Whole-brain causal connectivity during decoded neurofeedback: a meta study

Fahimeh Arab[1], AmirEmad Ghassami[2], Hamidreza Jamalabadi[3], Megan A. K. Peters[4,5,6], and Erfan Nozari[1,7,8,*]

[1]Department of Electrical and Computer Engineering, University of California, Riverside, USA
[2]Department of Mathematics and Statistics, Boston University, USA
[3]Department of Psychiatry and Psychotherapy, Phillips University of Marburg, Germany
[4]Department of Cognitive Sciences, University of California, Irvine, USA
[5]Center for the Neurobiology of Learning & Memory, University of California, Irvine, USA
[6]Program in Brain, Mind, & Consciousness, Canadian Institute for Advanced Research, Canada
[7]Department of Mechanical Engineering, University of California, Riverside, USA
[8]Department of Bioengineering, University of California, Riverside, USA
[*]Corresponding author (email: erfan.nozari@ucr.edu)

Abstract

Decoded Neurofeedback (DecNef) represents a pioneering approach in human neuroscience that enables modulation of brain activity patterns without subjective conscious awareness through the combination of real-time fMRI with multivariate pattern analysis. While this technique holds significant potential for clinical and cognitive applications, the causal mechanisms underlying successful DecNef regulation and the neural dynamics that distinguish successful learners from those who struggle remain poorly understood. To address this question, we conducted a meta-study across functional magnetic resonance imaging (fMRI) data from five DecNef experiments, each with multiple fMRI sessions, to reveal causal network dynamics associated with individual differences in neurofeedback performance. Using the newly proposed CaLLTiF causal discovery method, we computed causal maps to identify causal network patterns that distinguish DecNef regulation from baseline and account for variations in neurofeedback success. We found that enhanced connectivity within the bilateral control network–particularly stronger connections involving the posterior cingulate and precuneus cortex–predicted neurofeedback success across all five studies. Whole-brain causal connectivity during DecNef further exhibited distinct network reorganizations, characterized by reduced average path lengths and increased right-limbic nodal degrees. Further, comparisons across cognition- and perception-targeted DecNef revealed a remarkable separation in connections to and from the somatomotor network, where connections between somatomotor and control-default-attention networks are larger during cognitive neurofeedback while causal effects between somatomotor and subcortical-visual-limbic networks are larger during perceptive DecNef. This is despite the fact that none of the involved studies targeted or involved motor activity. Overall, our results demonstrated the key role of bilateral medial control network in successful DecNef regulation regardless of the DecNef targets, a clear separation in somatomotor involvement between cognitive and perceptive DecNef, and general promise of whole-brain causal discovery in understanding complex neural processes such as decoded neurofeedback.

Keywords: fMRI, causal discovery, brain networks, statistical algorithms, cognitive neuroscience, decoded neurofeedback

Introduction

Twenty years have passed since (Weiskopf et al., 2004)'s pioneering demonstration of the feasibility of using real-time fMRI as a brain-computer interface, enabling participants to self-regulate brain activity via feedback. More recently, decoded neurofeedback (DecNef) has been proposed as a novel technique combining implicit neurofeedback and multivariate pattern analysis (Shibata et al., 2011; Taschereau-Dumouchel et al., 2021). Unlike traditional methods that rely on overall signal amplitude and explicit strategies, DecNef

induces specific signal patterns in target brain regions, altering these neural patterns (and subsequently impacting behavior) without participants' awareness of the exact content and purpose of the manipulation (Cortese et al., 2021; Shibata et al., 2019, 2011). As a result, DecNef can help reduce potential confounding effects from cognitive processes or awareness of the specific dimension being manipulated. (Cortese et al., 2021). These characteristics have made DecNef especially well-suited for developing new clinical applications, particularly in the treatment of neuropsychiatric disorders (Chiba et al., 2019; Koizumi et al., 2016; Taschereau-Dumouchel et al., 2018, 2020; Yamada et al., 2017). DecNef has also proved valuable beyond clinical applications, offering insights in systems and cognitive neuroscience to explore fundamental brain functions in diverse areas such as visual sensitivity (Shibata et al., 2011), color perception (Amano et al., 2016), fear memory (Koizumi et al., 2016; Taschereau-Dumouchel et al., 2018), facial preference (Shibata et al., 2016), and perceptual confidence (Cortese et al., 2016).

The precise neural mechanisms underlying DecNef, however, are poorly understood. Recent research has started to delve into this question through a variety of methods, including meta-analyses, computational models, and neural network simulations (Emmert et al., 2016; Haugg et al., 2020; Oblak et al., 2017, 2019; Pereira et al., 2024; Sepulveda et al., 2016; Shibata et al., 2019; Skottnik et al., 2019). One plausible mechanism that has been suggested is reinforcement learning. For example, Shibata and colleagues (Shibata et al., 2019) proposed the "targeted neural plasticity model," suggesting that DecNef induces plasticity at the neuronal level in specific brain regions, leading to behavioral changes. Empirical evidence from previous studies supports this model. The findings by (Shibata et al., 2019) indicate that DecNef likely drives neural plasticity through reinforcement learning mechanisms, with significant activation in reward-related brain regions such as the ventral striatum and putamen in response to feedback signals. This suggests that DecNef engages the brain's reward-processing circuits and may share neural foundations with conventional neurofeedback and brain-machine interfaces. However, while these results highlight specific regional activations, they leave unexplored *how* broader brain connectivity and interactions contribute to the neural dynamics of the induction process. Our work addresses this gap by identifying causal interactions between brain regions during DecNef induction sessions compared to baseline. By examining these connectivity patterns, we aim to provide a connectivity-based understanding of neural dynamics and offer insights into the mechanisms that drive DecNef's effects on brain function.

Causal discovery provides an invaluable opportunity for uncovering brain mechanisms from purely observational data, such as fMRI. fMRI possesses a major advantage for causal discovery because of its potential for whole-brain coverage, but it also poses significant challenges. fMRI's low temporal resolution, combined with the computational complexity of analyzing large-scale networks, makes it difficult to accurately discern directional relationships between brain regions. Traditional methods like Granger Causality (GC)(Barnett and Seth, 2014; Granger, 1969) and Dynamic Causal Modeling (DCM) (Friston et al., 2014) are common choices, but struggle to handle these complexities, especially in extensive fMRI networks. In our previous work, we developed CaLLTiF (Causal discovery for Large-scale Low-resolution Time-series with Feedback) (Arab et al., 2023) to address these challenges by utilizing both lagged and contemporaneous variables to identify causal connections. When applied to synthetic fMRI data, CaLLTiF outperformed state-of-the-art methods in both accuracy and scalability. Applied to resting-state human fMRI, CaLLTiF uncovered causal connectomes that are highly consistent across individuals, revealing a top-down causal flow from attention and default mode networks to sensorimotor regions, Euclidean distance-dependence in causal interactions, and a strong dominance of contemporaneous effects.

Building on these insights, our current study applies CaLLTiF to DecNef induction sessions and compares them to baseline to explore how specific causal interactions shape neural dynamics during neurofeedback. We conducted a meta-study across five previously-published DecNef experiments, each involving multiple fMRI sessions per participant, to identify core causal mechanisms underlying DecNef across varied neurofeedback tasks. Using CaLLTiF, we derived causal graphs from fMRI data collected during both neurofeedback (NF) and decoder construction (DC) sessions (used to train the machine learning models which are then applied during real-time neurofeedback), uncovering brain interactions that either enhance or diminish neurofeedback performance. This meta-analysis integrates data from 45 participants across five distinct tasks, allowing us to isolate causal mechanisms that are fundamental to DecNef and not specific to any single task. Our findings reveal distinct patterns in causal dynamics, with mechanisms differing between tasks targeting cognitive functions and those focused on perceptual processes.

Results

Causal discovery from decoded neurofeedback. Figure 1a illustrates the general framework of Decoded Neurofeedback (DecNef), comprising decoder construction (DC) sessions and neurofeedback (NF) sessions. In DC sessions, multivariate pattern analysis (MVPA) is used to train a decoder on brain activity patterns. In NF sessions, participants use the results of this decoder applied to current patterns of brain response to self-regulate specific brain regions based on real-time feedback. Figure 1a summarizes the general experimental design employed in the "DecNef collection" (Cortese et al., 2021), whose data we used for our analyses. Figure 1b presents the pipeline for the causality analysis conducted in this paper. Using our recently proposed algorithm, CaLLTiF (Arab et al., 2023), we derived causal brain connectomes from whole-brain fMRI data collected during both NF and DC sessions, with the latter used as a subject- and task-specific baseline for the former. Starting with fMRI data for each subject, we performed an automated parcellation that divided the brain into 100 cortical (Schaefer et al., 2018) and 16 subcortical (Yeo et al., 2011)) regions. We then applied CaLLTiF to the data from each session type to construct causal graphs, capturing directional relationships between parcels and revealing network dynamics specific to NF and baseline sessions.

To enhance interpretability, we further combined parcels belonging to the same "functional networks" (Yeo et al., 2011) into 7 cortical (Schaefer et al., 2018) and 1 subcortical (Tian et al., 2020) subnetworks, each separated across the left and right hemispheres. This allowed us to generate subnetwork-level causal graphs, which we then statistically compared between NF and baseline sessions. This comparison enabled us to identify key differences in subnetwork connectivity patterns, shedding light on how neurofeedback impacts functional brain networks relative to baseline conditions.

