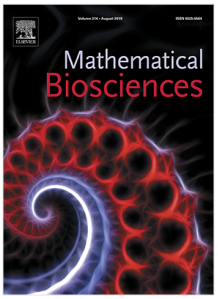
Journal Pre-proof



DDQN-based optimal targeted therapy with reversible inhibitors to combat the Warburg effect

Jose M. Sanz Nogales, Juan Parras, Santiago Zazo

PII: S0025-5564(23)00085-8
DOI: https://doi.org/10.1016/j.mbs.2023.109044
Reference: MBS 109044

To appear in: *Mathematical Biosciences*

Received date : 25 February 2023
Revised date : 9 May 2023
Accepted date : 23 June 2023

Please cite this article as: J.M.S. Nogales, J. Parras and S. Zazo, DDQN-based optimal targeted therapy with reversible inhibitors to combat the Warburg effect, *Mathematical Biosciences* (2023), doi: https://doi.org/10.1016/j.mbs.2023.109044.

DDQN-based optimal targeted therapy with reversible inhibitors to combat the Warburg effect

Jose M. Sanz Nogales, Juan Parras, Santiago Zazo

Information Processing and Telecommunications Center, Universidad Politécnica de Madrid, ETSI Telecomunicación, Av. Complutense 30, 28040 Madrid, Spain.

Abstract

We cover the Warburg effect with a three–component evolutionary model, where each component represents a different metabolic strategy. In this context, a scenario involving cells expressing three different phenotypes is presented. One tumour phenotype exhibits glycolytic metabolism through glucose uptake and lactate secretion. Lactate is used by a second malignant phenotype to proliferate. The third phenotype represents healthy cells, which performs oxidative phosphorylation. The purpose of this model is to gain a better understanding of the metabolic alterations associated with the Warburg effect. It is suitable to reproduce some of the clinical trials obtained in colorectal cancer and other even more aggressive tumours. It shows that lactate is an indicator of poor prognosis, since it favours the setting of polymorphic tumour equilibria that complicates its treatment. This model is also used to train a reinforcement learning algorithm, known as Double Deep Q-networks, in order to provide the first optimal targeted therapy based on experimental tumour growth inhibitors as genistein and AR-C155858. Our in silico solution includes the optimal therapy for all the tumour state space and also ensures the best possible quality of life for the patients, by considering the duration of treatment, the use of low-dose medications and the existence of possible contraindications. Optimal therapies obtained with Double Deep Q-networks are validated with the solutions of the Hamilton-Jacobi-Bellman equation.

Keywords: The Warburg effect, Optimal inhibition targeted therapy, Genistein, AR-C155858, Double Deep Q-Networks.

1. Introduction

Metabolism can be understood as a cell strategy for the production of the energy that is needed to survive and proliferate. Healthy cells obtain energy from glucose through two major metabolic pathways known as glycolysis and oxidative phosphorylation (OXPHOS) [1]. In glycolysis, cells consume 2 molecules of adenosine triphosphate (ATP) and obtains 4 molecules of ATP, by breaking down one glucose molecule into two pyruvate molecules [2]. Colloquially speaking, an ATP molecule can be understood as the elementary form of energy for the cell. Thus, during glycolysis, cells obtain an average energy of 2 ATP molecules per glucose molecule. In OXPHOS, pyruvate enters the citric acid cycle (Kreb's cycle)

Email address: sanz_nogales@yahoo.es (Jose M. Sanz Nogales)

in the mitochondria and 24–28 ATP molecules are generated from one glucose molecule converted into pyruvate (see Chapter 12 in [3]). Glycolysis is an anaerobic metabolic pathway, but OXPHOS necessarily requires oxygen. Under normal oxygen concentration, normoxic conditions, healthy cells make use of both metabolic pathways and obtain a net yield of 26–30 ATP molecules per glucose molecule. Only under hypoxia, such as intense physical exercise conditions, do cells shift their metabolism from OXPHOS towards anaerobic glycolysis in order to cover the punctual energy demand that occurs when the oxygen is scarce.

The Warburg effect is a metabolic alteration, known as aerobic glycolysis, where tumour cells avoid OXPHOS and base their entire metabolism on glycolysis even in normoxic conditions. As a reminder, conventional metabolism (glycolysis plus OXPHOS) produces up to 28 ATP molecules under normoxic conditions, whereas glycolysis produces only 2 ATP molecules. It is still unclear why tumour cells prefer this inefficient metabolism. It is thought that tumour cells consume large amounts of glucose, possibly in order to compensate for energy deficiencies, and ferment lactic acid as well. Lactic acid (or simply lactate) contributes to the acidification of the environment, which is harmless to tumour cells but detrimental to healthy ones [4–7]. In addition, it is now known that lactate, long considered a waste product of glycolytic metabolism, is used by malignant cells as an extra energy fuel (see e.g. [8, 9]) to proliferate and reproduce. Uncontrolled cell growth also leads to vascularization problems for healthy cells. In contrast, malignant cells are able to overcome these vascularization problems through sustained angiogenesis. In sustained angiogenesis, also considered in [10, 11] as a hallmark of cancer, some tumour cells secrete vascular endothelial growth factor to stimulate the growth of nearby blood vessels, which ensures the continuous supply of nutrients and oxygen. In addition, access to the bloodstream allows tumour cells to establish distant niches [12, 13], to later reach other organs, thus generating secondary tumours in the form of metastases [14]. Therefore, the Warburg effect can be understood as a very complex evolutionary process, which alters the environmental conditions to provide competitive advantage to malignant cells. Moreover, the Warburg effect triggers other unwanted alterations, such as sustained angiogenesis and metastasis. Even though the Warburg effect is not universal [15], it has been observed in a wide range of cancer types, including colorectal cancer [16–18], glioblastoma [19, 20], bladder [21], kidney [22, 23], breast [24, 25], melanoma [26, 27], pancreatic cancer [28–30], lung [31, 32], prostate [33], thyroid [34, 35], liver [36–38] and stomach [39–42]. All of these reasons contribute to our belief in this paper that the elimination of the Warburg effect, or at the very least, its mitigation, may be relevant to the cure of cancer.

From seminal work [43], Evolutionary Game Theory (EGT) has gained much popularity in cancer research, due to its ability to model cell populations that express different phenotypes, and that compete with each other according to the metabolic strategy that they show. The replicator equation (RE) is probably the most widespread deterministic dynamics in EGT, which has also been used in [44–47] to cover complex interactions that take place in a cell population subjected to the Warburg effect. RE states that the growth rate in the number of types that express a strategy depends on the fitness of

such strategy within the population. In this paper we propose a three component evolutionary model, a component per each metabolic strategy (glycolytic, non-glycolytic and oxidative), whose dynamic is governed by RE. In our approach, cells with glycolytic strategy express phenotypes that uptake glucose and secrete lactate. In contrast, non-glycolytic cells simply absorb lactate, while oxidative cells perform conventional phosphorylation with oxygen. Diffusible factors as glucose, lactate and oxygen stimulate non-linear cell responses (see e.g. [48]) when their ligands bind into the cell receptors . This stimulation is considered in RE through the fitness of each metabolic strategy. Similarly to [49–52], here we consider the Michaelis–Menten equation as a plausible way to model the fitness.

The vast majority of studies focused on EGT propose therapeutic treatments based on radiotherapy, chemotherapy and immunotherapy. Experimental treatments based on tumour growth inhibitors have received much less attention. In contrast to cytotoxic treatments, inhibitors target to cells that express specific phenotypes, by preventing the ligands of some diffusible factors, such as glucose and lactate, from binding the cell receptors. In this way, inhibitors seek to cancel cellular responses, thus avoiding the development of malignant phenotypes. In this paper, we differentiate two types of inhibitors: competitive and non–competitive inhibitors. Competitive inhibitors compete with the diffusible factors for binding into the cell receptors. In this way, a cell receptor cannot bind a diffusible factor when is blocked by a competitive inhibitor. In contrast, non–competitive inhibitors block the cell responses by bidding into the ligands of the diffusible factors. The ligand of a diffusible factor cannot bind a cell receptors once it is bound to the inhibitor. To the best of our knowledge, [44] was the first to mathematically formulate the potential effect of these drugs in cancer therapies. However, we miss the design of effective therapeutic treatment based on these drugs. In this paper, we propose the first optimal targeted therapy based on the combination of experimental tumour growth inhibitors to annul the Warburg effect. We also provide in silico results with application to colorectal cancer and other more aggressive tumours.

Standard cancer treatment consists of alternating drug sessions at Maximum Tolerated Dose (MTD) followed by time off (drug holidays). This approach seeks to kill as many cancer cells as possible with MTD, while controlling the toxic burden of drugs and side effects through rest days. However, MTD only succeeds in eradicating therapy-sensitive tumour cells, thereby providing competitive advantage to resistant cells [53]. In such a case, it may occur an uncontrolled growth of tumour cell whose traits are resistant to therapy (competitive release [54]). To avoid this, the doctor can change the medication to attack the resistant cells, producing a rebound of the cells sensitive to the previous medication. This is falling into a vicious circle that is not recommended, since it may lead to disease chronification. On the other hand, it can also happen that sensitive traits recover during drug-free times. In this other case, the physician may reduce resting times, thus increasing the toxic effect of the drugs in the patient's body. In this paper, we think that formulating therapy programming as if it were an optimal control problem can be a much better alternative to the conventional one. The general idea is to drive tumour dynamics to a safe state, which implies the cure of patients, or at least their long-term survival, at the lowest possible cost. With this purpose, the physician has to decide the dose concentration of each drug,

which needs to be supplied according to the tumour state. This approach results more attractive than an MTD-based alternative, since the optimal control problem allows doctors to act actively against the tumour. In other words, optimal control transforms the therapy problem into a Stackelberg Evolutionary Game [55], where the physician is the leader who anticipates the tumour state and acts as a rational player seeking to minimize the cost of an objective function.

The main concern in any therapy is to remove the presence of malignant cells. However, meeting this objective does not necessarily guarantee the safeguarding of the patients' life or the improvement of their quality of life. It may also be relevant to consider other factors such as the duration of the treatment, the toxicity of the drugs, the intensity of adverse side effects, the patient's pathologies, age, weight, etc. All of these factors make it difficult to formulate a problem aimed at providing the best possible therapy. Furthermore, in case of formulating such a problem, there are no guarantees of finding the optimal solution (or at least a good enough one) due to possible non-convexities. We can find very recent efforts with in silico results in the field of optimal cancer therapy in [56–60]. The authors of [56–58] achieve optimal chemotherapy and immunotherapy by applying Pontryagin's maximum (or minimum) principle. Similarly, [59] applies Forward Backward Sweep, an algorithm based on Pontryagin's maximum principle, to deliver optimal doses of abiraterone in prostate cancer. Importantly, Pontryagin's maximum principle provides necessary conditions for optimal control. However, these conditions are not sufficient, unless the problem meets certain convexity conditions. Convexity limits the formulation of a general therapeutic optimal control problem that may be more effective. In contrast, authors in [60] propose a bang-bang control by solving the Hamilton-Jacobi-Bellman (HJB) equation of a tumor dynamics, subjected to a cost function that penalizes the duration of the treatment, the delivery of chemotherapy doses and specific terminal states. Different from Pontryagin's maximum principle, the HJB equations provide necessary and sufficient conditions for an optimum, regardless of the problem's convexity. However, HJB requires perfect knowledge of tumour dynamics equations. This condition can be relaxed by applying model–free controllers, which do not require explicit knowledge of the environment or system. Several of these controllers have already been used in the treatment of cancer (see e.g. [61–63]), but we miss quantitative results to show their effectiveness as compared to the optimal solution.

In this paper, we present the best therapy obtained with Double Deep Q-Network (DDQN), which is one of the most popular model–free algorithms. Far from conventional approaches, in this article we do not propose the implementation of optimal therapies based on chemotherapy, immunotherapy or radiotherapy, but rather on experimental tumor growth inhibitors. Our solution is more complete and complex than a bang-bang control, since it allows the combination of different drugs in different doses. In addition, our solution is targeted therapy because it allows physicians to attack several Warburg effect symptoms simultaneously or separately. Our solution is complete because it applies to the entire problem state space. We validate the resulting therapeutic solutions by comparing them with those obtained with HJB, because our cost function is non-convex to consider the toxicity of the treatment, possible contraindications, and adverse side effects in a more general way than is usually done in the literature.

This paper is structured as follows. Section 2 presents a novel three-component evolutionary model to represent the Warburg effect. The effect of inhibitors over malignant phenotypes is also considered in this section. Section 3 formally introduces the problem of optimal therapy. In Section 4, the problem of optimal therapy is restated from the perspective of Markov decision processes (MDP). The use of DDQN to solve this problem is introduced and justified. In this section, we also introduce the solutions provided by HJB. This section ends with the parametrization of the optimal control problem and with the parametrization of most of our tumour dynamics. In Section 5, we discus the results of our evolutionary tumour growth model and the optimal therapy obtained with DDQN. Concretely, it explores the conditions that favour the establishment of polymorphic equilibria and discuss whether lactate toxicity plays a relevant role in tumour development. This section also compares the performance of DDQN with other more conventional therapeutic strategies. It shows the solution with the optimal targeted therapy in the full state space, which is provided by DDQN. This section ends by comparing the performance of DDQN with HJB in a specific case. Conclusions are presented in Section 6.

2. The model

2.1. Tumour state

The tumour state or state of the system refers to the diversity of the phenotypes of a cell population.

Let $\mathcal{L} \triangleq \{S, L, O\}$ denote the set of diffusible factors S, L, and O, which represent glucose, lactate, and oxygen, respectively. Let G^S denote a glycolytic phenotype with anaerobic metabolism. These cells consume glucose through glucose transporters (GLUT), and secrete lactate as an end product, regardless of the oxygen concentration that is available in the cell population. Let R^L denote a phenotype which absorbs lactate in presence of monocarboxylate transporters (MCT). These cells uptake lactate as an extra energy fuel. Let H^O denote a normal phenotype which uptakes oxygen. These cells base their metabolism on conventional oxidative phosphorylation. Let $\mathcal{M} \triangleq \{G^S, R^L, H^O\}$ denote the set of all phenotypes expressed by the cells. This set includes the metabolic strategies which are available in the cell population. Let $x_m(t)$ denote the relative frequency of phenotype $m \in \mathcal{M}$ in the cell population. The state of the population is given by:

$$\boldsymbol{x}(t) \triangleq (x_m(t))_{m\in\mathcal{M}} \in \boldsymbol{\Delta}^{|\mathcal{M}|}, \tag{1}$$

where $\boldsymbol{\Delta}^{|\mathcal{M}|} \triangleq \{\boldsymbol{x}(t) \in \Re^{|\mathcal{M}|} : 0 \leq x_m(t) \leq 1, \sum_{m\in\mathcal{M}} x_m(t) = 1\}$ denotes the simplex of $(|\mathcal{M}| - 1)$ dimensions in $\Re^{|\mathcal{M}|}$.

