LONG-HORIZON PHYSIOLOGICAL ANALYSIS VIA NATIVE CALIBRATION OF A MULTIMODAL WAVEFORM FOUNDATION MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Bedside monitors record PPG and ECG waveforms continuously, but laboratory values such as hemoglobin and potassium are measured only when blood is drawn: hours apart in intensive care, months apart between emergency-department visits. We study long-horizon physiological analysis: estimating from the waveforms how a lab value has changed since the previous draw. Our method, *native calibration*, uses the patient's own history: each earlier draw provides a measured value and a waveform recorded at the same time. It estimates the current value as an earlier measured value plus the change read from the current and past waveforms, both embedded by a frozen waveform foundation model (FM). We also introduce a new multimodal FM, a transformer pretrained without labels to relate ECG and PPG and to match a patient's PPG across time. We evaluate on two kinds of task: 1) tracking lab changes, namely hemoglobin between emergency visits and potassium within intensive-care stays; and 2) reading values from the current waveform alone, namely first-draw emergency labs, intra-operative haemodynamics and 24 standard PPG benchmark tasks. On held-out patients, native calibration improves the estimated change in both hemoglobin and potassium over strong non-waveform baselines, and part of this gain is specific to the patient: it shrinks when the patient's past waveforms are replaced by those of other patients with the same lab values. Our FM achieves state-of-the-art results on first-draw labs and the standard PPG tasks, ranking first among ten encoders, and is competitive with the best public encoder on intra-operative haemodynamics.

1 INTRODUCTION

A monitored patient produces photoplethysmography (PPG) and ECG waveforms continuously, but laboratory values such as hemoglobin (Hgb) and potassium (K) are known only when blood is drawn. *Long-horizon physiological analysis* aims to track these values across the gaps between draws, which last hours in an intensive care unit (ICU) and months between emergency-department (ED) visits. Waveform foundation models (FMs) provide general-purpose embeddings of windows lasting seconds to minutes (Pillai et al., 2025; Nie et al., 2026; Guo et al., 2026; Li et al., 2025), general time-series FMs can embed the same windows (Goswami et al., 2024; Ansari et al., 2024; Feofanov et al., 2026a), and recent work predicts lab values from such embeddings (Wang et al., 2025; Kataria et al., 2025; Gustafsson et al., 2026). Two things are missing: an encoder trained for patient-level, multimodal readout, and a readout that uses the patient's own history.

Native calibration. Cuffless blood-pressure estimation calibrates per person, because the map from pulse transit time to pressure is subject-specific (Mukkamala et al., 2015). A patient's lab history offers similar calibration points: each earlier draw pairs a measured value y_j with a waveform window e_j recorded at the time. Native calibration estimates the current value as $\hat{y}_i = y_j + g(e_i) - g(e_j)$, with the same map g applied to the current window e_i and the past window: the measured value sets the level, and the waveforms supply only the change. This is first-differencing (Wooldridge, 2010) on FM embeddings; equivalently, a population prediction $g(e_i)$ plus the patient's past residual $y_j - g(e_j)$, so what matters is how g is fitted. We fit g with a small attention network over several past windows (NC-attn); its linear special case (NC-lin) is an ablation.

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.

To test whether the patient's own past waveforms matter, a donor control retrains the same pipeline with each past waveform replaced by that of another patient who had the same lab value at that draw. Whether the gain comes from calibration in the strict sense (each past window paired with its own value) or from the patient's recent waveform is examined in Section 5.3.

Evaluation. We evaluate on four task families. 1) Long-horizon tasks, which test native calibration, predict how a lab value changes between draws: Hgb between ED visits, from PPG, and K within an ICU stay, from ECG. The other three families test the FM alone, predicting a value from the current waveform only; we call them level-readout tasks. 2) First-draw labs: seven analytes in the first blood draw of an ED visit. 3) Intra-operative haemodynamics: stroke volume, cardiac output and stroke-volume variation during surgery. 4) Standard PPG tasks: 24 short-window benchmark tasks covering heart rate, blood pressure, atrial fibrillation, respiration and SpO_2.

Contributions.

1. **A multimodal waveform encoder** (Sections 4.1 and 5.1) that ranks first among ten encoders on first-draw labs and standard PPG tasks, ties the best public encoder on intra-operative haemodynamics, and beats a ResNet backbone trained with the identical recipe (validation patients).
2. **Native calibration, tested by the donor control** (Sections 4.2 and 5.2). The patient's own past beats the donor control for ED hemoglobin and ICU potassium with our encoder, and for 10 of 13 other encoder–task pairs.
3. **A dose-response in the patient's own history** (Section 5.2). For hemoglobin, the own-past gain grows with the amount of past-visit waveform read, from +0.0198 for 30 s to +0.0377 for the full visit, and does not decay over gaps of weeks to months.

2 Related work

Within-person estimation and calibration. In panel data, first-differencing and fixed effects remove time-invariant individual effects (Wooldridge, 2010, Ch. 10), and random-effects models capture person-specific deviations in longitudinal data (Laird & Ware, 1982). Blood-pressure estimation from pulse transit time typically requires subject-specific calibration (Mukkamala et al., 2015). Native calibration applies first-differencing to frozen FM embeddings, with measured lab values as calibration points; what it adds is a map shared between current and past embeddings, and the donor control. Weight-tied (siamese) networks compare two inputs through identical sub-networks (Bromley et al., 1993); our linear map is the simplest case. **Waveform FMs and within-subject pretraining.** PaPaGei (Pillai et al., 2025), AnyPPG (Nie et al., 2026) and SIGMA-PPG (Guo et al., 2026) are pretrained PPG encoders, ECGFounder (Li et al., 2025) is a public ECG FM, and MOMENT, Chronos with its Chronos-Bolt release, and Mantis with MantisV2 (Goswami et al., 2024; Ansari et al., 2024; Amazon Science, 2024; Feofanov et al., 2026a;b) are general time-series FMs. Our encoder uses a LeJEPA-style objective (Balestriero & LeCun, 2025) with a patient-contrastive term in the spirit of PCLR (Diamant et al., 2022). Patient-contrastive (Diamant et al., 2022) and temporal-neighbourhood (Tonekaboni et al., 2021) objectives use within-subject structure during *pretraining*; native calibration uses it at *inference*, together with measured labels. **Labs from waveforms.** Continuous noninvasive hemoglobin monitors show small mean bias but wide limits of agreement with laboratory hemoglobin (Kim et al., 2014). Deep learning on the ECG can screen for hyperkalemia (Galloway et al., 2019). UNIPHY+Lab (Wang et al., 2025) estimates ICU lab trajectories from PPG with a patient-aware Mamba model. Kataria et al. (2025) benchmark generalist and specialist FMs on 51 PPG-focused tasks, including lab-value estimation. SignalMC-MED (Gustafsson et al., 2026) benchmarks FMs on single-lead ECG and PPG from the ED, including lab regression, without a lab-history baseline. PulseOx-FM (Kohn et al., 2026) relates wearable oximetry embeddings to health and cardiometabolic markers. We count only what waveforms add beyond the strongest non-waveform model we could build, and we guard against leakage by splitting by patient and fitting every model on training patients only (Kaufman et al., 2012; Kapoor & Narayanan, 2023).

3 Setting and evaluation

Data. *ED*: MC-MED (Kansal et al., 2025a;b), visits with monitor PPG and ECG lead II. *ICU*: the MIMIC-III (Johnson et al., 2016; Moody et al., 2020) patients with linked bedside waveform

Table 1: Floors: the non-waveform model that every waveform result must improve on. Each is a gradient-boosted model selected on training patients.

Floor	Contents
ED cross-visit	previous value of the analyte, log gap, the current visit's case-mix floor, change in vitals since the previous visit
ED case mix	demographics, arrival mode, triage acuity, chief-complaint tokens, triage and monitor vitals, every analyte's value at earlier visits; for Hgb also waveform-monitoring timing terms
ICU lab history	last values and ages of every analyte (including every anchor value used by native calibration), demographics, acuity (organ-failure score, vasopressors, transfusion, insulin), monitor numerics, vital-sign trends
OR haemodynamics	case context (age, sex, BMI, ASA class, surgery and anaesthesia type, elapsed time), 10-min vital-sign summaries, device flag; the *strong* variant adds arterial systolic, diastolic and pulse pressure and noninvasive pressures

records; chart times are corrected to the waveform clock. *Operating room (OR)*: VitalDB (Lee et al., 2022), surgical cases with PPG and cardiac-output monitoring. All splits are by patient (60/20/20; cohort sizes in Section A). In the ED and OR, stored lab times are result times, not draw times; ED windows therefore end 30 min before the visit's first lab result, and OR windows 15 min before the result. In the ICU, the ECG window is the hour before the recorded draw; the value is never an input.

Floors. We report every waveform result as an *increment* over a *floor*: the strongest model we could build from non-waveform information, selected by patient-grouped cross-validation on training patients (Table 1; floor performance in Section B). Floors use only information recorded before the prediction time: in the ICU, labs, treatments and monitor numerics recorded before the draw; in the ED, triage information, monitor vitals and values from earlier visits. The ED hemoglobin floor also includes the time from arrival to the start of waveform monitoring (arrival hour and weekday carry no clock information, as timestamps are shifted per patient). The waveform model is fitted to the floor's out-of-fold residual on training patients, by nested patient-grouped cross-validation in which the null model is always a candidate. For level targets we report ΔR^2, the gain in R^2 over the floor. For change targets we report ΔR^2_Δ, the gain in R^2_Δ, the R^2 of the change $y_i - y_{\text{prev}}$ from the previous measured value. Intervals are 95% paired patient-cluster bootstrap intervals (2,000 resamples). All configurations, including each family's public comparator and each encoder's export (whitened or unwhitened principal components, chosen by inner cross-validation on training patients), were fixed on validation patients; each family was then evaluated once on held-out test patients (Section G). For the two native-calibration contrasts of our encoder we also report Bonferroni-adjusted intervals ($\alpha = 0.05/2$, exact cluster-bootstrap percentiles).

Tasks. First-draw labs are scored over the ED case-mix floor. Intra-operative haemodynamics uses 10-min PPG windows; its targets are stroke volume (SV), cardiac output (CO), stroke-volume variation (SVV) and SVV$\geq$13%. The 24 standard PPG tasks come from 15 datasets, drawn from the benchmark of Kataria et al. (2025) (Table 5; heart rate 5, blood pressure 13, atrial fibrillation 2, respiration and SpO_2 4), and are read by frozen linear probes. The public comparators are AnyPPG, PaPaGei-S, MantisV2 (Feofanov et al., 2026b), MOMENT-base, Chronos-Bolt-small (Ansari et al., 2024; Amazon Science, 2024) and SIGMA-PPG, plus ECGFounder for ECG, frozen with fixed front-ends (Section D).

4 METHOD

4.1 OUR ENCODER

Architecture. The input is a 30-s PPG window at 40 Hz, or an ECG lead II window. A multi-branch convolutional stem feeds a pre-norm transformer trunk with rotary position encoding (patch size 8). The readout concatenates pooled token statistics of an intermediate and the final layer with a slot-competition aggregate, giving a 5378-d embedding (17.3M parameters; Section D). Training uses no labels and has two stages.

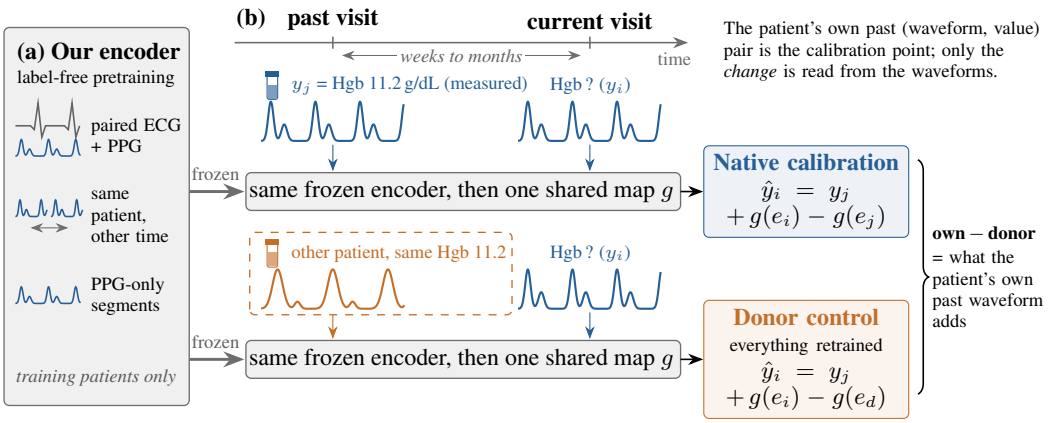

Figure 1: Native calibration on one patient's timeline. (a) Our encoder is pretrained without labels on three waveform streams from training patients, then frozen. (b) Top: the patient's past visit provides a waveform and a measured hemoglobin value. Both windows pass through the same frozen encoder and shared map g; the estimate is the past value plus the change read from the two waveforms. Bottom: the donor control is identical and fully retrained, except that the past waveform comes from another patient who had the same value at that draw. Own − donor measures what the patient's own past waveform adds. NC-attn, used for held-out results, extends this to several past windows combined by attention, added to a floor that holds the past lab values.

Stage 1 trains with a LeJEPA-style objective (Balestriero & LeCun, 2025) on PPG crops, with ECG lead II as an auxiliary view. We keep its checkpoint as the comparator *Ours (initial)*. **Stage 2** continues for 24,000 steps on three streams, all drawn from training patients (Figure 1a; Section C). (i) *Paired*: time-aligned ECG and PPG segments from ICU and ED patients (661k and 95k). Each modality has its own LeJEPA term, projector and isotropy regulariser, and a cross-modal term predicts each modality's view centroid from the other's. (ii) *Patient-contrastive*: triplets of PPG segments from ED patients (442k segments), trained with a symmetric InfoNCE loss in which a window of the same patient recorded at another time is the positive and other patients' windows are the negatives. (iii) *PPG-only*: ICU segments (10.3M, 85.8k hours) that enter the PPG LeJEPA term alone. Each step draws a fixed share from each stream, and the loss is a weighted sum of the stream terms. No validation or test patient of any evaluation task appears in any stream. We also compare two single-stream stage-2 variants, *Ours (patient-contrastive)* and *Ours (ICU multimodal)*.

4.2 Native Calibration

Let e_i be the pooled embedding of the current window. Each earlier draw of the patient provides an *anchor* $(e_j, y_j, \Delta t_{ij})$: the window recorded at that draw, its measured value, and the time since. The floor already sees these values, so native calibration adds only the past windows, not lab information. The linear form, **NC-lin**, fits one ridge regression on all within-patient pairs of training patients,

$$y_i - y_j \;\approx\; w^\top (e_i - e_j) + c^\top [\log \Delta t_{ij},\, 1], \tag{1}$$

with the penalty chosen by patient-grouped cross-validation. Because w is shared, a patient offset u_p in $y = w^\top e + u_p$ cancels in Equation (1). For each current window and anchor k, NC-lin gives a calibrated estimate of the change, $d_k = y_{j_k} + w^\top (e_i - e_{j_k}) + c^\top [\log \Delta t_{ik}, 1] - y_{\text{prev}}$, and a level score $w^\top e_i$. These few features enter the model of the floor's residual. The map is *cross-fitted*: a training row's features come from a model that never saw that patient.

Attention form (NC-attn). NC-attn is used for all held-out results (Figure 1b). A small network passes the current and past embeddings through one shared map ϕ. For each anchor it reads $[\phi(e_i) - \phi(e_j),\ \phi(e_i) \odot \phi(e_j),\ \log \Delta t_{ij}]$ and outputs a prediction and an attention weight; the attention-weighted predictions are added to a head that reads the current window alone. This handles a variable number of anchors: one past visit in the ED, up to three earlier draws of the stay in the ICU. NC-attn never reads a lab value: it is trained on the floor's residual, and the past values enter only through the floor. It is cross-fitted like NC-lin and averaged over three initialisations (Section I).

