Dynamic Control of Stochastic Evolution: A Deep Reinforcement Learning Approach to Adaptively Targeting Emergent Drug Resistance

Dalit Engelhardt*

Department of Chemistry and Chemical Biology
Harvard University
Cambridge, MA 02138, USA

Abstract

The challenge in controlling stochastic systems in which random events can set the system on catastrophic trajectories is to develop a robust ability to respond to such events without significantly compromising the optimality of the baseline control policy. Drug resistance can emerge from random and variable mutations in targeted cell populations; in the absence of an appropriate dosing policy, emergent resistant subpopulations can proliferate and lead to treatment failure. Dynamic feedback dosage control holds promise in combatting this phenomenon, but cell population evolutionary dynamics are complex, stochastic, and often high-dimensional, posing significant challenges to system control. This paper presents CelluDose, a deep reinforcement learning closed-loop dynamic control prototype for automated precision drug dosing targeting stochastic and heterogeneous cell proliferation. Developing optimal dosing schedules for preventing therapy-induced drug resistance involves a tradeoff between the effective suppression of emergent resistant cell subpopulations and the use of conservative dosages and a preference for first-line drugs. CelluDose is trained on model simulations of cell population evolutionary dynamics that combine a system of stochastic differential equations and the additional occurrence of random perturbing events. Both the single-drug and combination therapy policies obtained in training exhibit a 100% success rate at suppressing simulated heterogeneous harmful cell growth and responding to diverse system fluctuations and perturbations within the allotted time and using conservative dosing. The policies obtained were found to be highly robust to model parameter changes and fluctuations not introduced during training.

1 Introduction

The development of drug resistance in response to therapy is a major cause of treatment failure worldwide across a variety of diseases, patient populations, and administered drugs. Resistance to a drug can be present prior to treatment or it can emerge during treatment [1–3] through a variety of mechanisms. Drug resistance evolves dynamically and often non-uniformly: intercellular variability can lead to faster adaptation to environmental pressures [4] and thus promote the rise of resistant subpopulations in a previously-susceptible population of cells. Tumor heterogeneity is now understood to be a major contributor to drug resistance in cancer [1, 5–7], and variability in resistance among bacterial cell populations has been tied to treatment failure [8, 9]. In bacterial evolution experiments, the

*dalit.engelhardt@gmail.com
emergence of multiple antibiotic-resistant cell lines has been observed [10, 11]. Such clonal diversity, even in a majority-susceptible cell population, can be the harbinger of drug resistance that ultimately leads to treatment failure: improper treatments can exert a selective evolutionary pressure that leads to the elimination of susceptible cell populations but leaves more resistant cells largely unaffected and able to proliferate due to reduced competition with susceptible cells [12].

The emergence and potential rise to dominance of resistant strains are stochastic processes driven by the dynamic and complex interaction between the influences and interplay of natural selection, demographic stochasticity, and environmental fluctuations. A predictive understanding of the likely evolutionary trajectories toward drug resistance is key to developing treatments that can effectively target and suppress resistant cell populations, but a fully predictive understanding of such processes remains a challenge. Control of evolution, however, does not in principle require full predictability or determinism in evolution. In a closed-loop setting, system feedback can mitigate the reliance on precise trajectory knowledge, so long as this feedback can be obtained at reasonable time intervals, the uncertainty and stochasticity can be approximated in an informed manner, and the controller is sufficiently robust to changes in system behavior and parameter fluctuations. In heterogeneous cell environments prone to resistance evolution, the development of such a control policy must balance the need for properly and efficiently targeting all cell types that may either exist at the onset of treatment or that may spontaneously emerge during therapy, while maintaining sufficiently low dosing for toxicity minimization.

The focus of this work is on the development of a closed-loop dynamic control prototype, CelluDose, for precision dosing that adaptively targets harmful cell populations of variable drug susceptibility and resistance levels based on discrete-time feedback on the targeted cell population structure. The time evolution of such systems is influenced by the potential presence, random emergence, and stochastic growth of multiple cell subpopulations, where demographic stochasticity arises from random processes on the single-cell level. Inappropriate dosing can lead to the proliferation of resistant cell populations, but this interplay is subject to significant levels of stochasticity that must be accounted for in the design of the control algorithm. The target goal of CelluDose is to combat the emergence of drug resistance during treatment by sensitively adjusting dosage and/or switching the administered drug in response to observed feedback on changes in the cell population structure while employing minimal overall drug dosage for toxicity reduction.

Significant and variable stochasticity, nonlinearity in the dynamics, and the potentially high-dimensional space of cell types pose significant challenges to the application of classical control methods to such systems. This extends beyond the scenarios considered in this paper: work on more complex disease progression and drug response models with anticipated higher predictive power has been on the rise in recent years, such as in systems pharmacology [13, 14], which brings systems biology approaches into integrative quantitative modeling in the drug discovery process. The expectation is that these increases in model complexity will result in more realistic descriptions of the dynamics that would offer significantly improved predictive power. However, as model complexity increases, the problem of disease progression control by way of model-informed precision dosing [15–19] becomes increasingly less tractable, necessitating a tradeoff between the ability to develop a controller for the system and the predictive accuracy of the system model [20].
A shift away from model-based control methods toward approaches that do not or only partly rely on model details will permit greater freedom in model selection and obviate the need for simplifications that may reduce predictability. The application of model-free reinforcement learning methods to continuous control tasks has seen significant advances in recent years\(^1\) and holds promise for the control of systems for which mathematical optimization and control may be intractable, with potential applications in diverse fields that include robotics, mathematical finance, and healthcare. However, for reinforcement learning to become a standard tool in control engineering and related fields more work is needed in the successful application of continuous reinforcement learning control to highly stochastic and realistic environments.

CelluDose combines a model-free deep reinforcement learning algorithm with model information in developing an adaptive dosing control policy for the elimination of harmful cell populations with heterogeneous drug responses and random yet significant demographic stochastic changes. Beyond presenting this implementation, the aims of this work are additionally to (1) motivate the use of model-free reinforcement learning control in the development of next-generation precision dosing controllers by demonstrating that a reliable and highly responsive dynamic control policy can be obtained for a complex and stochastic biological system that is intractable with traditional control methods, and (2) showcase how model knowledge may be capitalized on in such a scheme while still benefitting from the flexibility of its model-free learning approach.

2 Background

2.1 Reinforcement learning for continuous control

The aim of a therapeutic agent targeting harmful cell populations is to produce an effect that leads to the elimination of the harmful cell population while minimizing toxicity by avoiding the use of high doses. Given the possibility of emergence of high-resistance cell populations, this task involves a potentially complex tradeoff between acceptably low dosing levels, the need to effectively eliminate the targeted cell populations, and a preference for shorter treatment times to reduce patient morbidity and – in infectious diseases – the possibility of further contagion. As a result, the dosing problem can be cast in somewhat different ways that depend on the particular disease conditions and prevailing treatment preferences. For the purposes of this work, a successful treatment is defined as one in which all targeted harmful cell subpopulations were eliminated by some maximal acceptable treatment time. Within this measure of success optimality is defined as lowest expected cumulative dosing over the course of the treatment. A treatment course is designated as a failure if any targeted cells remain after the maximal treatment time has been reached, with the extent of failure dependent on the number of remaining cells.

Model-free reinforcement learning (RL) approaches aim to learn an optimal decision-making policy through an iterative process in which the policy is gradually improved by sequential tuning of parameters either of the policy itself or of a value (e.g. the action-value function) indicative of the policy’s optimality. The data on which learning is done is supplied in the form of transitions from a previous state \(s\) of the environment to the next

\(^1\)See, e.g., [21, 22] for recent surveys.
state $s'$ that occurs probabilistically (in stochastic environments) when a certain action $a$ is chosen. Each such transition is assigned a reward $r$, and reward maximization drives the optimization of the policy. Learning can be done on a step-by-step basis or episodically, where a single episode consists of multiple steps of transitions. In the case of the dosing problem considered here, an episode is a single-patient full course of therapy. It ends when either the maximal time has been reached or all cells have been eliminated, whichever comes first. Episodes are thus finite but can be long, with drugs administered at discrete time steps over a continuous range of dosages. At each time interval, a decision is made on which drugs at what doses should be administered based on observations of the current state of the disease, defined as the concentrations of the targeted cell types. However, only at the end of each episode does it become clear whether or not the sequence of therapeutic actions was in fact successful. This credit assignment problem [23] frequently plagues episodic tasks and will be addressed in Section 3.2.

The optimal policy for this decision-making process needs to provide next-time-step dosing guidelines under the optimality guidelines described above and based on potential stochastic evolutionary scenarios described by the model in Section 3.1. This system is continuous in its state space (cell population composition), involves time-varying and potentially high stochasticity, can be high-dimensional due to the large number of possibly-occurring mutant cell types, and involves one or multiple controls (drugs) that if administered intravenously (the scenario of interest here) can in principle take a continuous range of values. For reasons of mechanical operability and medical explainability, a deterministic dosing policy is needed in this case.

Deep deterministic policy gradient (DDPG) [24] is an off-policy actor-critic deterministic policy deep reinforcement learning algorithm suitable for continuous and high-dimensional observation and action spaces. Building on the deterministic policy gradient algorithm [25], DDPG employs neural network function approximation through the incorporation of two improvements used in Deep Q Networks [26, 27] for increasing learning stability: the use of target networks and a replay buffer from which mini-batches of $(s,a,s',r)$ transitions are sampled during training. We note that while the action space in the dosing problem is typically low-dimensional (only a few drugs are typically considered for treatment of a particular disease), the observation space (number of cell types) can be quite large, especially if finer observations are available on cellular heterogeneity (finer differentiations in dose responses). For these reasons, the dosing control scheme described here employs the DDPG algorithm.

