Computational mechanisms underlying thalamic regulation of prefrontal signal-to-noise ratio in decision making

Xiaohan Zhang[1], Michael M. Halassa[2] and Zhe Sage Chen[1,3,4] *

1. Department of Psychiatry, New York University Grossman School of Medicine, New York, NY, USA
2. Department of Neuroscience, Tufts University School of Medicine, Boston, MA, USA
3. Neuroscience Institute, New York University Grossman School of Medicine, New York, NY, USA
4. Department of Biomedical Engineering, New York University Tandon School of Engineering, New York, NY, USA.

* Email correspondence: zhe.chen@nyulangone.org (Z.S.C.)

Abstract

The mediodorsal (MD) thalamus is a critical partner for the prefrontal cortex (PFC) in cognitive flexibility. Animal experiments have shown that the MD enhances prefrontal signal-to-noise ratio (SNR) in decision making under uncertainty. However, the computational mechanisms of this cognitive process remain unclear. Here we use performance-optimized computational models to dissect these mechanisms. We find that the inclusion of an MD-like feedforward module increases robustness to sensory noise and enhances working memory maintenance in the recurrent PFC network performing a context-dependent decision-making task. Incorporating genetically identified thalamocortical pathways that regulate signal amplification and noise reduction further improves performance and replicates key neurophysiological findings of neuronal tuning. Our model reveals a key computational mechanism of context-invariant, cell-type specific regulation of sensory uncertainty in a task-phase specific manner. Additionally, it makes experimentally testable predictions that connect disrupted thalamocortical connectivity with classical theories of prefrontal excitation-inhibition (E/I) imbalance and dysfunctional inhibitory cell types.

MAIN

Successful execution of complex decision-making tasks requires identification and processing of multiple sources of uncertainty[1]. Task uncertainty may appear in the form of corrupted or incongruent sensory cues (sensory uncertainty)[2,3], their mapping to onto internal or behavioral variables (mapping uncertainty)[1,4,5], or their likelihood of resulting in reward (outcome uncertainty)[6]. Experiments across multiple species have sown that the MD thalamus is a critical partner for the PFC in resolving uncertainty in decision making[7-13]. Human neuroimaging studies have shown that MD activity tracks sensory uncertainty in a multi-attribute attention task[13], as well as a categorization task[15]. This process generalizes to non-human animals; in mice performing a decision-making task, the MD tracks sensory uncertainty and outputs a signal that appears to enhance prefrontal SNR[5,16]. Optical manipulations support this notion and delineate their causal roles in cognitive control[4,17]. However, the computational mechanism by which the MD enhances prefrontal activity in decision making under uncertainty remain unclear. In addition, how the newly discovered cellular diversity in MD thalamus contributes to such computations is unexplored. Biologically-inspired computational modeling provides a powerful approach to probe important questions that yield mechanistic insight into neural circuits[18-25]. Development of computational

circuit models that incorporate neuronal subtype and circuit pathway knowledge is vital in this direction[26-28].

Motivated by accumulating data in rodent and human experiments[4,13,14,16,17], here we develop biological constrained models for the PFC-MD network with cortical interneuron subtype and thalamocortical pathway specificity. Following the tenet 'structure determines functions', we ask the question whether and how cortical interneuron cell types, intracortical and thalamocortical connectivity may influence task performance and generate emergent task-specific neural representations. To explore these questions, we trained computational PFC-MD models to perform a context-dependent decision-making task under sensory cue uncertainty, from which we delineated the roles of interneuron subtype and thalamocortical-projection specificity in prefrontal computation. Our PFC-MD model replicated key behavioral data and neural tunings in task-performing mice, including prefrontal context-invariant rule-tuning and thalamic context tuning[4,5]. The MD subpopulations and cell-type specific MD-to-PFC projections have distinct functional roles in encoding sensory uncertainty driven by conflict and sparsity, maintaining working memory, and regulating prefrontal SNR in a task-phase specific manner. Furthermore, our models enabled us to parse the cognitive deficits of thalamocortical circuits that are related to prefrontal excitation-inhibition (E/I) imbalance, dysfunctional inhibitory cell types, and disrupted thalamocortical and corticothalamic connectivity.

RESULTS

Training biologically constrained PFC-MD models to perform decision-making tasks with parameterized task uncertainties. Behavioral flexibility to map the same rule under different contexts is critical in decision making. We adopted a performance-optimized training strategy of computational models to delineate the computational mechanism of decision making under uncertainty. We first modeled the PFC with a rate-based excitatory-inhibitory (E/I) recurrent neural network (E/I-RNN) (**Fig. 1a** and **Methods**), where cortical GABAergic inhibitory neurons consist of parvalbumin (PV), vasoactive intestinal peptide (VIP)-expressing and somatostatin (SOM) interneuron cell types (**Fig. 1b**). We assumed structured inhibitory-to-excitatory and inhibitory-to-inhibitory connectivity and disallowed weak connections that are negligible[28-30]. Specifically, the VIP interneurons mainly disinhibit other classes of interneurons[31,32]. We imposed key biological constraints on network connectivity; first, we modeled the MD itself as a feedforward structure devoid of lateral connectivity, where all neurons received excitatory inputs from the PFC and projected back to excitatory neurons. Second, guided by recent knowledge of genetically identified thalamocortical projections that differentially target distinct PFC interneuron types[4], we included these additional connectivity constraints into our model. One MD subpopulation, which is termed as MD_1, projects to prefrontal PV interneurons, while the other subpopulation, MD_2, projects to cortical VIP interneurons. To translate them to biological terms, $MD_1 \rightarrow PV$ corresponds to the kainite receptor (GRIK4)-expressing (i.e., MD_{GRIK4}) pathway with preferentially targeted PV+ neurons; $MD_{DRD2} \rightarrow VIP$ projection corresponds to the dopamine receptor (D2)-expressing (i.e., MD_{DRD2}) pathway with preferentially targeted VIP^+ neurons. We trained both the PFC-MD model, and as a comparison a 'PFC-alone' model (using identical stimulus inputs and cortical E/I configuration) to perform a cueing context-dependent decision-making task inspired by mouse experiments[5] (**Fig. 1c** and **Methods**). The two contexts essentially correspond to two independent subtasks that map different sensory inputs to one set of rules: attend-to-audition vs. attend-to-vision (**Fig. 1d**). Importantly, the cueing uncertainty was parameterized in both subtasks, where the first is a random dot motion (RDM) task, and the second is a conflicting acoustic cue (CAC) task (see **Methods**). Additionally, the sparsity of cues, as a generalized form sensory uncertainty, was parameterized and controlled. These two subtasks consisted of three common task phases: (i) identifying the rule by accumulating sensory evidence from ambiguous cues (800-ms cueing period); (ii) maintaining the rule information during working memory (800-ms task delay period);

(iii) making choices in the presence of both visual and auditory stimuli (200-ms target period). During the cueing period, the model first learned to accumulate sensory evidence of ambiguous cues to recognize the rule. Next, during the delay period, the model learned to preserve the rule information in working memory. Finally, during the target period, the network learned to attend to the right modality and made a rule-guided choice.

We trained PFC-MD models to learn the RDM and CAC subtasks where the sensory uncertainty was minimal. Upon reaching the desired task accuracy, we tested the performance-optimized PFC-MD networks by systematically varying the degree of cue conflicts or uncertainty. In the RDM context, this corresponded to changing the coherence of random dot motion; whereas in the CAC context, it corresponded to changing the congruence or conflict levels of low-pass (LP) vs. high-pass (HP) tones. For the first subtask, motion direction mapped onto the rule, whereas tone identity did so in the second.

PFC-MD model outperformed PFC-alone model in tasks with higher sensory uncertainty. In testing, we produced psychometric curves for two subtasks (**Fig. 1e**). We changed the level of coherence or congruence of the cue signals, and found that the PFC-MD model outperformed the PFC-alone model in both contexts, especially in the medium-to-high cue conflicting conditions. This was consistently observed in multiple independently performance-optimized models (**Extended Data Fig. 1a**), when all models achieved the same performance on similarly constructed training trials. Furthermore, we examined the percentage of error types in decision making. Given the outcome mismatch among four choices, task errors can be ascribed by either attending the wrong rule or attending the correct rule yet selecting the wrong choice. The rule-encoded executive error occurred in either the cueing or delay period, whereas the sensory error only occurred in the target period. While testing the performance-optimized PFC-MD model, the sensory error was stable regardless of the change in cue uncertainty, where the rule error changed according to the level of cue uncertainty (**Fig. 1f**).

The SNR of cueing context was determined by not only the ratio of conflicting cues, but also the density of cue signals (or equivalently the sparsity: the higher the density, the lower the sparsity). For instance, in the CAC context, a 4/2=2:1 ratio of LP/HP pulses indicates a relatively high level of noise---where signals are conflicting and sparse; an identical 8/4=2:1 ratio of LP/HP but with more pulses would have a higher SNR, where a higher density of pulses indicates a higher signal level (**Fig. 1g**). Increasing the cue sparsity had a negative impact on the task performance (**Fig. 1h**), but the PFC-MD model consistently outperformed the PFC-alone model (**Fig. 1i** and **Extended Data Fig. 1b**). In the RDM context, we changed the spatial density by varying the number of moving dots under the same cue coherence condition (**Fig. 1j**). Similarly, decreasing the signal density or increasing cue sparsity led to a degraded task performance (**Fig. 1k**), and the PFC-MD model also outperformed the PFC-alone model (**Fig. 1l** and **Extended Data Fig. 1c**).

PFC and MD units showed diverse tunings and population dynamics according to task variables, task phase and cue uncertainty. We examined the emerged tuning properties of single units with respect to task variables during both cueing and delay periods. A large majority (~95%) of PFC excitatory units exhibited tuning preference in rule, but not context, and some showed cueing context-invariant rule tuning during the delay period (**Fig. 2a,d** and **Extended Data Fig. 2a**). Similarly, a large majority (~90%) of PFC inhibitory units encoded rules during the delay period, among which some (~10%) were also modulated with the cueing context (**Fig. 2b,d,e** and **Extended Data Fig. 2b**). In contrast, nearly all (~95%) MD units showed modulation selectivity with respect to the cueing context, but not rule (**Fig. 2c,e** and **Extended Data Fig. 2c**). Furthermore, during the cueing period, we found that some MD_1 units (40-50%, across subtasks and multiple test conditions) modulated their firing rates with respect to cue uncertainty, whereas

MD_2 units didn't show such modulation (**Fig. 2f,g** and **Extended Data Fig. 2e**)., suggesting their functional distinction induced by prefrontal cell-type target specificity and connectivity (because we assumed no direct VIP→Exc connection in **Fig. 1b**). We also found firing modulation with respect to cue sparsity for the same MD_1 subpopulation (**Fig. 2h**), but not for MD_2 subpopulation (**Extended Data Fig. 2d,e**). Similar cue uncertainty tunings were also found in PFC inhibitory units (**Fig. 2i**). All of these emergent PFC and MD tuning properties were consistent with experimental findings in task-performing mice[4,5]. Additionally, increasing the cue uncertainty or cue sparsity reduced rule selectivity of PFC excitatory units in working memory (**Extended Data Fig. 3**), suggesting that reduced SNR of cue inputs may lead to an information loss during a cue-to-rule transformation. During the delay period, sorting the peak responses of PFC excitatory units produced a rule-specific sequence, replicating previous results[5,16,21,23,25]; moreover, the sequence was context-invariant for two subtasks (**Extended Data Fig. 4**).

We further examined pairwise correlation within the PFC population activity, and quantified the prefrontal SNR based on functional connectivity of prefrontal excitatory neurons (**Methods**). Independent of the rule or context, the prefrontal SNR estimate was greater during both cueing and delay periods when there was a higher degree of cue uncertainty (**Fig. 3a)**, implying that more prefrontal resources were demanded for engaging in a more difficult task. Furthermore, disconnecting PFC-MD connectivity (i.e., "MD lesioned") reduced prefrontal SNR (**Fig. 3b**), with accompanying reduction in task performance. We also computed the canonical correlation between PFC and MD activities during the cueing and delay periods, and found that the canonical correlation increased with increasing cue uncertainty and the canonical correlation was the highest value in the delay period under high cue uncertainty (**Fig. 3c**). To further explore the tuning properties of PFC-MD populations, we used decoding analyses to read out either rule or context information on a trial-by-trial basis based on the mean firing of population (**Methods**). The rule information could be reliably decoded from the PFC, but not from the MD (**Fig. 3d**). Under the same setting, increasing cue uncertainty led to a decrease in the PFC's rule-decoding accuracy, but had no effect on the MD's context-decoding accuracy (**Fig. 3e**).

Furthermore, dimensionality reduction of the PFC or joint PFC-MD population activity derived from the performance-optimized PFC-MD model revealed low-dimensional neural trajectories in the task-specific subspace (**Fig. 4a,d**). Increasing the level of cue uncertainty or sparsity led to changes of neural trajectories in the task-variable subspace (**Fig. 4b,c,e,f**). Interestingly, the neural trajectory converged to the fixed points faster under the condition of lowest cue uncertainty or cue sparsity, and the corresponding initial velocity was larger while dealing with lower cue uncertainty or sparsity (**Fig. 4g,h**). A fixed-point analysis on the three-dimensional PC subspace (**Methods**) revealed two 'line attractor-like' basins, with each one matching a rule (attend-to-vision vs. attend-to-audition) and aligning cue uncertainty from one end to the other (**Fig. 4i**). Rotating the view angle or collapsing these points could show two discrete 'fixed-point-like' attractors. Note that these two 'attractor lines' had finite lengths; the neural trajectory could diverge from the attractor basins when the cue uncertainty was too high.

