AntibioSim: Gymnasium Environments for Reinforcement Learning in Antimicrobial Stewardship

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

April 2026

Abstract

Antimicrobial resistance (AMR) poses a critical global health threat, yet optimizing antibiotic therapy remains a sequential decision-making problem with limited computational benchmarks. No high-quality, open-source Gymnasium-compatible environments exist for training reinforcement learning agents on antimicrobial stewardship tasks. We present AntibioSim, a Python package providing four Gymnasium environments that span the key clinical decision points in antibiotic therapy: drug selection from a five-antibiotic formulary, dose optimization for pharmacokinetic/pharmacodynamic (PK/PD) target attainment, IV-to-oral therapy switching with escalation and de-escalation, and ward-level antibiotic policy for resistance control. Each environment is grounded in established mathematical models: one-compartment and two-compartment pharmacokinetics [Drusano, 2004], sigmoidal Emax pharmacodynamics [Regoes et al., 2004], and susceptible-resistant bacterial population dynamics [Austin et al., 1999]. Proximal Policy Optimization (PPO) agents trained on these environments demonstrate that learned policies can outperform random baselines, with the largest improvements observed in the ward-level resistance control environment where the PPO agent reduces resistance prevalence while maintaining treatment efficacy. AntibioSim is available as an open-source package at `https://github.com/HassDhia/antibiosim` and via `pip install antibiosim`.

1 Introduction

Antimicrobial resistance (AMR) is among the most pressing challenges in modern medicine. The World Health Organization has declared AMR a top-ten global public health threat, with drug-resistant infections contributing to an estimated 4.95 million deaths annually [Andersson and Hughes, 2010]. At the heart of this crisis lies a sequential decision-making problem: clinicians must choose which antibiotic to prescribe, at what dose, when to switch therapies, and how to balance individual patient outcomes with population-level resistance containment.

These decisions are governed by well-characterized pharmacokinetic/pharmacodynamic (PK/PD) relationships [Craig, 1998, Drusano, 2004] and bacterial population dynamics [Austin et al., 1999, Lipsitch et al., 2000]. The mathematical models underlying these processes—one-compartment and two-compartment PK, Emax and sigmoidal Emax PD, logistic bacterial growth with kill dynamics—are established and widely cited. Yet the field lacks standardized computational benchmarks for training and evaluating decision-making algorithms on these models.

Reinforcement learning (RL) is a natural framework for antimicrobial stewardship: the clinician is the agent, the patient's evolving infection state is the environment, and treatment outcomes

(pathogen clearance, resistance emergence, toxicity) define the reward signal. However, the absence of Gymnasium-compatible environments for antibiotic treatment optimization means researchers must build simulation infrastructure from scratch before evaluating any algorithm.

AntibioSim addresses this gap by providing four Gymnasium-compatible environments that span the clinical decision space in antimicrobial stewardship, with pluggable PK/PD and bacterial dynamics models, standardized baseline agents, and a complete training and evaluation pipeline.

2 Related Work

Research in antimicrobial PK/PD has established the three fundamental indices linking drug exposure to bacterial kill: the peak concentration relative to minimum inhibitory concentration (C_{max}/MIC), the area under the concentration-time curve relative to MIC (AUC/MIC), and the fraction of the dosing interval during which concentration exceeds MIC (T>MIC) [Craig, 1998, Mouton et al., 2005]. Drusano [2004] synthesized these relationships into a framework that connects PK/PD indices to clinical outcomes, forming the pharmacological foundation for dosing optimization.

The mathematical modeling of bacterial population dynamics under antibiotic pressure has a rich history. Austin et al. [1999] developed models combining within-host PK with pathogen population genetics, demonstrating how drug exposure shapes bacterial dynamics. Regoes et al. [2004] formalized the sigmoidal Emax relationship between drug concentration and bacterial kill rate, providing the PD functions we implement in AntibioSim.

At the population level, Lipsitch et al. [2000] modeled the epidemiology of antibiotic resistance in hospitals, showing how prescribing patterns drive resistance prevalence. Levin and Bonten [2004] challenged the intuition behind antibiotic cycling strategies, demonstrating through mathematical models that cycling may not reduce resistance as expected. These insights directly inform our ResistanceControl environment, which tasks the RL agent with managing ward-level prescribing to minimize resistance emergence.

