An Optimal Bayesian Intervention Policy in Response to Unknown Dynamic Cell Stimuli

Seyed Hamid Hosseini, Mahdi Imani[a]

[a]*Northeastern University, 360 Huntington Ave, Boston, MA, 02115, U.S.*

Abstract

Interventions in gene regulatory networks (GRNs) aim to restore normal functions of cells experiencing abnormal behavior, such as uncontrolled cell proliferation. The dynamic, uncertain, and complex nature of cellular processes poses significant challenges in determining the best interventions. Most existing intervention methods assume that cells are unresponsive to therapies, resulting in stationary and deterministic intervention solutions. However, cells in unhealthy conditions can dynamically respond to therapies through internal stimuli, leading to the recurrence of undesirable conditions. This paper proposes a Bayesian intervention policy that adaptively responds to cell dynamic responses according to the latest available information. The GRNs are modeled using a Boolean network with perturbation (BNp), and the fight between the cell and intervention is modeled as a two-player zero-sum game. Assuming an incomplete knowledge of cell stimuli, a recursive approach is developed to keep track of the posterior distribution of cell responses. The proposed Bayesian intervention policy takes action according to the posterior distribution and a set of Nash equilibrium policies associated with all possible cell responses. Analytical results demonstrate the superiority of the proposed intervention policy against several existing intervention techniques. Meanwhile, the performance of the proposed policy is investigated through comprehensive numerical experiments using the p53-MDM2 negative feedback loop regulatory network and melanoma network. The results demonstrate the empirical convergence of the proposed policy to the optimal Nash equilibrium policy.

Keywords: Gene Regulatory Networks, Two-Player Zero-Sum Game, Bayesian intervention, Boolean networks, Nash Equilibrium.

1. Introduction

Recent genomics advances have deepened our understanding of complex biological systems, particularly gene regulatory networks (GRNs) [1, 2, 3, 4, 5, 6]. GRNs consist of several interacting genes whose activities control cellular processes, including DNA repair, stress response, and complex diseases like cancer [7]. In genomics intervention, the objective is to design effective intervention strategies that can alter the undesirable behavior of unhealthy cells (e.g., those associated with chronic diseases) and shift them into desirable ones.

Boolean networks have emerged as a powerful class of models for characterizing the temporal dynamics of GRNs [8, 9, 10, 11, 12, 13]. Several intervention strategies have been developed for Boolean network models in recent years. These include structural interventions, which aim to make a single-time, long-lasting change in the interaction between two or more genes [14, 15, 16, 17, 18], and dynamic interventions that perturb (e.g., overexpress or suppress) the activity of targeted genes over time [14, 15, 16, 17]. The most well-known method is the optimal stationary intervention derived in [19], which is later extended to include constraints [20, 21] and asynchronicity of the GRNs [22, 13]. Meanwhile, several intervention approaches are developed for GRNs with states observed indirectly through gene-expression data [23, 24, 25, 26, 27, 28], including robust intervention methods for domains with partially-known dynamics and costs [29, 30].

Most existing intervention methods are built on the assumption that cells are isolated and non-responsive to therapies. However, the dynamic and intelligent responses of cells to therapies, triggered by internal stimuli, often result in the short-term success of interventions at early stages and the recurrence of the unhealthy condition afterward. This paper models GRNs using Boolean networks with perturbation (BNp) [31, 32], and models the cell dynamic responses to interventions through a two-player zero-sum game [33, 34, 35]. There are two players in the game: the cell and the intervention, each with opposing goals. The cell aims to maintain the cell condition in unhealthy states using its internal stimuli, while the intervention's objective is to deviate the system from unhealthy conditions through therapies. Assuming incomplete information about the possible cell responses to interventions, this paper develops a recursive method for computing the posterior distribution of the cell responses. Given the quantified uncertainty in cell responses, we develop a Bayesian intervention policy. The proposed policy utilizes the combination of the Nash equilibrium policies for different cell responses and the posterior associated with them. The policy is fully adaptive; as new data appears, the posterior distribution of cell responses and the proposed intervention policy are updated.

The main contributions of this paper are as follows:

- Modeling the aggressive and dynamic responses of unhealthy cells during the intervention process, which enables deriving intervention solutions by accounting for and predicting possible cell responses to therapies.
- Develop an adaptive Bayesian intervention policy that can probabilistically reason about cell responses and incorporate such knowledge to make better intervention decisions.
- Analytically demonstrating the superiority of the proposed policy compared to existing intervention methods, along with numerical results indicating the empirical convergence of the proposed policy to the optimal Nash policy.

We analyze the performance of the proposed intervention policy using the p53-MDM2 and melanoma networks. The p53-MDM2 network is a crucial regulatory system that responds to cellular stresses such as DNA damage [36, 37]. The melanoma regulatory network also plays a crucial role in the development and progression of melanoma, a highly aggressive form of skin cancer [21, 38]. Through a comprehensive set of numerical experiments using these two networks, we compare the performance of the proposed policy with state-of-the-art intervention methods.

The article is organized as follows: The GRN model is briefly described in Section 2. Section 3 includes formulating the intervention process as a two-player zero-sum game, followed by the optimal Nash equilibrium policy for a two-player zero-sum game. The proposed Bayesian intervention policy and its matrix-form implementation are presented in Sections 4 and 5, respectively. The analytical and numerical results are presented in Section 6 and Section 7, respectively. Finally, Section 8 contains the concluding remarks.

2. Background

In this paper, a Boolean network with perturbation model [32, 39] is used to capture the dynamics of gene regulatory networks. The BNp model effectively incorporates the stochastic nature of GRNs and accounts for the uncertainty coming from unmodeled parts of the systems. Consider a GRN consisting of d components. The *state process* can be represented as $\{\mathbf{x}_k; k = 0, 1, \ldots\}$, where $\mathbf{x}_k \in \{0, 1\}^d$ denotes the activation or inactivation state of the genes at time k. The genes' state is influenced by a series of internal and external inputs/stimuli. At each discrete time point, the state of the genes evolves according to the following Boolean signal model [40]:

$$\mathbf{x}_k = \mathbf{f}(\mathbf{x}_{k-1}) \oplus \mathbf{a}_{k-1} \oplus \mathbf{u}_{k-1} \oplus \mathbf{n}_k, \quad k = 1, 2, \ldots \ , \tag{1}$$

where $\{\mathbf{a}_k; k = 0, 1, ...\}$ refers to a set of external interventions/therapies, $\{\mathbf{u}_k; k = 0, 1, ...\}$ represents internal inputs regulated by the cell, $\mathbf{n}_k \in \{0, 1\}^d$ represents Boolean transition noise at time k, "$\oplus$" denotes component-wise module-2 addition, and $\mathbf{f}$ is the *network function*. The noise value $\mathbf{n}_k(j) = 1$ alters the state of the jth gene at time step k; whereas for $\mathbf{n}_k(j) = 0$, the jth state follows the value predicted by the network function. The noise process $\mathbf{n}_k$ is assumed to have independent components modeled by a Bernoulli distribution with parameter $p > 0$. The Bernoulli parameter p represents the noise intensity, with higher values representing more chaotic systems and smaller values indicating nearly deterministic models. Note that the rest of the paper is applicable to a general class of Boolean network models of the form $\mathbf{f}(\mathbf{x}_{k-1}, \mathbf{a}_{k-1}, \mathbf{u}_{k-1}, \mathbf{n}_k)$.

The network function in GRNs is often represented through a Boolean logic model or a pathway diagram model [41, 40]. The Boolean logic model captures the genes' activities and interactions using logical operators such as AND, OR, XOR, and NOT, while the pathway diagram model parameterizes suppressive and activating interactions among genes to capture their dynamics. These models have shown success in capturing the temporal changes in gene activities and causal interactions among genes.

3. Battle of Cell and Intervention

3.1. Two-Player Zero-Sum Game

We represent the battle between the cell and intervention as a two-player zero-sum game [42, 33, 34, 35]. This can be characterized by a tuple $\langle \mathcal{X}, \mathcal{A}, \mathcal{U}, R^a, T \rangle$, where $\mathcal{X} = \{0, 1\}^d$ is the *state space*, $\mathcal{A}$ is the *intervention space*, $\mathcal{U}$ is the *cell control space*, R^a is the *intervention reward function*, and T is the *state transition probability function*. $T : \mathcal{X} \times \mathcal{A} \times \mathcal{U} \times \mathcal{X}$ is such that $p(\mathbf{x}' \mid \mathbf{x}, \mathbf{a}, \mathbf{u})$ represents the probability of moving to state $\mathbf{x}'$ according to the external and internal inputs $\mathbf{a}$ and $\mathbf{u}$ in state $\mathbf{x}$. Also, $R^a(\mathbf{x}, \mathbf{a}, \mathbf{u}, \mathbf{x}')$ denotes the immediate intervention reward gained if the system moves from state $\mathbf{x}$ to state $\mathbf{x}'$ according to the joint intervention and cell actions $(\mathbf{a}, \mathbf{u})$.

3.2. Optimal Nash Intervention Policy under Known Cell Responses

The diagram representing the fight between cell and intervention is shown in Fig. 1. For cells in cancerous conditions, the intervention objective is to decrease cell proliferation, whereas cells aim to increase such proliferation by fighting against interventions. The opposite objectives of the intervention and cell can be expressed by the cell reward R^u taking the negative of the intervention reward, i.e., $R^u(\mathbf{x}, \mathbf{a}, \mathbf{u}, \mathbf{x}') = -R^a(\mathbf{x}, \mathbf{a}, \mathbf{u}, \mathbf{x}')$.

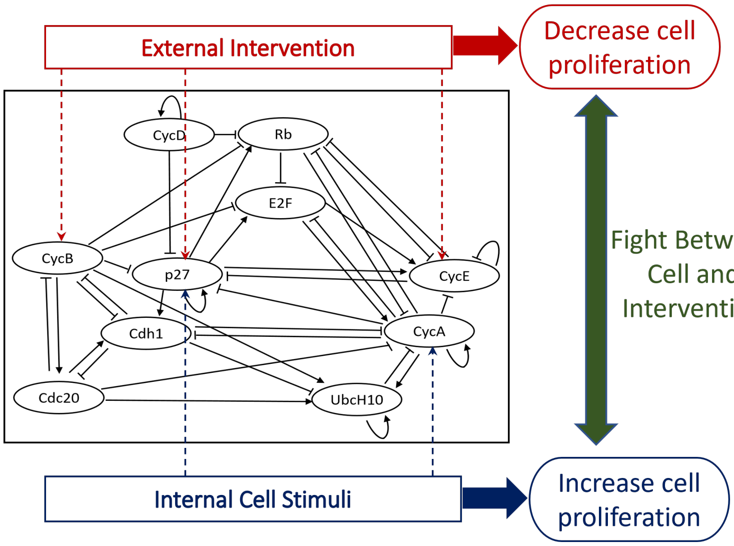

External Intervention
Decrease cell proliferation
CycD
Rb
E2F
CycB
p27
CycE
Cdh1
CycA
Cdc20
UbcH10
Fight Between Cell and Intervention
Internal Cell Stimuli
Increase cell proliferation

Figure 1: The fight between intervention and the cell dynamic response according to its internal stimuli.

