PNAS RESEARCH ARTICLE | MICROBIOLOGY OPEN ACCESS

Reinforcement learning informs optimal treatment strategies to limit antibiotic resistance

Davis T. Weaver[a,b], Eshan S. King[a,b], Jeff Maltas[b,1], and Jacob G. Scott[a,b,c,1]

Edited by Paul Turner, Yale University, New Haven, CT; received February 24, 2023; accepted February 23, 2024

Antimicrobial resistance was estimated to be associated with 4.95 million deaths worldwide in 2019. It is possible to frame the antimicrobial resistance problem as a feedback-control problem. If we could optimize this feedback-control problem and translate our findings to the clinic, we could slow, prevent, or reverse the development of high-level drug resistance. Prior work on this topic has relied on systems where the exact dynamics and parameters were known a priori. In this study, we extend this work using a reinforcement learning (RL) approach capable of learning effective drug cycling policies in a system defined by empirically measured fitness landscapes. Crucially, we show that it is possible to learn effective drug cycling policies despite the problems of noisy, limited, or delayed measurement. Given access to a panel of 15 β-lactam antibiotics with which to treat the simulated *Escherichia coli* population, we demonstrate that RL agents outperform two naive treatment paradigms at minimizing the population fitness over time. We also show that RL agents approach the performance of the optimal drug cycling policy. Even when stochastic noise is introduced to the measurements of population fitness, we show that RL agents are capable of maintaining evolving populations at lower growth rates compared to controls. We further tested our approach in arbitrary fitness landscapes of up to 1,024 genotypes. We show that minimization of population fitness using drug cycles is not limited by increasing genome size. Our work represents a proof-of-concept for using AI to control complex evolutionary processes.

antibiotic resistance | evolution | artificial intelligence

Significance

Drug-resistant pathogens are a wide-spread and deadly phenomenon that was responsible for nearly 5 million deaths worldwide in 2019. Our work highlights the development of a platform that leverages artificial intelligence to discover effective antibiotic cycling policies given limited knowledge about an evolving system. Extension and translation of technology inspired by these ideas may enable hospitals and clinicians to control drug-resistant pathogens without requiring a steady supply of new antibiotic drugs.

Drug-resistant pathogens are a wide-spread and deadly phenomenon that were responsible for nearly 5 million deaths worldwide in 2019 (1). Current projections suggest the global burden of antimicrobial resistance could climb to 10 million deaths per year by 2050 (2). In the United States alone, 3 million cases of antimicrobial-resistant infections are observed each year (3). Despite the significant public health burden of antibiotic resistance, development of novel antibiotics has slowed due to the poor return on investment currently associated with this class of drugs (4). Novel approaches to therapy design that explicitly take into account the adaptive nature of microbial cell populations while leveraging existing treatment options are desperately needed.

Evolutionary medicine is a rapidly growing discipline that aims to develop treatment strategies that explicitly account for the capacity of pathogens and cancer to evolve (5–11). Such treatment strategies, termed "evolutionary therapies," cycle between drugs or drug doses to take advantage of predictable patterns of disease evolution. Evolutionary therapies are often developed by applying optimization methods to a mathematical or simulation-based model of the evolving system under study (12–22). For example, in castrate-resistant prostate cancer, researchers developed an on–off drug cycling protocol that allows drug-sensitive cancer cells to regrow following a course of treatment. Clinical trials have shown this therapy prevents the emergence of a resistant phenotype and enables superior long-term tumor control and patient survival compared to conventional strategies (23, 24).

Current methods for the development of evolutionary therapies require an enormous amount of data on the evolving system. For example, many researchers have optimized treatment by using genotype–phenotype maps to define evolutionary dynamics and model the evolving cell population (16, 25–33). For instance, Nichol et al. modeled empirical drug fitness landscapes measured in *Escherichia coli* as a Markov chain to show that different sequences of antibiotics can promote or hinder resistance. However, defining the Markov chain framework required exact knowledge of the high-dimensional genotype–phenotype map under many drugs (16). Most published methods for optimization of these models require a complete understanding of the underlying system dynamics (15, 16, 34–36). Such detailed knowledge is currently unobtainable in the clinical

Author affiliations: [a]Case Western Reserve University School of Medicine, Cleveland, OH 44106; [b]Translational Hematology Oncology Research, Cleveland Clinic, Cleveland, OH 44106; and [c]Department of Physics, Case Western Reserve University, Cleveland, OH 44106

Author contributions: D.T.W., J.M., and J.G.S. designed research; D.T.W., E.S.K., and J.M. performed research; D.T.W. and J.M. contributed new reagents/analytic tools; D.T.W., J.M., and J.G.S. analyzed data; and D.T.W., E.S.K., J.M., and J.G.S. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1] To whom correspondence may be addressed. Email: jeff.maltas@gmail.com or scottj10@ccf.org.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2303165121/-/DCSupplemental.

Published April 12, 2024.

setting. Approaches that can approximate these optimal policies given only a fraction of the available information would fill a key unmet need in evolutionary medicine. We hypothesize that reinforcement learning algorithms can develop effective treatment policies in systems with imperfect information. Reinforcement learning (RL) is a well-studied subfield of machine learning that has been successfully used in applications ranging from board games and video games to manufacturing automation (34, 37–39). Broadly, RL methods train artificial intelligence agents to select actions that maximize a reward function. Importantly, RL methods are particularly suited for optimization problems where little is known about the dynamics of the underlying system. While previous theoretical work has studied evolutionary therapy with alternating antibiotics, none have addressed the problems of noisy, limited, or delayed measurement that would be expected in any real-world applications (12–14, 17, 40–42). Further, RL and related optimal control methods have been previously applied for the development of clinical optimization protocols in oncology and anesthesiology (21, 43–48).

In this study, we developed an approach to discovering evolutionary therapies using a well-studied set of empirical fitness landscapes as a model system (26). We explored "perfect information" optimization methods such as dynamic programming in addition to RL methods that can learn policies given only limited information about a system. We show that it is possible to learn effective drug cycling treatments given extremely limited information about the evolving population, even in situations where the measurements reaching the RL agent are extremely noisy and the information density is low.

1. Methods

1.1. Model System of Evolving Microbial Populations. As a model system, we simulated an evolving population of *E. coli* using the well-studied fitness landscape paradigm, where each genotype is associated with a certain fitness under selection (16, 26, 29). To parameterize our evolutionary model, we relied on data from a previously described fitness landscape of the *E. coli* β-lactamase gene (26). In this study, Mira et al. assayed the growth rates of 16 genotypes, representing a combinatorially complete set of 4 point mutations in the β-lactamase gene, against a set of 15 β-lactam antibiotics (Table 1). We used these data to define 15 different fitness landscapes on the same underlying genotype space, each representing the selective effects of one of the 15 drugs.

Table 1. Reference codes for drugs under study

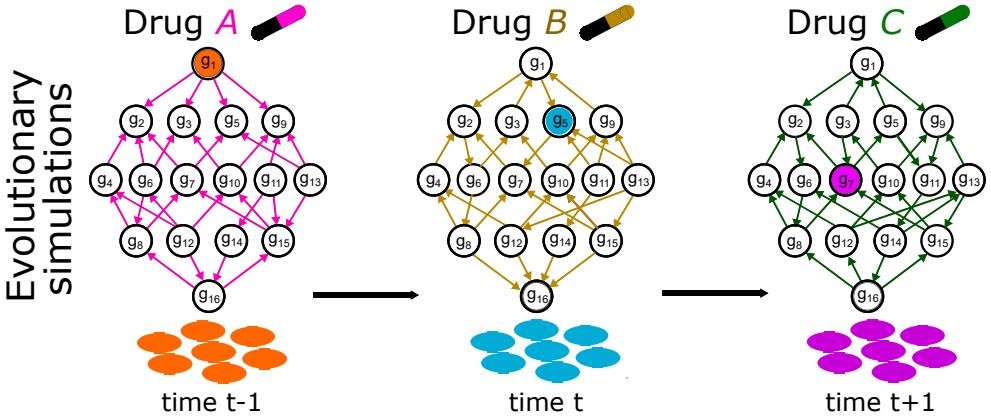
Index	Drug code	Drug
1	AMP	Ampicillin
2	AM	Amoxicillin
3	CEC	Cefaclor
4	CTX	Cefotaxime
5	ZOX	Ceftizoxime
6	CXM	Cefuroxime
7	CRO	Ceftriaxone
8	AMC	Amoxicillin + Clavulanic acid
9	CAZ	Ceftazidime
10	CTT	Cefotetan
11	SAM	Ampicillin + Sulbactam
12	CPR	Cefprozil
13	CPD	Cefpodoxime
14	TZP	Piperacillin + Tazobactam
15	FEP	Cefepime

We applied this well-studied *E. coli* model system because it is one of the few microbial cell populations for which a combinatorially complete genotype–phenotype mapping has been measured (26, 29). We extended this paradigm to larger procedurally generated landscapes as a sensitivity analysis (described in *SI Appendix*). By simulating an evolving *E. coli* cell population using the described fitness landscape paradigm, we were able to define an optimization problem on which to train RL agents (Fig. 1).

