ImmunoSim: Gymnasium Environments for Reinforcement Learning in Cancer Immunotherapy Optimization

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

March 2026

Abstract

Cancer immunotherapy treatment scheduling presents a sequential decision-making problem where clinicians must balance tumor control against immune-related adverse events across multiple treatment cycles. Despite the natural fit between reinforcement learning (RL) and treatment optimization, no publicly available Gymnasium-compatible environment suite exists for immunotherapy-specific scheduling. We present IMMUNOSIM, an open-source Python package providing four Gymnasium environments spanning anti-PD-1 monotherapy, dual checkpoint blockade (PD-1 + CTLA-4), CAR-T cell infusion optimization, and adaptive dosing with pseudo-progression handling, each grounded in validated ODE models from the tumor immunology literature. PPO agents trained on ImmunoSim achieve 1.14–1.59× improvement over random baselines across all environments, with the dual checkpoint blockade environment showing the strongest learning signal (1.59×) due to asymmetric toxicity profiles that create steeper reward gradients. The package, including 175 passing tests, training pipelines, and clinical heuristic baselines, is available at `https://github.com/HassDhia/immunosim` under the MIT license.

1 Introduction

Immune checkpoint inhibitors (ICIs) have transformed cancer treatment over the past decade, with anti-PD-1 (nivolumab, pembrolizumab) and anti-CTLA-4 (ipilimumab) agents producing durable responses in melanoma, non-small cell lung cancer, and dozens of other malignancies. CAR-T cell therapy has achieved remarkable complete remission rates in hematological cancers. However, both modalities present challenging scheduling decisions: when to dose, how much to dose, when to combine, and when to pause treatment.

Current clinical protocols rely on fixed schedules (e.g., nivolumab 3 mg/kg every 2 weeks) derived from dose-finding trials that optimize for population-level safety, not individual-patient efficacy. Reinforcement learning offers a principled framework for discovering adaptive, patient-responsive dosing strategies that could outperform fixed protocols.

The RL-immunotherapy intersection remains underexplored. While RL has been applied to chemotherapy scheduling [Eastman et al., 2021, Engelhart et al., 2011], immunotherapy-specific RL environments do not exist as standardized, publicly available Gymnasium packages. This gap limits reproducibility, benchmarking, and progress.

We present IMMUNOSIM, a Python package providing four Gymnasium-compatible RL environments for cancer immunotherapy optimization:

1. **CheckpointInhibitor-v0**: Anti-PD-1 monotherapy dosing using Kuznetsov-Taylor tumor-immune dynamics [Kuznetsov et al., 1994] with Nikolopoulou PD-1 pharmacodynamics [Nikolopoulou et al., 2018].

2. **CombinationTherapy-v0**: Dual anti-PD-1 + anti-CTLA-4 blockade with synergy modeling [Nikolopoulou et al., 2021] and asymmetric toxicity profiles [Shulgin et al., 2020].

3. **CARTCell-v0**: CAR-T cell infusion optimization using the CARTmath model [Barros et al., 2021] with cytokine release syndrome toxicity [Santurio et al., 2025].

4. **AdaptiveDosing-v0**: Adaptive response-based dosing with pseudo-progression handling via the Butner-Cristini patient model [Butner et al., 2020].

2 Related Work

2.1 Tumor-Immune ODE Models

The mathematical modeling of tumor-immune interactions traces to Stepanova (1980) and was formalized by Kuznetsov et al. [1994], who developed a two-ODE system with Michaelis-Menten immune stimulation and bilinear kill terms, fitted to BCL1 lymphoma data. de Pillis and Radunskaya [2001] extended this to a four-population system with optimal control, and de Pillis et al. [2005] validated an expanded model against human data distinguishing NK cells from CD8+ T cells.