2.2 Model-informed treatment planning with reinforcement learning

The benefit of a mechanistic model-informed approach to treatment planning is the ability to explore therapeutic decisions prior to the start of or as an accompaniment to clinical trials, which can then be used to inform clinical trials. The simulation-based use of RL here thus differs from work that employs reinforcement learning for making decisions based on patient data and clinical outcomes, where exploration of treatment parameter space is constrained by data available from previously-attempted treatments. When this data becomes available during and after clinical trials, these approaches and the approach presented here could be used in a complementary manner to inform and improve individualized treatments.
The use of RL for mechanistic model control in treatment planning is less common than data-driven approaches. We note below a few applications of interest. Q-learning [28], a discrete state and action space RL algorithm, was employed in [29] for closed-loop control of propofol anesthesia and in [30] for control of cancer chemotherapy based on an ordinary differential equations (ODE)-based deterministic model of tumor growth. In [31] a natural actor-critic algorithm [32] was employed for cancer chemotherapy control of a similar ODE-based model, with a binary action space (drug or no drug). Tumors were taken to be uniform and resistance to treatment was not considered as part of the models used in [30] and [31]. In [33] a DDPG-based algorithm was applied to an agent-based model of sepsis progression for obtaining an adaptive cytokine therapy policy.

Interest in adaptive control of therapy that takes resistance evolution into account has been on the rise in recent years (see, e.g., [34, 35]) but these efforts have been restricted to simple models that are typically low dimensional and/or deterministic. The approach presented here provides a mechanism for determining and automating dosing in a responsive and robust way that can be generalized to arbitrarily heterogeneous cell populations exhibiting complex and realistic dynamics. Both single-drug and combination therapy control policies are developed in this context, and observations on population composition are supplied at discrete time intervals, as would be expected in a laboratory or clinical scenario. By extension, this work seeks to motivate efforts in the high-resolution tracking of cell-level drug resistance progression within an individual patient. From a control engineering perspective, the incorporation of model information into model-free learning presented here may be usefully transported and adapted to the control of other stochastic systems with an equations-based description for improved RL learning stability and convergence. For reinforcement learning insights, a detailed analysis is included of policy features that were observed in training.

2.3 Implementation and workflow

Training and implementation rely on three key components: model and simulation development, selection of the subset of drugs of interest, knowledge of the range of likely resistance levels that have been recorded in response to these drugs, and access to diagnostics that can provide temporal data on cell population structure. In this paper, all data supplied in training is simulated; we describe below how such data may be obtained for future proof-of-concept validation.

For demonstrative purposes and facility of near-future validation, the focus of this work is on drug resistance in evolving bacterial populations subjected to growth-inhibiting antibiotic drugs. We note that evolutionary models of cell population dynamics are widely used to understand and model cancer progression and the development of resistance to chemotherapy treatment, and that another promising application area for a CelluDose-based platform is in cancer chemotherapy control.

Future clinical implementation of a CelluDose-based controller could be as an integrated system with the ability to track and supply the trained control model with sensitive measurements of the presence of small populations of cells and equipped with an automated intravenous infusion mechanism. This implementation would be particularly useful in intensive care units (ICUs), where the development of resistance to therapy is a major concern,
therapy is performed in the controlled environment of the clinic, and substantial genetic diversity in bacterial strains has been observed [37]. Alternatively, in longer-term therapies, it could be implemented as a decision support tool providing informed dosing recommendations to clinicians. In both cases, observations at the chosen fixed time intervals for which training was done would be supplied in the form of the respective targeted cell type concentrations, where a distinct cell type is (phenotypically) defined by non-negligible differences in the dose response across the possible ranges of dosages that are permitted for administration.

2.3.1 Modeling and simulation

Although training implements a model-free reinforcement learning algorithm, model knowledge is used in two ways: to provide the training data (as simulation data) and for algorithm design via feature engineering and reward assignment. This feature engineering is made possible by trajectory estimation from the equations-based component of the growth-mutation model implemented here. This model combines (1) a stochastic differential equations-based system for approximating the growth of existing cell subpopulations with (2) external random events that perturb this system by creating new subpopulations of resistant cells. These events, which increase the effective dimensionality and alter the composition of the system at random points in time, can be thought of as either de novo resistant mutations that occur during therapy or as the above-detection-threshold rise of mutant populations. The resulting system exhibits variable and substantial levels of stochasticity that can significantly alter its trajectory and potentially lead to treatment failure (see Fig. 1). Since, as shown in Section 4, the learned policies are highly robust to changes in the key mutation parameters, detailed knowledge of mutation rates is not an essential training parameter.

Stochasticity can arise from both demographic and environmental fluctuations. The focus here is on demographic stochasticity arising from single-cell variability in birth, death, and mutation processes. For simplicity, spatial homogeneity and a constant environment (other than changes in drug administration) will be assumed here. To obtain a quantifiable estimate of demographic stochasticity, a stochastic physics approach was employed here (Appendix A) in deriving the system of stochastic differential equations describing the heterogeneous cell population time evolution. This permits an informed estimate of the extent of demographic noise from individual cells’ probabilistic behavior and a subsequent quantification of an important contribution to risk in disease progression.

The universal parameters that enter the model are the drugs that may be used with preference information (e.g. which ones are first-line vs. last-line), the appropriate discrete time intervals in which observations become available and dosing can be altered, and the likely spectrum of dose responses. In laboratory evolution experiments [38] this spectrum (e.g. see Fig. 2) can be obtained by allowing the bacteria to naturally evolve under different drug levels and identifying subsequent cell lines accompanied by a characterization of their dose responses through parameter fitting of their growth curves at different drug concentrations. We note that it is not in general necessary to identify all possible and potential mutations prior to training. In particular, knowledge of all possible genetic changes is not in principle necessary: it suffices to characterize the expected relevant diversity in drug response.

\footnote{Such studies, however, are frequently accompanied by genotype analysis.}
idea is to provide enough training data for the model so that new variants are unlikely to exhibit dose responses considerably different from all those already identified. During an experimental run of the trained RL model, when a newly-emergent cell type subpopulation is detected it can thus be binned without significant loss of accuracy into the state-space dimension (Section 3.2) of a previously-identified variant.

2.3.2 Observations and diagnostics

Early, fast, and high-resolution detection of heterogeneity in drug response is crucial for identifying resistant cell subpopulations before such populations spread [8, 39, 40]. Resistance detection may be done through phenotype-based drug susceptibility testing or by capitalizing on advances in genomic sequencing in combination with predictions of resistance from genotype [11, 41]. A key aspect is the requisite ability to effectively detect low concentrations of resistant cells in an otherwise susceptible population. Single-cell analysis is showing mounting promise in this regard and is being used to resolve heterogeneity in tumors [42] and in bacterial populations [43, 44]. Direct phenotype-based detection of small resistant bacterial subpopulations in proportions as low as $10^{-6}$ of the total population was performed in [45]. For genotype detection, single-cell sequencing [46] can be combined with predictions of resistance from genotype for high-resolution detection of a population’s resistance profile.

These advances could be applied to near-future laboratory validation of an in vitro CelluDose controller via combination with software implementation into an automated liquid handling system in bacterial evolution experiments. Greater automation, efficiency, and standardization are nonetheless needed for these technologies to be routinely employed as part of a clinical controller apparatus, and, in part, the intention of this work is to motivate efforts in the clinical development and use of single-cell analysis for resistance profiling by demonstrating its potential utility for automated disease control.

3 Methods

3.1 Stochastic modeling and simulation of cell population evolution

The evolutionary fate of an individual cell during any given time interval depends on the timing of its cell cycle and its interaction with and response to its environment. The model of evolutionary dynamics employed here depends on three processes: cell birth, cell death, and inheritable changes in cell characteristics that lead to observable changes in growth under inhibition by an antibiotic. We consider here drugs that suppress cell growth by inhibiting processes essential for cell division by binding to relevant targets and thus leading to reduced cell birth rates. The dose-response relationship can be described by a Hill-type relationship [47, 48]. Here a modified Hill-type equation will be employed that includes the potential use of multiple drugs with concentrations given by $I = (I_1, \ldots, I_m)$.

The growth rate $g_i$ of a particular cell type $i$ as a function of the drug concentrations to which cells are exposed is thus taken to be

$$g_i(I) = \beta_i(I) - \delta_i = \frac{\beta_i,0}{1 + \sum_\alpha (I_\alpha / \rho_{i,\alpha})} - \delta_i$$

(1)
where $\beta_i$ is the rate of cell birth, $\beta_{i,0}$ is its drug-free birth rate, $\delta_i$ is the rate of cell death, and $\rho_{i,\alpha}$ describes the extent of resistance of cell type $i$ to drug $\alpha$. For simplicity, since drugs are assumed here to affect only birth rates, changes that confer resistance are assumed to affect only $\beta_i(I)$ rather than $\delta_i$. Resources for cell growth are assumed to be limited, with the environmental carrying capacity denoted by $K$, indicating the total bacterial cell population size beyond which no further population increases take place due to resource saturation.

