Evidence for implicit and adaptive deployment of precision weighting during passive listening: a simultaneous EEG/MEG study

Françoise Lecaignard[1,2,3 *], Olivier Bertrand[1,2], Anne Caclin[1,2**], Jérémie Mattout[1,2**]

Affiliations

[1] Lyon Neuroscience Research Center, CRNL; INSERM, U1028; CNRS, UMR5292; Brain Dynamics and Cognition Team, Lyon, F-69000, France
[2] University Lyon 1, Lyon, F-69000, France
[3] CERMEP Imaging Center, Lyon, F-69500, France

* Corresponding author
** Both authors equally contributed to this work.

Contact Information

Françoise Lecaignard
INSERM U1028 - CNRS UMR5292
Centre de Recherche en Neurosciences de Lyon
Equipe DYCOG
Centre Hospitalier Le Vinatier (Bât. 452)
95, Boulevard Pinel
69500 Bron, France
francoise.lecaignard@inserm.fr

Telephone: 0033 472138963
Fax: 0033 472138901

Highlights

- Hierarchical perceptual learning was studied with simultaneous EEG and MEG during passive listening of auditory oddball sequences.
- Neuro-computational modeling confirmed perceptual learning deploying within a fronto-temporal network.
- An implicit and computationally efficient adaptation of perceptual learning to the informational content of the sound sequence was revealed.
- Larger information relevance yielded a larger information integration in perceptual learning.
- The context-dependent tuning of sensory prediction errors was associated to cortical inhibitory mechanisms.

Keywords

Simultaneous EEG-MEG, Dynamic Causal Modeling, Computational Modeling, Fused Source Reconstruction, Mismatch Negativity, Deviance Processing, Implicit Learning, Predictive Coding, Hierarchical Perceptual Learning, Precision-weighted Prediction Error

ABSTRACT

In the predictive coding view of perception, the context-sensitive filtering of incoming sensations is achieved through the precision weighting (PW) of sensory prediction errors (PE). We aimed at seeking empirical evidence for this context-dependent, adaptive computational process and the respective neurophysiological encoding of PW and PE during passive listening. We hypothesized that the implicit learning of contextual (second-order) regularities would exert a top-down influence onto (first-order) sensory processing. We therefore manipulated the predictability of oddball sequences. Using learning models and dynamic causal models (DCM) of effective connectivity, we reveal the trial-to-trial dynamics of perceptual learning and its implementation within a fronto-temporal network. Essentially, we show an implicit adaptation of perceptual learning to the informational content of the environment. Under predictability, it translates into a larger information integration leading to an increased learning rate (PW), associated with a decreased neuronal self-inhibition, and a decreased (precision-weighted) PE, conveyed by decreased forward connectivity.

INTRODUCTION

Predictive brain theories have challenged cognitive neuroscience research in the recent years, by proposing a general model of human cognition based on a computational description of the brain as a Bayesian agent in constant interaction with its environment (Dayan et al. 1995; Friston 2012). In this view, information in the brain is processed probabilistically, in a way that accounts for uncertainty in an optimal fashion, by estimating the relative precision of the various incoming sensations and prior knowledge, respectively. This dynamical process would correspond to what is commonly referred to as perceptual learning and would express by the deployment of context-dependent relative weightings of prior beliefs and sensory observations. Predictive brain theories have already succeeded in providing new (and testable) hypotheses, leading to new or revisited experimental paradigms to address how the brain adapts under uncertainty (Behrens et al. 2007; Tenenbaum et al. 2011; Vossel et al. 2014; Dehaene et al. 2015; Auksztulewicz & Friston 2015) . A striking example is the growing interest in how stimulus predictability modulates some long-established perceptual responses (Schröger et al. 2015; Heilbron & Chait 2017). However, beyond the great insights that those theories have already provided, the precise mental processes and neurophysiological mechanisms that subsume perceptual learning remain elusive. The Bayesian brain hypothesis involves the neural coding of the relevance (or precision) of context-dependent information (Friston 2005). To date, no direct evidence of such a calculation and associated neuronal implementation has been reported. In this study, we sought to shed light on this important question by considering a simple form of perceptual learning, that could work during the passive listening of sound sequences. To that aim, we addressed the effect of the contextual

relevance of information (if it was to be learnt implicitly by the brain) on auditory processing, using physiological and psychological dynamical modeling fitted to simultaneous EEG-MEG data.

The view of perception as probabilistic inference has been augmented with predictive coding as a possible cognitive framework and a plausible neuronal implementation (Friston & Kiebel 2009; Kok & de Lange 2015; Markov & Kennedy 2013; Bastos et al. 2012). Contextual adaptation in predictive coding is driven by hierarchical *precision-weighted prediction errors* (PWPEs) induced by every incoming sensation (Mathys et al. 2011). At the first (lower) level of the hierarchy, sensory PWPE is computed as the combination of 1) a prediction error (PE) quantifying the unexpectedness of the current event, and 2) a precision weight (PW) balancing the expected reliability of the current sensation relative to the estimated reliability of internal predictions. Such context-sensitive PW enables to filter out errors that are irrelevant and, conversely, to gate errors that are relevant for optimal belief updating. This endows the brain with an efficient and flexible perceptual process. Interestingly, PW dysfunctions have been suggested to be at the heart of psychopathology (Adams et al. 2013). Hence, testing its neural account strongly deserves serious consideration. So far, few but very encouraging results from trial-by-trial analysis of neuroimaging data have started addressing the computation of hierarchical PWPEs in a visual cue-association learning task (Iglesias et al. 2013) or an auditory detection task (Ostwald et al. 2012; Lieder et al. 2013). Here, we extend such studies by addressing the distinct neural representations of sensory PE and PW.

We used an original design of the well-known oddball paradigm in order to elicit deviance responses (Escera et al. 2014) in different contexts of information relevance.

We conducted simultaneous EEG and MEG recordings (Lopes da Silva 2013) with the aim to reveal the subtle changes of brain activity induced by our contextual manipulation, if the brain was to encode perceptual learning. Oddball sequences incorporate unexpected deviant stimuli and are therefore particularly well suited to test and refine the theoretical hypothesis of a close link between evoked responses and PWPE (Friston 2005). Such a link is supported by neuroimaging findings, providing evidence at both the cognitive (Ostwald et al. 2012; Lieder et al. 2013) and neurophysiological (Garrido, Kilner, Kiebel, et al. 2009; Moran et al. 2013) levels. Here, we implemented a model-driven contextual manipulation of oddball sequences that, as we shall see, enables modulating sensory PE and PW in an opposite direction. The resulting and simple passive listening paradigm enabled us to disambiguate the cortical representations of PW and PE when the (Bayesian) brain processes auditory sequences. In a preliminary study, we evidenced the brain's ability to learn implicitly such different contexts (predictable and unpredictable) using a typical auditory oddball sequence (Lecaignard et al. 2015). We next examined how such contextual learning modulates sound processing. Our investigation required a step-by-step computational and neuronal modeling approach that is presented here. We found empirical evidence for Bayesian perceptual learning (optimization of PE and PW) within a fronto-temporal network. Moreover, and critically, predictability successfully exhibited changes in learning models at the cognitive level. Our findings reveal a larger integration of information in perceptual learning under a structured environment. The physiological counterpart of such implicit tuning of learning could also be measured in the synaptic connectivity established by deviance processing. Precisely, we observed an increase in self-inhibition within cortical areas (coding for PW) together with a decrease in forward connectivity between cortical areas (coding for PE or PWPE) with predictability.

RESULTS

The results presented here were obtained from a four-step modeling procedure. Its ultimate goal was to test the existence of implicit perceptual learning, during passive listening and, where appropriate, to disclose the neurophysiological signature of PE and PW computation. Trial-by-trial computational modeling and dynamic causal models (DCM) of evoked responses (Kiebel et al. 2009) were applied to characterize perceptual learning at both the cognitive and neurophysiological levels, respectively. First analysis aimed at testing the hypothesis of perceptual learning at play during oddball sound processing using computational models (*PL-oddball* analysis). Secondly, using DCM, we addressed the cortical implementation of this learning process (*DCM-oddball* analysis). Importantly, we manipulated the predictability of the auditory oddball sequences, in the aim of revealing and disentangling PE and PW representations. In a third step, we thus tested whether the brain adapts its learning style to the contextual manipulation (*PL-predictability* analysis). The latter provided clear predictions regarding the mapping of PE and PW onto physiology, which we tested in a final step, using Dynamic Causal Modeling (*DCM-predictability* analysis).

We report findings from 20 healthy volunteers undergoing a passive auditory oddball paradigm with simultaneous EEG-MEG recordings. Oddball sequences embedding frequency deviants were delivered either within a purely deterministic context (condition PC, predictable context), or a pseudo-random one (condition UC, unpredictable context) (Figure 1). The implicit learning of the two different statistical structures was demonstrated in a previous study (Lecaignard et al. 2015), where we reported at the EEG sensor level a significant reduction of mismatch responses (at different latencies) under predictability while none of the participants reported having

noticed our contextual manipulation. In the present study, computational learning models were fitted to single-trial data whereas DCM ones rested on the simultaneous inversion of standard (preceding a deviant) and deviant evoked responses (see Material and Methods).

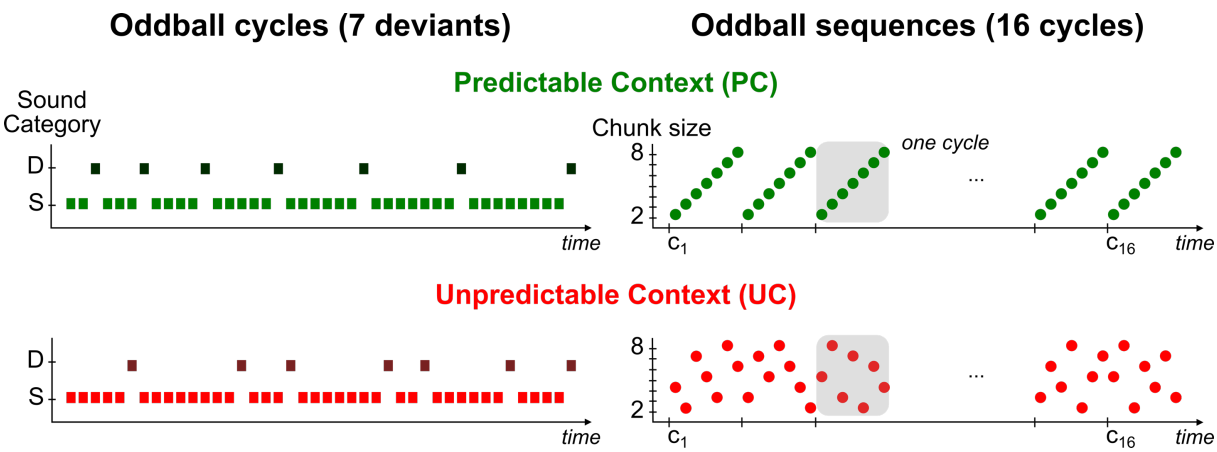

Figure 1. Experimental Design. *Left, Schematic view of typical oddball cycles in the predictable (green, top) and unpredictable (red, bottom) contexts. Both cycles correspond to the same deviant probability (p = 1/6) and the same distribution of deviants (D) among standards (S). They only differ in the transition between segments of repeating standards (chunks), which are sorted by increasing size in condition PC. Right, Representation of oddball sequences as a function of chunk size to reveal the different statistical structures applying on the same oddball rule in context PC (green, top) and UC (red, bottom). X-axis ticks indicate the beginning of each cycle (16 cycles comprise a sequence, from cycle* c_1 *to* c_{16}*). Grey rectangles delineate one exemplary cycle in both sequences.*

