GRAPEVINE: A Reinforcement Failing Framework for AI-Guided Discovery of Hidden Spatial 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-University Erlangen-Nürnberg, 91054 Erlangen, Germany
[*]E-mail: jona.kayser@mpzpm.mpg.de

March 31, 2025

Abstract

Adaptively modulated cancer therapies, integrating evolutionary models of intra-tumor competition with clinical surveillance, have emerged as a promising new approach to delay resistance-mediated therapy failure in metastatic malignancies, a primary cause of cancer related deaths. Yet, how to integrate the highly-complex, physics-driven collective dynamics shaping solid tumor growth and therapy response into this approach remains unclear. Reinforcement learning (RL) offers a powerful new paradigm for strategy optimization in such highly-complex environments. However, discovery and abstraction of emergent mechanism, often the key to the successful training of decision-making agents, represents a critical hurdle. Here, we present Group-Relative Agent Policy Evaluation for Varying INterpretable Environments (GRAPEVINE), an artificial scientific discovery framework for uncovering hidden emergent processes. At its core, GRAPEVINE builds on Reinforcement Failing (RF), a novel concept that systematically examines the failure modes of an ensemble of independently trained agents. We show how the resulting RL-guided, human-in-the-loop augmentation of the training environment not only boosts agent performance but additionally allows for the injection of existing scientific knowledge while maintaining mechanistic interpretability of key dynamics. We demonstrate the power of this approach in the context of adaptive tumor therapy by uncovering the critical role of position-dependent growth and mechanically-induced motion of resistant mutants. Together, these findings illustrate the potential of GRAPEVINE as a powerful artificial scientific discovery engine for deciphering high-complexity processes in personalized cancer treatment and beyond.

Introduction

Despite substantial advances in modern cancer treatment — including targeted therapies and immunotherapies — the emergence of therapy resistance remains one of the most urgent challenges in modern oncology, calling for innovative new approaches [1]. Evolution-based therapy strategies have recently emerged as a promising new paradigm that seeks to delay resistance-mediated failure by exploiting the competition between drug-sensitive and drug-resistant cancer cell populations [2, 3]. Combining mathematical models of cancer evolution with longitudinal tumor surveillance, these so-called adaptive therapies (AT) have already shown encouraging results in preclinical and early clinical settings [4–8].

However, current implementations of adaptive therapy do not yet incorporate essential biophysical characteristics of solid tumors, ranging from spatial population structure and environmental gradients to physical cell–cell interactions [9]. These interactions can drive emergent collective dynamics that fundamentally alter evolutionary trajectories and treatment response. Recent computational studies using cell-based simulations, alongside *in vitro* experiments in dense microbial and cancer cell populations, have underscored the importance of these physical mechanisms for resistance evolution in general, and the potential gains of adaptive scheduling in particular [10–21].

These findings reinforce the view that solid tumors must be understood as actively proliferating, granular matter systems in which evolutionary outcomes are emergent properties of spatial and mechanical constraints [22]. Consequently, designing effective adaptive therapies demands a mechanistic understanding of these collective effects. Yet, the very complexity required to accurately capture such dynamics renders traditional modeling approaches opaque and hampers intuitive strategy design.

Reinforcement learning (RL) presents a powerful solution to this challenge. By learning to optimize sequential decision-making strategies through interaction with an environment, RL has been successfully applied to domains ranging from robotics to large language models [23, 24]. In oncology, RL is increasingly used to optimize treatment regimens, including in radiotherapy or chemotherapy [25–30]. Notably, Gallagher *et al.* recently demonstrated that RL can outperform fixed-threshold adaptive therapy strategies in simulated patients by learning more nuanced dosing policies [31].

However, applying RL in complex tumor environments remains computationally demanding and conceptually challenging. Success often hinges on the careful abstraction of essential system features into simplified training models. Without principled guidance, this abstraction process becomes a bottleneck, potentially omitting key dynamics and lacking mechanistic interpretability [32]. The nascent field of multi-fidelity RL has begun to address these challenges by pairing low-fidelity training environments with policy evaluations in high-fidelity simulations [33, 34]. While this approach can improve training efficiency and performance of the learned strategy in the high-fidelity environment, it relies on a predesigned set of multi-fidelity environments.

Since, *a priori*, it is not known which emergent features might drive therapeutic outcomes in a tumor context, we argue that multi-fidelity RL must be extended to the systematic probing and dynamic augmentation of mechanistically interpretable training environments. To this end, we introduce Group-Relative Agent Policy Evaluation for Varying INterpretable Environments (GRAPEVINE). This human-in-the-loop RL framework builds on Reinforcement Failing (RF), a novel concept that leverages group-level variation in agent performance to expose hidden system features that critically shape therapeutic success in a high-fidelity tumor simulation. Using GRAPEVINE, we discover position-dependent growth and mechanically-induced cell motion to be key drivers of adaptive therapy outcomes. Incorporating these features into the training environment significantly enhances agent performance while facilitating a detailed understanding of the underlying mechanisms by human researchers.

