ORIGINAL ARTICLE

A physiologically-constrained neural network digital twin framework for replicating glucose dynamics in type 1 diabetes

Valentina Roquemen-Echeverri · Taisa Kushner · Peter G. Jacobs · Clara Mosquera-Lopez

Received: 25 July 2025 / Accepted: 4 February 2026

Abstract

Simulating glucose dynamics in individuals with type 1 diabetes (T1D) is critical for developing personalized treatment strategies and supporting informed, data-driven clinical decisions. Current modeling approaches often fail to capture all physiological aspects influencing glucose dynamics and are difficult to adapt to individuals. In this work, we present physiologically-constrained neural network (NN) digital twins for simulating glucose dynamics in T1D. To ensure that the digital twins are interpretable and consistent with known physiology, we first develop a population-level NN state-space model designed to align with a set of ordinary differential equations (ODEs) describing the glucoregulatory system in T1D. This model is formally verified to conform to established glucose regulation dynamics. Individual-specific glucose dynamics (i.e., digital twins) are then achieved by combining the population-level model with individual-level models that enhance predictions through the integration of personal glucose management and contextual data, capturing both inter- and intra-individual variability. We validated our approach using participants from the T1D Exercise Initiative (T1DEXI) study by simulating the real-world scenarios observed during the study. Two weeks of data were segmented into 5-h sequences, and we compared the simulated versus actual glucose profiles. Similarity was assessed using clinically relevant glucose outcomes, and paired equivalence t-tests were conducted with predefined margins based on clinical significance. Across 394 digital twins, we observed equivalent glucose outcomes between the simulated and real-world scenarios. Time in range (70–180 mg/dL) 75.1(95% CI 69.0–80.9)% for simulations and 74.4(95% CI 69.9–78.6)% for real-world data (P-value = < 0.001); time below range (< 70 mg/dL) was 2.5(95% CI 1.2–4.1)% versus 3.0(95% CI 2.2–4.0)% (P-value = 0.022), and time above range (> 180 mg/dL) was 22.4(95% CI 16.6–28.8)% versus 22.6(95% CI 18.4–27.3)% (P-value = < 0.001). Our proposed framework can incorporate unmodeled features, such as sleep and physical activity, while preserving key physiological dynamics. Results show that including these additional inputs leads to more accurate simulations than omitting them entirely. This physiologically-informed framework enables personalized *in-silico* testing of T1D treatment strategies, supports model-based insulin optimization, and integrates physics-based constraints with data-driven learning. Code and models are publicly available at: https://github.com/mosqueralopez/T1DSim_AI.

Keywords Artificial intelligence · Digital twin · Glucose regulation · Hybrid modeling · Neural network state-space model · Physiologically-constrained neural network · Type 1 diabetes

1 Introduction

Type 1 diabetes (T1D) is a chronic condition characterized by elevated glucose levels, resulting from the inability of the pancreas to produce enough insulin [1]. Thus, people with T1D must take exogenous insulin to enable their bodies to metabolize glucose. Achieving optimal glycemic control in T1D remains challenging as insulin therapy needs to be adjusted over time based on each individual's glucose response to meals, exercise, hormone cycles, stress, and other external disturbances [2–4]. This burden is increasingly alleviated through advancements in diabetes technology, such as accurate, nonadjunctive continuous glucose monitoring (CGM) [5], connected insulin pens and pumps, and automated insulin delivery (AID) and decision support systems (DSS) [6–8]. AID systems in particular have demonstrated significant improvements in glucose outcomes, increasing glucose time in range (TIR, 70–180 mg/dL) and reducing both time above range (TAR, >180 mg/dL) and time below range (TBR, < 70 mg/dL) [9].

Central to the development of AID and DSS have been virtual patient models, which are used as predictive models as well as the basis of *in-silico* simulation environments used for pre-clinical validation of the systems' efficacy and safety. However, approximating the human glucoregulatory system remains difficult and most virtual patient models are lacking the individual-specific adjustments needed to ensure that *in-silico* results translate to real-world individuals [10]. Digital twins build on this concept by creating personalized, data-driven models of individual patients, combining mechanistic knowledge with patient-specific data to better reflect their physiology and predict responses to therapy [11]. By tailoring the virtual representation to each patient, digital twins aim to overcome the limitations of current virtual patient models and support more precise prediction and optimization of treatment. Furthermore, digital twins are increasingly being applied across a wide range of healthcare domains, including drug discovery and development [12–14], disease diagnosis [15], preventive care [16], therapy optimization [17], personalized medicine [18], clinical research [19], education [20], and clinical decision support [21–23], highlighting their growing relevance and versatility in biomedical applications.

Herein, we provide a novel modeling approach for constructing physiologically-constrained neural network (NN)-based digital twin models architected based on a mechanistic model, enabling conformance of the NN model with known physiological constraints. We find these models are more accurate than comparator ordinary differential equations (ODE)-based mechanistic models in replicating real-world scenarios. Key contributions include:

1. *Novel model architecture*: A population-level NN state-space model architected to be consistent with the ODEs describing the glucoregulatory system of individuals with T1D. This model is interpretable such that its consistency with physiology can be observed and formally verified (Sect. 3.2).
2. *Formal verification of dynamics*: A methodology for verifying whether each sub-network of the population-level model replicates the dynamics present in the associated compartmentalized mechanistic model (Sect. 3.3).
3. *Digital twins*: A methodology for creating digital twins by combining the population-level model with individual-level models that augment predictions through integration of additional glucose management and contextual data to model inter- and intra-individual variability in glucose dynamics. By learning individual-level residuals directly from the rate of glucose change over time and dynamically adjusting glucose values at each simulation step, rather than learning the error for a given glucose prediction, our method ensures that the residuals remain independent of factors such as elapsed simulation time. We also created a large virtual population validated using real-world data from the T1D exercise initiative (T1DEXI) study (Sect. 3.4).

We organize this paper as follows: Sect. 2 reviews the related work to contextualize the relevance of our study. Section 3 describes our modeling and conformance verification approach, as well as the methodology to build

physiologically-constrained NN digital twins. Section 4 reports simulation accuracy results. In Sect. 5, we discuss our most relevant findings and the strengths and limitations of our work. Section 6 concludes the paper.

2 Related work

Most existing T1D metabolic simulators are based on mechanistic models governed by ODEs that model various aspects of glucose metabolism such as carbohydrate intake and absorption, physical activity, insulin kinetics, and glucose-insulin dynamics [19, 24–31]. While these models provide an approximation of the human glucoregulatory system, they do not capture all physiological aspects impacting glucose dynamics and are difficult to fit to specific individuals. Identifying a personalized model that best matches a real-world person is oftentimes referred to identifying a digital twin for that person [6]. There are a number of approaches for linking ODE models to individual data that have been described previously. Young et al. in [18] proposed an event-based approach to match a digital twin from a virtual population to real-world glucose traces before an exercise event based on insulin sensitivity and body weight [32]. The selected digital twin was used to simulate multiple interventions and select optimal recommendations specifically for enabling safe exercise in T1D. Cappon et al. [31] developed a two-stage digital twin-based simulation methodology that leverages observed glucose management data to identify a personalized model to approximate an individual's postprandial glucose response using Markov Chain Monte Carlo (MCMC). Then, the personalized model was used to simulate alternative insulin and carbohydrate therapies and evaluate their impact on glucose outcomes. MCMC is a robust approach to do model identification; however, it is time consuming and computationally expensive. Moreover, a more comprehensive evaluation of the proposed methodology on real-world data is lacking.

As an alternative to using ODE models to represent digital twins, a variety of groups have proposed data-driven methods for constructing digital twins. These data-driven approaches utilize artificial intelligence (AI) models, which can learn complex patterns from large datasets. However, these models are often not interpretable and risk mistaking patterns in data for causal relationships, unless designed with constraints [10, 33–35].

In recent years, there has been increasing interest in combining mechanistic or physics-based and data-driven AI-based modeling techniques. The domain knowledge built into mechanistic models is leveraged to inform the architectural design of AI models and to guide the learning process using aggregate supervision and constraints [36, 37]. This hybrid modeling approach, also known as physics-informed machine learning, yields models that can accurately capture complex patterns from the data while adhering to physics principles. Hybrid modeling has been applied in various fields such as geophysics [38], epidemiology [39], fluid dynamics [40], and more recently in a range of biomedical applications [41].

Hybrid approaches have been proposed for glucose prediction [42–45], most of which predict future glucose over a fixed prediction horizon rather than being designed to simulate the dynamics of the glucoregulatory system states over time given simulation scenarios. These methods typically require historical data to contextualize predictions, and the hybrid component comes from using compartmental models to preprocess insulin kinetics and carbohydrate absorption before learning a function to predict glucose levels.

Table 1 provides a detailed comparison of our approach with existing digital twin frameworks, highlighting the advantages of integrating data-driven flexibility with mechanistic constraints. Overall, our work leverages NNs to enable adaptability to capture complex patterns from real-world data while adhering physiological principles encoded through ODE-based models.

Table 1 Comparison of the proposed framework with existing digital twin approaches, highlighting architecture, learning process, inputs, code availability and data used for parameter identifications

Digital twins framework for T1D		
	Our approach	Other approaches
Architecture	First methodology using a neural network state space model architected based on an ODE-based mechanistic model	None are based on neural networks [25, 26, 31, 46–50]
Model fitting	Gradient descent	Markov chain Monte Carlo [25, 26, 31]; Latin-Hypercube Sampling [46]; Parameters obtained from literature [47]; Gradient-based method [48]; Least-square Fitting [49]; Bayesian Maximum a Posteriori [50]
Simulations	Digital twins can simulate real-world scenarios with no need for segmentation of events such as meals or exercise	Digital twins are created based on specific events such exercise [25], or constrained meal scenarios [26, 46, 47, 50].
Code availability	Our framework code is available on GitHub	Only [31, 47] have public code repositories
Inputs	Insulin, carbohydrate intake, heart rate, sleep data, and time-based contextual features	Usually include only insulin and carbohydrate intake [26, 31, 46, 48–50], only a few have included heart rate information [25, 47]
Flexibility	The framework is capable of (1) incorporating new inputs that may affect glucose levels, (2) modifying the loss function to adapt to individual needs (e.g., high glucose variability), or (3) computationally efficient continuous training as more data becomes available	No other digital twin approach offers this level of flexibility. Some existing methods are constrained to be fitted to each event, making them computationally intensive and impractical for real-time applications [31]

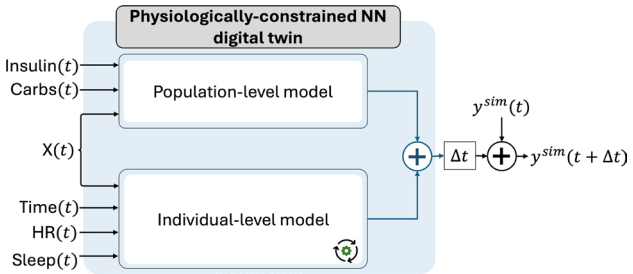

Fig. 1 Overview of the physiologically-constrained NN digital twin framework for simulating glucose dynamics in type 1 diabetes. A digital twin consists of a population-level neural network state-space model and an adaptive individual-level neural network that models residual dynamics. X(t), HR(t), and $y^{sim}(t)$ represent the system's states, heart rate and glucose level, respectively, at time t

3 Methods

The process for developing physiologically-constrained NN digital twins involves 3 steps: (1) developing a population-level NN state-space model of the glucose-insulin dynamics, (2) personalizing the population-level model for each individual; and (3) assessing their simulation accuracy. Our proposed framework for constructing digital twins is depicted in Fig. 1.

3.1 Datasets

We used both simulated and real-world data to develop and test the physiologically-constrained NN digital twins. We used an open source metabolic simulator [19] that enabled us to (1) observe all model states at any given time, (2) create a large and diverse dataset to capture the glucose dynamics under various real-world meal and insulin dosing scenarios, and (3) avoid potential errors found in free-living datasets, such as those related to self-reported events. Simulated data was used to develop the population-level model. Then, we used a real-world dataset for individual-specific augmentation of the population-level model (i.e., the creation of the digital twins).

To train the population-level model of the glucoregulatory system, we simulated glucose management data using a validated single-hormone simulator developed by Resalat et al. [19] referred as $T1DSim_{ODE}$ (code available at https://github.com/petejacobs/T1D_VPP).

The real-world data came from the T1DEXI study, a large observational study that collected free-living glucose management data from 497 physically individuals with T1D across the United States (mean age 37 ± 14 years; HbA1c 6.6±0.8% [89±157 mg/dL]) using multiple daily injections (MDI), standard insulin pump, or AID therapies. The aim of the T1DEXI study was to investigate the effects of different types of exercise (i.e., aerobic, interval, and resistance) on glucose outcomes. During four weeks, study participants wore a CGM and a fitness tracker, and used an insulin delivery device and a smartphone-based app to log food intake, glucose management data and relevant life events. An Institutional Review Board approved the T1DEXI Study and electronic informed consent was obtained from each participant (dataset available for research purposes at https://doi.org/10.25934/PR00008428). The study had 2 phases: the initial pilot data collection [51] followed by the subsequent main data collection [52]. We used the daily meal scenarios from the T1DEXI pilot phase to generate simulated data, and data from the T1DEXI main phase data to create individual-level models. As MDI data was self-reported, to ensure correct model identification and validation, we limit our analysis to participants using either open loop insulin pumps or AID systems. Specifics of this subset are provided in Appendix A.

