




Dual-stream cortical pathways mediate sensory prediction

Qian Chu ^{1,2,5,†}, Ou Ma^{2,3,†}, Yuqi Hang ^{2,4}, Xing Tian ^{1,2,3,*}

¹Shanghai Frontiers Science Center of Artificial Intelligence and Deep Learning, Division of Arts and Sciences, New York University Shanghai, Shanghai 200126, China,

²NYU-ECNU Institute of Brain and Cognitive Science at NYU Shanghai, Shanghai 200062, China,

³Shanghai Key Laboratory of Brain Functional Genomics (Ministry of Education), School of Psychology and Cognitive Science, East China Normal University, Shanghai 200062, China,

⁴Department of Administration, Leadership, and Technology, Steinhardt School of Culture, Education, and Human Development, New York University, New York, NY 10003, United States,

⁵Present address: Max Planck-University of Toronto Centre for Neural Science and Technology, Toronto, ON M5S 2E4, Canada.

*Corresponding author: Shanghai Frontiers Science Center of Artificial Intelligence and Deep Learning, Division of Arts and Sciences, New York University Shanghai, Shanghai 200126, China. Email: xing.tian@nyu.edu

†These authors contributed equally

Predictions are constantly generated from diverse sources to optimize cognitive functions in the ever-changing environment. However, the neural origin and generation process of top-down induced prediction remain elusive. We hypothesized that motor-based and memory-based predictions are mediated by distinct descending networks from motor and memory systems to the sensory cortices. Using functional magnetic resonance imaging (fMRI) and a dual imagery paradigm, we found that motor and memory upstream systems activated the auditory cortex in a content-specific manner. Moreover, the inferior and posterior parts of the parietal lobe differentially relayed predictive signals in motor-to-sensory and memory-to-sensory networks. Dynamic causal modeling of directed connectivity revealed selective enabling and modulation of connections that mediate top-down sensory prediction and ground the distinctive neurocognitive basis of predictive processing.

Key words: predictive processing; descending connections; sensorimotor integration; episodic memory; mental imagery.

Introduction

Generating predictions is a trait of adaptive organisms to efficiently interact with the environment (Conant and Ashby 1970; Schultz et al. 1997; Friston 2010). For example, a seminal trend in cognitive neuroscience considers perception to depend on dynamic predictions based on the internal models of external world (Rao and Ballard 1999; Bar 2007; de Lange et al. 2018). In contrast to the “ascending” information flow from sensory to nonsensory areas, coordinated “descending” projections from nonsensory to sensory areas provide a neural substrate for conveying top-down sensory predictions (Mumford 1992; Rao 1999; Rao and Ballard 1999; Bastos et al. 2012; Shipp 2016; Keller and Mrosovsky 2018).

How descending projections convey predictive signals in the human brain remains enigmatic. Theoretically, the action-perception loop that links an agent’s cognitive system and the environment necessitates multiple forms of predictions. One category of predictions is motor based. According to theories of motor control, the agent could use a copy of the endogenous motor command and a model of action-consequence coupling to predict the sensory consequences of actions (Wolpert and Ghahramani 2000; Shadmehr et al. 2010; McNamee and Wolpert 2019). Motor-based predictions could be used for world state estimation (Wolpert et al. 1995), and the resulting prediction error could drive immediate motor correction as well as long-term motor learning (Jordan and Keller 2020). Whereas, predictions that do not involve an agent’s actions are exemplified by the

suppression of neural response to statistically organized stimuli (e.g. structured sequences (Garrido et al. 2009; Todorovic et al. 2011) or associated pairs (Kok, Jehee, et al. 2012; Garner and Keller 2022)). Humans learn rich statistical regularities in the external world and utilize exogenous information by transforming memory traces into sensory predictions. The combination of motor-based and memory-based predictive algorithms constructs a dual-stream prediction model (DSPM) (Tian and Poeppel 2013; Tian et al. 2016)—motor and memory systems could reverse their traditionally assumed roles as receivers of sensory information to act as independent sources that provide endogenous and exogenous information for generating sensory prediction (Fig. 1a).

Methodological challenges also obstruct the investigation of the neural basis of prediction. This is partly because of the spatial-temporal overlapping between descending prediction and ascending input during perception (Keller and Mrosovsky 2018). Moreover, most studies investigate predictive processing by probing how prediction modulates perception, granting them only indirect access to descending predictive signals (Todorovic et al. 2011; Kok, Jehee, et al. 2012; Kok, Rahnev, et al. 2012). The perceptual modulation approach focuses on the local computation of prediction error in the sensory cortex. But this indirect assessment of predictive signals faces difficulty in revealing the neural origin and generation processes of descending predictions that constrain the cognitive computations as well as the neural implementation of predictive processing from a system perspective.

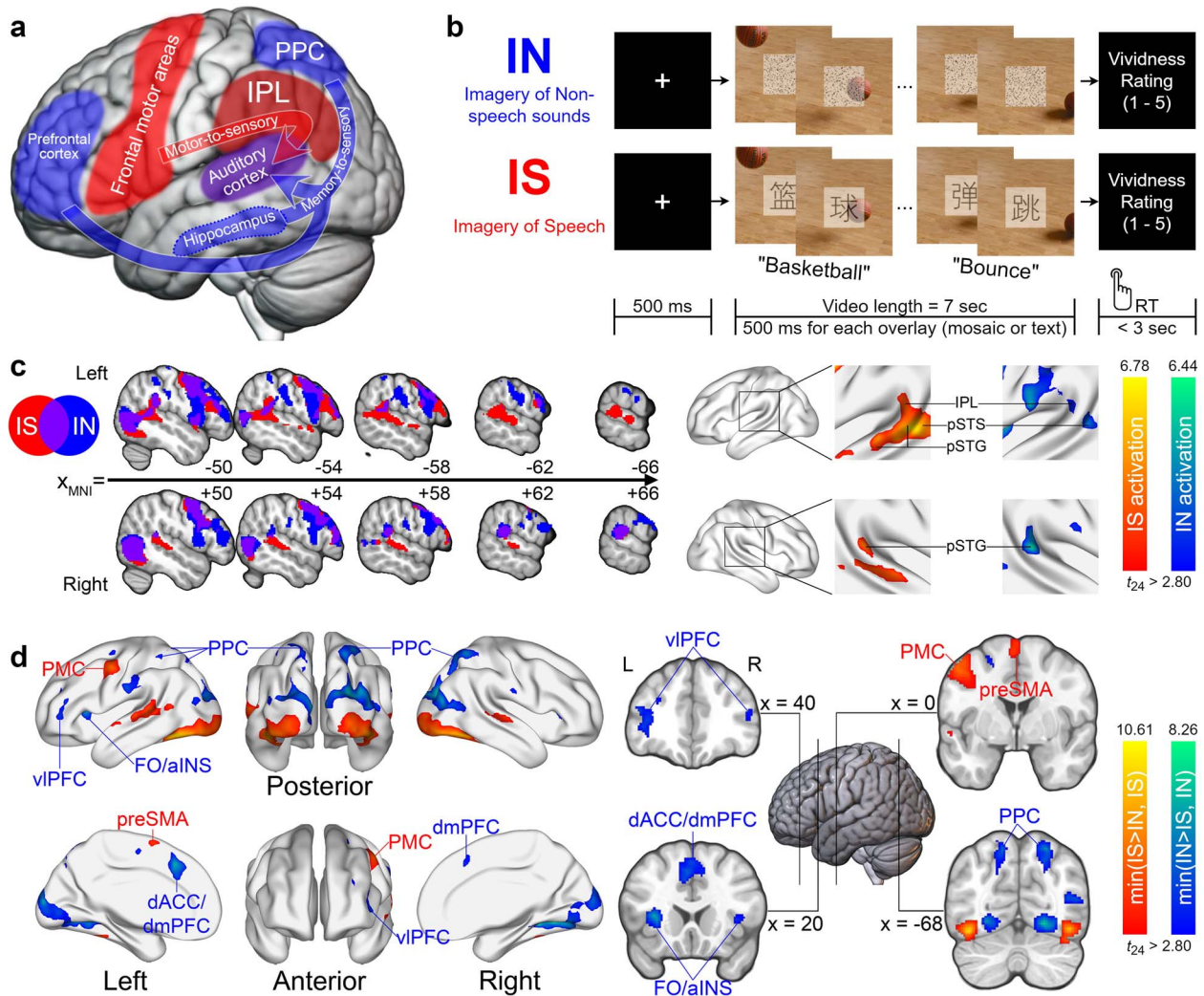


Fig. 1. Model of distinct pathways for generating prediction, experimental paradigms, and fMRI results of univariate analyses. a) The DSPM. The model posits that auditory representations in the temporal area can be established by two descending streams. The motor-to-sensory stream originates from the frontal motor network where speech plan encoding is carried out. A copy of the motor plan (efference copy), relaying via the IPL, establishes auditory representations in the auditory cortex to predict the sensory consequence of speech action. The memory-to-sensory stream, originating in a distributed memory network, including the prefrontal cortex, hippocampus, and superior parietal lobe, reconstructs auditory representations in the auditory system via memory retrieval. b) Experimental paradigm. Following a 500-ms fixation period, participants watched a muted video of objects in motion (frames from the bouncing basketball video are used for illustration). Participants were asked to imagine sounds ought to be in the video (e.g. the whomp of a basketball hitting the floor repeatedly) in the IN condition and imagine saying characters superimposed on the video in the IS condition. c) Activations in the inferior parietal and superior temporal regions during IS and IN. Top: Activations in the left hemisphere. Bottom: Activations in the right hemisphere. Left: The mosaic. Colored voxels were activated significantly in IS (red), IN (blue), or both (purple). Right: Thresholded surface rendering with t-value indicated by the color bar. See also Supplementary Fig. S1. d) Thresholded surface rendering showing the conjunctions (minimum statistic) between (i) IS > IN and IS and (ii) IN > IS and IN. IS induced stronger activations in the left PMC and preSMA, whereas IN induced stronger activations in the bilateral fronto-parietal and CONS.

