Inducing Causal Structure for Interpretable Neural Networks Applied to Glucose Prediction for T1DM Patients

Ana Esponera[1] ANAESPONERA@GMAIL.COM

Giovanni Cinà[1 2] G.CINA@AMSTERDAMUMC.NL

[1] *Medical Informatics Dept., Amsterdam University Medical Center, The Netherlands*
[2] *Institute for Logic, Language and Computation, University of Amsterdam, The Netherlands*

Editors: Biwei Huang and Mathias Drton

Abstract

Causal abstraction techniques such as Interchange Intervention Training (IIT) have been proposed to infuse neural network with expert knowledge encoded in causal models, but their application to real-world problems remains limited. This article explores the application of IIT in predicting blood glucose levels in Type 1 Diabetes Mellitus (T1DM) patients. The study utilizes an acyclic version of the *simglucose* simulator approved by the FDA to train a Multi-Layer Perceptron (MLP) model, employing IIT to impose causal relationships. Results show that the model trained with IIT effectively abstracted the causal structure and outperformed the standard one in terms of predictive performance across different prediction horizons (PHs) post-meal. Furthermore, the breakdown of the counterfactual loss can be leveraged to explain which part of the causal mechanisms are more or less effectively captured by the model. These preliminary results suggest the potential of IIT in enhancing predictive models in healthcare by effectively complying with expert knowledge.

Keywords: interchange intervention, causality, glucose prediction, interpretability, diabetes

1. Introduction

Despite the great amount of machine learning models developed in both academia and industry, there is often uncertainty regarding their compliance with the underlying causal structures of the problems they aim to solve, which might be problematic in high-stake scenarios. This misalignment becomes particularly concerning when models are deployed in critical applications, such as medical prediction, where compliance with causal mechanisms is essential for reliability and safety. Recent research emphasizes the need for integrating causal reasoning into predictive models for decision support to enhance their robustness and interpretability (van Geloven et al., 2024; van Amsterdam et al., 2022, 2024) . Causal models, specifically structural causal models (SCMs), play a pivotal role in this integration as they encode domain knowledge and illustrate how different variables influence one another. When the causal structure—represented as a DAG (Pearl, 2009) —and the corresponding causal mechanisms in a SCM are known, causal abstraction provides mathematical tools to test whether a given machine learning model aligns with this SCM. Causal abstraction is defined as the alignment of a low-level model's behaviour (e.g., a neural network) with the causal structure of a high-level model. By simplifying a complex system and focusing on the key causal relationships that drive outcomes, causal abstraction defines when a human-comprehensible, high-level causal explanation accurately represents opaque low-level details of a deep learning model

© 2025 A. Esponera & G. Cinà.

(Schölkopf et al., 2021). When this alignment is successful, models can reason more effectively, avoid spurious correlations, and comply with pre-existing knowledge.

An essential aspect of causal reasoning is counterfactual reasoning, which involves evaluating "what-if" scenarios by manipulating specific variables and observing the resulting changes in outcomes. IIT is a training method based on counterfactual behaviour optimization, designed to impose causal structures on neural networks (Geiger et al., 2022). While this technique was demonstrated on simple problems such as visual recognition and natural language processing tasks, it remains to be shown whether it can make a difference in a real-world setting.

One such use cases is the management of T1DM. T1DM is a prevalent chronic metabolic condition characterized by the insufficient absorption of glucose by the body's cells, leading to elevated levels of blood glucose (BG) (Dilworth, 2021; American Diabetes Association, 2009; Aloke, 2022). This condition results from the lack of insulin production, requiring external insulin administration which affects 8.75 million patients worldwide (Federation, 2022) and is estimated to cost 966 billion U.S. dollars globally in healthcare expenditures (Statista, 2024). BG prediction poses unique challenges due to the complex and dynamic nature of glucose metabolism, influenced by numerous factors such as food intake, physical activity, and insulin administration. Accurately predicting BG levels is essential for effective T1DM management, yet existing simulators may be computationally intensive and unsuitable for integration into lightweight, wearable medical devices, such as insulin pumps (Pereira et al., 2024; Nahavandi et al., 2022). A neural network that can serve as a computationally efficient, causal abstraction of BG prediction could enable real-time, on-device predictions, potentially improving T1DM management and reducing healthcare burdens.

In this paper, we apply IIT to the prediction of BG in T1DM patients. Our primary objective is to demonstrate the applicability of IIT in a real-world context, specifically in enhancing predictive models within the healthcare domain by leveraging expert knowledge encoded in causal models. We detail the improvements of IIT-trained models over standard models in terms of (a) model performance, (b) data efficiency, and (c) compliance with expert knowledge. Finally, we discuss the challenges encountered when integrating this technology into real-world healthcare scenarios.

2. Preliminaries

Interchange intervention training is a technique designed to inject causal structure into neural network models (NN) by leveraging counterfactual reasoning. The goal is to guide the model towards more robust learning by focusing on the underlying causal logic of the task, rather than merely capturing surface correlations. To this end, it is essential to establish a mapping between the causal DAG and the architecture of the neural network. This mapping specifies which parts of the neural network correspond to specific nodes in the DAG, effectively assigning roles within the network to components of the causal model. Different mappings can yield different results, as they influence how the high-level causal structure is embedded within the low-level computations of the NN. Exploiting this mapping, an interchange intervention involves swapping part of the input value while the rest stays constant, and then comparing the outcomes under both scenarios. Geiger et al. (2022) defines an interchange intervention as a model used to process two different inputs ($source$ and $base$) and then a particular internal state obtained by processing $source$ is used in place of the corresponding internal state obtained by $base$. This allows researchers to estimate the counterfactual effect of that part of the intervened value by effectively holding other variables constant or "interchanging" them. The process establishes and aligns a proposed causal structure with the model. The alignments ensure that (clusters of) lower-level variables accurately reflect or capture the high-level variables in the SCM. Importantly, these interventions are not limited to the input layer but

can be applied to any intermediate value within the causal structure. The loss in such counterfactual scenario is dubbed L_{INT} (Geiger et al., 2024):

$$L_{INT} = \sum_{b,s \in in} Loss(M_{H_{\tau(I \leftarrow_{b,s})}}, M_{L_{I \leftarrow_{b,s}}}) \tag{1}$$

where b and s are the $base$ and $source$ input values from the in input space swapped during the intervention, I is the variable being intervened on, $I \leftarrow b, s$ is the intervention mechanism, M_H is the high-level neural model, M_L is the low-level causal model, τ is a mapping of output values from the low to high-level and $Loss$ is the chosen function to quantify the distance. If the L_{INT} is reduced to zero, we can guarantee that the target causal model serves as a causal abstraction of the neural model (Geiger et al., 2022). Crucially, this procedure does not guarantees correctness when the starting knowledge is flawed: if the SCM is encoding incorrect knowledge then IIT will align the low-level model to the ill-specified SCM. By training the model with IIT and encouraging it to reduce L_{INT}, the desired causal structure can be effectively imposed on a neural network. Since interventions are local, i.e. they pertain to a subpart of the SCM, IIT also allows for a breakdown of which components of the SCM are successfully abstracted by the NN, adding to the interpretability of the model.

3. Related work

Geiger et al. (2022) introduce IIT and evaluate it across three tasks: a structural vision task (MNIST-PVR), a navigational language task (ReaSCAN), and a natural language inference task (MQNLI). They demonstrate that IIT outperforms multi-task training and data augmentation, yielding more interpretable neural models aligned with the intended causal structure. In the context of model distillation, Wu et al. (2022) apply IIT to BERT, achieving significant improvements. Specifically, IIT reduces perplexity on masked language modeling tasks and enhances performance on benchmarks such as GLUE, SQuAD, and CoNLL-2003. Furthermore, Huang et al. (2023) propose Type-level Interchange Intervention Training (TIIT) to induce character-level structures in subword-based models, improving robustness and the model's understanding of subword dependencies. As the most recent application of IIT, Gupta et al. (2024) present a framework that leverages IIT with known circuits for evaluating mechanistic interpretability methods. Their experimental results demonstrate that this approach not only enhances performance on challenging benchmarks but also provides deeper insights into the internal mechanisms driving model predictions.

In the field of machine learning for BG prediction in T1DM patients, (Liu et al., 2023) provide a meta-analysis showing that NN exhibit the highest performance across different prediction horizons (PH). Regarding models that incorporate glucose-insulin dynamics, we have specifically chosen to focus on the works of Karim et al. (2020) and Liu et al. (2019) because both report clear metrics that allow for direct comparison, employ the same PHs and comply with GLYFE benchmark (Bois et al., 2022). GLYFE establishes a reproducible benchmark for evaluating diverse machine learning models for personalized glucose forecasting in T1DM across both simulated, namely UVA/Padova Type 1 Diabetes Metabolic Simulator, and real patient dataset Ohio Type-1 Diabetes Mellitus. Karim et al. (2020) introduce a novel BG prediction method using an absorption model, achieving superior performance for 60- and 120-minute PHs. Finally, Liu et al. (2019) present a deconvolutional model based on glucose-insulin dynamics, outperforming conventional models for PHs beyond 60 minutes.

4. Data and Methods

4.1. Experimental set up and metrics

The predictions were evaluated on the 30, 45, 60 and 120-minute PH, following the patients' first meal of the day, standardized as breakfast. Patients were assumed not to consume any extra food between this meal and the subsequent glucose measurement, and were all treated with a specific dose of insulin. Glucose levels, which are continuous measurements typically ranging from 40 to 400 mg/dL in diabetic patients, were monitored throughout the simulation. The performance metrics used in the evaluation are: average absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE). Since not all prediction errors are the same when it comes to vital signs, we included as metric the percentage of predicted values in the "clinically acceptable" EGA classes A and B (Clarke et al., 1987). EGA grids can be used as an indicator of underfitting in a model, for instance when the predictions display a horizontal trend. Detailed information regarding the training configuration and evaluation metrics is reported in Appendix B.

4.2. Data acquisition, preprocessing and setting

The data used in this study is based on in-silico data from Xie (2018). The entire dataset consists of 30 T1DM patients (10 children, 10 adolescents, 10 adults) with characteristics derived from three joint parameter distributions. Each virtual patient is described by the patient's initial state x, multiple kinetics constants k, an action representing carbohydrate intake CHO and insulin dosage $insulin$ (see Fig. 1). In total, 64 parameters are provided per patient as a subject model parameter vector. As the provided number of patients is limited, we generate a specific number of observations comparable to existing cohorts used in previous studies in the literature (Liu et al., 2019). Therefore, 200 T1DM subjects were generated using Gaussian distributions conditioned by age from the FDA-approved in-silico joint distributions (7 children, 10 adolescents, 183 adults). More details about UVA/Padova T1DM joint distributions can be found in Xie (2018); Man et al. (2009) and more details about the data derivation can be found in Appendix A.

