Inferring context-dependent computations through linear approximations of prefrontal cortex dynamics

Joana Soldado-Magraner[1,3*], Valerio Mante[2†], Maneesh Sahani[1†]

[1]The Gatsby Computational Neuroscience Unit, University College London, London, UK;
[2]Institute of Neuroinformatics, ETH Zurich-University of Zurich, Zurich, Switzerland;
[3]Present address: Neuroscience Institute, Carnegie Mellon University, Pittsburgh, PA, USA;
[†] Denotes shared senior authorship.

[*]**Correspondence:**

Dr. Joana Soldado Magraner
Neuroscience Institute
Carnegie Mellon University
Pittsburgh, PA, USA
jsoldadomagraner@cmu.edu

Abstract

The complex activity of neural populations in the Prefrontal Cortex (PFC) is a hallmark of high-order cognitive processes. How these rich cortical dynamics emerge and give rise to neural computations is largely unknown. Here, we infer models of neural population dynamics that explain how PFC circuits of monkeys may select and integrate relevant sensory inputs during context-dependent perceptual decisions. A class of models implementing linear dynamics accurately captured the rich features of the recorded PFC responses. These models fitted the neural activity nearly as well as a factorization of population responses that had the flexibility to capture non-linear temporal patterns, suggesting that linear dynamics is sufficient to recapitulate the complex PFC responses in each context. Two distinct mechanisms of input selection and integration were consistent with the PFC data. One mechanism implemented recurrent dynamics that differed between contexts, the other a subtle modulation of the inputs across contexts. The two mechanisms made different predictions about the contribution of non-normal recurrent dynamics in transiently amplifying and selectively integrating the inputs. In both mechanisms the inputs were inferred directly from the data and spanned multi-dimensional input subspaces. Input integration likewise consistently involved high-dimensional dynamics that unfolded in two distinct phases, corresponding to integration on fast and slow time-scales. Our study offers a principled framework to link the activity of neural populations to computation and to find mechanistic descriptions of neural processes that are consistent with the rich dynamics implemented by neural circuits.

Introduction

A fascinating aspect of our daily existence is that, in a blink of an eye, we can effortlessly change our course of action, switch between tasks or wander in between lines of thought. To achieve this flexibility, brain circuits must be endowed with mechanisms to perform context-dependent computations, so that behavior is quickly adapted to each situation and the correct decisions can be taken. The mechanisms underlying this flexibility are still poorly understood.

A brain structure known to mediate flexible computations is the Prefrontal Cortex (PFC)[1]. PFC is part of an extensive and highly distributed network of cortical and subcortical areas comprising the decision-making circuitry of the brain[2]. It is involved in the generation of complex behaviors such as planning, selective attention and executive control[3,4]. PFC is thought to hold the representation of goals, contexts and task rules[5,6] and in primates is required to switch behaviors according to different task instructions[7]. Finally, PFC's crucial role in ignoring task distractors suggests that it actively filters out irrelevant information[8,9]. This makes PFC of special importance for studying contextual decision-making.

Previous work suggested that flexible prefrontal computations emerge from the concerted interaction of large, interacting neural populations[1]. Surprisingly, during contextual decisions requiring monkeys to integrate noisy sensory information towards a choice, irrelevant information did not appear to be gated at the level of inputs into PFC. Instead, irrelevant inputs may be dynamically discarded through recurrent computations occurring within PFC. A possible mechanism for such dynamical gating was revealed by reverse-engineering recurrent neural networks (RNNs) trained to solve the same contextual decision-making task as the monkeys. Remarkably, the trained RNNs reproduced key features of the PFC population activity, even though the networks were not explicitly designed to match the dynamics of the data. The match with the recorded data, however, was only qualitative, as the networks failed to reproduce many aspects of the rich and heterogeneous responses of individual PFC neurons. This raises the question of whether a model that captured the complex PFC dynamics would rely on the same contextual decision-making mechanism as the RNNs.

In this study, we took the approach of fitting linear dynamical system models (LDS) directly to the PFC data, allowing us to infer interpretable linear systems that approximate the neural population activity in each context. We characterized the nature of computations implemented in each context by analysing the properties of such models, whose dynamics closely matched those of the PFC population. To validate our assumption of linear dynamics, we compared the LDS to a novel low-rank factorization of the data, Tensor Factor Regression (TFR), which can capture non-linear dynamics. Both models performed comparably,

implying that a linear model is sufficient to explain the PFC activity in a given context.

We fitted different LDS model classes corresponding to different hypotheses about the nature of context-dependent computations in PFC. One class could implement context-dependent recurrent dynamics but received fixed inputs, mimicking the design of RNNs developed in past work[1]. Another class had fixed recurrent dynamics, but could implement context-dependent inputs. In both models, we inferred external input signals directly from the data. Surprisingly, these two model classes explained the PFC responses similarly well, meaning that both contextual decision-making mechanisms are consistent with the data. Both mechanisms shared some features with the RNN solution, but also differed from it in important ways, thus revealing previously unknown properties of contextual decision-making computations in PFC.

Our data-driven approach to analyzing neural dynamics, based on fitting LDS models to neural population responses, can be applied across different brain areas, neural data sets, and computational mechanisms, providing a general tool to test specific hypothesis about the nature of computations implemented by neural circuits.

Results

We analysed PFC recordings from two monkeys performing a contextual version of the classic random dots motion task[1,10]. The monkeys had to report the overall color or motion of the dots, depending on context (Fig. 1a). Since both types of sensory evidence were simultaneously presented, the monkeys had to actively ignore the irrelevant sensory input in order to form a decision based only on the relevant input. We analysed only correct trials and focused on the random dots presentation period, during which the motion and color evidence needed for a correct decision were presented[1]. In the next sections, we present an in-depth analysis of the PFC data from one of the monkeys (monkey A). Findings from monkey F are presented in the supplementary material, and confirm the key insights gained from monkey A (Supplementary Figs. 6 to 10).

PFC dynamics in each context are well approximated by a linear system

In order to infer the mechanisms underlying PFC population dynamics, we fitted several LDS models to the PFC responses (Fig. 1b). Each LDS was parameterised by three distinct components: a dynamics matrix A, which determined the recurrent contribution to the evolution of the low-dimensional (low-d) latent activity state $\mathbf{x}(t)$; by external motion and color inputs $\mathbf{u}_m(t)$ and $\mathbf{u}_m(t)$; and by motion and color input subspaces B_m and B_m, which specified the dimensions along which the external inputs modulated the state dynamics. The dynamics matrix and the input subspaces were fixed over time, whereas the external inputs

could be time-varying. An orthonormal "loading" matrix C mapped the low-d latent activity $\mathbf{x}(t)$ to the high-dimensional (high-d) "observations" $\hat{\mathbf{y}}_{PFC}(t)$, i.e. the condition-averaged z-scored peri-stimulus-time-histograms (PSTHs) of individual units in PFC. Therefore, the observations were reconstructed based on a linear combination of low-d latent activity, which approximated the dynamics of the high-d neural population.

We captured the observed changes in activity across contexts[1] by fitting an LDS model jointly to the PFC data from the two contexts. Some of the model parameters varied across contexts, while the remainder were shared across contexts. We implemented several distinct classes of LDS, which differed with respect to their context-dependent parameters. In a first class, the dynamics matrix A^{cx} could differ across contexts (Fig. 1b, cx = mot/col context), whereas the input parameters were fixed. In a second class, the dynamics matrix was fixed, but the motion and color subspaces $B^{cx}_{m,c}$ were allowed to vary across contexts. Both model classes can process inputs flexibly, but do so based on different mechanisms.

The A^{cx}, B model class retains some of the key properties of previously proposed RNNs[1]. As in the RNNs, the motion and color inputs are fixed across contexts, meaning that any context-dependent computations must be achieved by the recurrent dynamics (Fig. 1c). The A, B^{cx} model class instead relies on contextually modulated inputs, a mechanism that appeared unlikely based on past analyses of the PFC responses[1]. Both model classes differ from the RNNs in several ways. First, in the LDS all parameters were learned from the data (Fig. 1b, grey boxes), including the external, time-varying inputs $\mathbf{u}_m(t)$ and $\mathbf{u}_m(t)$. The RNN instead was trained on the task, with hand-crafted external inputs that were constant over time. Second, the LDS could learn multi-dimensional input subspaces $B_{m,c}$. Such subspaces could capture rich activity patterns arising under direct influence of the external inputs, which may be required to explain some aspects of the data[11,12]. In the RNNs, the inputs instead were one-dimensional. Importantly, we fitted the LDS models with a regularization favoring weak inputs, to avoid solutions that relied entirely on input driven activity. Activity patterns that do not directly represent the motion and color coherence, such as the integrated relevant evidence or activity related to the passage of time, would then have to emerge through the transformation of the inputs by the recurrent dynamics in all LDS models.

Surprisingly, we found that the two LDS model classes could explain the PFC responses similarly well (Fig. 2a, A^{cx}, B and A, B^{cx}; cold color lines), implying that two very different mechanisms could explain the observed activity. A third model class that had contextual flexibility in both the recurrent dynamics and the inputs (referred to as A^{cx}, B^{cx}) did not improve the fits. A model that could change only the initial conditions across contexts (Methods), but not the recurrent dynamics or the inputs (referred to as A, B) instead performed significantly worse. We estimated the dimensionality of the latent dynamics and inputs

based on generalization performance using leave-one-condition-out cross-validation (LOOCV[13], Fig. 2b). All models performed substantially better for input dimensionality higher than 1D (Fig. 2a), meaning that they required multi-dimensional input signals. The best performing LDS models required three dimensions for both the color and motion inputs. The LDS models needed between 13 and 18 latent dimensions to best fit the data (Supplementary Table 1), many more than the 4 dimensions required to describe the task (motion, color, context, and decision).

Several lines of evidence show that the two best LDS model classes provide accurate descriptions of the PFC responses. First, the two LDS models closely approximated the highly heterogeneous responses of individual PFC neurons (Extended Data Fig. 1). The fits captured a substantial fraction of the data variance (27%, corresponding to MSE=0.73 on z-scored responses, Fig. 2b) even though we did not smooth nor "de-noise" the data[1]. We included all neurons in the fits, even those with weak, sparse responses that could only be poorly captured by the models (Extended Data Fig. 1a,b, firing rates < 1Hz). Furthermore, we did not optimize our models to match the noise statistics of the data, since the fitted responses were trial-averaged.

Second, the best LDS model classes performed comparably to a more powerful model class that we refer to as Tensor Factor Regression (TFR). TFR is based on a new low-rank factorization of the data that partitions the data tensor into several low-d tensors, including a core tensor and an input tensor (Fig. 2c). The core tensor can be learned with independent parameters at each point in time, which provides TFR with a greater flexibility to capture temporal patterns compared to the LDS models, which are constrained to generate linear dynamics. The LDS classes are nested within the TFR model class (Methods), which places the two types of models on an equal footing and avoids potential pitfalls that can arise when performing model comparison across different model classes[14].

The observation that the LDS and TFR models performed comparably implies that the additional flexibility of the TFR model was not necessary to explain the data. PFC activity in each context thus appears to be well approximated by linear dynamical system models. The TFR model incorporated input parameters and contextual constraints equivalent to those of the LDS (Fig. 2c, Methods) and achieved comparable performance for a wide range of input and latent dimensionalities (Fig. 2a,b, warm color lines). The TFR model required a similar range of latent dimensionality than the LDS (13-18, Supplementary Table 2). For the TFR model, however, the optimal inputs were 2-dimensional, compared to the 3-dimensional inputs required by the LDS. This difference could imply that the LDS needs an extra input dimension to overcome the limitations of the linear dynamical constraints, a possibility that should be taken into account when

interpreting the parameters of the LDS model.

Third, the best LDS models qualitatively captured salient features of the population dynamics equally well. In particular, both the A, B^{cx} and the A^{cx}, B models reproduced the rich PFC dynamics revealed by population trajectories in the low-d activity subspace capturing most variance due to motion, color and choice[1] (Fig. 3). The TFR model resulted in comparable fits, both at the population (Extended Data Fig. 2a) and single neuron level (Extended Data Fig. 1).

The good qualitative match between the low-d PFC trajectories and the fits of the A, B^{cx} model are somewhat surprising. By design, the task-related activity subspace capturing most variance due to motion, color, and choice can distinguish between computations that rely on inputs that are stable across contexts from computations implementing a strong suppression of the irrelevant input[1]. Computations relying on stable inputs result in characteristic features in the trajectories, in particular how they depend on the strength of the motion and color inputs (Fig. 3a). These features approximately occur in the measured PFC trajectories (Fig. 3b), and indeed are reproduced by the A^{cx}, B model (Fig. 3d). However, the same features are also reproduced by the A, B^{cx} model (Fig. 3c), which by construction must rely on inputs that are variable across contexts.

One conclusion from these analyses is that the properties of population trajectories in the considered low-d activity subspace are not sufficient to rule out computations that rely on variable inputs across contexts. The very close similarity between trajectories for the two LDS models suggests that the strength of input modulation required to explain the PFC responses is likely small. To understand how such small input modulations may support context-dependent integration in PFC, below we first separately characterize the inputs and recurrent dynamics in the A, B^{cx} and the A^{cx}, B models, and then ask how their combined effects can account for contextual integration in PFC.

Input signals span curved manifolds and are largely stable across contexts

To better understand the mechanism of contextual integration implemented by the two LDS model classes, we fitted 100 models for each class, with random initialization and with latent and input dimensionality set by cross-validation (Fig. 2a, 3D inputs and latent dims 18 and 16). As we found similar model parameters across random initializations, we report parameter averages across the 100 models.

The context-dependent nature of computations in these models can be easily appreciated by considering the time-dependent norm of the latent activity, $||\mathbf{x}_m^{cx}(t)||$ and $||\mathbf{x}_c^{cx}(t)||$. For each context, we computed this norm

for activity in response to only one of the inferred inputs, with the other input set to zero (Fig. 4a). Here we refer to the resulting activity as the "output" of the LDS. In both models, the norm of the output activity increases over time within the trial and is much larger for the relevant compared to the irrelevant input, reflecting context-dependent integration (Fig. 4b,c, bottom panels, thick dotted vs. thin lines; significance, Wilcoxon rank-sum test, p<0.001 across 100 models, green bars; with $91 \pm 2\%$ increase for mot, $95 \pm 3\%$ for col in the A, B^{cx} model, and $94 \pm 3\%$ increase for mot, $93 \pm 3\%$ for col in the A^{cx}, B model, mean $\pm$ std across 100 models). The predicted norms are essentially identical across the two models, in agreement with the similarity of the low-d trajectories they produce (Fig. 3c,d). In the A^{cx}, B model this context dependence is entirely due to differences in the recurrent dynamics across contexts. In the A, B^{cx} model, the difference must entirely reflect contextual modulation of the overall input strength, of the input direction, or a combination of the two.

We first considered the time and context-dependence of the input strength, which we defined as $||B_m^{cx}\mathbf{u}_m(t)||$ and $||B_c^{cx}\mathbf{u}_c(t)||$ (motion and color strengths, respectively). Input strength is subtly different across the two models (Fig. 4b,c, top panels). After their initial rise at stimulus onset, the inputs were approximately sustained in the A, B^{cx} model, but somewhat transient in the A^{cx}, B model. The strength of the motion and color inputs was similar, but both were overall weaker in the A^{cx}, B compared to the A, B^{cx} model. The finding that the inputs differ across models, whereas the output norm does not, implies that the recurrent dynamics must also operate differently in the two models. In the A^{cx}, B model, input strength was the same across contexts, by definition. In the A, B^{cx} model, input strength was different across contexts (Fig. 4b, top panels, thick dotted vs. thin lines, Wilcoxon rank-sum test, p<0.001, green bars), but only modestly, with the irrelevant inputs being slightly weaker than the relevant ones ($38 \pm 14\%$ decrease for mot, $22 \pm 8\%$ for col, averaged across all time points starting at t=200ms, mean $\pm$ std across 100 models).

Next, we considered the input direction, by characterizing the structure of coherence representations within the inferred input subspaces. Even though our cross-validation procedure consistently inferred 3-dimensional input subspaces, we found that most of the inferred input variance was contained in a 2D plane (Extended Data Fig. 3a). This plane was spanned by dimensions that captured, respectively, variance related to input coherence (mot, col) and coherence magnitude (|mot|, |col|). As a result, input coherence was represented along a curved 1D-manifold within the input plane[15]. Similar curved representations were found by the two models (Fig. 4d,e) and for both contexts in the A, B^{cx} model (Fig. 4d). Analogous 2D coherence representations were present also in the PFC data (Extended Data Fig. 3b), whereas the 3rd input dimension contained very little input and data variance (Extended Data Fig. 4c,f). In fact, the LDS models with 2D inputs were very close in performance to the 3D models (Fig. 2a) and captured the population trajectories

nearly as well (Extended Data Fig. 2c,d). 1D input models, on the contrary, performed worse (Fig. 2a) and did not capture the trajectories along the input dimensions as well (Extended Data Fig. 2e,f).

The 2D input planes identified across the two model classes were highly aligned (16-31°, for averaged dimensions across 100 models from each model class, Extended Data Fig. 3c), an effect not expected by chance (Extended Data Fig. 3d). In the A, B^{cx} model, the motion and color planes varied across contexts, but only modestly ($33° \pm 10$ mot, $46° \pm 16$ |mot|, $25° \pm 5$ col, $27° \pm 9$ |col| dims, mean $\pm$ std across 100 models, Fig. 4d cartoons, Extended Data Fig. 3c) and less than expected by chance (Extended Data Fig. 3d). These relatively small changes in input direction across contexts, together with the concurrent, modest change in input strength (Fig. 4b, top), are entirely responsible for the context-dependent changes in the output of the A, B^{cx} model (Fig. 4b, bottom).

In both models, the time-course (Fig. 4b,c, top) and structure of the inputs (Fig. 4d,e) is thus relatively simple, with most of the variance captured by a 2D-input subspace. This finding alleviates a possible confound inherent in fitting an LDS with time-dependent inputs. In principle, the fitted inputs could be very rich, and effectively approximate on their own the dynamics of a very complex, non-linear dynamical system. Considering that we retrieved inputs that are of much lower dimensionality than the recurrent dynamics (3D vs. 16-18D), such a scenario appears unlikely. Indeed, we find that the same classes of LDS models can be refit to the data with only a small drop in performance when their inputs are constrained to be fixed over time (Extended Data Fig. 5a, Supplementary Note 1). The observed complexity of PFC responses thus need not be inherited from the external inputs, but rather can be explained as resulting from approximately linear, time-dependent recurrent dynamics.

Input integration relies on high-dimensional linear dynamics

To reveal the contributions of the recurrent dynamics to selective integration, we follow an approach motivated by the original analysis of RNNs trained to solve the same task as the monkeys[1]. The computation implemented by the RNNs can be fully understood at the level of local linear approximations of the dynamics (Fig. 5a, left panel). Specifically, selective integration reflects four key features of the linear approximations. First, the largest eigenvalue of the dynamics is close to one, with the rest of the eigenvalues being much smaller than one, implying that input integration is implemented as movement along a line-attractor[16]. Second, inputs are selected for integration based on their alignment with the leading left-eigenvector of the dynamics (the "input-mode" associated with the largest eigenvalue, i.e. slowest dynamics). The direction of this left-eigenvector is context-dependent, such that it is orthogonal to the contextually irrelevant input direction. Third, the direction of the leading right-eigenvector of the dynamics (the "output-mode" associated

with the largest eigenvalue), which determines the direction of the line-attractor, is fixed across contexts. Fourth, the dynamics is non-normal, as the leading right and left eigenvectors have different directions. We collectively refer to these features of the linear approximations as the "RNN-mechanism" and below compare them to the dynamics of the fitted models.

We computed the eigenspectrum of the dynamics matrices, and in both models found multiple slow dimensions associated with time-constants that were relatively long compared to the duration of a trial (eigenvalues with norm $|\lambda| > 0.8$, or decay time-constant $\tau > 224$ms, while the trial lasts 750 ms, Methods, Fig. 5b). The A, B^{cx} model had a larger fraction of slow modes than the A^{cx}, B model ($55 \pm 7\%$ vs. $35 \pm 8\%$ mot cx/ $41 \pm 8\%$ col cx, mean$\pm$std across 100 models). Many of the eigenvalues were imaginary, implying that the inferred recurrent dynamics was rotational[17] (Extended Data Fig. 6a,b). The largest eigenvalue was 0.98 ± 0.02 for the A, B^{cx} model and 0.96 ± 0.03 / 0.99 ± 0.03 (mot/col cx) for the A^{cx}, B model (mean norm$\pm$std across 100 models), corresponding to average decay time-constants of 2.5 and 1.2 / 5 seconds. While not strictly compatible with a line-attractor (an eigenvalue of 1), these time-constants are much longer than the duration of the trial. Additionally, the LDS models implemented many other slow modes, some with time-constants comparable to the duration of the trial (Fig. 5b). The large number of slow modes provides a first indication that PFC dynamics may be higher-dimensional than would be predicted by the RNN-mechanism.

We assessed context-dependent relations between the recurrent dynamics and the inputs by focusing on the coherence component of each input, while ignoring the absolute-coherence component (Fig. 4d,e). In the considered linear models, only the coherence component of the input can contribute to choice-dependent responses. We first examined the "load" (the non-normalized projection, Methods) of the coherence input onto each left-eigenvector, computed separately at each time instant and then averaged over the entire trial (Fig. 5c). Consistent with the RNN mechanism, the input load onto the left-eigenvectors was overall larger for the relevant vs. the irrelevant input, in both types of models (Fig. 5c, green bars, Wilcoxon rank-sum test, $p<0.05$). Furthermore, the load was close to zero for the irrelevant input along the slowest mode, implying an orthogonal arrangement of the irrelevant input directions with respect to the slowest left eigenvectors, as in the RNN mechanism. However, this difference in load across contexts resulted from very different mechanisms in the two models: in the A, B^{cx} model it entirely reflects changes in the inputs, whereas in the A^{cx}, B model it entirely reflects changes in the recurrent dynamics. Unlike in the RNN mechanism, in both models the coherence input does not preferentially load onto the left-eigenvectors associated with the largest eigenvalues ($|\lambda| > 0.9$, time-constant $\tau > 475$ms). Rather, the largest loads are consistently obtained

for eigenvectors with intermediate eigenvalues ($|\lambda| = 0.7 - 0.8$, $\tau = 140 - 224$ms), and thus relatively fast decay time-constants (Fig. 5b,c).

Non-normal dynamics makes model-specific contributions to selective integration

The qualitative similarities in the eigenspectra (Fig. 5b) and the input loads (Fig. 5c) of the A, B^{cx} and A^{cx}, B models masks a key difference in the recurrent dynamics implemented by these models. Specifically, we find that the two models implement dynamics with very different degrees of non-normality. We assess the strength of non-normality through one of its possible consequences, namely the transient amplification of perturbations of the activity[18,19] (Supplementary Note 2, Extended Data Fig. 7). We simulated the effect of injecting a short pulse of activity at trial onset, along random state-space directions. For the A, B^{cx} model, these perturbations gradually decay over the course of the trial (Fig. 6a, top, dashed lines, average across pulses in random directions). For the A^{cx}, B model, on the other hand, activity following a pulse is transiently amplified, i.e. the gradual decay is preceded by a transient increase in activity (Fig. 6a, bottom, dashed lines). When perturbations are applied selectively along the left-eigenvectors, transient amplification is even more pronounced in the A^{cx}, B model, but still largely absent in the A, B^{cx} model (Fig. 6a, doted lines). Dynamics is thus strongly non-normal in the A^{cx}, B model, as in the RNN mechanism, but less so in the A, B^{cx} model (Fig. 6c).

When dynamics is non-normal, the activity in response to a short pulse of input can reflect the combined effects of two processes operating at different time-scales: first, the transient, non-normal amplification of the input pulse; and second, its long-term integration towards a choice. The relative contributions from these processes differ across the two models, due to the different degree of non-normality of the underlying dynamics. In the A, B^{cx} model, coherence input pulses are not transiently amplified, but rather immediately decay, whether they are relevant or not (Fig. 6b, top, thick doted and thin lines). In the A^{cx}, B model, the relevant input is transiently "persistent", due to non-normal dynamics (Supplementary Note 2), whereas the irrelevant input quickly decays (bottom). Also at longer time-scales, the decay of a relevant input pulse is faster in the A, B^{cx} model compared to the A^{cx}, B model, indicating less accurate input integration. At both fast and slow time-scales, the recurrent dynamics of the A, B^{cx} model thus cannot sustain information provided along relevant input dimensions as well as the dynamics in the the A, B^{cx} model (Fig. 6b, top vs. bottom thick doted lines). As a consequence, the A, B^{cx} model must instead rely on inputs that do not decay towards the end of the trial (Fig. 4b, top) to explain the observed persistent activity in PFC, whereas the A^{cx}, B infers inputs that are more transient (Fig. 4c, top).

The pulse responses in Fig. 6b also illustrate that changes in the direction of the inputs across contexts are not sufficient to explain the differences in the representation of the relevant and irrelevant inputs in PFC. The employed pulses have unitary strength, and thus isolate the contribution of the input direction to the responses. The change in input direction across contexts in the A, B^{cx} model lead to only relatively small differences in the amount of integration between the relevant and irrelevant inputs (compare late responses). To explain context-dependent integration in the PFC data, the A, B^{cx} model in addition must thus rely on contextual modulation of the strength of the inputs throughout the duration of the trial (Fig. 4b, top; relevant input is stronger than irrelevant input).

The features of the dynamics considered so far imply that the two LDS models implemented mechanisms of selection and integration that share some key properties of the RNN mechanism. Like the RNN, all LDS models ultimately relied on a context-dependent realignment of the inputs and a subset of the modes of the recurrent dynamics, either through a change of the inputs (A, B^{cx}) or of the recurrent dynamics (A^{cx}, B). Like the RNN mechanism, the A^{cx}, B model (but not the A, B^{cx} model) implements strongly non-normal recurrent dynamics. However, while the RNN mechanism relies on a single or few slow modes that are well aligned with the relevant input (an approximate "line attractor"[1]), both LDS models instead implemented a large number of modes with different degrees of persistence, whereby the the inputs are not preferentially aligned with the slowest modes.

Input integration occurs in two distinct phases

The above analyses provide insights into two mechanisms of context-dependent selection and integration that can account for population dynamics in PFC. However, these analyses alone do not explain how the neural trajectories predicted by the models emerge from the interaction of the inputs and the recurrent dynamics. Such an explanation must include also the properties of the right eigenvectors of the dynamics matrix, which amount to the "output" dimensions of the LDS models. The right eigenvectors influence both "where" in activity space the inputs are mapped onto and "how" they are transformed over time.

To establish how the trajectories emerge from the two LDS mechanisms, here we separately consider condition-dependent (CD) and condition-independent (CI) components of the neural trajectories. CD components were the primary focus of past accounts of this data[1] and, particularly late in the trial, primarily capture choice-related activity. CI components, on the other hand, capture prominent structure in the neural responses that is related to the passage of time during a trial, and is common to all conditions and choices. To identify the modes of the dynamics that mostly account for CD or CI variance at a particular time in the trial, we computed the alignment between the right eigenvectors of the dynamics and the dimensions

capturing most CD and CI variance at different times in the trial (Fig. 7, for the A^{cx}, B model in the motion context, Extended Data Fig. 8, for all models and contexts). Only right eigenvectors that are well aligned with a given CD or CI dimension can contribute to the responses variance along that dimension.

The alignment between CD dimensions and right eigenvectors suggests that input integration occurs in two distinct phases, each dominated by distinct dynamics. Early in the trial, the CD responses occur primarily along right-eigenvectors corresponding to modes implementing relatively fast decay and fast rotations ($|\lambda| = 0.7 - 0.8$, decay time constant $\tau = 140 - 224$ms, rotation frequency $f > 1$Hz, Fig. 7a,b, yellow lines). Late in the trial, the CD responses instead occur along right-eigenvectors with very slow decay and weak or no rotations ($|\lambda| > 0.9$, $\tau > 475$ms, $f < 0.25$Hz, red lines). This transition was consistently observed across model classes (A, B^{cx} and A^{cx}, B), contexts and model initializations (Extended Data Fig. 8). The differences in decay constants and rotational frequencies of the best aligned modes early vs. late in the trial are highly significant (Fig. 7b, Extended Data Fig. 8b, Wilcoxon rank-sum test, p<0.001). These observations imply that the relevant input is initially integrated along multiple decaying and rotational modes, consistent with the fact that the relevant inputs are strongly loaded onto left eigenvectors with intermediate eigenvalues (Fig. 5c). Later in the trial, the input is further integrated and maintained along at set of different, persistent and non-rotational modes.

The CI variance in the responses, on the other hand, appears to be mediated by largely different dynamic modes compared to the CD variance (Fig. 7c,d, Extended Data Fig. 8c,d). Unlike for the CD variance, the alignment between the leading CI direction and the right-eigenvectors is largely preserved across the trial. At all times in the trial, the leading CI variance occurs along directions associated with modes decaying more slowly than the early CD-aligned modes, but more quickly than the late CD-aligned modes ($|\lambda| = 0.8 - 0.9$, $\tau = 224 - 475$ms). Likewise, these CI directions are associated with rotational frequencies that are smaller than those in early CD-aligned modes, but faster than late CD-aligned modes ($f = 0.25 - 1$Hz).

Overall, the inferred modes of the dynamics can thus be grouped into three distinct, non-overlapping sets, accounting for different components in the trajectories. The first and second set of modes account for early and late choice-related activity, while the third set accounts for choice-independent activity. The existence of these three different components of the PFC responses presumably explains why both models infer dynamics that is relatively high-dimensional and involving many modes associated with relatively slow decay.

To further validate the existence of multiple phases of the dynamics, we examined the activity trajectories along directions aligned with the relevant CD and CI components. Specifically, we defined an early and a

late CD direction, which are primarily aligned with the first and second set of dynamics modes, respectively (yellow and red lines in Fig. 7a), and a single CI direction, which is primarily aligned with the third set of modes (green line in Fig. 7a), and then averaged these across contexts. To simplify the comparison with trajectories in Fig. 3, we projected the trajectories into two-dimensional subspaces spanned by the late CD direction and one of the other task-related directions.

We find that subspaces that include the early CD direction and the CI direction reveal prominent features in the population trajectories that are not apparent in other subspaces (Fig. 8a), confirming their potential importance in explaining the observed dynamics. The late CD direction closely matched the choice axis identified by Mante et al. (average angular difference of 18° across contexts, much more than expected by chance, Extended Data Fig. 6b) and captured a steady build-up of decision signals in both contexts over time[1] (Fig. 8b, top panel, red dimension, dec). The early CD direction, on the other hand, captures an additional component of choice-related activity, which emerges early in the trial, but later decays (Fig. 8b, top panel, yellow dimension, dec 2). This decay is consistent with the above observation that early CD-directions are aligned with relatively fast decaying modes (Fig. 7b,c).

Together, the two CD directions thus capture decision-related activity evolving on different time-scales, whereby one component is transient and the other persistent. Notably, the projections along the early CD direction differ from those along the input directions (Fig. 8a,b, middle panels, black and blue mot and col dimensions, here the LDS-identified coherence input dimensions, averaged across models and contexts). Indeed, while activity along a given input dimension reflects the sign of a single input regardless of context, activity along the early CD dimension instead only reflects the sign of the contextually relevant input, and is not modulated by the irrelevant input (Fig. 8a,b, middle vs. top panels). Projections onto the CI direction likewise reveal additional components of the responses that are common to both choices (Fig. 8a,b, bottom panels). Finally, additional input-related dimensions can be defined based on the LDS fit, by considering variance due to coherence magnitude (|col| and |mot| in Fig. 4d,e, Extended Data Fig. 9a,b). All the inferred dimensions explained substantial fractions of the data variance (1-9%, Extended Data Fig. 9c) that are comparable to those captured by previously found task-related dimensions[1] (Extended Data Fig. 9d).

