GLUFORMER: LEARNING GENERALIZABLE REPRESENTATIONS FROM CONTINUOUS GLUCOSE MONITORING DATA

Guy Lutsker[1,2,3], **Gal Sapir**[1,4], **Smadar Shilo**[1,2,5,6], **Jordi Merino**[7,8,9], **Anastasia Godneva**[1,2], **Jerry R. Greenfield**[10,11,12], **Dorit Samocha-Bonet**[10,11], **Raja Dhir**[13], **Francisco Gude**[14,15], **Shie Mannor**[3], **Eli Meirom**[3], **Gal Chechik**[3], **Hagai Rossman**[4,*], **Eran Segal**[1,16,*]

[1]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel
[2]Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel
[3]NVIDIA, Tel Aviv, Israel
[4]Pheno.AI, Tel-Aviv, Israel
[5]Faculty of Medical and Health Sciences, Tel Aviv University, Tel-Aviv, Israel
[6]The Jesse Z and Sara Lea Shafer Institute for Endocrinology and Diabetes, National Center for Childhood Diabetes, Schneider Children's Medical Center of Israel, Petah Tikva, Israel
[7]Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, Copenhagen, Denmark
[8]Diabetes Unit, Endocrine Division, Massachusetts General Hospital, Boston, MA, USA
[9]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA
[10]Clinical Diabetes, Appetite and Metabolism Lab, Garvan Institute of Medical Research, Sydney, Australia
[11]St Vincent's Clinical School, University of NSW, Sydney, Australia
[12]Department of Endocrinology and Diabetes, St Vincent's Hospital, Sydney, Australia
[13]Swiss Institute of Allergy and Asthma Research (SIAF), University of Zurich, Davos, Switzerland
[14]Department of Medicine, University of Santiago de Compostela, Spain
[15]Concepción Arenal Primary Care Center, Santiago de Compostela, Spain
[16]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
[*]Corresponding authors

ABSTRACT

Continuous glucose monitoring (CGM) enables near-continuous measurement of glucose trends, offering detailed insight into metabolic health. However, existing CGM-based metrics (e.g., time in range, glucose management indicator) only partially capture the complexities of glycemic variability. In this work, we present *GluFormer*, a generative foundation model employing self-supervised representation learning on over 10 million CGM measurements from 10,812 participants without a known diabetes diagnosis. By predicting future tokens in an autoregressive fashion, GluFormer learns latent representations that generalize across 19 additional cohorts ($n = 6{,}044$) with differing devices, ethnicities, and clinical contexts (from prediabetes and gestational diabetes to type 1/2 diabetes). GluFormer outperforms standard CGM metrics in forecasting clinical measures (e.g., A1c, visceral adipose tissue, and liver function) and in risk stratification for longer-term outcomes such as incidence of diabetes and cardiovascular mortality. Beyond single-number CGM summaries, the model generates realistic glucose curves that align with real-world data, and its performance further improves when including discrete dietary tokens in a multimodal framework. These findings suggest that large-scale self-supervised learning on continuous physiological signals can improve our ability to identify and manage metabolic risks, as well as simulate personalized glycemic trajectories.

1 INTRODUCTION

With the growing accessibility of continuous glucose monitoring (CGM) technologies in both clinical and non-clinical settings, unprecedented volumes of densely sampled glucose data have become available. These rich time-series enable fine-grained representation learning of an individual's metabolic state (Battelino et al., 2019; Shilo et al., 2023), yet common CGM metrics such as time in range or glucose management indicator (GMI) often overlook subtle temporal patterns that may signal distinct metabolic risks (Broll & Etc., 2021; Bergenstal et al., 2018).

Parallel to these developments in CGM, the realm of self-supervised learning (SSL) has produced powerful "foundation models" that discover general-purpose representations from large-scale unlabeled data (Zhou et al., 2023; Saab et al., 2024; Lutsker et al., 2024). These approaches, exemplified by transformer architectures, have transformed fields like natural language processing (Devlin et al., 2019), and show promise in clinical and biomedical domains.

Here, we propose **GluFormer**, a transformer-based SSL model for CGM data. GluFormer is trained on over 10 million measurements from 10,812 non-diabetic adults, learning generative and latent representations of glycemic time-series through next-token (autoregressive) prediction. We show that GluFormer's learned representations:

1. Generalize effectively to 19 additional datasets from multiple countries, devices, and clinical populations (e.g., gestational diabetes, type 1/2 diabetes).
2. Predict both near-term (0–4 years) and long-term (up to 12 years) clinical endpoints (e.g., diabetes onset, cardiovascular mortality) more accurately than conventional CGM summaries such as GMI.
3. Generate synthetic glucose signals whose distributions of clinically relevant CGM metrics strongly match real patient data, suggesting potential applications in data augmentation or simulation.
4. Achieve further improvements in generative fidelity when extended to a *multimodal* form, adding discrete tokens of dietary intake.

We argue that large-scale representation learning approaches can exploit the rich structure of continuous biosignals, offering a more holistic view of metabolic health than the single-point glucose tests currently used in routine care. GluFormer highlights the broader possibility of foundation models for physiological data, with implications for risk stratification and personalized interventions.

