How Much Can Sampling-Aware Uncertainty Buy?
A Measurable Ceiling for Irregularly Sampled Time Series

Anonymous authors
Paper under double-blind review

Abstract

A large line of work conditions predictive uncertainty on observation density: a model should be less certain where a channel is sampled too sparsely to resolve its variation. We measure what that idea is worth on four irregularly-sampled benchmarks and report three findings. *Units decide the reading.* The density–variance sign inversion that motivates such heads is present on PhysioNet's raw per-channel scales and absent under the z-scoring the baselines use (-0.12 to $+0.58$, every channel positive). And a selective-risk comparison on a per-channel standardized loss must rank by the variance in the loss's units: ranked by the raw σ, as our own boards were, our head appeared to beat a deep ensemble by 34%; aligned, the two are indistinguishable. *A ceiling and a floor answer different questions.* The tower property caps a variance head reading the observation set S alone at $\mathrm{corr}(\hat{\sigma}^2, e^2) \leq \sqrt{s}$. We estimate the correlation ratio s_T of a coarsening T of S — channel, horizon decile, density tercile — with a split-half statistic over whole sequences: on PhysioNet $\sqrt{s_T}$ is 0.120 raw and 0.045 z-scored. A cross-fitted S-only abstention rule lowers the normalised risk–coverage area by at least 0.377 on PhysioNet and 0.208 on USHCN for the mean of five fitted models, and a small ceiling does not bound this floor. *Both are validated, including where they fail.* Against a closed-form truth, interval coverage holds in heavy tails only because two biases cancel, falls to 84–88% in light tails, and collapses when cells hold fewer than 50 targets. On six forecasters' residuals, on three benchmark settings, the ceiling is similar while the floor depends on the forecaster. Our own density-conditioned head establishes no gain from its density input, and on PhysioNet the S-only lookup outranks its variance.

1 Introduction

Irregularly sampled multivariate time series (IMTS) — clinical vitals and labs, climate stations, asynchronous sensors — break the regular-grid assumption behind most forecasting architectures. The widespread response is to make predictive uncertainty *sampling-aware*: a model should be less certain where a channel is observed too sparsely to resolve its variation. HeTVAE (Shukla & Marlin, 2022) conditions a heteroscedastic head on sparsity; Gaussian processes are density-adaptive by construction (Rasmussen & Williams, 2006); a line of IMTS forecasters feeds observation density to a variance head. This paper asks the question that should come first. *How much of the forecast error can the observation set explain at all?* We answer it for heads that read a coarse summary T of the observation set; for richer readings of it our estimate is a lower bound, so a small value bounds what those heads can track and says nothing about what methods reading more can gain.

The premise deserves a look before the question does. One motivation for density-conditioned variance, and the one that motivated our own head — predictive variance falling where sampling is sparse — is, on the benchmark where we observed it, a property of raw per-channel scale: z-scoring PhysioNet's channels as every baseline does moves the pooled head's density–variance correlation from -0.12 to $+0.58$, every channel positive (Appendix I.0.3). Whatever sampling-awareness buys, it does not buy a fix for that.

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

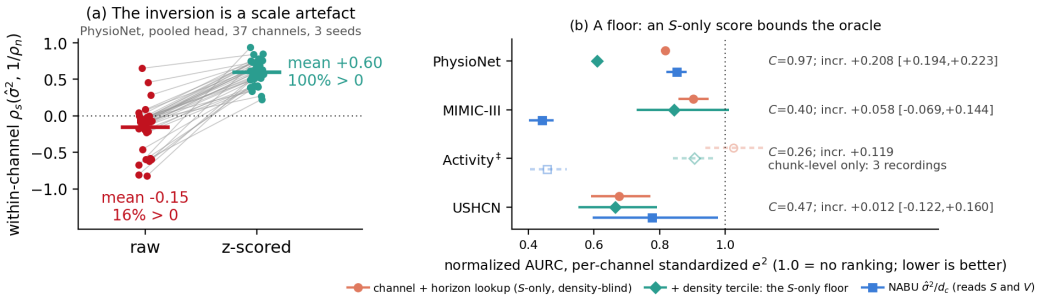


Figure 1: **(a) The density–variance inversion is a preprocessing artefact.** Within-channel $\rho_s(\hat{\sigma}^2, 1/\rho_n)$ of a pooled head (NABU off; it reads the raw per-channel density, values and time) on PhysioNet, one dot per channel averaged over three seeds, both arms trained the same day with `dataset.normalize` the only difference (Appendix I.0.3). On raw per-channel units the head is "inverted" — variance falls where sampling is sparse — on 84% of channels; under the per-variable z-scoring every baseline uses, all 37 channels are positive. **(b) A floor.** Normalized AURC in risk–coverage on per-channel standardized e^2 ($1.0 =$ no ranking), sequence-level cluster-bootstrap intervals. An S-only score bounds the oracle, so at the floor lowers normalised AURC by ≥ 0.377 on PhysioNet and 0.208 on USHCN for the mean of five fitted models, and is *not establishable* on MIMIC, whose interval crosses 1.0, or on Activity‡, whose test split is three recordings (dashed, hollow). The density increment is distinguishable from zero on PhysioNet only. NABU reads S *and* V, so its distance from either lookup identifies no density pathway; the contrast that does deletes the input from an otherwise identical head (Appendix J). C is the units-confound index of Appendix M. No relationship between the four rows is claimed; none is supportable at $n{=}4$ (Appendix H.0.2).

Writing S for the observation set and $v(S) = \mathbb{E}[e^2 \mid S] - \mathbb{E}[e^2]$, the tower property caps any variance head reading S alone at $|\operatorname{corr}(\hat{\sigma}^2, e^2)| \leq \sqrt{s}$ with $s = \operatorname{Var}[v]/\operatorname{Var}[e^2]$, attained at $\mathbb{E}[e^2 \mid S]$ — so $\sqrt{s}$ is exactly the prize. The inequality is two lines; all the difficulty is in estimating it. A naive contrast is inflated by its own noise floor, confounded with forecast horizon ($1/\rho_n$ correlates $+0.70$ with lead time on our data), and reported in a metric the inequality does not bound, and we made each of those mistakes in print before the measurement below (Appendix K). BandRate, the architecture the measurement is instantiated on, is described in Section 3.2 and Appendix N; it is not the contribution, and the measurement does not depend on it.

Our contributions:

- **Units decide the reading, twice.** The negative density–variance correlation that sampling-aware heads are built to repair is specific to raw per-channel units: z-scored, it is $+0.58$ on PhysioNet with every channel positive, and PhysioNet is the only benchmark of four whose raw statistic is negative (Appendix I.0.3). And a variance ranked against a per-channel standardized loss must be put in that loss's units, $\hat{\sigma}^2/d_c$. Our own boards ranked by the raw σ; aligned, our head's apparent 34% lead over a deep ensemble becomes an indistinguishable 8%, while no comparison with the S-only lookup reverses (Section 4.1.7, Table 13). Both are properties of the evaluation, not of any model.
- **A ceiling and a floor, which answer different questions.** We estimate s_T, the correlation ratio of $T = (\text{channel}, \text{horizon decile}, \text{density tercile})$, with a split-half U-statistic over whole sequences: $\sqrt{s_T}$ is the exact ceiling for a head reading only T and a *lower* bound on the ceiling for heads reading all of S. On PhysioNet it is 0.120 raw and 0.045 z-scored, and density's increment is distinguishable from zero on the raw scale only. An exhibited, cross-fitted S-only score bounds the oracle in risk–coverage (Corollary 2), so abstaining on the observation set alone lowers the normalised area by at least 0.377 on PhysioNet and 0.208 on USHCN for the mean of five fitted models; density adds 0.208 of that on PhysioNet only. A small ceiling does not bound the floor (Appendix O).
- **Validation, including where it fails.** Against a closed-form truth the ceiling's interval covers in heavy tails only because two biases cancel, undercovers in light tails because tercile edges are estimated, and fails below 50 targets per cell, which the released tool now flags (Appendix). A held-out lookup sits 0.0020 below its matched ceiling when both are

debiased alike (Section 4.1.3). On six forecasters' residuals on raw PhysioNet, z-scored PhysioNet and z-scored MIMIC-III, the ceiling is similar and the floor depends on the forecaster (Table 4). Our own head gains nothing establishable from its density input, and on PhysioNet the S-only lookup outranks its variance even recalibrated to the same resolution (Appendix J, Appendix Z).

We release the estimator as a framework-free tool (Appendix). Every falsified rescue, fired gate and retracted claim is recorded, with a ledger of all twenty-six checked by script (Appendix K); the body below states what we now find, and that appendix states what we found before and why it was wrong.

2 Related Work

IMTS forecasting and sampling-aware uncertainty. Recurrent, ODE, attention, graph and patch models handle irregularity differently (Che et al., 2018; Rubanova et al., 2019; Shukla & Marlin, 2021; Biloš et al., 2021; Schirmer et al., 2022; Yalavarthi et al., 2024; Li et al., 2025; Zhang et al., 2024; Liu et al., 2026); our measurement consumes residuals, whichever produces μ. Heteroscedastic heads have known optimization hazards (Seitzer et al., 2022; Stirn et al., 2023; Detlefsen et al., 2019); proper scoring rules underpin our evaluation (Gneiting & Raftery, 2007). Closest to us, HeTVAE (Shukla & Marlin, 2022) conditions a heteroscedastic layer on observation sparsity and *reports that the pathway pays*: removing its intensity encoding costs $0.554 \rightarrow 0.654$ NLL on PhysioNet and $0.666 \rightarrow 0.743$ on MIMIC-III, a drop its authors call large — it does not agree with our null, whatever we once said. LLapDiff (You et al., 2026) agrees: deleting its sampling-irregularity token is the largest degradation in its ablation (PhysioNet CRPS $0.318 \rightarrow 0.498$). Closest in aim, CircuITS (Klötergens et al., 2026) makes the joint predictive distribution over an IMTS forecast valid by construction; it is a probabilistic forecaster, not a density-conditioned variance head, and so sits outside what Corollary 1 bounds — but it is the kind of model a practitioner would weigh against this family, and our floor applies to any abstention rule it feeds. **This bounds the reach of our result and we state it rather than finesse it:** both condition on sampling in ways Corollary 1 does not cover — the signal feeds the *mean* as well as the variance, the task is interpolation or full-distribution forecasting, and the metric is a proper scoring rule. Our ceiling constrains a variance head reading the observation set *alone*, scored by correlation with realized error, forecasting off the sampling domain — narrower than "sampling-aware uncertainty". The two closest published systems sit outside it, and so does NABU itself, which reads observed values V as well: for NABU, as for them, s_T bounds only what the T-reading part of the head can track. Iversen et al. (2025) are distinct again, attributing a Gaussian head's *predicted* variance to input features where we bound correlation with *realized* error.

Variance decomposition and conformal prediction. s is a correlation ratio, a first-order Sobol index (Sobol', 1993; Saltelli et al., 2010), and the framing — measure a signal's share before adopting the family that conditions on it — has a longer history than we first credited: Song et al. (2010) bound mobility predictability by entropy, and Watson et al. (2023) and SAGE (Covert et al., 2020) give Shapley decompositions of predictive uncertainty and loss of which our density marginal is a single-ordering special case, the contrast being LOCO (Lei et al., 2018) on ρ and the R^2 increment a $\mathcal{V}$-information estimate (Xu et al., 2020; Ethayarajh et al., 2022). Our claim is narrower than novelty of framing: identifying the observation set as the signal whose share caps this pathway, an estimator for it on IMTS, and the measurement. Conformal supplies distribution-free coverage (Vovk et al., 2005; Lei et al., 2018; Papadopoulos et al., 2011; Boström & Johansson, 2020; Stankeviciute et al., 2021), exact conditional coverage being unattainable (Barber et al., 2021); baselines are deep ensembles (Lakshminarayanan et al., 2017), MC-dropout (Gal & Ghahramani, 2016), quantile/CQR (Koenker & Bassett, 1978; Romano et al., 2019) and a Gaussian head (Nix & Weigend, 1994), all identically calibrated.

3 Method

3.1 Problem and notation

An IMTS sample is $\mathcal{D} = \{(x_{n,i}, t_{n,i})\}_{n \le N, i \le L_n}$ with channel-specific irregular timestamps. Given history on $[0, T_h]$, we predict the distribution of $y_n(\tau)$ at arbitrary query times $\tau \in (T_h, T]$. All

components consume raw timestamps; no interpolation onto a grid occurs anywhere in the pipeline (Figure 3).

3.2 NABU: BAND-WISE UNCERTAINTY FROM A RATE-NORMALIZED RATIO

The decoder is a band-partitioned SIREN (Sitzmann et al., 2020) whose first-layer frequencies split into K disjoint bands $B_1, \ldots, B_K$ with per-band maxima $f_k^{\max}$, each owned by a learnable query token that aggregates the raw irregular observations without interpolation. That construction, the conformal layer, the objective and the experiments establishing this is a reasonable forecaster rather than a strawman are in Appendix N; the ceiling argument needs only the variance head's inputs.

Let $\rho_n(\tau)$ be the *per-channel* kernel density of channel n's observation times — pooled densities are confounded on clinical data, where densely measured vitals are noisy and sparse labs smooth. The **identifiability ratio** $r_{n,k}(\tau) = \rho_n(\tau)/2f_k^{\max}$ is the local rate normalized by band k's Nyquist rate. One shared MLP conditioned on $[\log(1+\rho_n), \log r_{n,k}, e_k, V_{n,k}, \phi(\tau)]$ with a learned band embedding e_k emits per-band log-variances summing to $\sigma_n^2(\tau)$; sharing it across bands forces a single identifiability response, modulated per band — the object our theory constrains.

Sampling-theoretic motivation, not a theorem about our setting. Landau's necessary density condition (Landau, 1967) motivates giving the head a per-band, rate-normalized feature rather than a pooled density. It is motivation and *not* a bound on our predictive error: at least four of its hypotheses fail here — forecasting is extrapolation *off* the sampling set, a minimax floor over a function ball does not lower-bound a conditional variance, our kernel density *upper*-bounds the Beurling density for bursty clinical timestamps so $r_{n,k} \geq 1$ certifies nothing, and the band-limitedness penalty is off in every reported run (Appendix Q). An earlier draft called NABU "theory-grounded"; we no longer do, we attach no threshold to $r_{n,k}$, and what survives is only that the ratio is a per-band rate normalization suggested by sampling theory, which Section 4.1 tests empirically against the density alone.

4 EXPERIMENTS

Setup. Benchmarks: Human Activity (12 channels), PhysioNet 2012 (41 channels, 88% missing; train/val/test = challenge sets a/b/c), USHCN (5) and MIMIC-III (96 channels, 96.7% missing, $23{,}457$ sequences). Baselines: GRU-D, mTAND, Latent-ODE, T-PatchGNN and a faithful reimplementation of the query-embedding forecaster of Lim (2026). All models trained 10 epochs on identical splits; seed counts vary by board and are given per table. **Three departures from the published protocols** (Appendix S): Activity and USHCN split each sequence at its midpoint, so USHCN forecasts about 100 months ahead where the literature forecasts one; some baselines run below their authors' capacity, which understates t-PatchGNN on PhysioNet; and the query-embedding row is our reimplementation, not the authors' code. Comparisons are internally consistent, not line-by-line comparable with published tables. Code, per-seed CSVs and a per-claim manifest are released.

On the p-values. We report 15 and had applied no multiplicity correction. Holm–Bonferroni at FWER$=0.05$ leaves exactly one (USHCN point MSE, $p=0.001$) and Benjamini–Hochberg at FDR$=0.05$ the five smallest — but that survivor is a paired t on $n=5$, whose validity rests on normality of five seed differences; an exact sign test on the same five seeds cannot go below 0.0625. The family was also enumerated after the analysis (Appendix K), so no correction here controls the choices of loss, scale and split made while revising. Treat each p as describing one comparison; the headline tables carry effect sizes and intervals. The family is enumerated in `experiments/multiplicity_family.yaml` and both corrections are recomputed by `scripts/multiplicity.py` (Appendix K records that this number was wrong in print).

4.1 NABU: DOES VARIANCE TRACK IDENTIFIABILITY?

On PhysioNet, realized squared error grows with per-channel sparsity: the within-channel correlation with $1/\rho_n$ is $+0.162$ raw, $+0.128$ with horizon partialled out (5 seeds, test split). Sparsity is genuinely hard. Yet a pooled variance head gets the *sign* wrong — within-channel $\rho_s(\sigma^2, 1/\rho_n) = -0.12$, seed-level CI $[-0.20, -0.04]$ — averaged over *four* seeds, not five, because on one the pooled σ^2 collapsed to a constant within every channel; imputing that seed at 0 gives -0.09. Details and both bootstrap conventions in Appendix F.

4.1.1 WHAT SURVIVES THE HORIZON CONFOUND

ρ_n is a density at forecast times while all observations precede them, so $1/\rho_n$ is itself rank-correlated with forecast distance (ρ_s=+0.70) and part of any density–variance correlation is variance growing with horizon. Partialling it out changes the diagnosis: the pooled head moves from -0.12 to $+\mathbf{0.03}$ — not inverted but *uninformative* about within-channel density — while NABU remains at $+\mathbf{0.53}$ (from $+0.70$). The pooled head is also an unstable abstention baseline (its $\hat{\sigma}^2$ collapses to 63 distinct values on one seed and one value on another), so the selective-prediction comparison against it is not evidence that NABU ranks error well; what would settle that is Appendix J, where no arm resolves the density input's contribution.

The inversion needed two conditions, and this section found only one of them. Partialling horizon already moved the pooled head from -0.12 to $+0.03$. Per-variable z-scoring — with no partialling at all — moves it to $+0.58$ with every channel positive (Appendix I.0.3). The negative number this subsection set out to explain therefore required *both* raw per-channel scale and an unpartialled horizon, and the preprocessing half went unexamined because `PhysioNet2012` had no way to z-score. A reader should take the inversion as a property of how we prepared this benchmark, not of irregular sampling.

4.1.2 HOW MUCH CAN *any* SAMPLING-AWARE HEAD BUY? A MEASURABLE CEILING

Lemma 1 (Appendix O) is the tower property: with $Y = e^2$, S collecting the query's own channel and time together with the observed timestamps and channel identities — not the observed values — the residual $R := Y - \mathbb{E}[Y \mid S]$ is orthogonal to every $g \in L^2(\sigma(S))$. The query must be in S for our own stratification by channel and horizon decile to be $\sigma(S)$-measurable. Two corollaries follow.

Corollary 1 (Correlation ceiling)**.** *With $v(S) := \mathbb{E}[Y \mid S] - \mathbb{E}[Y]$ and $s := \mathrm{Var}[v]/\mathrm{Var}[Y]$, every $g \in L^2(\sigma(S))$ with $\mathrm{Var}[g] > 0$ satisfies $|\mathrm{corr}(g(S), e^2)| \le \sqrt{s}$, attained at $g = \mathbb{E}[e^2 \mid S]$ when $\mathrm{Var}[v] > 0$.*

Corollary 2 (Selective-risk floor)**.** *Among all $\sigma(S)$-measurable abstention rules retaining a fixed fraction of queries, the rule retaining the smallest values of $\mathbb{E}[e^2 \mid S]$ minimizes retained risk. So any exhibited S-measurable score gives an* upper *bound on the oracle's risk–coverage area, hence a* lower *bound on what the observation set can buy.*

Proofs in Appendix O; neither needs an independence assumption. $s = \eta^2(e^2 \mid S)$ is *associational* — clinical acuity drives both measurement frequency and volatility — and a property of the *(data, model, preprocessing)* triple, since e^2 is a trained model's error.

Three quantities, which earlier versions of this paper conflated. s itself is never estimated here. What we estimate is s_T, the correlation ratio of the coarsening $T = (\text{channel}, \text{horizon decile}, \text{density tercile})$ of S. Because T is a function of S,

$$s_T = \frac{\mathrm{Var}(\mathbb{E}[e^2 \mid T])}{\mathrm{Var}(e^2)} \;\le\; s,$$

so $\sqrt{s_T}$ is the *exact* ceiling for a head that reads only T and a *lower* bound on the ceiling for heads reading all of S. The sparsest-versus-densest tercile contrast $\hat{s}_{\mathrm{ext}}$ that earlier versions called the ceiling is a diagnostic and bounds nothing: with $\rho \in \{1, 2, 3\}$ and $e^2 = 1 + \mathbf{1}\{\rho = 2\}$ a density-only function tracks e^2 perfectly while $\hat{s}_{\mathrm{ext}} = 0$ (Appendix). Its noise correction also ignores within-sequence correlation.

What no ceiling here covers: a *marginal* contribution — an $S \times V$ interaction can be arbitrarily large at $s = 0$ (Appendix E.0.1), and NABU's $r_{n,k} = \rho_n / 2 f_k^{\max}$ is one — a training gain, or a statistic aggregated differently (Appendix P).

Estimating s_T. A split-half U-statistic removes the plug-in's group-mean noise if its halves are *whole sequences*; splitting targets, as an earlier version did, is biased by $+0.022$ at a true zero on planted clustered data (`tests/test_ceiling_estimator.py`). One sequence bootstrap is applied jointly to every seed, conditional on the fitted checkpoints. On PhysioNet (5 seeds, 3,993 test sequences) $\sqrt{s_T} = \mathbf{0.120}$, with s_T in $[0.0061, 0.0993]$ and a one-sided 95% upper limit of $\mathbf{0.288}$ (a target-level split gives 0.121: the heavy-tailed interval, not the correction, changes the reading). Density's increment over channel and horizon is $+0.0079$, $[+0.0009, +0.0642]$, 55% of

the total. Z-scored, $\sqrt{s_T} = 0.045$ (upper 0.062) and the increment is $+0.0001$, $[-0.0013, +0.0015]$ (Appendix P). **A head reading only channel, horizon and density tercile can track at most about** 0.29 **of PhysioNet's raw squared error and** 0.06 **of its z-scored error, in correlation.** All four benchmarks are in Table 3. **Across forecasters the ceiling was similar and the floor was not:** $\sqrt{s_T}$ is 0.11–0.14 on six PhysioNet forecasters' raw residuals, 0.04 on the same six retrained z-scored, and 0.02–0.03 on six z-scored MIMIC-III ones, while raw PhysioNet's floor runs from 0.39 to nothing (Table 4). That is an observed similarity in three settings, not model invariance.

4.1.3 Testing the bound with an object it actually covers

Corollary 1 constrains a function of S alone and neither of our heads is one, so the admissible test object is an equality case: our stratified cell means are a plug-in $\hat{\mathbb{E}}[e^2 \mid T]$ — not $\mathbb{E}[e^2 \mid S]$ — fit on half the sequences and scored on the other half. Within a (channel, horizon) cell it is a function of the tercile alone, so its matched ceiling is the within-cell correlation ratio over all three terciles. Debiasing both sides by the same sequence split-half product, the oracle sits below the ceiling: 0.0607 against 0.0626, paired gap -0.0020, $[-0.0027, -0.0014]$ (Table 23). **Consistent with the bound, and not evidence of slack**: in simulation a half-sample lookup falls this far short when the population lookup attains the bound. We changed the oracle's debiasing after its first version exceeded the ceiling by $+0.0020$; the target-level permutation null it used is biased upward on clustered data, by that much in simulation (Appendix L). Debiasing also sets the ceiling itself: the nested η^2 plug-in is inflated $2.8\times$, so it is 0.120, not the 0.203 we published.

4.1.4 Removing the density input

Deleting the density input from the variance head and retraining moves the horizon-controlled statistic from $+0.53$ to $+0.60$ and error-tracking by -0.075 ± 0.049 ($p = 0.25$); it does not replicate (Section 4.1.6) and the arms are not matched (Appendix J), so it is not the "net harmful" reading it once carried. Appendix G has the arm-by-arm account.

4.1.5 One percent of *what*? The denominator

"Sampling geometry accounts for on the order of one percent of error variance" has a denominator we never measured. Measured out-of-fold and grouped by sequence, the e^2 scale cannot carry the statement: the R^2 is negative (-0.0041 ± 0.0049) while the *same* predictions correlate with e^2 at $\mathrm{corr}^2 = 0.0104 \pm 0.0054$ ($5/5$ seeds, $p = 0.013$) — signal with the wrong *scale*, not absent signal: at kurtosis 4.6×10^4 the top 1% of points carry 94.7–100% of $\mathrm{Var}[e^2]$, and a fitted conditional mean loses to the grand mean while still ranking. The *size* of the total is not identified on that scale (estimates span 0.002–0.025 with $\hat{s}$ inside), so we report the share on tail-tamed scales, where local density contributes $+0.05$ to $+0.08$ above channel identity and forecast horizon on both clinical benchmarks and, once each dataset's own permutation null is subtracted, $+0.006$ on Activity and -0.002 on USHCN. **We will claim** that density carries genuine out-of-fold information about forecast error on the clinical pair; **we will not** claim the near-regular datasets carry none, that dispersion *orders* the effect (MIMIC's marginal is largest at $8.5\times$, PhysioNet's smaller at $22\times$), or that the transform is immaterial. Appendix C has Table 6, the four published readings of this number, and why a difference of two R^2 lower bounds is not itself a bound.

4.1.6 Does this generalize? Five contrasts, none distinguishable from zero

Repeating the no-density contrast on all four benchmarks and on a second architecture gives marginals of -0.075, $+0.038$, -0.065, -0.006 and $+0.051$: *not one is distinguishable from zero* (Welch $p \geq 0.23$), and USHCN's standard error exceeds its own ceiling. These are not tests of the bound — it does not cover a marginal contribution (above) — and they have no power to fail; nor does their $n{=}4$ ordering against the ceiling ($\rho_s = +0.80$, permutation $p = 0.33$, floor 0.083) support anything, having refused the same argument two subsections earlier when it ran against us (Appendix E.0.1).

4.1.7 The same finding on a decision quantity

Everything above is a correlation between $\hat{\sigma}^2$ and realized error: a ratio of second moments on a target with kurtosis 4.6×10^4, inheriting an aggregation, a transform, a debiasing and a denominator

choice, each of which we got inconsistently wrong at least once. A bound on a decision quantity has none of them. Risk–coverage convention of Appendix D; every score fit on half the sequences and scored on the other half, since an in-sample lookup can over-fit its cells and beat the population oracle, in which case exhibiting it bounds nothing (that correction moved PhysioNet's floor from 0.589 to 0.611). **The loss is a declared choice.** We divide each channel's squared error by that channel's mean on the fitting half, which asks which queries are hard *relative to their own channel* and removes differences in physical units; a user who must weigh a missed glucose reading against a missed heart rate needs their own weights, and the floor is then re-measured on those.

Table 1: Normalized AURC on per-channel standardized e^2 (1.0 = no ranking, lower is better). Intervals are for the *mean of the five fitted models*: 200 sequence-bootstrap draws, one resample shared by every seed and the seed mean taken within each draw, conditional on the fitted lookups. An earlier version pooled the five models' draws into one mixture and read it as the spread for a randomly chosen model; a mixture percentile carries no nominal guarantee for that target, so it is now secondary (`results/v2_bound/aurc_cluster_bootstrap.csv`; PhysioNet floor ≥ 0.286). The prize is 1 minus the upper end of the floor's interval. The NABU column ranks by $\hat{\sigma}^2/d_c$, the variance in the loss's own units; an earlier version ranked by the raw $\hat{\sigma}^2$ against the standardized loss, comparing channels in mismatched units (0.803, 0.767, 0.536, 0.489). `results/v2_bound_joint`.

	normalized AURC, per-channel standardized e^2				
dataset	+horizon (lookup)	+density tercile (the S-only floor)	NABU $\hat{\sigma}^2$ (reads S *and* V)	density increment (95% CI)	prize $\geq$
PhysioNet	0.817	$\mathbf{0.609}$ $[.594, .623]$	0.852	$+0.208$ $[+.194, +.223]$	$\mathbf{0.377}$
USHCN	0.677	$\mathbf{0.665}$ $[.552, .792]$	0.778	$+0.012$ $[-.122, +.160]$	$\mathbf{0.208}$
Activity[‡]	1.025	0.905 $[.840, .966]$	*0.457*	$+0.119$ $[+.037, +.202]$	—
MIMIC-III	0.902	$\mathbf{0.844}$ $[.730, 1.011]$	*0.442*	$+0.058$ $[-.069, +.144]$	—

A floor that errs in one direction — as far as the data supports (Table 1). The density-tercile score is *exhibited* and S-measurable and the oracle minimizes selective risk among all such scores, so it bounds $\mathrm{AURC}_{f^\star}$ above and the prize below. **Which interval.** Table 1's intervals are for the *mean* of the five fitted models, one resample shared by every seed, conditional on their lookups. An earlier version pooled the five models' draws into one mixture and read it as the spread for a randomly chosen model. A mixture percentile carries no nominal guarantee for that target, so we report it only as the wider band it is (PhysioNet ≥ 0.286, USHCN 0.180). Neither is the training algorithm's average, which five seeds cannot establish. [‡]**On Activity the unit is a chunk, not a recording**: three recordings supply its test split, so a population interval cannot be reliably estimated from it (Appendix B) (MIMIC's increment reads $\pm.012$ across seeds and $[-.074, +.150]$ across resamples of the data). At the one-sided 97.5% limit **an S-only rule lowers normalised AURC by at least** 0.377 **on PhysioNet**, 0.208 on USHCN, *not establishable* on Activity once its unit is the recording, and **on MIMIC no prize is establishable** — its interval crosses 1.0, i.e. "no better than not ranking". Two distinct quantities: *the floor* is what any S-only score buys (channel, horizon, any other within-channel structure), and a small ceiling does not bound it (Appendix O); *the density increment* is what adding density to the best channel-and-horizon score buys, and on the unit the data supplies it is distinguishable from zero on PhysioNet only. USHCN having the second-largest floor and no establishable increment is no contradiction: monthly climate data with strong seasonality should have a large horizon component and little density signal.