3.2 Population-level model

To capture complex temporal relationships, even when using shallow NNs, we designed a NN state-space model based on the physiology knowledge embedded into the $T1DSim_{ODE}$ simulator, an open-source simulator developed by our group that was successfully validated against a clinical dataset [19]. This simulator is governed by a glucoregulatory model that consists of eight ODEs describing insulin kinetics (Eqs. 1a–1c), insulin dynamics (Eqs. 1d–1f), and glucose kinetics (Eqs. 1g and 1h). In addition to capturing complex temporal relationships, the NN state-space model design allows us to verify physiological conformance by adhering to the foundational dynamics of the $T1DSim_{ODE}$. This alignment allows the model to maintain consistency with real-world physiological behavior while accommodating variability in compartments where the dynamics are less certain. Furthermore, this population-level model enables us to track the evolution of the system's states over time starting from any given initial state that encodes the history of the system, similar to conventional ODE-based models.

$$\dot{S}_1(t) = u_I(t) - \frac{S_1(t)}{t_{max}} \tag{1a}$$

$$\dot{S}_2(t) = \frac{S_1(t)}{t_{max}} - \frac{S_2(t)}{t_{max}} \tag{1b}$$

$$\dot{I}(t) = \frac{S_2(t)}{t_{max} V_I} - k_e I(t) \tag{1c}$$

$$\dot{X}_1(t) = -k_{a1} X_1(t) + S_{f1} k_{a1} I(t) \tag{1d}$$

$$\dot{X}_2(t) = -k_{a2} X_2(t) + S_{f2} k_{a2} I(t) \tag{1e}$$

$$\dot{X}_3(t) = -k_{a3} X_3(t) + S_{f3} k_{a3} I(t) \tag{1f}$$

$$\begin{aligned} \dot{Q}_1(t) = &-X_1(t) Q_1(t) - F_{01}^c - F_R + k_{12} Q_2(t) \\ &+ U_G(t) + EGP_0(1 - X_3(t)) \end{aligned} \tag{1g}$$

$$\dot{Q}_2(t) = X_1(t) Q_1(t) - k_{12} Q_2(t) - X_2(t) Q_2(t) \tag{1h}$$

$S1$ and $S2$ represent the masses of insulin [mU/kg] in the subcutaneous compartment and an unobservable compartment, respectively. u_I represents the rate of insulin infusion into the subcutaneous space [mU/kg/min]. I represents the plasma insulin concentration [mU/L], and t_{max}, V_I and k_e are the time-to-maximum absorption [min], distribution volume [L/kg] and elimination rate [min^{-1}] of insulin, respectively.

$X_1[\mathrm{min}^{-1}]$, X_2 [min^{-1}] and X_3 [unitless] represent the effect of insulin on glucose distribution, disposal, and suppression of endogenous glucose production (*EGP*), respectively. $S_{f1}[(\mathrm{mU}\cdot\mathrm{L}\cdot\mathrm{min})^{-2}]$, $S_{f2}[(\mathrm{mU}\cdot\mathrm{L}\cdot\mathrm{min})^{-2}]$ and $S_{f3}[(\mathrm{mU}\cdot\mathrm{L}\cdot\mathrm{min})^{-1}]$ are the insulin sensitivity factors, and are the parameters used for representing inter-individual insulin sensitivity variability of the glucoregulatory system in T1D.

Q_1 and Q_2 are the masses of glucose in the accessible (i.e., plasma) and non-accessible (i.e., peripheral tissue) compartments given in [mmol/kg], respectively. EGP_0 is the basal endogenous glucose production at a theoretical zero insulin concentration [(mmol/kg)/min]. F_{01}^c and F_R are the non-insulin mediated glucose uptake and the renal glucose clearance rate, respectively [mmol/kg/min]. U_G represents the glucose absorption rate from meals [mmol/kg/min] (Eq. 2).

$$U_G(t) = \frac{D_G A_G (t - t_0) e^{-\frac{t-t_0}{t_{max,G}}}}{t_{max,G}^2} \tag{2}$$

In Eq. (2), $t_{max,G}$ is the time-to-maximum appearance rate of glucose in Q_1 [min], A_G is the carbohydrate bioavailability [unitless], t_0 is the meal onset time [min] and D_G is the estimated carbohydrate intake [mmol/kg]. For *in-silico* simulations, D_G is converted from grams to mmol/kg to be compatible with the variables of the glucose kinetics model (Eq. 1g).

The population-level model called $T1DSim_{NN}^{P}$ is a 10-dimensional NN state-space model with the general form given in Eq. (3).

$$\dot{\mathrm{x}}(t) = \mathrm{N}(\mathrm{x}(t), \mathrm{u}(t); \Theta^P) \tag{3a}$$

$$y(t) = Q_1(t) \tag{3b}$$

In Eq. (3), $\mathrm{x}(t) = \{x^i(t)\} = \{S_1(t), S_2(t), I(t), X_1(t), X_2(t), X_3(t), Q_1(t), Q_2(t), C_1(t), C_2(t)\}$ is the set of system's states at time *t*; $\mathrm{u}(t) = \{u_I(t), u_{carb}(t)\}$ is the set of system's inputs, $u_I(t)$ is the insulin infusion rate [U/h] and $u_{carb}(t)$ is the carbohydrate intake [g] at time *t*; and $\mathrm{N} = \{\mathbb{N}_{fi}\}$ is a set of fully connected NNs parameterized by $\mathbf{\Theta^P} = \{\theta_i^P\}$, with i = 1,2,...,10.

The architecture of the $T1DSim_{NN}^{P}$ is described by Eq. (4). Note that the $T1DSim_{NN}^{P}$ includes the C_1 and C_2 compartments to explicitly model carbohydrate absorption following the model structure described by Hovorka et al. in [53]. This definition allows the data driven model states to represent the ODE model states in the glucoregulatory model including the compartment states, metabolic fluxes, and dependencies among the system's inputs and states. This approach increases the number of parameters compared to the $T1DSim_{ODE}$, which may lead to longer inference time. However, this trade-off provides the advantage of not only achieving a physiologically meaningful representation of the system but also establishing a framework for future extensions.

$$\dot{S}_1(t) = \mathbb{N}_{f1}(S_1(t), u_I(t)) \tag{4a}$$

$$\dot{S}_2(t) = \mathbb{N}_{f2}(S_1(t), S_2(t)) \tag{4b}$$

$$\dot{I}(t) = \mathbb{N}_{f3}(S_2(t), I(t)) \tag{4c}$$

$$\dot{X}_1(t) = \mathbb{N}_{f4}(X_1(t), I(t)) \tag{4d}$$

$$\dot{X}_2(t) = \mathbb{N}_{f5}(X_2(t), I(t)) \tag{4e}$$

$$\dot{X}_3(t) = \mathbb{N}_{f6}(X_3(t), I(t)) \tag{4f}$$

$$\dot{Q}_1(t) = \mathbb{N}_{f7}(X_1(t), X_3(t), Q_1(t), Q_2(t), C_2(t)) \tag{4g}$$

$$\dot{Q}_2(t) = \mathbb{N}_{f8}(X_1(t), X_2(t), Q_1(t), Q_2(t)) \tag{4h}$$

$$\dot{C}_1(t) = \mathbb{N}_{f9}(C_1(t), u_{carbs}(t)) \tag{4i}$$

$$\dot{C}_2(t) = \mathbb{N}_{f10}(C_1(t), C_2(t)) \tag{4j}$$

3.2.1 Development dataset generation

To develop a diverse dataset, we conducted 7-day simulations using a variety of real-world meals, insulin dosing, and initial glucose conditions. Simulations were started at midnight using the $T1DSim_{ODE}$ with population-level parameters as presented in [19], with details on scenarios described below:

- *Meal scenarios*: Daily meal scenarios were obtained from the pilot phase of the T1DEXI study [51]. Data was pre-processed to remove unrealistic scenarios caused by errors in self-reported data by selecting those that were within the 5^{th} and 95^{th} percentiles of the distribution of number of meals per day (1-9) and total carbohydrate intake per day (30-359 g). For any given 7-day simulation, 7 daily meal scenarios were selected.
- *Insulin dosing*: The insulin-to-carb ratio is simulated using the "1700 rule" of Davidson et al. [54]. Multiple bolus timing scenarios were utilized including insulin given at meal onset, and delayed by 5 to 45 min (with 5-min incremental steps). In addition to varying bolus timing, bolus size was varied to include under- and over-dosing. For all simulations, basal insulin was delivered every 5 min.
- *Initial glucose values*: For each 7-day meal scenario, five different initial glucose values were drawn from a normal distribution, with the mean and standard deviation calculated using CGM values at midnight from the participant corresponding to each meal scenario. The mean of the midnight CGM was 156±45 mg/dL.

The final *in-silico* development dataset $\mathcal{D}^P = \{\mathrm{x}(t), \mathrm{u}(t)\}$, reflects 323,400 days of simulated glucose management data divided into 7-day simulation scenarios. $\mathcal{D}^P$ was split into $\mathcal{D}^P_{train}$ (60%), $\mathcal{D}^P_{validation}$ (20%), $\mathcal{D}^P_{test}$ (20%) subsets for model training, validation, and testing, respectively. The split was performed randomly, but with controls to ensure the same daily meal scenario appeared in no more than one subset, ensuring independence and avoiding data leakage.

To improve data balance across rare events, we included additional traces with glucose less than 70 mg/dL, as well as those with delayed meal boluses to better represent glucose levels above 250 mg/dL. Specifically we supplement scenarios with TBR greater than 20% or percentage time above 250 mg/dL greater than 40%. Additionally, robust scaling was utilized to reduce the effect of outliers on the model training (see Appendix B for more details).

3.2.2 Model training

$T1DSim^P_{NN}$ was trained following the truncated simulation error minimization methodology described by Forgione and Piga in [55]. Given $\mathcal{D}^P_{train}$, we created batches containing B^P sequences of size M with 75% overlap between consecutive sequences. We used batch training meaning that model parameters were updated iteratively using the average loss calculated after processing a given batch to find the optimal set of parameters Θ^P that minimized the loss function defined in Eq. (5).

$$\mathcal{L}^P_{total} = \mathcal{L}^P_{fit} + \alpha \mathcal{L}^P_{consistency} \tag{5}$$

In Eq. (5), $\mathcal{L}^P_{fit}$ is the loss associated with the measured glucose compartment Q_1, $\mathcal{L}^P_{consistency}$ is the loss associated with all system's states, and $\alpha \geq 0$ is a regularization parameter selected to balance $\mathcal{L}_{fit}$ and $\mathcal{L}_{consistency}$. The $\mathcal{L}^P_{fit}$ is defined as

$$\mathcal{L}^P_{fit} = \frac{1}{B^P M} \sum_{b=0}^{B^P-1} \sum_{m=0}^{M-1} (y^{sim}_{b,m} - y_{b,m})^2 \mathcal{P}(y^{sim}_{b,m}, y_{b,m}), \tag{6}$$

where B^P is the batch size, M is the training sequence length, y^{sim} is the simulated Q_1 sequence, and y is the actual Q_1 sequence. $\mathcal{P}(y^{sim}, y)$ is a penalty function defined in Eq. (7) according to the penalty proposed by Del Favero et al. in [56]. $\mathcal{P}(y^{sim}, y)$ was designed to penalize the errors made by the model on infrequent yet clinically significant events, such as hypoglycemia (glucose < 70 mg/dL) and extreme hyperglycemia (glucose >250 mg/dL); both of which pose challenges in glucose management for individuals with T1D [57, 58]. The penalty values in $\mathcal{P}(y^{sim}, y)$ were determined based on the ratio of the glucose risk function $\mathcal{R}_{BG}(g) = 22.77(ln(g)^{1.084} - 5.381)^2$ defined by Kovatchev et al. [59] at glucose levels $g = 70$ mg/dL ($\mathcal{R}_{BG}(70) = 7.8$) and $g = 250$ mg/dL ($\mathcal{R}_{BG}(250) = 22.4$), which is about 3. This selection ensures that the errors are appropriately penalized based on the risk of over- or under-predicting glucose in critical ranges.

$$\mathcal{P}(y^{sim}, y) = \begin{cases} 2 & y \leq 70 \wedge y^{sim} > y \\ 6 & y \geq 250 \wedge y^{sim} < y \\ 1 & otherwise \end{cases} \tag{7}$$

The consistency loss $\mathcal{L}^P_{consistency}$ is as a weighted sum of the errors calculated for each of the S compartments in the NN state-space model, and it is defined as

$$\mathcal{L}^P_{consistency} = \sum_{i=0}^{S-1} \omega_i \mathcal{L}^{P,i}_{consistency}, \tag{8}$$

where $S = 10$ is the number of compartments in the model, w_i is the weight associated with compartment i, with $\sum_{i=0}^{S-1} w_i = 1$. $\mathcal{L}^{P,i}_{consistency}$ is the loss associated with the i^{th} compartment in the state-space model, and it is defined in Eq. (9).

$$\begin{aligned} \mathcal{L}^{P,i}_{consistency} = & \frac{1}{B^P M} \sum_{b=0}^{B^P-1} \sum_{m=0}^{M-1} (x^{i,sim}_{b,m} - x^i_{b,m})^2 \\ & + \beta \sum_{b=0}^{B^P-1} \sum_{m=0}^{M-1} max[0, -(x^{i,sim}_{b,m} - min(x^i)] \end{aligned} \tag{9}$$

In Eq. (9), the first component of the consistency loss is the mean squared error (MSE) between $x^{sim,i}$, the simulated i^{th} state, and the actual state x^i. The second component is a soft constraint regularized by the parameter $\beta \geq 0$ added to prevent the states taking values that are less than those observed in the training set. Note that the $MSE^i = 0$ and $min(x^i) = 0$ for $i = 9$ corresponding to the compartment $x^9 = C_1$ because there is no ground truth data for this compartment, which we introduced to explicitly model the delay in carbohydrate absorption. However, by making $min(x^i) = 0$, we are including a penalty to prevent C_1 to take negative values, which is not physiologically possible as it represents a negative meal amount. This is aligned with the interpretation of the C_1 compartment as the amount of carbohydrates on board, a quantity inherently non-negative.

