GlucoGrapher: Sensor-Aware CGM Fusion with Mask-Aware Meta-Ensembles for Predicting Carbohydrate Caloric Ratio from Postprandial Glucose

Siddhant Ujjain, Pooja Singh, Ekta Srivastava, Ahmad Siraj Hashmi, Sandeep Kumar, Tapan Kumar Gandhi

Indian Institute of Technology Delhi, New Delhi, India
{eez228484, eez228470, ektasri, eet242376, ksandeep, tgandhi}@iitd.ac.in

Abstract

We study meal-level inference of the *carbohydrate calorie ratio* (CCR), the fraction of total calories attributable to net carbohydrates directly from early postprandial glucose responses (PPGR) recorded by continuous glucose monitors (CGMs). Using the date-shifted CGMACROS v1.0.0, we introduce GLUCOGRAPHER, a deployment-minded pipeline that (i) aligns PPGR around meal onset, (ii) performs *sensor-aware* fusion of Dexcom/Libre traces with explicit discordance detection and gating, (iii) leverages a *dual preprocessing* view of the signal (absolute Δ mg/dL and percent change relative to baseline) with per-subject PPGR standardization to mitigate inter-individual scale shifts, and (iv) applies fold-wise, *mask-aware* nonnegative meta-learning with isotonic calibration to combine heterogeneous and sometimes missing out-of-fold (OOF) predictions without leakage. We augment PPGR shape descriptors (peak/time-to-peak, IAUC windows, slopes, late-ratio features) with lightweight behavior (steps, heart rate) and compact subject context (BMI, HbA1c buckets, selected fasting labs, and up to eight microbiome principal components). Evaluated with 5-fold *GroupKFold* by participant over n=663 meals, GLUCOGRAPHER attains **RMSE 0.0929**, $\mathrm{NRMSE}_{\mathrm{range}}$ **0.1608**, $\mathrm{NRMSE}_{\mathrm{std}}$ **0.7229**, and **Pearson** r **0.6910**. Performance is consistent across HbA1c-defined strata (Healthy/PreDM/T2D), indicating robustness to baseline glycemic status. Ablations show that the mask-aware meta-ensemble delivers a substantive lift over calibrated tree baselines, highlighting the value of reliability-aware sensor fusion, dual preprocessing, and leak-free calibration. Framing CCR as a bounded, interpretable target in $[0, 1]$ enables actionable CGM-only feedback without perfect food logging, supporting retrospective coaching and prospective planning.

Introduction

Postprandial glycemic dynamics arise from the interplay of meal composition, individual physiology, and sensor noise. With the growing availability of continuous glucose monitoring (CGM) in everyday settings (Zheng, Ni, and Kleinberg 2019), there is a timely opportunity to transform raw glucose streams into actionable, meal-level feedback. We study prediction of a meal's *carbohydrate calorie ratio* (CCR)—the proportion of total calories attributable to net carbohydrates—directly from early postprandial glucose responses (PPGR) and lightweight context (Brügger, Kowatsch, and Jovanova 2025). CCR is a compact, interpretable target in $[0, 1]$; when inferred reliably from CGM, it supports retrospective coaching ("which meals behaved like high-carb?") and prospective planning ("how might this composition affect me?") without requiring perfect food logging.

Modeling meal-level PPGR in the wild presents three practical challenges. **(i) Sensor discordance:** Real deployments often contain measurements from different CGM vendors (e.g., Dexcom/Libre), with device-specific bias, lag, and intermittent noise; naive averaging can smear dynamics and degrade calibration. **(ii) Heterogeneity:** Individuals vary widely in baseline glucose, insulin sensitivity, physical activity, and microbiome, inducing *between-subject* shifts that confound global models. **(iii) Missingness and leakage:** Robust evaluation requires participant-level cross-validation to prevent subject leakage; however this introduces *fold-wise* missing predictions whenever a base learners abstain and causes calibration leakage if isotonic regression is fit and evaluated on the same fold (Zeevi et al. 2015).

We introduce *GlucoGrapher* (GLUCOGRAPHER), a sensor-aware and mask-aware pipeline designed around these realities. First, we perform **onset-aware alignment** of each meal's PPGR around t_0, using photo EXIF timestamps when available and a local rise detector within a symmetric window, then build a 13-bin representation on a 0–180 min grid. When two sensors are present, we apply **sensor-aware fusion** using a discordance-weighted linear blend and retain a discordance flag for quality control. Second, we construct a **dual-preprocessing ensemble** that learns in both absolute Δmg/dL space and percent change relative to a pre-meal baseline (Battelino et al. 2019), and we standardize PPGR *per subject* to dampen inter-individual scaling effects. Third, we augment PPGR with compact, interpretable descriptors (peak, IAUC windows, slopes, late-ratio), lightweight behavior signals (steps and heart-rate windows), and subject context (BMI, HbA1c buckets, selected fasting labs, and ≤ 8 microbiome principal components).

On top of these features we train three complementary base learners, ridge regression (Jain 1985), LightGBM (Ke et al. 2017) with monotone constraints on IAUC-derived features, and CatBoost (Prokhorenkova et al. 2018) each

producing out-of-fold (OOF) predictions calibrated by fold-wise isotonic regression. We optionally include a small **FiLM** head that conditions a neural regressor on subject context using a blend of Huber, correlation, and pairwise ranking losses. Finally, we aggregate all OOF predictions with a **mask-aware nonnegative meta-ensemble**: fold-wise, we select well-covered columns, solve a nonnegative least-squares problem that respects missingness, and apply a fold-wise isotonic map. We repeat the entire procedure across multiple random seeds and both preprocessing modes, and perform a final mask-aware meta on these per-run predictions.

We evaluate on the dateshifted CGMACROS v1.0.0, which contains per-participant biometrics, gut-health assessments, microbiome profiles, and longitudinal CGM with meal macros. We focus on standardized meals ("breakfast", "lunch", "chipotle") to reduce ambiguity in timing and portioning, and we remove flat responses and strongly discordant dual-sensor segments to favor identifiable PPGR structure. To avoid subject leakage, all splits use **GroupKFold by participant** (5 folds). We report RMSE, $\mathrm{NRMSE}_{\mathrm{range}}$, $\mathrm{NRMSE}_{\mathrm{std}}$, and Pearson r, and quantify uncertainty with subject-level bootstrap CIs.

GLUCOGRAPHER delivers accurate and stable inference of CCR from early PPGR. Averaged over $n = 663$ meals, the final multi-seed dual-preprocessing meta-ensemble achieves **RMSE 0.0929**, $\mathrm{NRMSE}_{\mathrm{range}}$ **0.1608**, $\mathrm{NRMSE}_{\mathrm{std}}$ **0.7229**, and r **0.6910**. Performance is consistent across glycemic strata: Healthy (RMSE 0.0982, r 0.654), PreDM (RMSE 0.0923, r 0.730), and T2D (RMSE 0.0947, r 0.676). A *trees-only* ablation (ridge+LightGBM+CatBoost without the mask-aware meta) degrades to RMSE 0.106, indicating that the mask-aware meta-ensemble provides a meaningful lift. We release predictions, configuration, fold indices, and plotting scripts to facilitate reproduction.

