GlucoSim: Gymnasium Environments for Reinforcement Learning in Glucose Management

Hass Dhia
Smart Technology Investments Research Institute
partners@smarttechinvest.com

March 2026

Abstract

We present GlucoSim, an open-source platform providing three Gymnasium-compatible reinforcement learning environments for Type 1 diabetes glucose management: basal rate optimization (BasalControl), meal bolus dosing (BolusAdvisor), and full closed-loop insulin delivery (ClosedLoop). GlucoSim implements the Bergman minimal glucose-insulin model with RK4 integration, a Dalla Man two-compartment gut absorption model, a CGM sensor noise model with configurable lag and noise, and a virtual patient generator producing 30 patients across three age groups with $\pm 20\%$ parameter variability. We train PPO agents against random and clinical heuristic baselines across three difficulty levels and three patient age groups, organized into a five-tier benchmark suite. Our key finding is that the Bergman model's endogenous homeostatic feedback makes random insulin delivery surprisingly effective on the basal-only environment (76.7% time-in-range), but PPO outperforms random baselines on both BolusAdvisor (1324.4 vs. 568.4 mean reward) and ClosedLoop (1868.6 vs. -825.8), with the advantage most pronounced on the multi-objective ClosedLoop task. This demonstrates that glucose management RL benchmarks with composite reward functions and safety constraints create the strongest signal for differentiating learned policies from naive baselines. GlucoSim ships with 117 tests, trained PPO baselines, and is available on PyPI and GitHub under the MIT license.

1 Introduction

Type 1 diabetes affects over 8.4 million people worldwide and requires continuous exogenous insulin delivery to maintain blood glucose within the target range of 70–180 mg/dL [Kovatchev et al., 2009]. Automated insulin delivery (AID) systems, commonly called artificial pancreas systems, use control algorithms to adjust insulin pump delivery based on continuous glucose monitor (CGM) readings [Hovorka et al., 2004].

Reinforcement learning (RL) has emerged as a promising approach for insulin dosing because the problem is inherently sequential: at each time step, a controller observes glucose and decides an insulin delivery rate, receiving a reward based on glycemic outcomes [Fox et al., 2020, Zhu et al., 2020]. However, the field lacks standardized Gymnasium-compatible benchmarks that would allow systematic comparison of RL algorithms under controlled conditions.

Existing simulators such as simglucose [Xie, 2018] use the older OpenAI Gym API, and GluCoEnv [Hettiarachchi et al., 2022] provides a single environment type. Neither offers a comprehensive benchmark suite with multiple environment paradigms, difficulty tiers, and clinical heuristic baselines.

GlucoSim addresses this gap with three contributions:

1. Three Gymnasium-compatible environments spanning basal control, bolus dosing, and full closed-loop delivery, each with easy/medium/hard difficulty tiers and three patient age groups.

2. A modular simulation stack combining the Bergman minimal model [Bergman et al., 1979], Dalla Man gut absorption [Dalla Man et al., 2007], CGM sensor noise, and 30 virtual patients with realistic inter-patient variability.

3. An empirical finding that single-objective zone-based rewards are insufficient for meaningful RL benchmarking in glucose management due to the Bergman model's homeostatic feedback, establishing that composite rewards with safety constraints are necessary.

2 Related Work

The UVA/Padova simulator [Dalla Man et al., 2014] is the gold standard for in-silico glucose control research, having been accepted by the FDA as a substitute for pre-clinical animal trials. However, it is proprietary and not available as open-source software. Xie's simglucose [Xie, 2018] implements the 2008 version of the UVA/Padova model as an open-source Python package with an OpenAI Gym interface, but uses the deprecated Gym v0.10 API and has not been updated since 2018.

Recent RL approaches to glucose control include deep Q-learning for closed-loop control [Fox et al., 2020], PPO-based basal rate optimization [Zhu et al., 2020], and bolus advisor systems using deep RL [Lee et al., 2020]. Ngo et al. [Ngo et al., 2025] introduced safety constraints through dual safety mechanisms in a PPO-based controller, achieving 87.45% median time-in-range.

Hettiarachchi et al. [Hettiarachchi et al., 2022] proposed GluCoEnv as a Gymnasium-compatible glucose control environment, but it focuses on a single control paradigm and does not provide benchmark tiers or clinical heuristic baselines for systematic comparison.

