Reinforcement Failing guides the discovery of emergent 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-Universität Erlangen-Nürnberg, 91054 Erlangen, Germany
[*]E-mail: jona.kayser@mpzpm.mpg.de

June 2, 2025

Abstract

Artificial intelligence is revolutionizing oncology by transforming how malignancies are diagnosed, cancer biology is understood, and therapeutics are discovered. A cornerstone of this progress has been the availability of extensive, carefully curated datasets. Similarly, advances in AI-guided therapeutic strategies via Reinforcement Learning (RL) hinge upon carefully designed computational training environments that are both efficient and sufficiently realistic to capture key dynamics of cancer growth and therapy response. However, designing suitable models remains challenging for solid tumors, where emergent physical phenomena significantly influence therapeutic outcomes. Here, we introduce Reinforcement Failing, an AI-guided scientific discovery framework designed to reveal emergent physical mechanisms in tumor therapy. Applying this approach to adaptive therapy in solid tumors, we identify the pivotal roles of position-dependent proliferation and mechanically driven collective motion of resistant cells. Our findings highlight how integrating tumor physics into therapeutic strategies can optimize outcomes while mitigating hidden pitfalls during translation. Together, these results demonstrate that Reinforcement Failing is a powerful artificial scientific discovery engine for deciphering high-complexity processes in personalized cancer treatment and beyond.

Introduction

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 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]. By 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, several critical questions remain on the path towards clinical translation[9]. For example, solid tumors are physical objects that are subject to a range of emergent physical properties and dynamics, ranging from spatial population structure and environmental gradients to mechanical cell–cell interactions. These interactions, inherent to any dense population of cells, can drive population-level collective dynamics that fundamentally alter evolutionary trajectories and treatment responses[10–14]. Recent computational and experimental studies have underscored the importance of these physical mechanisms for resistance evolution in general, and the potential gains of adaptive scheduling in particular [15–22].

These findings argue for a new perspective on solid tumors, understanding them as actively proliferating, granular matter systems in which evolutionary outcomes are emergent properties of spatial and mechanical constraints [23]. Consequently, designing effective adaptive therapies demands a mechanistic understanding of these collective effects. Yet, the very complexity required to accurately capture such dynamics renders simulations hard to interpret, hampering traditional model-driven 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 [24, 25]. In oncology, RL has recently been used to optimize treatment regimens, including in radiotherapy or chemotherapy [26–31]. Notably, Gallagher *et al.* recently demonstrated that RL can outperform fixed-threshold adaptive therapy strategies in simulated patients by learning more nuanced dosing policies [32]. Similar efforts, using the same dataset, were undertaken by Lu *et al.* [33].

However, applying RL to the realm of solid tumors remains elusive. Explicitly integrating spatial effects and inherent physical mechanisms, such as resource gradients or the growth-induced collective motion of cells, into tumor simulations necessitates a complexity that makes interpretation opaque while rendering them ill-suited for RL training due to high computational demands. Overcoming these barriers will require the abstraction of complex tumor growth and therapy dynamics into simplified, interpretable, and more efficient training models. The nascent field of multi-fidelity RL has begun to address these challenges by pairing low-fidelity training environments with subsequent evaluation and possibly transfer learning in high-fidelity simulations [34, 35]. However, without principled guidance on which emergent physical mechanisms shape therapeutic outcomes, this abstraction process becomes a bottleneck, potentially omitting key dynamics [36].

To overcome this hurdle, we extend multi-fidelity RL by introducing a novel framework that integrates high-fidelity simulations of physical tumor growth dynamics with RL to guide the systematic augmentation of the low-fidelity training environment. The core idea behind this approach is "Reinforcement Failing" (RF), a novel concept that leverages Group-Relative Agent Performance Evaluation (GRAPE) to expose hidden system features that shape therapeutic success. We show that the differential performance of some RL agents - those that "fail" relative to their peers - can serve as signal for human researches to guide them toward hidden emergent processes that are pivotal for therapy success. 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.

