Predictive coding in auditory processing : insights from advanced modeling of EEG and MEG mismatch responses

Françoise Lecaignard

N° d'ordre NNT : 2016LYSE1160

THÈSE DE DOCTORAT DE L'UNIVERSITÉ DE LYON

opérée au sein de

l'Université Claude Bernard Lyon 1

École Doctorale ED476

École Doctorale Neurosciences et Cognition

Spécialité de doctorat : Neurosciences

Soutenue publiquement le 28/09/2016, par :

Françoise Lecaignard

Predictive coding in auditory processing: insights from advanced modeling of EEG and MEG mismatch responses

---

Devant le jury composé de :

Luauté Jacques, PU-PH, Université Lyon 1	Président
Kiebel Stefan, Prof., Université de Dresde	Rapporteur
Escera Carles, Prof. Université de Barcelone	Rapporteur
Bénar Christian, CR, Inserm	Examinateur
Bertrand Olivier, DR, Inserm	Directeur de thèse
Caclin Anne, CR, Inserm	Co-encadrant de thèse
Mattout Jérémie, CR, Inserm	Co-encadrant de thèse

UNIVERSITE CLAUDE BERNARD - LYON 1

Président de l'Université	**M. le Professeur Frédéric FLEURY**
Président du Conseil Académique	M. le Professeur Hamda BEN HADID
Vice-président du Conseil d'Administration	M. le Professeur Didier REVEL
Vice-président du Conseil Formation et Vie Universitaire	M. le Professeur Philippe CHEVALIER
Vice-président de la Commission Recherche	M. Fabrice VALLÉE
Directeur Général des Services	M. Alain HELLEU

COMPOSANTES SANTE

Faculté de Médecine Lyon Est – Claude Bernard	Directeur : M. le Professeur J. ETIENNE
Faculté de Médecine et de Maïeutique Lyon Sud – Charles Mérieux	Directeur : Mme la Professeure C. BURILLON
Faculté d'Odontologie	Directeur : M. le Professeur D. BOURGEOIS
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : Mme la Professeure C. VINCIGUERRA
Institut des Sciences et Techniques de la Réadaptation	Directeur : M. le Professeur Y. MATILLON
Département de formation et Centre de Recherche en Biologie Humaine	Directeur : Mme la Professeure A-M. SCHOTT

COMPOSANTES ET DEPARTEMENTS DE SCIENCES ET TECHNOLOGIE

Faculté des Sciences et Technologies	Directeur : M. F. DE MARCHI
Département Biologie	Directeur : M. le Professeur F. THEVENARD
Département Chimie Biochimie	Directeur : Mme C. FELIX
Département GEP	Directeur : M. Hassan HAMMOURI
Département Informatique	Directeur : M. le Professeur S. AKKOUCHE
Département Mathématiques	Directeur : M. le Professeur G. TOMANOV
Département Mécanique	Directeur : M. le Professeur H. BEN HADID
Département Physique	Directeur : M. le Professeur J-C PLENET
UFR Sciences et Techniques des Activités Physiques et Sportives	Directeur : M. Y.VANPOULLE
Observatoire des Sciences de l'Univers de Lyon	Directeur : M. B. GUIDERDONI
Polytech Lyon	Directeur : M. le Professeur E.PERRIN
Ecole Supérieure de Chimie Physique Electronique	Directeur : M. G. PIGNAULT
Institut Universitaire de Technologie de Lyon 1	Directeur : M. le Professeur C. VITON
Ecole Supérieure du Professorat et de l'Education	Directeur : M. le Professeur A. MOUGNIOTTE
Institut de Science Financière et d'Assurances	Directeur : M. N. LEBOISNE

Remerciements

J'adresse tout d'abord mes remerciements les plus sincères à Anne Caclin et Jérémie Mattout pour avoir avoir accepté de diriger cette thèse. J'ai énormément appris à leurs côtés, grâce à leur qualités pédagogiques et scientifiques. Ils ont chacun bien sûr leur domaine d'expertise mais également une telle ouverture d'esprit que les interactions furent toujours particulièrement précieuses. Un immense merci. J'adresse toute ma gratitude à Olivier Bertrand pour avoir (aussi) accepté de diriger cette thèse. Depuis l'époque dite *du Cours Albert Thomas*, l'occasion m'est donnée ici de le remercier chaleureusement pour ses conseils avisés et son soutien sans faille dans mes projets scientifiques et professionnels. Je remercie très sincèrement Gérard Gimenez pour avoir écouté mes motivations professionnelles pour ce projet de thèse et pour m'avoir accordé en conséquent les conditions de travail nécessaires à son bon déroulement. J'associe bien sûr Claude Delpuech à ces remerciements.

Je remercie très chaleureusement tous les membres de l'équipe Dycog du CRNL, pour certains compagnons de longue date, auprès desquels il est remarquablement agréable scientifiquement et humainement de travailler. Je remercie tout particulièrement Manu, Romain, Aurélie, Dominique, Perrine, Patrick et Martine, sans oublier Marie-Hélène. Cette thèse s'est aussi déroulée avec le soutien du Cermep, et je remercie en particulier Jamila Lagha et Christian Pierre. Je remercie bien évidemment Sébastien Daligault pour son rôle précieux dans l'utilisation des ressources du centre de calcul de l'IN2P3. Je remercie sincèrement Pascal Calvat et Yonni Cardenas pour m'avoir donné accès à ces ressources qui ont été cruciales pour ce projet.

Merci à toute l'équipe du RTT, dynamique dans sa composition à travers les années mais aussi et surtout dans son état d'esprit. Je remercie Marie Gomot pour avoir elle aussi apporté sa contribution à ma formation « MMN ». Je remercie également Andreea Diaconescu, Lilian Weber et Christoph Mathys pour des discussions fructueuses lors de l'élaboration de nos modèles d'apprentissages.

J'ai une pensée toute particulière pour ma si chère famille.

Enfin, je remercie infiniment Greg pour m'avoir toujours comprise et encouragée tout au long de ce projet. Je n'oublie pas de remercier nos filles qui ont su s'adapter aux conditions logistiques (parfois hasardeuses) de cette *grosse* thèse.

Résumé

Cette thèse porte sur le codage prédictif comme principe général pour la perception et vise à en étayer les mécanismes computationnels et neurophysiologiques dans la modalité auditive. Ce codage repose sur des erreurs de prédictions pondérées par leur précision (quantifiant leur plausibilité) se propageant au sein de la hiérarchie corticale, et se reflétant dans des réponses neurophysiologiques au changement (ou déviance) telles que la Négativité de discordance (mismatch negativity, MMN). Dans ce cadre théorique, nous avons formulé des hypothèses précises quant aux mécanismes génératifs de ces réponses, à la fois d'un point de vue computationnel et neurophysiologique. Nous avons pu les tester finement grâce à une manipulation expérimentale de la prédictibilité du changement auditif d'une part, et à l'utilisation d'enregistrements électrophysiologiques (EEG, MEG) simultanés, d'autre part.

Une modulation des réponses à la déviance par la prédictibilité a été observée, permettant d'établir un lien avec les erreurs de prédictions. Cet effet démontre un apprentissage implicite des régularités acoustiques, dont l'influence sur le traitement auditif a pu être caractérisée par notre approche de modélisation. Du point de vue computationnel, un apprentissage a été mis en évidence au cours de ce traitement auditif, reposant sur une fenêtre d'intégration temporelle des informations dont la taille augmente avec la prédictibilité des déviants. Du point de vue neurophysiologique, cet effet est associé à des modifications de la connectivité effective (extrinsèque et intrinsèque) induites par une modulation des gains synaptiques, comme le montre une analyse par modèles causaux dynamiques.

Ces résultats soulignent l'importance d'une approche multimodale, combinant ici EEG et MEG. Ils montrent aussi l'importance des modèles, à la fois computationnels et neurophysiologiques qui concourent à révéler la hiérarchie corticale auditive et éclairent les processus d'apprentissage perceptif. En accord avec le codage prédictif, nous avons notamment identifié la modulation contextuelle de l'erreur de prédiction ainsi que de sa précision.

Mots clés:
Cerveau bayésien, erreurs de prédictions, précision, inférence perceptive, électroencéphalographie, magnétoencéphalographie, potentiels évoqués, sources distribuées, reconstruction de sources, fusion multimodale, modèles dynamiques causaux.

Abstract

This thesis aims at characterizing predictive coding during auditory perception. Predictive coding rests on *precision-weighted prediction errors* elicited by unexpected sounds that propagate along a hierarchical organization in order to maintain the brain adapted to a varying sensory environment. This general principle is thought to subsume perceptual learning. However, its precise computational underpinnings and its implications in terms of neurophysiological mechanisms remain unclear. Using the mismatch negativity (MMN), a brain response to unexpected stimuli (deviants) that reflects such prediction errors in the brain, we tackled this twofold question. Precisely, we manipulated the predictability of deviants and applied computational learning models and dynamic causal models (DCM) to electrophysiological responses (EEG, MEG) measured simultaneously.

Deviance responses were found to be modulated by deviant predictability, a result further supporting their interpretation as prediction errors. This effect reflects the (high-level) implicit learning of sound sequence regularities which would in turn influences auditory processing in lower levels of the hierarchy. Computational modeling of trial-by-trial variations of electrophysiological evoked responses revealed the perceptual learning of sounds at play. Importantly, increased predictability yielded an increase in the size of the information temporal integration window. In addition, DCM analysis indicated predictability changes in the synaptic connectivity established by deviance processing. Precisely, we observed an increase in self-inhibition (coding for precision weighting) together with a decrease in forward connectivity (coding for prediction error) with higher predictability. This is consistent with the computational findings.

These results confirm predictive coding predictions regarding both deviance processing and its modulation by predictability. They shed light on perceptual learning processes within the auditory hierarchy. They also emphasize the great power of multimodal approaches (here the combination of EEG and MEG), together with the essential contribution of advanced computational and neurobiological models in modern cognitive neuroscience.

Key words:
Bayesian brain, precision-weighted prediction errors, perceptual inference, electroencephalography, magnetoencephalography, evoked potentials, event-related responses, source reconstruction, distributed sources, fused inversion, dynamical causal model.

Preamble

Recent Bayesian theories of brain functions propose that the brain would be constantly anticipating (predicting) its interactions with the external world and that both expected and unexpected interactions - once experienced - would be exploited to stay optimally adapted to the ever-changing environment. These advanced theories imply a Bayesian treatment of information and are often referred to as the *Bayesian brain* hypothesis. Importantly, they aim at explaining brain functioning as a whole, from perception to action, as well as brain dysfunctions.

Regarding perception, the Bayesian brain requires the (perceptual) learning of sensory regularities in order to build up predictions from past sensations. The general principle of predictive coding has been proposed as a possible functional implementation for such an information processing framework and precise implications for its neurobiological underpinnings have been suggested. While several behavioral studies strongly support Bayesian computation in the brain, only few neuroimaging or electrophysiological studies attempted to relate these findings with the underlying neurophysiology.

This thesis work places itself in this broad framework in order to tackle both the computational and neurophysiological mechanisms underlying auditory processing and more specifically the well-known *Mismatch Negativity* (MMN).

The first part is dedicated to the theoretical background of this work. Chapter 1 provides the rationale for Bayesian Inference and chapter 2 is dedicated to the brief description of the Bayesian brain theory, with an overview of the main associated findings in the field of perception. Chapter 3 introduces the MMN and reviews the existing accounts for this component. Finally, chapter 4 gives an overview of most recent Bayesian models of the MMN.

The second part describes precisely the objectives of this work and presents the methodology used to attain them, with corresponding findings. In particular, chapter 5 describes the simultaneous EEG and MEG study that we designed to elicit a contextual modulation of the MMN, which we predicted under the predictive coding hypothesis. Sensor-level analysis provided positive evidence for this hypothesis. Chapter 6 is dedicated to the fine-grained reconstruction of the cortical sources of deviance responses using both modalities. Relying on these findings, alternative hypotheses were formalized regarding the generative mechanisms of the MMN, at both the physiological (chapter 7) and cognitive (chapter 8) levels. These competing models were tested against our EEG and MEG data.

Finally, the third and last part summarizes and discusses our findings.

Contents

Remerciements iii

Résumé v

Abstract vii

Preamble ix

I Theoretical Background 3

1 Elements of Bayesian inference 7

1.1 From Bayes's rule to Bayesian inference . . . 7

1.1.1 Bayes's rule . . . 7

1.1.2 A few general comments on the Bayesian framework . . . 7

1.1.3 Bayesian inference . . . 8

1.2 Models: definitions of useful concepts . . . 9

1.2.1 Generative models . . . 9

1.2.2 Hierarchical models . . . 10

1.2.3 Dynamic causal models . . . 12

1.3 Numerical methods for Bayesian inference . . . 13

1.3.1 Stochastic methods . . . 13

1.3.2 Deterministic methods . . . 14

1.4 Model comparison . . . 16

1.4.1 Trading model accuracy and model complexity . . . 16

1.4.2 Approximations of model evidence . . . 17

1.4.3 The Bayes Factor . . . 18

1.4.4 Bayesian Model Selection and Bayesian Model Averaging . . . 19

1.5 Summary . . . 20

2 The Bayesian brain 21

2.1 The Bayesian brain hypothesis . . . 21

2.1.1 Definition . . . 21

2.1.2 Implications for brain function . . . 22

2.1.3 Empirical evidence for Bayesian behavior . . . 23

2.1.4 Neural implementation: the Bayesian coding hypothesis . . . 24

2.2 The free energy principle . . . 24

2.2.1 Definition . . . 25

2.2.2 Hierarchical and dynamic causal models in the brain . . . 26
2.2.3 Perceptual Inference, Perceptual Learning and Active Inference . . . 28
2.2.4 The Bayesian brain and the free energy principle . . . 29
2.3 The predictive coding implementation for perception . . . 30
2.3.1 Definition . . . 30
2.3.2 Generalized predictive coding . . . 31
2.3.3 Empirical evidence . . . 33
2.4 Meta-Bayesian analysis . . . 36
2.4.1 Observing the observer . . . 36
2.4.2 Examples of response models . . . 37
2.4.3 The VBA toolbox . . . 39
2.5 Summary . . . 39

3 The Mismatch Negativity 41
3.1 Introduction to the MMN . . . 41
3.1.1 Presentation . . . 41
3.1.2 Brief overview of auditory evoked potentials . . . 42
3.1.3 A comment on refractoriness and neuronal adaptation . . . 44
3.2 Modulation of the MMN by experimental manipulations . . . 45
3.2.1 Key features . . . 45
3.2.2 The broad spectrum of deviance types . . . 46
3.3 Neurophysiological underpinnings of the MMN . . . 46
3.3.1 Sources of the MMN . . . 47
3.3.2 Characterization of cortical dynamics (premises) . . . 49
3.3.3 Auditory hierarchy for deviance processing . . . 50
3.4 What functional role for the MMN? . . . 52
3.4.1 The sensory memory account . . . 52
3.4.2 The adaptation model . . . 53
3.4.3 The predictive coding model . . . 54
3.5 Summary . . . 55

4 Advanced Bayesian modeling of mismatch responses 57
4.1 Expected physiological and functional dynamics under predictive coding . . . 57
4.2 Dynamic causal modeling (DCM) . . . 58
4.2.1 General presentation . . . 58
4.2.2 DCM for EEG and MEG evoked responses . . . 59
4.2.3 DCM of mismatch responses . . . 65
4.3 Computational learning models . . . 69
4.3.1 The MMN as a Bayesian surprise (Ostwald et al., 2012) . . . 69
4.3.2 Free energy principle models of the MMN (Lieder et al., 2013) . . . 74
4.4 Attempts of computationally-informed dynamic causal models . . . 75
4.5 Summary . . . 77

II Experimental work 79

5 Effect of deviant predictability on mismatch responses 83
5.1 Objectives . . . 83
5.2 EEG analysis - Article . . . 83
5.3 MEG analysis . . . 98
5.3.1 Material and methods . . . 98
5.3.2 Results . . . 98
5.3.3 Conclusion . . . 100
5.4 Attempts to identify electrophysiological markers of perceptual learning . . . 101
5.5 Conclusion . . . 103

