Lefort. EquiMod: An equivariance module to improve visual instance discrimination. In *International Conference on Learning Representations*, 2023.

Yann Dubois, Tatsunori Hashimoto, and Percy Liang. Evaluating self-supervised learning via risk decomposition. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 8779–8820. PMLR, 2023.

Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. Neural audio synthesis of musical notes with WaveNet autoencoders. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1068–1077, 2017.

Quentin Garrido, Laurent Najman, and Yann Lecun. Self-supervised learning of split invariant equivariant representations. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10975–10996. PMLR, 2023.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018.

Brian Gow, Tom Pollard, Larry A. Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Jonathan W. Waks, Parastou Eslami, Tanner Carbonati, Ashish Chaudhari, Elizabeth Herbst, Dana Moukheiber, Seth Berkowitz, Roger Mark, and Steven Horng. MIMIC-IV-ECG: Diagnostic electrocardiogram matched subset. PhysioNet, 2023. URL https://doi.org/10.13026/4nqg-sb35. Version 1.0.

Sharut Gupta, Joshua Robinson, Derek Lim, Soledad Villar, and Stefanie Jegelka. Structuring representation geometry with rotationally equivariant contrastive learning. In *International Conference on Learning Representations*, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Jiarui Jin, Haoyu Wang, Hongyan Li, Jun Li, Jiahui Pan, and Shenda Hong. Reading your heart: Learning ECG words and sentences via pre-training ECG language model. In *International Conference on Learning Representations*, 2025.

Alena Kalyakulina, Igor Yusipov, Viktor Moskalenko, Alexander Nikolskiy, Konstantin Kosonogov, Nikolai Zolotykh, and Mikhail Ivanchenko. Lobachevsky university electrocardiography database. PhysioNet, 2021. URL https://physionet.org/content/ludb/1.0.1/. Version 1.0.1.

Masanori Koyama, Kenji Fukumizu, Kohei Hayashi, and Takeru Miyato. Neural Fourier transform: A general approach to equivariant representation learning. In *International Conference on Learning Representations*, 2024.

Xiang Lan, Dianwen Ng, Shenda Hong, and Mengling Feng. Intra-inter subject self-supervised learning for multivariate cardiac signals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 4532–4540, 2022. doi: 10.1609/aaai.v36i4.20376.

Xiang Lan, Hanshu Yan, Shenda Hong, and Mengling Feng. Towards enhancing time series contrastive learning: A dynamic bad pair mining approach. In *International Conference on Learning Representations*, 2024.

Hankook Lee, Kibok Lee, Kimin Lee, Honglak Lee, and Jinwoo Shin. Improving transferability of representations via augmentation-aware self-supervision. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17710–17722, 2021.

Joohyung Lee, Changhun Kim, Hyunsu Kim, Kwanhyung Lee, and Juho Lee. Soft equivariance regularization for invariant self-supervised learning. In *International Conference on Learning Representations*, 2026.

Jun Li, Che Liu, Sibo Cheng, Rossella Arcucci, and Shenda Hong. Frozen language model helps ECG zero-shot learning. In *Medical Imaging with Deep Learning*, volume 227 of *Proceedings of Machine Learning Research*, pp. 402–415. PMLR, 2024.

Jun Li, Aaron D. Aguirre, Valdery Moura Junior, Jiarui Jin, Che Liu, Lanhai Zhong, Chenxi Sun, Gari Clifford, M. Brandon Westover, and Shenda Hong. An electrocardiogram foundation model built on over 10 million recordings. *NEJM AI*, 2(7), 2025. doi: 10.1056/AIoa2401033.

Yazhe Li, Jorg Bornschein, and Marcus Hutter. Evaluating representations with readout model switching. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2302.09579.

Che Liu, Zhongwei Wan, Cheng Ouyang, Anand Shah, Wenjia Bai, and Rossella Arcucci. Zero-shot ECG classification with multimodal learning and test-time clinical knowledge enhancement. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 31949–31963. PMLR, 2024. URL https://proceedings.mlr.press/v235/liu24bg.html.

Feifei Liu, Chengyu Liu, Lina Zhao, Xiangyu Zhang, Xiaoling Wu, Xiaoyan Xu, Yulin Liu, Caiyun Ma, Shoushui Wei, Zhiqiang He, Jianqing Li, and Eddie Ng Yin Kwee. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. *Journal of Medical Imaging and Health Informatics*, 8(7):1368–1373, 2018. doi: 10.1166/jmihi.2018.2442.

Yuhang Liu, Peng Zhang, and Fan Lin. Resting ECG segmentation dataset. figshare, 2025. URL https://doi.org/10.6084/m9.figshare.28892186. Dataset.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Nooshin Maghsoodi, Sarah Nassar, Paul F. R. Wilson, Minh Nguyen Nhat To, Sophia Mannina, Shamel Addas, Stephanie Sibley, David Pichora, David Maslove, Purang Abolmaesumi, and Parvin Mousavi. Domain knowledge is power: Leveraging physiological priors for self-supervised representation learning in electrocardiography. *IEEE Transactions on Biomedical Engineering*, pp. 1–10, 2026. doi: 10.1109/TBME.2026.3656904.

Dominique Makowski, Tam Pham, Zen J. Lau, Jan C. Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and S. H. Annabel Chen. NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behavior Research Methods*, 53(4):1689–1696, 2021. doi: 10.3758/s13428-020-01516-y.

Luca Mazzon, Yuma Koizumi, Masahiro Yasuda, and Noboru Harada. First order ambisonics domain spatial augmentation for DNN-based direction of arrival estimation. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, pp. 154–158, New York University, NY, USA, October 2019.

Kaden McKeen, Sameer Masood, Augustin Toma, Barry Rubin, and Bo Wang. ECG-FM: An open electrocardiogram foundation model. *JAMIA Open*, 8(5):ooaf122, 2025. doi: 10.1093/jamiaopen/ooaf122.

Aviv Navon, Idan Achituve, Haggai Maron, Gal Chechik, and Ethan Fetaya. Auxiliary learning by implicit differentiation. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2007.02693.

Zhuo Ouyang, Kaiwen Hu, Qi Zhang, Yifei Wang, and Yisen Wang. Projection head is secretly an information bottleneck. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/bf6939f9058a391c47014731b2486e2a-Abstract-Conference.html.

Archontis Politis, Kazuki Shimada, Parthasaarathy Sudarsanam, Sharath Adavanne, Daniel Krause, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, Yuki Mitsufuji, and Tuomas Virtanen. STARSS22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop*, 2022.

Antônio H. Ribeiro, Manoel Horta Ribeiro, Gabriela M. M. Paixão, Derick M. Oliveira, Paulo R. Gomes, Jéssica A. Canazart, Milton P. S. Ferreira, Carl R. Andersson, Peter W. Macfarlane, Wagner Meira, Thomas B. Schön, and Antonio Luiz P. Ribeiro. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature Communications*, 11:1760, 2020. doi: 10.1038/s41467-020-15432-4.

Alain Riou, Stefan Lattner, Gaëtan Hadjeres, and Geoffroy Peeters. PESTO: Pitch estimation with self-supervised transposition-equivariant objective. In *Proceedings of the International Society for Music Information Retrieval Conference*, 2023.

Zhenmei Shi, Jiefeng Chen, Kunyang Li, Jayaram Raghuram, Xi Wu, Yingyu Liang, and Somesh Jha. The trade-off between universality and label efficiency of representations from contrastive learning. In *International Conference on Learning Representations*, 2023.

Kazuki Shimada, Archontis Politis, Parthasaarathy Sudarsanam, Daniel A. Krause, Kengo Uchida, Sharath Adavanne, Aapo Hakala, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, Tuomas Virtanen, and Yuki Mitsufuji. STARSS23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/e6c9671ed3b3106b71cafda3ba225c1a-Abstract.html.

Nils Strodthoff, Juan Miguel Lopez Alcaraz, and Wilhelm Haverkamp. MIMIC-IV-ECG-Ext-ICD: Diagnostic labels for MIMIC-IV-ECG. *PhysioNet*, August 2024a. doi: 10.13026/ypt5-9d58. URL https://doi.org/10.13026/ypt5-9d58. Version 1.0.1.

Nils Strodthoff, Juan Miguel Lopez Alcaraz, and Wilhelm Haverkamp. Prospects for artificial intelligence-enhanced electrocardiogram as a unified screening tool for cardiac and non-cardiac conditions: an explorative study in emergency care. *European Heart Journal - Digital Health*, 5 (4):454–460, 2024b. doi: 10.1093/ehjdh/ztae039.

Chenyu Tang, Wentian Yi, Edoardo Occhipinti, Yanning Dai, Shuo Gao, and Luigi G. Occhipinti. A roadmap for the development of human body digital twins. *Nature Reviews Electrical Engineering*, 1(3):199–207, 2024. doi: 10.1038/s44287-024-00025-w.

Chenyu Tang, Wentian Yi, Muzi Xu, Yuxuan Jin, Zibo Zhang, Xuhang Chen, Caizhi Liao, Mengtian Kang, Shuo Gao, Peter Smielewski, and Luigi G. Occhipinti. A deep learning-enabled smart garment for accurate and versatile monitoring of sleep conditions in daily life. *Proceedings of the National Academy of Sciences*, 122(7):e2420498122, 2025. doi: 10.1073/pnas.2420498122.