1.2. Simulation of Evolution Using Fitness Landscapes. We used a previously described fitness-landscape-based model of evolution (16, 27). In brief, we model an evolving asexual haploid population with N mutational sites. Each site can have one of two alleles (0 or 1). We can therefore represent the genotype of a population using an N-length binary sequence, for a total of 2^N possible genotypes. We can model theoretical drug interventions by defining fitness as a function of genotype and drug. Different drugs can then be represented using N-dimensional hypercubic graphs (Fig 1*A*). Further, if we assume that evolution under drug treatment follows the strong selection and weak mutation (SSWM) paradigm, we can then compute the probability of mutation between adjacent genotypes (genotypes that differ by 1 point mutation) and represent each landscape as a Markov chain as described by Nichol et al. (16, 27, 49). Briefly, the population has an equal probability of evolving to all adjacent fitter mutants and a zero probability of evolving to less fit mutants. With a sufficiently small population size, we can then assume that the population evolves to fixation prior to transitioning to another genotype. At each time step, we sampled from the probability distribution defined by the Markov chain to simulate the evolutionary course of a single population. In the model system described above (Section 1.1), fitness is given by the growth rate of a population with a given genotype under a specific drug. As a sensitivity analysis, we also incorporated a previously described phenomenological model that better simulates clonal interference compared to the base-case SSWM model (27, 50). We incorporate a parameter "phenom," which biases the probability of transition to a given genotype based on the difference in fitness between the current genotype and all adjacent genotypes. When $phenom = 0$, the model is identical to SSWM, where all adjacent genotypes with fitness larger than current fitness have an equal probability of fixing. When *phenom* is large, the model becomes more deterministic, with the most-fit genotype becoming more and more likely to fix.

1.3. Optimization Approaches.

1.3.1. Markov decision process with perfect information. First, we employed a Markov Decision Process (MDP), a mathematical framework recently used to predict optimal drug policies capable of minimizing resistance acquisition in evolving bacterial populations (13, 34). In an MDP, the system stochastically transitions between discrete states (genotypes in our case) and at each time step a decision (action, drug choice) is made. This action defines both the instantaneous reward and influences which state will occur next. The goal of an MDP is to calculate a policy that optimizes the objective function (in our case, minimizing the population fitness over a given time period). In our system, the state s_t at time step $t \in \{0, 1, 2, \ldots\}$ is defined as the vector of fitness values associated with the population's genotype. For example, if the population currently occupies genotype g_5 , the state is defined as a vector that describes the



Decreasing available model information

Model details	Markov Decision Process (MDP)	RL - genotype	RL - fit
Inputs:	all TMs + $F(g_n)$ for all genotype drug pairs	g_n current genotype + current drug	$F(g_n)$ current fitness + previous drug
Algorithm:	Dynamic programming backwards induction	Deep Q Learning	Deep Q Learning
Output:	True optimal policy mapping the optimal action/drug to each possible state	Predicted optimal treatment given any current genotype	Predicted optimal treatment given the current fitness and previous drug

Fig. 1. Schematic evolutionary simulation and tested optimization approaches. *Top:* A simulated isogenic population of *E. coli* evolves on fitness landscapes in response to drug-imposed selection pressure. At each evolutionary time step, the population transitions to a neighboring available genotype indicated by directional edges between genotype nodes. *Bottom:* A table describing the inputs and outputs of the three tested optimization algorithms. The MDP condition receives the complete transition matrix for the evolving system and fitness of each genotype-action pair. It outputs a policy function that provides an action given the current system genotype. RL-genotype is batch trained using the instantaneous genotype as the only training input. It outputs a value function, which is used to select the prescribed action given a genotype as input. RL-fit is batch trained using instantaneous fitness and previous drug as the training inputs. Like RL-genotype, it outputs a value function with which we can derive a prescribed policy.

fitness of g_5 in each of the 15 drug landscapes. At each time step an action/drug is applied and the system transitions with probability $P_a(s_{t+1}|s_t a_t)$ to the new state s_{t+1}. In the previous MDP model, the transition probabilities were estimated from replicate evolution experiments. Here, we adapted the model to fitness landscapes, estimating the transition probabilities between genotypes using the well-characterized Strong Selection Weak Mutation (SSWM) limit which can be formally described by a Markov chain. The instantaneous reward function is defined as $1 - R_a(s)$, where $R_a(s)$ is the fitness of the current genotype in the current chosen drug/action. The optimal policy $(\pi(s))$ is a formal mapping of an optimal action/drug for each genotype. The benefit of an MDP formulation is the certainty of converging on the true optimal drug policy via dynamic programming techniques, in our case, backward induction. However, the MDP formulation requires perfect knowledge of the evolving system. To set up and solve the MDP formulation, we needed the transition probabilities between each genotype given each action and the fitness of each genotype-drug pair.

1.3.2. Reinforcement learners with imperfect information. Next, we leveraged reinforcement learning to train two distinct deep learning agents. Identical to the previously described MDP, these agents suggest an action (drug) based on the information it has about the system. Importantly, these learners are able to operate without perfect information, and as a result are more suitable for clinical application where perfect genotype and fitness information is impossible to achieve. In both models, the reinforcement learners begin with no knowledge of the state space (and its associated organization/geometry), range of possible fitness values, or transition probabilities between genotypes. Over the course of 500 episodes, each of which is 20 evolutionary time steps, the reinforcement learners gather information from their evolutionary environment in response to its actions (drugs) and converge on a suggested policy. We chose to implement Deep Q reinforcement learners, a well-studied and characterized method of reinforcement learning particularly suited for situations with very little a priori knowledge about the environment (34, 51). In the first learner, termed RL-genotype, we made the instantaneous

genotype of the population at each time step known to the learner. For this learner, the neural architecture was composed of an input layer, two 1d convolutional layers, a max pooling layer, a dense layer with 28 neurons, and an output layer with a linear activation function. In the second learner, termed RL-fit, we made only the previous action, a_{t-1}, and current fitness known to the learner at each time step. RL-fit received no information about the genotype of the population. This learner was composed of a neural network with an input layer, two dense hidden layers with 64 and 28 neurons, and an output layer with a linear activation function. In both cases, Q-value estimates are improved by minimizing the temporal difference between Q-values computed by the current model and a target model, which has weights and biases that are only updated rarely. We used mean squared error as the loss function.

1.3.3. Simulation of measurement noise in the RL-fit learner. As we approach a more realistic model, we consider a scenario where the fitness measurement is imperfect. More specifically, the RL-fit learner is given the previous action, a_{t-1}, as before; however, the current fitness is subject to measurement noise drawn from a zero-mean normal distribution with variance σ^2 according to $f'(s|a) = f(s|a) + \mathcal{N}(\mu, 0.05 \times \sigma^2)$. The variable σ determines the amplitude of the noise, and therefore the precision of our fitness measurement. Finally, we evaluated the performance of RL-genotype learners that were trained on delayed information to explore the viability of using outdated sequencing information to inform drug selection (described in *SI Appendix*).

All code and data needed to define and implement the evolutionary simulation and reinforcement learning framework can be found at https://github.com/DavisWeaver/evo_dm. The software can be installed in your local python environment using "pip install git+https://github.com/DavisWeaver/evo_dm.git". We also provide all the code needed to reproduce the figures from the paper at https://github.com/DavisWeaver/rl_cycling.

