Conformance Verification for Neural Network Models of Glucose-Insulin Dynamics

Taisa Kushner and Sriram Sankaranarayanan
firstname.lastname@colorado.edu
Univ. of Colorado Boulder, USA.

Marc Breton
mb6nt@virginia.edu
Univ. of Virginia, Center for Diabetes Technology, USA.

ABSTRACT

Neural networks present a useful framework for learning complex dynamics, and are increasingly being considered as components to closed loop predictive control algorithms. However, if they are to be utilized in such safety-critical advisory settings, they must be provably "conformant" to the governing scientific (biological, chemical, physical) laws which underlie the modeled process. Unfortunately, this is not easily guaranteed as neural network models are prone to learn patterns which are artifacts of the conditions under which the training data is collected, which may not necessarily conform to underlying physiological laws.

In this work, we utilize a formal range-propagation based approach for checking whether neural network models for predicting future blood glucose levels of individuals with type-1 diabetes are monotonic in terms of their insulin inputs. These networks are increasingly part of closed loop predictive control algorithms for "artificial pancreas" devices which automate control of insulin delivery for individuals with type-1 diabetes. Our approach considers a key property that blood glucose levels must be monotonically decreasing with increasing insulin inputs to the model. Multiple representative neural network models for blood glucose prediction are trained and tested on real patient data, and conformance is tested through our verification approach. We observe that standard approaches to training networks result in models which violate the core relationship between insulin inputs and glucose levels, despite having high prediction accuracy. We propose an approach that can learn conformant models without much loss in accuracy.

CCS CONCEPTS

• **Computing methodologies → Cross-validation**; • **Computer systems organization → Embedded and cyber-physical systems**; • **Applied computing → Systems biology**.

KEYWORDS

Conformance Verification, Neural Networks, Artificial Pancreas Systems

ACM Reference Format:
Taisa Kushner and Sriram Sankaranarayanan and Marc Breton. 2020. Conformance Verification for Neural Network Models of Glucose-Insulin Dynamics.

HSCC '20, April 22–24, 2020, Sydney, NSW, Australia

ACM ISBN 978-1-4503-7018-9/20/04…$15.00
https://doi.org/10.1145/3365365.3382210

In *23rd ACM International Conference on Hybrid Systems: Computation and Control (HSCC '20), April 22–24, 2020, Sydney, NSW, Australia.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3365365.3382210

1 INTRODUCTION

In the recent years, neural networks have risen in popularity due to their ability to "learn" complex phenomena from large data sets [28]. They are extensively being used for tasks such as classification, perception, and, increasingly, control of autonomous systems [35, 37]. In this work, we focus on the use of feed-forward neural networks as dynamic models for the physical, chemical and biological phenomena which underlie the glucose-insulin regulatory system in individuals with type-1 diabetes. An increasing number of papers in this area have proposed neural networks trained from patient data [42, 43, 48, 50, 53], in addition to a competition to improve prediction accuracy for blood glucose levels [8], and substantial investment in this area from the JDRF (formerly the Juvenile Diabetes Research Foundation) [36]. Some resulting models have been proposed for use in closed-loop artificial pancreas systems which automate the delivery of insulin to individuals with type-1 diabetes. Therein, neural networks serve as a prediction model for model predictive control (MPC) algorithms [22]. However, the question of whether neural network models are safe to use in such applications remains unanswered.

Mathematically, neural networks are complicated nonlinear functions learned through a process of regression on data. Today, given sufficient data, the training of these networks can be achieved using off-the-shelf tools such as Tensorflow [1] or PyTorch [51]. Although large amounts of data are often available, they can be laden with *biases*. This is particularly true in medical applications such as type-1 diabetes. In this work, we focus on identifying and explaining potential biases in data, developing a framework for rigorously testing if networks have incorrectly attributed correlations to causation, and propose methods for improving model conformance to correct causal relations based on known physiology.

Conformance to Scientific Facts: In the case of blood glucose prediction, neural network models often achieve good accuracy as measured by the root mean squared (RMSE) prediction error on test data. However, whether the models actually capture the key scientific facts governing regulation of glucose by insulin remains unknown. Within the human body, the hormone insulin is the key regulator of blood glucose levels. Insulin enables cells to uptake glucose from the blood, and also encourages glucose uptake by the liver, in the form of glycogen. Together, this results in a complex nonlinear decrease in blood glucose levels, and understanding this relation is central to safe control of T1D [12, 16, 31]. In this work, we focus on testing and ensuring model conformance to this key

property: *all else equal, increasing insulin input must cause a decrease predicted blood glucose levels.*

Failure of a model to conform to this property, especially one which is to be utilized in an advisory manner (eg. artificial pancreas systems), can have fatal consequences for an individual. Imagine a model that under some conditions predicts adding insulin increases blood glucose levels: if used in an advisory system, such a model could result in a controller *increasing* insulin when the blood glucose levels are low. This is potentially dangerous, and poses significant risks of coma or even death.

Domain of Validity. A related issue in utilizing neural networks pertains to their domain of validity. These can be roughly defined as the region of inputs for which the network produces valid outputs, and can be estimated by considering the convex hull of the set of learning points [14]. In general, how well a network is able to generalize depends on the distance of the unseen pattern to the domain of validity for the network. While previous work has focused on algorithms for estimating this domain [14, 52], our focus is on systematically quantifying how well a network performs within the domain of validity, and outside it. Rather than estimating the region, we give sensitivity values for a network's response to inputs increasingly outside the domain of validity and show how these values can differ between models and input ranges.

Contributions. In this work, we develop a range-estimation based approach for verifying conformance of neural network models to known scientific laws governing input-output relations within the model. Results are shown as applied to the real-world example of glucose-insulin regulatory models, a central component of artificial pancreas systems. While models show significant promise for prediction, their use in an advisory manner, such as within artificial pancreas systems, requires models be trustworthy and *conform* to known underlying physiology.

We focus first on explaining the data on which these models are trained, including patterns inherently present and potential biases which may arise. We then present our range-estimation based approach and protocol for testing model conformance, and demonstrate how it can be implemented to test conformance of multiple neural network models for blood glucose prediction. Models are built and tested on real patient data from the at-home phase of a previous clinical trial [6].

Remark: This work builds on an extended abstract which was recently presented at a special invited session of ICCAD 2019 [49]. This paper presents a detailed explanation of both the techniques involved and analysis of results on clinical trial data.

2 RELATED WORK

We cover related work focusing on definitions of conformance, in particular those used in formal verification, along with related work in verification of neural networks.

2.1 Conformance

Conformance, and conformance checking, is a broad term used to encompass methods for checking if an engineered artifact satisfies (conforms to) a desired specification. In numerous prior research studies, this term has been used when checking if a model satisfies properties of a reference model. Often times, the reference model has increased complexity that we desire to capture using a simplified model. Conformance specification checks whether the simpler model is able to capture all the key properties present in the more complex model. Multiple methods for doing so have been proposed including an event log based method by Carmona et al. [9], and an input-output conformance approach for model-based testing of algorithms by Jan Tretmans [62].

For cyber-physical systems, conformance testing has been used in applications ranging from checking similarity of traces obtained from a deployed system to an abstract formal model [67], to whether inputs can be crafted to identify behaviors of the implemented system which were not seen in the reference model [2]. This latter work of Abbas et al. extends the notion of conformance from a Boolean predicate, to a real-valued measure of distance between models. In our work, we likewise measure the degree of conformance, though do so through sensitivity analysis on inputs, rather than a measure distance. This enables us to study model conformance relative to specific input locations.

2.2 Neural Network Verification

Broadly, verification problems for neural networks can be separated into two categories: (1) properties defining an input-output relation for a single neural network; and (2) system-wide or end-to-end properties of a system with neural network components.

In terms of the first category, tools for checking networks tend to fall into categories of either property checking, or image computation. With respect to property checking, approaches have historically focused on checking if a condition, $\psi[y]$, on a network's outputs (y) holds whenever a precondition $\varphi[x]$ is satisfied by its inputs (x). Numerous approaches have been proposed in this domain, notably posing SMT problems over piece-wise linear abstractions of the activation functions proposed originally by Pulina and Tacchella [54]. Since the original publication, multiple improvements upon the SMT-based approach have since been proposed including the Reluplex tool [38], and Planet solver [24].

The second approach, image computation, refers to approaches which compute a range over the output y, given some pre-condition $\varphi[x]$. Towards this end, many tools have formulated a neural network's operation using mixed-integer linear programs (MILP) [20, 44], or have utilized abstract intepretation-based approaches, originally proposed for program analysis problems [15]. Key examples include an approach by Vechev et al which utilizes zonotopes as an abstract domain to perform image computation across a neural network [25], work by Xiang et al which uses an abstract domain consisting of the union of polytopes [70], and an approach which computes the abstract domain of symbolic intervals, which is implemented in the Reluval tool [65].