Table 2: Level readout on held-out test patients; one export rule, with configurations and comparators fixed on validation. First-draw labs: mean ΔR^2 over the case-mix floor across seven labs [95% CI]; intra-operative haemodynamics: mean increment over the vitals-and-case floor across four targets; standard PPG tasks: mean rank over 24 tasks, and family-balanced rank. Ranks (lower is better) are among 11 rows: the ten encoders shown and Ours (initial) with a second front-end. Per-task results in Section H.

	first-draw labs		intra-op. haemodynamics		standard PPG tasks	
encoder	mean ΔR^2	rank	mean incr.	rank	mean rank	family-bal.
Ours	+0.021 [+0.018, +0.025]	1.29	+0.095	3.00	3.33	3.86
Ours (patient-contrastive)	+0.020 [+0.017, +0.023]	2.00	+0.099	2.50	4.83	4.20
Ours (ICU multimodal)	+0.018 [+0.015, +0.021]	5.00	+0.098	2.25	4.96	5.24
Ours (initial)	+0.018 [+0.015, +0.021]	4.57	+0.092	5.50	5.54	5.99
AnyPPG	+0.020 [+0.017, +0.023]	3.14	+0.063	10.00	4.00	4.02
MOMENT-base	+0.012 [+0.0092, +0.014]	7.43	+0.094	4.00	7.17	6.74
MantisV2	+0.0078 [+0.0055, +0.010]	9.29	+0.094	5.25	6.21	5.23
Chronos-Bolt-small	+0.0086 [+0.0066, +0.011]	8.57	+0.076	8.50	7.12	7.24
SIGMA-PPG	+0.0070 [+0.0049, +0.0090]	9.43	+0.076	9.00	8.33	8.55
PaPaGei-S	+0.0053 [+0.0037, +0.0069]	10.29	+0.064	10.50	9.25	9.38

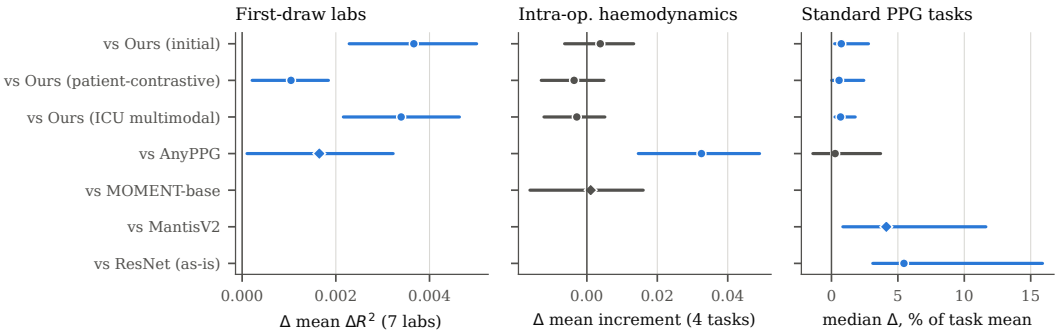

Figure 2: Our encoder minus each comparator on identical held-out rows. First-draw labs and intra-operative haemodynamics: mean difference, 95% paired patient-cluster bootstrap CI; standard PPG tasks: median over the 24 tasks of the difference in % of the task mean, 95% task-bootstrap CI. Diamonds: public encoder frozen as best on validation for that family. Blue: interval above zero; grey: interval spans zero. Comparators not run in a family are omitted. ResNet (as-is): the stage-1 checkpoint of Table 3, standard PPG tasks only, not ranked.

Controls. The *donor* control keeps the patient's own past values, gaps and current window, but replaces each past *window* with a window of a different patient from the same split (training donors for training rows, test donors for test rows). For each anchor, the donor is drawn at random from the 10 other patients whose value at that anchor is nearest to the patient's own, with replacement across patients. The swap is applied to training and evaluation rows alike, and the whole pipeline, including cross-fitting, is retrained. The control thus receives the same lab values and amount of waveform, and differs only in whose past waveform it sees; we report the *own* arm minus the *donor* arm, own − donor. We repeat the donor draw 6 times and average the 6 predictions per row. The bootstrap holds donors fixed; donor-draw noise is measured separately by the A/A contrast between two donor arms (Section I.1). The *current-only* control is the same network trained without past windows (own − current). Each task also has a negative-control analyte with no plausible waveform pathway (Section I.5). Multi-seed encoders' predictions are averaged over seeds before scoring.

Table 3: Architecture under a matched recipe: our encoder minus a ResNet backbone, both continued with the identical stage-2 recipe and compared at the same step. Validation patients, one seed [95% CI]; positive favours our architecture. Standard PPG tasks: mean sign-adjusted difference in % of the task mean. Last column: the two stage-1 checkpoints on held-out test patients. Details in Section E.

family	4k steps	8k steps	stage 1 only (test)
first-draw labs, mean ΔR^2	not scored	+0.0095 [+0.0076, +0.0114]	–
intra-op. haemodynamics, mean incr.	+0.0638 [+0.0445, +0.0834]	+0.0555 [+0.0357, +0.0752]	–
standard PPG tasks, % of task mean	+7.73% [+5.37, +10.04]	+7.27% [+5.29, +9.30]	+10.05% [+7.74, +12.49]

5 RESULTS

5.1 OUR ENCODER AS A FOUNDATION MODEL

First-draw labs. Table 2 summarises the level readout, and Figure 2 shows our encoder's paired differences to each comparator. The case-mix floor alone explains R^2 0.177–0.531 of the seven analytes (4,267–4,335 held-out visits per lab). Our encoder adds +0.021 [+0.018, +0.025] on average, most for hemoglobin (+0.049 [+0.040, +0.057]), and ranks first (mean rank 1.29, first on 6 of seven labs; AnyPPG 3.14). Paired on identical visits, it beats AnyPPG, the public encoder frozen as best, by +0.0016 [+0.0001, +0.0032], Ours (initial) by +0.0037 [+0.0023, +0.0050] (with an interval above zero for 6 of seven labs individually), and both single-stream variants (patient-contrastive +0.0010 [+0.0002, +0.0018]; ICU multimodal +0.0034 [+0.0022, +0.0046]).

Intra-operative haemodynamics. Every encoder adds substantially to the vitals-and-case floor (172 test patients). Our encoder ranks first on SV and on SVV$\geq$13%, 3rd on CO and 7th on SVV (mean rank 3.00). It ties MOMENT-base, the public encoder frozen as best (+0.0011 [$-$0.016, +0.016]), and Ours (initial) (+0.0038 [$-$0.0063, +0.013]), and beats AnyPPG (+0.033 [+0.015, +0.049]). Over the strong floor, which adds arterial pressures, it ties MOMENT-base (+0.0032 [$-$0.0062, +0.012]; Section H.2).

Standard PPG tasks. Our encoder has the best mean rank (3.33, task-bootstrap [2.38, 4.38]), the best family-balanced rank (3.86) and the most wins (9 of 24). By family it ranks 2.40 on heart rate, 3.31 on blood pressure and 2.75 on respiration and SpO_2, but 7.00 on the two atrial-fibrillation tasks, its weakest family. Because a few small tasks dominate mean differences, we compare encoders task by task. Our encoder is better than Ours (initial), both single-stream variants and MantisV2 (the public encoder frozen as best on validation) on most tasks: 17–19 of 24 by sign test, Wilcoxon signed-rank $p \leq 0.035$, with a positive median difference (Figure 2). It is on par with AnyPPG, the best public encoder on test: better on 13/24 tasks (Wilcoxon $p = 0.195$), median difference +0.3% [$-$1.4, +3.7].

Architecture. To separate the architecture from the stage-2 recipe, we took a ResNet backbone (5.37M parameters) whose stage-1 checkpoint our group had pretrained with a cross-group LeJEPA recipe, continued it with our stage-2 recipe, dropping only the slot terms (the ResNet has no slots), and compared both encoders at the same step (Table 3). At 8k steps our architecture is better on all three level-readout families: by +7.27% [+5.29, +9.30] of the task mean on the standard PPG tasks (better, tied and worse on 11/12/1 of 24 tasks), by +0.0555 [+0.0357, +0.0752] on intra-operative haemodynamics and by +0.0095 [+0.0076, +0.0114] on first-draw labs. The pattern is the same at 4k on the two families scored. Before stage 2, Ours (initial) already leads the ResNet checkpoint, by +10.05% [+7.74, +12.49] on the held-out standard tasks. Details are in Section E.

5.2 NATIVE CALIBRATION ON HELD-OUT PATIENTS

ED: hemoglobin between visits, from PPG. Consecutive ED visits are a median of 73 d apart, and the floor holds the previous value, the gap and the current visit's case mix. Native calibration on our encoder raises R^2_Δ of the hemoglobin change by +0.063 [+0.043, +0.084] over this floor (Table 4). The patient's own past windows add +0.030 [+0.017, +0.043] to the current window alone

Table 4: Native calibration (NC-attn) on held-out test patients: ΔR^2_Δ with 95% CI. Own − donor: own past minus value-matched donors' past (6 donor draws, pipeline retrained); own − current: own past minus current window only; own vs floor: gain over the floor. Top: ED hemoglobin change between consecutive visits (891 pairs / 528 patients), every PPG encoder with the same input treatment. Bottom: ICU potassium change from ECG (1,537 draws / 604 stays), ECG-capable encoders.

encoder	own − donor	own − current	own vs floor
ED hemoglobin from PPG			
Ours	+0.038 [+0.023, +0.053]	+0.030 [+0.017, +0.043]	+0.063 [+0.043, +0.084]
Ours (patient-contrastive, 3 seeds)	+0.037 [+0.020, +0.054]	+0.032 [+0.016, +0.047]	+0.065 [+0.042, +0.088]
Ours (initial)	+0.029 [+0.013, +0.046]	+0.027 [+0.012, +0.044]	+0.049 [+0.029, +0.070]
AnyPPG	+0.025 [+0.010, +0.040]	+0.021 [+0.0073, +0.035]	+0.045 [+0.026, +0.065]
MOMENT-base	+0.012 [+0.0023, +0.022]	+0.011 [+0.0014, +0.019]	+0.033 [+0.018, +0.047]
PaPaGei-S	+0.0042 [+0.0014, +0.0076]	+0.0037 [+0.0009, +0.0067]	+0.011 [+0.0057, +0.016]
Chronos-Bolt-small	+0.0029 [−0.0033, +0.0092]	−0.0013 [−0.0064, +0.0039]	+0.019 [+0.0089, +0.030]
MantisV2	+0.0027 [−0.0052, +0.011]	+0.0008 [−0.0049, +0.0067]	+0.015 [+0.0016, +0.029]
SIGMA-PPG	+0.0013 [−0.0010, +0.0039]	+0.0039 [+0.0008, +0.0072]	+0.0059 [+0.0021, +0.010]
ICU potassium from ECG			
Ours	+0.011 [+0.0057, +0.018]	+0.0079 [+0.0026, +0.014]	+0.037 [+0.022, +0.054]
Ours (ICU multimodal, 3 seeds)	+0.014 [+0.0073, +0.021]	+0.012 [+0.0050, +0.019]	+0.043 [+0.027, +0.061]
Ours (initial)	+0.0080 [+0.0027, +0.014]	+0.0058 [+0.0004, +0.012]	+0.032 [+0.018, +0.048]
ECGFounder	+0.011 [+0.0064, +0.018]	+0.011 [+0.0057, +0.016]	+0.020 [+0.010, +0.032]
MOMENT-base	+0.010 [+0.0058, +0.015]	+0.0077 [+0.0034, +0.012]	+0.030 [+0.021, +0.041]
MantisV2	+0.0035 [+0.0016, +0.0057]	+0.0028 [+0.0011, +0.0046]	+0.010 [+0.0061, +0.016]

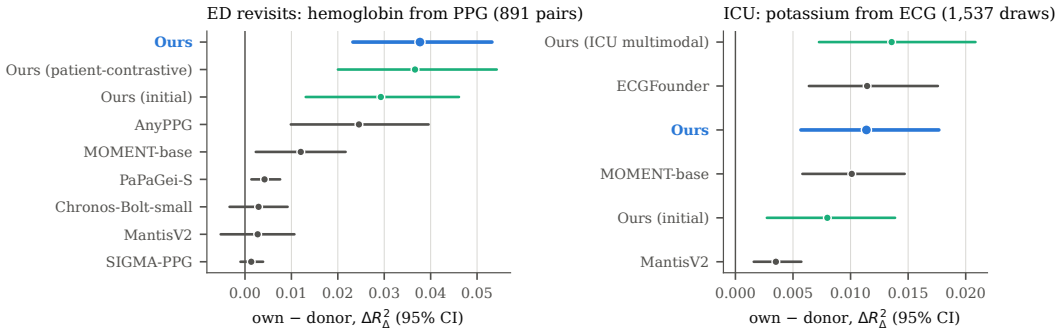

Figure 3: Own − donor for every encoder run on each task (NC-attn, held-out test, 95% patient- or stay-cluster CI), sorted by estimate. Ours in blue, our other encoders in green, public encoders in grey.

and beat the donor control by +0.038 [+0.023, +0.053] (Bonferroni-adjusted [+0.021, +0.055]). Paired on the same test pairs, our encoder's own − donor contrast is larger than AnyPPG's by +0.013 [+0.0031, +0.024] and than that of Ours (initial) by +0.0084 [+0.0010, +0.016], and matches that of the patient-contrastive variant (+0.0011 [−0.0060, +0.0078]). The contrast is above zero for 5 of the 8 other PPG encoders (Figure 3). Donor-draw noise, which the patient bootstrap does not capture, is small (A/A standard deviation 0.0035; Section I.4).

ICU: potassium within a stay, from ECG. Here the floor holds every recent measurement, and up to three earlier draws of the stay serve as anchors. With our encoder, the own past windows add +0.0079 [+0.0026, +0.014] to the current window and beat the donor control by +0.011 [+0.0057, +0.018] (Bonferroni-adjusted [+0.0047, +0.019]; +0.011 with 12 draws, against a draw-noise standard deviation of 0.0007). Paired on the same draws, our encoder's contrast minus that of Ours (ICU multimodal) is −0.0022 [−0.0056, +0.0011], and minus that of Ours (initial) +0.0034 [−0.0013, +0.0084]. The contrast is above zero for all 5 other ECG-capable encoders, including ECGFounder,

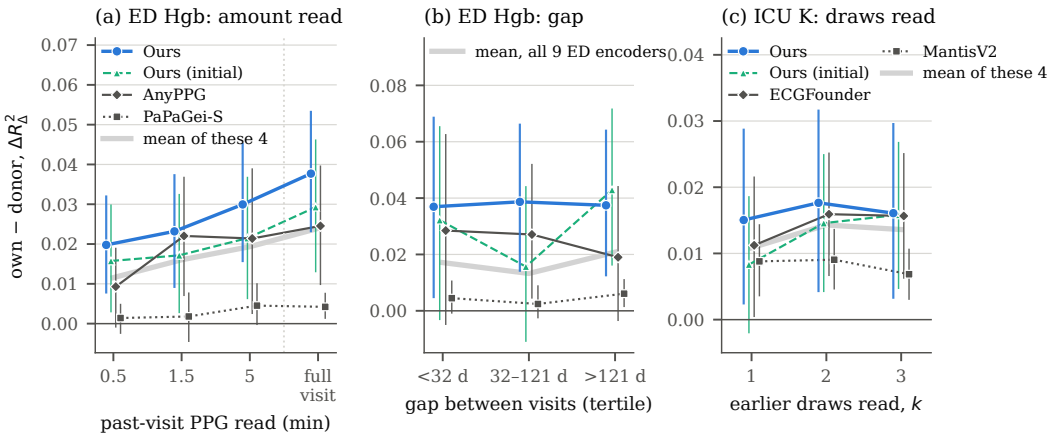

Figure 4: Own − donor against the amount of own history (NC-attn, held-out test, 95% patient- or stay-cluster CI). (a) ED hemoglobin, retrained with the m most recent 30-s windows of the past visit ($m = 1, 3, 10$; own, donor and training windows alike) or the full visit (up to 40 windows); 891 pairs. (b) Full-visit predictions of the main analysis by tertile of the gap between visits (median 11, 64 and 245 d); grey: mean over all 9 ED encoders of Table 22. (c) ICU potassium, retrained with the k most recent earlier draws, on the 514 draws / 160 stays with three valid earlier draws (a subset of Table 4).

MOMENT-base and MantisV2, and our ICU-pretrained encoders give the largest gains over the floor. The negative-control analytes are null with our encoder: ALT between ED visits +0.0005 [−0.0003, +0.0014], and sodium within an ICU stay −0.0002 [−0.0006, +0.0002] (Section I.5).

More own history, more gain. If the gain comes from the patient's own past, it should grow with how much of that past is read. We retrained the ED pipeline with only the m most recent 30-s windows of the past visit (Figure 4a). With our encoder, own − donor grows from +0.0198 at 0.5 min to +0.0232 at 1.5 min, +0.0300 at 5 min and +0.0377 with the full visit. Over $m = 1, 3, 10$, where every window set has exactly m windows, the slope per doubling of m is +0.0031 [+0.0010, +0.0053] (one-sided $p = .002$); for the mean of the 4 encoders it is +0.0023 [+0.0005, +0.0042] ($p = .005$). The growth comes from the own arm: its gain over the floor rises from +0.0462 to +0.0628, whereas the donor arm's does not rise (slope −0.0005 [−0.0008, −0.0002]). Nor does the contrast decay with the gap between visits (Figure 4b): it is +0.0369, +0.0387 and +0.0374 for gaps below 32 d, 32–121 d and above 121 d (slope +0.0003 [−0.0181, +0.0183]). In the ICU, the contrast is flat in the number of earlier draws read ($k = 1, 2, 3$: +0.0150, +0.0176, +0.0161; slope +0.0005 [−0.0032, +0.0045]; Figure 4c): the most recent draw carries it.

