Understanding adaptive immune system as reinforcement learning

Takuya Kato[1,*] and Tetsuya J. Kobayashi[2,1,3,4,†]

[1]*Department of Mathematical Informatics, Graduate School of Information and Science, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8654, Japan*

[2]*Institute of Industrial Science, The University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan*

[3]*Graduate School of Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan*

[4]*Universal Biology Institute, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8654, Japan*

(Received 31 January 2020; accepted 18 February 2021; published 9 March 2021)

The adaptive immune system of vertebrates can detect, respond to, and memorize diverse pathogens from past experience. While the clonal selection of T helper (Th) cells is the simple and established mechanism to better recognize new pathogens, the question that still remains unexplored is how the Th cells can acquire better ways to bias the responses of immune cells for eliminating pathogens more efficiently by translating the recognized antigen information into regulatory signals. In this work, we address this problem by associating the adaptive immune network organized by the Th cells with reinforcement learning (RL). By employing recent advancements of network-based RL, we show that the Th immune network can acquire the association between antigen patterns of and the effective responses to pathogens. Moreover, the clonal selection as well as other intercellular interactions are derived as a learning rule of the network. We also demonstrate that the stationary clone-size distribution after learning shares characteristic features with those observed experimentally. Our theoretical framework may contribute to revising and renewing our understanding of adaptive immunity as a learning system.

DOI: 10.1103/PhysRevResearch.3.013222

I. INTRODUCTION

The adaptive immunity of vertebrates is a complex adaptive system. The system constantly adapts to intruding pathogens by orchestrating the populations and responses of diverse immune cells, each type of which can have distinct roles [1–3]. For example, effector cells (innate cells, T killer cells, a part of innate lymphoid cells, B cells, etc.) are responsible for executing intrinsic pathogen-specific responses, whereas T helper (Th) cells mainly control and bias the activities of these effector cells. The diversity and activity of immune cells are modulated over the organisms' lifetimes through intercellular communications via hundreds of chemical messengers and subsequent adaptive changes in the population sizes or phenotypic states [4–7]. Even though young children are susceptible to infections [8–10], they may develop higher resistance to infections through the modulation. As evidenced by vaccination and immunization [11–13], such modulation may be achieved as an adaptive response to previous infections, which can be regarded as a type of learning from experience.

Despite the availability of the latest experimental technologies, revealing the principles of such complex learning dynamics is still intricate because the immunological dynamics is shaped and organized by the collective interactions of the entire immune cell population, which prevents us from simply reducing the problem down to the mere existence of specific cell types or molecules. In order to comprehend a complex learning system in neuroscience, Marr highlighted the importance of characterizing the system at three levels [14,15]: the goal of the system (the computational level), the process and computation to realize the goal (the algorithmic level), and the physical implementation of the process (the implementation level).

In the past decades, a substantial amount of effort has been devoted to understanding the immune system, especially at the implementation level [1,2]. Cellular and molecular immunology has identified hundreds of phenotypically and functionally distinct immune cells and associated molecular markers [5]. Concurrently, tens of cytokines and chemokines have been discovered as chemical messages to coordinate the communications between immune cells [6,16]. Moreover, with the advancement of high-throughput sequencing, it is now possible to measure the diversity of T and B cells, which constitutes an integral part of immunological recognition and memory [17,18]. Despite the accumulation of such data and knowledge at the implementation level, our understanding of the immune system at the algorithmic and computational levels lags far behind and still remains limited to conceptual theories such as the clonal selection theory [19,20]. In the face of the revealed complexity, the theory is neither

[*] takuya.kato.origami@gmail.com

[†] tetsuya@mail.crmind.net

sufficiently descriptive nor quantitative to draw new insights [21] and should be renewed to have a greater explanatory and predictive power by being endowed with a firm mathematical basis [22–25]. In particular, most theoretical works still focus only on antigen-induced T cell selection even though activated innate immune cells convey information to T cells about the origin and nature of antigens and pathogens via co-stimulation and cytokine signals [26]. The problem that still remains unsolved is how Th cells are modulated by these different signals not only to recognize antigens but also to induce and bias the activities of the groups of effector cells for evicting pathogens more efficiently by translating the recognized antigen information.

To this end, we revised the concept of immunological learning and bestowed it with a modern mathematical basis by focusing on the computational and algorithmic levels. At the computational level, the goal of the system may be to learn better ways from past experiences to bias the activities of the effector cells in response to infections, so as to evict the infected pathogens more promptly and specifically. We formulated this process as a reinforcement learning (RL) problem described using a Markov decision process (MDP) [27–29].

At the algorithmic level, the system has to find a better way to bias the activities of the effector cells to the infected pathogens. For example, activating T killer cells is effective in coping with virus-infected cells but not with bacteria. Th cells coordinate this process; they obtain the information of the infected pathogens from the pattern of antigens presented by antigen presenting cells (APCs). Then, the Th cells regulate the activities of groups of effector cells by secreting different kinds of cytokines. As a network, the Th cell population constitutes the middle layer between the pattern of antigens and that of the activated effector cells (Fig. 1). By following the recent advancements in the applications of neural networks for solving RL problems [30,31], we derive the learning dynamics of the Th cell population, which corresponds to the algorithm to achieve the goal formulated at the computational level. The derived learning dynamics has the form of a replicator equation, which can be interpreted at the implementation level as the clonal selection of the Th cells in response to antigen presentation. The derived learning rule also contains the terms that work as feedback from effector cells to Th cells. These results provide us with fruitful insights on the potential roles of molecular and cellular components for learning in real immune systems. The simulations of the MDP with the derived learning dynamics demonstrate that the clone-size distributions of the Th cell population after learning can show properties that are qualitatively consistent with those observed experimentally.

It should be noted that our formulation is not intended to account for all the details of immunity but to highlight the learning aspect of immunity. While we focus primarily on Th cells, we also discuss how other components and constraints of the immune system can be incorporated as possible extensions. Our approach can also complement more mechanistic investigations of the dynamics and regulation of immune responses by suggesting the functional roles of such dynamics at the computational and algorithmic levels.

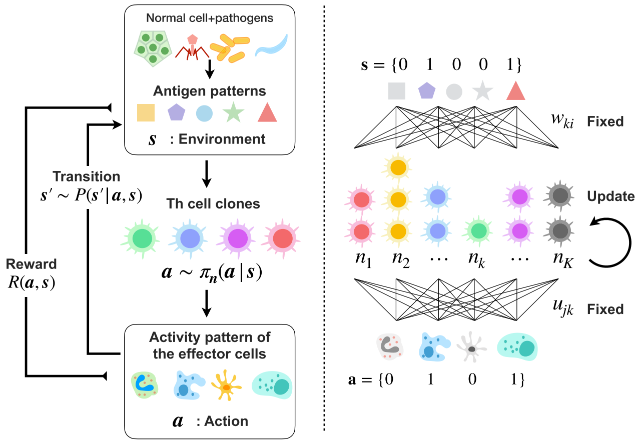

FIG. 1. Schematic diagram of the adaptive immune system in relation with network-based reinforcement learning. When an infection occurs, APCs engulf the pathogens and present multiple antigens as an antigen pattern $\mathbf{s}$. The Th cell population recognizes the antigen pattern $\mathbf{s}$ and biases the activities of the effector cells $\mathbf{a}$. The stochastic mapping $\pi_{\mathbf{n}}(\mathbf{a}|\mathbf{s})$ from $\mathbf{s}$ to $\mathbf{a}$ is regarded as the policy of the system parametrized by the abundance of the Th clones, $\mathbf{n}$. The effectiveness of the pattern of the effector activities $\mathbf{a}$ to the infection is represented as the reward $R(\mathbf{s}, \mathbf{a})$. Learning of the system is achieved not by adjusting weights but by the modulation of the clone-size distribution $\mathbf{n}$.

II. MODEL

A. Framing adaptive immune response and learning as reinforcement learning

Upon infection by a pathogen, the innate immune responses are initiated. Subsequently, the APCs that engulf the pathogen start presenting peptide fragments of the pathogen (antigens) to the Th cells. In general, multiple peptide fragments are derived from a pathogen and their pattern works as a fingerprint of the pathogen. Let N be the number of different types of antigens and $\mathbf{s} \in \{0, 1\}^N$ be the pattern of the antigens; $s_i = 1$ and $s_i = 0$ indicate the presence and absence of the ith type of antigen, respectively (Fig. 1). An antigen pattern $\mathbf{s} = \{0, 1, 1, 1\}$, for example, indicates that all but the first type of antigens exist among the four. This antigen pattern conveys information about the infected pathogens to the Th cells.

Upon receiving the information, the Th cell population secretes a pattern of cytokines, which may differ depending on the activities of the Th cells induced by the antigen pattern $\mathbf{s}$. In turn, depending on the cytokine pattern, different groups of effector cells, e.g, B cells, T killer cells, macrophages, etc., are activated or deactivated, which constitutes the response to the pathogen. It should be noted that these effector cells are different from the effector T cells. Let M be the number of different types of effector cells and $\mathbf{a} \in \{0, 1\}^M$ be the activation pattern of the effector cells.

Here, $a_j = 1$ and $a_j = 0$ indicate the activation and inactivation of the jth type of effector cells, respectively (Fig. 1). An activation pattern $\mathbf{a} = \{1, 1, 0\}$, for example, implies that the first and second types of effector cells are activated while the third one is inactivated. The Th cell population can bias the activation pattern $\mathbf{a}$ of the effector cells based on the information of the antigen pattern $\mathbf{s}$. We express this role of the Th cell population by a stochastic transition probability $\pi(\mathbf{a}|\mathbf{s})$ that determines which activation pattern $\mathbf{a}$ is likely to be realized when the Th cells are exposed to the antigen pattern $\mathbf{s}$. We call this conditional probability distribution π the policy of the Th cell population.

Patterns of the activated effector cells have different influences on the antigen patterns. If the activated effector cells are effective against the pathogen, the types of antigens that are specific to the pathogen should disappear from the antigen pattern with a high probability. Otherwise, the antigen pattern $\mathbf{s}$ may not change much. We express this stochastic transition of $\mathbf{s}$ with a transition probability $P(\mathbf{s}'|\mathbf{s},\mathbf{a})$, where $\mathbf{s}$ and $\mathbf{s}'$ denote the antigen patterns before and after being exposed to a pattern $\mathbf{a}$ of effector cells, respectively. It should be noted that this transition law itself is physically determined by the nature of effector cells and pathogens and that each transition may not necessarily be dependent on its immunological effectiveness. Immunological effectiveness is a vague but important factor with which Th cells can learn a better policy π to induce a better activity pattern $\mathbf{a}$ for a given antigen pattern $\mathbf{s}$. The immunological effectiveness of action $\mathbf{a}$ for antigen pattern $\mathbf{s}$ is modeled here using the reward function $R(\mathbf{s},\mathbf{a}) \in [0,\infty)$ (Fig. 1). The reward function can be a complicated function of the antigen and activation patterns in general, but it can be presumed to take a large value if the activity pattern of the effector cells is effective for the current state; otherwise, it takes a low value. The details and a biological counterpart of this reward signal will be discussed in a later section.

In summary, the learning dynamics of the immune cell population is modeled by the following five components: (1) a set of possible antigen patterns, $\mathcal{S} \subset \{0,1\}^N$; (2) a set of possible activity patterns of the effector cells, $\mathcal{A} \subset \{0,1\}^M$; (3) a transition probability $P(\mathbf{s}'|\mathbf{s},\mathbf{a})$, (4) a reward function $R(\mathbf{s},\mathbf{a}) \in \mathbb{R}$, and (5) a policy of Th cell population, $\pi(\mathbf{a}|\mathbf{s})$. In the terminology of the MDP, the first four components correspond to a set of states, a set of actions, transition probability, and reward, respectively.

By optimizing the policy $\pi(\mathbf{a}|\mathbf{s})$ via interactions with pathogens, the immune system can adaptively respond to infections. The optimal policy $\pi^\dagger$ is characterized as the policy that maximizes the expected cumulative reward $J[\pi] := \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t R^\pi(\mathbf{s}_t,\mathbf{a}_t)]$ with the discount rate $0 \leqslant \gamma \leqslant 1$ as $\pi^\dagger := \arg\max_\pi J[\pi]$, where

$$R^\pi(\mathbf{s}_t,\mathbf{a}_t) = R(\mathbf{s}_t,\mathbf{a}_t) - \frac{1}{\beta}\log\frac{\pi(\mathbf{a}_t|\mathbf{s}_t)}{\pi_0(\mathbf{a}_t|\mathbf{s}_t)}.$$

The additional term $\frac{1}{\beta}\log \pi/\pi_0$ represents the control cost to bias the activity of effector cells from their intrinsic behavior π_0 to π. $\beta \in (0,\infty)$ is a scaling parameter of the cost. Because of the functional form of the control cost, this formulation is also recognized as an entropy-regularized reinforcement learning (ERRL) [32] (see also Appendix B). For simplicity, we assume here that π_0 is uniform. The optimal policy can be explicitly represented as

$$\pi^\dagger(\mathbf{a}|\mathbf{s}) = \frac{\exp[\beta Q^\dagger(\mathbf{a},\mathbf{s})]}{\sum_{\mathbf{a}}\exp[\beta Q^\dagger(\mathbf{a},\mathbf{s})]},$$

where the optimal Q function $Q^\dagger$ is defined as $Q^\dagger := \max_\pi Q(\mathbf{s},\mathbf{a})$ and

$$Q(\mathbf{s},\mathbf{a}) := \mathbb{E}\left[R(\mathbf{s}_0,\mathbf{a}_0) + \sum_{t=1}^{\infty}\gamma^t R^\pi(\mathbf{s}_t,\mathbf{a}_t)\middle|\begin{matrix}\mathbf{s}_0 & = \mathbf{s}, \\ \mathbf{a}_0 & = \mathbf{a}\end{matrix}\right]. \quad (1)$$

While the optimal policy $\pi^\dagger$ is obtained theoretically, it is not clear how the immune system can implement it. Moreover, because the immune system does not have perfect information of $P(\mathbf{s}'|\mathbf{s},\mathbf{a})$ and $R(\mathbf{s},\mathbf{a})$ *a priori*, the optimal policy should be learned via interactions with pathogens *a posteriori*.

B. Implementation of policy by T helper cell population

Each T helper cell can be characterized by its T cell receptor (TCR) and the types of cytokines secreted, which roughly correspond to the phenotypic subtypes of the Th cells, e.g., Th1 and Th2 [2,33]. We abstractly characterize each Th clone by its TCR and by its interactions with effector cells presumably via cytokines. It should be noted that this definition is not identical to the biological subtypes of Th clones. It should be understood as an approximation. Let K be the number of different Th clones classified according to these criteria, and n_k be the population size of the kth clone. Each clone interacts with each type of antigen with a different strength, which is determined by the affinity of the TCR of the clone to the antigen and also by how the antigen is presented. Because each clone has a unique TCR, the interaction strength $w_{ki} \in \mathbb{R}$ of the kth clone to the ith antigen is the same among all cells of the kth clone. When the antigen pattern is $\mathbf{s}$, each Th cell of type k is supposed to receive stimulation $w_{ki}s_i$ from the ith antigen. The total stimulation that each Th cell of type k receives becomes $\Sigma_i w_{ki}s_i$. In reality, a Th cell encounters antigens probabilistically, and only a fraction of Th cells of each clone encounter antigens. Thus, $\Sigma_i w_{ki}s_i$ should be interpreted as an expected stimulus that each Th cell in a population can receive on average. A Th cell of type k activates itself, and its activity, $h_k(\mathbf{s})$, is assumed to be dependent sigmoidally on the strength of the stimulation as

$$h_k(\mathbf{s}) = \sigma\left(\sum_{i=1}^{N} w_{ki}s_i\right), \quad (2)$$

where σ is the sigmoid function and $\sigma(x) = 1/(1+\exp(-x))$. The activity h_k becomes either 1 or zero when the cell is fully activated or deactivated, respectively. Such monotonous and sigmoidal dependency is consistent with several experimental observations [2].

Depending on their activities, the Th cells release cytokines, which in turn bias the activities of the effector cells. All Th cells of type k are assumed to release the same types of cytokines, because they belong to the same effector subtype. A stimulus to the jth type of effector cells via cytokines released from a Th cell of the kth type on average is expressed as $\beta u_{jk}h_k$, where h_k is the activity of the kth clone and $u_{jk} \in \mathbb{R}$ defines the strength and sign of the stimulus. $\beta \in (0,\infty)$ is a global scaling parameter which can be identified with β introduced in reinforcement learning above.