Cell type-differential modulation of prefrontal computation in PFC-MD connectivity. Given genetically identified thalamocortical pathways from the MD to cortical interneuron subtypes[4], we further examined the roles of two distinct MD subpopulations (MD_1 and MD_2) and two specific thalamocortical projections (MD_1→PV and MD_2→VIP) in regulating prefrontal computation. Since a complete MD lesion has been shown to reduce prefrontal SNR (**Fig. 3b**), we first disconnected one of two thalamocortical projections while keeping the other intact, and in turn examined their individual impact on the task performance and prefrontal representations. We found that weakening MD_1→PV connectivity (by setting a small percentage of connections to zeros) reduced the performance rapidly for both RDM and CAC subtasks (**Fig. 5a**). To simulate an effect of evoked firing activity similar to optogenetic MD activation, we induced a transient external input

to the MD subpopulations to increase the phasic firing activity of MD units. Interestingly, enhancement of MD_1 activity could boost the task performance (especially in reducing rule-specific errors), but only within a narrow range (dashed line within the inset of **Fig. 5a**). Notably, MD activation or inactivation differentially modified PFC excitatory and inhibitory unit firing rates (**Extended Data Fig. 5a**), and MD activation also enhanced local PFC excitatory neuronal correlation (**Extended Data Fig. 5b**). In contrast, weakening $MD_2 \rightarrow$VIP connectivity caused the task performance to decay slowly (**Fig. 5b**); however, enhancement of MD_2 activity had no effect on the task performance (see inset). This difference may be because there was no direct VIP$\rightarrow$Exc pathway to prefrontal excitatory neurons.

Since MD_1 neurons encode cue uncertainty (but not for MD_2 neurons, **Fig. 2f,g**), we asked whether changes of MD_1 activity in the presence of cue uncertainty will affect the task performance. As a demonstration, we chose a high cue uncertainty setup for both RDM (cue coherence of 0.1) and CAC (cue congruence of 0.6) subtasks. Increasing the MD_1 subpopulation firing rate during the cueing period efficiently boosted the task performance during the cueing period, whereas increasing the MD_2 activity under a similar condition had little effect (**Fig. 5c**). Furthermore, we varied the level of cue uncertainty and sparsity in subtasks, and confirmed the benefits of MD_1 stimulation for task performance (**Fig. 5d,e**).

We also examined the role of cortical interneuron subtypes and changed the unidirectional connectivity between PFC excitatory units and respective SOM, PV and VIP interneurons in turn. Although the number of inhibitory units from these three subtypes were equal, their respective impacts on task performance were different (**Extended Data Fig. 6**).

MD enhances working memory maintenance in the PFC-MD network. The neural basis for persistent activity in working memory is thought to involve recurrent excitation among prefrontal pyramidal neurons. During working memory, the rule information is maintained in the thalamocortical communication, but the information may be subject to a loss with increased task delay duration or noise distraction. Therefore, we tested the robustness of performance-optimized PFC-MD model subject to potential information loss. In testing, we gradually increased the duration of delay period from the initial 800 ms to 1200 ms, and found that increasing delay duration had a negative impact on the task performance (**Fig. 6a**). To demonstrate the role of MD in working memory maintenance[16], we increased the phasic firing rate of MD_1 or MD_2 subpopulation during an elongated 1100-ms delay period (with coherence/congruence of 0.8 in respective RDM/CAC subtasks). In turn, this led to a slight boost in rule maintenance during the prolonged delay, but it required less firing rate elevation for MD_2 than MD_1 to achieve the same task improvement (**Fig. 6b,c**), probably because the $MD_2 \rightarrow$VIP pathway has an amplification effect on increasing the target sensitivity for working memory maintenance. In contrast, reducing the MD_1 activity during the delay period had an opposite effect on the task performance (**Fig. 6d**). Alternatively, we strengthened or weakened PFC excitatory-to-MD connection strengths and observed similar phenomena (**Extended Data Fig. 7a-d**). Overall, these results support that reciprocal thalamocortical communications can help maintain the rule information and two different MD cell types form functionally distinct thalamocortical pathways. As a comparison, the PFC-alone model also suffered from similar information loss while increasing the task delay duration (**Extended Data Fig. 8a**). Notably, direct enhancement of the prefrontal excitatory population firing rate had a similar impact on the task performance as indirect stimulation of $MD_1 \rightarrow$PV or $MD_2 \rightarrow$VIP thalamocortical pathway (**Extended Data Fig. 8b**), but modifying prefrontal excitatory-to-excitatory connectivity had a smaller effect (**Extended Data Fig. 8c**).

Parsing cognitive deficits and probing mechanistic causes in the PFC-MD network. Our performance-optimized PFC-MD models can serve as a platform to interrogate mechanisms of abnormal neural representations and cognitive deficits in working memory or behavioral

inflexibility. This was achieved by modifying the 'normal control' with specific 'computational dysfunction'. It is important to emphasize that even if two circuit manipulations lead to same task performance, the underlying mechanisms can still be different. Computational 'divide-and-conquer' strategies can help discriminate between different mechanistic causes that are inaccessible to interrogation through causal circuit manipulation.

Recurrent synaptic excitation among pyramidal neurons is approximately balanced by synaptic inhibition[33]. Thus, as our first computational dysfunction, we aimed to induce cortical excitation-inhibition (E/I) imbalance by changing the overall inhibition to excitatory neurons. This can be achieved by increasing the sparsity of prefrontal SOM→Exc connections to reduce the inhibition (**Fig. 5e**). Alternatively, MD inhibition also alters the prefrontal E/I balance[34] (**Fig. 5a**). Notably, these operations changed the rule tunings of some PFC excitatory units during the delay period and reduced the choice discriminability at the population level during the target period (**Fig. 7a,b** and **Extended Data Fig. 9**). This can be explained by the fact that the prefrontal network dynamics became more unstable and sensitive to external input perturbation, which pushed the neural trajectories out of their usual stable attractor regimes. Visualizing the population dynamics in the PCA space showed that E/I imbalance distorted the neural trajectory dynamics (**Fig. 7c** and **Extended Data Fig. 9**), so that the network might favor one rule over the other (i.e., 'cognitive biases' in decision making) --- a cognitive deficit observed in many mental disorders.

We further weakened prefrontal SOM→VIP and SOM→PV connectivity independently and observed a more robust task performance in SOM→PV manipulations (**Fig. 7d**), while a similar manipulation of VIP→SOM connectivity had little impact (**Extended Data Fig. 10a**). For bidirectional SOM-VIP connections, strengthening mutual inhibition between SOM and VIP neurons could amplify the sparse cue signal (blue curves in **Fig. 7e**; note that this was effective for bidirectional SOM-VIP amplification but not for unidirectional manipulation; **Extended Data Fig. 10b**), supporting the role of VIP-SOM motif in gain control[28]. A closer examination of single-unit and population responses revealed that bidirectional SOM-VIP amplification produced emergent rule tunings and enhanced rule discriminability (**Fig. 7f**). Importantly, MD_2→VIP stimulation further facilitated this amplification and improved prefrontal SNR (red curves in **Fig. 7e**), validating recent experimental data on the MD amplifier[4]. Notably, abnormal cortical GABAergic signaling or deficits in interneuron subtype has been implicated in cognitive impairment in autism spectral disorder (ASD), such as decreased responses to salient stimuli ('hypo-sensitivity') under a low SNR or decreased inhibitory gain control[35,36].

Unbalanced corticothalamic connectivity reshapes MD tunings. Finally, we investigated whether additional biological constraints in corticothalamic connectivity can change MD neuronal tunings. Data from mouse experiments have suggested that MD_{GRIK4} neurons received dense inputs from L5/6 of the prelimbic cortex, in contrast to MD_{D2} neurons that receive sparse prefrontal inputs (unpublished data, Halassa lab). To incorporate this prior knowledge, we further made two modifications. First, we assumed unbalanced MD_1 and MD_2 subpopulations (2:1 ratio in place of original 1:1 ratio). Second, we assumed that MD_2 neurons received a much sparser projection from prefrontal excitatory neurons than MD_1 neurons and imposed a sparsity constraint during the course of model training, where Exc→MD_2 connectivity was 90% sparser than Exc→MD_1 connectivity (**Fig. 8a**). Interestingly, we found that this modification of corticothalamic connectivity did not change psychometric curves of the performance-optimized PFC-MD model (**Fig. 8b**), but led to emergent tunings in some MD_2 units with respect to cue sparsity or cue uncertainty (**Fig. 8c**). However, MD_1 stimulation, but not MD_2 stimulation, improved the performance of two subtasks under cue uncertainty; yet either MD_2 or MD_1 stimulation could improve the task performance under cue sparsity (**Fig. 8d,e**). Importantly, when cue uncertainty and sparsity were both present, stimulation of both MD subpopulations was required to improve task performance (**Fig. 8f**). Together, these findings suggest distinct modulatory functions of cell type-specific thalamocortical projections.

DISCUSSION

The thalamus has various subnetwork motifs and diverse cortical target outputs. The first-order thalamic nuclei plays a main role in sensory transmission, whereas the higher-order thalamic nuclei such as the MD and pulvinar is critical for dynamic regulation of cortical activity in attention, executive control and perceptual decision-making[8,37,38]. Many computational models of thalamocortical networks have been developed involving the sensory thalamus[39-41] and higher order thalamus[18,42,43,]. An important question is how the additional MD thalamus contributes to decision making with sensory uncertainty? In this work, motivated by the accumulated data from animal experiments[4,44], we trained biologically inspired PFC-MD models with interneuron subtype and thalamocortical pathway specificity to perform decision-making tasks with sensory uncertainty. We found that the PFC-MD model outperformed the PFC-alone model in the presence of cue input uncertainty, while cell-type and pathway specificity play a vital role in regulating prefrontal computation during evidence accumulation (cueing period) and working memory (delay period). At different task phases, the MD regulates prefrontal resources and SNR for information transmission and maintenance (**Supplementary Fig. 1**) to encode the task variables (rule and context) in a coordinated manner. We found cell-type specific tunings in the MD subpopulation with respect to cue uncertainty and/or cue sparsity, and such MD tunings can be influenced by corticothalamic projection.

Our computational model provides direct support and validation for circuit mechanisms of PFC-MD network in decision making. Distinct cellular targeting among cortical interneurons underlies differential inhibitory effects on excitatory neurons[45]. Cortical interneuron-specific cells that specialize in the synaptic disinhibition of excitatory neurons may shape the way excitatory neurons integrate information, and exhibit context-specific and behavior-specific responses[33,46]. Optogenetic tagging experiments in mice revealed that MD_{GRIK4}→PV projection suppresses prefrontal noise when task inputs are dense but conflicting, and MD_{DRD2}→VIP projection amplifies prefrontal signals when task inputs are sparse[4]. The MD_1→PV projection is the primary drive for feedforward thalamic inhibition[47,48]. Cortical PV interneurons have been implied in the mechanisms of gamma oscillatory activity and cognitive function[49]; disturbances in this signaling contribute to altered gamma oscillations and working memory deficits in schizophrenia[50,51]. Notably, simulating prefrontal-hippocampal GABAergic projections through prefrontal long-range inhibition (which preferentially inhibits VIP-expressing interneurons) can also increase hippocampal feedforward inhibition and regulate the SNR[52]. Thalamocortical and corticothalamic pathways have distinct functional contributions to decision making[16]. Corticothalamic projection is known to adjust the gain and the tuning precision of thalamic neurons as required by behavioral demands[9]. Our analyses of corticothalamic and thalamocortical projections suggest an asymmetric information flow and distinct contributions to decision making at different stages. A complete delineation of cell-type specific connectivity of thalamocortical circuitry will enable us to improve our current understanding of cognitive control.

The capacity of information transmission between neurons or circuits is fundamentally limited by the SNR. The functional connectivity of neurons may serve as a proxy of information that is recruited and routed based on task difficulty and available resources. Furthermore, neuromodulation such as dopamine may enhance the SNR of prefrontal activity[53,54]. Since the MD is known to receive dopaminergic inputs[55], the MD_2 subpopulation in our model may be viewed as performing a similar function relevant to the dopamine type-2 receptor (D2)-expressing projection[4]. This action of dopamine is achieved by D1 and D2-receptor-mediated effects on pyramidal and local circuit neurons, which further mediate neuronal excitability and recurrent inhibition and thus contribute to the stability of cortical representations of external and internal stimuli[56]. In light of experimental findings[4], our modeling results on MD_2 neurons also support the role of MD in regulating prefrontal SNR. This is consistent with the idea that the MD can facilitate GABA shift and change prefrontal functional connectivity and E/I balance[57].

What insight do our thalamocortical models provide to the studies of cognition in a broader context? First, interareal communications requires temporal coordination between the source and target[58]. Our PFC-MD model provides a means to identify the role of thalamocortical and corticothalamic signaling and how populations of neurons interact at different task phases to engage in decision making. A deeper question that arises from our modeling is why the combination of a recurrent E/I structure (PFC) and a feedforward excitatory architecture (MD) plus neuron subtype specificity produces more flexible control and robustness (as compared to a PFC-alone model that achieves nearly the same computational power)? Our simulations suggest that the feedforward MD structure has a lower intrinsic dimensionality than the recurrent PFC structure (**Supplementary Fig. 2a**), yet the feedforward control strategy is effective for regulating recurrent prefrontal computation in a task-phase specific manner. Neurobiology may have good reverse-engineering lessons to learn and benefit from. One specific theoretical insight is that the MD may help 'slow down' prefrontal dynamics by increasing the time constant in order to improve information integration during cueing or working memory (**Supplementary Fig. 2b** and **Methods**). Second, computer simulations for all possible task conditions, which are usually difficult to accomplish in animal experiments due to limited resource constraints, may offer experimentally testable predictions on task performance and neural representation. Additionally, biologically constrained modeling can provide a paradigm to compare functions and computational efficiency derived from different intracortical, intra-thalamic, thalamocortical and corticothalamic connectivity (**Supplementary Fig. 3**). Lastly, our models with cell-type specificity provide a new strategy to dissect computational mechanisms of cognitive deficits encountered in neuropsychiatric disorders such as schizophrenia[59-62]. For instance, abnormal dopaminergic activation and lower cortical SNR have been implicated in schizophrenia[55]. Additionally, a lack of inhibitory control increases hyperexcitability and reduces cognitive flexibility in schizophrenia or ADHD[11,63,64]. Dysfunction of GABAergic inhibition also impacts synaptic E/I balance that is linked to pathophysiology of a wide range of disease phenotypes such as cognitive deficits and negative symptoms. Further dissection of these computational mechanisms may provide therapeutic insight into thalamic deep brain stimulation strategies to improve cognitive functions.

Our PFC-MD model replicated most key (but not all) experimental findings[4], which may be explained by several limitations in our current work and present new paths for future investigations. First, the current circuit model does not account for distinct corticothalamic projections to MD subpopulations (e.g., cortical projections to MD_2 populations are much sparser). A model of broader range of thalamocortical circuits can be further enriched by incorporating additional task-relevant cortical areas (such as the anterior cingulate cortex or orbitofrontal cortex)[65]. Second, our current model does not consider inter-trial feedback, and therefore is insufficient to encode consecutive trial information. This issue may be overcome by accommodating a trial-and-error reinforcement learning paradigm. Short-term plasticity in bidirectional PFC-MD pathways may also correct cognitive impairment[66].

