PREDICTIVE UNCERTAINTY READS INSULIN RESISTANCE:

A LATENT-DYNAMICS READOUT FROM FROZEN CGM FORECASTERS AND A CONTROLLED FEEDBACK SYSTEM

Anonymous authors
Paper under double-blind review

ABSTRACT

Probabilistic time-series forecasters output uncertainty, but it is usually treated as an error bar rather than a signal. We ask whether predictive uncertainty carries information about a latent property of the data-generating process, using a frozen continuous glucose monitoring forecaster whose representation tracks gold-standard insulin resistance when read by a supervised probe. We define *relative forecast uncertainty* (RFU), the predictive scale normalized by recent within-person glucose variability, and find that it discriminates insulin resistance defined by the steady-state plasma glucose (SSPG) reference test in pooled gold-standard cohorts (AUROC 0.78, two-sided permutation $p{=}0.0009$, $n{=}47$); binary discrimination remains above chance after linear adjustment for glucose level, variability, and cohort (cross-fitted AUROC 0.70 $[0.53, 0.85]$, nominally significant). The sign is clinically counterintuitive: more insulin-resistant subjects receive *lower* relative uncertainty, and the forecaster has greater realized skill relative to persistence in this group. The readout is sensitive to within-window time shuffling, overlaps substantially with long-range persistence, and is reproduced by four frozen forecasters, including a general-purpose time-series model applied zero-shot (subject rankings agree at 0.90 to 0.96), with protocol-dependent magnitude. RFU aligns with the representation-based readout it was derived to explain, and on a controlled feedback system with subject-level mean and variance matched it ranks a held-out latent gain ($+0.90$). RFU thus provides a scalar readout of SSPG-defined insulin-resistance status without label-based fitting; its physiological interpretation and the added value of the learned scale beyond normalization remain unresolved.

1 INTRODUCTION

Probabilistic forecasters emit an estimate of their own uncertainty, almost always treated as an error bar rather than as a signal in its own right. For a time series generated by a regulated physiological system, however, predictability depends on the system, and a forecaster's uncertainty is a learned estimate of it. Can the predictive uncertainty of a self-supervised forecaster carry information about latent properties of the process that generated the data?

We study this question in glucose forecasting. Recent work (Anonymous, 2026b) reports that a continuous glucose monitoring (CGM) forecaster, trained only to predict glucose, yields a frozen representation from which a linear probe reads gold-standard insulin resistance, measured by the steady-state plasma glucose (SSPG) reference test, in a cohort it never saw, with the signal residing in glucose dynamics rather than level. A model optimized only for prediction thus carries a clinically meaningful latent phenotype (Reaven, 1988; Tabák et al., 2009) it was never labelled to represent; which quantity in the model carries it remains open.

Two clues constrain a candidate. The object is a probabilistic forecaster, so the natural carrier of prediction-related structure is its uncertainty; and the signal lives in dynamics, not level, so the candidate should be normalized by the signal's own variability. Both point to the same scalar, the predictive scale divided by recent within-person glucose variability, which we call *relative forecast uncertainty* (RFU; Section 3). We test, on pooled gold-standard cohorts, whether RFU carries the readout and what it could be confounded with (Figure 1).

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.

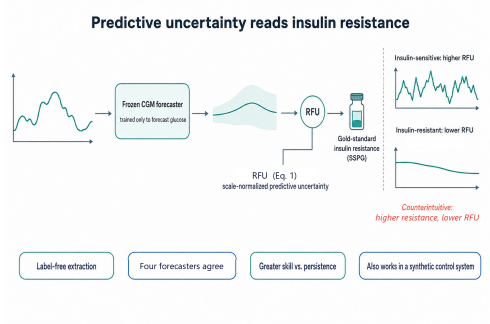


Figure 1: **Overview.** A frozen probabilistic continuous glucose monitoring forecaster, trained only to predict glucose, yields a scale-normalised predictive uncertainty (relative forecast uncertainty, RFU) that discriminates gold-standard insulin resistance without labels. The direction is clinically counterintuitive: more insulin-resistant glucose receives lower relative uncertainty. RFU aligns with the supervised readout of the model's 320-dimensional representation, and the readout is reproduced by four frozen forecasters, including a general-purpose model applied zero-shot.

RFU discriminates SSPG-defined insulin resistance (AUROC 0.78, two-sided permutation $p{=}0.0009$, $n{=}47$), and binary discrimination remains above chance after linear adjustment for glucose level, variability, and cohort (nominally significant; Appendix C); more insulin-resistant subjects receive *lower* relative uncertainty, and the forecaster has greater realized skill relative to persistence in this group. The readout is sensitive to within-window time shuffling, overlaps substantially with long-range persistence, and is reproduced by four frozen forecasters, including a general-purpose time-series model applied zero-shot. On a controlled feedback system with no glucose, with subject-level mean and variance matched, the same statistic ranks a held-out latent controller gain ($+0.90$). Secondary analyses, including the adjusted continuous gradient and the added value of the learned scale beyond normalization, are directional but inconclusive at this sample size, and we report them as such.

Contributions.

1. RFU, a label-free scalar that for a window-standardizing forecaster is computed from its own normalized predictive scale, discriminates SSPG-defined insulin resistance in pooled gold-standard cohorts and remains above chance after linear adjustment, with lower uncertainty in the resistant group.
2. The signal is sensitive to time shuffling, overlaps with long-range persistence, accompanies better realized skill relative to persistence, and tops an ablation ladder whose learned-scale increment is not established.
3. The readout is reproduced across four frozen forecasters, including a general-purpose model applied zero-shot (ranking agreement 0.90 to 0.96), with protocol-dependent magnitude.
4. RFU aligns with the supervised representation-based readout it was derived to explain, and ranks a latent gain in a controlled feedback system, consistent with an idealized monotonicity result.

2 Related work

CGM foundation models and glucose representation learning. Relevant work spans generative and transformer foundation models and joint-embedding predictive approaches (Lutsker et al., 2026; Luo et al., 2025; Lu et al., 2025; Muhammad et al., 2026), distributional representations of glucose profiles (Matabuena et al., 2021; 2025), general-purpose time-series foundation models (Ansari et al., 2024; Das et al., 2024; Jain & Sen, 2026; Goswami et al., 2024), and forecasting benchmarks (Sergazinov et al., 2024); recent CGM foundation models also report insulin-resistance phenotypes from label-fitted probes (Li et al., 2026). Our starting point is a reported property of one such model (Anonymous, 2026b): a linear probe on a frozen glucose forecaster's representation, fitted to a fasting-proxy label in one cohort, reads gold-standard insulin resistance in another. We

identify a scalar derived from predictive uncertainty and context variability that aligns with that supervised readout, recomputed here (Section 4.4), and discriminates SSPG-defined insulin-resistance status across four frozen forecasters.

System identification and latent-parameter inference. Given a model structure, Bayesian filters, system identification, and neural latent-variable models estimate hidden states and parameters (Kalman, 1960; Gordon et al., 1993; Ljung, 1999; Chen et al., 2018; Rubanova et al., 2019). RFU posits no physiological model and fits no parameter estimator; the forecasters are trained only to predict glucose, and RFU uses no insulin-resistance label. A physiological simulator enters only as a tool in Section 4.5.

Predictive uncertainty, dynamics measures, and probing. Probabilistic forecasters model predictive distributions (Salinas et al., 2020), and calibration and proper scoring rules assess whether that uncertainty is trustworthy (Guo et al., 2017; Gneiting & Raftery, 2007); we treat the uncertainty as a signal about the data-generating physiology and use calibration as a control. Long-range persistence (Peng et al., 1994) and regularity (Pincus, 1991; Richman & Moorman, 2000) are natural comparators, and lower CGM complexity (sample entropy) has been associated with insulin resistance measured by HOMA-IR (Crenier et al., 2016), so the direction of our readout has precedent; we benchmark against these measures directly (Table 2) and focus our contribution on frozen forecasters' own outputs, gold-standard validation, and consistency across forecasters. Linear probes ask what a representation encodes (Belinkov, 2022; LeCun, 2022; Assran et al., 2023; He et al., 2022); we ask what a forecaster's own uncertainty output carries.

3 METHOD

Frozen probabilistic forecasters. Our primary CGM forecaster (Anonymous, 2026a) was pretrained self-supervised to predict glucose on data from about $1{,}000$ individuals with type 1 diabetes (Cinar & Maleshkova, 2025). Its convolutional–attention architecture outputs a Student-t predictive distribution for each of the next $H{=}12$ five-minute steps, with mean $\hat{\mu}_h$, scale $\hat{\sigma}_h$, and degrees of freedom $\hat{\nu}_h$ (> 2 by parameterization). Three further forecasters follow the identical protocol: a patch Transformer (PatchTST) and a neural ordinary differential equation model (NeurODE), trained on the same corpus and also emitting a Student-t scale, and TimesFM-3, a general-purpose time-series foundation model applied zero-shot (Jain & Sen, 2026), whose uncertainty is given as nine quantiles; for it we take $\hat{\sigma}_h$ as the 10th-to-90th percentile width divided by 2.563, the normal-equivalent standard deviation. Each numerator is thus the model's own uncertainty summary (variants in Appendix F). All forecasters are frozen for every RFU computation; the only retrained models are the matched controls of Section 4.2 (Appendix H).

Relative forecast uncertainty (RFU). For a context window w of subject i (the preceding L glucose samples) the forecaster produces predictive scales $\hat{\sigma}_h$ over the next hour. The window's relative uncertainty is the horizon-averaged ratio of predictive scale to the window's own glucose standard deviation, and a subject's RFU is the logarithm of its mean over all qualifying windows:

$$r_{i,w} = \frac{1}{H}\sum_{h=1}^{H}\frac{\hat{\sigma}_{i,w,h}}{\mathrm{SD}(g_{i,w})}, \qquad \mathrm{RFU}_i = \log\Big(\frac{1}{W_i}\sum_{w=1}^{W_i} r_{i,w}\Big). \tag{1}$$

r is dimensionless, and $\exp(\mathrm{RFU}_i)$ is the model's mean uncertainty as a fraction of the subject's own variability. Windows lie on a 5-minute grid with a 2-hour stride; a window qualifies when all $L{+}H$ samples are valid (40–400 mg/dL, no gaps); every qualifying window is used, with no per-subject cap, and subjects with fewer than three windows are excluded. For the window-standardizing forecasters $r_{i,w}$ equals the mean of the model's own normalized-space scale head up to a constant (Appendix A). Higher insulin resistance corresponds to lower RFU. All analyses use this logged definition; an alternative aggregation (the median over windows of the log root-mean-square ratio) and per-subject window caps are protocol variants in Appendix F.