Results

Bridging tumor physics and adaptive therapy via Reinforcement Learning

The goal of this work is to use reinforcement learning to uncover how emergent physical mechanisms shape

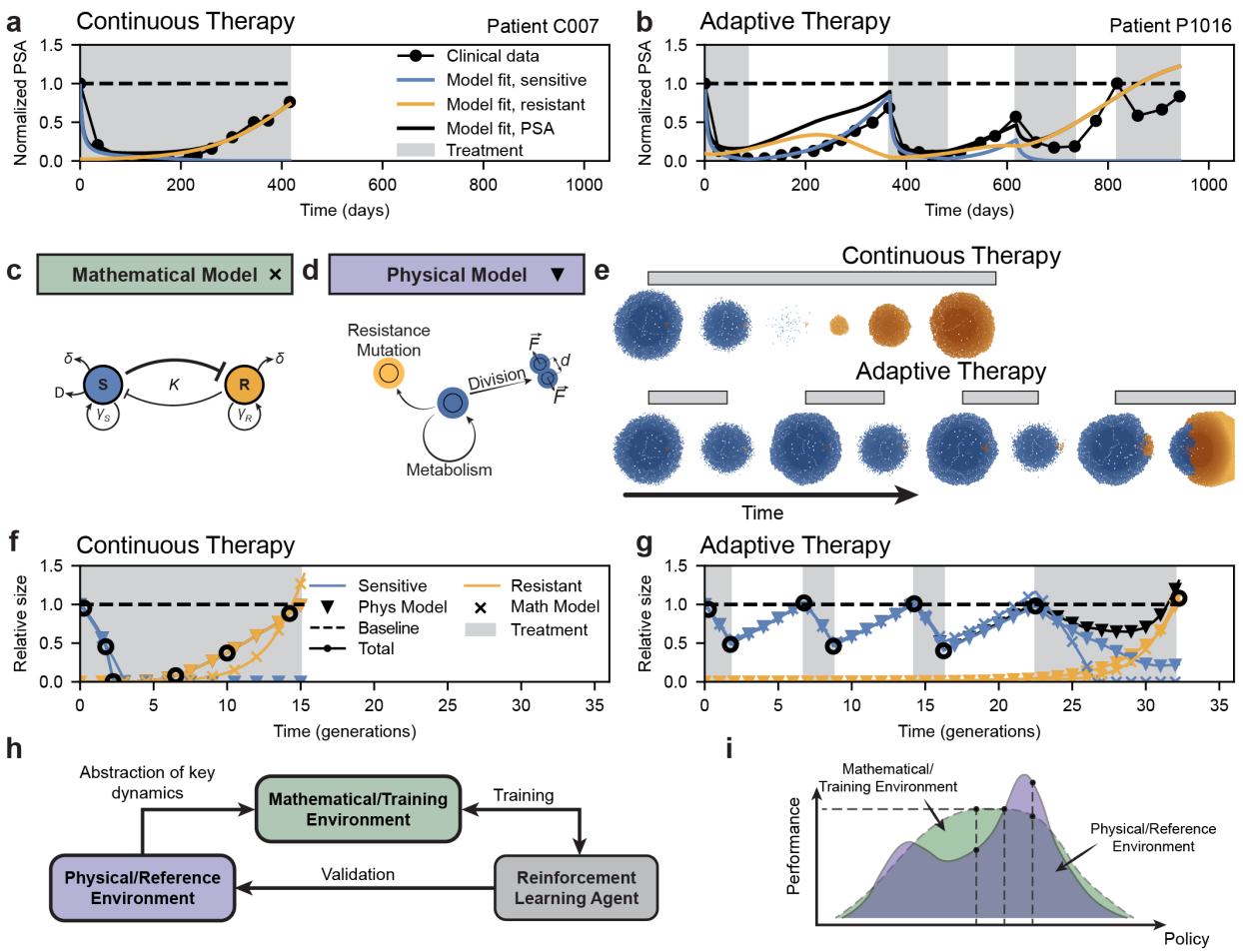

Figure 1: **A physical model for adaptive tumor therapy**. **a**, **b** Patient data of normalized PSA values for continuous and adaptive therapy AT_{50} (adapted from Ref. [5]). Solid lines represent a mathematical model fit to individual patient data as reported in that study. **c** Schematic of mathematical tumor growth models (see main text for details). **d** Schematics of physical tumor growth model. **e** Snapshots of the physical simulation under Continuous Therapy (CT, Supplementary Movie 2) and Adaptive Therapy (AT_{50}, Supplementary Movie 3). Sensitive cells are shown in blue, resistant cells are shown in orange. Brighter colors indicate higher proliferation rates of the cells. Grey bars correspond to regions of treatment application. **f** Dynamics of the sensitive (blue) and resistant (orange) population sizes, relative to the total initial population size, in the physical simulation (triangles) and fitted mathematical model (crosses), under CT. Grey regions indicate treatment application. **g** Population size dynamics under AT_{50}. Black circles mark the time points shown in the snapshots in **e**. **h** Conceptual outline of multi-fidelity reinforcement learning. **i** Schematic of cross-environment performance differences (CEPD) of policies between the mathematical and the physical models.

therapy failure dynamics in solid tumors as a gateway towards physics-informed adaptive tumor therapies.

