Attack and defence in cellular decision-making: lessons from machine learning

Thomas J. Rademaker [1], Emmanuel Bengio [2] and Paul François [1]
[1]*Department of Physics, McGill University, Montreal, Quebec, Canada*
[2]*School of Computer Science, McGill University, Montreal, Quebec, Canada*
(Dated: June 15, 2019)

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 for 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.

INTRODUCTION

Machine learning is becoming increasingly popular with major advances coming from deep neural networks [1]. Deep learning has improved the state-of-the-art in automated tasks like image processing [2], speech recognition [3] and machine translation [4], 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 [5]. 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 [6]. Real world scenarios exist, like adversarial road signs fooling computer vision algorithms (Fig. 1 A) [7], or adversarial perturbations on medical images triggering incorrect diagnosis [8]. Worse, adversarial examples are often transferable across algorithms (see [9] for a recent review), and certain universal perturbations fool any algorithm. [10].

Categorization and inference are also tasks found in cellular decision-making [11]. 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 [12, 13]. Decision-making in an immune context is equally prone to detrimental perturbations in a phenomenon called ligand antagonism [14]. Antagonism appears to be a general feature of cellular decision-makers: it has been observed in T cells [15], mast cells [16] and other recognition processes like olfactory sensing [17, 18].

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 [19–21]. Recent examples might include neoantigens in cancer [22, 23] which are implicated in tumour immunoediting and escape from the immune system. Those medical examples are reminiscent of how adversaries could generate black box attacks aimed at fooling neural networks [7]. Strategies for provable defenses or robust detection of adversarial examples [24, 25] 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 [26] 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 perturba-

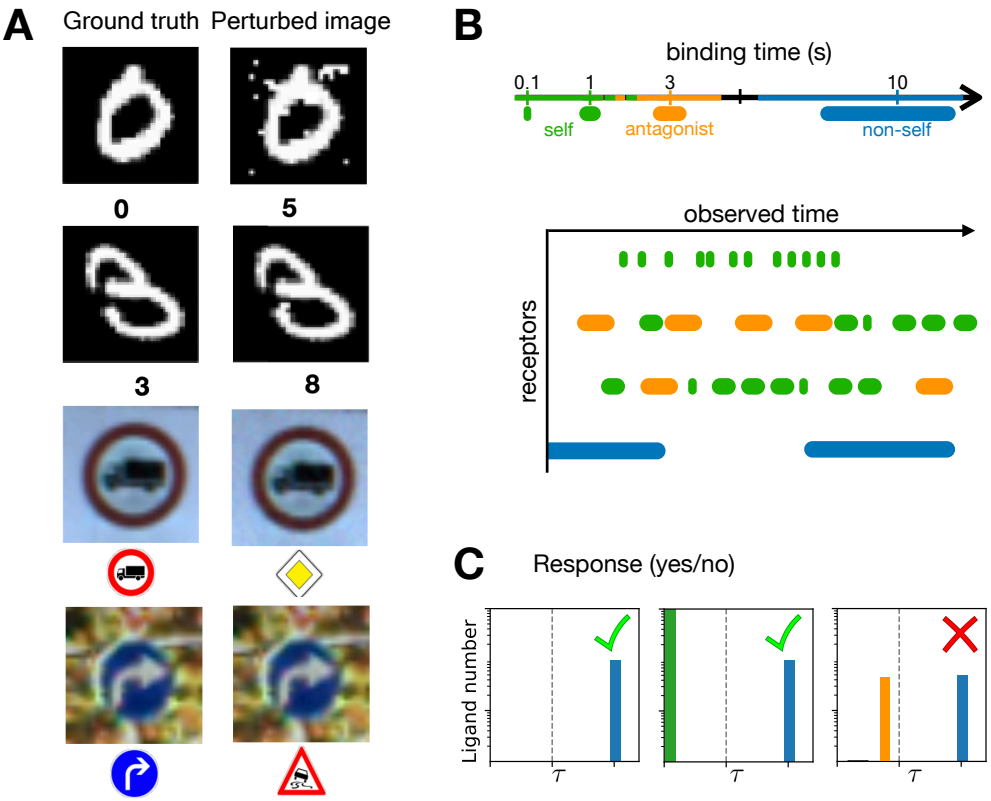

FIG. 1. **Ligand discrimination and digit recognition tasks** (A) Adversarial examples on digits and roadsigns. Reproduced from [7]. Left column displays original images with categories recognized by machine learning algorithms, right column displays images containing targeted perturbations leading to 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 $k_i^{\rm on}$. (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 responds 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.

tions 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 [13, 27, 28]. 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 [13], 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 [29–31]). 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. 1 B). The challenge is to ignore many subthreshold ligands ($\tau < \tau_d$) while responding to few agonist ligands with $\tau > \tau_d$ [13, 15, 32]. In particular, it is known experimentally in many different contexts that addition of antagonistic subthreshold ligands can impair proper decision-making (Fig. 1 C) [15–17].

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 [14, 33], and can be shown to capture all relevant features of such cellular decision-making close to a

decision threshold [34]. An example of such a model is displayed in Fig. 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 [35]. 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. 2 A is presented in Appendix S1.

We assume an idealized situation where a given receptor i, upon ligand binding (on-rate $k_i^{\rm on}$, binding time τ_i) can exist in N biochemical states (corresponding to phosphorylation stages of the receptor tails in the immune context [36, 37]). Those states allow the receptor to effectively compute different quantities, such as $c_n^i = k_i^{\rm on}\tau_i^n, 0 \leq n \leq N$, which can be done with kinetic proofreading [36, 38, 39]. 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) $C_n = \sum_i c_n^i = \sum_i k_i^{\rm on}\tau_i^n$. 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 $k_i^{\rm on}$ and τ_i are functions of time and stochastic treatments are required [11, 40, 41], but on the time-scale of primary decision-making it is reasonable to assume that the ligand distribution does not change much [15].

Adaptive proofreading models rely on an incoherent feed-forward loop, where an output is at the same time activated and repressed by bound ligands via two different branches in a biochemical network (Fig. 2 A). An explicit biochemical example is shown on the right panel of Fig. 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 $k_{\rm on}$ 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 $k_{\rm on}L\tau^n$. 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 $K \propto (k_{\rm on}L\tau^m)^{-1}$ 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 [42]). 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

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

and at steady state

$$C_N \propto \frac{k^{\rm on}L\tau^N}{k^{\rm on}L\tau^m} = \tau^{N-m} \tag{2}$$

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 [15, 33].