We found the maximum similarity within the NF graphs of the same subject. To quantify the consistency of causal graphs and assess the robustness of causal structures across different conditions, we computed a set of correlation measures comparing causal graphs across studies, subjects, sessions, and runs, as illustrated in Figure 1c. Our analysis revealed that NF graphs from the same subject within their own NF sessions exhibited the highest similarity scores, suggesting stable and individualized causal connectivity patterns during NF. This high within-subject similarity was followed by the similarity between NF graphs and baseline graphs from the same subjects. The partial similarity between NF and baseline sessions implies that, while individualized patterns persist, NF sessions introduce unique causal dynamics that set them apart from baseline sessions. Next in similarity were neurofeedback graphs across different subjects within the same study, suggesting that some shared causal features may be driven by study-specific protocols or task demands. The lowest similarity was observed between NF graphs from subjects across different studies, reflecting the influence of study-specific factors—such as targeted brain regions, neurofeedback paradigms, and participant characteristics—on the resulting causal network patterns. As shown in Figure 1c, these findings reveal a hierarchy in the consistency of causal connectivity, with the strongest patterns occurring within individual subjects' NF sessions and the greatest variability seen across different studies. This hierarchical pattern underscores the personalized nature of neurofeedback's impact on brain network organization while also highlighting the role of study design in shaping causal connectivity structures. Such insights are valuable for refining neurofeedback interventions by balancing individualized approaches with study-specific factors across experiments.

Neurofeedback (NF) graphs show greater heterogeneity across subjects and sessions compared to baseline graphs. To quantify this variability, we calculated correlations between each pair of baseline graphs and each pair of NF graphs across all studies, subjects, sessions, and runs. As shown in Figure 1d, NF graphs exhibited significantly more variability, while baseline graphs were more consistent across subjects and sessions. This difference likely arises from the nature of the NF task, where subjects aim to modulate target brain activity to achieve higher scores, allowing flexibility in the brain dynamics they engage. In contrast, baseline sessions follow a relatively consistent task design across all studies, providing fewer opportunities for individual deviations. This distinction highlights the adaptive and personalized nature of neurofeedback, where each subject's unique neural responses contribute to greater variability.

During NF sessions, we observed increased engagement of control, limbic, and visual networks, along with diminished involvement of attention networks. We next examined the strengths of iden-

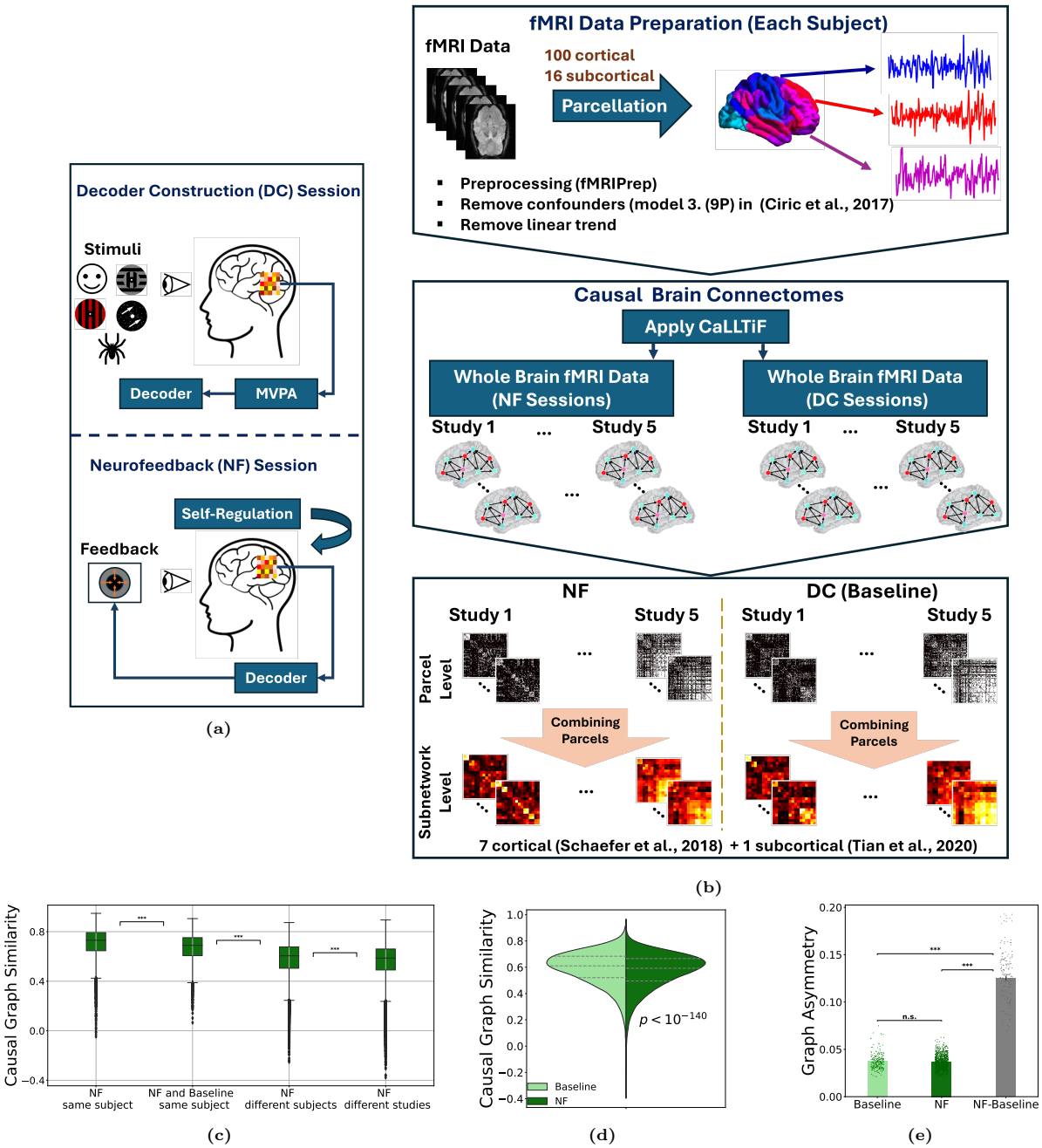


Figure 1: Overview of the decoded neurofeedback (DecNef) experimental pipeline and data preprocessing. (a) During the Decoder Construction session (DC, used as baseline in the present study), participants view stimuli while their while brain fMRI is recorded and used (offline) to construct decoders through multivariate pattern analysis (MVPA). In the Neurofeedback (NF) session, participants engage in self-regulation of brain activity, guided by real-time feedback based on decoders trained on data from the DC session. **(b)** For each subject, fMRI data is parcellated into 100 cortical and 16 subcortical regions. Preprocessing with fMRIPrep includes skull stripping, motion correction, spatial normalization, and smoothing, followed by additional steps "consisting of" confound removal (model 3, 9P in (Ciric et al., 2017)) and linear detrending. CaLLTiF is then applied to whole-brain fMRI data from both NF and baseline sessions to generate causal connectivity maps across multiple studies, sessions and runs. These connectivity matrices are combined into subnetwork-level representations, organized into 7 cortical and 1 subcortical subnetworks. **(c)** Hierarchy of causal connectivity consistency across conditions. Strongest similarity is observed within individual subjects' NF sessions, while the greatest variability occurs across different studies. **(d)** NF graphs exhibit greater variability across subjects and sessions compared to baseline. To quantify this, we calculated the Pearson correlation as a measure of similarity between each pair of NF causal graphs and, separately, each pair of baseline causal graphs across all studies, subjects, sessions, and runs. **(e)** Distribution of causal graph asymmetries computed for NF graphs, baseline graphs, and the relative (NF-baseline) graphs.

tified causal edges, as measured by the weighted subnetwork graphs using partial correlations (cf. Methods). Both NF and baseline graphs predominantly exhibited excitatory connections, as shown in Supplementary Figures 2a and 2b. For each pair of subnetworks (each edge in the subnetwork graph), statistical tests were conducted to compare the distribution of edge weights across all baseline and NF graphs from all studies. Since edge weights in each causal graph are derived from partial correlations, the sign of these correlations (positive or negative) can provide insights into whether connections are excitatory or inhibitory.

In general we can have four edge types with varying effects in neurofeedback: excitatory edges becoming more or less excitatory (Figures 2a and 2c) and inhibitory edges becoming more or less inhibitory (Figures 2b and 2d). Figure 2e displays the set of edges that strengthen during NF sessions compared to baseline sessions (whether positive or negative), while Figure 2f illustrates the set of edges that weaken in NF sessions compared to baseline sessions (positive or negative). We observed heightened engagement of control, limbic, and visual networks, with reduced involvement of attention networks during NF sessions. However, these differences tend to "average out" when examining more summarized network measures, where few significant distinctions remain. Among the global metrics–graph density, shortest path length, assortativity, modularity, spectral radius, and synchronizability–only the average shortest path length showed a significant reduction in NF graphs (Figure 2i). No other global metrics exhibited significant differences between conditions. For nodal centralities, we found a significant difference in nodal degree only in the right limbic system, with NF graphs showing a higher degree compared to baseline (Figure 2g). Additionally, for causal flow, only the right ventral attention subnetwork revealed a significant difference, where NF graphs had a weaker sink strength than baseline graphs (Figure 2h).