More recently, the image computation approaches mentioned above have been extended for use within a closed loop system where neural network components are used. The simplest such model consists of a neural network applying a feedback control to a physical process, modeled as an ordinary differential equation. This setup has often been used for performing reachability analysis for resulting closed loop behaviors [19, 32, 34, 59, 64, 68, 69].

2.3 Adversarial Inputs and Falsification

The susceptibility of neural networks to adversarial inputs was first shown in [60], spurring the development of a large range of techniques for producing adversarial images to dupe image classifiers. A related, though slightly differing, line of research involving searching for adversarial inputs has been the falsification and testing approach for systems. Similar to an adversarial input, the falsification problem consists in finding an execution of a system which violates a specified property. We mention a brief sampling of work in this field, especially as it related to control-specific tasks [3, 18, 63, 71].

In this work, we utilize an optimization-based approach to find what may loosely be considered as pairs of adversarial inputs to a network. However, our work differs from the above approaches as our concern is not on the identification of individual inputs which result in a property violation, but rather on testing model robustness across a range of inputs, and identifying specifically those violations which have high likelihood to be encountered in the field.

3 BACKGROUND: ARTIFICIAL PANCREAS

In this section, we will provide a brief background on type-1 diabetes mellitus (T1DM) [57], and the artificial pancreas. Further details are available from one of the many surveys in this area [11, 13, 29, 42]. T1DM is an autoimmune disease in which the body targets and destroys the pancreatic β-cells. Without these cells, the human body is unable to produce insulin, the hormone required in order for cells to be able to uptake glucose from the blood, the main source of energy for cells. This leads to a dangerous cycle of cell starvation, coupled with increasingly elevated blood glucose levels, as the body breaks down glycogen stores to release *more* glucose in order to feed the starving cells. If left untreated, this results in increasingly acidification of blood leading to coma and even death [10, 57]. Consequently, individuals with T1DM must take external insulin analogs to counteract the lack of insulin. In order to properly dose insulin, an individual must constantly monitor their blood glucose levels, anticipating future changes in the blood glucose levels due to impending meals and physical activities. This process is incredibly burdensome, and error prone [56]. On one hand, too little insulin results in elevated blood glucose levels (hyperglycemia), which can lead to long-term organ damage. On the other hand, too much insulin leads to extremely low blood glucose levels (hypoglycemia), which risks coma or even death.

Artificial pancreas systems refer to a closed or semi-closed loop system of medical devices which serve to automate the delivery of insulin to individuals with T1DM [13, 17, 43]. The advent of artificial pancreas is considered one of the most promising treatment strategies to improve patient health, with the potential to free up time for these individuals and lessen human error [41]. As they stand now, these systems consist of two devices, a continuous glucose monitor (CGM) which measures blood glucose levels, and an insulin pump, along with an algorithm connecting the two [13, 17, 43]. Central to these systems, is a model of human physiology which allows the algorithm to predict an individual's future blood glucose levels and dose insulin accordingly [4, 11, 17, 58]. Developing such models has been a nontrivial task to due to the complexity of the human-glucose regulatory system, individual's changing sensitivity to insulin, and the long (up to 90minute) delay of activity onset of insulin analogs, as well as their persistence in the body for upwards of 7 hours. Insulin doses are typically delivered in two different forms during open as well as closed loop insulin delivery:

(1) **Basal insulin** - this is insulin delivered continuously in the background at levels < 0.1 Units. Basal insulin is delivered to combat the rise in glucose levels resulting from the liver's "endogenous" release of glucose in the blood between meals.
(2) **Bolus insulin** - a single large dose of insulin, typically at levels > 1 Units. Boluses are subdivided into two types: (a) *meal boluses* that are delivered in anticipation of rising blood glucose levels due to a meal. This can be seen in data as a large insulin dose, followed by (or coinciding with) a rise in blood glucose from the meal. The fall in blood glucose caused by the insulin is not observed until 60-120 minutes later; and (b) *correction boluses*: these are given to bring down a high blood glucose level. They tend to be smaller in size than meal boluses, and are most often followed by a decrease in blood glucose levels. Such doses can vary but are typically of $0.1 - 2$ Units.

One-Sided Control: While the hormone insulin and its analogs can be dosed through the device to enable cells to uptake glucose from the blood, thereby lower blood glucose levels, the counter-regulatory hormone glucagon which serves to increase blood glucose level, cannot be dosed outside of a research setting. This is due to the current lack of availability of shelf-stable glucagon in a form which can be variably dosed for commercial use. As a result, if too much insulin is delivered to a patient, the device lacks the ability to counteract this dose and it is left to the individual to counteract the affects of excess insulin through a fast-acting external glucose supply (eg. drinking juice). This makes current commercially viable artificial pancreas systems *one-sided*.

3.1 Modeling Insulin Glucose Regulation

We will briefly overview mathematical models of insulin-glucose regulation. A detailed survey of mathematical modeling approaches can be found in Kushner et al [42].

Differential equation models have been important for modeling insulin-glucose regulation, starting with the pioneering work of Bergman [5]. Since then, many differential equation models have been proposed and refined through studies on patients using tracer labeled foods [12]. Prominent modeling efforts include the Hovoroka et al model [30, 31, 66] and the Dalla-Man et al model [16, 46]. The latter model is part of the UVa-Padova simulator for Type-1 Diabetes which has been approved by the US Food and Drug Administration (FDA) to replace animal trials for new closed loop control devices [40]. While such high fidelity ordinary differential equation (ODE) models of the glucose-insulin regulatory system exist, translating these models to specific patients, as is needed in artificial pancreas systems, has had limited success. A key reason lies in the large number of patient-specific parameters that govern the behavior of the models and must be identified. This is further complicated by the use of state variables that are hard if not impossible to measure. These include states that attempt to capture blood glucose concentration in a fictitious "remote chamber" [5] or the

plasma insulin concentration which requires radioactive studies to measure directly.

The nonlinear nature of ODEs and the difficulties of identifying parameters has spurred interest in a number of data-driven approaches for predicting blood glucose values using historic CGM and insulin pump data have been proposed, many of which utilize neural networks [26, 27, 33, 47, 53]. Additionally, interest in such neural network based models continues to grow as can be noted by a recent NIH supported "Blood Glucose Prediction Challenge" held as part of the KDH workshop in 2018 [8] and recent JDRF investment [36].

4 NEURAL-NETWORK BASED MODELS

While data-driven models have shown promise for developing patient-specific models towards use in artificial pancreas systems [23, 26, 27, 33, 47, 53], conformance of such models to known underlying physiological properties has not yet been tested. In this section, we explain a core conformance requirement for such models. Next, we present an overview of the type of data these models are trained on, and potential incorrect causal relations and biases which may potentially lead to models that fail to be conformant.

4.1 Conformance Property

In this work, we verify models with inputs of glucose and insulin and an output of future blood glucose levels. A key conformance property states that: "Insulin should *cause* blood glucose levels to decrease". In particular, we test that given all other factors remaining equal, if an insulin input is increased monotonically, predicted blood glucose levels must decrease. At a physiological level, insulin binds to receptors in cells that increase the uptake of glucose and causes the storage of blood glucose in the liver as glycogen. While nonlinear, this relationship is known to be monotonic.

However, data-driven models, such as neural networks, need not capture this natural *causal* relationship even if we have "sufficient" data. This is due to reliance of such data-driven methods on the most common *correlations* within the data, rather than underlying *causality*. Here we describe one common pattern in data which could result in models learning non-conformant dynamics (insulin results in blood glucose rise) due to the correlation between meals and insulin bolus:

(1) Large insulin boluses are commonly given before meals.
(2) The nutrients in the meal cause blood glucose levels to rise.
(3) The insulin bolus is input to the network but meal inputs are often unreliable or not present as input to the network (they are difficult to collect in a reliable manner).
(4) As a result, the data driven model "attributes" the rise in blood glucose levels to the insulin.

Since neural networks are opaque models, such an improperly learned causal relationship cannot explicitly be seen in a human readable format, however the danger this poses is easy to see: if a neural network model incorrectly attributes a rise in blood glucose to a large insulin dose rather than a meal, a network may lead to a decision that treating hypoglycemia with an insulin bolus is the optimal course of action since it causes blood glucose levels to rise. In this work, we utilize our conformance testing methodology to identify if such relations exists.

4.2 Structure

Figure 1 depicts the structure of neural network models analyzed in this paper. The neural networks considered here are feedforward neural network models, with inputs and structures selected to be in line with recently proposed models of blood glucose prediction [26, 27, 33, 47, 53]. The general network structure analyzed is a two-layer feed forward network utilizing ReLU (rectified linear unit) activation functions. The networks input longitudinal blood glucose data, as measured by a continuous glucose sensor, along with insulin pump doses. Inputs are restricted to the past 30 minutes, with readings obtained every 5 minutes. The output of the network is the predicted blood glucose value T minutes into the future, where T is the prediction horizon. Here, we will set $T = 60$ minutes. As observed in the literature, smaller networks perform better on the test data.