The increment is a comparison, not a decomposition. $\mathrm{AURC}_{\text{+horizon}} - \mathrm{AURC}_{\text{+tercile}} = +0.208 \pm 0.048$ is a difference of two upper bounds, and a difference of bounds is not a bound. What it is: *a score built from channel, horizon and density achieves* 0.208 ± 0.048 *more risk reduction than the best score from channel and horizon alone at the same resolution* ($5/5$ seeds). Both sides are exhibited; nothing is decomposed.

Where NABU sits — and what that cannot say. Ranked by $\hat{\sigma}^2/d_c$, the variance in the loss's own units, NABU is worse than the density lookup on PhysioNet by 0.243, $[0.210, 0.275]$, indistinguishable on USHCN (-0.113, $[-0.328, +0.081]$), and better on MIMIC ($+0.402$, $[+0.274, +0.609]$) and Activity ($+0.448$, chunk-level, Appendix B); paired differences, same draws. Recalibrated to the lookup's resolution, a lookup on channel, horizon and the tercile of $\hat{\sigma}^2/d_c$, it still loses on PhysioNet by 0.033, $[0.021, 0.045]$. Earlier versions ranked by the raw $\hat{\sigma}^2$ against the standardized loss, comparing channels in mismatched units. Aligning the units moved NABU's area by 0.011 to 0.078

and reversed none of these comparisons (`results/v2_bound_joint`). **This is not a density contrast.** The lookup reads S; NABU reads S *and* V, and such a difference is unbounded even at zero ceiling (Section 4.1.6, applied to us). It cuts both ways: MIMIC's and Activity's gains cannot be credited to density, and PhysioNet's null never licensed "we cannot detect our head collecting any of the prize" either, since V can add what density subtracts. The four-dataset spread did not create that defect; it exposed one that was there when the number was zero. The identified contrast deletes the input with both arms reading V (Appendix J). NABU's per-seed *share* of the increment is not a reportable quantity: its denominator nearly vanishes and the ratio runs from -1411% to $+1098\%$ across seeds.

The raw column is a defect in our own boards. On native scales a lookup on *channel identity alone* captures 97% of achievable abstention benefit and NABU loses to it six-fold: abstention there is mostly a contest over which score sorts channels by error magnitude, glucose in mg/dL against pH near 7.4. Table 10's shift-robustness ordering, the deployed-UQ comparison and the selective-prediction gain are all confounded this way; all three are re-derived under standardization and reported on both scales, with the confound measured per dataset, in Appendix M.

Per-channel standardization also withdraws three earlier falsifications — two of which favoured us — and the largest gain we ever reported, $+65\%$ on selective prediction, becomes $+12.0\%$ at $p = 0.074$ with the variance in the loss's units; the record is in Appendix K.1.

4.1.8 Four further checks (Appendix I)

The density diagnostic moves $17\times$ while error-tracking barely moves. The negative raw density–variance statistic is PhysioNet-specific, so the diagnosis is scoped to *PhysioNet* rather than offered as a law, and not to "clinical IMTS with strongly heterogeneous channel rates", because rate dispersion is now known to order neither the floor nor the density increment (Appendix H.0.2), so a criterion built on it has no basis. And permuting the per-band Nyquist constants retains 84% of the effect — a pre-registered arm we had never run until this revision — so the mechanism is a per-band offset and "Nyquist-aware" is not earned.

4.2 Is conformal alone the contribution? UQ baselines (E2b)

Split conformal is model-agnostic, so calibration alone is not a contribution: we wrap four standard UQ methods — deep ensembles (Lakshminarayanan et al., 2017), MC-dropout (Gal & Ghahramani, 2016), quantile regression (Koenker & Bassett, 1978; Romano et al., 2019) and a heteroscedastic Gaussian head (Nix & Weigend, 1994) — in the identical conformal layer on a shared GRU-D backbone. All five reach 0.90 coverage, so the discriminator is *sharpness*. Three things from the full board (Appendix D). **(i)** At matched coverage BandRate's intervals are narrowest on PhysioNet (14% under the best baseline) and on *Activity*, where Table 6 finds no density signal ($+0.004$), and not on MIMIC, where it finds the most ($+0.075$): sharpening and error-ranking come apart, and we have no account of why. **(ii)** A deep ensemble, reading no density at all, does *not* rank error better than NABU on either clinical dataset once errors are per-channel standardized and the variance put in the same units; the native-scale reading that said otherwise, and that we had offered as confirmation of the ceiling thesis, is withdrawn in Appendix K.1, and we claim no reverse. **(iii)** On heavy-tailed PhysioNet BandRate's narrow intervals overshoot the occasional extreme, so wider Gaussian heads win the interval score there. No UQ method dominates across the four.

5 Limitations and honest negatives

Pre-registered gates. Thresholds were frozen in `experiments/claim_gates.yaml` before the external-validity runs and three fired against us. *Generality* (pooled correlation ≤ -0.10 on ≥ 2 datasets) is met by *no* dataset under the horizon control, so we executed the pre-declared fallback and removed every general-IMTS claim. *Decision utility* ($\geq 20\%$ AURC gain on ≥ 2 datasets) now fails at PhysioNet 12.0% (loss-aligned) / MIMIC 1.3% (raw-σ ranking, not rerun) once the AURC is per-channel standardized; on native scales PhysioNet read 65%, so the gate failed then on the ≥ 2 requirement and fails now on both. Its fallback was to *delete* the utility claim; we scoped it instead, an override of our own pre-registration that we flag rather than let the frozen file imply compliance. *The ratio gate* asked for a shuffled-$f^{\max}$ arm, now run, which fires against us (84% survives permutation).

A methods note. The correlation screen and the risk–coverage increment are *different estimands* that both ask what density adds given channel and horizon: the first is a ratio of second moments of the squared error, the second a functional of the ranking and the retention level. That they resolve on different datasets is therefore not a contradiction. They also behave differently as estimators: the correlation form needs a debiasing term that depends on the group counts and is $2.5\times$ the estimate, while the decision form scores an exhibited, cross-fitted rule and needs none. *Where the target is heavy-tailed, prefer the estimand whose errors have a known sign* — a preference we can motivate, not a result.

Limitations of the measurement. $\hat{s}$ bounds heads reading a local *rate*, not gap-since-last-observation or burst structure; it attenuates by a dataset-dependent factor; Corollary 1 is transform-specific; and s is associational — clinical acuity drives both measurement frequency and volatility — so it caps what an S-only head can track, not what sampling causes. **And the ceiling depends on the fit, not the data alone**: re-estimated on the same seeds at 30 epochs, $\sqrt{s_T}$ on the main board's seeds 1–3 moves from 0.122 to 0.097, but the two intervals overlap almost entirely (Appendix B), so we cannot say which way it moves with the fit. **Its upper limits are not guarantees.** Against a known truth the interval covers in heavy tails only because two biases cancel, the estimate runs 9–15% low in light tails because tercile edges are cut from the data, and below 50 targets per cell density is not read at all (Appendix). Four datasets is too few to order anything (Appendix H.0.2). **The matched deletion pair on PhysioNet we had listed as missing has now been run, and returns a null whose seed spread is $43\times$ its mean**, so the dataset the paper is built around still does not settle our own head's contribution. Appendix B has the full list.

6 Conclusion

This paper began as a method for sampling-aware uncertainty in IMTS and ends as a measurement of what that idea is worth. The inequality is elementary — the tower property caps a variance head reading S alone at $\sqrt{\eta^2(e^2 \mid S)}$ — and all the work is in measuring it honestly. It took five rounds and twenty-six ledgered claims, twenty-three retracted in print, all checked by script (Appendix K).

What survives is small and measured. $\sqrt{s_T}$ is between 0.03 and 0.12 on all four benchmarks, with one-sided upper limits up to 0.29, and density's share of it is distinguishable from zero on raw PhysioNet only; the extreme-tercile screen we used to call the ceiling is a diagnostic and resolves nothing. The floor, being one-sided and cross-fitted, is robust to the estimation choices that defeated the screen, and lowers normalised AURC by at least 0.377 on PhysioNet though not establishably on MIMIC. Where the prize is we do not know. Setting rate dispersion experimentally on one dataset (five levels, ten seeds) left the retrain contrast dominated by variation between fitted models, with per-seed intervals excluding zero in opposite directions at three levels. A lower-variance eval-only estimand resolved an effect at every level, but its pre-registered trend ran $+0.90$ on one held-out dataset and -0.70 on the other (Appendix I). Our own head's one positive contrast, $+0.029$ on Activity, rests on three test recordings, too few for a reliable population interval; the other arms lacked the power to detect an effect of the size observed, and on the z-scored scale no arm resolves one in either direction — including PhysioNet's, run for this revision, whose seeds disagree in sign. That is an unresolved contrast, not evidence that density does not help.

Rank in the loss's units. The lesson that transfers furthest is about evaluation, not about density. If channel c is rescaled, $y_c \mapsto a_c y_c + b_c$, then a loss standardized by the channel's error scale d_c is unchanged, and so is the ranking by $\hat{\sigma}^2/d_c$; the ranking by the raw $\hat{\sigma}^2$ is not, because it compares channels in different units. Every standardized risk–coverage board in earlier versions of this paper ranked by the raw variance. Aligning the units turned our head's 34% lead over a deep ensemble into an indistinguishable 8%, removed an apparent gain under thinning, and reversed no comparison with the S-only lookup (Table 13). Any evaluation that normalizes a loss per channel and then ranks by a model's variance should check this.

What to carry away, and what would overturn it. The second transportable part is the measurement (Appendix N.9): put channels on a common scale, partial out horizon, estimate s_T with the sequence as the unit, and measure the cross-fitted floor on the loss you will use. The inequality is not falsifiable; our estimates are. A T-reading head whose $\mathrm{corr}(\hat{\sigma}^2, e^2)$ clears our upper limit would indict the estimator, and one reading more of S than T can legitimately exceed it, since s_T only bounds s from below.

ETHICS STATEMENT

Data. Two of the four benchmarks are clinical. PhysioNet Challenge 2012 and MIMIC-III are distributed by PhysioNet under credentialed access: MIMIC-III requires CITI human-subjects training and a signed data use agreement, and both prohibit re-identification and redistribution. We accessed them under those terms and redistribute no record-level data — this paper's released artefacts are per-claim aggregates, per-seed summary CSVs and model checkpoints, and the download scripts fetch from PhysioNet with the reader's own credentials rather than bundling data. Human Activity and USHCN are public and non-clinical.

Use. This paper measures how much predictive uncertainty the observation set can explain; it does not propose a clinical decision rule. The selective-prediction experiments abstain on a retained-risk metric and are not a triage proposal. A head whose variance is well calibrated on aggregate can still be wrong on an individual patient, and nothing here supports deployment in care.

Risks. The main foreseeable harm is the opposite of over-claiming: a ceiling reported without its interval could be read as licence to stop improving sampling-aware uncertainty. We therefore report the ceiling with its one-sided upper limit rather than as a point estimate, and state that it is a lower bound on what heads reading more of the observation set could track.

REPRODUCIBILITY STATEMENT

Every headline number is tied to a board, a git SHA and checkpoint hashes in `REPRODUCIBILITY.md`, generated by `scripts/make_manifest.py` and checked by `scripts/validate_results.py` (55 boards). Twenty-three of the paper's 38 tabular displays are *recomputed* from the boards behind them by `scripts/check_paper_tables.py` — 880 cells, the printed values parsed out of the source so the specification cannot drift into agreeing with a stale table. One more is a descriptive summary of current claims (Appendix A); the other 14 carry captions but no label, so they cannot be declared to that tool and their numbers are checked by hand only; the tool now says so on every run. It also reports how many cells each table compared, which is how we found two tables silently skipping a row. Appendix K documents the gates, including the ones that failed and what they missed. Seeds are fixed per run and the seed count is stated per table. Three caveats a reader should have before attempting replication. *(i)* Attention and FFT kernels on GPU are non-deterministic, so a rerun reproduces our numbers to seed-level precision, not bitwise. *(ii)* `pyproject.toml` declares version floors, which every NumPy release that removed a function this code called would still have satisfied; `requirements-lock.txt` pins the environment the boards were produced in and should be installed first. *(iii)* The clinical datasets require credentialed access, so those boards cannot be regenerated from this repository alone.

AI USE STATEMENT

Required by the ICLR 2027 author guidelines and AI policy; the same disclosure is entered on the submission form. An LLM-based coding assistant (Claude, Anthropic) was used substantially throughout this work. Against the policy's list of uses requiring disclosure, it: **helped develop the theoretical framework, formulated mathematical claims, and assisted in writing the proofs** — the tower-property lemma and the three corollaries of Appendix O, and the partial ceiling of Appendix P, were developed with its assistance, as was the appendix stating why the sampling-theoretic sketch is not claimed as a theorem; **implemented methods** — every analysis script in `scripts/`, the data loaders, and the released tool; **proposed and refined hypotheses** — among them that per-channel scale heterogeneity might be the trigger the paper looks for, and that the eval-only contrast grows with rate dispersion, both recorded as hypotheses in `experiments/*_prereg.md` and both subsequently unsupported; **designed experiments and gave feedback on methodology** — the dispersion sweep, the change of estimand, the minimum-detectable-effect analysis, and the pre-registrations that fix a reading before a board is scored; **generated synthetic data** — the Bernoulli masks that make the regular benchmarks irregular, and the planted-truth fixtures; **cleaned and reformatted datasets** — the loaders in `src/timeprism/data/`; **created the scientific figures** — every figure here is output by a script in `scripts/` written with its assistance, and the label placement in Figure 5 is its work; **retrieved and discovered related work** — the search behind

Section 2 and the list of methods we were required to cite, the scouting of which recent baselines were runnable on our datasets, and the verification of every entry in this bibliography against publisher records, which found three wrong entries including two fabricated co-authors; and **interpreted results** throughout. It also drafted and revised the manuscript text. Two items on the policy's list it was *not* used for, stated rather than omitted so the absence is visible: translation, and qualitative or thematic data analysis.

It produced several of the errors this paper retracts, including the in-sample lookup that invalidated an earlier statement of the floor, the divergent risk–coverage implementation of Appendix K, and — in the most recent revision — three readings taken from raw-scale boards that further measurement showed to be properties of the preprocessing rather than of the methods.

How the output was checked. Every experimental number is produced by code in this repository and tied to a board, a git SHA and checkpoint hashes by `scripts/make_manifest.py`, with `scripts/validate_results.py` checking 55 boards and `scripts/check_retractions.py` checking that no retracted claim is asserted outside its retraction. The test suite pins the properties that outputs alone do not reveal, including several that caught defects in LLM-written analysis code: an influence function missing a term, a power curve computed against the wrong null, and an interval that did not match its estimand. The authors verified the claims retained in this paper and take full responsibility for its content, including its errors.

REFERENCES

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference*, 10(2):455–482, 2021.

Marin Biloš, Johanna Sommer, Syama Sundar Rangapuram, Tim Januschowski, and Stephan Günnemann. Neural flows: Efficient alternative to neural odes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. arXiv:2110.13040.

Henrik Boström and Ulf Johansson. Mondrian conformal regressors. In *Conformal and Probabilistic Prediction and Applications (COPA)*, volume 128 of *Proceedings of Machine Learning Research*, pp. 114–133. PMLR, 2020. URL `https://proceedings.mlr.press/v128/bostrom20a.html`.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8:6085, 2018. doi: 10.1038/s41598-018-24271-9.

Ian Covert, Scott Lundberg, and Su-In Lee. Understanding global feature contributions with additive importance measures. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Xilin Dai, Zhijian Xu, Wanxu Cai, and Qiang Xu. From samples to scenarios: A new paradigm for probabilistic forecasting. In *International Conference on Learning Representations (ICLR)*, 2026. URL `https://openreview.net/forum?id=SsJ6bZUmfU`.

Nicki S. Detlefsen, Martin Jørgensen, and Søren Hauberg. Reliable training and estimation of variance networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

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

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5988–6008. PMLR, 2022. URL `https://proceedings.mlr.press/v162/ethayarajh22a.html`.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning (ICML)*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059. PMLR, 2016. URL `https://proceedings.mlr.press/v48/gal16.html`.

Yonatan Geifman, Guy Uziel, and Ran El-Yaniv. Bias-reduced uncertainty estimation for deep neural classifiers. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://arxiv.org/abs/1805.08206.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Pascal Iversen, Simon Witzke, Katharina Baum, and Bernhard Y. Renard. Identifying drivers of predictive aleatoric uncertainty. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 5453–5462, 2025. doi: 10.24963/ijcai.2025/607.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning (ICML)*, 2020. arXiv:2006.16236.

Seyed Mehran Kazemi, Rishab Goel, Sepehr Eghbali, Janahan Ramanan, Jaspreet Sahota, Sanjay Thakur, Stella Wu, Cathal Smyth, Pascal Poupart, and Marcus Brubaker. Time2Vec: Learning a vector representation of time. *arXiv preprint*, 2019. arXiv:1907.05321v1.

Christian Klötergens, Vijaya Krishna Yalavarthi, and Lars Schmidt-Thieme. Probabilistic circuits for irregular multivariate time series forecasting. *arXiv preprint arXiv:2604.27814*, 2026.

Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

H. J. Landau. Necessary density conditions for sampling and interpolation of certain entire functions. *Acta Mathematica*, 117:37–52, 1967. doi: 10.1007/BF02395039.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018. doi: 10.1080/01621459.2017.1307116.

Boyuan Li, Yicheng Luo, Zhen Liu, Junhao Zheng, Jianming Lv, and Qianli Ma. Hyperimts: Hypergraph neural network for irregular multivariate time series forecasting. In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2505.17431.

Junghoon Lim. QuITE: Query-based irregular time series embedding. In *International Conference on Machine Learning (ICML)*, 2026. URL https://openreview.net/forum?id=ILQGHFvEoo.

Xvyuan Liu, Xiangfei Qiu, Xingjian Wu, Zhengyu Li, Chenjuan Guo, Jilin Hu, and Bin Yang. Rethinking irregular time series forecasting: A simple yet effective baseline. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 40, pp. 23873–23881, 2026. doi: 10.1609/aaai.v40i28.39563.

Yicheng Luo, Bowen Zhang, Zhen Liu, and Qianli Ma. Hi-Patch: Hierarchical patch GNN for irregular multivariate time series. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 41494–41519. PMLR, 2025. URL https://proceedings.mlr.press/v267/luo25r.html.

David A. Nix and Andreas S. Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, volume 1, pp. 55–60, 1994. doi: 10.1109/ICNN.1994.374138.

Harris Papadopoulos, Vladimir Vovk, and Alexander Gammerman. Regression conformal prediction with nearest neighbours. *Journal of Artificial Intelligence Research*, 40:815–840, 2011.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006. doi: 10.7551/mitpress/3206.001.0001.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Yulia Rubanova, Ricky T. Q. Chen, and David K. Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. URL https://papers.nips.cc/paper_files/paper/2019/hash/42a6845a557bef704ad8ac9cb4461d43-Abstract.html.

Andrea Saltelli, Paola Annoni, Ivano Azzini, Francesca Campolongo, Marco Ratto, and Stefano Tarantola. Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Computer Physics Communications*, 181(2):259–270, 2010.

Mona Schirmer, Mazin Eltayeb, Stefan Lessmann, and Maja Rudolph. Modeling irregular time series with continuous recurrent units. In *International Conference on Machine Learning (ICML)*, 2022. arXiv:2111.11344.

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

Satya Narayan Shukla and Benjamin M. Marlin. Multi-time attention networks for irregularly sampled time series. In *International Conference on Learning Representations (ICLR)*, 2021.

Satya Narayan Shukla and Benjamin M. Marlin. Heteroscedastic temporal variational autoencoder for irregularly sampled time series. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=Az7opqbQE-3.

Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020. URL https://papers.nips.cc/paper_files/paper/2020/hash/53c04118df112c13a8c34b38343b9c10-Abstract.html.

Ilya M. Sobol'. Sensitivity estimates for nonlinear mathematical models. *Mathematical Modelling and Computational Experiments*, 1(4):407–414, 1993. English translation of Matem. Modelirovanie 2(1):112–118, 1990.

Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

Kamile Stankeviciute, Ahmed M. Alaa, and Mihaela van der Schaar. Conformal time-series forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Andrew Stirn, Hans-Hermann Wessels, Megan Schertzer, Laura Pereira, Neville E. Sanjana, and David A. Knowles. Faithful heteroscedastic regression with neural networks. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 206 of *Proceedings of Machine Learning Research*, pp. 5593–5613. PMLR, 2023. URL https://proceedings.mlr.press/v206/stirn23a.html.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005. doi: 10.1007/b106715.

David S. Watson, Joshua O'Hara, Niek Tax, Richard Mudd, and Ido Guy. Explaining predictive uncertainty with information theoretic Shapley values. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021. URL https://papers.nips.cc/paper_files/paper/2021/hash/bcc0d400288793e8bdcd7c19a8ac0c2b-Abstract.html.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *International Conference on Learning Representations (ICLR)*, 2020.

Vijaya Krishna Yalavarthi, Kiran Madhusudhanan, Randolf Scholz, Nourhan Ahmed, Johannes Burchert, Shayan Jawed, Stefan Born, and Lars Schmidt-Thieme. Grafiti: Graphs for forecasting irregularly sampled time series. In *AAAI Conference on Artificial Intelligence*, 2024. arXiv:2305.12932.

Zinuo You, Jin Zheng, and John Cartlidge. Latent Laplace diffusion for irregular multivariate time series. In *International Conference on Machine Learning (ICML), Spotlight*, 2026. arXiv:2605.19805v2, Table 3.

Weijia Zhang, Chenlong Yin, Hao Liu, Xiaofang Zhou, and Hui Xiong. Irregular multivariate time series forecasting: A transformable patching graph neural networks approach. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 60179–60196. PMLR, 2024.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pp. 11106–11115, 2021. doi: 10.1609/aaai.v35i12.17325.

A WHAT THIS PAPER CLAIMS, AND ON WHAT

Every claim the paper currently makes, with the loss and scale it is measured on, the board behind it, its independent unit and its status. Withdrawn readings are in Appendix K and nowhere else.

Table 2: Current claims. "Seeds" is the number of trained models; "unit" is what the interval resamples. Every interval is conditional on the fitted checkpoints unless stated.

claim	loss / scale	board	seeds	unit	status
PhysioNet's pooled head is "inverted" only on raw scale ($-0.12 \to +0.58$)	within-channel ρ_s; raw vs z-scored	`v2_w1_gate_norm`	3+3	chan.	measured; one dataset
PhysioNet is the only benchmark with a negative raw statistic	within-channel ρ_s, raw	`v2_w2_regime_pooled`	5	chan.	measured
S-only floor: normalised AURC lowered by $\geq$ 0.377 (PhysioNet), 0.208 (USHCN)	e^2 per-channel standardised, cross-fitted	`v2_bound_joint`	5	seq.	mean of five fitted models, conditional on them; not a risk reduction at a retention level
AURC density increment $+0.208$ on PhysioNet only	same	`v2_bound`	5	seq.	measured
s_T ceilings for heads reading only T	e^2, raw (and z-scored, PhysioNet)	`v2_ceiling_seqsplit`	3–5	seq.	*lower* bound on the full-S ceiling
Density-deletion contrast resolves on no arm	normalised AURC	`v2_bound_nodensity*`	3	seq.	8 of 9 arms underpowered; Activity meets the criterion only for its 3 recordings
s_T depends on the fit	e^2, raw vs z-scored; 10 vs 30 epochs	`v2_ceiling_seqsplit`	3	seq.	z-scored far lower; epochs unresolved; no monotonicity result
s_T similar across six forecasters; the floor is not	e^2; raw and z-scored PhysioNet, z-scored MIMIC-III	`v2_ceiling_across_models`	5	seq.	observed similarity in three settings, not invariance
A variance ranked against a standardized loss must be $\hat{\sigma}^2/d_c$; aligned, our head and a deep ensemble are indistinguishable	standardized AURC	`v2_cr_uq_*`	3	seed	measured; the raw-σ lead of 34% is withdrawn
Point accuracy: not significantly different on 2 datasets, a few percent behind on 2	test MSE	`repro`	5	seed	protocol departures in Appendix S

B LIMITATIONS OF THE INSTANTIATION

The measurement, in full. (a) $\hat{s}$ comes from a density-*tercile* contrast, so it bounds heads reading a local rate; gap-since-last-observation, burst structure and inter-arrival dispersion are in S but not in

this $\hat{s}$. (b) The estimator attenuates by a dataset-dependent factor, so the calibration constant is not transportable, and the raw value is not a lower bound on the ceiling — which ceiling applies depends on the statistic's aggregation (Appendix P). (c) Corollary 1 is transform-specific. (d) s is associational: clinical acuity drives both measurement frequency and volatility, so it caps what an S-only head can track, not what sampling causes. (e) It is a property of the (data, model, preprocessing) triple. (f) Four datasets is too few to order anything, which is why Appendix H.0.2 retires the ordering question outright. (g) The matched deletion pair on PhysioNet, listed here for three rounds as the one experiment that would settle our own head's contribution, has since been run (Appendix J); its seed spread is $43\times$ its mean, so it does not settle it either. (h) **Activity's intervals understate their own uncertainty — which removes both of our positive results.** The cluster bootstrap's unit is the dataset sample, and Activity is the one dataset whose samples are 30-row chunks of a longer recording: its test split is 607 chunks from three recordings (A02, C03, C04, and those from only *two* subjects), so roughly two hundred dependent units are resampled per independent one. The split is clean — whole recordings go to one split — so this is a width rather than a leak. But the consequence has to be stated, not just the cause. With the recording as the unit Activity has *three* independent units, too few to estimate a reliable population interval, so: the density increment $+0.119$ $[+.032, +.213]$ and the prize ≥ 0.029 in Table 1 are not establishable; neither is the deletion arm's $3/3$ in Appendix J, which is the one row Appendix I calls powered. Those were the two positive results this paper once reported *for our own head's density input*, and on the unit the data actually supplies neither is established. The results that do stand rest on other evidence: the PhysioNet floor, the scale artefact and the s_T estimates (Appendix A). We mark them rather than delete them because the chunk-level intervals are what we computed and a reader should see both; the recording-level reading is the one to believe. (j) **Every ceiling in this paper is a 10-epoch quantity.** s is defined through *the model's* error, so a better-fit forecaster leaves less S-explainable error and the ceiling moves with the training budget — it is not a fixed prize that better methods approach. Our budget check (Appendix S) cuts BandRate's PhysioNet MSE by 20% at 30 epochs. Re-estimating s_T on those checkpoints — same three seeds, same script, sequence-level split (`results/v2_ceiling_seqsplit`) — moves $\sqrt{s_T}$ on the main board's seeds 1–3 (the paired board's raw arm, separately trained, reads 0.123) from 0.122 to 0.097, but s_T's intervals are $[0.0063, 0.0925]$ and $[0.0039, 0.0926]$: the point estimate fell and the data do not establish that it did. Every ceiling in this paper is therefore a property of the 10-epoch fit, and we do not know which way a better fit moves it; no general result fixes that direction either (Appendix P). Concretely we retract "the density increment is distinguishable from zero on PhysioNet and Activity", carried by four drafts, and the 0.029 prize with it.