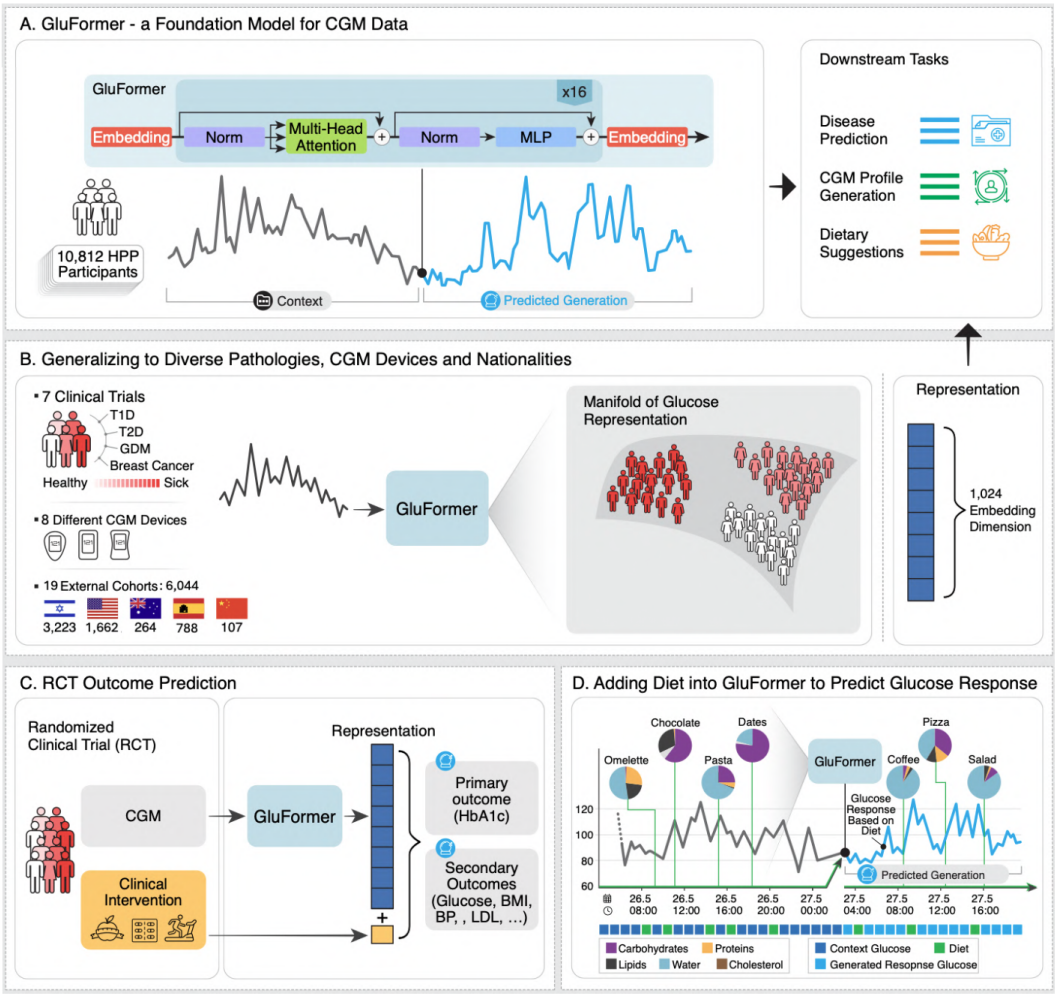

Figure 1: **Conceptual overview of GluFormer's training and capabilities.** (A) We first convert continuous glucose time-series into discrete tokens and train a causal transformer to predict subsequent tokens (autoregressive language modeling). (B) Once trained, GluFormer yields flexible *representations* for each individual's CGM profile, which can be used for downstream tasks like forecasting clinical measures or identifying risk. (C) GluFormer representations can also be used to predict outcomes of RCTs. (D) By adding discrete "diet tokens," a multimodal extension accurately simulates personalized glycemic responses to foods.

2 Related Work

Representation Learning in Health.
Self-supervised learning has rapidly expanded in biomedical fields, from analyzing large-scale pathology images (Zhou et al., 2023) and wearables data (Yuan et al., 2023), to structured clinical records. By focusing on tasks like next-token prediction or masked data reconstruction, these methods discover domain-relevant features without relying on labor-intensive labeling.
Limitations of Existing CGM Metrics: While CGM data can reveal glycemic patterns, clinicians often rely on summary statistics (e.g., GMI, time in range), which may not fully capture inter-day variability or early markers of metabolic deterioration (Broll & Etc., 2021). Traditional CGM analyses also offer limited capacity for generative modeling or for transferring to new populations.
Transformers for Time Series: Transformer architectures excel at capturing long-range dependencies (Vaswani et al., 2017), making them well-suited for multi-day CGM. Several prior works show success applying transformers to continuous signals by discretizing them into tokens (like words in NLP) and then learning next-token prediction (van den Oord et al., 2016a; Rabanser et al., 2023).

3 METHODS

3.1 DATA AND PARTICIPANTS

HPP Training Dataset. We constructed our core training set from a large cohort of 10,812 adults who wore Abbott Freestyle Libre CGM devices for approximately two weeks, capturing glucose measurements every 15-minutes. The vast majority of individuals had no diabetes diagnosis at enrollment. Following standard removal of first day measurements (due to device warm-up adjustments) and linear interpolation for rare missing data, the final set exceeded 10 million glucose values (Shilo et al., 2021).

3.2 TOKENIZATION OF CGM VALUES

We adopt a discretization strategy that maps continuous glucose readings into integer tokens, drawing on research indicating that binning real-valued signals can improve forecasting by regularizing inputs, enhancing robustness, and aligning with transformer vocabulary mechanisms (van den Oord et al., 2016a; Rabanser et al., 2023; Ansari et al., 2024). Specifically, glucose values from 40–500 mg/dL are uniformly split into 460 possible bins, yielding a vocabulary of size 460. Each CGM sample can thus contain up to 1,200 tokens, representing about 10–12 days of measurements, and shorter sequences are padded with a special `<MASK>` token. During training, we apply a causal mask to ensure that each token only attends to preceding tokens, enabling autoregressive next-token prediction. This token-based representation trades away some continuous detail, but can often yield better generalization. Indeed, beyond CGM, discretization has benefited generative models in domains such as audio (van den Oord et al., 2016a) and images (van den Oord et al., 2016b), where binning or quantization helps stabilize training and simplifies modeling.