Out of those 64 parameters, only 9 were selected for input to the model. The parameter vector consists of 51 pharmacokinetic constants and 13 patient state variables at the initial time t_n. The 51 pharmacokinetic constants were excluded from the input dataset because they are fixed for each patient and our objective is to develop a model that generalizes across patients without this fine-grained information. While the 13 patient state variables are patient-specific, they represent dynamic states that could in principle be available for the model. In practice, the model is only fed with the information that can reasonably be measured, allowing it to infer the remainder. Four patient state parameters, x_1, x_2, x_3, and x_7, were also excluded as they were initialized to 0 for all patients. They represent the amount of glucose and insulin at the initial time t_n which is negligible for the breakfast scenario. The final input data comprised 20 parameters: 9 from the patient's initial state, 9 from pre-meal glucose levels, and 2 for insulin and CHO intake. Preprocessing included Z score standardization for patient's state parameters, and glucose levels were scaled at a 100:1 ratio to improve training stability without distorting relationships between values. The data was split 80/20, with 160 patients in the training set and 40 in the validation set. The test set only included the original 30 FDA-approved in-silico patients to evaluate model generalization to FDA-approved data.

4.3. The *Simglucose* simulator

To use IIT, a causal model is required. We selected the UVA/Padova T1DM simulator, *simglucose*, a collaborative effort between the University of Virginia and the University of Padova. It simulates physiological processes and glucose-insulin dynamics in T1DM patients, serving as a substitute for preclinical trials of insulin treatments, including closed-loop algorithms.

Three versions of the simulator have been released so far: S2008, S2013 and S2017. Each version incorporates an improved modelling of glucose-insulin interactions. Visentin et al. (2014) detected that the incidence of hypoglycemic events projected by the S2008 simulator did not entirely align with those recorded in clinical trials. Therefore, the S2013 version incorporated a new insulin-dependent compartment that improves the model of glucose kinetics in hypoglycemia. Later on, the S2017 release improved the intra-day glucose variability and the nocturnal glucose increase. The three versions have been validated and accepted by FDA. However, S2013 and S2017 official implementations in *MatLab* do not have an academic license. For this reason, we have chosen the S2008 version, which is open source for Matlab and Python and is summarized in Fig 1. Beside the limitation in hypoglycemic scenarios, the S2008 simulator faithfully represents glucose dynamics in T1DM patients for euglycemic and hyperglycemic scenarios. More details on the model are available in Xie (2018); Man et al. (2009). At its full complexity, the *simglucose* causal model for PH_n corresponds to a DAG unfolding through n steps in time. In order to obtain alignment with a non-recurrent NN - which effectively predicts only the BG at the PH in one leap - we elected to simplify this DAG compressing the time dimension. This operation produced two cyclic relationships between equations dx_4dt/dx_5dt and $dx_6dt/dx_{10}dt$; acyclicity was recovered by removing the terms colored in red in the figure. We refer to this as the "amended" *simglucose*. The decision to experiment with a time-compressed causal model was driven by the need for simplicity and the fact that the structural equations themselves are stable over time. While this approach may not fully capture temporal dependencies, it provides a tractable starting point for applying IIT in a healthcare setting.

4.4. NN and alignment to causal model

We use a minimal building block to approximate each node of the DAG representing the causal structure. Each building block consists of a linear layer followed by a leaky ReLU activation function and a dropout layer with a rate of 0.3 (See Fig. 2). By assembling these building blocks, we construct neural network models that approximate the causal structure at varying complexity levels.

Our main model is the **MLP tree architecture**. The architecture of this model establishes connections between the 13 sequential modules, imitating the blueprint of the causal model (See Fig 2(a)). Each sequential module corresponds to one of the patient state parameters from x_1 to x_{13}. For this NN, two versions of this architecture exist: one with $hidden_size$ 128 and another with more flexibility of $hidden_size$ 256.

In this model, each sequential module approximates the function that computes a variable in the structural causal model: variable x_1 from the computational model is aligned to the sequential module $X1$ in the NN, and so on. Because of the compression of the DAG on the time dimension, each module's output corresponds to the value of the corresponding causal variable at the prediction horizon after the meal.

To assess the effect of alternative architectures and the impact of the aforementioned operations on the DAG, we also consider two variations.

Figure 1: *Simglucose* specific kinetic constants and details on the model

Equations:

$$
\begin{aligned}
dx_1dt &= -k_{\text{max}} \cdot x_1 + \text{CHO} \\
dx_2dt &= k_{\text{max}} \cdot x_1 - x_2 \cdot k_{\text{gut}} \\
dx_3dt &= k_{\text{gut}} \cdot x_2 - k_{\text{abs}} \cdot x_3 \\
dx_4dt &= EGP_t + Ra_t - U_{\text{iit}} - E_t - k_1 \cdot x_4 \\
&\quad + k_2 \cdot x_5 \\
dx_5dt &= -U_{\text{idt}} + k_1 \cdot x_4 - k_2 \cdot x_5 \\
dx_6dt &= -(k_{\text{m2}} + k_{\text{m4}}) \cdot x_6 + k_{\text{m1}} \cdot x_{10} \\
&\quad + k_{\text{a1}} \cdot x_{11} + k_{\text{a2}} \cdot x_{12} \\
dx_7dt &= -k_{\text{p2u}} \cdot x_7 + k_{\text{p2u}} \cdot (I_t - k_{\text{Ib}}) \\
dx_8dt &= -k_i \cdot (x_8 - I_t) \\
dx_9dt &= -k_i \cdot (x_9 - x_8) \\
dx_{10}dt &= -(k_{\text{m1}} + k_{\text{m3}}) \cdot x_{10} + k_{\text{m2}} \cdot x_6 \\
dx_{11}dt &= \text{action}_{\text{insulin}} - (k_{\text{a1}} + k_d) \cdot x_{11} \\
dx_{12}dt &= k_d \cdot x_{11} - k_{\text{a2}} \cdot x_{12} \\
dx_{13}dt &= -k_{\text{sc}} \cdot x_{13} + k_{\text{sc}} \cdot x_4
\end{aligned}
$$

Where:

x_1 : Amount of glucose in the stomach (solid-state) (mg)
x_2 : Amount of glucose in the stomach (liquid state) (mg)
x_3 : Glucose mass (GM) in the intestine (mg)
x_4 : GM in plasma and rapidly equilibrating tissues (ET) (mg/kg)
x_5 : GM in tissue and slowly ET (mg/kg)
x_6 : Insulin mass in plasma (pmol/kg)
x_7 : Insulin in the interstitial fluid (pmol/L)
x_8 : Plasma insulin concentration (pmol/L)
x_9 : Delayed insulin (pmol/L)
x_{10} : Insulin mass in the liver (pmol/kg)
x_{11} : Nonmonomeric insulin in subcutaneous subcutaneous space (pmol/kg)
x_{12} : Monomeric insulin in subcutaneous space (pmol/kg)
x_{13} : Subcutaneous glucose level (mg/kg)

Model Variables:

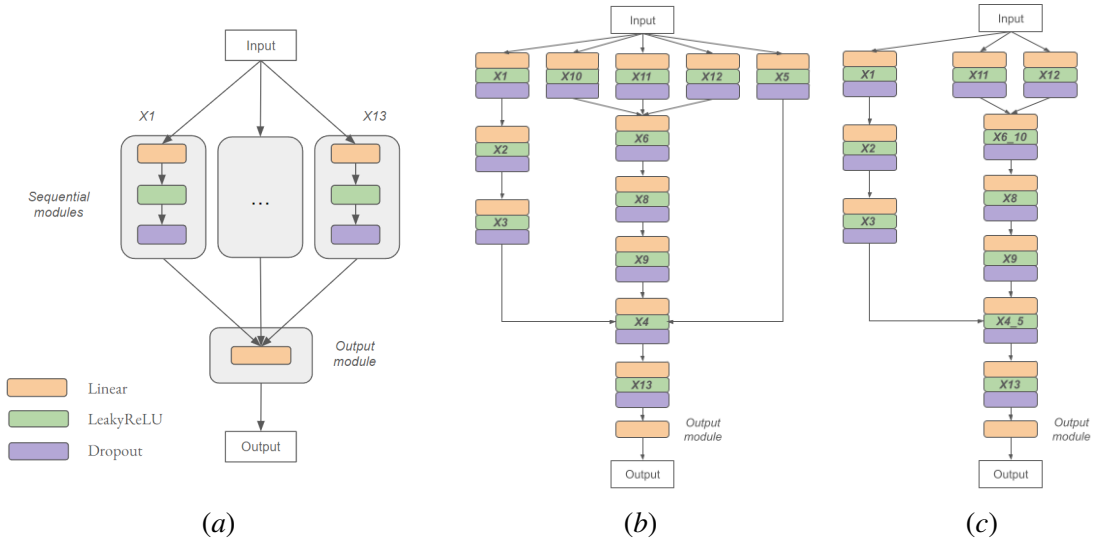
U_{idt} : insulin-dependent glucose utilization (mg/kg/min)
EGP_t : Endogenous glucose production (mg/kg/min)
Ra_t : Glucose rate of appearance in plasma (pmol/kg/min)
U_{iit} : Insulin-independent glucose utilization (mg/kg/min)
E_t : Renal excretion (mg/kg/min)
I_t : Plasma insulin concentration (pmol/liter)
k_{sc} : Amount of nonmonomeric and monomeric insulin in the subcutaneous space (pmol/kg)
k_{Ib} : Plasma insulin concentration at basal state (pmol/liter)

Constants: (min^{-1})

k_{max} : Rate of grinding
k_{gut} : Rate gastric emptying
k_{abs} : Rate intestinal absorption
k_{p2u} : Rate insulin action on peripheral glucose utilization
k_i : Rate parameter accounting for delay between insulin signal and insulin action
k_d : Rate insulin dissociation
k_1, k_2 : Rate parameter of glucose kinetics
$k_{\text{m1}}, k_{\text{m2}},$: Rate parameter of insulin kinetics
$k_{\text{m3}}, k_{\text{m4}}$: Rate parameter of insulin kinetics
k_{a1} : Rate nonmonomeric insulin absorption
k_{a2} : Rate monomeric insulin absorption

The first one is **MLP parallel architecture**. It is composed of 13 identical sequential modules ($X1$ to $X13$) arranged in parallel, with no connections between them (See Fig 2(*b*)). This model does not capture the causal dependencies among the variables and serves as a baseline to evaluate the importance of incorporating the causal structure. The alignment follows the same alignment as in the MLP tree model.