Chenyu Tang, Xuemeng Li, Liang Qi, Wentian Yi, Zibo Zhang, Edoardo Occhipinti, Shuo Deng, Yu Pan, Luigi G. Occhipinti, and Shuo Gao. Robust sleep behavior monitoring under motion artefacts via motion-conditioned learning. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 34:3610–3620, 2026. doi: 10.1109/TNSRE.2026.3715694.

Yuanyuan Tian, Zhiyuan Li, Yanrui Jin, Mengxiao Wang, Xiaoyang Wei, Liqun Zhao, Yunqing Liu, Jinlei Liu, and Chengliang Liu. Foundation model of ECG diagnosis: Diagnostics and explanations of any form and rhythm on ECG. *Cell Reports Medicine*, 5(12):101875, 2024. doi: 10.1016/j.xcrm.2024.101875.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5–6):544–557, 2009. doi: 10.1016/j.neunet.2009.06.042.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I. Lunze, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7:154, 2020. doi: 10.1038/s41597-020-0495-6.

Ling Yang and Shenda Hong. Unsupervised time-series representation learning with iterative bilinear temporal-spectral fusion. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 25038–25054. PMLR, 17–23 Jul 2022.

Yuzhe Yang, Xin Liu, Jiang Wu, Silviu Borac, Dina Katabi, Ming-Zher Poh, and Daniel McDuff. SimPer: Simple self-supervised learning of periodic targets. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.03115.

Wenrui Zhang, Shijia Geng, and Shenda Hong. A simple self-supervised ECG representation learning method via manipulated temporal-spatial reverse detection. *Biomedical Signal Processing and Control*, 79:104194, 2023. doi: 10.1016/j.bspc.2022.104194.

Jianwei Zheng, Jianming Zhang, Sidy Danioko, Hai Yao, Hangyuan Guo, and Cyril Rakovski. A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Scientific Data*, 7:48, 2020. doi: 10.1038/s41597-020-0386-x.

A EXPERIMENTAL DETAILS

We specify the shared response-learning setup here; clinical training is described in Appendix C.1. Complete numerical results, including the tables not reproduced below, are provided in the accompanying CSV and JSON files.

A.1 SOURCE GROUPS AND DATA ROLES

The ECG pool contains quality-filtered records with all 12 measurement coordinates available. We assign each source group to exactly one role: encoder training, reader fitting, reader development, or evaluation. The groups are patients for ECG, instruments for NSynth, and recording clips for STARSS22. Assignment uses seed 20260910 and target fractions 0.60/0.15/0.10/0.15, with no group shared across roles. STARSS22 clip assignment is stratified by site/room, dominant class, and the modal 45-degree azimuth bin. Strata with fewer than seven clips are merged by dropping azimuth and then class while retaining site/room.

Source and group counts for each role appear in Table 5. The principal NSynth fitting set uses fixed pitch support; a separate extension includes every note in the fitting role. Both use the same held-out evaluation groups.

Table 5: **Sources and groups assigned to each data role.** Patients, instruments, or clips are disjoint across roles. Counts cover the complete roles; the principal NSynth full-support reader uses a 32-pitch subset of its fitting role.

Domain	Unit	Encoder train	Reader fit	Reader dev	Evaluation
ECG	Records	335,912	84,735	56,112	84,144
	Patients	80,985	20,246	13,497	20,247
NSynth	Notes	169,349	44,816	31,216	43,824
	Instruments	571	142	95	145
STARSS22	Windows	4,753	871	932	1,491
	Clips	67	16	11	19

A.2 SIGNAL OPERATIONS AND RESPONSE COORDINATES

ECG dilation. Each twelve-lead source contains 1,000 samples at 100 Hz, clipped at 5 mV. Under dilation α, output sample $j \in \{0, \dots, 799\}$ linearly interpolates the source at $499.5 + (j - 399.5)/\alpha$. At identity, the view covers source samples 100 through 899.

The 12 source coordinates comprise RR, P duration, PR, QRS duration, QT, Bazett QTc, and sine/cosine encodings of the P, QRS, and T axes. Timing is measured in milliseconds; angles are converted to radians before encoding. In $\mathcal{A}_\alpha(m)_j = \alpha^{k_j} m_j$, the first five timing exponents are one, the QTc exponent is one-half, and the six axis exponents are zero. The targets transform full-record provider annotations under the digital dilation rule. Appendix C.4 evaluates measurements recomputed on the actual views.

We use the following dilation grids:

$$\begin{aligned} \text{training:} &\quad \{0.8, 0.9, 1.1, 1.2\}, \\ \text{interpolation:} &\quad \{0.82, 0.85, 0.875, 0.925, 0.95, 0.975, 1.025, 1.05, 1.075, 1.125, 1.15, 1.175\}, \\ \text{extrapolation:} &\quad \{1.225, 1.25, 1.275, 1.3\}. \end{aligned}$$

Primary evaluation pools the listed interpolation and extrapolation actions, all absent from training.

NSynth pitch translation. NSynth targets are the official training-release metadata pitches plus the imposed semitone shifts. We extract seconds 0.5 to 2.5 of each 16-kHz mono source and divide by source RMS, floored at 10^{-4}. Pitch shifting uses `pedalboard.time_stretch` with stretch factor 1.0, the specified shift, `high_quality=False`, `transient_mode="crisp"`,

and `preserve_formants=True`. Duration is preserved, after which a centered one-second crop supplies the 16,000-sample view.

The training shifts are $\{-12, -6, 6, 12\}$ semitones and the unseen shifts are $\{-8, -4, 4, 8\}$. Cached float16 views are converted to FP32 for computation. Targets outside the MIDI range are retained.

STARSS22 spatial rotation. We select stable single-source windows from all classes in the original STARSS22 development splits. The center frame and four neighbors on each side must each contain exactly one source with the same class and identifier; their directions must lie within 10° of the center direction. We greedily select complete 0.8-second windows at least ten frames apart, without a per-clip cap. A view contains four first-order Ambisonic channels and 19,200 samples at 24 kHz. We normalize by RMS across channels and time, with floor 10^{-4}.

Azimuth a and elevation e, in radians, define $m = (\cos e \cos a, \cos e \sin a, \sin e)$. The azimuth action ϕ maps (x, y, z) to $(\cos\phi\, x - \sin\phi\, y, \sin\phi\, x + \cos\phi\, y, z)$. The action rotates the recording coordinate system.

Signal channels are stored in W, Y, Z, X order. With Y denoted by c_1 and X by c_3, the corresponding rotation is $c_1' = \cos\phi\, c_1 + \sin\phi\, c_3$ and $c_3' = -\sin\phi\, c_1 + \cos\phi\, c_3$. W and Z remain unchanged.

Training rotations are $\{-60, -30, 0, 30, 60\}$ degrees; unseen rotations are $\{-90, -45, 45, 90\}$. The action target is $(\sin\phi, \cos\phi)$.

A.3 Encoders, objectives, and normalization

We use a ResNet1D signal encoder (He et al., 2016). The stem uses a convolution with kernel size seven and stride two, followed by batch normalization and ReLU. Four residual blocks follow, with kernel size five and widths 32, 64, 64, and 128. Blocks that increase width have stride two. Global average pooling gives a 128-dimensional vector. The width study uses this vector directly at $d = 128$ and adds a learned linear projection at $d = 2$. The encoder receives 12 ECG, one NSynth, or four STARSS22 input channels.

Encoder initialization seeds are 42, 3407, 2027, 1013, and 11. We train with AdamW, an initial learning rate of 0.002, weight decay 10^{-4}, and cosine decay to 10^{-5} (Loshchilov & Hutter, 2019). Batch sizes are 256 for ECG and 32 for audio; incomplete final batches are retained. Computation uses FP32 with TF32 disabled. We use the final checkpoint after 40 encoder forwards per source; the longer-exposure ablation uses 80.

We schedule actions deterministically using index (source offset + forward pass) $\bmod K$, where K is the training-grid size. This gives each source ten exposures to each ECG or NSynth action and eight to each STARSS22 action. Two-view objectives pair adjacent scheduled passes.

Each supervised branch with D target coordinates uses its own normalization, fitted on encoder-training targets. For observed target values y_{ij} and availability q_{ij}, masked means and population standard deviations give $t_{ij} = (y_{ij} - \mu_j)/\sigma_j$. We floor standard deviations at 10^{-6} and define the constant-predictor scale

$$C = \frac{1}{D} \sum_{j=1}^{D} \frac{\sum_i q_{ij} t_{ij}^2}{\sum_i q_{ij}}, \tag{5}$$

with C floored at 10^{-12}. Every training-set coordinate must have at least one observation.

Equation 2 applies these fixed branch scales to observed entries, without inverse-probability correction for missingness. Evaluation uses MAE or angular error rather than the training MSE.

Each response-only prediction head is Linear(d,128)–GELU–Linear(128,output width), without normalization or dropout. State and response widths are 12, one, and three for ECG, NSynth, and STARSS22; action widths are one, one, and two. The contrastive projector has output width 128, and relative-action heads receive concatenated features of width $2d$. Table 6 lists the supervised branches and their coefficients.

Table 6: **Prediction targets, paths, and losses.** Branch normalization is fixed from encoder-training data. The four factorial cells have response coefficient one; F+Comp uses one-half. Factorized composes its factor predictions at evaluation. Direct+Aux retains a direct response head.