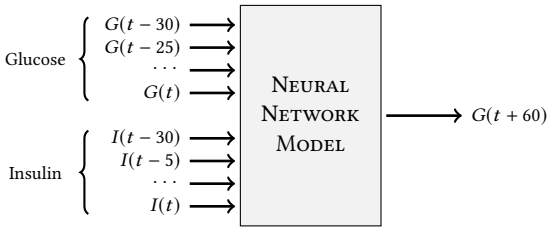

Figure 1: Structure of the predictive model for future blood glucose (BG) levels from past BG and insulin levels.

We consider three different types of network structures: M1-M3.

Basic Neural Network (M1): The basic structure M1 has two dense layers with 8 neurons per layer, with a single output neuron. This model was selected as a generalization of various feedforward neural network models recently proposed models of blood glucose prediction [26, 27, 33, 47, 53].

Split Structure (M2): Figure 2 shows a split first layer topology first considered by Dutta et al. [22]. This network "splits" the first hidden layer into two parts: one part connected just to the glucose inputs and the other to the insulin inputs before these are connected to a joint second hidden layer. The reason for the split is to mimic physiological models of insulin-glucose regulation, wherein the insulin inputs are combined to calculate a "insulin-on-board" that affects the future course of blood glucose levels. This model is included to test how conformance changes, and if it improves, when compared to a "standard" model, M1.

Split Structure with Monotonicity Constraints (M3): Finally, we consider the split structure in model M2 and additionally constrain the network to be *monotonic* with respect to the insulin inputs. Let N be a neural network with inputs $(x_1, \ldots, x_n)$, and whose activation functions are monotone with respect to their inputs.

Definition 4.1 (Negative Monotonic Neural Networks). An input x_i is said to be *negative monotonic* with respect to an output x of a network with monotonic activation functions iff the product of weights along each path from x_i to the output x is non-positive.

Let $F_N(x_1, \ldots, x_n)$ be the function computed by a neural network N.

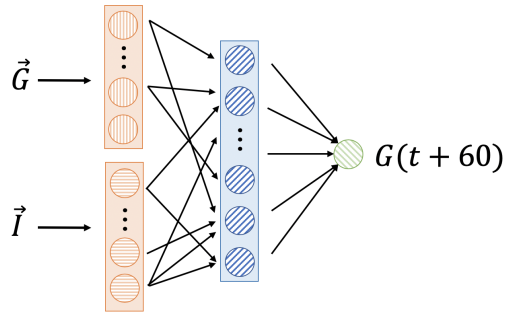

Figure 2: Neural network structure consisting of a split first layer with a fully connected second layer, as proposed in [22].

Lemma 4.2. *If an input x_i is negative monotonic (Def. 4.1) with respect to to an output x of a network N, then $F_N(x_1, \ldots, x_n)$ is a monotonically non-increasing function of x_i: i.e, for all $(x_1, \ldots, x_n)$ and $\hat{x}_i$,*

$$(x_i \leq \hat{x}_i) \implies F_N(x_1, \ldots, x_n) \geq F_N(x_1, \ldots, \hat{x}_i, \ldots, x_n)\,.$$

To ensure partial monotonicity (negative monotonicity with respect to insulin inputs), we follow a common procedure of imposing weight constraints on paths from monotone inputs to output. [55, 72]. Utilizing the split network structure of model M2 enables us to maintain unconstrained weights for paths from glucose inputs to the first layer, while placing non-positivity constraints on weights from insulin inputs to the first layer of *separate* neurons, and non-negativity constraints for the second layer and output neuron. To ensure that the constraints are respected during the backpropagation process, we perform a "projection" operation during the gradient descent that clips weights which violate the monotonicity constraints to 0 weights. The resulting model is denoted M3. This model is included to both demonstrate how weight constraints may be utilized to improve conformance, and test prediction accuracy against unconstrained models.

4.3 Data, Model Training and Test Accuracy

Models M1-M3 are trained and evaluated using previously collected real-world data from continuous glucose monitor (CGM) and insulin pump data from a cohort of twenty-four subjects. The data set was collected during the observation period of a clinical trial in individuals with type 1 diabetes mellitus (T1DM). The data set consists of average 37.8 ± 14 days of CGM and insulin logs from 29 adults and adolescents with T1DM, 11-57 years of age, and was collected during the observational period of a clinical trial [6]. Additionally, 358.0 ± 242.4 associated blood glucose monitor measurements are present. For each individual, 211.1 ± 181.5 meals and snacks are consumed and 258.2 ± 197.0 U of insulin boluses are injected. All measurements are taken at 5 minute intervals. To obtain training and test data, we clean and process this data to clean and extract sequences of contiguous glucose and insulin input values after ensuring that no meals occurred during the interval between input data and output prediction (since this would constitute a large external disturbance the model would be unable to capture).

In order to train the network, the data is separated using a standard 80/20 division between training and test data. Due to the high temporal correlation in blood glucose data, we ensure at least a 180 minute separation between data placed into the training versus testing bins. Data is otherwise divided randomly, after this condition is met. Models are trained using backpropogation using the Adam optimizer in Tensorflow[1], with training done to 100,000 epochs with an average time of 10min 42sec to train a model on a Macbook Pro with 16GB of RAM.

Table 1 shows the results along with a standard linear control model that simply predicts $G(t+\Delta) = G(t)$ (such a linear prediction has been shown to be surprisingly effective for predictions of up to 30 minutes). With respect to prediction accuracy, we find very little substantial difference between models M1, M2, as measured by the two main metrics for blood glucose prediction accuracy, root-mean square error (RMSE) and the number of predictions falling within 20% of the actual value.

Table 1: Model accuracy networks M1-M3, as well as a standard linear control model. Note accuracy does not differ significantly between models M1-M3, and prediction accuracy is improved over the control.

Model	RMSE (mg/dL)	Predictions within 20%
M1	46	63%
M2	47	62%
M3	46	61%
Control	56	58%

5 VERIFYING MONOTONICITY THROUGH RANGE ESTIMATION

We will now describe the setup for a formal verification approach for determining if the predictive networks have the appropriate relationship that increased insulin inputs decrease glucose predictions. Let N be a network whose inputs include a vector of $K_g > 0$ past glucose values $\vec{G}$ and a vector of $K_i > 0$ past insulin values $\vec{I}$, as depicted in Figure 1. In particular, the values $K_g = K_i = 7$ representing 30 minutes of history at intervals of 5 minutes.

To verify that a network N computing a function $F_N(x_1, \ldots, x_n)$ is monotonically non-increasing with respect to input x_i over some domain D, we would like to show that $\frac{\partial F_N}{\partial x_i} \leq 0$ for all inputs $(x_1, \ldots, x_n) \in D$. For functions computed by neural networks, this is doubly hard: (a) they are not differentiable everywhere, in particular if non-differentiable activation functions are used (eg. ReLu); and (b) the verification problem is quite hard to solve.

In order to solve this problem, we define the following notion of δ-Monotonicity, which enables us to analyze dynamics of networks which are not necessarily differentiable.

Definition 5.1 (δ-Monotonicity). Let $\delta > 0$ be a fixed limit and $F(x_1, \ldots, x_n)$ be any function over a domain D. Then, the function F defined over a domain D is δ monotonically increasing over input x_j iff for all inputs $(x_1, \ldots, x_n) \in D$ and for all increments $0 \leq s_j \leq \delta$, if $(x_1, \ldots, x_j + s, \ldots, x_n) \in D$, then

$$F(x_1, \ldots, x_j, \ldots, x_n) \leq F(x_1, \ldots, x_j + s, \ldots, x_n)\,.$$

Likewise, we say that F is δ monotonically decreasing iff

$$F(x_1, \ldots, x_j, \ldots, x_n) \geq F(x_1, \ldots, x_j + s, \ldots, x_n)\,.$$

In other words, upon increasing the input x_j by a quantity $s \in [0, \delta]$, the output monotonically increases (decreases for a monotonically decreasing function). The δ-monotonicity requirement makes no demands that that function F be continuous or differentiable.

It is easy to see that any monotonic function is also δ-monotonic but not vice-versa, since the domain D may not necessarily be connected. Nevertheless, we use δ-monotonicity as a working requirement for the insulin-glucose regulatory model. Specifically, fixing a minimal increment for insulin $\delta = 0.1$U, we wish to check that for the same glucose history, increasing any of the insulin inputs by more than δ units yields an overall decrease in the predicted blood glucose value.

