Modeling Defensive Response of Cells to Therapies: Equilibrium Interventions for Regulatory Networks

Seyed Hamid Hosseini and Mahdi Imani

***Abstract*—A major objective in genomics is to design interventions that can shift undesirable behaviors of such systems (i.e., those associated with cancers) into desirable ones. Several intervention policies have been developed in recent years, including dynamic and structural interventions. These techniques aim at making targeted changes to cell dynamics upon intervention, without considering the cell's defensive mechanisms to interventions. This simplified assumption often leads to early and short-term success of interventions, followed by partial or full recurrence of diseases. This is due to the fact that cells often have dynamic and intelligent responses to interventions through internal stimuli. This paper models gene regulatory networks (GRNs) using the Boolean network with perturbation. The dynamic and adaptive battle between intervention and the cell is modeled as a two-player zero-sum game, where intervention and the cell fight against each other with fully opposite objectives. An optimal intervention policy is obtained as a Nash equilibrium solution, through which the intervention is stochastic, ensuring the optimal solution to all potential cell responses. We analytically analyze the superiority of the proposed intervention policy against existing intervention techniques. Comprehensive numerical experiments using the p53-MDM2 negative feedback loop regulatory network and melanoma network demonstrate the high performance of the proposed method.**

***Index Terms*—Gene Regulatory Networks, Biological Interventions, Nash Equilibrium, Dynamic Programming.**

I. Introduction

Recent technological advancements in genomics have significantly enhanced our understanding of these complex biological systems. Gene regulatory networks (GRNs) are comprised of a number of interacting genes whose interactions control the ecosystem functioning and various cellular processes, such as stress response, DNA repair, and other mechanisms involved in complex diseases such as cancer [1–4]. A major goal in genomics is to find intervention strategies to alter undesirable behavior of these systems, such as those associated with chronic diseases.

Several intervention strategies have been developed for the systematic intervention of GRNs. These include dynamic perturbations [5–10], which aim at providing time-dependent intervention solutions, and structural interventions [7, 11–15], which aim at making a single-time change in interactions between the genes to shift their dynamics properly. Most existing intervention methods assume that cells are isolated with no defensive response to selected interventions. However, cells have highly complex, dynamic, and robust responses to abnormality, stress, or external therapies. This is achieved through internal stimuli controlled by cells, which protect us against diseases and ensure the proper functioning of cells by controlling gene activities and protein production. For unhealthy cells, such as those associated with autoimmune diseases (e.g., cancer), the cell's defense mechanisms fight against itself, which often leads to the uncontrolled proliferation of cells. To create effective therapeutic solutions, we need to take into account the defense mechanisms of cells against interventions and their dynamic responses and willingness to recur in cancerous conditions.

This paper models GRNs using the well-known Boolean network with perturbation (BNp) model [2, 16]. The Boolean networks have shown tremendous success in representing the causal relationship among genes as well as designing targeted therapies to alter the system behavior. This paper models the dynamic and intelligent defensive response of cells to therapies using a two-player zero-sum game. A non-cooperative game is previously considered for genomics interventions in [17], where a commonly centralized intervention is extended to multiple interventions/therapies without accounting for cell dynamic response. By contrast, this paper considers the cell and intervention as two players with opposite goals; the cell aims at keeping the cell in an unhealthy condition using its internal stimuli, whereas the interventionist aims at properly selecting drugs/therapies that deviate the system from unhealthy states. Therefore, success for one of the cells and interventionists is a failure in achieving the other player's goal.

This paper derives optimal infinite-horizon Nash equilibrium intervention policy for GRNs with known system dynamics. We show that there is an equilibrium policy for an interventionist, where any deviation from that leads to more recurrence of disease and a better cell defensive response. Unlike most available deterministic intervention policies, the proposed equilibrium policy is stochastic and guarantees to achieve the best therapeutic solutions under the most aggressive response of the cells. The optimal Nash equilibrium policy is computed using the min-max theorem and dynamic programming technique. We analytically analyze the difference between the proposed equilibrium policy and state-of-art intervention methods and describe how the equilibrium solution can help design interventions and analyze the long-term impact of therapies on the treatment. The high performance of the proposed framework in terms of intervention performance and robustness is demonstrated through comprehensive numerical experiments using a p53-MDM2 negative feedback loop network and melanoma cell cycle network. Our future work includes studying finite horizon cases where interventions are

S. H. Hosseini and M. Imani are with the Department of Electrical and Computer Engineering at Northeastern University. Emails: hosseini.ha@northeastern.edu, m.imani@northeastern.edu

conducted over a fixed period of time, as well as GRNs with partially known or unknown regulatory networks.

The article is organized as follows. Section II describes the regulatory network model. Section III includes formulating the intervention process as a two-player zero-sum game, followed by developing an algorithm using dynamic programming for finding the optimal Nash equilibrium intervention policy. The analysis of performance and complexity is presented in Section IV. Finally, Section V presents results for the numerical experiments and Section VI contains the concluding remarks.

II. Background - Regulatory Network Model

This paper employs a Boolean network with perturbation (BNp) model [16] for capturing the dynamics of gene regulatory networks. This model properly captures the stochasticity in GRNs, coming from intrinsic uncertainty or unmodeled parts of systems. Consider a system consisting of d components. The *state process* can be expressed as $\{\mathbf{x}_t; t = 0, 1, \ldots\}$, where $\mathbf{x}_t \in \{0,1\}^d$ represents the activation/inactivation state of the genes at time t. The state of the genes is affected by a sequence of internal and external inputs. The genes state are updated at each discrete time through the following Boolean signal model:

$$\mathbf{x}_t \;=\; \mathbf{f}(\mathbf{x}_{t-1}) \,\oplus\, \mathbf{a}_{t-1} \,\oplus\, \mathbf{u}_{t-1} \,\oplus\, \mathbf{n}_t\,, \tag{1}$$

for $t = 1, 2, \ldots$, where $\{\mathbf{a}_t; t = 0, 1, ...\}$ is an external set of interventions/therapies, $\{\mathbf{u}_t; t = 0, 1, ...\}$ is an internal inputs controlled by the cell, $\mathbf{n}_t \in \{0,1\}^d$ is a Boolean transition noise at time t, "$\oplus$" indicates component-wise module-2 addition, and $\mathbf{f}$ is the *network function*. Therefore, if $\mathbf{n}_t(j) = 1$, the state of the jth gene at time step t is flipped; otherwise, it is determined by the network function. The noise process $\mathbf{n}_t$ is assumed to have independent components distributed as Bernoulli(p), where parameter $p > 0$ corresponds to the amount of "perturbation" to the Boolean state process. Larger values of p correspond to more chaotic systems, whereas small values of p models are nearly deterministic models. It should be noted that without loss of generality, the rest of paper holds for a general class Boolean network models and more complex network function of form $\mathbf{f}(\mathbf{x}_{t-1}, \mathbf{a}_{t-1}, \mathbf{u}_{t-1}, \mathbf{n}_t)$.

The network function in GRNs is often expressed through either a Boolean logic model or a pathway diagram model [7, 18, 19]. The logic model represents the complex relationships between genes using operators such as AND, OR, XOR, and NOT, while the pathway diagram model parameterizes the suppressive and activating interactions between the elements to represent the system dynamics. These two models have been successful in representing temporal changes in gene activities and capturing complex relationships among genes.

III. Proposed Intervention Policy

In unhealthy or cancerous cells (e.g., those associated with autoimmune diseases), cells continuously fight against themselves through internal stimuli. This self-harm is often due to mutations, stress, or other unknown factors and, in most cases, leads to changes in gene activity and excessive proliferation of cancerous cells. Since cells contain complex and dynamic defense mechanisms, fighting against them through intervention is extremely challenging. In practice, most existing interventions provide only a short-term reduction in cancerous cell proliferation, as the cells find new ways to fight against the interventions and proliferation recurs. This paper models the defense mechanism of cells and their dynamic response to therapies, and develops an optimal intervention policy given the cell's defensive response. In this section, we first outline the model for representing the battle between the intervention and the cell, followed by developing the optimal intervention policy.

A. Battle of Cell and Interventionist - Two-Player Zero-Sum Game

This paper models the battle between the cell and interventionist, as a two-player zero-sum game [20–22]. 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 *action (e.g., intervention) space*, $\mathcal{U}$ is the *internal cell control (i.e., internal stimuli) space*, $T : \mathcal{X} \times \mathcal{A} \times \mathcal{U} \times \mathcal{X}$ is the *state transition probability function* 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 $\mathbf{x}$. $R^a : \mathcal{X} \times \mathcal{A} \times \mathcal{U} \times \mathcal{X}$ denotes the reward functions for an interventionist, where $R^a(\mathbf{x}, \mathbf{a}, \mathbf{u}, \mathbf{x}')$ denotes the immediate shift from the cancerous states (i.e., reduction in cell proliferation) if the system moves from state $\mathbf{x}$ to state $\mathbf{x}'$ after the intervention $\mathbf{a}$ and the internal cell response $\mathbf{u}$. The cell aims at increasing cell proliferation in cancerous cells, while the interventionist aims at reducing cell proliferation. Thus, the the reward for the cell R^u takes negative of the interventionist reward function, i.e., $R^u(\mathbf{x}, \mathbf{a}, \mathbf{u}, \mathbf{x}') = -R^a(\mathbf{x}, \mathbf{a}, \mathbf{u}, \mathbf{x}')$.

B. Optimal Nash Equilibrium Policy

This paper focuses on stationary Markov Nash equilibria in GRNs modeled by the infinite-horizon discounted Markov game. Let $\pi^a(\mathbf{a} \mid \mathbf{x})$ denote the stationary (Markov) intervention strategy, which specifies the probability of action $\mathbf{a} \in \mathcal{A}$ in any given state $\mathbf{x} \in \mathcal{X}$. Let also $\pi^u(\mathbf{u} \mid \mathbf{x})$ be the cell policy, specifying the probability of input $\mathbf{u} \in \mathcal{U}$ in state $\mathbf{x} \in \mathcal{X}$. We define the expected value function of interventionist and cell under the joint stochastic policy (π^a, π^u) as:

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

for $\mathbf{x} \in \mathcal{X}$, where $0 < \gamma < 1$ is a discount factor that indicates the importance of early-stage rewards compared to future ones, $\mathbf{a}_{0:\infty} = \{\mathbf{a}_0, \mathbf{a}_1, \ldots, \mathbf{a}_\infty\}$ and $\mathbf{u}_{0:\infty} = \{\mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_\infty\}$ are the sequence of interventions and cell stimuli over an infinite horizon (i.e., from the initial time step 0 to infinity). It can be seen in (2) that the state values for the cell and interventionist

are intertwined. In fact, the solution of a Markov game is different from a Markov Decision Process (MDP) since the optimal performance of each agent is controlled not only by its own policy but also by the choices of both cells and interventionists in the game. Using the fact that cell and interventionist 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 $\mathbf{x} \in \mathcal{X}$. The joint optimal policy $\pi^* = (\pi^a_*, \pi^u_*)$ is a Nash equilibrium policy, which satisfies [23]

