Discovering biophysical principles in latent space representations of immune recognition

Thomas J. Rademaker

Department of Physics
McGill University
Montreal, Québec, Canada

February, 2021

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

To my family

Abstract

The adaptive immune system is a complex biological system acting over many length and time scales. T cells, effector cells of the adaptive immune system, are capable of recognizing minute amounts of pathogens and mounting a grotesque response, while not responding at all to large amounts of self antigens. The response is tightly regulated at the intra-, inter- and extracellular level through intricate protein-protein interaction networks. While the interactions have been described qualitatively, a quantitative understanding is often lacking.

In this thesis, we present computational approaches inspired by physics and machine learning to quantitatively study different aspects of immune recognition. First, using fitness-based parameter reduction, we extract the core module from the intracellular network of immune recognition. Second, using machine learning techniques, we study the sensitivity of immune recognition networks to antagonism, a perturbation to the antigen distribution that prevents T cells from responding to pathogens. We find that the output function of robust immune recognition networks contains a critical point, a finding that informs the design of robust machine learning classifiers.

Finally, we predict antigen quality from cytokine dynamics. We represent the cytokine profile in a latent space and parameterize the latent space using piecewise ballistic models. We validate our model against diverse experimental configurations, providing us with a biological basis for the model parameters. Using these parameters, we predict antigen quality independent of antigen quantity and initial T cell number, providing a reference antigen quality that known baselines cannot provide with a single measurement.

Résumé

Le système immunitaire adaptatif est un système biologique complexe fonctionnant sur de nombreuses longueurs et échelles de temps. Les cellules T, cellules effectrices du système immunitaire adaptatif, sont capables de reconnaître des quantités infimes d'agents pathogènes et de monter une réponse très fort, tout en ne répondant pas du tout à de grandes quantités de soi-antigènes. La réponse est étroitement régulée au niveau intra-, inter- et extracellulaire par des réseaux complexes d'interaction protéine-protéine. Bien que les interactions aient été décrites de manière qualitative, une compréhension quantitative fait souvent défaut.

Dans cette thèse, nous présentons des approches informatiques inspirées de la physique et de l'apprentissage automatique pour étudier quantitativement différents aspects de la reconnaissance immunitaire. Premièrement, en utilisant la réduction des paramètres basée sur la fitness, nous extrayons le module de base du réseau intracellulaire de reconnaissance immunitaire. Deuxièmement, à l'aide de techniques d'apprentissage automatique, nous étudions la sensibilité des réseaux de reconnaissance immunitaire à l'antagonisme, une perturbation de la distribution de l'antigène qui empêche les cellules T de répondre aux agents pathogènes. Nous constatons que la fonction de sortie des réseaux de reconnaissance immunitaire robustes contient un point critique, une découverte qui informe la conception de classificateurs d'apprentissage automatique robustes.

Enfin, nous prédisons la qualité de l'antigène à partir des dynamiques de cytokines, molécules messagères extracellulaires. Nous représentons le profil des cytokines dans un espace latent, paramétrons l'espace latent à l'aide de modèles balistiques par morceaux et études des configurations expérimentales, à partir desquelles nous extrayons une base biologique pour les paramètres du modèle. À partir de ces paramètres, nous prédisons la qualité de l'antigène indépendamment de la quantité d'antigène et du nombre initial de lymphocytes T, fournissant une qualité d'antigène de référence que les lignes de base connues ne peuvent pas fournir avec une seule mesure.

Acknowledgments

First and foremost, I want to express gratitude to my thesis advisor Paul François. He has been a scientific father to me like no other. His forward thinking, his support in dire times, his encouragement on any of my endeavours have shown me what a model scientific advisor can be like. I am especially thankful to Paul for providing me with opportunities wherever possible, and for helping me push forward not just scientific projects but also outreach initiatives.

Then, I want to thank François Bourassa and Félix Proulx-Giraldeau, labmates with whom I had the pleasure to collaborate. My favorite parts of the PhD were our shared coding experiences and discussions on how to proceed following meetings with Paul. I also want to thank other current and former labmates Laurent, Allen, Jimmy, Adrien, Juliette, Andres, Louis, Vincent and Robin for sharing an office with me and shaping the group dynamics.

Thanks to my collaborators, each of whom has helped me form my perspective on science in general, and on our projects specifically. Sooraj Achar and Grégoire Altan-Bonnet for showing me how exciting it is to be at the forefront of an experimental field. Emmanuel Bengio for keeping me on the right track on the machine learning side.

Thanks to Judith Mandl and Nikolas Provatas for serving on the supervisory committee

Regarding the writing, thanks to François and Paul for the detailed feedback, and to Yony Bresler for the LaTeX template of your thesis.

Friends from Rutherford, Physics Outreach, McGill Outdoors Club, those scattered

around Montreal and Geneva, thank you all for being part of my life and forming me as a person. While leaving out many, I would like to specifically mention my dear friends Joel and Saad, and Florestan, David, Céline, Keli and Justin, friends I was fortunate to cohabitate with.

Thanks, Deb and David for taking me into your home and making me part of the family when COVID-19 forced the world indoors.

Papa, mama and Siebrand, thank you for your continuous support from across the Atlantic despite my promise that I would spend my PhD years in Europe. I hope you will bear with me wherever the tide may take me. I will promise I will find my way back when the time is ripe.

Finally, thank you Amy, love of my life, for sharing your life with me. Thank you for supporting me and kicking my ass when I needed it (also when I did not need it). I am excited about our future together.

Contributions and Originality

The contents of this dissertation is based on one first-author publication, one first co-author publication, and on work that features in a manuscript that is currently in preparation for submission. None of this material has been used before in another dissertation. Conform the thesis guidelines of the McGill University - Graduate and Postdoctoral Studies, excerpts from these publications are presented verbatim and in quotations in deviating Helvetica font with the first word of each paragraph indented, adjusted margins and the source of the quote below the last paragraph (*Parameter reduction* [1] and *Attack and Defence* [2] respectively). It is understood that figures and captions in Chapter 2 were taken from [1] and those in Chapter 3 were taken from [2]. All were used with permission of their respective copyright. The manuscript, author contributions and contribution of original knowledge are given below.

- “Untangling the hairball: Fitness-based asymptotic reduction of biological networks” by *Thomas J. Rademaker, Félix Proulx-Giraldeau, and Paul François* [1], published in *Biophysical Journal*.

  I developed and implemented an algorithm for parameter reduction, and reduced three immune recognition models. *FPG* further developed the algorithm and applied it to models of biochemical adaptation. *PF* conceptualized the project, assisted in analysis of the reductions, and wrote the manuscript. *FPG* and I prepared the figures and assisted in writing of parts of the manuscript corresponding to our sections. The publication introduced a novel algorithm for parameter reduction for models based on rate-equations. It shows how immune recognition models are related, and how they are built from the ground up capturing more features the more involved they get. Chapter 2 is based on this manuscript and supplement.

- "Attack and defence in cellular decision-making: Lessons from machine learning" by *Thomas J. Rademaker, Emmanuel Bengio and Paul François* [2], published in *Physical Review X*.

  I developed and implemented the immune recognition model and machine learning algorithms, prepared the figures, and wrote the supplementary material. *PF* conceptualized the project and assisted with analyzing simulation results. *PF* and I together did the analytical modelling and wrote the main text. *EB* assisted with the machine learning part and the writing. In this work, we establish a connection between fooling mechanisms in computer vision and fooling mechanisms in models of immune recognition. We demonstrate how insights in non-neural machine learning translates to other sensory systems. Chapter 3 is based on this manuscript and supplement.

- "Learning the immune manifold from robotic cytokine multiplexing" by *Sooraj Achar, Thomas J. Rademaker, François Bourassa, Paul François and Grégoire Altan-Bonnet* [3], in preparation for submission.

  SA and *GAB* performed all experiments. *FB* designed the processing pipeline assisted by me and *SA*. I designed and validated all aspects of the classification procedure assisted by *FB* and *SA*, I designed the ballistic model assisted by *FB*, and I did the FIM analysis. *PF* and *GAB* conceptualized the project, assisted in the analysis and wrote the manuscript, assisted by *SA, FB* and me. We also prepared the figures and wrote the supplementary material. In this work, we designed a method to classify antigen quality using detailed cytokine kinetics. With our pipeline guided by interpretable machine-learning, we model cytokine dynamics, allowing us to understand diverse experimental configurations. Due to the universal usage of predicting antigen quality in systems of immune activation, we expect this work to see wide applicability. Parts of this manuscript and supplement are based on Chapter 4.

Contents

1

Introduction

This thesis presents work at the interface of biophysics and immunology. To provide perspective on how interdisciplinary work typically proceeds, we highlight the potential of thinking across scientific fields before introducing the immune system, a history of immune recognition models, and the research projects.

1.1 Interdisciplinarity

To explore the relevance of interdisciplinarity, we start off by wondering if a biologist can fix a radio [4]. Engineers have worked for a long time on a comprehensive representation of a multitude of electronic devices. Lazebnik argued that biologists do not have a representation for the building blocks of their systems, and with their toolset, will thus not be able to fix a radio, except in isolated cases when an individual component causes the radio to break down. Molecular biologists might have more success fixing a radio if they would formulate a quantitative language to study the structure and dynamics of function, rather than isolated parts of the cell [5]. Such considerations led to the birth of the quickly broadening field of systems biology. Due to its popularity, systems biology suffered from a definition issue, but in a sense, it allowed biologists to quantitatively assess cellular and organismal functioning, deeply impacting molecular biology [6]. It also led immunology towards systems, computational and quantitative immunology away from the notion that the effect of gene or protein X and Y on Z ought to be studied in isolation with the rest of the system [7].

1.1 Interdisciplinarity

Quantitative studies require appropriate experimental techniques and novel analysis tools, often borrowed from other fields, leading directly into interdisciplinary work.

Oftentimes, breakthroughs in science occur by approaching a problem from a new direction, taking inspiration from methods or insights from a related field. For instance, artificial neural networks were inspired by neurons processing information in the brain, optical microscopy below the diffraction limit was invented by physicists using cutting-edge laser techniques, and mathematical modelling of HIV dynamics provided quantitative understanding of the infection and led to improved therapy [8].

Not just have molecular biologists adopted engineering approaches or have computer scientists borrowed ideas from neuroscience, leading to the boom of machine learning [9], the inverse also occurs. In biomimicry, engineers use nature's solution in modern technology, for instance by equipping the bullet train with the aerodynamic shape of a kingfisher's beak to make it more quiet, or reliably transmitting data underwater over long distances emulating the frequency-modulating acoustics of dolphins.

There are not only success stories: Jonas & Kording showed that current computational methods from neuroscience, when applied to a simulated microprocessor, will not lead one to the circuit diagram of a microprocessor [10]. Most understanding of the global dynamics of the microprocessor was gained through dimensional reduction, where the discovered components led to variables of interest. Krakauer et al. agreed that neuroscience needs to let go of the reductionist approach and focus on function and behavior [11] through three levels of analysis (computational, algorithmic and implementational level), as defined by Marr & Poggio [12].

Inspired by Lazebnik [4] and Jonas & Kording [10], and recognizing that the ultimate goal of the immune system is to protect an organism, we could ask ourselves if an immunologist could support a cyber-security system. This idea is not new: artificial immune systems have been used as classifiers since 1986 [13], specifically as virus detectors [14] or network intrusion detectors [15]. Without going into any detail on the cyber-security side of the analogy, let us proceed regardless in broad strokes and focus on where current approaches in immunology fall short.

The different cell types are the fundamental units that constitute the organism's defence. Through precise developmental schemes, learning algorithms, regulatory mechanisms and homeostatic processes using two separate but connected systems, the immune system is able to withstand unknown threats the environment poses. Sometimes, the danger comes from within through auto-immune disorders or evasive tumor cells. Complex interacting systems are implemented, regulated and fine-tuned at many levels (intracellular, extracellular, tissue, organismal), allowing Höfer and Altan-Bonnet to observe that "immunology is such a quintessential science that 'systems immunology' could be considered a tautology" [16]. Here we reveal the limits of immunology as described by immunologists: descriptions are well-established for the intracellular protein-protein interaction networks and the population dynamics between cell types, but there is at most an incomplete understanding of how interaction networks cause immune activation and population-level dynamics on time scales from hours to years. Immunologists have the tools capable of dissecting a system in detail, but relating the parts to the whole is a problem of a different dimension and requires quantitative thinking and tools and ideas from across disciplines.

In conclusion, it seems like an immunologist can assist the immune system, and equally well might be able to assist a cyber-security system, although there exists threats (coronaviruses, pancreatic cancer, certain metastasized cancers, autoimmune disorders) for which there is currently no defence (cure). Moreover, bridging the connections between molecular, cellular and organismal scales is hard. Immunology is a field in progress, as it has been for as long as medicine exists, and will continue to provide solutions to challenges posed by the environment. Contributing to this progress are interdisciplinary approaches, which form the heart and soul of the work described in this thesis.

In the remainder of the introduction we provide a short overview of the adaptive immune system, leading into the research questions and an outline of the thesis.

1.2 Short overview of the adaptive immune system

Introducing well-known and basic immunology, this section is mainly inspired by [17] and [18]. All organisms have a rudimentary form of an immune system to mitigate threats from the environment, but only vertebrates have an adaptive immune system. The adaptive immune system contrasts the innate immune system in that it is not fixed at birth, but is able to adapt to challenges posed by the environment during an organism's lifetime. A unique feature of the adaptive immune system is that following exposure to antigens (also called peptides or ligands) of foreign origin, immunological memory is formed. The next time antigens of this type are encountered, an immune response will be invoked much faster than the first time this antigen was encountered.

The adaptive immune system consists mainly of B cells and T cells, both of which can turn into memory cells. Before activation, B cells and T cells are morphologically indistinguishable, but they differ in their function. The main effector function of B cells is to produce antibodies, which, when bound, neutralize viruses or mark pathogens, providing a target for phagocytes. Phagocytes are a part of the innate immune system, consisting among others of macrophages, which are most efficient in engulfing pathogens, and dendritic cells, professional antigen presenting cells (APCs). Upon engulfing pathogens, dendritic cells degrade the material, and chop it up into small chains of amino acids, the antigens. They then migrate to lymph nodes where T cells reside for antigen presentation. This action provides the link between the innate and adaptive immune system.

APCs display antigens from the pathogens on large extracellular constructs called major histocompatibility complexes (MHCs) and express co-stimulatory molecules to support T cell activation. MHCs are designed specifically to display antigens (or peptides) to T cells via a peptide-MHC (pMHC) complex. T cells scan APCs for pMHCs to bind to with their T cell receptors (TCRs), and when the displayed antigen is specific to the TCR, and the co-stimulatory molecules on the APCs have bound to the co-stimulatory receptors on the T cell, the T cell activates. T cell activation results in a range of actions like altered gene expression, growth, proliferation, cytokine production and migration to the site of infection.

The T in T cells is short for thymus, which is where they mature[1]. T cell precursors are derived from multipotent hematopoietic stem cells in the bone narrow that migrate to the thymus. In the thymus, the TCR is formed through a combinatorial process called V(D)J

[1] In contrast, B cells mature in the bone marrow.

recombination and selected if the strength of this newly-formed TCR to a series of self pMHCs is neither too strong (negative selection), nor too small (positive selection) [19]. Having passed these tests, the T cell becomes part of the pool of naive T cells that resides in lymphoid organs and circulates the lymphatic system in search of antigens specific to their TCR. This forms the basis of clonal selection theory: every antigen activates T cell clones with the TCR the antigen is specific to, and every TCR is specific to only one antigen.

There is a caveat to that, which is that T cells do not actually respond to only one amino acid sequence [20]. It was shown experimentally that T cells respond on average to $3 \cdot 10^4$ antigens [21]. This is needed to provide full coverage of the antigenic shape space, because there are many more possible pMHCs class I peptides (a conservative estimate is $\sim 10^{16}$ [22]) than a human has unique clonotypes ($\sim 10^{10}$ [23]). Too much cross-reactivity would result in too strong negative selection, making it seem “plausible that evolutionary pressures might have optimized this trade-off and determined the degree to which TCR can respond to multiple pMHC” [24].

So far, we have only introduced self or negatively selecting antigens (nonagonists) and not self or positively selecting antigens (agonists). According to Feinerman et al. [25], there exists a third category of antigens called antagonists, to which T cells mount a small response if encountered by themselves, but which lower the response to agonists when T cells encounter antagonists together with agonists. This begs the question: is the T cell response binary (specific or not specific) or does it allow for a continuum of responses? Throughout the thesis, we approach this question from various angles.

1.3 History of immune recognition models

Over the past few decades, a variety of models has been proposed to explain the mechanism for enhanced specificity in immune recognition. The first model we examine is the kinetic proofreading (KPR) model proposed by Timothy McKeithan [26]. His proposal is inspired by the KPR mechanism explaining the high fidelity in DNA replication [27, 28]. The basepair d bound to ATP into a dATP complex associates with the basepair a of the DNA strand that is replicated (Fig. 1.1). Before the basepair d is accepted, during an intermediate step, a diphosphor molecule is removed from the a - dATP complex. Right after association and after this intermediate step, the dATP or dAMP complex may dissociate from the basepair a. Since correct matching basepairs have a higher affinity for the basepair of the replicated DNA strand than incorrect basepair, they are more likely to stay bound. The proofreading step squares the fidelity rate of DNA replication from 10^{-4} to 10^{-8}.

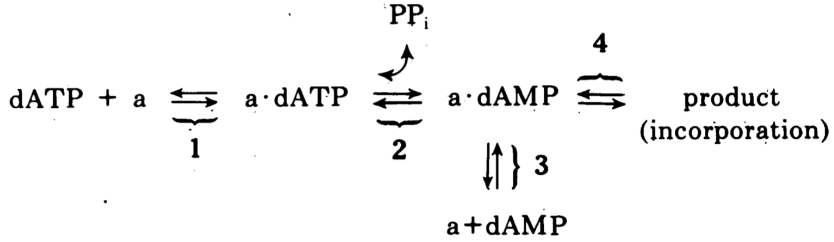


Figure 1.1: **Kinetic proofreading in DNA replication.** The basepair ATP complex dATP binds to the DNA strand's basepair a (1), from where it can either dissociate, lose a diphosphor molecule (2) before dissociating as an dAMP (3) or be incorporated (4). Figure from [27].

Despite a minor difference in specificity between nonagonists and agonists, multiple phosphorylations of intracellular chains extending from the TCR (the proofreading steps) allow for a specific response over many orders of magnitude of antigen quantity. McKeithan already noted that additional negative feedback must exist to explain antagonism, inhibition of the immune response by peptides called antagonists that by themselves do not trigger the T cell but inhibit the response to agonists. This feature that was known to exist from studies in HIV [29]. The mechanism that he proposed was that through binding of antagonists small clusters of TCRs are formed, which pass on an inhibitory signal [30], and may negate the activating signal transmitted through large clusters of TCR that form when bound by agonists.

It is commonly accepted that a KPR cascade forms the fundament of immune recognition through TCR activation. A natural extension of this model is to characterize all interactions occurring in the TCR signalling pathway. This heroic effort was completed by Altan-Bonnet and Germain [31]. Through a mix of experimental and computational work, they estimated hundreds of reaction rates and were able to simulate a complete immune recognition model. A summary of the model is given in Fig. 1.2. The activation signal travels through the network through multiple phosphorylations of the pMHC - TCR complex (KPR steps), while modulated by negative and positive feedback. A simpler version of this model was proposed by Lipniacki et al. [32], capturing the main features and preserving the main motifs of the model. A next iteration on this model is provided by François et al. [33]. Here the negative feedback module modulated by SHP-1 is coupled to a KPR backbone. It is understood that a subsequent positive threshold on the output could create the bistability observed in Refs. [31, 32].

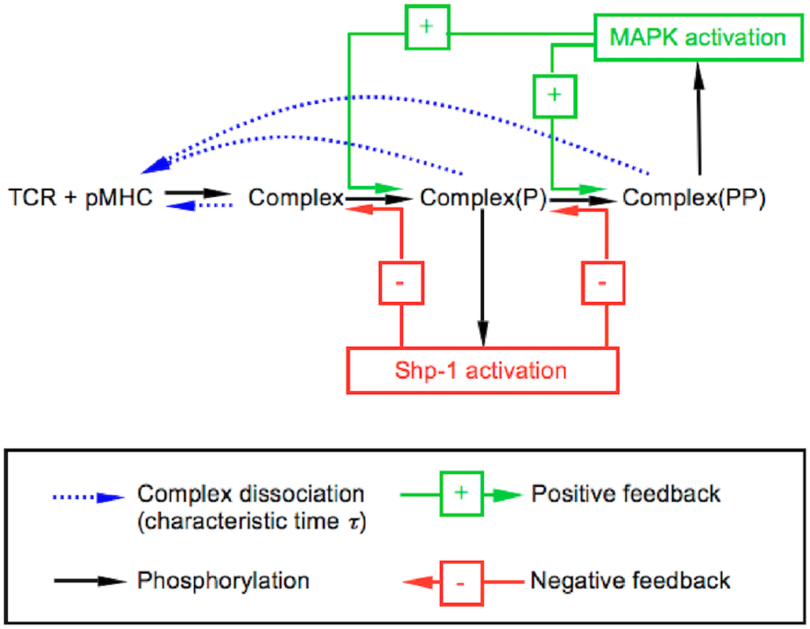


Figure 1.2: **Coarse-grained version of the model** A tug of war between a positive feedback system modulated by MAPK activation and negative feedback modulated by SHP-1 along the KPR cascade determines T cell activation by the pMHC - TCR complex. Figure from [31].

The adaptive module that couples to the KPR cascade to be able to discriminate agonists from nonagonists over a large range in antigen quantity has been investigated using various approaches, i.e. through an evolutionary search [34], a comparison between phenotypic models [35] and a systematic search of possible models [36]. In these approaches, the authors did not specify the protein that modulates the feedback, although suggestions were given. Recent experimental and theoretical work assigns importance to the role of ZAP70 in immune recognition. Optogenetics revealed that ZAP70 recruitment occurs before the KPR cascade activates and was thus solely dependent on binding of the pMHC to the TCR, not on the difference between nonagonist and nonagonist [37], and thus does not play a role in antigen discrimination. Similarly, a recent model proposes a mechanism for the role of ZAP70 in the reset of the microenvironment around the intracellular TCR chains following dissociation of the pMHC-TCR complex [38]. In contrast, Ganti et al. argue that feedback loops through ZAP70 recruitment and related proteins are at the core of the information processing in T cell signalling network [39].

Others have considered biophysical interactions to lie at the fundament of immune recognition. The biophysical interactions of TCR triggering can be summarized as aggregation of the TCR chains, conformational changes of the TCR and segregation of the TCR [40]. Some of these mechanisms were refuted by considering a system where TCRs were introduced in a nonimmune cell and the ligand-specific pathway was reconstituted [41]. Recently, it was also found that through dynamic TCR bonds (so called catch and slip bonds) T cells could distinguish between nonagonists and agonists during thymic selection [42].

Finally, also models containing both biophysical interactions and TCR signalling have been proposed. For instance, TCR triggering can be extended through coupling of the positive feedback loop of ZAP70 binding to TCR chains with a crowded intracellular environment [43]. Moreover, an alternative for enhanced sensitity of immune recognition with a KPR cascade alone is by considering rebinding of dissociated pMHCs [44, 45].

The latest developments show that the field has still not reached a consensus on what parts of the T cell signalling network are responsible for immune recognition, and if it is even due to the signalling network, biophysical interactions or a mix. For each proposed mechanism, the authors have carefully considered observations based on experimental evidence generated within their lab or by their collaborators and determined what model could best explain such data. The diversity of proposed mechanisms could mean that some of the groups pursue the wrong approach and only one of them has the correct idea. More likely is that the researchers are all correct within the regime they investigate. For instance, the information-theoretic model validation by Ganti et al. only considers discrimination between nonagonist and agonists [39], while it is well possible that the T cell also has the ability to discriminate antagonists. Moreover, as has been noted by Lever et al. *published data are incomplete, with experiments typically using only a small panel of TCRs (or pMHC complexes) with a limited range of affinities, and a single or just a few different doses of antigen or pMHC* [35]. This is why an attempt to quantify the discriminatory power of the T cell through a comparison between published experimental data is so compelling [46]. With more advanced methods to manipulate T cells becoming readily available, as well as the ability to generate large timeseries of immune responses using robotic measurements, the last word on the matter has not been spoken.

1.4 Outline

Immunologists have been highly successful in describing isolated parts of their system, like the signal transduction pathways concerned with T cell activation following pMHC-TCR binding [31]. Yet, falling in the pitfall described by Krakauer et al. for neuroscience [11], knowing all interactions between proteins in the TCR signalling pathway does not answer the question how a T cell discriminates between a self antigen and a not self antigen, which is the function encoded in this network. In essence, this is a simple question that does not require an answer using high-dimensional data. The manifold hypothesis states that high-dimensional data can often be described in a lower dimensional space. In statistics (nowadays machine learning), manifold learning techniques are routinely used to uncover low-dimensional descriptions [47]. In physics, effective field theories and renormalization groups are used to reduce the number of degrees in a freedom in a model allowing for a simpler description (approximate or exact at certain energy, time or length scales) [48]. The low-dimensional description (be it of the data or of the model) are referred to as latent spaces in this thesis. Inspired by renormalization group, we use parameter reduction to find a latent space description of the TCR signalling pathway, allowing us to give a simple answer to the question how a T cell discriminates between self and not self (Chapter 2). From this latent space we derive how T cells alleviate the effect of antagonists (Chapter 3), and relate this to a similar phenomenon in machine learning. Finally, we connect immune recognition by individual T cells to the cytokine response at the population level. We uncover a low-dimensional description of the cytokine response that is most informative of the antigen with which the T cells were stimulated, we build a phenomenological model of the latent space, and we show that the specificity of T cells to antigen is a continuous measure (Chapter 4).

The main body of this thesis is structured as follows. In Chapter 2, using parameter reduction, we map three immune recognition networks to a network consisting of two simple modules. These modules are the minimum requirements for an immune network to recognize antigen quality independent of antigen quantity. The finding that antagonistic effects are mitigated only when the immune recognition network is of sufficient complexity led us to further investigate antagonism in Chapter 3. We study this through the lens of machine learning. In computer vision, a small, deliberate perturbation that does not change the input visually, ten out of ten times changes the classification of the input for naive classifiers. We analyze the conditions immune recognition networks must satisfy to become less sensitive to antagonistic effects, and relate these conditions to the sensitivity of neural networks to such perturbations. In Chapter 4 we study the cytokine response of T cells for a variety of experimental setups. We first represent the cytokine dynamics in a latent space, next we model the curves using ballistic equations, and interpret the fitted parameters in terms of the experimental setup. These parameters are then used to predict antigen quality. Finally, we provide an outlook and discuss future work in Chapter 5.

2

Parameter reduction

“ Complex mathematical models of interaction networks are routinely used for prediction in systems biology. However, it is difficult to reconcile network complexities with a formal understanding of their behavior. Here, we propose a simple procedure (called $\bar{\phi}$) to reduce biological models to functional submodules, using statistical mechanics of complex systems combined with a fitness-based approach inspired by *in silico* evolution. $\bar{\phi}$ works by putting parameters or combination of parameters to some asymptotic limit, while keeping (or slightly improving) the model performance, and requires parameter symmetry breaking for more complex models. We illustrate $\bar{\phi}$ on biochemical adaptation and on different models of immune recognition by T cells. An intractable model of immune recognition with close to a hundred individual transition rates is reduced to a simple two-parameter model. $\bar{\phi}$ extracts three different mechanisms for early immune recognition, and automatically discovers similar functional modules in different models of the same process, allowing for model classification and comparison. Our procedure can be applied to biological networks based on rate equations using a fitness function that quantifies phenotypic performance. ”

(Parameter reduction [1])

2.1 Introduction

Immune recognition by T cells can be posed as a simple problem: “Is there a not self ligand present in this mixture of ligands?” If so, activate, if not remain quiescent. This deceptively simple problem is constrained from various angles: the T cell response needs to be specific, sensitive and fast [49]. Complicating factors are the structural similarity between self and not self ligands, the composition of the mixture and inherent biochemical constraints. Self and not self ligands may differ by a single amino acid causing the pMHC-TCR binding time to reduce by an order of magnitude or less, while few not self ligands, if any, are present among an abundance of self ligands. Finally, T cells needs to make a decision based on the collective state of its TCRs. It cannot measure individual pMHC-TCR binding times. It may not come as a surprise then that the T cell activation pathway has evolved to a network with hundreds of parameters and equations [31]. Yet, it remains unclear what part of this network actually implements the core functional module. More coarse-grained networks of immune recognition exists [32–34], making us wonder what the similarities between the networks are, and if there exists a common mechanism for measuring binding time independent of ligand concentration.

A technique for systematically reducing the number of parameters of a model fit on experimental or simulated data is parameter space compression [50], as demonstrated by Transtrum et al. through a boundary manifold approach on the EGFR signalling pathway [51] or enzyme kinetics [52]. Motivated by their success, we propose a simple method to coarse-grain phenotypic models based on rate-equations by optimizing fitness. Fitness quantifies how well a biological function is implemented. This allows us to extract the core functional modules of various models of immune recognition that are responsible for the specific and sensitive response. We then classify and categorize the models varying in complexity, providing insight into the principles and constraints of immune recognition.

2.2 Materials and methods

An algorithm for fitness based asymptotic reduction

> “ Transtrum & Qiu [51, 52] studied the problem of data fitting using cellular regulatory networks modelled as coupled ordinary differential equations. They proposed that models can be reduced by following geodesics in parameter space, using error

fitting as the basis for the metric. This defines the Manifold Boundary Approximation Method (abbreviated as MBAM) that extracts the minimum number of parameters compatible with data [51].

While simplifying models to fit data is crucial, it would also be useful to have a more synthetic approach to isolate and identify functional parts of networks. This would be especially useful for model comparison of processes where abstract functional features of the models (e.g. the qualitative shape of a response) might not correspond to one another, or where the underlying networks are different while they perform the same overall function [53]. We thus elaborate on the approach of [51] and describe in the following an algorithm for FItness Based Asymptotic parameter Reduction (abbreviated as FIBAR or $\bar{\phi}$). $\bar{\phi}$ does not aim at fitting data, but focuses on extracting functional networks, associated to a given biological function. To define biological function, we require a general fitness (symbolized by ϕ) to quantify performance. Fitness is broadly defined as a mathematical quantity encoding biological function in an almost parameter independent way, which allows for a much broader search in parameter space than traditional data fitting (examples are given in the next sections). The term fitness is inspired by its use in evolutionary algorithms to select for coarse-grained functional networks [54]. We then define model reduction as the search for networks with as few parameters as possible optimizing a predefined fitness. There is no reason *a priori* that such a procedure would converge for arbitrary networks or fitness functions: it might simply not be possible to optimize a fitness without some preexisting network features. A more traditional route to optimization would rather be to increase the number of parameters to explore missing dimensions, rather than decrease them (see discussions in [55, 56]). We will show how $\bar{\phi}$ reveals network features in known models that were explicitly designed to perform the fitness of interest.

Due to the absence of an explicit cost function to fit data, there is no equivalence in $\bar{\phi}$ to the metric in parameter space in the MBAM allowing to incrementally update parameters. However, upon further inspection, it appears that most limits in [51] correspond to simple transformations in parameter space: single parameters disappear by putting them to 0 or ∞, or by taking limits where their product or ratio are constant while individual parameters go to 0 or ∞. In retrospect, some of these transformations can be interpreted as well-known limits such as quasi-static assumptions or dimensionless reduction, but there are more subtle transformations, as will appear below.

Instead of computing geodesics in parameter space, we directly probe asymptotic limits for all parameters, either singly or in pair. Practically, we generate a new parameter set by multiplying and dividing a parameter by a large enough rescaling factor f (which is a parameter of our algorithm, we have taken $f = 10$ for the simulations presented here), keeping all other parameters constant, or doing the same operation on a couple of parameters.

At each step of the algorithm, we compute the behavior of the network when changing single parameters, or any couple of parameters by factor f in both directions. We then compute the change of fitness for each of the new models with changed parameters. In most cases, there are parameter modifications that leave the fitness unchanged or even slightly improve network behavior. Among this ensemble, we follow a conservative approach and select (randomly or deterministically) one set of parameter modifications that minimizes the fitness change. We then implement parameter reduction by effectively pushing the corresponding parameters to 0 or ∞, and iterate the method until no further reduction enhances the fitness or leaves it unchanged, or until all parameters are reduced. The evaluation of these limits effectively removes parameters from the system while keeping the fitness unchanged or incrementally improving it. There are technical issues we have to consider: for instance, if two parameters go to ∞ some numerical choices have to be made about the best way to implement this. Our choice was to keep the reduction simple : in this example, instead of defining explicitly a new parameter, we increase both parameters to a very high value, freeze one of them, and allow variation of the other one for subsequent steps of the algorithm. Another issue with asymptotic limits for rates is that corresponding divergence of variables might occur. To ensure proper network behavior, we thus impose overall mass conservation for some predefined variables, e.g. total concentration of an enzyme (which effectively adds fluxes to the free form of the considered biochemical species). We also explicitly test for convergence of differential equations and discard parameter modifications leading to numerical divergences. Details on the implementation of the reduction rules for specific models

> are presented in Appendix A and can be automatically implemented for any model based on rate equations.
>
> These iterations of parameter changes alone do not always lead to simpler networks. This is also observed in the MBAM when it is sometimes no longer possible to fit all data as well upon parameter reduction. However, with the goal to extract minimal functional networks, we can circumvent this problem by implementing what we call “symmetry breaking" of the parameters (Fig. 2.1 B-C): in most networks, different biochemical reactions are assumed to be controlled by the same parameter. An example is a kinase acting on different complexes in a proofreading cascade with the same reaction rate. However, an alternative hypothesis is that certain steps in the cascade are recognized to activate specific pathways, or targeted for removal (e.g. in “limited signalling models", the signalling step is specifically tagged, thus having dual specificity [35]). So to further reduce parameters, we assume that those rates, which are initially equal, can now be varied independently by $\bar{\phi}$ (Fig. 2.1 C). Symmetry breaking in parameter space allow us to reduce models to a few relevant parameters/equations, and as explained below are necessary to extract simple descriptions of network functions. Note that symmetry breaking transiently expand the number of parameters, allowing for a more global search for a reduced model in the complex space of networks. Fig. 2.1 A summarizes this asymptotic reduction. ”
>
> *(Parameter reduction [1])*

Defining the fitness

To illustrate the $\bar{\phi}$ algorithm, we apply it to the problem of absolute discrimination. In this section we briefly describe absolute discrimination and define the associated fitness function.

> “ Absolute discrimination is defined as the sensitive and specific recognition of signalling ligands based on one biochemical parameter. Possible instances of this problem can be found in immune recognition between self and not self for T cells [25, 49] or mast cells [57], and recent works using chimeric DNA receptor confirm sharp thresholding based on binding times [58]. More precisely, we consider models where

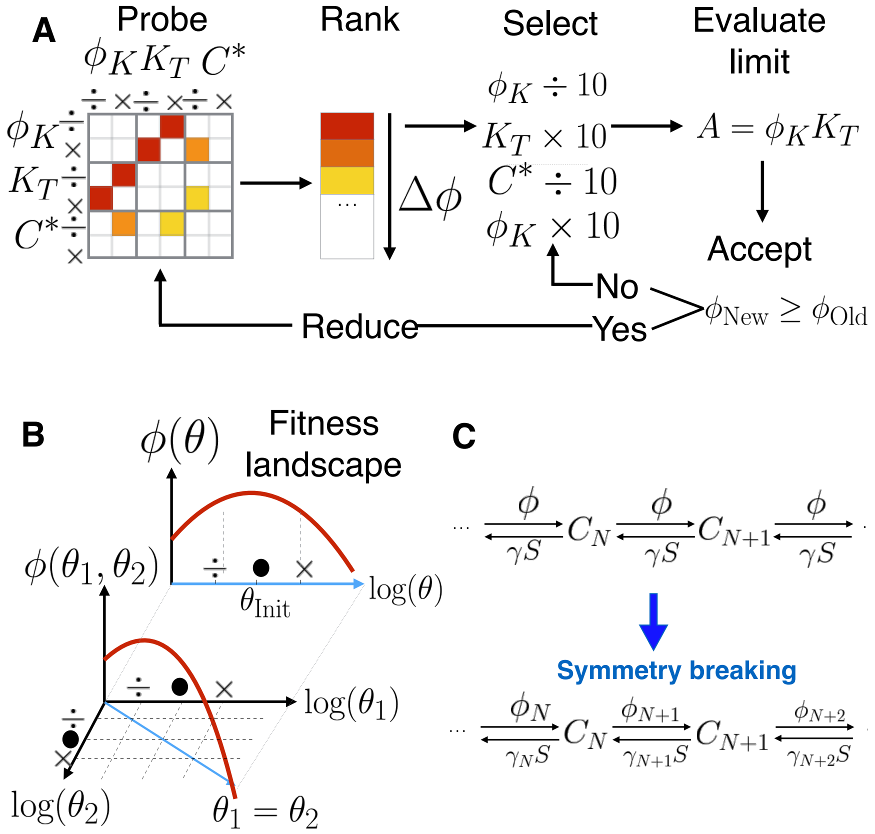

A
Probe
$\phi_K K_T C^*$
÷ × ÷ × ÷ ×
ϕ_K
K_T
C^*
Rank
...
$\Delta\phi$
Select
$\phi_K \div 10$
$K_T \times 10$
$C^* \div 10$
$\phi_K \times 10$
Evaluate limit
$A = \phi_K K_T$
Accept
No
Yes
Reduce
$\phi_{\mathrm{New}} \geq \phi_{\mathrm{Old}}$
B
$\phi(\theta)$
Fitness landscape
$\phi(\theta_1, \theta_2)$
÷ ● ×
θ_{Init}
$\log(\theta)$
$\log(\theta_1)$
$\log(\theta_2)$
$\theta_1 = \theta_2$
C
$\cdots \underset{\gamma S}{\overset{\phi}{\rightleftarrows}} C_N \underset{\gamma S}{\overset{\phi}{\rightleftarrows}} C_{N+1} \underset{\gamma S}{\overset{\phi}{\rightleftarrows}}$
Symmetry breaking
$\cdots \underset{\gamma_N S}{\overset{\phi_N}{\rightleftarrows}} C_N \underset{\gamma_{N+1} S}{\overset{\phi_{N+1}}{\rightleftarrows}} C_{N+1} \underset{\gamma_{N+2} S}{\overset{\phi_{N+2}}{\rightleftarrows}}$

Figure 2.1: **Summary of the $\bar{\phi}$ algorithm.** (A) Asymptotic fitness evaluation and reduction: for a given network, the values of fitness ϕ are computed for asymptotic values of parameters or couples of parameters. If the fitness is improved (warmer colors), one subset of improving parameters is chosen and pushed to its corresponding limits, effectively reducing the number of parameters. This process is iterated. See main text for details. (B) Parameter symmetry breaking: a given parameter present in multiple rate equations (here θ) is turned into multiple parameters (θ_1, θ_2) that can be varied independently during asymptotic fitness evaluation. (C) Examples of parameter symmetry breaking, considering a biochemical cascade similar to the model from [33]. See main text for comments.

> a cell is exposed to an amount L of identical ligands, where their binding time τ defines their quality. Then the cell should discriminate only on τ, i.e. it should decide if τ is higher or lower than a critical value τ_c *independently* of ligand concentration L. This is a nontrivial problem, since many ligands with binding time slightly lower than τ_c should not trigger a response, while few ligands with binding time slightly higher than τ_c should. Absolute discrimination has direct biomedical relevance, which explains why there are models of various complexities, encompassing several interesting and generic features of biochemical network (biochemical adaptation, proofreading, positive and negative feedback loops, combinatorics, etc.). Such models serve as ideal tests for the generality of $\bar{\phi}$. " *(Parameter reduction [1])*

To quantify how well a network performs absolute discrimination, we commence with computing the dose-response curves of the output O (Fig. 2.2 A). Absolute discrimination is only possible if few values of τ correspond to a given Output value $O(L, \tau)$ (as detailed in [49]). Intuitively, this is not possible for monotonic dose response curves (top panel of Fig. 2.2 A): for any value of output O, one can find many associated couples of (L, τ). Thus, ideal performance corresponds to separated horizontal lines, encoding different values of O for different τ independent of L (bottom panel Fig. 2.2 A). A function that maps the amount of overlap to a continuous variable is the mutual information between O and τ.

To compute mutual information, we set up a probabilistic framework. Let us sample from ligands with binding time τ_i, their concentrations L log-uniformly distributed. For every pair (L, τ_i), there exists a correponding O. The histogram of network outputs O (Fig. 2.2 B) is our approximation of the marginal probability distribution $p(O|\tau)$. We sample from two equiprobable τ_i and compute the mutual information between O and τ as the difference between the classical Shannon entropy $H(\tau)$ and the conditional entropy $H(\tau|O)$

$$\mathcal{I}(O, \tau) = H(\tau) - H(\tau|O). \tag{2.1}$$

Here, $H(\tau) = -\sum_i \log_2 p(\tau_i) = -2 \cdot \log_2 1/2 = 1$ and

$$H(\tau|O) = -\sum_{i,O} p(\tau_i|O) \log\left(p(\tau_i|O)\right) = -\sum_{i,O} \frac{p(O|\tau_i)p(O)}{p(\tau_i)} \log \frac{p(O|\tau_i)p(O)}{p(\tau_i)} \tag{2.2}$$

where we have used Bayes theorem.

> " Mutual information measures how much information we can recover from one variable knowing the other. For instance, when $\mathcal{I}(O, \tau) = 0$, it means we cannot recover information on the value of τ by observing O, which would be the case when both distributions are equal $p(O|\tau_1) = p(O|\tau_2)$. Conversely, when the two distributions are fully separated, we can fully recover τ by observing O. Then the mutual information is at its maximum of 1 bit (bottom panel Fig. 2.2 B). For partially overlapping distributions (top panel Fig. 2.2 B), the mutual information varies gradually between 0 to 1. Mutual information allows us to focus only on the respective positions of the distributions, and not on their shape or moments. We thus quantify the discriminatory phenotype irrespective of other parameters.
>
> During the reduction, we typically sampled 50 log-uniformly distributed L on the interval $[1, 10^4]$ and binned the resulting outputs O in 40 log-uniformly distributed bins in the range $[10^{-2}, 10^2]$. The results are largely independent from the number of bins or the range of the bins, as long as O remains in the neighborhood of biologically feasible values, the working range of the initial networks. Partly due to this loose constraint, the output of the reduced networks was near the output of the initial networks. " *(Parameter reduction [1])*

We have run $\bar{\phi}$ on three different models of absolute discrimination:

> " adaptive sorting with one proofreading step [34], a simple model based on feedback by phosphatase SHP-1 from [33] ("SHP-1 model"), and a complex realistic model accounting for multiple feedbacks from [32] ("Lipniacki model"). Initial models are described in more details in following sections. We have taken published parameters as initial conditions. Those three models were all explicitly designed to describe absolute discrimination, modelled as sensitive and specific sensing of ligands of a given binding time τ [49], so ideally those networks would have perfect fitness. However due to various biochemical constraints, these three models have very good initial (but not necessarily perfect) performance for absolute discrimination. We see that after some initial fitness improvement, $\bar{\phi}$ reaches an optimum fitness within a

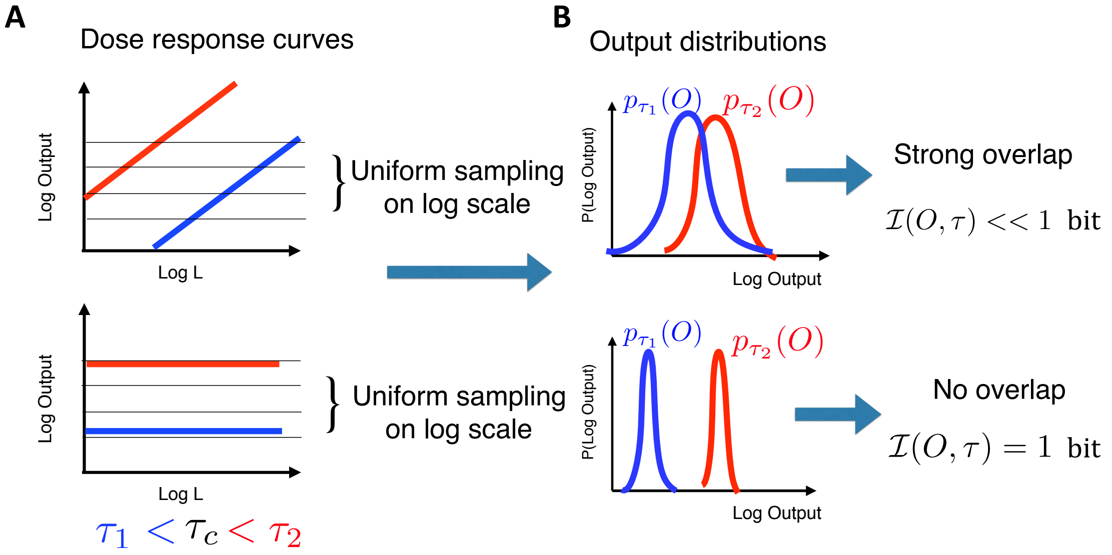


Figure 2.2: **Illustration of the fitness.** (A) Schematics of response line for absolute discrimination. We represent expected dose response curves for a "bad" (top) and a "good" (bottom) model. Response to different binding times τ are symbolized by different colors. For the "bad" monotonic model (e.g. kinetic proofreading [26]), by setting a threshold (horizontal dashed line), multiple intersections with different lines corresponding to different τs are found, which means it is not possible to measure τ based on the Output. Bottom corresponds to absolute discrimination: flat responses plateau at different Output values easily measure τ. Thus, the network can easily decide the position of τ with respect to a given threshold (horizontal dashed line). (B) For actual fitness computation, we sample the possible values of the Output with respect to a predefined Ligand distribution for different τs (we have indicated threshold similar to panel (B) by a dashed line). If the distribution are not well separated, one can not discriminate between τs based on Outputs and $\mathcal{I}(O,\tau) \ll 1$. If they are well separated, one can discriminate τs based on Output and $\mathcal{I}(O,\tau) = 1$.

few steps and thus merely simplifies models while keeping constant fitness. We have tested $\bar{\phi}$ with several parameters of the fitness functions, and we give in the following for each model the most simplified networks obtained with the help of those fitness functions. Complementary details and other reductions are given in Appendix A.

For each model, $\bar{\phi}$ succeeds in fully reducing the system to a single equation with essentially two effective parameters, final model is given in the FINAL OUTPUT formula, and discussion of the effective parameters in the section "Comparison and categorization of models"). However, to help understanding the mathematical structure of the models, it is helpful to deconvolve some of the reduction steps from the final model. In particular, this helps to identify functional submodules of the network that perform independent computations. Thus for each example below, we give a small set of differential equations capturing the functional mechanisms of the reduced model . In Figures we show in the "FINAL" panel the behaviour of the full system of ODEs including all parameters (but potentially very big or very small values after reduction), and thus including local flux conservation. "

(Parameter reduction [1])

2.3 Results

$\bar{\phi}$ for adaptive sorting

" We now proceed with applications of $\bar{\phi}$ to the more challenging problem of absolute discrimination. Adaptive sorting [34] is one of the simplest models of absolute discrimination. It consists of a one-step kinetic proofreading cascade [26] (converting complex C_0 into C_1) combined to a negative feedforward interaction mediated by a kinase K, see Fig. 2.3 A for an illustration. A biological realization of adaptive sorting exists for FCR receptors [57].

This model has a complete analytic description in the limit where the backward rate from C_1 to C_0 cancels out [34]. The dynamics of C_1 is then given by:

$$\dot{C}_1 = \phi_K K C_0(L) - \tau^{-1} C_1 \qquad \text{with} \qquad K = K_T \frac{C^*}{C_0(L) + C^*} \tag{2.3}$$

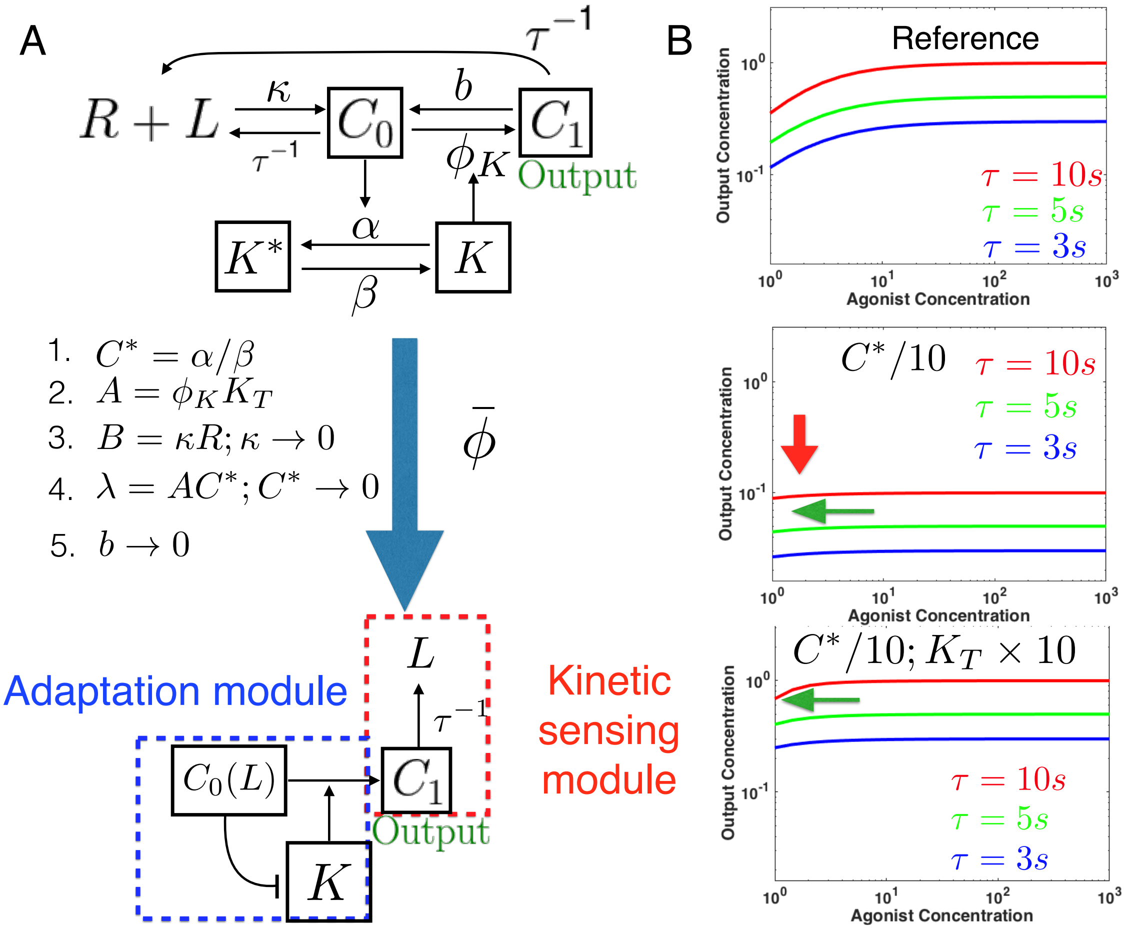


Figure 2.3: **Reduction of Adaptive sorting.** (A) Sketch of the network, with 5 steps of reductions by $\bar{\phi}$. Adaptation and kinetic sensing modules are indicated for comparison with reduction of other models. (B) Illustration of the specificity/response trade-off solved by Step 4 of $\bar{\phi}$. Compared to the reference behavior (top panel), decreasing C^* (middle panel) increases specificity with less L dependency (horizontal green arrow) but globally reduces signal (vertical red arrow). If K_T is simultaneously increased (bottom panel), specificity alone is increased without detrimental effect on overall response, which is the path found by $\bar{\phi}$.

K is the activity of a kinase regulated by complex $C_0(L)$, itself proportional to ligand concentration L. K activity is repressed by C_0 (Fig. 2.3, Eq. 2.3), implementing an incoherent feedforward loop in the network (full system of equations are given in Appendix A).

Absolute discrimination is possible when C_1 is a pure function of τ irrespective of L (so that C_1 encodes τ directly) as discussed in [34, 49]. *A priori*, both C_0 and C_1 depend on the input ligand concentration L. If we require C_1 to be independent of L, the product KC_0 has to become a constant irrespective of L. This is possible because K is repressed by C_0, so there is a “tug-of-war" on C_1 production between the substrate concentration C_0, and its negative effect on K. In the limit of large enough C_0, K is indeed becoming inversely proportional to C_0, giving a production rate of C_1 independent of L. τ dependency is then encoded in the dissociation rate of C_1 so that in the end C_1 is a pure function of τ.

The steps of $\bar{\phi}$ for adaptive sorting are summarized in Fig. 2.3 A. The first steps correspond to standard operations: step 1 is a quasi-static assumption on kinase concentration, step 2 brings together parameters having similar influence on the behavior, and step 3 is equivalent to assuming receptors are never saturated. Those steps are already taken in [34], and are automatically rediscovered by $\bar{\phi}$. Notably, we see that during reduction several effective parameters emerge, e.g. parameter $A = K_T\phi$ can be identified in retrospect as the maximum possible activity of kinase K.

Step 4 is the most interesting step and corresponds to a nontrivial parameter modification specific to $\bar{\phi}$, which simultaneously reinforces the two tug-of-war terms described above, so that they balance more efficiently. This transformation solves a trade-off between sensitivity of the network and magnitude in response, illustrated in Fig. 2.3 B. If one decreases only parameter C^*, the dose response curves for different τs become flatter, allowing for better separation of τs (i.e. specificity), Fig. 2.3 B, middle panel. However, the magnitude of the dose response curves is proportional to C^* so that if we were to take $C^* = 0$, all dose response curves would go to 0 as well and the network would lose its ability to respond. It is only when both C^* and the parameter $A = K_T\phi_K$ are changed in concert that we can increase specificity without losing response, Fig. 2.3 B, bottom panel. This ensures that $K(L)$ becomes always proportional to L without changing the maximum production rate AC^* of C_1. $\bar{\phi}$ finalizes the reduction by putting other parameters to limits that do not significantly change C_1's value. There is no need to perform symmetry breaking for this model to reach optimal behavior and one-parameter reduction.

This simple example illustrates that not only is $\bar{\phi}$ able to rediscover automatically classical reduction of nonlinear equations, but also, as illustrated by step 4 above, it is able to find a nontrivial regime of parameters where the behavior of the network can be significantly improved. Here this is done by reinforcing simultaneously the weight of two branches of the network implicated in a crucial incoherent feedforward loop, implementing perfect adaptation, and allowing to define a simple adaptation submodule. τ dependency is encoded downstream this adaptation module in C_1, defining a kinetic sensing submodule. A general feature of $\bar{\phi}$ is its ability to identify and reinforce crucial functional parts in the networks, as will be further illustrated below.

