Proceedings of Machine Learning Research 225:607–622, 2023 Machine Learning for Health (ML4H) 2023

Interpretable Mechanistic Representations for Meal-level Glycemic Control in the Wild

Ke Alexander Wang ALXWANG@CS.STANFORD.EDU
Stanford University

Emily B. Fox EBFOX@STANFORD.EDU
Stanford University & CZ Biohub

Abstract

Diabetes encompasses a complex landscape of glycemic control that varies widely among individuals. However, current methods do not faithfully capture this variability at the meal level. On the one hand, expert-crafted features lack the flexibility of data-driven methods; on the other hand, learned representations tend to be uninterpretable which hampers clinical adoption. In this paper, we propose a hybrid variational autoencoder to learn interpretable representations of CGM and meal data. Our method grounds the latent space to the inputs of a mechanistic differential equation, producing embeddings that reflect physiological quantities, such as insulin sensitivity, glucose effectiveness, and basal glucose levels. Moreover, we introduce a novel method to infer the glucose appearance rate, making the mechanistic model robust to unreliable meal logs. On a dataset of CGM and self-reported meals from individuals with type-2 diabetes and pre-diabetes, our *unsupervised* representation discovers a separation between individuals proportional to their disease severity. Our embeddings produce clusters that are up to 4x better than naive, expert, black-box, and pure mechanistic features. Our method provides a nuanced, yet interpretable, embedding space to compare glycemic control within and across individuals, directly learnable from in-the-wild data.

Keywords: interpretable representations, mechanistic differential equations, CGM, diabetes, T2D, glycemic variability, disease subtyping, disease progression, variational autoencoder, hybrid modeling

1. Introduction

Diabetes manifests through a multitude of pathways, resulting in a highly heterogeneous population of individuals grappling with the condition (Ha and Sherman, 2020). Traditional metrics for characterizing diabetes, such as HbA1C levels and time-in-range, offer only a limited perspective (Brown et al., 2019). They fall short in capturing the nuanced variations in glycemic control at the meal-level, a critical aspect of daily life that is often targeted as the intervention lever to control disease progression. In contrast, continuous glucose monitors (CGMs), which sample every few minutes, have opened a new avenue into understanding glycemic control. The advent of CGMs heralded a host of promising possibilities, from personalized nutrition plans attuned to individual glycemic control (Zeevi et al., 2015; Hall et al., 2018; Berry et al., 2020; Ben-Yacov et al., 2021), to remote monitoring systems that deliver patient care based on their glycemic profiles and disease severity (Prahalad et al., 2022; Ferstad et al., 2021). All of these applications require a fine-grained method to subtype glycemic control and diabetic severity, a task which has proven difficult.

Mechanistic models of glucose dynamics have been considered in the past as tools for characterizing glycemic control (Hovorka et al., 1998). These ordinary differential equations (ODEs) provide *clinically interpretable* axes like insulin sensitivity and glucose effectiveness, which quantitatively summarize the disease severity across patients. However, these models must be fit for each person individually, requiring an intensive and invasive lab test such as an Oral Glucose Tolerance Test (OGTT), a Mixed Meal Tolerance Test (MMTT), or an Intravenous Glucose Tolerance Tests (IVGTT). These tests are expensive with long appointment backlogs, and require fasting, intravenous injections or insertions, and hours of immobility in a lab setting. As a result, these mechanistic measures of glycemic control are largely unused in diabetes diagnosis and management, relegating these

© 2023 K.A. Wang & E.B. Fox.

mechanistic models to be used only in scientific studies (Sugiyama et al., 2022) and bioengineering applications (Hovorka, 2011).

Beyond the challenge of acquiring high-quality data to fit such mechanistic models to an individual, the rigid form of these models has furthermore limited their translation to real-world, in-the-wild deployments. First, these ODE-based models are highly sensitive to inaccuracies in self-reported meal information, such as misestimated carbohydrate content, missing meal timestamps, and misreported meal timestamps, all of which are the norm (Landry et al., 2021). Likewise, these models cannot readily incorporate additional patient covariates or contextual information — such as demographics or fine-grained macronutrition data — requiring a bespoke mechanism for each new physiological pathway (Man et al., 2014).

In this paper, we use physiologically-grounded causal pathways of a mechanistic model to learn an *interpretable* representation of post-meal glycemic responses, while simultaneously addressing the challenges of *in-the-wild data*. The result is a framework that can characterize glycemic control *using only CGM data and self-reported meal information*, potentially enabling a new low-cost and non-invasive tool for characterizing diabetes severity.

More specifically, we introduce a hybrid variational autoencoder (VAE) that unites a flexible neural network encoder with a causally-grounded mechanistic ODE decoder. The latent variables act as the inputs to our mechanistic ODE, defining the axes of our clinically-interpretable embedding space. Simultaneously, our encoder infers an effective glucose appearance rate from CGM and meal logs, making the mechanistic model robust to erroneous meal data.

Our contributions are as follows:

1. We introduce a hybrid VAE that incorporates a mechanistic differential equation as its decoder. This allows us to ground the neural network on the underlying physiological processes governing glucose dynamics, pruning the function space to only the mechanistically plausible ones.

2. We use amortized variational inference to eliminate the need to refit the model for every person-meal pair, overcoming scalability issues of traditional mechanistic models. Our approximate Bayesian inference procedure ensures that the inferred parameters align with our prior knowledge of physiologically plausible parameter ranges.

3. We show that our hybrid model learns an interpretable mechanistic representation that characterizes glycemic control within and across individuals, resulting in patient stratification that agrees with known symptoms of diabetes pathogenesis, *despite not having access to diagnosis information and labwork data.*

By synergizing the flexibility of deep neural networks with the parsimony of mechanistic models, we produce a novel representation space that aligns with clinical understanding and opens a new avenue for scalable, non-invasive, yet fine-grained profiling of glycemic control[1].

2. Background

2.1. Mechanistic model of insulin-glucose dynamics.

The core of our mechanistic model is Bergman's minimal model of glucose regulation (Bergman et al., 1979) augmented by a standard two-compartment model of glucose appearance (Hovorka et al., 2002)

$$\dot{G}(t) = -X(t)G(t) + S_G(G - G_b) + G_2(t)/\tau_m \quad (1)$$

$$\dot{X}(t) = -p_2 X(t) + p_2 \cdot S_I I(t) \quad (2)$$

$$\dot{G}_1(t) = -G_1(t)/\tau_m + U_G(t) \quad (3)$$

$$\dot{G}_2(t) = G_1(t)/\tau_m - G_2(t)/\tau_m, \quad (4)$$

where I is the insulin concentration in the body, G_b is the basal glucose level, X is a mediator variable that describes the effect of insulin on glucose, and $U_G(t)$ is the rate of appearance of glucose in the body from meal consumption. The two-compartment model of glucose appearance smooths out and delays glucose appearance from when food is eaten. Food enters the system through $U_G(t)$, which is often called the glucose appearance rate. U_G is often parameterized via fixed functional forms and rescaled according to the amount of carbs that is eaten (Wang et al., 2022). Accurately learning and estimating $U_G(t)$ for generic meals is an active research area (Herrero et al., 2012; Eichenlaub et al., 2021; Wang et al., 2022). Here we use the standard assumption that $U_G(t) = u_G(t)/V_G$ where $u_G(t)$ is the rate of carbohydrate intake in mg/min and V_G is the accessible glucose volume. We set V_G to a constant due to its non-identifiability.

1. Our implementation is available at https://github.com/KeAWang/interpretable-cgm-representations

For individuals with pre-diabetes or type-2 diabetes who are not on external insulin therapy, I is simply the endogenous insulin I_{endo}. In our setting, we do not observe I, commonly assumed by prior works that seek to identify these model parameters (Kanderian et al., 2009; Hovorka et al., 2002), since we wish to apply these models outside of a lab setting. Hence, we must directly account for the unobserved insulin in our mechanistic model. We assume the endogenous insulin is proportional to the glucose deviation from its basal levels, a common model for insulin-producing individuals (Ruan et al., 2015):

$$I(t) = I_{\text{endo}}(t) = M_I \cdot \max(G(t) - G_b, 0). \tag{5}$$

Interpreting mechanistic parameters. Our method's interpretability comes from the causal grounding behind our model's mechanistic parameters:

- S_G is known as glucose effectiveness, governing the body's self-regulating ability to bring glucose levels down to baseline levels without insulin. Diabetes is associated with lower glucose effectiveness.
- M_I and S_I are known as insulin *productivity* and insulin *sensitivity*, respectively. Diabetes is associated with lower insulin productivity and sensitivity.
- $1/p_2$ and τ_m govern physiological timescales, corresponding to the peak effect of insulin and the peak appearance of glucose from the meal, respectively. Diabetes is associated with longer timescales, as it takes longer for blood glucose to be brought down to baseline levels.