The second one is **MLP joint architecture** . It addresses the cyclic connections in the compressed DAG by merging the related modules (dx_4dt with dx_5dt, and dx_6dt with $dx_{10}dt$) into joint



(*a*) (*b*) (*c*)

Figure 2: NN architecture diagram for (a) MLP parallel, (b) MLP tree, and (c) MLP joint

modules (See Fig 2(*c*)). This model can therefore align with the model described in Figure 1 including the red component, albeit not with an injective mapping. The latter is adjusted as follows: the joint sequential module $X4_5$ is aligned to both x_4 and x_5, and $X6_10$ is aligned to both x_6 and x_{10}. For the rest of the sequential modules, the alignment follows the same alignment as before.

5. Results

5.1. Performance of MLP tree with and without IIT

The box plot in Figure 3 compares RMSE across PHs for the MLP tree model with 256 hidden units, using amended *simglucose* as causal model. Solid-coloured columns represent IIT-trained model, and trace-filled columns represent standard-trained model. Lower RMSE values indicate better predictions. Overall, IIT training consistently achieves lower RMSE values than Standard training across all PHs, highlighting its effectiveness. The median RMSE values are visibly lower for IIT models. This is particularly pronounced at the 30-minute and 45-minute PHs, where IIT training shows significant reductions in RMSE (approximately 16 and 23 mg/dL, respectively), compared to the standard training counterparts, which have higher medians and a wider distribution of errors. At 60 and 120-minute PHs, the RMSE differences between IIT and standard models are less pronounced, but IIT-trained models still maintain a lower error trend. The interquartile ranges are narrower for IIT training, indicating more consistent model performance across multiple runs, whereas standard training tends to have wider distributions, implying greater variability. Detailed values for MSE, MAE and EGA performance are reported in Appendix F. Similarly, the IIT-trained model achieves lower MSE, lower MAE and higher EGA across all PHs compared to standard-trained model. The difference is substantial for PH 60 (MSE reduction = -95.89 $(mg/dL)^2$) and 120 (MSE reduction = -128.34 $(mg/dL)^2$). In addition, Appendix G shows Clarke error grid analysis to visualize predictions within clinically accepted ranges. Only the representative 120 PH grid is

shown. The model MLP tree with 256 as hidden size Fig. 10(*a*), 10(*b*) shows dispersed predictions along the diagonal for both IIT-trained and standard-trained model.

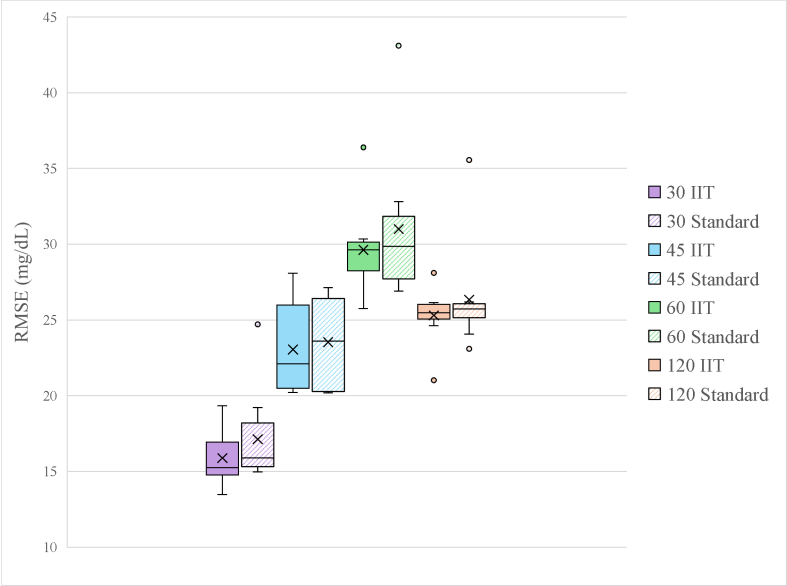


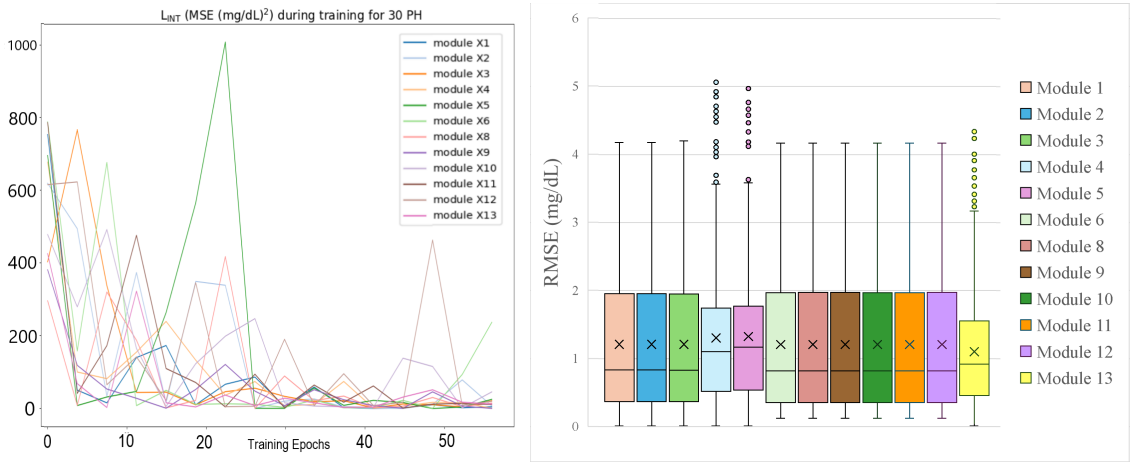
Figure 3: RMSE (mg/dL) prediction error of the model MLP tree (256) IIT using amended *simglucose* across the four prediction horizons (PH) for the test (n=30) in-silico T1DM patients and 10 different random seeds. The solid bars refer to IIT trainings while the striped bars refer to standard training. The mean is indicated by the cross.

5.2. Causal abstraction analysis

We tracked the counterfactual loss (L_{INT}) during the training process across different PHs for the MLP tree model with 256 hidden units. Figure 4(*a*) displays the results for 30 PH, showing that L_{INT} consistently decreases as training epochs progress; the same holds for the other PHs. The corresponding plots are included in Appendix E Fig 6. With more detail, Figure 4(*b*) visualizes the locality, spread, and skewness of L_{INT} for the 30 PH to facilitate analysis of causal abstraction per module. Module $X4$ and module $X5$ tend to have a slightly higher L_{INT} median than the rest of the modules. Modules $X4$ and $X5$ are aligned to the glucose mass in plasma and in rapidly/slowly ET. Similarly, module $X13$ also presents a slightly higher L_{INT} median. This module is aligned to the subcutaneous glucose level. Therefore, the model has shown a lower capacity to abstract the mechanisms for $X4$, $X5$ and $X13$ than the rest of the modules. Similar trends are observed across the other PHs, with detailed plots included in Appendix E Table 5.

5.3. Sensitivity analysis with regular *simglucose*

Figure 5 contrasts the four different models: MLP parallel, MLP tree (hidden sizes 128 and 256), and MLP joint; trained with IIT using regular (i.e. without removal of red terms in Figure 1) *simglucose* as causal model and standard training. Detailed values for MSE, MAE and EGA performance are reported in Appendix C. For Clarke error grid analysis, the model MLP parallel Figs. 8(*a*), 8(*b*)



(*a*) (*b*)

Figure 4: L_{INT} tracking for the MLP tree model for PH 30. The L_{INT} is grouped by modules. a) during training and b) during testing.

show a tendency to predict within the range of 140 and 180, nearly horizontal. On the other hand, the rest of the models show more dispersed predictions along the diagonal. This holds for both IIT-trained and standard-trained models.

Overall, no model outperforms others consistently across all PHs or metrics (lowest MSE, MAE, RMSE, or highest EGA A-B). In addition, none of the four models presents a better performance defined as improvement in all metrics over the four PHs when IIT is applied to the regular *simglucose* with respect to its standard training. Generally, errors increase as PH increases.

5.4. Comparison with previous studies

We performed a comparative analysis with prior research on BG for T1DM employing prediction models that incorporate expert knowledge. The findings are reported in Table 1. For this comparison, we consider the MSE as the only metric due to the unavailability of MAE and the percentage of predictions within zones A or B of the EGA in most previous studies. The best results per PH are emphasized in bold. The table shows in the first row the results obtained from the MLP tree model (with 256 as hidden size) trained through IIT and using the regular *simglucose* as causal model. Although this model is not the best-performing model in this study, it is the best model using the regular *simglucose* and therefore comparable with the published models. Additionally, Appendix D Table 4 shows a more detailed version of Table 1, reporting the results of each individual study that Liu et al. (2023) aggregates in theirs.

6. Discussion

The objective of the experiments is to determine if IIT can be applied in real healthcare use cases that are more complex and present different challenges than those addressed in recent publications. Since the evolution of BG levels in T1DM patients can be modelled via the *simglucose* simulator, we elected to employ the latter as a causal model for the IIT training of a neural network.

