STRUCTURED TRAJECTORY SUPERVISION FOR RELIABLE PHYSIOLOGICAL TIME-SERIES REGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Many regression targets are summary statistics of an unobserved structured trajectory: the scalar label is a reduction of a higher-resolution signal the model never sees. We ask whether estimating the trajectory as the model's representation—and reducing it to the scalar afterwards—is a beneficial form of dense structured supervision, and what can be learned about a model's behaviour *from* that representation. Using wrist photoplethysmography (PPG) heart-rate (HR) estimation as the testbed, our estimator, PPI-Net, predicts a 128-point pulse-to-pulse-interval trajectory with per-point uncertainty and derives HR by inverse-variance-weighted aggregation. On an *identical* backbone, replacing a scalar head by the trajectory head lowers participant-level MAE from 5.11 to 4.04 BPM under leave-one-subject-out evaluation (paired -1.07 BPM, better on 13/15 subjects, $P{=}.010$)—three times a re-run noise floor of $\pm$0.34 BPM that we measure and treat as the threshold of interpretability. The representation also supports mechanistic analysis that a scalar cannot: the signed error decomposes into shrinkage toward the cohort mean under subject-level shift (unbiased point $\approx$92 BPM, slope -0.18), and the delta-method uncertainty inherits a $\hat{h}^2$ factor that inverts its ranking in the high-HR stratum, which we derive and then verify. Under one corrected protocol, PPI-Net outperforms a standard 1D-CNN and a classical adaptive filter and is on par with the strongest estimator we could retrain from public code, BeliefPPG (3.64 vs. 3.83$\pm$0.20; $P{=}.17$), with a complementary error profile; the shared failure reproduces in three held-out cohorts, and a calibration–test-separated confidence score turns it into selective regression (MAE $4.04{\rightarrow}1.43$ at 50% coverage). We state explicitly what the evidence does not support, including a sensor ablation that reverses our own earlier conclusion.

1 INTRODUCTION

Consider regression from a signal window $\mathbf{x}$ to a scalar y that is not a primitive measurement but a reduction $y = g(\mathbf{z})$ of an unobserved structured trajectory $\mathbf{z} = (z_1, \ldots, z_T)$: an average over a sub-interval, a rate, a count. The standard recipe fits $f_\theta : \mathbf{x} \mapsto \hat{y}$ directly. When $\mathbf{z}$ is observable at training time, an alternative is to fit $f_\theta : \mathbf{x} \mapsto \hat{\mathbf{z}}$ and apply g. We call the second option *structured trajectory supervision*. It changes two things at once: the learning signal (many targets per window instead of one) and the *representation* the model exposes (a trajectory instead of a number). This paper is about both: whether the change helps, and what the representation lets us conclude about how and why the model fails.

Our testbed is heart-rate (HR) estimation from wrist photoplethysmography (PPG), where the structure is physiological. The reference HR of a window is, up to labelling conventions, the reciprocal of the mean pulse-to-pulse interval (PPI) over its central seconds, and the PPI trajectory is recoverable from a simultaneously recorded ECG. Our estimator, PPI-Net, predicts a 128-point PPI trajectory with a heteroscedastic scale per point and reduces it to HR by inverse-variance weighting (Section 3). Three mechanisms distinguish this from a change of output format: (i) *denser supervision*—128 targets per window on the quantity that generates the label, in the spirit of deep and auxiliary supervision (Lee et al., 2015; Szegedy et al., 2015; Caruana, 1997); (ii) a *physiological prior*—successive intervals vary little within the central seconds of a window (we measure the ir-

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.

reducible reference gap at 0.46 BPM), which a smoothness penalty exploits as a soft constraint; (iii) *selective aggregation*—the per-point scale down-weights corrupted segments when the scalar is formed.

The representation is what makes the second half of the paper possible. Because the model exposes a physiological trajectory and a per-point scale, we can *derive* how its uncertainty must behave under the reduction (a delta-method factor of $\hat{h}^2$), predict where that behaviour must fail (the high-HR stratum), and verify the prediction; and we can decompose the signed error and show that the dominant failure under subject-level shift is shrinkage toward the cohort mean (Galton, 1886), shared by every method we test. None of this is available from a scalar output.

The credibility of both halves rests on evaluation. We organise the experiments as a *claims ladder* in which each experiment isolates one question, interpret no paired effect smaller than a re-run noise floor that we measure ($\pm$0.34 BPM, fold level), and treat the participant as the unit of inference. Along the way we identified three defects in our own earlier pipeline—a pooling stage that collapsed trajectory resolution, a $+4$ s misalignment between labels and windows, and model selection on the test subject—and we quantify each rather than fix it silently. One consequence is instructive: on the defective architecture, a sensor ablation had suggested that accelerometry was essential; on the corrected pipeline the same ablation is null. An ablation on a defective architecture measures the defect, not the component.

Claims. (a) Trajectory supervision beats scalar supervision on an identical backbone (-1.07 BPM, 13/15, $P{=}.010$); (b) the estimator beats a standard 1D-CNN and a classical adaptive filter on one corrected protocol; (c) it is on par with BeliefPPG (Bieri et al., 2023), the strongest estimator we could retrain on identical targets, with a complementary error profile (more recent methods are discussed, not retrained); (d) its accuracy does not depend on the accelerometer; (e) the dominant failure is shrinkage toward the cohort mean, shared by all tested methods and derivable from the representation; (f) a model-agnostic, calibration–test-separated confidence score recovers reliability by withholding. We do *not* claim superiority over BeliefPPG, superiority of the propagated uncertainty, or any benefit of accelerometer fusion.