Mental imagery serves as a promising paradigm for directly scrutinizing what and how descending projections convey predictive signals. Imagery, a cognitive capacity to endogenously create episodic mental states (Langland 2020), has been widely reported to elicit perceptual-like neural representations (Zatorre et al. 1996; Kosslyn et al. 1999; O'Craven and Kanwisher 2000; Bunzeck et al. 2005; Kraemer et al. 2005; Hubbard 2010) that arise from top-down connectivity (Dentico et al. 2014; Dijkstra et al. 2017; Pearson 2019).

Definitions of sensory prediction commonly have 2 components: the top-down generation of perceptual representations, and the interaction between prediction and sensory afference. Because both sensory prediction and imagery reactivate perceptual representations of possible sensory outcomes, imagery has

been argued to be a mental realization of prediction (Moulton and Kosslyn 2009) and to exploit the same set of internal models as implemented in predictive processing (Langland-Hassan 2016; Williams 2021). In the perspective of prediction-afference interaction, mental imagery has also been shown to modulate perceptual processing in a wide range of modalities similar as prediction does (Tian and Poeppel 2013, 2015; Kilteni et al. 2018; Tian et al. 2018; Ma and Tian 2019). Imagery thus exhibits significant parallels to prediction in both aspects.

Therefore, we leveraged mental imagery to investigate descending projections that establish auditory representations in the absence of confounding ascending signals to trace the neural origin of predictions. Moreover, our novel dual-imagery paradigm maximized the differences between motor-based and

memory-based prediction as participants were asked to imagine speech or natural sounds that human articulators cannot produce (Fig. 1b). The DSPM model and preliminary empirical findings (Tian and Poeppel 2010; Tian et al. 2016; Ma and Tian 2019; Li et al. 2020) derive three major experimental predictions. First, both motor-based and memory-based predictions in different types of imagery would reactivate the auditory cortex without external acoustic stimulation. Second, the upstream networks for generating sensory predictions should be distinct. Motor-based imagery would activate the frontal motor network, whereas memory-based imagery would involve the frontal-parietal and hippocampal networks. Third, and most importantly, information would flow directionally from motor or memory upstream systems to auditory areas in distinct functional descending networks that mediate the generation of prediction. The parietal lobe in particular would relay descending projections, with posterior parietal cortex (PPC) subserving memory-based prediction (Dijkstra et al. 2017; Sestieri et al. 2017) and inferior parietal lobe (IPL) as a sensorimotor interface in speech (Hickok and Poeppel 2007; Hickok 2012) subserving motor-based prediction (Tian and Poeppel 2010; Tian et al. 2016; Li et al. 2020). These hypotheses were tested using fMRI. We performed whole-brain statistical parametric mapping and multivariate pattern analysis to identify regions of interest (ROIs) for subsequent dynamic causal modeling (DCM) of directed (i.e. effective) connectivity. We obtained evidence that supported our hypotheses and revealed the origin, structure, and endpoint of dual-stream descending connections in generating predictions.

Materials and methods

Ethics statement

The experimental protocol was approved by the Institutional Review Board at New York University Shanghai (IRB00009975/FWA#00022531) in accordance with policies and regulations found in The Common Rule (45 CFR part 46).

Participants

Twenty-nine right-handed, native Mandarin speakers participated in the experiment with informed consent and received monetary incentives. No participant reported a history of neurological or psychological illness. All participants had normal or corrected-to-normal vision. Data from four participants were removed from analyses due to excessive head motion (>5 mm in any session) or drowsiness during scanning. The remaining 25 participants were included in the analyses (12 females; mean age \pm SD = 21.3 \pm 2.3).

Materials

Ten different 7 s video clips with their corresponding audio tracks were selected and used as the stimuli in the experiment. All video clips were about scenes or objects and none of them contained human speech. Examples included a basketball bouncing on the wooden floor, a train quickly passing by, and a ringing telephone. Our motivation was to choose videos with sounds that were hard to simulate with human vocal organs but easy to imagine with the aid of visual scenes. Every 500 ms, a square image patch was superimposed on the center of the video, making a total of 14 patches. These images were either Chinese characters (black, against a white background) constituting a sentence that described the content of the video (e.g. “一个篮球在木质地板上反复弹跳”; “A basketball bounces on the wooden floor over and over”; see [Supplementary Table S1](#) for sentences describing all 10 videos), or mosaics made by randomly

shuffling pixels of the Chinese characters, thus serving as nonlinguistic visual controls that share equal net luminance as the character images. We also created synthesized speech of the sentences in a male’s voice using the VoiceGen toolbox (<https://github.com/ray306/VoiceGen>).

Procedure

We presented participants with 12 sessions of videos following a structural scanning session. During the first three sessions, participants were presented with videos with the original audio tracks with mosaics overlaid on them. We refer to this condition as “Hearing of Nonspeech” sounds (HN). Each “HN” session consisted of 22 trials, which included two catch trials featuring a pure tone (frequency = 1,000 Hz, duration = 715 ms) played at a random time point of a random video. The other 20 trials consisted of 10 videos each played twice in random order. After watching each video, participants were asked to report if they heard the pure tone in the video by pressing button 1 (for yes) or button 2 (for no) on an MRI-compatible response pad. HN condition was designed for localizing auditory areas and allowed participants to encode auditory memory of the nonspeech sounds for later retrieval.

Three sessions of “Imagery of Nonspeech” (IN) followed. In these sessions, videos were muted, and mosaics were overlaid in the center. Participants were instructed to imagine the sounds they heard during the preceding HN sessions, thus inducing memory-based auditory reactivation. Participants rated the vividness of imagery (rating range = 1–5) with the response pad at the end of each trial. This visually aided imagery of the nonspeech task was similar to previous studies (Bunzeck et al. 2005). Thereafter, came three sessions of “Imagery of Speech” (IS) where the videos were also muted and Chinese characters were overlaid on the videos. Participants were instructed to imagine saying the characters and gave a vividness rating afterward. “IS” was hypothesized to also recruit auditory representations but through a motor-based pathway. Every IS or IN session consisted of 20 videos with each of the 10 videos randomly played twice.

The task in the last three sessions was “Hearing of Speech” (HS), which was designed to localize auditory areas responsive to verbal stimuli. During the video presentation, the original audio track was replaced with synthesized speech. Similar to the HN sessions, two catch trials were included in each “HS” session in which two nearby characters in the synthesized speech were reversed (e.g. 鞭炮 to 炮鞭; firecracker to “crackerfire”). Participants indicated whether they heard a reversal using the response pad in a similar manner as in HN sessions.

Trials in every session shared a similar procedure, starting with a 500-ms fixation period, followed by 7 s of video presentation, a button response from the participant for vividness rating (IN and IS) or catch trial detection (HN and HS), and an intertrial interval of either 4.44 or 6.66 s (2 or 3 repetition times [TRs] for fMRI scanning) minus the response time for rating or detection. The participants were asked to make the button response within 3 s. A trial would be considered invalid if the participant did not respond in the time limit or made an incorrect response for catch trial in HN and HS. Invalid trials were separately modeled and thus excluded from formal analyses.

The order of sessions (HN–IN–IS–HS) was designed to simplify instructions while reducing confounds. The HN and IN sessions allowed encoding and subsequent retrieval of nonspeech sounds, and they proceeded IS and HS such that participants were less likely to perform IS during IN. IS proceeded HS because there otherwise existed an alternative strategy for participants to retrieve their memory of the synthesized speech they had listened to,

maximally separating the putative motor-based and memory-based mechanisms for perceptual prediction.

fMRI data acquisition

MRI images were collected on a Siemens MAGNETOM Prisma System (Erlangen, Germany) at East China Normal University. Anatomical images were acquired using a T1-weighted magnetization-prepared rapid acquisition gradient echo sequence (192 sagittal slices; field of view [FOV]=240 mm × 240 mm; flip angle [FA]=8°; TR=2,300 ms; echo time [TE]=2,320 ms; voxel size=0.9375 × 0.9375 × 0.9000 mm³). Functional images were acquired using a T2*-weighted echo-planar imaging pulse sequence (38 even-first interleaved slices; FOV = 192 mm × 192 mm; FA = 81°; TR/TE = 2,220/30 ms; voxel size = 3.0 × 3.0 × 3.6 mm³; interslice gap = 0.6 mm). Functional slices were oriented to an approximately 30° tilt toward coronal from AC–PC alignment to maximize the coverage of individual brain volumes.

Preprocessing

Preprocessing of fMRI data and subsequent analyses were implemented via SPM12 (<https://www.fil.ion.ucl.ac.uk/spm/>, version 7771) and custom-written scripts with MATLAB R2021a (MathWorks Inc., Natick, MA, United States). Preprocessing followed the standard procedure in SPM12.

All functional images from each participant were temporally interpolated to the first slice of each volume and were spatially realigned to the mean image. The structural image was coregistered with functional images. For univariate and connectivity analyses, functional images were then spatially normalized to the Montreal Neurological Institute (MNI) standard brain space (resampled voxel size = 2 mm isotropic) and were smoothed with a 6-mm full width, half maximum (FWHM) Gaussian kernel. For multivoxel pattern analysis (MVPA), the functional images were neither normalized nor smoothed to preserve information patterns in the individual's native brain space.

Univariate analysis

Events were modeled as sustained boxcar epochs spanning their corresponding duration. They included the presentation of fixation points, videos (in which participants performed the imagery and hearing tasks), instructions for vividness rating or catch trial detection, and button presses. Events in catch trials, no-response, or wrong-response trials, were modeled separately to improve the model sensitivity. All events were convolved with a canonical hemodynamic response function implemented in SPM12 and were entered as regressors into a general linear model (GLM) for each individual. Each GLM also included head motion regressors and session-wise baseline regressors. The GLM was then estimated using functional data high-pass filtered at 1/128 Hz. Individual-level contrasts were constructed using the beta estimates of regressors of interest and were subject to a one-sample t-test for group-level inference. For whole-brain mapping throughout the paper, we used cluster-wise $P_{FDR} < 0.05$ to determine the statistical significance, where clusters were defined with an uncorrected threshold of $P = 0.005$.

To examine common activation in imagery and comparable hearing conditions, we used a conjunction (i.e. minimum statistics) approach (Nichols et al. 2005). We obtained thresholded ($t_{24} > 2.80$, $P < 0.005$, $P_{FDR} < 0.05$) t-value maps from imagery and hearing conditions (IS and HS, IN and HN), computed the smaller t-value of the 2 conditions for each voxel, and reported only voxels that were significant ($t_{24} > 2.80$) after the operation. Similarly, to

examine differential activations in IS and IN, we took the minimum t-value from the IS and IS > IN contrasts as well as IN and IN > IS contrasts. Therefore, all significant voxels showed both significant activities during one type of imagery and significant difference over the other.

