

418 Supplemental Materials

419 In the following please find the supplemental materials for the manuscript entitled: "Reinforcement Learning informs optimal
420 treatment strategies to limit antibiotic resistance"

421 1.1 Procedurally generated drug landscapes

422 The 4 curated amino acid substitutions used to generate fitness landscapes in this study likely represent the most important
423 mutations in the evolution of drug resistance to β -lactam antibiotics in this model system. However, there may be other
424 "off-landscape" mutations in genes such as drug efflux pumps that also significantly impact resistance. Furthermore, other
425 organisms or drug combinations may demand much larger landscapes to effectively predict evolution^{53,54}. To evaluate the
426 feasibility of reinforcement learning based drug cycle optimization in larger state spaces, we generated correlated landscapes
427 with arbitrary N (where N is the number of alleles), following a procedure adapted from previous work²⁷. We first generated an
428 index landscape L by sampling a vector of length 2^N (where N is the number of alleles) from a uniform distribution with range
429 $(-1, 1)$. We then introduced epistasis by applying a gaussian noise vector (with $\mu = 0$ and $\sigma = 0.5$) to each element of the
430 vector. This process is traditionally referred to as a "rough Mt Fuji." Next, we generated a set of n correlated landscapes L_c
431 with correlations to the original landscape ranging from -1 (perfectly anti-correlated) to 1 (perfectly correlated). Briefly, we
432 generate a Gaussian random vector (with zero mean and variance) of length 2^N . We subtract from this vector its projection onto
433 the original landscape vector L , making our new vector orthogonal to L . It then follows that any vector L_c is a linear combination
434 of L and our new orthogonal vector. In practice, anti-correlated drug landscapes display striking collateral sensitivity, while
435 correlated drug landscapes display collateral resistance. From the set of correlated landscapes L_c , we selected a subset of
436 landscapes demonstrating the full range of correlations to ensure that collateral sensitivity and collateral resistance were both
437 present.

438 1.2 Measurement Delay

439 In this sensitivity analysis, we aimed to assess the viability of RL-based drug cycling policies in a setting where actions are
440 taken based on "out-of-date" information. For example, if DNA sequencing takes multiple days to process, drugs applied based
441 on that data would be reacting to a version of the population that no longer exists. We therefore defined a delay parameter d
442 which controlled the number of time steps removed the action was from the measurement of environmental state. Put another
443 way, s_{t-d} informed a_t . If $d = 0$ (as in our base case), s_t informed a_t .

444 We limited this analysis to the RL-genotype condition. Our rationale was that fitness or growth rate measurements are much
445 easier to obtain and such delays wouldn't be as common, even in complex *in vitro* settings. We hypothesized that out-of-date
446 state vectors would lack sufficient information content to effectively inform the reinforcement learning agents.

447 1.3 Hyperparameter tuning

448 We varied key hyperparameters one at a time in order to identify optimal values to promote learning in this setting. Parameter
449 ranges and the selected value are shown in **Table S1**. Due to the long run-times of the training process, we were unable to make
450 use of more formal hyper-parameter optimization approaches. Future work will increase the efficiency of training reinforcement
451 learners in this setting, opening up a number of interesting follow-on studies.

Table S1. Key Hyperparameters for reinforcement learner

Parameter	Value	Range
gamma	0.99	0-1
learning rate	0.0001	0.000001-0.1
minibatch size	60	20-500
update target model frequency	310	100-1000

452 **1.4 Additional performance data for RL agents**

453 As mentioned in the main text, we tested both the RL-fit and RL-genotype conditions 100 times each. In **Fig S1**, we show
454 the performance of all 100 RL-fit and RL-genotype replicates. In 98/100 replicates, RL-fit outperformed the random drug
455 cycling case (**Fig S1A**). The very best RL-fit replicates still fell short of the MDP-derived optimal policy (**Fig S1B**). In all 100
456 replicates, RL-genotype outperformed the random drug cycling case (**Fig S1C**). RL-genotype performance approached the
457 performance of the optimal policy (**Fig S1D**).

