Basal Glucose Control in Type 1 Diabetes using Deep Reinforcement Learning: An *In Silico* Validation

Taiyu Zhu, *Student Member, IEEE*, Kezhi Li, *Member, IEEE*,
Pau Herrero, *Member, IEEE*, and Pantelis Georgiou, *Senior Member, IEEE*

***Abstract*—People with Type 1 diabetes (T1D) require regular exogenous infusion of insulin to maintain their blood glucose concentration in a therapeutically adequate target range. Although the artificial pancreas and continuous glucose monitoring have been proven to be effective in achieving closed-loop control, significant challenges still remain due to the high complexity of glucose dynamics and limitations in the technology. In this work, we propose a novel deep reinforcement learning model for single-hormone (insulin) and dual-hormone (insulin and glucagon) delivery. In particular, the delivery strategies are developed by double Q-learning with dilated recurrent neural networks. For designing and testing purposes, the FDA-accepted UVA/Padova Type 1 simulator was employed. First, we performed long-term generalized training to obtain a population model. Then, this model was personalized with a small data-set of subject-specific data. *In silico* results show that the single and dual-hormone delivery strategies achieve good glucose control when compared to a standard basal-bolus therapy with low-glucose insulin suspension. Specifically, in the adult cohort (n=10), percentage time in target range $[70, 180]$ mg/dL improved from 77.6% to 80.9% with single-hormone control, and to 85.6% with dual-hormone control. In the adolescent cohort (n=10), percentage time in target range improved from 55.5% to 65.9% with single-hormone control, and to 78.8% with dual-hormone control. In all scenarios, a significant decrease in hypoglycemia was observed. These results show that the use of deep reinforcement learning is a viable approach for closed-loop glucose control in T1D.**

***Index Terms*—Deep learning, reinforcement learning, neural networks, dual-hormone delivery, artificial pancreas, diabetes.**

I. Introduction

Diabetes is a chronic disease which affects millions of people worldwide. It is characterised by elevated blood glucose (BG) which in the long term can lead to complications such as cardiovascular disease, retinopathy and nephropathy. Its global prevalence rate has reached epidemic proportions, doubling in the last 20 years [1]. There are two main types of diabetes, Type 1 and Type 2. Type 2 diabetes is characterised by the body ineffectively using insulin and can be usually treated with lifestyle interventions and oral medication. Type 1 diabetes (T1D) however is distinguished by insufficient insulin production by the pancreatic β-cell and therefore requires exogenous insulin administration. The standard insulin replacement therapy for T1D includes a bolus of fast-acting insulin to compensate the fast glucose increase after meal ingestion, and a basal insulin delivery through an injection of slow-acting insulin to keep glucose levels within target range in fasting conditions. Alternatively, basal insulin can be delivered through continuous infusion using an insulin pump with fast-acting insulin. Although software tools such as bolus calculators exist to support people with T1D to self-administer insulin, they still fall short to achieve optimal glycemic control [2]. Therefore, realising an automated system to deliver optimal insulin doses is one of the long-standing challenges in glucose management over the past decades [3].

Recent improvements in accuracy and reliability of continuous glucose monitoring (CGM) systems has allowed the development of a closed-loop insulin delivery system, also known as the artificial pancreas (AP), to automatically control BG levels in T1D [4]. An AP consists of, at least, a CGM sensor, a control algorithm, and an insulin pump. Additionally, some AP systems might also incorporate a glucagon pump to counter-regulate the action of insulin [5] and an activity monitor to quantify physical exercise [6]. Glucose measurements are captured by the CGM device every five minutes and are sent to the control algorithm which calculates the corresponding dose of insulin aiming at maintaining glucose level in a target range, which is then delivered by the infusion device. To date, most existing AP systems that have been evaluated in clinic have used a control engineering approach [7], such as the model predictive control [8], [9], and proportional integral control [10]. Other groups have also employed a bio-inspired approach [11] and an artificial intelligence approach (fuzzy logic) [12]. In particular, two of them, the Medtronic 670G and the Tandem Control-IQ have reached the commercialization stage. However, although these systems have been proven to improve glycemic control [10], [13], challenges remain, and further work is needed to achieve optimal therapeutic targets.

In recent years, powered by the large scale of available medical data and the rapid advances in computational power, machine learning, in particular deep learning, has increasingly been used in many healthcare applications that were out of reach in the past [14], especially in diagnostics and medical imaging [15], [16].

In the field of diabetes, the use of machine leaning has also attracted significant attention [17]. In particular, neural networks (NN) have achieved success in glucose forecasting [18] (fully-connected neural networks), [19], [20] (convolu-

This work was supported by EPSRC EP/P00993X/1. T. Zhu and K. Li have equal contribution.

T. Zhu, P. Herrero, P. Georgiou are with Centre for Bioinspired Technology, Imperial College London, London, United Kingdom. (e-mail: {taiyu.zhu17, p.herrero-vinias, pantelis}@imperial.ac.uk).

K. Li is with Institute of Health Informatics, University College London, London, United Kingdom. (e-mail: ken.li@ucl.ac.uk).

tional neural networks), [21]–[23] (recurrent neural networks (RNN)), and [24] (physiological-based networks). Of note, dilated RNN (DRNN) has performed particularly well in processing long-term dependencies and future glucose prediction [21], [23], [25]. Recently, another technique under the spotlight in the field of automatic insulin delivery is reinforcement learning (RL) [26]. RL is a machine learning framework for learning sequential decision-making tasks. Combining the techniques of RL and deep learning, deep RL improved the state of the art in various high-dimensional tasks [27], [28]. Many healthcare problems, such as drug delivery, and in particular, insulin delivery, can be seen as closed-loop sequential action-selection problems, which is what RL focuses on [27]. In the recent systematic review by Tejedor *et al.* on the application of RL to blood glucose control [26], almost all the included studies (i.e., 29 out of 30) employed traditional RL approaches, except for a recent work using a deep RL algorithm [29], which is compared with our work in Section IV-A. In our work, many of the latest deep RL advancements are applied for the first time to the problem of glycemic control. In fact, the use of deep RL in healthcare has been limited by several practical issues. Unlike successful deep RL applications in the virtual world, such as Atari video-games [27] or board-game Go [28], where an agent dynamically interacts with a virtual environment, performing such exploration on human subjects can be dangerous without proper safety supervision. Alternatively, deep RL algorithms can learn from existing collected data using experience replay. This process is called off-policy learning and plays an important role in practical RL algorithms. However, collecting the training data required is expensive and time consuming [30]. Fortunately, an FDA-accepted T1D simulator developed in collaboration between the University of Virginia (US) and the University of Padova (Italy) is available for developing and evaluating insulin and glucagon delivery strategies [31].

In this paper, we explore, *in silico*, the use of deep RL for closed-loop control of BG levels in T1D. The paper is organized as follows. Section II describes the architecture and algorithms of the proposed deep RL framework for glucose control. The performance of the proposed method is evaluated in Section III. Section IV compares the results with existing work and discusses the future work. Finally, we summarize the work in Section V.

II. Methodology

In this section, we state the problem of basal blood glucose closed-control in terms of deep RL. Then, we introduce a two-step framework, adapted from transfer learning, to develop, *in silico*, single and dual-hormone glucose controllers to be potentially used in clinical practice.

