nature communications

Article

https://doi.org/10.1038/s41467-025-58011-1

Mediodorsal thalamus regulates task uncertainty to enable cognitive flexibility

Received: 23 October 2023

Accepted: 7 March 2025

Published online: 18 March 2025

Check for updates

Xiaohan Zhang[1], Arghya Mukherjee[2], Michael M. Halassa[2] & Zhe Sage Chen[1,3,4,5]

The mediodorsal (MD) thalamus is a critical partner for the prefrontal cortex (PFC) in cognitive control. Accumulating evidence has shown that the MD regulates task uncertainty in decision making and enhance cognitive flexibility. However, the computational mechanism of this cognitive process remains unclear. Here we trained biologically-constrained computational models to delineate the mechanistic role of MD in context-dependent decision making. We show that the addition of a feedforward MD structure to the recurrent PFC increases robustness to low cueing signal-to-noise ratio, enhances working memory, and enables rapid context switching. Incorporating genetically identified thalamocortical connectivity and interneuron cell types into the model replicates key neurophysiological findings in task-performing animals. Our model reveals computational mechanisms and geometric interpretations of MD in regulating cue uncertainty and context switching to enable cognitive flexibility. Our model makes experimentally testable predictions linking cognitive deficits with disrupted thalamocortical connectivity, prefrontal excitation-inhibition imbalance and dysfunctional inhibitory cell types.

Cognitive flexibility is fundamental in decision making, referring to the ability to adjust the behavioral strategy in response to changing contexts or rules[1]. Successful execution of complex decision-making tasks requires identification and processing of multiple sources of uncertainty[2]. Dealing with uncertainty, such as resolving conflicts and switching tasks is an important component of cognitive control[3]. Task uncertainty may appear in the form of corrupted or incongruent sensory cues ("cue uncertainty")[4,5], and their mapping onto internal or behavioral variables ("mapping uncertainty")[2,6,7]. Cognitive flexibility to map the same rule under different contexts or map different rules with the same cues is critical in decision making. Executive functions central to this cognitive process may involve working memory, attention, context, and error monitoring[8]. Experiments across multiple species have shown that the mediodorsal (MD) thalamus is an important partner for the prefrontal cortex (PFC) in resolving task uncertainty and enhancing cognitive flexibility[9–16]. Human neuroimaging studies have shown that MD activity tracks cue uncertainty in a multi-attribute attention task[15] and a categorization task[17]. This process generalizes to non-human animals; in mice performing a context-dependent decision-making task, the MD tracks cue uncertainty and enables a rapid switching of cue-to-rule transformation[6,18]. Optical manipulations support this notion and delineate the causal roles of MD thalamus in cognitive control[5,19]. However, the computational mechanism by which the MD enhances prefrontal activity to enable cognitive flexibility in context-dependent decision making remain unclear. In addition, how the newly discovered cellular diversity in MD thalamus contributes to such computations is unexplored. Biologically-inspired neural network modeling that incorporates neuronal subtype and circuit pathway knowledge[20–22] may provide a reverse-engineering approach to probe these questions and yield mechanistic insight[23–31].

[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]Department of Neuroscience and Physiology, New York University Grossman School of Medicine, New York, NY, USA. [4]Neuroscience Institute, New York University Grossman School of Medicine, New York, NY, USA. [5]Department of Biomedical Engineering, New York University Tandon School of Engineering, Brooklyn, NY, USA. e-mail: zhe.chen@nyulangone.org

A central tenet of the neuron doctrine is that structure/connectivity determines function. Here, we ask a fundamental question: how does a feedforward MD-like structure facilitate neural computation in a recurrent PFC network to enhance cognitive flexibility or improve cognitive control under task uncertainty? More specifically, how does thalamocortical connectivity contribute to working memory maintenance and improve cognitive control under a low signal-to-noise ratio (SNR) in comparison to a PFC-alone network? What are the roles of interneuron cell type-specific thalamocortical projections in regulating prefrontal computation and generating emergent task-specific neural representations?

In this work, to investigate these questions, we trained biologically-constrained PFC-MD models to perform two versions of context-dependent decision-making tasks, one in a cross-modal setting with cue uncertainty and the other in a cue-to-rule switching setting with both cue and mapping uncertainties. Our models incorporate the anatomic and functional connectivity knowledge derived from mouse prelimbic (PL-PFC) and MD circuits, with genetically identified thalamocortical pathways and prefrontal interneuron cell type-specific connectivity[6]. Our PFC-MD model replicated key behavioral data and neuronal tunings in task-performing mice[6,7,18], including prefrontal context-invariant rule-tuning, thalamic context tuning, the impact of MD activation or suppression on task accuracy, and context-switching. In addition to the emerging findings from the trained model, our model further made experimental predictions on various testing conditions of task difficulty as well as on behavioral deficits induced by abnormality in thalamocortical circuits. The MD subpopulations and cell-type specific MD-to-PFC projections had distinct functional roles in regulating SNR driven by either cue uncertainty ("noise") or cue sparsity ("signal,") maintaining working memory, and mediating prefrontal computation in a task-phase specific manner. In the presence of cue-to-rule mapping uncertainty, we showed that synaptic plasticity of thalamocortical connections enables rapid context switching to perform rule-invariant prefrontal computation. Our analysis suggests that the feedforward MD regulates recurrent prefrontal computation to improve information integration ("working memory maintenance") and cognitive flexibility ("cue-to-rule remapping"). We also provide geometric insight into MD thalamic control in regulating task uncertainty and context switching based on neural subspace analysis. Furthermore, the PFC-MD models enabled us to parse cognitive deficits of thalamocortical circuits in various scenarios of low SNR or behavioral inflexibility, which may be induced by prefrontal excitation-inhibition (E/I) imbalance, dysfunctional inhibitory cell types, and disrupted thalamocortical and corticothalamic connectivity. Finally, our modeling approach, in light of computational psychiatry, provides a framework to link circuit mechanisms to algorithmic processes to understand cognitive deficits underlying psychiatric disorders, such as schizophrenia.

Results

Training PFC-MD models for context-dependent decision-making

Recurrent neural networks (RNNs) have become an important reverse-engineering tool for dissecting circuit functions[24–30]. In the literature, several thalamocortical circuit models have been proposed based on RNNs. Incorporating biological details (such as the thalamus size) and thalamocortical connectivity constraints has been proven important for learning. We adopted a "task-training-followed-by-generalization-testing" strategy to delineate the computational mechanism of PFC-MD circuits in flexible, context-dependent decision making. We first modeled the PFC with a rate-based excitatory-inhibitory (E/I) RNN (Methods), where cortical GABAergic inhibitory neurons consisted of parvalbumin (PV), vasoactive intestinal peptide (VIP)-expressing, and somatostatin (SOM) interneuron cell types (Fig. 1a). We assumed structured inhibitory-to-excitatory and inhibitory-to-inhibitory connectivity and disallowed weak cell-type connections that are negligible. Specifically, the VIP interneurons mainly disinhibit other classes of interneurons[32,33]. We imposed multiple key biological constraints on network connectivity. First, we imposed Dale's principle onto the RNN, specially for recurrent PFC connectivity. Second, we modeled the MD as a feedforward structure devoid of lateral connectivity, where all MD neurons received excitatory corticothalamic inputs from the PFC and projected back to excitatory neurons. Third, guided by recent knowledge of genetically identified thalamocortical projections that differentially target distinct PFC interneuron types[6], we imposed additional connectivity constraints onto 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→PV corresponds to the kainite receptor (GRIK4)-expressing (i.e., MD_{GRIK4}) pathway with preferentially targeted PV+ neurons; MD_{DRD2}→VIP projection corresponds to the dopamine receptor (D2)-expressing (i.e., MD_{DRD2}) pathway with preferentially targeted VIP^+ neurons. We trained both PFC-MD models and "PFC-alone" models (as a control, with an identical setup in stimulus input, the total number of units, and cortical E/I configuration) to perform a cueing context-dependent four-alternative forced choice (4AFC) task (Fig. 1b and Methods). Briefly, the task is to attend to either a visual or an auditory target that are simultaneously presented, and the targets are cued by cross-modal cues that signal the "attend-to-vision" or "attend-to-audition" rule. The presentation of the cue and the targets are temporally separated by a stimulus-free delay period. When the cueing signal is uncertainty free, the task becomes easier to learn, but when the uncertainty level of the cueing signal increases, the task difficulty will increase.

In the first version of cueing context-dependent decision-making task, task uncertainty contained cue uncertainty only, in which uncertainty is conceptually linked to the SNR in the cueing signal. At the noise level, uncertainty is characterized by either cue coherence or congruence in the presence of competitive cues. A higher cue uncertainty implies a lower SNR. At the signal level, uncertainty is characterized by the cue density, or alternatively the cue sparsity (a higher density or lower sparsity level implies a greater SNR). The cue uncertainty and sparsity levels are controlled independently and jointly influence the SNR. The two contexts corresponded to two cue-independent subtasks that mapped different modalities of sensory cues to one set of rules: attend-to-audition vs. attend-to-vision (Fig. 1c). In our setting, we parameterized the cue uncertainty in the two contexts: a random dot motion (RDM) context that mimics visual modality, and a conflicting auditory cue (CAC) context that mimics auditory modality (Methods). The cue-to-rule transformation was determined by the motion direction of random dots in the RDM context, and was determined by the tone majority in the CAC context. These two contexts 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 followed by 200-ms choice period). Therefore, the task-optimized model first needs to accumulate sensory evidence from conflicting cues to identify the rule, then preserve the rule information in working memory, and ultimately attend to audition/vision to choose the correct target.

In the training phase, we trained PFC-MD models to learn two contexts jointly where the trials were interleaved, and cue uncertainty was minimal. Upon reaching the desired performance accuracy, then in the testing phase, we tested the PFC-MD model performance by varying the degree of cue uncertainty. The level of cue uncertainty is controlled by the coherence of RDM in the RDM context or the congruence (i.e., degree of agreement) of low-pass (LP) vs. high-pass (HP) tones in the CAC context.

https://doi.org/10.1038/s41467-025-58011-1

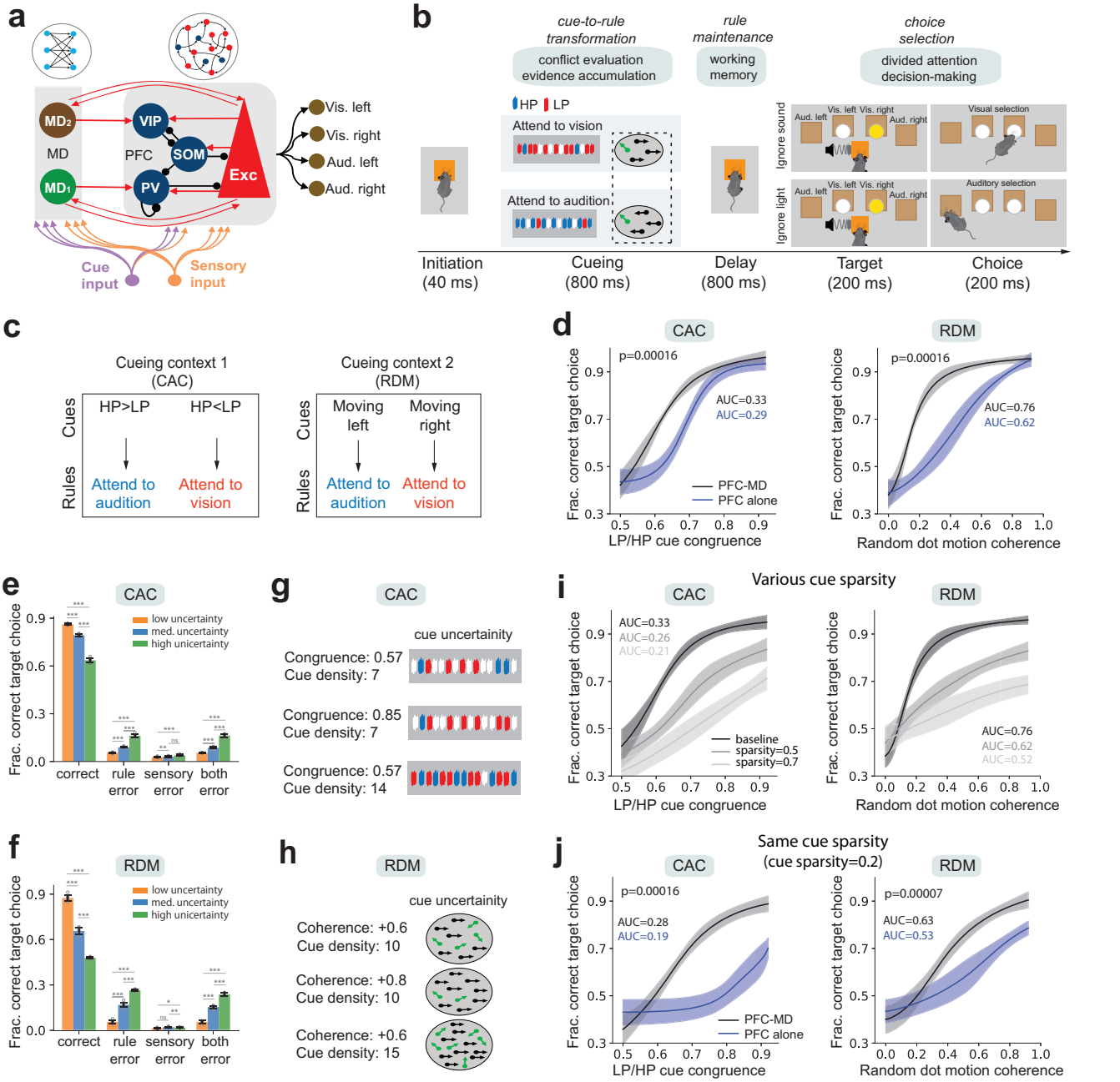


Fig. 1 | The task-optimized PFC-MD model in decision making with parameterized cue uncertainty. a Excitation-inhibition (E/I) recurrent neural network (RNN) for modeling the PFC-alone network, where three major prefrontal interneuron cell types were specified. The PFC-MD model consisted of two target-specific non-recurrent excitatory MD subpopulations and bidirectional MD-PFC projections. **b** Schematic of a cross-modal, cueing context-dependent decision-making task with working memory and divided attention components. Sensory cues were either conflicting LP/HP pulses or moving random dots. **c** Schematic of cueing context-dependent rule encoding. The RDM and CAC learned the same rule from different cueing modalities: attend-to-audition vs. attend-to-vision. **d** Psychometric curves of RDM and CAC contexts. Shaded areas denote SD, and the lines denote the mean derived from the parameter fit of a logistic function. Each condition was repeated 10 times with different input realizations yet with identical summary statistics. The PFC-MD model (black) outperformed the PFC-alone model (blue) in the presence of intermediate-to-high cue uncertainty, achieving a greater area under the psychometric curve (AUC). $p = 0.00016$, two-tailed rank-sum test for both panels. **e** Percentage of task error types in the CAC context under various cue uncertainty conditions. Error bar denotes SD ($n = 10$). For each correct/error group, all pairwise p-values were computed based on Bonferroni-corrected two-tailed rank-sum test (**$p < 0.01$, ***$p < 0.001$). **f** Similar to (**e**) but for the RDM context. Error bar denotes SD ($n = 10$). **g** Illustration of three types of cues with two congruence levels and two densities in the CAC context. **h** Illustration of three types of cues with two coherence levels and two densities in the RDM context. **i** Psychometric curve of the PFC-MD model under three different cue sparsity levels in the CAC and RDM contexts. The AUC statistics decreased with an increasing sparsity level. Shaded areas denote SD. **j** Psychometric curve comparison between the PFC-MD and PFC-alone models in the CAC and RDM contexts in a low cue sparsity condition. The PFC-MD outperformed the PFC-alone model in AUC (CAC: $p = 0.00016$; RDM: $p = 0.00007$, two-tailed rank-sum test). Shaded areas denote SD.

PFC-MD outperformed PFC-alone model in ambiguous cue integration

We tested the task-optimized model's generalization and produced psychometric curves independently for two contexts (Fig. 1d). The psychometric curve characterizes the fraction of correct target choice under different stimulus conditions. For a fair comparison, we kept the number of total units equal between two models (Methods). By changing the level of coherence or congruence of the cues, we found that the PFC-MD model outperformed the PFC-alone model on ensemble average (Supplementary Fig. 1a), especially in the medium-to-high cue uncertainty (i.e., medium-to-low SNR) regime. Comparisons on the area under the psychometric curve (area under the curve (AUC)) showed statistically significant improvement in the PFC-MD model (Fig. 1d; $p = 0.00016$, one-sided rank-sum test). Furthermore, we examined the percentage of error types in decision making. Given the outcome mismatch among four choices, task errors could be ascribed by either attending the wrong rule ("rule error" during the cueing or delay period) or attending the correct rule but with a wrong choice ("sensory error" during the target or choice period). Notably, the rule error increased with increasing cue uncertainty in the task-optimized PFC-MD model (Fig. 1e, f; $p < 0.001$; Bonferroni-corrected rank-sum test).

The SNR of cueing was determined by not only the ratio of conflicting cues, but also the density of cues (or alternatively, the cue sparsity: higher density is equivalent to lower 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 cueing signals are both conflicting and sparse; an identical 8/4 = 2:1 ratio of LP/HP has a higher density of pulses (Fig. 1g). In the RDM context, the cue density was determined by the number of moving dots in the fixed stimulus space. Even under the same coherence condition, the SNR would be higher when the number of dots was large (Fig. 1h). In both contexts, increasing cue sparsity had a negative impact on the task accuracy (Fig. 1i), but the PFC-MD model consistently outperformed the PFC-alone model under the same cue sparsity and uncertainty (Fig. 1j, $p < 0.005$, rank-sum test; see also Supplementary Fig. 1b, c). Together, these results suggest that the MD plays a role in sensory evidence integration and boosts prefrontal computation in lower SNR during the cueing period.

PFC and MD showed diverse unit tunings and population dynamics

We examined the emergent tuning properties of single units of the trained PFC-MD model with respect to task variables during both cueing and delay periods. A majority (~55–65%) of PFC excitatory units exhibited tuning preference in rule, but not context; some showed cueing context-invariant rule tuning during the delay period (Fig. 2a, d, and Supplementary Fig. 2a for additional examples). Similarly, some (~40%) of PFC inhibitory units encoded rules during the delay period (Fig. 2d). In contrast, ~80% of MD units showed modulation selectivity with respect to the cueing context, but not rule (Fig. 2b, c, e and Supplementary Fig. 2b for additional examples). To simulate an effect of evoked firing activity similar to optogenetic MD activation, we induced a transient input to the specific MD subpopulation to increase the phasic firing activity of MD units (Methods). During the baseline (i.e., complete absence of stimuli within a trial), increasing the MD_2 unit firing tended to amplify the firing rate of task-relevant, rule-tuned PFC units, whereas increasing the MD_1 unit firing tended to suppress the firing rate of task-irrelevant PFC units (especially those with lower firing rates) (Supplementary Fig. 2c).