6 Spatial characterization of the cortical network for auditory deviance processing 105
6.1 Objectives . . . 105
6.2 Article . . . 106

7 Neurophysiological modeling of deviance responses: insights from predictability manipulation 127
7.1 Introduction . . . 127
7.2 Material and methods for all DCM studies . . . 128
7.3 DCM structure for deviance processing . . . 132
7.3.1 Methods . . . 132
7.3.2 Results . . . 134
7.4 Neural connectivity for deviance processing . . . 135
7.4.1 Methods . . . 135
7.4.2 Results . . . 136
7.5 Predictability effect on deviance processing . . . 138
7.5.1 Methods . . . 138
7.5.2 Results . . . 139
7.6 Conclusion . . . 139
7.7 Annex: Fused EEG-MEG DCM attempts . . . 143
7.7.1 Methods . . . 143
7.7.2 Results with simulated data . . . 145
7.7.3 Results with real data . . . 146
7.7.4 Conclusion . . . 147

8 Computational single-trial analysis of auditory responses: evidence for Bayesian learning at the MMN latency 149
8.1 Introduction . . . 149
8.2 Material and methods . . . 150
8.2.1 Analysis 1: Modeling auditory mismatch responses . . . 151
8.2.2 Analysis 2: Assessing the effect of predictability . . . 154
8.3 Results . . . 155
8.3.1 Implicit perceptual treatment of the oddball sequence . . . 155
8.3.2 Predictability effect on the forgetting value τ_t . . . 156

8.4 Conclusion 158

III General Discussion 163

9 Discussion and perspectives 165

9.1 Summary of the main results 165
9.2 Implications for future research 166
9.2.1 The auditory hierarchy serving Bayesian inference 166
9.2.2 Precision-weighted prediction errors 167
9.3 Towards model-driven clinical applications of the MMN 168
9.4 Concluding remarks 170
9.4.1 Related works performed during this PhD 170
9.4.2 List of publications related to this PhD work 171
References 172

Part I

Theoretical Background

Estimation of unknowns from knowns defines well the principle of which is statistical inference, one of the key step of empirical science. In the field of cognitive neuroscience, *knowns* may refer to measures of reaction time in behavioral studies, or to measures of electric potential differences or magnetic fields in neuroimaging studies. These measures (the *knowns*) should contain information to estimate the unobservable properties of the brain (the *unknowns*) that directly or indirectly have influenced their generation. For instance, *unknowns* could refer to cognitive processes (attentional load, musical perception,...) or biophysical properties like in cortical source reconstruction. Importantly, statistical inference requires the hypothesis driving the study to be formalized in the form of alternative models that will entail various assumptions and inevitable simplifications. These lead to *uncertainty* in the information manipulated by the model that, combined with the variability of observed data, may affect the *plausibility* or the degree of confidence to place in the inferred estimates. Bayesian inference, as we will see in the following section, is a mathematical framework that is very much appropriate to account for both prior knowledge and sources of uncertainty.

Estimation of unknowns from knowns, this is also what the brain could do according to the Bayesian brain. Leading contributors in the field such as Karl Friston suggest that the brain would be equipped to process statistical inference using (or approximating) a Bayesian scheme. Considered by some as revolutionary, the Bayesian brain provides a general theory for brain functioning, unifying different psychological models of mental processes and bridging the gap between psychology and physiology. It naturally opens the way to a vast field of research in neuroscience and has progressively given rise to a growing amount of dedicated empirical and fundamental studies over the past decade or so. The work presented in this thesis builds on this recent line of research and attempts to shed light on the specific field of auditory perception.

Bayesian inference is thus at the core of this work. On the one hand the brain is approached as a Bayesian engine observing the world. On the other hand we, as scientist observing the brain at work, adopt a Bayesian methodology to formulate and test our hypothesis. In other words, we conform to the Bayesian brain hypothesis when it comes to modeling brain functions (perceptual learning in the auditory domain). Besides, we behave as Bayesian scientist in our experimental and methodological approach in order to fit and compare alternative models of our data. We thus start this introduction by a brief recall of the theoretical principles of Bayesian inference that will be used throughout this work (chapter 1), followed by a presentation of the Bayesian brain hypothesis (chapter 2). The last two parts pertain to the auditory MMN (chapter 3), which was central to this work to investigate the predictive coding account of auditory processing as reflected by evoked responses, and to the existing (and promising) Bayesian neurobiological and cognitive models of this brain component (chapter 4).

Chapter 1

Elements of Bayesian inference

1.1 From Bayes's rule to Bayesian inference

1.1.1 Bayes's rule

Bayes's rule (or Bayes's Theorem) is the cornerstone of Bayesian inference. It simply derives from the expression of the *joint* probability $p(A, B)$ (or $p(A \cap B)$) of concurrently observing two events A and B:

$$p(A, B) = p(A|B)p(B) \tag{1.1}$$

where $p(A|B)$ denotes the *conditional* probability of observing A, given B (or having observed B). $p(B)$ is called the *marginal* probability of B and represents the probability to have B regardless of A. Importantly, the joint probability $p(A, B)$ depends on A and B in a symmetric manner and Eq. (1.1) can be reformulated by marginalizing on A, leading to the following equality:

$$p(A|B)p(B) = p(B|A)p(A) \tag{1.2}$$

This leads to the following expression known as Bayes's rule:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} \tag{1.3}$$

This form reflects that reverse probabilities (*e.g.* $p(B|A)$) can be obtained from direct ones (*e.g.* $p(A|B)$) and is precisely of utmost importance in Bayesian inference when one wants to infer causes from consequences.

1.1.2 A few general comments on the Bayesian framework

Before describing the basics of Bayesian inference, two central concepts of the Bayesian framework must be introduced. First, probability p does not refer to the frequency of occurence of an event (*e.g. the number of tails in N tosses of a coin, N being infinite*) but to its degree of plausibility. Plausibility is quantified by a real number normalized between 0 and 1, where 0 means that the event can never be (or is *false*, according to a logical formalism) and 1 means it always happens (always *true*). Hence, in a Bayesian scheme, the probability distribution of a variable (*i.e.* the function that assigns a probability to each value of the set of possible values for the given variable)

formalizes our knowledge (or equivalently the *uncertainty* associated with) of this variable. For instance, let J be the activity of a cortical source at time t. Using a gaussian distribution, we can model both our knowledge (or belief) that this activity is (for instance) null and the confidence we place on this knowledge, by informing accordingly the mean and the variance of the distribution respectively.

The second key notion in a Bayesian framework is the fact that every available information about the phenomenon of interest has to be accounted for (with associated uncertainty represented in the probability distribution). These two concepts highlight the fact that in a Bayesian setting, it is all about knowledge and uncertainty.

1.1.3 Bayesian inference

From these introducing remarks, one can see that Bayesian inference, allowing inferring parameter estimates from observations, is nothing else than belief updating from new observations. It is also termed as *Bayesian learning*. Precisely, Bayesian inference rests on the central notions of *prior* and *posterior* knowledges. Every unknown is treated as a random variable whose knowledge evolves with the confrontations to observations. Before having observed the data, this knowledge is represented in the form of a *prior distribution*, and the integration of information contained in the data leads to an update formalized by a *posterior distribution*. Posterior distribution is by essence conditioned to the data and reflects learning.

Let y and θ denote the observations and the unknowns respectively (θ could also refer to an hypothesis or a model, as will be seen in §1.2). Bayes's rule (Eq. (1.3)) provides the following relation between the prior distribution $p(\theta)$ and the posterior distribution $p(\theta|y)$:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{1.4}$$

Under this form, $p(y|\theta)$ is referred to as the *likelihood* distribution and represents how likely it is to observe y given θ. $p(y)$ is called the *evidence* and corresponds to the marginal distribution of y:

$$p(y) = \int p(y|\theta)p(\theta)d\theta \tag{1.5}$$

$p(y)$ is thus a constant term independent from θ, whose value is obtained by integrating out (marginalizing) $p(y|\theta)$ over the domain of possible values for θ. Eq. (1.4) expresses how the knowledge and uncertainty about θ are updated using the rules of probability and given the observed data. Importantly, the respective uncertainty associated with the prior $p(\theta)$ and likelihood $p(y|\theta)$ will affect this update, as illustrated in Figure 1.1. The lack of knowledge about an information is formalized by the variance of its distribution. This gives rise for instance to the following: large value of the likelihood variance (or equivalently, low confidence or precision), reflecting unreliable observation, will force the posterior to stick to the prior (whatever the data), whereas low value (high precision) allows posterior to differ from prior, being mostly informed by the (reliable) data.

In practice, Bayesian inference starts with the specification of likelihood and prior distributions that together define a *model*. This critical step relies on a subjective decision that depends on

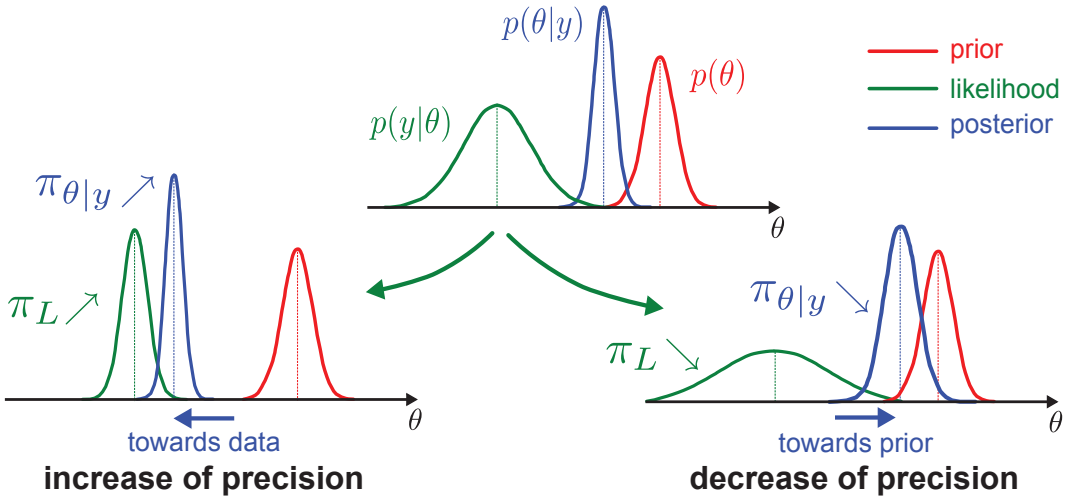

Figure 1.1 – Typical schematic view of Bayes's rule application to derive the posterior distribution of parameter θ. π_θ, π_L, and $\pi_{\theta|y}$ denote the precision (inverse variance) of the prior (red), likelihood (green) and posterior (blue) distributions. Lower: increasing the likelihood precision (right) results in the posterior distribution being mostly informed by the observed data, whereas a decrease (left) gives more weight to the prior. The same reasoning applies to changes in prior precision.

the degree of knowledge (before observing the data) about both the relation between data and parameters, and the plausibility of parameters. It may not be achievable easily (in particular when the phenomenon of interest is not well understood or implies a large number of data and/or parameters) and yet, it directly affects the estimation of posterior parameters. However, if one considers several models corresponding to competing hypothesis, parameter estimates given each model can be formally compared by means of their respective posterior distributions. This point is exactly where Bayesian inference derives its strength from: it allows both parameter estimation (in terms of its full distribution) and a principled model selection. These aspects will be covered in the following sections.

1.2 Models: definitions of useful concepts

The purpose of this section is to present different model families that were used in this thesis. These pertain to generative (first paragraph), hierarchical (second paragraph) and dynamical (last paragraph) models.

1.2.1 Generative models

Generative models are probabilistic models that formalize the generation of observable data y as a function of parameters θ (and possibly states x in the particular case of dynamic system models, as will be described in the following section). Put simply, a generative model m allows to predict how data should look like, given hidden causes. The *real* (and unknown) generative mechanism of y may not be necessarily a random process but its modeling - aiming at a better understanding of this mechanism - appeals to a probabilistic framework to account for uncertainty associated with both observed data and underlying parameters. Data predicted by m are consequently always wrong and the difference with observed data represents the amount of data that is not explained by the model, referred to as the error term and denoted ε. Possible origins of uncertainty entering

this residual term include model assumptions that may be inherited from a limited knowledge about the generative mechanism of the data. Furthermore, data are typically contaminated with noise originating from the measurement equipment. The Bayesian framework can flexibly account for these different uncertainties and Bayesian learning aims at removing part of the residual term associated with limited knowledge before seeing the data, hence providing better predictions.

Definition of a Bayesian generative model m requires the specification of a likelihood and a prior distributions:

$$p(y, \theta|m) = p(y|\theta, m)p(\theta|m) \tag{1.6}$$

In the case of the generative models used throughout this thesis to account for brain signals, the likelihood distribution $p(y|\theta, m)$ derives from:

- a deterministic part embodying the generative mechanism *per se*, called the *forward* or the *observation* model, denoted g
- a prior distribution characterizing the error term, denoted $p(\varepsilon|m)$.

The expression of y given θ is then:

$$y = g(\theta) + \varepsilon \tag{1.7}$$

The likelihood function $p(y|\theta, m)$ can be obtained by combining Eq. (1.7) and the prior distribution of ε, $p(\varepsilon|m)$. Two prior assumptions thus enter the likelihood function: the prior about the generative mechanism and the prior regarding the source of variability affecting the data. For instance, in the field of source reconstruction from electrophysiological data (EEG or MEG), g could indicate the lead-field operator that allows simulating sensor data from cortical activity given specific quantities (dipolar source parameters, anatomical and biophysical assumptions...) represented by θ. A gaussian distribution is usually assumed for ε, with zero mean and variance σ.

The specification of the prior distribution $p(\theta|m)$ should entail prior knowledge inherited from previous studies or expert statements for instance. In some cases where no explicit knowledge about θ is provided, *non-informative priors* can be used: typically, these are priors whose distribution has an nearly infinite variance. However, the use of such *flat* priors constitutes in itself a prior and may favor overfitting in parameter estimation. Another important category of priors is referred to as *empirical priors*. Empirical priors impose a hierarchical structure of the generative model m, which will be described in the following section.

1.2.2 Hierarchical models

Hierarchical models emerge when one considers priors about priors. As already mentioned, it may be difficult to fully specify the prior distribution of θ. Once the form of the parametric distribution has been chosen given the nature of the phenomenon to be modeled (*e.g.* a Bernoulli distribution to model a parameter taking two distinct possible values, a Gaussian distribution to model the activity of a cortical source), parameters of this distribution need to be defined (*e.g.* mean and variance) but in most cases, they remain unknown. Treating these parameters as random variables allows them to be estimated by Empirical Bayesian (EB) inference. Pursuing with the example of source reconstruction, with $y = g(\theta) + \varepsilon$, and θ being the amplitude of cortical sources, one

could model θ as a multivariate Gaussian variable with zero mean and covariance Σ_s. This would lead to a two-level hierarchical model of the form:

$$\begin{cases} y = g(\theta) + \varepsilon_n \quad \text{with} \quad p(\varepsilon_n|m) \sim \mathcal{N}(0\,;\Sigma_n) \\ \\ \theta = 0 + \varepsilon_s \quad \text{with} \quad p(\varepsilon_s|m) \sim \mathcal{N}(0\,;\Sigma_s) \end{cases} \tag{1.8}$$

where ε_n and ε_s reflect the uncertainty at each level of the hierarchy, namely the error at the sensor-level and source-level, respectively. The Gaussian probability distributions of the variables thus express as follows:

$$\begin{cases} p(y|\theta, m) \sim \mathcal{N}(g(\theta)\,;\Sigma_n) \\ \\ p(\theta|m) \sim \mathcal{N}(0\,;\Sigma_s) \end{cases} \tag{1.9}$$

In the general case, a hierarchical model is composed of N levels, each associated with a parameter $\theta_i, i \in [1, N]$ such that prior distribution of θ_i is a function of θ_{i+1} (Figure 1.2). Data y can be seen as the parameter of the lowest level with $y = \theta_0$ and prior distribution of θ_N, at the highest level is a function of θ_{N+1} that is supposed to be known (otherwise, a level $N + 1$ should be defined to account for its uncertainty). At each level i, hyperparameters θ_{i+1} represent the parameters of the distribution over θ_i. The full generative model now expresses as:

$$p(y, \theta_1, ..., \theta_N, \theta_{N+1}|m) = p(y|\theta_1, ..., \theta_N, \theta_{N+1}, m)p(\theta_1, ..., \theta_N, \theta_{N+1}|m) \tag{1.10}$$

Interestingly, the hierarchical nature of the model yields the following factorization and convenient rewriting:

$$p(y, \theta_1, ..., \theta_N|m) = p(y|\theta_1)p(\theta_1|\theta_2)...p(\theta_N|\theta_{N+1})p(\theta_{N+1}|m) \tag{1.11}$$

where $p(y|\theta_1)$ is the likelihood function. This formulation highlights the relations between parameters that will be exploited within the hierarchical Bayesian inference, when one wants to infer $\theta_1, ..., \theta_N, \theta_{N+1}$ from data y. Indeed, applying Eq. (1.4) to such a hierarchical model m yields:

$$p(\theta_1, ..., \theta_N, \theta_{N+1}|y, m) = \frac{p(y|\theta_1)p(\theta_1|\theta_2)...p(\theta_N|\theta_{N+1})p(\theta_{N+1}|m)}{p(y|m)} \tag{1.12}$$

$p(y|m)$ is a constant term and can be temporarily omitted so that simply:

$$p(\theta_1, ..., \theta_N, \theta_{N+1}|y, m) \propto p(y|\theta_1)p(\theta_1|\theta_2)...p(\theta_N|\theta_{N+1})p(\theta_{N+1}|m) \tag{1.13}$$

The causal relationships between these hierarchical levels can be visualized using *directed graphical models*. These models are composed of nodes which entail a random variable (or a vector of random variables), whose distribution is conditioned to the variable of the parent's node. The key feature is that each node thus receives constraints from its parents and itself provides constraints to its children. Such models are widely used in Bayesian statistics as they furnish an intuitive way to design the structure of a probabilistic model required for model inversion. They also conveniently and strikingly conform to often encountered structures in the environment as presumably in the brain.