2.2. Diffusible factors

The diffusible factor concentrations stimulate the proliferation of the phenotypes. Let $\boldsymbol{a} \triangleq (a_\ell)_{\ell\in\mathcal{L}} \in \Re^{|\mathcal{L}|}_{\geq 0}$ denote the growth factor concentrations in normal conditions. This vector represents external resources provided by the host. Let $b_L \geq 0$ denote the amount of lactate secreted by each cell that expresses G^S. We introduce a definition describing the growth factor concentrations.

Definition 1. *The expected growth factor concentration which stimulates to phenotype $m \in \mathcal{M}$, denoted $d_m : \boldsymbol{\Delta}^{|\mathcal{M}|} \mapsto \Re_{\geqslant 0}$, is defined as:*

$$d_{G^S}\left(\boldsymbol{x}(t)\right) \triangleq a_S, \tag{2}$$

$$d_{R^L}\left(\boldsymbol{x}(t)\right) \triangleq a_L + b_L x_{G^S}(t), \tag{3}$$

$$d_{H^O}\left(\boldsymbol{x}(t)\right) \triangleq a_O. \tag{4}$$

Expression (2) shows the expected glucose concentration which stimulates to the growing of G^S. Expression (3) depicts the lactate concentration which stimulate to R^L. Recall that a_L in (3) is the lactate concentration which is present in the cell population in normal conditions. Lactate is uniformly spread in the cell population, and $b_L x_{G^S}$ expresses in (3), the part of lactate which is produced by G^S and that stimulates to R^L. Equation (4) indicates the expected oxygen concentration which stimulates to H^O.

2.3. Reversible inhibitors

We now introduce the concentration of therapeutic drugs in the host's organism. Let C and $\bar{C}$ denote the set of *competitive* and *non-competitive* inhibitors, respectively. Let $\mathcal{K} \triangleq C \cup \bar{C}$ denote the set of reversible inhibitors, which meets $C \cap \bar{C} = .$ Let $\Re_{\geq 0}$ denote the set of non-negative real numbers. Let $u_{mk}(t)$ denote the concentration of inhibitor $k \in \mathcal{K}$, which targets to phenotype $m \in \mathcal{M}$. The concentration of drugs is given by:

$$\boldsymbol{u}(t) \triangleq \left(u_{mk}(t)\right)_{m\in\mathcal{M},k\in\mathcal{K}} \in \Re_{\geq 0}^{|\mathcal{M}||\mathcal{K}|}. \tag{5}$$

Expression (5) can be understood as the control action that indicates the dose concentration of inhibitors applied by the physician.

2.4. Fitness and lactate toxicity

Fitness refers to the benefit obtained by phenotypes, when diffusible factors ligands bind to cell receptors. Fitness determines the reproductive capacity of phenotypes. Usually, diffusible factors have hyperbolic effects on the fitness of phenotypes. Reversible inhibitors may neutralize the fitness in two different ways. Competitive inhibitors block cell receptors where diffusible factors bind. Non-competitive inhibitors neutralize the cellular response when they bind to diffusible factor ligands. Let $\beta_m \geqslant 0$ denote the affinity of phenotype $m \in \mathcal{M}$ for the diffusible factor that stimulate it. Let $\beta_{mk} \geqslant 0$ denote the inhibitory constant. We first introduce a definition with the fitness of phenotypes G^S and R^L.

Definition 2. *The fitness of glycolytic and abnormal oxidative phenotypes, denoted $f_m : \boldsymbol{\Delta}^{|\mathcal{M}|} \times \Re_{\geq 0}^{|\mathcal{M}||\mathcal{K}|} \mapsto$*

$\Re_{\geqslant 0}$, *is defined for all* $m \in \mathcal{M}, m \neq H^O$, *as follows:*

$$f_m\left(\boldsymbol{x}(t), \boldsymbol{u}(t)\right) \triangleq \begin{cases} \dfrac{1}{1+\left(\dfrac{\beta_m}{d_m\left(\boldsymbol{x}(t)\right)}\right)\left(1+\dfrac{u_{mk}(t)}{\beta_{mk}}\right)} & \text{if } k \in C, \quad (6a) \\ \dfrac{1}{\left(1+\dfrac{\beta_m}{d_m\left(\boldsymbol{x}(t)\right)}\right)\left(1+\dfrac{u_{mk}(t)}{\beta_{mk}}\right)} & \text{if } k \in \bar{C}, \quad (6b) \\ \dfrac{1}{1+\dfrac{\beta_m}{d_m\left(\boldsymbol{x}(t)\right)}} & \text{otherwise.} \quad (6c) \end{cases}$$

Expression (6c) is a normalized version of the Michaelis–Menten equation [64]. This equation represents the fitness of phenotypes in absence of inhibitors, and shows a hyperbolic shape which is linear when the diffusible factor concentration is very low. Interestingly, reference [65] predicts GLUT1 (glucose transporter 1) kinetics with a normalized version of the Michaelis–Menten equation and [66] states the same for MCT1 (monocarboxylate transporter 1) kinetics. Thus (6c) seems to be a plausible function for the fitness of phenotypes G^S and R^L. Term β_m, affinity in (6a) – (6c), represents the diffusible factor concentration which makes $m \in \mathcal{M}$ responds with half the maximum. Equations (6a) and (6b) are accepted in [64, 67–72] as formal expressions that characterize the impact of reversible inhibitors. Expression (6a) depicts the fitness of phenotypes in the presence of competitive inhibitors. Competitive inhibitors reduce the affinity from β_m in (6c) to $\beta_m(1 + u_{mk}(t)/\beta_{mk})$ in (6a). Expression (6b) indicates the fitness of phenotypes in the presence of non-competitive inhibitors. This class of drug reduces the maximum fitness from 1 in (6c) to $\beta_{mk}/(u_{mk}(t) + \beta_{mk})$ in (6b).

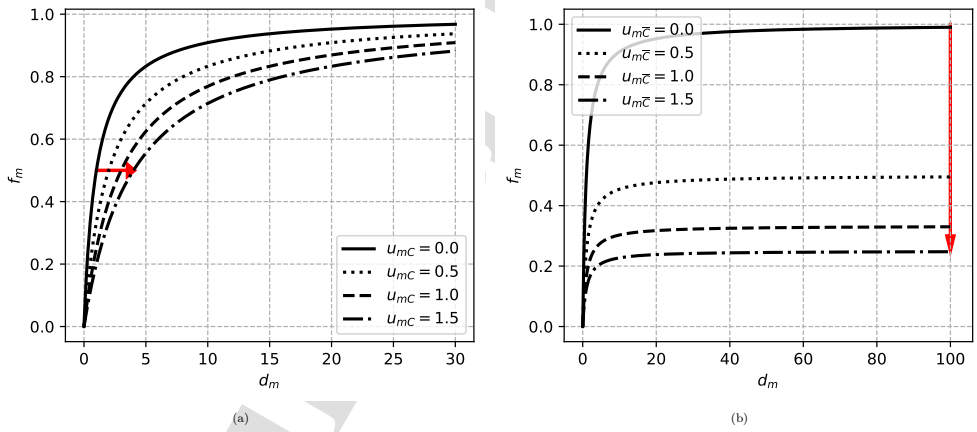


Figure 1: Cellular fitness vs. dose concentration of reversible inhibitors. Settings: $\beta_m = 1, \beta_{mC} = \beta_{m\bar{C}} = 0.5$. **(a)** Effect of competitive inhibitors on cellular fitness. **(b)** Effect of non-competitive inhibitors on cellular fitness. Competitive inhibitors shift the affinity to the right while non-competitive inhibitors reduce the maximum fitness. In the presence of inhibitors, tumour phenotypes need more diffusible factor to keep the same fitness.

Let $\ell^{sup} \geqslant 0$ denote the toxicity threshold for lactate. Let $\theta_m \geqslant 0$ denote the impact of lactate on phenotype $m \in \mathcal{M}$. Similarly to (6a)–(6c) and without loss of generality, the fitness of H^O is a function with domain $\boldsymbol{\Delta}^{|\mathcal{M}|} \times \Re_{\geq 0}^{|\mathcal{M}||\mathcal{K}|}$, which is defined as:

Definition 3. *The fitness of conventional oxidative phenotypes, denoted* $f_m : \boldsymbol{\Delta}^{|\mathcal{M}|} \times \Re_{\geq 0}^{|\mathcal{M}||\mathcal{K}|} \mapsto \Re$*, is defined for all* $m \in \mathcal{M}, m = H^O$*, as follows:*

$$f_m\left(\boldsymbol{x}(t), \boldsymbol{u}(t)\right) \triangleq \begin{cases} \dfrac{1}{1 + \dfrac{\beta_m}{d_m\left(\boldsymbol{x}(t)\right)}} & \text{if } d_{R^L}\left(\boldsymbol{x}(t)\right) < \ell^{sup}, \quad (7a) \\ \dfrac{1}{1 + \dfrac{\beta_m}{d_m\left(\boldsymbol{x}(t)\right)}} - \theta_m\left(d_{R^L}\left(\boldsymbol{x}(t)\right) - \ell^{sup}\right) & \text{otherwise .} \quad (7b) \end{cases}$$

Reference [73] states that healthy phenotypes respond with a Michaelis–Menten function. Similarly, expression (7a) provides the fitness of H^O under normal conditions, i.e. when the acidity of the environment does not prevent or hinder the normal development of this phenotype. Scalar ℓ^{sup} in (7a) and (7b) can be understood as a tolerance threshold of H^O to lactate. Lactate concentrations higher than ℓ^{sup} penalize the fitness of H^O with an extra cost $\theta_{H^O} d_{R^L}\left(\boldsymbol{x}(t)\right)$ in (7b). This cost is only considered in the fitness of cells with conventional oxidative metabolism, because authors in [4–7] suggest that acidosis given by glycolysis may result toxic to healthy cells, and harmless to cancerous cells. Term $\theta_m \ell^{sup}$ in (7b) provides a smoother and more natural response in the presence of lactate, by avoiding the discontinuity at $d_{R^L}\left(\boldsymbol{x}(t)\right) = \ell^{sup}$.

2.5. Cell population dynamic

RE is the deterministic differential equation most extended in EGT. Let $N_m \in \Re_{\geq 0}$ denote the number of cells that express phenotype $m \in \mathcal{M}$. RE states that the per capita growth rate in the number of cells that express a phenotype is equal to the fitness of the phenotype [74]:

$$\frac{\dot{N}_m(t)}{N_m(t)} \triangleq f_m\left(\boldsymbol{x}(t), \boldsymbol{u}(t)\right), \forall m \in \mathcal{M}. \quad (8)$$

In this way, RE proposes the reproduction and survival of the fittest types. However, conventional RE does not include the effects of external agents to the population, and states that fitness only depends on state $\boldsymbol{x}(t)$. In coherence with previous subsections, expression (8) also includes the effect of therapeutic actions through drug concentrations represented by $\boldsymbol{u}(t)$. Let $\dot{x}_m$ denote the dynamic of $m \in \mathcal{M}$ that matches RE. A straightforward calculus allows to express (8) in relative frequencies (see Appendix A):

$$\dot{x}_m(t) \triangleq x_m(t)\left(f_m\left(\boldsymbol{x}(t), \boldsymbol{u}(t)\right) - \sum_{n \in \mathcal{M}} f_n\left(\boldsymbol{x}(t), \boldsymbol{u}(t)\right) x_n(t)\right), \forall m \in \mathcal{M}. \quad (9)$$

Equation (9) states that the growth rate expressed in relative frequency of a phenotype $m \in \mathcal{M}$ is given by the difference between its fitness $f_m\left(\boldsymbol{x}(t), \boldsymbol{u}(t)\right)$, and the averaged fitness of the population:

$$F\left(\boldsymbol{x}(t), \boldsymbol{u}(t)\right) \triangleq \sum_{n \in \mathcal{M}} f_n\left(\boldsymbol{x}(t), \boldsymbol{u}(t)\right) x_n(t). \quad (10)$$

Recall the per capita growth rate remains being (8). Thus (8) and (9) are two ways of expressing the same idea (the survival and reproduction of individuals depend on their fitness within the group), but (8) is a map $\dot{N}_m : \Re^{|\mathcal{M}|} \times \Re_{\geqslant 0}^{|\mathcal{M}||\mathcal{K}|} \mapsto \Re$, while (9) is a map $\dot{x}_m : \boldsymbol{\Delta}^{|\mathcal{M}|} \times \Re_{\geqslant 0}^{|\mathcal{M}||\mathcal{K}|} \mapsto \boldsymbol{\Delta}$.

It is well known that (9) is the most common way to express the RE (see e.g. [75, 76]), since it provides dynamic responses which are bounded on simplex $\boldsymbol{\Delta}$. This property results interesting, specially in those cases where the difference in the number of individuals which express each of the types is high.

3. The problem

In this section we present the problem of optimal therapy.

