Interchange Intervention Training Applied to Post-meal Glucose Prediction for Type 1 Diabetes Mellitus Patients

Ana Esponera Gómez[1] **Giovanni Cinà**[1, 2, 3]

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

Abstract

This research explores the application of Interchange Intervention Training (IIT) in predicting blood glucose levels in Type 1 Diabetes Mellitus (T1DM) patients by leveraging expert knowledge encoded in causal models. 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 abstract causal relationships. Results show that the model trained with IIT effectively abstracted the causal structure and it outperformed the standard one in terms of predictive performance across different prediction horizons (PHs) post-meal. These preliminary results with the acyclic model suggest the potential of IIT in enhancing predictive models in healthcare by effectively complying with expert knowledge.

1 INTRODUCTION

Theories of causal abstraction provide the mathematical tools needed to analyze a deep learning model by defining when a human-comprehensible, high-level causal explanation accurately represents opaque low-level details. For example, interventionist theories of causality [Geiger et al., 2021, Potts, Wu et al., 2021] allow us to model deep learning models and algorithms as the same mathematical object: an acyclic causal model also referred to as directed acyclic graphs (DAGs). Recently, Geiger et al. [2021] have succeeded in applying some of these structural analysis methods to several models for both visual recognition and natural language processing tasks. This has opened the door to the possibility of applying them in the healthcare field.

Taking into account the foregoing, the prediction of glucose levels in patients suffering from T1DM is a potential use case for several reasons: firstly, there is much prior knowledge of how glucose, insulin and carbohydrate intake interact in this type of patients. This knowledge can be encoded in a DAG. Secondly, the complexity of these interactions makes them very difficult to compute in a conventional way. Finally, while deep learning models have been proposed for glucose prediction [Liu et al., 2023, Karim et al., 2020, Liu et al., 2019], providing interpretability for deep learning model predictions can facilitate their approval by the healthcare community.

We showcase the applicability of IIT in a medical scenario by training a neural network to predict glucose levels of T1DM patients after a meal, enforcing the causal structure encoded in the *simglucose* simulator, acting as the DAG for this problem. Our results suggest that this type of training improves upon standard training in terms of performance, data efficiency and compliance to clinical knowledge.

2 METHODS

Simglucose Man et al. [2009] was chosen as the causal model. Approved by the FDA, it serves as a model for glucose-insulin dynamics in individuals with T1DM and its newest version is used for pre-trial explorations of medical interventions. For the purpose of these experiments, we modified the original computational model to conform to a DAG: we have removed two cyclic relationships between two pairs of equations. As for the neural network, we have chosen a simple MLP model. The architecture consists of 13 identical sequential modules connected in a tree structure inspired by the causal model. Each module consists of a linear layer with 256 as hidden size, followed by a leakyReLU activation function and a dropout layer with a rate of 0.3.

Interchange intervention training (IIT) is used for the optimization. An interchange intervention involves the evaluation of the model in various counterfactual intervention scenarios. These are created by mapping variables in the DAG to modules of the MLP, swapping part of the module values while the rest stays constant, and then comparing the outcomes under both scenarios. Intervention locations are not restricted only to the inputs but to any intermediate value in the causal structure. The loss in such counterfactual scenario is dubbed L_{INT} [Geiger et al., 2023]. It guarantees that the target causal model serves as a causal abstraction of the neural model, provided the L_{INT} is reduced to zero [Geiger et al., 2021].

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

where b and s are the $base$ and $source$ input values from the in input space swapped during the intervention, M_H is the high-level neural model, M_L is the low-level causal model, I is the variable being intervened on, τ is a mapping of output values from the low to high-level and $Loss$ is the chosen function to quantify the distance.

3 RESULTS

Table 1 shows the results achieved by the MLP model using the acyclic *simglucose* as the causal model and FDA-approved in-silico data as test dataset. The best results for each PH are highlighted in boldface comparing the IIT training versus the standard training for five different initializations. Models trained via IIT achieve lower errors than standard-trained ones.

Table 1: Results of the MLP model across the four PHs for the test (n=30) in-silico T1DM patients (best results in bold). Estimation from five different initializations.