Perceptual learning of environmental regularities (*PL-oddball* analysis)

Under predictive brain theories, mismatch responses (including the Mismatch Negativity, MMN) elicited by unexpected deviants reflect prediction errors of a continuously updated model of the acoustic environment (Friston 2005). We tested this hypothesis against alternative cognitive processes that did not involve perceptual

learning (presented below), and corresponding models were all defined as a two-level linear model of the form:

$$\begin{cases} y = X\theta_1 + \theta_2 + \varepsilon_1 \\ \theta_1 = 0 + \varepsilon_2 \\ \theta_2 = 0 + \varepsilon_3 \end{cases} \quad (1)$$

Where y indicates the reconstructed cortical activity (informed by a fused EEG-MEG source inversion) in the form of a vector of trial-by-trial activity at a particular sample of the peristimulus time; X is defined for each model and represents the predicted trajectory of PWPE over the sound sequence; $\{\theta_1, \theta_2\}$ and $\{\varepsilon_1, \varepsilon_2, \varepsilon_3\}$ refer to Gaussian observation parameters and Gaussian noise, respectively. Vector y was defined for each time sample of the [-50 +350] ms epoch and for each cluster of the MMN network identified through group and fused EEG-MEG source reconstruction (Lecaignard et al., in prep). Six bilateral clusters were precisely located in the Heschl's Gyrus (HG), the Planum Polare (PP) and the Inferior Frontal Gyrus (IFG) (Figure 2.A). We considered a model space of seven cognitive models partitioned into three families, largely inspired by models used in a previous tactile oddball study (Ostwald et al. 2012):

- Family fam_{null} is made of a single model, the null model (M0) assuming that the brain response to every tones is the same way, or equivalently that EEG fluctuations only reflect random noise (i.e. $\theta_1 = 0$ in model Equation Eq.1).
- Family fam_{noL} contains two non-learning or static models, namely "change detection" (CD) and "linear change detection" (LinCD). Both assume that the brain simply compares each incoming sensation to the preceding one. In model CD, X_k at trial k is assigned to 0 if the kth stimulus is equal to the preceding one, and 1 otherwise. Model LinDC is similar to CD but assigns a prediction error proportional to the number of preceding sounds that differ from the one being

currently observed (Ostwald et al. 2012; Lieder et al. 2013).

- Family fam_L includes Bayesian perceptual learning models assuming that the brain estimates the probability μ to hear a standard (under a Bernoulli distribution, see Material and Methods). These models involve a temporal integration window that allows down-weighting the influence of past events and whose size is defined by a parameter (τ). We considered four different values for τ (2, 6, 10 and 100), leading to four models in fam_L. PWPE computed at each trial quantifies the belief update about μ and was defined as the Bayesian Surprise (BS; the Kulback-Leibler divergence between the prior and the posterior distribution of μ, see Material and Methods).

These models were all fitted to the reconstructed cortical activity in both PC and UC conditions. For each source and at each time sample of the peristimulus interval, model inversions were performed following a meta-Bayesian approach using the VBA toolbox (Daunizeau et al. 2014; Daunizeau et al. 2010) and model families were compared to each other using an RFX family-level inference (Penny et al. 2010). As shown in Figure 2.B, the Null hypothesis (fam_{null}) was found to outperform the other families, at every time sample and every cortical source, except for the time interval of the MMN. Precisely, posterior exceedance probability of family fam_L was found to be significantly greater than the ones of the other families between 150 ms and 200 ms (6 samples), in the left and right Heschl's Gyrus (HG), between 130 ms and 180 ms (6 samples) and at 150 ms (1 sample) in the left and right Planum Polare (PP), respectively, and from 130 ms to 200 ms (8 samples), and from 140 ms to 190 ms (6 samples) in the left and right Inferior Frontal Gyrus (IFG), respectively.

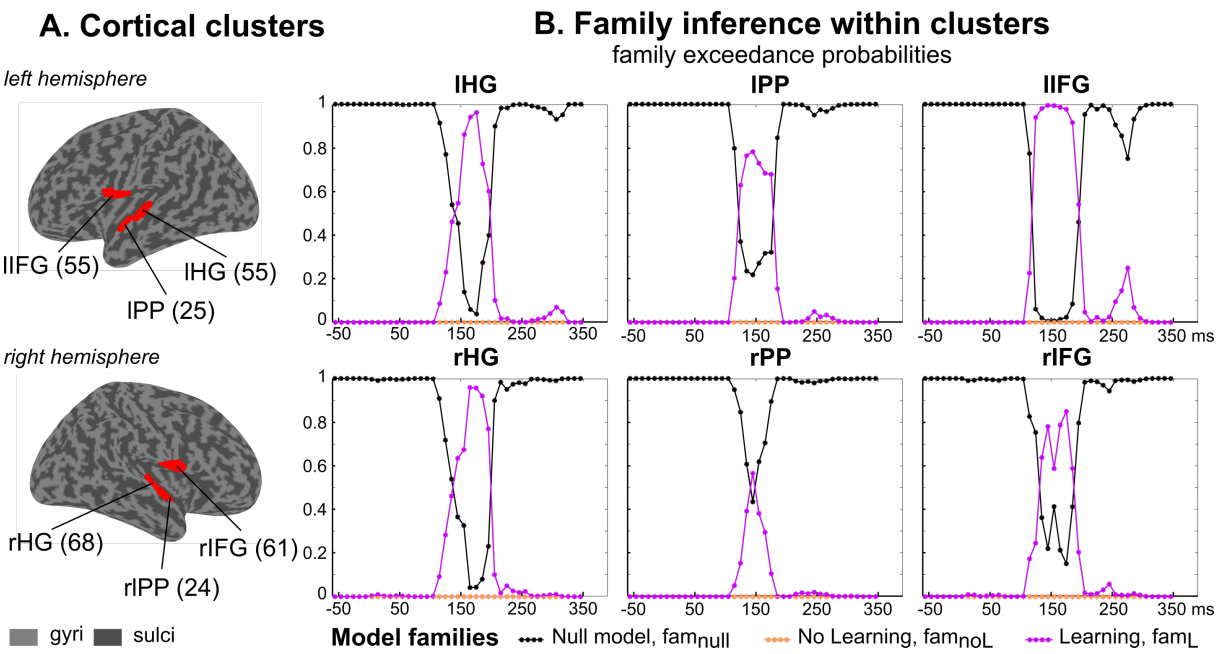

Figure 2. Computational modeling of deviance processing. *A) Cortical clusters involved in* PL-oddball *modeling are represented (red) over the inflated cortical surface of the SPM template brain* (Mattout et al. 2007)*. These six clusters are left, right Heschl's Gyrus (lHG, rHG), left, right Planum Polare (lPP, rPP) and left, right Inferior Frontal Gyrus (lIFG, rIFG). Total number of nodes in each cluster is indicated in parenthesis (see also Figure S2.B). Within each cluster, reconstructed single-trial traces (informed by EEG and MEG) were averaged over nodes (black dots) and confronted to trial-by-trial model predictions (for every model of* PL-oddball *model space) at each time sample within [-50 +350] ms. B) Family-wise Bayesian model comparison. For each cluster and at each time sample, family inference provides the estimated posterior family exceedance probability of each model family (fam_{null}: black, fam_{noL}: orange, fam_L: pink).*

Neurophysiological implementation of perceptual learning (*DCM-oddball* analysis)

The previous analysis revealed the implicit perceptual learning of acoustic regularities during an oddball sequence, at the latency of the MMN. This raises the question of the corresponding underlying mechanisms that, following a predictive coding view, could

rest on ascending error and descending prediction messages within a cortical hierarchy. DCMs for evoked responses have been recently augmented with a biologically plausible mesoscopic model of the cortical laminar circuitry (*Canonical Micro-Circuit*, CMC) which enables testing formally the mapping of computational learning quantities onto cortical representations (Bastos et al. 2012; Moran et al. 2013). Precisely, within each source, DCM with CMC assigns a prediction error unit to the superficial pyramidal (SP) cells, and accounts for the precision weighting of PEs by means of a self-inhibition parameter governing the excitability of subpopulation SP. Inspired by previous DCMs of the MMN (Garrido, Kilner, Kiebel, et al. 2009; Auksztulewicz & Friston 2016), and guided by current learning model predictions, we addressed the two following questions: what is the structure of the network engaged in typical oddball sequence processing? Within this auditory network, what modulation of effective connectivity supports deviant compared to standard tone processing? This DCM analysis was conducted with data from condition UC and combined EEG and MEG signals (see Material and Methods).

Network characterization. Here, we assessed the network architecture and input. Regarding the former, we hypothesized a four-level hierarchy composed of eight sources distributed bilaterally over (from the lowest to the highest level) HG, PP, IFG and a superior frontal area SF (Figure 3.A). We connected these sources with intra-hemispherical bidirectional (forward and backward) connections. Alternative hypotheses entailed two- and three-level networks allowing to test the hierarchical depth as well as the contribution of PP and SF sources, leading to five model families (*A1*, *A2*, *A3*, *A4* and *A5;* see Material and Methods). Regarding DCM inputs, all models included a direct thalamic inputs to bilateral HG. In addition, inputs targeting IFG sources (known to receive direct thalamic afferents) were tested as the current EEG

and MEG evoked responses suggested that frontal regions were activated prior to temporal ones (echoing a still-debated issue in the MMN literature, (Deouell 2007). The input factor thereby included two levels (*HG* and *HG-IFG*). Considering the integration over different forms of standard-to-deviant modulations of connectivity, the resulting model space comprised a total of 36 DCMs (for a full description see Material and Methods, *Network-Structure* model space). Each model inversion was performed on standard (the one preceding the deviant) and deviant evoked responses, for EEG and MEG data separately. Resulting explained variance averaged across models and subjects was equal to 92.5% (±10.5) and 78.1% (±11.6) in EEG and MEG, respectively. Posterior fusion of EEG and MEG DCMs (denoted p-MEEG) was achieved by applying Bayes' rule (see Material and Methods). In the p-MEEG modality, family level inference revealed that family *A5* outperformed other model families with a posterior confidence probability (*pcp*) and posterior exceedance probability (*pep*) of 0.68 and >0.99, respectively (Figure 3.B). Regarding the DCM inputs (Figure 3.C), family level inference was clearly in favor of models with inputs arriving in both HG and IFG sources (*pcp* / *pep*: 0.82 / >0.99).

Deviant-related changes on standard connectivity. Considering the winning DCM structure (architecture *A5* and input *HG-IFG*), we next examined the extrinsic and the intrinsic connectivity that govern the neural coupling between and within cortical sources, respectively. Importantly, every extrinsic and intrinsic connection strength accounted for standard processing and could be modulated by a trial-specific gain accounting for deviant-related changes (Trial-specific modulation, TSM). The present analysis rests on a model space composed of 14 models (*Coupling-Change* model space, see Material and Methods) designed to especially address TSMs. Explained variance was equal to 96.5% (±5.1) for EEG, and 87.1% (±7.4) for MEG. We found

evidence for forward, backward and intrinsic TSMs (*pcp* / *pep*: 0.68 / 0.96; 0.52 / 0.82; 0.82 / >0.99, resp.) in line with pioneering DCMs of the MMN (Garrido, Kilner, Stephan, et al. 2009; Auksztulewicz & Friston 2016). In order to characterize the deviant-related direction of change, we further examined group-level posterior estimates of TSM obtained with Bayesian model averaging (BMA) over model space and subjects (see Material and Methods). Over the entire network, we found an increase of the forward (1.145 ± 0.057; present in 5/6 forward connections; Figure 3.D) and the backward TSM (1.179 ± 0.073; 5/6 backward connections). Conversely, we found a decrease of the intrinsic TSM (0.947 ± 0.064; 5/8 self-inhibition connections).