3.2.3 Network architecture

For this work, we selected the simplest architecture for each sub-network $\mathbb{N}_{fi}$ which corresponds to a NN of one hidden layer with a rectified linear unit (ReLU) activation function. The batch size was $B^P = 128$, sequence length $M = 5$ h with sampling period $\Delta m = \Delta t = 5$ min, regularization constants $\alpha = 0.7$ and $\beta = 0.08$, state error weights $w^i = 0.08\overline{3}$ i = 1,2,3,4,5,6,10, $w^7 = 0.208\overline{3}$, $w^8 = 0.1\overline{6}$, $w^9 = 0.041\overline{6}$, starting learning rate $\lambda^P = 10^{-3}$ scheduled to be reduced by $e^{-0.1}$ every epoch. The training process for the $T1DSim^P_{NN}$ model is detailed in the Algorithm 1. We used the training dataset $\mathcal{D}^P_{train}$ and gradient-based optimization using the Adam optimizer with recommended default parameters [60].

The optimal number of neurons per sub-network hidden layer, $\mathbb{N}_{fi}$, was found using Bayesian optimization [61], from an initial wide range of 4 to 256 neurons. The optimization target was to minimize, over 20 epochs, the balanced simulation error in Eq. (10) on $\mathcal{D}^P_{validation}$, and this was performed with 50 exploration and 50 exploitation iterations. Our $\mathcal{D}^P_{validation}$ was split into sequences of size $M = 5$ h with no overlap.

$$\mathcal{L}^P_{BayesOpt} = \frac{1}{3}\sum_{j=1}^{3}\sqrt{\frac{1}{M}\sum_{m=0}^{M-1}(y^{sim}_{j,m} - y_{j,m})^2} \tag{10}$$

In Eq. (10), $j = 1, 2, 3$ are the groups of sequences in $\mathcal{D}^P_{validation}$ with the following characteristics:

- *Group 1*: Sequences with TIR greater than 70% and TBR less than 20%.
- *Group 2*: Sequences with TBR greater than 20%.
- *Group 3*: The remaining sequences not included in Group 1 or Group 2.

Algorithm 1 $T1DSim^P_{AI}$ training process

Require: Training dataset $\mathcal{D}^P_{train}$ split into $n_{sequences}$ of size M, number of epochs n_{epochs}, batch size B^P, learning rate λ^P
Ensure: Optimized $T1DSim^P_{AI}$ parameters Θ^P

initialize the $T1DSim^P_{AI}$ parameters Θ^P
set number of iterations
  $n_{iterations} \leftarrow \frac{n_{epochs} n_{sequences}}{B^P}$
for $j \leftarrow 0$ to $n_{iterations} - 1$
  select B^P M-sized sequences to form a training batch
  simulate each training sequence in the batch
    Euler's ODE integration method
    Initial states: $\mathbf{x}^0_{seq}$
    System's inputs: $\mathbf{u}_{seq}(t)$
    $\mathbf{x}^{sim}_{seq}(0) \leftarrow \mathbf{x}^0_{seq}$
    for $m \leftarrow 0$ to $M - 1$
      $\dot{\mathbf{x}}^{sim}_{seq}(m) \leftarrow \mathbb{N}(\mathbf{x}^{sim}_{seq}(m), \mathbf{u}(m); \Theta^P)$
      $\mathbf{x}^{sim}_{seq}(m + \Delta m) \leftarrow \mathbf{x}^{sim}_{seq}(m) + \dot{\mathbf{x}}^{sim}_{seq}(m)\Delta m$
    end for
  compute $\mathcal{L}^P_{total}$ using Equation 5
  evaluate the gradients $\nabla_{\Theta^P}\mathcal{L}_{total}$
  update Θ^P
    $\Theta^P \leftarrow \Theta^P - \lambda^P \nabla_{\Theta^P}\mathcal{L}^P_{total}$
end for

3.3 Conformance verification for ensuring neural network models match known physics

In order to verify that the dynamics of the learned models match known physiology, we leveraged our previously developed approach of verifying δ-monotonicity through range estimation [33] (Algorithm 2). This approach enables us to determine whether or not each sub-network $\mathbb{N}_{fi}$ defined in Eq. (4) replicate the same general dynamics (e.g., rates of change and direction of change) present in the associated ODE-based compartment model in Eq. (1), while allowing variability on the individual-level model in terms of specific parameters. Specifically, we constructed input-output properties for each set of equations based on the partial derivatives of the corresponding ODE. If we take as an example Eq. (1a), $\frac{\partial \dot{S_1}(t)}{\partial S_1} = -\frac{1}{t_{max}}$; then, we have that the corresponding NN in Eq. (4a) ($\mathbb{N}_{f1}$) must have the property that *all else equal*, the output from Eq. (4a) should monotonically decrease as S_1 increases. Similarly, the output of $\mathbb{N}_{f1}$ should increase as u_I increases. This approach allows us to formally verify whether the NN models conform to known physiological properties, while providing flexibility in the specific rates of change observed.

Following this rationale, we can define a set of properties for each NN in Eqs. (4a)–(4h), based on their corresponding ODE in Eqs. (1a)–(1h). For the compartments defined in Eqs. (4i) and (4j), the properties were defined based on the two-compartment ODE model described by Hovorka et al. in [53] for carbohydrate absorption.

Using these properties as the ground truth, we set up two copies of the same network $\mathbb{N}_{fi-left}$ and $\mathbb{N}_{fi-right}$ in parallel, with all inputs but the one input being tested (e.g., x_{test} or u_{test}) held equal, as shown in Fig. 2. If we were evaluating x_{test}, at the input location of x_{test}, $\mathbb{N}_{fi-left}$ gets x_{test} and $\mathbb{N}_{fi-right}$ gets $x_{test} + \epsilon$, with $0 \leq \epsilon \leq \delta$. We then construct a range estimation optimization problem to maximize (or minimize, when appropriate) the difference z_i between the outputs of $\mathbb{N}_{fi-left}$ and $\mathbb{N}_{fi-right}$. This pushes the networks to extremes, allowing us to easily and formally identify any locations where the network dynamics break. When testing for a monotonically decreasing property ($\downarrow$), we set $z_i = \mathbb{N}_{fi-right} - \mathbb{N}_{fi-left}$ and expect $z_i \leq 0$. For a monotonically increasing property ($\uparrow$), we set $z_i = \mathbb{N}_{fi-left} - \mathbb{N}_{fi-right}$ and expect $z_i \leq 0$.

Algorithm 2 Neural network conformance verification

Require: Neural network $\mathbb{N}_{fi}$, minimum and maximum value of all $\mathbb{N}_{fi}$ inputs, input to be tested (x_{test} or u_{test}), property to be tested ($\uparrow$: monotonic increase or $\downarrow$: monotonic decrease), δ, minimum-viable change Δx^i_{min}

Ensure: Conformance verification result

1: **create** two replicas of $\mathbb{N}_{fi}$ as $\mathbb{N}_{fi-left}$ and $\mathbb{N}_{fi-right}$
2: **execute** Gurobi solver to find the objective values $objVal$ and $objVal_\epsilon$, which given to $\mathbb{N}_{fi}$ would result in the minimum value of z_i
3: **evaluate** $\mathbb{N}_{fi}$ at $objVal$ and $objVal_\epsilon$
4: **if** property to be tested $= \uparrow$
$z_i \leftarrow \mathbb{N}_{fi}(objVal) - \mathbb{N}_{fi}(objVal_{\epsilon)}$
else
$z_i \leftarrow \mathbb{N}_{fi}(objVal_{\epsilon)} - \mathbb{N}_{fi}(objVal)$
end if
5: **if** $z_i < \Delta x^i_{min}$
$\mathbb{N}_{fi}$ is conformant w.r.t x_{test} or u_{test}
else
$\mathbb{N}_{fi}$ is non-conformant w.r.t x_{test} or u_{test}
$\Delta x^i_{critical} \leftarrow z_i$
end if

The range estimation problem we just described asks for a conservative over-approximation of the output of a network given constraints on the inputs. For our approach, we cast the range estimation problem in a mixed

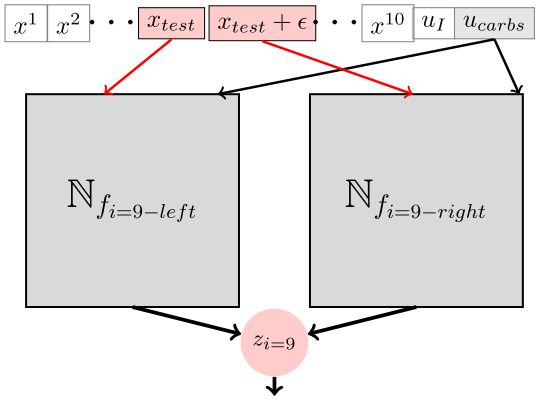

Fig. 2 Scheme for checking δ-monotonicity of a network $\mathbb{N}_{fi}$ with respect to a specific input location x_{test} adapted from Kushner et al. [33]. In this example, conformance verification is performed on $\mathbb{N}_{f9}$, $x_{test} = x^9 = C_1$, and u_{carbs} is the other input to the network

integer linear programming (MILP) optimization framework, and solved it via the Gurobi solver [62] to estimate the objective values *objVal* and $objVal_\epsilon$ that would result in the minimum values of z_i. To validate the results of the MILP result, we checked the best-case solution (e.g., the example which falsifies the conformance problem) by evaluating the network $\mathbb{N}_{fi}$ with inputs *objVal* and $objVal_\epsilon$ and calculating the actual value of z_i. In all cases, we desire $z_i \leq 0$. Thus, if we find a solution where $z_i > 0$, we note the network $\mathbb{N}_{fi}$ is non-conformant. In essence, we took a falsification approach to solving the verification problem.

3.3.1 Controlling for minimum-viable change in neural network compartments

When verifying conformance of the each $\mathbb{N}_{fi}$, the NNs can demonstrate dynamics which are below the discrepancy threshold for the Gurobi optimizer. To address this issue, we included a minimum-viable change Δx^i_{min} based on the absolute value of smallest non-zero rate of change measured at the *i*th compartment over 1 time unit (i.e., $\Delta t = \Delta m = 5$ min) in $\mathcal{D}^P_{train}$. We considered Δx^i_{min} to be effectively zero based on physiology knowledge from the ODE-based glucoregulatory model and used it to determine whether or not a sub-network was conformant. Therefore, if we found a solution where $z_i < \Delta x^i_{min}$, we noted the network $\mathbb{N}_{fi}$ is conformant; otherwise, we reported the critical error $\Delta x^i_{critical}$ below which the network output would be considered partially conformant.

3.4 Individual-level model

To construct the individual-level augmentation models, we developed a NN model $\mathbb{N}_{Ind}$, parameterized by θ^I_k, $k = 1, ..., K$, called $T1DSim^I_{NN,k}$ (Eq. 11). K is the total number of digital twins to be constructed.

$$\begin{aligned} \dot{Q}^I_1(t) = \mathbb{N}_{Ind}(X_1(t), X_3(t), Q_1(t), \\ Q_2(t), C_2(t), \mathbf{u_{Ind}}(t); \theta^I_k) \end{aligned} \tag{11}$$

We construct this model to be trained over time on the data generated by each individual using gradient descent, allowing the network to act as an additional compartment that captures the inter- and intra-individual variability in glucose dynamics not modeled by the population-level model $T1DSim^P_{NN}$. $\mathbb{N}_{Ind}$ has as inputs the same states as the compartment Q_1 (Eq. 4g) along with the input vector $\mathbf{u_{Ind}}(t)$ which includes sleep efficiency, heart rate, and contextual time information (e.g., hour of day, and weekday versus weekend) specific to a given individual. To capture complex individual patterns, we designed $\mathbb{N}_{Ind}$ as a multi-layer NN that consists of 3 fully connected hidden dense layers with 128, 64, and 32 neurons, respectively. Each

hidden layer has a ReLU activation function. $T1DSim^{I}_{NN,k}$ is integrated in the simulation as shown in Eq. (12).

$$y^{sim}_{Ind}(t+\Delta t) = Q_1(t) + (\dot{Q}_1(t) + \dot{Q}^{I}_1(t))\Delta t \tag{12}$$

3.4.1 Development dataset preprocessing

The development dataset $\mathcal{D}^{I}_{k} = \{\mathrm{x}(t), \mathrm{u}(t), \mathbf{u_{Ind}}(t)\}, k = 1, ..., K$, was generated using data from the main phase of the T1DEXI study (see Appendix A).

All the system's states ($\mathrm{x}(t)$) were simulated using the $T1DSim^{P}_{NN}$ during the training of the individual model, except for $x^{7}(t) = Q_1(t)$ which is the actual glucose measured by the CGM sensor. The system's input $u_I(t)$ was obtained from the insulin pumps and the carbohydrate intake $u_{carb}(t)$ from self-reported meal events. The self-reported meal events were confirmed by a validated meal detection algorithm [63]. We first reviewed the predicted meals and included those that (1) had a bolus reported within a 90-min window (for these, we used the estimated carbohydrate amount and the bolus timestamp) or (2) had a meal reported within a 90-min window but no bolus reported (for these, we used the carbohydrate amount reported in that range with timestamp detected by the missed-meal algorithm). Next, we processed the remaining bolus events, checking for meal events within a 90-min window and including them where applicable. Throughout this process, we ensured that no meal event was included more than once. Meals that did not meet any of these preprocessing criteria were discarded as unreliable. Initially, a total of 44,133 meals were reported across all datasets. After applying the described process, the number of meal events increased to 51,497. Of the initial events, only 11,205 retained their original size and timing without modification. The meal-detection algorithm was utilized to mitigate the problem whereby some individuals did not accurately report meal events, thereby improving the quality of the dataset.