2 Related work

Dense, auxiliary, and intermediate supervision. Adding objectives at intermediate representations improves optimisation and generalisation (Lee et al., 2015; Szegedy et al., 2015), and auxiliary tasks act as inductive bias when they share structure with the main task (Caruana, 1997). Representation learning has long argued that the *form* of what a model exposes matters as much as its accuracy (Bengio et al., 2013). Trajectory supervision is a particular case in which the auxiliary target is the structured quantity whose reduction *is* the label, so the two objectives cannot conflict, and the exposed representation is physically interpretable by construction. Our contribution is to isolate its effect on a parameter-matched backbone and to use the representation for mechanistic analysis.

Selective prediction and the reject option. Abstaining on low-confidence inputs trades coverage for risk (Chow, 1970; El-Yaniv & Wiener, 2010); for deep networks, selective classification and integrated reject heads make the trade-off explicit (Geifman & El-Yaniv, 2017; 2019), and conformal methods give distribution-free guarantees (Vovk et al., 2005; Angelopoulos & Bates, 2021). Selective *regression* for physiological signals is less studied; we use risk–coverage curves with strict calibration–test separation at the subject level and a confidence score that does not depend on the estimator.

Uncertainty in regression. Heteroscedastic heads model input-dependent noise (Kendall & Gal, 2017), but negative-log-likelihood training can degrade accuracy (Seitzer et al., 2022); MC-dropout (Gal & Ghahramani, 2016) and ensembles (Lakshminarayanan et al., 2017) estimate epistemic uncertainty, and calibration is imperfect (Guo et al., 2017). Less discussed is what happens to a per-point uncertainty passed through a non-linear reduction; we show that the delta method makes it inherit the prediction itself.

Group-level distribution shift. Leave-one-subject-out evaluation is a group shift (Koh et al., 2021): the held-out participant is a group unseen in training. Domain-adversarial training (Ganin

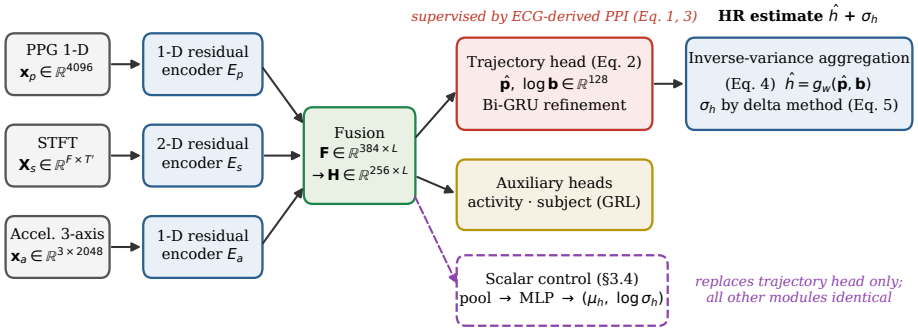


Figure 1: PPI-Net. Three encoders map the PPG, its spectrogram, and accelerometry to a common length L=128; the fused representation drives a trajectory head that outputs a 128-point PPI trajectory $\hat{\mathbf{p}}$ and per-point scales $\mathbf{b}$, refined by a bidirectional GRU. HR is the inverse-variance-weighted reduction of the central sub-trajectory (Eq. 4). The scalar control (dashed) replaces only the trajectory head; every other module is identical.

& Lempitsky, 2015) and group-robust optimisation (Sagawa et al., 2020) target invariance or worst-group performance. Our analysis identifies the dominant failure under such shift for regression—shrinkage of the held-out group's predictions toward the training cohort's mean—and shows it is shared across architectures, including a classical non-learned estimator.

Wearable PPG heart-rate estimation. Classical pipelines cancel motion with accelerometry and track spectral peaks (Zhang et al., 2015; Temko, 2017). Deep learning on the PPG-DaLiA benchmark was established by Reiss et al. (2019); subsequent architectures include CorNET (Biswas et al., 2019). BeliefPPG (Bieri et al., 2023) discretises HR, models it as a hidden Markov process, and refines per-window posteriors by temporal belief propagation, reporting a leave-one-subject-out MAE of 3.57 BPM with calibrated uncertainty; we reproduce it from its public code. Two recent methods report lower numbers on the same benchmark under the 8-s/2-s protocol: KID-PPG (Kechris et al., 2025), whose often-quoted 2.85 BPM is the error *after retention* by its error classifier (10-BPM threshold; 27% and 53% of two participants' samples rejected) while its best full-coverage model reports 3.70 BPM, and EnhancePPG (Benfenati et al., 2024) (3.54 BPM, self-supervised pretraining on an unreleased corpus). Both use the dataset's native labels and windowing, which differ from ours (Section 4.1), so we quote them as context and make no ranking claim. Evaluation is also moving to unstructured recordings—WildPPG (Meier et al., 2024) and GalaxyPPG (Park et al., 2025)—on which laboratory-trained estimators degrade; our cohorts are controlled, and we bound our claims accordingly. All of these methods regress a scalar HR; the beat-interval sequence, the object of heart-rate-variability analysis (Task Force of the ESC and NASPE, 1996), has not been used as the supervised representation.