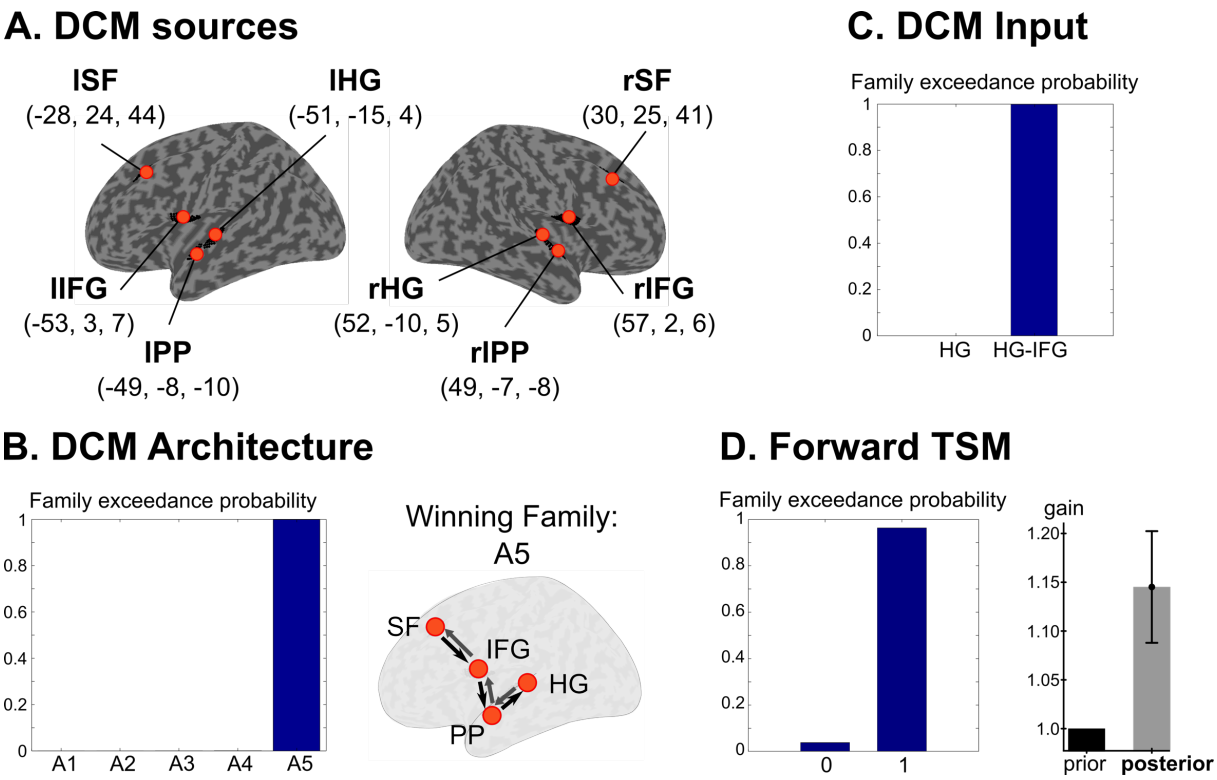

Figure 3. Dynamic causal modeling of deviance processing, using p-MEEG. *A) Cortical sources for DCM analysis. Each source is indicated schematically with orange dots on the inflated cortex (SPM template brain), with corresponding MNI coordinates in parenthesis. B) Family inference for DCM architecture: family exceedance probabilities (left) and corresponding network for the winning family A5 (right) following the notations and color codes of Figure S3. C) Family inference for DCM inputs (family*

exceedance probabilities). D) Family inference for standard-to-deviant trial-specific modulation (TSM) for the forward connectivity. Family exceedance probabilities of model family with disabled (0) and enabled TSM (1) is provided (left) and group-level BMA estimates of TSM values averaged over the DCM network is represented (right) with standard-error bar relative to network variability. Prior value of TSM (black bar) was set to 1 assuming no standard-to-deviant modulation. Labels for the cortical sources and model families are provided in the main text.

Differential effect of predictability on PE and PW.

The above *DCM-oddball* analysis confirmed the usefulness of oddball sequences and validity of the DCM approach to address the neural representation of perceptual learning within the auditory hierarchy. In particular, the larger forward coupling observed with deviants supports its assignation to ascending PWPEs (being larger for unexpected stimuli; Figure S1). To further refine the respective mapping of learning quantities (PE and PW) onto DCM representations, we compared the two experimental conditions (PC and UC) varying the informational value carried by similar stimuli, hence influencing sensory PE and PW differently. Precisely, lower PE and larger PW are generated under PC, due to the reduced surprise and the more structured environment, respectively. In other words, predictability yields smaller but more relevant PE signaled as such by larger PW. Passive listening was used to avoid presumable confounding effects of voluntary attention on PW (Feldman & Friston 2010). Our design manipulation allows to reveal the contextual (predictability) effect on deviance processing, if the brain was to capture the predictable rule and adapt sensory processing accordingly. This implicit learning process was already evidenced from the modulation of mismatch responses observed at the sensor-level on averaged ERPs (Lecaignard et al. 2015). The next important step here consisted in testing formally whether the perceptual learning evidenced in *PL-oddball* analysis adapted to

the predictable context, and how it manifests in terms of the underlying effective connectivity. For both levels of analysis, we conducted an ANOVA described in Material and Methods.

Contextual adaptation of perceptual learning (*PL-predictability* analysis)

As briefly mentioned in *PL-oddball* analysis, we introduced a temporal integration window in fam_L models as proposed by Ostwald et al. (2012), whose size is an increasing function of τ (see Material and Methods). Crucially, beyond the memory-based interpretation (the brain may arguably not be able to deal with long-standing information), this parameter endows the learning models with a flexible way to integrate past information and formalizes brain adaptation to its environment. Precisely, the decrease of BS with τ (see also Material and Methods) and the reduced mismatch responses measured in the predictable context led us to expect that passive sound processing under context PC would rest on a larger temporal integration window compared to the one under context UC. Prior to testing this prediction, we simulated the grand-averaged MMN amplitude for each condition and different values of τ (the MMN was computed as the difference of the average BS measured for deviants and for standards preceding a deviant; see Material and Methods). Results show that the (simulated) MMN amplitude decreases as τ increases. Interestingly, they also show a similar amplitude across contexts, for each τ value (Figure 4.A). This suggests that the sequence alone cannot explain the MMN reduction in condition PC and a reduced τ under a more predictable sequence (demonstrating an adaptive learning) is thus expected in order to explain the observed sensor-level difference in ERPs between UC and PC. Using EEG-MEG reconstructed cortical data, we performed model inversion separately for condition PC and UC, at every time sample exhibiting statistical significance in the *PL-oddball* analysis. Crucially, τ was no longer fixed here but

estimated (see Material and Methods). For each source and each parameter, posterior estimates were averaged across samples. As demonstrated by the ANOVA, larger τ values were found with condition PC compared to condition UC ($F_{(1,19)} = 10.13$; $p = 0.002$). On average across sources, τ was equal to 16.0 and to 21.3 with UC and PC, respectively (Figure 4.B). The ANOVA showed no other significant effect. As can be seen in Figure 4.C, the different τ values inferred in UC and PC generate different down-weightings of past observations, with precisely a larger amount of information integrated during perceptual learning in the predictable context.

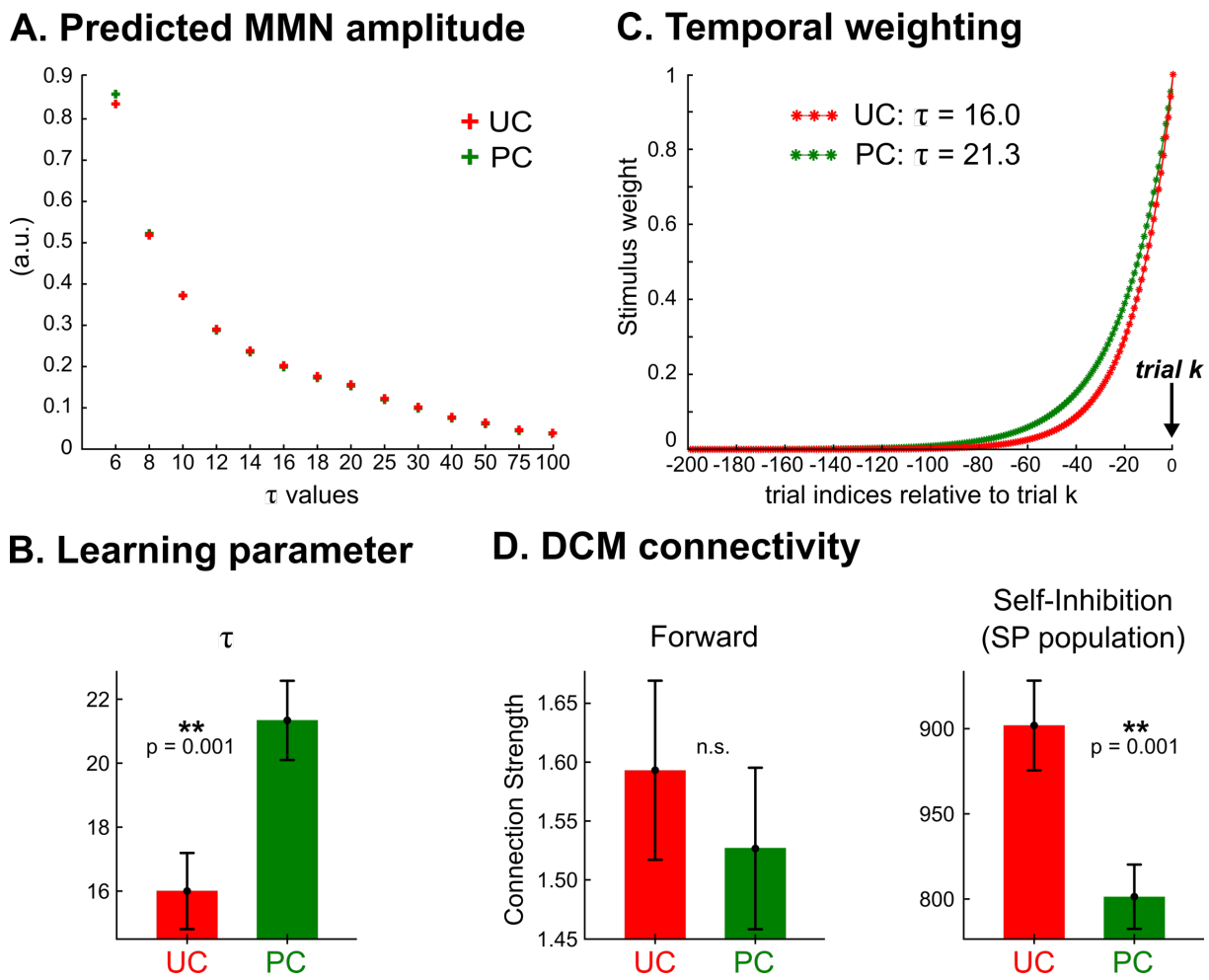

Figure 4. Effect of predictability on auditory processing. *Panels A, B and C refer to the cognitive modeling of our contextual manipulation (PL-predictability), and panel D to its physiological counterpart (DCM-predictability). A) Learning model predictions*

of the MMN amplitude (group average) as a function of τ for UC (red) and PC (green) sound sequences; arbitrary units (a. u.). These simulated MMN correspond to the difference in Bayesian Surprise elicited by deviants and by standards preceding a deviant (see Material and Methods). Note that in both contexts, the MMN amplitude decreases as τ increases. B) Effect of predictability on learning parameter τ. Posterior estimates of τ averaged at the group-level and over the six clusters exhibited a significant difference between conditions. C) Downweighting of past observations obtained with the posterior estimates of τ for conditions UC (red) and PC (green), respectively. D) Effect of predictability onto effective connectivity, using p-MEEG. Posterior estimates of forward (left) and self-inhibition (right) strengths measured in both conditions (UC: red, PC: green); averaged over the group and over the connections within the (DCM) network. SP: superficial pyramidal.

