# Reinforcement Failing guides the discovery of emergent physical dynamics in adaptive tumor therapy

Serhii Aif[1,2], Maximillian Eiche[1,2], Nico Appold[1,2], Elias Fischer[1,2], Timon Citak[1,2], and Jona Kayser[1,2,*]

[1]Max-Planck-Zentrum für Physik und Medizin & Max Planck Institute for the Science of Light, 91058 Erlangen, Germany
[2]Department of Physics, Friedrich-Alexander-Universität Erlangen-Nürnberg, 91054 Erlangen, Germany
[*]E-mail: jona.kayser@mpzpm.mpg.de

September 18, 2025

## Abstract

**Artificial intelligence is revolutionizing scientific discovery in medicine, with reinforcement learning (RL) emerging as a promising tool for optimizing therapeutic strategies. Yet applying RL to complex scenarios such as therapy dynamics in solid tumors is constrained by the challenge of constructing training environments that are both computationally efficient and mechanistically interpretable. Here we introduce Reinforcement Failing, an AI-guided, human-in-the-loop discovery framework that shifts the focus from agent policy optimization to the refinement of the training environment itself. By combining multi-fidelity RL with group-relative performance evaluation across agent cohorts, Reinforcement Failing systematically reveals emergent mechanisms that first-principles models overlook. We apply this framework to adaptive therapy in solid tumors, which seeks to delay resistance-mediated treatment failure. In this setting, Reinforcement Failing uncovered a coupling between the mechanically driven collective motion of cells and spatially-heterogeneous proliferation that strongly influences therapy outcomes. Incorporating these emergent physical mechanisms into an augmented training environment improved cross-environment therapeutic performance and exposed potential pitfalls in translation. More broadly, these findings position Reinforcement Failing as a powerful artificial scientific discovery framework, capable of deciphering high-complexity processes at the interface of physics, machine learning, and medicine.**

## Main

Despite substantial advances in modern cancer treatments — including targeted therapies and immunotherapies — the emergence of therapy resistance remains one of the most urgent challenges in modern oncology, calling for innovative approaches

[1]. While many patients initially respond to anti-tumor therapy, long-term tumor control often remains elusive due to the presence of resistant subpopulations. Building on mathematical models of tumor dynamics, preclinical and clinical pilot studies of evolution-based adaptive therapies (AT) have recently demonstrated how intratumor competition between susceptible and resistant subpopulations may be exploited to extend the time to disease progression (Fig. 1 a) [2–11].

The transition from fixed dosing regiments toward adaptive, evolution-informed strategies reflects a broader shift in medicine toward flexible, data-driven interventions — a shift increasingly enabled by advances in artificial intelligence [12]. Within ongoing efforts to optimize dosing in chemo- and radiotherapy [13–19], recent studies by Gallagher *et al.* [20] and Lu *et al.* [21] have applied reinforcement learning (RL) directly to adaptive therapy optimization via more nuanced dosing strategies (Fig. 1b).

Since training of RL agents is typically performed in simulations, a central challenge lies in identifying which system characteristics must be captured by these training environments to ensure transferability of the obtained policy to the original context. In packed cellular aggregates, including solid tumors, this "abstraction" is particularly challenging due to the intricate coupling of spatial population structure, resource gradients, and mechanical cell–cell interactions, all of which can drive collective phenomena that alter evolutionary trajectories and treatment responses [22–25]. Computational and experimental studies have shown that these mechanisms strongly influence resistance evolution and the potential benefits of adaptive dose scheduling [26–34].

Modeling such processes explicitly makes simulations both opaque for scientific interpretation and computationally prohibitive for RL. The nascent field of multi-fidelity RL aims to address these challenges by pairing low-fidelity training environments with subsequent evaluation and possibly transfer learning in high-fidelity simulations [35, 36]. However, without principled guidance on which emergent physical mechanisms shape therapeutic outcomes, this abstraction process becomes a bottleneck, potentially omitting key dynamics [37].

To overcome this critical hurdle, this work introduces a novel framework that integrates high-fidelity simulations of physical tumor growth dynamics with group RL to guide the systematic augmentation of the low-fidelity training environment (Fig. 1c-d). The core idea behind this approach is "Reinforcement Failing" (RF) (Fig. 1e), a novel concept that leverages Group-Relative Agent Performance Evaluation (GRAPE) to expose hidden system features that shape therapeutic success. Instead of discarding the policies of presumably "failing" RL agents, we show that these can serve as informative signal for human researches, guiding them toward hidden emergent processes that differentiate between therapy success and failure. Using this RL-guided human-in-the-loop approach, we uncover that position-dependent growth and mechanically-induced cell motion are key drivers of therapy outcomes. Augmenting the training environment accordingly yields superior performing agents while facilitating a detailed understanding of the underlying mechanisms and ensuing therapeutic pitfalls by human researchers. In the following, we first outline the general architecture of the framework before applying it to the specific scenario of adaptive tumor therapy.

# Results

## Reinforcement Failing: A framework for RL-guided augmentation of abstracted training environments

The Reinforcement Failing framework consists of four sequential main blocks: construction of the reference environment and its abstraction into a training environment, training of agent ensemble via reinforcement learning, cross-environment group-relative agent performance evaluation, and the human-in-the-loop augmentation of the training environment by previously hidden features (Fig. 1 e).

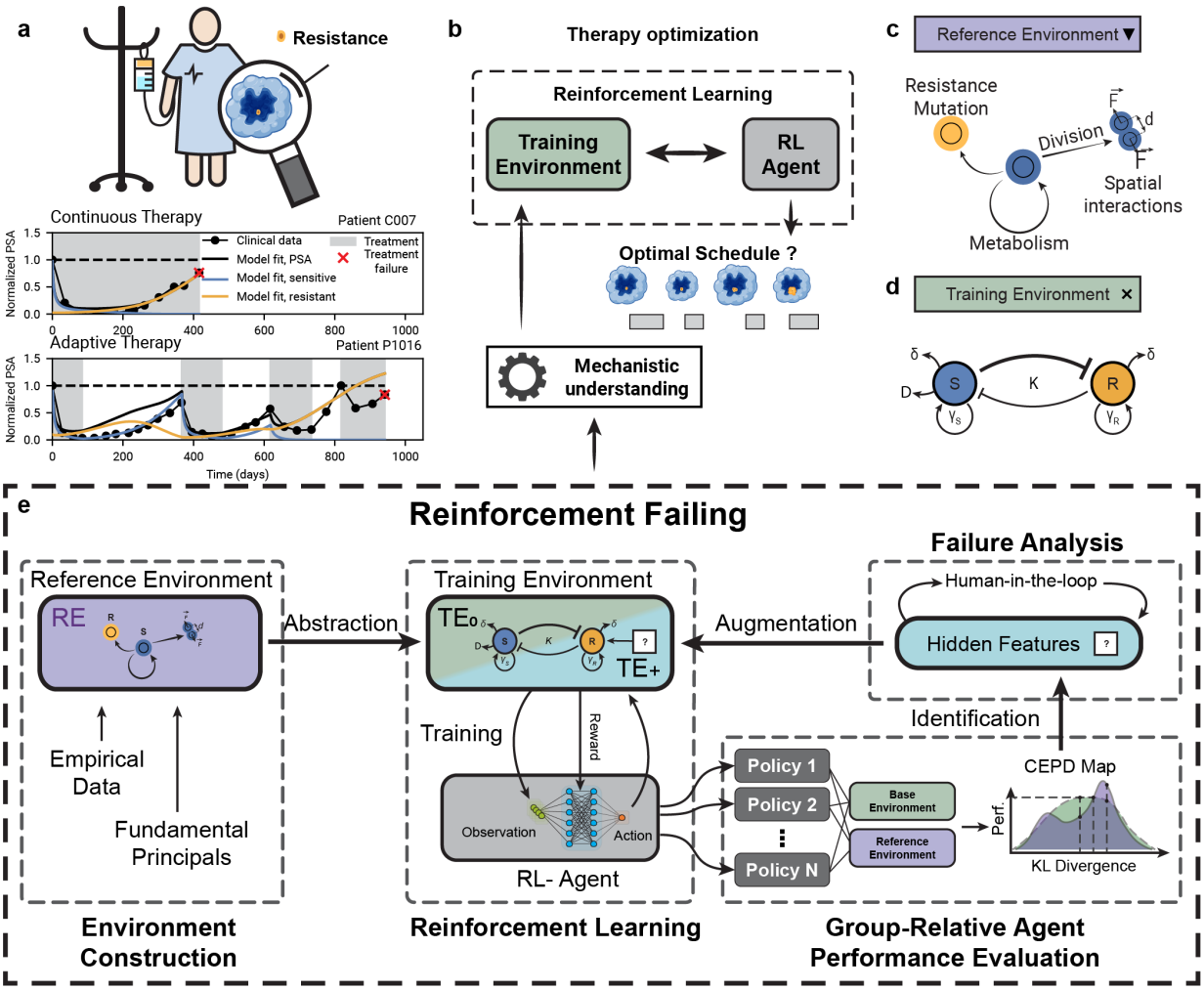



Figure 1: **Overview of the Reinforcement Failing framework**. **a** Illustration of a cancer patient with a tumor that has developed resistance. Prostate cancer patients data - normalized prostate-specific antigen (PSA) values for continuous and adaptive therapy $AT_{50}$ (adapted from Ref. [3]). Solid lines represent a mathematical model fit to individual patient data as reported in that study. **b** Schematic of therapy optimization via Reinforcement Learning using a mechanistically interpretable training environment. **c** Schematic of the reference environment, explicitly simulating the growth, division, mutation, and force-induced mutations of thousands of cells. **d** Schematic of mathematical tumor growth model based on coupled ordinary differential equations, used as the training environment (see main text for details). **e** Overview of the Reinforcement Failing framework, consisting of four main blocks. Environment Construction: Based on empirical data and fundamental physical principles, a high-fidelity simulation is designed as reference environment (RE). The complex reference environment is then abstracted into a mathematical base training environment ($TE_0$). Reinforcement Learning: An ensemble of $N$ independent agents is trained in $TE_0$. Group-Relative Agent Performance Evaluation (GRAPE): Evaluating all agents in both $TE_0$ and RE yields a cross-environment performance difference (CEPD) map. Failure Analysis: CEPD-guided analysis of differentially performing agents by a human researcher enables the identification of performance-critical but previously hidden system features and their integration into a refined augmented training environment ($TE_+$). The augmentation is followed by the training and evaluation of second-generation agents. Tumor schematic created in BioRender. Kayser, J. (2025) `https://BioRender.com/3jkoj49`.

The initial step is the construction of the high-fidelity reference environment (RE), designed to capture the intricacies of the complex real system, followed by its subsequent abstraction into a low-fidelity but high-efficiency base training environment ($TE_0$). The base training environment is then used to independently train an ensemble of treatment-decision-making agents with a diverse range of policies via reinforcement learning.

A key innovation of the framework is the use of Group-Relative Agent Performance Evaluation (GRAPE) to uncover hidden but outcome-critical emergent features. Rather than aiming to find the single best policy, GRAPE moves beyond absolute performance metrics as the central goal but instead aims to systematically explore the decision space more broadly. To this end, each agent is evaluated in both the training and reference environments to generate a cross-environment performance difference (CEPD) map of the policy space (see Methods for details).

Using the Kullback–Leibler divergence between state–action probability distributions as a distance metric to hierarchically order policies in behavior space reveals the most informative agents and guides human researchers in identifying key behavioral patterns underlying performance mismatches. Understanding these failure modes provides the foundation for augmenting the training environment with the newly identified features. The augmented environment ($TE_+$) is then used to train a cohort of second-generation agents, which are subsequently analyzed with GRAPE. If necessary, this cycle can be repeated for iterative model refinement until performance in the training and reference environments converges.

## Bridging tumor physics and adaptive therapy via Reinforcement Learning

To assess the role of emergent physics-driven dynamics in a therapy setting, we construct a high-fidelity reference environment (RE) based on the state-of-the-art PhysiCell cancer simulator [38–42]. In this model, individual cells are represented as mechanically interacting spheres that move freely in continuous space in response to external forces (Fig. 1c and Fig. 2). Cells proliferate by consuming diffusible resources and can be killed upon drug exposure. Any higher-order population phenomena, such as competition, is solely mediated via the abundance of local resources (Supplementary Movie 1).