Cohorts, labels, and representation probe. The reference measure is SSPG from the insulin-suppression test (Greenfield et al., 1981; Pei et al., 1994; DeFronzo et al., 1979), available in two cohorts (Metabolic Subphenotypes and Hall) (Metwally et al., 2025; Hall et al., 2018; Park et al.,

2025); subjects with a valid SSPG and at least three qualifying windows form the pooled set (n=47 at 24 hours, n=32 at 48 hours). The binary endpoint is insulin-resistance status as each cohort defines it (the published SSPG classification in Metabolic Subphenotypes; $\mathrm{SSPG} \geq 150$ mg/dL in Hall), giving 16 of 47 resistant (single-threshold sensitivity in Appendix C). For the alignment analysis (Section 4.4) a ridge probe reads insulin resistance from the primary encoder's pooled 320-dimensional representation, fitted on AI-READI windows (AI-READI Consortium, 2025; 2024; Owsley et al., 2025) against log HOMA-IR (Matthews et al., 1985; Wallace et al., 2004) as in the study we seek to explain (Anonymous, 2026b); the probe is supervised by that label, whereas RFU uses no insulin-resistance label. The disposition index (DI) (Kahn et al., 1993) serves in Section 4.5 as a measure of β-cell function; cohort counts are detailed in Appendix C.

Analysis hierarchy and statistics. The 24-hour context (L=288) is primary and 48 hours (L=576) a sensitivity analysis. Binary discrimination is the area under the receiver operating characteristic curve (AUROC) when ranking pooled subjects by $-\mathrm{RFU}$, with 95% subject-bootstrap intervals (2,000 resamples (Efron & Tibshirani, 1993)) and a within-cohort label permutation test that is two-sided in orientation ($\max(\mathrm{AUROC}, 1 - \mathrm{AUROC})$; 20,000 permutations). Adjusted binary discrimination residualizes cohort-standardized RFU on cohort-standardized mean glucose, glucose variability (coefficient of variation), and a cohort indicator and scores subjects by the negative residual, cross-fitted over five cohort-stratified folds with all constants estimated on the training folds (fixed fold assignment, seed 0, plus the mean over 50 assignments); its permutation test asks whether residualized RFU still discriminates status, not whether RFU adds to a model of the covariates (Appendix C). Continuous associations use cohort-internal percentile ranks: Spearman correlation with SSPG, and partial correlation after residualizing both RFU and SSPG on mean glucose and variability with a single pooled regression, tested by Freedman–Lane permutation (Freedman & Lane, 1983) (two-sided; 20,000 permutations). The binary AUROC is the primary endpoint and the adjusted continuous association corroboration. Sign convention: higher resistance, lower RFU, so a negative correlation or an AUROC above 0.5 reads insulin resistance.

4 Results

4.1 A frozen forecaster's relative uncertainty reads gold-standard insulin resistance

We first ask whether RFU carries information about insulin resistance as measured by SSPG. Because gold-standard cohorts are individually small, we pool Metabolic Subphenotypes and Hall (Appendix C). The 24-hour context is primary and 48 hours a sensitivity analysis.

Binary discrimination. Ranked by $-\mathrm{RFU}$, this single scalar separates insulin-resistant from insulin-sensitive subjects with an area under the receiver operating characteristic curve (AUROC) of 0.780 (pooled n=47, 16 resistant; 95% bootstrap CI $[0.64, 0.90]$; permutation p=0.0009; Figure 2a; 0.772 with cohort-internal standardization). Discrimination remains above chance after linear adjustment for mean glucose, glucose variability, and cohort (5-fold cross-fitted AUROC of the negative residual: fixed fold assignment 0.70, 95% CI $[0.53, 0.85]$, permutation p=0.029; mean over 50 splits 0.71); this adjusted test is nominally significant and does not survive family-wise correction (Appendix C).

Continuous gradient and direction. RFU also ranks SSPG as a continuous variable (Spearman $\rho = -0.363$ on cohort-internal percentiles; n=47; permutation p=0.014; Figure 2b). The direction may appear clinically counterintuitive, but has precedent in CGM complexity studies (Crenier et al., 2016): *higher* insulin resistance goes with *lower* uncertainty assigned by the frozen forecaster relative to observed glucose variability; whether this corresponds to lower realized forecast error is addressed in Section 4.2. After partialling out mean glucose and variability the gradient keeps its sign but is no longer conclusive (partial $\rho = -0.258$, 95% CI $[-0.52, +0.04]$, p=0.092). Binary discrimination is our primary endpoint; the adjusted continuous association remains inconclusive.

Sensitivity, pooling, and claim boundary. At 48 hours, in the n=32 subjects with a complete 48-hour context, the estimates are larger: the binary AUROC is 0.840 (95% CI $[0.69, 0.95]$; cross-fitted adjusted 0.84 $[0.59, 0.95]$, p=0.0008; mean over 50 splits 0.81) and the adjusted gradient is

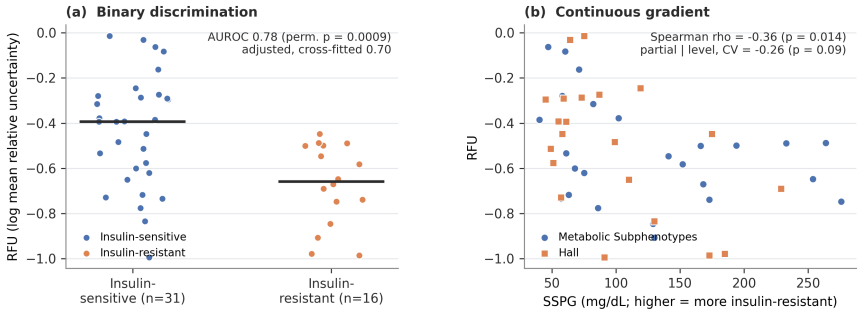


Figure 2: **A frozen forecaster's relative uncertainty discriminates gold-standard insulin resistance** (pooled Metabolic Subphenotypes + Hall, main protocol, 24-hour context, n=47). **(a)** RFU by insulin-resistance status (bars: medians); AUROC 0.78. **(b)** RFU versus SSPG by cohort (Spearman $\rho = -0.36$). Higher insulin resistance corresponds to lower RFU.

-0.437 ($[-0.71, -0.07]$, p=0.018). We report 48 hours as a sensitivity analysis, not as a context effect, because the two samples are not the same subjects. Pooling is supported in Appendix C (per-cohort AUROC 0.750 and 0.824; cohort indicator in the adjusted analyses). Binary discrimination of SSPG-defined insulin-resistance status is the primary result, and it remains above chance after linear adjustment for glucose level and variability at nominal significance; it does not resolve peripheral versus hepatic contributions (Section 5).

4.2 RFU IS SENSITIVE TO TIME SHUFFLING AND OVERLAPS WITH CLASSICAL MEASURES OF GLUCOSE DYNAMICS

RFU against classical summaries. We benchmark RFU against linear autoregression error, sample entropy, detrended fluctuation analysis (DFA), and lag-1 autocorrelation, each as a partial correlation with SSPG controlling for mean glucose and variability under identical cohort-internal percentile de-batching (Table 2, Appendix F). At 24 hours RFU has the largest adjusted association (partial $\rho = -0.258$) but, like every classical measure, its 95% interval includes zero; at 48 hours RFU (-0.437) and long-range persistence (DFA, $+0.436$) are the two statistics whose intervals exclude zero. The two are strongly coupled (Spearman $\rho = -0.83$ and -0.85 at 24 and 48 hours) and each attenuates when the other is controlled (Appendix F): they carry substantially overlapping information, RFU retaining more at 24 hours. Because the closest classical analogue of RFU is a probabilistic innovation variance (Section 4.6), we also benchmark a fixed-order autoregressive innovation ratio, $\mathrm{SD}_{\mathrm{AR}(3)}/\mathrm{SD}_{\mathrm{context}}$ per window with the same aggregation: it discriminates insulin resistance (AUROC 0.712; partial $\rho = -0.190$), below the neural RFU (0.780; -0.258), so a linear innovation variance already carries part of the signal.

Time-shuffle control. A within-subject *time-shuffle* isolates order from window marginals: permuting the sample order inside each context window preserves that window's mean, standard deviation, and every order statistic exactly while destroying all temporal structure, including within-window diurnal shape. Under this shuffle the trained forecaster's RFU no longer discriminates insulin resistance (AUROC $0.780 \rightarrow 0.456 \pm 0.023$ over three realizations; partial ρ from -0.258 to $+0.194 \pm 0.041$) and its predictive scale increases (median per-window scale $9.9 \rightarrow 29.8$ mg/dL); the same shuffle collapses the 320-dimensional probe's readout (AUROC $0.724 \rightarrow 0.566$; Appendix E). The RFU readout is sensitive to within-window time shuffling.

Lower RFU accompanies better forecast skill relative to persistence. A natural worry is that RFU is phenotype-dependent *miscalibration*: a smaller predictive scale for insulin-resistant subjects without their glucose being forecast any better. We therefore compare *realized* error between groups. Mean absolute error shows no clear difference (9.7 versus 8.8 mg/dL, p=0.71), but relative to each subject's own dynamics the forecaster does better on the resistant group: error normalized by within-window variability is lower (0.47 versus 0.53, p=0.067), and the mean absolute scaled error (MASE; the model's error divided by a persistence forecast's error on the same windows, horizons, and

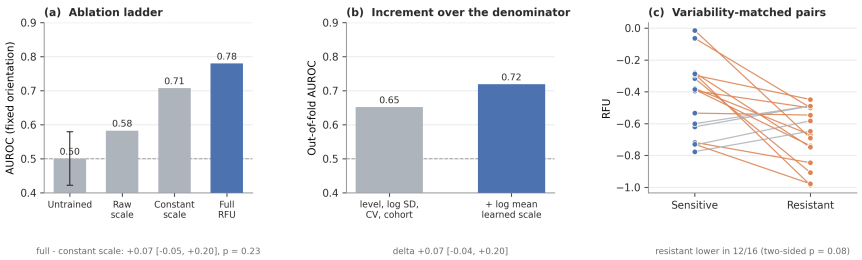


Figure 3: **Which components of RFU carry the signal** (main protocol, pooled n=47; orientation fixed in advance, lower score = more resistant). **(a)** Ablation ladder: untrained forecaster, raw learned scale, constant-scale control, full RFU. **(b)** Out-of-fold logistic AUROC of a covariate model (mean glucose, log window SD, CV, cohort) before and after adding the log mean learned scale. **(c)** Variability-matched insulin-resistant/sensitive pairs: the resistant member has lower RFU in 12 of 16.

targets) differs clearly (0.86 versus 1.08, Mann–Whitney p=0.003). 80% predictive intervals cover 0.81 and 0.86 in the two groups with centred probability-integral transforms, so we observe no under-coverage, though scale miscalibration is not excluded. Across subjects RFU tracks realized normalized error (Spearman $+0.71$; the two share the variability denominator) and realized skill (MASE, $+0.59$): lower RFU is associated with better realized performance relative to persistence.