References

1. Bach, D. R. & Dolan, R. J. Knowing how much you don't know: a neural organization of uncertainty estimates. *Nat. Rev. Neurosci.* 13, 572-586 (2012).
2. Fetsch, C. R., Pouget, A., DeAngelis, G. C. & Angelaki, D. E. Neural correlates of reliability-based cue weighting during multisensory integration. *Nat. Neurosci.* 15, 146-154 (2011).
3. Zhou, Y., Acerbi, L. & Ma, W. J. The role of sensory uncertainty in simple contour integration. *PLoS Comput. Biol.* 16, e1006308 (2020).
4. Mukerjee A, Lam NH, Wimmer RD, & Halassa MM. Thalamic circuits for independent control of prefrontal signal and noise. *Nature*, 600, 100-104 (2021).
5. Rikhye RV, Gilra A, & Halassa MM (2018). Thalamic regulation of switching between cortical representations enables cognitive flexibility. *Nat. Neurosci.* 21, 1753-1763 (2018).

6. O'Neill, M. & Schultz, W. Coding of reward risk by orbitofrontal neurons is mostly distinct from coding of reward value. *Neuron* 68, 789-800 (2010).
7. Marton TF, Seifikar H, Luongo FJ, et al. Roles of prefrontal cortex and mediodorsal thalamus in task engagement and behavioral flexibility. *J. Neurosci.* 38, 2569-2578 (2018).
8. Bolkan SS, Stujenske JM, Parnaudeau S, Spellman TJ, Rauffenbart C, Abbas AI, Harris AZ, Gordon JA, & Kellendonk C. Thalamic projections sustain prefrontal activity during working memory maintenance. *Nat. Neurosci.* 20, 987-996 (2017).
9. Halassa MM, & Kastner S. Thalamic functions in distributed cognitive control. *Nat. Neurosci.* 20, 1669-1679 (2017).
10. Wolff M & Vann S.D. The cognitive thalamus as a gateway to mental representations. *J. Neurosci.* 39, 3-14 (2019).
11. Chakraborty S, Kolling N, Walton ME, & Mitchell AS. Critical role for the mediodorsal thalamus in permitting rapid reward-guided updating in stochastic reward environments. *eLife* 5, e13588 (2016).
12. Parnaudeau S, Bolkan SS, & Kellendonk C. The mediodorsal thalamus: an essential partner of prefrontal cortex for cognition. *Biol. Psychiatry* 83, 648-656 (2018).
13. DeNicola AL, Park M-Y, Crowe DA, MacDonald 3rd AW, & Chafee MV. Differential roles of mediodorsal nucleus of the thalamus and prefrontal cortex in decision-making and state representation in a cognitive control task measuring deficits in schizophrenia. *J. Neurosci.* 40, 1650-1667 (2020).
14. Kosciessa, J.Q., Lindenberger, U. & Garrett, D.D. Thalamocortical excitability modulation guides human perception under uncertainty. *Nat. Commun.* 12, 2430 (2021).
15. Grinband, J., Hirsch, J. & Ferrera, V. P. A neural representation of categorization uncertainty in the human brain. *Neuron* 49, 757-763 (2006).
16. Schmitt LI, Wimmer RD, Nakajima M, Happy M, Mofakham S, & Halassa MM. Thalamic amplification of cortical connectivity sustains attentional control. *Nature* 545, 219-223 (2017).
17. Alcaraz F, Fresno V, Marchand AR, Kremer EJ, Coutureau E & Wolff M. Thalamocortical and corticothalamic pathways differentially contribute to goal-directed behaviors in the rat. *eLife* **7**, e32517 (2018).
18. Jaramillo J, Mejias JF, & Wang X-J. Engagement of pulvino-cortical feedforward and feedback pathways in cognitive computations. *Neuron*, 101, 321-336 (2019).
19. Yang, G.R., Joglekar, M.R., Song, H.F. et al. Task representations in neural networks trained to perform many cognitive tasks. *Nat. Neurosci.* 22, 297-306 (2019).
20. Song, H.F., Yang, G.R., & Wang, X.-J. Training excitatory-inhibitory recurrent neural networks for cognitive tasks: A simple and flexible framework. *PLoS Comput. Biol.* 12, e1004792 (2016).
21. Zhang, X., Liu, S. & Chen, Z.S. A geometric framework for understanding dynamic information integration in context-dependent computation. *iScience*, 24, 102919 (2021).
22. Zhang X, Long X, Zhang S-J, & Chen Z.S. Excitatory-inhibitory recurrent dynamics produced robust visual grids and stable attractors. *Cell Rep*., 41, 111777 (2022).
23. Xue, X., Wimmer, R.D., Halassa, M.M. & Chen, Z.S. Spiking recurrent neural networks represent task-relevant neural sequences in rule-dependent computation. *Cogn. Comput.* (2022). https://doi.org/10.1007/s12559-022-09994-2
24. Mante, V., Sussillo, D., Shenoy, K. et al. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* 503, 78-84 (2013).
25. Orhan, A.E. & Ma, W.J. A diverse range of factors affect the nature of neural representations underlying short-term memory. *Nat. Neurosci.* 22, 275-283 (2019).
26. Kepecs, A. & Fishell, G. Interneuron cell types are fit to function. *Nature* 505, 318-326 (2014).
27. Pulvermüller, F., Tomasello, R., Henningsen-Schomers, M.R. et al. Biological constraints on neural network models of cognitive function. *Nat. Rev. Neurosci*. 22, 488-502 (2021).
28. Hertag L & Sprekeler H. Amplifying the redistribution of somato-dendritic inhibition by the interplay of three interneuron types. *PLoS Comp. Biol.* 15, e1006999 (2019).
29. Keller AJ, Dipoppa M, Roth MM, et al. A disinhibitory circuit for contextual modulation in primary visual cortex. *Neuron*, 108, 1181-1193 (2020).
30. Poort J, Wilmes KA, Blot A, et al. Learning and attention increase visual response selectivity through distinct mechanisms. *Neuron*, 110, 686-697 (2022).
31. Pi, HJ., Hangya, B., Kvitsiani, D. et al. Cortical interneurons that specialize in disinhibitory control. *Nature* 503, 521-524 (2013).

32. Wood KC, Blackwell JM, & Geffen MN. Cortical inhibitory interneurons control sensory processing. *Curr. Opin. Neurobiol*. 46, 200-207 (2017).
33. Tao HW, Li Y-T, & Zhang LI. Formation of excitation-inhibition balance: inhibition listens and changes in tune. *Trends Neurosci.* 37, 528-530.
34. Ferguson BR, Gao W-J. Thalamic control and social behavior via regulation of gamma-aminobutyric acidergic signaling and excitation/inhibition balance in the medial prefrontal cortex. *Biol. Psychiatry*, 83, 657-669 (2018).
35. Contractor, A., Ethell, I.M. & Portera-Cailliau, C. Cortical interneurons in autism. *Nat Neurosci* 24, 1648-1659 (2021).
36. Fu Y, Tucciarone JM, Espinosa JS, et al. A cortical circuit for gain control by behavioral state. *Cell* 156, 1139-1152 (2014).
37. Halassa MM & Sherman SM. Thalamocortical circuit motifs: a general framework. *Neuron*, 103, 762-770 (2019).
38. Roy DS, Zhang Y, Halassa MM & Feng G. Thalamic subnetworks as units of function. *Nat. Neurosci.* 25, 140-153 (2022).
39. Shine, J.M., Müller, E.J., Munn, B., et al. Computational models link cellular mechanisms of neuromodulation to large-scale neural dynamics. *Nat Neurosci* 24, 765-776 (2021).
40. Izhikevich EM, Edelman GM. Large-scale model of mammalian thalamocortical systems. *Proc. Natl. Acad. Sci. USA*, 105, 3593-3598 (2008).
41. Krishnan GP, Ghauvette S, Shamie I, et al. Cellular and neurochemical basis of sleep stages in the thalamocortical network. *eLife*, 5, e18607 (2016).
42. Quax S, Jensen O, & Tiesinga P. Top-down control of cortical gamma-band communication via pulvinar induced phase shifts in the alpha rhythm. *PLoS Comput Biol* 13, e1005519 (2017).
43. Hummos A, Wang BA, Drammis S, Halassa MM, & Pleger B. Thalamic regulation of frontal interactions in human cognitive flexibility. *PLoS Computat. Biol.* 18, e1010500 (2022).
44. Anastasiades PG, Collins DP & Carter AG. Mediodorsal and ventromedial thalamus engage distinct L1 circuits in the prefrontal cortex. *Neuron*, 109, 314-330 (2021).
45. Hattori, R., Kuchibhotla, K., Froemke, R. et al. Functions and dysfunctions of neocortical inhibitory neuron subtypes. *Nat. Neurosci*. 20, 1199-1208 (2017).
46. Guet-McCreight A, Skinner FK, & Topolnik L. Common principles in functional organization of VIP/Calretinin cell-driven disinhibitory circuits across cortical areas. *Front. Neural Circuits*, 14, 32 (2020).
47. Delevich K, Tucciarone J, Huang ZJ, & Li B. The mediodorsal thalamus drives feedforward inhibition in the anterior cingulate cortex via parvabumin interneurons. *J. Neurosci.* 35, 5743-5753 (2015).
48. Delevich K, Jaaro-Peled H, Penzo M, Sawa A, & Li B. Parvalbumin interneuron dysfunction in a thalamo-prefrontal cortical circuit in Disc1 locus impairment mice. *eNeuro* 7, ENEURO.0496-19.2020 (2020).
49. Sohal VS, Zhang F, Yizhar O, & Deisseroth K. Paravalbumin neurons and gamma rhythms enhance cortical circuit performance. *Nature*, 459, 698-702 (2009).
50. Lewis, D., Hashimoto, T. & Volk, D. Cortical inhibitory neurons and schizophrenia. *Nat. Rev. Neurosci.* 6, 312-324 (2005).
51. Dienel SJ & Lewis DA. Alterations in cortical interneurons and cognitive function in schizophrenia. *Neurobiol. Dis.* 131, 104208 (2019).
52. Malik R, Li Y, Schamiloglu S, & Sohal VS. Top-down control of hippocampal signal-to-noise by prefrontal long-range inhibition. *Cell* 185, 1602-1617 (2022).
53. Vander Weele, C.M., Siciliano, C.A., Matthews, G.A. et al. Dopamine enhances signal-to-noise ratio in cortical-brainstem encoding of aversive stimuli. *Nature* 563, 397-401 (2018).
54. Mininni, C.J., Caiafa, C.F., Zanutto, B.S. et al. Putative dopamine neurons in the ventral tegmental area enhance information coding in the prefrontal cortex. *Sci. Rep.* 8, 11740 (2018).
55. Garcia-Cabezas MA, Martinez-Sanchez P, Sanchez-Gonzalez MA, et al. Dopamine innervation in the thalamus: monkey versus rat. *Cereb. Cortex* 19, 424-434 (2009).
56. Winterer G & Weinberger DR. Genes, dopamine and cortical signal-to-noise ratio in schizophrenia. *Trends Neurosci*. 27, 683-690 (2004).

57. Ferguson BR & Gao W-J. Development of thalamocortical connectivity between the mediodorsal thalamus and the prefrontal cortex and its implication in cognition. *Front. Hum. Neurosci*. 8, 1027 (2014).
58. Kohn A, Jasper AI, Semedo JD, Gokcen E, Machens CK, & Yu BM. Principles of corticocortical communication: proposed schemes and design considerations. *Trends Neurosci.* 43, 725-737 (2020).
59. DeNicola AL, Park M-Y, Crowe DA, MacDonald 3rd AW, & Chafee MV. Differential roles of mediodorsal nucleus of the thalamus and prefrontal cortex in decision-making and state representation in a cognitive control task measuring deficits in schizophrenia. *J. Neurosci.* 40, 1650-1667 (2020).
60. Murray JD & Anticevic A. Toward understanding thalamocortical dysfunction in schizophrenia through computational models of neural circuit dynamics. *Schizophr. Res.* 180, 70-77 (2017).
61. Lewis DA, Hashimoto T, Volk DW. Cortical inhibitory neurons and schizophrenia. *Nature Rev. Neurosci.* 6, 312-324 (2015).
62. Mukherjee A, Halassa MM. The associative thalamus: a switchboard for cortical operations and a promising target for schizophrenia. *The Neuroscientist*. (2022) doi:10.1177/10738584221112861
63. Ouhaz Z, Flemming H, & Mitchell AS. Cognitive functions and neurodevelopmental disorders involving the prefrontal cortex and mediodorsal thalamus. *Front. Neurosci.* 12, 33 (2018).
64. Benoit, L.J., Holt, E.S., Posani, L. et al. Adolescent thalamic inhibition leads to long-lasting impairments in prefrontal cortex function. *Nat. Neurosci.* 25, 714-725 (2022).
65. Mukherjee A, Bajwa N, Lam NH, Porrero C, Clasca F, & Halassa MM. Variation of connectivity across exemplar sensory and associative thalamocortical loops in the mouse. *eLife,* 9, e62554 (2020).
66. Bulin SE, Hohl KM, Paredes D, Silva JD, & Morilak DA. Bidirectional optogenetically-induced plasticity of evoked responses in the rat prefrontal cortex can impair or enhance cognitive set-shifting. *eNeuro* 7, ENEURO.0363-19.2019 (2019).