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 $k_{\rm on}L\tau^n$ can be replaced by $\sum_i k_i^{\rm on}L_i\tau_i^N$, calling L_i the quantity of ligands with identical $k_i^{\rm on}, \tau_i$. We then define the generalized output of the biochemical network as

$$T_{N,m} = \frac{\sum_i k_i^{\rm on}L_i\tau_i^N}{\sum_i k_i^{\rm on}L_i\tau_i^m}. \tag{3}$$

Similar equations for an output $T_{N,m}$ can be derived for many types of networks, as described in [35]. 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 $N = 1, m = 0$ [17] (see also another example in [34]) . 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. 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. $\tau = 3\,$s), one never reaches the detection threshold (dotted line on Fig. 2 B, left panel) even for many ligands. For slightly bigger $\tau = 10\,{\rm s} > \tau_{\rm d}$, 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 $L_{\rm a}$ antagonists with lower binding time $\tau_{\rm a} < \tau$ and equal on-rate $k^{\rm on}$, we have $T_{N,m} = \frac{L\tau^N + L_{\rm a}\tau_{\rm a}^N}{L\tau^m + L_{\rm a}\tau_{\rm a}^m}$, which

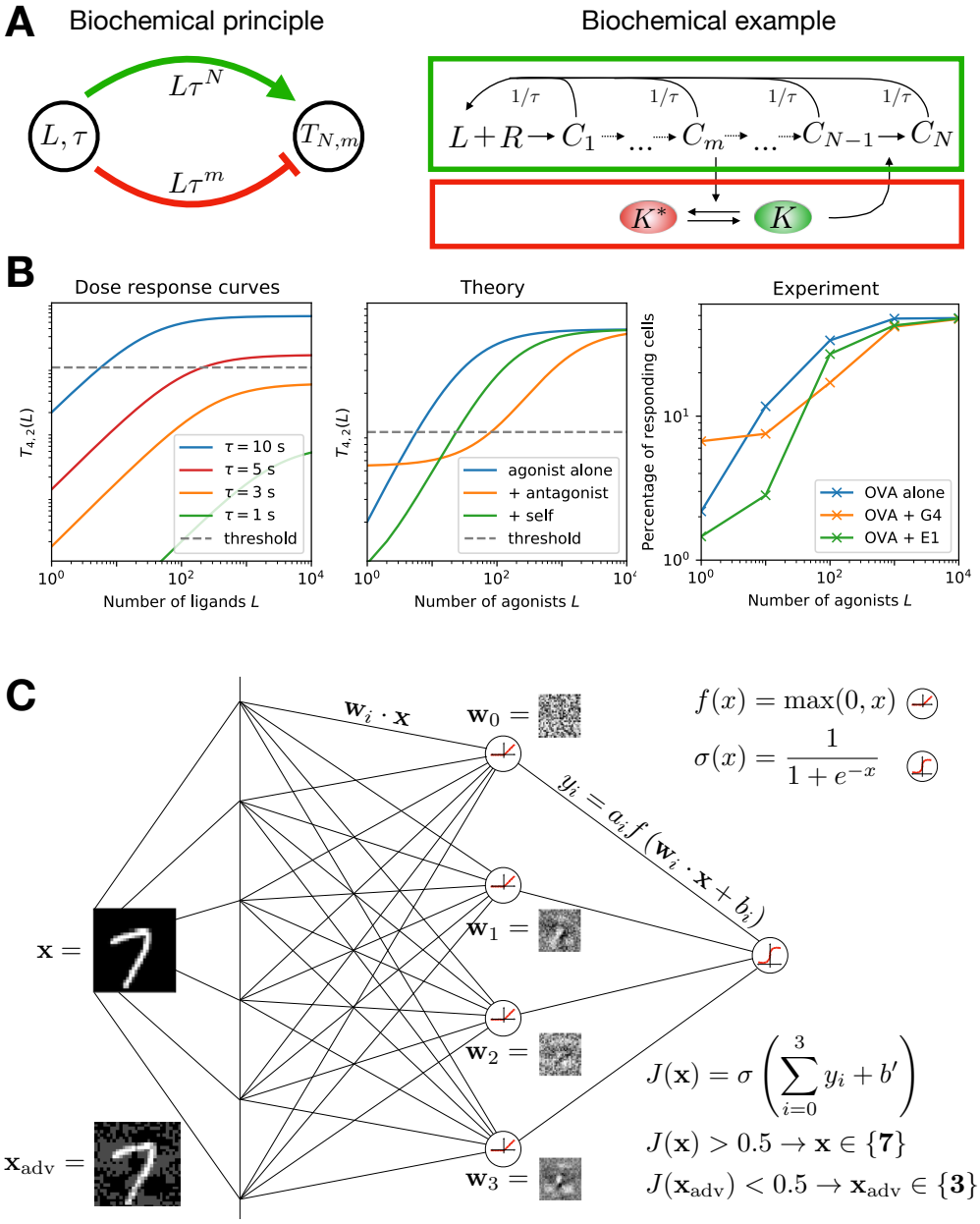

FIG. 2. **Adaptive proofreading and neural network** (A) Left: Adaptive proofreading networks have an activating and repressing branch with different weights on τ. Right: detailed adaptive proofreading network adapted from [34]. Ligand L binds to receptor R to form unphosphorylated complex C_0. The receptor 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 [32]). 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 $\mathbf{x}_{\text{adv}}$.)

is smaller than the response τ^{N-m} for a single type of ligands, corresponding to ligand antagonism (Fig. 2 B, middle panel) [14, 15, 43, 44]. 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. 2 B, right panel, data redrawn from [32]). 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 [14, 33, 34].

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 [8]), using as a case-study image classification from two types of digits. These images are taken from MNIST [45], 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 [6]. Fig. 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 [46], see Appendix S2. As an example, in Fig. 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 ($J(\mathbf{x}_{\text{adv}}) < 0.5$).

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 $28 \times 28 = 784$, while in immunology ligands can bind to roughly 30000 receptors [15]. 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 [6]. The FGSM computes the local maximum adversarial perturbation $\eta = \epsilon \,\mathrm{sgn}\,(\nabla_x J)$ (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 [6]). 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. 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 $k_i^{\rm on}$ (corresponding to the rate at which ligands bind to receptors, also called on-rate [47]) and τ_i (corresponding to quality). If a receptor i is unoccupied, we set its k_i and τ_i to 0 [48]. 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 [17], with the role of $k^{\rm on}$ 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 $\tau_{\rm avg}$ of ligands presented weighted by $k_i^{\rm on}$ (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 [14, 17] for explicit examples.

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

$$\eta = \epsilon\, \mathrm{sgn}\begin{pmatrix} \partial_{\tau_i} J \\ \partial_{k_i^{\rm on}} J \end{pmatrix} = \epsilon\, \mathrm{sgn}(A)\mathrm{sgn}\begin{pmatrix} k_i^{\rm on} \\ \tau_i - T_{1,0} \end{pmatrix}, \qquad (4)$$

where $A = \frac{J'(T_{1,0})}{\sum k_i^{\rm on}} > 0$. 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 A).

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

The key relation to adversarial examples from [6] comes from considering what happens to the unbound receptors for which both $k_i^{\rm on}$ and τ_i are initially 0. Let us consider a situation with L identical bound ligands with ($k^{\rm on} = 1$, binding time τ) giving response $T_{1,0}^{\rm before} = \tau$ where τ itself is of order 1 (i.e. much bigger than the ϵ-sized perturbation on binding time considered in Eq. 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 $k^{\rm on} = \epsilon$, with small binding time ϵ. We compute the new response to be

$$T_{1,0}^{\rm after} = \frac{L(\tau - \epsilon) + \epsilon R \epsilon}{L + \epsilon R} = \frac{\tau - \epsilon + \frac{\epsilon R}{L}\epsilon}{1 + \frac{\epsilon R}{L}} \qquad (5)$$

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

$$\frac{T_{1,0}^{\rm after} - T_{1,0}^{\rm before}}{T_{1,0}^{\rm before}} \simeq -\frac{\frac{\epsilon R}{L}}{1 + \frac{\epsilon R}{L}} \qquad (6)$$

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 [16], a so-called "dog in the manger" effect. Another example is likely found in detection by NK cells [27]. A similar effect called "competitive antagonism" is also observed in olfaction where ligands with strong inverse affinity can impinge action of other ligands [17]. 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 $k^{\rm on}$ is low. Since we consider the product of both terms, both situations lead to similar effects, but our focus on a small change of $k^{\rm on}$ 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 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. 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 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 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.