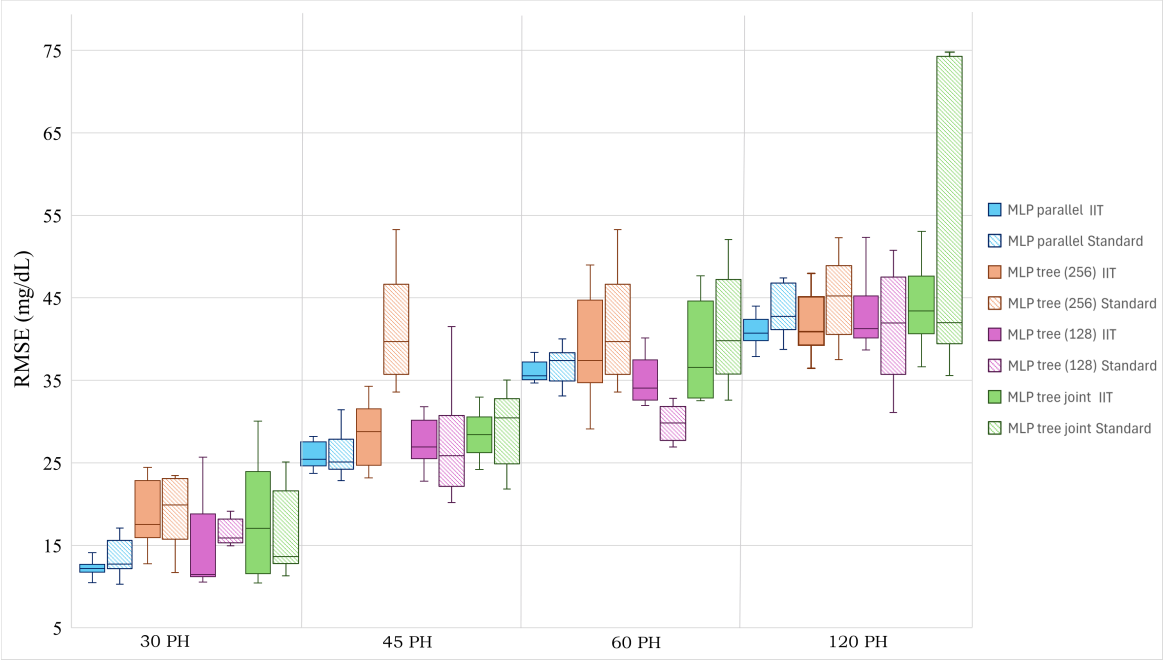


Figure 5: RMSE (mg/dL) prediction error of the different MLP architectures (4 models) across the four PHs for the test (n=30) in-silico T1DM patients. The causal model used is the regular *simglucose* and the boxplots are produced with 10 different random seed. The solid bars refer to the models being trained through interchange intervention training while the striped bars refer to conventional training; without IIT.

Reference	Includes expert knowledge	UvA/Padova dataset	MSE $(mg/dL)^2$			
			30 PH	45 PH	60 PH	120 PH
Our model	Yes	Yes	181.50	857.40	1340.30	1521.00
Liu et al. (2019)	Yes	Yes	**101.20**	-	508.95	1206.86
Karim et al. (2020)	Yes	No	-	-	**495.96**	**1081.28**
Liu et al. (2023)	No	No	457.96	**452.41**	900.60	-

Table 1: Comparison between the model that achieved the best performance during this research MLP tree (256) IIT using regular *simglucose* causal model (first row) and previous models sourced from the literature. All four prediction horizons are compared with the best results in bold.

Firstly, we adapted the *simglucose* model by compressing the time dimension and imposed acyclicity as described in Section 4.3, obtaining a SCM. Figure 3 shows significant improvements in IIT performance across all PHs when using this amended *simglucose*. IIT achieves low errors and performs better than standard training. It is worth noting that standard training is generally more unstable than IIT, as it almost doubles the standard deviation for PHs 30, 60, and 120. However, the overlapping boxes for each PH, especially in the interquartile range, suggest that the groups have similar medians and variability. Due to the small sample size, it was not possible to test whether the performances with and without IIT were significantly different via statistical hypothesis testing.

Regarding this results, the absolute IIT RMSE gains may appear modest. The causal model used in this study is a simplified version of the original *simglucose* model, reducing the number of input parameters per patient from 64 to 13. While this abstraction lowers computational complexity and aligns with the IIT framework, it may also introduce inaccuracies by omitting important physiological dynamics. The choice of a simplified architecture was intentional to focus on evaluating the applicability of IIT. In addition, the improvements from IIT diminish as the prediction horizon increases (e.g., at 120 minutes) due to the growing influence of unmodeled factors such as behavioral variability, physical activity, and stress, which are generally challenging to capture.

As for the causal abstraction analysis, we observe in Figure 4(*a*) that the L_{INT} lowers as the training progresses. This metric quantifies how much the predictive model is able to align with causal knowledge. These graphs indicate that the MLP tree (256) model is learning from the counterfactual amended *simglucose* behaviour, producing better counterfactual predictions, and therefore lowering the L_{INT} after each training epoch. While the L_{INT} does not reach zero, and thus the model does not achieve a complete causal abstraction, the value of L_{INT} does reach single digit across all horizons, suggesting a good degree of causal abstraction. Appendix E Table 5 reports L_{INT} for each IIT MLP tree (256) module for the test dataset. Interestingly, we observe that for all the PHs, module $X4$ and module $X5$ tend to have a slightly higher L_{INT} median than the rest of the modules. Modules $X4$ and $X5$ are aligned to the variables encoding glucose mass in plasma and in rapidly/slowly equilibrating tissues. The causal mechanism is especially complex at $dxdt_4$, involving the EGP_t endogenous glucose production at time t, Rat glucose rate of appearance in plasma at time t, U_{iit} insulin-independent glucose utilization at time t and E_t renal excretion at time t. Likewise, the mechanism at $dxdt_5$ utilizes U_{idt} insulin-dependent glucose utilization at time t. The fact that L_{INT} is higher in these modules may be due to the fact that they are concerned with time-dependent factors that are endogenous to the SCM, which could be particularly influential and difficult to abstract. Similarly, module $X13$ also presents a slightly higher L_{INT} median. This module is aligned to the variable representing subcutaneous glucose level, which is directly receiving input from module $X4$, thus higher L_{INT} values are likely due to propagation from $X4$.

Comparing the regular and amended causal models, Appendix G Fig. 10(*a*) and Fig. 10(*b*) show differences between BG targets generated by the amended simulator and those from the regular one. These figures reveal that the absence of cyclical relationships, which are essential for capturing the feedback loops in glucose-insulin dynamics, significantly alters the model's output, leading to targets that do not fully reflect clinical behavior in T1DM patients. Nevertheless, the deviation produced is within the safe clinical areas for the patient (EGA areas A and B). For this reason, we believe experiments with the amended model are still worth conducting and sharing. This suggests that the amended simulator, while allowing for a successful abstraction with the MLP tree (256), may not fully reflect the glucose and insulin dynamics in T1DM patients as represented by the regular *simglucose*.

Next, we evaluated the four MLP models using the regular *simglucose* as the causal model. In the case of MLP parallel model, Figure 5 shows lower prediction errors with IIT at PH 30, 60 and 120, but higher errors at PH 45 compared to standard training. This model lacks connections between modules, unlike the causal model, which may explain the similar results for both training methods. As a consequence, the model is underfitted, as seen in the horizontal prediction trends for the error grid analysis in Appendix G Figures 8(*a*) and 8(*b*). Regarding the MLP tree model with 256 hidden size, IIT outperforms standard training: MLP tree with IIT consistently shows lower RMSE values across all PHs compared to the standard version, just like in the case of the amended simulator

(albeit with higher errors across the board). These results suggest that the connections between modules help the IIT-trained MLP tree model to make more accurate predictions. Therefore, besides the mapping, an architectural resemblance to the causal model is helpful for the training. On the other hand, the MLP tree (128) model shows slightly lower RMSE, especially at PH 30 and 60. This suggests that a larger number of nodes is needed for a module to capture the complexity of the corresponding causal mechanism. Moving to MLP joint model, its architecture design should be addressing the cyclic connections on the regular *simglucose*. However, it does not significantly outperform the MLP tree model (Table 6). These results suggest that encapsulating the cycles of the time-compressed DAG in a single module does not solve the problem.

Finally, in Table 1 we compared state-of-the-art results with our MLP tree (256) using the regular *simglucose*. The study by Liu et al. (2019) achieves superior perfomance at 30-minute PH, with a MSE difference of 80.30 $(mg/dL)^2$ in favor of their model. However, Liu et al. (2019) utilizes a more constrained testing dataset with lower variability, as it includes only 10 out of the 30 available UvA/Padova in-silico patients. This limitation may impact the generalizability of their results and complicate direct performance comparisons. 45-minute PH comparison is missing as the other two studies did not report their perfomance at that PH. At PH 45, (Liu et al., 2023) ranks first, aggregating results from 11 different ML models. For 60-minute PH, Karim et al. (2020)'s model outperforms our model but their results are based on testing only on one T1DM patient, whereas we included more patient variability, possibly explaining the performance gap. At 120-minute PH, our performance is lower ($\triangle$MSE=439.72$(mg/dL)^2$). This result underscores the increasing complexity of longer prediction horizons, where maintaining predictive accuracy becomes more challenging. In summary, our model does not perform as well as current state-of-the-art models, but nonetheless achieves results in the same ballpark with a very simple architecture and a more challenging test set.

Beyond predictive performance, it is important to consider the cost-benefit trade-offs of incorporating expert knowledge. First of all, the causal model we used is well-documented in the T1DM domain, reducing the effort required for model development. Secondly, the primary advantage of IIT is not merely the predictive improvement but also the alignment of the neural network with causal reasoning, enhancing interpretability and trustworthiness—critical factors in clinical applications. In addition, by integrating causal knowledge, we reduce reliance on patient-specific pharmacokinetic parameters, which are expensive and time-consuming to obtain. This trade-off could make the model more scalable and practical in real-world settings. Finally, the effort to develop a causal model is a one-time investment. Even slight improvements, when applied consistently over time on a large population, could have meaningful implications for long-term diabetes management.

6.1. Limitations and future work

First, the model used as a causal model is the S2008 version, the most outdated. Using the S2008 version has provided us with a causal structure to work with but it does not faithfully reflect all real clinical scenarios. In the event that in the future S2017 becomes open to the public, it is recommended to update the causal model. Furthermore, this project has been limited to a simple MLP architecture, to ease the exploration of different alignments and configurations. Future experiments should include the investigation of a more complex MLP architecture, such as an MLP where the time dimension is not collapsed and each time-indexed variable is modelled distinctly. This would allow for a fine-grained representation of temporal dynamics. However, we have already observed that this model introduces substantial complexity due to the increased number of variables and in-

terdependencies, which can only be investigated in the presence of much larger datasets. It is also interesting to include in this list models designed for temporal tasks such as long short-term memory or DRNN. As a potential extension of our current findings, one could explore the idea of deliberately injecting incorrect knowledge into the model to evaluate its robustness. This approach could provide insights into the resilience of predictive models when faced with potential errors or inaccuracies in expert knowledge, helping to assess their reliability under less-than-ideal conditions. We would also like to highlight that the data generated for the training comes from distributions created from 30 in-silico subjects. It is recommended to obtain the real distributions of *simglucose* in the future, under a paid license. In addition, it would be appropriate to test the models on real data, such as the Ohio dataset (Marling and Bunescu, 2020). Finally, if the data allows, it is recommended to run statistical tests to determine the superiority of IIT training.