Clinical relevance. With our encoder, the mean absolute error (MAE) changes by −0.048 g/dL (−4.6%) for hemoglobin between ED visits (−0.013 [−0.024, −0.0024] relative to AnyPPG), by −0.056 g/dL (−4.5%) for first-draw hemoglobin and by −0.011 mmol/L (−3.4%) for ICU potassium. Detection of first-draw hemoglobin below 8 g/dL and of ICU potassium ≤3.5 mmol/L also improves (Section J).

5.3 What carries the gain (validation analyses)

These exploratory analyses use validation patients with Ours (initial) or public encoders.

Estimator. Among estimators reading the same inputs (Table 21), tying the map is the only ingredient with a significant effect: separate penalties for the current and past windows do worse by +0.020 [+0.0059, +0.035]. First-difference least squares and a carried-residual model with NC-lin's features and cross-fitting match NC-lin. Partial pooling, gradient boosting and a generic reader of $[e_i, e_j]$ recover less of the own-past contrast. Adding the native-calibration features to the full current embedding leaves own − donor at +0.029, so the contrast is not produced by handicapping the donor arm.

Which past. With both estimators, the own past also beats donors matched on value, demographics and gap, and look-alike donors nearest in patient-mean embedding. Hand-crafted PPG features give no own-past gain (Table 25). The same holds for AnyPPG on ED hemoglobin and for ECG-Founder on ICU potassium (Table 26), so it is not a property of our encoders. The look-alikes are half as far from the patient in embedding space as value-matched donors, yet do no better, so the gain is specific to the patient's own past rather than to a patient type. Pairing each window with its own value adds +0.014 [+0.0023, +0.024] on the 1,443 rows with two earlier visits (one estimator only), and the recency of the own window matters: the older window gives +0.0036 with the previous value, against +0.024 with its own value.

6 LIMITATIONS

Effect size and clinical value. The patient-specific part of native calibration is small: most of the gain over the floor survives with a donor's past (Table 4). The clinical gains are modest: a better prior for the next draw, not a substitute measurement (Section J). On first-draw labs, the margin over AnyPPG is small: positive in point estimate for 6 of seven labs, and significant for the seven-lab mean and for 1 lab individually. On the standard PPG tasks, mean differences against the ranked comparators are not significant; those claims rest on ranks and on consistency across tasks. **Mechanism.** Pairing of window and value is not established: the pairing contrast (Section 5.3) rests on one estimator and on the rows with two earlier visits. The own previous window is closer to the current window in embedding space than any donor's (distance 18.3 against 21.1–23.1; likewise for AnyPPG and ECGFounder), so a persistent device, skin or recording signature remains an alternative to calibration. Donors are matched on the anchor value only, so their windows may also differ in case mix and recording conditions; the stricter donors and the current-only control do not depend on this. In NC-attn, the product term means that additive patient offsets no longer cancel exactly. **What the long-horizon tasks can rank.** Across 34 checkpoints of our encoders on ED hemoglobin (validation, NC-lin), own − donor has a standard deviation of 0.0019, below its bootstrap standard error of 0.0115 (ICU potassium: 0.0023 against 0.0031; Section I.6). These tasks thus cannot rank close checkpoints; our encoder claims rest on paired comparisons on identical test rows and on the level-readout families. In the ICU, our encoder's contrast is not significantly larger than that of Ours (initial); on ED hemoglobin, the own-past contrast is indistinguishable from zero for MantisV2, Chronos-Bolt-small and SIGMA-PPG. **Encoder evidence.** One pretraining seed; the three stage-2 streams are not separated. The architecture comparison uses validation patients and 8k of the 24k-step schedule; the ResNet's BatchNorm statistics are re-estimated on training PPG, and it exports its native 384-d embedding against our 512 principal components (Section E). Over the strong haemodynamics floor, our encoder's mean rank is 5.75, and our single-stream variants lead (Table 16). **Pretraining exposure.** In the ED hemoglobin task, 98% of test patients have unlabelled visits in an MC-MED waveform collection that may have fed our stage-1 pretraining. In VitalDB, all but 12 of 172 test patients are PulseDB subjects whose waveforms, not labels, are in our stage-1 corpus. AnyPPG's pretraining lists MC-MED and PulseDB, which is built from MIMIC-III and VitalDB (Nie et al., 2026); PaPaGei's lists VitalDB and the MIMIC-III waveform matched subset (Pillai et al., 2025). The haemodynamics comparison therefore favours the VitalDB-exposed rows (our encoders, AnyPPG and PaPaGei-S). No ICU validation or test stay appears in our pretraining data or in PulseDB's MIMIC-III part (Section K). **Scope.** One analyte per setting, one site per dataset, no external cohort.

7 CONCLUSION

We introduced native calibration, which anchors on an earlier measured lab value and reads only the change from the patient's current and past waveforms, and a multimodal waveform FM pretrained without labels. On held-out patients, native calibration improved estimates of hemoglobin and potassium change, and part of this gain is specific to the patient's own history. The FM ranked first among ten encoders on first-draw labs and standard PPG tasks and was competitive on intraoperative haemodynamics. The patient-specific gains are modest and their mechanism is open (Section 6).

ETHICS STATEMENT

All data are de-identified research releases (MIMIC-III, VitalDB, MC-MED) used under their data-use agreements; no new data were collected. We made no attempt at re-identification, and figures show only aggregate statistics. The models are research artifacts, not validated for clinical use, and must not replace laboratory testing. A waveform-based estimate could fail silently on devices or populations not represented here.

REPRODUCIBILITY STATEMENT

Section 3, Table 1 and Section 4.2 specify splits, floors, estimators and controls; Sections A to D, I and L give cohorts, floor performance, our encoder's corpus, objective and hyperparameters, the encoders and front-ends, the native-calibration specification, and compute. Every number in the paper is generated from result files by a script that we will release, together with the scoring code, the floors and the encoder weights. An anonymized repository is uploaded to OpenReview as a zip file.

AI USE STATEMENT

In this work, we used generative AI tools for 1) brainstorming and providing feedback on research methodology and experiments, 2) identifying relevant literature, and 3) implementing non-critical components of the code. The following uses do not apply to this work: 1) generating synthetic datasets; 2) helping develop theoretical models or conceptual frameworks; 3) formulating mathematical claims; 4) providing critical ingredients for proving mathematical claims; 5) assisting in the writing of proofs; 6) assisting with translation, or cleaning and reformatting datasets; 7) supporting qualitative and thematic data analysis or interpreting results. Additionally, we used generative AI to edit the paper and format references. We have reviewed all AI-assisted work; in particular, we double-checked every key citation and carefully verified the core code. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Amazon Science. Chronos-Bolt model release: chronos-bolt-small. Hugging Face model card `https://huggingface.co/amazon/chronos-bolt-small`; code at `https://github.com/amazon-science/chronos-forecasting`, 2024.

Santhalakshmi Angappan, Satyen Parida, Arumugam Vasudevan, and Ashok S. Badhe. The comparison of stroke volume variation with central venous pressure in predicting fluid responsiveness in septic patients with acute circulatory failure. *Indian Journal of Critical Care Medicine*, 19(7): 394–400, 2015. doi: 10.4103/0972-5229.160278.

Abdul Fatir Ansari, Lorenzo Stella, Ali Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Bernie Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics, 2025.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a "Siamese" time delay neural network. In *Advances in Neural Information Processing Systems*, volume 6, 1993.

Charles Carlson, Vanessa-Rose Turpin, Ahmad Suliman, Carl Ade, Steve Warren, and David E. Thompson. Bed-based ballistocardiography: Dataset and ability to track cardiovascular parameters. *Sensors*, 21(1):156, 2021. doi: 10.3390/s21010156.

Nathaniel Diamant, Erik Reinertsen, Steven Song, Aaron D. Aguirre, Collin M. Stultz, and Puneet Batra. Patient contrastive learning: A performant, expressive, and practical approach to electrocardiogram modeling. *PLOS Computational Biology*, 18(2):e1009862, 2022. doi: 10.1371/journal.pcbi.1009862.

Vasilii Feofanov, Songkang Wen, Shifeng Xie, Simon Roschmann, Marius Alonso, Hongbo Guo, Romain Ilbert, Malik Tiomoko, Quentin Bouniot, Zeynep Akata, Lujia Pan, Jianfeng Zhang, and Ievgen Redko. Mantis: Lightweight foundation model for time series classification. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026a.

Vasilii Feofanov, Songkang Wen, Jianfeng Zhang, Lujia Pan, and Ievgen Redko. MantisV2: Closing the zero-shot gap in time series classification with synthetic data and test-time strategies, 2026b.

Conner D. Galloway, Alexander V. Valys, Jacqueline B. Shreibati, Daniel L. Treiman, Frank L. Petterson, Vivek P. Gundotra, David E. Albert, Zachi I. Attia, Rickey E. Carter, Samuel J. Asirvatham, Michael J. Ackerman, Peter A. Noseworthy, John J. Dillon, and Paul A. Friedman. Development and validation of a deep-learning model to screen for hyperkalemia from the electrocardiogram. *JAMA Cardiology*, 4(5):428–436, 2019. doi: 10.1001/jamacardio.2019.0640.

Ainara Garde, Parastoo Dehkordi, Walter Karlen, David Wensley, J. Mark Ansermino, and Guy A. Dumont. Development of a screening tool for sleep disordered breathing in children using the Phone Oximeter™. *PLoS ONE*, 9(11):e112959, 2014. doi: 10.1371/journal.pone.0112959.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 16115–16152, 2024.

Zongheng Guo, Tao Chen, Yang Jiao, Yi Pan, Xiao Hu, and Manuela Ferrario. SIGMA-PPG: Statistical-prior informed generative masking architecture for PPG foundation model. In *International Conference on Machine Learning*, 2026.

Fredrik K. Gustafsson, Xiao Gu, Mattia Carletti, Patitapaban Palo, David W. Eyre, and David A. Clifton. SignalMC-MED: A multimodal benchmark for evaluating biosignal foundation models on single-lead ECG and PPG, 2026.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016. doi: 10.1038/sdata.2016.35.

Aman Kansal, Emma Chen, Boyang Tom Jin, Pranav Rajpurkar, and David A. Kim. MC-MED, multimodal clinical monitoring in the emergency department. *Scientific Data*, 12:1094, 2025a. doi: 10.1038/s41597-025-05419-5.

Aman Kansal, Emma Chen, Boyang Tom Jin, Pranav Rajpurkar, and David A. Kim. Multimodal clinical monitoring in the emergency department (MC-MED) (version 1.0.1). PhysioNet, 2025b.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023. doi: 10.1016/j.patter.2023.100804.

Saurabh Kataria, Yi Wu, Zhaoliang Chen, Hyunjung Gloria Kwak, Yuhao Xu, Lovely Yeswanth Panchumarthi, Ran Xiao, Jiaying Lu, Ayca Ermis, Anni Zhao, Runze Yan, Alex Federov, Zewen Liu, Xu Wu, Wei Jin, Carl Yang, Jocelyn Grunwell, Stephanie R. Brown, Amit Shah, Craig Jabaley, Tim Buchman, Sivasubramanium V. Bhavani, Randall J. Lee, and Xiao Hu. Generalist vs specialist time series foundation models: Investigating potential emergent behaviors in assessing human health using PPG signals, 2025.

Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, 6(4):1–21, 2012. doi: 10.1145/2382577.2382579.

Sang-Hyun Kim, Marc Lilot, Linda Suk-Ling Murphy, Kulraj S. Sidhu, Zhaoxia Yu, Joseph Rinehart, and Maxime Cannesson. Accuracy of continuous noninvasive hemoglobin monitoring: A systematic review and meta-analysis. *Anesthesia & Analgesia*, 119(2):332–346, 2014. doi: 10.1213/ANE.0000000000000272.

Sarah Kohn, Guy Lutsker, Alon Diament, Smadar Shilo, Adam Gabet, Gil Sasson, Gili Wolf, Adva Wolf, Anastasia Godneva, Adina Weinberger, Hagai Rossman, and Eran Segal. A foundation model of wearable pulse oximetry reveals physiological signatures of health and cardiometabolic risk. medRxiv preprint, 2026.

Nan M. Laird and James H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4): 963–974, 1982. doi: 10.2307/2529876.

H. Lee, H. Chung, and J. Lee. Motion artifact cancellation in wearable photoplethysmography using gyroscope. *IEEE Sensors Journal*, 19(3):1166–1175, 2019. doi: 10.1109/JSEN.2018.2879970.

Hyung-Chul Lee, Yoonsang Park, Soo Bin Yoon, Seong Mi Yang, Dongnyeok Park, and Chul-Woo Jung. VitalDB, a high-fidelity multi-parameter vital signs database in surgical patients. *Scientific Data*, 9:279, 2022. doi: 10.1038/s41597-022-01411-5.

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):AIoa2401033, 2025. doi: 10.1056/AIoa2401033.

Yongbo Liang, Zhencheng Chen, Guiyong Liu, and Mohamed Elgendi. A new, short-recorded photoplethysmogram dataset for blood pressure monitoring in China. *Scientific Data*, 5:180020, 2018. doi: 10.1038/sdata.2018.20.

Zeng-Ding Liu, Ye Li, Yuan-Ting Zhang, Jia Zeng, Zu-Xian Chen, Zhi-Wei Cui, Ji-Kui Liu, and Fen Miao. Cuffless blood pressure measurement using smartwatches: A large-scale validation study. *IEEE Journal of Biomedical and Health Informatics*, 27(9):4216–4227, 2023. doi: 10.1109/JBHI.2023.3278168.

Rebecca Mieloszyk, Hope Twede, Jonathan Lester, Jeremiah Wander, Sumit Basu, Gabe Cohn, Greg Smith, Dan Morris, Sidhant Gupta, Desney Tan, Nicolas Villar, Moni Wolf, Sailaja Malladi, Matt Mickelson, Lauren Ryan, Lindsey Kim, Jeffrey Kepple, Susanne Kirchner, Emma Wampler, Riena Terada, Joel Robinson, Ron Paulsen, and T. Scott Saponas. A comparison of wearable tonometry, photoplethysmography, and electrocardiography for cuffless measurement of blood pressure in an ambulatory setting. *IEEE Journal of Biomedical and Health Informatics*, 26(7): 2864–2875, 2022. doi: 10.1109/JBHI.2022.3153259.

Benjamin Moody, George Moody, Mauricio Villarroel, Gari D. Clifford, and Ikaro Silva. MIMIC-III waveform database matched subset (version 1.0). PhysioNet, 2020.

Ramakrishna Mukkamala, Jin-Oh Hahn, Omer T. Inan, Lalit K. Mestha, Chang-Sei Kim, Hakan Töreyin, and Survi Kyal. Toward ubiquitous blood pressure monitoring via pulse transit time: Theory and practice. *IEEE Transactions on Biomedical Engineering*, 62(8):1879–1901, 2015. doi: 10.1109/TBME.2015.2441951.

Andrea Nemcova, Enikö Vargova, Radovan Smisek, Lucie Marsanova, Lukas Smital, and Martin Vitek. Brno University of Technology smartphone PPG database (BUT PPG): Annotated dataset for PPG quality assessment and heart rate estimation. *BioMed Research International*, 2021: 3453007, 2021. doi: 10.1155/2021/3453007.

Andrea Nemcova, Radovan Smisek, Enikö Vargova, Lucie Maršánová, Martin Vitek, Lukas Smital, Marina Filipenska, Pavlina Sikorova, and Pavel Gálík. Brno University of Technology smartphone PPG database (BUT PPG) (version 2.0.0). PhysioNet, 2024.

Alexander Neshitov, Konstantin Tyapochkin, Evgeniya Smorodnikova, and Pavel Pravdin. Wavelet analysis and self-similarity of photoplethysmography signals for HRV estimation and quality assessment. *Sensors*, 21(20):6798, 2021. doi: 10.3390/s21206798.

Guangkun Nie, Xiaocheng Fang, Gongzheng Tang, Yujie Xiao, Jun Li, Bo Liu, Hongyan Li, and Shenda Hong. AnyPPG: An ECG-guided PPG foundation model trained on over 100,000 hours of recordings for holistic health profiling. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Volume 2 (KDD '26)*, pp. 11717–11727. ACM, 2026. doi: 10.1145/3770855.3818911.

Arvind Pillai, Dimitris Spathis, Fahim Kawsar, and Mohammad Malekzadeh. PaPaGei: Open foundation models for optical physiological signals. In *International Conference on Learning Representations*, 2025.

Marco A. F. Pimentel, Alistair E. W. Johnson, Peter H. Charlton, Drew Birrenkott, Peter J. Watkinson, Lionel Tarassenko, and David A. Clifton. Toward a robust estimation of respiratory rate from pulse oximeters. *IEEE Transactions on Biomedical Engineering*, 64(8):1914–1923, 2017. doi: 10.1109/TBME.2016.2613124.

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. doi: 10.3390/s19143079.

Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. Introducing WESAD, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI '18)*, pp. 400–408. ACM, 2018. doi: 10.1145/3242969.3242985.