(i) On raw-scale MIMIC-III one of three seeds diverged at evaluation (the depth-3 SC-INR extrapolating to extreme values at a few query times, inflating MSE $3\times$). We traced this to the dataset's heavy-tailed raw targets (std 180, max 12200): per-variable z-normalization — the standard preprocessing of the baselines we compare to — maps targets to $\mathcal{N}(0, 1)$ and removes the divergence at its source (test-split 3-seed MSE 0.630 ± 0.003, no divergent seed) while *preserving* the NABU density–variance correlation ($+0.59$). A soft output clamp also stabilizes accuracy but destroys that correlation ($\to 0$), since clamping the mean blinds the variance head; we therefore adopt normalization, not clamping. Table 28 reports MIMIC on this normalized scale, where all six models are stable. (ii) A low-rank temporal FiLM decoder variant destabilized PhysioNet training ($+88\%$ MSE) despite an exact identity initialization; we report this negative result and exclude it. (iii) Depth-3 SC-INR decomposition remains seed-sensitive due to SIREN optimization instability — a failure mode upstream of, and not fixed by, $\mathcal{L}_{\text{band}}$. (iv) Conformal guarantees are marginal (per Mondrian bin) under exchangeability; temporal distribution shift can erode them, and our rate-shift experiments quantify but do not certify that regime. (v) Raw (pre-conformal) PIT remains miscalibrated; we rely on CONFIDE for interval validity rather than claiming a calibrated likelihood. (vi) Two comparison methods are reimplemented from their papers. Both have since been published — APN (Liu et al., 2026) at AAAI 2026 and the query-embedding method (Lim, 2026) at ICML 2026 — and the latter now has a public implementation, released after ours was written. **We did not re-run against it**, so our rows are faithful-in-design reimplementations rather than reproductions, and a reader comparing them with the authors' own code should expect them to differ. We run them under our protocol, and report MIMIC on the per-variable normalized scale they use, but do not replicate their per-dataset display scalings or the exact $12{\to}36/24{\to}24/36{\to}12$ horizon splits, so our numbers are internally consistent rather than line-by-line comparable to their tables. We are explicit that for the query-embedding method (Lim, 2026) *numerical parity with its paper is not establishable in principle*, not merely unverified: that work presents a *backbone-agnostic, plug-and-play embedding module*, and

where it does report standalone errors — eleven IMTS forecasters, all four of our baselines among them — they sit three orders of magnitude from ours — its PhysioNet MSEs run 3.8–10.7 against our 11,512–16,813 — because it adopts Hi-Patch's preprocessing with per-variable normalization while we report native clinical units, so the two boards cannot be placed side by side whatever the method; and it uses a $60/20/20$ random split rather than the PhysioNet challenge set-a/b/c partition. Our "QuITE" row is therefore a standalone forecaster built from that paper's tokenization and variable-query aggregation — a faithful-in-design but not line-by-line-comparable baseline — and should be read as such rather than as a reproduction of its published numbers. (vii) Activity, PhysioNet and USHCN are reported on their native scales and MIMIC on the per-variable z-scored scale, and this asymmetry is load-bearing rather than cosmetic: Appendix H.0.2 shows that normalizing PhysioNet removes the sign inversion the paper is built around, and Appendix P that it cuts $\sqrt{s_T}$ from 0.123 to 0.045. PhysioNet now exists on both scales (`results/v2_w1_gate_norm`); `PhysioNet2012` and `HumanActivity` gained the `normalize` parameter they lacked, which is why neither comparison could be run before. All four now exist on both scales for the deletion contrast (`results/v2_bound_nodensity_otherscale` adds USHCN z-scored, MIMIC raw and, for this revision, PhysioNet z-scored; Appendix J) — which four drafts of this item asserted while the PhysioNet z-scored arm had not yet been run, so the claim was false as printed then; the board's ninth row is ETTm1, and the raw MIMIC arm that item (i) abandoned for divergence evaluated cleanly on all three seeds this time, so that divergence was seed-dependent rather than intrinsic — though on raw units neither MIMIC arm ranks error at all. Given what the PhysioNet contrast returned, every native-scale number in this paper should be read as specific to that scale rather than as a property of the benchmark.

C THE PHYSIONET SHARE DECOMPOSITION

The nested decomposition behind Section 4.1.5; Table 6 gives the density marginal on all four benchmarks.

The extreme-tercile diagnostic, and a withdrawn ceiling. Within each (channel, horizon-decile) stratum $\hat{s}_{\text{ext}}$ contrasts the sparsest and densest density terciles and subtracts target-level sampling noise: PhysioNet 0.0084 ($\sqrt{\cdot} = 0.092$, 343 strata, 5 seeds), Activity 0.0059 (0.077), USHCN 0.0046 (0.068), MIMIC-III 0.0008 (0.029), with sequence-clustered intervals $[-0.0248, +0.0416]$, $[-0.0020, +0.0139]$, $[-0.0034, +0.0126]$ and $[-0.0097, +0.0113]$ — none distinguishable from zero. **We retract** the reading earlier versions gave the upper limits of these intervals as ceilings, "the density pathway can contribute at most 0.204 in correlation on PhysioNet". $\hat{s}_{\text{ext}}$ ignores the middle tercile and so can be zero when a density-only function tracks the error perfectly (Appendix); its upper limit is not an upper limit on anything. The ceilings we now report are the s_T of Section 4.1.2, which include every tercile.

Table 3: s_T, the correlation ratio of $T = (\text{channel}, \text{horizon decile}, \text{density tercile})$, on all four benchmarks: the exact ceiling for a head reading only T and a lower bound on the full-S ceiling. Debiased by a split-half U-statistic over whole sequences; intervals from one sequence bootstrap (200 draws) applied jointly to every seed, conditional on the fitted checkpoints (`results/v2_ceiling_seqsplit`). "Upper" is the one-sided 95% limit on $\sqrt{s_T}$. Activity's unit is a 30-row chunk, not one of its three recordings, so its intervals are too narrow. MIMIC's total is too close to zero for a density share to mean anything.

			s_T and its ceiling			density increment		
dataset	seeds	$\sqrt{s_T}$	s_T lo	s_T hi	upper	Δ	lo	hi
PhysioNet, raw	5	0.120	0.0061	0.0993	0.288	+0.0079	+0.0009	+0.0642
PhysioNet, z-scored	3	0.045	0.0009	0.0045	0.062	+0.0001	−0.0013	+0.0015
USHCN, raw	5	0.111	0.0036	0.0218	0.140	+0.0035	−0.0044	+0.0108
Activity, raw (chunks)	5	0.102	0.0037	0.0194	0.133	+0.0038	−0.0011	+0.0104
MIMIC-III, z-scored	5	0.034	0.0002	0.0192	0.132	−0.0008	−0.0038	+0.0089

The same diagnostic on six forecasters' residuals, in three settings. Both quantities use only residuals and the observation set, so they can be computed on any model's errors. We did so for all six models on raw PhysioNet, on PhysioNet retrained with the per-variable z-scoring the baselines use (all 30 runs trained for this revision, `results/repro_norm`) and on z-scored MIMIC-III,

building T from a fixed-bandwidth density identical across models, so the rows of a block differ only in whose residuals they score (`scripts/ceiling_across_models.py`). This protocol differs from Table 1 in three declared ways. The density is fixed rather than the model's learned one; tercile edges come from the fitting half only; and the floor averages both cross-fit directions rather than scoring one, which roughly halves its variance. Both report the mean of the five fitted models. On the same checkpoints the point floors agree (PhysioNet 0.385 against 0.391, MIMIC-III 0.187 against 0.156). The intervals do not: MIMIC-III's floor is establishable under this protocol and not under that of Table 1.

Table 4: s_T and the cross-fitted floor on six forecasters' residuals (5 seeds each; z- marks the z-scored retrain; T from a fixed-bandwidth density shared by all rows). Intervals are 95% from a joint sequence bootstrap over each model's seeds (100 draws), conditional on its checkpoints. The floor's lower limit is the 2.5th percentile, the rule behind the floors of Table 1; an earlier version printed the minimum over seeds of per-seed limits, a different rule. Split-half point estimates average six random sequence splits and move by up to 0.011 in $\sqrt{s_T}$ between random streams; the intervals include that variation. `results/v2_ceiling_across_models`.

	s_T			density increment			floor	
dataset, model	$\sqrt{s_T}$	lo	hi	Δ	lo	hi	point	lower
PhysioNet, BandRate	0.116	0.0065	0.0867	$+0.0070$	-0.0007	$+0.0396$	0.385	0.375
PhysioNet, T-PatchGNN	0.109	0.0055	0.0856	$+0.0060$	$+0.0002$	$+0.0507$	0.229	0.221
PhysioNet, QuITE	0.125	0.0066	0.0849	$+0.0093$	$+0.0017$	$+0.0460$	0.255	0.239
PhysioNet, GRU-D	0.130	0.0090	0.1043	$+0.0074$	$+0.0017$	$+0.0591$	0.017	-0.038
PhysioNet, mTAND	0.126	0.0053	0.0877	$+0.0094$	-0.0008	$+0.0496$	0.060	0.038
PhysioNet, Latent-ODE	0.142	0.0091	0.1211	$+0.0108$	$+0.0039$	$+0.0709$	-0.195	-0.236
z-PhysioNet, BandRate	0.041	0.0010	0.0059	-0.0002	-0.0012	$+0.0025$	0.159	0.123
z-PhysioNet, T-PatchGNN	0.038	0.0007	0.0045	$+0.0002$	-0.0014	$+0.0014$	0.057	-0.003
z-PhysioNet, QuITE	0.043	0.0008	0.0050	$+0.0005$	-0.0012	$+0.0023$	0.125	0.082
z-PhysioNet, GRU-D	0.043	0.0007	0.0049	$+0.0002$	-0.0012	$+0.0023$	0.136	0.080
z-PhysioNet, mTAND	0.041	0.0007	0.0084	$+0.0004$	-0.0009	$+0.0053$	0.209	0.066
z-PhysioNet, Latent-ODE	0.042	0.0006	0.0078	$+0.0002$	-0.0011	$+0.0049$	0.246	0.070
MIMIC-III, BandRate	0.033	0.0002	0.0200	-0.0008	-0.0034	$+0.0089$	0.187	0.109
MIMIC-III, T-PatchGNN	0.033	0.0003	0.0232	-0.0005	-0.0051	$+0.0107$	0.216	0.165
MIMIC-III, QuITE	0.032	0.0001	0.0245	-0.0010	-0.0056	$+0.0090$	0.211	0.121
MIMIC-III, GRU-D	0.024	0.0001	0.0127	-0.0008	-0.0038	$+0.0050$	0.174	0.099
MIMIC-III, mTAND	0.032	0.0002	0.0136	-0.0010	-0.0038	$+0.0056$	0.137	0.078
MIMIC-III, Latent-ODE	0.027	0.0001	0.0102	-0.0002	-0.0040	$+0.0047$	0.139	0.062

We withdraw the reading an earlier version gave this table, that the ceiling is not a property of our model. s_T is defined through a model's errors, and changing the residuals changes both its numerator and its denominator. Similar point estimates under wide intervals are an observed similarity, not invariance. What the table does show is this. On PhysioNet $\sqrt{s_T}$ is 0.109–0.142 on z-scored PhysioNet 0.038–0.043, and on z-scored MIMIC-III 0.024–0.033. Density's increment excludes zero for four of the six raw PhysioNet models, all but ours and mTAND, whose lower limits sit just below zero, and for none of the twelve z-scored ones. The floor depends on the forecaster on PhysioNet: its lower limit is above zero for four models, from 0.04 to 0.38, and not for GRU-D or Latent-ODE, where a lookup fitted on half the sequences ranks the other half no better than chance, and for Latent-ODE worse. Under z-scoring the floor's lower limit is above zero for five of six PhysioNet models (0.07 to 0.12; T-PatchGNN -0.003) and all six MIMIC-III ones (0.06 to 0.17), and since density adds nothing there, it comes from channel and horizon.

The denominator, and the four-benchmark table. "Sampling geometry accounts for on the order of one percent of error variance" has a denominator we never measured. Regressing the error out-of-fold onto every query-time feature, grouped by sequence, shows the e^2 scale cannot carry the statement: the R^2 is negative (-0.0041 ± 0.0049, 5 seeds) while the *same* predictions correlate with e^2 at $\mathrm{corr}^2 = 0.0104 \pm 0.0054$ ($5/5$ seeds, $p = 0.013$). Signal with the wrong *scale*, not absent signal: at kurtosis 4.6×10^4 the top 1% of points carry 94.7–100% of $\mathrm{Var}[e^2]$, so the fitted conditional mean loses to the grand mean in squared error while still ranking. This is the fourth reading of the number we have published and the first that is identified; the other three, and why the *size* of the total is not identified on this scale at all (estimates span 0.002–0.025 with $\hat{s}$ inside and no route separating from it, $p \geq 0.20$), are in Appendix C. We therefore report the share on the tail-tamed scales where the estimator is stable.

Table 5: Out-of-fold predictable share of the forecast error (PhysioNet, 5 seeds), under nested feature sets. Columns are cumulative. On e^2 — the scale Corollary 1 is stated on — *nothing* is predictable: the full-information regression does not beat the mean. On any tail-tamed transform the same features explain 20–47%, and local density contributes $+0.04$ to $+0.05$ *on top of* channel identity and horizon, so the density signal is neither a units artifact nor a lead-time artifact.

target	channel	+ horizon	+ density	full (S, V)	density marginal
e^2	$+0.004$	-0.014	-0.010	-0.004	$+0.004$
$\log e^2$	$+0.321$	$+0.345$	$+0.398$	$+0.458$	$\mathbf{+0.053}$
$\lvert e \rvert$	$+0.091$	$+0.063$	$+0.105$	$+0.196$	$+0.042$
$\operatorname{rank} e^2$	$+0.351$	$+0.371$	$+0.424$	$+0.470$	$\mathbf{+0.052}$

Table 6: Density's marginal contribution to the out-of-fold predictable share, above channel identity and forecast horizon; four benchmarks, 5 seeds, full test split. *Seed standard deviations are not printed*: all five seeds score the same points, so they are the wrong variance component, and a number labelled wrong in its own caption is still read as a number. The right ones are a sequence bootstrap and a per-dataset permutation null on ρ, which calibrate the marginals to MIMIC $\mathbf{+0.076}$, PhysioNet $\mathbf{+0.052}$, Activity $+0.006$ (raw CI **straddling zero**) and USHCN $\mathbf{-0.002}$. The null floor is neither zero nor common, so raw marginals are not on one scale; calibrated they form a gradient, not the two-versus-two partition an earlier draft claimed (Appendix C).

	rate dispersion	$\log e^2$	$\lvert e \rvert$	$\operatorname{rank} e^2$	total η^2, rank
MIMIC-III	$8.5\times$	$\mathbf{+0.071}$	$+0.036$	$\mathbf{+0.075}$	0.31
PhysioNet	$22\times$	$+0.053$	$+0.042$	$+0.052$	0.47
Activity	$1.3\times$	$+0.008$	$+0.035$	$+0.004$	0.17
USHCN	$1.05\times$	-0.020	-0.032	-0.021	0.47

We will claim: local density carries genuine out-of-fold information about forecast error on both clinical datasets, above channel identity, above horizon, and far above each dataset's own permutation null. **We will not:** that the near-regular datasets carry *none*; that dispersion *orders* the effect (MIMIC's marginal is largest at $8.5\times$, PhysioNet's smaller at $22\times$); or that the transform is immaterial — the $|e|$ column makes MIMIC and Activity indistinguishable ($p = 0.82$). Two caveats: Corollary 1 is transform-specific, so a practitioner should estimate the share on the scale their loss uses, and a *difference* of two R^2 lower bounds is not itself a bound — an earlier draft claimed the marginal "could only be larger", which does not follow, and ρ's $+0.70$ rank correlation with forecast distance is a mechanism by which it could be smaller.

The withdrawn ceiling intervals. PhysioNet's published $[0.019, 0.146]$ is exactly $\min$ of five per-seed lower bounds and $\max$ of five upper bounds — 0.019 from seed 4, 0.146 from seeds 2 and 5. A union of five intervals has coverage far above nominal and an uninterpretable width. The per-seed intervals themselves run $[0.026, 0.127]$ to $[0.042, 0.146]$ around ceilings of 0.077 to 0.100. Pooling all 1715 cells into one bootstrap gives $[0.068, 0.114]$, narrower by $2.7\times$ in the s domain, and is wrong for the mirror-image reason: the five seeds score the *same* test points, so pooling inflates the effective n fivefold. **The sequence bootstrap does not rescue it either, and the reason is specific to this estimator.** Resampling sequences gives $[0.059, 0.318]$, $2.5\times$ the width of the cell bootstrap — so cell resampling does understate substantially, as expected — but its *centre* moves from 0.086 to 0.156, and a resampling unit must not move the point estimate. The cause is the analytic noise floor: it subtracts var/n assuming the n points in a tercile are independent, while a with-replacement sequence resample duplicates sequences, so n counts copies, the floor is *under*-subtracted and the corrected contrast comes out too high. The width ratio is therefore informative about how badly cell resampling understates; the location is not. **What we report is one seed's interval with the seed-to-seed spread stated separately** — PhysioNet seed 1 gives $[0.026, 0.127]$ around 0.085, and the five seed ceilings span 0.077 to 0.100 — and no pooled interval, because we do not have a construction that earns one. Activity's published $[0.045, 0.104]$ does not reproduce even as an envelope of the checkpoints we now hold, whose per-seed intervals run $[0.042, 0.108]$; that is a provenance defect of its own. USHCN's $[0.000, 0.104]$ was the case worth checking first, because a correctly built interval

excluding zero would have reversed a claim in the body: it does not, and every one of the five seeds returns $[-0.0001, +0.0111]$ in s (`scripts/bound_validation.py`).

Estimating s, in full. Within each (channel, horizon-decile) stratum we contrast the sparsest and densest density terciles, removing the horizon confound by construction. The raw contrast is inflated by within-cell sampling noise, subtracted analytically ($0.028 \to 0.0084$ on PhysioNet); on synthetic data with a planted share the corrected estimator attenuates by $\approx 60\%$. *We no longer report a calibrated value beside the raw one.* The calibration was introduced to close a gap between $\hat{s}$ and a measured correlation, and that gap no longer exists now that the interval spans zero; a calibration factor applied to an estimate not distinguishable from zero is not interpretable, and multiplying a null by 2.5 gives a larger null. A reviewer predicted the within-stratum numerator against a pooled denominator would deflate $\hat{s}$ several-fold and invert our conclusion; measured, the between-stratum share of $\mathrm{Var}[e^2]$ is 1.5% — heavy-tailed error puts the variance inside the cells — so the ceiling moves from 0.091 to 0.092. With a stratum bootstrap: PhysioNet $\hat{s} = 0.0084$ ($\sqrt{\hat{s}} = 0.092$, 343 strata, 5 seeds), Activity 0.0061 (0.078), USHCN 0.0044 (0.066), MIMIC-III 0.00077 (0.028), the last three on 3 seeds. The intervals previously printed here were min/max envelopes over per-seed bootstraps and are withdrawn (Section 4.1.2); Activity's $[0.045, 0.104]$ does not even reproduce as the envelope of the checkpoints we now hold, whose per-seed intervals run $[0.042, 0.108]$, which is a provenance defect of its own. USHCN's interval reaches zero, so there we report no detectable sampling-attributable component rather than a small one.

The four readings of "one percent", and why the size is not identified. The number has been published four times: as "one percent of error variance" with no denominator; as "the prize does not exist in that metric", from the negative out-of-fold R^2 on e^2; as $2.4\hat{s}$, from a single route; and as the calibration diagnosis of Section 4.1.5. Only the last is identified, and even it does not fix the *size* of the total: across losses and retransformations the point estimates span 0.002–0.025, $\hat{s} = 0.0084$ lies inside that interval, and no route separates from it ($p \geq 0.20$). The fraction of the e^2-scale prize attributable to sampling is therefore not identified and we claim none. The e^2 and $|e|$ columns of Table 6 are additionally sign-unstable to subsampling with a spread wider than $\hat{s}$, which is why the reported share is on $\log e^2$ and $\mathrm{rank}\, e^2$ (`scripts/tail_dominance.py`, `scripts/total_predictable_share.py`).

The sequence bootstrap and the permutation nulls, in full. MIMIC $+0.0747$, CI $[+0.067, +0.081]$, null -0.001, calibrated $+0.076$; PhysioNet $+0.0504$, CI $[+0.046, +0.055]$, null -0.001, calibrated $+0.052$; USHCN -0.0203, CI $[-0.029, -0.012]$, null -0.018 ± 0.006, calibrated -0.002; Activity $+0.0062$, CI $[-0.001, +0.013]$, null -0.000 ± 0.002, calibrated $+0.006$. All four are from our own runs; earlier drafts quoted a reviewer's (`scripts/share_uncertainty.py`).

D THE FULL UQ BOARD

Summarized in Section 4.2. **Risk–coverage convention.** We rank test points by a score, abstain on the top fraction f, and average the risk on the retained points. The risk–coverage curve and selective risk are due to El-Yaniv & Wiener (2010); we use the area under the empirical curve as Geifman et al. (2019) define it, the *rectangular* mean $\frac{1}{n}\sum_\theta \hat{r}(f, g_\theta)$ over thresholds, which is the discretization every board here uses and the one the fourth copy of our own code drifted from (Appendix K). Our normalization is not theirs: they subtract the oracle's AURC to form an excess measure, whereas we divide by the no-ranking value so that 1.0 means a score that orders nothing.

D.1 IS CONFORMAL ALONE THE CONTRIBUTION? UQ BASELINES (E2B)

A natural challenge is that split conformal is model-agnostic — *any* forecaster can be wrapped to hit nominal coverage — so calibration by itself is not the contribution. We test this directly: four standard UQ methods — Deep Ensembles (Lakshminarayanan et al., 2017), MC-Dropout (Gal & Ghahramani, 2016), quantile regression (Koenker & Bassett, 1978; Romano et al., 2019), and a heteroscedastic Gaussian head (Nix & Weigend, 1994) — on a shared GRU-D backbone, each given the identical split conformal (Table 7). Conformal does bring all five methods to 0.90 coverage, confirming the challenge; the discriminator is therefore *sharpness*. At matched coverage on PhysioNet BandRate's

Table 7: UQ baselines vs BandRate under the *same* model-agnostic split conformal (test split, 3 seeds, matched 10-epoch budget; baselines share a GRU-D backbone, so this is a separate apples-to-apples board from Table 29). Conformal brings every method to ∼nominal 90% coverage, so the discriminators are interval *width* and the *interval score* (IScore, Winkler: width plus a coverage-violation penalty). CRPS is pre-conformal. Best width/IScore per block in bold. BandRate's PhysioNet width here (88.6) exceeds its main-board value (∼ 72, Table 29) because this board retrains BandRate fresh under the shared matched budget; each board is internally consistent and the cross-method comparison within this table is the valid one.

Data	Method	CRPS↓	Cov@90_{cf}	Width@90↓	IScore@90↓
PhysioN.	Deep Ensemble	20.1	.904	108.8	181
	MC-Dropout	19.2	.904	103.3	**178**
	Gaussian NLL	19.5	.904	105.8	179
	Quantile/CQR	28.0	.903	164.6	285
	BandRate	22.2	.898	**88.6**	277
Activity	Deep Ensemble	.115	.919	.678	1.04
	MC-Dropout	.129	.926	.787	1.14
	Gaussian NLL	.118	.921	.722	1.09
	Quantile/CQR	.120	.918	.752	1.15
	BandRate	**.107**	.911	**.632**	**1.00**
USHCN	Deep Ensemble	.364	.900	2.04	3.36
	MC-Dropout	.360	.899	1.98	3.33
	Gaussian NLL	.361	.899	2.00	3.35
	Quantile/CQR	.318	.899	**1.48**	3.23
	BandRate	**.276**	.894	1.60	**2.74**
MIMIC	Deep Ensemble	**.355**	.901	**1.88**	**3.11**
	MC-Dropout	.371	.899	1.97	3.27
	Gaussian NLL	.365	.901	2.00	3.22
	Quantile/CQR	.357	.899	1.97	3.42
	BandRate	.359	.898	1.95	3.24

intervals are 14% narrower than the best baseline (88.6 vs MC-Dropout's 103.3) and 46% narrower than quantile regression, with a similar margin at the 50% level (∼ 27 vs 32.1 for the narrowest baseline at that level, quantile/CQR — note the 90% comparison uses the best baseline and this one a different method); on Human Activity the margin extends — BandRate has both the narrowest intervals (0.632 vs 0.678) *and* the lowest CRPS (0.107), best on every probabilistic metric. The near-regular USHCN is the honest exception that proves the mechanism: BandRate still has the best CRPS (0.276) but *not* the narrowest interval (quantile regression is, 1.48 vs 1.60), exactly as the σ-swap predicts (Appendix N.6) — where sampling density barely varies, NABU's shape cannot sharpen, so the width advantage is specific to the genuinely irregular datasets. The reason is NABU: an identifiability-ordered $\hat{\sigma}$ lets the single conformal multiplier yield tighter typical-case intervals where sampling is irregular. Honest caveat: on PhysioNet the simple Gaussian heads have marginally lower *pre*-conformal CRPS (19–20 vs 22; not so on Activity), so BandRate's edge is, at minimum, calibrated sharpness rather than raw likelihood. The interval score — which also charges for how far the missed 10% fall — turns this into an honest trade-off: BandRate wins it on Activity (1.00) and USHCN (2.74), but on heavy-tailed PhysioNet its confident narrow intervals overshoot the occasional extreme, so the wider Gaussian heads score better there (178 vs 277). Density-adaptive sharpness thus buys the narrowest *typical* width at a tail-risk on heavy-tailed targets — a limitation we state plainly rather than hide behind width alone. It appears intrinsic to narrow density-adaptive intervals rather than an artifact of the Gaussian assumption: retraining the head with a heavy-tailed Student-t likelihood and post-hoc σ-reshaping both *worsen* the interval score, and the quantile/CQR baseline scores worse still. A plausible mechanism: NABU's variance tracks density more strongly ($\rho_s = +0.70$) than realized error does ($+0.10$, Section 4.1), so it is somewhat *over*-density-responsive, sharpening confidently in dense regions where a rare extreme is then missed by a wide margin. The broader takeaway still corrects an earlier overclaim — BandRate is not "the only calibrated model" (anyone can conformalize); it is the model whose variance *shape* gives the tightest *typical-case* intervals. The fourth dataset keeps us honest: on normalized MIMIC neither width nor interval score favors BandRate (a GRU-D deep ensemble is best, its agreeing members yielding a tight well-scored

band), so across all four *no* UQ method dominates. We visualize this frontier in Figure 5. BandRate's distinguishing property is thus not universal sharpness but that its variance is the only explicitly *density-adaptive* one (we no longer call it theory-grounded; Section 3.2) — the within-channel signal a pooled head does not carry (Section 4.1) — while it stays competitive on width and interval score on every dataset.

E THE FIVE DELETION CONTRASTS

E.0.1 DOES THIS GENERALIZE? FIVE CONTRASTS, NONE DISTINGUISHABLE FROM ZERO

A single backbone on a single dataset cannot support a claim about a family of methods, so we repeated the no-density contrast on all four benchmarks and on a second architecture — a plain GRU-D Gaussian head with the portable density term of Appendix X ($\log \sigma^2 \mathrel{+}= \mathrm{softplus}(\gamma_n) \cdot z(-\log \rho_n)$, whose sign is architecturally forced). Each arm is identical to its control in backbone, head, optimizer and budget; the PhysioNet/SC-INR row is the exception and is *not* matched. Its no-density arm is 3 seeds from one board and its with-density comparator 5 seeds from another, and the two configurations also differ in early-stopping patience (20 against none) — so the row varies at least three things at once, not the density input alone, and it is the row the body quotes most often. We mark it here and refuse it in Appendix J, where a configuration diff gates the comparison rather than a reader's memory. The GRU-D row is a single seed.

Table 8: The density-deletion contrast on five backbone–dataset arms. [§]The 95% upper limit of the extreme-tercile contrast $\hat{s}_{\mathrm{ext}}$, a diagnostic and not a ceiling (Section 4.1.2); an earlier version headed this column "ceiling: at most". The PhysioNet/SC-INR row is *not* matched — its two arms differ in seed count, board and early-stopping patience — and is refused in Appendix J.

data	backbone	$\hat{s}_{\mathrm{ext}}$ upper[§]	no density	with density	marginal	SE	p vs 0
PhysioNet[†]	SC-INR	0.204	$+0.217 \pm .055$	$+0.142 \pm .060$	-0.075	0.049	0.25
Activity	SC-INR	0.118	$+0.317 \pm .032$	$+0.355 \pm .009$	$+0.038$	0.024	0.23
USHCN	SC-INR	0.112	$+0.184 \pm .038$	$+0.119 \pm .114$	-0.065	0.085	0.51
MIMIC-III	SC-INR	0.106	$+0.296 \pm .004$	$+0.290 \pm .014$	-0.006	0.010	0.61
PhysioNet	GRU-D	0.204	-0.011	$+0.040$	$+0.051$	—	—

[†]unpaired: 3-seed no-density arm vs a 5-seed with-density comparator from a different board, differing also in early-stopping patience (Appendix J). ± and SE are seed standard deviations, n=3 except the PhysioNet with-density arm (n=5) and the GRU-D row (n=1) — the *wrong* variance component, since the seeds score the same points. The ceiling column is the 95% *upper limit* on the partial quantity, sequence-clustered (Section 4.1.2); the point estimates it replaced were 0.092, 0.078, 0.066 and 0.028, and none of the four is distinguishable from zero. They are kept here only because the p column is computed from them, and the direction is safe: a seed sd understates the sampling uncertainty, so every "not distinguishable from zero" in this table is conservative rather than generous. Appendix K records the sweep.

An earlier draft presented this table as "five tests of the bound, five passes". That reading was wrong twice over and we retract it.