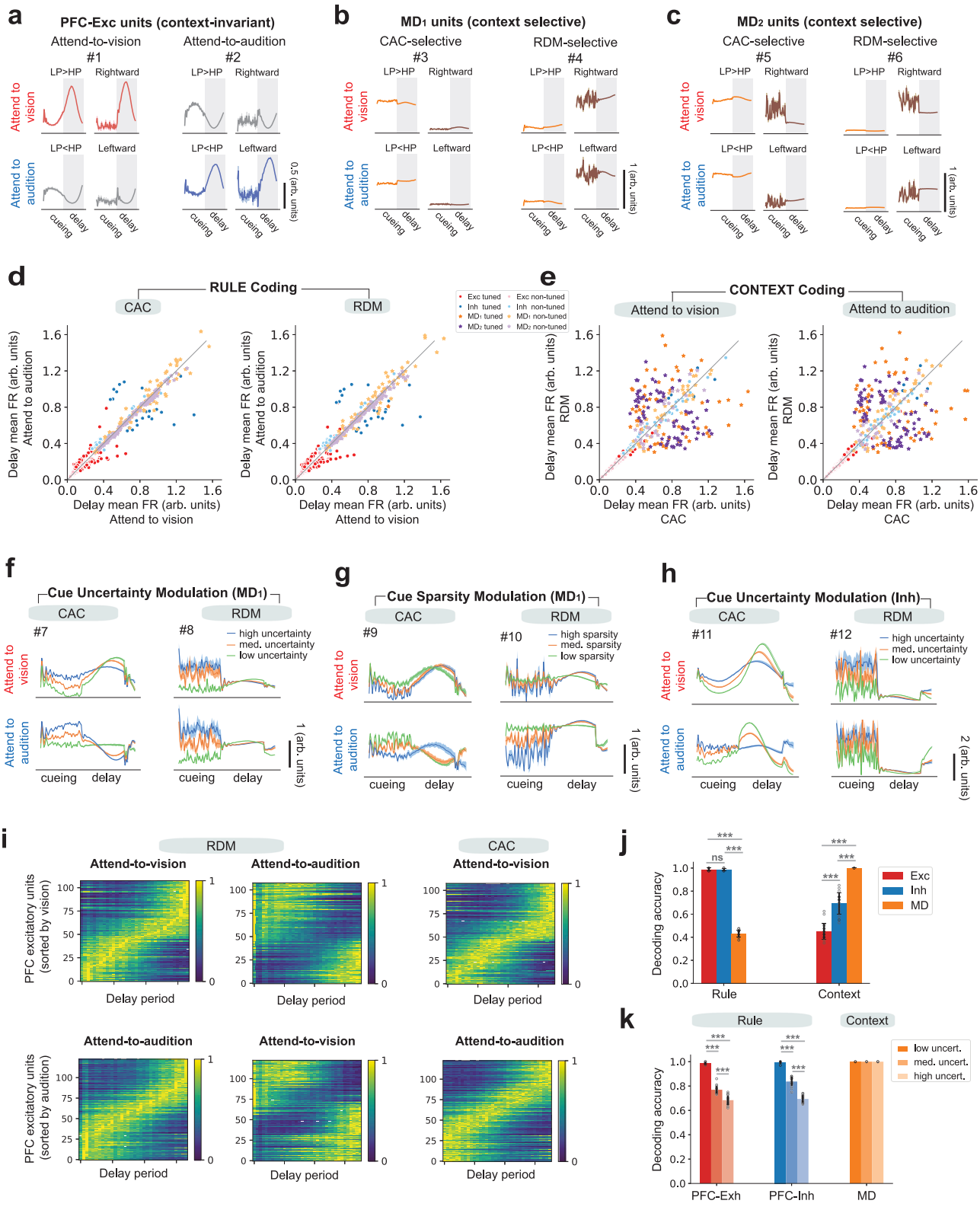
Additionally, during the cueing period, a large percentage of MD_1 units (30–40%) modulated their firing rates with respect to cue uncertainty regardless of the context (Fig. 2f, Supplementary Fig. 2d and 3a, b; "Methods"). We also found firing modulation with respect to cue sparsity for the MD_1 subpopulation (Fig. 2g and Supplementary Fig. 3c, d), but not for MD_2 units (Supplementary Fig. 2d, e). Among MD_1 units, tuning properties with respect to the context and cue uncertainty were relatively orthogonal since very few MD_1 units showed conjunctive coding (Supplementary Fig. 2f). Similar cue uncertainty tunings were also found in PFC inhibitory units (Fig. 2h). It is noteworthy that not only these emergent PFC and MD tuning properties were in line with neurophysiological findings in task-performing mice[6,7], intensive computer simulations also enabled us to investigate the dependency of these tunings on cell type-specific thalamocortical connectivity (Supplementary Table 1), providing insight into "why" and "how" questions of MD regulation (see "Discussion").

Increasing cue uncertainty or cue sparsity reduced rule selectivity of PFC excitatory units in working memory (Supplementary Fig. 4), suggesting that lowering the SNR of cueing signals may lead to an information loss during cue-to-rule transformation. During the delay period, sorting the peak responses of rule-tuned PFC excitatory units produced a rule-specific sequence[18,25,27,29]; moreover, the neural sequence was similar across RDM and CAC contexts (Fig. 2i). Therefore, a large overlapping population of PFC units engaged in context-invariant rule encoding via temporal codes. To further explore the tuning properties of PFC-MD populations, we employed population decoding analyses to read out either rule or context information on a trial-by-trial basis based on the mean firing of the population (Methods). The rule information could be reliably decoded from the PFC, but not from the MD, whereas the context information could be inferred perfectly from the MD, but less reliably from the PFC (Fig. 2j). Increasing cue uncertainty led to the PFC's decreased rule-decoding accuracy but had no effect on the MD's context-decoding accuracy (Fig. 2k; $p < 0.001$, Bonferroni-corrected rank-sum test). While recurrent PFC units showed robust rule tunings, systematic investigations of the network architecture and cell-type connectivity could reveal the impact of thalamocortical loops and network connectivity on MD unit tunings (Supplementary Table 1).

We further examined the PFC population representations in both PFC-MD and PFC-alone networks. Dimensionality reduction of the PFC population activity revealed low-dimensional neural trajectories in the task-specific subspace (Fig. 3a, b). Increasing the level of cue uncertainty or sparsity led to changes of neural trajectories (Fig. 3c–f). While the trajectories remained separated at the end point, the neural trajectory converged to fixed points faster under the lowest cue uncertainty, and the initial "neural velocity" (Methods) was greater in the presence of lower cue uncertainty (Fig. 3g, h). All neural trajectories reached to a steady state where the velocity was close to zero. A fixed-point analysis on the three-dimensional PC subspace (Methods) revealed two "line attractor-like" basins, with each one matching a rule and aligning cue uncertainty from one end to the other end (Fig. 3i). Note that these attractor-like basins have a finite length, and the neural trajectory could diverge from the attractor basins when the cue uncertainty level was too high. To link such neural representations with the network behavior, we conducted similar analyses on the population activity in PFC-alone networks, showing that well-separated neural trajectories were associated with good psychometric curves (see Fig. 3j for an illustration) and associated with faster convergence speed to fixed points.

To compare the PFC-MD and PFC-alone networks, we further computed the intrinsic time constant (Methods) derived from the maximum eigenvalue of their respective recurrent weight matrices. We found that the time constant was significantly greater in the PFC-MD than in the PFC-alone network ($p = 5 \times 10^{-48}$, signed-rank test, $n = 20$) (Fig. 3k). From a dynamic system perspective, the MD may help "slow down" prefrontal dynamics by increasing the time constant to improve cue or rule information integration during cueing or working memory. Additionally, the non-normality degree of the network connectivity (Methods) was significantly greater in the PFC-MD than PFC-alone network ($p = 1 \times 10^{-18}$, signed-rank test, $n = 20$) (Fig. 3l), suggesting that the eigenspectrum in the PFC-MD structure contains richer

https://doi.org/10.1038/s41467-025-58011-1



eigenmodes that allow stronger transient amplification in the complex plane; this computational mechanism may help improve the SNR in the presence of cue uncertainty and/or cue sparsity. Finally, we sought to compute the principal angle (Methods) between two PCA subspaces spanned by PFC-only units and MD-only units during the cueing period. We found that in both CAC and RDM contexts, the principal angle was the smallest (or largest) with the lowest (or highest) cue uncertainty (Fig. 3m). This result suggests that in a task-optimized PFC-MD network, the MD activity was more aligned with the PFC activity (e.g., ~10 deg) so that the MD could better regulate prefrontal computation; however, when the activity of two areas became more orthogonal (e.g., 70–90 deg), the MD's partnership role was compromised, and the final performance degraded substantially. Notably, the principal angle distributions of MD-PFC subspaces shared a similar trend in the CAC and RDM contexts despite the differences in cue complexity, density, and dimensionality.

Fig. 2 | Emergent neural representations of PFC and MD units from the task-optimized PFC-MD network. a Two representative PFC excitatory units encoding two rules under two contexts. These two units were cue invariant units that encoded the same rule. Shaded areas around the PSTH denote the SD. **b** Two representative MD_1 units encoding the cueing context. Shaded area around the PSTH denote the SD. **c** Two representative MD_2 units encoding the cueing context. Shaded areas around the PSTH denote the SD. **d** Population statistics of mean firing rates (during the task delay period) of excitatory PFC units and MD units for encoding rule. Markers in dark/light color shade represent tuned/non-tuned units, respectively. **e** Similar to panel **d**, except 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. Shaded areas around the PSTH denote the SD. **g** The same MD_1 units also showed firing rate modulation with respect to cue sparsity. Shaded areas around the PSTH denote the SD. **h** Two representative PFC inhibitory units that showed firing rate modulation with respect to cue uncertainty. Shaded areas around the PSTH denote the SD. **i** Prefrontal neural sequences showed rule specificity and context invariance. Each heatmap shows the normalized peri-stimulus time histogram (PSTH) of selected prefrontal excitatory units of the task-optimized PFC-MD model during the delay period. In the first row, all units of all panels were sorted in the same order according to attend-to-vision tuning. In the second row, all units of all panels were sorted in the same order according to attend-to-audition tuning. The first and second columns demonstrated rule specificity, whereas the first and third columns demonstrated context invariance. **j** Population decoding analysis showed that the PFC population better encoded rule, whereas the MD population better encoded context. Accuracy is presented by mean ± SD ($n = 20$ Monte Carlo runs, 50 independent trials per run). $^{***}p < 0.001$, Bonferroni-corrected rank-sum test. **k** Increased in cue uncertainty caused decreased rule decoding accuracy (mean ± SD) for the PFC but did not affect context decoding accuracy for the MD. Error bar denotes SD ($n = 20$ Monte Carlo runs, 50 independent trials per run). In rule decoding, statistical tests were conducted independently for each group; all paired comparisons were statistically significant. $^{***}p < 0.001$, Bonferroni-corrected rank-sum test.

MD enhances working memory maintenance

Next, we focused our investigation on the properties of task-optimized PFC-MD network during the delay period. The neural basis for persistent activity in working memory is thought to involve recurrent excitation among prefrontal excitatory neurons. During working memory, the rule information is maintained in recurrent PFC connectivity, but the information may be subject to a loss with increased task delay duration or lower SNR. Therefore, we tested the robustness of the PFC-MD model subject to a potential information loss. In so doing, we gradually increased the duration of delay period from the initial 800–1200 ms and found that increasing delay duration had a negative impact on task accuracy (Fig. 4a). To demonstrate the role of MD in working memory maintenance[18], we selected the condition of coherence/congruence of 0.8 in respective RDM/CAC contexts (see the box symbol in Fig. 4a) and increased the firing rate of MD_1 or MD_2 subpopulation by a multiplicative gain during an elongated 1100-ms delay period while keeping the other MD subpopulation unchanged. Such MD activations led to a slight boost in rule maintenance and subsequent decision-making accuracy under a wide range of cue uncertainty levels. Notably, it required less firing rate elevation for MD_2 than MD_1 to achieve the same level of task improvement or AUC statistic (Fig. 4b, c), probably because the MD_2→VIP pathway has an amplification effect on increasing the target sensitivity for working memory maintenance. In contrast, suppressing the MD_1 activity while keeping the MD_2 activity unchanged during the delay period had a negative effect on task accuracy or AUC statistic (Fig. 4d). Alternatively, we strengthened or weakened prefrontal Exc-to-MD corticothalamic connection strengths and observed qualitatively similar effects on performance (Supplementary Fig. 5a–d). Together, these results support that reciprocal PFC-MD communications help maintain the rule information through two functionally distinct thalamocortical and corticothalamic pathways.

To distinguish the role of two specific thalamocortical projections (MD_1→PV and MD_2→VIP) in regulating prefrontal computation, we disconnected one of two thalamocortical projections or set the respective thalamocortical connections to zeros (i.e., "computational lesion") while keeping the other intact. 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 contexts (Fig. 4e). Within a narrow range (20–50% sparsity, shaded area in the left panel of Fig. 4e), enhancement of phasic MD_1 activity could boost task performance in a wide range of sparsity conditions (Fig. 4e, dashed lines of two smaller panels). In contrast, weakening MD_2→VIP connectivity degraded the task accuracy slowly (Fig. 4f), and enhancement of MD_2 activity had little effect on task accuracy. This was possibly because there was no direct VIP→Exc pathway to prefrontal excitatory neurons in our model assumption.

Furthermore, we investigated the impact of various MD manipulations on the network time constant (Fig. 4g). Compared to the PFC-MD baseline, computational MD-lesion significantly decreased the time constant ($p < 0.0001$, Bonferroni-corrected rank-sum test, based on $n = 20$ independently trained PFC-MD models), whereas strengthening MD_1→PV and MD_2→VIP connections significantly increased the time constant ($p < 0.0001$ in both cases, Bonferroni-corrected rank-sum test), which may explain why the MD_1 or MD_2 activation strategy can rescue the low SNR in the presence of cue uncertainty.

Probing mechanistic causes of PFC-MD circuit in cognitive deficits

Our task-optimized PFC-MD models may serve as a platform to interrogate mechanisms of abnormal neural representations and cognitive deficits in working memory or cognitive inflexibility. This was achieved by modifying the "normal control" with a specific "computational lesion or dysfunction."

Recurrent synaptic excitation among pyramidal neurons is approximately balanced by synaptic inhibition[34]. Thus, we first changed the overall inhibition to excitatory neurons and induced cortical E/I imbalance in the task-optimized PFC-MD model; this was done by increasing the sparsity of prefrontal SOM→Exc connections to reduce inhibition. Alternatively, MD inhibition also alters prefrontal E/I balance[35]. Notably, these operations changed rule tunings of some PFC excitatory units during the delay period and reduced choice discriminability (characterized by the mean firing rate difference between two rules) during the target period (Fig. 5a–d and Supplementary Fig. 6a, b). Specifically, in the reduced inhibition condition, the mean firing rates of rule-tuned PFC excitatory units became nearly identical (Fig. 5d), so that the distributions of $\Delta FR_{v\text{-}a}$ = FR (attend-to-vision) − FR (attend-to-audition) dramatically differed from the control condition (delay period: $p = 4 \times 10^{-7}$; target period: $p = 4.8 \times 10^{-15}$, Kolmogorov–Smirnov test). This can be also explained by the fact that the PFC dynamics became more unstable and sensitive to noise perturbation, which pushed neural trajectories out of the stable attractor space. Population dynamics in the PCA subspace further showed that E/I imbalance distorted the neural trajectories (Fig. 5e and Supplementary Fig. 6a, b).

We further weakened prefrontal SOM→VIP and SOM→PV connectivity independently and observed a more robust task performance in SOM→PV manipulations (Fig. 5f), while a similar manipulation of VIP→SOM connectivity had little impact (Supplementary Fig. 6c). For bidirectional SOM-VIP connections, strengthening mutual inhibition between SOM and VIP units in the PFC-MD model could amplify the sparse cue signal (gray curves in Fig. 5g and replot in 5 h). Furthermore, this was effective only for bidirectional SOM-VIP amplification but not for unidirectional manipulation; Supplementary Fig. 6d), supporting

https://doi.org/10.1038/s41467-025-58011-1

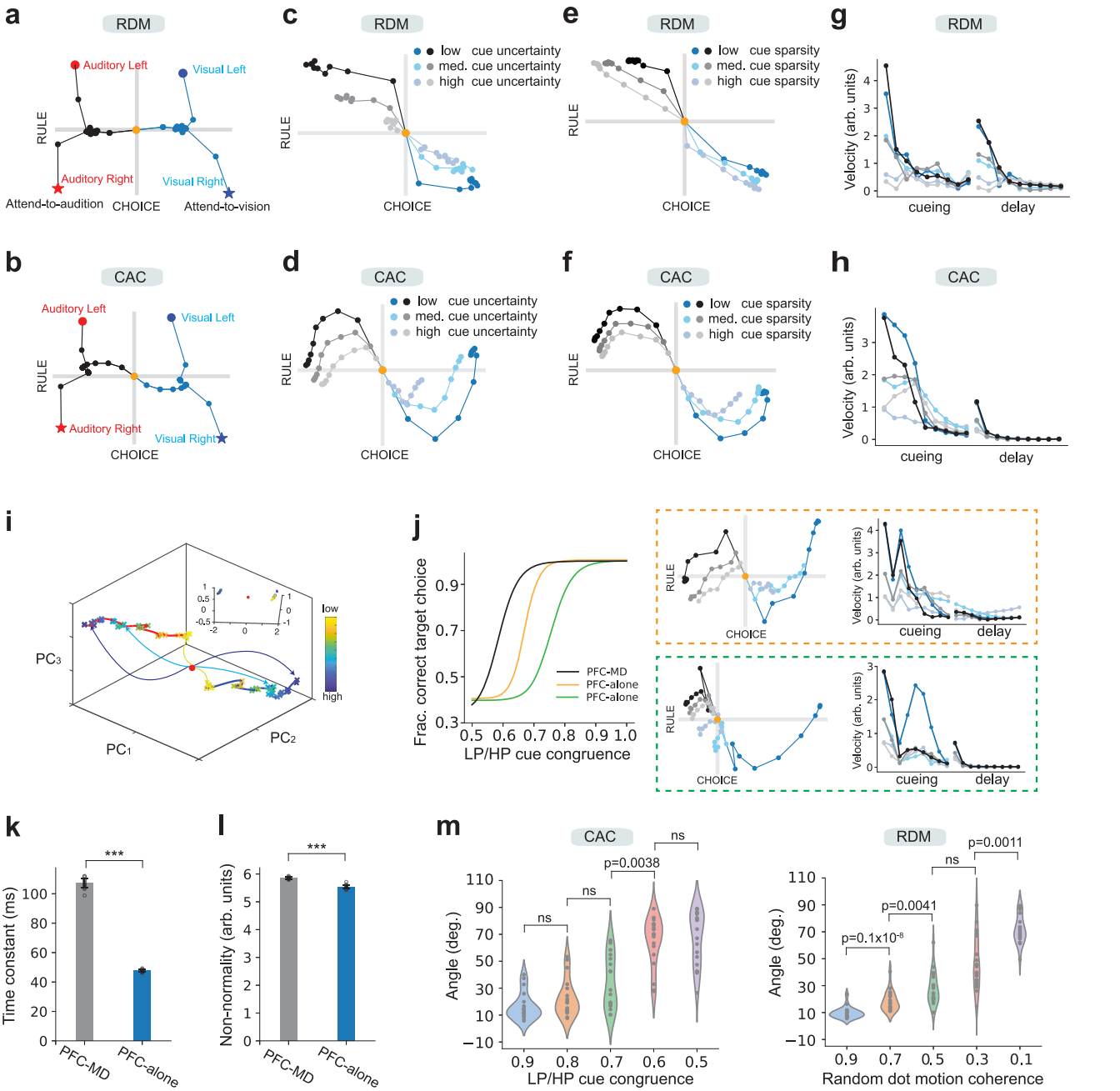


Fig. 3 | Population representations from task-optimized PFC-MD and PFC-alone networks. a Dynamics of population responses in the RDM context. The average population trajectory for a given condition and time was represented as a point in the state space. Responses from correct trials only were shown from 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 (**a**) but for the CAC context. **c**, **d** Similar to (**a**, **b**), except for during the cueing period only. For a better illustration, three levels of cue uncertainty were shown by different shades of gray or blue color (dark/intermediate/light shade: low/medium/high cue uncertainty, respectively). **e**, **f** Similar to (**c**, **d**), except for a fixed cue uncertainty but different levels of cue sparsity. Three levels of sparsity were shown by different shades of gray or blue color (dark/intermediate/light shades: low/medium/high sparsity, respectively). **g**, **h** Comparison of neural velocity during both cueing and delay periods for different levels of cue uncertainty for the RDM and CAC contexts, respectively. The change of neural activity reduced to a low level during the delay period, reaching a fixed-point regime. **i** Fixed-point analysis in the three-dimensional PCA 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. **j** Similar to (**d**, **f**) in the CAC context, but from two PFC-alone networks with different psychometric curves. As expected, a better performance curve corresponded to better-separated neural trajectories and faster convergence speed to fixed points in the high cue uncertainty regime. Group statistics comparison between PFC-MD and PFC-alone networks ($n = 20$ models per group) on intrinsic network time constant (**k**) and degree of non-normality (**l**) estimated from their respective recurrent weight matrices. Center denotes mean, and error bar denotes SD. One-sided signed-rank tests showed their median statistics were significantly greater in the PFC-MD network (***, $p = 5 \times 10^{-48}$ for (**k**), $p = 1 \times 10^{-18}$ for **l**). **m** Quantification of the principal angle between MD and PFC subspaces with respect to various levels of cue uncertainty. There was an increasing trend in principal angle with increasing cue uncertainty ($n = 20$ Monte Carlo runs for each condition). Comparisons between neighboring conditions were highlighted (Bonferroni-corrected two-tailed rank-sum test).