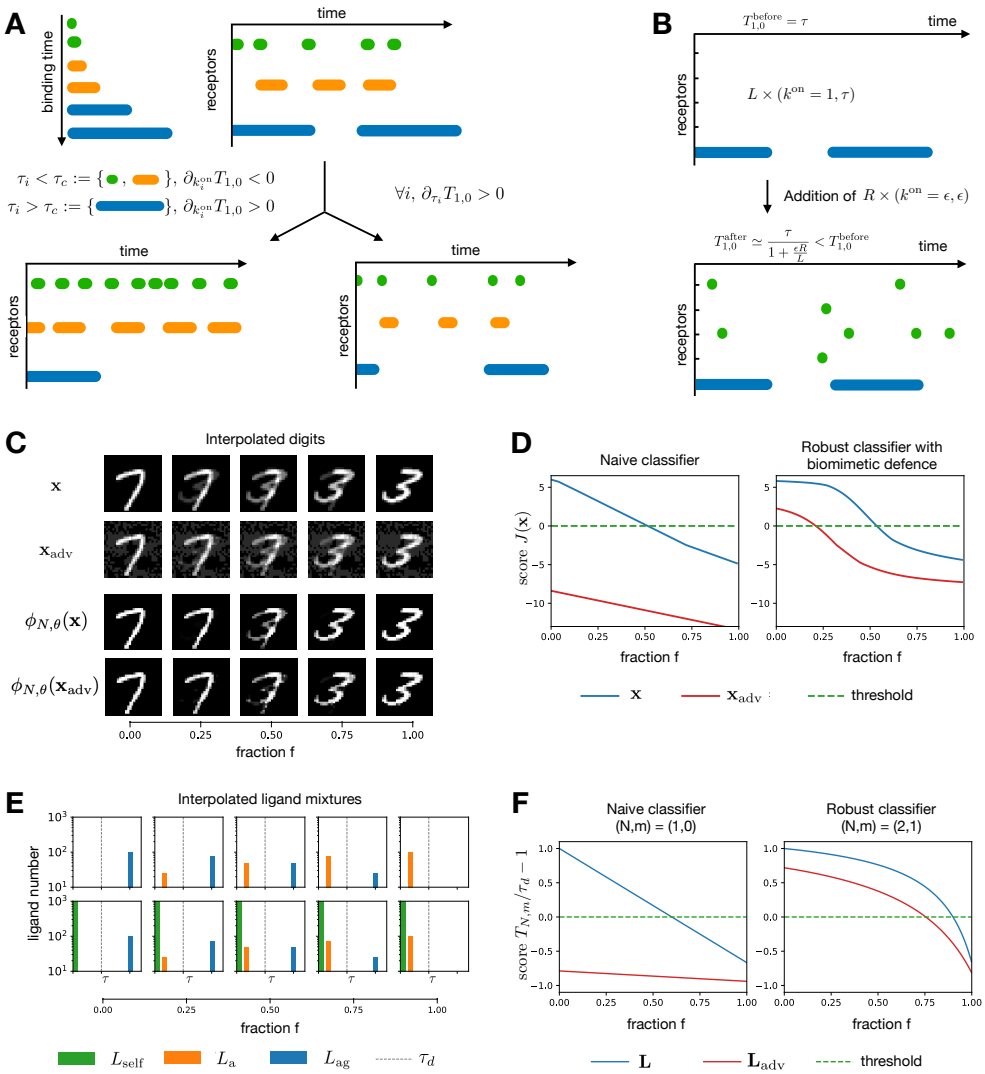

FIG. 3. **Schematics of FGSM applied to immune recognition.** (A) We compute how to lower the response for the receptor occupancy through a given period of time by changing $k_i^{\rm on}$ and τ_i. Bottom left: increasing $k_i^{\rm on}$ for ligands with $\tau_i < \tau_d$ and decreasing $k_i^{\rm on}$ 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 $T_{1,0}^{\rm before}$ to $T_{1,0}^{\rm after}$ 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, $(L_{\rm ag}, \tau_{\rm ag}) = (100, 6)$; $(L_{\rm a}, \tau_{\rm a}) = (100, 1)$; $(L_{\rm self}, \tau_{\rm self}) = (1000, 0.1)$ (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.

Goodfellow et al. [6] proposed the linearity hypothesis as an explanation for this adversarial effect: adding $\eta = \epsilon\ \mathrm{sgn}\,(\nabla_x J)$ 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 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 [49] 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.

$$y \in \{-1, +1\}, \quad x_1 \sim \begin{cases} +y, & \text{w.p. } p \\ -y, & \text{w.p. } 1-p \end{cases} \quad x_2, \ldots, x_{d+1} \in \mathcal{N}(\eta y, 1) \tag{7}$$

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 $\eta = O(1/\sqrt{d})$ [49].

We observe a very similar effect in the simplest adaptive proofreading model. The strong shift of the average $T_{1,0}$ in Eq. 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 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 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 [38, 39], has been proposed as a mechanism for ligand discrimination [36]. In the adaptive proofreading models we are studying here, kinetic proofreading allows the encoding of distinct τ dependencies in the activation/repression branches [33]. 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

$$x_i \leftarrow \phi_{N,\theta}(x_i) = \frac{x_i^N}{x_i^N + \theta^N}, \tag{8}$$

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 C bottom panels.

In Fig. 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 [50] (reminiscent of the iterative FGSM introduced in [51]). 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 [26, 50].

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

$$\tau \leftarrow \tau - \epsilon \frac{\partial T_{N,m}}{\partial \tau} \tag{9}$$

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 [19, 20] and associated vaccine failure [21]. Similar mechanisms might also be implicated in the process of tumour immunoediting [23].

From a given ligand mixture with few ligands above threshold and many ligands below thresholds, we follow the dynamics of Eq. 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. 9 (Fig. 4, see also Fig. S1 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. 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 for 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. 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. 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 for 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. 9. For $m \geq 2$, due to the new local flatness of this gradient, ligands at $\tau = 0$, the dynamical critical point of Eq. 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. 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" [52] (Fig. S2 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. 9 [50]. 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 [50] or many weakly bound ligands for the adaptive proofreading case, also similar to the meaningless changes in $x_2, \ldots x_d$ described above in Eq. 7 [49].

Defence against the FGSM perturbation is implemented through a higher degree n of the rectified polynomials in [50], while in adaptive proofreading, this is done through critical slowing down of the dynamics of Eq. 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 [50]. 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. 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 [50] 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 [15, 32], 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 [26], 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