Let σ be a positive or null scalar, which penalizes the duration of treatment. Let $u_{mk}^{max} \geq 0$ denote the maximum tolerated concentration of inhibitor $k \in \mathcal{K}$, which targets to $m \in \mathcal{M}$, and that is set at physician discretion, according to the characteristics of the patient (age, weight, medical history, etc.) Let σ_{mk} be a positive scalar, which weights the relative toxicity of mk – inhibitor with respect to others. It is used to consider medical contraindications or side effects of some drugs over others. Let $x_{H^O}^{min} \in (0,1)$ denote the minimum allowable relative frequency for phenotype H^O, which is compatible with the life of the patient. Lower relative frequencies imply the death or the administration of palliative care to the patient. Let $x_{H^O}^{max} \in (0,1)$ denote the threshold that x_{H^O} has to overcome in order for the patient to be cured. We now formally introduce the cost function of the problem:

Definition 4. *The cost function, denoted* $J : \boldsymbol{\Delta}^{|\mathcal{M}|} \times \Re_{\geq 0}^{|\mathcal{M}||\mathcal{K}|} \mapsto \Re_{\geqslant 0}$*, is defined as follows:*

$$J\left(\boldsymbol{x}(t), \boldsymbol{u}(t)\right) \triangleq K(\boldsymbol{x}(t)) + \int_{t_0}^{t_f} \mathcal{L}\left(\boldsymbol{x}(t), \boldsymbol{u}(t)\right) dt, \tag{11}$$

such that:

$$K(\boldsymbol{x}(t)) \triangleq \begin{cases} \infty & \text{if } x_{H^O}(t_f) < x_{H^O}^{min}, \quad (12a) \\ 0 & \text{otherwise,} \quad (12b) \end{cases}$$

and

$$\mathcal{L}\left(\boldsymbol{x}(t), \boldsymbol{u}(t)\right) \triangleq h(\boldsymbol{x}(t)) + e(\boldsymbol{u}(t)), \tag{13}$$

where

$$h(\boldsymbol{x}(t)) \triangleq \begin{cases} 0 & \text{if } x_{H^O}(t) > x_{H^O}^{max}, \quad (14a) \\ \sigma & \text{otherwise,} \quad (14b) \end{cases}$$

$$e(\boldsymbol{u}(t)) \triangleq \sum_{m \in \mathcal{M}, k \in \mathcal{K}} \sigma_{mk} \frac{u_{mk}(t)}{u_{mk}^{max}}. \tag{15}$$

Then $K(\boldsymbol{x}(t))$ is a terminal cost function that severely penalizes therapy failure. The second term in (11) is the cost-to-go function or the trajectory cost from state $\boldsymbol{x}(t_0)$ to $\boldsymbol{x}(t_f)$, and is used to evaluate the way by which a final state is reached from an initial state. Concretely, $h(\boldsymbol{x}(t))$ increases the cost of the treatment during the time it takes place, while $e(\boldsymbol{u}(t))$ regulates both the doses and the toxicity.

Let $T \geqslant 0$ denotes the time that treatment lasts. The goal consists on finding the optimal control therapy $\boldsymbol{u}^*\left(\boldsymbol{x}(t)\right)$, which minimizes the cost function (11) from $\boldsymbol{x}(t=0)$ to $\boldsymbol{x}(t=T)$:

$$\begin{aligned} \boldsymbol{u}^*\left(\boldsymbol{x}(t)\right) &\triangleq \arg\min J\left(\boldsymbol{x}(t), \boldsymbol{u}(t)\right), \\ \text{s.t.} \quad \dot{x}_m(t) &= x_m(t)\left(f_m\left(\boldsymbol{x}(t), \boldsymbol{u}(t)\right) - F\left(\boldsymbol{x}(t), \boldsymbol{u}(t)\right)\right), \forall m \in \mathcal{M}. \end{aligned} \tag{16}$$

In the next section, we indicate how to solve (16) for all the state space of the problem.

4. Materials and methodology

4.1. Double Deep Q-Network

In this subsection we focus on solving (16), with a reinforcement learning (RL)-based technique called Double Deep Q-Network (DDQN).

Let us frame the problem described in Section 3 in the context of Markov decision process (MDP). An MDP is an extension of Markovian processes, where the agent (a physician in our case) takes actions sequentially (by administering different concentrations of competitive and/or non-competitive inhibitors) to drive the system (the cell population) to a specific state (state involving cure of the patient). A discrete-time MDP can be defined with a tuple $\langle \mathcal{X}, \mathcal{A}, \mathcal{P}, \mathcal{R} \rangle$, where:

- $\mathcal{X}$ is the set of states.
- $\mathcal{A}$ is the set of actions.
- $p : \mathcal{X} \times \mathcal{A} \to [0, 1]$ is the transition probability function, where $p\left(x_{t+1}|x_t, a_t\right)$ denotes the probability of getting state $x_{t+1} \in \mathcal{X}$, given that the state of the system is $x_t \in \mathcal{X}$ and the agent plays action $a_t \in \mathcal{A}$.
- $r : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to \Re$ is the reward function, where $r\left(x_t, a_t, x_{t+1}\right)$ denotes the reward perceived by the agent, when the system goes from state x_t to sate x_{t+1} after playing action a_t.

The goal of any MDP is to maximize the expected cumulative reward in an infinite time horizon:

$$\max \mathbb{E}\left(\sum_{t=0}^{\infty} \gamma^t r\left(x_t, a_t, x_{t+1}\right)\right), \tag{17}$$

where $\gamma \triangleq (0, 1)$ is called *the discount factor*.

We now reformulate the problem described in Section 3 and pose it as an MDP problem. Suppose that $\mathcal{X}$ is a discrete version of $\mathbf{\Delta}^{|\mathcal{M}|}$, and that $p\left(x_{t+1}|x_t, a_t\right)$ is a deterministic transition probability function set by (9). Let $\mathcal{U}_{km}$ denote a discrete set with the dose concentrations of inhibitor $k \in \mathcal{K}$ that targets to phenotype $m \in \mathcal{M}$. The set of actions, $\mathcal{A} \triangleq \prod_{k \in \mathcal{K}, m \in \mathcal{M}} \mathcal{U}_{km}$, represents the Cartesian product of sets $\mathcal{U}_{km}$, and includes all possible combination of drugs that can be used in a inhibitor–based therapy. In this way, any action $a_t \in \mathcal{A}$ is unequivocally defined by a specific combination of drugs. We also define two terminal state sets. The first terminal state set, $\mathcal{X}_{end1} \triangleq \{x_t \in \mathcal{X} : x_{HO} < x_{HO}^{min}\}$, occurs when therapy fails. In that case, when therapy fails, the agent perceives a high enough penalty $c \in \Re_{<0}$. The second terminal state set is $\mathcal{X}_{end2} \triangleq \{x_t \in \mathcal{X} : x_{HO} > x_{HO}^{max}\}$ and implies that therapy succeed. We are ready to reformulate cost function (11) as an MDP reward function:

$$r\left(x_t, a_t, x_{t+1}\right) \triangleq \begin{cases} c & \text{if } x_{t+1} \in \mathcal{X}_{end1}, \quad (18a) \\ 0 & \text{if } x_t \in \mathcal{X}_{end2}, \quad (18b) \\ -\left(\sigma + e(a_t)\right) & \text{otherwise.} \quad (18c) \end{cases}$$

With the problem posed according to an MDP, the solution can be obtained by solving the Bellman equation [77, 78]. The Bellman equation, denoted as $v : \mathcal{X} \to \Re$, is defined as follows:

$$v\left(x_t\right) \triangleq \max_{a_t \in \mathcal{A}} \left(r\left(x_t, a_t, x_{t+1}\right) + \gamma \sum_{x_{t+1} \in \mathcal{X}} p\left(x_{t+1} | x_t, a_t\right) v\left(x_{t+1}\right) \right), \forall x_t \in \mathcal{X}. \quad (19)$$

Expression (19) provides the value function or the maximum expected long term return of state x_t. In short, it indicates how good or bad such a state is.

The Bellman equation, given by (19), is called state value function, to emphasize that it is defined in terms of states. We can also find another function called q–function, which expresses a similar idea as (19) does, but in terms of state–action pairs instead of just states. The q–function, or state-action value function, is denoted as $q : \mathcal{X} \times \mathcal{A} \to \Re$ and is defined as follows:

$$q\left(x_t, a_t\right) \triangleq r\left(x_t, a_t, x_{t+1}\right) + \gamma \sum_{x_{t+1} \in \mathcal{X}} p\left(x_{t+1} | x_t, a_t\right) \max_{a_{t+1} \in \mathcal{A}} q\left(x_{t+1}, a_{t+1}\right), \forall x_t \in \mathcal{X}, \forall a_t \in \mathcal{A}. \quad (20)$$

Analogously to the Bellman equation, now q–function provides the maximum expected long term return of playing action $a_t \in \mathcal{A}$ at state $x_t \in \mathcal{X}$.

Let us recall that expressions (19) and (20) are equivalent, in the sense that (19) refers to states (indicates how good a state is), whereas (20) refers to state-action pairs (indicates the effectiveness of executing an action in a certain state). In fact, both expressions can be derived from each other as follows:

$$v\left(x_t\right) = \max_{a_t \in \mathcal{A}} q\left(x_t, a_t\right), \quad (21)$$

$$q\left(x_t, a_t\right) = r\left(x_t, a_t, x_{t+1}\right) + \gamma \sum_{x_{t+1} \in \mathcal{X}} p\left(x_{t+1} | x_t, a_t\right) v\left(x_{t+1}\right), \forall x_t \in \mathcal{X}, \forall a_t \in \mathcal{A}. \quad (22)$$

So far, we have taken the problem posed in Section 3, formulated it as an MDP, and solved it with either (19), (20), (21), or (22). However, this approach requires us to know transition probabilities $p\left(x_{t+1} | x_t, a_t\right)$ and we may not have access to this information. Therefore, we need a methodology to approximate the Bellman equation or the q–function without knowing the transition probabilities or the tumour's dynamic equations. RL offers a framework to address this problem.

Fig. 2 depicts a typical RL-based algorithm schematic. In this scenario, an agent observes state x_t and executes action a_t. As a consequence of the action, the scenario changes from state x_t to x_{t+1} and the agent receives a reward $r\left(x_t, a_t, x_{t+1}\right)$. According to our scenario, the environment in Fig. 2 is defined by (9), the reward is defined by (18), and the actions are defined by combination of drugs listed later in Table 3. In this way, q–function can be estimated with a model–free techniques as Q–learning, which

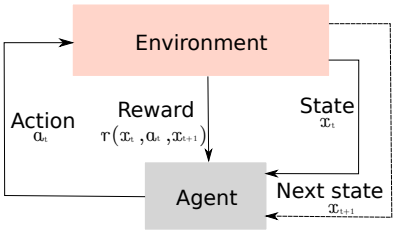


Figure 2: Generalized RL–scheme.

is probably the simplest RL–based algorithm that can be used to solve complex problems as (16). Let $\alpha \in (0,1)$ denote a scalar called *the learning rate*. Q-learning updates the value of playing action $a_t \in \mathcal{A}$ when the state of the system is $x_t \in \mathcal{X}$, as follows (see e.g. [77]):

$$q\left(x_t, a_t\right) \leftarrow (1-\alpha)\, q\left(x_t, a_t\right) + \alpha \left(r\left(x_t, a_t, x_{t+1}\right) + \gamma \max_{a_{t+1} \in \mathcal{A}} q\left(x_{t+1}, a_{t+1}\right) \right), \forall x_t, x_{t+1} \in \mathcal{X}. \tag{23}$$

Note that (23) estimates (20) without considering transition probabilities and without access to dynamic (9). The convergence of (23) is achieved by applying the scheme of Fig. 2 iteratively. Once expression (23) converges, the optimal control is given by the actions that deliver the maximum expected return at each state.

Q-learning requires a discrete state-action set, while $\mathbf{\Delta}^{|\mathcal{M}|}$, the state space in the problem at hand, is continuous. A discretized version of $\mathbf{\Delta}^{|\mathcal{M}|}$ would lead to a state-action set with such a large dimensionality that it would made (23) be computationally infeasible. This drawback is overcome with neural networks, considered as universal approximation functions, with the ability to map from continuous states $\boldsymbol{x}(t) \in \mathbf{\Delta}^{|\mathcal{M}|}$ to discrete actions $a_t \in \mathcal{A}$. In this way, we move from Q-learning to Deep Q-learning (or Deep Q-networks) by posing (16) as an MDP, with direct access to the continuous state space. Let θ and θ' denote the weights of the policy and the target nets respectively. Different from Q-learning, Deep Q-networks (DQN) now estimates the state-action value function (20) with a Q-function:

$$Q\left(x_t, a_t; \theta\right) \triangleq r\left(x_t, a_t, x_{t+1}\right) + \gamma \max_{a_{t+1} \in \mathcal{A}} Q\left(x_{t+1}, a_{t+1}; \theta'\right), \tag{24}$$

where

$$L\left(\theta\right) \triangleq \mathbb{E}\left(r\left(x_t, a_t, x_{t+1}\right) + \gamma \max_{a_{t+1} \in \mathcal{A}} Q\left(x_{t+1}, a_{t+1}; \theta'\right) - Q\left(x_t, a_t; \theta\right) \right)^2 \tag{25}$$

is the loss function, which updates the weights of the policy net by using the gradient descent algorithm.

However, the original DQN algorithm may overestimate the value of the actions under certain conditions [79]. DDQN can alleviate this overestimation by introducing a slight change in (24). Different from DQN, DDQN involves both the policy and target networks in maximizing the estimation of the Q-function. Concretely, the policy net selects the action to be used (i.e., the one that maximizes the value), but the value is taken from the target network:

$$Q\left(x_t, a_t; \theta\right) \triangleq r\left(x_t, a_t, x_{t+1}\right) + \gamma Q\left(x_{t+1}, \arg\max_{a_{t+1} \in \mathcal{A}} Q\left(x_{t+1}, a_{t+1}; \theta\right); \theta' \right). \tag{26}$$

We make use of this variation of DQN, as it does introduce a negligible increment on the computational load and may improve significantly the algorithm results. Our network topology consists on a single fully–connected hidden layer with 72 neurons and *relu* activation function. We set $\gamma = 0.9992, \alpha = 0.001$ and also implement ϵ*–greedy policy* from $\epsilon_{max} = 1.0$ to $\epsilon_{min} = 0.0001$ with decay $2.222 \cdot 10^{-5}$ per episode. Our *replay memory* consists on 512 experiences which update the *target network* every 10 episodes. An episode ends when the problem trajectory reaches one of the terminal states defined above. Full details with Python code implementation of DDQN, environment (9) and reward function (18) are available at https://github.com/jsanno/ddqn.

4.2. The Hamilton–Jacobi–Bellman equation

The Hamilton–Jacobi–Bellman (HJB) equation is the most important equation in non-linear control theory and one of the most popular methodologies, for solving deterministic continuous time optimal controls. It was also used in [80] to solve problems similar to (16). In this paper, we solve the HJB equation in order to compare the goodness of the solutions obtained with DDQN and validate them. The HJB equation states the condition for the value function:

$$\frac{\partial V}{\partial t} \triangleq - \min_{\boldsymbol{u}(t) \in \mathcal{U}} \left(\frac{\partial V}{\partial \boldsymbol{x}}^{\top} f\left(\boldsymbol{x}(t), \boldsymbol{u}(t)\right) + \mathcal{L}\left(\boldsymbol{x}(t), \boldsymbol{u}(t)\right) \right). \tag{27}$$

As the value function represents the minimum expected long term cost subject to a dynamic $f\left(\boldsymbol{x}(t), \boldsymbol{u}(t)\right)$, then problem (16) can be reformulated in terms of the value function, as follows:

$$\begin{aligned} & V\left(\boldsymbol{x}(t_0), t_0, t_f\right) \triangleq \min_{\boldsymbol{u}(t) \in \mathcal{U}} J\left(\boldsymbol{x}(t), \boldsymbol{u}(t), t_0, t_f\right), \\ & \text{s.t.} \quad \frac{d\boldsymbol{x}}{dt} = f\left(\boldsymbol{x}(t), \boldsymbol{u}(t)\right), \end{aligned} \tag{28}$$

where $\mathcal{U}$ denotes the set of admissible controls, which is equal to $\Re_{\geq 0}^{|\mathcal{M}||\mathcal{K}|}$ for the problem we are dealing with, and $f\left(\boldsymbol{x}(t), \boldsymbol{u}(t)\right)$ denotes the dynamic under control, i.e. the RE introduced in Subsection 2.5.