The bound does not cover a marginal contribution. Corollary 1 constrains $\mathrm{corr}(g(S), e^2)$ for a head reading S alone and says nothing about $\mathrm{corr}(f(S,V), e^2) - \mathrm{corr}(f_0(V), e^2)$ for two heads that both read V. A pure $S \times V$ interaction makes that gap arbitrarily large at $s = 0$: with S, V independent Rademacher and $e^2 = 1 + SV$ (positive, as a squared error must be), $\mathbb{E}[e^2 \mid S] = 1$ is constant so the ceiling is 0, yet $f(S,V) = SV$ correlates 1 with e^2. NABU's ratio $r_{n,k} = \rho_n / 2f_k^{\max}$ *is* such an interaction — density scaled by a bandwidth read from the value context — so the bound was applied to the one functional form it cannot constrain. The earlier draft's checkmark column also took an absolute value the theory never licenses, since nothing bounds a negative marginal; both are removed.

The contrasts have no power to fail. No marginal is distinguishable from zero (Welch $p \geq 0.23$), and USHCN's standard error (0.085) is of the same order as the largest effect the pathway could produce there at all (0.112, the 95% upper limit). An earlier draft compared that SE against the *point* estimate 0.066 and said it exceeded the ceiling; that was an understated standard error against a

number with no interval, which is two errors that happened to point the same way (Appendix K). "Five passes" records that deleting an input did not destabilize training. We withdraw the accompanying $\rho_s(\sqrt{\hat{s}}, |\text{marginal}|) = +0.80$ for the same reason: its exact two-sided permutation p is 0.33, the minimum attainable at $n{=}4$ is 0.083, and rows 1 and 5 are one dataset carrying one ceiling. We had refused an $n{=}4$ relationship two subsections earlier when it went against us; applying a softer standard to the flattering one was the error.

What the table does support is weaker and ours to keep: the sign of the density pathway's contribution is unstable, so no claim that removing density helps survives; and the magnitude is uniformly small, every effect within one seed standard deviation of zero. That is an absence of signal consistent with a small ceiling, not a test of one. The real test is Table 23.

F Stability of the pooled baseline

F.0.1 Is the baseline itself broken?

A fair objection to diagnosing our own pooled ablation is that its variance head may simply be unstable, in which case we would be describing an optimization failure. It is partly true and we quantify it: across five seeds the pooled $\hat{\sigma}^2$ takes $\approx 110{,}000$ distinct values on three seeds but collapses to 63 on one and to a *single* value on another — under the same β-NLL stabilizer NABU uses, so it is a genuine instability of the pooled parameterization. The diagnosis survives it: restricting to the three non-degenerate seeds gives -0.14 raw and $+0.038$ horizon-controlled, against -0.12 and $+0.033$ with all five. What the degeneracy does undermine is the pooled head's role as an *abstention* baseline — a collapsed σ^2 ranks nothing — so the selective-prediction comparison against it is not evidence that NABU ranks error well. The deployed-UQ board was our evidence there, and it was native-scale; standardized the reading does not survive (Appendix I), so nothing currently settles it.

G The no-density arm on PhysioNet

Summarized in Section 4.1.4. Testing the bound needs an arm with the density pathway removed, which neither of our heads is: both the per-band NABU head and the "pooled" ablation consume ρ_n. We trained a third arm — same backbone, head, optimizer and budget — with the density input deleted (116,450 vs 116,514 parameters, the difference being exactly that one input dimension). It is *not* seed-matched to its comparator: the no-density arm is 3 seeds from this board while the with-density numbers are 5 seeds from a different board, so the contrast is unpaired and we mark it wherever it appears. A second scope caveat, raised in review: each arm has its own e^2 and therefore its own s and its own ceiling, whereas we compare all three arms and the oracle against one. On this data the arms' ceilings agree within seed noise ($\sqrt{\hat{s}} = 0.086 \pm 0.008$ for the no-density arm against 0.091 ± 0.010 for NABU, per-cell 0.188 ± 0.027 against 0.183 ± 0.017), so the error is one of scope rather than of arithmetic — but the comparison is licensed only arm-by-arm.

Table 9: The two arms of the no-density control on PhysioNet, each scored against realized error. Their ceilings agree within seed noise, so the earlier error was one of scope rather than of arithmetic; the comparison is licensed arm-by-arm only.

variance head	density input	$\rho_s(\hat{\sigma}^2, e^2)$
pooled	yes (per-channel)	$+0.142$
NABU	yes (per-band)	$+0.152$
no-density	**none**	$\mathbf{+0.217}$

Deleting the density input *raises* the horizon-controlled correlation with realized error by roughly half ($+0.230$ uncontrolled). An earlier draft called this the sharpest result in the paper, read it as the density pathway's marginal contribution being negative, and used it to predict the paper's other negative results. We retract all three readings.

Why. Corollary 1 bounds $\text{corr}(g(S), e^2)$ for a head reading S alone and says nothing about a difference between two heads that both read the value context — the Rademacher counterexample in Section 4.1.6 shows that difference is unbounded even when the ceiling is zero. Opening this

Table 10: Normalized AURC under test-time thinning (PhysioNet, 3 seeds; 1.0 = each method's own value at the native sampling rate, lower is more robust). **Neither the native-scale ordering nor the raw-σ standardized one survives.** On native scales abstention is 97% a contest over which score sorts channels by error magnitude (Section 4.1.7). The middle block standardizes the loss but ranks by the raw σ, comparing channels in mismatched units, as earlier versions did; the bottom block ranks by $\hat{\sigma}^2/d_c$. Aligned, every method stays within 6% of its own native-rate value and BandRate at keep=0.2 reads $1.025/0.980/1.040$ across seeds, $+0.010 \pm 0.047$ against the ensemble. We withdraw "BandRate is the least shift-robust of five methods" and do not replace it with the reverse: this experiment separates nothing. MC-dropout samples at evaluation, so its rows move slightly between runs.

keep prob.	1.0	0.8	0.6	0.4	0.2
native scale (confounded by channel units)					
BandRate (NABU)	1.000	1.012	1.368	1.232	1.974
ensemble (GRU-D)	1.000	1.068	1.130	1.164	1.484
MC-dropout	1.000	1.012	0.847	1.058	1.098
Gaussian head	1.000	0.985	0.996	0.964	1.127
quantile regression	1.000	1.004	0.969	0.976	0.986
per-channel standardized loss, raw-σ ranking (units mismatched)					
BandRate (NABU)	1.000	0.933	0.924	0.882	0.943
ensemble (GRU-D)	1.000	0.999	0.999	0.996	1.000
MC-dropout	1.000	1.000	1.002	1.005	1.012
Gaussian head	1.000	1.001	1.001	1.003	1.004
quantile regression	1.000	1.000	1.001	1.003	1.004
per-channel standardized loss, loss-aligned $\hat{\sigma}^2/d_c$ ranking					
BandRate (NABU)	1.000	0.987	0.993	0.955	1.015
ensemble (GRU-D)	1.000	0.987	0.979	0.964	1.005
MC-dropout	1.000	1.006	1.005	1.017	1.040
Gaussian head	1.000	0.998	1.001	0.999	1.028
quantile regression	1.000	1.001	1.002	1.008	1.001

appendix with "the bound constrains the marginal contribution of the sampling pathway" was exactly the category error retracted there. The effect also does not replicate: on four benchmarks and a second architecture (Appendix E.0.1) the marginals run -0.075 to $+0.051$, unstable in sign and, on every arm, within one seed standard deviation of zero (Welch $p \geq 0.23$). And this row is the unpaired one.

What survives, and it is what the paper's headline now rests on: a head reading no density at all reaches $+0.217$ against $+0.142$ and $+0.152$. We do *not* assert that ordering: the difference is -0.075 ± 0.049 ($p = 0.25$), the same statistic Section 4.1.6 reports as indistinguishable from zero, and this row is additionally unpaired. What it does establish is that we have no evidence for crediting any part of NABU's 0.152 to the density pathway, since the value context alone produces a number in the same range — an absence of evidence, not evidence of absence.

H TWO FALSIFIED CLAIMS, IN FULL

Summarized in Appendix K.1.

H.0.1 A FALSIFICATION WE NOW WITHDRAW: THE SHIFT-ROBUSTNESS ORDERING WAS A UNITS ARTIFACT

The pre-registered rescue hypothesis was that a variance grounded in observation geometry should *extrapolate* under a sampling-rate shift where an ensemble — whose members all saw the same sampling distribution — can be confidently wrong together. We Bernoulli-thin the test history mask and track each method's risk–coverage AURC, normalized to its own value at the native rate. Earlier drafts asserted the falsification in five places without showing the numbers; Table 10 shows them, on both scales, and the falsification does not survive the standardized one.

What the aligned panel shows is not a reversal but an absence. Every method's ranking is nearly insensitive to thinning, within 6% of its native-rate value at every level, and BandRate moves no differently from the ensemble (1.025, 0.980, 1.040 at keep$=0.2$). The raw-σ standardized panel, which earlier versions read, suggested BandRate *improved* under thinning while the baselines stood still; that too was the unit mismatch. The native-scale story, that the methods degrading least are the ones that never read ρ, was a statement about which score happens to order channels by error magnitude rather than about extrapolation. We reported that story as one of our own falsified rescues for three rounds. Retracting a *negative* result is unusual and we flag it: a falsification computed on a confounded metric is no more reliable than a confirmation would have been.

H.0.2 Does the ceiling predict where the mechanism pays? We retire the question

We initially read the per-dataset ceilings as the applicability test this line of work has lacked, because their ordering (PhysioNet 0.092, Activity 0.078, USHCN 0.066, MIMIC-III 0.028) appeared to match where NABU helped. That reading used a *different* benefit metric on each dataset — AURC on PhysioNet and MIMIC, the ratio ablation on Activity, the σ-swap width on USHCN — and we withdrew it. Re-measuring all four with one metric reversed the pattern, and we published the reversal (PhysioNet: largest ceiling, worst rank $4/5$) as an *anti*-ordering. **We withdraw that too, and we do not replace it with a third answer.** The uniform metric was native-scale risk-normalized AURC, which Section 4.1.7 shows is 97% a contest over which score sorts channels by error magnitude and which Appendix M measures per dataset. PhysioNet's $4/5$ is that confound: a channel lookup table captures 97% of everything abstention can buy there, against 40% on MIMIC.

Table 11: The per-dataset ordering on both scales. The native-scale column is withdrawn: abstention there is 97% a contest over which score sorts channels by error magnitude, a confound Appendix M measures per dataset.

		native (withdrawn)		std., raw σ (withdrawn)		std., aligned	
dataset	ceiling $\sqrt{\hat{s}}$	vs best other	rank	vs best other	rank	vs best other	rank
PhysioNet	0.092	$+464\%$	4/5	-33.8%	**1/5**	-7.8%	**1/5**
Activity	0.078	-15.7%	**1/5**	-17.8%	**1/5**	-15.9%	**1/5**
USHCN	0.066	-14.7%	**1/5**	$+13.4\%$	4/5	$+6.5\%$	4/5
MIMIC-III	0.028	$+5.0\%$	3/5	-0.3%	**1/5**	$+0.9\%$	2/5

All three columns, kept so the withdrawals show their work. Standardization changed our rank on three of the four datasets — PhysioNet $4\to1$, MIMIC $3\to1$, USHCN $1\to4$ — and that column was itself flawed: it ranked by the raw σ against a standardized loss. Put in the loss's units, the variance ranks $1, 1, 4, 2$ and the PhysioNet margin shrinks from 34% to 8%. Earlier versions printed PhysioNet's native gap as $+341\%$ from a board that predated the paper's single risk–coverage implementation. The aligned column is a *fifth* answer to the ordering question. That is the reason to retire the question rather than re-derive it: an answer that arrives in our favour, on a metric we changed this revision, at $n{=}4$, is precisely what we refused when $\rho_s = +0.80$ arrived against us. We print both columns because a withdrawal that replaces a number is indistinguishable from harvesting unless both numbers stay visible (`scripts/regime_uniform_metric.py`).

And the ordering is not a property of the data. We had told a clinical-versus-near-regular story since round 3. Two metrics give two different pairs:

Table 12: Two metrics, two different orderings of the same four datasets — which is why the clinical-versus-near-regular story we told from round 3 onward is retired rather than restated. ‡Activity's increment excludes zero only when its 607 test chunks are resampled as if independent; they come from three recordings (Appendix B).

dataset	Table 6 rank marginal (null-calibrated)	AURC density increment
MIMIC-III	$+\mathbf{0.076}$ (strongest)	*not establishable*
PhysioNet	$+0.054$	establishable
Activity	$+0.004$	*not establishable*‡
USHCN	-0.001	*not establishable*

MIMIC goes from the paper's strongest dataset to one where nothing can be established, and the two metrics resolve together on PhysioNet alone; an earlier version marked Activity's increment establishable, which rested on the chunk-level resampling withdrawn in Appendix B. That is a stronger reason to retire the question than $n{=}4$ was: it is not that four points cannot establish an ordering, it is that the ordering is not stable across the ways we measure it, so a fifth dataset would not settle it either.

Why we are not re-deriving it under standardization. The question has now been asked three times and produced three answers — the ceiling orders the benefit, anti-orders it, and neither — and a fourth would face the arithmetic that disposed of the first three: at $n{=}4$ no distribution-free test can return below 0.083, and we withdrew $\rho_s(\sqrt{\hat{s}}, |\text{marginal}|) = +0.80$ at $p = 0.33$ on exactly that ground (Appendix E). A favourable fourth answer would put us back inside the double standard we removed; an unfavourable one would be a fifth retraction. Four datasets cannot settle whether the ceiling predicts where the mechanism pays, and we no longer offer a reading either way. What takes its place in Figure 1 is the floor of Section 4.1.7 — a quantity one dataset establishes, which errs in a known direction, and which needs no ordering across datasets at all.

Two readings that rested on the withdrawn ordering go with it: that NABU "harms most exactly where density varies most", and that the near-regular datasets are where it ranks best. Both were statements about channel units. The no-density arm above already shows the sampling input has no reliable sign, and that result stands on its own contrast rather than on this ordering.

I FOUR FURTHER CHECKS ON THE MECHANISM

Summarized in Section 4.1.8; the evidence is here.

I.0.1 DOES THE DIAGNOSTIC TRACK UNCERTAINTY *quality*?

A reviewer-grade objection to $\rho_s(\sigma^2, 1/\rho_n)$ is that NABU consumes ρ_n directly, so the statistic partly measures whether a head echoes its own input rather than whether its uncertainty is good. We therefore measured the quantity that does matter — the within-channel correlation between predicted σ^2 and *realized* squared error, same horizon control, 5 seeds — and the answer is sobering: $+0.142$ for the pooled head versus $+0.152$ for NABU. The density statistic moves from $+0.03$ to $+0.53$, a seventeen-fold change, while the error-tracking that a practitioner actually cares about is statistically indistinguishable between the two heads. We take this as the central negative result of the paper: supplying the identifiability signal demonstrably changes what the variance encodes about *sampling*, and demonstrably does not, on this dataset, make the variance a better predictor of *error*. It is also the mechanism behind the two results that would otherwise look unrelated — NABU ranking fourth of five on *native-scale* PhysioNet AURC, an ordering Appendix I withdraws and losing interval score on heavy-tailed data (Section 4.2). Readers should treat $\rho_s(\sigma^2, 1/\rho_n)$ as a diagnostic of *what a head represents*, not as a proxy for uncertainty quality, and we no longer claim the latter.

Is there a deficit to repair at all? The premise needs an oracle: does realized error actually rise with within-channel sparsity once the same horizon control is applied? It does. Partialling horizon out of the realized-squared-error-vs-sparsity correlation gives $+0.13$ (5 seeds, $+0.16$ uncontrolled; $+0.12$ measured on the NABU checkpoints), so the signal a variance head ought to carry is genuinely positive and the pooled head's $+0.03$ genuinely misses most of it. The same oracle indicts our own fix, however: NABU's $+0.53$ *overshoots* the target roughly four-fold. Scored as distance from the oracle the two heads are comparable (0.10 for pooled, 0.40 for NABU, in opposite directions), and this over-response is, we believe, the mechanism behind NABU's interval-score loss on heavy-tailed PhysioNet (Section 4.2): variance that tracks density more strongly than error does will be too narrow where data is dense. One statistical caveat cuts the other way: a single realized squared error is a noisy estimate of the conditional variance, which attenuates the oracle correlation, so $+0.13$ is a lower bound on the ideal target rather than the target itself. We report both readings and do not claim NABU's magnitude is calibrated — only that the signal exists and the pooled head does not carry it. We therefore do not claim the pooled head is directionally wrong once horizon is controlled; the defensible claim is that it carries almost no within-channel density signal and that NABU supplies it. We previously offered Shukla & Marlin (2022) as agreeing, which inverts what that paper reports: the near-constant uncertainty it documents belongs to *homoscedastic* output layers, and its intensity

encoding is the repair. Our pooled head reads density and still carries almost none of the signal, so this is a contrast with that result rather than support from it — which is the point, the failure being architectural rather than a missing input. On the uncontrolled statistic NABU reads $+0.70$ (5-seed; CI $[+0.65, +0.75]$; 92% positive) against the pooled head's -0.12, with disjoint intervals; under the horizon control that both numbers require, the comparison is $+0.53$ against $+0.03$. The correction has decision value, but far less than we reported. On a risk–coverage test (abstain on the highest predicted variance, measure error on the retained), NABU cuts the area under the risk–coverage curve by 28–88% versus the pooled head on all 5 seeds — *on native scales* (mean $65\pm28\%$; normalized AURC 0.314 against 0.920; paired t p=0.008). That board carries the channel-units confound of Section 4.1.7 exactly like every other native-scale AURC in this paper, and it is the one the confound flatters *us* on. Per-channel standardized and ranked by the variance in the loss's units, $\hat{\sigma}^2/d_c$, the gain falls to $12.0\pm11.1\%$ (0.882 against 1.003), NABU wins on $4/5$ seeds rather than $5/5$, and the paired p is 0.074 — above the 0.0625 floor that n=5 imposes on any distribution-free test, so this is not a significant gain (ranked by the raw variance, as earlier versions did, it read $11.3\pm19.1\%$ at $p = 0.269$), and it fails the pre-registered $\geq 20\%$ utility gate that the native number passed. *These figures moved when an audit found this board computing a different risk–coverage metric from the rest of the paper* — averaging over $f \in [0, 0.8]$ by trapezoid rather than $[0, 0.9]$ by rectangular mean, resolving ties by array order although the pooled baseline's σ^2 collapses to a constant on one seed, and standardizing each channel in-sample. The earlier reading was $+73\%$ and $+15.9\%$ at $p = 0.079$; the correction leaves the conclusion and moves the significance well away from the boundary (Appendix K). We report the standardized figure as the real one. The direction survives — the pooled head's variance points at the *easy* points, so abstaining on it *raises* error, while NABU's flags the hard ones — but the magnitude does not.

I.0.2 Against deployed UQ: the comparison we published was a units artifact

Ranking the same PhysioNet test set with the four baselines of Table 7 (3 seeds, identical protocol), BandRate's variance is the *fourth* of five error-rankers on native scales: AURC 3823 against 702 for a GRU-D deep ensemble, 1091 for a plain Gaussian head, 1136 for MC-dropout, beating only quantile regression (8557). We reported that for four rounds, including in the abstract, as one of our own honest negatives, and recommended ensembling over NABU on its strength. (Earlier versions printed 3022, 998 and 1246: that board predated the paper's single risk–coverage implementation with random tie-breaking, and NABU's variance has many exact ties. The ordering is unchanged.)

Neither reading survives a fair comparison. Native-scale abstention on PhysioNet is 97% a contest over which score orders channels by error magnitude (Section 4.1.7), so we had standardized each channel's loss — and then ranked by the raw σ, which compares channels in mismatched units. The fair ranking puts the variance in the loss's units, $\hat{\sigma}^2/d_c$.

Table 13: Normalized AURC of five error-rankers on the two clinical datasets, mean $\pm$ sd over 3 seeds. "Raw σ" ranks by the model's σ against the per-channel standardized loss, as earlier versions did; "aligned" ranks by $\hat{\sigma}^2/d_c$, the variance in the loss's units. MIMIC-III is z-scored, as its baselines use. `results/v2_cr_uq_physionet`, `results/v2_cr_uq_mimic3`.

	PhysioNet			MIMIC-III		
	native	raw σ	aligned	native	raw σ	aligned
BandRate (NABU)	3823	0.656 ± 0.199	0.854 ± 0.092	0.2715	0.527 ± 0.005	0.494 ± 0.007
ensemble (GRU-D)	702	0.991 ± 0.009	0.926 ± 0.024	0.2585	0.528 ± 0.004	0.489 ± 0.003
MC-dropout	1136	1.003 ± 0.002	1.014 ± 0.004	0.2980	0.559 ± 0.008	0.528 ± 0.010
Gaussian head	1091	1.000 ± 0.002	0.978 ± 0.015	0.2841	0.560 ± 0.004	0.515 ± 0.003
quantile regression	8557	1.001 ± 0.000	1.001 ± 0.000	0.3237	0.590 ± 0.006	0.575 ± 0.021

We withdraw the reading earlier versions gave the raw-σ column: that on PhysioNet BandRate beats every baseline by 34–35% and the baselines rank within-channel error essentially not at all. Both came from the unit mismatch. Aligned, the ensemble and the Gaussian head do rank within-channel error (0.926, 0.978), and BandRate is nominally first on PhysioNet but distinguishable from the ensemble by no paired test at three seeds (7.8% lower, $2/3$ seeds, $p = 0.34$). On MIMIC-III the ensemble is nominally first and the two are indistinguishable (0.9%, $p = 0.53$); BandRate is better than MC-dropout and quantile regression on $3/3$ seeds ($p = 0.009$, 0.020). What survives is narrow:

on neither clinical dataset does a deep ensemble rank error better than our head, and on neither does our head rank it better than the ensemble.

MIMIC's confound was "already mostly gone" because we had normalized it. The sentence above attributes the PhysioNet/MIMIC difference to MIMIC's scale confound being smaller without saying why it is smaller. The reason is the one Appendix H.0.2 uncovers: MIMIC was the one board trained with `dataset.normalize=true`. Retraining the whole five-method board on PhysioNet the same way (`results/v2_cr_uq_physionet_norm`, 3 seeds) makes that a demonstration rather than an inference, and it *corrects* what we wrote above:

Table 14: The five-method UQ board retrained end-to-end with `dataset.normalize=true` (`results/v2_cr_uq_physionet_norm`, 3 seeds), next to the raw board. Standardized AURC; "raw σ" ranks by the model's σ against the standardized loss, as earlier versions did, and "aligned" by $\hat{\sigma}^2/d_c$, the variance in the loss's units.

	raw board		z-scored board	
method	raw σ	aligned	raw σ	aligned
BandRate	0.656 ± 0.199	0.854 ± 0.092	0.739 ± 0.007	$\mathbf{0.654 \pm 0.010}$
ensemble GRU-D	0.991 ± 0.009	0.926 ± 0.024	0.770 ± 0.020	0.729 ± 0.018
MC-dropout GRU-D	1.003 ± 0.002	1.014 ± 0.004	0.782 ± 0.005	0.776 ± 0.025
Gaussian GRU-D	1.000 ± 0.002	0.978 ± 0.015	0.804 ± 0.004	0.754 ± 0.019
quantile GRU-D	1.001 ± 0.000	1.001 ± 0.000	0.875 ± 0.005	0.798 ± 0.006

"The baselines cannot rank error at all on PhysioNet" was a statement about raw clinical units, not about the baselines. Trained on z-scored values all four rank error genuinely (0.729–0.798 aligned against 1.0 for no ranking), the native-scale spread across them collapses from $11.7\times$ to $1.24\times$, and the native and standardized orderings almost agree where on the raw board they inverted. This is the third reading in this revision that we took from a raw board and that turned out to be a property of the scale, which is why Appendix K now names every claim resting on `results/repro`.

The control that made the raw reading look solid fails with it. On the raw board a per-channel mean-error lookup, fit on half the sequences and scored on the other half — a table that reads no input at all — beat two of the four baselines on $3/3$ seeds and our own method on $3/3$, on the native metric their authors would report. On the z-scored board it loses to all five on $0/3$ (`results/v2_metric_confound_normboard`), and the shared-backbone native spread is $1.11\times$. So the sharpest version of this finding — that native risk–coverage on these benchmarks is beaten by a constant table — is *also* raw-scale, and we withdraw it rather than let it stand on the one scale where it holds. What remains is narrow and worth keeping: on native clinical units, abstention is mostly a contest over which score sorts channels by error magnitude, so a native-scale risk–coverage number on un-normalized IMTS should be reported next to the channel-identity lookup that bounds it.

What it leaves standing is a narrower claim than the one we set out to make, and one we can state cleanly. BandRate ranks error best of the five on the z-scored board, beating all four baselines on $3/3$ seeds with the variance in the loss's units (0.654 against 0.729–0.798); on the raw board it is nominally first but indistinguishable from the ensemble (Table 13). That is the *architecture*, not the density mechanism, because NABU loses to BandRate's own pooled head there ($3/3$ paired seeds, Appendix H.0.2) and the density input's deletion contrast is null on all four datasets (Appendix J). By our own caveat the paired p-values on this board (0.003–0.042 aligned) are not the claim either; $3/3$ seeds is.

Is a metric change that withdraws two of our own negatives convenient? It is the obvious objection — for several revisions the reporting of results against ourselves was this paper's distinguishing asset, and a metric correction now withdraws two of them — so we make the case rather than leave it implicit. **First, the correction was not chosen to rescue these results.** Per-channel standardization follows from Corollary 2 *once a loss is declared*: the corollary forces the declaration, not this particular choice, which we defend on its own terms; it was introduced to establish that floor, not to re-score this board; and it is validated by a control with no bearing on either negative — a channel-identity lookup, which must and does score 0.994 ± 0.027 on standardized errors, i.e. nothing. **Second, the**

same correction cost us elsewhere. It exposed a tie-breaking defect in our own AURC estimator and forced the retraction of the density increment as a *decomposition* (Section 4.1.7); a fix chosen for convenience does not do that. **Third, one of the two withdrawals is to *nothing*, not to its reverse**: standardized, the shift-robustness experiment separates no method from any other (Appendix H.0.1), and that is what we report. We claim no reverse on this board either — whether BandRate is the better error ranker is *not* established by three seeds on a metric we changed this revision, and we do not claim it. Both scales are reported side by side for every affected result, permanently, so that a reader can see what the correction did rather than take our word for its direction.

One further note. The 0.655 here is not directly comparable to the 0.783 in Section 4.1.7: different checkpoints (the UQ board's 3 seeds against the main board's 5) under different training configurations.

I.0.3 SCOPE: THE NEGATIVE RAW STATISTIC IS PHYSIONET-SPECIFIC

Retraining the pooled head on MIMIC-III yields $+0.58$ (95% channel-bootstrap CI $[+0.55, +0.61]$; 97% of channels positive) — there is *no* negative statistic to correct there, and correspondingly a negligible ($< 2\%$) selective-prediction gain. In fact PhysioNet is the *only* one of the four datasets whose pooled head is negative on the raw statistic: the pooled within-channel ρ_s is $+0.58$ on MIMIC-III, $+0.35$ on Activity and $+0.00$ on USHCN — positive or neutral, i.e. the sign is already right — versus -0.12 on PhysioNet. PhysioNet also has the widest per-channel rate dispersion ($22\times$ inter-decile rate ratio, vs MIMIC $8.5\times$, Activity $1.3\times$, USHCN $1.05\times$), so the sole dataset with a negative raw statistic is the most dispersed one; but the relationship is not graded (MIMIC, the second-most dispersed, is the *most* positive, and its rate coefficient-of-variation matches PhysioNet's), so the magnitude of heterogeneity alone is not the trigger. A within-PhysioNet dispersion-compression sweep (retraining the pooled head while thinning fast channels toward the median rate) supports even this only weakly and non-monotonically ($\rho_s = -0.15 \to -0.02 \to -0.07$ as dispersion falls), so we treat rate dispersion as an *associated regime indicator, not a demonstrated cause*. We had gone on to scope the diagnosis to "PhysioNet-like clinical IMTS with strongly heterogeneous channel rates"; that scope is withdrawn along with the ordering it rested on (Appendix H.0.2), since dispersion orders neither the floor nor the density increment. The honest scope is the dataset we observed it on. Note that no string and no number was retracted here — what lost its basis was a *criterion* — which is why our checker cannot catch this class and `scripts/claim_siblings.py` has to: both this sentence and the withdrawn ordering stand on `results/v2_w2_regime`.

Asking it by intervention: a dispersion sweep, and why it still cannot answer. Refusing rate dispersion above was the right call and was also not evidence against it. At $n{=}4$ the smallest attainable two-sided permutation p is $2/4! = 0.083$: the four-dataset ordering read Spearman $+0.80$ at exact $p = 0.33$, so *no* effect size could have been rejected there. More datasets do not fix that if the predictor is left wherever it happens to fall — our first attempt added three artificially-IMTSed regular benchmarks and left all three at one observation rate per channel, where the density input has nothing to read and three rows at the same x contribute nothing to an ordering (`results/v3_panel`, kept and declared superseded). What fixes it is *setting* the predictor, which a drawn mask allows.