Overall, these low-dimensional projections of the activity support the existence of the two phases of integration inferred from the analysis of the right eigenvectors. Moreover, these projections illustrate how the fits of the LDS models can be used to define a novel set of dimensions that appear to isolate the meaningful components of the computations implemented by the neural population.

Trained RNNs do not capture all features of the PFC data

The properties of the inputs and dynamics inferred by the two LDS classes appear to differ in several ways from those expected by a simple line attractor mechanism, of the kind previously shown to be implemented by RNNs trained to solve the contextual integration task. In particular, both LDS models implement a decision process that unfolds in two phases (early vs. late choice dimensions), rely on the contextual modulation of a large number of dynamics modes across a wide range of decay constants, and infer multi-dimensional inputs modulated both by signed and unsigned input coherence. However, it is not immediately clear that these features reflect meaningful differences between the mechanisms implemented by the LDS models and the RNN. While the RNN are non-linear, the LDS are linear, meaning that some LDS features may simply reflect somewhat trivial consequences of approximating non-linear dynamics with a linear system.

To evaluate this possibility, we repeated all the analyses we performed on the PFC responses also on simulated responses of a trained RNN, and then directly compared the two (Supplementary Figs. 1 to 5). This comparison shows that the features of the PFC responses highlighted above are not captured by the trained RNN. First, contextual modulation of the dynamics in the RNN is most pronounced in modes that are persistent or slowly decaying (Supplementary Fig. 3b), whereas in PFC it is strongest in relatively quickly decaying modes (Fig. 5c). In the RNN, as in PFC, the inferred slow dynamics is not limited to a single mode, unlike in a perfect line attractor (Supplementary Fig. 3a and Fig. 5b). This observation is expected, as the RNN tend to implement integration along a one-dimensional manifold that is curved, rather than perfectly straight, and thus cannot be approximated by a single linear mode. Second, the LDS fits do not provide any evidence of multiple phases of input integration in the RNN (Supplementary Fig. 4), whereas they reveal distinct early and late phases in the PFC responses (Extended Data Fig. 8). Finally, the LDS fits of the RNN responses learn inputs that are largely one-dimensional, whereby input signals are modulated by absolute input strength only weakly or not at all (Supplementary Fig. 2c-e). This finding contrasts with the robust encoding of absolute input strength in PFC (Fig. 4d,e).

The analyses of the RNN responses also reiterate the challenges in establishing which of the two mechanisms implemented by the LDS models is more likely to be implemented by PFC. Indeed, as for the PFC data, also the RNN data can be well fit by both model classes (Supplementary Fig. 1), even though arguably only one of the two classes matches the RNN in how it selects and integrates the sensory inputs. In the trained RNN, the inputs are not modulated by context, and input-related responses in the network are likewise largely constant across contexts. This property matches the design of the A^{cx}, B model, but not the A, B^{cx} models, and yet both models fit the RNN data equally well. However, the A, B^{cx} fits of the RNN responses do display some idiosyncratic properties suggestive of a very precise fine-tuning of parameters. In particular, the A, B^{cx}

model required many more dimensions than the A^{cx}, B model to fit the data (Supplementary Table 5) and presented extreme levels of amplification for dimensions other than the input dimensions (Supplementary Fig. 3c-e). Such a fine-tuning may reflect the mismatch between the underlying mechanisms of integration. Notably, we did not find such evidence of precise fine-tuning in the fits of the PFC responses, meaning that also in this respect both model classes are equally valid descriptions of the PFC data.

Discussion

The complex and highly heterogeneous activity patterns observed in prefrontal areas are thought to reflect the specific computations implemented in these regions[20]. In this study, we inferred candidate mechanisms for one of such computation, contextual decision-making, directly from PFC responses. By fitting several LDS models to the PFC data, we inferred interpretable dynamics that linearly approximate the neural activity in each context. We found that two distinct mechanism of contextual-integration were consistent with the PFC activity: a switch in recurrent dynamics and a modulation of inputs. The key features of these mechanisms were consistently found across the motion and color inputs in monkey A, and the motion input in monkey F (Supplementary Figs. 6 to 10). As previously reported[12], representations of color inputs were instead weak or absent in monkey F.

The first LDS mechanism is broadly consistent with past accounts of PFC responses in this task[1], in that the input selection relies on non-normal, context-dependent recurrent dynamics. In addition to this role in inputs selection, our analysis revealed that non-normal dynamics might additionally result in the transient amplification of relevant inputs in PFC. Non-normal transient amplification had previously been proposed to play a role in the processing of external inputs[21–23], as well as computations as varied as maintaining inputs in working-memory[24], generating the transient neural activations required for generating movements[25] and mediating robustness to perturbations[26]. Our observation of two distinct stages in PFC dynamics during decision formation is evocative of a recently proposed mechanisms relying on transient amplification to optimally load information onto an attractor[22]. In contrast to the predictions of such optimal loading, however, we found that the inputs where not preferentially aligned with the most amplifying dimensions of the dynamics (Extended Data Fig. 6e-g).

The second LDS mechanism relies on modulation of the inputs, and could be implemented via top-down influences on sensory areas. Our LDS fits reveal how strong such top-down modulation would have to be to explain context-dependent responses in PFC. The inferred modulation strengths ($38 \pm 14\%$ mot, $22 \pm 8\%$ col, Fig. 4b) are in the range of some attentional effects observed in sensory areas[27,28], although other studies have

reported substantially weaker or stronger feature-based modulation[29–33], potentially reflecting differences in task design, cognitive demands, sensory features, and areas. In particular, our findings differ from recent modeling of sensory and prefrontal responses during auditory, contextual-decision making[33], in that the irrelevant input is not completely gated out before reaching PFC, and the relevant input is integrated entirely within PFC. Notably, we inferred not just a modulation of the input strength, but also of its direction. Such a change in direction could be achieved with top-down modulation if each input was multi-dimensional (i.e. originated in multiple neural populations or areas; Supplementary Note 3) and individual dimensions were modulated independently. Alternatively, input amplitude and direction could both be modulated by non-linear dynamics occurring within PFC,[34], a possibility that we did not explicitly model here.

Our fits revealed novel features of the population responses in PFC that were consistent across both LDS mechanisms. Both models implemented input integration in two distinct phases, whereby choice-related signals first emerged along relatively fast decaying dimensions with rotational dynamics, and then transitioned towards orthogonal dimensions with slower, non-rotational dynamics. As a result, individual task-related signals were encoded dynamically along multiple dimensions at different time-scales, consistent with previous analyses of this data[12]. Beyond describing the structure of these signals, here we show how they could emerge dynamically from the interaction of inputs and recurrent dynamics. The dynamics we inferred differs in several ways from that implemented by previously proposed one-dimensional line-attractors[1,16]. Nonetheless, in agreement with such simpler models, we found that at longer time-scales decision signals emerged predominantly along a single integration dimension that was common across contexts[1] (Extended Data Fig. 10).

The LDS models provide several insights into the properties of potential inputs into PFC, beyond their contextual modulation. Both mechanisms inferred multi-dimensional inputs carrying information about both signed coherence and coherence magnitude. As a result, the inputs defined curved manifolds with respect to coherence, in agreement with findings in parietal and frontal areas[11,12,15]. Our results strengthen these previous findings, as in our models the different input components were inferred entirely from the data, rather than being hand-designed[12]. Second, both models inferred inputs that were somewhat transient, even though the fits penalized large magnitude inputs. The fits inferred inputs that became weaker (A, B^{cx} mechanism) or progressively decayed (A^{cx}, B mechanism) late in the trial (Fig. 4b,c). However, models with time-invariant inputs cannot be ruled out, as they resulted in comparable performance (Extended Data Fig. 5b) and captured the neural trajectories nearly as well (Extended Data Fig. 2g,h). Critically, these models relied on mechanisms that were analogous to those described above (data not shown), confirming that the complexity of the PFC responses is well approximated by linear dynamics and not necessarily inherited from

inputs with rich dynamics.

Our models provide an alternative to previously proposed approaches to inferring the properties of inputs into an area. One advantage over past approaches[35,36] is that we make minimal assumption about the properties of the inputs, and in particular about their dimensionality. While several studies have emphasized the importance of inferring inputs to understand cortical dynamics and function[11,35,37–39] such efforts are complicated by unavoidable model degeneracies that arise when attempting to distinguish inputs from recurrent contributions without access to the upstream areas from which the inputs originate[35,40,41]. Our finding that two fundamentally different mechanisms of input selection explain PFC responses equally well is a reflection of such degeneracy. Indeed, both the inputs and the choice-related signals inferred from PFC activity may reflect computations distributed across several cortical areas[2].

Our modeling approach has the advantage of decomposing the dynamics of a complex system into simpler linear parts that are amenable to analysis and interpretation, similar to switching LDS models[42]. A previous application of such models has led to the discovery of line attractor dynamics in the hypothalamus of mice during decisions underlying aggression[43]. In combination with methods from control theory, LDS can also be used to infer inputs that are optimal for a given task, like bringing brain activity into healthy regimes in biomedical applications[44] or optimally configuring cortical dynamics during movement preparation[37–39]. Here, we found that our fitted LDS models are fully controllable[37] (data not shown), and applied methods from control theory to identify the most amplifying dimensions of the dynamics[22], but an exhaustive analysis of this type is beyond the scope of our study.

We validated the assumption of linear dynamics by comparing the LDS fits to the fits from our novel TFR model. The LDS models explained the data essentially as well as TFR, which sets an upper bound to the goodness of fit achievable by an LDS. The success of the LDS models imply that, in PFC, intuitive linear descriptions apply to all regions of state space, and not only to local regions around fixed points[1]. While here we fitted activity from only a relatively short time window from each trial (the 750ms of random dots presentation), recent findings suggests that linear models may not be outperformed by non-linear models in capturing cortical dynamics even over longer time-windows[45]. Nonetheless, analyses based on non-linear models are becoming increasingly common, given their flexibility in capturing very complex neural data[35] and the interest in modeling biological constraints that cannot be captured by linear models[46] (but see[26]).

A crucial aspect of our data-driven modeling approach is that we tested multiple model designs corresponding to specific computations underlying the measured activity. No single model can be expected to perfectly

explain the rich dynamics observed in areas like PFC. Thus, it is important to test multiple alternatives hypotheses and identify all models that are plausible explanations of the data, rather than committing to the best model as the "correct" one[14,47]. Indeed, we found that several LDS mechanisms explained the data similarly well (A, B^{cx} and A^{cx}, B models with time-varying 3D inputs, Fig. 2a, Fig. 3c,d, and 2D inputs Extended Data Fig. 2c,d; and time-constant 3D inputs, Extended Data Fig. 5b, Extended Data Fig. 2g,h), whereas others explained the data less well (models with time-varying 1D inputs, Fig. 2a, Extended Data Fig. 2e,f) or only poorly (a A, B model, fully constrained across contexts, with time-varying 3D inputs, Fig. 2a, Extended Data Fig. 2b). The mechanisms we identified as plausible explanations of the PFC responses share key features with mechanisms of context-dependent integration that were recently described in rats[48]. Notably, that study demonstrated the advantage of pulsatile inputs in distinguishing between different mechanisms of input selection and integration. Similarly, one approach for distinguishing between the two candidate mechanisms we identified would rely on studying the dynamics following perturbations along random state-space directions, which would evolve differently under the two mechanisms due to the different degree of non-normality in the dynamics (Fig. 6a). Alternatively, in simultaneous recordings from large groups of neurons, input and recurrent contributions to the dynamics may sometimes be distinguished based on the properties of trial-by-trial variability of the population responses[40].

Methods for inferring neural population dynamics of the kind proposed here will likely play a key role in uncovering the neural computations underlying behavior. While abstract mental processes were originally hypothesized to reflect structural changes at the level of single neurons (Santiago Ramón y Cajal, see[49]), more recent evidence suggest that cognitive functions arise at the neural population level and depend critically on the ability of neural circuits to flexibly switch between dynamical regimes[17,50–52]. Ultimately, a complete description of neural computations will also explain how neural dynamics emerges from the rich and dynamic structural components of biological circuits[53–55]. The lawful characterization of population level dynamics amounts to a theoretical abstraction of the neural computations emerging from such a rich neural circuit, and provides a key bridge in linking lower-level biological structure to behavior.

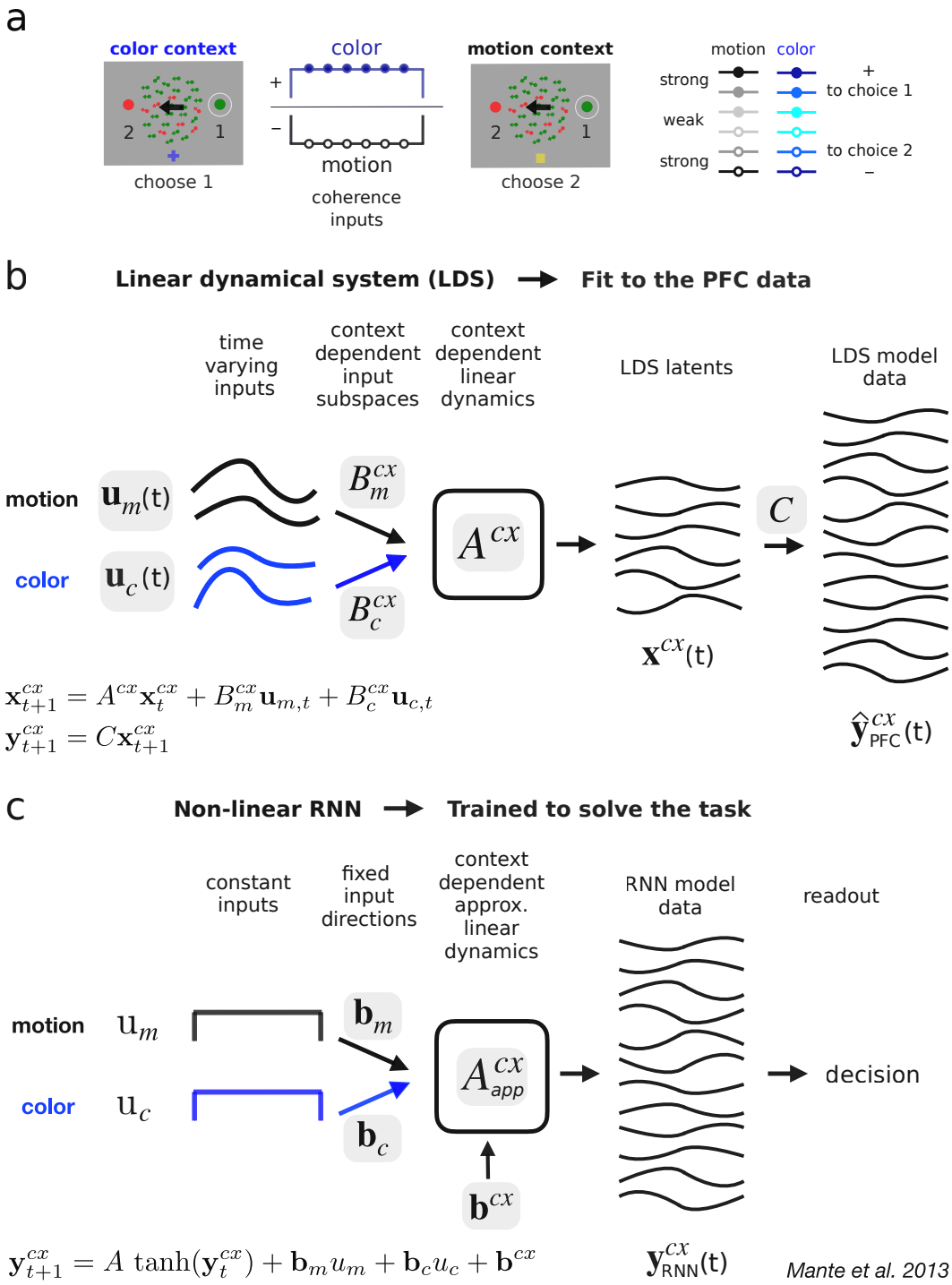


Figure 1 . **Linear dynamical systems (LDS) model-fitting approach to study contextual decision-making computations. a,** Task. Monkeys where trained to choose the target indicated by either the motion or the color of a random dots display depending on context, cued by the fixation point (blue cross, yellow square). The sensory evidence could point towards (choice 1, filled circles/positive values) or away from (choice 2, hollow circles/negative values) one of the two choice targets, placed on the receptive field of the recorded neurons (white circles). The evidence strength was modified by changing the color/motion coherence level of the random dots (3 levels, color shades), yielding 6 coherence conditions in total. Here, strong color and motion evidence were simultaneously presented and pointed at opposite targets (positive color, negative motion). In the color context, the monkey should choose the target matching the overall color of the dots, the green right target (choose 1), while in the motion context it should choose the target indicated by the overall direction of motion of the dots (left arrow), the left red target (choose 2). **b,** An LDS model is fitted to the PFC data from both contexts learning either joint or context-dependent linear dynamics A^{cx} and input dimensions $B^{cx}_{m,c}$. The external inputs are also learned, are fixed across contexts and can vary in time $\mathbf{u}_{m,c}(t)$, and so capture motion and color coherence-related signals throughout the trial. A different $\mathbf{u}_{m,c}(t)$ is learned for each coherence level and direction (6 total, see **a**), and are paired across motion and color coherence conditions to recreate the 36 task conditions (Methods). These parameters define a low-d latent process $\mathbf{x}(t)$ that approximates the dynamics of the high-d PFC data $\hat{\mathbf{y}}_{PFC}(t)$. The orthonormal mapping C from latents $\mathbf{x}$ to observations $\mathbf{y}$ is assumed fixed across contexts. **c,** The non-linear RNN was trained by Mante et al. on the same task as the monkeys. Motion and color sensory evidence were modeled as noisy input signals with mean $u_{m,c}$ constant over time and proportional to the strength of the coherence evidence. Input signals reached the circuit through two fixed input directions across contexts $\mathbf{b}_{m,c}$. The model had the flexibility to learn different contextual input vectors $\mathbf{b}^{cx}$, which changed the dynamics of a fixed, non-linear recurrent network (bottom left equation) between two approximately linear regimes A^{cx}_{app}. A linear readout pooled network responses to generate a decision signal. Network population responses $\mathbf{y}_{RNN}(t)$ were qualitatively compared to the PFC responses. Grey shadings: learned parameters due to training or data fitting.

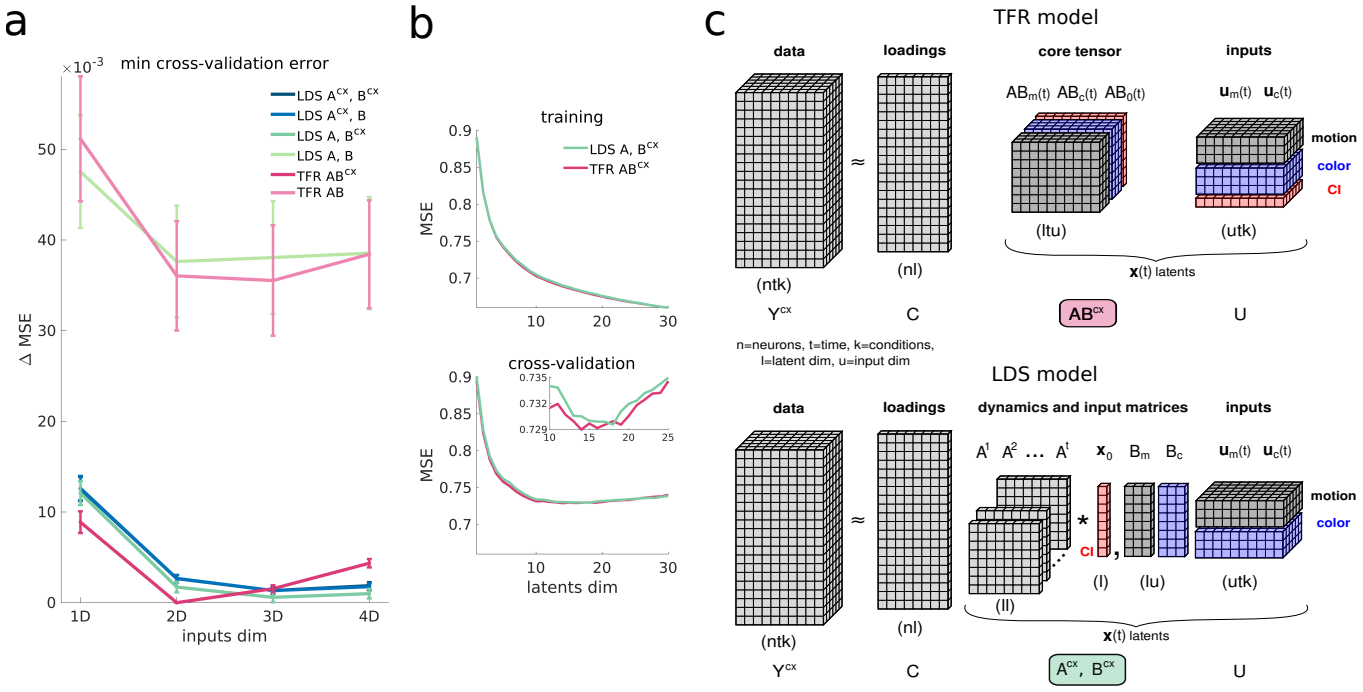


Figure 2 . **Several LDS model classes capture the data equally well, and comparably to a more flexible Tensor Factor Regression (TFR) model. a,** Leave-one-condition-out cross-validation performance (LOOCV)[13] for LDS and TFR models with different input dimensionalies and contextual constraints (36 task conditions, Methods). We report the minimum cross-validation mean squared errors (MSE) across all latent dimensionalities, shown relative to the best performing model (A, B^{cx} model with 3D inputs). The latent dimensionality for which the minimum was attained is documented in Supplementary Table 1. Error bars indicate the standard error mean (sem) across LOOCV folds. A^{cx}, B^{cx} line is below A^{cx}, B's. **b,** Best LDS and TFR models training and LOOCV performance for different latent dimensionalities (TFR 2D AB^{cx} and LDS 3D A, B^{cx}, min LOOCV latent dim = 14 and 18). Data is from monkey A. **c,** TFR model (top). The data tensor Y is factorized into 3 low-rank tensors, all learned. The loadings C (an orthonormal matrix) sets the rank of the factorization and maps the low-d core tensor AB into the high-d neural space. The low-d latents $\mathbf{x}(t)$ are generated by multiplying the core tensor and the input tensor U, which captures motion, color and condition independent (CI) signals. For clarity, we omitted two indicator tensors, one recreating an LDS-like temporal convolution of the core tensor and inputs, and another one used to repeat the inputs across the 36 task conditions (Methods). To generate context-dependent activity Y^{cx} the core tensor can change across contexts AB^{cx}. The LDS model (bottom) is nested within the class of TFR models (Methods). The core tensor AB is replaced by a smaller set of parameters, A and B. The data temporal structure is restricted to be captured by powers of A over time. The latents are recreated by convolving (asterisk symbol) the dynamics matrix powers with the inputs and adding the initial conditions (Methods). Inputs are also repeated across task conditions.

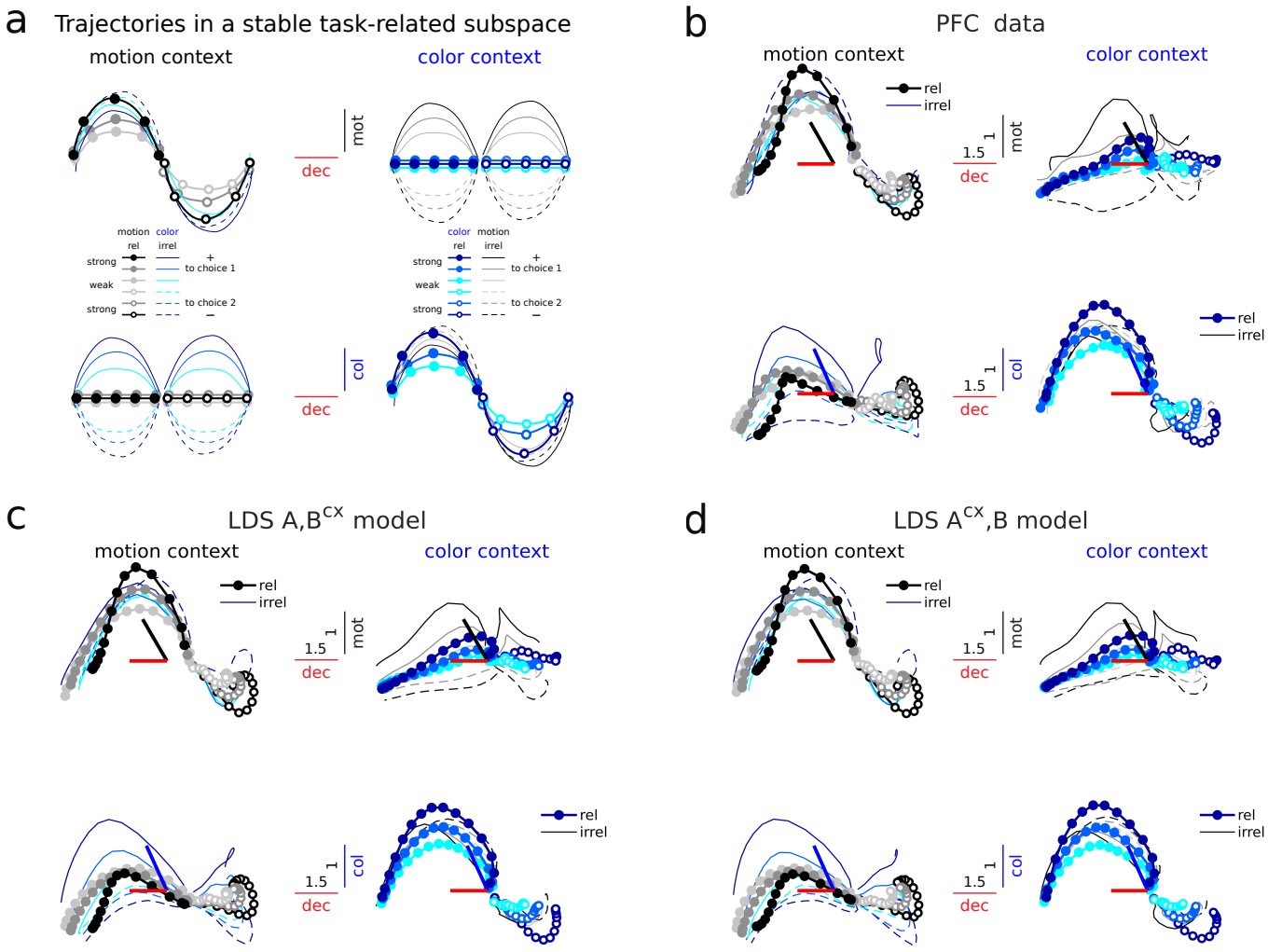


Figure 3 . LDS model classes with both fixed and context-dependent input dimensions can capture the PFC trajectories in a stable task-related subspace. a, Expected trajectories in a task-related subspace that is stable across contexts. Trajectories are sorted by choice and coherence conditions (as in Fig. 1a). The same trajectories in each context are sorted twice, by either motion or color, and then plotted along different task-related dimensions (top, motion and choice, bottom, color and choice). A stable input dimension is expected to reflect the sign and the strength of a specific input regardless of context, but not of other inputs (top, motion dimension encodes motion coherence, in black, but not color coherence, in blue; bottom, the reverse); i.e., stable input dimensions should reflect the input signals regardless of whether these are relevant or irrelevant in a given context (thick doted vs. thin lines). In contrast, a stable decision-related dimension should reflect the sign (and the integrated strength[1]) of the relevant input in each context, but not the irrelevant one—a signature of selective integration (decision axes separate filled vs. hollow circles, the relevant input sign, but not filled vs. dashed lines, the irrelevant; corresponding to motion in the motion context, left, and color in the color context, right); Thus, the movement of the trajectories along the choice axis is coupled to the behavior along the relevant input axis, but not the irrelevant. This suggests that choice-related activity emerges from the relevant evidence signal. Input signals are assumed transient along the input dimensions. **b,** PFC trajectories in the task-related subspace found by Mante et al.[1] using targeted dimensionality reduction (TDR), for monkey A. The subspace captures motion, color and choice-related variance along a set of orthonormal axes that are stable across contexts. Colored thick bars indicate the angle between the found TDR axes before orthogonalization. Numbers on bars are a scaling factor, to ease visualization as in[1]. Trajectories are sorted by choice and motion/color coherence conditions, with color/motion conditions averaged out[1]. **c** Cross-validated model trajectories (LOOCV) for the LDS AB^{cx} model. **d** Same for the A^{cx}, B model. All trajectories have been smoothed with a Gaussian filter for visualization (sliding window size, 5-bins). This step did not change the LDS trajectories much, since they are inherently smooth.