Multivoxel pattern analysis

To test whether the activated areas represent imagery contents, we trained support vector machine (SVM)-based classifiers to decode the imagery associated with 10 video categories in IS and IN (MVPA) using The Decoding Toolbox (TDT, version 3.999E) (Hebart et al. 2015). It is noteworthy that there exists another approach for characterizing brain-specific representations, namely the representational similarity analysis (RSA) (Kriegeskorte et al. 2008). We deem it less suitable for the present study in that (i) RSA relies on a prespecified representational dissimilarity matrix derived from experimental conditions, usually measured by physical properties of behavior or stimulus. Yet, it is impossible to determine such dissimilarity, given the nature of imagery as a purely subjective experience. (ii) Even if we compute dissimilarity from the physical counterparts of imagery (e.g. articulatory or acoustic, as in Zhang et al. 2020) as an approximation, the long-duration imagery contents (7 s) and the slow dynamics of BOLD signals would likely make the resulting video-wise dissimilarity matrix insensitive to crucial features that unfold over time. Therefore, we adopted MVPA to utilize spatial information to decode imagery free of constraints on any representational dimension.

One additional participant was excluded from MVPA due to his lack of response to the coin video in all HS sessions, making the sample size $n = 24$. Beta estimates of each video category from all three sessions in imagery (IS and IN) or hearing (HS and HN) conditions were obtained and were used to train and test a L2-norm SVM available through LIBSVM (Chang and Lin 2011). We used a regularization parameter $C = 1$ and scaled the data at a range of 0–1. To efficiently test which voxels across the brain could be used for accurate classification, we moved spherical searchlights (Kriegeskorte et al. 2006) throughout the brain. To avoid the choice of radius from biasing our results, we conducted searchlight analyses with varying radii from 1 to 8 voxels. The accuracy maps obtained using a radius of four voxels were visualized as surface renderings.

To decode video categories within a condition, we used a leave-one-session-out crossvalidation scheme. In each decoding step, two out of three sessions in the condition were used to train an SVM classifier, and the remaining session was used as test data to decode the 10 video categories from multivoxel patterns. The average classification accuracy from all three decoding steps, each having a different test session and two corresponding training sessions, was calculated and assigned to the center voxel of the searchlight to generate a decoding accuracy map.

To decode video categories across IS and IN, we used a two-way leave-one-session-out crossclassification scheme. Similar to the previous scheme, two sessions from the IS condition were used to train a classifier, but the test data, this time, were one session from the IN condition. This procedure was iterated for all three IN sessions, each using a different combination of two IS sessions as training data. Next, the IS sessions were used as test data for classifiers trained with IN sessions. A cross-classification accuracy map was generated using the average accuracy obtained from a total of 6 decoding steps.

To perform group-level inference, we normalized the individual-level accuracy maps into MNI space and smoothed them with

a 6-mm FWHM Gaussian kernel to account for the individual neuroanatomical difference. These accuracy maps were then brought to one-sample t-tests, and the mean accuracy level of each significant cluster (voxel-wise threshold: $P < 0.005$; cluster-wise threshold: $P_{FDR} < 0.05$) was displayed. For better visualization, the range of data for display was controlled at 10%–20% (0%–10% above the chance level of 10%) because accuracy $> 20\%$ was mostly observed in visual areas. To validate the significance of decoding in specific ROIs across searchlight radii, we also performed Wilcoxon signed-rank tests with decoding accuracy from each MNI coordinate from each patient. These nonparametric tests over specific regions complemented the parametric t-tests for whole-brain mapping.

Time series extraction from ROIs

Based on univariate and MVPA results, we selected ROIs that showed content-specific activations. For each ROI, voxel-wise time courses in IS and IN were high-pass filtered at 1/128 Hz and the estimated effects of nonimagery regressors (e.g. fixation cue, button press, and head motion) were subtracted out. This adjustment should increase model sensitivity in the connectivity analysis by excluding activities induced by nonimagery events. The resulting first principal component of each ROI was used for DCM analyses.

Dynamic causal modeling

To test our central hypothesis about the descending projections in generating prediction, we used DCM (Friston et al. 2003), a well-established method that allows the inference of directional brain connectivity modulated by an experimental condition (IS and IN in the present study). DCM features a neuronal state equation, which is coupled to a biophysically plausible model to explain BOLD signals. We used the bilinear DCM that features the following state equation:

$$\dot{z} = (A + Bu)z + Cu, \quad (1)$$

where z denotes the hidden neural activity from all ROIs, and the dot notation denotes change per unit time. The A matrix represents baseline connectivity in the absence of external stimulation. The B matrix represents the modulatory effects of an experimental input u (IS or IN in the present study) on connectivity between regions. The C matrix represents the direct driving effect of each u on neuronal activity.

We first specified motor-to-sensory and memory-to-sensory network models for IS and IN, respectively. Each imagery condition could drive the activity of a brain area (C matrix) or modulate cortico-cortical connectivity between areas (B matrix). We specified an all-1s A matrix (i.e. enabled baseline connectivity between every ROI pair) for both motor-to-sensory and memory-to-sensory models because we did not have any prior hypothesis regarding baseline connectivity. Enabled parameters had Gaussian priors with zero mean and non-zero variance, while the others had zero variance. The neural activity z was coupled with a biophysically informed forward model (Friston et al. 2003; Zeidman, Jafarian, Corbin, et al. 2019) to predict the BOLD time series. The standard (single-state, deterministic) DCM was used. A slice timing model was used in alignment with the slice timing correction performed during preprocessing.

For subject-level model inversion, our goal was to find parameter estimates that maximize log model evidence. DCM uses a variational Laplace scheme to approximate model evidence with negative variational free energy (Friston et al. 2007). This

estimation scheme also penalizes model complexity calculated as the Kullback–Leibler divergence between the priors and the posteriors. Thus, DCM evaluates how well the model achieves a trade-off between accuracy and complexity.

The expected parameter values and the posterior covariances at the subject level were then brought to a parametric empirical Bayes (PEB) analysis to make inferences about the group-level effects (Friston et al. 2016; Zeidman, Jafarian, Corbin, et al. 2019; Zeidman, Jafarian, Seghier, et al. 2019). In terms of the between-subject design matrix, since our experimental design involves no between-subject factors, the design matrix was simply an all-1s vector $X = [1 \ 1 \ \dots \ 1 \ 1]^T$ to model commonalities across subjects. In addition, random effects (unexplained between-subject variability) on parameters were assumed to account for individual differences.

Having estimated parameters of the motor-to-sensory and memory-to-sensory full models and specified candidate reduced models by “switching off” some parameters, we then performed Bayesian model reduction (BMR) (Friston et al. 2016) to analytically derive the evidence and parameters of the reduced models. We compared the evidence of each reduced model to find the winning model as well as pooled evidence of models belonging to each model family. We also plotted parameters that had positive evidence (posterior probability, $P_p > 0.75$) of being present versus absent, assessed by the Bayesian model average (BMA) on all reduced models.

After mapping out the functional motor-to-sensory and memory-to-sensory descending networks using IS and IN data, respectively, we tested the hypothesis that the two motor-to-sensory and memory-to-sensory networks were differentially implemented in two types of imagery with 2 approaches.

First, we “swapped” the data-model combination by reinverting the full motor-to-sensory and memory-to-sensory DCMs using data from the other imagery condition. That is, we used IN data (BOLD time series and imagery events in the condition) as input to the specified motor-to-sensory DCM and used IS data for memory-to-sensory DCM. We hypothesized a difference in the explained variance and/or parameter estimate between the pairs of DCMs with the same model structure but fitted with different data. This approach keeps the model structure as specified in the previous separate DCMs. The explained variance from the pairs of DCMs was subject to a two-sided Wilcoxon signed-rank test. For model parameter estimates, because they correspond to a multivariate Gaussian density, we computed each parameter’s mean and variance with the leading diagonal of the covariance matrix. To compare the posterior distributions yielded by a model with different data, we performed z-tests using the mean and variance of each parameter estimate.

Second, we designed a “fully mixed” model with all motor (left PMC), memory (bilateral PPC), and sensory ROIs (left IPL and bilateral pSTG). In the full model, IS and IN modulated both motor-to-sensory (left PMC to left IPL then to bilateral pSTG) and memory-to-sensory (bilateral PPC to left IPL and bilateral pSTG) connections. We ran Bayesian model comparison (BMC) to compare the full model with reduced models where (i) IS specifically modulates motor-to-sensory connections and only drives left PMC, (ii) IN specifically modulates memory-to-sensory connections and only drives bilateral PPC, and (iii) the combination of (i) and (ii) a.k.a. no mixing at all. If the reduced models explain the data more efficiently as measured by free energy, it will support the distinctness of the two networks, as enabling modulatory effects of 1 imagery condition on the noncorresponding connections would not explain the data better. This approach,

however, requires fitting a DCM with a larger parameter space (6 nodes, 36 A parameters, 26 B parameters, and 6 C parameters) and thus might be vulnerable to underfitting.

Results

Behavioral results

The completion and success of mental imagery are hard to assess behaviorally because imagery is an internal experience. We relied on the timeliness of the participants' vividness report to infer whether they performed the imagery tasks instructed. Participants actively engaged in IN and IS as the response rate of vividness rating after each trial was at ceiling (mean = 98.20%, SD = 3.23%). Mean vividness score in IS (mean = 3.53, SD = 0.48) was significantly larger than that in IN (mean = 2.90, SD = 0.60) as revealed by a two-sided paired *t*-test ($t_{24} = 5.57, P = 10^{-5}$). Accuracy of detecting the catch trials (hit and correct rejection) was also high in both HN (mean = 98%, SD = 3.49%) and HS (mean = 95%, SD = 5.17%).