$\bar{\phi}$ for SHP-1 model

This model aims at modelling early immune recognition by T cells [33] and combines a classical proofreading cascade [26] with a negative feedback loop (Fig. 2.4 A, top). The proofreading cascade amplifies the τ dependency of the output variable, while the variable S in the negative feedback encodes the ligand concentration L in a nontrivial way. The full network presents dose response-curves plateauing at different values for different τs, allowing for approximate discrimination as detailed in [33] (Fig. 2.4 B, step 1). Full understanding of the steady state requires solving an $N \times N$ linear system in combination with a polynomial equation of order $N - 1$, which is analytically possible if N is small enough (see Appendix A). Behavior of the system can only be intuitively grasped in limits of strong negative feedback and infinite ligand concentration [33]. The logic of the network appears superficially similar to the previously described adaptive sorting network, with a competition between proofreading and feedback effects compensating for L, thus allowing for approximated kinetic discrimination based on parameter τ. Other differences include the sensitivity to ligand

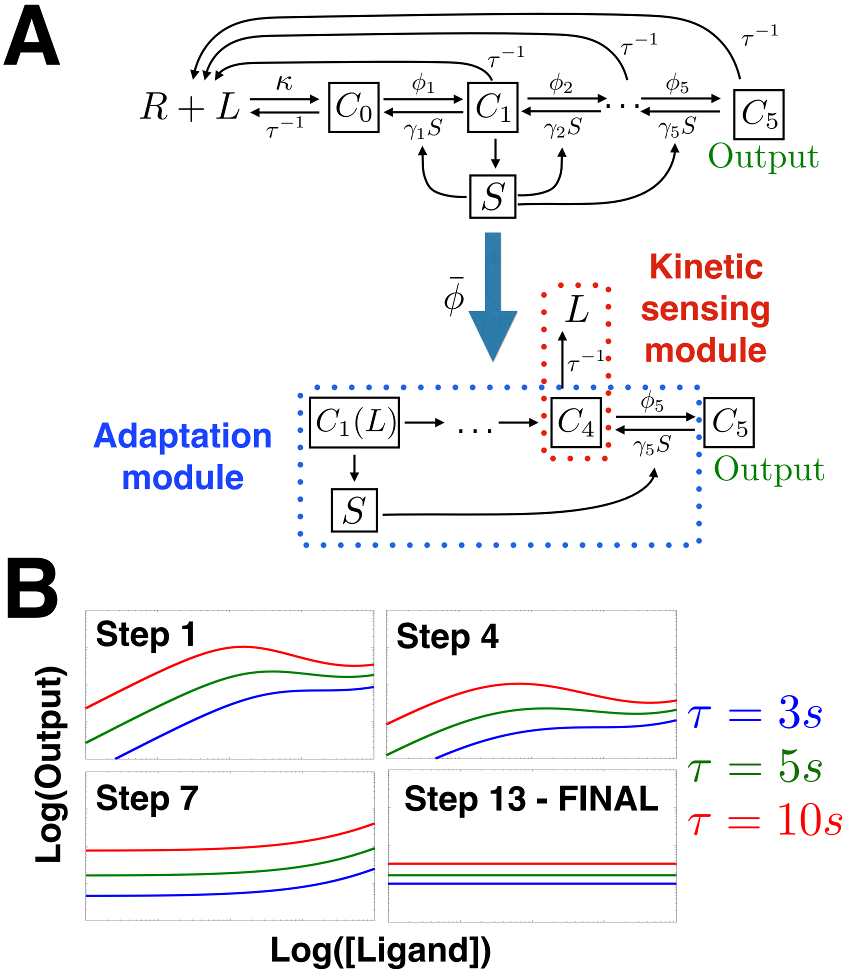


Figure 2.4: **Reduction of SHP-1 model.** (A) Initial model considered and final reduced model (bottom). Step 1 shows the initial dynamics from Eqs. A.6-A.11 in Appendix A. $\bar{\phi}$ (with parameter symmetry breaking) eliminates most of the feedback interactions by S, separating the full network into an adaptation module and a kinetic sensing module. See main text for discussion. (B) Dose response curves for 3 different values of $\tau = 3, 5, 10s$ and different steps of $\bar{\phi}$ reduction, showing how the curves become more and more horizontal for different τ, corresponding to better absolute discrimination. Corresponding parameter modifications are given in Appendix Table A.5. FINAL panel shows behavior of Eqs. A.12-A.18 in the Appendix (full system including local mass conservation).

antagonism because of the different number of proofreading steps, discussed in [49].

When performing $\bar{\phi}$ on this model, the algorithm quickly gets stuck without further reduction in the number of parameters and corresponding network complexity. By inspection of the results, it appears that the network is too symmetrical: variable S acts in exactly the same way on all proofreading steps at the same time. This creates a strong nonlinear feedback term that explains why the nonmonotonic dose-response curves are approximately flat as L varies as described in [33], as well as other features, such as loss of response at high ligand concentration that is sometimes observed experimentally. This also means the output can never be made fully independent of L (see details in 2.2). But it could also be interesting biologically to explore limits where dephosphorylations are more specific, corresponding to breaking symmetry in parameters.

We thus perform symmetry breaking, so that $\bar{\phi}$ converges in less than 15 steps, as shown in one example presented in Fig. 2.4. The dose-response curves as functions of τ become flatter while the algorithm proceeds, until perfect absolute discrimination is reached (flat lines on Fig 2.4 B, step 13).

A summary of the core network extracted by $\bar{\phi}$ is presented in Fig. 2.4 A. In brief, symmetry breaking in parameter space concentrates the functional contribution of S in one single network interaction. This actually *reduces* the strength of the feedback, making it exactly proportional to the concentration of the first complex in the cascade C_1, allowing for a better balance between the negative feedback and the input signal in the network.

Eventually, the dynamics of the last two complexes in the cascade are given by:

$$\dot{C}_4 = \phi_4 C_3 + \gamma_5 S C_5 - (\phi_5 + \tau^{-1}) C_4 \qquad \text{with} \qquad C_3 \propto C_1 \tag{2.4}$$

$$\dot{C}_5 = \phi_5 C_4 - \gamma_5 S C_5 \qquad \text{with} \qquad S \propto C_1 \tag{2.5}$$

Now at steady state, $\phi_5 C_4 = \gamma_5 S C_5$ from Eq. 2.5 so that those terms cancel out in Eq. 2.4 and we get that at steady state $C_4 = \phi_4 \tau C_3$, with C_3 proportional to C_1 via

C_2 in the cascade. Looking back at Eq. 2.5, it means that at steady state both the production and the degradation rates of C_5 are proportional to C_1 (respectively via C_3 for production and S for degradation) . This is another tug-of-war effect, so that at steady state C_5 concentration is independent of C_1 and thus from L. However, there is an extra τ dependency coming from C_4 at steady state (Eq. 2.4), so that C_5 concentration is simply proportional to a power of τ (see full equations in Appendix A).

Again, $\bar{\phi}$ identifies and focuses on different parts of the network to perform perfect absolute discrimination. Symmetry breaking in the parameter spaces allows to decouple identical proofreading steps and effectively makes the behavior of the network more modular, so that only one complex in the cascade is responsible for the τ dependency (“kinetic sensing module" in Fig. 2.4) while another one carries the negative interaction of S (“Adaptation module" in Fig. 2.4) .

When varying initial parameters for reduction, we see different possibilities for the reduction of the network (see examples in Appendix A). While different branches for degradation by S can be reinforced by $\bar{\phi}$, eventually only one of them performs perfect adaptation. Similar variability is observed for τ sensing. Another reduction of this network is presented in the Appendix A.

$\bar{\phi}$ for Lipniacki model

While the $\bar{\phi}$ algorithm works nicely on the previous examples, the models are simple enough so that in retrospect the reduction steps might appear as natural (modulo nontrivial effects such as mass conservation or symmetry breaking). It is thus important to validate the approach on a more complex model which can be understood intuitively but is too complex mathematically to assess without simulations, a situation typical in systems biology. It is also important to apply $\bar{\phi}$ to a published model not designed by ourselves.

We thus consider a much more elaborated model for T cell recognition proposed in [32] and inspired by [31]. This models aims at describing many known interactions of receptors in a realistic way, and accounts for several kinases such as Lck, ZAP70, ERK, and phosphatases such as SHP-1, multiple phosphorylation states of the internal ITAMs. Furthermore, this model accounts for multimerization of receptors with the enzymes. As a consequence, there is an explosion of the number of cross-interactions and variables in the system, as well as associated parameters (since all enzymes modulate variables differently), which renders its intractable without numerical simulations. It is nevertheless remarkable that this model is able to predict a realistic response line (e.g. Fig. 3 in [32]), but its precise quantitative origin is unclear. The model is specified in Appendix A by its twenty-one equations that include a hundred odd terms corresponding to different biochemical interactions. With multiple runs of $\bar{\phi}$ we found two variants of reduction. Figs. 2.5 and 2.6 illustrate examples of those two variants, summarizing the behavior of the network at several reduction steps. Due to the complexity of this network, we first proceed with biochemical reduction. Then we use the reduced network and perform symmetry breaking.

The network topology at the end of both reductions is shown in Figs. 2.5 and 2.6 with examples of the network for various steps. Interestingly, the steps of the algorithm correspond to successive simplifications of clear biological modules that appear in retrospect unnecessary for absolute discrimination (multiple runs yield qualitatively similar steps of reduction). In both cases, we observe that biochemical optimization first prunes out the ERK positive feedback module (which in the full system amplifies response), but keeps many proofreading steps and cross-regulations. The optimization eventually gets stuck because of the symmetry of the system, just like we observed in the SHP-1 model from the previous section (Fig. 2.5 B and Fig. 2.6 A).

Symmetry breaking is then performed, and allows is to considerably reduce the combinatorial aspects of the system, reducing the number of biochemical species and fully eliminating one parallel proofreading cascade (Fig. 2.5 C) or combining two cascades (Fig. 2.6 B). In both variants, the final steps of optimization allow for further reduction of the number of variables keeping only one proofreading cascade in combination with a single loop feedback via the same variable (corresponding to phosphorylated SHP-1 in the complete model).

Further study of this feedback loop reveals that it is responsible for biochemical adaptation, similarly to what we observed in the case of the SHP-1 model. However,

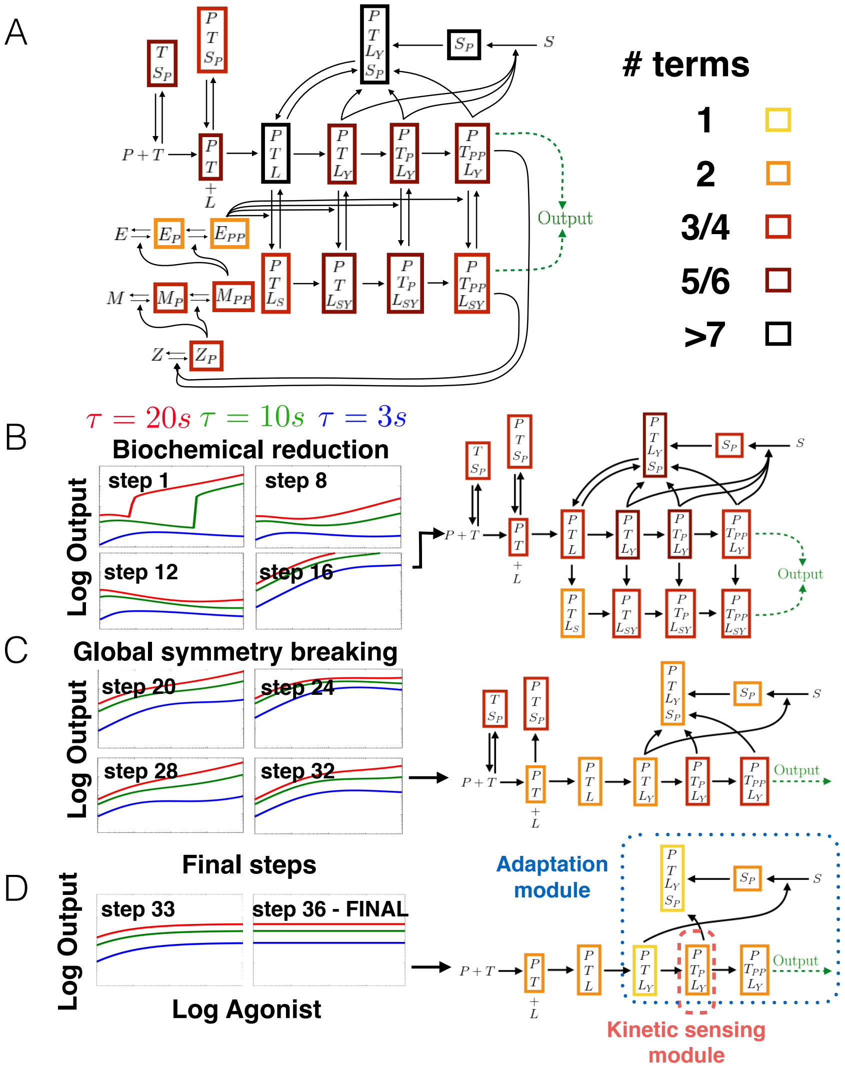


Figure 2.5: **Reduction of the Lipniacki model.** (A) Initial model considered. We indicate complexity with coloured squared boxes that correspond to the number of individual reaction rates in each of the corresponding differential equations for a given variable. (B) to (D) Dose response curves for different reduction steps. Step 1 shows the initial dynamics. From top to bottom, graphs on the right column displays the (reduced) networks at the end of steps 16 (biochemical reduction), 32 (symmetry breaking), 36 (final model). The corresponding parameter reduction steps are given in Appendix A. FINAL panel shows behavior of Eqs. A.31-A.37 in the Appendix (full system including local mass conservation).

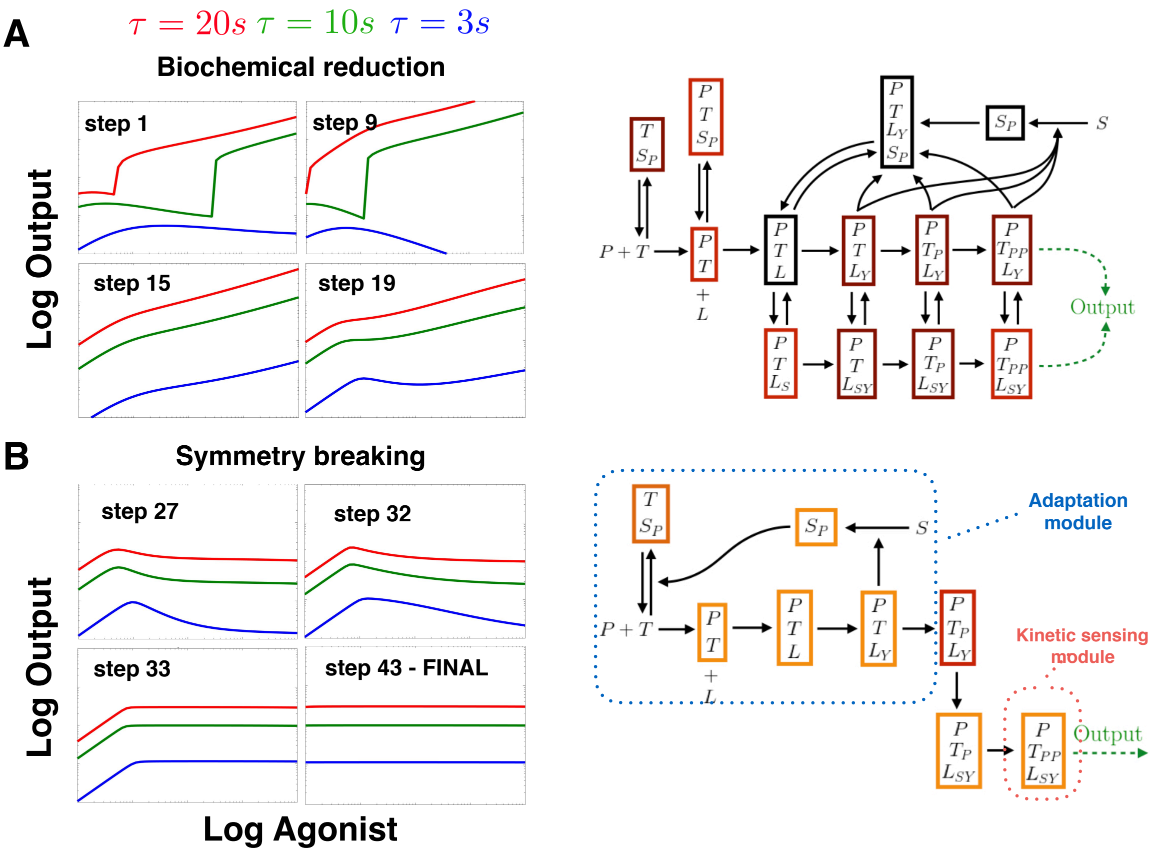


Figure 2.6: **Another reduction of the Lipniacki model.** Starting from the same network as in Fig. 2.5 A but leading to a different adaptive mechanism. The corresponding parameter reduction steps are given in Appendix A. (A) Initial biochemical reduction suppresses the positive feedback loop in a similar way (compare with Fig. 2.5 B). (B) Symmetry breaking breaks proofreading cascades and isolates different adaptive and kinetic modules (compare with Fig. 2.5 D). FINAL panel shows behavior of Eqs. A.38-A.45 in the Appendix (full system including local mass conservation).

the mechanism for adaptation is different for the two different variants and corresponds to two different parameter regimes.

For the variant of Fig. 2.5, the algorithm converges to a local optimum for the fitness. However upon inspection, the structure appears very close to the SHP-1 model reduction, and can be optimized by putting three additional parameters to 0. The Output of the system of Fig. 2.5 is then governed by three variables out of the initial twenty-one and is summarized by:

$$\dot{C_7} = \phi_1 C_5(L) - \phi_2 C_7 - \gamma S C_7 \tag{2.6}$$

$$\dot{S} = \lambda C_5(L) - \mu R_{tot} S \tag{2.7}$$

$$\dot{C_N} = \phi_2 C_7 - \tau^{-1} C_N \tag{2.8}$$

Here $C_5(L)$ is one of the complex concentrations midway of the proofreading cascade (we indicate here L dependency that can be computed by mass conservation but is irrelevant for the understanding of the mechanism). S is the variable accounting for phosphatase SHP-1 in the Lipniacki model, and R_{tot} the total number of unsaturated receptors (the reduced system with the name of the original variables is given in Appendix A).

At steady state S is proportional to $C_5(L)$ from Eq. 2.7. We see from Eq. 2.6 that the production rate of C_7 is also proportional to $C_5(L)$. Its degradation rate $\phi_2 + \gamma S$ is proportional to S if $\phi_2 \ll \gamma S$ (which is the case). So both the production and degradation rates of C_7 are proportional (similar to what happens in the SHP-1 model, Eq. 2.5), and the overall contribution of L cancels out. This corresponds to an adaptation module.

One τ dependency remains downstream of C_7 through Eq. 2.8 (realizing a kinetic sensing module) so that the steady state concentration of C_N is a pure function of τ , thus realizing absolute discrimination. Notably, this model corresponds to a parameter regime where most receptors are free from phosphatase SHP-1, which actually allows for the linear relationship between S and C_5.

For the second variant, when the system has reached optimal fitness the same feedback loop in the model performs perfect adaptation, and the full system of equations in both reductions have similar structure (compare Eqs. A.31-A.37 to Eqs. A.38-A.45 in the Appendix). But the mechanism for adaptation is different: this second reduction corresponds to a regime where receptors are essentially all titrated by SHP-1. More precisely, we have (calling R_f the free receptors, and R_p the receptors titrated by SHP-1):

$$\begin{aligned} \dot{R}_p &= \mu R_f(L)S - \epsilon R_p && (2.9) \\ \dot{S} &= \lambda C_5 - \mu R_f(L)S && (2.10) \\ \dot{C}_5 &= C_3(L) - lSC_5 && (2.11) \end{aligned}$$

Now at steady state, ϵ is small so that almost all receptors are titrated in the form R_p, and thus $R_p \simeq R_{tot}$. This fixes the product $R_f(L)S \propto R_{tot}$ to a value independent of L in Eq. 2.9, so that at steady state of S in Eq. 2.10, $C_5 = \epsilon R_{tot}/\lambda$ is itself fixed at a value independent of L. This implements an "integral feedback" adaptation scheme [59]. Down C_5, there is a simple linear cascade where one τ dependency survives, ensuring kinetic sensing and absolute discrimination for the final complex of the cascade. " *(Parameter reduction [1])*

Recovering antagonism

In their full form,

" the models we reduce all capture the phenomenon of ligand antagonism, where the response of agonist ligands in the presence of high amounts of well chosen subthreshold ligands (i.e. with binding time lower than critical binding time τ_c triggering response) is antagonized. Throughout the reduction, ligand antagonism has remained as a feature, but the hierarchy of antagonism has changed. In the simplest systems, antagonism is maximum for minimum τ, while for more complex models maximum antagonism is reached closer to threshold τ_c [60]. It turns out we can recover this property by adding two terms to the final reduced equations.

An overview of antagonism is presented in Fig. 2.7. We draw the response line as a binary activation by choosing a threshold of the final output for activation (we know from our previous works [33, 34] that adding stochasticity for a probabilistic view does not fundamentally change this picture). The response at lowest ligand concentration always comes from agonist alone (red line Fig. 2.7). Immune cells presented with OVA + G4 are start activating at higher agonist concentration than the OVA + E1 (Fig. S2.7 A). G4 peptides are strong antagonists with a binding time close to threshold than weak antagonists E1. This hierarchy is typical for experimentally observed antagonism: antagonism strength is large just below τ_c, the critical binding time above which a response is elicited.

Similarly, in the full models for SHP-1 and Lipniacki (Fig. S2.7 B - C), we find the same hierarchy. However, for the same binding times in reduced SHP-1 (Fig. 2.7 E) and reduced Lipniacki (Fig. 2.7 F), we find an inverted hierarchy, where ligands further below are more antagonizing, so closer to the naive models discussed in [60].

It turns out that the position of the adaptive module m in the kinetic proofreading cascade of N complexes determines the antagonism strength, like in Fig. 4 of [60]. We retrieve the correct hierarchy of antagonism by adding kinetic terms τ^{-1} to the equations. We illustrate this on the second SHP-1 reduction. The antagonism hierarchy is initially absent from the reduced model (Fig. 2.7 G). When we add τ^{-1} terms to Eqs. A.20 and A.21, it is retrieved (Fig. 2.7 D), because $m = 4$ is large enough. When m is too low ($m = 2$, Figs. 2.7 E - F), antagonism strength peaks for $\tau \ll \tau_c$ and we can not recover the hierarchy observed experimentally. ”

(Parameter reduction [1])

Comparison and categorization of models

“ An interesting feature of $\bar{\phi}$ is that reduction allows to formally classify and connect models of different complexities. We focus here on absolute discrimination only. Our approach allows us to distinguish at least four levels of coarse-graining for absolute discrimination, as illustrated in Fig. 2.8.

At the upper level, we observe that all reduced absolute discrimination models

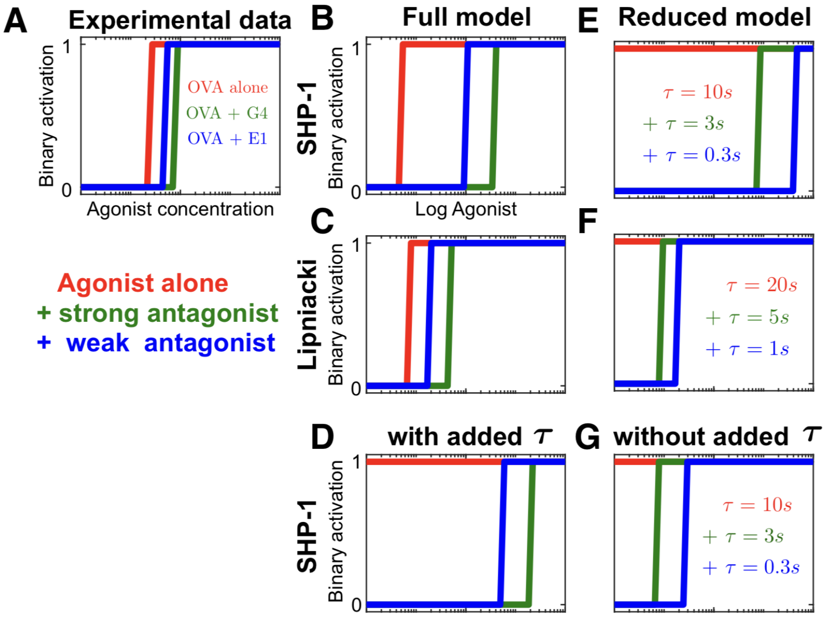


Figure 2.7: **Overview of antagonism.** Red corresponds to agonists alone, green to agonists in the presence of a fixed number of strong antagonist G4 ligands and blue to agonists with weak antagonists E1 ligands. The output is shown as an binary activation depending on threshold crossing. (A) Experimental data, reproduced from [33]. (B)-(C) Full SHP-1/Lipniacki model, showing typical antagonistic hierarchy with binding times as in (E),(F), which show reduced variants of the SHP-1/Lipniacki model via global symmetry breaking. (D),(G) Second variant of the reduced SHP-1 model. Upon adding back terms in τ to Eqs. A.20-A.21, we retrieve the proper hierarchy of antagonism. We added 10^4 antagonist ligands to SHP-1 models, 10^2 antagonist ligands to Lipniacki models and 10 μmol antagonist ligand concentration in the experiments.

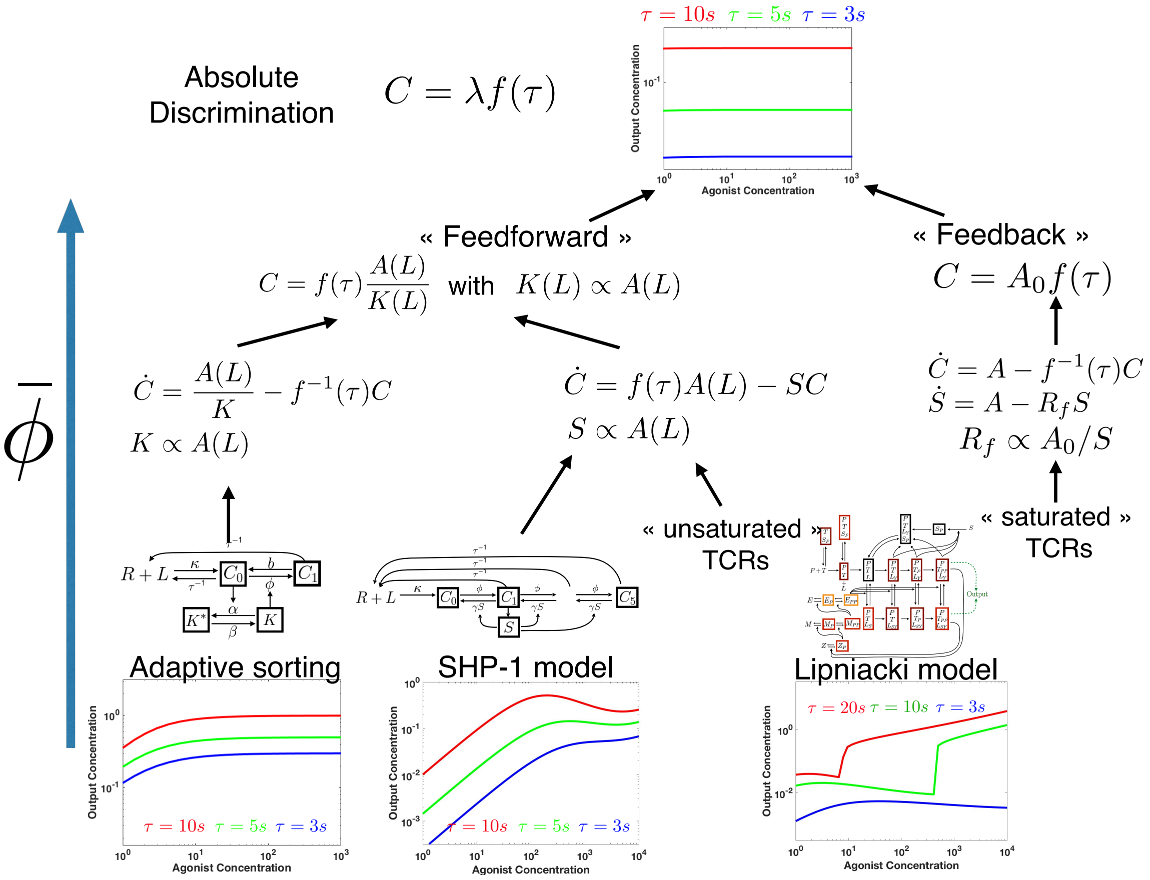


Figure 2.8: **Categorization of networks based on $\bar{\phi}$ reduction.** Absolute discrimination models considered here (bottom of the tree) can all be coarse-grained into the same functional forms (top of the tree). Intermediate levels in reduction correspond to two different mechanisms, “feedforward" based and “feedback” based. See main text for discussion.

considered can be broken down into two parts of similar functional relevance. In all reduced models, we can clearly identify an adaptation module realizing perfect adaptation (defining an effective parameter λ in Fig. 2.8), and a kinetic sensing module performing the sensing of τ (function $f(\tau)$ in Fig. 2.8). If $f(\tau) = \tau$, we get a two-parameter model, where each parameter relates to a submodule.

The models can then be divided in the nature of the adaptatation module, which gives the second level of coarse-graining. With $\bar{\phi}$, we automatically recover a dichotomy previously observed for biochemical adaptation between feedforward and feedback models [61, 62]. The second variant of Lipniacki relies on an integral feedback mechanism, where adaptation of one variable (C_5) is due to the buffering of a negative feedback variable ($S(L)$) (Eqs. 2.9 - 2.11, Fig. 2.8). Adaptive sorting, the SHP-1 model and the first variant of Lipniacki model instead rely on a "feedforward" adaptation module where a tug-of-war between two terms (an activation term $A(L)$ and feedforward terms K / S in Fig. 2.8) exactly compensates.

The tug-of-war necessary for adaptation is realized in two different ways, which is the third level of coarse-graining. In adaptive sorting, this tug-of-war is realized at the level of the *production* rate of the Output, that is made ligand independent by a competition between a direct positive contribution and an indirect negative one (Eq. 2.3, Fig. 2.8). In the reduced SHP-1 model, the *concentration* of the complex C upstream the output is made L independent via a tug-of-war between its production and degradation rates. The exact same effect is observed in the first variant of the Lipniacki model: at steady state, from Eqs. 2.6 and 2.7 the production and degradation rates of C_7 are proportional (Fig. 2.8) which ensures adaptation. So $\bar{\phi}$ allows to rigorously confirm the intuition that the SHP-1 model and the Lipniacki model indeed work in a similar way and belong to the same category in the unsaturated receptor regime. We also notice that $\bar{\phi}$ suggests a new coarse-grained model for absolute discrimination based on modulation of degradation rates, with fewer parameters and simpler behavior than the existing ones, by assuming specific dephosphorylation in the cascades (we notice that some other models have suggested specificity for the last step of the cascade, e.g. in limited signalling models [35]).

Importantly, the variable S, encoding for the same negative feedback in both the

SHP-1 and the first reduction of Lipniacki model, plays a similar role in the reduced models, suggesting that two models of the same process, while designed with different assumptions and biochemical details, nevertheless converge to the same class of models. This variable S also is the buffering variable in the integral feedback branch of the reduction of the Lipniacki model, yet adaptation works in a different way for this reduction. This shows that even though the two reductions of the Lipniacki model work in different parameter regimes and rely on different adaptive mechanisms, the same components in the network play the crucial functional roles, suggesting that the approach is general. As a negative control of both the role of SHP-1 and more generally of the $\bar{\phi}$ algorithm, we show in Appendix A on the SHP-1 model that reduction does not converge in the absence of the S variable (Fig. S3).

Coarse-graining further allows us to draw connections between network components and parameters for those different models. For instance, the outputs are functions of $K(L)C_0(L)$ for adaptive sorting and of $\frac{C(L)}{S(L)}$ for SHP-1/Lipniacki models, where $C_0(L)$ and $C(L)$ are in both models concentrations of complex upstream in the cascade. So we can formally identify $K(L)$ with $S(L)^{-1}$. The immediate interpretation is that deactivating a kinase is similar to activating a phosphatase, which is intuitive but only formalized here by model reduction.

At lower levels in the reduction, complexity is increased, so that many more models are expected to be connected to the same functional absolute discrimination model. For instance, when we run $\bar{\phi}$ several times, the kinetic discrimination module on the SHP-1 model is realized on different complexes (see several other examples in Appendix A). Also, the precise nature and position of kinetic discriminations in the network might influence properties that we have not accounted for in the fitness. We illustrate this on ligand antagonism [60]: depending on the complex regulated by S in the different reduced models, and adding back kinetic discrimination (in the form of τ^{-1} terms) in the remaining cascade on the reduced models, we can observe different antagonistic behaviour, comparable with the experimentally measured antagonism hierarchy. Finally, a more realistic model might account for nonspecific interactions (relieved here by parameter symmetry breaking), which might only give approximate biochemical adaptation (as in [33]) while still keeping the same core principles (adaptation + kinetic discrimination) that are uncovered by $\bar{\phi}$.

2.4 Discussion

When we take into account all possible reactions and proteins in a biological network, a potentially infinite number of different models can be generated. But it is not clear how the level of complexity relates to the behavior of a system, nor how models of different complexities can be grasped or compared. For instance, it is far from obvious whether a network as complex as the one from [32] (Fig. 2.5 A) can be simply understood in any way, or if any clear design principle can be extracted from it. We propose $\bar{\phi}$, a simple procedure to reduce complex networks, which is based on a fitness function that defines network phenotype, and on simple coordinated parameter changes.

$\bar{\phi}$ relies on the optimization of a predefined fitness that is required to encode coarse-grained phenotypes. It performs a direct exploration of the asymptotic limit on boundary manifolds in parameter space. *In silico* evolution of networks teaches us that the choice of fitness is crucial for successful exploration in parameter spaces and to allow for the identification of design principles [54]. Fitness should capture qualitative features of networks that can be improved incrementally; an example used here is mutual information [34]. While adjusting existing parameters or even adding new ones (potentially leading to overfitting) could help optimizing this fitness, it is not obvious *a priori* that systematic *removal* of parameters is possible without decreasing the fitness, even for networks with initial good fitness. For both cases of biochemical adaptation and absolute discrimination, $\bar{\phi}$ is nevertheless efficient at pruning and reinforcing different network interactions in a coordinated way while keeping an optimum fitness, finding simple limits in network space, with submodules that are easy to interpret. Reproducibility in the simplifications of the networks suggests that the method is robust.

In the examples of SHP-1 and Lipniacki models, we notice that $\bar{\phi}$ disentangles the behavior of a complex network into two submodules with well identified functions, one in charge of adaptation and the other of kinetic discrimination. To do so, $\bar{\phi}$ is able to identify and reinforce tug-of-war terms, with direct biological interpretation. This allows for a formal comparison of models. The reduced SHP-1 model and the first reduction of the Lipniacki model have a similar feedforward structure, controlled by a variable corresponding to phosphatase SHP-1 defining the same biological interaction. This is reassuring since both models aim to describe early immune recognition; this was not obvious *a priori* from the complete system of equations or the considered network topology (compare Fig. 2.4 with Fig. 2.5A). These feedforward dynamics discovered by $\bar{\phi}$ contrast with the original feedback interpretation of the role of SHP-1 from the network topology only [31–33]. Adaptive sorting, while performing the same biochemical function, works differently by adapting the production rate of the output, and thus belongs to another category of networks (Fig. 2.8).

$\bar{\phi}$ is also able to identify different parameter regimes for a network performing the same function, thereby uncovering an unexpected network plasticity. The two reductions of the Lipniacki model work in a different way (one is feedforward based, the other one is feedback based), but importantly, the crucial adaptation mechanism relies on the same node, again corresponding to phosphatase SHP-1, suggesting the predictive power of this approach irrespective of the details of the model. From a biological standpoint, since the same network can yield two different adaptive mechanisms depending on the parameter regime (receptors titrated or not by SHP-1), it could be that both situations are observed. In mouse, T Cell Receptors (TCRs) do not bind to phosphatase SHP-1 without engagement of ligands [63], which would be in line with the reduction of the SHP-1 model and the first variant of the Lipniacki model reduction. But we cannot exclude that a titrated regime for receptors exists, e.g. due to phenotypic plasticity [64], or that the very same network works in this regime in another organism. More generally, one may wonder if the parameters found by $\bar{\phi}$ are realistic in any way. In cases studied here, the values of parameters are not as important as the regime in which the networks behave. For instance, we saw for the feedforward models that some specific variables have to be proportional, which requires nonsaturating enzymatic reactions. Conversely, the second reduction of the Lipniacki model requires titration of receptors by SHP-1. These are direct predictions on the dynamics of the networks, not specifically tied to the original models.

Since $\bar{\phi}$ works by sequential modifications of parameters, we get a continuous mapping between all the models at different steps of the reduction process, via the most simplified one-parameter version of the model. By analogy with physics, $\bar{\phi}$ thus “renormalizes" different networks by coarse-graining [50], possibly identifying universal classes for a given biochemical computation, and defining subclasses [65]. This allows us to draw correspondences between networks with very different topologies, formalizing ideas such as the equivalence between activation of a phosphatase and repression of a kinase (as exemplified here by the comparison of influences of $K(L)$ and $S(L)$ in reduced models from Fig. 2.8). In systems biology, models are neither traditionally simplified, nor are there systematic comparisons between models, in part because there is no obvious strategy to do so. The approach proposed here offers a solution for both comparison and reduction, which complements other strategies such as the evolution of phenotypic models [54] or direct geometric modelling in phase space [66].

To fully reduce complex biochemical models, we have to perform symmetry breaking on parameters. Similar to parameter modifications, the main roles of symmetry breaking is to reinforce and adjust dynamical regimes in different branches of the network, e.g. imposing proportionality to tug-of-war terms. Intuitively, symmetry breaking embeds complex networks into a higher dimensional parameter space allowing for better optimization. Much simpler networks can be obtained with this procedure, which shows in retrospect how the assumed nonspecificity in interactions strongly constrains the allowed behavior. Of course, in biology, some of this complexity might also have evolutionary adaptive values, corresponding to other phenotypic features we have neglected here, such as signal amplification. A tool like $\bar{\phi}$ allows for a reductionist study of these features by specifically focusing on one phenotype of interest to extract its core working principles. Once the core principles are identified, it should be easier to complexify a model by accounting for other potential adaptive phenotypes (e.g. as is done to reduce antagonism in [34]).

Finally, there is a natural evolutionary interpretation of $\bar{\phi}$. In both evolutionary computations and evolution, random parameter modifications in evolution can push single parameters to 0 or potentially very big values (corresponding to the ∞ limit). However, it is clear from our simulations that concerted modifications of parameters are needed, e.g. for adaptive sorting, the simultaneous modifications of the kinetics and the efficiency of a kinase regulation is required in Step 4 of the reduction. Evolution might select for networks explicitly coupling parameters that need to be modified in concert. Conversely, there might be other constraints preventing efficient optimizations in two directions in parameter space at the same time, due to epistatic effects. Gene duplications provide an evolutionary solution to relieve such trade-offs, after which previously identical genes can diverge and specialize [67]. This clearly bears resemblance to the symmetry breaking proposed here. For instance, having two duplicated kinases instead of one would allow to have different phosphorylation rates in the same proofreading cascades. We also see in the examples of Figs. 2.4, 2.5, and 2.6 that complex networks that cannot be simplified by pure parameter changes, can be improved by parameter symmetry breaking via decomposition into independent submodules. Similar evolutionary forces might be at play to explain the observed modularity of gene networks [68]. More practically, $\bar{\phi}$ could be useful as a complementary tool for artificial or simulated evolution [54] to simplify complex simulated dynamics [69]. ” *(Parameter reduction [1])*

3 Attack and defence

In Chapter 2 we have shown how models of immune recognition varying in complexity contain the same core functionality, but differ in additional features they may describe. Next, we focus our attention on ligand antagonism, a feature in immune recognition models that we touched on before. Ligand antagonism is the inhibition of a response variable by ligands with a subthreshold binding time that by themselves do not trigger a response. It has been proven that ligand antagonism is unavoidable for certain classes of immune recognition models [49], although the strength of antagonism by ligands with a given subthreshold binding time $\tau \leq \tau_c$ depends on biochemical parameters which determine the immune recognition model.

A similar fooling mechanism exists in machine learning. Machine learning algorithms routinely misclassify regular samples when a specific, imperceptible perturbation called an adversarial perturbation is added to them. Such adversarial examples are known to exist across a wide variety of models and are transferable across architectures [70, 71]. There even exists universal adversarial perturbation that fool all algorithms [72]. It has been shown that adversarial examples are inevitable [73], similar to ligand antagonism in immune recognition models. Hypotheses for the existence of adversarial examples range from the nonlinearity of machine learning algorithms [70] to the high-dimensionality of the data [74], but the field does not unequivocally agreed on this.

Attack and defence

In this Chapter, we draw a connection between ligand antagonism and adversarial examples. We use techniques for generating adversarial perturbations for machine learning to compute the most efficient antagonizers in immune recognition, and show that additional nonlinearity aids in mitigating the effect of adversarial perturbations for both a simple machine learning model and the immune recognition model. Finally, by inspecting samples at the decision boundary, we show that through tuning the biochemical parameters determining the immune recognition model we can expect most robustness. Our conclusions highlight design constraints machine learning algorithms robust to adversarial examples must adhere to, as well as to what extend adversarial examples may be limited.

> “ Machine learning algorithms can be fooled by small well-designed adversarial perturbations. This is reminiscent of cellular decision-making where ligands (called antagonists) prevent correct signalling, like in early immune recognition. We draw a formal analogy between neural networks used in machine learning and models of cellular decision-making (adaptive proofreading). We apply attacks from machine learning to simple decision-making models, and show explicitly the correspondence to antagonism by weakly bound ligands. Such antagonism is absent in more nonlinear models, which inspired us to implement a biomimetic defence in neural networks filtering out adversarial perturbations. We then apply a gradient-descent approach from machine learning to different cellular decision-making models, and we reveal the existence of two regimes characterized by the presence or absence of a critical point in the gradient. This critical point causes the strongest antagonists to lie close to the decision boundary. This is validated in the loss landscapes of robust neural networks and cellular decision-making models, and observed experimentally for immune cells. For both regimes, we explain how associated defence mechanisms shape the geometry of the loss landscape, and why different adversarial attacks are effective in different regimes. Our work connects evolved cellular decision-making to machine learning, and motivates the design of a general theory of adversarial perturbations, both for *in vivo* and *in silico* systems. ” *(Attack and defence [2])*

3.1 Introduction

“ Machine learning is becoming increasingly popular with major advances coming from deep neural networks [9]. Deep learning has improved the state-of-the-art in automated tasks like image processing [75], speech recognition [76] and machine translation [77], and has already seen a wide range of applications in research and industry. Despite their success, neural networks suffer from blind spots: small perturbations added to unambiguous samples may lead to misclassification [70]. Such adversarial examples are most obvious in image recognition, for example, a panda is misclassified as a gibbon or a handwritten 3 as a 7 [74]. Real world scenarios exist, like adversarial road signs fooling computer vision algorithms (Fig. 3.1 A) [78], or adversarial perturbations on medical images triggering incorrect diagnosis [79]. Worse, adversarial examples are often transferable across algorithms (see [80] for a recent review), and certain universal perturbations fool any algorithm. [72].

Categorization and inference are also tasks found in cellular decision-making [81]. For instance, T cells have to discriminate between foreign and self ligands which is challenging since foreign ligands might not be very different biochemically from self ligands [25, 82]. Decision-making in an immune context is equally prone to detrimental perturbations in a phenomenon called ligand antagonism [49]. Antagonism appears to be a general feature of cellular decision-makers: it has been observed in T cells [31], mast cells [57] and other recognition processes like olfactory sensing [83, 84].

There is a natural analogy to draw between decision-making in machine learning and in biology. In machine learning terms, cellular decision-making is similar to a classifier. Furthermore, in both artificial and cellular decision-making, targeted perturbations lead to faulty decisions even in the presence of a clear ground truth signal. As a consequence, arms races are observed in both systems. Mutating agents might systematically explore ways to fool the immune cells via antagonism, as has been proposed in the HIV case [29, 85, 86]. Recent examples might include neoantigens in cancer [87, 88] which are implicated in tumour immunoediting and escape from the immune system. Those medical examples are reminiscent of how adversaries

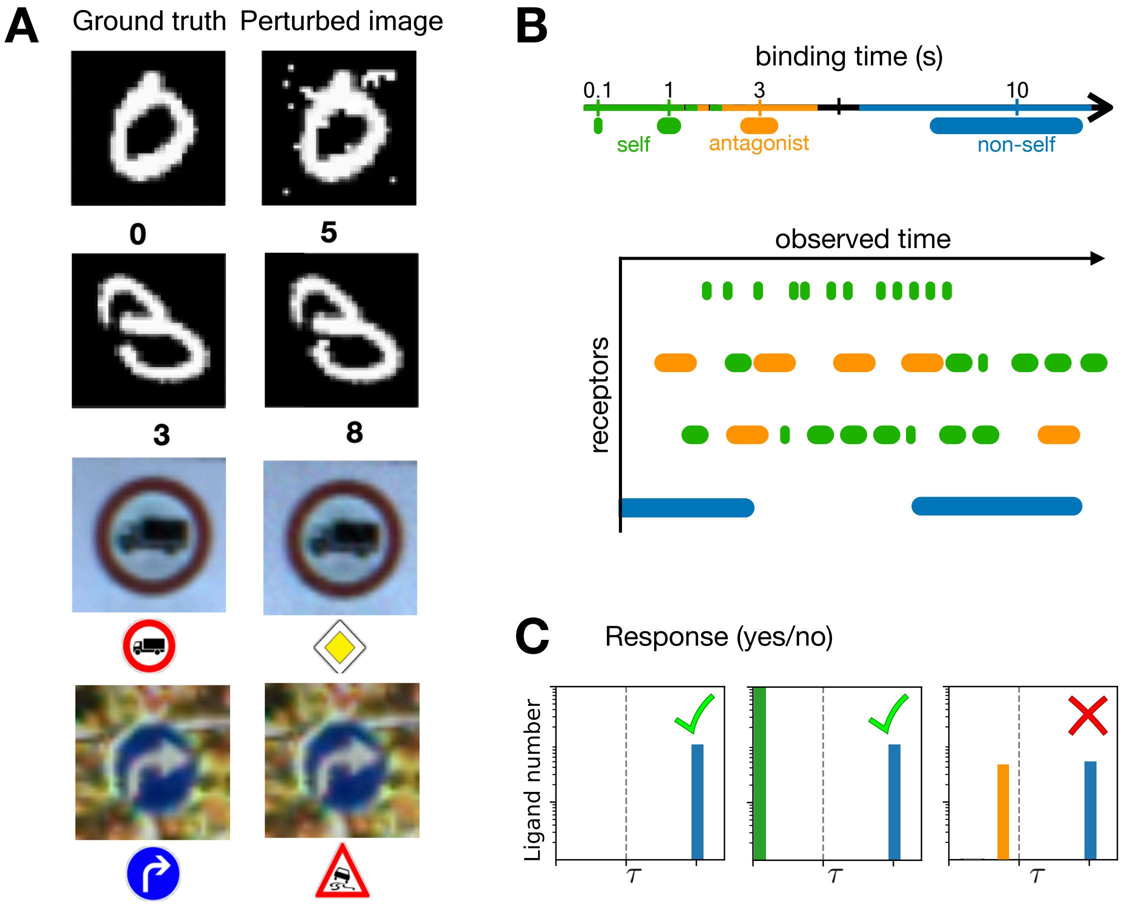


Figure 3.1: **Ligand discrimination and digit recognition tasks** (A) Adversarial examples on digits and roadsigns. Reproduced from [78]. Left column displays original images with categories recognized by machine learning algorithms, right column displays images containing targeted perturbations leading to misclassification. (B) Schematics of ligand binding events showing typical receptor occupancy through some observed time during cellular decision-making using T cell terminology ("self vs non-self"). The colored bars corresponds to self (green), antagonist (orange) and non-self (blue) ligands binding to receptors. Their lengths are indicative of the binding time τ_i, whereas their rate of binding measures the on-rate k_i^{on}. (C) Different ligand distributions give different response. The vertical dotted line indicates quality τ_d. Decision should be to activate if one observes ligands with $\tau > \tau_d$, so on the right of the dotted line. In an immune context, T cells respond to ligand distributions of agonists alone and agonists in the presence of non-agonists (with very small binding times τ), while the T cell fails to respond if there are too many ligands just below threshold τ_d.

could generate black box attacks aimed at fooling neural networks [78]. Strategies for provable defenses or robust detection of adversarial examples [89, 90] are currently developed in machine learning, but we are still far from a general solution.

In the following, we draw a formal correspondence between biophysical models of cellular decision-making displaying antagonism on the one hand, and adversarial examples in machine learning on the other hand. We show how simple attacks in machine learning mathematically correspond to antagonism by many weakly bound ligands in cellular decision-making. Inspired by kinetic proofreading in cellular decision-making, we implement a biomimetic defence for digit classifiers, and we demonstrate how these robust classifiers exhibit similar behavior to the nonlinear adaptive proofreading models. Finally, we explore the geometry of the decision boundary for adaptive proofreading, and observe how a critical point in the gradient dynamics emerges in networks robust to adversarial perturbations. Recent findings in machine learning [91] confirm the existence of two regimes, which are separated by a large nonlinearity in the activation function. This inspired us to define two categories of attack (high-dimensional, small amplitude and low-dimensional, large amplitude) both for models of cellular decision-making and neural networks. Our work suggests the existence of a unified theory of adversarial perturbations for both evolved and artificial decision-makers.

Adaptive proofreading for cellular decision-making

Cellular decision-making in our context refers to classification of biological ligands in two categories, e.g. “self vs non-self" in immunology, or “agonist vs non-agonist" in physiology [25, 92, 93]. For most of those cases, qualitative distinctions rely on differences in a continuously varying property (typically a biochemical parameter). Thus it is convenient to rank different ligands based on a parameter (notation τ) that we will call quality. Mathematically, a cell needs to decide if it is exposed to ligands with quality $\tau > \tau_d$, where τ_d is the quality at the decision threshold. Such ligands triggering response are called agonists. A general problem then is to consider cellular decision-making based on ligand quality irrespective of ligand quantity (notation L). An example can be found in immune recognition with the lifetime dogma [25], where it is assumed that a T cell discriminates ligands based on their characteristic binding time τ to T cell receptors (this is of course an approximation and other parameters might also play a role in defining quality, see [36, 43, 94]). Ligand discrimination is a nontrivial problem for the cell, which does not measure single-binding events but only has access to global quantities such as the total number of bound receptors (Fig. 3.1 B). The challenge is to ignore many subthreshold ligands ($\tau < \tau_d$) while responding to few agonist ligands with $\tau > \tau_d$ [25, 31, 33]. In particular, it is known experimentally in many different contexts that addition of antagonistic subthreshold ligands can impair proper decision-making (Fig. 3.1 C) [31, 57, 83].

To model cellular decision-making, we will use the general class of "adaptive sorting" or "adaptive proofreading" models, which account for many aspects of immune recognition [34, 49], and can be shown to capture all relevant features of such cellular decision-making close to a decision threshold [60]. An example of such a model is displayed in Fig. 3.2 A. Importantly, we have shown previously that many other biochemical models present similar properties for the steady-state response as a function of the input ligand distribution [1]. In the following we summarize the most important mathematical properties of such models. An analysis of the detailed biochemical kinetics of the model of Fig. 3.2 A is presented in Appendix S1.

We assume an idealized situation where a given receptor i, upon ligand binding (on-rate k_i^{on}, binding time τ_i) can exist in N biochemical states (corresponding to phosphorylation stages of the receptor tails in the immune context [26, 95]). Those states allow the receptor to effectively compute different quantities, such as $c_n^i = k_i^{\text{on}} \tau_i^n, 0 \leq n \leq N$, which can be done with kinetic proofreading [26–28]. In particular, ligands with larger τ give a relatively larger value of c_N^i due to the geometric amplification associated with proofreading steps. We assume receptors to be identical, so that any downstream receptor processing by the cell must be done on the sum(s) $C_n = \sum_i c_n^i = \sum_i k_i^{\text{on}} \tau_i^n$. We also consider a quenched situation in which only one ligand is locally available for binding to every receptor. In reality, there is a constant motion of ligands, such that k_i^{on} and τ_i are functions of time and stochastic treatments are required [81, 96, 97], but on the time-scale of primary decision-making it is reasonable to assume that the ligand distribution does not change much [31].

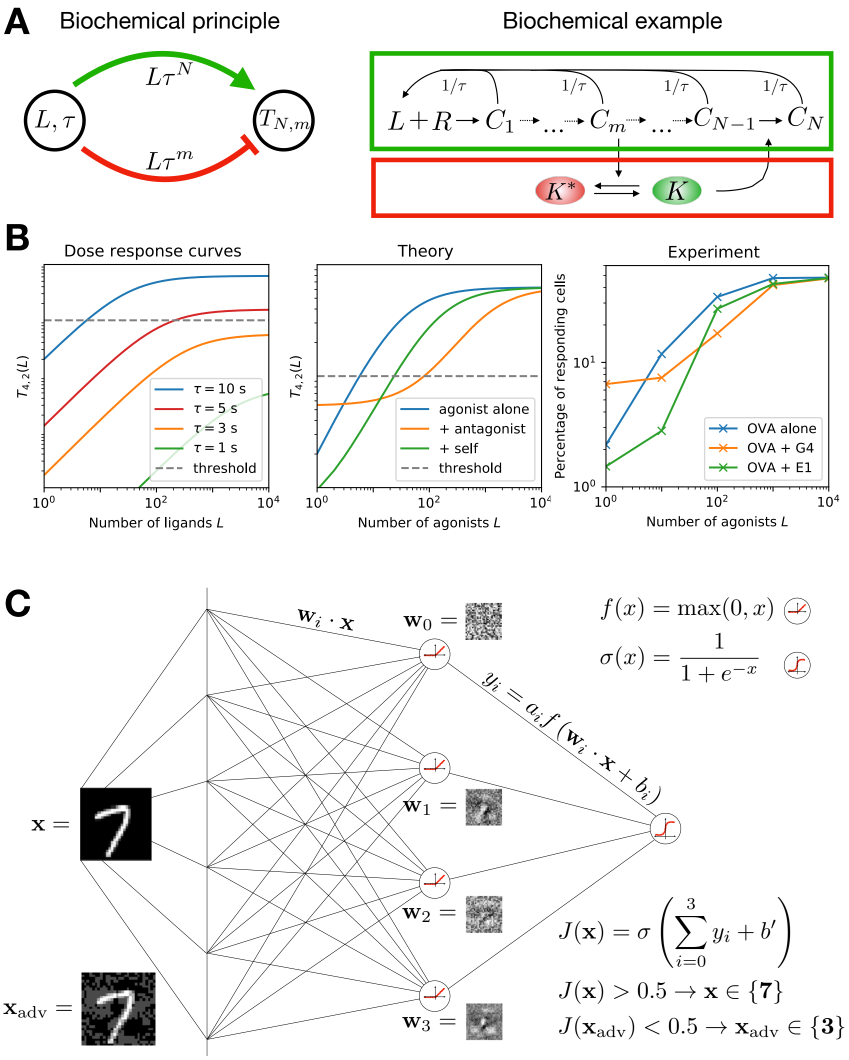


Figure 3.2: **Adaptive proofreading and neural network** (A) Left: Adaptive proofreading networks have an activating and repressing branch with different weights on τ. Right: detailed adaptive proofreading network adapted from [60]. Ligand L binds to receptor R to form unphosphorylated complex C_0. The receptor chain is iteratively phosphorylated until reaching state C_N along the activating branch (green). At every stage C_i, the ligand can unbind from the receptor with ligand-specific rate τ^{-1}. At C_m, the repressing branch (red) splits by inhibiting the kinase K, which mediates the feedforward mechanism. (B) Dose-response curves for pure ligand types and mixtures, in both adaptive proofreading models and experiments on T cells (redrawn from [33]). Details on models and parameters used are given in Appendix S2. For experiments, OVA are agonist ligands, G4 and E1 are ligands known to be below threshold, but showing clear antagonistic properties. (C) Schematic of the neural network used for digit recognition. We explicitly show the 4 weight vectors W_i learned in one instance of the training, the activation function J and an adversarially perturbed sample $\mathbf{x}_{\mathrm{adv}}$.)

3.1 Introduction

Adaptive proofreading models rely on an incoherent feedforward loop, where an output is at the same time activated and repressed by bound ligands via two different branches in a biochemical network (Fig. 3.2 A). An explicit biochemical example is shown on the right panel of Fig. 3.2 A. Here, activation occurs through a kinetic proofreading cascade (green arrow/box), and repression through the inactivation of a kinase by the same cascade (red arrow/box). The branches engage in a tug-of-war, which we describe below.

For simplicity, let us first assume that only one type of ligands with binding time τ and on rate k_{on} are presented. We call L the quantity of ligands. Then, in absence of saturation, the total number of n-th complex C_n of the proofreading cascade along the activation branch will be proportional to $k_{\text{on}} L \tau^n$. This is the activation part of the network where the response is activated.