This paper focuses on stationary Markov Nash equilibria in GRNs modeled by the infinite-horizon discounted Markov game. Let $\mathcal{U}$ contain a finite set of stimuli/actions that the cell could perform during the intervention process against therapies. Let also $\mathcal{A}$ be the set of actions/therapies available during the intervention process. We define the intervention policy $\pi^a(\mathbf{a} \mid \mathbf{x})$, representing the probability of taking action $\mathbf{a} \in \mathcal{A}$ in any given state $\mathbf{x} \in \mathcal{X}$. Similarly, the cell policy $\pi^u(\mathbf{u} \mid \mathbf{x})$ specifies the probability of selecting input $\mathbf{u} \in \mathcal{U}$ in state $\mathbf{x} \in \mathcal{X}$. For the joint stochastic policy (π^a, π^u), the expected value function of intervention and cell can be defined as:

$$V^a_{\pi^a,\pi^u}(\mathbf{x}) = \mathbb{E}\Big[\sum_{t\geq 0} \gamma^t R^a(\mathbf{x}_t, \mathbf{a}_t, \mathbf{u}_t, \mathbf{x}_{t+1}) \mid \mathbf{a}_{0:\infty} \sim \pi^a, \mathbf{u}_{0:\infty} \sim \pi^u, \mathbf{x}_0 = \mathbf{x}\Big],$$

$$V^u_{\pi^a,\pi^u}(\mathbf{x}) = \mathbb{E}\Big[\sum_{t\geq 0} \gamma^t R^u(\mathbf{x}_t, \mathbf{a}_t, \mathbf{u}_t, \mathbf{x}_{t+1}) \mid \mathbf{a}_{0:\infty} \sim \pi^a, \mathbf{u}_{0:\infty} \sim \pi^u, \mathbf{x}_0 = \mathbf{x}\Big], \tag{2}$$

for $\mathbf{x} \in \mathcal{X}$; where $0 < \gamma < 1$ is a discount factor that prioritizes the early-stage rewards compared to future ones. Given that cell and intervention reward functions are negative of each other, we have $V^a_{\pi^a,\pi^u}(\mathbf{x}) = -V^u_{\pi^a,\pi^u}(\mathbf{x})$, for any $\mathbf{x} \in \mathcal{X}$. Due to the interplay between state values for the cell and intervention, this problem differs from a Markov decision process (MDP). The optimal solution for a two-player zero-sum game can be expressed through the Markov game. This is expressed as the optimal Nash equilibrium policy $\pi^* = (\pi^a_*, \pi^u_*)$, which for any joint policy $\pi = (\pi^a, \pi^u)$ and $\mathbf{x} \in \mathcal{X}$ satisfies [33]:

$$V^a_{\pi^a_*,\pi^u_*}(\mathbf{x}) \geq V^a_{\pi^a,\pi^u_*}(\mathbf{x}) \text{ and } V^u_{\pi^a_*,\pi^u_*}(\mathbf{x}) \geq V^u_{\pi^a_*,\pi^u}(\mathbf{x}). \tag{3}$$

The optimal Nash equilibrium policy is the policy that the cell and intervention have no motivation to deviate from it. This policy can be expressed according to the min-max theorem as [43]:

$$(\pi^a_*, \pi^u_*) = \underset{\pi^a}{\operatorname{argmax}}\, \underset{\pi^u}{\operatorname{argmin}}\, V^a_{\pi^a,\pi^u}(\mathbf{x}) = \underset{\pi^u}{\operatorname{argmin}}\, \underset{\pi^a}{\operatorname{argmax}}\, V^a_{\pi^a,\pi^u}(\mathbf{x}), \text{ for all } \mathbf{x} \in \mathcal{X}. \tag{4}$$

Based on equation (2), any pair of (π^a, π^u) that achieves the supremum and infimum values in equation (4) forms an optimal Nash equilibrium.

4. Bayesian Intervention Policy under Unknown Cell Responses

4.1. Intervention Challenges of Unknown Cell Space

If the cell space $\mathcal{U}$, representing the internal cell stimuli, is fully known, then the optimal Nash policy could be achieved as a solution for the optimization in (4). However, in practice, the cell's internal stimuli are often unknown, preventing the computation of the optimal Nash policy. Therefore, this paper aims to derive an effective intervention policy that can be implemented despite incomplete knowledge about cell space. We present a systematic approach to probabilistically reason about the possible cell responses using the latest available data and use this knowledge for effective intervention selection.

Let $\mathcal{U}^1, ..., \mathcal{U}^M$ be the set of all possible cell spaces. This set depends on the size of the regulatory networks and the prior biological knowledge regarding the cell responses. Given a regulatory network consisting of d genes, there are 2^d possible cell actions. In this case, there are $\binom{2^d}{1}$ cell spaces containing 1 cell actions, $\binom{2^d}{2}$ sets with 2 cell actions, and $\binom{2^d}{m}$ sets containing m cell actions. This set can be large in large regulatory networks, but as described in the following paragraph, the posterior of many models approach zero as more data are observed.

If $\mathcal{U}^i$ is the true cell space, the optimal space-specific Nash policy can be expressed as $(\pi^{a,\mathcal{U}^i}_*, \pi^{u,\mathcal{U}^i}_*)$, where this policy can be computed using the optimization problem in (4) corresponding to the cell space $\mathcal{U}^i$. The Nash policy obtained under cell space $\mathcal{U}^i$ might significantly differ from $\mathcal{U}^j \neq \mathcal{U}^i$. Thus, given the limited or no knowledge about the true cell space, the space-specific intervention policies are not directly implementable. In fact, executing a wrong (non-optimal) intervention policy corresponding to $\mathcal{U}^j \neq \mathcal{U}^*$ could lead to poor intervention performance and the dominance of the cell.

4.2. Probability Model over Cell Spaces

This paper constructs a probabilistic model over the cell spaces. Let $p_0(i)$ be the prior probability of the ith cell space $\mathcal{U}^i$. The prior information about the set of cell spaces can be represented in a single vector as:

$$p_0 = [P(\mathcal{U}^1), ..., P(\mathcal{U}^M)]^T. \tag{5}$$

If no prior biological knowledge about cell space is available, a uniform prior can be considered over the cell spaces, i.e., $p_0 = [1/M, ..., 1/M]$.

Let $p_{k-1} = [p_{k-1}(1), ..., p_{k-1}(M)]$ be the posterior probability over the cell spaces obtained according to the sequence of observed states $\mathbf{x}_{0:k-1}$ obtained upon taking interventions $\mathbf{a}_{0:k-2}$. If intervention $\mathbf{a}_{k-1}$ is taken at time step $k-1$ and the state $\mathbf{x}_k$ is observed at time step k, the posterior probability of the cell spaces at time step k can be expressed as:

$$\begin{aligned} p_k(i) &= P(\mathcal{U}^* = \mathcal{U}^i | \mathbf{a}_{0:k-1}, \mathbf{x}_{0:k}) \\ &= \frac{p(\mathbf{x}_k, \mathcal{U}^i \mid \mathbf{a}_{0:k-1}, \mathbf{x}_{0:k-1})}{p(\mathbf{x}_k \mid \mathbf{a}_{0:k-1}, \mathbf{x}_{0:k-1})} \\ &= \frac{P(\mathbf{x}_k | \mathbf{a}_{0:k-1}, \mathbf{x}_{0:k-1}, \mathcal{U}^i) P(\mathcal{U}^* = \mathcal{U}^i | \mathbf{a}_{0:k-2}, \mathbf{x}_{0:k-1})}{\sum_{j=1}^{M} P(\mathbf{x}_k | \mathbf{a}_{0:k-1}, \mathbf{x}_{0:k-1}, \mathcal{U}^j) P(\mathcal{U}^* = \mathcal{U}^j | \mathbf{a}_{0:k-2}, \mathbf{x}_{0:k-1})} \\ &= \frac{p(\mathbf{x}_k \mid \mathbf{a}_{0:k-1}, \mathbf{x}_{0:k-1}, \mathcal{U}^i)\, p_{k-1}(i)}{\sum_{j=1}^{M} p(\mathbf{x}_k \mid \mathbf{a}_{0:k-1}, \mathbf{x}_{0:k-1}, \mathcal{U}^j)\, p_{k-1}(j)}, \end{aligned} \tag{6}$$

for $i = 1, ..., M$. The numerator term in (6) specifies the probability of observing the next state $\mathbf{x}_k$ given the sequence of interventions and states and the cell space $\mathcal{U}^i$. Further simplification of this term through marginalization of the joint distribution of the state $\mathbf{x}_k$ and the unobserved cell action $\mathbf{u}_{k-1}$ at time step k leads to:

$$\begin{aligned}
p(\mathbf{x}_k \mid \mathcal{U}^i, \mathbf{a}_{0:k-1}, \mathbf{x}_{0:k-1}) &= \sum_{\mathbf{u}\in\mathcal{U}^i} p(\mathbf{x}_k, \mathbf{u}_{k-1} = \mathbf{u} \mid \mathcal{U}^i, \mathbf{a}_{0:k-1}, \mathbf{x}_{0:k-1}) \\
&= \sum_{\mathbf{u}\in\mathcal{U}^i} p(\mathbf{x}_k \mid \mathbf{u}_{k-1} = \mathbf{u}, \mathbf{a}_{k-1}, \mathbf{x}_{k-1})\, p(\mathbf{u}_{k-1} = \mathbf{u} \mid \mathcal{U}^i, \mathbf{x}_{k-1}) \\
&= \sum_{\mathbf{u}\in\mathcal{U}^i} \left(\frac{p}{1-p}\right)^{||\mathbf{f}(\mathbf{x}_{k-1}) \oplus \mathbf{a}_{k-1} \oplus \mathbf{u} \oplus \mathbf{x}_k||_1} (1-p)^d\, \pi_*^{u,\mathcal{U}^i}(\mathbf{u} \mid \mathbf{x}_{k-1}) \,,
\end{aligned} \tag{7}$$

where $\pi_*^{u,\mathcal{U}^i}(\mathbf{u}_{k-1} = \mathbf{u}|\mathbf{x}_{k-1}) = p(\mathbf{u}_{k-1} = \mathbf{u} \mid \mathcal{U}^* = \mathcal{U}^i, \mathbf{x}_{k-1})$ is the probability that cell takes action $\mathbf{u}_{k-1} = \mathbf{u}$ at state $\mathbf{x}_{k-1}$ if the true cell action space is $\mathcal{U}^i$. The first line in the last expression in (7) is obtained using the Markovian properties of the state transition and the Bernoulli process noise. Replacing (7) into (6), the posterior probability of the cell space can be recursively computed using the last taken intervention and observed state.

4.3. Bayesian Intervention Policy

Let p_k be the posterior probability over the cell spaces obtained according to the states $\mathbf{x}_{0:k}$ and the sequence of intervention $\mathbf{a}_{0:k-1}$. The proposed Bayesian intervention policy at time step k can be expressed as:

$$\begin{aligned}
\mu_k^{a,\mathrm{B}}(\mathbf{a}|\mathbf{x}_k) :&= p(\mathbf{a}_k = \mathbf{a} \mid \mathbf{a}_{0:k-1}, \mathbf{x}_{0:k}) \\
&= \sum_{i=1}^{M} p(\mathbf{a}_k = \mathbf{a}, \mathcal{U}^* = \mathcal{U}^i \mid \mathbf{a}_{0:k-1}, \mathbf{x}_{0:k}) \\
&= \sum_{i=1}^{M} p(\mathbf{a}_k = \mathbf{a} \mid \mathcal{U}^i, \mathbf{a}_{0:k-1}, \mathbf{x}_{0:k})\, p(\mathcal{U}^* = \mathcal{U}^i \mid \mathbf{a}_{0:k-1}, \mathbf{x}_{0:k}) \\
&= \sum_{i=1}^{M} p(\mathbf{a}_k = \mathbf{a} \mid \mathcal{U}^i, \mathbf{a}_{0:k-1}, \mathbf{x}_{0:k})\, p_k(i) \\
&= \sum_{i=1}^{M} \pi_*^{a,\mathcal{U}^i}(\mathbf{a}|\mathbf{x}_k)\, p_k(i),
\end{aligned} \tag{8}$$

for $\mathbf{a} \in \mathcal{A}$; where the cell space is augmented and marginalized out in the second line. One can see that if the uncertainty over the cell spaces goes to zero, the Bayesian policy $\mu^{a,\mathrm{B}}(.|\mathbf{x}_k)$ becomes the optimal Nash equilibrium policy under the known cell space $\pi^{a,\mathcal{U}^*}(.|\mathbf{x}_k)$.

The Bayesian policy in (8) is stochastic and provides the best intervention solution given the available data. Let $\{\mathbf{u}^1, ..., \mathbf{u}^N\}$ be all unique cell actions in the set of cell spaces, i.e., $\{\mathbf{u}^1, ..., \mathbf{u}^N\} = \mathcal{U}^1 \cup ... \cup \mathcal{U}^M \subset \{0,1\}^d$. The Bayesian modeling of the cell defense policy at time step k can be expressed as:

$$\begin{aligned}
\mu_k^{u,\mathrm{B}}(\mathbf{u}|\mathbf{x}_k) &= p(\mathbf{u}_k = \mathbf{u} \mid \mathbf{a}_{0:k-1}, \mathbf{x}_{0:k}) \\
&= \sum_{i=1}^{M} p(\mathbf{u}_k = \mathbf{u}, \mathcal{U}^* = \mathcal{U}^i \mid \mathbf{a}_{0:k-1}, \mathbf{x}_{0:k}) \\
&= \sum_{i=1}^{M} p(\mathbf{u}_k = \mathbf{u} \mid \mathcal{U}^i, \mathbf{a}_{0:k-1}, \mathbf{x}_{0:k})\, p(\mathcal{U}^* = \mathcal{U}^i \mid \mathbf{a}_{0:k-1}, \mathbf{x}_{0:k}) \\
&= \sum_{i=1}^{M} p(\mathbf{u}_k = \mathbf{u} \mid \mathcal{U}^i, \mathbf{a}_{0:k-1}, \mathbf{x}_{0:k})\, p_k(i) \\
&= \sum_{i=1}^{M} \pi_*^{u,\mathcal{U}^i}(\mathbf{u}|\mathbf{x}_k)\, p_k(i),
\end{aligned} \tag{9}$$

for $\mathbf{u} \in \{\mathbf{u}^1, ..., \mathbf{u}^N\}$. Note that the cell defense response in (9) represents the intervention belief about the cell policy since the cell performs the optimal Nash policy corresponding to the true cell space.