The individual input vector ($\mathbf{u_{Ind}}(t)$) included the following inputs:

- *Timed-based contextual features*: Hour of day coded as $\cos\left(2\pi\frac{hour}{24}\right)$, $\sin\left(2\pi\frac{hour}{24}\right)$; and a binary variable to indicate whether or not the simulation instance corresponds to a weekend day.
- *Heart rate*: Change in heart rate measurements expressed in beats per minute (BPM), relative to the individual's baseline. The individual's baseline was estimated as the average heart rate during periods of rest, when no exercise was being performed.
- *Sleep efficiency*: Fraction of time within a 5-min window spent sleeping.

The heart rate was scaled using a robust scaler as done for the states in the population-level model. The remaining inputs were constructed to be between 0 and 1 or between -1 and 1 (e.g., the hour feature), thus avoiding the need for scaling.

$\mathcal{D}^{I}_{k}$ was divided into a training subset ($\mathcal{D}^{I}_{train,k}$) comprising the first two weeks of the study and a testing set ($\mathcal{D}^{I}_{test,k}$) corresponding to the remaining days. We chose a two-week training window because previous research has shown it to be the optimal duration for identifying individual-level glucose dynamics using machine learning, while also minimizing the risk of overfitting [64]. If for a given participant, there were no insulin or carbohydrates reported in either of the two subsets, the participant was excluded from the virtual population.

3.4.2 Model training

The $T1DSim^{I}_{NN,k}$ models constructed for all digital twins were trained following the methodology outlined in Sect. 3.2.2. No modifications were made to the methods apart from those indicated in this section.

Given $\mathcal{D}^I_{train,k}$, we created batches containing 5-h sequences, with each consecutive sequence overlapping 90% with the previous one. Note that this overlap is higher than the 75% used for the population-level model, reflecting the smaller amount of individual-specific training data. The overlap ensures that the training data contains a larger number of samples, as a new sequence is generated by shifting the starting point by 10% of the sequence length (e.g., 30 min for a 5-h sequence). This approach increases the diversity of training samples by exposing the model to slightly shifted but highly similar input sequences, improving its robustness to variations in temporal alignment. The augmentation process is useful, as small shifts help capture subtle temporal patterns. Since we penalized each timestamp in the sequence, we required sequences to have 100% of the CGM values present to be included in the training phase.

The loss function for the training individual-level models is defined in Eq. (13).

$$\mathcal{L}^I_{total} = \mathcal{L}^I_{fit} + \alpha^I \mathcal{L}^I_{consistency} \tag{13}$$

$\mathcal{L}^I_{fit}$ was defined as in Eq. (6). However, we redefined the penalty function as presented in Eq. (14) to penalize more heavily errors in the low glucose range. This penalty in the training process is important for ensuring that the digital twins are accurately representing glucose data in all ranges, including low glucose values which are important to model for any interventional system. High times below range are associated with high glucose variability, which is an increasingly relevant clinical marker of daily glucose control [65]. We searched for the penalty value for the condition $y \leq 70 \wedge y^{sim} > y$ in $\mathcal{P}^I(y^{sim}, y)$ that would better approximate global TBR statistics for all digital twins in $\mathcal{D}^I_{train,k}$.

$$\mathcal{P}^I(y^{sim}, y) = \begin{cases} 8.5 & y \leq 70 \wedge y^{sim} > y \\ 6 & y \geq 250 \wedge y^{sim} < y \\ 1 & otherwise \end{cases} \tag{14}$$

$\mathcal{L}^I_{consistency}$ was defined as the mean squared error between the first-order differences of the simulated 5-h sequence and the first order differences observed in the actual glucose data. This error term was included to avoid unrealistic changes within a 5-min window in the simulation. And $\alpha^I \geq 0$ is a regularization parameter selected to balance $\mathcal{L}^I_{fit}$ and $\mathcal{L}^I_{consistency}$.

For training the individual-level models, we used a learning rate of $\lambda_I = 10^{-4}$ and a regularization constant $\alpha^I = 10$. Each individual model was trained for 150 epochs.

3.5 Simulation of real-world scenarios using physiologically-constrained neural network digital twins

3.5.1 Estimating initial states

To initialize the simulator when only the initial CGM data was available, we developed an approach based on steady-state assumptions derived from the CGM readings at the start of the simulation ($t = 0$). Our technique involves obtaining the set of initial states $\mathrm{x}(0)$ calculated when assuming $\dot{\mathrm{x}}(0) = 0$, given the model $T1DSim^P_{NN}$. This was achieved using a gradient-based optimization approach with the stochastic gradient descent optimizer. The number of epochs and the learning rate were empirically set to 10,000 and 0.1, respectively.

3.5.2 Conducting simulations

Given the initial states $\mathrm{x}(0)$, and inputs $\mathrm{u}(t)$ and $\mathbf{u_{Ind}}(t)$, $t = 0, ..., T$. A $T-$length simulation of the glucose dynamics of the k^{th} digital twin parameterized by θ^I_k can be performed using the Euler's ODE integration method (Eq. 12).

3.6 Evaluation of simulation accuracy

3.6.1 ODE-based digital twins used as control models

We compared our approach with other ODE-based methodologies. Similar published digital twin frameworks in T1D are all based on mechanistic models governed by ODEs describing glucose-insulin dynamics [66]. Thus, we prioritized comparisons between our proposed methodology and the clinically validated ODE-based glucoregulatory model developed by Resalat et al. [19], as follows:

- *ODE-based population-level model*: We simulated real-world scenarios using the population model parameters presented in [19].
- *Closest ODE-based digital twin from existing virtual population*: For each participant in the T1DEXI study included in our analysis, we identified the best digital twin from the population of 99 virtual patients described in [19] by selecting the twin that over the first two weeks of the T1DEXI study minimized the error used in the loss function optimized during the individual-model optimization phase (Eq. 13).
- *Bayesian optimization-based digital twin*: Similar to the ReplayBG twinning technique by Cappon et al. [31], we used Bayesian optimization to find the optimal subset of the ODE-based model parameters related to insulin sensitivity (i.e.,S_{f1}, S_{f2}, and S_{f3}) that minimized simulation error for each participant in the T1DEXI study included in our analysis. Insulin sensitivity parameters are the most important parameters that capture inter-individual variability. We used Bayesian optimization [61] instead of MCMC because this methodology requires substantial time and computational resources. MCMC is a sampling algorithm that construct Markov Chains to approximate distributions and is used for Bayesian inference to estimate posterior distribution of a model's parameters. Bayesian optimization uses probabilistic modeling to optimize expensive functions and it is use for parameter tuning or black-box function optimization. In our work, the optimization target was to minimize the simulation error using the first two weeks of the T1DEXI study.

3.6.2 Statistical analysis of similarity between simulated and real-world glucose outcomes

The accuracy of the simulated glucose profiles was evaluated by comparing average outcome metrics between each digital twin and real-world data. Specifically, we assessed TIR, TAR, and TBR. Additionally, we computed clinically relevant risk indices, including the high and low blood glucose indices (HBGI and LBGI), as well as mean glucose (MG). These metrics are widely used to assess clinical performance and glucose control [67].

To ensure consistency with the model training data, all metrics were calculated over 5-h sequences. This sequence length was chosen to reflect a realistic simulation window, capturing key physiological dynamics such as those associated with meals or physical activity.

We used paired equivalence t-tests to determine whether simulated and real-world outcomes were statistically equivalent within predefined clinically meaningful margins. The equivalence margins for each metric were: $\delta_{TIR} = 5\%$, $\delta_{TAR} = 5\%$, $\delta_{TBR} = 1\%$, $\delta_{HBGI} = 1$, $\delta_{LBGI} = 0.5$ and $\delta_{MG} = 10$ mg/dL [67–69]. All tests were two one-sided paired t-tests (TOST) with a significance level of 0.050. We also estimated the 95% confidence intervals (CIs) of the mean glucose outcomes for both real and simulated data. The 95% CIs were obtained using a bootstrapping procedure with replacement, drawing samples of size 50 across 10,000 iterations. We expect the 95% CIs of the real and simulated mean to be comparable. Results are reported as mean (95% CI), along with the larger of the two one-sided *P*-values.

3.7 Evaluation of model robustness

In the analysis of similarity between real-world and simulated glucose outcomes, we aimed to verify that the digital twins capture the overall glucose dynamics. However, to further deepen the analysis, this section focuses

on the methodology used to identify and characterize edge cases for each digital twin in which substantial discrepancies arise in clinically relevant glucose outcomes (i.e., TIR, TBR, and TAR).

Outliers for each digital twin were identified separately for over- and under-prediction of a given glucose outcome using Tukey's criterion, defined by the interval $[Q_1 - 1.5(Q_3 - Q_1),\ Q_3 + 1.5(Q_3 - Q_1)]$, where Q_1 and Q_3 denote the first (25%) and third (75%) quartiles, respectively.

Once these cases were identified, we investigated the underlying causes of errors. In order to categorize the issues preventing the model to accurately reproduce observed glucose dynamics, we defined five possible sources of discrepancy detailed below. After classification, the percentage of occurrences of each category was computed to identify the most prevalent sources of prediction errors across digital twins.

1. *Category* 1: *Initialization error*. The simulation is highly dependent on the definition of the initial states. If divergence between the simulated and actual glucose trajectories occurs from the beginning of the sequence, the error is attributed to initialization. Specifically, this problem is identified when the RMSE increases during the first hour of simulation and exceeds the median RMSE across all digital twins within the initial hour.
2. *Category* 2: *Unreported meal*. Since meals are self-reported, errors or omissions may occur. Despite the use of a meal detection algorithm, some meals may remain undetected. To identify these cases, we detect periods of rising glucose levels accompanied by an insulin bolus occurring within ±1 h of the inferred meal time.
3. *Category* 3: *Meal without insulin bolus report*. In some cases, meals may have been recorded for example following the use of a bolus calculator, but insulin delivery was not confirmed or done using another insulin delivery device. These cases are identified by detecting meals for which no bolus was recorded within ±1 h, leading to glucose excursions in the predicted data that are not reflected in the actual measurements.
4. *Category* 4: *Unmodeled dynamics or missing input data*. Discrepancies may also arise due to physiological dynamics not captured by the model or due to unreported inputs. These cases are identified by time frames in which predicted and actual glucose outcomes differ (e.g., one sequence remains in range while the other is above range) and the squared error increases over time, indicating diverging trajectories.
5. *Category* 5: *Correct dynamics with differing glucose outcomes*. In contrast to the previous categories, some sequences exhibit similar underlying dynamics but differ in glucose outcomes. This scenario is identified by comparing point-to-point glucose variations between predicted and actual trajectories. The dynamics are considered similar if these variations differ by less than 10 mg/dL for at least 70% of the sequence. Differences in outcomes in these cases typically occur near clinical boundaries (e.g., one sequence in 70–180 mg/dL range and the other >180 mg/dL).

3.8 Additional experiments

To further evaluate the utility and robustness of the digital twins, we conducted three complementary experiments:

3.8.1 Contribution of individual-level inputs

To assess the contribution of each component of the individual input vector $\mathbf{u_{Ind}}(t)$, we trained multiple versions of the digital twin model using different subsets of the input features. Specifically, we trained models using only one input at a time-heart rate, sleep efficiency, weekend day or time-based features—as well as a model without addition inputs (i.e., $\mathbf{u_{Ind}}(t) = \{\}$). They where compared with the NN-based digital twin proposed.

Each model configuration was evaluated across three key physiological events known to impact glucose levels (i.e., meals, physical activity, and sleep). We used the same outcome metrics described above (TIR, TAR, TBR, HBGI, LBGI, and MG). This experiment aimed to determine whether incorporating all individual-level inputs provides additive value and improves simulation accuracy compared to models trained with partial or no individual-level information.

Table 2 Number of units in the hidden layers of subnetworks in the population- and individual-level models

Population-level model ($T1DSim^{P}_{NN}$)							
Glucose kinetics		Insulin kinetics		Insulin dynamics		Carbohydrate absorption	
$\dot{Q}_1$	248	$\dot{S}_1$	174	$\dot{X}_1$	191	$\dot{C}_1$	149
$\dot{Q}_2$	80	$\dot{S}_2$	60	$\dot{X}_2$	101	$\dot{C}_2$	140
		$\dot{I}$	197	$\dot{X}_3$	127		
Individual-level model ($T1DSim^{I}_{NN,k}$)							
$\dot{Q}^{I}_1$ [128,64,32]							

Table 3 Results of the neural network state-space model conformance verification: summary of the maximum critical error $\Delta x^{i}_{critical}$ for each sub-network of the population-level model ($T1DSim^{P}_{AI}$)

$\Delta x^{i}_{critical}$							
Glucose kinetics		Insulin kinetics		Insulin dynamics		Carbohydrate absorption	
Q1 [mg/dL]	$2e^{-2}$	S1 [mU/kg]	$1e^{-2}$	X1 [min^{-1}]	$5e^{-9}$	C1 [-]	✓
Q2 [mmol/kg]	$2e^{-5}$	S2 [mU/kg]	$9e^{-6}$	X2 [min^{-1}]	✓	C2 $\left[\frac{mmol}{kg \cdot min}\right]$	$2e^{-8}$
		I [mU/L]	$2e^{-5}$	X3 [unitless]	$4e^{-7}$		

If the result in the sub-network is a checkmark (✓), it indicates that the sub-network is fully conformant for all tested inputs

3.8.2 Evaluation as a predictive model

Although digital twins are designed for scenario-based simulation (i.e., forecasting glucose trajectories under varied conditions), we also evaluated their performance in a traditional predictive setting (i.e., forecasting glucose values over a short-term prediction horizon). Specifically, we compared the performance of the digital twins against state-of-the-art glucose prediction models [33, 35, 70] using a 30-, 60- and 120-min prediction horizon.