3 Method

3.1 Problem statement

Let $\mathbf{x}$ be a 64-s multichannel window and $\mathbf{z} \in \mathbb{R}^T$ the pulse-to-pulse-interval trajectory over the window, obtained at training time from the R-peaks of a simultaneous ECG and resampled to T=128 points. The reference label is a reduction of its central portion,

$$y = g(\mathbf{z}) = \frac{60}{\mathrm{mean}_{t \in \mathcal{C}}\ z_t} \quad [\mathrm{BPM}], \qquad \mathcal{C} = \{57, \ldots, 72\}, \tag{1}$$

the 16 points covering the central 8 s. Scalar supervision fits $\hat{y} = f_\theta(\mathbf{x})$ against y; trajectory supervision fits $\hat{\mathbf{z}} = f_\theta(\mathbf{x})$ against $\mathbf{z}$ and reports $g_w(\hat{\mathbf{z}})$, a weighted variant of g defined below. The label convention matters: we found that the dataset's reference HR corresponds to the interval 4 s *after* the window centre and we correct for it (Section 4.1); Eq. 1 is stated for the corrected alignment.

3.2 ARCHITECTURE

Three residual encoders (Figure 1) map the 1-D PPG at 64 Hz ($\mathbf{x}_p \in \mathbb{R}^{4096}$), its STFT magnitude ($\mathbf{X}_s \in \mathbb{R}^{F\times T'}$), and tri-axial accelerometry ($\mathbf{x}_a \in \mathbb{R}^{3\times 2048}$) to features $\mathbf{F}_p, \mathbf{F}_s, \mathbf{F}_a \in \mathbb{R}^{128\times L}$ with a common temporal length $L{=}128$ (the natural output length of the PPG encoder; an earlier version pooled to 32 and replicated, which collapsed the trajectory's resolution—Section 4.1). Residual convolutions fuse the concatenation into $\mathbf{H} \in \mathbb{R}^{256\times L}$. The trajectory head is a $1{\times}1$ convolution producing a coarse trajectory and a log-scale per point, refined by a residual bidirectional GRU,

$$(\hat{\mathbf{p}}^{(0)},\, \log \mathbf{b}) = \mathrm{Conv}_{1\times 1}(\mathbf{H}), \qquad \hat{\mathbf{p}} = \hat{\mathbf{p}}^{(0)} + \mathrm{FC}\big(\mathrm{BiGRU}(\mathbf{H})\big), \tag{2}$$

with $\hat{\mathbf{p}} \in \mathbb{R}^{128}$ in seconds and $\mathbf{b} \in \mathbb{R}^{128}_{>0}$ the scale of a Laplace observation model per point. An activity-conditioned gate can rescale $\hat{\mathbf{p}}$ by a factor in $[0.85, 1.15]$ from predicted activity probabilities (it self-deactivates under its regulariser; Section 4.3). Two auxiliary heads on the pooled representation predict activity class and, through a gradient-reversal layer (Ganin & Lempitsky, 2015), subject identity.

3.3 SUPERVISION AND AGGREGATION

The trajectory loss is a smooth-ℓ_1 term on $\hat{\mathbf{p}}$ against $\mathbf{z}$, a smoothness penalty encoding the physiological prior, and a residual-regression term that fits $\mathbf{b}$ to the magnitude of the detached residual (the negative-log-likelihood alternative degraded accuracy in pilot folds, consistent with Seitzer et al. (2022)):

$$\mathcal{L}_{\mathrm{traj}} = \sum_t \ell_1^{\delta}(\hat{p}_t - z_t) \;+\; \lambda_s \sum_t (\hat{p}_{t+1} - \hat{p}_t)^2 \;+\; \lambda_b \sum_t \ell_1^{\delta}\big(b_t - |\hat{p}_t - z_t|_{\mathrm{sg}}\big), \tag{3}$$

where sg denotes stop-gradient. The total loss adds an HR-level term $\ell_1^{\delta}(g_w(\hat{\mathbf{p}}, \mathbf{b}) - y)$ computed with the same reduction used at inference, a gate regulariser $\lambda_g|\log s|$ on the gate scale s, and the auxiliary activity and adversarial-subject cross-entropies. The reduction is an inverse-variance-weighted mean over the central points,

$$g_w(\hat{\mathbf{p}}, \mathbf{b}) \;=\; \frac{60}{\bar{p}}, \qquad \bar{p} = \frac{\sum_{t\in\mathcal{C}} w_t\, \hat{p}_t}{\sum_{t\in\mathcal{C}} w_t}, \qquad w_t = b_t^{-2}, \tag{4}$$

and its uncertainty follows from the delta method,

$$\sigma_{\bar{p}}^2 = \Big(\sum_{t\in\mathcal{C}} w_t\Big)^{-1}, \qquad \sigma_h \;=\; \frac{60}{\bar{p}^2}\,\sigma_{\bar{p}} \;=\; \frac{\hat{h}^2}{60}\,\sigma_{\bar{p}}, \tag{5}$$

which Section 4.4 uses: whatever $\sigma_{\bar{p}}$ does, σ_h inherits a factor $\hat{h}^2$. Two caveats: b_t is a Laplace scale (variance $2b_t^2$), so $\sigma_{\bar{p}}^2 = 2(\sum_t w_t)^{-1}$—the constant cancels in the weights and is absorbed into a per-fold multiplier fitted on the validation participant; and Eq. 5 assumes independent per-point errors, which the GRU and the smoothness penalty violate (adjacent errors are positively correlated), so the aggregate variance is underestimated—consistently, the nominal 68% interval covers 63% of held-out errors. Both, with the $\hat{h}^2$ factor, motivate the model-agnostic score of Section 4.5. Using the same operator in training and inference removed a train/test mismatch of our earlier version (mean in training, median at inference), and the weighting improved MAE by 0.46 BPM over the unweighted mean in pilot folds.