The RE provides a detailed representation of emergent system characteristics, but its computational demands limit its suitability for training. Therefore, we first abstract the high-fidelity physics-based RE into a more tractable, low-fidelity mathematical training environment (Fig. 2c).

As initial base training environment ($TE_0$), we use a system of coupled ordinary differential equations that describe the competition between drug-sensitive ($S$) and drug-resistant ($R$) populations (Fig. 1d), following previous work [20, 43]:

$$\frac{dS}{dt} = \gamma_S S \cdot (1 - \frac{S+R}{K}) - \delta \cdot S - D \tag{1a}$$

$$\frac{dR}{dt} = \gamma_R R \cdot (1 - \frac{R + c \cdot S}{K}) - \delta \cdot R \tag{1b}$$

The environment models logistic growth of the two competing populations, with growth rates $\gamma_S$ and $\gamma_R$, limited by a shared carrying capacity $K$ and an asymmetric competition coefficient $c$. Cell death is modeled through a stochastic rate $\delta$, and the susceptible population additionally experiences a drug-induced death $D$, which is set to zero in the absence of treatment. In contrast to previous implementations, the drug effect is modeled independently of population size to account for the resource-limited periphery-to-core killing observed in the reference environment (see Methods for details).

We evaluated two treatment protocols in the RE: continuous therapy (CT) (Supplementary Movie 2) and adaptive therapy ($AT_{50}$) (Supplementary Movie 3) (Fig. 2 d,e). The adaptive therapy protocol is analogous to the design of the clinical study by Zhang *et al.*, pausing treatment when the tumor burden falls below 50% of its baseline value and resuming treatment when the burden returns to 100% [2, 3]. The dynamics observed in our *in silico* model qualitatively capture the trajectories reported in the clinical trial, including an extended control of tumor growth (compare Fig. 1 a). This is an inherent behavior of the

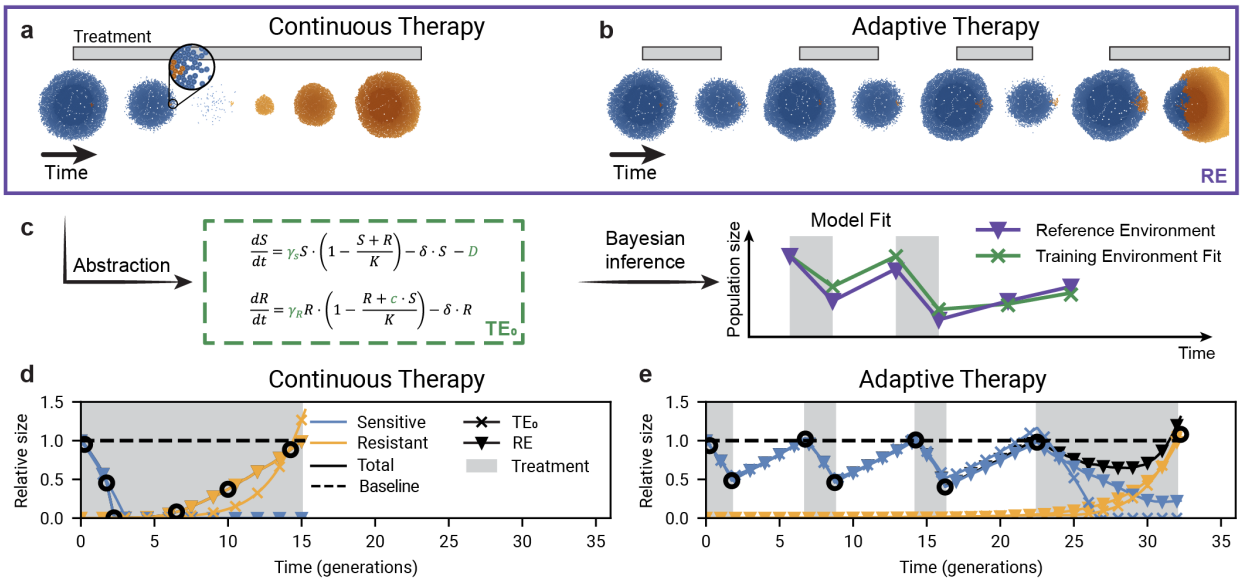



Figure 2: **A cell-based tumor simulation as reference environment integrates emergent physical effects and therapy dynamics**. **a**, **b** Snapshots of the RE simulation under Continuous Therapy (CT, Supplementary Movie 2, **a**) and Adaptive Therapy (AT50, Supplementary Movie 3, **b**). Sensitive cells are shown in blue, resistant cells in orange; brighter colors indicate higher proliferation rates. Gray bars indicate treatment application. **c** Schematic of the abstraction of reference environment dynamics into a fitted training environment. **d**, **e** Dynamics of sensitive (blue) and resistant (orange) population sizes, relative to the total initial population, are shown for RE (triangles) and the fitted $TE_0$ (crosses), under CT (**d**) and AT50 (**e**). Grey regions denote treatment application, and black circles mark the time points shown in the snapshots.

system without any fitted parameters. Notably, the observed benefit of adaptive scheduling is not contingent on a resistance-associated fitness costs, corroborating previous reports on adaptive therapy in lattice-based spatial models [43, 44]. The parameters of the base training environment ($TE_0$) are fitted to the dynamics under both strategies observed in RE via Bayesian inference (see Methods for details).

While it is clear that the simplified $TE_0$ may not capture the whole spectrum of system dynamics, we want to test how well the complex behavior of the physical reference environment can be abstracted into *effective* parameters. For example, in the reference environment, many cells are confined to the resource-depleted tumor core, resulting in a lower effective proliferation rate $\gamma_S$ than the actual division rate in the physical simulation. This abstraction process is analogous to the fitting of mathematical models to clinical patient data (Fig. 1a, b) [2, 3, 20, 30]. Note, these effective parameters inferred from this simplification should not be interpreted as reflecting the true biological rates in the system.

Despite its inherent limitations, the parameterized training environment qualitatively captures the trajectories of the reference environment (Fig. 2 a). To quantitatively compare strategy performance, we measure the time to therapy failure (TTF), defined as the time at which resistant cells reach the initial tumor size, rendering continued confinement elusive (see Methods).

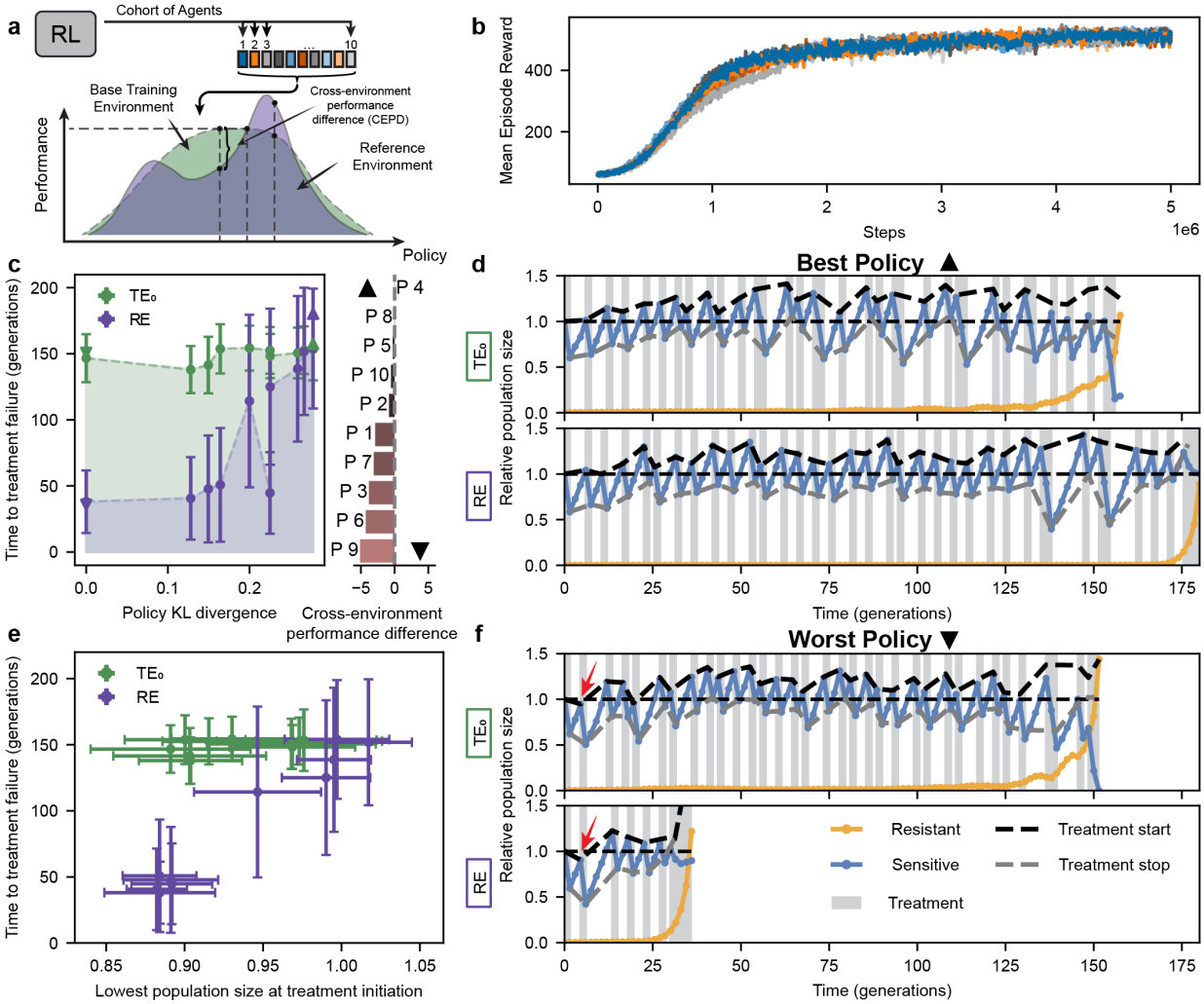



Figure 3: **Reinforcement learning and GRAPE reveal critical agent behavior**. **a** Schematic of agent policy space exploration and cross-environmental performance difference. **b** Mean episode reward over training steps of $N = 10$ trained agents in the $TE_0$. **c** Left: Time to treatment failure (TTF) in $TE_0$ (green) and RE (purple) for policies trained in $TE_0$ ($N = 10$). Dots represent averages over $n = 50$ runs, error bars indicate the standard deviation. The Kullback-Leibler divergence of each policy from the worst-performing policy is used as behavior space coordinate. Right: Standardized cross-environment performance differences (Cohen's d) are shown for all ($N = 10$) policies. Identified best (▲) and worst (▼) policies are indicated. **d**, **f** Examples of the dynamics of the sensitive (blue) and resistant (orange) relative population sizes for the best and worst performing policies in the reference environment, shown for both $TE_0$ and RE (Supplementary Movies 4 and 5). Shown runs are indicated in **c** with ▲ (best) and ▼ (worst), colored according to the environment. Grey regions indicate treatment application. Dashed curves show the tumor burdens at treatment initiation (black) and treatment pause (gray). Arrows point to the application of treatment at low tumor burden by the agent using the worst policy (see main text). **e** Time to therapy failure (TTF) in $TE_0$ (green) and RE (purple) for $N = 10$ policies trained in $TE_0$ as a function of the lowest total population size at which treatment was initiated. Dots represent averages over $n = 50$ runs. Error bars indicate standard deviation.

We find that $AT_{50}$ significantly outperforms CT, approximately doubling TTF in both the mathematical and the physical environments (CT: $15.3 \pm 0.2$ generations in $TE_0$ and $14.9 \pm 0.5$ generations in RE; $AT_{50}$: $33.8 \pm 0.8$ generations in $TE_0$ and $30.2 \pm 2.9$ generations in RE; Extended Data Fig. 1). Based on these observations, we form the null-hypothesis that $TE_0$ sufficiently captures the dynamics of the RE for reinforcement learning.