5.1 Verifying Monotonicity in Neural Networks

In order to check if a neural network is δ-monotonic with respect to some input x_j, we leverage recent approaches for output range estimation of neural networks, given constraints over the inputs. The definition follows:

First, let N be a neural network computing function F_N with inputs $x_1, \ldots, x_n$ and a single output x, and let $\varphi[x_1, \ldots, x_n]$ represent linear inequality constraints over the network inputs.

Definition 5.2 (Range Estimation). The range propagation problem asks for an over-approximate interval $[\ell, u]$ such that the lower bound l satisfies

$$\ell \leq \min\ \{F_N(x_1, \ldots, x_n) \mid \varphi[x_1, \ldots, x_n] \text{ holds}\}\ ,$$

and the upper bound u satisfies

$$u \geq \max\{F_N(x_1, \ldots, x_n) \mid \varphi[x_1, \ldots, x_n] \text{ holds}\}\ .$$

The range estimation problem asks for a conservative over-approximation of the output of a network given constraints on the inputs. In recent years, a profusion of approaches to solve this problem has been proposed, employing ideas ranging from interval analysis to mixed integer optimization [7, 21, 39, 45, 61].

In order to solve the δ-monotonicity problem, we take a network, N, and create a composite network consisting of two identical copies of the network placed side-by-side, Fig. 3. We then restrict the input space such that the two copies share all inputs except the input at a single location at which we wish to test the "δ-monotonically decreasing" property (eg. index x_j), where they differ by amount ϵ, such that $0 \leq \epsilon \leq \delta$. Next, we take the difference of output ranges between these two instances of the network and define this difference as z. The input constraints φ are given as follows:

$$\begin{gathered}(x_1, \ldots, x_n) \in D \wedge \\ (x_1, \ldots, x_{j-1}, x_j + \epsilon, x_{j+1}, \ldots, x_n) \in D \wedge \\ x_j \leq x_j + \epsilon \leq x_j + \delta\end{gathered} \tag{1}$$

We will use the range estimation approach to estimate the maximum and minimum values of the output z.

Lemma 5.3. *A network N is δ-monotonically decreasing with respect to input x_j over a domain D if the maximum value of the output z of the composite network under the input constraints* (1) *is non-positive.*

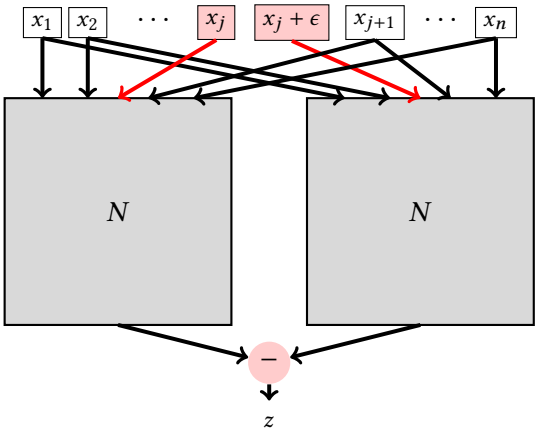

Figure 3: Scheme for checking δ-monotonicity of a network with respect to a specific input location, x_j. Here $0 \leq \epsilon \leq \delta$.

To apply this approach to the insulin-glucose regulation networks, we utilize the framework presented in Fig. 3, and perform analysis for each insulin input location by iterating through the locations in sequence, eg. $x_j = I(j)$ where $j = t - 30, t - 25, \ldots, t$.

In order to consider model dynamics under physiologically reasonable conditions, we employ constraints to the insulin and glucose profiles, described below. For each case we employ physiological constraints on the glucose domain D as:

(1) Blood glucose inputs must be with a "reasonable range" of $[40, 400]$ mg/dl.

$$\vec{G}_i \in [40, 400],\ i \in \{1, \ldots, 7\}\,. \tag{2}$$

(2) The maximum change of blood glucose levels over a 5 minute period is bounded:

$$|\vec{G}_i - \vec{G}_{i+1}| \leq 25,\ i \in \{1, \ldots, 6\}\,. \tag{3}$$

As we wish to test insulin sensitivity across a wide range of values, we allow insulin input at test location $I(j)$ to be within a bolus range of $[0, 5]$ Units, stepping through at $\delta = 0.1$ Unit steps. This range is identified to be within the domain of validity of the network through an analysis of the data set described in Section 4.3. In general, it is considered "unreasonable" to see two large insulin doses (boluses) within a 30 minute period. Thus, we restrict basal insulin rates to lie in the range $[0, 0.1]$ Units.

$$I_j \in [0, 5]\ \wedge\ \bigwedge_{k \neq j} (I_k \in [0, 0.1])\ . \tag{4}$$

This procedure enables us to test conformance of each insulin location independently. We then compute the *sensitivity* of each input location as the max change in output range for each δ change in input. The maximum positive sensitivity (increased output range due to increased insulin) and negative sensitivity (decreased output due to increased insulin) are computed independently.

5.2 Verification Results

We implemented the verification procedure for δ-monotonicity using a mixed integer linear programming (MILP) encoding along the lines used in many neural network analysis tools [7, 21]. The

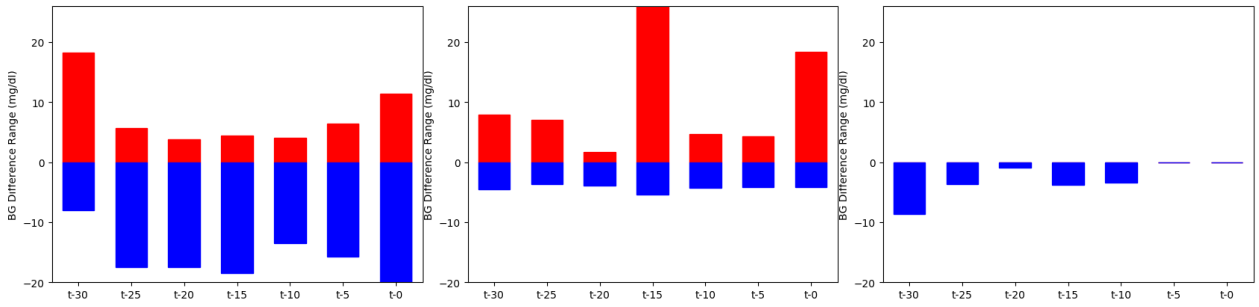

Figure 4: Plot showing sensitivity of each insulin input location for networks M1 (left), M2 (middle), and M3 (right). Differences above 0 are shown in red since they correspond to a violation of the conformance property. Note that all plots have the same y-axis ranges. The range of blood glucose values are $[40, 400]$ mg/dL, the maximum 5 minute change in blood glucose is restricted to 25 mg/dL and insulin boluses are allowed in the range $[0, 5]$ U.

verification is performed on the three networks M1-M3 described in section 4.2.

We obtain a picture of overall conformance by observing the maximum positive and negative sensitivities across the entire domain of insulin and glucose values. In Figure 4, we show how sensitivity differs across each insulin input location within the three networks. The insulin inputs are labeled $t-30, \ldots, t$ to denote when the insulin was given relative to the current time t. We note that networks M1 and M2 clearly violate the property of interest, at times exhibiting changes of almost 25 mg/dL Unit *increases* in glucose due to a 0.1 Unit increase in insulin. Note the alarming sensitivity of model M2 where blood glucose can rise by as much as 250 mg/dL/Unit of insulin at the $t-15$ input location. The use of a small value $\delta = 0.1$ that is comparable to basal insulin values further highlights the lack of conformance.

Unlike networks M1, M2, we find network M3 verifiably conformant in that the insulin inputs are verified to monotonically decrease the blood glucose predictions. Note that the observed sensitivity of the network's output to the change in insulin is also smaller for network M3 when compared to M1 and M2.

In addition to overall sensitivity, the use of MILP solvers allows us to obtain concrete examples of glucose and insulin traces which result in violations to the conformance property that increased insulin should decrease glucose levels. Table 2 provides one such example.

	−30	−25	−20	−15	−10	−5	0	+60 (Pred)
Ins.	0	0	0.1	0.09	0	0	0	
Gluc.	176	151	176	151	127	102	77	287
Ins.	0	0	0.1	0.2	0	0	0	
Gluc.	176	151	176	151	127	102	77	313

Table 2: Counterexample trace showing conformance violation for network M2 showing two different inputs: the same glucose values are input in both cases, the insulin values differ only at times $t-15$. However, the increased insulin results in an increased blood glucose level.

Although this counterexample is valid, it also demonstrates limitations of the formal approach thus far: (a) the pattern of BG values observed in this counterexample is seldom seen in actual patient data; and (b) the counterexample involves zero insulin delivered when the patient's BG levels are near the upper limit of the normal range. This scenario is also unrealistic in clinical practice.

In the following section, we address this limitation and check if conformance violations exist in more realistic scenarios by using the patterns of blood glucose values that are actually observed in the user data.