Implementation details (target interpolation, STFT, gate mapping, loss weights, pooling, and the SQI and motion definitions) are given in Appendix **??**.

3.4 SCALAR-HEAD CONTROL

To isolate the trajectory target we instantiate the identical network, delete the trajectory head and its refinement, and attach a scalar head, $(\mu_h, \log \sigma_h) = \mathrm{MLP}(\mathrm{pool}(\mathbf{H}))$, trained with the same smooth-ℓ_1 HR term, the same residual-regression term for σ_h, and the same auxiliary losses. Every shared module matches parameter-for-parameter (encoders 297,472 / 185,856 / 83,840; fusion 1,609,472; auxiliary heads 33,412 / 34,960); only the head differs (105,666 vs. 33,154 parameters). The control

Table 1: Claims, the experiment that isolates each, and the evidence (PPG-DaLiA, 15 participants, corrected protocol, paired Wilcoxon over participants). Fold-level re-run noise floor ± 0.34 BPM.

Claim	What isolates it	Result
Trajectory target beats scalar regression	Identical backbone, data, seed, early stopping; head only differs	-1.07 BPM, 13/15, $P{=}.010$ ($3\times$ floor)
Beats standard deep and classical practice	1D-CNN retrained; adaptive filter re-evaluated; same targets	-2.76 (14/15, $P{=}.0003$); -25.9 (15/15, $P{=}.0001$)
On par with state of the art, complementary	BeliefPPG retrained from public code, native pipeline, same targets	$+0.40$ n.s. ($P{=}.28$); better on 12/15 subjects (descriptive), worse on the 2 hardest
Accuracy does not need the accelerometer	ACC zeroed, retrained; zone-wise paired	-0.52 n.s. ($P{=}.93$); null in every zone
Failure is shrinkage to the cohort mean, shared	Signed-error decomposition; HR bands; mixed effects; 5 methods; 3 held-out cohorts	$r = -0.73$ across subjects; 4.3-fold across HR; reproduces in 3/3 cohorts
Propagated σ inverts at high HR	Derived from Eq. 5; verified by stratified ranking	$\rho(\sigma_h, \hat{h}) \geq 0.95$; selective efficiency -0.81 above 110 BPM
Reliability is recoverable by withholding	Confidence thresholds calibrated on other participants	Attainment $89.6 \rightarrow 97.2\%$; AURC 2.11 vs. 4.04 random

receives a heteroscedastic head *and* MC-dropout (Gal & Ghahramani, 2016) at inference, so that its uncertainty channel is at least as capable as ours and the comparison is not won by handicapping the competitor. One confound is disclosed: the control receives no beat-interval supervision, so the comparison changes the output structure and the supervision richness together; it isolates the value of the trajectory design, not of the head in isolation.

4 Experiments

We first fix the protocol and the defects we corrected in it (Section 4.1), then climb the claims ladder: the isolated trajectory-versus-scalar comparison and the full six-method table (Section 4.2); component ablations (Section 4.3); the mechanistic analysis that the representation makes possible—signed-error shrinkage and the uncertainty inversion, verified in-cohort and under cohort shift (Section 4.4); and selective regression (Section 4.5). Table 1 summarises each claim, the experiment that isolates it, and the result against the noise floor.

4.1 Protocol, corrected defects, and the noise floor

Data. PPG-DaLiA (Reiss et al., 2019): 15 participants, wrist PPG with chest-ECG reference during daily activities; 64-s windows with 8-s stride (16,076 windows; reference HR median 85 BPM, IQR 74–102; Appendix **??**). **Splits.** Leave-one-subject-out: the test participant is held out, one further participant is held out for early stopping, and the rest train. **Inference unit.** Windows are nested within participants, so window-level pooling is pseudoreplication (Hurlbert, 1984); we report participant-level MAE, paired Wilcoxon tests over participants, subject-clustered bootstrap intervals, and Fisher-z aggregation of within-participant correlations.

Corrected defects. (a) *Resolution collapse.* A pooling stage in our earlier head replicated a 32-point representation four-fold; a numerical-rank probe on random inputs gave rank exactly 32 and only 4 distinct values among the 16 central points. The corrected head (Eq. 2) restores 16 distinct values (rank 64). (b) *Label alignment.* An alignment probe that shifts the label grid by δ and evaluates $|y_\delta - g(\mathbf{z})|$ found its minimum at $\delta{=}-4$ s independently in every probed participant (2.47/2.20/2.71 BPM at $\delta{=}0$ vs. 0.48/0.33/0.46 at -4 s): the reference HR had been that of an interval 4 s later than the window centre. Correcting it reduced the irreducible reference gap from 2.61 to 0.46 BPM. It changes the evaluation target for every method, so corrected and legacy numbers are never mixed. (c) *Model selection.* Our earlier pipeline selected checkpoints on the test participant. Logging the test curve without using it, the optimism was $+0.125$ BPM over three folds—below the noise floor, so earlier numbers are not attributed to leakage; the corrected protocol nevertheless selects on the inner validation participant.

Noise floor. Two identical runs (same configuration and seed; GPU non-determinism) differed by 0.34 BPM on average at the fold level (maximum 0.95); across three seeds the 15-fold mean had SD 0.20. We report the primary result as mean$\pm$SD over seeds and interpret no paired effect below the fold-level floor.