Neural correlates of PE and PW (*DCM-predictability* analysis)

This last analysis addressed the effect of predictability on effective connectivity during deviance processing, with the aim of clarifying the neural representation of key learning quantities (PE, PW). Based on the findings from the *PL-predictability* analysis, we focused specifically on extrinsic forward and intrinsic self-inhibition connections, and we expected:

- No difference between UC and PC in the forward and intrinsic TSMs. Indeed, the simulated MMN amplitudes (Figure 4.A) did not reveal any interaction between the contextual adaptation and stimulus deviance.
- Lower forward connection strength in condition PC, if it was to encode PE (or PWPE).
- Lower SP self-inhibition connection strength in condition PC, if it was to encode PW.

Note that these predictions apply to the entire DCM network for a twofold reason: learning models in fam_L are not hierarchical; and computational findings in *PL-*

deviance analysis did not exhibit any spatial specificity. BMA posterior estimates of DCM parameters were obtained for each subject and each condition over the whole *Coupling-Change* model space. Model inversions in the *DCM-deviance* analysis had been previously performed in condition PC and they provided explained variances equal to 97.8% (±2.7) for EEG, and to 87.5% (±5.5) for MEG. Using p-MEEG, the neural coupling at play during deviance processing was found very similar to the one in condition UC with deviant-related changes on the forward and intrinsic connections (*pcp* / *pep*: 0.73 / 0.99; 0.59 / 0.82, resp.). Likewise, larger forward trial-specific gain (in 4/6 connections; network average = 1.027 ± 0.042) and lower self-inhibition trial-specific gain (in 5/8 connections; network average = 0.956 ± 0.047) were measured. An ANOVA was performed on resulting BMA posterior estimates (p-MEEG modality) to test for the above-listed predictions (see Material and Methods). Regarding TSM, we found no significant main effect of *Condition* (forward: $F_{(1,19)}$=2.50; p=0.12; intrinsic: $F_{(1,19)}$=0.01; p=0.90). Secondly, a reduced forward connection strength was evidenced in condition PC compared to condition UC (average across the network with standard error: 1.53 ± 0.07 and 1.59 ± 0.08, resp.; maximum difference at the lower level of the network: 1.67 and 1.90, resp.) but did not reach significance (Figure 4.D, left). Finally, there was a reduced self-inhibition connection strength in population SP in condition PC compared to condition UC (801.4 ± 18.9 and 901.9 ± 26.3, resp.; $F_{(1,19)}$ = 10.52; p = 0.001), as shown in Figure 4.D (right).

DISCUSSION

Using a twofold computational modeling approach to explain the effect of an implicit contextual manipulation of auditory oddball sequences, the present findings reveal the occurrence of perceptual learning during passive listening and shed light on its underlying computational and neurophysiological mechanisms. It provides empirical support to current but largely qualitative predictive brain theories. Strikingly, this work also demonstrates the implicit adaptability of this perceptual learning to the changes in the statistical structure of the environment, in a way that optimizes the integration of relevant sensory information over time. Computationally speaking, this adaptation relies on the tuning of the precision weighting of prediction errors, a process which is known to be a cornerstone of Bayesian information processing (Mathys et al. 2014). For the first time, the neural encoding of this precision weight could be distinguished from the prediction error *per se* and directly attributed to inhibitory mechanisms.

Combining cognitive and physiological modeling.

Generative models (and their inversion) of EEG-MEG data in response to standard and deviant inputs were necessary to formally show that learning is a likely cause of the reduced mismatch responses measured at the sensor-level when manipulating sound predictability (Lecaignard et al. 2015). Our goal-directed step-by-step approach, combining psychological and physiological generative models made possible the distinction between PE and PW neural representations and thereby echoes David Marr's levels of analysis to provide a comprehensive description of the brain as an information processing system (Marr 1982).

At the cognitive level, trial-by-trial modeling enabled us to tackle the barely addressed question of a *memory trace formation* pertaining to the standard stimulus (Näätänen

et al. 2005; Recasens et al. 2015). Only the learning models could by essence predict the trial-to-trial fluctuations in brain responses, contrary to change detection models that overlook the information conveyed by all sounds but deviants and standards preceding a deviant (these latter models are equivalent to a typical MMN analysis on averaged evoked responses). Note that, in the *PL-oddball* statistical analysis, the learning models were selected at the latency of the MMN and no other latency. Beyond the traditional definition of the MMN, this temporally-resolved result sheds light on the processes at play during deviance processing (learning) and demonstrates both the feasibility and the potential of single trial analysis. It also highlights the importance of dynamic models to reveal these computational processes and neurophysiological mechanisms.

Regarding DCM, despite simplifications in modeling extrinsic and intrinsic connectivity, the recent CMC appears powerful to test the neural implementation of predictive coding (its message passing in particular), provided the functional role attributed to model parameters is valid (Bastos et al. 2012; Pinotsis et al. 2017). In addition to recent efforts to assess and reinforce DCM reliability (Pinotsis et al. 2017; Phillips et al. 2016), the fact that we could establish empirically a direct link between PW and self-inhibition of subpopulation SP constitutes a critical step in validating CMC-based DCM. Recent DCM studies strongly supported this PW / self-inhibition mapping, as they reported consistent modulations of SP self-inhibition by some experimental manipulations expected to influence PW under predictive coding (namely a cholinergic neuromodulation (Moran et al. 2013), sensory precision (Brown & Friston 2012), selective attention (Brown & Friston 2013) and predictability (Auksztulewicz et al. 2017)). Here, it is the implicit learning processed that we revealed and proved adaptable to the context, that fills the missing link enabling to relate PW and SP self-inhibition directly. This important finding calls for further studies aiming at

characterizing predictive coding specific implementation at different stages of the cortical hierarchy, namely to address what mechanisms subsume the adaptation of the learning process that we observed.

Perceptual learning within the auditory hierarchy.

We investigated the psychological and physiological mechanisms underlying regularity processing, a cognitive function that is widely admitted to contribute to mismatch responses. Our report of perceptual learning evidenced at the MMN latency and deploying within a hierarchical fronto-temporal network is very much in line with previous trial-by-trial computational modeling (Ostwald et al. 2012; Lieder et al. 2013) and DCM (Garrido, Kilner, Stephan, et al. 2009) of the MMN. The present findings thus reinforce the predictive coding account of hierarchical auditory processing.

In hierarchical learning, higher levels are thought to compute errors and related precisions that pertain to more stable or slow-changing (contextual) stimulus features (Friston & Kiebel 2009). This allows adapting the learning style to the contextual relevance (put simply, should I learn or not?). In the auditory modality, there is a growing convergence of findings supporting hierarchical error processing within a fronto-temporal (Garrido et al. 2007; Chennu et al. 2013; Escera et al. 2014; Lecaignard et al. 2015; Dürschmid et al. 2016). Similarly, direct computational evidence for low- and high-level errors could be established in the visual modality using a cue-association learning task (Iglesias et al. 2013). In the same vein, Bayesian learning was evidenced during a counting task in Ostwald et al. (2012) and showed a spatio-temporal progression, starting with early effects in somatosensory regions, followed by later ones in frontal and medial cingulate cortices. In the current *PL-oddball* analysis however, no similar differentiation could be observed. This is arguably inherent to our passive paradigm that was not designed to reveal early and late

mismatch responses (see discussion in Lecaignard et al., 2015).

Nevertheless, evidence for hierarchical learning could be reported thanks to our contextual manipulation of predictability which was designed to test the descending influence of (second-order) context-related expectations onto (first-order) sensory processing. Increasing second-order statistic reliability has been theoretically demonstrated to increase the precision-weighted of sensory errors (Friston 2008; Kanai et al. 2015). The observed twofold effect of predictability in this study (an increase in the size of the temporal integration window, and changes in ascending and intrinsic connectivity) does conform to this prediction and supports biological hierarchical inference in the brain. Note that second-order learning was not addressed here and that, at the sensory level, we considered a single-level learning based on a simple Bernoulli distribution (other, slightly more complex transition probability models could be considered (e.g. see Meyniel et al. (2016)).

Predictability and Attention

The larger τ in the predictable condition (more information is integrated over time) indicates that the brain can implicitly capture the larger informational value of predictable stimuli, compared to when stimulations are presented randomly. According to our learning model, this is associated with a larger PW. A computational process classically associated with the role of attention whose general purpose is to select incoming sensations in a contextual dependent manner in order to optimize perception and learning (Friston 2005; Parr & Friston 2018). Under this view, *attention* should be distinguished from *salience,* with the former being related to the gain control of sensory channels (through precision weighting) (Feldman & Friston 2010), while the latter refers to actions engaged by the brain to seek out-of-range sensory information. The view of attention as precision tuning has received great support from DCM findings

where larger self-inhibition (hence larger PW) proved induced by attention orienting in a Posner task (Brown & Friston 2013). Moreover, this view has also enabled addressing formally the opposite influence of attention (an increase) and predictability (a decrease) on ERP amplitudes (reflecting PWPEs). In Chennu et al. (2013) and in Auksztulewicz and Friston (2015), both factors could be manipulated orthogonally using different task instructions during perception of oddball-like sound sequences. These two studies revealed different modulations of mismatch responses at different latencies, and related attention to self-inhibitory mechanisms using DCM, respectively. Interestingly, Chennu and colleagues reported a reduced MMN when attention was explicitly engaged towards tone transitions compared to multi-tone patterns, which fits with the reduced MMN we observed in the predictable condition. Precisely, a more informed learning of the oddball rule (providing better predictions) could be at play either through an explicit attentional engagement or, in our case, through the implicit learning of the contextual information. This highlights the fact that the contextual modulation of learning we observed reflects an implicit optimization of perceptual learning through automated precision tuning. Therefore, likewise but without the voluntary orientation of attention, predictability here acts as an implicit attentional process, enhancing the efficiency of sensory processing.

IFG and perceptual learning.

The present IFG-related findings obtained at both levels of modeling appear indicative of the IFG being a key contributor in regularity learning. Precisely, the IFG exhibited bilaterally perceptual learning at the MMN latency and was also found to receive a direct thalamic input. Similar twofold DCM inputs in low (HG) and high (IFG) levels have already been reported recently in oddball studies by Phillips et al. (2015; 2016). These authors speculated that it may convey bottom-up sensory signals and top-down

predictions from higher levels, respectively. Here, IFG was also found to receive a feedback connection from the frontal SF level (see the network characterization findings in *DCM-oddball* analysis). This result led us to favor a true thalamic origin of the IFG input. This is also in agreement with recent research on thalamo-cortical interactions (Briggs & Usrey 2008; Sherman 2016). More specifically, cortico-thalamo-cortical loops have been proposed to control thalamic responsiveness and the synchronization of spatially distinct cortical areas (such as HG and the IFG in the present case) under a higher-level cortical influence. At the functional level, the thalamus and the IFG could be involved in computing the precision of sensory-level predictions (Kanai et al. 2015) and its backward influence onto lower-level predictions (Auksztulewicz et al. 2017), respectively. Our findings regarding the IFG are remarkably consistent with these learning-related assumptions. Note that a recent report of the implication of the right inferior / middle frontal gyrus in hierarchical learning in an active probability learning task also contribute to shed light on the role of this region in the Bayesian perspective of brain functions (Meyniel & Dehaene 2017). More generally, all these promising findings encourage to further investigating the cortical (and subcortical) connectivity at play during auditory processing and how Bayesian learning manifests at the different levels of the auditory hierarchy.