Objective	Training path	Factor targets	Loss
Matched factorial: response coefficient one			
RePs	Direct	No	L_r
Direct+Aux	Direct	Yes	$L_r + \frac{1}{4}L_m + \frac{1}{4}L_g$
CompOnly	Composed	No	L_c
Comp+Aux	Composed	Yes	$L_c + \frac{1}{4}L_m + \frac{1}{4}L_g$
Original factor-based objectives			
Factorized	State/action heads	Yes	$\frac{1}{2}L_m + \frac{1}{2}L_g$
F+Comp	Composed	Yes	$\frac{1}{2}L_c + \frac{1}{4}L_m + \frac{1}{4}L_g$

Composition first restores the predicted state and action to their original units, computes $\widetilde{r}_i = \mathcal{A}_{\widehat{g}_i}(\widehat{m}_i)$, and then standardizes the response for the same loss as direct prediction. State and response branches use source-availability masks; action coordinates are fully observed. ECG composition clamps predicted dilation at 10^{-6}; NSynth adds the predicted shift to the source pitch. For STARSS22, the predicted sine/cosine pair is normalized by its Euclidean norm with floor 10^{-8} before rotating the predicted XYZ state, which is not unit-normalized. These operations remain in the differentiation graph, including the clamps and normalizations.

For shared-base training, we calibrate λ_{base} on encoder-training batches using four independently reset models at seed 42. The median initial response-to-NT-Xent encoder-gradient norm ratio gives 0.113815 for ECG, 0.275155 for NSynth, and 0.122894 for STARSS22.

A.4 Baseline adaptations

All adapted objectives use the signal preprocessing and ResNet1D backbone above. The external comparison matches 40 encoder forwards per source: 20 view pairs for shared-base RePs, SimCLR, AugSelf-style, and SER-style, or ten quartets for EquiMod-style and CARE-style.

SimCLR and AugSelf-style. SimCLR uses two views and NT-Xent with temperature 0.1 (Chen et al., 2020). AugSelf-style adds a unit-weight MSE loss for relative actions predicted from concatenated embeddings (Lee et al., 2021). The target is a log scale ratio for ECG, a shift difference for NSynth, or sine/cosine of relative yaw for STARSS22, standardized on the training grid. A signal backbone and one transformation family replace the original image backbone and augmentation set.

EquiMod-style. Four transformed ECG views supply all six pairs to NT-Xent with temperature 0.5 (Devillers & Lefort, 2023). The equivariant term, with temperature 0.2, predicts the last three embeddings from the first embedding and standardized relative log actions. We use a transformed reference and six-pair invariant loss in place of the original unaugmented reference and two augmented views.

SER-style. The ECG adaptation combines global NT-Xent with token-level contrastive learning (Lee et al., 2026). After the first three residual blocks, an 800-sample view gives a 64×200 feature map. A shared 1×1 convolution, 128-channel batch normalization, ReLU, and a second 1×1 convolution produce 128-dimensional tokens. The global branch uses the fourth block and average pooling.

Relative dilation b/a aligns the first token sequence to the second by linear interpolation with `align_corners=True` and zero padding. We sample an all-ones mask and retain values at least 0.999 to exclude invalid boundaries. Matched positions are positives; negatives come only from other records, excluding other positions of the same record. The bidirectional token loss has temperature 0.4 and global NT-Xent has temperature 0.1, both with unit weight. Temporal tokens replace the original SER implementation's ViT spatial tokens.

CARE-style. Two STARSS22 view pairs supply an NT-Xent term and a Gram-equivariance term weighted by 0.01 (Gupta et al., 2024). The Gram term averages action-offset groups with at least two sources.

T-SSL is an internal action-prediction control: a single transformed view supervises the applied action.

A.5 Reader fitting and evaluation

Finite reader-fitting budgets contain 32, 64, 128, 256, 512, or 1,024 distinct sources. For each available budget, we use three source draws paired across methods.

Full support contains 84,735 ECG records, 12,349 NSynth notes in 32 pitch strata, or all 871 STARSS22 fitting windows. The STARSS22 fitting role is too small for a 1,024-source condition. Using all 44,816 NSynth fitting notes changes both source count and pitch coverage.

The 32 NSynth strata are chosen at equally spaced integer indices of the sorted encoder-training pitch support. Each draw covers all strata first, then selects the least-used eligible group. Source order is randomized within groups using seed $20260910 +$ draw index. Encoder-training and reader-development sources never enter the fitting budget.

We fit response readers and select their hyperparameters on the encoder-training action grids. We evaluate seen and unseen actions separately, using unseen actions for the principal results. Unseen actions occur in none of encoder training, reader fitting, or reader development. Known-action interfaces and spatial-factorial readers follow the same split.

Architectures and optimization. Ridge selects regularization from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$ by development MSE. The RBF extension applies a Nyström map with $\gamma = 1/d$ and $\min(512, N_{\text{fit}})$ components, where N_{fit} counts source–action rows. Poly2 instead uses degree-two polynomial features without a bias term. Both extensions fit Ridge on the mapped features with the same selection rule.

MLP-1 and MLP-2 use one and two 128-unit GELU hidden layers, respectively. The bilinear reader is $h(z) = \ell(z) + o(a(z) \odot b(z))$, with width-64 branches, affine map ℓ, and bias-free linear map o.

Neural readers use AdamW with learning rate 0.001 and weight decay 0.01. Each minibatch samples $\min(512, N_{\text{fit}})$ source–action rows with replacement from the concatenated fitting features. Regression minimizes MSE and classification minimizes cross-entropy.

Reader selection uses at most 4,000 updates, with feature statistics fitted on the assigned source draw. We check development loss every 200 updates, begin checkpoint selection at update 1,000, and stop after five checks without improvement. We then reset the reader and train for the selected number of updates on the same draw, with the same initialization and random generator. Development sources remain outside this refit.

We average reader starts within each encoder seed. Full-support MLP-2 and bilinear readers for RePs, Factorized, F+Comp, and Direct+Aux use four starts with seeds encoder seed $+ 17000 +$ start index. CompOnly and Comp+Aux also use four starts for each full-support nonlinear reader. The Identity and Wrong RePs controls use one start.

A.6 Metrics and uncertainty

For spatial predictions, let $\bar{u} = u/\max(\|u\|_2, 10^{-8})$ for both predicted and target direction vectors. Angular error in degrees is

$$\frac{180}{\pi} \arccos\big(\text{clip}(\hat{\bar{r}}^\top \bar{r}, -1, 1)\big).$$

An exactly zero prediction therefore incurs $90°$. We average over actions and windows within each recording clip, then equally across clips.

ECG uses patient-equal standardized MAE across six timing coordinates, NSynth pitch MAE in semitones, and STARSS22 clip-macro angular error over 19 clips (1,491 windows). After averaging source draws and reader starts within each encoder, the five initializations define the replication units for paired differences and pointwise 95% $t(4)$ intervals on the fixed split and action schedule.

Holm correction (Holm, 1979) is applied separately to 12 principal RePs–Factorized/F+Comp nonlinear-reader tests, 12 Identity/Wrong tests, four spatial-factorial tests, and 12 width tests. Intervals remain pointwise; native tasks, external evaluations, measurement sensitivity, and optimization studies use secondary pointwise intervals.

B Additional response-learning results

B.1 Comparison across readers

Table 7 evaluates every adapted external objective with Ridge, MLP-1, MLP-2, and bilinear readers. RePs has lower mean response error across all four reader classes in each domain, extending the MLP-2 comparison in Figure 2. Entries average reader starts within each encoder and report variation across five encoder seeds. Figure 5 visualizes the comparison with one fixed comparator per domain.

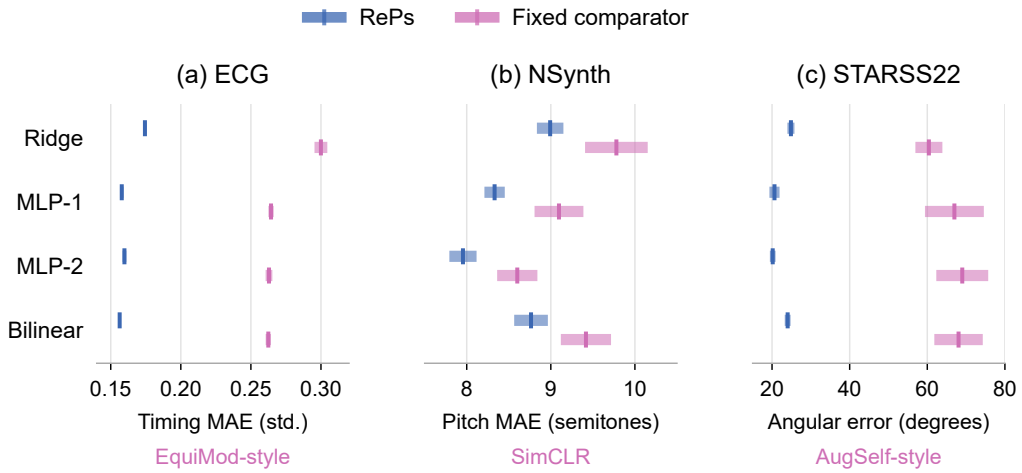


Figure 5: **Four readers with one fixed comparator per domain.** EquiMod-style (ECG), SimCLR (NSynth), and AugSelf-style (STARSS22) are selected from the principal MLP-2 comparison and held fixed across readers. Vertical marks are means; horizontal intervals are pointwise 95% intervals over five encoders. Methods are offset within rows. Domain-specific units are retained; lower is better. RePs uses source measurements and the shared-base objective.