2. Results

In this study, we explored the viability of developing effective drug cycling policies for antibiotic treatment given less and less information about the evolving system. To this end, we developed a reinforcement learning framework to design policies that limit the growth of an evolving *E. coli* population in silico. We evaluated this system in a well-studied *E. coli* system for which empirical fitness landscapes for 15 antibiotics are available in the literature (26). A given RL agent could select from any of these 15 drugs when designing a policy to minimize population fitness. We defined three experimental conditions. In the first, we solved a Markov decision process formulation of the optimization problem under study. In doing so, we generated true optimal drug cycling policies given perfect information of the underlying system (described in Section 1.2). In the second, RL agents were trained using the current genotype of the simulated *E. coli* population under selection (RL-genotype). Then, further constricting the information of the learner, an agent was trained using only observed fitness of the *E. coli* population along with the previous action taken (RL-fit). Finally, we introduced noise into these measures of observed fitness to simulate real-world conditions where only imprecise proxy measures of the true underlying state (genotype) may be available. Each experimental condition was evaluated based on its ability to minimize the fitness of the population under study in 20 time-step episodes. We compared these conditions to two negative controls; a drug cycling policy that selects drugs completely at random (which we will refer to as "random"), and all possible two-drug cycles (i.e.,

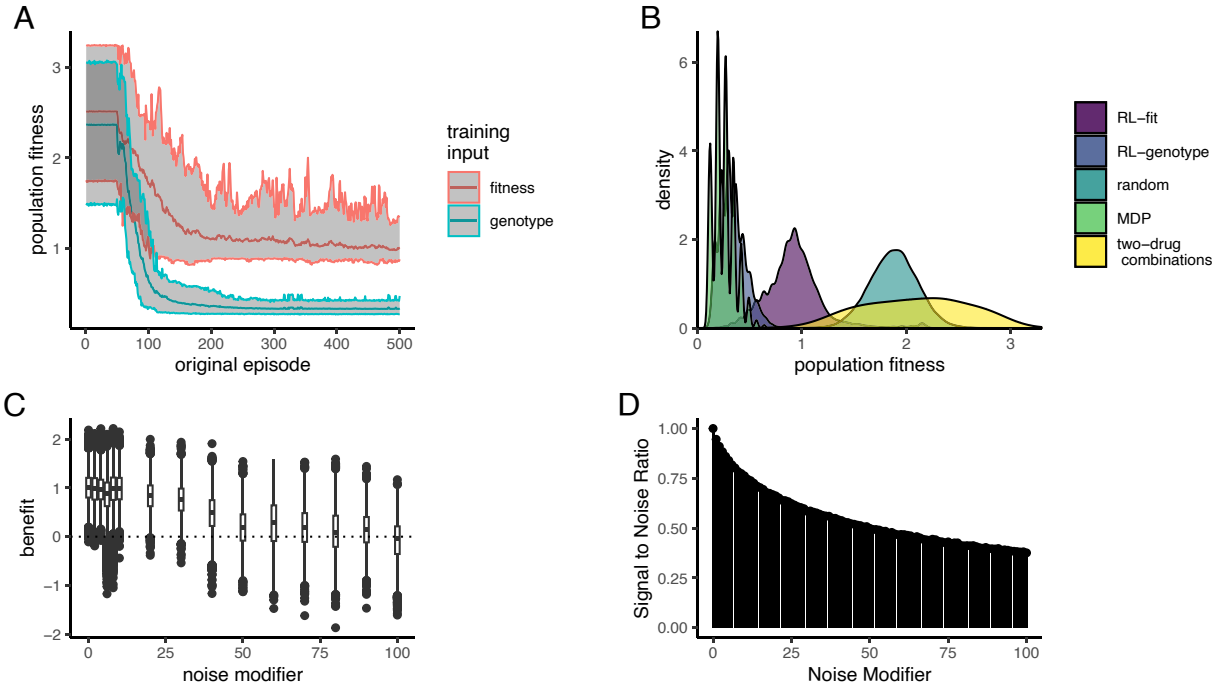


Fig. 2. Performance of RL agents in a simulated *E. coli* system. (*A*): Line plot showing the effectiveness (as measured by average population fitness) of the average learned policy as training time increases on the x-axis for RL agents trained using fitness (red) or genotype (blue). (*B*): Density plot summarizing the performance of the two experimental conditions (measured by average population fitness) relative to the three control conditions. (*C*): Boxplot showing the effectiveness of 10 fully trained RL-fit replicates as a function of noise. Each data point corresponds to one of 500 episodes per replicate (5,000 total episodes). The width of the distribution provides information about the episode by episode variability in RL-fit performance. (*D*): Signal to noise ratio associated with different noise parameters. Increasing noise parameter decreases the fidelity of the signal that reaches the reinforcement learner.

Table 2. Example drug sequences

Drug Sequence	Replicate	Condition
CTX,AMC,CTX,CPR,CTX,CPR,CTX,CPR,CTX,CPR	53	RL-fit
CTX,CPR,CPR,CPR,CTX,CPR,CPR,CPR,CTX,SAM	53	RL-genotype
CTX,AMC,CTX,AMC,CTX,AMC,CTX,AMC,CTX,CPR	23	RL-fit
CTX,AMC,CTX,AMC,CTX,AMC,CTX,AMC,CTX,AMC	23	RL-genotype
CTX,AMC,CTX,AMC,CTX,CPR,CTX,AMC,CTX,CPR	96	RL-fit
CTX,SAM,CTX,SAM,CTX,CPR,CTX,CPR,CTX,CPR	96	RL-genotype

Here, we show the first 10 selected drugs for representative episodes of the three top-performing replicates.

AMP-AM-AMP-AM-AMP...). We tested 100 replicates of RL-fit and RL-genotype against each of these conditions. Each replicate was trained for 500 episodes of 20 evolutionary steps (10,000 total observations of system behavior). We chose 500 episodes as the training time after hyperparameter tuning showed decreased or equal effectiveness with additional training.

2.1. Comparison of RL Drug Cycling Policies to Negative Controls. We found that both RL conditions dramatically reduced fitness relative to the random policy. In both cases, the RL conditions learned effective drug cycling policies after about 100 episodes of training and then fine-tuned them with minimal improvement through episode 500 (Fig. 2*A*). As expected, RL-genotype learned a more effective drug cycling policy on average compared to RL-fit. RL-genotype had access to the instantaneous genotype of the evolving population, while RL-fit was only trained using indirect measurement (population fitness). We define population fitness as the instantaneous growth rate of the current genotype. In 98/100 replicates, we observed a measurable decrease in population fitness under the learned RL-fit policy versus a random drug cycling policy (*SI Appendix*, Fig. S1*A*). Further, we found that the average RL-fit replicate outperformed all possible two-drug cycling policies (Fig. 2*B*). RL-genotype outperformed both negative controls in all 100 replicates (Fig. 2*B*). In some replicates, RL-genotype achieved similar performance compared to the MDP policy (*SI Appendix*, Fig. S1*D*). In addition, the distribution of performance for RL-genotype policies nearly overlapped with MDP performance (Fig. 2*B*). Introduction of additional noise to the training process for RL-fit led to degraded performance (Fig. 2*C*). However, even with a large noise modifier, RL-fit still outperformed the random drug cycling condition. For example, with a noise modifier of 40, RL-fit achieved an average population fitness of 1.41 compared to 1.88 for the random drug cycling condition (Fig. 2*C*).