Together with the basal glucose level G_b, these parameters provide a low-dimensional and succinct description of diabetes severity and heterogeneity. However, estimating these parameters outside of lab settings remains a challenge.

2.2. Latent Variable Modeling

In latent variable modeling, we assume there exists a joint distribution $p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z})p(z)$ between our observed data $\mathbf{x}$ and a set of latent variables $\mathbf{z}$. Then, for every distribution $q_\phi(\mathbf{z})$, there exists a variational lower bound on the log evidence/marginal likelihood:

$$\log p_\theta(\mathbf{x}) \geq \mathcal{L}(\theta, \phi; \mathbf{x}) = \mathbb{E}_{\mathbf{z}\sim q_\phi}\left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z})}\right]. \tag{6}$$

The evidence lower bound (ELBO) can be rewritten to be amenable to stochastic optimization with reparameterized samples (Kingma and Welling, 2013; Rezende et al., 2014; Titsias and Lázaro-Gredilla, 2014):

$$\begin{aligned}\mathcal{L}(\theta, \phi; \mathbf{x}) = \mathbb{E}_{\mathbf{z}\sim q_\phi}\left[\log p_\theta(\mathbf{x} \mid \mathbf{z})\right] \\ - \beta \cdot \mathrm{D}_{\mathrm{KL}}\left(q_\phi(\mathbf{z}) \,||\, p(\mathbf{z})\right)\end{aligned} \tag{7}$$

for $\beta \geq 1$. The ELBO decomposes into a sum of two terms, a reconstruction term and a regularization term. Tuning β trades off accurate reconstruction and sample quality for disentangled representations (Bowman et al., 2016; Higgins et al., 2016).

The decoder p_θ and encoder q_ϕ are usually parameterized by neural networks with parameters θ and ϕ that are trained by maximizing Equation 7 with reparameterized sampling. Solving this optimization problem by learning these parametric functions is known as amortized variational inference. The amortization comes from reusing these learned parametric functions to do posterior inference on new observations, instead of having to solve an optimization problem every time (Blei et al., 2017),

In this work, we instead parameterize the decoder with our mechanistic model, providing us with a low-dimensional and interpretable latent space.

3. Mechanistic Hybrid Variational Autoencoder

Our goal is to embed high-dimensional observations of glycemic response in a clinically-interpretable latent space. Such representations can be used for downstream tasks such as studying patient heterogeneity and subtypes, characterizing disease progression, or designing personalized nutrition. More formally, let $\mathbf{x} = (x_1, \ldots, x_T)$ denote a sequence of observations, such as blood sugar measurements from a continuous glucose monitor (CGM). Assume that every observation is associated with a context variable $\mathbf{c}$, such as demographics or a sequence of meals an individual ate at the corresponding timesteps $(c_1, \ldots, c_T)$. We want to infer a latent variable $\mathbf{z}$ associated with each $\mathbf{x}$. To provide clinical interpretability, we propose that $\mathbf{z}$ (or a transform thereof) correspond to the parameters of our mechanistic model.

Notice that we can freely replace the distribution $p_\theta(\mathbf{x} \mid \mathbf{z})$ and $q_\phi(\mathbf{z})$ in Equation 7 with conditional distributions, resulting in the following conditional

ELBO:

$$\mathcal{L}(\theta, \phi; \mathbf{x}, \mathbf{c}) = \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z} \mid \mathbf{c})} \left[\log p_\theta(\mathbf{x} \mid \mathbf{z}, \mathbf{c})\right] - \beta \cdot \mathrm{D}_{\mathrm{KL}} \left(q_\phi(\mathbf{z} \mid \mathbf{c}) \,||\, p(\mathbf{z} \mid \mathbf{c})\right). \tag{8}$$

To reduce the number of trainable parameters in our model for our data-scarce application, we make the simplifying assumption that $p(\mathbf{z} \mid \mathbf{c}) = p(\mathbf{z})$, though past works have found the conditional prior beneficial in other applications (Sohn et al., 2015). In our experiments, we found that it was very important to initialize the prior to a fixed expert prior based on what is physiologically plausible. See Section A for more details.

Grounding the latent space So far, we have kept the parameterization of our VAE completely general. Indeed, typical parameterizations use a neural network for both the encoder q_ϕ and the decoder p_θ to provide enough flexibility to approximate the posterior distribution and generate accurate reconstructions. Unfortunately, a purely black-box parameterization will produce an uninterpretable latent space. For our health application, we would like interpretable representations that also make use of our existing causal understanding of the body.

To ground the latent space in physiologically meaningful quantities, we return to the mechanistic ODEs defined by Equation 1 through Equation 5. *We parameterize the decoder with the Bergman model with endogenous insulin*, using a sequence of discrete Euler integration steps to autoregressively model p_θ. That is, we have $p_\theta(\mathbf{x} \mid \mathbf{z}, \mathbf{c}) = p_\theta(x_1 | \mathbf{z}, \mathbf{c}) \prod_{i=2}^{T} p_\theta(x_i | \mathbf{z}, x_{i-1}, \mathbf{c})$, where each factor is a normal distribution centered at an integration step output. To retain sufficient flexibility, we parameterize the encoder q_ϕ with a factorized normal distribution with means and scales output by an LSTM.

Our parameterization constrains the latent space to be the inputs of our mechanistic ODE. These inputs include the ODE initial state $\mathbf{x}_0 = [G(0), X(0), G_1(0), G_2(0)]$ and the ODE parameters $\mathbf{w} = [\tau_m, G_b, S_G, p_2, S_I, M_I]$.

Inferring effective glucose appearance Recall that a key component of the Bergman minimal model is the rate of appearance of glucose $U_G(t) = u_G(t)/V_G$ from Equation 3. However, in our in-the-wild setting the carbohydrate consumption rate u_G is known only through self-reported data, which frequently contain inaccurate meals or misreported timestamps. Such errors are especially challenging for mechanistic ODEs; for example, meals that are logged after the glucose rises violate the causal structure of the ODE.

To make our model robust to data issues with meal information, we propose to use the encoder to infer an "effective" carb consumption $\mathbf{u}$ from data, by including it as a latent variable. Here $\mathbf{u} = (u_1, \ldots, u_T)$ where $u_i = u_G(t_i)$. This is possible because rises and falls in CGM data contain information about when a meal was eaten and how much was eaten. By inferring $\mathbf{u}$, our encoder learns to extract this information from CGM, assisted by the meal information, without being restricted to the reported carb consumption. See Figure 1 for a visualization of this mechanism.

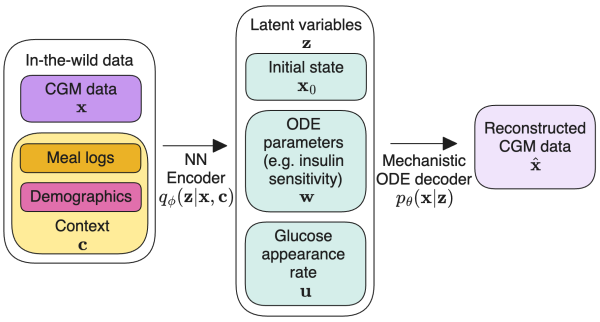

Figure 1: Visualization of our mechanistic hybrid VAE. Notably, our unsupervised method can be applied directly to in-the-wild data.