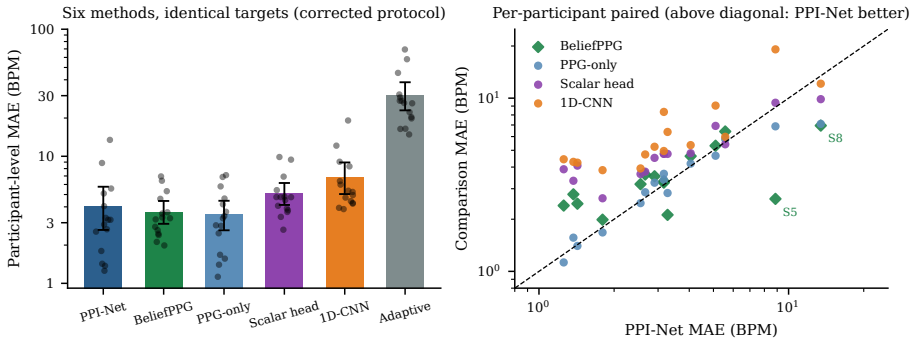


Figure 2: Six methods on identical evaluation targets and LoSo splits. Left: participant-level MAE (log scale) with subject-clustered bootstrap 95% CIs and per-participant points. Right: per-participant paired comparison against PPI-Net; BeliefPPG (diamonds) lies above the diagonal for 12 participants and far below it for S5 and S8.

4.2 Trajectory versus scalar supervision, and the six-method comparison

Under the corrected protocol (seed 42) PPI-Net attains participant-level MAE 4.04 BPM; across three seeds with HR-aware oversampling, 3.83±0.20. The scalar-head control (Section 3.4) attains 5.11: a paired difference of -1.07 BPM, PPI-Net better in 13/15 participants, Wilcoxon $P{=}.010$, three times the noise floor (Figure 2). Selective accuracy favours the trajectory model likewise (MAE at 50% coverage 2.35 vs. 3.10, $P{=}.010$). The gain is not bought with a weaker confidence channel: the control's learned σ_h ranks errors at least as well as our propagated one (selective efficiency 0.49 vs. 0.43, $P{=}.93$), and its MC-dropout channel is markedly worse than its own heteroscedastic head (0.32 vs. 0.49, 15/15 folds, $P{=}.0001$).

A standard 1D-CNN (PPG and accelerometry as four input channels, smooth-ℓ_1 on HR, same splits and early-stopping rule) attains 6.80±4.11 (-2.76, 14/15, $P{=}.0003$; the exception is the cohort's least stable fold). A training-free adaptive filter (per-axis NLMS cancellation, harmonic-aware spectral peak selection with temporal continuity) attains 29.98±15.73 (15/15, $P{=}.0001$); its two worst folds are the two highest-HR participants, where spectral tracking loses lock.

State of the art. BeliefPPG was retrained from the authors' public code on the same folds with its native pipeline (8-s frames, 2-s stride, temporal belief propagation over the recording; default hyperparameters; no pretrained weights, which would leak DaLiA). Its native-protocol MAE reproduced the published value (3.62 vs. 3.57). To compare on identical targets, its per-frame estimates were mapped onto the central 8 s of each of our windows; the mapping is deterministic and was verified by exact agreement of the two *label* sequences (offset $+6$ frames, residual 0.00 BPM in all 15 participants; Appendix **??**). It attains 3.64±1.53. The aggregate difference from PPI-Net is not significant in either direction ($P{=}.28$ vs. 4.04; $P{=}.17$ vs. 3.83±0.20), but the paired structure is informative: PPI-Net is more accurate on 12 of 15 participants, and on the 13 participants other than S5 and S8 by $+0.52$ BPM on average (12/13)—reported descriptively, since the subset is defined by the two outliers and a test on it would be post hoc—whereas BeliefPPG is far better on those two—the highest-HR participant (2.6 vs. 8.9) and the least stable fold (6.9 vs. 13.5)—which Section 4.4 identifies as PPI-Net's shrinkage failures. Its between-participant spread is correspondingly smaller (SD 1.53 vs. 3.27). Three caveats: BeliefPPG uses temporal context across the whole recording, which our single-window estimator does not; it was run at one seed against our three, an asymmetry that leaves the 0.2-BPM aggregate difference within the noise floor either way; and the more recent KID-PPG and EnhancePPG (Section 2) were not retrained on our targets. We therefore claim parity with the strongest method we could reproduce, with complementary failure profiles, not superiority over the field.

4.3 Component ablations

All ablations retrain under the corrected protocol (seed 42) and are paired against PPI-Net (4.04). **Accelerometer input.** Zeroing the accelerometer channels and retraining (architecture unchanged;

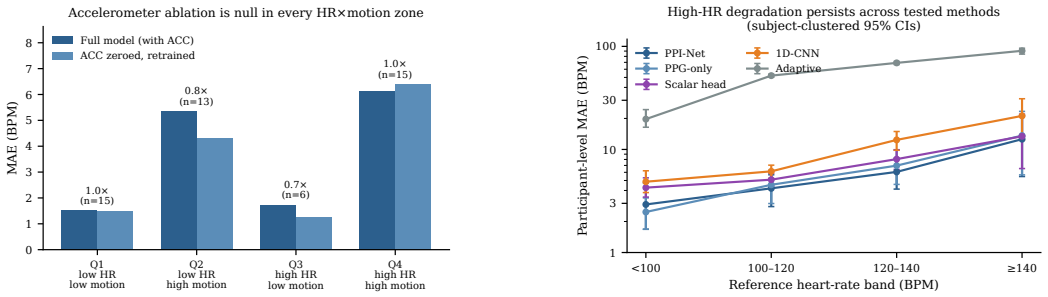


Figure 3: Left: accelerometer ablation by HR×motion zone (per-participant median motion split; paired; min. 5 windows per cell). Right: participant-level MAE by reference HR band for five methods (log scale; subject-clustered 95% CIs).