Common activations in auditory cortices accompanied by differential motor and memory activations

To test the hypothesis that the auditory cortex is activated as the sensory endpoints of descending signaling, we first carried out whole-brain univariate analyses. We found overlapping activations in both IS and IN in the bilateral posterior part of superior temporal gyri and sulci (pSTG and pSTS). The common activation in both IS and IN also extended to the left IPL that anatomically covered parts of parietal operculum, posterior supramarginal gyrus, and planum temporale (Fig. 1c; for whole-brain surface rendering, see Supplementary Fig. S1). In IS, activations also extended to left anterior STG (aSTG), which is consistent with previous findings of aSTG harboring higher-level linguistic representations (e.g. phonemes and words; DeWitt and Rauschecker 2012). Activations at pSTG and IPL were observed in the hearing conditions (Supplementary Fig. S2), further supporting that these regions mediate auditory-like representations.

Next, we contrasted IS with IN to examine differential activations that would likely distinguish upstream networks underlying prediction generation (Fig. 1d). We took a minimum statistics approach (Nichols et al. 2005) to select voxels that showed both significant activity during 1 type of imagery and significant difference over the other (e.g. IS > IN masked with IS activations). IS induced stronger effects than IN in the frontal motor network, including the left premotor cortex (PMC) and presupplementary motor area (preSMA). IN activated the frontoparietal network comprising the left ventrolateral prefrontal cortex (vlPFC) and bilateral PPC and activated the cingulo-opercular network (CON) comprising the dorsal anterior cingulate cortex (dACC) and bilateral frontal operculum/anterior insular (FO/aINS).

Motor, memory, and auditory systems represent imagery contents

High decoding accuracy observed in the visual cortex demonstrated the validity of our decoding method since the videos differed in visual stimulation. Moreover, we found above-chance accuracy (chance level = 10%) in bilateral pSTG and left IPL in both IS (Fig. 2a), IN (Fig. 2b), and comparable hearing conditions (Supplementary Fig. S3). These results support our hypothesis that specific auditory representations were activated in a top-down manner as auditory endpoints in the descending networks.

Consistent with univariate results, significant decoding of videos was found in the left PMC in IS. This decoding of imagery contents in the frontal motor region without participants' overt movement suggests a motor representation space in the motor upstream network (Fig. 2a).

For IN, decoding accuracy was significantly above chance in bilateral PPC, but not in vlPFC nor in the CON (Fig. 2b). Despite significant decoding observed in bilateral PPC during IS, two-sided paired *t*-tests (for whole-brain mapping) and Wilcoxon signed-rank tests (for data from ROIs) showed that the decoding accuracy in parts of PPC (left intraparietal sulcus and right superior parietal lobule) was significantly higher in IN than that in IS reliably across searchlight radii (Fig. 2c), suggesting memory representations in PPC in addition to putatively visual representations commonly available in both conditions (confirmed by a cross-classification analysis, Supplementary Fig. S4).

Putting together the univariate and MVPA results, the selective activations and content specificity of PMC in IS, PPC in IN, and the auditory cortex in both conditions supported our first hypothesis of common sensory endpoint and our second hypothesis of differential upstream systems for motor-based and memory-based predictions. We next tested our last hypothesis about the descending structures mediating the two types of predictions by examining the cortico-cortical connectivity with DCM.

Motor-to-sensory and memory-to-sensory networks assessed by DCM

For connectivity analyses, we selected ROIs based on univariate and MVPA results. The representative voxel coordinate of each ROI and their associated *t*-values for each contrast are reported in Supplementary Table S2 and all selected voxels are visualized in Fig. 3a. Our criteria are summarized below. For auditory ROIs, we selected areas that showed increased BOLD magnitude and representational patterns in both IS and IN as well as hearing conditions, leading to our choice of left pSTG (sphere center $x = -50, y = -46, z = 12$) and its right homolog (sphere center $x = 62, y = -36, z = 18$). Given its consistent appearance revealed by multiple analyses, left IPL (sphere center $x = -54, y = -38, z = 24$) was also selected to test whether it serves as a mediating hub for motor-to-sensory and/or memory-to-sensory descending networks. As for the motor ROI, we included left PMC (sphere center $x = -38, y = 0, z = 36$) based on its significantly higher activity during IS than IN and its content-selective pattern during IS. All ROIs mentioned above consisted of gray matter voxels within spheres with a radius of 4 mm. Specifically, the small radius ensured that the left pSTG and left IPL ROIs, despite their spatial proximity (Euclidean distance = 14.97 mm), had no shared voxels nor smoothing-induced (FWHM = 6 mm) data contamination. Left and right PPCs were selected as memory ROIs, and due to their being large and nonspherical clusters, we used the conjunction of the following contrasts to select all PPC voxels that showed significant effects: IN, IN > IS, IN MVPA, and IN > IS MVPA. The resulting left PPC ROI entailed 120 voxels (centroid $x = -20, y = -72, z = 40$) and right PPC ROI entailed 548 voxels (centroid $x = 24, y = -60, z = 54$).

We used DCM to model activities in these ROIs to infer effective connectivity between the motor, memory, and sensory nodes, which could potentially mediate two distinct types of prediction. We set IS as the driving input to left PMC in the motor-to-sensory model and IN as the driving input to bilateral PPC in the memory-to-sensory model. Most importantly, we specified a priori imagery-modulated connections that reflect changes in connectivity during a specific mental operation.

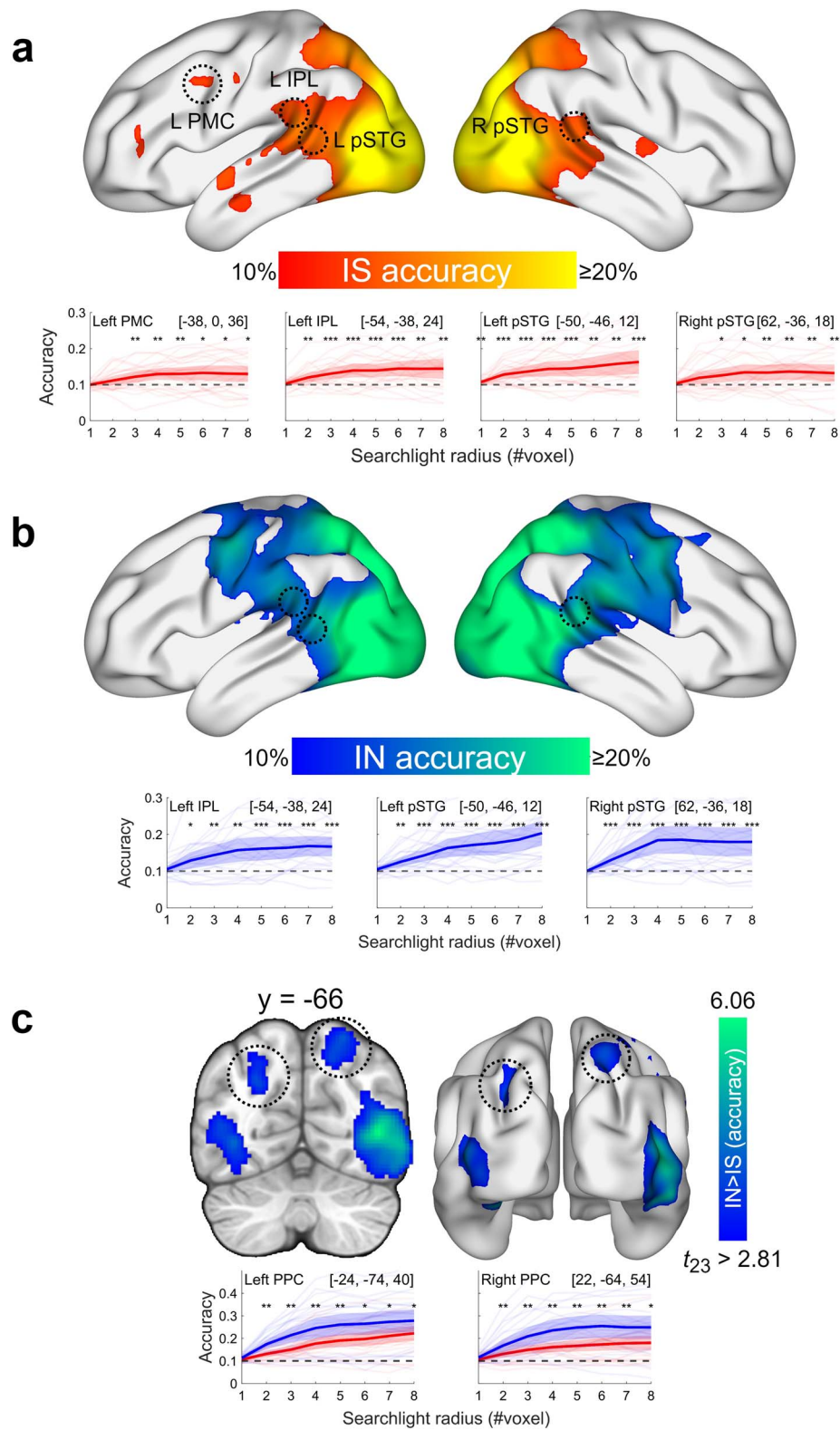


Fig. 2. Results of MVPA. a) Decoding of video categories in IS. Top: Thresholded surface rendering of decoding accuracy using a moving searchlight with a radius of 4 voxels. Bottom: Decoding accuracy at ROIs across different radii (1–8 voxels). The triplet numbers in brackets denote MNI coordinate of the searchlight center. Asterisks denote significance level of decoding accuracy above-chance level (10%) evaluated by a Wilcoxon signed-rank test. b) Similar to (a) but for classification in IN. c) Top: A coronal view and a surface rendering of areas showing higher decoding accuracy in IN than IS. Bottom: Classifier performance in bilateral PPC during IS and IN across searchlight radii. Asterisks denote the significance level of decoding accuracy higher in IN than IS evaluated by a Wilcoxon signed-rank test. For all panels, error bars indicate 95% confidence interval. * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$.

