Articles in PresS. J Neurophysiol (September 12, 2012). doi:10.1152/jn.00208.2012

Learning selective top-down control enhances performance in a visual categorization task

Top-down control in learning a categorization task

Mario Pannunzi[1], Guido Gigante[2,5], Maurizio Mattia[2], Gustavo Deco[1,3], Stefano Fusi[4] and Paolo Del Giudice[2,6]

[1]Universitat Pompeu Fabra, Barcelona (Spain),

[2]Istituto Superiore di Sanitá, Rome (IT),

[3]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona (Spain),

[4]Center for theoretical neuroscience, Columbia university, NY (USA),

[5]Mperience s.r.l., Rome (IT),

[6]Istituto Nazionale di Fisica Nucleare, Rome (IT),

Corresponding author: Mario Pannunzi, DTIC-Universitat Pompeu Fabra (55.120) C/ Roc Boronat, 138, 08018 Barcelona, Spain; mario.pannunzi@gmail.com

We thank Jochen Braun for a critical reading of the manuscript

Abstract

We model the putative neuronal and synaptic mechanisms involved in learning a visual categorization task, taking inspiration from single-cell recordings in inferior temporal cortex (ITC). Our working hypothesis is that learning the categorization task involves both bottom-up, ITC to pre-frontal cortex (PFC), and top-down (PFC to ITC) synaptic plasticity, and that the latter enhances the selectivity of the ITC neurons encoding the task-relevant features of the stimuli, thereby improving the signal-to-noise ratio. We test this hypothesis by modeling both areas and their connections with spiking neurons and plastic synapses, ITC acting as a feature-selective layer, and PFC as a category coding layer. This minimal model gives interesting clues as to properties and function of the selective feedback signal from PFC to ITC which help solving a categorization task. In particular, we show that, when the stimuli are very noisy because of a large number of non-relevant features, the feedback structure helps getting better categorization performance and decreasing the reaction time. It also affects the speed and stability of the learning process and sharpens tuning curves of ITC neurons. Furthermore, the model predicts a modulation of neural activities during error trials, by which the differential selectivity of ITC neurons to task-relevant and task-irrelevant features diminishes or is even reversed, and modulations in the time course of neural activities which appear when, after learning, corrupted versions of the stimuli are input to the network.

1 Introduction

A large body of knowledge has accumulated about the brain areas involved in categorization across multiple sensory modalities auditory (Vallabha et al. (2007)), somatosensory (Romo and Salinas (2001)), olfactory (Howard et al. (2009)) systems, categorization of visual stimuli being the most studied (Knoblich et al. (2002)).

Several open questions remain as to the specific roles played by the areas participating in the association between sensory stimuli and categories (Freedman and Assad (2011); Swaminathan and Freedman (2012)), and the learning mechanisms involved. The neural correlates of the acquired association carry multiple traces of category-related modulations (Freedman et al. (2001); Sigala and Logothetis (2002); Freedman et al. (2003); De Baene et al. (2008); Meyers et al. (2008)); however it is not always clear (especially for feature-encoding areas) whether such modulations are epiphenomenal reflections of category-specific neural signals generated elsewhere, or they have an important computational role in the categorization process.

Among the brain areas involved in categorization, prefrontal cortex (PFC) plays a fundamental role, notably for learning novel associations and encoding abstract rules (Seger and Miller (2010)). Neurons in PFC show sharp response properties across boundaries between categories largely independent of stimulus similarity (Freedman et al. (2003)). However, category-related actions/decisions involve multiple areas: premotor cortex for action planning (Boettiger and D'Esposito (2005); Muhammad et al. (2006)); parietal cortex to implement visuospatial processing linking perceptual information with potential responses; basal ganglia to gate selectively cortical areas for choice of action (Humphries et al. (2006); Seger (2008)); hippocampus and medial temporal lobe to encode and learn items to be categorized (Myers et al. (2003); Shohamy and Wagner (2008)); the dopaminergic system and the associated plasticity of striatal and corticostriatal synapses, to support reward-modulated learning; inferotemporal cortex (ITC), where a modular, feature-based representation has been observed (Tsunoda et al. (2001); Yamane et al. (2006)), and where neurons with category-related tuning properties have been reported (Vogels (1999); Freedman et al. (2003)).

A general plausible computational principle underlying the categorization process is that it must rely on the selection of relevant features in a specific behavioural context, and neurophysiological studies in ITC have indeed shown that the activity of neurons encoding sensory features relevant for the task are maximally modulated (see Sigala and Logothetis (2002); De Baene et al. (2008).

Starting from this evidence, the basic assumption of the present modeling work is that a mutual interaction between a feature-encoding (e.g. ITC) and a category-encoding (e.g. PFC) brain areas is the fundamental neural substrate of categorization. We implement a learning scenario for the acquisition of an association between stimuli and categories, and explore the consequences of top-down selective modulation of sensory representations.

We take as a relevant testing ground of our model the results of Sigala and Logothetis (2002); De Baene et al. (2008), extending the scope of our previous work Szabo et al. (2006) to the domain of dynamic, online learning, and from this we address more general questions about the computational role of learned selective feedback in a categorization task. As a significant progress of Szabo et al. (2006) in the present work we face the full complexity of the ongoing dynamic coupling between spiking activity induced by stimuli and the spike-driven, local synaptic dynamics. Beyond the appeal in terms of biological plausibility this entails inter alia coping with the finite-size effects which are important in determining learning histories (see Del Giudice et al. (2003)), due to the distribution of firing rates and the consequent distribution of rates of synaptic changes.

Based on simulations of a multi-modular architecture composed of spiking (integrate-and-fire) neurons and plastic, spike-driven synapses, we will indeed show that successful learning histories emerge naturally through a combination of Hebbian plasticity for correct trials and partially anti-Hebbian plasticity for error trials. The learnt top-down synaptic structure produces better performances and faster response. Besides reproducing, in correct trials, modulation of neural activity in ITC modules, qualitatively similar to the one observed in Sigala and Logothetis (2002), the model generates specific neurophysiological and behavioral predictions, including a different and specific tuning during error trials: the selectivity of the task-relevant feature neurons is diminished or even reversed.

2 Methods

2.1 Neuronal and synaptic dynamics

Our neural model is the single compartment linear integrate-and-fire (IF) neuron (Fusi and Mattia (1999)). The subthreshold dynamics of the membrane potential of neuron i is: $\dot{V}_i(t) = -\beta + I_i(t)$, (assuming units such that the membrane capacitance $C = 1$) with a reflecting barrier condition such that if $V_i(t)$ is driven below 0, it stays 0. β is a constant leakage term. When the membrane potential reaches the threshold $\Theta = 1$ the neuron emits a spike and the membrane

potential V_i is set and kept to a reset potential V_r for a refractory period τ_{arp}. $I_i(t)$ is the total synaptic current afferent to neuron i and it is the sum of the external excitatory current I_{ext}, the recurrent excitatory and inhibitory currents:$I_i(t) = \sum_j J_{ij} \sum_k \delta(t - t_j^k - \delta_j) + I_{ext}$. J_{ij} represents the amplitude of the instantaneous change of the postsynaptic potential (positive for excitatory synapses, negative for inhibitory synapses). The sums are over all presynaptic neurons j, and for each j over all the emitted spikes at times t_j^k, reaching the target neuron i with delay δ_j. Delays are randomly sampled from a truncated exponential distribution with a minimum and a maximum value, respectively δ_j^m and δ_j^M. Parameters are listed in Table 1.

Plastic synapses in the model are bistable and stochastic as motivated and described in Fusi et al. (2000). The synaptic efficacy J takes one of two values — J_- (depressed) and J_+ (potentiated). Learning evolves as a sequence of random transitions between J_- and J_+, triggered by the arrival of pre-synaptic spikes; the direction of the transition (potentiation or depression) is determined by the instantaneous value of the post-synaptic potential (Fusi et al. (2000)), as detailed in the following.

Each synapse has an internal dimensionless variable ('synaptic potential', X_J), ranging in the interval [0; 1]. This range is split in two by a threshold Θ_J; when X_J is above this threshold, the synapse is in the potentiated state and X_J constantly moves toward 1, with a constant drift α_X; below Θ_J, the synapse is depressed and X_J moves toward 0, with the same drift α_X. Thanks to the drift, in the absence of pre-synaptic spikes, the synaptic state (potentiated or depressed) never changes (thus we have long-term potentiation, LTP, or long-term depression, LTD). Transitions can happen only upon the arrival (at time t_k) of a presynaptic spike (with index k), which causes a sudden jump in X_J. If the postsynaptic potential is found above a threshold Θ_V, the jump will be positive and of size dX_+, otherwise it will be negative and of size dX_-. A positive jump can take X_J above Θ_J, making thus the synapse switch to the potentiated state; conversely, a negative jump can make the synapse switch to the depressed state. In formulae:

$$\frac{dX(t)}{dt} = \alpha_X \,\mathrm{sign}(X(t) - \Theta_J) \; + \; \sum_k [\Theta(V_{post}(t_k) - \Theta_V)\, dX_+ - \Theta(\Theta_V - V_{post}(t_k))\, dX_-]$$

LTP results from high pre-synaptic actitivy (high rate of triggering pre-synaptic spikes) and high post-synaptic activity (which on average implies high values of post-synaptic membrane potential). LTD occurs for highly active pre-synaptic neuron and poorly active post-synaptic one.

Parameter values for the synaptic dynamics are listed in Table 3.

2.2 Network architecture and stimuli

We set up a network with two layers, analogous to the one described in Szabo et al. (2006) (see Fig. 1), meant to describe respectively a cortical area coding for the visual features defining each stimulus, and a higher cortical area coding for the category assignment to stimuli according to a rule to be learnt. Each layer includes 6000 neurons, divided into several selective populations of excitatory neurons (see below), one non-selective excitatory 'background' population, and one inhibitory population.

The first layer ('ITC') comprises $N_F + 1$ feature-selective populations of 240 neurons each (denoted with the letter D and O in Fig. 1, see section 2.3 for the role played by the different populations). In the absence of stimuli, all the neurons in each population receive a background (excitatory) Poissonian synaptic input of base-rate λ_0. Each stimulus to ITC is defined by the activation of one of the two values of each feature (*e.g.*, small *vs* large distance between eyes in a face, see Fig. 3); upon stimulation two disjoint subsets of 120 neurons for each ITC population (D1-D2 and O1-O2: $N_F = 1$ in Fig. 1) will receive a differential current (except for section 3.4 in which population D will be divided into four subsets of 120 neurons each, to code for four values); the active (inactive) value corresponds to an input Poisson spike train with rate $\lambda + \Delta\lambda$ ($\lambda - \Delta\lambda$), with λ a reference value; $\lambda \pm \Delta\lambda > \lambda_0$. The value of $\Delta\lambda$ can be varied, (see Section 3.2). Corrupted stimuli are implemented (see Section 3.3) by reducing the size of selective populations to a fraction $x<1$ of the original size, the remaining $(1 - x)\,240$ neurons being given a stimulus-independent current of rate λ_0. N_F too will be varied in the following, to study the effects of a more or less complex feature space; $N_F = 16$ unless otherwise specified.

Stimulation parameters are listed in Table 1.

The second layer ('PFC') has a winner-take-all structure, similar to the one described in Wang (2002), in which two cooperating-competing populations (480 neurons each) encode the two categories, C1 and C2; in the regime of interest, upon stimulation of the ITC populations, the network dynamics always leads to a stable state where C1 is firing high and C2 is almost silent or viceversa, signalling a decision of the network as to which category the presented stimulus is assigned to.

All neurons have a probability $c = 25\%$ of being synaptically connected with any other neuron in the network, with the exception that excitatory-to-inhibitory and inhibitory-to-excitatory connections are restricted to each of the two layers ('local' inhibition).

C1 and C2, as well as the subsets in ITC, are self-excited and mutually excited; besides, they

are reciprocally connected with the excitatory background and inhibitory populations in the corresponding layers (see Table 1 and Fig.1).

The bi-directional synaptic connections between the ITC and PFC layers, are the only plastic synapses (see Section 2.1) and are thus shaped by learning (see Section 2.3).

Values chosen for the fixed synaptic efficacies are in Table 2.

(Figure 1 about here)

2.3 Task and learning

We define the task after Sigala and Logothetis (2002) and De Baene et al. (2008). In the experiments, monkeys were shown schematic visual stimuli, defined by a fixed number of features, and grouped in two categories, to which the monkey is trained to assign stimuli by trials and errors. Only a subset of features were relevant for the categorization, and neurons in ITC selective for those "diagnostic" features turned out to be maximally modulated depending on the feature values.

We chose only one of the $N_F + 1$ features to be relevant ('diagnostic') for the categorization: expect for section 3.4, the task the network has to learn is to associate stimuli with one of the two values (e.g., distant eyes D1) of this one feature to category C1, and stimuli taking the other value (e.g., close eyes D2) to C2, regardless of the values taken by the remaining N_F non-diagnostic features.

At the beginning of training, we generate a random pattern of activation $\lambda \pm \Delta\lambda$ for each of the $N_F + 1$ ITC populations. If the answer of the network, as read from the pattern of activity of C1 and C2 in the PFC layer, is correct (the correct classification being determined by the value of the diagnostic feature), we go on generating a new stimuls encoded by a new random choice of all $N_F + 1$ features for the subsequent trial. If not, in the subsequent trial the network is presented with a new random stimulus belonging, however, to the same class as the preceding (wrongly classified) one (N_F random values for the non-diagnostic features). This is consistent with the training strategy adopted in Sigala and Logothetis (2002) [1].

Each stimulus lasts for 2 sec. At 1.5 sec from stimulus onset, the network response is 'read' and a signal Reward/No Reward is determined, which activates the appropriate synaptic plasticity mode (see below), until the end of the stimulus.