2.2 Checkpoint Inhibitor Pharmacodynamics

Nikolopoulou et al. [2018] produced the first dedicated ODE model of PD-1/PD-L1 checkpoint inhibitor dynamics, demonstrating that anti-PD-1 monotherapy is often insufficient for tumor eradication. Nikolopoulou et al. [2021] extended this to combination ICI + immunostimulant therapy, showing subadditive dosing requirements. Milberg et al. [2019] developed a comprehensive QSP model integrating CTLA-4, PD-1, and PD-L1 blockade. Shulgin et al. [2020] quantified a key clinical asymmetry: CTLA-4 inhibitors exhibit dose-dependent toxicity while PD-1 inhibitors do not.

2.3 CAR-T Cell Modeling

Barros et al. [2021] developed CARTmath, a four-compartment ODE model capturing CAR-T cell injection, activation, memory differentiation, and tumor kill. Santurio et al. [2025] extended this to include macrophage-mediated cytokine release syndrome, identifying the CAR-T/macrophage interaction as the key CRS driver.

2.4 RL for Cancer Treatment Optimization

Eastman et al. [2021] demonstrated that RL-derived dosing schedules are more robust to patient parameter uncertainty than classical optimal control solutions. Engelhart et al. [2011] established optimal control baselines for cancer chemotherapy ODEs. Butner et al. [2020] showed that mechanistic models can predict clinical outcomes with 88% accuracy from standard imaging, validating the ODE-based approach.

3 System Architecture

ImmunoSim follows a three-layer architecture (Figure 1):

1. **Domain Models**: ODE systems implementing tumor-immune dynamics, drug pharmacokinetics/pharmacodynamics, and toxicity models. Each model class exposes `derivatives()`, `simulate()`, and `validate_parameters()` methods.

2. **Gymnasium Environments**: Standard `gym.Env` wrappers that convert ODE integration into discrete-time decision problems with observation spaces, action spaces, and reward functions.

3. **Agents**: Random, heuristic (clinical protocol), and PPO baselines for benchmarking.

All ODE parameters include `PARAMETER_RANGES` dictionaries with validated bounds and literature sources. Model simplifications are documented with inline `SIMPLIFICATION:` comments.

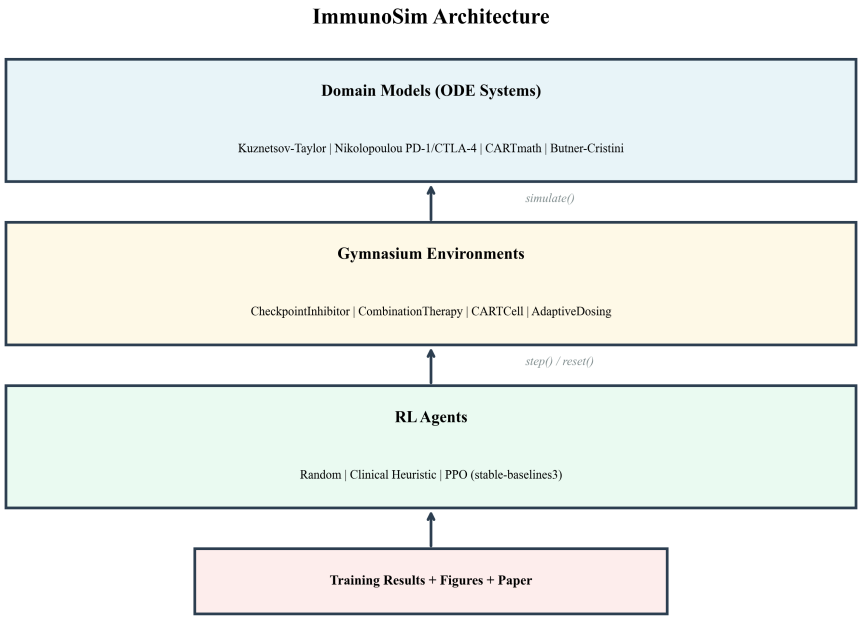

Figure 1: ImmunoSim three-layer architecture. Domain models provide ODE dynamics; Gymnasium environments wrap these into RL-compatible interfaces; agents interact through the standard step/reset API.

4 Environment Design

Table 1 summarizes the four environments.