Online Methods

Computational tasks with parameterized uncertainties for context-dependent cognition. The computational task was designed and motivated by a combination of a version of the two-context two-modality four alternative forced choice (4AFC) task[16] and a single-modality cue-conflicting two alternative force choice (2AFC) task[4]. **Figure 1a** illustrates the two cueing contexts that were mapped to the same rule: attend-to-audition (rule 1) or attend-to-vision (rule 2). Each trial started with a 40-ms trial initialization period, followed by 800-ms stimulus cueing. In the cueing period, the neural network received cues from either visual or auditory modality, each of which had parameterized cue uncertainty or conflicting cues. In the visual modality context, the random dot motion (RDM) subtask used cues as random dots in motion, where the coherence of moving dots in time indicates the degree of uncertainty: with 1 meaning all dots moving with the same direction (rightward, rule 1 vs. leftward, rule 2) and 0 meaning completely random direction. The default dimensionality of cue input in RDM was 64. In the auditory modality context, the conflicting acoustic cue (CAC) subtask used cues as a sequence of conflicting pulses: low-pass (LP) pulse (+1) or high-pass (HP) pulse (-1). The {+1,-1} pulses were task-relevant cues, where the background was represented by 0. The number of pulses during the 800-ms cueing period was 32 in the training dataset, giving a signal density 40 s^{-1}. The mixture of sensory cues determined the uncertainty. Accumulating the sensory evidence (i.e., the total or relative number of LP and HP pulses) would produce the indication of rule 1 (more +1's) or rule 2 (more -1's). During the cueing period, the neural network first learned to accumulate the sensory evidence of ambiguous cues to recognize the rule. During the follow-up 800-ms delay period, the neural network learned to preserve the rule information during working memory. During the 200-ms target period, the auditory and visual stimuli (which were used differently from the cueing period) were simultaneously presented (e.g., upsweep/downsweep tone that was mapped to auditory left/right choice, respectively; yellow/blue light that was mapped to visual left/right choice, respectively)[16], and each attended target was associated with a unique choice. Therefore, the network learned to attend to the correct modality and select one of four choices (**Fig. 1a**). Together, the neural network learned a cross-modal context-dependent working memory and decision-making task, which mapped the ambiguous cueing signals and rules to decision. As a result, the task error could be ascribed to either rule error or sensory error, or both. Specifically, errors related to rule encoding was termed the "rule error" (or "executive error", which occurred in cueing and delay periods), whereas errors related to target stimulus perception was termed the "sensory error" (which occurred in the post-delay period)[16].

	Rule and Context-Dependent Cueing			
	Rule 1: attend-to-audition Cue: A_{LP} or V_{right}		Rule 2: attend-to-vision Cue: A_{HP} or V_{left}	
Response \ Target	Auditory left	Auditory right	Visual left	Visual right
Choice #1 (auditory left)	correct	sensory	rule	both
Choice #2 (auditory right)	sensory	correct	both	rule
Choice #3 (visual left)	rule	both	correct	sensory
Choice #4 (visual right)	both	rule	sensory	correct

In theory, the chance-level accuracy for the 4AFC task is 25%. However, the rule error and sensory error are independent in the task design. For the rule error alone, the chance-level accuracy is 50%. To introduce a proper degree of sensory error in the target period, we assumed two competing sensory inputs were stochastic and drawn from normal distributions with two different mean but the same variance. Increasing the variance would introduce the task difficulty and increase the sensory error; we selected a variance value that matched most to the experimental data. In the extreme case where sensory inputs were deterministic, the model prediction would lead to only rule error and minimal sensory error.

Computational models. We first constructed a PFC-alone network as the baseline model. The PFC network is an excitatory-inhibitory (E/I) recurrent neural network (RNN) with N_{PFC}=256 fully interconnected units described by a standard firing-rate model[20,22]. We adopted a continuous-time formulation of RNN dynamics as follows:

$$\tau\dot{\mathbf{x}}^{\mathrm{PFC}} = -\mathbf{x}^{\mathrm{PFC}} + \mathbf{W}^{\mathrm{rec}}\mathbf{r}^{\mathrm{PFC}} + \mathbf{W}^{\mathrm{in}}\mathbf{u} + \sigma\boldsymbol{\xi}$$

where τ denotes the time constant (we used τ=20 ms), $\boldsymbol{\xi}$ denotes additive N_{PFC}-dimensional Gaussian noise, each independently drawn from a standard normal distribution, σ defines the scale of the noise

standard deviation; $\mathbf{W}^{\mathrm{rec}}$ is an $N_{\mathrm{PFC}}\times N_{\mathrm{PFC}}$ matrix of recurrent connection weights, and $\mathbf{W}^{\mathrm{in}}$ denotes an $N_{\mathrm{PFC}}\times N_{\mathrm{in}}$ matrix of connection weights from the input to network units (N_{in} =72). The network output consisted a set of linear readout units that produced a 4-dimensional target estimate: $\mathbf{z} = \mathbf{W}^{\mathrm{out}}\,\mathbf{r}$, where $\mathbf{W}^{\mathrm{out}}$ is a $4\times N_{\mathrm{PFC}}$ matrix, and the N_{PFC}-dimensional neuronal firing rate vector $\mathbf{r}$ is defined by a nonnegative rectified linear unit (ReLu) function: $\mathbf{r} = \phi(\mathbf{x})$. The final decision layer employed a softmax function to select the maximum activation that yielded one out of four choices. We assumed 4:1 ratio of the number of excitatory to inhibitory neurons and imposed Dale's principle on synaptic connectivity. Therefore, the recurrent weight matrix consisted of functionally distinct connectivity according to the cell types: excitatory-to-excitatory, inhibitory-to-excitatory, excitatory-to-inhibitory, and inhibitory-to-inhibitory connections. Among cortical inhibitory neurons, we further considered three major neuron subtypes: parvalbumin (PV), vasoactive intestinal peptide (VIP)-expressing and somatostatin (SOM) interneurons. The interneuron cell types had distinct inhibitory-to-excitatory and inhibitory-to-inhibitory connectivity[26-29]. The most common PV interneurons target excitatory neuronal cell bodies, and the second most common SOM interneurons target the distal dendrites of postsynaptic excitatory neurons. VIP interneurons preferentially target SOM cells, and to a lesser extent PV cells, so we did not consider VIP→PV connectivity here; therefore, VIP interneurons can modulate cortical excitatory neurons through disinhibition of SOM cells (**Fig. 1b**). We assumed that the percentages of three interneuron cell types were equal and these subtypes were determined by the synaptic interconnectivity. To discretize continuous-time dynamics, we used a bin size of Δt=20 ms in all numerical simulations.

We further constructed a PFC-MD model based on the knowledge of genetically identified thalamocortical projections and cell type specificity[4]. In light of experimental findings[4], we assumed that the MD primarily consists of non-recurrent excitatory neurons (N_{MD}=200) and have two specific MD→PFC pathways: where the MD_1 subpopulation projects to cortical PV interneurons, whereas the MD_2 subpopulation projects to cortical VIP interneurons. Meanwhile, MD neurons received reciprocal projections from PFC excitatory neurons. We assumed an equal number of excitatory units in both MD_1 and MD_2 subpopulations. The MD and PFC units both received an N_{in}-dimensional stimulus input, but only the PFC units were used to generate a 4-dimensional choice output. There were fully connected corticothalamic connections (of size $N_{\mathrm{PFC}}\times N_{\mathrm{MD}}$) from PFC units to MD_1 and MD_2 subpopulation units, separately. For convenience, we let $\mathbf{W}^{\mathrm{eff}}$ represent the effective PFC-MD connectivity matrix of size $(N_{\mathrm{PFC}}+N_{\mathrm{MD}})\times(N_{\mathrm{PFC}}+N_{\mathrm{MD}})$. In this case, the prefrontal dynamics were controlled by both intracortical weights and thalamocortical weights[67]

$$\tau\dot{\mathbf{x}}^{\mathrm{PFC}} = -\mathbf{x}^{\mathrm{PFC}} + \mathbf{W}^{\mathrm{rec}}\mathbf{r}^{\mathrm{PFC}} + \mathbf{W}^{\mathrm{MD}\to\mathrm{PFC}}\mathbf{r}^{\mathrm{MD}} + \mathbf{W}^{\mathrm{in}}\mathbf{u} + \sigma\boldsymbol{\xi}$$

And the MD dynamics were controlled by corticothalamic weights:

$$\tau\dot{\mathbf{x}}^{\mathrm{MD}} = -\mathbf{x}^{\mathrm{MD}} + \mathbf{W}^{\mathrm{PFC}\to\mathrm{MD}}\mathbf{r}^{\mathrm{PFC}} + \mathbf{V}^{\mathrm{in}}\mathbf{u} + \sigma\boldsymbol{\xi}$$

where $\mathbf{V}^{\mathrm{in}}$ denotes an $N_{\mathrm{MD}}\times N_{\mathrm{in}}$ matrix. The output setup was identical to the PFC-alone model, where the readout only depended on the PFC units but not MD units.

We trained the PFC-alone and PFC-MD models by supervised learning. The two subtasks, CAC and RDM tasks, were learned separately or simultaneously. The CAC or RDM subtask consisted of a batch size of 64 simulated trials used for training. The number of training trials for two rules (attend-to-audition vs. attend-to-vision) was balanced. We used the mean-squared-error (MSE) plus weight decay as the regularized cost function and optimized the model parameters using a gradient-descent algorithm[20], with default configuration of hyperparameters (learning rate parameter 0.0005; weight decay regularization parameter 0.1). The optimization continued until the desired accuracy was reached. In total, we trained more than 20 independent PFC-MD networks (and an equally matched number of PFC-alone networks), and selected their respective best models to present the results (however, the psychometric curves and neural representations were quite robust among most performance-optimized models). In testing, we varied the degree of cue uncertainty and simulated the 500 independent random test trials per condition in order to compute the psychometric curve. In the CAC subtask, we changed the ratio of LP/HP pulses for different cue uncertainties, and varied the timing and order of the LP and HP pulses for a fixed uncertainty (**Fig. 1g**). In the RDM subtask, we varied the percentage of dots moving rightward or leftward for different cue uncertainties, and selected different subsets of dots in time (**Fig. 1j**). We also observed that some performance-optimized PFC-MD and PFC-alone models did not show good generalization across all tested cue uncertainty and sparsity conditions (e.g., making more one type of rule-specific error), which we have excluded in the analyses.

Behavioral analysis and psychometric curve. The model performance was assessed by a psychometric curve, where the accuracy (percentage of correct) was calculated with varying degrees of cue uncertainty. In the CAC task, the LP/HP congruence level (range: [0.5,1.0]) was defined by a ratio: $\max\left\{\frac{\#\text{LP pulses}}{\#\text{LP pulses}+\#\text{HP pulses}}, \frac{\#\text{HP pulses}}{\#\text{LP pulses}+\#\text{HP pulses}}\right\}$, with 0.5 being maximally uncertain and 1 being maximally certain. In the RDM task, the cue uncertainty was calculated by the degree of coherence (range: [-1.0, +1.0]) of random dot motion (with absolute coherence of 1 indicating maximal certainty and noise-free, and coherence of 0 being completely uncertain). Furthermore, we used the algebraic sign to represent the motion direction (with positive/negative values moving rightward/leftward, respectively)[24]. In testing, we generated 100 random realizations of input sequences and calculated the task errors for producing performance curves (mean ± s.e.m.). Based on the percentage of correct choices, we further fit a sigmoid-shaped psychometric curve function using maximum likelihood estimation (**Fig. 1e**). Note that in all psychometric curves, only the rightmost cue condition in the x-axis (i.e., lowest cue uncertainty) was used for training the models, and the remaining cue conditions were never used in training.

Single-unit tuning curve analysis. We computed the trial-averaged peri-stimulus time histogram (PSTH) from single unit firing activity for specific task condition (rule/context/cue uncertainty/cue sparsity). We identified rule or context modulation by comparing their respective mean firing rate as well as temporal PSTH profiles during both cueing and delay periods (**Fig. 2a-e**). To define the cue uncertainty tuning, we varied the cue coherence or congruence level (high/medium/low) (**Fig. 2f,g,i**) and compared the respective PSTH profiles. A unit was called to have cue uncertainty modulation if and only if the PSTH comparisons were significantly different between all three paired cue uncertainty levels (high vs. medium, medium vs. low, high vs. low) during the cueing period.

Neural sequence analysis. To examine potential neural sequences during the delay period, we normalized the PSTHs of PFC excitatory units between 0 and 1 and sorted by the latency of their peak firing rates. Units with peak firing rate smaller than 0.2 during the delay period were excluded in this analysis. We compared the neural sequence among the selected subpopulation between rules and between contexts using the same rank order (**Extended Data Fig. 4**).

Unit pairwise correlation analysis. We computed the zero-lag cross-correlation of demeaned firing activities of multiple single units on a trial-by-trial basis. Context-dependent, rule-specific correlation was defined based on temporally concatenated single-trial activity of all training trials (**Extended Data Fig. 5b**).

Canonical correlation analysis. We employed canonical correlation analysis (CCA) to examine the trial-by-trial canonical correlation between the MD and PFC population activity from the performance-optimized PFC-MD model. Briefly, CCA finds pairs of dimensions, one for MD and the other for PFC, such that the correlation between the projected activity onto these dimensions is maximally correlated:

$$\arg\max_{\mathbf{a},\mathbf{b}} \operatorname{corr}(\mathbf{X}_{\mathrm{PFC}}\mathbf{a}, \mathbf{X}_{\mathrm{MD}}\mathbf{b})$$

where $\mathbf{X}_{\mathrm{PFC}}$ and $\mathbf{X}_{\mathrm{MD}}$ denote the matrices consisting of temporally binned N_{PFC} -dimensional PFC and N_{MD}-dimensional MD population activities, respectively. The vectors **a** and **b** have respective dimensions $N_{\mathrm{PFC}}\times 1$ and $N_{\mathrm{MD}}\times 1$, defining dimensions in the population activity space of each area. In MATALAB script, we used [**A**,**B**,**r**] = canncorr($\mathbf{X}_{\mathrm{PFC}}$, $\mathbf{X}_{\mathrm{MD}}$), such that **a** = **A**(:,1), **b** = **B**(:,1), and **r**(1) corresponded to the maximum canonical correlation (**Fig. 3c**). We used the maximum canonical scores at each trial to measure the overall population correlation for a given experimental condition or task period. A zero time-lag was used in our analysis.