Supreeth Prajwal Shashikumar, Amit J. Shah, Qiao Li, Gari D. Clifford, and Shamim Nemati. A deep learning approach to monitoring and detecting atrial fibrillation using wearable technology. In *2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pp. 141–144, 2017. doi: 10.1109/BHI.2017.7897225.

Sana Tonekaboni, Danny Eytan, and Anna Goldenberg. Unsupervised representation learning for time series with temporal neighborhood coding. In *International Conference on Learning Representations*, 2021.

Jessica Torres-Soto and Euan A. Ashley. Multi-task deep learning for cardiac rhythm detection in wearable devices. *npj Digital Medicine*, 3(1):116, 2020. doi: 10.1038/s41746-020-00320-4.

Pieter-Jan Toye. Vital Videos: A dataset of videos with PPG and blood pressure ground truths, 2023. arXiv:2306.11891v1.

Minxiao Wang, Runze Yan, Carol Li, Saurabh Kataria, Xiao Hu, Matthew Clark, Timothy Ruchti, Timothy G. Buchman, Sivasubramanium V. Bhavani, and Randall J. Lee. Estimating clinical lab test result trajectories from PPG using physiological foundation model and patient-aware state space model – a UNIPHY+ approach, 2025.

Weinan Wang, Pedram Mohseni, Kevin L. Kilgore, and Laleh Najafizadeh. PulseDB: A large, cleaned dataset based on MIMIC-III and VitalDB for benchmarking cuff-less blood pressure estimation methods. *Frontiers in Digital Health*, 4:1090854, 2023. doi: 10.3389/fdgth.2022.1090854.

Jeffrey M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, MA, 2 edition, 2010.

APPENDIX

The appendix is organised as follows. Section A describes the datasets and cohorts, Section B the floors, Section C the encoder's two-stage training, Section D the encoders, front-ends and export rule, Section E the matched architecture comparison, Section F a context-length analysis on MIMIC-III vitals, and Section G the evaluation protocol. Section H gives full per-task test results for the three level-readout families, Section I the native-calibration specification and its controls, Section K the pretraining-overlap audit, and Section L compute and reproducibility. Section J reports clinical-utility metrics.

A DATASETS AND COHORTS

MC-MED (ED). Emergency-department visits with bedside monitor PPG and ECG lead II (Kansal et al., 2025a;b). Waveforms are cut into 30-s windows, PPG at 40 Hz and ECG at 120 Hz, min-max normalised per window. A visit's embedding is the mean over its window embeddings. The split is by patient (medical record number), 60/20/20, and is shared by the ED long-horizon task, the first-draw labs and every pretraining stream that draws on MC-MED. For the first-draw labs, a visit enters a lab's panel when that lab is among the visit's first resulted labs; waveform windows end 30 min before the visit's first resulted lab. For the hemoglobin revisits of the long-horizon tasks, a pair is two consecutive visits of the same patient that both carry a first-draw hemoglobin.

MIMIC-III (ICU). Bedside waveforms of the MIMIC-III patients with linked waveform records (Johnson et al., 2016; Moody et al., 2020). Chart times are corrected to the waveform clock. The input for a potassium draw is the ECG lead II of the hour before it, in 30-s segments; up to three earlier potassium draws of the same stay whose hour has at least 10 valid segments serve as past windows. The split is by patient.

VitalDB (OR). Surgical cases with PPG and cardiac-output monitoring (Lee et al., 2022). Each anchor is a target reading; the input is the 10-min PPG window ending 15 min before it. Targets are stroke volume, cardiac output, stroke-volume variation and SVV$\geq$13%, a cutoff reported for stroke-volume variation in mechanically ventilated septic patients (Angappan et al., 2015). The split is by patient.

Standard PPG tasks. 24 short-window tasks from 15 public or data-use-agreement datasets, drawn from the short-window PPG tasks of the benchmark of Kataria et al. (2025), in four families: heart rate (5 tasks, MAE), blood pressure (13, systolic and diastolic MAE), atrial fibrillation (2, macro-F1), and respiration and SpO_2 (4, MAE). Each task comes with fixed train, validation and test splits; WESAD is split into five subject folds whose scores are averaged. Windows are resampled to 40 Hz and min-max normalised, with lengths from 2.1 s (PPG-BP) to 30 s. Table 5 lists the tasks.

Validation cohorts: ED hemoglobin revisits 872 pairs / 537 patients; ICU potassium 1,677 draws / 624 stays. Training: ED revisits 2,776 pairs / 1,668 patients; ICU potassium 8,008 draws. Per-lab test counts for the first-draw labs are in Tables 11 and 12.

B FLOORS

Every floor is a gradient-boosted model fitted on training patients only, with its configuration selected by patient-grouped cross-validation; the contents are listed in Table 1. The training rows' floor predictions are out-of-fold, so the waveform stage is fitted to honest residuals, and validation and test predictions come from the fit on all training rows; for the ED case-mix floor, the out-of-fold predictions use 5 patient-grouped folds and the held-out predictions average three seeds. The ICU floor includes every anchor value that native calibration uses, and the ED cross-visit floor includes the previous visit's value, so native calibration never adds lab information. Table 7 reports each floor on the held-out test patients.

C OUR ENCODER: CORPUS, OBJECTIVE AND TRAINING

Corpus. All three streams are drawn from training patients only (Table 8). The paired stream uses the ECG-bearing rows of the ICU training pack (MIMIC-III training patients of the potassium task)

Table 5: The 24 standard PPG tasks: source dataset, target, metric and number of windows (train / validation / test; WESAD summed over its five folds).

task	dataset	target	metric	windows (train / val / test)
BUT PPG HR	BUT PPG (Nemcova et al., 2021; 2024)	heart rate	MAE	2,670 / 228 / 990
DaLiA HR	PPG-DaLiA (Reiss et al., 2019)	heart rate	MAE	38,677 / 8,664 / 17,356
Gyro-Acc HR	Gyro-Acc-PPG (Lee et al., 2019)	heart rate during exercise	MAE	1,512 / 504 / 504
Welltory HR	Welltory smartphone PPG (Neshitov et al., 2021)	heart rate	MAE	176 / 59 / 117
WESAD HR	WESAD (Schmidt et al., 2018)	heart rate	MAE	176,374 / 36,256 / 54,105
Aurora-A DBP	Aurora-BP, auscultatory (Mieloszyk et al., 2022)	diastolic BP	MAE	1,506 / 503 / 503
Aurora-A SBP	Aurora-BP, auscultatory (Mieloszyk et al., 2022)	systolic BP	MAE	1,511 / 504 / 504
BCG DBP	Bed-based BCG (Carlson et al., 2021)	diastolic BP	MAE	428 / 53 / 54
BCG SBP	Bed-based BCG (Carlson et al., 2021)	systolic BP	MAE	428 / 53 / 54
BUT PPG SBP	BUT PPG (Nemcova et al., 2021; 2024)	systolic BP	MAE	2,634 / 216 / 990
CAS-BP DBP	CAS-BP (Liu et al., 2023)	diastolic BP	MAE	31,987 / 3,978 / 4,111
CAS-BP SBP	CAS-BP (Liu et al., 2023)	systolic BP	MAE	31,987 / 3,978 / 4,111
Aurora-O DBP	Aurora-BP, oscillometric (Mieloszyk et al., 2022)	diastolic BP	MAE	11,298 / 3,766 / 3,766
Aurora-O SBP	Aurora-BP, oscillometric (Mieloszyk et al., 2022)	systolic BP	MAE	12,246 / 4,082 / 4,083
PPG-BP DBP	PPG-BP (Liang et al., 2018)	diastolic BP	MAE	393 / 131 / 131
PPG-BP SBP	PPG-BP (Liang et al., 2018)	systolic BP	MAE	393 / 131 / 131
Vital Videos DBP	Vital Videos (Toye, 2023)	diastolic BP	MAE	109 / 28 / 59
Vital Videos SBP	Vital Videos (Toye, 2023)	systolic BP	MAE	109 / 28 / 59
Simband AF	Samsung Simband (Shashikumar et al., 2017)	atrial fibrillation	macro-F1	479 / 120 / 258
Stanford AF	Stanford wrist PPG (DeepBeat) (Torres-Soto & Ashley, 2020)	atrial fibrillation	macro-F1	30,502 / 7,626 / 2,986
BIDMC RR	BIDMC (Pimentel et al., 2017)	respiratory rate	MAE	2,880 / 960 / 960
BIDMC SpO_2	BIDMC (Pimentel et al., 2017)	SpO_2	MAE	2,764 / 922 / 922
MIMIC-III RR	MIMIC-III Waveform (matched) (Moody et al., 2020)	respiratory rate	MAE	12,000 / 12,000 / 12,000
Sleep SpO_2	Pediatric sleep (Phone Oximeter) (Garde et al., 2014)	SpO_2	MAE	48,000 / 48,000 / 48,000

Table 6: Held-out test cohorts. Patients are counted by medical record number (ED), stay (ICU) or subject (OR).

family / task	source	unit	held-out test
long-horizon: ED hemoglobin revisits	MC-MED	visit pairs	891 pairs / 528 patients
long-horizon: ICU potassium	MIMIC-III	draws	1,537 draws / 604 stays
first-draw labs (7)	MC-MED	visits	4,267–4,335 visits / 3,394–3,444 patients per lab
intra-operative haemodynamics (4)	VitalDB	anchors	4,746 anchors per task / 172 patients (174 cases)
standard PPG tasks	24 task datasets	windows	fixed per-task test splits

and of the ED training pack (MC-MED training patients of the ED long-horizon task and the first-draw labs). The patient-contrastive stream uses all PPG rows of the ED training pack, with the patient defined by medical record number. The PPG-only stream indexes the MIMIC-III bedside PPG (Moody et al., 2020) of the 2,128 stays that are in the training split of both the potassium task and our vital-sign split. Rows are quantised and normalised exactly as the packs used for evaluation.

Objective. For modality $m \in \{\text{PPG}, \text{ECG}\}$, $\mathcal{L}_m$ is the stage-1 objective applied to that modality with its own projector: a LeJEPA-style invariance term between 2 global and 4 local views of each row, pulled towards their centroid, with an isotropy regulariser (SIGReg, weight $\lambda = 0.05$, 1,024 random slices), plus a slot-level invariance term (weight 1) and a slot-diversity term (weight 0.5). PPG rows of both the paired and the PPG-only streams enter $\mathcal{L}_{\text{PPG}}$; ECG rows enter $\mathcal{L}_{\text{ECG}}$. The cross-modal term predicts each modality's view centroid from the other's with two small predictors, $\mathcal{L}_{\text{xm}} = \frac{1}{2}(\|p_{E\to P}(z_E) - z_P\|^2 + \|p_{P\to E}(z_P) - z_E\|^2)$, on the paired rows. The patient-contrastive term embeds an anchor window and a window of the same patient recorded at another time (log-uniform gap, at least 4×60 s; median positive gap 0.12 h, 3.8% at least 2 h) through a state projector, and applies a symmetric InfoNCE loss at temperature $\tau = 0.1$ over the batch, with SIGReg on the state projections (weight 0.05). The total loss is

$$\mathcal{L} = \mathcal{L}_{\text{PPG}} + \mathcal{L}_{\text{ECG}} + 1 \cdot \mathcal{L}_{\text{xm}} + 1 \cdot \big(\mathcal{L}_{\text{InfoNCE}} + 0.05 \cdot \text{SIGReg}(z_{\text{state}})\big).$$

Optimisation. AdamW (weight decay 0.05), trunk learning rate 0.0001, head learning rate 0.0005, gradient-norm clipping at 1. The trunk is frozen for the first 300 steps while the new heads warm up; its learning rate then warms up linearly over 300 steps and decays by a cosine to 1% of the peak. Training ran for 24,000 steps on 3 GPUs with data-parallel training (seed 1), in bfloat16 autocast. Per step the 3 GPUs together process 768 LeJEPA rows and 288 triplets. Over stage 2 the encoder saw about 18.4M PPG rows, 9.2M ECG rows and 6.9M triplets; stage 1 saw about 10M windows. The final checkpoint was used throughout, fixed in advance.

Table 7: Floor performance on held-out test patients. R^2 of the level on the modelling scale; R^2_Δ of the change from the previous value; AUROC for SVV$\geq$13%. MAE in the analyte's units.

task	floor	floor performance
first-draw Hgb, Alb, BUN, Ca	ED case mix	R^2 0.477, 0.423, 0.446, 0.177
first-draw Cr, Glu, Na	ED case mix	R^2 0.531, 0.279, 0.177
ED hemoglobin revisits	ED cross-visit	R^2_Δ 0.281; MAE 1.042 g/dL
ICU potassium	ICU lab history	R^2_Δ 0.345; level R^2 0.406; MAE 0.331 mmol/L
SV, CO, SVV, SVV$\geq$13%	OR vitals and case	R^2 0.378, 0.462, 0.209; AUROC 0.771
same, strong floor	+ arterial pressures	R^2 0.706, 0.723, 0.434; AUROC 0.846

Table 8: Stage-2 pretraining streams. Rows per step are per GPU; there are 3 GPUs.

stream	source (training patients only)	rows	rows / step	terms
paired ECG II + PPG	ICU pack; ED pack	661k; 95k	64 + 64	LeJEPA per modality, cross-modal
patient-contrastive	ED pack, all PPG rows	442k	96 triplets	InfoNCE + SIGReg
PPG-only	MIMIC-III bedside PPG, 85.8k h	10.3M	128	PPG LeJEPA

Leakage assertions. The PPG-only index excludes, by assertion, every validation and test patient of the potassium task and of our vital-sign split; the ICU training pack's split hash matches the potassium task's; the ED pack holds training patients of the ED split only. A snapshot taken during the frozen phase (step 100) has a trunk identical to the stage-1 checkpoint (maximum absolute difference 0); scored through the identical harness, it provides the Ours (initial) comparator of the level-readout families, so the two differ only in stage 2.

D ENCODERS, FRONT-ENDS AND EXPORT

Architecture and stage 1: Ours (initial). 17.3M parameters. Input: a 30-s PPG window at 40 Hz (1,200 samples), min-max normalised per window. A multi-branch stem (convolutions with kernel widths 41 and 9 on the signal, a convolution on its first difference, and per-patch mean and standard-deviation features) feeds a pre-norm rotary-position transformer trunk (width 384, two stages of depth 4, patch 8). The readout concatenates pooled statistics of an intermediate and the final layer, a slot-competition aggregate with 8 slots, and summary statistics, giving a 5378-d embedding. Stage 1 is a checkpoint of this architecture that our group pretrained earlier; stage 2 warm-starts from it rather than training both stages in one run. Stage 1 was label-free on 3M windows for 40,000 steps (batch 256, learning rate 0.0005). The objective, in the spirit of LeJEPA (Balestriero & LeCun, 2025), combines a centroid-invariance term over PPG crops and time-aligned ECG lead II views at 120 Hz (the auxiliary modality), an isotropy regulariser (SIGReg, weight 0.05, 1,024 random slices), and a slot-level invariance term with a slot-diversity term. The same trunk embeds ECG lead II segments for the ICU task.

Ours (ICU multimodal). Warm-started from stage 1 and trained for 12k further steps on the 30-s segments of MIMIC-III *training* patients that carry time-aligned ECG lead II and PPG (128 segments per step). The objective applies the pretraining terms to both modalities with a separate projector and isotropy regulariser per modality, adds cross-modal prediction between the two modalities' view centroids, and adds a within-encounter time-contrastive term on triplets of segments 1 min to 96 h apart. Three seeds. No labels are used. For the ICU task, an input is the mean embedding of up to 60 ECG segments in the hour before the draw, reduced by a PCA fit on training draws.

Ours (patient-contrastive). Warm-started from stage 1 and trained for 12k further steps on ED training patients with the patient-contrastive term alone (symmetric InfoNCE between windows of the same patient, SIGReg on the projections). Three seeds.

Ours. Stage 2 of the same architecture; Section C.

Public PPG and time-series encoders. Each public encoder receives our 30-s, 40-Hz window resampled to its own native rate with an exact polyphase resampler, passes through its documented preprocessing, is tiled into its native window length, and is mean-pooled over tiles (length-weighted

where the last tile is padded). PaPaGei-S: band-pass 0.5–12 Hz, 125 Hz, three 10-s tiles, per-tile z-normalisation, 512-d. AnyPPG: 125 Hz, three 10-s tiles, per-tile z-score, 512-d. SIGMA-PPG: band-pass 0.5–8 Hz, 50 Hz, one pass, 360-d. MantisV2 (Feofanov et al., 2026b): the window linearly interpolated to 512 samples, layer-2 combined token. MOMENT-base: 512-sample tiles with a masked tail, instance normalisation, length-weighted tile mean, 768-d. Chronos-Bolt-small (Ansari et al., 2024; Amazon Science, 2024): the 1,200 samples as context, mean over patch tokens. All are frozen and deterministic, and every front-end was fixed before any result.