In particular, a deep Q-learning model [27] is employed to optimize insulin and glucagon delivery. Insulin and glucagon dose deliveries are treated as actions (a) taken by a stochastic policy, glycemic outcomes (e.g. percentage time in glucose target) are considered as rewards (r), and physiological variables are seen as states (s). A deep neural network (DNN) is used as a non-linear function approximator to estimate action-values, also referred to as a deep Q-network (DQN). Unlike previous artificial pancreas systems using traditional RL, our proposed method does not require prior knowledge of the glucose-insulin-glucagon metabolism. Instead, a stack of recurrent layers is used for processing multi-dimensional time series data. According to our previous studies [21], [23], the dilated connections improved RNNs performance in terms of BG level prediction with supervised learning and similar multi-dimensional input. Hence, in this work, we exploited DRNN layers to develop DQNs for glycemic control with deep RL. Because of its enlarged receptive field, the DRNN is able to capture the complexity of glucose-insulin-glucagon dynamics. Section VII-A in the Appendix also explains how the DRNN model was selected over other neural network architectures.

Fig. 1 depicts an overview of the system architecture used to develop the DQN controllers evaluated on the T1D *in silico* environment and to be potentially used in clinical trials. Algorithm 1 and Algorithm 2 correspond to the two-step learning framework in Section II-B.

A. Problem Formulation

The problem of basal glucose closed-loop control in T1D can be formulated as an infinite-state Markov decision process with noise, which is defined by a tuple $\langle S, \mathcal{P}, A, R, \gamma \rangle$ consisting of a state S (i.e., physiological state), a state transition function $\mathcal{P}$ (i.e., physiological model), an action A (i.e., insulin and glucagon control actions), a reward function R (i.e., glycemic outcomes), and a discount factor $\gamma \in [0, 1]$ (i.e., the importance of future glycemic outcomes). The agent in the environment takes an action $a \in A$ at each time step (i.e., each CGM measurement), and then its state $s \in S$ turns into the successor state s' according to $\mathcal{P}$. The policy to select action for given states is denoted by π. Maximizing the accumulation of expected reward $r_t = R(s_t, a_t)$ at each time step t is the target of RL. An action-value (Q-function) $Q^\pi(s, a)$ can be defined to computed this reward:

$$Q^\pi(s, a) = \mathbb{E}[\sum_{t'=t}^{\infty} \gamma^{t'-t} r_{t'} | s_t = s, a_t = a, \pi]. \quad (1)$$

The optimal action-value function $Q^*(s, a) = \max_\pi Q^\pi(s, a)$ offers the maximal values, which can be determined by solving the Bellman equation defined by

$$Q^*(s, a) = \mathbb{E}_{s'} \left[R(s, a) + \gamma \max_{a'} Q^*(s', a') \right], \quad (2)$$

The optimal action-value at the current state s is obtained by selecting the action that maximizes expected return with the optimal $Q^*(s', a')$ at the next state s'. Although this recursive equation can be estimated by an iterative update, linear and non-linear approximators are commonly used in RL for better generalization [27]. In this paper, DQNs are employed to approximate the action-values $Q(s, a; \theta) \approx Q^*(s, a)$ where θ represents the parameters of the neural networks.

1) Agent states: In the closed-loop glucose control problem, we collect the multi-modal data from the control system, as shown in Fig. 1, to form a multi-dimensional input vector D to approximate physiological state S. Specifically, D comprises the real-time continuous blood glucose levels G (mg/dL)

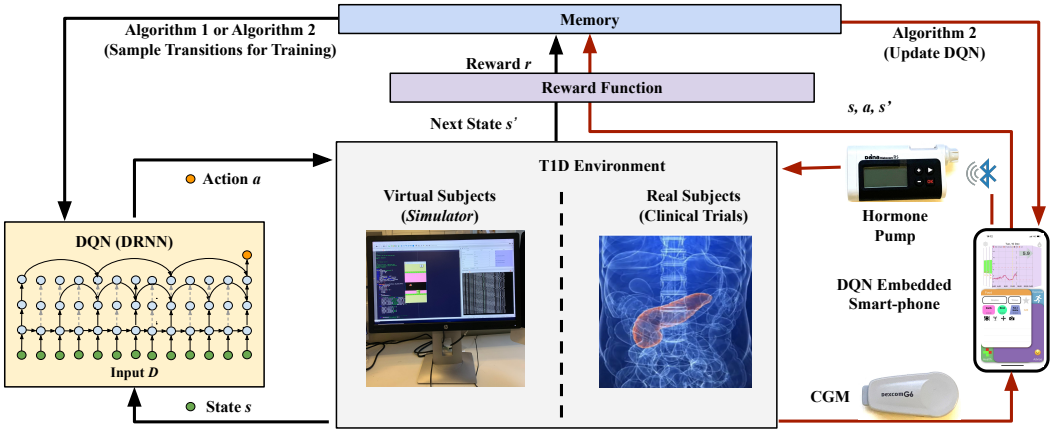


Fig. 1: The system architecture to implement deep RL on T1D in the simulator (black arrows) and clinical trials (red arrows).

measured with a CGM sensor, the carbohydrate estimation of meal ingestion M (g) recorder through a smartphone application, and hormone doses delivered by the infusion pumps, including the meal bolus insulin B, basal insulin Bas, and glucagon dose C. Thus, we have $D = \{G, M, I, C\} = [d_{t+1-L}, \cdots, d_t]^{\mathbb{T}} \in \mathbb{R}^{\mathbb{L}\times 4}$, where L is the length of the time steps vector, $I = B + Bas$ (Unit) represents the sum of meal bolus insulin and basal insulin. The approximated observation $o_t = s_t + e_t$ takes into account the errors or miss-estimations e_t in glucose measurements G, carbohydrate meal estimation M, and the meal insulin bolus B. Here B is computed from M with a standard bolus calculator [32]. From a deep RL perspective, the problem can be seen as an agent interacting with an environment over sequential time steps. Every five minutes, an observation o_t can be obtained from the environment, and an action a_t can be taken according to the agent's policy. We choose a five-minute time scale because this is the common sampling frequency for many commercial CGMs (e.g. Dexcom G6; Medtronic Guardian) and a typical setting for AP systems [4]. Please not that glucose-insulin-glucagon dynamics are quite slow; hence a shorter sampling period is unlikely to improve the outcomes of an artificial pancreas system.

2) Actions: Following the same framework, we provide two types of delivery strategies for different pump settings. For people with T1D wearing insulin pumps, the action space is defined by modifying the basal insulin rate (BR) as follows: {suspension of BR, 0.5*BR, BR, 1,5*BR, 2*BR}. For those wearing dual-hormone pumps, the action space is defined by the following options: {suspension of BR, 0.5*BR, BR, 1,5*BR, 2*BR, delivering glucagon}. Note that the value of BR is subject-specific and is known in advance. Based on previous works, we fix glucagon doses to 0.3 µg/kg for all individuals and constraint the total amount of delivered glucagon to a maximum of one mg per day [33]. This dosage has also been tested in clinical trials with two formulations of glucagon, which demonstrates efficacy and safety [34].

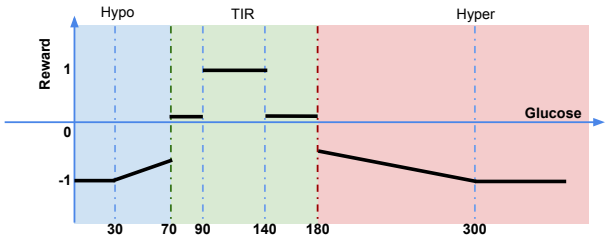


Fig. 2: Visualization of the employed reward function in terms of the glucose level (mg/dL) in the next state.

