Behavioral/Cognitive

Multiple Concurrent Predictions Inform Prediction Error in the Human Auditory Pathway

Alejandro Tabas[1,2,3] and Katharina von Kriegstein[2,3]

[1]Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, United Kingdom, [2]Department of Psychology, Technische Universität Dresden, 01062 Dresden, Germany, and [3]Max Planck Institute for Human Cognitive and Brain Sciences, 04103 Leipzig, Germany

The key assumption of the predictive coding framework is that internal representations are used to generate predictions on how the sensory input will look like in the immediate future. These predictions are tested against the actual input by the so-called prediction error units, which encode the residuals of the predictions. What happens to prediction errors, however, if predictions drawn by different stages of the sensory hierarchy contradict each other? To answer this question, we conducted two fMRI experiments while female and male human participants listened to sequences of sounds: pure tones in the first experiment and frequency-modulated sweeps in the second experiment. In both experiments, we used repetition to induce predictions based on stimulus statistics (stats-informed predictions) and abstract rules disclosed in the task instructions to induce an orthogonal set of (task-informed) predictions. We tested three alternative scenarios: neural responses in the auditory sensory pathway encode prediction error with respect to (1) the stats-informed predictions, (2) the task-informed predictions, or (3) a combination of both. Results showed that neural populations in all recorded regions (bilateral inferior colliculus, medial geniculate body, and primary and secondary auditory cortices) encode prediction error with respect to a combination of the two orthogonal sets of predictions. The findings suggest that predictive coding exploits the non-linear architecture of the auditory pathway for the transmission of predictions. Such non-linear transmission of predictions might be crucial for the predictive coding of complex auditory signals like speech.

Key words: **auditory midbrain; auditory pathway; cortico-thalamic interactions; predictive coding; sensory processing; sensory thalamus**

Significance Statement

Sensory systems exploit our subjective expectations to make sense of an overwhelming influx of sensory signals. It is still unclear how expectations at each stage of the processing pipeline are used to predict the representations at the other stages. The current view is that this transmission is hierarchical and linear. Here we measured fMRI responses in auditory cortex, sensory thalamus, and midbrain while we induced two sets of mutually inconsistent expectations on the sensory input, each putatively encoded at a different stage. We show that responses at all stages are concurrently shaped by both sets of expectations. The results challenge the hypothesis that expectations are transmitted linearly and provide for a normative explanation of the non-linear physiology of the corticofugal sensory system.

Introduction

Predictive coding (Rao and Ballard, 1999; Friston, 2003b, 2005) is the leading theoretical framework for understanding how expectations are integrated in our experience of reality (Keller and Mrsic-Flogel, 2018). Its central assumption is that sensory processing is mediated by the computation of prediction error: the residual between expectations and the sensory input (Spratling, 2017; Keller and Mrsic-Flogel, 2018).

Predictive coding follows the hierarchical organization of sensory systems (Keller and Mrsic-Flogel, 2018). Units computing prediction error at each processing stage are generally assumed to test the predictions drawn by the level above (Spratling, 2017; Keller and Mrsic-Flogel, 2018; e.g., a pitch processing stage tests predictions drawn from a stage encoding melodic phrases). However, it is unclear how prediction error at each processing stage depends on predictions drawn at even higher levels of the hierarchy, and what happens when predictions from different levels are inconsistent.

Received Dec. 2, 2022; revised Sept. 8, 2023; accepted Sept. 16, 2023.

Author contributions: A.T. and K.v.K. designed research; A.T. performed research; A.T. analyzed data; A.T. and K.v.K. wrote the paper.

The work was funded by the European Research Council [Grant SENSOCOM (647051); Grant Beneficiary: K.v.K] and the Sächsisches Staatsministerium für Wissenschaft, Kultur und Tourismus (Grant Beneficiary: K.v.K).

The authors declare no competing financial interests.

Correspondence should be addressed to Alejandro Tabas at at2045@cam.ac.uk.

https://doi.org/10.1523/JNEUROSCI.2219-22.2023

One line of research suggests that predictions drawn by one level are only tested by the level immediately below. This is the converging conclusion of the studies using the so-called *local–global* paradigm (Bekinschtein et al., 2009). In this paradigm, participants hear successive repetitions of a melodic phrase of tones *AAAAB* rarely substituted by a deviant phrase *AAAAA*. Prediction error units testing (*local*) predictions from a stage encoding pitch elicit prediction error to the fifth tone in the *AAAAB* phrase, and no prediction error to the *AAAAA* phrase. *Local* prediction error has been reported in primary auditory cortex (AC). Conversely, prediction error units testing (so-called) *global* predictions from a stage encoding melodic phrases only show prediction error to the fifth tone in the *AAAAA* phrase. This kind of *global* prediction error has been reported in the frontal areas of the cerebral cortex (Bekinschtein et al., 2009; Wacongne et al., 2011; Chennu et al., 2013; Recasens et al., 2014; El Karoui et al., 2015; Dürschmid et al., 2016; Nourski et al., 2018). Together, these results indicate that predictions do not propagate along the processing hierarchy: otherwise, one would have also expected AC to show prediction error responses to the *AAAAA* phrase.

Another line of research, however, suggests that predictions drawn by one level are tested by more than one lower-level processing stage: task instructions drive the encoding of prediction error in human AC (Stein et al., 2022) and two stages of the subcortical pathway (Tabas et al., 2020, 2021): the auditory thalamus [medial geniculate body (MGB)] and midbrain [inferior colliculus (IC)]. High-resolution studies in non-human primates also reported that the *global* predictions from the *local–global* literature are tested by prediction error units in AC (Uhrig et al., 2014, 2016; Jiang et al., 2022) and MGB (Jiang et al., 2022). Unlike the first line of research, these findings imply that predictions are further transmitted downstream in the hierarchy, all the way to the subcortical pathways.

How can one reconcile the results of the two lines of research? A likely possibility is that the nature of the predictions plays a role for determining whether or not they are further transmitted downstream in the hierarchy. Indeed, while *global* predictions are based on local statistics of the stimulus sequences, *task-induced* predictions stem from a holistic understanding of the structure of the sensory input. The brain may use the better-informed *task-induced* predictions across the sensory pathway and restrict the use of predictions based on the statistics of a representation like the *global* predictions to the immediately lower level. To understand the hierarchical interplay of predictions in predictive coding, it is thus crucial to study scenarios beyond the *local–global* paradigm and consider the interplay of predictions of different natures.

In the present study, we investigate how *stats-informed* and *task-informed* predictions are used to compute prediction error in the human AC, MGB, and IC. We consider three alternative scenarios: first, that prediction error is computed only with respect to the *stats-informed* predictions, consistent with the first line of research (Bekinschtein et al., 2009; Wacongne et al., 2011; Chennu et al., 2013; Recasens et al., 2014; El Karoui et al., 2015; Dürschmid et al., 2016; Nourski et al., 2018); second, that prediction error is computed only with respect to the *task-informed* prediction, consistent with the second line of research (Tabas et al., 2020, 2021; Stein et al., 2022); and third, that prediction error is computed with respect to a combination of both sets of predictions. The third scenario is the one that best reflects the non-linear anatomy of the descending auditory pathway: for instance, the IC receives feedback connections from both the AC and the MGB (Hackett and Kaas, 2004; Lee and Sherman, 2011; Schofield, 2011).

Methods

Experimental design and statistical analysis

Experimental paradigm

In this study, we reanalyze data from two previous experiments (Tabas et al., 2020, 2021). Both experiments used the same task. A trial consisted of a sequence of eight sounds: seven repetitions of a standard and one deviant (Fig. 1*A*). Participants were instructed to monitor the sequences and to report, as accurately and fast as possible, the position of the deviant within the sequence.

The experimental paradigm was designed to elicit two sets of independent predictions on the sensory input. One set of predictions (stats-informed) would be drawn by a population x_k that monitors the stimulus statistics. Within a given trial, we assumed that this population would expect a deviant in position n with a probability $P^{stats}_{nstd} = 1/8$ and a standard in position n with a probability $P^{stats}_{n,std} = 1 - P_{n,std} = 1/8$. This probability distribution P^{stats} defines the stats-informed predictions on the sensory input. Trials were arranged in blocks of 10 that kept the same deviant and standard to ensure that the local representations had sufficient time to infer which sound was the standard and which sound was the deviant.

To elicit the task-informed set of predictions, we introduced two abstract rules: (1) there will always be a deviant and (2) the deviant could only be located at positions 4, 5, or 6. These rules were disclosed to the participants at the beginning of the experiment, who could use it to infer the likelihood of the position of the deviant during each trial. The rule renders $P^{task}_{n,devm} = 0\ \forall n \in \{1, 2, 3, 7, 8\}$, independently of the actual location of the deviant $m \in \{4, 5, 6\}$. The position of the deviant in each trial was pseudorandomized across the experiment so that all deviant positions were equally likely, which means that $P^{task}_{4,devm} = 1/3\ \forall m \in \{4, 5, 6\}$. However, if participants did not find the deviant in position 4, the deviant could only be located in positions 5 or 6; namely, $P^{task}_{5,devm} = 1/2\ \forall m \in \{5, 6\}$. If the deviant is also not present in position 5, it necessarily occurs at position 6, and therefore $P^{task}_{6,dev6} = 1$. This probability distribution P^{task} defines the *task-informed predictions* on the sensory input. Note that the stats-informed and the task-informed predictions do not co-vary; namely, their predicted probabilities of finding a deviant along the different locations of the trial are weakly correlated ($\rho(P^{stats}_{:,dev4}\ P^{task}_{:,dev4}) = 0.14$, $\rho(P^{stats}_{:,dev5}\ P^{task}_{:,dev5}) = 0.21$, and $\rho(P^{stats}_{:,dev4}\ P^{task}_{:,dev4}) = 0.25$).

The inter-trial-interval (ITI) was jittered so that deviants were separated by an average of 5 s, up to a maximum of 11 s, with a minimum ITI of 1500 ms. This maximized the efficiency of the response estimation of the deviants (Friston et al., 1999) while keeping a sufficiently long ITI to ensure that the sequences belonging to separate trials were not confounded.

The experiment consisted in several runs of the same task. Each run contained 6 blocks of 10 trials. The 10 trials in each block contained the same standard–deviant combination, so that within a block only the position of the deviant was unknown, while the identity of the deviant was known. Each of the blocks in a run used one of the six standard–deviant combinations. The order of the blocks within the experiment was randomized. The position of the deviant was pseudorandomized across all trials in each run so that each deviant position happened 60 (pure tone experiment) or 180 (sweep experiment) times per subject but an unknown amount of times per run. This constraint allowed us to keep the same prior probability for all deviant positions in each block (i.e., $P = 1/3$). In addition, there were 23 silent gaps of 5300 ms duration (i.e., null events of the same duration as the tone sequences) randomly located in each run (Friston et al., 1999). Each run lasted around 10 min, depending on the reaction times of the participant.

Stimuli

All stimuli were 50 ms long, including 5 ms ramp-in and ramp-out Hanning windows. Stimuli were arranged in each sequence with a fixed inter-stimulus-interval of ISI = 700 ms.