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

for any $\pi = (\pi^a, \pi^u)$ and $\mathbf{x} \in \mathcal{X}$. The optimal Nash equilibrium policy is the policy that the cell and interventionists do not have any incentive to deviate from their policies. This policy can be obtained using the min-max theorem in matrix form zero-sum games [24]. Let $\mathbf{V}^a_{\pi^a,\pi^u} = [V^a_{\pi^a,\pi^u}(\mathbf{x}^0), \ldots, V^a_{\pi^a,\pi^u}(\mathbf{x}^{2^d})]^T$ be the vector-form of the state value function associated with policy (π^a, π^u). One can define the optimal Nash equilibrium policy as:

$$(\pi^a_*, \pi^u_*) = \underset{\pi^a}{\operatorname{argmax}} \underset{\pi^u}{\operatorname{argmin}} \mathbf{V}^a_{\pi^a,\pi^u} = \underset{\pi^u}{\operatorname{argmin}} \underset{\pi^a}{\operatorname{argmax}} \mathbf{V}^a_{\pi^a,\pi^u}. \quad (4)$$

According to (3), any pair of (π^a, π^u) that attains the supremum and infimum in (4) constitutes a Nash equilibrium. It is impossible or computationally challenging to go through all possible policies to find the optimal policy in (4). We define the s-simplex Δ_s as:

$$\Delta_s = \{[v_1, ..., v_s] \in \mathbb{R}^s : v_1 + v_2 + ... + v_s = 1, v_i \geq 0\},$$

for $i = 1, 2, ..., s$. The search space for policy π^a contains 2^d simplexes of size $|\mathcal{A}|$, meaning $2^d \times \Delta_{|\mathcal{A}|}$. Similarly, the search space for the policy of the cell is $2^d \times \Delta_{|\mathcal{U}|}$, which is also a large continuous space. In the following paragraphs, we describe a dynamic programming approach to find the optimal Nash equilibrium policy in a two-player zero-sum game, which resembles the dynamic programming solution for MDPs.

For any state value function $\mathbf{V} : \mathcal{X} \to \mathbb{R}$, we define the state joint actions value function for the interventionist as:

$$Q^a_{\mathbf{V}}(\mathbf{x}, \mathbf{a}, \mathbf{u}) = \mathbb{E}_{\mathbf{x}'|\mathbf{x},\mathbf{a},\mathbf{u}}\left[R^a(\mathbf{x}, \mathbf{a}, \mathbf{u}, \mathbf{x}') + \gamma \mathbf{V}(\mathbf{x}')\right], \quad (5)$$

for $\mathbf{x} \in \mathcal{X}, \mathbf{a} \in \mathcal{A}$ and $\mathbf{u} \in \mathcal{U}$, where $Q_{\mathbf{V}}(\mathbf{x}, ., .)$ can be regarded as a matrix in $\mathbb{R}^{|\mathcal{A}|\times|\mathcal{U}|}$, and the expectation with respect to the next system state. The Q-value specifies the expected reward for the interventionist if the joint actions $(\mathbf{a}, \mathbf{u})$ are selected at state $\mathbf{x}$ and the policy associated with the state value function $\mathbf{V}$ is followed afterward.

We define the Bellman operator $\mathcal{T}^*$ by solving a matrix form zero-sum game for $Q_{\mathbf{V}}(\mathbf{x}, ., .)$ as the payoff matrix, i.e., for any $\mathbf{x} \in \mathcal{X}$, one can define

$$\begin{aligned}(\mathcal{T}^*\mathbf{V})(\mathbf{x}) &= \text{Value}[Q^a_{\mathbf{V}}(\mathbf{x}, ., .)] \\ &= \max_{\pi^a} \min_{\pi^u} \sum_{i=1}^{|\mathcal{A}|} \sum_{j=1}^{|\mathcal{U}|} \pi^a(\mathbf{a}^i|\mathbf{x})\, \pi^u(\mathbf{u}^j|\mathbf{x})\, Q^a_{\mathbf{V}}(\mathbf{x}, \mathbf{a}^i, \mathbf{u}^j).\end{aligned} \quad (6)$$

The Bellman operator defined in (6) consists of Value$[Q_{\mathbf{V}}(\mathbf{x}, ., .)]$ and can be computed using linear programming techniques [25–27].

The Bellman operator $\mathcal{T}^*$ is γ-contractive in the L-norm and the unique solution to the Bellman equation corresponds to the optimal value function, i.e., $\mathbf{V}^* = \mathcal{T}^*\mathbf{V}^*$ [28, 29]. The value iteration algorithm, described later in this section, provides a recursive procedure to find the fixed-point solution of the Bellman operator in (6).

We define the *joint action transition matrix* of size $2^d \times 2^d$ associated with actions $(\mathbf{a}, \mathbf{u})$ as:

$$\begin{aligned}(M(\mathbf{a}, \mathbf{u}))_{ij} &= P\left(\mathbf{x}_t = \mathbf{x}^j \mid \mathbf{x}_{t-1} = \mathbf{x}^i, \mathbf{a}_{t-1} = \mathbf{a}, \mathbf{u}_{t-1} = \mathbf{u}\right) \\ &= p^{||\mathbf{f}(\mathbf{x}^i) \oplus \mathbf{a} \oplus \mathbf{u} \oplus \mathbf{x}^j||_1} (1-p)^{d - ||\mathbf{f}(\mathbf{x}^i) \oplus \mathbf{a} \oplus \mathbf{u} \oplus \mathbf{x}^j||_1},\end{aligned} \quad (7)$$

for $i, j = 1, \ldots, 2^d$, where $P(.)$ is the probability mass function, p is the Bernoulli noise parameter, $||.||_1$ is the absolute L-1 norm of a vector, $\mathbf{f}(\mathbf{x}^i) \oplus \mathbf{a} \oplus \mathbf{u}$ is the noise-free predictive state of genes in the next time step, and $||\mathbf{f}(\mathbf{x}^i) \oplus \mathbf{a} \oplus \mathbf{u} \oplus \mathbf{x}^j||_1$ measures the number of flips caused by noise if the system moves from state $\mathbf{x}^i$ to state $\mathbf{x}^j$. The computation of the transition matrix in (7) requires full knowledge of the system dynamics (i.e., network function) and the intensity of the Bernoulli noise, p. It should be noted the cell and intervention spaces, as well as the temporal changes in genes' activities, are all accounted for in the computation of the transition matrices corresponding to each pair of $(\mathbf{a}, \mathbf{u}) \in \mathbb{A} \times \mathcal{U}$.

Let $\mathbf{R}^a(\mathbf{a}, \mathbf{u})$ be a matrix-form of the interventionist reward function associated with control inputs $\mathbf{a}$ and $\mathbf{u}$ expressed as:

$$(\mathbf{R}^a(\mathbf{a}, \mathbf{u}))_{ij} = R^a\left(\mathbf{x}^i, \mathbf{a}, \mathbf{u}, \mathbf{x}^j\right), \text{ for } i, j = 1, ..., 2^d. \quad (8)$$

Meanwhile, for joint action $(\mathbf{a}, \mathbf{u})$, we define the expected interventionist reward function, which can be expressed in a vectored form as $R^a_{\mathbf{a},\mathbf{u}} = [R^a(\mathbf{x}^1, \mathbf{a}, \mathbf{u}), \cdots, R^a(\mathbf{x}^{2^d}, \mathbf{a}, \mathbf{u})]^T$. The ith row of this vector can be calculated according to the controlled transition matrix in (7) and the matrix-form reward function in (8) as:

$$\begin{aligned}R^a(\mathbf{x}^i, \mathbf{a}, \mathbf{u}) &= \mathbb{E}_{\mathbf{x}'|\mathbf{x}^i,\mathbf{a},\mathbf{u}}[R^a(\mathbf{x}^i, \mathbf{a}, \mathbf{u}, \mathbf{x}')] \\ &= \sum_{j=1}^{2^d} R^a(\mathbf{x}^i, \mathbf{a}, \mathbf{u}, \mathbf{x}^j) \\ &\qquad \times P(\mathbf{x}_t = \mathbf{x}^j \mid \mathbf{x}_{t-1} = \mathbf{x}^i, \mathbf{a}_{t-1} = \mathbf{a}, \mathbf{u}_{t-1} = \mathbf{u}) \\ &= \sum_{j=1}^{2^d} (\mathbf{R}^a(\mathbf{a}, \mathbf{u}))_{ij}\, (M(\mathbf{a}, \mathbf{u}))_{ij},\end{aligned}$$

for $i = 1, .., 2^d$. The vector-form of expected interventionist reward function, $R^a_{\mathbf{a},\mathbf{u}}$, can be computed in a compact form as:

$$R^a_{\mathbf{a},\mathbf{u}} = (\mathbf{R}^a(\mathbf{a}, \mathbf{u}) \odot M(\mathbf{a}, \mathbf{u}))\, \mathbf{1}_{2^d \times 1}, \quad (9)$$

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

Using the controlled transition matrix and the vector-form reward function, the Q-values defined in (5) for any given state value function $\mathbf{V}$ can be calculated as:

$$\begin{bmatrix} Q^a_{\mathbf{V}}(\mathbf{x}^1, \mathbf{a}, \mathbf{u}) \\ \vdots \\ Q^a_{\mathbf{V}}(\mathbf{x}^{2^d}, \mathbf{a}, \mathbf{u}) \end{bmatrix} = R^a_{\mathbf{a},\mathbf{u}} + \gamma M(\mathbf{a}, \mathbf{u})\, \mathbf{V}, \quad (10)$$

for $\mathbf{a} \in \mathcal{A}$, and $\mathbf{u} \in \mathcal{U}$.

Algorithm 1 Optimal Nash Equilibrium Intervention Policy

Intervention reward function $R^a(\mathbf{x}^i, \mathbf{a}, \mathbf{u}, \mathbf{x}^j)$, controlled transition matrix $M(\mathbf{a}, \mathbf{u})$ for $\mathbf{a} \in \mathcal{A}, \mathbf{u} \in \mathcal{U}$, threshold $\epsilon > 0$.

Matrix-form intervention reward function: $(\mathbf{R}^a(\mathbf{a}, \mathbf{u}))_{ij} = R^a(\mathbf{x}^i, \mathbf{a}, \mathbf{u}, \mathbf{x}^j)$, for $\mathbf{a} \in \mathcal{A}, \mathbf{u} \in \mathcal{U}$, and $i, j = 1, ..., 2^d$.

Vector-form intervention reward function: $R^a_{\mathbf{a},\mathbf{u}} = (\mathbf{R}^a(\mathbf{a}, \mathbf{u}) \odot M(\mathbf{a}, \mathbf{u}))\, \mathbf{1}_{2^d \times 1}$, for $\mathbf{a} \in \mathcal{A}, \mathbf{u} \in \mathcal{U}$.

Set $\mathbf{V}' = \mathbf{0}_{2^d \times 1}$.

repeat

$\mathbf{V} = \mathbf{V}'$.

$$\begin{bmatrix} Q^a_{\mathbf{V}}(\mathbf{x}^1, \mathbf{a}, \mathbf{u}) \\ \vdots \\ Q^a_{\mathbf{V}}(\mathbf{x}^{2^d}, \mathbf{a}, \mathbf{u}) \end{bmatrix} = \left[R^a_{\mathbf{a},\mathbf{u}} + \gamma M(\mathbf{a}, \mathbf{u})\, \mathbf{V}\right], \text{ for } \mathbf{a} \in \mathcal{A} \text{ and } \mathbf{u} \in \mathcal{U}.$$

Bellman Operator: $\mathbf{V}'(\mathbf{x}^i) = \text{Value}[Q^a_{\mathbf{V}}(\mathbf{x}^i, ., .)]$, for $i = 1, ..., 2^d$ — Eq. (4)

until $\max_{i \in \{1,...,2^d\}} |\mathbf{V}(i) - \mathbf{V}'(i)| < \epsilon$

$\mathbf{V}^* = \mathbf{V}'$.