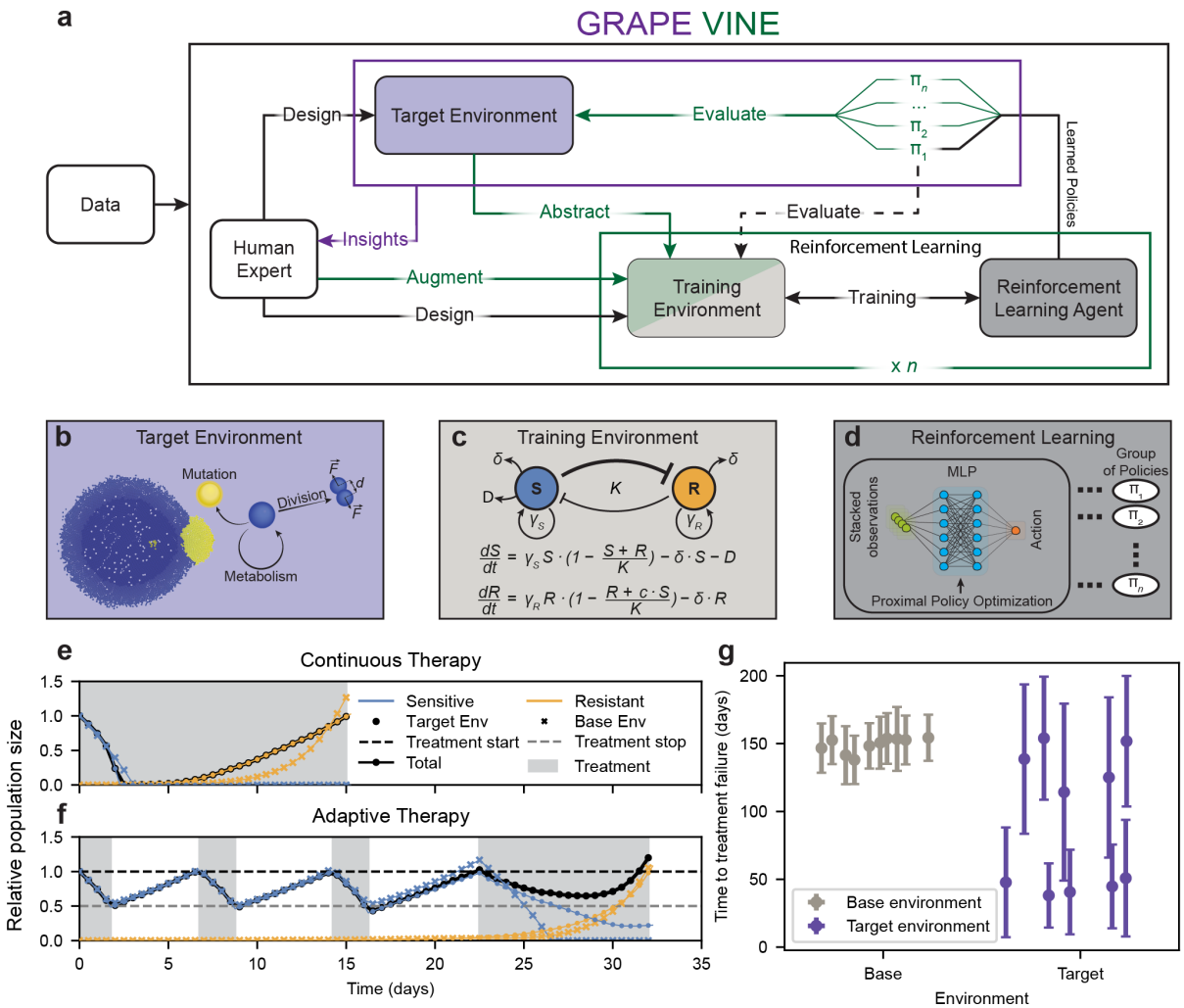

Figure 1: **The GRAPEVINE Framework: A Systematic Approach to Feature Discovery via Reinforcement Learning**. **a** The overall GRAPEVINE framework, with **b** depicting the target environment (TE), **c** - the training base environment (BE) and **d** - Reinforcement Learning (see Methods for details). **e**, **f** Dynamics of the sensitive (blue) and resistant (orange) population sizes, relative to the total initial population size, in TE (dots) and fitted BE (crosses), under continuous therapy CT **e** (Supplementary Movie 2) and adaptive therapy AT_{50} **f** (Supplementary Movie 3). Grey regions indicate treatment application. Dashed lines show the tumor burden thresholds for treatment initiation (black) and treatment pause (grey). **g** Time to treatment failure (TTF) in BE (grey) and TE (green) for n=10 policies trained in BE. Dots represent averages over n=50 runs, and error bars indicate the standard deviation.

Results

A major challenge in applying adaptive therapy is integrating the complex physical and biological properties of solid tumors into treatment optimization strategies. Spatial population structure, mechani-

cal cell-cell interactions, and emergent collective dynamics shape tumor evolution and therapy response in ways that remain difficult to capture in standard reinforcement learning (RL) training environments. To address this, we introduce GRAPEVINE (Group-Relative Agent Policy Evaluation for Varying INterpretable Environments), a structured framework designed to iteratively refine RL training environments by identifying missing emergent features. We will first explain the core concepts of the GRAPEVINE pipeline on a general level before demonstrating it on the concrete example of adaptive tumor therapy.

The GRAPEVINE framework: A systematic approach to feature discovery via Reinforcement Learning

The GRAPEVINE framework consists of two interlinked parts, GRAPE and VINE (Fig. 1 a). It operates by systematically comparing the performance of multiple RL agents across two environments: a high-fidelity (HiFi) target environment, which captures physical complexities of the system, and a low-fidelity (LoFi) training environment, which serves as a simplified but computationally efficient representation. Typically, the target environment will be represented by a detailed *in silico* simulation of a real-world system, often referred to as a digital twin. The training environment is a coarse-grained mathematical abstraction of system dynamics that is designed by the researchers based on *a priori* assumptions and - if available - real-world data. In addition to its computational efficiency, the mathematical representation facilitates the interpretation by human researchers.

Agents are first trained in the LoFi environment, allowing them to discover policies that optimize performance under the given model assumptions. These trained agents are then evaluated in the HiFi environment, providing a means to assess whether strategies that were successful in a simplified setting remain effective under more realistic conditions.

A key innovation of the framework is the use of Group-Relative Agent Policy Evaluation (GRAPE) to uncover missing emergent features. Instead of relying on a single policy, GRAPEVINE trains a diverse cohort of agents, each of which learns a distinct but similarly successful policy π_n in the training environment. Depending on the reward landscape, policy diversity can either develop spontaneously during stochastic training or be induced via a diversity-encouraging reward term. By deploying this ensemble of agents in the target environment, the framework moves beyond absolute performance metrics and instead examines the spread in agent performance relative to the group. If all policies remain comparably effective when evaluated in the target environment, this suggests that the training environment adequately captured the key mechanisms. However, if agent performance diverges significantly in the target environment, despite similar success in training, this indicates the presence of emergent effects that were not accounted for during training. We term this approach "Reinforcement Failing," leveraging the hidden insights found in the seemingly failing behavior of RL agents. Notably, by actively exploring the system's decision landscape, GRAPE can expose previously overlooked but outcome-critical interactions, revealing mechanistic dependencies that were not evident through conventional model-driven reasoning alone.

To systematically integrate these missing features into the training environment, the VINE part of the framework employs a human-in-the-loop augmentation step. Guided by the gradient in observed agent performances, a human expert can identify potential underlying mechanisms by comparing patterns in agent behavior in the context of existing scientific knowledge. Based on these insights, the initial base training environment is augmented to incorporate the newly identified emergent features while preserving computational efficiency. The process can then be repeated with a new set of RL agents trained under the updated conditions, allowing for progressive model refinement until cross-environment-performance in training and target environments converge.