Practical implementation We would like our latent variable $\mathbf{z}$ to represent three sets of variables: $\mathbf{u}$, $\mathbf{x}_0$, and $\mathbf{w}$. However, all of these variables must be constrained to be non-negative and fall in physiologically plausible ranges, which is inconvenient to directly model probabilistically. Instead, we model these parameters in an unconstrained space. Thus we define $\mathbf{u} = \sigma_{\mathbf{u}}(\mathbf{u}')$, $\mathbf{x}_0 = \sigma_{\mathbf{x}_0}(\mathbf{x}_0')$, and $\mathbf{w} = \sigma_{\mathbf{w}}(\mathbf{w}')$ where each $\sigma_{(\cdot)}$ is an invertible transformation (e.g. rescaled sigmoid) that maps $\mathbb{R}$ to a constrained space $[a_{(\cdot)}, b_{(\cdot)}]$. Thus, we finally arrive at our latent variable $\mathbf{z} = (\mathbf{u}', \mathbf{x}_0', \mathbf{w}')$, which allows us to do variational inference over an unconstrained space. When decoding, we map the unconstrained latent variables into a physiologically-plausible constrained space for both the parameters and the initial states. See Section B for details on the constraint ranges.

To account for missing values in $\mathbf{x}$ and $\mathbf{c}$, we project each timestep into separate 8-dimensional embedding spaces and mask the entire timestep in embedding space when the corresponding timestep has a missing channel as done in (Yue et al., 2021).

4. Dataset

In our experiments we use the Keto-Med dataset (Landry et al., 2021), a randomized crossover trial (NCT03810378) that placed participants on two different low-carb diets, Ketogenic and Mediterranean. The participants were given controlled meals during the first four out of twelve weeks for each diet, but were responsible for their own diet adherence for the last eight weeks. The study targeted adults from the San Francisco Bay area with Type-2 diabetes or prediabetes. Only a subset of recorded meals have accompanying glucose measurements, since participants were only given CGMs periodically during the study. The meal information in this dataset is representative of data-collection challenges for in-the-wild conditions, including missing meals, inaccurate meal information, and mistimed reporting. From each meal, we extract the total amount of food eaten, total carb, total sugar, total dietary fiber, total fat, and total protein as the meal covariates. Note that not all meals contain all of these macronutritional information.

After preprocessing and joining the data, we end up with a total of 964 postprandial glucose response (PPGR) time-series from 33 individuals. We define a PPGR to be a sequence of length $T = 60$ (5 hours), containing CGM and self-reported information, with a food intake occurring at timestep $t = 12$ (hour 1). We include 1 hour of context before the meal to give the model information about glucose excursions prior to the meal. The 4-hour horizon following the meal is set to be sufficiently long to capture the temporal effects of the meal. Note that often other foods are eaten during this 5 hour period, making the problem of estimating mechanistic parameters even more difficult.

5. Results

We apply our hybrid VAE model to the CGM and meal dataset described in Section 4. We compare to a baseline non-hybrid VAE model with a similar number of parameters (53,561) as our model (41,454). Both models use a 2-layer bidirectional LSTM encoder with a hidden state size of 32. The non-hybrid VAE uses a 1-layer uni-directional LSTM decoder with the same hidden state size as well, as a 32-dimensional black-box latent space $\mathbf{z}$. Both models use $\hat{\beta} = 0.01$ as defined in Higgins et al. (2016), a batch size of 64, and the ADAM (Kingma and Ba, 2015) optimizer with a constant learning rate of 0.01. Our model converged quickly in 100 epochs of training, while the black-box model was trained for 1000 epochs. Importantly, we use a high dropout rate of 0.5 following the latent $\mathbf{u}$ to regularize the inferred $\mathbf{u}$ away from non-plausible meal sequences (Srivastava et al., 2014). We tune parameters for both models to train until convergence and achieve similar reconstruction losses. Both models are trained on CGM data, meal data, and demographics data (age, weight, and sex). Static variables like demographics are repeated and concatenated with the per-timestep meal data as $\mathbf{c}$. We leave out additional contextual variables such as BMI and HbA1C from model training since these variables are highly correlated with the hidden diagnosis information, potentially biasing the models' representations.

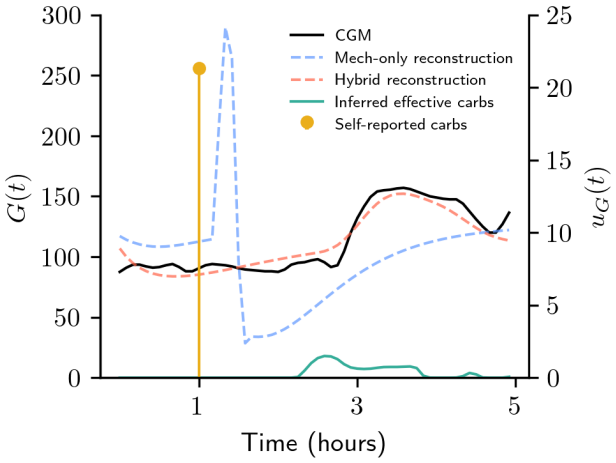

Figure 2: The mechanistic-only model is brittle to incorrect self-reported meal data. Here, it is unable to explain the delayed rise in glucose due to the early reporting. However, our hybrid model can compensate for this and infer an effective carb rate, leading to a more accurate reconstruction.

Robustness to self-reported meals We first show the robustness of our model to mistimed meal reports in Figure 2. We train the mechanistic model by doing direct gradient descent on its parameters and initial state with respect to mean-squared error. The reported meal sequence here is typical of a standard PPGR in-the-wild. Although macronutritional components, such as high fat content, can delay the rise of glucose (Ann Ahern et al., 1993), Figure 2 is a clear case of a meal that was logged *too early* with

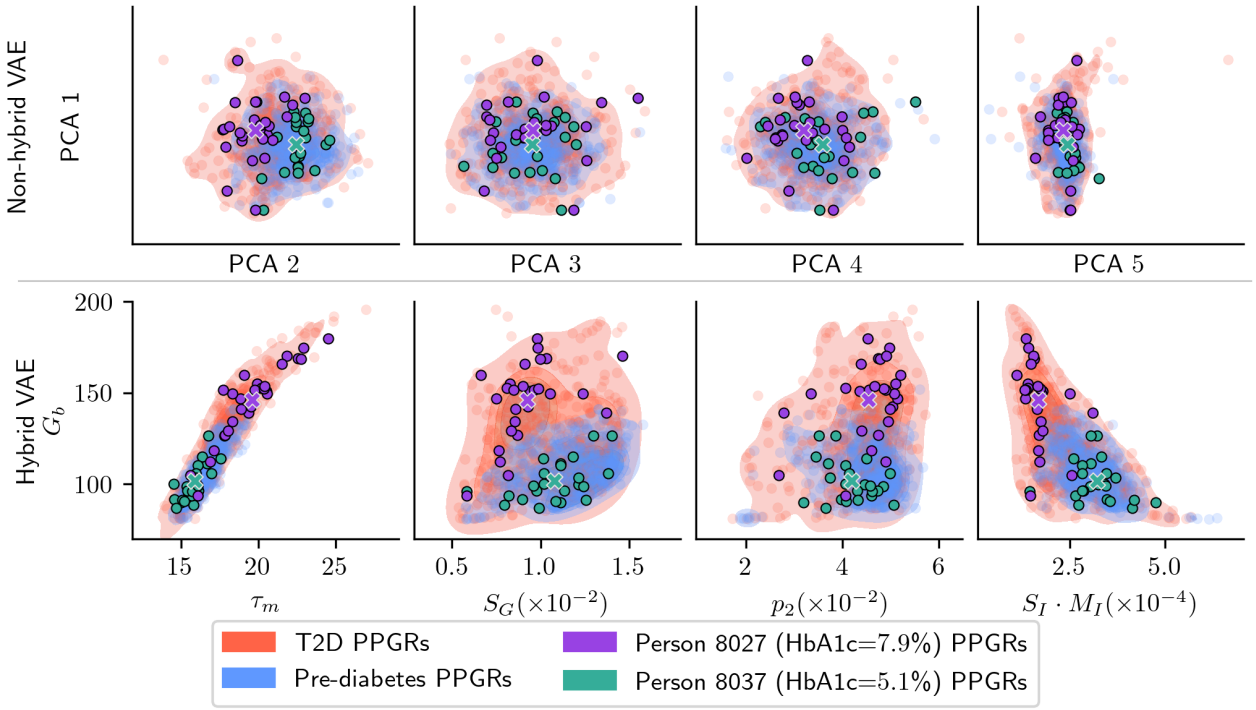

Figure 3: Post-prandial glucose responses (PPGRs) for all meals projected into a non-hybrid VAE's embedding space (top) and our mechanistic embedding space (bottom). Our embeddings are learned *without* any diagnosis information or diabetes indicators besides CGM. The embeddings of the two groups overlap as expected. Notably, the distinction between the two groups aligns with our understanding of diabetes disease progression; individuals with diabetes have higher basal glucose G_b, more extended meal responses τ_m, and lower insulin sensitivity and insulin productivity ($S_I \cdot M_I$). We also show the average embedding for two individuals on opposite ends of the disease spectrum (8027 and 8037), outlined in white.

a possible carb overestimate. As a result, the brittle mechanistic-only model must contort the mechanistic parameters to describe the observed CGM data, leading to poor reconstructions. In contrast, our hybrid VAE infers an "effective" meal consumption rate u_G that can explain the data much better. Although not shown here, the mechanistic-only model also fails to account for rises in glucose when carb information is missing entirely or reported late. Late and missing meals logs are especially challenging for mechanistic-only models, since they violate the causality of the ODE.