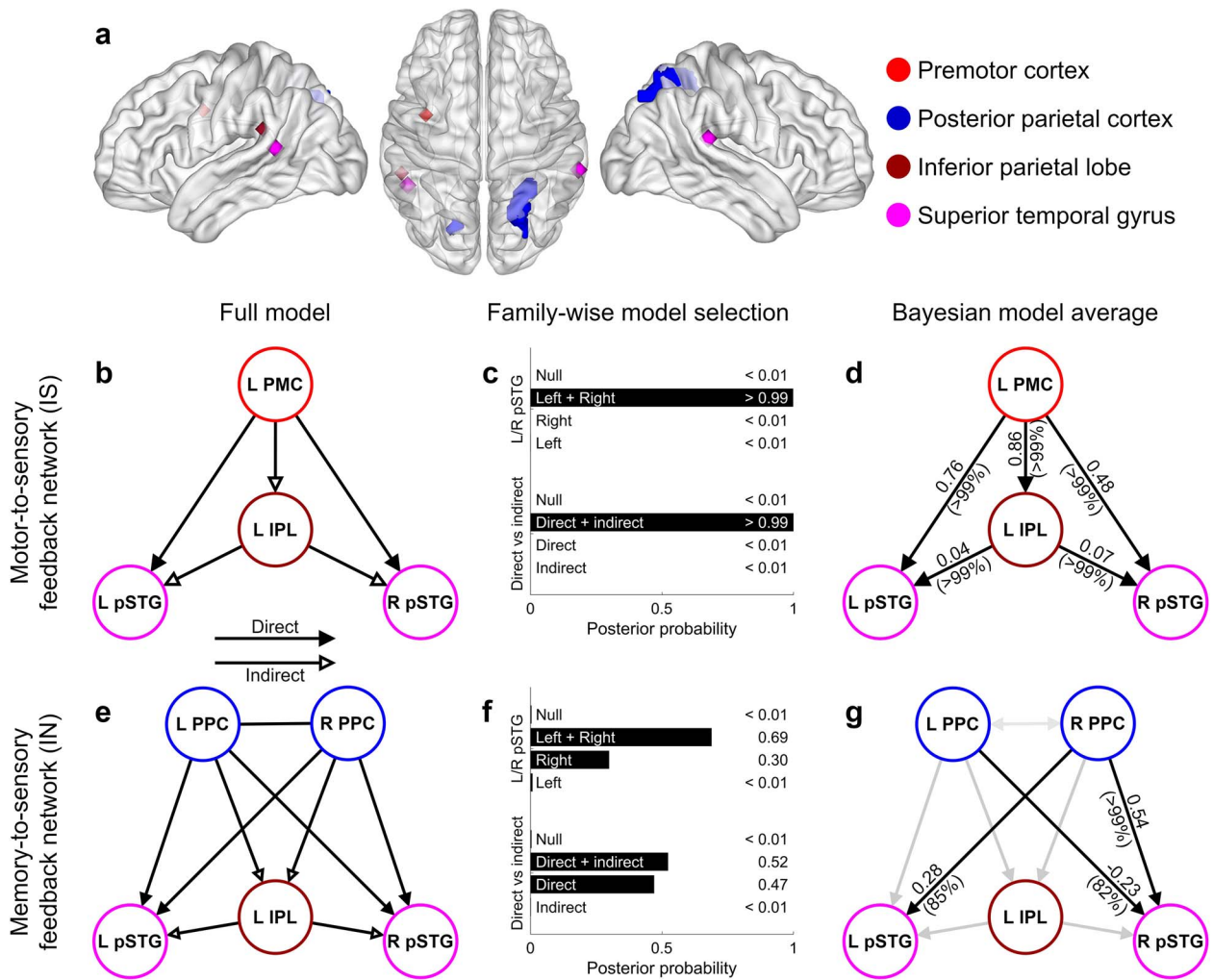


Fig. 3. Motor-to-sensory and memory-to-sensory descending networks assessed by DCM. a) ROIs used for DCM. ROIs were spherical with a radius of 4 mm except for bilateral PPC ROIs, which were selected based on contrast conjunctions. b) Graphical illustration of the full model of the motor-to-sensory descending network. The arrows pertain to connections modulated by IS. The baseline connections and driving inputs are not presented in the plot. c) Family-wise BMC of the two motor-to-sensory descending model factors (descending structure and sensory endpoints). Numbers on the right denote posterior probability. d) BMA of effective connectivity parameters for the motor-to-sensory descending network. Parameters that reached the significance level of posterior probability (Pp) > 0.75 were shown in black and otherwise in gray. Numbers out of parentheses denote parameter estimate in the unit of Hertz and numbers in parentheses denote posterior probability. e–g) DCM results similar as (b–d) but for the memory-to-sensory descending network.

To construct a full motor-to-sensory DCM, we allowed IS to modulate 5 connections: “direct” connections from left PMC to bilateral pSTG, and “indirect” connections from left PMC to left IPL and then to bilateral pSTG (Fig. 3b). We then constructed 11 reduced models with a subset of these connections “switched off” according to two factors concerning the descending architecture (direct-only/indirect-only/direct and indirect/null) and auditory endpoint (left pSTG/right pSTG/left and right pSTG/null). The null model contained no modulated descending connection and thus offers the null hypothesis. A graphical illustration of all reduced models is shown in Supplementary Fig. S5.

Under the BMR scheme (Friston et al. 2016; Zeidman, Jafarian, Corbin, et al. 2019; Zeidman, Jafarian, Seghier, et al. 2019), the free energy (lower bound on model evidence) (Friston et al. 2007) of each reduced model was derived. This allowed us to perform BMC to systematically infer whether motor-to-sensory connections were enhanced in IS, and if so, through what route (direct vs. indirect) and in which hemisphere they ended. BMC returned the single winning model to be the full model itself with a posterior

probability (Pp) > 0.99. We pooled reduced models according to the two factors to perform family-wise Bayesian model selection (Fig. 3c), which revealed that the hybrid architecture entailing both direct and indirect connections to bilateral pSTG was the most likely (Pp > 0.99 for both families). We then summarized model parameters across all models by taking the weighted average of parameters from each model with the weight determined by each model’s Pp, an approach known as BMA (Jennifer et al. 1999). The BMA results (Fig. 3d) confirmed the essence of all 5 IS-modulated connections that all had a positive mean and Pp > 0.99. All these results together suggest a motor-to-sensory descending architecture originating at the left PMC, mediated by left IPL, and ending at bilateral pSTG during IS.

A similar procedure was applied to construct and evaluate memory-to-sensory DCMs using data from the IN session. This DCM model specified entailed IN-modulated connections between bilateral PPC, “direct” connections from bilateral PPC to pSTG, and “indirect” connections from PPC to left IPL and then to bilateral pSTG (Fig. 3e); 112 reduced models were constructed

according to 4 factors: descending origin (left PPC/right PPC/left and right PPC), auditory endpoint (left pSTG/right pSTG/left and right pSTG/null), descending architecture (direct-only/indirect-only/direct and indirect/null), and PPC mutual connection (present/absent). A graphical illustration of key connections in reduced models is shown in [Supplementary Fig. S6](#).

BMC over reduced models of the memory-to-sensory DCM showed that the most probable (despite the relatively low $P_p=0.18$) model entailed descending connections initiating from bilateral PPC (without mutual connection) to bilateral pSTG via both direct and indirect pathways. Results of family-wise model selection over the two most important factors are shown in [Fig. 3f](#). Regarding the descending architecture, evidence near equally supported the direct-only architecture ($P_p=0.47$) and the hybrid architecture with both direct and indirect connections ($P_p=0.52$). Bilateral PPC ($P_p=0.65$) was more probable than right PPC alone ($P_p=0.35$) to be the descending source, and bilateral pSTG ($P_p=0.69$) was more probable than right pSTG alone ($P_p=0.30$) to be the auditory endpoints. When summarizing individual parameter estimates using BMA, we found three significant ($P_p > 0.75$) connections along the direct route ([Fig. 3g](#)): left PPC to right pSTG (mean = -0.23 Hz, $P_p=0.82$); right PPC to left pSTG (mean = 0.28 Hz, $P_p=0.85$); and right PPC to right pSTG (mean = 0.54 Hz, $P_p > 0.99$). Overall, these results spoke for the existence of a memory-to-sensory projection from bilateral PPC to bilateral pSTG. IN modulated the left PPC to pSTG connection in an inhibitory manner while enhancing right PPC to pSTG connections, suggesting a hemispheric division of function. The lack of evidence in family-wise model selection and BMA did not support any mediating role of left IPL in memory-to-sensory transformation.

Distinct motor-to-sensory and memory-to-sensory descending networks in generating predictions

To test the functional distinctness of the motor-to-sensory and memory-to-sensory networks, we first “swapped” the data-model combination and then compared variance explained by the DCM as well as PEB parameter estimates in each model fitted with IS and IN data. To complement the analysis, we fitted a mixed model with all motor-to-sensory and memory-to-sensory connections modulated by both IS and IN.

We found that the motor-to-sensory DCM fitted with IS data yielded significantly higher explained variance (mean = 14.96%) than with IN data (mean = 6.86%), as revealed by a two-sided Wilcoxon-signed rank test ($P=0.01$, [Fig. 4a](#)). However, no significant difference in the mean parameter estimates ($P > 0.09$ for all five parameters, two-sided z-test) was found ([Fig. 4b](#)). These results suggest that the motor-to-sensory model cannot effectively explain IN data despite the fact that “forced” modeling fitting yielded similar parameter estimates.

On the other hand, we did not see a significant difference ($P=0.90$) in explained variance when fitting the memory-to-sensory DCM with IS and IN data (mean explained variance = 7.28 and 7.82) ([Fig. 4c](#)). The insignificance persisted after removing the obvious outlier ($P=0.63$). Significant difference in several PEB parameters was observed ([Fig. 4d](#)). Notably, the modulated connections from right PPC directly to bilateral pSTG were significantly higher in IN than in IS (right PPC to left pSTG, $P=0.033$; right PPC to right pSTG, $P < 0.001$). Such differences suggest that the memory-to-sensory architecture identified in the previous section does not explain activities during motor-based prediction. Whereas, several connections involving left

IPL in the indirect pathway yielded higher parameter estimates using IS data (left PPC to left IPL, $P < 0.001$; left IPL to right pSTG, $P=0.006$). These results were consistent with the indirect pathway found in the motor-to-sensory DCM, as left IPL exerted excitatory connectivity to pSTG even in a memory-to-sensory DCM where no motor node was included. These results also explained why there was no significant decrease in explained variance when fitting the memory-to-sensory DCM with IS data, as some pSTG activity might have been explained by IPL-exerted connectivity.