Through this iterative feedback loop, GRAPEVINE enables the discovery and systematic integration of critical emergent properties into RL-based strategy optimization. By leveraging group-level policy evaluation, the framework ensures that reinforcement learning remains inter-

pretable and grounded in mechanistic understanding, bridging the gap between computational strategy optimization and the complexity of many real-world scenarios, including tumor growth and therapy response.

Using GRAPEVINE to improve adaptive therapies for solid tumors

To illustrate the power of GRAPEVINE, we apply it to the optimization of adaptive cancer therapy. To model the complex, physics-driven tumor dynamics underlying therapy response, we construct a target environment (TE) based on the state-of-the-art PhysiCell cancer simulator (Fig. 1b) [35]. In this model, individual cells are represented as mechanically interacting spheres that move freely in continuous space in response to external forces. Cells consume diffusing resources, which regulate their proliferation, and can be killed upon drug exposure. The initial tumor environment is generated by growing a tumor from a single cell over 15 days (Supplementary Movie 1), resulting in a population of 3584 cells with a compact subpopulation of 5 drug-resistant cells located approximately halfway between the tumor center and its surface.

To construct the base training environment (BE), we use a system of coupled ordinary differential equations that describe the competition between drug-sensitive (S) and drug-resistant (R) populations (Fig. 1c), following previous work [31, 36]:

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

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

The model captures logistic growth of the two competing populations, with growth rates γ_S and γ_R, limited by a carrying capacity K and an asymmetric competition coefficient c. Cell death is modeled through a stochastic rate δ, and the susceptible population additionally experiences a drug-induced death D, which is set to zero in the absence of treatment. This drug effect is modeled independently of population size as an abstraction of the periphery-to-core killing observed in the target environment (see Methods for details).

To parameterize the training environment, we assess therapy dynamics in the target environment under two reference treatment protocols: continuous therapy (CT) and adaptive therapy (AT_{50}) (Fig. 1d-e). The adaptive therapy protocol follows the clinical pilot study by Zhang *et al.* [4], in which treatment is paused once the tumor burden falls below 50% of its baseline value and resumed when the burden returns to 100%. The resulting tumor dynamics are used to calibrate the training environment parameters via Bayesian inference [37, 38].

Despite inherent simplifications, this parameterized training environment captures the trajectories observed in the target environment reasonably well by abstracting its complexity into effective parameters (Fig. 1e-f). For example, in the target environment, many mutant cells are confined to the resource-depleted tumor core, resulting in a lower effective proliferation rate γ_S than the actual division rate in the physical simulation. This is analogous to previous approaches where mathematical models were fit to clinical patient data for adaptive therapy optimization [18, 31, 36].

Therapy performance is evaluated based on the time to therapy failure (TTF), defined as the time at which resistant cells reach the initial tumor size. This metric differs from the commonly used time to progression (TTP), which is typically defined by a threshold on total population size [18, 31, 36]. We find that AT_{50} significantly outperforms CT, approximately doubling TTF in both the training and target environments (CT: 15.3±0.2 days in BE and 14.9±0.5 days in TE, AT_{50}: 33.8±0.8 days in BE and 30.2±2.9; Supplementary Fig. S1). Importantly, this benefit is observed even in the absence of resistance-associated fitness costs, in agreement with previous studies on adaptive therapy in spatially structured models [36, 39].

Using this parameterized training environment, we trained a cohort of $n = 10$ RL agents via proximal policy optimization (PPO) [40] (Fig. 1d). Actor and critic networks shared the same architecture, consisting of two fully connected layers of 128 units. Agents received observations based on a stacked history of

64 state-action pairs, encoding total population sizes and applied therapy actions. Multi-objective training was implemented by penalizing cumulative tumor burden while rewarding extended TTF, with agents receiving a constant reward for each time step in which the average tumor size within the observation buffer remained below baseline (see Methods for details).

During training, all agents successfully converged to effective policies (Supplementary Fig. S2), achieving an average TTF of 149 ± 19 days in the training environment, close to the maximum simulation time of 180 days (Fig. 1g). This represents an almost fivefold improvement over the fixed-threshold AT_{50} strategy (33.8 ± 0.8 days). However, applying the GRAPE evaluation framework revealed substantial performance divergence when these agents were transferred to the target environment (Fig. 1g; Supplementary Fig. S3). While some agents retained performance levels comparable to those observed in the training environment, others exhibited significantly worse outcomes. The observed degeneracy in the training environment, coupled with its collapse in the target environment, suggests that a critical — but initially hidden — dimension of tumor therapy dynamics was missing from the base training environment. Notably, this gap only became apparent once the RL agents had fully explored the decision space, exposing previously overlooked mechanisms that appear to significantly shape therapy outcomes.

Learning from failure: Systematically analyzing the mechanisms of agent failure in the target environment

To understand why certain agents fail to generalize to the target environment, we systematically analyze their decision-making behavior and environment-state trajectories in comparison to those of agents that generalize well. Within the GRAPEVINE framework, this step serves to uncover features that were overlooked in the initial design of the base training environment. Specifically, our goal is to identify the key differences in decision-making between successful and failing agents before link these differences to an emergent feature of the system that was not properly accounted for in the training environment abstraction. In this section, we focus on the first step.

Figs. 2a and b illustrate the dynamics of tumor burden and treatment decisions for the best- and worst-performing agents, evaluated in both training and target environments. We find that the RL-derived therapy protocols are not simple on-off strategies but exhibit dynamic, context-dependent behavior that deviates from the fixed-threshold approaches used in prior adaptive therapy studies [4, 31, 36].

A key distinction between high- and low-performing agents lies in their treatment initiation patterns. Poorly performing agents frequently apply treatment when tumor sizes are smaller than the initial baseline, particularly during the early phases of therapy. In contrast, successful agents initiate treatment at progressively higher tumor burdens, delaying aggressive therapy until tumor regrowth accelerates. This tendency to apply treatment at low tumor burdens — often inducing sharp tumor reductions — frequently precedes therapy failure in the target environment (see arrows in Figs 2a and b). While in the training environment such reductions have little consequence as long as they are followed by sufficient tumor regrowth, the same is not true in the target environment, where once a sharp reduction in tumor burden occurs, the system is unable to recover, ultimately leading to therapy failure (compare early trajectories in Fig. 2b).