Key edges within bilateral control network linked to the posterior cingulate and precuneus cortex showed stronger connectivity in NF sessions and positively correlated with neurofeedback score. Next, we asked whether subject-specific causal graphs can be used to predict eash subject's success in self-regulation, measured by their trial-by-trial DecNdef scores. To ensure comparability of scores across studies and eliminate study-specific biases, we applied a preprocessing pipeline (Figure 3a). This involved transforming scores with an inverse sigmoid function (the last layer of the logistic regression models used in score generation), z-scoring them across studies, averaging scores within each run, and further averaging over consecutive runs to associate each pair of runs with a single causal graph. We then conducted a correlation analysis at the level of causal edges within the subnetwork-level graphs. This detailed examination revealed many edges that were significantly correlated with the neurofeedback score. At the same time, we computed a differential graph to identify edges that were significantly stronger during NF sessions compared baseline. We then examined the intersection of two graphs, which gave rise to three sets of edges, as shown in Figure 3b. The first set includes edges that are stronger in NF sessions and positively correlate with the score (red), indicating these edges enhance the neurofeedback score. The second set comprises edges that are stronger in NF sessions but negatively correlate with the score (blue), suggesting these edges detract from the neurofeedback score. Lastly, the third set consists of edges that are stronger in NF sessions but do not have a significant impact on the neurofeedback score (gray, no significant correlation with the score). These findings highlight specific patterns of connectivity that may underlie the efficacy of neurofeedback training and provide a more nuanced understanding of the relationship between brain network dynamics and neurofeedback performance.

As shown in Figure 3b (red edges), connections within the control network in each hemisphere are the only edges at the subnetwork level that are statistically stronger during NF *and* positively correlate with neurofeedback scores. Further parcel-level analysis revealed specific control network parcels driving this effect. Figure 3c shows a zoomed-in view of the bilateral control networks, where we can see that (1) there are no causal connections that were stronger in NF compared to baseline *and* negatively correlated with NF score (i.e., no blue edges), and (2) all the edges that are stronger in NF and positively correlate with score connect to bilateral posterior cingulate (PCC) and precuneus cortices. Our results thus suggest that bilateral PCC and precuneus function as a medial control hub in DecNef. The control network is widely recognized as a core system supporting high-level cognitive functions such as attention, task management, and goal-directed behavior (Cole et al., 2013; Seeley et al., 2007). It facilitates the integration of information across distributed brain regions, allowing for adaptive responses to dynamic task demands (Cole et al., 2013). Within this network, PCC and precuneus play essential roles in orienting attention, maintaining focus, and coordinating between self-referential and externally directed processes (Cavanna and Trimble,

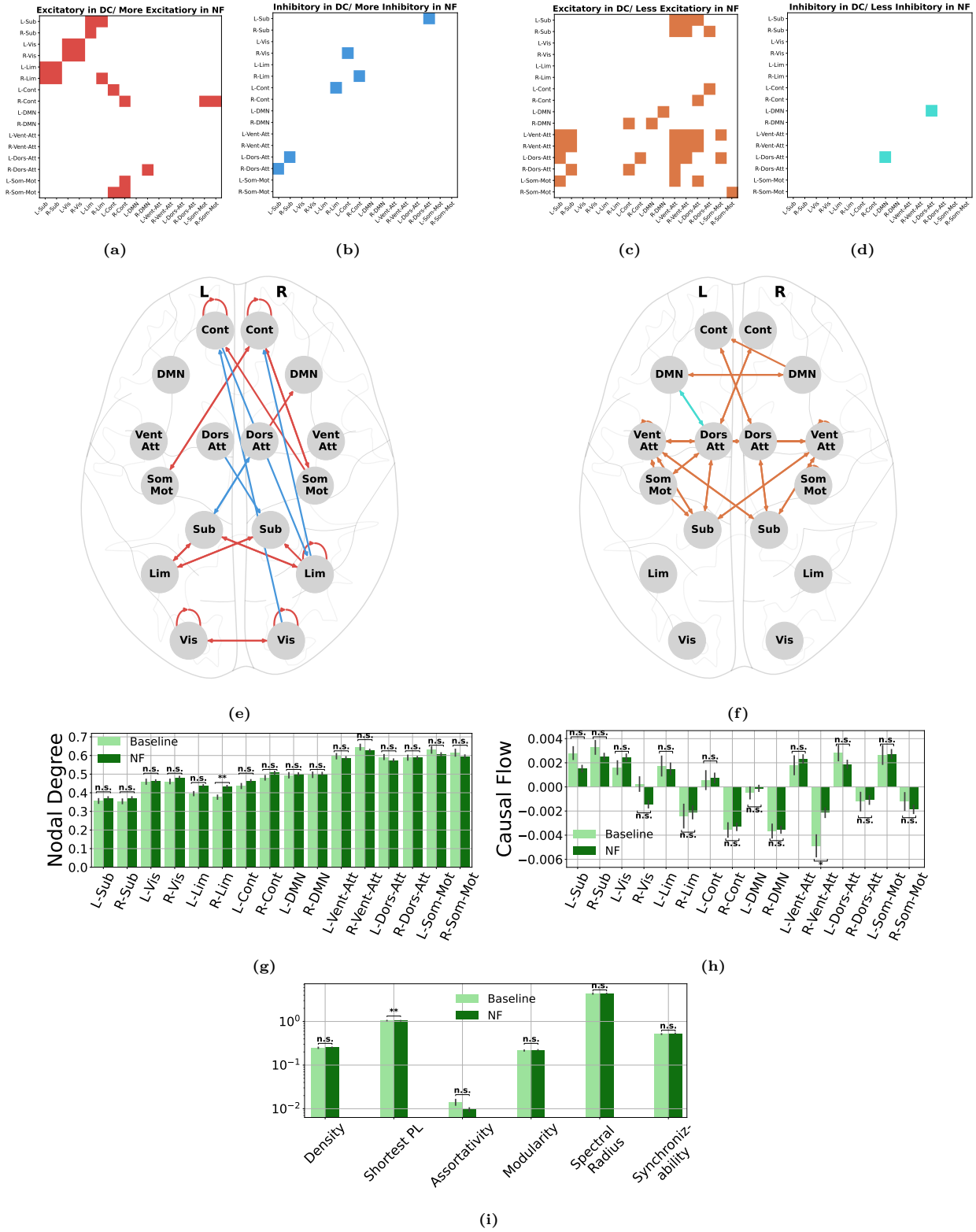


Figure 2: Neurofeedback involves strengthening of control, limbic, and visual causal connectivity while weakening causal connections involving attention networks. **(a)** Excitatory edges in baseline sessions that become more excitatory in NF sessions. **(b)** Inhibitory edges in baseline sessions that become more inhibitory in NF sessions. **(c)** Excitatory edges in baseline sessions that weaken in NF sessions. **(d)** Inhibitory edges in baseline sessions that weaken in NF sessions. **(e)** Schematic topographic visualization of edges from (a) and (b). **(f)** Similar to (e) but for edges in (c) and (d). **(g)** Distribution of nodal degrees across different subnetworks for NF and baseline sessions. **(h)** Similar to (g) but for nodal causal flows. **(i)** Distribution of global network measures for NF and baseline causal graphs.

2006; Leech and Sharp, 2014). The observed strengthening of connections in this medial control hub during neurofeedback suggests that these regions serve as integrative centers, facilitating adaptive, goal-oriented adjustments necessary for effective neurofeedback performance. This enhanced connectivity likely supports the ability to modulate brain states in response to feedback, positioning the PCC and precuneus as critical components that link cognitive flexibility with targeted brain dynamics during neurofeedback.

As seen in Figure 3c, there is noticeable hemispheric asymmetry in the existing nodes and connections within the control network. To ensure that this asymmetry is not due to how the parcels within high-level association cortices are assigned across different brain networks, particularly between control and default mode networks (DMN), we also examined the edges that are stronger in NF and positively correlate with the score within the DMN. Interestingly, only one additional edge appeared (Supplementary Figure 4), indicating that the hemispheric asymmetry observed within the control network is likely an intrinsic characteristic rather than an artifact of network parcellation. Although some parcels in the right control network have corresponding counterparts in the left DMN (e.g., several prefrontal cortex parcels), Supplementary Figure 4 shows that the edges strengthened in NF and positively correlated with the score are primarily confined to the control network, and predominantly within the right control network. Similar to earlier comparisons, we observed no significant correlations between average neurofeedback scores and the global measures of causal graphs, including graph density, shortest path length, assortativity, modularity, spectral radius, and synchronizability (Supplementary Figure 3a). Similarly, analyses of nodal centralities, including nodal degree and causal flows for each node, showed no significant associations with neurofeedback scores (see Supplementary Figures 3b,3c).

Somatomotor causal connectivity distinctly separates perceptive from cognitive neurofeedback. In the five studies we analyzed, two studies (Study 2 and Study 3 in the DecNef dataset) targeted early visual areas for neurofeedback, whereas the other three studies (Studies 1, 4, and 5 in the DecNef dataset) targeted higher-level brain regions—namely, the cingulate cortex, inferior parietal cortex, dorsolateral prefrontal cortex, and ventral temporal areas. We categorized these studies into two sets. Studies 2 and 3 were designated as "perception experiments," regulating lower-level cortices, while Studies 1, 4, and 5 were designated as "cognition experiments," regulating higher-level areas. We then tested whether significant differences in the causal graphs between these two sets of neurofeedback experiments are observable. As with previous analyses, we did not find major differences at the level of global network measures (See Supplementary Figure 5a), or nodal centralities (See Supplementary Figures 5b, 5c). At the edge level, however, distinct patterns emerged between perceptive and cognitive NF sessions (Figure 4a). Connections involving the somatomotor network reveal a distinct pattern: those that strengthen during cognitive NF distinctly link the somatomotor network the control, defaul mode, and attention networks (hierarchically higher-order networks), whereas those that intensify during perceptive NF distinctly connect the somatomotor network to subcortical, visual, and limbic networks (Figure 4b). Notably, this strong division appears even though motor regions were not directly targeted in any of the studies. A similar but subtler pattern is observed in the dorsal attention network's connectivity ((Figure 4c), while a reversed pattern occurs in subcortical connections (Figure 4d). Finally, connections among other networks, namely, the ventral attention, limbic, visual, control, and default mode networks, are predominantly stronger during cognitive NF (Figure 4e).