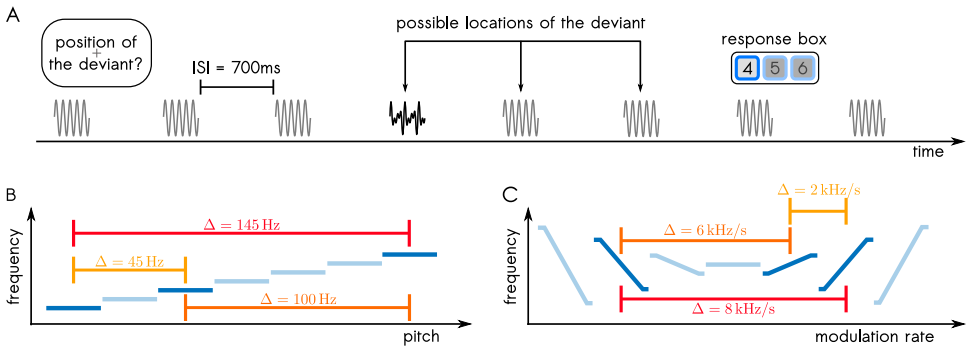

Figure 1. Experimental design. ***A***, Example of a trial. Each trial consisted of a sequence of seven repetitions of one standard (gray) and a single instance of a deviant (black). The deviant could occur in positions 4, 5, or 6 of the sequence. Participants reported, in each trial, the position of the deviant immediately after they identified it. Within a sequence, stimuli were separated by 700 ms ISIs. ***B***, The three pure tones used in the *pure tone* experiment are displayed in dark blue. Trials were characterized by the absolute difference between the frequency of the standard and the deviant Δ. ***C***, The three FM-sweeps used in the *FM-sweep* experiment are displayed in dark blue. Trials were characterized by the absolute difference between the modulation rate of the standard and the deviant Δ. The stimuli schematically shown in light blue in panels ***B*** and ***C*** were not used in the experiments and are plotted here only to contextualize the used stimuli within a family characterized by a continuously varying property (frequency in ***B*** and modulation rate in ***C***).

There were two sets of stimuli, one based on pure tones (experiment 1) and one based on frequency-modulated (FM)-sweeps (experiment 2). Pure tones and FM-sweeps are two of the three information-bearing-elements (IBEs) (Suga, 2012) in which meaningful acoustic signals can be linearly decomposed. We used these two sets to test whether the same principles operate across different IBE types and thus generalize to information-bearing auditory signals.

The pure tone set consisted of three pure tones of frequencies $f_1 = 1455$Hz, $f_2 = 1500$Hz, and $f_3 = 1600$Hz. With these three pure tones, we built six standard–deviant combinations characterized by the absolute frequency difference between deviant and standard $\Delta = |f_{dev} - f_{std}|$. Across the experiments, participants encountered trials with three different values of $\Delta \in \{45, 100, 145\}$ Hz (Fig. 1*B*).

The FM-sweep set consisted of three linear FM-sweeps, one with a descending FM (down) and two with ascending FM (up), with modulation rates $\nu_1 = -4$ kHz/s, $\nu_2 = +2$ kHz/s, and $\nu_1 = +4$ kHz/s. Specifically, the *fast up* sweep had a starting frequency $f_0 = 1000$ Hz and ending frequency $f_1 = 1200$Hz ($\Delta f = 200$ Hz); the *slow up* sweep had $f_0 = 1070$ Hz and $f_1 = 1170$ Hz ($\Delta f = 100$ Hz); and the fast down sweep with $f_0 = 1280$ Hz and $f_1 = 1080$ Hz ($\Delta f = -200$ Hz). The FM-sweeps were designed so that they elicited the same pitch percept and the same average activity across the tonotopic axis, ensuring that participants had to rely on their perception of the modulation rate to tell them apart (see Tabas et al., 2021 for details). Analogously to the pure tones, we used the FM-sweeps to build six standard–deviant combinations characterized by $\Delta = |\nu_{dev} - \nu_{std}| \in \{2, 4, 8\}$kHz/s (Fig. 1*B*).

Description of the data

Data for each experiment were acquired with different MRI-machines and different participant cohorts. Here we describe shortly the key characteristics of each data-set; full descriptions are detailed in Tabas et al. (2020, experiment 1, pure tones) and Tabas et al. (2021, experiment 2, FM-sweeps). Data collection of the pure tone data-set was approved by Ethics Committee of the Medical Faculty of the University of Leipzig, Germany (ethics approval number 273/14-ff). Data collection of the FM-sweep data-set was approved by the Ethics Committee of the Technische Universtät Dresden, Germany (ethics approval number EK 315062019). All listeners provided written informed consent and received monetary compensation for their participation.

Data from 19 (12 female) and 18 participants (12 female) were included in the pure tone and FM-sweeps data-sets, respectively. All participants had normal hearing (thresholds equal of below 25 dB in the range 250 Hz and 8 kHz, as measured by pure tone audiometry) and scores within the neurotypical range in screenings for developmental dyslexia (rapid automatized naming test of letters, numbers, and objects; Denckla and Rudel, 1976) and autism spectrum disorder (AQ; Baron-Cohen et al., 2001; all screenings conducted in German).

Stimuli were presented using MATLAB (The Mathworks) with the Psychophysics Toolbox extensions (Brainard, 1997). Loudness was adjusted independently for each subject before starting the data acquisition to a comfortable level. In the pure tone experiment, stimuli were delivered through an MrConfon amplifier and headphones (MrConfon GmbH). In the FM-sweep experiment, stimuli were delivered through an Optoacoustics (Optoacoustics or Yehuda) amplifier and headphones equipped with active noise-cancellation.

Data from the pure tone data-set were collected using a 7-Tesla Magnetom (Siemens Healthineers) with a spatial resolution of 1.5 mm isotropic and temporal resolution of TR = 1.6 s. Data from the FM-sweep data-set were collected using a 3-Tesla Trio (Siemens Healthineers) with a spatial resolution of 1.75 mm isotropic and temporal resolution of TR = 1.9 s. In both cases, we used EPI sequences with partial coverage. Slices were oriented in parallel to the superior temporal gyrus such that the volumes encompassed the IC, the MGB, and the superior temporal gyrus.

Participants from the pure tone data-set completed 4 runs in a single session (240 trials in total, 80 per deviant position). All but one participant from the FM-sweep data-set completed 9 runs of the main experiment across 3 sessions (540 trials in total, 180 per deviant position); subject 18 completed only 8 runs due to technical reasons. Due to an undetected bug in the presentation code, the first three runs of subjects 1, 2, 4, and 5 and the first six runs of subject 3 were discarded.

During fMRI data acquisition, we also recorded the respiration (in the pure tone data-set) and heart rate (in both the pure tone and FM-sweep data-sets) of the participants. We recorded structural MR-images of each participant using either MP2RAGE (Marques et al., 2010; pure tone data-set; parameters: TE = 2.45 ms, TR = 5000 ms, TI1 = 900 ms, TI2 = 2750 ms, flip angle 1 = 5°, flip angle 2 = 3°, FoV = 224mm × 224mm, GRAPPA acceleration factor 2) or MPRAGE (Brant-Zawadzki et al., 1992; FM-sweep data-set; parameters: TE = 1.95 ms, TR = 1000 ms, TI = 880 ms, flip angle 1 = 8°, FoV = 256mm × 256mm) protocols with nominal resolutions of 0.7 mm and 1.0 mm isotropic, respectively.

All data were preprocessed using Nipype (Gorgolewski et al., 2011), and analyses were carried out using tools of the Statistical Parametric Mapping toolbox, version 12 (SPM); Freesurfer, version 6 (Fischl et al., 2002); the FMRIB Software Library, version 5 (FSL; Jenkinson et al., 2012); and the Advanced Normalization Tools, version 2.2.0 (ANTs; Avants et al., 2011). All second level analyses were performed in Montreal Neurological Institute (MNI) MNI152 1 mm isotropic asymmetric template.

Data were first realigned and unwarped with SPM. The transformation between the functional runs and the structural image was computed with Freesurfer's *BBregister*, which fits the boundaries between gray and white matter of the structural data to the functional images using the whole-brain

EPI as an intermediate step. We computed the transformation between the structural image and the MNI template fitting a concatenation of a rigid, affine, and B-spline non-linear volume-based mappings with ANTs. ANTs simultaneously fit the direct and inverse transform between the two spaces: we used the direct transform to map the data to MNI space, and the inverse transform to map the region of interests (ROIs) to the subject space to validate the registrations (see below).

Physiological (respiration or/and heart rate) data were processed by the PhysIO Toolbox (Kasper et al., 2017) that computes the Fourier expansion of each component along time and adds the coefficients as covariates of no interests in the model's design matrix. All the preprocessing parameters, including the smoothing kernel size, were fixed before we started fitting the general linear model and remained unchanged during the subsequent steps of the data analysis.

Regions of interest

We used two atlases to identify which voxels belonged to each subcortical and cortical ROI. For the subcortical ROIs, we used an in vivo atlas (Sitek et al., 2019) that identified which voxels of the MNI space most likely cover bilateral IC and MGB. To test for potential functional specializations of the subdivision of the MGB, we used masks calculated in Mihai et al. (2019) detailing the average location of the ventral tonotopic axis of the nucleus across 28 participants. This is, to-date, the best existing approximation of the location of primary (ventral) MGB (Mihai et al., 2019).

For the cerebral cortex ROIs, we used the Morosan atlas (Morosan et al., 2001), which subdivides AC in four bilateral cortical areas using cytoarchitectural considerations. Cortical areas are identified as Te1.0, Te1.1, Te1.2, and Te3. Areas Te1.0, Te1.1, and Te1.2 are mostly located on Heschl's gyrus (Te1.1 most posterio-medial, Te1.2 most antero-lateral, and Te.1.0 in between), and Te3 is located on the lateral surface of the superior temporal gyrus (Morosan et al., 2001). Te1.0 includes areas analogous to the core of the AC; i.e., primary AC (Moerel et al., 2014).

To empirically validate the registration procedure, we used the inverse transforms provided by ANTs to project the subcortical (Fig. 2) and cortical (Figs. 3, 4) ROIs to the native space of the structural images. The plots confirm that our registration pipeline successfully mapped the anatomical ROIs of each participant into the MNI space with an orientation that resembles the orientation of the different AC regions.

Bayesian model comparison

To evaluate whether neural responses in each of the ROIs corresponded to prediction error with respect to the stats-, task-informed predictions, or both (Fig. 1), we used Bayesian model comparison. Bayesian model comparison allows to calculate the evidence for a given model of the response profile in each voxel of the ROI. We used three models that capture three different hypotheses. *Stat*: neural responses encode prediction error with respect to P^{stats}; *task*: neural responses encode prediction error with respect to P^{task}; *combined*: neural responses encode a linear combination of the prediction errors with respect to each, P^{stats} and P^{task}. In addition, we used a *control* model that encoded the structure of the paradigm and served as baseline. The numeric definitions of these models are described below (Methods, Definition of the models).

All regressors corresponding to each of the model were normalized to have a mean of zero and variance of one across each run before convolution and model fitting. Note however that SPM orthogonalizes the regressors before fitting them to the data. Moreover, since we applied the same procedure to all models, preprocessing of the regressors cannot bias model comparison.

We first computed the log-evidence for each of the three models in each voxel of the ROIs per each participant using SPM via nipype. Given the model amplitude(s) a_n and the timecourse of a voxel y, SPM calculates the log-evidence of the linear model $y = \beta_0 + \sum_n \beta_n a_n + \varepsilon$, where β_n are the linear coefficients of each regressor and ε are noise terms. To avoid inspecting the data more than once (and thus incur into multiple-comparison problems), we used the default uninformed priors (unweighted graph Laplacian, which impose a soft constraint that values for the coefficients change smoothly across nearby voxels) and default hyperparameters of the implementation of the method in SPM for the computation of the first level Bayesian analysis (Penny et al., 2003).

Log-evidence maps were then combined across participants for each stimulus set using custom scripts (see Data and code availability) following a two-step procedure: first, we combined the log-evidences across sessions for each individual subject assuming fixed-effects (i.e., summing across the log-evidences for each subject) and then computed a posterior distribution at the group level using the random-effects procedure described in Rosa et al. (2010) and Stephan et al. (2009). Uninformed priors (i.e., uniform distribution across models) were used for the second level Bayesian analysis. This procedure resulted in an estimation of the log-evidence of each model for each voxel. Group-level log-evidence maps were then subtracted to compute the Bayes factor of the comparison of any two models m_1 and m_2: $K_{m_1/m_2} = e^{\log Ev_{m_2} - \log Ev_{m_1}}$. Since K-maps are only used for the computation of possible correlations between the relative posteriors of the models with the temporal signal-to-noise ratio (tSNR), we did not threshold the resulting k-maps. The predictive power of each of the main three models m is quantified as the K factor given m and the control model. We consider that there is substantial evidence that a voxel is better explained by m than by the control model when $K > \sqrt{10}$.