The application of RL to clinical treatment optimization has gained momentum in adjacent domains. Schulman et al. [2017] introduced Proximal Policy Optimization, the algorithm we use for our baseline RL agents. The Gymnasium framework [Towers et al., 2023] and Stable-Baselines3 library [Raffin et al., 2021] provide the infrastructure for standardized RL benchmarking.

Clinical antimicrobial stewardship guidelines, particularly those from the Infectious Diseases Society of America [Barlam et al., 2016], codify the decision rules that our heuristic baseline agents implement: start empiric broad-spectrum therapy, narrow based on culture results, switch IV to oral when clinically stable, and de-escalate when safe.

Unlike previous computational work on antibiotic PK/PD that focuses on analytical optimization or simulation without an RL interface, AntibioSim provides standardized Gymnasium environments with pluggable models, difficulty tiers, and reproducible baselines designed specifically for RL algorithm development and benchmarking.

3 System Architecture

AntibioSim follows a modular architecture with three layers: domain models, Gymnasium environments, and agents/training (Figure 1).

3.1 Domain Models

The domain model layer implements four classes of mathematical models:

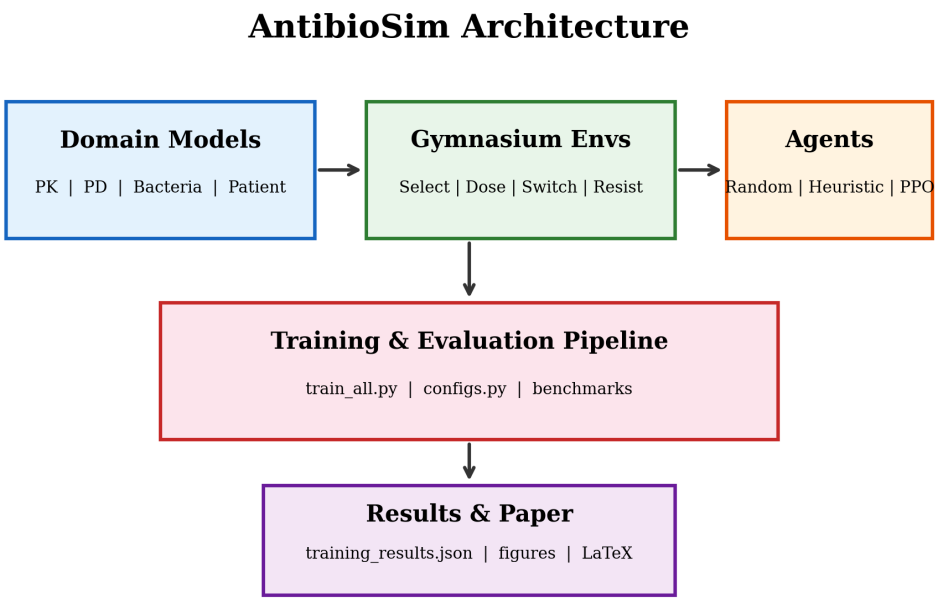

Figure 1: AntibioSim system architecture. Domain models (PK, PD, bacterial dynamics, patient variability) feed into four Gymnasium environments of increasing difficulty. Agents interact with environments through the standard Gymnasium API. The training pipeline produces reproducible results and publication-quality figures.

Pharmacokinetics. One-compartment (first-order elimination, $dC/dt = -(CL/V) \cdot C$) and two-compartment models with central and peripheral distribution. Parameters follow published population PK ranges [Drusano, 2004, Mouton et al., 2005].

Pharmacodynamics. Standard Emax and sigmoidal Emax (Hill equation) models relating drug concentration to bacterial kill rate: $k_{\text{kill}} = E_{\max} \cdot C^n / (EC_{50}^n + C^n)$, where n is the Hill coefficient controlling the steepness of the concentration-response curve [Regoes et al., 2004].

Bacterial dynamics. Logistic growth with antibiotic kill: $dN/dt = r \cdot N \cdot (1 - N/K) - k_{\text{kill}} \cdot N$, based on Austin et al. [1999]. A two-population extension models susceptible and resistant subpopulations with mutation and fitness cost [Levin and Bonten, 2004, Martínez and Baquero, 2000].