5.3 Data-Based Verification Results

In the previous section, we formulated optimization problem that checked for δ monotonicity for a range of "reasonable" glucose inputs and found networks M1 and M2 have significant conformance violations. The linear inequality constraint approach enabled us to properly test the networks by ensuring we capture all physiologically possible glucose and insulin profiles. However, as is highlighted by the counterexample in 2, the notion of what is a "reasonable" input is difficult to capture through linear inequality constraints alone. This begs the question: can we demonstrate lack of conformance under more realistic inputs?

In order to address this question, we perform a secondary data-based verification approach wherein the initial glucose traces are constrained to be those observed in data.

Data Source: As we are testing sensitivity to glucose traces and not model accuracy, initial glucose traces vectors consisting of 7 continuous values are pulled from both the training and testing subsets of our clinical trial dataset described previously in Section 4.3. We regard these values as samples from an underlying true distribution of values. The data set yielded $N = 10,800$ such sample inputs.

Analysis Approach: The analysis approach is identical to that presented in section 5.1, which formulates constraints described in (2), (3) and (4). However, rather than allow the glucose inputs to be decision variables of the resulting MILP problem, we fix the 7 glucose inputs to a sample $\vec{g}_j$ taken from the test data. In other words, the constraints (2) and (3) are removed and replaced simply

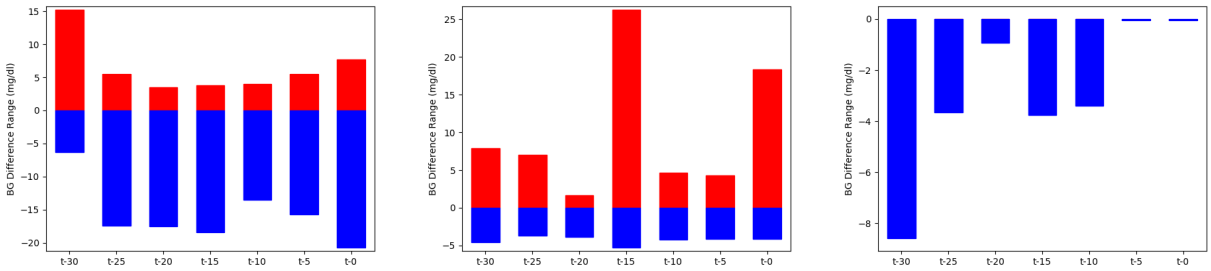

Figure 5: Plots showing network sensitivity at input locations $t-30,\ldots t$ for networks M1 (left), M2 (middle), and M3 (right), when glucose inputs are restricted to samples from clinical trial data. Differences above 0 correspond to a violation of the conformance property and are hence shown in red.

	−30	−25	−20	−15	−10	−5	0	+60 (Pred)
Ins.	0.1	0	0	0.21	0.1	0.1	0.1	
Gluc.	81	79	77	76	75	72	43	181.4
Ins.	0.1	0	0	0.31	0.1	0.1	0.1	
Gluc.	81	79	77	76	75	72	43	207.7

Table 3: Counterexample trace showing conformance violation for network M2 for differing insulin inputs at time $t-15$ using actual blood glucose measurements from clinical trial data-set: the same glucose values are input in both cases, the insulin values differ only at times $t-15$. However, the increased insulin results in an increased blood glucose level.

by a constraint: $\vec{G} = \vec{g}_j$, where $\vec{g}_j$ is a sample from the data. The insulin inputs, on the other hand, is allowed to vary as described in the constraints (4). As a result, we solve N different MILP instances, one for each sample $\vec{g}_j$, and compute the maximum/minimum over the outputs obtained from each MILP.

Results: Fig.5, presents the results of maximum positive and negative sensitivity at each insulin input location for models M1-M3 when computed over the 10,800 sample glucose inputs taken from the clinical trial data. We observe that the bounds are nearly identical to those seen in Figure 4 which used constraints on glucose ranges and difference between successive glucose inputs. This indicates that the alarming conformance violations detected in models M1-M2 were not edge cases and are observable under clinically observed glucose input values. One such case for model M2 is shown in Table 3 (compare with violation reported in Table 2). Note that the blood glucose levels in this counterexample are taken from actual clinical trial data and are thus viable in practice. Also note that the relatively small insulin dose is consistent with insulin delivery when blood glucose levels are low. However, we see an increase of 0.1 in the insulin given causes an increase in the predicted blood glucose values.

5.4 Domain of Validity

The domain of validity for a network is defined as the region of inputs for which the network produces valid outputs, which can be estimated roughly by considering the convex hull of the set of learning points [14]. Importantly, how well a network is able to generalize depends on the distance of the unseen pattern to the domain of validity for the network.

In this section, we present a detailed discussion of network dynamics and conformance results when tested within various restricted input ranges both within and outside the convex hull of learning points for both insulin and glucose values, and demonstrate how dynamics, even within the domain of validity, may vary significantly depending on distributions of underlying data.

	Mode	**Mean**	**Median**	**Max**	**Min**
Glucose	400	191.3	175	40	400
Insulin	0.0208	0.0614	0.0208	0	5.17

Table 4: Descriptive statistics for the data set on models M1-M3 were trained, separated by glucose (CGM) and insulin inputs.

Training Data: We first present details of the dataset available for training the predictive models. In the case of both insulin and glucose data, we find that while a wide range of values are represented, Table. 4.

However, we note distributions within these ranges are highly skewed, Fig. 7.

For insulin data, we find that 98% of values fall into the *basal* insulin category of < 0.1 Units. This is to be expected as basal insulin is typically delivered every 5-minutes throughout the day, while bolus values occur only around meal time (pre-meal or corrective post-meal bolus). With respect to glucose values, we note the particularly uneven distribution across the three clinically defined ranges:

(1) The *hypoglycemic* range of $[40, 70]$ mg/dL.
(2) The normal *euglycemic* range of $[70, 180]$ mg/dL.
(3) The *hyperglycemic* range of $[180, 400]$ mg/dL.

We note these ranges are bounded to $[40, 400]$ mg/dL due to sensor bounds and that levels about 300 mg/dL correspond to a dangerous condition called diabetic ketacidosis (DKA).

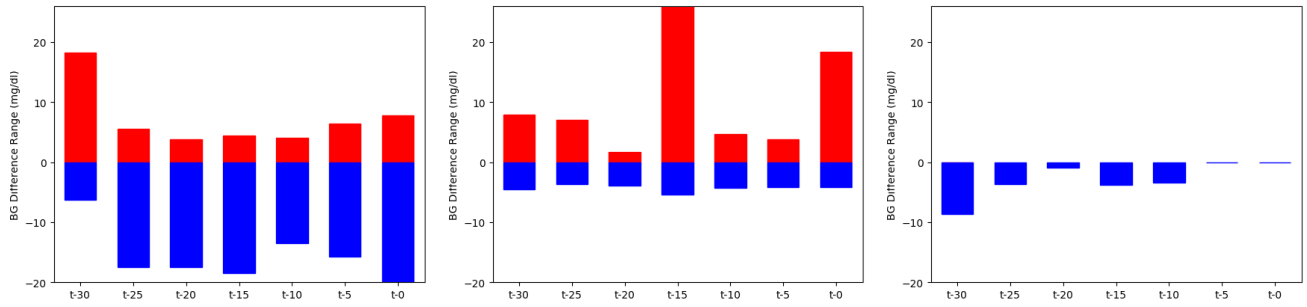

Figure 6: Effects of changing the 5 minute difference in BG values to 25 mg/dL for networks M1 (left), M2 (middle) and M3(right). The range of glucose inputs is [70, 180] mg/dL. Compare with the middle row Figure 8.

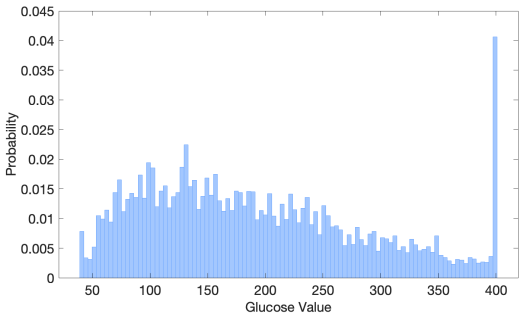

Figure 7: Distribution of glucose values within training dataset. Note sensor range is [40, 400].

Overall we find that 44% of values fall into the *euglycemic* range, 48% fall into the *hyperglycemic* range, and only 6% fall into the *hypoglycemic* range.

Effects of Restricting Glucose Profiles: While the training data ranges from 40 to 400mg/dL, the distribution of data within this range is quite skewed towards high glucose values, Fig. 7. Herein we demonstrate how model dynamics (sensitivities) are affected by the underlying distribution of training data.