Definition of the models

Modeling prediction error

In all models, we assume that neural responses encode prediction error with respect to a set of predictions. The models disregard contributions of other factors (e.g., neural habituation) that, although are expected to influence the signal, would not differ same across sets of predictions.

We defined prediction error as the mismatch between the expected stimulus and the actual stimulus weighted by the likelihood of encountering the stimulus in each position $\Pi = P^{stats/task}_{n,std/dev_m}$. We assumed that the mismatch between the expected and actual stimulus would be a monotonically increasing function of the absolute difference between the deviant and standard Δ; we approximated this function to be locally linear in a neighborhood of the set of values of Δ considered in our experiments and to be zero if the expected and the presented stimuli were the same. As such, prediction error ξ is defined as follows:

$$\xi = \sum_{s \in stimuli} \Pi_s f(\Delta, s, \text{input})$$
$$f(\Delta, s, \text{input}) = \begin{cases} 0 & \text{if } s = \text{input} \\ b_0 + b_1\Delta & \text{if } s \neq \text{input} \end{cases} \quad (1)$$

where $s \in$ *stimuli* are all the stimuli that could plausibly be heard in the next location: the standard and the deviant. For instance, if the prediction for a given tone is $P_{std} = 2/3$ $P_{dev} = 1/3$ and the tone is actually a standard, the prediction error would be $\xi = P_{std} \times 0 + P_{dev}(b_0 + b_1\Delta) = 1/3(b_0 + b_1\Delta)$.

We modeled the prediction error responses using two regressors:

$$a_1^{\xi} = \sum_s \Pi_s \delta_{s,\text{input}}, \quad (2)$$

$$a_2^{\xi} = \sum_s \Pi_s \delta_{s,\text{input}} \Delta_s, \quad (3)$$

where δ_{sinput} is the Kronecker delta. While the regressor a_1 represents a constant response to unexpected stimuli (b_0 in Eq. 1), a_2 captures the dependence between prediction error and the mismatch between the expected and presented stimuli (b_1 in Eq. 1). Although a_2 is not strictly necessary to differentiate between the *stats*, *task*, *combined* models, these additional regressors allow the models to fit different amounts of prediction error in trials that have different standard–deviant combinations. Moreover, using these two regressors yields the linear model $y = \beta_0 + \beta_1 \sum_s \delta_{s,\text{input}} \Pi_s + \beta_2 \sum_s \delta_{s,\text{input}} \Pi_s \Delta_s \propto \beta_0 + \varepsilon$ (compare with Eq. 1), where β_0 accounts for the baseline constant BOLD response of the voxel (i.e., independently of the stimulation and predictions).

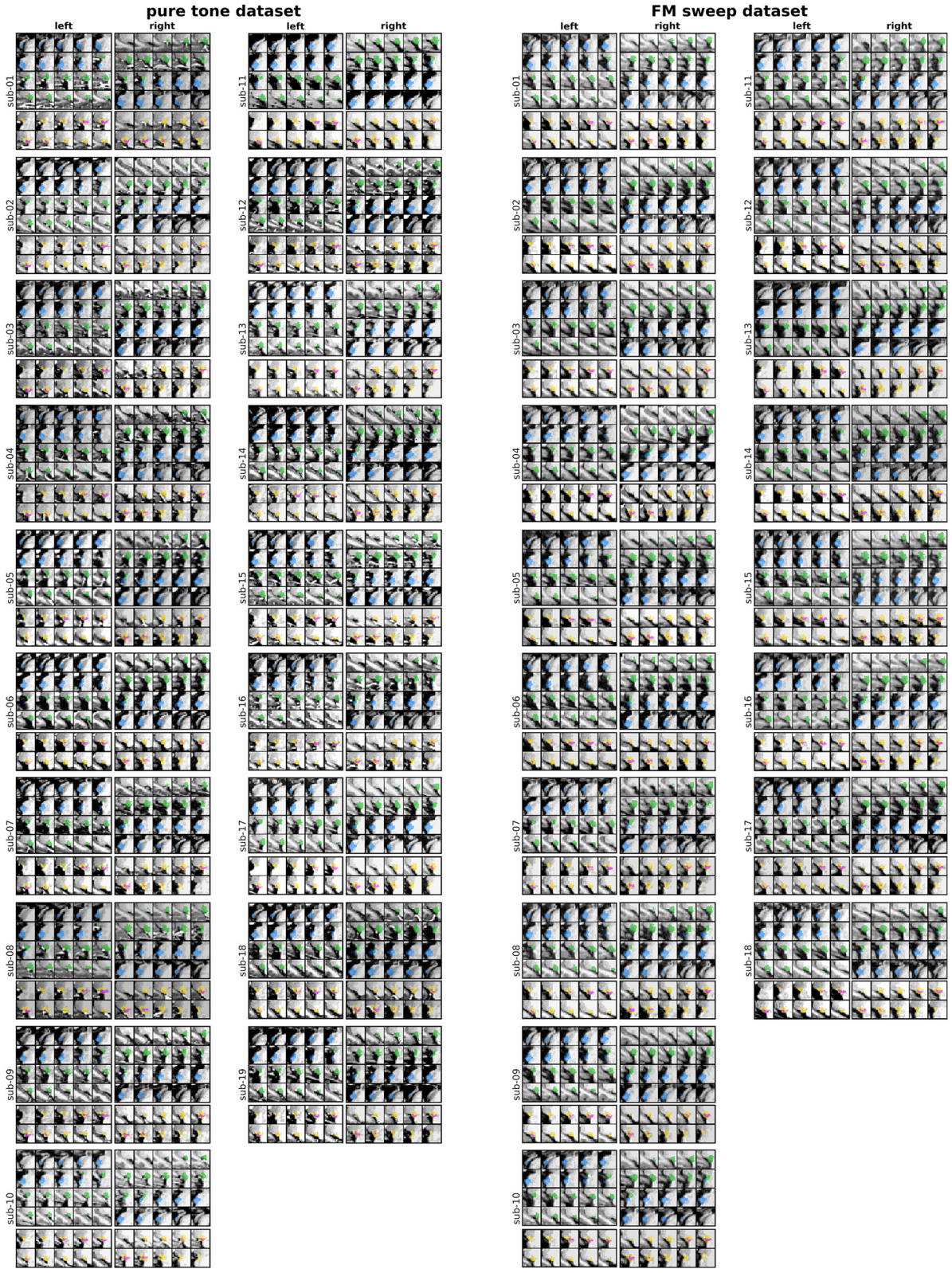

Figure 2. Anatomical location of the subcortical ROIs in each participant. Each panel plots the location of each ROI projected from MNI to the structural space of the participant using the coregistration inverse transform.

Regressors in Equations 2 and 3 can capture the prediction error to stimuli in positions 2–8; however, they cannot capture the responses to the first standard in each sequence. The first standard elicits prediction error with respect to the task-informed predictions not because its identity is unknown, but because its onset time is unknown. It also elicits prediction error with respect to the stats-informed predictions because it interrupts the silence that precedes it in the local stimulus history. To take into account the contributions of the first standard without tweaking

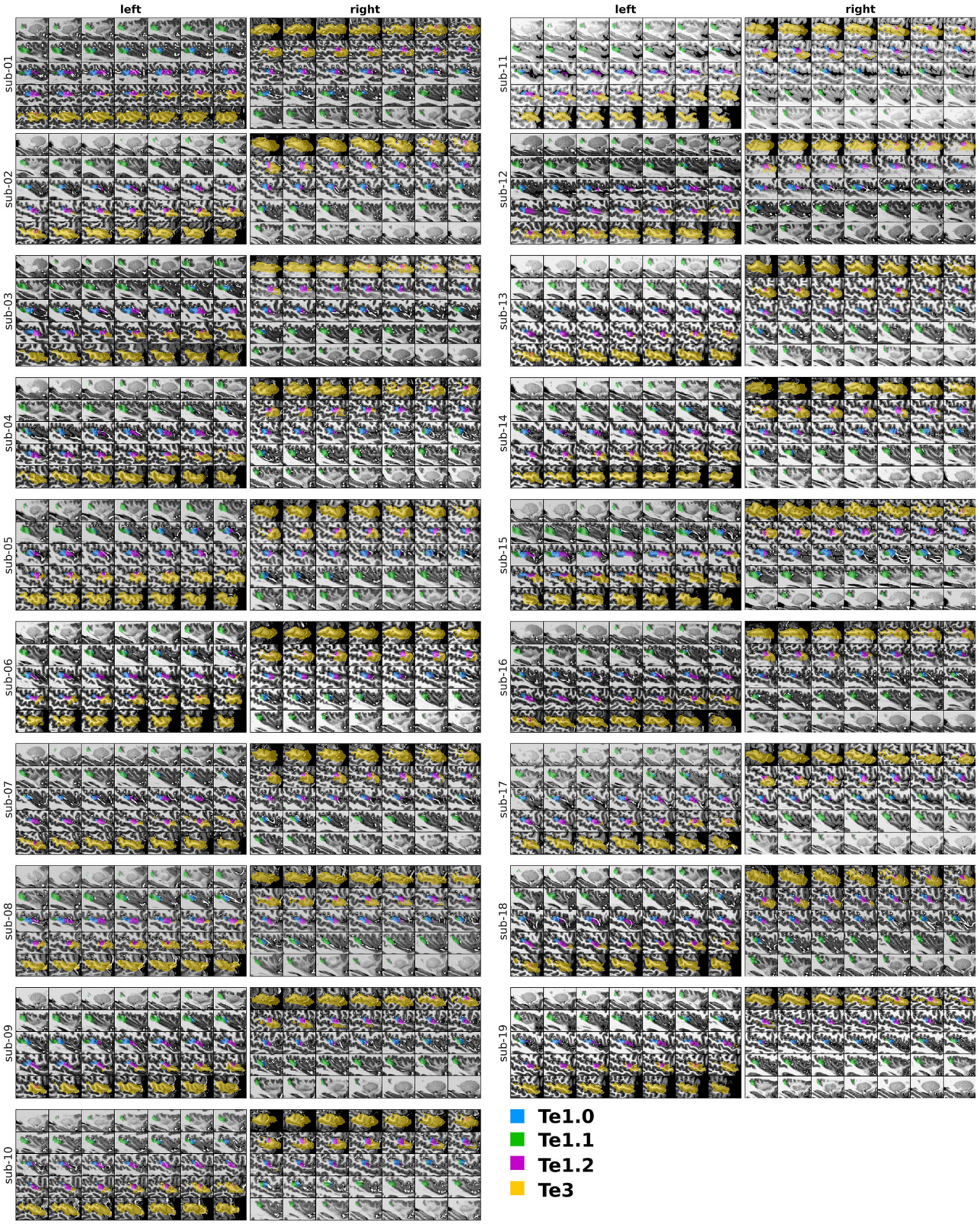

Figure 3. Anatomical location of the cortical ROIs in each participant (pure tone experiment). Each panel plots the location of each ROI projected from MNI to the structural space of the participant using the coregistration inverse transform.

the definition of ξ from Equation 1, we added another regressor $a_3^{\xi} = \delta_{n,1}$; namely, $a_3^{\xi} = 1$ for the first standard of each sequence and $a_3^{\xi} = 0$ for the remaining of the sounds in each sequence. Conversely, a_1^{ξ} and a_2^{ξ} are non-zero only for positions 2–8 within each sequence.