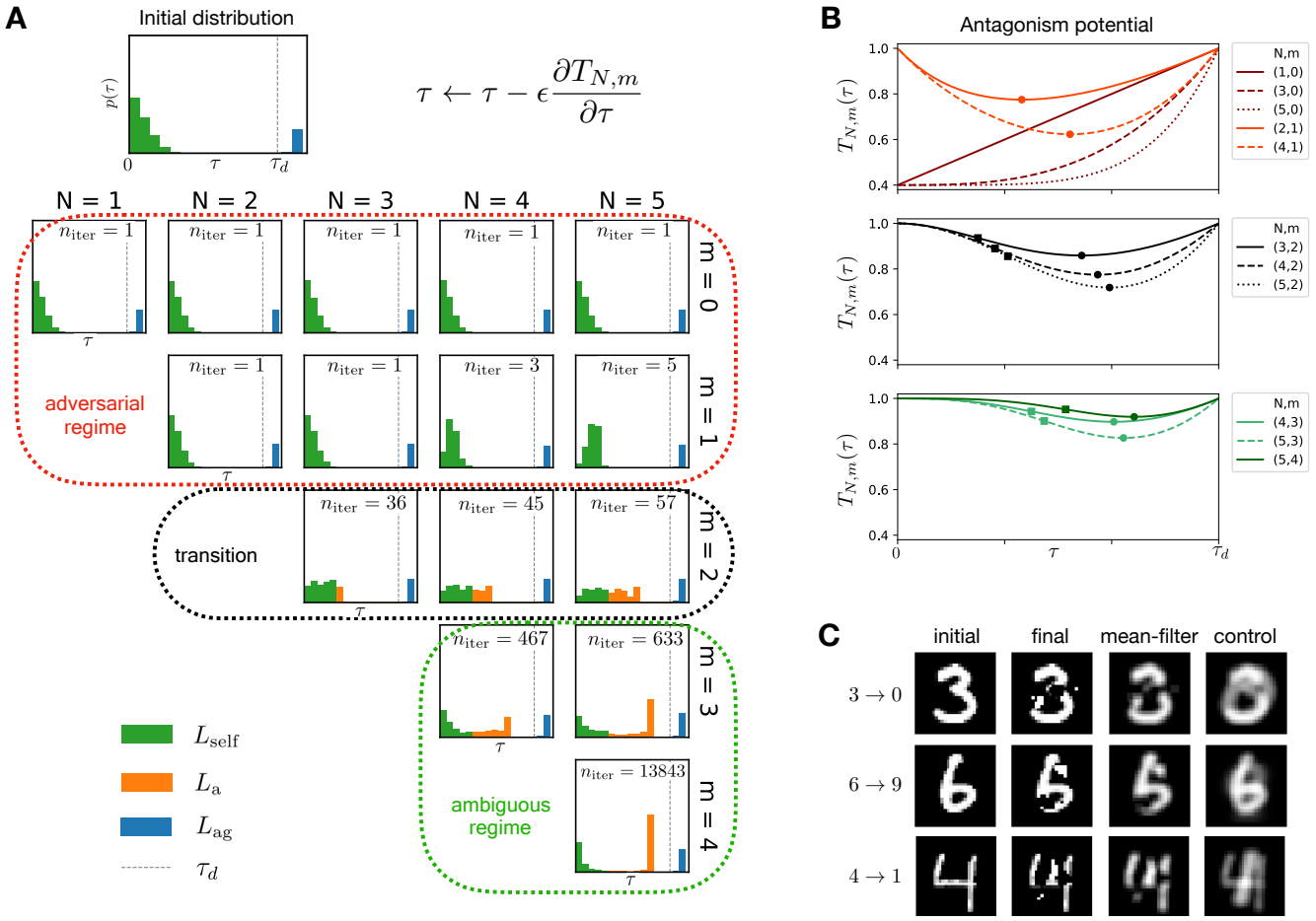

FIG. 4. **Characterization of the decision boundary following gradient descent dynamics.** (A) Ligand distribution at the decision boundary by applying iterative gradient descent (top right of the panel) to an initial distribution (top left). For various cases (N, m) we change the binding time of self 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.

biomimetic defences. The natural analogy is to implement attacks based on strong modification of few pixels [53].

For this problem, we choose to implement a two-tier biomimetic defence: we implement first the transformation defined in Eq. 8, that will remove influence of the FGSM types of perturbations by flattening the local landscape as in Fig. 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 [32, 33], 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 [54], 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 [53], 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 [50], 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. 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. S3 and details on the behaviour of scoring functions in Fig. S4. 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. S3). 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 [50]. 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 (or a Japanese ち). This observation is consistent with previous work attempting to interpolate in latent space between digits [55], where at the boundary a third digit corresponding to another category may appear. We also compare in Fig. 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 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 [49] similar to results of [50] and our observation for adaptive proofreading models above.

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 [33]. Strong antagonism naturally appeared in the simplest simulations, and required modification of objective functions very similar to adversarial training [6].

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 [56]. Other current explorations include new biomimetic learning algorithms, giving rise to prototype-like classification [57]. Adversarial defence strategies, including non-local computation and nonlinearities in the neural network are also currently under study [54]. 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 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. 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. 4 C with our digit classifier with biomimetic defence. Examples in image classification might include the meaningful adversarial transformations between samples found in [49] or the perturbed animal pictures fooling humans [58] 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 [15]. We interpret this property as a consequence of the flat landscape far from the decision threshold leading to a steeper gradient close to it [32, 34].

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 [2]. 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. 7) and weakly correlated features ($x_2, \ldots x_d$ in Eq. 7) [49], 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. 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 [34], as further identified here using the FGSM transformation. This latter result can be reformulated in terms of machine learning [49] 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. 8. Elaborating on the distinction between robust and weak features proposed in [49], nonlinear transformations should specifically target weak correlated features. Explorations of generalized nonlinear transformations in image feature space [26, 50] 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. 4, inspired by gradient descent in machine learning [50], 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 [42]. 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 [17]. 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 [19–21] or, more generally, could provide explanations on the diversity of altered peptide ligands [59]. 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 [60, 61]. 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 [62] (pushing towards lower m in our model) .

Our correspondence could also be useful for the theoretical modelling and understanding of cancer immunotherapy [22]. 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 [63]. 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 [23]. 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 [64].

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 [65], in size estimation of animals [66], or might be needed for proper detection of physical 3D objects [67] and road signs [68]. 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.

ACKNOWLEDGEMENTS

We thank Joelle Pineau and members of the François group for useful discussions. We thank three anonymous reviewers for comments and suggestions. P.F. is supported by a Simons Investigator in Mathematical Modelling of Living Systems award, an Integrated Quantitative Biology Initiative award and a Natural Sciences and Engineering Research Council award (Discovery Grant). T.J.R. receives funding from the Centre for Applied Mathematics in Bioscience and Medicine (graduate award), McGill Physics (Schulich award), and the Fonds de Recherche du Québec - Nature et Technologies (graduate award). E.B. acknowledges support from the Samsung Advanced Institute of Technology and the Fonds de Recherche du Québec – Nature et Technologies (graduate award).

[1] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep learning," Nature **521**, 436 (2015).

[2] 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.

[3] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," IEEE Signal processing magazine **29**, 82–97 (2012).

[4] 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.

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

[6] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples," arXiv preprint arXiv:1412.6572 (2014).

[7] 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.

[8] Samuel G Finlayson, Isaac S Kohane, and Andrew L Beam, "Adversarial attacks against medical deep learning systems," arXiv preprint arXiv:1804.05296 (2018).

[9] Naveed Akhtar and Ajmal Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," arXiv preprint arXiv:1801.00553 (2018).

[10] 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.

[11] Eric D Siggia and Massimo Vergassola, "Decisions on the fly in cellular sensory systems," Proceedings of the National Academy of Sciences **110**, E3704–E3712 (2013).

[12] Nicholas RJ Gascoigne, Tomasz Zal, and S Munir Alam, "T-cell receptor binding kinetics in T-cell development and activation," Expert Reviews in Molecular Medicine **3**, 1–17 (2001).

[13] Ofer Feinerman, Ronald N Germain, and Grégoire Altan-Bonnet, "Quantitative challenges in understanding ligand discrimination by $\alpha\ \beta$ T cells," Molecular Immunology **45**, 619 (2008).

[14] Paul François and Grégoire Altan-Bonnet, "The case for absolute ligand discrimination: modeling information processing and decision by immune T cells," Journal of Statistical Physics **162**, 1130–1152 (2016).