The integral stimulus received by the jth effector cells from all Th cells is then represented as $\Sigma_k n_k h_k u_{jk}$. In response to this integral stimulus, the probability that the jth type of effector cells is activated ($a_j = 1$) or deactivated ($a_j = 0$) is modulated according to the following conditional probability:

$$p(a_j = 1|\mathbf{s}) = \sigma\left(\sum_{k=1}^{K} \beta u_{jk} n_k h_k(\mathbf{s})\right), \tag{3}$$

where we suppose that the cytokines are the major biasing factors of the activities of effector cells by Th cells. The positive and negative biases may be associated with inflammatory and anti-inflammatory cytokines, respectively. It should be noted that, in this formulation, the effector cells have the autonomous ability to be activated or deactivated, which is biased by the signal from the Th cells. While we can consider that such autonomous activity is directly modulated by the pathogens as in the case of trained immunity, we assume here that the activities of effector cells are modulated only via Th cells for simplicity and for focusing mainly on the roles of the Th cells.

Therefore, the Th cell population translates the antigen pattern $\mathbf{s}$ that it receives into an activation pattern of the effector cells, $\mathbf{a}$, with a probability $\pi_{\mathbf{n}}$:

$$\pi_{\mathbf{n}}(\mathbf{a}|\mathbf{s}) = \prod_{j=1}^{M} p(a_j|\mathbf{s}) = \frac{\exp(\beta \tilde{Q}_{\mathbf{n}}(\mathbf{s}, \mathbf{a}))}{\sum_{\mathbf{a}\in\mathcal{A}} \exp(\beta \tilde{Q}_{\mathbf{n}}(\mathbf{s}, \mathbf{a}))}, \tag{4}$$

where $\tilde{Q}_{\mathbf{n}}$ is defined as

$$\tilde{Q}_{\mathbf{n}}(\mathbf{s}, \mathbf{a}) = \sum_{k=1}^{K} n_k h_k(s) \sum_{j=1}^{M} u_{jk} a_j. \tag{5}$$

This conditional probability is the policy of the immune system implemented by the Th cell population, and the role of the Th cell population is to update the policy over time by modulating the clone-size distribution $\mathbf{n}$ so as to make $\pi_{\mathbf{n}}$ closer to $\pi^{\dagger}$ in order to receive a greater reward. Note that we model the Th clones being homogeneous with respect to their parameters for simplicity and for focusing on general properties of the model. Depending on the objectives, we can include specific subtypes and structures among them as extensions.

C. Learning dynamics of Th cell population

Similarly to $\pi^{\dagger}$, the policy $\pi_{\mathbf{n}}(\mathbf{a}|\mathbf{s})$ is represented in the form of a Boltzmann distribution with respect to $\mathbf{a}$ in which $\tilde{Q}_{\mathbf{n}}(\mathbf{s}, \mathbf{a})$ and β are the negative energy and the global scaling parameter, respectively. Because of this form, the policy π is likely to select an activity pattern of effector cells, $\mathbf{a}$, with a greater value of $\tilde{Q}_{\mathbf{n}}(\mathbf{s}, \mathbf{a})$ than the others. If $\tilde{Q}_{\mathbf{n}}(\mathbf{s}, \mathbf{a})$ represents the value of choosing $\mathbf{a}$ in response to $\mathbf{s}$, the policy $\pi_{\mathbf{n}}(\mathbf{a}|\mathbf{s})$ implemented by the Th cell population can be interpreted as a strategy to choose the activity pattern of a higher $\tilde{Q}_{\mathbf{n}}(\mathbf{s}, \mathbf{a})$ value with higher probability than the others. In terms of maximizing the reward, the immune system should select the activity pattern of the effector cells, $\mathbf{a}$, that returns a higher reward $R(\mathbf{s}, \mathbf{a})$ in response to an antigen pattern $\mathbf{s}$. Therefore, intuitively, the policy $\pi_{\mathbf{n}}(\mathbf{a}|\mathbf{s})$ becomes better when $\tilde{Q}_{\mathbf{n}}(\mathbf{s}, \mathbf{a})$ has been updated to represent the reward $R(\mathbf{s}, \mathbf{a})$ more faithfully. This intuitive interpretation can be rationalized by considering $\gamma = 0$ for Eq. (1). The policy in the Boltzmann form of Eq. (4) is shown to be optimal when $\tilde{Q}_{\mathbf{n}}(\mathbf{s}, \mathbf{a})$ becomes identical to the optimal $Q^{\dagger}$ function, which is equal to $R(\mathbf{s}, \mathbf{a})$ in this case. Therefore, optimizing $\pi_{\mathbf{n}}(\mathbf{a}|\mathbf{s})$ is equivalent to learning $\tilde{Q}_{\mathbf{n}}(\mathbf{s}, \mathbf{a})$, which is an estimate of the reward function $R(\mathbf{s}, \mathbf{a})$, from the past experiences of interactions with pathogens [32].

Therefore, the learning dynamics can be reduced to updating $\tilde{Q}_n(\mathbf{s}, \mathbf{a})$ to be closer to the reward function $R(\mathbf{s}, \mathbf{a})$ than before by modulating the clone-size distribution $\mathbf{n}$. One way to derive such an update rule is to minimize the following cost function with respect to the parameter $\mathbf{n}$ in $\tilde{Q}_{\mathbf{n}}(\mathbf{s}, \mathbf{a})$ for each episode $\mathbf{s}$, $\mathbf{a}$, and $r = R(\mathbf{s}, \mathbf{a})$:

$$L_{\mathbf{n}}(\mathbf{s}, \mathbf{a}) = \tfrac{1}{2}(r - \tilde{Q}_{\mathbf{n}}(\mathbf{s}, \mathbf{a}))^2. \tag{6}$$

$\tilde{Q}_{\mathbf{n}}(\mathbf{s}, \mathbf{a}) = R(\mathbf{s}, \mathbf{a})$ is achieved when this cost function $L_{\mathbf{n}}$ takes the minimum value of zero for all pairs of $(\mathbf{s}, \mathbf{a})$. If a biological learning system were equipped with a versatile memory that could store the experienced rewards for all pairs of $(\mathbf{s}, \mathbf{a})$, $\tilde{Q}_{\mathbf{n}}(\mathbf{s}, \mathbf{a}) = R(\mathbf{s}, \mathbf{a})$ would seem to be achieved trivially. However, such implementation is not feasible either biologically or computationally. Storing such information requires a large memory, the capacity of which is of the order of 2^{M+N}. Moreover, owing to the lack of generalization in this implementation, i.e., the experienced rewards are not exploited to infer rewards of not-yet-experienced pairs of $(\mathbf{s}, \mathbf{a})$, the system needs an extraordinarily long time to experience all pairs. Recent advancements in network-based reinforcement learning have demonstrated that the implementation of the Q function by a neural network is efficient in terms of both memory usage and generalization [30].

The Th cells form a network similar to a neural network (Fig. 1) and can potentially approximate $R(\mathbf{s}, \mathbf{a})$ in the form of $\tilde{Q}_{\mathbf{n}}$ in Eq. (5). However, experimental evidence suggests that Th cells realize learning mainly by adjusting their clone-size distribution $\mathbf{n}$ and that the other parameters such as the weights $\mathbf{w}$ of the Th clone-antigen interactions may not be changed. This is in sharp contrast to the case of neural networks in which the interaction weights can be directly modulated to achieve learning. Thus, it is not obvious whether learning can be achieved in the immune system only by adjusting $\mathbf{n}$.

In addition, a simple update of $\mathbf{n}$ along the gradient of the cost function, $\nabla_{\mathbf{n}} L_{\mathbf{n}}(\mathbf{s}, \mathbf{a})$, may not necessarily be relevant biologically because the learning dynamics should satisfy the invariance constraint with respect to the subdivision of a clone, which is imposed by the interpretation of $\mathbf{n}$ as the clone-size distribution. Suppose that the kth clone accommodates n_k cells and is subdivided into two subclones, k_1 and k_2, as $n_k = n_{k_1} + n_{k_2}$. Such a subdivision should not change the learning dynamics as long as the two subclones have the same properties as those before the subdivision. To satisfy this invariance rule, the following metric of the parameter space $\mathbf{n}$ should be considered (see Appendix C):

$$g_{ij}(\mathbf{n}) = \delta_{ij}/n_i. \tag{7}$$

With this metric, the appropriate gradient can be derived as

$$\left\{\nabla^{g}_{\mathbf{n}} L_n(s, a)\right\}_k = \sum_m g^{-1}_{km} \frac{\partial L_n(\mathbf{s}, \mathbf{a})}{\partial n_m} = n_k \frac{\partial L_n(\mathbf{s}, \mathbf{a})}{\partial n_k}, \tag{8}$$

which is the natural gradient in the parameter space $\mathbf{n}$ [34]. Thus, when the Th population with a clone-size distribution $\mathbf{n}(t)$ at time t experiences an antigen pattern $\mathbf{s}(t)$ and an activation pattern $\mathbf{a}(t)$, the update rule of $\mathbf{n}$ can be derived in the form of a replicator equation as

$$n_k(t+1) = n_k(t) + \alpha\, n_k(t)\lambda_k(t), \tag{9}$$

where the positive constant α is the learning rate,

$$\lambda_k(t) := \frac{\partial L_n(\mathbf{s}, \mathbf{a}, r)}{\partial n_k} \tag{10}$$

$$= [r(t) - \tilde{Q}(\mathbf{s}(t), \mathbf{a}(t))]h_k(\mathbf{s}(t))\sum_j u_{jk}a_j(t), \tag{11}$$

and $r(t) := R(\mathbf{s}(t), \mathbf{a}(t))$. This rule of learning dynamics is similar to that of state action reward state action (SARSA) or Q learning with a linear functional approximation [28]. The details of the dynamics mean that the Th cell population can learn an effective response to each pathogen if each clone of type k proliferates or dies by following the growth rate λ_k. The self-replicative nature of the dynamics originates from the metric $g_{ij}(\mathbf{n})$, whereas the functional form of the growth rate $\lambda_k(t)$ is determined by the gradient of $L_n(\mathbf{s}, \mathbf{a}, r)$. Therefore, the self-replicative nature of Eq. (9) is invariant to changes in the details as long as the Th clone size works as the learning parameter. We note that the clonal selection is derived here as the learning rule. It is an important open problem how the phenotypic switching can be derived as a part of the learning rule.

D. Biological interpretation of learning dynamics

The derived learning dynamics can be interpreted biologically by introducing the following decomposition of the growth rate λ_k of the kth clones:

$$\lambda_k(\mathbf{s}, \mathbf{a}, r) = f_k(\mathbf{s}, \mathbf{a})[r - \tilde{Q}_{\mathbf{n}}(\mathbf{s}, \mathbf{a})], \tag{12}$$

where

$$\tilde{Q}_{\mathbf{n}}(\mathbf{s}, \mathbf{a}) = \sum_l n_l f_l(\mathbf{s}, \mathbf{a}), \quad f_k(\mathbf{s}, \mathbf{a}) := h_k(\mathbf{s})\sum_j u_{jk}a_j. \tag{13}$$

$[r - \tilde{Q}_{\mathbf{n}}(\mathbf{s}, \mathbf{a})]$ is common to all clones and can be interpreted as a global signal to all Th cells. In contrast, $f_k(\mathbf{s}, \mathbf{a})$ determines the clone-specific sensitivity to the global signal. Further, $h_k(\mathbf{s})$ in $f_k(\mathbf{s}, \mathbf{a})$ is the antigen-dependent activity of the kth clone, whereas $\sum_j u_{jk}a_j$ is the feedback from the active effector groups. This indicates that the kth clone has a high sensitivity to the global signal when it receives a strong antigenic signal from the current antigen pattern and also has feedback from the effector groups [3,35]. Such feedback requires local interactions between the Th cells and the effector groups as cytokines mediate the paracrine communications. Biologically, the Th cell population is known to proliferate only when being exposed to both signals, namely, the stimulus to TCR and the co-stimulus from APCs or innate immune cells [2,36]. There are additional pieces of evidence that indicate interactions between the signals from TCR and cytokines that might be released by the effector groups [37,38]. Moreover, proinflammatory cytokines are secreted by myeloid cells to activate naive Th cells [39]. It should be noted that $f_k(\mathbf{s}, \mathbf{a})$ can be negative if the kth clone has an inhibitory effect on certain effector groups. Such a situation might be related to the activation-induced cell death (AICD) of T cells [40,41]. Th cells express Fas ligands as they are activated. While these Fas ligands have no significant effect on inactive T cells, they can induce apoptosis on the activated Th cells, which is considered a mechanism of immune tolerance [40]. Additionally, the negative $f_k(\mathbf{s}, \mathbf{a})$ may also be interpreted as the action of anti-inflammatory cytokines. Our result suggests that these positive and negative back-propagating controls may be responsible for modulating the relative contributions of the T cell clones depending on the consistency between the activities of the Th cells and the effector groups. Moreover, the derived learning rule indicates that such local modulations are not sufficient for learning because the individual Th clones are blind to whether their activities and those of the induced effector groups have actual immunological impact. The global signal $[r - \tilde{Q}_{\mathbf{n}}(\mathbf{s}, \mathbf{a})]$ is indispensable for conveying that information and might be associated with the damage signal [42] or physiological conditions of the body such as temperature or endocrine signals. While the biological interpretations of the local and global signals are not decisive, the learning rule highlights the roles and necessity of these two different types of interactions among immune cells for achieving an appropriate learning.

III. NUMERICAL SIMULATIONS AND CLONE-SIZE DISTRIBUTION AFTER LEARNING

We conducted simulations to confirm that the derived dynamics could actually learn because the steepest descent method does not always guarantee their resulting performance. We also investigated stationary clone-size distributions of the Th cell population to draw insights on the behavior of an appropriately trained learning system and to check the consistency with experimentally observed results. Because individual antigens, Th clones, and effector types are modeled explicitly in our formulation, simulations with realistic parameter values are prohibitive. For example, the varieties of Th clones, K, can be of the order of 10^6 or more [43]. While we lack a reliable estimate, the variety of antigens, N, can be of a similar order as that of K because each antigen is characterized by the peptide sequence of length 8 or 9. The number of effector types, M, should be of a much smaller order because the effector types are determined by genetically encoded cell types and their phenotypic states. To circumvent this difficulty, we instead focus on the general properties of the learning dynamics and stationary distribution for a tractable parameter set with smaller values and investigate the scaling property with respect to changes in the parameter values.

We assume that there are N distinct antigens and P different antigen patterns $\{\mathbf{s}^1, \mathbf{s}^2, \dots, \mathbf{s}^P\}$ representing uninfected and infected states. $\mathbf{s}^1$ corresponds to the antigen pattern of the uninfected state and $\mathbf{s}^i$ corresponds to that of the ith pathogen. We also assume that there are M different types of effector cells and the associated P possible activity patterns $\{\mathbf{a}^1, \mathbf{a}^2, \dots, \mathbf{a}^P\}$, each of which represents the most effective activity pattern of effector cells to the corresponding pathogen. While the possible antigen and activity patterns, in

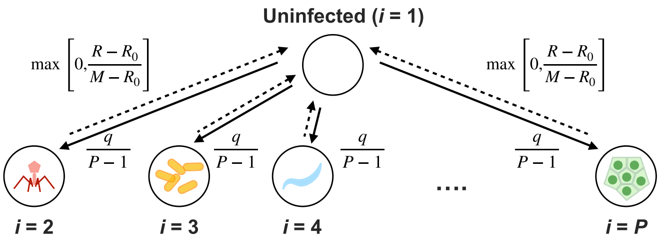

FIG. 2. Schematic diagram of the stochastic transition dynamics of the infected and uninfected states (the environment in the MDP) defined by Eq. (15). In all simulations, $q = 0.9$ and $R_0 = M/2$ are used.

principle, depend on the kind of pathogens considered, such detailed information is experimentally available only for exceptional cases [44]. Thus, both antigen and activity patterns are generated randomly by following the maximum entropy principle under noninformative situations. Depending on our purposes, we can include more specific dynamics and similarities of pathogens in the model. For each $i \in \{1, \dots, P\}$, the reward function $R(\mathbf{s}, \mathbf{a})$ is determined as

$$R(\mathbf{s}^i, \mathbf{a}) = M - \mathrm{ham}(\mathbf{a}^i, \mathbf{a}), \tag{14}$$

where $\mathbf{a}^i$ is the most effective activity pattern for the antigen pattern $\mathbf{s}^i$ and $\mathrm{ham}(\mathbf{a}, \mathbf{a}')$ is the Hamming distance between two binary vectors $\mathbf{a}$ and $\mathbf{a}'$. This functional form indicates that the immune system receives the highest reward M when the activity pattern $\mathbf{a}$ matches the most effective one for the antigen pattern $\mathbf{s}^i$. If $\mathbf{a}$ deviates from the most effective $\mathbf{a}^i$, the immune system experiences a loss of reward by the deviation $\mathrm{ham}(\mathbf{a}^i, \mathbf{a})$. Because M is the maximum Hamming distance for a pair of vectors with length M, $R(\mathbf{s}, \mathbf{a})$ is always positive for any pair of antigen and activity patterns.