3) Rewards: The desired performance of closed-loop glucose control is to maintain BG in a target range of 70-180 mg/dL. By using an empirical approached aiming at maximising time in range (TIR) and minimising hypoglycemia, the following piece-wise reward function was selected.

$$r_t = \begin{cases} 1, & 90 \le G_{t+1} \le 140 \\ 0.1, & 70 \le G_{t+1} < 90 \ \& \ 140 < G_{t+1} \le 180 \\ -0.4 - (G_{t+1} - 180)/200, & 180 < G_{t+1} \le 300 \\ -0.6 + (G_{t+1} - 70)/100, & 30 \le G_{t+1} < 70 \\ -1, & \text{else.} \end{cases} \tag{3}$$

As depicted in Fig. 2, the agent receives a positive reward if the BG level for the next state is in the target range and a negative reward otherwise. If the BG is below 30 mg/dL or above 300 mg/dL, we terminate exploration and restart the simulator. Different evaluated reward functions are presented in Section VII-B of the Appendix.

B. Two-step Learning Framework

First, we perform long-term generalized training to obtain a population model for the hormone delivery strategies. We use dilated recurrent neural networks [21] for modeling the multi-dimensional time series including glucose levels, hormone doses, and meal intake. Note that other inputs affecting glucose levels, such as physical exercise, could also be considered. To

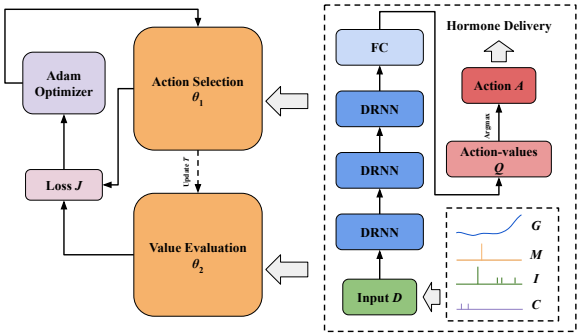


Fig. 3: The diagram of the propose double DQN. The structure of the neural network is the same for both action selection and value evaluation, which consists of an input layer, a stack of DRNN layers, a fully-connected (FC) layer and output. The input data includes BG series from CGM G, meal M, insulin I and glucagon C.

train the model, each basal hormone delivery (at five-minute intervals) is regarded as an action taken by the agent, while the glucose level on the next time step is set to the reward by the criteria of time in range (Equation 3). Secondly, by initializing the weights obtained from the population model, we have a model with good initial performance. With a transfer learning process, we individualize the DQNs according to personal characteristics and safety constraints with a small subject-specific data-set. Safety constraints in the AP refer to a set of safety measures based on the observations by monitoring systems (e.g., CGM measurements), estimation of the metabolic state of the subject (e.g. insulin on board), and meal ingestion, to prevent or mitigate possible harmful BG events [35]. A safety supervision system can comprise multiple safety constraints tasked with potentially dangerous events that may arise in a clinical setting (e.g. manual inputs constraints, glucose sensor saturations, insulin and glucagon delivery limits).

During clinical trials, the data for training is usually very limited, thus we aim at fast learning performance. Therefore, we use a double DQN with modified importance sampling to further optimize approximated action values. A state-of-art technique is employed to accelerate learning processes, where prioritized experience replay samples important transitions more frequently [36], [37]. To avoid overestimating the action values, a double DQN decouples action selection and value evaluation by two separate neural networks [38], as shown in Fig. 3. The second step is suitable for a clinical trial setting, where the model is able to adjust itself in a relatively short period of time.

C. Generalized DQN Training

In the first step, we use the simulator to generate an environment by using the average T1D subject for each one of the virtual cohorts (i.e. adult and adolescent). Compared to standard RNNs, DRNNs are preferred as DQNs for learning the delivery strategies. The large receptive field brought by dilation is powerful to extract features from glucose time series, where the dilated skip connection can be represented as

$$c_t^{(l)} = f\left(n_t^{(l)}, c_{t-d^{(l)}}^{(l)}\right), \tag{4}$$

where $c_t^{(l)}$ is the cell in layer l at time t, $n_t^{(l)}$ is the input to layer l at time t, $d^{(l)}$ denotes the dilation of layer l, and $f(\cdot)$ represents the output function of RNN cells. As shown in Fig. 3, we use three DRNN layers with exponentially increasing dilation, to process the multi-dimensional time-aligned sequence and extract high-level features. Then training is carried out in the simulator with double DQN weights θ_1, θ_2, where action selections θ_1 and value evaluations θ_2 are obtained from two separate neural networks. According to Equation (2), the action-selection networks are trained with the loss as

$$\begin{aligned} J_{DQ}(Q) = \mathbb{E}_{(o,a,r,o')\sim\rho}[(r+ \\ \gamma Q(o', a'; \theta_2) - Q(o, a; \theta_1))^2], \end{aligned} \tag{5}$$

where ρ is a mini-batch with transitions (o, a, r, o') sampled from the memory pool, and $a' = \arg\max'_a Q(o', a'; \theta_1)$ is chosen by the action selection DQN in Fig. 3. Thus, the Q-function can be updated as

$$\begin{aligned} Q_{\theta_1}(o, a) \leftarrow Q_{\theta_1}(o, a) + \alpha(r+ \\ \gamma Q_{\theta_2}(o', \arg\max_a Q_{\theta_1}(o', a)) - Q_{\theta_1}(o, a)), \end{aligned} \tag{6}$$

where α is the learning rate, and the weights of θ_1 are copied to θ_2 with a fixed period. We optimize the learning rate by Adam method at each iteration [39]. The corresponding pseudo-code is presented in Algorithm 1.

Algorithm 1 Generalized DQN Training

Input: the environment E with average T1D subject parameters Is provided by the simulator, update period T_G, ε-greedy
Initialize DQNs with random weights θ_1, θ_2, replay memory $\mathcal{B}$
for steps t$\in 1, 2, ..k$ **do**
    Sample action from $a \sim \pi(Q_{\theta_1}, \varepsilon)$, observe o' in E_{Is}, calculate r, store (o, a, r, o') into $\mathcal{B}$
end for
repeat
    Sample action from $a \sim \pi(Q_{\theta_1}, \varepsilon)$, observe o' in E_{Is}, calculate r, store (o, a, r, o') into $\mathcal{B}$
    Sample a mini-batch uniformly from $\mathcal{B}$ and calculate loss $J_{DQ}(Q)$
    Perform a gradient descent to update θ_1
    if t mod $T_G = 0$ **then** $\theta_2 \leftarrow \theta_1$ **end if**
until converge

For each meal, a standard dose of bolus insulin is delivered, and the agent explores random hormone delivery actions (single or dual) under policy π that is ε-greedy with respect to Q_{θ_1}. Human intervention could reduce training time and improve initial performance, but it would cause potential bias during the training process [40]. For *in silico* trials, random actions are tested with great flexibility and no safety concerns, as a great advantage of using a simulation environment. In

this case, we can train the agent for a long time until the loss converges, so human intervention is not necessary. At the end of the generalized training, a population model consisting of a double DQN with weights θ_1 and θ_2 is obtained.