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

For any given $\mathbf{x} \in \mathcal{X}$, use linear programming approach over $Q^a_{\mathbf{V}^*}(\mathbf{x}, ., .)$ to obtain $\pi^a_*(.|\mathbf{x}^i)$ and $\pi^u_*(.|\mathbf{x}^i)$.

Given that the Bellman operator is a γ-contraction mapping for the Markov game, one can start from any arbitrary $\mathbf{V}$ and iteratively apply $\mathbf{V}_{t+1} = \mathcal{T}^*[\mathbf{V}_t]$ for $t = 0, 1, ...$ until a fixed-point solution is contained. The fixed-point solution is an optimal Nash equilibrium solution for the Markov game. Let $\mathbf{V}_0 = [0, \cdots, 0]^T$ be the initial value vector with all elements 0. At iteration r of value iteration, one needs to find $\mathbf{V}_{r+1}$ from $\mathbf{V}_r$ as:

$$\mathbf{V}_{r+1}(\mathbf{x}^i) = \text{Value}[Q^a_{\mathbf{V}_r}(\mathbf{x}^i, ., .)], \text{ for } i = 1, ..., 2^d, \quad (11)$$

where $Q_{\mathbf{V}}(\mathbf{x}, ., .)$ consists of Q-values at state $\mathbf{x}$ and all joint pairs of $(\mathbf{a}, \mathbf{u})$. In practice, the iterations continue till the time that the maximum difference between elements of value vectors in two consecutive iterations falls below a small pre-specified threshold, i.e., $\max_{i \in \{1,..,2^d\}} |\mathbf{V}_T(i) - \mathbf{V}_{T-1}(i)| < \epsilon$.

Let $\mathbf{V}^*$ be the fixed-point solution obtained by the value iteration method. One can compute $Q^a_{\mathbf{V}^*}(., ., .)$ as:

$$\begin{bmatrix} Q^a_{\mathbf{V}^*}(\mathbf{x}^1, \mathbf{a}, \mathbf{u}) \\ \vdots \\ Q^a_{\mathbf{V}^*}(\mathbf{x}^{2^d}, \mathbf{a}, \mathbf{u}) \end{bmatrix} = R^a_{\mathbf{a},\mathbf{u}} + \gamma M(\mathbf{a}, \mathbf{u})\, \mathbf{V}^*, \quad (12)$$

for $\mathbf{a} \in \mathcal{A}$ and $\mathbf{u} \in \mathcal{U}$.

The optimal policy for cell and interventionist can be represented as a matrix saddle point problem involving the following matrix for any $\mathbf{x} \in \mathcal{X}$:

$$\pi^*(.|\mathbf{x}) = (\pi^a_*(.|\mathbf{x}), \pi^u_*(.|\mathbf{x})) = \underset{\pi^a}{\text{argmax}}\, \underset{\pi^u}{\text{argmin}} \sum_{i=1}^{|\mathcal{A}|} \sum_{j=1}^{|\mathcal{U}|} \pi^a(\mathbf{a}^i|\mathbf{x})\, \pi^u(\mathbf{u}^j|\mathbf{x})\, Q^a_{\mathbf{V}^*}(\mathbf{x}, \mathbf{a}^i, \mathbf{u}^j), \quad (13)$$

where $\pi^a(.|\mathbf{x}) \geq 0$, $\pi^u(.|\mathbf{x}) \geq 0$, and $\sum_{i=1}^{|\mathcal{A}|} \pi^a(\mathbf{a}^i|\mathbf{x}) = 1$ and $\sum_{j=1}^{|\mathcal{U}|} \pi^u(\mathbf{u}^j|\mathbf{x}) = 1$. The optimal policy can be easily obtained using linear programming techniques for matrix $Q^a_{\mathbf{V}^*}(\mathbf{x}, ., .)$, for $\mathbf{x} \in \mathcal{X}$. The existence of a Nash Equilibrium in a two-player zero-sum game is guaranteed by the Minimax theorem [30]. The entire process of the proposed Nash equilibrium intervention policy is provided in Algorithm 1. The complexity of each iteration of Algorithm 1 is $O(2^{2d} \times \mathcal{A} \times \mathcal{U})$, since for each state, all possible joint cell and intervention actions need to be considered.

The detailed iterative steps of computation of the optimal Nash equilibrium intervention policy are provided in Algorithm 1. The process requires a full knowledge of the transition probabilities, which means the knowledge about the system dynamics and Bernoulli process noise, representing the system's stochasticity. The process resembles the dynamic programming for MDP, with the key difference that the actions are defined as joint intervention and cell response, and the bellman operator is applied iteratively as the minimax operator in (6), which can be achieved through linear programming.

To better understand the optimal Nash equilibrium policy, consider the following simple example consisting of a single component:

$$\mathbf{x}_t = \overline{\mathbf{x}_{t-1}} \oplus \mathbf{a}_{t-1} \oplus \mathbf{u}_{t-1}, \quad (14)$$

where the state transition is deterministic. The state space is $\mathcal{X} = \{\mathbf{x}^1 = 0, \mathbf{x}^2 = 1\}$, the intervention space is $\mathcal{A} = \{\mathbf{a}^1 = 0, \mathbf{a}^2 = 1\}$, and the cell action space is $\mathcal{U} = \{\mathbf{u}^1 = 0, \mathbf{u}^2 = 1\}$. The cell aims at keeping the gene state at $\mathbf{x} = 1$, whereas the interventionist aims at reducing the activation of the gene (i.e., keeping the gene at state $\mathbf{x} = 0$). This can be reflected in the following interventionist reward functions: $R^a(\mathbf{x}, \mathbf{a}, \mathbf{u}, \mathbf{x}' = 0) = 1$ and $R^a(\mathbf{x}, \mathbf{a}, \mathbf{u}, \mathbf{x}' = 1) = -1$. Under no intervention, the cell can keep the gene at $\mathbf{x} = 1$ using positive and no stimuli at state $\mathbf{x} = 1$ and $\mathbf{x} = 0$, respectively. Also, under no cell action,

the interventionist can keep the gene at $\mathbf{x}=0$ using no and positive intervention at state $\mathbf{x}=1$ and $\mathbf{x}=0$, respectively. However, when there is both intervention and cell action, neither the interventionist nor the cell can always keep the gene at its desired state using a deterministic policy. The best choice for both of them is to use the Nash equilibrium policy, which in this scenario is choosing random actions. Despite its randomness, any deviation from this Nash policy by the interventionist will help the cell to gain more and more returns of the undesirable condition.

To enable the implementation of the proposed intervention policies in real biological settings, future studies should address several key practical considerations during genomics intervention. Firstly, the proposed strategy relies on direct access to all genes' states, which, in practice, are only partially observable through noisy gene expression data. Meanwhile, measuring cell stimuli may not be possible in practice, and limited knowledge might be available about the possible cell defensive responses. Furthermore, the complexity of genomics systems often introduces uncertainty in the pathways of gene regulatory networks. Finally, the scalability of such an approach to large gene regulatory networks is crucial in real-world implementations of these policies. By systematically addressing these challenges, future studies can pave the way for the practical implementation of such stochastic intervention policies in real experimental settings.

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

This section analyzes the performance of the proposed Nash equilibrium intervention policy with conventional intervention policies and the system without interventions. For the system with no intervention (i.e., $\mathcal{A}=\{\mathbf{0}\}$), the cell can aggressively push the system to the undesirable states. The cell policy in this case can be expressed as:

$$\vartheta_{a=0}(\mathbf{x})=\underset{\mathbf{u}\in\mathcal{U}}{\operatorname{argmin}}[R^a_{\mathbf{a}=\mathbf{0},\mathbf{u}}+\gamma M(\mathbf{a}=\mathbf{0},\mathbf{u})\mathbf{V}_{\mathbf{a}=\mathbf{0}}],\ \text{for } \mathbf{x}\in\mathcal{X}. \tag{15}$$

where $\mathbf{V}_{\mathbf{a}=\mathbf{0}}=\min_{\mathbf{u}\in\mathcal{U}}\left[R^a_{\mathbf{a}=\mathbf{0},\mathbf{u}}+\gamma M(\mathbf{a}=\mathbf{0},\mathbf{u})\,\mathbf{V}_{\mathbf{a}=\mathbf{0}}\right]$.

Most existing intervention strategies are deterministic, meaning they assume the cell has no defense mechanism against the interventions [5–10, 31–34]. This assumption allows the agent to naively decide about the deterministic interventions at various states. In this case, the Markov game is modeled through the Markov decision process with a single agent/player. The MDP can be defined as $\langle\mathcal{X},\mathcal{A},\mathcal{U}=\{\mathbf{0}\},\tilde{T},R^a\rangle$, where transition probability $\tilde{T}: p(\mathbf{x}'\mid\mathbf{x},\mathbf{a},\mathbf{u}=\mathbf{0})$ represent the probability of next state given the intervention $\mathbf{a}$ and no cell response $\mathbf{u}=\mathbf{0}$.

The control transition matrices perceived by the interventionist for the MDP are $(M(\mathbf{a},\mathbf{u}=\mathbf{0})_{ij}=p(\mathbf{x}'=\mathbf{x}^j\mid\mathbf{x}=\mathbf{x}^i,\mathbf{a},\mathbf{u}=\mathbf{0})$, for $i,j=1,...,2^d$ and $\mathbf{a}\in\mathcal{A}$. The immediate reward under this naive cell response assumption can also be represented by $R^a(\mathbf{x},\mathbf{a},\mathbf{u}=\mathbf{0},\mathbf{x}')$. We put the intervention reward into a matrix $(R^a(\mathbf{a},\mathbf{u}=\mathbf{0}))_{ij}=R^a(\mathbf{x}^i,\mathbf{a},\mathbf{u}=\mathbf{0},\mathbf{x}^j)$, for $i,j=1,...,2^d$ and $\mathbf{a}\in\mathcal{A}$. We define $R^a_{\mathbf{a},\mathbf{u}=\mathbf{0}}=(\mathbf{R}^a(\mathbf{a},\mathbf{u}=\mathbf{0})\odot M(\mathbf{a},\mathbf{u}=\mathbf{0}))\,\mathbf{1}_{2^d\times 1}$, for $\mathbf{a}\in\mathcal{A}$. Since the interventionist perceives itself as the only agent/player, the best policy for the interventionist is deterministic and it is the fixed-point solution of the following Bellman optimal equations:

$$\mathbf{V}_{\mathbf{u}=\mathbf{0}}=\max_{\mathbf{a}\in\mathcal{A}}\left[R^a_{\mathbf{a},\mathbf{u}=\mathbf{0}}+\gamma M(\mathbf{a},\mathbf{u}=\mathbf{0})\,\mathbf{V}_{\mathbf{u}=\mathbf{0}}\right], \tag{16}$$

where the maximum is applied row-wise. The fixed-point solution $\tilde{\mathbf{V}}=\mathbf{0}$ in (16) can be obtained using the Value Iteration algorithm by iteratively applying the Bellman operator starting from an arbitrary initial state vector. Upon computation of the fixed point solution $\mathbf{V}_{\mathbf{u}=\mathbf{0}}$, the naive intervention policy can be obtained as:

$$\mu_a(\mathbf{x})=\underset{\mathbf{a}\in\mathcal{A}}{\operatorname{argmax}}[R^a_{\mathbf{a},\mathbf{u}=\mathbf{0}}+\gamma M(\mathbf{a},\mathbf{u}=\mathbf{0})\mathbf{V}_{\mathbf{u}=\mathbf{0}}],\ \text{for } \mathbf{x}\in\mathcal{X}. \tag{17}$$