2.2. Overview of Learned Drug Cycling Policies for RL-Fit and RL-Genotype. We evaluated the learned drug cycling policies of RL-fit and RL-genotype for the 15 β-lactam antibiotics under study. Represented drug sequences for these conditions can be found in Table 2. We compared these to the true optimal drug cycling policy as a reference. For this system, we show that the optimal drug cycling policy relies heavily on Cefotaxime, Ampicillin + Sulbactam, and Ampicillin (Fig. 3*A*). Cefotaxime was used as treatment in more than 50% of time-steps, with Ampicillin + Sulbactam and Ampicillin used

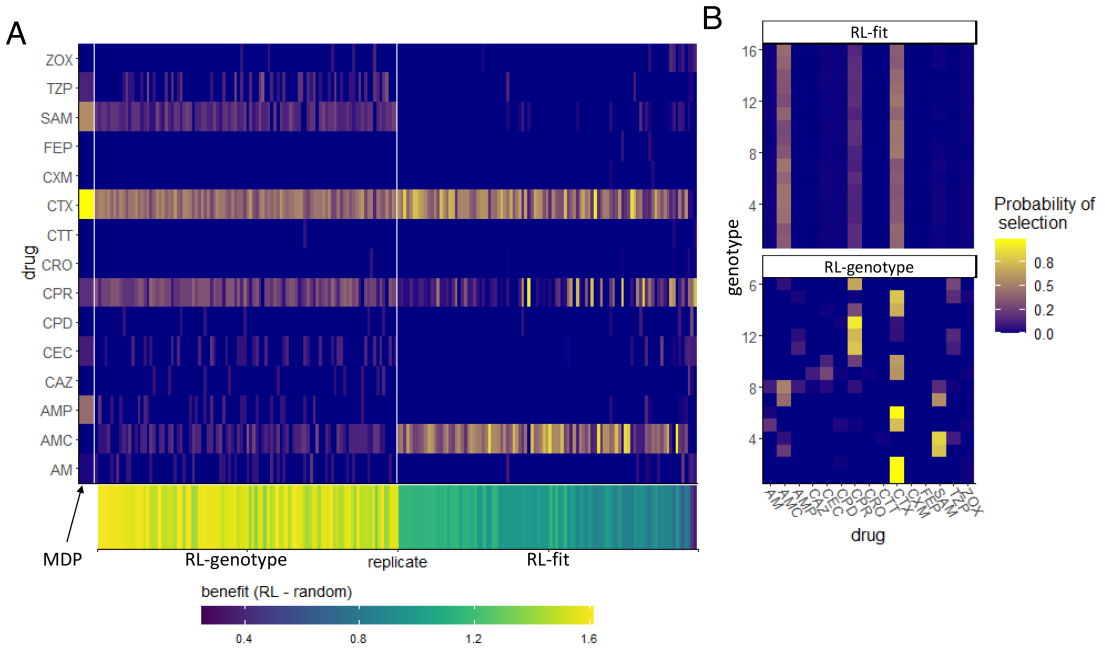


Fig. 3. Drug cycling policies learned by RL-genotype and RL-fit. (*A*): Heatmap depicting the learned policy for 100 replicates (on the x-axis) of the RL-genotype and 100 replicates of RL-fit. Far Left column (enlarged) corresponds to the optimal policy derived from the MDP condition. The Y-axis describes the β-lactam antibiotics each RL agent could choose from while the color corresponds to the probability that the learned policy selected a given antibiotic. Bottom heatmap shows the median fitness benefit observed under the policy learned by a given replicate. (*B*) Heatmap showing the average learned policy for RL-fit and RL-genotype. RL-genotype learns a more consistent mapping of genotype to action compared to RL-fit.

next most frequently. The optimal drug cycling policy used Cefprozil, Piperacillin + Tazobactam, and Cefaclor infrequently. The remaining drugs were not used at all. The different RL-fit replicates largely converged on a similar policy. They relied heavily on Cefotaxime and Amoxicillin + Clavulanic acid. However, they relied infrequently on Cefprozil. RL-genotype replicates also converged on a relatively conserved policy. Further, RL-genotype replicates showed a much more consistent mapping of genotype to action compared to RL-fit (Fig. 3*B*). Both RL-genotype and RL-fit identified complex drug cycles that use 3 or more drugs to treat the evolving population. The practical benefit of complex dosing protocols with 3+ drugs is immediately apparent as the MDP/RL-genotype/RL-fit all significantly outperform each two-drug combination. We show that policies that do not rely on Cefotaxime are suboptimal in this system. The three replicates that showed the least benefit compared to the random drug cycling case did not use Cefotaxime at all (Fig. 3*B*). The importance of Cefotaxime is likely explained by the topography of the CTX drug landscape (*SI Appendix*, Fig. S5). More than half of the available genotypes in the CTX landscape lie in fitness valleys, providing ample opportunities to combine CTX with other drugs and "trap" the evolving population in low-fitness genotypes.

2.3. Evolutionary Trajectories Observed under RL-Genotype, RL-Fit, and MDP Drug Policies. Next, we compared the evolutionary paths taken by the simulated *E. coli* population under the MDP, RL-fit, RL-genotype, and random policy paradigms. The edge weights (corresponding to the probability of observed genotype transitions) of the RL-genotype and MDP landscapes show a 0.96 Pearson correlation (*SI Appendix*, Fig. S2). In contrast, the edge weights of the RL-fit and MDP landscapes show a 0.82 Pearson correlation (*SI Appendix*, Fig. S2). During the course of training the MDP condition, the backward induction algorithm generated a value function $V(s, a)$ for all $s \in S$ and $a \in A$. In *SI Appendix*, Fig. S2*D*, we use this value function to show that certain genotypes (namely 1, 5, 6, and 13) were more advantageous to the evolving population than to the learner. These genotypes were frequented much more often under the random drug cycling condition compared to any of the experimental conditions (*SI Appendix*, Fig. S2*D*). We also show that other genotypes (namely 12 and 11) were particularly advantageous for the learner compared to the evolving population. These genotypes were frequented much more often under the experimental conditions compared to the random drug cycling condition (*SI Appendix*, Fig. S2*D*).

We also show that certain genotype transitions occur more frequently than others, independent of experimental conditions. For example, the population nearly always transitioned from genotype 5 to genotype 7 (Fig. 4). This transition highlights the way these learned policies use drug landscapes to guide evolution. Genotype 5 (0100) is a fitness peak in most of the drug landscapes used in the learned policies, and is therefore a very disadvantageous genotype for the controlling agent. CTX, the

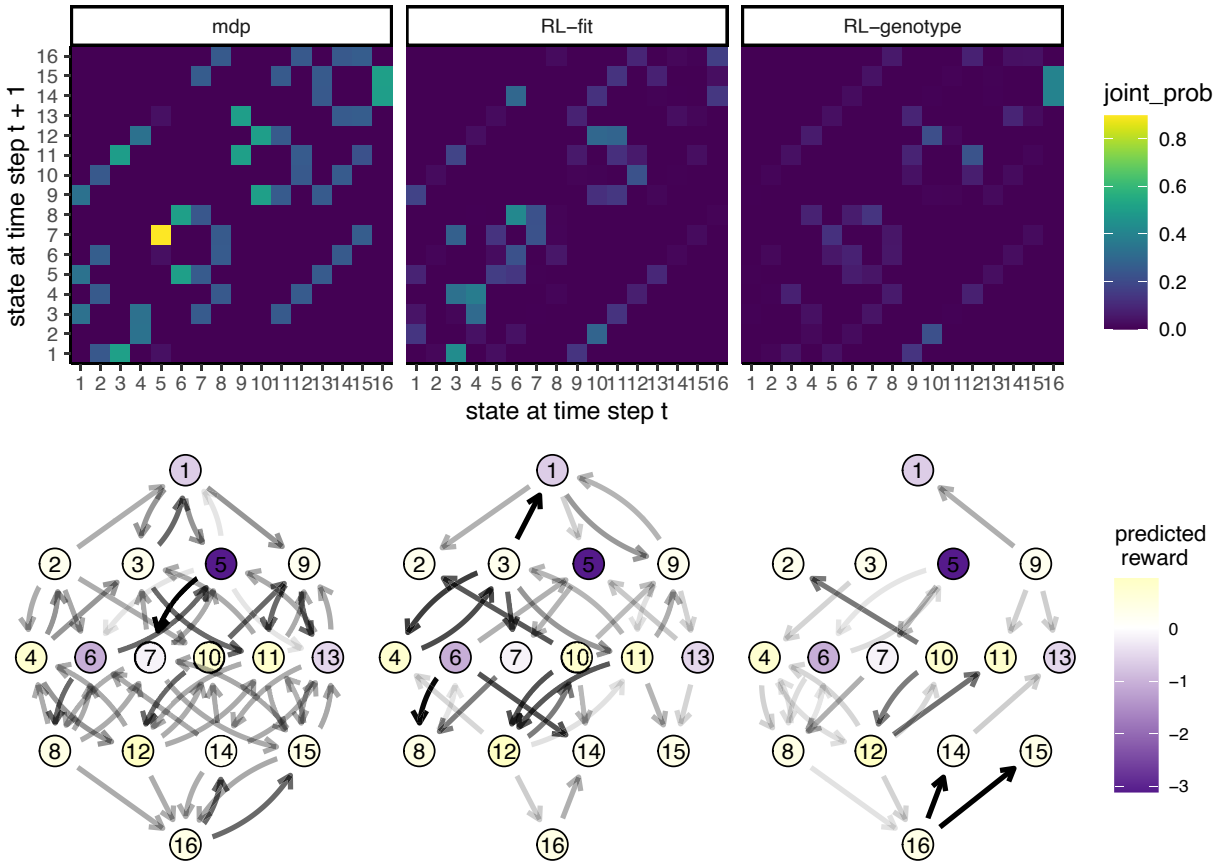


Fig. 4. Movement of simulated *E. coli* population through the genomic landscape. *Top row*: Heatmap depicting the joint probability distribution for each genotype transition under the different experimental conditions. The second two show the difference in genotype transition probability compared to the MDP condition. *Bottom row:* Graph depicting the fitness landscape, beginning with the wild type (*Bottom*) all the way to the quadruple mutant (*Top*). Size of the arrow depicts the frequency with which a genotype transition was observed under the labeled experimental condition. The color of each node corresponds with the predicted reward (to the learner) of being in that genotype. As above, the second two plots correspond to the observed difference between RL-Fit or RL-genotype and the MDP condition.

most commonly used drug in all effective policies, has a slightly higher peak at genotype 7 (0110), which forces the population away from genotype 5 (*SI Appendix*, Fig. S3).