Learning is semi-supervised and partially anti-Hebbian: the Hebbian synaptic dynamics (see Section 2.1) is activated only upon the correct completion of a trial, that is just after the network

[1] N.Sigala, private communication

has correctly classified a stimulus ('Reward' condition). When the network generates a wrong classification ('No Reward' condition), synapses that would undergo a positive jump dX_+ (up regulated) are subject to a negative jump dX_- (down regulated), while the ones that would be down-regulated are left unchanged. This way, synapses that would be potentiated if the classification provided was correct, will be depressed, while the other synapses are left unchanged.

Fig. 2 shows a cartoon of the average expected effect of learning on the synaptic structure of the network, both upon correct and wrong classification. This suggests that the synaptic connections between PFC and ITC, after learning, can be effectively described with 6 parameters (see Fig. 1): J_P, the average synaptic efficacy between D1 and C1, and D2 and C2 (correct association); J_D, the average synaptic efficacy between D2 and C1, and D1 and C2 (wrong association); J_n, the average synaptic efficacy between category populations and non-diagnostic features (C1/C2$\leftrightarrow$ O1/O2), in both directions.

As already noticed in Roelfsema et al. (2010), synaptic dynamics is required to change after an erroneous answer, in order to prevent the impairment of what was learned contingently to correct answers. Such strong assumptions are consistent with the experimentally observed reward-related modulation of synaptic plasticity by the dopaminergic neurons (Schultz (1998); Schultz and Dickinson (2000)).

(Figure 2 about here)

2.4 A rationale for a features-based representation in the ITC layer

The adopted model is certainly a gross oversimplification, both as to the type of neural representation in ITC and as to the areas involved. However, regarding the former, we emphasize that evidence reported in the literature provides at least a rationale for adopting a neural representation in ITC which is based on feature-selective populations, and for assuming that the collection of such segregated representations is what is forwarded in the first place to PFC for further processing. To substantiate this statement, in Fig. 3 we propose a close analogy between our ITC model and the results reported in Tsunoda et al. (2001). These authors explored extensively the representation of visual objects in ITC, and how such representations are altered when simplified versions of the same objects are presented (a simplified object being an object deprived of some features). Fig. 3, panel A, shows the observed patchiness of the neural representations of an object, and what those representations reduce to when some features are removed. It is seen that there are patches uniquely associated to some features, others are overlapping to various degrees. In the words of Tsunoda et al, "an object is represented by a combination of cortical

columns, each of which represents a visual feature (feature column)" (other aspects of the results from Tsunoda et al. will be commented on in the Discussion, in the light of the present model). Fig. 3, panel B illustrates the assumed patchy representation in ITC that, by way of analogy, we associate with the representation of the Brunswick faces used in the Sigala and Logothetis (2002) work. The scheme is then mapped onto the model architecture described above in panel C, where we suggest that the simplified representation we adopted could be imagined to be obtained from a selection of recorded neurons based on their feature selectivity (as exemplified in panel D Fig. 3).

(Figure 3 about here)

All the results we present are from simulations in which both network and learning dynamics run concurrently. We performed a preliminary exploration in the large parameter space by resorting to dynamic mean-field equations (Amit and Brunel (1997); Del Giudice et al. (2003); Fusi and Mattia (1999)).

The simulations have been carried out with a high-performance custom C program, implementing the event-based approach described in Mattia and Del Giudice (2000). In order to estimate instantaneous firing rates, the spikes from each neuronal population are sampled in a 10 ms sliding window.

3 Results

In the following we first illustrate the typical time course of a learning history, and describe the way it is affected by the build up of a selective top-down synaptic structure. We then move to illustrate how the network performance is affected by such top-down synaptic structure. Finally we study the network behaviour for corrupted versions of the training stimuli, and the time course of the neural activities in error trials, and formulate testable predictions in both cases.

3.1 Neural correlate of categorization

Fig. 4, panels A-B-C, shows the evolution of the network performance and synaptic configuration as learning proceeds. Before stimulation, all populations in the network are in a stable asynchronous state of low firing rate (few Hz). Upon stimulation, one of the two category populations in the PFC layer is in the end brought to an 'up state' ($\approx 40Hz$), resulting from winning the competition with the other category, which settles into a 'down state' $\approx 0Hz$.

Depending on whether the winning category is the correct one according to the defined

categorization rule, the plastic synapses are allowed to change following an Hebbian rule (for correct outcome) or a partially anti-Hebbian rule (for incorrect outcome), as described in Fig. 2 and in the Methods. We recall here that we repeat the stimulus after a wrong answer. In the figure, panel A shows the time course of the performance on the categorization task, i.e. the fraction of correct outcomes averaged over a non-overlapping sliding window of 30 trials, for the case in which only the bottom-up (ITC to PFC) synapses are plastic ('TD-off', black) and when both bottom-up and top-down synapses are plastic ('TD-on', grey). Performance is moderately affected by the presence of plastic top-down synapses: learning is seen to be slightly faster and more stable. As seen from Fig. 4, worse performance for the TD-off case is accompanied by larger fluctuations, as expected.

Moreover the differences between the bottom-up structures suggest that there is a better signal-to-noise ratio for the TD-on case even if the performances are similar for TD-on and TD-off, mainly due to the high stimulus contrast, that is the difference of stimulation to the two diagnostic population subsets $\Delta\lambda$ =0.3 Hz, used in these simulations. We will see in the following that as the stimulus contrast gets lower the top-down selective synaptic structure also entails differences in the performance and in the time needed for the categorization.

Panels B-C show the corresponding evolution of the fraction of potentiated synapses for the different bottom-up (B panel) and top-down (C panel) synapses, grouped according to the feature/category populations they connect. Only a representative subset of synapse groups is shown in the figure.

Panel D of Fig. 4, the three checkerboards, illustrate the final synaptic configurations, for both plastic and non-plastic top-down synapses. Each square represents the fraction of potentiated synapses at the end of the learning period. The synaptic weights connecting a diagnostic feature population with its (anti-) correlated category were (depressed) potentiated. For the synapse groups involving non-diagnostic features, the case shown is the computationally advantageous one in which the final bottom-up synapses are slightly depressed (which helps in obtaining a stable learning trajectory), and the top-down ones are markedly depressed (which results in a sort of 'selective amplification' of the relevant information, improving the categorization performance, see Fig. 6 and discussion below). The final synaptic configuration is consistent with the expectations explained in Fig. 1.

(Figure 4 about here)

We show the neural correlate of the categorization process in Fig. 5. The first three panels on the left show, for three successive stages during learning, the time course of the firing activity of

five populations, averaged over 20 correct trials (outcome $C1$). In each panel we plot the activity for the two category populations, ($C1$, $C2$ respectively requested and non-requested class), the population encoding the active value of the diagnostic feature ($\Delta\lambda = 0.3Hz$), one non-stimulated population and one of the stimulated populations coding for the non-diagnostic features.

Here it is seen that, as learning proceeds, the activities of diagnostic populations stimulated and not-stimulated (cyan and orange) split, consistent with the experimental observations of Sigala and Logothetis (2002), and with theoretical results obtained by Szabo et al. (2006) in a simpler context. The last panel in Fig. 5 shows the time course of the average activities for the same populations at the end of learning, but with non plastic top-down synapses, and confirms that the splitting observed for the TD-on case is an exquisite effect of a selective top-down synaptic structure.

(Figure 5 about here)

3.2 Influence of selective top-down synapses on network performances

We showed in Fig. 4 that the implemented plasticity in the top-down synapses: 1) sharpens the selectivity of the Bottom-Up synaptic structure, 2) moderately affects the time course of the learning dynamics. We also found (Fig. 5) that the selective top-down synaptic structure entails breaking the symmetry between the neural activity encoding stimulated diagnostic and non-diagnostic features, consistent with the results of Sigala and Logothetis (2002). One can still ask to what extent the above results also imply important effects on the computational performance of the network, thereby helping formulating informed guesses about the mechanisms underlying the corresponding experimental findings.

To understand this, it is important first to remark that our system is a noisy one and in fact it is subject to two very different main sources of noise, one 'exogenous' and the other 'endogenous'. The first depends on the dimensionality of the feature space in which the stimuli to be categorized are defined. The very definition of diagnostic *vs* non-diagnostic features implies that during learning the former are consistently associated with the defined categories, trial after trial, while the latter implement a random category labeling for each trial. Consequently, the higher the number of non-diagnostic features, the larger the associated component of the total input to PFC, that would overwhelm the one from diagnostic features in the absence of plasticity in Top-Down synapses, which realizes an effective amplification of this signal-to-noise ratio (which is then to be understood as the ratio between the inputs coming from diagnostic and non-diagnostic features). The other, 'endogenous' noise source is due to stochastic nature of the

neuronal network dynamics. The network has sparse connectivity, a different realization of which is generated for each simulation. This quenched noise, together with the finite-size effects due to the finite number of neurons in each population and the consequent distribution of firing rates, affects both the dynamics of decision (through the firing rate fluctuations) and the dynamics of learning (since, as explained in the Methods, the dynamics of synaptic plasticity, though rate-dependent on average, is stochastic, to the extent that the neural activities are). This endogenous noise, besides being a realistic ingredient of any finite and sparse system, contributes to expand the dynamic repertoire of the network (e.g. the range of decision times). Of course, an additional 'fast ' noise component is due to the Poisson spike trains implementing the stimuli.

In the following we quantify the effect of a selective top-down synaptic structure in diminishing the effect of the 'exogenous' noise, thus improving learning performance. We also investigate how the top-down synaptic structure affects the dynamics and characteristic times of the classification process.

At the end of learning (obtained with $\Delta\lambda = 0.3\,Hz$, $N_F = 16$), we freeze the synaptic structure and then apply new stimuli with different levels of contrast $\Delta\lambda = 0.05; 0.1; 0.2\ Hz$ ($\lambda = 4.2\ Hz$), and different numbers of stimulated non-diagnostic features ($N_F = 4$, 8 and 16), and test the performance. For each $\Delta\lambda$ and N_F we also test the performance of the network in which we substitute the learned, selective top-down synaptic structure with a uniform set of synapses ($J_P^{FB} = J_D^{FB} = J_n^{FB}$) whose efficacies are drawn from the same probability distribution, with an average such as to match the top-down synaptic efficacy averaged over each post-synaptic population of the structured TD-on case.

Fig. 6, panel A, shows the network performance (percentage of correct classification) for different numbers N_F of non-diagnostic features, and for three values of stimulus contrast $\Delta\lambda$, with selective and uniform top-down synaptic structure.

(Fig. 6 about here.)

For the same number of non-diagnostic features, performance increases with $\Delta\lambda$, as expected; the most relevant result is that, for given $\Delta\lambda$, the network with structured top-down synapses retains high performance even for a large number of non-diagnostic features. The effective signal-to-noise ratio which drives the competitive mechanism in the PFC layer is made almost independent from the number of 'distracting' non-diagnostic features, as a result of the depressed top-down synapses pointing to non-diagnostic features. In other words, if we call S and $S + \Delta S$ the total input to the two category populations, the structured top-down synapses determine a higher $\Delta S/S$ ratio: as the number of stimulated non-diagnostic features increases, the common

input S would be dominated by the activity of non-diagnostic populations, were it not for the combined effects of top-down depression of J_n^{FB} synapses, and the sharpened differentiation of (J_D^{FF} *vs* J_P^{FF}) and (J_D^{FB} *vs* J_P^{FB}) (see panel D of Fig. 6).

On the other hand we remark that the selective top-down synaptic structure does not increase the signal ΔS since, until the competition in the category layer sets in and a decision is taken, the diagnostic features populations receive equal feedback. This is consistent with the observation that, for low numbers of non-diagnostic features, the performance is essentially the same for structured and unstructured top-down synapses.

Fig. 6, panels B-D, illustrate the relationship between performance and decision times for the same values of N_F and $\Delta\lambda$ as in panel A, for both selective and uniform top-down synaptic structure. We defined the decision time (DT) as the instant when the absolute difference between the firing rates of the two category populations, divided by their sum, exceeds a given threshold D (we set $D = 0.7$ as in Marti et al. (2008)). The typical speed-vs-performance curve derived from psychophysics experiments is monotonically decreasing (as it derives from increasing the stimulus control parameter – $\Delta\lambda$ in our case – which entails increasing performance and decreasing decision time). The selective network (gray lines) is always faster than the uniform one (black lines) at equal performance. Thus the structured feedback not only favors better performance for the same input (panel A), but also makes the network more prompt to respond if the input current is adjusted to match the performance. In the case with uniform top-down, for given N_F, $C1$ and $C2$ receive greater symmetric input S, and are therefore less sensitive to the relative variations of their activities, such that they spend more time in a symmetric, 'undecided' state. The higher N_F, the greater the comparative advantage of the selective network (panel B to panel D). The DT gap between selective and uniform case for high performance is seen to be mostly due to a marked flattening of the DT *vs* performance curves for the uniform network: for the selective top-down network, DT preserves higher sensitivity to the selective input strength $\Delta\lambda$, and the expected DT *vs* performance fall-off is observed even for the highest N_F. The above observation can be understood again in terms of the larger increase of the symmetric input component S to $C1$ and $C2$ for the uniform network. As S increases with N_F, the dynamics spends longer and longer times around a state with almost equally high firing rates for both $C1$ and $C2$; $\Delta\lambda$ has little influence on this time, yet it still determines the performance of the network; taken together, these two effects explain the observed flattening of the RT *vs* performance curves. Indeed, an almost symmetric fixed point of the dynamics with high firing rates for $C1$ and $C2$ is expected to develop as S increases, around which the dynamics of decision is strongly distorted

and where the network stays for longer times.