Conclusion.

An oddball paradigm was used to test regularity learning during auditory processing and predictability was further manipulated to disentangle PE and PW neural representations. Precisely, this contextual effect was found to increase the extent of temporal integration of past information, which implies lower sensory prediction errors amplified by a larger precision weighting. Findings in this paper 1) demonstrate the conclusive power of modeling approaches combining neuronal and cognitive levels

and 2) emphasize the importance of accounting for the encoding of precision weighting when investigating perceptual learning and decision-making. Unfolding the mechanisms of precision tuning and encoding, especially at an implicit level, is a potentially critical step for clinical applications as alterations of these processes have been suggested to be at the core of several psychiatric disorders (Adams et al. 2013; Lawson et al. 2017). Applying such a simple oddball paradigm, only involving passive listening, could be of great value in that context.

MATERIAL AND METHODS

Experimental model and subject details

Twenty healthy volunteers (10 female, mean age 25±5 years, ranging from 18 to 35) participated in the study. All participants were free from neurological or psychiatric disorder, and reported normal hearing. All participants gave written informed consent and were paid for their participation. Ethical approval was obtained from the appropriate regional ethics committee on Human Research (CPP Sud-Est IV - 2010-A00301-38).

Experimental design

Predictable and unpredictable sound sequences embedding a typical frequency oddball rule (conditions PC and UC, respectively) were used in the present study. Participants were instructed to ignore the sounds and watch a silent movie of their choice with subtitles. Predictable sound sequences comprised 16 cycles that were each made of a repeating 42-tone pattern following the deterministic incrementing rule depicted in Figure 1. Unpredictable sequences corresponded to pseudo-random

oddball sequences typically used in oddball paradigms, with specific controls for the number of standards in between two deviants. Despite their differing statistical structure, both sequence types had the same deviant probability ($p = 0.17$) and the same distribution of deviants among standards (there were exactly the same number of chunks of repeating standards before a deviant in both conditions, with chunk size varying from 2 to 8 standards). Further details about stimuli and sequences can be found in Lecaignard et al. (2015). All stimuli were delivered using Presentation software (Neurobehavioral Systems, Albany, CA, USA).

Data acquisition and pre-processing

Simultaneous MEG and EEG recordings were carried out in a magnetically shielded room with a whole-head 275-channel gradiometer (CTF-275 by VSM Medtech Inc.) and the CTF-supplied EEG recording system (63 electrodes), respectively. Digitization of electrode locations (Fastrak, Polhemus, Colchester, VT, USA) and acquisition of individual T1-weighted magnetic resonance imaging images (MRIs, Magnetom Sonata 1.5 T, Siemens, Erlangen, Germany) were conducted for coregistration purposes in the distributed source reconstruction (see below).

Pre-processing of data included the following: rejection of data segments corrupted by head movements (larger than 15 mm relative to the average position over sessions) and SQUID jumps, power-line filtering (stop-band filters centered on 50 Hz, 100 Hz and 150 Hz with bandwidth of ± 2 Hz), independent component analysis (ICA) correction for ocular artifacts (EEGlab routines, http://sccn.ucsd.edu/eeglab/index.html), rejection of trial epochs (from -200 ms to 410 ms after stimulus onset) with signal amplitude range (over entire epoch) exceeding 2000 fT for MEG data and 150 μV for EEG data, and 2-45 Hz band-pass digital filtering

(bidirectional Butterworth, 4th order). Importantly, we only used time epochs that survived the procedures applied for artifact rejection for both modalities. Finally, trial epochs were imported in SPM (Wellcome Department of Imaging Neuroscience, http://www.fil.ion.ucl.ac.uk/spm) and were down-sampled (200 Hz) for data reduction and low-pass filtered (20 Hz low-pass digital filter, bidirectional Butterworth, 5th order).

Computational modeling (cognitive models)

In this section, we first present the perceptual learning model of family *fam*$_L$. We secondly provide the expression of the related trialwise PWPEs, followed by a description of how these errors vary with the size of the temporal integration window that we introduced. Details about MMN simulations are then provided. Finally, we describe the data for model inversion (single-trial cortical activity reconstructed using fused EEG-MEG source inversion), as well as the inversion procedure.

- Learning models

Learning models in *fam*$_L$ assume that the brain learns from each stimulus exposure the probability μ to have a deviant, in order to predict the next sound category u (with $u = 1$ in the case of a deviant and $u = 0$ in the case of a standard). We define $u \sim \mathcal{B}ern\ (\mu)$ with $\mathcal{B}ern$ the Bernoulli distribution, and $\mu \sim \mathcal{B}eta\ (\alpha, \beta)$ with α and β the parameters of the distribution $\mathcal{B}eta$, corresponding in the current case to deviant and standard counts, respectively. At trial k, we have:

$$\begin{cases} p(u_k|\mu_{k-1}) = {\mu_{k-1}}^{u_k}(1-\mu_{k-1})^{1-u_k} \\ p(\mu_k|u_k) = \dfrac{p(u_k|\mu_{k-1})p(\mu_{k-1})}{p(u_k)} \end{cases} \quad \text{(S2)}$$

The posterior distribution of μ is in the form of a $\mathcal{B}eta$ distribution ($\mathcal{B}eta$ distribution is

conjugate to the $\mathcal{B}ern$ distribution), leading to the following updated expression of μ at trial k:

$$\mu_k = \frac{\Gamma(\alpha_k)\Gamma(\beta_k)}{\Gamma(\alpha_k + \beta_k)} \tag{S3}$$

With Γ the Gamma Euler function, and α and β update equations:

$$\begin{cases} \alpha_k = u_k + \alpha_{k-1} \\ \beta_k = 1 - u_k + \beta_{k-1} \end{cases} \tag{S4}$$

PWPE in *fam*$_L$ is the Kullback-Leibler (KL) divergence between the prior and the posterior $\mathcal{B}eta$ distributions of μ, also referred to as a Bayesian Surprise (BS) in the MMN study by Ostwald et al. (2012). At trial k, it expresses as:

$$\begin{aligned} & BS(u_k) \\ & = \log\left(\frac{\Gamma(\alpha_{k-1} + \beta_{k-1})}{\Gamma(\alpha_k + \beta_k)}\right) + \log\left(\frac{\Gamma(\alpha_k)}{\Gamma(\alpha_{k-1})}\right) \\ & + \log\left(\frac{\Gamma(\beta_k)}{\Gamma(\beta_{k-1})}\right) + (\alpha_{k-1} - \alpha_k)[\psi(\alpha_{k-1}) - \psi(\alpha_{k-1} + \beta_{k-1})] + (\beta_{k-1} \\ & - \beta_k)\,[\psi(\beta_{k-1}) - \psi(\alpha_{k-1} + \beta_{k-1})\,] \end{aligned} \tag{S5}$$

With ψ the digamma Euler function.

- Variation of BS with the size of temporal integration window

The size of temporal integration window was parameterized by τ which enters standard and deviant count updates as follows:

$$\begin{cases} \alpha_k = u + e^{-\frac{1}{\tau}}\alpha_{k-1} \\ \beta_k = (1-u) + e^{-\frac{1}{\tau}}\beta_{k-1} \end{cases} \tag{S6}$$

From equation Eq. S5 we see that the larger the τ, the larger the weight applying to past observations, leading to a more informed learning (an illustration can be found in Figure 4.C). This leads to lower BS as can be seen in Figure S1.

- Simulations of group-level MMN

For each subject, we used the VBA routine to simulate individual trial-by-trial PWPE trajectories induced by UC and PC sequences for τ in $\{6,8,10,12,14,16,18,20,25,30,40,50,75,100\}$. We then selected PWPE values obtained for each standard preceding a deviant and for each deviant (in keeping with sensor-level trial rejection). Using exactly the same procedure that had been used to compute the event-related difference response at the sensor level, we could simulate the group-average MMN amplitude in conditions UC and PC. Results are provided in Figure 4.A.

- Trial-wise reconstructed cortical data

Model predictions were confronted to trial-by-trial timeseries collected with EEG and MEG. More precisely, single-trial data was obtained in a preparatory step involving the distributed source reconstruction of EEG-MEG data. Advanced methods were employed for source inversion with namely realistic forward models for both modalities (Gramfort et al. 2010), Bayesian framework enabling Multiple Sparse Priors (Mattout et al. 2006; Friston et al. 2008), EEG-MEG fusion (Henson et al. 2009) and group-level inference (Litvak & Friston 2008). Source inversions were all performed with the SPM software (SPM8 release). First, in a preliminary study (Lecaignard et al., in prep), six cortical clusters could be identified from the inversions of early and late (MMN) mismatch peaks. These sources were located precisely in the left and right Heschl's Gyrus (lHG and rHG, respectively), Planum Polare (lPP and rPP) and Inferior Frontal Gyrus (lIFG and rIFG) (see Figure 2.A). Critically, they subsequently served as spatial

priors to constrain the distributed inversion of entire single-trial epochs (from -200 ms to +410 ms, 674 trials per condition). Within each cluster and for each trial, reconstructed cluster-node activities were averaged to derive a cluster-level and single-trial trace being informed by both EEG and MEG data. As a control, we computed the grand-average difference response (using exactly the same procedure as for the sensor-level analysis and in keeping with individual bad trial rejection) for each cluster. Resulting group-level traces are presented in Figure S2 and exhibit the predictability effect (UC vs. PC) at the latency of the MMN (around 175 ms) at every cluster.

- Model inversion

Model inversions were performed with the VBA toolbox (Daunizeau et al. 2014) at each time sample of trial-wise cortical timeseries. To reduce the number of inversions, we restricted the time interval to -50 ms to 350 ms and considered one over two samples, leading to 41 samples. Consequently, in the first analysis (*PL-oddball*), given the 7 models, 6 clusters and 41 samples, 1722 inversions were carried out per subject (N=20). Individual UC and PC data (4 sessions) were processed all at once (multi-session inversions), and bad trials (with regard to sensor-level artifact rejection) were treated so that corresponding signals would not corrupt parameter optimization (note that related stimuli still entered learning dynamics because they were observed by the brain). Regarding model specifications in VBA, we assumed a deterministic evolution model (no state noise was introduced in the evolution model) and used VBA defaults priors for the measurement noise. Initial states values (involved in fam_{NoL} and fam_{L}) and prior distributions of learning quantities across models are given in Table S1. Model inversions in *PL-oddball* analysis provided in particular posterior estimates of observation parameters $\theta_1, \varepsilon_1, \varepsilon_2$ defined in Equation Eq.1 and the free energy at

convergence approximating posterior model evidence. In the subsequent *PL-predictability* analysis, model space was restricted to one model, corresponding to fam_L form augmented with an evolution parameter to be estimated (parameter τ). Temporal samples included in this analysis were those exhibiting fam_L as the winning family in the *PL-oddball* analysis (the number of samples consequently varied across sources). Condition UC and PC were treated separately and multi-session inversion was employed within each condition to fit the two reverse sessions at once. Overall, 66 inversions were conducted per subject (1 model, 2 conditions, 33 samples across the 6 clusters). Other model specifications were identical to those of *PL-oddball* analysis (described above).

Neural modeling (Dynamic Causal Modeling, DCM)

We start by describing the cortical sources considered throughout the different DCM analyses reported in the study. We then present two model spaces specifically designed to address the structure and the connectivity of the DCM for passive oddball sequence processing. Following subsections address EEG and MEG evoked responses entering model inversions, and model inversion specification. We finally describe the procedure developed to combine EEG and MEG DCMs that we propose.