D. Personalized DQN Training

After developing a generalized model, we fine-tune the model by transfer learning with regards to the personal characteristic. We fetch the weights and features from the generalized model, then train the personalized DQNs within a data-set corresponding to a short period of time with safety constraints. We can choose to fine-tune all layers of the generalized model or to retain the weights of some of the earlier layers and only fine-tune a higher-level portion of the network to avoid over-fitting. In experiments, we found that earlier layers contain more generic features (e.g. insulin suspension during the trend of hypoglycemia) that should be useful for all the subjects with T1D.

Here a method modified from [37] is used for calculating the loss of policy-generated data. Specifically, $J_n(Q)$ has an n-step returns (n= 12) to propagate values of actions to earlier states $r_t + \gamma r_{i+1} + \cdots + \gamma^{n-1} r_{i+n-1} + \max_a Q(o_{i+n}, a)$, and $J_{L_2}(Q)$ is an L2 regularization loss applied to θ to mitigate over-fitting. Prioritized experience replay samples the transitions with a probability Pr_i proportional to its importance priority [36], which is computed from previous data and normalized afterwards,

$$Pr_i = \frac{p_i^\alpha}{\sum_i p_i^\alpha}, \quad p_i = |\delta_i| + \epsilon', \tag{7}$$

where $\alpha \in [0, 1]$ determines the level of using prioritization, p_i is the priority of transition i calculated from last temporal-difference (TD) error δ_i and ϵ' is a small positive constant. It allows the DQN to more frequently replay transitions with higher TD error. In addition, to ensure that hormones are delivered safely in the clinical trial, constraints $\mathcal{C}$ are applied to the suggested action before execution. Here we use a simple strategy for the safety constraints: suspending basal insulin or glucagon when the current BG level is below 80 mg/dL or over 160 mg/dL, respectively. In practice, the trend and prediction of BG levels can also be used in the safety constraints for early interventions. With proper training of the generalized model and adequate safety constraints, this algorithm can be adopted in a clinical trial setting. The corresponding pseudo-code detailing the algorithm is presented in Algorithm 2.

Algorithm 2 Personalized DQN Training

Input: replay memory $\mathcal{B}$ and DQNs weights θ_1', θ_2' from generalized training; individual environment E, safety constraints $\mathcal{C}$, update period T_P, parameter λ_1, λ_2,
Initialize personalized DQNs weights $\theta_1 \leftarrow \theta_1'$, $\theta_2 \leftarrow \theta_2'$
Initialize replay memory $\mathcal{D}$, merging $\mathcal{B}$ with priorities
for steps i$\in 1, 2, ..N$ **do**
    Sample action from policy $a \sim \pi(Q_{\theta_1})$,
    if a subject to $\mathcal{C}$ **then** execute a **end if**
    Observe (o', r) in E
    Store (o, a, r, o') in $\mathcal{D}$, overwriting the samples previously merged from $\mathcal{B}$
    Sample a mini-batch from $\mathcal{D}$ by modified importance sampling Pr and update the transition priority
    Calculate loss $J(Q) = J_{DQ}(Q) + \lambda_1 J_n(Q) + \lambda_2 J_{L_2}(Q)$
    Perform a gradient descent to update θ_1
    if t mod $T_P = 0$ **then** $\theta_2 \leftarrow \theta_1$ **end if**
end for

III. Experiments

Following the architecture evaluation setup depicted in Fig 1, we conducted experiments to evaluate, *in silico*, the effectiveness of proposed deep RL framework with the UVA/Padova T1D Simulator [31]. As stated in Section II-A2, we use two settings of control actions in the proposed deep RL (DRL) algorithm: single-hormone (DRL-SH) and dual-hormone delivery (DRL-DH). Following a transfer learning strategy, we started with a long-term exploration with 1,500 simulated days to obtain a stable generalized model using Algorithm 1, then, we performed personalized training for each individual in the cohort (i.e. adult and adolescent) with 30 simulated days using Algorithm 2. Due to the significant amount of data required, the generalized model is meant to be trained in the simulator, whereas the personalized model training has the potential to be done in a clinical setting. Finally, the personalized models were tested in a period of 90 days.

A. Experimental Setup

1) In Silico environment: The UVA/Padova T1D simulator provides an interactive environment for the agent to explore and learn the policy. We introduced additional intra-subject variability in the meal protocol scenario and the parameters of the T1D model [41]. In particular, we selected four meals as the daily pattern (average cases: 7 am (70 g), 10 am (30 g), 2 pm (110 g), 9 pm (90 g)) with meal-time variability (STD = 60 min) and meal-size variability (CV = 10%). The meal-duration was set to 15 minutes. A misestimation of carbohydrate amount between -30% and $+10\%$ with uniform distribution was applied. The reason we used this skewed distribution is that the underestimation of carbohydrate content is more common than overestimation in real-life conditions, according to a cross-sectional study with 50 T1D subjects [42]. Variability for meal absorption and carbohydrate bioavailability were set to 30% and 10%, respectively. The variability of insulin sensitivity was considered to be 30% for adult cohort and 20% for adolescent cohort, which are created by the scenario function in the subjects' own profile. These values of variability were selected based on available physiological knowledge and to achieve the glycemic outcomes commonly observed in such populations when treated with standard therapy [43]. We saved intra-day and intra-person variability for each subject and used the same scenarios for all the evaluated methods, i.e. same daily events and variability time series, in order to have a fair comparison. We utilized

TABLE I: The testing performance of glucose control on the adult virtual cohort

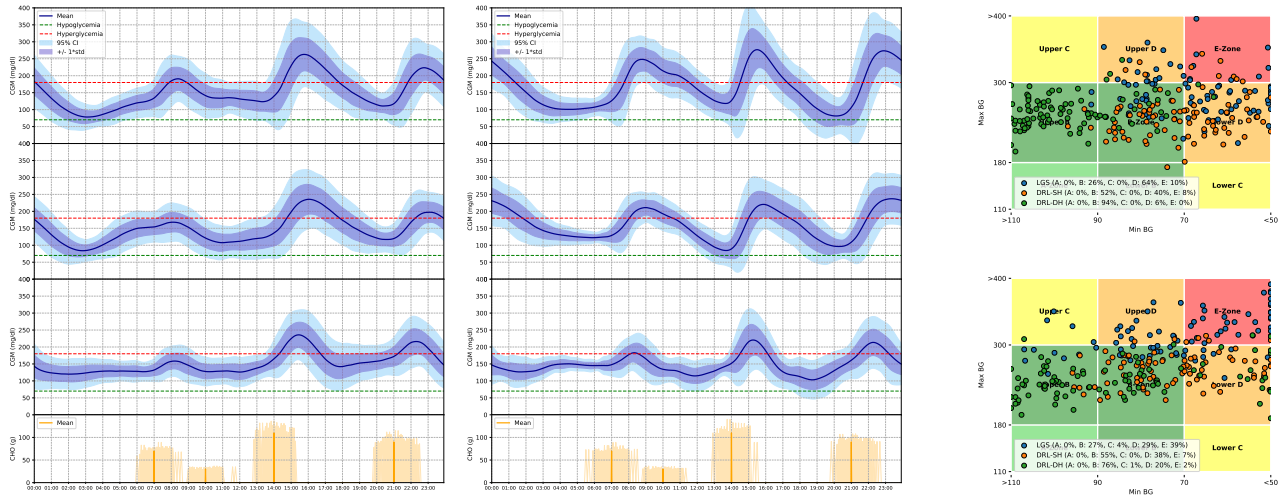
Method	TIR (%)	Hypo (%)	Hyper (%)	Mean (mg/dL)	RI
LGS	77.55±6.78	2.87±1.38	19.58±5.79	140.78±8.23	2.52±0.89
DRL-SH	80.94±7.00*	2.06±1.33*	17.00±5.82	140.36±5.98	2.28±0.72
DRL-DH	85.55±7.33**,†	1.92±1.90*	13.81±6.67**,†	140.12±8.13	2.16±0.65†

Symbol * indicates statistical significance ($p \leq 0.05$) with with respect to the low-glucose suspension (LGS) and † indicates statistical significance ($p \leq 0.05$) with with respect to the single-hormone DRL (DRL-SH). A double symbol (e.g. ‡) indicates statistical significance ($p \leq 0.01$).