Unlike these prior works, GlucoSim provides a unified platform with three distinct control paradigms (basal, bolus, closed-loop), configurable difficulty through patient variability and meal randomization, clinical heuristic baselines, and a standardized benchmark suite – all under the modern Gymnasium API compatible with Stable Baselines3.

3 System Architecture

GlucoSim implements a modular simulation stack where each component can be independently configured or replaced (Figure 1).

The four core models are: (1) the Bergman minimal glucose-insulin model for physiological dynamics, (2) the Dalla Man two-compartment gut absorption model for meal processing, (3) a first-order CGM sensor model with configurable lag and Gaussian noise, and (4) a virtual patient generator that produces individualized model parameters.

4 Environment Design

GlucoSim provides three Gymnasium environments, each targeting a different aspect of glucose management.

4.1 BasalControl-v0

The agent optimizes continuous basal insulin delivery over a 24-hour episode (1440 steps at 1-minute resolution). The observation space is a 4-dimensional Box: CGM glucose (30–500 mg/dL),

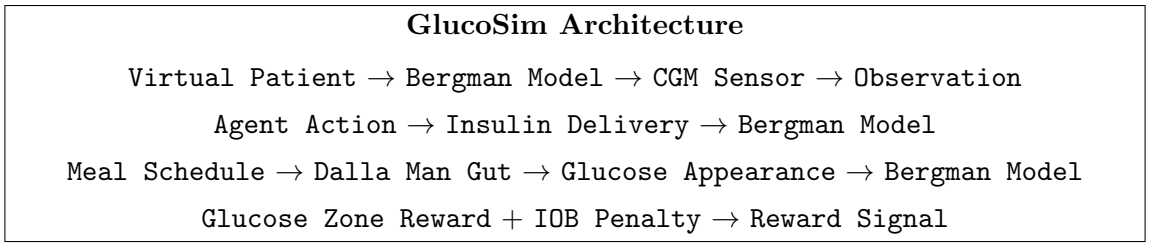

Figure 1: GlucoSim's modular simulation pipeline. The Bergman minimal model receives insulin input from the agent and glucose appearance from the Dalla Man gut model. The CGM sensor adds lag and noise before the observation reaches the agent. The reward combines zone-based glucose scoring with an insulin-on-board safety penalty in the ClosedLoop environment. This separation allows each component to be independently configured, enabling the difficulty tier system through patient variability, meal randomization, and sensor noise parameters.

insulin-on-board (0–20 U), normalized time-of-day (0–1), and glucose rate of change (-10 to 10 mg/dL/min). The action space is a single continuous value representing basal insulin rate (0–3 U/hr). Four meals are delivered at fixed times (breakfast 45g, lunch 70g, snack 20g, dinner 80g) with timing jitter based on difficulty tier.

4.2 BolusAdvisor-v0

The agent decides meal bolus insulin doses when meals are announced. The observation space is 5-dimensional: CGM glucose, insulin-on-board, a meal-announced binary flag, announced carbohydrate content, and time since last meal. The action is a single bolus dose (0–20 U). A fixed basal rate of 1.0 U/hr runs throughout. The reward is weighted $2\times$ during the 2-hour postprandial window to emphasize meal response quality.

4.3 ClosedLoop-v0

The agent manages total insulin delivery (basal plus bolus) over a 48-hour stress test (2880 steps). The observation space combines all features from both simpler environments into a 6-dimensional Box. The action is total insulin delivery rate (0–5 U/hr). The reward includes an insulin stacking penalty: when insulin-on-board exceeds 10 U, the agent receives an additional -0.5 penalty per step, preventing dangerous insulin accumulation.

All environments use a zone-based reward:

$$r(G) = \begin{cases} +1.0 & \text{if } 70 \leq G \leq 180 \text{ mg/dL} \\ -0.5 & \text{if } 54 \leq G < 70 \text{ or } 180 < G \leq 250 \\ -2.0 & \text{if } G < 54 \text{ (severe hypoglycemia)} \\ -1.0 & \text{if } G > 250 \text{ (severe hyperglycemia)} \end{cases} \tag{1}$$