While many patients initially respond to anti-tumor therapy, long-term tumor control often remains elusive due to the presence of resistant subpopulations (Fig. 1 a). Building on mathematical models

of tumor dynamics, 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 b, c)[5, 37].

However, strategy performance critically depends on the assumptions underlying the model design. How well existing mathematical frameworks can capture the dynamics in more complex settings, such as spatially structured solid tumors, or how they need to be augmented to capture the physical nature of solid tumors, is currently not known.

To assess the role of emergent physics-driven dynamics in a therapy setting, we construct a tailored *in silico* tumor, building on the state-of-the-art PhysiCell cancer simulator [38]. In this model, individual cells are represented as mechanically interacting spheres that move freely in continuous space in response to external forces (Fig. 1d). Cells consume diffusing resources, which regulate their proliferation, 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).

In this physical environment, we evaluated two treatment protocols: continuous therapy (CT) (Supplementary Movie 2) and adaptive therapy (AT_{50}) (Supplementary Movie 3) (Fig. 1e-g). The adaptive therapy protocol is analogous to the study design of 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% [37]. The dynamics observed in our *in silico* model qualitatively capture the trajectories reported in the clinical trial, including an extended control of tumor growth. This is an inherent behavior of the system without any fitted parameters. Notably, the observed benefit of adaptive scheduling is not contingent on a resistance-associated fitness costs, in agreement with previous reports on adaptive therapy in lattice-based spatial models [39, 40].

While this physical environment is ideal for capturing emergent system characteristics, its computational demands make it less suited for reinforcement learning. We therefore need to first abstract this physical "reference" environment (RE) into a mathematical "training" environment (TE).

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. 1c), following previous work [32, 40]:

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

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

The environment models logistic growth of the two competing populations, with growth rates γ_S and γ_R, limited by a shared 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. In contrast to previous implementations, the drug effect is modeled independently of population size to account for the periphery-to-core killing observed in the reference environment (see Methods for details). The parameters of the base training environment are fitted to the dynamics observed in the reference environment via Bayesian inference [41, 42].

While it is clear that this simplified environment 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 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 abstraction process is analogous to the fitting of mathematical models to clinical patient data (Fig. 1a, b) [5, 19, 32, 37].

Despite its inherent limitations, the parameterized training environment qualitatively captures the trajectories of the reference environment (Fig. 1). 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). We find that AT_{50} significantly outperforms CT, approximately doubling TTF in both the mathematical and the physical environments (CT: 15.3 ± 0.2 generations in TE_0 and 14.9 ± 0.5 generations in RE; AT_{50}: 33.8 ± 0.8 generations in TE_0 and 30.2 ± 2.9 generations in RE; Supplementary Fig. 1). Here, a generation is defined as the time required for a cell to divide under optimal conditions in RE.

The parameterized mathematical model can now serve as training environment in a reinforcement learning approach to therapy optimization (Fig. 1h). In short, an agent - typically a deep neural net (DNN) - learns to make optimal treatment decisions, based on the observed state of the training environment. Finally, the performance of the trained agent's policy, or strategy, needs to be validated post-training in the reference environment. To achieve this, we created PhysiLearning, an open repository that integrates the PhysiCell simulation platform with Stable Baselines 3 for reinforcement learning in PyTorch (see data availability statement).

However, the success of this scheme hinges on the adequate abstraction of the training environment. During training the agent can explore regions of the vast decision-making space that may have not been considered during the initial abstraction process, leading to local divergences in training and validation performance (Fig. 1i). Such variations in performance is a direct result of the hidden features of the reference environment that were omitted during abstraction. Rather than discarding poorly performing agents, we argue that this relation can be harnessed to guide the abstraction process itself, eventually resulting in superior performing agents. To this end, we introduce "Reinforcement Failing", a new conceptual framework that integrates RL-based strategy optimization with the guided refinement of the training environment (Fig. 2). In the following, we first give an overview of the general structure of the framework before applying it to the specific context of adaptive tumor therapy.

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

The Reinforcement Failing framework consists of four concatenated main blocks: Construction and abstraction of the reference environment, RL training, cross-environment performance evaluation, and the human-in-the-loop augmentation of the training environment by previously hidden features (Fig. 2). The initial step is the construction of the reference environment (RE), designed to capture the intricacies of the complex real system, and its subsequent abstraction into the base training environment (TE_0).