Table 7: **Full-support readout for every evaluated external objective.** Mean ± SD over five encoder-level averages; lower is better. RePs uses source measurements and a shared contrastive base. Domain headings specify units; columns distinguish the newly fitted readers.

Objective	Ridge	MLP-1	MLP-2	Bilinear
ECG – Timing MAE (standardized)				
RePs shared-base	0.174 ± 0.001	0.158 ± 0.001	0.160 ± 0.001	0.156 ± 0.000
SimCLR	0.718 ± 0.002	0.691 ± 0.007	0.691 ± 0.007	0.696 ± 0.005
AugSelf-style	0.304 ± 0.003	0.279 ± 0.005	0.279 ± 0.004	0.279 ± 0.006
EquiMod-style	0.300 ± 0.004	0.264 ± 0.002	0.263 ± 0.002	0.262 ± 0.002
SER-style	0.711 ± 0.003	0.680 ± 0.008	0.678 ± 0.008	0.686 ± 0.009
NSynth – Pitch MAE (semitones)				
RePs shared-base	8.99 ± 0.13	8.33 ± 0.10	7.96 ± 0.13	8.76 ± 0.16
SimCLR	9.78 ± 0.30	9.10 ± 0.23	8.60 ± 0.19	9.42 ± 0.24
AugSelf-style	9.23 ± 0.14	9.03 ± 0.16	8.61 ± 0.21	9.48 ± 0.31

Table 7 (continued)

Objective	Ridge	MLP-1	MLP-2	Bilinear
STARSS22 – Angular error (degrees)				
RePs shared-base	24.87 ± 0.73	20.63 ± 1.05	20.17 ± 0.59	24.02 ± 0.61
SimCLR	66.07 ± 3.04	73.15 ± 0.87	75.11 ± 0.41	72.64 ± 1.26
AugSelf-style	60.41 ± 2.80	66.98 ± 6.11	69.01 ± 5.39	68.06 ± 5.03
CARE-style	63.78 ± 4.68	70.34 ± 1.11	72.01 ± 0.96	71.33 ± 0.66

B.2 Supervision targets and source association

The response-only ablations keep source signals, measurements, and action schedules fixed. Identity predicts source measurements from identity views; Static predicts them from transformed views. Wrong RePs uses a fixed cyclic derangement of action labels to construct targets, while Shuffled RePs permutes complete source states and their masks between sources. These controls distinguish view choice from the association between a view, its action, and its source measurement.

Table 8 collects full-support MLP-2 results, including all target-association controls. Figure 3 gives the principal fitting-budget curves; the complete budget-by-reader matrices are available in the accompanying numerical files. The full-support tests form the two separate families in Table 9; intermediate budgets are descriptive, not extra tests in these families.

Table 8: **Full-support supervision and target-association controls.** MLP-2 means over five encoder-level averages; lower is better. All objectives from the full fitting-budget grid are retained. Full support uses 84,735 ECG records, 12,349 NSynth notes in 32 pitch strata, and 871 STARSS22 windows. The complete budget grid and reader matrices are supplied as numerical files.

Objective	ECG timing MAE (std.)	NSynth (semitones)	STARSS22 (degrees)
RePs	0.158	7.502	17.977
Factorized	0.208	7.779	23.431
F+Comp	0.183	7.528	20.347
Direct+Aux	0.168	7.559	19.097
Identity	0.170	7.984	20.813
Wrong RePs	0.181	7.849	20.680
Static	0.177	7.491	22.853
Shuffled RePs	0.503	11.205	51.643
SimCLR	0.691	8.601	75.106
T-SSL	0.580	10.296	53.007
AugSelf-style	0.279	8.612	69.012

The benefit of action–target association depends on the domain and reader. On NSynth, Static and RePs have MLP-2 errors of 7.491 and 7.502 semitones. Of the six nonlinear RePs–Wrong RePs contrasts, only NSynth bilinear passes the Identity/Wrong correction. These controls distinguish target choice from access to the source measurements.

Table 9: **Full-support contrasts in two test families.** RePs minus comparator at $d = 128$ on unseen actions; negative favors RePs. Units are standardized ECG timing MAE, NSynth semitones, and STARSS22 clip-macro degrees. Intervals are pointwise paired 95% $t(4)$ intervals. Holm correction is applied separately to each labeled 12-test family.

Domain	Reader	Comparator	Difference [95% CI]	p_H
Primary family: 12 nonlinear-reader tests				
ECG	MLP-2	Factorized	-0.0499 [-0.0607, -0.0391]	0.0021
ECG	MLP-2	F+Comp	-0.0255 [-0.0483, -0.0026]	0.1461
ECG	bilinear	Factorized	-0.0551 [-0.0663, -0.0439]	0.0018
ECG	bilinear	F+Comp	-0.0260 [-0.0494, -0.0027]	0.1461
NSynth	MLP-2	Factorized	-0.2770 [-0.3781, -0.1758]	0.0129
NSynth	MLP-2	F+Comp	-0.0253 [-0.2148, +0.1642]	0.7295
NSynth	bilinear	Factorized	-0.4237 [-0.6369, -0.2105]	0.0316
NSynth	bilinear	F+Comp	-0.3530 [-0.7606, +0.0545]	0.1479
STARSS22	MLP-2	Factorized	-5.4538 [-6.4170, -4.4906]	0.0011
STARSS22	MLP-2	F+Comp	-2.3702 [-3.7899, -0.9506]	0.0488
STARSS22	bilinear	Factorized	-6.9713 [-8.6853, -5.2572]	0.0032
STARSS22	bilinear	F+Comp	-3.3940 [-4.6458, -2.1422]	0.0129
Target-control family: 12 Identity/Wrong tests				
ECG	MLP-2	Identity	-0.0117 [-0.0135, -0.0099]	0.0006
ECG	MLP-2	Wrong RePs	-0.0234 [-0.0427, -0.0041]	0.0839
ECG	bilinear	Identity	-0.0149 [-0.0161, -0.0138]	< 0.0001
ECG	bilinear	Wrong RePs	-0.0247 [-0.0463, -0.0031]	0.0839
NSynth	MLP-2	Identity	-0.4816 [-0.7945, -0.1688]	0.0611
NSynth	MLP-2	Wrong RePs	-0.3461 [-0.5673, -0.1249]	0.0611
NSynth	bilinear	Identity	-0.3514 [-0.6890, -0.0137]	0.0839
NSynth	bilinear	Wrong RePs	-0.6077 [-0.8159, -0.3995]	0.0113
STARSS22	MLP-2	Identity	-2.8361 [-3.4483, -2.2239]	0.0021
STARSS22	MLP-2	Wrong RePs	-2.7030 [-4.1797, -1.2264]	0.0566
STARSS22	bilinear	Identity	-1.6086 [-2.4942, -0.7230]	0.0566
STARSS22	bilinear	Wrong RePs	-3.3348 [-5.1677, -1.5019]	0.0566

B.3 PREDICTION PATH AND FACTOR SUPERVISION

The STARSS22 factorial separates direct versus composed prediction from the presence of true-factor targets. Its direct cells are RePs and Direct+Aux; its composed cells are CompOnly and Comp+Aux. Each uses response coefficient one, and the auxiliary cells add $\frac{1}{4}L_m + \frac{1}{4}L_g$. CompOnly learns factor-shaped outputs through response loss alone. F+Comp uses response coefficient one-half and remains a separate objective (Table 6).

For encoder seed s, let $E_s(o)$ denote objective o's response error after averaging nested fits within an interface. The path effects without and with factor targets, and their interaction, are

$$\delta_{0,s} = E_s(\text{CompOnly}) - E_s(\text{RePs}), \tag{6}$$

$$\delta_{1,s} = E_s(\text{Comp+Aux}) - E_s(\text{Direct+Aux}), \tag{7}$$

$$I_s = \delta_{1,s} - \delta_{0,s}. \tag{8}$$

A positive I_s means that factor supervision raises error more on the composed path. The four primary tests are the two MLP-2 path effects and the MLP-2 and retained-prediction interactions. Both primary interactions pass Holm correction; the secondary Ridge and bilinear interaction intervals include zero. Tables 10 and 11 report the cells and paired contrasts.

Table 10: **The four STARSS22 factorial cells.** Full-support clip-macro angular error in degrees, mean $\pm$ SD over five encoders; lower is better. All cells use response coefficient one. Columns separate newly fitted readers from the retained prediction path.

	Refitted readers			Trained path
Objective	Ridge	MLP-2	Bilinear	Retained
Direct path				
RePs	24.00 ± 1.24	17.98 ± 0.44	21.49 ± 0.66	14.59 ± 0.58
Direct+Aux	25.48 ± 0.97	19.10 ± 0.58	23.68 ± 0.46	14.90 ± 0.37
Composed path				
CompOnly	24.17 ± 0.17	18.00 ± 0.52	21.41 ± 0.46	14.74 ± 0.28
Comp+Aux	25.76 ± 0.70	20.03 ± 0.55	24.03 ± 0.96	16.56 ± 0.23

Table 11: **Primary and secondary STARSS22 factorial contrasts.** Differences in degrees with pointwise paired 95% $t(4)$ intervals over five encoders. Auxiliary cost is factor-on minus factor-off; path effect is composed minus direct. The four numeric p_H entries form one primary Holm family. Dashes mark secondary contrasts.