https://doi.org/10.1038/s41467-025-58011-1

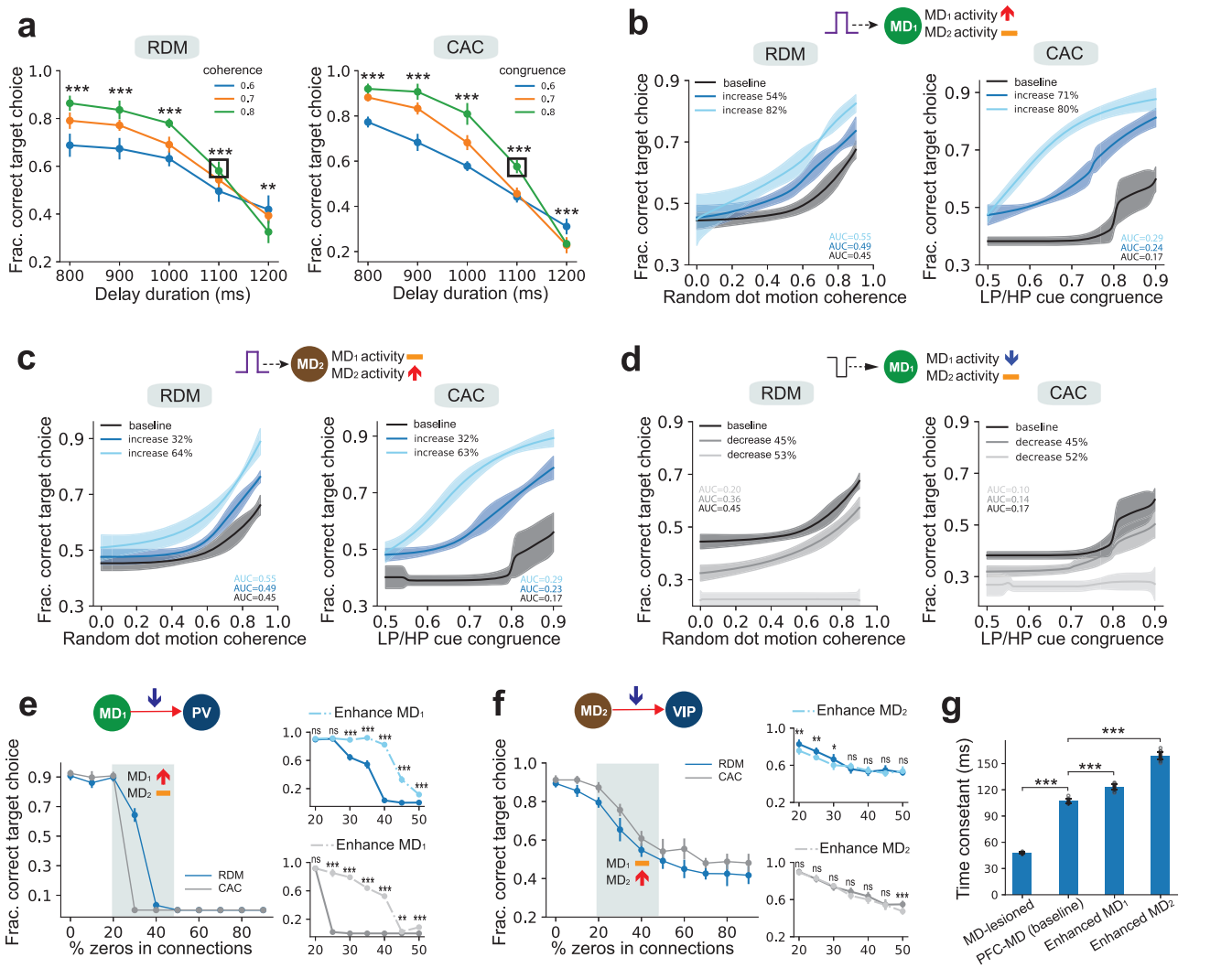


Fig. 4 | MD activation enhances working memory maintenance in the PFC-MD network. a Increasing the duration of task delay period reduced the task performance under cue uncertainty in both contexts. Statistics are presented in mean ± SD ($n = 10$ Monte Carlo runs). Statistical tests were conducted between green and blue curves (pointwise) for each delay duration condition (*$p < 0.05$, **$p < 0.01$, two-tailed rank-sum test). The task condition marked with "□" was used in the illustrations of the 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 curves (mean ± SD) in two contexts. Black curve denotes the baseline, and the number of % denotes the relative increase in subpopulation firing rate. Shaded area denotes SD ($n = 10$). With MD_1 activation, the derived AUC increased from the baseline (RDM: dark blue vs. black, $p = 0.016$; light blue vs. black, $p = 0.009$; CAC: dark blue vs. black, $p = 0.0001$; light blue vs. black, $p = 0.7 \times 10^{-4}$; all by the two-tailed rank-sum test). With MD_2 activation, the derived AUC increased from the baseline (RDM: dark blue vs. black, $p = 0.028$; light blue vs. black, $p = 0.009$; CAC: dark blue vs. black, $p = 0.0001$; light blue vs. black, $p = 0.0002$). **d** Decreasing the MD_1 population firing rate degraded the working memory and the psychometric curves of two contexts. Shaded area denotes SD ($n = 10$). With MD_1 suppression, the derived AUC increased from the baseline (RDM: dark gray vs. black, $p = 0.014$; light gray vs. black, $p = 0.009$; CAC: dark gray vs. black, $p = 0.0003$; light gray vs. black, $p = 0.7 \times 10^{-4}$). **e** *Left:* Weakening MD_1→PV connections (by increasing the percentage of zeros) during the cueing period quickly reduced the task performance in both RDM and CAC contexts. *Right:* increasing MD_1 activity (dotted line) could rescue each context's performance under a wide range of connectivity conditions (shaded area: 20%-50% of zeros). Statistics are presented in mean ± SD ($n = 10$ Monte Carlo runs). Statistical tests were conducted between dark blue and light blue curves (pointwise) for condition (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, two-tailed rank-sum test). **f** Weakening MD_2→VIP connections only degraded the task performance slowly in both RDM and CAC contexts, but increasing MD_2 activity had no or little effect on the change in task performance. Statistics are presented in mean ± SD ($n = 10$ Monte Carlo runs). Statistical tests are similar to (**e**). **g** Impact of network time constant under different thalamocortical connectivity manipulations. Statistics are presented in mean ± SD ($n = 20$ independently trained PFC-MD models). The MD_1 and MD_2 enhancement corresponded to the experiments where respective thalamocortical connections were strengthened during working memory. All statistical tests were two-tailed rank-sum tests against the PFC-MD baseline (MD-lesioned vs. baseline: $p = 5 \times 10^{-45}$; enhanced MD_1 vs. baseline: $p = 3 \times 10^{-17}$; enhanced MD_2 vs. baseline: $p = 1 \times 10^{-36}$; Bonferroni-corrected rank-sum test).

the role of VIP-SOM motif in gain control[20]. A closer examination of single-unit and population responses revealed that bidirectional SOM-VIP amplification produced emergent rule tunings and enhanced rule discriminability (Fig. 5h). Notably, the amplification factor of 1.2 in bidirectional SOM-VIP connectivity produced visible changes in unit tunings and population representation compared to the no-amplification baseline (i.e., scaling factor of 1). Additional MD_2→VIP stimulations further facilitated this SOM-VIP amplification (red curves in Fig. 5g), validating the notion of MD amplifier to regulate prefrontal computation and enhance SNR[6]. In animal experiments, abnormal cortical GABAergic signaling or deficits in interneuron subtype have been implicated in cognitive impairment in autism spectral disorder, such as decreased responses to salient stimuli ("hypo-sensitivity") under a low SNR or decreased inhibitory gain control[36,37].

https://doi.org/10.1038/s41467-025-58011-1

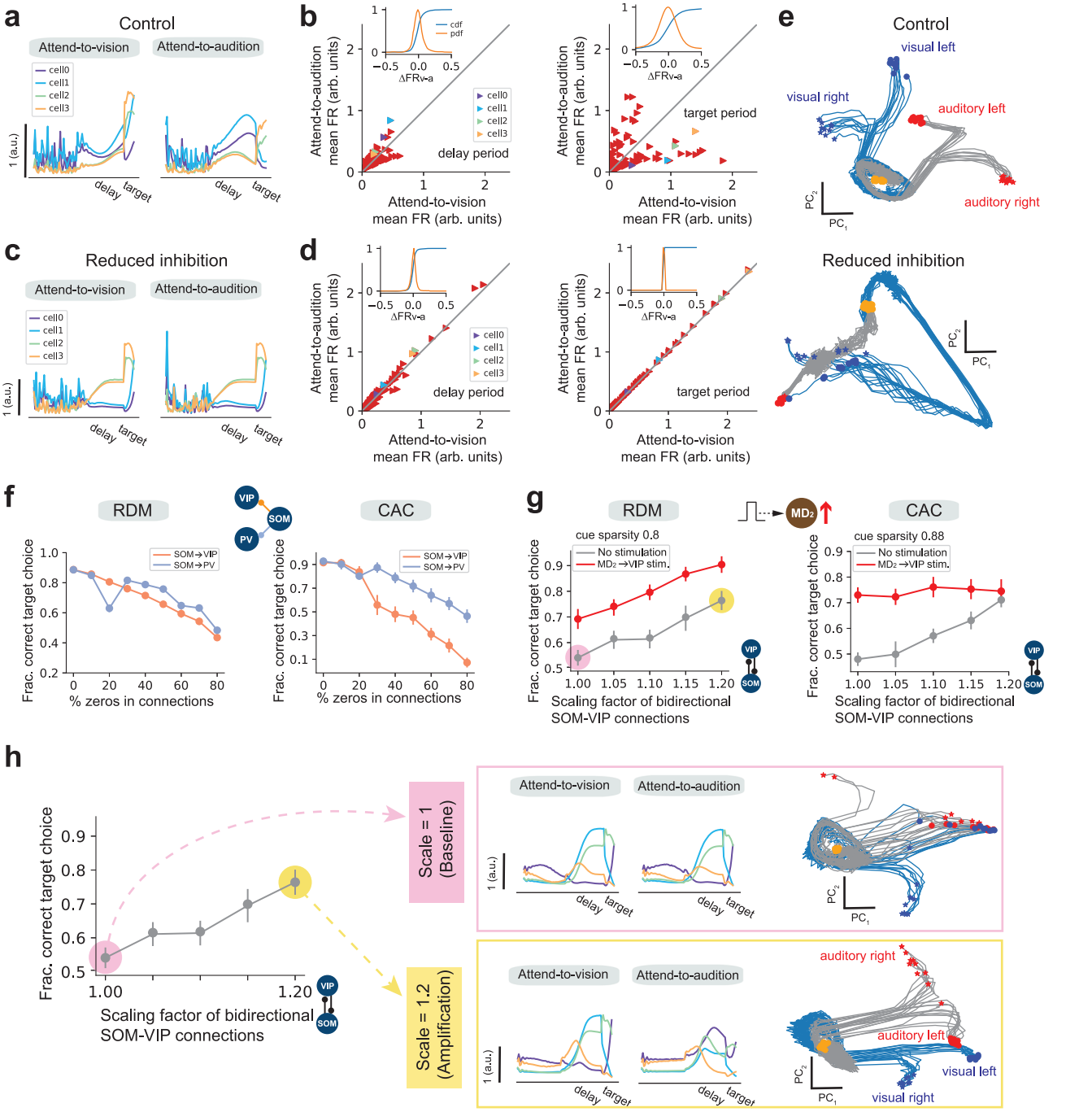


Fig. 5 | Parsing cognitive deficits and probing mechanistic causes in PFC-MD networks. a Tuning curves of four selective PFC excitatory units (indicated by different colors) during the control condition. These four units showed rule tunings ("attend-to-vision" vs "attend-to-audition"). **b** At the population level, mean firing rate (FR) comparison of PFC excitatory units between two rules during delay (left panel) and target (right panel) periods. Four units are labeled in the same color in (**a**). Inset: Curves of probability density function (pdf, orange) and cumulative distribution function (cdf, blue). **c** In the presence of prefrontal E/I imbalance (e.g., reduced inhibition), PFC excitatory units shown in (**a**) decreased their discriminability in rule tuning during both delay and target periods. **d** Similar to (**b**), except for the reduced inhibition condition. Comparing the cdfs of $\Delta FR_{v\text{-}a}$ = FR (attend-to-vision)−FR (attend-to-audition) between (**b**) and (**d**) showed statistically significant differences (delay period: $p = 4 \times 10^{-7}$; target period: $p = 4.8 \times 10^{-15}$, Kolmogorov–Smirnov test). **e** Two-dimensional neural trajectory representations in the control and E/I imbalance conditions. Trajectories were generated from all correct and error trials. Orange dots: cue off and start of the delay period. Blue and red end points represent two rule representations, whereas dot and star symbols represent left and right choices, respectively. **f** Task performance decreased with reduced prefrontal SOM→VIP and SOM→PV connectivity. Statistics are presented in mean ± SD ($n = 10$ Monte Carlo runs). **g** Increasing mutual inhibition strengths between SOM and VIP neurons amplified the gain under a high cue sparsity. Statistics are presented in mean ± SD ($n = 10$ Monte Carlo runs). Activating the MD_2→VIP pathway further facilitated amplification and improved task performance. Two shaded circles indicate the two conditions illustrated in (**h**). **h** In the case of RDM task of (**g**) (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.

https://doi.org/10.1038/s41467-025-58011-1

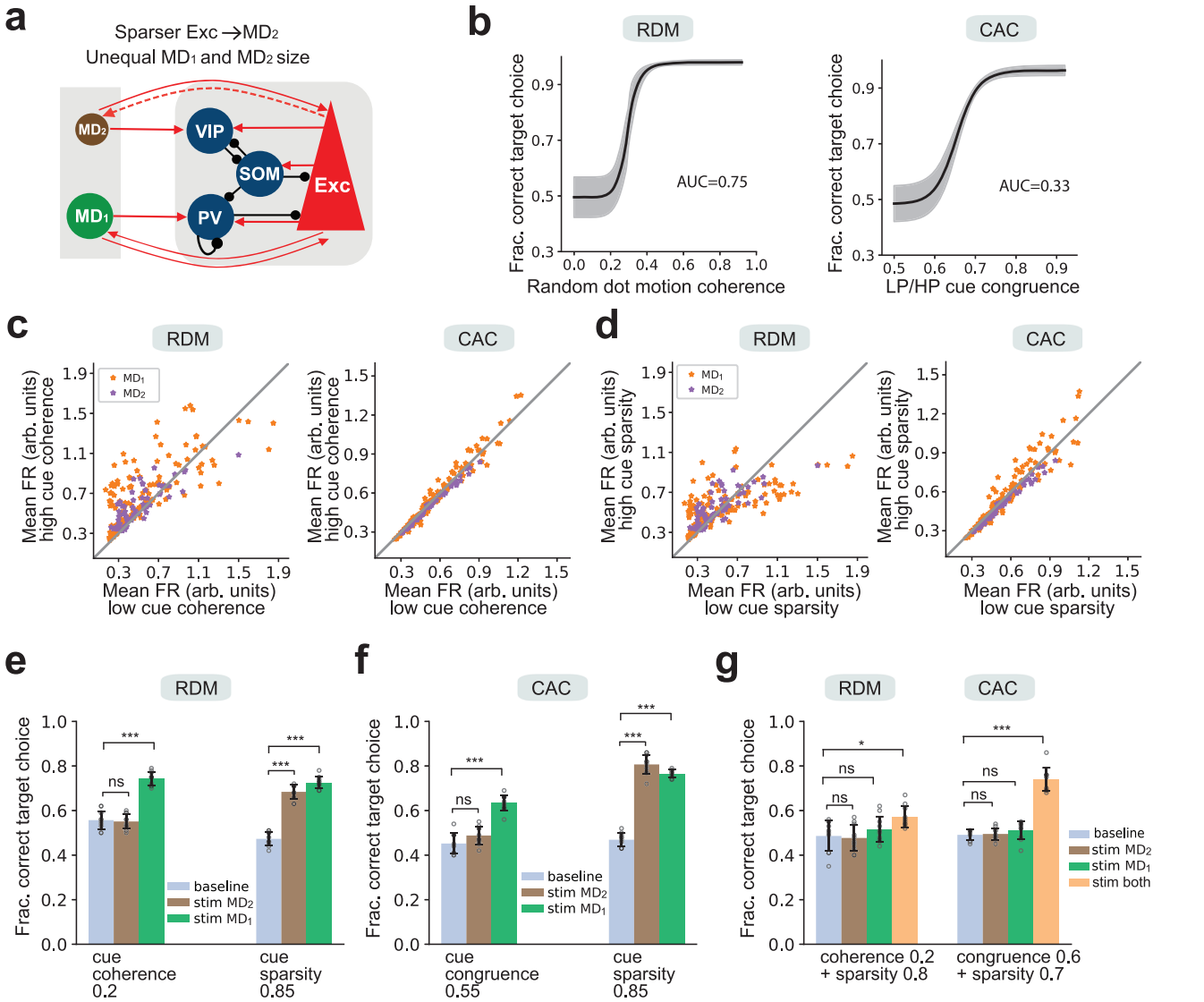


Fig. 6 | Modified PFC-MD model with imposed sparse corticothalamic connectivity. a The modified PFC-MD model with assumed unbalanced MD subpopulations (as shown by two different circle sizes, where the size of MD_1 is greater than the size of MD_2) and sparser Exc→MD_2 connectivity than Exc→MD_1 connectivity (as shown by a weaker connection in dashed line). **b** Psychometric curves of task-optimized modified PFC-MD model in two contexts. Shaded areas denote the SD ($n = 10$ realizations). Mean AUC showed similar or statistically non-significant values compared to the model described in Fig. 1b (RDM: $p = 0.496$; CAC: $p = 0.821$, two-tailed rank-sum test). **c** Population statistics of mean firing rates (during the cueing period) of MD_1 and MD_2 units for encoding cue uncertainty. **d** Population statistics of mean firing rates (during the cueing period) of MD_1 and MD_2 units for encoding cue sparsity. **e** In the RDM context, MD_1 activation, but not MD_2 activation, improved task accuracy under cue uncertainty. MD_1 or MD_2 activation improved task performance under cue sparsity. Statistics are presented in mean ± SD ($n = 10$ Monte Carlo runs). Cue uncertainty comparison: Baseline vs. MD_2 activation ($p = 0.82$); baseline vs. MD_1 activation ($p = 0.00016$); MD_2 vs. MD_1 activation ($p = 0.00016$). Cue sparsity comparison: Baseline vs. MD_2 activation ($p = 0.00016$); baseline vs. MD_1 activation ($p = 0.00016$); MD_2 vs. MD_1 activation ($p = 0.045$), two-tailed Wilcoxon rank-sum tests. **f** Similar to (**e**), except in the CAC context. Statistics are presented in mean ± SD ($n = 10$ Monte Carlo runs). Baseline vs. MD_2 activation ($p = 0.16$); baseline vs. MD_1 activation ($p = 0.00016$); MD_2 vs. MD_1 activation ($p = 0.00043$). Cue sparsity comparison: baseline vs. MD_2 activation ($p = 0.00016$); baseline vs. MD_1 activation ($p = 0.00016$); MD_2 vs. MD_1 activation ($p = 0.017$), two-tailed Wilcoxon rank-sum tests. **g** When both cue uncertainty and cue sparsity were present, MD_1 or MD_2 activation alone couldn't improve task performance, but activation of both MD subpopulations could. Statistics are presented in mean ± SD ($n = 10$ Monte Carlo runs). Baseline vs. co-activation, *, $p = 0.011$; ***, $p = 0.00015$, two-tailed Wilcoxon rank-sum tests. Other paired comparisons were n.s.

Unbalanced corticothalamic connectivity reshapes MD tunings

Next, we investigated whether modified corticothalamic connectivity can change MD unit tunings. Data from previous mouse experiments have suggested that MD_{GRIK4} neurons received dense inputs from L5/6 of the prelimbic cortex, in contrast to MD_{DRD2} neurons that receive sparse prefrontal inputs (unpublished data from the Halassa lab). To incorporate this prior knowledge into the PFC-MD model, we made two modifications. First, we assumed unbalanced MD_1 and MD_2 subpopulations (3:2 ratio in place of 1:1 ratio). Second, we assumed that MD_2 units received a much sparser projection from prefrontal excitatory units than MD_1 units and imposed a sparsity constraint that Exc→MD_2 connectivity was 75% sparser than Exc→MD_1 connectivity (Fig. 6a). Notably, this modified corticothalamic connectivity produced quantitatively similar AUC statistics in psychometric curve (Fig. 6b vs. Fig. 1d; $p = 0.496$ for the RDM context; $p = 0.821$ for the CAC context, rank-sum test), but led to emergent weak tunings in some MD_2 units with respect to cue uncertainty (Fig. 6c) and cue sparsity (Fig. 6d). Additionally, MD_1 activation, but not MD_2 activation,

improved the task accuracy under cue uncertainty (by "suppressing interference noise"); yet either MD_2 or MD_1 activation could improve the task performance under cue sparsity (by "enhancing the target signal") (Fig. 6e, f). Furthermore, when cue uncertainty and sparsity were simultaneously present, co-activation of both MD_1 and MD_2 subpopulations was required to achieve task improvement (Fig. 6g). Importantly, these results were also replicated when changing the proportions of inhibitory PFC cell types (Supplementary Table 1). Together, these findings further suggest distinct MD modulatory functions imposed by cell type-specific connectivity-dependent thalamocortical projections and support the "connectivity determines function" doctrine.