- DCM sources

DCM architecture involved the sources of the early and late (MMN) components that we previously reconstructed using fused EEG-MEG inversion (Lecaignard et al., in prep), with the current standard (preceding a deviant) and deviant evoked responses. In addition to the 6 bilateral clusters identified over Heschl's Gyrus (HG), the Planum Polare (PP), and the Inferior Frontal Gyrus (IFG), we assumed a bilateral superior frontal contribution (SF) motivated by our sensor-level findings but also by previous

electrophysiological studies (Marco-Pallarés et al. 2005; Fulham et al. 2014). SF clusters resulted from the fused source reconstruction of the difference between UC and PC deviance responses over the significant time intervals reported in Lecaignard et al. (2015). Each of the 8 resulting clusters led to an equivalent current dipole (ECD) located at the averaged position of local maxima over the different time intervals. MNI coordinates are provided in Figure 3.A.

- The *Network-Structure* model space

This model space is dedicated to the *DCM-oddball* analysis, to test the architecture and the inputs of DCM. It is partitioned over five architecture families (*A1*, *A2*, *A3*, *A4* and *A5)* and two input families (*HG* and *HG-IFG*), depicted in Figure S3. Here, we assumed forward and backward trial-specific modulations (TSM), as already reported in several MMN-DCM studies (Garrido, Kilner, Stephan, et al. 2009). For TSM applying on the self-inhibition connection strength of the superficial pyramidal cells (subpopulation SP), also referred so as the intrinsic modulation, we integrated over the two possible hypotheses (presence or absence of modulation). DCM with CMC also includes extrinsic *modulatory* connections to enable the top-down indexation of subpopulation SP excitability on the output activity of higher-level feedbacking sources. Like for the intrinsic modulation, model space was made to integrate over the two alternative hypotheses (presence or absence of modulatory connection). The resulting model space thus embedded 36 models (Figure S3).

- The *Coupling-Change* model space

The *Coupling-Change* model space was dedicated to address precisely the effective (directed) connectivity at play during passive listening within the winning DCM network (architecture *A5* and double-input *HG-IFG*). It was first employed in the *DCM-oddball*

analysis with data in condition UC, and secondly in the *DCM-predictability* analysis with data in condition PC. We addressed in particular TSM applying on forward, backward and intrinsic connections. For forward and intrinsic TSMs, two model families were considered each, corresponding to TSM enabled at all or none connections. Regarding backward TSMs, their effect depend on the above-mentioned extrinsic modulatory connections, if any. Specifically, they apply to backward extrinsic connections in the absence of modulatory connections, and to intrinsic connections otherwise (Bastos et al. 2012). This led us to rigorously consider three model families corresponding to backward TSM i) disabled, ii) enabled without and iii) enabled with modulatory connections. A total of 14 models composed this model space.

- Data for model inversion: EEG and MEG evoked responses

EEG and MEG averaged responses evoked by standards just preceding a deviant and by deviants were considered for all DCM analyses presented in the current study. EEG responses were re-referenced to the averaged mastoid electrodes. Time interval of 0 ms to 220 ms after sound onset was used for model inversion. It was defined from sensor-level (EEG and MEG) statistical analysis on deviance responses to ensure it encompasses the MMN (and no later components). A Hanning window was applied to time-series to ensure that system's dynamics was set to zero before being excited. Data reduction achieved within the inversion procedure led to the selection of 8 (±2.7) and 13 spatial modes with EEG and MEG data, respectively (on average across subjects).

- Model inversion

DCM analyses were performed with SPM12 (Wellcome Department of Imaging

Neuroscience, http://www.fil.ion.ucl.ac.ik/spm). We used the canonical micro-circuit (CMC) neural mass model (Bastos et al., 2012; Brown & Friston, 2013) to exploit its relevance to test predictive coding predictions. We used default values of SPM12 as prior expectations and prior variance for each DCM parameter to be estimated. Each DCM inversion involved standard response as the initial state of the system and deviant response resulting from the experimental perturbation. Regarding DCM sources, we maintained dipole locations fixed (but not their orientation) to inform spatially DCM inversion with the group-level MEG-EEG information. For each modality (EEG, MEG), the forward model used for DCM inversion was an advanced realistic Boundary Element Model (BEM) computed with Openmeeg software (Gramfort et al. 2010).

- Fusion of EEG and MEG DCMs

To date, the fusion of EEG and MEG data within the generative model of DCM is not implemented. Our approach consisted in performing EEG and MEG DCM inversions separately (using exactly the same parameter priors), and in a second step, exploiting the Bayesian framework of DCM to derive posterior estimates informed by both modalities (referred to as p-MEEG). Precisely, the proposed procedure rests on the assumption of conditional independence of EEG and MEG data under the quasi-static approximation of Maxwell equations, which is largely admitted for signals below 1kHz (as is the case here). Denoting EEG and MEG data by y_{EEG} and y_{MEG}, respectively, we have:

$$p(y_{EEG}, y_{MEG}) = p(y_{EEG})p(y_{MEG}) \quad \text{(S7)}$$

And posterior model evidence of model m can be approximated using unimodal EEG and MEG model evidences:

$$p(y_{EEG}, y_{MEG}|m) = p(y_{EEG}|m)p(y_{MEG}|m) \quad \text{(S8)}$$

Consequently, $\mathcal{F}_{p-MEEG}$ the variational free energy approximation to p-MEEG model log-evidence could be obtained by:

$$\mathcal{F}_{p-MEEG}(m) \approx \mathcal{F}_{EEG}(m) + \mathcal{F}_{MEG}(m) \quad \text{(S9)}$$

With $\mathcal{F}_{EEG}$ and $\mathcal{F}_{MEG}$ the free energy values for EEG and MEG respectively. Besides, the posterior distribution of some DCM parameter θ under model m writes given:

$$p(\theta|y_{EEG}, y_{MEG}) = \frac{p(y_{EEG}, y_{MEG}|\theta)p(\theta)}{p(y_{EEG}, y_{MEG})} \quad \text{(S10)}$$

Which can be re-formulated as follows to reveal the posterior distributions of θ deriving from unimodal inversion of EEG and MEG data:

$$p(\theta|y_{EEG}, y_{MEG}) = \frac{p(\theta|y_{EEG})p(\theta|y_{MEG})}{p(\theta)} \quad \text{(S11)}$$

DCM approach assumes every parameter θ to have of the form of a Gaussian distribution. Hence prior distribution expresses as $q(\theta) \sim \mathcal{N}(\mu_o, \sigma_o)$. We also denote $q(\theta, y_{EEG}) \sim \mathcal{N}(\mu_e, \sigma_e)$, $q(\theta, y_{MEG}) \sim \mathcal{N}(\mu_m, \sigma_m)$ and $q(\theta, y_{EEG}, y_{MEG}) \sim \mathcal{N}(\mu_p, \sigma_p)$ the posterior distribution of θ given EEG data, MEG data and EEG-and-MEG data respectively. We have μ_{em} and σ_{em} the mean and variance of the distribution resulting from the multiplication of $q(\theta, y_{EEG})$ and $q(\theta, y_{MEG})$. From equation Eq.S10 and the analytical expressions of μ_{em} and σ_{em} (detailed in most statistic books), we derive:

$$\begin{cases} \sigma_p = \dfrac{\sigma_o \sigma_{em}}{\sigma_o - \sigma_{em}} \\ \mu_p = \mu_{em} + \dfrac{(\mu_{em} - \mu_o)\sigma_p}{\sigma_o} \end{cases} \tag{S12}$$

Statistical analysis

PL-oddball analysis. We ran a family-level comparison (Penny et al. 2010) with a RFX model at each cluster and each time sample. Each comparison provided the posterior exceedance probability of each family (fam_{null}, fam_{NoL}, fam_{L}) that quantifies in a statistically principled way how the related family is likely to outperform others.

PL-predictability analysis. We assumed a constant value of τ within the time interval used for model inversion (spanning the MMN). We therefore averaged τ estimates across samples for each cluster. Predictability effect could thus be analyzed by conducting a repeated-measures ANOVA on these posterior estimates with factors *Condition* (UC, PC), *Hemisphere* (Left, Right) and *Sources* (HG, PP, IFG).

DCM-oddball analysis. For the first analysis over the *Network-Structure* model space with data in condition UC, we quantitatively evaluated the architecture and the input families using family level inference (Penny et al., 2010) with a RFX model, based on the p-MEEG approximations of model evidence ($\mathcal{F}_{p-MEEG}$). Second analysis with UC data and the *Coupling-Change* model space involved family level inference (RFX model) over the forward, backward and intrinsic TSM factor. As each of these three family comparisons indicated standard-to-deviant changes, we subsequently examined TSM estimates to determine the direction of change (a value larger than one would indicate larger connection strength in deviants than in standards, and *vice-*

versa). We used Bayesian model averaging (BMA) to derive group-level posterior estimates averaged across model space (with model-evidence weighting) and across subjects. For each connection type, BMA estimates were averaged over the entire network. Note that in the p-MEEG modality, BMA estimates were computed with the p-MEEG posterior distributions of parameters obtained with equation Eq.S11.

DCM-predictability analysis. Before comparing UC and PC BMA posterior estimates, we first replicated in condition PC the statistical analysis over the *Coupling-Change* model space described just above in condition UC. Next, to test the predictability effect *per se*, we focused on six DCM parameters because of their implication in the three model-driven predictions that we tested (presented in the main manuscript). Namely, they pertain to the forward, backward and intrinsic TSM, two forward connection strengths (in CMC, there are two types of forward connection both originating in subpopulation SP and targeting either the spiny stellate (SS) cells or subpopulation DP) and the strength of the self-inhibition connection within subpopulation SP. For each of these parameters, we conducted a repeated-measures ANOVA over individual BMA estimates with factors *Condition* (UC, PC), *Hemisphere* (Left, Right) and *Levels* (HG-PP, PP-IFG, IFG-SF); in the case of self-inhibition parameter, this latter factor was replaced by factor *Sources* (HG, PP, IFG, SF). For the forward connection strength parameter, we also included an additional *ConnectionType* factor (target_SS, target_DP). Reported averaged values of forward connection strength across the network account for both forward connection types.

Throughout the paper, ANOVA analyses were performed using R software (The R Foundation, https://www.r-project.org/).