	RMSE (mg/dL)	
PH	IIT (mean $\pm ST$)	Standard (mean $\pm ST$)
30	**15,70** $\pm 0,97$	15,84 $\pm 0,70$
45	**22,42** $\pm 1,69$	25,77 $\pm 2,74$
60	**29,32** $\pm 0,58$	30,86 $\pm 1,71$
120	**25,37** $\pm 0,72$	26,35 $\pm 0,47$

Figure 1 in the Appendix tracks the counterfactual training loss across the different PHs for the model trained through IIT for the acyclic *simglucose* as the causal model. As a general tendency, for each MLP module, the L_{INT} lowers as the training epochs increase, signalling an improvement in compliance with the causal model. Figure 2 reports L_{INT} across the different PHs for the IIT model for the test dataset. This metric quantifies to what extent the acyclic *simglucose* is a causal abstraction of the MLP model. The lower the value of L_{INT}, the more evidence that the causal model is an abstraction of the MLP model.

4 DISCUSSION AND CONCLUSION

In these preliminary results, we observe a difference between IIT and standard training performance. Setting the same initialization for IIT and standard, the models using IIT have lower errors. This is also supported by the causal abstraction analysis for the training and the testing. The causal abstraction analysis tracks the L_{INT} metric during the training process. This loss measures how much the causal model explains the MLP's reasoning, meaning a perfect causal abstraction when it is zero. For the models trained by IIT, it decreases as training progresses, suggesting effective counterfactual learning. Furthermore, L_{INT} is computed for the IIT MLP model using test dataset inputs. All modules are intervened for the entire test dataset. Thanks to this, we are able to measure the model abstraction by module, determining which modules are similar to the causal behaviour and which are not. Although perfect L_{INT} is not observed for any of the MLP modules, lower values indicate a certain degree of causal abstraction. Interestingly, we observe that for all the PHs, modules 4, 5 and 13 tend to have a higher L_{INT} than the rest of the modules. Modules 4 and 5 are aligned with the causal representation of glucose mass in plasma and tissues. Moreover, module 13 is aligned with the causal representation of the subcutaneous glucose level. Their relatively high L_{INT} values allow us to identify which modules of MLP have not been abstracted correctly and therefore justify the difference between predictions and true labels. For the rest of the modules, the counterfactual behaviour is lower and steady. The model shows a generally low RMSE, indicating near-perfect abstraction.

Therefore, the results demonstrate that leveraging expert knowledge through causal models, particularly via IIT, can improve predictive models in the healthcare domain. Besides, IIT inherently aims to comply with expert knowledge by encoding it into the high-level model's training process. In this case, it was possible to measure the extent to which the models complied with the expert knowledge through the counterfactual loss errors. The higher the causal abstraction of the DAG, the lower the L_{INT} results. The L_{INT} errors obtained are low, which indicates that the predictive model has been able to abstract the causal structure of the DAG almost entirely. However, complete compliance was not achieved, indicating room for further refinement.

In future work, the results presented here should be tested for statistical significance. Furthermore, the models should be re-trained and evaluated on real patient data (as opposed to in-silico patients). Finally, these estimates should be benchmarked against existing literature on glucose prediction.

Author Contributions

Both authors designed the experiments and wrote the article. AE carried out the data preparation and experiments. GC supervised the project.

Acknowledgements

AE would like to thank the Medical Informatics department from UvA for the opportunity to dive into this research.

References

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. 2021.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. Finding alignments between interpretable causal variables and distributed neural representations. 2023.

Rebaz A.H. Karim, Istvan Vassanyi, and Istvan Kosa. 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`.

Chengyuan Liu, Josep Vehi, 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`.

CD Man, MD Breton, and C Cobelli. Physical activity into the meal glucose-insulin model of type 1 diabetes: in silico studies. *J Diabetes Sci Technol*, 2009.

Christopher Potts. Analysis methods in nlp. *Natural language understanding*.