3.3 MODEL ARCHITECTURE AND TRAINING

Transformer Configuration: GluFormer is a 16-layer causal transformer with hidden size 1024, feedforward dimension 2048, and 16 attention heads (Vaswani et al., 2017). We process sequences of up to 1,200 glucose tokens, predicting the next token at each time step.
Self-Supervised Objective: We train GluFormer autoregressively via cross-entropy loss over the next-token distribution. Formally, we minimize: $\mathcal{L} = -\sum_{t=1}^{T} \log P(x_t \mid x_{<t})$. where x_t is the token at time t. This objective forces the model to learn a representation that captures glycemic patterns over many days.

3.4 GENERATING EMBEDDINGS

In our approach, each CGM sample is represented as a sequence of 1,200 glucose tokens (after tokenizing and possibly padding shorter sequences). We pass this entire sequence through the pretrained transformer, producing a 1,024-dimensional hidden state for each token at the final layer. Because some CGM traces may have fewer than 1,200 valid measurements, we first remove hidden states corresponding to any padded `<MASK>` tokens. We then apply *max-pooling over time*, which means that for each of the 1,024 dimensions, we select the maximum value across all remaining token positions. Formally, if $\{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_T\}$ denotes the set of final-layer hidden states for a sequence of length $T \leq 1200$, then our single-sample embedding is: $\mathbf{e} = \max_{1 \leq t \leq T} \mathbf{h}_t$, where the $\max$ is taken element-wise across the T vectors.

This final embedding vector $\mathbf{e} \in \mathbb{R}^{1024}$ serves as a learned, self-supervised summary of the entire CGM sequence. It can then be used as input to downstream models for classification or regression tasks, for instance to predict a clinical measure (e.g., A1C) or the likelihood of disease onset. Because the transformer learns to capture temporal dependencies among all tokens, pooling over the hidden states emphasizes the most salient features in the sequence and creates a single, fixed-size representation for each CGM sample.

3.5 MULTIMODAL EXTENSION WITH DIETARY TOKENS

Although glucose tokens alone capture broad glycemic trends, real-world trajectories are often strongly influenced by dietary intake. To better reflect the impact of meals, we extend GluFormer

into a multimodal model that jointly processes discretized glucose values (glucose tokens) and discrete bins of macronutrients (diet tokens). The diet tokens represent carbohydrates, protein, and other nutrients consumed at each meal or snack time, interleaved within the same sequence as glucose tokens. **Diet Tokenization:** We obtain time-stamped food logs (self-reported) and convert each macronutrient amount (e.g., grams of carbs) into a discrete bin. For example, we might have 15 bins for carbohydrate amounts, 10 bins for protein amounts, and so forth. We then insert these diet tokens at the appropriate time steps in the sequence, effectively interleaving them with glucose tokens in chronological order. **Unified Transformer Input:** Both types of tokens share a base embedding dimension (e.g., 1024), but to help the model distinguish modalities, we add a learned "modality embedding" (glucose vs. dietary). We also incorporate positional (time-of-day, day-of-week) encodings to handle the real time differences between consecutive measurements or meals. **Training and Generation:** The model *still focuses on predicting the next glucose token* (cross-entropy loss computed only over glucose bins), while diet tokens act as contextual input. In generation, we provide recent CGM data plus dietary tokens as "observed" context. The model autoregressively predicts future glucose values, factoring in mealtime macronutrient inputs. This design often yields more accurate capture of postprandial peaks and inter-meal patterns.

As shown in Section 3.11, the multimodal GluFormer significantly improves day-ahead predictions (lower MAE, higher correlation with real CGM) when dietary data is available. This framework can also be extended to incorporate other behavioral or sensor inputs (e.g., physical activity, sleep) in future work.

3.6 External Datasets and Generalization

While GluFormer is trained on a large, mostly non-diabetic cohort, we sought to test whether its embeddings and generative capabilities transfer to more diverse populations. Accordingly, we assembled 19 additional datasets (total $n = 6{,}044$) that include participants from multiple continents, various CGM devices, and different glycemic conditions, as summarized in Table 1 (see Supplementary). Each external cohort was preprocessed by applying the same binning scheme for glucose tokens (i.e., the 40–500 mg/dL range split into 460 discrete bins) so that the new data could be directly ingested by our pretrained GluFormer. We then generated 1,024-d embeddings for each participant's CGM sequence and used these embeddings to predict various clinical outcomes relevant to each dataset's phenotype (e.g., A1c, or future changes in glucose tolerance). This uniform pipeline allows us to systematically evaluate how well GluFormer representations extend across device types, populations, and disease states.

3.7 Learned Representations Stratify Glycemic Risk

Figure 1 provides an overview of the training pipeline. We first asked whether the GluFormer embeddings capture clinically relevant differences in CGM data. To visualize the internal representation space, we used UMAP projections (McInnes et al., 2018) (Figure 2)

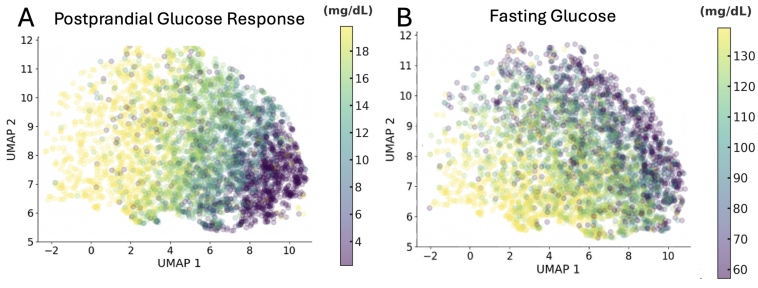

Figure 2: UMAP visualizations of GluFormer embeddings and coloring by independently collected glycemic health parameters.