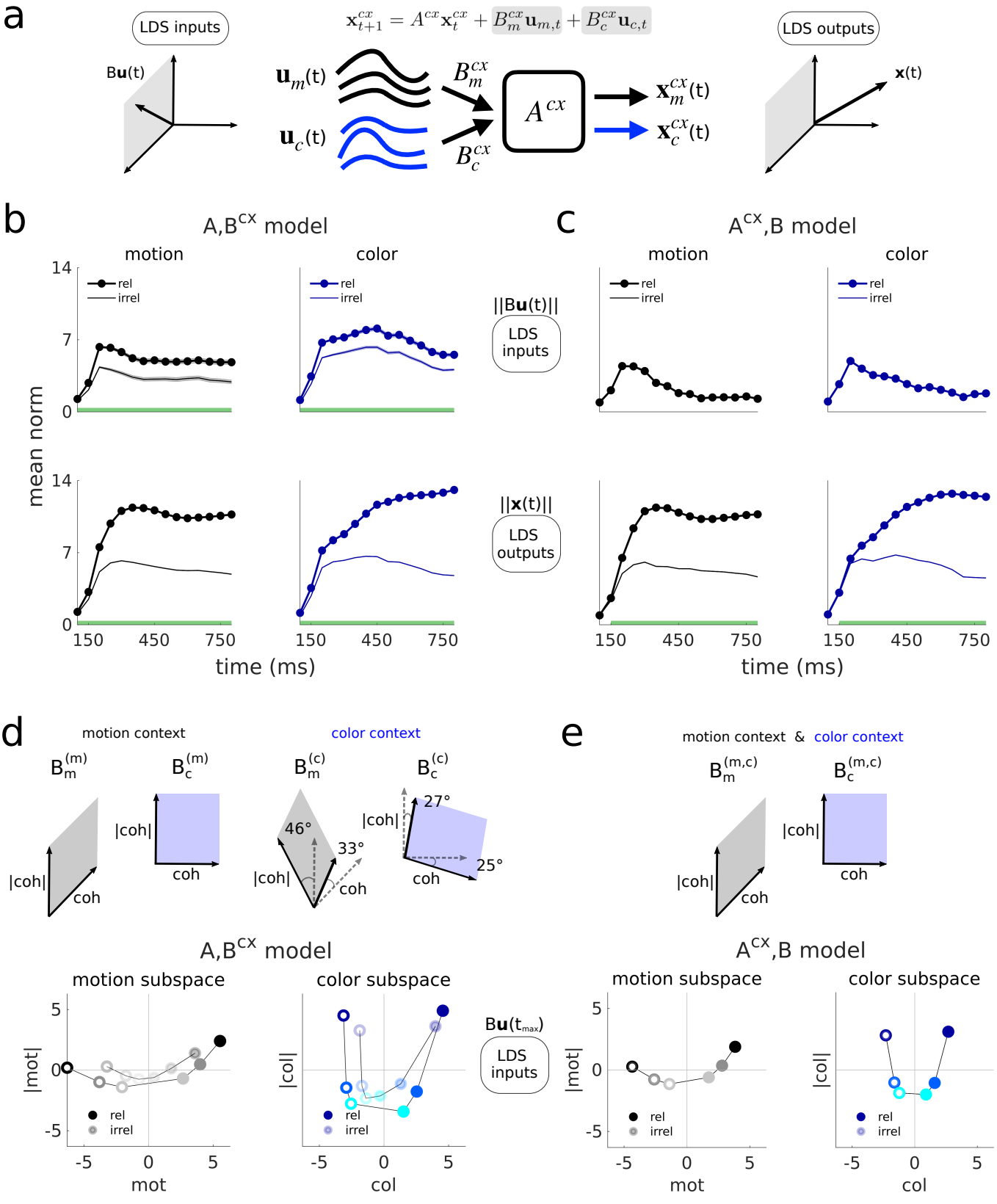
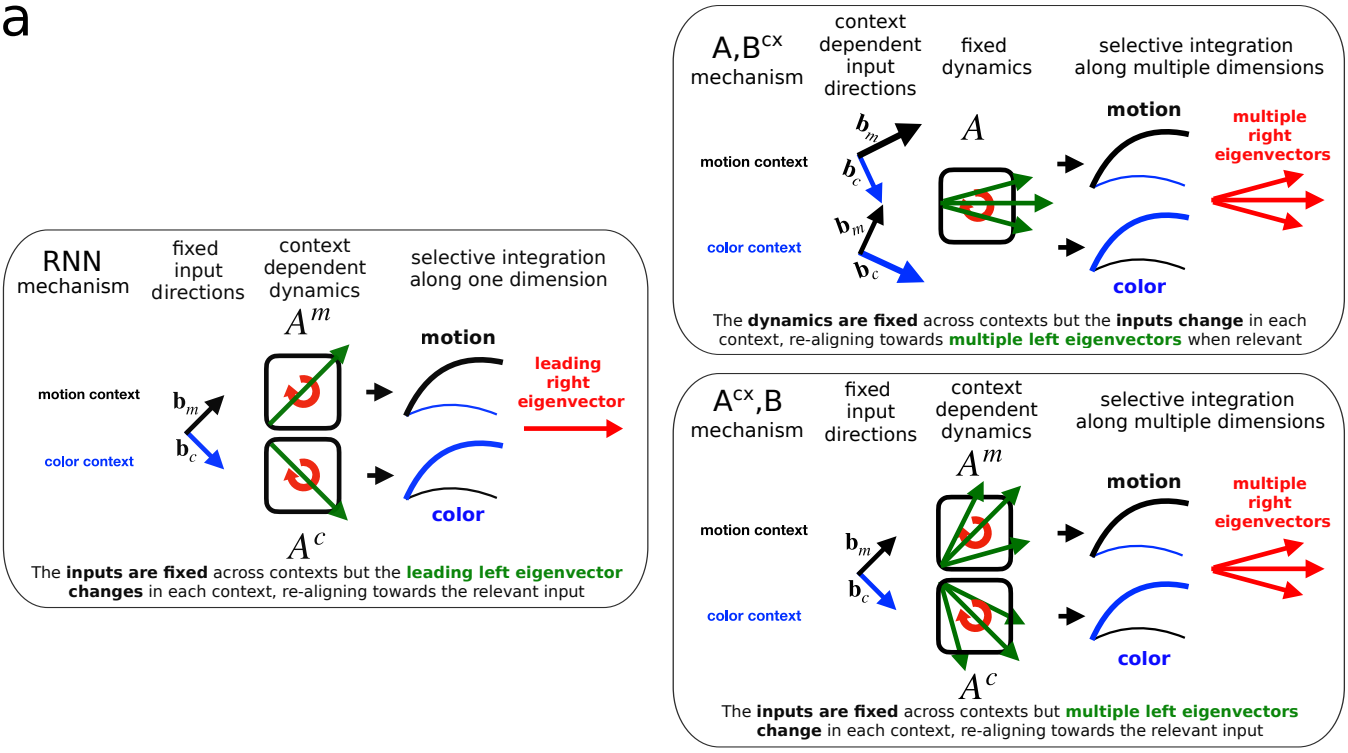
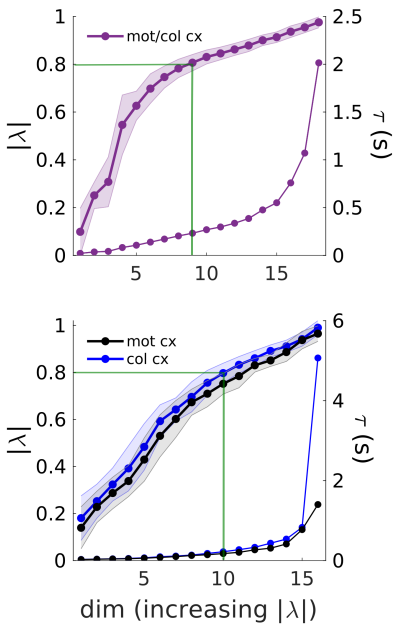


Figure 4 . LDS inputs are integrated selectively by both models, are largely stable across contexts and span curved manifolds. a, Schematic illustrating the input and output LDS signals. The strength of the external inputs over time is defined as the norm of the input vectors at each time step $||Bu(t)||$, which accounts for the norm of B. Input vectors live in the subspace spanned by the columns of B, and hence, the input signals are confined within the input subspaces (left, here a 2D subspace for illustration). The latents live in the full low-d LDS subspace (right, here 3D) and can be seen as the output of the LDS model, since they result from the convolution of the inputs with the dynamics. **b** Top, external input strength over time inferred for the A, B^{cx} model across contexts (relevant vs. irrelevant), here shown for the strongest positive coherence. Bottom, same but for the output signals (latents' norm). Mean across 100 models. Shades = sem (not visible in the outputs). Green bars indicate times when relevant and irrelevant inputs/outputs are significantly different (Wilcoxon rank-sum test, $p<0.001$). **c** Same but for the A^{cx}, B model. **d,e,** Orthonormal 2D subspaces that demix coherence and coherence magnitude variance (coh and |coh|). These dimensions are found within each 3D input subspace from both models by linearly regressing the inferred external input values against the experimental coherence values and their magnitudes. Shown are inputs inferred at t=250ms (after input norm pick strength, Fig. 4b, means over 100 models) for all coherences, projected onto the 2D coh-|coh| planes, which form a curved representation of coherence information. Lines are drawn to ease visualization. For the A, B^{cx} model input projections are shown onto the plane bisecting the two input planes found for each context, which were highly aligned (angles between dashed and filled lines). Color and motion planes were nearly orthogonal within each context for both models. For all input values the mean across conditions has been subtracted out to remove condition independent signals (CI). The latents are computed running through the dynamics the CI subtracted motion and color inputs independently. Monkey A data.

a



b



A,B^cx

A^cx,B

c

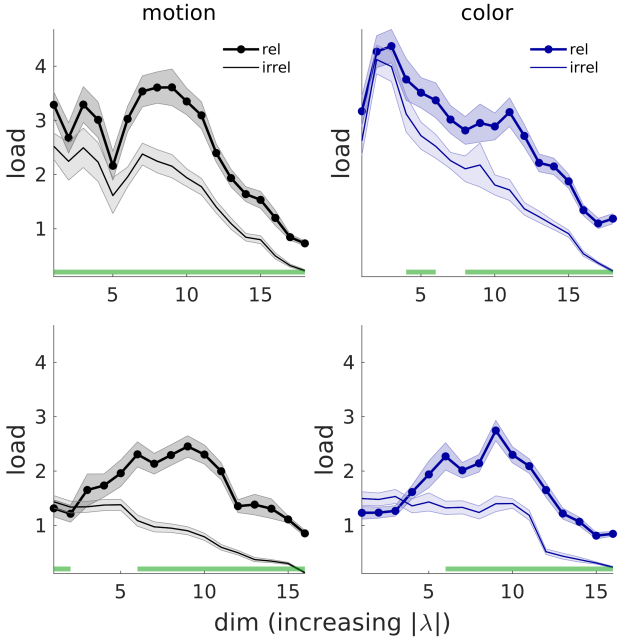


Figure 5 . **Selective integration requires multiple linear dynamics modes. a,** The RNN (left) was build with fixed input directions across contexts $b^{m,c}$. The dynamics in each context switched between two approximately linear regimes, represented by the linearized dynamics matrices A^m and A^c. The leading left eigenvector of $A^{m,c}$ was realigned towards the relevant inputs in each context, loading them onto the slowest output mode of the dynamics (the leading right eigenvector, with associated eigenvalue close to 1), which defined a one-dimensional integrator or line-attractor system[1]. The two LDS models (right) performed a realignment of either the inputs (A, B^{cx}) or the left eigenvectors (A^{cx}, B) across contexts, which loaded the inputs onto multiple modes. The A, B^{cx} model also increased the relevant input's norm (Fig. 4b, bigger input arrows in cartoon). **b** Average eigenvalues norm across 100 models initialized at random (shades=std). The norm sets the rate of decay of each mode (time constant τ), and determines the stability of the dynamics ($|\lambda| > 1$ expanding mode, $|\lambda| < 1$ decaying mode, $|\lambda| = 1$ integration mode). Slow modes have norms close to one ($0.8 < |\lambda| \leq 1$, $\tau > 224$ms, green bars, see Methods). The A^{cx}, B model learns similar eigenvalues across contexts. **c** Average coherence input loads onto the eigenmodes of the dynamics across 100 models (shades=sem). The loads were defined by the non-normalized projection of the coherence inputs onto the left eigenvectors, averaged across all time steps (Methods), here shown for the strongest positive coherence inputs only. For each context, the two models significantly increases the relevant input load onto multiple eigenmodes (green bars, Wilcoxon rank-sum test, $p<0.05$). Monkey A data.

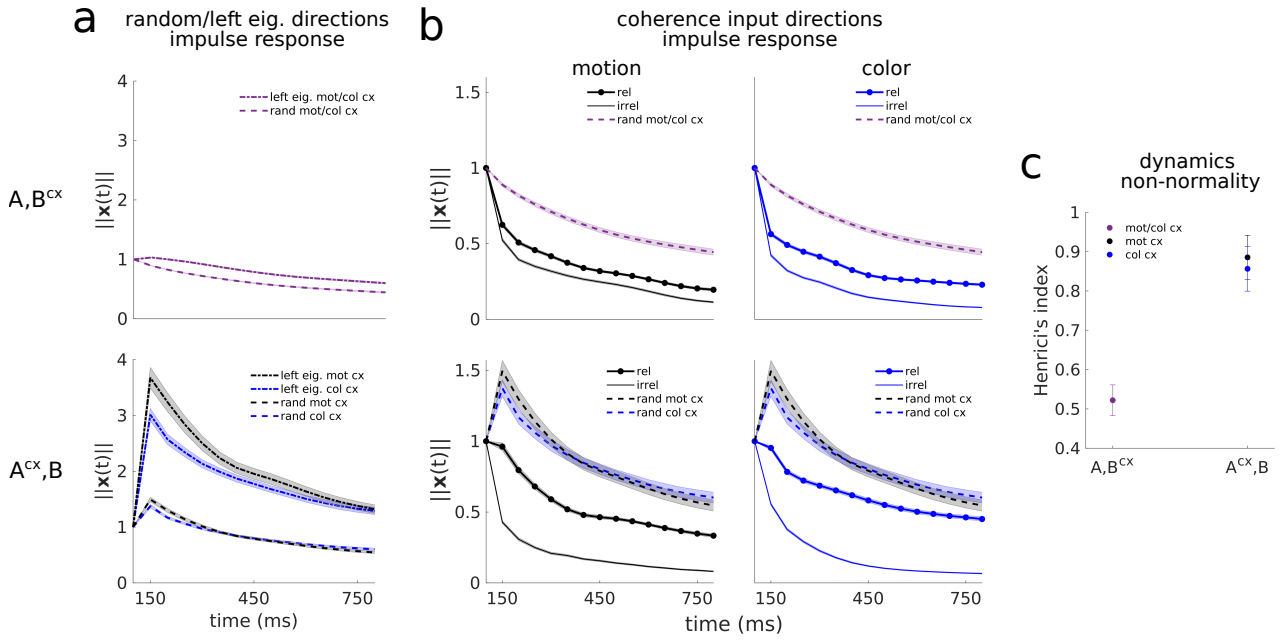


Figure 6 . **Non-normal transient dynamics contributes to selective integration in the A^{cx}, B model. a,** Models A^{cx}, B and A, B^{cx} mean impulse response for perturbations along random directions (dashed lines) and along the left eigenvectors (dotted lines), averaged across 100 models and across left eigenvectors or perturbations (num pert. = num left eigv. = 16/18). This measure shows how the dynamics matrix transforms a perturbation (or input) of unit norm, by tracking the state norm of the system $||\mathbf{x}(t)||$ over time. Note that the A^{cx}, B system has a different impulse response for each context, since the dynamics matrix A changes in each context. Shades, sem across 100 models. **b,** Impulse response for unit norm perturbations along the motion and color coherence input dimensions. For the A, B^{cx} model the dynamics matrix is the same across contexts, and thus, the difference in the impulse response between perturbations along the relevant and the irrelevant input dimensions arises due to the fact that these input dimensions subtly change across contexts. For the A, B^{cx}, the perturbations are applied along the same input directions across contexts, since these are fixed, but the dynamics matrix changes, which causes a different transformation of the same input pulse in each context. Note that the impulse response along the input directions is different from the average impulse response along random directions (dashed lines, same as in **a**), which indicates processing selectivity of the dynamics along the input directions. Shades, sem across 100 models. **c,** Degree of non-normality of the two model classes (Henrici's index, Methods). Error bars, std across 100 models. Monkey A data.

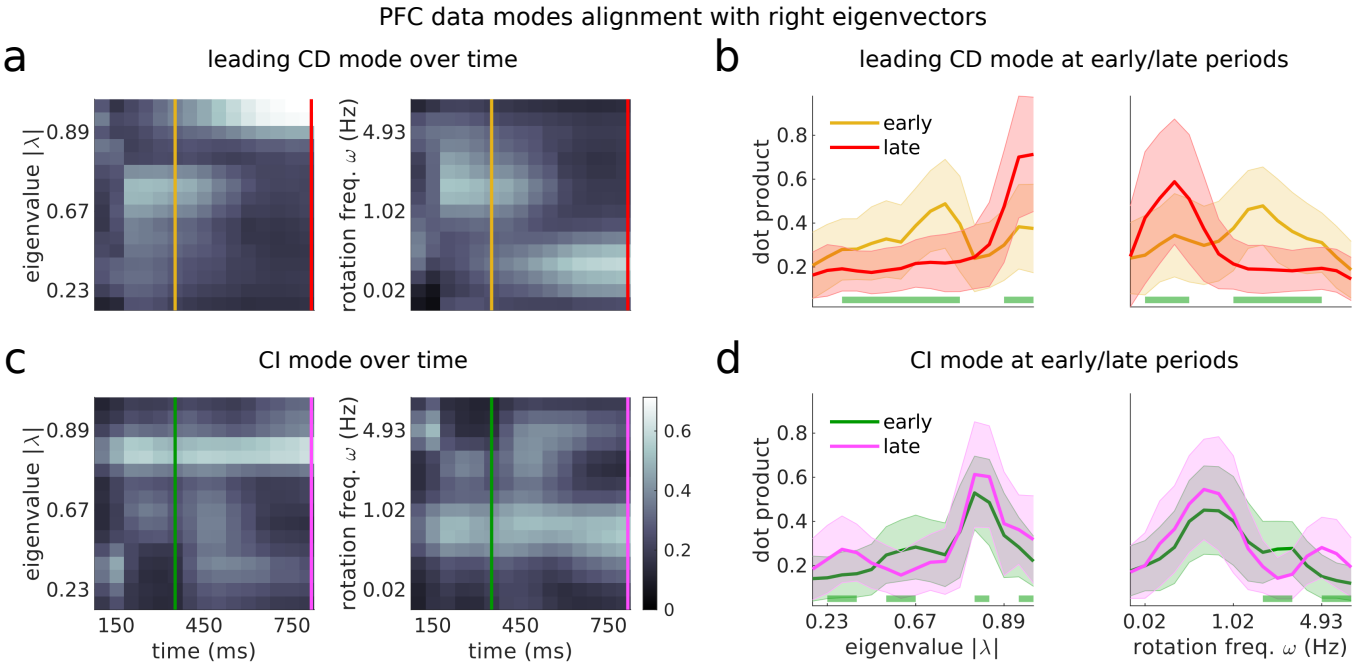


Figure 7 . **Integration of the relevant inputs occurs in two separate phases of the dynamics. a,** Largest variance dimension of the condition dependent (CD) data in the motion context (1st leading singular vector of the CD data, with condition independent CI effects subtracted) at each time step projected onto the right eigenvectors of the dynamics from the A^{cx}, B model (dot products for real eigenvectors, cosines of minimum subspace angles for complex conjugate pairs of eigenvectors, see Methods, averaged across 100 random models). Left/right panel shows dot products sorted by increasing eigenvalue norm/rotation frequency of the associated right eigenvectors (averaged across 100 models, see Methods). Yellow lines mark the early phase of the integration process, t=350ms, the time at which the integrated motion signal in Fig. 4b picks and saturates. Red lines indicate the late phase of the integration process (the last time step of the trial) where decision signals are the strongest[1]. **b,** Mean distribution of alignments across 100 random models at the early and late phases (at times marked in **a**). Shades = std. Green bars indicate the eigenvalues along which the early and late alignment distributions significantly differ (Wilcoxon rank-sum test, $p<0.001$). **c,d** Same as panel **a,b** but for the CI data vector (condition-averaged data vector). Green/purple lines mark the same periods as yellow/red lines. Monkey A data.

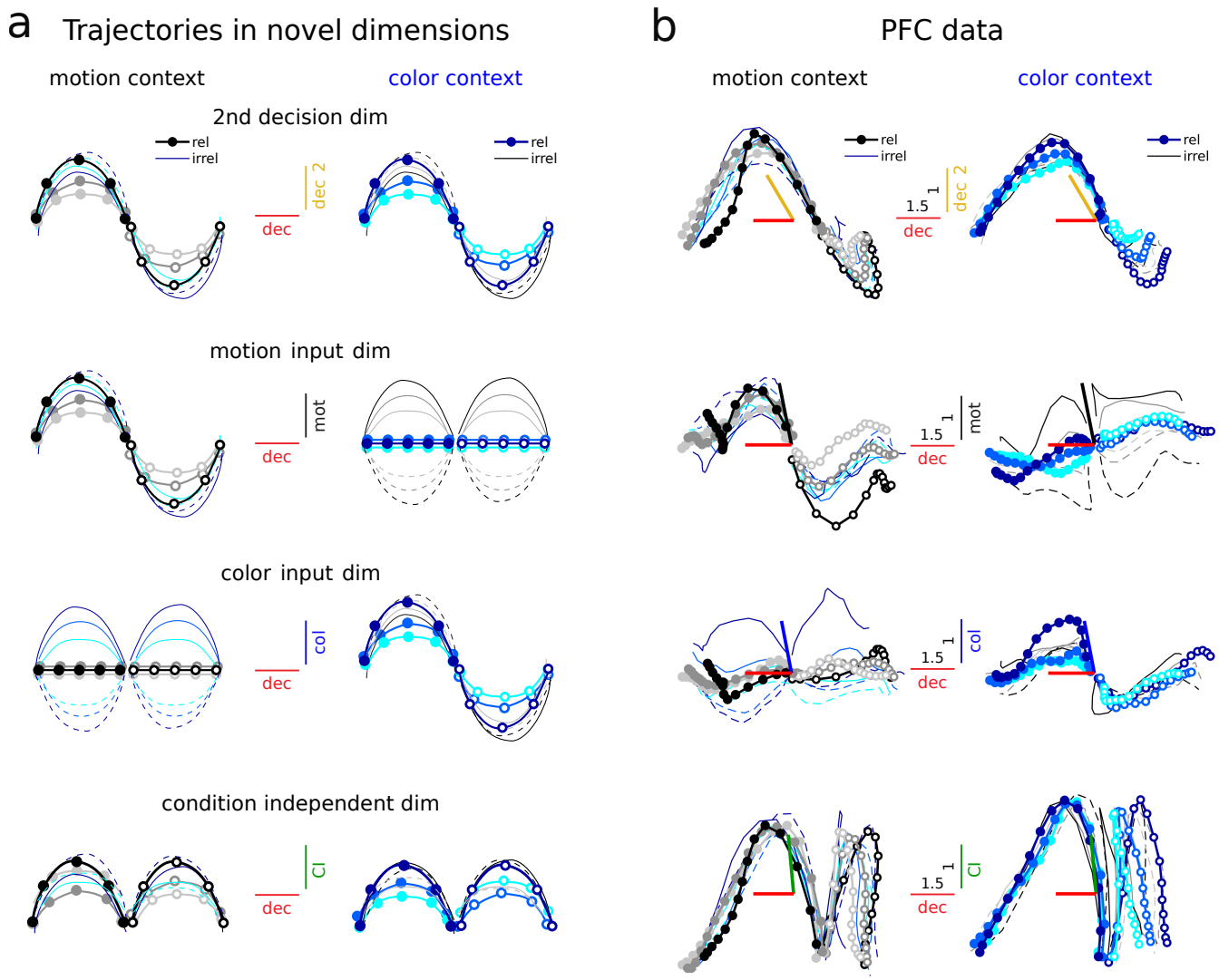


Figure 8 . **The LDS models help discover novel computational dimensions in PFC. a,** Expected trajectories along a novel secondary decision dimension (top), that might reflect transient decision signals, and a dimension that captures CI signals (bottom), plotted against the evolution along a persistent decision dimension (same plotting conventions as in Fig. 3a). Contrast them with known dimensions that reflect motion and color inputs (middle), as in Fig. 3a. The new dimensions capture novel features of the population trajectories. **b,** PFC data trajectories from monkey A along the early integration (secondary decision), decision and CI dimensions averaged across contexts. Same plotting conventions as in Fig. 3b. Middle panels show the trajectories along the LDS-identified input coherence dimensions, also averaged across contexts, and models from each LDS class. The data projections along them resembled the input projections found by TDR (Fig. 3b; TDR-LDS input alignments: A, B^{cx}, mot = 55°, col = 42°, for mean input coherence dimensions across contexts and across 100 models; A^{cx}, B, mot = 44°, col = 31°, for mean input coherence dimensions across 100 models; the alignments are higher than expected by chance, Extended Data Fig. 6b). CI variance has been subtracted to the trajectories in the middle panels to emphasise input-related variance. The dimensions have been orthogonalized with a QR-decomposition[1] (starting with decision, and then dec 2, motion, color and CI). Colored bars show the alignment between dimensions before the orthogonalization step. Trajectories have been smoothed with a Gaussian filter for visualization (sliding window size, 5-bins).

End Notes

Acknowledgements. We thank Renate Krause for providing the RNN data and for valuable discussions. We thank Lea Duncker, Asma Motiwala, Saray Soldado Magraner and Gabriela Michel for providing feedback on the manuscript and for valuable discussions. This work was supported by the Gatsby Charitable Foundation.

Author Contributions. J.S.M. and M.S. conceptualized the methodological approach, developed the models and performed data analysis. V.M. conceived the experiments, collected the data and conceptualized data analysis. All authors actively participated in the interpretation of the data. J.S.M and V.M wrote the paper.

Code Availability. The code will be made publicly available on GitHub upon peer-reviewed publication.

1 Methods

1.1 Experimental procedures and data

1.1.1 Subjects and task

Two adult male rhesus monkeys were trained in a contextual two-alternative forced-choice visual discrimination task. The monkeys had to discriminate either the color or the motion of a random dots display based on context, which was indicated by the fixation cue. The presentation of the random dots lasted for 750ms, after which the monkeys had to wait for a variable delay and report their decision. This was done by saccading to one of two diametrically opposite targets, as indicated by the color or motion evidence. The strength of the evidence was modified by varying the motion and color coherence of the random dots. This was determined by the percentage of dots moving coherently or that were colored the same. Six different coherence settings were used: three strength levels and two directions. The later indicated whether the evidence was pointing towards or away from one of two choice targets—placed at the receptive field (RF) location of the recorded neurons. When the evidence pointed towards the RF of the neurons, their FRs typically increased above baseline. Therefore, positive values were used to define the in-RF evidence. On the contrary, when the evidence pointed away from the RF of neurons, their FRs typically decreased, and hence negative values were used to define the out-RF evidence. Considering all possible motion and color coherence value pairings (6x6), 36 different random dots configurations were presented, which defined the 36 task conditions. Importantly, the motion and color evidence in a given trial could be congruent or incongruent. When incongruent, it was necessary for the monkey to ignore the irrelevant signals in order to perform the correct decision. For further details on the animal procedures and task we refer to the original study[1].

1.1.2 Neural data

Electrophysiological recordings were performed during the task in PFC regions, likely comprising the frontal eye fields (FEF) and surroundings. Both single-unit and multi-unit activity was isolated from the recordings. We referred to them as neurons, for simplicity. Only a few neurons were recorded simultaneously in each trial, but their activity was collected for multiple trials under the 36 different task conditions. Population responses were then constructed by pooling the condition-averaged activity of all neurons. For that, the firing rate of the neurons was computed in each trial using a 50ms sliding square window from spike trains sampled at 1ms. Activity was then averaged across trials under the same condition and z-scored, as in[1]. However, we did not apply any smoothing to the data prior to fitting the models (only in the analysis, for visualization purposes). Thus, the data consisted of a pseudo-population of per-condition averaged PSTHs. The population size was N=727 for monkey A and N=574 for monkey F. We included only neurons which

had recorded activity under all conditions and for all times. As in the original study, we focused our analysis on the period of random dots presentation (750ms, from 100ms after dots onset to 100ms after dots offset) and we analysed only correct trials.

1.2 Models

1.2.1 Linear Dynamical System (LDS) model

The linear dynamical system model considered was a non-probabilistic version of a standard LDS or state space model, with equations

$$\begin{aligned} \mathbf{x}_k(t) &= A\mathbf{x}_k(t-1) + B\mathbf{u}_k(t) \\ \mathbf{y}_k(t) &= C\mathbf{x}_k(t) + \mathbf{d} \end{aligned} \tag{1}$$

where the vector $\mathbf{x}_k(t)$ represents the latent state at time step t and task condition k, $\mathbf{y}_k(t)$ are the observations (a vector containing the PFC condition-averaged PSTHs) and $\mathbf{u}_k(t)$ the external input vector. The dynamics matrix A determines the transition between subsequent latent states. The matrix B defines the input dimensions. The external inputs drive the dynamical system at each time step and define input vectors ($B\mathbf{u}(t)$) that live in the latent subspace spanned by the columns of B. Therefore, the external inputs are assumed linearly mixed in the population at each time step. Note that the input vectors ($B\mathbf{u}(t)$) can point in different directions over time, but these changes are always confined within the input subspaces. The input term in equation (1) can be decomposed to make explicit its color and motion components $B_m\mathbf{u}_m(t) + B_c\mathbf{u}_c(t)$. The loading matrix C maps the low-dimensional latent state onto the high-dimensional neural space. The constant vector $\mathbf{d}$ acts as a bias. This LDS model can be seen as a low-dimensional RNN that reads-out onto a high-dimensional output space.

To capture changes in activity across contexts $\mathbf{y}_k^{cx}(t)$, we fitted an LDS model jointly to the PFC data from each context. The model could learn independent parameters for each context (based on the data from each context) or a single parameter across contexts (using the joint data from both contexts). Both the dynamics matrix A^{cx} and the motion and color subspaces $B_{m,c}^{cx}$ could be context-dependent (cx = mot or col context). The $B_{m,c}^{cx}$ matrices could have different norms, and hence, contextual modulation of inputs could be implemented through changes in both input subspace orientation and norm. The external input signals $\mathbf{u}_{m,c}(t)$ and the mapping C were assumed fixed across contexts.

For each motion and color input dimension, 6 external input time courses were learned, corresponding to the 6 different coherence values in the task (3 strength levels and 2 directions). These were inferred pooling data from all task conditions were a particular coherence level was presented, and therefore, were

shared across task conditions (36 task conditions). The model incorporated additional input constraints, which simplified their temporal structure and were found to improve generalization performance. Time courses were constrained to be the same for all coherence levels of the same direction. That is, a single time course was shared for positive coherences (in-RF evidence) and another one for negative coherences (out-RF evidence). The coherence strength level was learned as a scalar value that multiplied the time course $u(t) = T_{in,out}(t)\, coh_{1,...,6}$. We also fitted a model constrained to learn fixed inputs in time, with $T_{in,out}(1, ..., t) = 1$. The resulting input vectors ($B\mathbf{u}$) also lived in the input subspace defined by the input matrices B, but unlike the input vectors for the time varying input model ($B\mathbf{u}(t)$), these do not move within the input subspaces over time, and remain fixed throughout the trial both in strength and direction.

A different vector of initial conditions was also learned for each context $\mathbf{x}_0^{cx}$. This parameter helped the model recreate the separation of trajectories in state space found across contexts (contextual axis in the Mante et al. study[1]). Note that this feature cannot account for the contextual differences in input integration, since the model is linear, so the relationship between inputs and dynamics modes is the same everywhere in state space. Indeed, a fully constrained model across contexts, with flexibility only in the initial conditions, fails to selectively integrate and poorly reproduces the data (Fig. 2a A, B model, Extended Data Fig. 2b). The initial conditions simply add a shift to the overall dynamics in an input independent manner, since $\mathbf{x}_0$ is the same across all task conditions, so it could only capture baseline changes across contexts. This can be seen in the next equation, which illustrates the unfolding of the dynamics from the initial state and makes the dynamics and inputs convolution explicit

$$\mathbf{x}(t) = A^t\mathbf{x}_0 + \sum_{t'=1}^{t} A^{t-t'}\, B\mathbf{u}(t') \tag{2}$$

This equation also illustrates the presence of a summation degeneracy in the model. The first term defines condition independent (CI) effects, but these can also be captured by the input term. For this reason, in figure Fig. 4, Fig. 8, Extended Data Fig. 3b, Extended Data Fig. 4, Extended Data Fig. 9 and associated Supplementary Figures, we subtracted out CI effects from the input/data trajectories along the input dimensions.

The model was implemented in Python and optimized using gradient descent (ADAM algorithm) to minimize the data reconstruction mean squared error (MSE), with learning rate of 0.009, and the rest of parameters set to the default. The convergence criteria was set to $\Delta\text{MSE} < 10^{-5}$, maximum iterations to 10,000 and minimum iterations to 5,000. The cost function incorporated an input norm penalty to constrain the space of possible solutions and to favour learning small inputs. This encouraged that task-related variables in the data other than the inputs, in particular integration signals, were generated dynamically by the model. Incorporating the penalty minimally impacted performance and helped provide consistent solutions across

fits even when parameters were initialized at random. Therefore, we incorporated such penalty in all our model fits and randomly initialized all parameters. The resulting objective function was:

$$MSE = \frac{1}{NTKC} \sum_{t,k,cx} ||\mathbf{y}_{t,k}^{cx} - \hat{\mathbf{y}}_{t,k}^{cx}||_2^2 + \sum_{t,k,cx} ||B^{cx}\mathbf{u}_{t,k}^{cx}||_2^2 \tag{3}$$

Where N = number of neurons, T = trial duration, K = number of conditions and C = number of contexts. Since the data was z-scored, the MSE captured the fraction of unexplained variance in the data by the model.