Public encoders for the ICU task. One fixed frontend each, chosen before any result. MantisV2 (Feofanov et al., 2026b): the 30-s ECG lead II segment at 120 Hz is cut into three 10-s tiles, each interpolated to 512 samples; the layer-2 combined token is averaged over tiles (fp32). ECGFounder (Li et al., 2025), single-lead model: three 10-s tiles resampled to 500 Hz, z-scored per tile as in the official loader, final-stage features averaged over time and tiles; the single-lead model was trained on lead I with limb-lead augmentation and we feed lead II, the only lead available. MOMENT-base: 120-Hz segments in 512-sample tiles with a masked tail, instance normalisation, length-weighted tile mean.

Export convention. Our encoders' embeddings are stored as principal-component scores, so a probe that standardises each input dimension whitens them and gives every direction equal prior weight. One rule fixes the export for every encoder and every level family: the native export, its whitened principal components, or its unwhitened principal components after a fixed random rotation (which a standardising probe leaves spectrum-preserving), chosen by inner cross-validation on training patients of the first-draw panel and carried unchanged to all three level-readout families. The rule picks the unwhitened, rotated export for our encoders, AnyPPG and MOMENT-base and the native export for the other public encoders. On validation, whitening lowers our encoder's first-draw increments by 0.005 to 0.009 across labs, which is why the export is chosen rather than fixed. Native calibration uses, for every PPG encoder, the top 64 principal components fitted on training visits and standardised inside NC-attn; because the rotation is orthogonal, this input does not depend on the export.

E Architecture under a matched recipe

Comparator. A ResNet backbone of 5.37M parameters (its global-average-pooled 384-d output is the representation) whose stage-1 checkpoint our group pretrained with a cross-group LeJEPA recipe; the checkpoint's projector is not part of the representation and is discarded. **Stage 2.** The ResNet was continued with our stage-2 recipe verbatim (corpus and three streams, per-GPU shares, loss weights, optimiser, frozen-trunk phase, warm-up, cosine schedule to 24,000 steps, seed, number of GPUs), with only the changes the architecture forces: all heads are built at width 384; the slot-level invariance and slot-diversity terms are dropped, because the ResNet has no slots; the pooled backbone output feeds the LeJEPA and patient-contrastive terms; BatchNorm stays in training mode as in the rest of the recipe. The run was stopped after 8k steps, and both encoders were scored at 4k and 8k steps against our encoder's snapshots at the same steps, on validation patients, with the harness of Section 5.1. **Export.** Declared before any score: because the evaluation surfaces are PPG-only and the training batches end with ECG views, the ResNet's BatchNorm statistics were re-estimated label-free on 32,768 clean training PPG windows, with parameters unchanged; the row with its training statistics is also reported. The ResNet exports its native 384-d embedding with the fixed rotation of Section D; our encoder, the 512 leading principal components of its 5378-d export with the same rotation. **Stage 1 alone.** The ResNet checkpoint as-is was scored once on the held-out standard PPG tasks with the same harness; ranked in place of our encoder among the eleven rows, its mean rank is 7.21. Table 9 gives all contrasts; row-level intervals on the standard tasks resample rows within tasks and are optimistic.

F Context length on MIMIC-III vitals

Task. Charted heart rate (HR), respiratory rate (RR) and SpO_2 of MIMIC-III patients with bedside PPG (Moody et al., 2020), in subject-disjoint splits of our vital-sign cohort; held-out test 4,383, 4,394 and 4,370 chart values from 420–421 subjects. The input is the PPG before the chart time, from 30 s to 30 min, on the same anchors at every length. **Readouts.** *Windowed*: each 30-s window

Table 9: Architecture comparison in full. Standard PPG tasks: mean sign-adjusted difference in % of the task mean [95% row-bootstrap CI] and tasks better/tied/worse; intra-operative haemodynamics: difference in the mean increment over the vitals-and-case floor across four targets; first-draw labs: difference in the mean ΔR^2 over the case-mix floor across seven labs. Mean ranks are among eleven validation rows: the ResNet, our encoder at the same step, Ours (initial), the two single-stream variants and six public encoders. †Public encoder frozen as best on validation.

family	contrast	estimate	better/tied/worse
4k steps (validation)			
standard PPG tasks	Ours − ResNet, % of task mean	+7.73% [+5.37, +10.04]	10/14/0
	same, training BatchNorm statistics	+9.38% [+7.35, +11.56]	
	mean rank of 11 (ResNet; ours)	6.00; 3.33	
intra-op. haemodynamics	Ours − ResNet, mean incr.	+0.0638 [+0.0445, +0.0834]	
	mean rank of 11 (ResNet; ours)	9.75; 2.25	
first-draw labs	–	not scored at 4k	
8k steps (validation)			
standard PPG tasks	Ours − ResNet, % of task mean	+7.27% [+5.29, +9.30]	11/12/1
	same, training BatchNorm statistics	+7.76% [+5.62, +9.83]	
	mean rank of 11 (ResNet; ours)	5.98; 3.21	
intra-op. haemodynamics	Ours − ResNet, mean incr.	+0.0555 [+0.0357, +0.0752]	
	mean rank of 11 (ResNet; ours)	9.00; 2.25	
first-draw labs	Ours − ResNet, mean ΔR^2	+0.0095 [+0.0076, +0.0114]	
	mean ΔR^2 (ResNet; ours)	+0.0137; +0.0232	
stage 1 only: ResNet as-is, held-out test			
standard PPG tasks	Ours (initial) − ResNet, % of task mean	+10.05% [+7.74, +12.49]	15/5/4
	AnyPPG − ResNet, % of task mean	+7.38% [+5.45, +9.40]	15/4/5
	MantisV2† − ResNet, % of task mean	+4.67% [+2.79, +6.70]	12/6/6
	Ours − ResNet, % of task mean	+12.14% [+9.77, +14.55]	16/5/3

is embedded separately, whitened to 64 principal components fitted on training windows, and a causal GRU (hidden size 64) reads the sequence. *Direct*: the whole context enters the encoder as one input; for our encoder the readout is taken over the last 30 s after attention over the whole context, for the convolutional encoders it pools over the whole input. Heads are fitted on training anchors with a fixed configuration and five seeds; we report pooled R^2 with a 2,000-replicate subject-cluster bootstrap shared by all cells, and the slope of R^2 on log context length. Comparators are the ResNet of Section E (stage-1 checkpoint) and AnyPPG.

Results (Figure 5). With the windowed reader, our encoder has the highest R^2 at every length on every target (HR 0.872 at 30 s and 0.917 at 30 min; ResNet 0.830 and 0.899; AnyPPG 0.853 and 0.908), with intervals above zero against the ResNet on all three targets and against AnyPPG on HR and RR. On RR it also gains more from context than both comparators (difference in slope +0.0098 [+0.0056, +0.0141] against the ResNet, +0.0070 [+0.0023, +0.0117] against AnyPPG; Holm-adjusted over targets $p <.001$ and .003). On HR the ResNet gains more (-0.0056 [-0.0082, -0.0032]) from a lower starting level; on SpO_2 the slopes do not differ. For our encoder, direct long input is not better than the windowed reader at any length and falls behind from 2 min (HR at 30 min 0.889 against 0.917; RR at 15 min 0.094 against 0.176). Carrying the last chart value forward (at least 60 min old) gives R^2 0.319 for RR and 0.314 for SpO_2, above every PPG model (at most 0.177). Stage 2 of our encoder saw training-split PPG of this cohort, with validation and test subjects excluded.

G EVALUATION PROTOCOL

Every configuration was fixed on validation patients, and each family was evaluated once on held-out test patients. Comparators (the best public encoder of each family) were frozen on validation: AnyPPG for the first-draw labs and for ED hemoglobin, MOMENT-base for intra-operative haemodynamics, MantisV2 for the standard PPG tasks and ECGFounder for ICU potassium. Our encoder was evaluated at its final checkpoint, fixed in advance. Test results of two intermediate checkpoints (12k and 16k steps) were also computed to monitor training and did not affect the choice; Table 10 reports them. The ED task gives every PPG encoder the same input treatment (the top 64 training principal components). Validation-only analyses (Tables 21, 25 and 26, pairing, saturation) are

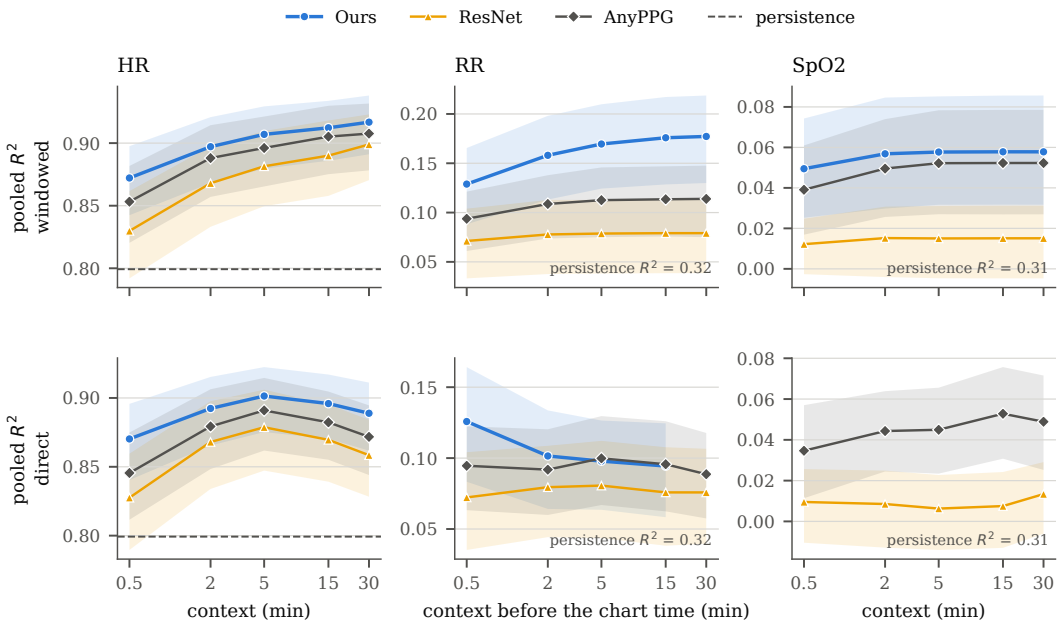

Figure 5: Pooled R^2 against context length on MIMIC-III vitals, held-out test (95% subject-cluster CI). Top: windowed reader; bottom: direct input. Dashed: carrying the last chart value forward (HR); for RR and SpO_2 its R^2 is printed. Our encoder's direct arm was not run for SpO_2 or for RR at 30 min.

exploratory. Test numbers exist only for the analytes, encoders and configurations reported in this paper.

Table 10: Our encoder at monitored stage-2 checkpoints, held-out test. First-draw labs: seven-lab mean ΔR^2 and paired differences [95% CI]; intra-operative haemodynamics: mean rank of 11 and paired differences in mean increment [95% CI]; standard PPG tasks: mean rank of 11 and the number of the 24 tasks on which it is better. The final checkpoint is the one reported throughout.

	first-draw labs			intra-op. haemodynamics			standard PPG tasks		
steps	mean ΔR^2	− initial	− AnyPPG	rank	− initial	− AnyPPG	rank	− initial	− AnyPPG
12k	+0.021	+0.0033 [+0.0020, +0.0046]	+0.0013 [−0.0002, +0.0029]	3.75	+0.0021 [−0.0077, +0.012]	+0.031 [+0.013, +0.048]	4.00	16/24	11/24
16k	+0.021	+0.0035 [+0.0022, +0.0049]	+0.0015 [+0.0000, +0.0031]	3.25	+0.0016 [−0.0085, +0.011]	+0.030 [+0.012, +0.047]	3.75	17/24	12/24
24k (final)	+0.021	+0.0037 [+0.0023, +0.0050]	+0.0016 [+0.0001, +0.0032]	3.00	+0.0038 [−0.0063, +0.013]	+0.033 [+0.015, +0.049]	3.33	18/24	13/24

H PER-TASK RESULTS FOR THE LEVEL-READOUT FAMILIES

Rows are the ten encoders of Table 2 and Ours (initial) under a second export pipeline, so ranks are among 11. Bold marks the best point estimate of a task.

H.1 FIRST-DRAW LABS

Table 11: First-draw labs, held-out test, part 1: ΔR^2 over the case-mix floor [95% patient-cluster CI].

encoder	Hgb	Alb	BUN	Ca
floor R^2	0.477	0.423	0.446	0.177
test visits / patients	4,317 / 3,426	4,267 / 3,394	4,335 / 3,444	4,328 / 3,440
Ours	**+0.049** [+0.040, +0.057]	**+0.019** [+0.012, +0.026]	**+0.016** [+0.0094, +0.021]	**+0.030** [+0.021, +0.038]
Ours (patient-contrastive)	+0.047 [+0.039, +0.056]	+0.016 [+0.011, +0.022]	+0.015 [+0.0085, +0.021]	+0.028 [+0.020, +0.037]
Ours (ICU multimodal)	+0.045 [+0.036, +0.054]	+0.016 [+0.011, +0.022]	+0.012 [+0.0064, +0.018]	+0.024 [+0.017, +0.032]
Ours (initial)	+0.042 [+0.033, +0.051]	+0.014 [+0.0083, +0.020]	+0.012 [+0.0067, +0.018]	+0.025 [+0.017, +0.033]
Ours (initial, alt. front-end)	+0.042 [+0.034, +0.051]	+0.014 [+0.0084, +0.020]	+0.012 [+0.0065, +0.018]	+0.024 [+0.017, +0.032]
AnyPPG	+0.045 [+0.036, +0.053]	+0.018 [+0.013, +0.024]	+0.014 [+0.0086, +0.020]	+0.027 [+0.019, +0.035]
MOMENT-base	+0.034 [+0.027, +0.041]	+0.011 [+0.0058, +0.017]	+0.0071 [+0.0029, +0.011]	+0.016 [+0.0091, +0.023]
MantisV2	+0.025 [+0.018, +0.032]	+0.0086 [+0.0031, +0.014]	+0.0031 [−0.0008, +0.0069]	+0.0048 [−0.0009, +0.011]
Chronos-Bolt-small	+0.029 [+0.022, +0.035]	+0.0086 [+0.0040, +0.013]	+0.0033 [−0.0005, +0.0067]	+0.0077 [+0.0022, +0.013]
SIGMA-PPG	+0.018 [+0.012, +0.024]	+0.010 [+0.0055, +0.015]	+0.0021 [−0.0016, +0.0053]	+0.0043 [−0.0005, +0.0092]
PaPaGei-S	+0.016 [+0.011, +0.021]	+0.0038 [+0.0014, +0.0062]	+0.0039 [+0.0009, +0.0066]	+0.0028 [−0.0001, +0.0059]

Table 12: First-draw labs, held-out test, part 2: ΔR^2 over the case-mix floor [95% patient-cluster CI] and the seven-lab mean.

encoder	Cr	Glu	Na	7-lab mean
floor R^2	0.531	0.279	0.177	–
test visits / patients	4,327 / 3,439	4,335 / 3,444	4,335 / 3,444	–
Ours	**+0.016** [+0.011, +0.022]	**+0.011** [+0.0033, +0.018]	+0.010 [+0.0032, +0.016]	+0.021 [+0.018, +0.025]
Ours (patient-contrastive)	+0.016 [+0.011, +0.022]	+0.010 [+0.0031, +0.017]	**+0.010** [+0.0049, +0.015]	+0.020 [+0.017, +0.023]
Ours (ICU multimodal)	+0.013 [+0.0079, +0.019]	+0.0080 [+0.0012, +0.014]	+0.0073 [+0.0018, +0.013]	+0.018 [+0.015, +0.021]
Ours (initial)	+0.013 [+0.0079, +0.018]	+0.0089 [+0.0018, +0.016]	+0.0091 [+0.0043, +0.014]	+0.018 [+0.015, +0.021]
Ours (initial, alt. front-end)	+0.013 [+0.0081, +0.018]	+0.0088 [+0.0016, +0.016]	+0.0090 [+0.0041, +0.014]	+0.018 [+0.015, +0.021]
AnyPPG	+0.016 [+0.011, +0.021]	+0.0078 [+0.0004, +0.015]	+0.010 [+0.0048, +0.015]	+0.020 [+0.017, +0.023]
MOMENT-base	+0.0041 [+0.0009, +0.0070]	+0.0056 [−0.0016, +0.012]	+0.0035 [+0.0002, +0.0067]	+0.012 [+0.0092, +0.014]
MantisV2	+0.0034 [−0.0007, +0.0078]	+0.0042 [−0.0023, +0.010]	+0.0049 [+0.0017, +0.0078]	+0.0078 [+0.0055, +0.010]
Chronos-Bolt-small	+0.0035 [+0.0001, +0.0067]	+0.0053 [−0.0006, +0.011]	+0.0035 [+0.0006, +0.0064]	+0.0086 [+0.0066, +0.011]
SIGMA-PPG	+0.0031 [−0.0004, +0.0065]	+0.0046 [−0.0013, +0.010]	+0.0069 [+0.0034, +0.010]	+0.0070 [+0.0049, +0.0090]
PaPaGei-S	+0.0032 [−0.0003, +0.0065]	+0.0043 [−0.0014, +0.010]	+0.0028 [+0.0009, +0.0047]	+0.0053 [+0.0037, +0.0069]