Bidirectional PFC-MD synaptic plasticity enabled rapid context switching

Learning changing contexts or context-dependent rule switching is a hallmark of cognitive flexibility. In the second version of context-dependent decision-making task, both cue and cue-to-rule mapping uncertainties were present. The context and rule were learned sequentially. Specifically, we introduced cue-to-rule mapping uncertainty and sequential context switching (*Context 1→Context 2→ Context 1'*) using the same CAC task and a slightly modified PFC-MD model with different network input dimensionality ($N_{in} = 6$) (Fig. 7a and "Methods"). At the first stage, Interestingly, PFC inhibitory-to-excitatory connectivity formed two distinct patterns according to the rule tuning preference of excitatory units (Supplementary Fig. 7a); a further examination of all connections between any PFC or MD units to postsynaptic PFC excitatory units revealed distinct four clusters according to their rule-tuning patterns (Supplementary Fig. 7b).

At the second stage, to highlight the role of the MD or thalamocortical connectivity, using the newly adopted PFC-MD model (Fig. 6a), we purposely kept the intracortical synaptic connectivity intact after successful completion of learning *Context 1*. Specifically, we adapted all other synaptic connections (except PFC-to-PFC connections) based on inter-trial error feedback to learn context switching on a trial-by-trial basis. Notably, updating thalamocortical and corticothalamic connections from back-propagating single-trial errors allowed the pretrained PFC-MD network to learn the new cue-to-rule transformation rapidly (Fig. 7b). We found that MD→Exc connectivity played a critical role in context switching and rule remapping (Fig. 7c). Interestingly, the change in thalamocortical connections also mapped to functional cell types in PFC excitatory units, the majority of which (>60%) displayed rule-invariant tunings despite context switching (Fig. 7d). Some PFC excitatory units did not change firing rates but shifted their peak firing rates temporally during the delay period (Supplementary Fig. 7c). Together, the PFC excitatory population displayed robust rule-invariant sequences during the delay period (Fig. 7e). Furthermore, the rule-tuning of PFC excitatory units tended to group with context-dependent E/I ratio and modified the E/I input accordingly while performing context switch successfully (Fig. 7f, left panel); in contrast, the PFC units would fail to discriminate the rule if context switching was unsuccessful (Fig. 7f, right panel). To quantify the degree of neural plasticity, we also computed the mean change in bilateral MD-PFC connection strengths during two consecutive context switches (Fig. 7g). In two context-switching conditions, we found that two dominant changes in MD-PFC plasticity were the Exc→MD_1 and MD_2→VIP pathways, suggesting distinct roles between corticothalamic and thalamocortical connectivity in cue-to-rule remapping, partly because of sparser Exc→MD_2 connectivity and unequal size between MD_1 and MD_2 subpopulations.

At the neural representation level, MD units showed context-invariant tuning between *Context 1* and *Context 1'* (Fig. 7h; and a unit tuning example in Fig. 7i). Additionally, the cue-to-rule transformation of the RNN at the steady state was qualitatively similar between *Context 1* and *Context 1'* (Supplementary Fig. 7d). Furthermore, a small subset of MD units increased firing with inter-trial error during context-switch learning: they increased the delay-period firing rates during the transient remapping state and resumed the baseline firing when succeeding context switching (Fig. 7j; top: single-unit example; bottom: mean firing rate statistics, $n = 8$ from one trained PFC-MD network). Imbalanced E/I or insufficient MD-to-PFC modulation led to a failure in cue-to-rule remapping at both single-unit and population representations (Fig. 7f, k).

Computational insight into MD feedforward control in context switching

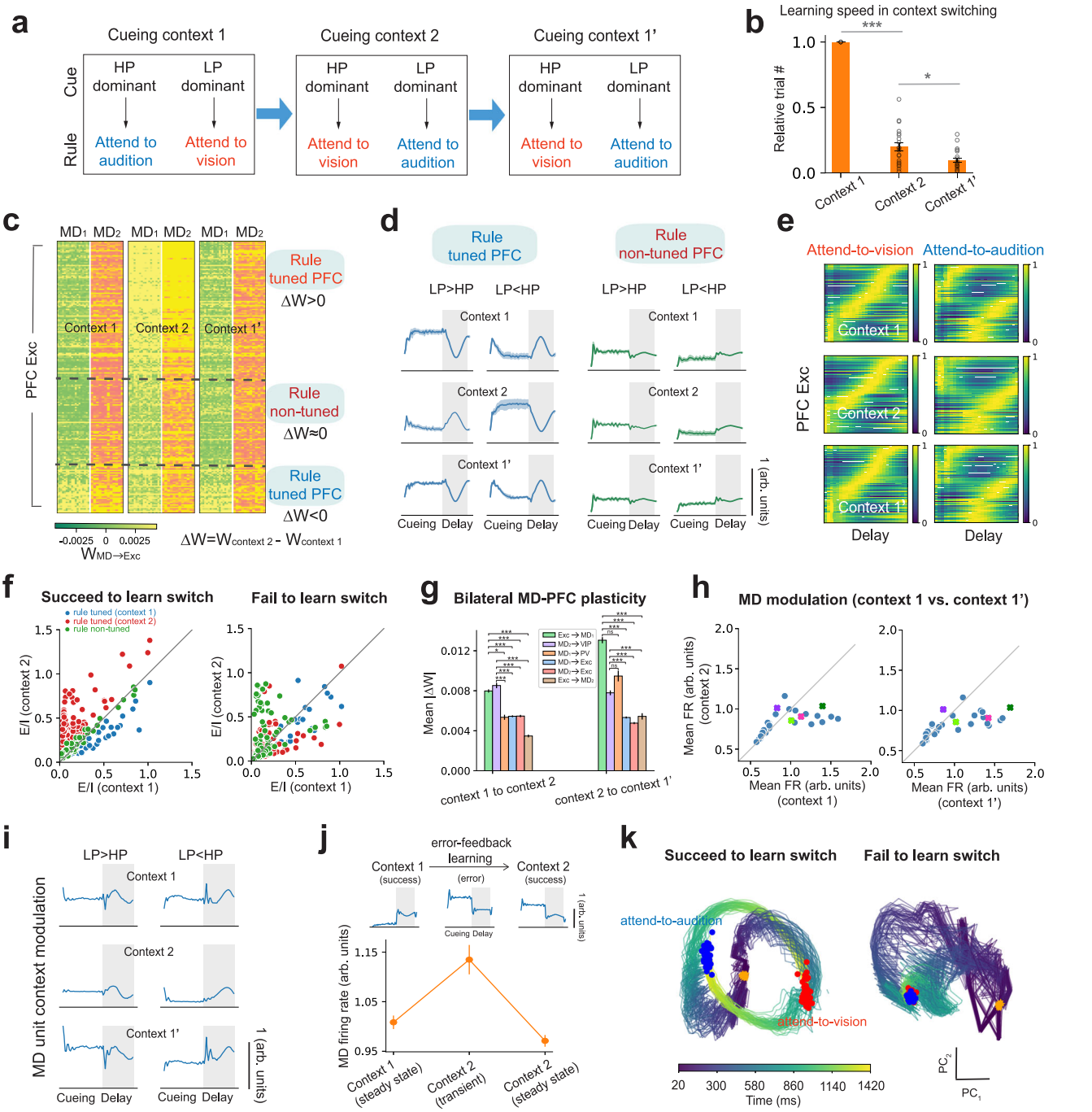
Our biologically-constrained computational models have provided a paradigm to test the role of MD in regulating task uncertainty in flexible and context-dependent decision-making (Fig. 8a and Supplementary Table 1). Bilateral interactions between the PFC and MD are critical for cognitive flexibility. Why and how does the combination of a recurrent E/I structure (PFC) and a feedforward excitatory architecture (MD) enable flexible cognitive control to deal with task uncertainty and context switching?

To further probe this question, we first investigated the impact of MD on the learning speed of context switching and network properties. We systematically varied the size of MD population ($N_{MD} = 2, 8, 16, 30$; $N_{PFC} = 256$) and retrained the respective PFC-MD networks, during which the input dimensionality remained constant ($N_{in} = 6$). Generally, adding sufficiently more MD units gradually improved the switch speed in terms of number of epochs to achieve convergence during trial-by-trial learning (Fig. 8b). Additionally, PCA showed that the feedforward MD structure had a lower intrinsic dimensionality than the recurrent PFC structure in task-optimized PFC-MD networks (Fig. 8c).

Second, as control experiments, in addition to keeping intracortical PFC-to-PFC connectivity intact, we systematically kept the unilateral or bilateral MD-PFC synaptic connectivity unchanged during context switching and only allowed synaptic plasticity among the remaining connectivity (Fig. 8d, top panels). Our result showed that the PFC rule-invariance property was lost when MD plasticity was disrupted, and the cue-to-rule remapping speed was the slowest when bilateral MD-PFC plasticity was disabled (Fig. 8d; compared to the baseline where intracortical connectivity was intact). In contrast, enabling MD→PFC plasticity significantly improved the switching speed, suggesting the crucial role of thalamocortical input in regulating prefrontal computation to enable cognitive flexibility. Interestingly, disabling corticothalamic plasticity had little effect on the change in switching speed, whereas disabling thalamocortical plasticity had a significant effect on the switch speed compared to the baseline ($p < 0.001$, rank-sum test). From a computational perspective of thalamic control (Methods), MD→PFC thalamocortical connectivity provides a computationally appealing strategy to control the cortical state and context-encoding MD firing in context switching.

Conceptually, we envisioned that the MD acts like an ON/OFF switch that modulates the respective E/I input of rule-tuned PFC units and changes prefrontal tunings according to the context. If the recurrent prefrontal circuit is viewed as an attractor network (Fig. 3i), the context-encoding MD can play the role of feedforward controller to usher the PFC to a proper attractor space, facilitating noise reduction (in the presence of conflicting cues) and cue-to-rule remapping (Fig. 8e). In a geometrical viewpoint, if the PFC dynamics is visualized as a two-dimensional (2D) system, the lower-dimensional (i.e., 1D) MD is acting like an actuator on the 2D system; rotating the system by 180 degrees through MD-PFC plasticity may enable cue-to-rule remapping while keeping the original PFC dynamics intact (Fig. 8f). This intuition was confirmed by our computer simulations, where prefrontal population dynamics under *Context 1* and *Context 2* were projected onto the same subspace (Fig. 8g). Additionally, the orthogonal relationship of PFC population dynamics was lost if the PFC-MD network failed to

https://doi.org/10.1038/s41467-025-58011-1



learn the context (Fig. 8h, left panel, dashed lines). Furthermore, when the MD was lesioned, the neural trajectory of PFC population (Fig. 8h, middle panel, green dash curve) was markedly different from that observed during successful context switching (Fig. 8h, middle panel, gray solid curve). Specifically, in the MD lesion condition, the end-points of trajectories that represent two rules were not separable at the first principal component axis. However, with additional MD activation in the other pathway (Fig. 8h, right panel), the neural trajectory of PFC activity recovered to the normal successful condition, suggesting that MD activation may rescue the cognitive deficit whereas MD lesion or inhibition has an opposite effect.

In light of our prior findings of MD and PFC representations in single-unit (Fig. 2) and population (Fig. 3) levels, our model can make further new predictions. Because the MD encodes the context, it shall reorient its neural subspace during context switching, therefore the principal angles between consecutive contexts shall be close to 90 degree such that the context-dependent MD subspaces are mutually orthogonal. Meanwhile, the neural subspace spanned by the PFC during consecutive context switching shall remain orthogonal. Indeed, detailed population analyses derived from our computational PFC-MD models confirmed this intuition and validated the predictions of MD-MD and PFC-PFC subspace angles: the angle between *Context 1* and *Context 2* and the angle between *Context 2* and *Context 1'* are relatively large, whereas the angle between *Context 1* and *Context 1'* is relatively small (Fig. 8i). Together, they provided clear geometric insight into thalamic regulation for prefrontal computation (Fig. 8j).

Experimental testing in a schizophrenia mouse model

Given the biophysical plausibility of our PFC-MD model, we wished to examine its utility in connecting circuit mechanism to algorithmic

Fig. 7 | Thalamocortical plasticity in the PFC-MD model enabled rapid context switching. a Schematic diagram of the context-switching task (*Context 1→Context 2→Context 1′*). **b** Relative learning speed in context switching, where neural plasticity of local MD-PFC connectivity was induced in trial-by-trial learning of *Context 2* and *Context 1′*. Error bar denotes s.e.m. ($n = 20$ independently trained models). Relative learning speed comparisons: *Context 1* vs. *Context 2* ($p = 1.9 \times 10^{-6}$, two-tailed rank-sum test); *Context 2* vs. *Context 1′* ($p = 0.003$, two-tailed rank-sum test). **c** Heatmap of connectivity of MD (MD_1 and MD_2) to prefrontal excitatory (Exc) units. Let W denote the MD-to-Exc connection matrix, units were sorted based on the change of $\Delta W = W_{context\ 2} - W_{context\ 1}$. According to the synaptic change ΔW, PFC excitatory units were mapped to two functional cell types: rule tuned (red and blue) vs. non-tuned (green) units. **d** Tuning curve examples of rule tuned (blue) and non-tuned (green) PFC units. Shaded area denotes the task delay period. Shaded areas around the PSTH denote the SD. **e** PFC excitatory units showed context-invariant rule-specific sequential activity during the delay period. Each row of the heatmap corresponded to the normalized trial-averaged firing activity. Units were sorted based on the location of peak firing rates and ranked in the same order in all six panels. **f** The E/I input of PFC excitatory units during the task delay period were clustered according to the PFC rule tuning properties when the PFC-MD network succeeded to learn the context switch. The cluster structure was lost when the PFC-MD network failed to learn the context switch. Units are color coded using the same color scheme based on rule-tuned (red and blue) or non-tuned (green) property in panel (**c**). **g** Quantification of mean synaptic plasticity (error bar: s.e.m., the number of connections for each colored bar, $n = 3690, 3690, 306, 306, 3690, 3690$, respectively) in bilateral MD-PFC connections during two consecutive context-switching conditions. In general, MD_1→Exc, MD_2→Exc, and Exc→MD_2 connections showed the overall smallest changes, whereas Exc→MD_1 and MD_2→VIP connections showed the overall dominant changes. All *p*-values from pairwise comparisons were based on the two-tailed rank-sum test (*$p < 0.05$, ***$p < 0.001$). **h** MD units showed context-invariant firing. Note that some MD units (overlaid "×" symbols) preserved their mean firing rates during the delay period between *Context 1* and *Context 1′*. **i** Turning curve illustrations of one MD unit during context switching. Shaded area denotes the task delay period. Note the strong context modulation between different contexts, but little modulation with respect to the rule. **j** A subset of MD units ($n = 8$ out of 30 units from one trained PFC-MD model) showed increased modulation with respect to decision error during the transient switching stage. Statistics are presented in mean ± SD. The mean MD firing rate was averaged across the delay period when the change of network state became very small (i.e., steady state). Tuning curves of one MD were shown on the top. **k** Comparison of neural trajectories of PFC population dynamics during cueing and delay periods between conditions in which the PFC-MD model succeeded or failed to learn context switching. Trajectories were color coded to represent time from the cue onset to the end of delay.

process relevant to computational psychiatry. To do so, we conducted computer simulations to reduce prefrontal cortical PV inhibition in the model (Fig. 6b and Methods), mimicking one of known changes in schizophrenia[38]. Dysfunction of GABAergic inhibition is known to impact synaptic E/I balance that is linked to pathophysiology of disease phenotypes including cognitive deficits and negative symptoms. Consistent with some known changes seen in schizophrenia patients, our PFC-MD model with reduced prefrontal inhibition also exhibited several important signatures, such as impaired working memory, sensitivity to noise (Fig. 9a), and switching deficits (Fig. 9b; see also Fig. 7f). These changes are reminiscent of reduced cognitive control seen in humans diagnosed with schizophrenia[39].

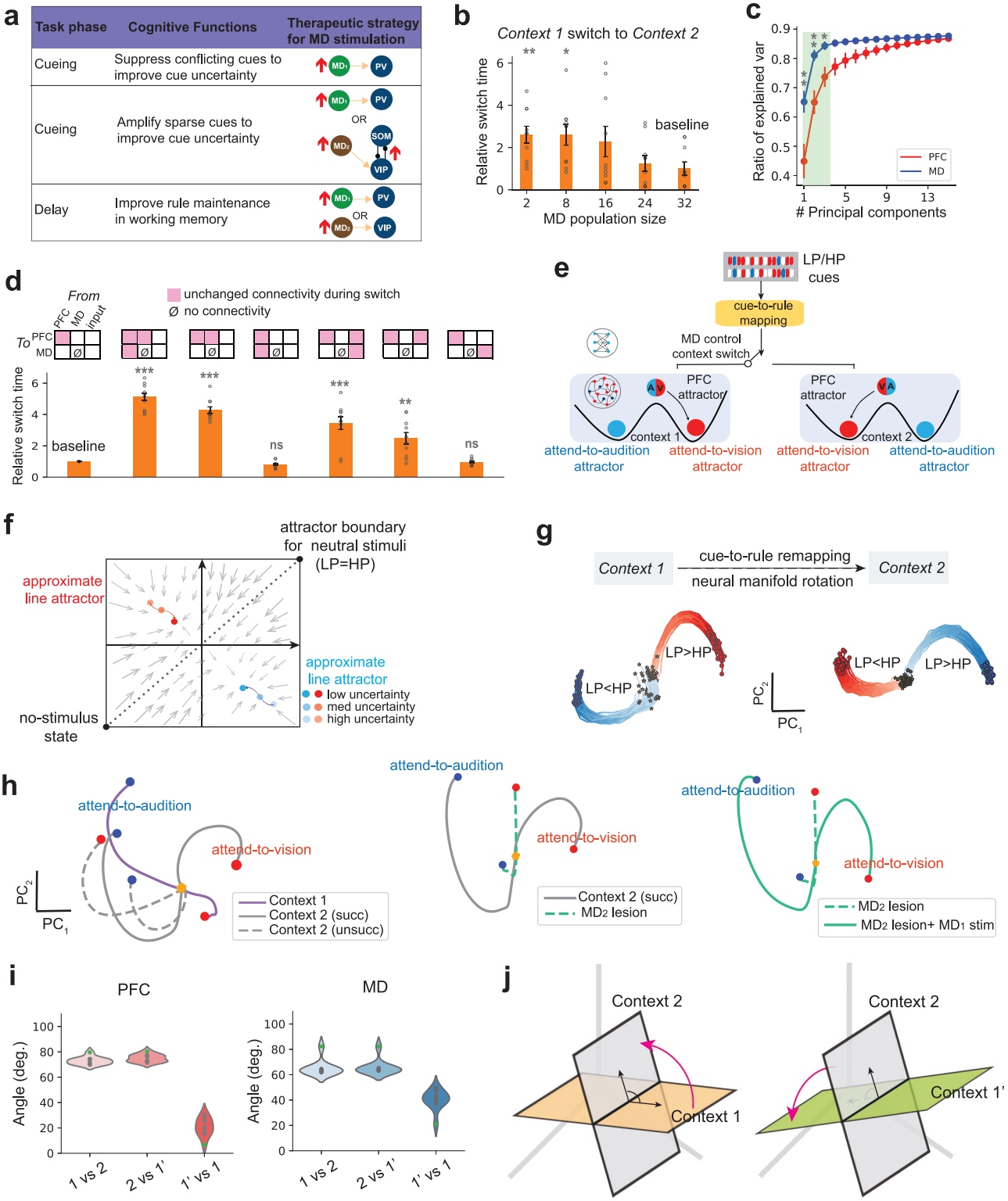
Recent data in both humans[40] and mice[41] have indicated that the MD thalamus is a putative target for cognitive deficits in schizophrenia. Specifically, Cascella and colleagues used deep brain stimulation (DBS) to target the basal ganglia output to the MD, which resulted in a profound symptomatic rescue in schizophrenia patients[40]. Zhou and colleagues used a mouse model relevant to schizophrenia and showed that optogenetic MD activation resulted in rescuing deficits in decision updating[41]. Strikingly, activating MD subpopulations in our model in the presence of aberrant PFC inhibition rescued the schizophrenia-like behavioral deficits (Fig. 9a; see also Fig. 4b, c and Fig. 5e, f for relevant results). These effects of rescuing in switching deficits were not only seen in the reversal CAC task (Fig. 9b; task described in Fig. 7a) but also in a cross-modal cueing context-switching task (Fig. 9c and Methods), as done previously in mice and shown to rely on the MD function[7]. Together, these results suggest that E/I balance is required for both maintaining task performance under noise disruption and adapting to the next context. Furthermore, in light of the observation that E/I imbalance deteriorates the reaction time in decision making[42], our computer simulations supported that MD activation rescued the reaction-time deficit in decision-making for the E/I-imbalanced model (Fig. 9d). Additionally, our modeling results showed that a decreased MD population size (Fig. 8b) or impaired MD-PFC pathway (Fig. 8d) reduced the speed of learning the context switch, resembling decreased MD volume[43] and cognitive inflexibility in schizophrenia[31,44].

