

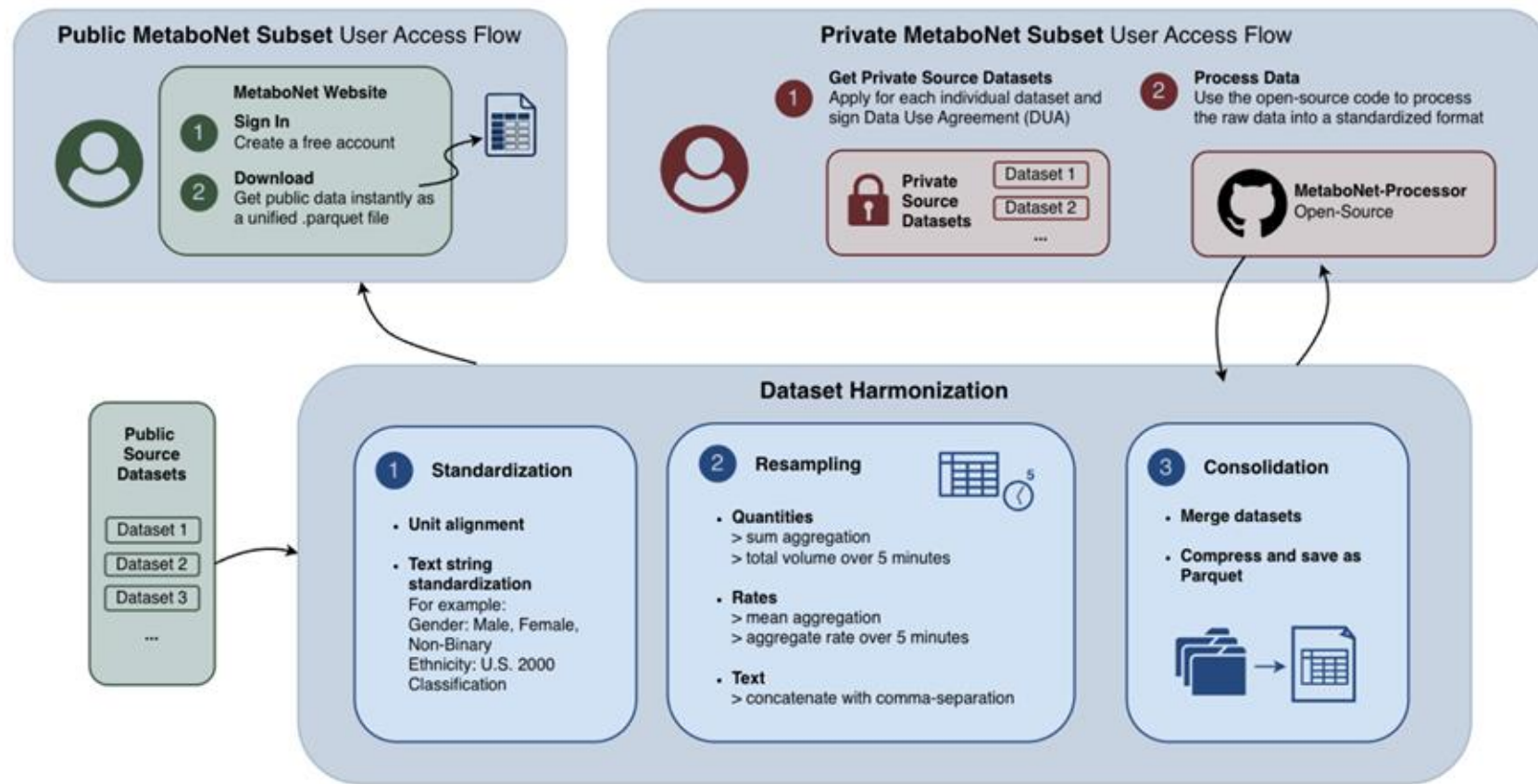
MetaboNet

Journal of Diabetes Science and Technology (JDST)

April 26, 2026

Citation: 3

MetaboNet Architecture



Dataset they combined

MetaboNet 2026 combines 21 existing Type 1 diabetes datasets.