As for the mixed model ([Fig. 4e](#)), despite an expected low explained variance (4.59% , $SD=5.34\%$), we found that a reduced version of the model explained the data most efficiently with a posterior probability of 99.97% via BMC. The reduced model is a “non-mixing” one where IS specifically modulates motor-to-sensory (PMC to IPL to pSTG) connections and IN specifically modulates memory-to-sensory (PPC to pSTG) connections. That the no-mixing model outperforms the full model or partially mixed models is another set of evidence supporting that the modulatory effects of imagery pertain to the specific networks. As the explained variance for the mixed model is poor in the first place, this result should be interpreted with caution and would be better viewed as complementary results that are largely consistent with the findings from data-model swapping.

Taking together the results from swapping data-model combinations and comparing a mixed model against reduced nonmixing models, we showed that distinct functional motor-to-sensory and memory-to-sensory descending networks and different subregions of parietal lobe (IPL vs. PPC) mediate the generation of content-specific auditory representations in IS and IN.

Discussion

Our study using fMRI with a dual imagery paradigm complements the existing findings on prediction-perception interactions (e.g. predictive cancellation). Although previous studies have demonstrated that predictions modulate sensory processing even at the lowest levels, it is methodologically challenging to study the neural origins that convey descending predictions in the presence of simultaneous sensory inputs. With respect to this research focus, imagery highly resembles prediction in terms of generating top-down sensory representations ([Moulton and Kosslyn 2009](#); [Williams 2021](#)) and thus serves as a useful paradigm for investigating predictive processing. Therefore, with the aid of the novel imagery paradigm, we have characterized the neural implementation of sensory prediction via descending projections from motor and memory systems to the auditory cortex. Our results revealed the motor and memory systems as independent sources of prediction. The differential involvement of IPL and PPC in the motor-based and memory-based prediction pathways further suggests a functional division of the parietal lobe for routing the generation processes. The interareal communicative neural structures mediate distinct predictive processes via representational transformation, converging motor and memory information into sensory format for adaptive behavior.

Motor-based prediction originates from the PMC

Significant activity was observed in the left PMC in the motor-based prediction task of IS and its representational specificity was supported by MVPA ([Figs. 1 and 2](#)). These results are consistent with previous studies that stress PMC’s role in speech planning ([Castellucci et al. 2022](#)) as well as studies on speech imagery ([Tian et al. 2016](#); [Li et al. 2020](#); [Proix et al. 2022](#)). In terms of lateralization, left PMC was more engaged in speech

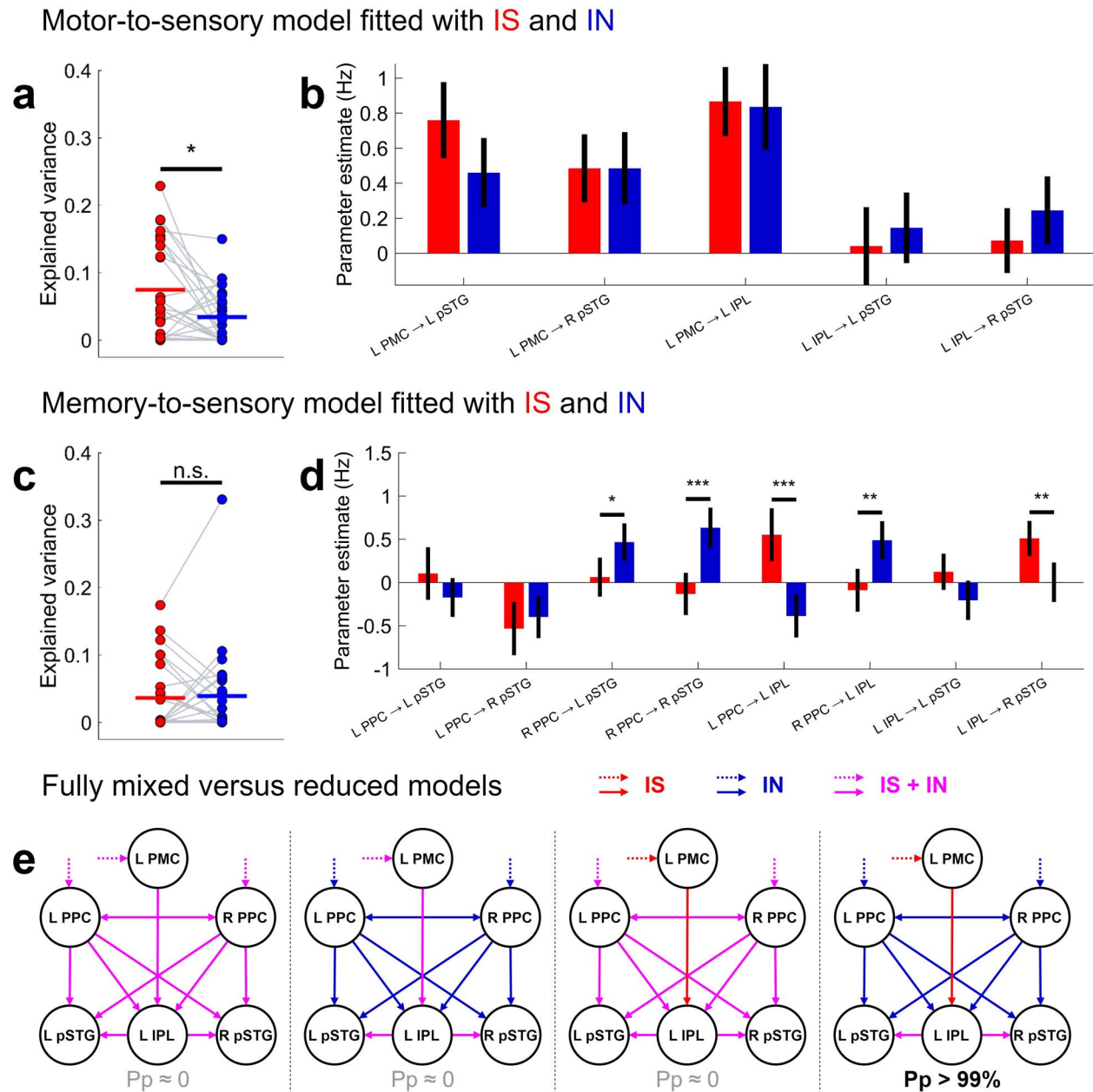


Fig. 4. Motor-to-sensory and memory-to-sensory networks differentially take place during IS and IN. **a)** Variance explained by the motor-to-sensory DCM. Red and blue circles denote results obtained with data from IS and IN sessions, respectively. Data points from each individual are joined by a gray line. The means of IS and IN data are indicated by thick solid lines. **b)** PEB estimates for all five imagery-modulated connections in the motor-to-sensory DCM. **c** and **d)** Similar to **a)** and **b)** but for explained variance and PEB estimates of the memory-to-sensory DCM. **e)** Results of BMC of fully mixed and reduced models. A full motor-to-sensory and memory-to-sensory mixed DCM that models concatenated data from IN and IS sessions (left-most illustration) was compared against 3 reduced versions of the full model (see main text for detailed specifications). Solid lines denote connections modulated by imagery conditions, and dotted lines denote the enabled driving effects of imagery conditions. Red, blue, and magenta represent the imagery conditions IS, IN, and both IS and IN. The strongest model evidence was obtained on the fourth model that comprised no mixing, suggesting the distinct motor-to-sensory and memory-to-sensory networks. Error bars indicate 95% confidence interval. * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$.

prediction. Crucially, the directed connectivity from PMC to pSTG was enhanced by motor-based imagery, thus revealing PMC's fundamental role as the upstream motor system in predicting the auditory consequence of speech. Two other common motor areas, preSMA and the inferior frontal gyrus (Brodmann areas 44 and 45, see [Supplementary Fig. S1](#)), were also activated. Yet, neither of them possessed significantly decodable representations.

Through the lens of predictive motor control, our results support the existence of the hypothesized efference copy signal ([Miall and Wolpert 1996](#); [Wolpert and Ghahramani 2000](#)). By definition, an efference copy is derived as a copy of the motor plan that

likely arises in PMC and conveys sensory prediction to the sensory cortex in a descending manner, which is the neural hierarchy we observed in motor imagery. In addition to the neural implementation, the functional implication of the efference copy is that it enables motor correction and state estimation. While the present study did not involve behavioral components to assess such functions, previous studies ([Tian and Poeppel 2013, 2015](#); [Kilteni et al. 2018](#); [Tian et al. 2018](#); [Ma and Tian 2019](#)) have shown that the brain compares motor-based sensory prediction with sensory afference and detects mismatch in basic features such as timing and intensity as well as phonological features. Overall,

by leveraging the imagery paradigm, we have provided evidence for both the descending neural architecture and functional properties of efference copy in speech.

The descending signal originating from the motor system can also be understood from the more contemporary framework of active inference (Adams et al. 2013; Parr et al. 2022). While the two frameworks commonly stress the importance of descending connections from the motor system to the sensory cortex, the interpretation of such connections are different (Friston 2011). Rather than viewing the descending predictive signal uniformly as a copy of the motor command as in predictive motor control, active inference divides predictions into proprioceptive and exteroceptive modalities. Proprioceptive predictions supply setpoints for motor reflexes in the spirit of the equilibrium point hypothesis (Feldman and Levin 1995; Feldman 2009), a.k.a. equivalent of motor commands. Exteroceptive (e.g. auditory) prediction on the other hand corresponds to the efference copy/corollary discharge and is integral to perceptual inference. It is worth noticing that since participants produced no overt speech in the present study, the activations we found over PMC would not correspond to the vocal motor command or proprioceptive predictions. Instead, they could support the generation about exteroceptive predictions, as discussed below.

IPL relays motor-to-sensory predictive signaling

Motor-to-sensory information flow from the PMC to pSTG was achieved by both direct and indirect routes (Fig. 3). The indirect route features IPL as a relaying hub (also referred to as the Sylvian parietal-temporal area). These results are consistent with previous reports of IPL activation in both speech perception and production (Buchsbaum et al. 2001; Hickok et al. 2003, 2009).