[15] Grégoire Altan-Bonnet and Ronald N Germain, "Modeling T cell antigen discrimination based on feedback control of digital ERK responses," PLOS Biology **3**, e356 (2005).

[16] Chikako Torigoe, John K Inman, and Henry Metzger, "An unusual mechanism for ligand antagonism," Science **281**, 568–572 (1998).

[17] 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," eLife **7**, e34958 (2018).

[18] Julia Tsitron, Addison D Ault, James R Broach, and Alexandre V Morozov, "Decoding complex chemical mixtures with a physical model of a sensor array." PLoS Comput Biol **7**, e1002224–e1002224 (2011).

[19] Paul Klenerman, Sarah Rowland-Jones, Steve McAdam, Jon Edwards, Susan Daenke, David Lalloo, Britta Köppe, William Rosenberg, Diana Boyd, Anne Edwards, *et al.*, "Cytotoxic T-cell activity antagonized by naturally occurring HIV-1 Gag variants," Nature **369**, 403 (1994).

[20] Ute-Christiane Meier, Paul Klenerman, Philip Griffin, William James, Britta Köppe, Brendan Larder, Andrew McMichael, and Rodney Phillips, "Cytotoxic T lymphocyte lysis inhibited by viable HIV mutants," Science **270**, 1360–1362 (1995).

[21] 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." The Journal of Immunology **158**, 807–815 (1997).

[22] Alexandra Snyder, Vladimir Makarov, Taha Merghoub, Jianda Yuan, Jesse M Zaretsky, Alexis Desrichard, Logan A Walsh, Michael A Postow, Phillip Wong, Teresa S Ho, Travis J Hollmann, Cameron Bruggeman, Kasthuri Kannan, Yanyun Li, Ceyhan Elipenahli, Cailian Liu, Christopher T Harbison, Lisu Wang, Antoni Ribas, Jedd D Wolchok, and Timothy A Chan, "Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma," N Engl J Med **371**, 2189–2199 (2014).

[23] Ton N Schumacher and Robert D Schreiber, "Neoantigens in cancer immunotherapy," Science **348**, 69–74 (2015).

[24] Kathrin Grosse, Praveen Manoharan, Nicolas Papernot, Michael Backes, and Patrick McDaniel, "On the (statis-

tical) detection of adversarial examples," arXiv preprint arXiv:1702.06280 (2017).

[25] 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.

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

[27] Jayajit Das, "Activation or Tolerance of Natural Killer Cells Is Modulated by Ligand Quality in a Nonmonotonic Manner," Biophysical Journal **99**, 2028–2037 (2010).

[28] 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," Environmental Health **14**, 13–a106 (2015).

[29] 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," Proceedings of the National Academy of Sciences , 201000966 (2010).

[30] Arup K Chakraborty and Arthur Weiss, "Insights into the initiation of TCR signaling," Nature immunology **15**, 798 (2014).

[31] Melissa Lever, Hong-Sheng Lim, Philipp Kruger, John Nguyen, Nicola Trendel, Enas Abu-Shah, Philip Kumar Maini, Philip Anton van der Merwe, and Omer Dushek, "Architecture of a minimal signaling pathway explains the T-cell response to a 1 million-fold variation in antigen affinity and dose," Proceedings of the National Academy of Sciences **113**, E6630–E6638 (2016).

[32] 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," Proceedings of the National Academy of Sciences , 201300752 (2013).

[33] Jean-Benoît Lalanne and Paul François, "Principles of adaptive sorting revealed by in silico evolution," Physical Review Letters **110**, 1–5 (2013).

[34] Paul François, Mathieu Hemery, Kyle A Johnson, and Laura N Saunders, "Phenotypic spandrel: absolute discrimination and ligand antagonism," Physical Biology **13**, 066011 (2016).

[35] Félix Proulx-Giraldeau, Thomas J Rademaker, and Paul François, "Untangling the hairball: Fitness-based asymptotic reduction of biological networks," Biophysical Journal **113**, 1893–1906 (2017).

[36] Timothy W McKeithan, "Kinetic proofreading in T-cell receptor signal transduction," Proceedings of the National Academy of Sciences **92**, 5042–5046 (1995).

[37] 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," Immunity **9**, 817–826 (1998).

[38] John J Hopfield, "Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity," Proceedings of the National Academy of Sciences **71**, 4135–4139 (1974).

[39] Jacques Ninio, "Kinetic amplification of enzyme discrimination," Biochimie **57**, 587–595 (1975).

[40] Jean-Benoît Lalanne and Paul François, "Chemodetection in fluctuating environments: Receptor coupling, buffering, and antagonism," Proceedings of the National Academy of Sciences **112**, 1898–1903 (2015).

[41] Thierry Mora, "Physical limit to concentration sensing amid spurious ligands," Physical review letters **115**, 038102 (2015).

[42] Martin Carballo-Pacheco, Jonathan Desponds, Tatyana Gavrilchenko, Andreas Mayer, Roshan Prizak, Gautam Reddy, Ilya Nemenman, and Thierry Mora, "Receptor crosstalk improves concentration sensing of multiple ligands." Physical Review E **99**, 022423 (2019).

[43] Ronald N Germain and Irena Stefanová, "The dynamics of T cell receptor signaling: complex orchestration and the key roles of tempo and cooperation," Annual review of immunology **17**, 467–522 (1999).

[44] 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," Immunity **11**, 289–298 (1999).

[45] Yann LeCun and Corinna Cortes, "The MNIST database of handwritten digits," (1998).

[46] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research **12**, 2825–2830 (2011).

[47] 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.

[48] 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.

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

[50] Dmitry Krotov and John J Hopfield, "Dense associative memory is robust to adversarial inputs," Neural computation , 1–17 (2018).

[51] Alexey Kurakin, Ian Goodfellow, and Samy Bengio, "Adversarial machine learning at scale," arXiv preprint arXiv:1611.01236 (2016).

[52] Thomas Tanay and Lewis Griffin, "A boundary tilting persepective on the phenomenon of adversarial examples," arXiv preprint arXiv:1608.07690 (2016).

[53] Jiawei Su, Danilo Vasconcellos Vargas, and Sakurai Kouichi, "One pixel attack for fooling deep neural networks," arXiv preprint arXiv:1710.08864 (2017).

[54] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan Yuille, and Kaiming He, "Feature denoising for improving adversarial robustness," arXiv preprint arXiv:1812.03411 (2018).

[55] David Berthelot, Colin Raffel, Aurko Roy, and Ian Goodfellow, "Understanding and improving interpolation in autoencoders via an adversarial regularizer," arXiv preprint arXiv:1807.07543 (2018).

[56] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, and Pascal Frossard, "Robustness via curvature regularization, and vice versa," arXiv preprint arXiv:1811.09716 (2018).

[57] Dmitry Krotov and John J Hopfield, "Unsupervised learning by competing hidden units," Proceedings of the National Academy of Sciences , 201820458 (2019).

[58] 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.

[59] Emil R Unanue, "Altered peptide ligands make their entry," The Journal of Immunology **186**, 7–8 (2011).

[60] Thomas C Butler, Mehran Kardar, and Arup K Chakraborty, "Quorum sensing allows T cells to discriminate between self and nonself," Proceedings of the National Academy of Sciences **110**, 11833–11838 (2013).

[61] 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," Cell reports **11**, 1208–1219 (2015).

[62] 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," Immunity **38**, 263–274 (2013).