The Bayesian policy in (8) yields the optimality with respect to the posterior distribution of the cell spaces. The schematic diagram of the proposed Bayesian intervention policy is shown in Fig. 2. As the next intervention is performed and the next state is observed, the posterior distribution over the cell spaces becomes updated, and the optimal Bayesian policy can also be computed according to the new posterior and the next observed state. The analysis of the proposed Bayesian policy and its comparison with the state-of-art intervention policies are described in Section 6.

5. Matrix-Form formulation of the Proposed Bayesian Intervention Policy

This section provides an efficient and recursive computation of the proposed Bayesian intervention policy. The process is divided into offline and online steps. The offline step consists of computing the space-specific optimal Nash policies associated with all cell spaces. Upon termination of the offline step, the online step computes the posterior distribution of all cell spaces given the last observed state, followed by the calculation of the Bayesian intervention policy. The details of these two steps are outlined below.

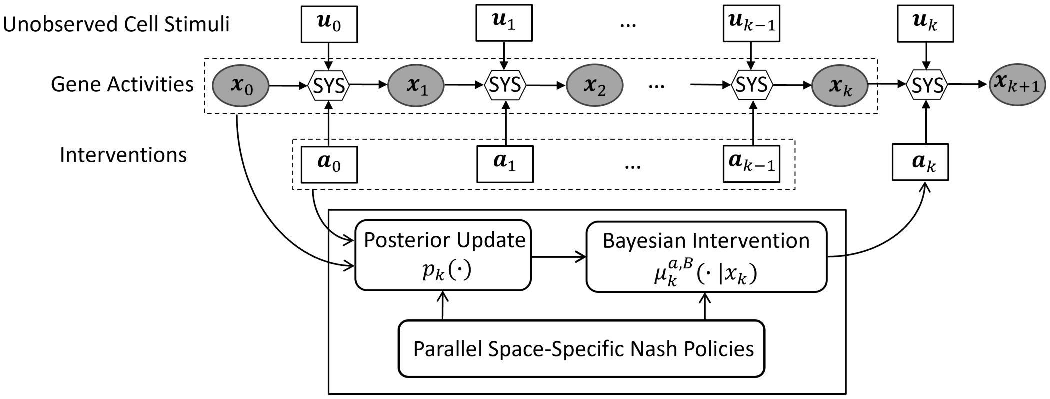


Figure 2: The schematic diagram of the proposed Bayesian intervention policy.

5.1. Offline Step Computation

The offline step computes the optimal space-specific optimal Nash equilibrium policy for all cell spaces, i.e., $\{\mathcal{U}^1, ..., \mathcal{U}^M\}$. This is achieved according to the value iteration method for a two-player zero-sum game [33]. For the ith cell space $\mathcal{U}^i$, we define the state joint-action value function for any state value function $\mathbf{V} : \mathcal{X} \to \mathbb{R}$ as:

$$Q_{\mathbf{V}}^{a,\mathcal{U}^i}(\mathbf{x}, \mathbf{a}, \mathbf{u}) = \mathbb{E}_{\mathbf{x}' \sim P(.|\mathbf{x},\mathbf{a},\mathbf{u})}\left[R^a(\mathbf{x}, \mathbf{a}, \mathbf{u}, \mathbf{x}') + \gamma \mathbf{V}(\mathbf{x}')\right], \tag{10}$$

for $\mathbf{x} \in \mathcal{X}, \mathbf{a} \in \mathcal{A}$ and $\mathbf{u} \in \mathcal{U}^i$. $Q_{\mathbf{V}}^{a,U^i}(\mathbf{x}, ., .)$ can be seen as a matrix in $\mathbb{R}^{|\mathcal{A}| \times |\mathcal{U}^i|}$, with elements representing the expected discounted accumulated rewards for the intervention when the joint actions $(\mathbf{a}, \mathbf{u})$ are performed at state $\mathbf{x}$ and the policy associated with the state value function $\mathbf{V}$ is followed.

We define the *joint-action transition matrix* associated with $(\mathbf{a}, \mathbf{u})$ in $\mathbb{R}^{2^d \times 2^d}$ as:

$$\begin{aligned}(M(\mathbf{a}, \mathbf{u}))_{lj} &= P\left(\mathbf{x}_k = \mathbf{x}^j \mid \mathbf{x}_{k-1} = \mathbf{x}^l, \mathbf{a}_{k-1} = \mathbf{a}, \mathbf{u}_{k-1} = \mathbf{u}\right) \\ &= p^{||\mathbf{f}(\mathbf{x}^l) \oplus \mathbf{a} \oplus \mathbf{u} \oplus \mathbf{x}^j||_1}(1-p)^{d-||\mathbf{f}(\mathbf{x}^l) \oplus \mathbf{a} \oplus \mathbf{u} \oplus \mathbf{x}^j||_1},\end{aligned} \tag{11}$$

for $l, j = 1, \ldots, 2^d$, $\mathbf{a} \in \mathcal{A}$, and $\mathbf{u} \in \mathcal{U}^i$, where $||.||_1$ is the absolute L-1 norm of a vector. Under zero noise and stochasticity, $\mathbf{f}(\mathbf{x}^l) \oplus \mathbf{a} \oplus \mathbf{u}$ represents the state of genes in the next time step. Thus, $||\mathbf{f}(\mathbf{x}^l) \oplus \mathbf{a} \oplus \mathbf{u} \oplus \mathbf{x}^j||_1$ counts the number of flips caused by the noise once the system moves from state $\mathbf{x}^l$ to state $\mathbf{x}^j$. The transition probability in (11) is computed based on the noise characteristics for each variable, modeled as independent Bernoulli variables with parameter p.

The matrix form representation of the intervention reward function associated with $\mathbf{a}$ and $\mathbf{u}$ can be expressed as:

$$(\mathbf{R}^a(\mathbf{a},\mathbf{u}))_{lj} \;=\; R^a\left(\mathbf{x}^l,\mathbf{a},\mathbf{u},\mathbf{x}^j\right)\;,\text{ for } l,j=1,...,2^d. \tag{12}$$

The expected intervention reward in state $\mathbf{x}^l$ after taking actions $(\mathbf{a},\mathbf{u})$ and before observing the next state can be computed as:

$$\begin{aligned} R^a(\mathbf{x}^l,\mathbf{a},\mathbf{u}) &= \mathbb{E}_{\mathbf{x}'|\mathbf{x},\mathbf{a},\mathbf{u}}[R^a(\mathbf{x}^l,\mathbf{a},\mathbf{u},\mathbf{x}')] \\ &= \sum_{j=1}^{2^d} P(\mathbf{x}_k=\mathbf{x}^j \mid \mathbf{x}_{k-1}=\mathbf{x}^l,\mathbf{a}_{k-1}=\mathbf{a},\mathbf{u}_{k-1}=\mathbf{u})R^a(\mathbf{x}^l,\mathbf{a},\mathbf{u},\mathbf{x}^j), \end{aligned} \tag{13}$$

for $l=1,..,2^d$. The expected reward in (13) can be rewritten according to (11) and (12) as:

$$R^a(\mathbf{x}^l,\mathbf{a},\mathbf{u}) = \sum_{j=1}^{2^d} (\mathbf{R}^a(\mathbf{a},\mathbf{u}))_{lj}\,(M(\mathbf{a},\mathbf{u}))_{lj}. \tag{14}$$

We define the expected intervention reward function in a vector form as $R^a_{\mathbf{a},\mathbf{u}} = [R^a(\mathbf{x}^1,\mathbf{a},\mathbf{u}),\cdots,R^a(\mathbf{x}^{2^d},\mathbf{a},\mathbf{u})]^T$. This vector can be computed using the following matrix-form computation:

$$R^a_{\mathbf{a},\mathbf{u}} = (\mathbf{R}^a(\mathbf{a},\mathbf{u})\odot M(\mathbf{a},\mathbf{u}))\,\mathbf{1}_{2^d\times 1}, \tag{15}$$

for $\mathbf{a}\in\mathcal{A}$ and $\mathbf{u}\in\mathcal{U}^i$; where $\mathbf{1}_{2^d\times 1}$ is a vector of size 2^d with all elements 1, and $\odot$ is the Hadamard product.

According to the controlled transition matrix $M(\mathbf{a},\mathbf{u})$ and the vector-form reward function $R^a_{\mathbf{a},\mathbf{u}}$, the Q-values defined in (10) can be calculated as:

$$\begin{bmatrix} Q^{a,\mathcal{U}^i}_{\mathbf{V}}(\mathbf{x}^1,\mathbf{a},\mathbf{u}) \\ \vdots \\ Q^{a,\mathcal{U}^i}_{\mathbf{V}}(\mathbf{x}^{2^d},\mathbf{a},\mathbf{u}) \end{bmatrix} = R^a_{\mathbf{a},\mathbf{u}} + \gamma M(\mathbf{a},\mathbf{u})\,\mathbf{V}, \tag{16}$$

for $\mathbf{a}\in\mathcal{A}$ and $\mathbf{u}\in\mathcal{U}^i$ and any given state value function $\mathbf{V}$.

Let π^a be 2^d-simplex of size $\mathcal{A}$, and π^u be 2^d-simplex of size $\mathcal{U}^i$. Consider $Q^{a,\mathcal{U}^i}_{\mathbf{V}}(\mathbf{x},.,.)$ as the payoff matrix in a matrix form zero-sum game. We define the Bellman operator $\mathcal{T}^*$ for any $\mathbf{x}\in\mathcal{X}$ as [33]:

$$\begin{aligned} (\mathcal{T}^*[\mathbf{V}])(\mathbf{x}) &= \text{Value}[Q^{a,\mathcal{U}^i}_{\mathbf{V}}(\mathbf{x},.,.)] \\ &= \max_{\pi^a}\min_{\pi^u}\sum_{\mathbf{a}\in\mathcal{A}}\sum_{\mathbf{u}\in\mathcal{U}^i}\pi^a(\mathbf{a}|\mathbf{x})\,\pi^u(\mathbf{u}|\mathbf{x})\,Q^{a,\mathcal{U}^i}_{\mathbf{V}}(\mathbf{x},\mathbf{a},\mathbf{u}), \end{aligned} \tag{17}$$

which should meet the condition $\sum_{\mathbf{a}\in\mathcal{A}} \pi^a(\mathbf{a}|\mathbf{x}) = \sum_{\mathbf{u}\in\mathcal{U}^i} \pi^u(\mathbf{u}|\mathbf{x}) = 1$. The solution for the min-max optimization in (17) can be obtained using a linear programming technique.

The Bellman operator $\mathcal{T}^*$ is a γ-contractive in the L_∞-norm, and the exclusive solution to the Bellman equation corresponds to the optimal value function, denoted as $\mathbf{V}^* = \mathcal{T}^*[\mathbf{V}^*]$ [33]. This fixed-point solution represents an optimal Nash equilibrium for the Markov game, associated with the cell space $\mathcal{U}^i$. Therefore, starting from any arbitrary $\mathbf{V}$, we can repeatedly apply $\mathbf{V}_{t+1} = \mathcal{T}^*[\mathbf{V}_t]$ for $t = 0, 1, ...$, and compute a fixed point solution for the value vector.