Interface	Contrast	Difference [95% CI]	p_H
MLP-2	Direct auxiliary cost	+1.1203 [+0.7045, +1.5361]	–
MLP-2	Composed auxiliary cost	+2.0279 [+1.4449, +2.6109]	–
MLP-2	Path without factors	+0.0203 [-0.4006, +0.4411]	0.9001
MLP-2	Path with factors	+0.9279 [+0.3231, +1.5326]	0.0392
MLP-2	interaction	+0.9076 [+0.1906, +1.6247]	0.0491
bilinear	Direct auxiliary cost	+2.1943 [+1.0403, +3.3483]	–
bilinear	Composed auxiliary cost	+2.6220 [+1.8597, +3.3843]	–
bilinear	Path without factors	-0.0790 [-1.1597, +1.0018]	–
bilinear	Path with factors	+0.3487 [-0.4963, +1.1937]	–
bilinear	interaction	+0.4277 [-0.8666, +1.7220]	–

Table 11 (continued)

Interface	Contrast	Difference [95% CI]	p_H
Ridge	Direct auxiliary cost	+1.4755 [+0.0988, +2.8521]	–
Ridge	Composed auxiliary cost	+1.5862 [+0.5834, +2.5889]	–
Ridge	Path without factors	+0.1731 [-1.5211, +1.8673]	–
Ridge	Path with factors	+0.2838 [-1.4076, +1.9752]	–
Ridge	interaction	+0.1107 [-1.9675, +2.1889]	–
retained	Direct auxiliary cost	+0.3159 [-0.3438, +0.9755]	–
retained	Composed auxiliary cost	+1.8162 [+1.4564, +2.1761]	–
retained	Path without factors	+0.1582 [-0.2777, +0.5940]	–
retained	Path with factors	+1.6586 [+1.2592, +2.0579]	–
retained	interaction	+1.5004 [+1.1585, +1.8422]	0.0010

B.4 Prediction interfaces and fitting support

The newly fitted readers h_r in Table 12 differ in access to the action. The signal-only reader predicts $h_r(z)$; the known-action reader concatenates z with the standardized true action. The state-then-law reader predicts the source state from z, restores its units, and applies the response law using the true action. Both action-supplied readers receive g during fitting, development, and evaluation; the encoder is unchanged.

Table 12: **Refitted readers with and without the true action.** Full-support MLP-2 means on unseen actions. Both action-supplied interfaces receive g in fitting, development, and evaluation. Lower is better: standardized ECG timing MAE, NSynth semitones, and STARSS22 clip-macro degrees. Companion matrices give Ridge results.

Domain	Method	Signal only $h(z)$	True action $h(z, g)$	State + law
ECG	RePs	0.1578	0.1522	0.2889
ECG	Factorized	0.2077	0.1706	0.5366
ECG	F+Comp	0.1833	0.1638	0.3947
ECG	Direct+Aux	0.1680	0.1557	0.3254
ECG	Identity	0.1695	0.1623	0.2950
ECG	Wrong RePs	0.1812	0.1686	0.4074
NSynth	RePs	7.5025	7.3627	9.6736
NSynth	Factorized	7.7795	7.6829	9.5190
NSynth	F+Comp	7.5278	7.3026	9.0095
NSynth	Direct+Aux	7.5593	7.4259	9.1920
NSynth	Identity	7.9841	7.6179	10.0349
NSynth	Wrong RePs	7.8486	7.5915	9.5150
STARSS22	RePs	17.9772	19.3062	58.4011
STARSS22	Factorized	23.4310	25.0658	37.4393
STARSS22	F+Comp	20.3474	21.8266	35.4665
STARSS22	Direct+Aux	19.0974	20.0145	35.9224
STARSS22	Identity	20.8133	22.2890	57.5052
STARSS22	Wrong RePs	20.6802	22.3651	43.1084

Concatenating the true action with the frozen features lowers error for the listed ECG and NSynth objectives but raises it for STARSS22. The state-then-law reader has higher error than the signal-only reader in all listed cases.

Retained prediction instead reuses the path trained jointly with the encoder (Table 4). F+Comp retains its composed path and Direct+Aux its direct head. These predictors use encoder-training data, whereas refitted readers use a separate fitting role. The NSynth ranking reversal therefore concerns

these complete model-use procedures, not a change of reader architecture alone. Action readout also differs from response readout: Factorized has lower action error on NSynth and STARSS22, while RePs has lower unseen-action but higher seen-action error on ECG.

The NSynth support extension uses all 44,816 fitting notes rather than 12,349 notes in 32 strata. Source count and pitch coverage change together. Table 13 reports every method under both conditions, with the held-out evaluation groups unchanged.

Table 13: **NSynth source support and pitch coverage.** MLP-2 pitch MAE in semitones, averaged over five encoders; lower is better. Source count and pitch coverage change together between the two conditions. Companion matrices contain all four readers.

Method	32 strata: 12,349 notes	All pitches: 44,816 notes
RePs	7.502	7.292
Factorized	7.779	7.629
F+Comp	7.528	7.449
Direct+Aux	7.559	7.421
Identity	7.984	7.735
Wrong RePs	7.849	7.697

C Transfer and measurement evaluation

C.1 Clinical pretraining and linear evaluation

Clinical RePs-SSL and its internal controls are pretrained separately from the response-learning configurations on 360,455 MIMIC-IV-ECG records from 87,431 patients, using the same five initialization seeds. Each final-step encoder has 219,328 parameters. Batch size 256 with incomplete batches dropped yields 1,408 updates per epoch and 14,080 over ten epochs.

Three independent random streams draw dilation from $\{0.8, 0.9, 1.1, 1.2\}$ for each source; actions may coincide. For RePs-SSL, two views enter NT-Xent at temperature 0.1 and the third enters the response branch. The response head is Linear(128,128)–ReLU–Linear(128,12); the projector uses the same hidden-layer width and activation, with output width 128. Response training uses Huber loss with threshold $\delta = 1$, averaged over observed batch–coordinate cells. All 12 coordinates must be available. Target means and population standard deviations are fitted on encoder-training records across the four actions; standard deviations at most 10^{-6} are rejected. For RePs-SSL, NT-Xent and Huber have unit coefficients, without constant-predictor risk normalization.

AdamW uses learning rate 0.001 and weight decay 10^{-4}, with epoch-wise cosine decay to 10^{-5} over ten epochs. There is no warmup or gradient clipping. FP16 autocasting starts with gradient scale 4,096; training stops if the scale decreases.

The matched SimCLR control uses the same data, initialization seeds, augmentations, optimizer, and training schedule as RePs-SSL, but optimizes NT-Xent alone. It retains the third encoder forward and auxiliary-head computation with zero auxiliary-loss weight to match the forward graph. Appendix C.5 specifies the other internal objectives.

For all clinical encoders, a linear reader is fitted on frozen features with per-dimension training standardization and standard-deviation floor 10^{-6}. AdamW uses learning rate 0.001, weight decay 10^{-4}, batch size 128, and 1,000 updates; evaluation uses the final update without validation-based early stopping. Multilabel tasks use binary cross-entropy with logits; Chapman uses four-class cross-entropy. All clinical metrics are macro-averaged over classes. We compute AUPRC using average precision and predict the Chapman class with the highest score. PTB-XL uses AUROC; Chapman uses F1.

Five repeats vary reader initialization and data order. For RePs-SSL and the internal controls, each seed also selects a separately pretrained encoder; released systems keep a single checkpoint fixed. The standard deviations of RePs-SSL and the internal controls include encoder and reader variation; released-system deviations include reader variation.

C.2 CLINICAL DATASETS AND INPUT ADAPTERS

All clinical encoders share each task's split and 10% training-label allocation, separate from the response-learning data roles. **PTB-XL Rhythm-12** uses official patient-grouped folds 1–8 for training, 9 for validation, and 10 for testing (Wagner et al., 2020); patient-level iterative sampling with seed 42 selects the 10% allocation. **CPSC2018** uses a deterministic multilabel-stratified 70/10/20% record split because reliable patient identifiers are unavailable (Liu et al., 2018). **CODE-15** uses a patient-disjoint split with 24,088 fitting records at 10% labels, 34,758 validation records, and 69,272 test records (Ribeiro et al., 2020).

Chapman uses the JS subset, the published four-class mapping (AFIB, GSVT, SB, and SR), and one record per subject (Zheng et al., 2020). Subject-stratified 70/10/20 splits use seed 20260817; seed 42 selects 10% of training records. **MIMIC cardiac labels** are drawn from MIMIC-IV-ECG-Ext-ICD v1.0.1 (Strodthoff et al., 2024a;b), with patient-disjoint downstream membership and seed 42 for the 10% allocation. We use `all_diag_all` and the fixed ALL2ALL vocabulary: codes are truncated to five characters, trailing X placeholders are removed, and three-to-five-character prefixes occurring at least 2,000 times in the source cohort are retained. Selecting Chapter IX codes from this 1,076-code vocabulary yields 158 labels. The 87,431 pretraining patients are disjoint from the 6,178 validation and 6,266 test patients.