Physiologically accurate meal-level embedding space Figure 3 shows the distribution of all meal responses (PPGRs) embedded into the latent space of both respective models. Since we do not observe the true insulin of the individuals, S_I and M_I are not identifiable on their own, only their product is identifiable. Hence we visualize $S_I \cdot M_I$ in our results.

By encoding the data with physiologically meaningful variables, our hybrid VAE can naturally separate individuals with pre-diabetes from individuals with type-2 diabetes *despite not being given any diagnosis information.* Moreover, the two populations are characterized in a way that is directly in line with our clinical understanding of the progression from pre-diabetes to Type-2 diabetes. Through its mechanistic grounding, the hybrid VAE can discover not just higher basal glucose levels G_b and longer elevated glucose levels τ_m among people with diabetes, *but also reduced glucose effectiveness S_G and insulin sensitivity $S_I \cdot M_I$.* In addition, we see a clear separation *in every dimension* between the average embedding for person 8037 and 8027, who have the lowest and highest levels of HbA1C in the dataset, respectively. Our results show that our meal-level mechanistic embedding space can serve as a powerful way to visualize *intrapatient* variations in glycemic control.

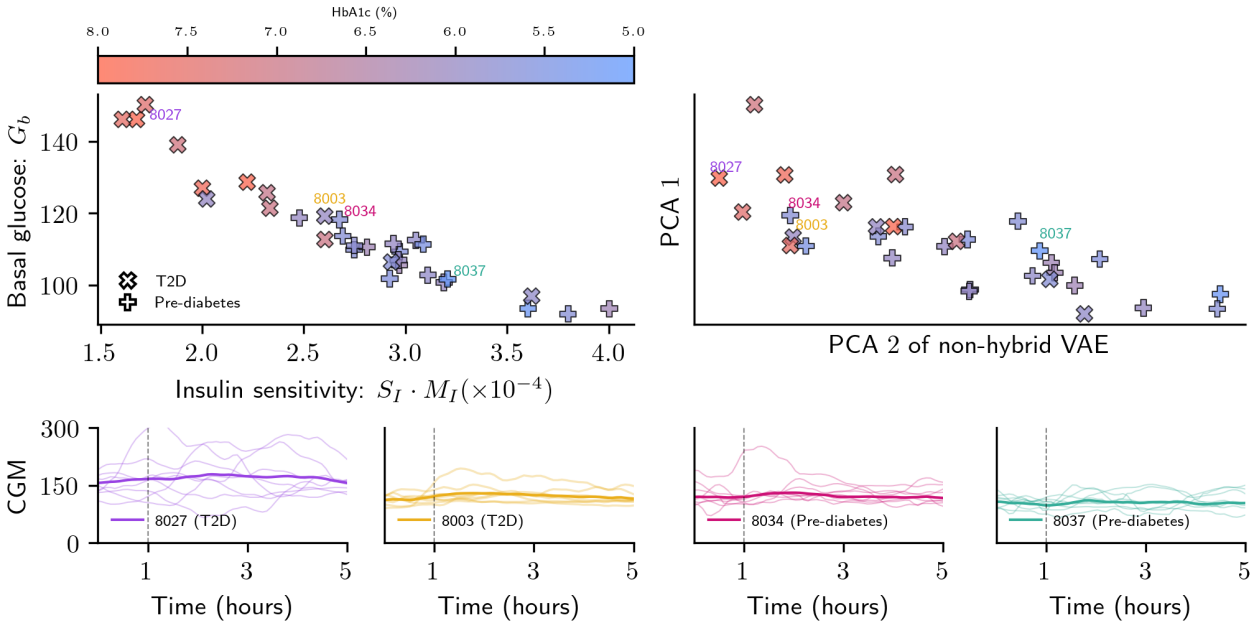

Figure 4: Here we show the average of each patient's PPGR embeddings in both representations. Notice that the mechanistic embeddings are not only more interpretable but also more consistent with HbA1C characterizations of disease progression. Patients that are similar in HbA1C tend to also be closer together in our mechanistic space, even though the model was not trained on HbA1C. Not only are our representations more interpretable, they also provide better separation. A linear decision boundary on this simple 2D plane can correctly classify 29 out of 33 individuals versus 26 on the black-box PCA plane. We choose a subset of individuals to visualize the disease heterogeneity as characterized by CGM data along this spectrum.

In contrast, 2D principal component analysis (PCA) projections of the non-hybrid model's latent space show a higher overlap between the two populations. In fact, in all but one of the projections (PCA 1 vs PCA 2), the PPGRs from both populations are completely interspersed. Although there is a slight separation in the PCA 1 vs PCA 2 projection, it is unclear what the two axes PCA 1 and PCA 2 correspond to physiologically.

Interpretable characterization of disease progression In addition to intrapatient variation, our mechanistic representation can effectively characterize *interpatient* variation in disease progression. Figure 4 (top) shows the average embedding by both models for every individual in the dataset along with their HbA1C levels, the current gold standard of characterizing diabetes severity. Note that HbA1C was not used to train either model.

From this visualization, we see that our hybrid-VAE's learned basal glucose G_b and insulin sensitivity $S_I \cdot M_I$ correlates well with increasing HbA1C levels. Although it is known that HbA1C can be well estimated by baseline glucose levels, here we also show its correlation with reduced insulin sensitivity.

Although the black-box embeddings of the non-hybrid VAE model shows a similar trend along its first two principal components, the model wrongly places a person with diabetes with one of the highest HbA1C levels in the study into a cluster of medium-level HbA1C individuals.

Evaluating Embedding Quality In Table 1 we quantitatively evaluate our Hybrid VAE's embedding space relative to 6 other baselines, separated into "naive", "expert", "black-box", and "graybox" categories. For each method, we represent an individual by averaging their embeddings across all their PPGRs. We then perform k-means clustering with $k = 2$. We evaluate each method by standard clustering metrics: normalized mutual information (NMI), adjusted mutual information (AMI), homogeneity, and completeness, as implemented in (Pedregosa et al., 2011). All metrics fall in the range of $[0, 1]$ with 1 be-

ing the highest. We use the diagnosis: "pre-diabetes" vs "type 2 diabetes as ground truth clusters. Our Hybrid VAE consistently produces the highest quality low-dimensional embedding space, with significantly better interpretability.

The naive baselines include "Raw CGM" and "Raw CGM + DTW". For these two baselines, we use the original raw CGM space as the embedding space. The latter uses k-means with Dynamic Time Warping (DTW), a standard time-series clustering method (Sakoe and Chiba, 1978), as implemented by Tavenard et al. (2020). Note that "CGM" and "CGM + DTW" do not involve a low-dimensional embedding space: they operate in the original raw sensor space.

The expert features include a set of 10 well-established CGM-derived features that are used to clinically measure glycemic variability and control (Kovatchev, 2017). The features are: mean glucose, glucose standard deviation, glucose coefficient of variability (CV) (Bergenstal et al., 2013), max glucose value, min glucose value, percent time in range (TIR) as recommended by the American Diabetes Association guidelines (American Diabetes Association Professional Practice Committee, 2021), glucose arc length (Peyser et al., 2018), J-index (Wójcicki, 1995), mean amplitude of glucose excursion (MAGE)(Service et al., 1970), incremental area under the glucose curve relative to the starting baseline (iAUC) (Brouns et al., 2005). We excluded other metrics such as eA1C (Bergenstal et al., 2018), CONGA (McDonnell et al., 2005), and MODD (Molnar et al., 1972), as they are long-term glycemic variablility metrics and require longer CGM sequences than our 6 hour PPGR period. Nevertheless, these excluded metrics are closely related to the metrics we chose; for example, eA1C is merely an affine transform of the mean glucose.