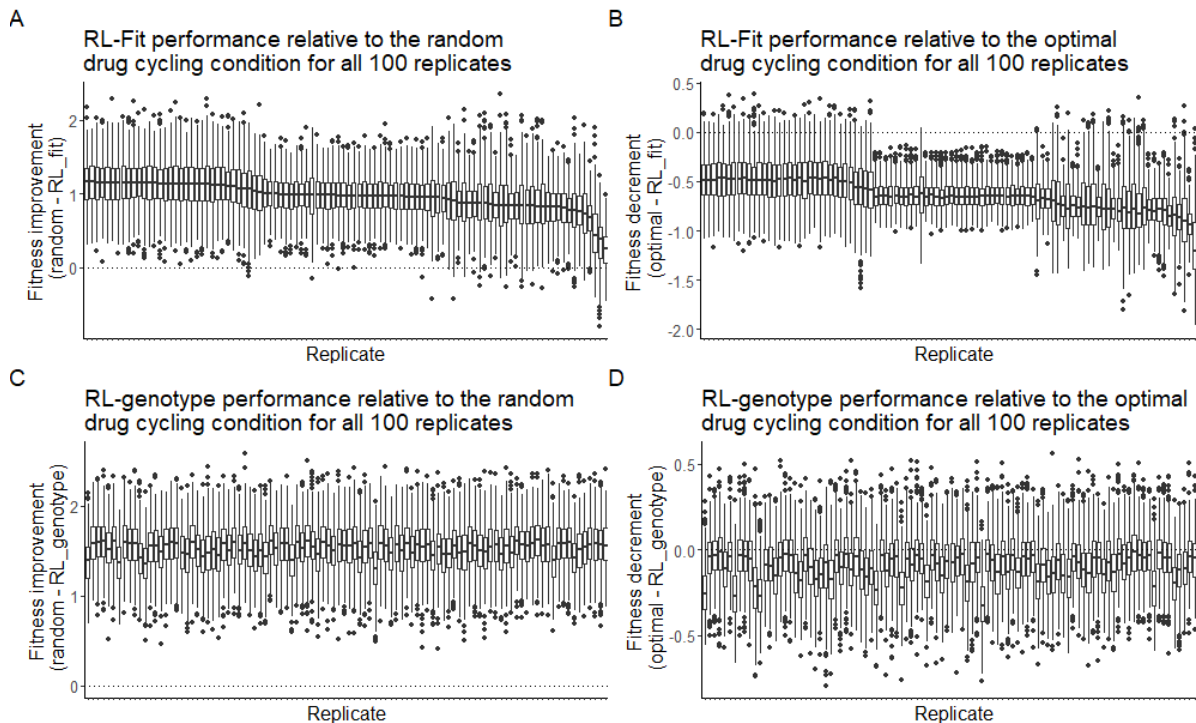


Figure S1. Performance of RL-fit and RL-genotype for each replicate. **A:** Fitness observed under RL-fit policy compared to random drug cycling condition. **B:** Fitness observed under RL-fit policy compared to fitness observed under optimal policy. **C:** Fitness observed under RL-genotype policy compared to random drug cycling condition. **D:** Fitness observed under RL-genotype policy compared to fitness observed under optimal policy.

458 **1.5 Additional evolutionary trajectory data**

459 We compared the state transition frequencies observed under different policy regimes, where a state transition is defined as the
 460 population evolving from genotype s_a to genotype s_b . We also show the frequency with which each state was visited under
 461 different conditions. Policy performance is closely tied to the frequency with which state 5 (a high fitness genotype in nearly all
 462 drugs) is visited. (**Fig S2**).