Notably, these projections show smooth gradients aligned with *important clinical markers* such as: **Fasting glucose:** The typical glucose level after an overnight fast, which can signal underlying insulin resistance. **Postprandial Glucose Response:** The glucose rise following meals, often indicative of beta-cell function and insulin response. **Day-to-day glycemic fluctuations:** The degree to which an individual's glucose levels vary from one day to the next, reflecting overall metabolic

resilience or instability.
These observations suggest that GluFormer's representation space differentiates between diverse, clinically significant glycemic profiles in an unsupervised manner.

A key question is whether these embeddings can *stratify* individuals by risk for adverse glycemic outcomes. In individuals whose baseline lab tests indicated prediabetes (HbA1c 5.7–6.4%), we used a simple linear model on top of GluFormer embeddings to produce a continuous "predicted A1c" score from multi-day CGM. Grouping participants by quartiles of this predicted A1c revealed a striking pattern: over a two-year follow-up, those in the top quartile showed substantially greater increases in actual blood-measured HbA1c compared to the bottom quartile. By contrast, grouping the same individuals by their baseline lab A1c measurement alone failed to separate those who would progress from those who would not. This indicates that multi-day CGM data (summarized in GluFormer embeddings) offers a richer signal for identifying who is likely to deteriorate metabolically than does a single blood HbA1c measurement (Keshet et al., 2023).

We further evaluated this observation in a larger, longer follow-up setting: a 12-year cohort study of 580 adults (Gude et al., 2017) that tracked both new-onset diabetes and cardiovascular mortality. Figure 3 summarizes our findings. Again, we took CGM for each person and derived an "predicted A1c" estimate from GluFormer representations. We then stratified participants into quartiles by this predicted A1c and compared long-term outcomes. The top quartile (highest predicted A1c) accounted for $\sim 65\%$ of all new diabetes cases and $\sim 69\%$ of cardiovascular deaths during the 12-year period. By contrast, stratifying by baseline *lab* A1c showed almost no separation in outcomes.

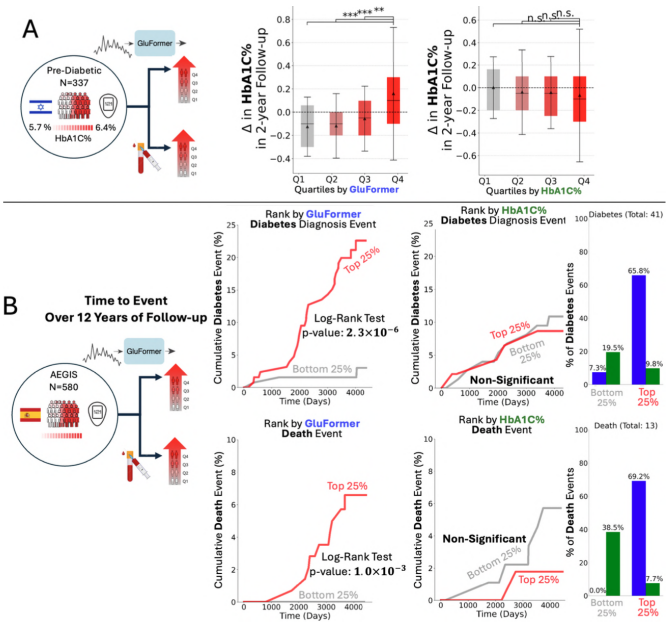

Figure 3: **GluFormer-derived A1c Outperforms Measured (blood) A1c for Future Risk.** (A) Among 337 prediabetic individuals (baseline A1c 5.7–6.4%) in the HPP cohort, stratifying by GluFormer-predicted A1c quartiles reveals significantly different 2-year A1c changes (p¡0.001): the top quartile (red) increases by +0.18 whereas the bottom quartile (gray) decreases by –0.13. By contrast, quartiles based on measured A1c alone show no significant difference. (B) Kaplan–Meier analysis in 580 adults from AEGIS over 12 years shows that the top quartile of GluFormer-predicted A1c has markedly higher diabetes incidence and cardiovascular mortality than lower quartiles, whereas measured A1c quartiles do not significantly separate outcomes. The bar plots indicate that 65.8% of diabetes cases and 69.2% of cardiovascular deaths occur in the GluFormer top quartile versus minimal events in the bottom quartile.

3.8 Predicting Clinical Measures and Risk

We first assessed how well GluFormer embeddings forecasted clinical metrics either measured at the time of CGM or several years later. Simple ridge or logistic regressions on these embeddings outperformed standard CGM summaries (GMI, time in range) across a range of outcomes:

- **A1c and Glucose Levels:** Even in non-diabetic and prediabetic ranges, GluFormer better predicted both current and future blood glucose measures, capturing nuanced glycemic fluctuations.
- **Visceral Adipose Tissue, Liver Enzymes, Renal Function, etc.:** Embeddings correlated more strongly with these metabolic markers than conventional summaries, suggesting that the model discovers clinically relevant signals beyond short-term glycemia.
- **Long-Horizon Events:** In a 12-year follow-up of 580 adults, the top quartile of GluFormer-derived A1c predictions captured $\sim 66\%$ of diabetes diagnoses and $\sim 69\%$ of cardiovascular deaths, whereas standard A1c or GMI quartiles showed weaker stratification.

We also performed decision-curve and survival analyses that consistently showed superior performance of GluFormer embeddings vs. standard CGM metrics.