Let $\mathbf{V}_0 = [0, \cdots, 0]^T$ denote the initial value vector with all elements set to 0. During the rth iteration of the value iteration method, the new vector $\mathbf{V}_{r+1}$ is obtained upon performing the Bellman operator to the previous value $\mathbf{V}_r$ as:

$$\mathbf{V}_{r+1}(\mathbf{x}^l) = \text{Value}[Q^{a,\mathcal{U}^i}_{\mathbf{V}_r}(\mathbf{x}^l, ., .)], \text{ for } l = 1, ..., 2^d, \tag{18}$$

where $Q^{a,\mathcal{U}^i}_{\mathbf{V}_r}(\mathbf{x}^l, ., .)$ consists of Q-values for all joint pairs of $(\mathbf{a}, \mathbf{u})$. In practice, the iterations continue until the maximum difference between the elements of the value vectors in two consecutive iterations becomes smaller than a predetermined threshold $\epsilon > 0$, expressed as:

$$\max_{l\in\{1,..,2^d\}} |\mathbf{V}_T(l) - \mathbf{V}_{T-1}(l)| < \epsilon.$$

Let $\mathbf{V}_T = \mathbf{V}^*$ be the fixed-point solution after conducting the value iteration method. The Q-values associated with $\mathbf{V}^*$ can be computed as:

$$\begin{bmatrix} Q^{a,\mathcal{U}^i}_{\mathbf{V}^*}(\mathbf{x}^1, \mathbf{a}, \mathbf{u}) \\ \vdots \\ Q^{a,\mathcal{U}^i}_{\mathbf{V}^*}(\mathbf{x}^{2^d}, \mathbf{a}, \mathbf{u}) \end{bmatrix} = R^a_{\mathbf{a},\mathbf{u}} + \gamma M(\mathbf{a}, \mathbf{u})\, \mathbf{V}^*, \text{ for } \mathbf{a} \in \mathcal{A}, \mathbf{u} \in \mathcal{U}^i. \tag{19}$$

After computation of the optimal Q-values, the optimal policy for intervention and cell can be calculated as:

$$\left(\pi^{a,\mathcal{U}^i}_*(.|\mathbf{x}), \pi^{u,\mathcal{U}^i}_*(.|\mathbf{x})\right) = \underset{\pi^a}{\text{argmax}}\, \underset{\pi^u}{\text{argmin}} \sum_{\mathbf{a}\in\mathcal{A}} \sum_{\mathbf{u}\in\mathcal{U}^i} \pi^a(\mathbf{a}|\mathbf{x})\, \pi^u(\mathbf{u}|\mathbf{x})\, Q^{a,\mathcal{U}^i}_{\mathbf{V}^*}(\mathbf{x}, \mathbf{a}, \mathbf{u}), \tag{20}$$

for any $\mathbf{x} \in \mathcal{X}$, where $\pi^{a,\mathcal{U}^i}(\mathbf{a}|\mathbf{x})$ and $\pi^{a,\mathcal{U}^i}(\mathbf{u}|\mathbf{x})$ are non-negative numbers that add up to 1 for any $\mathbf{x} \in \mathcal{X}$. The solution to the Nash equilibrium policy in (20) can be obtained using a linear programming technique. Repeating the above process for all cell spaces leads to the computation of the space-specific Nash policies in the offline step.

Algorithm 1 Bayesian Intervention Policy

```
Intervention space $\mathcal{A}$; Cell spaces, $\mathcal{U}^1, \ldots, \mathcal{U}^M$; intervention reward $(\mathbf{R}^a(\mathbf{a}, \mathbf{u}))_{lj} = R^a(\mathbf{x}^l, \mathbf{a}, \mathbf{u}, \mathbf{x}^j)$; controlled transition matrix $M(\mathbf{a}, \mathbf{u})$; threshold $\epsilon > 0$.
Offline Step
for $\mathcal{U}^i \in \{\mathcal{U}^1, \ldots, \mathcal{U}^M\}$ do
    Set $\mathbf{V}' = \mathbf{0}_{2^d \times 1}$.
    repeat
        $\mathbf{V} = \mathbf{V}'$.
        $\begin{bmatrix} Q^{a,\mathcal{U}^i}_{\mathbf{V}}(\mathbf{x}^1, \mathbf{a}, \mathbf{u}) \\ \vdots \\ Q^{a,\mathcal{U}^i}_{\mathbf{V}}(\mathbf{x}^{2^d}, \mathbf{a}, \mathbf{u}) \end{bmatrix} = \left[ (\mathbf{R}^a(\mathbf{a}, \mathbf{u}) \odot M(\mathbf{a}, \mathbf{u}))\, \mathbf{1}_{2^d \times 1} + \gamma M(\mathbf{a}, \mathbf{u})\, \mathbf{V} \right]$, for $\mathbf{a} \in \mathcal{A}$ and $\mathbf{u} \in \mathcal{U}^i$.
        Bellman Operator: $\mathbf{V}'(\mathbf{x}^l) = \text{Value}[Q^{a,\mathcal{U}^i}_{\mathbf{V}}(\mathbf{x}^l, ., .)]$, for $l = 1, \ldots, 2^d$ – Eq. (17)
    until $\max_{l \in \{1,\ldots,2^d\}} |\mathbf{V}(\mathbf{x}^l) - \mathbf{V}'(\mathbf{x}^l)| < \epsilon$
    For any given $\mathbf{x} \in \mathcal{X}$, use linear programming approach over $Q^{a,\mathcal{U}^i}_{\mathbf{V}'}(\mathbf{x}, ., .)$ to obtain $\pi^{a,\mathcal{U}^i}_*(.|\mathbf{x})$ and $\pi^{u,\mathcal{U}^i}_*(.|\mathbf{x})$.
end for
Online Step
Initial state $\mathbf{x}_0$, and initial probability of cell space: $p_0 = [P(\mathcal{U}^1), \ldots, P(\mathcal{U}^M)]$.
for $k = 0, 1, 2, \ldots,$ do
    Compute Bayesian Intervention $\mu^{a,\text{B}}_k(\mathbf{a} \mid \mathbf{x}_k) = \sum_{i=1}^{M} \pi^{a,\mathcal{U}^i}_*(\mathbf{a} \mid \mathbf{x}_k)\, p_k(i), \mathbf{a} \in \mathcal{A}$, and select action accordingly: $\mathbf{a}_k \sim \mu^{a,\text{B}}_k(. \mid \mathbf{x}_k)$.
    Apply the intervention $\mathbf{a}_k$ and receive the next system state, $\mathbf{x}_{k+1}$.
    Posterior Update:
    $p_{k+1}(i) = \dfrac{\left[ \sum_{\mathbf{u} \in \mathcal{U}^i} \left( \frac{p}{1-p} \right)^{||\mathbf{f}(\mathbf{x}_k) \oplus \mathbf{a}_k \oplus \mathbf{u} \oplus \mathbf{x}_{k+1}||_1} \pi^{u,\mathcal{U}^i}_*(\mathbf{u}|\mathbf{x}_k) \right] p_k(i)}{\sum_{j=1}^{M} \left[ \sum_{\mathbf{u} \in \mathcal{U}^j} \left( \frac{p}{1-p} \right)^{||\mathbf{f}(\mathbf{x}_k) \oplus \mathbf{a}_k \oplus \mathbf{u} \oplus \mathbf{x}_{k+1}||_1} \pi^{u,\mathcal{U}^j}_*(\mathbf{u}|\mathbf{x}_k) \right] p_k(j)}, i = 1, \ldots, M.$
end for
```

5.2. Online Step Computation

This section describes a recursive and online computation of the Bayesian intervention policy, obtained according to the space-specific Nash equilibrium policies computed during the offline step. Let p_k contain the posterior probability of the cell spaces and $\mathbf{x}_k$ be the system state at time step k. An intervention at time step k can be selected according to the Bayesian policy in (8) as:

$$\mathbf{a}_k \sim \mu_k^{a,\mathrm{B}}(. \mid \mathbf{x}_k), \tag{21}$$

where

$$\mu_k^{a,\mathrm{B}}(\mathbf{a} \mid \mathbf{x}_k) = \sum_{i=1}^{M} \pi_*^{a,\mathcal{U}^i}(\mathbf{a} \mid \mathbf{x}_k)\, p_k(i), \text{ for } \mathbf{a} \in \mathcal{A}. \tag{22}$$

Upon performing the intervention $\mathbf{a}_k$ and observing the next state $\mathbf{x}_{k+1}$, the posterior distribution of the cell spaces can be updated using (6) and (7) as:

$$p_{k+1}(i) = \frac{\left[\sum_{\mathbf{u}\in\mathcal{U}^i} \left(\frac{p}{1-p}\right)^{||\mathbf{f}(\mathbf{x}_k)\oplus\mathbf{a}_k\oplus\mathbf{u}\oplus\mathbf{x}_{k+1}||_1} \pi_*^{u,\mathcal{U}^i}(\mathbf{u}|\mathbf{x}_k)\right] p_k(i)}{\sum_{j=1}^{M}\left[\sum_{\mathbf{u}\in\mathcal{U}^j} \left(\frac{p}{1-p}\right)^{||\mathbf{f}(\mathbf{x}_k)\oplus\mathbf{a}_k\oplus\mathbf{u}\oplus\mathbf{x}_{k+1}||_1} \pi_*^{u,\mathcal{U}^j}(\mathbf{u}|\mathbf{x}_k)\right] p_k(j)}, \tag{23}$$

for $i = 1, ..., M$.

The diagram in Fig. 3 represents the processes of the computation of the proposed intervention policy in the offline and online steps. Algorithm 1 provides the details of the computations in both steps. Meanwhile, the complexity of each step is provided in Table 1. The offline step has a computational complexity of order $O(2^{2d} \times |\mathcal{A}| \times \max_{i=1,...,M} |\mathcal{U}^i| \times L)$, where 2^{2d} is due to the transition matrices involved, L represents the number of steps of the value iteration method before termination, $|\mathcal{A}|$ is the size of intervention space, and the $|\mathcal{U}^i|$ is the size of the ith cell space. In the online step, the computation of the Bayesian intervention has the complexity of order $O(M)$, whereas the posterior update's complexity is of order $O(M \times \max_{i=1,...,M} |\mathcal{U}^i|)$. Overall, the complexity of the online step is significantly lower than that of the offline step, enabling a recursive computation of the proposed intervention policy.

6. Performance Analysis and Comparison with State-of-Art Methods

This section analyzes the performance of the proposed Bayesian intervention policy with the system under no intervention and some of the existing intervention policies. First, consider a system with no intervention under the aggressive

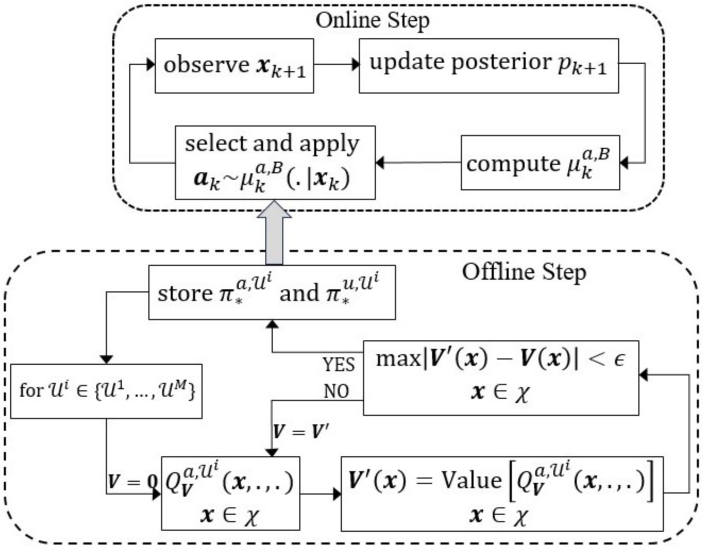

Online Step
observe $\boldsymbol{x}_{k+1}$
update posterior p_{k+1}
select and apply
$\boldsymbol{a}_k \sim \mu_k^{a,B}(.|\boldsymbol{x}_k)$
compute $\mu_k^{a,B}$
Offline Step
store $\pi_*^{a,\mathcal{U}^i}$ and $\pi_*^{u,\mathcal{U}^i}$
YES
NO
$\max|\boldsymbol{V}'(\boldsymbol{x}) - \boldsymbol{V}(\boldsymbol{x})| < \epsilon$
$\boldsymbol{x} \in \chi$
for $\mathcal{U}^i \in \{\mathcal{U}^1, \dots, \mathcal{U}^M\}$
$\boldsymbol{V} = \boldsymbol{V}'$
$\boldsymbol{V} = \boldsymbol{0}$
$Q_{\boldsymbol{V}}^{a,\mathcal{U}^i}(\boldsymbol{x},.,.)$
$\boldsymbol{x} \in \chi$
$\boldsymbol{V}'(\boldsymbol{x}) = \text{Value}\left[Q_{\boldsymbol{V}}^{a,\mathcal{U}^i}(\boldsymbol{x},.,.)\right]$
$\boldsymbol{x} \in \chi$

Figure 3: The schematic diagram of processes in the offline and online steps of the proposed Bayesian intervention policy.