Note that the LDS was simply optimized to minimize the MSE of the condition-averaged PSTHs. We did not learn any observations noise model or inferred a latent state distribution, contrary to more standard formulations of the LDS, which are fully probabilistic (and typically infer Gaussian latents, or Gaussian latents combined with Poisson observations[56]). We considered this simpler case given that our data was trial-averaged. Furthermore, our focus was to analyse the parameters of the dynamical model, which are part of the prior distribution over the latents in the probabilistic LDS, and not the data-corrected posterior distribution.

1.2.2 Tensor Factor Regression (TFR) model

The model consists of a factorization of the data tensor structure into three main low-rank tensors

$$Y_{ntk} \approx C_{nl}\, AB_{ltu}\, U_{utk} \tag{4}$$

where n = number of neurons, t = time steps, k = conditions, l = latent dimensionality, u = input dimensionality + 1D baseline. The tensor C (an orthonormal matrix) sets the rank of the factorization and maps the low-dimensional core tensor AB into the high-dimensional neural space. The inputs tensor U captures the condition-dependent effects in the data and acts as a regressor, when this is known. When learned, as it is the case here, it is used to capture task-related variables, such as motion and color input signals. Note that similar to the LDS, these signals are assumed linearly mixed in the population at each time step.

In the previous equation, for clarity (as in Fig. 2c), we omitted an indicator tensor T that emulates the LDS-like convolution of inputs and dynamics

$$Y_{ntk} \approx C_{nl}\, AB_{lt''u}\, T_{tt't''}\, U_{ut'k} \tag{5}$$

where $T_{tt't''} = \delta(t - t' = t'')$. One can see how this model encompasses the LDS by writing

$$AB_{lt''u}\, T_{tt't''} = \begin{cases} A_{ll}^{t-t'} B_{lu} & t \geq t' \\ 0 & otherwise \end{cases} \tag{6}$$

where A and B correspond to the LDS dynamics and input subspace matrices, respectively.

The inputs incorporated constraints analogous to the LDS. First, inputs were repeated across conditions with an additional indicator tensor Q

$$Y_{ntk} \approx C_{nl}\, AB_{lt''u}\, T_{tt't''}\, Q_{ukc}\, U_{ct'} \tag{7}$$

where $c = (6 \times u) + 1$ indexes the six coherence conditions, plus baseline (that captures CI effects). In this way, the tensor U is designed to extract common task-related variables across conditions. Second, the temporal structure of the inputs was constrained to be the same for coherences of the same direction. For that, the input tensor U was factorized further as follows

$$Y_{ntk} \approx C_{nl}\, AB_{lt''u}\, T_{tt't''}\, Q_{ukcd}\, P_c\, R_{dt'} \tag{8}$$

where $d = 2$ indexed the two possible coherence directions.

The parameters of the TFR model can be computed by alternating the estimation of the tensors $W = CAB$ and U. For that, one can consider the tensor unfolding $Y_{(n)(tk)}$ and compute C and AB via reduced rank regression, with fixed U. Then, knowing W, the least squares estimate of U can be computed. In practice, we estimated the parameters following the same optimization procedure we used for the LDS, which provided identical results. That is, the model was implemented in Python and optimized using ADAM, with objective given by the data reconstruction MSE.

The TFR model is related to existing regression-based methods that discover task-related variance in the data[1,12,57], but with the difference that TFR incorporates task regressors that are themselves learned from the data. Another key distinction is that TFR considers a joint factorization of the whole data tensor structure, similar to other studies[58], but the tensor components relate to the parameters of the task and are themselves low-dimensional.

1.2.3 Recurrent Neural Network (RNN) model

We generated data from a RNN model of the same type as used by Mante and colleagues[1].

$$\mathbf{y}(t) = A \tanh(\mathbf{y}(t-1)) + \boldsymbol{b}_m u_m + \boldsymbol{b}_c u_c + \boldsymbol{b}^{cx} \tag{9}$$

Briefly, the model was a non-linear RNN trained using back-propagation to solve the same contextual decision-making task as the monkeys. Contrary to the LDS, the RNN was not optimized to reproduce the complex and heterogeneous responses of PFC neurons, i.e. to match PFC's dynamics. This network

was designed with the same built-in assumptions as in the original model (Fig. 1c). Namely, that the external coherence input signals u_m and u_c were noisy but constant in time, with mean proportional to the strength of the coherence evidence, and that these reached the circuit through two fixed input dimensions across contexts $\mathbf{b}_m$ and $\mathbf{b}_c$. The model had the flexibility to learn different contextual input vectors $\mathbf{b}^{cx}$, whose activation changed the dynamics of a fixed, non-linear recurrent network (with connectivity A). This allowed the model to switch its state between two approximately linear regimes (A^{cx}_{app}=A^{mot}_{app}/A^{col}_{app}), performing different computations in each context. Namely, selecting the contextually-relevant input signals for integration towards a choice and dynamically discarding the irrelevant ones. In the original study, the RNN population activity $\mathbf{y}^{cx}_{RNN}$ was analysed and qualitatively compared with the PFC activity, revealing some shared features that were suggestive of a common contextual-integration mechanism between PFC and the network. The network could be "reverse-engineered" in order to understand the mechanism underlying such computation, by linearizing the dynamics around the identified fixed points of the system (obtaining different local $A^{mot/col}_{app}$, which however were similar in dynamics and could be averaged[1]). In this work, we instead focused on analysing the properties of LDS models fit to the RNN population activity $\mathbf{y}^{cx}_{RNN}$ (the z-scored condition-averaged responses, as in the PFC data, from a 100 RNN units). For further details on the RNN training and analysis we refer to the original study[1].

1.3 Dynamics analysis

1.3.1 Eigenspectrum and time constants

The eigenspectrum of the dynamics matrix contains both real and imaginary eigenvalues, which come in complex-conjugate pairs

$$\begin{aligned} \lambda &= \lambda_{re} + \lambda_{im} i \\ \lambda^{\dagger} &= \lambda_{re} - \lambda_{im} i \end{aligned} \tag{10}$$

The absolute value of the eigenvalues determines the rate of decay or growth of each dynamic mode[59]. Modes are stable if they either decay or persist

$$\begin{aligned} \lambda &\leq 1 \qquad \forall \lambda \; real \\ |\lambda| = \sqrt{\lambda^2_{re} + \lambda^2_{im}} &\leq 1 \qquad \forall \lambda \; complex \end{aligned} \tag{11}$$

The slower the decay, the slower or more persistent a given mode is, and the greater input information is preserved along it. The time constant measures the time at which the initial state will have decayed by 37%

(1/e=0.37) along a given mode. Considering that each time step is 50ms (the data binning size)

$$\begin{aligned} x_t &= |\lambda|^t x_0 \\ (1/e)x_0 &= |\lambda|^t x_0 \\ \tau &= \frac{\log(1/e)}{\log|\lambda|}50 \end{aligned} \tag{12}$$

We consider that a mode is slow if it has a norm close to one, that is, if $|\lambda| > 0.8$. This corresponds to a decay time constant of $\tau > 224$ms, which encompasses approximately a third of the trial duration. Given that the inferred external inputs in the two models are strong for the first third of the trial (Fig. 4a,b), inputs mapped onto such slow modes largely persist until the end of the trial, albeit with some decay for modes $|\lambda| = 0.8 - 0.9$. In particular, by the second third of the trial, inputs would have decayed by at most 37%. We consider the slowest modes to have $|\lambda| > 0.9$ and time constant $\tau > 475$ms. These are strongly persistent and preserve most input information until the end of the trial. The relatively fast decaying modes ($|\lambda| = 0.7 - 0.8$, $\tau = 140 - 224$ms) are somewhat persistent, but loose most input information by the end of the trial.

Many of the eigenvalues were imaginary, indicating the presence of rotational dynamics in the data[17]. Some of the eigenvalues were negative, which also indicate the presence of oscillations[44]. A few models identified slightly unstable eigenmodes (with eigenvalue norm slightly bigger than 1), but this is expected when learning from finite trial lengths and limited data samples[60]. However, the models inferred from monkey F data, in particular for the A, B^{cx} model, seemed to use instability properties of the dynamics in order to capture specific features of the data (Supplementary Fig. 8a,d).

1.3.2 Rotational dynamics measure

The existence of complex eigenvalues indicates the presence of rotational dynamics in the data. Rotations are confined to the planes defined by pairs of complex-conjugate eigenvectors, with directions spanned by the real and imaginary components of the vectors[59]. Considering first the phase plane representation of a complex-conjugate pair of eigenvalues in polar coordinates

$$\begin{aligned} \lambda_{re} &= |\lambda| \cos\omega \\ \lambda_{im} &= |\lambda| \sin\omega \end{aligned} \tag{13}$$

where

$$\omega = \arctan(\frac{\lambda_{im}}{\lambda_{re}}) \qquad \forall \lambda \; complex \tag{14}$$

Rotations on each plane are determined by the rotation matrix J, which derives from the dynamics matrix expressed in the Jordan normal form[59]. As an example, for a 2D system with two distinct complex eigenvalues

$$J = \begin{bmatrix} \lambda_{re} & -\lambda_{im} \\ \lambda_{im} & \lambda_{re} \end{bmatrix} = |\lambda| \begin{bmatrix} \cos\omega & -\sin\omega \\ \sin\omega & \cos\omega \end{bmatrix} \tag{15}$$

Rotations evolve in time following powers of J, with amplitude over time (the rate of decay or growth) given by the absolute value of the eigenvalues

$$J^t = \left(|\lambda| \begin{bmatrix} \cos\omega & -\sin\omega \\ \sin\omega & \cos\omega \end{bmatrix} \right)^t = |\lambda|^t \begin{bmatrix} \cos\omega t & -\sin\omega t \\ \sin\omega t & \cos\omega t \end{bmatrix} \tag{16}$$

Therefore, ω is the rotation frequency in the complex plane. Note that the frequency increases when the ratio $\frac{\lambda_{im}}{\lambda_{re}}$ is big. The rotation frequency ω is given in rad/s and $f = \omega/(2\pi)$ in Hz. Since the data was down-sampled at 20Hz (50ms bins), the frequency is given by $f = 20\omega/(2\pi)$ in Hz. For real modes, the rotation frequency is zero.

1.3.3 Non-normality measure

The Henrici's index measures the degree of non-normality of the the dynamics, and is given by[61]

$$H = \frac{\sqrt{||A||_F^2 - \sum_i |\lambda_i|^2}}{||A||_F} \tag{17}$$

This is a normalized metric with values between 0 and 1, with 0 indicating that the system is normal, and 1 that is maximally non-normal. A system is normal when its dynamics can be described with an orthonormal eigenvector basis. A system is non-normal when its eigenvectors do not necessarily form an orthonormal basis, and the transformation to eigenvector coordinates may involve a strong distortion of the phase space[61]. Importantly, in normal linear networks, the network responses are explained with a linear combination of exponentially decaying modes (if the system is stable), with timescales defined by the corresponding eigenvalue (equation (12)). In non-normal stable networks, however, more complex patterns can emerge, which often involve transient responses where the network activity temporarily grows, but eventually decays as in normal systems.

A crucial property of non-normal systems is that they have different left and right eigenvectors.

$$A = R\Lambda L \tag{18}$$

with $L = R^{-1}$, whereas for normal systems $L = R^{\dagger}$ (†=conjugate transpose). This non-normal property allowed the RNN to change the leading left eigenvectors across contexts, while keeping the right eigenvectors pointing in the same direction.

1.3.4 Input loads

To compute the input load onto the modes of the dynamics, we start by expressing the latents in the left eigenvectors basis.

$$\begin{aligned} \mathbf{x}(t) &= A\mathbf{x}(t-1) + B\mathbf{u}(t) \\ \mathbf{x}(t) &= (R\Lambda L)\mathbf{x}(t-1) + B\mathbf{u}(t) \\ L\mathbf{x}(t) &= \Lambda L\mathbf{x}(t-1) + LB\mathbf{u}(t) \end{aligned} \tag{19}$$

where we have taken the eigendecomposition of the matrix A, with R containing the right eigenvectors in its columns and $L = R^{-1}$ the left eigenvectors in its rows. We have then left-multiplied by L. Defining $\boldsymbol{\alpha}(t) = L\mathbf{x}(t)$ we obtain

$$\boldsymbol{\alpha}(t) = \Lambda\boldsymbol{\alpha}(t-1) + LB\mathbf{u}(t) \tag{20}$$

The evolution of the latents in this basis is independent, that is, decoupled from one another—given that the matrix Λ is diagonal. Unrolling this equation in time we obtain

$$\boldsymbol{\alpha}(t) = \Lambda^t L\mathbf{x}_0 + \sum_{t'=1}^{t} \Lambda^{t-t'} LB\mathbf{u}(t') \tag{21}$$

As the eigenmodes are independent, we can write down a set of uncoupled equations that describe the evolution of each eigenmode, one for each entry of the vector $\boldsymbol{\alpha}$, given by α_l with l indexing the latent variable dimension

$$\alpha_l(t) = \lambda_l^t\, \boldsymbol{l}_l^\intercal \mathbf{x}(0) + \sum_{t'=1}^{t} \lambda_l^{t-t'}\, \boldsymbol{l}_l^\intercal B\mathbf{u}(t') \tag{22}$$

and $\boldsymbol{l}_l$ being the lth left eigenvector. The input "loads" are defined by the last term of the summation, which correspond to the non-normalized projection of the inputs onto the left eigenvectors (note that neither the input vectors nor the left eigenvectors are unit norm).

$$load_l(t) = \boldsymbol{l}_l^\intercal B\mathbf{u}(t) \tag{23}$$

This term specifies how strongly the inputs are mapped onto the dynamic modes, at each time step t, before being processed by the dynamics (i.e., in this basis, before being scaled by λ). The extend to which the inputs are mapped or "loaded" onto each mode depends on the alignment between the input vectors and each left eigenvector, as well as the norm of both vectors. For each pair of complex modes, the load is given by

$$load_{l-l^\dagger}(t) = 2||\, \Re\{\boldsymbol{l}_l^\intercal\} B\mathbf{u}(t)\, \Re\{\boldsymbol{r}_l\} - \Im\{\boldsymbol{l}_l^\intercal\} B\mathbf{u}(t)\, \Im\{\boldsymbol{r}_l\}\, || \tag{24}$$

Where $\Re\{.\}$ and $\Im\{.\}$ take the real and imaginary components of their arguments. The rationale for the expression above comes from the following. For complex modes, equation (23) contains imaginary numbers,

since the left eigenvectors are complex, so we cannot interpret the loads in this basis. However, we can do it in the original state vector basis **x**(t), which is real. To change basis, we use $\boldsymbol{\alpha}(t) = L\mathbf{x}(t)$ and express $\mathbf{x}(t)$ as a linear combination of the state along each right eigenvector dimension. The coefficients of the linear combination are given by $\alpha_l(t)$, which contains the input loads

$$\mathbf{x}(t) = R\boldsymbol{\alpha}(t) = \sum_l \alpha_l(t)\boldsymbol{r}_l \tag{25}$$

We can now make explicit the contribution due to real eigenmodes and complex eigenmodes, which come in complex conjugate pairs ($l - l^\dagger$).

$$\mathbf{x}(t) = \sum_{l-l^\dagger, img} (\alpha_l(t)\boldsymbol{r}^l + \alpha_l^\dagger(t)\boldsymbol{r}^{l\dagger}) + \sum_{l, real} \alpha_l(t)\boldsymbol{r}_l \tag{26}$$

Due to the complex conjugacy, the imaginary numbers end up cancelling out in the summation, and only real terms survive. This is why in this basis, the state vector **x**(t) is real. In particular, the way the complex roots end up contributing to the state dynamics is given by their real and imaginary parts. This is because for each pair of complex conjugate roots, two complementary real solutions exist, which are given by the sum and difference modes $\alpha_{l^\pm}(t)$

$$\begin{aligned}
\alpha_{l^+}(t) &= \frac{1}{2}(\alpha_l(t) + \alpha_l^\dagger(t)) = \Re\{\alpha_l(t)\} \\
\alpha_{l^-}(t) &= \frac{1}{2i}(\alpha_l(t) - \alpha_l^\dagger(t)) = \Im\{\alpha_l(t)\}
\end{aligned} \tag{27}$$

This can be seen by expanding the complex term in the state equation

$$\begin{aligned}
\alpha_l(t)\boldsymbol{r}_l + \alpha_l^\dagger(t)\boldsymbol{r}_{l\dagger} &= (\Re\{\alpha_l(t)\} + i\Im\{\alpha_l(t)\})(\Re\{\boldsymbol{r}_l\} + i\Im\{\boldsymbol{r}_l\}) \\
&\quad + (\Re\{\alpha_l(t)\} - i\Im\{\alpha_l(t)\})(\Re\{\boldsymbol{r}_l\} - i\Im\{\boldsymbol{r}_l\}) \\
&= 2\Re\{\alpha_l(t)\}\Re\{\boldsymbol{r}_l\} - 2\Im\{\alpha_l(t)\}\Im\{\boldsymbol{r}_l\} \\
&= 2(\alpha_{l^+}(t)\Re\{\boldsymbol{r}_l\} - \alpha_{l^-}(t)\Im\{\boldsymbol{r}_l\})
\end{aligned} \tag{28}$$

Thus

$$\mathbf{x}(t) = \sum_{l-l^\dagger, img} 2(\Re\{\alpha_l(t)\}\Re\{\boldsymbol{r}_l\} - \Im\{\alpha_l(t)\}\Im\{\boldsymbol{r}_l\}) + \sum_{l, real} \alpha_l(t)\boldsymbol{r}_l \tag{29}$$

To understand how the inputs are loaded at each time step t into the dynamic modes to affect the latent state, we focus on the last term of the summation in the $\alpha_l(t)$ equation (22), as we did before

$$\mathbf{x}(t)_{input} = \sum_{l-l^\dagger, img} 2(\Re\{\boldsymbol{l}_l^\intercal\}B\mathbf{u}(t)\,\Re\{\boldsymbol{r}_l\} - \Im\{\boldsymbol{l}_l^\intercal\}B\mathbf{u}(t)\,\Im\{\boldsymbol{r}_l\}) + \sum_{l, real} \boldsymbol{l}_l^\intercal B\mathbf{u}(t)\boldsymbol{r}_l \tag{30}$$

The last term contains the input loads along each real mode, $\boldsymbol{l}_l^\intercal B\mathbf{u}(t)$, which gives equation (23). This value indicates how much of the input is mapped along each right eigenvector direction $\boldsymbol{r}_l$. Thus, considering only

this term, the latent state vector is reconstructed with a linear combination of real right eigenvectors, weighted by the input loads. Note however, that the right eigenvectors are not orthogonal, so the result of the sum could be non-trivial, if for instance some of this vectors cancel out, or give rise to amplification (Supplementary Notes, Extended Data Fig. 7). The total input contribution or load along each direction $\boldsymbol{r}_l$ is thus given by the norm of the vector $\boldsymbol{l}_l^\intercal B\mathbf{u}(t)\boldsymbol{r}_l$. Since the real right eigenvectors are normalized, this is equal to $\boldsymbol{l}_l^\intercal B\mathbf{u}(t)$, which gives equation (23). Similarly, the load for each complex conjugate pair of modes is given by the norm of the vector $2(\Re\{\boldsymbol{l}_l^\intercal\}B\mathbf{u}(t)\,\Re\{\boldsymbol{r}_l\}-\Im\{\boldsymbol{l}_l^\intercal\}B\mathbf{u}(t)\,\Im\{\boldsymbol{r}_l\})$, which gives equation (24). This vector lives within the 2D plane spanned by the real and imaginary components of the complex-conjugate right eigenvector pairs.

To compute the loads in Fig. 5c, we use the inferred inputs for the largest motion and color positive coherence values, and project them along the coherence dimension. So the loads are computed using the coherence component of $B\mathbf{u}(t)$, for all times and all 100 models, and then averaged across time and models. For complex modes, the same load is shared across both complex conjugate pairs, and is computed using equation (24).

1.3.5 Most amplifying dimensions

The most amplifying modes where found following[22], by computing the Observability Gramian and its associated eigenvectors. The most amplifying modes are defined by the eigenvectors with the largest associated eigenvalues. We computed the Observability Gramian by solving the following Lyapunov equation

$$A^T X + XA + C^T C = 0 \tag{31}$$

where A is the LDS models dynamics matrix and C is the loading matrix. We considered only stable models[22], which in our case were 90% of the 100 A, B^{cx} models and 85% (mot cx), 60% (col cx) of the A^{cx}, B models in monkey A.

1.4 Additional analysis methods

1.4.1 Alignment metrics

We report alignments between different dimensions using either dot products or angles (in degrees). When computing alignments between a given vector and complex eigenvector dimensions, we consider the plane spanned by the real and imaginary vector components of the pair of complex conjugate modes, and compute the minimum subspace angle between the vector and the plane.

1.4.2 Statistical tests

To asses statistical significance of differences between distributions, such as the relevant vs. irrelevant load distributions in Fig. 5c, we used a Wilcoxon rank-sum test with significance levels (p-values) set at $p<0.001$ (Fig. 4b,c, Fig. 7b, Extended Data Fig. 8b, and associated Supplementary Figures) or $p<0.05$ (Fig. 5c and associated Supplementary Figures, also Supplementary Fig. 4). This is a two-sided rank sum test of the null hypothesis that two independent samples come from distributions with equal medians.

References

1. Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. en. *Nature* **503,** 78–84. ISSN: 1476-4687 (Nov. 2013).
2. Siegel, M., Buschman, T. J. & Miller, E. K. Cortical information flow during flexible sensorimotor decisions. en. *Science* **348,** 1352–1355. ISSN: 0036-8075, 1095-9203 (June 2015).
3. Fuster, J. *The Prefrontal Cortex* en. ISBN: 978-0-12-407815-4. (2018) (Elsevier, 2015).
4. Miller, E. K. & Cohen, J. D. An integrative theory of prefrontal cortex function. eng. *Annual Review of Neuroscience* **24,** 167–202. ISSN: 0147-006X (2001).
5. Wallis, J. D., Anderson, K. C. & Miller, E. K. Single neurons in prefrontal cortex encode abstract rules. en. *Nature* **411,** 953–956. ISSN: 1476-4687 (June 2001).
6. Tanji, J. & Hoshi, E. Role of the lateral prefrontal cortex in executive behavioral control. eng. *Physiological Reviews* **88,** 37–57. ISSN: 0031-9333 (Jan. 2008).
7. Buckley, M. J. *et al.* Dissociable components of rule-guided behavior depend on distinct medial and prefrontal regions. eng. *Science (New York, N.Y.)* **325,** 52–58. ISSN: 1095-9203 (July 2009).
8. Katsuki, F. & Constantinidis, C. Unique and shared roles of the posterior parietal and dorsolateral prefrontal cortex in cognitive functions. *Frontiers in Integrative Neuroscience* **6.** ISSN: 1662-5145. (2018) (May 2012).
9. Suzuki, M. & Gottlieb, J. Distinct neural mechanisms of distractor suppression in the frontal and parietal lobe. eng. *Nature Neuroscience* **16,** 98–104. ISSN: 1546-1726 (Jan. 2013).
10. Newsome, W. T., Britten, K. H. & Movshon, J. A. Neuronal correlates of a perceptual decision. en. *Nature* **341,** 52–54. ISSN: 0028-0836 (Sept. 1989).
11. Soldado Magraner, J. *Linear Dynamics of Evidence Integration in Contextual Decision Making.* Doctoral thesis (University College London, Dec. 2018).
12. Aoi, M. C., Mante, V. & Pillow, J. W. Prefrontal cortex exhibits multidimensional dynamic encoding during decision-making. en. *Nature Neuroscience* **23,** 1410–1420. ISSN: 1546-1726 (Nov. 2020).
13. Elsayed, G. F. & Cunningham, J. P. Structure in neural population recordings: an expected byproduct of simpler phenomena? en. *Nature Neuroscience* **20,** 1310–1318. ISSN: 1546-1726 (Sept. 2017).
14. Chandrasekaran, C. *et al.* Brittleness in model selection analysis of single neuron firing rates. en. *bioRxiv.* Publisher: Cold Spring Harbor Laboratory Section: New Results, 430710 (Sept. 2018).

15. Okazawa, G., Hatch, C. E., Mancoo, A., Machens, C. K. & Kiani, R. Representational geometry of perceptual decisions in the monkey parietal cortex. English. *Cell* **184.** Publisher: Elsevier, 3748–3761.e18. ISSN: 0092-8674, 1097-4172 (July 2021).

16. Seung, H. S. How the brain keeps the eyes still. en. *Proceedings of the National Academy of Sciences* **93,** 13339–13344. ISSN: 0027-8424, 1091-6490 (Nov. 1996).

17. Churchland, M. M. *et al.* Neural population dynamics during reaching. en. *Nature* **487,** 51–56. ISSN: 0028-0836 (July 2012).

18. Murphy, B. K. & Miller, K. D. Balanced amplification: a new mechanism of selective amplification of neural activity patterns. *Neuron* **61,** 635–648. ISSN: 0896-6273 (Feb. 2009).

19. Hennequin, G., Vogels, T. P. & Gerstner, W. Non-normal amplification in random balanced neuronal networks. en. *Physical Review E* **86.** ISSN: 1539-3755, 1550-2376. doi:10.1103/PhysRevE.86.011909. https://link.aps.org/doi/10.1103/PhysRevE.86.011909 (2018) (July 2012).

20. Rigotti, M. *et al.* The importance of mixed selectivity in complex cognitive tasks. en. *Nature* **497,** 585–590. ISSN: 1476-4687 (May 2013).

21. Bondanelli, G. & Ostojic, S. Coding with transient trajectories in recurrent neural networks. *PLoS Computational Biology* **16.** ISSN: 1553-734X. doi:10.1371/journal.pcbi.1007655. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7043794/ (2021) (Feb. 2020).

22. Stroud, J. P., Watanabe, K., Suzuki, T., Stokes, M. G. & Lengyel, M. *Optimal information loading into working memory in prefrontal cortex* en. Tech. rep. (Dec. 2021), 2021.11.16.468360. (2022).

23. Christodoulou, G., Vogels, T. P. & Agnes, E. J. Regimes and mechanisms of transient amplification in abstract and biological neural networks. en. *PLOS Computational Biology* **18.** Publisher: Public Library of Science, e1010365. ISSN: 1553-7358 (Aug. 2022).

24. Goldman, M. S. Memory without Feedback in a Neural Network. English. *Neuron* **61.** Publisher: Elsevier, 621–634. ISSN: 0896-6273 (Feb. 2009).

25. Hennequin, G., Vogels, T. P. & Gerstner, W. Optimal Control of Transient Dynamics in Balanced Networks Supports Generation of Complex Movements. en. *Neuron* **82,** 1394–1406. ISSN: 08966273 (June 2014).

26. O'Shea, D. J. *et al. Direct neural perturbations reveal a dynamical mechanism for robust computation* en. Pages: 2022.12.16.520768 Section: New Results. Dec. 2022. doi:10.1101/2022.12.16.520768. https://www.biorxiv.org/content/10.1101/2022.12.16.520768v1 (2023).

27. Treue, S. & Maunsell, J. H. Attentional modulation of visual motion processing in cortical areas MT and MST. eng. *Nature* **382,** 539–541. ISSN: 0028-0836 (Aug. 1996).

28. Treue, S. & Trujillo, J. C. M. Feature-based attention influences motion processing gain in macaque visual cortex. en. *Nature* **399,** 575–579. ISSN: 1476-4687 (June 1999).

29. Katzner, S., Busse, L. & Treue, S. Attention to the color of a moving stimulus modulates motion-signal processing in macaque area MT: evidence for a unified attentional system. *Frontiers in Systems Neuroscience* **3.** ISSN: 1662-5137. https://www.frontiersin.org/articles/10.3389/neuro.06.012.2009 (2022) (2009).

30. Sasaki, R. & Uka, T. Dynamic readout of behaviorally relevant signals from area MT during task switching. eng. *Neuron* **62,** 147–157. ISSN: 1097-4199 (Apr. 2009).

31. Mirabella, G. *et al.* Neurons in Area V4 of the Macaque Translate Attended Visual Features into Behaviorally Relevant Categories. English. *Neuron* **54.** Publisher: Elsevier, 303–318. ISSN: 0896-6273 (Apr. 2007).

32. Bartsch, M. V. *et al.* Attention to Color Sharpens Neural Population Tuning via Feedback Processing in the Human Visual Cortex Hierarchy. en. *Journal of Neuroscience* **37.** Publisher: Society for Neuroscience Section: Research Articles, 10346–10357. ISSN: 0270-6474, 1529-2401 (Oct. 2017).

33. Barbosa, J. *et al. Flexible selection of task-relevant features through population gating* en. Pages: 2022.07.21.500962 Section: New Results. Oct. 2022. doi:10.1101/2022.07.21.500962. https://www.biorxiv.org/content/10.1101/2022.07.21.500962v2 (2022).

34. Langdon, C. & Engel, T. A. *Latent circuit inference from heterogeneous neural responses during cognitive tasks* en. Pages: 2022.01.23.477431 Section: New Results. Jan. 2022. doi:10.1101/2022.01.23.477431. https://www.biorxiv.org/content/10.1101/2022.01.23.477431v1 (2022).

35. Pandarinath, C. *et al.* Inferring single-trial neural population dynamics using sequential auto-encoders. en. *Nature Methods* **15.** Number: 10 Publisher: Nature Publishing Group, 805–815. ISSN: 1548-7105 (Oct. 2018).

36. Sylwestrak, E. L. *et al.* Cell-type-specific population dynamics of diverse reward computations. eng. *Cell* **185,** 3568–3587.e27. ISSN: 1097-4172 (Sept. 2022).

37. Kao, T.-C. & Hennequin, G. Neuroscience out of control: control-theoretic perspectives on neural circuit dynamics. en. *Current Opinion in Neurobiology. Computational Neuroscience* **58,** 122–129. ISSN: 0959-4388 (Oct. 2019).

38. Kao, T.-C., Sadabadi, M. S. & Hennequin, G. Optimal anticipatory control as a theory of motor preparation: A thalamo-cortical circuit model. English. *Neuron* **109.** Publisher: Elsevier, 1567–1581.e12. ISSN: 0896-6273 (May 2021).

39. Schimel, M., Kao, T.-C., Jensen, K. T. & Hennequin, G. *iLQR-VAE : control-based learning of input-driven dynamics with applications to neural data* en. Tech. rep. Section: New Results Type: article (bioRxiv, Oct. 2021), 2021.10.07.463540. doi:10.1101/2021.10.07.463540. https://www.biorxiv.org/content/10.1101/2021.10.07.463540v1 (2022).

40. Galgali, A. R., Sahani, M. & Mante, V. Residual dynamics resolves recurrent contributions to neural computation. en. *Nature Neuroscience*. Publisher: Nature Publishing Group, 1–13. ISSN: 1546-1726 (Jan. 2023).