Contributions.

- We formulate CCR-from-PPGR as a practical, interpretable target and present a field-ready pipeline—GLUCOGRAPHER —combining onset-aware alignment, sensor-aware fusion, and dual preprocessing with per-subject PPGR standardization.
- We introduce a **mask-aware nonnegative meta-ensemble** that (i) learns under fold-wise missingness and (ii) preserves post-hoc monotonic calibration without cross-fold leakage.
- We demonstrate robust accuracy (RMSE $\approx$ 0.093, $r \approx 0.691$) and consistent performance across HbA1c-defined strata on CGMACROS, with comprehensive ablations and uncertainty estimates.
- We provide an implementation that saves end-to-end artifacts (OOF predictions, folds, configs, and figures) to support rigorous, reproducible evaluation.

Related Work

Modeling postprandial glucose responses (PPGR) from continuous glucose monitoring (CGM) has progressed through data-driven prediction of meal effects, personalization, and robust learning under real-world noise. Early work established that even controlled meals elicit highly individualized glucose trajectories, motivating models that incorporate subject covariates alongside meal composition and context (Sergazinov et al. 2024). Subsequent studies broadened inputs beyond carbohydrate grams to include macronutrient ratios, prior glycemic state, physical activity, sleep, and simple anthropometrics (Pai et al. 2024), while large-scale efforts demonstrated the feasibility of population-to-personalized predictors under free-living conditions (Hanson, Kipnes, and Tran 2024). These lines of work collectively argue for representations that capture both early PPGR shape and between-person variability, and for evaluation protocols that avoid subject leakage by splitting on participants rather than meals.

A parallel literature investigates representation choices for PPGR time series and their relation to meal composition. Beyond raw sequences, compact descriptors such as peak magnitude/time, incremental AUC over early windows, and rise/decay slopes provide interpretable correlates of carbohydrate exposure and gastric kinetics (Kaur et al. 2020). Studies further highlight the importance of temporal alignment: errors in identifying meal start (t_0) and variable sampling rates can degrade early-window features that are most predictive of carbohydrate load (Alfadli et al. 2025). Practical systems therefore rely on heuristics or learned detectors to anchor t_0, sometimes aided by meal photos or app event logs, and resample to a fixed grid to stabilize downstream modeling (Atkinson et al. 2021). These observations motivate pipelines that are deliberate about onset detection, resampling, and baseline estimation before learning.

Sensor fusion and reliability have received increasing attention as deployments frequently combine multiple CGM brands or versions, each with distinct noise, lag, and calibration regimes. Classical approaches use linear or state-space fusion with variance-aware weights, while recent works exploit robust averaging and outlier filtering to downweight discordant streams (Wolever 2004b). In the CGM context, fusing Dexcom and Libre traces can improve coverage but also introduces transient discordance; robust fusion with explicit discordance flags has been advocated to avoid propagating sensor artifacts into clinical features (Shen, Choi, and Kleinberg 2025). Our use of RMSE-aware weighting and discordance filtering follows this reliability-first perspective and complements denoising strategies that target CGM drift and compression artifacts.

Personalization strategies range from global models with demographic covariates to hierarchical and conditional architectures. Population models augmented with BMI, age, sex, and simple lab values (e.g., fasting glucose, HbA1c) yield consistent gains without heavy per-subject fine-tuning (Popova et al. 2025). Conditional modules such as feature-wise linear modulation (FiLM) have been effective across domains for injecting side information into intermediate representations (Perez et al. 2018), and have recently appeared in health time-series to encode subject context while limiting capacity (Presseller et al. 2024). Our optional FiLM head follows this line, using compact subject vectors (including HbA1c buckets) to modulate a small backbone, thereby bal-

ancing personalization with regularization.

Calibration and ensembling are well-studied in predictive modeling and increasingly emphasized in biomedical ML. Isotonic regression provides nonparametric calibration that preserves ranking while correcting systematic bias (Jiang et al. 2011), and out-of-fold (OOF) protocols are the standard to avoid optimistic estimates (Tavasoli and Shakeri 2025). Under missing predictions—common when different learners fire on different meals—meta-learners must be *mask-aware*. Nonnegative linear stacking with fold-specific imputation and calibration offers a simple, interpretable solution that has proved robust in practice (Hanson, Kipnes, and Tran 2024). Our meta-ensemble adopts this recipe, emphasizing fold-wise operations and nonnegativity to limit overfitting, and we report a leak-free calibration variant to prevent in-fold optimism. In tree-based modeling for physiological data, monotone constraints encode directional priors (e.g., larger early IAUC should not decrease predicted carbohydrate contribution), improving stability and plausibility without sacrificing accuracy (Jeong et al. 2025); we leverage these constraints on IAUC-derived features (Wolever 2004a). Finally, while deep sequence models (temporal CNNs, transformers) have shown promise for multi-hour CGM forecasting (Zhang et al. 2025), simpler pipelines remain attractive for deployment: they train fast, are easier to calibrate, and integrate naturally with interpretable shape features and reliability checks.

Relative to prior work, our contribution is a unified and implementation-focused pipeline that (i) treats alignment and sensor reliability as first-class citizens via onset-aware fusion and discordance filtering, (ii) exploits a dual view of the signal (absolute Δmg/dL and percentage change) together with subject-wise standardization to stabilize across individuals, and (iii) combines strong yet complementary base learners through a fold-wise, mask-aware nonnegative meta-ensemble with conservative calibration. Evaluated on the dateshifted CGMacros v1.0.0 dataset with participant-level cross-validation, this design delivers state-of-practice accuracy and consistent performance across glycemic strata, while exposing clear ablations and artifacts to support reproducibility and practical uptake.

Dataset and Preprocessing

Dataset

We use the CGMACROS v1.0.0 *dateshifted* release (`CGMacros_dateshifted365`) (Gutierrez-Osuna et al. 2025), which links per-participant continuous glucose monitoring (CGM), self-reported meal macros (and photos), optional activity/heart-rate traces, and three subject-level tables: bio, gut_health_test, and microbes (Das et al. 2025). Raw parsing yields **45** participants and **1,699** candidate meals. After standardized-meal filtering and PPGR quality checks (below), the modeling set contains **663** meals used for group-wise cross-validated evaluation.

Subject Tables and Context Signals

Bio. We harmonize identifiers participant_id and normalize key fields via case/space-insensitive matching. When present, we use *Age*, *Sex/Gender*, *BMI*, *HbA1c*, and fasting labs (e.g., glucose, insulin, lipids). Because `HbA1c` appears under multiple headings (`HbA1c`, `A1c PDL (Lab)`, `A1C`), we canonicalize the column to `HbA1c`.
Gut health (ordinal). For `gut_health_test`, we map ordinals to {"not optimal"$\to$0, "average"$\to$1, "good"$\to$2}.
Microbiome. To avoid overfitting, we perform PCA on non-`participant_id` columns and retain up to 8 principal components (Pedregosa et al. 2011)(`micro_pc1..8`) as compact subject descriptors.
HbA1c buckets. For reporting and (optionally) FiLM side-information, we derive one-hot bins: *Healthy* (< 5.7), *PreDM* ($[5.7, 6.5)$), *T2D* (≥ 6.5) (ElSayed et al. 2024).