While models corresponding to the stats- and task-informed predictions had three regressors, the model that incorporates both sets of predictions had five regressors. Bayesian log-evidences penalizes the addition of extra regressors, meaning that the evidence for any model

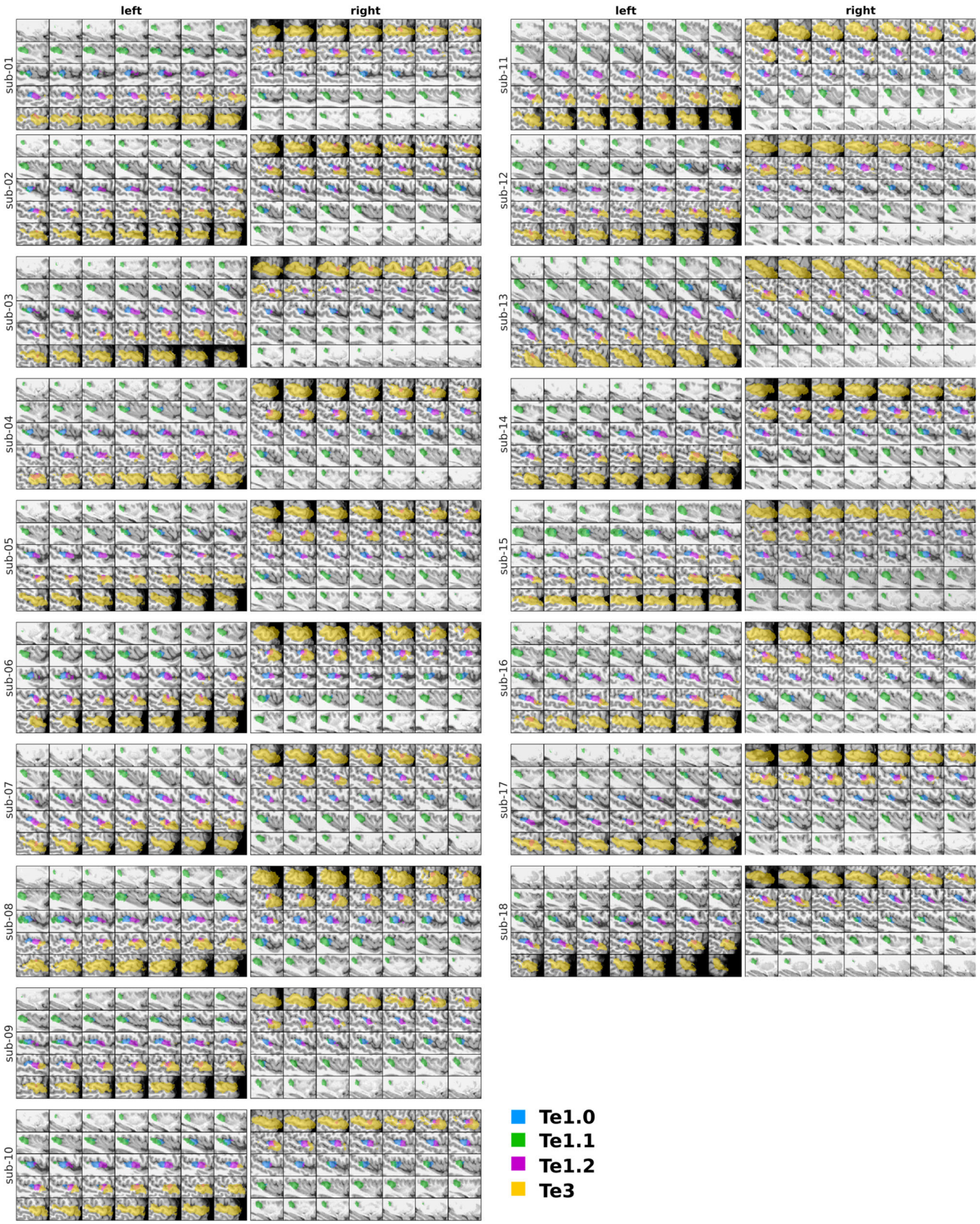

Figure 4. Anatomical location of the cortical ROIs in each participant (FM-sweeps experiment). Each panel plots the location of each ROI projected from MNI to the structural space of the participant using the coregistration inverse transform.

of a higher complexity would only be greater than the evidence for a model of lower complexity if the additional regressors explain the data better beyond what would have been expected due to overfitting of the extra free parameters.

Prediction error with respect to the stats-informed predictions

The stats-informed predictions were $P_{std}^{stats} = 7/8$ $P_{dev}^{stats} = 1/8$ for all tones in the sequence (Fig. 5*A*). Exact values for the regressors are detailed in Table 1.

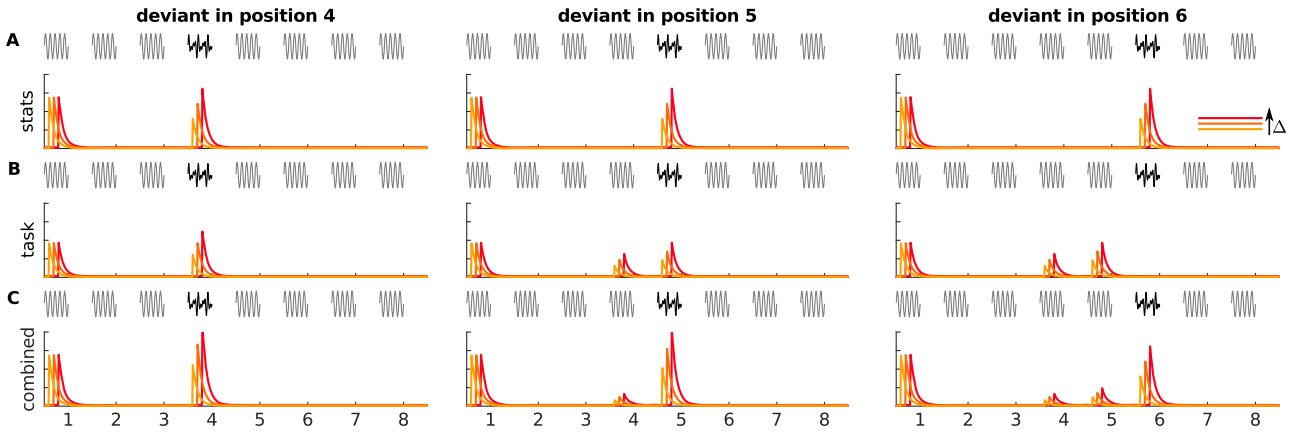

Figure 5. Schematics of t.eps models used for Bayesian model comparison. Each panel plots a possible linear combination of the regressors used in each of the three models for each of the nine trial types (three deviant positions × three values of Δ) of the experiments. Plots in panel ***A*** show the *stats* model, and in panel ***B***, the *task* model, and in ***C***, the *combined* model. Each colored line corresponds to one Δ value (red corresponds to the largest delta, yellow to the lowest). The apparent delay between colored lines is a visualization device: there was no such delay in the model. Note that the relative height of the first standard (in comparison to the deviant) and the relative weight that Δ has in the responses to the deviants are free parameters of the model.

Table 1. Amplitudes of the models used for Bayesian model comparison

.eps2lstats		1	2	3	4	5	6	7	8
a_1	Deviant at 4	0	1/8	1/8	7/8	1/8	1/8	1/8	1/8
	Deviant at 5	0	1/8	1/8	1/8	7/8	1/8	1/8	1/8
	Deviant at 6	0	1/8	1/8	1/8	1/8	7/8	1/8	1/8
a_2	Deviant at 4	0	1/8Δ	1/8Δ	7/8Δ	1/8Δ	1/8Δ	1/8Δ	1/8Δ
	Deviant at 5	0	1/8Δ	1/8Δ	1/8Δ	7/8Δ	1/8Δ	1/8Δ	1/8Δ
	Deviant at 6	0	1/8Δ	1/8Δ	1/8Δ	1/8Δ	7/8Δ	1/8Δ	1/8Δ
a_3	All deviants	1	0	0	0	0	0	0	0
task		1	2	3	4	5	6	7	8
a_1	Deviant at 4	0	0	0	2/3	0	0	0	0
	Deviant at 5	0	0	0	1/3	1/2	0	0	0
	Deviant at 6	0	0	0	1/3	1/2	0	0	0
a_2	Deviant at 4	0	0	0	2/3Δ	0	0	0	0
	Deviant at 5	0	0	0	1/3Δ	1/2Δ	0	0	0
	Deviant at 6	0	0	0	1/3Δ	1/2Δ	0	0	0
a_3	All deviants	1	0	0	0	0	0	0	0
combined		1	2	3	4	5	6	7	8
a_1	Deviant at 4	0	1/8	1/8	7/8	1/8	1/8	1/8	1/8
	Deviant at 5	0	1/8	1/8	1/8	7/8	1/8	1/8	1/8
	Deviant at 6	0	1/8	1/8	1/8	1/8	7/8	1/8	1/8
a_2	Deviant at 4	0	1/8Δ	1/8Δ	7/8Δ	1/8Δ	1/8Δ	1/8Δ	1/8Δ
	Deviant at 5	0	1/8Δ	1/8Δ	1/8Δ	7/8Δ	1/8Δ	1/8Δ	1/8Δ
a_3	Deviant at 4	0	0	0	2/3	0	0	0	0
	Deviant at 5	0	0	0	1/3	1/2	0	0	0
	Deviant at 6	0	0	0	1/3	1/2	0	0	0
a_4	Deviant at 4	0	0	0	2/3Δ	0	0	0	0
	Deviant at 5	0	0	0	1/3Δ	1/2Δ	0	0	0
	Deviant at 6	0	0	0	1/3Δ	1/2Δ	0	0	0
a_5	All deviants	1	0	0	0	0	0	0	0

Notes: Amplitudes of the linear models used for Bayesian model comparison. For each of the three models, we computed the log-evidence that each responses in each voxel $y \sim \sum_n \beta_n a_n$, where β_n are the free parameters of the model. *Stats* assumes that responses encode prediction error with respect to a set of predictions that are informed by local stimulus history and statistics. *Task* assumes that responses encode prediction error with respect to predictions drawn by the internal representation of the task instructions. *Combined* assumes that responses encode a linear combination of prediction error with respect to the stats-informed and task-informed predictions. All regressors were normalized (mean of zero and variance of one) prior fitting.

Prediction error with respect to the task-informed predictions

Task-informed predictions depended on the position of the incoming sound n (Fig. 5*B*). For positions $n \in [2, 3, 7, 8]$, $P^{task}_{n,std} = 1$, $P^{task}_{n,dev} = 0$; for position $n = 4$, $P^{task}_{4,dev} = 1/3$, $p^{task}_{4,std} = 2/3$. Predictions for positions $n \in \{5\ 6\}$ depended on the actual position of the deviant m. If the deviant was at position $m = 4$, participants would not expect any other deviant in positions 5 and 6, and thus the predictions would be $P^{task}_{n,dev} = 0$, $P^{task}_{n,std} = 1\ \forall n \in \{5, 6\}$. If the deviant was not in position $m = 4$, the predictions on $n = 5$ would be $P^{task}_{5,dev} = P^{task}_{5,std} = 1/2$. If the deviant was in position $m = 5$, predictions for $n = 6$ would be $P^{task}_{6,dev} = 0$, $P^{task}_{6,std} = 1$. Last, if the deviant was neither in position $m = 5$, the predictions on $n = 6$ would be $P^{task}_{6,dev} = 1$, $P^{task}_{6,std} = 0$. Exact values for the regressors are detailed in Table 1.

With the exception of the expected responses to the first standard, the regressor a_2 of this model is identical to the *predictive coding* hypothesis in Tabas et al. (2020, 2021) and Stein et al. (2022).

Prediction error with respect to a combination of both predictions
To test whether both, stats- and task-informed predictions, contribute to the computation of prediction error, we assumed that predictions would be a linear combination of the predictions of both models (Fig. 5*C*); or, similarly, that the neural responses would be a linear combination of the prediction error expected by each set of predictions (note that the dependence of ξ on Π in Eq. 1 is linear). We modeled this scenario by adding the regressors a_1^{ξ} a_2^{ξ} corresponding to each of the two sets of predictions. Since the responses to the first standard co-vary in both, stats- and task-informed models, we added only one regressor for the first standard of each sequence.

Control model
The control model includes three regressors, one per stimulus identity. Before normalization, a regressor corresponding to, for example, the pure tone with the highest pitch had a value of 1 when the tone was played and zero otherwise.