41. Feulner, B. *et al.* Small, correlated changes in synaptic connectivity may facilitate rapid motor learning. en. *Nature Communications* **13.** Number: 1 Publisher: Nature Publishing Group, 5163. ISSN: 2041-1723 (Sept. 2022).

42. Linderman, S. *et al. Bayesian Learning and Inference in Recurrent Switching Linear Dynamical Systems* en. in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* ISSN: 2640-3498 (PMLR, Apr. 2017), 914–922. https://proceedings.mlr.press/v54/linderman17a.html (2022).

43. Nair, A. *et al. An approximate line attractor in the hypothalamus that encodes an aggressive internal state* en. Tech. rep. Section: New Results Type: article (bioRxiv, Apr. 2022), 2022.04.19.488776. doi:10.1101/2022.04.19.488776. https://www.biorxiv.org/content/10.1101/2022.04.19.488776v1 (2022).

44. Yang, Y. *et al.* Modelling and prediction of the dynamic responses of large-scale brain networks during direct electrical stimulation. en. *Nature Biomedical Engineering* **5,** 324–345. ISSN: 2157-846X (Apr. 2021).

45. Sani, O. G., Pesaran, B. & Shanechi, M. M. *Where is all the nonlinearity: flexible nonlinear modeling of behaviorally relevant neural dynamics using recurrent neural networks* en. Pages: 2021.09.03.458628 Section: New Results. Sept. 2021. doi:10.1101/2021.09.03.458628. https://www.biorxiv.org/content/10.1101/2021.09.03.458628v1 (2022).

46. Perich, M. G. *et al.* Inferring brain-wide interactions using data-constrained recurrent neural network models. en. *bioRxiv*. Publisher: Cold Spring Harbor Laboratory Section: New Results, 2020.12.18.423348 (Dec. 2020).

47. Genkin, M. & Engel, T. A. Moving beyond generalization to accurate interpretation of flexible models. en. *Nature Machine Intelligence* **2.** Number: 11 Publisher: Nature Publishing Group, 674–683. ISSN: 2522-5839 (Nov. 2020).

48. Pagan, M. *et al. A new theoretical framework jointly explains behavioral and neural variability across subjects performing flexible decision-making* en. Pages: 2022.11.28.518207 Section: New Results. Nov. 2022. doi:10.1101/2022.11.28.518207. https://www.biorxiv.org/content/10.1101/2022.11.28.518207v1 (2022).

49. DeFelipe, J. Brain plasticity and mental processes: Cajal again. en. *Nature Reviews Neuroscience* **7,** 811–817. ISSN: 1471-0048 (Oct. 2006).

50. Machens, C. K., Romo, R. & Brody, C. D. Flexible Control of Mutual Inhibition: A Neural Model of Two-Interval Discrimination. en. *Science* **307,** 1121–1124. ISSN: 0036-8075, 1095-9203 (Feb. 2005).

51. Shenoy, K. V., Sahani, M. & Churchland, M. M. Cortical control of arm movements: a dynamical systems perspective. eng. *Annual Review of Neuroscience* **36,** 337–359. ISSN: 1545-4126 (July 2013).

52. Remington, E. D., Egger, S. W., Narain, D., Wang, J. & Jazayeri, M. A Dynamical Systems Perspective on Flexible Motor Timing. en. *Trends in Cognitive Sciences. Special Issue: Time in the Brain* **22,** 938–952. ISSN: 1364-6613 (Oct. 2018).

53. Marder, E. & Bucher, D. Central pattern generators and the control of rhythmic movements. en. *Current Biology* **11,** R986–R996. ISSN: 0960-9822 (Nov. 2001).

54. Hutcheon, B. & Yarom, Y. Resonance, oscillation and the intrinsic frequency preferences of neurons. en. *Trends in Neurosciences* **23,** 216–222. ISSN: 0166-2236 (May 2000).

55. Soldado-Magraner, S. *et al.* Conditioning by subthreshold synaptic input changes the intrinsic firing pattern of CA3 hippocampal neurons. *Journal of Neurophysiology* **123.** Publisher: American Physiological Society, 90–106. ISSN: 0022-3077 (Nov. 2019).

56. Macke, J. H. *et al.* in *Advances in Neural Information Processing Systems 24* (eds Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F. & Weinberger, K. Q.) 1350–1358 (Curran Associates, Inc., 2011). (2018).

57. Kobak, D. *et al.* Demixed principal component analysis of neural population data. en. *eLife* **5,** e10989. ISSN: 2050-084X (Apr. 2016).

58. Williams, A. H. *et al.* Unsupervised Discovery of Demixed, Low-Dimensional Neural Dynamics across Multiple Timescales through Tensor Component Analysis. *Neuron* **98,** 1099–1115.e8. ISSN: 0896-6273 (June 2018).

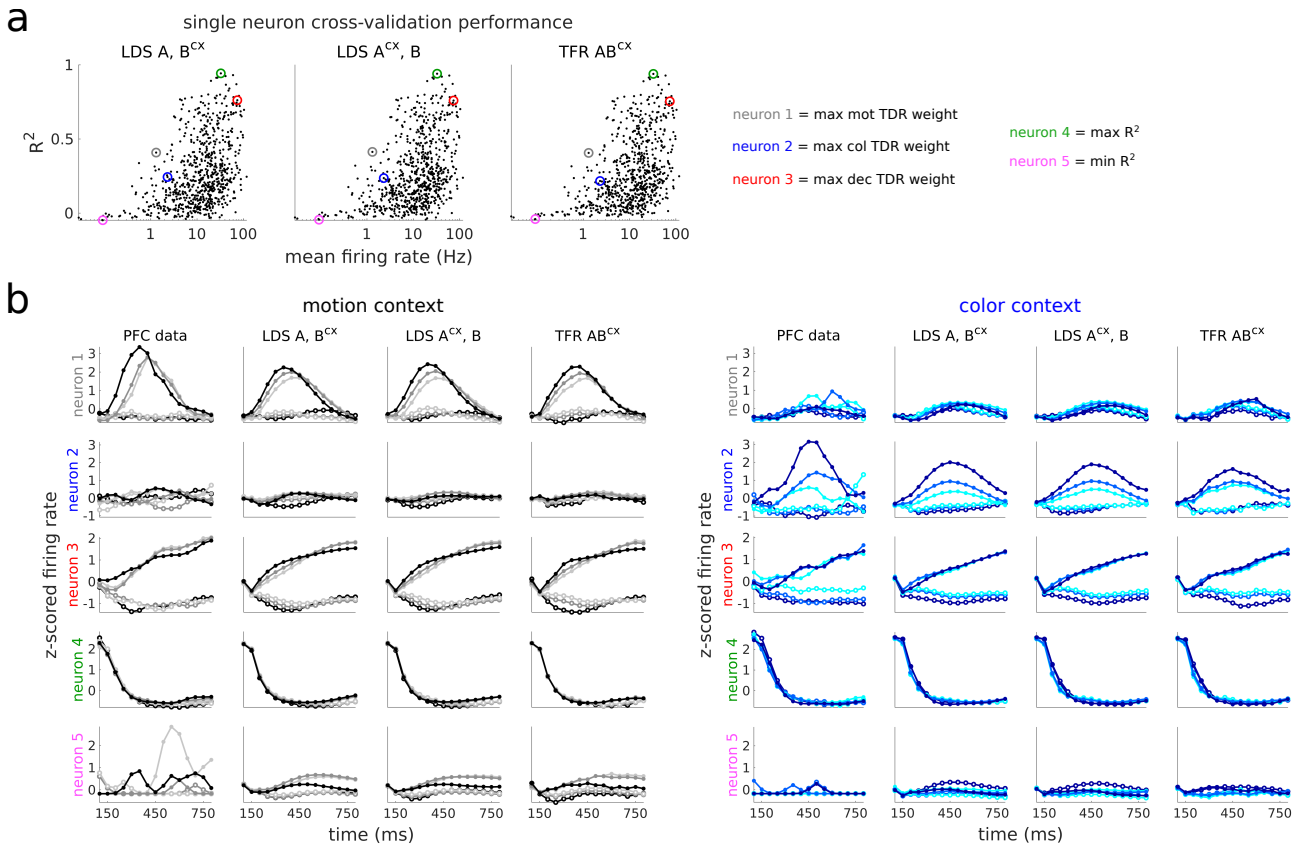
59. Galor, O. *Discrete Dynamical Systems* en. ISBN: 978-3-540-36776-5 (Springer, May 2007).

60. Buesing, L., Macke, J. H. & Sahani, M. Learning stable, regularised latent models of neural population dynamics. eng. *Network (Bristol, England)* **23,** 24–47. ISSN: 1361-6536 (2012).

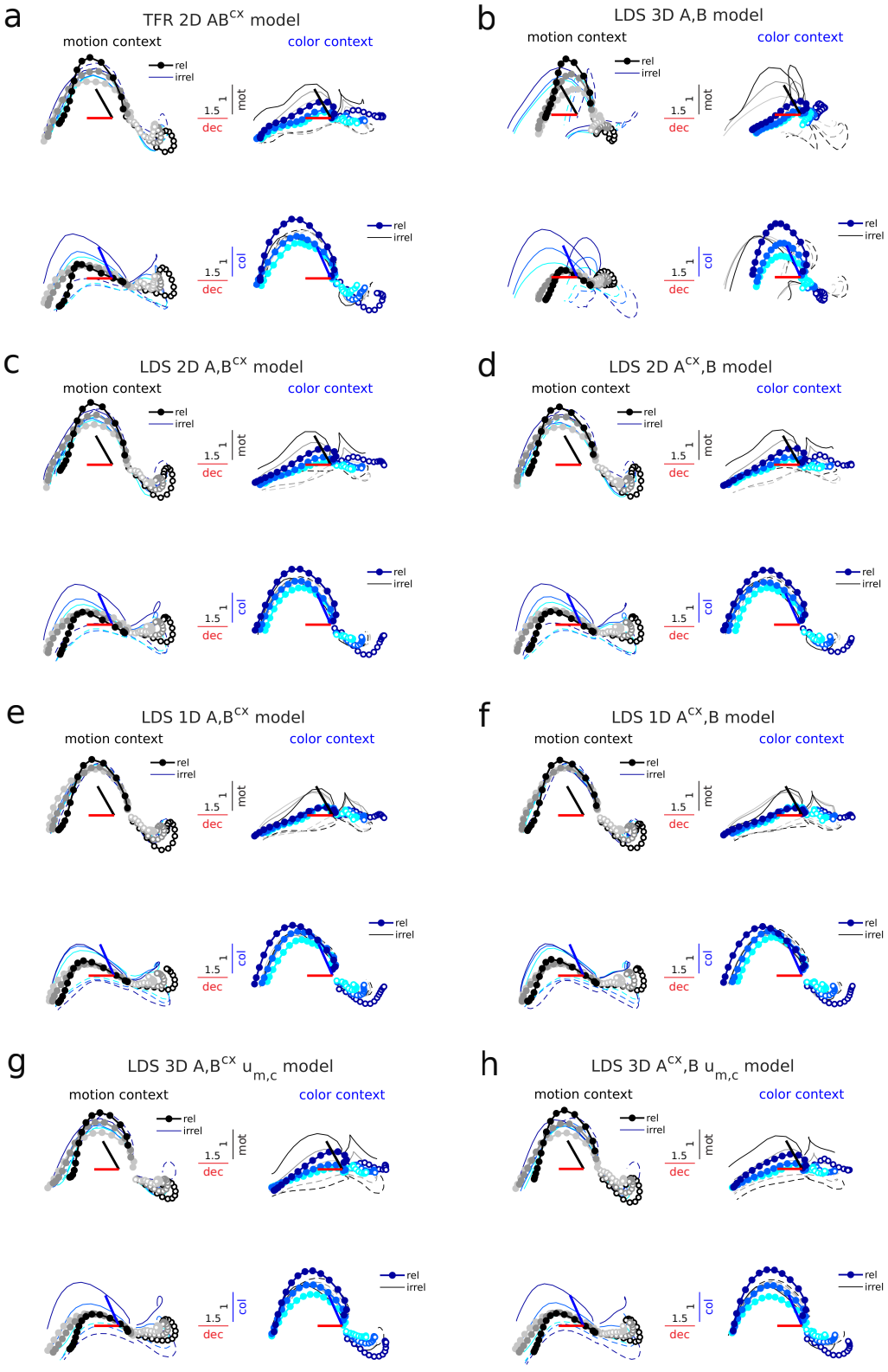
61. Asllani, M., Lambiotte, R. & Carletti, T. Structure and dynamical behavior of non-normal networks. en. *Science Advances* **4.** Publisher: American Association for the Advancement of Science Section: Research Article, eaau9403. ISSN: 2375-2548 (Dec. 2018).

62. Elsayed, G. F., Lara, A. H., Kaufman, M. T., Churchland, M. M. & Cunningham, J. P. Reorganization between preparatory and movement population responses in motor cortex. en. *Nature Communications* **7,** 13239. ISSN: 2041-1723 (Oct. 2016).

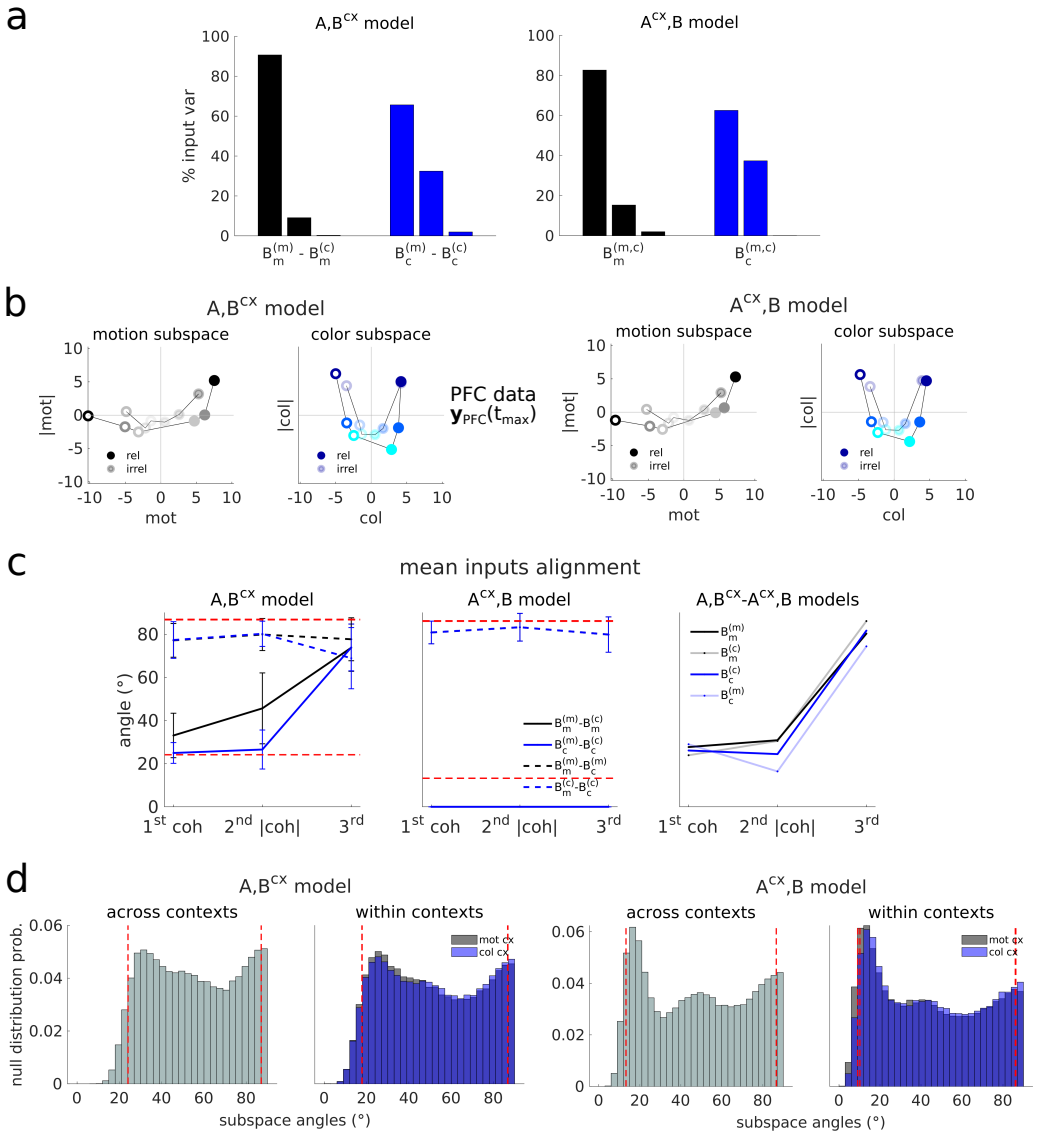
63. Valente, A., Ostojic, S. & Pillow, J. W. Probing the Relationship Between Latent Linear Dynamical Systems and Low-Rank Recurrent Neural Network Models. *Neural Computation* **34,** 1871–1892. ISSN: 0899-7667 (Aug. 2022).



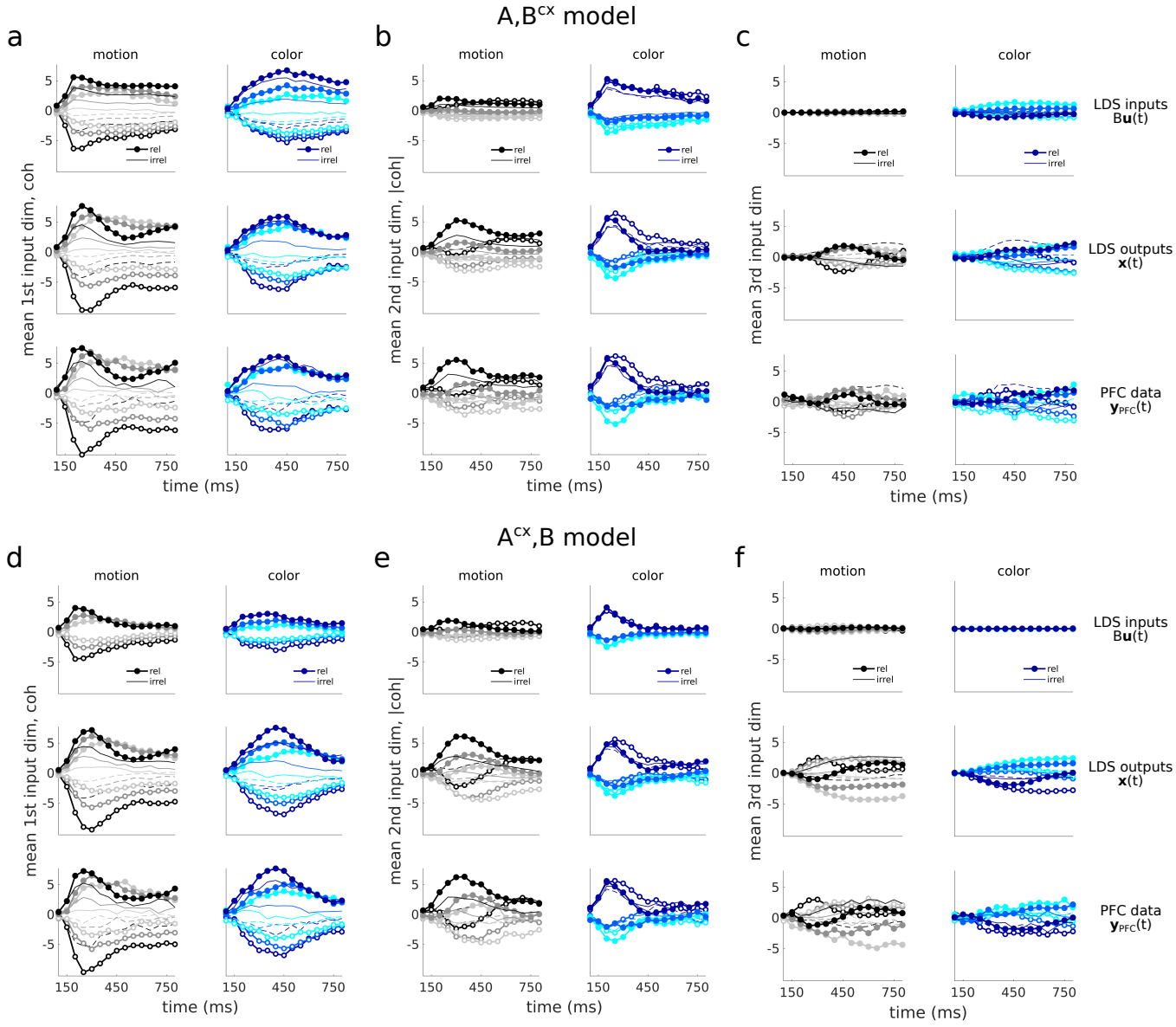
Extended Data Figure 1 . **Individual neurons performance for the LDS and TFR models. a,** LDS and TFR models R^2 for all individual neurons, sorted by their mean firing rate as in Aoi et al.[12]. Highlighted in colors are five example neurons. Three of them have maximum selectivity to either motion, color or decision (in black, blue, red), as measured by their weight onto the motion, color and decision population vectors found using targeted dimensionality reduction (TDR)[1]. The two other neurons were selected based on model performance (best neuron captured, max R^2, in green, and worse neuron captured, min R^2, in pink). **b,** PFC data and LDS/TFR models cross-validated PSTHs for the 5 example neurons. The PSTHs are computed from z-scored data/model responses sorted by the relevant coherence value in each context (motion in the motion context, left, and color in the color context, right) and averaged across irrelevant coherence conditions, as in Mante et al.[1]. Color shades and filled/hollow circles indicate the strength and direction of the coherence evidence, respectively (same notation as in Fig. 1a). PFC data responses have been smoothed with a Gaussian kernel (σ=40ms) for visualization[1]. Data is from monkey A.



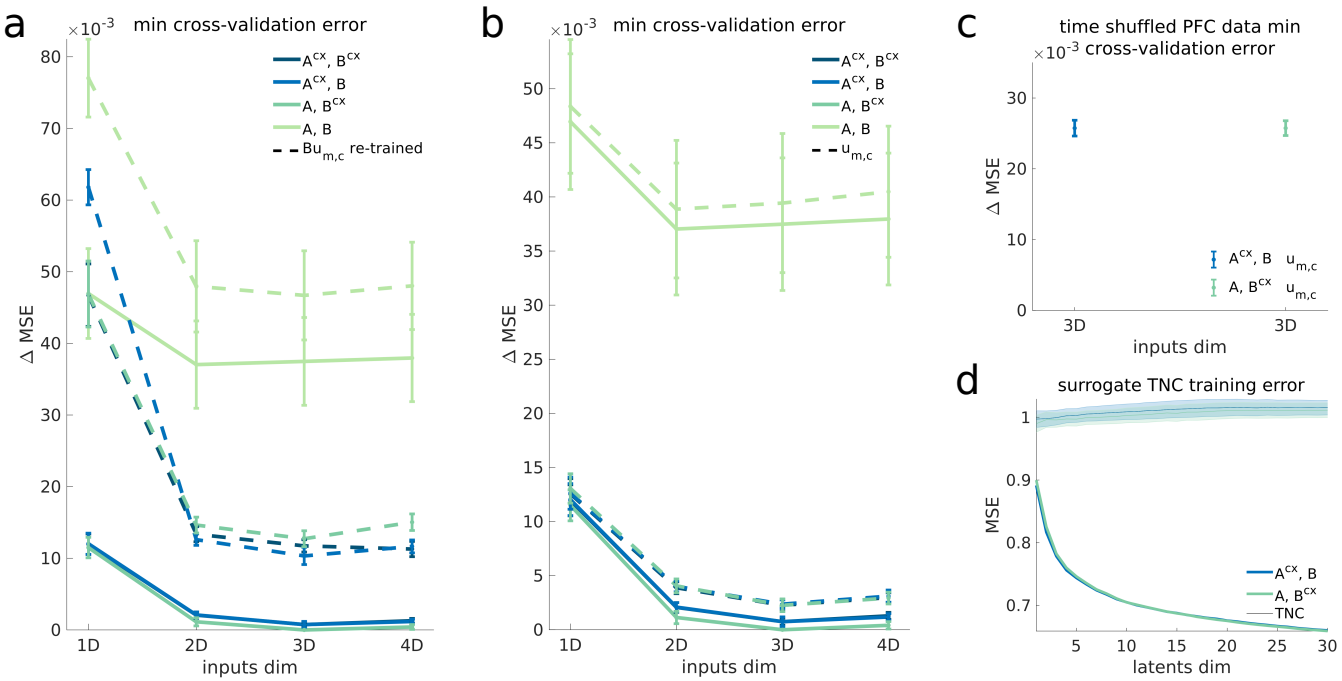
Extended Data Figure 2 . **Population trajectories of the TFR model and alternative LDS models in the task-related subspace.** Cross-validated model trajectories (LOOCV) for the best TFR model and additional LDS models with various input dimensionalities and input/contextual constraints. **a,** TFR AB^{cx} model with 2D inputs. **b,** LDS A, B model with 3D inputs. This model poorly captures the trajectories, specially along the decision dimension. **c,d,** LDS A, B^{cx} and A^{cx}, B models with 2D inputs. **e,f,** LDS A, B^{cx} and A^{cx}, B models with 1D inputs. Note that the trajectories along the input dimensions are not accurately captured, in particular for the irrelevant inputs, where trajectories poorly separate by coherence conditions. **g,h,** LDS $A, B^{cx}\mathbf{u}_{m,c}$ and $A^{cx}, B\mathbf{u}_{m,c}$ models with time-constant 3D inputs. Same conventions as in Fig. 3. All trajectories have been smoothed with a Gaussian filter for visualization (sliding window size, 5-bins). This step did not change much the LDS trajectories, since they are inherently smooth, but it helped smooth-out substantially the TFR model trajectories, given that this model has no dynamical constraints that enforce smoothness (data not shown). Monkey A data.



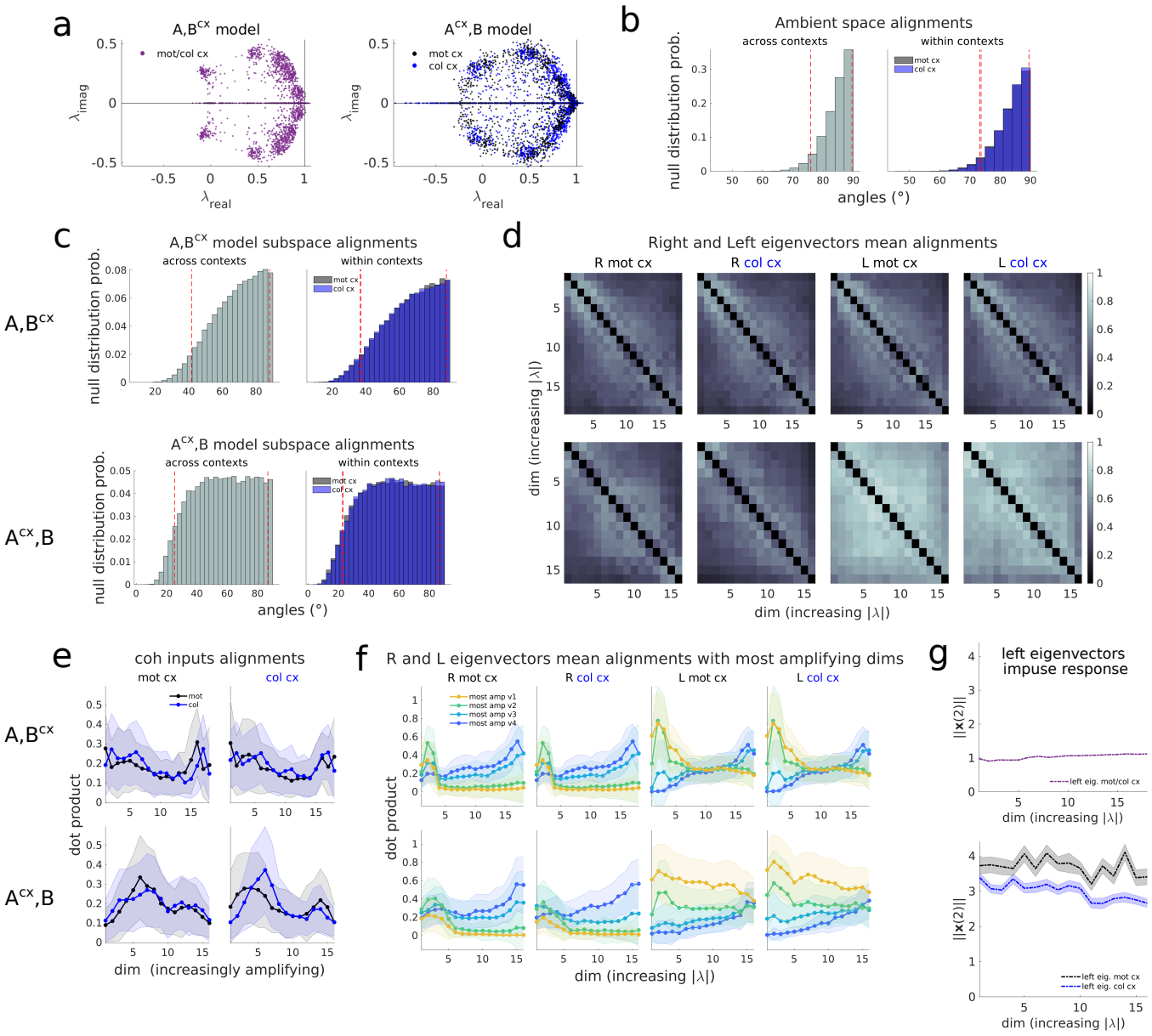
Extended Data Figure 3 . Input variance, PFC data in the 2D input subspaces and alignment statistics. a, External inputs variance in the three input dimensions from both LDS models. For the A, B^{cx} model variance is computed along the mean dimensions across contexts (first taking the average across contexts: mot avg. $B_m^{(m)} - B_m^{(c)}$, for dims 1-3, and col avg. $B_c^{(m)} - B_c^{(c)}$, for dims 1-3, then re-orthogonalizing the three input dimensions). Note that most of the input variance is concentrated in the first two input dimensions (a 2D plane) whereas the third dimension carry almost no input variance. Averages across 100 models. **b,** PFC data at t=250ms for all coherences projected onto the same 2D coh-|coh| input planes as in Fig. 4d,e. The PFC data contains a curved representation of coherence information. CI signals have been subtracted. **c,** Alignment between the motion and color input vectors within contexts (dashed lines), and between the motion or color input vectors across contexts (filled lines), for each of the three input dimensions, and the two LDS models (left and middle panels). Dashed red lines, 5th and 95th percentiles of a null distribution of alignments (see **d**). Error bars, std across a 100 randomly initialized models. Note that the inferred coherence and coherence magnitude dimensions for both motion and color are largely stable across contexts in the A, B^{cx} model (i.e. they are highly aligned, filled lines). However, the across-contexts alignments for the third input dimension are close to random, indicating that this dimension is not common across contexts. Right panel, alignments between the mean input directions (across 100 models) from each LDS model class. **d,** Null distribution of alignments (subspace angles) from randomly sampled 3D subspaces within and across contexts drawn aligned to the data covariance[62], orthonormalized, and then projected onto the low-d subspaces from each LDS model class (defined by the columns of the loading matrices C, left, right panels). The null distributions from the two model classes are different since they learned different C matrices. Random samples s≈33,000 orthonormal subspaces, or 100,000 vectors. Dashed red lines, 5th and 95th null distribution percentiles. Alignments not expected by chance fall in regions $\leq$ the 5th or $\geq$ the 95th percentiles of the control distributions. Additionally, given the binomial nature of the distribution, where both high and low alignments are expected, random alignments should on average lie around $50°$. This is not what is typically obtained from the data (panel **c**). Furthermore, in all 100 models from the A, B^{cx} class, the highest alignments consistently occurred between the two color and two motion dimensions across contexts, and not between the motion and color dimensions within contexts (first panel in **c**, filled vs. dashed lines). This was true only for the first two input dimensions, but not the third, where all alignments are very low. Similarly, the first two mean input dimensions were highly aligned across model classes, but not the third (**c**, third panel). Monkey A data.