Table 1: Computational complexity of the proposed Bayesian intervention policy.

Offline Step (Cell Space $\mathcal{U}^i$)	Bayesian Intervention	Posterior Update
$O(2^{2d} \times \lvert\mathcal{A}\rvert \times \lvert\mathcal{U}^i\rvert \times L_i)$	$O(M)$	$O(M \times \max\{\lvert\mathcal{U}^1\rvert, ..., \lvert\mathcal{U}^M\rvert\})$

response of cells, e.g., representing uncontrolled cancerous conditions. The best cell policy under no intervention is deterministic. Let $\pi^u : \mathcal{X} \to \mathcal{U}^*$ be a deterministic cell policy, which maps a cell action in $\mathcal{U}^*$ to each system state. The optimal cell response under no intervention can be computed as:

$$\pi_*^{u,\mathbf{a}=\mathbf{0}}(\mathbf{x}) = \underset{\pi^u}{\operatorname{argmin}}\, \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R^a(\mathbf{x}_t, \mathbf{a}_t = \mathbf{0}, \mathbf{u}_t, \mathbf{x}_{t+1}) \mid \mathbf{x}_0 = \mathbf{x}, \mathbf{u}_{0:\infty} \sim \pi^u\right], \tag{24}$$

where $\pi^u \in (\mathcal{U}^*)^{2^d}$ and the minimization is used since the reward of the intervention is negative of the cell reward function. The steady-state probability under no intervention can be expressed as:

$$\mathbf{\Pi}_{\mathbf{a}=\mathbf{0}}^{\infty}(j) = \lim_{k\to\infty} P(\mathbf{x}_k = \mathbf{x}^j \mid \mathbf{u}_{0:\infty} \sim \pi_*^{\mathbf{u},\mathbf{a}=\mathbf{0}}, \mathbf{a}_{0:\infty} = \mathbf{0}), \tag{25}$$

for $j = 1...., 2^d$. One can see $\mathbf{\Pi}_{\mathbf{a}=\mathbf{0}}^{\infty}$ as a long-term probability of the visitation of various states under no intervention.

Most conventional intervention methods assume non-responsive cells [19], wherein cells lack defense mechanisms to counteract interventions (i.e., $\mathcal{U} = \{\}$). In this scenario, the Markov game can be represented by an MDP with a single agent/player, and since the intervention is driven by no competition with cell responses assumption, the optimal intervention policy becomes deterministic. This policy can be expressed as:

$$\pi_*^{a,\mathbf{u}=\mathbf{0}}(\mathbf{x}) = \operatorname*{argmax}_{\pi^a} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R^a(\mathbf{x}_t, \mathbf{a}_t, \mathbf{u}_t = \mathbf{0}, \mathbf{x}_{t+1}) \mid \mathbf{x}_0 = \mathbf{x}, \mathbf{a}_{0:\infty} \sim \pi^a\right], \tag{26}$$

where the maximization is over all deterministic intervention policies, i.e., $(\mathcal{A})^{2^d}$. The cell's aggressive response to the naive and deterministic intervention in (26) can be expressed as:

$$\pi_*^{u,\pi_*^{a,\mathbf{u}=\mathbf{0}}}(\mathbf{x}) = \operatorname*{argmin}_{\pi^u} \mathbb{E}\Bigg[\sum_{t=0}^{\infty} \gamma^t R^a(\mathbf{x}_t, \mathbf{a}_t, \mathbf{u}_t, \mathbf{x}_{t+1}) \mid \mathbf{x}_0 = \mathbf{x},$$
$$\mathbf{a}_{0:\infty} \sim \pi_*^{a,\mathbf{u}=\mathbf{0}}, \mathbf{u}_{0:\infty} \sim \pi^u\Bigg], \text{ for } \mathbf{x} \in \mathcal{X}. \tag{27}$$

The expected value function for the intervention under no cell response policy in (26) and (27) can be expressed through $\mathbf{V}^a_{\pi_*^{a,\mathbf{u}=\mathbf{0}},\pi_*^{u,\pi_*^{a,\mathbf{u}=\mathbf{0}}}}$. The intervention gain obtained under this policy compared to no intervention case can be expressed as:

$$\mathbf{V}^a_{\pi_*^{a,\mathbf{u}=\mathbf{0}},\pi_*^{u,\pi_*^{a,\mathbf{u}=\mathbf{0}}}}(\mathbf{x}) - \mathbf{V}^a_{\mathbf{0},\pi_*^{u,\mathbf{a}=\mathbf{0}}}(\mathbf{x}) \geq 0, \tag{28}$$

for any $\mathbf{x} \in \mathcal{X}$. The positivity of the difference in the state values indicates that the intervention helps the system to experience less undesirable conditions, compared to cases with no intervention. Meanwhile, the comparison with the optimal Nash policy $(\pi_*^{a,\mathcal{U}^*}, \pi_*^{u,\mathcal{U}^*})$ can be achieved as:

$$\mathbf{V}^a_{\pi_*^{a,\mathbf{u}=\mathbf{0}},\pi_*^{u,\pi_*^{a,\mathbf{u}=\mathbf{0}}}}(\mathbf{x}) \leq \mathbf{V}^a_{\pi_*^{a,\mathbf{u}=\mathbf{0}},\pi_*^{u,\mathcal{U}^*}}(\mathbf{x}) \leq \mathbf{V}^a_{\pi_*^{a,\mathcal{U}^*},\pi_*^{u,\mathcal{U}^*}}(\mathbf{x}), \tag{29}$$

for any $\mathbf{x} \in \mathcal{X}$, where the inequalities are obtained due to the fact that deviation of the intervention from the optimal Nash policy leads to a reduction in the intervention performance (see (3)). More specifically, if the intervention policy deviates from the Nash policy, the cell can take advantage of this and further shift the system toward undesirable conditions. Note that the conventional intervention policies can achieve the same performance level as the optimal Nash policy if and only if the optimal Nash policy is deterministic, i.e., $\pi_*^{a,\mathcal{U}^*}(\pi_*^{a,\mathbf{u}=\mathbf{0}}(\mathbf{x})|\mathbf{x}) = 1$, for all $\mathbf{x} \in \mathcal{X}$.

In this part, the difference between the state-value function of the proposed Bayesian intervention policy and the optimal Nash policy is investigated. The proposed Bayesian policy is adaptive, meaning that its policy becomes updated according to the latest observed states. We represent the Bayesian policy after time step k as $\mu_{k:\infty}^{a,\mathrm{B}} = [\mu_k^{a,\mathrm{B}}, \mu_{k+1}^{a,\mathrm{B}}, ...]$, where $\mu_{k+1}^{a,\mathrm{B}}$ yields optimality with respect to the information up to time step $k+1$. Thus, we can express the difference between the state-value functions of the proposed Bayesian policy and the optimal Nash policy as:

$$\mathbf{V}^a_{\mu_{k:\infty}^{a,\mathrm{B}}, \pi_*^{a,\mathcal{U}^*}}(\mathbf{x}_k) - \mathbf{V}^a_{\pi_*^{a,\mathcal{U}^*}, \pi_*^{u,\mathcal{U}^*}}(\mathbf{x}_k) \leq 0. \tag{30}$$

It can be shown that the state value function of the Bayesian policy becomes close to the optimal Nash policy as time progresses. In fact, for a sufficiently large value of k, the posterior distribution over the cell spaces is expected to become peaked over the true cell space, and according to (8), the Bayesian policy becomes the same as the optimal Nash policy. In particular, the difference between the proposed Bayesian policy at time step k and the optimal Nash policy can be expressed as follows:

$$\begin{aligned}\mathrm{KL}(\pi_*^{a,\mathcal{U}^*}(.|\mathbf{x}_k), \mu_k^{a,\mathrm{B}}(.|\mathbf{x}_k)) &= \sum_{\mathbf{a}\in\mathcal{A}} \pi_*^{a,\mathcal{U}^*}(\mathbf{a}|\mathbf{x}_k) \log \frac{\pi_*^{a,\mathcal{U}^*}(\mathbf{a}|\mathbf{x}_k)}{\mu_k^{a,\mathrm{B}}(\mathbf{a}|\mathbf{x}_k)} \\ &= \sum_{\mathbf{a}\in\mathcal{A}} \pi_*^{a,\mathcal{U}^*}(\mathbf{a}|\mathbf{x}_k) \left[\log \pi_*^{a,\mathcal{U}^*}(\mathbf{a}|\mathbf{x}_k) - \log \mu_k^{a,\mathrm{B}}(\mathbf{a}|\mathbf{x}_k)\right],\end{aligned} \tag{31}$$

where KL indicates the Kullback-Leibler divergence. The KL approaches zero if the posterior peaks over a single cell space (i.e., the true cell space). Finally, unlike existing deterministic intervention policies, the stochastic nature of the proposed policy aligns with the stochastic nature of the optimal Nash policy. This stochasticity prevents the cell from predicting a single deterministic intervention in different cases, helping to ensure short-term and long-term success during the intervention process.

7. Numerical Experiments

In this section, the performance of the proposed intervention policy is assessed through two well-known gene regulatory networks: the p53-MDM2 Boolean network model and the melanoma regulatory network.

7.1. P53-MDM2 Negative Feedback Loop Network

This paper utilizes a simplified p53-MDM2 Boolean network [44] with DNA double-strand break (DNA-DSB) for the experiment. This network has been widely studied for assessing the performance of various intervention policies. The p53 tumor suppressor is a crucial transcription factor that regulates essential cellular processes, including DNA repair, cell cycle control, apoptosis, angiogenesis, and senescence [45]. Fig. 4(a) illustrates the diagram of this network, where solid and blunt arrows indicate activating and suppressive interactions, respectively. The network consists of four genes: ATM, p53, WIP1, MDM2, and DNA-DSB, which is an external stress to the cell. The system state is represented using the following vector: $\mathbf{x}_k = [\mathrm{ATM}_k, \mathrm{p53}_k, \mathrm{WIP1}_k, \mathrm{MDM2}_k]$. The Boolean model described in (1) represents the state transition of the healthy system as:

$$\mathbf{x}_k = \overline{\begin{bmatrix} 0 & 0 & -1 & 0 \\ +1 & 0 & -1 & -1 \\ 0 & +1 & 0 & 0 \\ -1 & +1 & +1 & 0 \end{bmatrix} \mathbf{x}_{k-1} + \begin{bmatrix} \text{dna\_dsb} \\ 0 \\ 0 \\ 0 \end{bmatrix}} \oplus \mathbf{a}_{k-1} \oplus \mathbf{u}_{k-1} \oplus \mathbf{n}_k, \tag{32}$$

where $\overline{\mathbf{v}}$ is a function that maps the element of the vector $\mathbf{v}$ greater than 0 to 1 and others to 0.

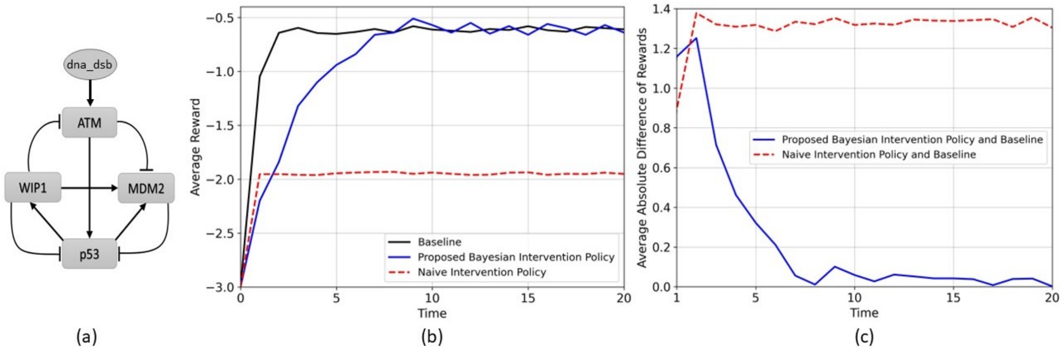


Figure 4: (a) The pathway diagram for the p53-MDM2 Boolean network. (b) The average reward gained by the Bayesian intervention policy, naive intervention policy, and the Baseline. (c) The average absolute difference of the rewards.