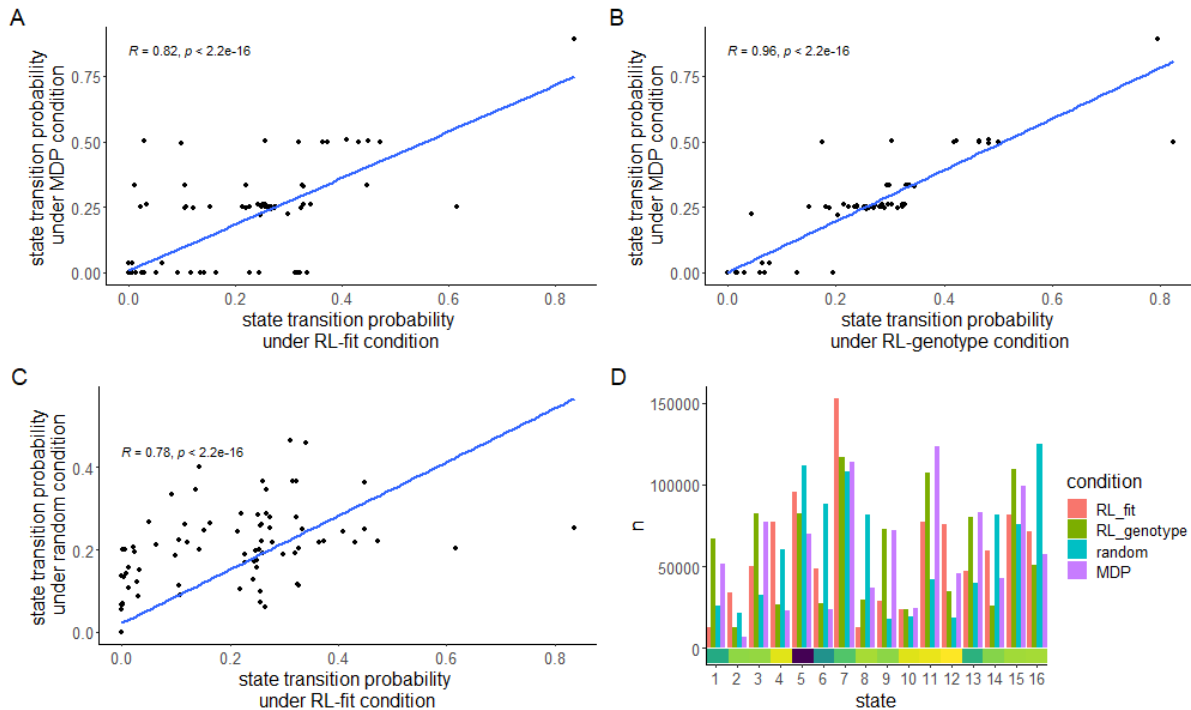


Figure S2. Comparison of evolutionary trajectories seen under different regimes A-C: Selected Pairwise comparisons of state transition frequency under different experimental conditions. State transition frequency is nearly identical for the RL-genotype and MDP conditions ($R=0.96$). In contrast, state-transition-frequency for the RL-fit and MDP conditions are related but less strongly correlated ($R=0.82$). As expected, state transition frequency were least similar between the RL-fit and random conditions ($R=0.78$). **D:** Bar chart comparing the frequency that states are observed under different experimental conditions. The value of each state (to the learner) is highlighted for each state by the bottom heatmap. High value states are observed more frequently in RL-fit, RL-genotype, and MDP conditions compared to the random condition.

463 **1.6 Opportunity Landscapes**

464 We define an opportunity landscape to be the most optimistic combination of n landscapes, formed by taking the minimum
465 possible fitness at each genotypic position. This construct can help us better understand how the learner uses different
466 combinations of drugs to maintain the evolving population at extremely low fitness values.