We consider the three clinical scenarios of hypo, hyper, and euglycemia separately. As values above 300mg/dL are consistent with DKA, we restrict the hyperglycemic range to be [180,300] mg/dL. We also restrict the maximum 5 minute change in blood glucose levels to a more conservative 15 mg/dL, to better represent the typically observed changes in glucose.

Figure 8 shows the range of differences in network sensitivities when the blood glucose levels are constrained within the hyperglycemic, euglycemic and hypoglycemic ranges, respectively. Under these conditions, we note that networks M1 and M2 continue to show conformance violations but they are markedly smaller especially in the hyperglycemic range. On the other hand, the violations become much larger for the hypoglycemic and euglycemic conditions.

We propose two aspects which may contribute to the "less bad" conformance results when glucose values are constrained to the hyperglycemic range. First, the distribution of glucose values within the training dataset is skewed towards the hyperglycemic range with a mean of 191.3 mg/dL and 48% of values falling into this range (32% within the more restrictive non-DKA hyperglycemic range of [180,300] mg/dL). Additionally, the type of glucose-insulin relation most often observed within this range is the correction bolus: an individual takes a larger dose of insulin due to high blood glucose values, with no meal occurring, resulting in a bolus followed by a *drop* in glucose values. This could contribute to the networks becoming more biased towards learning the desired *negative* sensitivity relation. Albeit we note this is not enough as networks M1,M2 still exhibit non-conformance.

Furthermore, we find the split structure network, M2, has a notable change in output sensitivity when we allow the 5-minute change in blood glucose to be further restricted from 25mg/dL (Fig. 4) to 15mg/dL (Fig. 8). While sensitivities of network M1 change very slightly, network M2 shows decrease in max sensitivity of about 15 mg/dL for a 0.1 U change in insulin. We note the accepted clinically reasonable change in blood glucose is 5mg/dL/min (or, 25mg/dL per 5 minutes).

Effects of Varying Max Insulin: Interestingly, the verification results in Figures 4 do not change when the maximum insulin bolus is varied beyond the domain of validity of the networks ($0 - 5.17$) to the range $1 - 10.0$ Units of insulin. This suggests that the worst cases are also achieved under small insulin doses (see Tables 3, 2, for instance). As a result, changing the limit on maximum insulin has no effect on the worst cases showing conformance violations.

6 CONCLUDING DISCUSSION

With the ever increasing use of neural-network-based process models within advisory settings, it is vital that these networks conform to known scientific laws which govern the underlying processes. Whereas significant work has been done to identify adversarial cases and "explain" such models utilizing simpler models, until now, little work has been done to formally test the time-dependent dynamics of these models, and check conformance to known physiological facts. In this work, we develop a formal verification-based approach for checking dynamics of neural network models enables the analysis of time-varying dynamics of trained networks, and quantification of the extent to which a network conforms to a specified property.

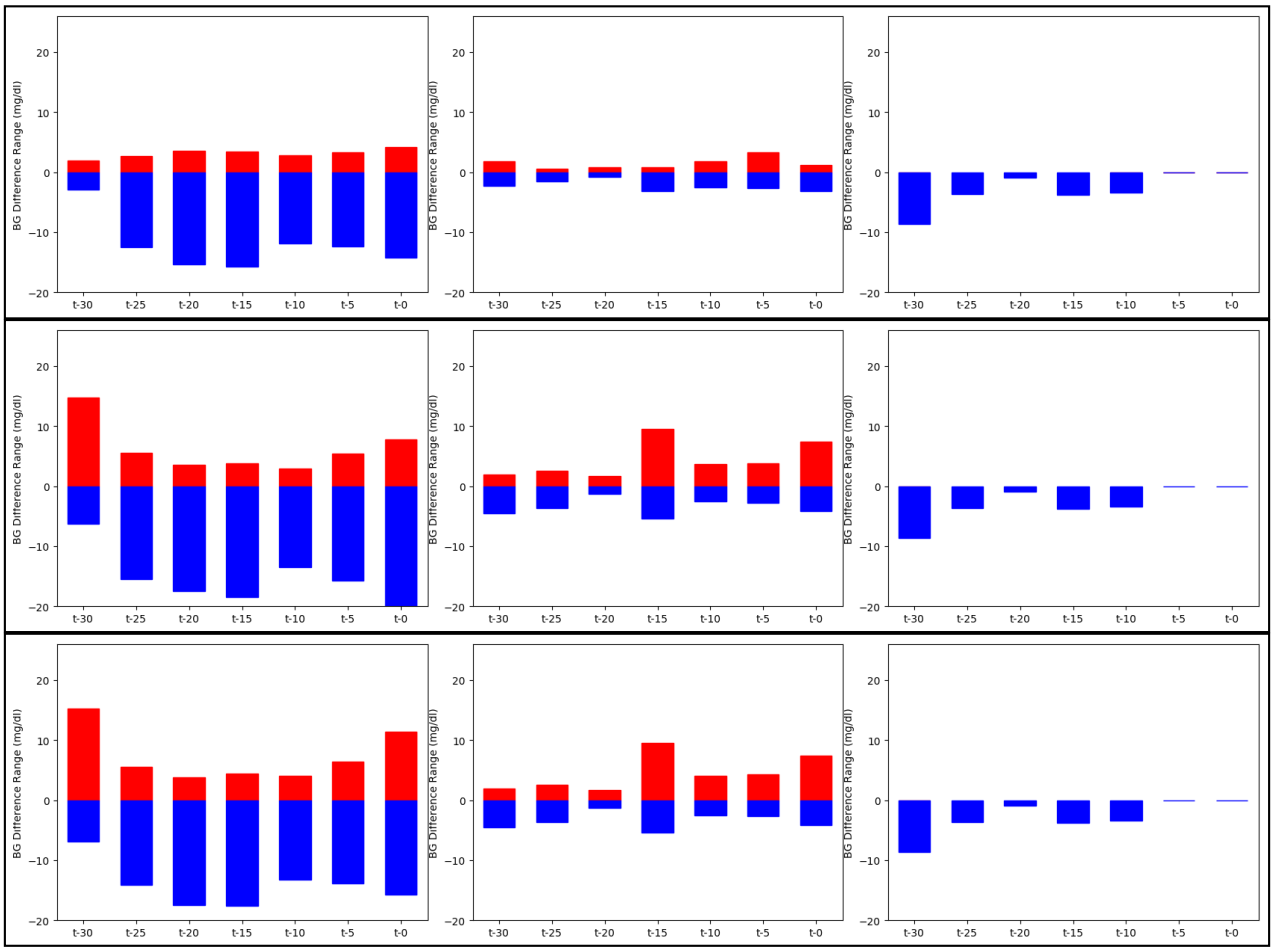

Figure 8: Plot showing model sensitivity of each input location for models M1 (left column), M2 (middle column), and M3 (right column). Top row: blood glucose ranges $[180, 300]$ **mg/dL, middle row: blood glucose ranges** $[70, 180]$ **mg/dL and bottom row** $[40, 70]$ **mg/dL. Max positive sensitivities are shown in red since they correspond to a violation of the conformance property. Note that all plots have the same y-axis ranges.**

We have demonstrated how this approach may be utilized to verify if the dynamics of neural network models used to predict blood glucose values conform to the key underlying process that increased insulin should result in decreased blood glucose values. We have shown how verification results may be utilized to understand sensitivity of networks, both for different locations in input history, as well as for different regions within the domain of validity of training data (eg. hypo, hyper, and euglycemic ranges).

The main conclusion of this paper is that obtaining monotonicity guarantees over a large range of possible inputs is hard for neural networks unless the monotonicity is guaranteed through a combination of a careful choice of network topology and constraints over the weights.

By providing a framework to rigorously test conformance of a network to key underlying physiology, we demonstrate how we may analyse and guarantee safer neural networks which may be used in an advisory manner even in cases of previously unseen patterns in data. This is a first step towards providing more formal guarantees to neural network models, and overcoming limitations in training data in the case where data may be limited, such as in medical research.

Acknowldegments: The authors are grateful to Dr. Rüdiger Ehlers for detailed comments and helpful discussions. We also acknowledge the efforts of Drs. Sergei Bogomolov and Bardh Hoxha in coordinating the repeatability evaluation for this paper. This work is supported by JDRF grant 1-SRA-2019-818-S-B and the US National Science Foundation (NSF) under award number 1932189. All opinions expressed are those of the authors and not necessarily of the NSF or JDRF.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, and et al. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. (2015). https://www.tensorflow.org/ Software available from tensorflow.org.

[2] Houssam Abbas, Georgios Fainekos, Sriram Sankaranarayanan, Franjo Ivancic, and Aarti Gupta. 2013. Probabilistic Temporal Logic Falsification of Cyber-Physical Systems. *Trans. on Embedded Computing Systems (TECS)* 12 (2013), 95–. Issue 12s.