However, in most cases it is very difficult and even intractable to obtain a classical solution of problem (28), i.e., a continuous and differentiable value function, since the minimization operator in (27) implies solving a system with non-linear partial differential equations. For this reason, we solve the HJB equation numerically with a tool called BocopHJB, which can be found available for free at [81]. BocopHJB proposes a numerical approximation, which consists of discretizing the state space, to later iteratively estimate the value function through dynamic programming. Concretely, let $N \in \mathbb{Z}$ denote the number of time steps. Let $h_0 \triangleq \frac{t_f}{N}$ represents the time step size. Then, the time for any step $k \in \mathbb{Z}$ if given by $t_k \triangleq h_0 k$. Algorithm 1 includes the BocopHJB's pseudocode to compute the value function at t_k.

Algorithm 1 BOCOPHJB: Compute value function at t_k

Require: $0 \leq k \leq N$

for $x \in$ Grid **do**
  if $k == N$ **then**
    $V_k(x) \Leftarrow K(t_f)$
  else
    $\boldsymbol{x}_{k+1} \Leftarrow x + h_0 f(\boldsymbol{u}, x)$
    $V_k(x) \Leftarrow \min_{\boldsymbol{u}\in\mathcal{U}} \left(h_0 \mathcal{L}(t_k, \boldsymbol{u}, x) + \mathbb{E}_x \left[V_{k+1}(\boldsymbol{x}_{k+1})\right]\right)$
  end if
end for

4.3. Control problem parametrization

Let us start with the parametrization of the functions introduced in Definitions 2 and 3. This parametrization can also be found summarized in Table 1. We first focus on the findings of [82], on the glucose and lactate concentrations in colorectal liver metastasis. Authors in this reference do not appreciate significant differences in the glucose concentrations of healthy and tumour tissues. They state 17.1 ± 3.6 mM and 17.2 ± 1.5 mM, for the glucose concentrations found in healthy and tumour tissues, respectively. For this reason we set $a_S = 13.5 - 20.7$ mM in Table 1. It can be found in the same reference, that the lactate concentrations in healthy tissues is 1.7 ± 0.3 mM. This parameter is later tuned in Subsection 5.1.1 to reproduce some clinical results.

Reference [4] indicates that blood lactate concentrations for healthy and cancerous tissues are $1.5 - 3$ mM and $10 - 30$ mM, respectively. Here we understand that lactate becomes toxic for H^O from 10 mM and assign $\ell^{sup} = 10$ mM in Table 1. However, we do not have medical data to characterize the negative impact of lactate over the healthy cells. For the moment, we set $\theta_{H^O} = 0.0 - 0.01$ M^{-1} in Table 1 and tune this parameter in Subsection 5.1.3 in order to reproduce the clinical results observed in the literature. Similarly, we set $b_L = 9.5 - 28.3$ in Table 1 and tune this parameter in Subsections 5.1.1 and 5.1.2.

Glucose transporter 1 (GLUT1) is overexpressed in colorectal cancer [83]. Recall that GLUT1 is one of fourteen proteins which are responsible for the uptake of glucose across the cell membrane. Reference [84] provides estimations with the affinity of GLUT1 for glucose in $3 - 7$ mM. For this reason we set $\beta_{G^S} = 3 - 7$ mM in Table 1. Monocarboxylate transporter 1 (MCT1) is also overexpressed in colorectal cancer [85]. Recall that MCT1 is associated to oxidative cells and lactate uptake [86–89]. According to [88–90], the affinity of MCT1 for lactate is about $3 - 10$ mM. For this reason we set $\beta_{R^L} = 3 - 10$ mM.

Reference [73] fits the affinity of healthy phenotypes for the oxygen at 14.37 μM. In addition, hypoxia and normoxia conditions are established in [91] with oxygen tensions 1–5% and 10–21%, respectively. By replacing these data and $\beta_{H^O} = 14.37$ μM in (7a), we obtain 0.14–0.75 μM and 1.6–3.82 μM as the oxygen concentrations in hypoxic and normoxic conditions, respectively. Since the Warburg effect is a metabolic alteration that occurs in normoxic conditions, we set $a_O = 3.82$ μM in Table 1, for simplicity

Diffusible factors		
Factor	**Value**	**Ref.**
Glucose (normal conditions)	$a_S = 13.5 - 20.7$ mM	[82]
Lactate (normal conditions)	$a_L = 1.4 - 2$ mM	[82]
Oxygen (normoxia)	$a_O = 3.82$ μM	[91]
Lactate tolerance	$\ell^{sup} = 10$ mM	[4]
Lactate toxicity	$\theta_{H^O} = 0.0 - 0.01$ M^{-1}	Section 5.1.3
Lactate production	$b_L = 9.5 - 28.3$ mM $x_{G^S}^{-1}$	Sections 5.1.1, 5.1.2
Phenotypes		
Metabolism	**Affinity**	**Ref.**
Glycolytic	$\beta_{G^S} = 3 - 7$ mM	[84]
Oxidative	$\beta_{R^L} = 3 - 10$ mM	[88–90]
	$\beta_{H^O} = 14.37$ μM	[73]
Reversible inhibitors		
Inhibitor	**Inhibitory constant**	**Ref.**
Isoflavone genistein	$\beta_{G^S C} = 7$ μM	[92]
AR-C155858	$\beta_{R^L \bar{C}} = 2.3$ nM	[93]

Table 1: Fitness parametrization.

and to ensure that the environment is well enough oxygenated.

Isoflavone genistein is a competitive GLUT1 inhibitor with inhibitory constant equal to 7 μM [92], while AR-C155858 is a non-competitive inhibitor of MCT1 whose inhibitory constant is about 2.3 ± 1.4 nM [93]. With this information we set $\beta_{G^S C} = 7$ μM and $\beta_{R^L \bar{C}} = 2.3$ nM in Table 1.

We now assign numerical values to the problem established in Definition 4, and also include additional comments, regarding the implementation of the DDQN algorithm introduced in Subsection 4.1. All the parametrization of the control problem can also be found in Table 2.

Authors in [94] state that genistein at concentrations $5 - 200$ μM can arrest cell cycle by modulating regulatory proteins. In contrast, according to [95], the efficacy of MCT1 can be modulated with AR-C155858 at concentrations ranging from 329 nM to 819 nM. In this paper we set $u_{G^S C}^{min} = 51.03$ μM and $u_{G^S C}^{max} = 102.06$ μM as the minimum and maximum doses of genistein, which can be applied to the patient. Similarly, we decide $u_{R^L \bar{C}}^{min} = 2.7$ nM and $u_{R^L \bar{C}}^{max} = 5.4$ nM for the dose concentrations of AR-C155858. Note that $u_{G^S C}^{max}$ and $u_{R^L \bar{C}}^{max}$ are too far from the maximum concentration of genistein and AR-C155858 established in [94] and [95], respectively. In addition, with this parametrization we provide similar weights to genistein and AR-C155858 in (15) and (18c), since $u_{G^S C}^{min}/u_{G^S C}^{max} \approx u_{R^L \bar{C}}^{min}/u_{R^L \bar{C}}^{max}$. We assume that AR-C155858 has more adverse effects or contraindications than genistein, by assigning $\sigma_{G^S C} = 1, \sigma_{R^L \bar{C}} = 10$. We also penalize the duration of the treatment with $\sigma = 0.01$.

For the running of DDQN, we set $x_{H^O}^{min} = 0.1$ and $x_{H^O}^{max} = 0.9$. This algorithm is trained for 90,000 episodes, with 300 as the maximum number of steps per episode, and with step size of 2 cell generations. An episode ends when $x_{H^O}(t) < x_{H^O}^{min}$, $x_{H^O}(t) > x_{H^O}^{max}$ or when the number of iterations is 300. We set an extra penalization with $c = -1,000$ in (18a), whether $x_{H^O}(t) < x_{H^O}^{min}$.

Description	Value	Ref.
Minimum genistein dose	$u_{G^SC}^{min} = 51.03\ \mu$M	[94]
Maximum genistein dose	$u_{G^SC}^{max} = 102.06\ \mu$M	[94]
Minimum AR-C155858 dose	$u_{R^L\bar{C}}^{min} = 2.7$ nM	[95]
Maximum AR-C155858 dose	$u_{R^L\bar{C}}^{max} = 5.4$ nM	[95]
Medical contraindications	$\sigma_{G^SC} = 1, \sigma_{R^L\bar{C}} = 10$	–
Failed terminal state	$x_{HO}^{min} = 0.1$	–
Safe terminal state	$x_{HO}^{max} = 0.9$	–
Treatment duration cost	$\sigma = 0.01$	–
Failed therapy penalty	$c = -1,000$	–

Table 2: Control problem parametrization.

The action space is given by the dose concentrations of genistein and AR-C155858, which are applied to the patient in each iteration. This action space is collected in Table 3. Finally, recall that $\mathbf{\Delta}^{|\mathcal{M}|}$ is the state space, since the population state is defined as (1).

	Action								
	1	2	3	4	5	6	7	8	9
Genistein dose	0.0	0.0	0.0	$u_{G^SC}^{min}$	$u_{G^SC}^{min}$	$u_{G^SC}^{min}$	$u_{G^SC}^{max}$	$u_{G^SC}^{max}$	$u_{G^SC}^{max}$
AR-C155858 dose	0.0	$u_{R^L\bar{C}}^{min}$	$u_{R^L\bar{C}}^{max}$	0.0	$u_{R^L\bar{C}}^{min}$	$u_{R^L\bar{C}}^{max}$	0.0	$u_{R^L\bar{C}}^{min}$	$u_{R^L\bar{C}}^{max}$

Table 3: Control problem action space.

5. Results and discussion

This section is divided into two different parts:

- Subsection 5.1 uses the model presented in Section 2, in order to reproduce some observations obtained from clinical trials of colorectal cancer and other tumours. We also explore the conditions that favour the establishment of polymorphic equilibria and discus whether lactate toxicity plays a relevant role in tumour development.

- Subsection 5.2 presents the optimal therapeutic policy (the optimal targeted therapy solution in the complete state space) provided by DDQN for 4 different tumor dynamics identified in Subsection 5.1. We illustrate the performance of these policies, comparing them with other more conventional therapeutic routines. Finally, we validate the results provided by DDQN by comparing them with the numerical solutions of the HJB equation.

5.1. Model results

5.1.1. Monomorphic populations in colorectal cancer

In this subsection we parametrize our model with data obtained in [82] about colorectal liver metastasis. We show that our model is able to provide the same clinical finding about lactate concentration that can be found at [82]. We also provide further results with the evolution of cell populations.

Fig. 3a and 3b show how the phenotypic composition of cell populations evolves. Black lines are different trajectories in order to represent the overall dynamic of the cell population. The background colors on the simplex represent the modulus of the gradient associated with the dynamics. Specifically, yellow colors represent fast dynamics, while the blue and purple colors represent slower ones. The filled and hollow red dots represent stable and unstable state equilibria respectively. Fig. 3c and 3d report the time course of lactate concentrations in the cell populations.

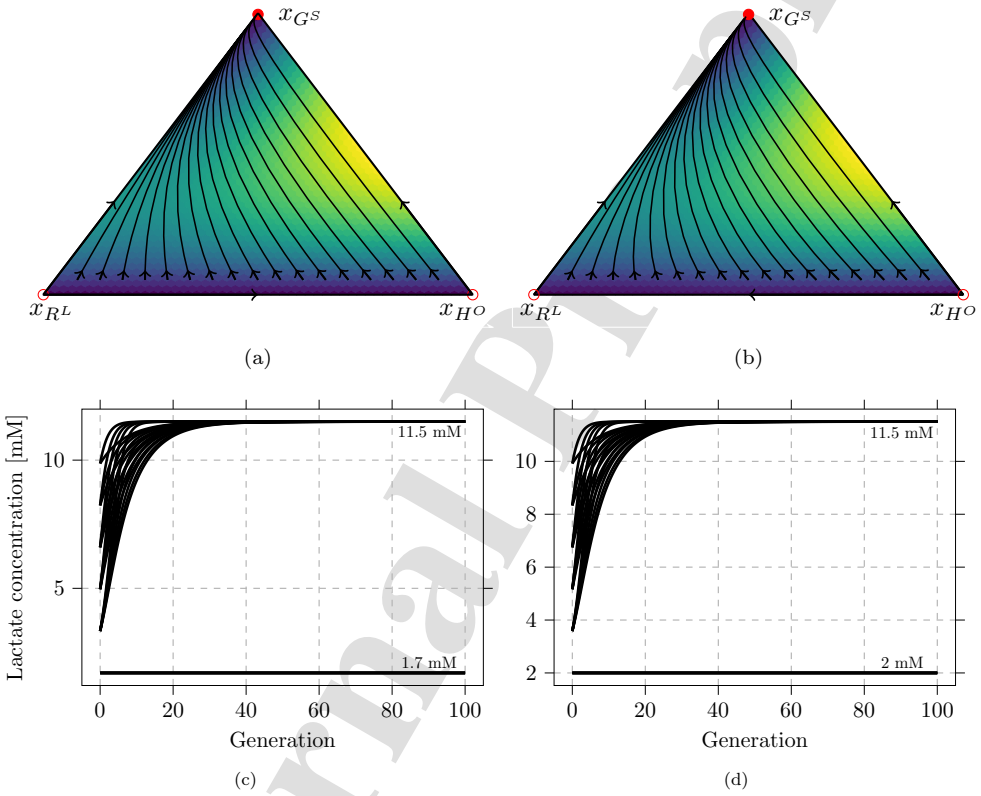


Figure 3: Reproducing some of the results provided by [82] regarding colorectal liver metastases. **(a),(b)** Tumour dynamics. **(c),(d)** Time course of lactate concentration. Settings in **(a),(c)** $a_L = 1.7$ mM, $b_L = 9.8$ mM $x_{G^S}^{-1}$. Settings in **(b),(d)** $a_L = 2$ mM, $b_L = 9.5$ mM $x_{G^S}^{-1}$. Settings in **(a),(b),(c),(d)**, $a_S = 17.1$ mM, $\beta_{G^S} = 5$ mM, $\beta_{R^L} = 6.5$ mM, $\theta_{H^O} = 0.01$ M^{-1} (the rest of parameters are included in Table 1).