To further experimentally test these modeling results, we further examined PFC circuitry and behavior in the 22Q11 deletion syndrome (DS) mouse, a genetic model with high construct validity relevance to schizophrenia[45]. It is well known that 22Q11DS mice exhibit deficits in set shifting, which have elements that overlap with cognitive flexibility known to rely on MD-PFC interactions. We first examined the cellular correlates of prefrontal E/I balance in relation to MD inputs. From MD→PFC anterograde labeling (Fig. 9e) and immunohistochemistry techniques (Fig. 9f; Methods), we found that MD neurons innervate fewer PV positive (PV^+) inhibitory interneurons in the PFC of 22Q11DS mice (Fig. 9g). In behavior, we trained mice to perform a cross-modal cueing context-switching task[7] (Fig. 9h; Methods), where 22Q11DS mice show prolonged switching latency compared to the published normative data[7]. Importantly, optogenetic MD activation during the first five post-switch trials resulted in normalization of these switching deficits (Fig. 9i; Methods), consistent with our modeling results and the broader experimental literature[14,46].

Discussion

The thalamus has various subnetwork motifs and diverse cortical target outputs, the higher-order thalamic nuclei, such as the MD is critical for dynamic regulation of cortical activity in attention, executive control, cognitive flexibility, and perceptual decision-making[10,46,47]. Several computational models of thalamocortical networks have been developed for the sensory thalamus[48,49] and higher-order thalamus[23,30,31,50]. However, it remains unknown how the MD thalamus regulates task uncertainty to enable flexible decision making. Complementary to experimental investigations, theory and modeling can help unravel key computational mechanisms[51]. Motivated by multiple lines of animal data[6,7,52,53], we trained biologically inspired PFC-MD models with interneuron subtype and thalamocortical pathway specificity to perform two versions of context-dependent decision-making tasks with various sources of task uncertainty. Our proposed PFC-MD model represents a conceptual advance from multiple previous modeling efforts[7,18,30,31,54] in that it allows us to dissect the functional roles of individual MD subpopulations or specific thalamocortical pathways in regulating task uncertainty and enabling cognitive flexibility (i.e., cue-to-rule remapping during context switching).

We found that the PFC-MD model outperformed the PFC-alone model in the presence of high cue uncertainty and cue sparsity, corresponding to lower SNR. Additionally, interneuron cell-type and thalamocortical pathway specificity play a vital role in regulating prefrontal computation during evidence accumulation (cueing period) and working memory (delay period) to enable or enhance cognitive flexibility. At different task phases, the MD regulates prefrontal resources for information transmission and maintenance to separately encode the task variables (PFC for rule and MD for context) in a



coordinated manner. Rule information is encoded by prefrontal attractor-line dynamics and preserved during the delay period, whereas MD activation enhances the SNR and improves working memory maintenance. We also found cell-type specific tunings in the MD subpopulation in the model with respect to cue uncertainty and cue sparsity, and such MD tunings may be influenced by corticothalamic projection. Additionally, adaptation of MD-PFC connectivity enabled the PFC-MD network to rapidly learn context switching and maintain context-invariant rule encoding, suggesting the role of MD as an ON/OFF switch actuator to enable cognitive flexibility. The notion of the thalamus as an ON/OFF switch to assist the thalamocortical control is also in line with the concept in the motor system[55–57].

Cognitive flexibility requires PFC-MD coordination to map context to behavior, as well as to regulate both cue and cue-to-rule mapping uncertainties. Our computational model provides direct support for circuit mechanisms of PFC-MD network in flexible decision making. Distinct cellular targeting among cortical interneurons underlies differential inhibitory effects on excitatory neurons[58]. Cortical interneuron-specific cells that specialize in synaptic disinhibition of excitatory neurons may shape the way excitatory neurons

Fig. 8 | Computational and geometric insight of MD in regulating prefrontal computation to learn context switching. a Schematic summary of cell type and task-phase specific regulatory roles of MD thalamus in decision making under cue uncertainty. **b** Comparison of relative context switch time with different sizes of MD population. All time was normalized with respect to the baseline (MD size of 32). Error bar denotes s.e.m. ($n = 10$ independently trained models). All statistical tests were two-tailed rank-sum tests against the baseline ($N_{MD} = 2$ vs. baseline: $p = 0.0036$; $N_{MD} = 8$ vs. baseline: $p = 0.0052$; Bonferroni-corrected rank-sum test; $N_{MD} = 16$, 24 vs. baseline, n.s.). **c** Ratio of explained variance of PFC and MD populations derived from task-optimized PFC-MD models. Our computer simulations and principal component analysis (PCA) showed that the recurrent PFC structure had a larger dimensionality than the feedforward MD structure. Error bar denotes s.e.m. ($n = 10$ independently trained models). Statistical tests were conducted between PFC and MD regarding the ratio of explained variance based on #principal components (PCs). The first three points were used to illustrate the differences based on two-tailed rank-sum tests ($p = 0.003$ for the first PC; $p = 0.0015$ for the second PC; $p = 0.012$ for the third PC). **d** Comparison of relative context switch time with different assumptions of modifiable connectivity during switch. In the top, unilateral or bilateral PFC, MD and input connectivity patterns are shown (∅ denotes no connectivity). In total, there are five unidirectional connections that can be adapted. In the bottom, all time in the y-axis was normalized with respect to the baseline (unchanged intracortical connectivity). Error bar denotes s.e.m. ($n = 10$ independently trained models). All statistical tests were two-tailed rank-sum tests against baseline (the respective p-values from left to right are 0.0002, 0.0002, 0.0233, 0.0007, 0.0025, and 0.4497). **e** Schematic illustration of the MD's role in context switch during cue-to-rule remapping, and the PFC is illustrated as a bistable attractor. **f** Two-dimensional (2D) vector field that illustrates the bistable PFC dynamics as an approximate line attractor. The diagonal dotted line represents an attractor boundary for neutral stimuli (e.g., #LP = #HP). Rotating the 2D vector field by 180 degrees is equivalent to switching the cue-to-rule transformation while preserving the PFC dynamics. **g** Comparison of PFC population dynamics under *Context 1* and *Context 2* while projecting them onto the same PCA subspace. The neural trajectories were nearly orthogonal to each other. Each color trace represents a single trial with specific cue input (#LP > #HP or #LP < #HP). Light to dark color represents the time evolution of the trial. **h** *Left panel*: Comparison of neural trajectories of PFC dynamics in *Context 1*, *Context 2* with successful context switching, and *Context 2* with unsuccessful context switching. *Middle panel*: Comparison of neural trajectories of PFC dynamics in *Context 2* with successful context switching and *Context 2* with MD_2 lesion. *Right panel*: Comparison of neural trajectories of PFC dynamics in *Context 2* with MD_2 lesion and *Context 2* with MD_2 lesion plus MD_1 stimulation. **i** Quantification of the principal angles of PFC-PFC and MD-MD subspaces during consecutive context switching ($n = 5$ independently trained PFC-MD models; orange dot represents the trained PFC-MD model used in Fig. 7 illustration). Principal angles were computed across trials during the cueing period. The angles were relatively large between two different contexts (1 vs. 2 or 2 vs 1′), whereas the angles were smaller between two similar contexts (1 vs 1′). **j** Geometric illustration of rotation of neural subspaces during two consecutive context switches. The angles of PFC-PFC and MD-MD changed in a coordinated manner.

integrate information, and exhibit context-specific and behavior-relevant responses[33,59]. 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[6]. The MD_1→PV projection is the primary drive for feedforward thalamic inhibition[60,61]. Cortical PV interneurons have been implied in the mechanisms of gamma oscillations and cognitive function[62]; disturbance in this signaling contributes to altered gamma oscillations and working memory deficits in schizophrenia[63,64]. Thalamocortical and corticothalamic pathways play different roles in decision-making processes[18]. Corticothalamic projection is known to adjust the gain and tuning precision of thalamic neurons as required by behavioral demands[11]. Our analyses of corticothalamic and thalamocortical projections suggest a continuous information flow and yet distinct contributions in cognitive control. Specifically, the MD→PFC projection can suppress competing cues or amplify sparse cueing signals to improve the SNR in a pathway-dependent manner (Fig. 4), whereas activating the PFC Exc→MD projection can achieve a similar effect (Supplementary Fig. 4). Together, enhancing the thalamocortical and corticothalamic loops can resolve the ambiguity in sensory cues and improve rule maintenance in working memory. During context switching, neural plasticity of thalamocortical connectivity seems to play a more crucial role in facilitating the switch than that of corticothalamic connectivity (Fig. 8d).

At the computational level, the regulatory role of MD can be explained in several aspects: increasing time constant of the network dynamics (through changing the eigenspectrum and increasing the degree of non-normality of the network connectivity matrix), aligning the MD activity subspace with the PFC activity subspace (through cue uncertainty tuning), and allowing thalamic feedforward control in context switching. In modulating cue uncertainty during the cueing period, the MD_1 subpopulation that targets on PV interneurons seems to play a more important role in cue evidence integration (Supplementary Fig. 2d and Supplementary Table 1). In working memory during the stimulus-free delay period, activating MD_1 or MD_2 subpopulation can enhance thalamocortical communications and boost working memory (Fig. 4b, c, g and Fig. 8a). In cue-to-rule remapping, from a geometric viewpoint, the MD orients its neural subspace to accommodate context switching while maintaining context-invariant prefrontal rule representations (Fig. 8i). Furthermore, lower-dimensional representations of the MD thalamus may facilitate compression of higher-dimensional cortical information and enable predictive coding in the context of cognitive flexibility. Specifically, since the MD modulates with cue uncertainty and context ("prior"), whereas the PFC encodes the rule ("likelihood"), integrating information from both the MD and PFC provides a natural paradigm for updating the posterior in a Bayesian framework[65].

The capacity of information transmission between neurons or circuits is fundamentally limited by the SNR. Neuromodulation, such as dopamine may enhance the SNR of prefrontal activity[66,67]. Since the MD is known to receive dopaminergic inputs[68], 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[6]. 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[69]. Our results are also consistent with the idea that the MD projections target both prefrontal glutamatergic pyramidal neurons and GABAergic interneurons can regulate the changes in GABA activity and further change prefrontal E/I balance and functional connectivity[70,71].

Adaptation to rule or behavior according to the task context is a fundamental property in cognitive control. Cognitive flexibility is facilitated by the inclusion of MD in a PFC-driven decision-making task[6,7,72]. While the PFC-alone can perform the task equally well in an ideal task condition, it lacks flexibility in dealing with various sources of task uncertainty. We reason that adaptive behaviors are driven by both fast and slow learning, and the feedforward thalamic structure is appealing to one-shot learning at a faster timescale, in contrast to the recurrent cortical structure that requires slower neural plasticity. Our computational simulations have shown that thalamocortical plasticity of a pretrained PFC-MD model enables a rapid and reversible cue-to-rule remapping for context switching. Moreover, we found that the thalamocortical MD→PFC connection plays a more vital role than the corticothalamic PFC→MD connections in the speed of context switching. During remapping, rule information is invariantly preserved by PFC population dynamics, resulting in a geometric rotation of neural trajectories. In biology, the indirect corticothalamic feedback

https://doi.org/10.1038/s41467-025-58011-1

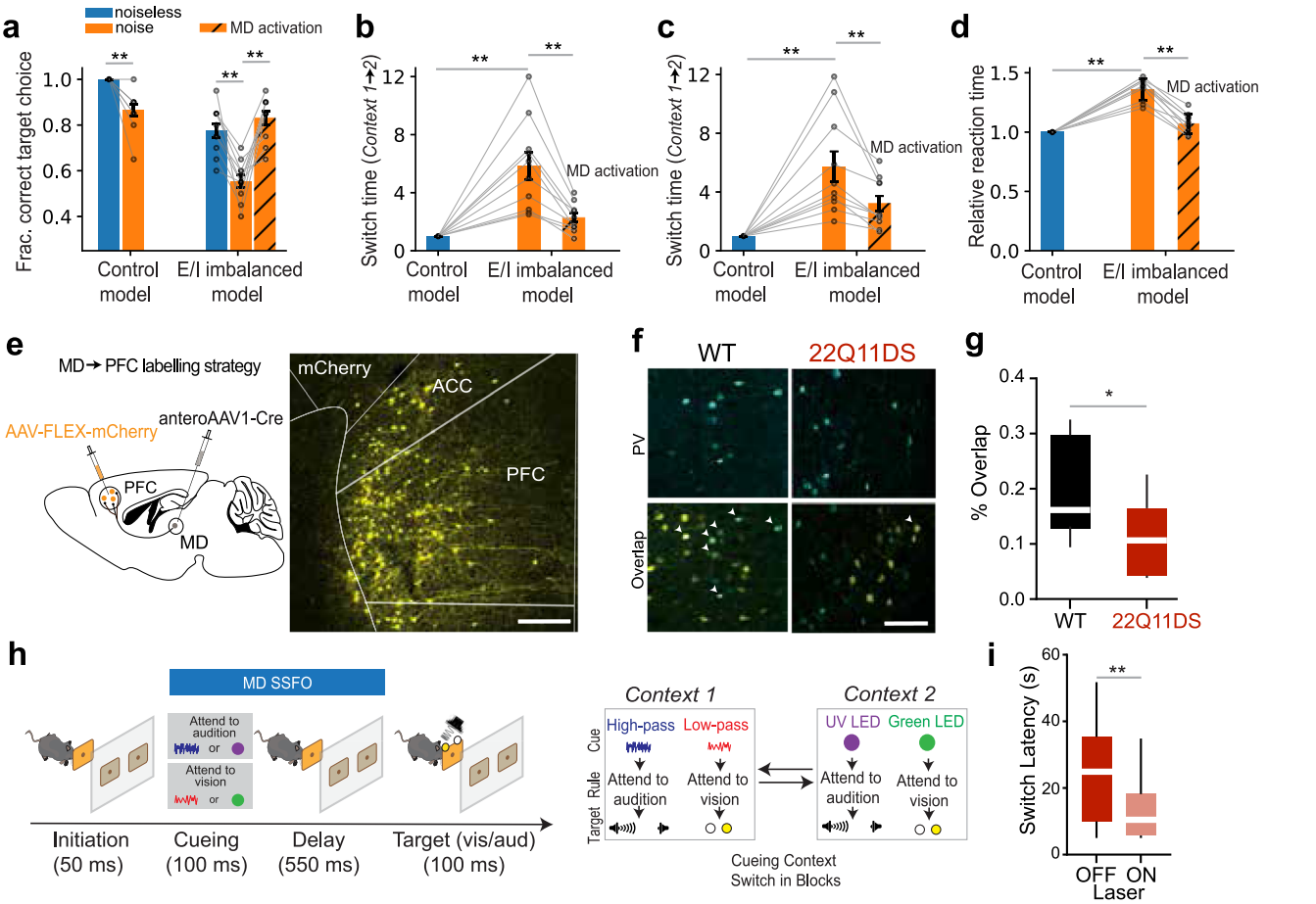


Fig. 9 | Computational prediction and experimental testing on a schizophrenia mouse model. a Reduced cortical inhibition or E/I imbalance in the trained PFC-MD model caused sensitivity to the noise during the task delay period, resulting in decreased task performance ($p = 0.0067$, noiseless vs. noise in control model, two-sided signed-rank test; $p = 0.0019$, noiseless vs. noise in E/I imbalance model, two-sided signed-rank test; the control model's performance was 1 in a noiseless condition). MD activation significantly improved the task performance in the presence of noise and E/I imbalance ($p = 0.0019$, noise vs. MD activation, two-sided signed-rank test). The illustration was shown for the CAC task, but the result was qualitatively similar for the RDM task. Statistics are presented in mean ± s.e.m. ($n = 10$ independent Monte Carlo runs). **b** Reduced cortical inhibition in the trained PFC-MD model caused switching deficit during the reversal context-switching task (*Context 1→Context 2*). The E/I imbalanced model had significantly greater switch time ($p = 0.0019$; two-sided signed-rank test), and MD activation rescued the switching deficit ($p = 0.0019$; two-sided signed-rank test). Statistics are presented in mean ± s.e.m. ($n = 10$ independent Monte Carlo runs). **c** Reduced cortical inhibition in the trained PFC-MD model caused switching deficit in a cross-modal cueing context switching task (CAC context→RDM context), which had significantly greater switch time than control model ($p = 0.0019$; two-sided signed-rank test). Similarly, MD activation rescued the switching deficit ($p = 0.0019$; two-sided signed-rank test). Statistics are presented in mean ± s.e.m. ($n = 10$ independent Monte Carlo runs). **d** Reduced cortical inhibition in the trained PFC-MD model caused slower relative reaction time ($p = 0.0019$; two-sided signed-rank test). MD activation improved reaction time ($p = 0.0019$; two-sided signed-rank test). Statistics are presented in mean±s.e.m. ($n = 10$ independent Monte Carlo runs). **e** Schematic of MD→PFC anterograde labeling strategy for output targets of MD neurons in the PFC of wild type (WT) and 22Q11DS mice. Representative image of PFC showing trans-synaptic labeling. Scale bar = 200 µm. **f** Representative images of immunohistochemically labeled parvalbumin (PV) neurons (*top panel*) and their overlap with mCherry+ MD output neurons (*bottom panel*) in WT and 22Q11DS mice. The lower degree of overlap in 22Q11DS mice suggests that reduced prefrontal inhibition. Scale bar = 200 µm. **g** Quantification of percentage of anterogradely labeled neurons which co-express PV revealed reduced innervation of PV neurons in 22Q11DS mice, which is equivalent to E/I imbalanced due to reduced cortical inhibition ($n = 3$ animals for each group; $^*p = 0.032$; two-tailed rank-sum test). **h** Schematic of task design with cross-modal cue switching component (adapted from REF[7]). **i** Optogenetic MD activation through SSFO (stabilized step function opsin) across the cueing and delay periods for 5 consecutive trials during the switch improved animals' behavioral switching latency in 22Q11DS mice ($n = 19$ sessions from 4 22Q11DS mice; $^{**}p = 0.0095$, two-tailed rank-sum test). For box plots in (**g**, **i**), boundaries, 25–75th percentiles; midline, median; whiskers, minimum–maximum.

(such as through other frontal cortical regions, including the orbitofrontal cortex and anterior cingulate cortex) may also play an important role in context switching[30,73].

What computational insight may our models provide to the studies of cognition in a broader context? First, our PFC-MD model provides a means to identify the role of thalamocortical and corticothalamic signaling and how populations of neurons interact between the source and target[74]. Second, computer simulations for various task difficulty conditions, which may be difficult to accomplish in animal experiments due to resource constraints, may produce experimentally testable predictions on behavior and neural representations (such as changes in neuronal tunings under task uncertainty). Lastly, our model with cell-type specificity provides a strategy to dissect computational mechanisms of cognitive deficits (e.g., sensory gating and top-down control) encountered in neuropsychiatric disorders, such as schizophrenia[15,75–78]. Behavioral and optogenetic stimulation experiments in a schizophrenia mouse model have confirmed our model prediction. Additionally, abnormal dopaminergic activation and lower cortical SNR have also been implicated in schizophrenia[66,67]. A lack of inhibitory control increases hyperexcitability and reduces cognitive flexibility in schizophrenia or ADHD[13,79,80]. Recent results have shown that bidirectional plasticity in

PFC-MD pathways may correct cognitive impairment[81]. To pursue these mechanistic inquiries, our modeling approach may provide a bridge to link circuit mechanisms to algorithmic processes in diseased brains. Finally, a complete dissection of these computational mechanisms may provide therapeutic insight into thalamic DBS strategies to improve cognitive functions in mental disorders.