We now assume that the m-th complex of the cascades are inactivating a kinase K specific to C_m, so that $K \propto (k_{\text{on}} L \tau^m)^{-1}$ for L big enough. This is the repression part of the network. K is assumed to diffuse freely and rapidly between receptors so that it effectively integrates information all over the cell (recent work quantified how this crosstalk can indeed improve detection [98]). m is an important parameter that we will vary to compare different models. K then catalyzes the phosphorylation of the final complex of the cascade so that we have for the total number C_N

$$\dot{C}_N = K C_{N-1} - \tau^{-1} C_N. \tag{3.1}$$

and at steady state

$$C_N \propto \frac{k^{\text{on}} L \tau^N}{k^{\text{on}} L \tau^m} = \tau^{N-m} \tag{3.2}$$

The L dependence cancels, and C_N is a function of τ alone. From this, it is clear that ligand classification can be done purely based on C_N, the total number of complexes, which is a measure of ligand quality. In this situation, it is easy to define a threshold τ_d^{N-m} that governs cell activation ($C_N > \tau_d^{N-m}$) or quiescence ($C_N < \tau_d^{N-m}$). Biochemically, this can be done via the digital activation of another kinase shared between all receptors [31, 34].

This model can be easily generalized to a mixture of ligands with different qualities. To do so, in the previous derivations all quantities accounting for the total complex C_n of the form $k_{\text{on}}L\tau^n$ can be replaced by $\sum_i k_i^{\text{on}} L_i \tau_i^N$, calling L_i the quantity of ligands with identical k_i^{on}, τ_i. We then define the generalized output of the biochemical network as

$$T_{N,m} = \frac{\sum_i k_i^{\text{on}} L_i \tau_i^N}{\sum_i k_i^{\text{on}} L_i \tau_i^m}. \tag{3.3}$$

Similar equations for an output $T_{N,m}$ can be derived for many types of networks, as described in [1]. For this reason we will focus in the following on the properties of $T_{N,m}$, forgetting about the internal biochemistry giving rise to this behaviour. Notice here that by construction $N > m > 1$, but other cases are posssible with different biochemistry, for instance examples in olfaction correspond to the case $N = 1, m = 0$ [83] (see also another example in [60]) . Also notice that if kinetic parameters of the ligands are not identical, the dependence on L_i does not cancel out, which will be the origin of most of the key phenomena described below.

Fig. 3.2 B shows theoretical and experimental curves of a realistic adaptive proofreading model (including minimum concentration for repression of kinase K, etc. see Appendix S2 for full model and parameter values). We have chosen $(N, m) = (4, 2)$ so that the qualitative features of the theoretical curves match the experimental curves best. Adaptive proofreading models give dose response curves plateauing at different values as a function of parameter τ, allowing to perform sensitive and specific measurement of this parameter. For small τ (e.g. $\tau = 3\,\text{s}$), one never reaches the detection threshold (dotted line on Fig. 3.2 B, left panel) even for many ligands. For slightly bigger $\tau = 10\,\text{s} > \tau_{\text{d}}$, the curve is shifted up so that detection is made even for a small concentration of agonists.

Nontrivial effects appear if we consider mixtures of ligands with different qualities. Then the respective computation made by the activation and repression branch of the network depend in different ways on the distribution of the presented ligand binding times. For instance, if we now add L_{a} antagonists with lower binding time $\tau_{\text{a}} < \tau$ and equal on-rate k^{on}, we have $T_{N,m} = \frac{L\tau^N + L_{\text{a}}\tau_{\text{a}}^N}{L\tau^m + L_{\text{a}}\tau_{\text{a}}^m}$, which is smaller than the response τ^{N-m} for a single type of ligands, corresponding to ligand antagonism (Fig. 3.2 B, middle panel) [31, 49, 63, 99]. In the presence of many ligands below the threshold of detection, the dose response curve are simultaneously moved to the right but with a higher starting point (compared to the reference curve for "agonist alone"), as observed experimentally (Fig. 3.2 B, right panel, data redrawn from [33]). Different models have different antagonistic properties, based on the strength of the activation branch (N) relative to the repression branch (m). More mathematical details on these models can be found in [34, 49, 60].

Neural networks for artificial decision-making

We will compare cellular decision-making to decision-making in machine learning algorithms. We will constrain our analysis to binary decision-making (which is of practical relevance, for instance in medical applications [79]), using as a case-study image classification from two types of digits. These images are taken from MNIST [100], a standard database with 70000 pictures of handwritten digits. Even for such a simple task, designing a good classifier is not trivial, since it should be able to classify irrespective of subtle changes in shapes, intensity and writing style (i.e. with or without a central bar for a 7).

A simple machine learning algorithm is logistic regression. Here, the inner product of the input and a learned weight vector determines the class of the input. Another class of machine learning algorithms are feedforward neural networks: interconnected groups of nodes processing information layer-wise. We chose to work with neural networks for several reasons. First, logistic regression is a limiting case of a neural network without hidden layers. Second, a neural network with one hidden layer more closely imitates information processing in cellular networks, i.e. in the summation over multiple phosphorylation states of the receptor-ligand complex (nodes) in a biochemical network. Third, such an architecture reproduces classical results on adversarial perturbations such as the ones described in [74]. Fig. 3.2 C introduces the iterative matrix multiplication inside a neural network. Each neuron i computes $\mathbf{w}_i \cdot \mathbf{x}$, $i \in [0, 3]$, adds bias b_i, and transforms the result with an activation function $f(x)$. We chose to use a Rectified Linear Unit (ReLU), which returns 0 when its input is negative, and the input itself otherwise. The resulting $f(\mathbf{w}_i \cdot \mathbf{x} + b_i)$ is multiplied by another weight vector with elements a_i, summed up with a bias, defining a scalar quantity $x = \sum_i a_i f(\mathbf{w}_i \cdot \mathbf{x} + b_i) + b'$. Finally, we obtain the score $J(\mathbf{x})$ (a probability between 0 and 1 for the input to belong to a class) by transforming x with the logistic function $\sigma(x)$. Parameters of such networks are optimized using classical stochastic gradient descent within a scikit implementation [101], see Appendix S2. As an example, in Fig. 3.2 C, a 7 is correctly classified by the neural network ($J(\mathbf{x}) > 0.5$), while the adversarial 7 is classified as a three ($J(\mathbf{x}_{\text{adv}}) < 0.5$).

3.2 Results

We first summarize the general approach followed to draw the parallel between machine learning and cellular decision-making. We will limit ourselves to simple classifications where a single decision is made, such as “agonist present vs no agonist present” in biology, or “3 vs 7” in digit recognition. As input samples, we will consider pictures in machine learning, and ligand distributions in biology. We define a ligand distribution as the set of concentrations with which the ligands with unique binding times are present. This corresponds to a picture that is presented as a histogram of pixel values; the spatial correlation between pixels is lost, but their magnitude remains preserved. Decision-making on a sample is then done via a scoring function (or score). This score is computed either directly by the machine learning algorithm (score J) or by the biochemical network, via the concentration of a given species (score $T_{N,m}$). For simple classifications, the decision is then based on the relative value of the score above or below some threshold (typically 0.5 for neural networks where decision is based on sigmoidal functions, or some fixed value related to the decision time τ_d for biochemical networks).

The overall performance of a given classifier depends on the behavior of the score in the space of possible samples (i.e. the space of all possible pictures, or the space of all possible ligand distributions). Both spaces have high dimensions: for instance the dimension in the MNIST picture correspond to number of pixels $28\times28 = 784$, while in immunology ligands can bind to roughly 30000 receptors [31]. The score can thus be thought of as a nonlinear projection of this high-dimensional space in one dimension. We will study how the score behaves in relevant directions in the sample space, and how to change the corresponding geometry and position of decision boundaries (defined as the samples where the score is equal to the classification threshold). We will show that similar properties are observed, both close to typical samples and to the decision boundary. It is important to notice at this stage that the above considerations are completely generic on the biology side and are not necessary limited to, say immune recognition. However, we will show that adaptive proofreading presents many features reminiscent of what is observed in machine learning.

Fast Gradient Sign Method recovers antagonism by weakly binding ligands

In this framework, from a given sample, an adversarial perturbation is a small perturbation in sample space giving a change in score reaching (or crossing) the decision boundary. We start by mathematically connecting the simplest class of adversarial examples in machine learning to antagonism in adaptive proofreading models. We follow the original Fast Gradient Sign Method (FGSM) proposed by [74]. The FGSM computes the local maximum adversarial perturbation $\eta = \epsilon \,\mathsf{sgn}\,(\nabla_x J)$ (where sign is taken elementwise). $\nabla_x J$ represents the gradient of the scoring function, categorizing images in two different categories (such as 3 and 7 in [74]). Its elementwise sign defines an image, that is added to the initial batch of images with small weight ϵ. Examples of such perturbations are shown in Fig. 3.2 C (bottom left) and Fig. S2 A for the 3 vs 7 digit classification problem. While to the human observer, the perturbation is weak and only changes the background, naive machine learning algorithms are completely fooled by the perturbation and systematically misclassify the digit.

Coming back to adaptive proofreading models, we apply FGSM for the computation of a maximally antagonistic perturbation. To do so, we need to specify the equivalent of pixels in adaptive proofreading models. A natural choice is to consider parameters associated to each pair (index i) of receptor/ligands, namely k_i^{on} (corresponding to the rate at which ligands bind to receptors, also called on-rate [1]) and τ_i

[1] The on-rate is easily confused with the unbinding rate, whose inverse we call the binding time, which indicates the lifetime of the ligand-receptor complex

(corresponding to quality). If a receptor i is unoccupied, we set its k_i and τ_i to 0 [2]. We then compute gradients with respect to these parameters.

As a simple example, we start with the case $(N, m) = (1, 0)$, which also corresponds to a recently proposed model for antagonism in olfaction [83], with the role of k^{on} played by inverse affinity κ^{-1}, the role of τ played by efficiency η, and the spiking rate of the olfactory receptor neurons is $J(T_{N,m})$, that can be interpreted as a scoring function in the machine learning sense. In this case, $T_{1,0}$ simply computes the average quality τ_{avg} of ligands presented weighted by k_i^{on} (models with $N > m > 0$ give less intuitive results as will be shown in the following). It should be noted that while this computation is formally simple, biochemically it requires elaborated internal interactions, because a cell can not easily disentangle influence of individual receptors, see [49, 83] for explicit examples.

Starting from the computation of $\nabla_x J$ with respect to parameters k_i^{on} and τ_i, the FGSM perturbation is:

$$\eta = \epsilon \,\text{sgn} \begin{pmatrix} \partial_{\tau_i} J \\ \partial_{k_i^{\text{on}}} J \end{pmatrix} = \epsilon \,\text{sgn}(A)\text{sgn} \begin{pmatrix} k_i^{\text{on}} \\ \tau_i - T_{1,0} \end{pmatrix}, \tag{3.4}$$

where $A = \frac{J'(T_{1,0})}{\sum k_i^{\text{on}}} > 0$. Notice in the above expression that since derivatives act on different parameters, an ϵ sized-perturbation of a given parameter is expressed in its corresponding unit. For simplicity we will not explicitly write the conversion factor between units (this is for mathematical convenience and does not impact our results). From the above expression, we find that an equivalent maximum adversarial perturbation is given by three simple rules (Fig. 3.3 A).

- Decrease all τ_i by ϵ
- Decrease k_i^{on} by ϵ for ligands with $\tau_i > T_{1,0}$
- Increase k_i^{on} by ϵ for ligands with $\tau_i < T_{1,0}$

The key relation to adversarial examples from [74] comes from considering what happens to the unbound receptors for which both k_i^{on} and τ_i are initially 0. Let us

[2] an alternative choice without loss of generality is to consider a situation where for unoccupied receptors, k_i is 0 but τ_i is arbitrary, corresponding to a ligand available for binding

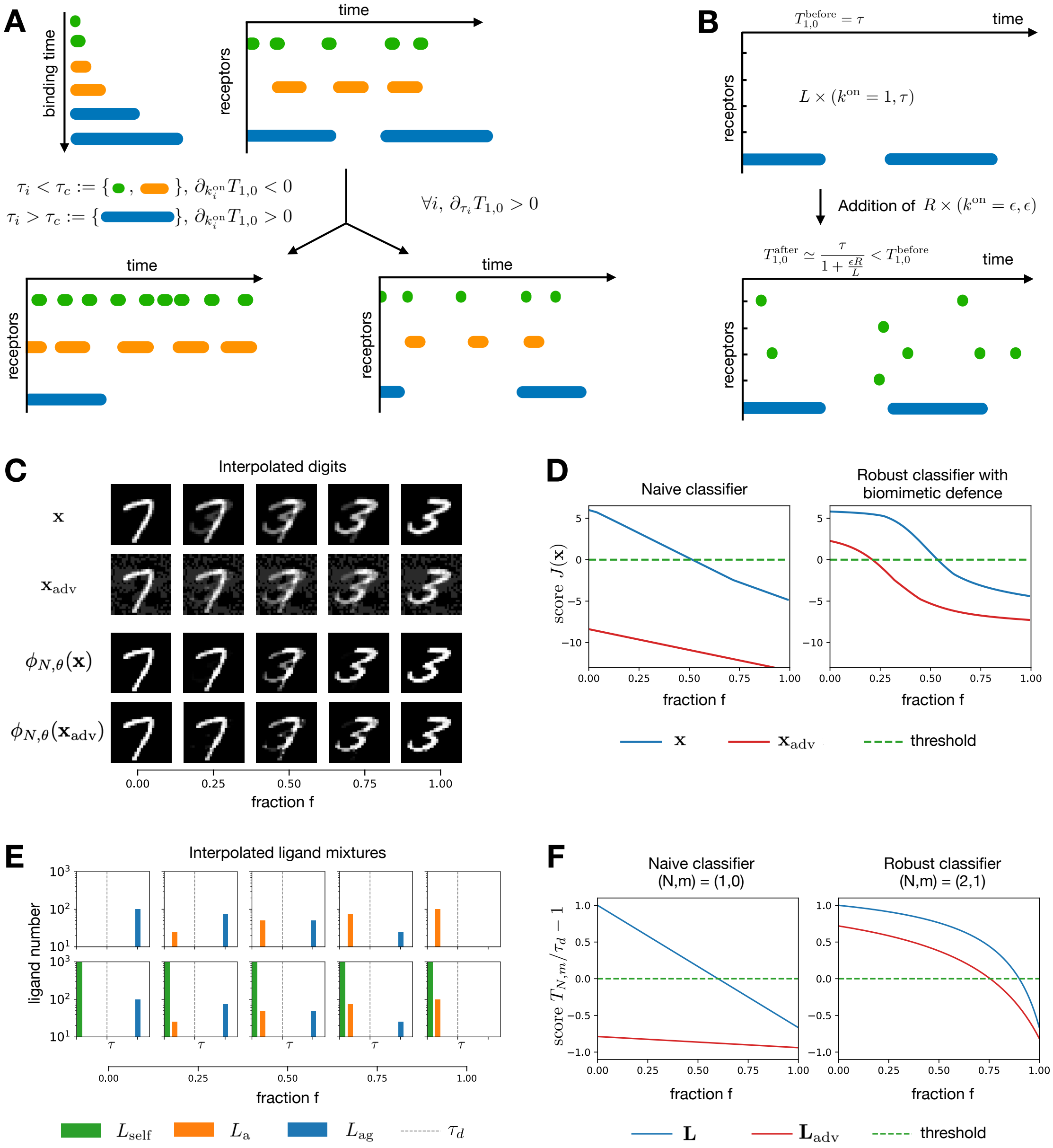


Figure 3.3: **Schematics of FGSM applied to immune recognition.** (Caption on the following page.)

Figure 3.3: (A) We compute how to lower the response for the receptor occupancy through a given period of time by changing k_i^{on} and τ_i. Bottom left: increasing k_i^{on} for ligands with $\tau_i < \tau_d$ and decreasing k_i^{on} for ligands with $\tau_i > \tau_d$ reduces the weighted average $T_{1,0}$ (change in frequency of the colored bars). Bottom right: decreasing τ_i for all ligands decreases $T_{1,0}$ (change in length of the colored bars). (B) Response to non-self ligands is lowered from $T_{1,0}^{\text{before}}$ to $T_{1,0}^{\text{after}}$ upon addition of R ligands with small binding time ϵ. (C) Interpolated digits with and without adversarial perturbation along the interpolation axis between $\vec{7}$ $(f = 0)$ and $\vec{3}$ $(f = 1)$. Adversarial perturbations are computed via the FGSM with $\epsilon = 0.2$. For the biomimetic defence $\phi(N, \theta)$, we choose $N = 5$ and $\theta = 0.5$. (D) Scoring function $J(\mathbf{x})$ on pictures of panel C without (left) and with (right) the biomimetic defence. The classification threshold is indicated by the dashed green line at $J = 0$. Samples with $J > 0$ are classified as 7, otherwise 3. (E) Interpolated ligand mixtures with and without self ligands along the interpolation axis between agonist $(f = 0)$ and antagonist $(f = 1)$. Here, $(L_{\text{ag}}, \tau_{\text{ag}}) = (100, 6)$; $(L_{\text{a}}, \tau_{\text{a}}) = (100, 1)$; $(L_{\text{self}}, \tau_{\text{self}}) = (1000, 0.1)$ (F) Scoring function on ligand mixtures of panel E for a naive immune classifier $(N, m) = (1, 0)$ (left) and a robust immune classifier $(N, m) = (2, 1)$ (right). The threshold is indicated by a dashed green line at $T_{N,m}/\tau_d - 1 = 0$. $T_{N,m}/\tau_d - 1 > 0$ corresponds to detection of agonists, below corresponds to no detection. In both digit recognition and ligand discrimination, the naive networks interpolate the score linearly and are sensitive to adversarial perturbations, while the score for robust networks is flatter, closer to the initial samples for longer, thus more resistant to perturbation.

consider a situation with L identical bound ligands with ($k^{\text{on}} = 1$, binding time τ) giving response $T_{1,0}^{\text{before}} = \tau$ where τ itself is of order 1 (i.e. much bigger than the ϵ-sized perturbation on binding time considered in Eq. 3.4). The three rules above imply that we are to decrease binding time by ϵ, and that all R previously unbound receptors are now to be bound by ligands with $k^{\text{on}} = \epsilon$, with small binding time ϵ. We compute the new response to be

$$T_{1,0}^{\text{after}} = \frac{L(\tau - \epsilon) + \epsilon R \epsilon}{L + \epsilon R} = \frac{\tau - \epsilon + \frac{\epsilon R}{L}\epsilon}{1 + \frac{\epsilon R}{L}} \tag{3.5}$$

If there are many receptors compared to initial ligands, and assuming $\epsilon \ll \tau$, the relative change

$$\frac{T_{1,0}^{\text{after}} - T_{1,0}^{\text{before}}}{T_{1,0}^{\text{before}}} \simeq -\frac{\frac{\epsilon R}{L}}{1 + \frac{\epsilon R}{L}} \tag{3.6}$$

is of order 1 when $\epsilon R \sim L$, giving a decrease comparable to the original response instead of being of order ϵ as we would naturally expect from small perturbations to all parameters. Thus, if a detection process is based on thresholding variable $T_{1,0}$, a significant decrease can happen with such perturbation, potentially shutting down response. Biologically, the limit where ϵR is big corresponds to a strong antagonistic effect of many weakly bound ligands. Examples can be found in mast cell receptors for immunoglobin: weakly binding ligands have been suggested to impinge a critical kinase thus preventing high affinity ligands to trigger response [57], a so-called “dog in the manger" effect. Another example is likely found in detection by NK cells [92]. A similar effect called “competitive antagonism” is also observed in olfaction where ligands with strong inverse affinity can impinge action of other ligands [83]. One difference in olfaction is that for competitive antagonism, the concentration C is of order 1 while the affinity κ^{-1} is big, conversely, here the concentration R is big while k^{on} is low. Since we consider the product of both terms, both situations lead to similar effects, but our focus on a small change of k^{on} makes the comparison with machine learning more direct.

Behaviour across boundaries in sample space and adversarial perturbations

To further illustrate the correspondence, we compare the behaviour of a trained neural network classifying 3s and 7s with the adaptive proofreading model $(N, m) = (1, 0)$ for more general samples. We build linear interpolations between two samples on either side of the decision boundary for both cases (Fig. 3.3 C–F, linear interpolation factor f varying between 0 and 1). This interpolation is the most direct way in sample spaces to connect objects in two different categories. The neural network classifies linearly interpolated digits, while the adaptive proofreading model classifies gradually changing ligand distributions.

We plot the output of the neural network x just before taking the sigmoid function σ defined in Fig. 3.2 C and similarly, we plot $T_{N,m}/\tau_d - 1$ for adaptive proofreading models. In both cases the decision is thus based on the sign of the considered quantity. In the absence of adversarial/antagonistic perturbations, for both cases, we see that the score of the system almost linearly interpolates between values on either side of the classification boundary (top panel of Fig. 3.3 D, F, blue curves). However, in the presence of adversarial/antagonistic perturbations, the entire response is shifted way below the decision boundary (top panel of Fig. 3.3 D, F, red curves), so that in particular the initial samples at $f = 0$ (image of 7 or ligand distribution above threshold) are strongly misclassified.

Goodfellow et al. [74] proposed the linearity hypothesis as an explanation for this adversarial effect: adding $\eta = \epsilon$ sgn $(\nabla_x J)$ to the image leads to a significant perturbation on the scoring function J of order ϵd, with d the usually high dimensionality of the input space. Thus many weakly lit up background pixels in the initial image can conspire to fool the classifier, explaining the significant shift in the scoring function in Fig. 3.3 D top panel. This is consistent with the linearity we observe on the interpolation line even without adversarial perturbations. A more quantitative explanation based on averaging is given in [102] on a toy-model, that we reproduce below to further articulate the analogy: after defining a label $y \in \{-1, +1\}$, a fixed probability p and a constant η, one can create a $(d + 1)$ dimensional feature vector x.

$$y \in \{-1,+1\}, \quad x_1 \sim \begin{cases} +y, & \text{w.p. } \; p \\ -y, & \text{w.p. } \; 1-p \end{cases} \tag{3.7}$$
$$x_2, \ldots, x_{d+1} \in \mathcal{N}(\eta y, 1)$$

From this, Tsipras et al. build a 100% accurate classifier in the limit of $d \to \infty$ by averaging out the weakly correlated features $x_2, \ldots x_d$, which gives the score $f_{\text{avg}} = \mathcal{N}(\eta y, \frac{1}{d})$. Taking the sign of f_{avg} will coincide with the label y with 99% confidence for $\eta \geq 3/\sqrt{d}$. But such classification can be easily fooled by adding a small perturbation $\epsilon = -2\eta y$ to every component of the features, since it will shift the average by the same quantity $-2\eta y$, which can still be small if we take $\eta = O(1/\sqrt{d})$ [102].

We observe a very similar effect in the simplest adaptive proofreading model. The strong shift of the average $T_{1,0}$ in Eq. 3.5 is due to weakly bound receptors ϵR, which play the same role as the weak features (components $x_2, \ldots, x_{d+1}$ above), hiding the ground truth given by ligands of binding time τ (equivalent to x_1 above) to fool the classifier. We also see a similar linearity on the interpolation in Fig. 3.3 F top panel. There is thus a direct intuitive correspondence between adversarial examples in machine learning and many weakly bound ligands. In both cases, the change of scoring function (and corresponding misclassification) can be large despite the small amplitude ϵ of the perturbation. Once this perturbation is added, the system in Fig. 3.3 still interpolates between the two scores in a linear way, but with a strong shift due to the added perturbation.

Biomimetic defence for digit classification inspired by adaptive sorting

Kinetic proofreading, famously known as the error-correcting mechanism in DNA replication [27, 28], has been proposed as a mechanism for ligand discrimination [26]. In the adaptive proofreading models we are studying here, kinetic proofreading allows the encoding of distinct τ dependencies in the activation/repression branches [34]. The primary effect of kinetic proofreading is to nonlinearly decrease the relative weight of weakly bound ligands with small binding times, thus ensuring defence against antagonism by weakly bound ligands. Inspired by this idea, we implement a simple defense for digit classification. Before feeding a picture to the neural network, we transform individual pixel values x_i of image $\mathbf{x}$ with a Hill function as

$$x_i \leftarrow \phi_{N,\theta}(x_i) = \frac{x_i^N}{x_i^N + \theta^N}, \tag{3.8}$$

where N (coefficient inspired by kinetic proofreading) and $\theta \in [0, 1]$ are parameters we choose. Similarly to the defence of adaptive proofreading where ligands with small τ are filtered out, this transformation squashes greyish pixels with values below threshold θ to black pixels, see Fig. 3.3 C bottom panels.

In Fig. 3.3 D, bottom panel, we show the improved robustness of the neural network armed with this defence. Here, the adversarial perturbation is filtered out efficiently. Strikingly, with or without adversarial perturbation, the score now behaves nonlinearly along the interpolation line in sample space: it stays flatter over a broad range of f until suddenly crossing the boundary when the digit switches identity (even for a human observer) at $f = 0.5$. Similarly, for adaptive sorting with $(N, m) = (2, 1)$, antagonism is removed, and the score exhibits the same behaviour of flatness followed by a sudden decrease on the interpolation line. Thus, similar defence displays similar robust behaviour of the score in sample space.

Gradient dynamics identify two different regimes

The dynamics of the score along a trajectory in sample space can thus vary a lot as a function of the model considered. This motivates a more general study of a worst-case scenario, i.e. gradient descent towards the decision boundary for different models. Krotov and Hopfield studied a similar problem for an MNIST digit classifier, encoded with generalized Rectified polynomials of variable degrees n [103] (reminiscent of the iterative FGSM introduced in [104]). The general idea is to find out how to most efficiently reach the decision boundary, and how this depends on the architecture of the decision algorithm. Krotov and Hopfield identified a qualitative change with increasing n, accompanied by a better resistance to adversarial perturbations [91, 103].

We consider the same problem for adaptive proofreading models, and study the potential-derived dynamics of binding times for a ligand mixture with identical k_{on} when following the gradient of $T_{N,m}$ (akin to a potential in physics). The adversarial goal is to fool the classifier with a minimal change in a given example (or in biological terms, how to best antagonize it). We iteratively change the binding time of non-agonist ligands $\tau < \tau_d$ to

$$\tau \leftarrow \tau - \epsilon \frac{\partial T_{N,m}}{\partial \tau} \tag{3.9}$$

while keeping the distribution of agonist ligands with $\tau > \tau_d$ constant. In the immune context, these dynamics can be thought of as a foreign agent selected by evolution to antagonize the immune system. Some biological constraints will force ligands to stay above threshold, so the only possible evolutionary strategy is to mutate and generate antagonists ligands to mask its non-self part. Such antagonistic phenomena have been proposed as a mechanism for HIV escape [29, 85] and associated vaccine failure [86]. Similar mechanisms might also be implicated in the process of tumour immunoediting [88].

From a given ligand mixture with few ligands above threshold and many ligands below thresholds, we follow the dynamics of Eq. 3.9, and display the ligand distribution at the decision boundary for different values of N, m as well as the number of steps to reach the decision boundary in the descent defined by Eq. 3.9 (Fig. 3.4, see also Fig. SB.1 for another example with a visual interpretation). We observe two qualitatively different dynamics. For $m < 2$, we observe strong adversarial effects, as the boundary is almost immediately reached and the ligand distribution barely changes. As m increases, in Fig. 3.4 A the ligands in the distribution concentrate around one peak. For $m = 2$, a qualitative change occurs: the ligands suddenly spread over a broad range of binding times and the number of iterations in the gradient dynamics to reach the boundary drastically increases. For $m > 2$, the ligand distribution becomes bimodal, and the ligands close to $\tau = 0$ barely change, while a subpopulation of ligands peaks closer to the boundary. Consistent with this, the number of ϵ-sized steps to reach the boundary is 3 to 4 orders of magnitude higher for $m > 2$ as for $m < 2$.

Qualitative change in dynamics is due to a critical point in the gradient

The qualitative change of behaviour observed at $m = 2$ can be understood by studying the contribution to the potential $T_{N,m}$ of ligands with very small binding times $\tau_\epsilon \sim 0$. Assuming without loss of generality that only two types of ligands are present (agonists $\tau_{\text{ag}} > \tau_d$ and spurious $\tau_{\text{spurious}} = \tau_\epsilon$), an expansion in τ_ϵ gives, up to a constant, $T_{N,m} \propto -\tau_\epsilon^m$ for small τ_ϵ (see Fig. 3.4 B for a representation of this potential and Appendix S3 for this calculation). In particular, for $0 < m < 1$, $\frac{\partial T_{N,m}}{\partial \tau_\epsilon} \propto -\tau_\epsilon^{m-1}$ diverges as $\tau_\epsilon \to 0$. This corresponds to a steep gradient of $T_{N,m}$ so that the system quickly reaches the boundary in this direction. The ligands close to $\tau_\epsilon \sim 0$ then quickly localize close to the minimum of this potential (unimodal distribution of ligand for small m on Fig. 3.4 A–B).

The potential close to $\tau_\epsilon \sim 0$ flattens for $1 < m < 2$, but it is only at $m = 2$ that a critical point in the *gradient* (i.e. characterized by $\partial^2 T_{N,m}/(\partial \tau_\epsilon)^2 = 0$) appears at $\tau_\epsilon = 0$. This qualitatively modifies the dynamics defined by Eq. 3.9. For $m \geq 2$, due to the new local flatness of this gradient, ligands at $\tau = 0$, the dynamical critical

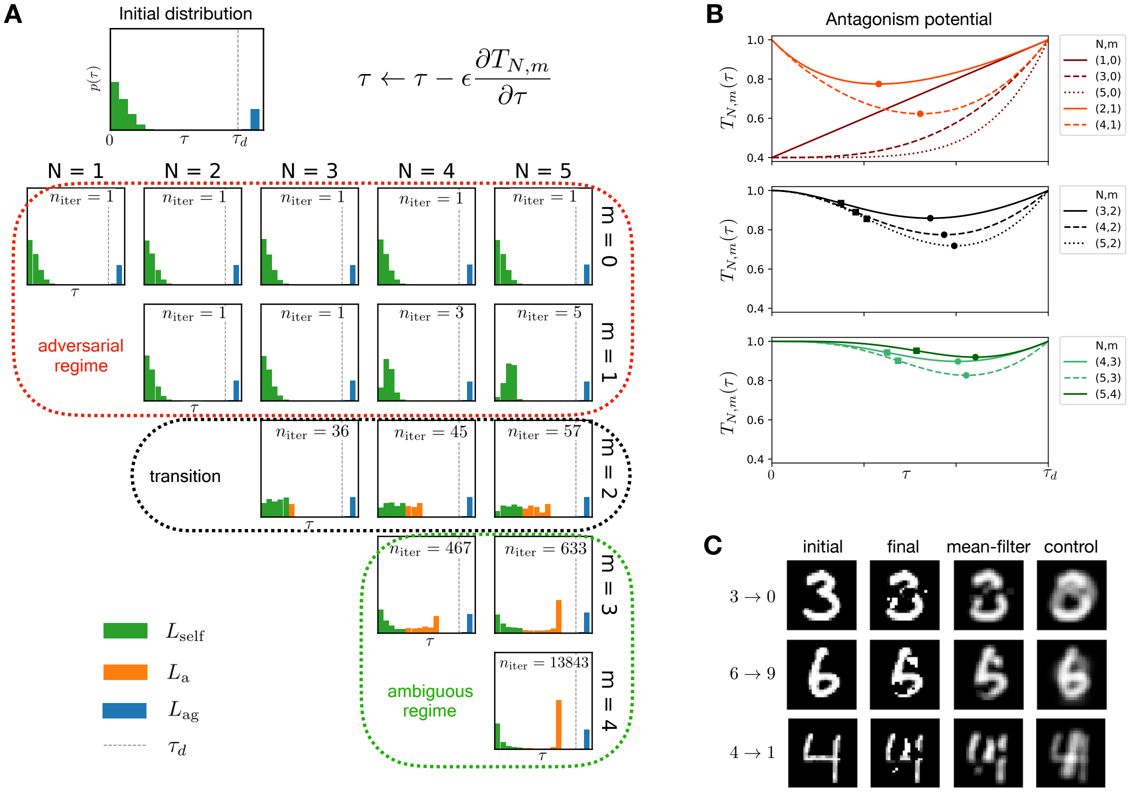


Figure 3.4: **Characterization of the decision boundary following gradient descent dynamics.** (A) Ligand distribution at the decision boundary by applying iterative gradient descent (top right of the panel) to an initial distribution (top left). For various cases (N, m) we change the binding time of self ligands along the steepest gradient until reaching the decision boundary. n_{iter} indicates the number of iterations needed to reach the decision boundary. We identify the adversarial regime (red), the ambiguous regime (green) and a transition (black) depending on m. (B) $T_{N,m}$ for mixtures of ligands at τ_d and ligands at τ, as a function of τ for various (N, m). Antagonism strength is maximal when $T_{N,m}$ is minimal. Minima and inflexion points are indicated with a circle and square. (C) Few-pixel attack as a way of circumventing proofreading or local contrast defence, while creating ambiguous digits. We add a 3x3 mean-filter to demonstrate the ambiguity of digits at the decision boundary. The control image is the mean filtered initial digit combined with the locally contrasted average target digit. Note that also the control is lacking a clear ground truth.

point of Eq. 3.9, are pinned by the dynamics. By continuity, dynamics of the ligands slightly above $\tau_\epsilon = 0$ are critically slowed down, making it much more difficult for them to reach the boundary. This explains both the sudden broadening of the ligand distribution, and the associated increase in the number of steps to reach the decision boundary. Conversely, an inflexion point (square) appears in between the minimum (circle) and $\tau_\epsilon = 0$ (Fig. 3.4 B). Ligands close to the inflexion point separate and move more quickly towards the minimum of potential, explaining the bimodality at the boundary (if we were continue the dynamics past the boundary, all ligands with non-zero binding times would collapse to the minimum of the potential). For both larger N and larger m we obtain flatter potentials, and a larger number of iterations. In Appendix S4, we further describe the consequence of adding proofreading steps on the position of the boundary itself, using another concept of machine learning called “boundary tilting” [105] (Fig. SB.2 and Table S1).

Categorization of attacks

The transition at $m = 2$ is strongly reminiscent of the transition observed by Krotov and Hopfield in their study of gradient dynamics similar to Eq. 3.9 [103]. In both our works, we see that there are (at least) two kinds of attacks that can bring samples to the decision boundary. The FGSM corresponds to small perturbations to the input in terms of L_∞ norm leading to modifications of many background pixels in [103] or many weakly bound ligands for the adaptive proofreading case, also similar to the meaningless changes in $x_2, \dots x_d$ described above in Eq. 3.7 [102].

Defence against the FGSM perturbation is implemented through a higher degree n of the rectified polynomials in [103], while in adaptive proofreading, this is done through critical slowing down of the dynamics of Eq. 3.9 for $m > 2$. The latter models are nevertheless sensitive to another kind of attack with many fewer perturbations of the inputs but with bigger magnitude. This corresponds to digits at the boundary where few well-chosen pixels are turned on in [103]. For adaptive proofreading models this leads to the ligand distribution becoming bimodal at the decision boundary. Three important features are noteworthy. First, the latter perturbations are difficult to find through gradient descent (as illustrated by the many steps to reach the boundary in Fig. 3.4A). Second, the perturbations appear to be meaningful: they correspond to interpretable features and interfere with the original sample. These perturbations make it difficult or even impossible to recover the ground truth by inspecting the sample at the decision boundary. Digits at the boundary for [103] appear indeed ambiguous to a human observer, and ligand distribution peaking just below threshold are potentially misinterpreted biologically due to inherent noise. This has actually been observed experimentally in T cells, where strong antagonists are also weak agonists [31, 33], meaning that T cells do not take reliable decisions in this regime. Lastly, it has been observed in machine learning that memory capacity considerably increases for high n in [91], due to the local flattening of the landscape close to memories (ensuring that random fluctuations do not change memory recovery). A similar effect in our case is observed: the antagonism potential is flattened out with increasing N, m so that any spurious antagonism becomes at the same time less important and lies closer to the decision boundary.

Biomimetic defenses against few-pixel attacks

It is then worth testing the sensitivity to localized stronger attacks of digit classifiers, helped again with biomimetic defences. The natural analogy is to implement attacks based on strong modification of few pixels [106].

For this problem, we choose to implement a two-tier biomimetic defence: we implement first the transformation defined in Eq. 3.8, that will remove influence of the FGSM types of perturbations by flattening the local landscape as in Fig. 3.3 D. In addition, we choose to add a second layer of defence where we simply average out locally pixel values. This can be interpreted biologically as a process of receptor clustering or time-averaging. Time-averaging has been shown to be necessary in a stochastic version of adaptive proofreading [33, 34], where temporal intrinsic noise would otherwise make the system cross the boundary back and forth endlessly. In the machine learning context, local averaging has been recently proposed as a way to defend against few pixel attacks [107], which thus can be considered as the analogous of defending against biochemical noise.

We then train multiple classifiers between different pairs of handwritten digits.

Following the approach of the “one pixel” attack [106], we consider digits classified in presence of this two-tier defence, then sequentially fully turn pixels on or off ranked by their impact on the scoring function, until we reach the decision boundary. Details on the procedure are described in Appendix S5. A good defence would manifest itself similarly to the Krotov-Hopfield case [103], where no recognizable (or ambiguous) digits are observed at the boundary.

Representative results of such few-pixel attacks with biomimetic defences are illustrated in Fig. 3.4 C. The “final" column shows the misclassified digits after the attack and the “mean-filter" column shows the local average of the “final" digits for further comparison, with other examples shown in Fig. SB.3 and details on the behaviour of scoring functions in Fig. SB.4. Clearly the attacked samples at the boundaries hide the ground truth of the initial digit, and as such can not be considered as typical adversarial perturbations. Samples at the boundary are out-of-distribution but preserve structure comparable to written characters (e.g. attacks from 0 to 1 typically look like a Greek ϕ, see Fig. SB.3). This makes them impossible to classify as Arabic digits even for a human observer. This is consistent with the ambiguous digits observed for big n by Krotov and Hopfield [103]. In other cases, samples at the boundary between two digits actually look like a third digit: for instance, we see that the sample at the boundary between a 6 and an 9 looks like a 5. This observation is consistent with previous work attempting to interpolate in latent space between digits [108], where at the boundary a third digit corresponding to another category may appear. We also compare in Fig. 3.4 C the sample seen by the classifier at the boundary after the biomimetic defences with a “control” corresponding to the average between the initial digit and the target of the attack (corresponding to the interpolation factor $f = 0.5$ in Fig. 3.3 C–D). It is then quite clear that the sample generated by the attack is rather close to this control boundary image. This, combined with the fact that samples at the boundary still look like printed characters without clear ground truth indicate that the few pixel attacks implemented here actually select for meaningful features. The existence of meaningful features in the direction of the gradient have been identified as a characteristic of networks robust to adversarial perturbation [102] similar to results of [103] and our observation for adaptive proofreading models above.

3.3 Discussion

Complex systems (*in vivo* or *in silico*) integrate sophisticated decision making processes. Our work illustrates common features between neural networks and a general class of adaptive proofreading models, especially with regards to mechanisms of defence against targeted attacks. Parallels can be drawn between these past approaches, since the models of adaptive proofreading presented here were first generated with *in silico* evolution aiming at designing immune classifiers [34]. Strong antagonism naturally appeared in the simplest simulations, and required modification of objective functions very similar to adversarial training [74].

Through our analogy with adaptive proofreading, we are able to identify the presence of a critical point in the gradient of response as the crucial mediator of robust adversarial defense. This critical point emerges due to kinetic proofreading for cellular decision network, and essentially removes the spurious adversarial directions. Another layer of defence can be added with local averaging. This is in line with current research on adversarial robustness in machine learning, showing that robust networks exhibit a flat loss landscape near each training sample [109]. Other current explorations include new biomimetic learning algorithms, giving rise to prototype-like classification [110]. Adversarial defence strategies, including non-local computation and nonlinearities in the neural network are also currently under study [107]. The mathematical origin of the effectiveness of those defences is not yet entirely clear, and identification of critical points in the gradient might provide theoretical insights into it.

More precisely, an interesting by-product of local flatness, where both the gradient and second derivative of the score are equal to zero, is the appearance of an inflexion point in the score, and thus a region of maximal gradient. This is visible in Fig. 3.3 D, F: while the score of non-robust classifiers is linear when moving towards the decision boundary, the scoring function of classifiers resistant to adversarial perturbations is flat at $f = 0$ and only significantly changes when the input becomes ambiguous near the inflexion point. The reason why this is important in general is that a combination of local flatness and an inflexion point is bound to strongly influence any gradient descent dynamics. For instance, for adaptive proofreading models, the ligand distribution following the dynamics of Eq. 3.9 changes from unimodal to bimodal at the boundary, creating ambiguous samples. For a robust classifier, such samples are thus expected to appear close to the decision boundary since they coincide with the larger gradients of the scoring function. As such they could correspond to meaningful features (contrasting the adversarial perturbations), as we show in Fig. 3.4 C with our digit classifier with biomimetic defence. Examples in image classification might include the meaningful adversarial transformations between samples found in [102] or the perturbed animal pictures fooling humans [111] with chimeric images that combine different animal parts (such as spider and snake), leading to ambiguous classifications. Similar properties have been observed experimentally for ambiguous samples in immune recognition: maximally antagonizing ligands have a binding time just below the decision threshold [31]. We interpret this property as a consequence of the flat landscape far from the decision threshold leading to a steeper gradient close to it [33, 60].

We used machine learning classification and implemented biomimetic defence by relying on a single direction, since that is what emerges in the most simple version of adaptive proofreading models that we considered here. In general, however, the space of inputs in machine learning is much more complex, and there are more than two categories, even in digit classification. One possible solution is to break down multilabel classification into a set of binary classification problems, but this might not always be appropriate. Instead, the algorithm effectively has to learn representations, such as pixel statistics and spatial correlations in images [75]. With a nonlinear transformation to a low-dimensional manifold description, one could still combine information on a global level in ways similar to parameter τ. The theory presented here could then apply once the mapping of the data from the full-dimensional space to such latent space is discovered.

Case-in-point, Tsipras et al. proposed a distinction in machine learning between a robust, but probabilistic feature (x_1 in Eq. 3.7) and weakly correlated features ($x_2, \dots x_d$ in Eq. 3.7) [102], both defining a single direction in latent space. They then observed a robustness-accuracy trade-off due to the fact that an extremely accurate classifier would mostly use a distribution of many weakly correlated features (instead of the robust – but randomized – feature) to improve accuracy. The weight to put in the decision on either feature (robust or weak) would depend on the training. Our work shows the natural connection between weak features in this theory and weak ligands in the biological models (see discussion below Eq. 3.7). In the biological context, the standard situation is that all ligands are treated equally. Then one can show mathematically that for such networks performing quality sensing irrespective of quantity, antagonism necessarily ensues [60], as further identified here using the FGSM transformation. This latter result can be reformulated in terms of machine learning [102] in the following compact way: perfectly robust classification (i.e. with no antagonism) is impossible in biology if all receptors are equivalent. But biology also provides evidence that robustness can nevertheless be improved by applying local nonlinear transformation such as the biomimetic defence of Eq. 3.8. Elaborating on the distinction between robust and weak features proposed in [102], nonlinear transformations should specifically target weak correlated features. Explorations of generalized nonlinear transformations in image feature space [91, 103] might lead to further insights into the possible nonlinear transformations defending against adversarial perturbations. We learn in particular from biology that the major effect of nonlinearity is to change the position of maximally adversarial perturbations in sample space. Perfect robustness might be impossible in general, yet similarly to cellular decision-making the most effective perturbations may shift from a pile of apparently unstructured features for naive classifiers to a combination of meaningful features for robust classifiers, giving ambiguous patterns at the decision boundary (allowing to further distinguish between ambiguous and adversarial perturbations).

From the biology standpoint, new insights may come from the general study of computational systems built via machine learning. In particular, systematic search and application of adversarial perturbations in both theoretical models and experiments might reveal new biology. For instance, our study of Fig. 3.4, inspired by gradient descent in machine learning [103], establishes that cellular decision-makers exist in two qualitatively distinct regimes. The difference between these regimes are geometric by nature through the presence or absence of a dynamical critical point in the gradient. The case $m < 2$ with a steep gradient could be more relevant in signalling contexts to separate mixtures of inputs, so that every weak perturbation *should* be detected [98]. For olfaction it has been suggested that strong antagonism allows for a rescaling of the distribution of typical odor molecules, ensuring a broad range of detection irrespective of the quantity of molecules presented [83]. The case $m \geq 2$ is much more resistant to adversarial perturbations, and could be most relevant in an immune context where T cells filter out antagonistic perturbations. This might be relevant for the pathology of HIV infections [29, 85, 86] or, more generally, could provide explanations on the diversity of altered peptide ligands [112]. We also expect similar classification problems to occur at the population-level, e.g. when T cells interact with each other to refine individual immune decision-making [113, 114]. Interestingly, there might be there a trade-off between resistance to such perturbations (in particular to self antagonism, pushing towards higher m in our model) and the process of thymic selection which relies on the fact that there should be sensitivity to some self ligands [115] (pushing towards lower m in our model) .

Our correspondence could also be useful for the theoretical modelling and understanding of cancer immunotherapy [87]. So-called neoantigens corresponding to mutated ligands are produced by tumours. It has been observed that in the presence of low-fitness neoantigens, the blocking of negative signals on T cells (via checkpoint inhibitor blockade) increases success of therapy [116]. This suggests that those neoantigens are ambiguous ligands: weak agonists acting in the antagonistic regime. Without treatment, negative signals prevent their detection (corresponding to an adversarial attack), but upon checkpoint inhibitor blockade those ligands are suddenly visible to the immune system, which can now eliminate the tumour. Importantly, differential responses are present depending on the type of cancer, environmental factors and tumour microenvironment [88]. This corresponds to different background ligand distributions in our framework, and one can envision that cancer cells adapt their corresponding adversarial strategies to escape the immune system. Understanding and categorizing possible adversarial attacks might thus be important to predicting the success of personalized immunotherapy [117].

We have connected machine learning algorithms to models of cellular decision-making, and in particular their defence strategies against adversarial attacks. More defences against adversarial examples might be found in the real world, for instance in biofilm-forming in bacteria [118], in size estimation of animals [119], or might be needed for proper detection of physical 3D objects [120] and road signs [121]. Understanding the whole range of possible antagonistic perturbations may also prove crucial for describing immune defects, including immune escape of cancer cells. It is thus important to further clarify possible scenarios for fooling classification systems in both cell biology and machine learning. ” *(Attack and defence [2])*

4 Cytokines

The cytokine project is the most explicit project where reducing data leads to a latent space representation. Here, obtaining the latent space representation is also the most straightforward: after processing the data and setting up the classification procedure, the latent space is just a linear transformation away. Yet, it is highly nontrivial to find the right basis function from which we can retrieve the latent space. The classification problem concerns predicting antigen class given a cytokine profile, an output of the immune response. Interestingly, through biophysical modelling of the latent space dynamics, we predict antigenicity quantitatively, instead of qualitatively as the initial classification problem was designed. This work serves as an example on how ideas borrowed from physics and techniques borrowed from machine learning can lead to progress in immunology.

4.1 Introduction

The introduction is structured as follows: First, I provide background on the cytokine response following activation of the adaptive immune system, complementing section 1.2 on the composition of the adaptive immune system. I then introduce the goal of the project and discuss its applicability and parallel approaches.

Background

Cytokine complexity

The fundamentals of cytokine-mediated communication can be compared to radiostations emitting and receiving radiosignals. Different immune cell types emit signals at different channels, intended for a subset of cell types. Upon receiving a signal through their cytokine receptors (receivers), the cells internally process the information, mostly through the JAK/-STAT pathway (note that the comparison between the processing units of a radio and a signal transduction pathway has been made previously [4]). Cytokines activate a unique set of JAKs and STATs; a mix of cytokines results in crosstalk between JAKs and STATs (interference) [122]. This pattern is recognized by the immune cell, and causes a corresponding response, for instance, through proliferation (build new radio stations) initiating or upregulating production of cytokines (enhance emission) or upregulation of cytokine receptors (enhance receiving). Cytokines signals are generally received by the producing T cell (autocrine signalling) as well as by surrounding cells (paracrine signalling), although T cells do not produce and consume IL-2 simultaneously [123, 124]. Finally, to conclude the comparison between communication through cytokines and radiowaves, the reach of the cytokines is limited (inverse square-law of intensity).

Despite the diversity of cytokines and complexity through internally shared processing units in the T cell, the premise is the same: molecular cues cause T cells and surrounding tissue to produce and consume cytokines according to

$$\frac{d[cy]}{dt} = +\kappa_+(t) - \kappa_-(t) \tag{4.1}$$

Here, $[cy]$ is the concentration of a given cytokine and $\kappa_+(t)$ and $\kappa_-(t)$ are the rates of production and consumption of the cytokines. The production term $\kappa_+(t)$ can be broken down further into

$$\kappa_+(t) = N_+(t)k_+(t)/(VN_A) \tag{4.2}$$

where $N_+(t)$ is the number of cells over time producing cytokines with rate $k_+(t)$. V is the extracellular volume and N_A is Avogadro's constant to convert numbers to concentration. The consumption term $\kappa_-(t)$ can be broken down further into

$$\kappa_-(t) = k_{\text{on}} N_-(t) R(t)/(V N_A) \tag{4.3}$$

where k_{on} is the affinity of cytokine and receptor, $N_-(t)$ is the number of consuming cells and $R(t)$ is the number of cytokine receptors per cell over time. N_+, N_-, R, k_+ can be arbitrarily complex over time, and may depend on $[cy]$ as well [18]. Furthermore, spatial considerations through diffusion and advection regulate the reach of the signals [18, 125]. Another regulating factor is competition for cytokines resulting in a balance between immune tolerance and immune response [125]. With 33 known interleukin families, many other non-interleukin cytokine families (i.e. interferons, tumor necrosis factor superfamily), and up to 10 or more members per interleukin family, a separate gene encoding for each member [126], it is clear that the cytokine code is of enormous complexity. Not all cytokines are of equal importance though; a good indicator of a cytokine's importance is if it has been used as a therapeutic target, a list that includes TNFα, IFNγ, IL-2, IL-6, IL-10 and IL-17A [127], not coincidentally the cytokines we study too. Yet, decoding messages sent in this code is a task whose surface we have hardly begun to scratch. Altan-Bonnet and Mukherjee underline the importance of quantitatively studying cytokines: "The ability of individual cells to process signals at multiple levels and to integrate the obtained information into a collective response at the population level is crucial for a coordinated immune response. As such, it is important to have a quantitative understanding of how immune cells integrate the large number of signals they receive into a tailored output" [125]. The goal of this work is to do exactly that. Before going there, we provide a contemporary qualitative understanding of the cytokines under study and introduce models aimed at understanding some of this decoding.

IL-6 and TNFα are among the most pleiotropic cytokines, meaning they are produced and consumed by many different cell types. They are typically known as proinflammatory mediators - a catch-all term for cytokines with many functions - and remain present througout the entirety of an acute response. IL-17A and IL-2 are more specialized cytokines produced by helper T cells (IL-17A, IL-2) and killer T cells (IL-2) [128]. IL-17A is known for inducing production of other cytokines like IL-6 and TNFα [129], and binding of IL-2 to the IL-2R is critical for inducing and regulating T cell expansion [130]. As such, IL-2 intuitively encodes information about antigen quality to determine the extent of T cell expansion. Finally, a major role of IFNγ concerns activation of macrophages [131] through its production by T cells, both in innate and adaptive immunity.

Regulation of the immune response

Regulation of the immune response occurs not only through cytokine receptor signalling, but also at the population level. Regulatory T cells (Tregs) consume IL-2 without producing any, inhibiting the expansion of helper and killer T cells [132–134]. T cell expansion is also limited by antigen availability [135]. Several studies have proposed models of regulation of T cell expansion through antigen consumption [136–138]. The last one proposes a simplification of the models of previous studies, aiming to preserve the observations that T cells proliferate exponentially at saturated antigen levels, and that antigens decay over time. The model is given by

$$\frac{dT}{dt} = \alpha \frac{TC}{K+T+C} - \delta T \tag{4.4}$$

$$\frac{dC}{dt} = -\mu C \tag{4.5}$$

where $T(t)$ and $C(t)$ represent the T cell number and antigen quantity at time t, K is the antigen quantity at which the T cells' response to the antigen is at half of its maximum response, also called EC$_{50}$. α is the proliferation rate, δ the rate at which T cells die, and μ the antigen decay rate. This model consists of two phases: exponential proliferation when antigen quantity saturates antigen quality and T cell number $C(t) \gg K, T(t)$ and a rapid slow-down of proliferation when either of the two conditions are no longer satisfied. In the competition-limited regime when $C(t) \sim T(t)$, the characteristic time t^* at which the transition between the two phases occurs is given by

$$t^* = \frac{1}{\alpha - \delta + \mu} \log \frac{C(0)}{T(0)}. \tag{4.6}$$

In the affinity-limited regime when $C(t) \sim K$, t^* is given by

$$t^* = \frac{1}{\mu} \log \frac{C(0)}{K}. \tag{4.7}$$

From these considerations, it emerges naturally that the fold expansion $f = T(t^*)/T(0)$ decreases with precursor frequency $N(0)$. In the competition-limited regime, Mayer et al. fitted experimental data and found the following power-law relation

$$f = \left(\frac{T(t^*)}{T(0)}\right) \propto T(0)^{-1/2}. \tag{4.8}$$

This raises the question of how the trade-off between resources spent on proliferating and minimum clone size following immune recognition was established. This model is beautiful in its simplicity, capturing the power-law relation between precursor frequency and fold expansion, yet provides no mechanistic insight into regulation of T cell expansion. It also ignores phenotypic variability, which causes substantial diversity of activation within a clonal population of T cells [64]. Tkach et al. developed a model taking into account the internal regulation of IL-2 production and consumption for CD4+ T cells (helper T cells) [124], and found evidence for an experimental scaling law in the peak IL-2 concentration

$$[\text{IL-2}]_{\max} \propto T_0^{-0.1}[Ag]^{0.8} \tag{4.9}$$

where T_0 is the precursor frequency and $[Ag]$ is the antigen quantity. With higher T_0, $[\text{IL-2}]_{\max}$ is smaller, and is reached earlier, because of the lower value and the higher number of activated T cells producing IL-2. Tkach et al. attributed the lower peak and higher number of activated cells needed to the per cell acceleration of IL-2 production over time [124]. With fixed proliferation time, we thus see a higher fold expansion for smaller T_0, which at least qualitatively reproduces the power-law of Eq. 4.8. As we are comparing in-vitro with in-vivo dynamics, the exact power may vary from the power $-\frac{1}{2}$ found in [138]. This is one case where Eq. 4.1 is solved explicitly for IL-2. Another example is given by Voisinne et al., who developed and experimentally tested a model of antigen discrimination for CD8+ T cells (killer T cells) using local cues (antigen quality and quantity) and global cues (cytokines) [114]. By considering phenotypic variability in the response threshold, they found that increased IL-2 levels engage otherwise weakly activated clones.

Models that focused on the internal regulation of a single cytokine already required numerical integration of multiple equations per simulated cell [114, 124]. One can only imagine the complexity of a model including multiple cytokines and their indirect effects on each other for a clonal population invariably including phenotypic variability. The best way forward to understanding cytokine communication might be the decoding problem: given the cytokine response for a population of T cells to an immune challenge, what signals can we extract from this?

Goal

The goal of the project is to do exactly what we ended the previous section with: decode parameters of an immune challenge from detailed cytokine kinetics. The parameter we are especially interested in is the antigen quality, a parameter of clinical interest, for instance in predicting a patient's survival chances following checkpoint inhibitor therapy [116, 139]. Thus, we are looking to design a classifier that is able to classify antigen quality independent of antigen quantity. With a classifier, we mean a classifier in the classical machine learning sense, one that processes an input and outputs a category, for instance a multi-layer perceptron (MLP) [140]. Our definition of antigen quality is the pMHC-TCR binding time. Antigen quantity corresponds to the number of antigens a T cell binds to while deciding (how strongly) to activate. Antigen quantity is referred to as ligand concentration in previous chapters. Finally, with cytokine kinetics we mean the concentration of cytokines over time from the start of the experiment (mixing of loaded antigen presenting cells and T cells) until three days later when T cell expansion has completed and T cells in the well start dying.