[63] Marta Łuksza, Nadeem Riaz, Vladimir Makarov, Vinod P Balachandran, Matthew D Hellmann, Alexander Solovyov, Naiyer A Rizvi, Taha Merghoub, Arnold J Levine, Timothy A Chan, Jedd D Wolchok, and Benjamin D Greenbaum, "A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy," Nature **551**, 517–520 (2017).

[64] Ugur Sahin and Özlem Türeci, "Personalized vaccines for cancer immunotherapy," Science **359**, 1355–1360 (2018).

[65] Jinyuan Yan, Maxime Deforet, Kerry E Boyle, Rayees Rahman, Raymond Liang, Chinweike Okegbe, Lars EP Dietrich, Weigang Qiu, and Joao B Xavier, "Bow-tie signaling in c-di-GMP: Machine learning in a simple biochemical network," PLOS Computational Biology **13**, e1005677 (2017).

[66] Andres Laan and Gonzalo de Polavieja, "Sensory cheating: adversarial body patterns can fool a convolutional visual system during signaling," bioRxiv , 326652 (2018).

[67] 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.

[68] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song, "Robust physical-world attacks on deep learning models," in *Proceedings of Conference on Computer Vision and Pattern Recognition* (2018).

APPENDIX

Appendix S1: Mathematical details of the adaptive proofreading models

Appendix S1 contains more details on the derivation of adaptive proofreading models (section *Biochemical kinetics*), referred to in section **Adaptive proofreading for cellular decision-making** in the main text. We also give the parameters and equations that are used to draw Fig. 2 B in the main text (section *Parameters for Fig. 2 B*).

Biochemical kinetics

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

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

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 $C_1 \to C_2$ (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

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

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

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

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. 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

$$\dot{C}_N = K C_{N-1} - \tau_{\text{ag}}^{-1} C_N \tag{5}$$

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

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

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

This expression for two types of ligands with same k_{on} can be clearly generalized to any types of ligands, giving Eq. 3 in the main text.

Appendix S2: Materials and methods

In Appendix S2, we give the parameters and equations that are used to draw Fig. 2 B in the main text and we give the hyperparameters used for training the neural networks classifying 3s and 7s. We referred to the latter in section **Neural networks for artificial decision-making** in the main text.

Parameters for Fig. 2 B

The curves on Fig. 2 B, left panel, come from the model given by

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

with parameter values $C_* = \beta/\alpha = 3000$, $\tau_d = 4s$ and τ as in the legend. The curves on the middle panel of Fig. 2 B come from

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

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.

Appendix S3: Ligand distribution at the decision boundary

In Appendix S3 we describe in detail the methods used in the gradient dynamics of changing a ligand distribution to the decision boundary (section *Methods*), we provide additional results when adding spatial correlation to the ligand distribution (section *MTL pictures*), and we calculate the leading order in small binding time τ_ϵ of the gradient $\frac{dT_{N,m}}{d\tau_\epsilon}$ (section *Behavior for small binding times*). We refer to Appendix S3 in the main text in section **Gradient dynamics identify two different regimes** and **Qualitative change in dynamics is due to a critical point for the gradient**, and in Fig. 3A.

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 [50] and adapted our approach from the iterative FGSM [51]. 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 $\tau_b < \tau_d$

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

where L_b is the number of ligands in the bth bin. We subtract this value multiplied by a small number ϵ from the exact binding times, as in Eq. 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 distribution. This results in what we term "MTL-pictures" (Fig. S1). 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. S1, 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. 4 A and Fig. S1 are available at https://github.com/tjrademaker/advxs-antagonism-figs/.

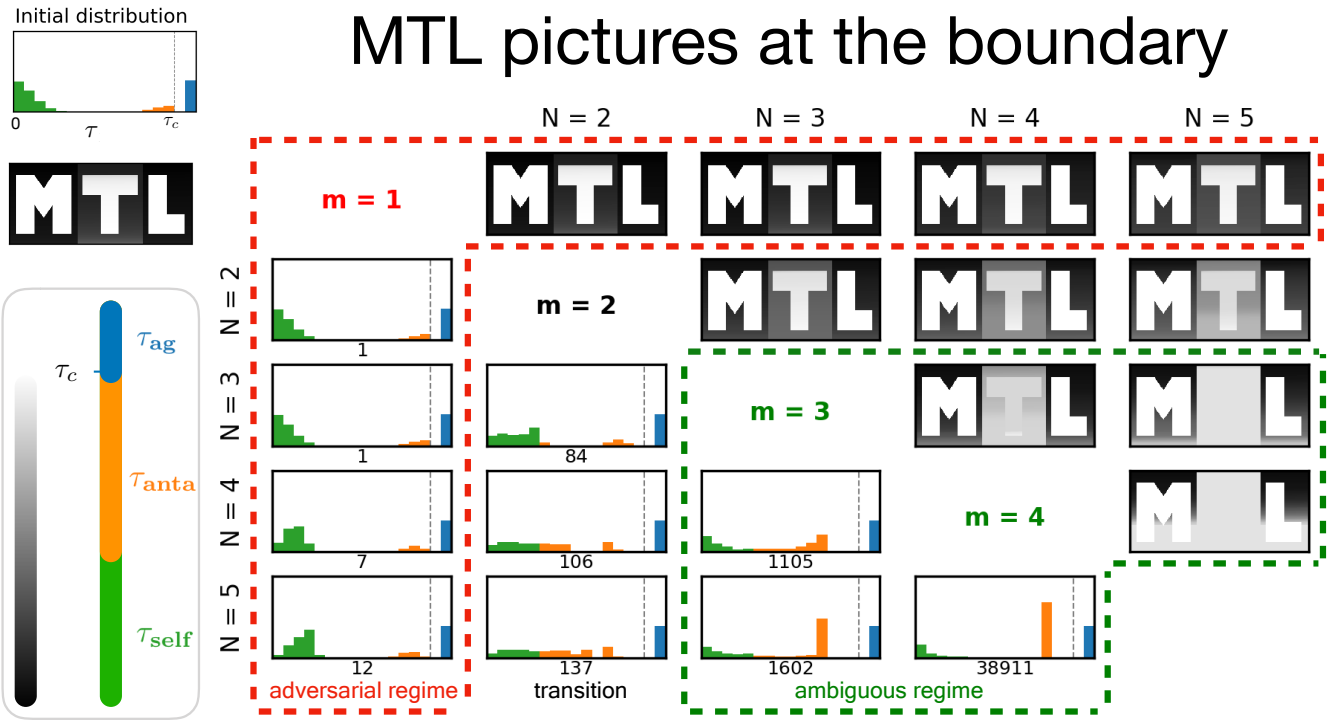

FIG. 1. **Figure S1.** MTL pictures. Explanation is found in the text

Behavior for small binding times

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

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

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

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

Appendix S4: 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 [52]. We will first illustrate this effect on the discrimination of the original MNIST 3 vs 7 problem MNIST from [6]) (section *Digit classification*), after which we will interpret boundary tilting via proofreading in ligand discrimination (*Boundary tilting and categorizing perturbations*), and finally, we will derive how the addition of a subthreshold ligand at the decision boundary changes the output (section *Gradient in the L_2 direction*). We refer to these results in the main text at the end of section **Qualitative change in dynamics is due to a critical point for the gradient**.