Evolution is an inherently stochastic process that depends on processes, as noted above, that individual cells undergo. However, individual (agent)-based simulations are generally very costly for larger numbers of cells and, moreover, obscure insight into any approximate deterministic information about the system. At very high population levels (that can be treated as effectively infinite) a deterministic description of system dynamics is appropriate. In the dynamic growth scenario considered here, large size discrepancies can exist at any time between the various concurrent cell populations, and a single population can decay or proliferate as the simulation progresses. The extent of demographic stochasticity in such systems will thus change in the course of evolution. For efficient modeling at higher population levels it is imperative to use a simulation model whose runtime does not scale with population size, but for accurate modeling a proper estimate of stochasticity must be included. This is done here by applying methods from stochastic physics [49–51] to derive a diffusion approximation of a master equation describing the continuous-time evolution of the probability distribution of the population being in some state $(n_1, n_2, \ldots, n_d)$, where $n_i$ are the (discrete) subpopulation levels for the different cell types. This approximation relies on the assumption that the environmental carrying capacity is large compared to individual subpopulation levels $n_i$ and results in the evolution of the system being described by a system of stochastic differential equations (SDEs) governing the time evolution of the subpopulation levels of the different cell types given by the continuous variables $x = (x_1, \ldots, x_d)$, where $d$ is the number of distinct phenotypes that may arise. We only consider phenotypes whose growth rate $g_i(I)$ at nonzero drug concentrations is higher than the baseline susceptible phenotype (wildtype) due to the negative evolutionary selection experienced by phenotypes with lower growth rates. The system evolves according to (Appendix A)

$$
\frac{dx_i(t)}{dt} = (\beta_i(I) - \delta_i) x_i(t) \left(1 - \frac{\sum_{j=1}^{d} x_j(t)}{K}\right) + \sqrt{(\beta_i(I) + \delta_i) x_i(t) \left(1 - \frac{\sum_{j=1}^{d} x_j(t)}{K}\right)} W_i(t),
$$

(2)

where $i = 1, \ldots, d$ and the white noise $W_i(t)$ satisfies

$$
\begin{align*}
\langle W_i(t) \rangle &= 0 \\
\langle W_i(t) W_i(t') \rangle &= \delta(t - t').
\end{align*}
$$

From a control standpoint, the advantage of using SDEs over an agent-based model is that we can capitalize on the equations-based description for trajectory estimation and use that information in feature engineering and reward assignment below. Note that the stochastic noise in these equations is not put in \textit{ad hoc} but instead emerges naturally from population demographic randomness that arises from demographic processes on the level

8
of single cells\textsuperscript{3}. For more realistic dynamics at very low cell levels, when the level of a subpopulation falls below 10 cells, the number of cells is discretized by a random choice (equal probability) to round up or down to the nearest integer cells after each stochastic simulation step; when cell numbers fall below 1 the subpopulation level is set to zero.

A few comments are due on the large-\(K\) approximation above (see Appendix A for details) in the context of the work here. In general, as \(x_j \to K\), this approximation ceases to provide a good description of the fluctuations in the system. In that regime, the large-\(K\) approximation leads to deterministic dynamics and neglects the effect of population fluctuations near carrying capacity, which are expected in biological systems. This is not a concern in the scenario considered here: since \(K\) is assumed here to be large, as \(x_j\) approaches \(K\) its dynamics become effectively deterministic and well-described by the logistic drift term alone; moreover, the treatment goal is to reduce the cell populations rather than maintain some distribution near carrying capacity, so that after the onset of treatment populations should spend little time in the vicinity of the resource capacity. As a result, any fluctuations at that point will have negligible effect on the progression of treatment. As the population decreases it moves farther away from the resource capacity and into the regime in which the diffusion term is a valid approximation of stochasticity, whose contribution to the dynamics also increases at lower population sizes due to the dependence on \(x_i\). The logistic deterministic dependence, however, is relevant for modeling accuracy at the initial stages of treatment (where it can affect the growth rate) as well as for training: a finite resource capacity introduces into training the issue of low growth despite population proliferation (and hence treatment failure). This needed to be addressed with a particular reward assignment scheme, discussed in Section 3.2, and was done for potential future generalization to scenarios where the population levels may stay for longer closer to the carrying capacity due to treatment constraints. Lastly, setting a carrying capacity motivated a non-arbitrary scale in the problem for feature rescaling (Appendix B).

For clarity, in this paper the term decision time step will be used to refer to the length of time between each dosing decision and thus defines the RL time step, whereas SDE simulation step will be used to refer to a stochastic simulation step. A single decision time step thus contains a large number of SDE simulation steps. The evolution of Eq. (2) was simulated via the Euler-Maruyama method with a step size of 0.01 over the 4-hour decision time step (unless all populations drop to zero before the end of the decision time step is reached).

Mutations are modeled here as random discrete external events that perturb the system of Eq. 2 by injecting new population members into a particular subpopulation \(x_i\). For simplicity, resistance is assumed to be acquired as a one-step process by mutation from wildtype; while resistance has been observed to increase with the accumulation of mutations, single mutations can suffice to confer resistance [52, 53]. During training (Section 3.2) there is a \(P_{\text{epis,mut}}\) probability that any mutations will occur during an episode, so that mutations are precluded from occurring in most episodes. This is done in order to permit the policy to identify an optimal baseline (no mutation) dosing schedule.

When mutations are allowed, they are permitted to occur at any time step up until 24 hours before the maximal time allowed for treatment, where the maximal time is set at

\textsuperscript{3}Other sources of noise due to environmental factors would in general be expected to exist as well and can be included in the model. Here we restrict to demographic stochasticity.
substantially longer than the time necessary to eliminate the cell population if no mutations were to occur. This restriction is set in place as any mutations at the very end of treatment will always necessitate an extension past the maximal time, which is already built in by setting the maximal time at longer than the necessary time. Within these bounds mutations may occur to any (or all) of a potential set of prespecified phenotypes at any decision time step with some fixed small probability $P_{\text{step,mut}}$. When a resistant population randomly emerges in this manner, its population size is assigned through a combination of random number generation and size estimation based on the expected deterministic increase in the baseline (bl) population level if it were the only present cell type:

$$N_{\text{mut,expected}} = \lceil \zeta \times \frac{x_{\text{bl}}e^{\beta_{bl}(t)r}}{1 + \frac{x_{\text{bl}}}{K}(e^{\beta_{bl}(t)r} - 1)} \rceil,$$

where $r$ is the decision time interval and $\zeta$ is generated randomly (in steps of $5 \times 10^{-6}$) from a uniform distribution between $5 \times 10^{-6}$ and $5 \times 10^{-5}$, in keeping with typical approximate observed rates of occurrence of resistant mutant cells in bacterial populations [8]). The rationale behind Eq. (3) is that larger populations will naturally give rise to more mutant cells. Rounding up is done both due to the typically low cell numbers (individual statistics modeling, as noted above), and in order to partially offset the low probability of mutation occurrence. If a mutation occurs to a particular phenotype, then $N_{\text{mut,expected}}$ is added to that phenotype’s subpopulation and, as an approximation, removed from the baseline phenotype’s subpopulation, as fewer cells are now available for its further growth. These population adjustments are done prior to simulating the stochastic evolution of the system via Eq. (2) for that decision time step.

Fig. 1 shows several model simulations for a constant low dose case that is above the susceptibility level of the initially-dominant phenotype but below that of three potential variants. All dose responses are shown in their respective colors in Fig. 2; the (simulated) parameters used are not specific to any particular organism or drug but fall within common concentration and growth rate ranges.

### 3.2 Learning an adaptive dosing policy

As discussed in Section 2.1, DDPG [24] was employed for training. Below we describe the state and action spaces and reward assignment. The neural network architecture and hyperparameter choices for the neural network and RL components are given in Appendix B.

In training, the maximal time for an episode was set at 7 days, with dosing decision steps of 4 hours. Although time scales for observations will be technology-dependent and possibly longer, this time interval was put in place in order to demonstrate the ability of the algorithm to handle episodes with many decision time steps, and hence a near-continuous dose adjustment, and produce a robust control policy able to adaptively and responsively adjust to changes in the cell population composition. To put in clinical context the maximal treatment time chosen, we note that a 2000 study [54] found that in a studied patient cohort typical ICU antibiotic courses lasted for 4-20 days, with an average length of 10 days; after 7 days of a standard antibiotic course 35% of patients in the study were found to have developed antibiotic resistance (new resistant patterns in the original pathogen) or superinfections (newly-detected pathogenic organisms not originally present).
Figure 1: Sample evolutionary simulations for four phenotypes with dose-response curves shown in Fig. 2. The horizontal and vertical axes represent, respectively, time (hours) and cell concentration. The initial state in any simulation was a population comprised entirely of the (most susceptible) baseline phenotype with a population of between $10^6$–$10^7$ (random initialization), but mutations were allowed from the first time step and could occur at any subsequent time step (4 hour intervals). The carrying capacity $K$ was set at $1.2 \times 10^7$. The dosage was set at a constant $0.5 \, \mu g/mL$ for the duration of treatment, which is higher than $\rho$ for the black phenotype but lower than $\rho$ for the blue, red, green phenotypes. Even when a phenotype is sufficiently strong to survive the dosing, random demographic fluctuations may eliminate its nascent population, as shown in several of the plots. Even when no mutations occur, variability in initial population as well as demographic randomness can lead to significantly different extinction times for the susceptible phenotype. Populations falling below 1 cell are set to zero; for plotting purposes, due to the log scale this is shown as having a population of just below 1 cell (0.98).