As another example, the evolving population very rarely transitioned from genotype 1 to genotype 9 in the RL-fit condition. This genotype transition occurred commonly in the MDP and RL-genotype conditions (Fig. 4). This difference is explained by the policies shown in Fig. 3*B*. Under the RL-genotype policy, CTX was selected every time the population was in genotype 1 (the initial condition). The CTX landscape topography allows transition to 3 of the 4 single mutants, including genotype 9 (1000) (*SI Appendix*, Fig. S5). Under the RL-fit policy, CTX and AMC were used in about equal proportion when the population is in genotype 1. Unlike the CTX landscape, the AMC landscape topography does not permit evolution from genotype 1 to genotype 9 (*SI Appendix*, Fig. S5)).

2.4. Characteristics of Selected Drug Policies. To better understand why certain drugs were used so frequently by RL-genotype, RL-fit, and the MDP policies, we developed the concept of an "opportunity landscape." We computed each opportunity landscape by taking the minimum fitness value for each genotype from a given set of fitness landscapes. This simplified framework gives a sense of a potential best case scenario if the drugs in a given combination are used optimally. For example, the MDP policy relied heavily on CTX, CPR, AMP, SAM, and TZP to control the simulated *E. coli* population. The resultant opportunity landscape (Fig. 5*A*) contains only a single fitness peak, with 15/16 of the genotypes in or near fitness valleys. In Fig. 5*B*, we show the actual genotype transitions observed during evolution under the MDP policy. We also color the nodes based on the value function estimated by solving the MDP. As expected, the value function estimated by the MDP aligns closely with the topography of the opportunity landscape. There is only one genotype that the value function scores as being very poor for the learner, corresponding to the single peak in the opportunity fitness landscape (Fig. 5). Interestingly, the opportunity landscape predicted that the population would evolve to the single fitness peak and fix. In contrast, the observed genotype transitions suggest that the MDP policy was able to guide the population

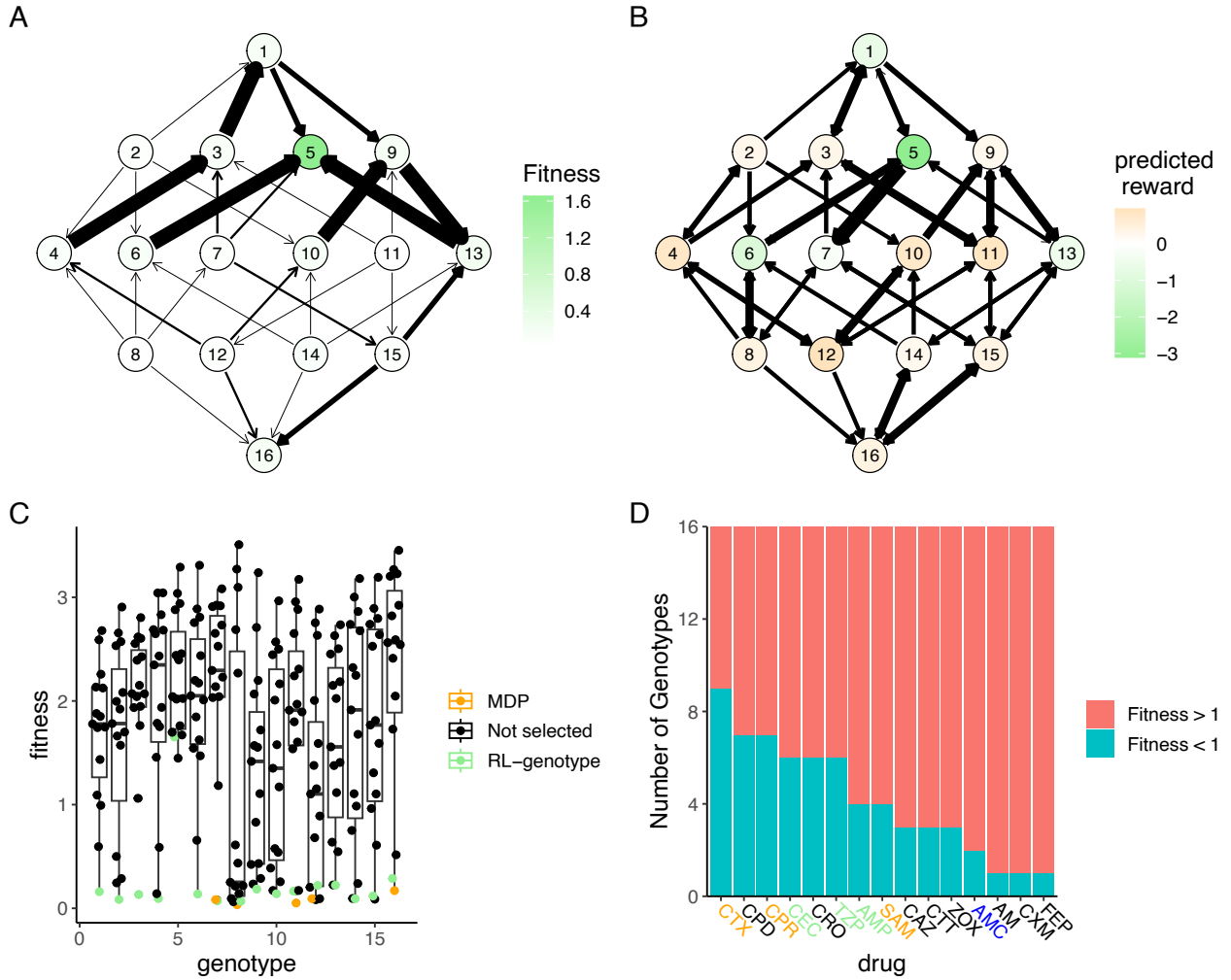


Fig. 5. MDP value function closely matches opportunity landscape for drugs commonly used under MDP policy. Panels *A* and *B* show the 16-genotype fitness landscape under study, starting with the wild type at the *Top*, progressing through the single mutants, double mutants, triple mutants, and finally the quadruple mutant at the *Bottom*. (*A*) Opportunity landscape for the five drugs most commonly used under the MDP policy (CTX, CPR, AMP, SAM, and TZP). (*B*) Observed genotype transitions under the MDP policy. The node color corresponds to the value function estimated by solving the MDP. Lower values correspond to genotypes the MDP policy attempts to avoid while higher values correspond to genotypes the MDP policy attempts to steer the population. (*C*) Scatter Plot showing the distribution of fitness with respect to genotype for the 15 β-lactam antibiotics under study. The drug selected by RL-genotype in a given genotype is highlighted in light blue. In cases where the MDP selected a different drug than RL-genotype, that drug is highlighted in orange. (*D*) Number of genotypes with fitness above or below 1 for each drug under study. Drugs that are used by both the MDP and RL-genotype are highlighted in orange. Drugs that are used by only the MDP are highlighted in green. Drugs that are used by only RL-genotype are highlighted in blue.

away from that single fitness peak. A more detailed discussion of opportunity landscapes can be found in *SI Appendix*.