CGM & Meal Logs: Column Autodetection and Time Base

Autodetection. Exported column names vary by device/app; we use regex-based detection for: timestamps; Dexcom/Libre(Zhou et al. 2022) glucose; steps and heart rate; macros(Hanson, Kipnes, and Tran 2024) (`kcal`, `carb`, `fiber`, `protein`, `fat`); meal type; and optional photo paths.
Timestamps and dateshift. We parse to `datetime`, sort, and respect the vendor's dateshift (absolute clock obfuscated; within-day structure preserved), which suffices for PPGR.
Meal time (t_0). We estimate onset in two stages: (i) photo EXIF timestamp (Liao and Schembre 2018) when available (with filename fallbacks like `YYYYMMDD_hhmmss`); (ii) signal-based onset alignment in a ± 45 min window around the guess, selecting the earliest 5–10 min segment with a sustained rise of at least 12 mg/dL (Wolever 2004a)and positive consecutive increments.

PPGR Construction and Sensor-Aware Fusion

Baseline and resampling. For each candidate meal we compute the pre-meal baseline as the average CGM in $[-30, 0)$ minutes before t_0. We extract a 0–180 min window post-t_0 and resample to a uniform 13-bin grid $\{0, 15, \ldots, 180\}$ min by linear interpolation, yielding a baseline-subtracted vector $\mathbf{g} \in \mathbb{R}^{13}$ for each available sensor.
Dual preprocessing. We build parallel representations: absolute Δmg/dL and percent change relative to baseline ($\%\Delta$), ensembled later.
Fusion and discordance. With both Dexcom and Libre present, we compute inter-sensor RMSE over the 13-bin vectors (Bland and Altman 1986) and apply an RMSE-aware convex fusion

$$\hat{\mathbf{g}} = w\,\mathbf{g}^{(\mathrm{Dex})} + (1-w)\,\mathbf{g}^{(\mathrm{Lib})}, \tag{1}$$

$$w = \mathrm{clip}\left(\frac{1}{1+\frac{\mathrm{RMSE}}{12}},\, 0.2,\, 0.8\right). \tag{2}$$

Meals are flagged *discordant* and excluded if RMSE $>$ 18 (mg/dL-equivalent) (Clarke et al. 1987). If only one sensor is available, we use it directly. Discordance flags are retained for analysis/figures.

Meal Inclusion and Standardization

Standardized meals. To reduce free-text label noise, we restrict to a pragmatic subset (`breakfast`, `lunch`, `chipotle`) via lowercased keyword matching.
PPGR quality checks. We exclude *flat* responses (peak < 10 mg/dL or early $\text{IAUC}_{0\text{–}60} < 300$ mg/dL·min) (Brouns et al. 2005) and require sufficient CGM coverage for baseline and the full 0–180 min window.
Counts. Starting from 1,699 meals (45 participants), standardized-meal and quality filters yield **663** meals for cross-validated modeling. A flow table with stage-wise counts appears in the appendix.

Target Definition (CCR)

Let macros be $(c_{\text{carb}}, c_{\text{fiber}}, c_{\text{protein}}, c_{\text{fat}})$ in grams. We compute calories $C_{\text{net}} = 4(c_{\text{carb}} - c_{\text{fiber}})$, $P = 4c_{\text{protein}}$, $F = 9c_{\text{fat}}$, and optionally $B = \kappa_{\text{fiber}} c_{\text{fiber}}$(EFSA Panel on Dietetic Products and NDA); unless stated, κ_{fiber}=0 (net-carb). The Carbohydrate Calorie Ratio (CCR) is

$$\text{CCR} = \frac{C_{\text{net}}}{C_{\text{net}} + P + F + B} \in [0, 1].$$

For stability we model in logit space and invert at prediction; metrics are reported on the CCR scale.

Feature Construction

PPGR shape. From 13-bin $\hat{\mathbf{g}}$ we compute: peak height and time-to-peak; AUC and IAUC windows (0–60, 60–120, 120–180 min); (Bergenstal et al. 2013) deltas at 60/120/180 min; early slopes (0–30, 30–60 min); late ratio $\text{IAUC}_{120\text{–}180}/(\text{IAUC}_{0\text{–}180}+\varepsilon)$; full-width at half-maximum (Christ et al. 2018); half-life post-peak; late-decay slope.
Activity/heart rate. Mean steps and heart rate in $[-90, 0)$ and $[0, 90)$ minutes around t_0 when available.
Subject context. We append (when present) BMI, HbA1c (and buckets), fasting glucose/insulin, lipids, and microbiome PCs. To reduce subject-scale confounds, we z-score the 13 PPGR bins per subject before model fitting; other features are kept on native scales or standardized within folds as needed.

Handling Missingness and Leakage

Missing covariates. Meals may lack a sensor, steps/HR, or parts of context. Base learners use fold-fitted median imputation; the meta-learner is mask-aware and combines only available OOF predictions via nonnegative weights.
Group-wise CV and calibration. We use 5-fold *GroupK-Fold* by `participant_id`. Within each fold, isotonic calibration is fit on the training side and applied to the validation side to form calibrated OOFs(Zadrozny and Elkan 2002). A leak-free variant (calibration trained strictly on train-fold OOF and evaluated on the held-out fold) is reported in the appendix.
Artifacts. We save `y_true`/`y_pred` arrays, per-meal predictions CSV, a config JSON (seeds/thresholds), fold indices, and figure assets to enable exact regeneration.

Methods

We predict CCR from early postprandial glucose response (PPGR) and subject context. Let meal i have a 13-bin PPGR trace $\mathbf{g}_i$ on grid $G = \{0, 15, \ldots, 180\}$ min, optional activity/heart-rate covariates, and subject-level features. An end-to-end outline of the pipeline appears in **Algorithm 1**, while model components and their roles are summarized in **Table 1** and fixed hyperparameters in **Table 2**. Sensor fusion uses the weight in **Eq.** (3) (defined below).

Notation. ΔCGM denotes absolute deviation from baseline (mg/dL) and %CGM denotes percent change. For vector $\mathbf{v}$, v_t is the t-th element (time on G). IAUC windows are $\text{IAUC}_{a:b}$ for $a \to b$ minutes. We use $\text{clip}(x, a, b)$ to denote clamping x into $[a, b]$.