motion metadata retained for evaluation) gives 3.52±1.90—statistically indistinguishable (P=.93): 13/15 folds lie within ±0.5 BPM of parity (mean +0.04) and the aggregate difference is produced entirely by the two least stable folds, which favour the ablated variant. With each participant split at their own median motion, the ablation is null in every HR×motion cell (error ratios 0.7–1.0; Figure 3, left), so the parity is genuine rather than a cancellation. On our earlier, resolution-collapsed head the same ablation had suggested a two-fold degradation. "Null" deserves one qualification: S8's −6.35 BPM is within its seed SD (2.5), but S5's −1.98 BPM is an order of magnitude beyond its seed SD (0.15), so accelerometry may actively hinder on the highest-HR participant—a single-fold, single-seed observation flagged for multi-seed confirmation. **Instance normalisation.** Removing input instance normalisation shifts the mean by +0.39 BPM with no consistent direction (worse in 8/15, P=.36); the shift concentrates in the two highest-HR folds (+4.6, +1.9), which we note as suggestive only. **Oversampling.** HR-aware oversampling is null at full scale (−0.24, P=.30). **Gate.** The activity gate self-deactivates under its regulariser (mean absolute deviation of $\log s$: 0.017 against an allowed 0.15), so its contribution is bounded by that measurement. **Adversarial subject invariance.** Removing the gradient-reversal branch and retraining leaves accuracy unchanged (4.03 vs. 4.04 BPM; −0.01, worse in only 6/15 folds, P=.72)—the cleanest null of the set; the one fold that moved by more than 1 BPM (S15, +1.39) stopped at epoch 5 and is an early-stopping instability rather than a GRL effect. **Reduced screen.** A 5-participant screen (80 epochs) showed that removing GRU refinement or the HR-level loss collapses training (MAE 24.84 and 24.41 vs. 4.89), while oversampling and gating had mixed effects (4.61–5.27). The picture is coherent: trajectory supervision carries the improvement, recurrent refinement and the HR loss are load-bearing, and the auxiliary mechanisms—accelerometer input, instance normalisation, oversampling, gating, and adversarial subject invariance—are not individually essential at the measured floor (per-fold tables in Appendix **??**).

4.4 From the representation to the mechanism of failure

Where the error is. Participant-level MAE rises monotonically with reference HR in every method (Figure 3, right): PPI-Net 2.93→12.58 BPM from <100 to ≥140 (4.3-fold), PPG-only 2.47→13.63, scalar 4.26→13.41, CNN 4.86→21.16, adaptive 19.68→90.56. Within-participant Spearman correlations are positive for every method for HR (ρ 0.24–0.80, positive in 13–15 of 15 participants, all $P \leq .0006$) and motion (ρ 0.28–0.69, all $P \leq .0001$). A mixed-effects model on $\log(1+\text{error})$ for PPI-Net (random intercept and motion slope per participant) gives HR β=+0.153 (95% CI 0.131–0.174) and motion β=+0.265 (0.126–0.405) with a negligible interaction (−0.010, P=.22); adjusting for a signal-quality index attenuates but does not remove either effect (HR +0.107, $P<10^{-21}$; motion +0.220, P=.002).

What the error is: shrinkage toward the cohort mean. The signed error tells the mechanism (Figure 4). Let μ_i be participant i's mean reference HR and $\bar{e}_i$ their mean signed error. Across participants, $\bar{e}_i$ is a decreasing linear function of μ_i,

$$\bar{e}_i \;\approx\; -\lambda\,(\mu_i - \mu_0), \qquad \hat{\lambda} = 0.18, \quad \hat{\mu}_0 = 92.6\,\text{BPM}, \quad r = -0.73\ (P{=}.002), \tag{6}$$

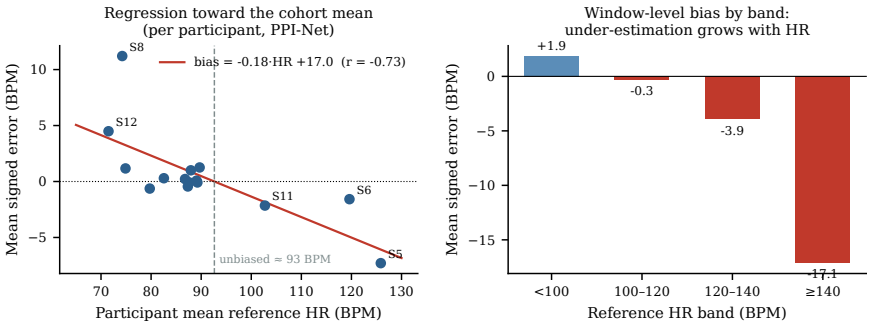


Figure 4: Signed error reveals shrinkage toward the cohort mean (PPI-Net, corrected protocol). Left: per-participant mean signed error against mean reference HR, with the fitted line (Eq. 6) and its unbiased point. Right: window-level mean signed error by reference band.

where μ_0 coincides with the cohort centre: high-HR participants are under-estimated and low-HR participants over-estimated, and the three worst folds owe 82–88% of their MAE to this bias alone. At the window level the bias is $+1.9$ BPM below 100 and -17.1 BPM at $\geq$140, negative in 11/15 participants above 110 BPM. Eq. 6 is regression toward the mean (Galton, 1886) induced by group shift: the held-out participant is a group whose marginal the model never saw, and the estimator shrinks toward the marginal it did see. The same signature appears in the classical filter (its two worst participants are the two highest-HR ones) and in the CNN, so it is a property of the task under leave-one-subject-out evaluation, not of one architecture. It also explains the complementarity with BeliefPPG: a recording-level temporal prior fitted to the training cohort but propagated through the test recording resists exactly this shrinkage on outlying participants.