## Application of Reinforcement Failing to adaptive tumor therapy

To apply Reinforcement Failing in adaptive tumor therapy, we trained a cohort of $N = 10$ independent agents in the parameterized base training environment $TE_0$ to explore the therapy decision space under the given model assumptions (Fig. 3a). Agents were trained via proximal policy optimization (PPO) and a customized reward function, using a sliding window of only the total normalized tumor burden as observation (see Methods for details). During training, all agents successfully converged to effective policies (Fig. 3 b). By learning how to use adaptive treatment application to optimally balance overall population growth and the risk of competitive release of resistant cells, trained agents achieved an average TTF of $149 \pm 19$ generations in the training environment (Fig. 3c). This represents an almost fivefold improvement over the fixed-threshold $AT_{50}$ strategy ($33.8 \pm 0.8$ generations).

However, applying the GRAPE evaluation framework reveals a substantial heterogeneity in agent performance when evaluated in the reference environment (Fig. 3c; Extended Data Fig. 2). This lack of transferability despite successful training rejects our null-hypothesis and instead suggests that the $TE_0$ base training environment is insufficient.

To understand cause of the observed performance heterogeneity, we analyze the decision-making behavior and environment-state trajectories of poorly performing agents and compare them to those of agents that generalize well (Figs. 3d-f). We find that a key distinction lies in their treatment initiation patterns. High-performing agents initiate treatment at progressively higher tumor burdens, delaying aggressive therapy until tumor regrowth accelerates. In contrast, poorly performing agents have at least one early decision in which they apply treatment when tumor sizes are already reduced (see arrows in Fig. 3f). This episode is of little consequence in the training environment as long as it is followed by sufficient tumor regrowth (Fig. 3f, top). However, in the reference environment, once a sufficiently sharp reduction in tumor burden occurs, the system is unable to recover, resulting in uncontrolled growth of resistant cells and ultimately therapy failure (Fig. 3f, bottom).

To quantitatively test whether aggressive treatment at low tumor burdens is a predictor of agent failure, we measure the smallest tumor size at which each agent initiates treatment and analyze its correlation to the time of therapy failure (Fig. 3e). No significant correlation is observed in the training environment (Pearson's correlation coefficient, $r = 0.45$). In contrast, we find a strong correlation ($r = 0.98$) in the reference environment, with well- and poorly-performing agents forming distinct clusters. Conversely, we find that a high average total burden - the conventionally proposed driver of resistance confinement - strongly correlated with long TTF in the training environment ($r = 0.91$) but was a much weaker predictor in the reference environment ($r = 0.77$) (Extended Data Fig. 3)[44, 45]. These observations suggest additional mechanisms as determinants of therapy failure dynamics (see Supplementary Information and Extended Data Fig. 4 for a detailed analysis).

## Growth-driven cell motion accelerates resistance escape

Having pinpointed the agent behaviors responsible for divergent performance, we can now investigate the underlying mechanisms by analyzing the full spatio-temporal dynamics of relevant subtrajectories in a human-in-the-loop manner (Fig. 4a; Supplementary Movie 6).

We find that during the post-treatment expansion phase, resistant cells within the bulk can be physically displaced by the collective growth dynamics following a therapy cycle, especially if preceded by a substantial reduction of the susceptible population

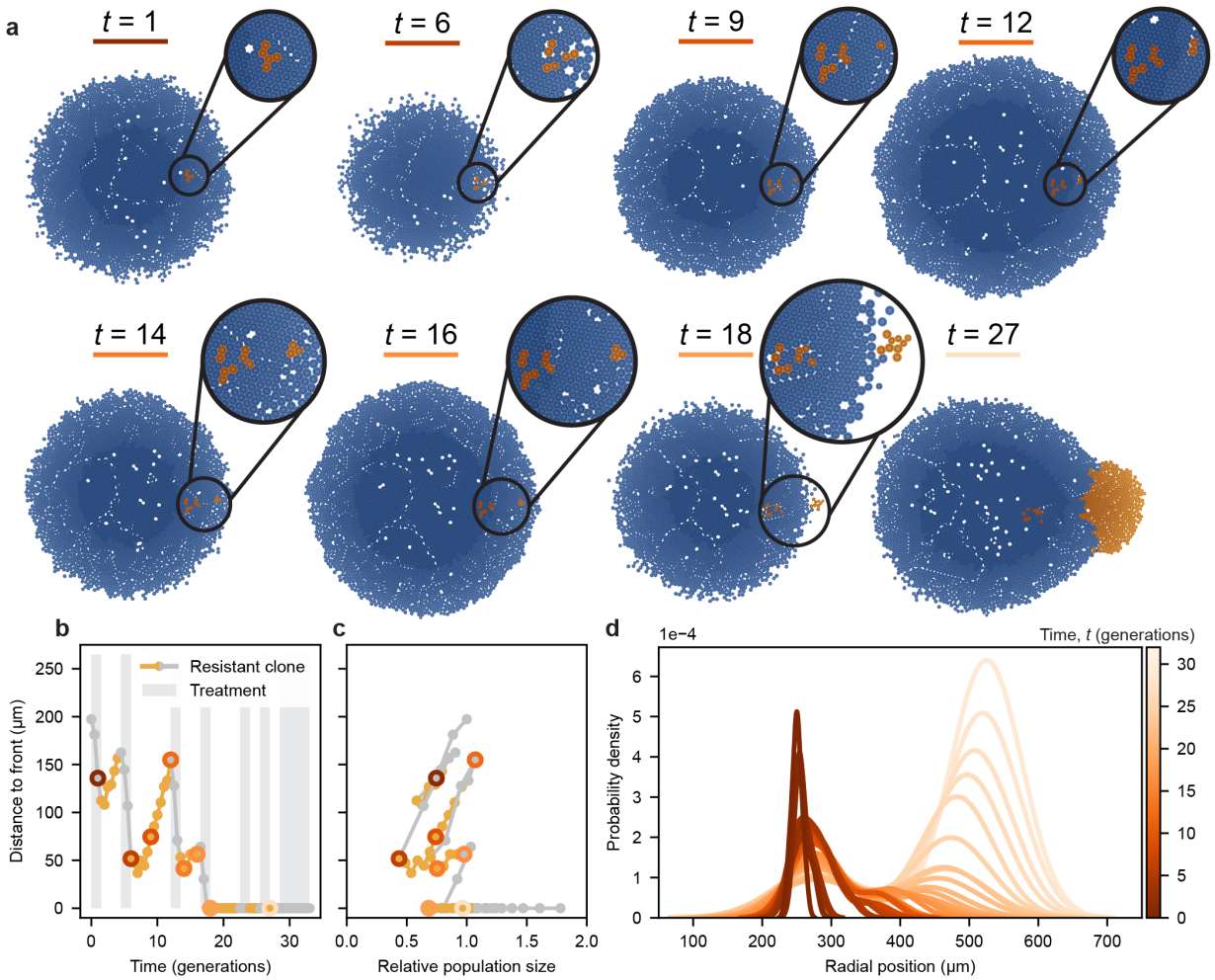



Figure 4: **A mechanical ratchet effect drives the competitive release of resistant cells.** **a** Selected simulation snapshots of the RE under the worst-performing policy. Sensitive cells in blue; resistant cells in orange. Brighter colors indicate larger growth rates (see Supplementary Movie 6). **b** Closest distance from the outer-most resistant cell to the nearest point on the tumor front over time for the simulation run in **a**. Orange lines indicate treatment holidays, gray indicates that treatment was applied. **c** Relationship between the distance to the front and the total population size, relative to the initial tumor burden. Colored circles around markers indicate the time points corresponding to the snapshots in **a**. **d** Radial distribution of resistant cells in time for the same simulation run. Brighter shades represent later time points.

(Fig. 4a). Importantly, once the distance of a resistant cell to the front is sufficiently decreased, even less intensive subsequent therapy cycles continue this process, gradually shifting the cell closer to the tumor front (Fig. 4b, c). This ratchet-like motion progresses in a stepwise but irreversible fashion until resistant cells reach the edge. As their distance to the tumor front decreases, resistant cells experience more favorable growth conditions and begin to expand more rapidly (Fig. 4d).

This interplay between position-dependent growth and mechanically induced motion of resistant cells is not captured by the base training environment. A refinement of the training environment is necessary to enable RL agents to account for these effects.

## Spatial augmentation of the training environment allows agents to avoid hidden therapy pitfalls

To gradually augment the training environment with the most essential spatial dynamics first, we focus on updating the representation of the resistant population while keeping the growth equation for susceptible cells unchanged (Fig. 5a). Two primary candidates for environment augmentation, suggested by the observed ratchet-like motion of resistant cells, are spatially heterogeneous proliferation and growth-induced pushing of cells.

In the reference environment, proliferation rates decline with increasing distance from the tumor edge as a result of nutrient consumption of peripheral cells (Extended Data Fig. 5a). When peripheral cells die in response to therapy, resource availability in the tumor core improves, allowing previously quiescent cells to resume proliferation and contribute to the tumor regrowth. In addition, proliferating cells within the tumor bulk generate mechanical forces that push neighboring cells outward, leading to a cumulative collective displacement that increases towards the front (Extended Data Fig. 5b).

To augmented the training environment accordingly, we first introduce a distance-dependent growth rate for the resistant population, $\gamma_R(d)$, where $d$ denotes the distance of a resistant clone to the tumor front (Fig. 5a). Second, we abstract the mechanically induced outward motion of resistant clones by introducing a distance-dependent velocity term, $v(d)$, that governs the clone's radial displacement toward the tumor edge. The resulting dynamics of the resistant population are described by the following system of equations:

$$\begin{aligned} \frac{dR}{dt} &= \gamma_R(d) \cdot R - \delta \cdot R \\ \frac{dr_R}{dt} &= v(d), \end{aligned} \tag{2}$$

with $d = r_T - r_R$, where $r_R$ is radial position of the resistant clone and the tumor radius is approximated as $r_T \sim \sqrt{S+R}$. Both $\gamma_R(d)$ and $v(d)$ are modeled as quadratically decaying functions of the form $f(d) = b \cdot (a-d)^2$, where parameters $a$ and $b$ are fitted to the position-dependent growth and velocity values measured in the reference environment (see Methods for details). Notice that the position-dependent growth rate in the augmented environment effectively replaces the role of the carrying capacity in the original base training environment.

We test the effects of these augmentations by evaluating the original agents — trained in $\mathrm{TE}_0$ — in the augmented environment and find that individual agent performance in the augmented environment correlates strongly with their performance in the reference environment (Extended Data Fig. 6a). Interestingly, both reference and augmented environments show a biphasic growth pattern that is in line with reports from previous *in vitro* studies (Fig. 5b)[26].

Finally, we train a new cohort of agents in the augmented training environment (Fig. 5c,d) (Extended Data Fig. 6b, Supplementary Movie 7). Cross-environment evaluation via GRAPE now show that these second-generation agents not only converge during training but now also exhibit superior and consistent performance in the reference environment (Fig. 5e).

Analyzing the strategies of these second-generation agent, we find that they consistently avoid treatment at low tumor burden, precluding the ratchet-like forward motion of resistant cells is prevented (Fig. 5c, d; Extended Data Fig. 7). Instead we observe a stable oscillation in clone distance (Fig. 5f, g). This success is further confirmed by the distribution of resistant cells remaining narrowly confined within the population bulk (Fig. 5h).