Figure 2: Dose-response curves for four sample (simulated) phenotypes, shown as growth rate as a function of drug concentration. The phenotype shown in black has the highest growth rate at zero drug inhibition. It is taken to be the baseline phenotype dominating the population at the onset of drug administration in the simulations (e.g. Fig. 1) that are used as the RL training data. The death rate was set at $\delta = 0.2 \, h^{-1}$ for all phenotypes.

treatment times (3 days) in that study correlated with a lower rate of antibiotic resistance (15% at 7 days). Stopping antibiotic treatment prematurely, however, runs the risk of failing to eliminate the pathogenic bacteria, suggesting that continuous monitoring and proper dose adjustments – as recommended here – are essential in combatting treatment failure in severe infections.

All code (simulation and RL) was written in Python, with the use of PyTorch [55] for the deep learning components.

3.2.1 State space: model-informed feature engineering

At each decision time step an observation of the current composition of the cell population, i.e. types and respective concentrations $x = (x_1, \ldots, x_d)$ is made. Cell types that are known
to potentially arise but are presently unobserved are assigned $x_i = 0$. As a result of model knowledge in the form of the SDE system (Eq. 2), however, additional feature combinations of $x$ can be used to supplement and improve learning. In particular, the current growth rate of the total cell population is highly indicative of drug effectiveness, but it is typically difficult to measure and necessitates taking multiple observations over some fixed time interval (e.g. 1 hour) prior to the end of a decision time step. Instead, the instantaneous growth rate can be calculated as

$$g_{\text{all}}(t) = \frac{\sum_i \dot{x}_i(t)}{\sum x_i(t)}$$  \hspace{1cm} (4)

and approximated directly from observed features, $x_t$, at time $t$ by noting that in the deterministic limit the numerator is simply given by

$$\sum_i \dot{x}_i(t) = \left(1 - \frac{\sum_i x_i}{K}\right) \times \sum_i g_i(I) x_i(t)$$  \hspace{1cm} (5)

where $g_i$ is fixed and known (Eqn. 1) during any decision time step. Hence $g_{\text{all}}$ is a function exclusively of the action taken (doses chosen) and observations of cell concentrations $x_t$.

Use of this feature combination, which would not be known in a blackbox simulation, was found to be instrumental in shaping the reward signal at non-terminal time steps and in leading to optimal policy learning, suggesting that such incorporation of information could in general assist in driving policy convergence when model knowledge is available but a model-free approach is preferred due to the complexity of the dynamics.

### 3.2.2 Action space

Training was performed with both a single drug $I_t$ and two drugs $(I^{\alpha}_t, I^{\beta}_t)$ under the same training hyperparameters (with the exception of drug preference). Drugs were set to take a continuous range of values from zero up to some maximal value $I_{\text{max}}$, and the action of drugs on the cell populations is expressed through the dose-response as in Eq. (1) and the SDE system of Eq. (2). This maximal value was used for reward scaling as shown in Section 3.2.3 rather than for placing a hard limit on the amount of drug allowed to be administered. It was set for both drugs at 8 times the highest $\rho$ value in the set $\{\rho_i, \alpha\}$ for all drugs $\alpha$. This very high value was chosen in order to allow enough initial dosing climb during training; in practice, conservative dosing was implemented through the reward assignment alone.

### 3.2.3 Reward assignment

While the state space includes information on separate cell subpopulations, the terminal (end-of-episode) reward is assigned based on only the total cell population. As noted above, a successful episode concludes in the elimination of the targeted cell population by the maximal allowed treatment time $T_{\text{max}}$. In training, different drugs are assigned preference through penalty weights $w_\alpha$, so that use of a last-line drug will incur a higher penalty than that of a first-line drug, $\alpha = 1$. To penalize for higher cumulative drug dosages, the reward
at the end of a successful episode is assigned as

\[ r_{success} = c_{end,success} \left( 1 - \sum_{\alpha} \frac{m_{\alpha}}{m_{\alpha} + 1} \sum_{n=1}^{T_{max}/\tau} I_{\alpha,n} \right), \]  

where \( I_{\alpha,n} \) is the dosage of drug \( \alpha \) administered at the \( n \)th decision time step, \( c_1, c_{end,success} > 0 \) are constants, and \( \tau \) is the length of each decision time step. When the episode fails (the cell population is above zero by \( T_{max} \)) a negative reward is assigned with an additional penalty that increases with a higher ratio of remaining cell concentrations to the initial cell population in order to guide learning:

\[ r_{failure} = -c_{end, fail} \left[ 1 + \log \left( 1 + \frac{\sum_i I_i(T_{max})}{\sum_i I_i(0)} \right) \right], \]  

where \( c_{end, fail} > 0 \). Since the episodes can be long, in order to address the credit assignment problem [23] a guiding signal was added at each time step in the form of potential-based reward shaping [56]. Policy optimality was proved to be invariant under this type of reward signal, and it can thus provide a crucial learning signal in episodic tasks. It is given by

\[ r_{shape}(s_t, a, s_{t+\tau}) = \gamma \Phi(s_{t+\tau}) - \Phi(s_t), \]  

where \( \gamma \) is the RL discount factor, \( \Phi: S \rightarrow \mathbb{R} \) is the potential function, and \( S \) is the state space. Since the drugs are assumed here to affect cell mechanisms responsible for cell birth, population growth provides a direct indicator of the efficacy of the drug. The potential function was therefore set here to

\[ \Phi = -c_{\Phi} \frac{g_{all}}{g_{max}}, \]  

where \( c_{\Phi} > 0, g_{all} \) is given by Eq. 4, and \( g_{max} \) is the zero-drug full-population deterministic growth rate at \( t = 0 \) (see also Eq. 5):

\[ g_{max} = \left( 1 - \frac{\sum_i x_i(t = 0)}{K} \right) \frac{\sum_i^{d} x_i(t = 0) g_i(I_\alpha = 0, \forall \alpha)}{\sum_i^{d} x_i(t = 0)}. \]  

Although incentives for low dosing were built into the terminal reward (Eq. 6), for guiding training toward low dosing it was necessary to provide some incentive for low dosing at each decision time step. The step reward was therefore assigned as the sum of the shaping reward and a dosage penalty:

\[ r_{step} = r_{shape} - \Theta \left( \sum_i x_i \right) \sum_\alpha w_{\alpha} \left( \frac{I_\alpha}{I_{max}} \right)^2 \]  

where \( \eta_\alpha \) is the specific penalty for each drug, and \( \Theta \left( \sum_i x_i \right) \) is a binary-valued function explained below. For a previous implementation of step-wise action penalties with potential-based reward shaping see [33].

An issue peculiar to the problem of a population growing under limited resources is that if insufficient dosing is applied and the population subsequently quickly grows and reaches
its carrying capacity, very little change in growth (or population) will occur yet the state of infection will be at its maximal worst state. However, the shaping reward (8) – which is based on changes from one decision time step to the next – will provide insufficient penalty for this (with the penalty arising solely from the $\gamma$ weighing). On the other hand, as a result of the inhibitor penalty in (10), the algorithm may continue to reduce the dosing even further rather than increasing it to escape this suboptimal policy. The result is that the targeted cell population proliferates at maximal capacity while dosing fluctuates around the minimal value of zero drug administration. This zero-dosage convergence problem was successfully resolved by assigning a significantly lower (20-fold) weight to the mid-episode dosage penalty if the total cell population exceeded the initial cell population by more than 1 cell for every initially present $10^4$ cells. The incorporation of the binary penalty function led to significantly improved stability and reduced sensitivity to the exact choice of $w_\alpha$.

### 4 Training and results

Up to four cell types with different dose responses were permitted to occur in each training scenario, with the most susceptible type initially dominating the population but with different cell populations allowed to emerge as early as the first time step. For single-drug training the dose responses shown in Fig. 2 were assumed. For combination therapy (two-drug) training, one drug was defined as the “first-line” drug and the second as the “last-line” drug via respective choices of $w_\alpha$. The responses of the black and blue phenotypes to both the first-line and last-line drugs were taken to be identical to those shown in Fig. 2. The more resistant phenotypes (red and green) were assumed to be completely resistant to the first-line drug (by assigning them a very large value of $\rho$) and to have the dose response shown in Fig. 2 to the last-resort drug.

In the single-drug case, in the early stages of training the dosages were increased rapidly and uniformly in all time steps; at that point the dosages were too high for any failures to occur. After about 180 episodes considerable decreases in dosing began to occur (also uniformly). The rate of decrease slowed down once the dosage reached the approximate level necessary to eliminate the most resistant cell subpopulations (at about 500 episodes), with dosing still uniform over the entire course of treatment. At that point increasing specialization in dosage at different time intervals in response to different mutations began to occur, and the baseline (no-mutation) dosing continued to further decrease in decision time steps following the first time step. Dosing during the first time step remained high even as subsequent doses were brought substantially lower. For these doses training gradually alternated for a long time between a gradual monotonic decrease, uniform dosing, and a gradual monotonic increase, eventually establishing the monotonic increase characteristic of the baseline dosing observed after training converged (top row of Fig. 4). Dosing during the first time step was eventually brought further down. In the early stage of training episode failures would occasionally occur; this ceased to happen later in training, at which point further improvements in learning were on dosing optimization.