What the representation predicts about its own uncertainty. Eq. 5 states $\sigma_h \propto \hat{h}^2 \sigma_{\bar{p}}$. If the learned per-point scales vary little between windows—which we find: the between-window coefficient of variation of $\bar{b}$ is only 0.02–0.12 per fold, although its within-window variation (0.48–0.69) is what makes the weighting in Eq. 4 useful—then σ_h is dominated by $\hat{h}^2$ and the derivation predicts $\rho(\sigma_h, \hat{h}) \to 1$. Observed: 0.957–1.000 in every fold. The derivation further predicts a sign reversal: in the high-HR stratum, shrinkage (Eq. 6) makes the *most* under-estimated windows have the *smallest* $\hat{h}$ and hence the smallest σ_h, so ranking by σ_h should be actively harmful there and helpful where $\hat{h}$ and error are positively associated. Observed: the selective efficiency of σ_h is -0.81 above 110 BPM (negative in 12/13 folds) and $+0.48$ below (positive in 15/15). Removing the propagation factor, $\sigma_h \cdot 60/\hat{h}^2$, restores a positive ranking in the high stratum ($+0.46$), confirming that the learned head carries signal that the reduction obscured. This chain—derivation from the representation, prediction, verification—is not available from a scalar estimator, whose uncertainty has no algebraic relation to its prediction; it is also why Section 4.5 uses a model-agnostic confidence score rather than σ_h.

Reproduction under cohort shift. A secondary leave-one-cohort-out run on PPG-DaLiA, WESAD, and BAMI-1 (native labels, pre-correction pipeline; direction only, since a misaligned target can itself induce shrinkage) found the HR–error association positive in 13/15, 12/15, and 24/24 held-out participants (Appendix **??**).

4.5 Selective regression with calibration–test separation

Because the failure is condition-dependent, it is predictable from observable conditions. We use a model-agnostic confidence score, $c = \mathrm{SQI}/(1+\mathrm{motion})$, and choose the withholding threshold τ_α as the α-quantile of c on the *other* 14 participants before applying it to the held-out one, so no test data influence any threshold. Selective risk at coverage α is the MAE over retained windows, $R(\alpha) = \mathbb{E}\big[\,|\hat{h} - y| \,\big|\, c \geq \tau_\alpha\big]$, and AURC is its integral over α (Geifman & El-Yaniv, 2017). Withholding low-confidence estimates raises held-out criterion attainment (within 5 BPM or 10% of reference) from 89.6% to 96.4% at 40% withheld and 97.2% at 50%, lowers MAE from 4.04 to 1.43 BPM, and gives AURC 2.11 BPM against 4.04 for random withholding at matched coverage—

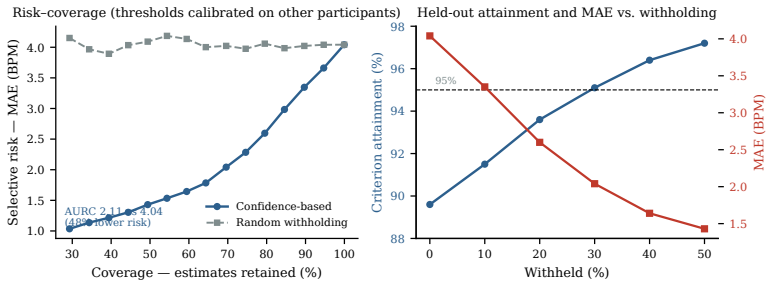


Figure 5: Selective regression with calibration–test separation. Left: risk–coverage against random withholding at matched coverage. Right: held-out criterion attainment and MAE as low-confidence estimates are withheld.

48% lower selective risk (Figure 5). The score works for the other estimators too (selective efficiency positive in 15/15 folds for the CNN), which is a feature: the reliability mechanism does not depend on the estimator that produced the value.

5 Discussion

Scope. The evidence supports claims (a)–(f) of Section 1 and not their stronger neighbours: not superiority over BeliefPPG, whose recording-level prior rescues exactly the participants a single-window estimator loses; not superiority of the propagated uncertainty as a confidence channel—we derive and confirm the opposite; and not a benefit of accelerometer fusion on this cohort.

Why the representation matters beyond accuracy. A scalar output leaves nothing to reason about once it is wrong; a trajectory with per-point scales lets us derive how the reduction transforms uncertainty (Eq. 5), predict its failure, and confirm it, and it exposes shrinkage as a linear law across participants (Eq. 6). When the label is a reduction of structure, estimating the structure buys accuracy *and* a handle on mechanism.

Why the evaluation matters. Two corrected defects moved numbers by more than any architectural comparison here: label alignment changed the irreducible floor by 2.2 BPM, and the resolution collapse reversed a sensor ablation. A re-run floor of ± 0.34 BPM implies that many single-seed differences in this literature are uninterpretable.