The transition probability $P(\mathbf{s}'|\mathbf{s}, \mathbf{a})$ is selected as

$$P(\mathbf{s}^i|\mathbf{s}^j, \mathbf{a}) = \begin{cases} 1 - q & \text{if } j = 1 \text{ and } i = 1 \\ q/(P-1) & \text{if } j = 1 \text{ and } i \neq 1 \\ \min\left[1, \frac{M - R(\mathbf{s}^i, \mathbf{a})}{M - R_0}\right] & \text{if } j \neq 1 \text{ and } i = j \\ \max\left[0, \frac{R(\mathbf{s}^i, \mathbf{a}) - R_0}{M - R_0}\right] & \text{if } j \neq 1 \text{ and } i = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

The transition probability represents the dynamics depicted in Fig. 2. Here, q is the probability to be infected by a pathogen, which is randomly chosen from $P - 1$ pathogens equally. If infected, the pathogen is swept out with the probability $\max\,[0, \frac{R(\mathbf{s}^i, \mathbf{a}) - R_0}{M - R_0}]$. This means that if the reward is its maximum M, the pathogen is eliminated with probability 1. On the other hand, the pathogen cannot be removed if the reward is less than or equal to the threshold R_0. As the reward increases from R_0, the chance of recovery increases linearly. If R_0 is small, the pathogen can be removed with a certain probability even without effective control from the Th cells due to the intrinsic ability of effector cells. If R_0 is close to the maximum reward, the learning is typically hampered by being trapped in one of the infected states. This suggests that the innate ability to recover from infections is a requisite for adaptive learning. We can consider more complicated situations by extending the space of $\mathbf{s}$ according to our purposes. For example, by assigning several states of $\mathbf{s}$ to each pathogen, we can represent the stages of infection, the transitions between which are dependent on the action of the immune system. Such a model can be effectively used to analyze more complicated infections such as chronic ones. We may also model adversarial ones where the next infection is dependent on the current infection due to the coevolution of pathogens (see Appendix D and Fig. 6). The inhomogeneous effectiveness of effector cells on different pathogens can also be incorporated.

Finally, we suppose that the distribution of the interaction strength w_{ki} of the kth clone with the ith antigen follows the normal distribution with mean zero and variance σ_w^2 to represent the cross-reactivity of TCRs [45]. Similarly, the effect of the stimulus on the jth effector cells from the kth clone, u_{jk}, is sampled from a normal distribution with mean zero and variance σ_u^2, because we lack quantitative information on this parameter [6]. It should be noted that the affinity of a TCR towards antigens is expected to be sparse such that a TCR reacts to a small fraction of all possible antigens. Even if this sparsity is considered, the following results are not affected qualitatively (see also Appendix E and Fig. 7). In contrast to the training of neural networks, these interaction parameters are fixed in our model and the Th clone size, $\mathbf{n}$, is the only tunable parameter for learning.

Figures 3(a) and 3(b) show the transient dynamics of the Th clone sizes and the rank-abundance distribution during a learning process starting from the uniform clone-size distribution. We observe that the clone-size distribution fluctuates transiently during the learning with switching of ranks of the clones, and $\mathbf{n}(t)$ eventually converges into a stationary distribution. The early fluctuation is due to the small β in the learning process (Appendix A). This early fluctuation promotes exploration of the system, which might be related to the downregulation of the Th function in the early infancy periods [46]. It should be noted that, in a real biological situation, the learning also starts with the Th clone distribution pretrained in the thymus. Such pretraining may be optimized to facilitate and expedite subsequent learning, possibly by evading the very early exploring stage of the learning [47]. Figure 3(c) shows the statistics of 100 independent learning curves and their fluctuations. The monotonous increase in the average reward demonstrates that the derived learning dynamics can actually work to obtain a greater reward over time by updating the clone-size distribution $\mathbf{n}(t)$ based on previous experiences. Figure 3(d) shows the stationary rank-abundance distributions for the 100 independent learning trials. Owing to the stochastic nature of the learning process, the rank-abundance distribution does not perfectly converge into an identical distribution; instead it fluctuates, which is prominent in the abundances of the highly ranked clones (i.e., Rank < 100) in Fig. 3(d).

The simulations are conducted using a set of parameter values chosen as a representative situation in which learning is effectively achieved with minimum diversities of the antigens and Th cells (see Appendix F). As shown in Figs. 8(a) and 8(b), further increase in either N or K does not improve the performance considerably, which indicates that the diversities of the antigens and clone types are sufficiently large under this condition. In contrast, the performance starts declining if either P or M increases [Figs. 8(c) and 8(d)]. The decline induced by the increase in the number of pathogen types, P,

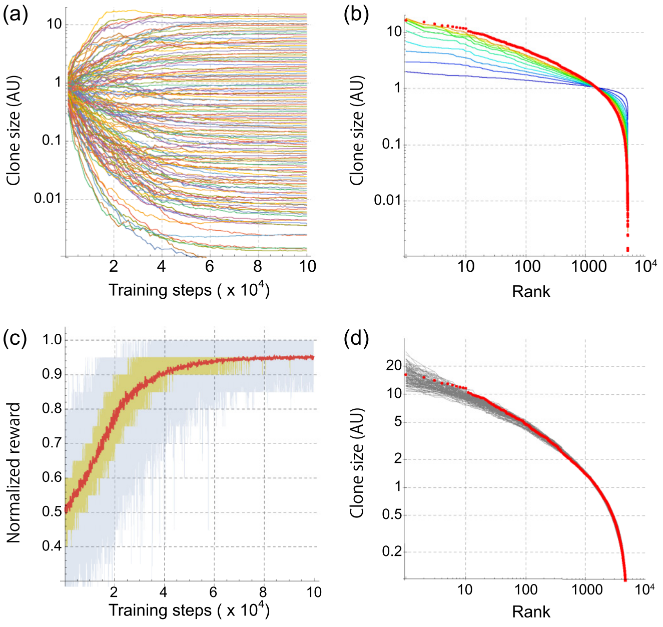

FIG. 3. (a) Trajectories of the Th clone size $\mathbf{n}$ along a learning trial. The parameter values are $N = 100$, $K = 5000$, $M = 20$, and $P = 30$. The clone size $\mathbf{n}$ is normalized by the initial abundance. This figure shows only 200 trajectories sampled evenly out of 5000 different clones at the stationary state to avoid complication of the plot. (b) The dynamics of rank-abundance distributions along the learning trial, which were calculated from the trajectories of $\mathbf{n}$ in (a). The red dots represent the stationary distribution after the learning, and the colored curves are the transient distributions calculated at training steps from 1×10^3 (blue) to 28×10^3 (yellow). (c) Statistics of learning curves of the Th cell population. The rewards normalized by its maximum value are shown as functions of the training step for 100 independent learning trials. The red curve is the average reward, and the yellow and blue regions show the range between 25th and 75th percentiles of the rewards and that between minimum and maximum of the rewards at each training step, respectively. (d) The stationary rank-abundance distributions for 100 independent learning trials in (c) are shown by gray curves. The red dots are the same as those in (b). See also Appendix A.

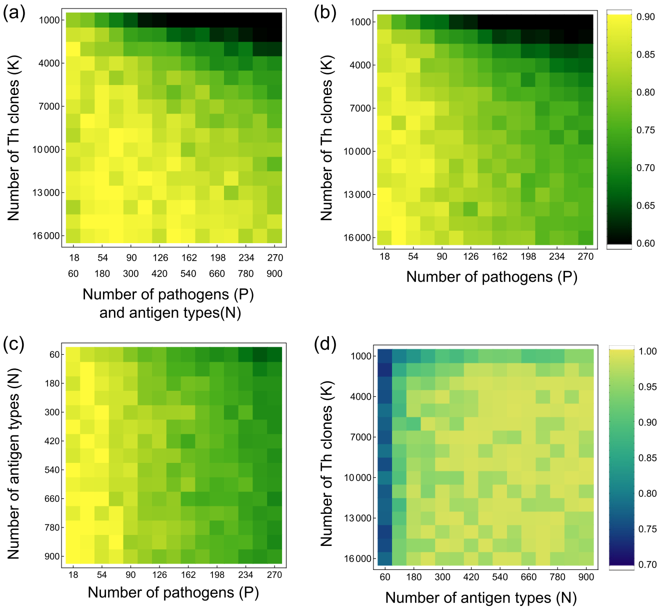

FIG. 4. Heat map plots of the stationary normalized reward after learning as functions of (a) K and $\{N, P\}$, (b) K and P, (c) N and P, and (d) K and N. The other parameters are the same as those in Fig. 3. The stationary reward was calculated as the moving average of the last 10^4 steps. (a)–(c) use the same color code.

is natural because learning more pathogens should be more difficult if the numbers of antigens and Th clone types are fixed. In contrast, the decline due to the increased M highlights the importance of constraining possible actions for an efficient learning [48].

To investigate how N, K, and P can be scaled to a greater size while maintaining learning performance, we calculated the stationary reward after learning by changing N, K, and P in Fig. 4. Figure 4(a) shows that the performance is approximately kept constant when N, K, and P are scaled as ξN, ξK, and ξP, where ξ is the scaling parameter. In contrast, if only either antigen diversity N or Th clone diversity K is increased while the other is kept constant, as in Figs. 4(b) and 4(c), the increased variety of pathogens (larger P) cannot be handled, which indicates the importance of both diversities for learning. This property should be linked to the learning capacity of the network, which has been intensively analyzed for deep neural networks [49].

Next, we investigate the effect of parameters on the shapes of the abundance distributions of the Th clones when the learning is conducted appropriately. For the set of parameter values in Fig. 3, the rank abundance distribution in the log-log plot is relatively flat for abundant clones (from rank 10 to 10^3, approximately) but shows a sharp decline in clone size for less abundant clones (rank $>10^3$), which results in a concave distribution. The sharp decline in the abundance is mainly due to the limited number of Th clone types, K, which works as a boundary condition for the rank-abundance distribution. By conducting learning for a larger K (i.e., $K = 10^5$) as in Fig. 5(a), we observe that the relatively flat region stretches, which enhances the power-law-like property of the distribution as shown in Fig. 5(b). If the total number of Th cells or observed samples is limited, very low abundance clones at the boundary are rarely observed, which can lead to flat clone-size and rank-abundance distributions, as shown in Figs. 5(c) and 5(d). Such flat distributions have been observed in several TCR sequencing experiments [17,50–52] even though Th clone types are discriminated only by the TCR sequences in the experiments. To compare the abundance distributions obtained using our model with the experimental ones, we use the TCR sequences of CD4+ T cells collected from peripheral blood of two healthy human donors (HV01 and HVD4) reported in Ref. [53]. In this data set, molecular barcodes were added to each cDNA molecule to correct for PCR amplification and errors, which can lead to more accurate and quantitative estimates. The coincidence seems to be fairly good especially with Data 2 [Figs. 5(c) and 5(d)]. We also note that qualitatively the same result was obtained for sparse $\{w_{ki}\}$ (Fig. 7).

This coincidence implies that a simple mechanism underlies the generation of power-law-like distributions irrespective of the details of the dynamics. In previous theoretical

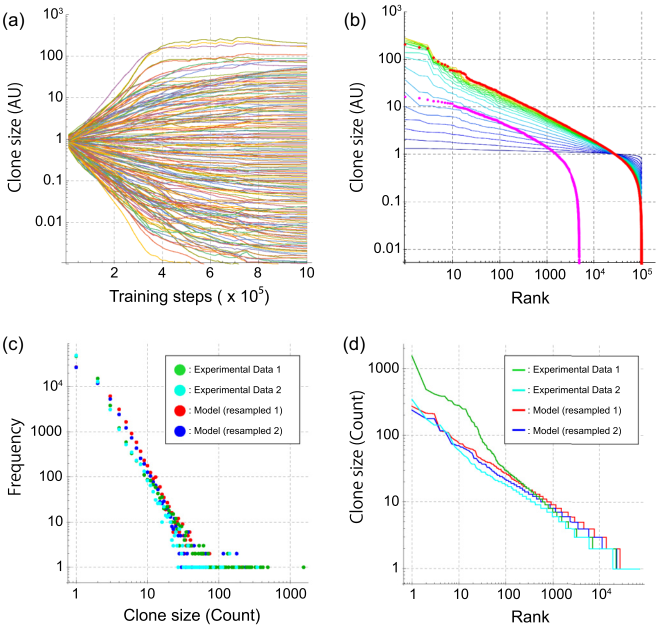

FIG. 5. (a) Trajectories of the Th clone size **n** along a learning trial for $K = 100\,000$. The weights, **w**, are also scaled to **w**/10 for comparison with the experimental data. The other parameter values are the same as those in Fig. 3. The clone size **n** is normalized by the initial abundance. The figure shows only 200 trajectories sampled evenly out of 100 000. (b) The dynamics of the rank-abundance distributions along the learning trial, which were calculated from the trajectories of **n** in (a). The red dots represent the stationary distribution after the learning, and the colored curves are the transient distributions calculated at training steps from 1×10^4 (blue) to 77×10^4 (yellow). The magenta dots show the same stationary distribution as in Fig. 3(b), which is shown here for comparison. (c) Clone-size distributions and (d) rank-abundance distributions obtained from the model and an experiment. The green and cyan points in (c) are the clone-size distributions obtained by counting the TCR sequences of CD4+ T cells collected from peripheral blood of two healthy human donors (HV01 and HVD4) in Ref. [53], and the curves in (d) are the corresponding rank-abundance distributions. Red and blue points in (c) are the clone-size distributions obtained by resampling the clones from the rank distributions in (b) for the same numbers of total counts as in experiments 1 (red) and 2 (blue), and the red and blue curves in (d) are the corresponding rank-abundance distributions.

analyses, the symmetric variation in the fitness of clones or external fluctuating stimuli to fitness was proposed as the mechanism of the power-law distribution [54–56]. The symmetric variation means that fitness fluctuation of clones is biased to neither high fitness nor low fitness by assuming symmetric distribution such as a Gaussian distribution. Similar symmetric fitness variation is also observed in Fig. 5(a), indicating that our model shares the same property as the previous ones under this learning condition (see also Appendix G).

The next question is the mechanism of the symmetric fitness variation, which was just assumed in the previous analyses. Our model demonstrates that such a variation is not an automatic consequence of efficient learning. Our result suggests that the symmetric variation can appear when Th clones have far more diversity than minimally required for achieving efficient learning under a given pathogen diversity (Fig. 9). If the pathogen diversity is far beyond the capacity of the Th diversity, a part of the Th clones dominates in fitness over the others, which results in asymmetric fitness variation (Fig. 9). While this problem is still open both theoretically and experimentally, our learning framework may provide a perspective from the viewpoint of computational and algorithmic levels.

IV. SUMMARY AND DISCUSSION

The learning dynamics of the adaptive immune system has not yet been fully understood due to the lack of approaches at the computational and algorithmic levels despite the accumulated evidence at the implementation level. Based on the framework of network-based reinforcement learning, we constructed a mathematical model, which may bridge this gap. From our model, the clonal selection of Th cells is naturally derived as a learning rule, which enables the system not only to recognize new pathogens but also to acquire the appropriate way to bias the responses to the pathogens. Even though the simulations were conducted under an abstract and simple situation, we found a good scaling property among K, N, and P, which enables us to extrapolate our results to a more realistic scale. In addition, our model could successfully reproduce the experimental clone-size distributions to a certain extent when a sufficiently diverse Th clone type was assumed.