A key innovation of the framework is the use of Group-Relative Agent Performance Evaluation (GRAPE) to uncover hidden but outcome-critical emergent features. Instead of only training an individual agent with the goal to find the single best policy, GRAPE aims to more widely explore the decision space via a group of independently trained RL agents with a diverse set of policies. Each of these agents can then be evaluated in both training and reference environment to generate a cross-environment performance difference (CEPD) map of the policy space.

Going beyond pure validation, a human expert can then use this information as guidance when comparing patterns in agent behavior and environment response in the context of existing scientific knowledge. By pinpointing policy weak points, Reinforcement Failing helps to identify potential underlying mechanisms that were missed during the initial reference environment abstraction. Based on these insights, the training environment is augmented to incorporate newly identified emergent features while preserving computational efficiency and mechanistic interpretability. The process can then be repeated with a new set of RL agents trained in the augmented environment (TE_+), allowing for progressive model refinement until performance in training and reference environments converge.

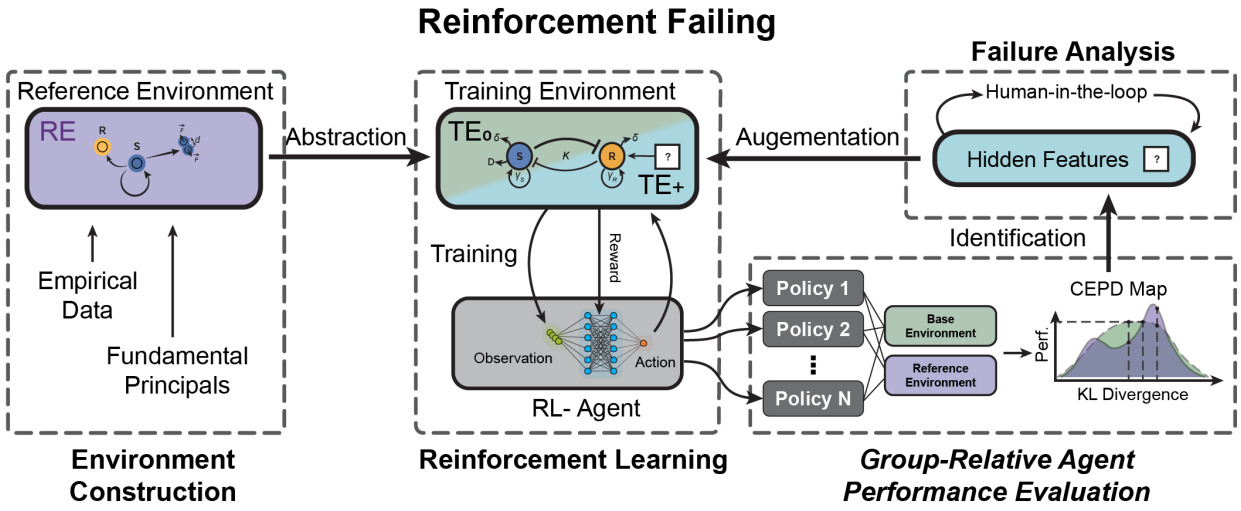

Figure 2: **Schematic of the Reinforcement Failing framework**. The Reinforcement Failing framework consists of four main blocks. Environment Construction: Based on empirical data and fundamental physical principles, a high-fidelity simulation is designed as reference environment (RE). The complex reference environment is then abstracted into a mathematical base training environment (TE_0). Reinforcement Learning: An ensemble of multiple independent agents with distinct policies is trained to optimize outcome based on a predefined reward function. Group-Relative Agent Performance Evaluation (GRAPE): Evaluating all agents in both TE_0 and RE yields a cross-environment performance difference (CEPD) map, using the Kullback-Leibler divergence from the least performing agent as policy space coordinate. Failure Analysis: Detailed analysis of differentially performing agents by a human researcher guides the refinement of the training environment to yield a more accurate representation in an augmented training environment (TE_+), followed by the training and evaluation of second-generation agents.

Application of Reinforcement Failing to adaptive tumor therapy

Applying Reinforcement Failing to the context of adaptive tumor therapy, we trained a cohort of $N = 10$ independent agents in the parameterized base training environment. During training, all agents successfully converged to effective policies (Supplementary Fig. 2), achieving an average TTF of 149±19 generations in the training environment (Fig. 3a). This represents an almost fivefold improvement over the fixed-threshold AT_{50} strategy (33.8±0.8 generations).