Once we have found an antigen quality classifier, there are many directions along which we could gain novel insight in immune response. First, we quantify how many different antigen qualities a naive CD8+ T cell can reliably detect. We also limit the range of antigen quantity within which two antigens of similar quality can be told apart. In other words, we quantify the quantity-independence of T cell response. We also learn about the minimal architecture required to accurately classify antigens by their quality. Similarly, with a feature analysis, we find the minimum inputs required for the system to still classify well providing insight into the role of various cytokines for communication between T cells.

In line with the title of this thesis, by using a neural network as a classifier we gain access to a latent space[1], which we analyze in detail. In machine learning terms, this could be called in-depth feature analysis on out-of-distribution samples, inspired by feature analysis in random forest classifiers [141], but as far as we are aware, there is no equivalent, as the details of this analysis are highly specific to the problem.

If we believe that nodes in the classifier's latent space correspond to information processing inside the T cell (locally) or in a population of T cells (globally), each of these experimental conditions should result in interpretable latent space dynamics. Assuming that we have indeed captured ongoing biology, we parameterize the latent space dynamics, and observe how various experimental setups change the parameters of the model. Suddenly, we have expanded the classification problem of naive OT-1 CD8+ T cells responding to four well-known antigens to parameterizing any T cell - APC interaction with just four parameters. That is, if we may say so, no small achievement, which goes beyond classical machine learning, allowing us to understand macroscopic behavior of detailed microscopic interactions at many time-scales (from minutes for pMHC-TCR interactions to days for the population-wide response), and setting an example for novel methods of interpretable machine learning.

Complementary approaches

We conclude the introduction by outlining other approaches people have taken in predicting antigen quality.

Predicting antigen quality happens routinely by predicting the likelihood of pMHC

[1] Like in an autoencoder, this latent space is only informative if the input is sent through a bottleneck so that the information on the output hidden in the input is compressed to few dimensions (preferably two which has the added benefit of being visually interpretable)

presentation through from peptide-MHC binding affinity. Some have argued that pMHC stability is actually more predictive than pMHC binding affinity [142], but this is difficult to measure experimentally. State-of-the-art algorithms like NetMHCpan [143] use the Immune Epitope Database (IEDB) [144] - a dataset of binding affinities between pMHCs - to predict pMHC binding affinity for unseen ligands. This has been used to predict the likelihood of presentation of neoantigens, and thus the immunogenicity of a tumor. Such information is of interest in clinical settings in predicting disease progression. Based on the sequence alignment between neoantigens and the closest match in the IEDB, Luksza et al. predicted a patient's survival prediction given the tumor's neoantigens [116, 139]. The underlying assumption is that once a pMHC is presented to the T cells, there are T cell clones in the repertoire that are specific to this ligand. This brings us back to the introduction: even though there might be an immune response, it is not necessarily a strong response.

The most common approach for quantifying the strength of an immune response is with IFNγ Elispot [145]. Scientists, using strict guidelines [146] set parameters for automated Elispot readers, counting the number and size of spots on a sample, corresponding to T cell clusters producing IFNγ 12-24 hours after activation following T cell expansion. The number and size of the spots provide information on the strength of the immune response. IFNγ Elispot is used in clinical settings for its low expense, ease of use and fast turnover time. It is also used across studies to measure a patient's response to new treatment. Other Elispot devices measure IL-2 or even a combination of cytokines with Fluorospot [147]. A drawback of Elispot measurements is that there is no baseline reference and that parameters for the Elispot reader need to be adjusted by the scientists, introducing a level of subjectivity.

Our contribution is clear: predicting antigen quality with a reference using detailed cytokine kinetics, improving Elispot and complementing algorithms predicting pMHC binding affinity.

4.2 Materials and methods

In this section, we describe the experimental setup, data processing, and classification procedure.

Experimental setup

The TECAN robotic platform in the Immunodynamics lab of Dr. Altan-Bonnet at the National Cancer Institute, Bethesda, MD, is designed to automate pipetting, allowing its handlers to run many conditions in parallel. It allows for the measurements of detailed time-kinetics of multiday experiments. Our collaborators extract immune cells from transgenic OT-1 mice: CD8+ T cells with the OT-1 TCR and APCs with a single MHC allele for peptide loading. APCs from the spleens of B6 mice - splenocytes - are loaded in various quantities with the chicken ovalbumin (OVA) or SIINFEKL antigen [148, 149]. Alteration of the prototypical SIINFEKL antigen N4 include A2, Y3, Q4, T4, G4 and E1. Here, the letter corresponds to the amino acid substituted at the position in the peptide chain indicated by the number, e.g. Y3 corresponds to the SIYNFEKL antigen. Typical numbers are 10^5 CD8+ T cells, $3 \cdot 10^5$ splenocytes and an antigen concentration of 1μM mixed in 200 μL supernatant. The splenocytes and T cells are prepared separately and mixed at the start of the experiment on 96, 192 and 384 well plates. At regular intervals, 20 μL supernatant is taken from the wells and stored for postprocessing. This is replaced with 20 μL fresh media so the overall volume stays the same. To measure cytokine concentrations from the samples of supernatant, a small amount of supernatant is mixed with beads coated with antibodies that are specific to the cytokine. Then, a second set of beads is added to the solution. These beads are specific to the cytokine-antibody bond and tagged with a fluorescent protein. Beads specific to different cytokines have different colors, which is how the flow cytometer decomposes the fluoresence of the mixture into cytokine-specific signals. Absolute cytokine concentrations are obtained from the geometric Mean Fluorescence Intensity using calibration curves. The seven cytokines that were measured initially are IFNγ, IL-2, IL-4, IL-6, IL-10, IL-17A and TNFα. As IL-4 and IL-10 rarely gave a response, we excluded them from our analysis. In the final stages of the project, our collaborators started experimenting with assays allowing multiplexing of 30 cytokines. In the future, we could decide to train a classifier using all these, but we found that we need only two dimensions to accurately determine antigen quality, and these two dimensions can be created using the initial five cytokines.

In earlier experiments at every timepoint, the contents of the entire well were stored, effectively stopping the experiment in this well. A timeseries was formed from wells with the same setup, but measured at unique times. This protocol allowed us to characterize the surface markers on the cells over time too. Once we focused our attention on cytokines alone, in subsequent experiments a timeseries was generated by extracting a small amount of supernatant from the same well at regular intervals, and replenishing it with fresh medium after. Removal of cytokines could have slowed down the immune response, but after finetuning the amount of supernatant extracted, we detected no effect, which we tested through crossvalidation in the classifier (Appendix Fig. C.1). The new protocol also significantly reduced the number of sacrificed mice, making the experiments cheaper and animal-friendlier.

Data processing

Much of the data processing procedure is developed by François Bourassa, lab member and collaborator on this project. It is described in detail in his M.Sc. thesis [150]. During the procedure, we replace missing data, smooth the data and gain access to the data in between the experimental timepoints.

To understand the variability on the cytokine dynamics, we assume an exponential proliferation rate α for a population of T cells $T(t)$ with an initial number of T cells T_0.

$$T(t) = \hat{T}_0 e^{(\alpha+\eta_\alpha)t}. \tag{4.10}$$

$\hat{\alpha} = \langle\alpha\rangle + \eta_\alpha$ and $\hat{T}_0 = \langle T_0\rangle + \eta_{T_0}$ are sampled from α and T_0, which follow a normal distribution. We assume that a T cell produces cytokine cy with constant rate λ, also normally distributed.

$$cy(t) = \int_{t'=0}^{t} cy'(t')dt' = \int_{t'=0}^{t} \hat{\lambda}\hat{T}_0 e^{\hat{\alpha}t} = \frac{\hat{\lambda}\hat{T}_0}{\hat{\alpha}}\left(e^{\hat{\alpha}t} - 1\right) \tag{4.11}$$

where $cy'(t)$ is the instantaneous amount of cytokines produced at time t. This expression can be approximated by

$$\log(cy(t)) \simeq \log\left(\frac{\hat{\lambda}\hat{T}_0}{\hat{\alpha}}\right) + \hat{\alpha}t \tag{4.12}$$

which follows the lognormal distribution

$$\log(X(t)) = \log(\mu_X(t)) + \eta(t),\ \eta(t) \sim \mathcal{N}(0, \sigma_X^2(t)) \tag{4.13}$$

where $\mu_X(t) = \log\left(\frac{\hat{\lambda}\hat{T}_0}{\hat{\alpha}}\right) + \langle\alpha\rangle t$ is the average cytokine timecourse that we are interested in, and $\eta(t) = \eta_\alpha t$ is the experimental noise with variance $\sigma_X^2(t)$ that we want to "average away". The assumptions on α and λ are approximately valid in the cytokine production phase lasting up to 24 hours following stimulation (or longer with fewer initial T cells T_0). More details on where additional noise could arise are given in Appendix B of [150]. In any case, logarithms are implemented routinely in biological systems [151], as such, the logarithms of cytokine concentrations are a reasonable starting point for our analysis. Practically, the log transformation is given by

$$c(t) = \log_{10}\left(\frac{cy(t)}{LOD}\right) \tag{4.14}$$

where $cy(t)$ is the cytokine concentration, $c(t)$ is the logtransformed cytokine concentration and LOD is the lower limit of detection. The LOD is equal or close to the minimum measured concentration, which we use instead of the LOD when the LOD is not available.

Before applying the log transformation, we look for measurements where cytokine levels decreased by a factor of 10 or more between consecutive timepoints. This is evidence of a missing datapoint, which may have occurred for various reasons, like dried up supernatant, issues with the cytokine beads or fallen plates. Cytokines like IL-2 naturally disappear from the system, but it does so gradually, which is why we require a factor 10 decrease. Missing data is replaced by a linear interpolation through the previous and the next timepoint. We now log transform the data and smooth the data with a moving average applied to the center points. The timepoints are spaced evenly such that this smoothing is applicable to the whole timeseries. Finally, we fit B cubic splines (piecewise polynomials of order 3) to the smoothed cytokine concentrations using a variable number of knots, a procedure that is detailed in [150]. The effect of each of these steps of the data processing is visible in Fig. 4.1.

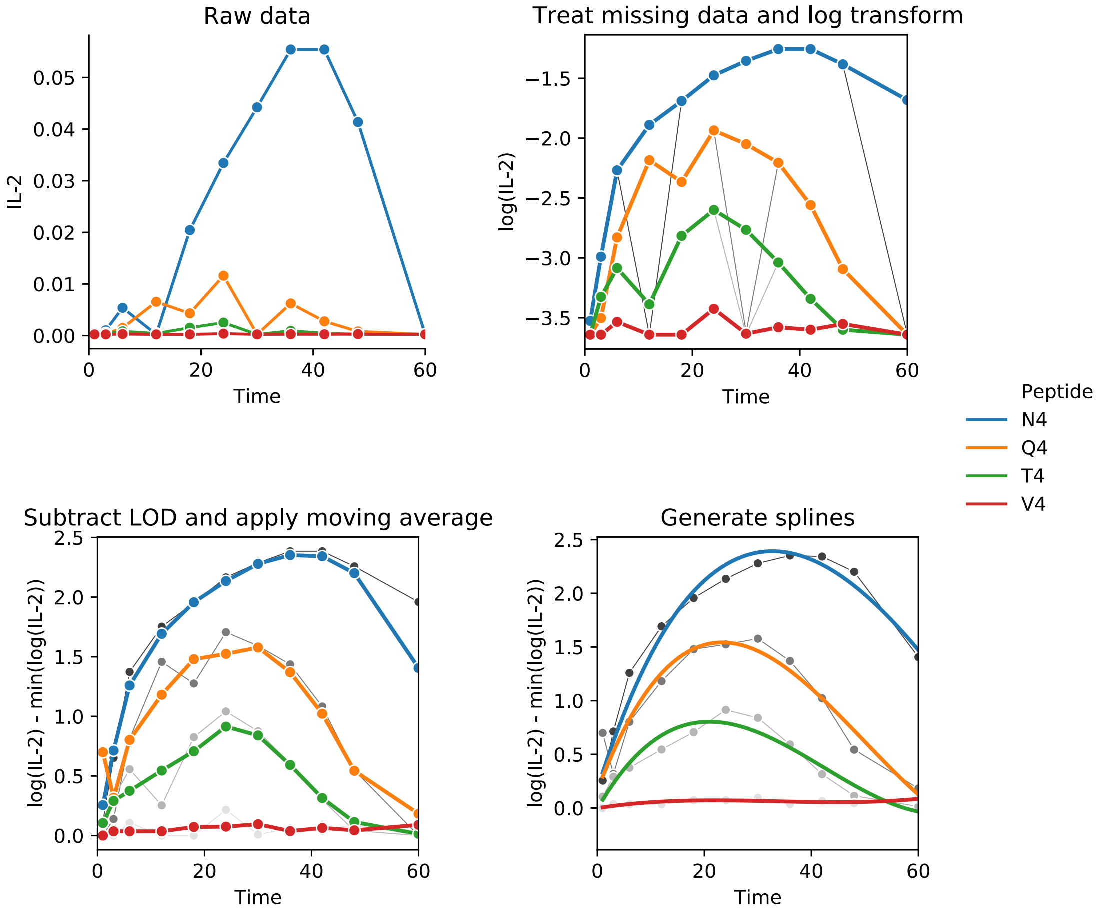


Figure 4.1: **Transforming raw data to splines.** Starting from raw data (top left), we treat missing data points and apply the log transform (top right). Grey lines indicate the raw data log transformed for a comparison on how the missing datapoints are interpolated. Then we subtract the LOD and apply a moving average (bottom left), and compute splines (bottom right).

Having obtained splines, we now have access to cytokine concentrations at every desired point in time. Splines are differentiable and integrable, which means we also have access to the derivatives and integrals of the cytokines. In the main body of the work in [150] is detailed how the evolutionary algorithm φ_{evo} [152] is used to generate plausible biochemical networks that produce an output that differentiates between agonists, partial agonists and nonagonists. The most promising network effectively computes an integral over the product of IL-2 and IL-6. This finding inspired us to also use integrals as our basis functions for classifying antigens. Integrals are more robust to small fluctuations in cytokines, fluctuations that continue to exist despite the smoothing procedure. Moreover, IL-2, the cytokine that is known to correlate with antigen quality [124], is consumed and disappears from the system. With integrals over the cytokine concentrations, at later time is still reflected how much IL-2 was initially present. Moreover, Elispot measures integrated values too by potentially capturing every cytokine that is produced, removing the consumption component from the system. We cannot do that with our measurements, but cytokine integrals, although overcompensating, resemble Elispot measurements more closely than cytokine concentrations do.

Training a classifier

There are multiple ways to design an antigen classifier using cytokines. For instance, one could train a deep neural network (DNN) [9] that takes as an input one large vector of cytokine concentrations at various times, processes this through various layers, and then outputs a category. One could also train a shallow neural network with a small number of inputs allowing for interpretation of the computation that the neural network performs. A main issue concerns the amount of data. Regardless of whether the robotic platform allows for measuring many experimental conditions in parallel, when complete timeseries are used as individual samples, the number of samples would be limited to hundreds, barely enough to merit training a DNN, and not nearly enough to hope for its generalization. Guided by the data limitation and our desire for interpretatibility, we chose to work with an MLP with one hidden layer, where we use every five-dimensional timepoint as a sample. For timepoints from the same timeseries, each timeseries has the same label, namely from the antigen that gave rise to this timeseries. Conveniently, through the splines we have access to an unlimited number of datapoints, although many of these datapoints are strongly correlated. To be explicit, a generalized input $\mathbf{I}(t) \in \mathcal{R}_{+}^{5}$ is given by

$$\mathbf{I}(t) = \int_{t'=0}^{t} \mathbf{c}(t')dt',\, t \in [0, 72]. \tag{4.15}$$

Here the input $\mathbf{I}(t)$ integrates each of the cytokines from $t' = 0$ to $t' = t$. The next datapoint $\mathbf{I}(t+1)$ is a new sample, highly correlated with $\mathbf{I}(t)$ as

$$\mathbf{I}(t+1) = \int_{t'=0}^{t+1} \mathbf{c}(t')dt' = \mathbf{I}(t) + \int_{t'=t}^{t+1} \mathbf{c}(t')dt'. \tag{4.16}$$

This is trivial when written out like this, but requires a shift of thinking. Each of the vectors $\mathbf{I}(t)$ are now equivalent samples. The goal of training the classifier is to uncover a relationship between the integrals of cytokines that provides information on the antigen quality. We assume that this relationship is preserved over time, so that each of the timepoints, highly correlated as they may be, is another example of this relationship in a slightly different part of the input space. With strongly correlated samples, we have to carefully split the data. For instance, we cannot distribute timepoints of the same timeseries across training, validation and test set. The classifier could have "hardcoded" the logic that values of the sample $\mathbf{I}(t)$ in the training set correspond to a given antigen. Then the sample $\mathbf{I}(t+1)$ is an easy guess, because in absolute value it is similar to $\mathbf{I}(t)$. We avoid this by assigning separate experiments with multiple timeseries to each of the datasets.

The last preparatory step before setting up the neural network is the normalization procedure. Because the range in concentration covered by each of the cytokines varies, we want to normalize each of the features to be within the same range. Options for normalization are to make each feature normally distributed (mean 0, variance 1) or set the range of each feature between $[-1, 1]$ or $[0, 1]$. The last option works best for us, as integrals are nonnegative and uniformly distributed over their range. We tried normalizing each dataset by their individual minima and maxima, but found that this introduced a dependence on the experiment (Appendix Fig. C.2). Instead, we take the absolute mininum and maximum per cytokine from the training data and use this to normalize the validation and test sets.

An MLP with one hidden layer goes through two processing steps, which we describe here in detail. The first one is simply linearly transforming the values in the hidden layer $\mathbf{h}(t) \in \mathcal{R}^N$ with N nodes by the learned weights in the matrix $W \in \mathcal{R}^{5\times N}$.

$$\mathbf{h}(t) = \mathbf{I}(t) \cdot W + \mathbf{b}. \tag{4.17}$$

The dependence on time serves as a reminder that we process a whole timeseries by sending timepoints through the MLP one by one, and stitching them back together later. The values in the hidden layer are processed by an activation function (hyperbolic tangent) and multiplied by the second matrix of learned weights $W' \in \mathcal{R}^{N\times M}$ where M is the number of nodes in the output layer, resulting in

$$\mathbf{h}'(t) = \tanh\left(\mathbf{h}(t)\right) \cdot W', \tag{4.18}$$

where $\mathbf{h}'(t) \in \mathcal{R}^M$ are the arguments in the softmax in the output layer. Finally, switching from vector to index notation, the Q^{th} node in the output layer $p(Q,t)$ is computed through a softmax activation function given by

$$p_Q(t) = \frac{e^{h'_Q(t)}}{\sum_{Q=1}^{M} e^{h'_Q(t)}}. \tag{4.19}$$

The softmax ensures that the sum $\sum_{Q=1}^{M} p_Q(t) = 1$. The result following the softmax is taken as a probability distribution, i.e. $p_Q(t)$ is the probability that the cytokine integrals $\mathbf{I}(t)$ are due to the stimulation of T cells with antigens of quality Q.

The next step in setting up the training procedure is to determine the loss function, whose gradient with respect to each of the weights will determine how they are updated. We chose to use the cross-entropy loss, although the mean squared error would have worked too. The cross-entropy loss aims to minimize sum of the distance between the current probability distribution p and the desired probability distribution p' for each of the samples. Here, p' consists of a one for the correct class and zeros for the incorrect classes, while p is initially randomly distributed, and will, throughout the training procedure, imitate p' more closely. For a single sample the cross entropy $H(p', p)$ is given by

$$H(p', p) = -E_{p'}(\log{(p)}) = -\sum_{Q} p'_Q \log{(p_Q)}, \tag{4.20}$$

where the sum runs over the all outputs Q in the output layer. This is ultimately summed over all samples to determine the total loss. However, when predicting the class of $\mathbf{I}(t)$, the node with the highest value $\max_Q (p_Q(t))$ is chosen as the correct label.

We train the classifier using scikit-learn [101] for 3000 iterations with a regularization rate $\lambda = 0.1$ and two nodes in the hidden layer. As a training set, we use hourly sampled datapoints from 78 timeseries distributed over 6 datasets from the old protocol for a total of $78 \times$ 72hrs $= 5616$ datapoints (Table 4.1). For crossvalidation we initially removed timeseries from this dataset and tested on these. Once our collaborators created more datasets with the new protocol, we trained on all conditions on these 6 training sets from the old protocol and tested on the new experiments.

A schematic of the workings of the classifier is shown in Fig. 4.2. As an example, we take integrals of the timecourse for the antigen with quality Q4 and quantities 1μM and 1nM. At every timepoint, $\mathbf{I}(t)$ is classified individually resulting in a $p_Q(t)$. For 1μM the probability of the Q4 node rises until it is the highest before decreasing slightly. From time $t \sim 15$ hours, Q4 is the class with the highest p_Q. Timepoints in this timeseries are for the most part classified correctly. This is in contrast to the timeseries with 1nM where the majority of the timepoints have seems to have a highest $p_Q(t)$ for T4. The final question then is how to turn the classification of 72 hourly sampled timepoints into a single timeseries prediction, which we show in the next section.

In this section, we described how the data is obtained and processed, under what constraints the classifier is designed, and how it works conceptually. In the next section, we report the performance and limitations of the classifier, we analyze the effect of input features and learned weights of the classifier, and work our way towards finding an interpretation for the dynamics of the nodes in the hidden layer, from now on called latent space.

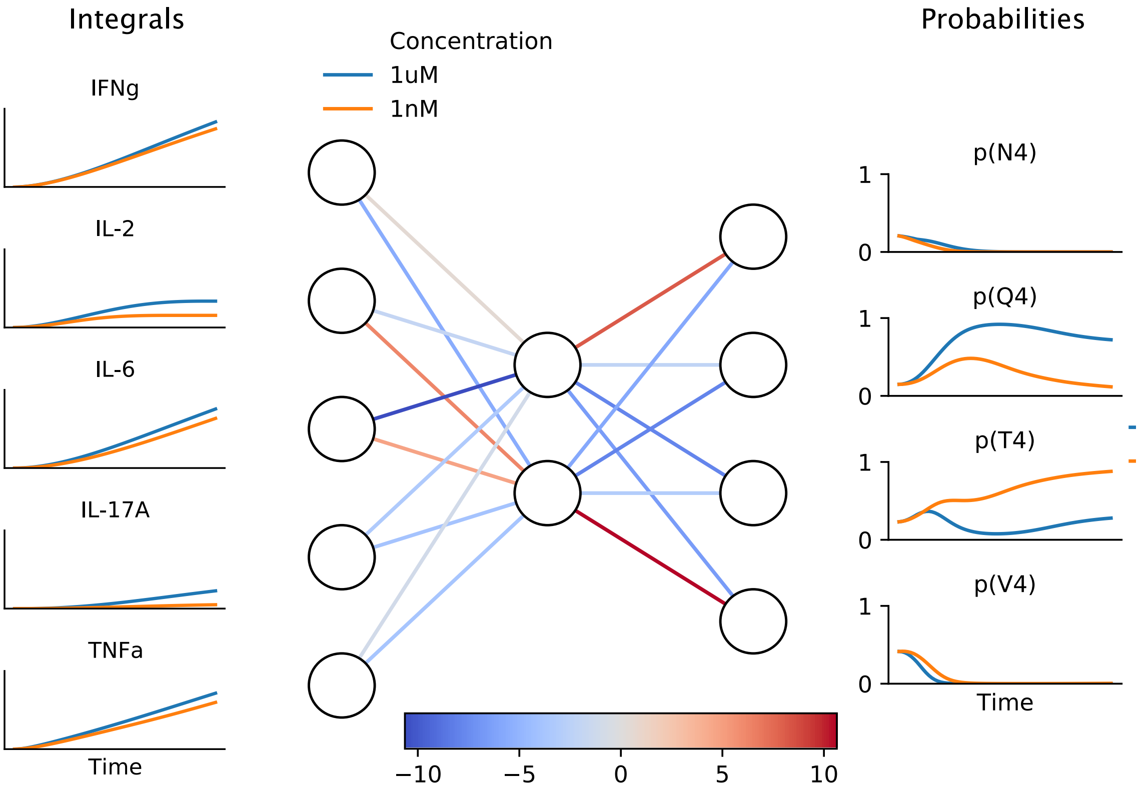


Figure 4.2: **Architecture and processing of the classifier.** Cytokine timeseries for Q4 1μM and 1nM are processed by the MLP with more intense red (blue) lines corresponds to stronger positive (negative) weights. For the integral values at every timepoint, the MLP returns a probability vector corresponding to the antigen class for that timepoint. Stitching the individual probabilities back together results in a continuous probability vector per antigen class over time.

Table 4.1: Number of times a condition was present in the standard training set with six datasets

	1μM	100nM	10nM	1nM
N4	6	5	4	6
Q4	6	5	5	6
T4	6	5	5	6
V4	4	3	3	3

4.3 Results

In the results section, we discuss classifier performance, network analysis, latent space parameterization, interpretation of the model parameters, and prediction of quality.

Classifier performance

The classification procedure of whole timeseries is shown in Fig. 4.3. Individual timepoints of a single timeseries are predicted in the rows (left panel). Circles correspond to correct prediction, crosses to incorrect prediction. The predictions are summed per antigen class and shown as a histogram (right panel). The antigen class with the highest number of points is chosen as the prediction for the timeseries. The vertical line in the histograms indicates the actual antigen class. If the line is green, the prediction is correct, if it is red, the prediction is incorrect. Classification of timeseries from another representative experiment where the same effect is visible is given in Appendix Fig. C.3. With this procedure, depending on the dataset, about 80% of the timeseries are classified accurately, and 20% are not. This directly reflects the limits of precision. That is, N4 1nM is sometimes classified as Q4, Q4 1nM often as T4, and T4 1nM as V4. It is impossible to assign the right quality to these timeseries, because the cytokine response of Q4 1nM and T4 1μM are equivalent (Fig. 4.4, top panel). The reason that the cytokine responses are the same is because the T cells individually measure the same output. Adaptive kinetic proofreading models measure an output that is independent of antigen quantity over many decades [33, 34] (Chapter 2). Here we see in practise that when the antigen quality of Q4 and T4 differs by a factor of 3-5 (Table 4.2), it can be determined over a range of two decades (a factor 100) for antigen quantity.

The T cell population cannot measure quality alone, it always measures a convolution of quality with quantity. Precisely because of the adaptive kinetic proofreading mechanism, it is much more sensitive to antigen quality than it is to antigen quantity. The same quality-quantity convolution appears when inspecting dose-response curves in Fig. 4.21. We want to discretize continuous and overlapping classes, which means that the classifier inevitably cuts off at a point where antigens of higher quality at low quantity will be classified in the lower quality class (Fig. 4.4, bottom panel). Taking into account this constraint means we have achieved maximum possible accuracy with our procedure.

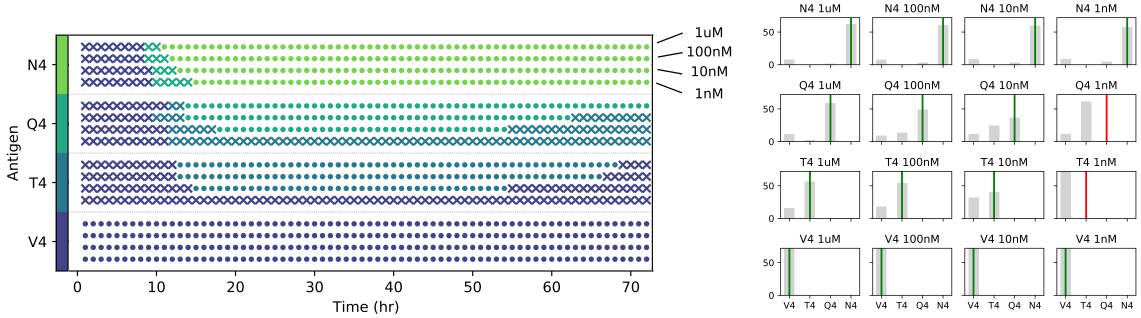


Figure 4.3: **Timeseries classification procedure.** Left panel: classification of timeseries of a given quality (set of four rows indicated by antigen name and color on the left) and quantity (four subrows per antigen quality indicated by antigen quantity on the right). Circle (cross) is correct (incorrect) classification. Color of the marker indicates what antigen was predicted. Right panel: summing individual timepoints per timeseries of given quality (rows) and quantity (columns). The timeseries prediction is the antigen with the most timepoints, indicated by the vertical line. Green (red) line indicates a correct (incorrect) prediction.

There are inherent limits on the reproducibility of the experiments. The training procedure is designed to be robust to some variability by using integrals and learning from experiments with as wide of a distribution as possible. The experiments are performed carefully and repeated when there are obvious errors, yet the absolute cytokine concentrations may vary strongly between experiments. Two experiment were proposed to find the origin of this variability. The first experiment tests the preparation and measurement by running four replicates on T cells from the same mice. We show the IL-2 concentration of the conditions in Fig. 4.5. Except for two deviating conditions of N4 1μM, all replicate conditions follow each other closely (same colored lines on top of each other), which

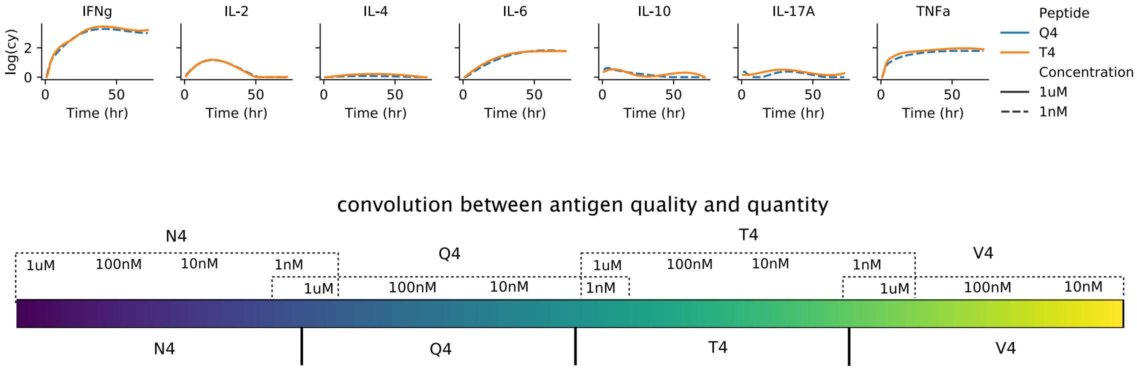


Figure 4.4: **Limits of deconvolution.** Top panel: Comparison between cytokine response of Q4 1nM of T4 1μM. Bottom panel: Schematic discretization of antigen qualities. Text above the colorbar indicates actual division with overlaps at saturating concentration (1uM and 1nM). Text below the colorbar indicates the classifier's cutoff to create four distinct antigen classes, causing timeseries with 1nM to be classified often as the lower antigen class.

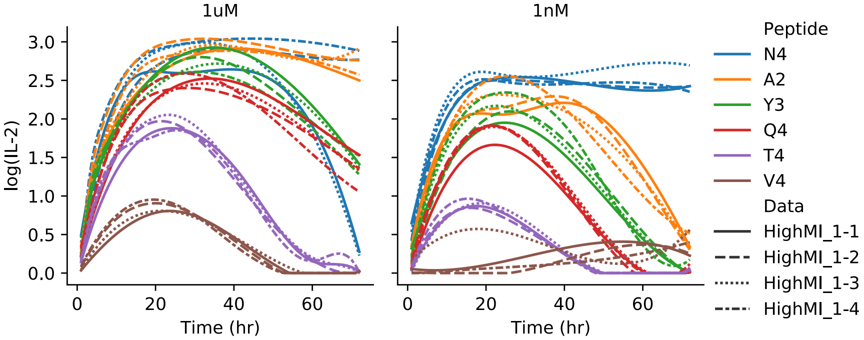


Figure 4.5: **Reproducibility of experiments**. IL-2 concentration of four replicate conditions of the same mice. Colors indicate antigen, linestyle indicates replicate. Left columns show conditions with antigen quantity 1μM, right columns show conditions with antigen quantity 1nM.

means that the variability does not come from measurement errors or variability in cytokine production. Instead, it might come from the T cells themselves. The same type of T cells are used in every experiment, taken from genetically equivalent mice, which were raised under the same circumstances to approximately the same age. This means their phenotype is controlled for as much as possible. But if a mouse starts suffering from a minor infection, some T cells change to a pre-activated state, which makes them respond differently from the entirely naive T cells. In the future, our collaborators will explicitly test this by taking T cells from different mice, using the same sample preparation, and running the same conditions on these.

Network analysis

In this section, we introduce the latent space dynamics, dicuss the computation that the neural network performs, and do a feature analysis.

The latent space is found by multiplying the input with the 5×2 matrix that is visualized in Fig. 4.6, top left, providing an explicit visualization of the colorcoded weights in Fig. 4.2. The x-axis are the input cytokines, and the y-axis shows what value they are multiplied with. The blue weights go into node 1, orange into node 2. These input values are summed like in Eq. 4.17. Projecting all sampled times in a timeseries on the latent space results in Fig. 4.6, right panel. Every line corresponds to a timeseries with antigen and concentration indicated by color and linestyle. The markers are spaced in intervals of 5 hours. All timeseries start in the origin with 0 accumulated cytokines, value 0 in node 1 and 2 and diverge from there. A common feature across conditions is that at a given time, later for antigens of higher quality, the lines curve down. This happens when all IL-2 is consumed and its integral remains constant. IFNγ, IL-6 and TNFα then cancel out IL-17A and node 1 remains constant. Looking more closely at the top left panel of Fig. 4.6, we see that node 1 is dominated by IL-2, and includes small contributions from IL-6 (positive) and TNFα (negative). Node 2, on the other hand, consists of positive contributions of IL-17A and IL-2, offset by smaller negative contributions of IFNγ, IL-6, TNFα. With the IL-2 integral at a constant level, node 2 slowly tends towards negative values.

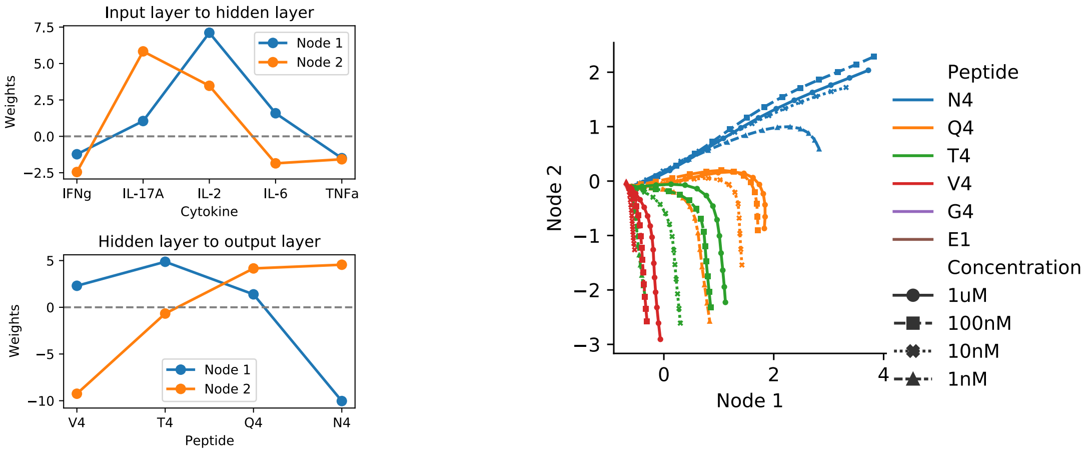


Figure 4.6: **Analysis of weights.** Integral cytokine timeseries multiplied by the weights from input layer to hidden layer (top left) results in latent space (right). Node 1 and node 2 are transformed with the activation function and multiplied by weights from hidden layer to output layer (bottom left) resulting in arguments for the softmax in the output layer.

The weights from the hidden layer to the output layer (Fig. 4.6, bottom left panel) are optimized such that when multiplied with coordinates in the region of antigen Q, they return an argument to the softmax in output layer that is maximum for antigen Q. This is why cytokine dynamics are constrained to certain regions. It also explains why classification at early timepoints does not work well: the initial response of all antigens occurs approximately in the same region. It is only after tens of hours that the dynamics are reliably confined to their respective regions.

Following our first exposure to the latent space, we now proceed with analyzing this thoroughly in the remainder of this chapter. We start with a feature analysis to understand how the latent space is constructed. A standard feature analysis in machine learning is “leave-one-out” cross-validation, where one leaves out an input, retrains the classifier with the same settings, and records performance differences. As we are less interested in absolute classification accuracy, and more in qualitative functions encoded in the latent space, we let our analysis be inspired by building models from the ground up, as we did in Chapter 2 for models of immune recognition. First, we find the unique functions in latent space that combinations of cytokines can encode. Note that the latent space is translational and rotational invariant, as any orientation can be achieved by multiplying the weights W with a rotation matrix and updating W' accordingly. We are thus looking for variations beyond this. Armed with this information, we then construct the rich latent space dynamics from the ground up one cytokine at at time.

To obtain the unique latent space functions, we go through all permutations of cytokines with which we train a new classifier using the standard architecture, hyperparameters and training data. We categorize the resulting latent spaces as shown in Fig. 4.7. In the text in the left margin we name the function and the cytokines required to encode it. Columns show the latent spaces for four datasets with an example of this function, allowing us to assess the robustness of this function across datasets. In the top row, all data has collapsed onto a single line, showing that the information on peptide quality is one-dimensional. This is obviously the case for any cytokine alone, but also for IL-6 and TNFα, and surprisingly, also for IL-2 and IL-17A, encoding quality effectively in 1d. The classifier measures how far the timeseries has progressed, a nonrobust quantity that depends strongly on time and details of the dataset. In the second row, we obtain swirls, encoded by IFNγ and IL-6 or TNFα. This behavior is near one-dimensional towards the end of the timeseries, but two dimensional during the production phase. This is because IFNγ and IL-6/TNFα have a distinct quality dependence during the production phase is different, while in the steady state, the quality dependence on these cytokines is similar. By using cytokine integrals instead of cytokine concentration, inputs at later times have "memorized" what happened at earlier stages, allowing for separation in the latent space. There is still significant variation across datasets, visible in the details of the swirl and the angle of the steady state. In the third row, we obtain straight lines in two dimensions by training the classifier on IL-17A and any of IFNγ, IL-6 or TNFα. For all antigens, the angle with the vertical remains preserved. Differently said, the ratio of Node 1 and Node 2 remains constant over time, requiring a constant quality dependence during the production phase and the steady-state of the cytokines, but a dependence that is different for IL-17A and IFNγ/IL-6/TNFα. IL-17A determine the angle dependence on antigen quality at the start of the timeseries. Spatially, timeseries of the same antigens across datasets occupy more or less the same regions, meaning we have found a representation that is robust to variability across datasets. Finally, in the fourth row, using IL-2 and IFNγ/IL-6/TNFα, we retrieve the base of the typical latent space that we find with the full classifier, containing an initial “rise” followed by a “descent” that occurs at times dependent on quality. This representation uses most of the space, allowing for a spatial division by quality that is insensitive to the dataset and the time of the input, for $t \simeq$ 20 hours.

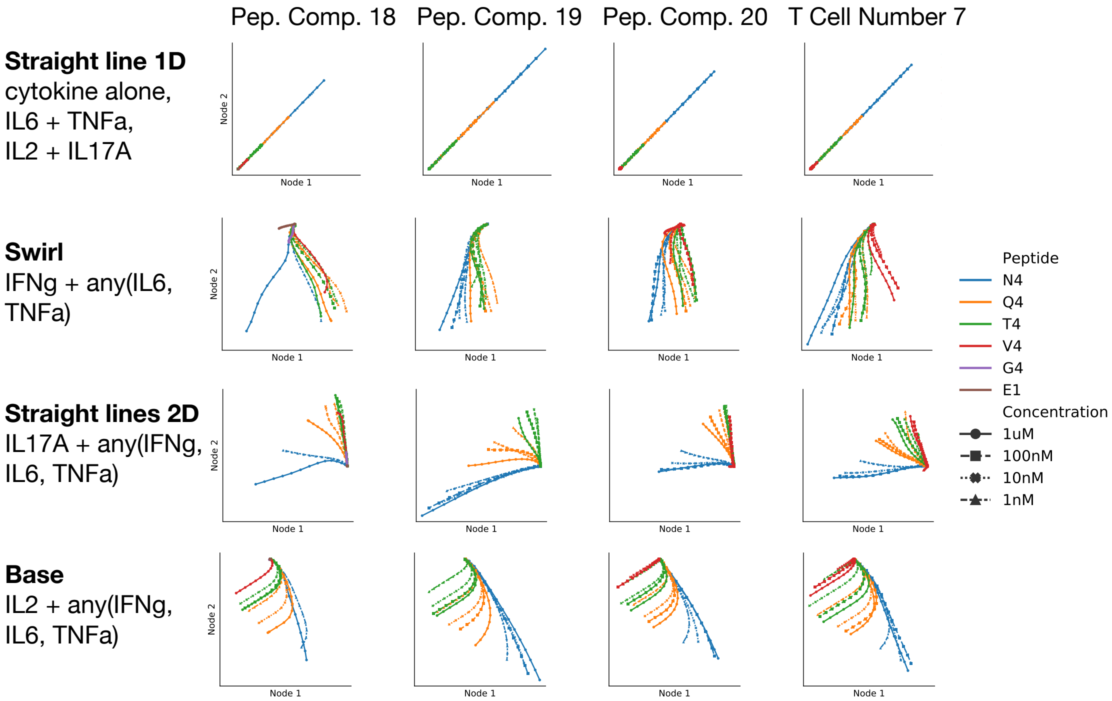
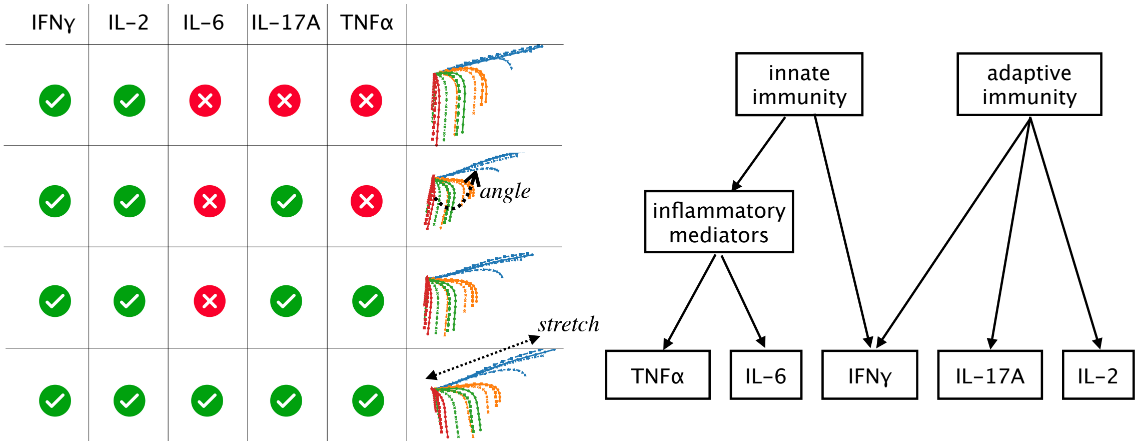


Figure 4.7: **Fundamental functions of cytokines.** Rows show straight lines in 1d, straight lines in 1d with a swirl, straight lines in 2d and base of the standard latent space. Columns show four datasets projected on latent spaces for the following setups from top to bottom: IL-2 alone, IFNγ + TNFα, IFNγ + IL-17A, IL-2 + IL-6. Combinations of input that encode a similar function are indicated in the margin. The classifier was trained using standard architecture and hyperparameters.

The latent space is built up from these four fundamental functions, shown in Fig. 4.8, left panel. Starting with just IL-2, we then add IFNγ, obtaining a 2d representation. Next, we add IL-17A for an early angle dependence on quality. Finally, we stretch the latent space by adding IL-6 and TNFα, slightly enhancing classification accuracy by making the classifier more robust to variability between experiments. Despite the importance of IL-6 [153] and TNFα [154] in the immune response, we learn from the incremental change to the latent space design upon their inclusion that they do not provide much information about encoding quality that other cytokines do not already provide. The causal relationships between the cytokines are hard to establish, and are not unidirectional, but an attempt is made to hierarchically link the cytokines (Fig. 4.8, right panel). The first level roughly distinguishes between cytokines produced through the innate immune response and adaptive immune response. The second level divides the innate immune response into inflammatory mediators TNFα/IL-6 and IFNγ and the adaptive immune response into IFNγ, IL-2 and IL-17A. As the response is two-dimensional, picking a cytokine from either of the two branches provides most of the information on quality contained in the cytokine response. The details matter, and choosing IL-2 or IL-17A makes a difference (Fig. 4.7), as does picking IL-6/TNFα over IFNγ.



IFNγ	IL−2	IL−6	IL−17A	TNFα
✓	✓	✗	✗	✗
✓	✓	✗	✓	✗
✓	✓	✗	✓	✓
✓	✓	✓	✓	✓

Figure 4.8: **Hierarchy of cytokines.** The latent space is built from the ground up by training a classifier on IL-2 and IFNγ, and incrementally including IL-17A, TNFα and IL-6 (left). Cytokines from both innate and adaptive immunity are required to obtain the base 2d latent space. This is explained through a crude hierarchy of cytokines (right). The latent space is given more features by adding IL-17A (angle) and TNFα/IL-6 (stretching), both aid in classification accuracy.

To conclude, we have analyzed what features are present in the cytokines by categorizing latent space functions, and building the latent space from the ground up. This allowed us to draw a hierarchical tree of cytokines, where branching points can be connected to an existing hierarchy of cytokines in the immune system. It is remarkable that we can retrieve these details by measuring only five cytokines. One could imagine that by including more cytokines, this tree can be completed in more detail, providing a visual interpretation of the cytokine response to T cell activation. New cytokines most likely become new leaves to already existing branches, pointing towards deeper levels of fine-graining in T cell activation. Having qualitatively described through what features the latent space is created, we now proceed with quantitatively modelling the dynamics.

Parameterizing latent space

The latent space dynamics of all timeseries have similar features, which we describe now. First, there is an initial rise at an angle correlated with antigen quality. Moreover, the rate of increase in cytokine integrals (=cytokine concentration) seems to increase with antigen quality. When IL-2 consumption exceeds IL-2 production, earlier for antigens of lower quality, a drop in the vertical latent space coordinate is initiated, resulting in a straight line pointing mostly down. The angle with the vertical of this steady state is constant for timeseries within an experiment, but varies between experiments.

These features remind us of ballistic trajectories with an initial propulsion phase followed by a free fall with drag. We get to do rocket science! We study two piecewise models that describe the dynamics well. The first model is the constant velocity model, named after the constant velocity that is held in the propulsion phase, and is described by

$$\mathbf{r}(t) = \begin{cases} \mathbf{v}_0 t & t \leq t_0 \\ \frac{\mathbf{v}_0 - \mathbf{v}_\infty}{2k}\left(1 - e^{-2k(t-t_0)}\right) + \mathbf{v}_\infty (t - t_0) + \mathbf{v}_0 t_0 & t > t_0. \end{cases} \tag{4.21}$$

Here, $\mathbf{r}(t)$ are the latent space coordinates. The constant velocity $\mathbf{v}_0$ can be decomposed into $\mathbf{v}_0 = (v_0 \cos(\theta), v_0 \sin(\theta))^T$ where θ is the angle relative to the vertical and v_0 is the average logarithmic cytokine concentration from $[0, t_0]$. The parameter t_0 determines when to end the propulsion and enter the free fall phase, the drag coefficient k determines how far the forward momentum from the propulsion is being carried over until the terminal velocity $\mathbf{v}_\infty$ is reached. The terminal velocity can be decomposed into $\mathbf{v}_\infty = (v_m, v_t)^T$, where v_m is the velocity of the medium corresponding to the steady state velocity in the x-direction, and v_t is the terminal velocity in the y-direction, which is reached when the drag force equals the gravitational force. Following standard convention, we assumed gravity points down, and the medium moves horizontally. The gravitational constant is implicit in v_t. The second model is the constant force model, taking into account finite acceleration during the propulsion phase. To find the velocity in the propulsion phase, we solve Newton's second law

$$\mathbf{r}''(t) = \mathbf{F} - k\mathbf{r}'(t). \tag{4.22}$$

Here $\mathbf{F}$ is decomposed into $\mathbf{F} = (F\cos\theta, F\sin\theta)^T$ where F is is the constant force per unit of mass and θ is the angle with the vertical, k is the drag coefficient per unit of mass and $\mathbf{r}'(t)$ and $\mathbf{r}''(t)$ are the instantaneous velocity and acceleration in latent space. Solutions for the velocity $\mathbf{r}'(t)$ during the propulsion phase are given by

$$\mathbf{r}'(t) = \frac{\mathbf{F}}{k}\left(1 - e^{-kt}\right) + \mathbf{B} \tag{4.23}$$

from which it follows that

$$\mathbf{r}(t) = \frac{\mathbf{F}}{k^2}\left(kt - \left(1 - e^{-kt}\right)\right) + \mathbf{B}t + \mathbf{C} \tag{4.24}$$

where $\mathbf{B}, \mathbf{C} = 0$ reflect that the timeseries starts in the origin without momentum. The second phase remains unchanged compared to the constant velocity model and is given by

$$\mathbf{r}(t) = \frac{\mathbf{r}'(t_0) - \mathbf{v}_\infty}{2k}\left(1 - e^{-2k(t-t_0)}\right) + \mathbf{v}_\infty(t - t_0) + \mathbf{r}(t_0) \tag{4.25}$$

with $\mathbf{r}'(t_0) = \frac{\mathbf{F}}{k}\left(1 - e^{-kt_0}\right)$ and $\mathbf{r}(t_0) = \frac{\mathbf{F}}{k^2}\left(kt_0 - \left(1 - e^{-kt_0}\right)\right)$. We solve both models in dimensionless time, i.e. $t \to kt$, $t_0 \to kt_0$, $\mathbf{v}_0 \to \mathbf{v}_0/k$, $\mathbf{v}_\infty \to \mathbf{v}_\infty/k$ and $\mathbf{F} \to \mathbf{F}/k^2$, removing the need to fit parameter k.

We settled on the following fitting procedure. First, we set k to a reasonable value and made time dimensionless. We had good results with $k = 1/20\text{hr}^{-1}$. Next, we determined the ratio v_t/v_m by taking the median from the distribution of v_t/v_m coming from all timeseries in the experiment, which is immediately available from the latent space. By taking the median, we remove the effect that outliers have on the mean. We use this ratio twice.

Once to determine v_m as a function of v_t, and a second time to offset the vertical axis that sets $\theta = 0$. We now have four parameters that fix the model: the scalars v_0 or F, the angle θ, the time t_0 and the scalar v_t. We then proceed with fitting the equations to the integrals of node 1 and node 2 separately using SciPy's curve_fit method [155]. We add a regularization term to the loss function penalizing the absolute value of the parameters and nudging them towards zero in case they are undetermined. This is of importance for timeseries with low antigen quality where t_0 is small, causing v_0 or F and θ to be undetermined, as well as for timeseries with high antigen quality where t_0 approaches the experimental times, causing v_t to be undetermined. In Appendix Fig. C.4 we show how regularization helps in constraining parameters.

A comparison of the two models is given in Fig. 4.9. In the integral latent space (left panel) are slight discrepancies, for instance, in the constant velocity model, the initial rise is overestimated, while around the transition from the propulsion phase to the ballistic phase, the model compensates by underestimating the integrals. This is a subtle effect in latent space coordinates. For a better comparison, we look at the derivatives of node 1 and node 2 over time (center and right panel). We now see much better that the constant velocity model does not take into account the bounded exponential in the propulsion phase, approximating it by a straight line. The constant force captures the exponential rise well, but around the transition overcompensates to accommodate the sharp transition.

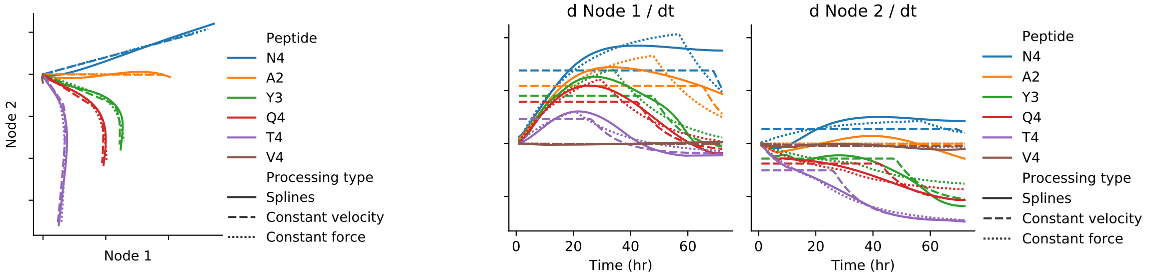


Figure 4.9: **Comparison of ballistic models.** Integral latent space (left panel) and concentration latent space for node 1 (center panel) and node 2 (right panel) over time. Solid linestyle are from splines, dashed lines are from constant velocity model and dotted lines are from constant force model

A natural extension of the constant force model is to introduce an interpolation parameter λ that smoothes the piecewise functions into a continuous function, for instance, by weighting the propulsion phase by $\sigma(t) = \frac{1}{1+e^{-\lambda(t-t_0)}}$ and the ballistic phase by $\sigma(t) = \frac{1}{1+e^{-\lambda(t_0-t)}}$. An immediate concern that arises from this definition is that the ballistic phase strongly decreases for $t < t_0$, while the propulsion phase remains constant for $t > t_0$. Because of this asymmetry, a simple interpolation of the entire function might smooth the curve, but changes its shape too strongly. To circumvent this, one could smooth both phases with the latent space coordinates $\mathbf{r}(t_0)$ at the transition as $\frac{\mathbf{r}(t_0)}{1+e^{\pm(t-t_0))}}$. We do not further discuss this, because the constant force model works well enough for our purposes, failing only near t_0 for the derivative of the fitted function. Note that the derivatives of the latent space coordinates are equal to the latent space representation of the cytokine concentrations

$$\frac{d\,\mathbf{h}(t)}{dt} = \frac{d\,\mathbf{I}(t)\cdot W}{dt} = \frac{d\,\mathbf{I}(t)}{dt}\cdot W = \mathbf{c}(t)\cdot W. \tag{4.26}$$

When one desires a more accurate model for the derivatives of the latent space representation, for instance, to recover the cytokine concentrations, it is worthwhile to consider an interpolating scheme that smoothes the discontinuous transition of the constant force model. François Bourassa worked on a more detailed model that also better takes into account node 2 dynamics [3].