To further investigate whether early 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 relationship to therapy failure time. Fig. 2c shows that, in the target environment, these two values are strongly correlated (Pearson's correlation coefficient, $r = 0.98$), with well- and poorly-performing agents forming distinct clusters. In contrast, no significant correlation is observed in the training environment ($r = 0.45$), suggesting that the influence of a temporary decrease in tumor burden on long-term outcomes was not adequately represented in the training environment.

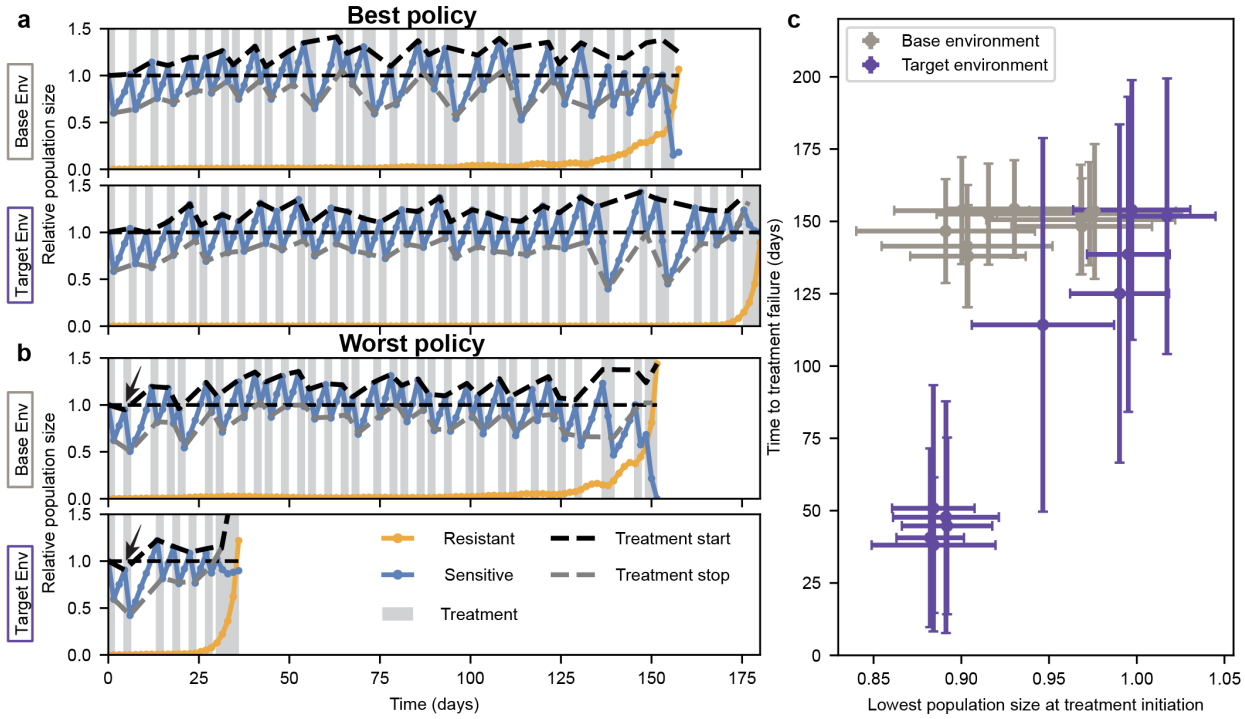

Figure 2: **Identifying the Causes of Agent Failure in the Target Environment. a**, **b** Examples of the dynamics of the sensitive (blue) and resistant (orange) population sizes - relative to the total initial population size - under the best and worst policies in the target environment, shown for both TE (Supplementary Movies 4 and 5) and BE. Grey regions indicate treatment application. Dashed lines show the tumor burdens at treatment initiation (black) and treatment pause (grey). Arrows point to the application of treatment at low tumor burden by the agent using the worst policy. **c** Time to treatment failure (TTF) in BE (grey) and TE (green) for n=10 policies trained in BE as the function of the lowest total population size at which treatment was initiated. Dots represent averages over n=50 runs, and error bars indicate the standard deviation.

Control of resistant population is not governed by total number of sensitive cells in the target environment

To investigate why early treatment at low tumor burden irreversibly leads to therapy failure in the target environment, we analyze the relationship between resistant and total population sizes over time (Fig. 3a-b). In the base training environment (Fig. 3a), resistant cells initially expand rapidly during the early treatment phase when tumor burden is low. However, as the total population size increases later in the therapy cycle, resistant cell growth slows and can even reverse, leading to partial suppression. Between days 30 and 80, therapy cycling maintains the resistant subpopulation in a controlled oscillatory state around 2%. While initiating treatment at lower tumor burdens in later stages weakens this control, resistance remains suppressed whenever the overall tumor burden is sufficiently high.

In contrast, the target environment exhibits fundamentally different behavior (Fig. 3b). Following aggressive early treatment, resistant cells initially expand at a high rate but are temporarily held in check. However, despite moderate tumor burdens in later treatment cycles, resistant cells eventually resume