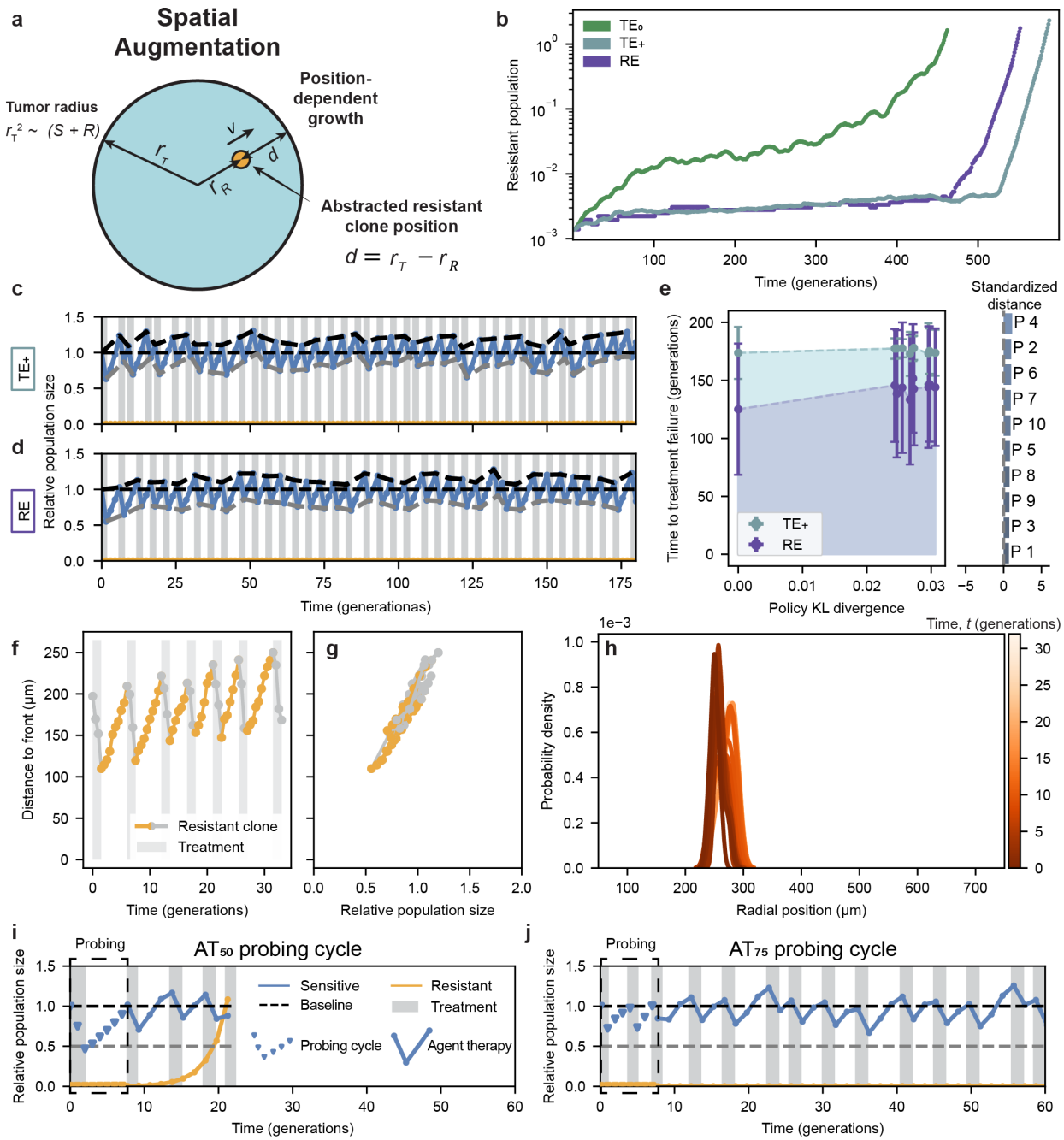



Figure 5: **Augmenting the training environment to reflect spatial dynamics yields superior agent performance.** **a** Schematic of the augmented environment ($TE_+$). **b** Resistant population dynamics over time in the three environments, $TE_0$ (green), $TE_+$, (gray-blue), and RE (purple). **c**, **d** Relative population size dynamics of sensitive (blue) and resistant (orange) cells for a median-performing agent trained in the augmented environment evaluated in both $TE_+$ and RE, respectively (Supplementary Movie 7). Gray bars indicate treatment application. Dashed curves show the tumor burdens at treatment initiation (black) and treatment pause (gray). **e** Time to treatment failure (TTF) in $TE_+$ (gray-blue) and RE (purple) for policies trained in $TE_+$ ($N = 10$). Dots represent averages over $n = 50$ runs, error bars indicate the standard deviation. Policies are spaced on the horizontal axis based on their Kullback-Leibler divergence, using the poorest-performing policy as reference. **f** Closest distance from outer-most resistant cell to the nearest point on the tumor front over time for the simulation run shown in **d**. gray indicates treatment application, orange indicates treatment holidays. **g** Distance of resistant clone to the front as a function of total population size. **h** Radial distribution of resistant cells in time for the same simulation run. Brighter shades represent later time points. **i**, **j** Dynamics of sensitive and resistant populations under the probing cycles for 8 generations of $AT_{50}$ and $AT_{75}$ therapies, respectively, followed by treatment decisions of the best performing agent in $TE_+$.

In addition to uncovering the importance of physical mechanisms for therapy dynamics, our findings point out a potential pitfall for the future design of adaptive therapy trials. An important insight from the adaptive therapy pilot study was the large variation of model parameters obtained from post hoc fitting [2, 3]. To more finely tailor therapy to individual patients, future trials might use an initial "probing cycle" for patient-specific model parameterization [46]. In such a scenario, patients would constitutively receive an initial round of fixed-threshold-therapy (e.g. $AT_{50}$) to gauge the patient-specific tumor response for model parameterization. Tailored models can then be used to suggest optimal personalized scheduling, including RL-based approaches.

However, our findings suggest that such a probing cycle should be applied very carefully, as the associated decrease in tumor burden carries the risk of an unrecoverable competitive release of a pre-existing resistant clone. Fig. 5i shows a trajectory for an initial probing cycle with a 50% reduction in burden, followed by optimal RL-based therapy decisions (TTF $= 26 \pm 14$ generations, $n = 50$). To mitigate the risk, data acquisition for parameterization could be split across multiple cycles of a higher-threshold-strategy (e.g. $AT_{75}$) substantially lowering the probability of premature competitive release (Fig. 5j; TTF$= 147 \pm 52$ generations, $n = 50$).

# Discussion

In this work, we introduce Reinforcement Failing, a framework for the AI-guided, human-in-the-loop refinement of abstracted training environments for complex dynamics, such as tumor therapy. Using cross-environment group-relative agent performance evaluation, Reinforcement Failing helps to expose previously hidden emergent processes that impact therapeutic outcomes. This approach extends the application of RL beyond classic policy optimization toward artificial scientific discovery, offering a systematic way to identify and integrate emergent phenomena into interpretable training environments for enhanced transferability.

Applied to tumor therapy, the framework reveals that position-dependent growth and motion of resistant cells can drive a ratchet-like conveyance of resistant clones to the tumor front. As a result, even short treatment pulses can lead to irreversible loss of resistance control. Physics-informed environment augmentation enables agents to avoid these pitfalls and achieve superior therapeutic performance. Finally, our findings highlighting potential risks of ill-timed probing cycles for patient-specific model parameterization and outline mitigation strategies during translation.

Our findings add a new layer to recent advances in describing the population dynamics of densely packed cellular population, such as tissues and microbial biofilms, as emergent phenomena in actively proliferating granular matter[24, 25, 30, 47]. By tying these phenomena directly to adaptive tumor therapies via RL policy failures, we offer a structured way to incorporate these features into interpretable mathematical models of tumor therapy dynamics — a process that resonates with broader efforts in multi-fidelity and group reinforcement learning to overcome the limitations of "one-size-fits-all" training environments [35, 36, 48].

As with any modeling framework, Reinforcement Failing has inherent limitations that warrant careful consideration. First, detecting missing features depends on sufficient diversity among RL agents; without carefully designed reward functions and training initiation, the ensemble may converge to near-identical strategies, limiting our ability to glean insights on the wider behavior space structure. Second, although we employed a state-of-the-art physics-driven cancer simulator (PhysiCell), real tumors involve many additional factors such as vascular heterogeneity, immune infiltration, and metabolic feedback loops. Third, in a tumor context, the initial positioning and abundance of resistant cells in our reference environment, as well as the assumption of an isolated lesion, does not reflect the complexity encountered clinically [33]. The presented results should therefore be primarily interpreted as a conceptual demonstration of the Reinforcement Failing approach with a limited translational perspective. To overcome these limitations, future work will have to take these biological complexities into account while

pairing them with technical advances of RF, such as diversity-promoting mechanism during agent training.

One key avenue will be the integration of more physiologically detailed simulations that better incorporate features like angiogenesis or the mechanical tumor microenvironment. A particularly exciting avenue of research will be the integration of immunotherapy into the adaptive therapy framework[49].

Looking beyond oncology, employing Reinforcement Failing in other complex systems — such as ecological management[50], autonomous driving[51], or robotics[52] — could similarly reveal hidden emergent processes that hamper straightforward RL-based optimization.

On the artificial scientific discovery level, refining methods that intentionally promote policy diversity during training will help to ensure that the group-level performance analysis yields rich insights into underexplored parts of the decision space. On the long run, combining RL-guided mechanistic discovery with reasoning Large Language Models, akin to the recently introduced algorithmic discovery by AlphaEvolve[53], has the potential to extend exploration to realms not attainable by conventional research.

In summary, Reinforcement Failing reframes the perceived "failure" of RL policies to generalize as a potent discovery mechanism for uncovering cryptic emergent phenomena in high-dimensional, physics-driven environments. By systematically identifying and integrating missing processes, researchers can iteratively refine their mechanistic models and develop more robust, interpretable strategies for disease management. This work thus bridges the gap between realistic tumor behavior and simplified mathematical abstractions, laying a foundation for the RL-guided development of more nuanced therapy strategies. Ultimately, Reinforcement Failing offers a versatile tool for artificial scientific discovery in other domains as well, underscoring the value of embracing failure to achieve deeper insights.

# Methods

## Reference environment

The reference environment is modeled using an off-lattice agent-based cancer model (ABM) implemented in PhysiCell (v1.12.0) [38] with BioFVM (v1.1.6) [54], an open-source state-of-the-art cancer simulation platform [39–42]. The model captures cellular growth, physical interactions, and the diffusion of resources and therapeutics within a 2D domain of 2 × 2 mm, in which individual cells have a radius of 8.41 µm.

Cell proliferation follows a single-phase cycle regulated by oxygen availability, which diffuses from Dirichlet boundary nodes arranged in a circular pattern near the simulation domain's perimeter. Cellular oxygen consumption establishes a radial gradient, restricting proliferation to an outer growth layer (Extended Data Fig. 5e-g), beyond which cells become quiescent. The parameters governing this oxygen-dependent behavior are informed by experimental data from Grimes et al. [55]. A full list of parameters is provided in the shared repository.

Cell mechanics and motion are governed by the inertialess equation of motion ($m_i \vec{a}_i = 0$), as described in Ghaffarizadeh *et al.* [38]. Adhesion and repulsion potentials are modeled using distant-dependent potential functions [54, 56]:

$$\varphi_A = \begin{cases} (1 - \frac{r}{R_A})^2 \frac{\vec{r}}{r}, & \text{if } 0 \leq r \leq R_A \\ 0, & \text{otherwise,} \end{cases}$$

where $\varphi_A$ denotes the adhesion potential, $r$ is the distance between two cells, and $\frac{\vec{r}}{r}$ is the unit vector from one cell to another, and $R_A$ is the maximum adhesive interaction distance.

Similarly, the repulsion potential is defined as:

$$\varphi_R = \begin{cases} (1 - \frac{r}{R})^2 \frac{\vec{r}}{r}, & \text{if } 0 \leq r \leq R \\ 0, & \text{otherwise,} \end{cases}$$

where $\varphi_R$ represents the repulsion potential and $R$ is the cell radius.

The initial tumor environment is generated by growing a tumor from a single cell over 15 generations (Supplementary Movie 1), resulting in a population of 3584 cells. We then locate a compact subpopulation of 5 drug-resistant cells approximately halfway between the tumor center and its surface.

Therapy is simulated by binary switching of Dirichlet boundary conditions, toggling between drug absence and full-dose exposure. Drug effects scales linearly with its concentration, largely increasing the apoptosis rate in proliferating cells. Cells growing slower than 10% of the maximum speed are not affected by the drug. Given the oxygen-dependent proliferation pattern, this results in preferential cell kill at the tumor periphery, where growth and drug concentrations are highest. In the absence of treatment, the baseline cellular turnover rate remains at 1%.

To apply RL policies we wrapped the aforementioned PhysiCell simulations in a custom Gymnasium environment [57]. Communication between the RL environment and the simulation was implemented using ZeroMQ [58], employing local inter-process communication (IPC) on Unix systems and

Transmission Control Protocol (TCP) on Windows (see github repository for a full description).

# Base training environment

For the computationally efficient base training environment, we employ a competition model based on the system of two ordinary differential equations (1), previously well-studied in the context of adaptive therapy [20, 30, 43].