7. Conclusion

This study investigated the applicability of IIT for predicting BG levels in T1DM patients using neural network models. Our primary objective was to determine whether IIT could be effectively utilized in complex healthcare scenarios, which pose greater challenges than previously explored applications. Unlike traditional simulators, a NN model capable of predicting BG levels offers the potential for computationally efficient causal abstraction, making it viable for integration into lightweight, wearable devices with insulin pumps (Pereira et al., 2024; Nahavandi et al., 2022). Such an approach could lead to real-time, on-device BG prediction that is both accurate and causally informed. Although the immediate application of our methodology focuses on BG prediction, the underlying approach holds potential for broader use across healthcare domains, demonstrating how a theoretical framework like IIT can be adapted to a real-world use case with promising outcomes.

By training NN with IIT to impose the causal structure of the *simglucose* model, we showed that some of the IIT-trained models obtain lower prediction errors and demonstrate a closer alignment with the causal structure, as evidenced by reduced L_{INT} values. Thanks to the breakdown of the L_{INT} loss into components, we were also able to recognize which aspects of the causal mechanisms were captured less well.

In conclusion, we have provided evidence that IIT can be successfully applied to complex medical prediction tasks like BG level forecasting in T1DM patients. This work indicates a path for future research to further refine neural network architectures and alignment to causal models.

8. Code and data availability

The code used for this study can be accessed at: `https://github.com/aespogom/IIT_simglucose`. For any questions or additional information regarding the code, please contact the authors.

Acknowledgments

AE would like to thank the Medical Informatics department from UvA for the opportunity to dive into this research. The authors also thank Thomas Icard for his support and feedback on earlier drafts of this work.

References

Chinyere Aloke. Current advances in the management of diabetes mellitus. *Biomedicines*, 10(10): 2436, 2022. ISSN 2227-9059. doi: 10.3390/biomedicines10102436. URL https://www.mdpi.com/2227-9059/10/10/2436. PubMed ID: 36289697.

Yotam Amar, Smadar Shilo, Tal Oron, Eran Amar, Moshe Phillip, and Eran Segal. Clinically accurate prediction of glucose levels in patients with type 1 diabetes. *Diabetes Technology & Therapeutics*, 22(8):562–569, 2020. doi: 10.1089/dia.2019.0435. URL https://doi.org/10.1089/dia.2019.0435. PMID: 31928415.

American Diabetes Association. Diagnosis and classification of diabetes mellitus. *Diabetes Care*, 32(Supplement 1):S62–S67, 01 2009. ISSN 0149-5992. doi: 10.2337/dc09-S062. URL https://doi.org/10.2337/dc09-S062.

Maxime De Bois, Mounîm A. El Yacoubi, and Mehdi Ammi. GLYFE: Review and benchmark of personalized glucose predictive models in type 1 diabetes. *Medical & Biological Engineering & Computing*, 60(1):1–17, January 2022. ISSN 1741-0444. doi: 10.1007/s11517-021-02437-4. URL https://doi.org/10.1007/s11517-021-02437-4.

William L Clarke, Daniel Cox, Linda A Gonder-Frederick, William Carter, and Stephen L Pohl. Evaluating clinical accuracy of systems for self-monitoring of blood glucose. *Diabetes Care*, 10 (5):622–628, 09 1987. ISSN 0149-5992. doi: 10.2337/diacare.10.5.622. URL https://doi.org/10.2337/diacare.10.5.622.

William L. Clarke, Stacey Anderson, Leon Farhy, Marc Breton, Linda Gonder-Frederick, Daniel Cox, and Boris Kovatchev. Evaluating the clinical accuracy of two continuous glucose sensors using continuous glucose–error grid analysis. *Diabetes Care*, 28(10):2412–2417, 10 2005. ISSN 0149-5992. doi: 10.2337/diacare.28.10.2412. URL https://doi.org/10.2337/diacare.28.10.2412.

John Daniels, Anaïs G. Herrero, David J. Warrington, Nick G. Oliver, and Pantelis Georgiou. A multitask learning approach to personalized blood glucose prediction. *IEEE Journal of Biomedical and Health Informatics*, 26(1):436–445, January 2022. doi: 10.1109/JBHI.2021.3100558. URL https://ieeexplore.ieee.org/document/9497711.

Lowell Dilworth. Diabetes mellitus and its metabolic complications: The role of adipose tissues. *International Journal of Molecular Sciences*, 22(14):7644, 2021. ISSN 1422-0067. doi: 10.3390/ijms22147644. URL https://www.mdpi.com/1422-0067/22/14/7644.

Federico D'Antoni, Mario Merone, Vincenzo Piemonte, Giulio Iannello, and Paolo Soda. Autoregressive time delayed jump neural network for blood glucose levels forecasting. *Knowledge-Based Systems*, 203:106134, 2020. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2020.106134. URL https://www.sciencedirect.com/science/article/pii/S0950705120303890.

International Diabetes Federation. Idf type 1 diabetes index 2022 report. Online, 2022. URL https://diabetesatlas.org/idfawp/resource-files/2022/12/IDF-T1D-Index-Report.pdf. Accessed: 2024-09-15.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah D. Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. *Proceedings of the 39th International Conference on Machine Learning (ICML)*, July 2022. URL https://arxiv.org/abs/2112.00826.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. Finding alignments between interpretable causal variables and distributed neural representations, 2024. URL https://arxiv.org/abs/2303.02536.

Rohan Gupta, Iván Arcuschin, Thomas Kwa, and Adrià Garriga-Alonso. Interpbench: Semi-synthetic transformers for evaluating mechanistic interpretability techniques, 2024. URL https://arxiv.org/abs/2407.14494.

Jing Huang, Zhengxuan Wu, Kyle Mahowald, and Christopher Potts. Inducing character-level structure in subword-based language models with type-level interchange intervention training, 2023. URL https://arxiv.org/abs/2212.09897.

Rebaz A.H. Karim, István Vassányi, and István Kósa. After-meal blood glucose level prediction using an absorption model for neural network training. *Computers in Biology and Medicine*, 125:103956, 2020. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2020.103956. URL https://www.sciencedirect.com/science/article/pii/S0010482520302900.

Kezhi Li, Chengyuan Liu, Taiyu Zhu, Pau Herrero, and Pantelis Georgiou. Glunet: A deep learning framework for accurate glucose forecasting. *IEEE Journal of Biomedical and Health Informatics*, 24(2):414–423, 2020. doi: 10.1109/JBHI.2019.2931842.

Chengyuan Liu, Josep Vehí, Parizad Avari, Monika Reddy, Nick Oliver, Pantelis Georgiou, and Pau Herrero. Long-term glucose forecasting using a physiological model and deconvolution of the continuous glucose monitoring signal. *Sensors*, 19(19), 2019. ISSN 1424-8220. doi: 10.3390/s19194338. URL https://www.mdpi.com/1424-8220/19/19/4338.

Kui Liu, Linyi Li, Yifei Ma, Jun Jiang, Zhenhua Liu, Zichen Ye, Shuang Liu, Chen Pu, Changsheng Chen, and Yi Wan. Machine learning models for blood glucose level prediction in patients with diabetes mellitus: Systematic review and network meta-analysis. *JMIR Med Inform*, 11:e47833, Nov 2023. ISSN 2291-9694. doi: 10.2196/47833. URL https://doi.org/10.2196/47833.

Chiara Dalla Man, Marc D. Breton, and Claudio Cobelli. Physical activity into the meal glucose—insulin model of type 1 diabetes: In silico studies. *Journal of Diabetes Science and Technology*, 3(1):56–67, 2009. doi: 10.1177/193229680900300107. URL https://doi.org/10.1177/193229680900300107. PMID: 20046650.

Cindy Marling and Razvan Bunescu. The ohiot1dm dataset for blood glucose level prediction: Update 2020. In *Proceedings of the 5th International Workshop on Knowledge Discovery in Healthcare Data (KDH)*, pages 71–74, September 2020. URL https://pubmed.ncbi.nlm.nih.gov/33584164/. CEUR Workshop Proceedings, Volume 2675.

Ali Mohebbi, Alexander R. Johansen, Nicklas Hansen, Peter E. Christensen, Jens M. Tarp, Morten L. Jensen, Henrik Bengtsson, and Morten Mørup. Short term blood glucose prediction based on continuous glucose monitoring data. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pages 5140–5145, 2020. doi: 10.1109/EMBC44109.2020.9176695.

Darius Nahavandi, Roohallah Alizadehsani, Abbas Khosravi, and U Rajendra Acharya. Application of artificial intelligence in wearable devices: Opportunities and challenges. *Computer Methods and Programs in Biomedicine*, 213:106541, 2022. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2021.106541. URL https://www.sciencedirect.com/science/article/pii/S0169260721006155.

J. Pearl. *Causality*. Causality: Models, Reasoning, and Inference. Cambridge University Press, 2009. ISBN 9780521895606. URL https://books.google.nl/books?id=f4nuexsNVZIC.

Carlos Vinicius Fernandes Pereira, Edvard Martins de Oliveira, and Adler Diniz de Souza. Machine learning applied to edge computing and wearable devices for healthcare: Systematic mapping of the literature. *Sensors*, 24(19), 2024. ISSN 1424-8220. doi: 10.3390/s24196322. URL https://www.mdpi.com/1424-8220/24/19/6322.

C. Pérez-Gandía, A. Facchinetti, G. Sparacino, C. Cobelli, E.J. Gómez, M. Rigla, A. de Leiva, and M.E. Hernando. Artificial neural network algorithm for online glucose prediction from continuous glucose monitoring. *Diabetes Technology & Therapeutics*, 12(1):81–88, 2010. doi: 10.1089/dia.2009.0076. URL https://doi.org/10.1089/dia.2009.0076. PMID: 20082589.

Francesco Prendin, Simone Del Favero, Martina Vettoretti, Giovanni Sparacino, and Andrea Facchinetti. Forecasting of glucose levels and hypoglycemic events: Head-to-head comparison of linear and nonlinear data-driven algorithms based on continuous glucose monitoring data only. *Sensors*, 21(5), 2021. ISSN 1424-8220. doi: 10.3390/s21051647. URL https://www.mdpi.com/1424-8220/21/5/1647.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Towards causal representation learning. *CoRR*, abs/2102.11107, 2021. URL https://arxiv.org/abs/2102.11107.

Statista. Global health care costs due to diabetes by region 2021-2045. Online, 2024. URL https://www.statista.com/statistics/241831/health-care-costs-due-to-diabetes-worldwide-by-region/. Accessed: 2024-09-15.

Wouter A. C. van Amsterdam, Pim A. de Jong, Joost J. C. Verhoeff, Tim Leiner, and Rajesh Ranganath. From algorithms to action: improving patient care requires causality. *BMC Medical Informatics and Decision Making*, 24, 2022. URL https://api.semanticscholar.org/CorpusID:252284130.