Fig. 7 illustrates the different neural dynamics during the decision process for the selective and uniform networks, for $N_F = 16$ and $\Delta\lambda = 0.5Hz$, which underlie the behaviour described above. Panels A and B show the time course of firing rates of $C1$ and $C2$ (black and grey curves, respectively) during the fastest (thicker curves) and the slowest (thinner curves) trials. It is seen that the spread in decision times is much larger larger for the uniform case (see also the distributions of decision times in panels C and D). One possible interpretation, consistent with known phenomenology of cooperative-competitive models of decision making, can be obtained if we picture the network dynamics as the motion on an 'energy landscape', with each point being identified by the 'energy' level and the average firing rates of $C1$ and $C2$. The stable decision states of the network, induced by a stimulus, are identified by two asymmetric minima of the landscape (high firing of $C1$, low firing of $C2$, and viceversa). The diagnostic component of the stimulus determines a force driving the dynamics from the initial saddle the systems starts from (almost equal firing for $C1$ and $C2$) towards one of the two decision states, and it is expected to be maximally effective for the selective case, in which the component related to diagnostic features is amplified; for the uniform case the strong common component of the input to $C1$ and $C2$ makes the system roll on a flatter saddle, the departure from which, towards one of the asymmetric minima, take more time on average.

(Fig. 7 about here.)

Of course the chosen value of uniform top-down synapses affects the characteristic times of the decision dynamics; by varying this value between 0 (no feedback) and $2\,J^{FB}$ (beyond which the state of spontaneous activity of the network with no external stimuli is disrupted) we checked that: 1) performance is only slightly affected ($< 6\%$); 2) the qualitative features of the lines in Fig. 6 B-D are preserved, though the exact values of DT obviously change.

3.3 Corrupting learned stimuli: a footprint of categorization in ITC

The multi-population model system is a complex recurrent network, with local intra-modules feedback and inter-modules PFC-ITC feedback. One might then ask whether certain characteristic properties of recurrent networks could play a role in the situation under study, such as the pattern completion ability of attractor networks.

Starting from a network configuration obtained at the end of successful learning, we went on to stimulate the network with corrupted versions of the stimuli, i.e. stimuli for which only a fraction x of the neurons in the currently activated diagnostic population receive the usual

increased external current, as explained in the Methods Section 2.

In this situation, we studied the time course of the neural activities of the stimulated and non-stimulated neurons in the diagnostic population defining the current stimulus, and how it is affected by the learned, selective synaptic top-down structure. The expectation is that, as the decision process in the PFC matures, the unstimulated neurons belonging to the activated diagnostic feature would be recruited as a result of the combined effect of the selective top-down input and the recurrent interaction with the stimulated neurons in the same population.

Fig. 8 illustrates example results, obtained for low value of $\Delta\lambda = 0.05Hz$ (where we expect the recurrent synaptic structure to play a comparatively greater role), and $x = 0.75$. Neural activities are reported for three trials which happen to have (for the same network configuration and stimulation parameters) different decision times DT. Firstly we note that, as expected, as the decision process matures the unstimulated population gets recruited (green trace). Secondly, and interestingly, the recruitment occurs with a latency determined by the time it takes for the decision process to complete (compare the three panels in Figure 8). This latter observation is interesting because it constitutes a specific reflection of the time course of the decision process developing in PFC, in a time-dependent modulation occurring in the unstimulated neurons in ITC, at the single trial level.

(Fig. 8 about here.)

The same reasoning would also apply to a situation in which each stimulus is defined by multiple diagnostic features, in which case corrupting a stimulus might mean excluding one or more diagnostic features from the ones the stimulus is supposed to activate.

In other words, assuming the feature selectivity of ITC neurons is experimentally well characterized, so that one can identify the 'unstimulated' diagnostic neurons that correspond to a specific corruption of the stimuli, the point in time when those neurons would start to sharply increase their firing activity during the trial would signal the completion of the decision process, in the absence of simultaneous recording from PFC.

We remark that the observed recruitment of non-stimulated neurons coding for a diagnostic feature would not be observed in the absence of a category-related top-down information flow, which is consistent with the data reproduced in Fig. 3 where (for the anesthetized animals which do not perform decisions and hence top-down projection would be unavailable and pattern completion should not occur) an impoverished stimulus makes the corresponding feature representations disappear.

3.4 Selective Top-Down synapses sharpen tuning curve

For all the numerical experiment described so far we adopted a minimum feature representation (one diagnostic feature, with two values). It is known that feature-selective neurons in the temporal lobe exhibit a variety of tuning curves, see e.g. Freiwald et al. (2009). As already noted, we do not aim at reproducing specific aspects of the activity in the infero-temporal cortex; however, studying the implications of tuning curves in our 'ITC' layer is relevant from the computational point of view that concerns us here. Therefore we adopted a simple generically plausible shape of tuning curves and explore how they are affected by learning. We report results obtained for a network in which the diagnostic feature possesses four values (therefore four populations D1, ..., D4 code for it in the ITC layer); furthermore each stimulus is encoded by a profile of activation of the four diagnostic populations, with the maximal activation for the active value of the feature for that stimulus (as before we have N_F non-diagnostic features with two values each). In this way we implement a crude representation of tuning curves in this feature layer. Stimuli with maximal activation of D1 and D2 are to be mapped to class C1, D3 and D4 to C2. Stimuli with maximal activation of D1 and D4, for which the activation profiles of the four diagnostic populations has the smallest overlap, are 'easier' to classify with respect to D2 and D3; we will therefore label them as 'easy' and 'difficult' stimuli. One can expect that the build-up of the selective top-down synaptic structure could affect the activation profile of the diagnostic populations. In fact we observed (see Fig. 9) a marked sharpening of the tuning curves for both easy and difficult stimuli (compared panels A and B *vs* C and D), for which however we obtained slightly different final performances (see percentages of correct responses reported on the right panels of Fig. 9). The results shown in this figure constitute a further evidence of the computational consequence of a learned selective top-down synaptic structure and establish a specific experimental prediction. We remark that in our model the sharpening develops after the decision is taken, which suggests that if one could monitor the tuning curves in successive time intervals during trials, depending on when their sharpening occurs with respect to the decision time, one could get an indication about their origin being in a task-specific top-down signal.

(Fig. 9 about here.)

3.5 Counter-modulation in error trials

In Fig. 10 we show how network activities are differently modulated according to the correct (upper panel) or wrong (lower panel) outcome of the trial. Solid lines are the average activities

of the relevant populations over 5 consecutive trials approximately at an intermediate point of the learning process. The upper panel shows the expected (see Fig. 5) ranking of the stimulated diagnostic *vs* non-diagnostic features, which is destroyed for the wrong trials. Notice in particular the non-stimulated diagnostic feature value (orange trace), which as expected shows and increase of activity, being tightly correlated to the wrong decision state. This is because in the former case the stimulated diagnostic population receives, in addition to the stimulation, a coherent reinforcement from the correctly winning PFC population, with which learning has already formed strengthened synapses, while in the latter case the wrongly winning PFC population feeds back on it through synapses which have been weakened by learning. This observation suggests that the operation of a task-related selective feedback could be tested in *in-vivo* experiments, by comparing the modulation of the activities of neurons with different selectivities in correct and wrong trials. A hint that such a strategy could be viable and informative is provided in Mirabella et al. (2007), where evidence is provided that the activity of neurons in V4 during a visual selection task involving attention is differently modulated depending on the trial being correct or wrong.

(Fig. 10 about here.)

(Table 1, 2 and 3 about here)

4 Discussion

The idea that perceptual learning, visual categorization and/or selective attention, involve an effective suppression of irrelevant sensory input is not new (i.e. see Riesenhuber and Poggio (1999); Bar (2003); Spratling and Johnson (2004); Roelfsema and van Ooyen (2005)). For the specific case of visual categorization, even if experimental evidence is still inconclusive as to whether top-down control by PFC is needed to accomplish it (see Minamimoto et al. (2010); Buckley and Sigala (2010)), several reported results suggest a role of ITC-PFC mutual interaction in the arbitrary stimulus-category association. In the present work we put this idea in a specific context and address a general computational issue, i.e. the implication of learned, selective top-down projections between a category-aware area and a sensory-coding one, taking visual classification as a relevant case in point. With a simplified ITC-PFC model designed to account for the key results shown in Sigala and Logothetis (2002) and in De Baene et al. (2008), we showed that a semi-anti-Hebbian, spike-driven learning mechanism generates a selective amplification of task-relevant neural representations in ITC and, because of this, enhanced classification performance

and faster response. In the present work we adopted the simplest choice for the feature space defining the stimuli, i.e. just one diagnostic feature with two possible values. We remark, and checked in simulation in a few cases, that as long as the classification problem remains linearly separable, enlarging the dimension of our one-dimensional 'diagnostic sub-space' (as in the case of the two-dimensional sub-space of Sigala and Logothetis (2002)) does not spoil the classification ability of the network (it acts in fact as a Perceptron, Rosenblatt (1958)), nor the mechanism of selective amplification of diagnostic information. However, for the higher-dimensional case the synaptic configuration generated by on-line learning does depend on the choice of the training stimuli, on the presentation sequence and on the initial condition (as it is the case for the Perceptron). This work is one of the few so far addressing dynamic learning effected by the ongoing spike-driven synaptic dynamics of LTP and LTD, coupled in closed loop with the stimulus-driven spiking activity in a multi-modular network. We showed how robust learning histories lead the global network to perform well in the face of the many sources of instabilities that affect dynamic learning (see Del Giudice and Mattia (2001); Del Giudice et al. (2003); Amit and Mongillo (2003)). In particular, finite-size effects are important, and bring about deviations from the the predictions of the mean field theory which guide simpler approaches like the one we adopted in Szabo et al. (2006). For a finite number of synaptic connections per neuron each population has a distribution of emission rates. We remind that our synapses are stochastic, as long as neural activities are (see Methods). Considering for example the synapses connecting populations of neurons stimulated by the same stimulus, and therefore supposed to get potentiated, the high- and low-rate tails of the actual frequency distribution corrupt the homogeneity of the pattern of synaptic transition probabilities, such that in the same synaptic group some synapses will have too high LTP probability, while others will be unexpectedly unchanged. Similarly, finite-size effects can provoke unwanted synaptic transitions where they are not expected and harmful (such as a potentiation of synapses involving post-synaptic background neurons, which can become the seed of instability for the spontaneous state). One ingredient which makes finite-size effects more or less harmful is the character of the 'synaptic transfer function', meaning the function giving the LTP/LTD transition probabilities as functions of the pre- and postsynaptic emission rates. The sensitivity of this function in the critical region where the rate distributions involved overlap is an important factor in determining how serious finite-size effects are going to be. These and other effects make online learning with realistic, spike-driven synaptic dynamics a major challenge that we faced in this work which, besides the value of the new results and predictions we provide, hopefully contributes to advance a domain of modeling studies that needs progress. The

assumed plasticity for the top-down (PFC-to-ITC) synapses turns out to ensure slightly faster, and significantly more robust, learning histories (Fig. 4).

The value of a model lies of course in its ability to generate testable predictions. We list in the following some speculative implications of our model. The learned selective top-down synaptic structure essentially lowers the effect of the 'noise' associated with the non-diagnostic features, by amplifying the 'signal' component associated with the diagnostic features; it is therefore expected to matter more as the number N_F of such features increases. This is indeed what we showed in Fig. 6, where it is seen how the improvement in the classification performance increases markedly with N_F. Note that, because of the non-linearity of network responses, the additive top-down feedback can in fact produce essentially multiplicative effects, as it has been shown in the context of attention modeling in Deco and Rolls (2005).

We expect the predicted task-specific modulation not to depend much on the details of the chosen model, and to be shared by any model based on selective amplification of task-relevant sensory information. We can then speculate that, if the global neural activity induced by a stimulus in ITC is measured, for instance by the BOLD signal, it would be more evenly distributed in the naive subject, before learning the task, and more concentrated in the well trained subject: before learning, the patchy ITC representation would generate signal spots of comparable intensity, while after learning the signals associated with the amplified, task-relevant features would pop up. Also, one can predict that turning a non-diagnostic feature into a diagnostic one in a previously learned set of stimuli would result in an expansion of the dominant signal spots.

Previous works (Op De Beeck et al. (2006); Gillebert et al. (2009)) studied the FMRi correlate of learning categorization, showing that the BOLD signal associated with categorized images is enhanced after learning. Our results are compatible with those findings, in that the global activity in our feature layer after learning is indeed higher. However our results together with the electrophysiological results of Sigala and Logothetis (2002); De Baene et al. (2008), suggest the above prediction which goes beyond these findings, i.e. the spatial variability of the BOLD signal should reflect the differential activation corresponding to diagnostic and non-diagnostic features.

Our results shown in Fig.9 suggest that if the proposed mechanism of selective amplification of diagnostic features representation operates, learning the categorization task would result in sharpen tuning curves for diagnostic features. This effect is qualitatively compatible with the differential category-tuning reported by De Baene et al. (2008).

We showed in Fig. 10 that error trials entail a qualitatively different modulation of neural activities. This generates a testable prediction, in line with previous suggestions like Mirabella et al. (2007), in which evidence is provided that the activity of neurons in V4 during a visual selection task (involving attention) is modulated depending on the trial being correct or wrong. It is clear that such an option would be viable only for a version of the task allowing for a non negligible error rate even in well trained subjects.