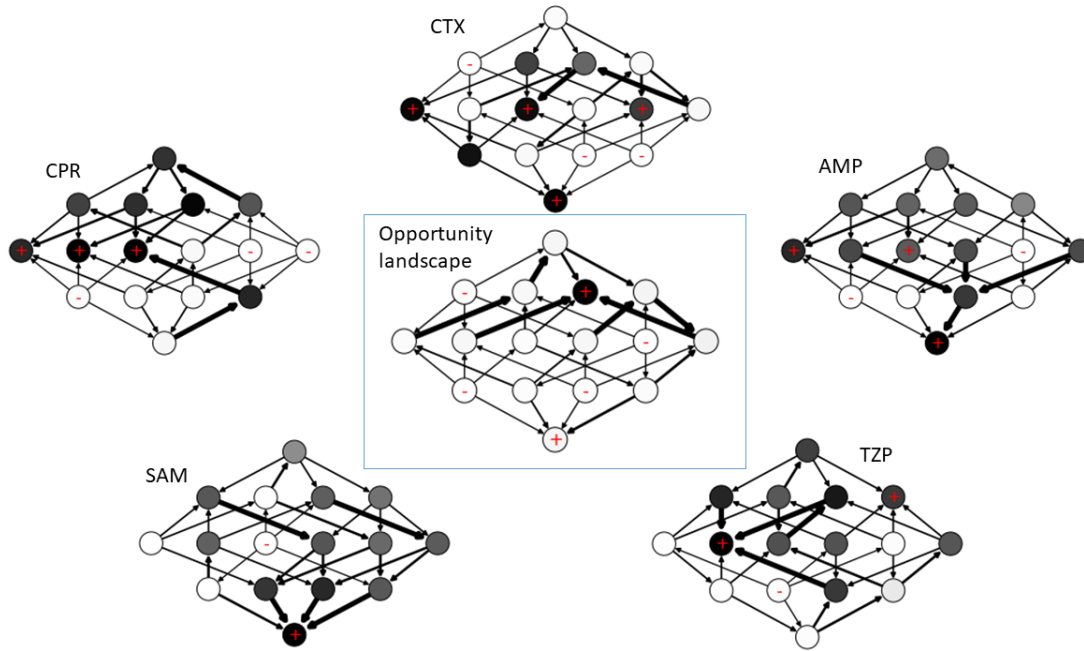


Figure S3. Opportunity Landscape for MDP-derived policy. Opportunity landscape is an optimistic combination of 5 empirically measured drug landscapes. Just 1/16 genotypes is near a fitness peak on the opportunity landscape, helping to explain the extremely low fitness observed in the simulated *E. coli* population when the MDP-derived policy is applied.

467 **Fig. S3** describes the opportunity landscape discovered by the MDP condition. As noted in the main text, the MDP primarily
468 uses 5 drugs (CTX, CPR, AMP, SAM, and TZP) in combination to trap the evolving population of *E. coli* at extremely low
469 fitness genotypes. In the combined opportunity landscape, just one genotype (0100) had a high fitness in all 5 drugs. As
470 expected, the opportunity landscape closely matches the value function estimated by the MDP (**Fig 4**).