Which components of RFU carry the signal. Because RFU combines a learned predictive scale with normalization by the window's standard deviation, we ask whether either component alone reads insulin resistance, every statistic oriented by the sign convention of the full RFU, fixed in advance (Figure 3a). An untrained forecaster with the same architecture, denominator, and aggregation reads it at chance on average (AUROC 0.50 ± 0.08 over five initializations); the raw learned scale without normalization at 0.58; the constant-scale control (the logarithm of the mean inverse window standard deviation) at 0.71; the full RFU at 0.78. The increment of the full RFU over the constant-scale control is small and not distinguishable from zero (paired subject-bootstrap ΔAUROC $+0.07$, 95% CI $[-0.05, +0.20]$, p=0.23), and the increment over the raw scale larger but also inconclusive ($+0.20$, $[-0.01, +0.41]$, p=0.065). Conditioning the learned scale on a subject-level summary of the denominator gives the same picture (partial $\rho = -0.28$ $[-0.53, +0.03]$; ΔAUROC $+0.07$ $[-0.04, +0.20]$ in a cross-fitted logistic model; Figure 3b), and among resistant/sensitive pairs matched on variability the resistant member has lower RFU in 12 of 16 (sign test p=0.08; Figure 3c). The full RFU thus has the highest point estimate on this ladder, but the added value of the learned scale beyond normalization is not established at this sample size.

Not a volatility ratio, and not dependent on the explicit volatility input. Because the decoder receives a 6-hour rolling standard deviation, RFU could reflect a handcrafted local-to-global volatility ratio. Computed directly from glucose with the same windows, aggregation, and orientation, the ratio of the last-6-hour to the 24-hour window standard deviation does not discriminate insulin resistance (AUROC 0.51 $[0.35, 0.68]$), is only weakly related to RFU (Spearman $+0.21$), and controlling for it leaves the RFU association unchanged (partial $\rho = -0.255$ versus -0.258). Under matched retraining at a reduced budget (Appendix H), the readout persists with the rolling-volatility input zeroed (AUROC 0.764 versus 0.772 with it; partial $\rho = -0.237$ versus -0.284), with almost identical subject rankings (Spearman $\rho = 0.996$). Three training seeds give AUROC 0.770 ± 0.003 (seed-to-seed RFU $\rho = 0.995$). The handcrafted ratio does not account for the readout and the explicit volatility input is not necessary for it under matched training; these controls do not identify the source of the signal within the forecaster.

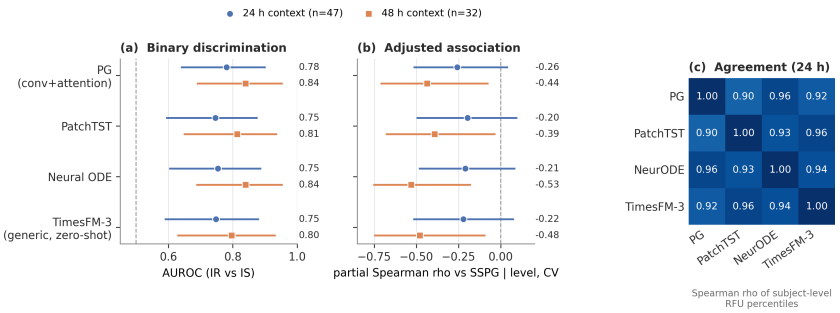


Figure 4: **Four frozen forecasters read insulin resistance from their relative forecast uncertainty, including a general-purpose model applied zero-shot** (main protocol; 95% subject-bootstrap intervals). **(a)** Binary discrimination (AUROC) at 24 hours (n=47) and 48 hours (n=32); dashed line, chance. **(b)** Level- and variability-adjusted association with SSPG (partial Spearman ρ); intervals exclude zero at 48 hours and include it at 24 hours for every forecaster. **(c)** Subject-level agreement between forecasters (Spearman ρ of within-cohort RFU percentiles, 24 hours). Protocol variants: Appendix F.

4.3 RFU READS INSULIN RESISTANCE CONSISTENTLY ACROSS FOUR FORECASTERS, INCLUDING A GENERAL-PURPOSE MODEL APPLIED ZERO-SHOT

We recompute RFU for the same 47 subjects on the same windows with three additional frozen forecasters: a patch Transformer and a neural ordinary differential equation model, both trained on the same glucose corpus, and TimesFM-3, a 330M-parameter general-purpose time-series foundation model (Jain & Sen, 2026), not designed for glucose and applied zero-shot. At 24 hours all four discriminate insulin-resistant from insulin-sensitive subjects (AUROC 0.780, 0.746, 0.754, and 0.748 in that order; permutation $p \leq 0.004$ each; Figure 4), and their subject-level rankings are highly correlated (Spearman $\rho = 0.90$, 0.96, and 0.92 with the primary forecaster). After adjustment for glucose level and variability all four yield negative associations with SSPG of similar magnitude (partial $\rho = -0.258$, -0.197, -0.211, and -0.221); the direction is consistent but inconclusive at this sample size.

TimesFM-3 differs from the primary forecaster in corpus, architecture (a decoder-only Transformer over 32-step patches), and uncertainty output (nine quantiles rather than a parametric scale), yet its subject rankings closely track the primary forecaster's ($\rho = 0.92$). At 48 hours (n=32) both again discriminate insulin resistance (AUROC 0.840 and 0.797) with adjusted associations of similar magnitude (partial $\rho = -0.437$ and -0.480, $p = 0.018$ and 0.009). The association is thus reproducible across the four tested forecasters, including a general-purpose model applied zero-shot.

These estimates depend on how per-window uncertainties are aggregated. Under the alternative protocols of Appendix F (median of per-window log-ratios; per-subject window caps of 30 or 40), every forecaster retains above-chance discrimination in every variant (AUROC 0.69–0.80 across the 16 forecaster-by-variant cells, permutation $p \leq 0.024$ in each), but effect sizes shift, and the adjusted association weakens for all four under the median-of-logs aggregation, most for TimesFM-3 (partial $\rho = -0.10$ to -0.16). The main protocol (all windows, mean ratio) was adopted after this audit (Appendix F). Cross-forecaster consistency is therefore robust in direction and in binary discrimination, and protocol-dependent in magnitude.

4.4 RFU ALIGNS WITH THE REPRESENTATION'S INSULIN-RESISTANCE READOUT

The emergent readout we set out to explain lives in a 320-dimensional representation. In the Metabolic Subphenotypes cohort (n=26), a ridge probe on that representation and RFU discriminate insulin resistance similarly (AUROC 0.79 and 0.75) and are strongly coupled at the subject level (Spearman $\rho = -0.84$; RFU is lower in the subjects the probe scores as more resistant); partialling either out of the other reduces both to near chance (0.60 and 0.58). In AI-READI (n=2181), where gold-standard labels are unavailable, RFU tracks a five-fold out-of-fold probe score at $\rho = -0.83$ (-0.55 after adjusting for glucose level and variability). The probe, however, reads insulin resis-

tance about as well from an untrained encoder as from the trained one (pooled AUROC 0.715 versus 0.724; Appendix G), whereas RFU from an untrained forecaster is at chance (Section 4.2). RFU is therefore a label-free scalar that aligns closely with the representation's readout, a surrogate for that readout rather than a mediator.

4.5 Level-matched analyses and simulations probe possible origins of the RFU signal

We ask where a level-adjusted signal could come from, using level-matched analyses in the real cohorts and a minimal-model simulator in which insulin sensitivity is set explicitly.

Real cohorts. Statistical adjustment assumes a functional form, so we add two assumption-light tests. Among subjects whose mean glucose falls in the central cohort-internal tercile (n=15), RFU ranks SSPG at $\rho = -0.39$ (two-sided permutation p=0.15); in caliper-matched resistant/sensitive pairs within cohort ($|\Delta\bar{g}|$<8 mg/dL; 13 pairs) the resistant member has lower RFU in 9 of 13 (sign test p=0.27). Both are directionally consistent with the adjusted primary result and inconclusive at these subset sizes.

Simulator. We evaluate a Bergman-type minimal-model simulator in four exploratory settings that differ in secretion model and in how mean glucose relates to insulin sensitivity S_I (48-hour context; Appendix D). With single-phase secretion and only S_I varying, RFU tracks S_I at Spearman $+0.96$ ($+0.20$ after adjusting for mean glucose). With biphasic secretion rescaled per subject so that mean glucose is near 100 mg/dL, it gives -0.11 (-0.12 adjusted). With S_I and the two biphasic gains sampled independently, it gives $+0.66$ ($+0.32$ after mean glucose, $+0.21$ after mean glucose and variability); this setting leaves mean glucose correlated with S_I ($\rho = -0.88$). With single-phase secretion rescaled inversely to S_I, the explicit compensation setting, it gives $+0.10$ (unchanged by adjustment). These settings differ in several respects at once and do not isolate a compensatory-secretion mechanism.

Resistance versus β-cell function. Because secretion enters the simulated mechanism, we ask whether RFU reads insulin resistance rather than β-cell function. In Metabolic Subphenotypes subjects with a disposition index (DI; n=24), the SSPG association given DI is -0.26 (permutation p=0.26) and the DI association given SSPG $+0.08$, which does not establish specificity. In the venous OGTT subset (n=20), early-phase glucose-shape proxies of secretion track SSPG ($\rho = +0.32$ to $+0.49$) and relate weakly to moderately to RFU after level adjustment (-0.08 to -0.26). These checks neither establish nor exclude secretion shape as the carrier; a direct test would require insulin or C-peptide time courses the OGTT does not provide (Matsuda & DeFronzo, 1999).

4.6 The phenomenon generalizes beyond glucose: a controlled latent-dynamics benchmark

Is reading a latent parameter from predictive uncertainty specific to glucose? We test a controlled system with no glucose: subjects simulated from a first-order feedback-controlled process, $x_t = x_{t-1} - k_i\,(x_{t-1} - B_0) + a_{t-1} + \varepsilon_t$, with k_i a subject-specific latent controller gain, a_t an exogenous impulse train, and ε_t noise (Appendix D). A probabilistic forecaster is trained on one population and RFU evaluated on a *subject-disjoint* population whose gains are drawn from the same range (ten simulation and training seeds; Figure 5).

Matched marginals. We z-score each subject's full series to match marginal means and variances across subjects (local-window statistics may still differ). Here a trained forecaster's RFU ranks the held-out gain at Spearman $+0.90$ ($[0.85, 0.93]$ across seeds) with a Gaussian head and $+0.88$ $[0.81, 0.93]$ with a Student-t head, whereas an untrained forecaster of the same architecture shows no clear positive association (-0.19 $[-0.63, +0.13]$).

Natural regime. When level and variance vary with k, the raw association is large but carried mostly by marginal statistics: after adjusting for series level and variability the partial Spearman is $+0.21$ ($[0.02, 0.38]$; $+0.18$ Student-t). Across the grid (Appendix D) the adjusted association is small in every cell and none is established at three seeds; the largest is $+0.28$ (T=1800).

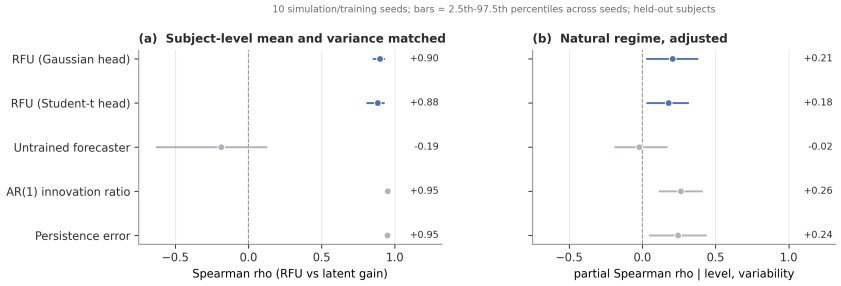


Figure 5: **Controlled latent-dynamics benchmark** (subject-disjoint test subjects; ten simulation and training seeds; brackets and error bars are the 2.5th–97.5th percentiles across seeds). **(a)** Subject-level mean and variance matched: a trained forecaster's RFU ranks the held-out latent gain, the untrained result is inconclusive, and an AR(1) innovation ratio and the persistence error read it at $+0.95$. **(b)** Natural regime, adjusted for series level and variability.