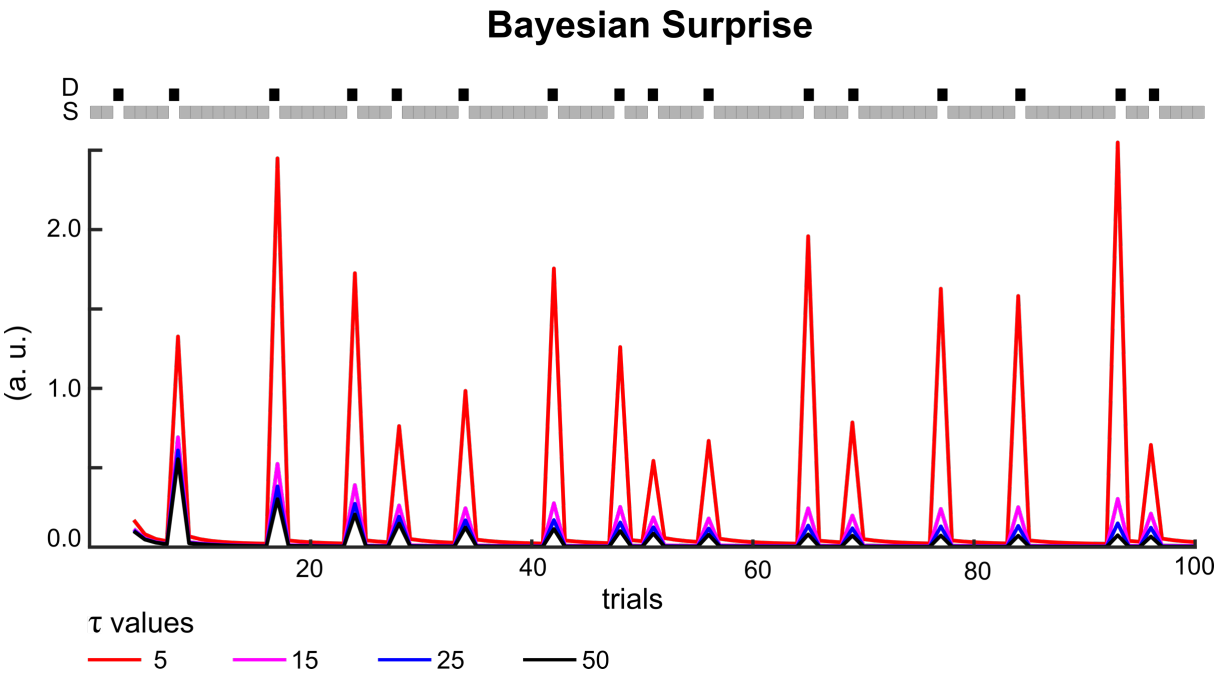

Figure S1. Bayesian surprise as a function of τ. (*arbitrary units, a. u.) Illustration of different BS trajectories obtained with varying τ, induced with the first 100 stimuli of a typical UC oddball sequence. Two comments should be made: 1) BS decreases as τ increases and 2) whatever τ, BS is larger for deviants (D, black rectangles) than for standards (S, grey rectangle) . Related to Figure 2.*

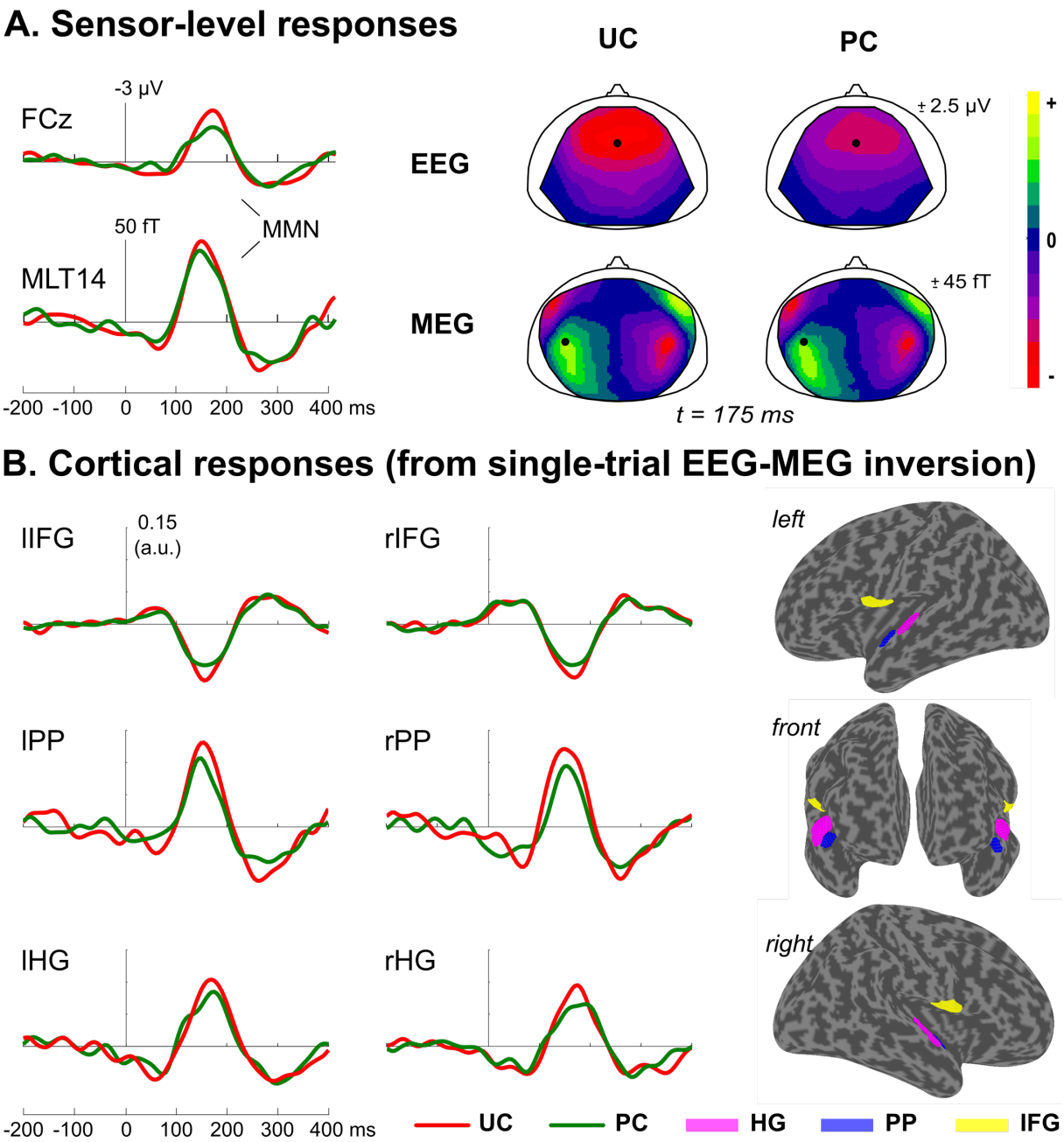

Figure S2. Mismatch responses for model inversion. *A) Group-average difference responses measured with EEG (upper row, 63 electrodes, averaged-mastoid reference) and MEG (lower row, 275 gradiometers). A typical MMN is indicated around 175 ms at electrode FCz and gradiometer MLT14 (left), in both contexts (UC: red, PC: green). Right, corresponding group-level scalp maps at 175 ms. Black dots refer to FCz (EEG maps, upper row) and MLT14 (MEG maps, lower row). B) Reconstructed mismatch response in condition UC and PC (red, green, resp.) obtained at the group-*

level from the individual EEG-MEG source reconstruction of single-trial traces, at six clusters engaged in MMN processing. Cluster labeling is provided in the main text. These bilateral clusters are described (right) on the inflated template cortex (SPM) under left, front and right views, using the bottom-right color code. Related to Figure 2 and Figure 4.

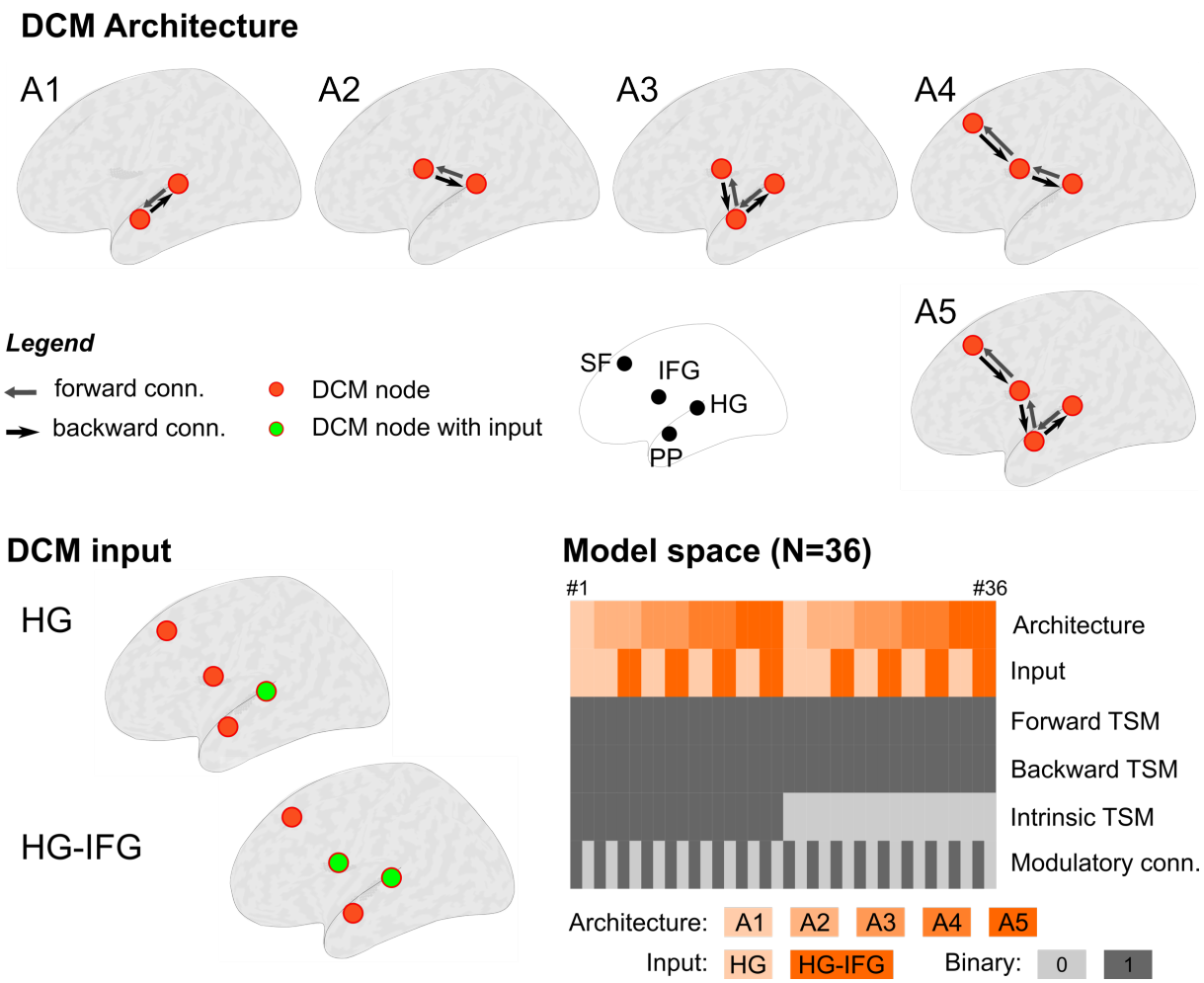

Figure S3. Network-Structure model space. *Upper row, schematic view of the five model families designed to test DCM architecture in deviance processing. Bottom left, the two model families of DCM thalamic input, HG and HG-IFG. Color codes of extrinsic connections (conn.) and DCM source (or node) are provided in the legend. Source labels HG, PP, IFG and SF: Heschl's Gyrus, Planum Polare, Inferior Frontal Gyrus and Superior Frontal. Bottom Right Table: DCM specifications for each of the 36 models*

(in columns). TSM: Trial-Specific Modulations accounting for standard-to-deviant changes in connectivity. Frontal, Backward and Intrinsic TSM, as well as Modulatory connections correspond to binary options (enabled = 1, disabled= 0) applying to the entire network. Related to main-text Figure 3.