Table 13: First-draw labs: Ours minus each comparator, paired on identical test visits [95% CI].

contrast	Hgb	Alb	BUN	Ca
Ours − Ours (initial)	+0.0066 [+0.0031, +0.010]	+0.0052 [+0.0018, +0.0088]	+0.0032 [+0.0014, +0.0049]	+0.0048 [+0.0013, +0.0082]
Ours − Ours (initial, alt. front-end)	+0.0064 [+0.0029, +0.010]	+0.0052 [+0.0018, +0.0088]	+0.0033 [+0.0015, +0.0050]	+0.0053 [+0.0017, +0.0087]
Ours − Ours (patient-contrastive)	+0.0012 [−0.0009, +0.0034]	+0.0030 [+0.0004, +0.0057]	+0.0010 [−0.0002, +0.0021]	+0.0017 [−0.0002, +0.0035]
Ours − Ours (ICU multimodal)	+0.0038 [+0.0004, +0.0073]	+0.0030 [−0.0002, +0.0065]	+0.0035 [+0.0017, +0.0052]	+0.0053 [+0.0023, +0.0082]
Ours − AnyPPG	+0.0037 [−0.0007, +0.0081]	+0.0010 [−0.0028, +0.0051]	+0.0013 [−0.0015, +0.0039]	+0.0023 [−0.0018, +0.0064]

contrast	Cr	Glu	Na	7-lab mean
Ours − Ours (initial)	+0.0034 [+0.0016, +0.0052]	+0.0018 [+0.0000, +0.0036]	+0.0007 [−0.0030, +0.0046]	+0.0037 [+0.0023, +0.0050]
Ours − Ours (initial, alt. front-end)	+0.0032 [+0.0014, +0.0050]	+0.0019 [+0.0001, +0.0037]	+0.0008 [−0.0029, +0.0047]	+0.0037 [+0.0023, +0.0051]
Ours − Ours (patient-contrastive)	+0.0004 [−0.0007, +0.0015]	+0.0002 [−0.0007, +0.0012]	−0.0002 [−0.0028, +0.0022]	+0.0010 [+0.0002, +0.0018]
Ours − Ours (ICU multimodal)	+0.0029 [+0.0011, +0.0047]	+0.0027 [+0.0008, +0.0045]	+0.0025 [−0.0008, +0.0058]	+0.0034 [+0.0022, +0.0046]
Ours − AnyPPG	+0.0005 [−0.0020, +0.0032]	+0.0029 [+0.0003, +0.0057]	−0.0001 [−0.0038, +0.0034]	+0.0016 [+0.0001, +0.0032]

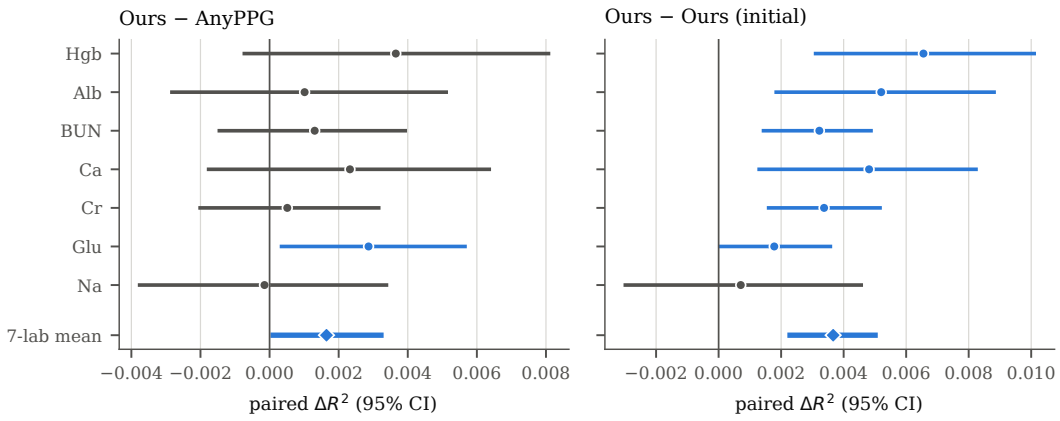

Figure 6: First-draw labs per lab: Ours minus AnyPPG (the public encoder frozen as best on validation) and minus Ours (initial), paired on identical test visits (95% patient-cluster CI). Blue: interval above zero.

Table 14: First-draw labs: rank of each encoder per lab (1 = best of 11), mean rank and wins.

encoder	Hgb	Alb	BUN	Ca	Cr	Glu	Na	mean rank	wins
Ours	1	1	1	1	1	1	3	1.29	6
Ours (patient-contrastive)	2	3	2	2	2	2	1	2.00	1
Ours (ICU multimodal)	4	4	6	6	4	5	6	5.00	0
Ours (initial)	6	5	4	4	6	3	4	4.57	0
Ours (initial, alt. front-end)	5	6	5	5	5	4	5	5.00	0
AnyPPG	3	2	3	3	3	6	2	3.14	0
MOMENT-base	7	7	7	7	7	7	10	7.43	0
MantisV2	9	9	10	9	9	11	8	9.29	0
Chronos-Bolt-small	8	10	9	8	8	8	9	8.57	0
SIGMA-PPG	10	8	11	10	11	9	7	9.43	0
PaPaGei-S	11	11	8	11	10	10	11	10.29	0

H.2 INTRA-OPERATIVE HAEMODYNAMICS

Table 15: Intra-operative haemodynamics, held-out test, vitals-and-case floor: ΔR^2 for SV, CO and SVV, ΔAUROC for SVV$\geq$13% [95% patient-cluster CI], mean rank and wins. First rows: the floor's own performance and the number of anchors / patients.

encoder	SV	CO	SVV	SVV$\geq$13%	mean rank	wins
floor	R^2 0.378	R^2 0.462	R^2 0.209	AUROC 0.771		
rows / patients	4,746 / 172	4,738 / 172	4,745 / 172	4,746 / 172		
Ours	**+0.087** [+0.050, +0.130]	+0.063 [+0.034, +0.106]	+0.149 [+0.069, +0.221]	**+0.083** [+0.054, +0.112]	3.00	2
Ours (patient-contrastive)	+0.084 [+0.048, +0.127]	+0.063 [+0.034, +0.106]	+0.168 [+0.095, +0.232]	+0.082 [+0.054, +0.109]	2.50	0
Ours (ICU multimodal)	+0.084 [+0.047, +0.128]	+0.059 [+0.030, +0.102]	**+0.168** [+0.100, +0.230]	+0.082 [+0.054, +0.110]	2.25	1
Ours (initial)	+0.079 [+0.042, +0.122]	+0.057 [+0.027, +0.100]	+0.154 [+0.084, +0.215]	+0.077 [+0.048, +0.104]	5.50	0
Ours (initial, alt. front-end)	+0.079 [+0.042, +0.122]	+0.056 [+0.028, +0.098]	+0.152 [+0.082, +0.213]	+0.077 [+0.051, +0.104]	5.50	0
AnyPPG	+0.062 [+0.030, +0.100]	+0.043 [+0.021, +0.075]	+0.100 [+0.048, +0.149]	+0.047 [+0.024, +0.071]	10.00	0
MOMENT-base	+0.083 [+0.047, +0.125]	**+0.064** [+0.037, +0.104]	+0.157 [+0.087, +0.223]	+0.075 [+0.047, +0.103]	4.00	1
MantisV2	+0.076 [+0.043, +0.112]	+0.056 [+0.032, +0.088]	+0.165 [+0.100, +0.225]	+0.079 [+0.049, +0.108]	5.25	0
Chronos-Bolt-small	+0.070 [+0.040, +0.101]	+0.047 [+0.024, +0.076]	+0.123 [+0.063, +0.178]	+0.062 [+0.038, +0.086]	8.50	0
SIGMA-PPG	+0.048 [+0.015, +0.079]	+0.042 [+0.020, +0.069]	+0.148 [+0.078, +0.213]	+0.067 [+0.041, +0.093]	9.00	0
PaPaGei-S	+0.037 [+0.012, +0.060]	+0.035 [+0.019, +0.054]	+0.122 [+0.068, +0.172]	+0.061 [+0.038, +0.084]	10.50	0

Table 16: Intra-operative haemodynamics, held-out test, strong floor (adds arterial systolic, diastolic and pulse pressure and noninvasive pressures). Layout as Table 15.

encoder	SV	CO	SVV	SVV$\geq$13%	mean rank	wins
floor	R^2 0.706	R^2 0.723	R^2 0.434	AUROC 0.846		
rows / patients	4,746 / 172	4,738 / 172	4,745 / 172	4,746 / 172		
Ours	+0.029 [+0.015, +0.045]	+0.026 [+0.016, +0.041]	+0.042 [−0.014, +0.090]	+0.040 [+0.024, +0.056]	5.75	0
Ours (patient-contrastive)	+0.029 [+0.016, +0.045]	**+0.027** [+0.016, +0.044]	**+0.070** [+0.028, +0.108]	+0.042 [+0.025, +0.059]	1.75	2
Ours (ICU multimodal)	**+0.029** [+0.015, +0.046]	+0.027 [+0.015, +0.045]	+0.065 [+0.022, +0.104]	+0.043 [+0.025, +0.061]	1.75	1
Ours (initial)	+0.028 [+0.014, +0.044]	+0.026 [+0.014, +0.044]	+0.059 [+0.014, +0.098]	+0.041 [+0.024, +0.057]	4.00	0
Ours (initial, alt. front-end)	+0.028 [+0.014, +0.043]	+0.026 [+0.014, +0.044]	+0.055 [+0.011, +0.093]	+0.041 [+0.024, +0.057]	4.50	0
AnyPPG	+0.022 [+0.010, +0.036]	+0.021 [+0.011, +0.035]	+0.041 [+0.014, +0.065]	+0.023 [+0.010, +0.035]	8.25	0
MOMENT-base	+0.020 [+0.0062, +0.035]	+0.020 [+0.0092, +0.035]	+0.047 [−0.0014, +0.091]	+0.037 [+0.021, +0.053]	7.50	0
MantisV2	+0.013 [+0.0002, +0.026]	+0.017 [+0.0078, +0.029]	+0.062 [+0.022, +0.100]	**+0.046** [+0.027, +0.064]	5.00	1
Chronos-Bolt-small	−0.0000 [−0.0010, +0.0007]	+0.016 [+0.0072, +0.027]	+0.033 [−0.014, +0.071]	+0.028 [+0.014, +0.041]	10.25	0
SIGMA-PPG	+0.0093 [−0.0039, +0.021]	+0.012 [+0.0018, +0.023]	+0.053 [+0.011, +0.089]	+0.038 [+0.023, +0.052]	8.00	0
PaPaGei-S	+0.0077 [−0.0002, +0.016]	+0.010 [+0.0013, +0.018]	+0.051 [+0.016, +0.084]	+0.030 [+0.016, +0.044]	9.25	0

Table 17: Intra-operative haemodynamics: Ours minus each comparator, paired on identical anchors [95% CI], under both floors. †Public encoder frozen as best on validation.

contrast	SV	CO	SVV	SVV$\geq$13%	4-task mean
vitals-and-case floor					
Ours − Ours (initial)	+0.0076 [+0.0013, +0.014]	+0.0064 [+0.0006, +0.012]	−0.0051 [−0.040, +0.027]	+0.0064 [−0.0015, +0.014]	+0.0038 [−0.0063, +0.013]
Ours − Ours (patient-contrastive)	+0.0028 [−0.0029, +0.0089]	−0.0002 [−0.0052, +0.0045]	−0.019 [−0.049, +0.0074]	+0.0018 [−0.0041, +0.0082]	−0.0036 [−0.013, +0.0049]
Ours − Ours (ICU multimodal)	+0.0028 [−0.0029, +0.0080]	+0.0040 [−0.0029, +0.010]	−0.019 [−0.052, +0.0090]	+0.0012 [−0.0048, +0.0069]	−0.0029 [−0.012, +0.0051]
Ours − AnyPPG	+0.025 [+0.011, +0.040]	+0.020 [+0.0084, +0.034]	+0.049 [−0.0078, +0.097]	+0.036 [+0.019, +0.053]	+0.033 [+0.015, +0.049]
Ours − MOMENT-base†	+0.0040 [−0.010, +0.019]	−0.0006 [−0.016, +0.013]	−0.0077 [−0.054, +0.031]	+0.0086 [−0.0054, +0.023]	+0.0011 [−0.016, +0.016]
strong floor (adds arterial pressures)					
Ours − Ours (initial)	+0.0012 [−0.0017, +0.0038]	−0.0000 [−0.0040, +0.0029]	−0.017 [−0.042, +0.0059]	−0.0004 [−0.0040, +0.0030]	−0.0040 [−0.010, +0.0018]
Ours − Ours (patient-contrastive)	−0.0001 [−0.0027, +0.0028]	−0.0014 [−0.0045, +0.0009]	−0.029 [−0.055, −0.0066]	−0.0018 [−0.0052, +0.0017]	−0.0080 [−0.015, −0.0022]
Ours − Ours (ICU multimodal)	−0.0001 [−0.0034, +0.0028]	−0.0009 [−0.0046, +0.0021]	−0.023 [−0.052, +0.0003]	−0.0022 [−0.0060, +0.0014]	−0.0065 [−0.014, −0.0006]
Ours − AnyPPG	+0.0069 [−0.0016, +0.015]	+0.0052 [−0.0001, +0.011]	+0.0013 [−0.042, +0.036]	+0.017 [+0.0075, +0.028]	+0.0077 [−0.0040, +0.018]
Ours − MOMENT-base†	+0.0088 [+0.0008, +0.018]	+0.0059 [−0.0005, +0.014]	−0.0052 [−0.034, +0.023]	+0.0033 [−0.0051, +0.013]	+0.0032 [−0.0062, +0.012]

H.3 STANDARD PPG TASKS

Table 18: Standard PPG tasks, held-out test scores per task (MAE: lower is better; F1: macro-F1 in %, higher is better), with summary rows. Task sources in Table 5. Columns: Ours; Ours (patient-contrastive, PC); Ours (ICU multimodal, ICU); Ours (initial, init.); Ours (initial), second front-end (init.′); AnyPPG; MOMENT-base; MantisV2; Chronos-Bolt-small; SIGMA-PPG; PaPaGei-S. ‡ResNet backbone as-is (stage-1 checkpoint of Table 3; native 384-d export with the fixed rotation, not chosen by the export rule), not ranked.

task	metric	Ours	PC	ICU	init.	init.′	AnyPPG	MOMENT	Mantis	Chronos	SIGMA	PaPaGei	*ResNet*‡
BUT PPG HR	MAE↓	**10.11**	10.36	10.18	10.40	10.45	10.13	10.55	10.16	10.37	11.01	10.88	*10.73*
DaLiA HR	MAE↓	**7.88**	8.13	8.37	7.98	7.99	9.46	10.64	10.26	11.23	12.88	16.46	*11.46*
Gyro-Acc HR	MAE↓	**19.03**	19.05	19.08	19.20	19.22	19.12	19.43	19.29	19.51	19.61	20.11	*19.54*
Welltory HR	MAE↓	6.58	6.80	5.95	6.80	6.91	6.24	5.98	**4.28**	5.67	6.64	7.84	*5.35*
WESAD HR	MAE↓	8.90	8.97	8.92	9.03	9.03	**8.78**	8.97	8.86	9.04	8.99	9.69	*8.95*
Aurora-A DBP	MAE↓	9.93	9.40	9.54	9.19	9.24	**8.91**	9.40	10.03	9.71	10.22	9.57	*12.13*
Aurora-A SBP	MAE↓	13.03	13.23	12.97	13.12	13.22	**12.68**	14.10	14.94	13.15	13.54	13.10	*17.89*
BCG DBP	MAE↓	**5.27**	6.30	6.25	6.79	6.77	6.12	5.50	6.62	6.20	5.36	6.79	*7.27*
BCG SBP	MAE↓	4.48	4.38	5.23	**3.51**	3.51	4.84	6.50	7.64	5.95	5.33	6.25	*6.22*
BUT PPG SBP	MAE↓	14.02	14.09	14.15	14.08	14.07	13.99	14.29	14.32	14.37	**13.85**	14.05	*13.80*
CAS-BP DBP	MAE↓	8.44	8.49	8.49	8.42	8.43	**8.13**	8.49	8.46	8.46	8.51	8.53	*8.54*
CAS-BP SBP	MAE↓	13.08	13.11	13.07	13.14	13.12	**12.93**	13.34	13.20	13.45	13.58	13.75	*13.66*
Aurora-O DBP	MAE↓	10.29	10.20	10.15	10.31	10.30	**9.42**	13.88	11.18	11.56	20.64	19.93	*18.37*
Aurora-O SBP	MAE↓	**19.37**	19.38	19.47	19.41	19.41	19.94	20.46	20.28	20.56	23.12	23.00	*21.91*
PPG-BP DBP	MAE↓	7.29	7.22	7.45	**6.99**	7.04	9.00	9.13	8.73	8.80	9.17	9.28	*7.98*
PPG-BP SBP	MAE↓	**12.69**	13.02	13.04	13.32	13.40	14.89	15.85	13.64	14.24	15.06	16.78	*14.20*
Vital Videos DBP	MAE↓	9.27	9.70	11.40	9.77	9.90	8.48	10.17	9.60	**8.16**	8.24	8.95	*8.23*
Vital Videos SBP	MAE↓	**19.26**	20.70	19.82	20.62	19.87	19.89	19.90	20.52	19.99	20.31	20.42	*19.89*
Simband AF	F1↑	53.43	66.37	60.75	55.83	60.87	**68.86**	61.46	67.39	62.20	63.68	65.84	*61.67*
Stanford AF	F1↑	89.41	**90.41**	88.99	89.15	89.29	89.55	82.57	78.98	74.82	74.57	70.48	*85.94*
BIDMC RR	MAE↓	**1.10**	1.17	1.14	1.17	1.16	1.62	1.63	1.63	1.86	2.11	2.09	*1.27*
BIDMC SpO_2	MAE↓	**0.83**	1.00	0.85	0.92	0.92	1.69	1.75	1.55	1.97	2.39	2.46	*1.19*
MIMIC-III RR	MAE↓	4.43	4.45	4.46	4.46	4.46	4.50	4.45	**4.23**	4.38	4.46	4.53	*4.56*
Sleep SpO_2	MAE↓	5.35	5.34	5.41	5.34	5.34	5.37	**5.04**	5.20	5.45	5.80	5.98	*5.63*
mean rank		3.33	4.83	4.96	5.54	5.25	4.00	7.17	6.21	7.12	8.33	9.25	–
fam.-bal. rank		3.86	4.20	5.24	5.99	5.55	4.02	6.74	5.23	7.24	8.55	9.38	–
wins		9	1	0	2	0	7	1	2	1	1	0	–