Given that μ_a is the naive interventionist policy under the non-defensive cell, we aim to analyze how the cell responds to this naive policy. This analysis provides the rationale behind the early success of the conventional intervention, followed by cell domination and recurrence of the disease in the long term. The best defense policy for the cell against a known and deterministic intervention policy μ_a in (17) can be formulated as:

$$\mu_u(\mathbf{x})=\underset{\mu}{\operatorname{argmin}}\,\mathbb{E}\Bigg[\sum_{t=0}^{\infty}\gamma^t R^a(\mathbf{x}_t,\mathbf{a}_t,\mathbf{u}_t,\mathbf{x}_{t+1})\mid\mathbf{x}_0=\mathbf{x},\ \mathbf{a}_{0:\infty}\sim\mu_a,\mathbf{u}_{0:\infty}\sim\mu\Bigg], \tag{18}$$

where the minimization occurs across the entire cell policy space, which has a cardinality of $|\mathcal{U}|^{2^d}$. Determining μ_u by exhaustively considering all potential policies is computationally demanding. However, it can be accomplished using the dynamic programming approach outlined below, which provides an efficient solution.

Let $\mathbf{R}^a(\mu_{\mathbf{a}},\mathbf{u})$ be a matrix of size $2^d\times 2^d$ associated with intervention policy μ_a and cell control input $\mathbf{u}$ with the element in the ith row and $j$the column as $(\mathbf{R}^a(\mu_a,\mathbf{u}))_{ij}=R^a(\mathbf{x}^i,\mu_a(\mathbf{x}^j),\mathbf{u},\mathbf{x}^j)$. The control transition matrix under intervention policy $\mathbf{a}\sim\mu_a$ can be expressed as:

$$(M(\mu_a,\mathbf{u}))_{ij}\ =\ P\left(\mathbf{x}_t=\mathbf{x}^j\mid\mathbf{x}_{t-1}=\mathbf{x}^i,\ \mathbf{a}_{t-1}=\mu_a(\mathbf{x}^i),\mathbf{u}_{t-1}=\mathbf{u}\right), \tag{19}$$

for $i,j=1,...,2^d$ and $\mathbf{u}\in\mathcal{U}$. Using (9), the vector-form reward function under cell action $\mathbf{u}$ and intervention policy μ_a can be expressed as:

$$\begin{aligned}R^a_{\mu_a,\mathbf{u}}&=\begin{bmatrix}R^a(\mathbf{x}^1,\mu_a(\mathbf{x}^1),\mathbf{u})\\ \vdots\\ R^a(\mathbf{x}^{2^d},\mu_a(\mathbf{x}^{2^d}),\mathbf{u})\end{bmatrix}\\&=(\mathbf{R}^a(\mu_a,\mathbf{u})\odot M(\mu_a,\mathbf{u}))\,\mathbf{1}_{2^d\times 1},\end{aligned} \tag{20}$$

for $\mathbf{u}\in\mathcal{U}$. The optimal cell policy in response to naive intervention can be obtained using the Value Iteration method as:

$$\mu_u(\mathbf{x})=\underset{\mathbf{u}\in\mathcal{U}}{\operatorname{argmin}}[R^a_{\mu_a,\mathbf{u}}+\gamma M(\mu_a,\mathbf{u})\,\mathbf{V}^*_{\mu_a,\mu_u}],\ \text{for } \mathbf{x}\in\mathcal{X}, \tag{21}$$

where $\mathbf{V}^*_{\mu_a,\mu_u}$ is a fixed-point solution of the following Bellman equation:

$$\mathbf{V}^*_{\mu_a,\mu_u} = \min_{\mathbf{u}\in\mathcal{U}} \left[R^a_{\mu_a,\mathbf{u}} + \gamma M(\mu_a,\mathbf{u})\, \mathbf{V}^*_{\mu_a,\mu_u} \right]. \tag{22}$$

Here, we compare the performance of the proposed Nash equilibrium policy and the naive intervention policy in terms of the state value function and the steady state probability. The difference in expected discounted rewards if the system starts at state $\mathbf{x} \in \mathcal{X}$ and follows the optimal Nash equilibrium policy (i.e., (π^a_*, π^u_*)) and the naive intervention policy (μ_a, μ_u) can be expressed as:

$$e(\mathbf{x}) = \mathbf{V}^*_{\mu_a,\mu_u}(\mathbf{x}) - \mathbf{V}^*_{\pi^a_*,\pi^u_*}(\mathbf{x}). \tag{23}$$

As described in (3), any deviation by the interventionist from the optimal equilibrium policy results in fewer accumulated rewards, meaning that the cell can find ways to recur the system to cancerous/unhealthy conditions. The deviation occurs if the optimal stochastic intervention policy, π^a, differs from the naive intervention policy, μ_a, which can also be expressed if $\pi^a_*(\mu_a(\mathbf{x}) \mid \mathbf{x}) < 1$, for at least one $\mathbf{x} \in \mathcal{X}$. In this case, $e(\mathbf{x}) < 0$ for one or more $\mathbf{x} \in \mathcal{X}$, meaning that the interventionist's deviation from the optimal equilibrium policy will ultimately give the cell an opportunity to increase its profit (e.g., increase cancerous cell proliferation). On the other hand, it can be shown that the expected return of intervention is smaller than that of the Nash equilibrium policy, i.e., $\mathbf{V}^a_{\mu^a_*,\mu^u_*} \leq \mathbf{V}^a_{\pi^a_*,\pi^u_*}$. The performance of these policies is demonstrated in the numerical experiments.

We can also analyze the performance of the proposed intervention policy in terms of the steady-state probability. Let $\mathcal{X}^u \subset \mathcal{X}$ be the subset of undesirable states (more information is provided in the numerical experiment section). The steady-state probability indicates the long-term state visitation of systems in desirable and undesirable conditions. Let's start with the no intervention case, where the system is under the cell policy ϑ_u defined in (15). The steady-state probability under this policy can be expressed as:

$$\mathbf{\Pi}^\infty_{\vartheta_u}(j) = \lim_{t\to\infty} P(\mathbf{x}_t = \mathbf{x}^j \mid \vartheta_u, \mathbf{a} = \mathbf{0}), j = 1, ..., 2^d, \tag{24}$$

where $\mathbf{\Pi}^\infty$ specifies the long-term probability of the visitation of various states. One can compute $\mathbf{\Pi}^\infty_{\vartheta_u}$ as a unique solution of the following equations:

$$\mathbf{\Pi}^\infty_{\vartheta_u}(i) = \sum_{j=1}^{2^d} \left(M(\vartheta_u, \mathbf{a} = \mathbf{0})\right)_{ji} \mathbf{\Pi}^\infty_{\vartheta_u}(j),\ \sum_{i=1}^{2^d} \mathbf{\Pi}^\infty_{\vartheta_u}(i) = 1. \tag{25}$$

Similarly, the steady-state probability under the naive intervention policy, i.e., (μ_a, μ_u), is the solution to the following equations:

$$\mathbf{\Pi}^\infty_{\mu_a,\mu_u}(i) = \sum_{j=1}^{2^d} \left(M(\mu_a,\mu_u)\right)_{ji} \mathbf{\Pi}^\infty_{\mu_a,\mu_u}(j),\ \sum_{i=1}^{2^d} \mathbf{\Pi}^\infty_{\mu_a,\mu_u}(i) = 1. \tag{26}$$

The Nash equilibrium policy is stochastic, meaning that the transition matrix needed for the computation of the steady state probability belongs to the probability of actions. Given the Nash equilibrium policy (π^a_*, π^u_*), the steady state distribution can be computed as:

$$\begin{aligned}\mathbf{\Pi}^\infty_{\pi^a_*,\pi^u_*}(i) = \sum_{j=1}^{2^d} \sum_{\mathbf{a}\in\mathcal{A}} \sum_{\mathbf{u}\in\mathcal{U}} \left(M(\mathbf{a},\mathbf{u})\right)_{ji} \\ \pi^a_*(\mathbf{a}|\mathbf{x}^j)\pi^u_*(\mathbf{u}|\mathbf{x}^j)\mathbf{\Pi}^\infty_{\pi^a_*,\pi^u_*}(j),\end{aligned} \tag{27}$$

where $\sum_{i=1}^{2^d} \mathbf{\Pi}^\infty_{\pi^a_*,\pi^u_*}(i) = 1$. We can compute the steady-state probability of undesirable states under no-intervention, naive and Nash equilibtrim policies according to their steady-state distribution as:

$$\begin{aligned}\lim_{t\to\infty} p(\mathbf{x}_t \in \mathcal{X}^{\mathrm{u}} \mid \vartheta_u) &= \sum_{i=1}^{2^d} 1_{\mathbf{x}^i\in\mathcal{X}^{\mathrm{u}}} \mathbf{\Pi}^\infty_{\vartheta_u}(i), \\ \lim_{t\to\infty} p(\mathbf{x}_t \in \mathcal{X}^{\mathrm{u}} \mid \mu_a, \mu_u) &= \sum_{i=1}^{2^d} 1_{\mathbf{x}^i\in\mathcal{X}^{\mathrm{u}}} \mathbf{\Pi}^\infty_{\mu_a,\mu_u}(i), \\ \lim_{t\to\infty} p(\mathbf{x}_t \in \mathcal{X}^{\mathrm{u}} \mid \pi^a_*, \pi^u_*) &= \sum_{i=1}^{2^d} 1_{\mathbf{x}^i\in\mathcal{X}^{\mathrm{u}}} \mathbf{\Pi}^\infty_{\pi^a_*,\pi^u_*}(i),\end{aligned} \tag{28}$$

where $1_{\mathbf{x}^i\in\mathcal{X}^u}$ takes 1 if $\mathbf{x}^i$ contained in set $\mathcal{X}^u$, and 0 otherwise. The next section includes the numerical experiments examining the changes in steady-state probability under different policies.

V. 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. The parameters used throughout the numerical experiments are provided in Table 1. All results are averaged over 100 runs, where, for each run, an initial state of genes is selected randomly from all possible states. The random initial state represents the cell at different initial conditions, which aids in the efficient analysis of the performance of the proposed intervention policy. The proposed intervention policy is trained offline. During online execution, at any given time, the policy recommends an intervention based on the true genes' state. Cell internal stimuli are applied simultaneously as interventions, which shift the genes' state. Subsequently, the next intervention in the next time step is performed based on the state values of the genes. It should be noted that the sequential performance of interventions is similar to most well-known intervention techniques. Our proposed intervention policy is stochastic and accounts for the cell response, whereas existing approaches fail to consider such dynamic and intelligent cell responses.