Wouter A. C. van Amsterdam, Nan van Geloven, Jesse H. Krijthe, Rajesh Ranganath, and Giovanni Ciná. When accurate prediction models yield harmful self-fulfilling prophecies, 2024. URL https://arxiv.org/abs/2312.01210.

Nan van Geloven, Ruth H Keogh, Wouter van Amsterdam, Giovanni Cinà, Jesse H. Krijthe, Niels Peek, Kim Luijken, Sara Magliacane, Paweł Morzywołek, Thijs van Ommen, Hein Putter, Matthew Sperrin, Junfeng Wang, Daniala L. Weir, and Vanessa Didelez. The risks of risk assessment: causal blind spots when using prediction models for treatment decisions, 2024. URL https://arxiv.org/abs/2402.17366.

Roberto Visentin, Chiara Dalla Man, Boris Kovatchev, and Claudio Cobelli. The university of virginia/padova type 1 diabetes simulator matches the glucose traces of a clinical trial. *Diabetes Technology & Therapeutics*, 16(7):428–434, 2014. doi: 10.1089/dia.2013.0377. URL https://doi.org/10.1089/dia.2013.0377. PMID: 24571584.

Zhengxuan Wu, Atticus Geiger, Joshua Rozner, Elisa Kreiss, Hanson Lu, Thomas Icard, Christopher Potts, and Noah Goodman. Causal distillation for language models. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4288–4295, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.318. URL https://aclanthology.org/2022.naacl-main.318/.

Jinyu Xie. Simglucose v0.2.1. Online, 2018. URL https://github.com/jxx123/simglucose. Accessed on: 02-17-2025.

Chiara Zecchin, Andrea Facchinetti, Giovanni Sparacino, Giuseppe De Nicolao, and Claudio Cobelli. Neural network incorporating meal information improves accuracy of short-time prediction of glucose concentration. *IEEE Transactions on Biomedical Engineering*, 59(6):1550–1560, 2012. doi: 10.1109/TBME.2012.2188893.

Taiyu Zhu, Kezhi Li, Jianwei Chen, Pau Herrero, and Pantelis Georgiou. Dilated recurrent neural networks for glucose forecasting in type 1 diabetes. *Journal of healthcare informatics research*, 4(3):308—324, September 2020. ISSN 2509-4971. doi: 10.1007/s41666-020-00068-2. URL https://europepmc.org/articles/PMC8982716.

Appendix A. Data acquisition

In order to generate 200 virtual subject model parameter vectors, several steps are needed. The distributions for each parameter are conditioned by age, considering children from 0 to 13 years, adolescents from 14 to 20 years, and adults older than 20 years old. Each joint distribution conforms to Gaussian probability distributions. In other words, each parameter is described by three different joint distributions; one for the age range of children, one for the age range of adolescents, and one for the age range of adults. Next, we generate a random list of ages between 0 and 100 years (child n=7, adolescent n=10, adult n=183). To sample the data, the Gaussian mixture model probability distribution from the Python *sklearn* library was used for each age range. More details about UVA/Padova T1DM joint distributions can be found in Xie (2018) and Man et al. (2009) although Table 2 shows the means and standard deviation of the ones included in the input dataset used for this study. This approach is novel but aligns with the methodology used in the simulator, which extracts subjects from these joint distributions. It is important to note that the 200 T1DM-generated subjects were used only during the training. The original 30 T1DM simulator subjects were used for the testing, ensuring that the evaluation reflects the characteristics of the original dataset.

	Child		Adolescent		Adult	
	Mean	SD	Mean	SD	Mean	SD
x_4	260.42	25.16	282.12	26.33	258.58	12.90
x_5	95.66	35.97	371.76	17.92	194.25	29.02
x_6	5.60	1.23	5.08	1.58	5.94	1.25
x_8	106.72	11.31	109.18	8.82	105.03	16.71
x_9	106.72	11.31	109.18	8.82	105.03	16.71
x_{10}	2.92	1.30	3.44	1.10	3.50	1.19
x_{11}	67.33	15.78	67.87	9.30	87.84	33.76
x_{12}	60.43	22.37	66.62	24.62	112.09	30.73
x_{13}	260.42	25.16	282.11	26.33	258.58	12.90

Table 2: Mean and standard deviation of each joint distribution for the parameters included in the data input. Each parameter is described by three different distributions depending on the age range.

Appendix B. Training configuration and evaluation metrics

The settings used to obtain the results of the experiments are now described. The implementation of a PyTorch Trainer class orchestrates the training process for a NN given a causal model. It includes functionality for performing forward passes, optimizing the model parameters, and evaluating the trained model.

Regarding the optimizer, AdamW optimizer is used with a learning rate of 0.01, epsilon of $1e^{-6}$, and betas parameters (0.9, 0.98). The learning rate scheduler is based on linear scheduling with warm-up steps. The optimizer is minimizing the Mean Squared Error (MSE) loss. In the case of IIT experiments, causal loss and standard loss have a 0.75 and a 0.25 coefficient, respectively. In the case of standard trainings, the coefficients are 0 and 1, respectively. These coefficients represent how much each loss contributes to the global model loss. The higher the coefficient, the more effort would be dedicated to minimizing the corresponding loss. In short, a conventional training would have a causal loss coefficient of 0 and a standard loss coefficient of 1 while a purely IIT would have a

causal loss coefficient of 1 and a standard loss coefficient of 0. Also, an early stopping functionality is incorporated to halt training if the validation loss does not improve for 20 consecutive epochs. The maximum number of epochs is 300. In terms of batch size, it is set to 2 so that the interchange intervention is performed between two input patients. However, the gradient accumulation steps value is 20 so the loss is computed as if the batch size was 20. Also, gradient clipping is applied to prevent exploding gradients during backpropagation. A set of 10 random seeds were used to calculate the mean of the estimate and the standard deviation. The selection of these seed values is random and ensures a different initialization for each experiment.

The following performance metrics were used in the evaluation.

- Average absolute error (MAE) is defined as the magnitude of difference between the prediction of an observation y_i and the true value of that observation $\hat{y}_i$ within a dataset of length n.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

- MSE measures the average squared difference between the predicted y_i and the actual target values $\hat{y}_i$ within a dataset of length n.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3}$$

- RMSE measures the square root of the squared difference between predicted values and the actual target values. Essentially, RMSE is a measure of the magnitude of the residuals, making interpretation straightforward.

$$RMSE = \sqrt{MSE} \tag{4}$$

- Percentage of predicted values in the 'clinically acceptable' EGA classes A and B (Clarke et al., 1987). CG-EGA classifies predictions into 5 classes A-B-C-D-E with respect to the clinical outcome based on the predicted BG level. The error grid divides the plot into five regions. Points in Zone A are within 20% of the reference sensor, showing a clinically accurate zone. Points in Zone B would not lead to inappropriate treatment. Points in Zone C would lead to unnecessary treatment, whereas the points in Zone D are potentially dangerous, and failed to detect hypoglycemia or hyperglycemia events correctly. A predicted value is termed "clinically acceptable" if it is classified into either the A or B EGA class. CG-EGA is a variation of the EGA grid in which the "accurate" domain is roughly equivalent to the EGA "clinically acceptable" classification Clarke et al. (2005).

The predictions were evaluated on the 30, 45, 60 and 120-minute horizon. The EGA and RMSE results were evaluated for completeness and to comply with GLYFE benchmark Bois et al. (2022).

Appendix C. Performance using with amended *simglucose* causal model

Table 3 presents the results for the MLP tree model with 256 hidden units, using amended *simglucose* as causal model. The best results for each PH are highlighted, comparing IIT training with standard training. Mean and standard deviation are reported due to different random seeds used.

	MSE $(mg/dL)^2$					MAE (mg/dL)					EGA A-B (%)				
PH	IIT (Mean ± ST)		Standard (Mean ± ST)		△	IIT (Mean ± ST)		Standard (Mean ± ST)		△	IIT (Mean ± ST)		Standard (Mean ± ST)		△
30	**255.60**	**63.83**	301.47	118.15	-45.87	**12.04**	**1.67**	12.93	2.87	-0.89	**99.67**	**1.00**	99.00	1.53	0.67
45	**538.63**	**133.18**	560.33	126.24	-21.69	**16.98**	**2.31**	17.55	1.82	-0.58	**99.33**	**2.00**	98.33	2.24	1.00
60	**885.20**	**175.13**	981.09	327.36	-95.89	**22.01**	**2.85**	23.30	5.12	-1.29	99.67	1.00	99.67	1.00	0
120	**642.95**	**85.93**	704.50	202.76	-128.34	**19.03**	**1.57**	19.39	3.50	-0.36	**99.33**	**2.00**	98.67	2.21	0.66

Table 3: Results of the MLP tree (256) model across the four PHs for the test (n=30) in-silico T1DM patients (best results in bold). 10 different seeds were used to obtain the estimations. The causal model used is the amended *simglucose*. The "IIT" column refers to the models being trained through interchange intervention training while "Standard" refers to conventional training; without IIT, and △ to the difference of the means between the IIT and Standard result. A green △ indicates that IIT achieves better performance than standard training for that metric. A red △ indicates that IIT achieves worse performance than standard training for that metric.

Appendix D. Comparison with previous studies

Reference	Includes expert knowledge	UvA/Padova dataset	MSE $(mg/dL)^2$			
			30 PH	45 PH	60 PH	120 PH
Our model	Yes	Yes	181.50	857.40	1340.30	1521.00
Liu et al. (2019)	Yes	Yes	101.20	-	508.95	1206.86
Karim et al. (2020)	Yes	No	-	-	**495.96**	**1081.28**
Pérez-Gandía et al. (2010)	No	No	324.00	729.00	-	-
Prendin et al. (2021)	No	No	490.62	-	-	-
Zhu et al. (2020)	No	No	357.21	-	-	-
D'Antoni et al. (2020)	No	No	349.69	-	-	-
Amar et al. (2020)	No	No	561.22	-	1706.52	-
Li et al. (2020)	No	Yes	115.13	-	513.02	-
Zecchin et al. (2012)	No	Yes	**88.36**	-	-	-
Mohebbi et al. (2020)	No	No	433.61	851.06	1340.09	-
Daniels et al. (2022)	No	No	353.44	**640.09**	1011.24	2227.84

Table 4: Granular comparison between the model that achieved the best performance during this research MLP tree (256) IIT using regular *simglucose* causal model (first row) and previous models sourced from the literature. All four prediction horizons are compared with the best results in bold.

Appendix E. Causal abstraction analysis

Counterfactual loss (L_{INT}) during the training (Fig. 6) and testing (Table 5 and Fig. 7) processes across different PHs for the MLP tree model with 256 hidden units, using the amended *simglucose*. This involved interchanging interventions between the causal model and MLP tree model, calculating the loss between their output values.