3.8.3 Evaluation of digital twins behavior under variable insulin boluses and carbohydrate inputs

We demonstrate the potential of the physiologically-constrained NN digital twins to replay various carbohydrate ratios by simulating three different meal scenarios (i.e., meal carbohydrate consumption of 30 g, 60 g, and 90 g) across all digital twins. For each meal scenario, we simulated insulin boluses for three different insulin-to-carbohydrate ratios (i.e., 10 g/U, 20 g/U, and 30 g/U). These simulations qualitatively illustrate that the digital twins can, for example, be used to adjust the optimal insulin-to-carbohydrate ratio for each individual based on improving the postprandial glucose response.

To perform the prediction, we used the glucose value at time 0 of a given scenario as the initial condition and simulated the glucose trajectory over the following hour using the same scenario. The predicted glucose value at time 30, 60 and 120 min was then compared to the actual value. For comparison, we used point-wise root mean squared error (RMSE) as the evaluation metric. This comparison helps establish the digital twins' capabilities not only as simulators but also as effective short-term glucose predictors.

4 Results

4.1 Optimal sub-network architecture

Table 2 summarizes the number of hidden units for each sub-network that minimized validation error for the population-level model $T1DSim^{P}_{NN}$ and the individual-level models $T1DSim^{I}_{NN,k}$.

4.2 Population-level model performance and conformance analysis

The population-level model $T1DSim^{P}_{NN}$ closely matches the ODE-based reference model $T1DSim^{P}_{ODE}$, as evidenced by glucose metrics evaluated on the hold-out simulated dataset $\mathcal{D}^{P}_{test}$. Specifically, the two models show nearly identical performance across all clinically relevant metrics, including TIR: 74.1 (95% CI 64.3–83.0)% versus 72.9 (95% CI 62.7–82.1)%, TAR: 22.3 (95% CI 13.1–32.3)% versus 23.1 (95% CI 13.6–33.4)%, and TBR: 3.6 (95% CI 1.9–5.7)% versus 4.0 (95% CI 1.9–6.5)%. Similarly, risk indices and MG values are consistent between models, with LBGI: 1.4 (95% CI 1.0–1.8) versus 1.5 (95% CI 1.1–2.0), HBGI: 4.1 (95% CI 2.5–5.8) versus 4.2 (95% CI 2.6–6.0) and MG: 132.8 (95% CI 120.3–146.0) mg/dL versus 132.6 (95% CI 119.6–146.0) mg/dL. All metrics are statistically significant (P-value = < 0.001), as expected given that the population-level neural model was trained on data generated from the underlying ODE-based simulator.

The results of verifying the dynamics learned by each sub-network $\mathbb{N}_{fi}$ in $T1DSim^{P}_{NN}$ are summarized in Table 3. This table shows the maximum critical error found when testing different inputs and properties within each sub-network as well as the input that produced such error. The sub-networks corresponding to states X_2 and C_1 are fully conformant. The other sub-networks are partially conformant with insignificant critical errors. For example, the maximum critical error for $\dot{Q}_1$ is only $2e^{-2}$ mg/dL, which is a very small error in practice. A detailed version of the conformance analysis is presented in Appendix C.

4.3 Validation of the physiologically-constrained neural network digital twins

We created a total of 394 physiologically-constrained NN digital twins. Table 4 presents the glucose outcomes for 5-h simulations of both NN-based and ODE-based digital twins, along with the results of the population-level

Table 4 Glucose outcomes for 5-h sequences across 394 digital twins, comparing actual data, the NN-based population-level model, the ODE-based population-level model, the NN-based digital twins, the ODE-based digital twins, and the BayesianOpt-based digital twins

	Actual	ODE-based population-level (control)	ODE-based digital twins (control)	Bayesian opt-based digital twins (control)	NN-based population-level (this work)	NN-based digital twins (this work)
TIR [%]	74.4 (69.9–78.6)	66.4 (63.1–69.5) 1.000	64.7 (59.0–70.0) 1.000	71.9 (68.1–75.4) < **0.001**	65.6 (61.6–69.5) 1.000	75.1 (69.0–80.9) < **0.001**
TAR [%]	22.6 (18.4–27.3)	27.7 (24.4–31.2) 0.560	29.3 (23.1–35.8) 0.999	21.0 (17.2–25.4) <**0.001**	33.1 (29.4–37.1) 1.000	22.4 (16.6–28.8) <**0.001**
TBR [%]	3.0 (2.2–4.0)	5.9 (4.4–7.8) 1.000	6.0 (3.9–8.5) 1.000	7.1 (5.5–8.9) 1.000	1.3 (0.8–1.8) 1.000	2.5 (1.2–4.1) **0.022**
LBGI	0.9 (0.7–1.1)	1.7 (1.2–2.2) 1.000	1.6 (1.1–2.2) 0.992	1.9 (1.5–2.4) 1.000	0.4 (0.3–0.5) 0.251	0.7 (0.4–1.0) <**0.001**
HBGI	5.3 (4.2–6.6)	6.5 (5.7–7.4) 0.944	6.8 (5.3–8.6) 1.000	4.9 (4.0–5.9) **0.001**	7.6 (6.7–8.5) 1.000	5.3 (3.9–7.1) <**0.001**
MG [mg/dL]	147.2 (140.6–154.6)	153.2 (147.5–159.0) < **0.001**	153.8 (143.4–164.7) < **0.001**	140.8 (134.2–147.9) < **0.001**	165.0 (159.9–170.4) 1.000	149.2 (140.2–159.3) <**0.001**

P-values below this threshold are shown in bold

Values are presented as mean, confidence interval (CI) and *P*-value. Equivalence margins used for the test: $\delta_{TIR} = 5\%$, $\delta_{TAR} = 5\%$, $\delta_{TBR} = 1\%$, $\delta_{HBGI} = 1$, $\delta_{LBGI} = 0.5$ and $\delta_{MG} = 10$ mg/dL. Equivalence is concluded only if the TOST p-value is < 0.05 (i.e., both one-sided tests are significant)

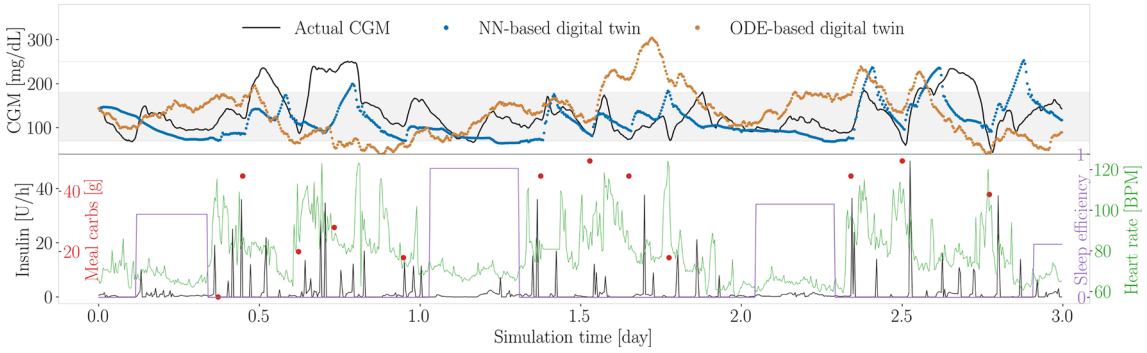

Fig. 3 Example of a 3-day glucose trace from the actual CGM versus simulated (NN-based digital twin vs. ODE-based digital twin) (Top panel). Glucose management scenario used for the simulation, which includes endogenous insulin, amount of carbohydrates, and heart rate and sleep efficiency (Bottom panel). TBR, TAR, and TIR for the actual trace are 1.9%, 15.0%, and 83.1%, respectively. For the NN-based digital twin, the values are 4.7%, 6.8%, and 88.4%, while for the ODE-based digital twin, they are 16.6%, 20.8%, and 62.6%

models. The NN-based digital twins consistently produce more accurate simulation results across all considered metrics when compared with ODE-based digital twins using different twinning techniques. Our results demonstrate that glucose outcomes calculated from NN-based digital twins closely match real-world glucose outcomes within clinically significant glucose ranges. The difference between simulated and observed outcomes are 0.7%, − 0.2%, − 0.5%, − 0.2, 0.0, 2.0 mg/dL for TIR, TAR, TBR, LBGI, HBGI, and MG, respectively (TIR 75.1 (95% CI 69.0–80.9)% versus 74.4 (95% CI 69.9–78.6)% (*P*-value = < 0.001), TAR 22.4 (95% CI 16.6–28.8)% versus 22.6 (95% CI 18.4–27.3)% (*P*-value = < 0.001), TBR 2.5 (95% CI 1.2–4.1)% versus 3.0 (95% CI 2.2–4.0)% (*P*-value = 0.022), LBGI 0.7 (95% CI 0.4–1.0) versus 0.9 (95% CI 0.7–1.1) (*P*-value = < 0.001), HBGI 5.3 (95% CI 3.9–7.1) versus 5.3 (95% CI 4.2–6.6) (*P*-value = < 0.001) and MG 149.2 (95% CI 140.2–159.3) mg/dL versus 147.2 (95% CI 140.6–154.6) mg/dL (*P*-value = < 0.001)).

Figure 3 shows an illustrative example of a 3-day simulation comparing an NN-based digital twin (this work), an ODE-based digital twin (control), and the actual CGM values. In this example, our approach exhibits a better match in glucose outcomes compared to the ODE-based digital twin, reproducing key outcome metrics (TIR, TAR, and TBR), as detailed in the caption. While some short-term deviations from the actual CGM trace are visible, these are expected and do not reflect clinically meaningful differences. The NN-based digital twin captures the underlying physiological dynamics, including post-meal glucose rises and reductions associated with elevated heart rate, possibly due to physical activity.

4.4 Model robustness analysis

Across the NN-based digital twins, events with larger model errors represented on average 28 ± 13% of sequences per digital twin, corresponding to a total of 6,309 sequences across all 394 digital twins. We evaluate these cases relative to the robustness categories identified in Sect. 3.7. Among these cases, the most prevalent sources of discrepancies were Category 1 (initialization error) and Category 4 (unmodeled dynamics or missing input data) issues, which accounted for 45 ± 20% and 40 ± 20% of events with large errors, respectively. The remaining cases were distributed among Category 2 (unreported meals), Category 3 (meals without bolus report), and Category 5 (similar dynamics with differing glucose outcomes), with lower percentages (12 ± 13%, 0 ± 2%, and 2 ± 5%, respectively). These results indicate that, while large errors can arise from multiple sources, initialization issues and unmodeled or missing dynamics are consistently the primary contributors across digital twins.

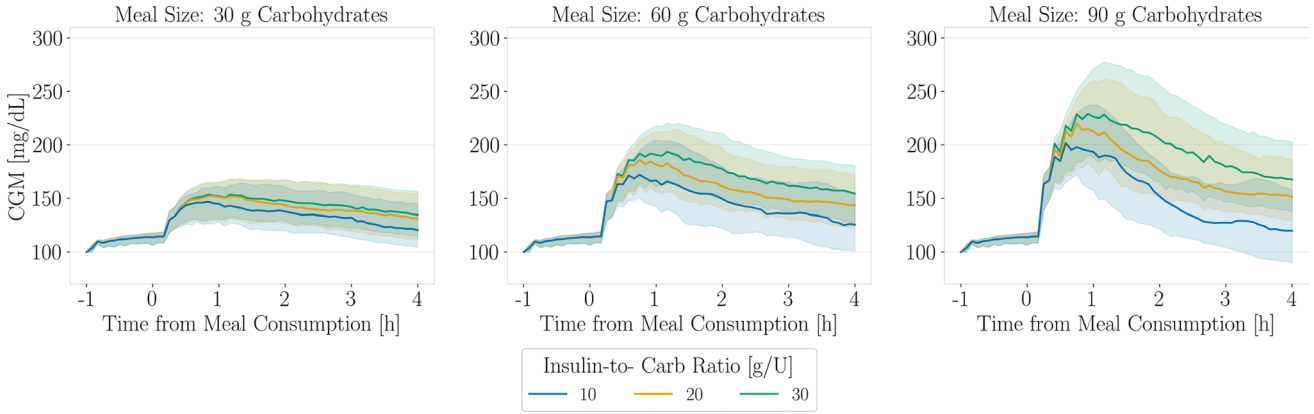

Fig. 4 Simulated responses of physiologically-constrained NN digital twins to three meal scenarios (from left to right: 30 g, 60 g, and 90 g of carbohydrate consumption). For each meal scenario, insulin boluses were simulated at three insulin-to-carbohydrate ratios (10 g/U, 20 g/U, and 30 g/U). Results for each simulated scenario in the set of physiologically-constrained NN digital twins are shown as the median (thick line) and interquartile range (shaded area). Note that increased carbohydrate consumption leads to higher postprandial glucose, while higher insulin-to-carbohydrate ratios reduce postprandial glucose for each bolus tested

4.5 Additional experiments

Figure 4 qualitatively illustrates the ability of the NN-based digital twins to adapt to varying inputs, showcasing their potential for individual-specific optimization of factors such as insulin-to-carbohydrate ratios. Furthermore, the results demonstrate variability in the response among the digital twins.