A. P53-MDM2 Negative Feedback Loop Network

The p53-MDM2 negative feedback loop gene regulatory network is responsible for suppressing the tumor in humans and represents the cell response to stress signals that might cause genome instability [35, 36]. This network includes four genes, ATM, p53, WIP1, and MDM2. The diagram for the network is shown in Fig. 1, where solid arrows represent the

activating rules and blunt arrows demonstrate the suppressive rules. This Boolean model in (1) can be represented for this systems as [37, 38]:

$$\mathbf{f}(\mathbf{x}_t) = \overline{\begin{bmatrix} 0 & 0 & -1 & 0 \\ +1 & 0 & -1 & -1 \\ 0 & +1 & 0 & 0 \\ -1 & +1 & +1 & 0 \end{bmatrix} \mathbf{x}_t}, \tag{29}$$

where $\mathbf{x}_t = [\text{ATM}_t, \text{p53}_t, \text{WIP1}_t, \text{MDM2}_t]$, and $\overline{\mathbf{v}}$ maps the element of the vector $\mathbf{v}$ greater than 0 to 1 and others to 0.

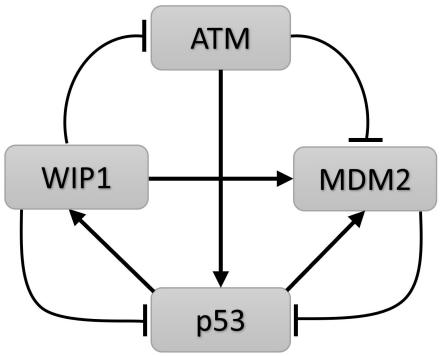


Fig. 1: The pathway diagram for the p53-MDM2 Boolean network.

TABLE I: Parameters used in numerical experiments.

Parameter	**Value**
Number of genes, d	4, 10
Disount Factor, γ	0.95
Process noise, p	0.05
Value Iteration Stopping threshold, ϵ	0.05
Initial State	$\mathbf{x}_0 \sim \text{Uniform}\{\mathbf{x}^1 ...., \mathbf{x}^{2^d}\}$

The p53-MDM2 system in the normal condition spends most of its time in the "0000" state, meaning that all genes are in inactivated states. In cancerous conditions, the system tends to show more gene activation, which often lead to uncontrolled proliferation of cells. These activities are due to internal stimuli within the cell, known as stress responses. Here, we consider the following internal stimuli for the cell's action space:

$$\mathcal{U} = \{\mathbf{u}^1 = [0\,0\,0\,0]^T, \mathbf{u}^2 = [1\,0\,0\,0]^T, \mathbf{u}^3 = [1\,1\,0\,0]^T\}, \tag{30}$$

where $\mathbf{u}^1$ corresponds to no-stress input, $\mathbf{u}^2$ alters the state value of the ATM gene, and $\mathbf{u}^3$ is capable of simultaneously flipping the state values of ATM and p53 genes. The impact of the cell control input is also investigated in the following paragraphs.

Intervention is critical for controlling cell proliferation in cancerous conditions and bringing the system closer to the normal condition (i.e., reducing the activation of genes). The intervention is achieved through the available drugs/therapies, which is expressed through the following intervention space:

$$\mathcal{A} = \{\mathbf{a}^1 = [0\,0\,0\,0]^T, \mathbf{a}^2 = [1\,0\,0\,0]^T, \mathbf{a}^3 = [0\,1\,0\,0]^T\}. \tag{31}$$

Given that the objective is to prevent the activation of the genes, the intervention reward function can be expressed as:

$$R^a(\mathbf{x}, \mathbf{a}, \mathbf{u}, \mathbf{x}') = -5||\mathbf{x}'||_1, \tag{32}$$

where $||\mathbf{x}'||_1$ counts the number of genes activation in the state vector $\mathbf{x}'$. The activation of each gene has a reward of -5, leading to the intervention reward taking in $\{-20, -15, -10, -5, 0\}$. The interventionist aims to maximize the accumulated intervention rewards by keeping the system in the "0000" state, while the cell with the opposite reward aims to increase the activation of the genes and move the system close to the "1111" state.

The optimal Nash equilibrium policy given the Bernoulli process noise $p = 0.05$, and the policy parameters $\gamma = 0.95$ and $\epsilon = 0.05$ is shown in Fig. 2. The blue bars represent the probability of each intervention (i.e., intervention policy), and the red bars indicate the cell policy. It can be seen that the Nash equilibrium policy is stochastic, meaning that the intervention and cell take actions according to the action probabilities indicated in Fig. 2. The action probabilities are different at various states; for instance, in state 16, the intervention and cell do not select $\mathbf{a}^1$ and $\mathbf{u}^1$, whereas, in state 1, all cell and intervention actions have non-zero probabilities.

The steady-state probabilities under the Nash equilibrium and no intervention policy are shown in Fig. 3. For the system under no intervention, the cell is capable of optimally using its internal stimuli to keep the genes activated, meaning that the cell stresses the system through internal stimuli to spend its time in state 16 (i.e., $\mathbf{x}^{16} = [1, 1, 1, 1]^T$). This can be seen as a large red bar in state 16, indicating the system spends 70% of the time in this state under no intervention. The steady-state probability under the Nash equilibrium policy is shown by blue bars. In this case, the state distribution has shifted significantly compared to the no-intervention case, and other states have been visited more frequently. This corresponds to the visitation of states with less activation of genes, which is an undesirable condition. In particular, the steady-state probability for state 16 has been reduced from 0.7 to 0.08, demonstrating the successful intervention outcome in response to the cell's aggressive policy of putting the system in a cancerous condition.

In this part, the performance of the proposed intervention policy is compared with the robust intervention policy [39]. This policy is widely used in systems biology when the system behavior is uncertain. In particular, the uncertainty comes from the cell's dynamic defensive response, which causes system behavior to change over time. Fig. 4 represents the average reward obtained by the following four policies: the Nash policy, naive intervention, no intervention, and robust intervention. It can be observed that the robust intervention policy outperforms both naive and no-intervention policies due to the adaptability inherent in such policy, considering possible uncertainties in the system model (i.e., dynamic cell responses). However, the proposed intervention policy, which takes into account the long-term impact of therapies and dynamic cell stimuli responses, surpasses the performance of the robust intervention policy. This superiority is evident in both short-term and long-

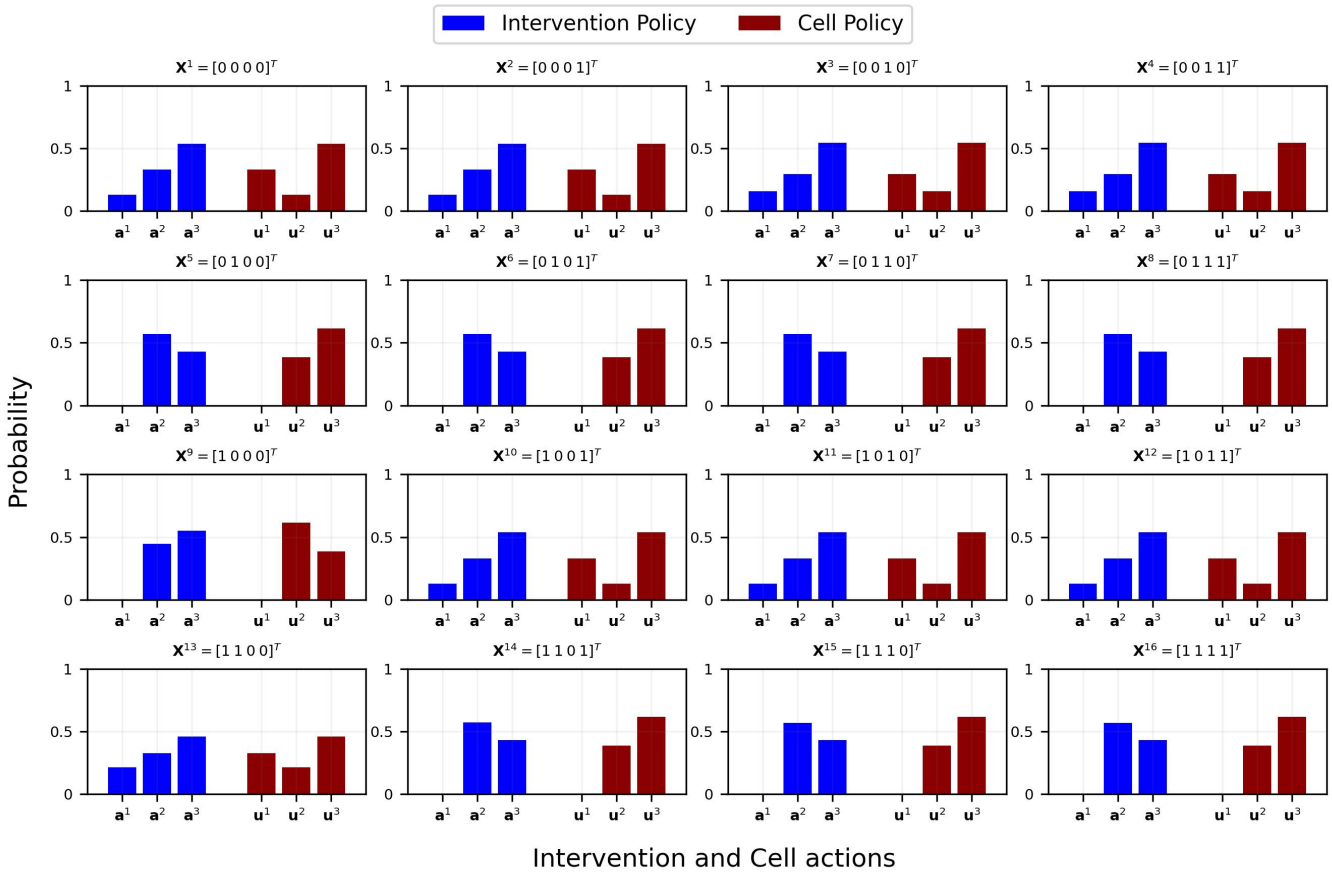


Fig. 2: The probability of intervention (blue) and cell (red) actions in different states under the proposed Nash equilibrium policy.

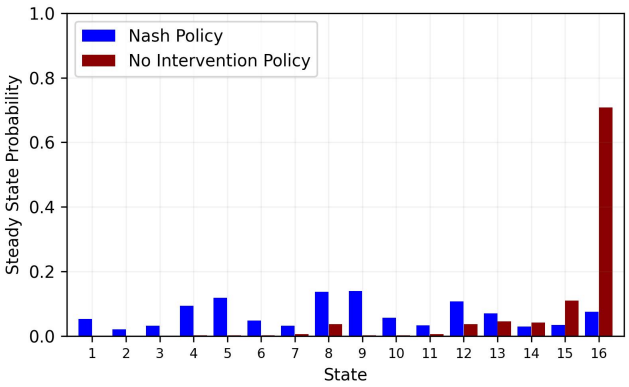


Fig. 3: The steady state probability under the proposed Nash equilibrium policy and no intervention policy.