[3] Yashwanth Annpureddy, Che Liu, Georgios E. Fainekos, and Sriram Sankaranarayanan. 2011. S-TaLiRo: A Tool for Temporal Logic Falsification for Hybrid Systems. In *TACAS*. 254–257.

[4] B. Wayne Bequette. 2013. Algorithms for a Closed-Loop Artificial Pancreas: The Case for Model Predictive Control. *J. Diabetes Science and Technology* 7 (November 2013), 1632–1643. Issue 6.

[5] R N Bergman, L S Phillips, and C Cobelli. 1981. Physiologic evaluation of factors controlling glucose tolerance in man: measurement of insulin sensitivity and beta-cell glucose sensitivity from the response to intravenous glucose. *Journal of Clinical Investigation* 68, 6 (Dec 1981), 1456–1467.

[6] Marc D. Breton, Stephen D. Patek, Dayu Lv, Elaine Schertz, Jessica Robic, Jennifer Pinnata, Laura Kollar, Charlotte Barnett, Christian Wakeman, Mary Oliveri, and et al. 2018. Continuous Glucose Monitoring and Insulin Informed Advisory System with Automated Titration and Dosing of Insulin Reduces Glucose Variability in Type 1 Diabetes Mellitus. *Diabetes Technology & Therapeutics* 20, 8 (Aug 2018), 531–540.

[7] Rudy Bunel, Ilker Turkaslan, Philip H. S. Torr, Pushmeet Kohli, and M. Pawan Kumar. 2017. Piecewise Linear Neural Network verification: A comparative study. *CoRR* abs/1711.00455 (2017). http://arxiv.org/abs/1711.00455

[8] Razvan Bunescu, Aili Guo, and Cindy Marling. 2018. The 3rd International Workshop on Knowledge Discovery in Healthcare. In *KDH*.

[9] J. Carmona, B. Van Dongen, A. Solti, and M. Weidlich. 2018. *Conformance Checking: Relating Processes and Models*. Springer.

[10] H. Peter Chase and David Maahs. 2011. *Understanding Diabetes (Pink Panther Book)* (12 ed.). Children's Diabetes Foundation. Available online through CU Denver Barbara Davis Center for Diabetes.

[11] Frederick Chee and Tyrone Fernando. 2007. *Closed-Loop Control of Blood Glucose*. Springer.

[12] Claudio Cobelli, David Foster, and Gianna Toffolo. 2000. *Tracer Kinetics in Biomedical Research*. Springer Science & Business Media.

[13] Claudio Cobelli, Chiara Dalla Man, Giovanni Sparacino, Lalo Magni, Giuseppe De Nicolao, and Boris P. Kovatchev. 2009. Diabetes: Models, Signals and Control (Methodological Review). *IEEE reviews in biomedical engineering* 2 (2009), 54–95.

[14] Pierre Courrieu. 1994. Three algorithms for estimating the domain of validity of feedforward neural networks. *Neural Networks* 7, 1 (1994), 169 – 174.

[15] Patrick Cousot and Rhadia Cousot. 1977. Abstract Interpretation: A unified Lattice Model for Static Analysis of Programs by Construction or Approximation of Fixpoints. In *ACM Principles of Programming Languages*. 238–252.

[16] Chiara Dalla Man, Robert A Rizza, and Claudio Cobelli. 2006. Meal simulation model of the glucose-insulin system. *IEEE Transactions on Biomedical Engineering* 1, 10 (2006), 1740–1749.

[17] Francis J. Doyle, Lauren M. Huyett, Joon Bok Lee, Howard C. Zisser, and Eyal Dassau. 2014. Closed-Loop Artificial Pancreas Systems: Engineering the Algorithms. *Diabetes Care* 37 (2014), 1191–1197.

[18] Tommaso Dreossi, Alexandre Donzé, and Sanjit A. Seshia. 2017. Compositional Falsification of Cyber-Physical Systems with Machine Learning Components. In *NASA Formal Methods (NFM)*. 357–372.

[19] Souradeep Dutta, Xin Chen, and Sriram Sankaranarayanan. 2019. Reachability Analysis for Neural Feedback Systems Using Regressive Polynomial Rule Inference. In *Proc. Hybrid Systems: Computation and Control (HSCC) (HSCC '19)*. ACM, New York, NY, USA, 157–168.

[20] Souradeep Dutta, Susmit Jha, Sriram Sankaranarayanan, and Ashish Tiwari. 2018. Output range analysis for deep feedforward neural networks. In *NASA Formal Methods Symposium*. Springer, 121–138.

[21] Souradeep Dutta, Susmit Jha, Sriram Sankaranarayanan, and Ashish Tiwari. 2018. Output Range Analysis for Deep Neural Networks. *Proceedings of NASA Formal Methods Symposium (NFM)* 10811 (2018), 121–138.

[22] Souradeep Dutta, Taisa Kushner, and Sriram Sankaranarayanan. 2018. Robust Data-Driven Control of Artificial Pancreas Systems Using Neural Networks. In *International Conference on Computational Methods in Systems Biology*. Springer, 183–202.

[23] Souradeep Dutta, Taisa Kushner, and Sriram Sankaranarayanan. 2018. Robust Data-Driven Control of Artificial Pancreas Systems using Neural Networks. In *Computational Methods in Systems Biology (Lecture Notes In Computer Science)*, Vol. 11095. Springer-Verlag, 183–202.

[24] Rüdiger Ehlers. 2017. Formal Verification of Piece-Wise Linear Feed-Forward Neural Networks. In *ATVA (Lecture Notes in Computer Science)*, Vol. 10482. Springer, 269–286.

[25] T. Gehr, M. Mirman, D. Drachsler-Cohen, P. Tsankov, S. Chaudhuri, and M. Vechev. 2018. AI2: Safety and Robustness Certification of Neural Networks with Abstract Interpretation. In *2018 IEEE Symposium on Security and Privacy (SP)*. 3–18.

[26] Eleni I Georga, Vasilios C Protopappas, Diego Ardigò, Michela Marina, Ivana Zavaroni, Demosthenes Polyzos, and Dimitrios I Fotiadis. 2013. Multivariate prediction of subcutaneous glucose concentration in type 1 diabetes patients based on support vector regression. *IEEE journal of biomedical and health informatics* 17, 1 (2013), 71–81.

[27] Eleni I Georga, Vasilios C Protopappas, Demosthenes Polyzos, and Dimitrios I Fotiadis. 2012. A predictive model of subcutaneous glucose concentration in type 1 diabetes based on random forests. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*. IEEE, 2889–2892.

[28] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

[29] Roman Hovorka. 2005. Continuous Glucose Monitoring and Closed-Loop Systems. *Diabetic Medicine* 23, 1 (2005), 1–12.

[30] R. Hovorka, V. Canonico, L.J. Chassin, U. Haueter, M. Massi-Benedetti, M.O. Frederici, T.R. Pieber, H.C. Shaller, L. Schaupp, T. Vering, and M.E. Wilinska. 2004. Nonlinear Model Predictive Control of Glucose Concentration in Subjects with Type 1 Diabetes. *Physiological Measurement* 25 (2004), 905–920.

[31] R. Hovorka, F. Shojaee-Moradie, P.V. Carroll, L.J. Chassin, I.J. Gowrie, N.C. Jackson, R.S. Tudor, A.M. Umpleby, and R.H. Hones. 2002. Partitioning Glucose distribution/transport, disposal and endogenous production during IVGTT. *Am. J. Physiol. Endocrinol. Metab.* 282 (2002), 992–1007.

[32] Chao Huang, Jiameng Fan, Wenchao Li, Xin Chen, and Qi Zhu. 2019. ReachNN: Reachability Analysis of Neural-Network Controlled Systems. *CoRR* abs/1906.10654 (2019). arXiv:1906.10654 http://arxiv.org/abs/1906.10654

[33] Janice J. Hwang, Lihong Jiang, Elizabeth Sanchez Rangel, Xiaoning Fan, Yuyan Ding, Wai Lam, Jessica Leventhal, Feng Dai, Douglas L. Rothman, Graeme F. Mason, and et al. 2018. Glycemic Variability and Brain Glucose Levels in Type 1 Diabetes. *Diabetes* 68, 1 (Oct 2018), 163–171.

[34] Radoslav Ivanov, James Weimer, Rajeev Alur, George J. Pappas, and Insup Lee. 2019. Verisig: Verifying Safety Properties of Hybrid Systems with Neural Network Controllers. In *Proc. Hybrid Systems: Computation and Control (HSCC) (HSCC '19)*. ACM, New York, NY, USA, 169–178.

[35] Kyle Julian and Mykel J. Kochenderfer. 2017. Neural Network Guidance for UAVs. In *AIAA Guidance Navigation and Control Conference (GNC)*.