The intermediate step of IPL in the motor-based prediction generation route could be an auditory-motor interface and computes the transformation between motor and auditory representations (Hickok 2012). Alternatively, because movement of articulators yield speech, the computation of auditory prediction could be mediated by predicting the sensorimotor status of articulators (Tian and Poeppel 2010, 2012). Thus, the IPL could be an intermediate stage for an abstract somatosensory prediction in a functional continuum between the somatosensory regions in the anterior part of parietal lobe to the final auditory prediction starting in the posterior part of temporal lobe. Somatosensory prediction has been observed in the secondary somatosensory area and extending to IPL (Kilteni and Ehrsson 2020). In the speech domain, the partially redundant predictions in the sensorimotor and auditory domains may provide computational benefits of detecting distinct sources of noise.

PPC mediates memory-to-sensory predictive signaling

PPC was active in the memory-based prediction task of IN and harbored imagery-specific codes in IN (Figs. 1 and 2). While it is possible that the strong activation and decoding in PPC could be explained by additional visual attention to aid auditory retrieval, the DCM results ruled out the possibility as they further revealed enhanced connectivity between right PPC and bilateral STG, suggesting right PPC is the crucial origin of the memory-to-sensory prediction network (Fig. 3). The role of PPC in episodic memory has been demonstrated in a broad range of studies employing paradigms such as N-back (Owen et al. 2005; Barch et al. 2013), retention (Kwak and Curtis 2022) and memory search (Sestieri et al. 2014). Directed connectivity from PPC to the sensory cortex has also been found in visual imagery (Dentico et al. 2014;

Dijkstra et al. 2017). Altogether, these findings further support PPC as a general episodic buffer in generating memory-based prediction across memory tasks and modality.

Another interesting property is that the left PPC to STG connectivity is reduced instead of enhanced as observed in its right PPC to STG counterpart. This could be due to a hemispheric division of PPC in the auditory memory or a functional-anatomical division of PPC, as the left PPC ROI we selected is majorly composed of the intraparietal sulcus, while the right PPC ROI majorly consists of the superior parietal lobule.

The prefrontal cortex and hippocampus were less supported by empirical evidence to be the origin in the memory-to-sensory network as they lacked significantly decodable patterns. As the role of vIPFC and hippocampus in memory maintenance and memory-based prediction has been described in the literature (Davachi and DuBrow 2015; Kumar et al. 2016), the discrepancy may arise from the experimental design and analysis scheme. Throughout our analyses, we modeled the imagery events as sustained boxcar events. Since participants may recall the soundtrack of the videos immediately after their initiation appearance, vIPFC and hippocampus could support the initial retrieval of auditory memory through visual-auditory association, which is then transferred to PPC for maintenance. The interpretation is, however, hard to assess due to the low temporal resolution of fMRI.

Outside of DSPM, we also found that the CON (including FO/aINS and dACC/dmPFC) was more active in IN but lacked decodable multivoxel patterns. This is consistent with previous studies reporting CON to have a more modulatory rather than the representational role in memory (Sestieri et al. 2014; Wallis et al. 2015). Because our study focuses on representational transformations in descending projections, we did not include CON in DCM to avoid complicating the model. Yet, our data suggest CON may have a role in modulating the memory-based prediction and imagery.

Common auditory reactivation via different descending projections

Common activation in both motor-based and memory-based imagery in the auditory cortex agrees with previous work on musical imagery (Halpern and Zatorre 1999; Li et al. 2020), speech imagery (Tian et al. 2016; Proix et al. 2022), and imagery of complex sounds (Bunzeck et al. 2005). Imagery induced similar activations in the auditory cortices as hearing controls, supporting the nature of sensory-like representation as the ending result of prediction. The commonality in auditory reactivation in IS and IN further suggests a sensory convergence of predictions originating in different upstream networks. Together, the common sensory activations by hearing and types of imagery hint at a neuroanatomical foundation for the integration of various predictive and stimulus-driven signals in the sensory system. At a more microscopic level, such integration may take place in distinct neural subpopulations in the sensory system, which differentially respond to ascending sensation and descending prediction (Bastos et al. 2012). A likely laminar organization for such functional populations involves descending prediction sent to the deep layers of the sensory cortex (Rao and Ballard 1999; Kok et al. 2016). Further investigations adapting our paradigm can aim to test this specific hypothesis, which will shed light on the microcircuitry that integrates ascending input and descending prediction.

One potential limitation of the dual imagery paradigm is the coengagement of multiple cognitive processes that could be confounds to the results. First, although our design of

intersession order (HN-IN-IS-HS) discourages participants to imagine linguistic elements during HN and IN, it could be possible that participants recalled sounds and imagined speaking simultaneously. However, while it is possible that common activations in bilateral pSTG could be due to this simultaneous imagery alternative, our DCM results suggest these activations are largely due to differential descending connections. The confound could also downgrade the statistical power when detecting the difference between IN- and IS-related brain activities. Nevertheless, our results spoke for such differences and allowed us to locate putative motor and memory nodes that contributed to the interareal communications. Second, due to the nature of the IS task, simultaneous engagement of the language processing in addition to the motor-to-sensory transformation is also possible, which might explain the stronger activation over the aSTG in IS (Fig. 1c and d). However, we believe that the contribution of linguistic processing to the revealed motor-to-sensory network is minimal since we uncovered motor-to-sensory effective connectivity that bottom-up visual word form processing would not predict.

Conclusion

In conclusion, using a dual imagery paradigm with fMRI, we found that motor and memory systems project to the sensory system via distinct network structures to generate sensory predictions. The neural origin and interareal communicative structures constrain the computations of representational transformation, creating the emergent properties of the distinct predictive neural networks for efficiently linking cognition with environment.

Acknowledgments

We thank Zheng Li, Xiao Ma, Wenjia Zhang, Yidan Gao, Lechuan Wang, Hao Zhu, Rui Tong, and Jialin Chen for their help in experimental design, data collection and analysis.

Authors' contributions

OM and XT designed the experiment. OM collected the data. QC, OM, and YH performed the analyses and drafted the manuscript. All the authors reviewed and corrected the manuscript. XT supervised the project.

CRedit authors statement

Qian Chu (Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing—original draft, Writing—review & editing), Ou Ma (Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Writing—original draft, Writing—review & editing), Yuqi Hang (Formal analysis, Investigation, Methodology, Software, Validation, Writing—original draft, Writing—review & editing), and Xing Tian (Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Software, Supervision, Writing—original draft, Writing—review & editing)

Supplementary material

Supplementary material is available at *Cerebral Cortex* online.

Funding

This work was supported by the National Natural Science Foundation of China (32071099 and 32271101 to XT), Natural Science

Foundation of Shanghai (20ZR1472100 to XT), Program of Introducing Talents of Discipline to Universities (Base B16018 to XT), NYU Shanghai Boost Fund (to XT), and NYU Shanghai Dean's Undergraduate Research Fund (to QC).

Conflict of interest statement: None declared.

Data availability

All relevant data (including preprocessed functional images, event files, statistical maps, and dynamic causal models) and codes for replicating the key findings are available at <https://doi.org/10.17605/OSF.IO/7492E>.

References

- Adams RA, Shipp S, Friston KJ. Predictions not commands: active inference in the motor system. *Brain Struct Funct.* 2013;218(3):611–643.
- Bar M. The proactive brain: using analogies and associations to generate predictions. *Trends Cogn Sci.* 2007;11(7):280–289.
- Barch DM, Burgess GC, Harms MP, Petersen SE, Schlaggar BL, Corbetta M, Glasser MF, Curtiss S, Dixit S, Feldt C, et al. Function in the human connectome: task-fMRI and individual differences in behavior. *NeuroImage.* 2013;80:169–189.
- Bastos AM, Usrey WM, Adams Rick A, Mangun George R, Fries P, Friston KJ. Canonical microcircuits for predictive coding. *Neuron.* 2012;76(4):695–711.
- Buchsbaum BR, Hickok G, Humphries C. Role of left posterior superior temporal gyrus in phonological processing for speech perception and production. *Cogn Sci.* 2001;25(5):663–678.
- Bunzeck N, Wuestenberg T, Lutz K, Heinze H-J, Jancke L. Scanning silence: mental imagery of complex sounds. *NeuroImage.* 2005;26(4):1119–1127.
- Castellucci GA, Kovach CK, Howard MA, Greenlee JDW, Long MA. A speech planning network for interactive language use. *Nature.* 2022;602(7895):117–122.
- Chang C-C, Lin C-J. Libsvm: a library for support vector machines. *ACM Trans Intell Syst Technol.* 2011;2(3):Article 27.
- Conant RC, Ashby WR. Every good regulator of a system must be a model of that system. *Int J Syst Sci.* 1970;1(2):89–97.
- Davachi L, DuBrow S. How the hippocampus preserves order: the role of prediction and context. *Trends Cogn Sci.* 2015;19(2):92–99.
- de Lange FP, Heilbron M, Kok P. How do expectations shape perception? *Trends Cogn Sci.* 2018;22(9):764–779.
- Dentico D, Cheung BL, Chang J-Y, Guokas J, Boly M, Tononi G, Van Veen B. Reversal of cortical information flow during visual imagery as compared to visual perception. *NeuroImage.* 2014;100:237–243.
- DeWitt I, Rauschecker JP. Phoneme and word recognition in the auditory ventral stream. *Proc Natl Acad Sci.* 2012;109(8):E505–E514.
- Dijkstra N, Zeidman P, Ondobaka S, van Gerven MAJ, Friston K. Distinct top-down and bottom-up brain connectivity during visual perception and imagery. *Sci Rep.* 2017;7(1):5677.
- Feldman AG. New insights into action–perception coupling. *Exp Brain Res.* 2009;194(1):39–58.
- Feldman AG, Levin MF. The origin and use of positional frames of reference in motor control. *Behav Brain Sci.* 1995;18(4):723–744.
- Friston K. The free-energy principle: A unified brain theory? *Nat Rev Neurosci.* 2010;11(2):127–138.
- Friston K. What is optimal about motor control? *Neuron.* 2011;72(3):488–498.
- Friston KJ, Harrison L, Penny W. Dynamic causal modelling. *NeuroImage.* 2003;19(4):1273–1302.