However, applying the GRAPE evaluation framework reveals substantial performance divergence when these agents are transferred to the reference environment (Fig. 3a; Supplementary Fig. 3). To understand why some agents fail to generalize to the reference environment, we analyze their decision-making behavior and environment-state trajectories in comparison to those of agents that generalize well (Figs. 3b-d).

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. 3d). This episode is of little consequence in the training environment as long as it is followed by sufficient tumor regrowth (Fig. 3d, top). However, in the reference environment, once a sufficiently sharp reduction in tumor burden occurs,

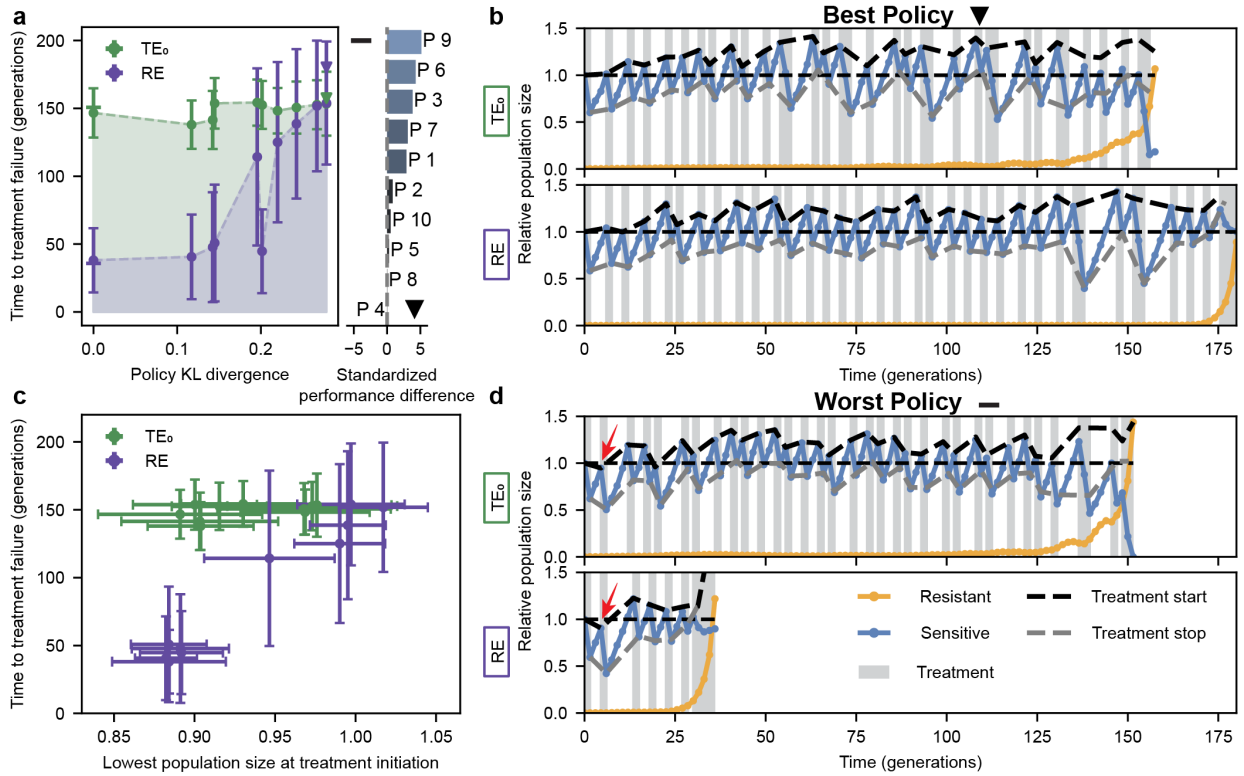

Figure 3: **Systematic analysis of agent failure modes**. **a** Left: Time to treatment failure (TTF) in TE_0 (green) and RE (purple) for policies trained in TE_0 ($N = 10$). Dots represent averages over $n = 50$ runs, error bars indicate the standard deviation. The Kullback-Leibler divergence of each policy from the worst policy is used as policy 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. **b**, **d** 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 **a** with ▼ (best) and $-$ (worst), colored according to the environment. 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 (see main text). **c** 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.

the system is unable to recover, resulting in resistance growth and ultimately therapy failure (Fig. 3d, 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. 3c). 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 found 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 ($r = 0.77$) (Supplementary Fig. 4)[39, 43]. These observations suggest additional mechanisms as determinants of therapy failure dynamics (see Supplementary Information and Supplementary Fig. 5 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). This was previously infeasible due to the sheer volume of data (> 45000 total simulated generations in the reference environment across all agents).