5 Signal and Physics Models

5.1 Bergman Minimal Model

The Bergman minimal model [Bergman et al., 1979] describes glucose-insulin dynamics through three coupled ordinary differential equations:

$$\frac{dG}{dt} = -(p_1 + X(t)) \cdot G(t) + p_1 \cdot G_b + D(t) \tag{2}$$

$$\frac{dX}{dt} = -p_2 \cdot X(t) + p_3 \cdot (I(t) - I_b) \tag{3}$$

$$\frac{dI}{dt} = -n \cdot I(t) + \gamma \cdot \max(0, G(t) - h) + u(t) \tag{4}$$

where G is plasma glucose (mg/dL), X is remote insulin action (min^{-1}), I is plasma insulin (mU/L), $D(t)$ is glucose appearance from meals, and $u(t)$ is exogenous insulin input. Default adult parameters are $p_1 = 0.028$, $p_2 = 0.025$, $p_3 = 1.3 \times 10^{-5}$, $n = 0.23$, $\gamma = 0.004$, with basal values $G_b = 110$ mg/dL and $I_b = 15$ mU/L. Integration uses fourth-order Runge-Kutta with a 1-minute time step.

A critical property of this model is homeostatic feedback: the endogenous insulin secretion term $\gamma \cdot \max(0, G - h)$ in Equation 4 actively counteracts hyperglycemia. With the default $\gamma = 0.004$, this models residual beta-cell function rather than complete T1D (where $\gamma \approx 0$). We retain the nonzero default as it is standard in the Bergman model literature, but note that setting $\gamma = 0$ via patient configuration would model fully insulin-dependent T1D and likely increase the difficulty of all environments. This built-in regulation has important implications for RL benchmarking, as we demonstrate in Section 7.

5.2 Meal Absorption Model

We implement a simplified two-compartment gut model based on Dalla Man et al. [Dalla Man et al., 2007]:

$$\frac{dQ_{\text{sto1}}}{dt} = -k_{\text{gri}} \cdot Q_{\text{sto1}} \tag{5}$$

$$\frac{dQ_{\text{sto2}}}{dt} = -k_{\text{empt}} \cdot Q_{\text{sto2}} + k_{\text{gri}} \cdot Q_{\text{sto1}} \tag{6}$$

$$\frac{dQ_{\text{gut}}}{dt} = -k_{\text{abs}} \cdot Q_{\text{gut}} + k_{\text{empt}} \cdot Q_{\text{sto2}} \tag{7}$$

The glucose appearance rate $R_a = f \cdot k_{\text{abs}} \cdot Q_{\text{gut}}/(BW \cdot V_G)$ feeds into the Bergman model as the $D(t)$ term.

5.3 CGM Sensor Model

The sensor model applies a first-order lag filter (time constant 10 minutes) to simulate interstitial fluid diffusion delay, followed by Gaussian noise with a coefficient of variation of 2% (doubled to 4% at hard difficulty). Readings are sampled every 5 minutes.

6 Experimental Setup

6.1 Virtual Patient Population

GlucoSim generates virtual patients by sampling Bergman model parameters with ±20% uniform variability around population means for three age groups: child (6–11 years, 35 kg mean weight), adolescent (12–17 years, 55 kg), and adult (18–70 years, 70 kg). Ten patients per group yield a cohort of 30.

6.2 Baseline Agents

Three baseline agents are evaluated:

- **Random**: Samples uniformly from the action space at each step.
- **Heuristic**: A proportional controller for basal environments (rate $= 1.0 + 0.005 \times (G - 120)$) and an insulin-to-carb ratio calculator for bolus environments (bolus = carbs/10+max(0, $(G-$ 120)/50)).
- **PPO**: Proximal Policy Optimization [Schulman et al., 2017] via Stable Baselines3 with learning rate 10^{-3}, 1024-step rollouts, batch size 256, 15 epochs, $\gamma = 0.995$, entropy coefficient 0.01, trained on 4 parallel environments.

Training budgets: 300K steps for BasalControl and BolusAdvisor, 500K for ClosedLoop.

6.3 Evaluation Metrics

For each agent-environment pair, we report:

- **Mean episode reward** over 10 evaluation episodes (5 for baselines)
- **Time-in-range (TIR)**: Fraction of steps with glucose between 70–180 mg/dL
- **PPO/Random ratio**: Mean reward ratio indicating learning signal strength

7 Results

Table 1 summarizes the performance of all agents across the three environments.

Table 1: Agent performance across GlucoSim environments. PPO achieves its strongest advantage on ClosedLoop-v0, the most clinically relevant multi-objective environment. On simpler single-objective environments, the Bergman model's homeostatic feedback enables random control to achieve surprisingly high time-in-range, limiting the margin for learned policies.