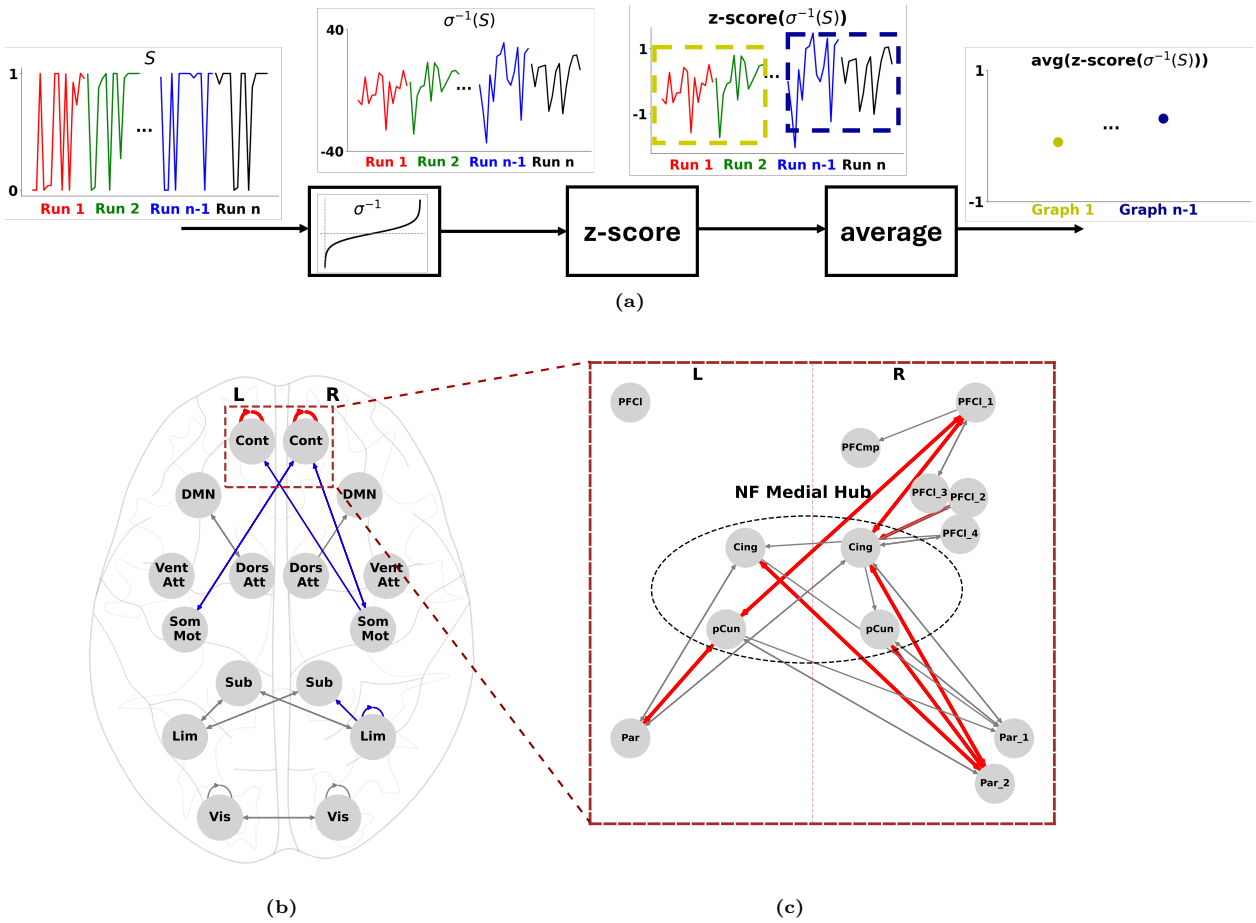
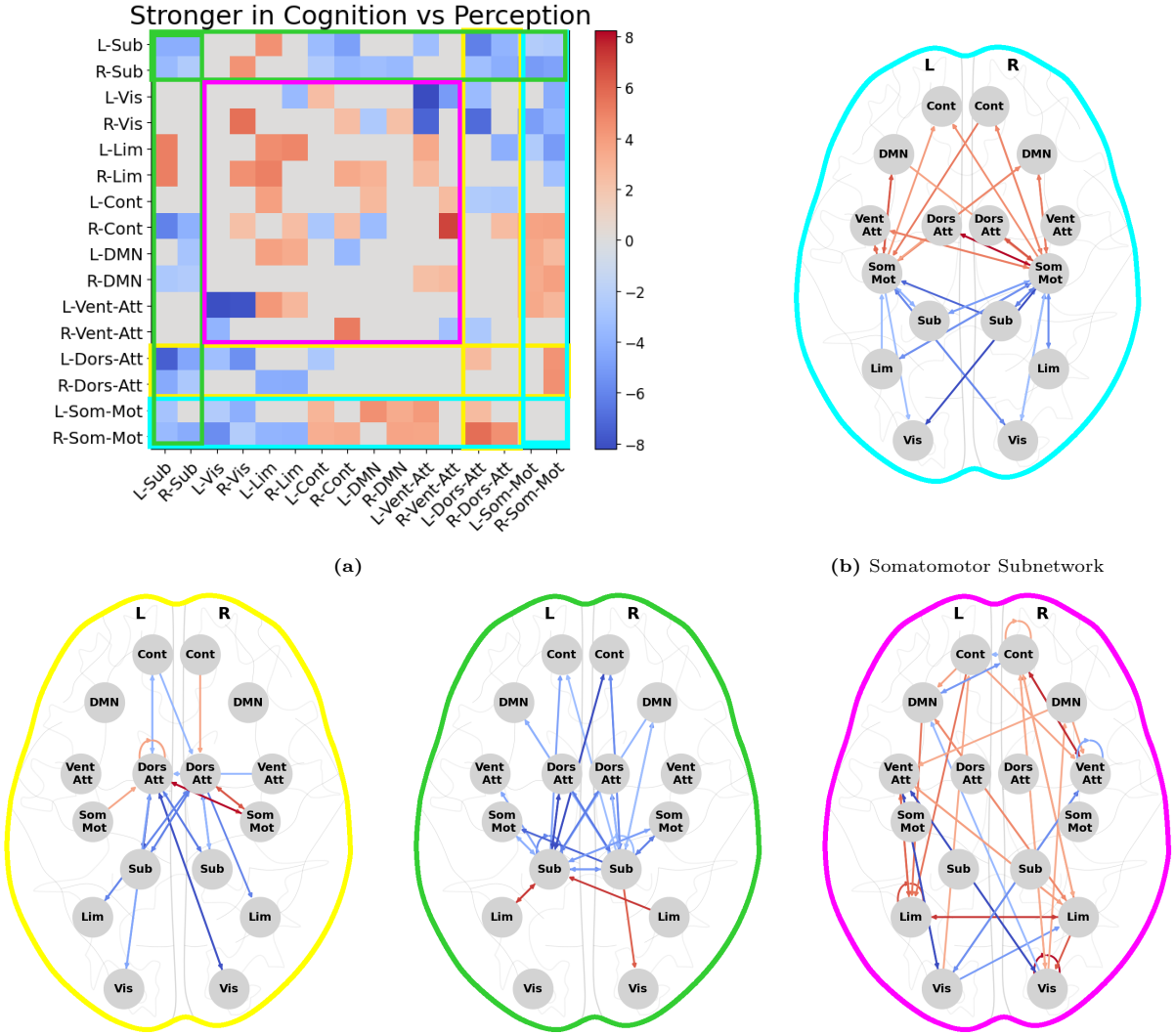


Figure 3: The medial control hub in decoded neurofeedback. We found that edges within bilateral control network linked specifically to the posterior cingulate cortex (PCC) and precuneus show stronger connectivity in NF sessions and positively correlated with neurofeedback score. **(a)** Score preprocessing steps to ensure comparability and minimize biases. We first applied an inverse sigmoid function to transform scores back to their original range. We then z-scored the scores for cross-study standardization, followed by averaging within each run, and then across two consecutive runs to yield a single representative score for each causal graph. **(b)** Edges that are significantly stronger in NF sessions compared to baseline and correlate with NF score positively (red), negatively (blue), or insignificantly (gray). **(c)** Zoomed-in view of the control network. Edge colors have the same meaning as in (b). All red edges connect to either PCC or precuneus, hence highlighting them as a medial control hub for DecNef. Also remarkably, we did not find any edges within the control network that are significantly stronger during NF but correlate negatively with NF score.



(a)

(b) Somatomotor Subnetwork

(c) Dorsal-Attention Subnetwork

(d) Subcortical Subnetwork

(e) Other Subnetworks

Figure 4: Somatomotor causal connectivity distinctly separates perceptive from cognitive neurofeedback. (a) Heatmap illustrating the significant differences in edge strength between cognitive neurofeedback (Studies 1, 4, 5, targeting higher brain functions) and perceptive neurofeedback (studies 2, 3, targeting lower brain functions). Red-colored edges indicate stronger connectivity in cognitive neurofeedback, while blue-colored edges represent stronger connectivity in perceptive neurofeedback. **(b)** Schematic diagram of only the subset of edges in (a) that connect to the bilateral somatomotor subnetwork. Edges linking the somatomotor network to higher brain areas, such as the control, default mode, and attention networks, exhibit greater strength during cognitive neurofeedback. In contrast, edges connecting the somatomotor network to the (hierarchically-lower) subcortical, visual, and limbic networks show stronger connectivity during perceptive neurofeedback. **(c)** Similar to (b) but for the subset of edges connecting to the dorsal attention subnetworks. A similar but less prominent pattern to (b) is observed. **(d)** Similar to (b) but for the subset of edges connecting to the subcortical subnetwork. Here we observe an approximately opposite pattern to that observed in (b). **(e)** Remaining connections in (a) other than those shown in (b-d). These consist of connections between the ventral attention, limbic, visual, control, and default mode networks and are largely stronger during cognitive neurofeedback.