Measuring the tSNR
To test whether the results were influenced by the tSNR of the data, we used nipype's native *confound* toolbox. We computed the tSNR in the exact same preprocessed data we used as input for the Bayesian model comparison analysis. A limitation of measuring the tSNR on raw data is that the analysis effectively interprets task-induced fluctuations, which are part of the signal, as noise. A popular alternative is to use the contrast-to-noise ratio (CNR) (Welvaert and Rosseel, 2013); however, the CNR is only defined with respect to a model of the data and cannot be used to measure the relative noise across different models. Moreover, given that fMRI is characterized by a low signal changes in high noise regimes (Welvaert and Rosseel, 2013), that we only use the tSNR estimations to compare model performances, and that both data-sets included in this study used the same task, the tSNR is a reasonable and unbiased estimation of the relative levels of noise of the data. Nevertheless, the results of the control analysis involving the estimations of the tSNR should be considered with caution.

Correlational analyses
To measure whether results were consistent across the two experiments and whether spatial variations in the winning models could be explained by tSNR heterogeneities, we computed Pearson's correlations between statistical maps across the voxels in each ROI. This means that the number of samples in each correlation is the number of voxel in the ROI. All *p*-values were Holm–Bonferroni corrected for the number of ROIs ($N = 4$ in the analyses on subcortical regions, $N = 10$ in the analyses of cerebral cortex). Results were deemed statistically significant when the corrected $p < 0.05$.

Data and code availability
All code and derivatives needed to reproduce the analyses and figures are openly available in osf.io/f5tsy.

Results
Multiple predictions are combined to compute prediction error in the subcortical auditory pathway
Large sections of the IC and MGB displayed responses that were best explained by the *task* and the *combined* models, both for pure tones (Fig. 6*A*) and FM-sweeps (Fig. 6*C*). To rigorously establish the prevalence of each model in each ROI, we computed the Bayes factors between each target and the *control* model (Fig. 6*B*,*D*). The prevalence of a model in a ROI was determined as the fraction of the voxels for which the model provides for a substantially better explanation of the data than the *control* model (i.e., $K(model/control) > \sqrt{10}$). The *combined* model was the most prevalent model in the four subcortical ROIs of the pure tone experiment (Table 2). Although the *task* model was the most prevalent model in the FM-sweep experiment, populations best explained by the *combined* model were also substantially present. The *stats* and *control* models were the best explanation of the data only in a few voxels scattered across the ROIs for both stimulus families.

We also computed the minimum K factor between each model and the remaining models (Table 3) to determine whether the responses in each voxel of each ROI were substantially better explained by any of the four models. The *combined* model provided for a substantially better explanation for the data than in 36%, 25%, 49%, and 60% of voxels of the IC-L, IC-R, MGB-L, and MGB-R, respectively. Results were less clear in the FM-sweep data, where the *combined* and *task* models seem to perform similarly well. The *control* and *stats* models had no substantial explanatory power in any of the ROIs.

In summary, both *combined* and *task* models were extremely prevalent in all ROIs for both experiments. While the *combined* model dominated the responses in the pure tone experiment, both *combined* and *task* models dominated the responses in the FM-sweep experiment.

Prediction error in the MGB is consistent across physiological subdivisions
The auditory pathway is subdivided into primary (central section of the IC and ventral MGB) and secondary (cortex of the IC, and medial and dorsal MGB) subdivisions (Hu, 2003). Neurons in primary subdivisions have narrowly tuned frequency responses and are responsible for the transmission of information between the periphery and the cerebral cortex; neurons in secondary subdivisions present wider tuned frequency responses and are thought to be involved in multisensory integration (Hu, 2003). One possibility is that the functional parcellations described in Figure 6 correspond to this physiological arrangement. Neural populations responding according to the *combined* model do indeed seem to be located toward the cortex of the ICs, although lower tSNRs are generally expected in the borders of the nuclei, whose signal has contributions from the adjacent cerebrospinal fluid.

Imaging subdivisions of the IC and MGB in humans are remarkably challenging (Moerel et al., 2015; Mihai et al., 2019). To-date, there is no available parcellation of the human IC into primary and secondary subdivisions; however, Mihai et al. (2019) managed to identify a ventral tonotopic gradient in the MGB that putatively corresponds to its primary subdivision. Here, we used this parcellation to assess whether neural populations in primary and secondary subdivisions of the MGB are signaling prediction errors related to the *task* or the *combined* model (Fig. 7). Results show that both models are similarly prevalent in both subdivisions, indicating that, at least in the MGB, the functional parcellation described in Figure 6 does not correspond to the physiological parcellations of the nuclei.

tSNR heterogeneity correlates with model performance
The prevalence of *stats* and *combined* models in different sections of the MGB and IC may indicate that there is not a unique strategy to propagate predictions on the sensory input downstream, but that different strategies are used in different neural populations. However, Figure 7 indicates that these neural populations do not correspond to specific subdivisions of the nuclei. Another possibility is that voxels best explained by the *task* model are those voxels in which the BOLD responses are noisier in comparison to those in other regions. Since the *combined* model has two free parameters more than the *task* model, the former needs to provide a better explanation of the data than the latter to yield a similar log-evidence. Voxels with poorer tSNR would present higher mean-square-errors with respect to the model fits, which might result in the winning of the *task* model.

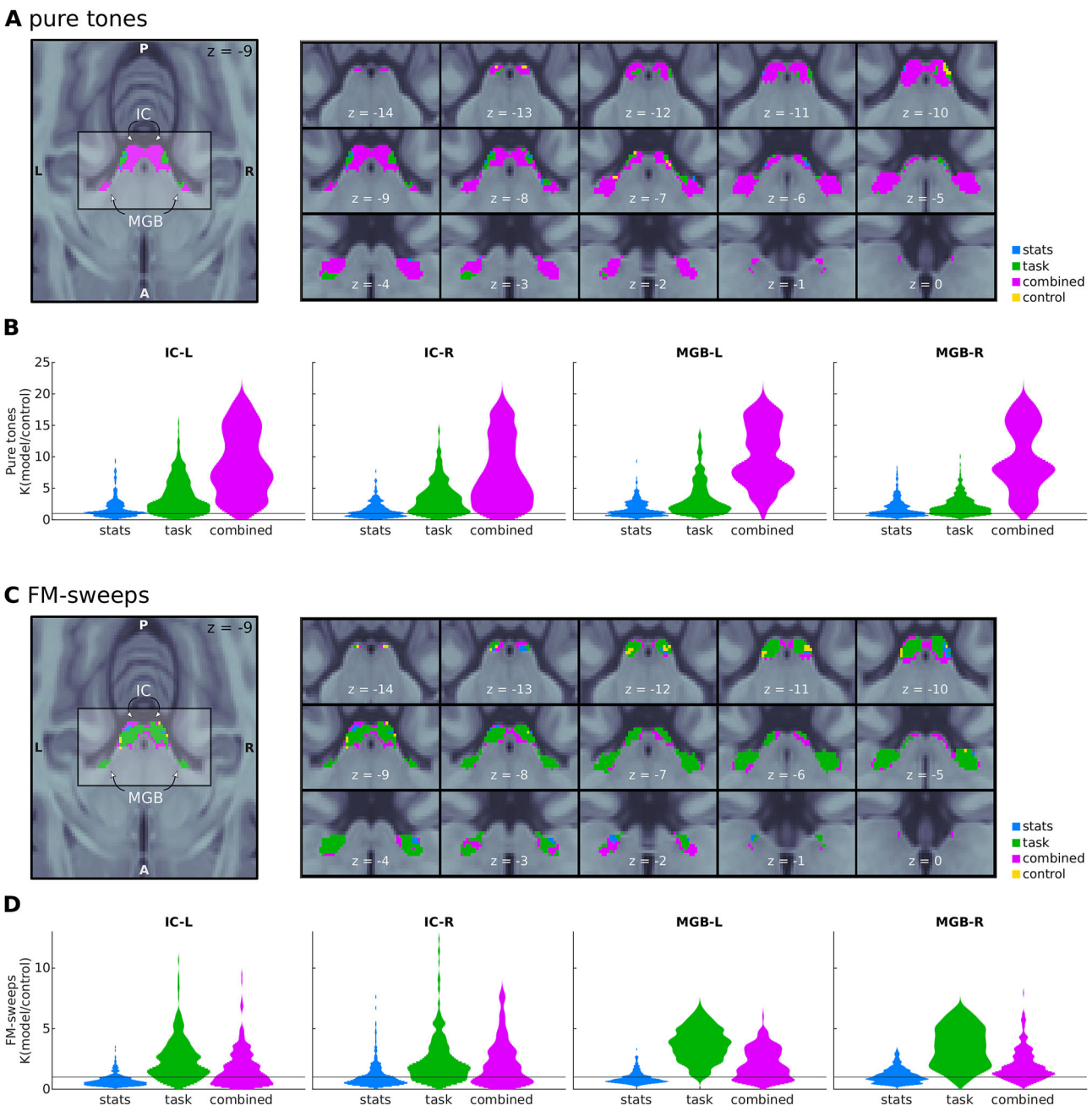

Figure 6. Bayesian model comparison in IC and MGB. ***A***, Experiment 1 (pure tones). Maps detailing which model best explained the responses to pure tones in each of the voxels of the IC and MGB ROIs. Colors indicate the model with the highest posterior density at each voxel of the IC and MGB ROIs. Blue voxels are best explained by the *stats* model, green voxels by the *task* model, purple voxels by the *combined* model, and yellow voxels by the *control* model, taken here as baseline. ***B***, Distributions (kernel-density estimations) of the *K* factors comparing the performance of each of the first three models again the control model across voxels of the IC and MGB rois. ***C***, The same as in *A*, but for Experiment 2 (FM-sweeps). ***D***, The same as in *B*, but for Experiment 2 (FM-sweeps).

To test whether that was the case, we computed the correlation between the tSNR and the posterior density of the *combined* model across the ROIs. We found a significantly positive correlation in three of the four ROIs for the pure tone data ($\rho \in [0.53, 0.71], p < 10^{-20}$ in IC-L, IC-R, and MGB-R; $\rho = -0.07, p = 0.2$ in MGB-L) and in all four ROIs for the FM-sweep data ($\rho \in [0.36, 0.71]\ p < 10^{-9}$). These results suggest that the lower tSNRs may be one of the reasons why the *combine.eps* model was not the best explanation for the data across the entire nuclei of the subcortical pathway.

Similar distribution of model performances for pure tones and FM-sweeps

Although the general prevalence of the *task* and *combined* models was different for the responses elicited by pure tones and FM-sweeps, they seem to follow a similar topographic organization on visual inspection: populations best explained by the *combined* model are located more centrally in the ICs and more dorsally in the MBGs. To quantify if the occurrence of the *task* model was consistent across the two stimulus families, we compared the distribution of the $K_{combined/task}$ associated to the responses to pure tones and FM-sweeps. Distribution of *K*

Table 2. Prevalence of each model as providing a substantially better explanation for the data than the control model

	Stats-informed		Task-informed		Combined	
	Pure tones	FM-sweeps	Pure tones	FM-sweeps	Pure tones	FM-sweeps
IC-L	0.12	0.01	0.47	**0.33**	**0.84**	0.14
IC-R	0.08	0.06	0.48	**0.28**	**0.76**	0.26
MGB-L	0.11	0.01	0.44	**0.62**	**0.99**	0.30
MGB-R	0.15	0.00	0.23	**0.62**	**0.90**	0.17
Te10-L	0.07	0.00	0.76	0.38	**0.81**	**0.54**
Te10-R	0.02	0.00	0.61	0.39	**0.98**	**0.50**
Te11-L	0.04	0.00	0.73	**0.55**	**0.90**	0.53
Te11-R	0.06	0.00	0.70	**0.55**	**0.89**	0.37
Te12-L	0.06	0.00	0.59	0.09	**0.89**	**0.53**
Te12-R	0.09	0.00	0.45	0.14	**0.99**	**0.69**
Te3-L	0.11	0.00	0.39	0.18	**0.90**	**0.51**
Te3-R	0.13	0.01	0.37	0.17	**0.74**	**0.49**

Notes: Each entry specifies the ratio of the voxels for which $K(\text{model}/\text{control}) > \sqrt{10}$ in each of the ROIs and for each of the set of stimuli. Entries in BOLD signal the model that provided for the substantially best explanation for the data in the largest amount of voxels in each ROI and stimulus set.