Table 14 summarizes the released systems and the shared architecture of the internal controls. Feature widths are 512 for MERL, 768 for HuBERT-ECG, 256 for ECG-FM, 768 for ECGFM-KED, and 128 for all internally pretrained encoders. WFDB adapters restore standard twelve-lead order and documented voltage units in millivolts, then resample and take a centered crop with zero padding as needed. ECG-FM additionally standardizes each lead; none uses sliding-window aggregation.

Table 14: **Resources and inputs of the clinical systems.** Parameters count encoder weights. Input gives the adapter's nominal tensor duration and sampling rate, not necessarily its effective signal context: for MIMIC, ECGFM-KED pads an eight-second signal to ten seconds. For MIMIC, inputs staged at 100 Hz are upsampled by the 500-Hz adapters. Pretraining columns describe the four selected releases and RePs-SSL. The internal controls share the RePs-SSL backbone, input shape, and pretraining corpus.

Encoder	Params (M)	Nominal input (s / Hz)	Reported pretraining	Source
MERL	3.849	2.5 / 500	MIMIC signals + reports	Liu et al. (2024)
HuBERT-ECG	92.798	5 / 100	9.1M ECGs; named multi-dataset mixture	Coppola et al. (2024)
ECG-FM	90.883	5 / 500	MIMIC + PhysioNet 2021	McKeen et al. (2025)
ECGFM-KED	7.866	10 / 100	MIMIC + report labels	Tian et al. (2024)
RePs-SSL	0.219	8 / 100	MIMIC-IV-ECG	This work

MIMIC preprocessing stages 1,000 samples at 100 Hz, clips at ± 5 mV, and selects samples 100–899 before the released-model adapters; ECGFM-KED therefore pads eight seconds of signal to its ten-second input. For RePs-SSL and the internal controls on the other WFDB tasks, preprocessing takes a centered 800-sample crop after resampling to 100 Hz. Their CODE-15 path resamples 4,096 samples at 400 Hz to 1,024 at 100 Hz and selects samples 112–911.

The released systems differ in pretraining data and supervision: MERL and ECGFM-KED use report-derived supervision (Liu et al., 2024; Tian et al., 2024), and ECG-FM reports pretraining on MIMIC and PhysioNet 2021 (McKeen et al., 2025). HuBERT-ECG reports 9.1 million ECGs from public and access-on-demand sources, including Ribeiro/CODE, PTB-XL, CPSC, Chapman-Shaoxing, and MIMIC-IV-ECG, with labels removed for self-supervised pretraining (Coppola et al., 2024). These sources include datasets used in the clinical panel.

C.3 Cross-task transfer and manual ECG timing

We evaluate response-only encoders on native ECG timing, audio classification, localization on STARSS23, and manual ECG annotations. MLP-2 supplies the audio predictions; pitch and instrument-family classification use macro-F1, and STARSS23 localization uses angular error. Official NSynth validation data, STARSS23 (Shimada et al., 2023), and the Resting ECG Segmentation Dataset (RDB; Liu et al., 2025) are secondary evaluation sets previously examined in exploratory analyses. STARSS23 uses stable single-source windows selected by the rule in Appendix A.2, after excluding every recording identifier in the STARSS22 development metadata. For STARSS23, the fitted readers transfer without tuning; the native-localization column reports identity views of STARSS22. Tables 15 and 16 show that benefits depend on the target: Identity has lower native ECG error than RePs (0.1865 versus 0.1918), and SimCLR has higher instrument-family F1.

Table 15: **Native ECG timing from aligned encoders.** Full-support MLP-2 predicts provider measurements on identity views. Lower standardized timing MAE is better. Differences are comparator minus RePs, with secondary pointwise paired 95% intervals over five encoders.

Objective	Mean error	Δ vs RePs	95% paired CI
RePs	0.1918	+0.0000	[+0.0000, +0.0000]
Factorized	0.1892	-0.0027	[-0.0039, -0.0014]
F+Comp	0.1920	+0.0002	[-0.0028, +0.0032]
Direct+Aux	0.1920	+0.0002	[-0.0013, +0.0017]
Identity	0.1865	-0.0053	[-0.0068, -0.0039]
Static	0.1933	+0.0015	[-0.0001, +0.0031]
Wrong RePs	0.1951	+0.0033	[+0.0027, +0.0039]
Shuffled RePs	0.4891	+0.2973	[+0.2498, +0.3447]
SimCLR	0.5475	+0.3557	[+0.3511, +0.3603]
T-SSL	0.5096	+0.3178	[+0.2982, +0.3373]
AugSelf-style	0.3356	+0.1438	[+0.1372, +0.1504]

Table 16: **Audio classification and localization from frozen features.** Mean ± SD over five encoders using MLP-2. Higher macro-F1 is better for pitch and instrument family; lower angular error is better for localization. These are secondary evaluations on datasets examined in earlier analyses.

Audio: classification and localization

	Macro-F1 ↑		Angular error (°) ↓	
Method	Pitch	Family	STARSS23	STARSS22 identity
RePs	0.3206 ± 0.0076	0.2278 ± 0.0075	20.6419 ± 0.7367	20.4259 ± 0.8018
Factorized	0.2779 ± 0.0122	0.2588 ± 0.0094	28.2493 ± 1.2107	26.2287 ± 1.2213
F+Comp	0.2768 ± 0.0118	0.2566 ± 0.0093	23.7908 ± 0.4869	23.3547 ± 0.6523
Direct+Aux	0.2810 ± 0.0122	0.2614 ± 0.0112	22.5751 ± 0.8414	21.9814 ± 1.0137
Identity	0.3023 ± 0.0109	0.2348 ± 0.0062	21.6710 ± 0.7857	20.8534 ± 1.4029
SimCLR	0.2503 ± 0.0705	0.2981 ± 0.0048	79.7841 ± 1.0407	73.0735 ± 1.1335
Wrong RePs	0.2610 ± 0.0075	0.2583 ± 0.0059	24.3562 ± 1.3890	23.1362 ± 1.4432
Static	0.2651 ± 0.0114	0.2509 ± 0.0050	26.8277 ± 0.6578	26.9174 ± 3.2309
Shuffled RePs	0.1241 ± 0.0105	0.1767 ± 0.0108	49.4949 ± 1.1979	54.4601 ± 0.5463
T-SSL	0.1968 ± 0.0183	0.2112 ± 0.0101	52.6690 ± 3.2546	52.6258 ± 1.0208
AugSelf-style	0.2158 ± 0.0197	0.2961 ± 0.0071	73.0475 ± 7.9943	68.9301 ± 3.7589

RDB eligibility requires finite, nonconstant twelve-lead waveforms and complete manually annotated P–QRS–T cycles within the central eight seconds. Cycles are matched across leads by QRS midpoint within 100 ms, retaining groups with at least six leads. PR spans P-onset to QRS-onset, QRS spans

QRS-onset to QRS-offset, and QT spans QRS-onset to T-offset; QRS and QT include the endpoint sample at 500 Hz. For each interval, we take the median across leads within a beat and then across beats to form a record-level target.

Manual timing evaluation uses five-fold out-of-fold Ridge prediction on 1,378 RDB records, with fold sizes 276, 276, 276, 275, and 275. For test fold f, fold $(f + 1) \bmod 5$ provides development data and the other three folds provide fitting data. Fitting folds determine normalization; development standardized MAE selects regularization from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$. We then re-estimate normalization and refit on the four outer training folds. The split is record-disjoint because reliable patient identifiers are unavailable. Table 17 reports standardized MAE and PR, QRS, and QT errors in milliseconds.

Table 17: **Manual ECG timing on 1,378 RDB records.** Five-fold out-of-fold Ridge with fold-local normalization and nested regularization selection. Entries average five encoders. Splits are record-disjoint. Lower is better in every error column.

Method	PR (ms) ↓	QRS (ms) ↓	QT (ms) ↓	Std. MAE ↓
RePs	6.405	4.023	9.220	0.261
Factorized	6.248	4.151	9.239	0.262
F+Comp	6.333	4.098	9.180	0.262
Direct+Aux	6.238	4.049	9.151	0.259
Identity	6.934	4.276	9.493	0.277
SimCLR	18.587	8.473	27.601	0.678

C.4 Measurements recomputed on transformed views

To examine the difference between nominal provider targets and measurements of the actual view, we construct a six-coordinate target r^{remeas} with NeuroKit2 0.2.12 on lead II at 100 Hz (Makowski et al., 2021). The pipeline uses `ecg_clean`, `ecg_peaks`, and `ecg_delineate(method="dwt")`, requiring at least three R peaks. For each beat, P-onset, P-offset, QRS-onset, QRS-offset, and T-offset must be finite, ordered, and inside the 800-sample view, with a previous R peak. Identity-view beats and their previous R peaks are mapped to each transformed view and matched within five samples. Coordinatewise medians use only beats measurable at every action required for that role.

From 1,000 sources sampled per role with seed 20260911, measurement succeeds on 569 fitting, 553 development, and 402 evaluation records. Fitting and development require success at identity and the four seen actions; evaluation also requires success at all 16 unseen actions. Readers fit and select on the four nonidentity seen actions and evaluate on all 20 actions. Table 18 compares provider and remeasured targets on these same successful cohorts with encoder-training normalization fixed.