Discussion

Comparison between NF and baseline sessions A comparative analysis between NF and baseline sessions highlights the unique neural dynamics fostered by NF. NF sessions showed significantly more variability in causal connectivity across subjects and sessions. This variability is likely due to the open-ended nature of NF training, where subjects are given real-time feedback to modify brain activity to achieve a target state without explicit task constraints. Consequently, subjects in NF sessions are free to explore various neural strategies to reach the desired brain state, leading to a wider range of causal network configurations. In contrast, baseline sessions are structured and task-oriented, requiring participants to complete specific cognitive or perceptual tasks designed to generate a consistent neural response across sessions and subjects. This structure inherently constrains the degree to which brain dynamics can vary, resulting in greater similarity in connectivity patterns across individuals. This finding can suggest that NF facilitates individualized brain dynamics exploration more, compared to more standardized neural response imposed during baseline. These differences underscore the adaptive nature of NF as an individualized training protocol and suggest that decoder construction sessions could serve as a baseline for understanding how neurofeedback reshapes brain networks. Furthermore, the higher correlation within NF causal graphs for the same subject—compared to NF graphs across subjects—suggests that while NF encourages flexible neural exploration, there are still stable, individualized patterns that characterize each participant's response to NF.

Based on our analysis, we observed significant differences in causal connectivity patterns between NF and baseline sessions, particularly at the level of individual edges. Both NF and baseline graphs were dominated by excitatory connections, yet the NF graphs exhibited unique alterations in connectivity strength. The global and nodal measure comparisons between NF and baseline graphs reveal that only a few metrics differ significantly, highlighting the selective nature of network reorganization during NF. Specifically, while NF graphs displayed a smaller average path length, suggesting more direct or efficient communication pathways, other global metrics—including graph density, assortativity, and modularity—did not show notable differences. This can indicate that the reconfiguration of brain networks during NF is targeted rather than widespread, adapting selectively to the demands of feedback-based learning. For nodal centralities, distinct differences emerged within the right limbic and ventral attention subnetworks. NF graphs demonstrated higher nodal degrees in the right limbic network, implying increased connectivity in areas associated with emotional engagement and motivational drive—key factors for maintaining focus and effort during neurofeedback. Additionally, NF graphs showed a decrease in causal flow within the right ventral attention network compared to baseline, potentially indicating a shift from external attention processing to more internally directed cognitive strategies. This shift likely supports the NF task's emphasis on internal modulation of brain states in response to feedback, aligning with the task's feedback-driven nature and reducing reliance on external attentional mechanisms.

The edge-level analysis comparing causal graphs between NF and baseline sessions highlights a clear reconfiguration of network engagement. Specifically, there is a stronger involvement of the control, limbic, and visual networks, paired with a reduced involvement of attention networks across interactions with every other subnetwork. This increased connectivity within and between the control, limbic, and visual networks during NF sessions likely reflects the core demands of NF, where participants strive to modulate their brain states to align with feedback targets. Enhanced connections between the control network and other networks may indicate an increased need for self-regulation and executive control, crucial for directing and sustaining focus on the internal goal of modulating neural activity. Similarly, the heightened involvement of the limbic network, with interactions across other subnetworks, suggests that emotional and motivational processes are integral to the NF task. The limbic network, often associated with engagement, reward, and motivation, may provide the motivational drive necessary for participants to stay engaged with the NF feedback, especially when attempting to achieve target states over prolonged sessions.

The visual network's increased connectivity with other subnetworks aligns with the feedback's visual nature, where visual processing is essential for participants to interpret the cues on screen. This added involvement of the visual network reinforces the role of visual feedback in guiding participants as they work toward their targets. In contrast, the reduced connectivity of attention networks across all interactions with other subnetworks suggests a shift from external attention toward internally focused regulation. During NF, participants may be less reliant on attentional processing as typically required in response-driven tasks, favoring an internally driven strategy. This reduction in attention network involvement may facilitate more

flexible approaches, allowing participants to explore different internal strategies rather than relying heavily on external, reactive attention. Overall, these patterns point to a distinct network reconfiguration in NF, prioritizing internal regulation and motivational support through stronger control, limbic, and visual network interactions, while decreasing reliance on externally oriented attention systems. This reorganization may represent a critical neural adaptation that enables effective neurofeedback learning.

Variability in neurofeedback performance and neural dynamics One of the notable observations in this study is the high degree of variability in NF performance across participants, reflected in the range of causal connectivity patterns identified in NF sessions. This variability likely arises from individual differences in the capacity to modify brain activity patterns in response to NF feedback. Some participants may be better equipped to recruit the bilateral control network, thereby achieving greater modulation of the target brain areas and higher NF scores. Conversely, others may rely on alternative neural strategies or fail to establish effective connectivity within the necessary networks, leading to lower NF performance.

The flexibility inherent in NF tasks, which do not impose strict constraints on the neural pathways that participants can engage, further allows for this variability. In NF sessions, participants are incentivized to achieve a higher score by matching their brain activity to a pre-specified pattern, but the strategies and networks they recruit are not explicitly dictated. This freedom results in diverse causal configurations, as individuals explore different neural pathways to meet the feedback criteria. In contrast, the baseline sessions likely yield more consistent causal structures, as they are designed around task-driven demands that are uniformly applied across participants. The findings indicate that this variability in NF might be a crucial factor contributing to individual differences in NF effectiveness. For clinical and research applications, understanding these individualized causal dynamics could help optimize NF training protocols by tailoring the feedback and task requirements to each participant's unique neural response pattern. The distinct connectivity patterns in high-performing versus low-performing participants suggest that monitoring these dynamics could serve as a biomarker for successful NF learning, potentially guiding personalized interventions that maximize NF's efficacy.

Assumptions and limitations. One limitation of this study arises from the constraints of the CaLLTiF method, particularly given the slower temporal resolution of the fMRI data (TR = 2s). With such a TR, many causal interactions are detected as contemporaneous rather than lagged, which can result in increased symmetry and reduced causal flow information in the resulting graphs. To address this, we modified CaLLTiF to yield more asymmetric graphs that better capture directional causality. However, this adjustment may increase the probability of Type I error. Nevertheless, we believe this trade-off is moderated by CaLLTiF's inherently conservative nature, which includes multiple comparison correction steps to control for false positives. As such, we are optimistic that the modified CaLLTiF provides accurate directional insights without substantially inflating Type I error rates. Another limitation concerns the meta-analysis across studies and subjects. Each DecNef study involved different participants, resulting in limited graph samples per subject. This constraint precluded robust per-subject analyses in some cases, as we lacked sufficient samples to explore individual-level causal dynamics comprehensively. While our combined dataset offers valuable insights at the group level, it limits our ability to make strong individual-specific conclusions. Additionally, due to the lack of resting-state fMRI data for each subject, we used data from each subject's DC sessions as a baseline for comparison. This approach allowed us to assess deviations of the NF graphs from the DC graphs, which effectively highlighted changes induced during NF sessions. Although this baseline choice added useful asymmetry and enhanced certain aspects of the analysis by making NF graphs more asymmetric, it may not provide the optimal baseline for detecting causal dynamics in NF sessions. A true resting-state baseline could offer a more neutral benchmark for assessing changes specific to NF interventions.

Summary. This study provides a detailed meta-analysis of whole-brain causal connectivity during decoded neurofeedback, applying CaLLTiF as a state-of-the-art causal discovery method across data from five studies. By constructing causal graphs for both NF and baseline sessions, we uncovered distinct causal network characteristics in NF sessions that correlate with successful neurofeedback. In particular, enhanced connectivity within the bilateral control network, particularly those involving the posterior cingulate and precuneus cortex, emerged as a key factor linked to improved neurofeedback scores. Furthermore, NF sessions displayed

unique network reorganization patterns, such as reduced path lengths and increased right-limbic connectivity, setting them apart from the more structured baseline sessions. Additionally, somatomotor connectivity patterns were found to vary between cognitive-focused and perception-focused DecNef tasks, highlighting task-specific neural modulation. Together, these findings contribute to a deeper understanding of the neural dynamics in DecNef, with implications for refining its application in both clinical and cognitive neuroscience.