We also show that both the MDP and RL-genotype conditions select the drug with the lowest fitness for most genotypes (Fig. 5*C*). There are a few notable exceptions to this rule, which highlight RL-genotype's capacity for rational treatment planning. A greedy policy that selects the lowest drug-fitness combination for every genotype would select Amoxicillin (AM) when the population is identified as being in genotype 5. The AM drug landscape then strongly favors transition back to the wild-type genotype ($g = 1$). From genotype 1, most available drugs encourage evolution back to the genotype 5 fitness peak. As we see in Fig. 5*B*, genotype 5 is by far the least advantageous for the learner. The greedy policy therefore creates an extremely disadvantageous cycle of evolution. In fact, none of the tested policies rely heavily on AM in genotype 5 (Fig. 3*B*), instead taking a reward penalty to select Cefotaxime (CTX). The CTX drug landscape encourages evolution to the double mutant, which has access to the highest-value areas of the landscape. Finally, we rank drug landscapes based on the number of genotypes with a fitness value < 1 (Fig. 5*D*). Based on the defined reward function, these genotypes would be considered advantageous to the learner. We show that drugs identified as useful by the optimal policy or RL-genotype tend to have more advantageous genotypes in their drug landscape. The only two highly permissive landscapes (CPD, CPR) that aren't used have extremely similar topography to CTX, which most policies were built around.

2.5. Impact of Landscape Size, Measurement Delay, and Clonal Interference. To understand the impact of larger fitness landscapes on the ability of our method to develop effective policies, we simulated random correlated landscapes of size N alleles, from $N = 4$ to $N = 10$, representing a range of 16 to 1,024 genotypes. Using a previously described technique, we tuned the correlation between landscapes to generate a range of collateral resistance and collateral sensitivity profiles (27). We found that reinforcement learners trained on fitness and genotype were able to outperform the random cycling control across a wide range of landscapes sizes (Fig. 6*A*). RL-genotype policies consistently outperformed RL-fit policies, which outperformed random drug cycling policies. Further, the MDP-derived optimal policy achieved similar performance on larger landscapes compared to smaller ones, suggesting that increasing genome size does not make drug cycling for evolutionary control less feasible.

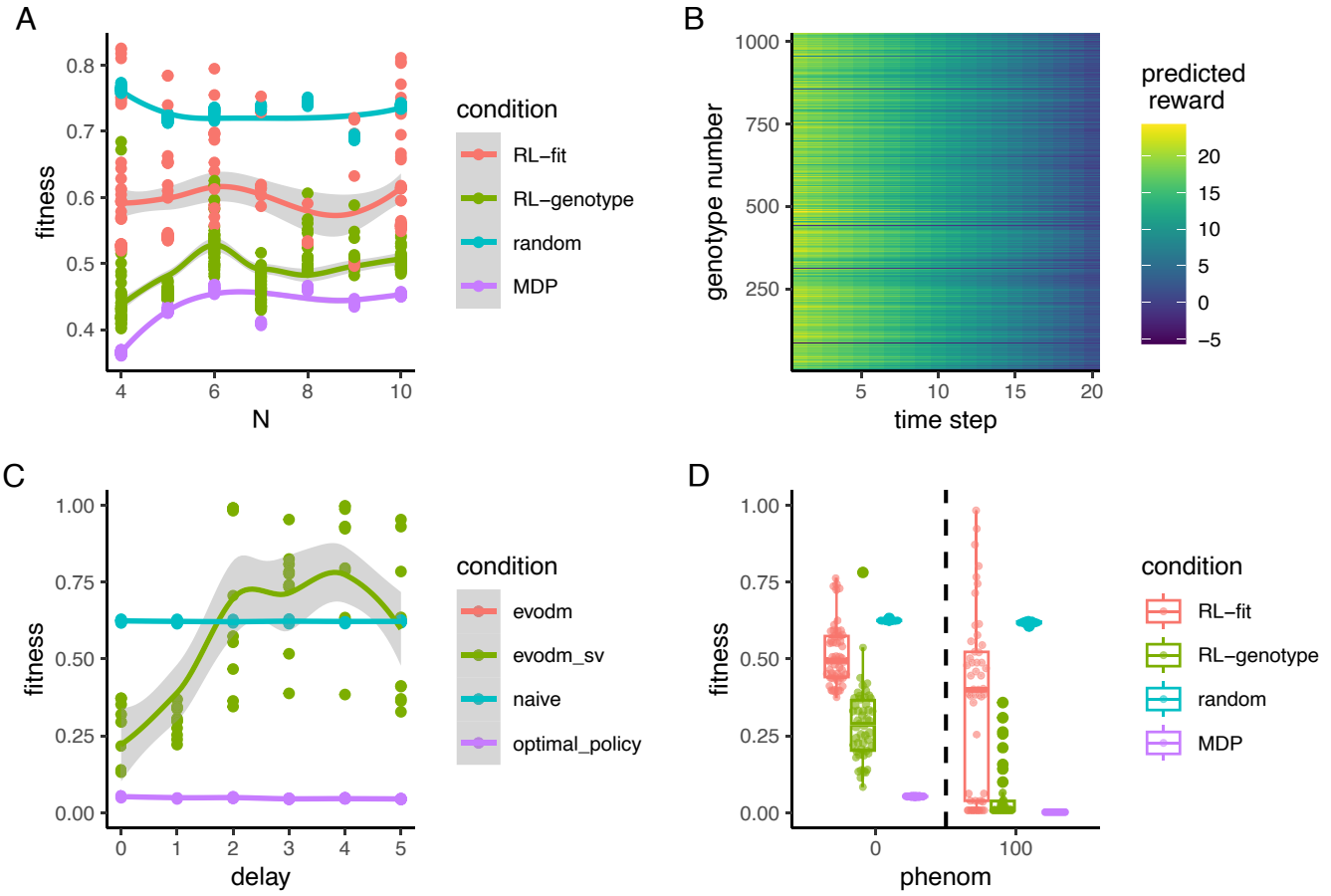


Fig. 6. Reinforcement learners learn improved policies, independent of landscape size. (*A*) Line plot describing the relationship between the number of alleles modeled and the average fitness observed under different policy regimes. The total number of genotypes in a landscape is given by 2^N. Each point represents the average fitness of a population under control of an agent trained on the set of landscapes for 500 episodes. The same set of landscapes was used for each condition. (*B*) Line plot describing the relationship between measurement delay and the observed fitness under different policy regimes. Delay only affects the RL_genotype policy (not the MDP or random conditions). RL_genotype still learned effective policies with a measurement delay of one time step. Performance decayed substantially with additional measurement delays. (*C*) Boxplot describing the relationship between phenom (a variable corresponding to the likelihood that the population will evolve to the most-fit genotype rather than any more fit genotype) and observed fitness under different policy regimens. (*D*) Heatmap showing the value function learned by solving the MDP for the $N = 10$ landscapes. The Y-axis corresponds to numerically encoded genotype, and the X-axis corresponds to the time step within a given episode. Bright cells correspond to genotypes that were advantageous to the agent while dark cells correspond to genotypes that were disadvantageous to the agent. The value space is rugged, with many peaks and valleys.

Finally, we investigated the genotype values for the largest landscape ($N = 10$), finding that the genotype value space, and therefore policy space, is rugged, with many peaks and valleys (Fig. 6*B*).

Time delays between when information is sampled from a system and when action is taken based on that information may be unavoidable in real-world applications. To understand how the practical limitation of time delays impacts our approach, we next tested the effect of measurement delays on the ability of reinforcement learners to generate effective policies. In this study, the delay parameter d controlled the "age" of the information available to the learner. With a delay of 0 ($d = 0$) time steps, the learner had instantaneous information about the system. If $d = 5$, the learner was using genotype information from time t to inform an action taken on time $t + 5$. We tested a set of delay parameters $d \in 0, 5$. For this experiment, we tested only the RL-genotype condition against the random and MDP conditions, arguing that growth rate estimates are much easier to obtain compared to sequencing, and thus measurement delays are less likely to be a practical limitation for the RL-fit condition. Because the MDP is our control/optimal policy it remains without a delay. We found that when $d <= 1$, average performance of the RL-genotype condition was equivalent to that observed in the base case. If $d > 1$, the performance of RL-genotype decreased to worse than or similar to the random condition (Fig. 6*C*).