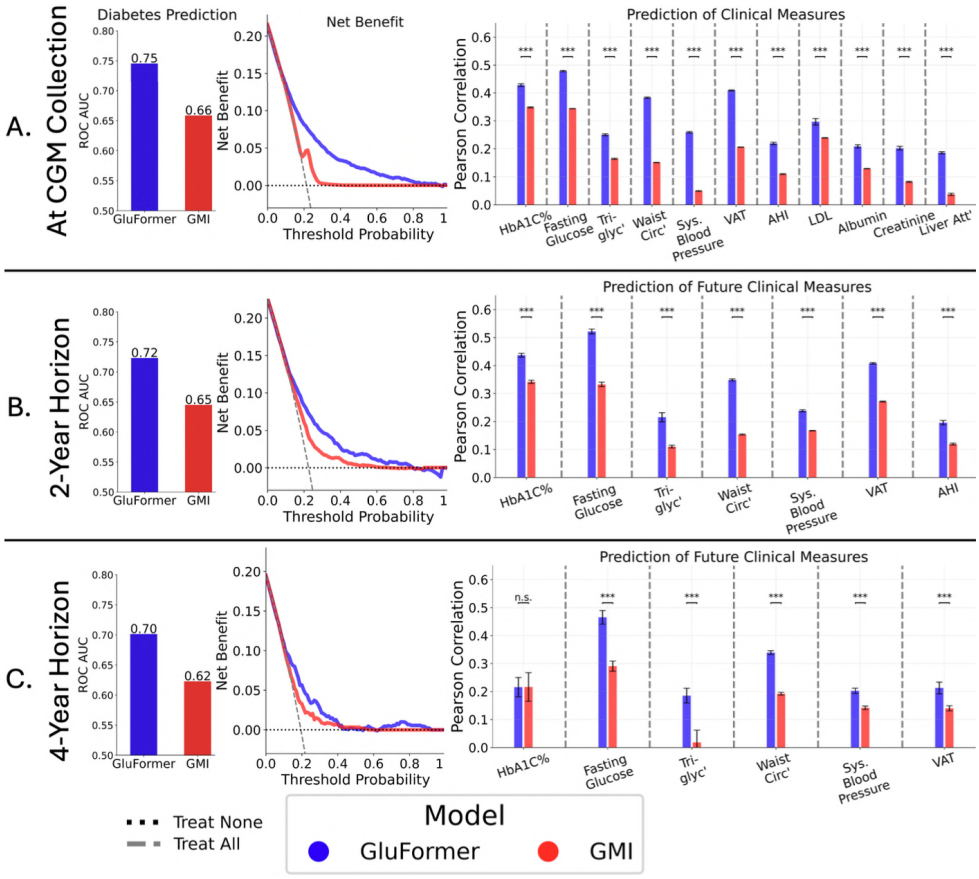

Figure 4: **Improved clinical predictions via GluFormer embeddings.** (A) Example: Predicting various metabolic measures (A1c, fasting glucose, visceral adipose tissue, systolic blood pressure) at the time of CGM. GluFormer-based regressions (blue) yield higher correlation with ground truth than GMI-based (red). (B) Similarly, at a two-year horizon, GluFormer continues to provide stronger performance, reflecting its ability to embed more stable and predictive signals from multi-day CGM data. Error bars represent standard deviation across multiple runs.

3.9 Performance on External Cohorts

To test out-of-distribution generalizability, we evaluated GluFormer embeddings on 19 external cohorts (Section 3.6). Figure 6 highlights that for numerous clinical endpoints (e.g., future changes in A1c, organ function biomarkers, or risk classifications), simple linear models on GluFormer embeddings generally matched or exceeded the performance of conventional CGM-based measures.

These cohorts varied widely in location, device, sample size, and disease phenotype, suggesting that the token-based representation learned from healthy populations can still robustly transfer to other contexts.

3.10 Generative Modeling of CGM

Because GluFormer is trained autoregressively, it can also sample plausible CGM trajectories. We tested its ability to generate day-scale glucose curves given some initial context (e.g., the prior day's data). Figure 7 shows that synthetic curves preserve individualized metrics such as mean glucose, time in range, and glucose variability, with correlation of $r = 0.80$–0.98 between real vs. generated composite measures.

3.11 Multimodal Extension Using Dietary Tokens

To better model diet-induced glycemic excursions, we extended GluFormer to a multimodal variant that ingests both glucose tokens and binned macronutrient tokens (carbohydrates, protein, etc.). The model's next-token objective still focuses on predicting glucose tokens, but it can attend to dietary tokens as additional context. Including diet tokens improved day-ahead CGM generation accuracy, raising correlation from ~ 0.22 to ~ 0.50. Figure 5 illustrates how the multimodal approach captures postprandial glucose spikes more accuratly.