Besides these results, our model still has room for accommodating more detailed quantitative information, if provided in the future, to make the simulation more realistic for more specific purposes. For example, the dynamics of antigen patterns can be more detailed for describing specific infectious and pathological situations. We may represent chronic infections by introducing hidden states in the dynamics of antigen patterns. For such cases, more predictive behaviors with $\gamma > 0$ might be related to actual immunological learning. The problem of cross immunity may be addressed by considering how the system can generalize the past experiences. We can also introduce pretraining to mimic and analyze the thymic selection [57]. Furthermore, if comprehensive quantitative data on the interactions between antigens and Th clones are obtained by future measurement technologies [58–61], we can include that information on the weights of the network.

Nonetheless, we acknowledge that there are several discrepancies between the actual immune system and our mathematical model. First, the dynamics of the pathogen is implicit in our model because the framework of the MDP requires the agent (Th cell population) to be accessible to the environmental state [28,29,62]. Such a problem can be addressed by extending the model to the partially observable MDP [29]. Second, the Th clones should be classified explicitly by the TCR and phenotypic state to directly compare the simulation with the experimental data. Third, while the derived learning dynamics was qualitatively consistent with the clonal selection theory, the local feedback interactions from the effector cells to the Th cells should be associated with actual cell types and interacting molecules [37,38]. Similarly, the biological counterpart of the global signal and reward should be identified. Because the system cannot learn without the global reward signal, its identification can be a pivotal target for the verification of theoretical prediction. In addition, the effector cells have an innate ability to recognize and respond to pathogens. Such an effect is abstractly represented by the stochastic activation and inactivation of the effector cells in our model and is important for preventing the learning from becoming stuck in a certain infectious state. We may improve our model to involve more detailed and active behaviors of the effector cells as self-supporting agents. Such a hierarchical architecture resembles the memetic algorithm used for optimization [63], and its investigation may deepen our understanding of the interrelationship between innate and adaptive immunity.

Finally, while our model can share the characteristic feature of experimentally observed clone-size distributions, it is not yet clear how the feature is related to the general property of learning dynamics. Revealing the general aspects of abstract learning systems is also essential for understanding both universal and problem-specific properties of the immune system.

ACKNOWLEDGMENTS

We thank Yuki Sughiyama and Ryo Yokota for fruitful discussions. We also greatly appreciate the comments and suggestions by the first anonymous referee who reviewed this work. This research is supported by Japan Society for the Promotion of Science (JSPS) Grants No. 19H05799, No. 18H04814, and No. 18H05096, and by Japan Science and Technology Agency (JST) CREST Grant No. JPMJCR2011, and PRESTO Grant No. JPMJPR15E4, Japan.

T.K. and T.J.K. designed the research, conducted mathematical analysis and numerical simulations, and wrote the paper.

APPENDIX A: SIMULATION

The simulation starts with a set of initializations. P different realizable antigen patterns are initialized by random selections of N-dimensional binary vectors. Additionally, the P most effective activity patterns of the effector cells are initialized by randomly selecting M-dimensional binary vectors. The weights of the TCR-antigen interactions, $\mathbf{w}$, are generated randomly by sampling each element of the matrix from a normal distribution $\mathcal{N}(0,\sigma_w^2)$, where the variance $\sigma_w^2 = 2/N$ is determined by the He normal initialization method that is widely used in the context of deep learning [64]. Each element in the weights, $\mathbf{u}$, representing the strength of the signals from the Th clones to the effector cells is also sampled from a normal distribution $\mathcal{N}(0,\sigma_u^2)$, where $\sigma_u^2 = 2/K$. The initial population size of each clone is uniformly set to 1. The initial antigen pattern $\mathbf{s}(t)$ at $t = 0$ is chosen uniformly at random.

The simulation was conducted by iterating the following steps. At each time step t, the activation pattern of the effector cells, $\mathbf{a}(t)$, was determined by sampling from the policy $\pi_{\mathbf{n}(t)}(\mathbf{a}(t)|\mathbf{s}(t))$ calculated based on the antigen pattern $\mathbf{s}(t)$. The calculation of the policy depends on the global scaling parameter $\beta(t)$, which was gradually increased from 1.0 to 20.0 linearly towards the end of the iterations. Based on the activation pattern of the effector cells, $\mathbf{a}(t)$, the reward $r(t)$ was determined as shown in Eq. (14). The population size of each clone changes according to Eq. (9) with a learning rate of $\alpha = 0.1$. If a clone size becomes lower than zero, which is possible due to the time discretization in the simulation, the clone size is set to be zero. Finally, the subsequent iterations were started after sampling the next antigen pattern $\mathbf{s}(t+1)$ from the transition probability [Eq. (15)].

All simulations were implemented either in MATLAB (R2018b, The MathWorks, Natick, MA) or PYTHON using the standard scientific libraries NUMPY and SCIPY.

APPENDIX B: OUTLINE OF ENTROPY-REGULARIZED MDP AND RL

In order to outline the entropy-regularized Markov decision process (ERMDP) and ERRL, we first introduce the conventional MDP and RL [65], and then extend them to the ERMDP and ERRL [66].

1. A brief introduction of conventional MDP and RL

In this section, we briefly introduce the conventional and entropy-regularized MDP and RL to supplement the background knowledge for their use to understand adaptive immune systems in the main text.

Similarly to the definition of MDP in the main text, let $\boldsymbol{s} \in \{0,1\}^N$ and $\boldsymbol{a} \in \{0,1\}^M$ be the state of the environment and the action of the agent (system) we are focusing on. Suppose that $P(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})$ is the Markov transition probability of the environment from $\boldsymbol{s}$ to $\boldsymbol{s}'$ when the action $\boldsymbol{a}$ is taken by the agent. We also define $R(\boldsymbol{s}';\boldsymbol{a},\boldsymbol{s})$ to be the reward that the agent obtains when the transition $\boldsymbol{s} \rightarrow \boldsymbol{a} \rightarrow \boldsymbol{s}'$ occurs. Note that $R(\boldsymbol{s}';\boldsymbol{a},\boldsymbol{s})$ has a more general form here than that in the main text, because R can depend not only on $\boldsymbol{a}$ and $\boldsymbol{s}$ but also on $\boldsymbol{s}'$. Finally, $\pi_t(\boldsymbol{a}_t|\boldsymbol{s}_t)$ is the policy of the agent to choose the action $\boldsymbol{a}_t$ when the environment is $\boldsymbol{s}_t$ at time t with probability $\pi_t(\boldsymbol{a}_t|\boldsymbol{s}_t)$. The explicit dependency on t will be removed later when we consider $t \rightarrow \infty$.

In order to relate the adaptive immune system with MDP in the main text, we associate $\boldsymbol{s}$ with the antigen pattern and $\boldsymbol{a}$ with the activity pattern of effector cells. Then $P(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})$ can be interpreted as the stochastic law of the change in the antigen pattern induced by the action of the effector cells and dynamics of the pathogens in our body. Also, $\pi_t(\boldsymbol{a}_t|\boldsymbol{s}_t)$ characterizes how Th cells control and bias the activities of the effector cells when exposed to an antigen pattern $\boldsymbol{s}_t$.

The MDP is the basis of optimal control and learning problems. If all the ingredients above are given explicitly and fixed, the MDP simply defines a stochastic joint dynamics of $\boldsymbol{s}$ and $\boldsymbol{a}$. The joint path probability of the action history $\mathcal{A}_{\tau:t} := \{\boldsymbol{a}_{t'}|t' \in [\tau,t]\}$ and the state history $\mathcal{S}_{\tau:t} := \{\boldsymbol{s}_{t'}|t' \in [\tau,t]\}$ is obtained by using $P(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})$ and $\pi_t(\boldsymbol{a}|\boldsymbol{s})$ as

$$\begin{aligned}&\mathbb{P}[\mathcal{A}_{1:t-1},\mathcal{S}_{1:t}|\boldsymbol{a}_0,\boldsymbol{s}_0]\\ &\quad := \mathbb{P}[\mathcal{S}_{1:t}||\mathcal{A}_{1:t-1},\boldsymbol{a}_0,\boldsymbol{s}_0]\Pi[\mathcal{A}_{1:t-1}||\mathcal{S}_{1:t-1}],\end{aligned}$$

where

$$\mathbb{P}[\mathcal{S}_{\tau+1:t}||\mathcal{A}_{\tau+1:t-1},\boldsymbol{a}_\tau,\boldsymbol{s}_\tau] := \prod_{t'=\tau}^{t-1} P(\boldsymbol{s}_{t'+1}|\boldsymbol{s}_{t'},\boldsymbol{a}_{t'}),$$

$$\Pi[\mathcal{A}_{\tau:t-1}||\mathcal{S}_{\tau:t-1}] := \prod_{t'=\tau}^{t-1} \pi_{t'}(\boldsymbol{a}_{t'}|\boldsymbol{s}_{t'}).$$

Similarly, we define

$$\mathbb{P}[\mathcal{A}_{0:t-1}, \mathcal{S}_{1:t}|\boldsymbol{s}_0] := \mathbb{P}[\mathcal{S}_{1:t}||\mathcal{A}_{1:t-1}, \boldsymbol{a}_0, \boldsymbol{s}_0]\Pi[\mathcal{A}_{0:t-1}||\mathcal{S}_{0:t-1}].$$

In the case of optimal control theory, by contrast, we are given all but the policy $\pi(\boldsymbol{a}|\boldsymbol{s})$. The goal of the optimal control is to find out the optimal policy $\pi^\dagger(\boldsymbol{a}|\boldsymbol{s})$ that maximizes the following expected cumulative reward:

$$\mathbb{J}_{0:t}[\pi] := \langle \mathbb{R}_\gamma[\mathcal{A}_{0:t-1}, \mathcal{S}_{0:t}]\rangle_{\mathbb{P}[\mathcal{A}_{0:t-1}, \mathcal{S}_{1:t}|\boldsymbol{s}_0]p(\boldsymbol{s}_0)},$$

where $p(\boldsymbol{s}_0)$ is the initial distribution of $\boldsymbol{s}_0$, and

$$\mathbb{R}_\gamma[\mathcal{A}_{\tau:t-1}, \mathcal{S}_{\tau:t}] := \sum_{t'=\tau}^{t-1} \gamma^{t'-\tau} R(\boldsymbol{s}_{t'+1}; \boldsymbol{a}_{t'}, \boldsymbol{s}_{t'})$$

is the pathwise reward function with discount rate $0 \leqslant \gamma \leqslant 1$, and $\langle f(x)\rangle_{p(x)}$ means the average of a function $f(x)$ with respect to a probability measure $p(x)$.

$$\max_{\{\pi_{t'}\}_{t'\in[0,t-1]}} \mathbb{J}_{0:t}[\pi] := \left\langle \max_{\pi_0} \left\langle \max_{\{\pi_{t'}\}_{t'\in[1,t-1]}} Q^\pi_{0:t}(\boldsymbol{a}_0, \boldsymbol{s}_0) \right\rangle_{\pi_0(\boldsymbol{a}_0|\boldsymbol{s}_0)} \right\rangle_{p(\boldsymbol{s}_0)},$$

where we define the Q function:

$$Q^\pi_{\tau:t}(\boldsymbol{a}_\tau, \boldsymbol{s}_\tau) := \langle \mathbb{R}_\gamma[\mathcal{A}_{\tau:t-1}, \mathcal{S}_{\tau:t}]\rangle_{\mathbb{P}[\mathcal{A}_{\tau+1:t-1}, \mathcal{S}_{\tau+1:t}|\boldsymbol{a}_\tau, \boldsymbol{s}_\tau]}.$$

The Q function can be interpreted as the expected and discounted cumulative reward up to time t if the agent took the action $\boldsymbol{a}_\tau$ when the environmental state was $\boldsymbol{s}_\tau$ at time τ. Thus, the Q function measures the future value of taking action $\boldsymbol{a}_\tau$ when the state is $\boldsymbol{s}_\tau$.

The Q function is known to satisfy the Bellman equation:

$$Q^\pi_{0:t}(\boldsymbol{a}_0, \boldsymbol{s}_0) = \left\langle R(\boldsymbol{s}_1; \boldsymbol{a}_0, \boldsymbol{s}_0) + \gamma \left\langle Q^\pi_{1:t}(\boldsymbol{a}_1, \boldsymbol{s}_1)\right\rangle_{\pi_1(\boldsymbol{a}_1|\boldsymbol{s}_1)}\right\rangle_{P(\boldsymbol{s}_1|\boldsymbol{s}_0, \boldsymbol{a}_0)}.$$

Similarly, by defining the optimal Q function as

$$Q^\dagger_{\tau:t}(\boldsymbol{a}_\tau, \boldsymbol{s}_\tau) := \max_{\{\pi_{t'}\}_{t'\in[\tau+1,t-1]}} Q^\pi_{\tau:t}(\boldsymbol{a}_\tau, \boldsymbol{s}_\tau),$$

we can obtain the Bellman optimality equation,

$$\begin{aligned} Q^\dagger_{0:t}(\boldsymbol{a}_0, \boldsymbol{s}_0) &= \left\langle R(\boldsymbol{s}_1; \boldsymbol{a}_0, \boldsymbol{s}_0) + \gamma \max_{\pi_1} \left\langle \max_{\{\pi_t\}_{t'\in[2,t-1]}} Q^\pi_{1:t}(\boldsymbol{a}_1, \boldsymbol{s}_1) \right\rangle_{\pi_1(\boldsymbol{a}_1|\boldsymbol{s}_1)} \right\rangle_{P(\boldsymbol{s}_1|\boldsymbol{s}_0, \boldsymbol{a}_0)} \\ &= \left\langle R(\boldsymbol{s}_1; \boldsymbol{a}_0, \boldsymbol{s}_0) + \gamma \max_{\pi_1} \langle Q^\dagger_{1:t}(\boldsymbol{a}_1, \boldsymbol{s}_1)\rangle_{\pi_1(\boldsymbol{a}_1|\boldsymbol{s}_1)} \right\rangle_{P(\boldsymbol{s}_1|\boldsymbol{s}_0, \boldsymbol{a}_0)} \\ &= \left\langle R(\boldsymbol{s}_1; \boldsymbol{a}_0, \boldsymbol{s}_0) + \gamma \max_{\boldsymbol{a}_1} Q^\dagger_{1:t}(\boldsymbol{a}_1, \boldsymbol{s}_1) \right\rangle_{P(\boldsymbol{s}_1|\boldsymbol{s}_0, \boldsymbol{a}_0)}. \end{aligned}$$

For the case of the infinite time horizon situation where $t \to \infty$, we have

$$Q^\pi(\boldsymbol{a}, \boldsymbol{s}) = \left\langle R(\boldsymbol{s}'; \boldsymbol{a}, \boldsymbol{s}) + \gamma \langle Q^\pi(\boldsymbol{a}', \boldsymbol{s}')\rangle_{\pi(\boldsymbol{a}'|\boldsymbol{s}')}\right\rangle_{P(\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{a})}$$

and

$$Q^\dagger(\boldsymbol{a}, \boldsymbol{s}) = \left\langle R(\boldsymbol{s}'; \boldsymbol{a}, \boldsymbol{s}) + \gamma \max_{\boldsymbol{a}'} Q^\dagger(\boldsymbol{a}', \boldsymbol{s}')\right\rangle_{P(\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{a})}.$$

Then, the optimal policy can be obtained as the deterministic policy:

$$\begin{aligned} \pi^\dagger(\boldsymbol{a}|\boldsymbol{s}) &:= \arg\max_\pi \langle Q^\dagger(\boldsymbol{a}, \boldsymbol{s})\rangle_{\pi(\boldsymbol{a}|\boldsymbol{s})} \\ &= \begin{cases} 1, & \boldsymbol{a} = \arg\max_{\boldsymbol{a}} Q^\dagger(\boldsymbol{a}, \boldsymbol{s}) \\ 0 & \text{otherwise,} \end{cases} \end{aligned} \tag{B1}$$

where the optimal policy chooses the action $\boldsymbol{a}^\dagger(\boldsymbol{s}) := \arg\max_{\boldsymbol{a}} Q^\dagger(\boldsymbol{a}, \boldsymbol{s})$, with probability 1, that maximizes the optimal Q function $Q^\dagger(\boldsymbol{a}, \boldsymbol{s})$ for the current environmental state $\boldsymbol{s}$. By inserting this optimal policy into the Bellman optimality equation, we have

$$Q^\dagger(\boldsymbol{a}, \boldsymbol{s}) = \left\langle R(\boldsymbol{s}'; \boldsymbol{a}, \boldsymbol{s}) + \gamma Q^\dagger(\boldsymbol{a}^\dagger(\boldsymbol{s}'), \boldsymbol{s}')\right\rangle_{P(\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{a})}. \tag{B2}$$

By solving this nonlinear equation with respect to $Q^\dagger(\boldsymbol{a}, \boldsymbol{s})$, we can obtain both $Q^\dagger(\boldsymbol{a}, \boldsymbol{s})$ and $\boldsymbol{a}^\dagger(\boldsymbol{s})$, because we are supposed to know the explicit functional form of $R(\boldsymbol{s}'; \boldsymbol{a}, \boldsymbol{s})$ and $P(\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{a})$ in the setting of the optimal control problem.