In cells under normal conditions, the stress response is zero (i.e., dna_dsb $=$ 0), whereas under stressed conditions, the stress is present (i.e., dna_dsb $=$ 1). For no stressed cells, the genes' states are mostly at rest, i.e., the system remains in the "0000" state. In stressed conditions, the activation and inactivation of p53 help the system control the genes' activities and cell proliferation. However, when p53, a tumor suppressor gene, undergoes a loss of function, other genes can exhibit excessive activations and cell proliferation, leading to transitioning from a healthy to a cancerous condition.

The cell defensive responses are modeled using single-gene and double-gene perturbations. This represents realistic situations in which cells have the capability to respond to therapies by altering the states of multiple genes simultaneously. Therefore, the possible cell responses can be expressed through the following 7 actions:

$$\begin{gathered}\mathbf{u}^1 = [0\,0\,0\,0]^T, \mathbf{u}^2 = [1\,0\,0\,0]^T, \mathbf{u}^3 = [0\,0\,1\,0]^T, \mathbf{u}^4 = [0\,0\,0\,1]^T, \\ \mathbf{u}^5 = [1\,0\,1\,0]^T, \mathbf{u}^6 = [1\,0\,0\,1]^T, \mathbf{u}^7 = [0\,0\,1\,1]^T.\end{gathered} \tag{33}$$

The cell might utilize one or multiple stimuli in response to interventions. In our experiment, we consider the following cell space to be true but unknown:

$$\mathcal{U}^* = \{\mathbf{u}^2, \mathbf{u}^6\}, \tag{34}$$

where $\mathbf{u}^2$ alters the state value of ATM, and $\mathbf{u}^6$ simultaneously alters the state of ATM and MDM2.

Toward modeling the possible cell spaces, we consider cell spaces to contain any subset of one, two, and three elements from the above 7 possible cell actions in (33). This leads to $M = \binom{7}{1} + \binom{7}{2} + \binom{7}{3} = 63$ possible cell spaces. Among them, 7 contain a single action, denoted by $\mathcal{U}^1$ to $\mathcal{U}^7$, 21 contain two actions indicated by $\mathcal{U}^8$ to $\mathcal{U}^{28}$, and 35 consist of 3 actions, indicated by $\mathcal{U}^{29}$ to $\mathcal{U}^{63}$. Note that the true cell space in (34) is the 17th space (i.e., $\mathcal{U}^* = \mathcal{U}^{17}$), which is unknown during the intervention.

The space of intervention (i.e., drugs/therapies) is assumed to be:

$$\mathcal{A} = \{\mathbf{a}^1 = [0\,0\,0\,0]^T, \mathbf{a}^2 = [1\,0\,0\,0]^T, \mathbf{a}^3 = [0\,0\,0\,1]^T\}, \tag{35}$$

where the first intervention $\mathbf{a}^1$ corresponds to no therapy, whereas the second and third interventions alter the state value of the ATM and MDM2 genes, respectively.

Intervention aims to reduce cell proliferation in cancerous situations and restore the system to a normal condition. For the p53-MDM2 network, this can be achieved by reducing the activation of ATM, WIP1, and MDM2. This can be expressed through the following intervention reward function:

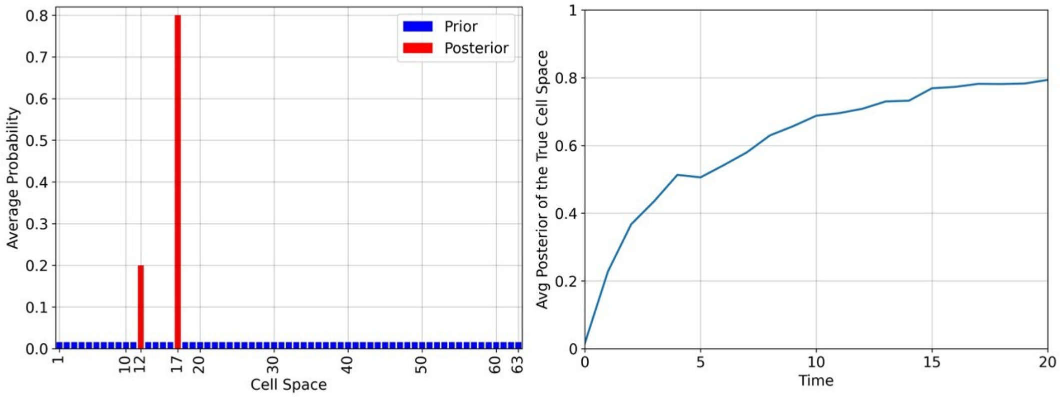
$$R^a(\mathbf{x}, \mathbf{a}, \mathbf{u}, \mathbf{x}') = -\mathbf{x}'(1) - \mathbf{x}'(3) - \mathbf{x}'(4). \tag{36}$$

The activation of each of ATM, MDM2, and WIP1 yields a negative reward of -1, resulting in an immediate reward ranging from -3 to 0. The objective of the intervention is to maximize cumulative intervention rewards by maintaining ATM, MDM2, and WIP1 in an inactivated state. Conversely, the cell with the opposing reward seeks to increase the activation of these genes and drive the system closer to states leading to uncontrolled cell proliferation.

We consider the optimal Nash policy associated with true cell space (i.e., $\pi_*^{u,\mathcal{U}^*}$, and $\pi_*^{a,\mathcal{U}^*}$) as a Baseline policy. The Baseline provides the best intervention outcomes that could be achieved by any intervention policy (since it assumes the full knowledge of true cell space). The following parameters are used for the numerical experiments: $p = 0.05$, $\gamma = 0.95$, and $\epsilon = 0.01$, the initial state "1011", representing the cancerous condition.

The average reward over 100 independent runs obtained by the proposed Bayesian intervention policy, the naive intervention policy, and the Baseline is presented in Fig. 4(b). As can be seen, the reward gained by the proposed Bayesian policy becomes closest to the Baseline after a few steps (i.e., a few numbers of interventions). The performance of the naive intervention policy is notably poor, with an average 2 out of 3 genes remaining activated. In contrast, the Bayesian intervention policy demonstrates a significant improvement by effectively deactivating approximately 2.4 of the genes, which highlights the superiority of the proposed approach. Furthermore, Fig. 4(c) shows the average absolute difference between the rewards obtained by the Baseline and the proposed Bayesian policy and the Baseline and the naive intervention policy. As can be seen, a much smaller absolute reward difference is achieved for the proposed intervention policy. In particular, the absolute reward difference approaches zero for the proposed Bayesian policy as time progresses, which means the proposed method achieves intervention performance (i.e., reward) similar to the Baseline. On the other hand, one can see the poor performance of the naive policy with a large absolute reward difference over time.

The prior and average posterior probability over cell spaces is shown in Fig. 5(a). A uniform prior is considered over cell spaces (blue bars). The average posteriors after 20 steps are shown with red bars. As can be seen, the proposed method has been almost able to discern the true cell space, i.e., $\mathcal{U}^{17}$. Aside from the true cell space, another cell space (i.e., $\mathcal{U}^{12} = \{\mathbf{u}^1, \mathbf{u}^6\}$) has a large posterior probability. This set shares a single cell action with the true cell space, making it probabilistically indistinguishable from the true cell space, given 20 observed states. Furthermore, the average posterior of the true cell space over time is shown in Fig. 5(b). The average posterior of the true cell space is increasing over time. The reason for not approaching 1 is the existence of another cell space, $\mathcal{U}^{12}$, with a similar space-specific Nash policy.



(a) (b)

Figure 5: (a) The prior and posterior (after 20 steps) probability over cell spaces. (b) The average posterior of the true cell space over time.

Fig. 6(a) represents the probability assigned to each intervention ($\mathbf{a}^1$, $\mathbf{a}^2$, and $\mathbf{a}^3$) by both the optimal Nash equilibrium policy and the proposed Bayesian policy in a single run. It can be seen that the proposed Bayesian policy and Baseline behave similarly after a few initial steps. In fact, the average result reveals that the Bayesian intervention policy empirically converges toward the optimal Nash intervention policy after approximately 7 steps.

In this part, the KL divergence is used as a distance measure between the optimal Nash equilibrium policy and the proposed Bayesian intervention policy. Fig. 6(b) represents the average KL divergence performed over 100 independent runs. The results indicate that these two policies become close to each other not only in individual runs (as shown in Fig. 6(a)), but also on average. This indicates the empirical convergence of the proposed policy to the optimal Nash policy as more interventions are taken, and more data are observed.

In this part of the experiment, we investigate the reason for obtaining a large posterior probability for a non-true cell space in Fig. 5(a). Fig. 7(a) illustrate the space-specific Nash policies under the true cell space $\mathcal{U}^*$ and the cell space $\mathcal{U}^{12}$. The blue bars represent the probability assigned to each intervention at the 16 states under the true cell space's Nash equilibrium policy, while the red bars represent the corresponding probabilities under the Nash policy associated with $\mathcal{U}^{12}$. One can see the similarity between these two policies in different states.

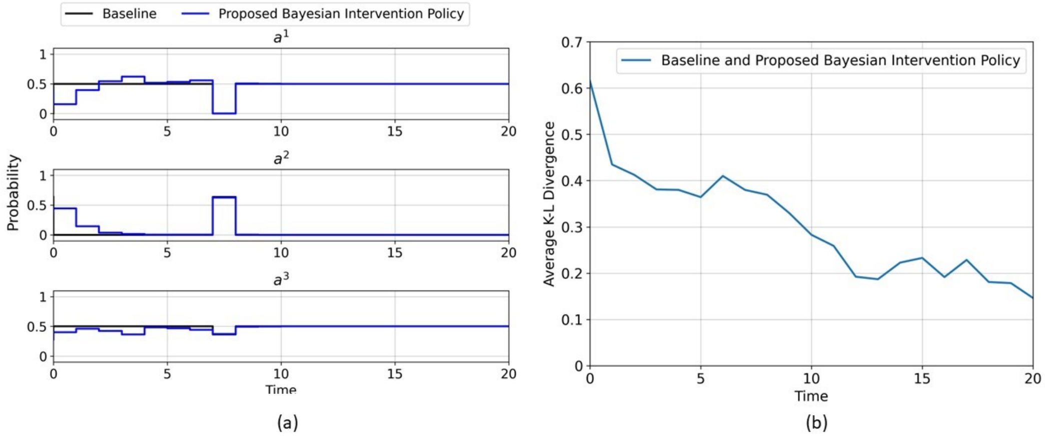


Figure 6: (a) The proposed Bayesian intervention policy and the optimal Nash equilibrium intervention policy (both stochastic) in one single run. (b) The average KL divergence between the true Nash intervention policy and the proposed Bayesian intervention policy.

The average rate of state visitations under the proposed Bayesian policy is shown in Fig. 7(b). One can see the subset of states $\{\mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^{10}, \mathbf{x}^{12}\}$ are the most frequently visited states. At these most visited states, we can see the similarity between the space-specific Nash policies associated with $\mathcal{U}^*$ and $\mathcal{U}^{12}$ in Fig. 7(a). This explains the reason behind the similar performance of the proposed Bayesian policy to the Baseline, despite a large posterior probability for a non-true cell space.

This section analyzes the impact of the system stochasticity on the performance of the proposed Bayesian policy. Fig. 8(a) illustrates the average posterior of the true cell space under two levels of state stochasticity. The solid line corresponds to the small noise level, characterized by a Bernoulli process noise with $p = 0.001$, whereas the dashed line represents a higher noise level with $p = 0.15$. The results indicate that when there is less randomness in the system (low stochasticity), the average posterior of the true cell space becomes closer to 1. However, when the stochasticity level increases (high stochasticity), there is greater uncertainty in determining the true cell space. Therefore, as expected, the proposed method performs better for less chaotic systems.