A linear baseline reads the gain comparably. Because the process is linear, we also compute a first-order autoregressive normalized innovation ratio (AR(1) residual standard deviation over series standard deviation); it ranks the gain as well as or better than RFU ($+0.26$ natural, $+0.95$ matched), so we do not present RFU as a superior estimator of latent dynamics; the narrower observation is that a general forecaster trained only to predict, with no model of the system, ranks the latent gain through its uncertainty at a level approaching a linear statistic fitted to each series. An idealized lemma (Appendix B; optimal forecaster, true conditional and marginal standard deviations, no impulse train) predicts a positive relation between the mean-ratio statistic and the gain on $k \in (0, 1)$; the simulation results, with a learned forecaster and finite-window standard deviations, are consistent with it.

5 Discussion

Interpretation. Across the four frozen forecasters we tested, including a general-purpose model applied zero-shot, relative predictive uncertainty discriminates SSPG-defined insulin resistance, and binary discrimination remains above chance after adjustment for glucose level and variability (nominally significant); in these tests the readout is not specific to the primary forecaster or to glucose-specific pretraining. The direction has a possible control-systems interpretation: in the idealized feedback model of Appendix B, higher regulatory gain yields higher relative uncertainty, as reproduced by the controlled benchmark. This interpretation remains untested in humans. We hypothesize that the association may weaken when external control (automated insulin delivery) dominates the observed dynamics.

Limitations. The gold-standard cohorts are small; the covariate-adjusted binary test is nominally significant only, and several secondary results are directional but inconclusive (Table 1, Appendix C). Effect sizes depend on the aggregation protocol, most for the general-purpose forecaster, and the adjusted binary result on fold assignment (fixed and averaged versions are reported). Medication status is unavailable in the public releases, so glucose-lowering therapy is an unadjusted potential confound. The time-shuffle control cannot separate diurnal from other temporal structure. The findings do not resolve peripheral from hepatic contributions; the control-systems interpretation is untested, and the simulator explores possible generative settings without identifying the human mechanism.

Reproducibility. All human analyses use public or data-use-agreement data under their original terms. The submission includes the RFU analysis code, protocol module, and outputs for all four forecasters, supporting reanalysis of the reported RFU results. Primary and retrained-control checkpoints are withheld pending clarification of redistribution permissions (Appendix H). TimesFM-3 weights carry a non-commercial license.

USE OF LARGE LANGUAGE MODELS

Large language models were used as assistants for writing and revising analysis code, for drafting and editing manuscript text, and as an adversarial reviewer of intermediate drafts. All analyses were specified by the authors; every reported number was recomputed by the authors from the released code and checked against the analysis outputs; and all scientific claims and their scope were decided by the authors, who take full responsibility for the content.

REFERENCES

AI-READI Consortium. AI-READI: rethinking AI data collection, preparation and sharing in diabetes research and beyond. *Nature Metabolism*, 6:2210–2212, 2024. doi: 10.1038/s42255-024-01165-x.

AI-READI Consortium. Flagship dataset of type 2 diabetes from the AI-READI project (v3.0.0). FAIRhub dataset, 2025. URL https://doi.org/10.60775/fairhub.3. Version 3.0.0; dataset reports 2280 participants.

Anonymous. Companion manuscript S1 (title withheld for double-blind review): a probabilistic forecaster for continuous glucose monitoring pretrained on type 1 diabetes data, 2026a. Under review.

Anonymous. Companion manuscript S2 (title withheld for double-blind review): insulin resistance read from a frozen glucose forecaster's representation, 2026b. Under review.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, et al. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=gerNCVqqtR.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Assran_Self-Supervised_Learning_From_Images_With_a_Joint-Embedding_Predictive_Architecture_CVPR_2023_paper.html.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, pp. 6571–6583, 2018. URL https://papers.neurips.cc/paper/7892-neural-ordinary-differential-equations.

Beyza Cinar and Maria Maleshkova. DiaData: An integrated large dataset for type 1 diabetes and hypoglycemia research. In *BIO Web of Conferences*, volume 195, pp. 03001, 2025. doi: 10.1051/bioconf/202519503001.

Laurent Crenier, Maria Lytrivi, Annelien Van Dalem, Bart Keymeulen, and Bernard Corvilain. Glucose complexity estimates insulin resistance in either nondiabetic individuals or in type 1 diabetes. *The Journal of Clinical Endocrinology & Metabolism*, 101(4):1490–1497, 2016. doi: 10.1210/jc.2015-4035.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10148–10167. PMLR, 2024. URL https://proceedings.mlr.press/v235/das24c.html.

R. A. DeFronzo, J. D. Tobin, and R. Andres. Glucose clamp technique: a method for quantifying insulin secretion and resistance. *American Journal of Physiology*, 237(3):E214–E223, 1979. doi: 10.1152/ajpendo.1979.237.3.E214.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993. ISBN 0-412-04231-2.

David Freedman and David Lane. A nonstochastic interpretation of reported significance levels. *Journal of Business & Economic Statistics*, 1(4):292–298, 1983. doi: 10.1080/07350015.1983.10509354.

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.

N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F (Radar and Signal Processing)*, 140(2):107–113, 1993. doi: 10.1049/ip-f-2.1993.0015.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 16115–16152. PMLR, 2024. URL https://proceedings.mlr.press/v235/goswami24a.html.

M. S. Greenfield, L. Doberne, F. Kraemer, T. Tobey, and G. Reaven. Assessment of insulin resistance with the insulin suppression test and the euglycemic clamp. *Diabetes*, 30(5):387–392, 1981. doi: 10.2337/diab.30.5.387.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Heather Hall, Dalia Perelman, Alessandra Breschi, Patricia Limcaoco, Ryan Kellogg, Tracey McLaughlin, and Michael Snyder. Glucotypes reveal new patterns of glucose dysregulation. *PLoS Biology*, 16(7):e2005143, 2018. doi: 10.1371/journal.pbio.2005143.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/He_Masked_Autoencoders_Are_Scalable_Vision_Learners_CVPR_2022_paper.html.

Ayush Jain and Rajat Sen. TimesFM-3: A zero-shot foundation model for multivariate forecasting. Google Research Blog, August 2026. URL https://research.google/blog/timesfm-3-a-zero-shot-foundation-model-for-multivariate-forecasting/. Model weights: https://huggingface.co/google/timesfm-3.0-pytorch (330M parameters; TimesFM Non-Commercial License v1.0). Accessed 2026-09-10.

S. E. Kahn, R. L. Prigeon, D. K. McCulloch, E. J. Boyko, R. N. Bergman, M. W. Schwartz, J. L. Neifing, W. K. Ward, J. C. Beard, J. P. Palmer, and D. Porte. Quantification of the relationship between insulin sensitivity and β-cell function in human subjects: evidence for a hyperbolic function. *Diabetes*, 42(11):1663–1672, 1993. doi: 10.2337/diab.42.11.1663.

R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(1):35–45, 1960. doi: 10.1115/1.3662552.

Yann LeCun. A path towards autonomous machine intelligence (version 0.9.2). OpenReview, 2022. URL https://openreview.net/pdf?id=BZ5a1r-kVsf.

Zechen Li, Keerthana Natarajan, Weizhi Zhang, Menglian Zhou, Simon A. Lee, Yuwei Zhang, Maxwell A. Xu, Zeinab Esmaeilpour, Flora D. Salim, Mark Malhotra, Lindsey Sunden, Shwetak Patel, Yuzhe Yang, and Ahmed A. Metwally. GlucoFM: A dual-stream foundation model for continuous glucose monitoring. arXiv preprint, 2026. URL https://arxiv.org/abs/2605.30865. arXiv:2605.30865.

Lennart Ljung. *System Identification: Theory for the User*. Prentice Hall PTR, Upper Saddle River, NJ, 2nd edition, 1999. ISBN 0-13-656695-2.

Yurun Lu, Dan Liu, Zhongming Liang, Rui Liu, Pei Chen, Yitong Liu, Jiachen Li, Zhanying Feng, Lei M. Li, Bin Sheng, Weiping Jia, Luonan Chen, Huating Li, and Yong Wang. A pretrained transformer model for decoding individual glucose dynamics from continuous glucose monitoring data. *National Science Review*, 12(5):nwaf039, 2025. doi: 10.1093/nsr/nwaf039.

Junjie Luo, Abhimanyu Kumbara, Mansur Shomali, Rui Han, Anand Iyer, Grazia Aleppo, Ritu Agarwal, and Gordon Gao. A large sensor foundation model pretrained on continuous glucose monitor data for diabetes management. *npj Health Systems*, 2:35, 2025. doi: 10.1038/s44401-025-00039-y.

Guy Lutsker, Gal Sapir, Smadar Shilo, Jordi Merino, Anastasia Godneva, Jerry R. Greenfield, Dorit Samocha-Bonet, Raja Dhir, Francisco Gude, Shie Mannor, Eli Meirom, Eric P. Xing, Gal Chechik, Hagai Rossman, and Eran Segal. A foundation model for continuous glucose monitoring data. *Nature*, 650:978–986, 2026. doi: 10.1038/s41586-025-09925-9.

Marcos Matabuena, Alexander Petersen, Juan C. Vidal, and Francisco Gude. Glucodensities: a new representation of glucose profiles using distributional data analysis. *Statistical Methods in Medical Research*, 30(6):1445–1464, 2021. doi: 10.1177/0962280221998064.

Marcos Matabuena, Rahul Ghosal, Javier Enrique Aguilar, Ayya Keshet, Robert Wagner, Carmen Fernández Merino, Juan Sánchez Castro, Vadim Zipunnikov, Jukka-Pekka Onnela, and Francisco Gude. Glucodensity functional profiles outperform traditional continuous glucose monitoring metrics. *Scientific Reports*, 15:33662, 2025. doi: 10.1038/s41598-025-18119-2.

Masafumi Matsuda and Ralph A. DeFronzo. Insulin sensitivity indices obtained from oral glucose tolerance testing: comparison with the euglycemic insulin clamp. *Diabetes Care*, 22(9):1462–1470, 1999. doi: 10.2337/diacare.22.9.1462.

D. R. Matthews, J. P. Hosker, A. S. Rudenski, B. A. Naylor, D. F. Treacher, and R. C. Turner. Homeostasis model assessment: insulin resistance and β-cell function from fasting plasma glucose and insulin concentrations in man. *Diabetologia*, 28(7):412–419, 1985. doi: 10.1007/BF00280883.