Prefrontal signal-to-noise ratio (SNR) metric. The SNR measure has been adopted in neuroscience applications[68,69]. For a rule-modulated PFC unit, the neural SNR was defined as the ratio between the peak of tuning curve and the baseline activity. For neuronal populations, correlation between individual units may encode information[70]. We defined the functional interaction of units (as defined by a cross-covariance matrix) as 'signal' during the task phase and 'noise' during the baseline, and then perform spatial filtering and projection onto a latent space. We further defined a neural SNR measure based on functional connectivity of population activity: SNR=$\frac{\mathbf{w}^T\mathbf{C}_s\mathbf{w}}{\mathbf{w}^T\mathbf{C}_n\mathbf{w}}$, where $\mathbf{C}_s$ denotes trial-averaged covariance matrix during working memory, and $\mathbf{C}_n$ denotes the covariance matrix in baseline. In our experiment, $\mathbf{C}_s$ was computed from the

population firing activity from concatenated correct trials during the delay period (with a specific cueing input), whereas $\mathbf{C}_n$ was computed from the population firing activity from concatenated trials by setting the cueing input to zero in a null model (i.e., spontaneous state baseline). We maximized the neural SNR by optimizing the spatial filters by solving a generalized eigenvalue decomposition problem, where $\frac{\mathbf{w}^T\mathbf{C}_s\mathbf{w}}{\mathbf{w}^T\mathbf{C}_n\mathbf{w}}$ also denotes the generalized Rayleigh quotient, where **w** was derived from the dominant eigenvector that defined the optimal filter for producing the maximal Rayleigh quotient. In MATALAB script, the generalized eigenvalue decomposition operation was written as [**V, D**] = eig($\mathbf{C}_s$,$\mathbf{C}_n$) such that $\mathbf{C}_s\mathbf{V} = \mathbf{C}_n\mathbf{V}\mathbf{D}$, with the largest eigenvalue $\lambda = \mathbf{D}(1,1)$ and the dominant eigenvector **w** = **V**(:,1). We computed $\mathbf{C}_s$ and $\mathbf{C}_n$ based on temporally binned firing rate activity across trials. In experiments, we computed $\mathbf{C}_s$ separately for trials with low and high conflicts, and used the common $\mathbf{C}_n$ to derive the respective spatial filter **w**. We then computed the prefrontal SNR for specific cue or task conditions (**Fig. 3a,b**). Assuming that prefrontal neurons are coordinated to regulate information transmission, the proposed prefrontal SNR measure defines an upper bound of the ratio of task-dependent population variance against baseline variance.

Population decoding analysis. We employed a binary support vector machine (SVM) classifier to achieve trial-by-trial classification to decode task rule or context, based on the simulated population activity of PFC or MD units. The software was implemented under the Sciket-learn Python environment. The network received an *N*-by-1 input, with each representing the mean firing activity of one unit. Among the total 200 simulated trials, we divided them into halves: 50% for training and remaining 50% for testing. We reported the average cross-validation decoding accuracy based on 20 independent Monte Carlo runs. We conducted with decoding analysis under various cue uncertainty conditions (**Fig. 3d,e**).

Dimensionality reduction analysis. To extract low-dimensional representation of population activity from the PFC-MD network, we employed principal component analysis (PCA) to visualize neural trajectories during the cueing and delay periods. Briefly, we first defined the task-related axes and grouped neural population activity into a matrix $\mathbf{X} \in \mathbb{R}^{N\times CT}$, where *N* denotes the number of units, *C* denotes the number of stimulus conditions, and *T* denotes the number of time bins. We performed PCA separately on temporally-binned data matrix $\mathbf{X}$ at specific or combined task periods, and further extracted two dominant principal subspaces associated with the two largest eigenvalues. The derived low-dimensional neural trajectory showed task-variable-nonspecific representations.

To further visualize the neural trajectory in task-variable-axes (e.g., choice of rule, cue, uncertainty), we further defined the task-relevant state space using a published method[24]. First, we used the 20 largest principal components (PCs) and denoised the representation of neural activity. Second, we constructed a regression model for each unit based on task variables for two context-dependent subtasks: cue and choice of rule. The regression coefficients across units defined an axis in the neural state space representing each task variable. After performing QR orthogonalization, these axes formed the task-relevant state space onto which the population response in the 20-dimensional PC subspace was projected, providing a denoised and demixed task-relevant neural representation. Unless stated otherwise, we used only correct trials to produce trial-averaged curves in two-dimensional subspace (**Fig. 4a-f**). From a neural trajectory $\boldsymbol{p}(t)$, we further defined the time-varying neural velocity by computing the change rate in neural trajectory[21]: $v(t) = \|\dot{\boldsymbol{p}}(t)\|$ (**Fig. 4g**).

Modeling phasic changes in MD neuronal firing. To model optogenetic activation or suppression effect on the MD neuronal firing, we introduced a depolarizing or hyperpolarizing single-pulse input to the targeted MD neurons, which further led to an increase or decrease in their phasic firing activities, respectively (**Fig. 6b-d**).

Fixed-point analysis. Similar to our previous analysis[21,22], we identified fixed-points or slow points of the performance-optimized PFC-MD model by numerically solving the optimization problem (https://github.com/mattgolub/fixed-point-finder)

$$\min_{\tilde{\mathbf{x}}} q(\tilde{\mathbf{x}}), \quad \text{where } q(\tilde{\mathbf{x}}) = \left\|-\tilde{\mathbf{x}} + \mathbf{W}^{\text{eff}}\phi(\tilde{\mathbf{x}}) + [\mathbf{W}^{\text{in}}, \mathbf{V}^{\text{in}}] \cdot \mathbf{u}\right\|^2$$

where $\tilde{\mathbf{x}} = [\mathbf{x}^{\mathrm{PFC}}, \mathbf{x}^{\mathrm{MD}}]$. We collected a set of fixed-points by randomly initializing the network. The dimensionality of fixed points was the same as dim($\tilde{\mathbf{x}}$)=N_{PFC}+N_{MD}. Once the numerical optimization was completed, we applied PCA to visualize the fixed points in three-dimensional PC subspace (**Fig. 4i**).

Eigenvalue analysis and time constant estimation. A square matrix $\mathbf{A}$ is non-normal if it satisfies $\mathbf{A}\mathbf{A}^T \neq \mathbf{A}^T\mathbf{A}$. For the non-normal matrix $\mathbf{W}^{\mathrm{eff}}$ (in the PFC-MD model) or $\mathbf{W}^{\mathrm{rec}}$ (in the MD-lesioned model), we conducted eigenvalue analysis to analyze the nonlinear dynamics with linear approximation[22,71], where the maximum of the real (or real-part) eigenvalues characterizes the long-term behavior for the speed of network's steady state decaying to zero[72]. We further we adapted a published method to estimate the intrinsic time constant of dynamical systems[73], for both PFC-MD and MD-lesioned structures (**Supplementary Fig. 2b**):

$$\tau_{\mathrm{intrinsic}} = \max_k \left(-\frac{\Delta t}{\log\ |\lambda_k|} \right)$$

where $|\lambda_k|$ denotes the absolute value of the *k*-th real or complex-valued eigenvalue. Since a large time constant is beneficial for temporal summation, a larger time constant can help information integration (during the cueing period) and working memory maintenance (during the delay period).

Statistics and reproducibility. No statistical methods were used to predetermine sample sizes. Data collection and analysis were not performed blind to the conditions of the experiments. Statistical analysis was performed using Python or MATLAB. Error bars are shown as means ± s.e.m. or means ± SD. We used two-sided Wilcoxon signed-rank or rank-sum test for all paired and unpaired statistical tests, respectively. These tests are nonparametric and do not assume a specific distribution for the data.

Reporting Summary. Further information on research design is available in the Reporting Summary linked to this article.

Data availability

The data generated from computer simulations that support the findings of this study will become publicly available.

Supplementary references

67. Logiaco L, Abbott LF & Escola S. Thalamic control of cortical dynamics in a model of flexible motor sequencing. *Cell Rep.* 35, 109090 (2021).
68. Zohary E, Shadlen MN & Newsome WT. Correlated discharge rate and its implication for psychophysical performance. *Nature* 370, 140-143 (1994).
69. Czanner G, Sarma SV, Ba D, et al. Measuring the signal-to-noise ratio of a neuron. *Proc. Natl. Acad. Sci. USA*, 112, 7141-7146 (2015).
70. Averbeck, B., Latham, P. & Pouget, A. Neural correlations, population coding and computation. *Nat. Rev. Neurosci*. 7, 358-366 (2006).
71. Rajakumar A, Rinzel J, & Chen ZS. Stimulus-driven and spontaneous dynamics in excitatory-inhibitory recurrent neural networks for sequence representation. *Neural Computat.* 33, 2603-2645 (2021).
72. Asllani M, Lambiotte R, & Carletti T. Structure and dynamical behavior of non-normal networks. *Science Adv.* 4, sciadv.aau9403 (2018).
73. Spitmaan M, Seo H, Lee D, & Soltani A. Multiple timescales of neural dynamics and integration of task-relevant signals across cortex. *Proc. Natl. Acad. Sci. USA* 117, 22522-22531 (2020).

Acknowledgements

We thank A. Mukherjee, N. Lam, R.D. Wimmer and Y. Liu for valuable feedback on the manuscript. The work was supported by grants MH118928 (Z.S.C.), DA056394 (Z.S.C.) from the National institutes of Health.

Author contributions

Z.S.C. conceived and supervised experiments, analyzed and interpreted the data, and wrote the paper. X.Z. developed the computational models, performed experiments, analyzed and interpreted the data.

M.M.H. provided experimental data that supported the computational model and interpreted the data. Z.S.C. acquired funding.

Competing interests
The authors declare no competing interests.

FIGURE LEGENDS

Figure 1. The performance-optimized PFC-MD model in decision making with sensory uncertainty. (**a**) Excitation-inhibition (E/I) recurrent neural network (RNN) for modeling the PFC-alone network, where three major prefrontal interneuron cell types were specified. (**b**) The PFC-MD model consisted of two target-specific non-recurrent excitatory MD subpopulations and bidirectional MD-PFC projections. (**c**) Schematic of a cross-modal, context-dependent decision-making task with working memory and divided attention components. Sensory cues were either conflicting LP/HP pulses or moving random dots. (**d**) Schematic of context-dependent rule encoding. The RDM and CAC subtasks learned the same rule: attend-to-audition vs. attend-to-vision. (**e**) Psychometric curves of RDM and CAC subtasks. Error bar was shown in mean $\pm$ s.e.m. Each condition was repeated 10 times with different input realizations yet with identical summary statistics. The PFC-MD model outperformed the PFC-alone model in the presence of intermediate-to-high cue conflict. (**f**) Percentage of task error types in two subtasks under various cue uncertainty conditions. (**g**) Illustration of three types of cues with two congruence levels and two densities in the CAC subtask. (**h**) Psychometric curve of the PFC-MD model under three different cue sparsity levels in the CAC subtask. (**i**) Psychometric curve comparison between the PFC-MD and PFC-alone models in the CAC subtask in one cue sparsity condition. (**j**) Illustration of three types of cues with two coherence levels and two densities in the RDM subtask. (**k**) Psychometric curve of the PFC-MD model under three different cue sparsity levels in the RDM subtask. (**i**) Psychometric curve comparison between the PFC-MD and PFC-alone models in the RDM subtask in one cue sparsity condition.

Figure 2. Neural representations of PFC and MD units from the performance-optimized PFC-MD model. (a) Two representative PFC excitatory units encoding two rules under two contexts. These two units were cue invariant units that encoded the same rule. (**b**) Two representative PFC inhibitory units encoding the cueing context. (**c**) Two representative MD units encoding context but no rule. (**d**) Population statistics of mean firing rates (during the task delay period) of excitatory PFC units and MD units for encoding rule. (**e**) Population statistics of mean firing rate (during the task delay period) of excitatory PFC units and MD units for encoding context. (**f**) Two representative MD_1 units that showed firing rate modulation with respect to cue uncertainty during both cueing and delay periods. (**g**) Two representative MD_2 units that showed no firing rate modulation with respect to cue uncertainty. (**h**) The same MD_1 units also showed firing rate modulation with respect to cue sparsity. (**i**) Two representative PFC inhibitory units that showed firing rate modulation with respect to cue uncertainty during both cueing and delay periods.

Figure 3. Information estimates in prefrontal and MD populations. (**a**) Comparison of signal-to-noise ratio (SNR) in prefrontal computation under different cue uncertainty and task periods. Error bar denotes s.e.m. (n=60 runs with random seeds, 100 simulated trials per run). *P*-values from two-tailed Wilcoxon rank-sum tests were used to compare between conditions (**, P<0.01; ***, P<0.001). (**b**) Disconnecting MD-PFC connectivity reduced the SNR statistics (n=60, $P=7\times10^{-9}$, RDM cueing and delay periods; $P=3\times10^{-8}$, CAC cueing and delay periods; Wilcoxon signed-rank test), where each trial was generated from a random seed. (**c**) Increased cue uncertainty resulted in increased canonical correlation between the PFC and MD activities during cueing and delay periods (n=60 runs, 100 simulated trials per run; two-tailed Wilcoxon rank-sum tests, *****, $P< 10^{-10}$). Error bar denotes s.e.m. (**d**) Population decoding analysis showed that the PFC population better encoded rule, whereas the MD population better encoded context. Error bar denotes s.e.m. (n=20). (**e**) Increased in cue uncertainty caused decreased rule decoding accuracy for the PFC, but did not affect context decoding accuracy for the MD. Error bar denotes SD (n=20). In rule decoding, all paired comparisons were statistically significant ($P<7\times10^{-8}$, two-tailed Wilcoxon rank-sum tests).