Table 5 summarizes the similarity between simulated and actual glucose profiles under varying input configurations across three scenarios: physical activity, meals, and sleep. Digital twins that included all available inputs produced glucose outcomes more closely aligned with real-world data. For physical activity and meal inputs, when all variables are included, all metrics passed the equivalence test except for TBR; for sleep, all metrics passed except for HBGI. In other input configurations, the behavior was not the same, as they failed at least two of the tests. For example, in the physical activity scenario, when no additional inputs were included, the MG, TBR, LBGI, and HBGI were not within the predefined equivalence margins. However, when models were trained using heart rate or time-based features, MG and HBGI became equivalent to real-world outcomes. Furthermore, when all individual-level inputs were included, LBGI also reached equivalence. This is a significant finding, as it indicates that incorporating these inputs adds value and improves the accuracy of simulated outcomes.

Finally, our point-based RMSE for a 30-, 60- and 120-min prediction horizon is 27±7 mg/dL, 41±12 mg/dL and 54±18 mg/dL, respectively, which is in alignment with other state-of-the-art models [33, 35, 70]. These results demonstrate that our proposed digital twins can function not only as a simulation tool but also as effective short-term glucose predictors.

5 Discussion

We described a framework for constructing physiologically-constrained NN digital twins capable of replicating the glucose dynamics of people living with T1D. Our proposed approach combines a population-level model $T1DSim^{P}_{NN}$ that is verified to conform to known glucose-insulin dynamics within very low critical errors bounds with an individual-level model $T1DSim^{I}_{NN}$ that captures residual dynamics not modeled by $T1DSim^{P}_{NN}$. The digital twins constructed in this work not only demonstrated good agreement with actual glucose values when

Table 5 Performance comparison of digital twins trained with varying input configurations across three scenarios: physical activity, meals, and sleep

	Actual	All individual inputs	Heart rate	Sleep efficiency	Weekend day	Time-based features	No additional inputs
Scenario: physical activity							
TIR [%]	73.9 (69.1–78.4)	73.1 (66.2–79.7) **< 0.001**	73.7 (66.4–80.5) **< 0.001**	74.1 (66.6–81.1) **< 0.001**	73.5 (66.0–80.4) **< 0.001**	73.3 (66.1–80.1) **< 0.001**	73.5 (66.0–80.5) **< 0.001**
TAR [%]	22.2 (17.7–27.1)	24.2 (17.5–31.3) **< 0.001**	24.7 (17.7–32.0) **0.002**	24.8 (17.7–32.3) **0.003**	25.5 (18.4–32.9) **0.020**	24.6 (17.7–31.8) **< 0.001**	25.7 (18.6–33.1) **0.039**
TBR [%]	3.8 (2.7–5.3)	2.7 (1.1–4.8) 0.679	1.6 (0.6–2.9) 1.000	1.0 (0.3–2.2) 1.000	1.0 (0.3–2.1) 1.000	2.1 (0.7–4.5) 0.989	0.8 (0.2–1.8) 1.000
LBGI	1.0 (0.8–1.4)	0.7 (0.4–1.2) **0.004**	0.5 (0.3–0.8) 0.656	0.4 (0.2–0.7) 0.999	0.4 (0.2–0.7) 0.999	0.6 (0.3–1.0) 0.383	0.3 (0.1–0.6) 1.000
HBGI	5.1 (4.0–6.4)	5.7 (4.1–7.5) **0.013**	5.6 (4.1–7.3) **0.002**	5.7 (4.1–7.5) **0.017**	5.8 (4.3–7.6) 0.059	5.7 (4.2–7.4) **0.009**	5.8 (4.3–7.5) 0.071
MG [mg/dL]	145.5 (138.4–153.6)	152.0 (141.6–163.4) **0.002**	152.7 (142.5–163.7) **0.008**	153.9 (143.7–165.2) 0.091	154.8 (144.6–165.9) 0.260	153.1 (142.9–163.9) **0.015**	155.4 (145.4–166.3) 0.463
Scenario: meals							
TIR [%]	70.5 (65.8–74.9)	70.3 (63.1–77.0) **< 0.001**	71.6 (63.9–78.9) **< 0.001**	72.1 (64.4–79.5) **< 0.001**	71.4 (63.7–78.7) **< 0.001**	70.7 (63.2–77.8) **< 0.001**	71.7 (63.9–79.1) **< 0.001**
TAR [%]	26.1 (21.6–31.1)	27.9 (21.1–35.1) **< 0.001**	27.2 (20.0–35.0) **< 0.001**	26.8 (19.7–34.6) **< 0.001**	27.3 (20.1–35.1) **< 0.001**	27.6 (20.6–35.3) **< 0.001**	27.3 (20.1–35.2) **< 0.001**
TBR [%]	3.4 (2.5–4.4)	1.9 (0.8–3.3) 0.999	1.2 (0.5–2.2) 1.000	1.1 (0.4–2.1) 1.000	1.3 (0.4–2.5) 1.000	1.7 (0.6–3.2) 0.999	1.0 (0.3–2.0) 1.000
LBGI	0.9 (0.7–1.2)	0.5 (0.3–0.9) **0.012**	0.4 (0.2–0.6) 0.891	0.4 (0.2–0.6) 0.953	0.4 (0.2–0.7) 0.779	0.5 (0.2–0.8) 0.114	0.3 (0.2–0.6) 0.987
HBGI	5.9 (4.8–7.3)	6.6 (4.9–8.7) **0.028**	6.4 (4.7–8.3) **< 0.001**	6.4 (4.6–8.4) **< 0.001**	6.4 (4.7–8.5) **< 0.001**	6.4 (4.8–8.4) **< 0.001**	6.4 (4.7–8.4) **< 0.001**
MG [mg/dL]	151.2 (144.1–158.9)	158.6 (148.2–169.9) **0.005**	158.3 (148.1–169.4) **0.002**	158.1 (147.6–169.6) **0.002**	158.4 (147.9–170.1) **0.003**	157.9 (147.8–169.0) **< 0.001**	158.8 (148.6–170.1) **0.009**
Scenario: sleep							
TIR [%]	76.3 (71.0–81.2)	77.9 (71.4–83.6) **< 0.001**	80.0 (73.4–86.0) 0.064	79.4 (72.8–85.4) **0.011**	81.7 (75.3–87.4) 0.695	78.7 (72.0–84.8) **< 0.001**	81.3 (74.6–87.2) 0.489
TAR [%]	21.1 (16.2–26.5)	18.4 (12.7–24.9) **0.002**	17.2 (11.4–23.9) 0.069	17.8 (11.9–24.6) **0.014**	15.6 (10.1–22.1) 0.720	18.7 (12.7–25.6) **< 0.001**	16.2 (10.5–23.0) 0.430
TBR [%]	2.6 (1.6–4.0)	3.7 (1.7–6.2) **0.001**	2.8 (1.1–4.8) **< 0.001**	2.8 (1.1–4.9) **0.017**	2.6 (1.0–4.7) **0.004**	2.5 (0.9–4.6) **0.005**	2.5 (0.8–4.6) **0.009**
LBGI	0.8 (0.6–1.1)	1.1 (0.7–1.5) **0.001**	1.0 (0.6–1.4) **<0.001**	0.9 (0.5–1.3) **<0.001**	0.9 (0.6–1.3) **<0.001**	0.9 (0.6–1.2) **<0.001**	0.9 (0.5–1.3) **<0.001**

Table 5 (continued)

	Actual	All individual inputs	Heart rate	Sleep efficiency	Weekend day	Time-based features	No additional inputs
HBGI	5.1	4.3	3.8	4.1	3.7	4.2	3.7
	(3.9–6.5)	(3.0–5.8)	(2.6–5.2)	(2.8–5.5)	(2.5–5.2)	(2.9–5.7)	(2.5–5.0)
		0.132	0.943	0.568	0.992	0.179	0.996
MG	146.0	138.3	135.9	137.9	134.8	138.8	135.1
[mg/dL]	(138.3–154.4)	(128.7–148.7)	(126.5–145.9)	(128.4–148.3)	(125.3–145.1)	(129.2–149.1)	(125.8–145.4)
		0.039	0.535	0.056	0.859	**0.008**	0.775

P-values below this threshold are shown in bold

Values are presented as mean, confidence interval (CI) and *P*-value. Equivalence margins used for the test: $\delta_{TIR} = 5\%$, $\delta_{TAR} = 5\%$, $\delta_{TBR} = 1\%$, $\delta_{HBGI} = 1$, $\delta_{LBGI} = 0.5$ and $\delta_{MG} = 10$ mg/dL. Equivalence is concluded only if the TOST *P*-value is < 0.05 (i.e., both one-sided tests are significant)

simulating meal, insulin, exercise, and sleep scenarios (Table 4), but also demonstrated their potential as a tool for simulating multiple interventions and optimizing treatment for individuals living with T1D (Fig. 4).

Using our proposed framework, digital twins can be constantly fine-tuned or extended by adding new inputs as new data from the twined individuals become available. This means the glucose dynamics of each digital twin can evolve along with the real-world twin using simple and fast methods involving gradient descent optimization techniques in contrast to the complex design and model identification techniques used for developing mechanistic models.

Our approach provide advances of significant utility by enabling long-term simulations across a multi-hour window, these digital twins can enable *in-silico* experiments that can enhance treatment strategies for individuals with T1D. We expect that the accuracy of the model will be sufficient for use in providing decision support around alternate treatment strategies such as alternative carbohydrate ratios or other settings within diabetes devices, but this will need to be tested in future research.

A natural question arising from this work is why we chose to adopt a NN state-space model design instead of simply creating a set of NN-based digital twins derived from the $T1DSim_{ODE}$ model. Our approach provides a physiologically meaningful and flexible framework for modeling the glucoregulatory system in T1D. Unlike traditional ODE-based models, the NN state-space model offers several advantages. First, it allows for continuous computationally efficient training as new data becomes available, ensuring adaptability to evolving clinical datasets. Second, it facilitates seamless integration into decision support systems, enhancing its applicability in personalized diabetes management. Third, the framework can be easily extended by incorporating additional compartments to account for other physiological factors affecting glucose regulation. These advantages position the NN state-space approach as a robust alternative for the creation of digital twins that can continuously evolve alongside their real-world counterparts.

While the number of parameters to be adjusted is larger than other approaches, we mitigate the risk of overfitting by the availability of extensive training data. Specifically, the population-level model contains 6,782 trainable parameters, representing less than 0.7% of the approximately 1 million 5-h sequences used for training. For the individual-level models, although they are more complex relative to the limited training data available, have significant reason to believe the observed improvements are not solely due to the model's complexity. Rather, factors such as the architecture design, the quality of the data, and the use of regularization techniques like weight decay play a critical role in preventing overfitting. Moreover, a key advantage of these digital twins is their ability to model additional features not included in the deterministic ODE model while continuing to train as more individual data becomes available, enhancing their performance and adaptability over time.

It is noteworthy that our initial approach to creating the individual-level models involved modeling the residual errors between the actual glucose data and the population-level model. However, this approach proved ineffective, as it tended to learn an average mean error. This was likely due to factors such as the intrinsic dependency on the duration of the population-level simulation, which made it challenging to capture the unmodeled phenomena driven by factors like physical activity or insulin sensitivity. We found our current approach improves robustness,

as it directly adjusts each digital twin's simulation to reflect changes in glucose dynamics, enabling a more accurate representation of individual variability.

In terms of model robustness, we found that the most prevalent causes of significant model errors were Category 1 (Initialization error) and Category 4 (Unmodeled dynamics or missing input data), occurring 45 ± 20% and 40 ± 20% of the time, respectively. These results indicate that discrepancies were primarily associated with limitations in model initialization and incomplete or missing input data rather than failures of the underlying physiological model.

Incomplete or unreliable input data, particularly unreported or inaccurately timed meal events, contributed to extreme prediction errors. Because meal ingestion critically influences postprandial glucose dynamics, missing or misaligned inputs can result in substantial divergence between simulated and observed trajectories. While meal detection algorithms mitigate some issues, undetected or falsely inferred meals remain a challenge and propagate errors into the simulations.

Another key factor in determining model performance was the initialization of model states. The accuracy of the digital twin simulations depended heavily on appropriate state initialization, particularly in capturing residual effects from prior meals or insulin on board, which has an impact on glucose trends. In several high-discrepancy cases, initial states did not adequately reflect recent physiological context, resulting in simulations that diverged from observed glucose patterns during the early sequence. In this study, we initialized the states using a steady-state assumption, and one possible approach to improve initialization is to consider the glucose trend during the 30 min preceding the start of the simulation. This emphasizes that state initialization remains a limitation of this framework.

Even with these data limitations, the methodology consistently reproduced overall glucose dynamics and clinically relevant trends. These findings demonstrate the robustness of the NN-based digital twin framework under typical data conditions while highlighting clear avenues for improvement, including advanced data curation, improved state initialization, and increased personalization through refined hyperparameter tuning.

6 Conclusion

This work presents a framework for constructing physiologically-constrained NN digital twins that can simulate glucose dynamics of individuals with T1D. We developed a novel NN state-space model architecture that allows for observability and interpretability of simulation outputs. This model adheres to known glucose-insulin dynamics as verified by our conformance verification analysis. When augmented with individual-level data, the resulting adaptive digital twin models can capture both inter- and intra-individual variability and incorporate various factors influencing glucose response, such as sleep and physical activity. We show that incorporating this data significantly improves simulation accuracy compared to digital twins trained with limited or varying input configurations, and that our framework outperforms ODE-based digital twins, highlighting the benefits of data-driven modeling combined with comprehensive input integration. Unlike traditional mechanistic models constrained by fixed parameters, our approach offers the flexibility to continuously adapt to new data, tasks, or architectures, enabling a dynamic and scalable framework for personalized modeling.