Ahmed A. Metwally, Dalia Perelman, Heyjun Park, Yue Wu, Alokkumar Jha, Seth Sharp, Alessandra Celli, Ekrem Ayhan, Fahim Abbasi, Anna L. Gloyn, Tracey McLaughlin, and Michael P. Snyder. Prediction of metabolic subphenotypes of type 2 diabetes via continuous glucose monitoring and machine learning. *Nature Biomedical Engineering*, 9(8):1222–1239, 2025. doi: 10.1038/s41551-024-01311-6.

Hada Melino Muhammad, Zechen Li, Flora Salim, and Ahmed A. Metwally. CGM-JEPA: Learning consistent continuous glucose monitor representations via predictive self-supervised pretraining. arXiv preprint, 2026. URL `https://arxiv.org/abs/2605.00933`. arXiv:2605.00933.

Cynthia Owsley et al. Cross-sectional design and protocol for Artificial Intelligence Ready and Equitable Atlas for Diabetes Insights (AI-READI). *BMJ Open*, 15(2):e097449, 2025. doi: 10.1136/bmjopen-2024-097449.

Heyjun Park, Ahmed A. Metwally, Alireza Delfarah, Yue Wu, Dalia Perelman, Caleb Mayer, Curtis McGinity, Majid Rodgar, Alessandra Celli, Tracey McLaughlin, Emmanuel Mignot, and Michael Snyder. High-resolution lifestyle profiling and metabolic subphenotypes of type 2 diabetes. *npj Digital Medicine*, 8:352, 2025. doi: 10.1038/s41746-025-01728-6.

D. Pei, C. N. Jones, R. Bhargava, Y.-D. I. Chen, and G. M. Reaven. Evaluation of octreotide to assess insulin-mediated glucose disposal by the insulin suppression test. *Diabetologia*, 37(8):843–845, 1994. doi: 10.1007/BF00404344.

C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger. Mosaic organization of DNA nucleotides. *Physical Review E*, 49(2):1685–1689, 1994. doi: 10.1103/PhysRevE.49.1685.

Steven M. Pincus. Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences*, 88(6):2297–2301, 1991. doi: 10.1073/pnas.88.6.2297.

Gerald M. Reaven. Banting Lecture 1988: Role of insulin resistance in human disease. *Diabetes*, 37(12):1595–1607, 1988. doi: 10.2337/diab.37.12.1595.

Joshua S. Richman and J. Randall Moorman. Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology–Heart and Circulatory Physiology*, 278(6):H2039–H2049, 2000. doi: 10.1152/ajpheart.2000.278.6.H2039.

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://proceedings.neurips.cc/paper/2019/hash/42a6845a557bef704ad8ac9cb4461d43-Abstract.html.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3): 1181–1191, 2020. doi: 10.1016/j.ijforecast.2019.07.001.

Renat Sergazinov, Elizabeth Chun, Valeriya Rogovchenko, Nathaniel Fernandes, Nicholas Kasman, and Irina Gaynanova. GlucoBench: Curated list of continuous glucose monitoring datasets with prediction benchmarks. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/4c12e97f2e05304a451e18c9c945036f-Abstract-Conference.html.

Adam G. Tabák, Markus Jokela, Tasnime N. Akbaraly, Eric J. Brunner, Mika Kivimäki, and Daniel R. Witte. Trajectories of glycaemia, insulin sensitivity, and insulin secretion before diagnosis of type 2 diabetes: an analysis from the Whitehall II study. *The Lancet*, 373(9682): 2215–2221, 2009. doi: 10.1016/S0140-6736(09)60619-X.

Tara M. Wallace, Jonathan C. Levy, and David R. Matthews. Use and abuse of HOMA modeling. *Diabetes Care*, 27(6):1487–1495, 2004. doi: 10.2337/diacare.27.6.1487.

A Normalization algebra

The primary and comparison forecasters standardize each context window before predicting: the window is centred by its mean and scaled by its standard deviation, the model predicts in that normalized space, and the output scale is rescaled to glucose units by the same window standard deviation, $\hat{\sigma}_h = s_{i,w}\,\hat{\sigma}_h^{\mathrm{norm}}$. Substituting into Eq. (1), the per-window ratio is $r_{i,w} = \frac{1}{H}\sum_h \hat{\sigma}_h^{\mathrm{norm}} \cdot s_{i,w}/\mathrm{SD}(g_{i,w})$, i.e. the mean of the model's own normalized-space scale head times the ratio of the two standard-deviation conventions. The model's internal $s_{i,w}$ uses the unbiased $(L-1)$ estimator and the RFU denominator the biased (L) estimator, so the factor is $\sqrt{L/(L-1)}$, equal to 1.0017 at $L{=}288$ and identical for every window; it does not affect any ranking. RFU is therefore a reading of the forecaster's native normalized predictive scale, not an externally imposed normalization.

B Analytical account: what the two RFU aggregations estimate

Setting. Let $\{x_t\}$ be a covariance-stationary process with past $\mathcal{F}_t$, marginal variance $\mathrm{Var}(x)$, and h-step conditional variance $v_h = \mathbb{E}[\mathrm{Var}(x_{t+h} \mid \mathcal{F}_t)]$. Write $R_h^2 = 1 - v_h/\mathrm{Var}(x)$ for the fraction of marginal variance an optimal h-step predictor explains, so $0 \le R_h^2 \le 1$. Throughout, "optimal forecaster" means one whose predictive scale $\hat{\sigma}_h$ equals $\sqrt{v_h}$ and whose denominator equals $\sqrt{\mathrm{Var}(x)}$.

Identity for the log-RMS form (Eq. 8). For the per-window statistic $\log\big(\bar{\sigma}/\mathrm{SD}\big)$ with $\bar{\sigma}^2 = \frac{1}{H}\sum_h \hat{\sigma}_h^2$, the optimal forecaster gives

$$\exp\big(2\log(\bar{\sigma}/\mathrm{SD})\big) = \frac{1}{H}\sum_{h=1}^{H}\frac{v_h}{\mathrm{Var}(x)} = 1 - \bar{R}^2_{\mathrm{pred}}, \qquad \bar{R}^2_{\mathrm{pred}} = \frac{1}{H}\sum_h R_h^2. \tag{2}$$

This identity is exact for the root-mean-square aggregation and does not hold for the mean-ratio form used as the main statistic; for that form we prove the weaker, but sufficient, property below.

Lemma (mean-ratio statistic on a first-order feedback process). Consider $x_t - \mu = \phi\,(x_{t-1} - \mu) + \varepsilon_t$ with ε_t i.i.d., $\mathrm{Var}(\varepsilon) = \sigma_\varepsilon^2$, and write $\phi = 1 - k$ for the feedback gain k. For $k \in (0, 1)$ we have $\phi \in (0, 1)$, the process is stationary with $\mathrm{Var}(x) = \sigma_\varepsilon^2/(1 - \phi^2)$, and the optimal h-step predictive scale satisfies $\hat{\sigma}_h/\sqrt{\mathrm{Var}(x)} = \sqrt{1 - \phi^{2h}}$. Hence the per-window mean ratio of the main statistic equals

$$r(k) = \frac{1}{H} \sum_{h=1}^{H} \sqrt{1 - (1 - k)^{2h}}. \tag{3}$$

Each summand is strictly increasing in k on $(0, 1)$ because $(1 - k)^{2h}$ is strictly decreasing there; therefore $r(k)$, and $\log r(k)$, are strictly increasing in k on $(0, 1)$ for every horizon count H. The statement is restricted to $k \in (0, 1)$: for $k \in (1, 2)$ the process is still stationary but $\phi < 0$ and $(1-k)^{2h}$ increases with k, so monotonicity reverses; $k \geq 2$ is non-stationary. Within this model the lemma gives (i) the ordering of RFU by gain: a high-gain loop has large innovation relative to its own spread and thus higher RFU, a low-gain loop lower RFU. The human direction (lower RFU with insulin resistance) is consistent with this ordering under the reading that insulin resistance lowers the regulatory gain, a reading the lemma does not establish. It also gives (ii) scale invariance: $r(k)$ is dimensionless, so per-subject standardisation of the series leaves it unchanged, which is why the matched-marginal synthetic readout survives. Both aggregations are monotone in k on this model; they differ only in how they weight horizons.

From the ideal model to the empirical RFU. The lemma concerns an optimal forecaster on a linear, stationary, first-order process. The empirical statistic differs in four respects, each of which we state rather than absorb into the lemma. (1) The empirical numerator is the model's own uncertainty summary, not $\sqrt{v_h}$. Even for a perfectly calibrated Student-t head the predictive standard deviation is $\hat{\sigma}_h\sqrt{\hat{\nu}_h/(\hat{\nu}_h - 2)}$ rather than $\hat{\sigma}_h$ (Appendix F reports RFU under that and other dispersion summaries), and a learned forecaster adds estimation error on top; the calibration analysis (Section 4.2) provides empirical diagnostics rather than an analytical bound. (2) The denominator is the empirical standard deviation of a finite context window, not the marginal standard deviation; for persistent series a short window under-estimates the marginal spread, and the shortfall shrinks with context length. This is a bias of the estimator, not a property of the process, and whether it matters for the human readout is not tested here: the 24- and 48-hour analyses use different subject sets, so their difference is not evidence for or against it. (3) The subject-level statistic is the logarithm of the mean of window ratios; the alternative in Eq. 8 is the median of window log-ratios. These differ in both horizon and window aggregation and need not coincide; their empirical agreement and disagreement are reported in Appendix F. (4) Glucose is neither linear nor stationary over a day (meals, biphasic secretion, circadian modulation), so $\bar{R}^2_{\mathrm{pred}}$ of real glucose need not be captured by any single persistence exponent or linear innovation ratio; the lemma does not say how much of it they capture, and the partial overlap of RFU with DFA and the AR(3) baseline (Section 4.2) is an empirical observation, not a consequence of the lemma. The account is therefore first-order and correlational: it identifies what the statistic estimates and when it tracks a latent gain in the idealized model, and it makes no claim that glucose follows that model.

C Pooling, per-cohort estimates, cross-fitting, and multiple comparisons

Pooling and sample accounting. All continuous analyses use one pipeline (cohort-internal percentile de-batching, partial Spearman controlling mean glucose and variability) on the clean set (valid SSPG, at least three qualifying windows): n=47 at 24 h (26 Metabolic Subphenotypes, 21 Hall; 16 resistant) and n=32 at 48 h (22 and 10; 11 resistant). Binary discrimination is computed on raw RFU pooled across cohorts; cohort-internal standardization changes the 24-h AUROC from 0.780 to 0.772.

Labels and a common-threshold sensitivity. The binary endpoint follows each cohort's convention (Section 3). Applying one threshold to both cohorts gives AUROC 0.808 at SSPG $\geq$ 120 mg/dL (17 of 47 resistant) and 0.719 at $\geq$ 150 mg/dL (13 of 47), versus 0.780 with cohort-defined labels (16 of 47).