Populations of sensitive $S$ and resistant $R$ cells grow with rates $\gamma_S$ and $\gamma_R$. Their effective proliferation is regulated by the extend to which the total population $S+R$ approaches the carrying capacity $K$. To capture the competitive disadvantage of resistant cells in the early phase of treatment, we introduce an asymmetric competition term $c$, which, if greater than 1, strengthens the suppressive effect of $S$ on $R$. This adjustment is necessary to align the simplified model with the resistant suppression dynamics observed in the reference environment during reference $\mathrm{AT}_{50}$ therapy.

Both populations experience a constant death rate $\delta$, independent of carrying capacity. Treatment is modeled as a fixed depletion $D$ of sensitive cells, representing a constant number of cells reached and killed by the drug. This approximation reflects cell death dynamics in the reference environment when sensitive cells dominate and resistant fraction remains low. The fixed-cell-number killing term is motivated by resource-limited therapy: only cells within the well-perfused growth layer receive sufficient oxygen and drug exposure to be eliminated. In the tumor size regime relevant for this study, this approximation is justified by the approximately constant thickness of the growth layer (see Extended Data Fig. 5e-g). It should be noted, however, that neither this nor the more common proportional-kill formulation fully captures sensitive-cell death dynamics simultaneously across all scenarios, particularly under continuous or adaptive therapy (Fig. 1f,g).

Deviations between the non-spatial base and the spatial reference environments emerge in two key scenarios:
1. Small total population size, where a larger fraction of cells exhibit high growth rates, leading to higher-than-expected drug-induced death.
2. High resistant cell burden, which reduces the number of drug-exposed sensitive cells due to spatial constraints.

These discrepancies suggest potential refinements to the mechanistic mathematical model, incorporating spatial constraints and heterogeneous drug exposure.

# Calibration of the base training environment

We performed Bayesian inference to parametrize the base training environment using the PyMC package v5.9.2 [59]. Markov Chain Monte Carlo (MCMC) sampling was conducted with the Differential Evolution Metropolis (DEMetropolis) algorithm, running 8 parallel chains to ensure robust posterior estimation.

The carrying capacity $K$ was fixed at 18,000 cells, corresponding to the estimated total cell count within the full simulation space. Parameter fitting followed an iterative process to ensure convergence and stability of parameter estimates:
1. The resistant cell growth rate $\gamma_R$ was first fitted to a continuous therapy scenario, keeping all other parameters fixed.
2. The sensitive cell growth rate $\gamma_S$, treatment strength $D$, and competition coefficient $c$ were then fitted to the adaptive therapy ($\mathrm{AT}_{50}$) reference scenario, using the updated $\gamma_R$.
3. This process was repeated iteratively until parameter changes between successive iterations were below 1%, ensuring model stability and accuracy in representing tumor dynamics across both reference therapy regimes.

For the final inference step, we used 2000 tuning steps followed by 1000 posterior draws. Priors were initialized as truncated Gaussian distributions, with mean values guided by previous analyses. Since priors evolved iteratively, we report the final prior means used in the last fitting iteration, which also represent the inferred parameter values used to train the RL agents: $\gamma_S = 0.04$, $\gamma_R = 0.134$, $D = 330$, $c = 4.78$. The inferred parameter values indicate a strong fitness advantage of resistant cells, while also highlighting a significant suppressive effect of sensitive cells on resistant cell growth.

For continuous (CT) and adaptive ($\mathrm{AT}_{50}$) therapies, treatment decisions were made every quarter generation to maintain cell counts near the treatment thresholds. For RL-based agents, this decision frequency was reduced to 1.5 simulated generations to increase clinical feasibility.

# Therapy performance metrics

As a robust metric for therapy performance, we measure the time to therapy failure (TTF), which we define as the time at which resistant cells reach the initial tumor size. TTF is analogous to the commonly used time to progression (TTP), defined via a fixed threshold on total population size relative to the baseline burden. TTF more robustly predicts the potential impact of continued treatment. For example, it precludes the previously reported "premature progression" scenario, in which a trajectory is falsely defined as progressed despite low resistance prevalence, due to a temporary fluctuation in tumor burden[46]. It should be noted, however, that the use of TTF is exclusive to *in silico* scenarios, where the abundance of resistant cells is readily available.

# Reinforcement learning

Treatment decision agents were implemented using Stable Baselines3 [60], a PyTorch-based library providing reliable RL algorithm implementations. We trained the agents using Proximal Policy Optimization (PPO) [61], with identical policy and value networks — both simple MLPs with two hidden layers of 128 units.

To account for history dependence and enable learning beyond simple threshold-based treatment rules, the observation space included stacked past 64 cell counts and treatment decisions, concatenated into a single observation. Hyperparameters (e.g., budget, entropy coefficient, clip range, batch size) remained consistent across agents and are available in the shared

code repository.

Agents were trained for $5 \times 10^6$ environmental steps and underwent $\sim 4,900$ policy update iterations, experiencing approximately $10^5$ full episodes throughout training. The reward function was designed to maximize time to progression while preventing excessive tumor burden. Specifically, agents received a reward of $+1$ per time step if the average tumor size remained below the initial level:

$$r_N = \begin{cases} 1 & \text{if } \sum_i^N \frac{(S_i+R_i)}{N} < (S_0 + R_0) \\ 0 & \text{otherwise.} \end{cases}$$

The base mathematical model and augmented mechanistic model were implemented in a gymnasium environment and solved using Euler's method for training. To enhance agents' policies robustness, Gaussian noise of mean 0 and $\sigma = 0.01 \cdot N$ was added to the calculated cell number $N$ for both resistant and sensitive populations after solving the equations.

To assess the strength of the relationship between agent policy performance and the smallest tumor size at which each agent initiated treatment (Fig. 3d), we computed Pearson's correlation coefficient ($r$) using the NumPy function corrcoef.

On 2 cores of Intel Xeon IceLake Platinum 8360Y processors, training times per agent were approximately 1 hr 58 min in $\text{TE}_0$ and 2 hr 14 min in $\text{TE}_+$ per agent.

## Group relative agent performance evaluation (GRAPE)

GRAPE is a performance evaluation framework based on relative cross-environment validation of agents. A cohort of equivalent agents (differing only in weight initialization) is first trained, and their performance is assessed in both the training and reference environments. This shifts the RL objective from identifying a single best-performing agent to exploring a broader decision space and enabling more reliable policy transfer to complex settings.

Cross-environment performance is quantified using a cross-environment performance difference (CEPD) map, which measures the standardized mean difference (Cohen's $d$) between reference and training environments for each agent:

$$d = \frac{\overline{p}_{RE} - \overline{p}_{TE}}{s},$$

where $\overline{p}$ is the average performance of the agent and $s$ is the pooled standard deviation.

In addition, GRAPE incorporates Kullback–Leibler (KL) divergence to compare policies directly. KL divergence quantifies the behavioral distance between agents, allowing us to determine whether policies achieving similar performance in TE rely on different decision strategies, and to visualize how performance differences between environments shift with changes in agent behavior.

Together, these measures reveal agents with strong or weak cross-environment generalization and help identify decision patterns contributing to poor transfer performance.

## Augmented training environment

To incorporate spatial dynamics into the mechanistic model, we modified the base ODE system (1). The sensitive cell population remained governed by its original equation (1a), while the resistant cell dynamics were updated with a position-dependent growth function (2).

Since oxygen availability defines cell growth rates in the reference environment, we approximated this dependence using distance from the tumor front — a reasonable proxy given that growth layer width remains relatively stable in large colonies (Extended Data Fig. 5e-g). We fitted this growth-distance relationship using free-growth simulations, excluding treatment cycles to avoid artifacts from loss of colony integrity. Similarly, we determined radially outward velocity as a function of cell position.

For position-dependent simulations, we estimated the total tumor radius ($r_\text{T}$) based on total population size, assuming a densely packed tumor structure with cells of radius $r_\text{c}$:

$$r_\text{T}^2 = (S + R) \cdot r_\text{c}^2$$

The resistant population was modeled using an effective radial coordinate $r_R$, evolving over time as:

$$r_R^{t+1} = r_R^t + v(d) \cdot t$$

where velocity depends on distance to tumor front $v = f(d)$.

Position-dependence of the growth rate $\gamma_R$ and of the velocity $v$ both have the form:

$$f(d) = \begin{cases} b \cdot (a - d)^2, & \text{if } d < a \\ 0, & \text{otherwise.} \end{cases}$$

with $a$ and $b$ being fit parameters. Fits to the data from reference environment simulations are presented in Extended Data Fig. 5c,d.

# Data availability

Scripts and data used to produce the figures together with the trained networks are available at `https://github.com/KayserLab/Aif_Reinforcement_Failing`. Reference environment data, including cell positions, are available upon request due to their large size. Patient data from Ref. [3] used in Fig. 1 is available on GitHub (`https://github.com/cunninghamjj/Evolution-based-mathematical-models-significantly-prolong-response-to-Abiraterone-in-mCRPC`).

# Code availability

The code used for simulations, training, and analysis in this study is available at `https://github.com/KayserLab/PhysiLearning`. Scripts for plotting are available at `https://github.com/KayserLab/Aif_Reinforcement_Failing`.

# References


1. Vasan, N., Baselga, J. & Hyman, D. M. A view on drug resistance in cancer. en. *Nature* **575,** 299–309 (Nov. 2019).
2. Zhang, J., Cunningham, J. J., Brown, J. S. & Gatenby, R. A. Integrating evolutionary dynamics into treatment of metastatic castrate-resistant prostate cancer. en. *Nature Communications* **8,** 1816 (Nov. 2017).
3. Zhang, J., Cunningham, J., Brown, J. & Gatenby, R. Evolution-based mathematical models significantly prolong response to abiraterone in metastatic castrate-resistant prostate cancer and identify strategies to further improve outcomes. *eLife* **11** (ed Perry, G. H.) e76284 (June 2022).
4. Gatenby, R. A. & Brown, J. S. Integrating evolutionary dynamics into cancer therapy. en. *Nature Reviews Clinical Oncology* **17,** 675–686 (Nov. 2020).
5. Gatenby, R. A., Brown, J. & Vincent, T. Lessons from Applied Ecology: Cancer Control Using an Evolutionary Double Bind. en. *Cancer Research* **69,** 7499–7502 (Oct. 2009).
6. Enriquez-Navas, P. M. *et al.* Exploiting evolutionary principles to prolong tumor control in preclinical models of breast cancer. en. *Science Translational Medicine* **8,** 327ra24–327ra24 (Feb. 2016).
7. Kim, E., Brown, J. S., Eroglu, Z. & Anderson, A. R. Adaptive Therapy for Metastatic Melanoma: Predictions from Patient Calibrated Mathematical Models. *Cancers* **13** (Feb. 2021).
8. Hockings, H. *et al.* Adaptive therapy achieves long-term control of chemotherapy resistance in high grade ovarian cancer. eng. *bioRxiv: The Preprint Server for Biology,* 2023.07.21.549688 (July 2023).
9. Brady-Nicholls, R. & Enderling, H. Range-bounded adaptive therapy in metastatic prostate cancer. *Cancers* **14,** 5319 (2022).
10. West, J. *et al.* Towards multi-drug adaptive therapy. en. *Cancer Research* (Jan. 2020).
11. West, J., Ma, Y. & Newton, P. K. Capitalizing on competition: An evolutionary model of competitive release in metastatic castration resistant prostate cancer treatment. *Journal of Theoretical Biology* **455,** 249–260 (2018).
12. Papachristou, N. *et al. Digital transformation of cancer care in the era of big data, artificial intelligence and data-driven interventions: navigating the field* in *Seminars in oncology nursing* **39** (2023), 151433.
13. Padmanabhan, R., Meskin, N. & Haddad, W. M. Reinforcement learning-based control of drug dosing for cancer chemotherapy treatment. en. *Mathematical Biosciences* **293,** 11–20 (Nov. 2017).
14. Jalalimanesh, A., Shahabi Haghighi, H., Ahmadi, A. & Soltani, M. Simulation-based optimization of radiotherapy: Agent-based modeling and reinforcement learning. *Mathematics and Computers in Simulation. Biomath 2014 and Biomath 2015* **133,** 235–248 (Mar. 2017).
15. Weaver, D. T., Maltas, J. & Scott, J. G. *Reinforcement Learning informs optimal treatment strategies to limit antibiotic resistance* en. Jan. 2023.
16. Yauney, G. & Shah, P. *Reinforcement Learning with Action-Derived Rewards for Chemotherapy and Clinical Trial Dosing Regimen Selection* en. in *Proceedings of the 3rd Machine Learning for Healthcare Conference* (PMLR, Nov. 2018), 161–226.
17. Huo, L. & Tang, Y. Multi-Objective Deep Reinforcement Learning for Personalized Dose Optimization Based on Multi-Indicator Experience Replay. en. *Applied Sciences* **13,** 325 (Jan. 2023).
18. Eastman, B., Przedborski, M. & Kohandel, M. Reinforcement learning derived chemotherapeutic schedules for robust patient-specific therapy. en. *Scientific Reports* **11,** 17882 (Sept. 2021).