In dual-drug training, similar uniform increases were observed in early training for both drugs, with the first-line drug (lower penalty) initially increasing more, but with this increase reversed after about 100 episodes (recall that lower dosages of the first-line drug are needed to suppress the mutations that are not completely resistant to it). By $\sim 220$ episodes into
Figure 3: MSE loss during training for the training with one drug and two drugs. Training was generally smoother and the MSE loss flattened somewhat earlier in the one-drug case. A low-pass filter was applied to obtain the black curves. Mutation parameters for the training shown were $P_{\text{epis},\text{mut}} = 0.3$, $P_{\text{epis},\text{mut}} = 0.03$, $\zeta = 5 \times 10^{-6}$, and $w_1 = 16$, $w_2 = 3w_1$ for the first-line and last-line drugs respectively.

Training the first-line drug was not administered at all - but treatments were still successful since the last-line drug is capable of suppressing all cell types, albeit at a higher penalty. The last-line dosage continued to be uniformly decreased until treatment failures started occurring and was subsequently increased to the level at which it could suppress mutant populations without specialization. Decreases again began occurring at that point, and the policy also started applying the first-line drug again. Training from that point on involved specialization in response to mutation and experimentation with various amounts of the first-line and last-line drugs.

Training was stopped after no further reduction was observed in the MSE loss and the actor policy loss had stabilized. The MSE loss flattened at $\sim O(10^{-3}) - O(10^{-2})$, and this was observed to occur in single-drug training after a somewhat smaller number of training episodes than in dual-drug training for the same parameters (Fig. 3). The policy parameters were saved every 100 episodes for later reference and the resulting policies were analyzed after training. Following training convergence (determined as described above) dosing was generally similar for policy parameters recorded at different episodes, with some minor fluctuations observed. In the dual-drug case, when no mutations were experienced, the learned policies either administered no last-line drug at all or negligibly small amounts of it only in the first and last time step (Fig 5).

Responsive adaptation to random events and system perturbations

A hallmark of both the single-drug and combination therapy policies is the responsive adaptation to random population composition changes (Fig. 4 and Fig. 5) insofar as both the appearance of a new cell type and stochastic fluctuations of already-present cell types. On a test set of 1000 episodes simulated randomly with training parameters a 100% success rate was obtained in both the single-drug and dual-drug cases (all episodes recorded involved the occurrence of mutations). Multiple concurrent subpopulations are handled well by the
Itot = 153.8
0
2
4
6
8
10
12
Tend = 88
0 20 40 60 80 100 120 140
100
101
102
103
104
105
106
107

Itot = 263.5
0
2
4
6
8
10
12
Tend = 72
0 20 40 60 80 100 120 140
100
101
102
103
104
105
106
107

Itot = 170.8
0
2
4
6
8
10
12
Tend = 80
0 20 40 60 80 100 120 140
100
101
102
103
104
105
106
107

Itot = 167.5
0
2
4
6
8
10
12
Tend = 80
0 20 40 60 80 100 120 140
100
101
102
103
104
105
106
107

Itot = 625.4
0
2
4
6
8
10
12
Tend = 84
0 20 40 60 80 100 120 140
100
101
102
103
104
105
106
107

Itot = 327.8
0
2
4
6
8
10
12
Tend = 80
0 20 40 60 80 100 120 140
100
101
102
103
104
105
106
107

Itot = 290.9
0
2
4
6
8
10
12
Tend = 76
0 20 40 60 80 100 120 140
100
101
102
103
104
105
106
107

Itot = 171.6
0
2
4
6
8
10
12
Tend = 92
0 20 40 60 80 100 120 140
100
101
102
103
104
105
106
107

Figure 4: Examples of simulation results for the single-drug policy on a test set that included both training mutation parameters ($P_{\text{step.mut}} = 0.03$, $\zeta = 5 \times 10^{-6}$), elevated emergent cell numbers ($\zeta = 10^{-4}$), and elevated probability of mutation in any given time step ($P_{\text{step.mut}} = 0.1$). In each run the top plot shows drug concentration (µg/mL) as a function of time (hours) and the bottom plot shows the cell concentration as a function of time, with colors corresponding to the dose-response curves shown on the top right. In all cases the dosing was increased adaptively over the baseline only when necessary and to the extents compatible with the dose responses shown in the plot on the top right (note that a green population can merit higher dosing than a red one if it is observed at sufficiently high concentrations). The baseline population (black) was initialized randomly between $10^6$ and $10^7$ in each simulation with the remaining populations initialized at zero but resistant subpopulations were allowed to emerge as early as the first time step. The carrying capacity $K$ was set at $1.2 \times 10^7$. Policy parameters used in these simulations were obtained after 95,000 training episodes from the training shown in Fig 3a.
Figure 5: Examples of simulation results for the combination therapy policy on a test set that included both training mutation parameters, elevated emergent cell numbers, and an elevated probability of mutation $P_{\text{step,mut}}$. The gray designates the first-line drug and the magenta designates the last-line drug. In each run the top plot shows drug concentration ($\mu$g/mL) as a function of time (hours) and the bottom plot shows the cell concentration as a function of time. The blue and black phenotypes exhibit the dose-response shown on the top right of Fig. 4 to both drugs. The green and red were assigned a $p$ value of 10,000 (completely resistant) to the gray drug and the same dose-response as shown on the top right of Fig. 4 to the magenta drug. Both dosing adjustments of a given drug as well as drug switching took place as appropriate. Since baseline dosing is lower in the combination therapy case, the appearance of a blue subpopulation necessitated mild increases in dosage. Policy parameters used in these simulations were obtained after 95,000 training episodes from the training shown in Fig 3b.
policies learned. In single-drug therapy, higher drug pulses are only administered after observing an emergent resistant subpopulation, and baseline dosing is restored following the elimination of this subpopulation (Fig 4). The dosage administered in these pulses directly increases with the resistance and population of the observed cell type. In combination therapy (Fig 5), the policy switches to the last-line drug when the red and green phenotypes are observed and switches back to the first-line drug after these cells are eliminated. For the low-resistance blue phenotype appropriate dosing increases are done exclusively with the first-line drug.

This is particularly remarkable given that the black and blue phenotypes are susceptible to the same extent to both the first- and last-line drugs. The sharp difference in drug administration is thus attributable to the differing incentives supplied in the reward assignment ($w_2 = 3w_1$).

**Robustness to changes in mutation parameters**

In a given episode, the rate of resistant subpopulation emergence and the extent of it were controlled by $P_{\text{step,mut}}$ and $\zeta$, respectively. To assess the robustness of the policy to variations in these parameters the success rates of policies shown in Fig. 4 (one drug) and Fig. 5 (two drugs) were tested over parameter combinations (100 trials for each combination) that significantly exceeded the training parameters. No deterioration in performance was found for parameter combinations within and substantially above the expected biological values: 100% success was recorded for testing of up to a 10-fold increase in $P_{\text{step,mut}}$ and more than a 1000-fold increase in $\zeta$, which controls the number of resistant mutants observed. No failures were recorded for $\zeta$ of up to 1% of the expected increase in susceptible cells in any time interval (recall that the actual number of cells is randomly generated according to Eq. 3 with respect to $\zeta$). Slight performance deterioration was found at the very high end of the $\zeta$ range tested ($\zeta = 5\%$), but success rates of $\geq 99\%$ were still recorded for these parameter combinations; we note that this high end of the range is in significant excess of resistant cell proportions that may be naturally expected [8].

Generally, the robustness of the learned policies to such strong parameter deviations implies that it may be possible to apply them to cases where the detection threshold for newly emergent populations is higher due to lower-resolution diagnostics. However, modifications accounting for the resolution at subsequent time steps will likely be needed in such cases.

**Policy behavior in the administration of baseline dosing**

Typical single-drug baseline (no-mutation episode) dosing exhibits slightly elevated dosages at the start of each episode that then mildly drop. Subsequent decision time steps involve approximately-constant but mildly monotonically increasing dosages. An episode can terminate at various points in time even in the absence of mutations (top rows of Figs. 4 and 5) due to variations in the initial cell population size of up to an order of magnitude as well

---

4Tests were performed with the policy parameters obtained after 95,000 episodes of training for both the single-drug and dual drug policies, shown in Fig. 3.
as demographic fluctuations at low populations toward the end of treatment. The dosing is seen to adjust in accordance with these variations.

We can gain insight into the reason for this mild monotonic increase by comparing the single-drug dosing with the dual-drug dosing policies, where we can isolate the effects of two of the mutations due to the differing responses to the two drugs. In the combination therapy case, after the first time step the first-line drug is administered at a nearly-constant value with no increase in dosing as the treatment proceeds. On the other hand, as noted above, occasional minor last-line drug dosages at the very beginning and end of non-mutation episodes can be observed in some runs\(^5\). One explanation for these dosages is that they may arise from learning that if resistant phenotypes emerge at the beginning or the very end of treatment then an immediate response via last-line dosing is paramount to achieving an optimal treatment not involving higher-than-necessary dosing. In the single-drug case a dosage increase can combat these mutations, and the observed monotonic rise may thus emerge from the policy’s anticipation of more resistant phenotypes. It is therefore appears plausible that the anticipation of late-stage higher-resistance mutations drives this increase rather than any aspects of the susceptible cell population growth.