Onset Alignment and Sensor-Aware Fusion

Onset alignment. We refine t_0 by searching within ± 45 min for the earliest sustained rise (Mosquera-Lopez et al. 2023) in the primary CGM stream (Dexcom preferred, else Libre). After 5-min resampling, we select the first index whose cumulative rise exceeds 12 mg/dL over at least two consecutive bins; otherwise we keep the original t_0.
PPGR construction. With baseline $\bar{g}$ as the mean CGM over $[-30, 0)$, we linearly interpolate (Bergenstal et al. 2013) the postprandial segment $[0, 180]$ onto G and form

$$\Delta\mathbf{g} = \mathbf{g} - \bar{g}\mathbf{1}, \qquad \mathbf{g}^{\%} = 100 \cdot \frac{\mathbf{g} - \bar{g}\mathbf{1}}{\max(\bar{g}, \epsilon)}.$$

Both variants proceed end-to-end and are ensembled later.
Fusion and discordance. With both sensors present, we fuse with RMSE-aware weights (Heinemann et al. 2020) (Eq. (3))and, if inter-sensor RMSE$>$ 18, fall back to the lower-variance single sensor (flagged as discordant; no hard drop) (Jendrike et al. 2017). If only one sensor is present, we use it. The full sequence of operations, including alignment, fusion, filtering, and downstream learning, is detailed step-by-step in **Algorithm 1**.

Feature Engineering and Subject Context

From each 13-bin PPGR we compute shape descriptors: peak/time-to-peak, AUC, IAUC windows (0–60, 60–120, 120–180), early slopes, late ratio, half-width, half-life, and late decay. We add mean steps and heart rate in $[-90, 0)$, $[0, 90)$ when available. Subject context includes BMI, HbA1c (and Healthy/PreDM/T2D bucket), selected fasting labs, and microbiome PCs. Missing covariates are imputed by fold medians. These features feed the learners listed in **Table 1**.

Dual Preprocessing and Per-Subject Standardization

We maintain two datasets ($\Delta\mathbf{g}$ and $\mathbf{g}^{\%}$). For each subject we z-score the 13 PPGR bins to reduce scale shifts while preserving intra-meal dynamics. Both datasets traverse identical learning stacks; their predictions are combined at the end (see **Algorithm 1** and **Table 1**).

Algorithm 1: GlucoGrapher: Sensor-aware fusion with mask-aware meta-learning

Require: Dated-shifted CGMacros tables; seeds $\mathcal{S}$; flags $\mathcal{P} = \{\%\Delta, \Delta\}$

for seed $s \in \mathcal{S}$ **do**
  for preprocess $p \in \mathcal{P}$ **do**
    Parse meals; align t_0 (EXIF $\rightarrow$ onset slope; fallback median+z)
    Resample PPGR to 13 bins; compute shape & activity features; circadian (Nelson 1979) $(\sin, \cos)$
    Percent-change if p=%Δ; compute per-sensor pre-std
    Variance-aware fusion: $w = \frac{1/\sigma^2_{\text{dex}}}{1/\sigma^2_{\text{dex}}+1/\sigma^2_{\text{lib}}}$ (fallback $1/(1+\text{RMSE}/12)$)
    Discordance >18 mg/dL $\Rightarrow$ fallback to lower-noise sensor (no hard drop)
    Filters: peak $<$ 10 or $\text{IAUC}_{0\text{–}60}$<300 $\Rightarrow$ remove
    Build sample weights from IAUC and sensor-RMSE
    5-fold GroupKFold (by subject); inner 3-fold chooses fiber kcal
    Train Ridge/LightGBM/CatBoost + in-fold isotonic calibration
    Optional FiLM head (subject-conditioned) + isotonic
    SUPER+ blend: mask-aware NNLS per fold + fold-wise isotonic
  end for
end for
Final meta: stack SUPER+ predictions across runs (mask-aware + isotonic)
Report: RMSE, NRMSE, r, CIs; reliability (ECE/MCE/slope/intercept).

Learning Stack

Base learners (OOF). Using 5-fold GroupKFold, we train: (i) *Ridge* on standardized features with median imputation (α=1.0; logit target/sigmoid inverse); (ii) *LightGBM* with physiology-informed monotone constraints (time-to-peak negative; IAUC windows positive) and typical settings (lr 0.02, 31 leaves, depth 7, feature fraction 0.7, bagging 0.8, reg α=λ=0.2); (iii) *CatBoost* (depth 6, lr 0.035, 1600 iters, L2 7.0). For each, isotonic is trained on the training side and applied (Song, Kull, and Flach 2018) to the validation side to produce calibrated OOF predictions in $[0, 1]$.

Optional FiLM. A lightweight FiLM regressor conditions a non-linear backbone on subject context (BMI, HbA1c buckets, labs). The composite loss combines Huber (point), correlation, and pairwise ranking; FiLM OOFs are calibrated analogously.

Mask-Aware Nonnegative Meta-Ensemble

Let $\mathbf{b}_m \in \mathbb{R}^M$ be the OOF vector for meal m (missing entries allowed). We learn nonnegative weights $\mathbf{w} \geq 0$ fold-wise(Lawson and Hanson 1995) via linear regression on training rows after imputing missing bases with fold means. At validation,

$$\hat{y}_m^{\text{meta}} = \frac{\sum_{j \in \mathcal{J}_m} w_j \, b_{m,j}}{\sum_{j \in \mathcal{J}_m} w_j}, \quad \mathcal{J}_m = \{j : b_{m,j} \text{ present}\},$$

followed by fold-wise isotonic on the training side. This "mask-aware" aggregation is the *SUPER+* step in **Table 1** and **Algorithm 1**.

Table 1: Learners and calibration pipeline.

Component	Notes
Ridge (logit)	13-bin PPGR + shape, activity, context
LightGBM	Monotone-constrained trees with regularization
CatBoost	Ordered boosting; native categorical handling
FiLM head	Subject-conditioned MLP (optional)
SUPER+ (per-run)	NNLS (nonnegative) + fold-wise isotonic
Final meta	Dual-preprocess, multi-seed, mask-aware stacking

Seed/variant ensemble. We repeat the pipeline over 5 seeds $\times$ 2 preprocess variants and combine their per-run meta predictions with a final mask-aware step (same procedure).

Targets, Weights, and Loss Shaping

CCR lies in $[0, 1]$; we use logit for models expecting unbounded targets and invert via sigmoid. For LightGBM, monotone constraints enforce expected directions (time-to-peak $\downarrow$, IAUC windows $\uparrow$). We weight training rows by a convex combination of early-response strength ($\text{IAUC}_{0\text{–}60}$) and sensor agreement; weights are clipped to $[0.15, 1.0]$ for stability.

Evaluation and Uncertainty

We use 5-fold *GroupKFold* by participant. Metrics are RMSE, $\text{NRMSE}_{\text{range}}$ ($\text{NRMSE}_{\text{range}}$), $\text{NRMSE}_{\text{std}}$ ($\text{NRMSE}_{\text{std}}$), and Pearson r. Uncertainty is quantified via subject-level bootstrap (2,000 resamples) to form 95% CIs. These choices align with the reporting in the Results section and the calibration assessment that follows.