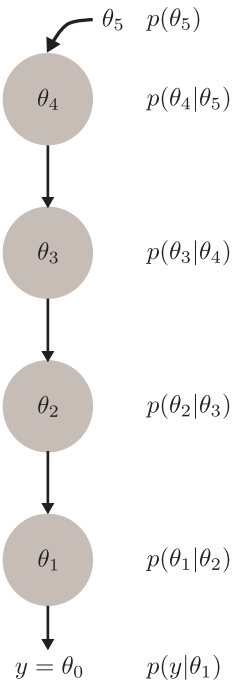

Figure 1.2 – Example of a hierarchical model composed of 4 levels.

1.2.3 Dynamic causal models

The final part of this section is dedicated to dynamic causal system modeling that allows describing the activity of a phenomenon of interest over time as a system perturbed by external (sensory or contextual) inputs. This system receives an input $u(t)$ at time t that influences its internal (hidden) states $x(t)$ leading to the observable outputs $y(t)$. *Causality* refers to the fact that present activity derives from past - but not future - activity. Dynamical system theory is a vast field of research and only the relevant aspects needed to understand the work presented in this thesis will be introduced here. In particular, we will refer to dynamical models resting on a *state-space* representation (also referred to as Hidden Markov Models), which is one of the widely used family of models in system theory.

Using a state-space representation, a dynamical model - presented here for continuous time - has the following structure:

$$\begin{cases} \dot{x}(t) = f(x(t), u(t), \theta) + \varepsilon_s \\ \\ y(t) = g(x(t), u(t), \psi) + \varepsilon_n \end{cases} \tag{1.14}$$

where $x(t)$ denotes the *hidden states* of the system, that describe its internal activity at time t (*hidden* refers to the fact that these states are not directly observable). The first equation describes the dynamical part of the model and is called the *evolution model*. The second equation refers to the static part of the model and is called the *observation model*.

- *The evolution model.* This model f rests on ordinary differential equations to calculate $\dot{x}(t)$, the update of state x at time t as a function of its value $x(t)$, the input $u(t)$ and parameters θ. The term ε_s is called the state noise and is set to 0 in the case of a deterministic evolution model. It can be shown that the trajectory of x has the following Markov property: $x(t)$ depends only on $x(t_0)$ (the initial state of x at time t_0), inputs u from t_0 to t and parameters θ (see 1.3.1 for the definition of the Markov property). Linear (non-linear) state-space

models refers to models with f being linear (non-linear). Using discrete time, the evolution equation is formulated as follows: $x_{t+1} = f(x_t, u_t, \theta) + \varepsilon_s$.

- *The observation model.* This model allows mapping internal states $x(t)$ to observations $y(t)$. Function g corresponds to the forward model described above, and gives $y(t)$ as a function of $x(t)$, input $u(t)$ and observation parameters ψ. The term ε_n refers to the error term already introduced.

1.3 Numerical methods for Bayesian inference

The previous section focused on describing generative models meant to predict observations, i.e. going forward, *from causes to consequences*, to derive predictions of observed data. Naturally, these models are made to be evaluated and used against empirical data, thus going *from consequences to causes*, namely to estimate the unknown parameters (and states) from observations. This is obtained using Bayes's rule and the ensuing posterior distribution over unknowns:

$$p(\theta|y, m) = \frac{p(y|\theta, m)p(\theta|m)}{p(y|m)} \tag{1.15}$$

which involves marginalization process, hence integral computation (see Eq. (1.5)). In some very simple (and rare) cases, the resolution of these integrals can be derived analytically leading to an exact solution. However, in the frequent case of complex models involving a large number of parameters and states and/or non-linear models - the brain for instance - it remains intractable and numerical methods approximating the parameters of the true Bayesian posterior distribution have to be employed. The two main categories of numerical approaches are presented below.

1.3.1 Stochastic methods

Stochastic methods (or sampling methods) allow approximating any kind of integration and in particular the posterior distribution $p(\theta|y)$, by collecting N samples from this distribution, $\{\theta_1, \theta_2, ..., \theta_N\}$[1]. As N tends towards infinity, this approximation becomes the exact distribution (under the Central Limit Theorem). In practice, sampling methods require drawing samples of θ from its set of possible values to derive a sample of $p(\theta|y)$ using the product $p(y|\theta)p(\theta)$.

Monte-Carlo methods rest on the specification of a *proposal* distribution q, from which samples θ are drawn. Each sample, denoted α will enter a decision criterion (involving $q(\alpha)$ and $p(\alpha|y)$), to reject or accept $p(\alpha|y)$ as a new sample for $p(\theta|y)$. If the decision criterion is too conservative, the procedure will tend to reject a lot of samples, hence increasing the total duration of the process.

For models dealing with high-dimensional θ, samples must be collected in an efficient manner (they should be picked preferentially in high-probability intervals) to avoid excessive time-consuming process. This can be addressed with Markov-Chain-Monte-Carlo (MCMC) methods. A distribution of random variables $\{X_1, X_2, ..., X_N\}$ is defined as a Markov-Chain if it verifies the following

[1] Subscript indices here no longer refer to hierarchical levels but to samples.

Markov property:

$$p(X_n = x_n|X_{n-1} = x_{n-1}, ..., X_1 = x_1, X_0 = x_0) = p(X_n = x_n|X_{n-1} = x_{n-1}) \tag{1.16}$$

This property states that the entire sequence $X_1, ..., X_N$ is obtained by simply knowing the initial value of X_0 and the transition probabilities $p_{ab} = p(X_k = a|X_{k-1} = b)$ with a, b being possible values of X_k.

MCMC methods consist in collecting a Markov chain of samples α from q defined as a transition probability. For sample k, $\alpha_k = q(\alpha_k|\theta_{k-1})$ and α_k will be accepted as the sample θ_k under some decision criterion. The two main MCMC methods are the Metropolis-Hastings and the Gibbs Sampler algorithms. MCMC methods have the advantage to possibly furnish the exact posterior distribution (whatever the distribution form) as N becomes large enough. However, in practice, the chain size (N) is not necessarily easy to define and results depend on the initial values of the chain (usually M chains of N samples are built up to avoid this issue). Finally, despite advances in computer performances, these powerful methods still remain computationally intensive in comparison with deterministic procedures introduced below.

1.3.2 Deterministic methods

Deterministic or *variational* methods rely on variational calculus, that is finding a function f that optimizes the value of a specific integral $I(f)$. Precisely, variational methods for Bayesian inference consist in finding $q(\theta)$, an approximate distribution to the posterior distribution $p(\theta|y)$ that maximizes a certain cost function $F(q)$, the *variational free energy*. Hence these methods transform the (intractable) problem of estimating $p(\theta|y)$ into (tractable) optimization problem over q.

Let $q(\theta)$ be the approximation to the posterior $p(\theta|y, m)$, under model m. To assess how well this approximation fits the targeted distribution, one could estimate the Kullback Leibler (KL) divergence between $q(\theta)$ and $p(\theta|y, m)$, $D_{KL}(q(\theta), p(\theta|y, m))$. This information-theoretic measure quantifies the dissimilarity between two probabilistic distributions (with this non-negative measure being such that the larger the more dissimilar q and p are). It writes:

$$D_{KL}(q(\theta), p(\theta|y, m)) = \int q(\theta) \ln \frac{q(\theta)}{p(\theta|y, m)} d\theta \tag{1.17}$$

Rearranging this equation gives:

$$D_{KL}(q(\theta), p(\theta|y, m)) = \int q(\theta) \ln q(\theta) d\theta - \int q(\theta) \ln p(\theta|y, m) d\theta \tag{1.18}$$

Applying Bayes's rule (Eq. (1.1)) to the term $p(\theta|y, m)$ gives:

$$D_{KL}(q(\theta), p(\theta|y, m)) = \int q(\theta) \ln q(\theta) d\theta - \int q(\theta) \ln p(\theta, y|m) d\theta + \int q(\theta) \ln p(y|m) d\theta \tag{1.19}$$

- The last term of this equation does not depend on θ and is equal to $\ln p(y|m)$, which corresponds to the *log-evidence* of the model (model evidence has been defined in Eq.

(1.5)). The term $-\ln p(y|m)$ is also defined as the *statistical surprise*.

- The first two terms correspond to the expectation of $(\ln q(\theta) - \ln p(\theta, y|m))$ under $q(\theta)$ which is commonly noted $\langle \ln q(\theta) - \ln p(\theta, y|m) \rangle_q$. It refers precisely to the expression of the negative *variational free energy* of q so that:

$$-F(q) = \langle \ln q(\theta) - \ln p(y|\theta, m) - \ln p(\theta|m) \rangle_q \tag{1.20}$$

Finally, the KL divergence between $q(\theta)$ and $p(\theta|y, m)$ becomes:

$$D_{KL}(q(\theta), p(\theta|y, m)) = -F(q) + \ln p(y|m) \tag{1.21}$$

The calculation of $\ln p(y|m)$ and $D_{KL}(q(\theta), p(\theta|y, m))$ is not feasible but by noticing that a KL divergence is always positive, one can see that the free energy becomes a lower bound[1] on model log-evidence ($\ln p(y|m)$) for any given distribution $q(\theta)$:

$$\ln p(y|m) \geq F(q) \tag{1.22}$$

Hence, finding the distribution $q^*(\theta)$ (sometimes referred to as the variational or approximate posterior or also the *recognition density*) that maximizes $F(q)$ ensures maximizing the evidence of the model, $p(y|m)$, and minimizing the KL difference between the approximate $q(\theta)$ and the true posterior distribution $p(\theta, y|m)$. This approach is referred to as *Variational Bayes* (VB).

A closed-form for $q^*(\theta)$ can be obtained using variational calculus, with $q^*(\theta)$ being a solution of $\frac{\partial F}{\partial q} = 0$. This problem turns out to simplify nicely when adopting the *mean-field assumption*. This approximation, typically applied in such procedures, consists in assuming that for θ being a vector of parameters, the approximate posterior $q(\theta)$ can factorize over N partitions of θ:

$$q(\theta) \approx \prod_{i=1}^{N} q_i(\theta_i) \tag{1.23}$$

with each $q_i(\theta_i)$ being an approximate of the posterior $p(\theta_i|y, m)$. When using fixed-form for q_i, VB allows inferring the hyperparameters of the corresponding parametric distribution (*e.g*: μ and σ if $q_i(\theta_i) \sim \mathcal{N}(\mu\,; \sigma)$; a and b if $q_i(\theta_i) \sim Gamma(a, b)$; ...). The use of the mean-field approximation borrows its strength to the fact that for each parameter θ_i, the approximate distribution q_i^* maximizing $F(q)$ is of the form[2]:

$$q^*(\theta_i) = \underset{q_i}{\text{argmax}} F(q) \tag{1.24}$$

Critically, by substituting $q(\theta)$ into $\prod_{i=1}^{N} q_i(\theta_i)$ in Eq. (1.20), we can derive an analytical expression for the posterior distributions q_i, also referred to as *variational updates*:

$$q^*(\theta_i) \propto \langle \ln p(y, \theta) \rangle_{q_{\setminus i}} \tag{1.25}$$

[1] This can be equivalently reformulated as the free energy being an upper bound on the statistical surprise.

[2] The *argmax* operator applied to a function f returns the values of the domain of f at which maxima are attained.

where $q_{\setminus i}$ refers to all but q_i. A complete demonstration can be found in Flandin et al. (2007). The general scheme of VB can thus be summarized as follows:

1. initialization of each $q_i(\theta_i)$ (using prior knowledge)
2. for each iteration k,
   for each parameter θ_i
   optimization of $q_i(\theta_i)$ using:
   the variational update equation for q_i
   the current estimates of $q_{\setminus i}$
3. iterations are repeated until convergence of $F(q)$

It should be noted that each update of $q_i(\theta_i)$ contributes to maximizing $F(q)$.

The Laplace approximation is also typically used in VB in the case $\langle \ln p(y,\theta) \rangle_{q_{\setminus i}}$ is intractable. This consists in assuming a Gaussian distribution for $p(y,\theta)$. The key aspect here is that both approximations (Laplace and mean-field assumptions) lend VB variational updates that express with closed-form, leading to an efficient procedure for Bayesian inversion. VB is a general scheme encompassing expectation maximization (EM, Dempster et al., 1977) and restricted maximum likelihood (ReML), that can be applied to linear and nonlinear models, as well as to generative, hierarchical models, and to dynamical models. We refer the reader to Friston et al. (2007) for a complete overview of VB.

1.4 Model comparison

In addition to model parameter estimation, model comparison is the other type of inference that Bayesian framework provides and that makes it so powerful. Model comparison aims at evaluating the plausibility of models within a specific model space (the domain of all possible models) and at simply providing quantitative (relative) clues to answer the question *"How good is my model ?"*.

1.4.1 Trading model accuracy and model complexity

The *Occam's Razor* (also referred to as the Principle of Parsimony) is a general principle that indicates that when several competing models may explain observations, one should favor the *simpler* model, that is the one embedding the smaller number of unnecessary hypothesis. Applied to Bayesian inference, this principle can be illustrated using three schematic models of different complexity (Figure 1.3):

1. Model m_1 (red in the figure). The evidence of m_1 indicates that this model is very good at describing y (it should be associated with a high value of *goodness of fit* for instance). However, one can see that it explains y and only y: it may not be generalizable to other data (acquired from another group of subjects for instance). In fact, model m_1 comprises a large number of parameters that allow fitting perfectly the data and possibly irrelevant information (noise) contained in the data. Such models are referred to as *complex* model, that should be penalized as they favor overfitting and hardly generalize.