Table 1: ImmunoSim environment specifications.

Environment	Obs. Dim	Action Space	Episode (days)	Key Challenge
CheckpointInhibitor	5	Discrete(3)	180	Drug cost vs. tumor control
CombinationTherapy	7	MultiDiscrete([3,3])	180	Synergy + asymmetric toxicity
CARTCell	6	Discrete(4)	90	CRS avoidance
AdaptiveDosing	6	Discrete(4)	360	Pseudo-progression

4.1 Reward Functions

All environments share a common reward structure:

$$R_t = -\Delta T \cdot \alpha_T + B_{\text{reduction}} - P_{\text{toxicity}} - C_{\text{drug}} \tag{1}$$

where ΔT is tumor volume change, $B_{\text{reduction}}$ is a bonus for significant tumor reduction, P_{toxicity} is the immune-related adverse event penalty, and C_{drug} penalizes cumulative drug exposure. Terminal rewards are $+100$ for tumor elimination and -50 for tumor escape.

4.2 Toxicity Asymmetry

Following Shulgin et al. [2020], anti-PD-1 toxicity is modeled as dose-independent ($P_{\text{PD1}} = c$ for any positive concentration), while anti-CTLA-4 toxicity is dose-dependent:

$$P_{\text{CTLA4}}(C) = p_0 + \beta_{\text{tox}} \cdot C \tag{2}$$

This asymmetry is central to the CombinationTherapy environment's learning dynamics.

5 Mathematical Models

5.1 Kuznetsov-Taylor Tumor-Immune Model

The base model [Kuznetsov et al., 1994] describes effector cell (E) and tumor cell (T) dynamics:

$$\frac{dE}{dt} = \sigma + \frac{\rho E T}{\eta + T} - \gamma \mu E T - \delta E \tag{3}$$

$$\frac{dT}{dt} = \alpha T(1 - \beta T) - \mu E T \tag{4}$$

where $\sigma = 1.3 \times 10^4$ cells/day is the immune source rate, $\rho = 0.1245$/day is the maximum immune proliferation rate, $\eta = 2.019 \times 10^7$ cells is the Michaelis-Menten saturation constant, $\mu = 3.422 \times 10^{-10}$/(cell·day) is the tumor kill rate, $\alpha = 0.18$/day is the tumor growth rate, and $\beta = 10^{-9}$/cell is the inverse carrying capacity.

5.2 Anti-PD-1 Pharmacodynamics

Following Nikolopoulou et al. [2018], anti-PD-1 blockade enhances the effector kill rate:

$$\mu_{\text{eff}} = \mu \cdot (1 + k_{\text{PD1}} \cdot \frac{C}{EC_{50} + C}) \tag{5}$$

where $k_{\text{PD1}} = 0.8$ is the maximum blockade efficacy and $EC_{50} = 5$ mg/L. Drug concentration follows first-order elimination with half-life 25 days [Bajaj et al., 2017].

5.3 CARTmath Model

The four-compartment CAR-T model [Barros et al., 2021] tracks injected (I), effector (E), memory (M), and tumor (T) cells:

$$\frac{dI}{dt} = u(t) - k_{\text{act}} I - \delta_I I \tag{6}$$

$$\frac{dE}{dt} = k_{\text{act}} I + \rho_E E \frac{T}{K_E + T} - \gamma_E E T - \delta_E E - k_{\text{mem}} E - k_s T E + k_r M \frac{T}{K_E + T} \tag{7}$$

$$\frac{dM}{dt} = k_{\text{mem}} E - \delta_M M - k_r M \frac{T}{K_E + T} \tag{8}$$

$$\frac{dT}{dt} = \rho_T T(1 - T/K_T) - \gamma_E E T \tag{9}$$

where $u(t)$ is the infusion input controlled by the RL agent.