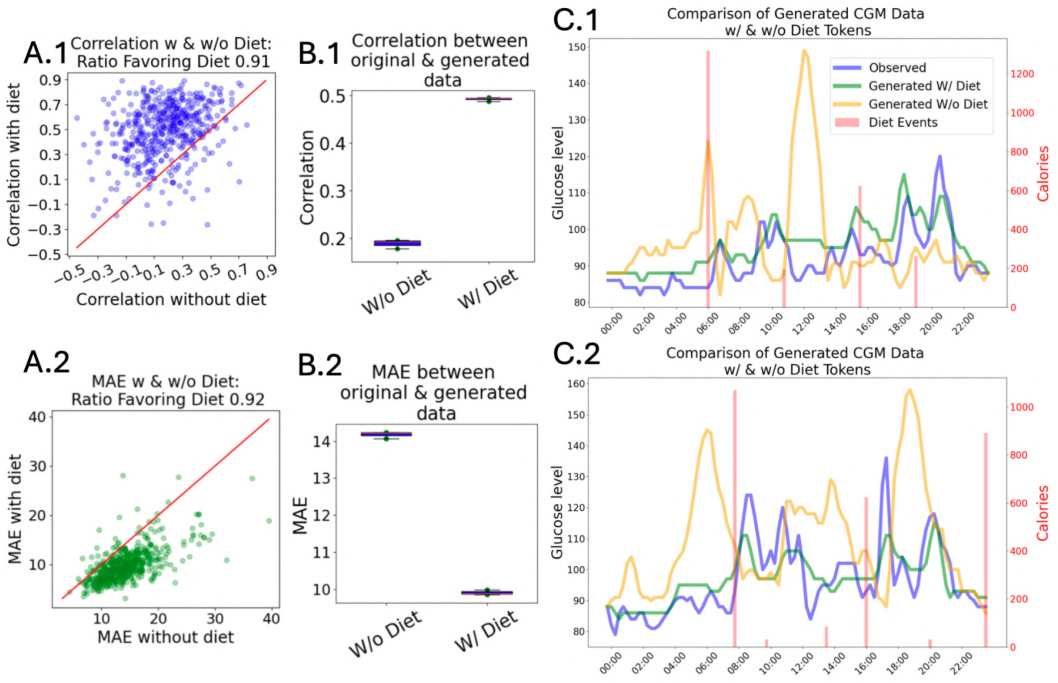

Figure 5: **Impact of Dietary Data on GluFormer Model Performance.** A. Comparison of Pearson correlation A.1 and mean absolute error (MAE) A.2 between the original and generated CGM data, with and without the inclusion of dietary data. Scatter plots show the improvements in correlation and MAE when dietary data is included, indicated by the majority of points falling above the diagonal line on correlation, and below on MAE metrics. B. Box plots summarizing the overall performance, showing the average correlation B.1 and MAE B.2 across all test participants, for 5 different random seeds (used for generation) with lower MAE and higher correlation for models including dietary data. C. Time series plots demonstrating glucose level predictions for two example participants C.1 and C.2. The observed CGM data (blue line) is compared to data generated with dietary tokens (green line) and without dietary tokens (orange line). Red bars indicate times of dietary events, highlighting the model's improved performance in capturing glucose spikes when dietary information is included.

4 Discussion and Conclusion

We introduced GluFormer, a transformer-based foundation model for CGM data that learns powerful latent representations of glycemic patterns via next-token prediction. Trained on a large sample of predominantly healthy adults, GluFormer successfully generalizes to external datasets including type 1/2 diabetes, gestational diabetes, and prediabetes populations. Its embeddings exceed standard CGM summaries in predicting diverse clinical outcomes—from near-term metabolic measures to 12-year follow-up events such as diabetes onset and cardiovascular mortality.

The ability to synthesize realistic CGM signals opens up possibilities for data simulation, while the multimodal variant incorporating dietary tokens offers more accurate modeling of postprandial changes. By leveraging principles from natural language processing and discretizing continuous time-series data, GluFormer underscores the promise of large-scale self-supervised learning in metabolic health. Future directions include incorporating additional sensors (activity, sleep), expanding to other populations, and exploring interpretability tools to further integrate CGM-based modeling into routine clinical care.

Limitations. Our training data, though large, focuses on mostly non-diabetic adults. Some device types and advanced T1DM populations are underrepresented. In addition, self-reported dietary logs may contain inaccuracies. Lastly, we have not performed a formal prospective intervention trial to see whether GluFormer-based predictions meaningfully improve clinical outcomes when integrated into treatment decisions.

Despite these caveats, our findings suggest that large-scale, token-based transformers can learn robust models of glycemic health that generalize across device types and diverse clinical contexts—a first step toward comprehensive foundation models of continuous biosignals.

Meaningfulness Statement

"Meaningful representations of life" should capture the hidden but predictive aspects of physiological signals that matter for human health, while generalizing to varied populations. GluFormer's foundation model approach to CGM data exemplifies this: it uncovers subtle glycemic patterns that foreshadow significant outcomes (e.g., diabetes onset, cardiovascular events). By aligning with real-world eating behaviors and inter-day variability, it yields individual-level embeddings that extend beyond basic summary metrics. We envision that similar token-based, self-supervised strategies can be applied to other continuous biosignals, ultimately providing cohesive, personalized windows into complex metabolic processes and informing interventions tailored to each patient's unique biology.

References

Aman Fazlur Rehman Ansari et al. Chronos: Learning the language of time series. arXiv:2403.07815, 2024.

Tadej Battelino et al. Clinical targets for continuous glucose monitoring data interpretation. *Diabetes Care*, 42(8):1593–1603, 2019.

Richard M Bergenstal et al. Glucose management indicator (gmi): a new term for estimating a1c from continuous glucose monitoring. *Diabetes Care*, 41(11):2275–2280, 2018.

S Broll and Etc. Interpreting blood glucose data with r package iglu. *PLOS ONE*, 16(3):e0248560, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

Francisco Gude et al. Glycemic variability and its association with demographics and lifestyles in a general adult population. *Journal of Diabetes Science and Technology*, 11(4):780–790, 2017.

A Keshet et al. Cgmap: Characterizing continuous glucose monitor data in thousands of non-diabetic individuals. *Cell Metabolism*, 35:758–769, 2023.

Guy Lutsker, Hagai Rossman, Godneva, and Eran Segal. COMPRER: A Multimodal Multi-Objective Pretraining Framework for Enhanced Medical Image Representation. *arXiv*, 2024.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426, 2018.

Gerhard Rabanser, Nick Heard, and Manuel Rodriguez. Ts or not ts: That is the question – testing the utility of time series models for real-world forecasting. arXiv:2305.10111, 2023.

Khaled Saab et al. Capabilities of gemini models in medicine. arXiv:2404.18416, 2024.

S Shilo et al. Continuous glucose monitoring and intrapersonal variability in fasting glucose. *Nature Medicine*, 30:1424–1431, 2023.