TABLE II: The testing performance of glucose control on the adolescent virtual cohort

Method	TIR (%)	Hypo (%)	Hyper (%)	Mean (mg/dL)	RI
LGS	55.50±14.68	6.93±4.69	37.57±11.64	162.15±20.46	4.76±2.70
DRL-SH	65.85±16.30**	5.51±3.37	28.63±14.36**	151.18±18.26**	3.99±2.43**
DRL-DH	78.83±6.60**,†	2.64±1.96**,‡	18.53±6.48**,†	149.96±8.83**	2.94±0.99**,‡

Symbol * indicates statistical significance ($p \leq 0.05$) with with respect to the low-glucose suspension (LGS) and † indicates statistical significance ($p \leq 0.05$) with with respect to the single-hormone DRL (DRL-SH). A double symbol (e.g. ‡) indicates statistical significance ($p \leq 0.01$).



(a) An adult subject simulation (b) An adolescent subject simulation (c) CVGA plots

Fig. 4: Visualization of the experiment results for T1D subjects. **(a)** and **(b)**: Performance of the three methods on an adult subject and an adolescent subject over the three-month testing period: (Top-to-bottom) LGS, DRL-SH, DRL-DH, carbohydrate distribution. The average BG levels are shown in solid blue lines, and the hypo/hyperglycemia thresholds are shown in dotted green/red lines. Blue shaded regions show the 95% confidence interval (CI), and the purple shaded regions indicate the standard deviation. **(c)**: The control variability grid analysis (CVGA) plot for the adult (Top) and the adolescent (Bottom). The blue, orange and green dots represent the LGS, DRL-SH and DRL-DH results, respectively.

the 10 virtual adults and 10 virtual adolescents, plus the corresponding average subjects, for generalized training.

2) Baseline method: As a baseline method, a low-glucose insulin suspension (LGS) strategy, commonly found in sensor-augmented insulin pumps, was employed [44]. LGS systems have been proven to reduce hypoglycemia by suspending basal insulin delivery [45]. For meal bolus calculation, a standard bolus calculator was used [32].

B. Results

To evaluate the performance of the proposed algorithms and compare them against the baseline method, we selected five standard glycemic metrics commonly employed by the diabetes technology community [46]. These include: percentage time in the glucose target range of $[70, 180]$ mg/dL (TIR), percentage time below 70 mg/dL (i.e. hypoglycemia) (Hypo), percentage time above 180 mg/dL (i.e. hyperglycemia) (Hyper), mean BG levels (Mean), and risk index (RI). Results are expressed by mean values and standard deviations (mean $\pm SD$).

Table I and Table II shows the results of the three tested methods evaluated on the adult and adolescent cohorts, respectively. Compared to LGS therapy, both single-hormone and dual-hormone DRL models improve the glucose control performance by reducing hypoglycemia, hyperglycemia and increasing TIR in the two cohorts. Of note, the dual-hormone DRL model significantly increases the mean TIR with a notable decrease of risk index, achieving the best performance.

Mean BG levels are maintained in the adult cohort, while the improvement is significant in the adolescent cohort.

For demonstration purposes, Fig. 4 graphically displays the performance of the three evaluated methods for a chosen adult and a chosen adolescent over a three-month testing period. In particular, the glucose profile over 24 hours (mean $\pm SD$) (i.e. ambulatory glucose profile) and the control variability grid analysis (CVGA), a commonly used tool for evaluating closed-loop insulin delivery techniques, were employed [47]. Note that the displayed results in Fig. 4 are consistent with the numerical results corresponding to the overall population presented in Tables I and II. Regarding the CVGA, it is worth noting the significant improvement achieved by DRL-DH when compared to LGS. In particular, the percentage of points in the A+B zones increases from 26% to 94% for the adult cohort and from 27% to 76% for the adolescent cohort.

IV. Discussion

A. Comparison with State-of-the-art

In the presented *in silico* experiments, when compared against a low-glucose insulin suspension technique, the proposed methodology based on deep RL achieves superior performance in terms of glycemic outcomes. Comparing the proposed technique with existing closed-loop insulin delivery techniques, although interesting, is a challenging task due to the difficulty in replicating the testing scenarios and the tuning of the controllers. Hence, a head-to-head comparison has not been performed. However, although not directly comparable, an informal comparison with existing works in the literature on RL for insulin and glucagon delivery has been done. In [48], the authors propose an RL-based controller and achieve the adult TIR of 89% on the UVA/Padova simulator, which is close to the performance achieved by our DRL-DH model. Note that in this previous work both basal and bolus insulin delivery are optimized, while in our work only basal insulin delivery is optimized using different variability in the simulator. In [49], Ngo and colleges use RL to optimize control parameters in glycemic models without providing comparable TIR results. In a later paper [29], the authors propose a DQNs algorithm to control single hormone (insulin) delivery and they evaluate it on the previous version of the UVA/Padova simulator (version S2008). However, no comparable glycemic outcomes are provided. Therefore, our work not only proposes a novel deep RL algorithm for insulin and glucagon delivery but also serves as a benchmark for the future evaluation of other control algorithms. In W3PHIAI-20 workshop [50], we briefly reported some preliminary results corresponding to the dual-hormone delivery configuration. In this paper, we extend this preliminary work by developing a new model for single-hormone delivery, improve the previous algorithms, use more realistic scenarios, and introduce a new baseline method for comparison purposes. To the best of our knowledge, this is the first study that systematically evaluates, *in silico*, a deep RL algorithm to control blood glucose levels with single-hormone and dual-hormone delivery, using the latest T1D simulator (version S2013) [31] and additional intra-subject variability.

B. Limitations and Future Work

Although the DQN models achieved superior control performance *in silico*, clinical validation is still required. There are many uncertainties and perturbations in real-world scenarios, and the main limitation of the simulator is over-estimating the efficacy of glycemic interventions. Despite being able to reliably model glucose-insulin dynamics in T1D, the current version of FDA-accepted version of the UVA/Padova simulator [31] lacks the effect of physical exercise and health conditions (e.g., recurrent illness), which are known to significantly influence insulin sensitivity in people with T1D. In particular, the effect of physical activity has been proven to be very complicated to model. Thus, the modelling of the insulin-mediated and non-insulin-mediated effect on muscle glucose need to be further assessed and developed through more research [51]. Therefore, in this work, physical activity and heath conditions have not yet been taken into account. However, in future work, if new features become available in the simulator, we will incorporate them with the proposed models. Meanwhile, there is rapid development in deep RL, and we plan to explore the latest advances in this area, such as model-based RL [52], which have the potential to further improve glycemic outcomes and accelerate the training process. Although we selected DNN architectures following the steps in Section VII-A (Appendix), it is worthy to test alternative DNNs in the future, such as one-dimensional convolutional neural networks (CNNs), convolutional RNNs, and bidirectional RNNs. Following the proposed setup and framework, it is convenient to implement other deep learning or RL techniques in basal glucose control.