Claim	Evidence (24 h primary; 48 h sensitivity)	Status
RFU discriminates SSPG-defined insulin resistance	AUROC 0.78 $[0.64, 0.90]$, p=0.0009, n=47; 48 h 0.84 $[0.69, 0.95]$, n=32	load-bearing
after linear adjustment (level, CV, cohort)	cross-fitted AUROC 0.70 $[0.53, 0.85]$, p=0.029 (Holm 0.17; mean over 50 splits 0.71); 48 h 0.84, p=0.0008	load-bearing (nominal)
continuous, adjusted	partial $\rho{=}-0.26$ $[-0.52, +0.04]$, p=0.09; 48 h -0.44, p=0.018	corroborating
Sensitive to time shuffling	within-window shuffle: AUROC $0.78 \rightarrow 0.46{\pm}0.02$	load-bearing
Better realized skill on resistant glucose	MASE 0.86 vs 1.08, p=0.003; RFU tracks realized skill $+0.59$	load-bearing
Overlaps with long-range persistence	$\rho(\mathrm{RFU}, \mathrm{DFA}){=}-0.83$; mutual adjustment attenuates both	descriptive
Learned scale adds over normalization	ladder: untrained 0.50, raw 0.58, constant-scale 0.71, full 0.78; Δ full$-$constant $+0.07$ $[-0.05, +0.20]$	inconclusive
Not the explicit volatility input	no-volatility retraining 0.764 vs 0.772; subject rankings ρ=0.996	load-bearing
Reproduced across four forecasters	AUROC 0.75–0.78 (all $p{\leq}0.004$); rankings agree 0.90–0.96; magnitude protocol-dependent	load-bearing
Aligns with representation readout	MSP $\rho{=}-0.84$ (n=26); AI-READI $\rho{=}-0.83$, adjusted -0.55 (n=2181)	load-bearing
Level-matched human analyses	central band $\rho{=}-0.39$, p=0.15; pairs $9/13$, p=0.27	corroborating
Specificity vs β-cell function	SSPG|DI -0.26, p=0.26; DI|SSPG $+0.08$ (n=24)	not established
Controlled system reads latent gain	matched marginals $+0.90$ $[0.85, 0.93]$; natural, adjusted $+0.21$ $[0.02, 0.38]$	load-bearing

Table 1: **Claims, evidence, and status.** Pooled gold-standard cohorts (Metabolic Subphenotypes + Hall) unless stated; AUROC p-values are two-sided orientation-free permutation tests, partial-correlation p-values Freedman–Lane permutation tests, intervals 95% subject bootstraps (Section 3), except the controlled-system row, whose brackets are 2.5th–97.5th percentiles across ten seeds. Status describes the role of each result in the argument rather than a fixed significance threshold: load-bearing results are those the paper's claims rest on (the primary endpoint, its adjusted form, and the controls and replications that qualify it); corroborating results are directionally consistent with the primary result but inconclusive at this sample size; descriptive results carry no inferential claim; inconclusive marks an increment whose interval includes zero; not established marks a test that did not support the claim. Under Holm correction of the inferential family (Appendix C), binary discrimination and the MASE contrast survive.

Per-cohort estimates (no pooling). At 24 h binary discrimination is present in both cohorts (Metabolic Subphenotypes AUROC 0.750, n=26, 12 resistant; Hall 0.824, n=21, 4 resistant). The adjusted continuous association is carried mainly by Metabolic Subphenotypes (partial $\rho = -0.324$) and is weak in Hall (-0.106), which has four resistant subjects and a compressed SSPG range; we therefore do not claim independent per-cohort replication of the adjusted gradient.

Cross-fitted adjustment. The adjusted binary AUROC uses five folds stratified by cohort only, so fold assignment does not depend on the insulin-resistance labels. Within each split, RFU, mean glucose, and glucose variability (coefficient of variation) are standardized per cohort with training-fold means and standard deviations, RFU is regressed on the standardized covariates and a cohort indicator on the training folds, and the held-out fold is scored by the negative residual; the out-of-fold scores are pooled and ranked. The permutation test permutes the labels of the residual score within cohort (20,000 permutations, two-sided in orientation), so it asks whether the residualized RFU still discriminates status, not whether RFU adds to a covariate model. The bootstrap interval re-runs the entire procedure (fold-internal standardization, regression, and scoring) on each of 2,000 subject resamples; within a resample all copies of a resampled subject are assigned to the same fold, so the held-out fold never contains a copy of a subject used to fit (`code/run_19d_adjusted_bootstrap_grouped.py`). For the AI-READI representation probe, folds are formed at the subject level, so all windows of a subject fall in the same fold. With

the fixed fold assignment (seed 0) the 24-h AUROC is 0.696 (95% bootstrap CI $[0.53, 0.85]$, two-sided permutation p=0.029); over 50 splits the mean is 0.712 (SD 0.016, range 0.685–0.748). At 48 h the corresponding values are 0.840 ($[0.59, 0.95]$, p=0.0008) and 0.814 (SD 0.015). A row-level bootstrap that lets copies of one subject fall in different folds gives $[0.55, 0.84]$ and $[0.61, 0.94]$; the subject-grouped intervals reported here are the ones stated throughout.

Freedman–Lane permutation for partial correlations. Continuous analyses use cohort-internal percentile ranks of RFU and SSPG. The partial correlation residualizes both on mean glucose and variability with a single pooled regression and reports the Pearson correlation of the residualized percentile scores. For the permutation test, the SSPG percentile is regressed on the same covariates with that pooled regression; its residuals are permuted within cohort and added back to the fitted values to form a permuted outcome, the partial correlation is recomputed, and the two-sided p-value is the fraction of $20{,}000$ permutations whose absolute partial correlation is at least the observed one (Freedman & Lane, 1983).

Analysis hierarchy and multiple comparisons. No written analysis plan preceded these analyses, so no test is presented as confirmatory in the pre-registered sense. The primary association, SSPG-defined status versus RFU, is analysed as a model sequence, crude (p=0.0009) and covariate-adjusted (p=0.029), in the way an epidemiological analysis reports crude and adjusted estimates of one association. The secondary tests (MASE contrast 0.003, variability-matched pairs 0.077, adjusted continuous gradient 0.092, level-matched subset 0.146, level-matched pairs 0.267, full-versus-constant-scale ablation 0.233) are corrected together with the primary test by the Holm procedure at α=0.05: the primary test and the MASE contrast survive. If the adjusted binary test is added as an eighth member of that family, its Holm-adjusted p is 0.17; we therefore describe it as nominally significant, and the statement that discrimination remains above chance after adjustment rests on the point estimate and its nominal interval (0.70 $[0.53, 0.85]$ at 24 hours; 0.84 $[0.59, 0.95]$ at 48 hours), not on a family-wise significant test. Neither that interval nor the stability of the estimate across 50 fold assignments substitutes for multiplicity correction: the former is nominal-level inference and the latter describes sensitivity to fold assignment.

D SIMULATOR AND CONTROLLED-SYSTEM SPECIFICATIONS

Synthetic benchmark: full specification. *Generative model.* Each subject i has a latent controller gain $k_i \sim \mathrm{LogUniform}(0.02, 0.30)$ and series $x_t = x_{t-1} - k_i(x_{t-1} - B_0) + a_{t-1} + \varepsilon_t$, with baseline B_0=100, noise $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ (σ=1.5), and an exogenous impulse train a_t (Poisson arrivals, rate $1/120$; gamma-shaped absorption) modelling meals; T=1200. *Protocol.* We draw 120 training and 80 *disjoint* test subjects with independent gains; the forecaster (2-layer GRU, per-window RevIN normalization, Gaussian or Student-t head, 20 epochs) is trained on the training subjects only, and RFU (main protocol: mean per-window ratio, logged) is computed on the never-seen test subjects. We report the mean over ten simulation/training seeds and, in brackets, the 2.5th–97.5th percentiles of the ten per-seed estimates, a seed spread rather than a subject bootstrap (`code/run_05_synthetic.py`; log `results/v3_05_synthetic_log.txt`). *Conditions.* In the *natural* regime, level and variance vary with k; the level/variability-adjusted partial Spearman is $+0.206$ ($[0.02, 0.38]$, Gaussian head; $+0.179$ $[0.03, 0.32]$, Student-t head), and an untrained forecaster gives -0.022 $[-0.19, +0.17]$, i.e. most of the (large) raw association runs through marginal statistics. In the *matched* regime, each subject's series is z-scored so all subjects share identical marginal mean and variance; RFU then ranks k at $+0.899$ $[0.85, 0.93]$ (Gaussian) and $+0.882$ $[0.81, 0.93]$ (Student-t), versus -0.186 $[-0.63, +0.13]$ untrained. *Baselines.* An AR(1) normalized innovation ratio reads k at $+0.262$ $[0.11, 0.41]$ (natural) and $+0.950$ $[0.93, 0.97]$ (matched), and the persistence error at $+0.243$ and $+0.949$; on this linear system a bespoke system-identification statistic therefore reads k as well as or better than RFU, and the benchmark shows that a general forecaster's uncertainty recovers k without a model of the system, not that it is a superior estimator. *Robustness grid* (natural regime, Gaussian head, three seeds, level/variability-adjusted partial Spearman; the three-seed default cell is $+0.041$ $[-0.07, +0.15]$). Observation noise $\sigma \in \{0.75, 1.5, 3.0\}$: $+0.044$, $+0.041$, $+0.049$. Impulse rate $1/240$, $1/120$, $1/60$: $+0.152$ $[0.10, 0.21]$, $+0.041$, $+0.044$ $[0.005, 0.08]$. Series length $T \in \{800, 1200, 1800\}$: $+0.125$ $[0.03, 0.20]$, $+0.041$, $+0.276$ $[0.13, 0.36]$. Gain range narrow, wide, high: $+0.045$, $+0.041$, -0.000 $[-0.17, +0.11]$. Brackets are three-seed ranges, not confidence intervals, and are not used as a significance test; the adjusted

association is small and non-negative throughout, largest at sparser impulses and longer series. The grid is a separate experiment from the main ten-seed estimate: it draws different simulated populations (population seeds $2000{+}s$ and $8000{+}s$ versus $1000{+}s$ and $9000{+}s$), trains for 18 rather than 20 epochs, and uses three seeds whose brackets are the percentiles of three values. A paired re-run with the same ten seeds (`code/run_05b_seed_decomposition.py`; per-seed values in `results/v3_05b_seed_decomposition.csv`) separates the two differences. It reproduces the main estimate exactly ($+0.206$ $[0.02, 0.38]$ natural; $+0.899$ $[0.85, 0.93]$ matched). The grid's default cell ($+0.041$) is the mean of its three seeds ($+0.156$, -0.078, $+0.044$); over ten seeds the grid population gives $+0.147$ $[-0.05, +0.38]$ at 18 epochs and $+0.170$ $[-0.03, +0.42]$ at 20 epochs. Paired within seed, 20 versus 18 epochs changes the natural-regime estimate by $+0.013$ (main population) and $+0.023$ (grid population), and the main-versus-grid population difference is $+0.036$ with a seed range of $[-0.27, +0.32]$; per-seed values have a standard deviation of 0.13–0.16. Expanding the grid from three to ten seeds reduces the discrepancy; the observed paired differences associated with training duration and population draws are smaller. Model capacity and analysis choices were not tuned on the evaluation seeds.