Having fit the latent space data to our satisfaction, we now proceed with discussing the parameter estimates. In the next section we provide an extensive discussion on the interpretation of the parameters, but first we want to understand why in certain regimes some parameters are not well determined. Fig. 4.10 shows two 2d parameter spaces (v_0 or F vs. t_0 and θ vs. v_t) for the constant velocity model (left panels) and the constant force model (right panels). Quality correlates with v_0, F, t_0 and θ, and after correcting for quality, with quantity too. The parameters v_0 / F, t_0 and θ correlate too, meaning that a high F implies a high t_0 and θ. Error bars show one standard deviation on the parameter estimates. In the constant velocity model, error bars on t_0 are especially high for antigens of high quality (N4, A2, Y3), as t_0 approaches experimental times. Moreover, in both models v_t is not well determined for antigens with high t_0 (A2, Y3, Q4), because little time is spent in the free fall phase. v_t has a narrow uncertainty interval for N4, because of regularization pushing the parameter values down, and t_0 being large enough that the value of v_t does not affect the parameter fit. Finally, in the constant force model, parameters F, t_0 and θ are

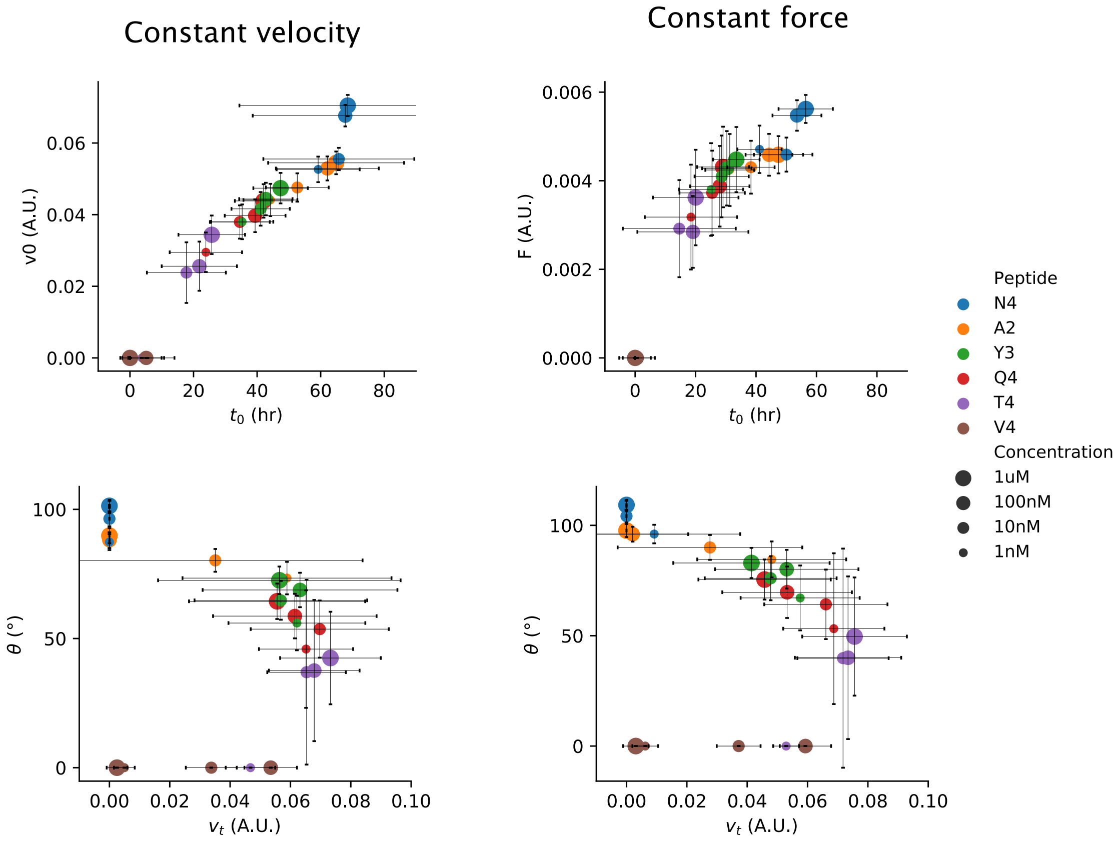


Figure 4.10: **Parameter space for ballistic model.** Parameters for the constant velocity model (left) and constant force model (right) against one another: Magnitude (F or v_0) vs. time to transition between propulsion and free fall phase t_0 (top panels) and angle θ vs. v_t (bottom panels). Error bars are one standard deviation coming from the covariance of the fit.

determined better for antigens of higher quality than for antigens of intermediate quality, likely because a small discrepancy in F or θ is amplified over a longer production phase. These parameters are very well determined for low quality antigens, because regularization sets them to zero. There thus seems to be a trade-off that in the higher quality regime, the model is very well able to capture the timeseries, while in the other regimes some features of the model are redundant. In these regimes, some of the parameters are "sloppy", the central concept in Chapter 2 that allowed us to do parameter reduction. We can find important ("stiff") and unimportant ("sloppy") parameters by inspecting the eigenvectors of the Fisher Information Matrix (FIM) [50, 156, 157]. The FIM can be considered a metric $g_{\mu\nu}$ tracking the curvature of parameter space. It is found by computing second derivatives of the log-likelihood $\log(P(\vec{r},\vec{\theta}))$ with respect to the parameters θ_μ, θ_ν

$$g_{\mu\nu} = \frac{\partial^2 \log\left(P(\vec{r},\vec{\theta})\right)}{\partial\theta_\mu \partial\theta_\nu}. \tag{4.27}$$

Here, $P(\vec{r},\vec{\theta}))$ is the probability that the parameters $\vec{\theta}$ are the best fit given the Gaussian-distributed residuals $\vec{r} = \{r_i\}$, $r_i = \hat{y}_i - y_i(\vec{\theta})$, where $\hat{y}_i$ is a data point, $y_i(\vec{\theta})$ is the corresponding model prediction, and i runs over all datapoints. This is equal to the square of the Jacobian of the model prediction [157], independent of the residuals

$$g_{\mu\nu} = \frac{\partial y(\vec{\theta})}{\partial\theta_\mu}\frac{\partial y(\vec{\theta})}{\partial\theta_\nu}. \tag{4.28}$$

The eigenvectors of $g_{\mu\nu}$ corresponding to the ranked eigenvalues are orthogonal directions in parameter space to which the model is decreasingly sensitive. The square root of the eigenvalues corresponds to the curvature in the direction of the eigenvector. Local changes in parameters along stiff directions have a strong effect on the model outcome, while local changes in parameters along sloppy directions have little effect on the model outcome.

We compute the eigenvalue spectrum of the FIM with Eq. 4.28 for each timeseries separately. The Jacobian is the square root of the Hessian matrix, which we find from the inverse of the covariance matrix, returned by SciPy's curve_fit [155]. The eigenvalue spectrum of the constant velocity model is given in Fig. 4.11, top panel. We plot the eigenvalues relative to the largest one per timeserie following Machta et al. [50]. Eigenvalues can become arbitrarily small; those that are not plotted are smaller than $5 \cdot 10^{-6}$ and can be considered zero, because we reach numerical precision computing the inverse of the covariance matrix. When inspecting the eigenvalue spectrum of the FIM, we first look for a hierarchy of eigenvalues, which is clearly present here, although the details depend on the antigen quality. We note that in the low-quality and high-quality regime, there are one or more eigenvectors with eigenvalue 0, which means that parameters of the corresponding eigenvector cannot be determined. In the low-quality regime (V4) we find a single most important eigenvector, meaning that this regime can be described well with only one effective parameter; in the other timeseries the first two eigenvectors have close to the same eigenvalue, which means the regime is described by two effective parameters. Finally, in the intermediate regime, three, or even four effective parameters are required to describe the data well using the constant velocity model. Seen across a range of conditions, all parameters are of importance, which means that our model is a parsimonious description of the data.

In the bottom panel of Fig. 4.11 we show the squared eigenvector components ranked by eigenvalue. The eigenvectors are orthonormal, so their components sum to one when squared. As such, the eigenvectors can be visualized and compared easily this way, although we lose information on the sign of the components. Rows show the eigenvectors corresponding to the eigenvalue of the timeseries denoted on the x-axis. We find that for antigens N4, A2, Y3, and Q4 parameters θ and v_0 are most important. t_0 is of no importance to N4 and A2 1μM and 100nM, because t_0 is larger than the experimental time of the system. Changing it slightly does not affect the model prediction. Similarly, v_t does not influence these timeseries, which is indeed the second of the two unimportant parameters. Timeseries with antigen T4 are at a turning point where the presence of v_0 in the most important eigenvector decreases with quantity, and is taken over by v_t. It also has the strongest mixed eigenvectors, which means there is no clear hierarchy in individual parameters in terms of importance on fitting the model. Mixed eigenvectors can provide hints on how to reduce the model into one with actual effective parameters. We would then have to inspect the signs of the eigenvector components, and do the parameter reduction, which we do not get into here.

Fig. 4.12 shows the eigenvalue spectrum (top panel) and eigenvector decomposition (bottom panel) for the constant force model. We again find that we need at least two effective parameters to fit the data for most timeseries, although all parameters matter. The exact make-up of the important parameters differs with antigen quality and quantity. There are modest differences between Fig. 4.12 and Fig. 4.11, most importantly that t_0 is well defined in the high quality regime, because it is smaller than the experimental time. This means that in the constant force model, except for the low-quality regime, all parameters

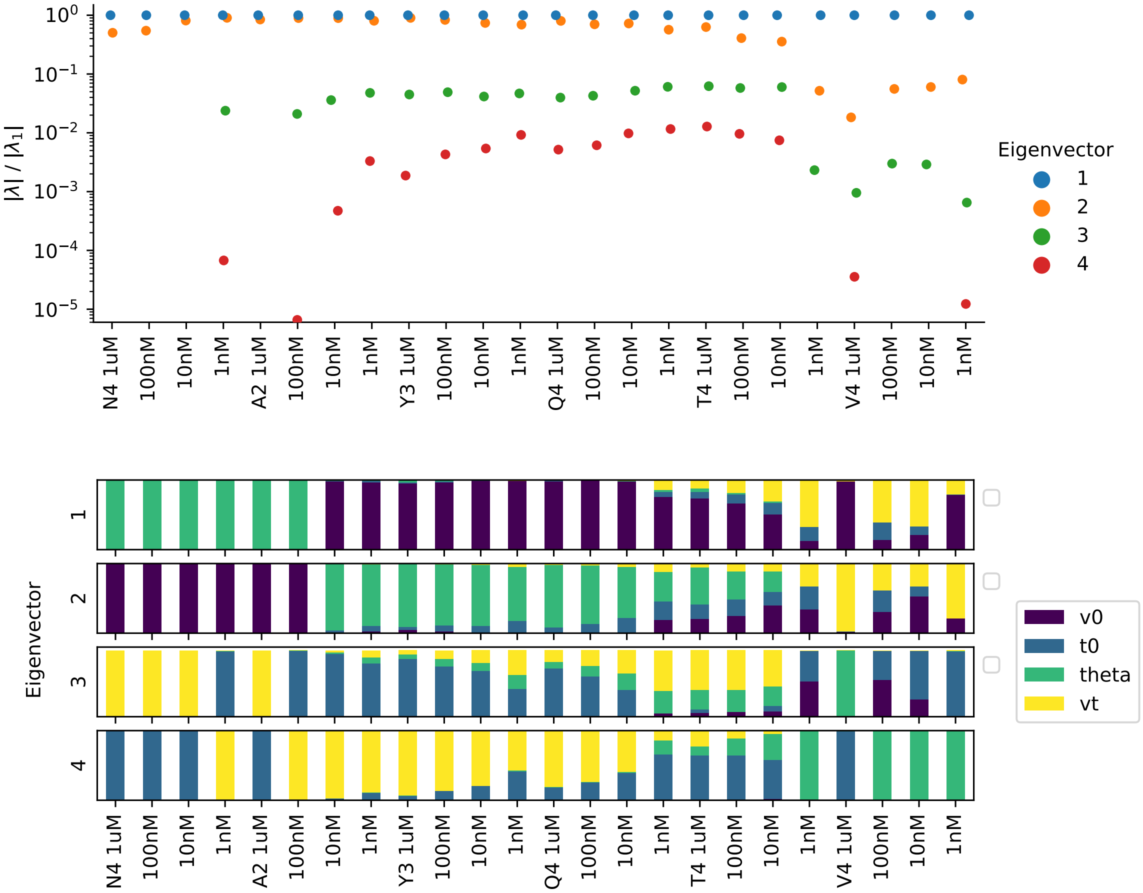


Figure 4.11: **Eigenvector decomposition of FIM for constant velocity model.** Eigenvalue spectrum (top panel) and squared eigenvector components (bottom panel). Eigenvalues are scaled to the largest eigenvalue per condition. The eigenvectors are orthonormal, so that their squared components sum to one.

are well-determined. From the eigenvector decomposition, we learn that for all antigens except V4 θ is the most important and F the second most important parameter. In the constant velocity model, θ and v_0 switched importance A2 100nM for no particular reason. In general, the eigenvalue decomposition of the FIM for the constant force model is smoother than the one for the constant velocity model. Besides fitting the concentration latent space better, these are more factors in favor of the constant force model, which why moving forward, we mainly discuss the constant force model. There is one experimental configuration where one could argue that the constant velocity model fits the data better, which is where we discuss both ballistic models.

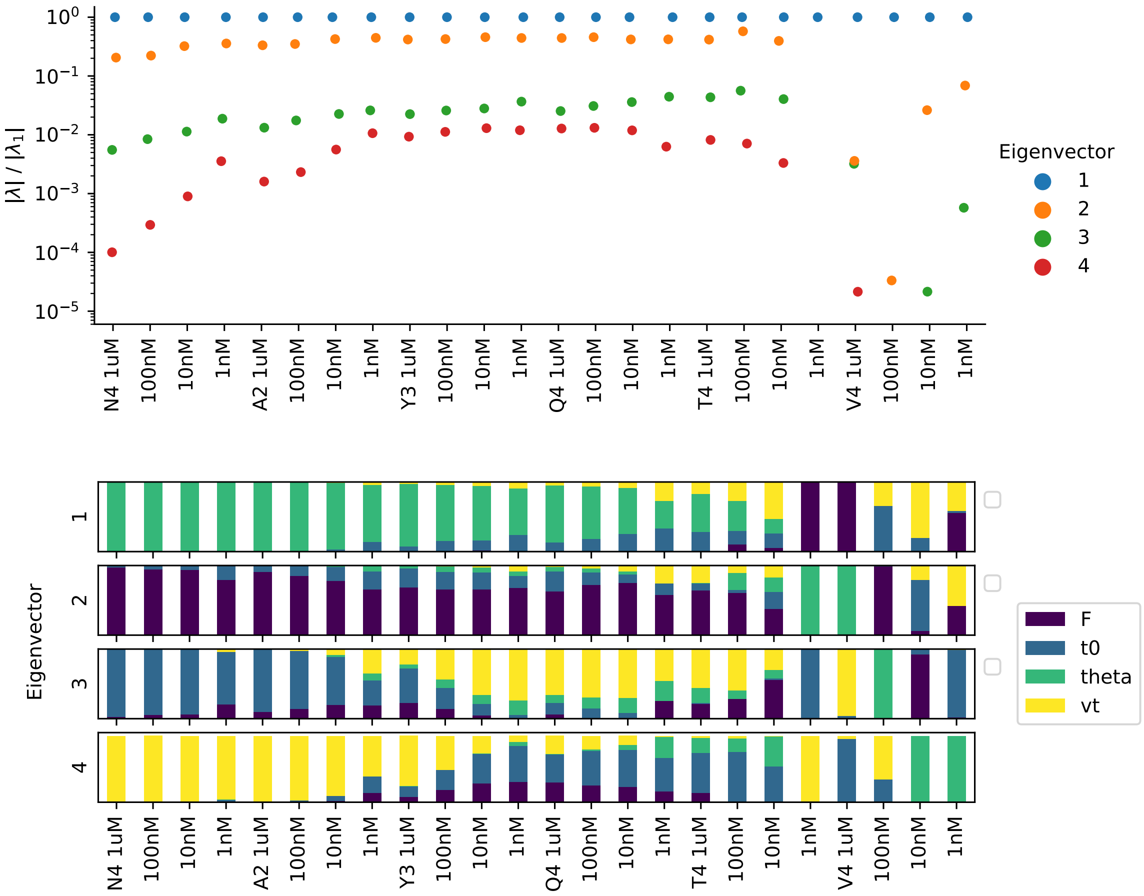


Figure 4.12: **Eigenvector decomposition of FIM for constant force model.** Eigenvalue spectrum (top panel) and squared eigenvector components (bottom panel). Eigenvalues are scaled to the largest eigenvalue per condition. The eigenvectors are orthonormal, so that their squared components sum to one.

The main takeaway from the FIM analysis is that the ballistic models and associated parameters can confidently be used for interpreting a range of experiments in the next setups. We have formalized that t_0 is better determined in the constant force model, and that both ballistic models contain parameters that can be hierarchically ordered in importance, and lack obviously redundant parameters. In short, this just shows that our approach for modelling the latent space is sound, and that we may proceed with relating these parameters to immunology.

Interpreting parameters of immune response

In this section, we study the behavior of the parameters of the ballistic models for various experimental setups. This serves as an opportunity to discuss the biological interpretation of these parameters and to explore the generality of the latent space and the robustness of the ballistic model. The parts that we can change in the experimental setup are antigens, T cells and APCs. A first test of robustness is to reduce the precursor frequency so that fewer cytokines are produced. The second test is to change the antigens from OVA antigens specific to OT-1 T cells to e.g. LCMV antigens specific to P14 T cells, as well as others. We also change the T cell type from CD8+ T cells to CD4+ T cells and we consider pre-activated T cells called T cell blasts instead of naive T cells. We also change the APCs delivering the antigens from splenocytes to macrophages, dendritic cells and tumors. Finally, we consider drug-experiments to directly connect biology to model parameters.

Precursor frequency

The first experimental configuration concerns precursor frequency. Experiments have been prepared with $\{3 \cdot 10^3, 10^4, 3 \cdot 10^4, 10^5\}$ T cells, where 10^5 cells is the default. Immune responses with a smaller precursor frequency take longer to mount to give the T cell population a chance to expand sufficiently. The relationship between the fold expansion and the precursor frequency follows a power-law with factor $-\frac{1}{2}$ [135], as discussed in 4.1, meaning that if the precursor frequency is 100 times smaller, the fold expansion is 10 times larger, and the final T cell population will be a factor 10 smaller. Naturally, it takes longer for the population with a smaller precursor frequency to reach the time at which proliferation ceases. We thus expect a larger t_0 that we may not be able to estimate well for antigens of higher quality within the fixed experimental time of 72 hours. In Fig. 4.13 we show the typical latent space dynamics (left panel) and associated F, t_0 parameter space (right panel). Dynamics for conditions with a lower precursor frequency do not have the range of conditions with similar antigen quality and quantity but larger precursor frequency. It is fascinating how similar the dynamics are for conditions with the same quality (linecolor) but different precursor frequency (linestyle), especially, because the latent space was optimized for classifying antigen quality independent of antigen quantity at fixed precursor frequency. Inspecting the parameter space, we note that with decreasing precursor frequency, t_0 consistently increases while F decreases.

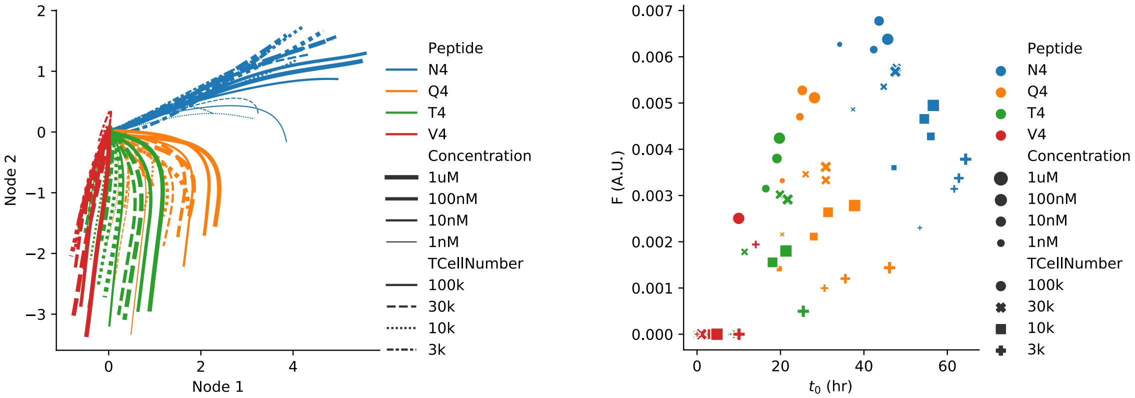


Figure 4.13: **Precursor frequency experiment.** Data projected in latent space (left) and F, t_0 plot of the parameter fit of the ballistic model (right).

The relationship between F, t_0 and precursor frequency makes us wonder if we could relate our model to an existing model of T cell expansion [138]. Similarities are that both models are biphasic and contain a characteristic time (t_0 in our case, t^* for Mayer et al.) at which the expansion switches from exponential proliferation to a slow decay. The main difference is that our model describes an in-vitro immune response, while Mayer's model describes an in-vivo immune response. We used the number of events that the flow cytometer records as a proxy for T cell numbers and fitted the data given in Appendix Fig. C.5 with Eq. 4.4 using the precursor frequency $T(0)$, quantity $C(0)$ and quality K as initial values. 1μM and 100nM correspond to a saturating number of antigens $C(0) = 10^5$ when the APCs are fully loaded, and decreases linearly after that. The antigen with the highest quality N4 has $K \approx 10^{-11} \sim 0.1$ in absolute number. EC_{50} or K for the other antigens follow from [149]. During the staining and washing about 90% of the T cells are lost; we account for this by normalizing the whole timeseries by a factor such that the number of counts at the first timepoint (1hr) corresponds to $T(0)$, preserving relative proportions. We fit the data using SciPy's least_squares method, and find $\alpha = 0.85$, $\delta = 0$, $\mu = 0$. We gathered that our experiments do not run for long enough to observe and accurately estimate the rate of antigen consumption μ and T cell decay δ, neither of which is actually zero. This does not come as a surprise, inspecting the data. Indeed, only for some conditions does the last timepoint at $t = 72hr$ shows sign of a decay in T cell number. Although we find that T cells with smaller precursor frequency proliferate for longer, and that V4 slows down proliferation noticeably earlier than N4, A2 or Y3, the antigen quantity dependence is minimal. With this being the driving force of the model of Eq. 4.4, we did not continue the analysis. We conclude that t_0 and t^* are timescales for different processes. Most likely, t_0 is a timescale related to the upregulation of IL-2 receptors consuming IL-2 [114, 124]. Given the complexity of this model, we have decided to not further pursue this path, but accept the correspondence as is. Drawing a rigorous connection between the ballistic models and [114, 124] remains an open problem.

T cell type

The second experimental configuration concerns changing T cell type. There are many aspects of the naive OT-1 CD8+ transgenic T cells that can be changed. We start with using preactivated T cells or T cell blasts instead of naive T cells. Naive T cells are mixed with beads coupled to anti-CD3 and anti-CD28 antibodies to obtain preactivated T cells. The CD3 antibodies bind to the TCR and the CD28 antibodies bind to the CD28 receptor, providing primary and costimulatory signals to the T cells, effectively activating them. After two days the antibodies are washed off and human IL-2 is added to the culture so that T cells can start expanding for a total of four days. Finally, the dead cells are removed, and the experiment is started the usual way by mixing the T cell blasts with antigens loaded on APCs. We also obtain T cell blasts by activating them with Concanavalin A (ConA). ConA is a lectin, a carbohydrate binding protein, which binds and crosslinks TCRs, activating T cells that way. Creating T cell blasts is an art, as overstimulation with IL-2 might cause T cells to no longer produce IL-2 when re-stimulated, potentially due to the elusive exhausted T cell type [158].

In Fig. 4.14, we compare the integral latent space dynamics (top panels) and model parameters for the constant force model (center panels) for naive and preactivated T cells.

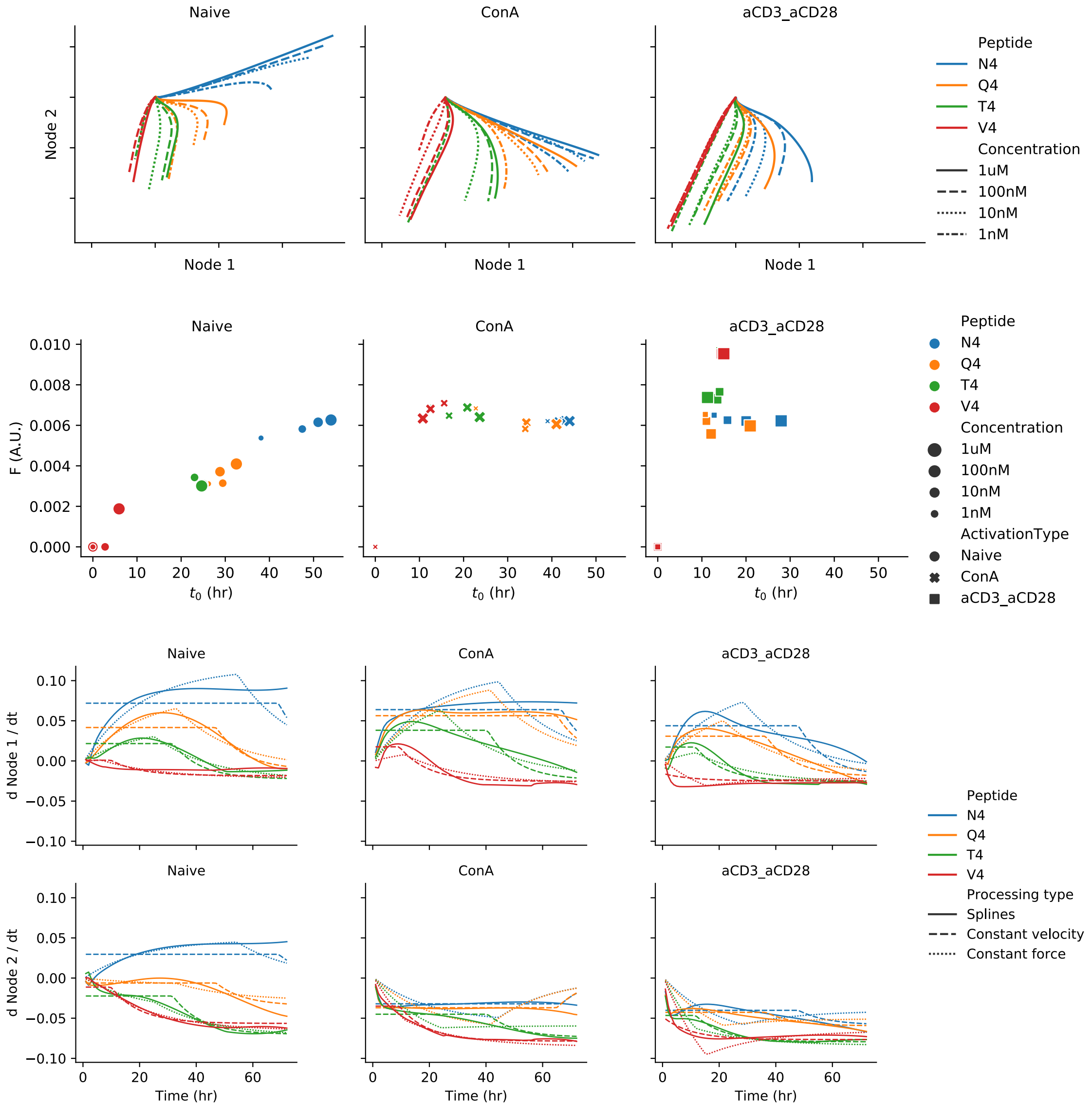


Figure 4.14: **Activation experiment.** Data projected in latent space (top), derivatives of node 1 and node 2 over time compared with fit of constant velocity and constant force model (center) for timeseries with quantity 1μM, and F, t_0 plot of the parameter fit of the ballistic model (bottom). Columns show from left to right experiments with naive, ConA activated and CD3/CD28-antibody activated cells.

Curves from the latter T cell type are turned clockwise in the direction of decreasing θ. The difference in response for the stronger antigen qualities (N4, Q4 and T4) has shrunk significantly, especially in the early response where θ is visually the same for N4, Q4 and T4. Indeed, F is constant (ConA) and near-constant (aCD3/aCD28) for all four qualities. The cytokine response for T cell blasts, the in-vitro equivalent of memory cells, is in accordance with the immunological paradigm of the existence of only three antigen types: non-agonists, partial agonist and agonists [25]. For naive cells there is no agreement, as there exists a continuum of cytokines response to antigens of varying quality.

For T cell blasts, we also compare fits for the constant velocity model and the constant force model. The biology is sufficiently different in this experimental configuration that the constant force model is not by default superior. For instance, T cell blasts ramp up cytokine production much faster than naive T cells due to enhanced chromatin accessibility [159]. This pronounced difference is visible in concentration latent space (Fig. 4.14 bottom panels). First, let us consider the derivatives of node 1 (top) and node 2 (bottom) for the naive cells. Modulo the sharp transition between propulsion and free-fall phase, the constant force model captures the dynamics very well, only underestimating t_0 for N4. The constant velocity model approximates the rise with a constant velocity in integral latent space coordinates over the distance travelled until time t_0, so a straight line in concentration latent space. This does no justice to the intricate dynamics at the start and end of the cytokine production phase, and generally overestimates t_0. Moving on to the concentration latent space coordinates of ConA and aCD3/aCD28 activated cells, we note that the ballistic models fail most obviously by assuming a discrete transition from production to consumption phase. If the transition is continuous, the estimate for t_0 is bound to be inaccurate. The ballistic models capture the initial production phase differently. The constant velocity model still starts off too high, but less so than for naive cells, which makes this model a better candidate for describing the cytokine response of preactivated cells. Yet, it preserves the dependence between v_0 and t_0 for different antigen qualities, which is in stark contrast to the actual response, where the first 10 hours of the cytokine production phase are very similar for antigens of higher quality, but the consumption phase starts at different times. The constant force model cannot capture the strong initial production phase, but assigns equal F to each of the antigen quantities, preserving the relative dependence. Finally, t_0 changes from a dependent measure of antigen strength (correlated with θ and F) to an independent measure of antigen strength for preactivated cells, a measure that may not correspond precisely to the transition time between two phases, but is correlated with antigen quality. For these reasons, the constant force model still is our preferred ballistic model.

The next aspect of the T cells we change is the TCR and corresponding antigens. We have access to a range of antigens of varying quality specific to transgenic P14, PMEL and HY T cells. Top panels in Fig. 4.15 show the latent spaces, bottom panels show the F, t_0 parameter spaces. It is worth reiterating that the integral latent space dynamics are found by training on cytokine integrals from the response of OT-1 T cells to SIINFEKL antigens. The cytokine response is messy for unknown experimental reasons, which is reflected in the integral latent space. We expect a future experiment with new supernatant and new mice to give cytokine responses and integral latent spaces similar to previous experiments. Despite that the curves in integral latent space do not precisely look like ballistic curves, the fitted parameters F and t_0 exhibit the hierarchy of antigen qualities for all four TCRs. It must be said that the specificity of the P14 and PMEL antigens is binary, while we find more of a continuum for OT-1 and HY. This point towards an inherent TCR property: certain transgenic T cells may respond in a more binary fashion while others may exhibit a continuum of antigen qualities.

The final part in changing T cell type is substituting CD8+ T cells with CD4+ T cells. As mentioned before, CD4+ T cells are helper T cells whose effector functions include regrouping other T cell types to the site of infection, while CD8+ T cells, the killer T cells, do the majority of the killing. The experiments are done with primed naive CD4+ T cells from 5CC7 transgenic mice, using antigens that are specific to this TCR. The integral latent space dynamics and corresponding F, t_0 parameter space are given in Fig. 4.16 (top rows). This again serves as an important validation that the latent space we have found from the cytokine response of CD8+ T cells is also informative of the cytokine response of CD4+ T cells. We find a similar hierarchy in antigen quality, but for CD4+ T cells, a much stronger dependence on quantity exists, corresponding with the scaling law of Eq. 4.9 [124], and recapitulated with the parameters of the constant force model. Interestingly,

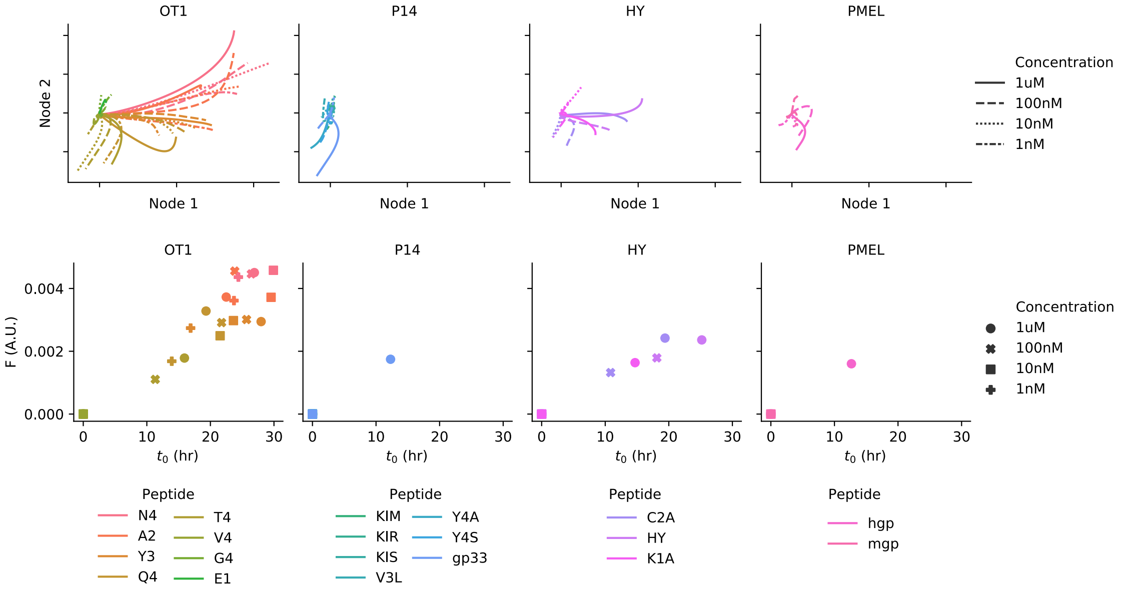


Figure 4.15: **TCR experiment.** Data projected in latent space (top) and F, t_0 plot of the parameter fit of the ballistic model (bottom). Panels from left to right show experiments with OT-1, P14, HY, PMEL T cells. Invisible markers in the bottom panel are plotted on top of each other at $(F, t_0) = (0, 0)$.

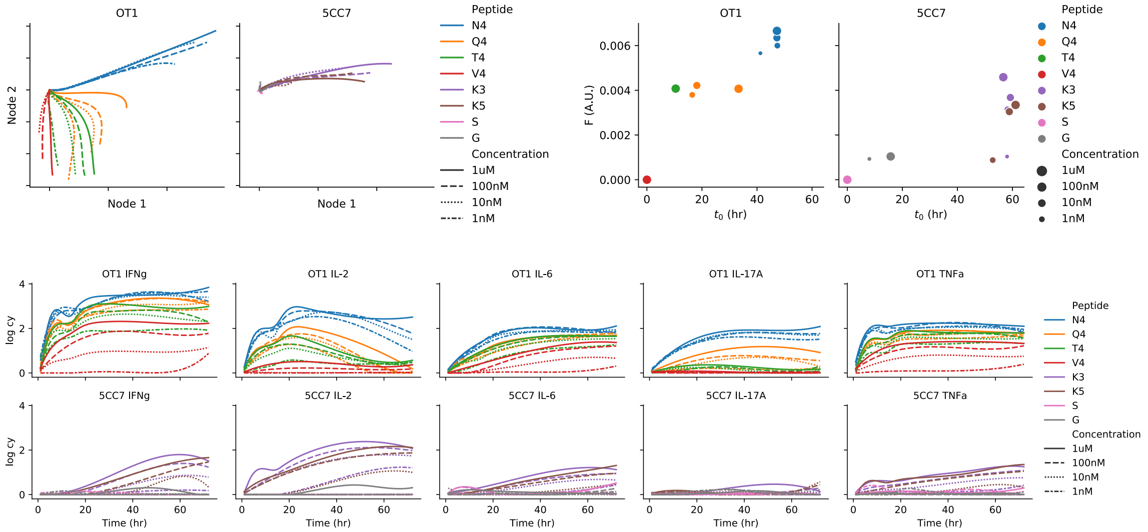


Figure 4.16: **CD-4 experiment.** Data projected in latent space (top left panels) and F, t_0 plot of the parameter fit of the ballistic model (top right panels). Left subpanels show experiments with CD8+ T cells (OT-1), right subpanels show CD4+ T cells (5CC7). Bottom panels show cytokine responses for CD8+ T cells (top) and CD4+ T cells (bottom).

with lower antigen quantity, t_0 remains constant while F decreases, which is due to t_0 being saturated and in reality larger than the experimental time. The curves in integral latent space do not descend due to the lack of IL-2 consumption (Fig. 4.16, bottom panels). The ballistic models are thus overparameterized for experiments with CD4+ T cells, as t_0 and v_t are indetermined. Yet, F alone suffices to capture the differences in strength of response. The cytokine responses between CD4+ T cells and CD8+ T cells (bottom panel) are quite different. For instance, CD4+ T cells do not produce IL-17A at all, causing the lack of variation in θ. Despite the differences we are able to fit the curves with the ballistic models and interpret the parameters, which is testament to the generality of ballistic models in fitting the cytokine response.

APC type

Next, we show results for experiments where we change the antigen presenting cells from splenocytes to macrophages, dendritic cells (DCs) and tumors. We discuss macrophages and DCs first. DCs link the innate and adaptive immune system through their effector function of uptaking antigens from the environment and presenting them to T cells [160]. Macrophages do this too, but their main function concerns the phagocytosis of pathogens, not the presentations of antigens [2] Finally, splenocytes are immune cells harvested from the spleen. The splenocytes used in the experiments have not been characterized at the time of writing, but it is known that the spleen contains mainly B cells, immune cells whose main effector function is to produce antibodies. B cells also present antigens, though not as efficiently as professional APCs like macrophages and DCs.

We compare the immune response to OT-1 antigens enhanced with three adjuvants to the response without adjuvants. A schematic for the experiment is given in Fig. 4.17. An adjuvant is an immunological agent that is designed to aid (*adjuvare*, latin) the immune response. The adjuvants concern lipopolysaccharide (LPS) + human IFNγ (hIFNγ) targeting Toll-like receptor 4 (TLR4) and polyinosinic : polycytidylic acid (poly I:C) targeting TLR3.

[2] For his discovery of DCs and its role in the adaptive immune system the 2011 Nobel Prize in Physiology and Medicine was awarded to Ralph Steinman, the third Nobel Prize that was awarded posthumously since he passed away three days before the announcement, unbeknownst to the Nobel Prize committee.

The last adjuvant is CD28 antibody (aCD28) targeting the costimulatory receptor CD28. TLRs are an essential part of the innate immune system and are found predominantly on APCs, although T cells also express several TLRs. Adjuvants like the ones described above are added to vaccines to enhance the quality of the immune response, reducing the number of vaccine doses required to achieve immunity and sometimes making vaccination possible at all [161]. This is an important field that has been given renewed interest due to the ongoing pandemic, which we do not discuss further. We are mainly interested in the latent space representation of such experimental configurations.

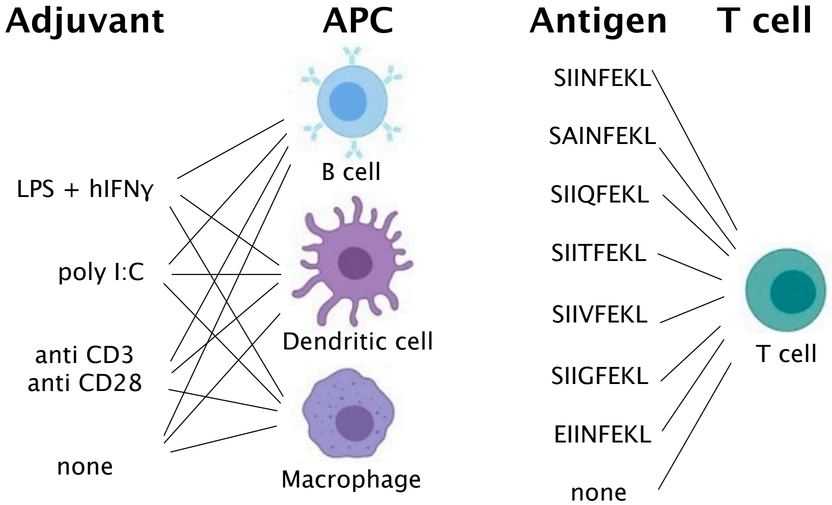


Figure 4.17: **Experimental setup for APC type experiment.** One of four adjuvants stimulates one of three APC types. Eight different antigens are loaded on the APCs and stimulate the T cells. Elements of this picture are taken from ibiology.org.

In Fig. 4.18, top panels, we show the F, t_0 parameter space for the fits to the integral latent space of all timeseries, given in Appendix Fig. C.6. Columns correspond to TLR agonists, colors indicate the antigen administered at 1μM and marker indicates APC type. Broadly speaking, we observe similar latent space dynamics regardless of APC type and adjuvant though the enhancement of the response varies in magnitude. Addition of adjuvants results in nearly all cases in a stronger response. This is clearly visible for antigens of lower quality (E1, G4, V4). The difference between macrophages and DCs is that for macrophages especially F increases, while for DCs, F increases a little, if anything, while

θ increases strongly (Appendix Fig. C.6), showing that macrophages and DCs respond differently to stimulation by TLR agonists. The F, t_0 parameter space for macrophages stimulated with adjuvants is reminiscent of T cell blasts, where antigen quality is no longer correlated with F but still with t_0. This effect is especially visible for LPS + hIFNγ but also for poly I:C. Regarding adjuvants, poly I:C is specific for macrophages, while LPS + hIFNγ enhances the response of all APC types to antigens of low to intermediate quality.

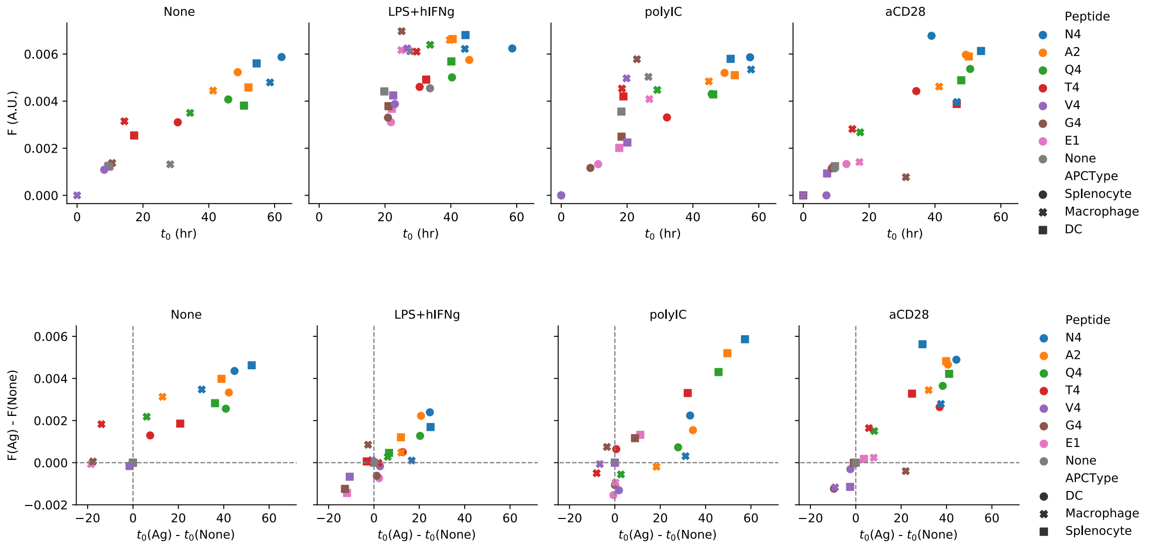


Figure 4.18: **Parameter space for APC experiment.** Top panels: F, t_0 plot of the constant force model of the integral latent space given in Appendix Fig. C.6. APCs used in these experiments are splenocytes (as usual), macrophages and dendritic cells (markers). The APCs are stimulated by the following TLR agonists: None, LPS + human IFNγ, polyIC and CD28 antibody (columns from left to right). Colors indicate antigen. Bottom panels: visualizing synergistic effects by considering parameters $F(\text{Ag})$ and $t_0(\text{Ag})$ relative to F and t_0 for timeseries without antigen administration ("None").

Finally, to our surprise, we found evidence for antagonism by comparing the model parameters to None (grey marker, no antigen administered). We visualize the potentially antagonistic effects by plotting $F(\text{Ag}) - F(\text{None})$ versus $t_0(\text{Ag}) - t_0(\text{None})$ for each antigen (Fig. 4.18, bottom panel). Here, Ag indicates the specific administered antigen. We comment on the findings from left to right in the columns. For "None" TLR agonist, there appears to be antagonism in macrophages for E1 and G4 and perhaps even T4 through parameter t_0. Upon inspection of the F, t_0 plot above, $t_0(\text{None})$ does not follow the trend $F \propto t_0$. Here, t_0 is surprisingly large given F. We cannot say with certainty that this is an error, but something might have gone wrong with the experiment or in the fit. This shows that one has to proceed with care in considering antagonism using this method, because it strongly relies on having the correct reference with the "None" agonist.

Moving on, for splenocytes stimulated with LPS + hIFNγ, both F and t_0 are negatively affected by antagonists V4, G4 and E1. Interestingly, professional APCs do not suffer from antagonism through LPS + hIFNγ activation. Then, for poly I:C, we find no antagonism in the splenocytes, but we do for macrophages and DCs. For macrophages, parameter F is antagonized for T4, V4 and G4, while for DCs t_0 is antagonized for V4, G4 and E1. The "None" agonist reference follows the trend of the other conditions in this experiments unlike in the "None" TLR agonist experiments, and thus seems more reliable. It is not known what causes the localized inhibition of the initial cytokine production in macrophages and the timescale of cytokine consumption in DCs, but that it happens is clear. There is likely a distinction in parallel processing of T cell activation for macrophages and DCs, i.e. through TCR signalling, acting on F and cytokines, acting on t_0.

In Chapter 3, we described antagonism for mixtures of antigens in detail. Here we show that T cell antagonism is not confined to mixtures of antigens. APCs activated through TLR agonists activate T cells via costimulators and cytokines even without presenting antigens. A T cell activated through these indirect channels will have its internal TCR chains phosphorylated regardless, triggering downstream pathways that result in T cell activation. It has been proposed that antagonists like V4, G4 and E1 bind to the TCR weakly enough to not further induce TCR chain phosphorylation, but strong enough to activate SHP-1, a global inhibitor of TCR chain phosphorylation [33]. Without being conclusive on the (family of) proteins mediating antagonism, an accepted general statement is that for ligand antagonism to occur, one requires activation of some sort, as well as an antagonist targeting a specific inhibitory mechanism against this activation. These experiments are further evidence against the hypothesis that antagonism in T cell activation is caused by competition for TCRs.

Moving on, we now discuss experiments with tumor cell lines. Due to an impressive feat of engineering, we can substitute splenocytes with tumor cells, while keeping the rest of the setup the same. The tumor cells are transduced with a retrovirus expressing the SIINFEKL protein, so they express only OVA antigens instead of tumor antigens [162]. Moreover, tumor cells do not naturally present antigens through their pMHCs, but they do once they are pulsed with an interferon, like IFNγ. Finally, we expect T cells to kill tumor cells, as long as the tumor presents antigen of sufficient quality and in sufficient quantity. This introduces a population level feedback loop to the cytokines. All this is done for several tumor cell lines (melanoma, lymphoma and adenocarcinoma), which introduces additional variation. For instance, melanoma is a fast-replicating cancer, so the population-level feedback might become important rapidly. At the start of the experiment, the tumor cells are pulsed with IFNγ at various concentrations, correlating with antigen quantity. Throughout the experiment, the pMHC expression varies depending on the amount of IFNγ produced. This is different from using nontumor APCs, where the number of antigens decreases exponentially over time.

We summarize the experiments in integral latent space (Fig. 4.19, top half) with corresponding latent space parameters (bottom half). We show two repeats of the same experimental setups done with 30000 tumor cells. We note that θ is much different between repeats with the same tumor and antigen, more different than between tumors within the same repeat. Clearly, there is experimental variability due to tumor and T cell preparation, variability that is more pronounced due to the tumor - T cell interaction. We may speak of a success with regards to the robustness of the integral latent space dynamics and the ability to fit these using the constant force model. We find the correct hierarchy emerge through F or t_0 for each of the tumor types and experiments, except for TumorTimeseries_2 with B6 APC type (splenocytes), where there was a mix-up of antigens.

Drugs

Up until now, we have been mainly concerned with validations of the integral latent space and the ballistic model through interpreting experimental configurations in the latent space. Although we have seen parameters shift in certain experimental configurations (i.e. an-

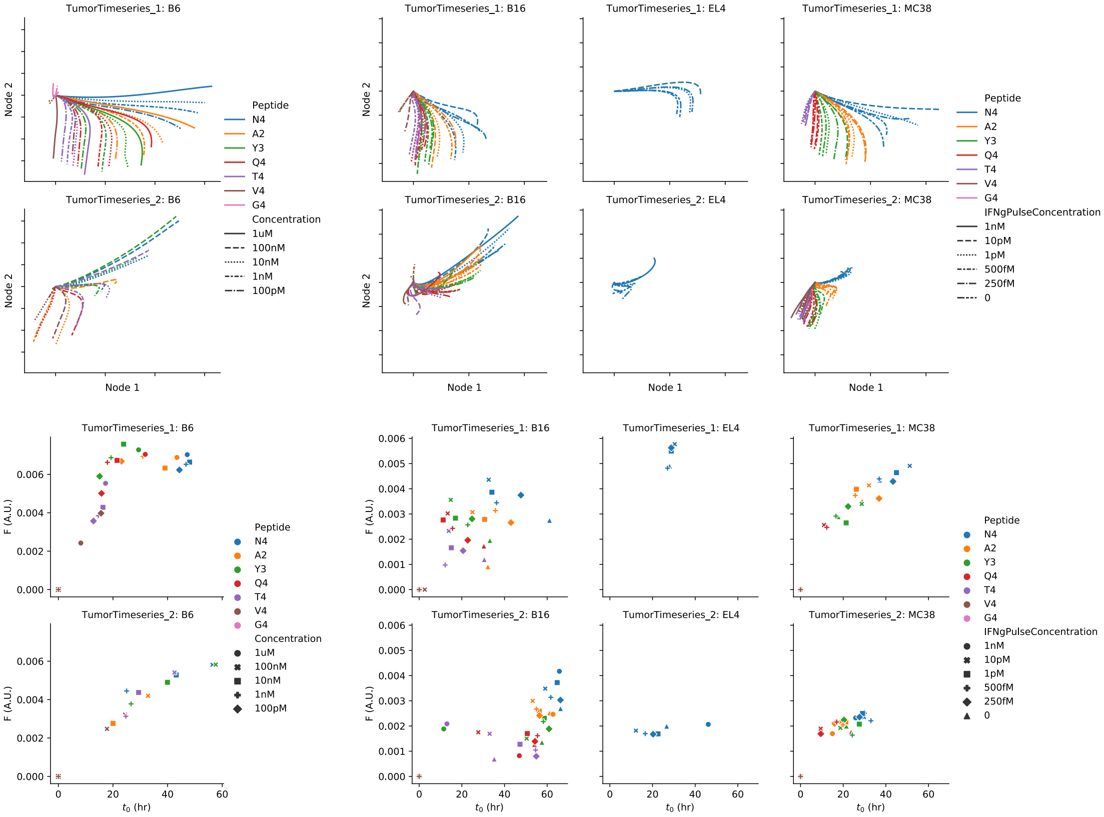


Figure 4.19: **Tumor experiments.** Data projected in latent space (top panels) and F, t_0 plot of the parameter fit of the ballistic model (bottom panels). Rows for both latent space and parameter space show two experiments done with tumors and T cells from different mice. Columns show from left to right splenocytes (B6), skin cancer cells (B16), lymphoma cells (EL4) and lung cancer cells (MC38). Tumors are genetically engineered to express SIINFEKL antigens, and require pulsing with IFNγ for pMHC expression.

tagonism in macrophages), making it plausible that these parameters capture underlying biology, we never explicitly intended to change the model parameters. The last series of experiments is designed to do exactly that. An intriguing observation of the parameters is that F is generally strongly correlated with t_0 and θ, except in special cases. The question is now: can we decorrelate F, t_0 and θ? Marchingo et al. quantified how many divisions T cells go through during an immune response [163]. They found that the effect of signal 1, 2 and 3 (antigen binding, costimulation and cytokines) sum linearly. Taking away one of these signals reduces the number of division T cells go through. This proportionally decreases the cytokine response, as cytokine concentrations are proportional to the size of the T cell population. Through drugs experiments, we change signal 2 and 3, and wonder how this affects the fitted parameters. Experiments are set up with antigen N4, Q4 and T4 at 1μM. Drugs are added 4 hours after the start of the experiment. In these experiments, we report the percentual change in F, t_0 and θ with respect to when no drugs is added, similar to how we reported antagonistic effects in bottom panel of Fig. 4.18. Now we can reliably report percentual change, because the "None" condition for each of the antigen gives nonzero parameter values for F, θ and t_0.

Our collaborators tested 24 drugs in this drugs experiment. Because the graphs get messy with so many dimensions (24 drugs, 3 antigens, 3 parameters), we decided to sort the drugs from largest negative effect to strongest positive effect. In the left panel of Fig. 4.20 we see how each of the drugs affects the parameters. The % difference for each of the drug averaged over the three peptides and parameters ranges from -83% (antibody for IL-2 receptor, aIL-2) to +7% (Resiquimod). In the right panels we have sorted drugs in three groups (strong inhibitors, weak inhibitors and weak activators). Comparing the effect of the drugs across parameters and antigens on the left panel, we observe that it is much easier to change parameters for the T4 response, especially positively. For instance, drugs are able to increase θ only for T4. This makes sense: N4 and Q4 already have large angles, so it is hard to increase them even further given the constraints on cytokine production. The second observation is that it is hard to completely shut down the response following drugs administration after 4 hours. Dasatinib shuts down TCR signalling, but at 4 hours, the immune response is already underway, and all three parameters can only be reduced so much. Only aIL-2 manages to completely shut down the response to Q4 and T4. aIL-2 prevents already produced IL-2 to signal through the IL-2 receptor, thereby putting a stop on cytokine production.

Coming back to our interest in decorrelating parameters, we now inspect 2d plots of the percentage change of ΔF(Drug) vs Δt_0(Drug) (top right panels) and ΔF(Drug) vs $\Delta\theta$(Drug) (bottom right panels). Colors indicate antigen. For the weak inhibitors, the drugs affect the parameters differently depending on the antigen: For N4, t_0 is slightly decreased,

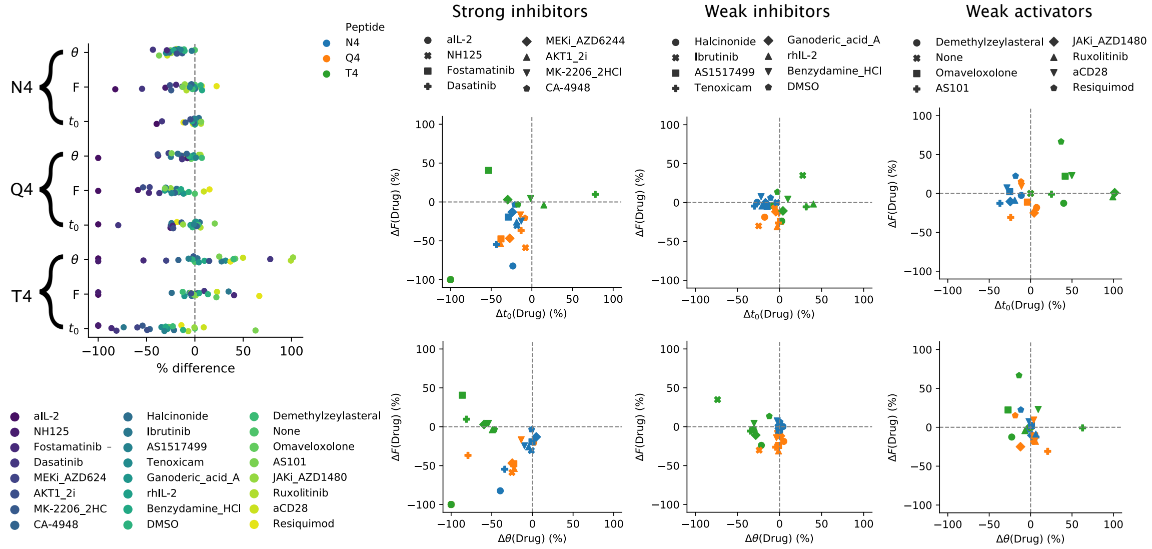


Figure 4.20: **Parameter spaces of the drugs experiment**. Cytokine response of the drugs experiment is projected in latent space and fit with the constant force model. Left panel shows % difference w.r.t. no drugs added for each parameter and antigen. Drugs are sorted by color based on their % difference averaged over each antigen and parameter. 24 drugs are divided into strong inhibitors, weak inhibitors and weak activators. Right panels show 2d space of % difference of F vs. t_0 (top) and F vs. θ (bottom) for each of the three categories of drug inhibition/activation. Color indicates antigen, marker indicates drug, which is different for each column of panels.

for Q4, F is slightly decreased, and for $T4$, θ is mostly decreased, although t_0 also increases. The weak inhibiting drugs are the closest we get to decorrelating the parameters. To our surprise, we find that the effect changes per antigen, meaning that parameters get fixed at in different ways depending on the antigen. This means that to study the effect of drugs on inhibiting the immune response, the antigen quality for the experiment matters. The reason it does is likely that the time scales are different per antigen. We have already seen that t_0 relates to the time at which we move from the IL-2 production phase to the IL-2 consumption phase, which ranges from 20 hours (T4) to 60 hours (N4). Not just is t_0 different per antigen, it is likely that the time at which t_0 is set is different per antigen. Similarly, for θ and F, when θ is high in the first hours, it is difficult to bend the curve all the way down, while for smaller θ, like with T4, there is more flexibility in this parameter.