In the case of RL, however, we address the situation that we do not know the explicit functional forms of $R(\boldsymbol{s}'; \boldsymbol{a}, \boldsymbol{s})$ and $P(\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{a})$. Instead, the agent is given the value r' of $R(\boldsymbol{s}'; \boldsymbol{a}, \boldsymbol{s})$ when the agent took an action $\boldsymbol{a}$ when the environment was $\boldsymbol{s}$ and then changes to $\boldsymbol{s}'$. Thus, the agent has to infer the optimal Q function, $Q^\dagger(\boldsymbol{a}, \boldsymbol{s})$, from the past history of the action and state pairs, $\{\mathcal{A}_{0:t}, \mathcal{S}_{0:t}\}$, and the reward samples $\mathcal{R}_{0:t} := \{R(\boldsymbol{s}_{t'+1}; \boldsymbol{a}_{t'+1}, \boldsymbol{s}_{t'})|t' \in [0, t-1]\}$.

To this end, we typically approximate $Q^{\dagger}(\boldsymbol{a}, \boldsymbol{s})$ with a tentative function $\tilde{Q}^{k}(\boldsymbol{a}, \boldsymbol{s})$ where k represents the number of the update of this function. We have a collection of algorithms to update $\tilde{Q}^{k}(\boldsymbol{a}, \boldsymbol{s})$ so that $\tilde{Q}^{k}(\boldsymbol{a}, \boldsymbol{s})$ is expected to converge to $Q^{\dagger}(\boldsymbol{a}, \boldsymbol{s})$ as $\tau \to \infty$. One such example is to update $\tilde{Q}^{k}(\boldsymbol{a}, \boldsymbol{s})$ to $\tilde{Q}^{k+1}(\boldsymbol{a}, \boldsymbol{s})$ so that a loss function defined based on Eq. (B2),

$$\mathcal{L} := \left[\tilde{Q}^{k}(\boldsymbol{a}, \boldsymbol{s}) - r' - \gamma \max_{\boldsymbol{a}'} \tilde{Q}^{k}(\boldsymbol{a}', \boldsymbol{s}')\right]^{2}, \qquad \text{(B3)}$$

is reduced where $\boldsymbol{s}$, $\boldsymbol{a}$, $\boldsymbol{s}'$, and r' are obtained as the sample. If the possible state and action spaces are not large and we have a versatile memory storage, we can represent $\tilde{Q}^{k}(\boldsymbol{a}, \boldsymbol{s})$ by a table. For the case that the possible state and action are huge, we have recently witnessed the success of network-based reinforcement learning in which the function $\tilde{Q}^{k}(\boldsymbol{a}, \boldsymbol{s})$ is approximated by a neural network (NN). Also, the representation by a NN can benefit from the generalization power of the NN.

Another issue that arises in RL is the choice of policy during learning of the optimal Q function. Because the optimal policy $\boldsymbol{a}^{\dagger}(\boldsymbol{s}) = \arg\max_{\boldsymbol{a}} Q^{\dagger}(\boldsymbol{a}, \boldsymbol{s})$ is defined based on the optimal Q function, we cannot use it during the learning. If we naively use the policy $\boldsymbol{a}^{k}(\boldsymbol{s}) = \arg\max_{\boldsymbol{a}} \tilde{Q}^{k}(\boldsymbol{a}, \boldsymbol{s})$, which maximizes the tentative Q function $\tilde{Q}^{k}(\boldsymbol{a}, \boldsymbol{s})$, the learning can be easily stuck by a local minimum, because $\tilde{Q}^{k}(\boldsymbol{a}, \boldsymbol{s})$ is not accurate at all, especially at the early stage of the learning. To obtain a better estimate of $Q^{\dagger}(\boldsymbol{a}, \boldsymbol{s})$, the agent should experience a variety of $(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s}', r')$ samples by exploring the action-state space. Several heuristic approaches have been proposed to encourage the agent to choose exploring actions rather than the exploiting action that maximizes the tentative $\tilde{Q}^{k}(\boldsymbol{a}, \boldsymbol{s})$ function. One such approach is the soft-max policy where the next action $\boldsymbol{a}$ is stochastically chosen by following the Boltzmann-type distribution

$$\tilde{\pi}^{k}(\boldsymbol{a}|\boldsymbol{s}) = \frac{\exp[\beta \tilde{Q}^{k}(\boldsymbol{a}, \boldsymbol{s})]}{\sum_{\boldsymbol{a}} \exp[\beta \tilde{Q}^{k}(\boldsymbol{a}, \boldsymbol{s})]}.$$

Under this heuristic policy, the agent is expected to balance choosing better action based on a current estimate of Q function (exploitation) and the other actions (exploration).

2. A brief introduction of entropy-regularized MDP and RL

One drawback of the conventional RL is that the way to update the Q function and the way to choose the policy are not theoretically linked. In addition, when we apply RL to biological systems, we observe that biological agents may not always choose their action deterministically even when trained enough. One potential reason why biological systems do not behave deterministically is that the biological systems are stochastic by nature and that controlling them deterministically may require a large cost of control. Thus, we should consider such cost of control when we work on biological systems

The entropy-regularized MDP (ERMDP) is an approach to bridge the gap we mentioned and also to incorporate cost of control in a stochastic system. In ERMDP, we additionally consider an entropic control cost $\ln \frac{\pi(\boldsymbol{a}|\boldsymbol{s})}{\pi_r(\boldsymbol{a}|\boldsymbol{s})}$ for each action $\boldsymbol{a}$ given $\boldsymbol{s}$ where $\pi_r(\boldsymbol{a}|\boldsymbol{s})$ is a reference policy. By considering the pathwise entropic cost,

$$\mathbb{D}_{\gamma}[\mathcal{A}_{\tau:t-1}, \mathcal{S}_{\tau:t-1}] := \sum_{t'=\tau}^{t-1} \gamma^{t'-\tau} \ln \frac{\pi_{t'}(\boldsymbol{a}_{t'}|\boldsymbol{s}_{t'})}{\pi_r(\boldsymbol{a}_{t'}|\boldsymbol{s}_{t'})},$$

we modify the expected cumulative reward as

$$\mathbb{J}_{0:t}[\pi] := \Big\langle \mathbb{R}_{\gamma}[\mathcal{A}_{0:t-1}, \mathcal{S}_{0:t}] - \frac{1}{\beta}\mathbb{D}_{\gamma}[\mathcal{A}_{0:t-1}, \mathcal{S}_{0:t-1}] \Big\rangle_{\mathbb{P}[\mathcal{A}_{0:t-1}, \mathcal{S}_{1:t}|\boldsymbol{s}_0]p(\boldsymbol{s}_0)}.$$

One interpretation of the entropic cost is that $\mathbb{D}_{\gamma}[\mathcal{A}_{0:t-1}, \mathcal{S}_{0:t-1}]$ accounts for the control cost to drive the agent away from its intrinsic and reference behavior $\pi_r(\boldsymbol{a}|\boldsymbol{s})$. Biological systems typically show autonomous stochastic behaviors even when not being controlled. For the case of the immune system, for example, the innate immunity has intrinsic ability to respond to pathogens without the control by the adaptive immunity. The entropic cost can flexibly represent such information. Also, as shown below, we can derive the soft-max policy as the optimal policy for ERMDP.

By defining the entropy-regularized Q and the optimal entropy-regularized Q functions as

$$\mathbb{Q}^{\pi}_{\tau:t}(\boldsymbol{a}_{\tau}, \boldsymbol{s}_{\tau}) := \Big\langle \mathbb{R}_{\gamma}[\mathcal{A}_{\tau:t-1}, \mathcal{S}_{\tau:t}] - \frac{\gamma}{\beta}\mathbb{D}_{\gamma}[\mathcal{A}_{\tau+1:t-1}, \mathcal{S}_{\tau+1:t-1}] \Big\rangle_{\mathbb{P}[\mathcal{A}_{\tau+1:t-1}, \mathcal{S}_{\tau+1:t}|\boldsymbol{a}_{\tau}, \boldsymbol{s}_{\tau}]},$$

$$\mathbb{Q}^{\dagger}_{\tau:t}(\boldsymbol{a}_{\tau}, \boldsymbol{s}_{\tau}) := \max_{\{\pi_{t'}\}_{t'\in[\tau+1,t-1]}} \mathbb{Q}^{\pi}_{\tau:t}(\boldsymbol{a}_{\tau}, \boldsymbol{s}_{\tau}),$$

we have

$$\max_{\{\pi_{t'}\}_{t'\in[0,t-1]}} \mathbb{J}_{0:t}[\pi] = \Big\langle \max_{\pi_0} \Big\langle \Big[\max_{\{\pi_t\}_{t'\in[1,t-1]}} Q^{\pi}_{0:t}(\boldsymbol{a}_0, \boldsymbol{s}_0) \Big] - \frac{1}{\beta} \ln \frac{\pi_0(\boldsymbol{a}_0|\boldsymbol{s}_0)}{\pi_r(\boldsymbol{a}_0|\boldsymbol{s}_0)} \Big\rangle_{\pi_0(\boldsymbol{a}_0|\boldsymbol{s}_0)} \Big\rangle_{p(\boldsymbol{s}_0)}$$

$$= \Big\langle \Big\langle Q^{\dagger}_{0:t}(\boldsymbol{a}_0, \boldsymbol{s}_0) - \frac{1}{\beta} \ln \frac{\pi_0^{\dagger}(\boldsymbol{a}_0|\boldsymbol{s}_0)}{\pi_r(\boldsymbol{a}_0|\boldsymbol{s}_0)} \Big\rangle_{\pi_0^{\dagger}(\boldsymbol{a}_0|\boldsymbol{s}_0)} \Big\rangle_{p(\boldsymbol{s}_0)} = \big\langle V^{\dagger}_{0:t}(\boldsymbol{s}_0) \big\rangle_{p(\boldsymbol{s}_0)},$$

where the optimal policy $\pi_0^{\dagger}(\boldsymbol{a}_0|\boldsymbol{s}_0)$ and the optimal value function $V^{\dagger}_{0:t}(\boldsymbol{s}_0)$ are defined as

$$\pi_0^{\dagger}(\boldsymbol{a}_0|\boldsymbol{s}_0) := \arg\max_{\pi_0} \Big\langle Q^{\dagger}_{0:t}(\boldsymbol{a}_0, \boldsymbol{s}_0) - \frac{1}{\beta} \ln \frac{\pi_0(\boldsymbol{a}_0|\boldsymbol{s}_0)}{\pi_r(\boldsymbol{a}_0|\boldsymbol{s}_0)} \Big\rangle_{\pi_0(\boldsymbol{a}_0|\boldsymbol{s}_0)}$$

$$= \frac{e^{\beta Q^{\dagger}_{0:t}(\boldsymbol{a}_0, \boldsymbol{s}_0)} \pi_r(\boldsymbol{a}_0|\boldsymbol{s}_0)}{\sum_{\boldsymbol{a}_0} e^{\beta Q^{\dagger}_{0:t}(\boldsymbol{a}_0, \boldsymbol{s}_0)} \pi_r(\boldsymbol{a}_0|\boldsymbol{s}_0)} = e^{\beta [Q^{\dagger}_{0:t}(\boldsymbol{a}_0, \boldsymbol{s}_0) - V^{\dagger}_{0:t}(\boldsymbol{s}_0)]} \pi_r(\boldsymbol{a}_0|\boldsymbol{s}_0)$$

and

$$V_{0:t}^{\dagger}(\boldsymbol{s}_0) := \max_{\pi_0}\left\langle Q_{0:t}^{\dagger}(\boldsymbol{a}_0,\boldsymbol{s}_0) - \frac{1}{\beta}\ln\frac{\pi_0(\boldsymbol{a}_0|\boldsymbol{s}_0)}{\pi_r(\boldsymbol{a}_0|\boldsymbol{s}_0)}\right\rangle_{\pi_0(\boldsymbol{a}_0|\boldsymbol{s}_0)} = \left\langle Q_{0:t}^{\dagger}(\boldsymbol{a}_0,\boldsymbol{s}_0) - \frac{1}{\beta}\ln\frac{\pi_0^{\dagger}(\boldsymbol{a}_0|\boldsymbol{s}_0)}{\pi_r(\boldsymbol{a}_0|\boldsymbol{s}_0)}\right\rangle_{\pi_0^{\dagger}(\boldsymbol{a}_0|\boldsymbol{s}_0)}.$$

Similarly to the normal Q function, the entropy-regularized Q function satisfies the Bellman equation:

$$\mathbb{Q}_{0:t}^{\pi}(\boldsymbol{a}_0,\boldsymbol{s}_0) = \left\langle R(\boldsymbol{s}_1;\boldsymbol{a}_0,\boldsymbol{s}_0) + \gamma\left\langle \mathbb{Q}_{1:t}^{\pi}(\boldsymbol{a}_1,\boldsymbol{s}_1) - \frac{1}{\beta}\ln\frac{\pi_1(\boldsymbol{a}_1|\boldsymbol{s}_1)}{\pi_r(\boldsymbol{a}_1|\boldsymbol{s}_1)}\right\rangle_{\pi_1(\boldsymbol{a}_1|\boldsymbol{s}_1)}\right\rangle_{P(\boldsymbol{s}_1|\boldsymbol{s}_0,\boldsymbol{a}_0)}.$$