Mechanism simulator: full specification. The simulator is a Bergman-type minimal-model closed loop integrated by explicit Euler at $\Delta t{=}1$ min. The state is plasma glucose G (mg/dL), remote insulin action X, and plasma insulin I (μU/mL):

$$G_t = G_{t-1} + \big[-(S_G + X_{t-1})\, G_{t-1} + S_G G_b + R_{a,t-1} \big] \Delta t, \tag{4}$$

$$X_t = \big[X_{t-1} + (-p_2 X_{t-1} + p_2\, S_I\, (I_{t-1} - I_b))\, \Delta t \big]_+, \tag{5}$$

$$I_t = \big[I_{t-1} + (S_t - n\, (I_{t-1} - I_b))\, \Delta t \big]_+, \tag{6}$$

where $[\cdot]_+ = \max(0, \cdot)$, S_G is glucose effectiveness, p_2 the insulin-action rate, n insulin clearance, S_I insulin sensitivity, and S_t insulin secretion. Meals enter through the rate of glucose appearance as gamma-shaped absorption, $R_{a,t} = \sum_m D_m\, k_a^2\, (t - \tau_m)\, e^{-k_a (t - \tau_m)}\, \mathbb{1}[t > \tau_m]$, with three meals per day (07:00/12:00/19:00; carbohydrate D_m and timing τ_m jittered) and stochastic snacks. The observed CGM is G down-sampled to 5 min with additive sensor noise $\mathcal{N}(0, 2.5^2)$ and clipped to $[40, 400]$. Secretion takes two forms, both evaluated from the previous glucose samples. *Single-phase*: $S_t = \beta\, [G_{t-1} - h]_+$. *Biphasic*: a first phase proportional to the rate of glucose rise plus a slow second-phase provision P,

$$S_t = k_{\mathrm{dyn}}\, [G_{t-1} - G_{t-2}]_+ + P_t, \qquad P_t = \big[P_{t-1} - \alpha \left(P_{t-1} - k_{\mathrm{stat}}\, [G_{t-1} - h]_+ \right) \Delta t \big]_+. \tag{7}$$

Settings. Four settings were simulated, each with its own script and virtual population; they differ in the secretion model and in how mean glucose is allowed to relate to S_I, and they are not a nested ladder. *(a) Single-phase free sweep* (`run_11a`, $n{=}48$): $S_I = S_I^{\mathrm{ref}} e^U$, $U \sim \mathrm{Unif}(\log 0.3, \log 2.2)$, secretion $\beta{=}0.06$ fixed; mean glucose drifts with S_I (88.8–102.3 mg/dL). *(b) Compensated single phase* (`run_11a`, $n{=}48$): the same draw with $\beta = 0.06\, S_I^{\mathrm{ref}} / S_I$, so less sensitive subjects secrete proportionally more; mean glucose 92.1–94.2 mg/dL and nearly uncorrelated with S_I ($\rho = -0.03$). *(c) Mean-clamped biphasic* (`run_11c`, $n{=}60$): biphasic secretion with S_I drawn over 0.35–1.9 $\times\, S_I^{\mathrm{ref}}$, $k_{\mathrm{dyn}} \sim \mathrm{Unif}(0.3, 2.5)$, $k_{\mathrm{stat}} \sim \mathrm{Unif}(0.03, 0.11)$, and a per-subject secretion multiplier found by bisection so that mean glucose approaches 100 mg/dL (achieved range 97.0–100.2). *(d) Independently randomized biphasic* (`run_11b`, $n{=}80$): S_I over 0.3–2.0 $\times\, S_I^{\mathrm{ref}}$ with $k_{\mathrm{dyn}} \sim \mathrm{Unif}(0.2, 2.5)$ and $k_{\mathrm{stat}} \sim \mathrm{Unif}(0.03, 0.11)$ sampled independently of S_I; no clamping, so mean glucose spans 78.8–106.0 mg/dL and remains correlated with S_I ($\rho = -0.88$). Settings (c) and (d) both use biphasic secretion, so their comparison does not add a secretion phase, and only (b) rescales secretion by S_I. Parameter values (min^{-1} where applicable): $S_G{=}0.015$, $p_2{=}0.025$, $n{=}0.15$, $G_b{=}90$, $I_b{=}8$, $h{=}85$, $\alpha{=}0.02$, $k_a{=}0.03$, $S_I^{\mathrm{ref}}{=}1.2{\times}10^{-3}$; 108 h simulated per subject, evaluated at a 48-hour (576-sample) context. On each virtual CGM we compute RFU exactly as on human data and correlate it with the *known* S_I. Raw and mean-glucose-adjusted Spearman correlations are $+0.96/+0.20$ (a), $+0.10/+0.10$ (b), $-0.11/-0.12$ (c) and $+0.66/+0.32$ (d); adjusting for mean glucose *and* variability gives $+0.32$, $+0.10$, $+0.20$ and $+0.21$ respectively. These are exploratory generative settings; they do not isolate a mechanism, and none of them is a validated human mechanism.

Statistic	**24 h (primary),** n=47 partial ρ vs SSPG [95% CI]	**48 h (sensitivity),** n=32 partial ρ vs SSPG [95% CI]
Relative forecast uncertainty (RFU)	-0.258 $[-0.52, +0.04]$	-0.437 $[-0.71, -0.07]^{\dagger}$
AR(3) normalized innovation ratio	-0.190 $[-0.48, +0.10]$	-0.233 $[-0.58, +0.16]$
Linear autoregression (normalized error)	-0.077 $[-0.34, +0.18]$	-0.060 $[-0.45, +0.29]$
Sample entropy	-0.093 $[-0.39, +0.17]$	-0.060 $[-0.46, +0.28]$
Detrended fluctuation analysis (persistence)	$+0.198$ $[-0.09, +0.46]$	$+0.436$ $[+0.11, +0.68]^{\dagger}$
Lag-1 autocorrelation	$+0.098$ $[-0.20, +0.39]$	$+0.160$ $[-0.24, +0.50]$
Glucose variability (coefficient of variation)	$+0.180$ $[-0.14, +0.49]$	$+0.068$ $[-0.32, +0.47]$

Table 2: **RFU against classical summaries of glucose dynamics.** Partial Spearman correlation with steady-state plasma glucose (SSPG), controlling for mean glucose and variability (the variability row is controlled for mean glucose only), in the pooled cohorts under identical cohort-internal percentile de-batching; bootstrap 95% intervals. †: interval excludes zero. At 24 hours no statistic's interval excludes zero and RFU has the largest point estimate; at 48 hours RFU and long-range persistence (DFA) do. The two are strongly coupled (ρ=-0.83 at 24 h, -0.85 at 48 h) and each attenuates when the other is controlled (RFU given DFA: -0.175 / -0.229; DFA given RFU: $+0.042$ / $+0.228$).

E WITHIN-SUBJECT TIME-SHUFFLE CONTROL

Within each context window we permute the sample order, which preserves the window mean, standard deviation, and every order statistic exactly and destroys all temporal structure, including any within-window diurnal shape; RFU is recomputed on the shuffled windows under the main protocol and averaged over three independent realizations. Trained RFU: pooled AUROC 0.780 (real) $\rightarrow$ 0.456$\pm$0.023 (shuffled); partial ρ vs SSPG -0.258 $\rightarrow$ $+0.194\pm0.041$; the forecaster's median per-window predictive scale increases from 9.9 to 29.8 mg/dL. Subject-level Spearman between real and shuffled RFU is 0.41–0.53. The representation-probe results under the same shuffle (untrained-encoder probe AUROC 0.715 $\rightarrow$ 0.539, adjusted 0.642 $\rightarrow$ 0.538; trained-encoder probe 0.724 $\rightarrow$ 0.566) use the probe protocol of Section 3 and are unchanged from the probe analysis. Script: `code/run_03_time_shuffle.py`.

F SENSITIVITY TO THE UNCERTAINTY SUMMARY AND TO THE AGGREGATION PROTOCOL

Window protocols of the classical statistics. Sample entropy (m=2, r=0.2), the detrended fluctuation exponent (scales $8, 16, 32, 64$), lag-1 autocorrelation and the AR(12) one-step normalized error are computed on the first 20 valid context windows at stride 48 (4 h), requiring only the context to be valid, and averaged per subject; the AR(3) normalized innovation ratio uses the RFU windows themselves (all qualifying windows, stride 24, context and horizon valid) and the RFU aggregation. All statistics then enter the same cohort-internal percentile de-batching. The comparison therefore holds the de-batching and the outcome fixed, not the observation windows.

Details referenced from Section 4.2. Mutual adjustment of RFU and long-range persistence (DFA), cohort-internal percentiles, partial Spearman given mean glucose and variability: RFU given DFA -0.175 (24 hours) and -0.229 (48 hours); DFA given RFU $+0.042$ and $+0.228$ (`results/v3_06_classical_ctx{288,576}.txt`). The untrained forecaster's AUROC on the ablation ladder ranges from 0.40 to 0.62 over five random initializations (`v3_15`). The handcrafted local-to-global volatility ratio has partial $\rho = -0.05$ with SSPG (`v3_02`). Binary discrimination of each classical statistic under one orientation rule, with paired differences to RFU, is given in Table 3.

Uncertainty summary. RFU uses the Student-t scale. Recomputing it from the predictive standard deviation $\hat{\sigma}\sqrt{\hat{\nu}/(\hat{\nu}-2)}$, the 80% predictive-interval half-width, and the predictive entropy (window value mean_h entropy $-$ log SD) gives partial $\rho = -0.272$, -0.250, and -0.237 (AUROC 0.764, 0.766, 0.746), each rank-correlated with the scale-based RFU at ≥ 0.98.

Statistic	AUROC [95% CI]	RFU − statistic [95% CI]
24-hour context, n=47 (16 resistant)		
Relative forecast uncertainty (RFU)	0.780 [0.64, 0.90]	—
AR(3) normalized innovation ratio	0.712 [0.55, 0.85]	+0.069 [−0.04, +0.19]
Linear autoregression, normalized error	0.681 [0.51, 0.84]	+0.099 [−0.03, +0.24]
Sample entropy	0.677 [0.50, 0.84]	+0.103 [−0.04, +0.26]
DFA (long-range persistence)	0.756 [0.60, 0.89]	+0.024 [−0.07, +0.13]
Lag-1 autocorrelation	0.706 [0.53, 0.86]	+0.075 [−0.05, +0.20]
Mean glucose (adjustment covariate)	0.661 [0.48, 0.82]	+0.119 [−0.05, +0.30]
Coefficient of variation (adjustment covariate)	0.641 [0.48, 0.80]	+0.139 [−0.00, +0.28]
48-hour context, n=32 (11 resistant)		
Relative forecast uncertainty (RFU)	0.840 [0.69, 0.95]	—
AR(3) normalized innovation ratio	0.736 [0.53, 0.91]	+0.104 [−0.04, +0.25]
Linear autoregression, normalized error	0.697 [0.48, 0.90]	+0.143 [−0.03, +0.31]
Sample entropy	0.680 [0.46, 0.87]	+0.160 [+0.02, +0.30]
DFA (long-range persistence)	0.779 [0.60, 0.92]	+0.061 [−0.05, +0.18]
Lag-1 autocorrelation	0.749 [0.54, 0.92]	+0.091 [−0.03, +0.23]
Mean glucose (adjustment covariate)	0.628 [0.41, 0.84]	+0.212 [−0.00, +0.43]
Coefficient of variation (adjustment covariate)	0.623 [0.42, 0.81]	+0.216 [+0.03, +0.40]

Table 3: **Binary discrimination of classical statistics and of the adjustment covariates under one orientation rule fixed in advance.** Each dynamics statistic is scored so that the more regular or more predictable direction (lower sample entropy, lower autoregressive error or innovation ratio, higher persistence, higher lag-1 autocorrelation, lower RFU) counts as insulin resistant, the direction reported for continuous glucose monitoring complexity (Crenier et al., 2016) and the sign convention of RFU; the two covariates are scored higher = resistant. AUROC on raw pooled values with 95% subject-bootstrap intervals (2,000 resamples, the same draw as Section 4.1); the difference column is the paired bootstrap difference, RFU minus the statistic, on the same resamples (`results/v3_20_baseline_auroc.txt`).