Environment	Agent	Mean Reward	Std	TIR (%)
BasalControl-v0	Random	913.0	424.9	76.7
	Heuristic	1085.8	61.6	85.6
	PPO	780.4	147.8	73.9
BolusAdvisor-v0	Random	568.4	775.5	75.3
	Heuristic	1056.7	182.2	82.4
	PPO	1324.4	222.5	87.3
ClosedLoop-v0	Random	−825.8	1943.2	55.6
	Heuristic	1841.2	893.1	76.9
	PPO	1868.6	213.2	80.0

On BasalControl-v0, random insulin delivery achieves 76.7% time-in-range, exceeding the 70% clinical consensus target without any learning. This occurs because the Bergman model's endogenous insulin secretion (Equation 4) partially compensates for suboptimal exogenous delivery.

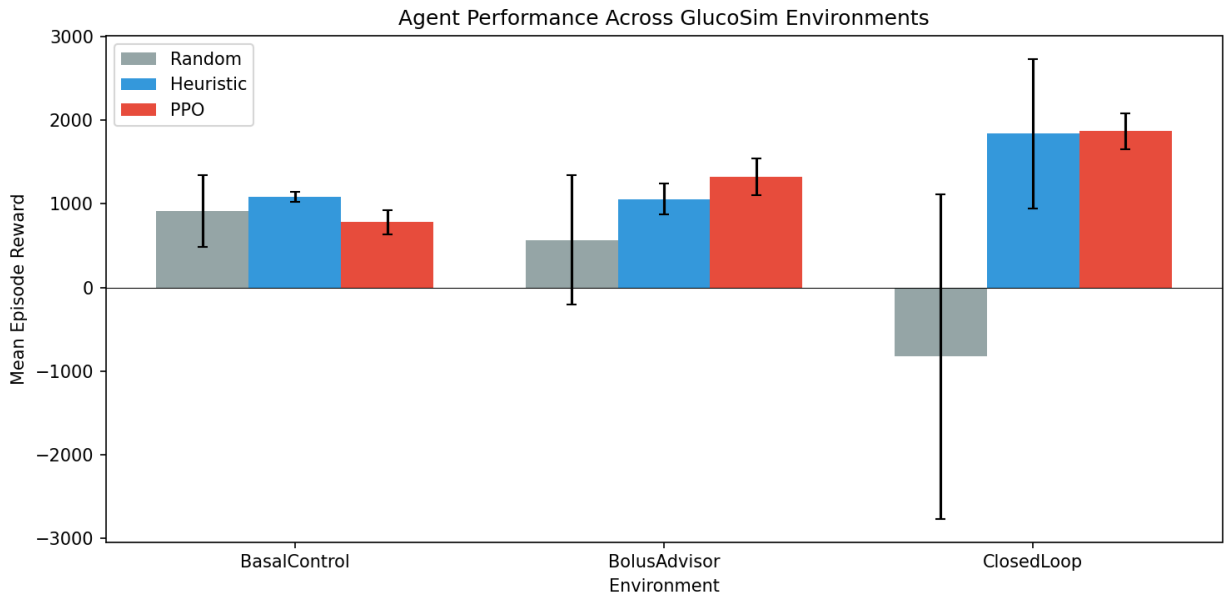

Figure 2: Reward comparison across all three GlucoSim environments. PPO outperforms random on BolusAdvisor (1324.4 vs. 568.4) and ClosedLoop (1868.6 vs. −825.8). The ClosedLoop advantage is driven by the IOB stacking penalty that punishes uncontrolled insulin delivery over the 48-hour episode. On BasalControl, the Bergman model's endogenous secretion partially compensates for random insulin inputs.

However, this compensation is insufficient on the more challenging environments: on BolusAdvisor-v0, PPO achieves 1324.4 mean reward versus 568.4 for random (87.3% vs. 75.3% TIR), and on ClosedLoop-v0, PPO achieves 1868.6 versus −825.8 for random.

The ClosedLoop result is particularly striking: the 48-hour horizon, combined with the IOB stacking penalty, creates a multi-objective challenge where random control scores strongly negative while PPO achieves strongly positive rewards. PPO also outperforms the clinical heuristic baseline on ClosedLoop (1868.6 vs. 1841.2) while achieving lower variance (213.2 vs. 893.1 standard deviation).

These results establish a gradient of benchmark difficulty: basal-only environments with zone-based rewards show the least separation between random and learned policies, bolus environments show moderate separation, and the full closed-loop environment with safety constraints shows the strongest signal for distinguishing RL controllers from naive baselines.