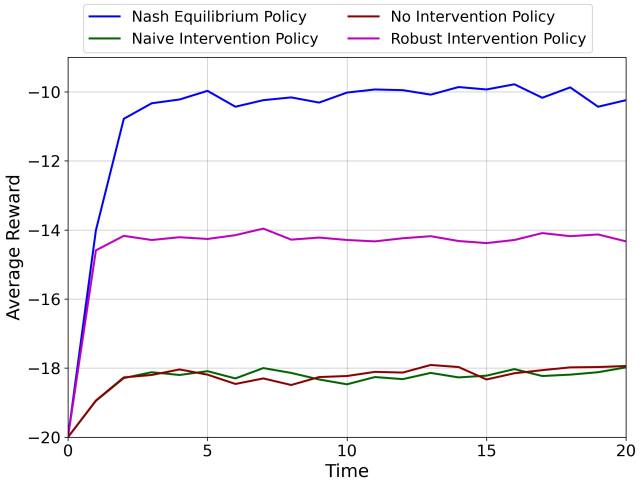


Fig. 4: Average reward over time obtained by different policies.

term behavior, as indicated by significantly higher average rewards obtained under the proposed policy.

In this part of the numerical experiments, the performance of the proposed intervention policy is compared with the naive intervention policy, which is obtained under the assumption of a non-responsive cell. Fig. 5 shows the naive intervention policy, where the deterministic intervention and cell actions are indicated in blue and red, respectively. One can see that the action $\mathbf{a}^1 = [0000]^T$ in state 1 is selected by the interventionist because, under the assumption of no response from the cell, the interventionist perceives that the system stays in the "0000" state upon taking this action. However, in reality, the cell has an intelligent response and takes action $\mathbf{u}^3 = [1100]^T$ to activate as many genes as possible. From the equilibrium perspective, one can compare the Nash policy in Fig. 2 with the naive policy in Fig. 5 as follows: the naive intervention policy in Fig. 5 can be seen as a policy deviated from the Nash policy shown in Fig. 2. As noted in (3), the Nash policy is the policy that neither the intervention nor the cell has

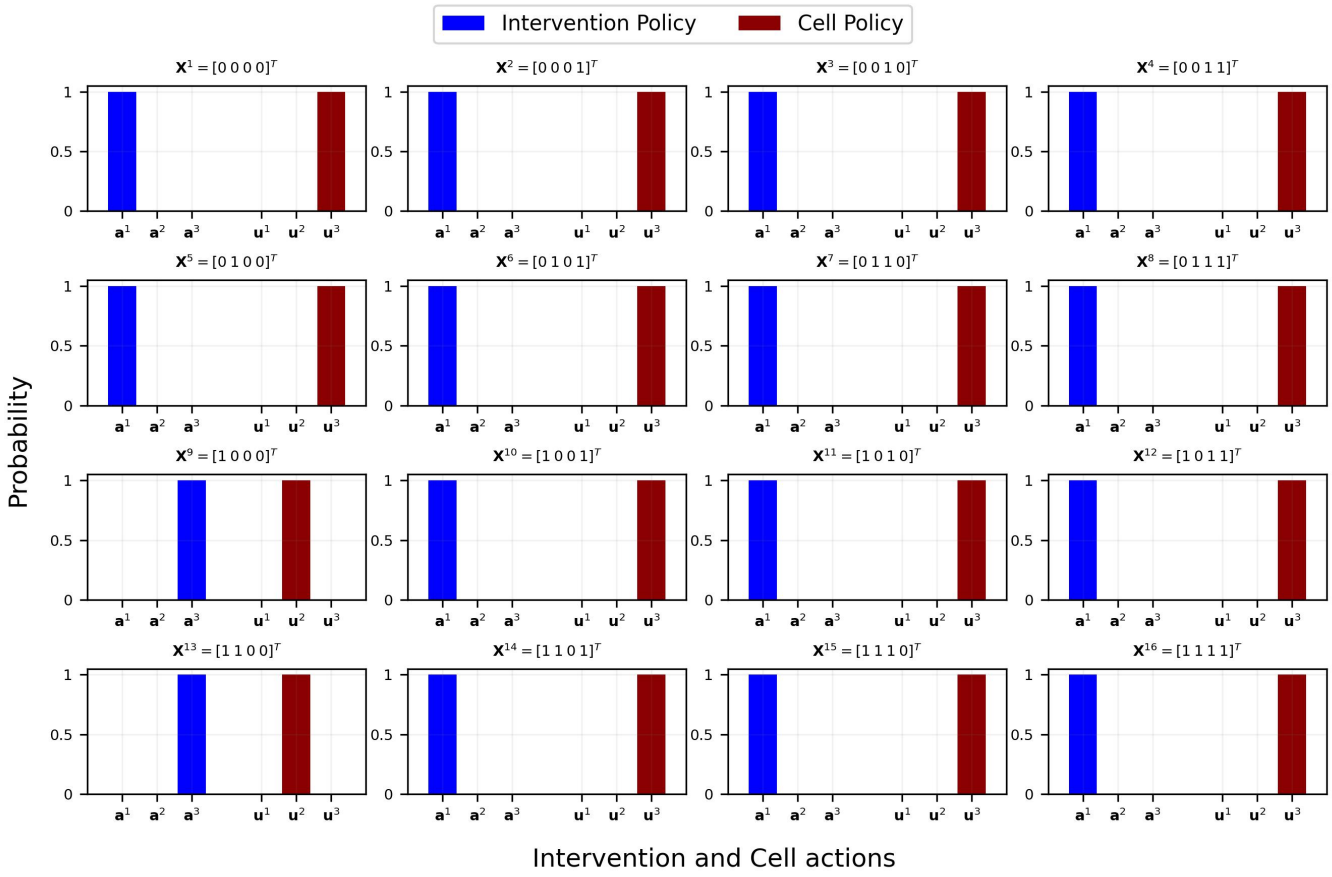


Fig. 5: The naive intervention policy obtained under non-responsive cell assumption (blue) and cell defensive policy (red).

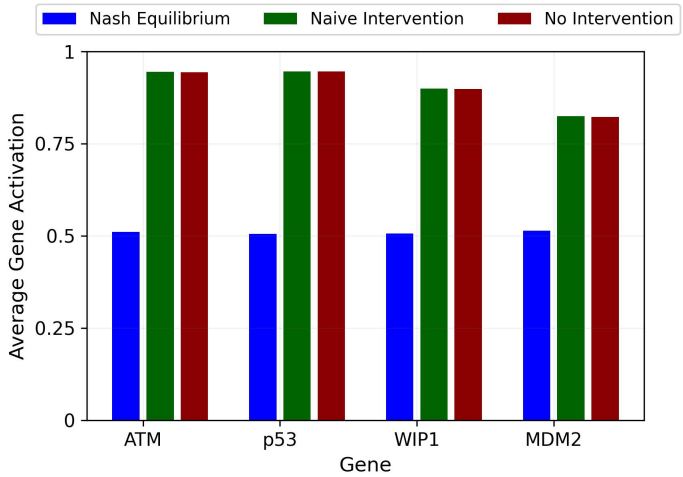


Fig. 6: The average gene activation obtained under optimal Nash equilibrium, naive intervention, and no intervention policies.

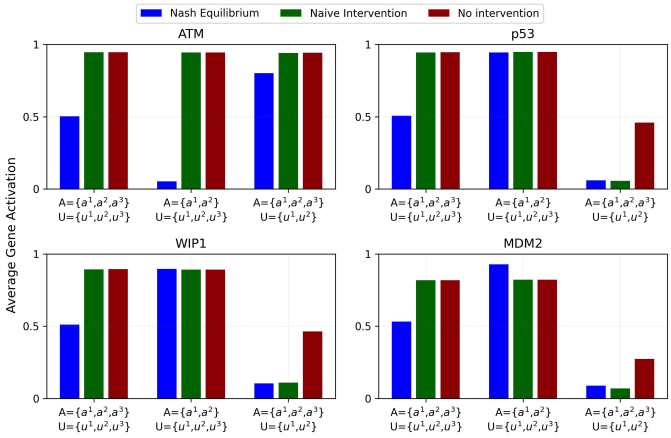


Fig. 7: The average gene activation obtained by various policies under three different cell and intervention spaces.

any incentive to deviate from. Therefore, as the intervention deviates from Nash, this provides the opportunity for the cell to find a better policy and enhance its profit. Therefore, the profit that the cell achieves in activating more cells under the naive intervention policy is the same profit that the intervention loses while deviating from the Nash policy.

To better understand the consequences of naive intervention without accounting for cell response in the recurrence of cancerous conditions, Fig. 6 illustrates the activation of genes in a steady state under various policies. Three policies are presented: Nash equilibrium, naive intervention, and no intervention. All four genes are mostly in an activated state under no intervention, as a result of the cell's aggressive response in activating all genes. For the system under the Nash equilibrium policy, gene activation has been significantly reduced by about 50%. Interestingly, gene activation is similar to no intervention under the naive intervention, demonstrating how cell response

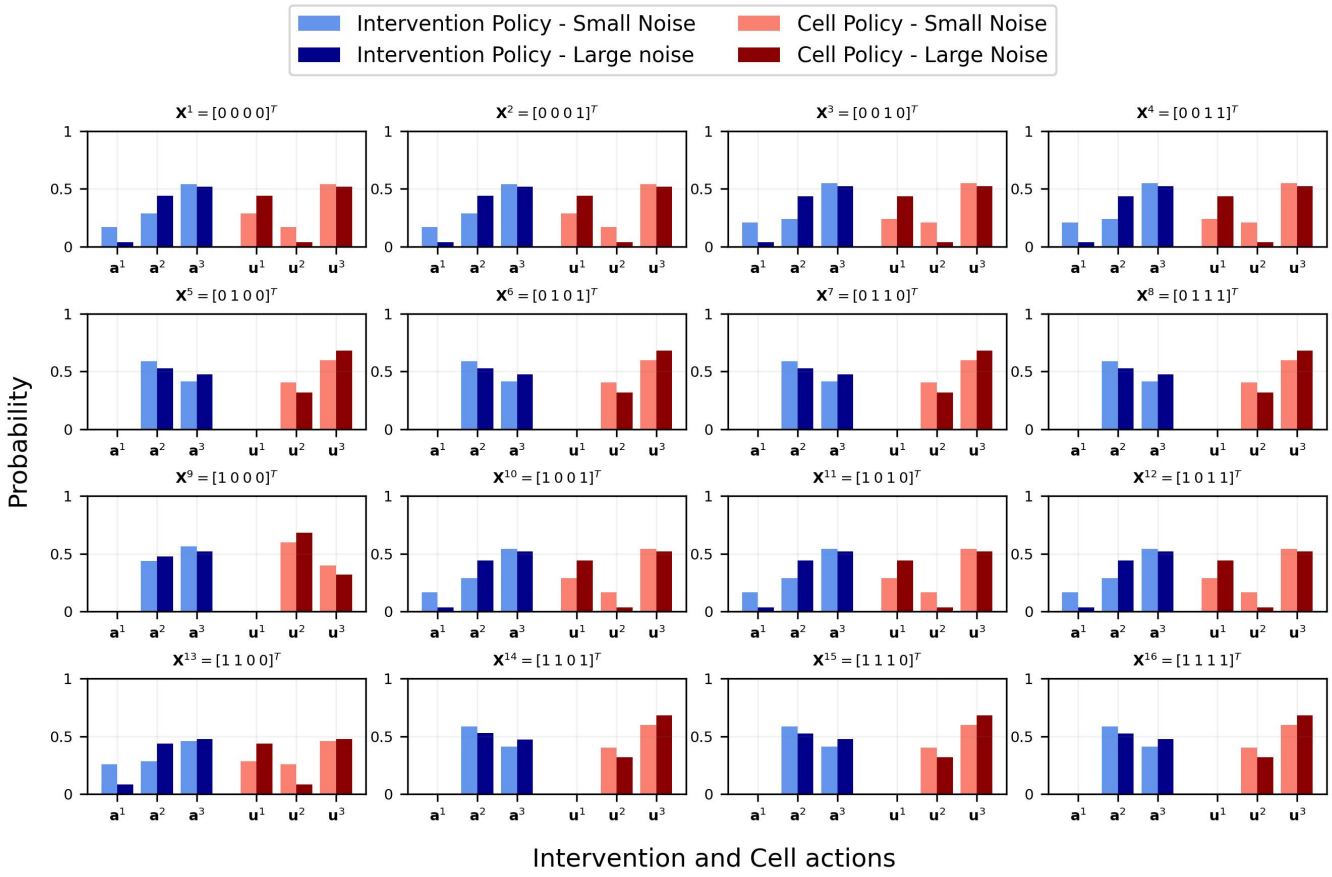


Fig. 8: Optimal Nash equilibrium policies for systems with low ($p = 0.001$) and high ($p = 0.2$) levels of stochasticity.