C. Towards Clinical Trials

In the past years, the technological advances in the field of diabetes technology and mobile phones have increased the connectivity between mobile apps, CGM and insulin pumps. As a result, many researchers have integrated control algorithm (single-hormone and dual-hormone) into apps to automatically administer or recommend hormone delivery and have evaluated them in clinical trials [11], [53]–[56] .

We have developed the deep RL models using TensorFlow, hence it is easy to implement such models on smartphones, or embedded devices, by means of TensorFlow Lite converter. This has been previously done by our group for implementing a DNN model on an app for T1D management [20], [22]. This algorithm has the potential to be continuously be trained and refined by the new incoming data from devices (e.g. CGM, pump, activity monitor) and user input (e.g. meals). Therefore, the algorithm proposed in this work can be implemented in a mobile app without much extra work (Fig. 1).

V. Conclusion

With the aim of overcoming the challenge of blood glucose control in T1D, we propose a novel deep RL algorithm for optimizing basal insulin and glucagon delivery. Dilated RNNs are applied to the structure of double DQNs to develop personalized models through a two-step framework that involves transfer learning. When compared to the baseline method with low-glucose insulin suspension, the proposed methodology

significantly improves glycemic outcomes in a virtual adult and adolescent population. This works shows that the proposed approach has the potential to be adopted in a clinical setting.

VI. Acknowledgement

We would like to thank Chengyuan Liu and Mariam Sarfrazc for their help and assistance. The work is supported by EPSRC EP/P00993X/1 and the President's PhD Scholarship at Imperial College London (UK).

VII. Appendix

A. Neural Network Selection

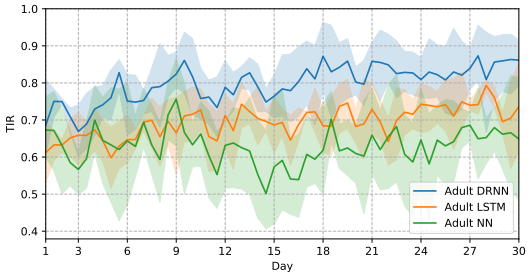
Fig. 5 shows the TIR results achieved with the different neural network architectures that we evaluated in the experiments. Considering that the input data is a multi-dimensional time-aligned sequence, we assumed that an RNN-based model would be a good candidate to map the multiple-step historical data. Therefore, we explored conventional long short-term memory (LSTM), NNs with five fully-connected layers and DRNNs as the potential structure of DQNs. The LSTM architecture has recently achieved great success in time-aligned tasks, but in our case, it obtains lower TIR results than the DRNN. NNs are commonly used in DQNs as a basic structure. However, the NN curve in Fig. 5 shows large variability and lower mean TIR. Less variability indicates a better capability to account for within-subject variability. Thus, the NN structure was discarded. Regarding the DRNNs, the generalized model achieves a good initial performance at the beginning of personalized training. In addition, the DRNN curve has a positive trend and small variability, which indicates its effectiveness at adjusting the models for a specific subject through a short period of time. Finally, DRNN prediction models ranked top in Blood Glucose Level Prediction Challenge in 2018 [21], outperforming various DNNs (e.g. one-dimensional CNNs and bidirectional RNNs). Therefore, DRNNs were naturally selected as the DQN modules for this work.

B. Reward Function

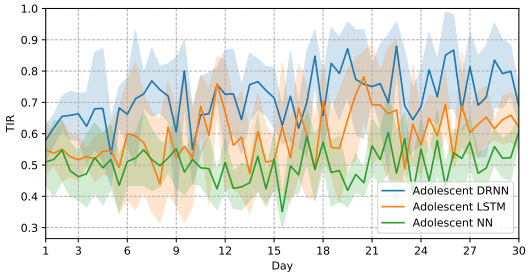
GL Range (mg/dL)	Reward Scheme 1	Reward Scheme 2	Reward Scheme 3	Reward Scheme 4
0-30	-10	-1	-1	-1
30-70	-1	-0.5	$-0.5+\frac{GL-70}{80}$	$-0.6+\frac{GL-70}{100}$
70-90	+0.1	+0.1	+0.1	+0.1
90-140	+1	+1	+1	+1
140-180	+0.1	+0.1	+0.1	+0.1
180-300	-1	-0.5	$-0.5-\frac{GL-180}{240}$	$-0.4-\frac{GL-180}{200}$
300+	-10	-1	-1	-1
TIR Score (%)	75	88	86	93

TABLE III: Reward functions and corresponding scores

Crafting reward functions for RL models is one of the most crucial factors determining the model performance. Note that this performance evaluation is only applicable when the TIR score converges to a fixed value through the training stage for the adult with Algorithm 1 and dual-hormone. Different reward schemes tested, together with their corresponding model scores, can be seen in Table III. In experiments, we started with a piece-wise step function referred to as Reward Scheme 1, then we constrained the reward function range within [-1, 1] to improve stability (Reward Scheme 2). Afterwards, we introduced slopes into the reward function to make the agent's response to glucose changes smoother (Reward Scheme 2). Note that the agent faces an increasing penalty as glucose moves up in the hyperglycemic range, or down in the hypoglycemic range. Finally, with minor adjustments on the slopes, the best TIR score was obtained by Reward Scheme 4, where a TIR of 93% can be achieved after 1.3 million training steps.



(a) Adult Cohort



(b) Adolescent cohort

Fig. 5: TIR results (mean, 95% CI) corresponding to DRL-DH during the personalized training for the adult and adolescent cohorts. The blue, orange and green lines show the results of DRNN, LSTM and NN models, respectively.

C. Hyper-parameters

In Table IV, we list the hyper-parameters that have been used in this work. For each parameter, we performed limited tuning based on the state-of-art DQN [57]. All the parameters are identical across all the virtual subjects.

References

[1] N. Cho, J. Shaw, S. Karuranga, Y. Huang, J. da Rocha Fernandes, A. Ohlrogge, and B. Malanda, "IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045," *Diabetes research and clinical practice*, vol. 138, pp. 271–281, 2018.

[2] S. Schmidt, M. Meldgaard, N. Serifovski, C. Storm, T. M. Christensen, B. Gade-Rasmussen, and K. Nørgaard, "Use of an automated bolus calculator in mdi-treated type 1 diabetes: the boluscal study, a randomized controlled pilot study," *Diabetes care*, vol. 35, no. 5, pp. 984–990, 2012.

[3] G. Quiroz, "The evolution of control algorithms in artificial pancreas: A historical perspective," *Annual Reviews in Control*, 2019.

[4] B. Kovatchev, "A century of diabetes technology: signals, models, and artificial pancreas control," *Trends in Endocrinology & Metabolism*, 2019.

[5] A. Haidar, "Insulin-and-glucagon artificial pancreas versus insulin-alone artificial pancreas: A short review," *Diabetes Spectrum*, vol. 32, no. 3, pp. 215–221, 2019.