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. S2 A. Tanay and Griffin [52] pointed out that the adversarial perturbation generated with the Fast Gradient Sign Method (FGSM) proposed in [6] can also be found via $D = \text{sign}\,(\bar{3} - \bar{7})$, Fig. S2 A (v). Note the similarity to the adversarial perturbation from the FGSM $\text{sgn}(w) = \text{sgn}\,(\nabla_x J)$ (Fig. S2 A (vi)). To reveal the linearity of binary

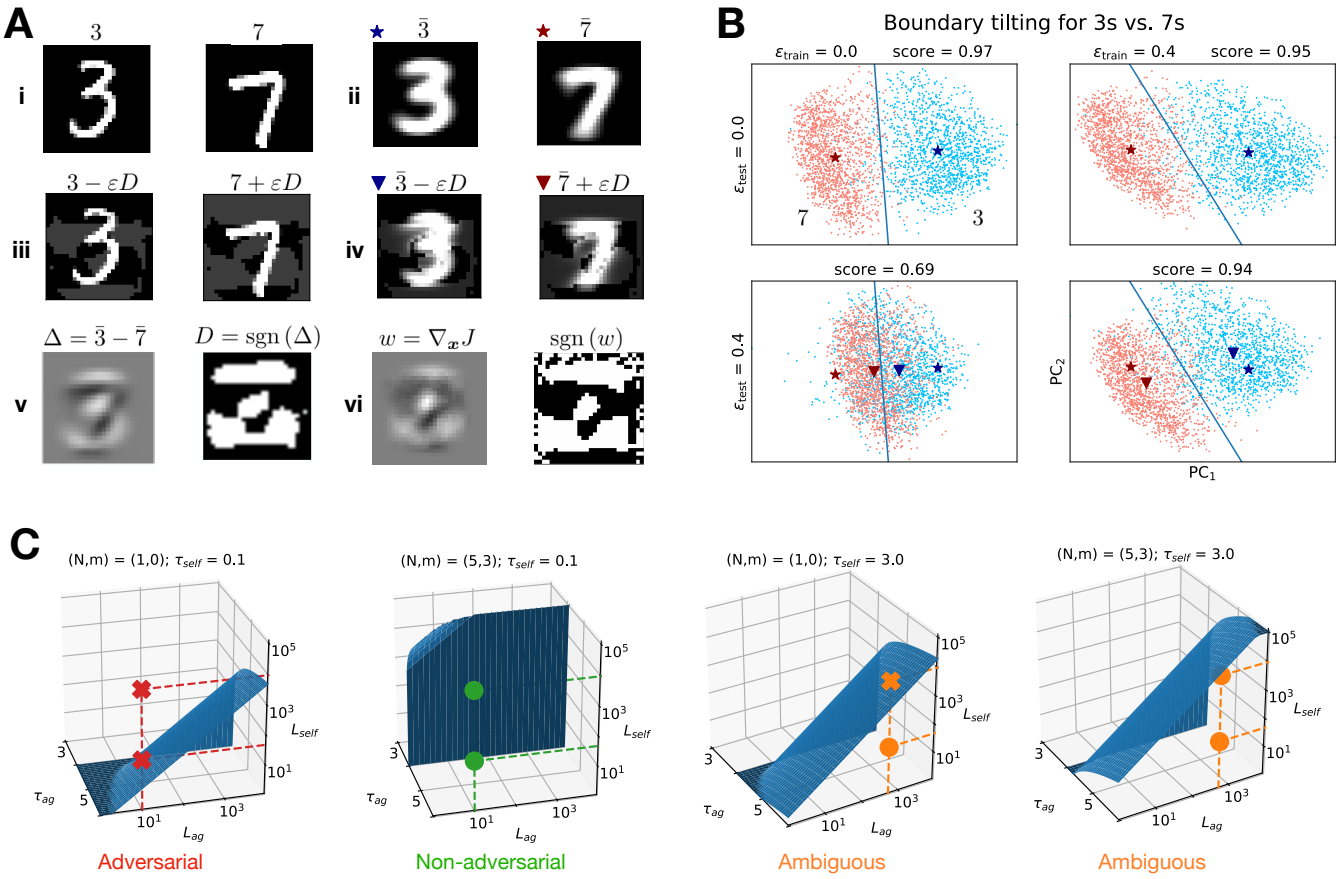

FIG. 2. **Figure S2.** Boundary tilting in one-dimensional digit classification. (A) (i) Typical 3 and 7 from MNIST. (ii) Average 3, 7 of the traditional test set, (iii, iv) with adversarial perturbation, found by (v) subtracting the sign of $\bar{3}$ from $\bar{7}$, which corresponds to (vi), the perturbation 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.

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. S2 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 [15]). The red and blue star in Fig. S2 A denote the average digit $\bar{3}, \bar{7}$.

Next, we transformed the test set as $3 \to 3' = 3 - \epsilon_{\text{test}} D$, $7 \to 7' = 7 + \epsilon_{\text{test}} D$, where $\epsilon_{\text{test}} = 0.4$ is the strength of the adversarial perturbation (Fig. S2 A (iii)). $\bar{3}'$ and $\bar{7}'$ moved closer in Fig. S2 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 accuracy. This is a illustration example of the more general phenomenon studied in [52].

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. 2 in the main text gives

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

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 the main text 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 [33]. Biochemically, this is done with kinetic proofreading [15, 32, 36], 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 [34]. An extreme case has been proposed for immune recognition where the strongest antagonists are found just below the threshold of activation [15].

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

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

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. S2 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. S2 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, [14]), which is again of the order of the response $T_{1,0} = \tau_{\text{ag}} \simeq \tau_d$ 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. S2 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 $\tau_{\text{self}} \simeq 0$ [14]. 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. S2 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

TABLE I. **Table S1**: 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 ($\mathcal{O}(\epsilon)$)

case of Fig. S2 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 [15]. Our categorization of perturbations is presented in Table SI. 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 [34] 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

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

The main result of [34] is the linear response of $T_{N,m}(L, \tau_d)$ to the addition of $\{L_a, \tau_d - \epsilon\}$ subthreshold ligands.

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

where we used the definition

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

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

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

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

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

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 the main text. For higher m, the linear approximation holds only for ligands at τ_a close to threshold.

Appendix S5: Few-pixel attack

In Appendix S5, we describe in detail the procedure for the few-pixel attack. We used this to come to our conclusion in section **Biomimetic defenses against few-pixel attacks** and Fig. 4 C in the main text.

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. S3. We first computed what we term pixelmaps. Pixelmaps contain the change of score when making a pixel white or black. In Fig. S3, 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. S3 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 [54], which is a 3x3 convolutional block that computes mean pixel values as

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

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. 8 ($N = 3; \theta = 0.5$) on the average of all digits in the target class, then mean-filtered (Fig. S3 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. S4 for the digits in Fig. S3. The behavior of the score is remarkably similar to the interpolation between ligand mixtures (Fig. 3F, bottom panel in the main text). A nonlinear filtering method proposed in [54] 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 [53] 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 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.