- Friston KJ, Mattout J, Trujillo-Barreto N, Ashburner J, Penny W. Variational free energy and the Laplace approximation. *NeuroImage*. 2007;34(1):220–234.
- Friston KJ, Litvak V, Oswal A, Razi A, Stephan KE, van Wijk BCM, Ziegler G, Zeidman P. Bayesian model reduction and empirical bayes for group (DCM) studies. *NeuroImage*. 2016;128:413–431.
- Garner AR, Keller GB. A cortical circuit for audio-visual predictions. *Nat Neurosci*. 2022;25(1):98–105.
- Garrido MI, Kilner JM, Stephan KE, Friston KJ. The mismatch negativity: a review of underlying mechanisms. *Clin Neurophysiol*. 2009;120(3):453–463.
- Halpern AR, Zatorre RJ. When that tune runs through your head: a pet investigation of auditory imagery for familiar melodies. *Cereb Cortex*. 1999;9(7):697–704.
- Hebart MN, Görger K, Haynes J-D. The decoding toolbox (tdt): a versatile software package for multivariate analyses of functional imaging data. *Front Neuroinform*. 2015;8:88.
- Hickok G. Computational neuroanatomy of speech production. *Nat Rev Neurosci*. 2012;13(2):135–145.
- Hickok G, Poeppel D. The cortical organization of speech processing. *Nat Rev Neurosci*. 2007;8(5):393–402.
- Hickok G, Buchsbaum B, Humphries C, Muftuler T. Auditory–motor interaction revealed by fmri: speech, music, and working memory in area spt. *J Cogn Neurosci*. 2003;15(5):673–682.
- Hickok G, Okada K, Serences JT. Area spt in the human planum temporale supports sensory-motor integration for speech processing. *J Neurophysiol*. 2009;101(5):2725–2732.
- Hubbard TL. Auditory imagery: empirical findings. *Psychol Bull*. 2010;136(2):302–329.
- Jennifer AH, David M, Adrian ER, Chris TV. Bayesian model averaging: a tutorial. *Stat Sci*. 1999;14(4):382–417.
- Jordan R, Keller GB. Opposing influence of top-down and bottom-up input on excitatory layer 2/3 neurons in mouse primary visual cortex. *Neuron*. 2020;108(6):1194–1206.e1195.
- Keller GB, Mrsic-Flogel TD. Predictive processing: a canonical cortical computation. *Neuron*. 2018;100(2):424–435.
- Kilteni K, Ehrsson HH. Functional connectivity between the cerebellum and somatosensory areas implements the attenuation of self-generated touch. *J Neurosci*. 2020;40(4):894–906.
- Kilteni K, Andersson BJ, Houborg C, Ehrsson HH. Motor imagery involves predicting the sensory consequences of the imagined movement. *Nat Commun*. 2018;9(1):1617.
- Kok P, Jehee JCM, de Lange FP. Less is more: expectation sharpens representations in the primary visual cortex. *Neuron*. 2012;75(2):265–270.
- Kok P, Rahnev D, Jehee JFM, Lau HC, de Lange FP. Attention reverses the effect of prediction in silencing sensory signals. *Cereb Cortex*. 2012;22(9):2197–2206.
- Kok P, Bains Lauren J, van Mourik T, Norris David G, de Lange FP. Selective activation of the deep layers of the human primary visual cortex by top-down feedback. *Curr Biol*. 2016;26(3):371–376.
- Kosslyn SM, Pascual-Leone A, Felician O, Camposano S, Keenan JP, Thompson WL, Ganis G, Sukel KE, Alpert NM. The role of area 17 in visual imagery: convergent evidence from PET and rTMS. *Science*. 1999;284(5411):167–170.
- Kraemer DJ, Macrae CN, Green AE, Kelley WM. Musical imagery: sound of silence activates auditory cortex. *Nature*. 2005;434(7030):158.
- Kriegeskorte N, Goebel R, Bandettini P. Information-based functional brain mapping. *Proc Natl Acad Sci*. 2006;103(10):3863–3868.
- Kriegeskorte N, Mur M, Bandettini P. Representational similarity analysis—connecting the branches of systems neuroscience. *Front Syst Neurosci*. 2008;2.
- Kumar S, Joseph S, Gander PE, Barascud N, Halpern AR, Griffiths TD. A brain system for auditory working memory. *J Neurosci*. 2016;36(16):4492–4505.
- Kwak Y, Curtis CE. Unveiling the abstract format of mnemonic representations. *Neuron*. 2022;110(11):1822–1828.e5.
- Langland-Hassan P. *Explaining imagination*. New York, NY: Oxford University Press; 2020
- Langland-Hassan P. On choosing what to imagine. In: Kind A, Kung P, editors. *Knowledge through imagination*. Oxford: Oxford University Press; 2016. pp. 61–84
- Li Y, Luo H, Tian X. Mental operations in rhythm: motor-to-sensory transformation mediates imagined singing. *PLoS Biol*. 2020;18(10):e3000504.
- Ma O, Tian X. Distinct mechanisms of imagery differentially influence speech perception. *eneuro*. 2019;6(5):ENEURO.0261-0219.2019.
- McNamee D, Wolpert DM. Internal models in biological control. *Annu Rev Control Robot Auton Syst*. 2019;2(1):339–364.
- Miall RC, Wolpert DM. Forward models for physiological motor control. *Neural Netw*. 1996;9(8):1265–1279.
- Moulton ST, Kosslyn SM. Imagining predictions: mental imagery as mental emulation. *Philos Trans R Soc B Biol Sci*. 2009;364(1521):1273–1280.
- Mumford D. On the computational architecture of the neocortex. *Biol Cybern*. 1992;66(3):241–251.
- Nichols I, Brett M, Andersson J, Wager T, Poline J-B. Valid conjunction inference with the minimum statistic. *NeuroImage*. 2005;25(3):653–660.
- O’Craven KM, Kanwisher N. Mental imagery of faces and places activates corresponding stimulus-specific brain regions. *J Cogn Neurosci*. 2000;12(6):1013–1023.
- Owen AM, McMillan KM, Laird AR, Bullmore E. N-back working memory paradigm: a meta-analysis of normative functional neuroimaging studies. *Hum Brain Mapp*. 2005;25(1):46–59.
- Parr T, Pezzulo G, Friston KJ. *Active inference: the free energy principle in mind, brain, and behavior*. Cambridge, MA: The MIT Press; 2022.
- Pearson J. The human imagination: the cognitive neuroscience of visual mental imagery. *Nat Rev Neurosci*. 2019;20(10):624–634.
- Proix T, Delgado Saa J, Christen A, Martin S, Pasley BN, Knight RT, Tian X, Poeppel D, Doyle WK, Devinsky O, et al. Imagined speech can be decoded from low- and cross-frequency intracranial EEG features. *Nat Commun*. 2022;13(1):48.
- Rao RPN. An optimal estimation approach to visual perception and learning. *Vis Res*. 1999;39(11):1963–1989.
- Rao RPN, Ballard DH. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat Neurosci*. 1999;2(1):79–87.
- Schultz W, Dayan P, Montague PR. A neural substrate of prediction and reward. *Science*. 1997;275(5306):1593–1599.
- Sestieri C, Corbetta M, Spadone S, Romani GL, Shulman GL. Domain-general signals in the cingulo-opercular network for visuospatial attention and episodic memory. *J Cogn Neurosci*. 2014;26(3):551–568.
- Sestieri C, Shulman GL, Corbetta M. The contribution of the human posterior parietal cortex to episodic memory. *Nat Rev Neurosci*. 2017;18(3):183–192.
- Shadmehr R, Smith MA, Krakauer JW. Error correction, sensory prediction, and adaptation in motor control. *Annu Rev Neurosci*. 2010;33(1):89–108.
- Shipp S. Neural elements for predictive coding. *Front Psychol*. 2016;7:1792.

- Tian X, Poeppel D. Mental imagery of speech and movement implicates the dynamics of internal forward models. *Front Psychol*. 2010;1(166):166.
- Tian X, Poeppel D. Mental imagery of speech: linking motor and perceptual systems through internal simulation and estimation. *Front Hum Neurosci*. 2012;6(314):314.
- Tian X, Poeppel D. The effect of imagination on stimulation: the functional specificity of efference copies in speech processing. *J Cogn Neurosci*. 2013;25(7):1020–1036.
- Tian X, Poeppel D. Dynamics of self-monitoring and error detection in speech production: evidence from mental imagery and meg. *J Cogn Neurosci*. 2015;27(2):352–364.
- Tian X, Zarate JM, Poeppel D. Mental imagery of speech implicates two mechanisms of perceptual reactivation. *Cortex*. 2016;77:1–12.
- Tian X, Ding N, Teng XB, Bai F, Poeppel D. Imagined speech influences perceived loudness of sound. *Nat Hum Behav*. 2018;2(3):225–234.
- Todorovic A, van Ede F, Maris E, de Lange FP. Prior expectation mediates neural adaptation to repeated sounds in the auditory cortex: an MEG study. *J Neurosci*. 2011;31(25):9118–9123.
- Wallis G, Stokes M, Cousijn H, Woolrich M, Nobre AC. Frontoparietal and cingulo-opercular networks play dissociable roles in control of working memory. *J Cogn Neurosci*. 2015;27(10):2019–2034.
- Williams D. Imaginative constraints and generative models. *Australas J Philos*. 2021;99(1):68–82.
- Wolpert DM, Ghahramani Z. Computational principles of movement neuroscience. *Nat Neurosci*. 2000;3Suppl(11):1212–1217.
- Wolpert DM, Ghahramani Z, Jordan MI. An internal model for sensorimotor integration. *Science*. 1995;269(5232):1880–1882.
- Zatorre RJ, Halpern AR, Perry DW, Meyer E, Evans AC. Hearing in the mind's ear: a PET investigation of musical imagery and perception. *J Cogn Neurosci*. 1996;8(1):29–46.
- Zeidman P, Jafarian A, Corbin N, Seghier ML, Razi A, Price CJ, Friston KJ. A guide to group effective connectivity analysis, part 1: first level analysis with DCM for fMRI. *NeuroImage*. 2019;200:174–190.
- Zeidman P, Jafarian A, Seghier ML, Litvak V, Cagnan H, Price CJ, Friston KJ. A guide to group effective connectivity analysis, part 2: second level analysis with PEB. *NeuroImage*. 2019;200:12–25.
- Zhang W, Liu Y, Wang X, Tian X. The dynamic and task-dependent representational transformation between the motor and sensory systems during speech production. *Cogn Neurosci*. 2020;11(4):194–204.