We found a common denominator for some of the drugs that change θ for T4: they reduce or enhance general inflammation by acting on cells in the innate immune system. Tenoxicam is an anti-inflammatory drug, ibrutinib dampens the B cells response, AS101 inhibits anti-inflammatory IL-10 and augments TNFα [164] and, fostamatinib reduces inflammation of APCs like macrophages. With drugs, it is hard to target specifically one parameter because they are tightly connected to the antigen quality. For instance, dasatinib disrupts TCR signalling and thus reduces F, t_0 and θ. The drugs that target cells in the innate immune system target θ specifically, as F and t_0 are purely linked to adaptive immunity. This leads us to hypothesize that θ measures the ratio between activation of adaptive immunity and innate immunity.

We can now try to relate the parameters explicitly to signal 1, 2 and 3. F relates to TCR signalling, θ relates to costimulation and t_0 relates to the timescale of IL-2 consumption. This connection is based on the observation that θ is affected by the activation of APCs, and that APCs contribute mostly via costimulation. APCs are responsible for part of the cytokine production too, but specifically not for IL-2, whose production and consumption phase determine t_0. Marchingo et al. demonstrated the tunability of these signals changing the number of divisions T cells go through. That we can establish connections between the parameters of the ballistic model and fundamental immune processes shows that we have developed a complete model of the population immune response.

Predicting quality

Quality is a measure of both the pMHC - TCR binding strength, characterized by the binding time τ, and the magnitude of the immune response, characterized by the EC$_{50}$. EC$_{50}$ is the antigen quantity at which a half max of a dose-response curve is observed. It is one of the three parameters in the Hill equation that the dose-response curve is fit to

$$C = \frac{AL^m}{L^m + EC_{50}^m}. \tag{4.29}$$

Here, A is the amplitude of the response, m is the Hill coefficient that sets the steepness of the curve, L is the antigen quantity, and C is an observable that is measured repeatedly, like

IFNγ concentration or the number of activated T cells after a certain time following stimulation. Our collaborators measured reference EC$_{50}$s, given in Table 4.2. This is roughly in correspondence with EC$_{50}$s of the same OVA antigens in [148] and [149]. The dose-response curves from which the EC$_{50}$s are estimated are shown in Fig. 4.21. From left to right, Daniels et al. measured the % of CD69+ thymocytes 16 hours post-activation, Zehn et al. measured intracellular IFNγ 24 hours post-activation, and our collaborators measured the size of a specialized cluster of T cells high in several markers of activation 10 hours post-activation. The differences between Zehn et al. and Achar et al. can be partly explained by the different observables (intracellular IFNg vs. cluster of specialized activated T cells). The differences between Daniels et al. and Achar et al. can be partly explained by the difference in T cells (naive cells vs. thymocytes, cells from the thymus in the state before they become naive cells). Another factor comes from the preparation of the antigens, which may affect the antigen loading on the APC. For instance, our collaborators found that the Q7 antigens (antigens with a mutation of the OVA antigen at the anchor residue) loaded on the APCs inconsistently, which is why they were excluded from this study. It is a given that it is difficult to quantitatively reproduce experiments across labs. That we find a correspondence within a factor of two for most antigens is already promising.

Table 4.2: Reference EC$_{50}$s scaled to N4, the antigen with lowest EC$_{50}$, and log transformed for three datasets from different authors and obtained in different ways.

Antigen	Daniels et al. (2006)	Zehn et al. (2009)	Achar et al. (2020)
N4	1	1	1
A2		2.7	2.4
Y3		4.1	7.6
Q4	39	18.3	21.5
T4	122	70.7	150
V4		680	1336
G4	7515		28292

The EC$_{50}$s given in Table 4.2 are quantities relative to the strongest antigen N4. For a good estimate of the EC$_{50}$, many measurements in the non-saturated regime are required. This is challenging if done for a range of antigens, because the non-saturating regime shifts

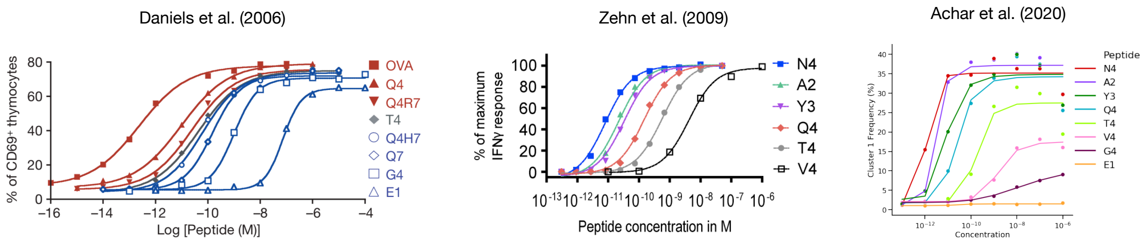


Figure 4.21: **Dose-responses curves to estimate EC$_{50}$ of OVA antigens.** Figures taken from [3, 148, 149].

towards higher concentration for lower quality antigens, and is per antigen at most two orders of magnitude wide. For an antigen of unknown quality, one would have to cover four decades, and multiple points per decade.

In this section, we attempt to predict EC$_{50}$s from a single cytokine timecourse and compare this to several baselines. We start by recognizing that four parameters in the constant force model summarize the data well, as we exhaustively showed in previous sections. We set up a multilayer perceptron using the model parameters as input and the relative EC$_{50}$s as a continuous output. Instead, we are only interested in one quantity, the EC$_{50}$, thus there is only one node O in the output layer, instead of one node per label. Matrix multiplication without activation function leads to

$$O = \mathbf{h} \cdot \mathbf{w}. \tag{4.30}$$

As usual, $\mathbf{h}$ are the values in the hidden layer and $\mathbf{w}$ is the vector of learned weights connecting the hidden and output layer. The typical loss function is the mean squared error (MSE)

$$MSE(O) = (O - EC_{50})^2, \tag{4.31}$$

where O is the predicted EC$_{50}$ and EC_{50} is the actual value. The goal of the learning algorithm is to minimize this quantity summed over all samples. To remove bias for high EC$_{50}$s in the mean squared error, we take the log transformed $\log_{10}(EC_{50})$ as output, measuring the difference between the antigen and $N4$ in orders of magnitude in antigen quantity to reach the half max.

Hints on what architecture may be required for good regression performance can be found by inspecting the score and projecting the input on the nodes in the hidden layer for an incrementally increasing number of nodes starting with a single hidden layer. With one node, we find a score of about 0.75 on the validation set, depending on the training run. With two nodes, the score for the best training run can reach over 0.80, and it does not improve with more nodes or more hidden layers. The projection of the test data on Node 1 and Node 2 of the two-node MLP regressor are given in Fig. 4.22 (left panel). It is no coincidence that the one-node MLP regressor already works well: the two nodes exhibit a strong linear dependence. Predicting EC_{50} from model parameters is a near one-dimensional problem, but we stick to a two-node hidden layer to exploit some of the nonlinearity at the boundary.

It is not surprising that a one-node hidden layer MLP regressor works well. First, we cannot expect a complex model to generalize well, because of the reduced amount of data. We are left with one sample per timeseries for the wildtype experiments with 10^5 precursors, for a total of 66 samples. Second, in previous sections, we have shown that the parameters of the ballistic model are correlated (Fig. 4.10). That the latent space of the model parameters is also close to one-dimensional then only makes sense. With the architecture presented on the right panel of Fig. 4.22, we train the MLP for 3000 iterations with a regularization rate of 0.1 (system is not sensitive to these hyperparameters) on the training data, and predict the EC_{50} of antigens in several other datasets, including unseen antigens A2 and Y3. Before we present the results, we introduce several baseline comparisons.

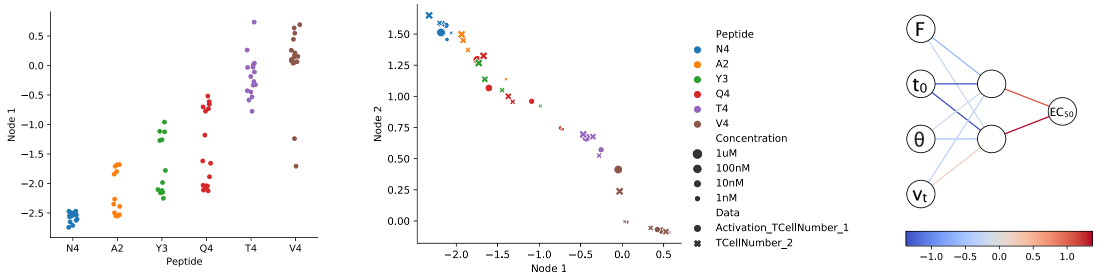


Figure 4.22: **Architecture MLP regressor.** Latent spaces for a one-node hidden layer MLP regressor (left) and two-node hidden layer MLP regressor (center). Right panel shows architecture of the MLP regressor with latent space in the center panel.

One may wonder if we even need an MLP regressor to predict EC_{50}. The simplest regressor is the linear regressor, which optimizes parameters a, b, c, d, e in the linear model

$$\log EC_{50} = aF + bt_0 + c\theta + dv_t + e. \tag{4.32}$$

We use L1 regularization (LASSO) with rate 0.1 to enhance generalizability, and fit the parameters using the same training data. Finally, in the true baseline, we predict EC_{50}s using the cytokine integral at a single timepoint, a measurement that is very similar to IFNγ or IL-2 Elispot. We fit a linear function to the cytokine integral $[cy]$

$$EC_{50} = a \cdot [cy] + b. \tag{4.33}$$

This requires fitting only two parameters. The model could be made more complex by including higher order polynomials, but as the linear model estimates EC_{50}s reasonably accurate, we see no need to go beyond this. The model parameters are set as to minimize the mean squared error between predicted and actual EC_{50}, like in Eq. 4.31. In the generalized linear model, we measure multiple cytokines, like we would for Fluorospot [147]

$$EC_{50} = \sum_{i=1}^{N} a_i \cdot [cy_i] + b. \tag{4.34}$$

We choose to measure IFNγ and IL-2 at 12 hours. Results for the baseline are shown in the three leftmost panels of Fig. 4.23. We show conditions with 10^5 precursors. Predictions for the EC_{50} per antigen using IFNγ at 12 hours range over at least one and even two decades for N4. Apparently, the IFNγ integral at 12 hours is too sensitive to antigen quantity to robustly predict EC_{50}. The EC_{50} prediction using IL-2 at 12 hours is confined to at most one decade across datasets and concentrations, except for T4 1nM, which is already better than we could do with IFNγ. The hierarchy within a dataset is preserved for the most part, but not entirely. Fluorospot gives almost equivalent predictions as IL-2 Elispot would, meaning that IFNγ does not contain information about EC_{50} that IL-2 already contain. The right panels show the EC_{50} prediction of the linear regressor and MLP regressor using the model parameters. The linear regressor makes precise prediction using the model parameter but operates in a regime that is too narrow, unable to predict a low enough EC$_{50}$ for N4 or high enough EC$_{50}$ for V4. This demonstrates that $\log(EC_{50})$ is not linear in the model parameters. The nonlinear activation function in the MLP regressor (hyperbolic tangent) allows us to reach the lower limit, at the expense of making less compact predictions per antigen. The MLP regressor remains limited in its capacity to predict the high EC$_{50}$ of V4. Inspecting the learned weights, we find that the first hidden node decreases with F, t_0 and θ, and increases with v_t, while the main node 2 contribution comes from t_0 and θ. The small cytokine response coming from V4 is fitted best by setting all F, t_0 and θ near zero and v_t to a small value, so that Node 1 is slightly positive and Node 2 is 0 (Fig. 4.22, center panel). It has not progressed far enough along Node 1 so that its positive contribution can push the EC$_{50}$ to the actual value of EC$_{50}$(V4). Predicting EC$_{50}$ from smaller cytokines responses by self antigens like G4 and E1 gives similar issues. This unveils a structural inability that this procedure suffers from: without much of a cytokine response, model parameters are small, and we cannot predict an EC$_{50}$ $> 10^3$.

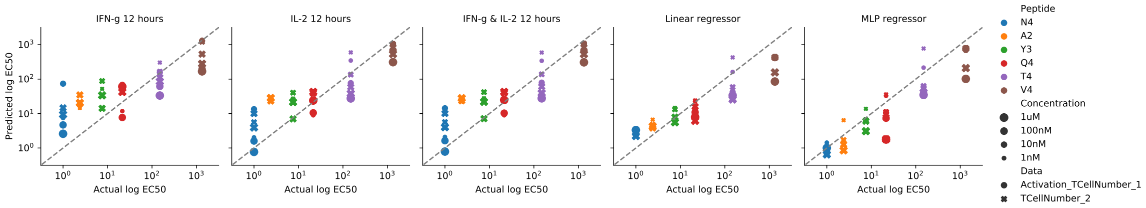


Figure 4.23: **Predicted EC$_{50}$ plotted versus the actual EC$_{50}$.** Columns show the predictions using different methods. Dashed line indicates the diagonal where the actual and predicted EC$_{50}$ are equal. As usual, color indicates antigen quality, size indicates antigen quantity and markers indicate different datasets.

We now proceed with the more challenging scenario where both precursor frequency and antigen quantity are unknown. How well do each of methods do in predicting EC$_{50}$? We quantify how well the predicted EC$_{50}$s correlate with the actual EC$_{50}$s through R squared, defined as

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}},\ SS_{\text{res}} = \sum_i (y_i - f_i)^2\,,\ SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2\,. \tag{4.35}$$

Here, SS_{res} is the residual sum of squares from the model prediction. SS_{tot} is the total

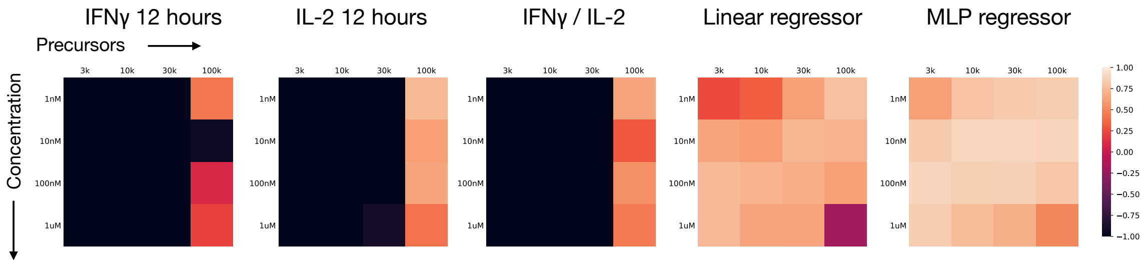


Figure 4.24: **R squared for experimental configurations from EC$_{50}$ predictions.** From left to right: IFNγ, IL-2, IFNγ + IL-2, linear regressor and MLP regressor. Within a subpanel, rows indicate concentration and columns indicate precursor frequency For the squares with 10nM and 100nM four contributions are summed, for 1nM and 1μM, there were 12 contributions. The methods are optimized with 10^5 precursor T cells.

sum of squares, measured as the difference between the actual values and the mean value, proportional to the variance of y. It is useful to discuss the limits to understand R^2. First, when $SS_{\text{res}} \to 0$, $R^2 \to 1$, meaning that the model perfectly describes the data. When $SS_{\text{res}} \to SS_{\text{tot}}$, $R^2 \to 0$, the model does as good as predicting the average of the values for every sample, which is not very good. When the model does worse than that, despite being a "squared value", R^2 becomes negative. We report values down to -1, although R^2 is not bounded from below. We compute R^2 for data that the functions were fitted to (10^5 precursors), as well as for out-of-distribution data ($3 \cdot 10^3$, 10^4, and $3 \cdot 10^4$ precursors). The results are shown in Fig. 4.24. From left to right, we observe the IFNγ, IL-2 and IFNγ/IL-2 baseline, the linear regressor and MLP regressor trained on parameters of the constant force model. For the regressors, R^2 approaches 1 for some pairs of concentration and T cell numbers, and is positive everywhere, even for $3 \cdot 10^3$ cells with 1nM. The MLP regressor gives a higher R^2 than the linear regressor with values $R \geq 0.5$ for all conditions. The baselines give acceptable results for 10^5 precursors, IL-2 better than IFNγ, but not at all with fewer precursors, because the predicted quality is highly dependent on absolute levels. This is precisely why one usually computes dose-response curves. Then, a change in cytokine concentration, be it due to the level of pre-activation of the T cells or the sample preparation (i.e. number of T cells in a well), affects the amplitude of the Hill equation 4.29, but not the EC$_{50}$. As long as for each of the measurements, the same sample preparation is used, one can recover the EC$_{50}$. The difficulty with predicting the EC$_{50}$ of an antigen of unknown quality using dose-response curves is first that the antigen quantity needs to be known, and second, that the antigen quantity needs to be varied in the nonsaturating regime. For an antigen of unknown quality it is not known where this is, meaning a large regime has to be sampled. Our procedure allows for the prediction of EC_{50}s by measuring the cytokine response of a single timecourse independent of antigen quantity and precursor frequency.

4.4 Discussion

In this chapter, we analyzed detailed time-kinetics of the cytokines response following T cell stimulation aiming to recover antigen quality independent of antigen quantity. We designed a procedure that allowed us to predict antigen quality from a timeseries of five cytokines. The procedure started with processing the data through a log transformation of the cytokine concentrations, smoothing with a moving average, spline-fitting, integration of the splines, and sampling from the integrals at hourly intervals. We trained an MLP with five input nodes, two nodes in the hidden layer and six output nodes. We predicted every sampled timepoint individually, and applied a majority rule to determine the antigen class of the timeseries. This procedure allowed us to correctly predict the antigen class of unseen timeseries over three orders of magnitude in EC_{50} and two orders of magnitude in quantity (10nM - 1μM). We reached the limits of precision, as the cytokine profile of an antigen at concentration 1nM is equivalent to cytokine responses from an antigen of the class just below at concentrations of 1μM and 100nM. From this we conclude that T cells measure antigenicity in a continuous fashion as a convolution of quality and quantity.

We made many choices in setting up the pipeline, and acknowledge that our procedure is not the only way to predict antigenicity from cytokines. For instance, one may wonder if it is necessary to use cytokine integrals, an observable that is difficult to interpret. During the development of the pipeline, we started off using cytokine concentrations, but did not manage to design a neural network that was robust to IL-2 disappearing from the system. Inspired by work by François Bourassa, we made a breakthrough in predicting antigen quality using cytokine integrals. We sampled 72 timepoints from a timeseries and treated each datapoint as an individual sample. Naturally, the datapoints were highly correlated, so we demanded our machine learning procedure to consist of a minimal architecture allowing for maximal interpretability to not overfit. This also allowed us to make connections to cytokine production in T cell decision-making. These design constraints guided the choices that we made in setting up our procedure in terms of data processing, hyperparameters, training data, measuring accuracy, etc.

Through a feature analysis, we discovered how the latent space was build from the ground up by combining incrementally more cytokines. We proposed a hierarchy of cytokines, and found out that for good antigen classification, one needs a cytokine that is predominantly produced by immune cells from the innate immune system and one that is predominantly produced by cells from the adaptive immune system.

Continuing along the lines of interpretability, we then treated the latent space as a literal space, parameterizing the latent space dynamics with a piecewise ballistic model consisting of a propulsion phase and a free fall phase. We found that the constant force outperformed the constant velocity model by better capturing the cytokine production phase and better estimating t_0, the parameter determining when to switch phase. We then performed a sensitivity analysis by considering the eigenvalue spectrum of the Fisher Information Matrix, and considered regions in which the analysis indicated the model had sloppy parameters. We understood that for timeseries with N4, v_t is ill-determined, as the free-fall phase has not started at the end of the experiment, and that for timeseries with V4, F and θ are ill-determined because t_0 is small, so there is just the free-fall phase. We noted that the first eigenvectors of the FIM consisted mainly of a combination of F, t_0 and θ, as if they were strongly correlated, and all determined at the start of the experiment. We then seeked an interpretation for each of the parameters. F corresponds to the magnitude of the cytokine response, t_0 is the timescale at which IL-2 consumption takes over IL-2 production, θ measures the importance of adaptive immunity relative to innate immunity, and v_t is the steady-state response independent of adaptive immunity, what we could call the amount of chronic inflammation.

Next, we represented the cytokine response of many different experimental configurations in latent space, parameterized the latent space dynamics, and provided a biological interpretation. We also looked for applications for our pipeline. First, we saw that the latent space dynamics are virtually independent of precursor frequency. We found that with decreasing precursor frequency, F decreases while t_0 increases, corresponding to the notion that T cell expansion continues for longer when the precursor frequency is smaller. Next, we considered different T cell types. We interpreted the latent space representation for T cell blasts, which are better described by the constant velocity model. We created a similar hierarchy of antigen quality for antigens specific to T cells from other transgenic mice. Finally, we noted a stronger quantity dependence in helper T cells, and that with decreasing quantity F decreased, while t_0 remained constant.

Moving on, we changed the APC type to macrophages, DCs and tumors. Stimulation by adjuvants enhanced the immune response in different ways for macrophages (increased t_0) and DCs (increased F), especially for antigens of low to intermediate quality. Unexpectedly, we found evidence for T cell antagonism at the cytokine level. Experiments with tumors resulted in population-level dynamics through T cells killing tumors depending on their activation and the tumor type. Despite this additional level of response, we retrieved the hierarchy of OVA antigens for each of the tumor cell lines. Finally, interested in the correlation between F, θ, and t_0, we analyzed parameters for drugs experiments. We found that the immune response is hard to shut down entirely following drugs administration after 4 hours, and that the effect of the drugs on each of the parameters differed per antigen. The strongest inhibitor was the IL-2 antibody, the strongest activitor was resiquimod. Drug inhibitors that affect the APCs specifically target θ, tilting the balance between adaptive and innate immunity. We found that many drug inhibitors targeted a combination of F and t_0, and that weak inhibitors targeted F (Q4) or t_0 (N4) specifically. They would always decrease the parameter though. Only for T4 could drugs systematically enhance the immune response. This points towards inherent constraints on T cell activation: it is easier to slow down cytokine consumption than it is to accelerate the response, especially for high quality antigens that already cause a strong response by themselves. The parameter θ, the ratio between activation of the adaptive immune system and inflammation, cannot be changed through the T cells, but through stimulating or inhibiting APCs, causing stronger or lighter inflammation. In future work, a systematic approach to drug perturbations could allow disentangling the effect of specific inhibitors on the immune response, and provide a more detailed biological interpretation of the parameters. We have already given a taste of what is possible by identifying θ as responding mostly to APC inhibition or activation. The same could be done for F and t_0.

Finally, we designed a procedure to predict EC_{50} with a MLP regressor using parameters of the ballistic model as inputs. We compared our results to baselines of IFNγ and IL-2 measured at single timepoints. We found that for a fixed precursor frequency, IL-2 at 12 hours did comparatively well across datasets, but when the precursor frequency changed, the performance of this baseline classifier dropped quickly. IFNγ at 12 hours was already too sensitive to variability in quantity in the experiments for fixed precursor frequency. This came to our surprise, as in clinical settings IFNγ Elispot 12 hours after activation is used to assess the T cell response. Apparently, this measure needs no reference. The MLP regressor was robust to precursor frequency, being able to quantitatively predict quality independent of quantity and precursor frequency based on a single timecourse.

This leads us into a discussion of the application of our pipeline. The reason Elispot is predominantly used in clinical settings is because it is cheap and simple. These are clear advantages over our system, which currently is more expensive and requires expertise. Mabtech has spent many years optimizing Elispot. Oftentimes, an Elispot reader is sold to laboratories who make extensively use of Elispot, making it easy to use. One could imagine that in the future it will be simpler to run the assays our collaborators do to obtain detailed cytokine time-kinetics. For the moment, it requires installation of a robot, expertise to program the robot and a flow cytometer to measure the cytokines. Still, there are currently already scenarios where it is crucial to obtain a more precise assessment of antigen quality with respect to a known reference.

An applications that comes to mind is in adoptive T cell therapy, one of several upcoming immunotherapies. Here, T cells and tumor cells are extracted from the blood and tumor of a patient. The T cells are selected for tumor-specificity, expanded in-vitro, and injected into the patients with the hope that the cancer-specific T cells attack the cancer [165]. With the robotic platform, tumor specificity can be determined accurately and should allow for careful selection of T cells. Moreover, following expansion, T cells can again be tested to make sure they are not overstimulated. Only then should they be injected into the patient. Finally, efforts are made to develop combination therapies of immune checkpoint blockade and adoptive T cell therapy [166]. This again is a step that can be tested in-vitro to understand how much more reactive the T cells become following immune checkpoint blockade.

Our work concerned testing T cells from transgenic mice, mice whose T cells have only one type of TCR. This relies on the hypothesis that within the entire immune repertoire, only a single TCR is specific to the antigen. This is an outdated hypothesis: cross-reactivity matters. Cross-reactivity is the reason why humans in principle have a specific response against any pathogen one can imagine. It is of interest to understand with our pipeline how cross-reactivity affects the cytokine response. Unfortunately, it is not feasible to test an organism's entire immune repertoire in the controlled environment of a well in an incubator: the number of T cells do not fit. One has to resort to using the mouse as a test tube, in which it is a lot more difficult to accurately measure cytokines. One could extract blood from the mice and measure the cytokines, but ethical principles prevent bleeding a mouse more often than once every 6 hours. These measurements are also not nearly as precise as cytokines in a well are, if only for the fact that the blood is not a complete representation of the current state of inflammation.

In this work, we focused exclusively on the cytokine response, while there is a non-explored world of cell surface marker data. A direction for future work is to devise a model that takes into account both T cell states and global output like the cytokine response. Examples of this focusing on IL-2 alone include [114, 124]. We have seen glimpses of connection to models of microscopic interactions, i.e. through antagonism at level of TCR interactions and cytokine response following addition of adjuvants, drugs that inhibit or activate microscopic aspects of the response, resulting in macroscopic changes. Future work could bridge the gap between models describing microscopic interactions resulting in a heterogeneous population of T cells, and macroscopic outputs that are measured here.

5

Outlook

In this thesis, we have discovered and explored latent spaces of immune recognition in the broadest sense of the word. We introduced $\bar{\phi}$, a fitness-based parameter reduction algorithm in Chapter 2 with which we extract "functional latent spaces": core modules in a biochemical network that implement a desired function in a biochemical network. We mapped three models of immune recognition onto each other and discovered how antigen discrimination is implemented in each of the models. We found that an adaptive and a kinetic sensing module are required for discrimination of antigen quality independent of antigen quantity. The adaptive module is implemented via a feedforward mechanism (positive or negative) and a negative feedback loop, corresponding to what's known in the literature about minimal models of adaptation [1, 61, 62]. We then were able to reconstruct a hierarchy of models. Further down the hierarchy, models are more detailed, containing additional features or modules, retrieving antagonism, nonmonotonicity, and bistability.

The fitness function to evolve the adaptive proofreading model was the same as the fitness function to reduce all three immune recognition networks. That means that building models from the ground up can be seen as a complementary process to reducing models from the top down. In the latest version of the evolutionary algorithm ϕ_{evo} [152], one can evolve networks by fitting data using the chi-squared function. The Manifold Boundary Approximation Method utilizes the chi-squared function as well, minimizing the difference between modelled and simulated data [51, 52]. Proulx-Giraldeau and François found that the choice of fitness function affects the parameter space one traverses through evolving or reducing networks (personal communication, April 5, 2019). They reduced oscillatory models with different fitness functions, and showed that more coarse-grained functions that compute a biological function result in a less rugged parameter space than fine-grained fitting functions that measure the distance between modelled and simulated or real data. With a fitness function, local minima are less ubiquitous, and it may be easier to retrieve the global minimum. Similar ideas have been proposed in the neuroscience literature [11].

A straightforward extension of FIBAR is to apply it to models of different processes, for instance to the T cell proliferation model that we discussed in Chapter 4. Mayer et al. based their model on [136, 137]. They also manually reduced the model of [136] to retrieve the typical power-law behavior in fold expansion and precursor frequency (personal communication, August 1, 2018). The manual reduction was reproduced with FIBAR, but it could not retrieve the further simplification of [138] (unpublished results). This was attributed to the constraint that the fitness function cannot decrease per reduction step, making the reduction get stuck in a local minimum. Allowing for a probability to accept a reduction step inversely proportional to the decrease in fitness could help FIBAR traverse local minima. Moreover, we chose the fitness function to represent a power-law relation, which may not have been the optimal function. This example shows that determining what fitness function best represents the biological function is not trivial and requires domain knowledge.

An interesting, related approach is “experimental parameter reduction” applied to a complete model of circadian redox oscillators to extract the core motif [167]. del Olmo et al. “clamped” variables one by one, meaning they set an oscillating variable to its mean value, “resembling conditions of constitutive expression from the wet lab” [167, 168]. If with a clamped variable, the network continues oscillating, this variable is not required for the generation of self-sustained oscillations, and can be fixed or removed. This techniques is specific for network exhibiting oscillatory behavior: a mean value in a timeseries is otherwise not well-defined or not impactful if it reaches a steady-state.

Moving on, we now summarize the latent space of antagonism as explored in Chapter 3. Antagonists or adversarial examples are small, specific perturbations of the input space that aim to fool the classifier. We have shown how the Fast Gradient Sign Method [74] applied to a naive immune classifier increases the number of weakly binding ligands, which is exactly ligand antagonism as we know it. We then show how through transforming the digit classifier with a Hill function of power N, we retrieve qualitatively similar behavior for the classification of interpolated samples as the robust immune classifier with proofreading and feedback $N > m > 1$: the output of the classifier is flat near the input sample and decreases sharply near the decision boundary. Following [103], we move a ligand distribution in the direction of the decision boundary, and study the resulting distribution. Naive classifiers $m \in [0, 1]$, are antagonized by weakly binding ligands. For $m = 2$ we find a flat distribution due to the emergence of a critical point, and for $m > 2$, the ligand distribution peaks with antagonists with binding times just below the threshold. It also takes many iterations to reach this distribution, owing to the flat antagonism potential for ligands of small binding time. Finally, we inspect what digits reside at the decision boundary of a robust digit classifier. Using a few-pixel attack, and turning pixels on or off sequentially in order of their effect on the scoring function, we find ambiguous digits at the decision boundary, in line with results from [103]. It shows how through adversarial robustness, the machine learning classifier has been made more interpretable. Indeed, it should not need to correctly classify out-of-distribution digits that are seemingly of another class.

In this work, we have shown the characteristics any classifier robust to small, specific perturbation should possess: a flat landscape around the distribution of samples belonging to certain classes that steeply descends towards the decision boundary in between those classes. The two regimes are separated by a critical point. The qualitative observation of flatness corresponds to the observation on machine learning classifiers that flat landscapes around the classes and strongly curved at the boundaries gives rise to adversarial robustness [109]. Moosavi-Dezfooli achieved this using curvature regularization. It was not known that the presence of a critical point causes this behavior.

Another continuation of this work is the characterization and interpretation of the class of adversarially robust machine learning classifiers. Krotov and Hopfield showed in models of Dense Associative Memory (DAM) that using Rectified Polynomials of high degree as activation functions leads to memorization of training digits in the DAM [103]. We were able to retrieve the characteristics of adversarial robustness that Krotov and Hopfield found using a Hill function of high power as an activation function in a multilayer perceptron (MLP), but this did not result in interpretable weights in the hidden layer. It would be of interest to find out under what conditions prototypes appear, and if this is a necessary condition for adversarial robustness. It is most natural to compare digits to other digits when attempting to classify them, rather than to compare digits to low-level features that may statistically relate to digit classes, but are not interpretable by any means. It is possible that we misinterpret human classification and that humans may not necessarily recognize digits as a whole, but only through combining features of the digits. Yet, deep convolutional neural networks hierarchically construct features: the first layers are low-level edge-detectors and the final layers are high-level object recognition [75]. That a DAM can be made to do object recognition instead of feature-detection through a simple change in activation function makes it likely that there exists a regime where this transition also appears for a standard MLP. Such work may provide an incentive to add the "ambiguous" class to the labels of classifiers.

Immune classifiers are an example of non-neural machine learning classifiers (decision-makers). It would be of interest to study fooling mechanisms of other biological, sensory systems that also make decisions, for instance, in yeast [169, 170] or bacteria [118]. Learning and pattern recognition is a prominent aspect of the immune system, and through comparisons with machine learning, more insights can be gained on the algorithms and implementation. A recent example is for negative selection, which is shown to possess generalizability of unseen self peptides when tested in a machine learning model of the T cell repertoire [171]. Another example is on framing the dynamics of clone size distributions in the adaptive immune system as a reinforcement learning system [172]. Finally, a connection on evolutionary time-scales could be made through the implementation of the evolution of immune-virus dynamics through generative adversarial networks (GANs) [173]. In a GAN, the generator creates samples that are classified by the discriminator as coming from the real data or the generator. Both generator and discriminator are optimized independently, until the discriminator can no longer discern between the artificial and the real data. Similarly, viruses mutate to escape the immune system while antibodies of the immune system "evolve" [1] to continue recognizing viruses. This could provide a new perspective on the conditions under which the desired broadly neutralizing antibodies (generalists) evolve [174].

The work in Chapter 4 started with the goal of predicting the antigen class given a cytokine response. We found that to optimally use the data, we required several processing steps, including log-transforming, interpolating missing data, removing noise, and fitting splines. This gave us access to an unlimited number of timepoints of cytokine concentration, as well as their integral and derivatives. We then found that integrals were most robust to experimental variability. We classified hourly sampled timepoints of integrals of cytokines one by one using an MLP with one hidden layer. To determine the antigen class of the entire timeseries, we applied the majority rule to the classified timepoints, which gave us a simple answer. We found that we can accurately determine antigen quality over two decades in antigen quantity. By changing the antigen quantity more than two decades, the effect of lower antigen quantity pushed the classification into the antigen class below.

We then wanted to understand the classification mechanism, specifically through the latent space. With a feature analysis, we found that the foundations of the integral latent space can be reproduced using IL-2 and IFNγ, cytokines that are produced majoritarily by cell types from the adaptive immune sytem and innate immune system, respectively. We then proceeded with parameterizing the integral latent space with piecewise ballistic models. The constant force model better reproduced the initial production phase than the constant velocity model, an observation we made in the concentration latent space. We then tested several experimental configurations to interpret the parameters F, t_0, θ, v_t. t_0 corresponds to a timescale of upregulation of IL-2 receptors, F is a convolution of antigen quality, quantity and precursor frequency, and determines the initiation of cytokine production. θ only becomes apparent when IL-17A is included and might indicate the ratio between activation of innate and adaptive immunity. Finally, v_t is the terminal speed, or the steady-state of all cytokines except IL-2. The first three parameters are strongly correlated with antigen quality, which means they are determined at the start of the experiment, similar to how the number of divisions of the T cells is determined at TCR-pMHC binding without additional

[1] More precisely, undergo B cell affinity maturation.

costimulation or cytokines [163]. Using drugs experiments, we were able to decorrelate F, t_0 and θ individually depending on the antigen quality.

Changing T cell type or APC type did not change the qualitative dynamics of timeseries in integral latent space, demonstrating that we found a universal latent space in which one can represent and interpret cytokine timeseries for a wide range of in-vitro T cell experiments. For instance, we found that CD4+ T cells were much more sensitive to antigen quantity than CD8+ T cells, and that we retrieve similar dynamics using antigen-expressing tumors instead of splenocytes. This opens up the possibility for experimentally determining salient neo-antigens on a tumor.

Finally, we predicted antigenicity quantitatively through EC_{50} instead of qualitatively through antigen class. We compared the prediction using an MLP trained on model parameters to linear fits on IFNγ and IL-2. At the same precursor frequency, accuracy for predicting with model parameters and with IL-2 was similar. However, when we tested on timeseries with a different number of precursors than we trained on, the prediction using model parameters remained good, while the linear fit on IL-2 became useless. This shows that the robotic platform and analysis pipeline is well-suited to predict antigenicity for any in-vitro setup, and gives more accurate and better reference results than current tests to predict antigenicity using Elispot. Advantages for using Elispot are that it is cheap, easy, and fast. Recently, FluoroSpot assays were proposed to measure multiple cytokines at once [147, 175], which while still measuring cytokines only at one timepoint, allows for multiplexing.

As an intermediate step towards more robustly predicting antigenicity, one could imagine formalizing the procedure for interpreting the number of dots and their sizes using machine learning. Despite precise guidelines, setting parameters for Elispot readers remains subjective work, depending on preparation and plate [146]. For that reason, automated comparison to a database of Elispot images with reference immunogenicities would be a welcome addition to predicting immunogenicities for a range of medical applications.

6

References

[1] Félix Proulx-Giraldeau, Thomas J Rademaker, and Paul François. “Untangling the Hairball: Fitness-Based Asymptotic Reduction of Biological Networks”. In: *Biophysical Journal* 113.8 (2017), pp. 1893–1906.

[2] Thomas J Rademaker, Emmanuel Bengio, and Paul François. “Attack and defense in cellular decision-making: lessons from machine learning”. In: *Physical Review X* 9.3 (2019), p. 031012.

[3] Sooraj Achar, Thomas J. Rademaker, François Bourassa, Paul François, and Grégoire Altan-Bonnet. “Learning the immune manifold from robotic cytokine multiplexing”. In preparation for submission. 2020.

[4] Yuri Lazebnik. “Can a biologist fix a radio?—Or, what I learned while studying apoptosis”. In: *Cancer cell* 2.3 (2002), pp. 179–182.

[5] Hiroaki Kitano. “Systems biology: a brief overview”. In: *science* 295.5560 (2002), pp. 1662–1664.

[6] Han-Yu Chuang, Matan Hofree, and Trey Ideker. “A decade of systems biology”. In: *Annual review of cell and developmental biology* 26 (2010), pp. 721–744.

[7] Christophe Benoist, Ronald N Germain, and Diane Mathis. “A plaidoyer for ‘systems immunology’”. In: *Immunological reviews* 210.1 (2006), pp. 229–234.

[8] Alan S Perelson, Avidan U Neumann, Martin Markowitz, John M Leonard, and David D Ho. "HIV-1 dynamics in vivo: virion clearance rate, infected cell lifespan, and viral generation time". In: *Science* 271.5255 (1996), pp. 1582–1586.

[9] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *Nature* 521.7553 (2015), p. 436.

[10] E Jonas and KP Kording. "Could a Neuroscientist Understand a Microprocessor?" In: *PLoS Computational Biology* 13.1 (2017), e1005268.

[11] John W Krakauer, Asif A Ghazanfar, Alex Gomez-Marin, Malcolm A MacIver, and David Poeppel. "Neuroscience needs behavior: correcting a reductionist bias". In: *Neuron* 93.3 (2017), pp. 480–490.

[12] David Marr and Tomaso Poggio. "From understanding computation to understanding neural circuitry". In: *AI Memo* 357 (1976).

[13] J Doyne Farmer, Norman H Packard, and Alan S Perelson. "The immune system, adaptation, and machine learning". In: *Physica D: Nonlinear Phenomena* 22.1-3 (1986), pp. 187–204.

[14] Stephanie Forrest, Alan S Perelson, Lawrence Allen, and Rajesh Cherukuri. "Self-nonself discrimination in a computer". In: *Proceedings of 1994 IEEE computer society symposium on research in security and privacy*. Ieee. 1994, pp. 202–212.

[15] Steven A Hofmeyr and Stephanie Forrest. "Architecture for an artificial immune system". In: *Evolutionary computation* 8.4 (2000), pp. 443–473.

[16] Thomas Höfer and Grégoire Altan-Bonnet. *Editorial overview: Systems immunology 2020: The art of putting it all together*. 2019.

[17] Kenneth Murphy and Casey Weaver. *Janeway's immunobiology*. Garland science, 2016.

[18] Grégoire Altan-Bonnet, Thierry Mora, and Aleksandra M Walczak. "Quantitative immunology for physicists". In: *Physics Reports* 849 (2020), pp. 1–83.

[19] Ludger Klein, Bruno Kyewski, Paul M Allen, and Kristin A Hogquist. “Positive and negative selection of the T cell repertoire: what thymocytes see (and don’t see)”. In: *Nature Reviews Immunology* 14.6 (2014), pp. 377–391.

[20] Don Mason. “A very high level of crossreactivity is an essential feature of the T-cell receptor”. In: *Immunology Today* 19.9 (1998), pp. 395–404.

[21] Jeffrey Ishizuka et al. “Quantitating T cell cross-reactivity for unrelated peptide antigens”. In: *The Journal of Immunology* 183.7 (2009), pp. 4337–4345.

[22] Andrew K Sewell. “Why must T cells be cross-reactive?” In: *Nature Reviews Immunology* 12.9 (2012), pp. 669–677.

[23] Grant Lythe, Robin E Callard, Rollo L Hoare, and Carmen Molina-Parıs. “How many TCR clonotypes does a body maintain?” In: *Journal of Theoretical Biology* 389 (2016), pp. 214–224.

[24] Andrew Yates. “Theories and quantification of thymic selection”. In: *Frontiers in Immunology* 5 (2014), p. 13.

[25] Ofer Feinerman, Ronald N Germain, and Grégoire Altan-Bonnet. “Quantitative challenges in understanding ligand discrimination by α β T cells”. In: *Molecular Immunology* 45.3 (2008), p. 619.

[26] Timothy W McKeithan. “Kinetic proofreading in T-cell receptor signal transduction”. In: *Proceedings of the National Academy of Sciences* 92.11 (1995), pp. 5042–5046.

[27] John J Hopfield. “Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity”. In: *Proceedings of the National Academy of Sciences* 71.10 (1974), pp. 4135–4139.

[28] Jacques Ninio. “Kinetic amplification of enzyme discrimination”. In: *Biochimie* 57.5 (1975), pp. 587–595.

[29] Paul Klenerman et al. “Cytotoxic T-cell activity antagonized by naturally occurring HIV-1 Gag variants”. In: *Nature* 369.6479 (1994), p. 403.

[30] Yousuke Takahama, Harumi Suzuki, Kenneth S Katz, Michael J Grusby, and Alfred Singer. "Positive selection of CD4+ T cells by TCR ligation without aggregation even in the absence of MHC". In: *Nature* 371.6492 (1994), pp. 67–70.

[31] Grégoire Altan-Bonnet and Ronald N Germain. "Modeling T cell antigen discrimination based on feedback control of digital ERK responses". In: *PLOS Biology* 3.11 (2005), e356.

[32] T Lipniacki, B Hat, JR Faeder, and WS Hlavacek. "Stochastic effects and bistability in T cell receptor signaling". In: *Journal of Theoretical Biology* 254.1 (2008), pp. 110–122.

[33] Paul François, Guillaume Voisinne, Eric D Siggia, Grégoire Altan-Bonnet, and Massimo Vergassola. "Phenotypic model for early T-cell activation displaying sensitivity, specificity, and antagonism". In: *Proceedings of the National Academy of Sciences* (2013), p. 201300752.

[34] Jean-Benoît Lalanne and Paul François. "Principles of adaptive sorting revealed by in silico evolution". In: *Physical Review Letters* 110.21 (2013), pp. 1–5.

[35] M Lever, PK Maini, PA van der Merwe, and O Dushek. "Phenotypic models of T cell activation". In: *Nature Reviews Immunology* 14.9 (2014), pp. 619–629.

[36] Melissa Lever et al. "Architecture of a minimal signaling pathway explains the T-cell response to a 1 million-fold variation in antigen affinity and dose". In: *Proceedings of the National Academy of Sciences* 113.43 (2016), E6630–E6638.

[37] Doug K Tischer and Orion David Weiner. "Light-based tuning of ligand half-life supports kinetic proofreading model of T cell signaling". In: *Elife* 8 (2019), e42498.

[38] Jesse Goyette et al. "Regulated unbinding of ZAP70 at the T cell receptor by kinetic avidity". In: *BioRxiv* (2020).

[39] Raman S Ganti, Wan-Lin Lo, Darren B McAffee, Jay T Groves, Arthur Weiss, and Arup K Chakraborty. "How the T cell signaling network processes information to discriminate between self and agonist ligands". In: *Proceedings of the National Academy of Sciences* 117.42 (2020), pp. 26020–26030.

[40] P Anton Van Der Merwe and Omer Dushek. “Mechanisms for T cell receptor triggering”. In: *Nature Reviews Immunology* 11.1 (2011), pp. 47–55.

[41] John R James and Ronald D Vale. “Biophysical mechanism of T-cell receptor triggering in a reconstituted system”. In: *Nature* 487.7405 (2012), pp. 64–69.

[42] Jinsung Hong et al. “A TCR mechanotransduction signaling loop induces negative selection in the thymus”. In: *Nature immunology* 19.12 (2018), pp. 1379–1390.

[43] Arup K Chakraborty and Arthur Weiss. “Insights into the initiation of TCR signaling”. In: *Nature immunology* 15.9 (2014), p. 798.

[44] Omer Dushek, Raibatak Das, and Daniel Coombs. “A role for rebinding in rapid and reliable T cell responses to antigen”. In: *PLoS Comput Biol* 5.11 (2009), e1000578.

[45] Omer Dushek and P Anton van der Merwe. “An induced rebinding model of antigen discrimination”. In: *Trends in immunology* 35.4 (2014), pp. 153–158.

[46] Anna Huhn, Daniel B Wilson, P Anton van der Merwe, and Omer Dushek. “The discriminatory power of the T cell receptor”. In: *bioRxiv* (2020).

[47] Yunqian Ma and Yun Fu. *Manifold learning theory and applications*. CRC press, 2011.

[48] Cliff P Burgess. “An introduction to effective field theory”. In: *Annu. Rev. Nucl. Part. Sci.* 57 (2007), pp. 329–362.

[49] Paul François and Grégoire Altan-Bonnet. “The case for absolute ligand discrimination: modeling information processing and decision by immune T cells”. In: *Journal of Statistical Physics* 162.5 (2016), pp. 1130–1152.

[50] BB Machta, R Chachra, MK Transtrum, and JP Sethna. “Parameter space compression underlies emergent theories and predictive models”. In: *Science* 342.6158 (2013), pp. 604–607.

[51] MK Transtrum and P Qiu. “Model reduction by manifold boundaries”. In: *Physical Review Letters* 113.9 (2014), pp. 098701–098701.

[52] MK Transtrum and P Qiu. "Bridging Mechanistic and Phenomenological Models of Complex Biological Systems". In: *PLoS Computational Biology* 12.5 (2016), e1004915.

[53] AJ Krol, D Roellig, ML Dequeant, O Tassy, E Glynn, G Hattem, and et al. "Evolutionary plasticity of segmentation clock networks". In: *Development* 138.13 (2011), pp. 2783–2792.

[54] P François. "Evolving phenotypic networks in silico". In: *Seminars in cell & developmental biology* 35 (2014), pp. 90–97.

[55] BC Daniels and I Nemenman. "Automated adaptive inference of phenomenological dynamical models". In: *Nature Communications* 6 (2015), p. 8133.

[56] BC Daniels and I Nemenman. "Efficient inference of parsimonious phenomenological models of cellular dynamics using S-systems and alternating regression". In: *PLoS ONE* 10.3 (2015), e0119821.

[57] Chikako Torigoe, John K Inman, and Henry Metzger. "An unusual mechanism for ligand antagonism". In: *Science* 281.5376 (1998), pp. 568–572.

[58] MJ Taylor, K Husain, ZJ Gartner, S Mayor, and RD Vale. "A DNA-Based T Cell Receptor Reveals a Role for Receptor Clustering in Ligand Discrimination". In: *Cell* 169.1 (2017), 108–119.e20.

[59] TM Yi, Y Huang, MI Simon, and J Doyle. "Robust perfect adaptation in bacterial chemotaxis through integral feedback control". In: *Proceedings of the National Academy of Sciences* 97.9 (2000), pp. 4649–4653.

[60] Paul François, Mathieu Hemery, Kyle A Johnson, and Laura N Saunders. "Phenotypic spandrel: absolute discrimination and ligand antagonism". In: *Physical Biology* 13.6 (2016), p. 066011.

[61] François P and Siggia ED. "A case study of evolutionary computation of biochemical adaptation". In: *Physical Biology* 5.2 (2008), p. 26009.

[62] W Ma, A Trusina, H El-Samad, WA Lim, and C Tang. "Defining network topologies that can achieve biochemical adaptation". In: *Cell* 138.4 (2009), pp. 760–773.

[63] Bonnie N Dittel, Ronald N Germain, Charles A Janeway Jr, et al. “Cross-antagonism of a T cell clone expressing two distinct T cell receptors”. In: *Immunity* 11.3 (1999), pp. 289–298.

[64] O Feinerman, J Veiga, JR Dorfman, RN Germain, and G Altan-Bonnet. “Variability and Robustness in T Cell Activation from Regulated Heterogeneity in Protein Levels”. In: *Science* 321.5892 (2008), pp. 1081–1084.

[65] MK Transtrum, G Hart, and P Qiu. “Information topology identifies emergent model classes”. In: *arXiv preprint arXiv:1409.6203* (2014).

[66] F Corson and ED Siggia. “Geometry and epistasis and and developmental patterning”. In: *Proceedings of the National Academy of Sciences* 109.15 (2012), pp. 5568–5575.

[67] H Innan and F Kondrashov. “The evolution of gene duplications: classifying and distinguishing between models”. In: *Nature Review Genetics* 11.4 (2010), p. 4.

[68] R Milo, S Shen-Orr, S Itzkovitz, N Kashtan, D Chklovskii, and U Alon. “Network motifs: simple building blocks of complex networks”. In: *Science* 298.5594 (2002), pp. 824–827.

[69] D Sussillo and LF Abbott. “Generating coherent patterns of activity from chaotic neural networks”. In: *Neuron* 63.4 (2009), pp. 544–557.

[70] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. “Intriguing properties of neural networks”. In: *arXiv preprint arXiv:1312.6199* (2013).

[71] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. “Delving into transferable adversarial examples and black-box attacks”. In: *arXiv preprint arXiv:1611.02770* (2016).

[72] S. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. “Universal Adversarial Perturbations”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2017, pp. 86–94.

[73] Ali Shafahi, W Ronny Huang, Christoph Studer, Soheil Feizi, and Tom Goldstein. “Are adversarial examples inevitable?” In: *arXiv preprint arXiv:1809.02104* (2018).

[74] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. “Explaining and Harnessing Adversarial Examples”. In: *arXiv preprint arXiv:1412.6572* (2014).

[75] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Advances in Neural Information Processing Systems*. 2012, pp. 1097–1105.

[76] Geoffrey Hinton et al. “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups”. In: *IEEE Signal processing magazine* 29.6 (2012), pp. 82–97.

[77] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. “Sequence to sequence learning with neural networks”. In: *Advances in Neural Information Processing Systems*. 2014, pp. 3104–3112.

[78] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. “Practical black-box attacks against machine learning”. In: *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*. ACM. 2017, pp. 506–519.

[79] Samuel G Finlayson, John D Bowers, Joichi Ito, Jonathan L Zittrain, Andrew L Beam, and Isaac S Kohane. “Adversarial attacks on medical machine learning”. In: *Science* 363.6433 (2019), pp. 1287–1289.

[80] Naveed Akhtar and Ajmal Mian. “Threat of adversarial attacks on deep learning in computer vision: A survey”. In: *arXiv preprint arXiv:1801.00553* (2018).

[81] Eric D Siggia and Massimo Vergassola. “Decisions on the fly in cellular sensory systems”. In: *Proceedings of the National Academy of Sciences* 110.39 (2013), E3704–E3712.

[82] Nicholas RJ Gascoigne, Tomasz Zal, and S Munir Alam. “T-cell receptor binding kinetics in T-cell development and activation”. In: *Expert Reviews in Molecular Medicine* 3.6 (2001), pp. 1–17.

[83] Gautam Reddy, Joseph D Zak, Massimo Vergassola, and Venkatesh N Murthy. "Antagonism in olfactory receptor neurons and its implications for the perception of odor mixtures". In: *eLife* 7 (2018), e34958.

[84] Julia Tsitron, Addison D Ault, James R Broach, and Alexandre V Morozov. "Decoding complex chemical mixtures with a physical model of a sensor array." In: *PLoS Computational Biology* 7.10 (2011), e1002224–e1002224.

[85] Ute-Christiane Meier et al. "Cytotoxic T lymphocyte lysis inhibited by viable HIV mutants". In: *Science* 270.5240 (1995), pp. 1360–1362.

[86] Stephen J Kent, Philip D Greenberg, Mark C Hoffman, Robert E Akridge, and M Juliana McElrath. "Antagonism of vaccine-induced HIV-1-specific CD4+ T cells by primary HIV-1 infection: potential mechanism of vaccine failure." In: *The Journal of Immunology* 158.2 (1997), pp. 807–815.

[87] Alexandra Snyder et al. "Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma". In: *N Engl J Med* 371.23 (2014), pp. 2189–2199.

[88] Ton N Schumacher and Robert D Schreiber. "Neoantigens in cancer immunotherapy". In: *Science* 348.6230 (2015), pp. 69–74.

[89] Kathrin Grosse, Praveen Manoharan, Nicolas Papernot, Michael Backes, and Patrick McDaniel. "On the (statistical) detection of adversarial examples". In: *arXiv preprint arXiv:1702.06280* (2017).

[90] Eric Wong and Zico Kolter. "Provable defenses against adversarial examples via the convex outer adversarial polytope". In: *International Conference on Machine Learning*. 2018, pp. 5283–5292.

[91] Dmitry Krotov and John J Hopfield. "Dense associative memory for pattern recognition". In: *Advances in Neural Information Processing Systems*. 2016, pp. 1172–1180.

[92] Jayajit Das. "Activation or Tolerance of Natural Killer Cells Is Modulated by Ligand Quality in a Nonmonotonic Manner". In: *Biophysical Journal* 99.7 (2010), pp. 2028–2037.

[93] Fabien Lagarde, Claire Beausoleil, Scott M Belcher, Luc P Belzunces, Claude Emond, Michel Guerbet, and Christophe Rousselle. "Non-monotonic dose-response relationships and endocrine disruptors: a qualitative method of assessment". In: *Environmental Health* 14.1 (2015), 13–a106.

[94] Christopher C Govern, Michelle K Paczosa, Arup K Chakraborty, and Eric S Huseby. "Fast on-rates allow short dwell time ligands to activate T cells". In: *Proceedings of the National Academy of Sciences* (2010), p. 201000966.

[95] Gilbert J Kersh, Ellen N Kersh, Daved H Fremont, and Paul M Allen. "High-and low-potency ligands with similar affinities for the TCR: the importance of kinetics in TCR signaling". In: *Immunity* 9.6 (1998), pp. 817–826.

[96] Jean-Benoît Lalanne and Paul François. "Chemodetection in fluctuating environments: Receptor coupling, buffering, and antagonism". In: *Proceedings of the National Academy of Sciences* 112.6 (2015), pp. 1898–1903.

[97] Thierry Mora. "Physical limit to concentration sensing amid spurious ligands". In: *Physical Review Letters* 115.3 (2015), p. 038102.

[98] Martin Carballo-Pacheco et al. "Receptor crosstalk improves concentration sensing of multiple ligands." In: *Physical Review E* 99.2-1 (2019), p. 022423.

[99] Ronald N Germain and Irena Stefanová. "The dynamics of T cell receptor signaling: complex orchestration and the key roles of tempo and cooperation". In: *Annual Review of Immunology* 17.1 (1999), pp. 467–522.

[100] Yann LeCun and Corinna Cortes. "The MNIST database of handwritten digits". In: (http://yann.lecun.com/exdb/mnist/ 1998).