Subset	Datasets	Subjects	Total data	Overlapping CGM + insulin
Public subset	13	1,895	1,464.2 patient-years	867.4 patient-years
DUA-restricted subset	8	1,240	565.9 patient-years	360.8 patient-years
Total	21	3,135	2,030.1 patient-years	1,228.2 patient-years

Public datasets

CTR3, DCLP3, DCLP5, FLAIR, IOBP2, Loop Observational Study, PEDAP, ReplaceBG, AZT1D, BrisT1D, HUPA-UCM, Shanghai T1DM, T1D-UOM

DUA-restricted datasets

DiaTrend, OhioT1DM, OpenAPS Commons, T1DEXI, T1DEXIP, Tidepool HCL150, Tidepool PA50, Tidepool SAP100

Public datasets

Study-level monitoring periods

Dataset	Reported monitoring duration
CTR3	2 weeks
DCLP3	26 weeks
DCLP5	16 weeks
FLAIR	12 weeks
IOBP2	13 weeks
Loop Observational Study	12 months
PEDAP	13 weeks
ReplaceBG	6 months
AZT1D	6–8 weeks
BrisT1D	6 months
HUPA-UCM	At least 14 days
Shanghai T1DM	3–14 days
T1D-UOM	3 months

DUA-restricted datasets

“Varying” means the dataset did not have one fixed monitoring duration for every participant.

Dataset	Reported monitoring duration
DiaTrend	Varying
OhioT1DM	8 weeks
OpenAPS Commons	Varying
T1DEXI	4 weeks
T1DEXIP	10 days
Tidepool HCL150	Varying
Tidepool PA50	Varying
Tidepool SAP100	Varying

Why is the file Parquet instead of CSV?

metabonet_public.parquet

CSV	Parquet
Plain-text format	Binary data format
Stores data row by row	Stores data column by column
Does not reliably preserve data types	Stores an explicit schema and data types
Usually much larger	Uses efficient compression and encoding
Must often read the entire file	Can read only selected columns
Easy to open in a text editor	Requires a data tool such as DuckDB
Suitable for small tables	Designed for large analytical datasets



CGM

DOUBLE

24,523

15.5%



TOTAL

154,842,077 NON-NULL

130,872,839

NULL

23,969,238 UNIQUE

~24,773

DISTRIBUTION

27.5M

13.7M

0

1.80

251.35

500.90

APPROX. STATISTICS

Min

1.8

Max

500.9

5th %

74.62

25th %

105

50th % (median)

135.49

75th %

179.04

95th %

264.78

Mean

148.18

Std Dev

59.72

The Parquet file is one large table with about **154.8 million rows** and **37 columns**. These are **37 standardized columns**, not 37 measurements available for every participant.

1. Core glucose, insulin, and meal variables

Column	Meaning
CGM	Continuous glucose measurement in mg/dL
basal	Basal insulin delivered during that five-minute interval
bolus	Bolus insulin delivered during that interval
insulin	Total insulin; generally basal plus bolus
carbs	Carbohydrate amount recorded during the interval
meal_label	Description or type of meal, when reported

2. Device and treatment variables

Column	Meaning	Examples from the file
cgm_device	The CGM device used to measure glucose	Dexcom G6, Dexcom G4, Guardian 3, FreeStyle Libre 2
insulin_delivery_device	The physical insulin pump or delivery device	OmniPod, t:slim X2, MiniMed 670G, Beta Bionics Gen 4 iLet
insulin_delivery_algorithm	The algorithm or treatment strategy controlling insulin delivery	LoopAlgorithm, Control-IQ, SmartGuard, Basal-IQ, iLet Bionic Pancreas
insulin_delivery_modality	General method of insulin delivery	AID, SAP, MDI
insulin_type_basal	Type or brand of insulin used for basal delivery	Humalog (Lispro), Novolog (Aspart), Fiasp, Long-Acting
insulin_type_bolus	Type or brand of insulin used for bolus delivery	Humalog (Lispro), Novolog (Aspart), Fiasp, Apidra
treatment_group	Clinical trial or treatment group assigned to the participant	CLC, CGM Only, CGM+BGM, BP, AHCL, 670G