Fig. 8(b) shows the average reward obtained by the proposed Bayesian intervention policy and the naive intervention policy under low and high levels of stochasticity. The average rewards obtained by both policies have more fluctuation under a larger stochasticity level. The results indicate that the naive intervention policy performs poorly when the stochasticity level is low. Under a high stochas-

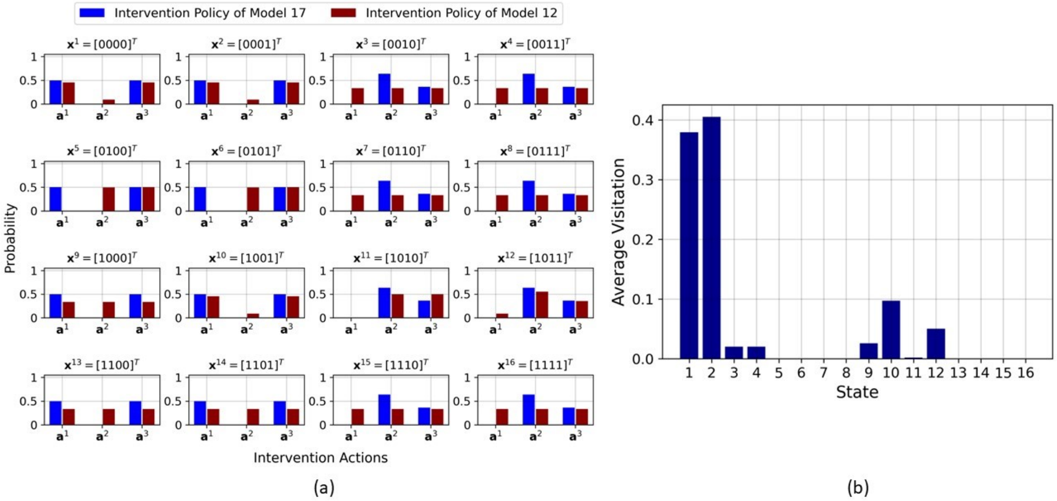


Figure 7: (a) The space-specific Nash equilibrium intervention policy associated with $\mathcal{U}^*$ and $\mathcal{U}^{12}$. (b) The average state visitation rate in 100 independent runs under the proposed Bayesian intervention policy.

ticity level, it takes longer for the proposed policy to achieve a performance similar to that of the optimal Nash equilibrium policy. However, the final average reward obtained by the proposed policy under low and high stochasticity levels is similar. This demonstrates that the proposed Bayesian policy exhibits greater robustness compared to the naive policy. In fact, in more chaotic systems characterized by higher levels of noise, decision-making becomes more challenging for both cells and intervention, resulting in similar performance regardless of changes in the noise level.

This section of numerical experiments investigates the robustness of the proposed policy with respect to different cell and intervention spaces. Table 1 presents the average reward obtained by various policies across 9 pairs of intervention and true cell spaces. The Bayesian policy and the Baseline outperform the naive policy in all cases. For a fixed intervention space (i.e., the results in a single row), a reduction in the reward can be seen for cell spaces with larger elements. This is due to the greater power of cells with larger cell space to resist intervention. Given a fixed true cell space (a column in the table), a stronger intervention space yields a larger or similar average reward. The improvement in the result is more significant when the size of the intervention space has increased from 2 to 3, and less significant once it is increased to 4.

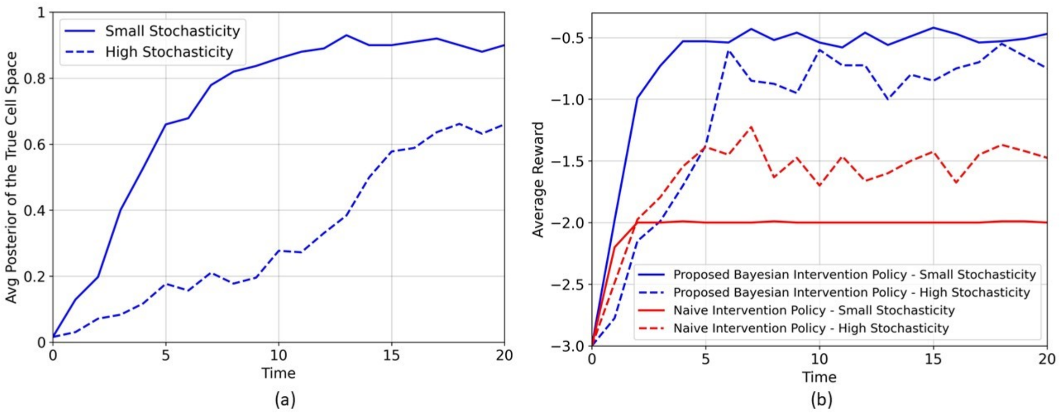


Figure 8: (a) Average posterior of the true cell space for systems with low ($p = 0.001$) and high ($p = 0.15$) levels of stochasticity. (b) The average reward gained by the Bayesian intervention policy and naive intervention policy under low ($p = 0.001$) and high ($p = 0.15$) levels of stochasticity.

7.2. *Melanoma Regulatory Network*

In this part of the numerical experiment, we evaluate the effectiveness of the proposed Bayesian intervention policy using the melanoma regulatory network. Melanoma is a deadly type of skin cancer arising from melanocytes' malignant conversion [21, 46, 47]. In this paper, we consider a well-known Boolean network model of melanoma network [21], which is widely studied in deriving genomics interventions. Fig. 9(a) illustrates the regulatory relationships among the genes in the network. This network consists of a total of 10 genes and 1,024 states. The state vector shows the activation/inactivation of the following genes in sequential order: WNT5A, pirin, S100P, RET1, MMP3, PHOC, MART1, HADHB, synuclein, and STC2. The network function can be expressed as:

Table 2: Average steady-state reward gained by different policies under different intervention sets and true cell spaces

	$\mathcal{U}^* = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$	$\mathcal{U}^* = \begin{bmatrix} 1 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}$	$\mathcal{U}^* = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$
$\mathcal{A} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$	Baseline: -0.402 ± 0.008 Bayesian: -0.415 ± 0.026 Naive: -1.319 ± 0.010	Baseline: -1.044 ± 0.013 Bayesian: -1.057 ± 0.036 Naive: -2.207 ± 0.012	Baseline: -1.802 ± 0.021 Bayesian: -1.885 ± 0.039 Naive: -2.602 ± 0.011
$\mathcal{A} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	Baseline: -0.288 ± 0.011 Bayesian: -0.297 ± 0.028 Naive: -1.188 ± 0.010	Baseline: -0.627 ± 0.016 Bayesian: -0.637 ± 0.041 Naive: -1.941 ± 0.011	Baseline: -0.833 ± 0.026 Bayesian: -0.846 ± 0.052 Naive: -2.131 ± 0.009
$\mathcal{A} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}$	Baseline: -0.193 ± 0.008 Bayesian: -0.209 ± 0.032 Naive: -1.051 ± 0.012	Baseline: -0.565 ± 0.018 Bayesian: -0.602 ± 0.053 Naive: -1.740 ± 0.014	Baseline: -0.725 ± 0.028 Bayesian: -0.744 ± 0.062 Naive: -1.969 ± 0.012

$$\mathbf{f}(\mathbf{x}_k) = [f_1(\mathbf{x}_k), f_2(\mathbf{x}_k), ..., f_{10}(\mathbf{x}_k)]^T =$$

$$\begin{bmatrix} (\text{S100P} \wedge \text{MMP3} \wedge \overline{\text{PHOC}}) \vee (\overline{\text{MMP3}} \wedge \text{PHOC}) \\ (\overline{\text{WNT5A}} \wedge \overline{\text{S100P}} \wedge \text{MMP3}) \vee (\text{WNT5A} \wedge \overline{\text{S100P}} \wedge \overline{\text{MMP3}}) \\ \text{MART1} \\ (\overline{\text{WNT5A}} \wedge \text{pirin} \wedge \text{RET1}) \vee (\overline{\text{pirin}} \wedge \text{RET1}) \\ (\text{RET1} \wedge \text{synuclein}) \vee \overline{\text{synuclein}} \\ (\overline{\text{RET1}} \wedge \overline{\text{MART1}}) \vee (\text{RET1} \wedge \text{MART1} \wedge \text{STC2}) \\ \text{MART1} \\ (\text{WNT5A} \wedge \text{MMP3}) \vee (\overline{\text{MMP3}} \wedge \overline{\text{synuclein}}) \vee (\text{WNT5A} \wedge \overline{\text{MMP3}} \wedge \text{synuclein}) \\ (\overline{\text{RET1}} \wedge \overline{\text{MART1}} \wedge \overline{\text{STC2}}) \vee (\text{RET1} \wedge \overline{\text{MART1}} \wedge \text{STC2}) \vee \text{MART1} \\ \overline{\text{S100P}} \end{bmatrix}.$$

The intervention objective is to reduce the activation of two genes: WNT5A and pirin. This can be expressed using the following intervention reward function:

$$R^a(\mathbf{x}, \mathbf{a}, \mathbf{u}, \mathbf{x}') = 2 - \mathbf{x}'(1) - \mathbf{x}'(2), \tag{37}$$

where the reward of 2 is reached if both genes are inactivated, 1 if one of them is activated, and 0 when both genes are in the inactivated state.

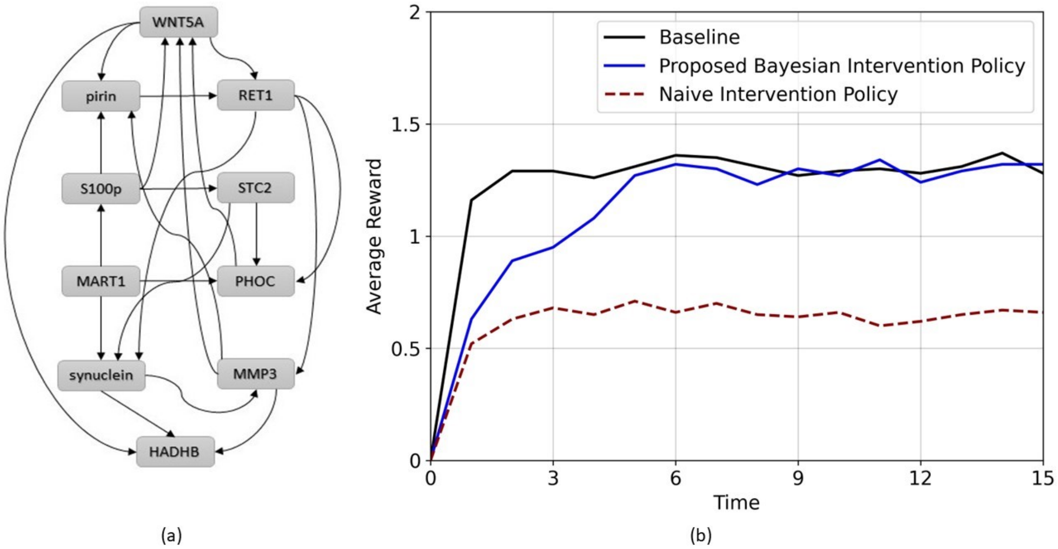


Figure 9: (a) The pathway diagram for the melanoma regulatory network. (b) The average reward gained by the Bayesian intervention policy, naive intervention policy, and the Baseline.

In our experiment, we consider modeling cell responses using single-gene perturbations, which lead to 11 distinct cell actions denoted as $\mathbf{u}^1$ to $\mathbf{u}^{11}$. The action $\mathbf{u}^1$ represents no cell stimuli, and $\mathbf{u}^2$ to $\mathbf{u}^{11}$ correspond to gene 1 to gene 10 stimuli, respectively. Similar to the previous experiment, cell spaces are assumed to contain one, two, or three cell actions, resulting in 231 possible cell spaces. We use the following true (unknown) cell space in our experiment:

$$\mathcal{U}^* = \mathcal{U}^{48} = \{\mathbf{u}^5, \mathbf{u}^8\}, \tag{38}$$

where the cell has the capability to alter the state value of the RET1 or MART1. The intervention space contains three possible actions as $\mathcal{A} = \{\mathbf{a}^1, \mathbf{a}^2, \mathbf{a}^3\}$, where $\mathbf{a}^1$ indicates no intervention, and $\mathbf{a}^2$ and $\mathbf{a}^3$ represent interventions targeting RET1 and PHOC, respectively. All the parameters are the same as in the previous experiment. The initial state is randomly selected from states with activated WNT5A and pirin.

Fig. 9(b) represents the average reward obtained by the proposed Bayesian intervention policy, naive intervention policy, and the Baseline. The average reward achieved by the Bayesian policy gradually converges towards the Baseline after a few steps. In contrast, the naive intervention policy performs poorly, with an average reward of approximately half of the Bayesian policy. This difference highlights the superiority of the Bayesian approach to probabilistically model the cell space and fight back against internal cell responses through stochastic policy.