Then, the optimal entropy-regularized Q function also satisfies the Bellman optimality equation:

$$\begin{aligned}
\mathbb{Q}_{0:t}^{\dagger}(\boldsymbol{a}_0,\boldsymbol{s}_0) &:= \max_{\{\pi_{t'}\}_{t'\in[1,t-1]}} \mathbb{Q}_{0:t}^{\pi}(\boldsymbol{a}_0,\boldsymbol{s}_0)\\
&= \left\langle R(\boldsymbol{s}_1;\boldsymbol{a}_0,\boldsymbol{s}_0) + \gamma\max_{\pi_1}\left\langle \max_{\{\pi_{t'}\}_{t'\in[2,t-1]}} \mathbb{Q}_{1:t}^{\pi}(\boldsymbol{a}_1,\boldsymbol{s}_1) - \frac{1}{\beta}\ln\frac{\pi_1(\boldsymbol{a}_1|\boldsymbol{s}_1)}{\pi_r(\boldsymbol{a}_1|\boldsymbol{s}_1)}\right\rangle_{\pi_1(\boldsymbol{a}_1|\boldsymbol{s}_1)}\right\rangle_{P(\boldsymbol{s}_1|\boldsymbol{s}_0,\boldsymbol{a}_0)}\\
&= \left\langle R(\boldsymbol{s}_1;\boldsymbol{a}_0,\boldsymbol{s}_0) + \gamma\max_{\pi_1}\left\langle \mathbb{Q}_{1:t}^{\dagger}(\boldsymbol{a}_1,\boldsymbol{s}_1) - \frac{1}{\beta}\ln\frac{\pi_1(\boldsymbol{a}_1|\boldsymbol{s}_1)}{\pi_r(\boldsymbol{a}_1|\boldsymbol{s}_1)}\right\rangle_{\pi_1(\boldsymbol{a}_1|\boldsymbol{s}_1)}\right\rangle_{P(\boldsymbol{s}_1|\boldsymbol{s}_0,\boldsymbol{a}_0)}\\
&= \left\langle R(\boldsymbol{s}_1;\boldsymbol{a}_0,\boldsymbol{s}_0) + \gamma V_{1:t}^{\dagger}(\boldsymbol{s}_1)\right\rangle_{P(\boldsymbol{s}_1|\boldsymbol{s}_0,\boldsymbol{a}_0)},
\end{aligned}$$

where

$$V_{\tau:t}^{\dagger}(\boldsymbol{s}_0) := \max_{\pi_\tau}\left\langle Q_{\tau:t}^{\dagger}(\boldsymbol{a}_\tau,\boldsymbol{s}_\tau) - \frac{1}{\beta}\ln\frac{\pi_\tau(\boldsymbol{a}_\tau|\boldsymbol{s}_\tau)}{\pi_r(\boldsymbol{a}_\tau|\boldsymbol{s}_\tau)}\right\rangle_{\pi_\tau(\boldsymbol{a}_\tau|\boldsymbol{s}_\tau)} = \left\langle Q_{\tau:t}^{\dagger}(\boldsymbol{a}_\tau,\boldsymbol{s}_\tau) - \frac{1}{\beta}\ln\frac{\pi_\tau^{\dagger}(\boldsymbol{a}_\tau|\boldsymbol{s}_\tau)}{\pi_r(\boldsymbol{a}_\tau|\boldsymbol{s}_\tau)}\right\rangle_{\pi_\tau^{\dagger}(\boldsymbol{a}_\tau|\boldsymbol{s}_\tau)},$$

and

$$\pi_\tau^{\dagger}(\boldsymbol{a}_\tau|\boldsymbol{s}_\tau) := \arg\max_{\pi_\tau}\left\langle Q_{\tau:t}^{\dagger}(\boldsymbol{a}_\tau,\boldsymbol{s}_\tau) - \frac{1}{\beta}\ln\frac{\pi_\tau(\boldsymbol{a}_\tau|\boldsymbol{s}_\tau)}{\pi_r(\boldsymbol{a}_\tau|\boldsymbol{s}_\tau)}\right\rangle_{\pi_\tau(\boldsymbol{a}_\tau|\boldsymbol{s}_\tau)} = e^{\beta[Q_{\tau:t}^{\dagger}(\boldsymbol{a}_\tau,\boldsymbol{s}_\tau)-V_{\tau:t}^{\dagger}(\boldsymbol{s}_\tau)]}\pi_r(\boldsymbol{a}_\tau|\boldsymbol{s}_\tau).$$

By considering the infinite time horizon ($t\rightarrow\infty$), we also have the Bellman equation:

$$\mathbb{Q}^{\pi}(\boldsymbol{a},\boldsymbol{s}) = \left\langle R(\boldsymbol{s}';\boldsymbol{a},\boldsymbol{s}) + \gamma\left\langle \mathbb{Q}^{\pi}(\boldsymbol{a}',\boldsymbol{s}') - \frac{1}{\beta}\ln\frac{\pi(\boldsymbol{a}'|\boldsymbol{s}')}{\pi_r(\boldsymbol{a}'|\boldsymbol{s}')}\right\rangle_{\pi(\boldsymbol{a}'|\boldsymbol{s}')}\right\rangle_{P(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})},$$

as well as the optimal value function $V^{\dagger}(\boldsymbol{s})$ and the optimal policy $\pi^{\dagger}(\boldsymbol{a}|\boldsymbol{s})$:

$$V^{\dagger}(\boldsymbol{s}) := \max_{\pi}\left\langle Q^{\dagger}(\boldsymbol{a},\boldsymbol{s}) - \frac{1}{\beta}\ln\frac{\pi(\boldsymbol{a}|\boldsymbol{s})}{\pi_r(\boldsymbol{a}|\boldsymbol{s})}\right\rangle_{\pi(\boldsymbol{a}|\boldsymbol{s})} = \left\langle Q^{\dagger}(\boldsymbol{a},\boldsymbol{s}) - \frac{1}{\beta}\ln\frac{\pi^{\dagger}(\boldsymbol{a}|\boldsymbol{s})}{\pi_r(\boldsymbol{a}|\boldsymbol{s})}\right\rangle_{\pi^{\dagger}(\boldsymbol{a}|\boldsymbol{s})},$$

and

$$\pi^{\dagger}(\boldsymbol{a}|\boldsymbol{s}) := \arg\max_{\pi}\left\langle Q^{\dagger}(\boldsymbol{a},\boldsymbol{s}) - \frac{1}{\beta}\ln\frac{\pi(\boldsymbol{a}|\boldsymbol{s})}{\pi_r(\boldsymbol{a}|\boldsymbol{s})}\right\rangle_{\pi(\boldsymbol{a}|\boldsymbol{s})} = e^{\beta[Q^{\dagger}(\boldsymbol{a},\boldsymbol{s})-V^{\dagger}(\boldsymbol{s})]}\pi_r(\boldsymbol{a}|\boldsymbol{s}). \tag{B4}$$

From this result, we can see that the soft-max policy is obtained as the optimal policy in Eq. (B4). If $\gamma = 0$, $R(\boldsymbol{s}';\boldsymbol{a},\boldsymbol{s}) = R(\boldsymbol{a},\boldsymbol{s})$, and π_r is assumed to be the uniform distribution, the optimal policy and the optimal Q function are reduced to those in the main text as

$$\pi^{\dagger}(\boldsymbol{a}|\boldsymbol{s}) = e^{\beta(\mathbb{Q}^{\dagger}(\boldsymbol{a},\boldsymbol{s})-V^{\dagger}(\boldsymbol{s}))},$$

$$\mathbb{Q}^{\dagger}(\boldsymbol{a},\boldsymbol{s}) = R(\boldsymbol{a},\boldsymbol{s}).$$

Similarly to the conventional RL, we have to approximate the entropy-regularized Q function $\tilde{\mathbb{Q}}^{\tau}(\boldsymbol{a},\boldsymbol{s})$ and update it to be closer to the optimal entropy-regularized Q function, $\mathbb{Q}^{\dagger}(\boldsymbol{a},\boldsymbol{s})$. We can consider the same loss function as in Eq. (B3) for the entropy-regularized case. When $\gamma = 0$, the loss function is reduced to that in Eq. (6) of the main text. In the main text, we showed that the Th immune network can work to represent $\tilde{\pi}(\boldsymbol{a}|\boldsymbol{s})$ and $\tilde{\mathbb{Q}}(\boldsymbol{a},\boldsymbol{s})$ and can be trained to approximate $\mathbb{Q}^{\dagger}(\boldsymbol{a},\boldsymbol{s})$.

While we assumed that the reference dynamics $\pi_r(\boldsymbol{a}|\boldsymbol{s})$ is uniform for simplicity and for highlighting the adaptive role of Th cells, $\pi_r(\boldsymbol{a}|\boldsymbol{s})$ can be used to represent the details of the intrinsic actions of the innate immunity. Such innate dynamics may play roles not only to induce quick response to pathogens, but also to promote learning by the adaptive immunity.

However, the direct action of the innate immunity is not antigen dependent but pathogen dependent. In order to account for the innate dynamics more appropriately, we should extend our MDP formulation to a partially observed Markov decision process (POMDP), where the pathogens and their dynamics are represented as the hidden state. Such an extension is interesting and promising as future work.

APPENDIX C: DERIVATION OF THE METRIC IN THE LEARNING RULE BASED ON INVARIANCE

We introduce a detailed derivation of the metric

$$g_{ij} = \frac{\delta_{ij}}{n_i}$$

on the parameter space $\mathbb{R}^K_{++}$, where $\mathbb{R}_{++} := \{x \in \mathbb{R} | x > 0\}$ and K is the number of different Th clones.

The two invariance rules imposed on the parameter space $\mathbb{R}^K_{++}$ in the main text can be summarized as the invariance under the Markov mappings [67]. Markov mappings are maps from a low-dimensional space $\mathbb{R}^K_{++}$ onto a higher-dimensional space $\mathbb{R}^L_{++}$ $(2 \leqslant K \leqslant L)$ constructed by the following steps:

(i) Partition a set $\{1, 2, \dots, L\}$ into nonempty K disjoint sets $\{C_1, C_2, \dots, C_K\}$.

(ii) For each $k \in \{1, 2, \dots, K\}$, associate a probability distribution over the set $\{1, 2, \dots, L\}$ as

$$\boldsymbol{q}_k := \left(q_k^1, q_k^2, \dots, q_k^L\right),$$

where $q_k^j = 0$ for $j \notin C_k$, and $\sum_{j=1}^L q_k^j = 1$.

(iii) Define a mapping $\mathbf{f} : \mathbb{R}^K_{++} \ni \mathbf{n} \mapsto \mathbf{m} \in \mathbb{R}^L_{++}$ by

$$m_l := \sum_{k=1}^K q_k^l n_k, \quad \text{for } l \in \{1, 2, \dots, L\}.$$

Imposing invariance under the Markov mappings is equivalent to imposing the two invariance rules described in the main text. In the following, we show why.

Consider a case where the dimensions of the domain and the range of the Markov mapping f are the same, $K = L$, and the associated partition sets $C_1, \dots, C_K$ are singletons $\{k\}$ for $k = 1, 2, \dots, K-2$ except for the last two cases $C_{K-1} = \{K\}$ and $C_K = \{K-1\}$. In this case, the resulting Markov mapping $\mathbf{f}$ is the identity mapping except for the following cases:

$$m_{K-1} = f_{K-1}(\mathbf{n}) = n_K,$$
$$m_K = f_K(\mathbf{n}) = n_{K-1}.$$

This mapping only swaps the labels of the components, n_{k-1} and n_k, and leaves the others untouched. Therefore, the invariance under the Markov mapping imposes the invariance under the swapping of labels of the clones.

The other case shows that the invariance under a Markov mapping is equivalent to imposing the invariance with respect to a subdivision of a clone. Suppose that $L = K+1$ and consider partition sets $C_1, \dots, C_K$ where $C_k = \{k\}$ for $k \in \{1, 2, \dots, K-1\}$ and $C_K = \{K, K+1\}$. In addition, suppose $q_K = (0, 0, \dots, 0, 1/2, 1/2)$. This results in a Markov mapping f such that almost all the components $n_1, \dots, n_{K-1}$ are mapped identically, but the last component n_K is divided in half:

$$m_K = f_K(\mathbf{n}) = n_K/2,$$
$$m_{K+1} = f_{K+1}(\mathbf{n}) = n_K/2.$$

This corresponds to a subdivision of a clone described in the main text.

Before deriving the metric with the invariance under the Markov mappings, we need to restate the invariance of the metric with respect to the Markov mapping in definitive terms. The parameter space $\mathbb{R}^K_{++}$ can be considered as a K-dimensional manifold M with a global coordinate $\{n_k\}_{k=1}^K$. Thus, its tangent space T_pM at any point $p = (n_1, n_2, \dots, n_K)$ has a basis $((\partial/\partial n_1)_p, (\partial/\partial n_2)_p, \dots, (\partial/\partial n_K)_p)$. The metric on this manifold on the point p is a map $g_p^K : T_pM \times T_pM \mapsto \mathbb{R}$. To state a relationship between two metrics defined on different manifolds, we need to define a relationship between the two tangent spaces defined on the two different manifolds. One way for that is called pushforward, in which a map f from a manifold M to the other N defines a map f_* from a tangent vector X of the tangent space T_pM on the manifold M to a tangent vector $f_*(X)$ of the tangent space T_pN on the manifold N as

$$f_*(X)(g) := X(g \circ f),$$

for any map $g : N \mapsto \mathbb{R}$. With this definition, the invariance under the Markov mapping is restated that the metric g_p^K on any point $p \in \mathbb{R}^K_{++}$ satisfies the following invariance of the metric under any Markov mappings f for any tangent vectors $X, Y \in T_p\mathbb{R}^K_{++}$:

$$g_p^K(X, Y) = g_{f(p)}^L(f_*X, f_*Y).$$

To derive the metric, we consider three special cases of Markov mappings to gradually restrict the possible form of the metric that is invariant under these Markov mappings. The first mapping to consider is a class of Markov mappings whose dimensions of the domain and the range are equal as $K = L$. This condition means that the Markov mappings only swap the labels of the parameters. Consider a metric g_p^K on a point $p = (n/K, \dots, n/K)$ $(n > 0)$. Since any permutation of the components of p does not change p, $f(p) = p$ holds. Also, pushforward f_* of a map f belonging to this class just changes the label i of $\partial/\partial n_i$ to another label $j \in \{1, 2, \dots, K\}$. The invariance under these mappings f first indicates that the value of the metric on identical tangent vectors,

$$g_p^K\left(\frac{\partial}{\partial n_i}, \frac{\partial}{\partial n_i}\right),$$

does not depend on the label i. This value only depends on the scaling parameter $n > 0$ and the dimension of the space K. We denote the value as $B_K(n)$:

$$g_p^K\left(\frac{\partial}{\partial n_1}, \frac{\partial}{\partial n_1}\right) = \dots = g_p^K\left(\frac{\partial}{\partial n_K}, \frac{\partial}{\partial n_K}\right) = B_K(n).$$

Another indication of the invariance under the mappings f is that the value of the metric on different basis tangent vectors,

$$g_p^K\left(\frac{\partial}{\partial n_i}, \frac{\partial}{\partial n_j}\right),$$

is identical for any pairs of (i, j) when $i \neq j$. We denote the difference between this value and $B_K(n)$ as $A_K(n)$. Combining these two implications dictates that we can express the value of the metric on a point $p = (n/K, \dots, n/K)$ as

$$g_p^K\left(\frac{\partial}{\partial n_i}, \frac{\partial}{\partial n_j}\right) = \delta_{ij}A_K(n) + B_K(n). \tag{C1}$$

The next class of Markov mappings to consider is a set of maps whose dimension of the range L is equal to the multiple of the dimension of the domain K. In this case, a Markov mapping f of this class maps a point $\mathbf{n} = (n_1, n_2, \dots, n_K) \in \mathbb{R}^K_{++}$ to

$$\begin{aligned} f(n_1, n_2, \dots, n_K) &= \left(\frac{n_1}{N}, \dots, \frac{n_1}{N}, \dots, \frac{n_K}{N}, \dots, \frac{n_K}{N}\right), \\ &= (m_{11}, \dots, m_{1N}, \dots, m_{K1}, \dots, m_{KN}) \\ &\in \mathbb{R}^L_{++}, \end{aligned}$$

where N is the integer satisfying $L = NK$. In this case, the pushforward of the basis tangent vector $\partial/\partial n_i$ becomes

$$f_*\left(\frac{\partial}{\partial n_i}\right) = \frac{1}{N}\sum_{r=1}^{N} \frac{\partial}{\partial m_{ir}}, \tag{C2}$$

because

$$\begin{aligned} &\frac{\partial}{\partial n_i}(g \circ f)(n_1, n_2, \dots, n_K) \\ &\quad = \frac{\partial}{\partial n_i} g\left(\frac{n_1}{N}, \dots, \frac{n_1}{N}, \dots, \frac{n_K}{N}, \dots, \frac{n_K}{N}\right), \\ &\quad = \sum_{r=1}^{N} \frac{\partial g}{\partial m_{ir}} \frac{\partial m_{ir}}{\partial n_i}, \\ &\quad = \left(\frac{1}{N}\sum_{r=1}^{N} \frac{\partial}{\partial m_{ir}}\right) g(m_{11}, \dots, m_{1N}, \dots, m_{KN}), \end{aligned}$$

for any mappings $g : \mathbb{R}^L_{++} \mapsto \mathbb{R}$. The invariance under these mappings f imposes the metric g^K_p on a point $p = (n/K, \dots, n/K)$ upon satisfying

$$\begin{aligned} g^K_p\left(\frac{\partial}{\partial n_i}, \frac{\partial}{\partial n_j}\right) &= g^L_{f(p)}\left(f_*\frac{\partial}{\partial n_i}, f_*\frac{\partial}{\partial n_j}\right) \\ &= g^L_{f(p)}\left(\frac{1}{N}\sum_{r=1}^{N}\frac{\partial}{\partial m_{ir}}, \frac{1}{N}\sum_{s=1}^{N}\frac{\partial}{\partial m_{js}}\right) \\ &= \frac{1}{N^2}\sum_{r,s=1}^{N} g^L_{f(p)}\left(\frac{\partial}{\partial m_{ir}}, \frac{\partial}{\partial m_{js}}\right) \\ &= \frac{1}{N^2}\sum_{r,s=1}^{N}(\delta_{ij}\delta_{rs}A_L(n) + B_L(n)) \\ &= \frac{\delta_{ij}}{N}A_L(n) + B_L(n). \end{aligned}$$