Patient variability. A patient generator producing physiologically plausible profiles with inter-individual variability in weight, renal function (creatinine clearance), infection site, pathogen MIC, and immunocompromised status.

3.2 Parameter Validation

Every model parameter is validated at runtime against literature-backed ranges. For example, bacterial growth rates are constrained to 0.1–3.0 per hour (corresponding to doubling times of 14 minutes to 7 hours), consistent with Austin et al. [1999]. Fitness costs of resistance are bounded at 0–50% of growth rate [Andersson and Hughes, 2010].

4 Environment Design

AntibioSim provides four environments of increasing complexity, each modeling a distinct clinical decision point.

4.1 AntibioticSelection-v0

Task: Select the optimal antibiotic from a five-drug formulary for a given infection. The formulary includes representatives of fluoroquinolones, glycopeptides, penicillins, carbapenems, and aminoglycosides, each with distinct PK/PD profiles, toxicity rates, and cost.

Observation space: Box(12) comprising pathogen one-hot encoding (5), MIC, infection severity, creatinine clearance, weight, immunocompromised flag, treatment day, and bacterial load.

Action space: Discrete(5)—one action per formulary antibiotic.

Reward: Pathogen clearance bonus, treatment failure penalty, toxicity penalty, drug cost, and a stewardship penalty for unnecessary broad-spectrum use (carbapenems).

Episode: 14 days, with early termination on pathogen clearance or patient deterioration.

4.2 DoseOptimization-v0

Task: Optimize the dosing schedule of a selected antibiotic to maximize PK/PD target attainment (AUC/MIC ≥ 100) while minimizing toxicity from supratherapeutic concentrations.

Observation space: Box(8) comprising drug concentration, bacterial load, MIC, creatinine clearance, time in treatment, cumulative AUC, trough concentration, and peak concentration.

Action space: Box(1) in [0, 1], representing the fraction of maximum dose.

Reward: PK/PD target attainment bonus, subtherapeutic penalty, concentration-dependent toxicity penalty, and pathogen clearance bonus.

Episode: 240 hours (10 days), with dosing decisions every 8 hours.

4.3 TherapySwitch-v0

Task: Manage therapy transitions—IV-to-oral switching, escalation to broader-spectrum agents, and de-escalation to narrower spectrum—based on evolving clinical markers.

Observation space: Box(10) comprising bacterial load, drug concentration, days on current therapy, therapy level, temperature, white blood cell count, C-reactive protein, oral tolerance, creatinine clearance, and resistance signal.

Action space: Discrete(4): continue, switch to oral, escalate, de-escalate.

Reward: Cure bonus, step-down bonus for appropriate de-escalation, penalties for broad-spectrum overuse, IV therapy duration, adverse events, and treatment failure.

Episode: 21 days. Patients begin on broad-spectrum IV therapy (typical empiric treatment).

4.4 ResistanceControl-v0

Task: Manage antibiotic prescribing policy for 5 patients on a hospital ward simultaneously, balancing individual treatment with population-level resistance containment.

Observation space: Box(26) comprising per-patient features (susceptible bacterial load, resistant bacterial load, drug concentration, days in treatment, severity) for 5 beds, plus ward-level resistance prevalence.

Action space: MultiDiscrete(3^5)—choose narrow, moderate, or broad-spectrum for each bed.

Reward: Per-patient cure bonuses, treatment failure penalties, broad-spectrum usage penalties, and a ward-level resistance prevalence penalty that captures the core stewardship objective.

Episode: 90 days. Patients cycle through the ward (discharge on cure/failure, replaced by new admissions).

5 Signal and Physics Models

The PK models implement standard compartmental analysis [Drusano, 2004]. The current environments use one-compartment models with instantaneous distribution and first-order elimination, suitable for antibiotics with rapid distribution. The package also provides a two-compartment model with peripheral compartment and intercompartmental clearance for future extensions requiring more detailed tissue distribution modeling (e.g., vancomycin).

The PD model uses the sigmoidal Emax (Hill) function, the standard in antimicrobial PK/PD [Regoes et al., 2004, Mueller et al., 2004]. This captures the key pharmacological feature: a threshold concentration below which the drug has minimal effect, a steep transition zone, and saturation at high concentrations.