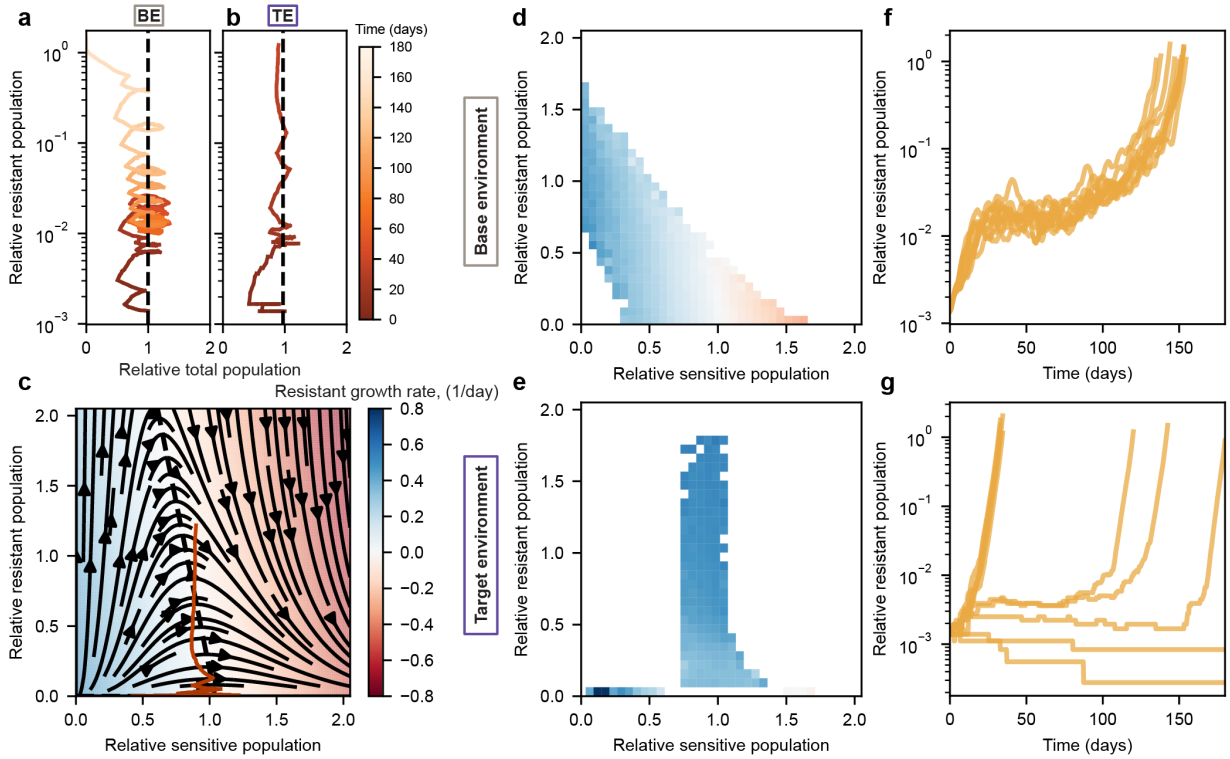

Figure 3: **Control of resistant population in the target environment is not governed by sensitive cell numbers. a**, **b** Relationship between resistant and total population sizes over time (color) for the base (BE) and target (TE) environments during therapy under the worst policy. Dashed lines indicate the average cell number kept by the policy. **c** Phase space of resistant vs. sensitive populations in BE. Streamlines represent trajectories in time under free growth, orange line shows the trajectory of the therapy run in TE, presented in **b**. Colors indicate the effective growth rate of the resistant population per day. **d**, **f** Regions of the phase space explored by the n=10 agents over n=50 evaluation runs on BE and TE. 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 BE and TE. All population sizes are shown relative to the initial tumor burden.

rapid growth with minimal suppression, even when total tumor burden remains high. The divergence between the two environments becomes particularly apparent when mapping population trajectories into a population phase space, using the relative sizes of the susceptible and resistant populations as dimensions (Fig. 3c). In the training environment, trajectories follow the expected flow field from the underlying equations, where resistant cells can be re-suppressed after expansion. In contrast, the target environment deviates sharply, suggesting that competitive release becomes effectively irreversible. These findings challenge a central assumption of the base environment — that resistance suppression is primarily dictated by high total cell numbers.

To further explore this discrepancy, we examine resistant cell growth under agent-driven therapies in the same phase space. In the training environment, observations align with model expectations, where resistance growth remains low or even negative when the susceptible population is large (Fig. 3d). However, in the target environment, the same phase-space

region consistently exhibits strong resistant growth (Fig. 3e), indicating that the expected suppression is absent. Temporal analysis of multiple resistance growth trajectories further highlights the contrast between environments: in the training model, growth follows predictable trajectories, while in the target environment, resistance growth surges at widely varying time points (Fig. 3f, g).

Together, these findings challenge the assumption that resistance suppression is primarily governed by the average abundance of susceptible cells. Instead, the loss of resistance confinement appears to be triggered by a transient reduction in tumor burden rather than absolute population numbers. In the next section, we will explore the underlying mechanisms driving these dynamics.

Position-dependent growth and collective cell motion are key features of therapy failure dynamics

Having identified the agent behaviors responsible for divergent performance as the first step of GRAPE, we now investigate the underlying mechanisms by manually analyzing the relevant subtrajectories. This step, previously infeasible due to the sheer volume of data (45284 total simulated days in the target environment across all agents), leverages the unique ability of humans to intuitively recognize spatio-temporal correlations in movies—a capability that remains difficult to replicate even with advanced AI algorithms. Using this human-in-the-loop approach, we visually inspect the trajectories of resistant clones that have been identified to exhibit critical behavior (Fig. 4a; Supplementary Movie 6).

We observe that cell proliferation rates gradually decline with increasing distance from the tumor edge, consistent with the resource gradient imposed by nutrient consumption in the simulations (Fig. 4b). 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 (Fig. 4c). 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 tumor regrowth.

Critically, we find that during this post-treatment expansion phase, resistant cells within the bulk can be physically displaced by the collective growth dynamics following a therapy cycle, especially if previous population shrinkage was substantial (see arrows in Fig. 4a). Importantly, once the distance of a resistant cell to the front is sufficiently decreased, subsequent therapy cycles — even if less extensive—continue this process, gradually shifting the cell closer to the tumor front. This ratchet-like motion progresses in a stepwise yet irreversible manner until resistant cells reach the edge (Fig. 4d, e). As their distance to the tumor front decreases, resistant cells gain access to more favorable growth conditions, ultimately triggering rapid resistance expansion (Fig. 4f).

This intricate interplay between position-dependent growth and mechanically induced motion of resistant cells fundamentally shapes therapy response dynamics. To enable RL agents to incorporate these effects and improve policy generalization to the target environment, the next step in the GRAPEVINE framework is to refine the training model accordingly.

Augmenting the training environment to reflect spatial dynamics

The goal of environment augmentation is to introduce minimal, interpretable modifications that capture essential spatial dynamics while maintaining the simplicity and efficiency of the original model. To this end, we focus on updating the representation of the resistant population, while keeping the growth equation for susceptible cells unchanged for now (Fig. 5a).