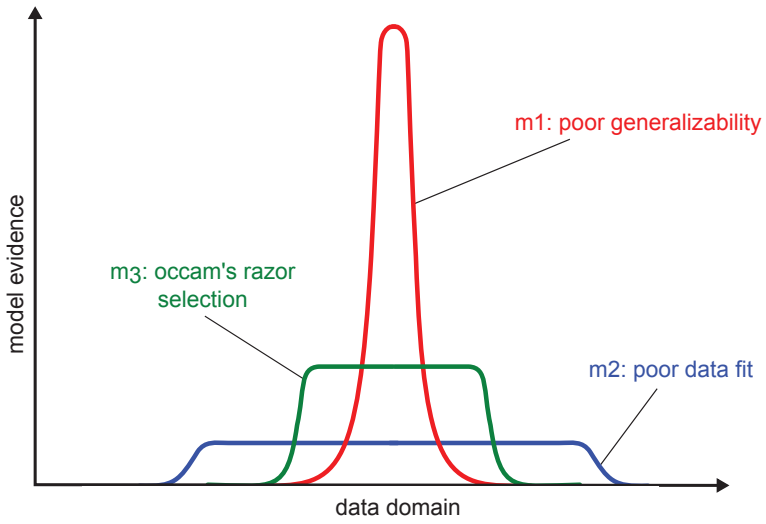

Figure 1.3 – Typical representation of Occam's razor principle. X-axis represents the data space and y-axis reports the model evidence $p(y|m)$.

2. Model m_2 (blue in the figure). This model obtains a similar but low evidence for a very large set of possible data values. It means that it can explain most observations. However, this model whose complexity is low provides a poor data fit, hence low model evidence.

3. Model m_3 (green in the figure). Finally, model m_3 appears as a good model as it allows describing y correctly without being too specific (hence precluding overfitting of y).

The principle of Occam's Razor highlights two properties that a *good* model should balance: *accuracy* (the capacity to describe the data) and *complexity* (inherited from a large number of parameters possibly interacting with each other).

1.4.2 Approximations of model evidence

As illustrated in Figure 1.3, model evidence provides information regarding how model fit and model complexity are balanced for any particular model. In fact, several approximations of model evidence have been proposed that formally account for this trade-off between these two requirements.

- *The AIC approximation.* The Akaike Information Criterion (Akaike, 1973) is given by:

$$AIC = \ln p(y|\theta^*, m) - p \tag{1.26}$$

  where θ^* represents the posterior estimate of θ and p is the number of parameters entering θ. The accuracy is thus formalized as the the likelihood of the data given θ^*, whereas complexity is represented in the form of the penalty term p.

- *The BIC approximation.* The Bayesian Information Criterion (Schwartz, 1978) is given by:

$$BIC = \ln p(y|\theta^*, m) - \frac{p}{2} \ln n \tag{1.27}$$

  where n is the number of observations. A model will be more complex in the BIC sense as either the dimension of the data or the number of parameters increases.

- *The free energy approximation.* As already introduced, the free energy is an approximation of the log-evidence. Interestingly, its expression can be rearranged to reveal an accuracy and a complexity term:

$$\begin{cases} F(m) = Accuracy(m) - Complexity(m) \\ \\ Accuracy(m) = \int q(\theta) \ln p(y|\theta, m) d\theta \\ \\ Complexity(m) = D_{KL}(q(\theta), p(\theta|m)) \end{cases} \tag{1.28}$$

  The demonstration can be found in Friston et al. (2010), Trujillo-Barreto et al. (2015). The complexity term is the KL divergence between the prior and the posterior distribution over θ. This amounts to associating a cost to letting the posterior get away from the prior. This cost has to be compensated by a better fit to provide the model a high evidence. In other words, priors will be reevaluated only if the data provide strong or accurate evidence. Contrary to the AIC and BIC criterion, the free energy is more flexible as it distinguishes parameters from one another and accounts for their interactions (covariance).

1.4.3 The Bayes Factor

Let m_i and m_j be two models. Introduced by Kass and Raftery (1995), the Bayes Factor B_{ij} allows comparing the evidence of each model:

$$B_{ij} = \frac{p(y|m_i)}{p(y|m_j)} \tag{1.29}$$

As detailed in section §1.1.3, model evidence $p(y|m_i)$ quantifies the plausibility of observing the current data y with model m_i having marginalized over model parameters θ. Interestingly, $p(m_i|y)$, the posterior probability of m_i, can also be obtained with Bayes's rule:

$$p(m_i|y) = \frac{p(y|m_i)p(m_i)}{p(y)} \tag{1.30}$$

where $p(y)$ corresponds to the plausibility of observing y having marginalized over the whole model space (it is a normalization constant). Modeling the prior distribution $p(m_i)$ as a uniform distribution is typically decided to formalize the belief that before seeing the data, all models from model space have equal probability. Under this condition, we have the equivalence:

$$p(y|m_i) \propto p(m_i|y) \tag{1.31}$$

Interpretation of the Bayes Factor is given in Kass and Raftery (1995) where for instance, a value of 20 for B_{ij} corresponds to $p(m_i|y) \approx 0.95$, and is typically interpreted as strong evidence in favor of model m_i relative to m_j. The Bayes Factor can thus be used for hypothesis testing, where H_0 could be rejected to accept H_1 if $B_{10} > 20$ (contrary to classical hypothesis testing, H_0 could also be accepted if $B_{01} > 20$).

1.4.4 Bayesian Model Selection and Bayesian Model Averaging

Bayesian Model Selection. The purpose of Bayesian Model Selection (BMS) is to select in model space, the model that is more likely than the others (the *winning* model) based on the comparison of their relative evidences. Importantly, just like Bayesian inference on parameters relies on the posterior distribution over parameters, Bayesian inference on models relies on the posterior distribution over models $p(m|y)$.
Let's consider a group of N subjects from whom observations $\{y_1, ..., y_N\}$ were collected, and M models $\{m_1, ..., m_M\}$, each corresponding to a specific hypothesis about the generation of the data. Bayesian inference for each model and each subject leads to the set of evidences $\{p(y_i|m_j)\}_{i,j}$. BMS can then be performed at the group level, either following a fixed-effect (FFX) or a random effect (RFX) type of analysis, as follows (Stephan et al., 2009):

- *Fixed-effect analysis (FFX).* This approach assumes that the same model m_i generated the data of every subjects. A multinomial distribution can be used to model prior distribution on models $p(m)$. The aim of Bayesian inference is thus to estimate the posterior distribution $p(m|y_1, ..., y_N)$.

- *Random-effect analysis (RFX).* This approach assumes that different models may have generated the different data $y_1, ..., y_N$. The parameter of the pre-cited multinomial distribution is now treated as a random variable (to be estimated) generated by a Dirichlet distribution (with a fixed parameter to be estimated). This leads to a two-level hierarchical model that can be inverted with Bayesian inference using MCMC or VB.

Note that when the number of models composing the model space become large, the sensibility of the BMS can be affected by the possibility that too many different models may be "used" by subjects. The *Family level inference* procedure proposed by Penny et al. (2010) allows addressing this issue. This approach consists in partitioning the model space according to a specific characteristic and is thereby dedicated to draw inference about this particular characteristic only. At some point, it constitutes an extension of the BMS to enable the selection of the *winning* family. This approach can deal with families having different numbers of models. Furthermore, the prior probability of each model is adjusted to account for family size. For instance, under a uniform distribution (each model having the same probability), such normalization precludes the family comprising the largest number of models to be favored.

Bayesian Model Averaging. The aim of Bayesian Model Averaging (BMA) is to provide the posterior distribution of parameter θ from different model inversions, hence with taking account of model uncertainty. It corresponds to an "average" of θ over models, denoted $\hat{\theta}$, in the sense that each posterior θ_i deriving from Bayesian inference on model m_i informs $\hat{\theta}$ in proportion to the evidence $p(y|m_i)$. Put simply, a general scheme of BMA rests on collecting N samples of θ and could be described as:

for each sample collection k
    sample a model m_i from the posterior distribution $p(m|y)$
        sample $\theta_{k,i}$ from the posterior distribution $p(\theta|y, m_i)$

The posterior distribution of θ is thus constructed from these N samples. Importantly, BMA can be applied at the individual level, at the group level and for Family level inference. We refer the reader to Penny et al. (2010) for further details about this method.

1.5 Summary

This chapter introduced the key elements of the Bayesian framework that are essential to understand this work. The rule of Bayes forms the basis of this large theoretical field and points to the four information types that are fundamental for Bayesian inference: the prior, the likelihood, the posterior and the evidence. This probabilistic framework is very well appropriate to explain data while accounting for their associated uncertainty. Formally speaking, this is achieved by describing every quantity by its probability distribution (instead of point estimate) where the variance (the inverse precision) characterizes the degree of confidence relative to our (or the system's) knowledge about these quantities. The Bayesian framework allows describing any kind of generative models (stochastic, deterministic, hierarchical, lineal, nonlinear, static, dynamic) and model inversion (being the aim of statistical inference) can involve exact methods in simple cases or most often numerical approaches, such as variational treatments. Importantly, Bayesian inference provides posterior estimates of unknown parameters (fully described by their distribution) and model evidence (or its approximation) that is necessary for model selection, a formal method to statistically compare hypothesis.

Chapter 2

The Bayesian brain

The Bayesian framework has progressively infiltrated neuroscience research as advances in the field of Artificial Intelligence, Computer Vision and Theory of Probabilities led to model human behavior using Bayesian systems. The last two decades have thus witnessed a growing interest for Bayesian models that appear promising to predict human behavior and to possibly understand the mechanisms underlying human brain function. By pointing out that the brain has to deal with uncertain information, it is not surprising that scientists considered this probabilistic framework as a valuable approach to address those issues. Uncertainty in the brain has various origins: information incoming to the brain from the external world may be noisy, partial and sometimes ambiguous, and may be coupled to internal noise (originating from the nervous system for instance). As will be detailed, in addition to persuasively describe human behavior, Bayesian models - gathered in the Bayesian brain hypothesis - also provide mechanistic assumptions regarding brain functioning. Both aspects appeal to empirical evidence to place reliability in this challenging theory.

This chapter is organized as follows: the first section is dedicated to the underpinnings of the Bayesian brain hypothesis, with a brief overview of empirical studies addressing the issue of predicting human behavior. The second section is dedicated to an even more ambitious framework, namely the free energy principle proposed by K. Friston. This principle introduces a generic computational criterion that the brain would optimize in order to survive in interaction with its environment. Importantly, this encompasses the Bayesian brain hypothesis and puts constraints on its neurobiological implementation. Finally, the third section introduces the predictive coding scheme that has been proposed as an implementation of Bayesian computation in the brain in the particular case of perception. The whole current work adheres to this general theoretical framework and aimed at testing some aspects of its neurobiological and computational underpinnings in the particular case of auditory processing.

2.1 The Bayesian brain hypothesis

2.1.1 Definition

The idea of the brain being able to perform probabilistic inference is at least two centuries old. It can be traced back to the concept of *unconscious inference* proposed by Helmholtz to explain

perception (Westheimer, 2008). Indeed, according to Helmholtz, perception would rest on recognizing the causes that have generated the sensations using previous knowledge about these causes (*e.g.* acquired from past experience) in order to disambiguate multiple interpretations (or percepts). This is exactly what Bayesian inference does and Helmholtz's concept can be formalized as follows:

$$p(cause|sensation) \propto p(sensation|cause)p(cause) \tag{2.1}$$

Replacing *cause* and *sensation* by *hypothesis* and *observation* allows generalizing this perceptual learning scheme to many brain functions spanning from perception to cognition. This is referred to as the *Bayesian brain hypothesis*, a general brain theory that rests on the two following features:

1. The brain accounts for uncertainty.

2. It does it by implementing Bayesian inference and learning.

The denomination of *Bayesian brain* was first employed in the review of Knill and Pouget (2004) but the foundation of this theory results from seminal works in various fields (including machine learning, computer vision, neuroeconomics) over the last decades, with notably the largely cited studies of Mumford et al. (1992) and Dayan et al. (1995). These pioneer works have been reviewed in Friston (2012).

2.1.2 Implications for brain function

Importantly, the Bayesian brain theory has two major implications:

- *Information processing.* Informations in the brain is coded in the form of probability distributions. As explained in §1.1.2, the lack of knowledge (or uncertainty) is formalized by the distribution variance (or inverse precision). Applying Bayes's rule allows the brain to manipulate information based on its reliability (Figure 1.1), that is by means of precision weighting. Each new observation can be used to update the current belief: the resulting posteriors would thus become the new priors. Hence, the brain *is learning* while interacting with its environment, and it does it in a Bayesian optimal way, that is by conforming to Bayes rule in order to update its belief about hidden causes in the surrounding world.

- *Computations.* To infer the unknown causes, the brain implements Bayesian calculations in order to invert and update its internal model of the world. The different necessary computations (on probability distributions) may include: combining different informations (with Bayes's rule), marginalizing over parameters, and estimating parameters from the posterior distribution (like the maximum a posteriori (MAP), the mean, ...).

Two main branches of empirical studies have emerged in order to address these central aspects of the Bayesian brain hypothesis. First aims at testing human behavior and asks: does a human observer behave as a Bayesian observer? Most of these studies involve a behavioral task and rest on comparing subject performances with predictions derived from Bayesian models of behavior. Main findings are presented in the following paragraph. The other class of studies concerns the neurobiological mechanisms that may underlie Bayesian inference in the brain. Paragraph 2.1.4 will summarize the main findings obtained at the neuronal level. At the brain scale, characterizing the feasibility of Bayesian processes requires *i)* making assumptions about the internal generative model the brain is entertaining and *ii)* formalizing this model in terms of mechanistic hypothesis

to be confronted with (real) brain activity. These assumptions refer to the neurophysiological substrates underlying Bayesian computation by the brain. Given the complex nature of the real world, non-linear hierarchical dynamical generative models have been proposed and tested in neuroimaging studies. So far, they rely on the free energy principle and the predictive coding scheme and will thereby be presented in their respective sections.

2.1.3 Empirical evidence for Bayesian behavior

Initial empirical studies investigating the Bayesian brain hypothesis were mostly conducted in the field of perception using psychophysical approaches. The work of Ernst and Banks (2002) using multisensory integration had a great influence by showing that Bayesian models can be efficient in predicting subject's performances. Precisely, in a forced-choice discrimination task, subjects had to estimate the height (h) of an object from visual (V) and haptic (H) observations, both of them being controlled by the experimenter. Visual stimuli were coupled to noise with different magnitudes to degrade accordingly the reliability of this modality. The precision of each modality was estimated from subject performances in unimodal - visual or haptic alone - sessions (from the parameters of the relative psychometric functions). These values allowed to model $p(u_V|h)$[1] and $p(u_H|h)$ the relative likelihood functions (describing the probability to observe u_V and u_H, the visual and haptic measures respectively, given h the height of the object). These functions entered a Bayesian model that predicted the performances in the multimodal session. Importantly, this session always comprised conflictual informations between modalities. Predictions were that accuracy under multimodal information should be larger than under unimodal information and that a perceptual bias should be observed toward the more reliable modality. The results did match both predictions, and were interpreted reflecting the way the brain optimally combines different sources of information (*ie*, in a Bayes-optimal manner).

Similar model-based approaches were further employed to compare human behavior with Bayesian predictions. A Bayesian ideal observer should for instance *i)* weight multiple sources of information (within or between different sensory modalities) in proportion to their relative reliability, *ii)* bias perception towards a prior knowledge in the case of ambiguous or poorly informed observations (this would account for illusory perceptions for instance) and *iii)* plan and control movements (actions) while accounting for estimated sensory and motor states (and associated uncertainties). All these predictions were evaluated empirically in several studies reviewed in Knill and Pouget (2004), with findings supporting Bayesian models (but see, O'Reilly et al., 2012, for multisensory integration). Bayes-optimal behavior were also reported in studies involving higher-order cognitive functions, such as decision-making (Behrens et al., 2007). In 2012, in a series of lectures dedicated to the Bayesian Brain (`http://www.college-de-france.fr/site/stanislas-dehaene/_course.htm`), Dehaene pointed to the review of Tenenbaum et al. (2011), reporting Bayesian models accounting for (unconscious) abstract reasoning such as language learning: for instance, inferring the meaning of a word from a limited number of expositions could rest on Bayesian model comparison to select the most plausible hypothesis. Infants may also behave *Bayes-optimally* according to the work of Teglas et al.(2011), where the authors could fit 12-month-old children performances in inductive reasoning tasks with Bayesian predictions.