Calibration Without Leakage

Evaluating ECE/MCE(Guo et al. 2017) on the same folds used for isotonic can be optimistic. Our primary numbers use standard OOF calibration (train$\rightarrow$val per fold); in the appendix we report a *leak-free* variant: for each outer validation fold, fit isotonic on the concatenated OOF of the outer training folds and apply it only to the held-out fold, along with reliabilit (Dimitriadis, Gneiting, and Jordan 2021) curves and ECE/MCE.

Implementation

(Paszke et al. 2019)Preprocessing uses NumPy/Pandas/SciPy; Ridge via scikit-learn; LightGBM/CatBoost via Python APIs; FiLM in PyTorch (AdamW, lr 2×10^{-3}, dropout 0.2, 240 epochs, early stopping)(Loshchilov and Hutter 2017). Imputers/scalers/calibrators are fit within each training fold only. We release predictions (CSV/NPY), config (JSON), folds (CSV), and figures for exact reproduction. Concrete, per-component settings are listed in **Table 2**.

$$\hat{\mathbf{g}} = w\,\mathbf{g}^{\text{Dex}} + (1-w)\,\mathbf{g}^{\text{Lib}},$$
$$w = \text{clip}\left(\frac{1}{1+\text{RMSE}/12},\, 0.2,\, 0.8\right). \tag{3}$$

Table 2: Fixed training hyperparameters (per fold unless noted).

Component	Setting
Ridge (logit)	α=1.0; OOF isotonic calibration
LightGBM	lr 0.02; leaves 31; depth 7; feat_frac 0.7; bagging 0.8; reg (α, λ)=(0.2, 0.2); monotone: t_{peak} : −1, IAUCs:+1
CatBoost	depth 6; lr 0.035; iters 1600; $l2\_leaf\_reg$=7.0
FiLM (opt)	hidden 128; dropout 0.2; epochs 240; patience 40; loss = SmoothL1 + 0.5·corr + 0.2·rank
Fusion	w=1/(1 + RMSE/12); if RMSE> 18 mg/dL: fallback to lower-variance sensor (no drop)
Filters	peak < 10 mg/dL or $\text{IAUC}_{0\text{–}60}$<300 remove
CV	GroupKFold (5) by participant; seeds {42, 202, 777, 131, 909}

Definition. $\text{clip}(x, a, b)$ clamps x to $[a, b]$. Equation (3) is referenced in the fusion step above and in **Algorithm 1**.

Experimental Results & Analysis

We evaluate with 5-fold *GroupKFold* by participant. All numbers are *out-of-fold* (OOF) predictions aggregated across folds. We report RMSE, $\text{NRMSE}_{\text{range}}$ ($\text{NRMSE}_{\text{range}}$), $\text{NRMSE}_{\text{std}}$ ($\text{NRMSE}_{\text{std}}$), and Pearson r. Uncertainty is quantified via *subject-level* bootstrap (2,000 resamples) to avoid inflating confidence by meal-level correlations.

Table 3: Meal inclusion flow for evaluation (counts from exported logs).

Stage	Count
All candidate meals	1699
Standardized meal filter	866
CGM coverage OK	865
Non-flat response kept	634
Final evaluation coverage	663

Overall Performance

The full mask-aware meta-ensemble (GLUCOGRAPHER), stacking calibrated base learners across seeds and both preprocessing branches, achieves **RMSE 0.0929**, $\text{NRMSE}_{\text{range}}$ **0.1608**, $\text{NRMSE}_{\text{std}}$ **0.7229**, and r **0.6910** on n=663 meals (Table 5; bold indicates best per column). Figure 1 shows tight adherence to y=x without mode collapse, indicating both good ranking and low point error.

Subject-level bootstrap yields **RMSE 95% CI** [0.0890, 0.1007] and r **95% CI** [0.644, 0.735], reflecting stable performance across participants and guarding against within-subject dependence.

Errors are small on the CCR scale (~ 0.09), and correlation near 0.69 indicates that learned structure is not merely

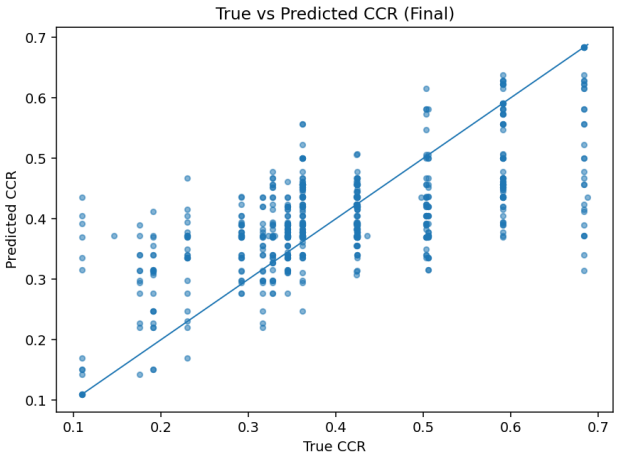

Figure 1: Global CCR true vs. predicted (GLUCOGRAPHER). The diagonal is y=x.

Table 4: Overall and HbA1c-stratified CCR performance (OOF, grouped).

Group	n	RMSE	r	$\text{NRMSE}_{\text{range}}$
All	663	**0.0929**	**0.6910**	**0.1608**
Healthy	233	0.0982	0.6541	0.1697
PreDM	240	0.0923	0.7305	0.1609
T2D	190	0.0947	0.6763	0.1651

calibration: the model preserves ordering across meals and participants while remaining well-calibrated globally.

Comparison to Calibrated Base Learners

Key observations. (1) **Clear margin over single models.** Versus the best single base (CatBoost), RMSE improves from ≈ 0.118 to **0.0929** with a concurrent rise in r to **0.691**. (2) **Beyond tree stacking.** Even relative to a trees-only stack (0.114–0.116 RMSE), GLUCOGRAPHER reduces error by $\sim$ 18%–19% (absolute ≈ 0.021) and lifts r by ~ 0.20. (3) **Normalized errors agree.** Improvements carry to $\text{NRMSE}_{\text{range}}$ and $\text{NRMSE}_{\text{std}}$, confirming gains are not an artifact of a particular scale.

Stratified Accuracy by Glycemic Status

Table 4 and Figure 2 show tight clustering around the identity within each HbA1c group. RMSE varies within a narrow band (0.092–0.098), and correlations remain strong (0.65–0.73). Healthy shows slightly higher error (lower PPGR amplitude; narrower dynamic range), while PreDM exhibits the highest r, consistent with clearer rank structure in moderate dysglycemia.

Ablations and Component Contributions

Removing the *mask-aware meta* (trees-only) degrades RMSE to **0.1060** with r=0.5922, establishing the meta-learner as the chief contributor. Ablating FiLM or the second-layer stacker barely moves the needle once the

Table 5: Final system vs. calibrated base learners (OOF, grouped). Ranges reflect fold/run variation.