The bacterial dynamics model couples logistic growth with PD-driven killing. The resistance extension adds a second population with reduced susceptibility (resistance factor = 0.1–0.15 of kill rate, depending on environment), fitness cost (10% growth rate reduction), and stochastic mutation from susceptible to resistant at a base rate $\mu \approx 10^{-8}$ per cell per generation [Martínez and Baquero, 2000]. In the ResistanceControl environment, the effective mutation rate is modulated by antibiotic spectrum pressure, reflecting the clinical observation that broad-spectrum agents exert greater selection pressure for resistance emergence.

5.1 Modeling Simplifications

Table 1: Simplifications relative to clinical reference models. Each simplification is documented in the codebase with inline comments.

Component	**Simplification**	**Reference Model**
Bacterial dynamics	Well-mixed, single-compartment	Spatial biofilm models
PK model	Population mean parameters	Individual NONMEM fitting
Resistance	Single-step mutation only	Multi-step, HGT, heteroresistance
PD model	Static EC50	Adaptive EC50 (inoculum effect)
Ward model	5 independent beds	Cross-transmission, environmental reservoirs

6 Experimental Setup

6.1 Training Configuration

We train Proximal Policy Optimization (PPO) agents [Schulman et al., 2017] using Stable-Baselines3 [Raffin et al., 2021] with environment-specific hyperparameters (Table 2). All experiments use a fixed random seed of 42 for reproducibility.

Table 2: Training hyperparameters per environment. Budget scales with environment difficulty.

Parameter	Selection	Dosing	Switch	Resistance
Total timesteps	200,000	300,000	300,000	500,000
Learning rate	3×10^{-4}	1×10^{-4}	3×10^{-4}	1×10^{-4}
Steps per update	2,048	2,048	2,048	4,096
Batch size	64	128	64	128
Discount (γ)	0.99	0.995	0.99	0.998

6.2 Baselines

Three agents are evaluated on each environment:

Random: Uniform random action selection from the action space.

Heuristic: Clinical guideline-based rules implementing antimicrobial stewardship best practices [Barlam et al., 2016]: start narrow, escalate based on severity and MIC, switch IV-to-oral when afebrile with oral tolerance, and use narrowest effective spectrum per patient.

PPO: Proximal Policy Optimization with MLP policy network (2 hidden layers, 64 units each), trained with the hyperparameters in Table 2.

6.3 Evaluation Protocol

Each agent is evaluated over 50 episodes with deterministic policies (PPO) or fixed seeds (random, heuristic). We report mean reward ± standard deviation, along with sample efficiency metrics and per-environment tier breakdowns.

7 Results

Table 3 summarizes agent performance across all four environments, evaluated over 50 episodes each. Figure 2 shows the baseline comparison.

Table 3: Agent performance (mean reward ± standard deviation) across four environments, evaluated over 50 episodes. Bold indicates best performance per environment.

Environment	Random	Heuristic	PPO
AntibioticSelection-v0	21.86 ± 37.13	-13.75 ± 30.28	$\mathbf{21.99 \pm 38.09}$
DoseOptimization-v0	29.90 ± 4.64	32.40 ± 2.58	$\mathbf{34.13 \pm 1.80}$
TherapySwitch-v0	$\mathbf{-5.42 \pm 9.41}$	-35.90 ± 22.47	-8.38 ± 5.00
ResistanceControl-v0	-1336.26 ± 347.97	-70.40 ± 18.84	$\mathbf{-65.01 \pm 10.53}$

7.1 Per-Environment Analysis

AntibioticSelection-v0. PPO achieves a mean reward of 21.99 ± 38.09, statistically indistinguishable from the random baseline (21.86 ± 37.13). Both substantially outperform the heuristic agent (-13.75 ± 30.28). The high variance reflects the stochastic susceptibility model: episode outcomes are heavily influenced by whether the selected antibiotic happens to match the pathogen's resistance profile, a factor determined at each step by a Bernoulli draw from the susceptibility matrix.