Parameter	Value
Exploration before learning k	2000
Generalized network update period T_G	1000
Generalized DQN ε-greedy	0.5→0.01
Personalized network update period T_P	100
Discount factor γ	0.9
Adam learning rate	1×10^{-5}
Batch size	32
Number of time steps L	12
Replay buff size $\mathcal{B}$	5000
Prioritization exponent α	0.3
Importance-sampling exponent β	0.4→1.0
Multi-step return λ_1	0.1
L2 regularization λ_2	1×10^{-5}
Cell type	Vanilla RNN
DRNN dilation	[1, 2, 4]
Hidden nodes of DRNN layers	[32, 64, 128]

TABLE IV: List of hyperparameters

[6] M. D. DeBoer, D. R. Cherñavvsky, K. Topchyan, B. P. Kovatchev, G. L. Francis, and M. D. Breton, "Heart rate informed artificial pancreas system enhances glycemic control during exercise in adolescents with t1d," *Pediatric diabetes*, vol. 18, no. 7, pp. 540–546, 2017.

[7] F. J. Doyle, L. M. Huyett, J. B. Lee, H. C. Zisser, and E. Dassau, "Closed-loop artificial pancreas systems: engineering the algorithms," *Diabetes care*, vol. 37, no. 5, pp. 1191–1197, 2014.

[8] R. Hovorka, "Artificial pancreas project at Cambridge 2013," *Diabetic Medicine*, vol. 32, no. 8, pp. 987–992, 2015.

[9] G. P. Forlenza, S. Deshpande, T. T. Ly, D. P. Howsmon, F. Cameron, N. Baysal, E. Mauritzen, T. Marcal, L. Towers, B. W. Bequette, *et al.*, "Application of zone model predictive control artificial pancreas during extended use of infusion set and sensor: a randomized crossover-controlled home-use trial," *Diabetes Care*, vol. 40, no. 8, pp. 1096–1102, 2017.

[10] L. H. Messer, G. P. Forlenza, J. L. Sherr, R. P. Wadwa, B. A. Buckingham, S. A. Weinzimer, D. M. Maahs, and R. H. Slover, "Optimizing hybrid closed-loop therapy in adolescents and emerging adults using the MiniMed 670G system," *Diabetes Care*, vol. 41, no. 4, pp. 789–796, 2018.

[11] P. Herrero, M. El-Sharkawy, J. Daniels, N. Jugnee, C. N. Uduku, M. Reddy, N. Oliver, and P. Georgiou, "The bio-inspired artificial pancreas for type 1 diabetes control in the home: system architecture and preliminary results," *Journal of diabetes science and technology*, vol. 13, no. 6, pp. 1017–1025, 2019.

[12] E. Atlas, R. Nimri, S. Miller, E. A. Grunberg, and M. Phillip, "Md-logic artificial pancreas system: a pilot study in adults with type 1 diabetes," *Diabetes care*, vol. 33, no. 5, pp. 1072–1076, 2010.

[13] G. P. Forlenza, L. Ekhlaspour, M. Breton, D. M. Maahs, R. P. Wadwa, M. DeBoer, L. H. Messer, M. Town, J. Pinnata, G. Kruse, *et al.*, "Successful at-home use of the Tandem Control-IQ artificial pancreas system in young children during a randomized controlled trial," *Diabetes technology & therapeutics*, vol. 21, no. 4, pp. 159–169, 2019.

[14] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, and Y. Wang, "Artificial intelligence in healthcare: past, present and future," *Stroke and vascular neurology*, vol. 2, pp. 230–243, 2017.

[15] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, p. 115, 2017.

[16] R. Gargeya and T. Leng, "Automated identification of diabetic retinopathy using deep learning," *Ophthalmology*, vol. 124, no. 7, pp. 962–969, 2017.

[17] I. Contreras and J. Vehi, "Artificial intelligence for diabetes management and decision support: literature review," *Journal of medical Internet research*, vol. 20, no. 5, p. e10775, 2018.

[18] C. Pérez-Gandía, A. Facchinetti, G. Sparacino, C. Cobelli, E. Gómez, M. Rigla, A. de Leiva, and M. Hernando, "Artificial neural network algorithm for online glucose prediction from continuous glucose monitoring," *Diabetes technology & therapeutics*, vol. 12, pp. 81–88, 2010.

[19] T. Zhu, K. Li, P. Herrero, J. Chen, and P. Georgiou, "A deep learning algorithm for personalized blood glucose prediction." in *The 3rd International Workshop on Knowledge Discovery in Healthcare Data, IJCAI-ECAI 2018*, 2018, pp. 64–78.

[20] K. Li, C. Liu, T. Zhu, P. Herrero, and P. Georgiou, "Glunet: A deep learning framework for accurate glucose forecasting," *IEEE Journal of Biomedical and Health Informatics*, 2019.

[21] J. Chen, K. Li, P. Herrero, T. Zhu, and P. Georgiou, "Dilated recurrent neural network for short-time prediction of glucose concentration." in *The 3rd International Workshop on Knowledge Discovery in Healthcare Data, IJCAI-ECAI 2018*, 2018, pp. 69–73.

[22] K. Li, J. Daniels, C. Liu, P. Herrero-Vinas, and P. Georgiou, "Convolutional recurrent neural networks for glucose prediction," *IEEE Journal of Biomedical and Health Informatics*, 2019.

[23] T. Zhu, K. Li, J. Chen, P. Herrero, and P. Georgiou, "Dilated recurrent neural networks for glucose forecasting in type 1 diabetes," *Journal of Healthcare Informatics Research*, pp. 1–17, 2020.

[24] A. Bertachi, L. Biagi, I. Contreras, N. Luo, and J. Vehi, "Prediction of blood glucose levels and nocturnal hypoglycemia using physiological models and artificial neural networks," in *The 3rd International Workshop on Knowledge Discovery in Healthcare Data, IJCAI-ECAI 2018*, Stockholm, Sweden, July 2018.

[25] S. Chang, Y. Zhang, W. Han, M. Yu, X. Guo, W. Tan, X. Cui, M. Witbrock, M. A. Hasegawa-Johnson, and T. S. Huang, "Dilated recurrent neural networks," in *Advances in Neural Information Processing Systems*, 2017, pp. 77–87.

[26] M. Tejedor, A. Z. Woldaregay, and F. Godtliebsen, "Reinforcement learning application in diabetes blood glucose control: A systematic review," *Artificial Intelligence in Medicine*, p. 101836, 2020.

[27] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529, 2015.

[28] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, *et al.*, "Mastering the game of go without human knowledge," *Nature*, vol. 550, no. 7676, p. 354, 2017.

[29] I. Fox and J. Wiens, "Reinforcement learning for blood glucose control: Challenges and opportunities," in *Reinforcement Learning for Real Life (RL4RealLife) Workshop in the 36th International Conference on Machine Learning (ICML)*, 2019.

[30] W. J. Artman, I. Nahum-Shani, T. Wu, J. R. Mckay, and A. Ertefaie, "Power analysis in a SMART design: sample size estimation for determining the best embedded dynamic treatment regime," *Biostatistics*, 2018.

[31] C. Dalla Man, F. Micheletto, D. Lv, M. Breton, B. Kovatchev, and C. Cobelli, "The UVA/PADOVA type 1 diabetes simulator," *Journal of diabetes science and technology*, vol. 8, no. 1, pp. 26–34, Jan. 2014.