Results shown in Fig. 6 B-D suggest that some of the direct implications of the selective feedback synaptic structure envisaged in the model might be amenable to investigation in psychophysics. Indeed, both the relevant parameters $\Delta\lambda$ and N_F playing a role in shaping the plots of Fig. 6 B-D can be experimentally varied (for N_F various choices are available in principle, including adding new features or making pre-existing non-diagnostic features variable among stimuli). The non-trivial, and interesting, experimental question relates to the interplay between $\Delta\lambda$ and N_F: $\Delta\lambda$ has the dual role of biasing the decision dynamics and also (when feedback is present and not yet selective) to cause an increase in the activity induced by non-diagnostic features. Therefore, by looking at the changes in the DT *vs* performance plots while the subject is "learning" to ignore one more non-diagnostic feature one could qualitatively check the prediction implied by a selective feedback buildup.

More in general, one can easily imagine a situation (frequently explored both in psychophysics and in electrophysiology) where a pair of features are consistently associated with the same category, or where stimuli are presented in different sensory modalities. In such situations, after training, a generic and testable prediction is that the presentation of one member of the pair, or the presentation of one sensory modality, would entail the partial activation of the other member of the pair, or the representation pertaining to the other sensory modality, respectively.

In Fig. 3 we suggested that the chosen architecture of our simplified model is consistent with reported evidence concerning visual objects representations in ITC. In Tsunoda et al. (2001), besides reporting the patchiness of object representation in IT cortex that we referred to in the Methods, the authors also notice that, as the visual appearance of objects is deprived of features that, though 'small', are key for its interpretation, not only the corresponding activity patches disappear, but also new ones appear. The authors suggest that the distributed object representation would result from the activation, and active suppression, of a constellation of (possibly overlapping) feature-specific neural populations. Based on our modeling results, we can speculate that the task-dependent modulation of selectivity induced by learning might reshape such regions of overlapping representations (see Fig. 3). Specifically, if a IT neuron, before

learning, has a mixed selectivity for two (values of) features, if the latter are mapped by the rule to the same category, the changes induced by learning in the synapses linking that neuron to the PFC category populations are consistent, and this will end up boosting the activities in the mixed selectivity neurons. On the other hand, if the mixed selectivity relates to features belonging to different categories, we can expect that the original differences in the neural responses to those features get amplified. If neurons in the overlapping patches have a nontrivial distribution of firing rates for the two features, learning would result in shrinking the overlapping regions. While the details will depend on several factors (such as the initial tuning curves for the involved features, the frequency of presentation of different stimuli, the nonlinearities in the dependence of synaptic changes on neural activity), we suggest that a task-dependent modulation of the overlaps in the patchy ITC representation would be expected. With reference to the categorization problem in Sigala and Logothetis (2002), in which the category boundary is linear in the two-dimensional diagnostic space, the proposed effect can be seen as a modulation of the classifier margin. Also, the differential modulation of diagnostic and non-diagnostic feature representations could in principle account for the appearance of previously suppressed features when key elements of the visual object are removed (observed in Tsunoda et al. (2001)), via the mutual inhibitory interactions in the ITC layer.

Finally, the experiments performed with corrupted stimuli, besides confirming expectations based on the recurrent network architecture, suggest that (thanks to the learned selective top-down synaptic structure) a specific neural correlate of the decision process being completed would be available in a feature-encoding area.

References

Amit, D. J. and Brunel, N. (1997). Dynamics of a recurrent network of spiking neurons before and following learning. Network, 8:373–404.

Amit, D. J. and Mongillo, G. (2003). Spike-driven synaptic dynamics generating working memory states. Neural Computation, 15(3):565–596.

Bar, M. (2003). Cortical mechanism for triggering top-down facilitation in visual object recognition. J. of Cogn. Neurosci., 15:600–609.

Boettiger, C. and D'Esposito, M. (2005). Frontal networks for learning and executing arbitrary stimulus-response associations. J. Neurosci., 25(10):2723–2732.

Buckley, M. and Sigala, N. (2010). Is top-down control from prefrontal cortex necessary for visual categorization? Neuron, 66:471–473.

De Baene, W., Ons, B., Wagemans, J., and Vogels, R. (2008). Effects of category learning on the stimulus selectivity of macaque inferior temporal neurons. Learning and memory, 15:717–727.

Deco, G. and Rolls, E. (2005). Neurodynamics of biased competition and cooperation for attention: A model with spiking neurons. J. of Neurophysiol., 94:295–313.

Del Giudice, P., Fusi, S., and Mattia, M. (2003). Modeling the formation of working memory with networks of integrate-and-fire neurons connected by plastic synapses. Journal of Physiology (Paris), 97:659–681.

Del Giudice, P. and Mattia, M. (2001). Long and short-term synaptic plasticity and the formation of working memory: A case study. Neurocomputing, 38-40(1-4):1175–1180.

Freedman, D. and Assad, J. (2011). A proposed common neural mechanism for categorization and perceptual decisions. Nature Neuroscience, 14(2):143–146.

Freedman, D., Riesenhuber, M., Poggio, T., and Miller, E. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. Science, 291(5502):312–6.

Freedman, D. J., Riesenhuber, M., Poggio, T., and Miller, E. K. (2003). A comparison of primate prefrontal and inferior temporal cortices during visual categorisation. J. of Neurosci., 23:5235–5246.

Freiwald, W., Tsao, D., and Livingstone, M. (2009). A face feature space in the macaque temporal lobe. Nat Neurosci, 12 (9):1187–1196.

Fusi, S., Annunziato, M., Badoni, D., Salamon, A., and Amit, D. J. (2000). Spike-driven synaptic plasticity: theory, simulation, VLSI implementation. Neural Computation, 12:2227–2258.

Fusi, S. and Mattia, M. (1999). Collective behavior of networks with linear (VLSI) integrate and fire neurons. Neural Computation, 11:633–652.

Gillebert, C. R., Op De Beeck, H. P., Panis, S., and Wagemans, J. (2009). Subordinate categorization enhances the neural selectivity in human object-selective cortex for fine shape differences. Journal of Cognitive Neuroscience, 21(6):1054–1064.

Howard, J., Plailly, J., Grueschow, M., Haynes, J., and Gottfried, J. (2009). Odor quality coding and categorization in human posterior piriform cortex. Nat. Neurosci., 12(7):12912–42.

Humphries, J., Stewart, R., and Gurney, K. (2006). A physiologically plausible model of action selection and oscillatory activity in the basal ganglia. J. Neurosci., 26:12921–42.

Knoblich, U., Riesenhuber, M., Freedman, D., Miller, E., and Poggio, T. (2002). Visual categorization: How the monkey brain does it. In Bulthoff, H., Wallraven, C., Lee, S.-W., and Poggio, T., editors, Biologically Motivated Computer Vision, volume 2525 of Lecture Notes in Computer Science, pages 305–327. Springer Berlin / Heidelberg.

Marti, D., Deco, G., Mattia, M., Gigante, G., and Del Giudice, P. (2008). A fluctuation-driven mechanism for slow decision processes in reverberant networks. PLoS ONE, 3(7):e2534.

Mattia, M. and Del Giudice, P. (2000). Efficient event-driven simulation of large networks of spiking neurons and dynamical synapses. Neural Computation, 12:2305–2329.

Meyers, E., Freedman, D., Kreiman, G., Miller, E., and Poggio, T. (2008). Dynamic population coding of category information in inferior temporal and prefrontal cortex. Journal of Neurophysiology, 100(3):1407–1419.

Minamimoto, T., Saunders, R. C., and J., R. B. (2010). Monkeys quickly learn and generalize visual categories without lateral prefrontal cortex. Neuron, 66(4):501–507.

Mirabella, G., Bertini, G. Samengo, I., Kilavik, B., Frilli, D., Della Libera, C., and Chelazzi, L. (2007). Neurons in area v4 of the macaque translate attended visual features into behaviorally relevant categories. Neuron, 54:303–318.

Muhammad, R., Wallis, J., and Miller, E. (2006). A comparison of abstract rules in the prefrontal cortex, premotor cortex, inferior temporal cortex, and striatum. J. Cogn. Neurosci., 18:974–989.

Myers, C., Shohamy, D., Gluck, M., Grossman, S., and A., K. (2003). Dissociating hippocampal versus basal ganglia contributions to learning and transfer. J. Cogn. Neurosci., 15:185–93.

Op De Beeck, H., Baker, C., DiCarlo, J., and Kanwisher, N. (2006). Discrimination training alters object representations in human extrastriate cortex. Journal of Neuroscience, 26(50):13025–13036.

Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex. Nat. Neurosci., 2:1019–1025.

Roelfsema, P. and van Ooyen, A. (2005). Attention-gated reinforcement learning of internal representations for classifiation. Neural computation, 17:2176–2214.

Roelfsema, P., van Ooyen, A., and Watanabe, T. (2010). Perceptual learning rules based on reinforcers and attentions. TINS review, 14:64–71.

Romo, R. and Salinas, E. (2001). Touch and go: decision-making mechanisms in somatosensation. Annu. Rev. Neurosci., 24:107–37.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev., 65(6):386–408.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. Journal of Neurophysiology, 80:1–27.

Schultz, W. and Dickinson, A. (2000). Neural coding of prediction errors. Annu. Rev. Neurosci., 23:43–500.

Seger, C. (2008). How do the basal ganglia contribute to categorization? Their roles in generalization, response selection, and learning via feedback. Neurosci. Biobehav. Rev., 32:265–78.

Seger, C. and Miller, E. (2010). Category learning in the brain. Ann. Rev. of Neurosci., 33:203–219.

Shohamy, D. and Wagner, A. (2008). Integrating memories in the human brain: Hippocampal-midbrain encoding of overlapping events. Neuron, 60(2):378–389.

Sigala, N. and Logothetis, N. K. (2002). Visual categorisation shapes feature selectivity in the primate temporal cortex. Nature, 415:318–320.

Spratling, M. W. and Johnson, M. H. (2004). A feedback model of visual attention. Journal of Cognitive Neuroscience, 16:219–237.

Swaminathan, S. and Freedman, D. (2012). Preferential encoding of visual categories in parietal cortex compared with prefrontal cortex. Nature Neuroscience, 15(2):315–320.

Szabo, M., Deco, G., Fusi, S., Del Giudice, P., Mattia, M., and Stetter, M. (2006). Learning to attend: Modeling the shaping of selectivity in infero-temporal cortex in a categorization task. Biological Cybernetics, 94:351–365.

Tsunoda, K., Yamane, Y., Nishizaki, M., and Tanifuji, M. (2001). Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns. Nat Neurosci, 4:832–838.

Vallabha, G., McClelland, J., Pons, F., Werker, J., and Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. Proc. Natl. Acad.Sci, 104:13273–78.

Vogels, R. (1999). Categorization of complex visual images by rhesus monkeys. Eur. J. Neurosci., 11:1239–55.

Wang, X. (2002). Probabilistic decision making by slow reverberation in cortical circuits. Neuron, 36(5):955–968.

Yamane, Y., Tsunoda, K., Matsumoto, M., Phillips, A., and Tanifuji, M. (2006). Representation of the spatial relationship among object parts by neurons in macaque inferotemporal cortex. J Neurophysiol, 96:3147–3156.

Figure 1: **Model architecture and expected learned synaptic structure** The dark gray rectangles labeled D and O represent the feature-selective populations in ITC ($N_F = 1$). Light gray blobs represent subsets coding for feature values in ITC and category-coding populations in PFC. Black blobs are the inhibitory populations, white blobs are the excitatory background, non-selective populations. D1, D2 and O1, O2 denote the two values of the 'diagnostic' and 'non-diagnostic' features respectively (see text); C1 and C2 denote the categories to be associated with stimuli. Category and feature populations are bi-directionally connected only by excitatory synapses J_P^{FF}, J_D^{FF}, J_n^{FF}, and J_P^{FB}, J_D^{FB}, J_n^{FB} which are the only plastic synapses in the network (FF/FB: ITC to/from PFC). The figure illustrates the expected synaptic structure reciprocally linking ITC and PFC layers as a result of learning the classification task: strong connections between $D1/D2$ and $C1/C2$ – high J_P^{FF}, J_P^{FB}; weak connections between $D1/D2$ and $C2/C1$ – low J_D^{FF}, J_D^{FB}. Synapses from/to non-diagnostic populations (J_n^{FF}, J_n^{FB}) will take on values intermediate between the corresponding J_P and J_D.

Figure 2: **Learning mechanism** The dynamics of the plastic synapses follows the model proposed in Fusi et al. (2000). The synapses connecting ITC and PFC neurons have two values of synaptic efficacies, which are determined by the state of an internal synaptic variable (see text). Synaptic changes occur in the final stage of each trial, when the winner-take-all mechanism has selected one of the two category-selective populations, which constitute the 'behavioural' outcome of the trial. Changes in the synaptic state variable are triggered by presynaptic spikes and the sign of the change (up- or down-regulation) is determined in a semi-supervised fashion by a threshold condition on the postsynaptic depolarization, and the outcome of the trial. If the 'right' or 'wrong' category was selected, the sign of synaptic changes is determined by an Hebbian or partially anti-Hebbian mechanism, respectively (in the latter case the synapses connecting active pre- and post-synaptic neurons are down-regulated, and no changes occur otherwise).

Figure 3: **Features-based representation in ITC** Panel A shows the observed patchiness of the neural representations of increasingly simplified versions of a visual object in ITC, as shown in Tsunoda et al. (2001). Panel B illustrates the assumed analogous patchy representation in ITC for the representation of the Brunswick faces used in Sigala and Logothetis (2002). The scheme is then mapped onto the model architecture described in panel C and in Fig. 1. Panel D illustrates the expected firing patterns of neurons belonging to non overlapping (blue or red) or overlapping (purple) patches of panel B, to illustrate the concept that a selection of recorded neurons based on their feature selectivity can be thought to give rise to the segregated representation in panel C.