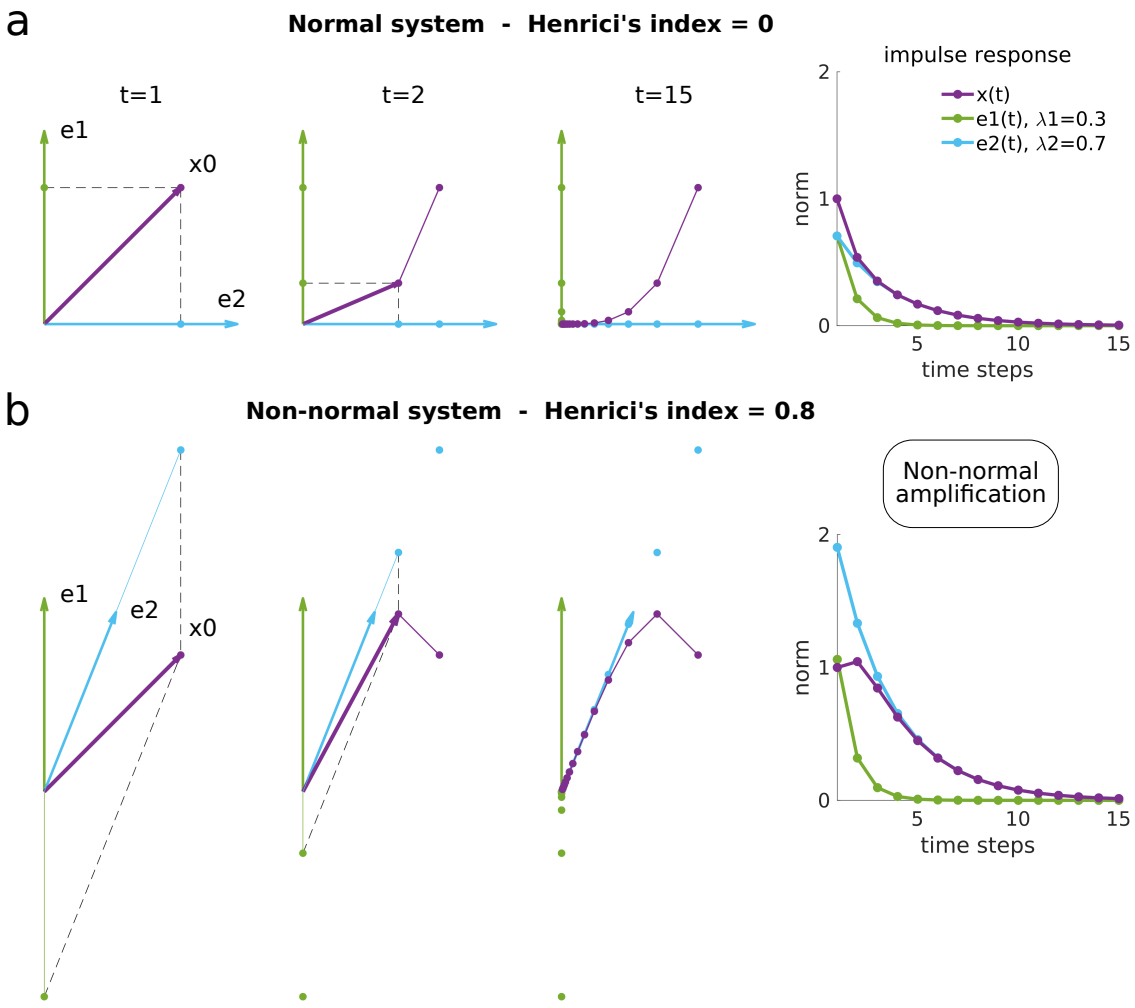
Extended Data Figure 4 . LDS models external inputs, latents and PFC data in the LDS input dimensions. LDS external inputs, LDS cross-validated latents (outputs) and PFC data trajectories projected along the three input dimensions found for the A, B^{cx} (**a-c**) and A^{cx}, B (**d-f**) models, for all coherence conditions and contexts (relevant vs. irrelevant). **a,d,** First input dimension, capturing coherence related variance (coh). **b,e,** Second input dimension, capturing coherence magnitude related variance (|coh|). **c,f,** Third input dimension, orthogonal to the coh and |coh| dimensions, capturing little input and relatively little output/data variance compared to the other dimensions (in particular, for color). For the A, B^{cx} model, projections are shown onto the direction bisecting the two color and two motion directions found for each context. All data is from means over 100 models initialized at random. The three mean color and motion input dimensions are orthogonalized with QR-decomposition. For all trajectories the mean across conditions has been subtracted out to remove condition independent signals (CI). Latents and data trajectories are generated for all 36 task conditions and plotted along the motion/color input dimensions with color/motion conditions averaged out. Same plotting conventions as in Fig. 3. Monkey A data.



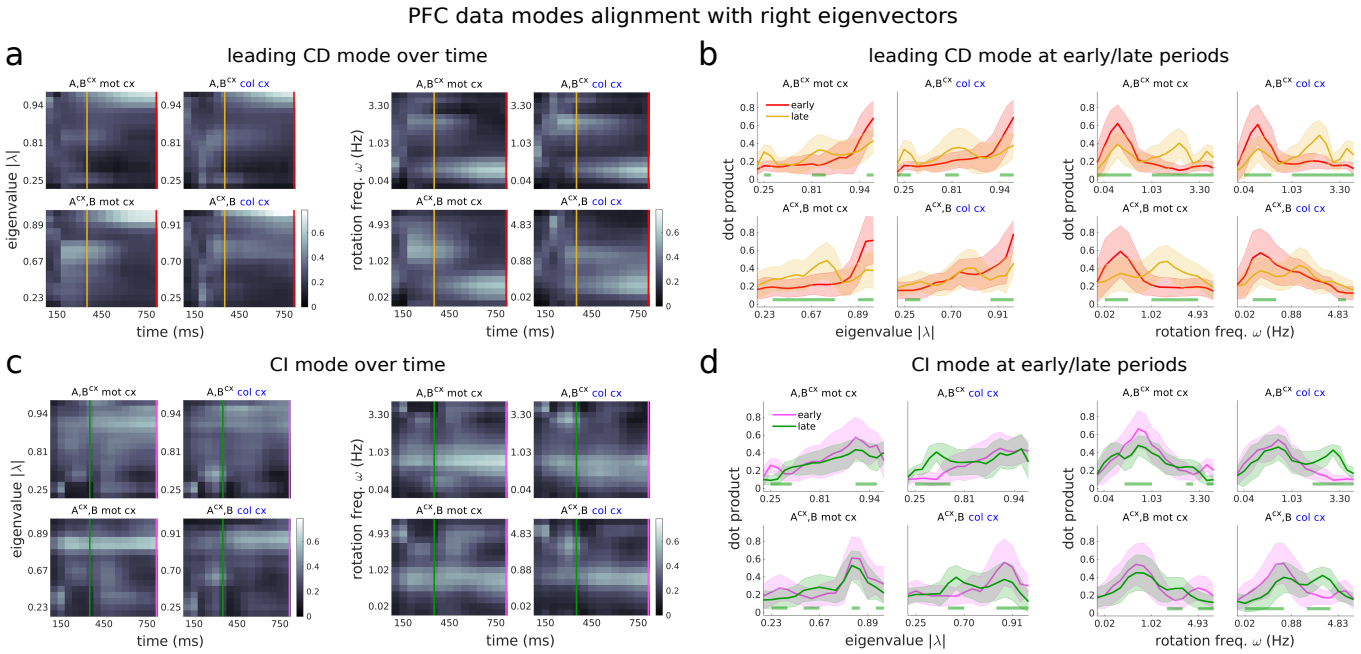
Extended Data Figure 5 . **Performance of LDS models with time-constant inputs and randomized data controls.** **a,** Leave-one-condition-out cross-validation performance (LOOCV) of the time-varying input LDS models in Fig. 2a (filled lines) and the same models after re-training their input parameters $B\mathbf{u}_{m,c}$, but constraining $\mathbf{u}_{m,c}$ to be constant in time, and with the rest of the parameters kept the same (dashed lines). See next panel for performance of a similar model but fully optimized (all parameters re-trained). **b,** Performance of LDS models with $\mathbf{u}_{m,c}$ constant in time where all parameters, including the dynamics matrix, are optimized to fit the data (dashed lines). The time-varying models from Fig. 2a are also shown for reference (filled lines). **c,** Performance of the best time-constant models ($A, B^{cx}\mathbf{u}_{m,c} - A^{cx}, B\mathbf{u}_{m,c}$ 3D models, dashed lines in **b**) when fitted to time-shuffled PFC data. For all three subpanels (**a**-**c**) the minimum cross-validation errors are shown relative to the best performing LDS model (the time-varying A, B^{cx} model with 3D inputs). Error bars indicate the standard error mean across LOOCV folds. Note that the performance of the best time-constant models when fitted to time-shuffled data drops substantially (**c**), being worse than the 1D input models and nearly as bad as the most contextually constrained A, B models (since Δ MSE = 26, see **b** for reference). **d,** Training performance of the best time-varying LDS models on surrogate data sets randomized across time, neurons and conditions (TNC), but designed to preserve the primary statistics of the data[13]. To obtain the randomized TNC data sets the tensor maximum entropy method (TME) was used. Shades indicate standard error mean across 30 surrogates. Monkey A data.



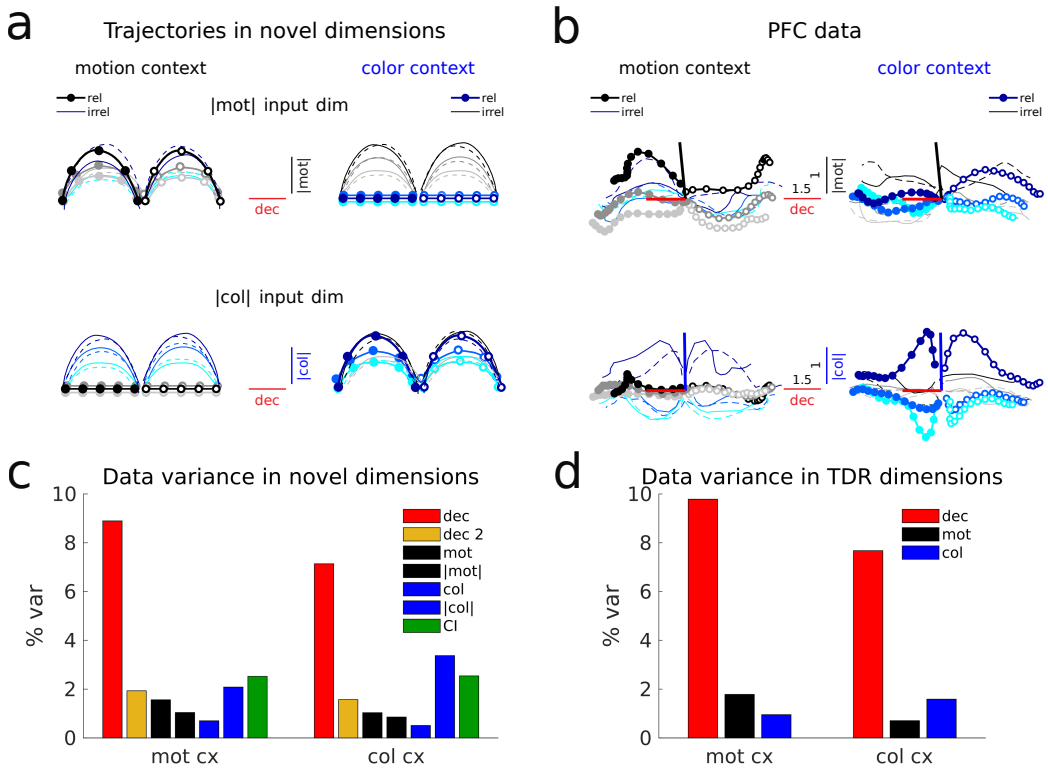
Extended Data Figure 6 . **LDS models dynamics properties. a,** LDS eigenspectrums for the 100 models from each model class. **b,c,** Null distribution of alignments for randomly sampled pairs of vectors, within contexts or across contexts, drawn aligned to the data covariance[62] in the ambient (high-d) space (**b**) or projected onto the low-d LDS subspaces from each model class (**c**, defined by the columns of the loading matrices C). Random samples s=100,000 vectors. Dashed red lines, 5th and 95th null distribution percentiles. Alignments not expected by chance fall in regions $<$ the 5th or $>$ the 95th percentiles of the control distributions. **d,** Mean alignments among the left/right eigenvectors from each model class (A, B^{cx}, top, A^{cx}, B, bottom, averages across 100 models from each class). **e,** Motion and color input coherence vector alignments with respect to dynamic dimensions of various degrees of amplification, sorted from the least to the most amplifying modes (Methods), for the two contexts (left and right panels) and the two LDS model classes (top, bottom). Mean $\pm$ std across 100 models. Motion and color inputs do not strongly align to the most amplifying dimensions. Note that in Fig. 5c we found that coherence inputs were strongly loaded onto the relatively fast decaying left eigenmodes. Indeed, these intermediate left eigenmodes do not align particularly strongly to the most amplifying modes, compared to the alignments for the fastest and the slowest eigenmodes (see **f**, L panels). **f,** Right (R) and Left (L) eigenvectors alignments with respect to the four most amplifying modes of the dynamics (Methods), for both models and both contexts. Mean $\pm$ std across 100 models. Note that all left eigenvector dimensions, from fast, to intermediate, to slow, have moderate to strong alignments with the most amplifying modes. In fact, all left eigenvector directions amplify inputs similarly (see next panel) **g,** Impulse response along each left eigenvector direction, measured at the time right after the perturbation (t=2), which was unit norm. The state at this time indicates the degree of transient amplification immediately after the pulse (see also Fig. 6a for response over time, averaged across all left eigenvectors). The state norm at t=2 is slightly bigger than 1 for all A, B^{cx} left eigenvectors, indicating that all these directions slightly amplify inputs. For the A^{cx}, B, in both contexts, the impulse response is much bigger than one for all left eigenvectors, indicating that all these directions strongly amplify inputs. This confirms that the dimensions where the inputs are mostly loaded, the intermediate or relatively fast decaying dimensions, are not particularly amplifying, relative to the fastest and slowest dimensions. Monkey A data.



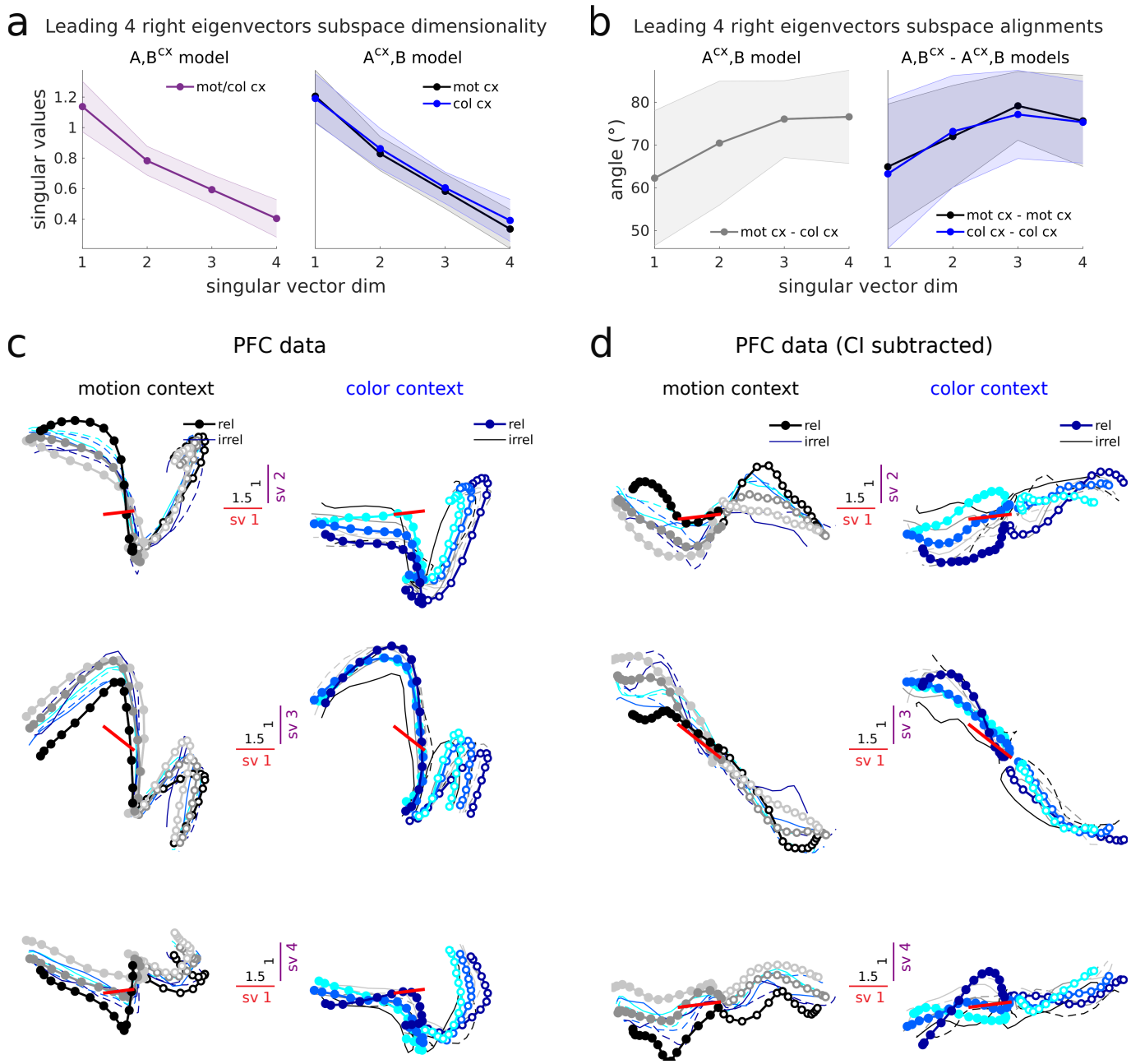
Extended Data Figure 7 . **Non-normal transient amplification of inputs.** Impulse response behavior of a 2D linear dynamical system with normal dynamics (**a**) and a similar but highly non-normal system (**b**). By construction, the normal system had orthogonal eigenvectors ($e1$, $e2$, left panels, in green and blue) and the non-normal system non-orthogonal, in this case set to be closely aligned to each other. In both systems $e1$, $e2$ correspond to the two right eigenvectors of the dynamics. Two additional left eigenvectors exist in the non-normal system, distinct from the right ones, but these are not shown since they are not crucial to understand this picture. In a normal system the right and left eigenvectors are the same (see Methods). Both models were set to have identical eigenvalues, one being small ($\lambda = 0.3$, fast dynamics, in green) and the other large ($\lambda = 0.7$, slower but also decaying dynamics, in blue). The impulse response of the system was defined as the dynamics of the system state $x(t)$ under a pulse input (or unit norm perturbation). The magnitude of the response was measured by the state norm over time after the pulse $||\mathbf{x}(t)||$ (right panels). The perturbation was given along a direction bisecting the plane spanned by the two right eigenvectors in the normal model ($x0$, left panels). For the normal system (**a**), the projection of the state onto the eigenvectors' orthonormal basis (dashed lines) at each time step gives the evolution along the dynamic modes (dots on $e1$ and $e2$ directions). For the non-normal system (**b**), the state cannot be decomposed using an orthogonal projection since the eigenvector basis is not orthogonal. Instead, the eigenvector components are given by the linear combination coefficients of two non-orthogonal basis vectors (here reconstructed using the parallelogram vector addition rule, dashed lines, middle panels). After the pulse, activity along the dynamic modes $e1$ and $e2$ decays exponentially for the normal system (right panel in **a**, in green and blue, given by the norm of the green and blue vectors in the middle panels, whose length at each time step is marked by dots), and so does the state norm (in purple). For the non-normal systems, the dynamic modes also decay exponentially (right panel in **b**, in green and blue). However, the state norm experiences a transient increase, followed by exponential decay (in purple), as a consequence of the non-orthogonality in the state-vector decomposition and the difference in decay rates of the two eigenmodes (left panels in **b**). This phenomenon is known as non-normal or transient amplification.



Extended Data Figure 8 . PFC integration phases for both LDS models across contexts. Same as in Fig. 7 but for both models and contexts. **a,b,** The two distinct integration phases are consistently observed across model classes (A, B^{cx} and A^{cx}, B), contexts (mot and col cx) and model instantiations (**b**, the distribution of alignments across 100 randomly initialized models are fairly narrow, mean $\pm$ std). Note that for the A^{cx}, B model in the color context the early and late alignment distributions are not significantly different along the intermediate set of eigenmodes (at significance level p$<$0.001, **b**), unlike for the other cases. However, the early distribution clearly peaks around the intermediate modes, which have relatively fast decaying dynamics (**b**, left) and fast rotations (**b**, right). The early and late alignment distributions are always significantly different along the last modes, which have the slowest dynamics (**b**, left) and very small rotation frequencies (**b**, right). **c,d,** CI signals are integrated along a different set of slow modes than CD signals, consistently across models and contexts (The peaks of the CI distributions, **d**, lie in different ranges of slow eigenvalues than the CD distribution peaks, **b**). The integration of CI signals do not clearly separate in two phases given that the alignments are largely steady across the trial (**c**). Indeed, the distribution of alignments early vs. late largely overlap (**d**, green and magenta distributions). Monkey A data.



Extended Data Figure 9 . **PFC data trajectories in coherence magnitude dimensions and data variance in the novel and TDR dimensions.** **a,b,** Same as in Fig. 8, but for the LDS inferred coherence magnitude dimensions (averaged across contexts and models). The condition independent (CI) variance has been subtracted to the trajectories in these panels to emphasise input-related variance. **c,** Variance in the novel decision, secondary decision, motion coherence, motion coherence magnitude, color coherence, color coherence magnitude and condition independent (CI) dimensions. Note that the variance that is reflected in each dimension is the total variance, and not the isolated task-related variance components. For instance, there is substantial CI variance in the input dimensions, which we removed in panel **b** to emphasize input-related features. **d,** Variance in TDR decision, motion and color dimensions. Monkey A data.



Extended Data Figure 10 . Choice signals in the slowest LDS subspace largely evolve along a single dimension across contexts. a, The four slowest dimensions inferred by the LDS models have very long time constants ($|\lambda| > 0.9$, time-constant $\tau > 475$ms, average across 100 models, Fig. 5b), so they are expected to define a 4D subspace that contains highly persistent signals. However, the associated right eigenvectors are not orthogonal (Extended Data Fig. 6d), so these signals could occupy fewer than four dimensions. We measured the effective dimensionality of the subspace spanned by the four slowest right eigenvectors by taking the SVD of the matrix containing them. We found that the right eigenvectors effectively span four dimensions, since all singular values are well above zero. Mean $\pm$ std across 100 models. The dominant dimension (1st singular vector) is also the one most highly aligned across contexts in the A^{cx}, B model and also across models (see next panel). **b,** Alignments of the slowest subspace dimensions found across contexts in the A^{cx}, B model (left panel) and across models (right panel). Mean $\pm$ std across 100 models. The alignments are moderate to weak and expected by chance from a control distribution of random vector alignments in the low-d LDS subspaces (alignments are within the 5th and the 95th percentiles of the control distributions, Extended Data Fig. 6c). However, the largest alignments occur precisely for the first dimension. **c,** PFC data projected in the four singular vector (sv) dimensions of the slowest subspace (averaged across models and contexts, and then orthonormalized). Importantly, the first dimension captures decision information, but dimensions 2 to 4 mostly capture condition independent (CI) variance, and some contextual variance (in particular for sv2, where the whole trajectories are slightly shifted vertically). **d,** Same as **c** but after subtracting CI signals. Red bars show the alignment of the decision dimension found by TDR with respect to the four averaged singular value dimensions (sv1: $36°$, sv2: $86°$, sv3: $65°$, sv4: $86°$). The first singular value dimension highly aligns with the decision dimension ($36°$), and the third one moderately aligns to it ($65°$). These alignments are not expected by chance, considering a control distribution of random alignments in the ambient (high-d) space (alignments $<$ 5th percentiles of the control distributions, Extended Data Fig. 6b). The 2D subspace spanned by these two singular value dimensions (middle panels) contains trajectories mainly evolving along a single dimension, when CI signals are subtracted out (compare to middle panels in **c**). This dimension is common across contexts and strongly aligns to the decision axis found by Mante et al. (red bars). Monkey A data.

Supplementary Information

Inferring context-dependent computations through linear approximations of prefrontal cortex dynamics

1. **Supplementary Notes 1–3**
2. **Supplementary Figures 1–10**
3. **Supplementary Tables 1–5**

1 Supplementary Notes

1.1 Additional LDS model-fitting controls

To rule out the possibility that the LDS models were learning dynamically complex external input signals $\mathbf{u}_{m,c}(t)$ to capture the data, we re-trained all input parameters and constrained $\mathbf{u}_{m,c}(t)$ to be constant in time $\mathbf{u}_{m,c}$. This resulted in a relatively small drop in performance (Extended Data Fig. 5a, for models with input dimensionalities higher than 2D). In particular, the performance of re-trained models with 3D constant inputs (dashed lines) dropped to the level of the 1D time-varying input models (filled lines). Newly fitted models with $\mathbf{u}_{m,c}$ constant in time, but where all parameters were optimized, performed nearly as well as the time-varying models (Extended Data Fig. 5b), and accurately captured the PFC trajectories (Extended Data Fig. 2g,h). Notably, in these two control model classes the optimal input dimensionality was consistently 3D, and the latent dimensionalty was also close to the time-varying input models' dimensionality (A^{cx}, B - A, B^{cx} xdim = 16-22 for the time-varying models with inputs retrained to be time-constant; 15-16 for the newly optimized time-constant models and xdim=16-18 for the original time-varying models). Most importantly, the time-constant input models could only rely on their recurrent dynamics to capture the temporal complexity of the PFC data. Therefore, the complexity of the PFC responses is well approximated by linear dynamics and is not necessarily inherited from the external inputs' dynamics.

We also quantitatively assessed whether the time-constant input models were indeed capturing any temporal structure in the data. For this, we asked how well these models performed on time-shuffled data, which had no correlational structure across time. If there was no drop in performance, this would indicate that the time-constant input models were uniquely capturing time-unrelated correlational structure, such as correlations across neurons and conditions. The performance of the best time-constant models (which had 3D inputs, Extended Data Fig. 5b) dropped substantially, being worse than the 1D input models and nearly as bad as the most contextually constrained A, B models (Extended Data Fig. 5c). This indicates that the simple, time-constant LDS models were indeed capturing time-related structure in the PFC data. However, these models still captured a substantial fraction of the time-shuffled data variance (24% on shuffled data vs 27% on the original data). This suggests that the LDS models might be primarily capturing correlational structure across neurons and conditions, besides time-related variance. Indeed, surrogate data sets randomized across conditions, neurons and time were very poorly captured, even by the best

time-varying LDS models (Extended Data Fig. 5d). These data sets were designed to preserve the primary statistics of the data, and thus these results also indicate that the LDS models were not merely capturing basic features of the data[13].

1.2 Understanding non-normal transient amplification

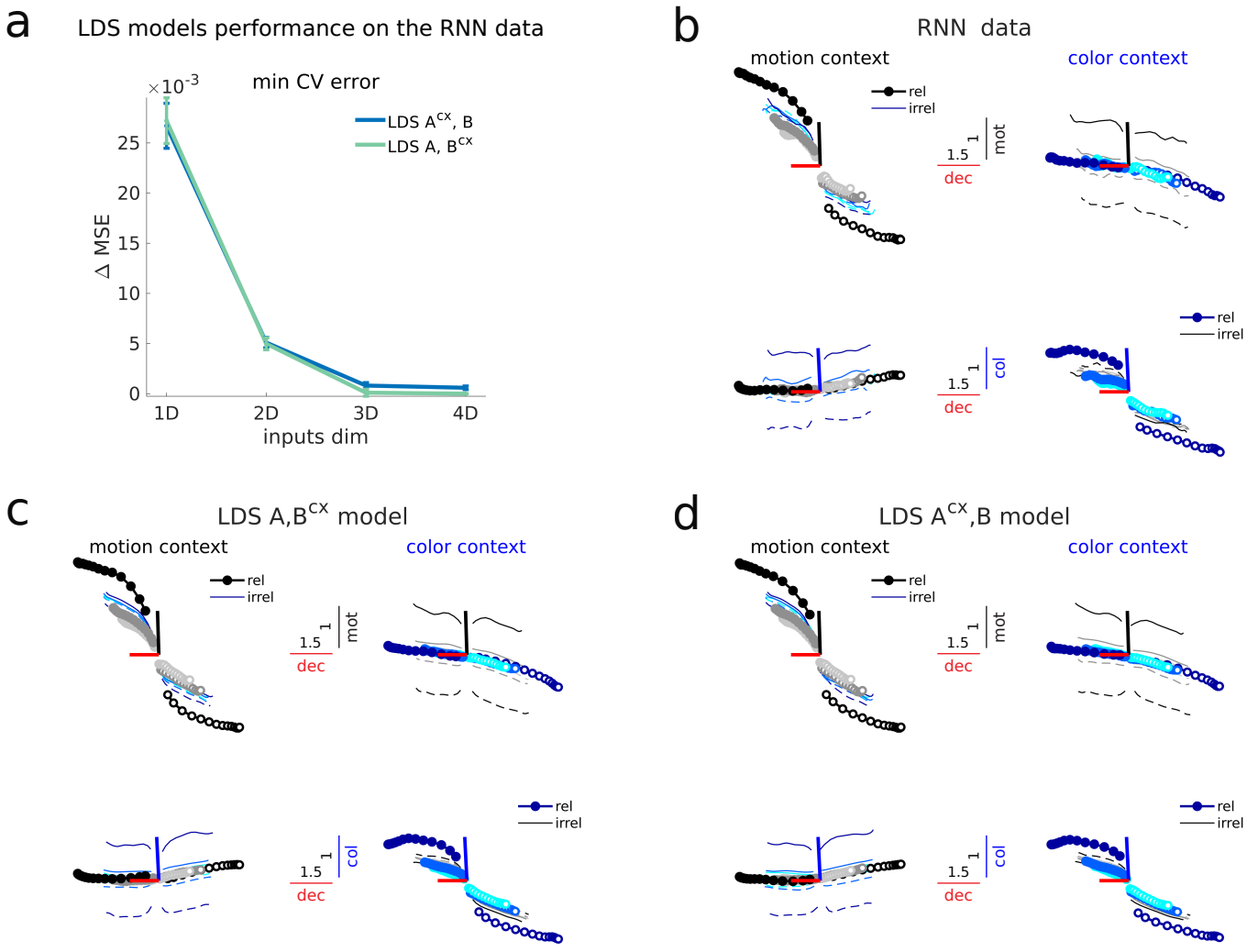
Transient amplification is a property of dynamical systems that have non-normal dynamics matrices[18,19]. These systems present non-trivial dynamical properties that are not predicted by their steady state behavior. In particular, such systems can transiently amplify inputs before decaying to a steady state. To illustrate how the transient amplification mechanism takes place, we built two simplified dynamical system models with identical specifications and only two dimensions, and made one normal (Extended Data Fig. 7a, degree of non-normality or Henrici's index=0, see Methods) and the other highly non-normal (Extended Data Fig. 7b, Henrici's index=0.8). In Extended Data Fig. 7a,b, we show the two right eigenvectors of the dynamics ($e1$, $e2$, in green and blue), for each system, which define the "output" dimensions were activity evolves over time. In the normal system, by definition, these are orthogonal vectors. In the non-normal system, they need not be, as it is the case in this example. Two additional left eigenvectors exist, but in a normal system they are the same as the right eigenvectors (see Methods). In a non-normal system the left eigenvectors are distinct from the right ones (see Methods), but these are not shown here since they are not crucial to understand this picture. We set one of the eigenvalues of the system to be small ($\lambda = 0.3$, fast dynamics) and the other large ($\lambda = 0.7$, slow dynamics) in both models. We then analysed the impulse response properties of both systems (for t=15 time steps, as in the PFC data). For this, we provided an input pulse of unit norm along a random direction $x0$ (in this case, for illustration purposes, a direction that bisected the plane spanned by the two right eigenvectors in the normal system, left panels in Extended Data Fig. 7a,b, in purple). We then looked at how each system processed such input by looking at the evolution of the system state over time $x(t)$ and its norm (middle and right panels, in purple). In the normal system, the input decayed exponentially towards its steady-state, at zero, as expected from a dynamical system with real eigenvalues smaller than one (right and middle panels). The evolution of the two modes of the dynamics was governed by the time-constant of the eigenvalues (right panel). The smallest eigenvalue mode decayed faster (in green) and the largest slower (in blue), with the slower eigenmode largely determining the evolution of the state at later stages (in purple). The evolution along the two dynamic modes could be obtained by projecting the state at each time step onto the orthonormal basis defined by the two right eigenvectors (left and middle panels). The non-normal system, however, had non-orthogonal eigenvectors. In particular, we purposely set the eigenvectors to be closely aligned to each other, to obtain non-normal dynamics. This meant that the system state could no longer be decomposed using an orthogonal projection. Instead, the state was constructed from a linear combination of a non-orthogonal eigenvector basis (left and middle panels, were the parallelogram vector composition rule is applied for this). We provided the same input perturbation as for the normal system. Despite the similarities in design (in particular, having the same eigenvalues), the impulse response properties of this system were very different. Most notably, right after the pulse, there was a transient increase in the state norm (right and middle panels). However, the state eventually decayed to zero. This is because the long-term behaviour of a non-normal system is still governed by its eigenvalues. The smallest eigenvalue mode decayed faster and the largest slower, and they did so exponentially, as in the normal system. Yet, initially a transient

amplification effect was observed. This happens precisely because of the difference in the decay rates of the modes, combined with the fact that the state is reconstructed with a non-orthogonal eigenvector basis (left and middle panels, note how the state vector, in purple, is reconstructed at each time step). The more aligned the eigenvectors are, and the stronger the difference between their eigenvalues, the larger the degree of non-normal amplification. Note that because of the non-trivial decomposition of the state along the non-orthogonal basis, the modes' initial norm was very large (left and right panel, specially for the mode in blue). This resulted in the state experiencing a more sustained decay than in the normal system (right panel, compare blue and purple lines in both systems), i.e. in the input pulse to be transiently "persistent".