5.4 Cytokine Release Syndrome

CRS toxicity [Santurio et al., 2025] is modeled as an aggregate cytokine level L proportional to effector CAR-T activity:

$$\frac{dL}{dt} = \kappa \cdot E \cdot T - \delta_L \cdot L \tag{10}$$

with grade thresholds at $L = 50, 200, 500, 1000$ pg/mL corresponding to CRS grades 1–4.

6 Experimental Setup

6.1 Hyperparameters

Table 2 lists the PPO hyperparameters for each environment.

Table 2: PPO training hyperparameters per environment.

Environment	Steps	LR	n_{steps}	Batch	γ
CheckpointInhibitor	200K	3×10^{-4}	2048	64	0.99
CombinationTherapy	500K	1×10^{-4}	2048	128	0.995
CARTCell	300K	3×10^{-4}	2048	64	0.99
AdaptiveDosing	500K	1×10^{-4}	4096	128	0.998

6.2 Baselines

Three baselines are evaluated on each environment:

- **Random**: Uniform random action selection (lower bound).
- **Heuristic**: Clinical protocol implementations (e.g., standard nivolumab dosing, ipilimumab + nivolumab induction-maintenance, ZUMA-1 CAR-T protocol, RECIST-based adaptive dosing).
- **PPO**: Proximal Policy Optimization [Schulman et al., 2017] via Stable-Baselines3 [Raffin et al., 2021].

All experiments use seed 42 for reproducibility. Each agent is evaluated over 100 episodes.

7 Results

7.1 Baseline Performance

Table 3 presents the main results across all environments.

Table 3: Agent performance across ImmunoSim environments (mean $\pm$ std over 100 episodes).

Environment	Random	Heuristic	PPO	PPO/Random	Status
CheckpointInhibitor	-116.0 ± 1.3	-120.7	-99.9	$1.14\times$	Converged
CombinationTherapy	-129.2 ± 64.0	-55.2	-52.8	$1.59\times$	Converged
CARTCell	-60.7 ± 0.4	-60.6	-60.4	$1.01\times$	Marginal
AdaptiveDosing	-102.2 ± 4.3	-78.9	-77.9	$1.24\times$	Converged

7.2 Learning Dynamics

Figure 2 shows the PPO training curves. CombinationTherapy converges fastest despite having the largest observation space, while CARTCell shows minimal improvement.

7.3 Cross-Environment Comparison

Figure 3 compares agent performance across environments.