Smadar Shilo et al. 10k: a large-scale prospective longitudinal study in israel. *European Journal of Epidemiology*, 36(11):1187–1194, 2021.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A generative model for raw audio. In *ISCA Speech Synthesis Workshop*, 2016a.

Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International Conference on Machine Learning (ICML)*, pp. 1747–1756, 2016b.

Ashish Vaswani et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

Han Yuan et al. Self-supervised learning for human activity recognition using 700,000 person-days of wearable data. *npj Digital Medicine*, 6(1):91, 2023.

Yuhang Zhou et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7962):156–163, 2023.

A Supplementary Figures

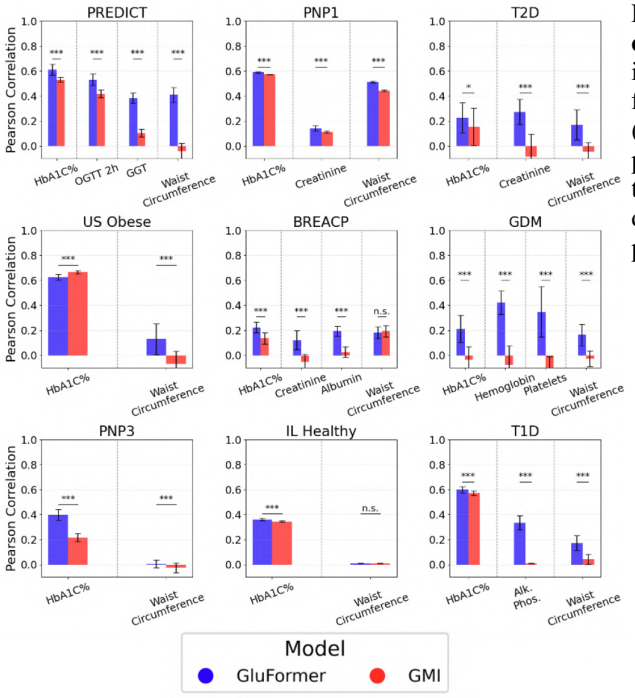

Figure 6: **Generalizing across diverse external cohorts.** We grouped participants by cohort (horizontal axis) and measured Pearson correlations for selected clinical outcomes. Bars compare GMI (red) to GluFormer embeddings (blue). Even for patient populations with type 1/2 diabetes, gestational diabetes, or obesity, and even when using different CGM devices, GluFormer typically outperforms or ties GMI.

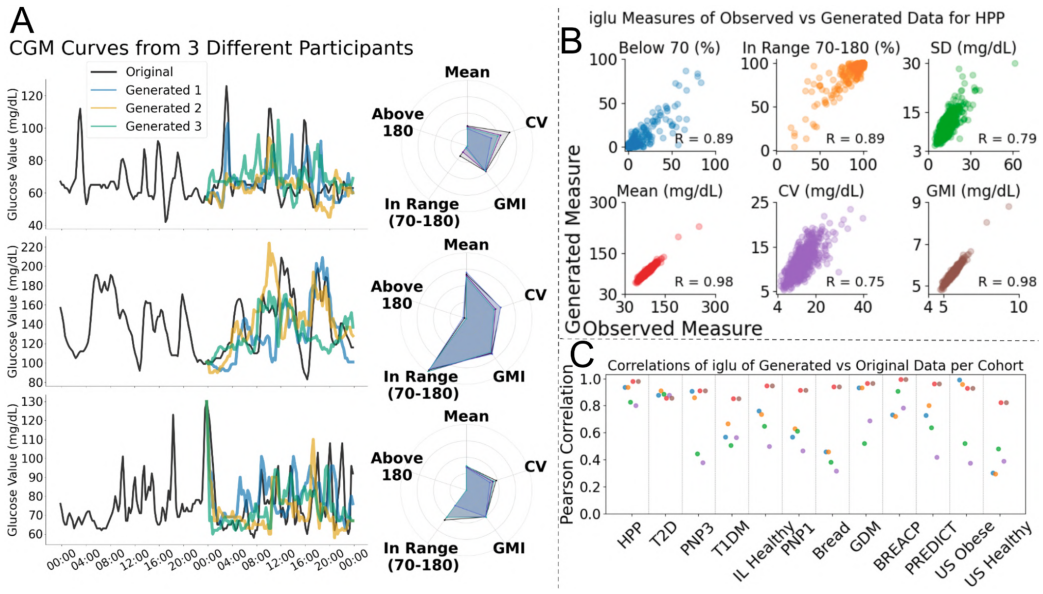

Figure 7: **GluFormer generates realistic day-to-day CGM signals.** (A) Representative examples from three individuals comparing true CGM profiles (black) vs. three generated trajectories (colored). Although exact alignment may vary, generated curves preserve overall patterns. (B) Radar plots show strong agreement between real and generated CGM metrics, e.g. for time in range, hyperglycemia, mean glucose, etc. (C) Across an entire validation set, we find high correlation ($r > 0.8$) between real vs. generated composite scores.