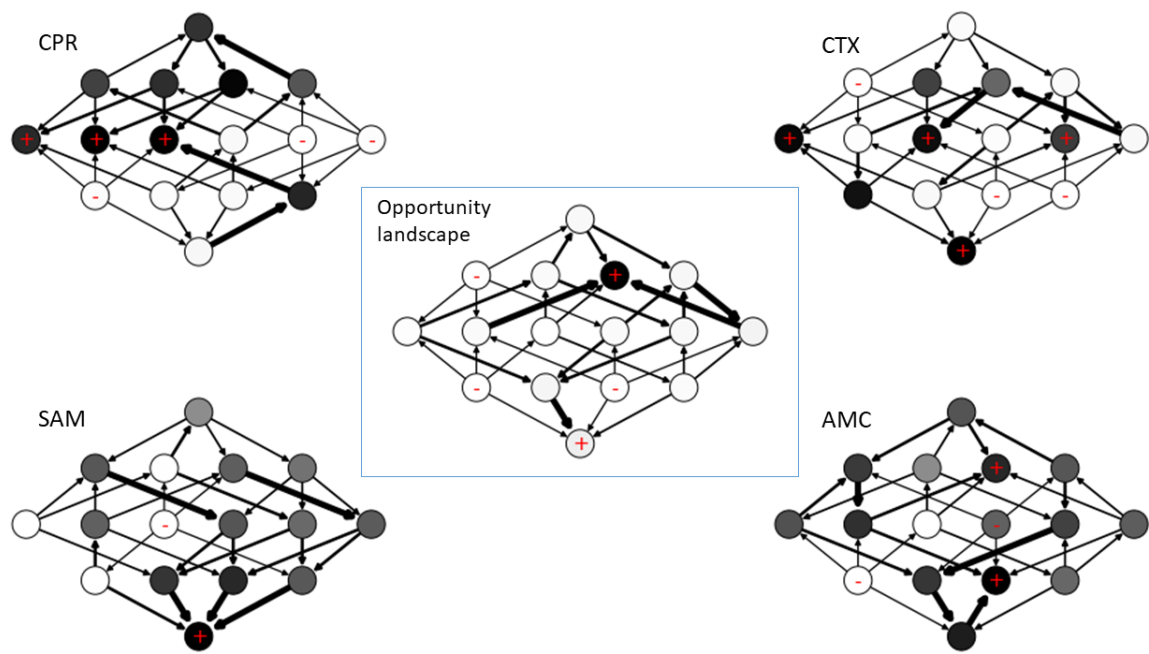


Figure S4. Opportunity landscape for most common policy identified in the RL-genotype condition. As in the MDP-derived policy, just 1/16 genotypes is near a fitness peak in the opportunity landscape.

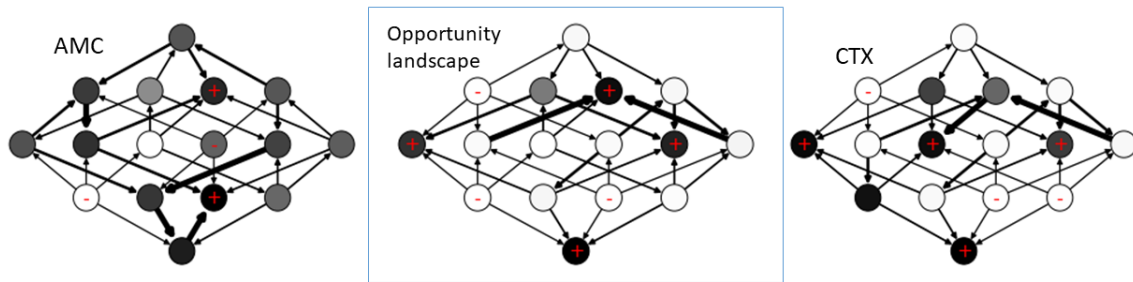


Figure S5. Opportunity landscape for the most common policy identified in the RL-fit condition. The most common RL-fit policy relies on AMC and CTX to control the E. coli population. Assuming the most optimistic combination of these two drug landscapes, 4/16 genotypes are near a fitness peak.

471 The opportunity landscape for the RL-genotype is almost identical to the opportunity landscape observed for the MDP
 472 policy (Fig S4). Interestingly, RL-genotype only uses 3 of the 5 drugs in the MDP policy; CPR, CTX, and SAM. RL-fit
 473 discovered policies that typically only used two drugs. The most effective RL-fit policies relied heavily on AMC and CTX. We
 474 present the resulting opportunity landscape in Fig S5. As expected, there are more genotypes with high fitness values under this
 475 two-drug paradigm compared to the 4 or 5 drug policies discovered by RL-genotype and the MDP, respectively.

476 **1.7 MDP policy**

477 As mentioned in the main text, we computed the MDP policy by formulating a Markov decision process of the strong selection,
478 weak mutation model of evolution under study. We then solved the MDP using backward induction, an algorithm designed to
479 identify an optimal policy for a finite time discrete MDP. The identified policy is a function of current state and current time
480 step, making it even more specific than the policies identified by the reinforcement learning conditions. We show the time and
481 state-specific MDP policy in **Fig S6**. Near the end of an episode (steps 19 and 20), we see a switch to a greedy policy that
482 simply selects the drug with the minimum fitness for a given genotype.