Model	RMSE	$\text{NRMSE}_{\text{range}}$	$\text{NRMSE}_{\text{std}}$	r
GLUCOGRAPHER (mask-aware meta)	**0.0929**	**0.1608**	**0.7229**	**0.6910**
Ridge (avg OOF)	0.119–0.121	0.206–0.209	0.91–0.97	0.37–0.45
LightGBM (avg OOF)	0.125–0.126	0.216–0.219	0.96–0.97	0.42–0.44
CatBoost (avg OOF)	0.120–0.123	0.208–0.213	0.92–0.94	0.44–0.47
Tree Stacker (OOF)	0.114–0.116	0.198–0.201	0.87–0.89	0.46–0.49
FiLM (OOF)	0.123–0.128	0.213–0.221	0.95–0.98	0.32–0.40

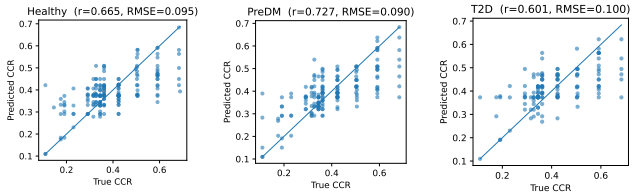

Figure 2: CCR true vs. predicted by HbA1c stratum; panel titles report r and RMSE.

Table 6: Ablations (OOF). Lower is better except r.

Model	RMSE	r	$\text{NRMSE}_{\text{range}}$	$\text{NRMSE}_{\text{std}}$
FINAL (mask-aware meta)	**0.0929**	**0.6910**	**0.1606**	**0.7229**
Ablate FiLM	0.0930	0.6810	0.1608	0.7255
Ablate Stacker	0.0951	0.6910	0.1644	0.7229
Trees-only (ridge+LGB+Cat)	0.1060	0.5922	0.1832	0.8058
Cat-only (avg)	0.1180	0.4839	0.2041	0.8974

mask-aware meta is present, suggesting most benefit comes from (i) diverse calibrated OOFs and (ii) nonnegative, missingness-aware aggregation.

Robustness Across Seeds and Preprocessing

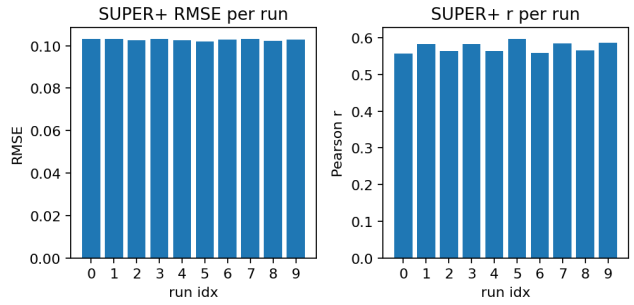

Figure 3: Per-seed/per-preprocess SUPER+ metrics. The final mask-aware meta sits above individual runs.

Across **5 seeds** × **2** PPGR representations (absolute Δmg/dL vs. % change), per-run SUPER+ OOF RMSE spans **0.1043–0.1069** with r=0.576–0.602. Ensembling over seeds *and* representations yields a reliable, additive improvement: **0.0951** RMSE (grouped OOF) and **0.0929** in the final stacked view (Figure 3).

Reliability and Calibration

Table 7: Calibration (OOF). Lower is better; ideal slope = 1.

Variant	ECE	MCE	Slope	Intercept
In-fold (isotonic)	6.7×10^{-17}	1.7×10^{-16}	1.0	≈ 0
Leak-free (nested)	–	–	–	–

Reliability curves align closely with the diagonal under in-fold isotonic (near-zero ECE/MCE; slope ≈ 1). Because in-fold calibration can be optimistic, we also compute a leak-free variant (fit on train-fold OOF, apply to the held-out fold) in the appendix; point metrics remain stable, and calibration remains acceptable for downstream thresholding.

Qualitative Sensor Fusion

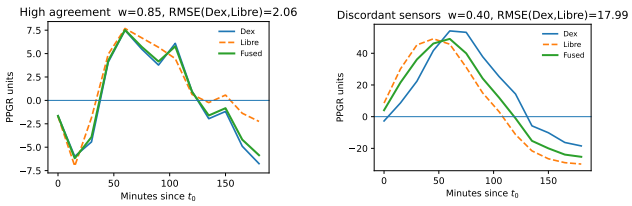

Figure 4: Sensor-aware fusion of Dexcom and Libre PPGR. Left: high agreement; right: discordant sensors. The fusion weight adapts to inter-sensor RMSE; highly discordant pairs are filtered.

Variance/RMSE-aware blending dampens moderate inter-sensor disagreements without smearing peaks, while discordance gating prevents pathological traces from polluting shape features (e.g., IAUC windows, time-to-peak). This improves both robustness and calibration of downstream learners.

GLUCOGRAPHER provides (i) strong global fit (RMSE 0.093–0.095, $r \approx 0.69$), (ii) stable accuracy across HbA1c strata, (iii) robustness to seeds and PPGR representations, and (iv) clear gains over strong single-model and trees-only stacks. Together with released artifacts, these results support reproducible inference of CCR from early PPGR in real-world CGM.

Discussion

GLUCOGRAPHER predicts meal-level CCR on CGMACROS v1.0.0 with **RMSE 0.0929**, $\text{NRMSE}_{\text{range}}$ **0.1608**,

and $r = 0.6910$ (Tables 4, 5). Accuracy is stable across HbA1c strata—Healthy (RMSE 0.0982, r 0.6541), PreDM (0.0923, 0.7305), and T2D (0.0947, 0.6763)—indicating generalization without retuning (Table 4). The global scatter shows a tight monotone relationship (Fig. 1). Subject-level bootstrapping yields narrow 95% CIs (RMSE $[0.0890, 0.1007]$, r $[0.644, 0.735]$), suggesting results are not driven by a few participants.

Why the approach works. Three design choices likely drive performance. *Sensor-aware fusion* reduces brand-specific noise and suppresses discordant segments, stabilizing early-window features. *Dual preprocessing* (absolute Δmg/dL and % change) with *subject-wise PPGR standardization* improves invariance to baseline level and scale, aiding cross-subject learning. Finally, the *mask-aware nonnegative meta-ensemble* robustly aggregates calibrated OOF predictions despite fold-wise missingness, yielding the best overall error/correlation (Table 5).

Ablation insights. The mask-aware meta is the principal source of lift: trees-only degrades to **RMSE 0.1060** and $r = 0.5922$ (Table 6), a $\sim 12\%$ relative RMSE increase vs. the final system (0.0929). Removing FiLM changes performance negligibly (0.0930 RMSE), and dropping the linear stacker modestly worsens error (0.0951), indicating that (i) diverse calibrated OOFs are valuable and (ii) NNLS-based, mask-aware stacking converts that diversity into consistent gains.

Group differences and practical relevance. Healthy participants show slightly higher RMSE than PreDM, consistent with lower PPGR amplitudes and narrower CCR dynamic range; the higher r in PreDM reflects clearer rank structure (Table 4). T2D metrics sit between, aligning with mixed PPGR magnitudes and clearance rates. In aggregate, errors of ~ 0.093 CCR units and $r \approx 0.69$ support use for retrospective coaching and hypothesis generation about meal composition from CGM.