[101] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[102] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. "Robustness may be at odds with accuracy". In: *arXiv preprint arXiv:1805.12152* 1 (2018).

[103] Dmitry Krotov and John J Hopfield. “Dense associative memory is robust to adversarial inputs”. In: *Neural computation* (2018), pp. 1–17.

[104] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. “Adversarial machine learning at scale”. In: *arXiv preprint arXiv:1611.01236* (2016).

[105] Thomas Tanay and Lewis Griffin. “A boundary tilting persepective on the phenomenon of adversarial examples”. In: *arXiv preprint arXiv:1608.07690* (2016).

[106] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. “One pixel attack for fooling deep neural networks”. In: *IEEE Transactions on Evolutionary Computation* 23.5 (2019), pp. 828–841.

[107] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. “Feature denoising for improving adversarial robustness”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 501–509.

[108] David Berthelot, Colin Raffel, Aurko Roy, and Ian Goodfellow. “Understanding and Improving Interpolation in Autoencoders via an Adversarial Regularizer”. In: *arXiv preprint arXiv:1807.07543* (2018).

[109] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, and Pascal Frossard. “Robustness via curvature regularization, and vice versa”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 9078–9086.

[110] Dmitry Krotov and John J Hopfield. “Unsupervised learning by competing hidden units”. In: *Proceedings of the National Academy of Sciences* (2019), p. 201820458.

[111] Gamaleldin Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alexey Kurakin, Ian Goodfellow, and Jascha Sohl-Dickstein. “Adversarial Examples that Fool both Computer Vision and Time-Limited Humans”. In: *Advances in Neural Information Processing Systems*. 2018, pp. 3911–3921.

[112] Emil R Unanue. “Altered peptide ligands make their entry”. In: *The Journal of Immunology* 186.1 (2011), pp. 7–8.

[113] Thomas C Butler, Mehran Kardar, and Arup K Chakraborty. “Quorum sensing allows T cells to discriminate between self and nonself”. In: *Proceedings of the National Academy of Sciences* 110.29 (2013), pp. 11833–11838.

[114] Guillaume Voisinne, Briana G Nixon, Anna Melbinger, Georg Gasteiger, Massimo Vergassola, and Grégoire Altan-Bonnet. “T cells integrate local and global cues to discriminate between structurally similar antigens”. In: *Cell Reports* 11.8 (2015), pp. 1208–1219.

[115] Judith N Mandl, João P Monteiro, Nienke Vrisekoop, and Ronald N Germain. “T Cell-Positive Selection Uses Self-Ligand Binding Strength to Optimize Repertoire Recognition of Foreign Antigens”. In: *Immunity* 38.2 (2013), pp. 263–274.

[116] Marta Łuksza et al. “A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy”. In: *Nature* 551.7681 (2017), pp. 517–520.

[117] Ugur Sahin and Özlem Türeci. “Personalized vaccines for cancer immunotherapy”. In: *Science* 359.6382 (2018), pp. 1355–1360.

[118] Jinyuan Yan et al. “Bow-tie signaling in c-di-GMP: Machine learning in a simple biochemical network”. In: *PLOS Computational Biology* 13.8 (2017), e1005677.

[119] Andres Laan and Gonzalo de Polavieja. “Sensory cheating: adversarial body patterns can fool a convolutional visual system during signaling”. In: *bioRxiv* (2018), p. 326652.

[120] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. “Synthesizing Robust Adversarial Examples”. In: *Proceedings of the 35th International Conference on Machine Learning*. Vol. 80. 2018, pp. 284–293.

[121] Kevin Eykholt et al. “Robust Physical-World Attacks on Deep Learning Models”. In: *Proceedings of Conference on Computer Vision and Pattern Recognition*. 2018.

[122] Martin Meier-Schellersheim, Rajat Varma, and Bastian Robert Angermann. “Mechanistic Models of Cellular Signaling, Cytokine Crosstalk and Cell-Cell Communication in Immunology”. In: *Frontiers in Immunology* 10 (2019), p. 2268.

[123] Meixiao Long and Adam J Adler. “Cutting edge: Paracrine, but not autocrine, IL-2 signaling is sustained during early antiviral CD4 T cell response”. In: *The Journal of Immunology* 177.7 (2006), pp. 4257–4261.

[124] Karen E Tkach et al. “T cells translate individual, quantal activation into collective, analog cytokine responses via time-integrated feedbacks”. In: *eLife* 3 (2014), e01944.

[125] Grégoire Altan-Bonnet and Ratnadeep Mukherjee. “Cytokine-mediated communication: a quantitative appraisal of immune complexity”. In: *Nature Reviews Immunology* 19.4 (2019), pp. 205–217.

[126] Charles A Dinarello. “Historical insights into cytokines”. In: *European Journal of Immunology* 37.S1 (2007), S34–S45.

[127] Stephen R Holdsworth and Poh-Yi Gan. “Cytokines: names and numbers you should care about”. In: *Clinical journal of the American Society of Nephrology* 10.12 (2015), pp. 2243–2254.

[128] Wei Liao, Jian-Xin Lin, and Warren J Leonard. “IL-2 family cytokines: new insights into the complex roles of IL-2 as a broad regulator of T helper cell differentiation”. In: *Current Opinion in Immunology* 23.5 (2011), pp. 598–604.

[129] Camille Zenobia and George Hajishengallis. “Basic biology and role of interleukin-17 in immunity and inflammation”. In: *Periodontology 2000* 69.1 (2015), pp. 142–159.

[130] Masanori Hatakeyama et al. “Interleukin-2 receptor beta chain gene: generation of three receptor forms by cloned human alpha and beta chain cDNA’s”. In: *Science* 244.4904 (1989), pp. 551–556.

[131] Alfons Billiau. “Interferon-γ: biology and role in pathogenesis”. In: *Advances in Immunology*. Vol. 62. Elsevier, 1996, pp. 61–130.

[132] Ofer Feinerman et al. “Single-cell quantification of IL-2 response by effector and regulatory T cells reveals critical plasticity in immune response”. In: *Molecular systems biology* 6.1 (2010), p. 437.

[133] Dorothea Busse, Maurus de la Rosa, Kirstin Hobiger, Kevin Thurley, Michael Flossdorf, Alexander Scheffold, and Thomas Höfer. “Competing feedback loops shape IL-2 signaling between helper and regulatory T lymphocytes in cellular microenvironments”. In: *Proceedings of the National Academy of Sciences* 107.7 (2010), pp. 3058–3063.

[134] Thomas Höfer, Oleg Krichevsky, and Grégoire Altan-Bonnet. “Competition for IL-2 between regulatory and effector T cells to chisel immune responses”. In: *Frontiers in immunology* 3 (2012), p. 268.

[135] Juan Quiel et al. “Antigen-stimulated CD4 T-cell expansion is inversely and log-linearly related to precursor number”. In: *Proceedings of the National Academy of Sciences* 108.8 (2011), pp. 3312–3317.

[136] Gennady Bocharov et al. “Feedback regulation of proliferation vs. differentiation rates explains the dependence of CD4 T-cell expansion on precursor number”. In: *Proceedings of the National Academy of Sciences* 108.8 (2011), pp. 3318–3323.

[137] Rob J De Boer and Alan S Perelson. “Antigen-Stimulated CD4 T Cell Expansion Can Be Limited by Their Grazing of Peptide–MHC Complexes”. In: *The Journal of Immunology* 190.11 (2013), pp. 5454–5458.

[138] Andreas Mayer, Yaojun Zhang, Alan S Perelson, and Ned S Wingreen. “Regulation of T cell expansion by antigen presentation dynamics”. In: *Proceedings of the National Academy of Sciences* 116.13 (2019), pp. 5914–5919.

[139] Vinod P Balachandran et al. “Identification of unique neoantigen qualities in long-term survivors of pancreatic cancer”. In: *Nature* 551.7681 (2017), pp. 512–516.

[140] Frank Rosenblatt. “The perceptron: a probabilistic model for information storage and organization in the brain.” In: *Psychological review* 65.6 (1958), p. 386.

[141] Leo Breiman. “Random forests”. In: *Machine learning* 45.1 (2001), pp. 5–32.

[142] Mikkel Harndahl, Michael Rasmussen, Gustav Roder, Ida Dalgaard Pedersen, Mikael Sørensen, Morten Nielsen, and Søren Buus. “Peptide-MHC class I stability is a better predictor than peptide affinity of CTL immunogenicity”. In: *European journal of immunology* 42.6 (2012), pp. 1405–1416.

[143] Vanessa Jurtz, Sinu Paul, Massimo Andreatta, Paolo Marcatili, Bjoern Peters, and Morten Nielsen. "NetMHCpan-4.0: improved peptide–MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data". In: *The Journal of Immunology* 199.9 (2017), pp. 3360–3368.

[144] Randi Vita et al. "The immune epitope database (IEDB): 2018 update". In: *Nucleic acids research* 47.D1 (2019), pp. D339–D343.

[145] Cecil C Czerkinsky, Lars-Åke Nilsson, Håkan Nygren, Örjan Ouchterlony, and Andrej Tarkowski. "A solid-phase enzyme-linked immunospot (ELISPOT) assay for enumeration of specific antibody-secreting cells". In: *Journal of immunological methods* 65.1-2 (1983), pp. 109–121.

[146] Sylvia Janetzki, Leah Price, Helene Schroeder, Cedrik M Britten, Marij JP Welters, and Axel Hoos. "Guidelines for the automated evaluation of Elispot assays". In: *Nature Protocols* 10.7 (2015), p. 1098.

[147] Niklas Ahlborg and Bernt Axelsson. "Dual-and triple-color fluorospot". In: *Handbook of ELISPOT*. Springer, 2012, pp. 77–85.

[148] Mark A Daniels et al. "Thymic selection threshold defined by compartmentalization of Ras/MAPK signalling". In: *Nature* 444.7120 (2006), pp. 724–729.

[149] Dietmar Zehn, Sarah Y Lee, and Michael J Bevan. "Complete but curtailed T-cell response to very low-affinity antigen". In: *Nature* 458.7235 (2009), pp. 211–214.

[150] François Bourassa. "Decoding cytokine dynamics with biochemical networks". MA thesis. McGill University, 2020.

[151] Mathieu Hemery and Paul François. "In silico evolution of biochemical log-response". In: *The Journal of Physical Chemistry B* 123.10 (2019), pp. 2235–2243.

[152] Adrien Henry, Mathieu Hemery, and Paul François. "φ-evo: A program to evolve phenotypic models of biological networks". In: *PLoS Computational Biology* 14.6 (2018), e1006244.

[153] Lauro Velazquez-Salinas, Antonio Verdugo-Rodriguez, Luis L Rodriguez, and Manuel V Borca. “The role of interleukin 6 during viral infections”. In: *Frontiers in Microbiology* 10 (2019), p. 1057.

[154] Narayanan Parameswaran and Sonika Patial. “Tumor necrosis factor-α signaling in macrophages”. In: *Critical Reviews™ in Eukaryotic Gene Expression* 20.2 (2010), pp. 87–103.

[155] Pauli Virtanen et al. “SciPy 1.0: fundamental algorithms for scientific computing in Python”. In: *Nature Methods* 17.3 (2020), pp. 261–272.

[156] MK Transtrum, BB Machta, KS Brown, BC Daniels, CR Myers, and JP Sethna. “Perspective: Sloppiness and emergent theories in physics and biology and and beyond”. In: *The Journal of Chemical Physics* 143.1 (2015), p. 010901.

[157] Chieh-Ting Jimmy Hsu, Gary J Brouhard, and Paul François. “Numerical parameter space compression and its application to biophysical models”. In: *Biophysical Journal* 118.6 (2020), pp. 1455–1465.

[158] Laura M McLane, Mohamed S Abdel-Hakeem, and E John Wherry. “CD8 T cell exhaustion during chronic viral infection and cancer”. In: *Annual review of immunology* 37 (2019), pp. 457–495.

[159] Christopher D Scharer, Alexander PR Bally, Bhanu Gandham, and Jeremy M Boss. “Cutting edge: chromatin accessibility programs CD8 T cell memory”. In: *The Journal of Immunology* 198.6 (2017), pp. 2238–2243.

[160] Ralph M Steinman. “Decisions about dendritic cells: past, present, and future”. In: *Annual Review of Immunology* 30 (2012), pp. 1–22.

[161] Alberta Di Pasquale, Scott Preiss, Fernanda Tavares Da Silva, and Nathalie Garçon. “Vaccine adjuvants: from 1920 to 2015 and beyond”. In: *Vaccines* 3.2 (2015), pp. 320–343.

[162] Ling Zhang et al. “Enhanced efficacy and limited systemic cytokine exposure with membrane-anchored interleukin-12 T-cell therapy in murine tumor models”. In: *Journal for immunotherapy of cancer* 8.1 (2020).

[163] Julia M Marchingo et al. “Antigen affinity, costimulation, and cytokine inputs sum linearly to amplify T cell expansion”. In: *Science* 346.6213 (2014), pp. 1123–1127.

[164] Yona Kalechman, Uzi Gafter, Ji Ping Da, Michael Albeck, Donato Alarcon-Segovia, and Benjamin Sredni. “Delay in the onset of systemic lupus erythematosus following treatment with the immunomodulator AS101: association with IL-10 inhibition and increase in TNF-alpha levels.” In: *The Journal of Immunology* 159.6 (1997), pp. 2658–2667.

[165] Karlo Perica, Juan Carlos Varela, Mathias Oelke, and Jonathan Schneck. “Adoptive T cell immunotherapy for cancer”. In: *Rambam Maimonides medical journal* 6.1 (2015).

[166] Aude G Chapuis et al. “Combined IL-21–primed polyclonal CTL plus CTLA4 blockade controls refractory metastatic melanoma in a patient”. In: *Journal of Experimental Medicine* 213.7 (2016), pp. 1133–1139.

[167] Marta Del Olmo, Achim Kramer, and Hanspeter Herzel. “A robust model for circadian redox oscillations”. In: *International Journal of Molecular Sciences* 20.9 (2019), p. 2368.

[168] J Patrick Pett, Anja Korenčič, Felix Wesener, Achim Kramer, and Hanspeter Herzel. “Feedback loops of the mammalian circadian clock constitute repressilator”. In: *PLoS computational biology* 12.12 (2016), e1005266.

[169] Mohan K Malleshaiah, Vahid Shahrezaei, Peter S Swain, and Stephen W Michnick. “The scaffold protein Ste5 directly controls a switch-like mating decision in yeast”. In: *Nature* 465.7294 (2010), pp. 101–105.

[170] Alejandro A Granados, Julian MJ Pietsch, Sarah A Cepeda-Humerez, Iseabail L Farquhar, Gašper Tkačik, and Peter S Swain. “Distributed and dynamic intracellular organization of extracellular information”. In: *Proceedings of the National Academy of Sciences* 115.23 (2018), pp. 6088–6093.

[171] Inge Wortel, Can Keşmir, Rob J de Boer, Judith N Mandl, and Johannes Textor. “Is T Cell Negative Selection a Learning Algorithm?” In: *Cells* 9.3 (2020), p. 690.

[172] Takuya Kato and Tetsuya J Kobayashi. “Understanding Adaptive Immune System as Reinforcement Learning”. In: *bioRxiv* (2020).

[173] Ian Goodfellow et al. “Generative adversarial nets”. In: *Advances in Neural Information Processing Systems*. 2014, pp. 2672–2680.

[174] Shenshen Wang and Lei Dai. “Evolving generalists in switching rugged landscapes”. In: *PLoS computational biology* 15.10 (2019), e1007320.

[175] Sylvia Janetzki, Markus Rueger, and Tomas Dillenbeck. “Stepping up ELISpot: multi-level analysis in FluoroSpot assays”. In: *Cells* 3.4 (2014), pp. 1102–1115.

A

Parameter reduction

Description of $\bar{\phi}$

MATLAB Algorithm

“ In the following, we illustrate $\bar{\phi}$’s MATLAB implementation. The following functions are used

- runLoop
- paramMODELtype
- calcFitness
- odeMODELtype
- evalLim
- updateParam
- catch_problems.

runLoop is the main script. paramMODELtype and odeMODELtype define parameters and associated ODEs which are problem specific. Parameters associated to the model are initially stored in variable *default*, then later modified parameters are stored in variable *param* and the list of removed parameters is stored in variable *removed*

Parameter reduction

A flowchart of the algorithm is presented in Fig. SA.1. In the following five steps we probe (1 - 3), rank, select, evaluate, accept (4), reduce and repeat (5).

1. Assign the parameter vector (PV) that paramMODELtype returns to *default.* This point in parameter space is going to be probed.

2. The fitness landscape around the initial PV *default* is characterized by the symmetric matrix *fitnessmap*, containing the fitness for modified parameters or couples of parameters. The fitness function, on the contrary, is problem specific, and is computed by calcFitness. Row by row, *fitnessmap* is filled by multiplying/dividing the parameters per entry by a reacaling *factor* ($f = 10$ from the main text). The performance of each of these entries is measured by computing the fitness with the new parameter combination relative to the initial fitness. A network with N parameters has $2N^2$ independent entries.

3. Removing a parameter is done in evalLim. With an estimate of the fitness landscape at hand found via the previous steps, the algorithm takes the corresponding limit (to 0 or ∞) for the parameters that were rescaled by f. We consider only changes of parameters giving identical or improved fitness. There exist four groups of two parameter limits θ_i, θ_j. In Table A.1, the groups are presented in order of importance. When several couples of parameters give favorable changes to the fitness, we evaluate the limit of all couples that fall in group 1 one by one.

4. When we encounter a parameter limit in which the fitness is improved, we eliminate corresponding parameters and return to step 1. If for none of the couples in the parameter limits the fitness is improved, we move to the members of group 2, the limits to infinity, and similarly when we find a parameter limit that improves the fitness, we reduce and move on. Otherwise we move to the parameter limits of groups 3 and, finally to group 4 with the same criteria. This natural order shows our preference for removing parameters one by one (set parameters values to zero), instead of simply rescaling them (as products). Notice that we take a very conservative approach where fitness can only be incrementally improved with this procedure.

The steps in evalLim are as follows:

- Find the least nonnegative elements in *fitnessmap*

Table A.1: Four groups of two-parameter limits

Group	Operation	Corresponding Limit taken
1	Division of two parameters by f	$(\theta_i, \theta_j) \to 0$
2	Multiplication of two parameters by f	$(\theta_i, \theta_j) \to \infty$
3	Division/multiplication by f	$\theta_i \to 0, \theta_j \to \infty$
4	Division/multiplication by f	Rescaling keeping product $\theta_i \cdot \theta_j = \text{constant}$

- Divide these in the groups defined above
- Pick a random element from the highest ranked nonempty group
- updateParam takes the PV *default* and a 2×2 block of *removed* as arguments and returns an updated PV to *param.*
- Compute a temporary fitness ϕ_{new} with *param.*
- Decide as follows:

```
If $\phi_{\text{new}} \geq \phi_{\text{init}}$.
    Accept removal
    Return *param* and *removed*
If $\phi_{\text{new}} < \phi_{\text{init}}$.
    Reject removal
    Set *fitnessmap*(picked element) to $\inf$.
    Repeat cycle at step A
```

The method we use to take asymptotic limits is described in the next section.

5. The returned PV becomes the new initial point in an $(N-1)$-dimensional plane that is embedded in N-dimensional parameter space. Around this new initial point, we will probe the fitness landscape in the next round. In *removed*, the removed parameters and their limits are stored such that $\bar{\phi}$ ignores directions of reduced parameters in subsequent rounds.

This procedure is repeated until no free parameters are left or until all directions will decrease the fitness.

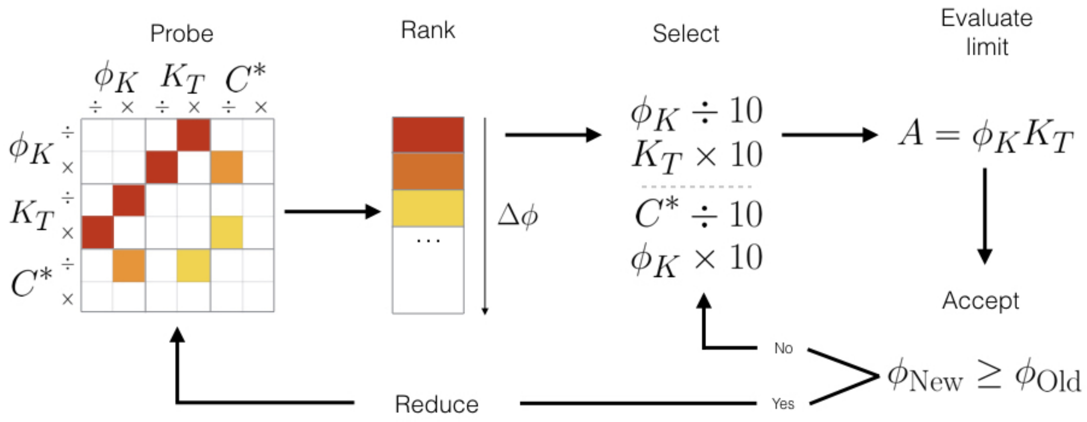


Figure A.1: **Flowchart of the algorithm.**

Taking asymptotic limits

There are two kinds of asymptotic limits: parameters are either taken to 0 or to ∞. The 0 case is trivial to deal with: when a parameter is chosen to be 0, we simply put and maintain it to 0 in the subsequent steps of $\bar{\phi}$.

In evaluating a limit to infinity, one cannot simply numerically set this parameter to infinity, like in the case of a zero-limit. Instead, we consider a limit where this parameter is increased to such an extend that it dominates other terms in a sum that affect the same variable; these other terms are then removed from the equations. More precisely, consider the following equation:

$$\dot{y}_2 = ay_1 - (b + c + dy_1)y_2. \tag{A.1}$$

In the limit of $b \to \infty$ we replace this equation by the following differential equation:

$$\dot{y}_2 = ay_1 - by_2, \tag{A.2}$$

where $b \to b' = fb$, where f is our multiplicative factor defined in the previous section. This implements the idea that the c and dy_1 terms are negligible compared to b.

It is important to define a vector of parameter coefficients to keep track of these infinities. The vector of coefficients is attached to the parameter vector and updated in updateParam similarly. When the limit of a parameter is taken to infinity, its coefficient becomes zero, and the other terms in the sum will disappear. Practically, Eq. A.1 is rewritten as

$$\dot{y}_2 = c_d a y_1 - (c_c c_d b + c_b c_d c + c_a c_b c_c d y_1) y_2. \tag{A.3}$$

The coefficients $c_{a,b,c,d}$ are initially set to 1. After evaluating the limit of $b \to \infty$, we set $c_b = 0$, and the simplification from Eq. A.1 to Eq. A.2 indeed takes place.

This can however create mass conservation problems in the rate equations. Consider the following equations for $\dot{y}_4$ and $\dot{y}_5$ where y_4 is turned into y_5 with rate r

$$\begin{aligned} \dot{y}_4 &= a y_3 - (r + q) y_4 \\ \dot{y}_5 &= r y_4 - d y_5 \end{aligned} \tag{A.4}$$

In the limit where parameter $q \to \infty$, parameter r will disappear from the equation of $\dot{y}_4$ potentially creating a divergence in the equations. A way to circumvent this is to impose global mass conservation: situations where y_4 is turned into y_5 correspond to signalling cascades where complexes are transformed into one another, so that we can impose that the total quantity of complex is conserved. This effectively adds a compensating term to the cascade. We also explicitly control for divergences and discard parameter sets for which variables diverge.

Choice of the path in parameter space

As shown in Section A, the matrix *fitnessmap* is analyzed in the function evalLim. This matrix is symmetrical since the upper triangular part of the matrix corresponding to parameters (k_1, k_2) and the lower triangular part corresponding to parameters (k_2, k_1) give similar limits for groups 2 and 4 in Table A.1. When given the choice between sending $(k_1, k_2) \to \infty$ or $(k_2, k_1) \to \infty$, FIBAR chooses randomly between the two, because the parameter combinations have the same change in fitness and in both cases a new parameter k_1/k_2 can be identified. However, because of FIBAR's design, choosing one will result in a different exploration of parameter space in the remaining steps. By choosing the first parameter combination, $\bar{\phi}$ will effectively freeze k_1 but allows $\bar{\phi}$ to keep exploring the logarithmic neighborhood of k_2. If the second combination is chosen, then the value of k_2 is frozen and it is the neighborhood of k_1 that will be probed. k_2 and k_1 may be present in different equations in the model, resulting in two not necessarily converging reductions.

A choice thus needs to be made in the final parameter reduced model. This allows for introduction of some kind of stochasticity in the produced networks in order to identify recurring patterns in the reduction. It can be a challenge in terms of reproducibility. One way to solve this problem is to set a fixed rule in the function evalLim (using variable seed) which is called the deterministic method in the main article. The method of choice (random or deterministic) is left at the discretion of the user. We indeed see differences in the way networks are reduced, but the final structure of the reduced networks in all these cases can easily be mapped onto one another as described in the main text. ” *(Parameter reduction [1])*

Adaptive Sorting

“ We perform parameter reduction on the Adaptative Sorting model without any symmetry breaking process. Initial equations for the adaptive sorting model are given by

$$\begin{aligned}
\dot{K} &= \beta(K_T - K) - \alpha K C_0 \\
\dot{C}_0 &= \kappa(L - \sum_i C_i)(R - \sum_i C_i) + bC_1 - (\phi K + \tau^{-1})C_0 \\
\dot{C}_1 &= \phi K C_0 - (\tau^{-1} + b)C_1.
\end{aligned}$$

Initial parameters are given in Table A.2. Steps of the reduction of this model are given in Table A.3.

Table A.2: Adaptative sorting initial parameters

Parameter	Value
ϕ	3×10^{-4}
K_T	10^3
α	1
β	1
κ	10^{-4}
R	10^4
b	5×10^{-2}

Table A.3: Adaptive sorting

Step	I_{init}	Parameters		Limit	Description per group
1	0.8131	(α, β)	$\rightarrow$	∞	$C^* = \beta/\alpha$
2	0.8131	(K_T, ϕ)	$\rightarrow$	$(0, \infty)$	$A = \phi K_T$
3	0.8131	(κ, R)	$\rightarrow$	$(0, \infty)$	$\kappa R \rightarrow \infty$
4	0.8131	R	$\rightarrow$	∞	
5	0.8645	(C^*, A)	$\rightarrow$	$(0, \infty)$	$\lambda = AC^*$
6	1	α	$\rightarrow$	∞	To undo the effect $C_1 \propto L$ for $L \leq 2$
7	1	b	$\rightarrow$	0	Uncluttering τ
				FINAL OUTPUT	$C_1 = \lambda\tau = \phi K_T \beta\tau/\alpha$

SHP-1 model

First reduction

We first perform parameter reduction on the SHP-1 model with global symmetry breaking. Initial equations for the SHP-1 model are given by

$$S = \alpha C_1(S_T - S) - \beta S \tag{A.5}$$

$$\dot{C}_0 = \kappa(L - \sum_i C_i)(R - \sum_i C_i) + \gamma_1 S C_1 - (\phi_1 + \tau^{-1})C_0 \tag{A.6}$$

$$\dot{C}_1 = \phi_1 C_0 + \gamma_2 S C_2 - (\gamma_1 S + \phi_2 + \tau^{-1})C_1 \tag{A.7}$$

$$\dot{C}_2 = \phi_2 C_1 + \gamma_3 S C_3 - (\gamma_2 S + \phi_3 + \tau^{-1})C_2 \tag{A.8}$$

$$\dot{C}_3 = \phi_3 C_2 + \gamma_4 S C_4 - (\gamma_3 S + \phi_4 + \tau^{-1})C_3 \tag{A.9}$$

$$\dot{C}_4 = \phi_4 C_3 + \gamma_5 S C_5 - (\gamma_4 S + \phi_5 + \tau^{-1}) C_4 \tag{A.10}$$

$$\dot{C}_5 = \phi_5 C_4 - (\gamma_5 S + \tau^{-1}) C_5. \tag{A.11}$$

Initial parameters for this model are given in Table A.4. Steps of the first reduction of this model are given in Table A.5. The final system is given by the following equations when the reduction steps of Table A.5 are applied.

Table A.4: SHP-1 model initial parameters

Parameter	Value
ϕ	9×10^{-2}
γ	1
S_T	7.2×10^{-1}
β	3×10^{2}
α	1
$\beta/\alpha = C^*$	3×10^{2}
κ	10^{-4}
R	3×10^{4}

Table A.5: SHP-1 First reduction

Step	I_{init}	Parameters		Limit	Description per group
1	0.7369	(κ, R)	$\rightarrow$	$(0, \infty)$	
2	0.7369	γ_1	$\rightarrow$	∞	
3	0.8468	(ϕ_2, ϕ_1)	$\rightarrow$	$(0, \infty)$	
4	0.8583	R	$\rightarrow$	∞	
5	0.8583	(γ_4, γ_5)	$\rightarrow$	$0, \infty$	Kinetic sensing module
6	1.0000	γ_2	$\rightarrow$	0	Uncluttering τ
7	1.0000	γ_3	$\rightarrow$	0	
8	1.0000	(ϕ_3, S_T)	$\rightarrow$	∞	Rescaling
9	1.0000	(ϕ_1, ϕ_4)	$\rightarrow$	∞	
10	1.0000	β	$\rightarrow$	∞	Adaptation module
11	1.0000	(ϕ_4, S_T)	$\rightarrow$	∞	
12	1.0000	(S_T, α)	$\rightarrow$	$(0, \infty)$	
				FINAL OUTPUT	$C_5 = \frac{\phi_2 \phi_5 \beta}{\gamma_5 S_T \alpha} \tau$

$$S = \alpha C_1 S_T - \beta S \tag{A.12}$$

$$\dot{C}_0 = \kappa R(L - \sum_i C_i) + \gamma_1 S C_1 - \phi_1 C_0 \tag{A.13}$$

$$\dot{C}_1 = \phi_1 C_0 - (\phi_2 + \gamma_1 S) C_1 \tag{A.14}$$

$$\dot{C}_2 = \phi_2 C_1 - \phi_3 C_2 \tag{A.15}$$

$$\dot{C}_3 = \phi_3 C_2 - \phi_4 C_3 \tag{A.16}$$

$$\dot{C}_4 = \phi_4 C_3 + \gamma_5 S C_5 - (\phi_5 + \tau^{-1}) C_4 \tag{A.17}$$

$$\dot{C}_5 = \phi_5 C_4 - \gamma_5 S C_5. \tag{A.18}$$

Second reduction

We then perform another reduction of the same model using a different binning for the computation of the mutual information. Initial parameters and equations are identical as in the previous reduction presented in section A. Steps for this reduction are given in Table A.6.

Table A.6: SHP-1 Second reduction

Step	I_{init}	Parameters		Limit	Description per group
1	0.6328	(κ, R)	$\rightarrow$	$(0, \infty)$	
2	0.6328	R	$\rightarrow$	∞	
3	0.6375	(γ_4, α)	$\rightarrow$	$(0, \infty)$	
4	0.6464	(γ_2, γ_1)	$\rightarrow$	$0, \infty$	
5	0.7264	γ_5	$\rightarrow$	∞	Adaptive module
6	1.0000	γ_3	$\rightarrow$	0	
7	1.0000	(ϕ_1, β)	$\rightarrow$	∞	
8	1.0000	ϕ_4	$\rightarrow$	∞	Kinetic sensing module
9	1.0000	(ϕ_3, S_T)	$\rightarrow$	∞	
10	1.0000	(S_T, β)	$\rightarrow$	∞	
11	1.0000	(ϕ_5, β)	$\rightarrow$	$(0, \infty)$	
12	1.0000	(β, ϕ_2)	$\rightarrow$	$(0, \infty)$	
		FINAL OUTPUT			$C_5 = \frac{\phi_2 \phi_5 \beta}{\gamma_5 S_T \alpha} \tau$

The final system is given by the following equations when the reduction steps given in Table A.6 are applied.

$$S = \alpha C_1 S_T - \beta S \tag{A.19}$$

$$\dot{C}_0 = \kappa R(L - \sum_i C_i) + \gamma_1 S C_1 - \phi_1 C_0 \tag{A.20}$$

$$\dot{C}_1 = \phi_1 C_0 - (\phi_2 + \gamma_1 S) C_1 \tag{A.21}$$

$$\dot{C}_2 = \phi_2 C_1 - \phi_3 C_2 \tag{A.22}$$

$$\dot{C}_3 = \phi_3 C_2 + \gamma_4 S C_4 - \phi_4 C_3 \tag{A.23}$$

$$\dot{C}_4 = \phi_4 C_3 + \gamma_5 S C_5 - (\phi_5 + \gamma_4 S + \tau^{-1}) C_4 \tag{A.24}$$

$$\dot{C}_5 = \phi_5 C_4 - \gamma_5 S C_5. \tag{A.25}$$

Third reduction

We perform another reduction of the same model using slightly different initial parameter values. All parameters are given in Table A.4 with $S_T \to 5S_T$. Initial set of equations is identical as in Sections A and A. Steps for this reduction are given in Table A.7.

Table A.7: SHP-1 Third reduction

Step	I_{init}	Parameters		Limit	Description per group
1	0.4946	(β, α)	$\to$	∞	
2	0.4946	(R, κ)	$\to$	$(0, \infty)$	
3	0.4946	(γ_1, γ_5)	$\to$	0	Kinetic sensing module
4	1.0000	(κ, ϕ_1)	$\to$	∞	
5	1.0000	ϕ_1	$\to$	∞	
6	1.0000	(ϕ_2, ϕ_4)	$\to$	$(0, \infty)$	
7	1.0000	(ϕ_5, γ_3)	$\to$	∞	Adaptation module
8	1.0000	γ_2	$\to$	0	
9	1.0000	S_T	$\to$	∞	
10	1.0000	γ_4	$\to$	0	
11	1.0000	(ϕ_4, ϕ_3)	$\to$	$(0, \infty)$	Rescaling
12	1.0000	(α, γ_3)	$\to$	$(0, \infty)$	
				FINAL OUTPUT	$C_5 = \frac{\phi_2 \phi_3 \phi_4 \beta}{\gamma_3 S_T \alpha} \tau^2$

The final system is given by the following equations when the reduction steps given in Table A.7 are applied.

$$
\begin{aligned}
S &= \alpha C_1 S_T - \beta S\\
\dot{C}_0 &= \kappa(L - \sum_i C_i)(R - \sum_i C_i) - \phi_1 C_0\\
\dot{C}_1 &= \phi_1 C_0 - (\phi_2 + \tau^{-1})C_1\\
\dot{C}_2 &= \phi_2 C_1 + \gamma_3 S C_3 - (\phi_3 + \tau^{-1})C_2\\
\dot{C}_3 &= \phi_3 C_2 + -\gamma_3 S C_3\\
\dot{C}_4 &= \phi_4 C_3 - \phi_5 C_4\\
\dot{C}_5 &= \phi_5 C_4 - \tau^{-1} C_5.
\end{aligned}
$$

Reduction without feedback

We perform a reduction of the SHP-1 model with the SHP-1 mediated feedback turned off. Parameter values are given in Table A.4 with $S_T = 0$. The network topology is as in Fig. S A.2A and the corresponding initial set of equations is identical as in Sections A and A. Fig. S A.2B shows that the reduction does not converge when crucial network elements (SHP-1 feedback) are missing.

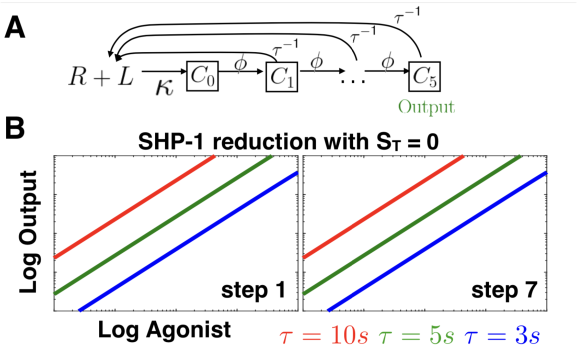


Figure A.2: **Negative control for SHP-1 model.** We attempt to reduce this model with $\bar{\phi}$ in absence of SHP-1 (corresponding to pure kinetic proofreading). The algorithm fails to optimize behavior and fitness, indicating that it is not possible to do so for arbitrary networks.

Analytical study

The full analytical study of the SHP-1 model is done in [33]. Assuming all $\phi_i = \phi$ and $\gamma_i = \gamma$ are equal, we get at lowest order

$$C_1 \simeq r_-(1-r_-)\frac{\kappa RL}{\kappa R + \nu_1} \tag{A.26}$$

with

$$r_\pm = \frac{\phi + S + \nu_1 \pm \sqrt{(\phi + S + \nu_1)^2 - 4\phi S}}{2S} \tag{A.27}$$

.

We can use the previous expression to get a closed equation for S as a function of $r_-(S)$ and C_*.

$$S = S_T \frac{C_1}{C_1 + C_*} = S_T \frac{r_-(1-r_-)}{r_-(1-r_-) + \frac{C_*(\kappa R + \nu_1)}{\kappa RL}} \tag{A.28}$$

This is a 4th order polynomial equation in S in terms of the parameters that can be conveniently solved numerically. Once this is done, we get the following expression for C_N, the final complex in the cascade as a function of $r_\pm$ to the lowest order in r_-^N.

$$C_N \simeq \frac{\kappa RL}{\kappa R + \nu_1}\left(1 - \frac{r_-}{r_+}\right) r_-^N \tag{A.29}$$

To see why this feedback hinders perfect adaptation, it is useful to consider the limit of big L and big S_T. In this limit, it is shown in [33] that the parameter r_- becomes inversely proportional to the feedback variable $1/S$, thus giving at lowest order a S^{-N} contribution in Eq. A.29, clearly coming from the coupling of N identical proofreading steps. Those equations can be approximately solved [33] so that

$$C_N \simeq \left(\frac{\phi\beta}{\alpha\gamma S_T}\right)^{N/2} (L)^{1-N/2}. \tag{A.30}$$

So we see that, unless $N = 2$, there is an unavoidable L dependency. The $L^{-N/2}$

dependency comes from the steady state value of the feedback variable $S \propto L^{1/2}$ appearing when we fully close this system.

Lipniacki model

In this section, we present two parameter reductions performed by $\bar{\phi}$. Initial equations for the Lipniacki model are:

$$
\begin{aligned}
\dot{X}_2 &= b_1\, pMHC_{free}\, TCR_{free} + (s_2+s_3)X_{23} - (lb\, LCK_{free} + ly_1 X_{29} + \tau^{-1})X_2\\
\dot{X}_3 &= lb\, LCK_{free}\, X_2 + ls_1 X_4 + (s_2+s_3)X_{24} - (ly_2 + X_{37} + ly_1 X_{29} + \tau^{-1})X_3\\
\dot{X}_4 &= X_{37}X_3 - (ly_2 + ls_1 + \tau^{-1})X_4\\
\dot{X}_5 &= ly_2 X_3 + ls_1 X_6 - (tp + X_{37} + ly_1 X_{29} + \tau^{-1})X_5\\
\dot{X}_6 &= ly_2 X_4 + X_{37}X_5 - (tp + ls_1 + \tau^{-1})X_6\\
\dot{X}_7 &= tpX_5 + ls_1 X_8 - (tp + X_{37} + ly_1 X_{29} + \tau^{-1})X_7\\
\dot{X}_8 &= tpX_6 + X_{37}X_7 - (tp + ls_1 + \tau^{-1})X_8\\
\dot{X}_9 &= tpX_7 + ls_1 X_8 - (\tau^{-1} + X_{37} + ly_1 X_{29} + \tau^{-1})X_9\\
\dot{X}_{10} &= tpX_8 + X_{37}X_9 - (ls_1 + \tau^{-1})X_{10}\\
\dot{X}_{22} &= ly_1 X_{29}\, TCR_{free} + \tau^{-1}(X_{23} + X_{24}) - (s_2+s_3)X_{22}\\
\dot{X}_{23} &= ly_1 X_{29} X_2 - (s_2 + s_3 + \tau^{-1})X_{23}\\
\dot{X}_{24} &= ly_1 X_{29}(X_3 + X_5 + X_7 + X_9) - (s_2 + s_3 + \tau^{-1})X_{24}\\
\dot{X}_{29} &= s_1(X_5 + X_7 + X_9)\, SHP_{free} + s_3(X_{22} + X_{23} + X_{24}) + s_0\, SHP_{free}\\
&\quad - ly_1(X_2 + X_3 + X_5 + X_7 + X_9 + TCR_{free})X_{29} - s_2 X_{29}\\
\dot{X}_{31} &= z_1(X_9 + X_{10})(m_1 - X_{31}) + z_0 m_1 - (z_0 + z_2)X_{31}\\
\dot{X}_{33} &= 2X_{31}(e_1 - X_{34}) + 2m_2 X_{34} - (m_2 + 3X_{31})X_{33}\\
\dot{X}_{34} &= X_{31}X_{33} - 2m_2 X_{34}\\
\dot{X}_{36} &= 2X_{34}(ls_2 - X_{37}) + 2e_2 X_{37} - (e_2 + 2X_{34})X_{36}\\
\dot{X}_{37} &= X_{34}X_{36} - 2e_2 X_{37}
\end{aligned}
$$

To ensure physical behavior throughout the reduction process, we manually implement the following mass conservation laws.

$$\begin{aligned}
pMHC_{free} &= pMHC - \left(\sum_{i=2}^{10} X_i + X_{23} + X_{24}\right) \\
TCR_{free} &= TCR - \left(\sum_{i=2}^{10} X_i + X_{22} + X_{23} + X_{24}\right) \\
LCK_{free} &= LCK - \left(\sum_{i=3}^{10} X_i + X_{24}\right) \\
SHP_{free} &= SHP - (X_{22} + X_{23} + X_{24} + X_{29}) \\
ZAP_{free} &= ZAP - X_{31} \\
MEK_{free} &= MEK - (X_{33} + X_{34}) \\
ERK_{free} &= ERK - (X_{36} + X_{37})
\end{aligned}$$

We also perform initial rescaling of equations X_{31} to X_{37} to save $\bar{\phi}$ steps:

$$\begin{aligned}
X_{31} &\to \frac{m_1 X_{31}}{ZAP} \\
X_{33} &\to \frac{e_1 X_{33}}{MEK} \\
X_{34} &\to \frac{e_1 X_{34}}{MEK} \\
X_{36} &\to \frac{ls_2 X_{36}}{ERK} \\
X_{37} &\to \frac{ls_2 X_{37}}{ERK}
\end{aligned}$$

Initial parameters are given in Table A.8.

Lipniacki model first reduction

We discarded all values of the output below the measurable threshold 10^{-2}, and used 40 log-uniformly distributed bins on the interval $[10^{-2}, 10^{2}]$ for the computation of the Output distribution. The Input concentrations were given by 50 log-uniformly

Table A.8: Lipniacki model initial parameters

Parameter	Value	Details
TCR	3×10^4	
LCK	10^5	
SHP	3×10^5	
ZAP	10^5	Can't be modified by $\bar{\phi}$
MEK	10^5	
ERK	3×10^5	
b_1	$3 \times 10^{-1}/TCR$	Agonist peptide binding
lb	$3 \times 10^{-1}/LCK$	LCK(s) binding
ly_1	$5/SHP$	pSHP complex binding
ly_2	3×10^{-1}	Theorine phosphorylation at complex
ls_1	10^{-1}	Spontaneous serine dephosphorylation
ls_2	5×10^{-1}	ppERK catalyzed serine phosphorylation
tp	5×10^{-2}	TCR phosphorylation
s_0	10^{-5}	Spontaneous SHP phosphorylation
s_1	$3 \times 10^2/SHP$	SHP phosphorylation
s_2	6×10^{-4}	SHP dephosphorylation
s_3	5×10^{-2}	SHP dissociation
z_0	2×10^{-6}	Spontaneous ZAP phosporylation
z_1	$5/ZAP$	ZAP phosphorylation
z_2	2×10^{-2}	ZAP dephosphorylation
m_1	$5 \times ZAP/MEK$	MEK phosphorylation
m_2	2×10^{-2}	MEK dephosphorylation
e_1	$5 \times MEK/ERK$	ERK phosphorylation
e_2	2×10^{-2}	ERK dephosphorylation

distributed values on the interval $[1, 10^4]$.

Steps of the first biochemical reduction of this model (odeLIPbasic.m in the MATLAB code) are given in Table A.9. The results of the biochemical reduction are given by

Table A.9: Lipniacki basic first variant

Step	I_{init}	Parameters		Limit	Description per group
1	0.45	(m_1, m_2)	$\rightarrow$	∞	
2	0.47	b_1	$\rightarrow$	∞	
3	0.47	(lb, LCK)	$\rightarrow$	$(0, \infty)$	$lb' = lb\,LCK$
4	0.47	ls_1, z_1	$\rightarrow$	$0, \infty$	Turning off the positive feedback
5	0.50	e_1, e_2	$\rightarrow$	0	
6	0.50	z_0, z_2	$\rightarrow$	0	
7	0.50	m_1	$\rightarrow$	0	
8	0.50	s_3	$\rightarrow$	0	
9	0.50	(ls_2, LCK)	$\rightarrow$	$(0, \infty)$	
10	0.50	ly_2	$\rightarrow$	∞	
11	0.50	(TCR, SHP)	$\rightarrow$	∞	
12	0.5017	s_0	$\rightarrow$	0	
13	0.5017	(s_2, ly_1)	$\rightarrow$	$(0, \infty)$	Products
14	0.5017	(SHP, s_1)	$\rightarrow$	$(0, \infty)$	$s_1' = s_1 SHP$
15	0.5017	(LCK, ly_1)	$\rightarrow$	$(0, \infty)$	
16	0.5216	s_1	$\rightarrow$	∞	

$$\begin{aligned}
\dot{X}_2 &= b_1\, pMHC_{free}\, TCR + s_{22} X_{23} - lb X_2 \\
\dot{X}_3 &= lb X_2 + s_{23} X_{24} - ly_{21} X_3 \\
\dot{X}_4 &= X_{37} X_3 - ly_{22} X_4 \\
\dot{X}_5 &= ly_{21} X_3 - (tp_1 + X_{37} + ly_{14} X_{29} + \tau^{-1}) X_5 \\
\dot{X}_6 &= ly_{22} X_4 + X_{37} X_5 - (tp_2 + \tau^{-1}) X_6 \\
\dot{X}_7 &= tp_1 X_5 - (tp_3 + X_{37} + ly_{15} X_{29} + \tau^{-1}) X_7 \\
\dot{X}_8 &= tp_2 X_6 + X_{37} X_7 - (tp_4 + \tau^{-1}) X_8 \\
\dot{X}_9 &= tp_3 X_7 - (\tau^{-1} + X_{37} + ly_{16} X_{29}) X_9 \\
\dot{X}_{10} &= tp_4 X_8 + X_{37} X_9 - \tau^{-1} X_{10}
\end{aligned}$$

$$\dot{X}_{23} = ly_{12}X_{29}X_2 - (s_{22} + \tau^{-1})X_{23}$$
$$\dot{X}_{24} = (ly_{13}X_3 + ly_{14}X_5 + ly_{15}X_7 + ly_{16}X_9)X_{29} - (s_{23} + \tau^{-1})X_{24}$$
$$\dot{X}_{29} = s_{11}X_5 + s_{12}X_7 + s_{13}X_9 - ly_{11}TCR\,X_{29}.$$

We then perform global symmetry breaking (odeLIPadvanced in the MATLAB code). Steps of reduction are given in Table A.10.

Table A.10: Lipniacki advanced first variant

Step	I_{init}	Parameters		Limit	Description per group
1	0.5837	(s_{22}, s_{23})	$\rightarrow$	0	
2	0.5837	(b_1, ly_{22})	$\rightarrow$	∞	$ly_{22} \rightarrow \infty$ makes no change
3	0.5837	(s_{21}, ly_{22})	$\rightarrow$	∞	
4	0.5837	ly_{22}	$\rightarrow$	∞	
5	0.5837	(TCR, ly_{11})	$\rightarrow$	$(0, \infty)$	$ly'_{11} = ly_{11}TCR$
6	0.5837	s_{12}	$\rightarrow$	0	
7	0.6097	s_{13}	$\rightarrow$	0	
8	0.6147	(tp_2, tp_3)	$\rightarrow$	$(0, \infty)$	
9	0.6231	(ly_{13}, lb)	$\rightarrow$	$0, \infty$	
10	0.6245	(tp_4, ls_2)	$\rightarrow$	$(0, \infty)$	Products
11	0.6246	(tp_3, ly_{16})	$\rightarrow$	$(0, \infty)$	
12	0.6354	(ly_{16}, ly_{15})	$\rightarrow$	$(0, \infty)$	
13	0.6563	(ly_{15}, ly_{13})	$\rightarrow$	$(0, \infty)$	
14	0.6699	ly_{21}	$\rightarrow$	∞	
15	0.6749	ls_2, ly_{12}	$\rightarrow$	$0, \infty$	
16	0.7405	(ly_{14}, s_{11})	$\rightarrow$	∞	

Global symmetry breaking results in the following system.

$$\dot{X}_2 = b_1\,pMHC_{free}\,TCR - lbX_2$$
$$\dot{X}_3 = lbX_2 - ly_2X_3$$
$$\dot{X}_5 = ly_2X_3 - ly_{13}X_{29}X_5$$
$$\dot{X}_7 = tp_1X_5 - (tp_2 + ly_{14}X_{29} + \tau^{-1})X_7$$
$$\dot{X}_9 = tp_2X_7 - (ly_{15}X_{29} + \tau^{-1})X_9$$
$$\dot{X}_{23} = ly_{12}X_{29}X_2 - \tau^{-1}X_{23}$$

$$\dot{X}_{24} = ly_{13}X_{29}X_5 - \tau^{-1}X_{24}$$
$$\dot{X}_{29} = s_1X_5 - ly_{11}TCR\,X_{29}$$

Only four more steps of reduction are needed to reach perfect adaptation, namely $(ly_{13}, ly_{15}) \to 0$, $(ly_{11}, ly_{14}) \to \infty$, $(tp_1, ly_{14}) \to \infty$ and finally $ly_{14} \to \infty$. We apply those steps of reduction by hand and reach the final following system.

$$\dot{X}_2 = b_1\,pMHC_{free}\,TCR - lbX_2 \tag{A.31}$$
$$\dot{X}_3 = lbX_2 - ly_2X_3 \tag{A.32}$$
$$\dot{X}_5 = ly_2X_3 - tp_1X_5 \tag{A.33}$$
$$\dot{X}_7 = tp_1X_5 - ly_{14}X_{29}X_7 - tp_2X_7 \tag{A.34}$$
$$\dot{X}_9 = tp_2X_7 - \tau^{-1}X_9 \tag{A.35}$$
$$\dot{X}_{24} = ly_{14}X_7X_{29} - \tau^{-1}X_{24} \tag{A.36}$$
$$\dot{X}_{29} = s_1X_5 - ly_{11}TCR\,X_{29} \tag{A.37}$$

Lipniacki model second reduction

Initial equations, parameters, mass conservation laws and equation transformations for this reduction are the same as for the previous Lipniacki reduction. For this reduction, we chose mutual information as the fitness with 40 bins log-uniformly distributed on the interval $[10^{-2}, 10^{2}]$, plus a lower bin for concentrations below 10^{-2} and a higher bin for concentrations above 10^{2}. We chose 50 log-uniformly distributed Input concentrations on the interval $[1, 10^{4}]$. Because of the binning choice, the fitness, was optimized quicker, while most reduction took place in the neutral fitness landscape of maximum fitness of $1\,bit$. The details of this biochemical reduction are given in Table A.11.

After the first reduction, the system is reduced to

$$\dot{X}_2 = b_1\,pMHC_{free}\,TCR_{free} + (s_{22} + s_{32})X_{23} - lb\,LCK\,X_2$$

Table A.11: Lipniacki basic second variant

Step	I_{init}	Parameters		Limit	Description per group
1	0.7583	(m_2, m_1)	$\to$	∞	Shutting down positive feedback
2	0.8337	(z_2, m_1)	$\to$	∞	
3	0.8337	LCK	$\to$	∞	
4	0.8777	lb	$\to$	∞	
5	0.8777	(ls_1, ly_2)	$\to$	$(0, \infty)$	
6	0.8777	(z_0, s_0)	$\to$	0	
7	0.8777	(m_1, s_1)	$\to$	∞	
8	0.8777	(ly_2, z_1)	$\to$	$(0, \infty)$	
9	0.8915	e_2	$\to$	0	
10	0.8915	z_1	$\to$	0	
11	0.8915	e_1	$\to$	∞	
12	0.8915	(s_1, SHP)	$\to$	$(0, \infty)$	Rescaling
13	0.8954	(b_1, SHP)	$\to$	∞	
14	0.9029	(s_3, s_2)	$\to$	$(0, \infty)$	
15	0.9278	(ls_2, tp)	$\to$	$(0, \infty)$	
16	0.9351	(tp, TCR)	$\to$	$(0, \infty)$	
17	0.9725	(s_2, ly_1)	$\to$	$(0, \infty)$	
18	1	(SHP, ly_1)	$\to$	$(0, \infty)$	

$$
\begin{aligned}
\dot{X}_3 &= lb\,LCK\,X_2 + ls_{11}X_4 + (s_{23} + s_{33})X_{24} - (ly_{21} + X_{37} + ly_{11}X_{29} + \tau^{-1})X_3 \\
\dot{X}_4 &= X_{37}X_3 - (ly_{22} + ls_{11} + \tau^{-1})X_4 \\
\dot{X}_5 &= ly_{21}X_3 + ls_{12}X_6 - (tp_1 + X_{37} + ly_{12}X_{29} + \tau^{-1})X_5 \\
\dot{X}_6 &= ly_{22}X_4 + X_{37}X_5 - (tp_2 + ls_{12} + \tau^{-1})X_6 \\
\dot{X}_7 &= tp_1X_5 + ls_{13}X_8 - (tp_3 + X_{37} + ly_{13}X_{29} + \tau^{-1})X_7 \\
\dot{X}_8 &= tp_2X_6 + X_{37}X_7 - (tp_4 + ls_{13} + \tau^{-1})X_8 \\
\dot{X}_9 &= tp_3X_7 + ls_{14}X_8 - (X_{37} + ly_{14}X_{29} + \tau^{-1})X_9 \\
\dot{X}_{10} &= tp_4X_8 + X_{37}X_9 - (ls_{14} + \tau^{-1})X_{10} \\
\dot{X}_{22} &= ly_{15}X_{29}\,TCR_{free} + \tau^{-1}(X_{23} + X_{24}) - (s_{21} + s_{31})X_{22} \\
\dot{X}_{23} &= ly_{16}X_{29}X_2 - (s_{22} + s_{32} + \tau^{-1})X_{23} \\
\dot{X}_{24} &= X_{29}(ly_{11}X_3 + ly_{12}X_5 + ly_{13}X_7 + ly_{14}X_9) - (s_{23} + s_{33} + \tau^{-1})X_{24} \\
\dot{X}_{29} &= (s_{11}X_5 + s_{12}X_7 + s_{13}X_9)\,SHP + (s_{31}X_{22} + s_{32}X_{23} + s_{33}X_{24}) \\
&\qquad - (ly_{16}X_2 + ly_{11}X_3 + ly_{12}X_5 + ly_{13}X_7 + ly_{14}X_9 + ly_{15}TCR_{free})X_{29} - s_{24}X_{29}
\end{aligned}
$$

$X_{37} = 0.05$

We then perform global symmetry breaking on this system. Steps of the biochemical reduction of this model are given in Table A.12.