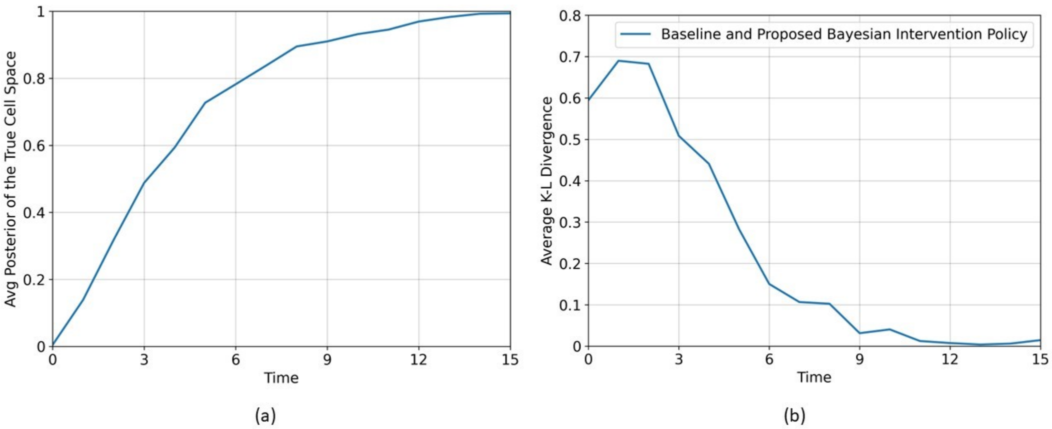


Figure 10: (a) The average posterior of the true cell space over time. (b) The average KL divergence between the true Nash intervention policy and the proposed Bayesian intervention policy.

Fig. 10(a) illustrates the average posterior of the true cell space over time. As can be seen, the true cell space has the largest posterior probability, and its probability approaches 1 after about 15 steps. Furthermore, Fig. 6(b) shows the average KL divergence between the true Nash equilibrium intervention policy and the proposed Bayesian intervention policy. The KL divergence approaching zero indicates the empirical convergence of the Bayesian policy converges to the optimal Nash policy.

8. Conclusion

This paper develops a Bayesian intervention policy for gene regulatory networks (GRNs) that takes into account cell defensive responses. The temporal dynamics of GRNs are modeled using a Boolean network with perturbation (BNp) model, and the interaction between the cell and the intervention is formulated as a two-player zero-sum game. Given incomplete information about cell responses, this paper provides a recursive and probabilistic method to capture the posterior distribution of cell defensive responses. The Bayesian policy is introduced using the combination of the cell-specific Nash policies for each cell space and the posterior distribution associated with them. Our analytical results demonstrate the superiority of the proposed intervention policy against several existing intervention techniques. Meanwhile, the superiority of the proposed intervention policy is demonstrated through comprehensive numerical experiments using the p53-MDM2 negative feedback loop regulatory network and melanoma network.

Our future studies will explore the extension of the proposed game-theoretic intervention policy to practical settings, including studying the partial observability of the genes' state through noisy gene-expression data, as well as addressing scalability issues related to large gene regulatory networks and cell stimuli spaces.

Acknowledgment

The authors acknowledge the support of the National Institute of Health award 1R21EB032480-01, National Science Foundation awards IIS-2311969 and IIS-2202395, ARMY Research Laboratory award W911NF2320179, ARMY Research Office award W911NF2110299, and Office of Naval Research award N00014-23-1-2850.

References

[1] H. Lähdesmäki, I. Shmulevich, O. Yli-Harja, On learning gene regulatory networks under the Boolean network model, Machine learning 52 (2003) 147–167.

[2] A. Paul, J. Sil, Optimized time-lag differential method for constructing gene regulatory network, Information Sciences 478 (2019) 222–238.

[3] **v**Z. Pu**v**snik, M. Mraz, N. Zimic, M. Mo**v**skon, Review and assessment of Boolean approaches for inference of gene regulatory networks, Heliyon (2022).

[4] Z. Zou, H. Chen, P. Poduval, Y. Kim, M. Imani, E. Sadredini, R. Cammarota, M. Imani, BioHD: an efficient genome sequence search platform using hyperdimensional memorization, in: Proceedings of the 49th Annual International Symposium on Computer Architecture, 2022, pp. 656–669.

[5] W.-P. Lee, Y.-T. Hsiao, Inferring gene regulatory networks using a hybrid ga–pso approach with numerical constraints and network decomposition, Information Sciences 188 (2012) 80–99.

[6] M. Alali, M. Imani, Inference of regulatory networks through temporally sparse data, Frontiers in control engineering 3 (2022) 1017256.

[7] E. R. Dougherty, R. Pal, X. Qian, M. L. Bittner, A. Datta, Stationary and structural control in gene regulatory networks: basic concepts, International Journal of Systems Science 41 (2010) 5–16.

[8] A. Yerudkar, E. Chatzaroulas, C. Del Vecchio, S. Moschoyiannis, Sampled-data control of probabilistic Boolean control networks: A deep reinforcement learning approach, Information Sciences 619 (2023) 374–389.

[9] M. Takizawa, K. Kobayashi, Y. Yamashita, Design of reduced-order and pinning controllers for probabilistic Boolean networks using reinforcement learning, Applied Mathematics and Computation 457 (2023) 128211.

[10] S. Dai, B. Li, J. Lu, J. Zhong, Y. Liu, A unified transform method for general robust property of probabilistic Boolean control networks, Applied Mathematics and Computation 457 (2023) 128137.

[11] J. A. Aledo, E. Goles, M. Montalva-Medel, P. Montealegre, J. C. Valverde, Symmetrizable Boolean networks, Information Sciences 626 (2023) 787–804.

[12] A. Ravari, S. F. Ghoreishi, M. Imani, Optimal inference of hidden Markov models through expert-acquired data, IEEE Transactions on Artificial Intelligence (2024).

[13] C. Su, J. Pang, CABEAN: a software for the control of asynchronous Boolean networks, Bioinformatics 37 (2021) 879–881.

[14] L. Van den Broeck, M. Gordon, D. Inzé, C. Williams, R. Sozzani, Gene regulatory network inference: connecting plant biology and mathematical modeling, Frontiers in genetics 11 (2020) 457.

[15] D. Mercatelli, L. Scalambra, L. Triboli, F. Ray, F. M. Giorgi, Gene regulatory network inference resources: A practical overview, Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms 1863 (2020) 194430.

[16] Y. You, Z. Hua, An intelligent intervention strategy for patients to prevent chronic complications based on reinforcement learning, Information Sciences 612 (2022) 1045–1065.

[17] J. Zhong, Y. Liu, J. Lu, W. Gui, Pinning control for stabilization of Boolean networks under knock-out perturbation, IEEE Transactions on Automatic Control 67 (2021) 1550–1557.

[18] S. H. Hosseini, M. Imani, Learning to fight against cell stimuli: A game theoretic perspective, in: 2023 IEEE Conference on Artificial Intelligence (CAI), IEEE, 2023, pp. 285–287.

[19] R. Pal, A. Datta, E. R. Dougherty, Optimal infinite-horizon control for probabilistic Boolean networks, IEEE Transactions on Signal Processing 54 (2006) 2375–2387.

[20] B. Faryabi, J.-F. Chamberland, G. Vahedi, A. Datta, E. R. Dougherty, Optimal intervention in asynchronous genetic regulatory networks, IEEE Journal of Selected Topics in Signal Processing 2 (2008) 412–423.

[21] X. Qian, E. R. Dougherty, Intervention in gene regulatory networks via phenotypically constrained control policies based on long-run behavior, IEEE/ACM Transactions on Computational Biology and Bioinformatics 9 (2011) 123–136.

[22] Q. Liu, Y. He, J. Wang, Optimal control for probabilistic Boolean networks using discrete-time Markov decision processes, Physica A: Statistical Mechanics and its Applications 503 (2018) 1297–1307.

[23] M. Imani, U. M. Braga-Neto, Control of gene regulatory networks using Bayesian inverse reinforcement learning, IEEE/ACM transactions on computational biology and bioinformatics 16 (2018) 1250–1261.

[24] M. Imani, U. M. Braga-Neto, Control of gene regulatory networks with noisy measurements and uncertain inputs, IEEE Transactions on Control of Network Systems 5 (2017) 760–769.

[25] M. Imani, U. M. Braga-Neto, Finite-horizon LQR controller for partially-observed Boolean dynamical systems, Automatica 95 (2018) 172–179.

[26] M. Imani, U. Braga-Neto, Multiple model adaptive controller for partially-observed Boolean dynamical systems, in: 2017 American Control Conference (ACC), IEEE, 2017, pp. 1103–1108.

[27] M. Imani, M. Imani, S. F. Ghoreishi, Optimal Bayesian biomarker selection for gene regulatory networks under regulatory model uncertainty, in: 2022 American Control Conference (ACC), IEEE, 2022, pp. 1379–1385.

[28] M. Imani, U. Braga-Neto, Point-based value iteration for partially-observed Boolean dynamical systems with finite observation space, in: 2016 IEEE 55th Conference on Decision and Control (CDC), IEEE, 2016, pp. 4208–4213.

[29] M. Imani, S. F. Ghoreishi, U. M. Braga-Neto, Bayesian control of large mdps with unknown dynamics in data-poor environments, Advances in neural information processing systems 31 (2018).

[30] M. Imani, U. Braga-Neto, Optimal control of gene regulatory networks with unknown cost function, in: 2018 Annual American Control Conference (ACC), IEEE, 2018, pp. 3939–3944.

[31] I. Shmulevich, E. R. Dougherty, W. Zhang, From Boolean to probabilistic Boolean networks as models of genetic regulatory networks, Proceedings of the IEEE 90 (2002) 1778–1792.

[32] L. E. Chai, S. K. Loh, S. T. Low, M. S. Mohamad, S. Deris, Z. Zakaria, A review on the computational approaches for gene regulatory network construction, Computers in biology and medicine 48 (2014) 55–65.

[33] K. Zhang, Z. Yang, T. Başar, Multi-agent reinforcement learning: A selective overview of theories and algorithms, Handbook of reinforcement learning and control (2021) 321–384.

[34] K. Zhang, S. Kakade, T. Basar, L. Yang, Model-based multi-agent RL in zero-sum Markov games with near-optimal sample complexity, Advances in Neural Information Processing Systems 33 (2020) 1166–1178.

[35] K. Zhang, Z. Yang, T. Basar, Policy optimization provably converges to nash equilibria in zero-sum linear quadratic games, Advances in Neural Information Processing Systems 32 (2019).

[36] I. Bose, B. Ghosh, The p53-mdm2 network: from oscillations to apoptosis, Journal of biosciences 32 (2007) 991–997.

[37] W. Abou-Jaoudé, M. Chaves, J.-L. Gouzé, A theoretical exploration of birhythmicity in the p53-mdm2 network, PLOS one 6 (2011) e17075.

[38] J. S. Chauhan, M. Hölzel, J.-P. Lambert, F. M. Buffa, C. R. Goding, The mitf regulatory network in melanoma, Pigment Cell & Melanoma Research 35 (2022) 517–533.

[39] A. Ravari, S. Ghoreishi, M. Imani, Structure-based inverse reinforcement learning for quantification of biological knowledge, in: IEEE Conference on Artificial Intelligence, 2023.

[40] A. Ravari, S. F. Ghoreishi, M. Imani, Optimal recursive expert-enabled inference in regulatory networks, IEEE Control Systems Letters 7 (2023) 1027–1032.

[41] M. Alali, M. Imani, Reinforcement learning data-acquiring for causal inference of regulatory networks, in: American Control Conference (ACC), IEEE, 2023.

[42] L. S. Shapley, Stochastic games, Proceedings of the national academy of sciences 39 (1953) 1095–1100.

[43] A. Rubinstein, H. W. Kuhn, O. Morgenstern, J. Von Neumann, Theory of Games and Economic Behavior, Princeton university press, 2007.

[44] E. Batchelor, A. Loewer, G. Lahav, The ups and downs of p53: understanding protein dynamics in single cells, Nature Reviews Cancer 9 (2009) 371.

[45] S. Nag, J. Qin, S. KS, M. Wang, R. Zhang, The mdm2-p53 pathway revisited, The Journal of Biomedical Research 27(4) (2013) 254–271.

[46] J. Paluncic, Z. Kovacevic, P. J. Jansson, D. Kalinowski, A. M. Merlot, M. L.-H. Huang, H. C. Lok, S. Sahni, D. J. Lane, D. R. Richardson, Roads to melanoma: Key pathways and emerging players in melanoma progression and oncogenic signaling, Biochimica et Biophysica Acta (BBA) - Molecular Cell Research 1863 (2016) 770–784.

[47] W. Guo, H. Wang, C. Li, Signal pathways of melanoma and targeted therapy, Signal Transduction and Targeted Therapy 6 (2021) 424.