Aggregation protocol. Three choices differ between the main protocol and an earlier version of this analysis: the aggregation (main: subject mean of per-window mean ratios; alternative: subject median of per-window log root-mean-square ratios, $\mathrm{RFU}_i^{\mathrm{alt}} = \mathrm{median}_w \log\big(\sqrt{\frac{1}{H}\sum_h \hat{\sigma}_h^2}/\mathrm{SD}(g_{i,w})\big)$, Eq. (8)); a per-subject cap on the number of windows (30 or 40, evenly spaced); and the hour input (clock hour versus a constant). Table 4 reports every forecaster under every variant. Binary discrimination is retained in all 16 cells at 24 h (AUROC 0.69–0.80, two-sided permutation $p \le 0.024$), but effect sizes shift, most for TimesFM-3, and the adjusted continuous association weakens for all four forecasters under the median-of-logs aggregation. The main protocol was adopted after this audit on the grounds that it introduces no per-subject cap and admits the monotonicity result of Appendix B.

$$\mathrm{RFU}_i^{\mathrm{alt}} = \mathrm{median}_w \, \log\Big(\frac{\big(\frac{1}{H}\sum_h \hat{\sigma}_{i,w,h}^2\big)^{1/2}}{\mathrm{SD}(g_{i,w})}\Big). \tag{8}$$

G Representation alignment and mechanism: full figures

Probe protocol. The ridge probe uses its own windows, which are shorter than the RFU windows: 39 five-minute samples, stride 24, all samples valid, up to 40 evenly spaced windows per subject, embedded by the frozen encoder (constant hour input; time-averaged to 320 dimensions) and mean-pooled per subject. It is fitted on the AI-READI training split (up to 20 windows per subject) against log HOMA-IR after per-fold standardization, with ridge penalty 10. The AI-READI out-of-fold score uses five subject-level folds (shuffled, seed 0), fitting on the training folds' windows and averaging the held-out subject's window predictions (`code/run_07_compression.py`).

Forecaster	Variant	AUROC	perm p	partial ρ	perm p
PG	main (all windows, mean ratio)	0.780	0.0009	−0.258	0.092
PG	cap 30, mean ratio	0.796	0.0004	−0.319	0.032
PG	median-of-logs, all windows	0.766	0.0014	−0.223	0.151
PG	median-of-logs, cap 40	0.774	0.0010	−0.244	0.114
PatchTST	main	0.746	0.0033	−0.197	0.204
PatchTST	cap 30, mean ratio	0.758	0.0017	−0.248	0.100
PatchTST	median-of-logs, all windows	0.716	0.0123	−0.121	0.437
PatchTST	median-of-logs, cap 40	0.698	0.0206	−0.072	0.648
NeurODE	main	0.754	0.0031	−0.211	0.176
NeurODE	cap 30, mean ratio	0.780	0.0018	−0.278	0.064
NeurODE	median-of-logs, all windows	0.728	0.0075	−0.181	0.242
NeurODE	median-of-logs, cap 40	0.722	0.0111	−0.115	0.463
TimesFM-3	main	0.748	0.0028	−0.221	0.151
TimesFM-3	cap 30, mean ratio	0.794	0.0002	−0.325	0.029
TimesFM-3	median-of-logs, all windows	0.692	0.0238	−0.100	0.519
TimesFM-3	median-of-logs, cap 40	0.720	0.0094	−0.162	0.288

Table 4: **Protocol variants, 24-hour context, clean pooled set (n=47).** AUROC permutation p-values are two-sided orientation-free; partial-correlation p-values are Freedman–Lane for the main protocol and within-cohort outcome permutations for the variants. All variants use the clock-hour input; the constant-hour variant changes the primary forecaster's AUROC by less than 0.02 (Appendix code).

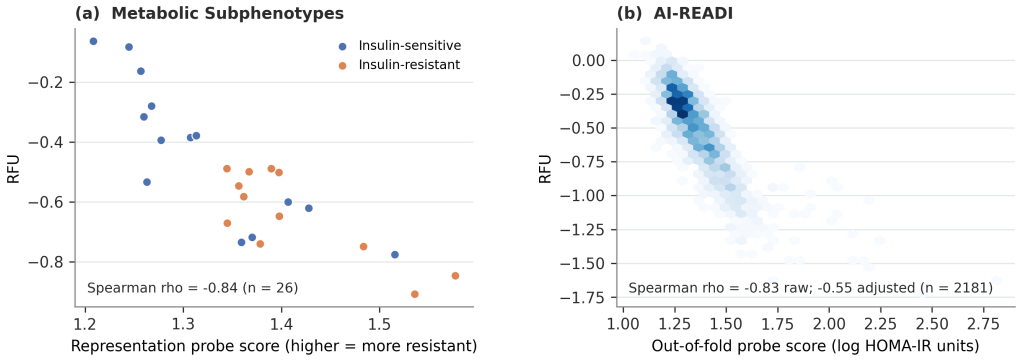


Figure 6: **RFU and the representation probe. (a)** Metabolic Subphenotypes (n=26): RFU versus the supervised probe's insulin-resistance score (Spearman ρ=−0.84). **(b)** AI-READI (n=2181): RFU versus the five-fold out-of-fold probe score (raw ρ=−0.83; −0.55 after adjusting for glucose level and variability).

H Primary forecaster specification, retraining controls, and one-directional readout

Primary forecaster: specification. The primary forecaster (Anonymous, 2026a) uses the following architecture and training procedure. Each context window of glucose at 5-minute cadence is instance-normalized by its own mean and standard deviation (the same statistics denormalize the output), and the model consumes five channels per step: the normalized glucose, its first and second differences, and the sine and cosine of the hour of day; no insulin, meal, or demographic input is used. A learned affine map projects the channels to 320 dimensions. Three causal one-dimensional convolutions (kernel size 3; dilations 1, 4, and 12, i.e. nominal per-layer spans of 15, 45, and 125 minutes) are composed as a residual cascade with GELU, dropout (0.12), and layer normalization, followed by a 4-layer Transformer encoder shared across horizons. The encoder's final state and a causal six-hour rolling standard deviation of the input enter a horizon-conditioned two-expert

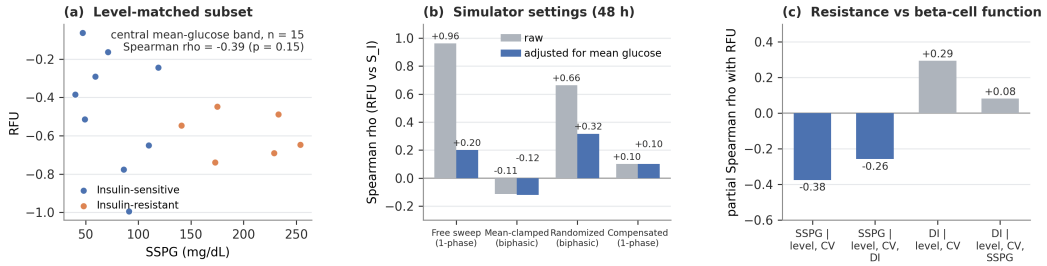


Figure 7: **Level-matched analyses and simulator settings. (a)** Central mean-glucose band (n=15, ρ=−0.39) and level-matched pairs (9/13). **(b)** Four exploratory simulator settings (Appendix D), raw and mean-glucose-adjusted: single-phase free sweep, mean-clamped biphasic, independently randomized biphasic, and compensated single phase. They differ in several respects at once and are not a nested ladder. **(c)** Resistance versus β-cell function in Metabolic Subphenotypes (n=24).

decoder (a learned per-horizon gate over two multilayer-perceptron experts with a horizon embedding), and a linear head emits Student-t parameters (μ_j, σ_j, ν_j) for each of the H=12 horizons, with $\sigma_j > 0$ and $\nu_j > 2$ enforced by softplus offsets and μ_j, σ_j mapped back to mg/dL by the window statistics. The model has 6,842,907 parameters. It was pretrained once on four constituent databases of the DiaData consortium (Cinar & Maleshkova, 2025) (1,052 participants with type 1 diabetes, about 52 million readings on a native 5-minute grid), at a 576-sample context only, minimizing the horizon-weighted Student-t negative log-likelihood (weight 1.20 on the first six horizons, 1.00 on the rest) with AdamW (learning rate 2×10^{-4}, weight decay 10^{-2}, one-cycle schedule, gradient clipping at 1.0, batch 24 per GPU, chunk stride 24) for 8 epochs under a participant-level split (seed 42). It is used frozen throughout this paper, at 288- and 576-sample contexts, without fine-tuning or adaptation.

Release scope. The DiaData cohorts carry their own data-use agreements, and whether these permit redistribution of derived model weights is unsettled; the pretrained checkpoint and the retrained control checkpoints are therefore not distributed with this submission, in line with the companion release. The submission includes the analysis code, RFU protocol module, per-window predictive scales and window standard deviations for all four forecasters, and per-subject control outputs. These support recomputation of the main RFU definition and reanalysis of the supplied control outputs. Realized-error and calibration analyses require additional predictions and corresponding targets; new inference requires the relevant checkpoint, and retraining requires training code and authorized data access. The withheld checkpoints will be released if the data-use terms permit.

Retraining controls. Retraining controls use the same data, split, and optimizer under a reduced budget (training stride 96, three epochs, batch 64); their AUROCs are below the primary checkpoint's. *No-volatility ablation*: with the decoder's 6-hour rolling standard deviation input zeroed, the gold-standard readout under the main protocol is AUROC 0.764 (partial $\rho = -0.237$) versus 0.772 (-0.284) with the input, and the two models' subject-level RFU rankings agree at Spearman 0.996. *Trained-seed robustness*: three independently seeded forecasters give AUROC 0.770±0.003 (partial $\rho = -0.265 \pm 0.014$), with seed-to-seed RFU Spearman 0.995 (0.993–0.997). These checkpoints are not distributed (release scope above); their per-subject outputs are.

One-directional readout (label conditioning, not a control loop). Conditioning the forecaster on a subject's known fasting-proxy insulin resistance does not improve held-out forecasts (incremental out-of-fold $R^2 = +0.003$; n=2181). We report this as a negative predictive-control result and do not infer causal one-directionality from it.