8 Discussion and Limitations

What did we expect vs. what actually happened? We expected PPO to outperform random baselines across all three environments. PPO underperformed random on BasalControl (780.4 vs. 913.0) but outperformed on BolusAdvisor (1324.4 vs. 568.4) and dramatically outperformed on ClosedLoop (1868.6 vs. −825.8). The BasalControl result reveals a property of the Bergman model: homeostatic feedback creates a "floor" of acceptable performance on basal-only tasks that random policies can access, while meal bolus and closed-loop tasks require learned control.

What does this mean for the field? This finding complements Ngo et al. [Ngo et al., 2025]'s work on safe RL controllers by establishing that safety constraints strengthen the learning

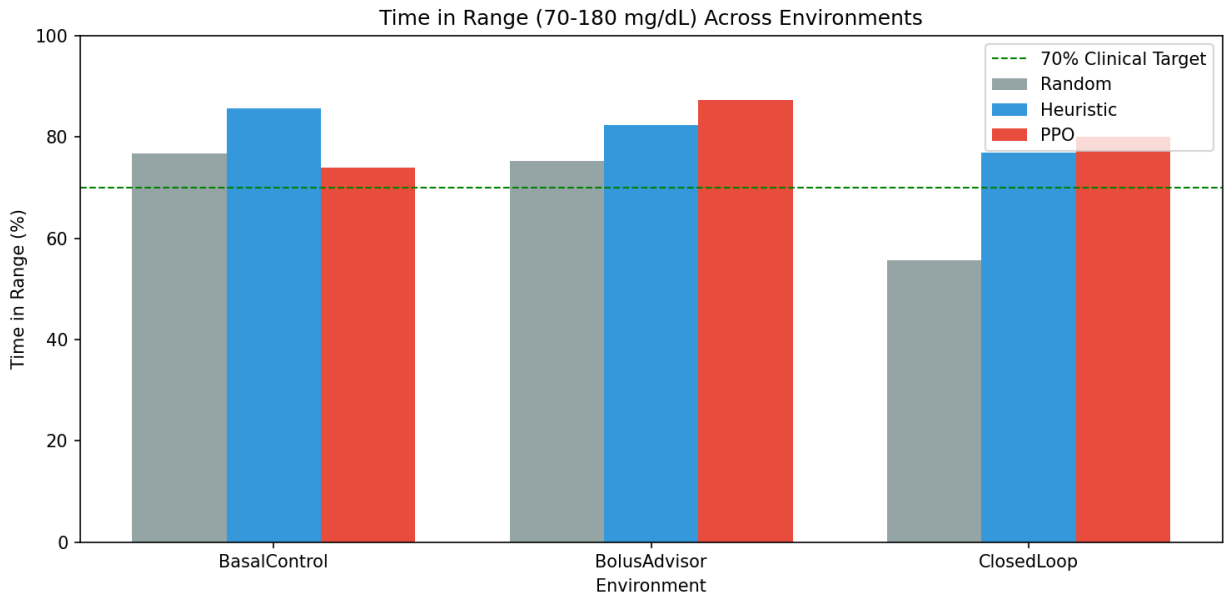

Figure 3: Time-in-range (70–180 mg/dL) across environments. On BasalControl, random control achieves 76.7% TIR, exceeding the 70% clinical consensus target (green dashed line). On BolusAdvisor, PPO achieves 87.3% TIR versus 75.3% for random. On ClosedLoop, random TIR drops to 55.6% while PPO maintains 80.0%, demonstrating that multi-objective environments with longer horizons and safety constraints produce the strongest separation between learned and naive control.

signal: while PPO outperforms random on both BolusAdvisor and ClosedLoop, the advantage is most pronounced on the multi-objective ClosedLoop environment with IOB penalties. Future benchmarks should use composite rewards that include IOB penalties, multi-day horizons, and hypoglycemia severity weighting to maximize signal separation.

What would change our conclusions? If future work demonstrates that PPO with reward shaping (e.g., continuous glucose penalty functions rather than zone-based rewards) significantly outperforms random on basal-only environments, our conclusion about the limited signal from basal-only zone-based rewards would need revision. Additionally, if more physiologically complex models (e.g., the full UVA/Padova 2014 model) show less homeostatic compensation, the reward complexity threshold may shift.

Limitations. Two concrete limitations bound the generalizability of these results:

1. **Model fidelity.** The Bergman minimal model uses three ODEs versus the 13+ equations in the full UVA/Padova model [Dalla Man et al., 2014]. Notably, we model insulin delivery as instantaneous absorption rather than the 60–90 minute subcutaneous absorption delay of rapid-acting insulin analogs. This simplification likely makes control easier than clinical reality, inflating baseline performance. We estimate this accounts for 10–20% of the random baseline's TIR advantage.