Table 3. Prevalence of each model as providing a substantially better explanation for the data than the remaining models

	Stats-informed		Task-informed		Combined	
	Pure tones	FM-sweeps	Pure tones	FM-sweeps	Pure tones	FM-sweeps
IC-L	0.00	0.00	0.00	**0.09**	**0.36**	0.04
IC-R	0.00	0.00	0.02	0.01	**0.25**	**0.09**
MGB-L	0.00	0.00	0.01	**0.11**	**0.49**	0.03
MGB-R	0.00	0.00	0.00	**0.16**	**0.60**	0.03
Te10-L	0.07	0.00	0.76	0.38	**0.81**	**0.54**
Te10-R	0.02	0.00	0.61	0.39	**0.98**	**0.50**
Te11-L	0.04	0.00	0.73	**0.55**	**0.90**	0.53
Te11-R	0.06	0.00	0.70	**0.55**	**0.89**	0.37
Te12-L	0.06	0.00	0.59	0.09	**0.89**	**0.53**
Te12-R	0.09	0.00	0.45	0.14	**0.99**	**0.69**
Te3-L	0.11	0.00	0.39	0.18	**0.90**	**0.51**
Te3-R	0.13	0.01	0.37	0.17	**0.74**	**0.49**

Notes: Each entry specifies the ratio of the voxels for which $\min_{m \neq \text{model}}(K(\text{model}/m) > \sqrt{10}$ in each of the ROIs and for each of the set of stimuli. Entries in BOLD signal the model that provided for the substantially best explanation for the data in the largest amount of voxels in each ROI and stimulus set. The prevalence of the control model, not listed in the table, was 0 for all ROIs and stimuli.

factors was significantly correlated in the left IC ($\rho = 0.2$, $p = 6 \times 10^{-4}$), right IC ($\rho = 0.31$, $p = 5 \times 10^{-8}$), and right MGB ($\rho = 0.21$, $p = 2.4 \times 10^{-3}$), but not in the left MGB ($\rho = -0.08$, $p = 0.16$). However, the tSNR distributions across both experiments were highly correlated in the four ROIs ($\rho \in [0.63, 0.89]$ $p < 10^{-32}$), indicating that the correlation between $K_{combined/task}$ in the pure tone and FM-sweep data could have also been driven by a similar distribution of the tSNR in both data-sets.

Multiple predictions are combined to compute prediction error in AC

Most of the AC responses (Te1.0, Te1.1, Te1.2, and Te3; Morosan et al., 2001) were best explained by the *combined* model (Table 2, Fig. 10): it was the best explanation of the data in more than half of the voxels across fields for pure tones (Fig. 8) and FM-sweeps (Fig. 9). The *task* model explained the responses of most of the remaining voxels, while the *stats* and *control* models were present only minimally.

To study whether the presence of the *task* model could also be related to the variations of the tSNR across the ROIs, we again computed the correlation between the tSNR and the posterior density of the *combined* model across the cerebral cortex ROIs. In the pure tone data, the posterior density was positively correlated with the tSNR in Te1.0-R, Te1.1-R, Te1.2-L, and bilateral Te3 ($\rho \in [0.13, 0.45]$ $p < 10^{-7}$), but not in the remaining ROIs ($\rho \in [-0.24, 0.07]$); in the FM-sweep data, correlations were significant in all cerebral cortex ROIs ($\rho \in [0.09, 0.70]$ $p < 0.02$). This indicates a partial contribution of tSNR heterogeneities to the prevalence of the *task* model.

To quantify if, as in the subcortical nuclei, the cortical organization of the *combined* and *task* models was consistent for both stimulus families across cortical fields, we computed the correlation between the Bayes' factor $K_{combined/task}$ associated with the responses to pure tones and FM-sweeps. We found significantly positive correlations in four of the cortical fields (Te1.0-R, bilateral Te1.1, and Te3-L; $\rho \in [0.04, 0.32]$ $p < 0.002$). However, the tSNR of the pure tone and FM-sweep data-sets was also positively correlated ($\rho \in [0.31, 0.73]$ $p < 10^{-25}$) across all cortical fields but Te1.1-R ($\rho = -0.07$), indicating that the correlations of

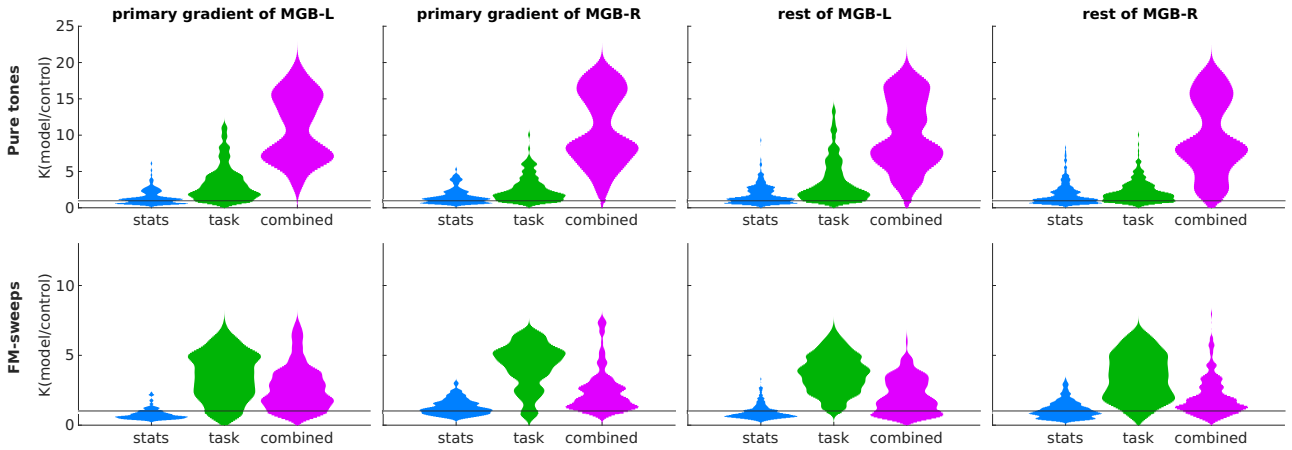

Figure 7. Bayesian model comparison in the primary and secondary subdivisions of the MGB. Distributions (kernel-density estimations) of the *K* factors comparing the performance of each of the first three models again the control model across voxels of the MGB subdivisions from Mihai et al. (2019) for the pure tone and FM-sweep stimuli. Distributions are qualitatively comparable in the primary MGB and the rest of the nucleus (secondary MBG) for both experiments.

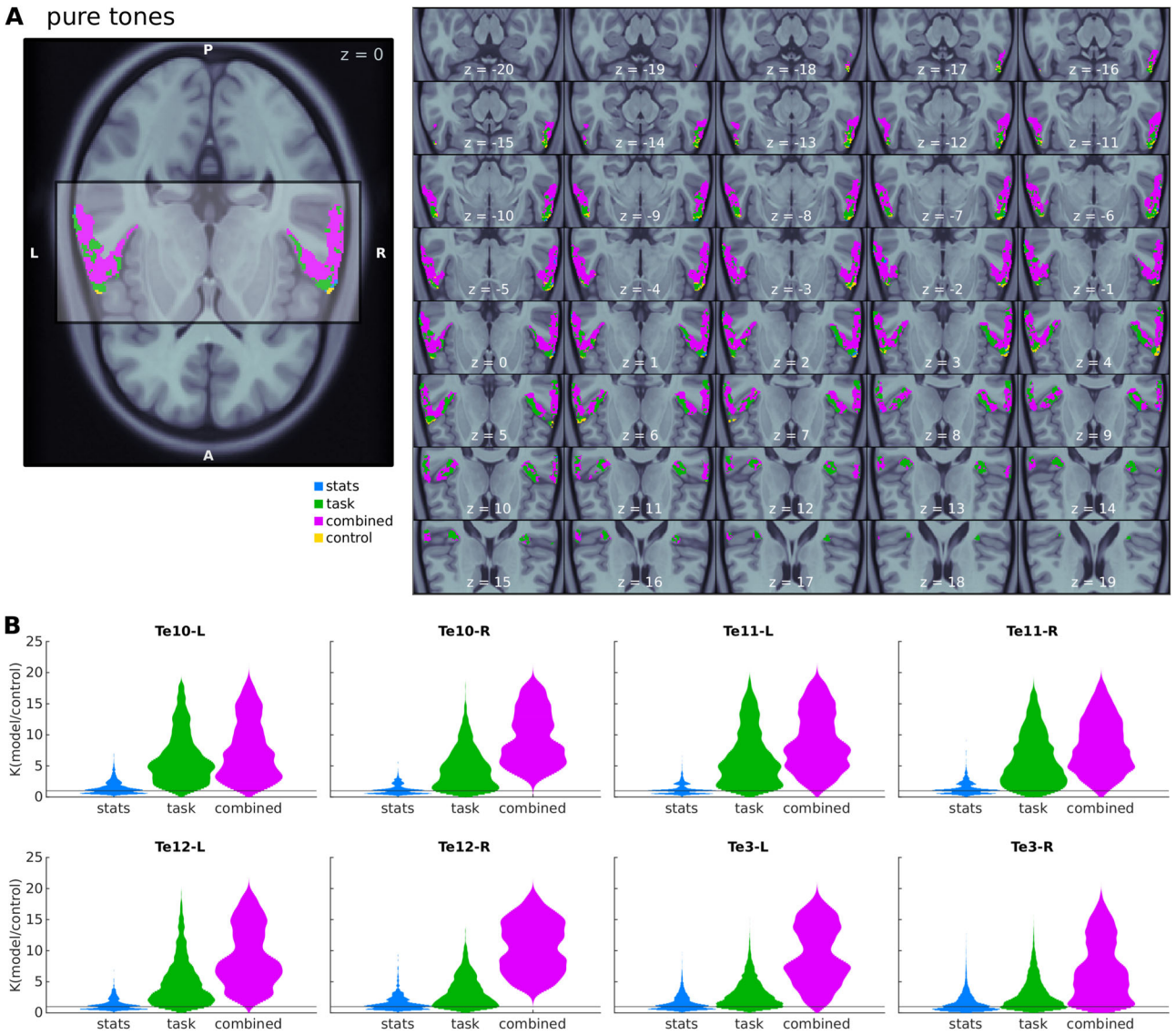

Figure 8. Prevalence of each model in AC for pure tones. ***A***, Map detailing which model best explains the responses to pure tones in each of the voxels of the AC. Colors indicate the model with the highest posterior density at each voxel. Blue voxels are best explained by the *stats* model, green voxels by the *task* model, and purple voxels by the *combined* model. ***B***, Distributions (kernel-density estimations) of the posterior densities of each model across voxels of each of the cortical fields for the pure tone stimuli.

$K_{combined/task}$ might, as the prevalence of the *task* model, be partially driven by tSNR heterogeneity in some of the ROIs.

Discussion

Hierarchical processing is the cornerstone of predictive coding (Friston, 2003a). Here we addressed the question whether inconsistent predictions derived from the task instructions and from the stimulus statistics are combined to compute prediction error. The main result is the robust presence of regions of the IC, MGB, and the AC that compute prediction error with respect to both sets of predictions. This result was consistent for pure tones and frequency sweeps. The relative size of these regions varied between the two families of stimuli: most voxels of bilateral IC, MGB, and AC encoded prediction error with respect to both sets of predictions in the pure tone experiment; this was also the case for the AC, but not for IC and MGB in the FM-sweeps experiment, where a majority of voxels in bilateral IC and MGB seemed to compute prediction error only with respect to the *task-informed* predictions. The different results in IC and MGB with the two sets of stimuli are however possibly driven by difference on the tSNR across studies. Independently of these differences, the presence of regions computing prediction error with respect to both sets of predictions in both experiments demonstrates that, at least in the auditory modality, predictive processing is powered by a complex system of transmission of predictions that escapes the linearity often assumed in the predictive coding literature (Spratling, 2017; Keller and Mrsic-Flogel, 2018). The corticofugal bundles that directly connect the AC with the MGB, IC, and superior olivary complex (Hackett and Kaas, 2004; Lee and Sherman, 2011; Schofield, 2011) might be responsible for the non-linear transmission of the task-informed predictions to nuclei of the subcortical pathways.