[1]From now on, u will refer to observations collected by the brain, and y to observations measured by the experimenter; this distinction points to the crucial aspect of meta-Bayesian analysis, described in §2.4

2.1.4 Neural implementation: the Bayesian coding hypothesis

The processing of Bayesian calculations by neural populations is of course a great challenge and to date, according to the pioneers in the field, only weak (few) evidences have been shown so far. As detailed in Knill and Pouget (2004), and more recently in Pouget et al. (2013), this neural implementation issue gives rise to the *Bayesian coding hypothesis*, which states that:

1. Neurons encode uncertainty.

2. Neural mechanisms underlie Bayesian inference in the brain.

The main findings addressing each point have been reviewed in the articles of Knill and Pouget (2004) and Pouget et al. (2013). A brief presentation is given in the two following paragraphs.

How neurons may encode uncertainty? Neural activity has long been believed to represent a single value (for instance neuron responses in V1 could represent the orientation of a visual stimulus). However, as described in Knill and Pouget (2004), this could be challenged by several studies using single-cell recordings with monkeys, reporting neuronal responses proportional to the probability of a feature of interest (for instance, the probability of the visual stimulus to be horizontal). Beside, several neural models have been proposed that could encode the probability distribution $p(\theta|y)$ (with θ a specific feature such as orientation, and y the neural activity) with different functional forms. Some have been supported by findings from recent studies (see Pouget et al., 2013) but still, Pouget et al. stress the need for collecting a large amount of data to better characterize the likelihood $p(y|\theta)$ that is required to infer $p(\theta|y)$.

How neurons may implement Bayesian inference? As described in chapter 1, probabilistic inference involves different computations manipulating probability distributions. Research in the field of neural circuit has led to several proposals to model the corresponding calculations, some of them being supported by electrophysiological recordings macaques (see Pouget et al., 2013). Here again, the biological plausibility of the implementation of the Bayesian brain has thus been evidenced and now requires empirical validations.

2.2 The free energy principle

In 2012, Friston wrote:

> *"The future of the Bayesian brain (in neuroimaging) is clear: it is the application of dynamic causal modeling to understand how the brain conforms to the free energy principle."*

The impact of this sentence will be clarified throughout this section, starting with the definition of the free energy principle as a general theory for brain function resting on a Bayesian framework, then presenting some important implications it entails.

2.2.1 Definition

The free energy principle addresses the issue of the exchanges the brain - as a biological system - operates with its environment. It assumes a Bayesian formalization of these interactions that accounts for brain adaptation to this (ever-changing) environment.

Minimizing the statistical surprise. The free energy principle rests on the key concept of statistical surprise which has been defined in section §1.3.2 as $-\ln p(u|m)$. This quantity is a measure of the unexpectedness of an event u, that represents how likely it is given a generative model m. It ranges from 0 (in the case of u being fully expected under model m: $p(u|m) = 1$) to infinity (in the case of u being not expected, $p(u|m) = 0$). Importantly, it is conditioned on model m: different models will result in different surprises. For instance, the surprise deriving from the observation of a rain of frogs may be smaller in individuals having the knowledge of this rare phenomenon than in individuals that can not even imagine it exists. Moreover, the idea that the brain should reduce surprise can intuitively be illustrated using an evolutionary perspective: a model that does not predict the existence of a predator (hence a model associated to a large surprise as $p(predator|m) = 0$) could hardly be selected as it threatens species survival. Minimizing surprise is equivalent to generating *good* predictions. Applied to brain function, it appeals to *(i)* the ability of the brain to specify a generative model of its environment and *(ii)* Bayesian inference in the brain to estimate the unknown environmental quantities that are necessary to provide, update and evaluate predictions.

Minimizing the free energy. As shown in Eq.(1.22), the free energy F is an upper bound on statistical surprise. Hence, minimizing F ensures reducing surprise. Precisely, the free energy principle states that the brain has to minimize its free energy to maintain its equilibrium within the external environment (Friston, 2010). To do so, the brain must change its internal quantities to stay adapted to environmental changes. Practically speaking, what are these quantities (the parameters entering F) that should be optimized? The idea of the brain being at equilibrium with its environment rests on bilateral exchanges comprising the effect of the environment on the brain (which corresponds to the sensory inputs that the brain receives) and the effect the brain may exert on the environment (through action). Importantly, this prescribes two ways for the brain to minimize its free energy: on the one hand, the brain can update its model of the environment from its sensory inputs. This implies Bayesian learning of the unknown environmental quantities. On the other hand, the brain can trigger an action to make its predictions come true. Both strategies contribute to minimize surprise. The first one corresponds to *perceptual inference and learning*, while the latter is also referred to as *active inference*. This twofold optimization can be cast in a single Variational Bayes (VB) scheme (Friston et al., 2010).

A comment on the thermodynamic free energy. The free energy is most well known from thermodynamics, where it quantifies the part of a system's internal energy that can be transferred into the environment in the form of useful work. Together with entropy H measuring the degree of disorder of a system, they constitute state variables describing the dynamics of a system. Friston and Stephan (2007a) drew a close parallel between the thermodynamic and the statistical (or information theoretic) free energy. This allowed interpreting the free energy principle as an account of how any biological system (and the brain in particular) maintains itself in an ordered state and prevents non-viable ones (associated with large entropy) that would compromise its

living. However, thermodynamics and information theory share equivalences but also differences. To prevent possible misinterpretation, one should refer to Friston et al. (2006) where the authors clarified that the free energy principle rests on the statistical formulation of the free energy.

2.2.2 Hierarchical and dynamic causal models in the brain

The cornerstone of the free energy principle remains the generative model of the external environment that the brain must entertain. The most advanced and generic form of models is a hierarchical dynamic causal model, which can capture the hierarchical, dynamical and non-linear nature of the world. From an evolutionary perspective, simple models may not survive natural "model" selection as they would fail explain the world and hence to minimize surprise. This paragraph addresses the issue of the biological plausibility for the brain to entail hierarchical dynamic causal models and then describes a general structure for such generative models that have been proposed in the literature and that will be used throughout this work.

Prerequisite. To represent hierarchical models, the brain should be itself hierarchically organized. Moreover, Bayesian model inversion (inference) demands the organization of connections between (cortical) areas to allow for two distinct (and segregated) pathways to convey messages within the hierarchy:

- a top-down pathway propagating information in the "generative" direction (from causes to consequences), leading to predictions or expected interactions with the environment,
- a bottom-up pathway propagating information in the "recognition" direction (from consequences to causes), leading to the recognition of the causes of current sensations.

Hierarchical organization of the cortex. A large number of studies (including the seminal work of Felleman & Van Essen, 1991b), reported hierarchical architecture in the macaque visual cortex. A key feature characterizing a cortical hierarchy pertains to the laminar patterns of the origins and the terminations of extrinsic[1] connections. Using the technique of neuronal tracing, an extensive work has pursued the pioneering study of Rockland and Pandya (1979) to categorize cortico-cortical connections in *feedforward*, *feedback* and *lateral* connections. The latter are acknowledged to have no laminar preferences: stemming from all cortical layers, they terminate in all layers of the target neuron. In the review of Markov and Kennedy (2013), feedforward and feedback pathways building up two counterstreams linking the upper (supra) and lower (infra) compartments of the cortical surface are detailed as follows:

- *feedforward connections*: they comprise connections originating in the supragranular layers (termed as supra-feedforward) or in the infragranular layers (infra-feedforward) and terminating predominantly in layer 4 (layer 6 is also reported). Supra-feedforward connections would be more dominant as the length of connection increases.
- *feedback connection*: they comprise connections originating in the supragranular layers (supra-feedback) principally targeting the supragranular layers, and those originating in the infragranular layers (infra-feedback) terminating in infra- and supragranular layers. Infra-feedforward would be more dominant as the length of connection increases.

[1] *extrinsic* refers to long-distance connections, and is to be opposed to *intrinsic* which indicates connections confined to a local neuronal circuit.

On the basis of the feedforward and feedback pathways proposed by Rockland and Pandya (1979), Van Essen and collaborators proposed a hierarchical cortical organization for the monkey visual cortex (Maunsell & Van Essen, 1983; Felleman & Van Essen, 1991b). To date, the hierarchical architecture of the cortex does not seem to be questioned but, as pointed by Markov and Kennedy (2013), its complexity lends to challenging perspectives to improve its characterization and increase our understanding of the functional role of connections. A detailed review of the cortical connectivity properties that support Bayesian inference is provided in Friston (2005), (and also in Friston & Kiebel, 2009; Friston & Stephan, 2007b), with in particular the notion that feedforward and feedback connections are associated with driving and modulatory effects respectively, (but see Bastos et al., 2012; Markov & Kennedy, 2013). A critical remaining question is how generalizable are the findings in the macaque visual cortex to other brain areas and other species? In Anderson and Martin (2015), the extensive search for a canonical scheme for local circuits in the cortex is reported, and a recent successful example of animal model applied to human is cited (Heinzle et al., 2007). In Bastos et al. (2012), a proposition for a canonical model accounting for the predictive coding prerequisites is described (as well as a detailed review of the literature regarding the hierarchical organization of the cortex). We will come back to the model of Bastos and collaborators in chapter 4, §4.2, as it is at the heart of the DCM that we used. Such attempts must however face *"the various morphological cell types and how this physiology differs across cortical layers and areas"* (Markov & Kennedy, 2013).

General structure for a hierarchical dynamic generative model of the environment. In Friston and Kiebel (2009) and in Feldman and Friston (2010), a very general structure of generative models possibly implemented in the brain is described (in line with the above findings and arguments). Precisely, the authors consider a generative model of the form $p(u, \vartheta) = p(u|\vartheta)p(\vartheta)$, u being the observations by the brain (the sensory inputs), and $\vartheta = \{x, \nu, \theta\}$ the hidden variables participating to the internal dynamics. First quantity, x, represents the hidden states of the model, and reflects the propagation of internal information over time (like a memory) as it links time samples together. The latter two refer to the directed graphical model structure employed to represent causal relationships between levels within the hierarchy (see §1.2.2): parameters within each level are denoted θ, and *causes* ν establish the links between levels. Both states and causes vary over time (they have their own trajectories), whereas θ are constants (or vary much more slowly than x and ν) and may vary across individuals. At each level i of the hierarchy, the dynamics of states and causes can be described using the following equations (the time dependency of x and ν has been omitted here for the sake of simplicity):

$$\begin{cases} \nu^{(\mathrm{i})} = {f_\nu}^{(\mathrm{i}+1)}(x^{(\mathrm{i}+1)}, \nu^{(\mathrm{i}+1)}, \theta) + {z_\nu}^{(\mathrm{i}+1)} \\ \\ \dot{x}^{(\mathrm{i})} = {f_x}^{(\mathrm{i})}(x^{(\mathrm{i})}, \nu^{(\mathrm{i})}, \theta) + {z_x}^{(\mathrm{i})} \end{cases} \tag{2.2}$$

with ${f_x}^{(\mathrm{i})}$ and ${f_\nu}^{(\mathrm{i})}$ being non-linear (deterministic) continuous state functions and ${z_x}^{(\mathrm{i})}$ and ${z_\nu}^{(\mathrm{i})}$ reflecting random fluctuations. These latter are assumed to be gaussian (with zero mean and covariances Σ_x, Σ_ν parameterized by hyperparameters γ_x, γ_ν respectively) and conditionally independent. At the lowest level, we have:

$$\nu^{(0)} = {f_\nu}^{(1)}(x^{(1)}, \nu^{(1)}, \theta) + {z_\nu}^{(1)} = u \tag{2.3}$$

and $f_\nu{}^{(1)}$ is thus an observer function, mapping hidden states $x^{(1)}$ and $\nu^{(1)}$ to $\nu^{(0)} = u$, the data observed by the brain. The random term $z_\nu{}^{(1)}$ here refers to the observation noise. At the highest level k we have:

$$\nu^{(\mathrm{k})} = \eta + z_\nu{}^{(\mathrm{k}+1)} \tag{2.4}$$

where η is the prior mean on cause $\nu^{(\mathrm{k})}$. Empirical Bayes (EB, §1.2.2) assumption accounts for the fact that priors on causes may not be generated *de novo* but have to be learned from past observations (for considerations regarding priors at the top of the hierarchy, see Friston, 2005). An illustrative example of such a generative model composed of three levels is described in the figure below, using a graph-based structure as can also be found in (Bastos et al., 2012):

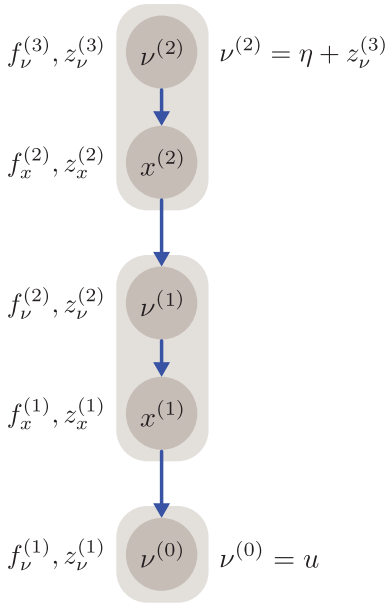

Figure 2.1 – Example of a hierarchical model composed of three levels.

Whether such generative model *truly* exists or not is not the important point, it is all about modeling the brain's representation of the world (however, according to Friston (2005), the brain's anatomy and physiology could have evolved to mirror the hierarchical and dynamic structure of the world). Importantly, the brain could specify several models and use Bayesian model selection to select the more likely given its observations and prior knowledge.

2.2.3 Perceptual Inference, Perceptual Learning and Active Inference

According to the free energy principle, the brain has two options to minimize surprise (and make its prediction match its sensory inputs).

First, it could adapt its internal quantities to suppress (reduce) the errors of its predictions. Using above formalism, this strategy amounts to optimizing ν and x. This is referred to as *perceptual inference*. It also rests on optimizing θ, which referred to as *perceptual learning*. As explained in Friston (2005), an expectation maximization (EM) approach constitutes a possible scheme to solve this optimization problem. Importantly, these considerations also apply to Bayesian models of perception that do not necessarily rest on the free energy principle. As will be described in

the next section, Bayesian models of perception could be implemented using a predictive coding scheme. The present work focuses on such models in the context of the auditory modality.

The second strategy pertains to action and rests on active inference: if the internal predictions do not match the sensory inputs, the only remaining possibility for the brain to minimize surprise is to trigger actions that will change or observe the environment such that predictions will be fulfilled. As can be seen in Eq.(2.2), action a does not enter the generative model but can influence inputs (u becomes $u(a)$). Importantly, active inference appears promising to account for different brain functions involved in reflexive behaviors such as saccades but also higher level intentional behaviors. These include for instance: eye movements (*e.g.* oculomotor pursuit, Adams et al., 2016), action observation and motor control (Friston et al., 2010), attribution of agency (Brown et al., 2013), decision-making (Friston, Rigoli, et al., 2015; Schwartenbeck, Fitzgerald, Mathys, Dolan, Kronbichler, & Friston, 2015), emotional adaptation (Joffily & Coricelli, 2013), not to mention several dysfunctions associated with psychiatric disorders (Stephan et al., 2015a). To date, active inference studies remain scarce and rest on simulations (which have proved useful but are limited to face validating the framework against behavioral or neurophysiological observations). Future studies should now address the predictive validity of active inference at different levels and scales.

2.2.4 The Bayesian brain and the free energy principle

The Bayesian brain hypothesis simply assumes that perception and learning by the brain conform to Bayes's rule, thus combining in an optimal fashion prior knowledge and incoming sensory evidence in order to update current belief. Although this is an important and far reaching statement, this does not prescribe what model of the world a given brain implements and neither how this model and the ensuing computation is implemented. These latter aspects are essentials in order to characterize brain functions and dysfunctions. Some authors have questioned Bayes optimality in the brain, even in healthy subjects (O'Reilly et al., 2012). Although this might well be the case, it is not trivial to demonstrate because of an indeterminacy which is worth noticing. Indeed, Bayesian inference and learning might well be optimal and conform to Bayes's rule while proceeding under a suboptimal or maladaptive model of the world, which may yield the conclusion that the brain is not optimal.

The free energy principle encompasses the Bayesian brain hypothesis and goes further in stating that a living organism optimizes free energy in order to survive. Optimizing free energy amounts to minimizing surprise, which for the brain can be achieved by both updating the model of the world and acting upon the world or its sampling. Moreover and beyond the Bayesian brain hypothesis, the free energy principle prescribes the way the functional anatomy is (hierarchically) organized in order to implement empirical hierarchical Bayes and surprise minimization. This yields important predictions that can be confronted to current knowledge of cortical organization and message passing in the brain.