Results in Fig. 3a show that H^O rejects invasions by R^L. In these cases, lactate level remains at 1.7 mM (see Fig. 3c). This level matches with the mean of lactate concentrations found in normal tissues [82]. In this way, a cell population composed by phenotype H^O corresponds to a healthy colorectal liver tissue in our model. Any other invasion collapses the population with phenotype G^S. In these other

cases, the lactate concentration grows up to 11.5 mM (see Fig. 3c), which matches with the mean of lactate concentrations found in cancerous tissues [82]. This result suggests that phenotype G^S counts with enough glucose to reproduce. In contrast, phenotype R^L do not have enough lactate under normal conditions, and phenotype G^S does not produce enough lactate to support R^L. Consequently, R^L tends to die out while G^S overcomes the cell population.

We now increase the lactate which is available in normal conditions (from $a_L = 1.7$ mM to $a_L = 2$ mM) and reduce the lactate producing capacity of G^S (from $b_L = 9.8$ mM $x_{G^S}^{-1}$ to $b_L = 9.5$ mM $x_{G^S}^{-1}$). Level 2 mM in Fig. 3d matches with the maximum lactate concentration which is found in healthy tissues [82]. Again, Fig. 3d also shows that general invasions drive G^S to fixation, while H^O and R^L are extinguished. Eventually, phenotype R^L is able to fixate in the population, but only in those invasions which do not include the presence of phenotype G^S.

Then it can be concluded that in the case of heterogeneous mutations, cells with glycolytic metabolism in colorectal liver tissues count with enough glucose resources to fixate in the population. In contrast, cells expressing other phenotypes tend to extinction. The end result in colorectal liver metastasis is a monomorphism given by populations with cells that express G^S.

5.1.2. Polymorphic populations

In this subsection we explore the conditions that favour polymorphisms in the Warburg effect. With this end, we take the same parameters as those used in Fig. 3a and 3c; but now, we increase the amount of lactate produced by G^S.

Fig. 4a and 4b show mixed strategy equilibria. Let $\boldsymbol{x}^* \in \boldsymbol{\Delta}$ denote a mixed strategy equilibria. These equilibria represent different polymorphisms that share two characteristics. First, phenotype H^O is extinct since it is not part of any of these equilibria (i.e. $x_{H^O}^* = 0$). Second, the fitness of G^S matches with the fitness of R^L (i.e. $f_{G^S}(\boldsymbol{x}^*) = f_{R^L}(\boldsymbol{x}^*)$). Please note that $x_{H^O}^* = 0$ and $f_{G^S}(\boldsymbol{x}^*) = f_{R^L}(\boldsymbol{x}^*)$ satisfy equilibrium conditions in (9). In addition, these equilibria are reached in Fig. 4c and 4d with lactate levels equal to 22.23 mM. Recall that cancerous tissues show lactate concentrations about $10-30$ mM (see e.g. [4]). Thus these polymorphisms represent tumour cell populations. The mixed equilibria in Fig. 4a and Fig. 4b are respectively at $\left(x_{H^O}^*, x_{G^S}^*, x_{R^L}^*\right) \approx (0, 0.95, 0.035)$ and $\left(x_{H^O}^*, x_{G^S}^*, x_{R^L}^*\right) \approx (0, 0.72, 0.26)$. Therefore, the producing capacity of G^S favours the presence of R^L in polymorphisms.

As it occurs in Fig. 3a and 3c, Fig. 4 shows that 1.7 mM is not enough lactate for R^L to proliferate. That is the reason why H^O rejects any invasion by R^L. Different from results provided in Fig. 3a and 3c, Fig. 4 reports now monomorphic equilibria with G^S, only in the case of invasions by this phenotype. In this way, 23 mM in Fig. 4c and 30 mM in Fig. 4d represent complete invasions by phenotype G^S.

In conclusion, according to data found at [82] and with the parametrization introduced in Subsection 4.3, it can be deduced that R^L needs 22.23 mM of lactate to match its fitness with G^S. This quantity of lactate is too high to be found in healthy tissues (see e.g. [4]). Thus, phenotype R^L requires G^S to produce extra lactate to ensure its survival, as well as the constitution of tumour polymorphisms.

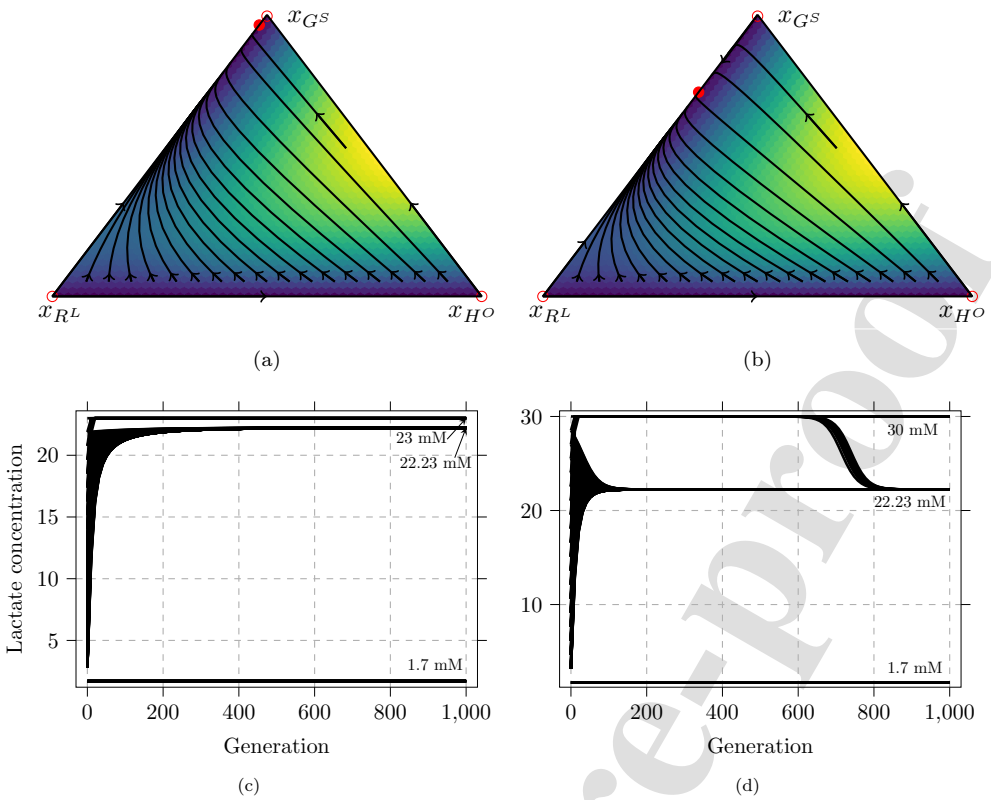


Figure 4: Lactate promotes polymorphic populations thus increasing the complexity of any targeted therapy. **(a),(b)** Tumour dynamics. **(c),(d)** Time course of lactate concentration. Settings in **(a),(c)** $b_L = 21.3$ mM $x_{G^S}^{-1}$. Settings in **(b),(d)** $b_L = 28.3$ mM $x_{G^S}^{-1}$. Settings in **(a),(b),(c),(d)** $a_S = 17.1$ mM, $a_L = 1.7$ mM, $\beta_{G^S} = 5$ mM, $\beta_{R^L} = 6.5$ mM, $\theta_{H^O} = 0.01$ M^{-1} (the rest of parameters are included in Table 1).

5.1.3. Lactate toxicity in colorectal cancer

Many authors as [4–7] argue that lactate can be poisonous to healthy cells and innocuous to cancer cells. In this subsection we review the influence of lactate toxicity on colorectal liver metastasis.

In previous subsections we set $\theta_{H^O} = 0.01$ M^{-1}. Suppose that lactate is safe for H^O regardless of its concentration in the population; that is, we now set $\theta_{H^O} = 0$. Suppose that under normoxic conditions, cells H^O are at the best possible scenario before invasion occurs. In such a scenario, the fitness of H^O should be as high as possible, while the fitness of R^L and G^S should be as low as possible. This situation can be considered by selecting from Table 1 the following parametrization: $a_S = 13.5$ mM, $a_L = 1.4$ mM, $a_O = 3.82$ μM, $\beta_{G^S} = 7$ mM, $\beta_{R^L} = 10$ mM, $\beta_{H^O} = 14.37$ μM, and by setting $\theta_{H^O} = 0$, as well. Now, by replacing these parameters in (6c), (7a) and (7a), we obtain the following fitness for phenotypes H^O and G^S:

$$f_{H^O}(\boldsymbol{x}(t)) = 0.21, f_{G^S}(\boldsymbol{x}(t)) = 0.66, \forall \boldsymbol{x}(t) \in \boldsymbol{\Delta}. \tag{29}$$

Thus, the metabolic strategy of G^S is strictly superior than H^O, regardless of the state of the population. In other words, strategy H^O is strictly dominated by G^S. Therefore, under normoxic conditions (recall

that the Warburg effect occurs under normoxic conditions), a population of cells that express phenotype H^O succumbs to any invasion by G^S, regardless of whether the lactate is toxic or not.

We now examine the quantity of lactate that R^L needs to get a higher fitness than H^O; i.e. we want to know the concentration given by (3), which satisfies:

$$f_{R^L}\left(\boldsymbol{x}(t)\right) \geq f_{H^O}\left(\boldsymbol{x}(t)\right). \tag{30}$$

By replacing the previous parameters in (30), we obtain that lactate levels have to meet $s_{R^L}\left(\boldsymbol{x}(t)\right) \geq 2.66$ mM. Therefore, phenotype R^L needs at least 2.66 mM of lactate to get a higher fitness than H^O. Recall that lactate concentrations in colorectal liver are about 1.4–2 mM and 8.7–14,3 mM in healthy and cancerous tissues, respectively. Also recall that in general, lactate concentrations in normal and tumour tissues are 1.5–3 mM and 10–30 mM, respectively (see e.g. [4]). Other references as [96] even observe tumours with lactate concentrations up to 40 mM. Therefore, in a tumour environment and under normoxic conditions, phenotype R^L has enough lactate to get a higher fitness that H^O.

In conclusion, lactate toxicity does not seem to be a determining factor in the aggressiveness of a tumour, since malignant phenotypes have sufficient resources under normal conditions to lead the population to collapse in the event of any mutation.

5.2. Optimal DDQN based control solution

In this section we cover the optimal therapy solutions obtained with DDQN.

Fig. 5 shows the optimal DDQN solutions for the tumour state space. Concretely, Fig. 5a and Fig. 5b represent the optimal therapeutic policies on the monomorphic populations found in colorectal cancer in Subsection 5.1.1, while Fig. 5c and Fig. 5d refers to the optimal policies on the polymorphic populations covered in Subsection 5.1.2. Recall that parametrization is collected in Tables 1 and 2 with the action space defined in Table 3. The white zone located to the left of the simplex represents $\mathcal{X}_{end1}$, with all those terminal states where we assume that therapy fails. On the contrary, the area of the same color that is on the right corresponds to the terminal states where the therapy is successful, i.e. $\mathcal{X}_{end2}$. The black lines in Fig. 5 illustrate tumour dynamics subject to DDQN's optimal policy. One can get a better idea of the effect of this therapy, by comparing these trajectories with therapy-free tumour dynamics, i.e. Fig. 5a vs. Fig. 3a, Fig. 5b vs. Fig. 3b, Fig. 5c vs. Fig. 4a and Fig. 5d vs. Fig. 4a. Based on this comparison, it can be verified that optimal therapy leads tumour dynamics to the set of safe states, $\mathcal{X}_{end2}$.

According to Fig. 5, actions 7, 8 and 9 are the only which take part in the optimal policies obtained with DDQN, i.e., the actions from 1 to 6 are not part of any optimal therapy. This result can be useful in the design of real therapies, because it suggests a significant simplification of the dose combinations to be used. Furthermore, the dose combinations are always the same, regardless of whether the tumour is monomorphic or polymorphic, which suggests a possible standardization of the inhibitor cocktails to be used.

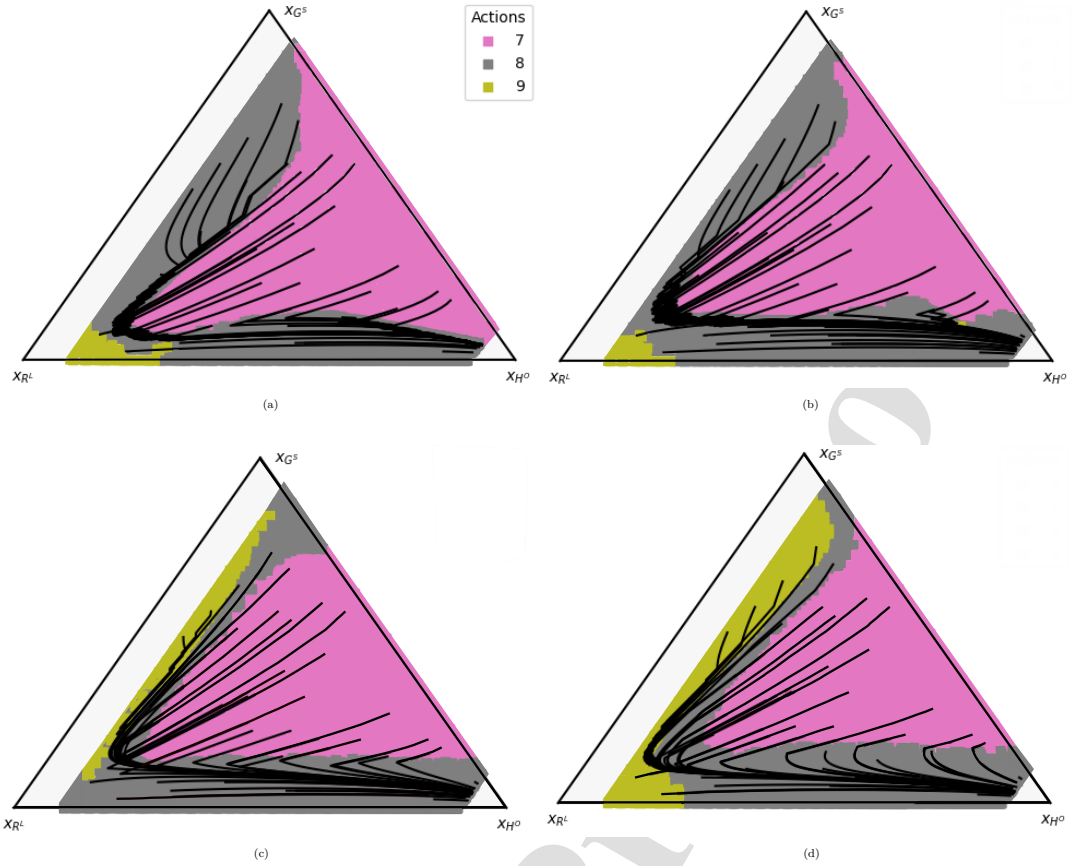


Figure 5: DDQN's optimal therapeutic policies (the optimal solution for the entire state space of the tumour). **(a),(b),(c)** and **(d)** shows the DDQN's optimal targeted therapy for tumour dynamics observed in Fig. 3a, 3b, 4a and 4b, respectively. Settings in **(a)** $a_L = 1.7$ mM, $b_L = 9.8$ mM $x_{G^S}^{-1}$. Settings in **(b)** $a_L = 2$ mM, $b_L = 9.5$ mM $x_{G^S}^{-1}$. Settings in **(c)** $a_L = 1.7$ mM, $b_L = 21.3$ mM $x_{G^S}^{-1}$. Settings in **(d)** $a_L = 1.7$ mM, $b_L = 28.3$ mM $x_{G^S}^{-1}$. Settings in **(a),(b),(c),(d)** $a_S = 17.1$ mM, $\beta_{G^S} = 5$ mM, $\beta_{R^L} = 6.5$ mM, $\theta_{H^O} = 0.01$ M^{-1} (the rest of parameters are included in Tables 1 and 2).