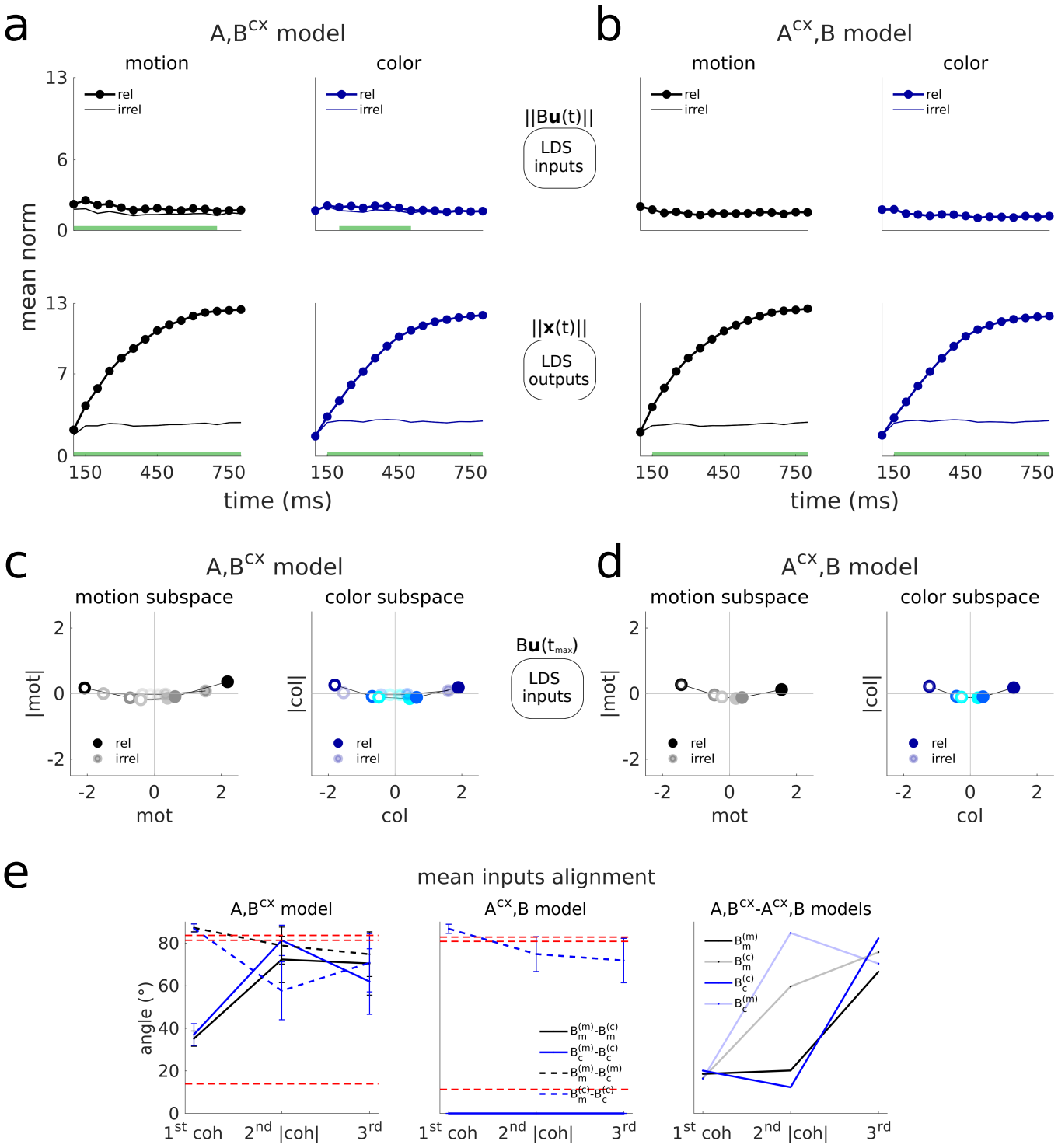
1.3 Alternative input modulation mechanism: fixed input subspaces with context-dependent external input signals

The A, B^{cx} mechanism can be implemented even when the input dimensions are fixed across contexts B. This is because the direction of the input vectors within the input subspaces $B\mathbf{u}$ can still be changed, provided that the external inputs are context-dependent $\mathbf{u}^{cx}$, and that the dimensionality of the input subspaces is higher than 2D. This can be achieved by changing the external inputs' strength in each context along each of the input dimensions (i.e. for 2D inputs, u_1^{cx} and u_2^{cx}). This effectively changes the input vector $B\mathbf{u}$ coordinates within the orthonormal input basis B. This change of coordinates can be used to rotate and stretch the inputs within the input subspace differently across contexts (i.e. u_1^{cx} and u_2^{cx} can be set to define any vector in the 2D input plane). This contextual mechanism might be closer to biology, given that long-range input projections are known to be anatomically stable.

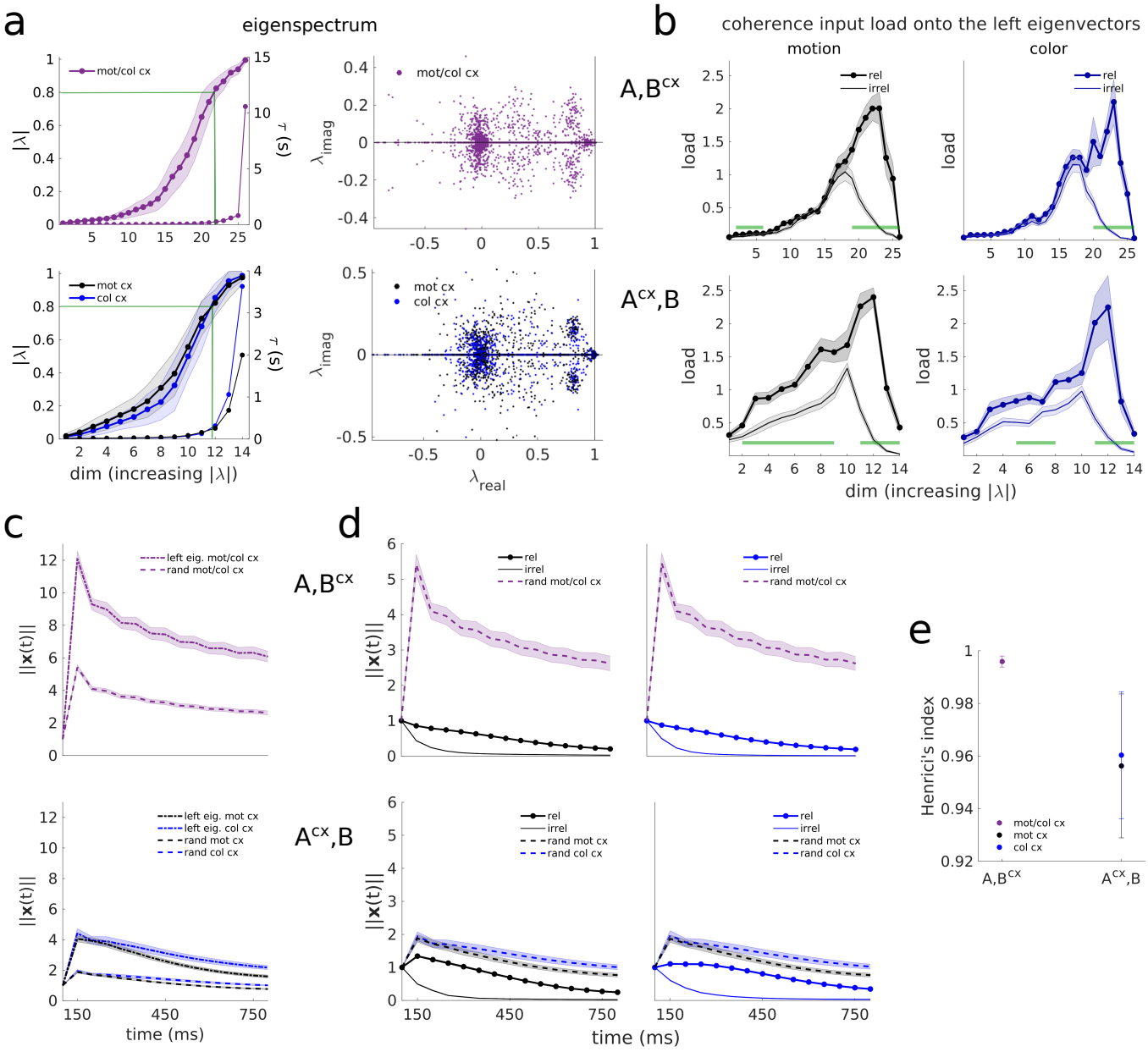
2 Supplementary Figures



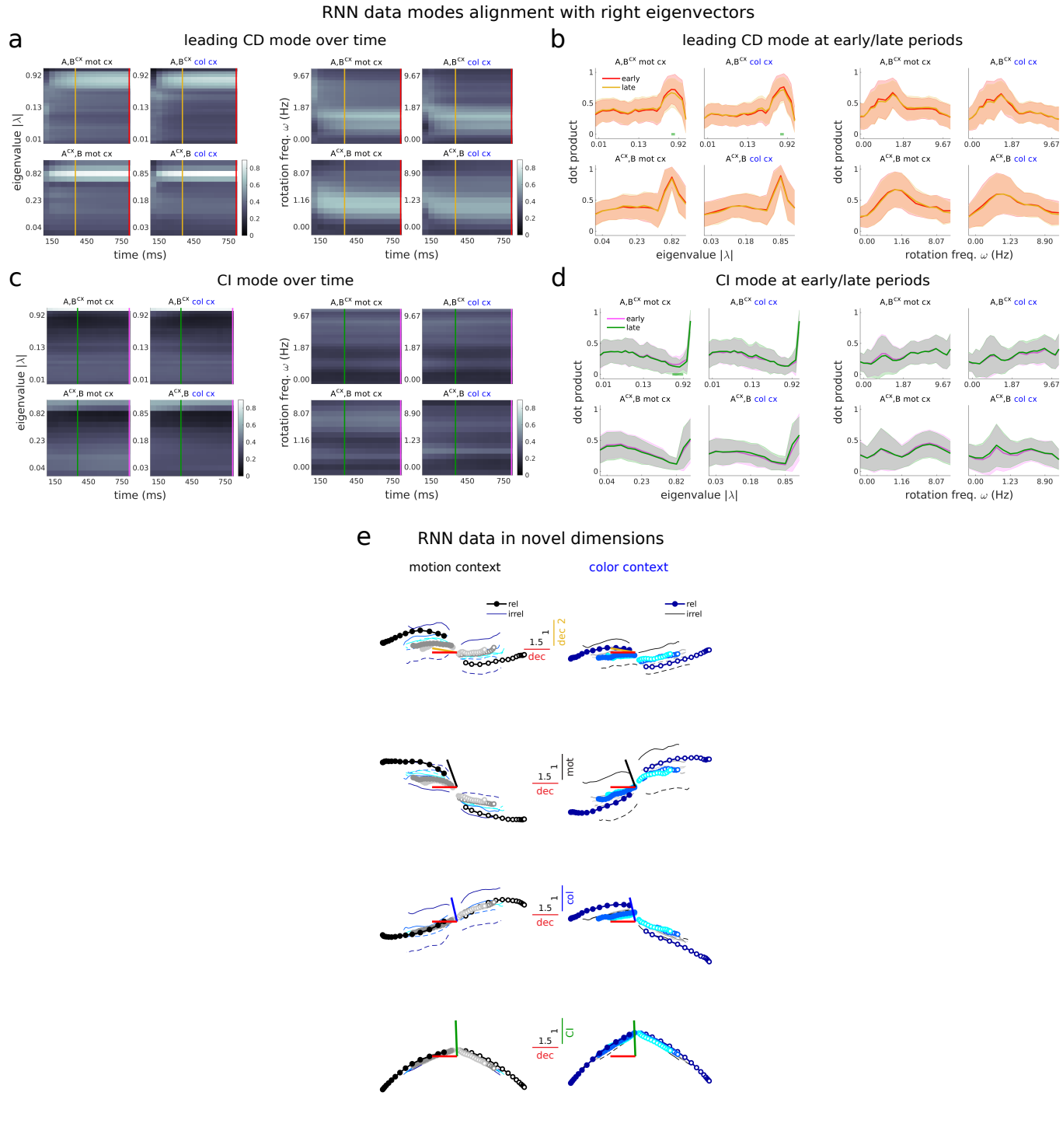
Supplementary Figure 1 . **The LDS models accurately capture the RNN data. a,** Same as Fig. 2a, but for the A, B^{cx} and A^{cx}, B models fitted to the RNN data. In this case the best performing model was the A, B^{cx} model with 4D inputs. However, the performance of the A, B^{cx} model with 3D inputs was not significantly different (Supplementary Table 5). Therefore, in the RNN data the best performing models required at least 3D inputs, as for the PFC data (Fig. 2a). This was greater than the actual input dimensionality of the RNN model, which had 1D inputs (Fig. 1c), but see Supplementary Fig. 2. The two LDS model classes performed equally well (Supplementary Table 5), as in the PFC data (Fig. 2a), but the optimal latent dimensionality for the 3D A, B^{cx} model was 26 and for the A^{cx}, B model 14 (Supplementary Table 5). The difference in dimensionality between the two model classes is larger than in the PFC data (18D vs. 16D, Supplementary Table 1). This suggests that the A, B^{cx} model struggles to capture the RNN data, since it needs more parameters. Note that the error for the two LDS models is close to zero (Supplementary Table 5, MSE$\approx$0, or $\approx 0\%$ of variance missed), unlike for the PFC data from both monkeys (MSE$\approx$0.73, or $\approx 73\%$ of variance missed, Supplementary Tables 1 and 3). Thus, the LDS models are able to approximate near perfectly the dynamics of the non-linear RNN in each context. **b-d,** Same as Fig. 3b-d but for the RNN data. Trajectories in the RNN task-related subspace are well captured by the both LDS models. Axes are defined by the RNN motion and color input vectors and the output vector (the decision readout) after training[1]. Note that the A, B^{cx} model can capture well the input variance along the RNN input dimensions, which by design are fixed across contexts (Fig. 1c). In spite of this, this LDS model must have changed the inputs to achieve contextual integration, since the variance along the decision axis is also well captured.



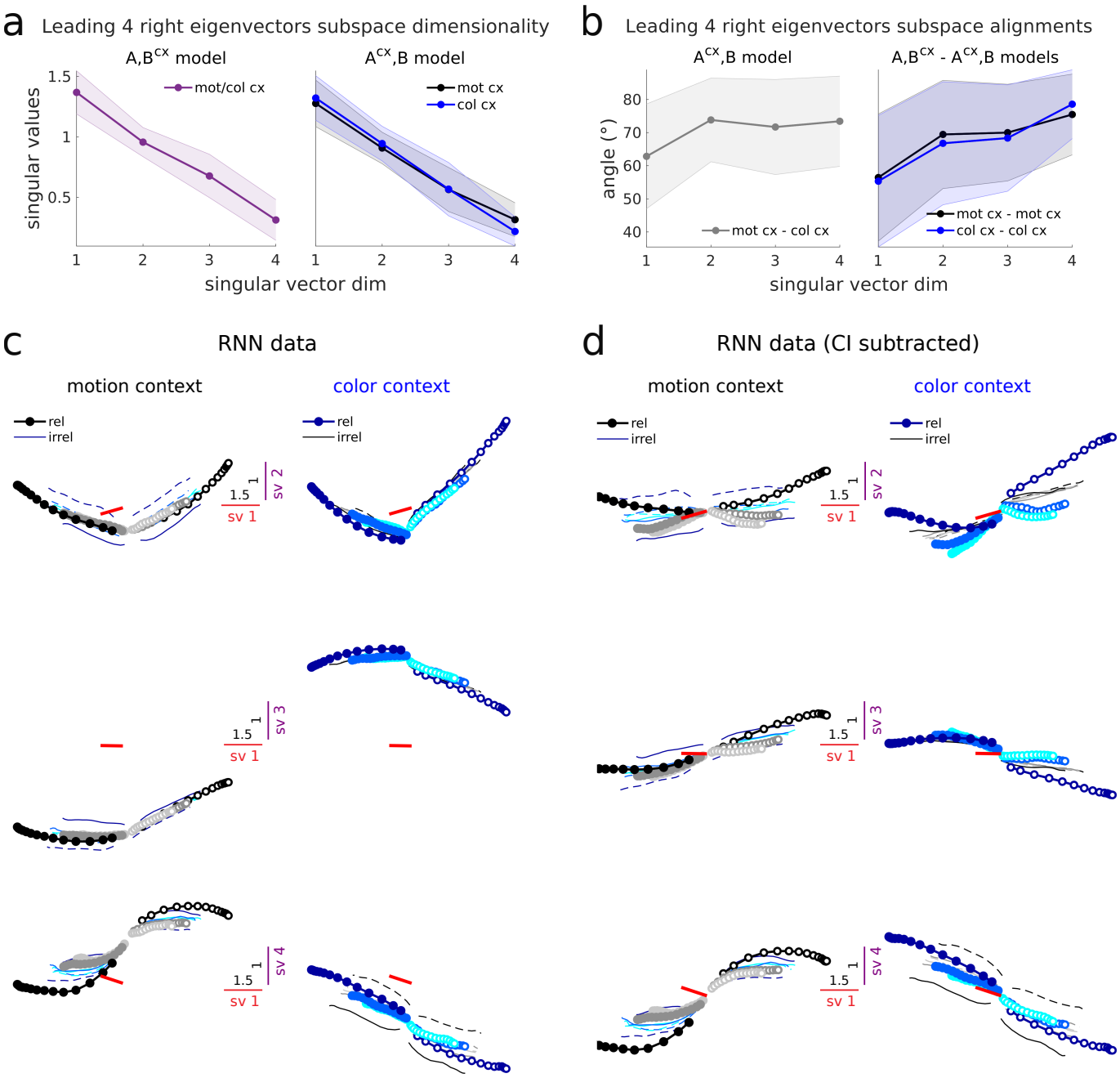
Supplementary Figure 2 . The LDS inputs inferred from the RNN data are largely constant over time and one-dimensional. a-d, Same as Fig. 4b-e but for the RNN data. **a,b,** Inputs are learned nearly flat in both the A, B^{cx} and A^{cx}, B models, and almost identical across contexts in the A, B^{cx} model, which is consistent with the ground truth RNN inputs being flat and fixed across contexts (Methods). Yet, both models, including the A, B^{cx} model, strongly amplify the relevant inputs across contexts, but not the irrelevant ones. Note that the relevant inputs are much more strongly amplified than the irrelevant ones in the RNN, as expected from an optimal selective integration strategy[1], but the same is not found in the PFC data (Fig. 4b,c). **c,d,** In the RNN the coherence representations are largely 1D since very little coherence magnitude modulation exists, unlike what is found for the PFC data (Fig. 4d,e). This is consistent with the ground truth RNN inputs being 1D and lacking a coherence magnitude component. **e,** Same as Extended Data Fig. 3c but for the RNN data. The alignments of the second inferred dimension (coherence magnitude) are not consistent across contexts in the A, B^{cx} model (first panel), and neither across models for the irrelevant inputs (third panel, transparent lines), unlike what is found for the PFC data (Extended Data Fig. 3c). The same is observed for the third input dimension, both in the RNN and the PFC data. This indicates that the coherence magnitude dimension in the RNN is not an invariant input feature. This is consistent with the ground truth RNN inputs being 1D and lacking a coherence magnitude component. Instead, the additional input dimensions might be learned to compensate for the linear approximation.



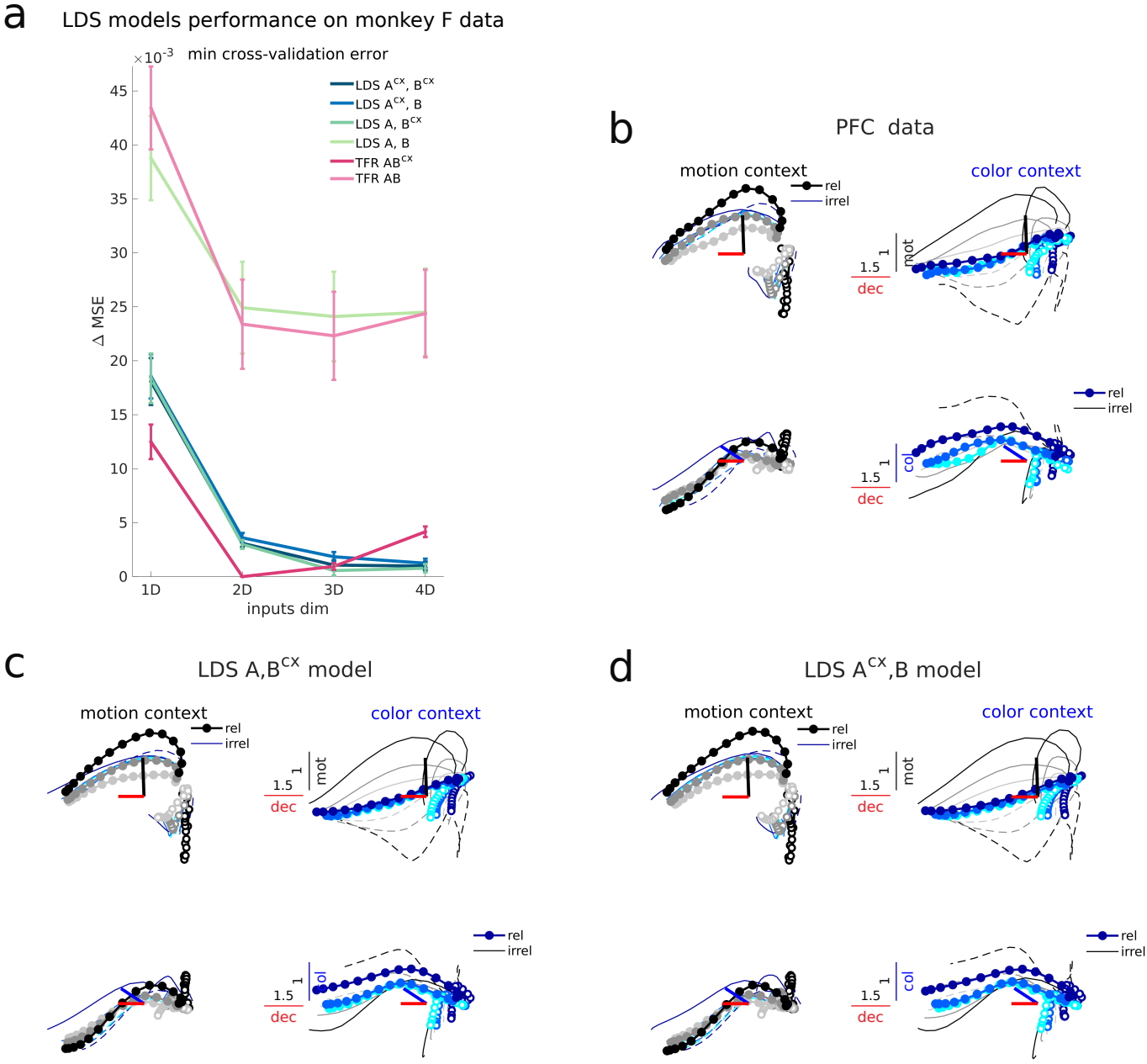
Supplementary Figure 3 . The LDS dynamics inferred from the RNN data are high-dimensional, mediate relevant input selection largely through slow modes, and implement transient amplification of relevant inputs in the two models, but are extremely non-normal for the A, B^{cx} model. a,b, Same as Fig. 5b,c and Extended Data Fig. 6a, but for the RNN data. **a,** The LDS models inferred from the RNN data are higher dimensional than expected for an idealized line attractor solution, with a single slow dimension and the rest of the dimensions fast decaying[16]. In particular, the LDS models learned several slow modes ($|\lambda| > 0.8$, green lines), as happened in the LDS data (Fig. 5b). However, the LDS models inferred from the RNN data had a smaller fraction of slow modes than the ones inferred from the PFC data (A, B^{cx}, $20 \pm 5\%$; A^{cx}, B, $21 \pm 7\%$ mot cx/ $22 \pm 5\%$ col cx, mean$\pm$std across 100 models; vs. $35 - 55\%$ in the PFC data). Furthermore, for the rest of the modes, most of them were very fast decaying ($|\lambda| < 0.4$, $\tau < 55$ms), unlike what was found in the PFC data, where most of the modes had either intermediate ($|\lambda| = 0.4 - 0.8$, $\tau = 98 - 224$ms) or slow eigenvalues ($|\lambda| > 0.8$). The largest eigenvalue was 1.00 ± 0.02 for the A, B^{cx} model and 0.98 ± 0.04 / 0.99 ± 0.04 (mot/col cx) for the A^{cx}, B model. The second largest eigenvalue was 0.94 ± 0.03 for the A, B^{cx} model and 0.93 ± 0.08 / 0.95 ± 0.06 (mot/col cx) for the A^{cx}, B model. The additional slow modes could have been learned to capture the curvature of the line attractor[1], or alternatively, to capture CI and contextual variance, as found in the PFC data (Extended Data Fig. 10, see also Supplementary Fig. 5). Another possibility is that the higher dimensionality is a necessary feature of the mapping between low-rank linear RNNs and LDS models[63] (note that the RNN dynamics was low-rank, and approximately linear in each context[1]). Similarly, this might also explain the dimensionality of the LDS models inferred from the PFC circuit. **b,** In both models, the coherence inputs inferred are most strongly loaded onto slow modes, rather than intermediate modes as in the PFC data (Fig. 5c). However, the inputs do not preferentially load onto the slowest modes ($|\lambda| > 0.9$, $\tau > 475$ms), as we found in the data. **c-e,** Same as Fig. 6a-c but for the RNN data. **c,** Left-eigenvector perturbations as well as random perturbations result in very strong amplification for the A, B^{cx} model, unlike what is found in the PFC data (Fig. 6a). Accordingly, the degree of non-normality for this model is extremely high (**e**), unlike what is found in the PFC data (Fig. 6c). On the contrary, the response behavior of the A^{cx}, B fitted to the RNN data is similar to the behavior of the same model fitted to the PFC data (Fig. 6a). **d,** Random perturbations are very strongly amplified by the A, B^{cx} model, but the inputs are not. This indicates a high degree of specificity in the model, which might imply fine-tuning. The A^{cx}, B model, on the contrary, process inputs in a similar way as found in the PFC data (Fig. 6b), with the difference that the relevant inputs in this case are transiently "amplified", rather transiently "persistent". This might be simply due to the fact that for the RNN data the inputs are learned much weaker than for the PFC data (Fig. 4c, Supplementary Fig. 2b), but the input pulse is provided with the same strength in both cases (unit norm). Also, the A^{cx}, B model is slightly more non-normal for the RNN data (**e**) than the PFC data (Fig. 6c).