As such, the free energy principle also encompasses the predictive coding (see §2.3) hypothesis which was first introduced to explain (visual) perception. In contrast with the Bayesian brain hypothesis, predictive coding emphasizes a computational scheme rather than an algorithmic theory, in the sense of David Marr's levels of analysis (Marr, 1982). Importantly though, the

free energy principle brings together those different aspects pertaining either to the underlying functional anatomy or to the underlying computation. In the current literature, those terms are sometimes used interchangeably. Nevertheless, the Bayesian brain hypothesis is often referred to in computational studies trying to explain mental processes and subsequent behavior. Predictive coding is rather invoked in electrophysiological and neuroimaging studies addressing the mechanisms underlying perceptual inference. To date, the free energy principle is rather mentioned in methods papers showing how a given behavior or neurophysiological phenomenon can be explained in the light of this framework.

These distinctions may explain some recurrent debates upon the validity of the free energy principle. Some have even questioned the empirical possibility to falsify this principle. The original work described hereafter did not intend to contribute to this somewhat philosophical debate. However, it took great advantage of all the recent and very advanced modeling approaches that originate from all the powerful ideas that the free energy principle has developed and greatly refined.

2.3 The predictive coding implementation for perception

2.3.1 Definition

Predictive coding refers to inference methods where the estimation process is driven by some prediction errors and stops when these have been suppressed. Precisely, at each level i of the

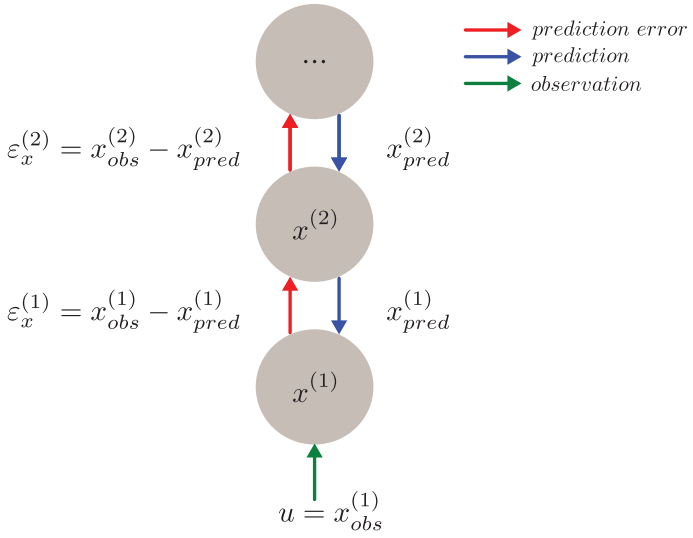

Figure 2.2 – General scheme for predictive coding

hierarchy, a prediction error relative to a specific quantity x_i is defined as the difference between its current value informed by the current observation u and its expected value derived from model predictions: ${\varepsilon_x}^{(\mathrm{i})} = {x_{obs}}^{(\mathrm{i})} - {x_{pred}}^{(\mathrm{i})}$. The key feature of predictive coding pertains to the *message-passing scheme* taking place within the hierarchy, that comprises the propagation of errors in the bottom-up (recognition) direction and the propagation of predictions in the top-down (generative) direction, until convergence (*i.e.* suppression of prediction errors). Figure (2.2) presents the general principle of predictive coding. Each time a new observation u is experienced,

the prediction errors, the internal (unknown) quantities and the predictions are updated according to specific equations embodying model assumptions. One of the most cited predictive coding scheme remains the model of Rao and Ballard (1999), initially designed for the visual system. This framework as well as others designed for perception have been reviewed in Spratling et al. (2016).

2.3.2 Generalized predictive coding

The predictive coding implementation for perception proposed by Friston and collaborators, referred to as *generalized predictive coding*, has been described in several articles (Friston, 2005; Friston & Kiebel, 2009; Feldman & Friston, 2010; Bastos et al., 2012) and rests on mathematical underpinnings which can be found for instance in (Friston, 2008). We will provide here the key principles that are required for the understanding of the hypothesis addressed in the current work.

The general form of the generative model considered here is the one described in Eq.(2.2). The first key aspect of this scheme pertains to the fact that it rests on Variational Bayes (VB) to estimate the unknown causes of sensory inputs and minimize free energy. This is a consequence of the equivalence between the minimization of the free energy and the minimization of prediction errors that has been introduced in §2.2.3. Indeed, minimizing F aims at eliciting predictions fulfilling incoming sensations, hence reducing prediction errors (the mathematical demonstration leading to the main equation of generalized predictive coding is provided in Friston, 2008). Putting this equivalence into practice, it means that the iterative VB scheme (until convergence in minimizing F is attained) is equivalent to message passing in the predictive coding scheme (which also operates iteratively until prediction error is suppressed).

In Friston and Kiebel (2009) and in Bastos et al. (2012), the authors focused on the recognition of states x and causes ν that serve perceptual inference. As described in §1.3.2, VB rests on the specification of a recognition density $q(x, \nu)$ that approximates the true posterior distribution $p(x, \nu|u)$ and for which the Laplace approximation is considered. The gaussian form $q \sim \mathcal{N}(\mu, C)$ thus gives the mean expectations μ_x and μ_ν as the conditional expectations for x and ν. Prediction errors relative to x and ν are thus defined at each level i of the hierarchy as the difference between the conditional expectations and the associated predictions prescribed by state functions ${f_x}^{(\mathrm{i})}$ and ${f_\nu}^{(\mathrm{i})}$:

$$\begin{cases} {\varepsilon_x}^{(\mathrm{i})} = {\mu_x}^{(\mathrm{i})} - {f_x}^{(\mathrm{i})}({\mu_x}^{(\mathrm{i})}, {\mu_\nu}^{(\mathrm{i})}, \theta) \\ \\ {\varepsilon_\nu}^{(\mathrm{i+1})} = {\mu_\nu}^{(\mathrm{i})} - {f_\nu}^{(\mathrm{i+1})}({\mu_x}^{(\mathrm{i+1})}, {\mu_\nu}^{(\mathrm{i+1})}, \theta) \end{cases} \tag{2.5}$$

Precision of predictions refer to the inverse variance of the random terms ${z_x}^{(\mathrm{i})}$ and ${z_\nu}^{(\mathrm{i})}$ in Eq.(2.2). They write ${\Pi_x}^{(\mathrm{i})}$ and ${\Pi_\nu}^{(\mathrm{i})}$ respectively. Given the expression for F, its gradients (the partial derivatives of F with respect to x and ν) can be derived; they constitute a system of differential equations for which the conditional expectations μ_x and μ_ν provide optimal solutions. These equations allow updating the internal quantities of the model, namely the precision-weighted predictions errors ${\xi_x}^{(\mathrm{i})} = {\Pi_x}^{(\mathrm{i})}.{\varepsilon_x}^{(\mathrm{i})}$ and ${\xi_\nu}^{(\mathrm{i})} = {\Pi_\nu}^{(\mathrm{i})}.{\varepsilon_\nu}^{(\mathrm{i})}$, and the conditional expectations ${\mu_x}^{(\mathrm{i})}$ and ${\mu_\nu}^{(\mathrm{i})}$. The update equations can be found in Bastos et al. (2012) (Eq.1) where they are expressed in generalized coordinates of motion that allow accounting for the different time derivatives of each variables. These equations reveal a message-passing scheme (Figure 2.3) with precision-

weighted prediction errors being informed by changes in conditional expectations, and vice-versa. The notion of *precision-weighted* prediction errors is a critical feature of this model that endows the brain with the ability to weight the information it manipulates according to its relative (estimated) uncertainty.

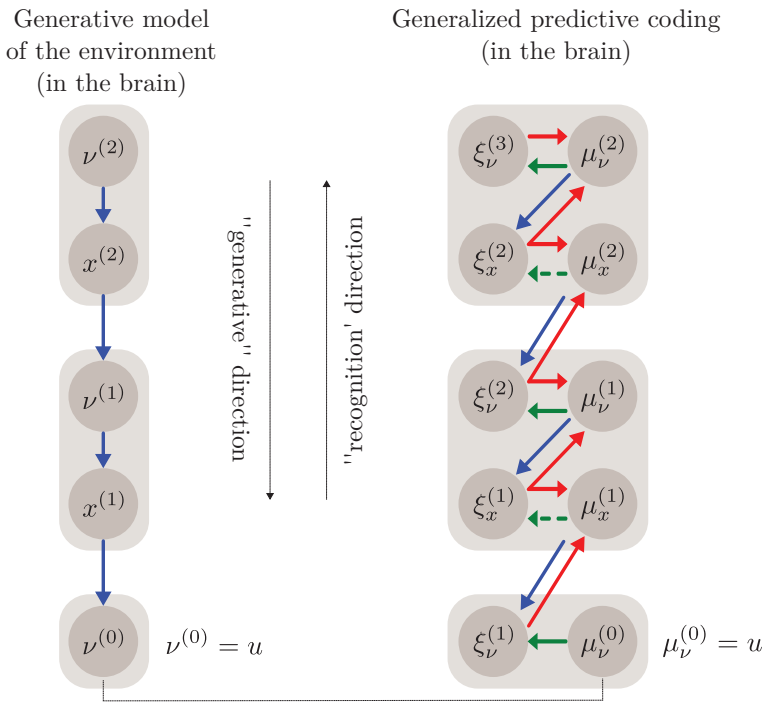

Figure 2.3 – Example of a generalized predictive coding scheme as proposed in Bastos et al. (2012), and the associated generative model. The color code for arrows is the same as in Figure (2.2); dotted green arrows indicate that both predictions and observations enter the target unit.

Figure (2.3) allows visualizing the key features of this predictive coding scheme:

- each level exchanges information with its first parent (level above) and its first children (level below),
- prediction errors integrate top-down and lateral (within source) information from the level above and the same level, respectively,
- updates in conditional expectations are guided by the prediction errors from the same level and from the level below, leading to lateral and bottom-up messages respectively.

These rules have implications on the neurobiological implementations of message passing:

- neuronal activity should encode the cited updates; therefore each level i of the hierarchy should comprise four neuronal units: two *error units* to compute ${\xi_x}^{(i)}$ and ${\xi_\nu}^{(i)}$ and a *state unit* and a *cause unit* to update conditional expectations ${\mu_x}^{(i)}$ and ${\mu_\nu}^{(i)}$ respectively.
- extrinsic connections should convey prediction and prediction error messages, with the former involving feedback connections, and the latter feedforward ones.
- intrinsic connections should embed the computations of conditional expectations and prediction errors within each level of the hierarchy.

Finally, the generalized predictive coding allowing the recognition of parameter θ is described in Friston (2008) and in Feldman and Friston (2010). This corresponds to perceptual learning. In Feldman and Friston (2010), the authors proposed that the neuronal activity encoding the conditional expectation of θ could be reflected by the gain of the post-synaptic responses of the error units. They further argue in favor of attention modulating the precision of prediction. Importantly, based on this scheme for perception, Friston (2005) suggested that electrophysiological responses should support this message-passing framework and proposed that evoked responses scale with prediction errors in the brain. In addition, some evidence has been reported in Bastos et al. (2012) suggesting that oscillations in the gamma band (>30 Hz) would be the biological marker of bottom-up prediction errors while oscillations in the beta band would encode top-down predictions. All these hypotheses hence allow relating neuroimaging and neurophysiological findings with (Bayesian) computational principles in the brain and thereby motivated a series of empirical studies whose main findings are presented in the following paragraph.

2.3.3 Empirical evidence

Predictive coding furnishes testable predictions regarding how brain activity (reflecting sensory information processing) should be modulated by experimental manipulations. Precisely, if the brain were to implement Bayesian inference, brain responses elicited by a specific stimulus should reflect the way predictions and precision-weighted prediction errors are updated.

Two recent comprehensive reviews (Summerfield & de Lange, 2014; Kok & de Lange, 2015, with the former focusing on visual perception) report the existing findings in line with predictive coding. Most of reported studies addressed the top-down influence of expectations (priors) on low-level sensory processing and were conducted with various paradigms (repetition suppression sequences (see below), cue-priming tasks, paired association learning, illusory contours or motions, stimulus omission) in humans and monkeys using different neuroimaging techniques (fMRI, EEG, MEG, single-cell recordings). A noticeable point is that predictive coding allows re-interpreting existing results, like for instance, those reported in Kok and de Lange (2015) regarding audiovisual integration for speech processing. Besides, it allows formulating new (mechanistic or model-based) hypothesis, to address for instance the influence of the precision of information (priors or sensory inputs) on brain responses.

An influential model-based fMRI study is the one by Summerfield et al. (2008), where the authors tested the predictive coding account for repetition suppression (RS). RS is a well-known effect characterized by the decrease of brain activity associated with the second (and subsequent) presentation(s) of a stimulus in comparison to the initial activity elicited by its first presentation. It can take many forms, and it has been observed with several neuroimaging techniques, at different timescales and in different species. RS has long been thought to reflect bottom-up mechanisms but a predictive coding interpretation would also imply top-down processes (Grill-Spector et al., 2006): these would generate sensory low-level predictions fulfilled by the repeated stimulus, leading to an absence of prediction-error related activity. Summerfield et al. (2008) designed a smart study dealing with expected and unexpected repetitions. This contextual manipulation was found to modulate RS amplitude as predicted by predictive coding, namely smaller RS amplitudes observed with unexpected repetitions. Following this, further studies were carried out to

better characterize this account, and have been recently reviewed in Auksztulewicz et al. (2016), Grotheer and Kovács (2016) (see also chapter 3, §3.4.3). Beside, the predictive coding account of degraded speech perception could be addressed in a recent EEG-MEG study (Sohoglu & Davis, 2016). Three factors were manipulated: the precision of sensory inputs (using vocoders to synthesize words pronounced with varying intelligibility), the precision of predictions (using matching or mismatching cues presented before the vocoded words) and the timescale of perceptual learning (using speech recognition tasks at a short "cue-stimulus" scale, and at a long "experimental-session" scale). Their results, including behavioral computational simulations, strongly support the predictive coding as a unifying approach to explain the effect of their experimental manipulations. Together with the above cited RS studies, these findings substantiate predictive coding but also, and importantly, they nicely illustrate how such mechanistic hypothesis can improve our understanding of a perceptual process.

A critical issue however remains: consistent evidence is not direct evidence of predictive coding operating in the brain. When it comes to assigning a functional role to the modulations of neuronal responses described in the studies cited here, one remains limited to speculative interpretations. For instance, as noticed in Kok and de Lange (2015), the activation of low-level visual area V1 observed using single-cell recordings in macaques for illusory triangles (hence in the absence of physical sensory input) could either reflect top-down predictions from V2 or a prediction error computed in V1. In fact, validating the predictive coding requires:

- the identification of the *separate* state and error units that compose a level in the hierarchy
- the characterization of their computational processes (namely, prediction and precision-weighted prediction error updates respectively) which involves intra-level and inter-level information exchanges.

Regarding the former point, some pieces of evidence of separate activities have been reported in Summerfield and de Lange (2014), Kok and de Lange (2015), based on the assumption that expected sensory input (matching priors) should induce an increase of state unit activity and a decrease of error unit activity. Kok and de Lange (2015) suggested from findings in macaque, that both units could be implemented in the different layers of a cortical column. The second aspect requires establishing the computational role of the intrinsic and extrinsic connectivities at play in the cortical hierarchy.

Today, addressing these two points remains great challenges that appeals to advanced modeling, aiming at describing from both a neurophysiological and computational perspectives, the internal circuitry of a level and its connections within the hierarchy. Dynamic Causal Modeling (Friston et al., 2003; Kiebel et al., 2009) - combined with the canonical microcircuit proposed by Bastos et al. (2012) - has been designed especially in order to tackle the neurophysiological-level description (see for instance this review of DCM findings for RS, Auksztulewicz & Friston, 2016). DCM is central to the present work, we will describe its underpinnings in chapter 4, §4.2. Besides, advanced computational learning models rests on a meta-Bayesian scheme (to be described in the next section). They are mandatory in order to test the mapping between computational and neurobiological processes. A exemplary study is the one by Iglesias et al. (2013): in this fMRI study using an associative learning task, the inversion of a hierarchical dynamic generative model (namely a hierarchical gaussian filter, HGF, Mathys et al., 2011, 2014) from BOLD

measures allowed to spatially characterize different prediction errors computed at different levels of a hierarchy. One remarkable aspect of these findings is that low-level prediction errors were found to involve the midbrain, known to enter the reward system. This study thus contributed to elaborate the emergent hypothesis of neuromodulation signaling predictive coding messages in the brain. Computational learning models were also crucial to our work, and will be described in chapter 4, §4.3.