3. Demographic and clinical variables

Column	Meaning	Examples from the file
age	Participant's age in years	
age_of_diagnosis	Age when the participant was diagnosed with Type 1 diabetes	
gender	Reported gender	Female, Male, Non-binary
ethnicity	Harmonized race and ethnicity information	White, Asian, Black/African American, white, Hispanic/Latino
height	Participant's height	
weight	Participant's weight	
is_pregnant	Whether pregnancy was reported	TRUE, FALSE
randomization_date	Date when a participant was randomized in an applicable clinical trial	
extension_date	Start date of an applicable study-extension period	

4. Activity, physiological, and environmental measurements

Column	Meaning
steps	Number of steps
calories_burned	Estimated calories burned
workout_duration	Duration of a recorded workout
workout_intensity	Recorded exercise intensity
workout_label	Type or description of exercise
heartrate	Average heart rate during the interval
skin_temp	Skin temperature
air_temp	Environmental temperature
galvanic_skin_response	Electrical skin response associated with sweating or physiological arousal

5. Identification, time, and data-split variables

Column	Meaning	Examples from the file
date	Timestamp associated with the row	
id	Participant identifier within the original dataset	
source_file	Name of the original dataset	Loop, ReplaceBG, IOBP2, DCLP3, PEDAP
is_test	Indicates whether the row belongs to a test partition	TRUE, FALSE
subject_split_across_training_and_testing	Indicates whether records from the same participant occur across both training and testing partitions	TRUE, FALSE

6. Additional notes

Column	Meaning	Examples from the file
misc_notes	Additional information that did not fit into the standardized numerical columns	Control-IQ Sleep Activity, Control-IQ Exercise Activity

Only **88,631,119 rows**, or about **57.2%**, contain CGM, insulin, and carbohydrates simultaneously before preprocessing.

Feature	Missing rows	Missing percentage
CGM	23,969,238	15.48%
Insulin	14,239,388	9.20%
Carbohydrates	51,425,299	33.21%
Basal insulin	18,478,643	11.93%
Bolus insulin	17,262,146	11.15%

Features used

Item	Details
Prediction target	CGM 30 minutes ahead
ML base features	3: CGM, total insulin, carbohydrates
Historical window	2 hours on a 5-minute grid
Approximate lagged inputs	About 24 observations per signal, or approximately 72 values; the exact input dimension was not reported
PBD/Loop inputs	CGM, separate basal and bolus insulin, carbohydrates
PBD history	8 hours
Other dataset features	Activity, demographics, devices, age, gender, etc. were not used

Prediction benchmark results

Model	Inputs/history	RMSE ↓	TG ↑	GM ↑	Clarke Zone A ↑
ZOH	Current CGM	25 mg/dL	0 min	0.70	82.0%
Naïve LE	Three recent CGM values	32 mg/dL	15 min	0.83	79.0%
Naïve AR	Three recent CGM values	23 mg/dL	10 min	0.72	85.8%
PBD/Loop	CGM, basal, bolus, carbohydrates; 8-hour history	26 mg/dL	10 min	0.82	81.7%
Linear NN	CGM, insulin, carbohydrates; 2-hour history	24 mg/dL	5 min	0.73	83.6%
SVR	CGM, insulin, carbohydrates; 2-hour history	23 mg/dL	10 min	0.77	86.0%
MLP	CGM, insulin, carbohydrates; 2-hour history	24 mg/dL	5 min	0.72	83.8%

Thank you

Questions