Besides comparing to the latent space of the black-box VAE described in Section 5, we also compare to Time Contrastive Learning (TCL) (Hyvarinen and Morioka, 2016), a strong baseline for learning black-box embeddings. We evenly divide each PPGR into 3 time-windows for TCL, using a 2 hidden-layer MLP with 32 hidden dimensions, corresponding to 5,667 trainable parameters. The output of the final pre-activation layer is used as the embedding space. Since TCL produces an embedding for each time-window, we compare averaging the time-window embeddings against concatenating the time-window embeddings.

Finally, we compare to "Mechanistic ODE", which uses the ODE parameters of a non-neural mechanistic model, i.e. the decoder of our hybrid ODE, as its embedding space. Table 1 confirms the stark contrast hinted by Figure 2. Hybridizing the mechanistic ODE results in 3 to 4 times higher clustering scores, showing the need to augment the original mechanistic model with more flexible components.

	NMI	AMI	Hom.	Comp.
Raw CGM ($d = 60$)	0.41	0.39	0.36	0.47
Raw CGM + DTW ($d = 60$)	0.54	0.53	**0.62**	0.51
Expert features ($d = 10$)	0.25	0.25	0.61	0.21
TCL + Average ($d = 32$)	0.24	0.21	0.19	0.33
TCL + Concat ($d = 3 \times 32$)	0.29	0.27	0.24	0.38
Black-box VAE ($d = 32$)	0.21	0.19	0.22	0.21
Mechanistic ODE ($d = 7$)	0.14	0.12	0.14	0.14
Hybrid VAE ($d = 7$)	**0.54**	**0.53**	0.51	**0.58**

Table 1: Embedding quality as measured by k-means clustering on each individual's average embedding, using normalized mutual information, adjusted normalized mutual information, homogeneity, and completeness. Higher is better. "Pre-diabetes" and "type 2 diabetes" labels form the ground truth clusters. Our embeddings produce up to 4x better clusterings than other methods.

6. Related Work

Until recently, mechanistic models have been primarily used for system identification in systems with few degrees of freedom (Hodgkin and Huxley, 1952; Varah, 1982; Bergman et al., 1979; Leander et al., 2014). Many recent works have explored ways to combine the inductive biases of mechanistic knowledge with the flexibility of black-box machine learning models. These models are also known as hybrid models (Miller et al., 2021; Levine and Stuart, 2021) or gray-box models (Duun-Henriksen et al., 2013), and they are especially relevant in healthcare and biological settings where data efficiency and interpretability are paramount (Baker et al., 2018).

In healthcare applications, recent works have shown that hybrid models can improve forecasting and prediction performance compared to black-box models when data is scarce. Miller et al. (2020) used a highly flexible mechanistic model for Type-1 diabetes to produce more physiologically plausible multi-hour CGM forecasts compared to purely black-box models and classic forecasting techniques. Hussain et al. (2021) used an attention-based hybrid model that ensembles multiple mechanistic pathways from pharmacological modeling, allowing them

to more accurately model cancer disease progression. In parallel, Qian et al. (2021) augments a mechanistic state space pharmacokinetics-pharmacodynamics model with black-box latent variables to model patient biomarkers longitudinally.

Outside of healthcare, hybrid models have been successfully applied in engineering and physics applications where the mechanistic models are better specified (Takeishi and Kalousis, 2021; Yin et al., 2021; Wehenkel et al., 2022; Karniadakis et al., 2021).

There have also been prior works that seek to predict post-prandial glycemic responses (Ben-Yacov et al., 2021), characterize diabetes disease progression (Wilson et al., 2023), and subtype glycemic control (Keshet et al., 2023). However, these works rely solely on expert hand-crafted features, which are lower quality than our learned embeddings, as demonstrated in Table 1.

In contrast to prior literature that primarily focus on predictive accuracy or disease subtyping with handcrafted features, our work *learns* an interpretable representation that is useful for comparing disease variation using *in-the-wild data*, *without needing label information*.

7. Discussion

The advent of continuous glucose monitors has opened a new frontier in understanding individualized glycemic control, offering a level of granularity previously unattainable. Yet, the full potential of this rich data source remains untapped due to the limitations of existing analytical tools. Our work addresses a part of this gap by marrying the flexibility of neural networks with the causal structure inherent in metabolic mechanistic models. Our hybrid approach enables us to learn physiologically grounded representations directly from in-the-wild data, without the constraints of invasive lab tests. Indeed, we can have *both* interpretability and separability in a single embedding space, all learned from out-patient data.

Traditional mechanistic models, while grounded in physiological understanding, often falter when applied to self-reported data. They struggle with missing meals, inaccurate meal estimates, and other imperfections that are common in outpatient settings. In contrast, our method is designed with these challenges in mind while remaining scalable and practical for real-world data.

Our hybrid VAE has shown to be a powerful tool for subtyping interindividual and intraindividual glycemic control, which is a key step in applications like personalized nutrition (Ben-Yacov et al., 2021) and remote health monitoring. As CGMs become ever more accessible, our framework has the promise of becoming a new low-cost component for the next generation of non-invasive patient-facing diabetes care.

Acknowledgments

This work was supported in part by AFOSR Grant FA9550-21-1-0397, ONR Grant N00014-22-1-2110, National Science Foundation Grant 2205084, the Stanford Institute for Human-Centered Artificial Intelligence (HAI). KAW was partially supported by Stanford Data Science as a Stanford Data Science Scholar. EBF is a Chan Zuckerberg Biohub – San Francisco Investigator.

We thank Matthew J. Landry and Christopher Gardner for providing us access to the Keto-Med dataset. We thank David Maahs, Priya Prahalad, Jiaxin Shi, and Rahul G. Krishnan for insightful discussions.

We thank the creators and maintainers of open-source software we relied on that made this work possible (Pedregosa et al., 2011; Virtanen et al., 2020; Van Rossum and Drake Jr, 1995; Harris et al., 2020; Paszke et al., 2019; Tavenard et al., 2020; Hunter, 2007).

References

American Diabetes Association Professional Practice Committee. 6. Glycemic Targets: Standards of Medical Care in Diabetes—2022. *Diabetes Care*, 45(Supplement_1):S83–S96, December 2021. ISSN 0149-5992. doi: 10.2337/dc22-S006. URL https://doi.org/10.2337/dc22-S006.

Jo Ann Ahern, Patricia M Gatcomb, NANcy A Held, William A Petit, Jr, and William V Tamborlane. Exaggerated Hyperglycemia After A Pizza Meal in Well-Controlled Diabetes. *Diabetes Care*, 16(4): 578–580, April 1993. ISSN 0149-5992. doi: 10.2337/diacare.16.4.578. URL https://doi.org/10.2337/diacare.16.4.578.

Ruth E. Baker, Jose-Maria Peña, Jayaratnam Jayamohan, and Antoine Jérusalem. Mechanistic

models versus machine learning, a fight worth fighting for the biological community? *Biology Letters*, 14(5):20170660, May 2018. doi: 10.1098/rsbl.2017.0660. URL https://royalsocietypublishing.org/doi/10.1098/rsbl.2017.0660. Publisher: Royal Society.

Orly Ben-Yacov, Anastasia Godneva, Michal Rein, Smadar Shilo, Dmitry Kolobkov, Netta Koren, Noa Cohen Dolev, Tamara Travinsky Shmul, Bat Chen Wolf, Noa Kosower, Keren Sagiv, Maya Lotan-Pompan, Niv Zmora, Adina Weinberger, Eran Elinav, and Eran Segal. Personalized Postprandial Glucose Response–Targeting Diet Versus Mediterranean Diet for Glycemic Control in Prediabetes. *Diabetes Care*, 44(9):1980–1991, July 2021. ISSN 0149-5992. doi: 10.2337/dc21-0162. URL https://doi.org/10.2337/dc21-0162.