**Learned preference for short treatment times**

All explicit optimization incentives that were given involved the dosage administered rather than the time of treatment (Section 3.2). As previously explained, the main challenge in the dosing problem is the balancing of the higher dosages needed to suppress resistant cell populations with the need to keep toxicity low. Rewards in the setup implemented here purposely did not incorporate any direct incentives for shorter treatment times, focusing instead on dosage minimization. Although conservative dosing was a strong feature of all learned policies, the policies learned in training involved a strong preference for shorter treatment times even at the expense of higher dosing, typically eliminating the cell population well in advance of the maximal treatment time. This is particularly evident from a comparison of the policy response in its dosing vs. its treatment time (Figs. 6a and 6b) to increases in the mutation rate per step \((P_{\text{step,mut}})\) and resistant cell supply via \(\zeta\) (Eq. 3): the policy compensates for the increased resistance almost exclusively through dosage increases, while treatment time is kept nearly constant. This behavior may arise through a combination of learned preferences based on dosing penalties and mutation parameters, and in training for a real clinical scenario a “clinician-in-the-loop” approach would permit informed choices in this regard.

5 Discussion

This paper presented a method for single-drug and combination therapy feedback control that results in policies capable of responsive and robust adaptation to changes in a stochastic dynamical system of evolving cell populations. Model-free deep reinforcement learning was supplemented with model information on system trajectory estimation into feature engineering and reward assignment, which was found to significantly improve learning. Various

\(^5\)At the beginning of the treatment this was only observed to coincide with a slight dip in the first-line dosage.
Figure 6: Policy preference for early termination times can be seen by analyzing the single-drug dosing behavior (Fig. 4) at different \((P_{\text{step,mut}}, \zeta)\) parameter combinations for episodes in which mutations occurred. (a): when \(\zeta\) is increased in testing over the training value (red dot), the mean total dosage in an episode (purple) increases significantly, but the time of the episode (orange) remains near-constant, exhibiting a strong policy preference for maintaining short treatment times at the cost of higher baseline dosage. (b): a preference for dosage rather than time increases is also observed for increases in \(P_{\text{step,mut}}\).

aspects of the resulting policies were investigated in this study and may help guide future efforts training similar systems in which some trajectory estimation is possible, the probability of random perturbing events may not be known, and a tradeoff between conflicting target goals (here: low toxicity with the need to properly target resistant cell populations) exists.

Given the demonstrated ability of the policies that were learned in robustly adjusting to large variations in certain parameters on which they were not trained, an interesting question for future exploration is how accurately the underlying structure of the model – here logistic growth dynamics – needs to be known for policy training. Higher flexibility in this regard would enable controller development with more limited predictive knowledge of the dynamics. The resolution of feedback observations and especially their frequency is likely to play an important role in this flexibility. Here, observations were assumed to be generated at discrete time points but be reflective of the population composition at these time points. Practical implementation would likely necessitate accounting for a certain time delay in obtaining such observations, and future directions for the development of CelluDose include the incorporation of delayed feedback and lower-resolution observations that lead to partial observability of the system.

As higher-resolution diagnostics, more sophisticated modeling techniques, and a recognition of the need for patient-specific care are making increasing gains in healthcare, the need for appropriate therapeutic control methods is likely to rise. The ultimate goal in precision dosing – a feedback loop in which drug absorption and response are tightly monitored and used to determine subsequent dosing – remains for now an elusive goal in all but a few cases [57]. The method presented here is intended for eventual use in this context under consistent monitoring of the targeted cell population composition in terms of the variability in its dose responses.

An important advantage of modeling and simulation in developing drug dosage controllers is the ability to investigate a large range of drug schedules even prior to clinical
trials and to thus potentially inform the development of clinical trials. The focus of the implementation presented here is on the drug responses and trajectories of an evolving population of cells. For clinical applications, modeling would need to additionally include patient-relevant models of drug absorption and immune system response. With appropriate diagnostics, the setup described here could be implemented for the automated control of laboratory bacterial evolution experiments and establish the roles and interplay of decision time intervals, diagnostic resolution, and model uncertainty as part of the controller design. This would be a crucial step toward future clinical application and a validation of the utility of artificial intelligence-driven approaches in simulation-based biomedical controller design.

Acknowledgements

I am grateful to Tommaso Biancalani, Roger Brockett, and Finale Doshi-Velez for useful discussions. This work was supported by NIH grant 5R01GM124044-02 to Eugene Shakhnovich.

A Master equation to stochastic differential equations

A.1 Master equation for the population dynamics of $d$ phenotypes

Let $A_i$ denote a single cell of phenotype $i$. We consider $d$ phenotype with equal resource utilization evolving subject to a resource capacity given by $K$. Following a simplified setup of [58], we initially consider a spatial grid permitting up to one individual of type $A_i$ per site. If empty, individual in site is denoted by $E$. The rates of birth and death of cells of type $i$ are denoted by $\tilde{\beta}_i$ and $\tilde{\delta}_i$ respectively:

Birth: $A_iE \xrightarrow{\tilde{\beta}_i} A_iA_i$

Death: $A_i \xrightarrow{\tilde{\delta}_i} E$

At each time step we sample the population. On a fraction $\eta$ of these events we randomly choose two individuals and allow them to interact, but since we do not consider direct competition effects other than through shared limited resources, we ignore combinations in which both $A_i$ cells were picked and therefore restrict to picking only the combination $A_i$ and $E$. In a fraction $1 - \eta$ of these events we choose only one individual randomly (if only $E$’s are drawn, the drawing is done again). The probabilities of picking the different combinations are given by

$$P(A_iE) : 2\eta \frac{n_i}{K} \left( \frac{K - \sum_k n_k}{K - 1} \right)$$

$$P(A_i) : (1 - \eta) \frac{n_i}{K}$$

where $n_i$ is the population of phenotype $i$. We will denote the transition probabilities from state $\mathbf{n} = (n_1, ..., n_d)$ to state $\mathbf{n}'$ by $T(\mathbf{n}'|\mathbf{n})$. They are thus given by

$$T(n_1, ..., n_i + 1, ..., n_d|\mathbf{n}) = 2\eta \tilde{\beta}_i \frac{n_i}{K} \left( \frac{K - \sum_k n_k}{K - 1} \right) \approx 2\eta \tilde{\beta}_i \frac{n_i(K - \sum_k n_k)}{K^2}$$

$$T(n_1, ..., n_i - 1, ..., n_d|\mathbf{n}) = (1 - \eta) \tilde{\delta}_i \frac{n_i}{K}$$
where for simplicity we replace \( K - 1 \rightarrow K \) as we will be considering the large-\( K \) limit for the subsequent diffusion approximation. The master equation is given by

\[
\frac{dP(n, t)}{dt} = \sum_{i=1}^{d} \left\{ T(n|i_1, \ldots, n_i + 1, \ldots, n_d)P(n_1, \ldots, n_{i+1}, \ldots, n_d, t)
+ T(n|i_1, \ldots, n_i - 1, \ldots, n_d)P(n_1, \ldots, n_{i-1}, \ldots, n_d, t)
- [T(n_1, \ldots, n_i - 1, \ldots, n_d|n) + T(n_1, \ldots, n_i + 1, \ldots, n_d|n)] P(n, t) \right\}
\]

(11)

subject to the initial condition \( P(n, 0) = \delta_{n0,n} \) (Kronecker delta, not to be confused with the death rates \( \delta_i \)). We must also impose boundary conditions,

\[
\begin{align*}
&T(n_1, \ldots, n_i = 0, \ldots, n_d|n_1, \ldots, n_i = -1, \ldots, n_d) = 0 \\
&T(\sum_i n_i = K|\sum_i n_i = K + 1) = 0.
\end{align*}
\]

A.2 Diffusion approximation: the Fokker-Planck equation and Ito stochastic differential equations

By assuming that the resource capacity \( K \) is large compared to the population levels \( n_i \), we can approximate \( \varphi_i = n_i/K \ll 1 \) as continuous variables; we denote \( P(n_1, \ldots, n_i \pm 1, \ldots, n_d, t) \equiv P_\varphi(\varphi_i \pm 1) \) and define

\[
\begin{align*}
f(\varphi_i) &\equiv \delta_i \varphi_i \\
g(\varphi_i) &\equiv \beta_i \varphi_i \left( 1 - \sum_j \varphi_j \right)
\end{align*}
\]

where we have rescaled the death and birth rates as \( \delta_i \equiv (1 - \eta) \tilde{\delta}_i \) and \( \beta_i \equiv \eta \tilde{\beta}_i \). Eqn. (11) can then be rewritten as

\[
\frac{dP_\varphi(\vec{\varphi}, t)}{dt} = \sum_{i=1}^{d} \left\{ \left[ f(\varphi_i + \frac{1}{K}) P_\varphi \left( \varphi_1, \ldots, \varphi_i + \frac{1}{K}, \ldots, \varphi_d, t \right) - f(\varphi_i) P_\varphi(\vec{\varphi}, t) \right]
+ \left[ g(\varphi_i - \frac{1}{K}) P_\varphi \left( \varphi_1, \ldots, \varphi_i - \frac{1}{K}, \ldots, \varphi_d, t \right) - g(\varphi_i) P_\varphi(\vec{\varphi}, t) \right] \right\}
\]