Figure 4: **Performances on the categorization task and evolution of ITC-PFC synaptic connectivity during learning.** Panel A shows the time course of the performances on the categorization task, i.e. the fraction of correct outcomes averaged over a sliding window of 30 trials with a step of 30 trials, for the case in which only the bottom-up (ITC to PFC) synapses are plastic ('TD-off', black) and when both bottom-up and top-down synapses are plastic ('TD-on', gray). The vertical dashed line marks the moment after which the difference between average black and grey curves becomes stastically significant (two standard errors). Panels B-C show the corresponding evolution of the fraction of potentiated synapses for the different bottom-up (center panel) and top-down (bottom panel) synapses, grouped according to the feature/category populations they connect. Only a representative subset of synapse groups are shown in the figure, even if the learning stimuli are composed of 16 non-diagnostic features. Panel D, the three checkerboards, illustrates the final synaptic configurations, for both plastic and non-plastic top-down synapses.

Figure 5: **Sample time course of neuronal firing rates for different populations in the network.** The figure shows, for three stages during learning (initial, intermediate and final from left to right), the time course of the firing activity during a correct trial (outcome $C1$), averaged over 20 trials. The shading is the s.e.m.. In each panel we plot the activity for the two category populations, ($C1$ or requested class, $C2$ or non-requested class), the population of the stimulated diagnostic feature value ($\Delta\lambda = 0.3Hz$), non-stimulated diagnostic feature value and , one of the stimulated non-diagnostic feature value out of 16 non diagnostic features in total respectively in blue, red, cyan, orange and black.

Figure 6: **The structure of top-down synapses influences performance and DTs.** Panel A shows the categorization performances *vs* number of non-diagnostic features, for different stimulus contrasts ($\Delta\lambda = .05; .1; .2\ Hz;\ \lambda = 4.2\ Hz$). Solid-gray/dashed-black lines are for selective/uniform Top-Down synapses. Different markers indicate the stimulus contrast $\Delta\lambda$. Panels B-D illustrate the behaviour of decision time (DT) vs performances for the same values of N_F and $\Delta\lambda$ as in panel A, for both selective and uniform Top-Down synaptic structure; for panel D we used $\Delta\lambda = .05; .1; .2; .3; .5\ Hz$. We recall that DT is defined as the instant when the difference in the firing rates of the two category populations, divided by their sum, exceeds a given threshold D (we put $D = 0.7$ as in Marti et al. (2008)).

Figure 7: **Examples of different neural dynamics for the selective and uniform Top-Down networks**, for $N_F = 16$ and $\Delta\lambda = 0.5Hz$. Panels A and B show the time course of firing rates of $C1$ and $C2$ (black and grey curves, respectively) during the fastest (thicker curves) and the slowest (thinner curves) trials. Panels C and D provide respectively the distribution of DT for the selective and uniform Top-Down networks.

Figure 8: **Time course of the network activity with corrupted versions of learned stimuli** The three panels show for low value of $\Delta\lambda = 0.05Hz$ the neural activities during three trials having different decision times DT, when the network is presented with a corrupted stimulus. In the cases shown the activated feature value is $D1$ and the fraction of activated neurons is $x = 0.75$. In each plot is reported the approximate value of DT. From left to right DT increases.

Figure 9: **Tuning curves in ITC before and after learning.** Each panel shows the activation profile of four value-selective subsets in ITC before (A, C) and after (B, D) learning, averaged over five trials (error bars are standard errors). We also report in panels B and D the final performance after learning. Panels A,B (C,D) show the activation profile following the presentation of a stimulus identified by the value D2 (D1) of the diagnostic feature. D2 corresponds to a difficult stimulus, D1 to an easy one (see text).

Figure 10: **Time course of the network activity in correct and wrong trials.** Time course of the network activity is plotted for correct trials (top) and wrong trials (bottom). Solid lines are averages over 5 trials, shaded strips are the standard errors of the mean. The plot refer to an intermediate stage of learning (when the rate of errors is still appreciable).

Single neuron parameters	**Value**
θ - Spike emission threshold	1.0 ua
V_r - Reset Potential	0.0 ua
τ_{arp} - Absolute refractory period	2.0 ms
β_E - Decay coefficient of excitatory neurons (ITC)	453 ua/ms
β_I - Decay coefficient of inhibitory neurons (ITC)	40 ua/ms
β_E - Decay coefficient of excitatory neurons (PFC)	54 ua/ms
β_I - Decay coefficient of inhibitory neurons (PFC)	20 ua/ms

Frequency of external stimulation	**Value [Hz]**
ν_{ext} - Extern neurons activity (PFC)	2.0
λ_0 - Extern neurons activity (ITC)	3.5
λ - Extern neurons activity (ITC, stimulus)	4.2
$\Delta\lambda$ - Extern neurons activity (ITC, extra-stimulus)	0.3

Net Parameter	**Value**
N_{Ex} - Extern excitatory neurons	1200
N_E - Excitatory neurons of the layer PFC	4800
N_E - Excitatory neurons of the layer ITC	4800
N_I - Inhibitory neurons of ITC and PFC	1200
$N_{category}$ - PFC selective excitatory neurons	480
$N_{feature}$ - ITC selective excitatory neurons	120
N_F - Non-diagnostic features	16
c_{in} - Internal connectivity (intra-layer)	0.25
c_{ex} - Extern connectivity (inter-layer)	0.25
δ_E^m - Minimum delay excitatory neurons	2 ms
δ_E^M - Maximum delay excitatory neurons	80 ms
δ_I^m - Minimum delay inhibitory neurons	0 ms
δ_I^M - Maximum delay inhibitory neurons	4 ms

Table 1: **Neuron parameters**

Synaptic parameters layer *PFC*	**Value $[\theta - V_r]$**
$J^{\mathrm{PFC}}_{\mathrm{E\,ext}}$ - Synaptic efficacy ext $\to E$	0.022
$J^{\mathrm{PFC}}_{\mathrm{I\,ext}}$ - Synaptic efficacy ext $\to I$	0.015
$J^{\mathrm{PFC}}_{\mathrm{EE}}$ - Synaptic efficacy $Bg \to x$, with $x \in (Bg, C1, C2)$	0.005
$J^{\mathrm{PFC}}_{\mathrm{IE}}$ - Synaptic efficacy $I \to E$	0.010
$J^{\mathrm{PFC}}_{\mathrm{EI}}$ - Synaptic efficacy $E \to I$	-0.012
$J^{\mathrm{PFC}}_{\mathrm{II}}$ - Synaptic efficacy $I \to I$	-0.028
$J^{\mathrm{PFC}}_{\mathrm{rec}}$ - Potentiated synaptic efficacy $x \to x$, with $x \in (C1, C2)$	0.011
$J^{\mathrm{PFC}}_{\mathrm{BE}}$ - Synaptic efficacy $x \to Bg$, with $x \in (C1, C2)$	0.040
$J^{\mathrm{PFC}}_{\mathrm{ME}}$ - Synaptic efficacy $x \to y$, with $x, y \in (C1, C2)$	0.0015

Synaptic parameters layer *ITC*	**Value $[\theta - V_r]$**
$J^{\mathrm{ITC}}_{\mathrm{E\,ext}}$ - Synaptic efficacy ext $\to E$	0.087
$J^{\mathrm{ITC}}_{\mathrm{I\,ext}}$ - Synaptic efficacy ext $\to I$	0.015
$J^{\mathrm{ITC}}_{\mathrm{EE}}$ - Synaptic efficacy $E \to E$	0.011
$J^{\mathrm{ITC}}_{\mathrm{IE}}$ - Synaptic efficacy $I \to E$	0.010
$J^{\mathrm{ITC}}_{\mathrm{EI}}$ - Synaptic efficacy $E \to I$	-0.070
$J^{\mathrm{ITC}}_{\mathrm{II}}$ - Synaptic efficacy $I \to I$	-0.031
$J^{\mathrm{ITC}}_{\mathrm{rec}}$ - Potentiated synaptic efficacy $x \to x$, with $x \in (C1, C2)$	0.015
$J^{\mathrm{ITC}}_{\mathrm{BE}}$ - Synaptic efficacy $x \to Bg$, with $x \in (C1, C2)$	0.040
$J^{\mathrm{ITC}}_{\mathrm{ME}}$ - Synaptic efficacy $x \to y$, with $x, y \in (C1, C2)$	0.0015

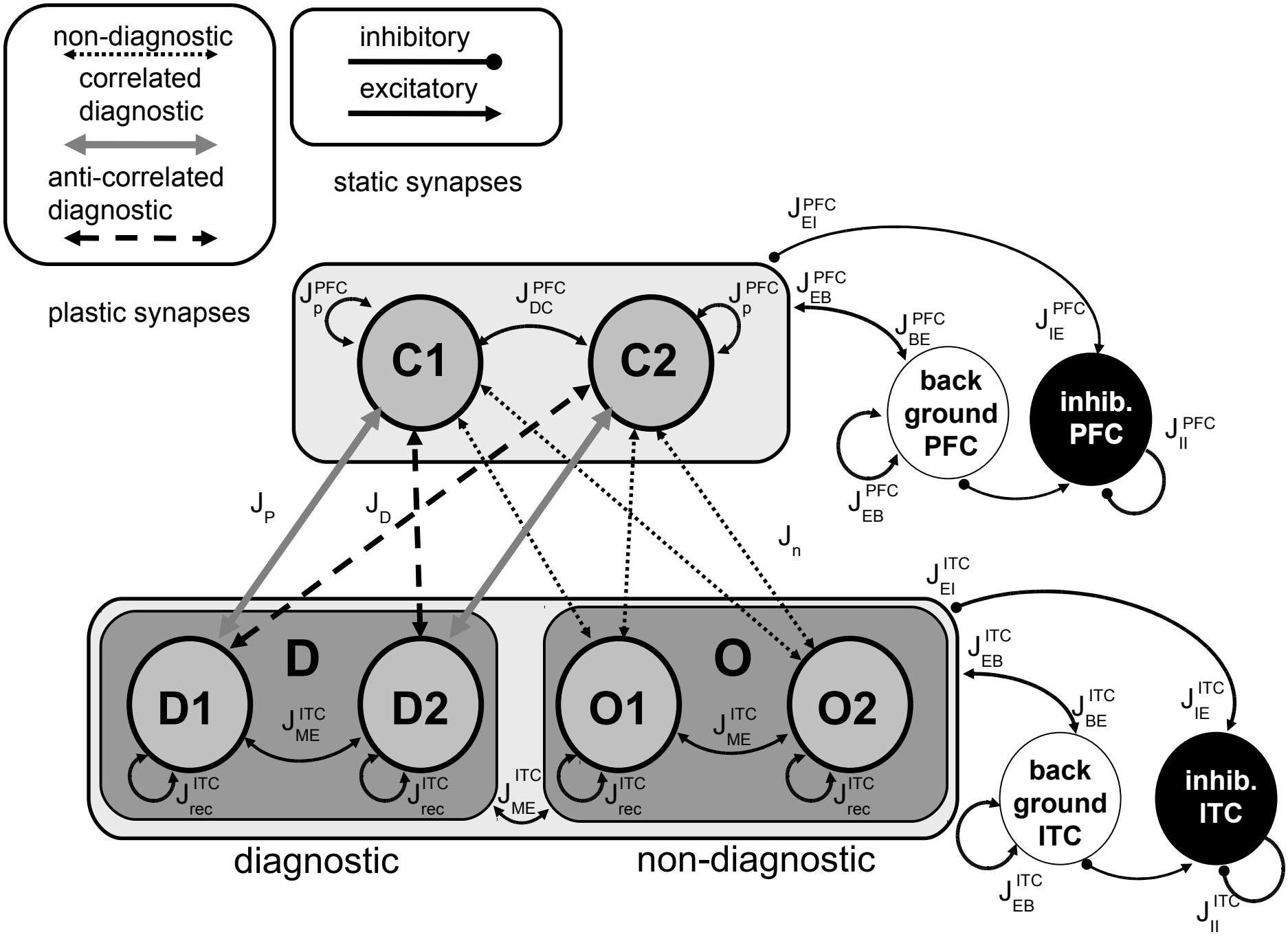
Synaptic parameters *ITC* $\leftrightarrow$ *PFC*	**Value $[\theta - V_r]$**
J^{FF}_{+} - Potentiated synaptic efficacy $ITC \to PFC$	0.007
J^{FB}_{+} - Potentiated synaptic efficacy $ITC \leftarrow PFC$	0.03
J^{FF}_{-} - Depressed synaptic efficacy $ITC \to PFC$	0.003
J^{FB}_{-} - Depressed synaptic efficacy $ITC \leftarrow PFC$	0