[32] S. Schmidt and K. Nørgaard, "Bolus calculators," *Journal of diabetes science and technology*, vol. 8, no. 5, pp. 1035–1041, 2014.

[33] P. Herrero, J. Bondia, N. Oliver, and P. Georgiou, "A coordinated control strategy for insulin and glucagon delivery in type 1 diabetes," *Computer methods in biomechanics and biomedical engineering*, vol. 20, no. 13, pp. 1474–1482, 2017.

[34] J. R. Castle, J. E. Youssef, D. Branigan, B. Newswanger, P. Strange, M. Cummins, L. Shi, and S. Prestrelski, "Comparative pharmacokinetic/pharmacodynamic study of liquid stable glucagon versus lyophilized glucagon in type 1 diabetes subjects," *Journal of diabetes science and technology*, vol. 10, no. 5, pp. 1101–1107, 2016.

[35] T. Peyser, E. Dassau, M. Breton, and J. S. Skyler, "The artificial pancreas: current status and future prospects in the management of diabetes," *Annals of the New York Academy of Sciences*, vol. 1311, no. 1, pp. 102–123, 2014.

[36] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," *International Conference on Learning Representations*, vol. abs/1511.05952, 2015.

[37] T. Hester, M. Vecerik, O. Pietquin, M. Lanctot, T. Schaul, B. Piot, D. Horgan, J. Quan, A. Sendonaris, I. Osband, *et al.*, "Deep Q-learning from demonstrations," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[38] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[40] Y. Gao, H. Xu, J. Lin, F. Yu, S. Levine, and T. Darrell, "Reinforcement learning from imperfect demonstrations," in *35th International Conference on Machine Learning (ICML)*, 2018.

[41] P. Herrero, P. Pesl, J. Bondia, M. Reddy, N. Oliver, P. Georgiou, and C. Toumazou, "Method for automatic adjustment of an insulin bolus

calculator: in silico robustness evaluation under intra-day variability," *Computer methods and programs in biomedicine*, vol. 119, no. 1, pp. 1–8, 2015.

[42] A. Brazeau, H. Mircescu, K. Desjardins, C. Leroux, I. Strychar, J. Ekoé, and R. Rabasa-Lhoret, "Carbohydrate counting accuracy and blood glucose variability in adults with type 1 diabetes," *Diabetes research and clinical practice*, vol. 99, no. 1, pp. 19–23, 2013.

[43] G. P. Forlenza, Z. Li, B. A. Buckingham, J. E. Pinsker, E. Cengiz, R. P. Wadwa, L. Ekhlaspour, M. M. Church, S. A. Weinzimer, E. Jost, *et al.*, "Predictive low-glucose suspend reduces hypoglycemia in adults, adolescents, and children with type 1 diabetes in an at-home randomized crossover study: results of the prolog trial," *Diabetes Care*, vol. 41, no. 10, pp. 2155–2161, 2018.

[44] C. Liu, P. Avari, Y. Leal, M. Wos, K. Sivasithamparam, P. Georgiou, M. Reddy, J. M. Fernández-Real, C. Martin, M. Fernández-Balsells, *et al.*, "A modular safety system for an insulin dose recommender: a feasibility study," *Journal of diabetes science and technology*, vol. 14, no. 1, pp. 87–96, 2020.

[45] T. Battelino, R. Nimri, K. Dovc, M. Phillip, and N. Bratina, "Prevention of hypoglycemia with predictive low glucose insulin suspension in children with type 1 diabetes: A randomized controlled trial," *Diabetes Care*, vol. 40, no. 6, p. 764, 2017.

[46] D. M. Maahs, B. A. Buckingham, J. R. Castle, A. Cinar, E. R. Damiano, E. Dassau, J. H. DeVries, F. J. Doyle, S. C. Griffen, A. Haidar, *et al.*, "Outcome measures for artificial pancreas clinical trials: a consensus report," *Diabetes Care*, vol. 39, no. 7, pp. 1175–1179, 2016.

[47] L. Magni, D. M. Raimondo, C. Dalla Man, M. Breton, S. Patek, G. De Nicolao, C. Cobelli, and B. P. Kovatchev, "Evaluating the efficacy of closed-loop glucose regulation via control-variability grid analysis," *Journal of diabetes science and technology*, vol. 2, no. 4, pp. 630–635, 2008.

[48] Q. Sun, M. Jankovic, J. Budzinski, B. Moore, P. Diem, C. Stettler, and S. G. Mougiakakou, "A dual mode adaptive basal-bolus advisor based on reinforcement learning," *IEEE Journal of Biomedical and Health Informatics*, 2018.

[49] P. D. Ngo, S. Wei, A. Holubová, J. Muzik, and F. Godtliebsen, "Control of blood glucose for type-1 diabetes by using reinforcement learning with feedforward algorithm," *Computational and mathematical methods in medicine*, vol. 2018, 2018.

[50] T. Zhu, K. Li, and P. Georgiou, "Personalized dual-hormone control for type 1 diabetes using deep reinforcement learning," in *International Workshop on Health Intelligence (W3PHIAI-20) in the 34th AAAI Conference on Artificial Intelligence*, 2020 in press.

[51] M. Schiavon, C. Dalla Man, Y. C. Kudva, A. Basu, and C. Cobelli, "In silico optimization of basal insulin infusion rate during exercise: implication for artificial pancreas," *Journal of diabetes science and technology*, vol. 7, no. 6, pp. 1461–1469, 2013.

[52] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi, "Dream to control: Learning behaviors by latent imagination," in *8th International Conference on Learning Representations, ICLR*, 2020.

[53] S. Deshpande, J. E. Pinsker, S. Zavitsanou, D. Shi, R. Tompot, M. M. Church, C. Andre, F. J. Doyle III, and E. Dassau, "Design and clinical evaluation of the interoperable artificial pancreas system (iAPS) smartphone app: interoperable components with modular design for progressive artificial pancreas research and development," *Diabetes technology & therapeutics*, vol. 21, no. 1, pp. 35–43, 2019.

[54] D. Lewis, S. Leibrand, and O. Community, "Real-world use of open source artificial pancreas systems," *Journal of diabetes science and technology*, vol. 10, no. 6, p. 1411, 2016.

[55] F. H. El-Khatib, C. Balliro, M. A. Hillard, K. L. Magyar, L. Ekhlaspour, M. Sinha, D. Mondesir, A. Esmaeili, C. Hartigan, M. J. Thompson, *et al.*, "Home use of a bihormonal bionic pancreas versus insulin pump therapy in adults with type 1 diabetes: A multicentre randomised crossover trial," *The Lancet*, vol. 389, no. 10067, pp. 369–380, 2017.

[56] J. R. Castle, J. El Youssef, L. M. Wilson, R. Reddy, N. Resalat, D. Branigan, K. Ramsey, J. Leitschuh, U. Rajhbeharrysingh, B. Senf, *et al.*, "Randomized outpatient trial of single-and dual-hormone closed-loop systems that adapt to exercise using wearable sensors," *Diabetes care*, vol. 41, no. 7, pp. 1471–1477, 2018.

[57] M. Hessel, J. Modayil, H. Van Hasselt, T. Schaul, G. Ostrovski, W. Dabney, D. Horgan, B. Piot, M. Azar, and D. Silver, "Rainbow: Combining improvements in deep reinforcement learning," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.