Table A.12: Lipniacki advanced second variant

Step	I_{init}	Parameters		Limit	Description per group
1	1	(m_{21}, ly_{22})	$\rightarrow$	0	Cleaning unnecessary parameters
2	1	(m_1, s_{24})	$\rightarrow$	0	
3	1	(ly_{11}, ls_{13})	$\rightarrow$	0	
4	1	(s_{12}, s_{31})	$\rightarrow$	0	
5	1	(ly_{13}, ls_{11})	$\rightarrow$	0	
6	1	(e_1, ls_{14})	$\rightarrow$	0	
7	1	(tp_2, s_{33})	$\rightarrow$	0	
8	1	(m_{22}, ls_{12})	$\rightarrow$	0	
9	1	(z_2, s_{22})	$\rightarrow$	0	
10	1	(s_{32}, s_{13})	$\rightarrow$	0	
11	1	ly_{16}	$\rightarrow$	0	
12	1	(s_{23}, ly_{14})	$\rightarrow$	0	
13	1	ls_2	$\rightarrow$	∞	
14	1	(s_{11}, tp_4)	$\rightarrow$	∞	Strengthening remaining reactions
15	1	(ly_{21}, ly_{15})	$\rightarrow$	∞	
16	1	(b_1, s_{21})	$\rightarrow$	∞	
17	1	tp_3	$\rightarrow$	0	Turning off one output
18	1	(lb, ly_{15})	$\rightarrow$	∞	Strengthening remaining reactions
19	1	(SHP, s_{21})	$\rightarrow$	∞	
20	1	(LCK, ly_{12})	$\rightarrow$	∞	
21	1	(ly_{12}, tp_4)	$\rightarrow$	∞	
22	1	(ly_{15}, tp_4)	$\rightarrow$	∞	
23	1	tp_4	$\rightarrow$	∞	
24	1	(TCR, tp_1)	$\rightarrow$	$(0, \infty)$	
		FINAL OUTPUT			$X_{10} = \frac{tp_1 s_{21} TCR}{s_{11} ly_{17}} \tau$

We can remove equations for X_4, X_6, X_{23} and X_{24} as they are dead ends in the network. $X_{37} = 0.5$ is held constant. The final expression of the output given in Table A.12 is extracted from remaining equations at steady-state; expanding the equations for the relevant cascade we get

$$\dot{X}_2 = b_1\, pMHC_{free}\, TCR_{free} - lb\, LCK\, X_2 \tag{A.38}$$

$$\dot{X}_3 = lb\, LCK\, X_2 - ly_{21} X_3 \tag{A.39}$$

$$\dot{X}_5 = ly_{21} X_3 - ly_{12} X_{29} X_5 \tag{A.40}$$

$$\dot{X}_7 = tp_1 X_5 - X_{37} X_7 \tag{A.41}$$

$$\dot{X}_8 = X_{37} X_7 - tp_4 X_8 \tag{A.42}$$

$$\dot{X}_{10} = tp_4 X_8 - \tau^{-1} X_{10} \tag{A.43}$$

$$\dot{X}_{22} = ly_{15} X_{29}\, TCR_{free} - s_{21} X_{22} \tag{A.44}$$

$$\dot{X}_{29} = s_{11} X_5\, SHP - ly_{15} TCR_{free} X_{29} \tag{A.45}$$

$$X_{37} = 0.5 \tag{A.46}$$

The output here is X_{10}. Variables X_{22}, X_{29} and X_5 respectively correspond to variables R_p, S and C_5 in the main text. The structure of Eqs. A.38 to A.45 is clearly very similar to the equations of the previous reduction A.31 to A.37, with a linear cascade for the second reduction $X_2 \rightarrow X_3 \rightarrow X_5 \rightarrow X_7 \rightarrow X_8 \rightarrow X_{10}$ and $X_2 \rightarrow X_3 \rightarrow X_5 \rightarrow X_7 \rightarrow X_9$ for the first reduction, modulated by a parallel loop via X_{29} and X_5. As described in the main text, the structural difference comes from the mechanism of this loop, the first reduction giving an effective feedforward adaptive system, while the second reduction is an integral feedback mechanism. "

(Parameter reduction [1])

B

Attack and defence

Mathematical details of the adaptive proofreading models

“ This section contains more details on the derivation of adaptive proofreading models (section **Biochemical kinetics**), referred to in subsection **Adaptive proofreading for cellular decision-making** in Chapter 3.1. We also give the parameters and equations that are used to draw Fig. 3.2 B (section **Parameters for Fig. 3.2 B**).

Biochemical kinetics

The kinetics for the biochemical network in Fig. 3.2 B in the simplest form ($(N, m) = (2, 1)$) are given by

$$\begin{aligned} \dot{C}_1 &= k^{\mathsf{on}} RL - (\phi K + \tau^{-1}) C_1 \\ \dot{C}_2 &= \phi K C_1 - \tau^{-1} C_2 \\ \dot{K} &= \beta (K_T - K) - \alpha C_1 K. \end{aligned} \tag{B.1}$$

Here, we assume the T cell has R receptors to which L ligands are bound to form ligand-receptor complexes C_1 and C_2. The parameters k^{on} and τ^{-1} denote ligand-specific rates, which correspond to an average number of events happening per second (mean of a Poisson-distributed variable). ϕ is the phosphorylation rate for the reaction $C_1 \rightarrow C_2$ (activation branch), which is activated by variable K, and which we will call a generic kinase. K itself is inhibited by C_1 (repression branch) with rate α. K_T here is the total number of kinase, and $K_T - K$ the number of inactive kinase. This kinase is shared between all receptors and assumed to diffuse freely and rapidly, so that since K is inactivated by C_1, (in)activity of K is a measure of the total number of receptors bound. Lastly, β is the activation rate of K. In the steady state, we can solve exactly for C_2 and find

$$C_2 = \phi K C_1 \tau = \frac{L\tau}{\beta/\alpha + L} \simeq \frac{L\tau}{L} = \tau. \tag{B.2}$$

Here $K = \frac{K_T \beta/\alpha}{\beta/\alpha + C_1}$, and as long as $L \gg \beta/\alpha$ the first-order approximation is exact and the ligand dependence in nominator and denominator cancels. Without loss of generality, we have set $\frac{\phi K_T \beta}{\alpha} = 1$.

When we consider an environment containing two ligand types with binding times τ_{ag} (agonists) and τ_{a} (antagonists) at concentrations L_{ag} and L_{a}, two types of ligand-receptor complexes can be formed. We call them C_i for agonists and D_i for antagonists. Full equations in the case of $(N, m) = (2, 1)$ are given by

$$\begin{aligned}
\dot{C}_1 &= k^{\mathsf{on}} R L_{\mathsf{ag}} - (\phi K + \tau_{\mathsf{ag}}^{-1}) C_1 \\
\dot{C}_2 &= \phi K C_1 - \tau_{\mathsf{ag}}^{-1} C_2 \qquad (\mathrm{B.3}) \\
\dot{D}_1 &= k^{\mathsf{on}} R L_{\mathsf{a}} - (\phi K + \tau_{\mathsf{a}}^{-1}) D_1 \\
\dot{D}_2 &= \phi K D_1 - \tau_{\mathsf{a}}^{-1} D_2 \qquad (\mathrm{B.4}) \\
\dot{K} &= \beta (K_T - K) - \alpha (C_1 + D_1) K.
\end{aligned}$$

where we have assumed that k^{on} is equal for both agonist and antagonist ligands. The main difference here is that variable K integrates global information from both ligand complexes, which results in the steady-state in $K = \frac{K_T \beta/\alpha}{\beta/\alpha + C_1 + D_1}$. Moreover, K acts locally on the phosphorylation of both C_1 and D_1. Finally, the output is given by $T_{2,1} = C_2 + D_2$.

We can generalize this case by assuming that inhibition of the variable K occurs further downstream a kinetic proofreading cascade, namely at the m-th complex $C_m = L_{\text{ag}}\tau_{\text{ag}}^m$ and $D_m = L_{\text{a}}\tau_{\text{a}}^m$. The output variable is then given by $T_{N,m} = C_N + D_N$. Fig. 3.2 A shows how information from a single ligand passes through the repression branch (red arrow and box) via K and through the activation branch (green arrow and box) via C_N. The global variable K integrates local information as $K = \frac{K_T\beta/\alpha}{\beta/\alpha + C_m + D_m} \propto \left(L_{\text{ag}}\tau_{\text{ag}}^m + L_{\text{a}}\tau_{\text{a}}^m\right)^{-1}$, and catalyzes the phosphorylation of $C_{N-1} = L_{\text{ag}}\tau_{\text{ag}}^{N-1}$ and $D_{N-1} = L_{\text{a}}\tau_{\text{a}}^{N-1}$ to final complex C_N and D_N as

$$\dot{C}_N = KC_{N-1} - \tau_{\text{ag}}^{-1}C_N \tag{B.5}$$

$$\dot{D}_N = KD_{N-1} - \tau_{\text{a}}^{-1}D_N. \tag{B.6}$$

In the steady-state, the solution for $T_{N,m}$ is then

$$T_{N,m} = C_N + D_N = \frac{L_{\text{ag}}\tau_{\text{ag}}^N + L_{\text{a}}\tau_{\text{a}}^N}{L_{\text{ag}}\tau_{\text{ag}}^m + L_{\text{a}}\tau_{\text{a}}^m}. \tag{B.7}$$

This expression for two types of ligands with same k_{on} can be clearly generalized to any types of ligands, giving Eq. 3.3.

Materials and methods

In this section, we give the parameters and equations that are used to draw Fig. 3.2 B and we give the hyperparameters used for training the neural networks classifying 3s and 7s. We referred to the latter in subsection **Neural networks for artificial decision-making** in Chapter 3.1.

Parameters for Fig. 3.2 B

The curves on Fig. 3.2 B, left panel, come from the model given by

$$T_{4,2}(L) = \frac{1}{\tau_d^2}\frac{L\tau^4}{C_* + L\tau^2} \tag{B.8}$$

with parameter values $C_* = \beta/\alpha = 3000,\ \tau_d = 4s$ and τ as in the legend. The curves on the middle panel of Fig. 3.2 B come from

$$T_{4,2}(L) = \frac{1}{\tau_d^2} \frac{L\tau^4 + L_a\tau_a^4}{C_* + L\tau^2 + L_a\tau_a^2} \tag{B.9}$$

with again $C_* = 3000,\ \tau_d = 4s$ and $\tau = 10s$. For blue "agonists alone", $L_a = 0$, for orange "+ antagonists" $L_a = 10^4$ and $\tau_a = 3s$, and for green "+ self" $L_a = 10^4$ and $\tau_a = 1s$.

Hyperparameters for training neural network

We have chosen our hyperparameters as follows: one hidden layer with four neurons feeding into an output neuron, a random 80/20 training/test split with a 10 percent validation split. The cross-entropy loss function is minimized via stochastic gradient descent in maximal 300 iterations with a batch size of 200 and an adaptive learning rate, initiated at 0.001. The tolerance is 10^{-4} and the regularization rate is 0.1. Most of these parameters are set to their default value, but we found that the training procedure is largely insensitive to the specific choice of hyperparameters.

Ligand distribution at the decision boundary

In this section, we describe in detail the methods used in the gradient dynamics of changing a ligand distribution to the decision boundary (subsection **Methods**), we provide additional results when adding spatial correlation to the ligand distribution (subsection **MTL pictures**), and we calculate the leading order in small binding time τ_ϵ of the gradient $\frac{dT_{N,m}}{d\tau_\epsilon}$ (subsection **Behavior for small binding times**). We refer to these sections in the main text in subsections **Gradient dynamics identify two different regimes** and **Qualitative change in dynamics is due to a critical point in the gradient** in Chapter 3.2, and in Fig. 3.3.

Methods

Adaptive proofreading is well-suited to characterize the decision boundary between two classes, because we can work with an analytical description. We want to know how to most efficiently change the binding time of the spurious binding ligand (with small τ) to cause the model to reach the decision boundary. We have taken inspiration from [103] and adapted our approach from the iterative FGSM [104]. At first, we sample the binding times τ_{self} for $L_{\text{self}} = 7000$ self ligands from a half-normal distribution $|\mathcal{N}(0, \frac{1}{3})|$ and τ_{ag} for $L_{\text{ag}} = 3000$ agonist ligands from a narrowly peaked normal distribution $|\mathcal{N}(\frac{7}{2}, \frac{1}{10})|$ just above $\tau_d = 3$. We fix the agonist ligand distribution, the "signal" in the immune picture. Next, we bin ligands in M equally spaced bins with center binding time τ_b, $b \in 1, \ldots, M$, and we compute the gradient for bins for which $\tau_b < \tau_d$

$$\frac{\partial T_{N,m}}{\partial \tau_b} = \frac{N\tau_b^{N-1}L_b - mT_{N,m}\tau_b^{m-1}L_b}{\sum_{i=1}^{M}\tau_i^m L_i} \tag{B.10}$$

where L_b is the number of ligands in the b^{th} bin. We subtract this value multiplied by a small number ϵ from the exact binding times, as in Eq. 3.6 in the main text, and we compute a new output $T_{N,m}$. We repeat this procedure until $T_{N,m}$ dips just below the response threshold τ_d^{N-m}. We then display the ligand distributions. We bin ligands and compute the gradient in batches to prevent the gradient from becoming negligibly small. If we would compute the gradient for each ligand with an individual binding time, there would be exactly one ligand with that specific binding time, and because the gradient scales with L, we would need to go through many more iterations. Decreasing the binsize and step size ϵ may enhance the resolution, but is not required. We found good results by considering bins with a binsize of 0.2s and $\epsilon = 0.2$.

MTL pictures

We can visually recast immune recognition as an image recognition problem by placing pixels on a grid and coloring them based on their binding time with a given scale. We chose to let white pixels correspond to not self ($\tau > \tau_d$), gray pixels to antagonist ligands ($\tau_a < \tau < \tau_d$) and black pixels to self ligands $\tau \ll \tau_a$. We are free to introduce any kind of spatial correlation to create "immune pictures" from a ligand

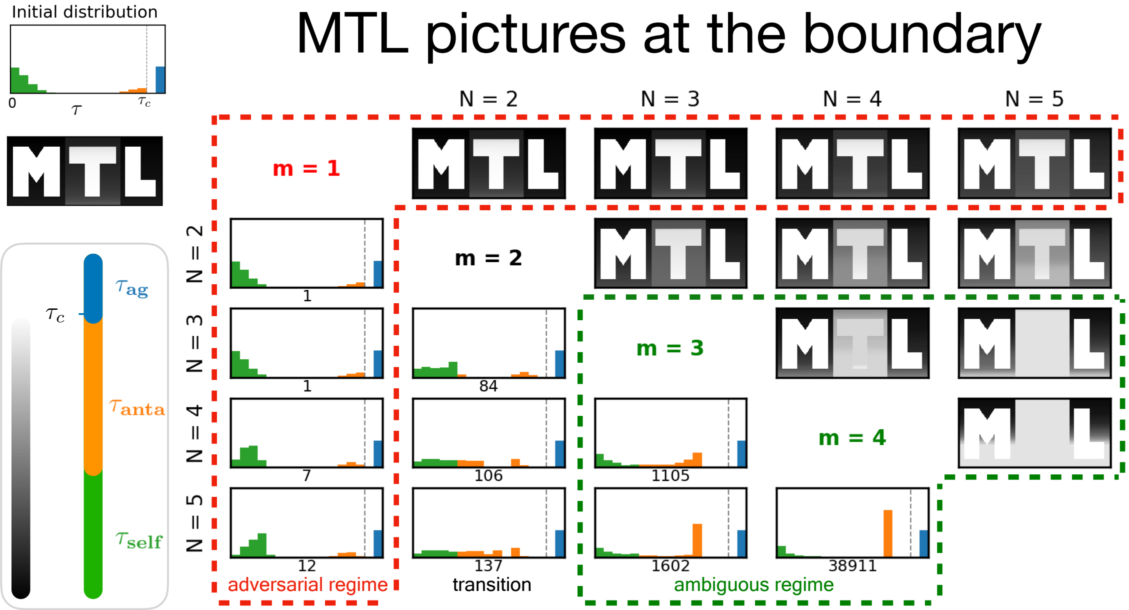


Figure B.1: **MTL pictures.** Explanation is found in the text

distribution. This results in what we term “MTL-pictures” (Fig. B.1). The initial ligand distribution, MTL picture and scale are given on the left. We perform iterative gradient descent like in the main text, and plot the ligand distribution and the corresponding immune pictures at the boundary for various (N, m). The results are striking. For a T cell operating in the adversarial regime, the “signal” MTL is unaltered at the decision boundary. At the transition $m = 2$, we see a slight change of color, while in the ambiguous regime, the signal actually changes from MTL to ML. As we desire for a robust decision-maker, the response should switch when the signal becomes significantly different. From this we conclude, *only in the robust regime can Montreal turn fully into the city of Machine Learning.*

For the MTL pictures in Fig. B.1, we have distributed the pixels in the 179×431 frame – equal to R, the number of receptors – as $L_{\text{self}} = 0.60R$, $L_{\text{a}} = 0.12R$, $L_{\text{ag}} = 0.28R$. We sampled τ_{self} from $|\mathcal{N}(0, \frac{1}{3})|$, τ_{a} from $\tau_d - |\mathcal{N}(0, 13)|$, τ_{ag} from $\tau_d + \mathcal{N}(\frac{1}{2}, \frac{1}{100})$, and we set $\tau_d = 3$. The picture is engineered such that the agonist ligands fill the M and the L, the antagonists fill the T (which is why the T is slightly darker than the M and L). The self ligands fill the area around the letters M, T and L, such that the self with highest binding time surround the T. We have chosen this example to make the effect of proofreading explicit (and of course because we are based in Montreal and study Machine Learning). This result is generic, and the ambiguity of instances at the decision boundary of a robust model can be visualized with any well-designed image. Scripts to reproduce Fig. 3.4 A and Fig. B.1 are available at `https://github.com/tjrademaker/advxs-antagonism-figs/`.

Behavior for small binding times

Consider a mixture with L_{ag} ligands at $\tau_{\text{ag}} > \tau_d$ and L ligands with small binding time $\tau_{\text{spurious}} = \tau_\epsilon \ll \tau_{\text{ag}}$. To understand the behaviour of $T_{N,m}$ as a function of τ_ϵ we expand $T_{N,m}$ in small variable $\epsilon = \frac{\tau_\epsilon}{\tau_{\text{ag}}}$ as

$$\begin{aligned}
T_{N,m}(\{L_{\text{ag}}, \tau_{\text{ag}}; L, \tau_\epsilon\}) &= \frac{\tau_{\text{ag}}^N L_{\text{ag}} + \tau_\epsilon^N L}{\tau_{\text{ag}}^m L_{\text{ag}} + \tau_\epsilon^m L} \\
&= \frac{1 + \epsilon^N \frac{L}{L_{\text{ag}}}}{1 + \epsilon^m \frac{L}{L_{\text{ag}}}} \tau_{\text{ag}}^{N-m} \\
&\simeq \left(1 + \epsilon^N \frac{L}{L_{\text{ag}}}\right)\left(1 - \epsilon^m \frac{L}{L_{\text{ag}}}\right) \tau_{\text{ag}}^{N-m} \\
&\simeq \tau_{\text{ag}}^{N-m} - \tau_{\text{ag}}^{N-m} \frac{L}{L_{\text{ag}}} \epsilon^m + O(\epsilon^N),
\end{aligned}$$

which confirms that up to a constant $T_{N,m} \propto -\epsilon^m \propto -\tau_\epsilon^m$ for $m \geq 1$ and $\tau_\epsilon \ll \tau_{\text{ag}}$, as well as that

$$\frac{dT_{N,m}}{d\tau_\epsilon} \simeq -m\tau_{\text{ag}}^{N-m-1} \frac{L}{L_{\text{ag}}} \epsilon^{m-1} \propto -\tau_\epsilon^{m-1}. \tag{B.11}$$

Boundary tilting

To further draw the connection between machine learning and adaptive proofreading models, we will study a framework to interpret adversarial examples called boundary tilting [105]. We will first illustrate this effect on the discrimination of the original MNIST 3 vs 7 problem MNIST from [74]) (subsection **Digit classification**), after which we will interpret boundary tilting via proofreading in ligand discrimination (subsection **Boundary tilting and categorizing perturbations**), and finally, we will derive how the addition of a subthreshold ligand at the decision boundary changes the output (subsection **Gradient in the L_2 direction**).

Digit classification

A typical 3 and 7 (i), the averages $\bar{3}$ and $\bar{7}$ (ii), and the corresponding adversarial examples (iii, iv) are shown in Fig. B.2 A. Tanay and Griffin [105] pointed out that the adversarial perturbation generated with the Fast Gradient Sign Method (FGSM) proposed in [74] can also be found via $D = \operatorname{sign}(\bar{3} - \bar{7})$, Fig. B.2 A (v). Note the similarity to the adversarial perturbation from the FGSM $\operatorname{sgn}(w) = \operatorname{sgn}(\nabla_x J)$ (Fig. B.2 A (vi)). To reveal the linearity of binary digit discrimination, we computed the principal components (PCs) of the traditional training set of 3s and 7s, and projected all digits in the test set on PC$_1$ and PC$_2$ (Fig. B.2 B). With a linear Support Vector Classifier (ordinary linear regression) trained on the transformed coordinates PC$_1$ and PC$_2$ of the training set, we achieve over 95% accuracy in the test set. While such accuracy is far from the state-of-the-art in digit recognition, it is much higher than typical detection accuracy for single cells (e.g. T cells present false negative rates of 10 % for strong antagonists [31]). The red and blue star in Fig. B.2 A denote the average digit $\bar{3}, \bar{7}$.

Next, we transformed the test set as $3 \rightarrow 3' = 3 - \epsilon_{\text{test}} D, \;\; 7 \rightarrow 7' = 7 + \epsilon_{\text{test}} D$, where $\epsilon_{\text{test}} = 0.4$ is the strength of the adversarial perturbation (Fig. B.2 A (iii)). $\bar{3}'$ and $\bar{7}'$ moved closer in Fig. B.2 B, orthogonal to the decision boundary and along the line between the initial averages. This adversarial perturbation moves the digits in what we call an adversarial direction perpendicular to the decision boundary, and reduces the accuracy of the linear regression model to a mere 69%.

Goodfellow et al. proposed adversarial training as a method to mitigate adversarial effects by FGSM. We implemented adversarial training by adding the adversarial perturbation $\epsilon_{\text{train}} D_{\text{train}} = \epsilon_{\text{train}}(\bar{3}_{\text{train}} - \bar{7}_{\text{train}})$ to the images in the training set, computing the new PCs and training the linear regression model. This effectively “tilts” the decision boundary, while preserving 95% accuracy. In the presence of the original adversarial perturbations, we see the effect of the tilted boundary: the perturbation moves digits parallel along the decision boundary, which results in good robust ac-

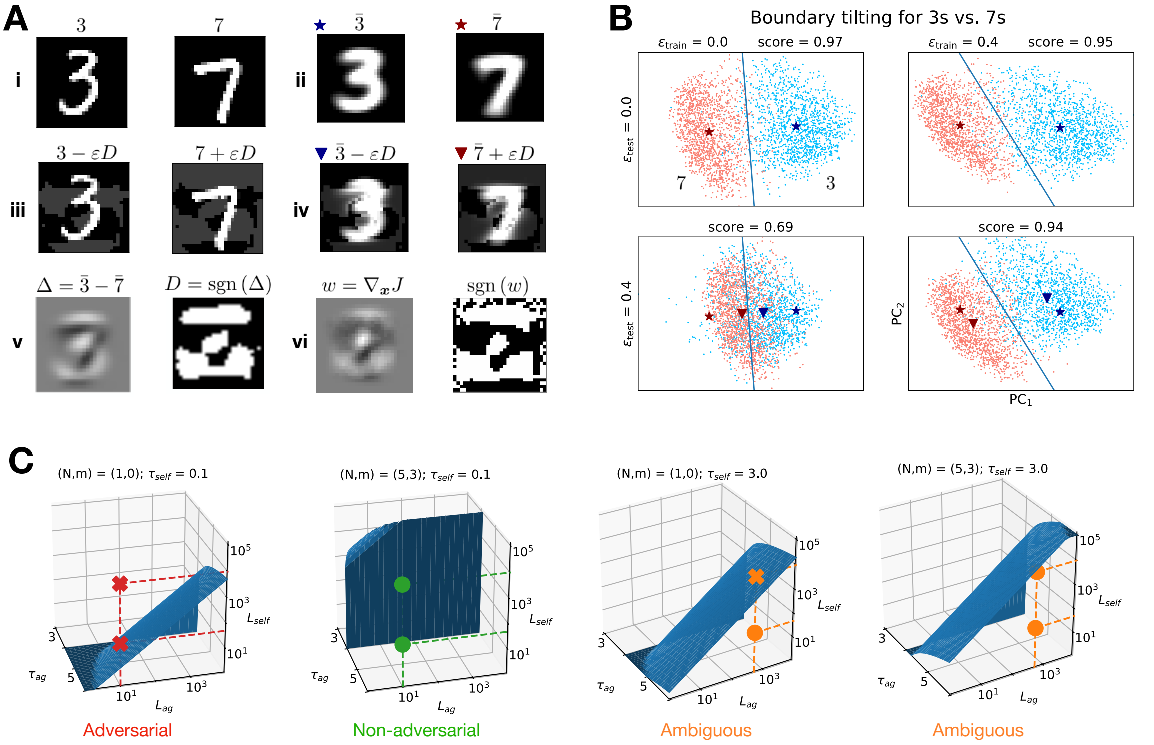


Figure B.2: **Boundary tilting in one-dimensional digit classification.** (A) (i) Typical 3 and 7 from MNIST. (ii) Average 3, 7 of the traditional test set, (iii, iv) with adversarial perturbation, found by (v) subtracting the sign of $\bar{3}$ from $\bar{7}$, which corresponds to (vi), the perturbation found with FGSM (B) Projection of the digits on the first principal components. The classes are separated by a linear Support Vector Classifier (blue), the average of the classes with and without adversarial perturbation is shown by the triangle and star. We have cycled through permutations of adversarial training and/or adversarial testing. Note how the boundary tilts on the right panels, and how the triangle moves parallel to the decision boundary. (C) Decision boundary of the immune model. The region under the surface is the response regime, the region above is the no-response regime. The classifier with a single proofreading step $(N, m) = (1, 0)$ fails to observe agonists in three of the four marked mixtures, while the robust classifier $(N, m) = (5, 3)$ correctly responds to each indicated mixture.

curacy. This is an illustration of the more general phenomenon studied in [105].

Boundary tilting and categorizing perturbations

We consider the change in $T_{N,m}$ for arbitrary N, m upon addition of many spurious ligands. Generalizing Eq. 3.2 gives

$$T_{N,m}^{\text{after}} = \frac{L(\tau-\epsilon)^N + \epsilon R \epsilon^N}{L\tau^m + \epsilon R \epsilon^m} = \frac{(\tau-\epsilon)^N + \frac{\epsilon^{N+1}R}{L}}{\tau^m + \frac{\epsilon^{m+1}R}{L}}. \tag{B.12}$$

From this expression, we note that $T_{N,m}$ is changing significantly with respect to its initial value upon addition of many weakly bound ligands as soon as $\epsilon^{m+1}R$ is of order L. Thus, the effect described in subsection **Adaptive proofreading for cellular decision-making** in Chapter 3 for weighted averages where $(N, m) = (1, 0)$ also holds for nonlinear computations as long as m is small. It appears that the general strategy to defend against this adversarial perturbation is by increasing m, as previously observed in [34]. Biochemically, this is done with kinetic proofreading [26, 31, 33], i.e. we take an output $T_{N,m}$ with $N > m \geq 1$. Here, the output is no longer sensitive to the addition of many weakly bound self ligands, yielding an inversion of the antagonistic hierarchy where the strongest antagonizing ligands exist closer to threshold [60]. An extreme case has been proposed for immune recognition where the strongest antagonists are found just below the threshold of activation [31].

We numerically compute how the decision boundary changes when L_{self} ligands at τ_{self} are added to the initial L_{ag} agonist ligands at τ_{ag}, i.e. we compute the manifold so that

$$T_{N,m}(\{L_{\text{ag}}, \tau_{\text{ag}}; L_{\text{self}}, \tau_{\text{self}}\}) = \frac{\tau_{\text{ag}}^N L_{\text{ag}} + \tau_{\text{self}}^N L_{\text{self}}}{\tau_{\text{ag}}^m L_{\text{ag}} + \tau_{\text{self}}^m L_{\text{self}}} \tag{B.13}$$

is equal to $T_{N,m}(\{L_{\text{ag}}, \tau_d\}) = \tau_d^{N-m}$. We represent this boundary for fixed τ_{self} and variable $L_{\text{ag}}, L_{\text{self}}, \tau_{\text{ag}}$ in Fig. B.2 C. Boundary tilting is studied with respect to the reference $L_{\text{self}} = 0$ plane corresponding to the situation of pure L_{ag} ligands at τ_{ag}, where the boundary is the line $\tau_{\text{ag}} = \tau_d$. The case $(N, m) = (1, 0)$ (Fig. B.2 C, left panel), corresponds to a very tilted boundary, close to the plane $L_{\text{self}} = 0$, and a strong antagonistic case. In this situation, assuming $\tau_{\text{ag}} \simeq \tau_d$, each new ligand added with τ_{self} close to 0 gives a reduction of $T_{1,0}$ proportional to $\frac{\tau_d}{L_{\text{ag}}}$ in the limit of small L_{self} (see next section, [49]), which is again of the order of the response

Table B.1: Categories of perturbations

	Boundary tilting	Gradient when adding one antagonistic ligand
Adversarial	yes	steep ($\mathcal{O}(1)$)
Non-adversarial	no	almost flat ($\mathcal{O}(\epsilon^m)$)
Ambiguous	yes	weak ($\mathcal{O}(\epsilon)$)

$T_{1,0} = \tau_{\text{ag}} \simeq \tau_d$ in the plane $L_{\text{self}} = 0$. This is clearly not infinitesimal, corresponding to a steep gradient of $T_{1,0}$ in the L_{self} direction. We call the perturbation in this case adversarial. This should be contrasted to the case for higher m (Fig. B.2 C, middle left) where the boundary is vertical, independent of L_{self}, such that decision-making is based only on the initially present L_{ag} ligands at τ_{ag}. Here, the change of response induced by the addition of each ligand with small binding time τ_{self} is τ_{self}^m, due to proofreading a very small number when $\tau_{\text{self}} \simeq 0$ [49]. Contrary to the previous case, the gradient of $T_{N,m}$ with respect to this vertical direction is almost flat and very small compared to the response in the $L_{\text{self}} = 0$ plane. We call the perturbation in this case non-adversarial.

Tilting of the boundary only occurs when τ_{self} gets sufficiently close to the threshold binding time τ_d (Fig. B.2 C, right panels). In this regime, each new ligand added with quality $\tau_{\text{self}} = \tau_d - \epsilon$ contributes an infinitesimal change of $T_{N,m}$ proportional to $\frac{\tau_d - \tau_{\text{self}}}{L_{\text{ag}}} = \epsilon / L_{\text{ag}}$, which gives a weak gradient in the direction L_{self}. But even with such small perturbations one can easily cross the boundary because of the proximity of τ_{self} to τ_d, which explains the tilting. The cases where the boundary is tilted and the gradient is weak are of a different nature compared to the adversarial case of Fig. B.2 C, left panel. Here the boundary is tilted as well, but the gradient is steep, not weak. For this reason we term the cases on the right panels ambiguous. Similar ambiguity is observed experimentally: it is well known that antagonists (ligands close to thresholds) also weakly agonize an immune response [31]. Our categorization of perturbations is presented in Table B.1. Scripts for boundary tilting in ligand discrimination and digit discrimination are available at `https://github.com/tjrademaker/advxs-antagonism-figs/`.

Gradient in the L_2 direction

We recall results from [60] to show how the addition of subthreshold ligands one at a time changes the output. We first consider $\{L, \tau_d\}$ threshold ligands with output

$$T_{N,m}(L, \tau_d) = \tau_d^{N-m}. \tag{B.14}$$

The main result of [60] is the linear response of $T_{N,m}(L, \tau_d)$ to the addition of $\{L_a, \tau_d - \epsilon\}$ subthreshold ligands.

$$T_{N,m}\left(\{L, \tau_d; L_a, \tau_d - \epsilon\}\right) \tag{B.15}$$
$$= T\left(L + L_a, \tau_d\right) - \epsilon L_a \mathcal{A}\left(L + L_a, \tau_d\right) \tag{B.16}$$
$$= \tau_d^{N-m} - \epsilon \frac{L_a}{L + L_a} \frac{d}{d\tau} T_{N,m}(L + L_a, \tau)\Big|_{\tau=\tau_d}, \tag{B.17}$$

where we used the definition

$$\mathcal{A}\left(L, \tau_d\right) = \frac{1}{L} \frac{d}{d\tau} T_{N,m}(L, \tau)\Big|_{\tau=\tau_d}. \tag{B.18}$$

for the coefficient in a mean-field description. As the derivative $\frac{d}{d\tau} T_{N,m}(L, \tau)\Big|_{\tau=\tau_d} > 0$, and $\epsilon = \tau_a - \tau_d$, each additional subthreshold ligand at τ_a decreases the output with a value proportional to

$$\frac{\tau_d - \tau_a}{L}. \tag{B.19}$$

In the case $(N, m) = (1, 0)$, the mean-field approximation is exact, i.e. the first derivative of $\frac{dT}{d\tau}$ is the only nonzero derivative, given by

$$\mathcal{A}(L, \tau_d) = \frac{1}{L} \frac{d}{d\tau} \tau \Big|_{\tau=\tau_d} = \frac{1}{L}. \tag{B.20}$$

With the addition of a single subthreshold ligand $\tau_a \simeq 0$, so that $\epsilon \simeq \tau_d$, the output is maximally reduced by $\frac{\tau_d}{L+1} \simeq \frac{\tau_d}{L}$, a finite quantity, as described in subsection **Fast Gradient Sign Method recovers antagonism by weakly binding ligands** in Chapter 3. For higher m, the linear approximation holds only for ligands at τ_a close to threshold.

Few-pixel attack

In this section, we describe in detail the procedure for the few-pixel attack. We used this to come to our conclusion in subsection **Biomimetic defenses against few-pixel attacks** and Fig. 3.4 C in Chapter 3.2.

The few-pixel attack connects to ligand antagonism in the sense that few pixels are needed to cause misclassification, corresponding to the addition of few maximally antagonizing ligands to a mixture fooling robust adaptive proofreading models. It is not the most efficient attack against a classifier without biomimetic defence, but it is the most efficient attack against classifiers with biomimetic defence, equivalent to adaptive proofreading models with $m > 1$. For these adaptive proofreading models, there exists a unique maximally antagonistic binding time, defined as the binding time that maximally reduces $T_{N,m}$.

With this in mind, we decided to make pixels black or white in a controlled manner, until the neural network classifies the perturbed, initial digit as the target class. In the following, we will refer to several stages of the few-pixel attack using Fig. B.3. We first computed what we term pixelmaps. Pixelmaps contain the change of score when making a pixel white or black. In Fig. B.3, blue colors correspond to pixels that will lower the score when turned white or black, while red colors are for pixels that will increase the score for the same operation. A grey color means the score is unchanged when whitening or blacking the pixel. The pixelmaps are scaled to the maximum change in score. We proceed in merging and sorting the pixelmaps from maximum to minimum change in score towards the target class, iteratively following the sorted list to decide which pixels in our digit to turn white or black. We do this until we reach the decision boundary (first iteration in which the digit is misclassified). The final digits in the row above the red rectangle in Fig. B.3 are the resulting boundary digits. They already contain perturbations corresponding to real features, but have an air of artificiality to them which allows us to fairly easily distill the ground truth. We remove this with a mean filtering [107], which is a 3x3 convolutional block that computes mean pixel values as

$$y_{i,j} = \frac{1}{9} \sum_{k,l=-1}^{1} x_{i+k,j+l}. \tag{B.21}$$

Biologically, this is pure receptor clustering, where a perturbation to a single receptor locally affects other ligands. Such digits are truly ambiguous digits that are tough to classify even as humans. These are the type of digits we expect to find on the decision boundary. Finally, we compare the mean-filtered digit at the decision boundary to the control: the sum of the initial digit and the hill function of Eq. 3.8 $(N = 3; \theta = 0.5)$ on the average of all digits in the target class, then mean-filtered (Fig. B.3 for a step-by-step composition). We apply the mean-filter to the control to again remove the artificiality of a digit plus an average, and make the comparison between boundary digit and control digit fairer. The similarity between mean-filtered boundary digit and control digit confirms our intuition that we are actually operating in the space between both classes when misclassification occurs.

We can also apply the mean-filter to the initial digit before generating the pixelmaps, and during the procedure, check the score on the mean-filtered perturbed image. This gives similar results, as we see by following the trajectory of the score for *boundary-null* and *boundary-mean*. We have shown the score explicitly in Fig. B.4 for the digits in Fig. B.3. The behavior of the score is remarkably similar to the interpolation between ligand mixtures (Fig. 3.3F, bottom panel). A nonlinear filtering method proposed in [107] is the median-filter, but this one works less well for black-and-white pixels.

We have shown examples that are generated when we select for instances where the number of iterations is large enough (20 suffices, we still consider this to be a few-pixel attack, keeping in mind that digits have 784 individual pixels). The authors of [106] specifically searched for single pixel attacks. Examples of single-pixel misclassification exist in our neural networks trained on two types of digits in MNIST too, but these we found non-informative. In cellular decision-making, this case corresponds to adding a single antagonist ligand to a ligand mixture to cause misclassification. This is only possible if the ligand mixture is already very close to the boundary. For such

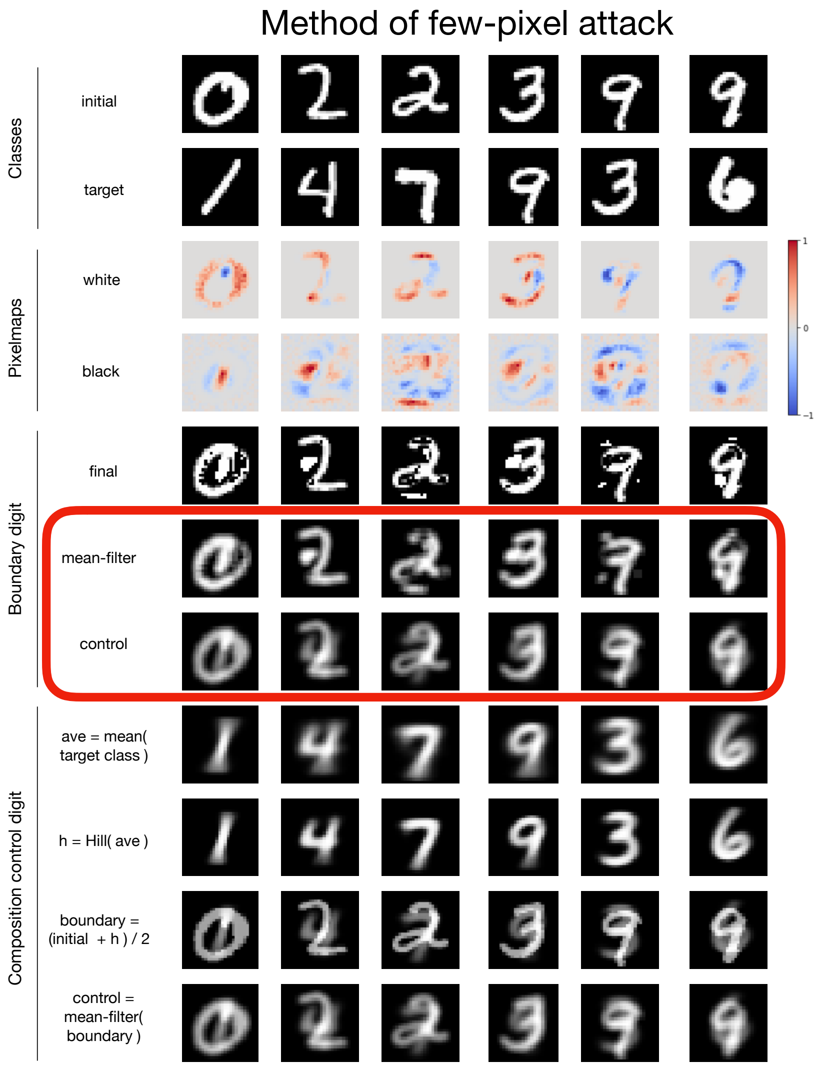

Method of few-pixel attack
Classes
initial
target
Pixelmaps
white
black
1
0
−1
Boundary digit
final
mean-filter
control
Composition control digit
ave = mean(target class)
h = Hill(ave)
boundary = (initial + h) / 2
control = mean-filter(boundary)

Figure B.3: **Method of few-pixel attack**. Each column show how a few-pixel attack causes misclassification of an initial digit to a target class. The important result are the pre-filtered boundary digits and the control in the red rectangle. Pixelmaps determine which pixels increase (red) or decrease (blue) the score when turning an individual pixel in the initial digit white or black. We merge the pixelmaps, sort this list of pixels, and go through it from maximum to minimum change in score until misclassification occurs, resulting in the pre-filtered digit. We apply a mean-filter to make them look more like real digits, and indeed, these mean-filtered boundary digits closely resemble our control digits at the boundary. The control digits are composed of the mean-filtered initial digit plus locally contrasted (with hill function ($N = 3; \theta = 0.5$) average digit of the target class.

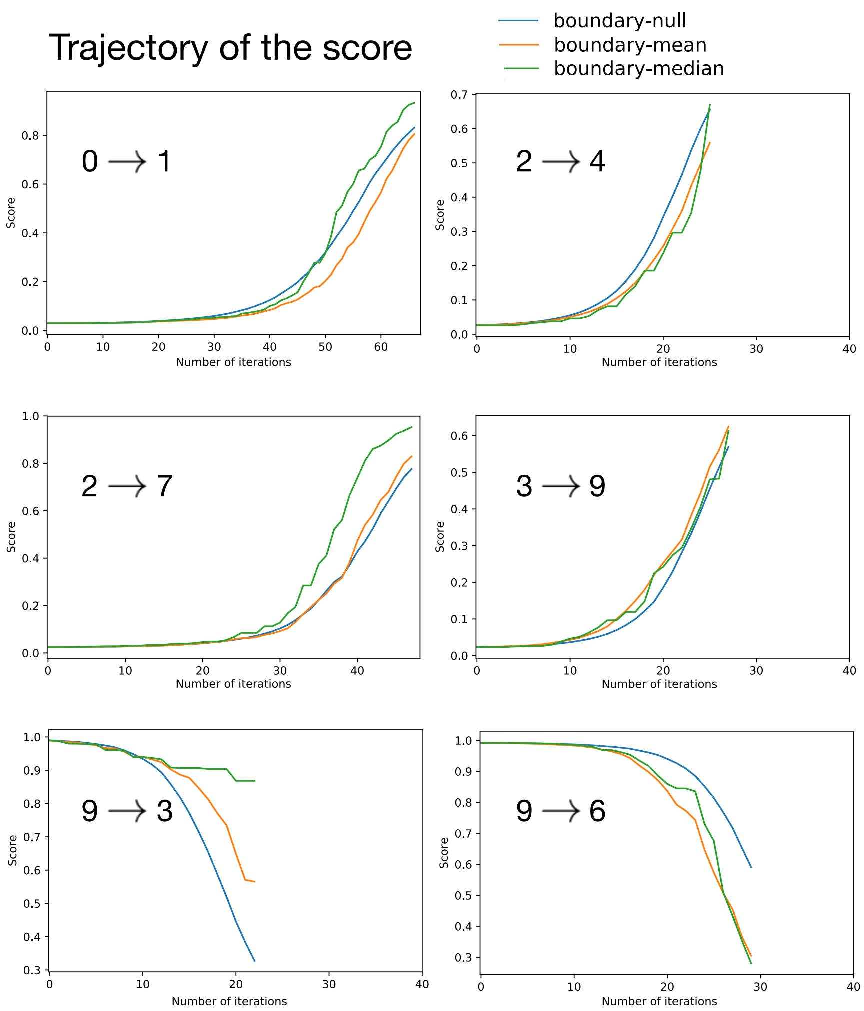


Figure B.4: **Trajectory of the scoring functions of the attacks in Fig.B.3.** The blue, orange and green line correspond to various digits (actual digit, mean-filtered digit, median-filtered digit) for which we check the score, and terminate when reaching the boundary. The trajectory of the score for the null digit and the mean-filtered digit is generally the same. Moreover, the behavior of the score looks similar to the behavior of $T_{N,m}$ upon addition of maximally antagonizing ligands to a mixture of only agonist ligands in Fig. 3.3 D.

samples, we do not expect ambiguity to appear. Remember that near the boundary, the score landscape is steep, and small additions have a large effect.

” *(Attack and defence [2])*

C Cytokines

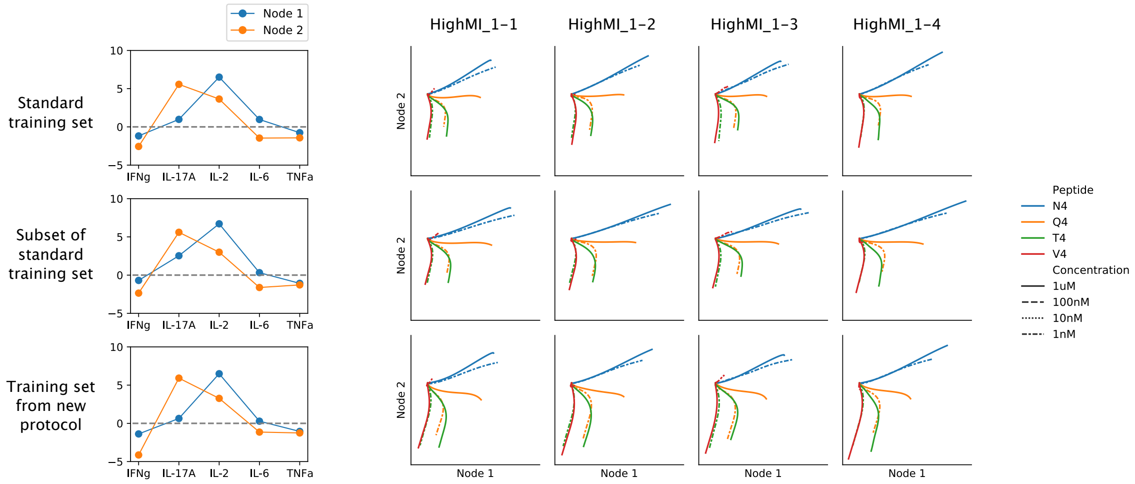


Figure C.1: **Comparison between training classifier on data from old protocol and new protocol.** Top row shows learned weights and latent space for classifier trained with standard dataset (6 datasets from old protocol). Middle row shows learned weights and latent space for classifier trained with subset of standard dataset (3 datasets from old protocol). Bottom row shows learned weights and latent space for classifier trained with 4 datasets from new protocol. Timeseries from the same reproducibility dataset are used to project on the latent spaces. There exist minor differences in the learned weights and the latent space, which is mainly due to different range for the normalization. This validates that the new protocol does not affect the results in any meaningful, and that it is okay to extract supernatant for measurement.

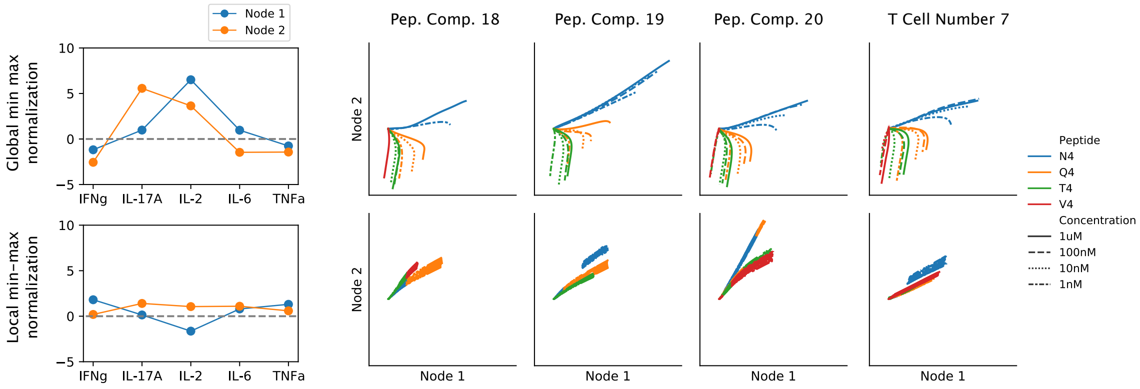


Figure C.2: **Comparison of normalization procedures.** Top panel: standard min-max normalization by dividing each training set by the global minimum and maximum per cytokine resulting in typical learned weights and latent space dynamics. Dividing each training set by minimum and maximum per cytokine per dataset gives warped learned weights and a non-interpretable latent space dynamics. Due to the diversity between conditions in an experiment, as well as within the same conditions between experiments, min-max normalization for each dataset individually strongly biases the training set towards dataset with unexpectedly low or high cytokine concentrations, and does not allow for the retrieval of correlation between cytokines and antigen quality that exists across datasets.

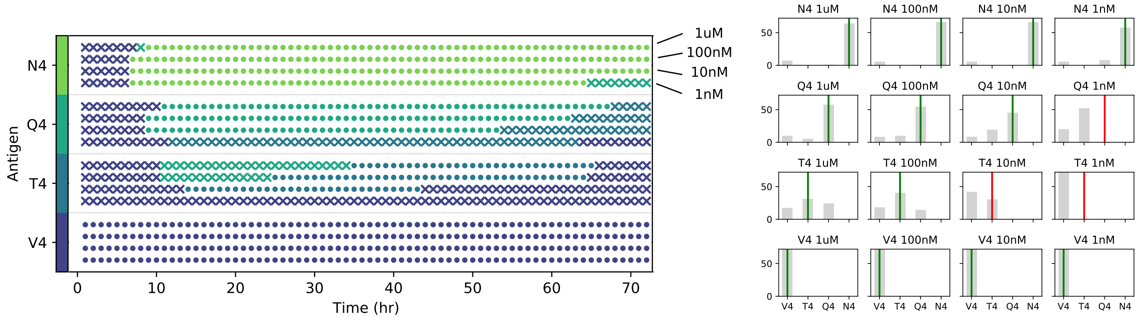


Figure C.3: **Timeseries classification procedure.** Left panel: classification of a timeserie of given quality (set of four rows indicated by antigen name and color on the left) and quantity (four subrows per antigen quality indicated by antigen quantity on the right). Circle (cross) is correct (incorrect) classification. Color of the marker indicates what antigen was predicted. Right panel: summing individual timepoints per timeseries of given quality (rows) and quantity (columns). The timeseries prediction is the antigen with the most timepoints, indicated by the vertical line. Green (red) line indicates a correct (incorrect) prediction.

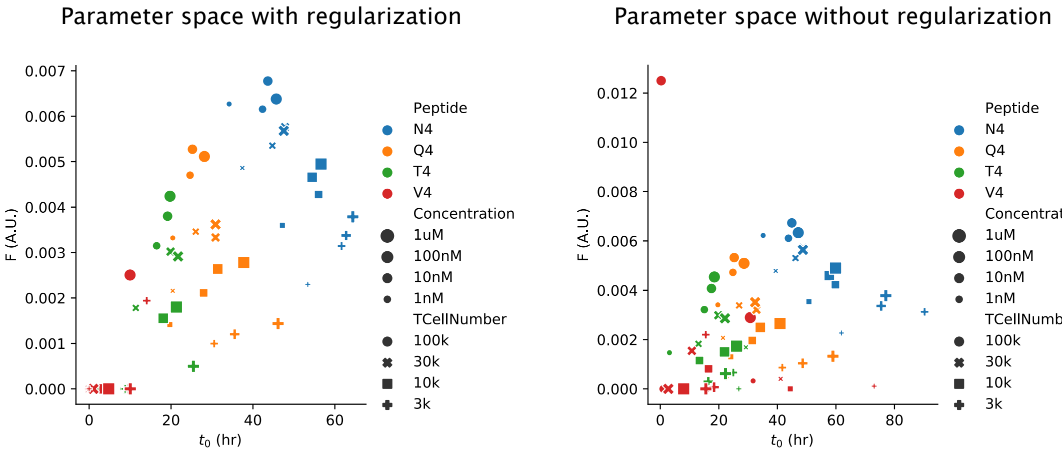


Figure C.4: **Comparison parameter spaces with and without regularization.** Left panel shows parameter space obtained by fitting constant force model including a regularization term, right panel shows parameter space without regularization. The difference is subtle but important. On the right, the y-axis is extended because of a timeseries that is fitted with very high F and $t_0 \simeq 0$. Moreover, timeseries with very small F may have quite large t_0 up to 40 hour on the right, while they are constraint to $t_0 \leq 10$ hour with regularization. L1 regularization forces indeterminate terms to near-zero values, which is the correct thing to do for conditions with a small response, meaning a small F and t_0. The value of the regularization constant is found to be not of importance; here it is set to 1.

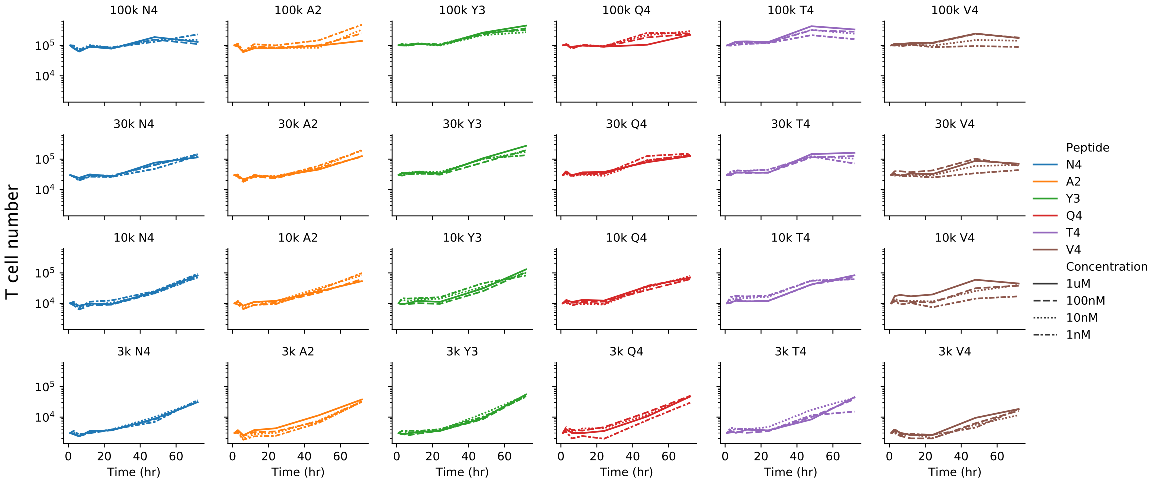


Figure C.5: **T cell numbers over time.** Number of events the flow cytometer records per measurement corrected by the relative proportion of events / initial T cell number per timeseries. 85 - 90 % of the T cells are lost following washes and stains.

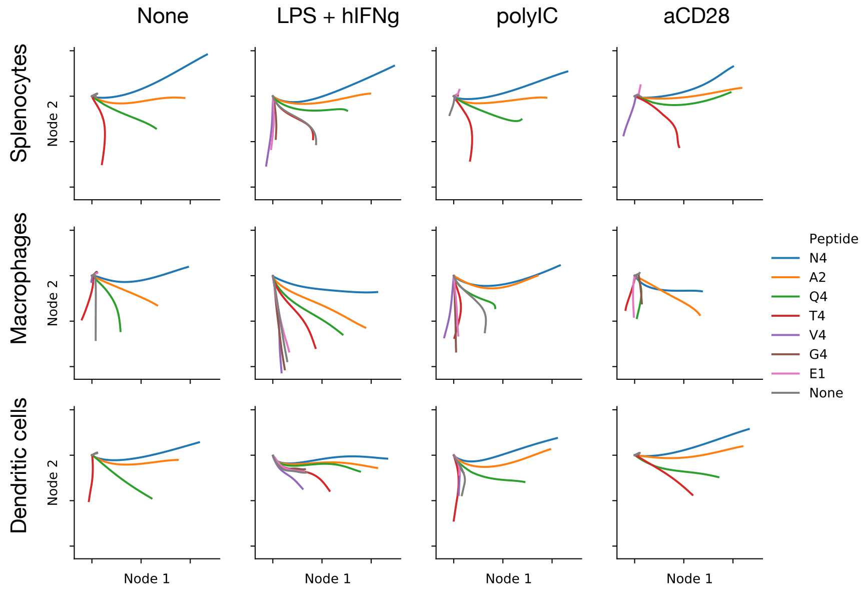


Figure C.6: **Integral latent spaces for APC experiment.** APC types are in the rows (from top to bottom: splenocytes, macrophages, dendritic cells), TLR agonist in the columns (from left to right: None, LPS + hIFNγ, poly I:C, aCD28), colors indicate antigen.