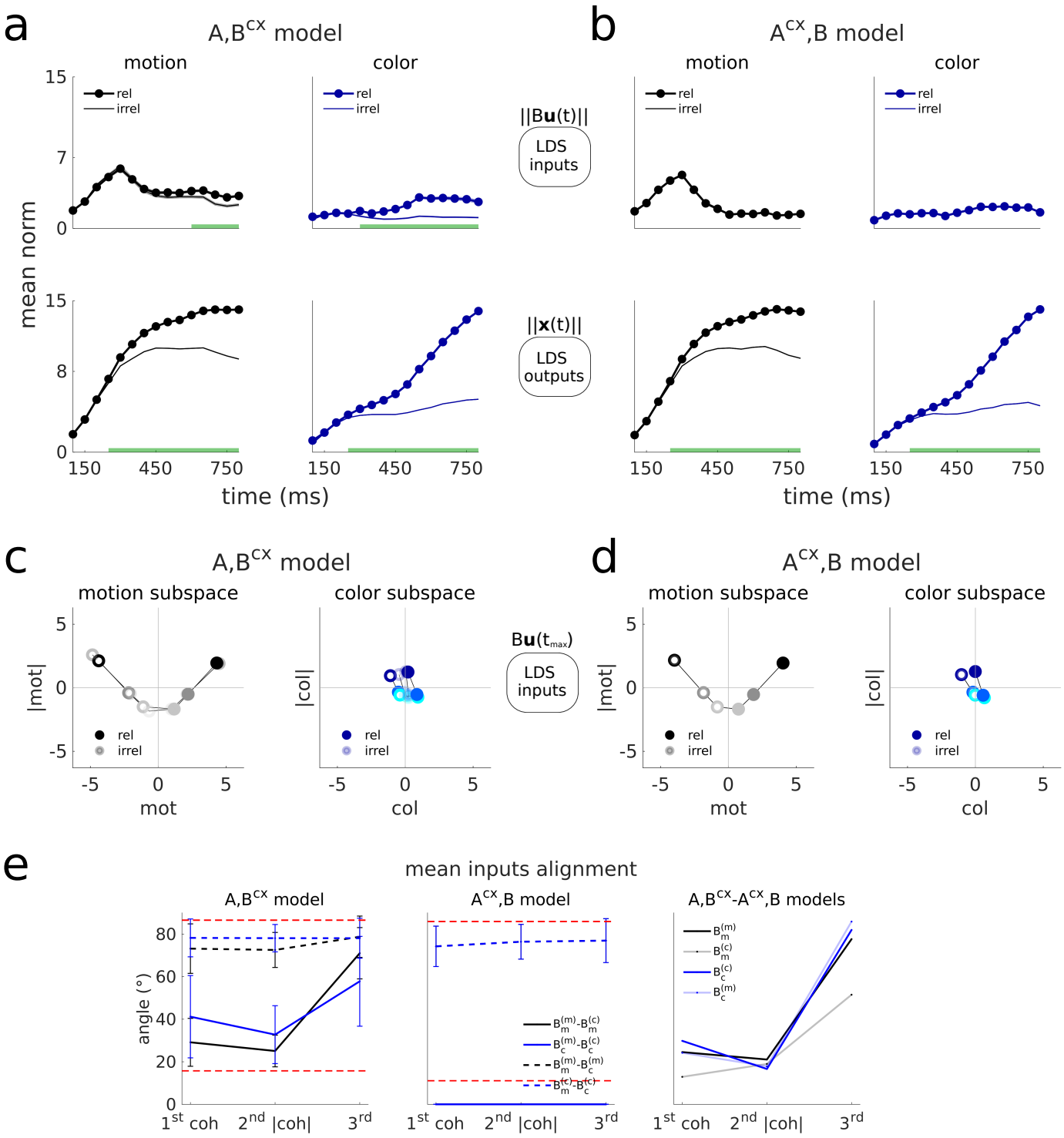
Supplementary Figure 4 . **The RNN integration process did not separate in two phases. a-d,** Same as Fig. 7, Extended Data Fig. 8, but for the RNN data. **a,b,** The largest CD mode for the RNN data projects to slow modes throughout the whole trial, unlike what is found in monkey A's PFC data, where the CD vector projects to relatively fast decaying modes first and to the slowest modes later in the trial. Thus, the CD projection pattern in the RNN does not change over the course of the trial. Indeed, the distributions of projection early vs. late in the trial are practically indistinguishable (absence of green bars in **b**, Wilcoxon rank-sum test, $p<0.05$). The CD vector aligns to slow modes ($|\lambda| > 0.8$), but not preferentially to the slowest modes ($|\lambda| > 0.9$), unlike what is found in monkey A's PFC data. The dynamics was also largely non-rotational ($\approx$1Hz), and not significantly different early vs. late in the trial (right panels in **b**). **c,d,** The largest CI mode for the RNN data projects most strongly to the slowest modes, unlike what is found for monkey A's PFC data, where the CI vector targets slow modes, but not the slowest. **e,** Same as Fig. 8b, but fore the RNN data. Top, the dimension inferred early in the trial highly aligns to the decision axis (small angle between red and yellow bars). This is consistent with the fact that early in the trial the RNN CD vector already projects to slow modes (**a,b**), and not to a different set of dimensions (the relatively fast decaying modes), as is found in monkey A's PFC data, which define a secondary decision dimension (Fig. 8b). Middle panels, trajectories in the LDS motion and color input coherence dimensions. The coherence input vectors found by the LDS models are only moderately aligned with the ground truth input vectors (A, B^{cx}: mot = 45°, col = 55°, for mean coherence input dimensions across 100 models and across contexts; A^{cx}, B, mot = 43°, col = 54°, for mean across 100 models), but higher than expected by chance (Extended Data Fig. 6b). However, the RNN trajectories along the LDS coherence input dimensions (middle panels) are very similar to the trajectories along the ground truth input vectors, and separate coherence information similarly well (Supplementary Fig. 1b).



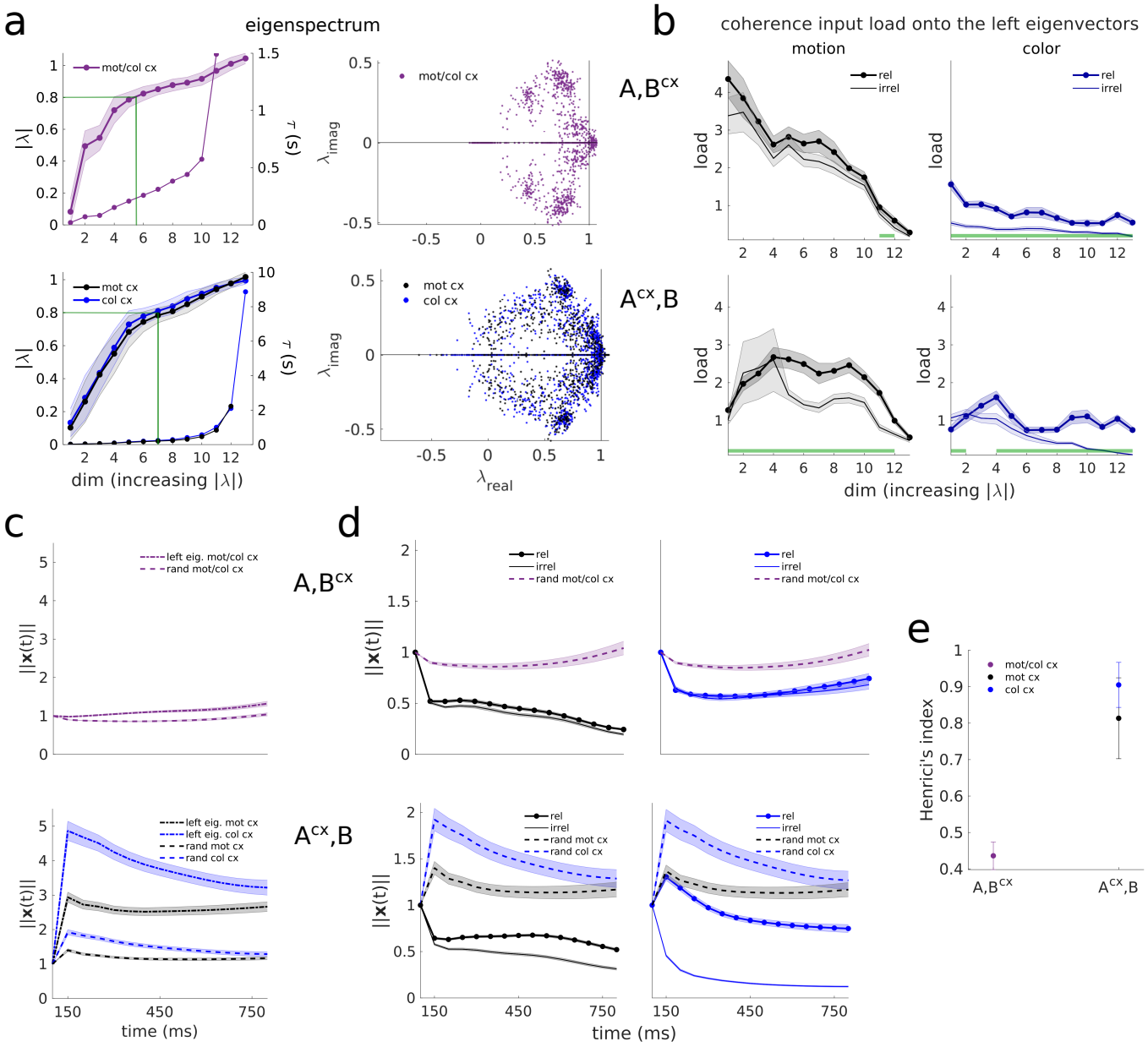
Supplementary Figure 5 . Choice signals in the slowest LDS subspace largely evolve along a single dimension across contexts in the RNN data. Same as Extended Data Fig. 10, but for the RNN data. **a,b,** The right eigenvectors effectively span four dimensions, since the singular values are not close to zero, as is found in the PFC data. Furthermore, the first first singular vector dimension is also the most aligned across contexts in the A, B^{cx} model and across models. **c,d,** The first singular vector dimension (sv) captures decision information. However, sv dimensions 2 to 4 mostly capture condition independent (CI) variance and contextual variance. Indeed, only the first dimension aligns well with the RNN decision dimension (red bars), which is defined by the RNN readout or output vector after training[1] (sv1: 61°, sv2: 85°, sv3: 90°, sv4: 84°). Thus, the RNN trajectories mainly evolve along a single dimension in this 4D subspace, which aligns with the decision (or output) axis of the network. The other dimensions are used to capture the small curvature of the approximate line attractor (sv4), CI features (sv2), and contextual separation (sv3, sv4).



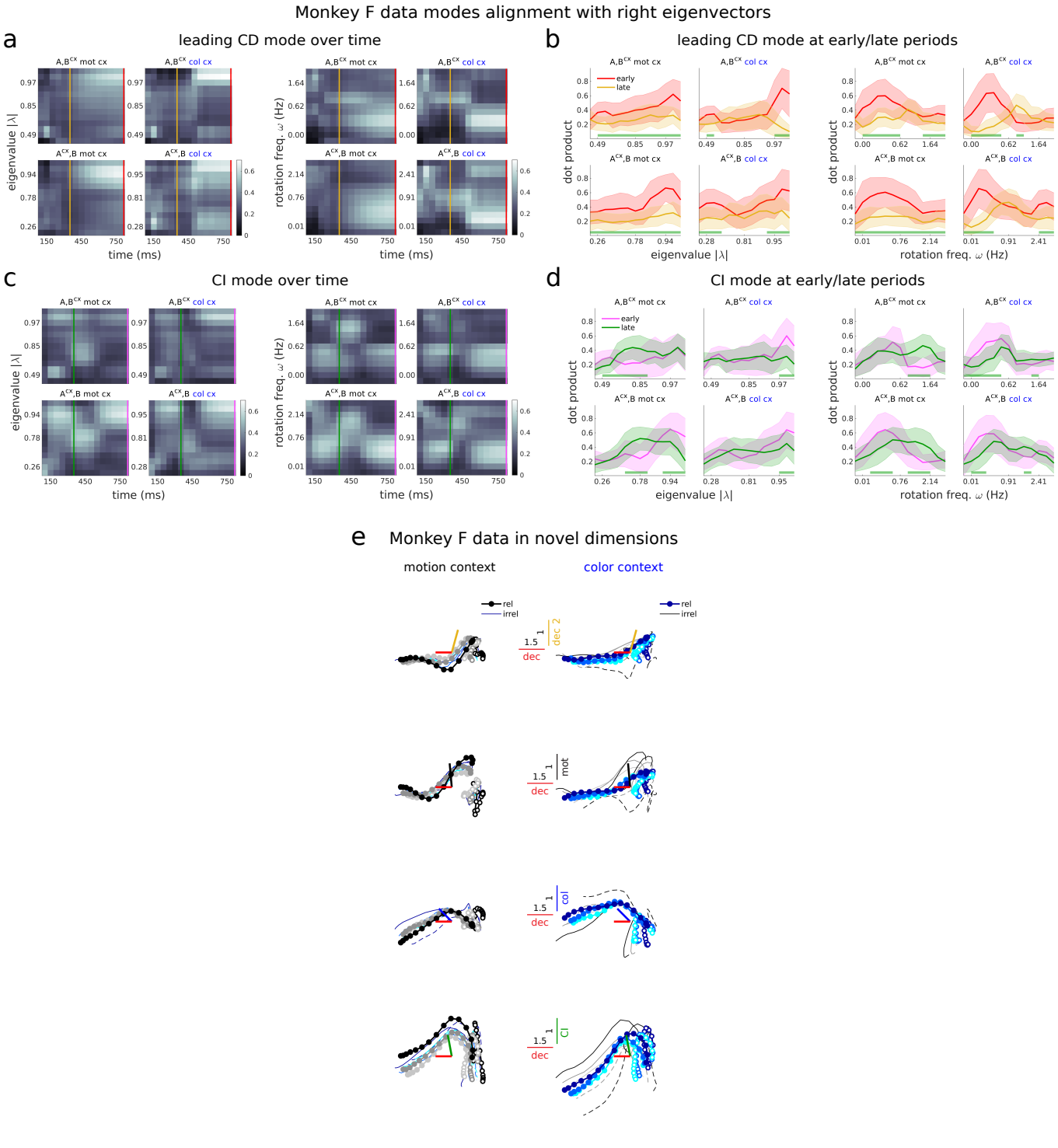
Supplementary Figure 6 . **The LDS models accurately capture monkey F's data.** Same as Supplementary Fig. 1, but for monkey F data. **a,** The pattern of errors across models and input dimensions closely follows the one obtained for monkey A (Fig. 2a). The best performing model is also the A, B^{cx} model with 3D inputs. The A^{cx}, B model with 3D inputs performs similarly well (Supplementary Table 3). **b-d** Same as Fig. 3b-d but for monkey F data. The LDS models accurately capture the trajectories in the stable task-relevant subspace.



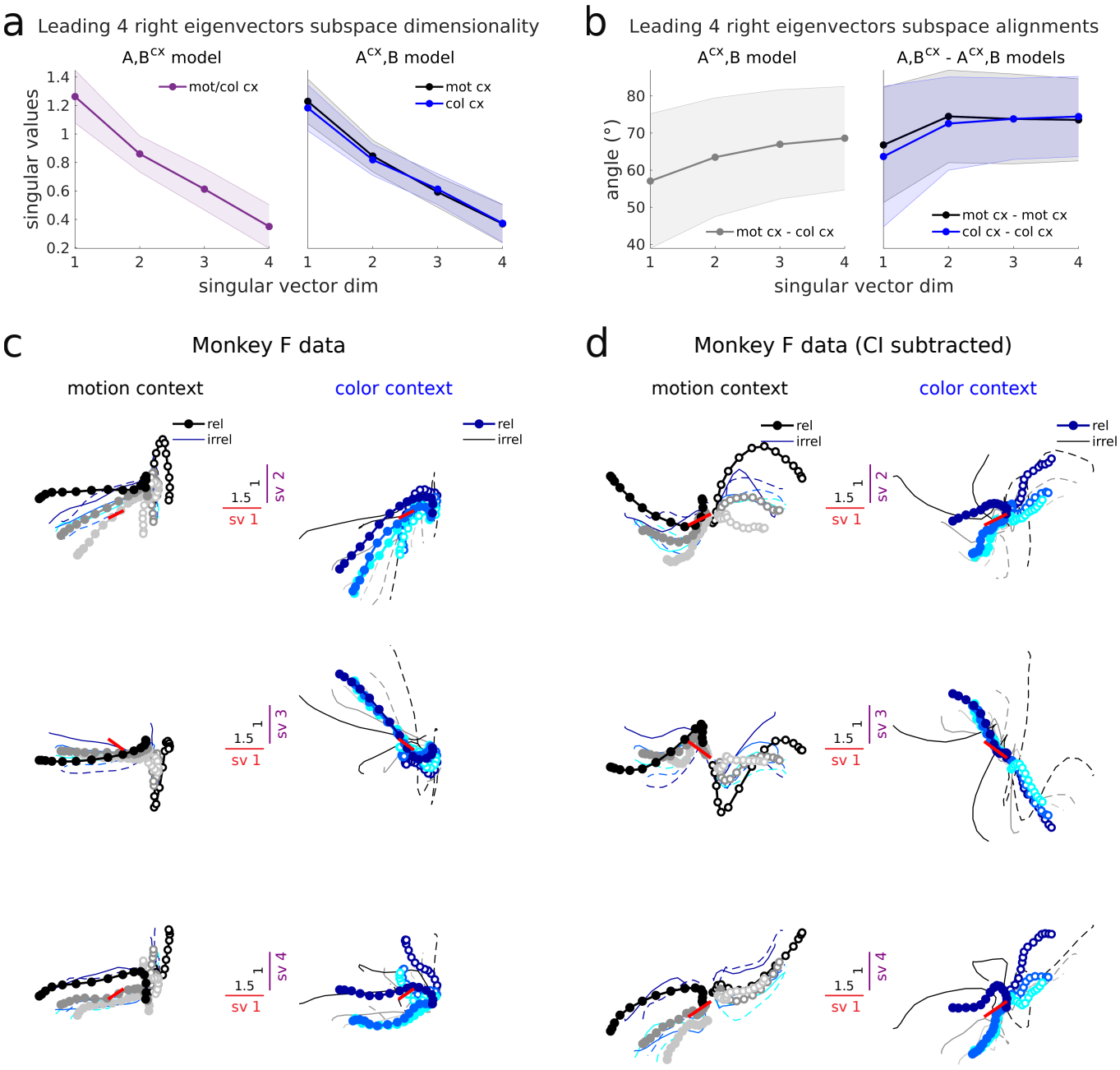
Supplementary Figure 7 . The inferred LDS inputs from monkey F's data are largely stable across contexts and span curved manifolds. Same as Supplementary Fig. 2, but for monkey F data. **a,b,** The motion inputs from both models are similar to the inputs inferred for monkey A, in that these are transient in both the A, B^{cx} and A^{cx}, B models, but more sustained in the A, B^{cx} model (top left panels). However, the A, B^{cx} model infers motion inputs that are nearly identical in strength across contexts (**a**, top left), unlike what is found for monkey A (Fig. 4b). Accordingly, the motion inputs are strongly integrated in both contexts, with the relevant outputs being only slightly stronger than the irrelevant ones (**a**, bottom left). Another difference is that the inferred color inputs are very weak, although these increase slightly towards the end of the trial when relevant (**a**, top right). Yet, the A, B^{cx} model selectively integrates the color inputs (**a**, bottom right). Both models generate identical outputs (**a,b**, bottom). The color outputs increase more sharply in the middle of the trial, and continue growing until the end of the trial. On the contrary, motion outputs saturate towards the end of the trial (bottom left). This saturation is also observed in monkey A, for both the motion and the color outputs (Fig. 4b,c). Thus, color inputs might be integrated later in the trial in monkey F. An alternative possibility could be that color signals arriving into monkey F's PFC circuit are already integration signals, and our LDS models learn this particular input-ouput solution due to the fact that they incorporate an input penalty which encourages learning weak inputs (Methods). In line with this interpretation, we found that the LDS coherence input dimension is highly aligned with the decision dimension (Supplementary Fig. 9e). **c,d,** The motion coherence representations inferred by both models are strongly curved in this monkey. The color inputs are weak along both the coherence and the coherence magnitude dimensions. **e,** The pattern of alignments between the different input dimensions across models and contexts is consistent across monkeys. The inferred coherence and coherence magnitude dimensions for both motion and color are largely stable across contexts in the A, B^{cx} model, but not the third dimensions (filled lines). One difference is that the motion coherence magnitude dimension is highly aligned across contexts in the A, B^{cx} model, but less so in monkey A (Extended Data Fig. 3c). This is consistent with the motion coherence representations having a stronger curvature in monkey F than in monkey A (**c,d** vs. Fig. 4d,e).



Supplementary Figure 8 . **The LDS dynamics inferred from monkey F's data are high-dimensional, mediate relevant input selection through both intermediate and slow modes, and implement transient amplification of relevant inputs in the A^{cx}, B model.** Same as Supplementary Fig. 3, but for monkey F data. **a,** The inferred eigenspectrum from both models presents multiple slow modes, as in monkey A. The A, B^{cx} model had also a larger fraction of slow modes than the A^{cx}, B model ($63 \pm 7\%$ vs. $46 \pm 8\%$ mot cx/ $53 \pm 8\%$ col cx, mean$\pm$std across 100 models). Note that the last two eigenvalues of the A, B^{cx} model are slightly larger than 1 (unstable), so the time constant is not shown. Same for the last eigenvalue of the A^{cx}, B model in the motion context. **b,** The coherence inputs are not preferentially loaded onto the slowest modes, but rather, intermediate and fast modes, as in monkey A's data. However, the A, B^{cx} model loads the motion inputs selectively only onto the second and third slowest modes (green bar). **c** In the A, B^{cx} model the average impulse response across random perturbations grows over time, indicating the presence of unstable dynamics (top), which was not found in monkey A. On the contrary, the A, B^{cx} model exhibited similar transient responses as monkey A (bottom). **d** The instability effect is also observed for perturbations along the color input dimension (top right panel). This feature might help integrate color input signals selectively, given that color inputs are very weak Supplementary Fig. 7a. The A, B^{cx} model response to motion and color pulses is broadly similar to that of monkey A's. However, the responses were more amplified for the relevant color inputs in this monkey, and the relevant motion inputs exhibited transient amplification effects later in the trial, rather than immediately after the pulse. **e,** The degree of non-normality of the LDS models is similar across monkeys (Fig. 6c).



Supplementary Figure 9 . **Monkey F's integration process did not separate in two phases.** Same as Supplementary Fig. 4, but for monkey F data. **a,b,** Monkey F did not present a clear separation of the integration process into different phases, unlike monkey A. In particular, early in the trial the leading CD variance vector did not preferentially project onto relatively fast decaying modes (**b**, left panels, yellow lines, the pattern of projections is nearly flat), unlike for monkey A (Fig. 7b, Extended Data Fig. 8b, left panels, yellow lines, projections pick at intermediate modes). However, the pattern of early vs. late projections was similar to that of monkey A when splitting the modes by their rotation frequency (**b**, right panels, yellow lines in the color context, and in the motion context for the A, B^{cx} model pick on intermediate rotation frequencies, as for monkey A, Fig. 7b, Extended Data Fig. 8b, right panels). The modes targeted during the late phase of the trial (red line projections) are consistent across monkeys both in time constant and rotation frequency. This late projections significantly differ from the early projections (green bars in **b**, Wilcoxon rank-sum test, $p<0.001$). **c,d,** The CI data vector for monkey F projects most strongly onto the slowest modes, in particular at the end of the trial (**d**, pink lines). This indicates that CI signals are integrated along the same dimensions as the CD signals, i.e., the dimensions that integrate inputs (compare with red lines in **b**), unlike what is found for monkey A (Fig. 7c,d, Extended Data Fig. 8c,d, left panels, pink lines pick onto slow modes, but not the slowest). This is consistent with the fact that CI signals are strongly present along the decision axis[1]. The block-like structure found in the pattern of projections in panels **a,c** comes from the fact that the variance along the first singular vector dimensions is close to the variance along the other singular value dimensions (i.e. that the covariance structure of the data is largely spherical), which leads to switches in the estimation of the dominant variance direction. **e,** The early-phase dimension or secondary decision dimension (yellow) does not capture much structure of the trajectories in monkey F (top). The late-phase dimension (red) is strongly aligned with the TDR decision axis (18°). The LDS coherence input dimensions were well aligned with the TDR dimensions in the A^{cx}, B model, specially for color, but not in the A, B^{cx} model (A, B^{cx}: mot = 54°, col = 71°, for mean coherence input dimensions across 100 models and across contexts; A^{cx}, B, mot = 44°, col = 31°, for mean across 100 models). The averaged coherence input dimension across contexts and models highly aligned with the decision dimension (angle between blue and red bars in middle panels). This alignments were higher than expected by chance (Extended Data Fig. 6b).



Supplementary Figure 10 . Choice signals in the slowest LDS subspace largely evolve along a single dimension across contexts in monkey F's data. Same as Supplementary Fig. 5, but for monkey F data. **a,b,** The right eigenvectors effectively span four dimensions, since the singular values are not close to zero, as found in the PFC data. Furthermore, the first singular vector dimension is also the most aligned across contexts in the A, B^{cx} model and across models (Extended Data Fig. 10a,b). **c,d,** The first singular vector dimension (sv) captures decision information. However, sv dimensions 2 to 4 mostly capture condition independent (CI) variance and contextual variance. Indeed, only the first dimension aligns well with the TDR decision dimension (red bars) (sv1: 58°, sv2: 80°, sv3: 75°, sv4: 76°).

3 Supplementary Tables

Supplementary Table 1. Monkey A data minimum MSE $\pm$ sem across k folds (36 conditions) and corresponding latent dimensionality for which it is achieved, for LDS models with different input dimensionalities and contextual constraints. Highlighted in black is the model that achieved the minimum MSE. It cannot be appreciated in the 4th column (min MSE) due to rounding error, but can be seen in the last column (min Δ(MSE $\pm$ sem)), where performance is given relative to the best performing model (TFR 2D AB^{cx} model, see Supplementary Table 2), so the differences across models, albeit small, can be revealed. This quantity is the one reported in Fig. 2a.

Input dim	Contextual constraints	Latent dim	min MSE	min ΔMSE ($\times 10^{-3}$)
1D	A^{cx}, B^{cx}	13	0.74 ± 0.02	13 ± 1
	A^{cx}, B	15	0.74 ± 0.02	13 ± 1
	A, B^{cx}	15	0.74 ± 0.02	12 ± 1
	A, B	16	0.78 ± 0.02	48 ± 6
2D	A^{cx}, B^{cx}	15	0.73 ± 0.02	2.7 ± 0.4
	A^{cx}, B	15	0.73 ± 0.02	2.7 ± 0.4
	A, B^{cx}	16	0.73 ± 0.02	1.7 ± 0.6
	A, B	17	0.77 ± 0.02	38 ± 6
3D	A^{cx}, B^{cx}	14	0.73 ± 0.02	1.3 ± 0.4
	A^{cx}, B	16	0.73 ± 0.02	1.3 ± 0.3
	$\mathbf{A}, \mathbf{B^{cx}}$	**18**	$\mathbf{0.73 \pm 0.02}$	$\mathbf{0.6 \pm 0.5}$
	A, B	17	0.77 ± 0.02	38 ± 6
4D	A^{cx}, B^{cx}	14	0.73 ± 0.02	1.9 ± 0.4
	A^{cx}, B	14	0.73 ± 0.02	1.8 ± 0.4
	A, B^{cx}	16	0.73 ± 0.02	1.0 ± 0.5
	A, B	17	0.77 ± 0.02	39 ± 6

Supplementary Table 2. Monkey A data minimum MSE $\pm$ sem, min Δ(MSE $\pm$ sem) and corresponding latent dimensionality for which it is achieved, for TFR models with different input dimensionalities and contextual constraints. Same conventions as in Supplementary Table 1. Performance in the last column is given relative to the best performing model (TFR 2D AB^{cx} model).

Input dim	Contextual constraints	Latent dim	min MSE	min ΔMSE ($\times 10^{-3}$)
1D	AB^{cx}	13	0.74 ± 0.02	9 ± 1
	AB	14	0.78 ± 0.02	51 ± 7
2D	$\mathbf{AB^{cx}}$	**14**	$\mathbf{0.73 \pm 0.02}$	$\mathbf{0 \pm 0}$
	AB	16	0.76 ± 0.02	36 ± 6
3D	AB^{cx}	14	0.73 ± 0.02	1.5 ± 0.4
	AB	18	0.76 ± 0.02	36 ± 6
4D	AB^{cx}	14	0.73 ± 0.02	4.3 ± 0.5
	AB	14	0.77 ± 0.02	38 ± 6

Supplementary Table 3. Monkey F data minimum MSE $\pm$ sem, min Δ(MSE $\pm$ sem) and corresponding latent dimensionality for which it is achieved, for LDS models with different input dimensionalities and contextual constraints. Same conventions as in Supplementary Table 1. Performance in the last column is given relative to the best performing model (TFR 2D AB^{cx} model, see Supplementary Table 4).

Input dim	Contextual constraints	Latent dim	min MSE	min ΔMSE ($\times 10^{-3}$)
1D	A^{cx}, B^{cx}	12	0.75 ± 0.02	18 ± 2
	A^{cx}, B	14	0.75 ± 0.02	19 ± 2
	A, B^{cx}	12	0.75 ± 0.02	18 ± 2
	A, B	15	0.77 ± 0.02	39 ± 4
2D	A^{cx}, B^{cx}	13	0.73 ± 0.02	3.1 ± 0.3
	A^{cx}, B	14	0.73 ± 0.02	3.6 ± 0.4
	A, B^{cx}	13	0.73 ± 0.02	3.0 ± 0.4
	A, B	15	0.75 ± 0.03	25 ± 4
3D	A^{cx}, B^{cx}	13	0.73 ± 0.02	1.1 ± 0.5
	A^{cx}, B	13	0.73 ± 0.02	1.8 ± 0.4
	A, B^{cx}	**13**	$\mathbf{0.73 \pm 0.02}$	$\mathbf{0.6 \pm 0.5}$
	A, B	14	0.75 ± 0.03	24 ± 4
4D	A^{cx}, B^{cx}	13	0.73 ± 0.02	1.0 ± 0.4
	A^{cx}, B	13	0.73 ± 0.02	1.2 ± 0.4
	A, B^{cx}	12	0.73 ± 0.02	0.8 ± 0.4
	A, B	13	0.75 ± 0.03	24 ± 4

Supplementary Table 4. Monkey F data minimum MSE $\pm$ sem, min Δ(MSE $\pm$ sem) and corresponding latent dimensionality for which it is achieved, for TFR models with different input dimensionalities and contextual constraints. Same conventions as in Supplementary Table 1. Performance in the last column is given relative to the best performing model (TFR 2D AB^{cx} model).

Input dim	Contextual constraints	Latent dim	min MSE	min ΔMSE ($\times 10^{-3}$)
1D	AB^{cx}	12	0.74 ± 0.02	12 ± 2
	AB	13	0.77 ± 0.02	43 ± 4
2D	AB^{cx}	**12**	$\mathbf{0.73 \pm 0.02}$	$\mathbf{0 \pm 0}$
	AB	13	0.75 ± 0.03	23 ± 4
3D	AB^{cx}	12	0.73 ± 0.02	1.0 ± 0.3
	AB	12	0.75 ± 0.03	22 ± 4
4D	AB^{cx}	12	0.73 ± 0.02	4.2 ± 0.5
	AB	12	0.75 ± 0.03	24 ± 4

Supplementary Table 5. RNN data minimum MSE $\pm$ sem, min Δ(MSE $\pm$ sem) and corresponding latent dimensionality for which it is achieved, for LDS models with different input dimensionalities and contextual constraints. Same conventions as in Supplementary Table 1. Performance in the last column is given relative to the best performing model (LDS 4D A, B^{cx} model).

Input dim	Contextual constraints	Latent dim	min MSE	min ΔMSE ($\times 10^{-3}$)
1D	A^{cx}, B	23	0.039 ± 0.003	27 ± 2
	A, B^{cx}	26	0.039 ± 0.003	27 ± 2
2D	A^{cx}, B	13	0.017 ± 0.001	5.1 ± 0.5
	A, B^{cx}	26	0.017 ± 0.001	5.0 ± 0.6
3D	A^{cx}, B	14	0.013 ± 0.001	0.8 ± 0.3
	A, B^{cx}	26	0.012 ± 0.001	0.1 ± 0.4
4D	A^{cx}, B	15	0.013 ± 0.001	0.6 ± 0.2
	A, B^{cx}	**24**	$\mathbf{0.012 \pm 0.001}$	$\mathbf{0 \pm 0}$