Taylor expanding around \( \varphi_i \) to second order [51] we have that

\[
h \left( \varphi_i \pm \frac{1}{K} \right) P_\varphi \left( \varphi_1, \ldots, \varphi_i \pm \frac{1}{K}, \ldots, \varphi_d, t \right) = h(\varphi_i) P_\varphi(\vec{\varphi}, t) \pm \frac{1}{K} \frac{\partial(h P)}{\partial \varphi_i} \bigg|_{\varphi} + \frac{1}{2K^2} \sum_j \frac{\partial^2(h P)}{\partial \varphi_j^2} \bigg|_{\varphi},
\]

which yields, after rescaling time as \( \tau = t/K \), the Fokker-Planck equation

\[
\frac{\partial P_\varphi(\vec{\varphi}, t)}{\partial \tau} = \sum_{i=1}^{d} \left[ \frac{\partial((f - g) P)}{\partial \varphi_i} \bigg|_{\varphi} + \frac{1}{2K} \sum_j \frac{\partial^2((f + g) P)}{\partial \varphi_j^2} \bigg|_{\varphi} \right],
\]

which corresponds to the system of Ito stochastic differential equations

\[
d\varphi_i = -(f(\varphi_i) - g(\varphi_i)) \, dt + \frac{1}{\sqrt{K}} \sum_m \delta_{mi} \sqrt{f(\varphi_m) + g(\varphi_m)} dW_m(t)
\]
where the white noise \( W_k(t) \) satisfies

\[
\begin{cases}
\langle W_k(t) \rangle = 0 \\
\langle W_k(t) W_l(t') \rangle = \delta_{kl} \delta(t - t')
\end{cases}
\]

and \( \delta_{mi} \) is the Kronecker delta (not to be confused with the death rate \( \delta_i \)). We have that

\[
f (\varphi_i) \pm g (\varphi_i) = \left( 1 - \sum_j^d \varphi_j \right) \delta_i \left( \frac{\varphi_i}{1 - \sum_j^d \varphi_j} - \delta_i \varphi_i + \delta_i \varphi_i \pm \beta_i \varphi_i \right)
\]

\[
= \varphi_i \left( 1 - \sum_j^d \varphi_j \right) \left( \delta_i \left( \frac{\sum_j^d \varphi_j}{1 - \sum_j^d \varphi_j} \right) + \delta_i \pm \beta_i \right).
\]

If we redefine \( x_i \equiv K \varphi_i \) as the levels (concentrations) rather than proportions of the populations, we observe that the fraction

\[
\frac{\sum_j^d \varphi_j}{1 - \sum_j^d \varphi_j} = \frac{1}{K} \frac{\sum_j^d x_j}{1 - \frac{1}{K} \sum_j^d x_j}
\]

approaches 1 in the large-\( K \) limit, in which case we have

\[
f (\varphi_i) \pm g (\varphi_i) \to \frac{1}{K} (\delta_i \pm \beta_i) x_i \left( 1 - \frac{\sum_j^d x_j}{K} \right)
\]

yielding the system (2)

\[
\frac{dx_i(t)}{dt} = (\beta_i - \delta_i) x_i(t) \left( 1 - \frac{\sum_{j=1}^d x_j(t)}{K} \right) + \sqrt{\left( (\beta_i + \delta_i) x_i(t) \left( 1 - \frac{\sum_{j=1}^d x_j(t)}{K} \right) \right) W_i(t)}
\]

for \( i = 1, \ldots, d \). To prevent the term in the square root of the noise from occasionally briefly dipping below zero and resulting in imaginary numbers (due to the cell subpopulations potentially dipping below zero during a simulation step before being set to zero), this term is clipped at zero.

**B Neural network architecture and hyperparameter choices**

The actor and critic network (and respective target network) architecture used is the same as in the original DDPG paper [24], with the notable difference that significant performance improvement was obtained with the addition of an extra hidden layer (45 units) in the actor network\(^6\) and that narrower hidden layers (60 and 45 units respectively) were implemented to avoid overfitting. All hidden layers employed rectified linear activation functions; the output layer of the actor was set to a hard tanh with range between 0 and a chosen maximum dose in order to allow for the dosage to drop to exactly zero (particularly important when

\(^6\)Notably, the addition of an extra layer in the critic network led to degraded performance.
multiple drugs are considered). Weights were initialized randomly but biases were set to a positive value informed by the evolution of the deterministic part of (2) (see Appendix C for a detailed description). This was found to prevent convergence to a suboptimal policy in which the minimal control (0) was applied at every time step early in the training.

No batch normalization was used at hidden layers, but features were rescaled prior to being added to the replay buffer in the following manner. Cell concentrations \( x_i \) were rescaled as

\[
x_i \rightarrow \frac{\log(x_i + 1)}{\log(K)}.
\]

Since cell concentrations can vary by multiple orders of magnitude, this was found to be necessary for smooth convergence. The growth rate \( g_{all} \) was rescaled as

\[
g_{all} \rightarrow \frac{g_{all} - g_{min}}{g_{max} - g_{min}}
\]

where \( g_{max} \) was defined in (9) and \( g_{min} \) was set to the negative of the death rate, \(-\delta\), which is the lowest growth rate (\( \delta \) is the highest rate of decline) possible in the system.

Adam optimization was used with learning rates of \( 10^{-5} \) for the actor network and \( 10^{-4} \) for the critic network as learning was found to be unstable for higher rates. Training was done with mini-batch size of 128 and a replay buffer of size \( 10^6 \). Soft updates to the target networks were done with \( \tau = 0.001 \) and the RL discount factor was set at \( \gamma = 0.99 \). Exploration was done with an Ornstein-Uhlenbeck process with \( \theta = 0.15 \), \( \sigma = 0.3 \), and \( \delta t = 10^{-2} \). Parameters used in the reward choice that resulted in successful training were \( c_{end, success} = 40 \), \( w_1 = w_2 = 1 \), \( c_1 = 2 \), \( c_{end, fail} = 20 \), \( c_\Phi = 10 \). Single-drug training was done with both \( \eta_1 = 16 \) and \( \eta_1 = 4 \). Combination therapy (two drugs) training was done with \( \eta_2 = 3 \times \eta_1 \) under the assumption that the drug able to effectively target the higher-resistance cells is a last-resort drug that also involves higher toxicity.

### C Bias initialization

If we assume the lowest-variance policy – constant uniform dosing – and permit no mutations to occur, then in deterministic evolution the number of baseline-phenotype cells under constant dose \( I \) starting from a population \( x_0 \) will be given at time \( t \) by the solution of the noise-free \( d = 1 \) system equivalent of (2):

\[
x(t) = \frac{x_0 e^{g(I)t}}{1 + \frac{x_0}{K} [e^{g(I)t} - 1]}.
\]

(13)

where \( g(I) = \frac{\beta}{1+\rho} - \delta \). If the population must be reduced to \( x_{min} \) (here, < 1 cell) within a time \( T_{max} \), then the dosage must be no less than

\[
I_{min} = \rho \left[ \frac{\beta T_{max}}{\delta T_{max} - \log \left( \frac{x_0(K-x_{min})}{x_{min}(K-x_0)} \right)} \right]
\]

for treatment to be successful. Initial cell populations were allowed to vary by up to an order of magnitude; by taking the maximum value allowed in this range we compute \( I_{min} \) and set...
the bias to five times this value or to 75% of the maximal inhibit - whichever is smallest (typically $5I_{\text{min}}$). $x_{\text{min}}$ was set to just under 1 cell (0.98) as the population falling below a single cell marked the conclusion of a successful episode. The full initialization for both actor and critic networks were saved and used in subsequent runs for training consistency.

References

[1] C. Holohan, S. Van Schaeybroeck, D. B. Longley, and P. G. Johnston, “Cancer drug resistance: an evolving paradigm,” *Nature Reviews Cancer* **13** no. 10, (2013) 714.

[2] N. Brusselaers, D. Vogelaers, and S. Blot, “The rising problem of antimicrobial resistance in the intensive care unit,” *Annals of intensive care* **1** no. 1, (2011) 47.

[3] J. M. Munita and C. A. Arias, “Mechanisms of antibiotic resistance,” *Microbiology spectrum* **4** no. 2, (2016).

[4] Z. Bódi, Z. Farkas, D. Nevozhay, D. Kalapis, V. Lázár, B. Csörgő, Á. Nyerges, B. Szamecz, G. Fekete, B. Papp, et al., “Phenotypic heterogeneity promotes adaptive evolution,” *PLoS biology* **15** no. 5, (2017) e2000644.

[5] D. L. Dexter and J. T. Leith, “Tumor heterogeneity and drug resistance.,” *Journal of clinical oncology* **4** no. 2, (1986) 244–257.

[6] C. Swanton, “Intratumor heterogeneity: evolution through space and time,” *Cancer research* **72** no. 19, (2012) 4875–4882.

[7] I. Dagogo-Jack and A. T. Shaw, “Tumour heterogeneity and resistance to cancer therapies,” *Nature reviews Clinical oncology* **15** no. 2, (2018) 81.

[8] M. Falagas, G. Makris, G. Dimopoulos, and D. Matthaiou, “Heteroresistance: a concern of increasing clinical significance?,” 2008.

[9] V. I. Band, S. W. Satola, E. M. Burd, M. M. Farley, J. T. Jacob, and D. S. Weiss, “Carbapenem-resistant klebsiella pneumoniae exhibiting clinically undetected colistin heteroresistance leads to treatment failure in a murine model of infection,” *Mbio* **9** no. 2, (2018) e02448–17.

[10] E. Toprak, A. Veres, J.-B. Michel, R. Chait, D. L. Hartl, and R. Kishony, “Evolutionary paths to antibiotic resistance under dynamically sustained drug selection,” *Nature genetics* **44** no. 1, (2012) 101.

[11] S. Suzuki, T. Horinouchi, and C. Furusawa, “Prediction of antibiotic resistance by gene expression profiles,” *Nature communications* **5** (2014) 5792.

[12] A. F. Read and R. J. Woods, “Antibiotic resistance management,” *Evolution, medicine, and public health* **2014** no. 1, (2014) 147.