Table 19: Standard PPG tasks: mean rank per task family, family-balanced mean rank, overall mean rank [task-bootstrap 95% CI] and wins, among 11 rows.

encoder	HR	BP	AF	RR/SpO_2	fam.-bal.	mean rank	wins
Ours	2.40	3.31	7.00	2.75	3.86	3.33 [2.38, 4.38]	9
Ours (patient-contrastive)	5.00	5.31	2.00	4.50	4.20	4.83 [3.88, 5.79]	1
Ours (ICU multimodal)	3.80	5.15	7.50	4.50	5.24	4.96 [3.96, 6.00]	0
Ours (initial)	6.40	5.08	7.50	5.00	5.99	5.54 [4.42, 6.67]	2
Ours (initial, alt. front-end)	7.00	4.69	6.00	4.50	5.55	5.25 [4.33, 6.21]	0
AnyPPG	3.60	3.46	1.50	7.50	4.02	4.00 [2.92, 5.12]	7
MOMENT-base	6.80	7.92	7.00	5.25	6.74	7.17 [6.12, 8.04]	1
MantisV2	4.00	7.92	5.00	4.00	5.23	6.21 [4.96, 7.38]	2
Chronos-Bolt-small	7.20	7.00	7.50	7.25	7.24	7.12 [6.08, 8.04]	1
SIGMA-PPG	9.00	7.69	7.50	10.00	8.55	8.33 [7.08, 9.38]	1
PaPaGei-S	10.80	8.46	7.50	10.75	9.38	9.25 [8.21, 10.21]	0

Mean differences are dominated by a few small tasks with large relative swings, so none against a ranked comparator is significant at the task level; the median and the counts of tasks are stable. Against Ours (initial), both single-stream variants and MantisV2, our encoder is better on most tasks with a positive median; against AnyPPG it is even.

Table 20: Standard PPG tasks: Ours minus each comparator per task, in % of the task's validation mean and sign-adjusted so that positive means Ours is better. Summary rows resample the 24 tasks (20,000 task-bootstrap resamples; the family-balanced mean resamples tasks within families and weights families equally); sign and Wilcoxon signed-rank tests are one-sided. †Public encoder frozen as best on validation. ‡ResNet backbone as-is (Table 18).

task	− MantisV2†	− AnyPPG	− Ours (initial)	− Ours (PC)	− Ours (ICU)	− ResNet‡
BUT PPG HR	+0.5	+0.2	+2.8	+2.4	+0.7	+5.8
DaLiA HR	+37.2	+24.7	+1.5	+3.9	+7.6	+56.1
Gyro-Acc HR	+1.3	+0.5	+0.9	+0.1	+0.3	+2.5
Welltory HR	−69.0	−10.0	+6.8	+6.7	−18.7	−37.0
WESAD HR	−0.4	−1.4	+1.5	+0.8	+0.2	+0.5
Aurora-A DBP	+1.1	−12.1	−8.8	−6.3	−4.7	+25.8
Aurora-A SBP	+15.2	−2.8	+0.7	+1.6	−0.5	+38.7
BCG DBP	+32.6	+20.4	+36.6	+24.9	+23.6	+48.3
BCG SBP	+48.3	+5.6	−14.8	−1.4	+11.5	+26.6
BUT PPG SBP	+1.9	−0.3	+0.4	+0.4	+0.8	−1.4
CAS-BP DBP	+0.2	−3.8	−0.2	+0.6	+0.6	+1.2
CAS-BP SBP	+0.9	−1.0	+0.4	+0.3	−0.0	+4.0
Aurora-O DBP	+6.9	−6.8	+0.2	−0.7	−1.1	+63.1
Aurora-O SBP	+4.5	+2.8	+0.2	+0.1	+0.5	+12.6
PPG-BP DBP	+16.9	+20.1	−3.6	−0.8	+1.8	+8.0
PPG-BP SBP	+7.0	+16.0	+4.6	+2.4	+2.6	+11.0
Vital Videos DBP	+3.7	−9.0	+5.7	+4.9	+24.3	−11.8
Vital Videos SBP	+7.4	+3.7	+8.0	+8.4	+3.3	+3.7
Simband AF	−17.0	−18.8	−2.9	−15.7	−8.9	−10.0
Stanford AF	+11.6	−0.2	+0.3	−1.1	+0.5	+3.9
BIDMC RR	+32.5	+32.1	+4.1	+4.4	+2.3	+10.2
BIDMC SpO_2	+43.2	+52.1	+5.2	+10.1	+1.2	+21.5
MIMIC-III RR	−4.4	+1.7	+0.8	+0.6	+0.7	+3.1
Sleep SpO_2	−2.8	+0.3	−0.2	−0.2	+1.0	+5.1
mean [task-bootstrap CI]	+7.5 [−2.2, +16.2]	+4.8 [−1.0, +11.3]	+2.1 [−1.0, +5.9]	+1.9 [−0.7, +4.8]	+2.1 [−1.2, +5.5]	+12.1 [+3.6, +21.3]
median [task-bootstrap CI]	+4.1 [+0.9, +11.6]	+0.3 [−1.4, +3.7]	+0.7 [+0.2, +2.8]	+0.6 [+0.0, +2.4]	+0.7 [+0.3, +1.8]	+5.5 [+3.1, +15.9]
family-balanced mean [CI]	+4.9 [−6.0, +15.0]	+4.4 [−2.3, +11.4]	+1.5 [−0.1, +3.5]	+0.2 [−2.6, +3.0]	−0.0 [−2.9, +2.7]	–
tasks better (sign test p)	19/24 (0.003)	13/24 (0.419)	18/24 (0.011)	17/24 (0.032)	18/24 (0.011)	20/24 (0.001)
Wilcoxon signed-rank p	0.005	0.195	0.020	0.035	0.021	0.002

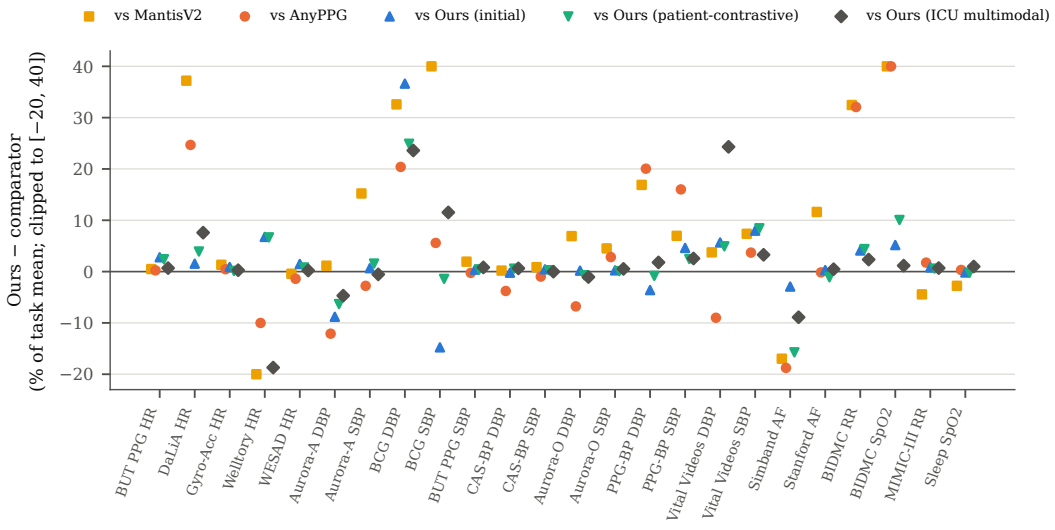

Figure 7: Standard PPG tasks per task: Ours minus each comparator in % of the task's validation mean (positive: Ours better). 4 of the 120 points lie outside the plotted range and are drawn at its edge; exact values are in Table 20.

I NATIVE CALIBRATION: SPECIFICATION, ALL ENCODERS AND CONTROLS

I.1 SPECIFICATION

Inputs. ED: the current visit's window and one past window, from the patient's previous visit with a hemoglobin value; each is the mean of the encoder's 30-s window embeddings, reduced to its top principal components fitted on training visits. ICU: the hour before the current draw and before each of up to three earlier potassium draws of the same stay whose windows have at least 10 valid segments, reduced by a PCA fit on training draws.

NC-attn. The first 64 standardised principal components of each window pass through a shared two-layer map ϕ (width 64, GELU). For each anchor, a two-layer head reads $[\phi(e_i) - \phi(e_j),\ \phi(e_i) \odot \phi(e_j),\ \log(1 + \Delta t_{ij})]$ and outputs a prediction and an attention logit (no lab value is an input); the softmax-weighted anchor predictions are added to a current-only head on $\phi(e_i)$. The target is the floor's out-of-fold training residual. Training uses AdamW (learning rate 10^{-3}, weight decay 10^{-2}, batch 256, at most 200 epochs) with early stopping (patience 15) on 15% of training patients. The model is cross-fitted in 5 patient-grouped folds and averaged over 3 initialisations; its scalar output enters the residual stage.

NC-lin. Ridge on all within-patient training pairs, penalty from 15 log-spaced values in $[10^{-1}, 10^{6}]$ by 5-fold patient-grouped cross-validation, cross-fitted as in Section 4.2. Its features are the calibrated change estimate, the mean over anchors, the level score and the number of anchors.

Residual stage. Low-dimensional feature blocks (NC-lin features, the NC-attn output, hand-crafted features) are fit to the floor's out-of-fold residual by ridge, with the penalty (8 log-spaced values in $[10^{-2}, 10^{5}]$) and a block scale chosen by 5-fold patient-grouped cross-validation, and the null model as a candidate. Embedding blocks (first-draw labs, VitalDB) use the standardised 512-dimensional PCA and choose, by nested patient-grouped cross-validation, between ridge (15 penalties in $[10^{-1}, 10^{6}]$), partial least squares (1 to 16 components) and the null model.

Donor control. For every recipient row and anchor slot, the own past window is replaced by a donor window; the recipient keeps its current window, its own anchor value, its gap and its floor features. *Donor pool:* rows of the same split (training recipients draw from training rows, held-out recipients from the same held-out split) that have an anchor at the same slot and belong to a different patient; the donor window is the donor row's own anchor window. *Matching:* on the anchor value only. Rows are sorted by value (stable sort), the candidates are the 60 sorted positions on either side minus same-patient rows, ordered by absolute value difference and then position, and the first 10 are kept; one is drawn uniformly with a seeded generator. Hemoglobin is reported to 0.1 g/dL, so most candidates are exact ties (96% exact matches on training rows, 80–83% on validation). *Reuse:* donors are drawn with replacement; 35% of training and 59% of validation recipients receive a distinct donor window (maximum reuse 11–14 and 5–6), and two donor seeds pick the same window for 11% of validation recipients. *Cross-fitting:* donors are drawn once for the training split before the patient-grouped 5-fold cross-fit; fold models are fit on fold-training (current, donor) pairs and produce held-out-fold features from those rows' donor windows. Donor windows are inputs only; the donor's value is never used. The penalty is re-selected on the donor pairs, and evaluation features come from the full-training fit with evaluation-split donors. *Averaging:* each donor seed is a separate end-to-end retrain; the donor arm is the mean of the 6 per-seed residual predictions, and own minus donor is scored on that mean. For NC-attn, each per-seed arm is itself the mean of 3 initialisations $\times$ 5 fold models. *Bootstrap:* paired patient-cluster (ICU: stay-cluster) bootstrap, 2,000 resamples of recipients with predictions fixed; donors are not resampled, and donor-draw variance is instead reduced by averaging and measured by the A/A contrast.

A pairing test with the past value supplied to the network as an input is not informative, because in the donor arm that input also carries the patient's recent lab trend; we do not use it.

Better-matched donors. Demographically matched donors are other patients in the same split within 0.2 g/dL of the anchor value, of the same sex, age band ($<$40, 40–59, 60–74, $\geq$75 years) and tertile of the gap to the previous visit (training cut points); the 10 value-nearest are kept and one is drawn at random. If fewer than 3 candidates remain, the gap, then the age, then the sex and value constraints are relaxed in turn; 353 validation rows (40%) are matched on all criteria. Identity-matched (look-alike) donors are drawn from the 30 value-nearest patients as one of the 3 nearest in

patient-mean embedding over their other visits, excluding the donated window and the recipient's replaced window.

Estimator baselines (Table 21). First-difference OLS regresses within-patient changes on the leading principal components (number chosen by grouped cross-validation). The untied reader is ridge on $[e_i, e_j, \log \Delta t]$ with separately tuned current and past penalties. The random-intercept model fits a level ridge on both visits and shrinks the patient's past residual by a factor estimated from the training residual covariance. The *carried-residual* model predicts the level from the current window with a population ridge and adds the patient's residual at the previous visit. Gradient boosting reads the first 32 components of each window, the past value and the log gap.