As a reductionist approach, our current PFC-MD model has several methodological limitations. First, the current model is simplified related to the real biological circuit in that it does not account for the cortical layer structure or thalamic inhibition, and it does not distinguish distinct somatic and dendrite targets from the cortical interneurons. Therefore, our model with only cell type specificity still has conceptual limitations. Second, the current model uses a rate-based point-neuronal model instead of a spiking-based biologically realistic compartment-based model and cannot model the local field potential oscillations, or between-region coherence. Third, the current model does not explicitly model inputs from cortico-cortical communications as well as top-down or modulatory input from other areas of the brain. Nevertheless, compared to the mean-field model, our cell-type specific model can validate MD and PFC tunings with respect to the network architecture and thalamocortical connectivity; additionally, the reductionist approach may provide a principled paradigm to explore the key computational mechanisms of the PFC-MD circuit during cognitive flexibility and presents new paths for future investigations. For instance, we may extend the current model to a reward-based decision-making setting to match more broadly with animal or human experiments. Biologically inspired neural plasticity rule can be further imposed. We may also modify the decision-making task to model different dimensions of cognitive control and test the model generalizability across species and task behaviors.

Methods

Context-dependent decision making with parameterized cue uncertainty

The computational task was modified from a combination of a two-context cross-modal 4AFC task[18] and a single-modal two alternative force choice (2AFC) task with ambiguous cues[6]. Figure 1b illustrates the two cueing contexts that were mapped to the same rule: attend-to-audition or attend-to-vision. During the 800-ms cueing period, the network received cueing signal from either visual (moving dots) or auditory modality (LP/HP pulses), each of which had various degrees of cueing uncertainty. Under visual modality, the RDM context used moving dots as cues, where the coherence of moving dots in time indicates the degree of cueing uncertainty: with coherence 1 meaning all dots moving with the same direction (rightward, rule 1 vs. leftward, rule 2) and coherence 0 meaning completely random direction. The dimensionality of cue input in RDM was 64, and each dimension represents one random dot. We used the set [−1, 1] to represent the direction of dot moving: -1 corresponding to leftward, -1 corresponding to rightward, and the value of other direction is selected randomly between −1 and 1. Under auditory modality, the CAC context used a sequence of conflicting pulses as cues: LP pulse (+1) or HP pulse (−1). The {+1, −1} pulses were task-relevant cues, and the background was represented by 0 ('blank'). In training, the number of pulses during the 800-ms cueing period was 32, yielding a cue signal density 40 s^{-1}. The ratio between the numbers of two pulses defines the auditory cue congruence. Accumulating the cueing evidence would produce the indication of one rule (more −1's) or the other (more +1's). During the cueing period, the evidence of ambiguous cues needs to be accumulated to identify the rule. During 800-ms delay period of working memory, rule information needs to be preserved. During the 200-ms target period with "divided attention", the auditory and visual stimuli (independent 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), and each attended target was associated with a unique outcome of four choices. Therefore, the goal was to attend to the correct modality and select the one correct choice during the 200-ms choice period (Fig. 1b). When errors occurred, the task error could be ascribed to either "rule error" (or "executive error", which occurred in the cueing or delay period) or "sensory error" (which occurred in the target period), or both (Supplementary Table 2).

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 level of sensory error, we assumed two competing sensory inputs were stochastic and drawn from normal distributions with two different mean values but the same variance. Increasing the variance would increase the sensory error; we selected a variance value that best matched to the experimental data. In the extreme case where sensory inputs were deterministic, the model prediction would lead to only rule error.

Computational models

We first constructed a PFC-alone network as the baseline model. The PFC network is an E/I RNN with $N_{\mathrm{PFC}} = 256$ fully interconnected units described by a standard firing-rate model[25,26]. We adopted a continuous-time formulation of RNN dynamics as follows:

$$\tau \dot{\mathbf{x}}^{\mathrm{PFC}} = -\mathbf{x}^{\mathrm{PFC}} + \mathbf{W}^{\mathrm{rec}}\mathbf{r}^{\mathrm{PFC}} + \mathbf{W}^{\mathrm{in}}\mathbf{u} + \sqrt{2\tau\sigma_{rec}^2}\boldsymbol{\xi}. \quad (1)$$

where τ denotes the time constant (we used $\tau = 20$ ms), $\boldsymbol{\xi}$ denotes additive N_{PFC}-dimensional Gaussian noise, each independently drawn from a standard normal distribution, and $\sqrt{\sigma_{rec}} = 0.05$ 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 ($N_{\mathrm{in}} = 72$), which includes four types of noisy input: $\mathbf{u} = \left(\mathbf{u}_{\mathrm{RDM}}, \mathbf{u}_{\mathrm{CAC}}, \mathbf{u}_{\mathrm{sensory}}, \mathbf{u}_{\mathrm{context}}\right) + \mathbf{u}_{\mathrm{noise}}$, where $\mathbf{u}_{\mathrm{noise}} = \sqrt{2\tau\sigma_{in}^2}\boldsymbol{\xi}$. Here, the stimulus input $\boldsymbol{u}_{\mathrm{RDM}}$ denotes the cueing input in the RDM context ($N_{\mathrm{RDM}} = 64$) and $\boldsymbol{u}_{\mathrm{CAC}}$ denotes cueing input in the CAC context ($N_{\mathrm{CAC}} = 2$). The context input $\mathbf{u}_{\mathrm{context}}$ indicates which task the network is performing in the current trial ($N_{\mathrm{context}} = 2$). The $\mathbf{u}_{\mathrm{sensory}}$ represents sensory input during the target period ($N_{\mathrm{sensory}} = 4$). The input noise strength is $\sqrt{\sigma_{in}} = 0.05$. The network output consisted a set of linear readout units that produced a 4-dimensional target estimate: $\mathbf{z} = \mathbf{W}^{\mathrm{out}}\mathbf{r} + \mathbf{b}$, where $\mathbf{W}^{\mathrm{out}}$ is a $4 \times N_{\mathrm{PFC}}$ matrix, $\mathbf{b}$ is the bias vector, and the N_{PFC}-dimensional neuronal firing rate vector $\mathbf{r}$ is defined by a softplus function: $\mathbf{r} = \phi(\mathbf{x})$. The network selected the maximum mean output from $\mathbf{z}(t)$ to yield one out of four choices in decision. We assumed 4:1 ratio of the number of excitatory to inhibitory neurons and imposed Dale's principle on synaptic connectivity. To make the RNN adhere to Dale's principle, we let $\mathbf{W}^{\mathrm{rec}} = \mathbf{W}^{\mathrm{rec},+}\mathbf{D}$, in which $\mathbf{W}^{\mathrm{rec},+}$ is a non-negative matrix and $\mathbf{D}$ is a diagonal matrix. For example, consider a 5-dimension RNN dynamic system with one inhibitory unit and four excitatory units; given a diagonal matrix $\mathbf{D} = \mathrm{diag}(1, 1, 1, 1, -1)$, the recurrent weight matrix can be written as follows:

$$\underbrace{\begin{pmatrix} 1 & + & + & + & - \\ + & 1 & + & + & - \\ + & + & 1 & + & - \\ + & + & + & 1 & - \\ + & + & + & + & -1 \end{pmatrix}}_{\mathbf{W}^{\mathrm{rec}}} = \underbrace{\begin{pmatrix} 1 & + & + & + & + \\ + & 1 & + & + & + \\ + & + & 1 & + & + \\ + & + & + & 1 & + \\ + & + & + & + & 1 \end{pmatrix}}_{\mathbf{W}^{\mathrm{rec},+}} \underbrace{\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 \end{pmatrix}}_{\mathbf{D}} \quad (2)$$

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: PV, VIP-expressing, and SOM interneurons. The interneuron cell types had distinct inhibitory-to-excitatory and inhibitory-to-inhibitory connectivity properties[27–30]. 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. 1a). For simplicity, we further assumed that the percentages of three interneuron cell types were equal and these subtypes were determined by the synaptic interconnectivity. According to Dale's principle, the ratio of Exc and Inh units (including PV, SOM, VIP) is 4:1. In our computer simulations, the total number of PFC units was 256 ($N_{\mathrm{PFC}} = 256$), and we set $N_{\mathrm{exc}} = 205$, and $N_{\mathrm{inh}} = 51$ ($N_{\mathrm{PV}} = N_{\mathrm{SOM}} = N_{\mathrm{VIP}} = 17$). To specify the pattern of recurrent connections $\mathbf{W}^{\mathrm{rec}}$, we imposed a mask matrix $\mathbf{M}^{\mathrm{rec}}$ on the cell-type connectivity to parametrize the weight matrix as follows.

$$\begin{aligned}\mathbf{W}^{\mathrm{rec}} &= \mathbf{W}^{\mathrm{rec,\,plastic}} \odot \mathbf{M}^{\mathrm{rec}} \\ &= \begin{pmatrix} \boldsymbol{w}_{\mathrm{E}\to\mathrm{E}} & \boldsymbol{w}_{\mathrm{PV}\to\mathrm{E}} & \boldsymbol{w}_{\mathrm{SOM}\to\mathrm{E}} & \boldsymbol{w}_{\mathrm{VIP}\to\mathrm{E}} \\ \boldsymbol{w}_{\mathrm{E}\to\mathrm{PV}} & \boldsymbol{w}_{\mathrm{PV}\to\mathrm{PV}} & \boldsymbol{w}_{\mathrm{SOM}\to\mathrm{PV}} & \boldsymbol{w}_{\mathrm{VIP}\to\mathrm{PV}} \\ \boldsymbol{w}_{\mathrm{E}\to\mathrm{SOM}} & \boldsymbol{w}_{\mathrm{PV}\to\mathrm{SOM}} & \boldsymbol{w}_{\mathrm{SOM}\to\mathrm{SOM}} & \boldsymbol{w}_{\mathrm{VIP}\to\mathrm{SOM}} \\ \boldsymbol{w}_{\mathrm{E}\to\mathrm{VIP}} & \boldsymbol{w}_{\mathrm{PV}\to\mathrm{VIP}} & \boldsymbol{w}_{\mathrm{SOM}\to\mathrm{VIP}} & \boldsymbol{w}_{\mathrm{VIP}\to\mathrm{VIP}} \end{pmatrix} \odot \begin{pmatrix} 1 & -1 & -1 & 0 \\ 1 & -1 & -1 & 0 \\ 1 & 0 & 0 & -1 \\ 1 & 0 & -1 & 0 \end{pmatrix} \end{aligned} \tag{3}$$

where $\odot$ denotes the Hadamard product (i.e., the element-wise product between two matrices or two submatrices), $\mathbf{1}$ denotes an all-ones matrix with proper dimensionality, and the algebraic sign +/- denotes respective excitatory/inhibitory connection, and $\mathbf{0}$ denotes an all-zeros matrix with proper dimensionality. $\mathbf{W}^{\mathbf{rec,\,plastic}}$ represents the trained recurrent weight matrix. The zero elements in mask $\mathbf{M}^{\mathbf{rec}}$ ensure the corresponding weights have no contribution, implying the lack of synaptic connection. For instance, $\boldsymbol{w}_{\mathbf{PV}\to\mathbf{VIP}} = 0$ implies that there is no connection between PV and VIP neurons. To discretize continuous-time dynamics, we used a bin size of $\Delta t = 20$ ms for update in all numerical simulations.

We further constructed a PFC-MD model based on the knowledge of genetically identified thalamocortical projections and cell type specificity. We have taken into consideration of several thalamocortical circuit models developed for different task behaviors in the literature[53–56]. In light of experimental findings in task-performing mice[6], we assumed that the MD primarily consists of a smaller number of non-recurrent excitatory neurons ($N_{\mathrm{MD}} < N_{\mathrm{PFC}}$) and have two specific MD→PFC pathways: 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 a 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[45]

$$\tau \dot{\mathbf{x}}^{\mathrm{PFC}} = -\mathbf{x}^{\mathrm{PFC}} + \mathbf{W}^{\mathrm{rec}}\mathbf{r}^{\mathrm{PFC}} + \mathbf{W}^{\mathrm{MD}\to\mathrm{PFC}}\mathbf{r}^{\mathrm{MD}} + \mathbf{W}^{\mathrm{in}}\mathbf{u} + \sigma\boldsymbol{\xi} \tag{4}$$

And the MD dynamics were controlled by corticothalamic weights:

$$\tau \dot{\mathbf{x}}^{\mathrm{MD}} = -\mathbf{x}^{\mathrm{MD}} + \mathbf{W}^{\mathrm{PFC}\to\mathrm{MD}}\mathbf{r}^{\mathrm{PFC}} + \mathbf{V}^{\mathrm{in}}\mathbf{u} + \sigma\boldsymbol{\xi} \tag{5}$$

where $\mathbf{V}^{\mathrm{in}}$ denotes an $N_{\mathrm{MD}} \times N_{\mathrm{in}}$ matrix. Together, we can rewrite Eqs. (4) and (5) with a unified equation:

$$\begin{aligned}\tau\begin{pmatrix}\dot{\mathbf{x}}^{\mathrm{PFC}} \\ \dot{\mathbf{x}}^{\mathrm{MD}}\end{pmatrix} &= -\begin{pmatrix}\mathbf{x}^{\mathrm{PFC}} \\ \mathbf{x}^{\mathrm{MD}}\end{pmatrix} \\ &+ \begin{bmatrix}\mathbf{W}^{\mathrm{rec}} & \mathbf{W}^{\mathrm{MD}\to\mathrm{PFC}} \\ \mathbf{W}^{\mathrm{PFC}\to\mathrm{MD}} & 0\end{bmatrix}\begin{pmatrix}\mathbf{r}^{\mathrm{PFC}} \\ \mathbf{r}^{\mathrm{MD}}\end{pmatrix} \\ &+ \begin{bmatrix}\mathbf{W}^{\mathrm{in}} & 0 \\ 0 & \mathbf{V}^{\mathrm{in}}\end{bmatrix}\begin{pmatrix}\mathbf{u} \\ \mathbf{u}\end{pmatrix} + \sigma\boldsymbol{\xi}\end{aligned} \tag{6}$$

The output setup was identical to the PFC-alone model, where the readout only depended on the PFC units but not MD units. We have also considered a scenario where there is no direct sensory input $\mathbf{u}$ in Eq. (5), then Eq. (6) can be further simplified. Since the results were qualitatively similar, we have only presented the results in the general case.

Training procedure

We trained the PFC-alone and PFC-MD models by supervised learning. The total number of units in the PFC-MD model was 256 ($N_{\mathrm{PFC}} + N_{\mathrm{MD}} = 256$), whereas the matching number of units in the PFC-alone model was 256 ($N_{\mathrm{PFC}} = 256$). The two contexts (CAC and RDM) were learned simultaneously and their trials were interleaved. The CAC or RDM context 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. The model was trained by minimizing a loss function using back-propagation through time. The loss function $\mathcal{L}$ is computed by mean-squared-error averaged over N_{trials} trials:

$$\mathcal{L} = \frac{1}{N_{\mathrm{trials}}}\sum_{n=1}^{N_{\mathrm{trials}}}\sum_{i}^{N_{\mathrm{out}}}\mathbf{m}_i(\mathbf{z}_i - \hat{\mathbf{z}}_i)^2 \tag{7}$$

where $\mathbf{z}_i$ denotes the i-th target output, and $\hat{\mathbf{z}}_i$ denotes the corresponding estimated network output, $\mathbf{m}_i$ denotes the error mask, which is a non-negative matrix that determines whether the squared error at time point t and output units i need to be computed. In the response epoch, $\mathbf{m}_i = 0$, in the other task epoch, $\mathbf{m}_i = 1$. We applied L_2 regularization in the loss function with a default regularization parameter 0.1. The model parameters were optimized using Adam optimizer, with default configuration of hyperparameters (learning rate parameter 0.0005; weight decay regularization parameter 0.1). The optimization procedure continued until the desired accuracy level of 0.98 was reached. The connection matrices $\{\mathbf{W}^{\mathrm{rec}}, \mathbf{W}^{\mathrm{MD}\to\mathrm{PFC}}, \mathbf{W}^{\mathrm{PFC}\to\mathrm{MD}}\}$ were initialized using Kaiming initialization with the parameter $a = 6$, which was designed to solve the vanishing and exploding gradient problems in practice[82]. The input weights were initialized using a uniform distribution between −1 and 1, and the output weights were initialized with a Gaussian distribution of mean 0 and standard deviation $0.4/\sqrt{256}$.

In total, we trained more than 20 independent PFC-MD networks (and an equally matched number of PFC-alone networks), and selected their respective optimal representative models to present the results (by jointly considering their task performance, generalization, and unit tunings). In testing, we varied the degree of cue uncertainty and simulated 10 independent random test trials per condition to compute the psychometric curve. In the CAC context, we changed the ratio of LP/HP pulses for different cue congruence, and varied the timing and order of the LP and HP pulses for a fixed uncertainty level (Fig. 1g). In the RDM context, we varied the percentage of dots moving rightward or leftward for different cue coherence, and selected different subsets of dots in time (Fig. 1h). We also observed that a few task-optimized PFC-MD and PFC-alone

outlier models did not show good generalization across all cue conditions (e.g., making one type of rule-specific error more frequently or showing specific choice biases); these outlier models have been excluded in the analyses.

Switching cue-to-rule transformation in the CAC task

Rule switching is a hallmark of behavioral flexibility. To introduce cue-to-rule mapping uncertainty, we considered a switching context task (*Context 1*→*Context 2*→ *Context 1'*) using the simulated LP/HP cues in the CAC task (with 800-ms cueing period and 600-ms delay period). The PFC-MD model remained similar (Fig. 6a), except that the input dimensionality was modified accordingly because the RDM task-related input was removed. In *Context 1*, more HP than LP pulses (#HP > #LP) would encode one rule: attend-to-audition. In *Context 2*, the mapping was reversed: more LP pulses (#LP > #HP) would encode the other rule: attend-to-vision (Fig. 7a). The cue and sensory inputs were identical to the previous setup. At the first stage of learning *Context 1*, we trained the PFC-MD model with batch learning to estimate all recurrent intracortical, thalamocortical, and corticothalamic connections. We assumed that the intracortical weights were relatively stable or adapted at a slow timescale, and the thalamocortical/corticothalamic weights were adapted at a faster timescale (such as in a trial-by-trial manner). To dissect the contribution of the MD during rule switching, we further assumed that the recurrent prefrontal connections remained unchanged. Therefore, at the second stage for learning the context switching, we froze intracortical connection weights $\mathbf{W}^{\mathrm{rec}}$, and adapted the other model parameters $\{\mathbf{W}^{\mathrm{MD}\to\mathrm{PFC}}, \mathbf{W}^{\mathrm{PFC}\to\mathrm{MD}}, \mathbf{W}^{\mathrm{in}}, \mathbf{V}^{\mathrm{in}}, \mathbf{W}^{\mathrm{out}}\}$ on a trial-by-trial basis. After each trial, the correct or error feedback was used and backpropagated to sequentially update the model parameters. We further learned *Context 2* under the new cue-to-rule transformation, and then relearned the *Context 1'* based on the original rule. To speed up computation, we used a smaller size of network, with $N_{\mathrm{PFC}} = 256$ and $N_{\mathrm{MD}} = 30$ ($N_{\mathrm{MD1}} = 18$, $N_{\mathrm{MD2}} = 12$). Additionally, we used a gradually annealing learning rule during online learning; that is, the initial learning rate was 0.0005 and then gradually reduced to 0.0001 after the accuracy reached 0.8.

Computer simulations in a schizophrenia mouse model

We conducted computer simulations to reduce prefrontal cortical PV inhibition in the PFC-MD model to mimic one of known changes in schizophrenia. In the CAC context switching task, we first trained the PFC-MD model to learn the task in *Context 1* (control model), and then manipulated E/I imbalance on the trained model (imbalanced model). We compared their task performances in the presence of noise corruption during working memory. We further activated MD subpopulation and reexamined the impact on task performance. Next, during context switching, we retrained both the control model and imbalanced model to learn *Context 2* and computed their relative switch time for comparison. We further activated MD subpopulation and reexamined the impact on the switch time. Finally, we repeated computer simulations in a cross-modal cueing context-switching task setting (e.g., CAC→RDM).

Behavioral analysis and psychometric curve

The model's performance and generalization were 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: $\max\left\{\frac{\#\mathrm{LP\ pulses}}{\#\mathrm{LP\ pulses}+\#\mathrm{HP\ pulses}}, \frac{\#\mathrm{HP\ pulses}}{\#\mathrm{LP\ pulses}+\#\mathrm{HP\ pulses}}\right\}$, 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 RDM (with absolute coherence of 1 indicating highest 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)[25]. In testing, we generated 10 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 smooth psychometric curve function using a logistic function (Fig. 1d). The logistic function was characterized by four parameters {*A, B, C, D*}:

$$f(x) = \frac{A-D}{1+\left(\frac{x}{C}\right)^{B}} + D \tag{8}$$

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 and only assessed in testing. To quantify the performance, we computed the AUC. A larger AUC value indicates a good model performance. To compare the AUC between the PFC-MD and PFC-alone models, we constructed the ensemble mean ± SD of two psychometric curves and derived the corresponding AUC samples; from which we calculated the *p*-value based on the Mann-Whitney rank-sum test.