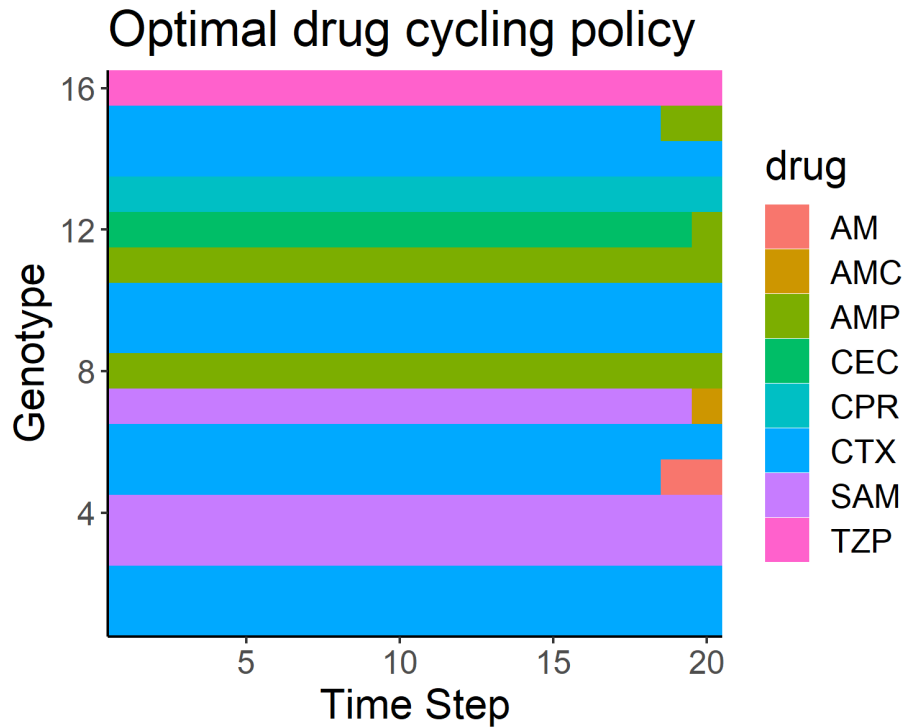


Figure S6. MDP-derived optimal policy for the empirical drug landscapes condition. The X-axis corresponds to numerically encoded genotype while the y-axis corresponds to time step within a given episode. Fill color corresponds to which drug the MDP-derived policy selects for each genotype-time step combination.

483 We also varied the discount rate (γ), between 0 and 1 during the hyperparameter tuning process. In **Fig S7**, we show the
484 effect of gamma on the average fitness achieved by the MDP policy. While gamma didn't have a large effect, likely due to the
485 relatively short length (20 time steps) of each episode, we show that increasing γ led to increased performance of the computed
486 MDP policy. We also show that increasing gamma led to increased use of CTX (drug 4).

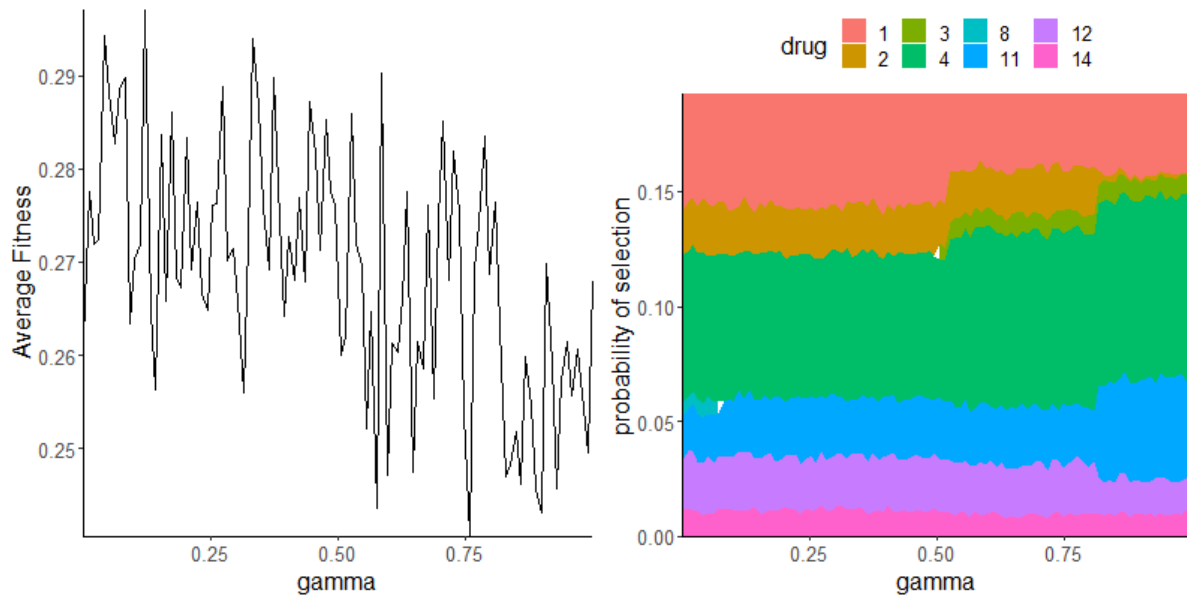


Figure S7. Effect of variation in γ on optimal policy performance and composition. We show that greedy policies (low γ) are slightly less effective compared to non-greedy policies. In panel B, we show how the actual policy changes as γ ranges from 0-0.9999