Table 21: Estimators on identical inputs and nesting (ED Hgb, validation, 872 pairs / 537 patients, Ours (initial)). Each row is retrained end to end, including its own donor arm (6 donor draws). Last column: NC-lin minus the row, own-past arms. †Same features, PCA, penalty grid and cross-fitting as NC-lin; only the estimation of the map differs (levels instead of within-patient differences).

estimator	own past over floor	own − donor	NC-lin − row
NC-attn	+0.044 [+0.022, +0.065]	+0.030 [+0.013, +0.047]	–
NC-lin: tied within-patient difference ridge	+0.054 [+0.030, +0.078]	+0.027 [+0.0061, +0.047]	–
first-difference OLS on leading PCs	+0.045 [+0.022, +0.068]	+0.031 [+0.010, +0.052]	+0.0087 [−0.0012, +0.018]
untied: separate current and past penalties	+0.034 [+0.0058, +0.063]	+0.020 [−0.0065, +0.046]	+0.020 [+0.0059, +0.035]
random-intercept (BLUP) level model	+0.041 [+0.020, +0.062]	+0.0089 [−0.0004, +0.018]	+0.013 [−0.015, +0.040]
carried residual, matched budget†	+0.038 [+0.018, +0.057]	+0.012 [+0.0019, +0.022]	+0.016 [−0.012, +0.042]
population model + carried residual	+0.038 [+0.017, +0.058]	+0.011 [+0.0009, +0.021]	+0.016 [−0.012, +0.043]
gradient boosting on $[e_i, e_j, y_j, \log \Delta t]$	+0.0087 [−0.0024, +0.019]	+0.0034	+0.045 [+0.022, +0.068]
generic ridge/PLS on $[e_i, e_j]$	–	+0.0071 [−0.0062, +0.022]	–

I.2 All encoders on the held-out tasks

Table 22: ED hemoglobin revisits, NC-attn, held-out test (891 pairs / 528 patients), every PPG encoder with the same input treatment: ΔR^2_Δ [95% patient-cluster CI]. The patient-contrastive encoder's row averages its three seeds' predictions; per-seed rows follow.

encoder	own − donor	own − current	own vs floor
Ours	+0.038 [+0.023, +0.053]	+0.030 [+0.017, +0.043]	+0.063 [+0.043, +0.084]
Ours (patient-contrastive, 3 seeds)	+0.037 [+0.020, +0.054]	+0.032 [+0.016, +0.047]	+0.065 [+0.042, +0.088]
seed 1	+0.036 [+0.019, +0.053]	+0.031 [+0.015, +0.047]	+0.064 [+0.041, +0.088]
seed 2	+0.036 [+0.020, +0.052]	+0.028 [+0.014, +0.042]	+0.065 [+0.043, +0.087]
seed 3	+0.037 [+0.019, +0.056]	+0.036 [+0.018, +0.054]	+0.064 [+0.040, +0.088]
Ours (initial)	+0.029 [+0.013, +0.046]	+0.027 [+0.012, +0.044]	+0.049 [+0.029, +0.070]
AnyPPG	+0.025 [+0.010, +0.040]	+0.021 [+0.0073, +0.035]	+0.045 [+0.026, +0.065]
MOMENT-base	+0.012 [+0.0023, +0.022]	+0.011 [+0.0014, +0.019]	+0.033 [+0.018, +0.047]
PaPaGei-S	+0.0042 [+0.0014, +0.0076]	+0.0037 [+0.0009, +0.0067]	+0.011 [+0.0057, +0.016]
Chronos-Bolt-small	+0.0029 [−0.0033, +0.0092]	−0.0013 [−0.0064, +0.0039]	+0.019 [+0.0089, +0.030]
MantisV2	+0.0027 [−0.0052, +0.011]	+0.0008 [−0.0049, +0.0067]	+0.015 [+0.0016, +0.029]
SIGMA-PPG	+0.0013 [−0.0010, +0.0039]	+0.0039 [+0.0008, +0.0072]	+0.0059 [+0.0021, +0.010]

Table 23: ICU potassium from ECG, NC-attn, held-out test (1,537 draws / 604 stays): ΔR^2_Δ [95% stay-cluster CI].

encoder	own − donor	own − current	own vs floor
Ours	+0.011 [+0.0057, +0.018]	+0.0079 [+0.0026, +0.014]	+0.037 [+0.022, +0.054]
Ours (ICU multimodal, 3 seeds)	+0.014 [+0.0073, +0.021]	+0.012 [+0.0050, +0.019]	+0.043 [+0.027, +0.061]
Ours (initial)	+0.0080 [+0.0027, +0.014]	+0.0058 [+0.0004, +0.012]	+0.032 [+0.018, +0.048]
ECGFounder	+0.011 [+0.0064, +0.018]	+0.011 [+0.0057, +0.016]	+0.020 [+0.010, +0.032]
MOMENT-base	+0.010 [+0.0058, +0.015]	+0.0077 [+0.0034, +0.012]	+0.030 [+0.021, +0.041]
MantisV2	+0.0035 [+0.0016, +0.0057]	+0.0028 [+0.0011, +0.0046]	+0.010 [+0.0061, +0.016]

Table 24: Native calibration: Ours minus each comparator, paired on identical held-out rows [95% CI]. For the ICU task only the own-minus-donor contrast was compared.

contrast	own − donor	own − current	own vs floor
ED hemoglobin from PPG (identical test pairs)			
Ours − AnyPPG	+0.013 [+0.0031, +0.024]	+0.0091 [+0.0002, +0.018]	+0.018 [+0.0064, +0.030]
Ours − Ours (initial)	+0.0084 [+0.0010, +0.016]	+0.0023 [−0.0061, +0.011]	+0.014 [+0.0035, +0.023]
Ours − Ours (patient-contrastive)	+0.0011 [−0.0060, +0.0078]	−0.0019 [−0.0087, +0.0050]	−0.0020 [−0.010, +0.0062]
ICU potassium from ECG (identical test draws)			
Ours − Ours (initial)	+0.0034 [−0.0013, +0.0084]	–	–
Ours − Ours (ICU multimodal)	−0.0022 [−0.0056, +0.0011]	–	–

In the ICU task, the multimodally pretrained encoders give the largest gains over the floor, and the own-past contrast is above zero for every ECG-capable encoder (Table 23). In the ED task, the gain over the floor and the own-past contrast track each other across encoders: encoders that add little over the floor also add little from the own past (Table 22).

I.3 Stricter donors with public encoders

Table 25: Own past against stricter donors and against hand-crafted features (ED Hgb, validation, Ours (initial), 6 donor draws each, retrained end to end). Entries: own minus the stated donor arm (or FM minus hand-crafted), ΔR^2_Δ with 95% CI. The same comparisons with AnyPPG and ECGFounder are in Table 26.

comparison	NC-lin	NC-attn
own − value-matched donor	+0.027 [+0.0061, +0.047]	+0.030 [+0.013, +0.047]
own − donor matched on value, sex, age band, gap	+0.026 [+0.0058, +0.046]	+0.028 [+0.011, +0.045]
own − look-alike donor (nearest patient-mean embedding)	+0.030 [+0.0086, +0.051]	+0.031 [+0.014, +0.048]
FM − hand-crafted features (own − donor)	+0.026 [+0.0076, +0.043]	+0.027 [+0.011, +0.043]

Table 26: Own past against stricter donors, validation, for Ours (initial) and two public encoders: own minus value-matched donor; own minus donor matched on value, sex, age band and gap; own minus look-alike donor (nearest in patient-mean embedding); and the own-minus-donor contrast of the FM minus that of hand-crafted features. ΔR^2_Δ [95% CI], 6 donor draws each, retrained end to end.

encoder	estimator	value-matched	demographically matched	look-alike	FM − hand-crafted
ED hemoglobin from PPG (validation, 872 pairs)					
Ours (initial)	NC-lin	+0.027 [+0.0061, +0.047]	+0.026 [+0.0058, +0.046]	+0.030 [+0.0086, +0.051]	+0.026 [+0.0076, +0.043]
Ours (initial)	NC-attn	+0.030 [+0.013, +0.047]	+0.028 [+0.011, +0.045]	+0.031 [+0.014, +0.048]	+0.027 [+0.011, +0.043]
AnyPPG	NC-lin	+0.024 [+0.0052, +0.042]	+0.021 [+0.0039, +0.038]	+0.023 [+0.0043, +0.040]	+0.022 [+0.0048, +0.039]
AnyPPG	NC-attn	+0.023 [+0.0061, +0.039]	+0.020 [+0.0041, +0.036]	+0.024 [+0.0072, +0.041]	+0.020 [+0.0056, +0.035]
ICU potassium from ECG (validation, 1,677 draws)					
ECGFounder	NC-lin	+0.014 [+0.0018, +0.026]	+0.015 [+0.0022, +0.028]	+0.014 [+0.0026, +0.026]	+0.017 [+0.0049, +0.031]
ECGFounder	NC-attn	+0.010 [+0.0044, +0.018]	+0.010 [+0.0038, +0.017]	+0.0089 [+0.0032, +0.016]	+0.011 [+0.0047, +0.018]

For both public encoders, every stricter donor leaves the own-past contrast essentially unchanged, and hand-crafted features give none of it. The own past window is closer to the current window than the donor's in each encoder's space (AnyPPG, NC-attn input: 9.2 against 10.6; ECGFounder: 5.4 against 10.8), the same pattern as for our encoder (Section 6).

I.4 Donor draws: sensitivity and draw noise

The patient-cluster bootstrap holds the drawn donors fixed, so it does not include the variability of the draws themselves. For our encoder we therefore drew 12 donors per anchor (the 6 of the main analysis and 6 more with new donor seeds) and retrained the donor arm for each. *Sensitivity:* in the ED, own minus donor is +0.038 with draws 1–6 (the frozen configuration and headline), +0.035 [+0.021, +0.049] with all 12, and between +0.030 and +0.039 (median +0.035) over all 924 subsets

of six; in the ICU the corresponding values are +0.011, +0.011 [+0.0051, +0.017], and +0.010 to +0.012. Donor draws depend only on the anchor values, patients and seeds, not on the embeddings, so every encoder in a task sees the same donors and the paired comparisons are draw-matched. *Draw noise:* Table 27 reports the standard deviation of the A/A contrast, one donor arm minus another built from disjoint draws, over all ways of splitting the 12 draws. With six draws per arm it is 0.0035 in the ED and 0.0007 in the ICU, about $1/\sqrt{6}$ of the single-draw value, as expected for pure draw noise; the own-minus-donor contrast with 12 draws is about 10 and 16 times this standard deviation. The observed A/A contrasts of the main analysis fall inside the 95th percentile of this spread.

Table 27: Donor-draw noise for our encoder, held-out test: standard deviation of the A/A contrast (difference between two donor arms built from disjoint draws) over all splits of 12 draws, with arms of 1, 3 and 6 draws (95th percentile of |A/A| in parentheses), and the observed A/A contrasts of draws 1–3 vs 4–6 and 1–6 vs 7–12. ΔR^2_Δ.

task	sd, 1 vs 1	sd, 3 vs 3	sd, 6 vs 6 (p95)	observed 3 vs 3	observed 6 vs 6
ED hemoglobin (ours)	0.0078	0.0048	0.0035 (0.0067)	−0.0072	−0.0057
ICU potassium (ours)	0.0015	0.0010	0.0007 (0.0013)	−0.0005	−0.0012

I.5 Negative controls

Each setting has a negative-control analyte with no plausible pathway to the waveform over the same horizon: alanine aminotransferase (ALT) between ED visits (873 pairs) and sodium within an ICU stay (1,469 draws). They use the same floors, estimators, donors and resamples as the primary analytes. All are null, for our encoder and for the other encoders (Table 28); its hemoglobin contrast is about 69 times its ALT contrast.

Table 28: Negative-control analytes: own minus donor, ΔR^2_Δ [95% CI]. *ED rows from the earlier input treatment (the encoder's own PCA basis for Ours (initial); AnyPPG's first native dimensions) rather than the common top-64 treatment of Table 22.

encoder	estimator, split	own − donor
ED: alanine aminotransferase (ALT) between visits, own − donor		
Ours	NC-attn, test	+0.0005 [−0.0003, +0.0014]
Ours (initial)*	NC-lin, test	−0.0003 [−0.0015, +0.0009]
Ours (initial)*	NC-attn, test	+0.0002 [−0.0011, +0.0014]
AnyPPG*	NC-lin, test	−0.0010 [−0.0037, +0.0014]
AnyPPG	NC-attn, validation	−0.0004 [−0.0013, +0.0006]
ICU: sodium within a stay, own − donor		
Ours	NC-attn, test	−0.0002 [−0.0006, +0.0002]
Ours (ICU multimodal, 3 seeds)	NC-attn, test	+0.0003 [−0.0000, +0.0006]
Ours (initial)	NC-attn, test	−0.0000 [−0.0009, +0.0009]
MantisV2	NC-attn, test	+0.0000 [+0.0000, +0.0000]
ECGFounder	NC-attn, test	+0.0005 [−0.0000, +0.0011]
MOMENT-base	NC-attn, test	−0.0000 [−0.0001, +0.0001]

I.6 Saturation across encoder checkpoints

To ask whether the long-horizon tasks can rank encoders, we scored native calibration (NC-lin, validation) for many checkpoints of our encoders: 34 checkpoints on ED hemoglobin and 21 on ICU potassium, spanning the stage-1 checkpoint, stage-2 variants and intermediate snapshots. On ED hemoglobin, own minus donor ranges from 0.0347 to 0.0444; its standard deviation across checkpoints is 0.0019, below the typical bootstrap standard error of one checkpoint (0.0115). On ICU potassium the spread is 0.0023 against a standard error of 0.0031. Differences between closely related checkpoints are therefore within the noise of a single task, which is why encoder comparisons in the paper use paired differences on identical draws.

I.7 MULTIPLICITY

The paper's two headline native-calibration contrasts for our encoder (ED hemoglobin and ICU potassium, own minus donor) are reported with Bonferroni-adjusted intervals at $\alpha = 0.05/2$, computed as exact percentiles of the same cluster bootstrap: [+0.021, +0.055] and [+0.0047, +0.019]. All other intervals are unadjusted 95% intervals and are descriptive.

J CLINICAL UTILITY

The clinical metrics are computed from the saved held-out predictions without refitting. Intervals use the patient-cluster bootstrap (2,000 resamples; stays in the ICU), and paired contrasts use the same resamples. Sensitivity at 90% specificity sets the threshold at the negatives' 90th percentile within each resample. For native calibration, "own" is the NC-attn prediction added to the floor; for the first-draw labs, our encoder's level readout added to the floor.

Table 29: Mean absolute error on held-out test patients: floor, our encoder (Ours), their difference [95% CI] with relative change, and Ours minus AnyPPG where AnyPPG was run.

task	floor	Ours	Ours − floor	Ours − AnyPPG
ED revisits, Hgb (g/dL)	1.042	0.995	−0.048 [−0.065, −0.030] (−4.6%)	−0.013 [−0.024, −0.0024]
First-draw labs, Hgb (g/dL)	1.242	1.185	−0.056 [−0.069, −0.044] (−4.5%)	−0.0028 [−0.010, +0.0033]
ICU, K (mmol/L)	0.331	0.319	−0.011 [−0.016, −0.0068] (−3.4%)	–

Table 30: Event detection on held-out test patients (events / rows): AUROC and sensitivity at 90% specificity for the floor and our encoder (Ours), Ours minus floor and Ours minus AnyPPG [95% CI]. AnyPPG was not run in the ICU task.

task	event	metric	floor	Ours	Ours − floor	Ours − AnyPPG
First-draw labs	Hgb <10 (629/4,317)	AUROC	0.868	0.888	+0.020 [+0.015, +0.027]	+0.0026 [+0.0002, +0.0050]
First-draw labs	Hgb <10 (629/4,317)	sens@90	0.652	0.668	+0.016 [−0.0016, +0.049]	−0.011 [−0.026, +0.011]
First-draw labs	Hgb <8 (175/4,317)	AUROC	0.867	0.891	+0.024 [+0.014, +0.034]	+0.0039 [+0.0004, +0.0077]
First-draw labs	Hgb <8 (175/4,317)	sens@90	0.623	0.703	+0.080 [+0.019, +0.122]	+0.040 [+0.0000, +0.076]
ICU	K ≤3.5 (227/1,537)	AUROC	0.800	0.822	+0.022 [+0.012, +0.033]	–
ICU	K ≤3.5 (227/1,537)	sens@90	0.414	0.489	+0.075 [+0.0088, +0.142]	–

The robust gains are the error reductions, detection of low potassium in the ICU and of first-draw hemoglobin below 8 g/dL. We do not report event-level metrics for hemoglobin change between ED visits or for high potassium, which is too rare in the ICU test set (48 draws ≥5.5 mmol/L) to assess.

K PRETRAINING OVERLAP AND EXPOSURE

We joined evaluation patients against every pretraining manifest we control. *ICU potassium:* none of the validation (1,677 draws) or test (1,537 draws) stays or patients appears in the MIMIC-III part of PulseDB (Wang et al., 2023) or in the stage-2 ICU data (the joins were verified on training stays, which do appear). *ED hemoglobin:* 518 of 528 test patients have visits in an MC-MED waveform collection (unlabelled) that may be part of the stage-1 pretraining corpus; the non-overlapping remainder is too small to analyse. *VitalDB:* 172 test patients, all but 12 of them PulseDB subjects whose waveforms, without labels, are in the stage-1 pretraining corpus; this exposure applies to every row of ours in the intra-operative haemodynamics family, including our encoder. The stage-1 corpus has no identifiers that join to MIMIC-III or MC-MED, so its overlap with those tasks cannot be determined. Stage 2 of our encoder and of its single-stream variants used training-split patients only. Public encoders are "independent" only in that we did not train them. AnyPPG's pretraining lists MC-MED and PulseDB, which is built from MIMIC-III and VitalDB (Nie et al., 2026; Wang et al., 2023). PaPaGei's pretraining lists VitalDB and the MIMIC-III waveform matched subset (Pillai et al., 2025). These are the sources their papers list; their overlap with our tasks at the patient level was not determined. ECGFounder's paper lists MIMIC-IV-ECG as a downstream set.

L COMPUTE AND REPRODUCIBILITY

Our encoder. Stage 2: 24,000 steps with data-parallel training on 3 NVIDIA RTX PRO 6000 Blackwell (96 GB) GPUs, at 0.51 s per step and 55.6 GB of memory per GPU, about 3.4 h of wall-clock time. **Evaluation.** Embedding and scoring run on one GPU and CPUs: every probe, floor, NC-attn model and bootstrap is small. Floors are gradient-boosted models on tabular features; NC-attn has two layers of width 64. **Determinism.** Every stochastic choice (splits, donors, bootstrap resamples, initialisations, the export rotation) uses a fixed seed. The donors and bootstrap resamples are shared across encoders within a task, so paired differences are computed on identical rows and resamples. Public encoders run in evaluation mode without stochastic operations. **Release.** Every number, table and figure in this paper is generated by one script from result files; the script, the scoring code, the floors, the donor-control code and the encoder weights will be released.