On an independent LUDB 1.0.1 check, identity-view measurement succeeds for 39 of the first 50 numeric record identifiers, with PR/QRS/QT errors of 31.70/62.13/63.94 ms on lead II at 100 Hz (Kalyakulina et al., 2021). The remeasured targets therefore provide a sensitivity analysis, not a gold standard: disagreement with provider targets can include delineation, crop, and transformation effects.

Table 18: **ECG targets on the common successful remeasurement cohort.** Evaluation covers 402/1,000 records (40.2%) and 20 actions; fitting and development contain 569 and 553 records. Both target types retain encoder-training normalization. Positive comparator-minus-RePs standardized six-timing MAE favors RePs. Intervals are secondary pointwise paired 95% intervals over five encoders.

		Comparator − RePs; positive favors RePs	
Reader	Comparator	Provider gap [95% CI]	r^{remeas} gap [95% CI]
Ridge	Factorized	+0.0592 [+0.0414, +0.0771]	+0.0015 [-0.0115, +0.0146]
MLP-2	Factorized	+0.0645 [+0.0456, +0.0834]	+0.0684 [+0.0434, +0.0934]
bilinear	Factorized	+0.0667 [+0.0455, +0.0878]	+0.1436 [+0.0892, +0.1979]
Ridge	F+Comp	+0.0271 [+0.0047, +0.0495]	+0.0031 [-0.0239, +0.0302]
MLP-2	F+Comp	+0.0373 [+0.0109, +0.0637]	+0.0472 [-0.0020, +0.0965]
bilinear	F+Comp	+0.0359 [+0.0089, +0.0630]	+0.0769 [-0.0390, +0.1928]
Ridge	Direct+Aux	+0.0100 [-0.0034, +0.0234]	-0.0022 [-0.0068, +0.0024]
MLP-2	Direct+Aux	+0.0145 [+0.0009, +0.0281]	+0.0242 [-0.0023, +0.0507]
bilinear	Direct+Aux	+0.0122 [-0.0081, +0.0325]	+0.0333 [-0.0146, +0.0812]
Ridge	Identity	+0.0260 [+0.0240, +0.0280]	-0.0034 [-0.0108, +0.0041]
MLP-2	Identity	+0.0179 [+0.0156, +0.0203]	-0.0012 [-0.0223, +0.0198]
bilinear	Identity	+0.0120 [+0.0056, +0.0183]	-0.0045 [-0.0409, +0.0319]
Ridge	Wrong RePs	+0.0207 [+0.0009, +0.0405]	+0.0084 [-0.0023, +0.0192]
MLP-2	Wrong RePs	+0.0331 [+0.0076, +0.0587]	+0.0381 [+0.0131, +0.0632]
bilinear	Wrong RePs	+0.0305 [+0.0094, +0.0516]	+0.0510 [+0.0129, +0.0891]

C.5 MATCHED CLINICAL PRETRAINING CONTROLS

The five clinical objectives share the 360,455-record corpus, ResNet1D backbone, initialization seeds, optimizer, and ten-epoch schedule in Appendix C.1. All use the same two-view NT-Xent term. SimCLR has zero auxiliary-loss weight; the other objectives add unit-weight Huber supervision on a third branch. Generic predicts an action-derived target, Identity predicts source measurements from the identity view, Static predicts source measurements from the transformed view, and RePs-SSL predicts the transformed measurements. Identity changes the auxiliary view; the remaining objectives use the same transformed third view. Target standardization is fitted for each target family using pretraining data.

Table 19 reports all five objectives under the fixed 10% clinical linear-probe protocol. The lower panel reports seed-paired RePs-SSL-minus-SimCLR differences with pointwise 95% $t(4)$ intervals on the fixed downstream splits, without multiplicity adjustment.

Table 19: **Clinical transfer under matched pretraining.** Five objectives use the same 360,455-record corpus, 0.219M-parameter ResNet1D, and ten-epoch schedule. Upper panel: mean ± sample SD across five encoder/reader seed pairs with 10% training labels; higher is better. Bold marks the highest mean among the five objectives. Lower panel: paired RePs-SSL-minus-SimCLR differences and pointwise 95% $t(4)$ intervals without multiplicity adjustment.

Objective	PTB-XL AUROC	CPSC AUPRC	Chapman F1	MIMIC AUPRC	CODE-15 AUPRC
SimCLR	0.6624 ±0.0297	0.4229 ±0.0120	0.6208 ±0.0170	0.0552 ±0.0013	0.1636 ±0.0048
Generic	0.7930 ±0.0219	0.5596 ±0.0114	0.8210 ±0.0222	0.0652 ±0.0011	0.2781 ±0.0132
Identity	0.8329 ±0.0296	0.6298 ±0.0112	**0.8774 ±0.0057**	0.0700 ±0.0012	0.3703 ±0.0120
Static	0.8259 ±0.0227	**0.6381 ±0.0086**	0.8424 ±0.0047	0.0699 ±0.0012	0.3538 ±0.0117
RePs-SSL	**0.8359 ±0.0185**	0.6358 ±0.0083	0.8730 ±0.0041	**0.0705 ±0.0006**	**0.3759 ±0.0231**
RePs-SSL minus SimCLR					
Mean difference	0.1735	0.2129	0.2521	0.0153	0.2123
95% interval	[0.1511, 0.1960]	[0.1925, 0.2333]	[0.2348, 0.2694]	[0.0137, 0.0169]	[0.1840, 0.2406]

RePs-SSL exceeds SimCLR on all five tasks, with each paired interval above zero. The measurement-supervised controls have mixed rankings: Identity has the highest Chapman mean (0.8774 versus 0.8730 for RePs-SSL), while Static has the highest CPSC mean (0.6381 versus 0.6358). This comparison supports adding measurement-based supervision to NT-Xent under the tested recipe; the advantage of the response-law target over other measurement targets depends on the clinical task.

D SENSITIVITY AND SUPPORTING ANALYSES

D.1 ENCODER AND READER TRAINING BUDGETS

The STARSS22 exposure ablation compares 40 and 80 encoder forwards per source with paired initializations, independent cosine schedules, and final checkpoints. The Factorized-minus-RePs MLP-2 gap changes from $5.45°$ to $5.17°$, and the F+Comp gap from $2.37°$ to $2.71°$. The mean ordering persists with greater encoder exposure (Figure 6a).

With the 40-forward encoders fixed, increasing the reader cap from 4,000 to 8,000 updates and patience from five to ten development checks tests sensitivity to stopping rules (Figure 6b). No fit reaches the 8,000-update cap; the largest attempted counts are 7,400 for MLP-2 and 6,400 for bilinear readers. Some fits select the same checkpoint under both rules. This extension covers STARSS22. ECG retains the 4,000-update protocol, where 12/90 MLP-2 and 18/90 bilinear fits select the cap checkpoint; these fits are nested within five encoders.

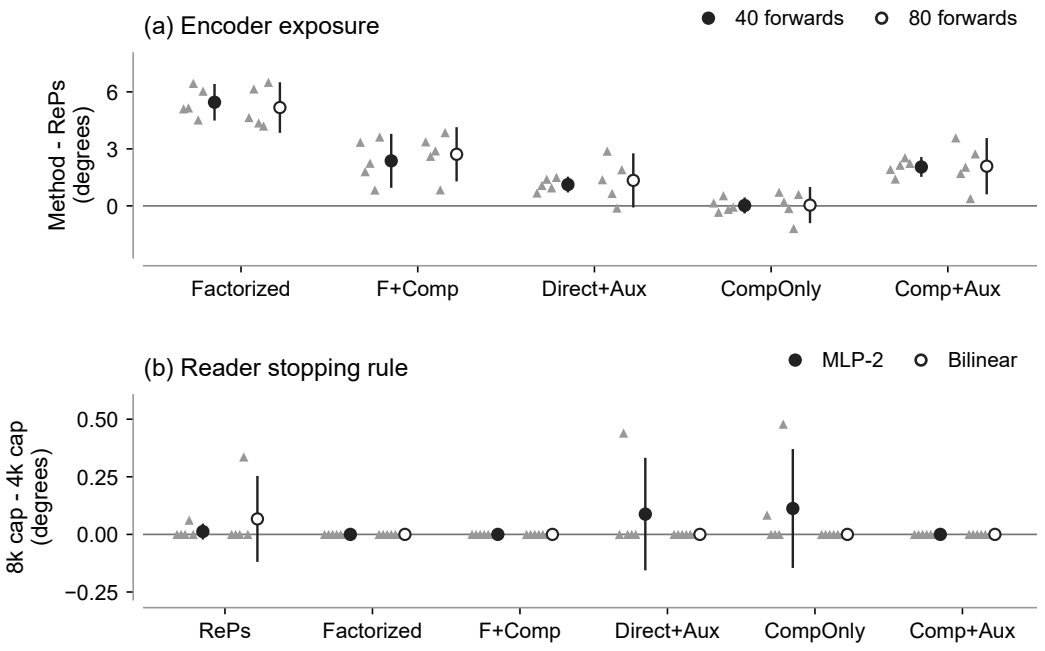


Figure 6: **Encoder exposure and reader stopping rules on STARSS22.** (a) Full-support MLP-2 method-minus-RePs error under independent 40- and 80-forward schedules. (b) Error with an 8k cap minus error with a 4k cap, using fixed 40-forward encoders and increased patience. Gray triangles are the five encoder-paired differences; circles and whiskers are means and pointwise paired 95% t_4 intervals. Panel (a) omits the defined-zero self-contrast; panel (b) retains observed zero changes.