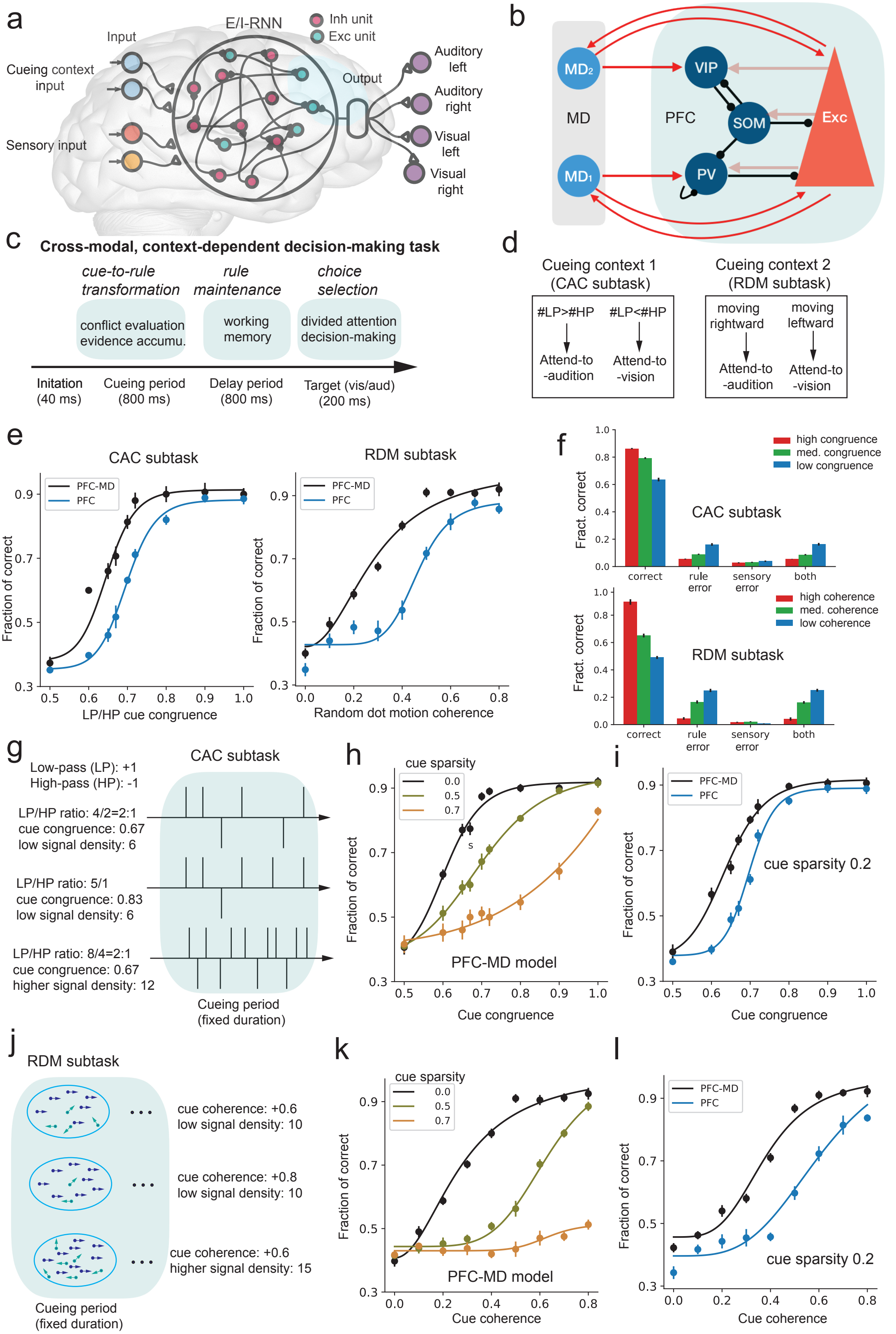
Figure 4. Population representations from the performance-optimized PFC-MD model. (**a**) Dynamics of population responses in the CAC subtask. The average population trajectory for a given condition and time was represented as a point in the state space. Responses from correct trials only were shown from the cue onset to the end of target period (80-ms step size), and were projected into the two-dimensional subspace capturing the variance according to the rule (attend-to-audition vs. attend-to-vision) and choice (see **Methods**). Units are arbitrary. The origin represents the cue onset. (**b**) Similar to panel **a**, except shown from the cue onset to the end of delay period. For better illustration, three levels of cue uncertainty were shown by different shades of grey or blue color (dark/intermediate/light shade: low/medium/high cue uncertainty, respectively). (**c**) Similar to panel **b**, except for a fixed cue uncertainty but different levels of cue sparsity. Three levels of sparsity were shown by different shades of grey or blue color (dark/intermediate/light shades: low/medium/high sparsity, respectively). (**d,e,f**) Similar to panels **a,b,c**, but for the RDM subtask. (**g,h**) Comparison of neural velocity for different levels of cue uncertainty for the CAC and RDM subtasks, respectively. (**i**) Fixed-point analysis in the three-dimensional PC subspace revealed two 'line-attractor-like' basins, with each basin representing a rule. Each cross symbol corresponded to a fixed point (color sorted by cue uncertainty). The red origin represents the cue onset. Inset: rotating the view angle and collapsing these points revealed two 'fixed-point-like' attractors.

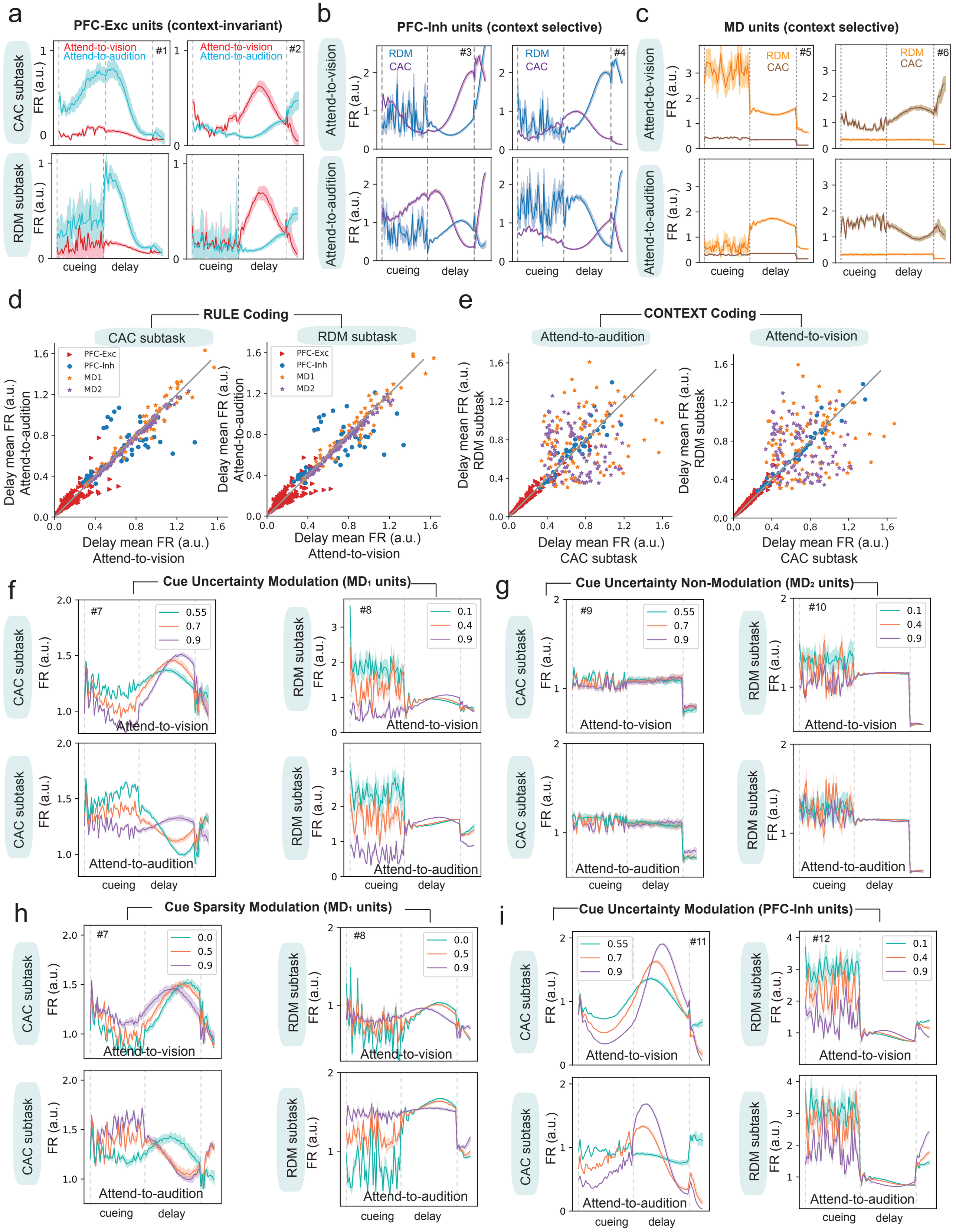
Figure 5. Cell type specific MD-PFC pathways shows differential functional roles. (a) Weakening $MD_1 \rightarrow PV$ connections (by increasing the percentage of zeros) during the cueing period quickly reduced the task performance in both RDM and CAC subtasks. Inset: increasing MD_1 activity could rescue each subtask performance under a wide range of connectivity conditions (shaded area: 20%-50% of zeros). Error bar denotes SD (n=10). **(b)** Weakening $MD_2 \rightarrow PV$ connections only degraded the task performance slowly in both RDM and CAC subtasks, but increasing MD_2 activity had no effect on the change in task performance (see inset). **(c)** Under cue uncertainty, increasing the MD_1 subpopulation mean firing rate (FR) during the cueing period could boost the task performance. In contrast, increasing the MD_2 subpopulation mean FR had little impact on the task performance. Zero denotes the baseline FR, and other numbers represent the percentages of increase relative to baseline. **(d)** MD_1 stimulation improved task performance in RDM and CAC subtasks under various levels of cue uncertainty. (**e**) MD_1 stimulation improved task performance in RDM and CAC subtasks under mixed cue uncertainty and cue sparsity conditions.

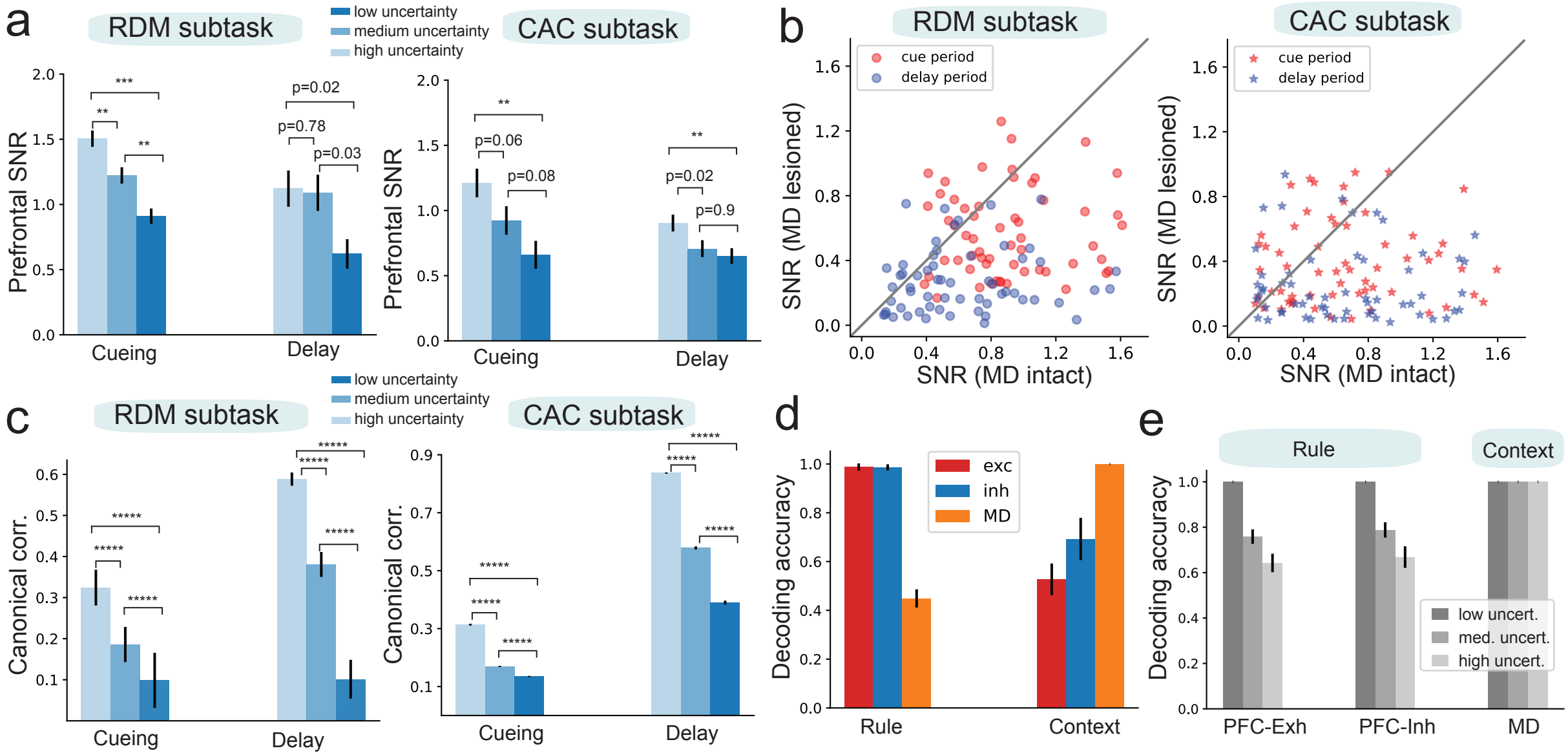
Figure 6. MD enhances working memory maintenance in the PFC-MD network. (a) Increasing the duration of task delay period reduced the task performance under cue uncertainty in both subtasks. Figure legend denotes the level of cue coherence or congruence. Error bar denotes SD (n=10). The task condition marked with "□" was used in the illustrations of three remaining panels. (**b,c**) Increasing the MD_1 or MD_2 population firing rate during an elongated delay period improved the working memory and the psychometric curve in both subtasks. Black curve denotes the baseline, and the number of % denotes the relative increase in subpopulation firing rate. (**d**) Decreasing the MD_1 population firing rate during an elongated delay period had an opposite effect on the task performance.