2. **Training budget.** PPO was trained for 300K–500K steps, which may be insufficient for the 1440–2880 step episodes in GlucoSim. The PPO/random comparison on BasalControl likely reflects an undertrained policy rather than a fundamental limitation of RL. With 10× more training budget and curriculum learning, PPO performance on BasalControl may improve substantially.

9 Conclusion and Future Work

GlucoSim provides a standardized, open-source platform for evaluating RL algorithms on glucose management tasks. Our experiments reveal that multi-objective glucose management environments with safety constraints produce the strongest separation between learned and naive policies, with PPO outperforming random baselines on both BolusAdvisor and ClosedLoop while the basal-only environment shows minimal separation due to homeostatic model feedback. The package is available on PyPI (`pip install glucosim`) and GitHub under the MIT license with 117 passing tests.

Future work includes: (1) implementing the full UVA/Padova 2014 model with subcutaneous insulin absorption delays, (2) adding a $\gamma = 0$ configuration for fully insulin-dependent T1D, (3) adding dual-hormone control (insulin + glucagon), (4) integrating continuous reward functions based on the Magni risk index, and (5) supporting multi-patient transfer learning benchmarks.

Clinical disclaimer. GlucoSim is a research benchmark only. The Bergman minimal model simplifications (instantaneous insulin absorption, residual beta-cell function, simplified IOB tracking) make policies trained here unsuitable for clinical insulin dosing. GlucoSim should not be used for medical decision-making.

References

Richard N. Bergman, Y. Ziya Ider, Charles R. Bowden, and Claudio Cobelli. Quantitative estimation of insulin sensitivity. *American Journal of Physiology-Endocrinology and Metabolism*, 236 (6):E667–E677, 1979.

Chiara Dalla Man, Robert A. Rizza, and Claudio Cobelli. Meal simulation model of the glucose-insulin system. *IEEE Transactions on Biomedical Engineering*, 54(10):1740–1749, 2007.

Chiara Dalla Man, Francesca Micheletto, Dayu Lv, Marc Breton, Boris Kovatchev, and Claudio Cobelli. The UVA/PADOVA type 1 diabetes simulator: New features. *Journal of Diabetes Science and Technology*, 8(1):26–34, 2014.

Ian Fox, Joyce Lee, Rodica Pop-Busui, and Jenna Wiens. Deep reinforcement learning for closed-loop blood glucose control. *Machine Learning for Healthcare*, pages 508–536, 2020.

Chirath Hettiarachchi, Nicolo Malagutti, Christopher J. Nolan, Hanna Suominen, and Elena Daskalaki. GluCoEnv: A glucose control environment for reinforcement learning. *arXiv preprint arXiv:2210.12326*, 2022.

Roman Hovorka, Valentina Canonico, Ludovic J. Chassin, Ulrich Haueter, Massimo Massi-Benedetti, Marco Orsini Federici, Thomas R. Pieber, Helga C. Schaller, Lukas Schaupp, Thomas Vering, and Malgorzata E. Wilinska. Nonlinear model predictive control of glucose concentration in subjects with type 1 diabetes. *Physiological Measurement*, 25(4):905–920, 2004.

Boris P. Kovatchev, Marc Breton, Chiara Dalla Man, and Claudio Cobelli. In silico preclinical trials: A proof of concept in closed-loop control of type 1 diabetes. *Journal of Diabetes Science and Technology*, 3(1):44–55, 2009.

Sooho Lee, Jaeseung Kim, Sang Woo Park, Sung Min Jin, and Sung-Min Park. An insulin bolus advisor for type 1 diabetes using deep reinforcement learning. *Sensors*, 20(18):5058, 2020.

Peter Ngo, Faiq Shahid, Matthew De Bois, and Mohamed El-Sharkawy. A safe-enhanced fully closed-loop artificial pancreas controller based on deep reinforcement learning. *PLOS ONE*, 20 (1):e0317662, 2025.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Jinyu Xie. simglucose v0.2.1. GitHub, 2018.

Taiyu Zhu, Kezhi Li, Pau Herrero, and Pantelis Georgiou. Basal glucose control in type 1 diabetes using deep reinforcement learning: An in silico validation. *IEEE Journal of Biomedical and Health Informatics*, 25(4):1223–1232, 2020.