One dataset (ETTm1 (Zhou et al., 2021), z-scored), five set dispersions, the deletion contrast at each, 10 seeds, 100 runs, everything but dispersion held fixed (`results/v3_dispersion`). Realised per-channel observation-count ratios are 1.10, 3.72, 8.53, 14.17, 21.86, spanning USHCN's 1.05 to PhysioNet's 22, with total observations flat to within 1% across levels — which took a correction: fixing the *geometric* mean of the per-channel rates let the arithmetic mean, and so the amount of data, rise $1.68\times$ from the lowest level to the highest, which would have confounded "dispersion helps" with "more data helps".

Table 15: The dispersion sweep at ten seeds. Observation count is held fixed across levels, so "dispersion helps" cannot be read off "more data helps".

set dispersion	gap mean	seed sd	seeds>0	sign-test p	paired CIs excluding 0
1.10	+0.0107	0.0465	6/10	0.754	4/10, **contradictory**
3.72	+0.0188	0.0298	7/10	0.344	2/10, agree
8.53	+0.0014	0.0312	6/10	0.754	2/10, agree
14.17	+0.0005	0.0523	5/10	1.000	6/10, **contradictory**
21.86	+0.0135	0.0537	6/10	0.754	3/10, **contradictory**

Ten seeds were chosen on the arithmetic: the exact two-sided sign test over seeds has floor $2/2^S$, which is 0.25 at three seeds — where a per-level p describes the design, not the data — and 0.002 at ten, so a small p is now *possible*. That is not power: whether an effect of a given size would be detected depends on its size against the seed spread, and we did not power the panel for a trend. None of the levels reaches it. No level reaches the $9/10$ that clears 0.05 at ten seeds (the best is $7/10$, $p = 0.344$), and the trend across levels is $\rho_s = -0.200$ at exact $p = 0.783$ against an attainable floor of 0.0167.

But this is not the negative result it resembles, and the reason is the part worth carrying forward. At three of the five levels the paired per-seed intervals exclude zero in *contradictory directions* — same setting, same data, opposite signs, both nominally significant — and tripling the seeds made that worse rather than better, from one level to three. So the contrast is set by which model was fitted, not by the dispersion: retraining at a fixed setting moves it more than moving the setting twentyfold does. The variance is in the fitting, not in the sampling: the effect is heterogeneous across training runs. More seeds would narrow the uncertainty of the *mean* effect, but that mean averages runs that disagree in sign, which is a different question from whether dispersion triggers the effect.

We therefore claim nothing about whether dispersion is a trigger. What we can say is sharper than the "we do not know" it replaces, and it is about the measurement: at the panel sizes a study like this can afford, the retrain contrast resolves the question poorly, and an estimand with a smaller seed footprint is the cheaper route than a much larger panel. Note also what this reproduces. The same signature — seed spread swamping the effect, per-seed bootstraps excluding zero in opposite directions — is what we report for PhysioNet's own deletion contrast (Appendix J), where the seed standard deviation is $43\times$ the mean. Seeing it again under a design where every nuisance was held fixed says it is a property of this measurement rather than a peculiarity of that dataset. The reading above was fixed in `experiments/dispersion_sweep_prereg.md` before the board was scored, with all three possible verdicts written out, because three earlier versions of this paper asserted a relationship between two four-point vectors and each was found after the fact.

A lower-variance estimand, and the same question asked again on held-out data. We read the sweep's failure as a property of the contrast rather than of the panel, so we measured what each candidate estimand could resolve before choosing one — on the hundred checkpoints already trained, since the variance that matters is in the fitting and the fitting was already on disk. Resampling each candidate's observed per-seed spread and running the same exact sign test gives the smallest effect it would catch at ten seeds with 80% power (`results/v3_dispersion/estimand_mde.csv`): 0.062 for the retrain contrast against observed effects of 0.002–0.021, 0.043 for averaging σ^2 over seeds within each arm, and 0.022 for an *eval-only* contrast — one trained model, its density input permuted within channel at inference, so fitting variance cancels exactly. Only the third can see its own effects.

On the sweep's own dataset the eval-only contrast then came out monotone in dispersion, $+0.034$ through $+0.113$. That is a post-hoc ordering found while measuring something else, so we fixed the reading in `experiments/eval_only_monotonicity_prereg.md` — statistic, exact permutation test, a per-level admissibility rule, a four-way decision, and a prohibition on pooling — and tested it on two datasets that had no part in forming it, ETTm2 (Zhou et al., 2021) and Weather (Wu et al., 2021), three seeds each (`results/v3_evalonly_heldout`).

Table 16: The pre-registered eval-only monotonicity test on two held-out datasets, ETTm2 and Weather, which had no part in forming the hypothesis (`results/v3_evalonly_heldout`, three seeds each).

dataset	per-level median eval-only contrast 1.10	3.72	8.53	14.17	21.86	ρ_s	exact p
ETTm2	$+0.019$	$+0.083$	$+0.138$	$+0.175$	$+0.169$	$+0.900$	0.083
Weather	$+0.082$	$+0.059$	$+0.055$	$+0.068$	$+0.040$	-0.700	0.233

Not confirmed, and the trend test had little power to confirm it. Every per-level median cleared its own detectable effect, which ran 0.0003 to 0.029 against observed medians of 0.040 to 0.175, so each *level's* effect is resolved. That says nothing about the power of the trend test across levels. With five levels the exact Spearman test rejects at 0.05 only a perfect ordering ($|\rho_s| = 1$, $p = 2/5! = 0.0167$); ETTm2's $\rho_s = +0.9$ already gives $p = 0.083$. We did not power the design for a trend. The two held-out datasets point in *opposite* directions. ETTm2 agreeing in sign with ETTm1 is not replication: they are two transformers of one ETT collection, gathered together over the same two years (Zhou et al., 2021), which is why the pre-registration forbade pooling them into a ten-point ordering, and Weather — an unrelated source — runs the other way. The ETTm1 ordering therefore stays an unreplicated post-hoc observation and we quote it as evidence nowhere.

What the exercise does establish is about the instrument rather than the mechanism, and it is worth separating. The eval-only contrast resolved an effect at 10 of 10 levels here, where the retrain contrast could do so at 1 of the 5 rows of our own deletion table. Changing the estimand was right; the hypothesis it was changed in order to test is not supported. And the two measure different things — whether a fitted model *uses* density, against whether having it available improves what can be *learned* — so no reading of either transfers to the other, and both datasets here are artificially masked regular series.

The low-variance estimand does not substitute for the high-variance one, and the gap between them is itself the finding. It would be convenient to answer this paper's question with the eval-only contrast, since that is the one with power. It cannot be done, and the reason is worth stating because the two quantities are easy to conflate. The retrain contrast asks what having the density input *available during fitting* buys; the eval-only contrast asks how much a model that was given it *leans on the point-level signal at inference*. A model can lean on density heavily and still lose nothing by never having had it — if a density-free model learns an equally good substitute from the value context — in which case the second is large and the first is zero. Nothing rules that out a priori, and it is not a hypothetical here: both were computed on the *same* checkpoints and the *same* test points.

Table 17: The two quantities a deletion contrast confounds, computed on the *same* checkpoints and the *same* test points: what a fitted model uses, and what having the input was worth to learn.

set dispersion	1.10	3.72	8.53	14.17	21.86	median
eval-only (uses it)	$+0.034$	$+0.054$	$+0.074$	$+0.102$	$+0.113$	$+0.074$
retrain (gains from it)	$+0.002$	$+0.021$	$+0.006$	-0.005	$+0.002$	$+0.002$

Eval-only exceeds retrain at $5/5$ levels and by $39\times$ at the median. So the models do read the density input — that is resolved at every level, well clear of its detectable effect — and removing it from training costs little that this contrast can detect. That is evidence about the models' *input dependence*, not an established claim that the information is substitutable: the retrain contrast is underpowered (Appendix J), and permuting density at inference also breaks its joint structure with time and values. It is consistent with Section 4.1.7, where NABU's $+0.51$ advantage over an S-only lookup turns out to be almost all value context while the density input's own contrast does not resolve on that dataset (Appendix J).

Two consequences we hold ourselves to. A large eval-only contrast is *not* evidence that the sampling pathway is valuable, so it cannot be quoted in support of the density mechanism — and we do not quote it so. And the question the title asks is the retrain question, whose detectable effect is 0.062 against observed effects of 0.002–0.021: the estimand we can measure is not the estimand we need, and buying power by changing the question would be the more comfortable error rather than a smaller one.

The inversion is a scale artefact: both arms, same day, `normalize` the only difference. The comparison above is not controlled, and we should have noticed.

`scripts/uq_baseline_board.sh` keyed per-variable z-scoring off the dataset *name* — `normalize=true` for `mimic3`, absent for everything else — and the checkpoints confirm it. So the two clinical boards differ in preprocessing as well as in dataset, and PhysioNet's raw per-channel value variance spans 6.45×10^7 (std 0.19 on one channel against 1526 on another; `results/v2_bound/channel_scale.csv`). Our appendix adopted normalization for MIMIC precisely because raw targets there were pathological, and verified that it *preserves* the fix — NABU stays at $+0.59$. It never asked whether normalization preserves the *problem*.

It does not. Retraining both arms on PhysioNet, same day and same code, with `dataset.normalize` the only difference (`results/v2_w1_gate_norm`, 3 seeds; `PhysioNet2012` had no such parameter until now, which is why this was never run):

Table 18: Both arms retrained on PhysioNet the same day with the same code and `dataset.normalize` the only difference — the controlled version of the comparison that suggested MIMIC had failed to replicate.

scale	head	$\rho_s(\hat{\sigma}^2, 1/\rho_n)$	$\rho_s(e^2, 1/\rho_n)$	$\rho_s(\hat{\sigma}^2, e^2)$
raw	pooled	-0.120 ± 0.104	$+0.147 \pm 0.033$	$+0.111 \pm 0.169$
raw	NABU	$+0.662 \pm 0.098$	$+0.070 \pm 0.060$	$+0.188 \pm 0.077$
z-scored	pooled	$\mathbf{+0.583 \pm 0.092}$	$+0.153 \pm 0.020$	$\mathbf{+0.275 \pm 0.006}$
z-scored	NABU	$+0.690 \pm 0.065$	$+0.143 \pm 0.037$	$+0.257 \pm 0.023$

The raw arm reproduces the pre-registered -0.154, so it is the normalization that moves the number, and $37/37$ channels are positive after it against $6/37$ before (cluster CIs $[-0.249, -0.057]$ and $[+0.541, +0.651]$). Normalized PhysioNet sits at $+0.598$ on the `dvcorr_ci` aggregation and MIMIC at $+0.58$: the same value. **The sign inversion this paper is built around is a property of raw per-channel scale, and MIMIC's non-replication — which fired our own W2 kill criterion — was a preprocessing difference, not a dataset difference.**

Three qualifications, two of them against us. The density signal itself is untouched: $\rho_s(e^2, 1/\rho_n)$ is $+0.147$ raw and $+0.153$ z-scored, so sparse regions genuinely do carry more error and it is the *diagnosis* that is scale-specific, not the premise. On the statistic that matters — whether $\hat{\sigma}^2$ tracks realized error — normalizing buys more than NABU does, and after normalizing NABU no longer helps ($+0.275$ against $+0.257$, pooled winning $3/3$ paired seeds). And the two arms are not equally good forecasters: the raw arm is 8–10$\times$ worse by per-channel median MSE on a common scale, with point-mean MSE whose seed standard deviation equals its mean, so we cannot separate "the inversion is caused by scale" from "the inversion is a symptom of badly under-fitting the small-scale channels, which raw scale induces". We state the intervention and decline the mechanism.

What we do *not* claim is that channel-scale heterogeneity is the general trigger. Across the four datasets it orders the statistic at Spearman -0.800, $p = 0.200$ (`results/v2_bound/channel_scale.csv`) — the same strength we refused for rate dispersion two paragraphs ago, and non-graded in the same way (Activity at $22\times$ is more positive than USHCN at $1.2\times$). One within-dataset intervention is what we have.

I.0.4 Is it *Nyquist*, or just a per-band constant?

Our pre-registration listed a shuffled-$f^{\max}$ arm and we had not run it; we have now. Permuting the per-band Nyquist constants among the bands—same architecture, same set of constants, only the assignment scrambled—retains most of the effect: horizon-controlled $\rho_s = +0.444$ (3 seeds) against $+0.529$ for the true assignment, i.e. roughly 84%. So what the head needs is a *per-band offset on log density*, and the specific Nyquist values contribute little beyond that. Under the criterion we froze in advance this is a failure of the Nyquist-specific claim, and we report it as such: the honest description of the mechanism is *per-band scale-conditioned density*, the name "Nyquist-aware" is aspirational rather than earned, and the sampling-theoretic story in Section 3.2 should be read only as the motivation that led us to a per-band normalization—not as an explanation of why it works.

Is the per-band normalization load-bearing? Ablating the ratio — feeding the head $\log(1+\rho_n)$ and the band embedding but *not* $\log r_{n,k}$ — collapses the PhysioNet correlation from $+0.65$ back to -0.06 ± 0.19 (pooled-head behaviour) and, with the shape gone, widens the calibrated intervals

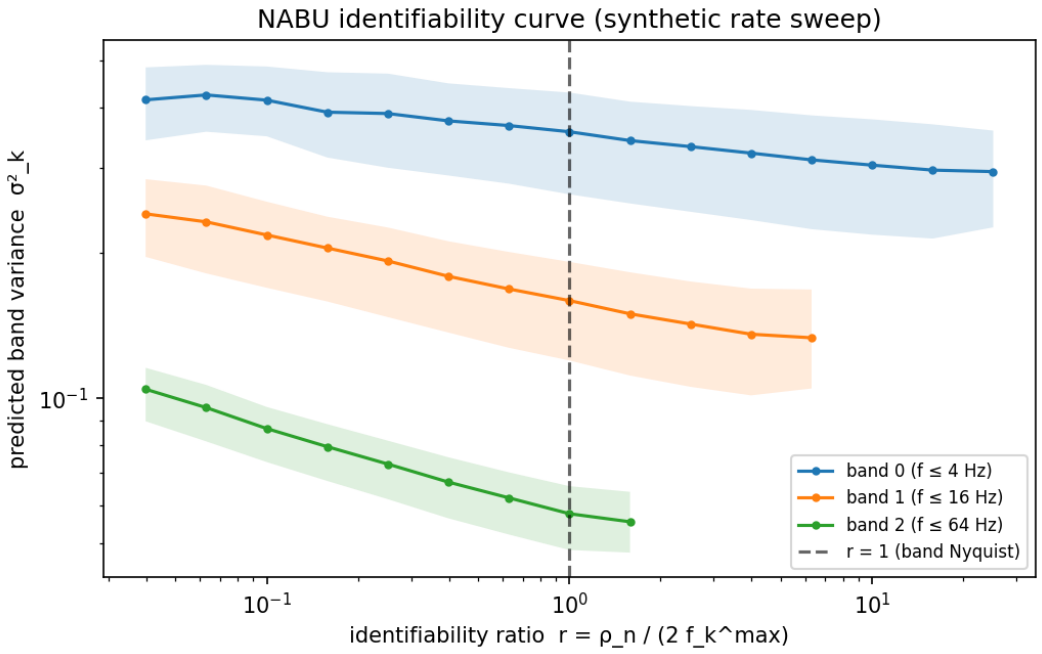


Figure 2: NABU identifiability curves (synthetic rate sweep). Predicted band variance σ_k^2 against the identifiability ratio $r = \rho_n / 2f_k^{\max}$; dashed line is drawn at r=1; note that under our own formula the implied threshold is $r^\star = |B_k|/2f_k^{\max} = 0.375$ for all three deployed bands (0.75 under the two-sided convention), so the line is a convention, not the boundary the motivation implies — another reason we do not read a regime change off this figure. Variance is flat in the identifiable regime and inflates below the boundary for every band.

29% (71.4 → 92.0) at a 33% worse interval score (217 → 289); on near-regular Activity it is neutral (+0.32 either way). So the identifiability ratio is empirically necessary exactly where per-channel rates are heterogeneous, as the rate-normalization motivation suggests — and although $\log r_{n,k} = \log \rho_n - \log 2f_k^{\max}$ is in principle recoverable from $[\log \rho_n, e_k]$, the network does not learn the per-band density scaling without it. On USHCN it is near zero (+0.03): with only five homogeneous-rate climate channels there is little per-channel density variation for the head to exploit. A controlled mask-thinning sweep on the trained model confirms this is the *absence of a signal to track*, not a failed head — injecting sampling irregularity (Bernoulli thinning of the observation mask) raises the within-channel correlation monotonically from +0.03 to +0.43 as density heterogeneity grows (ρ coefficient-of-variation 1.84→2.41), so NABU activates precisely when density varies.

J THE DELETION CONTRAST ON A DECISION QUANTITY

Section 4.1.7 compares NABU against a lookup on (channel, horizon-decile). That comparison is informative about whether our head beats the best score obtainable from the sampling set alone, and uninformative about the density pathway, because the lookup reads S while NABU reads S and V. The contrast that isolates the pathway holds the architecture fixed and deletes the density input, so both arms read V and only the input under test differs. It has been run as a correlation (Appendix E.0.1) and is run here on the same risk–coverage metric as the floor, per-channel standardized, with a sequence-level cluster bootstrap.

Withdrawn here. An earlier version quoted NABU's per-seed *share* of the density increment as "about 6%". We withdraw it: that ratio's denominator nearly vanishes and it runs from −1411% to +1098% across seeds, so no share is reportable, and the body no longer quotes one.

Table 19: The density-deletion contrast: normalised AURC on per-channel standardised e^2 with and without the density input, otherwise identical heads, 3 seeds each; gap is without minus with (positive: density helps), mean $\pm$ seed sd. The z-scored PhysioNet row was run for this revision.

dataset	AURC with density	without	gap	seeds with gap > 0
Activity	0.5167	0.5459	$\mathbf{+0.0292 \pm 0.0084}$	$3/3$
Activity, z-scored	0.5078	0.5061	$\mathbf{-0.0017 \pm 0.0146}$	$1/3$
PhysioNet	0.8604	0.8633	$+0.0029 \pm 0.1257$	$2/3$
PhysioNet, z-scored	0.7111	0.7345	$+0.0234 \pm 0.0373$	$2/3$
MIMIC-III	0.5305	0.5273	-0.0032 ± 0.0070	$1/3$
USHCN	0.7636	0.7613	-0.0023 ± 0.0280	$2/3$

Positive means deleting density made ranking *worse*, i.e. the input helps. On Activity it does, on $3/3$ seeds, and each seed's own sequence bootstrap excludes zero ($[+0.008, +0.036]$, $[+0.010, +0.043]$, $[+0.029, +0.053]$) — a statement about the data rather than about retraining, which matters because at n=3 an exact two-sided sign test over seeds cannot go below $p = 0.25$. **Those three intervals are not usable and we presented them as if they were.** That "sequence bootstrap" resamples Activity's 607 test *chunks*, which come from three recordings (A02, C03, C04) and two subjects; resampling roughly two hundred dependent units per independent one makes an interval as narrow as one likes. Three recordings cannot support a reliable population interval, so the $+0.0292$ is a point difference on three recordings and nothing more (Appendix B). On MIMIC and USHCN the gap is indistinguishable from zero in both directions.

PhysioNet, z-scored, run for this revision. The mean gap is $+0.0234 \pm 0.0373$: two seeds show a gain whose own sequence-bootstrap intervals exclude zero ($[+0.017, +0.083]$, $[+0.027, +0.062]$) and the third nearly excludes zero the other way ($[-0.042, +0.002]$). Those intervals are conditional on each fitted model; they say two of three trained models gain from the input on these patients, not that the training algorithm does on average, which three seeds cannot establish. It is neither evidence that density does not help on the z-scored scale nor evidence that it reliably does.

And the Activity result does not survive the preprocessing the baselines use. Appendix H.0.2 found PhysioNet's headline diagnosis to be an artefact of raw per-channel scale, so we asked the same of our one positive contrast rather than leaving it the benefit of the doubt (`HumanActivity` had no `normalize` parameter either; it does now). Retrained with `dataset.normalize=true`, same budget, the contrast carried by `model.var_use_density` alone, the gap is -0.0017 ± 0.0146 on $1/3$ seeds, and the per-seed intervals *contradict* each other — $[-0.023, -0.005]$ excludes zero on the negative side while the other two include it. **After normalization the density input is not shown to help on any of the four datasets** — which is weaker than "does not help", and the difference is the subject of the next paragraph. Three independent routes now point the same way where before they disagreed: the z-scored ceiling's density increment is $+0.0001$ of 0.0020, interval $[-0.0013, +0.0015]$ (Appendix P), and this contrast is null everywhere. Two caveats. The raw and z-scored arms of a dataset are not guaranteed to be equally good forecasters, so a vanishing contrast does not on its own separate "density stops helping" from "the model changed" (Appendix H.0.2).

And four of these five rows could not have shown an effect if there were one. We only asked this after the dispersion sweep made it unavoidable (Appendix I), and it applies to the table above. For each row, resampling that dataset's own per-seed gaps gives the smallest true effect at which the criterion we actually report — all seeds agreeing in sign — would be met 80% of the time. At three seeds no significance was attainable in either direction, the exact sign test's floor being $2/2^3 = 0.25$, so this is the power of the weaker criterion rather than of a test:

Table 20: Minimum detectable effects under the all-seeds-agree criterion the paper actually reports. At three seeds the exact sign test's floor is $2/2^3 = 0.25$, so no significance was attainable in either direction and this is the power of the weaker criterion. ‡ETTm1 is the synthetic-mask arm; four drafts printed seven of the board's then eight rows and omitted it, while the text counted all eight. The PhysioNet z-scored row was run for this revision.

dataset	seeds	median gap	seed sd	detectable effect	criterion met
Activity, raw	3	$+0.0270$	0.0084	$\mathbf{0.0047}$	yes
Activity, z-scored	3	-0.0048	0.0146	0.0097	no, $2.0\times$ short
MIMIC-III, z-scored	3	-0.0013	0.0070	0.0096	no, $7.5\times$ short
MIMIC-III, raw	3	-0.0017	0.0022	0.0026	no, $1.5\times$ short†
PhysioNet, raw	3	$+0.0468$	0.1257	0.1856	no, $4.0\times$ short
PhysioNet, z-scored	3	$+0.0448$	0.0373	0.0645	no, $1.4\times$ short
USHCN, raw	3	$+0.0061$	0.0280	0.0396	no, $6.5\times$ short
USHCN, z-scored	3	-0.0048	0.0395	0.0216	no, $4.5\times$ short
ETTm1, raw‡	3	-0.0367	0.0718	0.0797	no, $2.2\times$ short

Two limits on this table before reading it. The detectable effects are post-hoc, estimated by resampling three observed seed differences, so they are themselves uncertain; and "all seeds agree in sign" is the paper's reporting criterion, not a 5% test. With those understood: eight of nine rows are *underpowered, not negative*: the effect would have had to be 1.4 to 7.5 times the observed size for our own criterion to register it, and on PhysioNet $4\times$ raw, where the seed spread is 0.126 against a median gap of 0.047. With USHCN z-scored, MIMIC raw and now PhysioNet z-scored added, each of the four datasets has been run on both scales; the ninth row is ETTm1, a synthetic-mask dataset rather than a second scale. Until this revision **the PhysioNet z-scored arm had never been run** and the table printed seven rows, while four drafts said every dataset existed on both scales; we retract "every dataset now exists on both scales" as it was printed then. The new arm repeats the raw arm's signature: its per-seed intervals are $[+0.017, +0.083]$, $[-0.042, +0.002]$ and $[+0.027, +0.062]$, two excluding zero in one direction and the third nearly excluding it in the other. No scale rescues any of them. †MIMIC raw needs its own note: it is the arm limitation (i) records as having diverged on one seed, and here all three seeds evaluated cleanly — but both arms sit at normalized AURC 1.00 (1.0036 with density, 1.0016 without), which is the value of *no ranking at all*. Its small seed spread (0.0022, the lowest of the nine) and correspondingly low detectable effect are absence of signal, not stability of it, the same signature the raw PhysioNet UQ board showed (Appendix N.6); the row is not evidence of power. **Every "indistinguishable from zero" in this section should be read as "this contrast could not have distinguished it"**, and we say it here rather than leaving a reader to derive it. Activity is the one row whose seed spread is small enough to meet the criterion. That spread is a genuine property of retraining — the number of chunks does not change how much three retrained models disagree on a fixed test set, and an earlier version of this paragraph wrongly attributed it to the chunking. What three recordings limit is something else: generalisation to new recordings, which neither the seed spread nor a chunk-level bootstrap measures (Appendix B). So the row meets our criterion *for the recordings it has*, and licenses nothing beyond them. The retrain contrast is simply a low-powered estimand — on the sweep's own dataset at ten seeds its detectable effect is 0.062 against observed effects of 0.002 to 0.021 — which is why Appendix I moves to an eval-only contrast whose fitting variance is zero by construction (`results/v3_dispersion/estimand_mde.csv`). **Set this beside Section 4.1.7: NABU beats the S-only lookup by $+0.51$ on Activity and $+0.41$ on MIMIC, while the density input itself is worth $+0.029$ and -0.003 — the Activity figure a point difference on three recordings (Appendix B).** Almost all of that advantage is value context, which is the arithmetic form of why the lookup comparison identifies nothing about the sampling pathway.

Two instruments that share no machinery: no agreement between them. Beside the risk–coverage density increment of Section 4.1.7 — an architectural deletion against a contrast between two exhibited S-only scores, with no estimator, metric or board in common:

Table 21: Two instruments with no estimator, metric or board in common — a risk–coverage contrast between two exhibited S-only scores, and an architectural deletion. The datasets on which each one resolves anything do not overlap.

dataset	AURC density increment	deletion arm
Activity	not establishable	$+0.0292$; the per-seed CIs resample chunks, not recordings
MIMIC-III	not establishable	-0.0032
USHCN	not establishable	-0.0023
PhysioNet	establishable	$+0.003 \pm 0.126$, seeds contradict

Four drafts called this agreement. It is not agreement; it is two instruments that never resolve the same dataset. On the AURC side only PhysioNet resolves, and there the deletion arm has

none ($+0.003 \pm 0.126$, its three seeds contradicting one another). On the deletion side only Activity resolves, and its per-seed intervals resample 607 chunks drawn from three recordings, so on the unit the data supplies it does not resolve either. The other two rows are a null agreeing with a null — and this section has just said that eight of nine rows are *underpowered, not negative*. Two instruments that both lack the power to detect an effect corroborate nothing and cannot validate each other retroactively. We retract "agree wherever both exist" and the "Three for three where the deletion arm resolves anything" that followed it. The honest reading of this table is that we have no cross-instrument evidence either way, and that the AURC increment stands or falls on its own.

PhysioNet was refused for a reason that turned out not to exist, and the answer is a loud null. `scripts/deletion_arm_aurc.py` diffs the two checkpoints' configurations and declines any pair differing in more than the density flag. It declined PhysioNet on `train.patience` — present in one board's config snapshots, absent from the other's — and an audit then found that key *inert*: it is in no configuration file, no line of the training code reads it, and it was a stored override that nothing consumed, so no run ever early-stopped. A key that changes no computation cannot make a contrast uncontrolled. With it allowed the pair passes, and we report it.

The result is $+0.0029 \pm 0.1257$, and the standard deviation is $43\times$ the mean. The three per-seed gaps are -0.139, $+0.101$ and $+0.047$ — and *each seed's own sequence bootstrap excludes zero*, in contradictory directions. That is worth more than the point estimate: a within-seed interval on this contrast is confidently wrong about the effect, which is the same lesson as the paper's other variance-component corrections arriving at the level of a single experiment. Two caveats stay: the arms come from different boards, so they share a configuration but not a training run, and the with-density arm has five seeds against three, which we pair in order. The correlation-domain row the body quotes (-0.075 ± 0.049, $p = 0.25$) is unpaired for the seed-count reason alone, not for the patience reason we previously gave.