487 **1.8 Additional Analyses**

488 As noted in **Fig 2C** in the main text, we evaluated the performance of all A-B-A-B two-drug cycles to use as a comparison
489 group for RL-fit and RL-genotype. In **Fig S8A**, we examine these combinations in greater depth. We also show the landscape
490 correlation between the two drugs in every combination. We show that anti-correlated landscapes tend to make more effective
491 combinations, likely due to collateral sensitivity. Highly correlated landscapes tend to make ineffective drug combinations,
492 likely due to collateral resistance.

493 Finally, we evaluated the effect of starting population genotype on the performance of each two-drug combination. We
494 found that the starting genotype of the population had no effect on the overall distribution of performance for these two-drug
495 combinations (**Fig S8**).

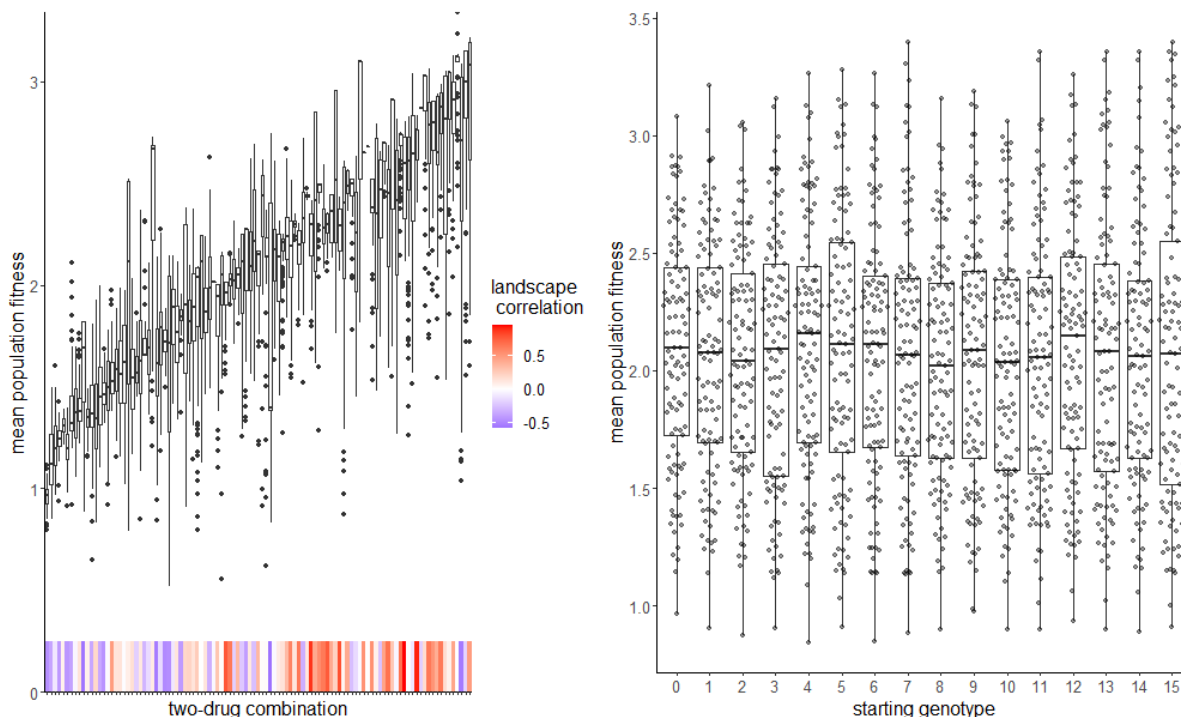


Figure S8. In the left panel, we show the fitness observed under every possible A-B-A-B two drug regime. The heatmap shows the correlation between the two landscapes in a pair, a measure of the expected collateral sensitivity or resistance. In the right panel, we show the effect of starting genotype on the performance of two-drug policies.