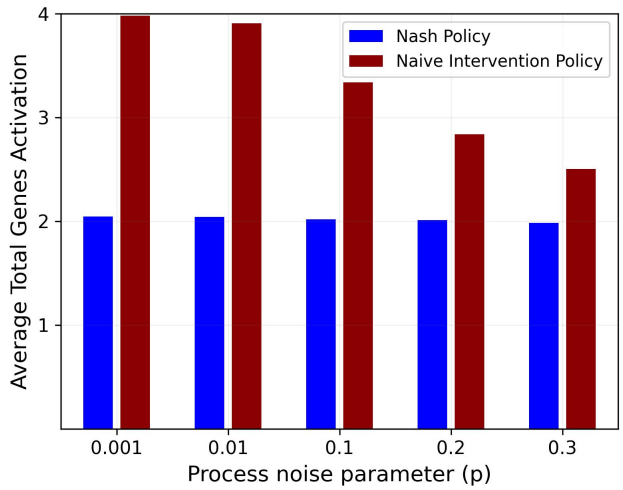
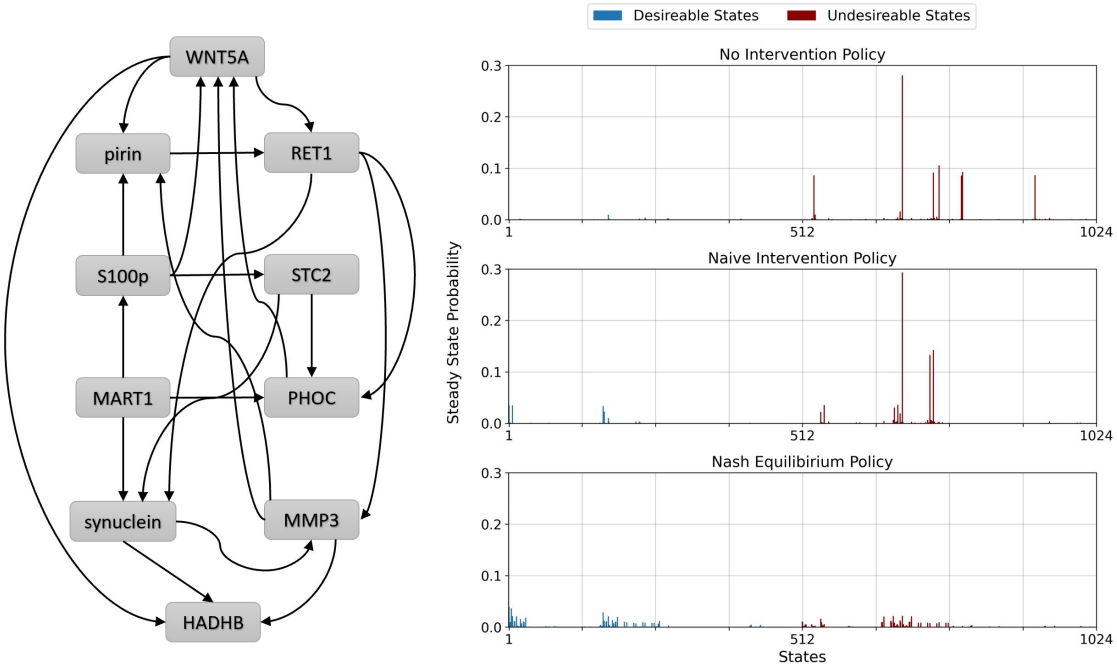


Fig. 9: Average total gene activation under the proposed Nash equilibrium and naive intervention policy with respect to the system stochasticity.

can nullify the impact of the naive intervention policy. Therefore, cells can fight against naive intervention and fully return the system to cancerous conditions, while under the Nash equilibrium and stochastic policies, a significant reduction in gene activation has been obtained.

The impact of the intervention and cell space on the performance of various policies is investigated in this part of numerical experiments. Three sets of action spaces are considered here. The first set consists of the full intervention and cell space indicated in (30) and (31). The second set's intervention space does not include $\mathbf{a}^3 = [0100]^T$, and the cell space in the third set does not include $\mathbf{u}^3 = [1100]^T$. Fig. 7 shows the results of various policies in terms of gene activation. Under the first set of action spaces, the Nash policy has reduced the activation of all genes in steady-state. For the second set with no $\mathbf{a}^3$ in the intervention space, the reduction in activation is only visible for ATM, whereas the other three genes remain mostly activated. This is due to the control power in intervention given a smaller intervention space, which impacts the overall performance of the intervention process (i.e., leading to a more overall activation of genes). For the naive intervention and no intervention policies, the results show the dominance of the cell in activating all genes, similar to the first action set. Finally, for the third set of action space, which corresponds to the scenario with limited cell stimuli, a more significant reduction in the activation of p53, WIP1, and MDM2 can be seen for Nash and naive intervention. Large activation can only be seen in ATM due to the $\mathbf{u}^2$ cell control that can directly control the ATM gene. One can also see much better performance of the naive intervention policy for the third action set, demonstrating that the naive intervention policy obtained under the non-responsive cell assumption becomes more effective in domains with fewer cell action spaces.



(a) (b)

Fig. 10: (a) The pathway diagram for the melanoma regulatory network; (b) the desirable and undesirable steady state probability under no intervention, naive intervention, and Nash equilibrium policies.

The impact of the system stochasticity on the proposed Nash equilibrium intervention policy is investigated in this part. Fig. 8 represents the Nash policy under two different stochasticity levels: 1) the small noise case corresponds to the Bernoulli process noise $p = 0.001$ and is indicated by the light colors; 2) a larger noise modeling more chaotic systems is associated with $p = 0.2$ and denoted with darker colors. It can be seen that the Nash policy is different under these two settings; for instance, the probability of taking intervention $\mathbf{a}^1$ in state 1 is larger under the small noise, coming from more chaotic activities under larger noise which demands taking other actions more often.

Meanwhile, the total gene activation in a steady state for five different levels of stochasticity is represented in Fig. 9. It can be seen that the results of the Nash policy are consistently similar in all conditions, demonstrating the robustness of the Nash policy with respect to the level of stochasticity. In fact, chaotic systems, i.e., those with larger levels of noise, can be perceived as scenarios where decision-making becomes more challenging for both cells and interventionists, resulting in similar performance regardless of changes in the noise level. Additionally, the performance of the Naive intervention policy decreases as the noise level increases. This reduction is due to the difficulty of cell response to naive intervention under chaotic systems.

B. Melanoma Regulatory Network

In this part of the numerical experiment, we analyze the performance of the proposed policy using the melanoma regulatory network [34, 40]. This network consists of complex interactions between genes and the signaling pathway that controls various cellular processes, including cell growth, differentiation, apoptosis, and migration. The regulatory relationships for this network are presented in Fig. 10(a), where the system consists of 10 genes. The genes represented in the state vectors are, in order: WNT5A, pirin, S100P, RET1, MMP3, PHOC, MART1, HADHB, synuclein, and STC2. Dysregulation of the network can lead to the uncontrolled growth of melanocytes and the development of melanoma. Key genes within these networks, such as WNT5A, play critical roles in melanoma development and progression.

Several signaling pathways such as Ras, B-Raf, MEK, PTEN, phosphatidylinositol-3 kinase (PI3Ks), and Akt, have been implicated in the development and progression of melanoma. This paper focuses on a widely researched Boolean network model of melanoma with 10 genes [34], known for its application in deriving dynamic interventions. The 10 genes in the melanoma regulatory network in Fig. 10(a) lead to $2^{10} = 1,024$ gene states. The Boolean function in this case can be expressed as [34]:

$$\mathbf{f}(\mathbf{x}_t) = [f_1(\mathbf{x}_t), f_2(\mathbf{x}_t), ..., f_{10}(\mathbf{x}_t)]^T$$
$$= \begin{bmatrix} (\text{S100P} \wedge \text{MMP3} \wedge \overline{\text{PHOC}}) \vee (\overline{\text{MMP3}} \wedge \text{PHOC}) \\ (\overline{\text{WNT5A}} \wedge \overline{\text{S100P}} \wedge \text{MMP3}) \vee (\text{WNT5A} \wedge \overline{\text{S100P}} \wedge \overline{\text{MMP3}}) \\ \text{MART1} \\ (\overline{\text{WNT5A}} \wedge \text{pirin} \wedge \text{RET1}) \vee (\overline{\text{pirin}} \wedge \text{RET1}) \\ (\text{RET1} \wedge \text{synuclein}) \vee \overline{\text{synuclein}} \\ (\overline{\text{RET1}} \wedge \overline{\text{MART1}}) \vee (\text{RET1} \wedge \text{MART1} \wedge \text{STC2}) \\ \text{MART1} \\ (\text{WNT5A} \wedge \text{MMP3}) \vee (\overline{\text{MMP3}} \wedge \overline{\text{synuclein}}) \vee (\text{WNT5A} \wedge \overline{\text{MMP3}} \wedge \text{synuclein}) \\ (\overline{\text{RET1}} \wedge \overline{\text{MART1}} \wedge \overline{\text{STC2}}) \vee (\overline{\text{RET1}} \wedge \overline{\text{MART1}} \wedge \text{STC2}) \vee \text{MART1} \\ \overline{\text{S100P}} \end{bmatrix}. \quad (33)$$

The activation of WNT5A has been explicitly linked to the development of metastatic conditions. Utilizing antibodies to bind to WNT5A and block it from activating its receptor has shown to be effective in deriving intervention. In particular,

reducing the activation of the WNT5A gene helps prevent melanoma from metastasizing and achieving a desirable outcome [40]. Consequently, the reward function for intervention can be formulated as follows:

$$R^a(\mathbf{x}, \mathbf{a}, \mathbf{u}, \mathbf{x}') = -5\mathbf{x}'(1), \tag{34}$$

which refers to the reward value of -5 for activation of WNT5A.