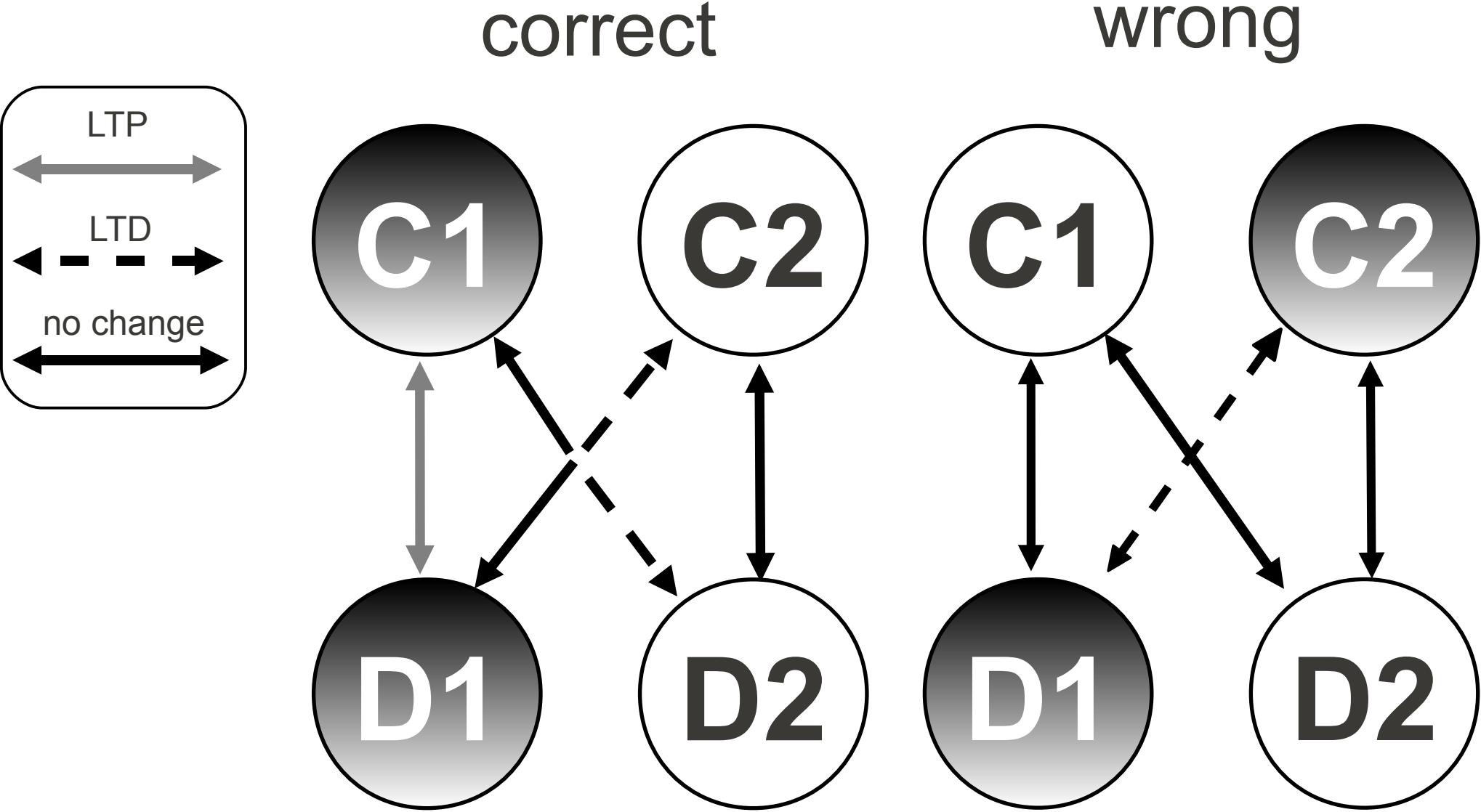
Table 2: **Synaptic efficacies**

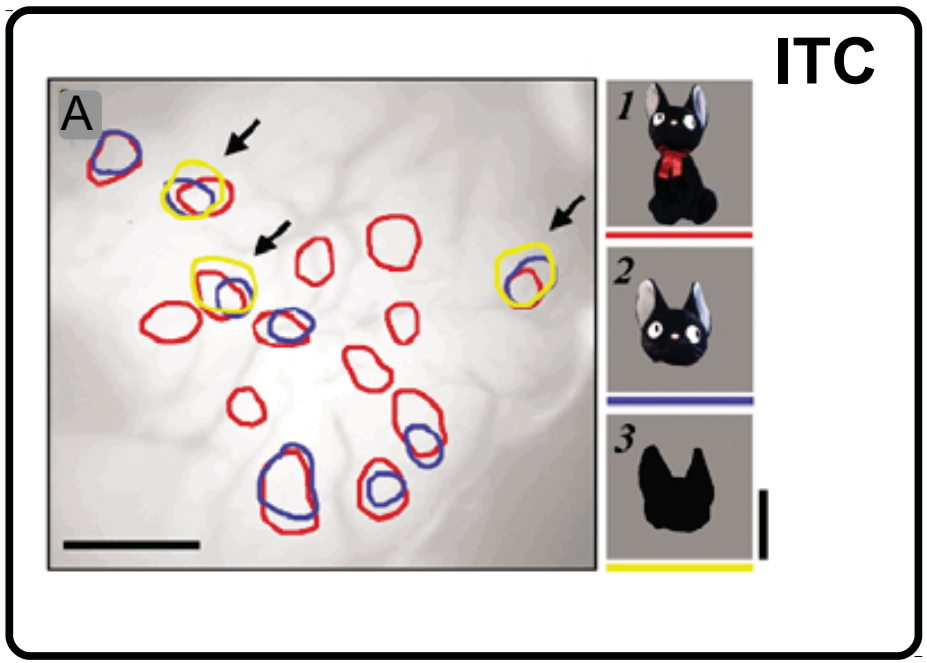
Internal synaptic variable parameters, $ITC \rightarrow PFC$	**Value** $[\theta - V_r]$
dX_- - Reward synaptic negative jump	0.085
dX_+ - Reward synaptic positive jump	0.080
dX_- - No-Reward synaptic negative jump	0.108
dX_+ - No-Reward synaptic positive jump	0.082
α_X - Synaptic positive/negative drift	0.0003
Θ_V - Threshold	0.245
Θ_J - Threshold	0.5

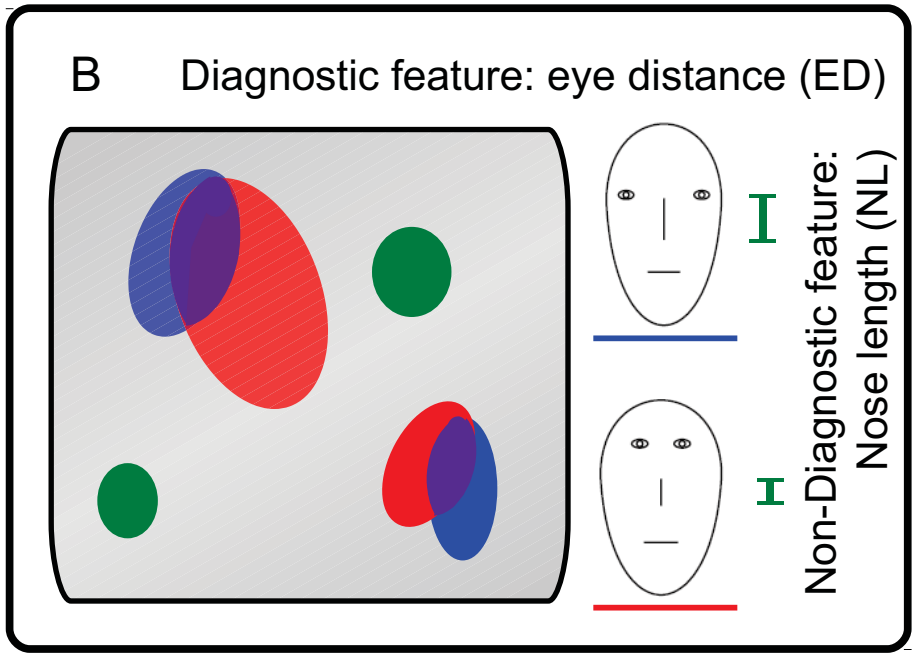
Internal synaptic variable parameters, $PFC \rightarrow ITC$	**Value** $[\theta - V_r]$
dX_- - Reward synaptic negative jump	0.095
dX_+ - Reward synaptic positive jump	0.096
dX_- - No-Reward synaptic negative jump	0.093
dX_+ - No-Reward synaptic positive jump	0.052
α_X - Synaptic positive/negative drift	0.002
Θ_V - Threshold	0.115
Θ_J - Threshold	0.5

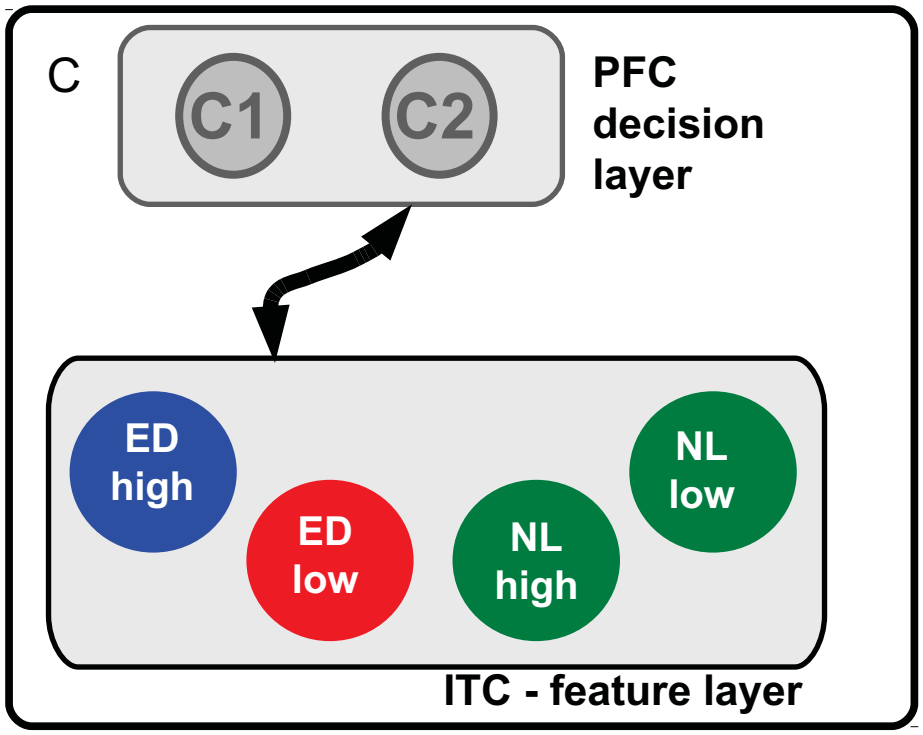
Table 3: **Parameters for synaptic dynamics**

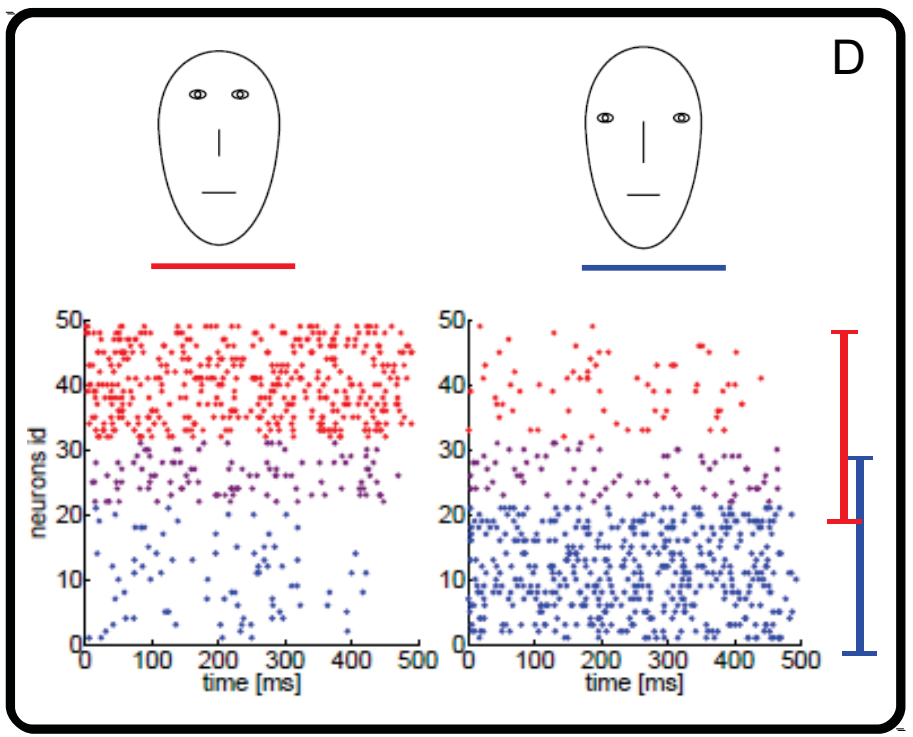

non-diagnostic
correlated
diagnostic
anti-correlated
diagnostic
plastic synapses
inhibitory
excitatory
static synapses
C1
C2
D
D1
D2
O
O1
O2
diagnostic
non-diagnostic
back ground PFC
inhib. PFC
back ground ITC
inhib. ITC