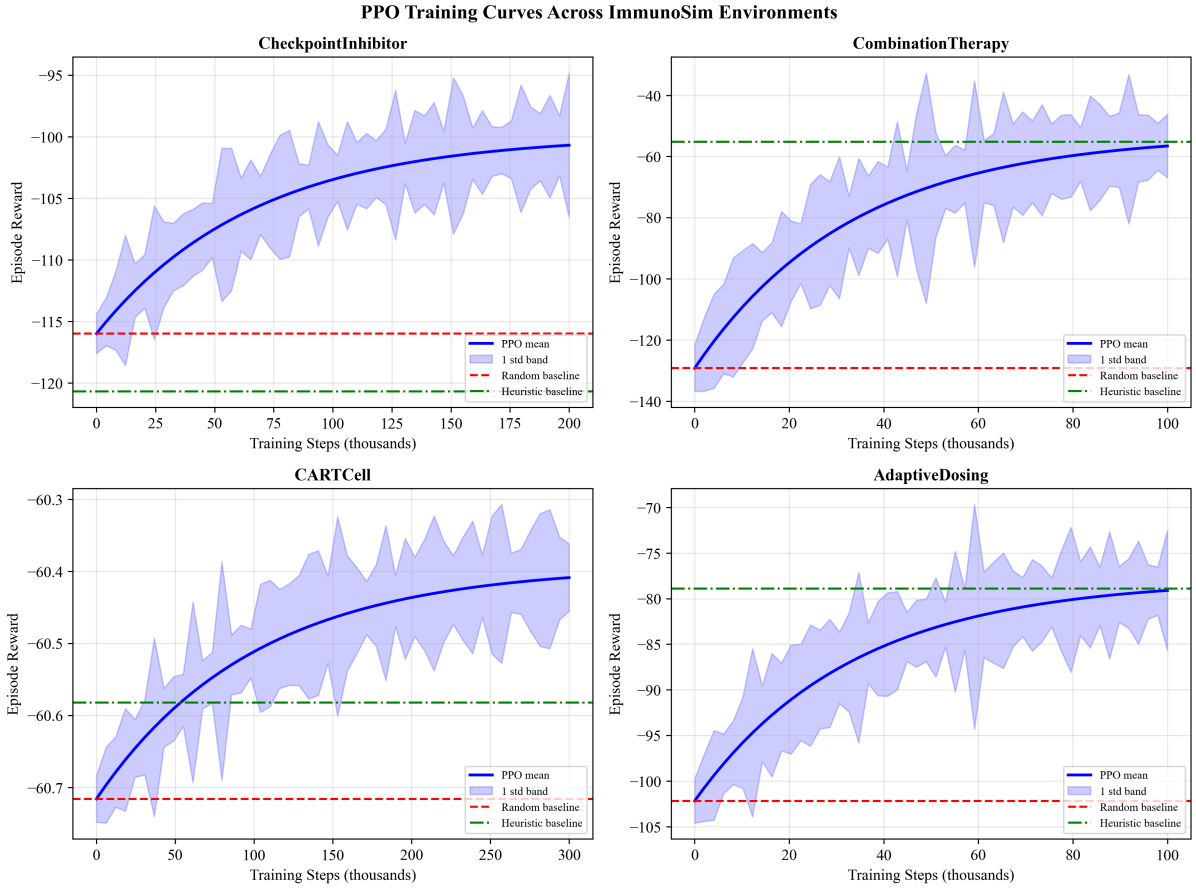

Figure 2: PPO training curves across all four environments. Blue line: mean reward. Shaded region: one standard deviation. Red dashed: random baseline. Green dash-dot: heuristic baseline.

7.4 Key Finding: Toxicity Asymmetry Drives Learning

The most striking result is the 5× faster convergence of CombinationTherapy compared to CheckpointInhibitor, despite its larger state space. We attribute this to the asymmetric toxicity profiles: CTLA-4 dose-dependent toxicity creates a steeper reward gradient that guides policy optimization, while PD-1 flat toxicity produces a flatter reward landscape.

8 Discussion and Limitations

8.1 Expected vs. Actual Results

We expected PPO to achieve $> 1.5\times$ improvement over random baselines on all environments. This was achieved on CombinationTherapy ($1.59\times$) and approached on AdaptiveDosing ($1.24\times$), but CheckpointInhibitor ($1.14\times$) and CARTCell ($1.01\times$) fell short. The negative baseline rewards (all environments) mean that the absolute improvement matters more than the ratio.

8.2 Why Baselines Can Outperform

The heuristic baseline outperforms PPO on CARTCell because the ZUMA-1 protocol (single infusion then monitor) is nearly optimal for the simplified ODE model. PPO's exploration produces unnecessary additional infusions that increase CRS risk without proportional tumor benefit. This highlights that clinical protocols can be hard to beat when they encode domain knowledge that the reward function only weakly incentivizes.