[13] V. Knight-Schrijver, V. Chelliah, L. Cucurull-Sanchez, and N. Le Novère, “The promises of quantitative systems pharmacology modelling for drug development,” *Computational and structural biotechnology journal* **14** (2016) 363–370.
[14] M. Danhof, “Systems pharmacology–towards the modeling of network interactions,” *European Journal of Pharmaceutical Sciences* **94** (2016) 4–14.

[15] T. M. Polasek, S. Shakib, and A. Rostami-Hodjegan, “Precision dosing in clinical medicine: present and future,” 2018.

[16] T. M. Polasek, C. R. Rayner, R. W. Peck, A. Rowland, H. Kimko, and A. Rostami-Hodjegan, “Toward dynamic prescribing information: Codevelopment of companion model-informed precision dosing tools in drug development,” *Clinical pharmacology in drug development* (2018).

[17] T. M. Polasek, A. Rostami-Hodjegan, D.-S. Yim, M. Jamei, H. Lee, H. Kimko, J. K. Kim, P. T. T. Nguyen, A. S. Darwich, and J.-G. Shin, “What does it take to make model-informed precision dosing common practice? report from the 1st asian symposium on precision dosing,” 2019.

[18] R. J. Keizer, R. ter Heine, A. Frymoyer, L. J. Lesko, R. Mangat, and S. Goswami, “Model-informed precision dosing at the bedside: Scientific challenges and opportunities,” *CPT: Pharmacometrics & Systems Pharmacology* (2018).

[19] A. Darwich, K. Ogungbenro, A. Vinks, J. Powell, J.-L. Reny, N. Marsousi, Y. Daali, D. Fairman, J. Cook, L. Lesko, et al., “Why has model-informed precision dosing not yet become common clinical reality? lessons from the past and a roadmap for the future,” *Clinical Pharmacology & Therapeutics* **101** no. 5, (2017) 646–656.

[20] R. S. Parker and F. J. Doyle III, “Control-relevant modeling in drug delivery,” *Advanced Drug Delivery Reviews* **48** no. 2-3, (2001) 211–228.

[21] B. Recht, “A tour of reinforcement learning: The view from continuous control,” *Annual Review of Control, Robotics, and Autonomous Systems* (2018).

[22] B. Kiumarsi, K. G. Vamvoudakis, H. Modares, and F. L. Lewis, “Optimal and autonomous control using reinforcement learning: A survey,” *IEEE Transactions on Neural Networks and Learning Systems* **29** no. 6, (2018) 2042–2062.

[23] M. Minsky, “Steps toward artificial intelligence,” *Proceedings of the IRE* **49** no. 1, (1961) 8–30.

[24] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, “Continuous control with deep reinforcement learning,” *arXiv preprint arXiv:1509.02971* (2015).

[25] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, “Deterministic policy gradient algorithms,” in *ICML*. 2014.

[26] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, “Playing atari with deep reinforcement learning,” *arXiv preprint arXiv:1312.5602* (2013).
[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* 518 no. 7540, (2015) 529.

[28] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[29] B. L. Moore, L. D. Pyeatt, V. Kulkarni, P. Panousis, K. Padrez, and A. G. Doufas, “Reinforcement learning for closed-loop propofol anesthesia: a study in human volunteers,” *The Journal of Machine Learning Research* 15 no. 1, (2014) 655–696.

[30] R. Padmanabhan, N. Meskin, and W. M. Haddad, “Reinforcement learning-based control of drug dosing for cancer chemotherapy treatment,” *Mathematical biosciences* 293 (2017) 11–20.

[31] I. Ahn and J. Park, “Drug scheduling of cancer chemotherapy based on natural actor-critic approach,” *BioSystems* 106 no. 2-3, (2011) 121–129.

[32] J. Peters and S. Schaal, “Natural actor-critic,” *Neurocomputing* 71 no. 7-9, (2008) 1180–1190.

[33] B. K. Petersen, J. Yang, W. S. Grathwohl, C. Cockrell, C. Santiago, G. An, and D. M. Faissol, “Precision medicine as a control problem: Using simulation and deep reinforcement learning to discover adaptive, personalized multi-cytokine therapy for sepsis,” *arXiv preprint arXiv:1802.10440* (2018).

[34] A. Fischer, I. Vázquez-García, and V. Mustonen, “The value of monitoring to control evolving populations,” *Proceedings of the National Academy of Sciences* 112 no. 4, (2015) 1007–1012.

[35] P. Newton and Y. Ma, “Nonlinear adaptive control of competitive release and chemotherapeutic resistance,” *Physical Review E* 99 no. 2, (2019) 022404.

[36] J. West, L. You, J. Brown, P. K. Newton, and A. R. Anderson, “Towards multi-drug adaptive therapy,” *bioRxiv* (2018) 476507.

[37] D. J. Roach, J. N. Burton, C. Lee, B. Stackhouse, S. M. Butler-Wu, B. T. Cookson, J. Shendure, and S. J. Salipante, “A year of infection in the intensive care unit: prospective whole genome sequencing of bacterial clinical isolates reveals cryptic transmissions and novel microbiota,” *PLoS genetics* 11 no. 7, (2015) e1005413.

[38] B. Van den Bergh, T. Swings, M. Fauvart, and J. Michiels, “Experimental design, population dynamics, and diversity in microbial experimental evolution,” *Microbiol. Mol. Biol. Rev.* 82 no. 3, (2018) e00008–18.

[39] C. U. Köser, M. J. Ellington, and S. J. Peacock, “Whole-genome sequencing to control antimicrobial resistance,” *Trends in Genetics* 30 no. 9, (2014) 401–407.

[40] O. M. El-Halfawy and M. A. Valvano, “Antimicrobial heteroresistance: an emerging field in need of clarity,” *Clinical microbiology reviews* 28 no. 1, (2015) 191–207.
[41] E. Ruppé, A. Cherkaoui, V. Lazarevic, S. Emonet, and J. Schrenzel, “Establishing genotype-to-phenotype relationships in bacteria causing hospital-acquired pneumonia: a prelude to the application of clinical metagenomics,” *Antibiotics* 6 no. 4, (2017) 30.

[42] D. A. Lawson, K. Kessenbrock, R. T. Davis, N. Pervolarakis, and Z. Werb, “Tumour heterogeneity and metastasis at single-cell resolution,” *Nature cell biology* 20 no. 12, (2018) 1349.

[43] K. Rosenthal, V. Oehling, C. Dusny, and A. Schmid, “Beyond the bulk: disclosing the life of single microbial cells,” *FEMS microbiology reviews* 41 no. 6, (2017) 751–780.

[44] S. O. Kelley, “New technologies for rapid bacterial identification and antibiotic resistance profiling,” *SLAS TECHNOLOGY: Translating Life Sciences Innovation* 22 no. 2, (2017) 113–121.

[45] F. Lyu, M. Pan, S. Patil, J.-H. Wang, A. Matin, J. R. Andrews, and S. K. Tang, “Phenotyping antibiotic resistance with single-cell resolution for the detection of heteroresistance,” *Sensors and Actuators B: Chemical* 270 (2018) 396–404.

[46] B. Hwang, J. H. Lee, and D. Bang, “Single-cell rna sequencing technologies and bioinformatics pipelines,” *Experimental & molecular medicine* 50 no. 8, (2018) 96.

[47] A. V. Hill, “The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves,” *j. physiol.* 40 (1910) 4–7.

[48] T.-C. Chou, “Derivation and properties of michaelis-menten type and hill type equations for reference ligands,” *Journal of theoretical biology* 59 no. 2, (1976) 253–276.

[49] N. G. Van Kampen, *Stochastic processes in physics and chemistry*, vol. 1. Elsevier, 1992.

[50] C. W. Gardiner, “Handbook of stochastic methods for physics, chemistry and the natural sciences,” *Applied Optics* 25 (1986) 3145.

[51] A. J. McKane, T. Biancalani, and T. Rogers, “Stochastic pattern formation and spontaneous polarisation: the linear noise approximation and beyond,” *Bulletin of mathematical biology* 76 no. 4, (2014) 895–921.

[52] A. C. Palmer, E. Toprak, M. Baym, S. Kim, A. Veres, S. Bershtein, and R. Kishony, “Delayed commitment to evolutionary fate in antibiotic resistance fitness landscapes,” *Nature communications* 6 (2015) 7385.

[53] J. V. Rodrigues, S. Bershtein, A. Li, E. R. Lozovsky, D. L. Hartl, and E. I. Shakhnovich, “Biophysical principles predict fitness landscapes of drug resistance,” *Proceedings of the National Academy of Sciences* 113 no. 11, (2016) E1470–E1478.
[54] N. Singh, P. Rogers, C. W. Atwood, M. M. Wagener, and V. L. Yu, “Short-course empiric antibiotic therapy for patients with pulmonary infiltrates in the intensive care unit: a proposed solution for indiscriminate antibiotic prescription,” *American journal of respiratory and critical care medicine* **162** no. 2, (2000) 505–511.

[55] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in pytorch.”

[56] A. Y. Ng, D. Harada, and S. Russell, “Policy invariance under reward transformations: Theory and application to reward shaping,” in *ICML*, vol. 99, pp. 278–287. 1999.

[57] G. T. Tucker, “Personalized drug dosage—closing the loop,” *Pharmaceutical research* **34** no. 8, (2017) 1539–1543.

[58] A. J. McKane and T. J. Newman, “Stochastic models in population biology and their deterministic analogs,” *Physical Review E* **70** no. 4, (2004) 041902.