Interestingly, all of the optimal policies in Fig. 5 target G^S with the maximum tolerated genistein dose. In contrast, phenotype R^L is never targeted with the maximum tolerated AR-C155858 dose. These results suggest that DDQN learns that R^L can be attacked indirectly through G^S (remember that R^L receives support from lactate released by G^S). The maximum dose of AR-C155858 is never administered to patients, as a result of this reason and in order to minimize the costs associated with the therapy's toxicity.

Table 4 summarizes the average costs of each of policies represented in Fig. 5. These costs are the result of averaging 512 different trajectories with uniformly distributed initial states in the non-terminal state space, i.e., each initial state is obtained by sampling the space $\boldsymbol{\Delta}^{|\mathcal{M}|} - \mathcal{X}_{end1} - \mathcal{X}_{end2} = \{\boldsymbol{x}(t) \in \boldsymbol{\Delta}^{|\mathcal{M}|}, \boldsymbol{x}(t) \notin \mathcal{X}_{end1}, \boldsymbol{x}(t) \notin \mathcal{X}_{end2}\}$ uniformly. Recall from Subsections 5.1.1 and 5.1.2 that lactate contributes to the heterogeneity of phenotypes in the cell population. We also suggest that heterogeneity may increase tumour aggressiveness and complicate treatment. Now, Table 4 provides therapeutic costs

that increase from the dynamics observed in Fig. 5a to Fig. 5d. Consequently, lactate is a reliable indicator of poor prognosis and high therapeutic costs.

	Optimal therapy			
	Fig. 5a	Fig. 5b	Fig. 5c	Fig. 5d
DDQN Cost	396.10	424.36	504.51	556.67

Table 4: Average costs of the optimal therapeutic policies represented in Fig. 5.

5.2.1. Optimal DDQN therapy vs conventional therapy

This subsection aims to compare DDQN-based therapies with other more conventional approaches. Suppose that in a conventional therapy, the doctor decides to apply the following protocol:

$$a_t \triangleq \begin{cases} a_6 & \text{if } x_{G^S} \leqslant x_{R^L}, \quad (31a) \\ a_8 & \text{otherwise .} \quad (31b) \end{cases}$$

According to Tables 2 and 3, the therapy defined by (31a) and (31a) involves attacking the dominant tumour phenotype with the corresponding maximum tolerated dose, while the secondary phenotype is attacked with the minimum tolerated dose. In this way, the aim is to attack both tumour phenotypes at the same time, avoiding the excessive costs of applying the maximum tolerated doses at the same time.

Fig. 6 and Table 5 show the results in the case that no therapy is applied to the patient, in the case of implementing the conventional therapy defined above and in the case of using the optimal therapy obtained with DDQN. All the trajectories start from the same initial state $(x_{H^O}, x_{G^S}, x_{R^L}) = (0.3, 0.6, 0.1)$. In any case, the absence of treatment implies the loss of the patient in two iteration steps. Note that conventional therapy also fails in the cases with more aggressive tumours, represented by Fig. 6c and 6d. Recall again, as discussed in Subsection 5.1.2, that Figs. 6c and 6d represent scenarios with polymorphic equilibria that are generated due to the presence of high lactate concentrations. The fact that conventional therapy succeeds in the cases represented by Figs. 6a and 6b and fails in the cases of Figs. 6c and 6c is indicative that lactate contributes to tumour aggressiveness. For this reason, conventional therapy fails earlier (uses fewer steps in Table 5) in the scenario represented by Fig. 6d.

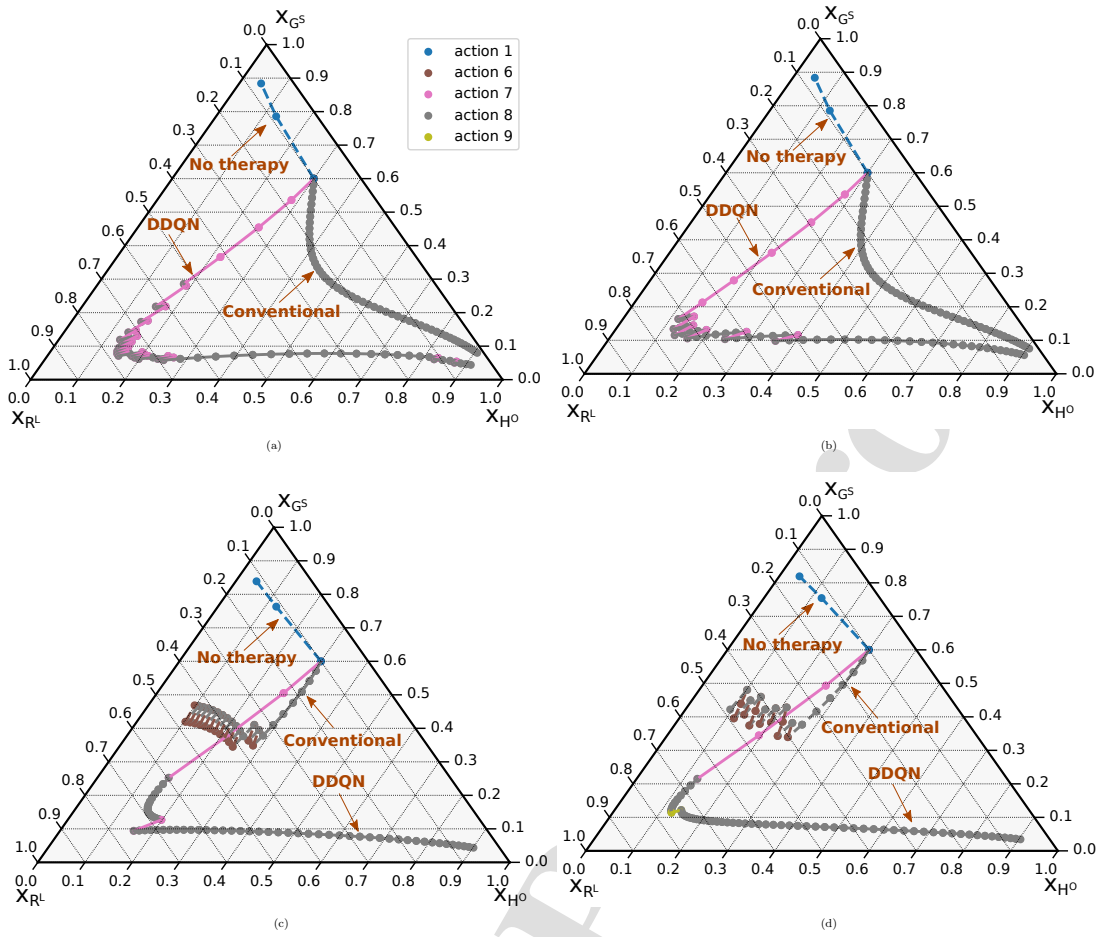


Figure 6: Illustration of tumour dynamics under no therapy, conventional therapy and DDQN's optimal therapy. No therapy always fails. Conventional therapy succeeds in **(a)** and **(b)**, but fails in more aggressive tumours **(c)** and **(d)**. DDQN's optimal targeted therapy always succeeds. Settings in **(a)** $a_L = 1.7$ mM, $b_L = 9.8$ mM $x_{G^S}^{-1}$. Settings in **(b)** $a_L = 2$ mM, $b_L = 9.5$ mM $x_{G^S}^{-1}$. Settings in **(c)** $a_L = 1.7$ mM, $b_L = 21.3$ mM $x_{G^S}^{-1}$. Settings in **(d)** $a_L = 1.7$ mM, $b_L = 28.3$ mM $x_{G^S}^{-1}$. Settings in **(a),(b),(c),(d)** $a_S = 17.1$ mM, $\beta_{G^S} = 5$ mM, $\beta_{R^L} = 6.5$ mM, $\theta_{H^O} = 0.01$ M^{-1} (the rest of parameters are included in Tables 1 and 2).

Fig. 6 shows that DDQN recovers the patient in all scenarios. Furthermore, Table 5 indicates lower therapeutic costs with DDQN, even though it approaches terminal failure states and employs a greater number of steps in patient recovery. DDQN would have obtained straighter trajectories towards the safe terminal state, without passing close to the failure terminal state, in the case of assigning greater relative weight to the penalty of treatment duration (parameter σ) over the toxicity of drugs (parameters σ_{mk} and u_{mk}^{max} for all $m \in \mathcal{M}, k \in \mathcal{K}$).

5.2.2. Validation of DDQN solutions with HJB

In this subsection we validate the optimal therapy solutions obtained with DDQN by comparing it with the solution provided by HJB.

In Fig. 7, we can compare the performance of DDQN vs. HJB. Figs. 5a and 7a respectively show the

	Fig. 6a		Fig. 6b		Fig. 6c		Fig. 6d	
	steps	cost	steps	cost	steps	cost	steps	cost
No therapy	2	1,000.02	2	1,000.02	2	1,000.02	2	1,000.02
Conventional therapy	52	625.04	53	637.06	34	1,504.66	24	1,366.46
DDQN therapy	62	545.24	58	587.16	62	705.24	66	773.32

Table 5: No therapy vs conventional therapy vs DDQN therapy: Iteration steps and therapeutic cost.

optimal policies obtained by DDQN and HJB, under the tumour dynamics represented in Fig. 3a. In Fig. 7b, we compare the therapeutic costs of 512 different trajectories. The initial state of each trajectory has been obtained by uniformly sampling the non-terminal state space. As it can be seen in Fig. 5a and Fig. 7a, the optimal therapies obtained with DDQN and HJB are apparently very different. Concretely, the optimal policy provided by HJB is much more complex, since it uses a significantly higher number of actions than DDQN.

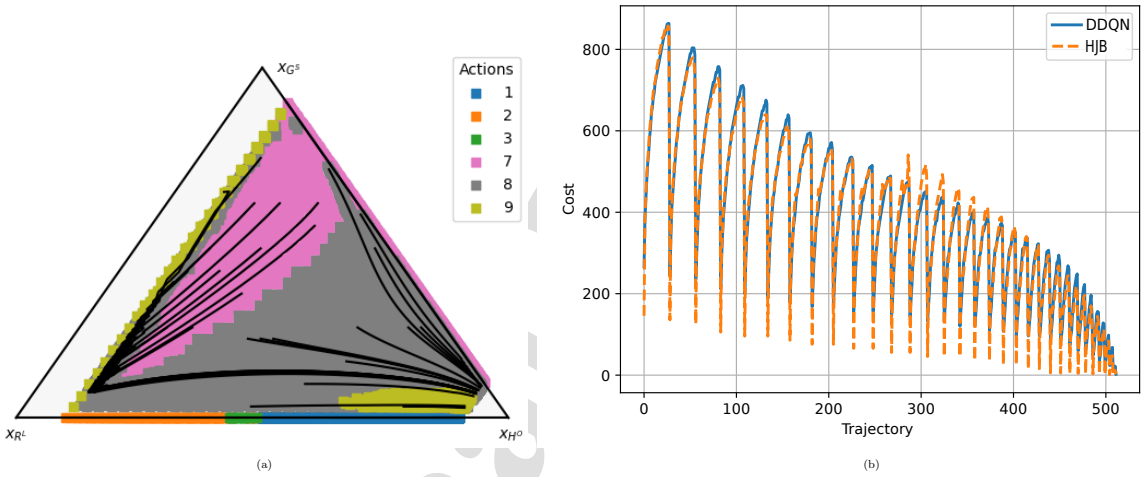


Figure 7: Optimal therapeutic policies: HJB vs. DDQN under tumour dynamics observed in Fig. 3a. **(a)** HJB's optimal targeted policy. HJB has an average cost of 394.37 compared to 396.10 (see Table 4) for DDQN. **(b)** Trajectory costs obtained with DDQN and HJB. Welch's t-test p-value: 0.87, thus there is no evidence that HJB and DDQN trajectory costs are different. Settings: $a_L = 1.7$ mM, $b_L = 9.8$ mM $x_{G^S}^{-1}$, $a_S = 17.1$ mM, $\beta_{G^S} = 5$ mM, $\beta_{R^L} = 6.5$ mM, $\theta_{H^O} = 0.01$ M^{-1} (the rest of parameters are included in Tables 1 and 2).

However, the results observed in Fig. 7b suggest that both policies are quite similar from the perspective of therapeutic costs. The average cost over the 512 trajectories are 396.10 and 394.37 for DDQN and HJB, respectively. Therefore, in this case, the HJB policy gets an improvement of 0.4 % over the DDQN policy, at the cost of increasing the complexity of the therapy, which can be a problem in the case of a real implementation.

Fig. 7b shows some trajectories where the cost obtained by HJB is greater than DDQN. This is surely

due to slight mismatches in the numerical approximations and interpolations applied by BocopHJB. In any case, a Welch's t-test on the samples in Fig. 7b provides a very high p-value equals to 0.87. Therefore, there is no evidence to support that the policies shown in Figs. 5a and 7a present different average costs, although visually they do not look alike. Very different policies can give similar cost results.