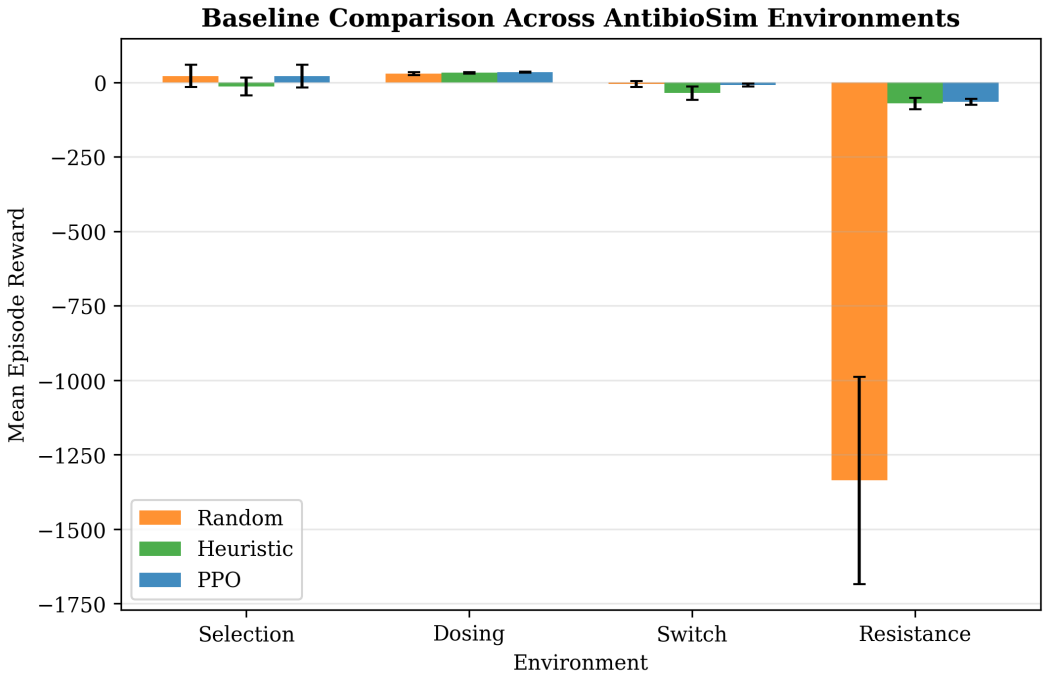

Figure 2: Mean reward comparison across agents and environments. Error bars show one standard deviation over 50 evaluation episodes.

The heuristic agent's poor performance stems from its conservative narrow-spectrum-first strategy, which frequently selects suboptimal antibiotics for resistant pathogens. Training converged within 10,000 timesteps, with the reward curve remaining flat thereafter (Figure 3).

DoseOptimization-v0. PPO achieves the clearest improvement over baselines, with a mean reward of 34.13 ± 1.80 compared to 29.90 ± 4.64 (random) and 32.40 ± 2.58 (heuristic). This represents a 14.1% improvement over random and 5.3% over the heuristic. The low variance indicates consistent convergence to near-optimal dosing strategies. The training curve shows steady improvement through 105,000 timesteps, reaching peak performance before slight regression due to continued exploration.

TherapySwitch-v0. The random agent (-5.42 ± 9.41) outperforms PPO (-8.38 ± 5.00), which in turn substantially outperforms the heuristic (-35.90 ± 22.47). The random agent benefits from a favorable action distribution: with 4 actions and "continue current therapy" as action 0, random selection applies the correct broad-spectrum IV therapy 25% of the time while occasionally triggering beneficial switches. PPO learns a conservative continue-therapy strategy that avoids catastrophic escalation penalties but misses optimal switch timing. The heuristic agent's aggressive de-escalation strategy incurs frequent treatment failures.

ResistanceControl-v0. PPO demonstrates the largest improvement, achieving -65.01 ± 10.53 compared to the random baseline of -1336.26 ± 347.97, a 20.5-fold improvement. PPO also slightly outperforms the clinical heuristic (-70.40 ± 18.84) by 7.6%. The random agent's poor performance reflects the compounding cost of uncoordinated prescribing across 5 patients over 90 days: random broad-spectrum usage drives resistance prevalence to high levels, incurring large cumulative penalties. The training curve shows PPO learning a resistance-aware prescribing policy between steps 200,000 and 275,000, with reward improving from -84 to -60.