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 (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 (Supplementary Fig. 6a). 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 (Supplementary Fig. 6b).

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

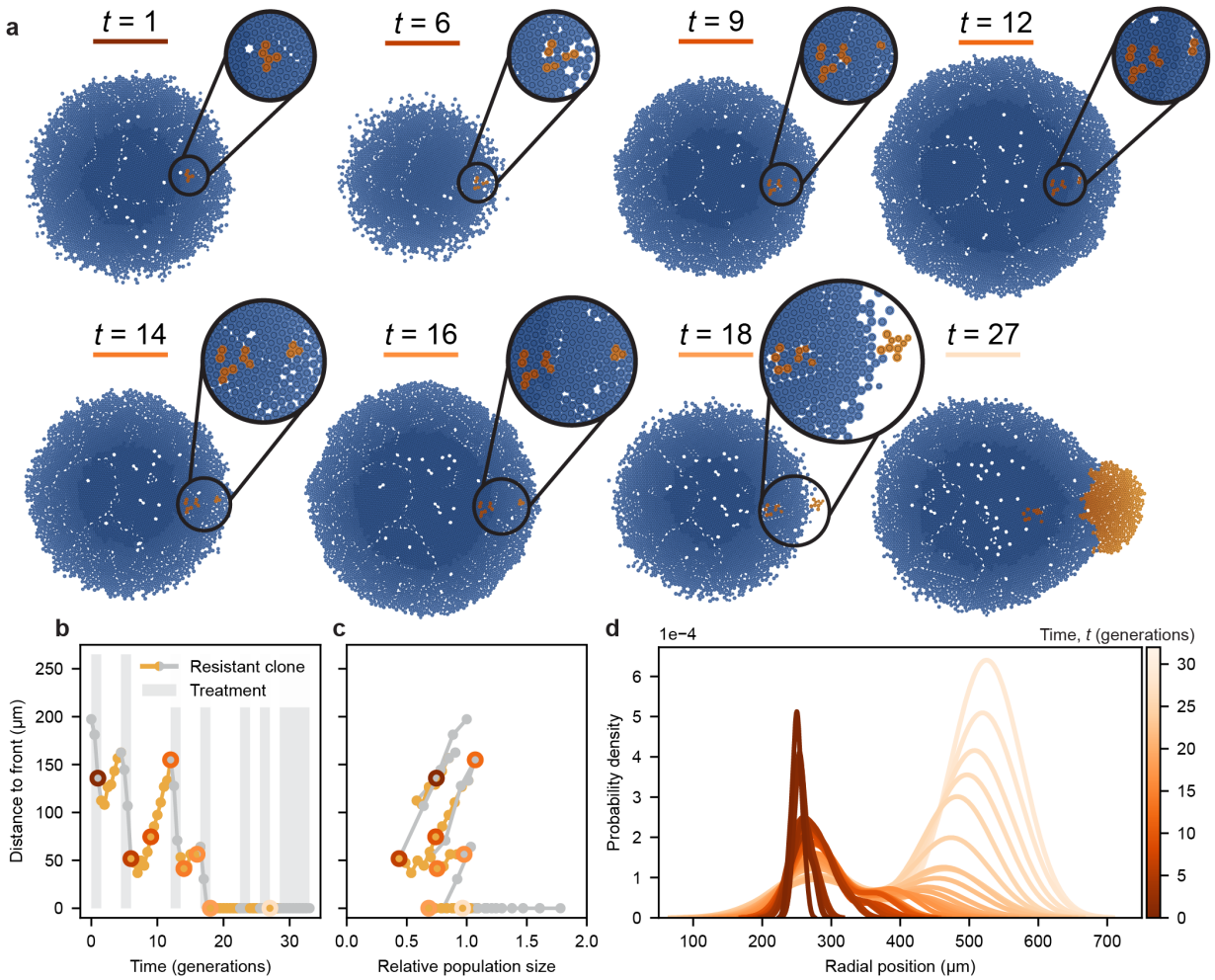

Figure 4: **A mechanical ratchet effect drives the competitive release of resistant cells.** **a** Selected simulation snapshots of the RE under the worst policy. Sensitive cells in blue; resistant cells in orange. Brighter colors indicate larger growth rates (see Supplementary Movie 6). **b** Closest distance from any resistant cell to the nearest point on the tumor front over time for the simulation run in **a**. Orange indicates treatment holidays, grey 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.

to the position-dependent growth and velocity values measured in the reference 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 training environment.