Material and Methods

Causal Discovery Algorithm (CaLLTiF)

In this work we used our recently developed causal discovery algorithm CaLLTiF (Arab et al., 2023) to extract causal connectivity graphs from fMRI data collected during NF and baseline sessions. Compared to the original algorithm in (Arab et al., 2023) here we slightly modified CaLLTiF to improve its effectiveness with even slower sampling of the fMRI data in this study (TR = 2s compared to TR = 0.72s in our earlier work). In CaLLTiF, a causal link is established from a node (parcel) X_i to node X_j with a lag of $\tau \geq 0$ samples if $X_i(t-\tau)$ is significantly correlated with $X_j(t)$ after conditioning on all other nodes and their lagged values, ensuring that correlation is not due to a common cause or mediation through other nodes. If $\tau = 0$, a bidirectional feedback connection is placed between X_i and X_j, unless at least one variable also causes the other with $\tau > 0$, in which case the direction of causality is determined based on the lagged effect(s). However, as noted in (Arab et al., 2023), lagged effects become exponentially harder to detect with increasing TR and finite samples, despite the presence of a statistically significant contemporaneous effect ($\tau = 0$), which is proof that a lagged effect must exist. To address this challenge, we adjusted CaLLTiF to handle the (even) slower sampling in this work. Specifically, for pairs of nodes with a statistically significant contemporaneous effect (detected at the originally suggested strict level $\alpha = 0.0025$), we relaxed the threshold of statistical significance on their lagged effects from $\alpha = 0.0025$ to $\alpha = 0.05$. Specifically, for pairs of nodes where only a statistically significant contemporaneous effect was detected (at the originally strict threshold of $\alpha = 0.0025$), we relaxed the significance level for detecting lagged effects from $\alpha = 0.0025$ to $\alpha = 0.05$. In CaLLTiF, a contemporaneous edge between two variables is typically considered bidirectional based on prior assumptions. By increasing the significance threshold for lagged edges, we aimed to uncover potential weaker lagged connections that may have been missed under the stricter α level. This adjustment allowed us to identify additional directional influences that would increase the asymmetry of the final causal graphs

Our analysis revealed that elevating the threshold parameter significantly enhances the asymmetry within the resulting causal graphs. This is achieved by detecting orientations from lagged edges that were initially weaker. This effect is depicted in Supplementary Figure 1, where it is evident that the peak asymmetry is observed at $\alpha = 0.5$. Despite this, we opted to limit our threshold to $\alpha = 0.05$ to ensure that the identified causal edges retained statistical significance. Asymmetry measures were computed across all causal graphs generated from both the baseline sessions and NF sessions within the scope of our studies. Another source of asymmetry in our graphs arises from our methodology, which treats the decoder graphs as the baseline. Figure 1e illustrates how this source of asymmetry contributes to the graphs' asymmetry in comparison to the original NF graphs we analyzed.

We investigated whole-brain causal connections using data from five DecNef studies (Cortese et al., 2021) through the CaLLTiF (Causal Discovery for Large-scale Low-Resolution Time-Series with Feedback) algorithm (Arab et al., 2023). For each participant, the data consists of a session used in the main experiment to train the machine learning decoder and several closed-loop fMRI neural reinforcement sessions. We computed one causal graph for each session, encompassing both baseline and NF sessions. Data were truncated to ensure the same sample size was used to compute each causal graph, and CaLLTiF was adapted to handle slower fMRI data. In total, we have 135 causal graphs for three NF sessions, each involving 9 subjects across 5 studies. For each session, one causal graph was computed. Additionally, for baseline session, we have 45 graphs encompassing all studies and subjects.

Data

Overview of DecNef experimental studies and targeted neural domains. The fMRI data used for causal connectivity analysis in this study were sourced from five distinct DecNef experiments, each examining

neural mechanisms in specific cognitive and perceptual domains (Cortese et al., 2021). For each participant, data includes a session for training the machine learning decoder and several (3 to 10) closed-loop fMRI neural reinforcement sessions.*Study 1* explored facial preference representation in the cingulate cortex (CC), showing that activation patterns within this region could be manipulated to alter preferences for initially neutral faces (Aharon et al., 2001; Chatterjee et al., 2009; Iaria et al., 2008; Said et al., 2011; Shibata et al., 2016). *Study 2* investigated associative learning between orientation and color in early visual areas, demonstrating that DecNef could induce long-term changes in color perception by linking specific visual features such as orientation and color in early visual areas (Amano et al., 2016). *Study 3* examined fear reduction through counter-conditioning in the visual cortex, leveraging DecNef to attenuate conditioned fear responses without explicit awareness (Koizumi et al., 2016). *Study 4* focused on the dissociation between subjective confidence and perceptual accuracy, using DecNef to manipulate confidence without affecting actual performance, challenging the prevailing view that confidence directly reflects perceptual reliability (Cortese et al., 2016; Fleming et al., 2012; Kepecs and Mainen, 2012; Koizumi et al., 2015; Meyniel et al., 2015; Rounis et al., 2010; Simons et al., 2010; Wilimzig et al., 2008). Finally, *Study 5* investigated the unconscious reprogramming of innate fear responses to spiders and snakes using hyperalignment-based neurofeedback, demonstrating a reduction in physiological fear indicators without conscious exposure to feared stimuli (Guntupalli et al., 2016; Haxby et al., 2011; Taschereau-Dumouchel et al., 2018). Collectively, these studies provide a rich dataset for examining causal brain dynamics across varied neural and behavioral domains, enhancing our understanding of individualized neurofeedback responses.

Unlike univariate approaches which measure overall activity levels within a region-of-interest (ROI) by treating each voxel independently, multivoxel pattern analysis (MVPA) using in DecNef (Kamitani and Tong, 2005; Norman et al., 2006) decodes information distributed across patterns of neural activity and can therefore result in higher target specificity. Recent advancements in DecNef include a method called hyperalignment (Haxby et al., 2011; Taschereau-Dumouchel et al., 2021), which allows the experimenter to infer the target neural representation indirectly from surrogate participants. Hyperalignment constructs a common, high-dimensional space from patterns of neural activity across participants using a series of linear transformations. These transformations align any new data patterns with the individual's brain coordinates and the model space coordinates. During the decoder construction session, participants performed tasks tailored to the study's focus, including a simple visual task (Studies 2 and 3), a preference task (Study 1), a perceptual task (Study 4), or a memory task (Study 5). In the NF sessions, participants consistently followed a similar procedure. They were instructed to adjust their brain activity to enlarge a feedback disc displayed on the screen at the end of each trial. The disc's size indicated the reward amount for that trial, contributing to a cumulative reward. Participants were told that the task's goal was to maximize their reward. However, they were unaware that the disc size—and thus the reward—was determined by how closely their current brain state matched a target state. The pre-trained decoder was used in real-time to evaluate this match. See (Cortese et al., 2021) for further details on DecNef and the present meta-dataset.

Decoded neurofeedback fMRI data collection. The fMRI data was acquired using Siemens MAGNETOM Verio and Prisma 3 Tesla MRI scanners. The scanning parameters included a repetition time (TR) of 2000 ms and a voxel size of $3 \times 3 \times 3.5$ mm^3 (See more details at (Cortese et al., 2021)). All participants across the five studies included in the analysis provided written informed consent. The recruitment procedures and experimental protocols were approved by the institutional review board at the Advanced Telecommunications Research Institute International (ATR, Kyoto, Japan), under the following approval numbers: 14–121, 12–120, 15–181, 14–140, and 16–181. The studies were conducted in accordance with the principles outlined in the Declaration of Helsinki.

fMRI data preprocessing. We initially preprocessed the fMRI data using standard steps implemented in fMRIPrep (Esteban et al., 2019). Subsequently, we eliminated 9 confounding factors from the time-series data of each voxel. We used Model 3. (9P) in (Ciric et al., 2017) which combines the 6 motion estimates, 2 physiological time series (mean White Matter and mean CSF signals), and the global signal. This model has been widely applied to functional connectivity studies (Ciric et al., 2017). For all subjects, we parcellated the brain into 100 cortical regions (Schaefer 100x7 atlas (Schaefer et al., 2018)) and 16 subcortical ones (Melbourne Scale I atlas (Tian et al., 2020)).

Computing Functional Graphs

To calculate the functional graphs for each subject, we consolidated the data from the four sessions of each subject in the HCP and computed the pairwise correlations among all pairs of parcels. To form a binary functional graph, we placed an edge between any two parcels displaying a statistically significant correlation coefficient ($p < 0.01$, t-test for Pearson correlation coefficient).

Computing Subnetwork Graphs from Parcel-Level Graphs

Subnetwork graphs were computed by aggregating parcel-level binary graphs into graphs between 16 subnetworks. These subnetworks consist of the standard 7 resting-state subnetworks (Yeo et al., 2011) plus one subcortical subnetwork, separately for the left and right hemispheres. A subnetwork-level graph is then computed for each subject, whereby the weight of an edge from subnetwork i to j is the number of nodes in subnetwork i that connect to nodes in subnetwork j, normalized by the number of all possible edges between these subnetworks.

Computing Degree and Causal Flow

To determine the degree and causal flow of a node i in a *binary* directed graph, its in-degree (number of edges pointing toward node i) and out-degree (number of edges originating from node i) are first computed and normalized by the total number of nodes in the graph. The degree of node i is then computed as the sum of the out-degree and in-degree, while the causal flow is obtained by subtracting the in-degree from the out-degree. The same process is followed for weighted graphs except that the calculation of in-degree and out-degree involves a weighted mean. Mathematically,

$$Causal\ Flow\ (i) = \frac{1}{N}\sum_{j=1}^{N} G(i,j) - \frac{1}{N}\sum_{j=1}^{N} G(j,i) \quad , \quad i = 1, 2, ..., N$$

$$Degree\ (i) = \frac{1}{N}\sum_{j=1}^{N} G(i,j) + \frac{1}{N}\sum_{j=1}^{N} G(j,i) \quad , \quad i = 1, 2, ..., N$$

where G denotes the graph's (binary or weighted) adjacency matrix.

Computing Global Network Measures

Density. This provides an overall measure of connectivity or density within the graph. While this measure in its definition cannot distinguish between a few edges with very large weights and many edges with smaller weights in the subnetwork graphs, since the weight of each edge in subnetwork graph reflects the number of parcels connecting the subnetworks, this density measure serves as a useful representation of the graph's general connectivity.