19. Gluzman, M., Scott, J. G. & Vladimirsky, A. Optimizing adaptive cancer therapy: dynamic programming and evolutionary game theory. *Proceedings of the Royal Society B* **287,** 20192454 (2020).

20. Gallagher, K. *et al.* Mathematical Model-Driven Deep Learning Enables Personalized Adaptive Therapy. en. *Cancer Research* **84,** 1929–1941 (June 2024).

21. Lu, Y. *et al.* Deep reinforcement learning identifies personalized intermittent androgen deprivation therapy for prostate cancer. *Briefings in Bioinformatics* **25,** bbae071 (2024).

22. Fusco, D., Gralka, M., Kayser, J., Anderson, A. & Hallatschek, O. Excess of mutational jackpot events in expanding populations revealed by spatial Luria–Delbrück experiments. en. *Nature Communications* **7,** 12760 (Nov. 2016).

23. Kayser, J., Schreck, C. F., Yu, Q., Gralka, M. & Hallatschek, O. Emergence of evolutionary driving forces in pattern-forming microbial populations. *Philosophical Transactions of the Royal Society B: Biological Sciences* **373,** 20170106 (May 2018).

24. Giometto, A., Nelson, D. R. & Murray, A. W. Physical interactions reduce the power of natural selection in growing yeast colonies. en. *Proceedings of the National Academy of Sciences* **115,** 11448–11453 (Nov. 2018).

25. Aif, S., Appold, N., Kampman, L., Hallatschek, O. & Kayser, J. Evolutionary rescue of resistant mutants is governed by a balance between radial expansion and selection in compact populations. en. *Nature Communications* **13,** 7916 (Dec. 2022).

26. Bacevic, K. *et al.* Spatial competition constrains resistance to targeted cancer therapy. en. *Nature Communications* **8,** 1995 (Dec. 2017).

27. Noble, R. *et al.* Spatial structure governs the mode of tumour evolution. en. *Nature Ecology & Evolution,* 1–11 (Dec. 2021).

28. Fu, X. *et al.* Spatial patterns of tumour growth impact clonal diversification in a computational model and the TRACERx Renal study. en. *Nature Ecology & Evolution* **6,** 88–102 (Jan. 2022).

29. Sottoriva, A. *et al.* A Big Bang model of human colorectal tumor growth. en. *Nature Genetics* **47,** 209–216 (Mar. 2015).

30. Strobl, M. A. R. *et al.* Spatial structure impacts adaptive therapy by shaping intra-tumoral competition. en. *Communications Medicine* **2,** 46 (Apr. 2022).

31. Gallaher, J. A., Enriquez-Navas, P. M., Luddy, K. A., Gatenby, R. A. & Anderson, A. R. A. Spatial Heterogeneity and Evolutionary Dynamics Modulate Time to Recurrence in Continuous and Adaptive Cancer Therapies. en. *Cancer Research* **78,** 2127–2139 (Apr. 2018).

32. Maltas, J. *et al.* Drug dependence in cancer is exploitable by optimally constructed treatment holidays. en. *Nature Ecology & Evolution* (Nov. 2023).

33. Gallaher, J. *et al.* Intermetastatic and Intrametastatic Heterogeneity Shapes Adaptive Therapy Cycling Dynamics. *Cancer Research* **83,** 2775–2789 (Aug. 2023).

34. West, J., Schenck, R. O., Gatenbee, C., Robertson-Tessi, M. & Anderson, A. R. A. Normal tissue architecture determines the evolutionary course of cancer. en. *Nature Communications* **12,** 2060 (Apr. 2021).

35. Agrawal, A. & McComb, C. Adaptive Multi-Fidelity Reinforcement Learning for Variance Reduction in Engineering Design Optimization. *arXiv preprint arXiv:2503.18229* (2025).

36. Agrawal, A. & McComb, C. Reinforcement learning for efficient design space exploration with variable fidelity analysis models. *Journal of Computing and Information Science in Engineering* **23,** 041004 (2023).

37. Metzcar, J., Jutzeler, C. R., Macklin, P., Köhn-Luque, A. & Brüningk, S. C. A review of mechanistic learning in mathematical oncology. *Frontiers in Immunology* **15,** 1363144 (2024).

38. Ghaffarizadeh, A., Heiland, R., Friedman, S. H., Mumenthaler, S. M. & Macklin, P. PhysiCell: An open source physics-based cell simulator for 3-D multicellular systems. en. *PLOS Computational Biology* **14,** e1005991 (Feb. 2018).

39. Ma, C. & Gurkan-Cavusoglu, E. A comprehensive review of computational cell cycle models in guiding cancer treatment strategies. *NPJ Systems Biology and Applications* **10,** 71 (2024).

40. L. Rocha, H., Aguilar, B., Getz, M., Shmulevich, I. & Macklin, P. A multiscale model of immune surveillance in micrometastases gives insights on cancer patient digital twins. *npj Systems Biology and Applications* **10,** 144 (2024).

41. Harkos, C. *et al.* Using mathematical modelling and AI to improve delivery and efficacy of therapies in cancer. *Nature Reviews Cancer,* 1–17 (2025).

42. Oliver, S., Williams, M., Jolly, M. K., Gonzalez, D. & Powathil, G. Exploring the role of EMT in ovarian cancer progression using a multiscale mathematical model. *npj Systems Biology and Applications* **11,** 36 (2025).

43. Strobl, M. A. *et al.* Turnover Modulates the Need for a Cost of Resistance in Adaptive Therapy. en. *Cancer Research* **81,** 1135–1147 (Feb. 2021).

44. Viossat, Y. & Noble, R. A theoretical analysis of tumour containment. en. *Nature Ecology & Evolution,* 1–10 (Apr. 2021).

45. Buhler, C. K., Terry, R. S., Link, K. G. & Adler, F. R. Do mechanisms matter? Comparing cancer treatment strategies across mathematical models and outcome objectives. *Mathematical biosciences and engineering: MBE* **18,** 6305 (2021).

46. Gallagher, K., Strobl, M., Gatenby, R., Maini, P. & Anderson, A. *Learning to Adapt - Deep Reinforcement Learning in Treatment-Resistant Prostate Cancer* en. Apr. 2023.

47. Kayser, J., Schreck, C. F., Gralka, M., Fusco, D. & Hallatschek, O. Collective motion conceals fitness differences in crowded cellular populations. en. *Nature Ecology & Evolution* **3,** 125–134 (Jan. 2019).

48. Guo, D. *et al.* Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948* (2025).

49. West, J. *et al.* A survey of open questions in adaptive therapy: Bridging mathematics and clinical translation. eng. *eLife* **12,** e84263 (Mar. 2023).

50. Lapeyrolerie, M., Chapman, M. S., Norman, K. E. & Boettiger, C. Deep reinforcement learning for conservation decisions. *Methods in Ecology and Evolution* **13,** 2649–2662 (2022).

51. Li, D. & Okhrin, O. A platform-agnostic deep reinforcement learning framework for effective sim2real transfer towards autonomous driving. *Communications Engineering* **3,** 147 (2024).

52. Ju, H., Juan, R., Gomez, R., Nakamura, K. & Li, G. Transferring policy of deep reinforcement learning from simulation to reality for robotics. *Nature Machine Intelligence* **4,** 1077–1087 (2022).

53. Novikov, A. *et al.* AlphaEvolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131* (2025).

54. Ghaffarizadeh, A., Friedman, S. H. & Macklin, P. BioFVM: an efficient, parallelized diffusive transport solver for 3-D biological simulations. *Bioinformatics* **32,** 1256–1258 (Apr. 2016).

55. Grimes, D. R., Kelly, C., Bloch, K. & Partridge, M. A method for estimating the oxygen consumption rate in multicellular tumour spheroids. *Journal of the Royal Society Interface* **11,** 20131124 (Mar. 2014).

56. Macklin, P., Edgerton, M. E., Thompson, A. M. & Cristini, V. Patient-calibrated agent-based modelling of ductal carcinoma in situ (DCIS): from microscopic measurements to macroscopic predictions of clinical progression. *Journal of theoretical biology* **301,** 122–140 (2012).

57. Towers, M. *et al.* Gymnasium: A standard interface for reinforcement learning environments. *arXiv preprint arXiv:2407.17032* (2024).

58. Hintjens, P. *ZeroMQ: messaging for many applications* (" O'Reilly Media, Inc.", 2013).

59. Abril-Pla, O. *et al.* PyMC: a modern, and comprehensive probabilistic programming framework in Python. en. *PeerJ Computer Science* **9,** e1516 (Sept. 2023).

60. Raffin, A. *et al.* Stable-Baselines3: Reliable Reinforcement Learning Implementations. *Journal of Machine Learning Research* **22,** 1–8 (2021).

61. Schulman, J., Wolski, F., Dhariwal, P., Radford, A. & Klimov, O. *Proximal Policy Optimization Algorithms* tech. rep. arXiv:1707.06347 (arXiv, Aug. 2017).

62. McGehee, C. & Mori, Y. A mathematical framework for comparison of intermittent versus continuous adaptive chemotherapy dosing in cancer. *npj Systems Biology and Applications* **10,** 140 (2024).

# Acknowledgements

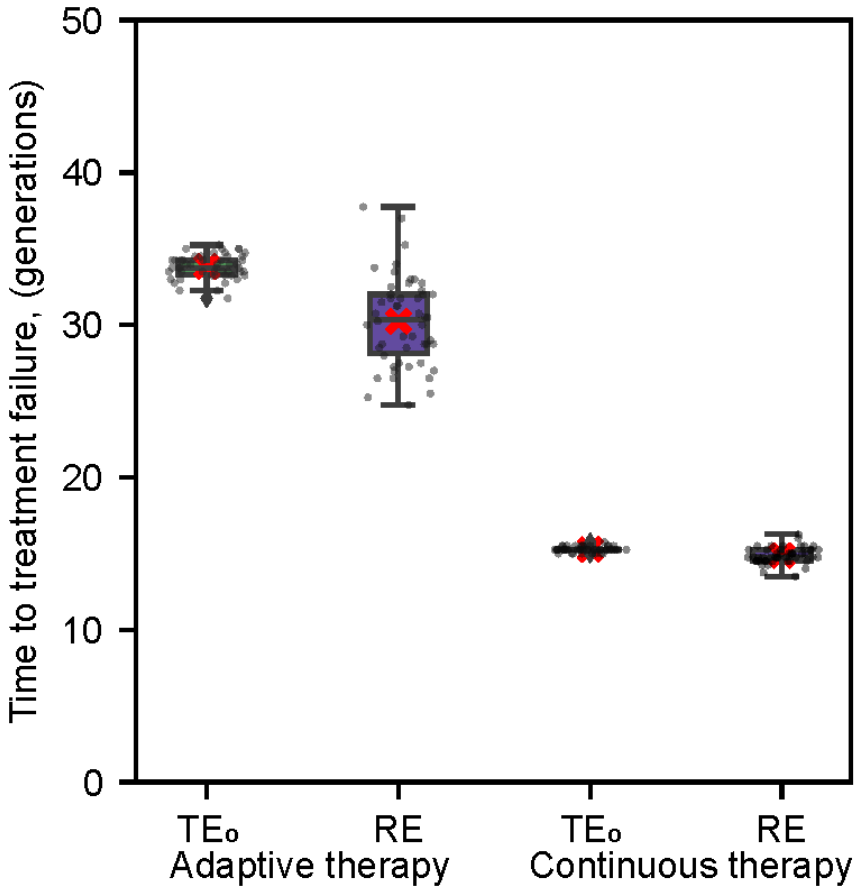
This work was supported by the Emmy Noether Programme of the German Research Foundation (project 455449456). ME acknowledges support from the German Academic Scholarship Foundation. The authors thank M. Dauber and L. Strampe for valuable discussions, I. Schneider for reviewing the manuscript, and J. Guck and his group for their invaluable support. We also acknowledge the open-source community, especially the contributors to Stable-Baselines3 and PhysiCell.