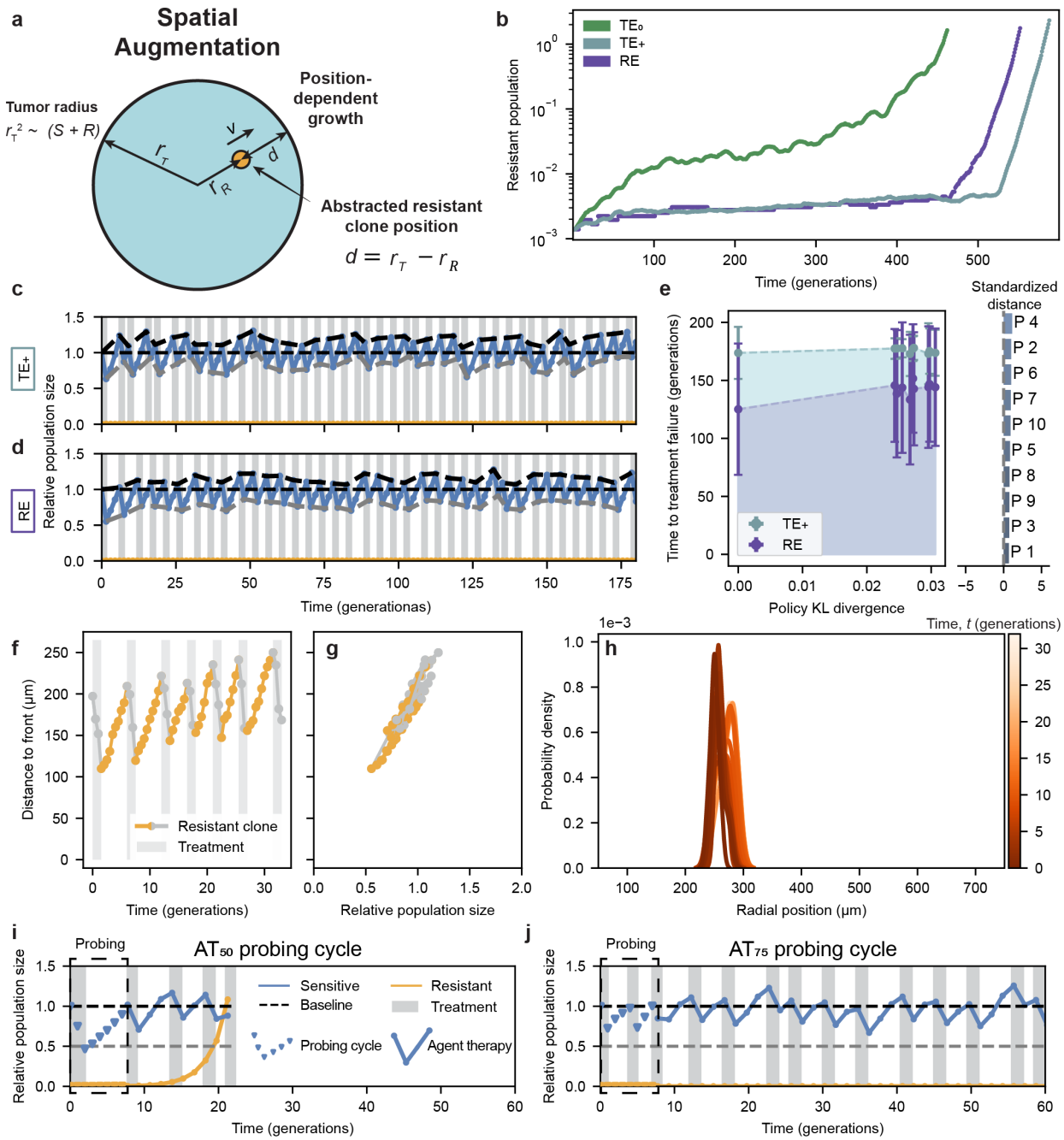

Figure 5: **Augmenting the training environment to reflect spatial dynamics.** **a** Schematic of the augmented environment (TE_+). **b** Resistant population dynamics over time in the three environments, TE_0 (green), TE_+, (grey-blue), and RE (purple). **c**, **d** Relative population size dynamics 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). Grey bars indicate treatment application. Dashed lines show the tumor burdens at treatment initiation (black) and treatment pause (grey). **e** Time to treatment failure (TTF) in TE_+ (grey-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 x-axis based on their Kullback-Leibler divergence, using the poorest-performing policy as reference. **f** Closest distance from any resistant cell to the nearest point on the tumor front over time for the simulation run shown in **d**. Grey 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_+.