Finally, we implemented a recent phenomenological model that simulates the effect of classic clonal interference by biasing the transition probabilities toward more and more fit adjacent mutants through increasing the *phenom* parameter. We find that with increasing clonal interference, the performance of RL-genotype and RL-fit dramatically improves (Fig. 6*D* and *SI Appendix*, S9). This is a result of significantly less stochasticity in the evolutionary system as the most-fit mutant is increasingly likely to fix in the population as *phenom* gets large. However, this model still treats the population as homogeneous and does not allow for the fixation of deleterious mutants or ecological interactions. It is possible that these factors would introduce much of the stochasticity back into the system, a focus of future work (52).

3. Discussion

The evolution of wide-spread microbial drug resistance is driving a growing public health crisis around the world. In this study, we show a proof of concept for how existing drugs could be leveraged to control microbial populations without increasing drug resistance. To that end, we tested optimization approaches given decreasing amounts of information about an evolving system of *E. coli*, and showed that it is possible to learn highly effective drug cycling policies given only empirically measurable information. To accomplish this, we developed a reinforcement learning approach to control an evolving population of *E. coli* in silico. We focused on 15 empirically measured fitness landscapes pertaining to different clinically available β-lactam antibiotics (Table 1). In this setting, RL agents selected treatments that, on average, controlled population fitness much more effectively than either of the two negative controls. Excitingly, we showed that RL agents with access to the instantaneous genotype of the population over time approach the MDP-derived optimal policy for these landscapes. Critically, RL agents were capable of developing effective complex drug cycling protocols even when the measures of fitness used for training were first adjusted by a noise parameter. This suggests that even imperfect measurements of an imperfect measure of population state (the kind of measurements we are able to make in clinical settings) may be sufficient to develop effective control policies. We also show that RL or MDP-derived policies consistently outperform simple alternating drug cycling policies. In addition, we performed a sensitivity analysis and showed how RL agents outperformed random controls with varying landscape size, measurement delay, and simulated clonal interference. Finally, we introduced the concept of the "Opportunity Landscape," which can provide powerful intuition into the viability of various drug combinations.

Our work expands a rich literature on the subject of evolutionary control through formal optimization approaches. Our group and others have developed and optimized perfect information systems to generate effective drug cycling policies (12, 13, 15, 17, 18, 53). Further, a limited number of studies have used RL-based methods for the development of clinical optimization protocols (21, 43–46, 48). These studies have been limited so far to simulated systems, including a recent study that introduced Celludose, a RL framework capable of controlling evolving bacterial populations in a stochastic simulated system (47).

Much like the studies noted above, we show that AI or MDP-based policies for drug selection or drug dosing dramatically outperform sensible controls in the treatment of an evolving cell population. We extend this literature in three key ways. We provide an optimization protocol capable of learning effective drug cycling policies using only observed population fitness (a clinically tractable measure) as the key training input. Importantly, the reinforcement learners have no prior knowledge of the underlying model of evolution. Second, we grounded our work with empirically measured fitness landscapes from a broad set of clinically relevant drugs, which will facilitate more natural extension to the bench. Third, we tested our approach in fitness landscapes of up to 1,024 genotypes, the largest state space that has been evaluated in the treatment optimization literature. We show that minimization of population fitness using drug cycles is not limited by increasing genome size.

There are several limitations to this work which bear mention. We assume that selection under drug therapy represents a strong-selection and weak mutation regime in order to compute transition matrices for our models. It is likely that other selection regimes emerge in cases of real-world pharmacokinetics or spatial regimes where the drug concentration fluctuates dramatically (54, 55). While we relax some of the strongest assumptions in the SSWM regime via a previously studied phenomenological model (27, 56), we still do not capture the possibility of deleterious or multiple simultaneous mutations to fix. We also assume the phenotype of the population is perfectly described by its genotype. As a result, contexts where epigenetic factors play a significant role will likely require a more careful treatment. In addition, we chose to keep drug concentration constant throughout our analysis, largely owing to the lack of robust empirical data linking genotype to phenotype under dose-varying conditions (sometimes called a fitness seascape) (57). As more empirical fitness seascape data become available, a natural extension would be to explore the efficacy of the RL system in controlling a population by varying both drug and dose.

While we present the most extensive genotype–phenotype modeling work to date on this subject, we still only modeled the effect of mutations at up to 10 genotypic positions. The real *E. coli* genome is approximately 5×10^6 base pairs (58). The evolutionary landscape for living organisms is staggeringly large, and not tractable to model in silico. It is possible that empirical

measures of fitness like growth rate or cell count may not provide a robust enough signal of the underlying evolutionary state on real genomes. In vitro implementations of reinforcement learning-based drug cycle optimization systems are needed to address this potential shortcoming. Another potential alternative would be to use the comparatively low-dimensional phenotype landscape of drug resistance (59).

In this work, we present a reinforcement-learning framework capable of controlling an evolving population of *E. coli* in silico. We show that RL agents stably learn complex multidrug combinations that are state specific and reliably outperformed a random drug cycling policy as well as all possible two-drug cycling policies. We also highlight key features of the types of drug landscapes that are useful for the design of evolutionary control policies. Our work represents an important proof-of-concept for AI-based evolutionary control, an emerging field with the potential to revolutionize clinical medicine.

Data, Materials, and Software Availability. All other data are included in the manuscript and/or *SI Appendix*. Previously published data were used for this work (https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0122283) (26).

ACKNOWLEDGMENTS. This work was made possible by the National Institute of Health 5R37CA244613-03 (J.G.S.) and 5T32GM007250-46 (D.T.W.), and T32CA094186 (J.A.M.) and American Cancer Society RSG-20-096-01 (J.G.S.). Fig. 1 was created with BioRender.com. Portions of the paper were included in the doctoral dissertation of DTW Weaver (2023).