Our previous (Tabas et al., 2020, 2021; Stein et al., 2022) and the present results contradict conclusions drawn by studies using the *local–global* paradigm in humans, which assume that predictions are only communicated to the immediately lower processing level (Bekinschtein et al., 2009). In these paradigms, *local*

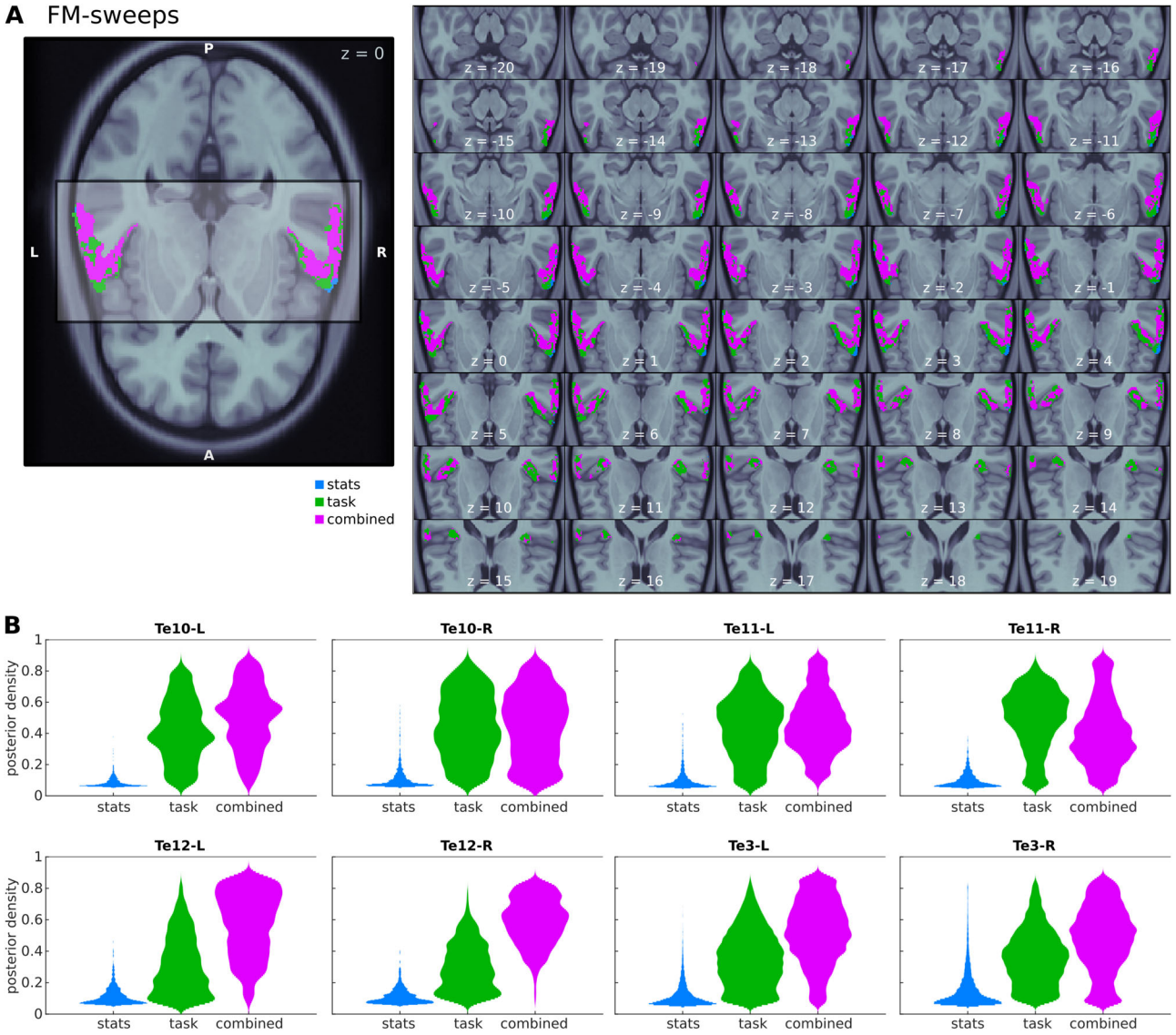

Figure 9. Prevalence of each model in AC for FM-sweeps. ***A***, Map detailing which model best explains the responses to FM-sweeps in each of the voxels of the AC. Colors indicate the model with the highest posterior density at each voxel. Blue voxels are best explained by the *stats* model, green voxels by the *task* model, and purple voxels by the *combined* model. ***B***, Distributions (kernel-density estimations) of the posterior densities of each model across voxels of each of the cortical fields for the FM-sweep stimuli.

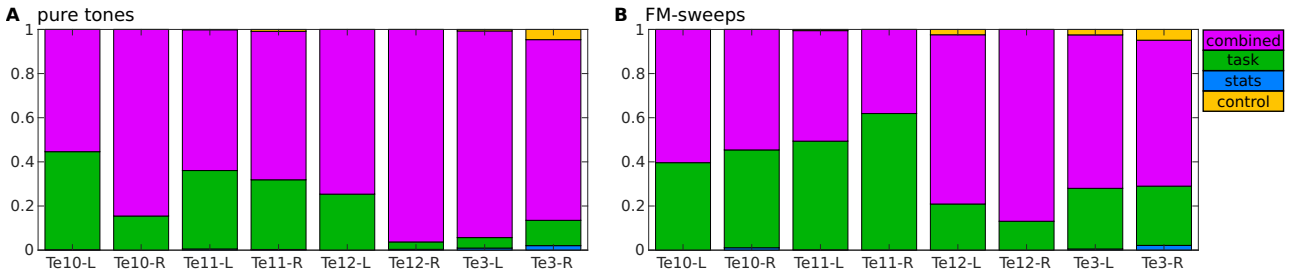

Figure 10. Prevalence of each model in each cortical field. Bars show the prevalence of each of the models across cortical fields for the pure tone ***A*** and FM-sweep ***B*** data. Blue bars correspond to voxels that are best explained by the *stats* model, green bars to voxels best explained by the *task* model, and purple bars to voxels best explained by the *combined* model.

predictions are based on the repetition of the tone *A*, while prediction referred as to *global* are based on the repetition of the melodic phrase *AAAAB*. Human E/MEG studies reported that prediction error to *local* predictions were present in primary sensory cortex whereas prediction error to *global* predictions were found in frontal cortex (Bekinschtein et al., 2009; Wacongne et al., 2011; Chennu et al., 2013; Recasens et al., 2014; El Karoui et al., 2015; Dürschmid et al., 2016; Nourski et al., 2018). Similar results have been shown using the same paradigm in primates (Chao et al., 2018) and in variations of the paradigm in

humans on the auditory (Uhrig et al., 2014, 2016; Maheu et al., 2019; Jiang et al., 2022), somatosensory (Naeije et al., 2016), and (auditory-cued) visual (Kouider et al., 2015) modalities. It is unclear why these previous studies did not find combined *local* and *global* prediction error in human AC. One possibility is that the *global* predictions from the *local–global* paradigm are functionally different from the *task-based* predictions elicited by our paradigm. This is likely the case, since *global* predictions are elicited by the local statistics of melodic phrases, whereas the *task-based* predictions are elicited by an inferential process that requires abstract reasoning. Another possibility is that prediction error to the *global* prediction was assigned only to frontal areas due to limitations of E/MEG (e.g., low spatial resolution). In non-human primates, three fMRI studies using the *local–global* paradigm have reported combined prediction error in AC (Uhrig et al., 2014, 2016; Jiang et al., 2022). Although two of the studies found only *local* prediction error in the MGB (Uhrig et al., 2014, 2016), one study found combined prediction error in MGB and only *local* prediction error in IC (Jiang et al., 2022). fMRI studies using the local–global paradigm in humans might be necessary to fully understand whether these divergences are species-specific differences or methods-related.

It is tempting to hypothesize that the *stats* model reflects habituation: if the receptive fields of deviant and standard would overlap significantly, then a model assuming habituation to the standard would show similar properties than the *stats* model, which assumes that the responses to the deviant will be stronger the larger the mismatch between deviant and standard Δ. However, in the pure tone experiment, with frequencies around $|f| \sim 1.5$ kHz, the expected equivalent rectangular bandwidth is $ERB \simeq 37.7$ Hz (Glasberg and Moore, 1990). Therefore, although the receptive fields might overlap subtly for deviant-standard combinations at the smallest $\Delta = 45$ Hz, habituation to the standard is unlikely to affect the responses to the deviant for $\Delta = 100$ Hz or $\Delta = 145$ Hz. This is also unlikely for the FM-sweep data, where deviant–standard combinations included FM-sweeps with opposite FM-modulation direction. Moreover, since habituation is a passive ubiquitous phenomenon in neural systems (Friauf et al., 2015), a model encoding habituation would affect all voxels in auditory ROIs. However, the *task* model is the best explanation for the data in some segments of the ROIs, even in regions for which there was no apparent correlation between the posterior density of the *combined* model and the tSNR (e.g., Te1.1-R, where this correlation was $\rho < 0$, but the *task* model was the best explanation of the data for around a third of the voxels; Fig. 10).

It could also be hypothesized that the *task-informed* and *combined* models could provide for good explanation for the data if the responses were modulated by attention-driven gain modulation (e.g., by mediation of the pulvinar; Kanai et al., 2015). Tones in positions 4 and 5 are indeed the most relevant for the task: if the responses were simply modulated by attention, the *task* model, where responses to positions 4 and 5 are higher than to the remaining tones, would explain the data better than the *stats* model, where all deviants 4–6 elicit the same response. However, previous analyses (Tabas et al., 2020, 2021) showed that (1) responses to deviants in positions 4 and 5, where participants were expected to show the same attention engagement, were significantly different and scaled by predictability (Tabas et al., 2020, 2021) and (2) the magnitude of the responses to deviants in position 6 and standards in positions 7 and 8 were statistically indistinguishable, even though, under non-attended listening, responses to deviants are always higher than to standards (Cacciaglia et al., 2015). The only explanation compatible with the different responses observed to the different deviant positions is that the activity encodes prediction error with respect to the task-informed predictions.

Our results show a generally higher prevalence of the *combined* model in pure tones than in FM-sweeps. One possibility is that these differences are driven by our decision of encoding Δ for FM-sweeps as differences in modulation rate only. Both FM-direction and rate are typically studied as independent features (Lui and Mendelson, 2003; Hsieh et al., 2012; Geis and Borst, 2013; Altmann and Gaese, 2014; Issa et al., 2016), and single neurons in the auditory pathway are usually either selective to FM-direction or to FM-rate. Therefore, FM might be encoded in a two-dimensional feature space in the brain. We could incorporate a contribution of FM-direction to Δ as an extra parameter in the models used to analyze the FM-sweep data, but adding another regressor would have increased the dependence of the log-evidence of the *combined* model on the tSNR even further.

Our study did not address whether subcortical pathways can adaptively track changes in the local statistics of the stimuli: in our paradigm, stimulus regularity is kept constant across the experiment, which arguably hampers the interpretability of the *stats* model. Future work could address whether the auditory pathway dynamically adapts to the local statistics using paradigms with varying stimulus regularities.

Another possible limitation of our study is the potential anatomical imprecision of the location subdivisions of the AC and the MGB. Due to the macroanatomical variability of the superior temporal plane in human subjects, it is possible that the mappings reported in Figures 3 and 4 do not exactly correspond to the microstructure boundaries of the auditory regions. Similarly, our 1.5 mm and 1.75 mm isotropic voxels might have been too coarse to precisely differentiate between primary and secondary subdivisions of the MGB. Therefore, the relatively homogeneous results we reported across subdivisions of the MGB (Fig. 7) and fields of AC (e.g., Fig. 10) should be considered with caution.