We incorporate two key features observed in the target environment: position-dependent growth and growth-induced motion of resistant cells. First, we 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. The tumor radius is approximated as $r_T \sim \sqrt{S+R}$. Second, we abstract the mechanically induced outward motion of resistant clones by introducing a

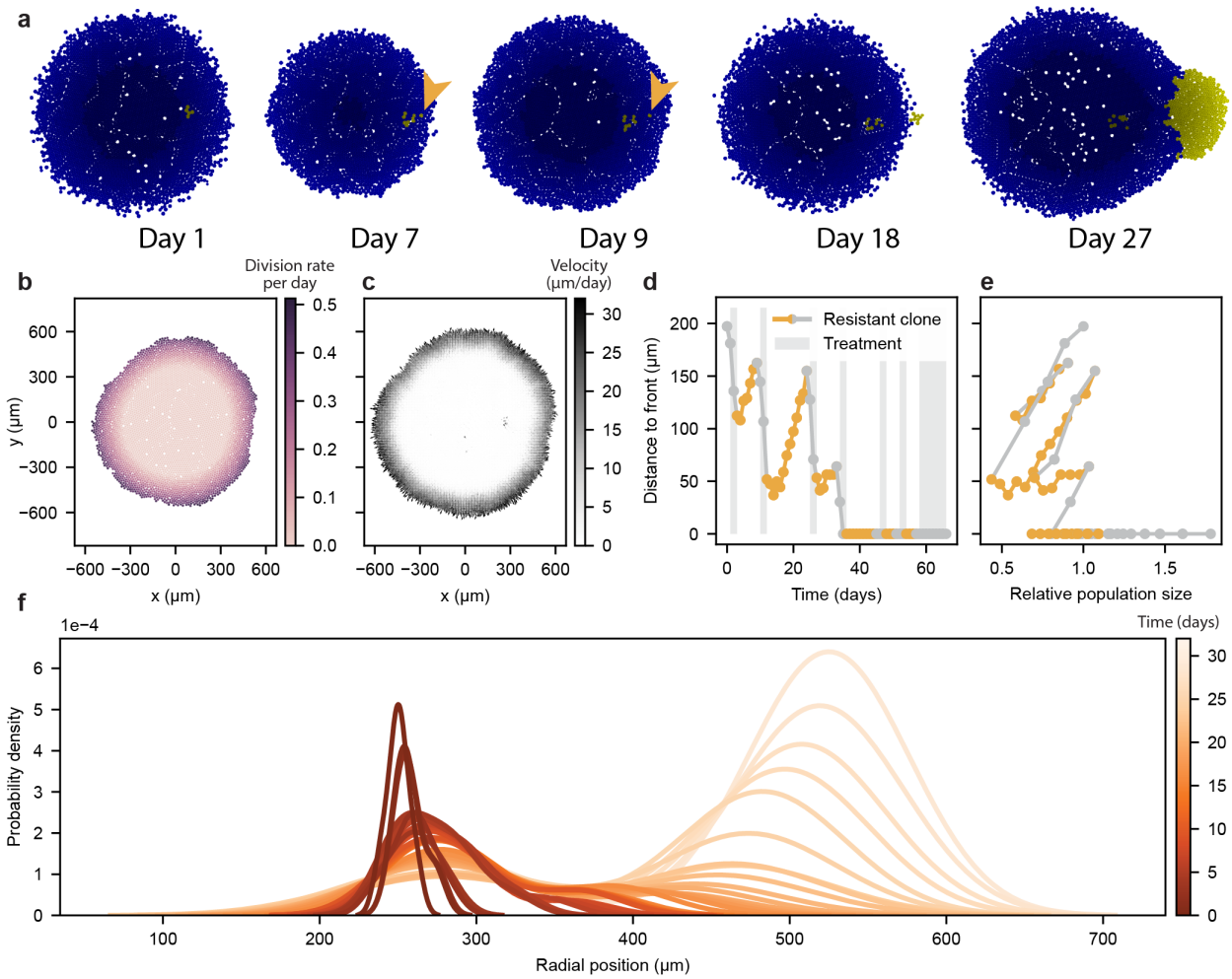

Figure 4: **Position-dependent growth and resulting motion gradients lead to history-dependent effects in the target environment a** Selected simulation snapshots of the TE under the worst policy. Sensitive cells in blue, resistant - in yellow. Brighter color corresponds to larger growth rate. Arrows point at a single resistant cell being pushed towards the tumor front (see Supplementary Movie 6). **b** Single-time snapshot of the 2-D growth rate distribution, where hue increases with growth rate. **c** Average velocity of cells at certain positions over 3 days under free growth (n=50 runs). Arrow magnitude is proportional to velocity. **d** Closest distance from any resistant cell to the cell exactly at the tumor front over time for the simulation run presented in **a**. Orange represents treatment holidays, while grey shows the time points when the treatment was applied. **e** Relationship between the distance to the front and the total population size, relative to the initial tumor burden. **f** Radial distribution of resistant cells in time for the same simulation run. Hue decreases with time.

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:

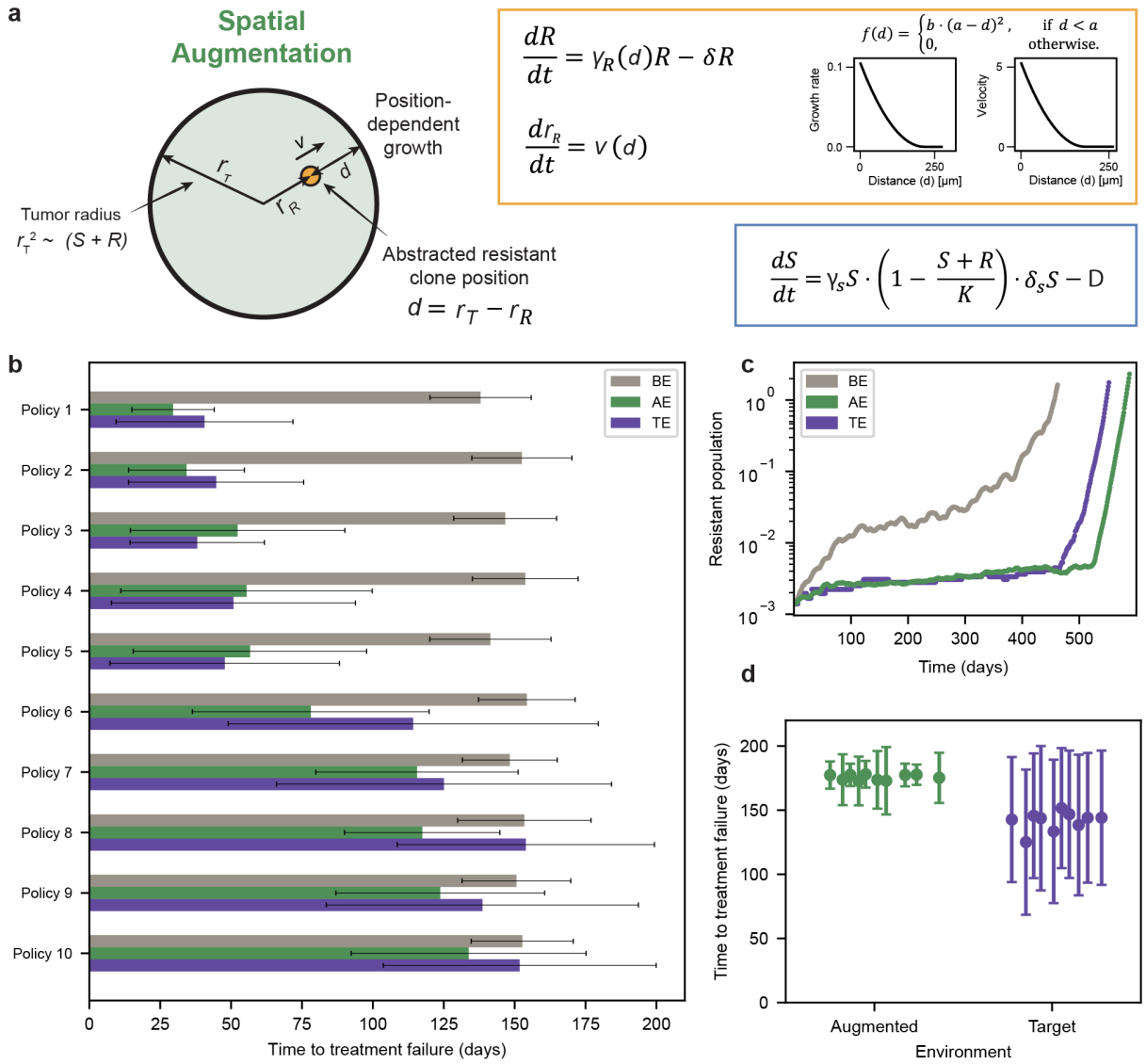

Figure 5: **Augmenting the training environment to reflect spatial dynamics.** **a** Schematic of the augmented environment (AE). **b** Average time to treatment failure (TTF) achieved by the RL agent with policies trained in BE, evaluated in BE (grey), AE (violet) and TE (green). n=50 runs, error bars indicate the standard deviation. **c** Resistant population dynamics over time in the three environments, BE - grey, AE - violet and TE - green. Population size is presented relative to the initial tumor burden. **d** Relative performance (TTF) of the group of agents trained in AE, evaluated in AE (violet) and TE (green). n=50 runs, error bars indicate the standard deviation.

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

with $d = r_T - r_R$. 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 target environment (see Methods for details). Notably, the position-dependent growth rate in the augmented environment effectively replaces the role of the carrying capacity in the original base environment.

Before retraining new agents, we first assess the impact of these augmentations by evaluating the original agents —trained in the base environment — on the newly augmented environment. We find that individual agent performance in the augmented environment correlates strongly with their performance in the target environment (Fig. 5b). Furthermore, the augmented environment successfully reproduces the characteristic biphasic growth dynamics observed in the spatial model and reported before (Fig. 5c)[14]. These results support our hypothesis that the spatially dependent growth and motion of resistant cells are key drivers of therapy failure in the target environment.

We then train a new cohort of agents in the augmented training environment (Supplementary Fig. S4). As before, agent performance in the training environment remains stable and consistent across replicates (Fig. 5d). Most importantly, agents trained in the augmented environment now exhibit strong generalization to the target environment.

The resulting RL-learned strategies deviate from the overly aggressive behaviors observed previously. Instead, they favor initiating treatment at higher tumor burdens, effectively delaying the onset of the ratchet-like forward motion of resistant cells (Supplementary Fig. S5). These results demonstrate that GRAPEVINE provides a powerful framework for guiding the abstraction of complex, spatially structured environments into RL-compatible training models in a way that preserves mechanistic interpretability.

Discussion

Solid tumors constitute complex, physics-driven systems whose treatment response can hinge on intricate, spatially dependent phenomena. Despite exciting advances in targeted drugs and immunotherapies, resistance evolution remains a primary cause of therapy failure. Addressing this challenge requires not just more potent treatments but also novel frameworks that can detect and integrate hidden mechanistic features governing tumor dynamics. Here, we sought to fill this gap by introducing GRAPEVINE (Group-Relative Agent Policy Evaluation for Varying INterpretable Environments). By systematically evaluating the differential policies of a group of therapy-applying reinforcement learning (RL) agents in a high-fidelity tumor simulations, GRAPEVINE exposes key emergent processes — enabling us to refine mechanistic models in a manner that both improves therapy strategies and advances our fundamental understanding of solid tumor evolution.

In this study, we use this strategy to uncover previously overlooked biophysical drivers of therapy outcomes. Instead of viewing inconsistent results between training (LoFi) and target (HiFi) environments as "failure", we leveraged these discrepancies to pinpoint which features were missing in the simplified model. By harnessing differences in RL agent performance, GRAPEVINE revealed that multiple policies can appear similarly effective in a minimal, low-dimensional environment yet diverge once transferred to a spatially explicit simulation. This divergence flags essential physics-based interactions absent from the original model, facilitating systematic environment augmentation. In doing so, we shift RL's focus from mere policy optimization toward environment refinement - where each "failing" agent essentially becomes an exploratory probe revealing hidden facets of the complex target environment.

Our proof-of-concept application of GRAPEVINE to adaptive cancer therapy underscores how spatial constraints and mechanical forces can dramatically alter the evolutionary course of drug-resistant subpopulations. While the field has long recognized that spatial organization matters for treatment outcomes, our findings uncover a new layer by showing

exactly how position-dependent growth and mechanically driven cell displacement lead to irreversible resistance expansion. These observations are consistent with prior *in silico* and *in vitro* studies highlighting how localized resource gradients and collective cell dynamics can shape selection pressures [11–13, 18]. However, by tying these phenomena directly to adaptive tumor therapies via RL policy failures, we offer a structured way to incorporate them 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 [24, 33, 34].

As with any modeling framework, GRAPEVINE has inherent limitations. First, detecting missing features depends on sufficient diversity among RL agents; without carefully designed reward functions or diversity-promoting mechanisms, the ensemble may converge to near-identical strategies, limiting our ability to glean insights. Second, the approach is only as robust as the realism of the target environment. Although we employed a state-of-the-art physics-driven cancer simulator (PhysiCell), real tumors involve additional factors such as vascular heterogeneity, immune infiltration, and metabolic feedbacks. Third, bridging the gap between simplified and fully realistic models can become computationally demanding, especially for large-scale, three-dimensional tumors. Finally, the initial positioning and abundance of resistant cells in our target environment, as well as the assumption of an isolated lesion, does not reflect the complexity encountered clinically, restricting immediate translational application. Future iterations would have to take these added complexities, as well as a more refined abstraction of susceptible cell behavior, into account.