Richard M. Bergenstal, Andrew J. Ahmann, Timothy Bailey, Roy W. Beck, Joan Bissen, Bruce Buckingham, Larry Deeb, Robert H. Dolin, Satish K. Garg, Robin Goland, Irl B. Hirsch, David C. Klonoff, Davida F. Kruger, Glenn Matfin, Roger S. Mazze, Beth A. Olson, Christopher Parkin, Anne Peters, Margaret A. Powers, Henry Rodriguez, Phil Southerland, Ellie S. Strock, William Tamborlane, and David M. Wesley. Recommendations for Standardizing Glucose Reporting and Analysis to Optimize Clinical Decision Making in Diabetes: The Ambulatory Glucose Profile. *Journal of Diabetes Science and Technology*, 7(2):562–578, March 2013. ISSN 1932-2968. doi: 10.1177/193229681300700234. URL https://doi.org/10.1177/193229681300700234. Publisher: SAGE Publications Inc.

Richard M. Bergenstal, Roy W. Beck, Kelly L. Close, George Grunberger, David B. Sacks, Aaron Kowalski, Adam S. Brown, Lutz Heinemann, Grazia Aleppo, Donna B. Ryan, Tonya D. Riddlesworth, and William T. Cefalu. Glucose Management Indicator (GMI): A New Term for Estimating A1C From Continuous Glucose Monitoring. *Diabetes Care*, 41(11):2275–2280, September 2018. ISSN 0149-5992. doi: 10.2337/dc18-1581. URL https://doi.org/10.2337/dc18-1581.

R. N. Bergman, G. Bortolan, C. Cobelli, and G. Toffolo. Identification of a Minimal Model of Glucose Disappearance for Estimating Insulin Sensitivity. *IFAC Proceedings Volumes*, 12(8):883–890, September 1979. ISSN 1474-6670. doi: 10.1016/S1474-6670(17)65505-8. URL https://www.sciencedirect.com/science/article/pii/S1474667017655058.

Sarah E. Berry, Ana M. Valdes, David A. Drew, Francesco Asnicar, Mohsen Mazidi, Jonathan Wolf, Joan Capdevila, George Hadjigeorgiou, Richard Davies, Haya Al Khatib, Christopher Bonnett, Sajaysurya Ganesh, Elco Bakker, Deborah Hart, Massimo Mangino, Jordi Merino, Inbar Linenberg, Patrick Wyatt, Jose M. Ordovas, Christopher D. Gardner, Linda M. Delahanty, Andrew T. Chan, Nicola Segata, Paul W. Franks, and Tim D. Spector. Human postprandial responses to food and potential for precision nutrition. *Nature Medicine*, 26(6):964–973, June 2020. ISSN 1546-170X. doi: 10.1038/s41591-020-0934-0. URL https://www.nature.com/articles/s41591-020-0934-0. Number: 6 Publisher: Nature Publishing Group.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877, April 2017. ISSN 0162-1459. doi: 10.1080/01621459.2017.1285773. URL https://doi.org/10.1080/01621459.2017.1285773. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2017.1285773.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. Generating Sentences from a Continuous Space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/K16-1002. URL https://aclanthology.org/K16-1002.

F. Brouns, I. Bjorck, K. N. Frayn, A. L. Gibbs, V. Lang, G. Slama, and T. M. S. Wolever. Glycaemic index methodology. *Nutrition Research Reviews*, 18(1):145–171, June 2005. ISSN 1475-2700, 0954-4224. doi: 10.1079/NRR2005100. URL https://www.cambridge.org/core/journals/nutrition-research-reviews/article/glycaemic-index-methodology/4AEE3FFEBBA327530129C2381907D4F2. Publisher: Cambridge University Press.

S. A. Brown, A. Basu, and B. P. Kovatchev. Beyond HbA1c: using continuous glucose monitoring metrics to enhance interpretation of treatment effect and improve clinical decision-making. *Diabetic Medicine*, 36(6):679–687, 2019. ISSN 1464-5491. doi: 10.1111/dme.13944. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/dme.13944.

Anne Katrine Duun-Henriksen, Signe Schmidt, Rikke Meldgaard Røge, Jonas Bech Møller, Kirsten Nørgaard, John Bagterp Jørgensen, and Henrik Madsen. Model identification using stochastic differential equation grey-box models in diabetes. *Journal of Diabetes Science and Technology*, 7 (2):431–440, March 2013. ISSN 1932-2968. doi: 10.1177/193229681300700220.

Manuel M. Eichenlaub, John G. Hattersley, Mary C. Gannon, Frank Q. Nuttall, and Natasha A. Khovanova. Bayesian parameter estimation in the oral minimal model of glucose dynamics from non-fasting conditions using a new function of glucose appearance. *Computer Methods and Programs in Biomedicine*, 200:105911, March 2021. ISSN 0169-2607. doi: 10.1016/j.cmpb.2020.105911. URL https://www.sciencedirect.com/science/article/pii/S0169260720317442.

Johannes O. Ferstad, Jacqueline J. Vallon, Daniel Jun, Angela Gu, Anastasiya Vitko, Dianelys P. Morales, Jeannine Leverenz, Ming Yeh Lee, Brianna Leverenz, Christos Vasilakis, Esli Osmanlliu, Priya Prahalad, David M. Maahs, Ramesh Johari, and David Scheinker. Population-level management of type 1 diabetes via continuous glucose monitoring and algorithm-enabled patient prioritization: Precision health meets population health. *Pediatric Diabetes*, 22(7):982–991, 2021. ISSN 1399-5448. doi: 10.1111/pedi.13256. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/pedi.13256.

Joon Ha and Arthur Sherman. Type 2 diabetes: one disease, many pathways. *American Journal of Physiology-Endocrinology and Metabolism*, 319(2):E410–E426, August 2020. ISSN 0193-1849. doi: 10.1152/ajpendo.00512.2019. URL https://journals.physiology.org/doi/full/10.1152/ajpendo.00512.2019. Publisher: American Physiological Society.

Heather Hall, Dalia Perelman, Alessandra Breschi, Patricia Limcaoco, Ryan Kellogg, Tracey McLaughlin, and Michael Snyder. Glucotypes reveal new patterns of glucose dysregulation. *PLOS Biology*, 16(7):e2005143, July 2018. ISSN 1545-7885. doi: 10.1371/journal.pbio.2005143. URL https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.2005143. Publisher: Public Library of Science.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585 (7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2. Publisher: Springer Science and Business Media LLC.

Pau Herrero, Jorge Bondia, Cesar C. Palerm, Josep Vehí, Pantelis Georgiou, Nick Oliver, and Christofer Toumazou. A Simple Robust Method for Estimating the Glucose Rate of Appearance from Mixed Meals. *Journal of Diabetes Science and Technology*, 6(1):153–162, January 2012. ISSN 1932-2968. doi: 10.1177/193229681200600119. URL https://doi.org/10.1177/193229681200600119. Publisher: SAGE Publications Inc.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. November 2016. URL https://openreview.net/forum?id=Sy2fzU9gl.

A. L. Hodgkin and A. F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, 117(4):500–544, 1952. ISSN 1469-7793. doi: 10.1113/jphysiol.1952.sp004764. URL https://onlinelibrary.wiley.com/doi/abs/10.1113/jphysiol.1952.sp004764.

Roman Hovorka. Closed-loop insulin delivery: from bench to clinical practice. *Nature Reviews Endocrinology*, 7(7):385–395, July 2011. ISSN 1759-5037. doi: 10.1038/nrendo.2011.32. URL https://www.nature.com/articles/nrendo.2011.32. Number: 7 Publisher: Nature Publishing Group.

Roman Hovorka, Ludovic Chassin, Stephen D. Luzio, Rebecca Playle, and David R. Owens. Pancreatic β-Cell Responsiveness during Meal Tolerance Test: Model Assessment in Normal Subjects and Subjects with Newly Diagnosed Noninsulin-Dependent Diabetes Mellitus1. *The Journal of Clinical Endocrinology & Metabolism*, 83(3):744–750, March 1998. ISSN 0021-972X. doi: 10.1210/jcem.83.3.4646. URL https://doi.org/10.1210/jcem.83.3.4646.