With Eq. (C1), we can conclude that $B_K(n)$ is not dependent on the dimension K,

$$B_K(n) = B_L(n) = \beta(n),$$

and $A_K(n)$ is proportional to the dimension K,

$$\begin{aligned} A_K(n) &= \frac{1}{N}A_L(n) = \frac{K}{L}A_L(n) \\ \Longleftrightarrow \quad \frac{A_K(n)}{K} &= \frac{A_L(n)}{L} = \alpha(n). \end{aligned}$$

As a result, the metric invariant under the Markov mapping takes the following form at a point $p = (n/K, \dots, n/K)$:

$$g^K_p\left(\frac{\partial}{\partial n_i}, \frac{\partial}{\partial n_j}\right) = K\alpha(n)\delta_{ij} + \beta(n). \tag{C3}$$

Consider a rational point p of $\mathbb{R}^K_{++}$,

$$p = (n_1, n_2, \dots, n_K) = \left(\frac{u_1}{v}, \dots, \frac{u_K}{v}\right),$$

where $v, u_1, \dots, u_K$ are all integers. Any rational points in $\mathbb{R}^K_{++}$ can be expressed in this form by multiplying the denominator of each element up to the least common denominator. The class of Markov mappings to examine $f : \mathbb{R}^K_{++} \mapsto \mathbb{R}^L_{++}$ is

$$\begin{aligned} f(n_1, \dots, n_K) &= \left(\overbrace{\frac{n_1}{u_1}, \dots, \frac{n_1}{u_1}}^{u_1}, \dots, \overbrace{\frac{n_K}{u_K}, \dots, \frac{n_K}{u_K}}^{u_K}\right) \\ &= (m_{11}, \dots, m_{1u_1}, \dots, m_{K1}, \dots, m_{Ku_K}), \end{aligned}$$

where we define the sum of all elements of p as n,

$$n = \sum_{k=1}^{K} n_k \in \mathbb{R}_{++},$$

and the sum of its numerators as L,

$$L = \sum_{k=1}^{K} u_k \in \mathbb{N}.$$

The invariance under these Markov mappings imposes the metric g^K_p at this point p upon satisfying

$$\begin{aligned} g^K_p\left(\frac{\partial}{\partial n_i}, \frac{\partial}{\partial n_j}\right) &= g^L_{f(p)}\left(f_*\frac{\partial}{\partial n_i}, f_*\frac{\partial}{\partial n_j}\right) \\ &= g^L_{f(p)}\left(\frac{1}{u_i}\sum_{r=1}^{u_i}\frac{\partial}{\partial m_{ir}}, \frac{1}{u_j}\sum_{s=1}^{u_j}\frac{\partial}{\partial m_{js}}\right) \\ &= \frac{1}{u_i u_j}\sum_{r=1}^{u_i}\sum_{s=1}^{u_j} g^L_{f(p)}\left(\frac{\partial}{\partial m_{ir}}, \frac{\partial}{\partial m_{js}}\right). \end{aligned}$$

Since the mapping f was defined to map the point p to a point with equivalent elements,

$$f(p) = \left(\frac{1}{v}, \dots, \frac{1}{v}\right),$$

Eq. (C3) implies

$$g^L_{f(p)}\left(\frac{\partial}{\partial m_{ir}}, \frac{\partial}{\partial m js}\right) = \delta_{ij}\delta_{rs}L\alpha(n) + \beta(n).$$

Therefore, g^K_p can be expressed as

$$\begin{aligned} g^K_p\left(\frac{\partial}{\partial n_i}, \frac{\partial}{\partial n_j}\right) &= \frac{1}{u_i u_j}\sum_{r=1}^{u_i}\sum_{s=1}^{u_j} g^L_{f(p)}\left(\frac{\partial}{\partial m_{ir}}, \frac{\partial}{\partial m_{js}}\right) \\ &= \frac{1}{u_i u_j}\sum_{r=1}^{u_i}\sum_{s=1}^{u_j}(\delta_{ij}\delta_{rs}L\alpha(n) + \beta(n)) \end{aligned}$$

$$
\begin{aligned}
&= \frac{\delta_{ij} L}{u_i}\alpha(n) + \beta(n)\\
&= \frac{\delta_{ij} L/v}{u_i/v}\alpha(n) + \beta(n)\\
&= \frac{\delta_{ij} n}{n_i}\alpha(n) + \beta(n).
\end{aligned}
$$

As a result, the metric invariant under the Markov mappings takes the following form at any rational point $p = (n_1, n_2, \ldots, n_K)$:

$$
g_p^K\left(\frac{\partial}{\partial n_i}, \frac{\partial}{\partial n_j}\right) = \frac{\delta_{ij}}{n_i}\gamma(n) + \beta(n), \tag{C4}
$$

where $\gamma(n) = n\alpha(n)$ and $n = \sum_{k=1}^{K} n_k$. The result can be extended to all points in $\mathbb{R}_{++}^K$ once we restrict the metric g_p^K to be continuous.

All the metrics in the form of Eq. (C4) satisfy the two invariance rules imposed on the parameter space $\mathbb{R}_{++}^K$. The derivation of the learning rule in the main text uses a specific case,

$$
\gamma(n) = 1,
$$
$$
\beta(n) = 0,
$$

to avoid introducing unnecessary complexities to our theory, in which case the metric is

$$
g_{ij} = g_p^K\left(\frac{\partial}{\partial n_i}, \frac{\partial}{\partial n_j}\right) = \frac{\delta_{ij}}{n_i}.
$$

APPENDIX D: AN EXTENSION TO ADVERSARIAL ENVIRONMENT

The dynamics of the antigen pattern can be more complicated by considering specific situations. As an example, let us consider the following transition probability $P(\mathbf{s}'|\mathbf{s}, \mathbf{a})$:

$$
P(\mathbf{s}^i|\mathbf{s}, \mathbf{a}) \propto \mathrm{ham}(\mathbf{a}^i, \mathbf{a}),
$$

which represents the probability that the antigen pattern transits to $\mathbf{s}^i$ when the Th cell population induces an activity pattern $\mathbf{a}$ in response to an antigen pattern $\mathbf{s}$. This transition probability roughly corresponds to an adversarial situation for the immune system such that the new antigen pattern is more likely to be a more distant one from the current pattern of the effector cells $\mathbf{a}$. Such a situation may represent a simplified version of immune-pathogen coevolution where the new pathogen attains the ability to evade current immunological responses. While the change in the immune system of only one antigen alone cannot affect the evolution of the pathogen, the pathogen evolution can be influenced if it infects many agents simultaneously. At the level of mean-field approximation, we may roughly use one agent as the proxy of the mean field of the others. As shown in Fig. 6, even under such a situation, learning can be achieved. We can obtain a similar learning profile and stationary distribution to those in Fig. 3 in the main text as shown in Fig. 6.

APPENDIX E: SPARSITY OF ANTIGEN-Th INTERACTION

While we assumed, in the main text, that the interaction strength w_{ki} of the kth clone with the ith antigen follows the normal distribution with mean zero and variance σ_w^2, it has been suggested that each TCR has a high affinity to a small fraction of antigens. While such a situation is approximately represented even by the normal distribution because most of w_{ki} are concentrated around zero, we tested the impact of the sparsity more directly by setting $w_{ki} = 0$ with probability $1 - q_w$. More specifically, we chose w_{ki} as

$$
w_{ki} = \begin{cases} 0 & \text{with probability } 1 - q_w \\ \omega & \text{with probability } q_w, \end{cases} \tag{E1}
$$

where ω is sampled from a normal distribution with mean zero and variance σ_w^2/q_w. The variance of the normal distribution is scaled such that the sample variance of the sparsified w_{ki} becomes independent of q_w. Figure 7 shows the learning dynamics for $q_w = 0.1$. Even though 90% of w_{ki} are zero, the learning dynamics and the stationary distribution are altered only a little as demonstrated in Fig. 7. Moreover, the

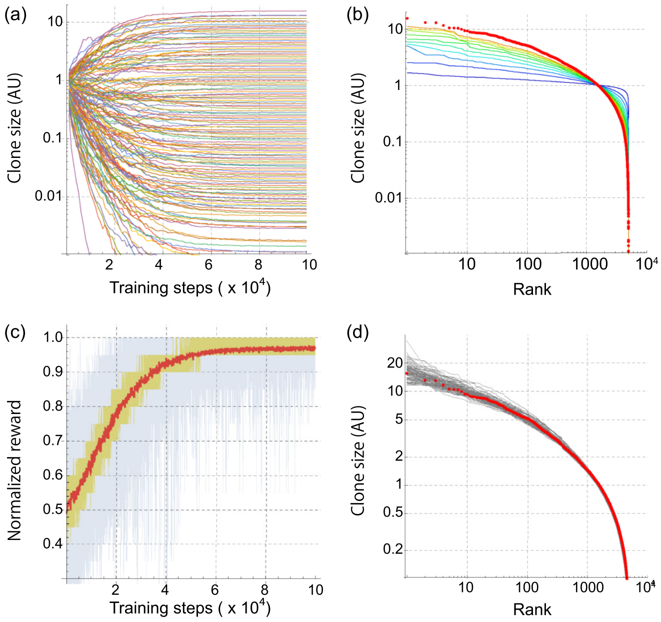

FIG. 6. Learning dynamics and stationary abundance distribution under the adversarial environment. (a) Time series of the Th clone size $\mathbf{n}$ along learning. The parameter values are the same as in Fig. 3 in the main text. The clone size $\mathbf{n}$ is normalized by the initial abundance. This figure shows only 200 trajectories sampled evenly out of 5000 at the stationary state to avoid complication of the plot. (b) The dynamics of the rank-abundance distributions along the learning trial, which were calculated from the trajectories of $\mathbf{n}$ in (a). The red circles represent the stationary distribution after the learning, and the colored curves are the transient distributions calculated at training steps from 1×10^3 (blue) to 28×10^3 (yellow). (c) Statistics of learning curves of the Th cell population. The rewards normalized by its maximum value are obtained as functions of the training step for 100 independent learning trials. The red curve is the average reward, and the yellow and blue regions show the range between 25th and 75th percentiles of the rewards and that between minimum and maximum of the rewards at each training step, respectively. (d) The stationary rank-abundance distributions for the 100 independent learning trials in (c) are shown by gray curves. The red dots are the same as those in (b).

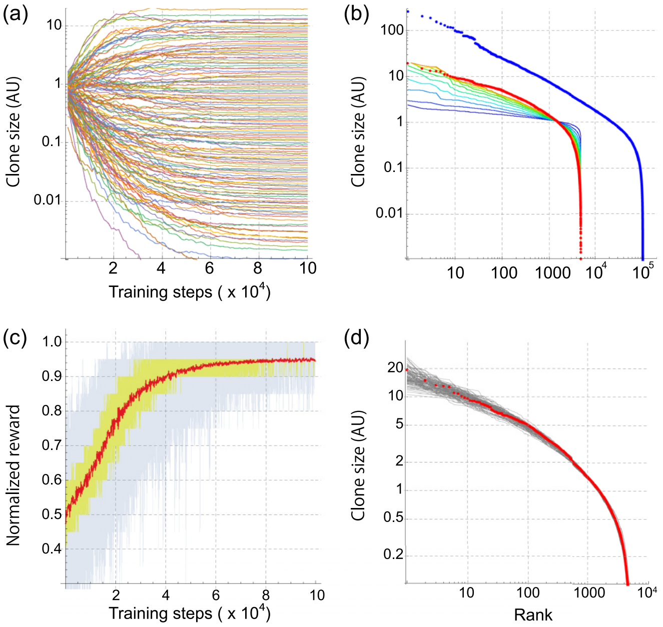

FIG. 7. Learning dynamics and stationary abundance distribution for sparse **w**. Each w_{kj} is determined by following Eq. (E1) for $q_w = 0.1$. (a) Time series of the Th clone size **n** along learning under the adversarial environment. The parameter values are the same as Fig. 3 in the main text. The clone size **n** is normalized by the initial abundance. This figure shows only 200 trajectories sampled evenly out of 5000 at the stationary state to avoid complication of the plot. (b) The dynamics of the rank-abundance distributions along the learning trial, which were calculated from the trajectories of **n** in (a). The red dots represent the stationary distribution after the learning, and the colored curves are the transient distributions calculated at training steps from 1×10^3 (blue) to 28×10^3 (yellow). The blue circles are the stationary distribution for $K = 100\,000$. All the parameters but the sparsity are the same as those in Fig. 5 in the main text. (c) Statistics of learning curves of the Th cell population. The rewards normalized by its maximum value are obtained as functions of the training step for 100 independent learning trials. The red curve is the average reward, and the yellow and blue regions show the range between 25th and 75th percentiles of the rewards and that between minimum and maximum of the rewards at each training step, respectively. (d) The stationary rank-abundance distributions for the 100 independent learning trials in (c) are shown by gray curves. The red dots are the same as those in (b).

power-law behavior discussed in Fig. 5 of the main text is also observed for sparse w_{ki} as in Fig. 7(b).

APPENDIX F: LEARNING PERFORMANCE AS FUNCTIONS OF K, N, M, AND P

In order to investigate how the learning performance is affected by K, N, M, and P, we plot the average reward as a function of either K, N, M, or P in Fig. 8. Figures 8(a) and 8(b) show that the average reward increases monotonously and gets saturated eventually as either the number of Th clones K or that of the antigen types N increases. This indicates that there are sufficient diversities of Th clones and antigen types to achieve sufficient learning performance for the given situation. In contrast, the average reward decreases monotonously as the number of pathogens increases. This means that the learning becomes more difficult when we have more diverse pathogens in the environment. If we increase the number of the effector types, M, the reward increases first and then turns to decrease. The first increase may correspond to the improvement of the ability of the immune system to handle different pathogens. The subsequent decline is due to the increased complexity to control the effector activity as the number of the effector types increases. As indicated by the dashed lines in Fig. 8, the parameter values used in Fig. 3 in the main text were chosen such that high learning performance is achieved with minimum diversity of the Th clones K and the antigen types N.

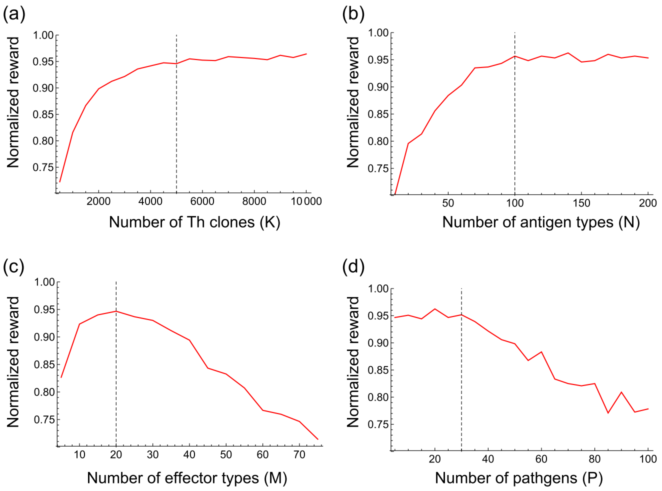

FIG. 8. The average reward as functions of (a) the number of the Th clone types, K; (b) the number of the antigen types, N; (c) the number of the effector types, M; and (d) the number of the environmental state (the number of pathogens), P. All the other parameters are the same as those as in Fig. 3 of the main text. The average reward for each parameter value is obtained by the arithmetic average of 20 independent learning trials. The dashed lines indicate the parameter values used in Fig. 3 of the main text.