[36] Juvenile Diabetes Research Foundation (JDRF). [n. d.]. Identification of Areas of Artificial Pancreas Algorithm Enhancements Through Big-Data Analysis (Part 1). ([n. d.]). http://grantcenter.jdrf.org/rfa/identification-of-areas-of-artificial-pancreas-algorithm-enhancements-through-big-data-analysis-part-1/.

[37] Guy Katz, Clark Barrett, David Dill, Kyle Julian, and Mykel Kochenderfer. 2017. Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks. (02 2017).

[38] Guy Katz, Clark Barrett, David L. Dill, Kyle Julian, and Mykel J. Kochenderfer. 2017. *Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks*. Springer International Publishing, Cham, 97–117.

[39] Guy Katz, Clark Barrett, David L. Dill, Kyle Julian, and Mykel J. Kochenderfer. 2017. Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks. In *Computer Aided Verification*. Springer, 97–117.

[40] Boris P Kovatchev, Marc Breton, Chiara Dalla Man, and Claudio Cobelli. 2009. In silico preclinical trials: a proof of concept in closed-loop control of type 1 diabetes. (2009).

[41] Aaron Kowalski. 2015. Pathway to Artificial Pancreas Revisited: Moving Downstream. *Diabetes Care* 38 (June 2015), 1036–1043. Issue 6.

[42] Taisa Kushner, B. Wayne Bequette, Faye Cameron, Gregory Forlenza, David Maahs, and Sriram Sankaranarayanan. 2019. *Models, Devices, Properties, and Verification of Artificial Pancreas Systems*. Springer, 93–131.

[43] Taisa Kushner, David Bortz, David Maahs, and Sriram Sankaranarayanan. 2018. A Data-Driven Approach to Artificial Pancreas Verification and Synthesis.. In *Intl. Conference on Cyber-Physical Systems (ICCPS'18)*. IEEE Press.

[44] Alessio Lomuscio and Lalit Maganti. 2017. An approach to reachability analysis for feed-forward ReLU neural networks. *CoRR* abs/1706.07351 (2017). arXiv:1706.07351 http://arxiv.org/abs/1706.07351

[45] Alessio Lomuscio and Lalit Maganti. 2017. An approach to reachability analysis for feed-forward ReLU neural networks. *CoRR* abs/1706.07351 (2017). arXiv:1706.07351 http://arxiv.org/abs/1706.07351

[46] Chiara Dalla Man, Francesco Micheletto, Dayu Lv, Marc Breton, Boris Kovatchev, and Claudio Cobelli. 2014. The UVA/PADOVA Type I Diabetes Simulator: New Features. *Journal of Diabetes Science and Technology* 8, 1 (2014), 26–34.

[47] Hrushikesh N. Mhaskar, Sergei V. Pereverzyev, and Maria D. van der Walt. 2017. A Deep Learning Approach to Diabetic Blood Glucose Prediction. *Frontiers in Applied Mathematics and Statistics* 3 (Jul 2017).

[48] Stavroula G Mougiakakou, Aikaterini Prountzou, Dimitra Iliopoulou, Konstantina S Nikita, Andriani Vazeou, and Christos S Bartsocas. 2006. Neural network based glucose-insulin metabolism models for children with type 1 diabetes. In *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE*. IEEE, 3545–3548.

[49] M. Narasimhamurthy, T. Kushner, S. Dutta, and S. Sankaranarayanan. 2019. Verifying Conformance of Neural Network Models: Invited Paper. In *2019 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*. 1–8.

https://doi.org/10.1109/ICCAD45719.2019.8942151

[50] Scott M Pappada, Brent D Cameron, Paul M Rosman, Raymond E Bourey, Thomas J Papadimos, William Olorunto, and Marilyn J Borst. 2011. Neural network-based real-time prediction of glucose in patients with insulin-dependent diabetes. *Diabetes technology & therapeutics* 13, 2 (2011), 135–141.

[51] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NIPS Workshop on Automatic Differentiation*.

[52] Marcello Pelillo. 1996. A relaxation algorithm for estimating the domain of validity of feedforward neural networks. *Neural Processing Letters* 3, 3 (Aug 1996), 113–121.

[53] Carmen Pérez-Gandía, A Facchinetti, G Sparacino, C Cobelli, EJ Gómez, M Rigla, Alberto de Leiva, and ME Hernando. 2010. Artificial neural network algorithm for online glucose prediction from continuous glucose monitoring. *Diabetes technology & therapeutics* 12, 1 (2010), 81–88.

[54] Luca Pulina and Armando Tacchella. 2012. Challenging SMT Solvers to Verify Neural Networks. *AI Commun.* 25, 2 (2012), 117–135.

[55] Joseph Sill. 1997. Monotonic Networks. In *Proceedings of the 10th International Conference on Neural Information Processing Systems (NIPS'97)*. MIT Press, Cambridge, MA, USA, 661–667.

[56] Jay S. Skyler. 2004. DCCT: The Study That Forever Changed the Nature of Treatment of Type 1 Diabetes. *British Journal of Diabetes and Vascular Disease* 4, 1 (2004). Cf. http://http://www.medscape.com/viewarticle/470738.

[57] Jay S. Skyler (editor). 2012. *Atlas of Diabetes: Fourth Edition*. Springer Science+Business Media.

[58] Garry M. Steil. 2013. Algorithms for a Closed-Loop Artificial Pancreas: The Case for Proportional-Integral-Derivative Control. *J. Diabetes Sci. Technol.* 7 (November 2013), 1621–1631. Issue 6.

[59] Xiaowu Sun, Haitham Khedr, and Yasser Shoukry. 2019. Formal Verification of Neural Network Controlled Autonomous Systems. In *Proc. Hybrid Systems: Computation and Control (HSCC) (HSCC '19)*. ACM, New York, NY, USA, 147–156.

[60] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *CoRR* abs/1312.6199 (2013). arXiv:1312.6199 http://arxiv.org/abs/1312.6199

[61] Vincent Tjeng and Russ Tedrake. 2017. Verifying Neural Networks with Mixed Integer Programming. *CoRR* abs/1711.07356 (2017). http://arxiv.org/abs/1711.07356

[62] Jan Tretmans. 2008. *Model Based Testing with Labelled Transition Systems*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1–38.

[63] Cumhur Erkan Tuncali, Georgios Fainekos, Hisahiro Ito, and James Kapinski. 2018. Simulation-based Adversarial Test Generation for Autonomous Vehicles with Machine Learning Components. In *2018 IEEE Intelligent Vehicles Symposium*. 1555–1562.

[64] Cumhur Erkan Tuncali, James Kapinski, Hisahiro Ito, and Jyotirmoy V. Deshmukh. 2018. Reasoning about safety of learning-enabled components in autonomous cyber-physical systems. In *Proc. Design Automation Conference, DAC 2018*. 30:1–30:6.

[65] Shiqi Wang, Kexin Pei, Justin Whitehouse, Junfeng Yang, and Suman Jana. 2018. Formal Security Analysis of Neural Networks using Symbolic Intervals. *CoRR* abs/1804.10829 (2018). arXiv:1804.10829 http://arxiv.org/abs/1804.10829

[66] M.E. Wilinska, L.J. Chassin, C. L. Acerini, J. M. Allen, D.B. Dunber, and R. Hovorka. 2010. Simulation Environment to Evaluate Closed-Loop Insulin Delivery Systems in Type 1 Diabetes. *J. Diabetes Science and Technology* 4 (January 2010). Issue 1.

[67] Matthias Woehrle, Kai Lampka, and Lothar Thiele. 2013. Conformance Testing for Cyber-physical Systems. *ACM Trans. Embed. Comput. Syst.* 11, 4, Article 84 (Jan. 2013), 23 pages.

[68] Weiming Xiang and Taylor T. Johnson. 2018. Reachability Analysis and Safety Verification for Neural Network Control Systems. *CoRR* abs/1805.09944 (2018). http://arxiv.org/abs/1805.09944

[69] Weiming Xiang, Hoang-Dung Tran, and Taylor T. Johnson. 2017. Reachable Set Computation and Safety Verification for Neural Networks with ReLU Activations. *CoRR* abs/1712.08163 (2017). arXiv:1712.08163 http://arxiv.org/abs/1712.08163

[70] Weiming Xiang, Hoang-Dung Tran, and Taylor T. Johnson. 2107. Reachable Set Computation and Safety Verification for Neural Networks with ReLU Activations. (2107). Cf. https://arxiv.org/pdf/1712.08163.pdf, posted on ArXIV Dec. 2017.

[71] Shakiba Yaghoubi and Georgios Fainekos. 2019. Gray-box adversarial testing for control systems with machine learning components. In *Proceedings of Hybrid Systems: Computation and Control*. 179–184.

[72] Seungil You, David Ding, Kevin Canini, Jan Pfeifer, and Maya Gupta. 2017. Deep Lattice Networks and Partial Monotonic Functions. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2981–2989.