Roman Hovorka, Fariba Shojaee-Moradie, Paul V. Carroll, Ludovic J. Chassin, Ian J. Gowrie, Nicola C. Jackson, Romulus S. Tudor, A. Margot Umpleby, and Richard H. Jones. Partitioning glucose distribution/transport, disposal, and endogenous production during IVGTT. *American Journal of Physiology-Endocrinology and Metabolism*, 282(5):E992–E1007, May 2002. ISSN 0193-1849. doi: 10.1152/ajpendo.00304.2001. URL https://journals.physiology.org/doi/full/10.1152/ajpendo.00304.2001. Publisher: American Physiological Society.

J. D. Hunter. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3): 90–95, 2007. doi: 10.1109/MCSE.2007.55. Publisher: IEEE COMPUTER SOC.

Zeshan M. Hussain, Rahul G. Krishnan, and David Sontag. Neural Pharmacodynamic State Space Modeling. In *Proceedings of the 38th International Conference on Machine Learning*, pages 4500–4510. PMLR, July 2021. URL https://proceedings.mlr.press/v139/hussain21a.html. ISSN: 2640-3498.

Aapo Hyvarinen and Hiroshi Morioka. Unsupervised Feature Extraction by Time-Contrastive Learning and Nonlinear ICA. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/hash/d305281faf947ca7acade9ad5c8c818c-Abstract.html.

Sami S. Kanderian, Stu Weinzimer, Gayane Voskanyan, and Garry M. Steil. Identification of intraday metabolic profiles during closed-loop glucose control in individuals with type 1 diabetes. *Journal of Diabetes Science and Technology*, 3(5): 1047–1057, September 2009. ISSN 1932-2968. doi: 10.1177/193229680900300508.

George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, June 2021. ISSN 2522-5820. doi: 10.1038/s42254-021-00314-5. URL https://www.nature.com/articles/s42254-021-00314-5. Number: 6 Publisher: Nature Publishing Group.

Ayya Keshet, Smadar Shilo, Anastasia Godneva, Yeela Talmor-Barkan, Yaron Aviv, Eran Segal, and Hagai Rossman. CGMap: Characterizing continuous glucose monitor data in thousands of non-diabetic individuals. *Cell Metabolism*, 35(5):758–769.e3, May 2023. ISSN 1550-4131. doi: 10.1016/j.cmet.2023.04.002. URL https://www.sciencedirect.com/science/article/pii/S1550413123001298.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. 2015. URL https://arxiv.org/abs/1412.6980.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. December 2013. URL https://openreview.net/forum?id=33X9fd2-9FyZd.

Boris P. Kovatchev. Metrics for glycaemic control — from HbA1c to continuous glucose monitoring. *Nature Reviews Endocrinology*, 13(7):425–436, July 2017. ISSN 1759-5037. doi: 10.1038/nrendo.2017.3. URL https://www.nature.com/articles/nrendo.2017.3. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7 Primary_atype: Reviews Publisher: Nature Publishing Group Subject_term: Diagnostic devices;Outcomes research;Prognostic markers;Type 1 diabetes;Type 2 diabetes Subject_term_id: diagnostic-devices;outcomes-research;prognostic-markers;type-1-diabetes-mellitus;type-2-diabetes-mellitus.

Matthew J. Landry, Anthony Crimarco, Dalia Perelman, Lindsay R. Durand, Christina Petlura, Lucia

Aronica, Jennifer L. Robinson, Sun H. Kim, and Christopher D. Gardner. Adherence to Ketogenic and Mediterranean Study Diets in a Crossover Trial: The Keto–Med Randomized Trial. *Nutrients*, 13(3):967, March 2021. ISSN 2072-6643. doi: 10.3390/nu13030967. URL https://www.mdpi.com/2072-6643/13/3/967. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

Jacob Leander, Torbjörn Lundh, and Mats Jirstrand. Stochastic differential equations as a tool to regularize the parameter estimation problem for continuous time dynamical systems given discrete time measurements. *Mathematical Biosciences*, 251:54–62, May 2014. ISSN 0025-5564. doi: 10.1016/j.mbs.2014.03.001. URL https://www.sciencedirect.com/science/article/pii/S0025556414000510.

M. Levine and A. Stuart. A Framework for Machine Learning of Model Error in Dynamical Systems. *ArXiv*, 2021.

Chiara Dalla Man, Francesco Micheletto, Dayu Lv, Marc Breton, Boris Kovatchev, and Claudio Cobelli. The UVA/PADOVA Type 1 Diabetes Simulator: New Features. *Journal of Diabetes Science and Technology*, 8(1):26–34, January 2014. ISSN 1932-2968. doi: 10.1177/1932296813514502. URL https://doi.org/10.1177/1932296813514502. Publisher: SAGE Publications Inc.

C.m. McDonnell, S.m. Donath, S.i. Vidmar, G.a. Werther, and F.j. Cameron. A Novel Approach to Continuous Glucose Analysis Utilizing Glycemic Variation. *Diabetes Technology & Therapeutics*, 7(2):253–263, April 2005. ISSN 1520-9156. doi: 10.1089/dia.2005.7.253. URL https://www.liebertpub.com/doi/10.1089/dia.2005.7.253. Publisher: Mary Ann Liebert, Inc., publishers.

Andrew C. Miller, Nicholas J. Foti, and Emily Fox. Learning Insulin-Glucose Dynamics in the Wild. In *Proceedings of the 5th Machine Learning for Healthcare Conference*, pages 172–197. PMLR, September 2020. URL https://proceedings.mlr.press/v126/miller20a.html. ISSN: 2640-3498.

Andrew C. Miller, Nicholas J. Foti, and Emily B. Fox. Breiman's Two Cultures: You Don't Have to Choose Sides. *Observational Studies*, 7(1):161–169, 2021. ISSN 2767-3324. doi: 10.1353/obs.2021.0003. URL https://muse.jhu.edu/article/799728. Publisher: University of Pennsylvania Press.

G. D. Molnar, W. F. Taylor, and M. M. Ho. Day-to-day variation of continuously monitored glycaemia: A further measure of diabetic instability. *Diabetologia*, 8(5):342–348, November 1972. ISSN 1432-0428. doi: 10.1007/BF01218495. URL https://doi.org/10.1007/BF01218495.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Thomas A. Peyser, Andrew K. Balo, Bruce A. Buckingham, Irl B. Hirsch, and Arturo Garcia. Glycemic Variability Percentage: A Novel Method for Assessing Glycemic Variability from Continuous Glucose Monitor Data. *Diabetes Technology & Therapeutics*, 20(1):6–16, January 2018. ISSN 1520-9156. doi: 10.1089/dia.2017.0187. URL https://www.liebertpub.com/doi/full/10.1089/dia.2017.0187. Publisher: Mary Ann Liebert, Inc., publishers.

Priya Prahalad, Victoria Y Ding, Dessi P Zaharieva, Ananta Addala, Ramesh Johari, David Scheinker, Manisha Desai, Korey Hood, and David M Maahs. Teamwork, Targets, Technology, and Tight Control in Newly Diagnosed Type 1 Diabetes: the Pilot 4T Study. *The Journal of Clinical Endocrinology & Metabolism*, 107(4):998–1008, April 2022. ISSN

0021-972X. doi: 10.1210/clinem/dgab859. URL https://doi.org/10.1210/clinem/dgab859.

Zhaozhi Qian, William Zame, Lucas Fleuren, Paul Elbers, and Mihaela van der Schaar. Integrating Expert ODEs into Neural ODEs: Pharmacology and Disease Progression. In *Advances in Neural Information Processing Systems*, volume 34, pages 11364–11383. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/5ea1649a31336092c05438df996a3e59-Abstract.html.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1278–1286. PMLR, June 2014. URL https://proceedings.mlr.press/v32/rezende14.html. ISSN: 1938-7228.

Yue Ruan, Hood Thabit, Malgorzata E. Wilinska, and Roman Hovorka. Modelling endogenous insulin concentration in type 2 diabetes during closed-loop insulin delivery. *BioMedical Engineering OnLine*, 14(1):19, March 2015. ISSN 1475-925X. doi: 10.1186/s12938-015-0009-5. URL https://doi.org/10.1186/s12938-015-0009-5.

H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, February 1978. ISSN 0096-3518. doi: 10.1109/TASSP.1978.1163055. Conference Name: IEEE Transactions on Acoustics, Speech, and Signal Processing.