1. C. J. Murray *et al*., Global burden of bacterial antimicrobial resistance in 2019: A systematic analysis. *Lancet* **399**, 629–655 (2022).
2. T. R. Walsh, A. C. Gales, R. Laxminarayan, P. C. Dodd, Antimicrobial resistance: Addressing a global threat to humanity. *PLOS Med* **20**, e1004264 (2023).
3. Centers for Disease Control and Prevention (U.S.), Antibiotic resistance threats in the United States, 2019 (Tech. Rep., Centers for Disease Control and Prevention U.S., 2019).
4. B. Plackett, Why big pharma has abandoned antibiotics. *Nature* **586** (2020).
5. S. C. Stearns, Evolutionary medicine: Its scope, interest and potential. *Proc. R. Soc. B: Biol. Sci.* **279**, 4305–4321 (2012).
6. D. Z. Grunspan, R. M. Nesse, M. E. Barnes, S. E. Brownell, Core principles of evolutionary medicine. *Evol. Med. Public Health* **2018**, 13–23 (2018).
7. G. H. Perry, Evolutionary medicine. *eLife* **10**, e69398 (2021).
8. D. I. Andersson *et al*., Antibiotic resistance: Turning evolutionary principles into clinical reality. *FEMS Microbiol. Rev.* **44**, 171–188 (2020).
9. S. Manrubia *et al*., From genotypes to organisms: State-of-the-art and perspectives of a cornerstone in evolutionary dynamics. *Phys. Life Rev.* **38**, 55–106 (2021).
10. M. Stracy *et al*., Minimizing treatment-induced emergence of antibiotic resistance in bacterial infections. *Science* **375**, 889–894 (2022).
11. M. Baym, L. K. Stone, R. Kishony, Multidrug evolutionary strategies to reverse antibiotic resistance. *Science* **351**, aad3292 (2016).
12. N. Yoon, R. Vander Velde, A. Marusyk, J. G. Scott, Optimal therapy scheduling based on a pair of collaterally sensitive drugs. *Bull. Math. Biol.* **80**, 1776–1809 (2018).
13. J. Maltas, K. B. Wood, Pervasive and diverse collateral sensitivity profiles inform optimal strategies to limit antibiotic resistance. *PLoS Biol.* **17**, e3000515 (2019).
14. J. Maltas, K. B. Wood, Dynamic collateral sensitivity profiles highlight challenges and opportunities for optimizing antibiotic sequences. bioRxiv (2021). https://www.biorxiv.org/content/10.1101/2021.12.19.473361v2.full.
15. M. Gluzman, J. G. Scott, A. Vladimirsky, Optimizing adaptive cancer therapy: Dynamic programming and evolutionary game theory. *Proc. R. Soc. B: Biol. Sci.* **287**, 20192454 (2020).
16. D. Nichol *et al*., Steering evolution with sequential therapy to prevent the emergence of bacterial antibiotic resistance. *PLOS Comput. Biol.* **11**, e1004493 (2015).
17. N. Yoon, N. Krishnan, J. Scott, Theoretical modeling of collaterally sensitive drug cycles: Shaping heterogeneity to allow adaptive therapy. *J. Math. Biol.* **83**, 47 (2021).
18. S. Iram *et al*., Controlling the speed and trajectory of evolution with counterdiabatic driving. *Nat. Phys.* **17**, 135–142 (2021).
19. J. Maltas *et al*., Drug dependence in cancer is exploitable by optimally constructed treatment holidays. *Nat. Ecol. Evol.* 1–16 (2023).
20. S. Chakrabarti, F. Michor, Pharmacokinetics and drug interactions determine optimum combination strategies in computational models of cancer evolution. *Cancer Res.* **77**, 3908–3921 (2017).
21. P. K. Newton, Y. Ma, Nonlinear adaptive control of competitive release and chemotherapeutic resistance. *Phys. Rev. E* **99**, 022404 (2019).
22. S. Kim, T. D. Lieberman, R. Kishony, Alternating antibiotic treatments constrain evolutionary paths to multidrug resistance. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 14494–14499 (2014).
23. J. Zhang, J. J. Cunningham, J. S. Brown, R. A. Gatenby, Integrating evolutionary dynamics into treatment of metastatic castrate-resistant prostate cancer. *Nat. Commun.* **8** (2017).
24. J. J. Cunningham, J. S. Brown, R. A. Gatenby, K. Staňková, Optimal control to develop therapeutic strategies for metastatic castrate resistant prostate cancer. *J. Theor. Biol.* **459**, 67–78 (2018).
25. D. M. Weinreich, R. A. Watson, L. Chao, Perspective: Sign epistasis and genetic costraint on evolutionary trajectories. *Evolution* **59**, 1165–1174 (2005), https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.0014-3820.2005.tb01768.x.
26. P. M. Mira *et al*., Rational design of antibiotic treatment plans: A treatment strategy for managing evolution and reversing resistance. *PLoS One* **10**, e0122283 (2015).
27. J. Maltas, D. M. McNally, K. B. Wood, Evolution in alternating environments with tunable inter-landscape correlations. *Evol. Int. J. Organ. Evol.* **75**, 10–24 (2021).
28. J. A. G. M. de Visser, J. Krug, Empirical fitness landscapes and the predictability of evolution. *Nat. Rev. Genet.* **15**, 480–490 (2014).
29. D. M. Weinreich, N. F. Delaney, M. A. Depristo, D. L. Hartl, Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science (New York, N.Y.)* **312**, 111–114 (2006).
30. C. B. Ogbunugafor, C. S. Wylie, I. Diakite, D. M. Weinreich, D. L. Hartl, Adaptive landscape by environment interactions dictate evolutionary dynamics in models of drug resistance. *PLoS Comput. Biol.* **12**, 1–20 (2016).
31. E. Toprak *et al*., Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. *Nat. Genet.* **44**, 101–105 (2012).
32. S. F. Greenbury, A. A. Louis, S. E. Ahnert, The structure of genotype-phenotype maps makes fitness landscapes navigable. *Nat. Ecol. Evo.* 1–11 (2022).
33. M. Baym *et al*., Spatiotemporal microbial evolution on antibiotic landscapes. *Science (New York, N.Y.)* **353**, 1147–1151 (2016).
34. R. S. Sutton, A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, 2018).
35. L. Du Plessis, G. E. Leventhal, S. Bonhoeffer, How good are statistical models at approximating complex fitness landscapes? *Mol. Biol. Evol.* **33**, 2454–2468 (2016).
36. D. Nichol *et al*., Antibiotic collateral sensitivity is contingent on the repeatability of evolution. *Nat. Commun.* **10** (2019).
37. D. Silver *et al*., Mastering the game of Go with deep neural networks and tree search. *Nature* **529** (2016).
38. V. Mnih *et al*., Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
39. O. Vinyals *et al*., Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* **575**, 350–354 (2019).
40. L. Imamovic, M. O. Sommer, Use of collateral sensitivity networks to design drug cycling protocols that avoid resistance development. *Sci. Transl. Med.* **5**, 204ra132-204ra132 (2013).
41. R. E. Mandt *et al*., Diverse evolutionary pathways challenge the use of collateral sensitivity as a strategy to suppress resistance. *eLife* **12**, e85023 (2023).
42. J. Maltas, D. M. McNally, K. B. Wood, Evolution in alternating environments with tunable interlandscape correlations. *Evolution* **75**, 10–24 (2021).
43. B. L. Moore *et al*., Reinforcement learning for closed-loop propofol anesthesia: A study in human volunteers. *J. Mach. Learn. Res.* **15**, 655–696 (2014).
44. B. K. Petersen *et al*., Deep reinforcement learning and simulation as a path toward precision medicine. *J. Comput. Biol.: A J. Comput. Mol. Cell Biol.* **26**, 597–604 (2019).
45. R. Padmanabhan, N. Meskin, W. M. Haddad, Reinforcement learning-based control of drug dosing for cancer chemotherapy treatment. *Math. Biosci.* **293**, 11–20 (2017).
46. I. Ahn, J. Park, Drug scheduling of cancer chemotherapy based on natural actor-critic approach. *Bio. Syst.* **106**, 121–129 (2011).
47. D. Engelhardt, Dynamic control of stochastic evolution: A deep reinforcement learning approach to adaptively targeting emergent drug resistance. *J. Mach. Learn. Res.* **21**, 1–30 (2020).
48. R. B. Martin, Optimal control drug scheduling of cancer chemotherapy. *Automatica* **28**, 1113–1123 (1992).
49. J. H. Gillespie, Molecular evolution over the mutational landscape. *Evolution*, 1116–1129 (1984).
50. L. Tan, J. Gore, Slowly switching between environments facilitates reverse evolution in small populations. *Evol. Int. J. Organ. Evol.* **66**, 3144–3154 (2012).
51. K. Arulkumaran, M. P. Deisenroth, M. Brundage, A. A. Bharath, Deep reinforcement learning: A brief survey. *IEEE Signal Process. Magaz.* **34**, 26–38 (2017).
52. R. Barker-Clarke, J. M. Gray, D. S. Tadele, M. Hinczewski, J. G. Scott, Maintaining, masking, and mimicking selection: the interplay of cell-intrinsic and cell-extrinsic effects upon eco-evolutionary dynamics. bioRxiv [Preprint] (2023). https://www.biorxiv.org/content/10.1101/2023.03.15.532871v2.
53. M. Wang, J. G. Scott, A. Vladimirsky, Stochastic optimal control to guide adaptive cancer therapy. bioRxiv [Preprint] (2022). https://www.biorxiv.org/content/10.1101/2022.06.17.496649v1.
54. N. Krishnan, J. G. Scott, Range expansion shifts clonal interference patterns in evolving populations (2019). Pages: 794867 Section: New Results.
55. E. S. King, B. Pierce, M. Hinczewski, J. G. Scott, Diverse mutant selection windows shape spatial heterogeneity in evolving populations. bioRxiv [Preprint] (2023). https://www.biorxiv.org/content/10.1101/2023.03.09.531899v3.
56. L. Tan, J. Gore, Slowly switching between environments facilitates reverse evolution in small populations. *Evolution* **66**, 3144–3154 (2012).
57. E. S. King *et al*., Fitness seascapes are necessary for realistic modeling of the evolutionary response to drug therapy. bioRxiv (2022). https://www.biorxiv.org/content/10.1101/2022.06.10.495696v2.
58. C. K. Rode, L. J. Melkerson-Watson, A. T. Johnson, C. A. Bloch, Type-specific contributions to chromosome size differences in *Escherichia coli*. *Infect. Immun.* **67**, 230–236 (1999).
59. J. Iwasawa *et al*., Analysis of the evolution of resistance to multiple antibiotics enables prediction of the *Escherichia coli* phenotype-based fitness landscape. *PLoS Biol.* **20**, e3001920 (2022).
60. D. T. Weaver, "Novel approaches for optimal therapy design in drug-resistant populations," Ph.D. thesis, Case Western Reserve University (2023). https://etd.ohiolink.edu/acprod/odbetd/etd/r/1501/10?clear=10&p10accessionnum=case168321594645388.