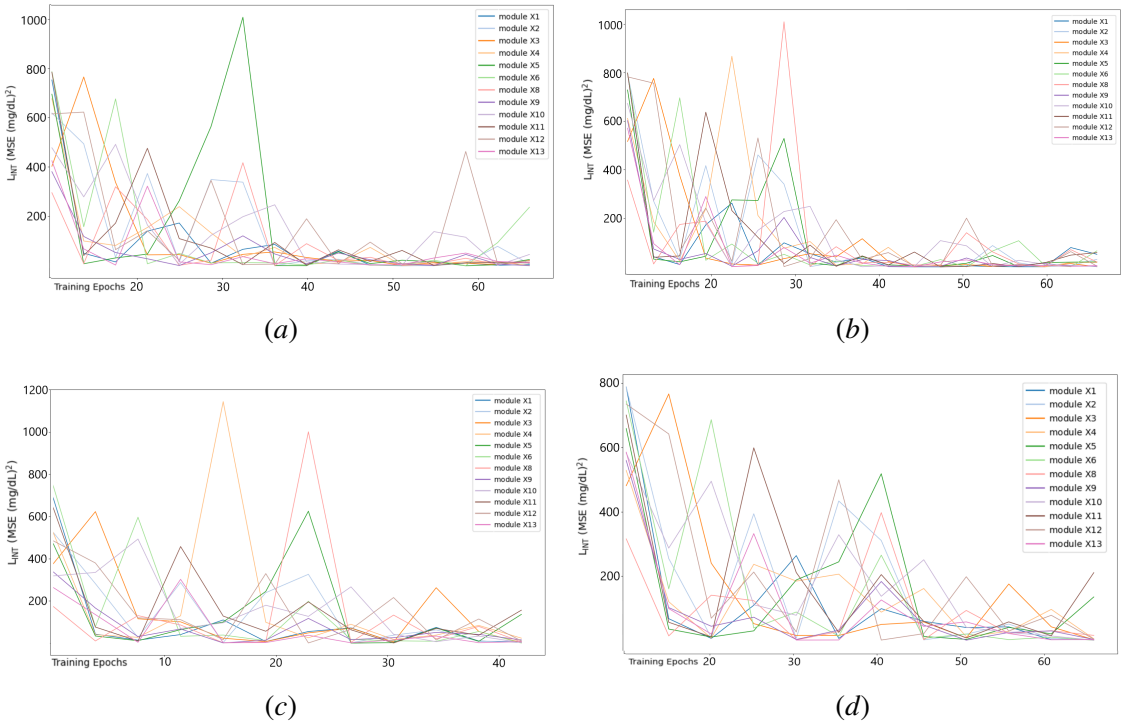


Figure 6: MLP IIT $L_{INT}(MSE(mg/dL)2)$ during the training using the amended *simglucose* as the causal model for PHs (a) 30, (b) 45, (c) 60 and (d) 120. The L_{INT} is grouped by modules.

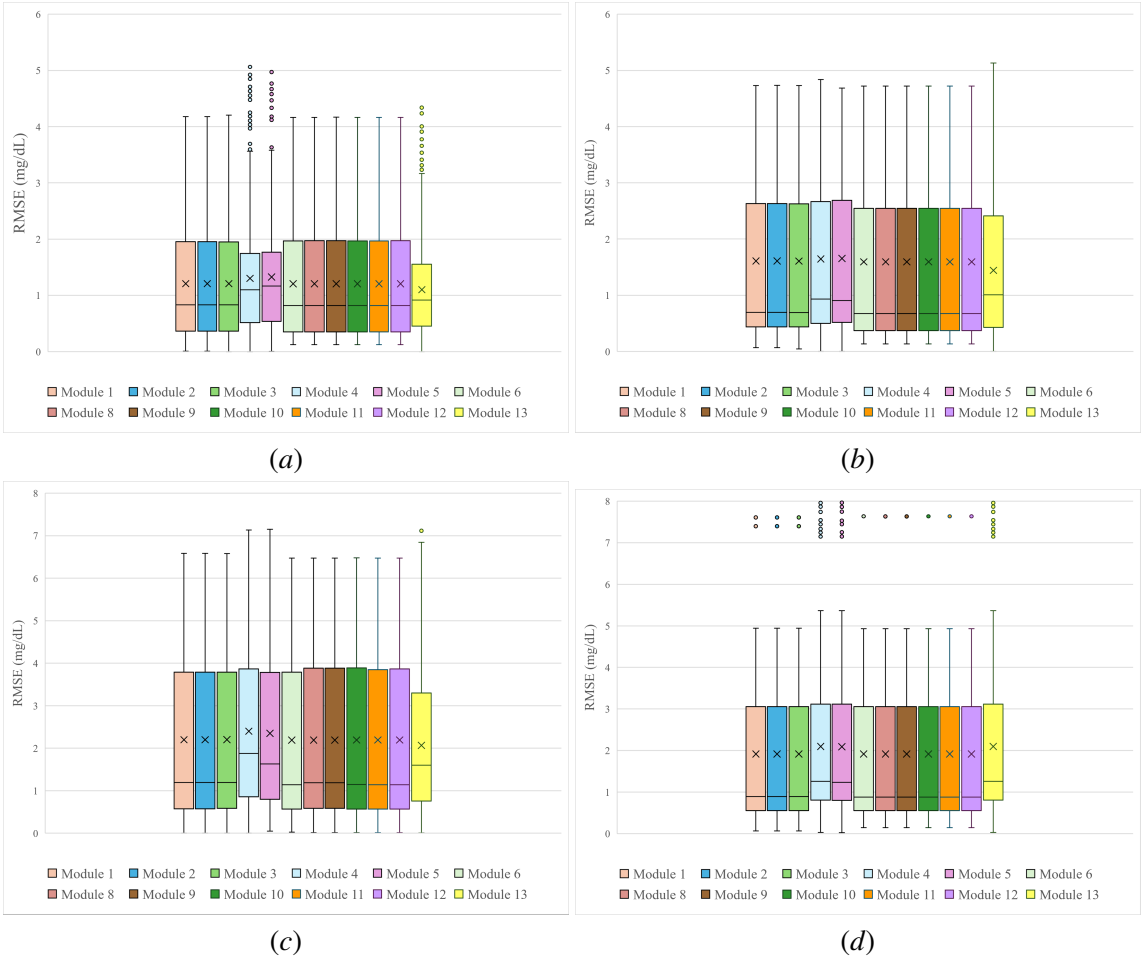


Figure 7: MLP IIT L_{INT} during the testing using the amended *simglucose* as the causal model for PHs (a) 30, (b) 45, (c) 60 and (d) 120. The L_{INT} is grouped by modules.

L_{INT} (RMSE mg/dL)				
PH	30	45	60	120
Module	Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD
X1	1.21 ± 1.04	1.61 ± 1.52	2.20 ± 1.95	1.92 ± 1.75
X2	1.21 ± 1.04	1.61 ± 1.52	2.20 ± 1.95	1.92 ± 1.75
X3	1.21 ± 1.04	1.64 ± 1.43	2.20 ± 1.95	1.92 ± 1.75
X4	1.30 ± 1.01	1.65 ± 1.47	2.40 ± 1.84	2.10 ± 1.70
X5	1.33 ± 1.02	1.60 ± 1.53	2.35 ± 1.87	2.09 ± 1.70
X6	1.21 ± 1.05	1.60 ± 1.53	2.19 ± 1.99	1.92 ± 1.78
X8	1.21 ± 1.05	1.60 ± 1.53	2.19 ± 1.95	1.92 ± 1.78
X9	1.21 ± 1.05	1.60 ± 1.53	2.19 ± 1.95	1.92 ± 1.78
X10	1.21 ± 1.05	1.60 ± 1.53	2.19 ± 1.98	1.92 ± 1.78
X11	1.21 ± 1.05	1.60 ± 1.53	2.19 ± 1.98	1.92 ± 1.78
X12	1.21 ± 1.05	1.60 ± 1.53	2.19 ± 1.98	1.92 ± 1.78
X13	1.10 ± 0.84	1.44 ± 1.20	2.06 ± 1.57	2.10 ± 1.70

Table 5: L_{INT} results of the MLP tree (256) model across the four prediction horizons (PH) for the test (n=30) in-silico T1DM patients. The random seed is 56. The causal model used is the DAG *simglucose*. All interchanged interventions are performed for the models trained through IIT, and then the loss for the two outputs is calculated. The lower L_{INT}, the higher the causal abstraction is achieved.

Appendix F. Performance using the regular *simglucose* causal model

Table 6 presents the results for the MLP tree model with 256 hidden units, using regular *simglucose* as causal model. The best results for each PH are highlighted, comparing IIT training with standard training.

		MSE $(mg/dL)^2$					MAE (mg/dL)					EGA A-B (%)				
MLP	PH	IIT (Mean ± ST)		Standard (Mean ± ST)		△	IIT (Mean ± ST)		Standard (Mean ± ST)		△	IIT (Mean ± ST)		Standard (Mean ± ST)		△
Parallel	30	**149.79**	**24.71**	185.30	58.15	-35.51	**9.83**	**1.55**	10.25	1.57	-8.28	97.58	4.96	**97.67**	**2.74**	-0.09
	45	**666.81**	**80.37**	680.52	140.10	-13.71	**19.52**	**1.65**	20.15	2.29	-0.63	**79.00**	**3.53**	76.67	7.37	2.33
	60	**1301.95**	**88.66**	1359.00	159.30	-93.05	**27.36**	**1.37**	28.00	2.38	-0.64	**72.33**	**2.25**	70.33	3.99	2.00
	120	**1682.48**	**150.68**	3696.86	5838.54	-35.51	**31.50**	**1.64**	43.39	33.00	-11.89	**62.67**	**4.39**	54.67	19.76	8.00
Tree (256)	30	**359.31**	**144.75**	379.01	145.52	-19.7	**14.68**	**2.92**	15.74	3.94	-1.06	**90.00**	**7.70**	89.00	8.02	1.00
	45	**814.44**	**205.59**	1710.04	554.81	-895.61	**22.87**	**3.32**	25.57	5.14	-2.7	**74.67**	**6.52**	67.67	11.97	7.00
	60	**1563.24**	**489.79**	1710.04	554.81	-146.8	**31.17**	**6.14**	33.10	6.67	-1.93	**65.33**	**10.68**	64.33	15.56	1.00
	120	**1756.24**	**302.96**	2033.94	431.97	-277.70	**32.73**	**2.95**	35.41	4.47	-2.68	**60.00**	**6.48**	57.67	7.04	2.33
Tree (128)	30	**245.14**	**179.81**	303.68	124.36	-58.54	**11.50**	**4.13**	12.93	3.03	-1.43	**95.00**	**6.14**	82.67	1.41	12.33
	45	**905.76**	**528.66**	814.23	492.97	-91.53	23.22	7.79	**18.64**	**2.49**	4.58	**70.00**	**16.56**	68.33	1.76	1.67
	60	1284.24	336.86	**976.58**	**336.21**	307.66	27.59	4.06	**23.20**	**5.45**	4.39	**70.67**	**7.83**	57.33	4.10	13.34
	120	1956.33	639.65	**1771.60**	**532.03**	184.73	34.16	6.47	**19.89**	**3.62**	14.27	**58.33**	**8.50**	58.18	4.05	0.15
Joint	30	359.06	255.69	**287.95**	**176.92**	71.11	15.00	6.36	**13.46**	**4.76**	1.54	90.33	15.59	**94.67**	**8.20**	-4.34
	45	**848.92**	**242.88**	874.63	247.22	-25.71	**23.07**	**4.12**	23.48	4.43	-0.41	73.00	10.71	73.00	9.36	0
	60	**1527.63**	**493.59**	1718.96	537.99	-191.33	**30.75**	**5.88**	33.21	6.04	-2.46	**65.67**	**10.31**	63.33	7.70	-2.34
	120	**1979.62**	**462.07**	5559.40	7901.60	-3579.78	**35.39**	**4.75**	54.01	44.38	-18.62	**56.33**	**7.61**	46.67	25.29	9.66