F John Service, George D Molnar, John W Rosevear, Eugene Ackerman, Lael C Gatewood, and William F Taylor. Mean Amplitude of Glycemic Excursions, a Measure of Diabetic Instability. *Diabetes*, 19(9):644–655, September 1970. ISSN 0012-1797. doi: 10.2337/diab.19.9.644. URL https://doi.org/10.2337/diab.19.9.644.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning Structured Output Representation using Deep Conditional Generative Models. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://papers.nips.cc/paper_files/paper/2015/hash/8d55a249e6baa5c06772297520da2051-Abstract.html.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. ISSN 1533-7928. URL http://jmlr.org/papers/v15/srivastava14a.html.

Seigo Sugiyama, Hideaki Jinnouchi, Kunio Hieshima, Noboru Kurinami, Katsunori Jinnouchi, Akira Yoshida, Tomoko Suzuki, Keizo Kajiwara, Fumio Miyamoto, and Tomio Jinnouchi. Potential Identification of Type 2 Diabetes with Elevated Insulin Clearance. *NEJM Evidence*, 1 (4):EVIDoa2100052, March 2022. doi: 10.1056/EVIDoa2100052. URL https://evidence.nejm.org/doi/full/10.1056/EVIDoa2100052. Publisher: Massachusetts Medical Society.

Naoya Takeishi and Alexandros Kalousis. Physics-Integrated Variational Autoencoders for Robust and Interpretable Generative Modeling. May 2021. URL https://openreview.net/forum?id=0p0gt1Pn2Gv.

Romain Tavenard, Johann Faouzi, Gilles Vandewiele, Felix Divo, Guillaume Androz, Chester Holtz, Marie Payne, Roman Yurchak, Marc Rußwurm, Kushal Kolar, and Eli Woods. Tslearn, a machine learning toolkit for time series data. *Journal of Machine Learning Research*, 21(118):1–6, 2020. URL http://jmlr.org/papers/v21/20-091.html.

Michalis Titsias and Miguel Lázaro-Gredilla. Doubly Stochastic Variational Bayes for non-Conjugate Inference. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1971–1979. PMLR, June 2014. URL https://proceedings.mlr.press/v32/titsias14.html. ISSN: 1938-7228.

Guido Van Rossum and Fred L Drake Jr. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

J. M. Varah. A Spline Least Squares Method for Numerical Parameter Estimation in Differential Equations. *SIAM Journal on Scientific and Statistical Computing*, 3(1):28–46, March 1982. ISSN 0196-5204. doi: 10.1137/0903003. URL https:

//epubs.siam.org/doi/10.1137/0903003. Publisher: Society for Industrial and Applied Mathematics.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Ke Alexander Wang, Matthew E Levine, Jiaxin Shi, and Emily Fox. Learning absorption rates in glucose-insulin dynamics from meal covariates. In *NeurIPS 2022 workshop on learning from time series for health*, 2022. URL https://openreview.net/forum?id=cbn7xvCCq6e.

Antoine Wehenkel, Jens Behrmann, Hsiang Hsu, Guillermo Sapiro, Gilles Louppe, and Joern-Henrik Jacobsen. Robust Hybrid Learning With Expert Augmentation. *Transactions on Machine Learning Research*, October 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=oe4dl4MCGY.

Darrell M. Wilson, Susan L. Pietropaolo, Maria Acevedo-Calado, Shuai Huang, Destiny Anyaiwe, David Scheinker, Andrea K. Steck, Madhuri M. Vasudevan, Siripoom V. McKay, Jennifer L. Sherr, Kevan C. Herold, Jessica L. Dunne, Carla J. Greenbaum, Sandra M. Lord, Michael J. Haller, Desmond A. Schatz, Mark A. Atkinson, Patrick W. Nelson, Massimo Pietropaolo, and the Type 1 Diabetes TrialNet Study Group. CGM Metrics Identify Dysglycemic States in Participants From the TrialNet Pathway to Prevention Study. *Diabetes Care*, 46(3):526–534, February 2023. ISSN 0149-5992. doi: 10.2337/dc22-1297. URL https://doi.org/10.2337/dc22-1297.

J. M. Wójcicki. "J"-Index. A New Proposition of the Assessment of Current Glucose Control in Diabetic Patients. *Hormone and Metabolic Research*, 27(1):41–42, January 1995. ISSN 0018-5043, 1439-4286. doi: 10.1055/s-2007-979906. URL http://www.thieme-connect.de/DOI/DOI?10.1055/s-2007-979906. Publisher: © Georg Thieme Verlag Stuttgart · New York.

Yuan Yin, Vincent Le Guen, Jérémie Dona, Emmanuel de Bezenac, Ibrahim Ayed, Nicolas Thome, and patrick gallinari. Augmenting Physical Models with Deep Networks for Complex Dynamics Forecasting. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=kmG8vRXTFv.

Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. TS2Vec: Towards Universal Representation of Time Series. *arXiv:2106.10466 [cs]*, September 2021. URL http://arxiv.org/abs/2106.10466. arXiv: 2106.10466.

David Zeevi, Tal Korem, Niv Zmora, David Israeli, Daphna Rothschild, Adina Weinberger, Orly Ben-Yacov, Dar Lador, Tali Avnit-Sagi, Maya Lotan-Pompan, Jotham Suez, Jemal Ali Mahdi, Elad Matot, Gal Malka, Noa Kosower, Michal Rein, Gili Zilberman-Schapira, Lenka Dohnalová, Meirav Pevsner-Fischer, Rony Bikovsky, Zamir Halpern, Eran Elinav, and Eran Segal. Personalized Nutrition by Prediction of Glycemic Responses. *Cell*, 163(5):1079–1094, November 2015. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2015.11.001. URL https://www.cell.com/cell/abstract/S0092-8674(15)01481-6. Publisher: Elsevier.

Appendix A. Initializing and fixing the expert prior

We found that it was important to initialize our prior distribution $p(\mathbf{z})$ to physiologically plausible values and kept fix during training. This is consistent with the traditional approach of specifying an expert-informed Bayesian prior over the latent variables.

For the prior over $\mathbf{u}' = \sigma_{\mathbf{u}}^{-1}(\mathbf{u})$, we set the mean to 0 and standard deviation to 10 for each timestep. For the prior over $\mathbf{x}_0' = \sigma_{\mathbf{x}_0}^{-1}(\mathbf{x}_0)$ where $\mathbf{x}_0 = [G(0), X(0), G_1(0), G_2(0)]$, we set the mean to be result of mapping $[30, 120, 10^{-2}, 5 \times 10^{-4}, 1/30, 1]$

to the unconstrained space; we set the standard deviation to be the result of mapping $[2, 1, 1, 1, 1, 1]$ to the unconstrained space. For the prior over $\mathbf{w}' = \sigma_{\mathbf{w}}^{-1}(\mathbf{w})$ where $\mathbf{w} = [\tau_m, G_b, S_G, p_2, S_I, M_I]$, we set the mean to the result of mapping $[120, 0.1, 0.1, 20]$ to the unconstrained space and the standard deviatino to the result of mapping $[1, 1, 1, 2]$ to the unconstrained space.

Appendix B. Parameter and state constraints

Since our hybrid VAE relies on a mechanistic ODE decoder, the initial state and the parameters of passed into the decoder must be physiologically-plausible. These constraints regularize the mechanistic ODE to be more numerically stable, more identifiable, and ensure that the learned latent space is physiologically interpretable and comparable to real values.

We use the following constraints on the initial state $\mathbf{x}_0 = [G(0), X(0), G_1(0), G_2(0)]$:

- $G(0) \in [50, 300]$ mg/dL
- $X(0) \in [0, 1]$ min^{-1}
- $G_1(0) \in [0, 1]$ mg/dL
- $G_2(0) \in [0, 100]$ mg/dL.

We use the following constraints on each ODE parameter $\mathbf{w} = [\tau_m, G_b, S_G, p_2, S_I, M_I]$:

- $\tau_m \in [10, 60]$ min
- $G_b \in [80, 200]$ mg/dL
- $S_G \in [5 \times 10^{-3}, 2 \times 10^{-2}]$ min^{-1}
- $S_I \in [10^{-4}, 10^{-3}]$ (L/mU)/min
- $p_2 \in [1/60, 1/15]$ min^{-1}
- $M_I \in [0.1, 3.0]$ (mU/L)/(mg/dL).

We constrain the carb amount u_t to be $[0, 1000]$ mg/min.