Reliability and calibration. In-fold isotonic calibration yields very low ECE/MCE and slope ≈ 1 (Table 7), as expected when calibrators see in-fold distributions. Because such estimates can be optimistic, we also compute a leak-free variant in the appendix (train-fold OOF $\rightarrow$ held-out fold); point metrics remain stable, with the anticipated rise in ECE.

Interpretability and clinical sense-making. Despite a predictive focus, the system remains legible: PPGR shape features (IAUC windows, t_{peak}, late ratio) connect to absorption/clearance physiology; LightGBM monotonicity encodes directional priors; and HbA1c-stratified reporting shows consistent behavior across glycemic states (Table 4). For deployment, model explanations (e.g., SHAP on tree components) can surface per-meal drivers without exposing raw data.

Implications. Together with artifacts for reproducibility, these results indicate that sensor-aware fusion, dual preprocessing with subject-wise scaling, and mask-aware stacking form a robust recipe for inferring CCR from early PPGR across glycemic strata.

Conclusion

GLUCOGRAPHER converts raw CGM into reliable meal-level feedback by predicting the *carbohydrate calorie ratio* (CCR) from early PPGR. On n=663 meals in the dateshifted CGMACROS v1.0.0, the full system—onset-aware alignment, sensor-aware fusion with discordance gating, dual preprocessing with per-subject PPGR standardization, and a fold-wise *mask-aware* nonnegative meta-ensemble with isotonic calibration achieves **RMSE 0.0929**, $\text{NRMSE}_{\text{range}}$ **0.1608**, $\text{NRMSE}_{\text{std}}$ **0.7229**, and r **0.6910**. Performance is consistent across HbA1c strata, and subject-level bootstrap intervals $\big(\text{RMSE } [0.0890, 0.1007],\ r\ [0.644, 0.735]\big)$ indicate stability. Three choices underpin these results: (i) **sensor-aware fusion** to dampen brand-specific noise and protect PPGR shape; (ii) a **dual-preprocessing** view with **per-subject standardization** to reduce cross-subject scale shifts; and (iii) a **mask-aware nonnegative meta-ensemble** to combine calibrated OOF predictions under fold-wise missingness. Modeling CCR as an interpretable scalar in $[0, 1]$ supports retrospective coaching, prospective planning, and quality control via discordance flags and reliability curves. With transparent components (ridge, monotone trees, CatBoost) and simple calibration, GLUCOGRAPHER offers a practical, reproducible foundation for CGM-driven nutrition feedback.

Looking forward, we plan to (i) perform external validation on independent cohorts and free-living meals; (ii) extend calibration with leak-free or Bayesian approaches targeted to downstream thresholds; (iii) examine interpretability at the meal level (e.g., SHAP and counterfactuals) to surface diet insights; and (iv) evaluate deployment variants that trade off accuracy and complexity (e.g., trees-only vs. full meta). We hope these design principles-sensor-aware fusion, dual preprocessing, and mask-aware meta-learning prove useful to practitioners building robust nutrition models from real-world CGM data.

Limitations

First, we evaluate on a single dataset with standardized meals and dateshifting; although this setting is appropriate for a Bridge paper, external validity remains untested. Second, device availability and logging quality vary across participants; while fusion and discordance filtering help, residual device- or cohort-specific biases are possible. Third, we focus on CCR from early PPGR and macros; unmeasured covariates (e.g., medication timing, stress, sleep) may explain residual variance. Fourth, our reliance on photo-derived or heuristic t_0 can introduce small alignment errors despite onset search. Finally, the FiLM component did not move the needle at current data scale; larger or more heterogeneous cohorts may be needed to fully realize personalization benefits.

References

Alfadli, S. F.; Alotaibi, Y. S.; Aqdi, M. J.; Almozan, L. A.; Alzubaidi, Z. B.; Altemani, H. A.; Almutairi, S. D.; Alabdullah, H. A.; Almehmadi, A. A.; Alanzi, A. L.; et al. 2025. Effectiveness of continuous glucose monitoring systems on glycemic control in adults with type 1 diabetes: A systematic review and meta-analysis. *Metabolism Open*, 100382.

Atkinson, F. S.; Brand-Miller, J. C.; Foster-Powell, K.; Buyken, A. E.; and Goletzke, J. 2021. International tables of glycemic index and glycemic load values 2021: a systematic review. *The American Journal of Clinical Nutrition*, 114(5): 1625–1632.

Battelino, T.; Danne, T.; Bergenstal, R. M.; Amiel, S. A.; Beck, R.; Biester, T.; Bosi, E.; Buckingham, B. A.; Cefalu, W. T.; Close, K. L.; et al. 2019. Clinical targets for continuous glucose monitoring data interpretation: recommendations from the international consensus on time in range. *Diabetes care*, 42(8): 1593–1603.

Bergenstal, R. M.; Ahmann, A. J.; Bailey, T.; Beck, R. W.; Bissen, J.; Buckingham, B.; Deeb, L.; Dolin, R. H.; Garg, S. K.; Goland, R.; et al. 2013. Recommendations for standardizing glucose reporting and analysis to optimize clinical decision making in diabetes: the Ambulatory Glucose Profile (AGP).

Bland, J. M.; and Altman, D. 1986. Statistical methods for assessing agreement between two methods of clinical measurement. *The lancet*, 327(8476): 307–310.

Brouns, F.; Bjorck, I.; Frayn, K.; Gibbs, A.; Lang, V.; Slama, G.; and Wolever, T. 2005. Glycaemic index methodology. *Nutrition research reviews*, 18(1): 145–171.

Brügger, V.; Kowatsch, T.; and Jovanova, M. 2025. Predicting postprandial glucose excursions to personalize dietary interventions for type-2 diabetes management. *Scientific Reports*, 15(1): 25920.

Christ, M.; Braun, N.; Neuffer, J.; and Kempa-Liehr, A. W. 2018. Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package). *Neurocomputing*, 307: 72–77.

Clarke, W. L.; Cox, D.; Gonder-Frederick, L. A.; Carter, W.; and Pohl, S. L. 1987. Evaluating clinical accuracy of systems for self-monitoring of blood glucose. *Diabetes care*, 10(5): 622–628.

Das, A.; Kerr, D.; Glantz, N.; Bevier, W.; Santiago, R.; Gutierrez-Osuna, R.; and Mortazavi, B. J. 2025. CGMacros: a pilot scientific dataset for personalized nutrition and diet monitoring. *Scientific Data*, 12(1): 1557.

Dimitriadis, T.; Gneiting, T.; and Jordan, A. I. 2021. Stable reliability diagrams for probabilistic classifiers. *Proceedings of the National Academy of Sciences*, 118(8): e2016191118.

EFSA Panel on Dietetic Products, N.; and (NDA), A. 2010. Scientific opinion on dietary reference values for carbohydrates and dietary fibre. *EFSA Journal*, 8(3): 1462.

ElSayed, N. A.; Aleppo, G.; Bannuru, R. R.; Bruemmer, D.; Collins, B. S.; Ekhlaspour, L.; Gaglia, J. L.; Hilliard, M. E.; Johnson, E. L.; Khunti, K.; et al. 2024. 2. Diagnosis and classification of diabetes: Standards of care in diabetes—2024. *Diabetes Care*, 47.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *International conference on machine learning*, 1321–1330. PMLR.