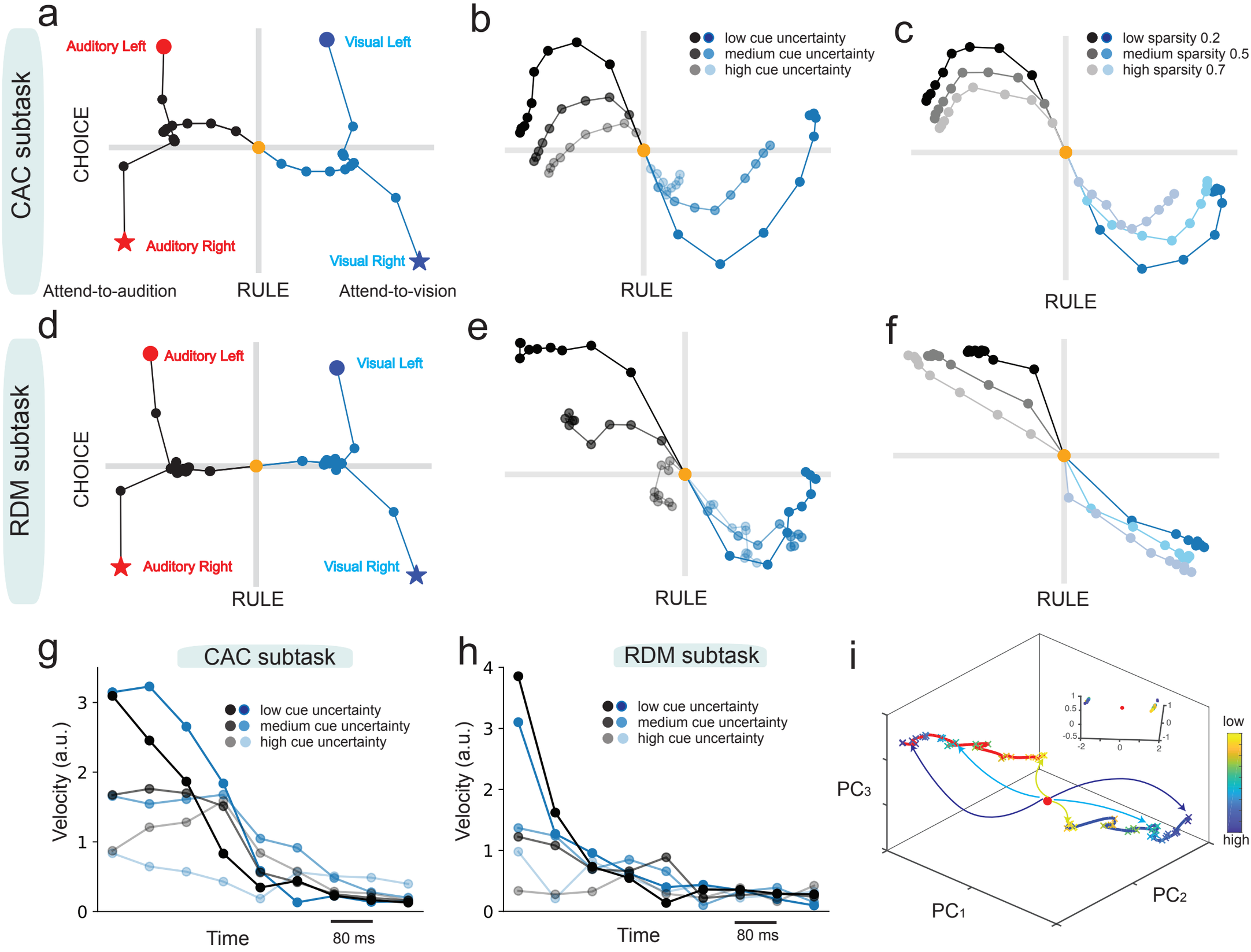
Figure 7. Parsing cognitive deficits and probing mechanistic causes in the PFC-MD networks. (a) During the control condition, PFC excitatory units (4 selected units are indicated by different colors) preserved rule tunings in both delay and target periods (left panel). Mean firing rate (FR) comparison of PFC excitatory units between two rules in delay (middle panel) and target (right panel) periods. Four units in the left panels are labeled in the same color in both middle and right panels. (**b**) In the presence of E/I imbalance (e.g., reduced inhibition), PFC excitatory units shown in panel **a** increased their firing rates but decreased their rule discriminability during the target period. (**c**) Two-dimensional neural trajectory representations in control and E/I imbalance conditions. Trajectories were generated from all correct and error trials. Orange dots: cue off and start of delay period. Blue and red end points represent two rule representations, whereas dot and star symbols represent left and right choices, respectively. **(d)** The task performance decreased with reduced prefrontal SOM→VIP and SOM→PV connectivity. **(e)** Increasing the mutual inhibition strength between SOM and VIP neurons amplified the gain under a high cue sparsity. Stimulating the MD_2→VIP pathway further facilitated the amplification and improved the task performance. Two shaded circles indicate the conditions illustrated in panel **f**. **(f)** In the case of RDM task of panel **e** (with cue sparsity 0.8), comparison of PFC excitatory unit tunings and two-dimensional neural trajectories between scale=1 (light gray, baseline) and scale = 1.2 for bidirectional SOM-VIP connection strengths. In the latter case, single-unit rule tunings emerged and population responses improved rule discriminability.

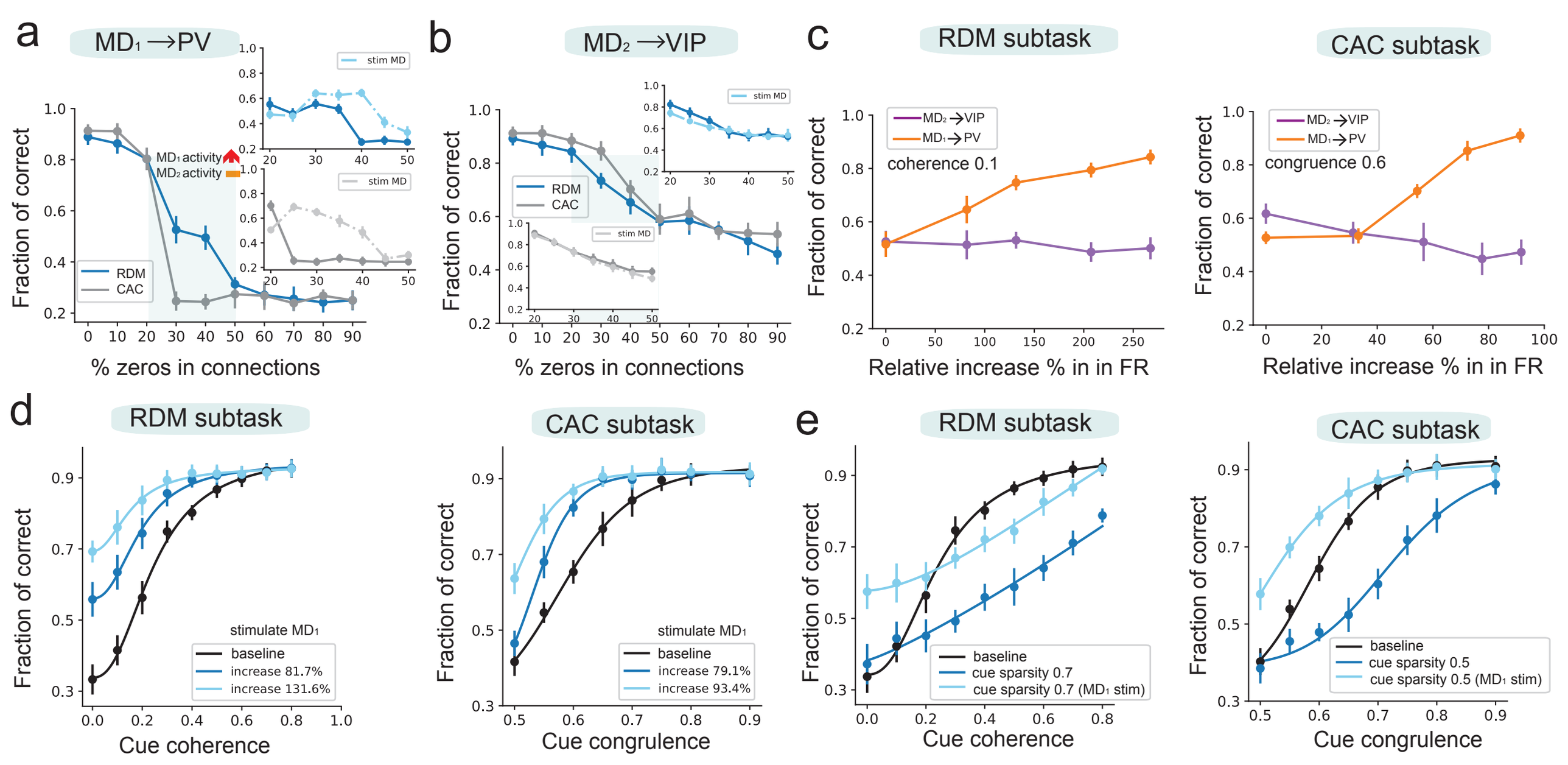
Figure 8. Modified PFC-MD model with additional sparse corticothalamic connectivity. (**a**) The modified PFC-MD model assumed that unbalanced MD subpopulations (as shown by two different circle sizes) and Exc→MD_2 connectivity was 90% sparser than Exc→MD_1 connectivity (as shown by a thinner connection). (**b**) The psychometric curves of performance-optimized modified PFC-MD model in two subtasks. Error bar was shown in mean ± s.e.m. (n=10 realizations). (**c**) Population statistics of mean firing rates (during the cue period) of MD_1 and MD_2 units for encoding cue uncertainty and cue sparsity. (**d,e**) MD_1 stimulation, but not MD_2 stimulation, improved task performance under cue uncertainty. MD_1 or MD_2 stimulation improved task performance under cue sparsity. Error bar denotes SD (n=10). Cue uncertainty comparison during the RDM (CAC) subtask: baseline vs. MD_2 stimulation, P=0.82 (0.16); baseline vs. MD_1 stimulation, P=0.00016 (0.00016); MD_2 vs. MD_1 stimulation, P=0.00016 (0.00043). Cue sparsity comparison during the RDM (CAC) subtask: baseline vs. MD_2 stimulation, P=0.00016 (0.00016); baseline vs. MD_1 stimulation, P=0.00016 (0.00016); MD_2 vs. MD_1 stimulation, P=0.045 (0.017), two-tailed Wilcoxon rank-sum tests. (**f**) When both cue uncertainty and cue sparsity were present, MD_1 or MD_2 stimulation alone couldn't improve task performance, but simulation of both subpopulations could. Baseline vs. stimulation both, *, P=0.011; ***, P= 0.00015, two-tailed Wilcoxon rank-sum tests. Other paired comparisons were n.s.

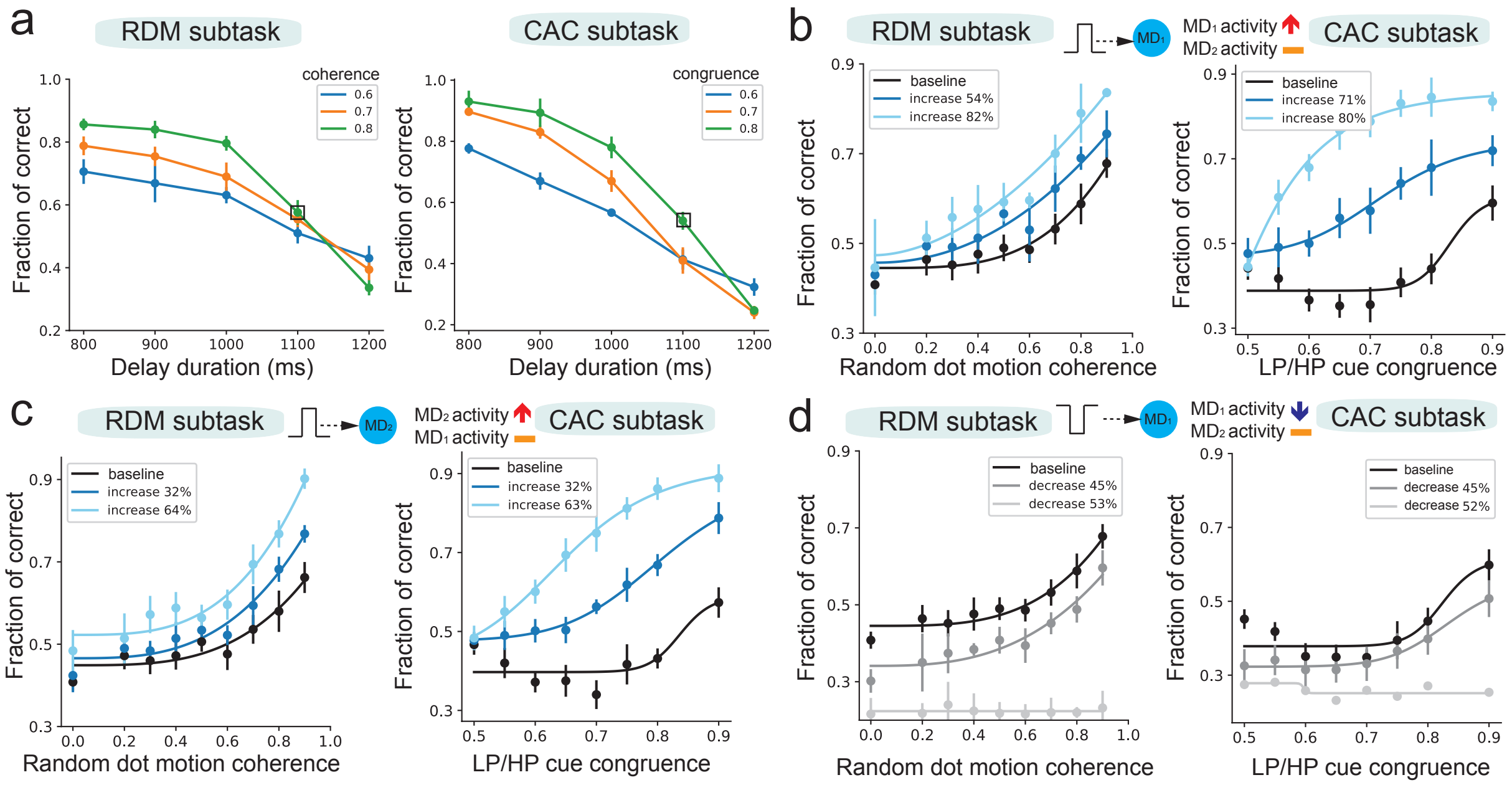

a
E/I-RNN
Inh unit
Exc unit
Input
Cueing context input
Sensory input
Output
Auditory left
Auditory right
Visual left
Visual right
b
MD₂
MD₁
MD
PFC
VIP
SOM
PV
Exc
c
Cross-modal, context-dependent decision-making task
cue-to-rule transformation
rule maintenance
choice selection
conflict evaluation evidence accumu.
working memory
divided attention decision-making
Initation (40 ms)
Cueing period (800 ms)
Delay period (800 ms)
Target (vis/aud) (200 ms)
d
Cueing context 1 (CAC subtask)
#LP>#HP
#LP<#HP
Attend-to -audition
Attend-to -vision
Cueing context 2 (RDM subtask)
moving rightward
moving leftward
Attend-to -audition
Attend-to -vision
e
CAC subtask
RDM subtask
PFC-MD
PFC
Fraction of correct
LP/HP cue congruence
Random dot motion coherence
f
high congruence
med. congruence
low congruence
Fract. correct
CAC subtask
correct
rule error
sensory error
both
high coherence
med. coherence
low coherence
RDM subtask
g
CAC subtask
Low-pass (LP): +1
High-pass (HP): -1
LP/HP ratio: 4/2=2:1
cue congruence: 0.67
low signal density: 6
LP/HP ratio: 5/1
cue congruence: 0.83
low signal density: 6
LP/HP ratio: 8/4=2:1
cue congruence: 0.67
higher signal density: 12
Cueing period (fixed duration)
h
cue sparsity
0.0
0.5
0.7
PFC-MD model
Fraction of correct
Cue congruence
i
PFC-MD
PFC
cue sparsity 0.2
Fraction of correct
Cue congruence
j
RDM subtask
cue coherence: +0.6
low signal density: 10
cue coherence: +0.8
low signal density: 10
cue coherence: +0.6
higher signal density: 15
Cueing period (fixed duration)
k
cue sparsity
0.0
0.5
0.7
PFC-MD model
Fraction of correct
Cue coherence
l
PFC-MD
PFC
cue sparsity 0.2
Fraction of correct
Cue coherence