Looking forward, there are several promising directions to build on these findings. One key avenue is extending GRAPEVINE to more physiologically detailed simulators that incorporate immunological components, angiogenesis, or patient-specific data. Such expansions could help refine adaptive therapy strategies in personalized oncology. Another important step is to develop methods that intentionally promote policy diversity, ensuring that the group-level performance analysis yields rich insights into underexplored parts of the decision space. Beyond oncology, employing GRAPEVINE in other complex systems — such as ecological management or robotics — could similarly reveal hidden emergent processes that hamper straightforward RL-based optimization. Integrating these insights into more sophisticated multi-fidelity pipelines can further reduce computational overhead while still capturing critical biophysical or biological details.

In summary, GRAPEVINE reframes the perceived "failure" of RL policies to generalize as a potent discovery mechanism for illuminating non-obvious dynamics in high-dimensional, physics-driven environments. By systematically identifying and integrating missing processes — rather than discarding failing agents — researchers can iteratively refine their mechanistic models and develop more robust, interpretable strategies for disease management. This work thus bridges the gap between simplified mathematical abstractions and realistic tumor behavior, laying a foundation for more nuanced adaptive therapies and offering a versatile tool for artificial scientific discovery in other domains as well. Ultimately, GRAPEVINE underscores the value of embracing, rather than avoiding, the complexities that drive real-world phenomena — and of harnessing those complexities to achieve deeper insights.

Methods

Target environment

The target environment is modeled using an off-lattice agent-based cancer model (ABM) implemented in PhysiCell (v1.12.0) [35] with BioFVM (v1.1.6) [41], an open-source cancer simulation platform. 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 (Supplementary Fig. S6), beyond which cells become quiescent. The parameters governing this oxygen-dependent behavior are informed by experimental data from Grimes et

al. [42]. A full list of parameters is provided in the shared repository.

Therapy is simulated by binary switching of Dirichlet boundary conditions, toggling between drug absence (0 mmHg) and full-dose exposure (38 mmHg). Drug effects scale with the concentration and the growth rate, largely increasing the apoptosis rate in proliferating cells. Given the oxygen-dependent proliferation pattern, this results in preferential cell kill at the tumor periphery, where growth is highest. In the absence of treatment, the baseline cellular turnover rate remains at 1%.

Base 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 [18, 31, 36].

Populations of sensitive S and resistant R cells grow with rates γ_S and γ_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 target environment during reference AT_{50} therapy.

Both populations experience a constant death rate δ, 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 target environment when sensitive cells dominate and resistant fraction remains low.

However, deviations between the non-spatial base and the spatial target 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 environment

We performed Bayesian inference to parametrize the base environment using the PyMC package v5.9.2 [37]. 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 γ_R was first fitted to a continuous therapy scenario, keeping all other parameters fixed.
2. The sensitive cell growth rate γ_S, treatment strength D, and competition coefficient c were then fitted to the adaptive therapy (AT_{50}) reference scenario, using the updated γ_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: $\gamma_S = 0.04$, $\gamma_R = 0.134$, $D = 330$, $c = 4.78$.

Posterior distributions for the fitted parameters are provided in Supplementary Fig. S7. 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 (AT_{50}) therapies, treatment decisions were made every 6 simulated hours to maintain cell counts near the treatment thresholds. For RL-based agents, this decision frequency was reduced to 1.5 simulated days to enhance clinical feasibility.

Reinforcement learning

Treatment decision agents were implemented using Stable Baselines3 [43], a PyTorch-based library providing reliable RL algorithm implementations. We trained the agents using Proximal Policy Optimization (PPO) [40], 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 underwent 4,882 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 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. 2c), we computed Pearson's correlation coefficient (r) using the NumPy function corrcoef.

Augmented 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 target 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 (Supplementary Fig. S6). 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_{T}) based on total population size, assuming a densely packed tumor structure with cells of radius r_{c}:

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

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 γ_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 target environment simulations are presented in Supplementary Fig. S8.

Data availability

Data used to produce figures is available at `https://github.com/sergiyayf/PhysiLearning`. Trained networks and target environment data, including cell positions, are available upon request due to their large size.

Code availability

The code used for simulations, training, and analysis in this study is available at `https://github.com/sergiyayf/PhysiLearning`. A new tag of this repository with detailed instructions and scripts for each figure panel will be created upon final submission.

References

1. Vasan, N., Baselga, J. & Hyman, D. M. A view on drug resistance in cancer. en. *Nature* **575,** 299–309 (Nov. 2019).
2. 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).
3. Gatenby, R. A. & Brown, J. S. Integrating evolutionary dynamics into cancer therapy. en. *Nature Reviews Clinical Oncology* **17,** 675–686 (Nov. 2020).
4. 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).
5. 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).
6. 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).
7. 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).
8. Brady-Nicholls, R. & Enderling, H. Range-bounded adaptive therapy in metastatic prostate cancer. *Cancers* **14,** 5319 (2022).
9. West, J. *et al.* A survey of open questions in adaptive therapy: Bridging mathematics and clinical translation. eng. *eLife* **12,** e84263 (Mar. 2023).

10. 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).

11. 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).

12. 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).

13. 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).

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

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

16. 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).

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

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

19. 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).

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

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

22. Hallatschek, O. *et al.* Proliferating active matter. *Nature Reviews Physics* **5,** 407–419 (2023).

23. 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).

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

25. 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).

26. 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).

27. Weaver, D. T., Maltas, J. & Scott, J. G. *Reinforcement Learning informs optimal treatment strategies to limit antibiotic resistance* en. Jan. 2023.

28. 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.

29. 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).

30. Eastman, B., Przedborski, M. & Kohandel, M. Reinforcement learning derived chemotherapeutic schedules for robust patient-specific therapy. en. *Scientific Reports* **11,** 17882 (Sept. 2021).

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

32. 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).

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

34. 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).

35. 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).

36. 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).

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

38. Brunkhorst, G. in *PyMC examples* (ed Team, P.) (2022).

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

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

41. 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).

42. 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).

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

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.