Table 6: Results of the four different MLP models (Parallel, Tree 256, Tree 128, Joint) models across the four PHs for the test (n=30) in-silico T1DM patients (best results in bold). 10 different seeds were used to obtain the estimations. The causal model used is the regular *simglucose*. The "IIT" column refers to the models being trained through interchange intervention training while "Standard" refers to conventional training; without IIT, and △ to the difference of means between the IIT and Standard result. A green △ indicates that IIT achieves better performance than standard training for that metric. A red △ indicates that IIT achieves worse performance than standard training for that metric.

Appendix G. EGAs

We used Clarke error grid analysis to visualize predictions within clinically accepted ranges. Only the 120 PH grid of predictions with random seed 56 is shown for each model. The model MLP parallel Figs. 8(*a*), 8(*b*) show a tendency to predict within the range of 140 and 180, nearly horizontal. On the other hand, the rest of the models Figs. 9(*a*), 9(*b*), 9(*c*), 9(*d*), 8(*c*), 8(*d*) show more dispersed predictions along the diagonal. This holds for both IIT-trained and standard-trained models.

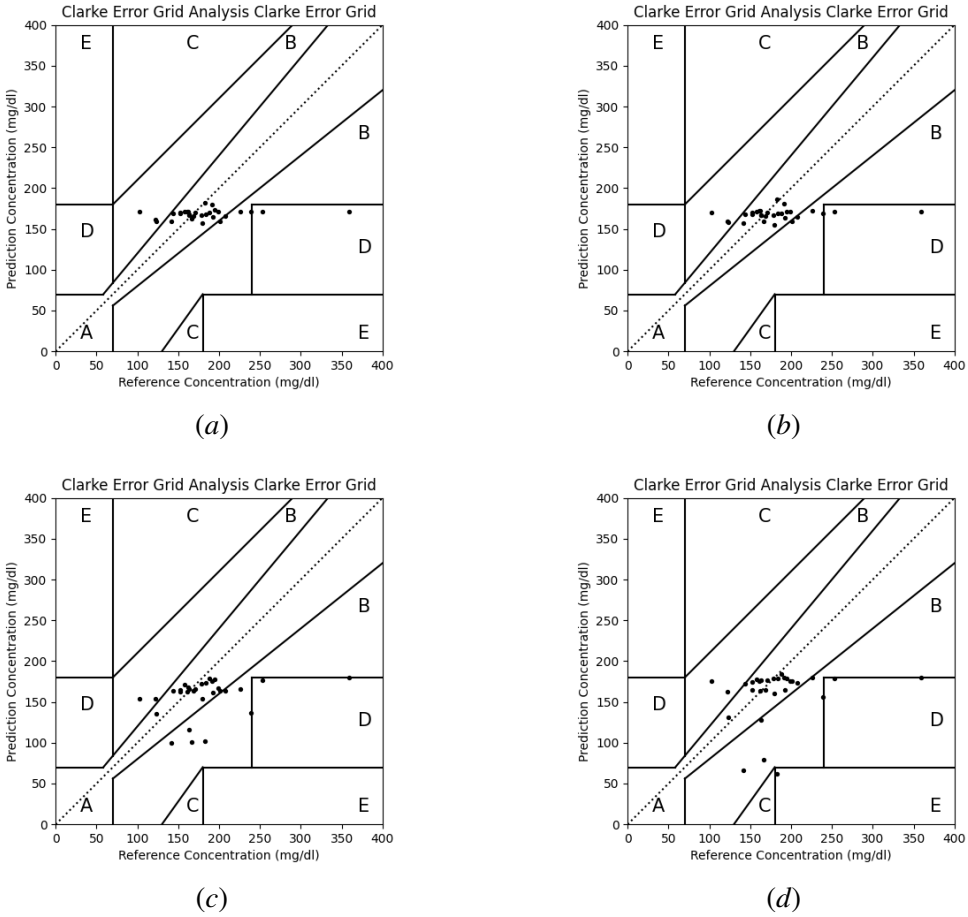


Figure 8: Error grid 120 PH of the MLP parallel architecture (a) and (b), and MLP joint architecture (c) and (d). The images on the left (a) and (c) refer to the models being trained through interchange intervention training using the regular causal model while the images on the right (b) and (d) correspond to conventional training; without IIT.

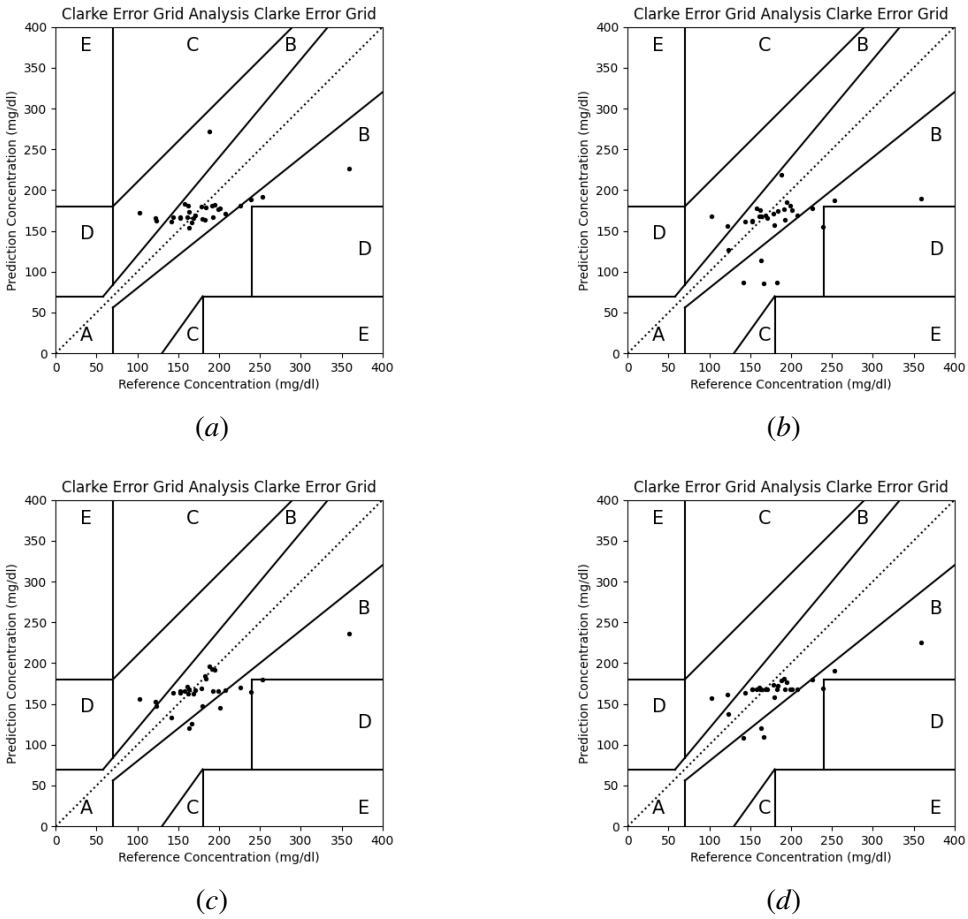
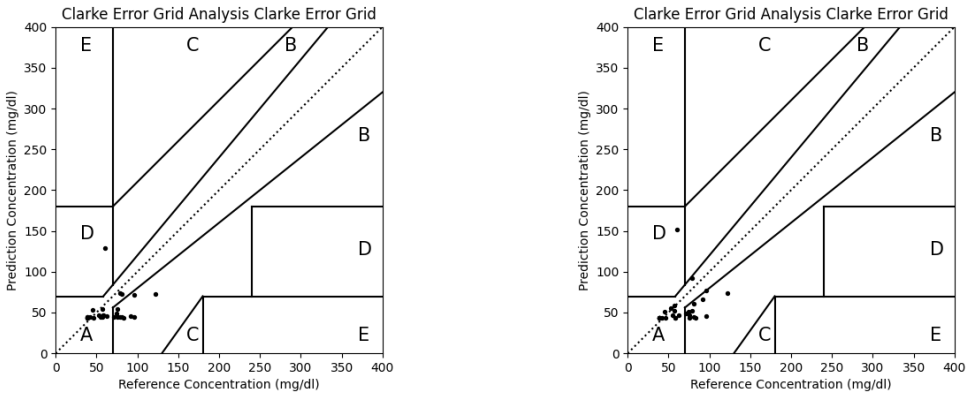


Figure 9: Error grid 120 PH of the MLP tree (128 hidden size) architecture (a) and (b), and MLP tree (256 hidden size) architecture (c) and (d). The images on the left (a) and (c) refer to the models being trained through interchange intervention training using the regular causal model while the images on the right (b) and (d) correspond to conventional training; without IIT.



(*a*) (*b*)

Figure 10: Error grid 120 PH of the MLP tree (256 hidden size) architecture. The images on the left (a) refer to the model being trained through interchange intervention training using the amended causal model while the image on the right (b) correspond to conventional training; without IIT.