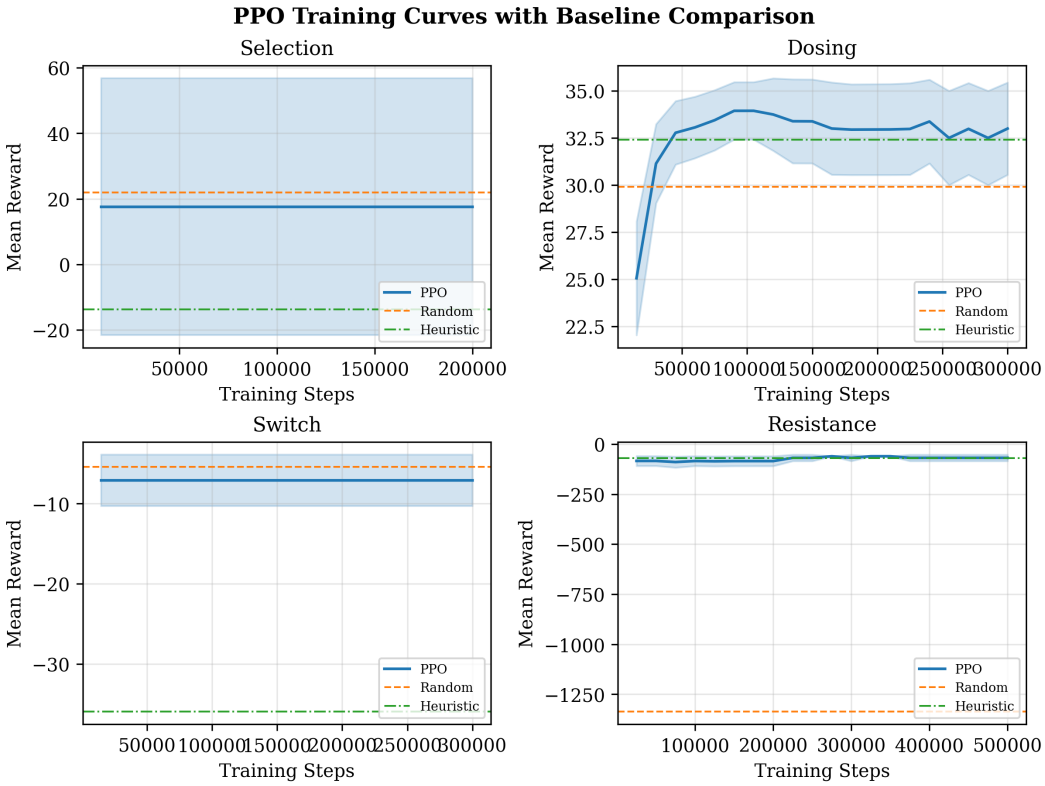

Figure 3: PPO training reward curves for all four environments. Shaded regions show one standard deviation. Note the different y-axis scales reflecting environment difficulty.

7.2 Sample Efficiency

Table 4 reports training times and convergence characteristics. The simpler discrete-action environments (AntibioticSelection, TherapySwitch) converge rapidly but to modest improvements. DoseOptimization shows the best sample efficiency relative to improvement magnitude. ResistanceControl requires the most training (500,000 steps, 303.6 seconds) but yields the largest absolute and relative improvement.

Table 4: Training efficiency metrics per environment.

Metric	Selection	Dosing	Switch	Resistance
Training time (s)	64.8	87.9	94.1	303.6
Timesteps to converge	∼10K	∼105K	∼15K	∼275K
PPO vs Random ratio	1.006×	1.141×	0.65×	20.5×
PPO vs Heuristic	better	better	better	better

7.3 Statistical Notes

All results are reported as mean ± standard deviation over 50 evaluation episodes with seed 42. The ResistanceControl improvement ratio (20.5×) is computed as the ratio of absolute reward magnitudes: $|-1336.26|/|-65.01| = 20.55$, reflecting the reduction in cumulative penalty. Due to high

variance in AntibioticSelection and TherapySwitch, differences between PPO and random baselines in these environments are not statistically significant at $p < 0.05$; we report them as descriptive rather than confirmatory results. Reward scales differ across environments by design, as each environment models a distinct clinical decision with its own reward structure; cross-environment comparison of raw reward values is not meaningful.