Shortest Path Length (PL). It is a measure of how efficiently information can travel across a network. It is computed by calculating the shortest path between all pairs of nodes, where the shortest path is defined as the minimum sum of edge weights connecting the nodes. We calculated this measure for each subnetwork graph using the NetworkX Python package (Hagberg et al., 2008), which efficiently computes the shortest paths for weighted graphs and averages them to produce a global measure of connectivity. It represents how well-connected the brain is. A lower average shortest path length indicates more efficient communication across the whole brain, meaning information can travel more quickly between nodes (subnetworks) (Milgram, 1967; Rubinov and Sporns, 2010). Conversely, a higher average shortest path length suggests less efficient connectivity, where information requires more steps to traverse between nodes (subnetworks).

Assortativity. It is a measure of the tendency of nodes in a network to connect to other nodes that are similar to themselves in some attribute, such as node degree or edge weight. In weighted networks, assortativity quantifies the correlation between the weights of edges connecting nodes. Positive assortativity

indicates that nodes are more likely to connect to others with similar attributes (e.g., similar node degrees or edge weights), while negative assortativity suggests that nodes with dissimilar attributes are more likely to be connected (Newman, 2002, 2003). We computed the degree assortativity coefficient for each subnetwork graph using a function from the NetworkX Python package (Hagberg et al., 2008). This function calculates the correlation between the degrees of connected nodes, specifically measuring degree assortativity. For weighted networks, we used the weight='weight' parameter, which ensures that the edge weights are taken into account when calculating the degree of each node. When applied to a weighted graph, the degree of a node is defined by the sum of the weights of the edges connected to it (i.e., the weighted degree). The assortativity coefficient then measures the correlation between the weighted degrees of pairs of connected nodes. This allows us to assess whether nodes with higher edge weights are more likely to be connected to other nodes with similarly high edge weights, providing insights into the subnetwork's structure. A positive assortativity coefficient suggests that nodes with higher weighted degrees tend to connect to each other, while a negative coefficient suggests that nodes with dissimilar weighted degrees are more likely to be connected.

Modularity. It is a measure of the strength of division of a network into communities, quantifying the difference between the observed density of edges within communities and the expected density in a random graph. A higher modularity value indicates a stronger community structure, where nodes within a community are more densely connected to each other than to nodes outside the community (Newman, 2006).. To compute the modularity for each subnetwork graph, we first converted the graph into an undirected format, as modularity optimization requires an undirected graph. After transforming the graph, we used the greedy modularity optimization algorithm to detect communities. This algorithm partitions the network into communities by maximizing the modularity score, which reflects the quality of the community structure. Finally, we calculated the modularity value for each subnetwork graph, which measures how well the nodes within each detected community are connected compared to what would be expected in a random graph with the same degree distribution. The resulting modularity score gives us an indication of the network's community structure. A higher modularity value suggests that the subnetwork has a more significant division into communities with dense intra-community connections and fewer connections between communities (Newman, 2006).

Spectral Radius. It is a global measure of network structure related to the spread of activity across the network. It is computed as the largest eigenvalue of the connectivity and represents the critical coupling strength required for synchronization. As the primary eigenvalue, the spectral radius provides insights into the structural properties, dynamical behavior, and stability of the underlying network. In network-based models of brain dynamics, the spectral radius has been linked to how easily the system can shift into an excited state. To compute the spectral radius for each subnetwork graph, we first calculated the eigenvalues of the weighted adjacency matrix. The spectral radius was determined by identifying the largest absolute eigenvalue from these eigenvalues. A higher spectral radius suggests a stronger, more dominant network structure, with greater potential for synchronization and transitions into excited state (Meghanathan, 2014; van Dam and Kooij, 2007; Wang et al., 2015, 2003).

Synchronizability. It is a measure of how easily a network can synchronize its components, reflecting the stability and collective behavior of the network when nodes attempt to synchronize (Arenas et al., 2008). To compute synchronizability for each subnetwork graph, we first calculated the Laplacian matrix of the directed graph, which was computed using the NetworkX Python package (Hagberg et al., 2008). This matrix captures the network's structural properties and the relationships between nodes. After computing the Laplacian matrix, we calculated its eigenvalues and sorted them in ascending order. Synchronizability is then assessed as the ratio of the second smallest eigenvalue to the largest eigenvalue of the Laplacian matrix. A higher value of this ratio indicates that the network is more easily synchronized, with less resistance to synchronization, as reflected by a low second smallest eigenvalue (Tang et al., 2014). We computed this ratio for each subnetwork graph, which provides insight into the network's ability to reach a synchronized state.

Computing Correlations Between Neurofeedback Scores and Causal Connectomes

We represented the strength of each parcel-level edge using the partial correlation values from CaLLTiF's causal graphs. The partial correlation value for each edge in the parcel-level causal summary graph (computed by temporal aggregation) was calculated as the partial correlation at the lag with the maximum absolute value, preserving its sign. We then condensed the original parcel-level graphs (116×116 matrix) into subnetwork-level graphs (16×16 matrix) by calculating a normalized edge weight for each pair of subnetworks. Specifically, for each pair of subnetworks, we summed the weights of all edges connecting parcels between the two subnetworks in the parcel-level partial correlation graph. To account for differences in parcel counts between subnetworks, we normalized this sum by dividing it by the total number of possible edges connecting those subnetworks. This normalization provided a consistent measure of connectivity strength between subnetworks, regardless of their size. Across all graphs from various studies, subjects, sessions, and runs, we compiled sequences of these edge strengths. For neurofeedback scores, we calculated an average by taking the mean of feedback samples reported during neurofeedback sessions for each subject. Since each causal graph was derived from fMRI data spanning two runs, we averaged feedback scores from these runs to align them with each causal graph. Finally, we computed Spearman correlations between edge strengths and average neurofeedback scores for each possible edge in the subnetwork graph. After applying FDR correction for multiple comparisons across all the edges, we retained the edges that showed significant correlations with neurofeedback scores for further analysis.

Computing

All the computations reported in this study were performed on a Lenovo P620 workstation with AMD 3970X 32-Core processor, Nvidia GeForce RTX 2080 GPU, and 512GB of RAM.

Additional Information

Author Contributions

EN designed and supervised the study; FA performed all analyses; AG assisted with the design and interpretations of the causal discovery algorithm; HJ and MAKP assisted in the analyses of human fMRI data; FA and EN drafted and all authors edited the manuscript.

Acknowledgments

The research conducted in this study was partially supported by NSF Award #2239654 to EN, by the Canadian Institute for Advanced Research (fellowship awarded to MAKP), and by the Air Force Office of Scientific Research under award number FA9550-20-1-0106 (to MAKP).

Competing financial interests

The authors declare no competing financial interests.

Data Availability Statement

All the fMRI data used in this work is publicly available. The fMRI data from DecNef studies can be accessed upon request (Cortese et al., 2021).

Code Availability Statement

The Python code for this study is publicly available at `https://github.com/nozarilab/2024DecNef_Causal_Connectome`.

References

Aharon, I., Etcoff, N., Ariely, D., Chabris, C. F., O'connor, E., and Breiter, H. C. (2001). Beautiful faces have variable reward value: fmri and behavioral evidence. *Neuron*, 32(3):537–551.

Amano, K., Shibata, K., Kawato, M., Sasaki, Y., and Watanabe, T. (2016). Learning to associate orientation with color in early visual areas by associative decoded fmri neurofeedback. *Current Biology*, 26(14):1861–1866.

Arab, F., Ghassami, A., Jamalabadi, H., Peters, M. A., and Nozari, E. (2023). Whole-brain causal discovery using fmri. *bioRxiv*, pages 2023–08.

Arenas, A., Díaz-Guilera, A., Kurths, J., Moreno, Y., and Zhou, C. (2008). Synchronization in complex networks. *Physics reports*, 469(3):93–153.

Barnett, L. and Seth, A. K. (2014). The mvgc multivariate granger causality toolbox: A new approach to granger-causal inference. *Journal of Neuroscience Methods*, 223:50–68.

Cavanna, A. E. and Trimble, M. R. (2006). The precuneus: a review of its functional anatomy and behavioural correlates. *Brain*, 129(3):564–583.

Chatterjee, A., Thomas, A., Smith, S. E., and Aguirre, G. K. (2009). The neural response to facial attractiveness. *Neuropsychology*, 23(2):135.

Chiba, T., Kanazawa, T., Koizumi, A., Ide, K., Taschereau-Dumouchel, V., Boku, S., Hishimoto, A., Shirakawa, M., Sora, I., Lau, H., et al. (2019). Current status of neurofeedback for post-traumatic stress disorder: a systematic review and the possibility of decoded neurofeedback. *Frontiers in human neuroscience*, 13:448460.

Ciric, R., Wolf, D. H., Power, J. D., Roalf, D. R., Baum, G. L., Ruparel, K., Shinohara, R. T., Elliott, M. A., Eickhoff, S. B., Davatzikos, C., et al. (2017). Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. *Neuroimage*, 154:174–187.

Cole, M. W., Reynolds, J. R., Power, J. D., Repovs, G., Anticevic, A., and Braver, T. S. (2013). Multi-task connectivity reveals flexible hubs for adaptive task control. *Nature neuroscience*, 16(9):1348–1355.

Cortese, A., Amano, K., Koizumi, A., Kawato, M., and Lau, H. (2016). Multivoxel neurofeedback selectively modulates confidence without changing perceptual performance. *Nature communications*, 7(1):13669.

Cortese, A., Tanaka, S. C., Amano, K., Koizumi, A., Lau, H., Sasaki, Y., Shibata, K., Taschereau-Dumouchel, V., Watanabe, T., and Kawato, M. (2021). The decnef collection, fmri data from closed-loop decoded neurofeedback experiments. *Scientific data*, 8(1):65.