Appendix A Description of the datasets used for model development and testing

See Table 6.

Table 6 Description of the datasets used for models' development and testing

Characteristic	Dataset	
	Simulated	T1DEXI main study
Use	Population	Personalization
Participants, N		394
Demographics		
Biological sex, N		
Female (Male)		295 (99)
Age, years		
Mean±SD		37±14
BMI, kg/m^2		
Mean±SD		25±4
Insulin therapy, N		
Pump		179
Closed loop		215
Overall glucose control, Mean [Min–Max] at participant level		
CGM between 70–180 mg/dL, %		
Mean	73.0	75.2
[Min–Max]	[70.0–76.6]	[20.7–99.1]
CGM >180 mg/dL, %		
Mean	23.2	21.8
[Min–Max]	[20.1–27.0]	[0.1–79.3]
CGM <70 mg/dL, %		
Mean	3.7	3.0
[Min–Max]	[2.8–6.1]	[0.0–17.0]

Appendix B Robust scaler

Robust scaling involves subtracting the median of the distribution of all values that a given compartment can take in the training dataset from the unscaled state value, and then dividing by the distribution inter-quartile range. For example, to scale the Q_1 compartment state value at time *t*, we used Eq. (B1). Scaling constants for all system's states were calculated from the training dataset and stored for subsequent use during the model validation and testing phases.

In Eq. (B1), $Q_1(t)_{sc}$ is the scaled version of $Q_1(t)$; $Q_{1,(50)}$ is the 50^{th} percentile or median value of the distribution of all values that the Q_1 compartment can take in the training set; $Q_{1,(25)}$ and $Q_{1,(75)}$ are the 25^{th} and 75^{th} percentiles, respectively.

$$Q_1(t)_{sc} = \frac{Q_1(t) - Q_{1,(50)}}{Q_{1,(75)} - Q_{1,(25)}} \tag{B1}$$

Appendix C Conformance verification results

For each sub-network $\mathbb{N}_{fi}$, we performed the conformance analyses over two regions:

1. *Training region*: This region considers minimum and maximum test inputs values within the distribution of values used for training.
2. *Generalization region*: This region considers values outside the training distribution to demonstrate model generalization.For both experiments, we set δ as 10 times the minimum change possible of the input to be tested. It is noteworthy that for some inputs tested, we had to test two properties. The reason is due to scaling: because we rescaled the values of the states using a robust scaler, we had both negative and positive values,

Table 7 Conformance verification results of each sub-network of the population-level model ($T1DSim^{P}_{AI}$)

Neural network	Input tested	Property tested	$\Delta x^{i}_{critical}$	
			Training region	Generalization region
Q1 [mg/dL]	X1	↓	✓	$1e^{-12}$
		↑	$\mathbf{2e^{-2}}$	$5e^{-3}$
	X3	↓	$2e^{-4}$	✓
	Q1	↓	$8e^{-6}$	$2e^{-5}$
		↑	$4e^{-5}$	✓
	Q2	↑	$6e^{-13}$	✓
	C2	↑	$1e^{-13}$	✓
Q2 [mmol/kg]	X1	↑	$6e^{-6}$	✓
		↓	$9e^{-6}$	$1e^{-5}$
	X2	↓	$1e^{-5}$	$\mathbf{2e^{-5}}$
		↑	$\mathbf{2e^{-5}}$	$2e^{-6}$
	Q1	↑	✓	✓
		↓	✓	$2e^{-7}$
	Q2	↓	✓	✓
		↑	$3e^{-7}$	$4e^{-7}$
S1 [mU/kg]	S1	↓	$1e^{-5}$	✓
	u_I	↑	$\mathbf{1e^{-2}}$	✓
S2 [mU/kg]	S1	↑	✓	$\mathbf{9e^{-6}}$
	S2	↓	✓	✓
I [mU/L]	S2	↑	✓	✓
	I	↓	✓	$\mathbf{2e^{-5}}$
X1 [min^{-1}]	I	↑	$7e^{-19}$	✓
	X1	↓	✓	$\mathbf{5e^{-9}}$
X2 [min^{-1}]	I	↑	✓	✓
	X2	↓	✓	✓
X3 [unitless]	I	↑	✓	✓
	X3	↓	$1e^{-7}$	$\mathbf{4e^{-7}}$
C2 [mmol/kg/min]	C1	↓	✓	$7e^{-12}$
	C2	↑	$9e^{-9}$	$\mathbf{2e^{-8}}$
C1 [-]	C1	↓	✓	✓
	u_{carbs}	↑	✓	✓

which affected the properties for ODEs where some components are defined as the interactions between two states (e.g., the interaction $-X_1(t)Q_1(t)$). Table 7 shows the detailed results of the conformance verification for each sub-network of the population-level model ($T1DSim^{P}_{NN}$).

Author contributions V.R.E: Conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing—original draft, visualization. T.K: Conceptualization, methodology, software, validation, formal analysis, investigation, writing—review & editing. P.G.J: Conceptualization, formal analysis, investigation, writing—review & editing. C.M.L: Conceptualization, methodology, validation, formal analysis, investigation, resources, writing—review & editing, supervision, project administration, funding acquisition.

Funding This research was funded by Breakthrough T1D, formerly JDRF (grant 2-SRA-2022-1273-S-B).

Data availability The code developed in this research can be found in the following repository: https://github.com/mosqueralopez/T1DSim_AI. The T1DEXI dataset is available for research purposes at https://doi.org/10.25934/PR00008428.

Declarations

Conflict of interest P.G.J. reports advisory board participation and research support from Eli Lilly and Dexcom. P.G.J. has a financial interest in Pacific Diabetes Technologies, a company that might have a commercial interest in the results of this research and technology. All other authors declare no conflicts of interest.

References

1. Katsarou A, Gudbjörnsdottir S, Rawshani A, Dabelea D, Bonifacio E, Anderson BJ, Jacobsen LM, Schatz DA, Lernmark Å (2017) Type 1 diabetes mellitus. Nat Rev Dis Primers 3(11):1–17. https://doi.org/10.1038/nrdp.2017.16
2. Visentin R, Dalla Man C, Kovatchev B, Cobelli C (2014) The university of virginia/padova type 1 diabetes simulator matches the glucose traces of a clinical trial. Diabetes Technol Ther 16(7):428–434. https://doi.org/10.1089/dia.2013.0377
3. Hassan K, Loar R, Anderson BJ, Heptulla RA (2006) The role of socioeconomic status, depression, quality of life, and glycemic control in type 1 diabetes mellitus. J Pediatr 149(4):526–531. https://doi.org/10.1016/j.jpeds.2006.05.039
4. Wellen KE, Hotamisligil GS (2005) Inflammation, stress, and diabetes. J Clin Investig 115(5):1111–1119. https://doi.org/10.1172/JCI25102
5. Castle JR, Jacobs PG (2016) Nonadjunctive use of continuous glucose monitoring for diabetes treatment decisions. J Diabetes Sci Technol 10(5):1169–1173. https://doi.org/10.1177/1932296816631569
6. Tyler NS, Jacobs PG (2020) Artificial intelligence in decision support systems for type 1 diabetes. Sensors 20(1111):3214. https://doi.org/10.3390/s20113214
7. Tyler NS, Mosquera-Lopez CM, Wilson LM, Dodier RH, Branigan DL, Gabo VB, Guillot FH, Hilts WW, Youssef JE, Castle JR, Jacobs PG (2020) An artificial intelligence decision support system for the management of type 1 diabetes. Nat Metab 2(7):612–619. https://doi.org/10.1038/s42255-020-0212-y
8. Breton M, Farret A, Bruttomesso D, Anderson S, Magni L, Patek S, Dalla Man C, Place J, Demartini S, Del Favero S, Toffanin C, Hughes-Karvetski C, Dassau E, Zisser H, Doyle I, Francis J, De Nicolao G, Avogaro A, Cobelli C, Renard E, Kovatchev B (2012) Study group: fully integrated artificial pancreas in type 1 diabetes: modular closed-loop glucose control maintains near normoglycemia. Diabetes 61(9):2230–2237. https://doi.org/10.2337/db11-1445 (https://diabetesjournals.org/diabetes/article-pdf/61/9/2230/561198/2230.pdf)
9. Wilson LM, Jacobs PG, Riddell MC, Zaharieva DP, Castle JR (2022) Opportunities and challenges in closed-loop systems in type 1 diabetes. Lancet Diabetes Endocrinol 10(1):6–8. https://doi.org/10.1016/S2213-8587(21)00289-8
10. Mujahid O, Contreras I, Beneyto A, Vehi J (2024) Generative deep learning for the development of a type 1 diabetes simulator. Commun Med 4(1):1–13. https://doi.org/10.1038/s43856-024-00476-0
11. Mosquera-Lopez C, Jacobs PG (2024) Digital twins and artificial intelligence in metabolic disease research. Trends Endocrinol Metab 35(6):549–557. https://doi.org/10.1016/j.tem.2024.04.019
12. An G, Cockrell C (2022) Drug development digital twins for drug discovery, testing and repurposing: a schema for requirements and development. Front Syst Biol 2:928387. https://doi.org/10.3389/fsysb.2022.928387
13. Bordukova M, Makarov N, Rodriguez-Esteban R, Schmich F, Menden MP (2024) Generative artificial intelligence empowers digital twins in drug discovery and clinical trials. Expert Opin Drug Discov 19(1): 33–42. https://doi.org/10.1080/17460441.2023.2273839
14. Li X, Lee EJ, Lilja S, Loscalzo J, Schäfer S, Smelik M, Strobl MR, Sysoev O, Wang H, Zhang H, Zhao Y, Gawel DR, Bohle B, Benson M (2022) A dynamic single cell-based framework for digital twins to prioritize disease genes and drug targets. Genome Med 14(1):48. https://doi.org/10.1186/s13073-022-01048-4
15. Wang T, Dremel J, Richter S, Polanski W, Uckermann O, Eyüpoglu I, Czarske JW, Kuschmierz R (2024) Resolution-enhanced multi-core fiber imaging learned on a digital twin for cancer diagnosis. Neurophotonics 11(S1):11505. https://doi.org/10.1117/1.NPh.11.S1.S11505
16. Lepper AGW, Buck CMA, Veer M, Huberts W, Vosse FN, Dekker LRC (2022) From evidence-based medicine to digital twin technology for predicting ventricular tachycardia in ischaemic cardiomyopathy. J R Soc Interface 19(194):20220317. https://doi.org/10.1098/rsif.2022.0317
17. Bahrami F, Rossi RM, De Nys K, Defraeye T (2023) An individualized digital twin of a patient for transdermal fentanyl therapy for chronic pain management. Drug Deliv Transl Res 13(9):2272–2285. https://doi.org/10.1007/s13346-023-01305-y
18. Young G, Dodier R, Youssef JE, Castle JR, Wilson L, Riddell MC, Jacobs PG (2024) Design and in silico evaluation of an exercise decision support system using digital twin models. J Diabetes Sci Technol 18(2):324–334. https://doi.org/10.1177/19322968231223217
19. Resalat N, El Youssef J, Tyler N, Castle J, Jacobs PG (2019) A statistical virtual patient population for the glucoregulatory system in type 1 diabetes with integrated exercise model. Plos One 14(7). https://doi.org/10.1371/journal.pone.0217301