Predictive coding has gone a long way since it was first explicitly theorized in the 1990s (Mumford, 1992), evolving from a theory explaining extra-classical receptive field properties in visual cortex (Rao and Ballard, 1999) to a full hierarchical theory of sensory processing (Friston and Kiebel, 2009; Keller and Mrsic-Flogel, 2018). Here we have taken a step forward by questioning the assumed linearity (Friston, 2003b; Friston and Kiebel, 2009; Spratling, 2017; Keller and Mrsic-Flogel, 2018) of its hierarchical architecture. Understanding the interplay between multi-level predictions is crucial to understand how natural sensory processing occurs. For instance, predictive speech processing involves contextual, semantic, grammatical, phonetic, and vocal predictions (Kuperberg and Jaeger, 2016; Heilbron et al., 2020; Choi et al., 2021). To extract meaningful messages from noisy and ambiguous speech signals, the human brain should be able to compute independent prediction errors to all those independent predictions. Our findings suggest that, at sensory stages of the processing hierarchy, prediction error units are indeed capable of testing multiple predictions. The auditory pathway might exploit the corticofugal lines directly connecting the AC with the MGB, IC, and superior olivary nucleus for the direct transmission of predictions, bypassing the linear hierarchy often assumed in the literature (Keller and Mrsic-Flogel, 2018). This intricate system of descending connections might be responsible for our exquisite capacity to decode predictable information from noisy sensory inputs.

References

Altmann CF, Gaese BH (2014) Representation of frequency-modulated sounds in the human brain. Hear Res 307:74–85.

Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC (2011) A reproducible evaluation of ANTs similarity metric performance in brain image registration. NeuroImage 54:2033–2044.

Baron-Cohen S, Wheelwright S, Hill J, Raste Y, Plumb I (2001) The "Reading the mind in the eyes" test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. J Child Psychol Psychiatry 42:241–251.

Bekinschtein TA, Dehaene S, Rohaut B, Tadel F, Cohen L, Naccache L (2009) Neural signature of the conscious processing of auditory regularities. Proc Natl Acad Sci U S A 106:1672–1677.

Brainard DH (1997) The psychophysics toolbox. Spat Vis 10:433–436.

Brant-Zawadzki M, Gillan GD, Nitz WR (1992) MP RAGE: a three-dimensional, T1-weighted, gradient-echo sequence—initial experience in the brain. Radiology 182:769–775.

Cacciaglia R, Escera C, Slabu L, Grimm S, Sanjuán A, Ventura-Campos N, Ávila C (2015) Involvement of the human midbrain and thalamus in auditory deviance detection. Neuropsychologia 68:51–58.

Chao ZC, Takaura K, Wang L, Fujii N, Dehaene S (2018) Large-scale cortical networks for hierarchical prediction and prediction error in the primate brain. Neuron 100:1252–1266.e3.

Chennu S, Noreika V, Gueorguiev D, Blenkmann A, Kochen S, Ibáñez A, Owen AM, Bekinschtein TA (2013) Expectation and attention in hierarchical auditory prediction. J Neurosci 33:11194–11205.

Choi HS, Marslen-Wilson WD, Lyu B, Randall B, Tyler LK (2021) Decoding the real-time neurobiological properties of incremental semantic interpretation. Cereb Cortex 31:233–247.

Denckla MB, Rudel RG (1976) Rapid automatized naming (R.A.N.): dyslexia differentiated from other learning disabilities. Neuropsychologia 14:471–479.

Dürschmid S, Edwards E, Reichert C, Dewar C, Hinrichs H, Heinze HJ, Kirsch HE, Dalal SS, Deouell LY, Knight RT (2016) Hierarchy of prediction errors for auditory events in human temporal and frontal cortex. Proc Natl Acad Sci U S A 113:6755–6760.

El Karoui I, et al. (2015) Event-related potential, time-frequency, and functional connectivity facets of local and global auditory novelty processing: an intracranial study in humans. Cereb Cortex 25:4203–4212.

Fischl B, et al. (2002) Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33:341–355.

Friauf E, Fischer AU, Fuhr MF (2015) Synaptic plasticity in the auditory system: a review. Cell Tissue Res 361:177–213.

Friston K (2003a) Functional integration in the brain. In: *Human brain function*, (Frackowiak RSJ, Friston KJ, Frith CD, Dolan RJ, Price CJ, Zeki S, Ashburner JT, Penny WD, eds) Ed 2, Vol. 68, pp 971–997. Cambridge, MA: Academic Press.

Friston K (2003b) Learning and inference in the brain. Neural Netw 16: 1325–52.

Friston K (2005) A theory of cortical responses. Philos Trans R Soc Lond B Biol Sci 360:815–36.

Friston K, Kiebel S (2009) Predictive coding under the free-energy principle. Philos Trans R Soc Lond B Biol Sci 364:1211–21.

Friston K, Zarahn E, Josephs O, Henson R, Dale A (1999) Stochastic designs in event-related fMRI. NeuroImage 10:607–619.

Geis H-RAP, Borst JGG (2013) Intracellular responses to frequency modulated tones in the dorsal cortex of the mouse inferior colliculus. Front Neural Circuits 7:2002–2016.

Glasberg BR, Moore BC (1990) Derivation of auditory filter shapes from notched-noise data. Hear Res 47:103–138.

Gorgolewski K, Burns CD, Madison C, Clark D, Halchenko YO, Waskom ML, Ghosh SS (2011) Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in Python. Front Neuroinform 5;13.

Hackett TA, Kaas JH (2004) Auditory cortex in primates: functional subdivisions and processing streams. In: *The cognitive neurosciences III*, Ed 3, pp 215–232. Cambridge, MA: MIT Press.

Heilbron M, Armeni K, Schoffelen J-M, Hagoort P, De Lange FP (2020) A hierarchy of linguistic predictions during natural language comprehension. bioRxiv:2020.12.03.410399.

Hsieh I-H, Fillmore P, Rong F, Hickok G, Saberi K (2012) FM-selective networks in human auditory cortex revealed using fMRI and multivariate pattern classification. J Cogn Neurosci 24:1896–1907.

Hu B (2003) Functional organization of lemniscal and nonlemniscal auditory thalamus. Exp Brain Res 153:543–549.

Issa JB, Haeffele BD, Young ED, Yue DT (2016) Multiscale mapping of frequency sweep rate in mouse auditory cortex. Hear Res 344:207–222.

Jenkinson M, Beckmann CF, Behrens TE, Woolrich MW, Smith SM (2012) Fsl. NeuroImage 62:782–790.

Jiang Y, Komatsu M, Chen Y, Xie R, Zhang K, Xia Y, Gui P, Liang Z, Wang L (2022) Constructing the hierarchy of predictive auditory sequences in the marmoset brain. eLife 11:1–21.

Kanai R, Komura Y, Shipp S, Friston K (2015) Cerebral hierarchies: predictive processing, precision and the pulvinar. Philos Trans R Soc Lond B Biol Sci 370:20140169.

Kasper L, et al. (2017) The PhysIO toolbox for modeling physiological noise in fMRI data. J Neurosci Methods 276:56–72.

Keller GB, Mrsic-Flogel TD (2018) Predictive processing: a canonical cortical computation. Neuron 100:424–435.

Kouider S, Long B, Le Stanc L, Charron S, Fievet AC, Barbosa LS, Gelskov SV (2015) Neural dynamics of prediction and surprise in infants. Nat Commun 6:8537.

Kuperberg GR, Jaeger TF (2016) What do we mean by prediction in language comprehension? Lang Cogn Neurosci 31:32–59.

Lee CC, Sherman SM (2011) On the classification of pathways in the auditory midbrain, thalamus, and cortex. Hear Res 276:79–87.

Lui B, Mendelson JR (2003) Frequency modulated sweep responses in the medial geniculate nucleus. Exp Brain Res 153:550–553.

Maheu M, Dehaene S, Meyniel F (2019) Brain signatures of a multiscale process of sequence learning in humans. eLife 8:1–24.

Marques JP, Kober T, Krueger G, van der Zwaag W, Van de Moortele PF, Gruetter R (2010) MP2RAGE, a self bias-field corrected sequence for improved segmentation and T1-mapping at high field. NeuroImage 49:1271–1281.

Mihai PG, Moerel M, de Martino F, Trampel R, Kiebel S, von Kriegstein K (2019) Modulation of tonotopic ventral medial geniculate body is behaviorally relevant for speech recognition. eLife 8:1–28.

Moerel M, De Martino F, Formisano E (2014) An anatomical and functional topography of human auditory cortical areas. Front Neurosci 8:1–14.

Moerel M, De Martino F, Uğurbil K, Yacoub E, Formisano E (2015) Processing of frequency and location in human subcortical auditory structures. Sci Rep 5:17048.

Morosan P, Rademacher J, Schleicher A, Amunts K, Schormann T, Zilles K (2001) Human primary auditory cortex: cytoarchitectonic subdivisions and mapping into a spatial reference system. NeuroImage 13:684–701.

Mumford D (1992) On the computational architecture of the neocortex II: the role of cortico-cortical loops. Biol Cybern 66:241–251.

Naeije G, Vaulet T, Wens V, Marty B, Goldman S, De Tiège X (2016) Multilevel cortical processing of somatosensory novelty: a magnetoencephalography study. Front Hum Neurosci 10:1–12.

Nourski KV, Steinschneider M, Rhone AE, Kawasaki H, Howard MA, Banks MI (2018) Processing of auditory novelty across the cortical hierarchy: an intracranial electrophysiology study. NeuroImage 183:412–424.

Penny W, Kiebel S, Friston K (2003) Variational Bayesian inference for fMRI time series. NeuroImage 19:727–741.

Rao RPN, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci 2:79–87.

Recasens M, Grimm S, Wollbrink A, Pantev C, Escera C (2014) Encoding of nested levels of acoustic regularity in hierarchically organized areas of the human auditory cortex. Hum Brain Mapp 35:5701–5716.

Rosa M, Bestmann S, Harrison L, Penny W (2010) Bayesian model selection maps for group studies. NeuroImage 49:217–224.

Schofield BR (2011) Central descending auditory pathways. In: *Auditory and vestibular efferents*, Springer handbook of auditory research (Ryugo D, Fay R, eds), ch. 9, pp 261–290. New York, NY: Springer.

Sitek KR, Gulban OF, Calabrese E, Johnson GA, Lage-castellanos A, Moerel M, Ghosh SS, Martino FD (2019) Mapping the human subcortical auditory system using histology, postmortem MRI and in vivo MRI at 7T. eLife 8:e48932.

Spratling MW (2017) A review of predictive coding algorithms. Brain Cogn 112:92–97.

Stein J, von Kriegstein K, Tabas A (2022) Predictive encoding of pure tones and FM-sweeps in the human auditory cortex. Cereb Cortex Commun 3:tgac047.

Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ (2009) Bayesian model selection for group studies. NeuroImage 46:1004–1017.

Suga N (2012) Basic acoustic patterns and neural mechanisms shared by humans and animals for auditory perception. In: *Listening to speech*, Vol. 36, pp 159–181. Mahwah, NJ: Psychology Press.

Tabas A, Kiebel S, Marxen M, von Kriegstein K (2021) Fast frequency modulation is encoded according to the listener expectations in the human subcortical auditory pathway. arXiv:2108.02066.

Tabas A, Mihai G, Kiebel S, Trampel R, Von Kriegstein K (2020) Abstract rules drive adaptation in the subcortical sensory pathway. eLife 9: 1–19.

Uhrig L, Dehaene S, Jarraya B (2014) A hierarchy of responses to auditory regularities in the macaque brain. J Neurosci 34:1127–1132.

Uhrig L, Janssen D, Dehaene S, Jarraya B (2016) Cerebral responses to local and global auditory novelty under general anesthesia. NeuroImage 141:326–340.

Wacongne C, Labyt E, Van Wassenhove V, Bekinschtein T, Naccache L, Dehaene S (2011) Evidence for a hierarchy of predictions and prediction errors in human cortex. Proc Natl Acad Sci U S A 108:20754–20759.

Welvaert M, Rosseel Y (2013) On the definition of signal-to-noise ratio and contrast-to-noise ratio for fMRI data. PLoS One 8:e77089.