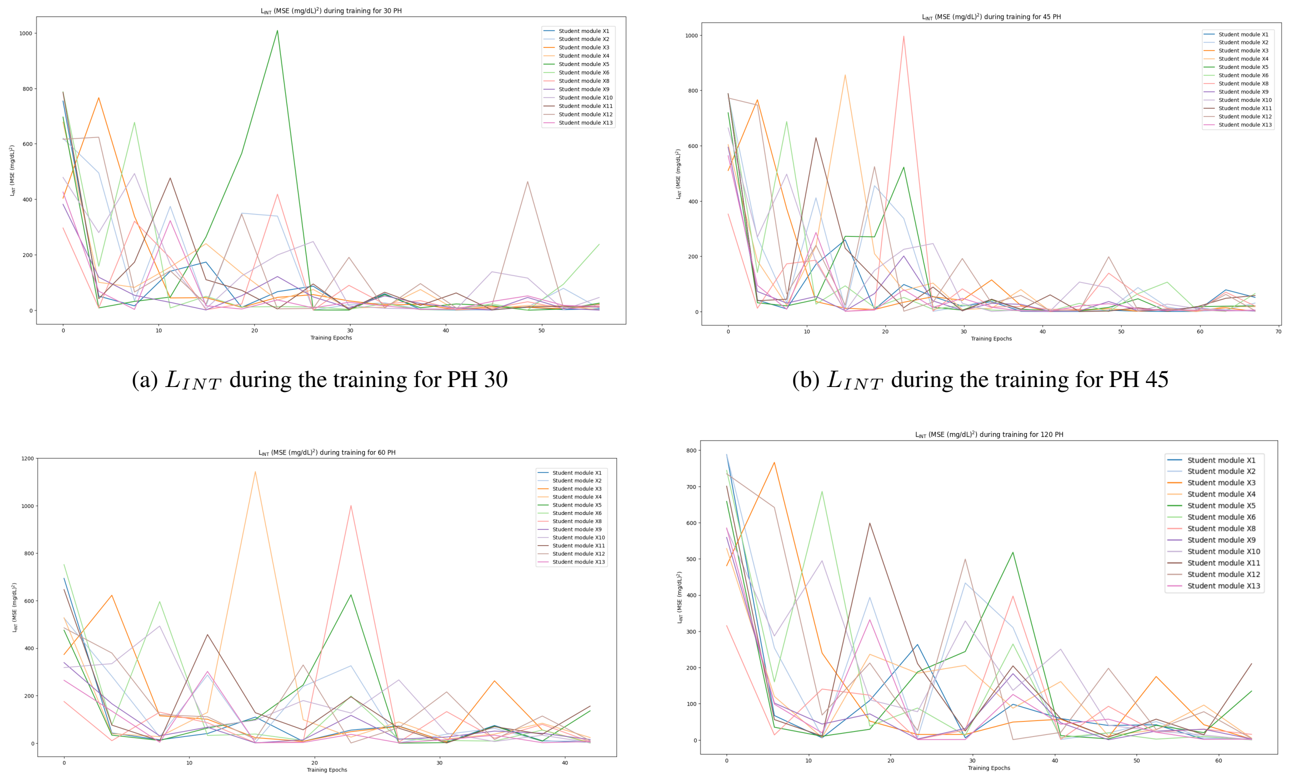
Zhengxuan Wu, Atticus Geiger, Josh Rozner, Elisa Kreiss, Hanson Lu, Thomas Icard, Christopher Potts, and Noah D. Goodman. Causal distillation for language models. 2021.

Interchange Intervention Training Applied to Post-meal Glucose Prediction for Type 1 Diabetes Mellitus Patients
(Supplementary Material)

Ana Esponera Gómez[1] **Giovanni Cinà**[1, 2, 3]