D.2 REPRESENTATION WIDTH

The width ablation replaces 128-dimensional pooled features with a learned two-dimensional linear projection, changing both width and the presence of this projection. Table 20 reports the difference between method gaps at $d = 2$ and $d = 128$. Two STARSS22 retained-path interactions pass the 12-test Holm correction: Factorized–RePs and F+Comp–Factorized. None of the six Direct+Aux–RePs interactions passes the correction, so the width effect depends on the objective and prediction path. Table 21 gives retained and MLP-2 means for all four objectives at both widths.

Table 20: **Change in method gaps under the width ablation.** Each gap at $d = 128$ is subtracted from its gap at $d = 2$, which adds a learned projection. Units are standardized ECG timing MAE, NSynth semitones, and STARSS22 degrees. Intervals are pointwise paired 95% $t(4)$ intervals over five encoders; p_H adjusts the 12 tests together.

Domain	Interface	Within-width gap	$\Delta_2 - \Delta_{128}$ [95% CI]	p_H
ECG	retained	D+Aux - RePs	-0.0136 [-0.0261, -0.0012]	0.3854
ECG	MLP-2	D+Aux - RePs	-0.0129 [-0.0264, +0.0006]	0.4504
ECG	retained	F+Comp - Fact.	-0.1677 [-0.3267, -0.0086]	0.3864
ECG	retained	Fact. - RePs	+0.1197 [-0.0356, +0.2751]	0.6939
NSynth	retained	D+Aux - RePs	+0.4407 [-1.4038, +2.2853]	1.0000
NSynth	MLP-2	D+Aux - RePs	-0.7028 [-2.1148, +0.7092]	1.0000
NSynth	retained	F+Comp - Fact.	+0.0393 [-0.3495, +0.4281]	1.0000
NSynth	retained	Fact. - RePs	+0.2998 [-1.6653, +2.2649]	1.0000
STARSS22	retained	D+Aux - RePs	+0.0077 [-2.9674, +2.9827]	1.0000
STARSS22	MLP-2	D+Aux - RePs	-0.9272 [-3.6654, +1.8110]	1.0000
STARSS22	retained	F+Comp - Fact.	-25.2454 [-31.1586, -19.3323]	0.0032
STARSS22	retained	Fact. - RePs	+27.6803 [+22.3693, +32.9912]	0.0016

Table 21: **Response prediction at two feature widths.** Five-encoder means; lower is better. The $d = 2$ condition adds a learned projection, whereas $d = 128$ uses pooled features directly. Units are standardized ECG timing MAE, NSynth semitones, and STARSS22 degrees. Companion files contain the other readers.

		Retained		MLP-2	
Domain	Objective	$d = 2$	$d = 128$	$d = 2$	$d = 128$
ECG	RePs	0.456	0.147	0.457	0.158
ECG	Factorized	0.669	0.240	0.681	0.208
ECG	F+Comp	0.448	0.187	0.450	0.183
ECG	Direct+Aux	0.453	0.158	0.454	0.168
NSynth	RePs	8.166	7.951	7.626	7.502
NSynth	Factorized	7.299	6.784	7.296	7.779
NSynth	F+Comp	7.038	6.483	7.103	7.528
NSynth	Direct+Aux	6.900	6.244	6.980	7.559
STARSS22	RePs	22.062	14.587	23.566	17.977
STARSS22	Factorized	54.830	19.674	55.870	23.431
STARSS22	F+Comp	26.924	17.014	27.799	20.347
STARSS22	Direct+Aux	22.386	14.903	23.759	19.097

D.3 Synthetic control with visible factors

The synthetic control tests supervision choice when both generating factors are explicitly available in the input. We draw $X = (M, U)$ uniformly from $[-1, 1]^2$. For addition, $G = U$ and $R = M + G$; for multiplication, $G = \exp(U)$ and $R = MG$. A $2 \to 128 \to d$ GELU encoder uses $d \in \{1, 2, 16\}$, with response MAE evaluated over five seeds.

The four roles contain 4,096 encoder-training, 1,024 fitting, 512 development, and 2,048 evaluation examples, sampled with seed 20260910. Forty passes at batch size 256 yield 640 AdamW updates, with learning rate 0.002, weight decay 10^{-4}, and cosine decay to 10^{-5}. Branch scales and loss coefficients follow Table 6.

At $d = 1$, all three response-containing objectives have lower error than Factorized for both maps (Table 22). At larger widths, the ordering depends on the map and prediction interface.

Table 22: **Response prediction with visible synthetic factors.** MAE over five encoder seeds; lower is better. Retained uses the trained path; Ridge, RBF, and Poly2 fit new readers. Displayed 0.00000 values are rounded.

Law	d	Method	Retained	Ridge	RBF	Poly2
add	1	RePs	0.00399	0.01291	0.00200	0.00332
add	1	Factorized	0.50146	0.59630	0.49368	0.55686
add	1	F+Comp	0.00467	0.01510	0.00475	0.00607
add	1	Direct+Aux	0.00312	0.01701	0.00282	0.00432
add	2	RePs	0.00398	0.00945	0.00150	0.00227
add	2	Factorized	0.00615	0.02420	0.00098	0.00359
add	2	F+Comp	0.00593	0.02300	0.00104	0.00350
add	2	Direct+Aux	0.00487	0.01931	0.00094	0.00286
add	16	RePs	0.00268	0.00012	0.00089	0.00000
add	16	Factorized	0.00346	0.00018	0.00090	0.00000
add	16	F+Comp	0.00310	0.00013	0.00094	0.00000
add	16	Direct+Aux	0.00302	0.00014	0.00090	0.00000
multiply	1	RePs	0.03780	0.04228	0.03166	0.04197
multiply	1	Factorized	0.37940	0.47714	0.37774	0.41307
multiply	1	F+Comp	0.04516	0.05336	0.02802	0.04481
multiply	1	Direct+Aux	0.03109	0.03746	0.02523	0.03708
multiply	2	RePs	0.01715	0.07892	0.00374	0.01658
multiply	2	Factorized	0.00685	0.28675	0.00124	0.04724
multiply	2	F+Comp	0.00586	0.27822	0.00115	0.04119
multiply	2	Direct+Aux	0.01525	0.18200	0.00144	0.02187
multiply	16	RePs	0.01140	0.00576	0.00089	0.00017
multiply	16	Factorized	0.00463	0.00610	0.00109	0.00024
multiply	16	F+Comp	0.00445	0.00619	0.00119	0.00026
multiply	16	Direct+Aux	0.01270	0.00745	0.00093	0.00018

D.4 Information-matched target coordinates

We compare the bijectively related targets (r, g) and (m, g) using the same backbone, initialization, source/action stream, training budget, and action-head architecture. The value branch predicts r for Response+Action and m for Factorized. Both outputs follow a shared hidden layer of width 127 for ECG and NSynth, or 126 for STARSS22, using ReLU for ECG and GELU for audio.

Let ℓ_{value} and ℓ_{action} denote the two regression losses. ECG standardizes the value and $\log \alpha$ targets using encoder-training statistics, then applies masked Huber losses with threshold one, averaged over observed cells. Its total loss is $L_{\text{NT-Xent}} + \frac{1}{2}(\ell_{\text{value}} + \ell_{\text{action}})$: NT-Xent uses two views at temperature 0.1, and both regression branches use a third view. NSynth standardizes the value and pitch-shift targets and minimizes $\frac{1}{2}(\ell_{\text{value}} + \ell_{\text{action}})$ using MSE. STARSS22 minimizes $\frac{1}{2}(\kappa\ell_{\text{value}} + \ell_{\text{action}})$, with MSE and fixed $\kappa = 4.332372$ for both objectives. Factorized uses raw XYZ and $(\sin\phi, \cos\phi)$ targets; Response+Action standardizes the response and action coordinates using encoder-training statistics. Thus the STARSS22 comparison also changes regression scaling. None of these losses uses the constant-predictor normalization in Eq. 2.

This separate five-seed protocol uses 360,455 ECG records. Batch size/training steps are 256/14,080 for ECG, 256/24,912 for NSynth, and 32/2,944 for STARSS22; data roles and schedules differ from the aligned response-only protocol. Regression induces different metrics in the two target coordinates. With Ridge regularization $\alpha = 1$, response readout favors Factorized on ECG and Response+Action on STARSS22; the NSynth interval includes zero (Table 23). Thus information-equivalent targets need not have the same ordering across domains.

Table 23: **Information-matched target coordinates.** Response+Action minus Factorized in standardized MAE, with paired 95% intervals over five seeds; negative favors Response+Action. The pairs (r, g) and (m, g) are bijective. This separate ablation uses its own data roles, schedules, and scalers.

Domain	Ridge readout target	Δ	95% paired CI
ECG	state	-0.00766	[-0.01313, -0.00220]
ECG	action	-0.10393	[-0.13439, -0.07347]
ECG	response	+0.01291	[+0.00914, +0.01667]
NSynth	state	+0.00189	[-0.00562, +0.00940]
NSynth	action	-0.01490	[-0.03261, +0.00281]
NSynth	response	+0.00393	[-0.00776, +0.01561]
STARSS22	state	-0.00197	[-0.02097, +0.01702]
STARSS22	action	-0.00747	[-0.02937, +0.01442]
STARSS22	response	-0.01788	[-0.01960, -0.01615]