Reaction time analysis

The model was trained to make a choice during the 200-ms choice period. We set an empirical threshold (such as 0.9) and marked the first moment that the model output crossed above the threshold as the shortest reaction time. We reported the relative reaction time (dimensionless) compared to the baseline model (reaction time 1).

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 density). We excluded the units with very low firing rates (lower 5–10 percentiles of population) in tuning curve analysis. 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). Across trials, mean ± SD firing rates of units were calculated, the rank-sum test was used to compare the statistics to determine significant modulation ($p < 0.05$). We computed the mean firing rate (FR) of each single unit based on its simulated activity averaged over time and across trials. If a unit was said to be tuned to a discrete task variable (e.g., rule or context) between two conditions, it must meet an empirical criterion: $|\text{mean FR}_{\text{task condition1}} - \text{mean FR}_{\text{task condition2}}| > 0.25 \times \text{mean overall FR}$.

For a continuous task variable (e.g., cue uncertainty or cue density), we assumed a linear regression model between the mean FR of a unit during the cueing period and the cue uncertainty (e.g., coherence or congruence)

$$\text{Mean FR} = \alpha + \beta \cdot \text{cue uncertainty level} \tag{9}$$

where α and β denote the regression coefficients. We fit the model prediction between mean FR and cue uncertainty level (Supplementary Fig. 3a), and further characterized the percentage of explained variance (EV) for both CAC and RDM contexts. The unit is strongly modulated with cueing uncertainty if it simultaneously meets the strong modulation criterion (e.g., %EV > 80%) for both contexts (Supplementary Fig. 3b). A similar procedure was also applied to the cue sparsity tuning analysis (Supplementary Fig. 3c, d).

Excitatory/inhibition (E/I) input

For each PFC excitatory unit, we computed the total excitatory input and inhibitory input

$$I_j^{\text{exc}} = \sum_{i\in \text{Exc}} W_{ij}^{\text{rec}} r_i + \sum_{k\in \text{MD}} W_{kj}^{\text{MD}\to\text{PFC}} r_k \tag{10}$$

$$I_j^{\text{inh}} = \sum_{i\in \text{Inh}} W_{ij}^{\text{rec}} r_i \tag{11}$$

where the total excitatory input consisted of excitatory units of both cortical and thalamic contributions, and the total inhibitory input consisted of input from cortical inhibitory units. If the synaptic strengths from cortical PV inhibitory neurons to cortical excitatory neurons are reduced, we will have reduced cortical inhibition (e.g., Figs. 5b and 9a) and cortical E/I imbalance.

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 (Fig. 2i).

Population decoding analysis

We employed a binary support vector machine (SVM) classifier to achieve trial-by-trial classification to decode the 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. 2j, k).

Dimensionality reduction and subspace 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, and uncertainty), we further defined the task-relevant state space using a published method[28]. 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: 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. 3a–f). From a neural trajectory $\mathbf{p}(t)$, we further defined the time-varying neural velocity by computing the change rate in neural trajectory: $v(t) = ||\dot{\mathbf{p}}(t)||$ (Fig. 3g, h).

We also computed the PCA subspaces separately for the PFC and MD populations derived from the trained PFC-MD model (Fig. 8c). We adopted an established method to compute the subspace angles ("principal angle")[83]. Briefly, let C_1 and C_2 be two d-dimensional PC projection matrices derived from respective higher-dimensional data. Note that these two matrices can be derived either from the PFC and MD populations of the same network (Fig. 3m), or from the same population of two different networks (e.g., MD-MD or PFC-PFC; Fig. 8i, j). The two d-dimensional subspaces have a total of d angles, denoted by $\{\theta_1, \ldots, \theta_d\}$. The angle θ_1 is defined as the largest angle defined by two vectors $\mathbf{u}_1$ and $\mathbf{v}_1$ within the range of $\mathbf{C}_1$ and $\mathbf{C}_2$, respectively. Similarly, the angle θ_2 is defined as the largest angle defined by vectors $\mathbf{u}_2$ and $\mathbf{v}_2$ constrained to be orthogonal to $\mathbf{u}_1$ and $\mathbf{v}_1$, respectively. This definition is continued iteratively until θ_d. In practice, these angles can be obtained as the cosine of the singular values of the inner product of two matrices: $\mathbf{C}_1^T\mathbf{C}_2$. For the principal angle, we chose the largest angle computed from the principal subspaces. An angle of 90 degree indicates the orthogonality of two subspaces.

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 led to an increase or decrease in their phasic firing activities, respectively (Fig. 4b–d).

Fixed-point analysis of recurrent network dynamics

Similar to our previous analysis[25,26], we identified fixed-points or slow points of the task-optimized PFC-MD model by numerically solving the optimization problem (https://github.com/mattgolub/fixed-point-finder)

$$\begin{aligned} &\min_{\tilde{\mathbf{x}}} q(\tilde{\mathbf{x}}), \text{where } q(\tilde{\mathbf{x}}) = \left|\left| -\tilde{\mathbf{x}} + \mathbf{W}^{\text{eff}}\boldsymbol{\phi}(\tilde{\mathbf{x}}) + [\mathbf{W}^{\text{in}}, \mathbf{V}^{\text{in}}] \cdot \mathbf{u} \right|\right|^2 \text{or} \\ &\min_{\mathbf{x}} q(\mathbf{x}^{\text{PFC}}), \text{where } q(\mathbf{x}^{\text{PFC}}) = \left|\left| -\mathbf{x}^{\text{PFC}} + \mathbf{W}^{\text{rec}}\boldsymbol{\phi}(\mathbf{x}^{\text{PFC}}) + \mathbf{W}^{\text{in}}\mathbf{u} \right|\right|^2 \end{aligned} \tag{12}$$

where $\tilde{\mathbf{x}} = [\mathbf{x}^{\text{PFC}}, \mathbf{x}^{\text{MD}}]$. We collected a set of fixed points by randomly initializing the network. Once the numerical optimization was completed, we applied PCA to visualize the fixed points in three-dimensional PC subspace (Fig. 3i).

Eigenvalue analysis and time constant estimation

A square matrix $\mathbf{A}$ is non-normal if it satisfies $\mathbf{A}\mathbf{A}^T \neq \mathbf{A}^T\mathbf{A}$, and the degree of the matrix non-normality, is measured by the Henrici metric $||\mathbf{A}\mathbf{A}^T - \mathbf{A}^T\mathbf{A}||_F$ [84]. Because of Dale's principle, our inferred weight matrix is non-normal. For the non-normal matrix $\mathbf{W}^{\text{eff}} = \begin{bmatrix} \mathbf{W}^{\text{rec}} & \mathbf{W}^{\text{MD}\to\text{PFC}} \\ \mathbf{W}^{\text{PFC}\to\text{MD}} & 0 \end{bmatrix}$ (in the PFC-MD model) or $\mathbf{W}^{\text{rec}}$ (in the PFC-alone or MD-lesioned model), we conducted eigenvalue analysis on the overall weight matrix to analyze the nonlinear dynamics with linear approximation[25,85], where the maximum of eigenvalues characterizes the long-term behavior for the speed of network's steady state decaying to zero[86]. We further we adapted a published method to estimate the intrinsic time constant of dynamical systems[87], for both PFC-MD and PFC-alone networks (Fig. 3k):

$$\tau_{\text{intrinsic}} = \max_k \left(-\frac{\Delta t}{\log|\lambda_k|} \right) \tag{13}$$

where $|\lambda_k|$ denotes the absolute value of the k-th real or complex-valued eigenvalue. Since a large value of time constant is beneficial for temporal summation, a greater time constant can help information

integration (during the cueing period) and working memory maintenance (during the delay period).

Mathematical analysis of cue-to-rule transformation

In light of Eq. (4), we assumed that the PFC-MD network converged to a steady state (i.e., the velocity of the state, $|\dot{\mathbf{x}}^{\mathrm{PFC}}|$ approached to zeros towards the end of task delay period, see Fig. 3g, h). To gain some mathematical intuition, we assumed a linear RNN dynamics and let $\dot{\mathbf{x}}^{\mathrm{PFC}} = \boldsymbol{0}$. We approximated the Jacobian of the dynamics with the first-order difference equation or numerically solved the equation via the Euler method. Upon rearranging the terms, we approximated the steady-state solution $\mathbf{x}_{\mathrm{ss}}^{\mathrm{PFC}}$

$$\begin{bmatrix} \mathbf{x}_{\mathrm{ss}}^{\mathrm{PFC}} \\ \mathbf{x}_{\mathrm{ss}}^{\mathrm{MD}} \end{bmatrix} = \left(\mathbf{I} - \mathbf{W}^{\mathrm{eff}}\right)^{-1} \begin{bmatrix} \mathbf{W}^{\mathrm{in}} \\ \mathbf{V}^{\mathrm{in}} \end{bmatrix} \mathbf{u} \text{ or } \mathbf{x}_{\mathrm{ss}}^{\mathrm{PFC}} = \left(\mathbf{I} - \mathbf{W}^{\mathrm{rec}}\right)^{-1} \mathbf{W}^{\mathrm{in}} \mathbf{u} \quad (14)$$

where the matrix $\Psi = \left(\mathbf{I} - \mathbf{W}^{\mathrm{eff}}\right)^{-1} [\mathbf{W}^{\mathrm{in}}, \mathbf{V}^{\mathrm{in}}]^T$ or $\Psi = \left(\mathbf{I} - \mathbf{W}^{\mathrm{rec}}\right)^{-1} \mathbf{W}^{\mathrm{in}}$ denotes the cue-to-rule transformation. Let Ψ_{context1} and $\Psi_{\mathrm{context1}}{}'$ denote the transformation matrices under *Context 1* and *Context 1′*, respectively; if the cue-to-rule tunings of PFC units are reversibly remapped, we will expect that these two transformation matrices are similar (Supplementary Fig. 7d). Note that if the input $\mathbf{u}$ is univariate, Ψ will reduce to a vector. Note that the MD activity and cueing input jointly influence the prefrontal neural dynamics $\dot{\mathbf{x}}^{\mathrm{PFC}}$. In the two-dimensional PCA subspace, the MD output can shift the original prefrontal dynamics and modify the vector field (Fig. 8f).

Feedforward thalamic control

From a control scenario, the MD thalamus can be viewed as a controller that operates on the cortical plant, which enables the context switch. This is reminiscent of the role of thalamus in optimal anticipatory control in motor preparation preparatory[53,54]. To gain computational insight, we assume that the controlled cortical state in *Context 2* at a fixed point (denoted by $\mathbf{x}^{*,\mathrm{PFC,\,context2}}$) was a rotated version of state in *Context 1* at the fixed point (denoted by $\mathbf{x}^{*,\mathrm{PFC,\,context1}}$; see a two-dimensional cartoon illustration in Fig. 8f) during the steady state (where $\dot{\mathbf{x}}^{\mathrm{PFC}} = \boldsymbol{0}$ and external input $\mathbf{u} = \mathbf{0}$). From Eq. (4), the naïve feedforward strategy to achieve the cortical target state is to set the net MD-to-PFC input as[53]

$$\mathbf{W}^{\mathrm{MD}\to\mathrm{PFC}} \mathbf{r}^{\mathrm{MD}} = \left(\mathbf{x}^{*,\mathrm{PFC,\,context2}} - \mathbf{W}^{\mathrm{rec}} \phi\left(\mathbf{x}^{*,\mathrm{PFC,\,context2}}\right)\right) \quad (15)$$

Provided that the intracortical connection $\mathbf{W}^{\mathrm{rec}}$ is intact, then a one-shot feedforward control strategy to learn the context switch (*Context 1→Context 2*) is to set $\mathbf{r}^{\mathrm{MD}}$ as

$$\mathbf{r}^{\mathrm{MD}} = \left(\mathbf{W}^{\mathrm{MD}\to\mathrm{PFC}}\right)^{-1} \left(\mathbf{x}^{*,\mathrm{PFC,\,context2}} - \mathbf{W}^{\mathrm{rec}} \phi\left(\mathbf{x}^{*,\mathrm{PFC,\,context2}}\right)\right) \quad (16)$$

Alternatively, $\mathbf{r}^{\mathrm{MD}}$ can be updated by modifying the thalamocortical weights $\mathbf{W}^{\mathrm{MD}\to\mathrm{PFC,\,context1}}$ and the corticothalamic weights $\mathbf{W}^{\mathrm{PFC}\to\mathrm{MD,\,context1}}$ under *Context 1*. This makes it computationally appealing since the MD thalamus has a feedforward network structure. The speed of switching between two contexts depends on the degree-of-freedom of the PFC-MD system and the initial cortical state.

Impact of noise on working memory

To investigate the impact of noise sensitivity on working memory, we injected a higher level of additive Gaussian noise into the PFC dynamics during the task delay period (Eq. (1)). We further quantified the task performance in five conditions (Fig. 9a): (i) the trained PFC-MD model under the noise-less condition (control); (ii) the trained PFC-MD model under the noise condition; (iii) the E/I imbalanced model with reduced cortical inhibition under the noiseless condition; (iv) the E/I imbalanced model with reduced cortical inhibition under the noise condition; (v) MD activation combined with condition (iv).

Training E/I imbalanced models for context switching

To investigate the impact of E/I imbalance (due to reduced PV inhibition) on switching deficit, we started with a pretrained PFC-MD model that completed training on *Context 1*. We further modified PV→PFC connectivity by setting a small percentage of connections to zeros, so that the PFC network with the new intra-cortical connectivity $\mathbf{W}^{\mathrm{rec}}\_\mathrm{new}$ had cortical E/I imbalance or reduced inhibition. To learn context switching for the E/I imbalanced PFC-MD model, we also kept some model parameters $\{\mathbf{W}^{\mathrm{rec}}, \mathbf{W}^{\mathrm{in}}\}$ unchanged (similar to the setting shown in Fig. 8d) and focused on thalamocortical and corticothalamic plasticity $\{\mathbf{W}^{\mathrm{MD}\to\mathrm{PFC}}, \mathbf{W}^{\mathrm{PFC}\to\mathrm{MD}}\}$. To quantify the switching deficit of the E/I imbalanced model, we compared the switching time to successfully learn the new context relative to the switching time required by the control model (Fig. 9b, c).

We tested the E/I imbalanced PFC-MD model in two context switching conditions. The first type was based on the auditory cueing only (LP/HP) but reversal rules (Fig. 7a). Specifically, we trained the E/I imbalanced model to learn the *Context 1→Context 2* switch, similar to the description of cue-to-rule remapping in the CAC context. The second type was based on cross-modal cueing, using both auditory cueing (CAC as *Context 1*) and visual cueing (RDM as *Context 2*). In this case, we also modified the task conditions (e.g., cue uncertainty, cue and delay duration) accordingly to match the mouse experiment (Fig. 9g).

To mimic the effect of MD activation during context switching, we modified the update of PFC dynamics by adding a scalar positive gain κ to compensate for reduced cortical inhibition introduced by the new recurrent weight connectivity $\mathbf{W}^{\mathrm{rec}}\_\mathrm{new}$

$$\tau \dot{\mathbf{x}}^{\mathrm{PFC}} = -\mathbf{x}^{\mathrm{PFC}} + \mathbf{W}^{\mathrm{rec}}\_\mathrm{new} \mathbf{r}^{\mathrm{PFC}} + \mathbf{W}^{\mathrm{MD}\to\mathrm{PFC}} (\mathbf{r}^{\mathrm{MD}} + \kappa) + \mathbf{W}^{\mathrm{in}} \mathbf{u} + \sigma \boldsymbol{\xi} \quad (17)$$

whereas the MD dynamics remained unchanged. The choice of κ depends on the degree of E/I imbalance. In our computer simulations, the value of κ was empirically chosen to be 0.5–0.8, which roughly matched the population firing rate in $\mathbf{r}^{\mathrm{MD}}$.

In all computer simulations to achieve successful context switching, the PFC-MD models were trained until reaching 100% performance accuracy.

Animal experiments

A total of 10 mice (50% female) were used in the current behavioral study. Adult C57Bl/6 (WT) or Df(h22q11)/+ (22Q11DS) mice aged 8–12 weeks old were purchased from Taconic Biosciences. All mice were kept in rooms with controlled temperature and ventilation (20–22 °C; 40–60% humidity) on a constant 12-h light–dark cycle. All animals used for histology were group housed and kept on a 12 h light-dark cycle, with ad libitum access to food and water. Behavioral animals were single housed, and food restricted to 90% of their free feeding body weight. All animal experiments were performed according to the guidelines of the US National Institutes of Health (NIH) and with approval from Tufts University IACUC committee.

Surgery. Mice were initially anesthetized in an induction chamber with a continuous supply of oxygen and 5% isoflurane. They were then placed on a heating pad within a stereotaxic frame (Kopf Instruments, Tujunga, California). During the surgery, anesthesia was maintained with 1–2% isoflurane delivered via a nose cone at a rate of 1 L/min. Analgesia was provided by subcutaneous injections of slow-release Buprenorphine (0.5 mg/mL) and Meloxicam (5 mg/mL). A midline incision was made on the scalp, which was then retracted, and a craniotomy was performed over the target region. The head was leveled, and a small burr hole was drilled over each target area using

coordinates from the established mouse brain atlas. The coordinates (in mm from Bregma) were the following: PFC: AP 2.6, ML ± 0.3, DV−1.9 and MD: AP−1.2, ML ± 0.6, DV−3.0 (from brain surface). For transsynaptic labeling of the MD output to the PFC, 90 nL of AAV1-hSyn-Cre-WPRE-hGH (Addgene) was injected unilaterally into the MD, and 400 nL of AAV2-hSyn-DIO-mCherry-WPRE (Addgene) was injected into the PFC ipsilateral to the MD injection. After 3–4 weeks, the mice were perfused, and their brains were extracted for visualization. Viruses were injected through a glass micropipette (Drummond Scientific) using a quintessential stereotactic injector (QSI, Stoelting) at a flow rate of 50 nl/min and given 10 min to spread after injection.

In behavioral experiments involving optogenetic manipulation, we used the following procedure for the initial surgical steps up to the virus injection. Briefly, after drilling the burr holes, 250 nL of AAV2-CaMKIIa-hChR2 (C128S/D156A)-mCherry (UNC Vector Core, Stabilized Step Function Opsin; SSFO) was bilaterally injected into the MD at coordinates (in mm from Bregma): AP −1.2, ML ± 0.6, and DV −3.0 (from brain surface) using a glass micropipette. Subsequently, two optic fibers were inserted bilaterally into the MD using the same burr holes as the virus injections, but at a DV coordinate 500 μm above the virus injection site. Three stainless steel screws were implanted around the optic fibers for mechanical stability, and the entire assembly was secured to the skull with dental cement (Parkell C&B Metabond). The skin was then sutured around the implant, and the animal was allowed to recover for 3 weeks to ensure complete healing and virus expression.

Behavioral training. Animal training was briefly described below[7]. First, 10 μl of evaporated milk (reward) was delivered randomly to each reward port for habituation to the setup. Reward availability was signaled by opening of the response port doors. Repeated illumination of the LED on the side spatially congruent to the rewarded response port was used to establish an association of the visual target to the reward while similar presentation of the auditory tone cloud target was used to build the association with the auditory target. Each trial ended 15 s after reward collection, and a new trial began 5 s later. The auditory and visual targets were initially presented for 500 ms and gradually reduced to 100 ms over multiple sessions. Once the mice were proficient at collecting rewards (30 rewards per hour), they were trained to nose poke to receive a reward, with all other parameters remaining constant and incorrect pokes having no negative consequence. By the end of this first training phase, the mice were able to nose poke to collect 20 rewards per 30-min session.

In the second stage, the mice were trained to self-initiate trials by briefly (50 ms) nose poking in the initiation port to trigger target stimulus presentation and render reward ports accessible. The trial rule ("Attend to Vision" or "Attend to Audition") was indicated by 4–8 kHz LP-filtered white noise (vision) or 12–40 kHz HP-filtered white noise cues (auditory cueing). Stimuli were presented in blocks of six trials consisting of single-modality stimulus presentation (no conflict). An incorrect response rendered the response port inaccessible. Rewards were available for 15 s after correct nose-poking, followed by a 5-s inter-trial interval. Incorrect poking was punished with a 30-s timeout, during which the mice could not initiate new trials. Trial availability was signaled by broadband white noise, which was immediately interrupted upon a nose poke.