2.4 Meta-Bayesian analysis

Under the assumption of the Bayesian brain, a close parallel can be drawn between the neuroscientist inferring unknowns from observed *brain* data, and the *brain* itself inferring unknowns from observed sensory inputs. In fact, it goes beyond noticing the similarity between these two types of Bayesian inference schemes, since the experimenter's level of analysis encapsulates the brain's level one. Therefore, when this experimenter's perspective is made explicit, it is referred to as *meta-Bayesian* approach. In this section, we will first introduce the meta-Bayesian scheme proposed by Daunizeau et al. (2010). Then, we will summarize the three model types used throughout this work with the aim to clarify this meta-Bayesian approach that emerges when working with a Bayesian framework to investigate Bayesian inference in the brain. Finally, we present the VBA Toolbox for nonlinear probabilistic model inversion (Daunizeau et al., 2014), which we used to perform meta-Bayesian analysis.

2.4.1 Observing the observer

In a series of two papers from Daunizeau and collaborators (Daunizeau, den Ouden, Pessiglione, Kiebel, Stephan, & Friston, 2010; Daunizeau, den Ouden, Pessiglione, Kiebel, Friston, & Stephan, 2010), a meta-Bayesian analysis for neuroscientists was proposed as a principled framework to tackle the *"observing the observer"* issue:

- the *observer* refers to a participant engaged in an experimental task (hence observing its environment). Neuronal mechanisms underlying perception and decision-making[1] are assumed to conform to Bayesian principle. During the task, the behavioral or neuroimaging data that should reflect such mechanisms are collected.

- *observing* refers to the experimenter analyzing these data with the aim to characterize the unobservable perceptual and decision-making processes. This characterization itself involves Bayesian inference in order to recognize the hidden states or parameters entering such processes.

Using the formalism introduced in these papers, the (neuroimaging or behavioral) data generative (meta-Bayesian) model is made of two main parts:

- *The perceptual model.* This model, $m^{(\mathrm{p})}$, pertains to the observer and describes how its sensory inputs u are generated given the unknown causes ν in the environment, and some hidden parameters $\vartheta^{(\mathrm{p})}$. Such a model can be expressed as a combination of the likelihood of observing the sensory inputs and some priors about the hidden causes and parameters:

$$p(u, \nu, \vartheta^{(\mathrm{p})}; m^{(\mathrm{p})}) = p(u|\nu, \vartheta^{(\mathrm{p})}; m^{(\mathrm{p})}) p(\nu, \vartheta^{(\mathrm{p})}|m^{(\mathrm{p})}) \tag{2.6}$$

  Model inversion rests on variational Bayes scheme, which entails minimizing the free energy under the free energy principle. The obtained posterior estimates (of ν and $\vartheta^{(\mathrm{p})}$) reflect the

[1] One of the strength of the meta-Bayesian framework presented here pertains to its ability to formalize and solve issues in the field of Bayesian decision theory. Because this aspect is beyond the scope of this work, we refer the interested reader to (Daunizeau, den Ouden, Pessiglione, Kiebel, Stephan, & Friston, 2010; Daunizeau, den Ouden, Pessiglione, Kiebel, Friston, & Stephan, 2010) for more detailed information, and to (Devaine et al., 2014) for an application in the context of theory of mind.

updated beliefs, and influence his or her (behavioral o neurophysiological) response through the response model.

- *The response model.* This model, $m^{(\mathrm{r})}$ describes how the subject's internal states which are hidden to the experimenter, map onto the observations y. This mapping depends upon the experimental design u and unknown parameters. It writes:

$$p(y, u, \vartheta^{(\mathrm{r})}; m^{(\mathrm{r})}) = p(y|u, \vartheta^{(\mathrm{r})}; m^{(\mathrm{r})}) p(\vartheta^{(\mathrm{r})}|m^{(\mathrm{r})}) \tag{2.7}$$

  As outlined by the authors, a critical aspect of meta-Bayesian analysis comes from the fact that the response model integrates the perceptual model and its inversion but also the mapping of associated quantities to measurable responses. Here again, the variational scheme can handle model inversion to provide posterior beliefs about $\vartheta^{(\mathrm{r})}$.

Importantly, each model inversion depends on the uncertainty associated with the likelihood and the priors that is passed into the free energy at convergence. As explained in Daunizeau et al. (2010), $F^{(\mathrm{p})}$ and $F^{(\mathrm{r})}$ should not be confused as they approximate model evidence of different observed data. The authors indicate the formal relationship between these two quantities and discuss how the uncertainty regarding the observer's knowledge can be estimated from the uncertainty in the experimental data.

2.4.2 Examples of response models

We describe here the general form of three Bayesian generative models corresponding to the three main types of models we used. Note that the first two models correspond to response models, and do not rely on any assumed perceptual model that the subject may entail. As such, they subsume a simple, classical Bayesian modeling approach for neuroimaging and do not require a meta-Bayesian scheme. Only the third model exploits the meta-Bayesian framework. Various Bayesian methods have been proposed for different types of analysis conducted with behavioral or neuroimaging data. The most popular toolbox for Bayesian neuroimaging data analysis is arguably the SPM package (`http://www.fil.ion.ucl.ac.uk/spm`). In the particular case of electrophysiological data, SPM proposes Bayesian methods that have been proved well suited (and successful) to solve the inverse problem of EEG-MEG source reconstruction (*case 1*), and to estimate the effective connectivity underlying the generation of electrophysiological data (DCM, *case 2*). The third case (*case 3*) refers to another type of dynamic causal models, referred to in this work as computational models as they aim at describing mental processes. Importantly and in contrast with the two first models, these are meta-Bayesian models since mental processes are not directly observed but through a mapping to electrophysiological measures.

1. *Static model for source reconstruction.* The method proposed in SPM for distributed source reconstruction rests on a two-level hierarchy of the form:

$$\begin{cases} Y = LJ + \varepsilon_n \quad with \quad \varepsilon_n \sim \mathcal{N}(0\,; \Sigma_n) \\ \\ \theta \sim \mathcal{N}(0\,; \Sigma_s) \end{cases}$$

where Y denotes the (static) observation by the experimenter (EEG or MEG data for instance), L the (known) leadfield operator, J the unknown parameter (the cortical source activity), Σ_n and Σ_s the unknown noise covariance at the sensor and source levels respectively. Model inversion (using VB) requires specifying the prior distributions of parameter J and hyperparameters λ_n and λ_s characterizing the noise distributions Σ_n and Σ_s respectively (Mattout et al., 2006).

2. *Dynamic causal model for effective connectivity estimation.* The generative model of DCM proposed in SPM, is of the form (Kiebel et al., 2006):

$$\begin{cases} \dot{x}(t) = f(x(t), u(t), \theta) \\ \\ y(t) = g(x(t), u(t), \psi) + \varepsilon_n \quad with \quad \varepsilon_n \sim \mathcal{N}(0\,; \Sigma_n) \end{cases}$$

where y denotes the observation by the experimenter (EEG or MEG data for instance), x the hidden states (describing neuronal activity), u the causes (the input of this dynamic system), θ and ψ the unknown evolution and observation parameters. Model inversion also rests on VB and requires specifying the prior distributions of the hyperparameter λ_n characterizing the distribution of the observation noise Σ_n (the evolution model is assumed to be deterministic), the evolution and observation parameters θ and ψ and the initial state value x_0. Note that the causes u entering the generative model are the experimental inputs, thereby controlled by the experimenter.

3. *Dynamic causal model for assessing mental processes.* We present here a general form for such models (examples of application to the cognitive processes behind the MMN will be described in chapter 4, §4.3). Usually, these models are expressed in discrete time since observations collected by the experimenter correspond to the brain activity elicited at each trial. Since these models are hierarchical, causal and dynamic, they write:

$$\begin{cases} x_{t+1} = f(x_t, u_t, \theta) \\ \\ y_t = g(x_t, u_t, \psi) + \varepsilon_n \end{cases}$$

where y denotes the observation by the experimenter (behavioral, EEG or MEG data for instance), x the hidden states (describing activity within computational units), u the causes (the input of the system), θ and ψ the unknown evolution and observation parameters. Model inversion requires specifying the prior distributions as for DCM. Again, the causes u entering the generative model are known by the experimenter and thereby are not to be estimated. Most importantly, f here entails the Bayesian inference that the subject is performing under his/her model of the world, while g is the observation model mapping the subject's (hidden) mental representations to available observations. As such, f embodies equations which result from the Bayesian inversion of the subject's perceptual model of the experimental task.

In the same manner as the brain *learns* (updates its prior knowledge) after having collected a new information, the experimenter *learns* from behavioral or neuroimaging data. From both perspectives, this learning process provides posterior estimates of unknowns that become the

priors for next observations. Additionally, Bayesian inference provides the free energy which is an approximation of the model log-evidence used for model comparison. Model comparison is at the heart of experimental science and is essential for the experimenter, as we shall see in this work. It might also be at the heart of human perception and decision-making (Friston et al., 2012).

2.4.3 The VBA toolbox

The aim of this subsection is to briefly describe the VBA toolbox that we used for our computational model analysis. This toolbox has been introduced in (Daunizeau et al., 2014) and can be downloaded from the website `http://mbb-team.github.io/VBA-toolbox`.

The VBA (*Variational Bayes Analysis*) toolbox is a Matlab package dedicated to the inversion of probabilistic nonlinear state-space models of behavioral and neuroimaging data. It thus addresses Bayesian inference from data observed *by the experimenter* (the causes u or inputs being controlled). The generative models can be static or dynamic, probabilistic or deterministic (in this case, state or observation noise is set to 0). The core of this toolbox rests on VB inversion using a Laplace approximation. Below are listed the four key features handled by VBA:

- *Data simulation.* This point deals with the generative model (from causes to consequences) and usually refers to preliminary analysis aiming at optimizing model specifications. Using the known trajectory of cause u (a time series or a discrete sequence of stimuli for instance), the trajectories of hidden states and causes can be computed, as well as predictions of neuroimaging data y_{pred}. This step thus allows checking whether an experimental modulation observed on the *real* data (for instance $y_A < y_B$, for two conditions A and B) can be captured by the model (can we predict $y_{pred}(u_A) < y_{pred}(u_B)$?). In the typical case of multiple models (embodying competing hypotheses), simulations helps figuring out which model should be best.

- *Parameter estimation.* This step concerns model inversion *per se*. The toolbox includes useful tools for diagnostic (*e.g.* to check the quality of fit of the data).

- *Model selection.* This step involves the functions that are necessary to conduct various statistical analysis, including Bayesian model comparison with between-group and between-condition comparisons.

- *Design optimization.* This point refers to design optimization given an objective, be it model selection or parameter estimation (Sanchez et al., 2014).

2.5 Summary

The different Bayesian models for brain function that have been described in this chapter share the common assumptions that the brain would specify a generative model of its environment and would learn, in the Bayesian way, while interacting with it. In other words, the brain would treat every observation that it receives to update its environmental knowledge as a Bayesian system would do. This requires the brain to represent and account for the uncertainty of each information that it manipulates (prior, likelihood and posterior). Bayesian models, the free energy principle and generalized predictive coding in particular, appear very convincing to describe the underpinnings

of perception, and to establish a distinction between perceptual inference and perceptual learning operating at different timescales. A large body of the literature is consistent with these ambitious hypotheses, which appears necessary but not sufficient for their acceptation. As we have seen, evidencing predictive coding in the brain requires relating functional (computational) processes to neurophysiology and promising efforts are gradually made to that aim. Today's advanced methods for meta-Bayesian analysis combined with neuroimaging model-based studies now enable investigating this challenging hypothesis.

Chapter 3

The Mismatch Negativity

The Mismatch Negativity (MMN) is a brain response that is central to cognitive neuroscience and that has accordingly motivated an outstanding piece of literature for almost four decades. As we will see, this interest in MMN derives from its rather ease of recording (notably because it does not require the explicit engagement of the participant's attention) but more importantly, because its investigation could help refining the characterization of the mechanisms underlying perception, including learning processes. Maybe the most significant feature of the MMN is the gap between its essential role in cognitive neuroscience and clinical research (Sussman et al., 2014) - not to mention its numerous applications (Näätänen, 2003; Morlet & Fischer, 2014) - and the fundamental issue of *understanding* the MMN that still remains an open question. By understanding we specifically refer to characterizing its underlying mechanisms as well as clarifying its functional role. This chapter starts with a presentation of the MMN, as well as a brief overview of empirical findings in relation to the current work. Following this, current knowledge about both the neurophysiological and the cognitive processes behind the MMN is summarized. The latter point will concern the different interpretations of the MMN that have been proposed, including its recent predictive coding account exploiting its potential usefulness for testing the Bayesian brain hypothesis.

3.1 Introduction to the MMN

3.1.1 Presentation

The Mismatch Negativity is an electrophysiological response that has been extensively described in the literature since it was first introduced by Näätänen in 1978 (Näätänen et al., 1978). It is elicited whenever a regular stream made of sensory events undergoes a change (for instance, the humming noise of the refrigerator that stops) and is largely assumed to reflect that the brain has detected this change, the so-called *mismatch*. It is evoked experimentally using *standard* stimuli building up the regularity, and *deviant* stimuli interrupting it. Using such experimental material, the MMN could be measured in several modalities (auditory, visual, tactile) (Winkler & Czigler, 2012; Restuccia et al., 2007) with the majority of MMN studies conducted using auditory stimuli. This chapter, as well as the work presented in this thesis, focuses on the auditory MMN.

The typical experimental design employed to measure the MMN is the two-tone *oddball* paradigm (Figure 3.1), which involves sequences of a repeating standard sound, with infrequent deviant

Figure 3.1 – Typical oddball sequence: standard sounds (grey) are being repeated, while deviant sounds (red) occurring infrequently violate the regularity established by the standards.

occurrences. The most intuitive way to design a deviant calls for changes in the physical attributes of the standard sound (like location, duration, intensity or frequency) but may also imply the temporal and statistical regularities present in the acoustic environment. Many deviant types were found to elicit an MMN, with findings contributing to a better understanding of the functional role of the MMN (see §3.2.2).

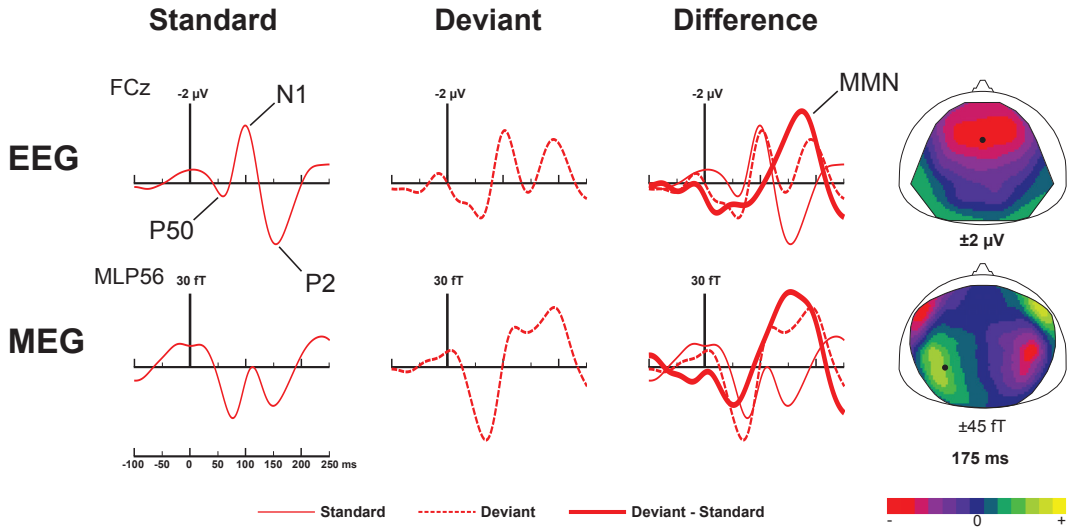

Figure 3.2 – Typical MMN measured with frequency deviants. First row corresponds to nose-referenced EEG responses, with standard (left), deviant (middle) and difference (deviant - standard) evoked responses (right). Bottom row corresponds to MEG signals (radial gradiometers). Right column indicates scalp topographies obtained at the peak latency of the MMN for both modalities (black dots indicate corresponding sensors, the range of values used for the color scale is mentioned for each map). For EEG map, the typical mastoid inversion can be revealed thanks to the nose reference.