6. Conclusion and future works

Aerobic glycolysis has been considered for a long time as an inefficient metabolic disorder in obtaining energy for the cell development. The lactate generated in aerobic glycolysis has also been considered as a by-product or metabolic waste with no apparent utility. However, nowadays it is known that lactate can be used as an extra energy source by some cellular phenotypes to proliferate. In this way, from a evolutionary perspective, glycolytic metabolism may make sense even under normal oxygen conditions, since it allows to increase the polymorphic heterogeneity of tumours and thus favour their aggressiveness. In this work, a simple model based on EGT has been proposed to represent this complex metabolic alteration, a.k.a. the Warburg effect, which is common to many types of cancer. This model has been adequately parametrized to reproduce the clinical observations obtained from different studies on colorectal cancer and other more aggressive tumours. This model has also been used as a training scenario for control systems based on recent deep learning algorithms.

In this work, we propose the first optimal therapy based on experimental tumour growth inhibitors, which have been obtained through the efficient implementation of control systems based on deep learning. The results have been compared with the solutions provided by HJB. The conclusion is that the policies obtained with HJB slightly outperform DDQN, at the cost of increasing the complexity of therapeutic routines. In real life, the implementation of simpler routines such as those obtained with DDQN may make more sense, although these are suboptimal compared to those obtained with HJB. Furthermore, solving HJB is conditioned on an exact knowledge of the system to be controlled, which is infeasible in most of the real-life cases. DDQN does not need to know the differential equations that govern tumour dynamics, but it requires a sufficiently reliable scenario to train. The quality of the scenario used in the training of any reinforcement learning algorithm is key to get realistic optimal policies. However, in our case, the implementation of a realistic scenario requires many clinical observations that provide clear and precise information on how the tumour evolves over time. For this reason, in future work, we plan to refine the model presented in this paper, as the literature provides chemical, biological, and medical data that allow a more accurate understanding of tumour dynamics.

Determining the system state is also an important detail to consider. In this paper, we have considered that the state is a vector with components that represent the relative frequency of the phenotypes expressed by the cells. A Markov decision process can be used to model tumour dynamics in this case, since the state of the system is observable. Nevertheless, in many real-world applications, the state cannot be directly observed or accessible, and estimates may be affected by noise. This may require posing the problem from the perspective of a partially observable Markov decision process. Deep recurrent Q-

networks [97], an extension of DQN with recurrent networks, could also be useful to address these types of problems.

In this paper we have modelled tumour dynamics with deterministic differential equations. This approximation is useful to address general or average dynamics. However, tumour dynamics may also be associated with stochastic components. A natural way to address the problem of obtaining therapeutic treatments in this type of systems would be by implementing stochastic optimal controls.

In conclusion, we highlight that the results obtained in this paper on optimal policies are in silico. Furthermore, the present study has the limitations described above. In this regard, the results derived from these therapies should be viewed with caution since much work remains to be done in order to obtain optimal treatments against cancer.

Acknowledgements

This work was supported by the Spanish Ministry of Science and Innovation under the grant PID2020-112502RB-C41 (NAUTILUS).

Appendix A. Obtaining RE from expression (8)

Let $N \in \Re_{\geq 0}$ denote the size of a cell population. Recall from Section 2.5, that $N_m \in \Re_{\geq 0}$ denotes the the number of cells that express phenotype $m \in \mathcal{M}$. Then, the relative frequency of any phenotype, introduced in Section 2.1, is given by:

$$x_m(t) \triangleq \frac{N_m(t)}{N(t)}, \forall m \in \mathcal{M}. \tag{A.1}$$

The derivative of (A.1) with respect to time results:

$$\dot{x}_m(t) = \frac{\dot{N}_m(t)N(t) - N_m(t)\dot{N}(t)}{N^2(t)}, \forall m \in \mathcal{M}. \tag{A.2}$$

The derivative of $N(t)$ with respect to time also satisfies:

$$\dot{N}(t) \triangleq \sum_{n \in \mathcal{M}} \dot{N}_n(t). \tag{A.3}$$

By replacing (A.3) in (A.2):

$$\dot{x}_m(t) = \frac{\dot{N}_m(t)N(t) - N_m(t)\sum_{n \in \mathcal{M}} \dot{N}_n(t)}{N^2(t)}, \forall m \in \mathcal{M}. \tag{A.4}$$

Equation (8) can be expressed as follows:

$$\dot{N}_n(t) = N_n(t) f_n\left(\boldsymbol{x}(t), \boldsymbol{u}(t)\right), \forall n \in \mathcal{M}. \tag{A.5}$$

By replacing (A.5) in (A.4):

$$\dot{x}_m(t) = \frac{N_m(t) f_m\left(\boldsymbol{x}(t), \boldsymbol{u}(t)\right)}{N(t)} - \frac{N_m(t)\sum_{n \in \mathcal{M}} N_n(t) f_n\left(\boldsymbol{x}(t), \boldsymbol{u}(t)\right)}{N^2(t)}, \forall m \in \mathcal{M}. \tag{A.6}$$

Finally, expression (9) can be directly obtained by replacing $N_n(t) = N(t)x_n(t), \forall n \in \mathcal{M}$, from (A.1), in (A.6).

References

[1] Z. Wang, D. Guan, S. Wang, L. Y. A. Chai, S. Xu, K.-P. Lam, Glycolysis and oxidative phosphorylation play critical roles in natural killer cell receptor-mediated natural killer cell functions, Frontiers in immunology 11 (2020) 202.

[2] R. Chaudhry, M. Varacallo, Biochemistry, glycolysis (2018).

[3] Y. Demirel, Nonequilibrium thermodynamics: transport and rate processes in physical, chemical and biological systems, Elsevier, 2007.

[4] D. la cruz López, K. Griselda, L. J. Castro-Muñoz, D. O. Reyes-Hernández, A. García-Carrancá, J. Manzo Merino, Lactate in the regulation of tumor microenvironment and therapeutic approaches, Frontiers in oncology 9 (2019) 1143.

[5] K. Garber, Energy deregulation: licensing tumors to grow (2006).

[6] A. Ramanathan, C. Wang, S. L. Schreiber, Perturbational profiling of a cell-line model of tumorigenesis by using metabolic measurements, Proceedings of the National Academy of Sciences 102 (17) (2005) 5992–5997.

[7] R.-h. Xu, H. Pelicano, Y. Zhou, J. S. Carew, L. Feng, K. N. Bhalla, M. J. Keating, P. Huang, Inhibition of glycolysis in cancer cells: a novel strategy to overcome drug resistance associated with mitochondrial respiratory defect and hypoxia, Cancer research 65 (2) (2005) 613–621.

[8] B. Jiang, Aerobic glycolysis and high level of lactate in cancer metabolism and microenvironment, Genes & Diseases 4 (1) (2017) 25–27.

[9] P. Danhier, P. Bański, V. L. Payen, D. Grasso, L. Ippolito, P. Sonveaux, P. E. Porporato, Cancer metabolism in space and time: beyond the warburg effect, Biochimica et Biophysica Acta (BBA)-Bioenergetics 1858 (8) (2017) 556–572.

[10] D. Hanahan, R. A. Weinberg, The hallmarks of cancer, cell 100 (1) (2000) 57–70.

[11] D. Hanahan, R. A. Weinberg, Hallmarks of cancer: the next generation, cell 144 (5) (2011) 646–674.

[12] Q. Liu, H. Zhang, X. Jiang, C. Qian, Z. Liu, D. Luo, Factors involved in cancer metastasis: a better understanding to "seed and soil" hypothesis, Molecular cancer 16 (1) (2017) 176.

[13] J. Sceneay, M. T. Chow, A. Chen, H. M. Halse, C. S. Wong, D. M. Andrews, E. K. Sloan, B. S. Parker, D. D. Bowtell, M. J. Smyth, et al., Primary tumor hypoxia recruits cd11b+/ly6cmed/ly6g+ immune suppressor cells and compromises nk cell cytotoxicity in the premetastatic niche, Cancer research 72 (16) (2012) 3906–3911.

[14] H. Peinado, H. Zhang, I. R. Matei, B. Costa-Silva, A. Hoshino, G. Rodrigues, B. Psaila, R. N. Kaplan, J. F. Bromberg, Y. Kang, et al., Pre-metastatic niches: organ-specific homes for metastases, Nature Reviews Cancer 17 (5) (2017) 302.

[15] M. Potter, E. Newport, K. J. Morten, The warburg effect: 80 years on, Biochemical Society Transactions 44 (5) (2016) 1499–1505.

[16] M. Sakashita, N. Aoyama, R. Minami, S. Maekawa, K. Kuroda, D. Shirasaka, T. Ichihara, Y. Kuroda, S. Maeda, , M. Kasuga, Glut1 expression in t1 and t2 stage colorectal carcinomas: its relationship to clinicopathological features, European journal of cancer 37 (2) (2001) 204–209.

[17] X. Zhong, X. He, Y. Wang, Z. Hu, H. Huang, S. Zhao, P. Wei, D. Li, Warburg effect in colorectal cancer: the emerging roles in tumor microenvironment and therapeutic implications, Journal of Hematology & Oncology 15 (1) (2022) 160.

[18] K. Offermans, J. C. Jenniskens, C. C. Simons, I. Samarska, G. E. Fazzi, K. M. Smits, L. J. Schouten, M. P. Weijenberg, H. I. Grabsch, P. A. van den Brandt, Expression of proteins associated with the warburg-effect and survival in colorectal cancer, The Journal of Pathology: Clinical Research 8 (2) (2022) 169–180.

[19] J.-H. Lai, H.-J. Jan, L.-W. Liu, C.-C. Lee, S.-G. Wang, D.-Y. Hueng, Y.-Y. Cheng, H.-M. Lee, H.-I. Ma, Nodal regulates energy metabolism in glioma cells by inducing expression of hypoxia-inducible factor 1α, Neuro-oncology 15 (10) (2013) 1330–1341.

[20] E. Michelakis, G. Sutendra, P. Dromparis, L. Webster, A. Haromy, E. Niven, C. Maguire, T.-L. Gammer, J. Mackey, D. Fulton, et al., Metabolic modulation of glioblastoma with dichloroacetate, Science translational medicine 2 (31) (2010) 31ra34–31ra34.

[21] J. E. Burns, C. D. Hurst, M. A. Knowles, R. M. Phillips, S. J. Allison, The warburg effect as a therapeutic target for bladder cancers and intratumoral heterogeneity in associated molecular targets, Cancer Science 112 (9) (2021) 3822–3834.

[22] B. Shuch, W. M. Linehan, R. Srinivasan, Aerobic glycolysis: a novel target in kidney cancer, Expert review of anticancer therapy 13 (6) (2013) 711–719.

[23] L. Yihan, W. Xiaojing, L. Ao, Z. Chuanjie, W. Haofei, S. Yan, H. Hongchao, Sirt5 functions as a tumor suppressor in renal cell carcinoma by reversing the warburg effect, Journal of Translational Medicine 19 (1) (2021) 1–12.

[24] M. Grover-McKay, S. A. Walsh, E. A. Seftor, P. A. Thomas, M. J. Hendrix, Role for glucose transporter 1 protein in human breast cancer, Pathology & Oncology Research 4 (1998) 115–120.

[25] A. Kalezic, M. Udicki, B. Srdic Galic, M. Aleksic, A. Korac, A. Jankovic, B. Korac, Tissue-specific warburg effect in breast cancer and cancer-associated adipose tissue—relationship between ampk and glycolysis, Cancers 13 (11) (2021) 2731.

[26] P. R. Kumar, J. A. Moore, K. M. Bowles, S. A. Rushworth, M. D. Moncrieff, Mitochondrial oxidative phosphorylation in cutaneous melanoma, British Journal of Cancer 124 (1) (2021) 115–123.

[27] Y. Kamenisch, T. S. Baban, W. Schuller, A.-K. von Thaler, T. Sinnberg, G. Metzler, J. Bauer, B. Schittek, C. Garbe, M. Rocken, et al., Uva-irradiation induces melanoma invasion via the enhanced warburg effect, Journal of Investigative Dermatology 136 (9) (2016) 1866–1875.

[28] F. Wang, H. Liu, L. Hu, Y. Liu, Y. Duan, R. Cui, W. Tian, The warburg effect in human pancreatic cancer cells triggers cachexia in athymic mice carrying the cancer cells, BMC cancer 18 (2018) 1–12.

[29] N. Rajeshkumar, P. Dutta, S. Yabuuchi, R. F. De Wilde, G. V. Martinez, A. Le, J. J. Kamphorst, J. D. Rabinowitz, S. K. Jain, M. Hidalgo, et al., Therapeutic targeting of the warburg effect in pancreatic cancer relies on an absence of p53 functionldh-a inhibition in pancreatic cancer, Cancer research 75 (16) (2015) 3355–3364.

[30] R. Li, H. Li, L. Zhu, X. Zhang, D. Liu, Q. Li, B. Ni, L. Hu, Z. Zhang, Y. Zhang, et al., Reciprocal regulation of loxl2 and hif1α drives the warburg effect to support pancreatic cancer aggressiveness, Cell Death & Disease 12 (12) (2021) 1106.

[31] M. Wu, A. Neilson, A. L. Swift, R. Moran, J. Tamagnine, D. Parslow, S. Armistead, K. Lemire, J. Orrell, J. Teich, et al., Multiparameter metabolic analysis reveals a close link between attenuated mitochondrial bioenergetic function and enhanced glycolysis dependency in human tumor cells, American Journal of Physiology-Cell Physiology 292 (1) (2007) C125–C136.

[32] L. Sha, Z. Lv, Y. Liu, Y. Zhang, X. Sui, T. Wang, H. Zhang, Shikonin inhibits the warburg effect, cell proliferation, invasion and migration by downregulating pfkfb2 expression in lung cancer, Molecular Medicine Reports 24 (2) (2021) 1–10.

[33] S. A. Dyshlovoy, D. N. Pelageev, J. Hauschild, K. L. Borisova, M. Kaune, C. Krisp, S. Venz, Y. E. Sabutskii, E. A. Khmelevskaya, T. Busenbender, et al., Successful targeting of the warburg effect in prostate cancer by glucose-conjugated 1, 4-naphthoquinones, Cancers 11 (11) (2019) 1690.

[34] S.-S. Wen, T.-T. Zhang, D.-X. Xue, W.-L. Wu, Y.-L. Wang, Y. Wang, Q.-H. Ji, Y.-X. Zhu, N. Qu, R.-L. Shi, Metabolic reprogramming and its clinical application in thyroid cancer, Oncology Letters 18 (2) (2019) 1579–1584.