Emmert, K., Kopel, R., Sulzer, J., Brühl, A. B., Berman, B. D., Linden, D. E., Horovitz, S. G., Breimhorst, M., Caria, A., Frank, S., et al. (2016). Meta-analysis of real-time fmri neurofeedback studies using individual participant data: How is brain regulation mediated? *Neuroimage*, 124:806–812.

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., et al. (2019). fmriprep: a robust preprocessing pipeline for functional mri. *Nature methods*, 16(1):111–116.

Fleming, S. M., Dolan, R. J., and Frith, C. D. (2012). Metacognition: computation, biology and function.

Friston, K. J., Kahan, J., Biswal, B., and Razi, A. (2014). A dcm for resting state fmri. *NeuroImage*, 94:396–407.

Granger, C. W. J. (1969). Investigatinag causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438.

Guntupalli, J. S., Hanke, M., Halchenko, Y. O., Connolly, A. C., Ramadge, P. J., and Haxby, J. V. (2016). A model of representational spaces in human cortex. *Cerebral cortex*, 26(6):2919–2934.

Hagberg, A., Swart, P. J., and Schult, D. A. (2008). Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Laboratory (LANL), Los Alamos, NM (United States).

Haugg, A., Sladky, R., Skouras, S., McDonald, A., Craddock, C., Kirschner, M., Herdener, M., Koush, Y., Papoutsi, M., Keynan, J. N., et al. (2020). Can we predict real-time fmri neurofeedback learning success from pretraining brain activity? *Human brain mapping*, 41(14):3839–3854.

Haxby, J. V., Guntupalli, J. S., Connolly, A. C., Halchenko, Y. O., Conroy, B. R., Gobbini, M. I., Hanke, M., and Ramadge, P. J. (2011). A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, 72(2):404–416.

Iaria, G., Fox, C. J., Waite, C. T., Aharon, I., and Barton, J. J. (2008). The contribution of the fusiform gyrus and superior temporal sulcus in processing facial attractiveness: neuropsychological and neuroimaging evidence. *Neuroscience*, 155(2):409–422.

Kamitani, Y. and Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature neuroscience*, 8(5):679–685.

Kepecs, A. and Mainen, Z. F. (2012). A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594):1322–1337.

Koizumi, A., Amano, K., Cortese, A., Shibata, K., Yoshida, W., Seymour, B., Kawato, M., and Lau, H. (2016). Fear reduction without fear through reinforcement of neural activity that bypasses conscious exposure. *Nature human behaviour*, 1(1):0006.

Koizumi, A., Maniscalco, B., and Lau, H. (2015). Does perceptual confidence facilitate cognitive control? *Attention, Perception, & Psychophysics*, 77:1295–1306.

Leech, R. and Sharp, D. J. (2014). The role of the posterior cingulate cortex in cognition and disease. *Brain*, 137(1):12–32.

Meghanathan, N. (2014). Spectral radius as a measure of variation in node degree for complex network graphs. In *2014 7th International Conference on u-and e-Service, Science and Technology*, pages 30–33. IEEE.

Meyniel, F., Sigman, M., and Mainen, Z. F. (2015). Confidence as bayesian probability: From neural origins to behavior. *Neuron*, 88(1):78–92.

Milgram, S. (1967). The small world problem. *Psychology today*, 2(1):60–67.

Newman, M. E. (2002). Assortative mixing in networks. *Physical review letters*, 89(20):208701.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, 45(2):167–256.

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582.

Norman, K. A., Polyn, S. M., Detre, G. J., and Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fmri data. *Trends in cognitive sciences*, 10(9):424–430.

Oblak, E. F., Lewis-Peacock, J. A., and Sulzer, J. S. (2017). Self-regulation strategy, feedback timing and hemodynamic properties modulate learning in a simulated fmri neurofeedback environment. *PLoS computational biology*, 13(7):e1005681.

Oblak, E. F., Sulzer, J. S., and Lewis-Peacock, J. A. (2019). A simulation-based approach to improve decoded neurofeedback performance. *NeuroImage*, 195:300–310.

Pereira, D. J., Morais, S., Sayal, A., Pereira, J., Meneses, S., Areias, G., Direito, B., Macedo, A., and Castelo-Branco, M. (2024). Neurofeedback training of executive function in autism spectrum disorder: distinct effects on brain activity levels and compensatory connectivity changes. *Journal of Neurodevelopmental Disorders*, 16(1):1–15.

Rounis, E., Maniscalco, B., Rothwell, J. C., Passingham, R. E., and Lau, H. (2010). Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. *Cognitive neuroscience*, 1(3):165–175.

Rubinov, M. and Sporns, O. (2010). Complex network measures of brain connectivity: uses and interpretations. *Neuroimage*, 52(3):1059–1069.

Said, C. P., Haxby, J. V., and Todorov, A. (2011). Brain systems for assessing the affective value of faces. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1571):1660–1670.

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X.-N., Holmes, A. J., Eickhoff, S. B., and Yeo, B. T. T. (2018). Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri. *Cerebral cortex*, 28(9):3095–3114.

Seeley, W. W., Menon, V., Schatzberg, A. F., Keller, J., Glover, G. H., Kenna, H., Reiss, A. L., and Greicius, M. D. (2007). Dissociable intrinsic connectivity networks for salience processing and executive control. *Journal of neuroscience*, 27(9):2349–2356.

Sepulveda, P., Sitaram, R., Rana, M., Montalba, C., Tejos, C., and Ruiz, S. (2016). How feedback, motor imagery, and reward influence brain self-regulation using real-time fmri. *Human brain mapping*, 37(9):3153–3171.

Shibata, K., Lisi, G., Cortese, A., Watanabe, T., Sasaki, Y., and Kawato, M. (2019). Toward a comprehensive understanding of the neural mechanisms of decoded neurofeedback. *NeuroImage*, 188:539–556.

Shibata, K., Watanabe, T., Kawato, M., and Sasaki, Y. (2016). Differential activation patterns in the same brain region led to opposite emotional states. *PLoS biology*, 14(9):e1002546.

Shibata, K., Watanabe, T., Sasaki, Y., and Kawato, M. (2011). Perceptual learning incepted by decoded fmri neurofeedback without stimulus presentation. *science*, 334(6061):1413–1415.

Simons, J. S., Peers, P. V., Mazuz, Y. S., Berryhill, M. E., and Olson, I. R. (2010). Dissociation between memory accuracy and memory confidence following bilateral parietal lesions. *Cerebral cortex*, 20(2):479–485.

Skottnik, L., Sorger, B., Kamp, T., Linden, D., and Goebel, R. (2019). Success and failure of controlling the real-time functional magnetic resonance imaging neurofeedback signal are reflected in the striatum. *Brain and Behavior*, 9(3):e01240.

Tang, Y., Qian, F., Gao, H., and Kurths, J. (2014). Synchronization in complex networks and its application–a survey of recent advances and challenges. *Annual Reviews in Control*, 38(2):184–198.

Taschereau-Dumouchel, V., Cortese, A., Chiba, T., Knotts, J., Kawato, M., and Lau, H. (2018). Towards an unconscious neural reinforcement intervention for common fears. *Proceedings of the National Academy of Sciences*, 115(13):3470–3475.

Taschereau-Dumouchel, V., Cortese, A., Lau, H., and Kawato, M. (2021). Conducting decoded neurofeedback studies. *Social Cognitive and Affective Neuroscience*, 16(8):838–848.

Taschereau-Dumouchel, V., Kawato, M., and Lau, H. (2020). Multivoxel pattern analysis reveals dissociations between subjective fear and its physiological correlates. *Molecular psychiatry*, 25(10):2342–2354.

Tian, Y., Margulies, D. S., Breakspear, M., and Zalesky, A. (2020). Hierarchical organization of the human subcortex unveiled with functional connectivity gradients. *bioRxiv*.

van Dam, E. R. and Kooij, R. (2007). The minimal spectral radius of graphs with a given diameter. *Linear Algebra and its Applications*, 423(2-3):408–419.

Wang, R., Zhang, Z.-Z., Ma, J., Yang, Y., Lin, P., and Wu, Y. (2015). Spectral properties of the temporal evolution of brain network structure. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(12).

Wang, Y., Chakrabarti, D., Wang, C., and Faloutsos, C. (2003). Epidemic spreading in real networks: An eigenvalue viewpoint. In *22nd International Symposium on Reliable Distributed Systems, 2003. Proceedings.*, pages 25–34. IEEE.

Weiskopf, N., Mathiak, K., Bock, S. W., Scharnowski, F., Veit, R., Grodd, W., Goebel, R., and Birbaumer, N. (2004). Principles of a brain-computer interface (bci) based on real-time functional magnetic resonance imaging (fmri). *IEEE transactions on biomedical engineering*, 51(6):966–970.

Wilimzig, C., Tsuchiya, N., Fahle, M., Einhäuser, W., and Koch, C. (2008). Spatial attention increases performance but not subjective confidence in a discrimination task. *Journal of vision*, 8(5):7–7.

Yamada, T., Hashimoto, R.-i., Yahata, N., Ichikawa, N., Yoshihara, Y., Okamoto, Y., Kato, N., Takahashi, H., and Kawato, M. (2017). Resting-state functional connectivity-based biomarkers and functional mri-based neurofeedback for psychiatric disorders: a challenge for developing theranostic biomarkers. *International Journal of Neuropsychopharmacology*, 20(10):769–781.

Yeo, T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., Roffman, J. L., Smoller, J. W., Zöllei, L., Polimeni, J. R., Fischl, B., Liu, H., and Buckner, R. L. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology*, 106(3):1125–1165.