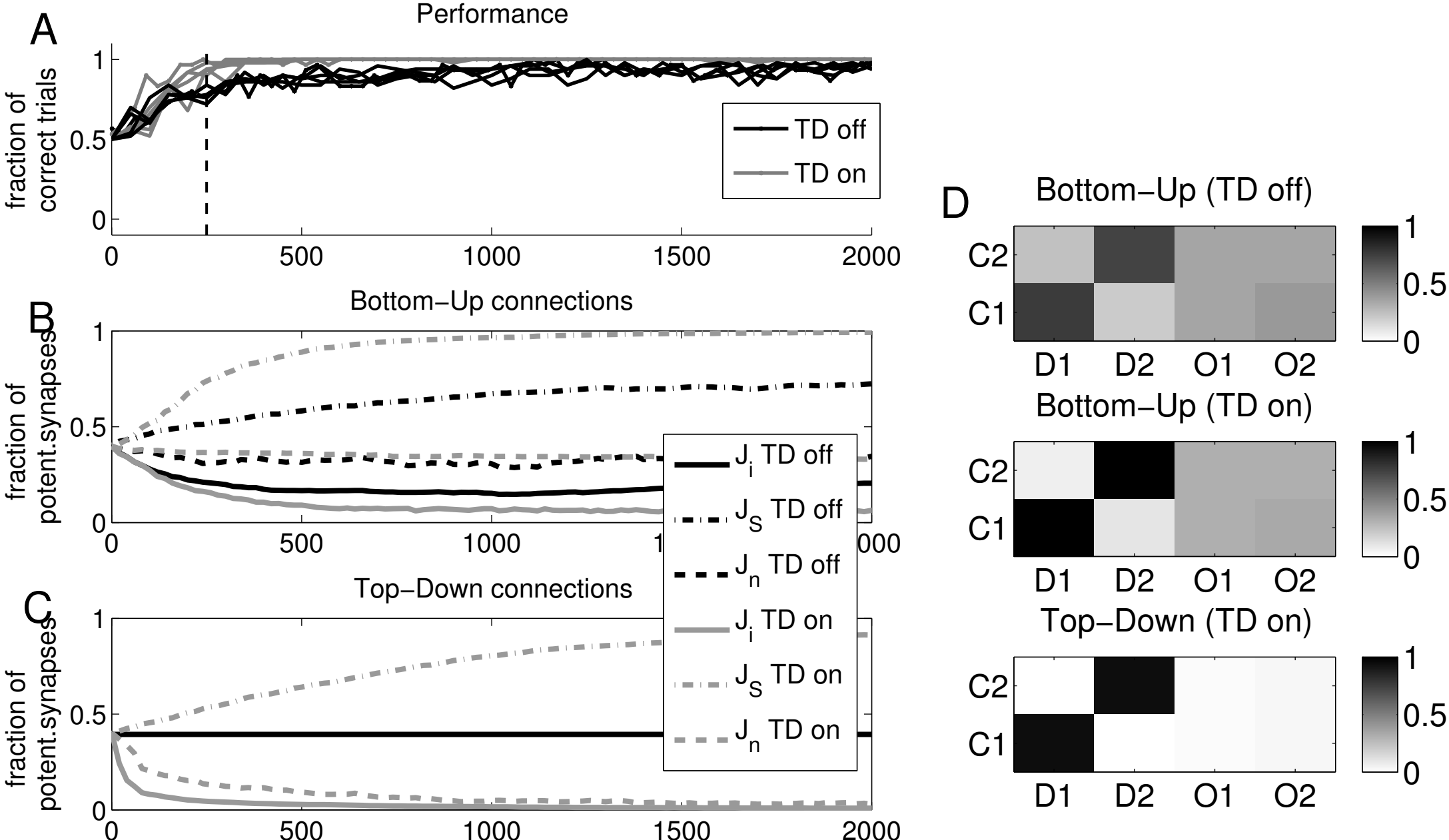

correct
wrong
LTP
LTD
no change
C1
C2
D1
D2
C1
C2
D1
D2

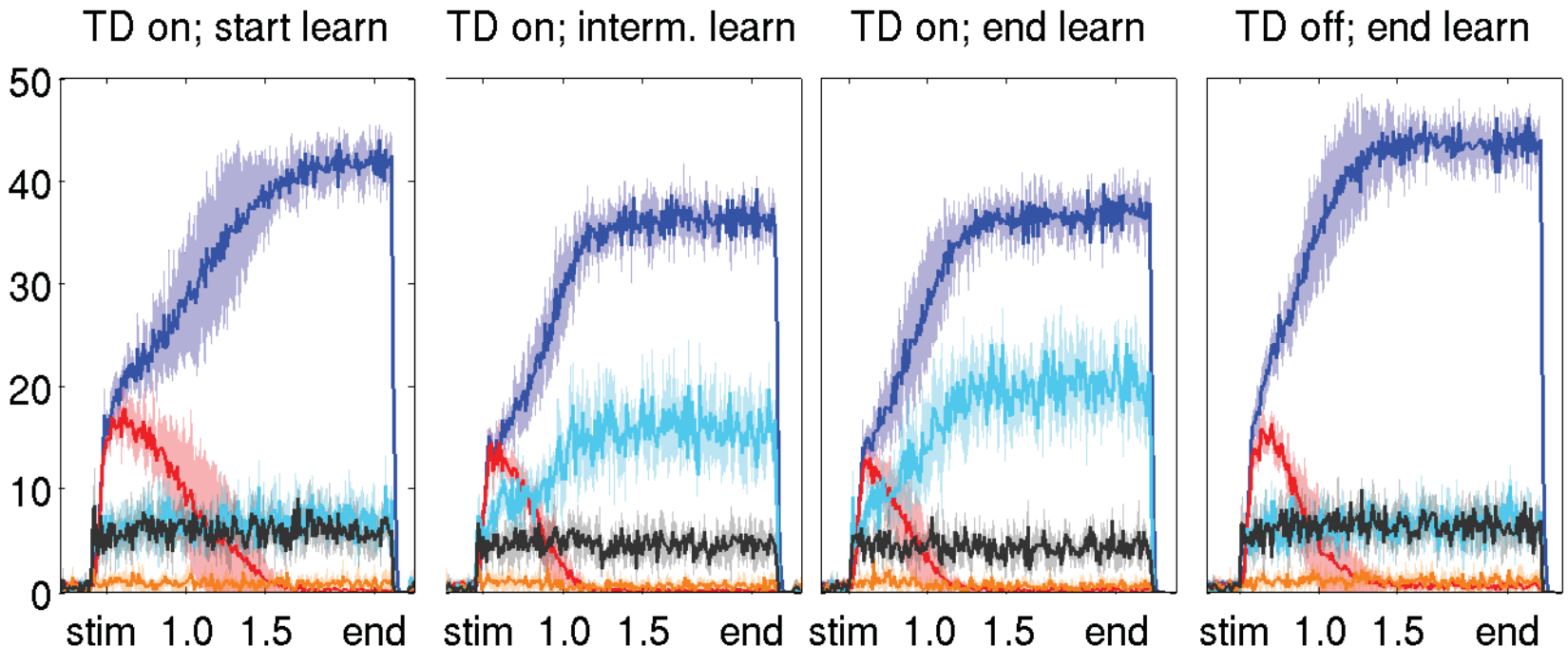

ITC
A
1
2
3

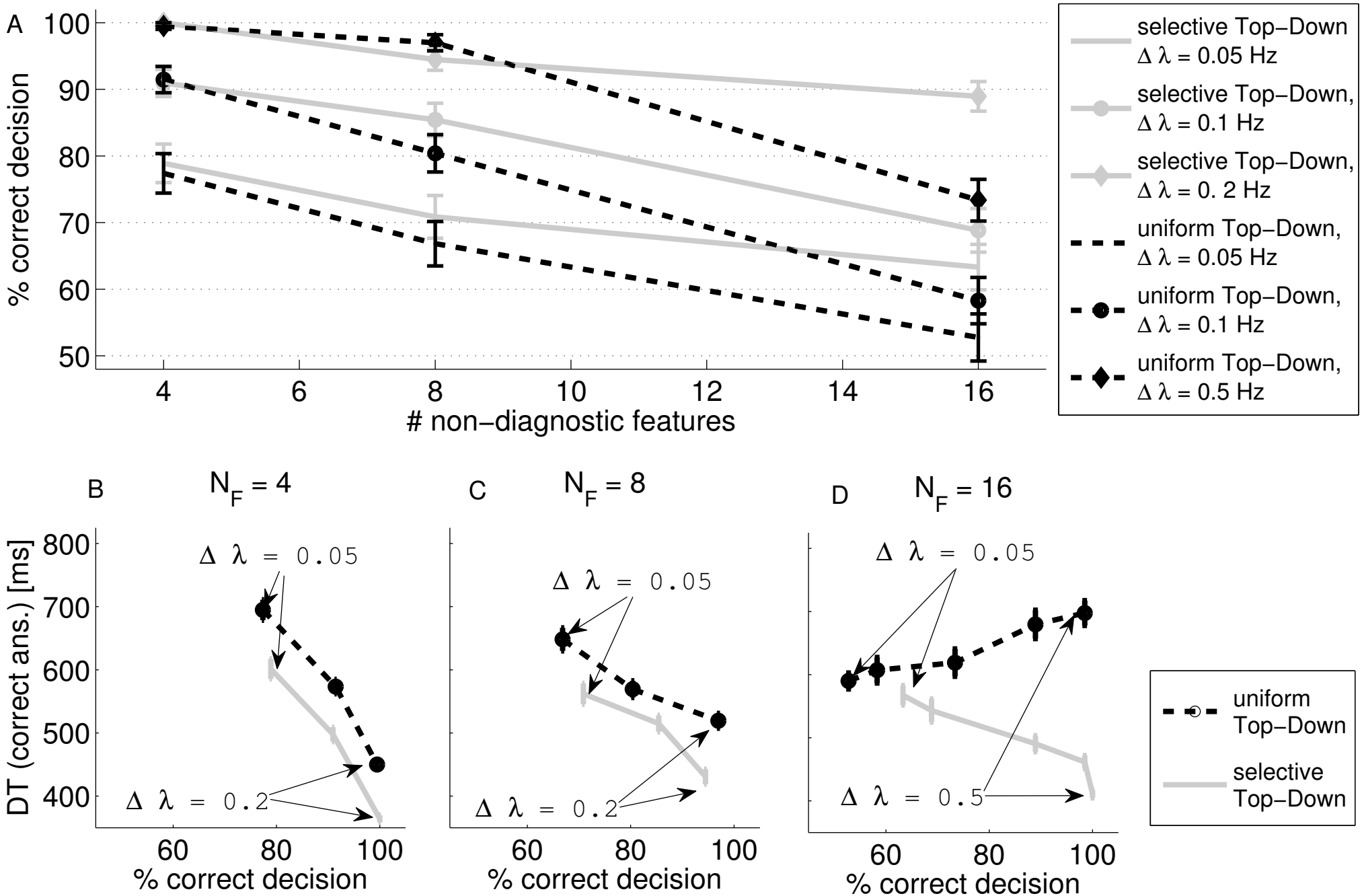

B
Diagnostic feature: eye distance (ED)
Non-Diagnostic feature:
Nose length (NL)