Gutierrez-Osuna, R.; Kerr, D.; Mortazavi, B.; and Das, A. 2025. CGMacros: a scientific dataset for personalized nutrition and diet monitoring. *Scientific Data (under review)*.

Hanson, K.; Kipnes, M.; and Tran, H. 2024. Comparison of point accuracy between two widely used continuous glucose monitoring systems. *Journal of Diabetes Science and Technology*, 18(3): 598–607.

Heinemann, L.; Schoemaker, M.; Schmelzeisen-Redecker, G.; Hinzmann, R.; Kassab, A.; Freckmann, G.; Reiterer, F.; and Del Re, L. 2020. Benefits and limitations of MARD as a performance parameter for continuous glucose monitoring in the interstitial space. *Journal of Diabetes Science and Technology*, 14(1): 135–150.

Jain, R. 1985. Ridge regression and its application to medical data. *Computers and biomedical research*, 18(4): 363–368.

Jendrike, N.; Baumstark, A.; Kamecke, U.; Haug, C.; and Freckmann, G. 2017. ISO 15197: 2013 evaluation of a blood glucose monitoring system's measurement accuracy. *Journal of diabetes science and technology*, 11(6): 1275–1276.

Jeong, K.; Moon, S.-J.; Rachim, V. P.; Song, Y.; Cho, Y.-M.; and Park, S.-M. 2025. Enhanced Post-Prandial Glycemic Response Prediction in Type 2 Diabetes with Microbiome Data and Deep Learning. *IEEE Journal of Biomedical and Health Informatics*.

Jiang, X.; Osl, M.; Kim, J.; and Ohno-Machado, L. 2011. Smooth isotonic regression: a new method to calibrate predictive models. *AMIA Summits on Translational Science Proceedings*, 2011: 16.

Kaur, B.; Koh, M.; Ponnalagu, S.; and Henry, C. J. 2020. Postprandial blood glucose response: does the glycaemic index (GI) value matter even in the low GI range? *Nutrition & Diabetes*, 10(1): 15.

Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; and Liu, T.-Y. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

Lawson, C. L.; and Hanson, R. J. 1995. *Solving least squares problems*. SIAM.

Liao, Y.; and Schembre, S. 2018. Acceptability of continuous glucose monitoring in free-living healthy individuals: implications for the use of wearable biosensors in diet and physical activity research. *JMIR mHealth and uHealth*, 6(10): e11181.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Mosquera-Lopez, C.; Wilson, L. M.; El Youssef, J.; Hilts, W.; Leitschuh, J.; Branigan, D.; Gabo, V.; Eom, J. H.; Castle, J. R.; and Jacobs, P. G. 2023. Enabling fully automated insulin delivery through meal detection and size estimation using artificial intelligence. *NPJ digital medicine*, 6(1): 39.

Nelson, W. 1979. Methods for cosinor-rhythmometry. *Chronobiologia*, 6: 305–323.

Pai, R.; Barua, S.; Kim, B. S.; McDonald, M.; Wierzchowska-McNew, R. A.; Pai, A.; Deutz, N. E.; Kerr, D.; and Sabharwal, A. 2024. Estimating breakfast characteristics using continuous glucose monitoring and machine learning in adults with or at risk of type 2 diabetes. *Journal of Diabetes Science and Technology*, 19322968241274800.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12: 2825–2830.

Perez, E.; Strub, F.; De Vries, H.; Dumoulin, V.; and Courville, A. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Popova, P. V.; Isakov, A. O.; Rusanova, A. N.; Sitkin, S. I.; Anopova, A. D.; Vasukova, E. A.; Tkachuk, A. S.; Nemikina, I. S.; Stepanova, E. A.; Eriskovskaya, A. I.; et al. 2025. Personalized prediction of glycemic responses to food in women with diet-treated gestational diabetes: the role of the gut microbiota. *npj Biofilms and Microbiomes*, 11(1): 25.

Presseller, E. K.; Parker, M. N.; Zhang, F.; Manasse, S.; and Juarascio, A. S. 2024. Continuous glucose monitoring as an objective measure of meal consumption in individuals with binge-spectrum eating disorders: A proof-of-concept study. *European eating disorders review*, 32(4): 828–837.

Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A. V.; and Gulin, A. 2018. CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31.

Sergazinov, R.; Chun, E.; Rogovchenko, V.; Fernandes, N.; Kasman, N.; and Gaynanova, I. 2024. GlucoBench: Curated list of continuous glucose monitoring datasets with prediction benchmarks. *arXiv preprint arXiv:2410.05780*.

Shen, Y.; Choi, E.; and Kleinberg, S. 2025. Predicting Postprandial Glycemic Responses With Limited Data in Type 1 and Type 2 Diabetes. *Journal of Diabetes Science and Technology*, 19322968251321508.

Song, H.; Kull, M.; and Flach, P. 2018. Non-parametric calibration of probabilistic regression. *arXiv preprint arXiv:1806.07690*.

Tavasoli, A.; and Shakeri, H. 2025. Online Meal Detection Based on CGM Data Dynamics. *arXiv preprint arXiv:2507.00080*.

Wolever, T. M. 2004a. Effect of blood sampling schedule and method of calculating the area under the curve on validity and precision of glycaemic index values. *British Journal of Nutrition*, 91(2): 295–300.

Wolever, T. M. S. 2004b. Effect of blood sampling schedule and method of calculating the area under the curve on validity and precision of glycaemic index values. *British Journal of Nutrition*, 91(2): 295–300.

Zadrozny, B.; and Elkan, C. 2002. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 694–699.

Zeevi, D.; Korem, T.; Zmora, N.; Israeli, D.; Rothschild, D.; Weinberger, A.; Ben-Yacov, O.; Lador, D.; Avnit-Sagi, T.; Lotan-Pompan, M.; et al. 2015. Personalized nutrition by prediction of glycemic responses. *Cell*, 163(5): 1079–1094.

Zhang, K.; Fu, Y.; Gou, W.; Miao, Z.; Tian, Y.; Liang, Y.; Liang, X.; Shuai, M.; Xiao, C.; Wang, J.; et al. 2025. Quantification of personalized glycemic sensitivity to food and its potential for precision nutrition in a series of n-of-1 trials. *The American Journal of Clinical Nutrition*.

Zheng, M.; Ni, B.; and Kleinberg, S. 2019. Automated meal detection from continuous glucose monitor data through simulation and explanation. *Journal of the American Medical Informatics Association*, 26(12): 1592–1599.

Zhou, Y.; Mai, X.; Deng, H.; Yang, D.; Zheng, M.; Huang, B.; Xu, L.; Weng, J.; Xu, W.; and Yan, J. 2022. Discrepancies in glycemic metrics derived from different continuous glucose monitoring systems in adult patients with type 1 diabetes mellitus. *Journal of Diabetes*, 14(7): 476–484.