K THE RETRACTION LEDGER, AND THREE GATES THAT COULD NOT FAIL

K.1 THREE WITHDRAWALS FROM ONE CAUSE, AND WHAT IT COST US

Every board scored on native-scale AURC is affected, including the ones in our favour, so **start with those**: the largest gain we report, selective prediction against the pooled head, is $+65\pm28\%$ on $5/5$ seeds at $p = 0.008$ (paired t) natively and $+12.0\pm11.1\%$ loss-aligned (earlier $+11.3\pm19.1\%$ at $p = 0.269$) on $4/5$ at $p = 0.074$ — not significant, and below the pre-registered $\geq 20\%$ gate the native number passed. **Then the three withdrawn falsifications.** *(1)* We predicted a geometry-grounded variance would degrade *less* than an ensemble under a sampling-rate shift and reported for three rounds that it degrades *most*; standardized, the experiment separates no method from any other, and we withdraw the falsification rather than replace it with its reverse (Appendix H.0.1). *(2)* We reported for four rounds that a deep ensemble ranks error better than NABU on both clinical datasets; standardized it is false on both, and we claim no reverse either (Appendix I). *(3)* We read the per-dataset ceilings as an applicability test, withdrew that when a uniform metric reversed the ordering, then published the reversal — on a metric that was itself native-scale AURC. Three answers at n=4; we retire the question (Appendix H.0.2). **Two further withdrawals the body no longer carries.** We retract "the bound is tight" — the reading that the oracle comes within 3% of the ceiling rested on a mean of per-cell correlations paired against the root of a ratio of weighted means, and once both sides are aggregated alike the oracle sits below the ceiling and the bound is consistent, not attained (Appendix L). And we retract having said that s can be computed before training anything: e^2 is a trained model's realized error, so s is a property of the (data, model, preprocessing) triple.

Why this is a correction and not a harvest. Per-channel standardization is forced by Corollary 2, was introduced to establish the floor rather than to re-score these boards, and is validated by a control with no bearing on any of them; and one withdrawal goes to *the experiment separates nothing*, not to its reverse. Both scales are reported throughout (Appendix M).

The ledger. Twenty-six claims are ledgered across five revisions, twenty-three of them retracted in print. `experiments/retraction_ledger.yaml` carries one row each: the claim, the reason, a regular expression that *asserts* it, and the contexts in which mentioning it is legitimate rather than assertive. `scripts/check_retractions.py` reads that file and scans the paper, every script, the released tool, the tests and every metadata file under `results/`; a match outside a

legitimate context fails the build, and `scripts/validate_results.py` runs it as a gate beside the numerical ones. It also checks the converse — that each withdrawal is *stated* somewhere, so a claim cannot leave by being quietly deleted. Three rows are marked as removed during revision rather than retracted in text, on the principle that a retraction note is owed to readers of a released version and no released version carried them; the row stays either way, so the claim cannot return unnoticed.

We built this because the same failure recurred: a retraction reaches the body prose and survives elsewhere. Round 4, an appendix preserved three claims the body had retracted, and duplicate labels made body cross-references resolve silently into appendices. Round 5, a withdrawal did not reach the section that used the withdrawn result as evidence, and the cascade into Figure 1 was not followed. For a paper whose asset is disclosure, an unpropagated retraction is worse than an uncorrected error: it shows the disclosure machinery does not reach the whole document.

The table checker went blind to a row while reporting "ok", during this revision. Marking Table 1's Activity row with a dagger changed its label from `Activity` to `Activity$^\ddagger$`, the row stopped matching the declared dataset names, and `check_paper_tables.py` went on passing — on 24 cells where it had compared 32. The per-table cell tally we added after an earlier round is what exposed it, and it exposed it only because a unit test asserts the row count independently; the tally alone would have shown a smaller number with nothing to compare it against. A footnote marker is an annotation, not a rename, and the checker now strips superscript markers before matching. Chasing that turned up a second silence in the same table: the spec declared columns one to four, so the *prize* column — and with it the number this paper quotes most, the floor of 0.286 (then phrased as a share of "marginal risk", which it is not) — had never been compared to anything. It is not a board row but a derived one, $1 - \mathrm{hi}$ of the floor interval, and being derived is why it was left undeclared rather than declared unverifiable. It is now checked, including that a dataset whose floor interval crosses 1.0 prints no prize at all. A third, found the same day and worse than either: the minimum-detectable-effect table printed *seven* of its board's eight rows — the ETTm1 arm absent from the table while the text beside it counted all eight — and no gate could have seen it, because a checker that walks printed rows never reaches a row nobody printed. That table is now declared and checked in both directions: every board row must appear, and every printed row must resolve. Declaring the ablation table the same afternoon — it had acquired a label, which made it visible to the gate and checked by nothing — immediately turned up a $K{=}2$ correlation printed as $+0.69$ where its board says 0.685, i.e. $+0.68$: wrong in every revision that printed it. Declared coverage went from thirteen tables to seventeen and from 323 compared cells to 461, and in this revision to twenty-three tables and 880 cells; the remaining fourteen displays are hand-checked only, and the three failures in this paragraph are what that is worth. *A gate that can go quiet when the paper changes is a gate that has to be tested against a planted failure*, which is the same lesson as the retraction checker's fixtures.

The variance-component sweep, and what it cost. Nearly every $\pm$ in earlier drafts was a standard deviation across model seeds, and the seeds score the *same* test points, so it measures how much retraining moves a number rather than how much a different sample of patients would. Every affected quantity was re-examined and the corrections were not uniform, which is the argument for having done it rather than reasoned about it. *The AURC board*: understated by up to an order of magnitude, and it cost two claims — MIMIC's density increment ($\pm.012$ across seeds, $[-.074, +.150]$ across resamples) and USHCN's "NABU clearly worse, $0/5$". *Table 23*: understated $2\times$, and the verdict survived and strengthened. *The ceiling intervals*: three constructions rejected for three different reasons, and no pooled interval published. *Table 6*: the seed sds are removed, since a number labelled wrong in its own caption is still read as a number. *Table 8*: the seed sds are kept, because the p column is computed from them, and annotated — there the direction is safe, since an understated standard error makes "not distinguishable from zero" conservative rather than generous. Two of these moved a headline and two did not, and the ones that did were not the ones we would have guessed.

Four copies of one function, and the fourth had drifted. A late audit found three defects in the selective-prediction board, all in the same file and none a reasoning error: it averaged risk–coverage over $f \in [0, 0.8]$ by trapezoid where the convention stated in Section 4.1.7 and used by every other board is a rectangular mean over $[0, 0.9]$; it resolved score ties by array position, although this appendix records that the *baseline* of that very comparison collapses to a constant σ^2 on one seed, which makes the entire ordering a tie; and it standardized each channel by that channel's mean over all points rather than on a sequence-disjoint half. Each is a defect the paper had already found and fixed elsewhere. What made them survive is that the risk–coverage code existed in four copies, and this one stopped being updated with its siblings.

The repair is one implementation (`scripts/aurc_common.py`) that the other three now delegate to, verified behaviour-preserving — the four-dataset decomposition reproduces to every printed digit — with a test that fails if any script re-derives the loop instead of delegating, and property tests for the two behaviours that separate the correct version from the drifted one: a constant score must score as no ranking, and the standardizer must honour its fit mask. *Neither our provenance gates nor the retraction ledger could have caught this*: no number was wrong on its own board and no retracted claim was restated. What was wrong was a convention, stated once in the paper and implemented four times.

The order in which a claim is withdrawn. We had been fixing the prose first and writing the ledger entry afterwards, which can only find what we already looked at: the pattern gets authored against the sentence just rewritten rather than against the claim as it stood, and that is how a withdrawal reached the abstract and an appendix while a live assertion of it survived in the section using it as evidence. `scripts/retract.py` inverts it — the entry is written from the claim as it stands, the checker then reports every occurrence because nothing is whitelisted yet, and the sites that remain after editing are ratified last.

Three gates that reported passes they had no coverage to support. The instruments failed the same way the estimators did, and the failure is the same shape as $\hat{s}$ reporting a number it had no resolution to support before we gave it a permutation null.

Table 22: Three of our own gates that reported passes they had no coverage to support. The instruments failed the same way the estimators did.

gate	why it could not fail
board validator	globbed `results/*/results.csv`, so a registered board missing that file was *invisible* rather than skipped; reported "0 errors" while four boards, including Table 28's, were unchecked
retracted-claim scan	matched line by line over LaTeX hard-wrapped near 110 characters, so any claim spanning a line break was unmatchable; reported "no retracted claim asserted" while one was live in the body, being used as evidence
page-limit check	printed the target page count whenever the bibliography heading was *absent* from the page it probed — including when the body had overrun onto later pages; reported compliance at 13 pages against a 9-page limit, in five commit messages

A fourth, in an estimator rather than a gate, found the same way. A lookup-table score is constant inside its group, so the whole within-group order is a tie, and `argsort` resolves ties in array order — which, for a loader that emits points grouped by sequence with correlated errors, credits the lookup with an ordering it never supplied. Permuting the *channel labels*, which destroys every real signal, still read 0.21 of the achievable benefit; that is what a null is for. Every AURC in this paper now breaks ties at random, averaged over five draws. It moved the native column materially (channel identity $0.049 \to 0.058$, omniscient $0.030 \to 0.031$) and the standardized column almost not at all (the floor moves $0.589 \to 0.611$ only because the lookups were re-fit out of sample, the increment $0.207 \to 0.208$), and the pipeline check it was found with now reads 1.0002 where it read 1.0035, and 0.994 ± 0.027 once the lookups were also re-fit out of sample. We note it because the same estimator produced every AURC we have published, including the ones in our favour.

Where the checking happens in the loop. Every error in the family this appendix describes was computed, written into the paper and committed before anyone checked it. The reviewers were not finding things we could not have found; they were finding them at the only point in the loop where anyone was looking. So a checkpoint now sits between *having a corrected number* and *writing a sentence about it*, with three questions, all answerable in a minute from the code just run and none answerable afterwards, because by then one is checking prose against belief rather than code against claim. *(i) For each side of this comparison, what is its aggregation, transform, denominator and null, and do they match?* — which is all four of the aggregation errors. *(ii) What does each side read?* — which is the V-confound, and would have fired three rounds ago on PhysioNet, when the answer was zero and looked like confirmation. *(iii) If this is one-sided, which estimator errs which way, and did I measure that or assume it?* — which is the in-sample floor.

Two one-sided claims we cannot currently support, listed rather than left to be found. `experiments/direction_registry.yaml` records, for every one-sided claim in the paper, what measures the direction it depends on; four have a fixture and two do not. That each out-of-fold R^2 lower-bounds its own η^2 is true of the *population* R^2 and not automatically of the out-of-fold estimate, whose fold split interacts with sequence structure exactly as the lookup's fit split did — no headline rests on it, since we already decline the stronger reading, but each column's single-sided reading does. And identification in the deletion contrast is an argument about what the two arms read rather than a measured bias; it is machine-checked by a configuration diff, which is weaker than a fixture and stronger than the prose it replaced.

A one-sided claim needs its direction measured, not declared. The floor of Section 4.1.7 is one-sided only if every estimator feeding it errs the safe way, and that argument lived in our prose while nothing checked the code implemented it — it did not, because the lookup was fit on the points it scored. The obvious guard is a declared bias sign per estimator, and it would have failed: whoever declared the lookup conservative would have been whoever believed it was exhibited. So the sign is *measured* on planted truth (`scripts/direction_fixtures.py`): plant cell means, and the in-sample lookup's AURC falls below the population oracle's on 100% of draws — an S-only score cannot beat the oracle in population, so $1 - \mathrm{AURC}$ overstated the floor — while the out-of-sample fit falls below on 0%. The fixture reproduces the defect from a direction it was not designed to look for, which is the test of whether such an instrument is worth building. Two of our estimators already had this and it is why their one-sided readings were sound: the synthetic planted-share validation of $\hat{s}$, and the within-cell permutation nulls, which measure each estimator at truth $= 0$. *Bias direction and variance are different axes*, and the cluster bootstrap showed why both are needed: the floor moving from 0.389 to 0.286, and MIMIC becoming unestablishable, is not a sign failure — every estimator could have erred conservatively and that would still have happened.

In all three, absence of evidence was coded as evidence of absence: the check looked for a failure signal, did not find it, and reported success without ever establishing that it could have found one. The rule we now apply — and the reason this appendix exists rather than a sentence — is that **a check may report success only by exhibiting the positive evidence it was looking for**, and that **every gate carries a planted-failure self-test that runs with it**: a fixture containing each ledgered claim, wrapped across a line break, inside the retraction appendix, in a log and in the manifest, all of which the scanner must find; and a deliberately overlong body the page check must call a violation. This is the permutation null moved from `numpy` to `pytest`. *An instrument that cannot fail is not evidence, whether it is an estimator or a test.*

What else inherits the preprocessing caveat. Appendix H.0.2 shows the sign inversion is a property of raw per-channel scale and Appendix P that $\sqrt{s_T}$ falls from 0.123 to 0.045 when values are z-scored. The consequence has to be followed rather than left at those two sections, which is what `scripts/claim_siblings.py` is for: it inverts `results/MANIFEST.json` from claim-to-board into board-to-claims, because a withdrawal's consequences can quote none of its wording and a string scan cannot find them. Run on this revision it reports that `results/repro` alone backs *nine* claims — `C5-ceiling-4dataset`, `C5-total-share`, `C6-aurc-decomposition`, `C6-debiased-hierarchy`, `C6-per-cell-ceiling`, `C6-share-uncertainty`, `C7-ceiling-interval`, `C7-confound-index` and `E1-main-accuracy` — and that board carries the same asymmetry as the UQ boards: PhysioNet, Activity and USHCN raw, MIMIC z-scored. So the caveat is not local to two sections. Every one of those nine is measured with three of four datasets on native scales, and on the one dataset where we have now run both, the scale moved the headline quantity threefold. We report them as measured and mark the scale rather than re-running the paper, but a reader should treat "on four datasets" throughout as "on these four datasets as we prepared them".

L THE ORACLE TEST: THREE WRONG VERSIONS, AND EVERY READING

Summarized in Section 4.1.3; the record is here, because the pattern of the errors is more useful than any one corrected number.

Table 23: Every ceiling in this paper with the aggregation that defines it, and the held-out lookup tested against its matched ceiling. PhysioNet, raw scale, 5 seeds, 340 cells. The lookup is $\hat{\mathbb{E}}[e^2 \mid T]$ fitted on half the sequences; inside a (channel, horizon) cell it is a function of the density tercile, so its ceiling is the within-cell correlation ratio η_c^2 over all three terciles. **The two matched rows are debiased by the same split-half product over whole sequences, on the same halves**, so their difference is paired; the interval is a sequence bootstrap of the held-out half, one resample for every seed and the seed mean taken within each draw (100 draws), conditional on the fitted lookups. The plug-in pair holds deterministically in sample and is an arithmetic check, not a test. The last matched row is the version we replaced: its target-level permutation null under-corrects clustered data. NABU is bounded by none of these — it reads V too. `per_cell_ceiling_sT.py`, `oracle_bias_sim.py`, `nested_eta2_debiased.py`.

	aggregation	s scale	correlation
diagnostic $\hat{s}_{\text{ext}}$, extreme terciles	ratio of weighted means	0.0084	0.092
ceiling s_T, debiased by sequence	η^2, split-half	0.0144	**0.120**
per-cell ceiling η_c^2, plug-in	mean_c	0.0733	—
lookup, plug-in	$\text{mean}_c\, r_c^2$	0.0699	—
per-cell ceiling η_c^2, split-half	mean_c	0.0626	—
lookup $\hat{\mathbb{E}}[e^2 \mid T]$, split-half	$\text{mean}_c\, r_c^2$	0.0607	—
paired gap, lookup − ceiling	*joint sequence bootstrap*	-0.0020	
	95% CI	$[-0.0027, -0.0014]$	
replaced: gap after permutation nulls	mean_c	$+0.0020$	
NABU $\hat{\sigma}^2$ (reads S *and* V)	single, pooled	—	0.183 ± 0.165

Three mismatches, each introduced by the correction of the last. The bound is Pearson while our measurements were Spearman; the bound is a single correlation while our oracle figures were averages of per-cell ones; and $\hat{s}$ conditions on strata that s does not.

(i) The bound is not violated, and the test cannot show more than that. Three drafts reported the oracle exceeding the ceiling, at 0.050, 0.183 and 0.177, each against $\sqrt{\hat{s}} = 0.092$; all three compared a mean of per-cell correlations with the root of a ratio of weighted means, which Corollary 1 does not relate. The chain that applies is the proposition inside each cell, aggregated the same way on both sides. Our first repair clipped, $\sqrt{\max(\hat{s}_c, 0)}$, which gave the ceiling a null floor three times the margin. Our second compared on the s scale unclipped, but a review found its per-cell ceiling was the extreme-tercile contrast $\hat{s}_{\text{ext}}$, which we had withdrawn as a ceiling because it ignores the middle tercile. The comparison in Table 23 replaces it with η_c^2 over all three terciles, computed on the same held-out points, cells and tercile edges the lookup uses.

(i) We changed the oracle's debiasing after seeing its result, and say so.* The first run of the corrected comparison subtracted a within-channel permutation null from each side, as every earlier version had. It put the lookup *above* the ceiling, $+0.0020$ with an interval excluding zero, which the proposition forbids in population. Before reading that as a violation we asked whether the estimators produce it when the truth obeys the bound. A simulation with PhysioNet's cell size and tail weight says they do, but only when targets cluster within sequences: with 8 targets per sequence sharing a tercile and an error level, the permutation-corrected lookup exceeds the ceiling on average in two of four settings, by $+0.0010$ and $+0.0016$ (standard errors 0.0002–0.0003), and with independent targets it falls short in all four. A target-level permutation breaks that clustering, so it estimates the lookup's upward bias at the wrong effective sample size. The split-half estimator, a product of covariances on two sequence-disjoint halves, has no such term and fell 0.002–0.005 short of the ceiling in every simulated setting (`results/v2_ceiling_oracle_sT/bias_sim.log`). That shortfall is the attenuation of a lookup fitted on half the data. The PhysioNet gap of -0.0020 is that size, so it is evidence of neither slack nor attainment. We withdraw the earlier reading that the bound is loose: a shortfall of this size is what the estimators produce when it is attained. The same simulation shows the per-cell ratio estimators run high under heavy tails at this cell size, because the in-sample variance understates a heavy-tailed denominator; that errs toward a larger ceiling.

(i′) What we can estimate is s_T, and it is a lower bound on s. Corollary 1 defines s for the *full* S; channel identity, query time and a density tercile are part of S, so the correlation ratio s_T of that coarsening satisfies $s_T \leq s$. Measuring it requires care: the plug-in $\sum_k p_k(m_k - \bar{g})^2$ is inflated by sampling noise in the group means, and the inflation grows with the number of groups. A split-half U-statistic whose halves are *whole sequences* removes that term in expectation; an earlier version split individual targets, which leaves one patient's correlated errors in both halves (Section 4.1.2). On PhysioNet at five seeds (`results/v2_ceiling_seqsplit`):

Table 24: The nested correlation ratio on PhysioNet at five seeds, plug-in against the sequence-split U-statistic. Intervals are in Table 3.

coarsening	groups	plug-in η^2	debiased η^2	debiased $\sqrt{\cdot}$
channel	37	0.0099	0.0068	0.083
+horizon-decile	369	0.0156	0.0065	0.081
+density-tercile	1055	0.0413	**0.0144**	**0.120**

An earlier version reported 0.203, the plug-in, and then 0.122 from the target-level split; on the same errors that split gives 0.121, so the correction for dependence is small here. One caveat we state rather than smooth over: the channel and horizon rungs are not separated by this estimator — the finer partition estimates slightly lower ($0.0068 \to 0.0065$), which cannot hold in population — so we report the total and the density increment and do not read the intermediate rungs as an ordered decomposition (`scripts/nested_eta2_debiased.py`).

The ceiling above is raw-scale, and the scale matters more than the debiasing did. Every number in that table is computed on e^2 in native clinical units. Because the baselines z-score per variable, we retrained both arms and recomputed (`results/v2_w1_gate_norm`, 3 seeds, only `dataset.normalize` differing). The raw arm reproduces the table on independently trained checkpoints — $0.0073/0.0070/0.0150$, $\sqrt{s_T} = 0.123$, density increment $+0.0080$, share 53% — which is what licenses reading the other column:

Table 25: The same correlation ratio on both scales, from one paired board where only `dataset.normalize` differs (`results/v2_w1_gate_norm`, 3 seeds, sequence split).

coarsening	groups	raw debiased η^2	z-scored debiased η^2	z-scored $\sqrt{\cdot}$
channel	37	0.0073	0.0013	0.037
+horizon-decile	369	0.0070	0.0019	0.044
+density-tercile	1055	**0.0150**	**0.0020**	**0.045**

Two readings, pointing opposite ways. $\sqrt{s_T}$ is *almost three times smaller* on the scale the baselines use — 0.045, not 0.123 — so the prize a T-reading head competes for is smaller than a raw-scale reading suggests. And density's share does not survive the change: z-scored, its increment is $+0.0001$ of a 0.0020 total, with interval $[-0.0013, +0.0015]$, against $+0.0080$, $[+0.0019, +0.0586]$ raw. *We predicted the opposite sign for this experiment*: raw e^2 has a kurtosis-4.6×10^4 tail carrying essentially all its variance, and we expected taming the denominator to raise η^2 and undercut our own headline.

Finally, a property of s that this makes unavoidable and that we had left implicit. $s = \mathrm{Var}[\mathbb{E}[e^2 \mid S]]/\mathrm{Var}[e^2]$ is defined through the *model's* error, so it was never a property of the data alone, and the z-scored arm is a much better forecaster (8–10$\times$ lower per-channel median MSE on a common scale, `results/v2_w1_gate_norm/arm_parity.csv`). The ceiling is therefore not a fixed prize that better methods approach. Nothing in general fixes the direction it moves as the fit improves: lower mean squared error changes both the numerator and the denominator of a ratio. On the two comparisons we have, the z-scored arm's $\sqrt{s_T}$ is far below the raw arm's (0.045 against 0.123, intervals disjoint), while at 30 against 10 epochs the point estimate is lower but the intervals overlap (Appendix B); that is one separated comparison, not a monotonicity result, and reporting one number for it understates how much it depends on what it was measured on.

(ii) The rank-domain comparison does not survive a null, and we withdraw it. We had read the oracle's 0.189 against a raw rank ceiling of 0.152 as attenuation. On synthetic data with a density share of *exactly zero* but a horizon effect non-monotone in τ, the same code reproduces a breach of

the same size: the rank measurement residualizes horizon only to first order in rank while the ceiling stratifies it away entirely. A statistic that fires at $s = 0$ says nothing about bias at $s > 0$.

(iii) "Loose" was never measurable from below anyway: two lower bounds (a coarsened oracle, an attenuated $\hat{s}$) can establish a violation but never looseness.

(iv) Nor can we show the head falls short of the oracle. Paired over five seeds the gap is $+0.037 \pm 0.052$ ($p = 0.18$; Wilcoxon returns the n=5 floor), carried by one seed on which NABU collapses to 0.031. Retracted by the standard that retired $\rho_s = +0.80$ — which we had again applied only to the unflattering result. Whether any head of ours collects part of the prize is not settled by that arm, whose contrast (-0.075 ± 0.049, $p = 0.25$) is underpowered in both directions; Appendix J is what speaks to it — the identified contrast, which resolves only on Activity. Section 4.1.7 cannot: it is not a density contrast, as that section says of itself.

M THE UNITS CONFOUND, MEASURED RATHER THAN INFERRED

Two falsified rescues in this revision were withdrawn because both rested on native-scale AURC, which Section 4.1.7 shows is largely a contest over which score sorts channels by error magnitude. An earlier version then observed that the correction was large on PhysioNet (native clinical units) and small on MIMIC (already normalized) and called that the diagnosis validating itself. That is an n=2 ordinal argument, found after the fact, running in our favour — and we had refused an n=4 one two subsections earlier when it ran against us. The confound is not a latent property to be inferred from an ordering; it is a number we already compute, so we compute it.

Confound index. $C = (1 - \mathrm{AURC}_{\text{chan}})/(1 - \mathrm{AURC}_{\text{omni}})$ on native-scale e^2, with AURC normalized so that no ranking = 1: the fraction of the achievable abstention benefit that a lookup table on *channel identity alone* captures. The lookup is fit on half the sequences and scored on the other half, so no channel earns AURC from its own points. Against it we put the size of the correction on the same checkpoints and the same metric, $D = \mathrm{AURC}_{NABU}(\text{native}) - \mathrm{AURC}_{NABU}(\text{standardized})$.

Table 26: The channel-units confound index C per dataset, fit on half the sequences and scored on the other half so no channel earns AURC from its own points, against D, the size of the correction on the same checkpoints and metric.

dataset	channels	C (native)	C (standardized)	C (labels permuted)	D
PhysioNet	37	$+\mathbf{0.972}\pm.003$	$+0.008\pm.032$	-0.035	$-0.463\pm.168$
USHCN	5	$+0.472\pm.008$	$+0.026\pm.110$	$+0.136$	$-0.216\pm.026$
MIMIC-III	96	$+0.399\pm.013$	$-0.054\pm.034$	-0.084	$-0.096\pm.007$
Activity	12	$+0.261\pm.015$	$-0.034\pm.035$	-0.043	$-0.112\pm.012$

Two controls behave as they must: on per-channel standardized errors C is within 0.06 of zero everywhere, and permuting the channel labels returns it to zero on the three datasets with at least twelve channels. On USHCN it does not ($+0.136$): with five channels a label permutation is one of 120 and leaves much of the assignment intact, so the permutation null is uninformative there and we do not read it as signal. *This control is also what found a defect in our own estimator* — before ties were broken at random it read $+0.21$ on PhysioNet, where the truth is zero (Section 4.1.7).

What we claim, and what we do not. C spans 0.26 to 0.97 and $|D|$ spans 0.10 to 0.46, and they are ordered together ($r = -0.97$, $\rho_s = -0.80$). We claim no relationship: at n=4 exact permutation test can return below 0.083, the Spearman permutation p is 0.200, and the parametric $p = 0.029$ assumes bivariate normality that four points cannot check. This is the same standard under which Appendix E refuses $\rho_s(\sqrt{\hat{s}}, |\text{marginal}|) = +0.80$, and it has to apply in both directions. What is *measured*, with no inference: on PhysioNet a channel lookup captures 97% of everything abstention can buy, and on MIMIC 40%. That is the confound, on the two datasets where the withdrawn comparison was made, as a quantity rather than a story. Note also that the naive reading fails on USHCN, whose five near-regular channels still carry $C = 0.47$ — a small channel count, not a small confound (`scripts/confound_index.py`).

Every affected result, on both scales, permanently. A withdrawal that replaces a number is indistinguishable from harvesting unless both numbers stay visible, so we keep them together rather than in the sections that retract them one at a time.

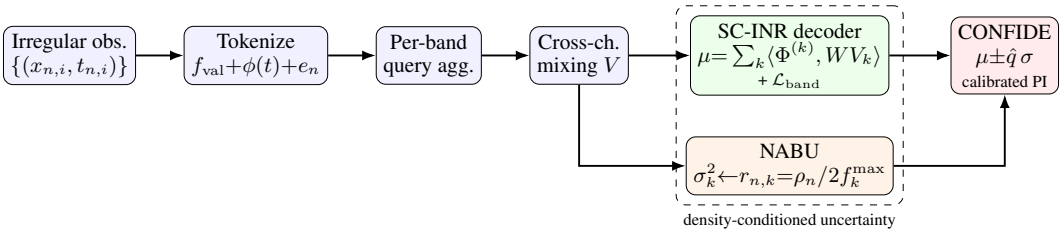


Figure 3: BandRate pipeline. Irregular observations are tokenized and aggregated by per-(channel,band) query tokens into context V, with no interpolation onto a grid. The SC-INR decoder reconstructs μ from K band-limited SIRENs (kept band-limited by $\mathcal{L}_{\text{band}}$); NABU derives each band's variance from its rate-normalized identifiability ratio $r_{n,k}$; CONFIDE calibrates the resulting distribution into prediction intervals with finite-sample coverage.

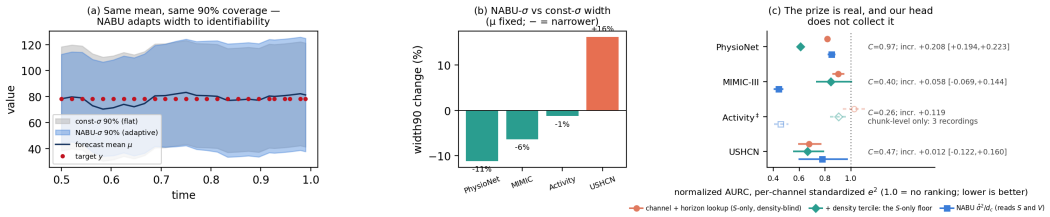


Figure 4: **What the variance shape buys.** (a) One PhysioNet channel — the *median* const-vs-NABU width saving among channels with ≥ 10 targets, not the best one — under split conformal: same μ, same marginal coverage, but the identifiability-driven band tightens where the channel is densely observed. (b) With μ held fixed, swapping only $\hat{\sigma}$ narrows the 90% width where sampling is irregular (PhysioNet -11%, MIMIC -6%) and not where it is near-regular. (c) is Figure 1, repeated here for context.