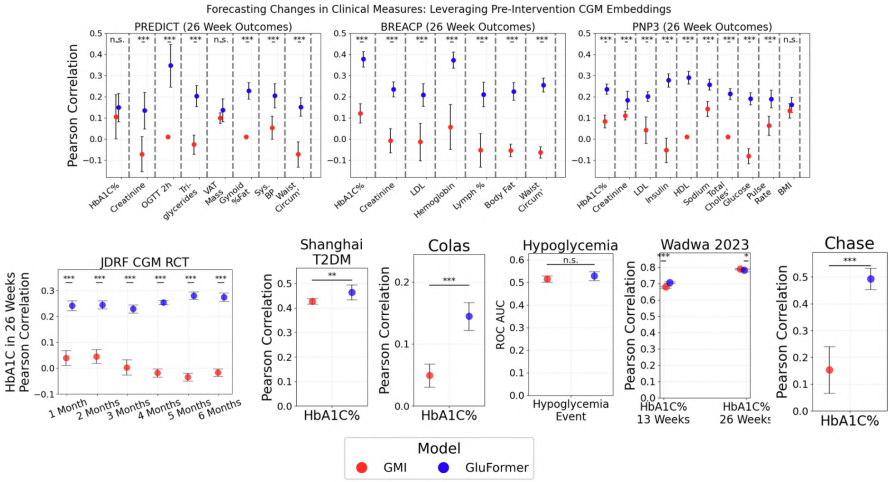

Figure 8: **Longitudinal RCT outcome prediction using GluFormer embeddings.** (Top) Pearson correlations for various post-intervention measures (A1c, waist circumference, body fat, etc.) in three clinical trials when predicting with GluFormer representations (blue) vs. GMI (red). Including a binary variable for the intervention arm, GluFormer's CGM embeddings consistently yield higher correlation. (Bottom) Additional validation on multiple open-access trials with distinct study populations, again showing that the learned embeddings can forecast future A1c or other endpoints more accurately than standard CGM metrics.

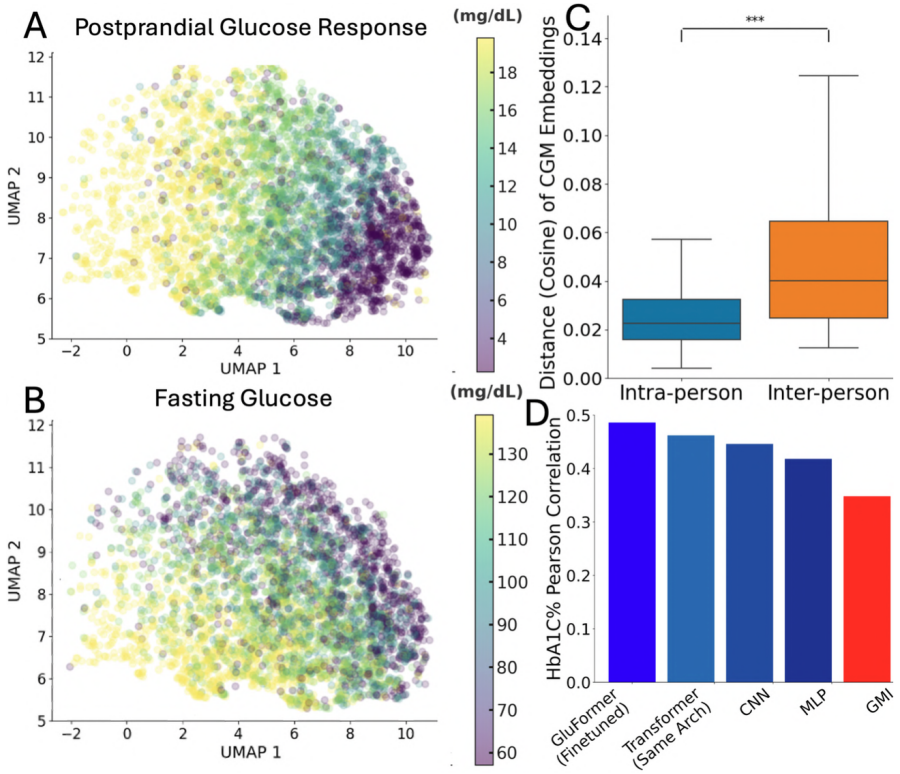

Figure 9: **Evaluation of GluFormer's Representations.**A. B. UMAP visualization of CGM representation from our model. Here we show how the representations relate to 2 clinical measurements not seen by the model during training. A. Plot shows a UMAP colored by Postprandial Glucose Response (PPGR), showing that UMAP dimension 1 captures the diversity in PPGR. Low PPGR values appear on the right, progressing to high PPGR on the left. B. Plot, colored by fasting glucose levels obtained from blood tests, shows that UMAP dimension 2 captures the range of fasting glucose levels, with lower levels on the right and higher levels on the left. These visualizations provide insights into how different clinical measures, crucial in endocrinology, could be associated with the learned CGM representations. C. A comparison of intra-participant and inter-participant cosine distances of CGM representations of the HPP. The "Intra Distances" (blue box plot) shows the distribution of cosine distances between representations of the same participant across different days (with no overlap), reflecting day-to-day variability in the individual's data. The "Inter Distances" (orange box plot) shows the distribution of distances between representations from different participants, showing variation across individuals. There is an overlap in inter, intra-participant embedding distances meaning that there are some instances of participants who are very similar and some who are highly variable. The significant difference between the two, indicated by three asterisks, was tested using the Mann-Whitney test. D. A plot of the effectiveness of different models in predicting HbA1C

B Implementation Details

For reproducibility, we provide code and hyperparameters in our GitHub repository. Briefly, the standard AdamW optimizer is used, with a 3×10^{-5} base learning rate. We train on 8 GPUs (batch size 256) for up to 100 epochs. For generation tasks, we use beam search or temperature sampling. Additional ablation experiments on tokenization choices, discrete vocabulary size, and pooling methods are provided in the repository.

C SUMMARY OF THE 19 EXTERNAL DATASET

Table 1: Summary of the 19 external datasets.

Characteristic	**Description**
Geography	Israel, Australia, the United States, China, and Europe
CGM Devices	Dexcom, Medtronic iPro2, Abbott Libre (including different generations), etc.
Glycemic Conditions	Type 1/2 diabetes, gestational diabetes, prediabetes, obesity, and non-diabetic controls
Study Designs	Randomized clinical trials, observational cohorts, and prospective follow-up studies (up to 12 years)