The internal cell stimuli space is $\mathcal{U} = \{\mathbf{u}^1, \mathbf{u}^2, \mathbf{u}^3\}$, where $\mathbf{u}^1$ corresponds to no stimuli, and $\mathbf{u}^2$ and $\mathbf{u}^3$ represent stimuli over the S100P and MMP3 genes, respectively. For the intervention space, we consider $\mathcal{A} = \{\mathbf{a}^1, \mathbf{a}^2\}$, where $\mathbf{a}^1$ represents no control and $\mathbf{a}^2$ represents intervention over the PHOC gene.

Fig. 10(b) represents the steady state probability for the system under the Nash equilibrium policy, the naive intervention, and the no intervention policies. The blue and red bars represent the desirable (i.e., inactivated WNT5A) and undesirable states (i.e., activated WNT5A). It can be seen that under no intervention, the system spends most of its time in undesirable states. For the naive intervention, there is a reduction in undesirable states compared to the no intervention cases. However, the highest reduction in undesirable states can be seen under the proposed Nash equilibrium policy, which most effectively reduces the steady state of undesirable states towards desirable ones.

VI. Conclusion

This paper developed an optimal intervention policy for gene regulatory networks with responsive cells. The GRNs are modeled using the Boolean network with perturbation, and the dynamic and adaptive battle between intervention and cells is modeled as a two-player zero-sum game. Most existing intervention policies are incapable of taking into account cell responses to the intervention, leading to early and short-term success of interventions, followed by partial or full recurrence of diseases. By contrast, this paper develops an optimal Nash equilibrium intervention policy that ensures the best possible intervention solutions under any cell response. We analytically analyze the superiority of the proposed intervention policy against existing intervention techniques. A comprehensive numerical experiment using the p53-MDM2 negative feedback loop regulatory network and the melanoma network demonstrates the high performance of the proposed method.

Our future research will investigate a more realistic context for genomics intervention, including the partial observability of genes' states through gene-expression data, the lack of partial knowledge about cell stimuli (or responses), and the pathway of the gene regulatory networks. Meanwhile, we will study the scalability of the proposed policy to larger networks consisting of several genes. These studies will aim at enabling the real-world application of such policies in real experimental settings.

Acknowledgment

The authors acknowledge the support of the National Institute of Health award 1R21EB032480-01, the 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] I. Shmulevich and E. R. Dougherty, *Probabilistic Boolean networks: the modeling and control of gene regulatory networks*. SIAM, 2010.

[2] I. Shmulevich, E. R. Dougherty, and W. Zhang, "From Boolean to probabilistic Boolean networks as models of genetic regulatory networks," *Proceedings of the IEEE*, vol. 90, no. 11, pp. 1778–1792, 2002.

[3] J. Liang and J. Han, "Stochastic Boolean networks: an efficient approach to modeling gene regulatory networks," *BMC systems biology*, vol. 6, no. 1, pp. 1–21, 2012.

[4] Y. Wang, C. Liu, Q. Xu, X. Han, and Z.-P. Liu, "PKI: A bioinformatics method of quantifying the importance of nodes in gene regulatory network via a pseudo knockout index," *Biochim Biophys Acta Gene Regul Mech*, 2023.

[5] R. Pal, A. Datta, and E. R. Dougherty, "Optimal infinite-horizon control for probabilistic Boolean networks," *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 2375–2387, 2006.

[6] S. Kharade, S. Sutavani, S. Wagh, A. Yerudkar, C. Del Vecchio, and N. Singh, "Optimal control of probabilistic Boolean control networks: A scalable infinite horizon approach," *International Journal of Robust and Nonlinear Control*, 2021.

[7] I. Shmulevich, E. R. Dougherty, and W. Zhang, "Gene perturbation and intervention in probabilistic Boolean networks," *Bioinformatics*, vol. 18, no. 10, pp. 1319–1331, 2002.

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

[9] N. S. Taou, D. W. Corne, and M. A. Lones, "Investigating the use of Boolean networks for the control of gene regulatory networks," *Journal of computational science*, vol. 26, pp. 147–156, 2018.

[10] G. Papagiannis and S. Moschoyiannis, "Deep reinforcement learning for control of probabilistic Boolean networks," *arXiv preprint arXiv:1909.03331*, 2019.

[11] X. Qian and E. R. Dougherty, "Effect of function perturbation on the steady-state distribution of genetic regulatory networks: Optimal structural intervention," *IEEE Transactions on Signal Processing*, vol. 56, no. 10, pp. 4966–4976, 2008.

[12] M. R. Yousefi and E. R. Dougherty, "Intervention in gene regulatory networks with maximal phenotype alteration," *Bioinformatics*, vol. 29, no. 14, pp. 1758–1767, 2013.

[13] J. Zhong, Y. Liu, J. Lu, and W. Gui, "Pinning control for stabilization of Boolean networks under knock-out perturbation," *IEEE Transactions on Automatic Control*, vol. 67, no. 3, pp. 1550–1557, 2021.

[14] X. Li, H. Li, Y. Li, and X. Yang, "Function perturbation impact on stability in distribution of probabilistic Boolean networks," *Mathematics and Computers in Simulation*, vol. 177, pp. 1–12, 2020.

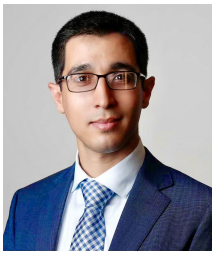
[15] M. Imani, R. Dehghannasiri, U. M. Braga-Neto, and E. R. Dougherty, "Sequential experimental design for optimal structural intervention in gene regulatory networks based on the mean objective cost of uncertainty," *Cancer informatics*, vol. 17, p. 1176935118790247, 2018.

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

[17] L. Wang and D. Schonfeld, "Game theoretic model for control of gene regulatory networks," in *2010 IEEE International Con-*

ference on Acoustics, Speech and Signal Processing, pp. 542–545, 2010.
[18] M. Alali and M. Imani, "Reinforcement learning data-acquiring for causal inference of regulatory networks," in *American Control Conference (ACC), IEEE*, 2023.
[19] A. Ravari, S. F. Ghoreishi, and M. Imani, "Optimal recursive expert-enabled inference in regulatory networks," *IEEE Control Systems Letters*, vol. 7, pp. 1027–1032, 2022.
[20] A. Sahoo and V. Narayanan, "Optimization of sampling intervals for tracking control of nonlinear systems: A game theoretic approach," *Neural Networks*, vol. 114, pp. 78–90, 2019.
[21] L. S. Shapley, "Stochastic games," *Proceedings of the national academy of sciences*, vol. 39, no. 10, pp. 1095–1100, 1953.
[22] K. Zhang, Z. Yang, and T. Basar, "Policy optimization provably converges to nash equilibria in zero-sum linear quadratic games," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
[23] K. Zhang, Z. Yang, and T. Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," *Handbook of reinforcement learning and control*, pp. 321–384, 2021.
[24] A. Rubinstein, H. W. Kuhn, O. Morgenstern, and J. Von Neumann, *Theory of Games and Economic Behavior*. Princeton university press, 2007.
[25] A. Sahoo and S. Jagannathan, "Stochastic optimal regulation of nonlinear networked control systems by using event-driven adaptive dynamic programming," *IEEE Transactions on Cybernetics*, vol. 47, no. 2, pp. 425–438, 2017.
[26] D. Fudenberg and D. K. Levine, "The theory of learning in games," *MIT press*, 1998.
[27] M. J. Osborne and A. Rubinstein, "An introduction to game theory," *Oxford University Press*, 2004.
[28] J. A. Filar and K. J. Vrieze, "Competitive Markov decision processes," *Springer Science & Business Media*, 1997.
[29] M. L. Puterman, "Markov decision processes: discrete stochastic dynamic programming," *John Wiley & Sons*, 1994.
[30] J. von Neumann and O. Morgenstern, "Theory of games and economic behavior," *Princeton University Press*, pp. 85–168, 1944.
[31] A. Datta, R. Pal, A. Choudhary, and E. R. Dougherty, "Control approaches for probabilistic gene regulatory networks-what approaches have been developed for addressing the issue of intervention?," *IEEE Signal Processing Magazine*, vol. 24, no. 1, pp. 54–63, 2007.
[32] B. Faryabi, J.-F. Chamberland, G. Vahedi, A. Datta, and E. R. Dougherty, "Optimal intervention in asynchronous genetic regulatory networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 3, pp. 412–423, 2008.
[33] R. Layek, A. Datta, R. Pal, and E. R. Dougherty, "Adaptive intervention in probabilistic Boolean networks," *Bioinformatics*, vol. 25, no. 16, pp. 2042–2048, 2009.
[34] X. Qian and 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*, vol. 9, no. 1, pp. 123–136, 2012.
[35] E. Batchelor, A. Loewer, and G. Lahav, "The ups and downs of p53: understanding protein dynamics in single cells," *Nature Reviews Cancer*, vol. 9, no. 5, p. 371, 2009.
[36] M. Imani and U. Braga-Neto, "Gene regulatory network state estimation from arbitrary correlated measurements," *EURASIP Journal on Advances in Signal Processing*, vol. 2018, no. 1, pp. 1–10, 2018.
[37] M. Imani and U. M. Braga-Neto, "Maximum-likelihood adaptive filter for partially observed Boolean dynamical systems," *IEEE Transactions on Signal Processing*, vol. 65, no. 2, pp. 359–371, 2017.
[38] M. Alali and M. Imani, "Inference of regulatory networks through temporally sparse data," *Frontiers in control engineering*, vol. 3, p. 1017256, 2022.
[39] R. Pal, A. Datta, and E. R. Dougherty, "Robust intervention in probabilistic boolean networks," *IEEE Transactions on Signal Processing*, vol. 56, no. 3, pp. 1280–1294, 2008.
[40] A. T. Weeraratna, Y. Jiang, G. Hostetter, K. Rosenblatt, P. Duray, M. Bittner, and J. M. Trent, "WNT5A signaling directly affects cell motility and invasion of metastatic melanoma," *Cancer cell*, vol. 1, no. 3, pp. 279–288, 2002.

Seyed Hamid Hosseini is a Ph.D. student in the Electrical and Computer Engineering Department at Northeastern University. He received his Bachelor's and Master's degrees in Electrical Engineering from Sharif University of Technology, Iran, in 2018 and 2021, respectively. His research interests include multi-agent reinforcement learning, game theory, and computational biology.



Mahdi Imani received his Ph.D. degree in Electrical and Computer Engineering from Texas AM University, College Station, TX in 2019. He is currently an Assistant Professor in the Department of Electrical and Computer Engineering at Northeastern University. His research interests include machine learning, Bayesian statistics, and decision theory, with a wide range of applications from computational biology to cyber-physical systems. He is the recipient of several awards, including the NIH NIBIB Trailblazer award in 2022, the Oracle Research Award in 2022, the NSF CISE Career Research Initiation Initiative award in 2020, the Association of Former Students Distinguished Graduate Student Award for Excellence in Research-Doctoral in 2019, and the Best Paper Finalist award from the American Control Conference in 2023 and the 49th Asilomar Conference on Signals, Systems, and Computers in 2015.