Table 27: Every result whose reading changed with the scale, collected in one place; each is retracted one at a time in the sections above. The standardized column ranks by $\hat{\sigma}^2/d_c$, the variance in the loss's units; earlier versions ranked by the raw σ.

result	native scale	per-channel standardized	C
Shift robustness (Table 10)	NABU worst of 5 (1.97)	separates nothing	0.97
Deployed UQ, PhysioNet	NABU 4/5 (3823 vs 702)	0.854 vs 0.926–1.014 (n.s.)	0.97
Deployed UQ, MIMIC-III	NABU 2nd	0.494 vs 0.489 (n.s.)	0.40
Selective prediction vs pooled	$+65\%$, p=0.008, 5/5	$+12.0\pm11.1\%$, p=0.074, 4/5	0.97
Per-dataset ordering (Appendix H.0.2)	PhysioNet 4/5, $+464\%$	question retired	0.97
The S-only floor (Section 4.1.7)	0.053	**0.611**	0.97

Three of the six were results we had been reporting *against* ourselves, and one — the selective-prediction gain — is the largest number in the paper's favour, which the correction cuts by a factor of 4.6 and below its own pre-registered gate. That asymmetry is the case that the correction was not selected to rescue anything: Corollary 2 forces a loss to be *declared*, and it is declared, and it was validated by a control with no bearing on any of these rows.

N THE INSTANTIATION WE MEASURED THIS ON

The main body keeps only what the ceiling argument needs: the variance head's inputs and the density estimator. The architecture those sit inside, and the experiments that establish it is a reasonable forecaster rather than a strawman, are collected here. Nothing in Section 4.1.2 depends on these choices — the measurement is model-agnostic, which is the point — but a bound measured on a broken model would not be worth reporting, so the evidence that it is not broken belongs in the record.

N.1 BAND-PARTITIONED CONTINUOUS REPRESENTATION (SC-INR)

Observations are tokenized as $z_{n,i} = f_{\text{val}}(x_{n,i}) + \phi(t_{n,i}) + e_n$ with a learnable harmonic time embedding ϕ (Kazemi et al., 2019) and channel identity e_n. For each (channel n, band k) a learnable query token cross-attends to channel n's tokens (masked; channels with no observations are zeroed, not imputed), and a per-band variable-axis transformer mixes channels, yielding context $V \in \mathbb{R}^{N\times K\times D}$.

The decoder is a bank of K SIRENs whose first-layer frequencies are initialized in disjoint bands $B_k = [2\pi f_k^{\min}, 2\pi f_k^{\max}]$:

$$\hat{y}_n(\tau) = \textstyle\sum_{k=1}^{K} \left\langle \Phi^{(k)}(\tau),\, WV_{n,k} \right\rangle, \tag{1}$$

where $\Phi^{(k)}$ are band-k features. Depth-≥ 3 SIRENs generate harmonics ($\sin \circ \sin$) that let low bands emit high-frequency content, silently breaking the decomposition. We therefore penalize the *out-of-band spectral energy fraction* of each $\Phi^{(k)}$, measured by windowed FFT on a regular τ grid:

$$\mathcal{L}_{\text{band}} = \tfrac{1}{K} \textstyle\sum_k E_{\text{out}}^{(k)} / E_{\text{tot}}^{(k)}. \tag{2}$$

This restores per-band recovery at depth 3 (NMSE $0.20 \to 0.036$ on the leaking band; Appendix N.8), tightening the band-limitedness the sampling-theoretic motivation assumes. It is *optional*: the band-partitioned ω initialization already yields approximately band-limited features, and our accuracy runs (Table 28) set its weight to zero because the penalty trades a small amount of point accuracy (Appendix U); we enable it only for the decomposition study. A band-orthogonality penalty on the same grid discourages redundant bands.

N.2 CONFIDE: DENSITY-MONDRIAN CONFORMAL CALIBRATION

NABU gives variance the right *shape*; conformal prediction anchors its *scale*. On a held-out calibration split we compute standardized scores $s_i = |y_i - \mu_i|/\hat{\sigma}_i$ and the finite-sample-corrected quantile $\hat{q}$ at level $1-\alpha$, yielding intervals $\mu \pm \hat{q}\,\hat{\sigma}$ with the standard distribution-free marginal guarantee under exchangeability (Vovk et al., 2005; Lei et al., 2018). Because score distributions differ across sparsity regimes, we additionally bin calibration points by ρ_n quantiles (Mondrian conformal) with one $\hat{q}_b$ per bin. The standardized score is Papadopoulos et al.'s normalized nonconformity (Papadopoulos et al., 2011) and the per-bin quantile is a Mondrian conformal regressor (Boström & Johansson, 2020); neither is ours, and exact conditional coverage is provably unattainable distribution-free (Barber et al., 2021), so what binning buys is coverage conditional on a coarsening, not conditional coverage. Our only choice here is the taxonomy variable (density). This gives a per-cell guarantee:

Proposition 1 (Per-bin conditional coverage)**.** *Partition calibration and test points into density bins $\{b\}$ by the ρ_n-quantile edges. If within each bin b the calibration and test standardized scores s_i are exchangeable, then Mondrian conformal with the finite-sample-corrected quantile $\hat{q}_b$ satisfies* $\Pr(y \in \text{PI} \mid \text{bin}{=}b) \geq 1 - \alpha$ *for every b, hence* $\Pr(y \in \text{PI}) \geq 1 - \alpha$ *marginally.*

The proof (Appendix R) applies the split-conformal exchangeability argument (Vovk et al., 2005; Lei et al., 2018) independently within each bin; a single global $\hat{q}$ gives only the marginal statement and can therefore under-cover the sparsest cell, which Appendix N.5 confirms and Mondrian repairs. The better $\hat{\sigma}$ orders difficulty, the narrower the calibrated intervals — this is where NABU pays.

N.3 TRAINING OBJECTIVE

$$\mathcal{L} = \mathcal{L}_{\beta\text{-NLL}} + \alpha\,\mathcal{L}_{\text{recon}} + \beta\,\mathcal{L}_{\text{ortho}} + \delta\,\mathcal{L}_{\text{band}} + \gamma\,\mathcal{L}_{\text{CRPS}} + \eta\,\mathcal{L}_{\text{cons}}. \tag{3}$$

$\mathcal{L}_{\beta\text{-NLL}}$ is the β-weighted Gaussian NLL of Seitzer et al. (2022) ($\beta{=}0.5$), which removed a severe seed instability of the variance head on PhysioNet (5-seed MSE $10377{\pm}2449 \to 6535{\pm}829$ on the *validation* split). $\mathcal{L}_{\text{recon}}$ is masked MSE on *observed* history points only. $\mathcal{L}_{\text{cons}}$ is our **rate-consistency** term: each step draws a Bernoulli-thinned view of the history mask and pulls its prediction toward the full view's detached prediction — the denser view acts as teacher, enforcing invariance of $\hat{y}$ to the sampling rate.

Table 28: Held-out **test**-split MSE. * = 5 seeds (BandRate and the best baseline per dataset, the two arms the significance test in Appendix V compares); all other cells are 3 seeds — we extended only the cells the test needed, and label rather than round over the difference. Every cell is recomputed from the released boards (`results/repro/main_table.csv`, `results/v2_cr_significance/`); an earlier draft quoted five cells that differed from them by 0.1–6%, none of which changed a ranking. Activity uses a leakage-free sequence-level group split; MIMIC uses the per-variable normalization standard to the baselines (different scale; see text); PhysioNet/USHCN native scale. Best per column in bold; BandRate underlined. QuITE: our reimplementation.

Model	Activity↓	PhysioNet↓	USHCN↓	MIMIC↓
GRU-D	0.0571	15358	0.620	0.657
mTAND	0.0631	12778	**0.400***	0.697
Latent-ODE	0.0729	16813	0.607	0.720
T-PatchGNN	0.0575	**11512***	0.605	0.624
QuITE	**0.0534***	12782	0.434	**0.614***
BandRate	<u>0.0543</u>*	<u>12627</u>*	<u>0.418</u>*	<u>0.630</u>*

N.4 Main accuracy (E1)

Table 28: at 5 seeds on the held-out test split BandRate places **second on Activity, PhysioNet and USHCN, and third on MIMIC** — top-2 on three of four datasets and top-3 on all four. That board predates the newest IMTS forecasters; Appendix W re-runs the comparison inside APN's own harness, where the ordering goes the other way by 1–2%, and reports both protocols rather than the more favourable one. No model dominates: the per-dataset best is QuITE (Activity, MIMIC), T-PatchGNN (PhysioNet), and mTAND (USHCN). Ordinal rank is seed-sensitive, however — PhysioNet moved from 4th at 3 seeds to 2nd at 5 — so the rigorous statement is a paired significance test across seeds (Appendix V): the MSE gap to the strongest baseline is *not* significant on Activity ($+1.6\%$, paired t p=0.21) or PhysioNet ($+9.7\%$, p=0.15), and significant but under 5% on USHCN ($+4.6\%$, p=0.001) and MIMIC ($+2.6\%$, p=0.007); an example-level test detects every gap given the large test sets, so BandRate is best read as accuracy-*competitive* — not significantly different on two datasets (no equivalence test was run, so not "tied"), a few percent behind on two, though *that reading is specific to the capacity we ran the baselines at*: at their authors' own settings the PhysioNet gap is 17.7% rather than 9.7% (Appendix S) — trading a small point-accuracy margin for the calibration and robustness developed next. Its distinguishing property is not a leaderboard win but that its uncertainty is the one whose variance is explicitly conditioned on per-band identifiability (Section 4.1), a property none of the compared UQ baselines targets — competitive-to-narrowest calibrated intervals with no method dominating all four (Section 4.2 and Appendix N.5), more rate-robust than the strong interpolation-based forecasters in point error (Appendix N.7) though the uncertainty-quality comparison under thinning, which an earlier draft reported as its degrading fastest of five methods, turns out to separate nothing once errors are standardized per channel (Appendix H.0.1), and a variance that tracks per-channel sampling density where a pooled head does not (Section 4.1). Earlier drafts reported validation-split numbers and a chunk-level Activity split, both of which inflated ranking; the corrected leakage-free test protocol with a seed-level significance test gives the calibrated picture above.

N.5 Calibration with guarantees (E2/E6)

On all four datasets CONFIDE brings 90% coverage to within ± 1.2 points of nominal (split: 0.892–0.912) and 50% coverage to 0.49–0.52, while *also* shrinking width 16–20% versus raw Gaussian intervals (Table 29) — no coverage/width trade. Raw Gaussian intervals systematically over-cover (0.917–0.938) because the un-anchored variance is miscalibrated. The Mondrian bins matter where score distributions differ across density cells: on PhysioNet the global quantile under-covers the sparsest density quintile (0.845) and Mondrian restores it (0.898), reducing the across-quintile coverage spread $17\times$ ($0.0751 \to 0.0044$); on the lower-channel datasets split and Mondrian coincide, consistent with the theory of when binning helps. Activity's slight over-coverage (0.912) is *not* the finite-sample quantile being conservative, as an earlier draft said: its calibration split holds $18{,}321$ points ($3{,}664$ per Mondrian bin), where the corrected level is 0.90001 — two orders of magnitude

Table 29: Coverage at nominal 90%/50% and 90% interval width on the **test** split (calibrated on val; mean over 3 seeds), all four datasets. Width units differ by dataset scale.

Dataset	Method	Cov@90	Cov@50	Width@90↓
Activity	Gaussian	.938	.691	0.709
	Split	.912	.519	**0.598**
	Mondrian	.912	.518	0.600
PhysioNet	Gaussian	.917	.722	90.1
	Split	**.899**	**.504**	**72.0**
	Mondrian	**.900**	**.504**	74.6
USHCN	Gaussian	.933	.678	2.00
	Split	**.892**	.490	1.65
	Mondrian	.891	.491	**1.64**
MIMIC	Gaussian	.931	.657	2.31
	Split	**.899**	**.506**	1.94
	Mondrian	**.899**	**.506**	**1.94**

Table 30: width90 at cov90≈0.90 under the same split conformal: strong-backbone Deep Ensembles over existing checkpoints of the best point models (K=5) vs single BandRate. Best per row bold; no retraining. BandRate stays narrowest on 3/4 despite the baselines' stronger backbones.

Data	T-PatchGNN-ens	QuITE-ens	BandRate
Activity	0.803	0.676	**0.590**
PhysioNet	**67.9**	83.6	74.4
USHCN	2.23	2.21	**1.54**
MIMIC	2.66	3.37	**1.96**

below the 1.2-point excess. The small group split causes something else. Two calibration recordings against three test recordings makes calibration and test scores exchangeable only under an assumption two clusters cannot support, and the excess moves with the seed (0.914, 0.915, 0.907) rather than sitting at a fixed floor.

N.6 Is the sharpness from μ or σ? Two confound controls (E2c)

Split-conformal width is $\hat{q}\,\hat{\sigma}$, which shrinks as μ improves — so a skeptic could ascribe BandRate's narrow intervals to a better *mean* rather than a better variance *shape*. We separate the two.

(i) σ-shape with μ held fixed. Keeping BandRate's own μ and swapping only $\hat{\sigma}$ from homoscedastic (constant) to NABU, under identical conformal, narrows the 90% width by 11.1% on PhysioNet and 6.3% on MIMIC at matched coverage — with μ *identical*, so the gain is purely the variance shape. On the near-regular datasets it is neutral or harmful (-1.2% Activity, $+16.2\%$ USHCN), exactly where density barely varies: NABU sharpens iff there is identifiability structure to exploit, note, however, that this is the *opposite* ordering to Appendix H.0.2, where under a uniform error-ranking metric PhysioNet is where NABU does worst; sharpening interval width and ranking error are different objectives and the two need not agree.

(ii) Strong-backbone UQ. The baselines of Table 7 share a GRU-D backbone; to rule out a weak-backbone artifact we give the baselines their best case — Deep Ensembles of independently trained checkpoints (five members) of the strongest point models (T-PatchGNN, QuITE): an excellent ensemble mean plus epistemic variance (Table 30). Under the same conformal, BandRate still yields the narrowest guaranteed intervals on three of four datasets — on MIMIC (1.96 vs 2.66) — and is second on PhysioNet, where T-PatchGNN's stronger mean (Table 28) wins the combined width. The sharpness is not an artifact of baseline backbone choice.

Table 31: Rate-shift robustness (synthetic diagnostic; train sparse/dense, test the other). **Main-config objective** (depth-3 SC-INR, NLL+recon+ortho+CRPS), 3 seeds; gap = shifted − matched MSE.

Model	sparse→dense	dense→sparse	mean |gap| ↓
GRU-D	+0.013	+0.010	**0.012**
BandRate (η=1)	−0.022	+0.054	0.038
Latent-ODE	+0.009	+0.090	0.049
T-PatchGNN	−0.064	+0.128	0.096
mTAND	−0.060	+0.154	0.107

N.7 Sampling-rate shift (E4)

Under the main-config objective (an earlier draft used a reduced depth-2 synthetic model that overstated the effect), rate-consistency training gives BandRate a *modest* shift-gap reduction ($0.050 \to 0.038$, $1.3\times$; per-seed std ~ 0.03, so this is suggestive not decisive). BandRate's mean shift gap is 2.5–3$\times$ smaller than the strongest point forecasters' (mTAND 0.107, T-PatchGNN 0.096; 3 seeds, per-seed std ~ 0.03, and at n=3 the distribution-free floor is 0.25, so no significance is attainable here either), consistent with grid/reference-point interpolation overfitting the training rate; the simpler GRU-D is the most robust of all. We therefore claim rate-robustness *relative to the strong interpolation-based baselines*, not overall, and report it as a controlled synthetic diagnostic rather than a real-data guarantee.

N.8 Decomposition and ablations (E5/E8)

With $\mathcal{L}_{\text{band}}$, a depth-3 SC-INR recovers all three planted components of a synthetic multi-band signal at per-band NMSE $\{0.044, 0.036, 0.078\}$ versus $\{0.091, 0.202, 0.055\}$ without it ($5.6\times$ on the leaking band). This is one seed (42), and the seed picture is worse than that number suggests: across seeds $\{0, 1, 2, 42\}$ only seed 42 clears our 0.1 NMSE gate, seed 0 improves but misses it ($0.166 \to 0.139$), and seeds 1 and 2 show SIREN optimization divergence ($0.692 \to 0.819$, $0.386 \to 0.364$) in which the decomposition never fits. We therefore claim only that $\mathcal{L}_{\text{band}}$ removes the leakage *when the depth-3 SC-INR converges*, and count the convergence failure itself as an open limitation. Ablations (Appendix U; PhysioNet test split, 3 seeds): at this budget the point-MSE gaps across ablations are within seed noise (per-seed std 0.6–5.1k), so we draw no accuracy ranking from them; the reproducible signal is *calibration*. Removing NABU (pooled head) drives the within-channel density–variance correlation negative ($+0.65 \to -0.11$) and removing cross-channel mixing collapses it ($\to +0.06$) — both are required by NABU's construction, so this is a check that the implementation does what the architecture says, not a test of necessity (3 seeds, PhysioNet) — while band count modulates it (K=1: $+0.43$; $K \geq 2$: $\geq +0.65$).

N.9 What to do with this if you train IMTS models

The transportable part of this paper is not NABU but the measurement, which costs one pass over a validation split and is model-agnostic. **(i) Put channels on a common scale first.** Train on per-variable z-scored values, as the baselines do; on raw clinical units a pooled head's density–variance correlation and every baseline's error ranking are dominated by channel units (Appendix I.0.3). **(ii) Partial out forecast horizon** from the within-channel correlation between $\hat{\sigma}^2$ and local sparsity $1/\rho_n$; without it the statistic mostly reports that variance grows with lead time ($\rho_s(1/\rho_n, \text{horizon}) = +0.70$ on our data). **(iii) Estimate s_T with the sequence as the unit** (`tools/imts_uq_ceiling.py`, Appendix). Read it as what a head reading only channel, horizon and density tercile can track, and as a *lower* bound on what a head reading the full observation set could; it does not bound heads that also read observed values, and a small value is not by itself a reason to abandon the family. **(iv) If the decision is abstention, measure the cross-fitted floor** on the loss you will actually use, and report it as a change in risk–coverage area, not as a risk reduction at a retention level. **(v) To ask whether an input helps, delete it and retrain** with matched budgets, several seeds and the sequence as the unit; permuting it at inference measures the model's dependence on the input, which is a different question. On the choice of UQ method our comparison settles little: on z-scored PhysioNet all five methods rank error (0.739–0.875 standardized AURC, Appendix I), a deep ensemble is not distinguishable

from NABU on MIMIC, and the narrowest intervals are not the best interval scores on heavy-tailed PhysioNet (Section 4.2).

O PROOFS

Lemma 1 (Orthogonality of the residual)**.** *Let $Y = e^2$ with $\mathbb{E}Y^2 < \infty$ and $\mathrm{Var}[Y] > 0$, and let $S = (n, \tau, \{(t_{m,i}, m)\})$ collect the query's own channel and time together with the observed timestamps and channel identities — not the observed values. Put $R := Y - \mathbb{E}[Y \mid S]$. Then $\mathbb{E}[R \mid S] = 0$, and hence $\mathrm{Cov}(g(S), R) = 0$ and $\mathbb{E}[R\,\mathbf{1}_A] = 0$ for every $g \in L^2(\sigma(S))$ and every $A \in \sigma(S)$.*

Lemma 1. For $g \in L^2(\sigma(S))$, $\mathbb{E}|g(S)R| < \infty$ by Cauchy–Schwarz and $\mathbb{E}Y^2 < \infty$, so the tower property applies: $\mathbb{E}[R \mid S] = 0$ gives $\mathrm{Cov}(g(S), R) = \mathbb{E}[g(S)\,\mathbb{E}[R \mid S]] = 0$. Measurability alone does not suffice — it does not make $\mathbb{E}|g(S)R|$ finite, and for $g \notin L^2$ the correlation is undefined. Taking $g = \mathbf{1}_A$ for $A \in \sigma(S)$ gives $\mathbb{E}[R\,\mathbf{1}_A] = 0$. □

Corollary 1. Write $Y = e^2$. By Lemma 1, $\mathrm{Cov}(g, Y) = \mathrm{Cov}(g, v) \le \mathrm{sd}(g)\,\mathrm{sd}(v)$ by Cauchy–Schwarz; dividing by $\mathrm{sd}(g)\,\mathrm{sd}(Y)$ gives the bound and $g \mapsto -g$ the two-sided form. At $g = \mathbb{E}[Y \mid S]$ both sides equal $\mathrm{sd}(v)/\,\mathrm{sd}(Y) = \sqrt{s}$, so the bound is attained whenever $\mathrm{Var}[v] > 0$; if $\mathrm{Var}[v] = 0$ that g is a.s. constant and its correlation is undefined rather than zero, which is the relevant case on USHCN, whose ceiling CI reaches zero. □

Corollary 2. Let $A \in \sigma(S)$ with $\Pr(A) = 1 - f$ be a retention set. By Lemma 1, $\mathbb{E}[Y\mathbf{1}_A] = \mathbb{E}[\mathbb{E}[Y \mid S]\mathbf{1}_A]$, so the retained risk depends on A only through $\mathbb{E}[Y \mid S]$; among sets of that measure it is minimized by retaining where $\mathbb{E}[Y \mid S]$ is smallest. Hence $f^\star = \mathbb{E}[Y \mid S]$ is the optimal S-measurable score at every coverage level, and any exhibited S-measurable score has risk–coverage area at least the oracle's — so its measured area is an upper bound on the oracle's, and one minus that area is a lower bound on what S can buy. □

A small ceiling does not bound the floor. The two quantities answer different questions, and one cannot stand in for the other. Let $T \sim \mathrm{Bernoulli}(1/2)$ and $B \sim \mathrm{Bernoulli}(p)$ be independent, and let $Y = e^2 = 1 + T + B/p$, realized as $e = \pm\sqrt{Y}$ with an independent random sign so that the forecast is unbiased. Then $\mathbb{E}[Y \mid T] = 2 + T$, $\mathbb{E}Y = 2.5$ and $\mathrm{Var}\,Y = 1/p - 3/4$, so

$$s_T = \frac{\mathrm{Var}\,\mathbb{E}[Y \mid T]}{\mathrm{Var}\,Y} = \frac{1/4}{1/p - 3/4} \;\to\; 0 \quad \text{as } p \to 0.$$

Yet retaining the half with $T = 0$ has risk 2 against 2.5, a 20% reduction for every p, and the oracle score T lowers the normalized risk–coverage area over abstaining 0–90% by 0.131, also for every p. At $p = 0.1$, 0.01 and 0.001 the ceiling $\sqrt{s_T}$ is 0.164, 0.050 and 0.016. The correlation ceiling is driven by a rare heavy tail the score cannot see, while the floor is driven by the conditional means the score does see. So a small $\sqrt{s_T}$ is not evidence that abstaining on the observation set buys little, which is why the paper reports the floor separately and never infers one from the other. The converse caution also holds: with small p a finite sample may see few of the rare large errors, so the floor's estimate is noisy exactly where the ceiling is smallest (`tests/test_tool_estimators.py` checks both numbers by simulation).

P THE PARTIAL CEILING

Corollary 3 (Partial ceiling)**.** *Let $Y \in L^2$ with $\mathrm{Var}[Y] > 0$, let S be the observation set, and let C be any $\sigma(S)$-measurable random element with $\mathbb{E}[\mathrm{Var}(Y \mid C)] > 0$. Put $\tilde{Y} := Y - \mathbb{E}[Y \mid C]$, $\tilde{g} := g - \mathbb{E}[g \mid C]$ and $s_{S|C} := \mathbb{E}[\mathrm{Var}(\mathbb{E}[Y \mid S] \mid C)]/\mathbb{E}[\mathrm{Var}(Y \mid C)]$. Then for every $g \in L^2(\sigma(S))$ with $\mathrm{Var}[\tilde{g}] > 0$, $|\,\mathrm{corr}(\tilde{g}, \tilde{Y})| \le \sqrt{s_{S|C}}$, attained at $g = \mathbb{E}[Y \mid S]$ when the numerator is positive.*

Proof. Since $\sigma(C) \subseteq \sigma(S)$ we have $\mathbb{E}[\mathbb{E}[Y \mid C] \mid S] = \mathbb{E}[Y \mid C]$, so $\mathbb{E}[\tilde{Y} \mid S] = \mathbb{E}[Y \mid S] - \mathbb{E}[Y \mid C] =: w$ and $\tilde{Y} - w = Y - \mathbb{E}[Y \mid S] = R$, the residual of Corollary 1. Conditional expectation is an L^2 contraction, so $\tilde{g} \in L^2(\sigma(S))$ and $\mathbb{E}|\tilde{g}R| < \infty$; hence $\mathrm{Cov}(\tilde{g}, R) = \mathbb{E}[\tilde{g}\,\mathbb{E}[R \mid S]] = 0$ and $\mathrm{Cov}(\tilde{g}, \tilde{Y}) = \mathrm{Cov}(\tilde{g}, w) \le \mathrm{sd}(\tilde{g})\,\mathrm{sd}(w)$. By the tower property $\mathrm{Var}[w] = \mathbb{E}[\mathrm{Var}(\mathbb{E}[Y \mid S] \mid C)]$

and $\mathrm{Var}[\tilde{Y}] = \mathbb{E}[\mathrm{Var}(Y \mid C)]$; dividing gives the bound, and $g \mapsto -g$ the two-sided form. At $g = \mathbb{E}[Y \mid S]$, $\tilde{g} = w$ and both sides equal $\mathrm{sd}(w)/\mathrm{sd}(\tilde{Y})$. □

Taking C trivial recovers Corollary 1; taking C = (channel, horizon-decile) gives $s_{S|C} = \hat{s}$'s estimand. Applying Corollary 1 to the conditional law given $C = c$ instead gives $|r_c| \le \sqrt{s_c}$ per cell, hence $|\sum_c w_c r_c| / \sum_c w_c \le \sum_c w_c \sqrt{s_c} / \sum_c w_c$ for any non-negative weights. The first two bound ratios of weighted means; the third bounds means of ratios. They are not ordered by Jensen in general and each governs only its own aggregation.

Q THE SAMPLING-THEORETIC MOTIVATION, AND WHY WE DO NOT CLAIM IT AS A THEOREM

We retain the sketch below because it is what suggested the ratio, but it is not a proof about our setting and we no longer present it as one. Two steps are specifically unsound as applied: Landau's condition is necessary for stable sampling *(a bounded inverse), not for* uniqueness*, so the step that infers an annihilating signal from sub-Landau density does not follow — and with finitely many samples such a signal exists at any density, making the inference vacuous rather than sharp; and the whole argument concerns the interval containing the samples, whereas all of our queries lie outside it. A correct version would need a frame-bound or Remez-type control of the annihilator's magnitude at τ, which we do not have.*

Necessity follows from Landau's density theorem (Landau, 1967): stable sampling of the Paley–Wiener space PW_{B_k} requires lower Beurling density at least $|B_k|$; below it, the observation set fails to be a set of uniqueness, so there exists $g \in PW_{B_k}$, $g \neq 0$, vanishing (up to window-boundary leakage controlled by $T f_k^{\min} \gg 1$) on all observed times in the window. The two-point construction $x'_k = x_k + \alpha g$ with g normalized at τ is observationally indistinguishable from x_k, and Le Cam's two-point bound yields the signal-power variance floor. Sufficiency uses quantitative frame bounds above the Landau rate for well-spread (separated) sampling sets; Gaussian noise then propagates through the bounded reconstruction operator as $O(\sigma_\varepsilon^2/(\rho T))$. Caveats: irregular sampling can resolve a *single* sinusoid below uniform Nyquist; the statement is therefore at the band/worst-case level, and asymptotic in the window–bandwidth product, which is weakest for the lowest band — matching the empirically noisier band-0 curve in Figure 2.

R BACKGROUND: PER-BIN VALIDITY OF MONDRIAN CONFORMAL

Fix a density bin b and let $\mathcal{C}_b = \{s_i\}_{i=1}^{n_b}$ be its calibration scores and s^* a test score in the same bin, all standardized $s = |y - \mu|/\hat{\sigma}$. Mondrian conformal sets $\hat{q}_b$ to the $\lceil (n_b{+}1)(1{-}\alpha) \rceil$-th smallest of $\mathcal{C}_b$. If $s^*, s_1, \ldots, s_{n_b}$ are exchangeable, then by symmetry the rank of s^* among the $n_b{+}1$ scores is uniform on $\{1, \ldots, n_b{+}1\}$, so $\Pr\big(s^* \le \hat{q}_b\big) \ge \lceil (n_b{+}1)(1{-}\alpha) \rceil / (n_b{+}1) \ge 1 - \alpha$ — the standard split-conformal argument (Vovk et al., 2005; Lei et al., 2018) restricted to bin b. Since $\{y \in \mathrm{PI}\} = \{s^* \le \hat{q}_b\}$, this is exactly $\Pr(y \in \mathrm{PI} \mid \mathrm{bin}{=}b) \ge 1 - \alpha$. Marginalizing over bins, $\Pr(y \in \mathrm{PI}) = \sum_b \Pr(\mathrm{bin}{=}b) \Pr(y \in \mathrm{PI} \mid \mathrm{bin}{=}b) \ge 1 - \alpha$. The guarantee is only as good as within-bin exchangeability: a global $\hat{q}$ pools bins with different score scales, so it holds marginally but can under-cover the heaviest-tailed (sparsest) cell, which is what Table 29 shows (0.845) and per-bin $\hat{q}_b$ repairs (0.898). Finite-sample caveat: each bin needs $n_b \gtrsim 1/\alpha$ for the corrected quantile to be attainable; we use 5 equal-mass bins so $n_b = n/5$.