We validate these augmentations by evaluating the original agents — trained in 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 (Supplementary Fig. 7). Interestingly, both reference and augmented environments show a biphasic growth pattern that is in line with reports from previous *in vitro* studies (Fig. 5b)[15].

Finally, we train a new cohort of agents in the augmented training environment (Fig. 5c,d) (Supplementary Fig. 8, Supplementary Movie 7). Cross-environment evaluation via GRAPE show that these second-generation agents not only converge during training but now also exhibit superior 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; Supplementary Fig. 9). 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).

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 [5, 37]. To more finely tailor therapy to individual patients, future trials might use an initial "probing cycle" for patient-specific model parameterization [44]. 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.

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 (5j; TTF$= 147 \pm 52$ generations, $n = 50$).

Discussion

In this work, we study how resistance-mediated therapy failure in solid tumors is shaped by emergent physical phenomena and how reinforcement learning can help to understand and harness these features to inform evolution-based adaptive therapy strategies. To this end, we introduce Reinforcement Failing, a framework for the RL-guided, human-in-the-loop refinement of abstracted training environments. Using cross-environment group-relative agent performance evaluation, Reinforcement Failing helps to expose previously hidden emergent processes that impact therapeutic outcomes.

Using this approach, we uncover that position-dependent growth and motion of resistant cells within the tumor mass can result in a treatment-induced ratchet-like conveyance of resistant cells to the front and subsequent disease progression. As a result, irreversible loss of therapeutic control can be triggered by even a single short burst of treatment. Using an augmented, physics-informed training environment, agents learn to avoid these hidden pitfalls, yielding superior therapeutic performance. In a clinical context, our results expose the potential risks of short but overly aggressive probing cycles for patient-specific model parameterization and outline mitigation strategies.

On a more general level, 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[12–14, 19]. 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 [25, 34, 35].

As with any modeling framework, Reinforcement Failing 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 reference environment. 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 feedbacks. Third, bridging the gap between simplified and fully realistic models can become computationally demanding, especially for large-scale, three-dimensional tumors. 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, restricting immediate translational application [22]. Future work will 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 our findings. One key avenue is to extend Reinforcement Failing to more physiologically detailed simulators that better incorporate immunological components, angiogenesis, or patient-specific data. A particularly exciting avenue of research will be the integration of immunotherapy into the adaptive therapy framework. Such expansions could help refine adaptive therapy strategies in personalized oncology[9].

On the artificial scientific discovery level, refining methods that intentionally promote policy diversity will help to ensure that the group-level performance analysis yields rich insights into underexplored parts of the decision space. Beyond oncology, employing Reinforcement Failing in other complex systems — such as ecological management[45], autonomous driving[46], or robotics[24] — could similarly reveal hidden emergent processes that hamper straightforward RL-based optimization. Integrating the obtained insights into more sophisticated multi-fidelity pipelines can further reduce computational overhead while still capturing critical details of the complex real-world system.

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) [47], 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. 10), beyond which cells become quiescent. The parameters governing this oxygen-dependent behavior are informed by experimental data from Grimes et al. [48]. A full list of parameters is provided in the shared repository.

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 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%.

Reinforcement learning (RL) was performed by wrapping the aforementioned PhysiCell simulations in a custom Gymnasium environment [49]. Communication between the RL environment and the simulation was implemented using ZeroMQ [50], 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 [19, 32, 40].

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 reference 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 reference environment when sensitive cells dominate and resistant fraction remains low.

However, 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 [41]. 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. 11. 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 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[44]. 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 [51], a PyTorch-based library providing reliable RL algorithm implementations. We trained the agents using Proximal Policy Optimization (PPO) [52], 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. Hyperparame-

ters (e.g., budget, entropy coefficient, clip range, batch size) remained consistent across agents and are available in the shared code repository.

Agents 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.

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 (Supplementary Fig. 10). 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 reference environment simulations are presented in Supplementary Fig. 12.

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. [5] 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. 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. 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).
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., 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).

12. 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).

13. 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).

14. 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).

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

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

17. 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).

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

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

20. 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).

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

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

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

24. 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).

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

26. 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).

27. 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).

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

29. 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.

30. 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).

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

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

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

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

35. 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).

36. 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).

37. 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).

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

40. 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).

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

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

43. 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).

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

45. 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).

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

47. 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).

48. 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).

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

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

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

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

53. 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.