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

A CAUSAL ABSTRACTION ANALYSIS



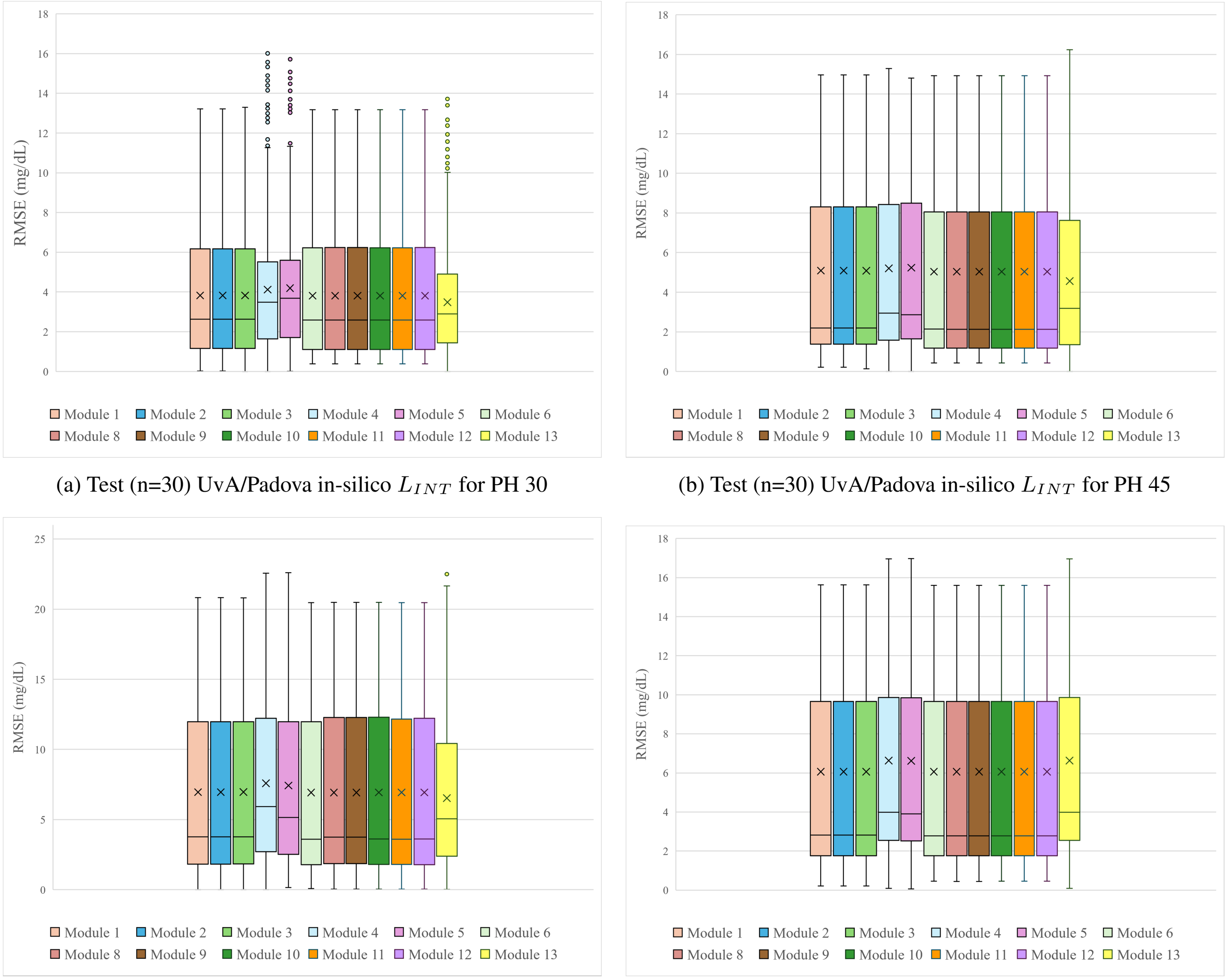
(a) L_{INT} during the training for PH 30

(b) L_{INT} during the training for PH 45

(c) L_{INT} during the training for PH 60

(d) L_{INT} during the training for PH 120

Figure 1: MLP IIT L_{INT} during the training using the acyclic *simglucose* as the causal model for PHs 30, 45, 60 and 120. The L_{INT} is grouped by modules.



(a) Test (n=30) UvA/Padova in-silico L_{INT} for PH 30

(b) Test (n=30) UvA/Padova in-silico L_{INT} for PH 45

(c) Test (n=30) UvA/Padova in-silico L_{INT} for PH 60

(d) Test (n=30) UvA/Padova in-silico L_{INT} for PH 120

Figure 2: MLP IIT L_{INT} during the testing using the acyclic *simglucose* as the causal model for PHs 30, 45, 60 and 120. The L_{INT} is grouped by modules.