[35] Z. Yang, R. Huang, X. Wei, W. Yu, Z. Min, M. Ye, The sirt6-autophagy-warburg effect axis in papillary thyroid cancer, Frontiers in Oncology 10 (2020) 1265.

[36] J. Feng, J. Li, L. Wu, Q. Yu, J. Ji, J. Wu, W. Dai, C. Guo, Emerging roles and the regulation of aerobic glycolysis in hepatocellular carcinoma, Journal of Experimental & Clinical Cancer Research 39 (1) (2020) 1–19.

[37] L. Ding, X. Liang, Ras related gtp binding d promotes aerobic glycolysis of hepatocellular carcinoma, Annals of Hepatology 23 (2021) 100307.

[38] X. Zhang, J. Guo, P. Jabbarzadeh Kaboli, Q. Zhao, S. Xiang, J. Shen, Y. Zhao, F. Du, X. Wu, M. Li, et al., Analysis of key genes regulating the warburg effect in patients with gastrointestinal cancers and selective inhibition of this metabolic pathway in liver cancer cells, OncoTargets and therapy (2020) 7295–7304.

[39] Z. Pu, M. Xu, X. Yuan, H. Xie, J. Zhao, Circular rna circcul3 accelerates the warburg effect progression of gastric cancer through regulating the stat3/hk2 axis, Molecular Therapy-Nucleic Acids 22 (2020) 310–318.

[40] Y.-L. Bin, H.-S. Hu, F. Tian, Z.-H. Wen, M.-F. Yang, B.-H. Wu, L.-S. Wang, J. Yao, D.-F. Li, Metabolic reprogramming in gastric cancer: Trojan horse effect, Frontiers in Oncology (2022) 5078.

[41] H. Sawayama, T. Ishimoto, H. Sugihara, N. Miyanari, Y. Miyamoto, Y. Baba, N. Yoshida, H. Baba, Clinical impact of the warburg effect in gastrointestinal cancer, International journal of oncology 45 (4) (2014) 1345–1354.

[42] J. Lu, M. Chen, S. Gao, J. Yuan, Z. Zhu, X. Zou, Ly294002 inhibits the warburg effect in gastric cancer cells by downregulating pyruvate kinase m2, Oncology letters 15 (4) (2018) 4358–4364.

[43] I. Tomlinson, W. Bodmer, Modelling the consequences of interactions between tumour cells, British journal of cancer 75 (2) (1997) 157.

[44] J. M. S. Nogales, S. Zazo, An evolutionary dynamics model for metastatic tumour growth based on public goods games, Communications in Nonlinear Science and Numerical Simulation 99 (2021) 105783.

[45] A. Kaznatcheev, R. Vander Velde, J. G. Scott, D. Basanta, Cancer treatment scheduling and dynamic heterogeneity in social dilemmas of tumour acidity and vasculature, British journal of cancer 116 (6) (2017) 785.

[46] M. Archetti, Heterogeneity and proliferation of invasive cancer subclones in game theory models of the warburg effect, Cell proliferation 48 (2) (2015) 259–269.

[47] M. Archetti, Evolutionary dynamics of the warburg effect: glycolysis as a collective action problem among cancer cells, Journal of theoretical biology 341 (2014) 1–8.

[48] M. Archetti, K. J. Pienta, Cooperation among cancer cells: applying game theory to cancer, Nature Reviews Cancer (2018) 1.

[49] P. Das, P. Das, S. Mukherjee, Stochastic dynamics of michaelis–menten kinetics based tumor-immune interactions, Physica A: Statistical Mechanics and its Applications 541 (2020) 123603.

[50] D. A. Drexler, T. Ferenci, A. Lovrics, L. Kovács, Comparison of michaelis-menten kinetics modeling alternatives in cancer chemotherapy modeling, in: 2019 IEEE 13th International Symposium on Applied Computational Intelligence and Informatics (SACI), IEEE, 2019, pp. 27–32.

[51] D. Lestari, E. Sari, H. Arifah, Dynamics of a mathematical model of cancer cells with chemotherapy, in: Journal of Physics: Conference Series, Vol. 1320, IOP Publishing, 2019, p. 012026.

[52] M. Shamsi, M. Saghafian, M. Dejam, A. Sanati-Nezhad, Mathematical modeling of the function of warburg effect in tumor microenvironment, Scientific reports 8 (1) (2018) 8903.

[53] R. H. Chisholm, T. Lorenzi, J. Clairambault, Cell population heterogeneity and evolution towards drug resistance in cancer: biological and mathematical assessment, theoretical treatment optimisation, Biochimica et Biophysica Acta (BBA)-General Subjects 1860 (11) (2016) 2627–2645.

[54] N. Farrokhian, J. Maltas, P. Ellsworth, A. Durmaz, M. Dinh, M. Hitomi, A. Kaznatcheev, A. Marusyk, J. Scott, Dose dependent evolutionary game dynamics modulate competitive release in cancer therapy, bioRxiv 2020 (18.303966) (2020).

[55] K. Staňková, Resistance games, Nature ecology & evolution 3 (3) (2019) 336–337.

[56] H. Zhang, J. Lei, Optimal treatment strategy of cancers with intratumor heterogeneity, Mathematical Biosciences and Engineering 19 (12) (2022) 13337–13373.

[57] P. Das, S. Das, P. Das, F. A. Rihan, M. Uzuntarla, D. Ghosh, Optimal control strategy for cancer remission using combinatorial therapy: a mathematical model-based approach, Chaos, Solitons & Fractals 145 (2021) 110789.

[58] P. Das, S. Das, R. K. Upadhyay, P. Das, Optimal treatment strategies for delayed cancer-immune system with multiple therapeutic approach, Chaos, Solitons & Fractals 136 (2020) 109806.

[59] J. Cunningham, F. Thuijsman, R. Peeters, Y. Viossat, J. Brown, R. Gatenby, K. Staňková, Optimal control to reach eco-evolutionary stability in metastatic castrate-resistant prostate cancer, Plos one 15 (12) (2020) e0243386.

[60] M. Gluzman, J. G. Scott, A. Vladimirsky, Optimizing adaptive cancer therapy: dynamic programming and evolutionary game theory, Proceedings of the Royal Society B 287 (1925) (2020) 20192454.

[61] D. Engelhardt, Dynamic control of stochastic evolution: A deep reinforcement learning approach to adaptively targeting emergent drug resistance, arXiv preprint arXiv:1903.11373 (2019).

[62] H.-H. Tseng, Y. Luo, S. Cui, J.-T. Chien, R. K. Ten Haken, I. El Naqa, Deep reinforcement learning for automated radiation adaptation in lung cancer, Medical physics 44 (12) (2017) 6690–6705.

[63] R. Padmanabhan, N. Meskin, W. M. Haddad, Reinforcement learning-based control of drug dosing for cancer chemotherapy treatment, Mathematical biosciences 293 (2017) 11–20.

[64] A. Cornish-Bowden, A. Cornish-Bowden, Fundamentals of enzyme kinetics, Vol. 510, Wiley-Blackwell Weinheim, Germany, 2012.

[65] P. Day, J. Cleal, E. Lofthouse, M. Hanson, R. Lewis, What factors determine placental glucose transfer kinetics?, Placenta 34 (10) (2013) 953–958.

[66] S. Bröer, B. Rahman, G. Pellegri, L. Pellerin, J.-L. Martin, S. Verleysdonk, B. Hamprecht, P. J. Magistretti, Comparison of lactate transport in astroglial cells and monocarboxylate transporter 1 (mct 1) expressing xenopus laevis oocytes expression of two different monocarboxylate transporters in astroglial cells and neurons, Journal of Biological Chemistry 272 (48) (1997) 30096–30102.

[67] H.-U. Son, E.-K. Yoon, C.-Y. Yoo, C.-H. Park, M. Bae, T.-H. Kim, C. H. Lee, K. W. Lee, H. Seo, K.-J. Kim, et al., Effects of synergistic inhibition on α-glucosidase by phytoalexins in soybeans, Biomolecules 9 (12) (2019) 828.

[68] N. S. Punekar, Enzymes: catalysis, kinetics and mechanisms, Springer, 2018.

[69] J. W. Pelley, Elsevier's Integrated Review Biochemistry E-Book: with STUDENT CONSULT Online Access, Elsevier Health Sciences, 2011.

[70] K. B. Storey, Functional metabolism: regulation and adaptation, John Wiley & Sons, 2005.

[71] R. A. Copeland, Enzymes: a practical introduction to structure, mechanism, and data analysis, John Wiley & Sons, 2004.

[72] V. Leskovac, Comprehensive enzyme kinetics, Springer Science & Business Media, 2003.

[73] M. Shan, D. Dai, A. Vudem, J. D. Varner, A. D. Stroock, Multi-scale computational study of the warburg effect, reverse warburg effect and glutamine addiction in solid tumors, PLoS computational biology 14 (12) (2018) e1006584.

[74] R. Cressman, Y. Tao, The replicator equation and other game dynamics, Proceedings of the National Academy of Sciences 111 (Supplement 3) (2014) 10810–10817.

[75] W. H. Sandholm, Population games and evolutionary dynamics, MIT press, 2010.

[76] J. W. Weibull, Evolutionary game theory, MIT press, 1997.

[77] R. S. Sutton, A. G. Barto, Reinforcement learning: An introduction, MIT press, 2018.

[78] D. Bertsekas, Dynamic programming and optimal control: Volume I, Vol. 1, Athena scientific, 2012.

[79] H. Van Hasselt, A. Guez, D. Silver, Deep reinforcement learning with double q-learning, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 30, 2016.

[80] M. Gluzman, J. G. Scott, A. Vladimirsky, Optimizing adaptive cancer therapy: dynamic programming and evolutionary game theory, arXiv preprint arXiv:1812.01805 (2018).

[81] BOCOP – the optimal control solver, `https://www.bocop.org/bocophjb-1-1-0/`, accessed: 2022-04-11 (2017).

[82] C. Harmon, M. W. Robinson, F. Hand, D. Almuaili, K. Mentor, D. D. Houlihan, E. Hoti, L. Lynch, J. Geoghegan, C. O'Farrelly, Lactate-mediated acidification of tumor microenvironment induces apoptosis of liver-resident nk cells in colorectal liver metastasis, Cancer immunology research 7 (2) (2019) 335–346.

[83] G. M. GabAllah, M. S. E.-d. Habib, S. E.-S. Soliman, Z. A. Kasemy, S. F. Gohar, Validity and clinical impact of glucose transporter 1 expression in colorectal cancer, Saudi journal of gastroenterology: official journal of the Saudi Gastroenterology Association 23 (6) (2017) 348.

[84] A. Zambrano, M. Molt, E. Uribe, M. Salas, Glut 1 in cancer cells and the inhibitory action of resveratrol as a potential therapeutic strategy, International journal of molecular sciences 20 (13) (2019) 3374.

[85] S. F. Martins, R. Amorim, M. Viana-Pereira, C. Pinheiro, R. F. A. Costa, P. Silva, C. Couto, S. Alves, S. Fernandes, S. Vilaça, et al., Significance of glycolytic metabolism-related protein expression in colorectal cancer, lymph node and hepatic metastasis, BMC cancer 16 (1) (2016) 535.

[86] D. Mishra, D. Banerjee, Lactate dehydrogenases as metabolic links between tumor and stroma in the tumor microenvironment, Cancers 11 (6) (2019) 750.

[87] S. J. Park, C. P. Smith, R. R. Wilbur, C. P. Cain, S. R. Kallu, S. Valasapalli, A. Sahoo, M. R. Guda, A. J. Tsung, K. K. Velpula, An overview of mct1 and mct4 in gbm: small molecule transporters with large implications, American journal of cancer research 8 (10) (2018) 1967.

[88] D. Benjamin, D. Robay, S. K. Hindupur, J. Pohlmann, M. Colombi, M. Y. El-Shemerly, S.-M. Maira, C. Moroni, H. A. Lane, M. N. Hall, Dual inhibition of the lactate transporters mct1 and mct4 is synthetic lethal with metformin due to nad+ depletion in cancer cells, Cell reports 25 (11) (2018) 3047–3058.

[89] J. Pérez-Escuredo, V. F. Van Hee, M. Sboarina, J. Falces, V. L. Payen, L. Pellerin, P. Sonveaux, Monocarboxylate transporters in the brain and in cancer, Biochimica et Biophysica Acta (BBA)-Molecular Cell Research 1863 (10) (2016) 2481–2497.

[90] S. Sun, H. Li, J. Chen, Q. Qian, Lactic acid: no longer an inert and end-product of glycolysis, Physiology 32 (6) (2017) 453–463.

[91] D. Ganten, K. Ruckpaul, Encyclopedic reference of genomics and proteomics in molecular medicine, Springer, 2006.

[92] C. Granchi, S. Fortunato, F. Minutolo, Anticancer agents interacting with membrane glucose transporters, MedChemComm 7 (9) (2016) 1716–1729.

[93] M. J. Ovens, A. J. Davies, M. C. Wilson, C. M. Murray, A. P. Halestrap, Ar-c155858 is a potent inhibitor of monocarboxylate transporters mct1 and mct2 that binds to an intracellular site involving transmembrane helices 7–10, Biochemical Journal 425 (3) (2010) 523–530.

[94] S. Ramos, Effects of dietary flavonoids on apoptotic pathways related to cancer chemoprevention, The Journal of nutritional biochemistry 18 (7) (2007) 427–442.

[95] X. Guan, M. A. Bryniarski, M. E. Morris, In vitro and in vivo efficacy of the monocarboxylate transporter 1 inhibitor ar-c155858 in the murine 4t1 breast cancer tumor model, The AAPS journal 21 (1) (2019) 1–10.

[96] S. Walenta, W. F. Mueller-Klieser, Lactate: mirror and motor of tumor malignancy, in: Seminars in radiation oncology, Vol. 14, Elsevier, 2004, pp. 267–274.

[97] M. Hausknecht, P. Stone, Deep recurrent q-learning for partially observable mdps, in: 2015 aaai fall symposium series, 2015.

- A model based on evolutionary game theory is proposed to represent the dynamics of cell populations subject to the Warburg effect. This model reproduces by computer simulation some clinical results observed in colorectal liver metastasis and other even more aggressive cancers.
- In silico results with optimal targeted therapies using Double Deep Q-networks is proposed. These therapies seek to attack cells that express specific cancerous phenotypes, with the combination of tumor growth reversible inhibitors in different doses. These therapies also consider the duration of treatment, drug toxicity, contraindications and harmful side effects in order to guarantee the patients' quality of life.
- Optimal therapies obtained with Double Deep Q-networks are validated with the solutions of the Hamilton-Jacobi-Bellman equation.

All authors declare that they have no conflicts of interest.