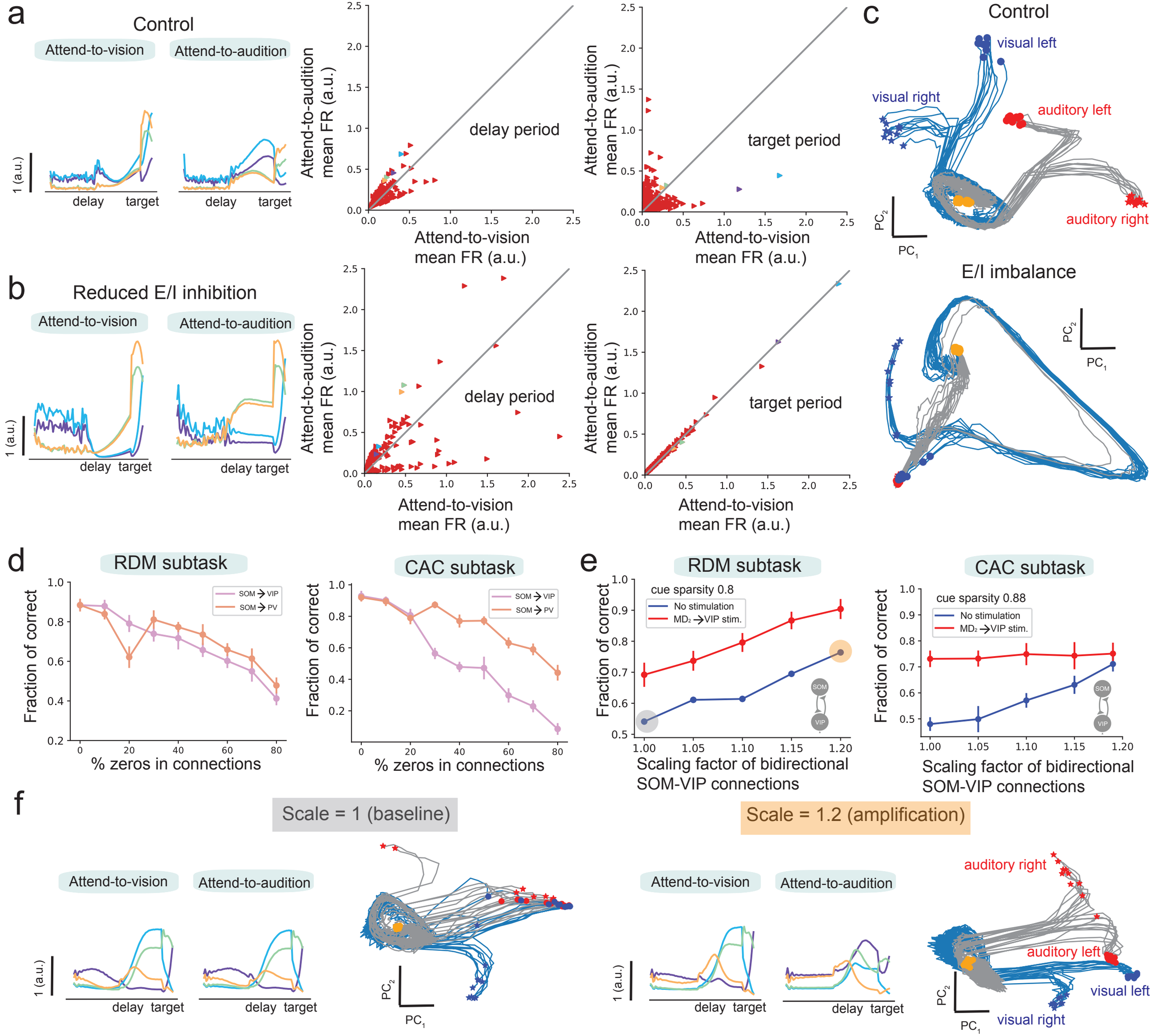

a
PFC-Exc units (context-invariant)
CAC subtask
RDM subtask
Attend-to-vision
Attend-to-audition
#1
#2
FR (a.u.)
cueing
delay
b
PFC-Inh units (context selective)
Attend-to-vision
Attend-to-audition
RDM
CAC
#3
#4
c
MD units (context selective)
#5
#6
d
RULE Coding
CAC subtask
RDM subtask
PFC-Exc
PFC-Inh
MD1
MD2
Delay mean FR (a.u.) Attend-to-audition
Delay mean FR (a.u.) Attend-to-vision
e
CONTEXT Coding
Attend-to-audition
Attend-to-vision
Delay mean FR (a.u.) RDM subtask
Delay mean FR (a.u.) CAC subtask
f
Cue Uncertainty Modulation (MD₁ units)
#7
#8
0.55
0.7
0.9
0.1
0.4
0.9
g
Cue Uncertainty Non-Modulation (MD₂ units)
#9
#10
h
Cue Sparsity Modulation (MD₁ units)
0.0
0.5
0.9
i
Cue Uncertainty Modulation (PFC-Inh units)
#11
#12

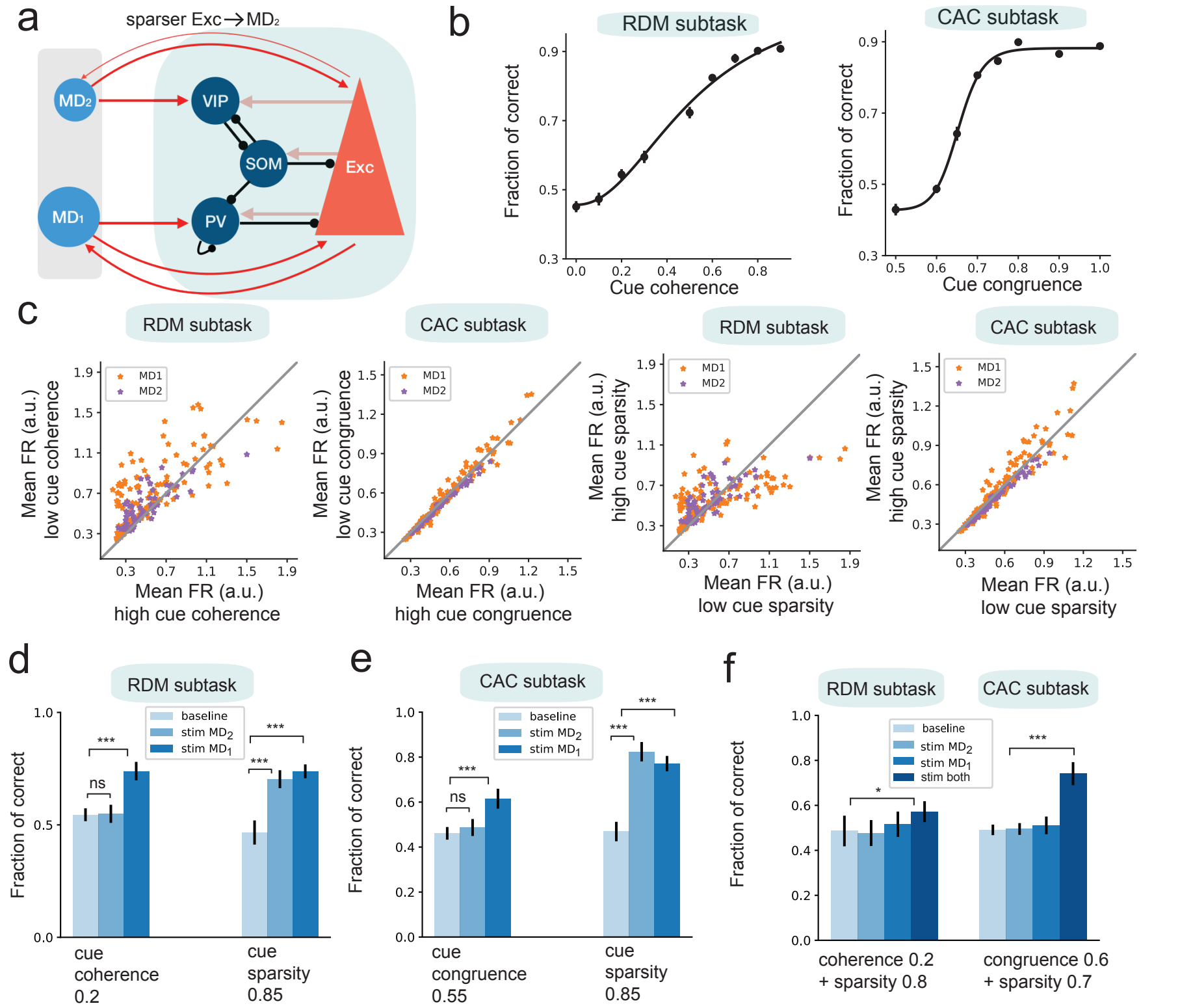

a
RDM subtask
low uncertainty
medium uncertainty
high uncertainty
CAC subtask
Prefrontal SNR
Cueing
Delay
***
**
**
p=0.02
p=0.78
p=0.03
**
p=0.06
p=0.08
**
p=0.02
p=0.9
b
RDM subtask
CAC subtask
cue period
delay period
SNR (MD lesioned)
SNR (MD intact)
c
RDM subtask
CAC subtask
Canonical corr.
*****
d
exc
inh
MD
Decoding accuracy
Rule
Context
e
Rule
Context
low uncert.
med. uncert.
high uncert.
PFC-Exh
PFC-Inh
MD


a
CAC subtask
Auditory Left
Visual Left
CHOICE
Auditory Right
Visual Right
Attend-to-audition
RULE
Attend-to-vision
b
low cue uncertainty
medium cue uncertainty
high cue uncertainty
RULE
c
low sparsity 0.2
medium sparsity 0.5
high sparsity 0.7
RULE
d
RDM subtask
Auditory Left
Visual Left
CHOICE
Auditory Right
Visual Right
RULE
e
RULE
f
RULE
g
CAC subtask
low cue uncertainty
medium cue uncertainty
high cue uncertainty
Velocity (a.u.)
Time
80 ms
h
RDM subtask
low cue uncertainty
medium cue uncertainty
high cue uncertainty
Velocity (a.u.)
Time
80 ms
i
PC3
PC1
PC2
low
high


a
MD₁ →PV
Fraction of correct
% zeros in connections
MD₁ activity
MD₂ activity
RDM
CAC
stim MD
b
MD₂ →VIP
Fraction of correct
% zeros in connections
RDM
CAC
stim MD
c
RDM subtask
CAC subtask
MD₂ →VIP
MD₁ →PV
coherence 0.1
congruence 0.6
Fraction of correct
Relative increase % in in FR
d
RDM subtask
CAC subtask
Fraction of correct
Cue coherence
Cue congrulence
stimulate MD₁
baseline
increase 81.7%
increase 131.6%
increase 79.1%
increase 93.4%
e
RDM subtask
CAC subtask
Fraction of correct
Cue coherence
Cue congrulence
baseline
cue sparsity 0.7
cue sparsity 0.7 (MD₁ stim)
cue sparsity 0.5
cue sparsity 0.5 (MD₁ stim)


a
RDM subtask
CAC subtask
coherence
0.6
0.7
0.8
congruence
Fraction of correct
Delay duration (ms)
b
RDM subtask
MD1
MD1 activity
MD2 activity
CAC subtask
baseline
increase 54%
increase 82%
increase 71%
increase 80%
Random dot motion coherence
LP/HP cue congruence
c
RDM subtask
MD2
MD2 activity
MD1 activity
CAC subtask
baseline
increase 32%
increase 64%
increase 63%
d
RDM subtask
MD1
MD1 activity
MD2 activity
CAC subtask
baseline
decrease 45%
decrease 53%
decrease 52%


a
Control
Attend-to-vision
Attend-to-audition
1 (a.u.)
delay
target
Attend-to-audition mean FR (a.u.)
Attend-to-vision mean FR (a.u.)
delay period
target period
c
visual left
visual right
auditory left
auditory right
PC2
PC1
b
Reduced E/I inhibition
E/I imbalance
d
RDM subtask
CAC subtask
Fraction of correct
% zeros in connections
SOM → VIP
SOM → PV
e
cue sparsity 0.8
cue sparsity 0.88
No stimulation
MD2 → VIP stim.
Scaling factor of bidirectional SOM-VIP connections
SOM
VIP
f
Scale = 1 (baseline)
Scale = 1.2 (amplification)


a
sparser Exc→MD2
MD2
MD1
VIP
SOM
PV
Exc
b
RDM subtask
CAC subtask
Fraction of correct
Cue coherence
Cue congruence
c
RDM subtask
CAC subtask
RDM subtask
CAC subtask
MD1
MD2
Mean FR (a.u.) low cue coherence
Mean FR (a.u.) high cue coherence
Mean FR (a.u.) low cue congruence
Mean FR (a.u.) high cue congruence
Mean FR (a.u.) high cue sparsity
Mean FR (a.u.) low cue sparsity
d
RDM subtask
baseline
stim MD2
stim MD1
Fraction of correct
ns
***
cue coherence 0.2
cue sparsity 0.85
e
CAC subtask
baseline
stim MD2
stim MD1
Fraction of correct
ns
***
cue congruence 0.55
cue sparsity 0.85
f
RDM subtask
CAC subtask
baseline
stim MD2
stim MD1
stim both
Fraction of correct
*
***
coherence 0.2 + sparsity 0.8
congruence 0.6 + sparsity 0.7