APPENDIX G: PARAMETER DEPENDENCE OF THE SHAPE OF RANK-ABUNDANCE DISTRIBUTION

In Fig. 5, we have shown that the power-law-like rank-abundance distribution appears if we increase K from 5000 to 100 000. Because the performance of learning starts saturating around $K = 5000$, $K = 100\,000$ means that the diversity of the Th clones is much larger than required to achieve sufficiently high learning performance for the given pathogen variety $P = 30$. This suggests that the power-law-like abundance distribution can appear if the Th clone diversity is much higher than required for the diversity of pathogens. The generality of this property can be verified by changing the number of pathogens P under the same setting as in Fig. 3 of the main text. As shown in Fig. 9(a), the linear region of the rank-abundance distribution in the log-log plot stretches if P is decreased from $P = 30$. The power-law property of the rank-abundance distribution is also associated with the symmetry of fitness distribution of clones. As

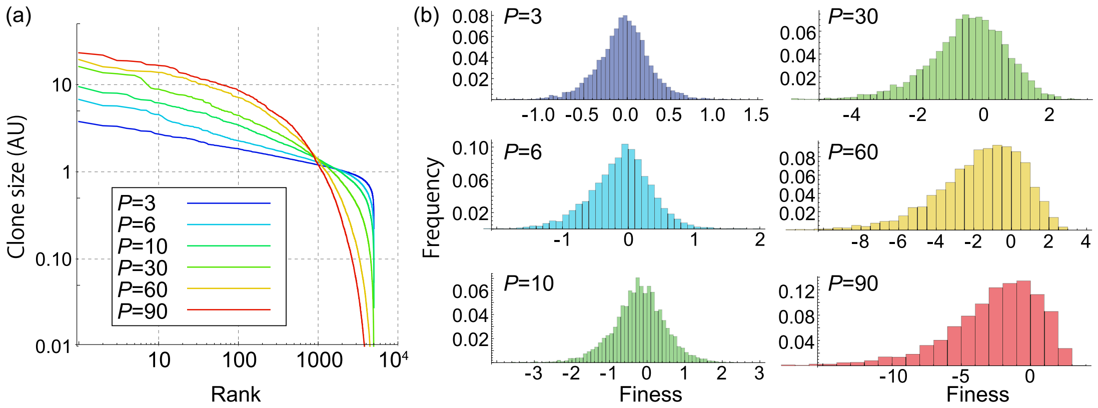

FIG. 9. (a) Stationary rank-abundance distributions for different values of P. All the other parameters are the same as in Fig. 3 in the main text. (b) The distributions of fitness of clones for different values of P. Fitness is defined by $\Lambda = \log(n_i(t)/n_i(0))$. The skewness of the distributions are -0.19, -0.35, -0.28, -0.79, -1.2, and -1.8 for $P = 3$, 6, 10, 30, 60, and 90, respectively.

shown in Fig. 9(b), the fitness distribution becomes negatively skewed as pathogen diversity P increases and thereby as the power-law property is lost. Because of this asymmetry for large P, there exist more clones that have smaller fitness than larger, which results in the deviation from the power-law distribution. While the mechanism of how the symmetric fitness distribution appears for small P relative to a given K is elusive at this moment, our learning framework suggests a connection with the capacity of the system to learn diverse pathogens.

[1] A. K. Abbas, A. H. H. Lichtman, and S. Pillai, *Cellular and Molecular Immunology* (Elsevier, Amsterdam, 2014).

[2] K. Murphy and C. Weaver, *Janeway's Immunobiology* (Garland Science, New York, 2016).

[3] F. Annunziato, C. Romagnani, and S. Romagnani, The 3 major types of innate and adaptive cell-mediated effector immunity, J. Allergy Clin. Immunol. **135**, 626 (2015).

[4] R. Satija and A. K. Shalek, Heterogeneity in immune responses: From populations to single cells, Trends Immunol. **35**, 219 (2014).

[5] A.-C. Villani, S. Sarkizova, and N. Hacohen, Systems immunology: Learning the rules of the immune system, Annu. Rev. Immunol. **36**, 813 (2018).

[6] K. Kveler, E. Starosvetsky, A. Ziv-Kenet, Y. Kalugny, Y. Gorelik, G. Shalev-Malul, N. Aizenbud-Reshef, T. Dubovik, M. Briller, J. Campbell, J. C. Rieckmann, N. Asbeh, D. Rimar, F. Meissner, J. Wiser, and S. S. Shen-Orr, Immune-centric network of cytokines and cells in disease context identified by computational mining of PubMed, Nat. Biotechnol. **36**, 651 (2018).

[7] P. Brodin, V. Jojic, T. Gao, S. Bhattacharya, C. J. L. Angel, D. Furman, S. Shen-Orr, C. L. Dekker, G. E. Swan, A. J. Butte *et al.*, Variation in the human immune system is largely driven by non-heritable influences, Cell **160**, 37 (2015).

[8] D. J. Dowling and O. Levy, Ontogeny of early life immunity, Trends Immunol. **35**, 299 (2014).

[9] A. K. Simon, G. A. Hollander, and A. McMichael, Evolution of the immune system in humans from infancy to old age, Proc. R. Soc. London Ser. B **282**, 20143085 (2015).

[10] E. Von Mutius and D. Vercelli, Farm living: Effects on childhood asthma and allergy, Nat. Rev. Immunol. **10**, 861 (2010).

[11] C. K. Baumgartner, H. Yagita, and L. P. Malherbe, A TCR affinity threshold regulates memory CD4 T cell differentiation following vaccination, J. Immunol. **189**, 2309 (2012).

[12] A. Rattan, K. A. Richards, Z. A. Knowlden, and A. J. Sant, Protein vaccination directs the CD4+ T cell response toward shared protective epitopes that can be recalled after influenza virus infection, J. Virol. **93**, e00947 (2019).

[13] M. Tian, H. Liang, Q. Qin, W. Zhang, and S. Zhang, Changes in T cell subpopulations after specific sublingual immunotherapy against Dermatophagoides farinae, Int. J. Clin. Exp. Med. **9**, 9411 (2016).

[14] D. Marr, *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information* (MIT Press, Cambridge, MA, 2010).

[15] D. Hassabis, D. Kumaran, C. Summerfield, and M. Botvinick, Neuroscience-inspired artificial intelligence, Neuron **95**, 245 (2017).

[16] J. C. Rieckmann, R. Geiger, D. Hornburg, T. Wolf, K. Kveler, D. Jarrossay, F. Sallusto, S. S. Shen-Orr, A. Lanzavecchia, M. Mann *et al.*, Social network architecture of human immune cells unveiled by quantitative proteomics, Nat. Immunol. **18**, 583 (2017).

[17] J. M. Heather, M. Ismail, T. Oakes, and B. Chain, High-throughput sequencing of the T-cell receptor repertoire: Pitfalls and opportunities, Briefings Bioinf. **19**, 554 (2018).

[18] E. Ruggiero, J. P. Nicolay, R. Fronza, A. Arens, A. Paruzynski, A. Nowrouzi, G. Ürenden, C. Lulay, S. Schneider, S. Goerdt *et al.*, High-resolution analysis of the human T-cell receptor repertoire, Nat. Commun. **6**, 8081 (2015).

[19] F. M. Burnet, A modification of Jerne's theory of antibody production using the concept of clonal selection, Ca-Cancer J. Clin. **26**, 119 (1976).

[20] A. S. Perelson, Immune network theory, Immunol. Rev. **110**, 5 (1989).

[21] G. I. Bell, Mathematical model of clonal selection and antibody production, Nature **228**, 739 (1970).

[22] R. J. de Boer and A. S. Perelson, Competitive control of the self-renewing T cell repertoire, Int. Immunol. **9**, 779 (1997).

[23] A. Mayer, V. Balasubramanian, T. Mora, and A. M. Walczak, How a well-adapted immune system is organized, Proc. Natl. Acad. Sci. USA **112**, 5950 (2015).

[24] A. Mayer, V. Balasubramanian, A. M. Walczak, and T. Mora, How a well-adapting immune system remembers, Proc. Natl. Acad. Sci. USA **116**, 8815 (2019).

[25] A. Mayer, Y. Zhang, A. S. Perelson, and N. S. Wingreen, Regulation of T cell expansion by antigen presentation dynamics, Proc. Natl. Acad. Sci. USA **116**, 5914 (2019).

[26] A. Jain and C. Pasare, Innate control of adaptive immunity: Beyond the three-signal paradigm, J. Immunol. **198**, 3791 (2017).

[27] E. O. Neftci and B. B. Averbeck, Reinforcement learning in artificial and biological systems, Nat. Mech. Int. **1**, 133 (2019).

[28] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA, 2018).

[29] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming* (Wiley, New York, 2014).

[30] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, Playing Atari with deep reinforcement learning, arXiv:1312.5602.

[31] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, Mastering the game of Go with deep neural networks and tree search, Nature **529**, 484 (2016).

[32] G. Neu, A. Jonsson, and V. Gómez, A unified view of entropy-regularized Markov decision processes, arXiv:1705.07798.

[33] W. Ellmeier and C. Seiser, Histone deacetylase function in CD4+ T cells, Nat. Rev. Immunol. **18**, 617 (2018).

[34] S.-I. Amari and S. C. Douglas, Why natural gradient?, in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing* (IEEE, Piscataway, NJ, 1998), pp. 1213–1216.

[35] P. Kalinski and M. Moser, Consensual immunity: Success-driven development of T-helper-1 and T-helper-2 responses, Nat. Rev. Immunol. **5**, 251 (2005).

[36] L. Chen and D. B. Flies, Molecular mechanisms of T cell co-stimulation and co-inhibition, Nat. Rev. Immunol. **13**, 227 (2013).

[37] G. Voisinne, B. G. Nixon, A. Melbinger, G. Gasteiger, M. Vergassola, and G. Altan-Bonnet, T cells integrate local and global cues to discriminate between structurally similar antigens, Cell Rep. **11**, 1208 (2015).

[38] K. E. Tkach, D. Barik, G. Voisinne, N. Malandro, M. M. Hathorn, J. W. Cotari, R. Vogel, T. Merghoub, J. Wolchok, O. Krichevsky, and G. Altan-Bonnet, T cells translate individual, quantal activation into collective, analog cytokine responses via time-integrated feedbacks, eLife **3**, e01944 (2014).

[39] M. Kopf, M. F. Bachmann, and B. J. Marsland, Averting inflammation by targeting the cytokine environment, Nat. Rev. Drug Discovery **9**, 703 (2010).

[40] S. Maher, D. Toomey, C. Condron, and D. Bouchier-Hayes, Activation-induced cell death: The controversial role of Fas and Fas ligand in immune privilege and tumour counterattack, Immunol. Cell Biol. **80**, 131 (2002).

[41] D. C. Huang, M. Hahne, M. Schroeter, K. Frei, A. Fontana, A. Villunger, K. Newton, J. Tschopp, and A. Strasser, Activation of Fas by FasL induces apoptosis by a mechanism that cannot be blocked by Bcl-2 or Bcl-x$_L$, Proc. Natl. Acad. Sci. USA **96**, 14871 (1999).

[42] T. Pradeu and E. L. Cooper, The danger theory: 20 years later, Front. Immunol. **3**, 287 (2012).

[43] D. J. Laydon, C. R. M. Bangham, and B. Asquith, Estimating T-cell repertoire diversity: Limitations of classical estimators and a new approach, Philos. Trans. R. Soc. B **370**, 20140291 (2015).

[44] D. B. Graham, C. Luo, D. J. O'Connell, A. Lefkovith, E. M. Brown, M. Yassour, M. Varma, J. G. Abelin, K. L. Conway, G. J. Jasso, C. G. Matar, S. A. Carr, and R. J. Xavier, Antigen discovery and specification of immunodominance hierarchies for MHCII-restricted epitopes, Nat. Med. **24**, 1762 (2018).

[45] T. P. Riley, L. M. Hellman, M. H. Gee, J. L. Mendoza, J. A. Alonso, K. C. Foley, M. I. Nishimura, C. W. Vander Kooi, K. C. Garcia, and B. M. Baker, T cell receptor cross-reactivity expanded by dramatic peptide-MHC adaptability, Nat. Chem. Biol. **14**, 934 (2018).

[46] K. Schmiedeberg, H. Krause, F.-W. Röhl, R. Hartig, G. Jorch, and M. C. Brunner-Weinzierl, T cells of infants are mature, but hyporeactive due to limited Ca^{2+} influx, PLoS ONE **11**, e0166633 (2016).

[47] Y. Elhanati, A. Murugan, C. G. Callan, T. Mora, and A. M. Walczak, Quantifying selection in immune receptor repertoires, Proc. Natl. Acad. Sci. USA **111**, 9875 (2014).

[48] T. Zahavy, M. Haroush, N. Merlis, D. J. Mankowitz, and S. Mannor, Learn what not to learn: Action elimination with deep reinforcement learning, arXiv:1809.02121.

[49] A. Wang, H. Zhou, W. Xu, and X. Chen, Deep neural network capacity, arXiv:1708.05029.

[50] O. V. Bolkhovskaya, D. Y. Zorin, and M. V. Ivanchenko, Assessing T cell clonal size distribution: A non-parametric approach, PLoS ONE **9**, e108658 (2014).

[51] S. DeWolf, B. Grinshpun, T. Savage, S. P. Lau, A. Obradovic, B. Shonts, S. Yang, H. Morris, J. Zuber, R. Winchester *et al.*, Quantifying size and diversity of the human T cell alloresponse, J. Clin. Invest. Insight **3**. 121256 (2018).

[52] T. Oakes, J. M. Heather, K. Best, R. Byng-Maddick, C. Husovsky, M. Ismail, K. Joshi, G. Maxwell, M. Noursadeghi, N. Riddell *et al.*, Quantitative characterization of the T cell receptor repertoire of naïve and memory subsets using an integrated experimental and computational pipeline which is robust, economical, and versatile, Front. Immunol. **8**, 1267 (2017).

[53] J. M. Heather, K. Best, T. Oakes, E. R. Gray, J. K. Roe, N. Thomas, N. Friedman, M. Noursadeghi, and B. Chain, Dynamic perturbations of the T-cell receptor repertoire in chronic HIV infection and following antiretroviral therapy, Front. Immunol. **6**, 644 (2016).

[54] R. Dessalles, Y. Pan, M. Xia, D. Maestrini, M. R. D'Orsogna, and T. Chou, How heterogeneous thymic output and homeostatic proliferation shape naive T cell receptor clone abundance distributions, arXiv:1906.07463.

[55] J. Desponds, T. Mora, and A. M. Walczak, Fluctuating fitness shapes the clone-size distribution of immune repertoires, Proc. Natl. Acad. Sci. USA **113**, 274 (2016).

[56] G. Altan-Bonnet, T. Mora, and A. M. Walczak, Quantitative immunology for physicists, Phys. Rep. **849**, 1 (2020).

[57] L. Klein, B. Kyewski, P. M. Allen, and K. A. Hogquist, Positive and negative selection of the T cell repertoire: What thymocytes see (and don't see), Nat. Rev. Immunol. **14**, 377 (2014).

[58] E. W. Newell and M. M. Davis, Beyond model antigens: High-dimensional methods for the analysis of antigen-specific T cells, Nat. Biotechnol. **32**, 149 (2014).

[59] J. Kisielow, F.-J. Obermair, and M. Kopf, Deciphering CD4+ T cell specificity using novel MHC–TCR chimeric receptors, Nat. Immunol. **20**, 652 (2019).

[60] V. I. Jurtz, L. E. Jessen, A. K. Bentzen, M. C. Jespersen, S. Mahajan, R. Vita, K. K. Jensen, P. Marcatili, S. R. Hadrup, B. Peters, and M. Nielsen, NetTCR: Sequence-based prediction of TCR binding to peptide-MHC complexes using convolutional neural networks, bioRxiv 433706 (2018).

[61] J. Glanville, H. Huang, A. Nau, O. Hatton, L. E. Wagar, F. Rubelt, X. Ji, A. Han, S. M. Krams, C. Pettus, N. Haas, C. S. L. Arlehamn, A. Sette, S. D. Boyd, T. J. Scriba, O. M. Martinez, and M. M. Davis, Identifying specificity groups in the T cell receptor repertoire, Nature **547**, 94 (2017).

[62] M. Hausknecht and P. Stone, Deep recurrent Q-learning for partially observable MDPs, arXiv:1507.06527.

[63] P. Moscato, On evolution, search, optimization, genetic algorithms and martial arts: Towards memetic algorithms, Caltech Concurrent Computation Program Report No. 826, 1989 (unpublished).

[64] K. He, X. Zhang, S. Ren, and J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in *Proceedings of the IEEE International Conference on Computer Vision* (IEEE, Piscataway, NJ, 2015), pp. 1026–1034.

[65] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. (MIT Press, Cambridge, MA).

[66] O. Nachum, M. Norouzi, K. Xu, and D. Schuurmans, Bridging the gap between value and policy based reinforcement learning, arXiv:1702.08892.

[67] L. L. Campbell, An extended Čencov characterization of the information metric, Proc. Am. Math. Soc. **98**, 135 (1986).