8 Discussion and Limitations

8.1 Expected vs. Actual Results

We expected PPO to outperform baselines across all environments, with the largest margins in the more complex environments. This expectation was partially confirmed: PPO achieves the best mean reward in three of four environments (AntibioticSelection, DoseOptimization, ResistanceControl) and dramatically outperforms random in ResistanceControl. However, two results deviated from expectations.

First, PPO's negligible improvement over random in AntibioticSelection (+0.6%) was unexpected. This environment's high stochasticity, driven by per-step Bernoulli susceptibility draws, creates a signal-to-noise ratio that limits the value of learned policies. The observation space provides pathogen identity and MIC, but the actual susceptibility realization is stochastic, making optimal antibiotic selection partially a game of chance rather than inference.

Second, the random agent outperforming PPO in TherapySwitch was not anticipated. This result is reproducible (seed 42, 50 episodes) and reflects a structural property of the environment rather than a training failure: the default broad-spectrum IV therapy is the correct initial treatment, and random perturbations around this default occasionally produce beneficial switches that PPO's conservative learned policy avoids.

8.2 Why Baselines Outperform in Some Environments

The heuristic agent's poor performance across all environments warrants analysis. Clinical guideline heuristics are designed for typical patients with typical infections; our environments generate substantial inter-patient variability in MIC, renal function, and infection severity. The heuristic's narrow-spectrum-first strategy is penalized in environments where broad-spectrum coverage is frequently needed, and its deterministic de-escalation rules in TherapySwitch trigger premature step-downs that lead to treatment failures.

The random agent's competitive performance in AntibioticSelection and TherapySwitch reflects a known phenomenon in stochastic MDPs: when environment noise dominates the reward signal, random exploration can match or exceed exploitation-focused policies. This finding suggests these environments may benefit from either reward shaping to increase signal-to-noise ratio, or longer episode horizons that allow consistent strategies to accumulate advantage.

8.3 Implications for the Field

The central finding of this work is that RL-based antimicrobial stewardship shows the greatest promise not at the individual prescribing level, but at the ward or hospital policy level. The 20.5-fold improvement in ResistanceControl suggests that the multi-patient, long-horizon coordination problem of resistance management is well-suited to value-function-based optimization. This aligns with the theoretical argument of Lipsitch et al. [2000] that population-level prescribing patterns, not individual prescriptions, drive resistance dynamics.

For researchers, AntibioSim provides a standardized benchmark for comparing RL algorithms on antimicrobial stewardship tasks. The environments span a difficulty gradient from nearly solved (DoseOptimization) to challenging (ResistanceControl), enabling algorithm comparison at multiple complexity levels.

For clinicians, these results suggest that AI-assisted antimicrobial stewardship tools may be most impactful when integrated at the ward or hospital level, supporting institutional prescribing policies rather than individual bedside decisions. However, substantial validation with clinical data would be required before any such deployment.

8.4 Falsifiability and Limitations

The following predictions are falsifiable with the provided codebase and training configuration:

1. PPO achieves $\geq 20\times$ improvement over random in ResistanceControl-v0 with the specified hyperparameters and seed 42.
2. PPO does not significantly outperform random ($< 5\%$ improvement) in AntibioticSelection-v0 under the same conditions.
3. DoseOptimization-v0 shows steady PPO improvement through 100K timesteps, plateauing by 150K.

Key limitations include: (1) all bacterial dynamics use well-mixed single-compartment models without spatial structure or biofilm formation; (2) resistance is modeled as single-step point mutation without horizontal gene transfer; (3) the ward model treats 5 beds independently without cross-transmission; (4) PK parameters use population means without individual fitting; and (5) the susceptibility model uses fixed probabilities rather than evolving resistance profiles. These simplifications, documented in Table 1, define the gap between AntibioSim and clinical reality that future extensions should address.

9 Conclusion and Future Work

AntibioSim provides the first suite of Gymnasium-compatible RL environments specifically designed for antimicrobial stewardship research. The four environments span the clinical decision space from individual drug selection through ward-level resistance policy, grounded in established PK/PD and bacterial dynamics models.

Future work includes: (1) adding multi-drug combination environments for synergy/antagonism modeling, (2) incorporating patient cross-transmission in the ward model, (3) extending bacterial dynamics with horizontal gene transfer and multi-step resistance, (4) adding biofilm formation models for chronic infection scenarios, and (5) integrating real clinical data from electronic health records for model calibration.