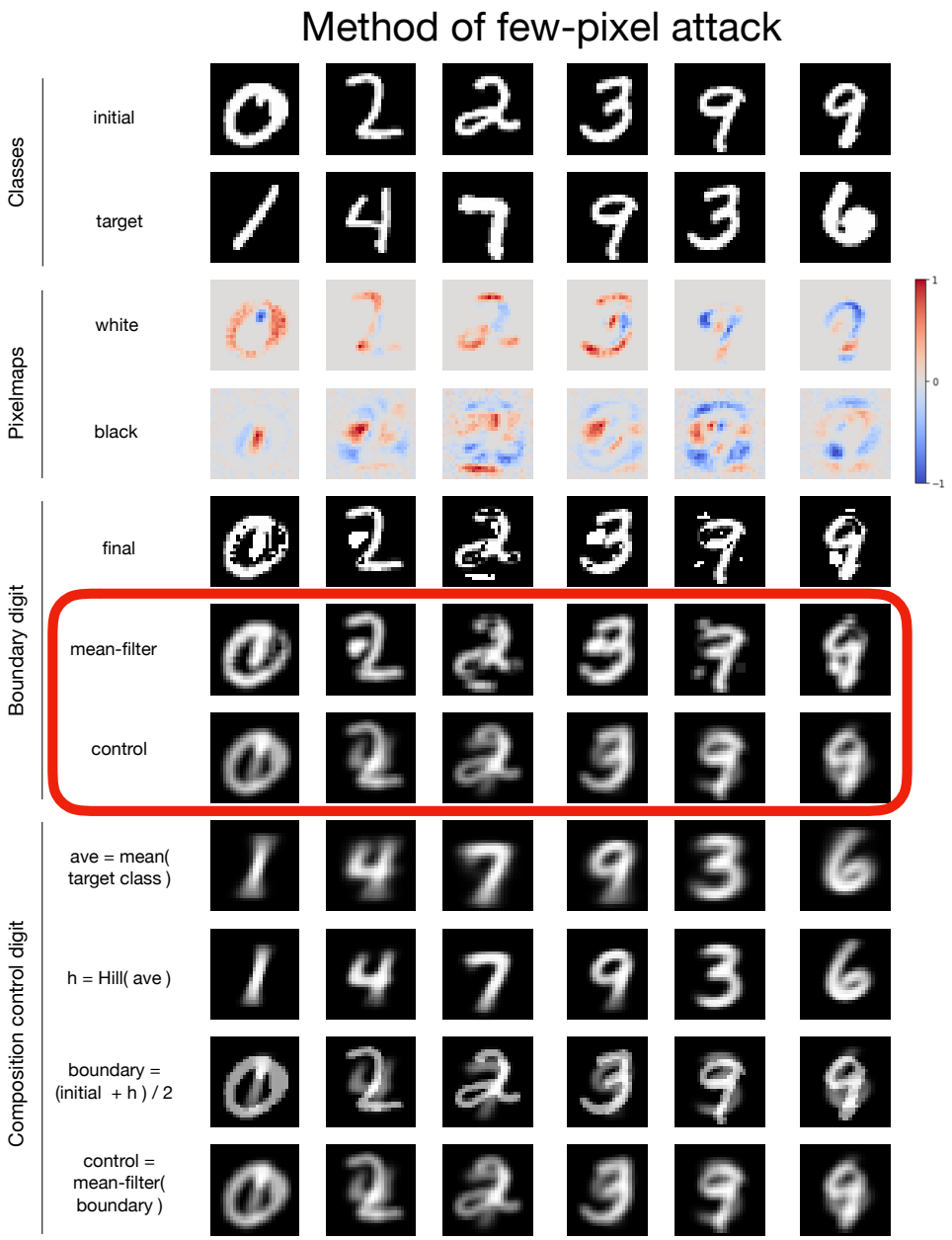

FIG. 3. **Figure S3.** Method of few-pixel attack. Each column show how a few-pixel attack causes misclassification of an initial digit to a target class. The important result are the pre-filtered boundary digits and the control in the red rectangle. Pixelmaps determine which pixels increase (red) or 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.

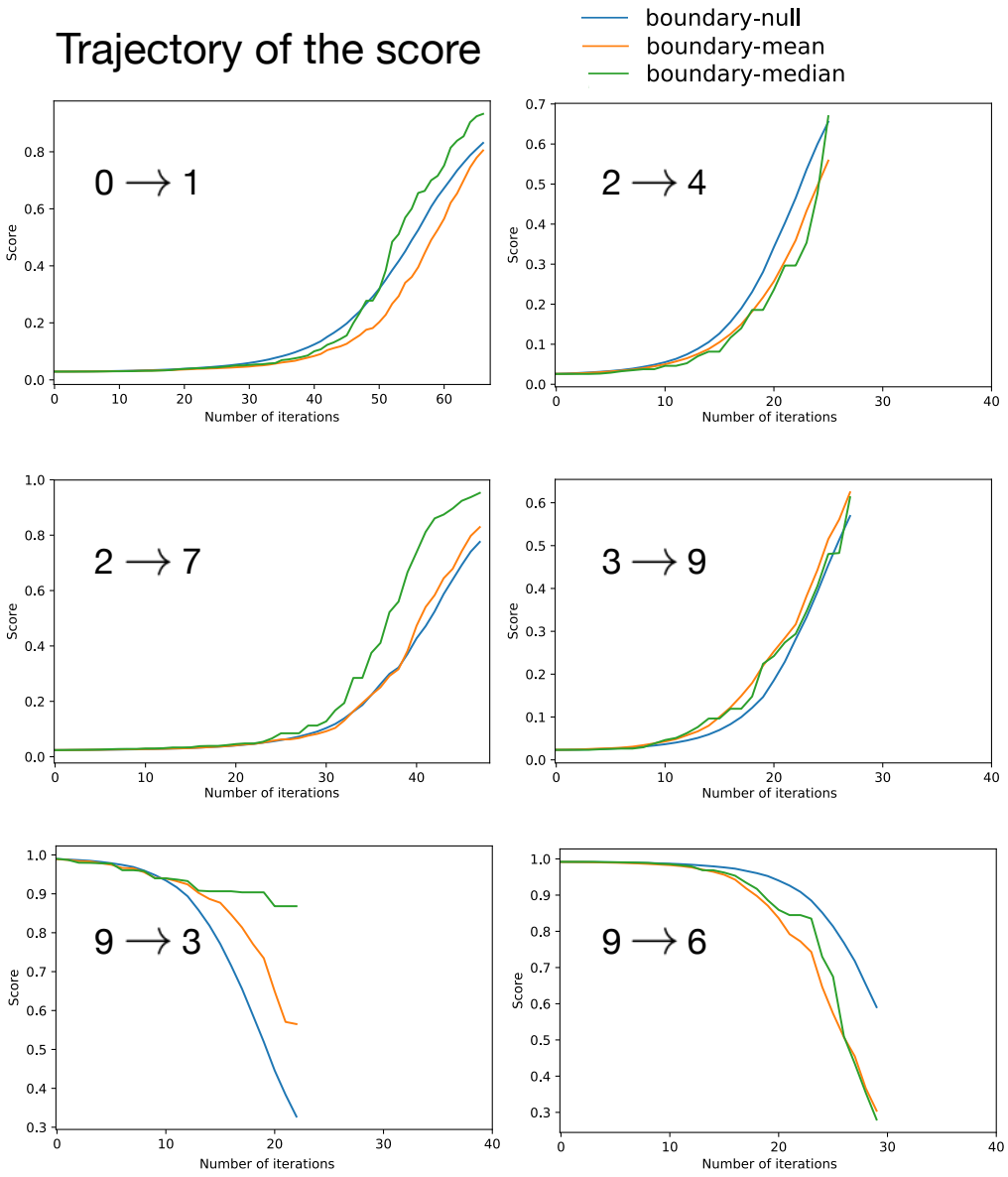

FIG. 4. **Figure S4.** Trajectory of the scoring functions of the attacks in Fig. S3. The blue, orange and green line correspond to various digits (actual digit, mean-filtered digit, median-filtered digit) for which we check the score, and terminate when reaching the boundary. The trajectory of the score 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 D in the main text.