# Author contributions

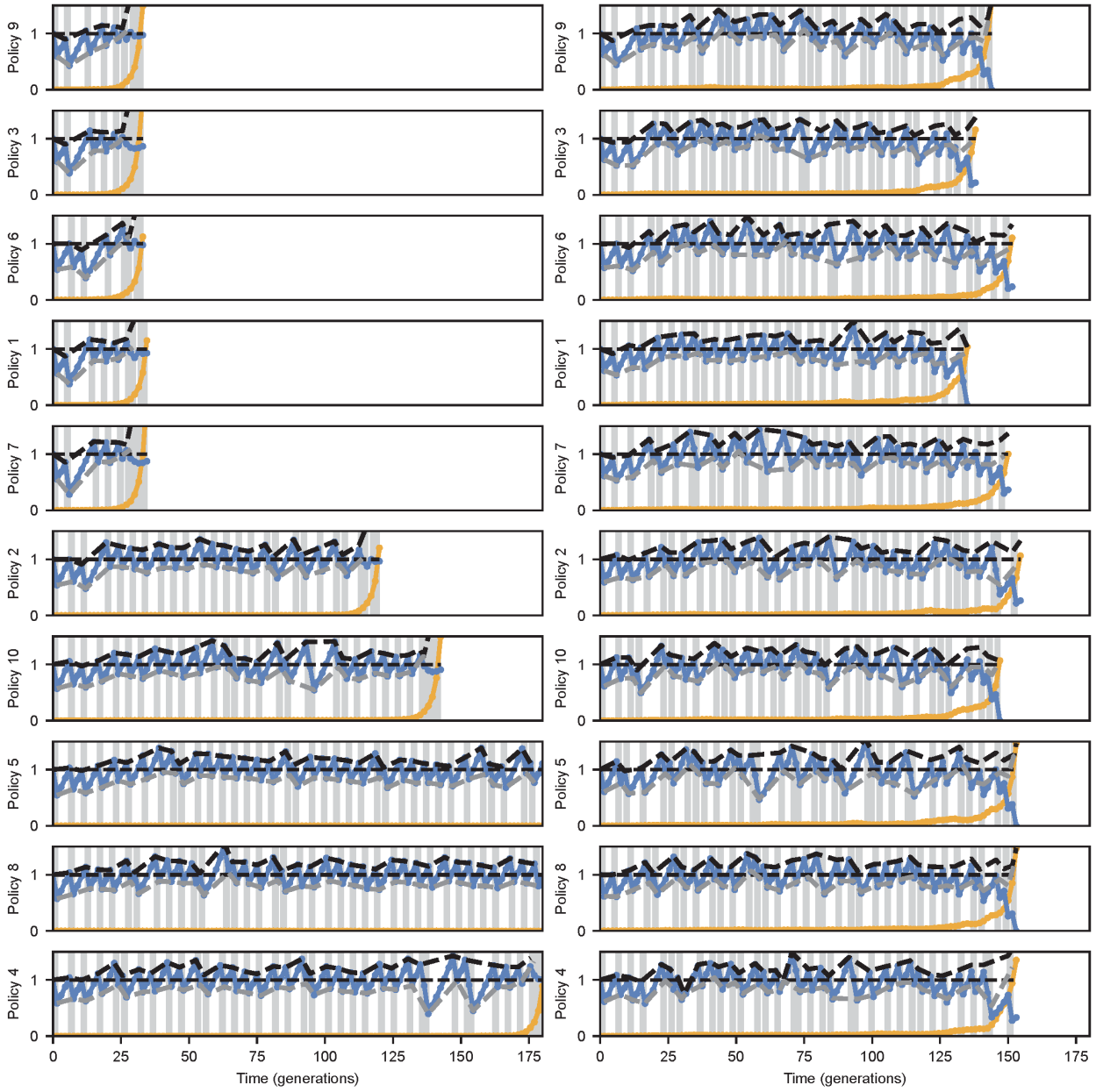
S.A., and J.K. conceived and designed the study. S.A. performed the simulations and data analysis. S.A. and J.K. wrote the manuscript. S.A., M.E., N.A., E.F., T.C. and J.K. discussed and interpreted the results, prepared figures and contributed to the final revision of the manuscript.

# Competing interests

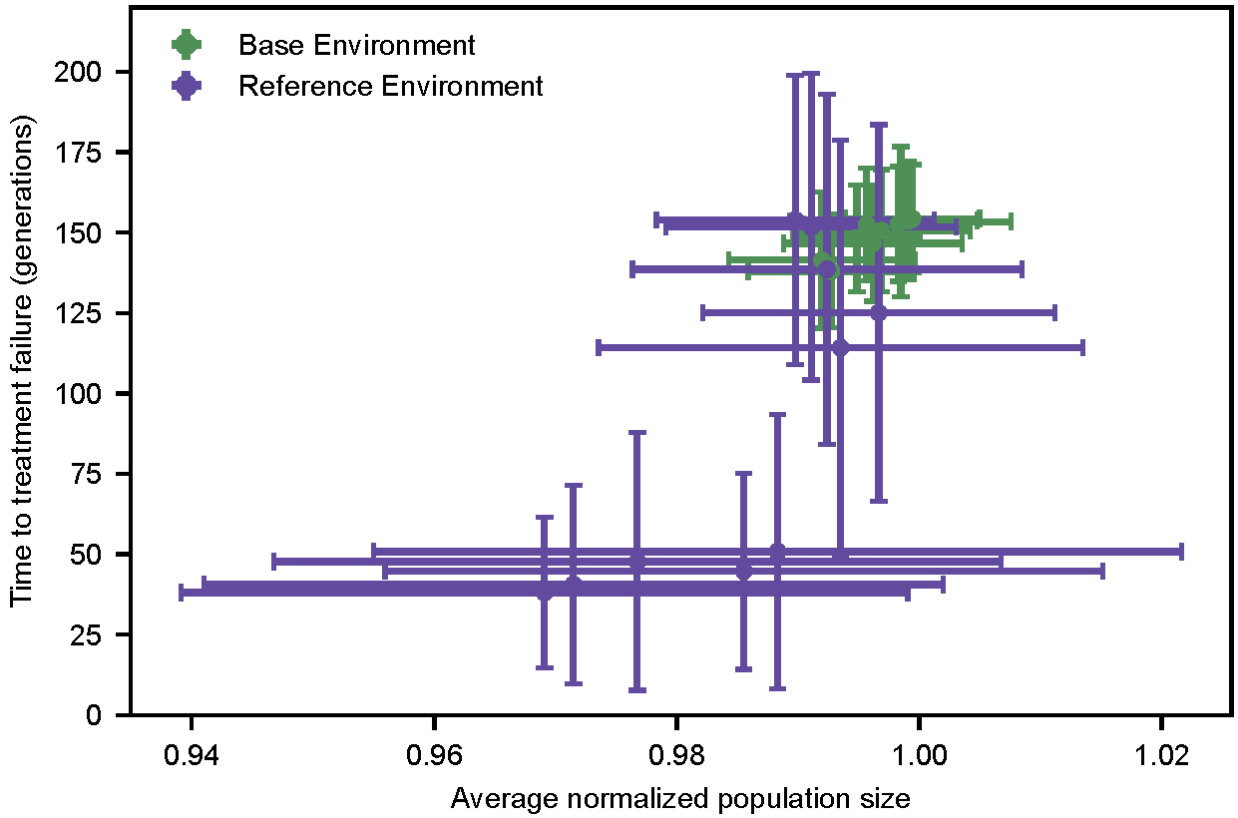
The authors declare no competing interests.



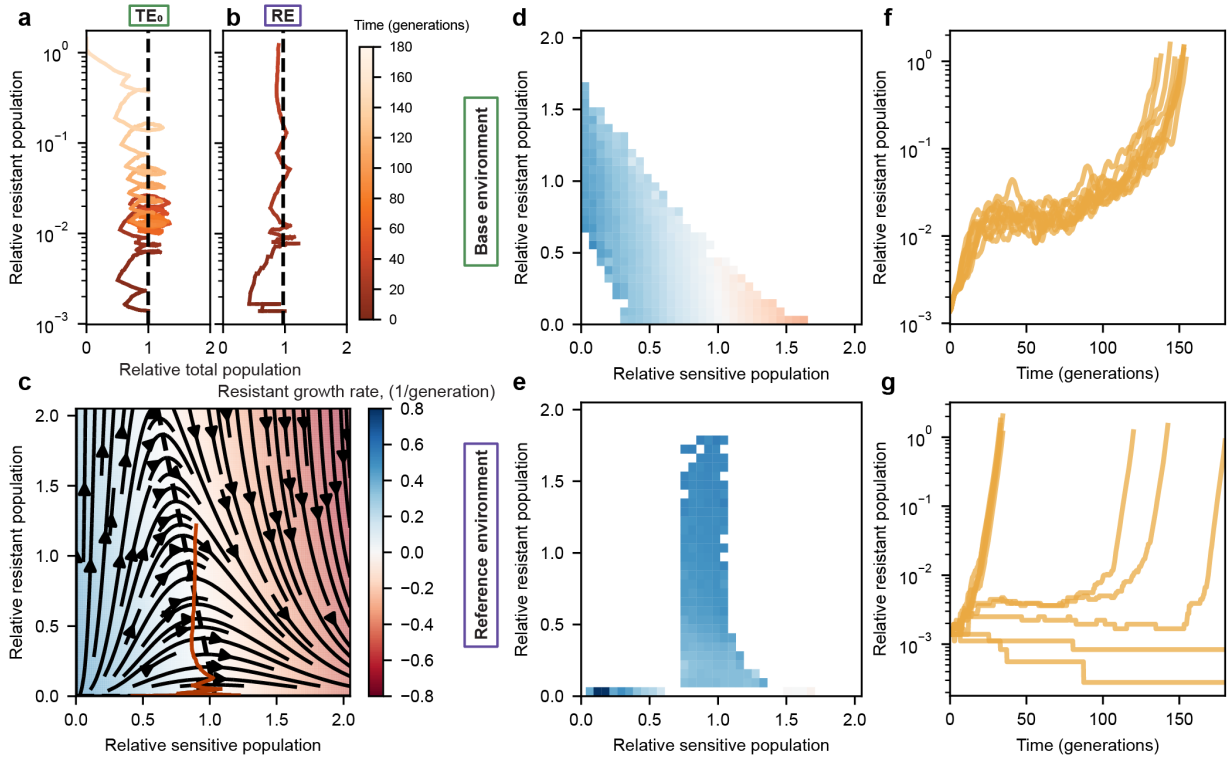

**Extended Data Figure** 1: **Reference therapies: Continuous Therapy (CT) and Adaptive Therapy ($AT_{50}$ in base ($TE_0$) and reference (RE) environments.** Figure shows the time to treatment failure (TTF) distribution for the CT and $AT_{50}$ therapy protocols applied to both the base and reference environments. Individual dots represent individual simulation runs, while the boxplot displays the median and quartiles. The whiskers extend to the rest of the distribution, and the red cross indicates the mean.