S EXPERIMENTAL DETAILS

A note on the name. The released code, every result board and every checkpoint path use `timeprism` as the model identifier. That was this method's name until we found that a model called TimePrism had been published at ICLR 2026 — Dai et al. (2026), in the same subfield — and renamed ours to avoid a reader taking the two for one. We did not rewrite the recorded outputs to match: the boards are the record of what was run, and a name chosen afterwards is not a reason to edit them. Read `timeprism` in the artefacts as BandRate.

No hyperparameter of BandRate was selected on test, and almost none was selected at all. The architecture constants — K, the band edges, d_{model}, SIREN depth, head count — and every loss weight ($\alpha_{\text{recon}}, \beta_{\text{ortho}}, \gamma_{\text{calib}}, \beta_{\text{NLL}}, \delta_{\text{band}}, \eta_{\text{cons}}$) were fixed when the model was first wired and never varied: each appears in the config history exactly once, as an addition, with no later edit. The three settings that *were* chosen from data — β-NLL β=0.5, NABU on by default, and the FiLM and output-bound variants left off — were each decided on the *validation* split (`results/betanll_validation`, `results/v2_w2_nabu_5seed`, `results/v2_w5_validation`; every evaluation log in those boards reads `split=val`). Table 33 corroborates this from the other direction: the shipped configuration is the *worst* MSE row in it and does not hold the best `dvcorr` either, which is not what a test-selected configuration looks like. This cuts against us as much as for us — our baselines ran at their authors' defaults while ours were never tuned at all — and we state it because "no tuning" is a claim a reader should be able to check rather than take.

The resampling unit, dataset by dataset. Our cluster bootstrap resamples whatever one dataset sample is, so we state what that is rather than leaving a reader to infer it. PhysioNet: one patient record, 4,000 in the test split — unit and sample coincide. USHCN: one station, and the test split holds 112 of them, which is why its floor interval $[.521, .820]$ is the widest of the four; the bootstrap is correctly specified, the n is simply small. Activity: one 30-row chunk, and the mismatch that costs us both positive results (Appendix B). MIMIC-III: one ICU *stay*. **We cannot verify that a stay is an independent unit here.** The MIMIC-III-Ext-tPatchGNN cohort we consume exposes only `ID` and `Time` over 194 columns, with IDs re-indexed contiguously, so the stay-to-subject mapping is not recoverable from it. In raw MIMIC-III, 18.8% of subjects have more than one stay and those account for 38.7% of all stays; if the upstream preparation kept repeat stays, MIMIC's intervals would be understated the way Activity's are, and one subject could sit in both splits. We have no evidence that it did — the preparation may well take one stay per subject — and we report stay-level independence as *assumed, not verified*, because an assumption we cannot test is not one a reader should have to discover.

All models: AdamW, lr 10^{-3}, batch 32 (16 on Activity), 10 epochs, gradient clipping 1.0, seeds $\{1..5\}$. BandRate: K=3 bands at $[1, 4], [4, 16], [16, 64]$ Hz over the normalized window, d_{model}=d_{band}=64, SIREN depth 3, β-NLL β=0.5, α=1, β_{ortho}=0.1, γ=0.1; δ_{band} and η_{cons} are 0 for accuracy boards and 0.5/1.0 for decomposition/robustness experiments respectively (stated wherever used). Conformal: 5 density-quantile bins, calibration on the validation split. Compute: 3× NVIDIA TITAN Xp (12 GB); a full 5-seed board runs in under one GPU-hour. PhysioNet uses the challenge's set-a/b/c partition; Activity and USHCN use deterministic 80/10/10 splits. **Horizons, and on USHCN we are not doing the published task.** PhysioNet and MIMIC use 24 h → 24 h, which matches the protocol these benchmarks are published under. Activity and USHCN instead split each sequence at its own midpoint (`history_frac` = 0.5), and measured against the data that is *not* the published window. Upstream (`lib/parse_datasets.py`, `scripts/run_all.sh` of the t-PatchGNN release) USHCN observes 24 months and predicts **one**; our USHCN sequences span a median 199 months, so we observe ∼100 and predict ∼100. **We are forecasting about a hundred months ahead where the literature forecasts one.** Activity runs the other way: upstream observes 3000 ms and predicts 1000, while our 30-row chunks last a median 946 ms, so we observe ∼473 and predict ∼473. This is not a fairness problem — every model in Table 28 runs the same protocol, so the comparison is internally consistent — but it is a comparability problem, and a sharper one than item (vi) of Appendix B admits, which disclaims only the 12→36/24→24/36→12 splits and so reaches neither of these. It also interacts with our own horizon confound ($\rho_s(1/\rho_n, \text{horizon}) = +0.70$): a hundred-month query window stretches that lever much further than a one-month window would, and USHCN is where the floor interval is widest and the density increment is not establishable.

Baseline capacity, measured rather than assumed. Our baselines run below the capacity their authors ship — *but not uniformly, and an earlier version of this paragraph said uniformly*. Read against t-PatchGNN's own per-dataset script rather than its argparse defaults, its `hid_dim` is halved only on PhysioNet and MIMIC (32 against 64) and **matches the authors' value on Activity and USHCN**; what is halved on all four is `node_dim` (5 against 10), and USHCN additionally runs at batch 32 against their 192. Our patching also differs (`n_patches`= 4, `patch_len`= 12 against their `patch_size`/`stride` of 8 on PhysioNet). Elsewhere: mTAND at `latent_dim` 16/`embed_time` 64 against 32/128, Latent-ODE at `gru_hidden` 32 against 100 — while BandRate uses $d_{\text{model}} = 64$. We had no rationale for that and do not offer one now. Re-running the strongest PhysioNet baseline at its authors' setting, changing nothing else (`results/v2_capacity_check`, 3 seeds), makes it

better: test MSE $11{,}009 \pm 204$ against the $11{,}398 \pm 699$ our board reports, so the gap to BandRate on this dataset widens from 13.6% to 17.7%. On MIMIC, the other dataset where the authors' script uses `hid_dim` 64, the same check at the scale Table 28 reports (z-scored) moves nothing: 0.6260 ± 0.0068 at the authors' setting against 0.6238 ± 0.0024 at ours, a difference well inside either seed spread — though t-PatchGNN is not MIMIC's strongest baseline (QuITE is, at 0.6123), so this leaves MIMIC's headline gap untested. Table 28 should therefore be read as understating t-PatchGNN on PhysioNet and not on MIMIC; the other baselines have not been re-run at their authors' capacity.

Budget sensitivity. To check that the 10-epoch budget does not distort rankings, we retrained BandRate and the strongest PhysioNet baseline (T-PatchGNN) for 30 epochs (3 seeds). Both improve — test MSE $12627 \to 10115$ (BandRate) and $11512 \to 9646$ (T-PatchGNN) — but the ordering is preserved (T-PatchGNN first, BandRate second) and the gap *narrows* from $+9.7\%$ to $+4.9\%$, so the shorter budget is if anything slightly conservative for BandRate rather than favorable.

T Computational cost (E9)

Provenance note: these are single-run wall-clock measurements taken during development; unlike every other table in this paper they are not backed by a released board in `results/`, and should be read as indicative rather than reproducible.

Table 32: Parameters and inference latency (batch 32, $N{=}41$ channels, forward pass, NVIDIA TITAN Xp, mean of 20 runs after warmup).

Model	Params	ms/batch↓
Latent-ODE	18.3K	34.9
GRU-D	25.2K	34.8
mTAND	56.0K	30.7
QuITE	72.1K	**16.8**
BandRate	117K	41.4
T-PatchGNN	159K	16.8

BandRate is mid-sized (117K parameters) and about $2.5\times$ slower per batch than QuITE or T-PatchGNN; these timings are a single development measurement on one machine, batch size and channel count, with no stored artefact, and say nothing about scaling in sequence length, channels or queries. Its forward latency (41 ms) is dominated by the SC-INR decoder's K SIREN evaluations at every query time; the cross-channel block is not a bottleneck at IMTS channel counts, so we keep full attention rather than a linear approximation (Katharopoulos et al., 2020) (the cost accounting of Appendix T applies, and a full 5-seed board completes in under one GPU-hour). The conformal layer adds negligible cost (a single quantile per density bin computed once on the calibration split).

U Full ablation table (E8)

V Accuracy significance test (W-C)

The seed-level test finds the gap to the strongest baseline insignificant on Activity and PhysioNet, and significant but under 5% on USHCN and MIMIC. The example-level Wilcoxon (per-target, matched order) rejects equality on *all* four datasets ($p < 10^{-17}$) simply because the test sets are large (10^4–10^6 targets) and detect even sub-percent systematic differences; we report it for completeness but read accuracy comparability from the seed-level test, the standard for "is method A better than B" under training stochasticity. Strongest baseline is chosen per dataset as the lowest-mean competitor; checkpoints re-evaluated on the held-out test split via `scripts/significance_test.py`.

W Recent baselines via a shared anchor (APN, Hi-Patch)

Table 28 predates the newest IMTS forecasters. To compare against them directly we ported BandRate into APN's public benchmark harness (Liu et al., 2026) — our model is imported unchanged and

Table 33: Ablations on the held-out **PhysioNet test** split (10 epochs, 3 seeds; mean±std). PhysioNet is the ablation dataset: Activity's leakage-free sequence-level group split leaves only 2/3 calibration/test sequences, too few for a stable ablation. $\text{dvcorr}_{\text{wc}}$ = *within-channel* density–variance correlation (the headline NABU metric). At this budget most point-MSE gaps are within seed noise, with one exception we flag rather than bury: removing NABU *lowers* MSE by 23.0% ($14979 \to 11537$; equivalently NABU costs $+29.8\%$ relative to the pooled head, the direction quoted in the body; 3/3 seeds, paired p=0.038), so NABU buys uncertainty ordering at a real accuracy cost. The other reproducible signal is calibration — removing NABU drives dvcorr negative and removing cross-channel mixing collapses it.

Ablation	PhysioNet MSE	Δ	$\text{dvcorr}_{\text{wc}}$
full (K=3)	14979±1216	—	+0.65
no cross-channel	13514±792	−9.8%	+0.06
no NABU (pooled head)	11537±595	−23.0%	−0.11
K=1	14473±5103	−3.4%	+0.43
K=2	12843±890	−14.3%	+0.68
K=4	12323±1411	−17.7%	+0.69
depth 2	13732±3341	−8.3%	+0.66
δ_{band}=0.5	12456±863	−16.8%	+0.35

Table 34: Paired significance of the test-MSE gap between BandRate and the strongest baseline per dataset (lowest-mean competitor). Seed-level: paired t-test over per-seed mean MSE — the test that respects training/seed variance. Example-level: paired Wilcoxon over per-target squared errors (matched order; large n, seed-blind). 5 seeds.

Data	Best base.	Gap	seed-t p	verdict
Activity	QuITE	+1.6%	0.21	n.s.
PhysioNet	T-PatchGNN	+9.7%	0.15	n.s.
USHCN	mTAND	+4.6%	0.001	base. ($< 5\%$)
MIMIC	QuITE	+2.6%	0.007	base. ($< 5\%$)

wrapped behind that harness's interface, so it is trained and scored by the *same* loop, data pipeline, early-stopping rule and metric code as the baselines — and ran APN (AAAI 2026), t-PatchGNN and BandRate on PhysioNet'12 under that harness's protocol (observe-36/predict-3, `tsdm` per-variable normalization, three seeds, identical seed set per model).

Table 35: **Same-run head-to-head** in APN's harness (PhysioNet'12, obs-36/pred-3, 3 seeds, mean±std test MSE). Paired t-test against BandRate over the shared seeds.

Model	test MSE ↓	gap vs BandRate	paired p
BandRate	$\mathbf{0.3065 \pm 0.0006}$	—	—
t-PatchGNN	0.3104 ± 0.0016	+1.3%	0.053
APN	0.3134 ± 0.0007	+2.2%	0.004

Table 35: under a genuinely common protocol BandRate has the lowest MSE, winning on $3/3$ seeds against both — by $+2.2\%$ against APN (p=0.004) and $+1.3\%$ against t-PatchGNN (p=0.053), p-values that, like all of ours, describe one comparison and are not corrected for the post-hoc family (Section 4). Three honest qualifications. First, the effect sizes are *small* (1–2%); this says BandRate is competitive with 2024–2026 methods, not that it is a decisively better forecaster. Second, this is one dataset under one horizon setting, so it does not override Table 28, where on our own protocol BandRate trails the best baseline on each dataset by 1.6–9.7%; the two protocols differ in split, horizon and normalization and we report both rather than the more favourable one. Third, no method was re-tuned inside this harness: the baselines use the hyperparameters their authors shipped *for this harness*, while BandRate uses the configuration it has had since it was first wired — which, as Appendix S records, was never tuned on any split (Appendix S), which if anything disadvantages us. A Hi-Patch (Luo et al., 2025) run in the same harness reached a comparable validation MSE (≈ 0.30) but did not converge within our compute budget, so we omit it rather than report a truncated number.

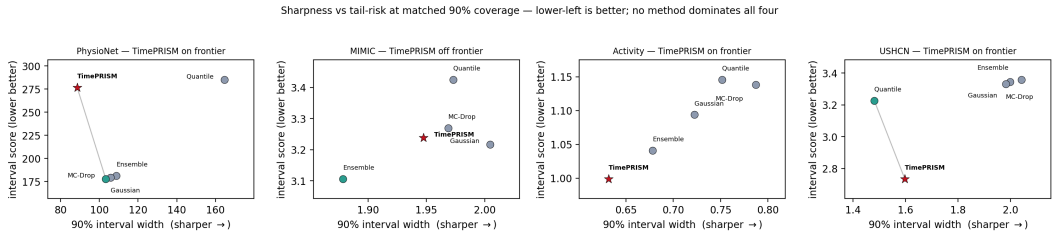


Figure 5: Sharpness (90% interval width, x) vs tail-risk (interval score, y) at matched conformal coverage; lower-left is better. BandRate ($\star$) is Pareto-optimal on Activity/USHCN, sharpest-but-tail-risky on PhysioNet, and dominated on MIMIC — no method wins all four. Mean over 3 seeds; grey line is the lower-left frontier.

The point of this table is not a leaderboard win — it is that the accuracy safety-check survives contact with the newest baselines under a protocol we did not design.

X NABU AS A PORTABLE PLUG-IN

Is NABU's density-conditioning tied to the SC-INR backbone, or a transferable module? We bolt a single density-conditioned variance term — $\log \sigma^2 \mathrel{+}= \mathrm{softplus}(\gamma_n) \cdot z(-\log \rho_n(\tau))$, with $z(\cdot)$ a standardized log-density — onto a plain GRU-D encoder, and compare it to the same encoder with a density-agnostic Gaussian head under identical NLL training (PhysioNet, 10 epochs). The plug-in flips the within-channel density–variance correlation from -0.01 to $+0.82$: NABU's density-conditioning *transfers* to a non-SC-INR backbone. Sharpness does not improve, and on two of three probabilistic metrics it degrades: interval score $182 \to 188$, width $110.2 \to 112.7$ ($+2.3\%$) and CRPS $19.7 \to 21.4$ ($+8.4\%$), at matched coverage (0.904 vs 0.905) — a single scalar density term moves the variance *shape* but the sharpness gain needs the full band-structured identifiability ratio $r_{n,k}$, not a pooled density scalar. We release the head (`scripts/nabu_plugin_demo.py`) as a reusable module for other IMTS backbones.

Y SHARPNESS–TAIL-RISK PARETO FRONTIER

Figure 5 plots every UQ method at matched 90% conformal coverage on the two axes that jointly matter: interval *width* (sharpness) and *interval score* (Winkler; width plus a tail-miss penalty). It makes the "no method dominates all four" claim of Section 4.2 visual and falsifiable. On Activity BandRate is Pareto-optimal (tightest *and* best-scored); on USHCN it attains the best interval score; on PhysioNet it is the sharpest (nothing beats its width) but the frontier is steep — the simpler Gaussian/ensemble heads trade $\sim 18\%$ width for a much lower interval score, the heavy-tailed tail-risk we state plainly; and on MIMIC it is off the frontier, dominated by a GRU-D deep ensemble whose agreeing members give a tight, well-scored band. This is the honest decision picture: BandRate's density-adaptive variance buys the narrowest *typical* interval, Pareto-optimally so on the light-tailed datasets, at a tail-risk cost on the heavy-tailed clinical ones.

Z RISK AT DECLARED RETENTION LEVELS

The floor is an area: normalized risk–coverage averaged over abstaining 0–90%. A review asked for the decision a practitioner actually faces. Keep a fixed share of the forecasts, under a stated loss, and report the risk of what is kept. We report $1 - \bar{\ell}_{\text{retained}}/\bar{\ell}_{\text{all}}$ at retention 90, 80 and 50% for three rankings, all on the same evaluation halves. The first is the cross-fitted (channel, horizon decile, density tercile) lookup of Appendix . The second is the variance our own head predicts, in the loss's units ($\hat{\sigma}^2/d_c$). The third recalibrates that variance at the lookup's resolution, a fitting-half lookup on channel, horizon and the tercile of $\hat{\sigma}^2/d_c$, so that the comparison is between what each input carries rather than between calibrations. An earlier version of this table ranked by the raw variance against the standardized loss, which compares channels in mismatched units. We use two losses. *Standardized* divides each channel's squared error by that channel's fitting-half mean,

which weights channels equally and is the loss of the floor. *Native* is the raw squared error, which lets channels with large physical units dominate. The checkpoints are the five NABU runs per dataset. The intervals resample test sequences jointly over seeds, conditional on the fitted models (`scripts/retention_risk.py`, `results/v2_retention_risk`).

Table 36: Risk reduction at fixed retention; positive is better. PhysioNet raw scale, MIMIC-III z-scored, $seedseach, 00$ and $00jointsequence - bootstrapdraws.Aligned''ranksby^2/d_c$; recal" by a fitting-half lookup on (channel, horizon, tercile of $\hat{\sigma}^2/d_c$), the density lookup's resolution. The minus" rows are paired differences on the same draws. Under the native loss the aligned and raw variance coincide by construction.

dataset, loss, ranking	keep 90%	keep 80%	keep 50%
PhysioNet, standardized, density lookup	$+0.163$ $[+.149, +.176]$	$+0.245$ $[+.225, +.263]$	$+0.412$ $[+.394, +.427]$
PhysioNet, standardized, NABU aligned	$+0.067$ $[+.058, +.076]$	$+0.102$ $[+.083, +.122]$	$+0.188$ $[+.161, +.213]$
PhysioNet, standardized, NABU recal	$+0.123$ $[+.115, +.130]$	$+0.195$ $[+.184, +.206]$	$+0.371$ $[+.357, +.385]$
PhysioNet, standardized, aligned minus lookup	-0.096 $[-.113, -.077]$	-0.142 $[-.170, -.114]$	-0.223 $[-.249, -.199]$
PhysioNet, standardized, recal minus lookup	-0.040 $[-.049, -.031]$	-0.050 $[-.064, -.038]$	-0.041 $[-.055, -.026]$
PhysioNet, native, density lookup	$+0.916$ $[+.826, +.959]$	$+0.976$ $[+.955, +.984]$	$+0.988$ $[+.978, +.992]$
PhysioNet, native, NABU raw	$+0.358$ $[+.302, +.431]$	$+0.530$ $[+.487, +.620]$	$+0.845$ $[+.691, +.863]$
MIMIC-III, standardized, density lookup	$+0.070$ $[+.057, +.095]$	$+0.120$ $[+.096, +.152]$	$+0.222$ $[+.154, +.288]$
MIMIC-III, standardized, NABU aligned	$+0.248$ $[+.192, +.287]$	$+0.351$ $[+.301, +.390]$	$+0.609$ $[+.572, +.643]$
MIMIC-III, standardized, NABU recal	$+0.143$ $[+.114, +.175]$	$+0.232$ $[+.198, +.263]$	$+0.473$ $[+.434, +.513]$
MIMIC-III, standardized, aligned minus lookup	$+0.178$ $[+.113, +.223]$	$+0.231$ $[+.163, +.284]$	$+0.387$ $[+.294, +.473]$
MIMIC-III, standardized, recal minus lookup	$+0.073$ $[+.042, +.102]$	$+0.112$ $[+.082, +.140]$	$+0.251$ $[+.191, +.301]$
MIMIC-III, native, density lookup	$+0.085$ $[+.057, +.111]$	$+0.133$ $[+.084, +.175]$	$+0.278$ $[+.144, +.380]$
MIMIC-III, native, NABU raw	$+0.252$ $[+.180, +.331]$	$+0.356$ $[+.292, +.432]$	$+0.645$ $[+.594, +.695]$

Three readings. *(i) The loss decides the answer.* On raw-scale PhysioNet, dropping the 10% the lookup ranks worst removes 92% of native risk, because a few large-unit channels carry almost all of it. Under the standardized loss the same rule removes 16%. This is the channel-units confound of Section 4.1.7 seen at a single retention level, and a user whose costs really are in native units would face the first number. *(ii) On PhysioNet an S-only lookup beats our head* at every retention level, by 0.10 to 0.22 against the aligned variance and by 0.04 to 0.05 against the recalibrated one, all paired intervals excluding zero; it also wins under the native loss. Aligning the units made our head *worse* here, not better. *(iii) On MIMIC-III the order reverses*: the aligned variance beats the lookup by 0.18 to 0.39, and still by 0.07 to 0.25 recalibrated. That head reads the observed values as well as the observation set, so this does not credit density (Section 4.1.7). Neither dataset licenses a general rule, and a practitioner should run both rankings under the loss that matches their costs. Under the native loss the aligned and raw variance are the same score; NABU's variance has many exact ties on PhysioNet, so its native rows there move by up to 0.05 between runs through random tie-breaking.

A STANDALONE CEILING ESTIMATOR

`tools/imts_uq_ceiling.py` computes both measurements from a table of predictions, with no dependency on BandRate, PyTorch or any IMTS framework: it takes targets y, predictions $\hat{y}$, channel ids, query times τ, a local density ρ and — required — a *sequence* id naming the independent unit (patient, station, recording). A review of the previous release found two defects, and this version is rebuilt around them; both counterexamples are now permanent tests (`tests/test_tool_estimators.py`).

The ceiling it reports is $\sqrt{s_T}$, and it says which ceiling that is. The earlier tool estimated s from the contrast between the sparsest and densest density terciles. That contrast is zero whenever the signal lives in the middle tercile: with $\rho \in \{1, 2, 3\}$ and $e^2 = 1 + \mathbf{1}\{\rho = 2\}$ a density-only function tracks e^2 perfectly while the contrast, and the tool's upper limit, read 0. It also advised against adopting the method family when the number was small, which does not follow. The tool now estimates the correlation ratio s_T of $T = (\text{channel}, \text{horizon decile}, \text{density tercile})$ with a sequence-level split-half U-statistic, which returns 1.000 on that example, and prints that $\sqrt{s_T}$ is the ceiling for a head reading only T and a *lower* bound on the ceiling for heads reading the full S.

The floor it reports is cross-fitted. The earlier `selective_risk_floor` fitted the $(\text{channel}, \text{horizon}, \text{tercile})$ lookup, the per-channel standardisation and the bin edges on the same points it scored, and broke ties by array position. On errors independent of everything it reported a floor of 0.23 on each of 30 datasets. It now fits all three on one half of the sequences and scores

the other half, in both directions, with the random tie-breaking of the paper's own risk–coverage code, which the tool carries a copy of and a test pins to. On the same 30 datasets it returns a mean of -0.001, with a one-sided 95% lower limit above zero on 1 of them; on errors that scale with $1/\rho$ the lower limit clears zero. The floor is a reduction in normalised risk–coverage area, averaged over abstaining 0–90%, not a risk reduction at any single retention level.

Coverage against a known truth, and where it fails. An earlier version of this study used the tool's own estimate on a large draw as the "truth"; a review pointed out that is not a known value. The generator's population s_T has a closed form (squared error $W(1+a\mathbb{1}\{\rho < 0.45\})$ with ρ persistent within a sequence and W log-normal with a sequence effect; `scripts/ceiling_coverage.py`), and we now compare against it. The closed form is for the population density tercile, not the finer indicator $\mathbb{1}\{\rho < 0.45\}$; a Monte Carlo check agrees to within 1% and a test pins it. Each world has 400 replicates of 200 sequences, 200 bootstrap draws each; the tool re-estimates the tercile edges every time, as it must on real data. The tool's partition and statistic are those of the estimator behind the paper's s_T numbers, and a test pins that equality. Binomial standard errors are about ± 0.01 to ± 0.02.

Table 37: Coverage of the tool's s_T interval against the population value. `results/v2_ceiling_coverage`.

world	targets per cell	true s_T	mean estimate	two-sided covers	upper $\geq$ truth
no signal, heavy tail	75	0	-0.0003	0.998	0.985
weak signal, heavy tail	75	0.0033	0.0032	0.998	0.993
strong signal, heavy tail	75	0.0151	0.0153	0.980	0.973
moderate tail	75	0.0145	0.0120	0.988	0.955
moderate tail, unbalanced clusters	75	0.0145	0.0127	0.965	0.935
light tail	75	0.0532	0.0455	0.875	0.835
light tail, dense cells	375	0.0532	0.0487	0.897	0.875
sparse cells, balanced	20	0.0145	0.0002	0.605	0.375
sparse cells, unbalanced	20	0.0145	0.0002	0.565	0.385

Three readings. *(i)* With heavy tails, which is what squared forecast errors have on every benchmark here, the interval covers at or above nominal and the mean estimate is close to the truth. *(ii)* With light tails and a strong signal the estimate runs 9–15% low and the one-sided upper limit covers only 84–88%. We traced the bias by replacing estimated ingredients with their population values (`scripts/coverage_bias_cause.py`, 300 replicates each). Population tercile edges remove it ($-14.4\% \to -0.8\%$ at 75 targets per cell, $-7.6\% \to -0.1\%$ at 375); a population denominator does not. Cutting terciles from the data blurs a sharp density effect, and the blur shrinks as cells grow. The same check changes reading (i): in the heavy-tailed world the edge blur is also present (-15.1% with a population denominator), and it is offset by the sample variance understating a heavy-tailed denominator ($+18.9\%$ with population edges), for a net $+3.0\%$. Coverage held there by cancellation, not because the estimator is unbiased, so we do not claim it transfers to other tail weights or cell sizes. The benchmarks' cells are larger than any simulated here (median 684 on PhysioNet), where the edge blur is smallest. *(iii)* Below 50 targets per (channel, horizon) cell the estimators do not read density at all, by design, so they estimate the ratio of a coarser T — still a lower bound on s, but not the s_T of the full T, and its upper limit does not bound that. This is a failure of the target, not of the interval, and the tool now warns when more than 5% of targets fall in such cells. On the benchmarks the median cell holds 684 targets on PhysioNet, 244 on USHCN, 224 on Activity and 76 on MIMIC-III, and the share of targets in cells below 50 is 0.05%, 0, 0 and 3.8% (`scripts/cell_occupancy.py`). MIMIC-III's cells are as small as those of the light-tailed worlds that undercovered, so its upper limits carry the most doubt.

A channel absent from the fitting half. The floor standardizes each channel's errors by its mean on the fitting half. A review found that a channel absent from that half fell back on its own evaluation-half mean, which put evaluation errors into the loss. It now falls back on the fitting half's overall mean, and a test pins this. The paper's floors come from a different script and never took this branch; the six-model floors of Table 4 were recomputed with the corrected tool.

Cost. On synthetic tables with 40 channels, one CPU process with two threads takes the following wall time; peak Python-heap memory stays under 110 MB. Time grows linearly in targets times bootstrap draws. The machine was shared and loaded during the measurement, so these are upper figures (`scripts/tool_cost_benchmark.py`, `results/v2_tool_cost`).

targets	ceiling, 50 draws	ceiling, 200 draws	floor, 50 draws	floor, 200 draws
10^5	9 s	27 s	9 s	29 s
3×10^5	23 s	79 s	21 s	83 s
10^6	69 s	273 s	92 s	329 s

Every interval resamples sequences. Targets from one sequence are dependent, so the split-half halves and the bootstrap draws are both drawn over sequences; the previous tool resampled strata, which treats one patient's points in different strata as independent and was optimistic by about $2.5\times$ on our benchmarks.

What it is for, and what it is not. Both numbers need realized errors, so they come after a model is trained, not before. $\sqrt{s_T}$ says how much a head reading only channel, horizon and a density tercile can track; heads reading gap-since-last-observation, burst structure or the observed values are bounded by larger quantities the tool does not estimate, so a small value is not a reason to abandon the family. The floor says how much an abstention rule built from those same features can buy under a declared loss. Neither is a go/no-go test, and the tool no longer offers one.