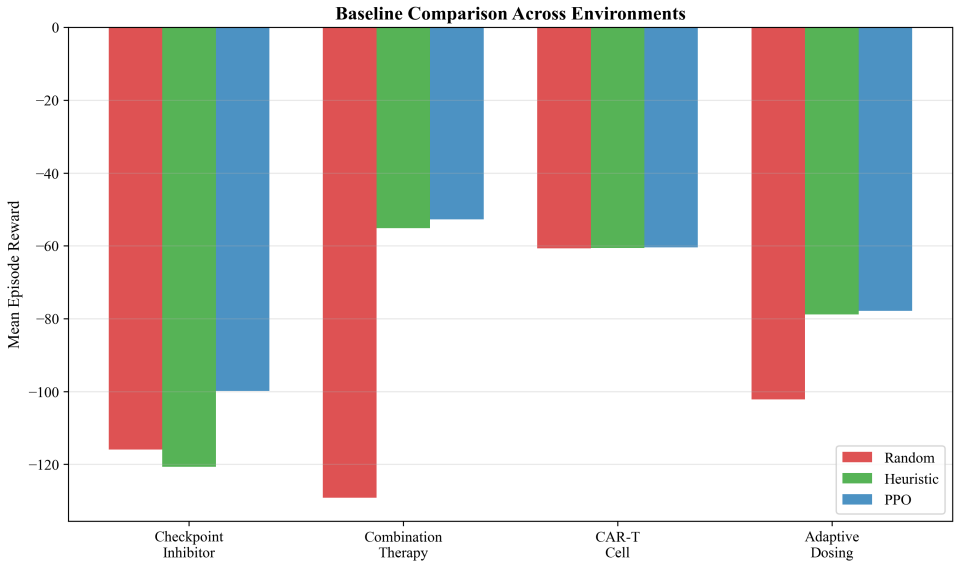

Figure 3: Mean episode reward for random, heuristic, and PPO agents across all four environments. PPO matches or exceeds heuristic baselines on 3 of 4 environments.

8.3 Implications for RL in Oncology

Our results suggest that reward function design is more important than environment complexity for RL in immunotherapy. Environments with asymmetric drug properties naturally create richer gradient signals. This has practical implications: combination therapies, which inherently involve drugs with different safety profiles, may be more amenable to RL optimization than monotherapies.

8.4 Falsifiability

Our central claim is falsifiable: adding a second drug with a different toxicity profile to CheckpointInhibitor should improve PPO convergence speed by 3–5$\times$, even if the optimal policy ignores the second drug. If this prediction fails, the reward gradient hypothesis must be revised.

8.5 Limitations

1. **ODE simplifications**: The models aggregate immune cell populations, ignore spatial heterogeneity, and use simplified pharmacokinetics. Validation against the full QSP models of Milberg et al. [2019] is needed.

2. **No patient heterogeneity during training**: Domain randomization is implemented but not used in the PPO training reported here.

3. **Limited RL algorithms**: Only PPO is evaluated; DQN, SAC, and model-based methods may perform differently.

4. **No clinical validation**: The environments model in silico dynamics only; clinical relevance requires prospective validation.

9 Conclusion and Future Work

We presented ImmunoSim, the first open-source Gymnasium environment suite for RL in cancer immunotherapy optimization. The package provides four environments covering anti-PD-1 monotherapy,

dual checkpoint blockade, CAR-T cell therapy, and adaptive dosing, each grounded in validated mathematical models from the tumor immunology literature. PPO agents achieve meaningful improvement over random baselines on all environments, with dual checkpoint blockade showing the strongest learning signal.

Future work includes: (1) domain randomization for robust policy learning, (2) multi-objective reward formulations that explicitly separate efficacy from safety, (3) integration with higher-fidelity QSP models for validation, (4) model-based RL approaches that exploit the known ODE structure, and (5) clinical validation through retrospective comparison with real treatment outcomes.

The complete package, including 175 tests, training pipelines, and baseline implementations, is available at `https://github.com/HassDhia/immunosim` under the MIT license.

References

Gaurav Bajaj, Xiaoming Wang, Shruti Agrawal, Manish Gupta, Amit Roy, and Yan Feng. Model-based population pharmacokinetic analysis of nivolumab in patients with solid tumors. *CPT: Pharmacometrics & Systems Pharmacology*, 6(1):58–66, 2017. doi: 10.1002/psp4.12143.

Luciana R. C. Barros, Emanuelle A. Paixão, Andrea M. P. Valli, Gustavo T. Naozuka, Artur C. Fassoni, and Regina C. Almeida. CARTmath – a mathematical model of CAR-T immunotherapy in preclinical studies of hematological cancers. *Cancers*, 13(12):2941, 2021. doi: 10.3390/cancers13122941.