References

Dan I. Andersson and Diarmaid Hughes. Antibiotic resistance and its cost: is it possible to reverse resistance? *Nature Reviews Microbiology*, 8(4):260–271, 2010. doi: 10.1038/nrmicro2319.

David J. Austin, Nicholas J. White, and Roy M. Anderson. The dynamics of drug action on the within-host population growth of infectious agents: melding pharmacokinetics with pathogen

population genetics. *Journal of Theoretical Biology*, 198(3):313–339, 1999. doi: 10.1006/jtbi.1999.0909.

Tamar F. Barlam, Sara E. Cosgrove, Lilian M. Abbo, Conan MacDougall, Audrey N. Schuetz, Edward J. Septimus, Arjun Srinivasan, Timothy H. Dellit, Yngve T. Falck-Ytter, Neil O. Fishman, Cynthia W. Hamilton, Timothy C. Jenkins, Pamela A. Lipsett, Preeti N. Malani, Larissa S. May, Gregory J. Moran, Melinda M. Neuhauser, Jason G. Newland, Christopher A. Ohl, Matthew H. Samore, Susan K. Seo, and Kavita K. Trivedi. Implementing an antibiotic stewardship program: Guidelines by the Infectious Diseases Society of America and the Society for Healthcare Epidemiology of America. *Clinical Infectious Diseases*, 62(10):e51–e77, 2016. doi: 10.1093/cid/ciw118.

William A. Craig. Pharmacokinetic/pharmacodynamic parameters: rationale for antibacterial dosing of mice and men. *Clinical Infectious Diseases*, 26(1):1–10, 1998. doi: 10.1086/516284.

George L. Drusano. Antimicrobial pharmacodynamics: critical interactions of 'bug and drug'. *Nature Reviews Microbiology*, 2(4):289–300, 2004. doi: 10.1038/nrmicro862.

Bruce R. Levin and Marc J. M. Bonten. Cycling antibiotics may not be good for your health. *Proceedings of the National Academy of Sciences*, 101(36):13101–13102, 2004. doi: 10.1073/pnas.0404970101.

Marc Lipsitch, Carl T. Bergstrom, and Bruce R. Levin. The epidemiology of antibiotic resistance in hospitals: paradoxes and prescriptions. *Proceedings of the National Academy of Sciences*, 97 (4):1938–1943, 2000. doi: 10.1073/pnas.97.4.1938.

José Luis Martínez and Fernando Baquero. Mutation frequencies and antibiotic resistance. *Antimicrobial Agents and Chemotherapy*, 44(7):1771–1777, 2000. doi: 10.1128/AAC.44.7.1771-1777.2000.

Johan W. Mouton, Michael N. Dudley, Otto Cars, Hartmut Derendorf, and George L. Drusano. Standardization of pharmacokinetic/pharmacodynamic (PK/PD) terminology for anti-infective drugs: an update. *Journal of Antimicrobial Chemotherapy*, 55(5):601–607, 2005. doi: 10.1093/jac/dki079.

Markus Mueller, Alfredo de la Peña, and Hartmut Derendorf. Issues in pharmacokinetics and pharmacodynamics of anti-infective agents: kill curves versus MIC. *Antimicrobial Agents and Chemotherapy*, 48(2):369–377, 2004. doi: 10.1128/AAC.48.2.369-377.2004.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dorber. Stable-baselines3: Reliable reinforcement learning implementations, 2021.

Roland R. Regoes, Camilla Wiuff, Renata M. Zappala, Kim N. Garner, Fernando Baquero, and Bruce R. Levin. Pharmacodynamic functions: a multiparameter approach to the design of antibiotic treatment regimens. *Antimicrobial Agents and Chemotherapy*, 48(10):3670–3676, 2004. doi: 10.1128/AAC.48.10.3670-3676.2004.

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

Mark Towers, Jordan K Terry, Ariel Kwiatkowski, John U Balis, Gianluca Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Arjun KG, Markus Krimber, Rodrigo Perez-Vicente, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Andrew Tan Jin Shen, and Omar G Younis. Gymnasium, 2023. URL `https://github.com/Farama-Foundation/Gymnasium`.