Limitations. One cohort for the corrected protocol; the cohort-shift run predates the corrections; BeliefPPG at one seed; KID-PPG and EnhancePPG not retrained on our targets; no in-the-wild cohort (WildPPG, GalaxyPPG); thin high-HR/low-motion cells (3–5 participants); and the shrinkage law is descriptive, untested by interventions such as group-robust objectives (Sagawa et al., 2020).

6 Conclusion

When a regression label is a reduction of an observable structured trajectory, supervising the trajectory is a better use of the same backbone: on wrist-PPG heart-rate estimation it lowers error by 1.07 BPM over an identical-architecture scalar head (three times a measured noise floor), beats standard deep and classical baselines, and matches the strongest method we could retrain, with five full-scale ablations showing no auxiliary mechanism is individually responsible. The trajectory is also a representation one can reason with: it lets us derive and confirm why the propagated uncertainty fails at high heart rate and exposes the dominant failure under subject-level shift as shrinkage toward the cohort mean; with a calibration–test-separated confidence score, that understanding turns an unreliable regressor into a selective one. The natural next step is to combine trajectory supervision with a recording-level temporal prior, targeting the outlying participants where shrinkage is largest.

REPRODUCIBILITY STATEMENT

All experiments use public datasets. Code, configurations, seeds, and the alignment and rank probes will be released with the camera-ready version; an anonymised repository can be provided during review. The noise-floor measurement, the three-seed primary result, and the strict separation of corrected from legacy protocols are described in Section 4.1. The BeliefPPG reproduction used the authors' public code with two documented environment shims (`TF_USE_LEGACY_KERAS=1`; a `wandb` import fallback), no changes to the method, and default hyper-parameters; its per-frame estimates were aligned to our targets by a label-only offset search (Appendix **??**). Appendices **??**–**??** give dataset statistics and per-fold tables.

ETHICS STATEMENT

The work uses publicly available, de-identified datasets collected under their original ethics approvals; no new human-subject data were collected. The method targets physiological monitoring, where silent high-magnitude errors carry risk; we characterise the failure mode explicitly and provide a calibrated abstention mechanism rather than reporting only average accuracy.

USE OF LARGE LANGUAGE MODELS

A general-purpose large language model was used for language editing and LaTeX formatting. All research questions, experimental design, code, analyses, figures, and claims were produced and verified by the authors, who take full responsibility for the content.

REFERENCES

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Luca Benfenati, Sofia Belloni, Alessio Burrello, Panagiotis Kasnesis, Xiaying Wang, Luca Benini, Massimo Poncino, Enrico Macii, and Daniele Jahier Pagliari. EnhancePPG: Improving PPG-based heart rate estimation with self-supervision and augmentation. *arXiv preprint arXiv:2412.17860*, 2024.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

Valentin Bieri, Paul Streli, Berken Utku Demirel, and Christian Holz. BeliefPPG: Uncertainty-aware heart rate estimation from PPG signals via belief propagation. In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 216 of *PMLR*, pp. 173–183, 2023.

Dwaipayan Biswas, Luke Everson, Muqing Liu, Madhuri Panwar, Bram-Ernst Verhoef, Shrishail Patki, Chris H. Kim, Amit Acharyya, Chris Van Hoof, Mario Konijnenburg, and Nick Van Helleputte. CorNET: Deep learning framework for PPG-based heart rate estimation and biometric identification in ambulant environment. *IEEE Transactions on Biomedical Circuits and Systems*, 13(2):282–291, 2019.

Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, volume 48 of *PMLR*, pp. 1050–1059, 2016.

Francis Galton. Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263, 1886.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37 of *PMLR*, pp. 1180–1189, 2015.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems 30*, pp. 4878–4887, 2017.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *PMLR*, pp. 2151–2159, 2019.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *PMLR*, pp. 1321–1330, 2017.

Stuart H. Hurlbert. Pseudoreplication and the design of ecological field experiments. *Ecological Monographs*, 54(2):187–211, 1984.

Christodoulos Kechris, Jonathan Dan, José Miranda, and David Atienza. KID-PPG: Knowledge informed deep learning for extracting heart rate from a smartwatch. *IEEE Transactions on Biomedical Engineering*, 2025. arXiv:2405.09559.

Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems 30*, pp. 5574–5584, 2017.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *PMLR*, pp. 5637–5664, 2021.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems 30*, pp. 6402–6413, 2017.

Chen-Yu Lee, Saining Xie, Patrick W. Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 38 of *PMLR*, pp. 562–570, 2015.

Manuel Meier, Berken Utku Demirel, and Christian Holz. WildPPG: A real-world PPG dataset of long continuous recordings. In *Advances in Neural Information Processing Systems 37, Datasets and Benchmarks Track*, 2024.

Sangjun Park, Dejiang Zheng, and Uichin Lee. A PPG signal dataset collected in semi-naturalistic settings using Galaxy Watch. *Scientific Data*, 12:892, 2025.

Attila Reiss, Ina Indlekofer, Philip Schmidt, and Kristof Van Laerhoven. Deep PPG: Large-scale heart rate estimation with convolutional neural networks. *Sensors*, 19(14):3079, 2019.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020.

Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, and Georg Martius. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. In *International Conference on Learning Representations (ICLR)*, 2022.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015.

Task Force of the ESC and NASPE. Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. *Circulation*, 93(5):1043–1065, 1996.

Andriy Temko. Accurate heart rate monitoring during physical exercises using PPG. *IEEE Transactions on Biomedical Engineering*, 64(9):2016–2024, 2017.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Zhilin Zhang, Zhouyue Pi, and Benyuan Liu. TROIKA: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise. *IEEE Transactions on Biomedical Engineering*, 62(2):522–531, 2015.