		Initial state values		**Ev. Par**		**Obs. Par**	
family	*model*						
fam_{null}	null	-	-	-	-	θ_2	$(0, 100)$
fam_{noL}	CD	previous observation	$(0, 0)$	-	-	θ_1	$(0, 100)$
						θ_2	$(0, 100)$
	LinCD	previous observation	$(0, 0)$	-	-	θ_1	$(0, 100)$
		observed standards	$(0, 0)$	-	-	θ_2	$(0, 100)$
fam_L	fixed τ	standard count	$(1, 0)$	-	-	θ_1	$(0, 100)$
		deviant count	$(1, 0)$	-	-	θ_2	$(0, 100)$
	free τ	standard count	$(1, 0)$	τ	$(\log(10), 10)$	θ_1	$(0, 100)$
		deviant count	$(1, 0)$			θ_2	$(0, 100)$

Table S1. Sufficient statistics (mean, variance) for initial conditions and model parameters. *Ev.Par : Evolution parameters; Obs. Par: Observation parameters. For family fam_{noL}, previous observation set to 0 correspond to a standard sound. θ_1 and θ_2 refers to Equation Eq.1. Initial states values were not estimated during model inversion (variance is set to 0). Related to Figure 2 and Figure 4.*

Acknowledgments

We thank Sébastien Daligault and Pascal Calvat for programming support in interfacing with the CC IN2P3. We acknowledge CC-IN2P3 for providing computing resources and services needed for this work. We thank Emmanuel Maby and Gaëtan Sanchez for helpful discussions. This work was supported by a grant from the Agence Nationale de la Recherche of the French Ministry of Research ANR-11-BSH2-001-01 to AC and FL and a grant from the Fondation pour la Recherche Médicale (FRM) to OB and JM. This work was conducted in the framework of the LabEx CeLyA ("Centre Lyonnais d'Acoustique", ANR-10-LABX-0060) and of the LabEx Cortex ("Construction, Function and Cognitive Function and Rehabilitation of the Cortex", ANR-10-LABX-0042) of Université de Lyon, within the program "Investissements d'avenir" (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR).

Author Contributions

Conceptualization, AC, FL and JM; Methodology, FL and JM; Software, FL and JM; Investigation, AC and FL; Writing, AC, FL and JM; Funding Acquisition, OB, AC and JM; Supervision, OB, AC and JM.

Declaration of Interest

The authors declare that they have no conflict of interests.

References

Adams, R.A. et al., 2013. The computational anatomy of psychosis. *Frontiers in psychiatry*, 4, p.47.

Auksztulewicz, R. & Friston, K., 2015. Attentional Enhancement of Auditory Mismatch Responses: a DCM/MEG Study. *Cerebral Cortex*.

Auksztulewicz, R. & Friston, K., 2016. Repetition suppression and its contextual determinants in predictive coding. *CORTEX*.

Auksztulewicz, R. et al., 2017. The Cumulative Effects of Predictability on Synaptic Gain in the Auditory Processing Stream. *Journal of Neuroscience*, 37(28), pp.6751–6760.

Bastos, A.M. et al., 2012. Canonical microcircuits for predictive coding. *Neuron*, 76(4), pp.695–711.

Behrens, T.E.J. et al., 2007. Learning the value of information in an uncertain world. *Nature Neuroscience*, 10(9), pp.1214–1221.

Briggs, F. & Usrey, W.M., 2008. Emerging views of corticothalamic function. *Current opinion in neurobiology*, 18(4), pp.403–407.

Brown, H.R. & Friston, K., 2012. Dynamic causal modelling of precision and synaptic gain in visual perception - an EEG study. *NeuroImage*, 63(1), pp.223–231.

Brown, H.R. & Friston, K., 2013. The functional anatomy of attention: a DCM study. *Frontiers in Human Neuroscience*, 7, p.784.

Chennu, S. et al., 2013. Expectation and attention in hierarchical auditory prediction. *Journal of Neuroscience*, 33(27), pp.11194–11205.

Daunizeau, J. et al., 2010. Observing the observer (I): meta-bayesian models of learning and decision-making. *PLoS ONE*, 5(12), p.e15554.

Daunizeau, J., Adam, V. & Rigoux, L., 2014. VBA: a probabilistic treatment of nonlinear models for neurobiological and behavioural data. *PLoS Computational Biology*, 10(1), p.e1003441.

Dayan, P. et al., 1995. The Helmholtz machine. *Neural computation*, 7(5), pp.889–904.

Dehaene, S. et al., 2015. The Neural Representation of Sequences: From Transition Probabilities to Algebraic Patterns and Linguistic Trees. *Neuron*, 88(1), pp.2–19.

Deouell, L.Y., 2007. The Frontal Generator of the Mismatch Negativity Revisited. *Journal of Psychophysiology*, 21(3-4), pp.188–203.

Dürschmid, S. et al., 2016. Hierarchy of prediction errors for auditory events in human temporal and frontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, pp.201525030–6.

Escera, C., Leung, S. & Grimm, S., 2014. Deviance detection based on regularity encoding along the auditory hierarchy: electrophysiological evidence in humans. *Brain Topography*, 27(4), pp.527–538.

Feldman, H. & Friston, K., 2010. Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*, 4, p.215.

Friston, K., 2005. A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456), pp.815–836.

Friston, K., 2008. Hierarchical Models in the Brain O. Sporns, ed., 4(11), p.e1000211. Available at: http://dx.plos.org/10.1371/journal.pcbi.1000211.

Friston, K., 2012. The history of the future of the Bayesian brain. *NeuroImage*, 62(2), pp.1230–1233.

Friston, K. & Kiebel, S., 2009. Predictive coding under the free-energy principle. *Journal of Mathematical Psychology*, 364(1521), pp.1211–1221.

Friston, K. et al., 2008. Multiple sparse priors for the M/EEG inverse problem. *NeuroImage*, 39(3), pp.1104–1120.

Fulham, W.R. et al., 2014. Mismatch negativity in recent-onset and chronic schizophrenia: a current source density analysis. *PLoS ONE*, 9(6), p.e100221.

Garrido, M.I. et al., 2007. Evoked brain responses are generated by feedback loops. *Proceedings of the National Academy of Sciences of the United States of America*, 104(52), pp.20961–20966.

Garrido, M.I., Kilner, J.M., Kiebel, S., et al., 2009. Dynamic causal modeling of the response to frequency deviants. *Journal of Neurophysiology*, 101(5), pp.2620–2631.

Garrido, M.I., Kilner, J.M., Stephan, K.E., et al., 2009. The mismatch negativity: a review of underlying mechanisms. *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, 120(3), pp.453–463.

Gramfort, A. et al., 2010. OpenMEEG: opensource software for quasistatic bioelectromagnetics. *Biomedical engineering online*, 9(1), p.45.

Heilbron, M. & Chait, M., 2017. Great expectations: Is there evidence for predictive coding in auditory cortex? *Neuroscience*, pp.1–20.

Henson, R., Mouchlianitis, E. & Friston, K., 2009. MEG and EEG data fusion: simultaneous localisation of face-evoked responses. *NeuroImage*, 47(2), pp.581–589.

Iglesias, S. et al., 2013. Hierarchical Prediction Errors in Midbrain and Basal Forebrain during Sensory Learning. *Neuron*, 80(2), pp.519–530.

Kanai, R. et al., 2015. Cerebral hierarchies: predictive processing, precision and the pulvinar. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1668), pp.20140169–20140169.

Kiebel, S. et al., 2009. Dynamic causal modeling for EEG and MEG. R. Salmelin & S. Baillet, eds. *Human Brain Mapping*, 30(6), pp.1866–1876.

Kok, P. & de Lange, F.P., 2015. Predictive Coding in Sensory Cortex. In E.-J. Wagenmakers & B. U. Forstmann, eds. *An Introduction to Model-Based Cognitive Neuroscience*. New-York, pp. 221–244.

Lawson, R.P., Mathys, C.D. & Rees, G., 2017. Adults with autism overestimate the volatility of the sensory environment. *Nature Neuroscience*, 15, pp.173–13.

Lecaignard, F. et al., 2015. Implicit learning of predictable sound sequences modulates human brain responses at different levels of the auditory hierarchy. *Frontiers in Human Neuroscience*, 9, p.505.

Lieder, F. et al., 2013. Modelling trial-by-trial changes in the mismatch negativity. *PLoS Computational Biology*, 9(2), pp.1–16.

Litvak, V. & Friston, K., 2008. Electromagnetic source reconstruction for group studies. *NeuroImage*, 42(4), pp.1490–1498.

Lopes da Silva, F., 2013. EEG and MEG: relevance to neuroscience. *Neuron*, 80(5), pp.1112–1128.

Marco-Pallarés, J., Grau, C. & Ruffini, G., 2005. Combined ICA-LORETA analysis of mismatch negativity. 25(2), pp.471–477.

Markov, N.T. & Kennedy, H., 2013. The importance of being hierarchical. *Current opinion in neurobiology*, 23(2), pp.187–194.

Marr, D., 1982. *Vision: A Computational Approach* W.H. Freeman and Company, San Francisco.

Mathys, C.D. et al., 2011. A bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience*, 5, p.39.

Mathys, C.D. et al., 2014. Uncertainty in perception and the Hierarchical Gaussian Filter. *Frontiers in Human Neuroscience*, 8(47), p.825.

Mattout, J. et al., 2006. MEG source localization under multiple constraints: an extended Bayesian framework. *NeuroImage*, 30(3), pp.753–767.

Mattout, J., Henson, R. & Friston, K., 2007. Canonical source reconstruction for MEG. *Computational Intelligence and Neuroscience*, p.67613.

Meyniel, F. & Dehaene, S., 2017. Brain networks for confidence weighting and hierarchical inference during probabilistic learning. *Proceedings of the National Academy of Sciences of the United States of America*, 114(19), pp.E3859–E3868.

Meyniel, F., Maheu, M. & Dehaene, S., 2016. Human Inferences about Sequences: A Minimal Transition Probability Model. *PLoS Computational Biology*, 12(12), p.e1005260.

Moran, R. et al., 2013. Free energy, precision and learning: the role of cholinergic neuromodulation. *Journal of Neuroscience*, 33(19), pp.8227–8236.

Näätänen, R., Jacobsen, T. & Winkler, I., 2005. Memory-based or afferent processes in mismatch negativity (MMN): a review of the evidence. *Psychophysiology*, 42(1), pp.25–32.

Ostwald, D. et al., 2012. Evidence for neural encoding of Bayesian surprise in human somatosensation. *NeuroImage*, 62(1), pp.177–188.

Parr, T. & Friston, K., 2018. Attention or salience? *Current opinion in psychology*, 29, pp.1–5.

Penny, W.D. et al., 2010. Comparing families of dynamic causal models. *PLoS Computational Biology*, 6(3), p.e1000709.

Phillips, H.N. et al., 2016. Convergent evidence for hierarchical prediction networks from human electrocorticography and magnetoencephalography. *CORTEX*, 82, pp.192–205.

Phillips, H.N. et al., 2015. Hierarchical Organization of Frontotemporal Networks for the Prediction of Stimuli across Multiple Dimensions. *Journal of Neuroscience*, 35(25), pp.9255–9264.

Pinotsis, D.A. et al., 2017. Linking canonical microcircuits and neuronal activity_ Dynamic causal modelling of laminar recordings. *NeuroImage*, 146(C), pp.355–366.

Recasens, M. et al., 2015. Repetition suppression and repetition enhancement underlie auditory memory-trace formation in the human brain: an MEG study. *NeuroImage*, 108, pp.75–86.

Schröger, E., Marzecová, A. & Sanmiguel, I., 2015. Attention and prediction in human audition: a lesson from cognitive psychophysiology. *The European journal of neuroscience*, 41(5), pp.641–664.

Sherman, S.M., 2016. Thalamus plays a central role in ongoing cortical functioning. *Nature Neuroscience*, 19(4), pp.533–541.

Tenenbaum, J.B. et al., 2011. How to grow a mind: statistics, structure, and abstraction. *Science (New York, NY)*, 331(6022), pp.1279–1285.

Vossel, S. et al., 2014. Spatial attention, precision, and Bayesian inference: a study of saccadic response speed. *Cerebral cortex (New York, NY : 1991)*, 24(6), pp.1436–1450.