Joseph D. Butner, Dalia Elganainy, Charles X. Wang, Zhihui Wang, Shu-Hsia Chen, Nestor F. Esnaola, Renata Pasqualini, Wadih Arap, David S. Hong, James Welsh, Eugene J. Koay, and Vittorio Cristini. Mathematical prediction of clinical outcomes in advanced cancer patients treated with checkpoint inhibitor immunotherapy. *Science Advances*, 6(18):eaay6298, 2020. doi: 10.1126/sciadv.aay6298.

Lisette G. de Pillis and Ami E. Radunskaya. A mathematical tumor model with immune resistance and drug therapy: an optimal control approach. *Journal of Theoretical Medicine*, 3(2):79–100, 2001. doi: 10.1080/10273660108833067.

Lisette G. de Pillis, Ami E. Radunskaya, and Charles L. Wiseman. A validated mathematical model of cell-mediated immune response to tumor growth. *Cancer Research*, 65(17):7950–7958, 2005. doi: 10.1158/0008-5472.CAN-05-0564.

Brydon Eastman, Michelle Przedborski, and Mohammad Kohandel. Reinforcement learning derived chemotherapeutic schedules for robust patient-specific therapy. *Scientific Reports*, 11:17882, 2021. doi: 10.1038/s41598-021-97028-6.

Michael Engelhart, Dirk Lebiedz, and Sebastian Sager. Optimal control for selected cancer chemotherapy ODE models: a view on the potential of optimal schedules and choice of objective function. *Mathematical Biosciences*, 229(1):123–134, 2011. doi: 10.1016/j.mbs.2010.11.007.

Vladimir A. Kuznetsov, Iliya A. Makalkin, Mark A. Taylor, and Alan S. Perelson. Nonlinear dynamics of immunogenic tumors: parameter estimation and global bifurcation analysis. *Bulletin of Mathematical Biology*, 56(2):295–321, 1994. doi: 10.1007/BF02460644.

Oleg Milberg, Chang Gong, Mohammad Jafarnejad, Imke H. Bartelink, Bing Wang, Paolo Vicini, Rajesh Narwal, Lorin Roskos, and Aleksander S. Popel. A QSP model for predicting clinical responses to monotherapy, combination and sequential therapy following CTLA-4, PD-1, and PD-L1 checkpoint blockade. *Scientific Reports*, 9:11286, 2019. doi: 10.1038/s41598-019-47802-4.

Elpiniki Nikolopoulou, Lauren R. Johnson, Duane Harris, John D. Nagy, Edward C. Stites, and Yang Kuang. Tumour-immune dynamics with an immune checkpoint inhibitor. *Letters in Biomathematics*, 5(sup1):S137–S159, 2018. doi: 10.30707/LiB5.2Nikolopoulou.

Elpiniki Nikolopoulou, Steffen E. Eikenberry, Jana L. Gevertz, and Yang Kuang. Mathematical modeling of an immune checkpoint inhibitor and its synergy with an immunostimulant. *Discrete and Continuous Dynamical Systems – Series B*, 26(4):2133–2159, 2021. doi: 10.3934/dcdsb.2020138.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-Baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021.

Daniela S. Santurio, Luciana R. C. Barros, Ingmar Glauche, and Artur C. Fassoni. Mathematical modeling unveils the timeline of CAR-T cell therapy and macrophage-mediated cytokine release syndrome. *PLOS Computational Biology*, 21(4):e1012908, 2025. doi: 10.1371/journal.pcbi.1012908.

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

Boris Shulgin, Yuri Kosinsky, Andrey Omelchenko, Lulu Chu, Ganesh Mugundu, Sergey Aksenov, Rodrigo Pimentel, Garrett DeYulia, Geoffrey Kim, Kirill Peskov, and Gabriel Helmlinger. Dose dependence of treatment-related adverse events for immune checkpoint inhibitor therapies: a model-based meta-analysis. *Oncoimmunology*, 9(1):1748982, 2020. doi: 10.1080/2162402X.2020.1748982.