In the third stage, conflict trials were introduced, where auditory and visual targets were co-presented, indicating rewards at opposing response ports. Trial types were presented in blocks of six visual target correct or auditory target correct trials, gradually reduced to blocks of three. The required duration for maintaining a continuous nose poke at the initiation port was gradually increased over sessions until it reached 0.8 s. The mice were trained on this stage until they responded to the correct cue on the first trial upon a block switch for 70% of the switches over three successive sessions.

Fourth, trial availability and task rules were dissociated. The broadband white noise indicating trial availability was separated from the auditory cue presentation indicating the rule by 100 ms. This was followed by a delay period (550 ms) before target stimuli presentation. All block structures were removed, and trial types were randomized. Mice were trained on this discrete cueing version of the task until mean performance plateaued and remained stable over four to five consecutive sessions (mean accuracy of 70 ± 3% correct). On a subset of trials, the two targets were shown on congruent sides to ensure that the mice did not develop a pro-anti strategy for a single cue.

Once the mice were fully familiarized with the main structure of the task and achieved consistent performance on the fourth stage of training, they were exposed to the visual cueing condition. To build an association of the visual cues to their respective rules, the animals were initially paired with the congruent auditory cues (LP with green LED and HP with UV LED) for a block of 50 trials. The volume of the auditory cue decayed over the course of trials, with full volume for the first 10 trials and up to 1/5 of the volume for the last few trials. The volume was reduced after a mouse made two consecutive correct responses, indicating understanding of the task. After 50 trials, the mice progressed to a visual-only block of 50 trials, where no auditory cues were played. Each session alternated between the two different blocks until an animal reached a performance criterion of over 60% correct for three consecutive sessions on the visual cue-only block. In subsequent sessions, the association blocks were removed, and alternating blocks of auditory cue-only (cueing Context 1) or visual cue-only (cueing Context 2) were presented until the animals performed equally on both blocks (over 60% over three sessions). Finally, the block switch was tied to performance, requiring the mice to reach a performance criterion of 80% correct over a moving window of 10 trials to trigger a block switch. Once the mice could successfully switch across two blocks per session for three consecutive sessions, they were considered experts and ready for testing.

Behavioral testing. After recovering from optic fiber implantation, the mice were retrained on the final stage of the task with performance-dependent switching between auditory (high-pitch/low-pitch; cueing Context 1) and visual (UV/Green; cueing Context 2) blocks until criterion (at least two block switches for three consecutive sessions). The identity of the first block (visual or auditory) was pseudorandomized daily. No significant difference was observed in switching from visual to auditory blocks versus auditory to visual blocks.

For optogenetic stimulation of the MD using SSFO, the first block switch was always non-manipulated. From the second block switch onwards, each switch had a 50% probability of being a stimulated switch (Laser ON vs OFF). In a stimulated block, the MD was activated (650 ms of SSFO activation) for the entire duration of the cue and delay periods of each trial for the first five trials post-switch. Switch latency was measured as the number of trials it took for an animal to achieve a post-switch performance of 70% correct over ten consecutive trials for each block switch. Switch latency was analyzed by an experimenter blinded to the manipulation (Laser ON vs. OFF).

Histology and immunohistochemistry

Mice were transcardially perfused with 30 ml of 0.1 M phosphate-buffered saline (PBS) followed by 20 ml of 4% paraformaldehyde in PBS. Brains were post-fixed in the same fixative overnight at 4 °C, then cryoprotected in 30% sucrose in PBS for 24 h. Serial 50-μm-thick coronal sections from the PFC region were prepared using a Thermo HM550 cryotome. For immunostaining PV-expressing neurons, sections were permeabilized and blocked in 10% bovine serum albumin (BSA, Sigma-Millipore) in PBS with 0.3% Triton X-100 (PBSTx) for 1 h. Sections were then incubated overnight at 4 °C with a rabbit anti-PV primary antibody (1:1,000, Swant, PV27a) in PBSTx with 3% BSA. After two washes in PBSTx, sections were incubated with an Alexa Fluor 488

donkey anti-rabbit secondary antibody (1:500, Thermo Fisher Scientific, A-21206) for 2 h at room temperature. Finally, sections were washed again, mounted, and prepared for imaging.

Image analysis

Images were acquired on a confocal microscope (LSM 710, Zeiss) with a 20×/0.80 numerical aperture objective (Zeiss). The images were manually overlaid with vectorized outlines from a modified version of the Reference Atlas from the Allen Brain Atlas (Unified Anatomical Atlas) using anatomical landmarks as guides. To determine the distribution of MD outputs to the PFC and their colocalization with PV neurons, the ImageJ *coloc* plugin (NIH) was used, followed by visual validation by the experimenter blinded to the datasets.

Statistics and reproducibility

The animal behavioral data were collected from 21 (laser OFF) and 26 switches collected from 19 independent sessions from 4 mice.

Statistical analyses and tests

All statistical analyses were performed using MATLAB or Python. For the animal behavioral experiment, power analyses were performed using the MATLAB function *sampsizepwr* to determine the number of switches needed to establish an effect. The effect size was established using a Cohen's *d* measure which for the given dataset was calculated to be 0.8975. With a significance value of 0.05 and a power of 0.7, we estimated 18 switches each for Laser ON and Laser OFF conditions to be sufficient and hence collected more than 20 switches for each condition (Laser ON vs Laser OFF) across 4 animals. Error bars are shown as means ± s.e.m. or means ± SD. We used nonparametric two-tailed Wilcoxon signed-rank tests for all paired statistical tests, and Mann–Whitney rank-sum tests for all unpaired statistical tests, respectively. For all tests, we used 0.05 as the minimum level of statistical significance.

Reporting summary

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

Data availability

Source data are provided with this paper. Additional animal experimental data will be made available upon request. Source data are provided with this paper.

Code availability

All software that supports the plots within this paper and other finding of this study are available at https://github.com/Xh-Zhang1/pfc-md and archived in https://doi.org/10.5281/zenodo.14602872.

References

1. Uddin, L. Q. Cognitive and behavioural flexibility: neural mechanisms and clinical considerations. *Nat. Rev. Neurosci.* **22**, 167–179 (2021).
2. 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).
3. Mushtaq, F., Bland, A. R. & Schaefer, A. Uncertainty and cognitive control. *Front. Psychol.* **2**, 249 (2011).
4. 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 (2012).
5. Zhou, Y., Acerbi, L. & Ma, W. J. The role of sensory uncertainty in simple contour integration. *PLoS Comput. Biol.* **16**, e1006308 (2020).
6. Mukerjee, A., Lam, N. H., Wimmer, R. D. & Halassa, M. M. Thalamic circuits for independent control of prefrontal signal and noise. *Nature* **600**, 100–104 (2021).
7. Rikhye, R. V., Gilra, A. & Halassa, M. M. Thalamic regulation of switching between cortical representations enables cognitive flexibility. *Nat. Neurosci.* **21**, 1753–1763 (2018).
8. Diamond, A. Executive functions. *Ann. Rev. Psychol.* **64**, 135–168 (2013).
9. Marton, T. F. et al. Roles of prefrontal cortex and mediodorsal thalamus in task engagement and behavioral flexibility. *J. Neurosci.* **38**, 2569–2578 (2018).
10. Bolkan, S. S. et al. Thalamic projections sustain prefrontal activity during working memory maintenance. *Nat. Neurosci.* **20**, 987–996 (2017).
11. Halassa, M. M. & Kastner, S. Thalamic functions in distributed cognitive control. *Nat. Neurosci.* **20**, 1669–1679 (2017).
12. Wolff, M. & Vann, S. D. The cognitive thalamus as a gateway to mental representations. *J. Neurosci.* **39**, 3–14 (2019).
13. Chakraborty, S., Kolling, N., Walton, M. E. & Mitchell, A. S. Critical role for the mediodorsal thalamus in permitting rapid reward-guided updating in stochastic reward environments. *eLife* **5**, e13588 (2016).
14. Parnaudeau, S., Bolkan, S. S. & Kellendonk, C. The mediodorsal thalamus: an essential partner of prefrontal cortex for cognition. *Biol. Psychiatry* **83**, 648–656 (2018).
15. DeNicola, A. L., Park, M.-Y., Crowe, D. A., MacDonald, A. W. 3rd & Chafee, M. V. 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).
16. Kosciessa, J. Q., Lindenberger, U. & Garrett, D. D. Thalamocortical excitability modulation guides human perception under uncertainty. *Nat. Commun.* **12**, 2430 (2021).
17. Grinband, J., Hirsch, J. & Ferrera, V. P. A neural representation of categorization uncertainty in the human brain. *Neuron* **49**, 757–763 (2006).
18. Schmitt, L. I. et al. Thalamic amplification of cortical connectivity sustains attentional control. *Nature* **545**, 219–223 (2017).
19. Alcaraz, F. et al. Thalamocortical and corticothalamic pathways differentially contribute to goal-directed behaviors in the rat. *eLife* **7**, e32517 (2018).
20. Hertag, L. & Sprekeler, H. Amplifying the redistribution of somato-dendritic inhibition by the interplay of three interneuron types. *PLoS Comput. Biol.* **15**, e1006999 (2019).
21. Keller, A. J. et al. A disinhibitory circuit for contextual modulation in primary visual cortex. *Neuron* **108**, 1181–1193 (2020).
22. Pulvermüller, F. et al. Biological constraints on neural network models of cognitive function. *Nat. Rev. Neurosci.* **22**, 488–502 (2021).
23. Jaramillo, J., Mejias, J. F. & Wang, X.-J. Engagement of pulvino-cortical feedforward and feedback pathways in cognitive computations. *Neuron* **101**, 321–336 (2019).
24. 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).
25. Zhang, X., Liu, S. & Chen, Z. S. A geometric framework for understanding dynamic information integration in context-dependent computation. *iScience* **24**, 102919 (2021).
26. 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).
27. 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.* **15**, 1167–1189 (2023).
28. Mante, V. et al. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).
29. 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).

30. Zheng, W. L., Wu, Z., Hummos, A., Yang, G. R. & Halassa, M. M. Rapid context inference in a thalamocortical model using recurrent neural networks. *Nat. Commun.* **15**, 8275 (2024).
31. Hummos, A., Wang, B. A., Drammis, S., Halassa, M. M. & Pleger, B. Thalamic regulation of frontal interactions in human cognitive flexibility. *PLoS Comput. Biol.* **18**, e1010500 (2022).
32. Pi, H. J. et al. Cortical interneurons that specialize in disinhibitory control. *Nature* **503**, 521–524 (2013).
33. Wood, K. C., Blackwell, J. M. & Geffen, M. N. Cortical inhibitory interneurons control sensory processing. *Curr. Opin. Neurobiol.* **46**, 200–207 (2017).
34. Tao, H. W., Li, Y.-T. & Zhang, L. I. Formation of excitation-inhibition balance: inhibition listens and changes in tune. *Trends Neurosci.* **37**, 528–530 (2014).
35. Ferguson, B. R. & 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).
36. Contractor, A., Ethell, I. M. & Portera-Cailliau, C. Cortical interneurons in autism. *Nat. Neurosci.* **24**, 1648–1659 (2021).
37. Fu, Y. et al. A cortical circuit for gain control by behavioral state. *Cell* **156**, 1139–1152 (2014).
38. Howes, O. D. & Shatalina, E. Integrating the neurodevelopmental and dopamine hypotheses of schizophrenia and the role of cortical excitation-inhibition balance. *Biol. Psychiatry* **92**, 501–513 (2022).
39. Nassar, M. R., Waltz, J. A., Albrecht, M. A., Gold, J. M. & Frank, M. J. All or nothing belief updating in patients with schizophrenia reduces precision and flexibility of beliefs. *Brain* **144**, 1013–1029 (2021).
40. Cascella, N. et al. Deep brain stimulation of the substantia nigra pars reticulata for treatment-resistant schizophrenia: A case report. *Biol. Psychiatry* **90**, e57–e59 (2021).
41. Zhou, T. et al. Enhancement of mediodorsal thalamus rescues aberrant belief dynamics in a novel mouse model of schizophrenia. BioRxiv preprint, https://doi.org/10.1101/2024.01.08.574745 2024.
42. Lam, N. H. et al. Effects of altered excitation-inhibition balance on decision making in a cortical circuit model. *J. Neurosci.* **42**, 1035–1053 (2022).
43. Pakkenberg, B. The volume of the mediodorsal thalamic nucleus in treated and untreated schizophrenics. *Schizophr. Res.* **7**, 95–100 (1992).
44. Pergola, G. et al. The regulatory role the human mediodorsal thalamus. *Trends Cogn. Sci.* **22**, 1011–1025 (2018).
45. Mukherjee, A., Carvalho, F., Eliez, S. & Caroni, P. Long-lasting rescue of network and cognitive dysfunction in a genetic schizophrenia model. *Cell* **178**, 1387–1402 (2019).
46. Wolff, M. & Halassa, M. M. The mediodorsal thalamus in executive control. *Neuron* **112**, 893–908 (2024).
47. Halassa, M. M. & Sherman, S. M. Thalamocortical circuit motifs: a general framework. *Neuron* **103**, 762–770 (2019).
48. Izhikevich, E. M. & Edelman, G. M. Large-scale model of mammalian thalamocortical systems. *Proc. Natl. Acad. Sci. USA* **105**, 3593–3598 (2008).
49. Krishnan, G. P. et al. Cellular and neurochemical basis of sleep stages in the thalamocortical network. *eLife* **5**, e18607 (2016).
50. 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).
51. Levenstein, D. et al. On the role of theory and modeling in neuroscience. *J. Neurosci.* **43**, 1074–1088 (2023).
52. Anastasiades, P. G., Collins, D. P. & Carter, A. G. Mediodorsal and ventromedial thalamus engage distinct L1 circuits in the prefrontal cortex. *Neuron* **109**, 314–330 (2021).
53. Roy, D. S., Zhang, Y., Halassa, M. & Feng, G. Thalamic subnetworks as units of function. *Nat. Neurosci.* **25**, 140–153 (2022).
54. Wang, M. B. & Halassa, M. M. Thalamocortical contribution to flexible learning in neural systems. *Net. Neurosci.* **6**, 980–997 (2022).
55. Logiaco, L., Abbott, L. F. & Escola, S. Thalamic control of cortical dynamics in a model of flexible motor sequencing. *Cell Rep.* **35**, 109090 (2021).
56. Kao, T. C., Sadabadi, M. S. & Hennequin, G. Optimal anticipatory control as a theory of motor preparation: A thalamo-cortical circuit model. *Neuron* **109**, 1567–1581 (2021).
57. Lakshminarasimhan, K. J. et al. Specific connectivity optimizes learning in thalamocortical circuits. *Cell. Rep.* **43**, 114059 (2024).
58. Hattori, R. et al. Functions and dysfunctions of neocortical inhibitory neuron subtypes. *Nat. Neurosci.* **20**, 1199–1208 (2017).
59. Guet-McCreight, A., Skinner, F. K. & Topolnik, L. Common principles in functional organization of VIP/Calretinin cell-driven disinhibitory circuits across cortical areas. *Front. Neural Circuits* **14**, 32 (2020).
60. Delevich, K., Tucciarone, J., Huang, Z. J. & Li, B. The mediodorsal thalamus drives feedforward inhibition in the anterior cingulate cortex via parvabumin interneurons. *J. Neurosci.* **35**, 5743–5753 (2015).
61. 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).
62. Sohal, V. S., Zhang, F., Yizhar, O. & Deisseroth, K. Parvalbumin neurons and gamma rhythms enhance cortical circuit performance. *Nature* **459**, 698–702 (2009).
63. Lewis, D., Hashimoto, T. & Volk, D. Cortical inhibitory neurons and schizophrenia. *Nat. Rev. Neurosci.* **6**, 312–324 (2005).
64. Dienel, S. J. & Lewis, D. A. Alterations in cortical interneurons and cognitive function in schizophrenia. *Neurobiol. Dis.* **131**, 104208 (2019).
65. Rikhye, R. V., Wimmer, R. D. & Halassa, M. M. Toward an integrative theory of thalamic function. *Ann. Rev. Neurosci.* **41**, 163–183 (2018).
66. Vander Weele, C. M. et al. Dopamine enhances signal-to-noise ratio in cortical-brainstem encoding of aversive stimuli. *Nature* **563**, 397–401 (2018).
67. Mininni, C. J. et al. Putative dopamine neurons in the ventral tegmental area enhance information coding in the prefrontal cortex. *Sci. Rep.* **8**, 11740 (2018).
68. Garcia-Cabezas, M. A. et al. Dopamine innervation in the thalamus: monkey versus rat. *Cereb. Cortex* **19**, 424–434 (2009).
69. Winterer, G. & Weinberger, D. R. Genes, dopamine and cortical signal-to-noise ratio in schizophrenia. *Trends Neurosci.* **27**, 683–690 (2004).
70. Ferguson, B. R. & 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).
71. Nakajima, M., Schmitt, L. I. & Halassa, M. M. Prefrontal cortex regulates sensory filtering through a basal ganglia-to-thalamus pathway. *Neuron* **103**, 445–458 (2019).
72. Scott, D. N., Mukherjee, A., Nassar, M. R. & Halassa, M. M. Thalamocortical architectures for flexible cognition and efficient learning. *Trends Cogn. Sci.* **28**, 739–756 (2023).
73. Lam, N. et al. Prefrontal transthalamic processing of uncertainty drives cognitive flexibility. *Nature* **637**, 127–136 (2025).
74. Kohn, A. et al. Principles of corticocortical communication: proposed schemes and design considerations. *Trends Neurosci.* **43**, 725–737 (2020).
75. Murray, J. D. & Anticevic, A. Toward understanding thalamocortical dysfunction in schizophrenia through computational models of neural circuit dynamics. *Schizophr. Res.* **180**, 70–77 (2017).
76. Mukherjee, A. & Halassa, M. M. The associative thalamus: a switchboard for cortical operations and a promising target for schizophrenia. *Neuroscientist* **30**, 132–147 (2024).

77. Anticevic, A. & Halassa, M. M. The thalamus in psychosis spectrum disorder. *Front. Neurosci.* **17**, 1163600 (2023).
78. Huang, A. S. et al. A prefrontal thalamocortical readout for conflict-related executive dysfunction in schizophrenia. *Cell Rep. Med.* **5**, 101802 (2024).
79. Ouhaz, Z., Flemming, H. & Mitchell, A. S. Cognitive functions and neurodevelopmental disorders involving the prefrontal cortex and mediodorsal thalamus. *Front. Neurosci.* **12**, 33 (2018).
80. Benoit, L. J. et al. Adolescent thalamic inhibition leads to long-lasting impairments in prefrontal cortex function. *Nat. Neurosci.* **25**, 714–725 (2022).
81. Bulin, S. E., Hohl, K. M., Paredes, D., Silva, J. D. & Morilak, D. A. 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).
82. He, K., Zhang, X., Ren, S., & Sun, J. Delving deep into rectifiers: surpassing human-level performance on imageNet classification. *Proceedings of IEEE International Conference on Computer Vision (ICCV)* 1026–1034 (IEEE Computer Society, 2015).
83. Knyazev, A. V. & Argentati, M. E. Principal angles between subspaces in an A-based scalar product: algorithms and perturbation estimates. *SIAM J. Sci. Comput.* **23**, 2008–2040 (2002).
84. Bouchard, K. E. & Kumar, A. *Feedback Controllability is a Normative Theory of Neural Population Dynamics*. Paper print at Researchsquare https://doi.org/10.21203/rs.3.rs-4102129/v1 (2024).
85. Rajakumar, A., Rinzel, J. & Chen, Z. S. Stimulus-driven and spontaneous dynamics in excitatory-inhibitory recurrent neural networks for sequence representation. *Neural Comput.* **33**, 2603–2645 (2021).
86. Asllani, M., Lambiotte, R. & Carletti, T. Structure and dynamical behavior of non-normal networks. *Sci. Adv.* **4**, sciadv.aau9403 (2018).
87. 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 N. Lam, R.D. Wimmer, and Y. Liu for valuable comments on the manuscript. The work was supported by grants R01-MH118928 (Z.S.C.), RF1-DA056394 (Z.S.C.), P50-MH132642, and R01-MH139352 (Z.S.C. and M.M.H.) from the US National Institutes of Health. We thank the Conte Center Team for scientific input and inspiration on cognitive thalamus. An earlier version of this work has appeared in BioRxiv preprint (https://biorxiv.org/cgi/content/short/2022.12.11.519975v1) and was presented in COSYNE'23.

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, and analyzed and interpreted the data. A.M. conducted the animal's behavioral experiments and image analysis. M.M.H. provided other experimental data and interpreted the data. Z.S.C. wrote the manuscript, with additional inputs from X.Z., A.M., and M.M.H. All authors read, edited, and approved the manuscript. Z.S.C. supervised the work. Z.S.C. acquired funding.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-58011-1.

Correspondence and requests for materials should be addressed to Zhe Sage Chen.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.