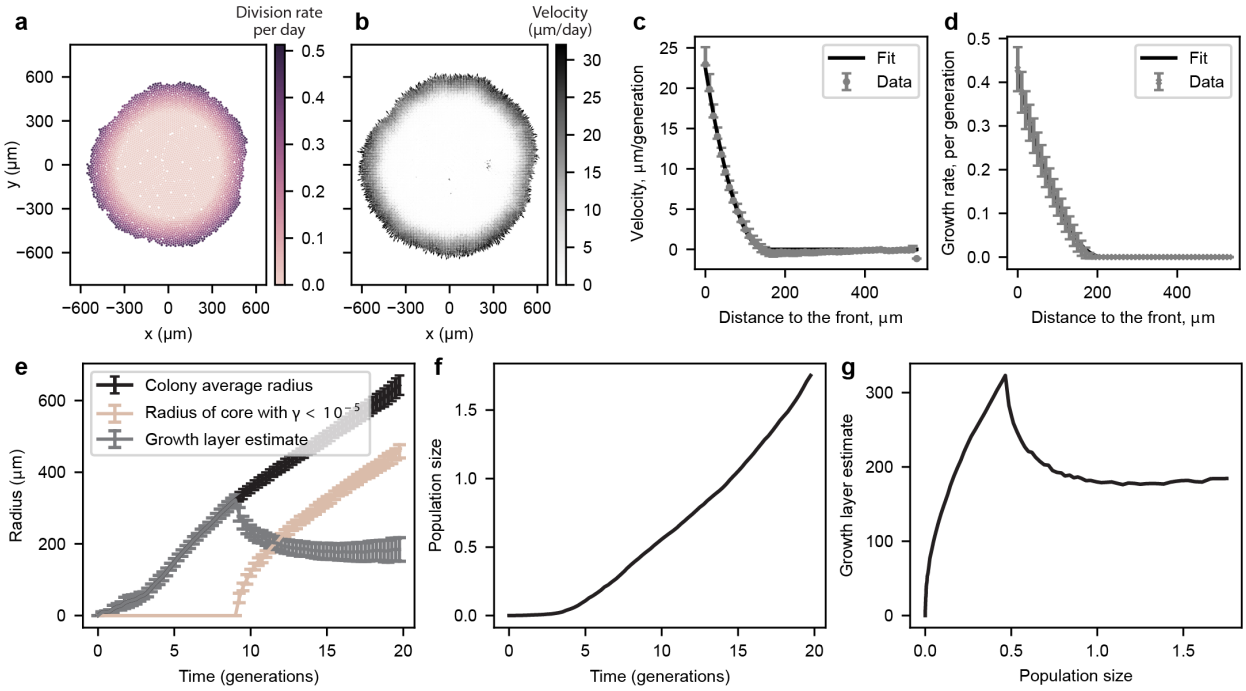

**Extended Data Figure 2: Sensitive and resistant population dynamics under agents policies trained in the base environment.** Dynamics of the sensitive (blue) and resistant (orange) population sizes - relative to the total initial population size - under all n=10 policies, which were trained in the base environment ($TE_0$). On the left - evaluation runs in the reference environment (RE), on the right - evaluation runs in the base training environment ($TE_0$). Grey regions indicate treatment application. Dashed lines show the tumor burdens at treatment initiation (black) and treatment pause (grey). Each run corresponds to a median time to failure of the evaluated policy.



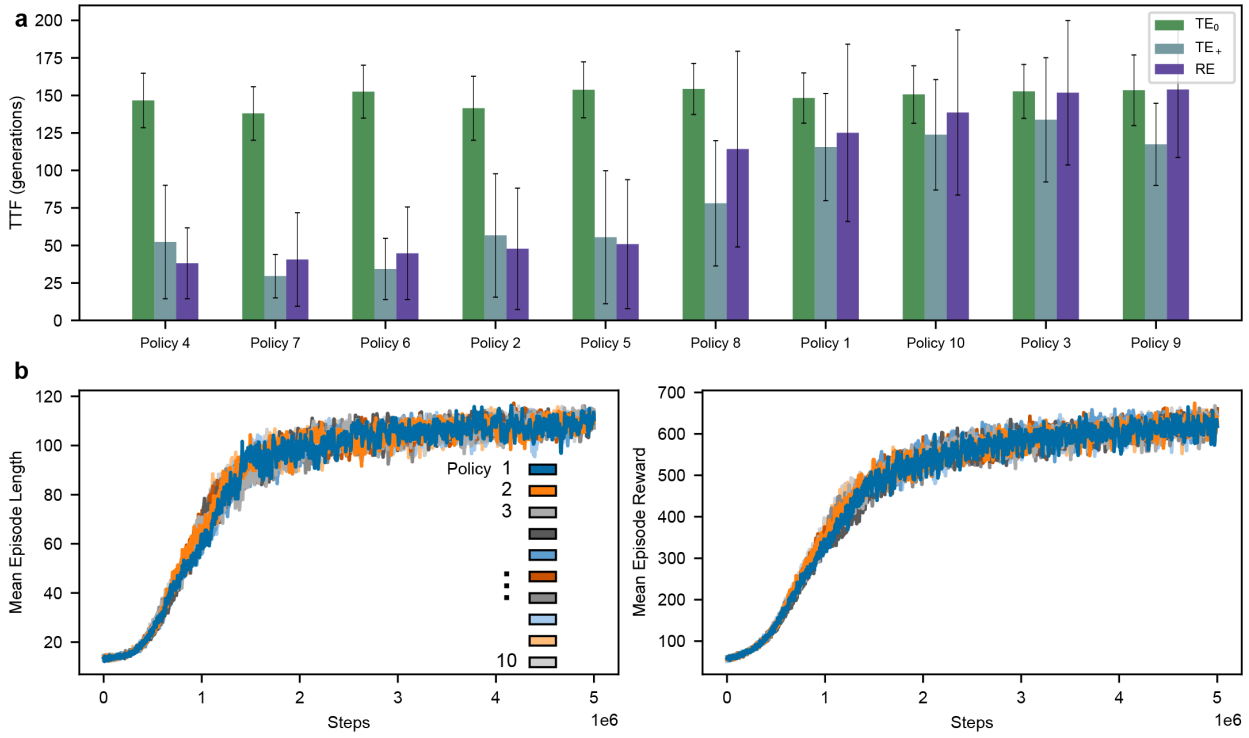

**Extended Data Figure** 3: Time to treatment failure (TTF) in $TE_0$ (green) and RE (violet) for n=10 policies trained in $TE_0$ as the function of the average total population size. Dots represent averages over n=50 runs, and error bars indicate the standard deviation.



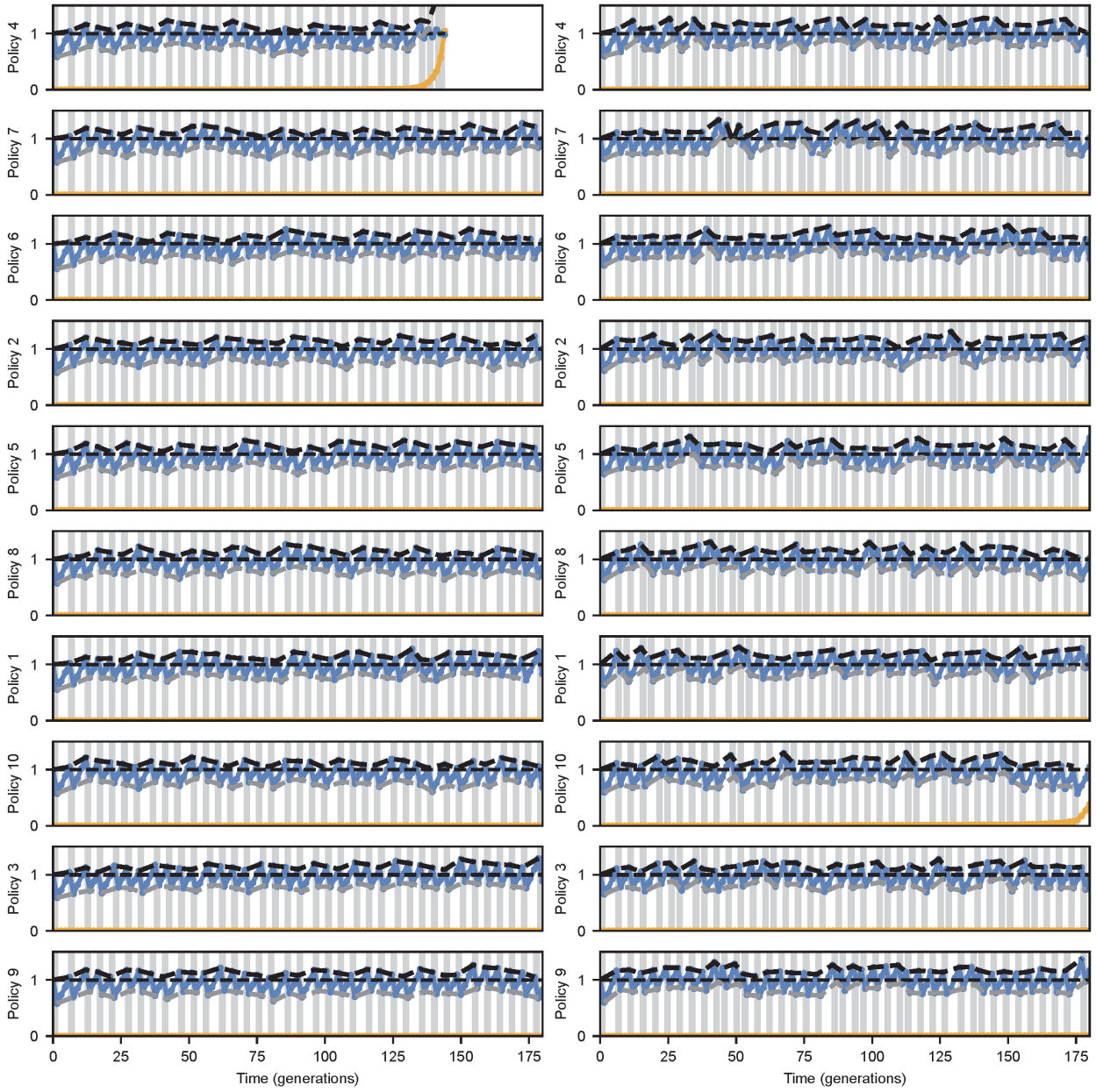

**Extended Data Figure** 4: **Control of resistant population in the reference environment is not governed by sensitive cell numbers. a**, **b** Relationship between resistant and total population sizes over time (color) for the base ($TE_0$) and reference (RE) environments during therapy under the worst policy. Dashed lines indicate the average cell number kept by the policy. **c** Phase space of resistant vs. sensitive populations in $TE_0$. Streamlines represent trajectories in time under free growth, orange line shows the trajectory of the therapy run in RE, presented in **b**. Colors indicate the effective growth rate of the resistant population per generation. **d**, **f** Regions of the phase space explored by the n=10 agents over n=50 evaluation runs on $TE_0$ and RE. The color scale from **c** is shared by **d** and **f**. **e**, **g** Resistant population dynamics over time for the evaluation run corresponding to the median TTF across all n=10 agents for $TE_0$ and RE. All population sizes are shown relative to the initial tumor burden.




**Extended Data Figure** 5: **Position-dependent growth and motion of the cells.** **a** Single-time snapshot of the 2-D growth rate distribution, where brightness decreases with growth rate. **b** Average velocity of cells at certain positions over 3 generations under free growth (n=50 runs). Arrow magnitude is proportional to velocity. **c**, **d** Fit of position dependent velocity and growth rate for the augmented environment. Data from the reference environment averaged over n=50 simulation runs over 3 generations of free growth is shown in grey. Solid black line represents the fit of a truncated quadratic function $f(d) = b(a-d)^2$. **c** Error bars represent 25th and 75th quantiles. **d** Error bars show standard deviations. **e** Dynamics of the estimated tumor radius (black), the radius of the non-growing ($\gamma < 10^{-5}$) core inside the tumor (beige), and the growth layer (grey), defined as the difference between the two, over time. The simulation is performed from a single cell seeded in the center of the simulation space over 20 generations. **f** Dynamics of the total population size, normalized to the initial population size used in training. **g** Relationship between the growth layer width and the relative population size.




**Extended Data Figure** 6: **a** Average time to treatment failure (TTF) achieved by the RL agent with policies trained in $TE_0$, evaluated in $TE_0$ (green), $TE_+$ (blue-green) and RE (violet). n=50 runs, error bars indicate the standard deviation. **b** Mean episode length and mean episode reward for each individual training run in the augmented environment as a function of simulated steps. One steps represents one treatment decision point and corresponds to 1.5 simulated generations. Policy updates are performed every 1024 steps.




**Extended Data Figure 7**: **Sensitive and resistant population dynamics under agents policies trained in the augmented environment.** Dynamics of the sensitive (blue) and resistant (orange) population sizes - relative to the total initial population size - under all n=10 policies, which were trained in the augmented training environment ($TE_+$). On the left - evaluation runs in the reference environment (RE), on the right - evaluation runs in the augmented training environment ($TE_+$). Grey regions indicate treatment application. Dashed lines show the tumor burdens at treatment initiation (black) and treatment pause (grey). Each run corresponds to a median time to failure of the evaluated policy.