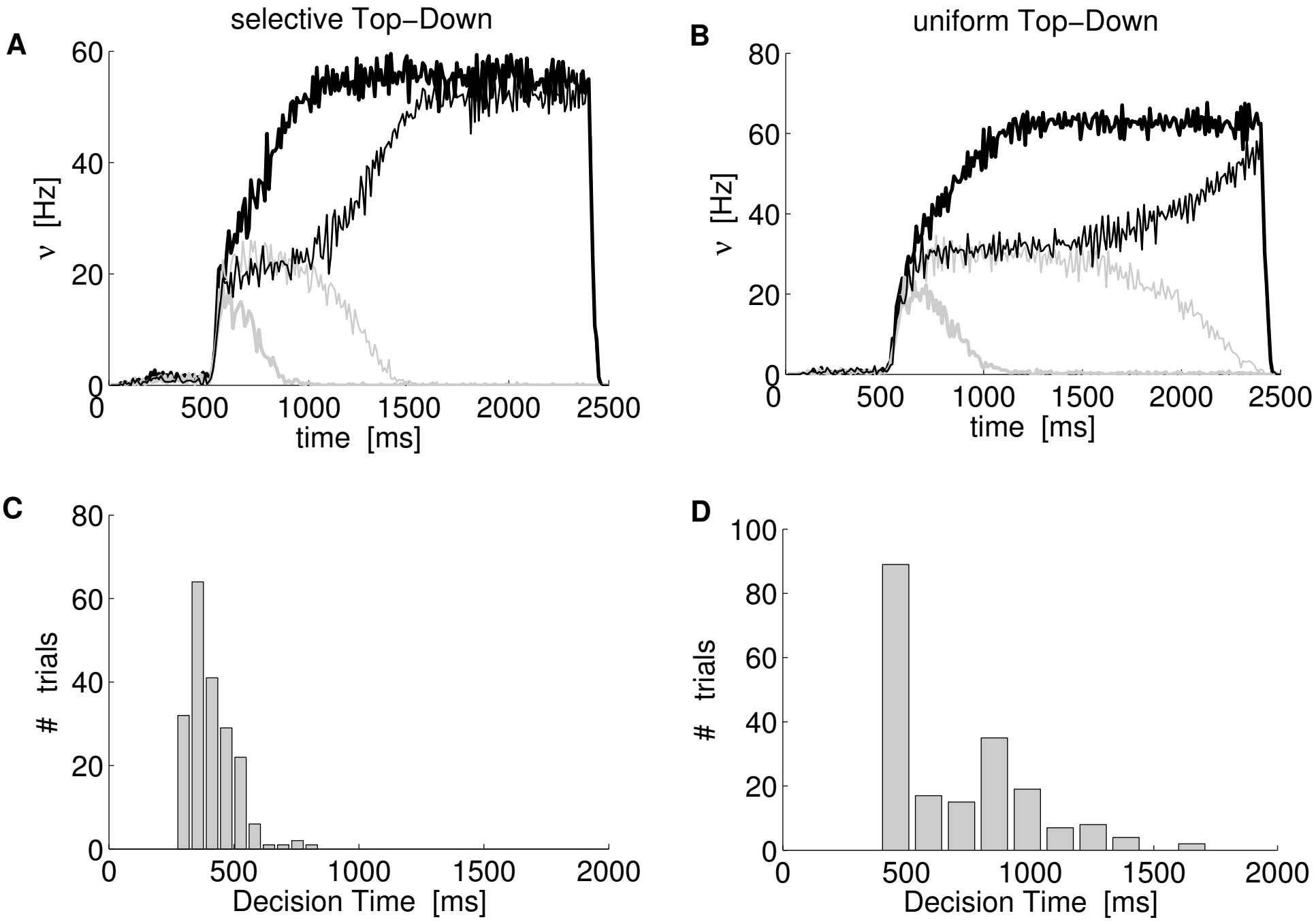

C
C1
C2
PFC
decision
layer
ED
high
ED
low
NL
high
NL
low
ITC - feature layer

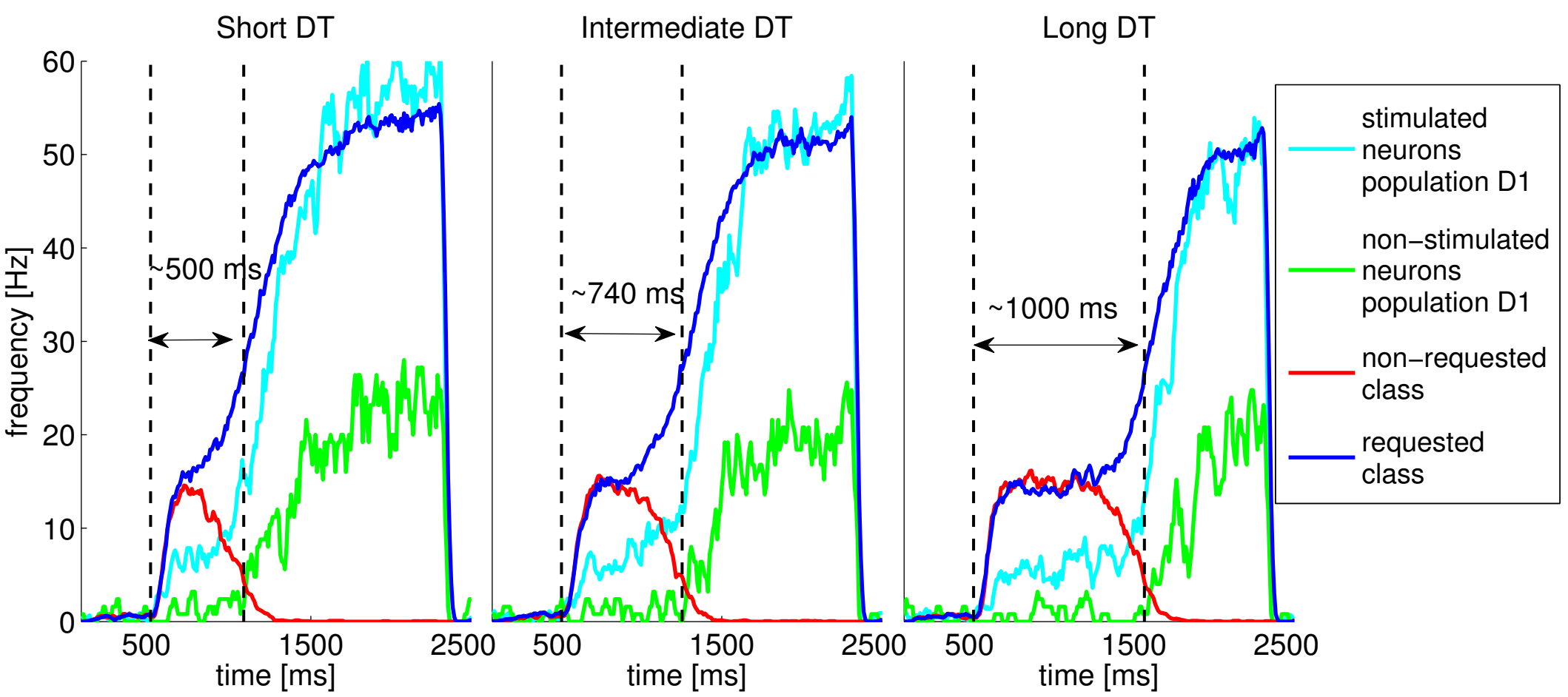

D
neurons id
time [ms]
time [ms]

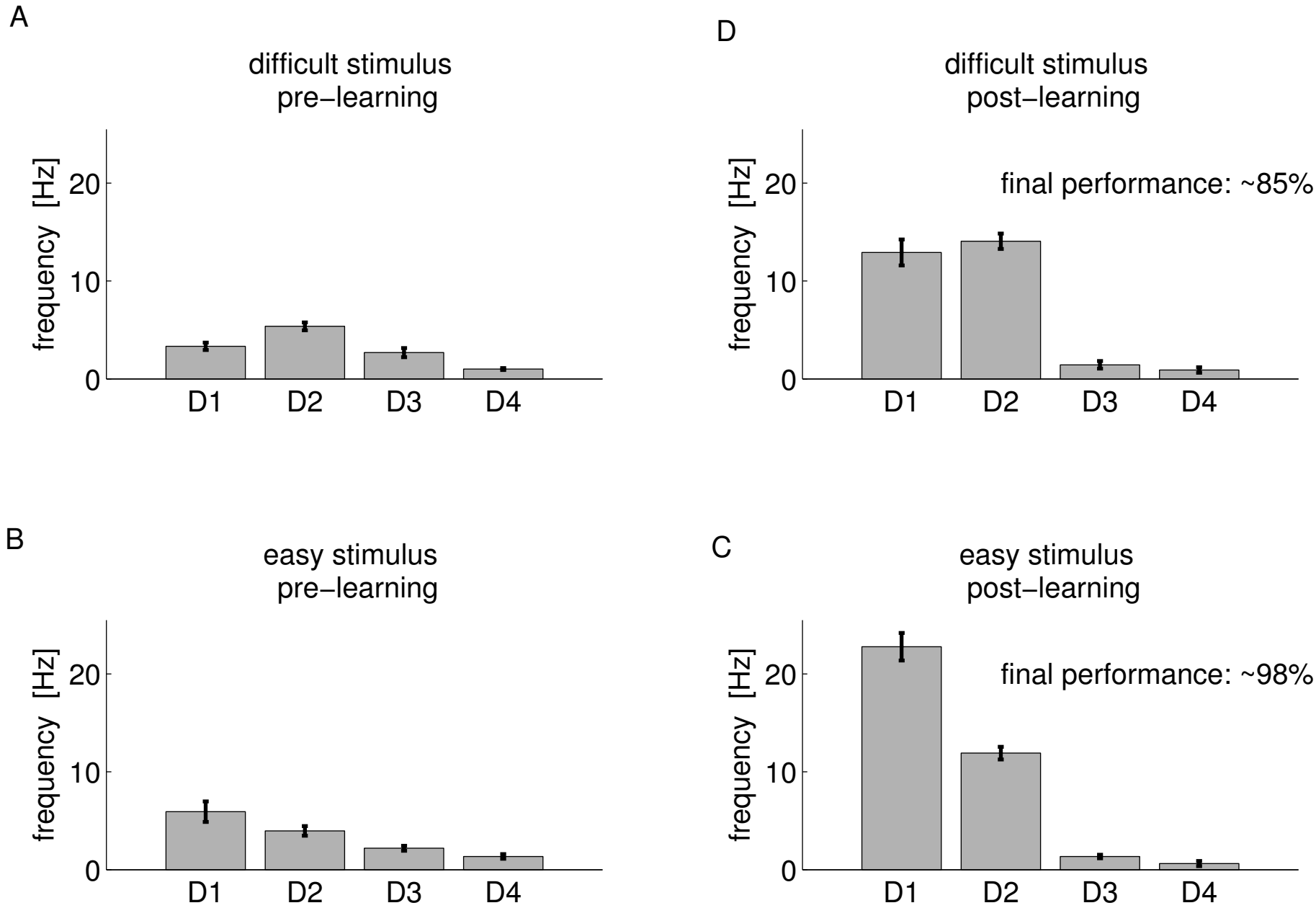

A
Performance
fraction of correct trials
TD off
TD on
B
Bottom−Up connections
fraction of potent.synapses
J_i TD off
J_S TD off
J_n TD off
J_i TD on
J_S TD on
J_n TD on
C
Top−Down connections
D
Bottom−Up (TD off)
Bottom−Up (TD on)
Top−Down (TD on)
C2
C1
D1
D2
O1
O2

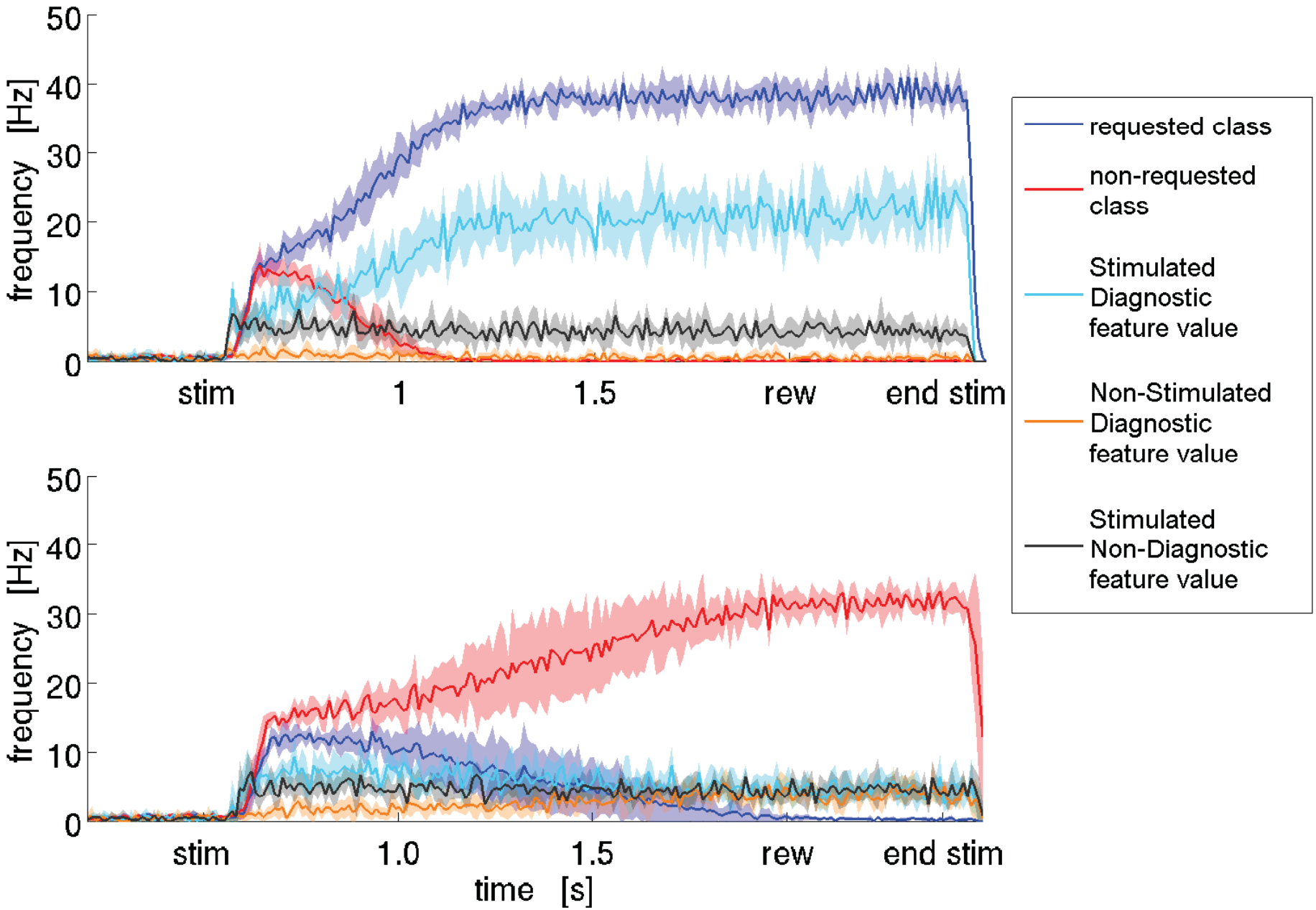

TD on; start learn
TD on; interm. learn
TD on; end learn
TD off; end learn
50
40
30
20
10
0
stim
1.0
1.5
end
requested class
non-requested class
stimulated diagnostic feature value
non-stimulated diagnostic feature value
stimulated non-diagnostic feature value


A
% correct decision
# non−diagnostic features
selective Top−Down
Δ λ = 0.05 Hz
selective Top−Down,
Δ λ = 0.1 Hz
selective Top−Down,
Δ λ = 0. 2 Hz
uniform Top−Down,
Δ λ = 0.05 Hz
uniform Top−Down,
Δ λ = 0.1 Hz
uniform Top−Down,
Δ λ = 0.5 Hz
B
C
D
DT (correct ans.) [ms]
% correct decision
Δ λ = 0.05
Δ λ = 0.2
Δ λ = 0.5
uniform
Top−Down
selective
Top−Down


A
selective Top−Down
ν [Hz]
time [ms]
B
uniform Top−Down
ν [Hz]
time [ms]
C
trials
Decision Time [ms]
D
trials
Decision Time [ms]


Short DT
Intermediate DT
Long DT
frequency [Hz]
60
50
40
30
20
10
0
~500 ms
~740 ms
~1000 ms
500
1500
2500
time [ms]
stimulated neurons population D1
non−stimulated neurons population D1
non−requested class
requested class


A
difficult stimulus
pre−learning
frequency [Hz]
0
10
20
D1
D2
D3
D4
D
difficult stimulus
post−learning
final performance: ~85%
frequency [Hz]
0
10
20
D1
D2
D3
D4
B
easy stimulus
pre−learning
frequency [Hz]
0
10
20
D1
D2
D3
D4
C
easy stimulus
post−learning
final performance: ~98%
frequency [Hz]
0
10
20
D1
D2
D3
D4


50
40
30
20
10
0
frequency [Hz]
stim
1
1.5
rew
end stim
50
40
30
20
10
0
frequency [Hz]
stim
1.0
1.5
rew
end stim
time [s]
requested class
non-requested class
Stimulated Diagnostic feature value
Non-Stimulated Diagnostic feature value
Stimulated Non-Diagnostic feature value