20. Rovati L, Gary PJ, Cubro E, Dong Y, Kilickaya O, Schulte PJ, Zhong X, Wörster M, Kelm DJ, Gajic O, Niven AS, Lal A (2023) Development and usability testing of a patient digital twin for critical care education: a mixed methods study. Front Med 10:1336897. https://doi.org/10.3389/fmed.2023.1336897
21. Casola L (ed) (2023) Opportunities and challenges for digital twins in biomedical research: proceedings of a workshop–in brief. The national academies collection: reports funded by national institutes of health. National academies press (US), Washington (DC). http://www.ncbi.nlm.nih.gov/books/NBK592664/
22. Vallée A (2023) Digital twin for healthcare systems. Front Digit Health 5: 1253050. https://doi.org/10.3389/fdgth.2023.1253050
23. Sarp S, Kuzlu M, Zhao Y, Gueler O (2023) Digital twin in healthcare: a study for chronic wound management. IEEE J Biomed Health Inform 27(11):5634–5643. https://doi.org/10.1109/JBHI.2023.3299028
24. Man CD, Micheletto F, Lv D, Breton M, Kovatchev B, Cobelli C (2014) The uva/padova type 1 diabetes simulator: new features. J Diabetes Sci Technol 8(1):26–34. https://doi.org/10.1177/1932296813514502
25. Young GM, Jacobs PG, Tyler NS, Nguyen T-TP, Castle JR, Wilson LM, Branigan D, Gabo V, Guillot FH, Riddell MC, El Youssef J (2023) Quantifying insulin-mediated and noninsulin-mediated changes in glucose dynamics during resistance exercise in type 1 diabetes. Am J Physiol-Endocrinol Metab 325(3):192–206. https://doi.org/10.1152/ajpendo.00298.2022
26. Haidar A, Wilinska ME, Graveston JA, Hovorka R (2013) Stochastic virtual population of subjects with type 1 diabetes for the assessment of closed-loop glucose controllers. IEEE Trans Biomed Eng 60(12):3524–3533. https://doi.org/10.1109/TBME.2013.2272736
27. Rashid M, Samadi S, Sevil M, Hajizadeh I, Kolodziej P, Hobbs N, Maloney Z, Brandt R, Feng J, Park M, Quinn L, Cinar A (2019) Simulation software for assessment of nonlinear and adaptive multivariable control algorithms: glucose-insulin dynamics in type 1 diabetes. Comput Chem Eng 130:106565. https://doi.org/10.1016/j.compchemeng.2019.106565
28. Estremera E, Cabrera A, Beneyto A, Vehi J (2022) A simulator with realistic and challenging scenarios for virtual t1d patients undergoing CSII and MDI therapy. J Biomed Inform 132:104141. https://doi.org/10.1016/j.jbi.2022.104141
29. Hovorka R, Canonico V, Chassin LJ, Haueter U, Massi-Benedetti M, Orsini Federici M, Pieber TR, Schaller HC, Schaupp L, Vering T, Wilinska ME (2004) Nonlinear model predictive control of glucose concentration in subjects with type 1 diabetes. Physiol Meas 25(4):905–920. https://doi.org/10.1088/0967-3334/25/4/010
30. Wilinska ME, Chassin LJ, Acerini CL, Allen JM, Dunger DB, Hovorka R (2010) Simulation environment to evaluate closed-loop insulin delivery systems in type 1 diabetes. J Diabetes Sci Technol 4(1):132–144
31. Cappon G, Vettoretti M, Sparacino G, Del Favero S, Facchinetti A (2023) Replaybg: a digital twin-based methodology to identify a personalized model from type 1 diabetes data and simulate glucose concentrations to assess alternative therapies. IEEE Trans Biomed Eng 70(11):3227–3238. https://doi.org/10.1109/TBME.2023.3286856
32. Young GM (2023) Exercise physiology in type 1 diabetes: development of metabolic models and decision support systems. PhD thesis, Ph.D. https://doi.org/10.6083/js956g59c. http://digitalcollections.ohsu.edu/record/10118
33. Kushner T, Sankaranarayanan S, Breton M (2020) Conformance verification for neural network models of glucose-insulin dynamics. In: Proceedings of the 23rd international conference on hybrid systems: computation and control. HSCC '20, pp 1–12. Association for computing machinery, New York, NY, USA. https://doi.org/10.1145/3365365.3382210
34. Erge O, Oort E (2022) Combining physics-based and data-driven modeling in well construction: hybrid fluid dynamics modeling. J Nat Gas Sci Eng 97:104348. https://doi.org/10.1016/j.jngse.2021.104348
35. Prendin F, Pavan J, Cappon G, Del Favero S, Sparacino G, Facchinetti A (2023) The importance of interpreting machine learning models for blood glucose prediction in diabetes: an analysis using shap. Sci Rep 13(1):16865. https://doi.org/10.1038/s41598-023-44155-x
36. Rai R, Sahu CK (2020) Driven by data or derived through physics? A review of hybrid physics guided machine learning techniques with cyber-physical system (cps) focus. IEEE Access 8:71050–71073. https://doi.org/10.1109/ACCESS.2020.2987324
37. Wang R, Yu R (2021) Physics-guided deep learning for dynamical systems: a survey. ACM Comput Surv 58(5):1–31. https://doi.org/10.48550/ARXIV.2107.01272
38. Yang J, Wang H, Sheng Y, Lin Y, Yang L (2024) A physics-guided generative ai toolkit for geophysical monitoring. https://doi.org/10.48550/ARXIV.2401.03131
39. Wu D, Gao L, Xiong X, Chinazzi M, Vespignani A, Ma Y-A, Yu R (2021) Deepgleam: a hybrid mechanistic and deep learning model for covid-19 forecasting. https://doi.org/10.48550/ARXIV.2102.06684
40. Wang R, Kashinath K, Mustafa M, Albert A, Yu R (2020) Towards physics-informed deep learning for turbulent flow prediction. https://doi.org/10.48550/ARXIV.1911.08655
41. Roquemen-Echeverri V, Mosquera-Lopez C (2025) Recent advancements and applications of physics-informed machine learning in biomedical research. Curr Opin Biomed Eng 35:100612. https://doi.org/10.1016/j.cobme.2025.100612
42. Zarkogianni K, Litsa E, Vazeou A, Nikita KS (2013) Personalized glucose-insulin metabolism model based on self-organizing maps for patients with type 1 diabetes mellitus. In: 13th IEEE international conference on bioinformatics and bioengineering, pp 1–4. https://doi.org/10.1109/BIBE.2013.6701604. https://ieeexplore.ieee.org/document/6701604

43. Contreras I, Oviedo S, Vettoretti M, Visentin R, Vehí J (2017) Personalized blood glucose prediction: a hybrid approach using grammatical evolution and physiological models. PLoS ONE 12(11):0187754. https://doi.org/10.1371/journal.pone.0187754
44. Balakrishnan NP, Samavedham L, Rangaiah GP (2013) Personalized hybrid models for exercise, meal, and insulin interventions in type 1 diabetic children and adolescents. Ind Eng Chem Res 52(36):13020–13033. https://doi.org/10.1021/ie402531k
45. Zou BJ, Levine ME, Zaharieva DP, Johari R, Fox EB (2024) Hybrid square neural ode causal modeling. https://doi.org/10.48550/arXiv.2402.17233. arXiv:2402.17233 [cs, stat]
46. Colmegna P, Wang K, Garcia-Tirado J, Breton MD (2020) Mapping data to virtual patients in type 1 diabetes. Control Eng Pract 103:104605. https://doi.org/10.1016/j.conengprac.2020.104605
47. Deichmann J, Bachmann S, Burckhardt M-A, Pfister M, Szinnai G, Kaltenbach H-M (2023) New model of glucose-insulin regulation characterizes effects of physical activity and facilitates personalized treatment evaluation in children and adults with type 1 diabetes. PLoS Comput Biol 19(2):1010289. https://doi.org/10.1371/journal.pcbi.1010289
48. Goodwin GC, Seron MM, Medioli AM, Smith T, King BR, Smart CE (2020) A systematic stochastic design strategy achieving an optimal tradeoff between peak BGL and probability of hypoglycaemic events for individuals having type 1 diabetes mellitus. Biomed Signal Process Control 57:101813. https://doi.org/10.1016/j.bspc.2019.101813
49. Hughes J, Gautier T, Colmegna P, Fabris C, Breton MD (2021) Replay simulations with personalized metabolic model for treatment design and evaluation in type 1 diabetes. J Diabetes Sci Technol 15(6):1326–1336. https://doi.org/10.1177/1932296820973193
50. Visentin R, Man CD, Cobelli C (2016) One-day bayesian cloning of type 1 diabetes subjects: toward a single-day UVA/Padova type 1 diabetes simulator. IEEE Trans Biomed Eng 63(11):2416–2424. https://doi.org/10.1109/TBME.2016.2535241
51. Riddell MC, Li Z, Beck RW, Gal RL, Jacobs PG, Castle JR, Gillingham MB, Clements M, Patton SR, Dassau E, Doyle Iii FJ, Martin CK, Calhoun P, Rickels MR (2021) More time in glucose range during exercise days than sedentary days in adults living with type 1 diabetes. Diabetes Technol Ther 23(5):376–383. https://doi.org/10.1089/dia.2020.0495
52. Riddell MC, Li Z, Gal RL, Calhoun P, Jacobs PG, Clements MA, Martin CK, Doyle FJ III, Patton SR, Castle JR, Gillingham MB, Beck RW, Rickels MR (2023) Examining the acute glycemic effects of different types of structured exercise sessions in type 1 diabetes in a real-world setting: the type 1 diabetes and exercise initiative (t1dexi). Diabetes Care 46(4):704–713. https://doi.org/10.2337/dc22-1721
53. Hovorka R, Canonico V, Chassin LJ, Haueter U, Massi-Benedetti M, Frederici MO, Pieber TR, Shaller HC, Schaupp L, Vering T, Wilinska ME (2004) Nonlinear model predictive control of glucose concentration in subjects with type 1 diabetes. Physiol Meas 25:905–920
54. Davidson PC, Hebblewhite HR, Steed RD, Bode BW (2008) Analysis of guidelines for basal-bolus insulin dosing: basal insulin, correction factor, and carbohydrate-to-insulin ratio. Endocr Pract 14(9):1095–1101. https://doi.org/10.4158/EP.14.9.1095
55. Forgione M, Piga D (2021) Continuous-time system identification with neural networks: model structures and fitting criteria. Eur J Control 59:69–81. https://doi.org/10.1016/j.ejcon.2021.01.008
56. Del Favero S, Facchinetti A, Cobelli C (2012) A glucose-specific metric to assess predictors and identify models. IEEE Trans Biomed Eng 59(5):1281–1290. https://doi.org/10.1109/TBME.2012.2185234
57. Cryer PE (2010) Hypoglycemia in type 1 diabetes mellitus. Endocrinol Metab Clin 39(3):641–654. https://doi.org/10.1016/j.ecl.2010.05.003
58. Kitabchi AE, Umpierrez GE, Miles JM, Fisher JN (2009) Hyperglycemic crises in adult patients with diabetes. Diabetes Care 32(7):1335–1343. https://doi.org/10.2337/dc09-9032
59. Kovatchev BP (2017) Metrics for glycaemic control: from hba1c to continuous glucose monitoring. Nat Rev Endocrinol 13(7):425–436. https://doi.org/10.1038/nrendo.2017.3
60. Kingma DP, Ba J (2017) Adam: a method for stochastic optimization. (arXiv:1412.6980) https://doi.org/10.48550/arXiv.1412.6980
61. Nogueira F (2014) Bayesian optimization: open source constrained global optimization tool for Python. GitHub
62. Gurobi optimization I (2016) Gurobi optimizer reference manual. http://www.gurobi.com
63. Mosquera-Lopez C, Wilson LM, El Youssef J, Hilts W, Leitschuh J, Branigan D, Gabo V, Eom JH, Castle JR, Jacobs PG (2023) Enabling fully automated insulin delivery through meal detection and size estimation using artificial intelligence. NPJ Digital Med 6(1):1–7. https://doi.org/10.1038/s41746-023-00783-1
64. Herrero P, Reddy M, Georgiou P, Oliver NS (2022) Identifying continuous glucose monitoring data using machine learning. Diabetes Technol Ther 24(6):403–408. https://doi.org/10.1089/dia.2021.0498
65. Wilmot EG, Choudhary P, Leelarathna L, Baxter M (2019) Glycaemic variability: the under-recognized therapeutic target in type 1 diabetes care. Diabetes Obes Metab 21(12):2599–2608. https://doi.org/10.1111/dom.13842
66. Cappon G, Facchinetti A (2024) Digital twins in type 1 diabetes: a systematic review. J Diabetes Sci Technol 19322968241262112. https://doi.org/10.1177/19322968241262112

67. ...Battelino T, Danne T, Bergenstal RM, Amiel SA, Beck R, Biester T, Bosi E, Buckingham BA, Cefalu WT, Close KL, Cobelli C, Dassau E, DeVries JH, Donaghue KC, Dovc K, Doyle I, Francis J, Garg S, Grunberger G, Heller S, Heinemann L, Hirsch IB, Hovorka R, Jia W, Kordonouri O, Kovatchev B, Kowalski A, Laffel L, Levine B, Mayorov A, Mathieu C, Murphy HR, Nimri R, Nørgaard K, Parkin CG, Renard E, Rodbard D, Saboo B, Schatz D, Stoner K, Urakami T, Weinzimer SA, Phillip M (2019) Clinical targets for continuous glucose monitoring data interpretation: recommendations from the international consensus on time in range. Diabetes Care 42(8):1593–1603. https://doi.org/10.2337/dci19-0028
68. ...Battelino T, Alexander CM, Amiel SA, Arreaza-Rubin G, Beck RW, Bergenstal RM, Buckingham BA, Carroll J, Ceriello A, Chow E, Choudhary P, Close K, Danne T, Dutta S, Gabbay R, Garg S, Heverly J, Hirsch IB, Kader T, Kenney J, Kovatchev B, Laffel L, Maahs D, Mathieu C, Mauricio D, Nimri R, Nishimura R, Scharf M, Del Prato S, Renard E, Rosenstock J, Saboo B, Ueki K, Umpierrez GE, Weinzimer SA, Phillip M (2023) Continuous glucose monitoring and metrics for clinical trials: an international consensus statement. Lancet Diabetes Endocrinol 11(1):42–57. https://doi.org/10.1016/S2213-8587(22)00319-9
69. Villa-Tamayo MF, Colmegna P, Breton M (2024) Validation of the UVA simulation replay methodology using clinical data: reproducing a randomized clinical trial. Diabetes Technol Ther 26(10):720–727. https://doi.org/10.1089/dia.2023.0595
70. Kushner T, Breton MD, Sankaranarayanan S (2020) Multi-hour blood glucose prediction in type 1 diabetes: a patient-specific approach using shallow neural network models. Diabetes Technol Ther 22(12):883–891. https://doi.org/10.1089/dia.2020.0061

Authors and Affiliations

Valentina Roquemen-Echeverri[1] · Taisa Kushner[2] · Peter G. Jacobs[1,3] · Clara Mosquera-Lopez[1]

✉ Valentina Roquemen-Echeverri
roquemev@ohsu.edu

✉ Clara Mosquera-Lopez
mosquera@ohsu.edu

[1] Artificial Intelligence for Medical Systems (AIMS) Lab, Department of Biomedical Engineering, Oregon Health and Science University, 3303 S. Bond Avenue, Portland, OR 97239, USA

[2] Galois, Inc., 421 SW 6th Avenue #300, Portland, OR 97204, USA

[3] School of Chemical, Biological and Environmental Engineering, Oregon State University, 1500 SW Jefferson Way, Corvallis, OR 97331, USA