In all cases, the MMN is revealed using the difference response obtained by subtracting the standard evoked response to the deviant one. The MMN generally occurs from about 130 ms to 210 ms after deviant onset with the following topographies: using nose-referenced EEG, a negativity at frontal sensors combined with positive inversion at temporal ones; using MEG with radial gradiometers, two anterior and two posterior poles with opposite signs, as represented in Figure 3.2.

3.1.2 Brief overview of auditory evoked potentials

The (auditory) MMN is a component of the auditory evoked responses at the latency range of the Long Latency Responses (LLR). To facilitate the understanding of the MMN in relation to other auditory components, we provide here a brief recall of the main auditory evoked potentials (AEP). These refer to the event-related potentials (ERP) induced by an auditory stimulus, and reflect the

electrical signal generated by this sound through the ascending auditory pathways (Figure 3.5). Auditory evoked responses have been initially described using EEG recordings (hence, potentials) but their magnetic counterpart also contributes to their characterization (notably due to the great sensibility of MEG channels to capture activity within the auditory cortex). In what follows, we present the three typical categories of AEP that are commonly used (Figure 3.3):

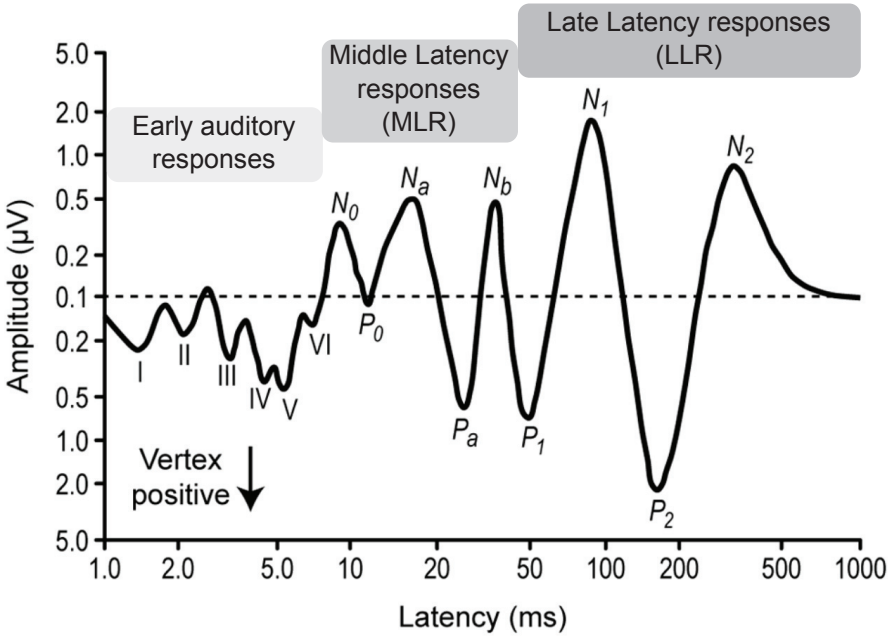

Figure 3.3 – Auditory evoked potentials (adapted from Pérez-González and Malmierca, 2014). Time is represented with a logarithmic scale.

Early Latency Responses. These early responses constitute the first responses to the stimulation and occurs within 10-12 ms after stimulus onset. They consist of seven positive deflections labelled wave I to wave VII, reflecting the auditory information carried out from the auditory nerve, through the cochlea and different relay nuclei upward the thalamus, being the last stop before the auditory cortex.

Middle Latency Responses (MLR). Early latency potentials are followed by the Middle Latency Responses (MLR), which comprises five components, N0, P0, Na, Pa and Nb, as described in (Picton, 1980). Their generators could predominantly be located in the supratemporal cortex (Pantev et al., 1995; Yvert et al., 2001), leading to the MLR being the earliest auditory cortical responses.

Late Latency Responses. Finally, the Late Latency Responses (LLR) occurs from 50 ms to 500 ms after stimulus onset. They comprise the *obligatory* components, namely the P50 (or P1 or Pb), a positive deflection peaking around 50 ms, and the N1-P2 complex entailing a negative component followed by a positive deflection peaking at around 100 ms and 200 ms after sound onset respectively (see response to standards in Figure 3.2). Both have larger amplitude in comparison to earliest components and are predominantly generated in the auditory cortex (Scherg et al., 1989). The LLR also comprise responses elicited by a change in the acoustic environment, such as the MMN but also the N2b (Näätänen, 1992) immediately following the MMN, and the P3a (Escera et al., 2000), a positive deflection peaking around 230 to 350 ms after stimulus onset.

In the particular case of active tasks, where the subjects have to process explicitly the auditory stimuli, additional components such as the N200 (or N2b), the P300 (or P3b) and the processing negativity (PN) are observed.

3.1.3 A comment on refractoriness and neuronal adaptation

This paragraph presents the refractoriness and the neuronal adaptation, two different processes both affecting the responsivity of neurons when a stimulus is being repeated and thereby involved in the auditory processing of oddball sequences.

Put simply, once a neuron has fired a spike, there is a delay (which inverse is called the spiking firing rate) before it fires again. Two terms in particular influence this delay: the *refractoriness* (or neuronal fatigue), a simple physiological mechanism resting on the refractory period of the neuron and the *neuronal adaptation*, relying upon more complex mechanisms explored by a large number of studies using single-cell recordings (for review, Pérez-González & Malmierca, 2014). Neuronal adaptation would derive from synaptic changes emerging when a stimulus is being repeated, that could modify the receptive field of neurons (relative to the stimulus physical features). Contrary to refractoriness, neuronal adaptation presents some sophisticated characteristics like for instance long-term effects (May & Tiitinen, 2010) and can take different forms over the different relays of the auditory hierarchy where it could be measured (Pérez-González & Malmierca, 2014). An illustrative example pertains to the N1, whose amplitude was often shown to be very sensitive to sound repetitions. A refractoriness account of this decrease would imply a rapid stabilization after sound repetitions, whereas neuronal adaptation would lead to a continuous and progressive diminution. Findings in Demarquay et al. (2011) where the maximal decrease of N1 amplitude was found to be attained after 2-3 sound repetitions, could reflect a refractoriness effect.

In their recent review, Pérez-González and Malmierca (2014) provide a comprehensive description of a particular form of neuronal adaptation, the stimulus-specific adaptation (SSA). A neuron is subject to SSA when the two following properties are encountered: *i)* the repetition of a stimulus induces the decrease of its firing rate and *ii)* the neuron fires when a different stimulus is presented (after the repeated-stimulus train). SSA could not be observed at the lowest level of the auditory hierarchy but with neurons in the inferior colliculus, the thalamus and the auditory cortex. Cortical SSA was found to entail larger temporal integration than subcortical ones, suggesting ascending levels of temporal processing. Underlying mechanisms remain unclear but a recent dynamic (deterministic) model was proposed (May et al., 2015), that provided convincing simulations of SSA emerging at the latency of the MMN within the tonotopically-organized auditory cortex in response to oddball-like sequence presentations (SSA could be generated with varying temporal integration windows). In addition to deviance processing, experimental manipulations at the cortical level also suggest a role of SSA in filtering the repetitive (hence non-informative) inputs to avoid sensory overload and recent findings reported in this review support adaptation as a proxy to encode the statistical distribution of stimuli.

Aside from SSA, basic forms of neuronal adaptation and refractoriness are unlikely to account for the MMN (but this hypothesis has been advanced, that will be presented in §3.4.2). They are thus commonly treated as undesirable effect to be minimized. Indeed, when deviants target

different neuronal populations than standards (for instance with frequency deviants due to the tonotopic organization of the auditory pathways), the difference response (deviant-standard) is contaminated by a difference in responsivity between both populations, predominantly at the latency of the MMN due to the refractoriness effect on the N1. Limiting the magnitude of deviance maximizes the chance that standard and deviant neuron populations overlap and thus remains strongly advised. The controlled paradigm proposed in Schröger and Wolff (1996) constitutes another attempt to reduce this unavoidable effect as much as possible.

It should be noted that another confounding effect might enter the difference response, with regards to the physical differences between stimuli (which might entail differences in their respectives evoked responses). This has led some authors to use the term of *genuine* MMN to make a clear distinction between the response triggered by the violation of the regularity and these undesirable effects. Methodological guidelines to optimize the observation of such genuine MMN are provided in Escera et al. (2014).

3.2 Modulation of the MMN by experimental manipulations

The purpose of this section is to describe the major findings about the modulation of the MMN amplitude that remain relevant in the context of the present work.

3.2.1 Key features

The MMN is an automatic response. The MMN elicitation is observed even in the absence of the subject's attention being engaged into the sound sequence. In fact, the MMN could be measured in coma (for review, Morlet & Fischer, 2014), during sleep (Ruby et al., 2008), under anesthesia (for review, Chennu & Bekinschtein, 2012). Other evidence supporting this fundamental property of the MMN have been reviewed in Näätänen et al. (2010). Hence, likewise a reflexive behavior, it seems that the MMN cannot be refrained, but its amplitude could be modulated by attention according to several studies reviewed in Fishman et al. (2014).

The MMN expresses very rapidly. Many studies have reported the MMN to be visible as soon as two standards were delivered, with the rule governing the sound sequence involving for instance a simple sound repetition (Sams et al., 1984) or specific transition probabilities (Bendixen et al., 2007). It should be noted that a single presentation is not sufficient to elicit the MMN (as recalled in Winkler et al., 1996), that the first standard following a deviant also generates a (small) MMN (Sams et al., 1984) and that repeated deviants exhibit a reduced MMN (Sams et al., 1984; Morlet et al., 2014).

The MMN does not habituate. Contrary to the N2b and P3a, the MMN amplitude has been shown to be invariant to the time exposure to the oddball sequence (Morlet et al., 2014).

3.2.2 The broad spectrum of deviance types

First study reporting a MMN (Näätänen et al., 1978) employed a duration deviance. Since then, MMN has been successfully measured with deviant stimuli embedding change relative to the standards that can be categorized as follows (for comprehensive reviews, see Näätänen et al., 2010; Fishman, 2014):

- physical changes: deviants differ from standards relative to the physical properties of sounds (such as frequency, duration, location), also comprising higher-level features such as phonetic properties in speech sounds. Ensuing MMN are found to be larger as the deviance magnitude increases.

- contextual changes: this time the physical content of both standards and deviants are the same but deviant occurrence induces a change in the temporal property of the sound sequence (with for instance, the time-interval duration between two sounds) or involves the statistical dependencies of sounds within the sequence. In this case, researchers often refer to *global* or *abstract* rules, that may rest on large time-scale (to induce a specific periodicity like in Alain et al., 1994; Bekinschtein et al., 2009), or to *local* time-scale in the case of specific transition probabilities or contingency rules (Bendixen et al., 2008; Todd & Mullens, 2011).

- omission: this is a specific case of deviance, where the deviant violating the rule is a standard not being delivered. This case of deviance is of particular interest as no auditory input is to be processed.

In addition, it is worthy to mention the numerous MMN findings observed using physical and contextual changes during high-level cognitive processes such as language and music processing (also reported in the above cited reviews).

Overall, these findings illustrate the large extent to which the MMN allows investigating perceptual processes in passive situation or without the explicit engagement of subject's attention towards the sound sequence. It is important to underscore the (trivial) statement that deviance can not be evoked without preceding standards to set up the regular acoustic environment. The range of rules (built up with standards) that were associated to an MMN when violated with appropriate deviants therefore suggest that the MMN is *"not so primitive"* (Todd et al., 2013), particularly considering its automatic nature. Its function is now accepted to be beyond the detection of an environmental change and to involve neural processes of regularity extraction. As will exposed in section §3.4, all these findings have contributed to clarify the (still unclear) role of the MMN.

3.3 Neurophysiological underpinnings of the MMN

In the present day, the mechanisms underlying the generation of MMN observed with electrophysiological recordings are not fully understood. These mechanisms are triggered by deviant occurrence and operate in a *"neurophysiological context"* established by preceding standards. They obviously involve auditory processing along the auditory pathways. The first step forward a better characterization of deviance processing concerns the identification of the related cortical

regions, for which a review of existing findings is presented below. We then describe preliminary attempts to assess the dynamics of cortical connectivity at play during this processing. Finally, we present the hypothesis of an auditory hierarchy spanning from sub-cortical structures to high-level cortical regions, proposed by Escera and Malmierca (2014).

3.3.1 Sources of the MMN

Equivalent current dipole (ECD) studies. The majority of the numerous studies dedicated to the identification of MMN generators aimed at characterizing a fronto-temporal network for which strong priors were provided from Current Source Density (CSD) studies (Giard et al., 1990; Deouell et al., 1998) and from lesioned-patient studies (Alain et al., 1998). Initial dipole modeling studies (reviewed in Alho, 1995) contributed to describe ECDs (location and orientation) within bilateral supratemporal planes (Scherg et al., 1989; Sams et al., 1991) that were different for different deviance features (Giard et al., 1995; Levänen et al., 1996). Altogether, these studies point to the spatial and temporal complexity of auditory processing achieved in temporal regions. Interestingly, a few recent EEG studies succeeded in modeling frontal generators (Schairer et al., 2001; Jemel et al., 2002; Rissling et al., 2014; MacLean et al., 2015).

Neuroimaging studies (fMRI, TEP). In the 2000's, a large number of MMN studies were conducted using neuroimaging techniques (fMRI, TEP, Optical Imaging) as they provide functional maps of brain activity with higher spatial (but lower temporal) resolutions than EEG and MEG data. Major findings of deviance-related activity were reviewed in (Deouell, 2007), with quasi-systematic implication of the Superior Temporal Gyrus (STG) and several reports of contributions of the Inferior Frontal Gyrus (IFG). Other frontal areas located in the Middle Frontal Gyrus (MFG), the Superior Frontal Gyrus (SFG) and the Anterior Cingulate Cortex (ACC), as well as parietal clusters (Molholm et al., 2005; Gomot et al., 2006) could be reported. Recently, Alho (2014) meta-analyzed fMRI studies showing activities within the auditory cortex and found median location (from 18 studies) of specific pitch-change processing bilaterally in the mid-STG close to the Planum Temporale (PT). Using duration deviance, a refined description of temporal activity, dissociating the STG, Heschl's Gyrus (HG) and PT could be obtained in a study combining fMRI and EEG recordings (Schönwiesner et al., 2007). More recently, temporal (HG and STG) and subcortical structures (Inferior Colliculus and Medial Geniculate Body) - but no frontal regions - were observed with a frequency deviance paradigm using fMRI (Cacciaglia et al., 2015). Critically, as already pointed by Deouell (2007), it remains difficult to summarize these numerous findings resulting from a large variety of experimental setups, each attempting to circumvent the inadequacy of metabolic techniques to measure genuine deviance response.

Distributed source studies (EEG, MEG). The 2000's also witnessed remarkable advances in distributed source modeling with electrophysiological data that now provides an adequate degree of spatial resolution. We briefly review below the main findings reported with EEG and MEG:

- Using EEG, clusters within supratemporal planes were measured (Waberski et al., 2001; Marco-Pallarés et al., 2005), including HG and PT (Fulham et al., 2014). MTG was also reported (Marco-Pallarés et al., 2005; Fulham et al., 2014). Frontal contributions were localized in ACC (Waberski et al., 2001; Marco-Pallarés et al., 2005) and in bilateral IFG (Fulham et al., 2014; Hanna, 2014) with the former study also showing MFG activity.

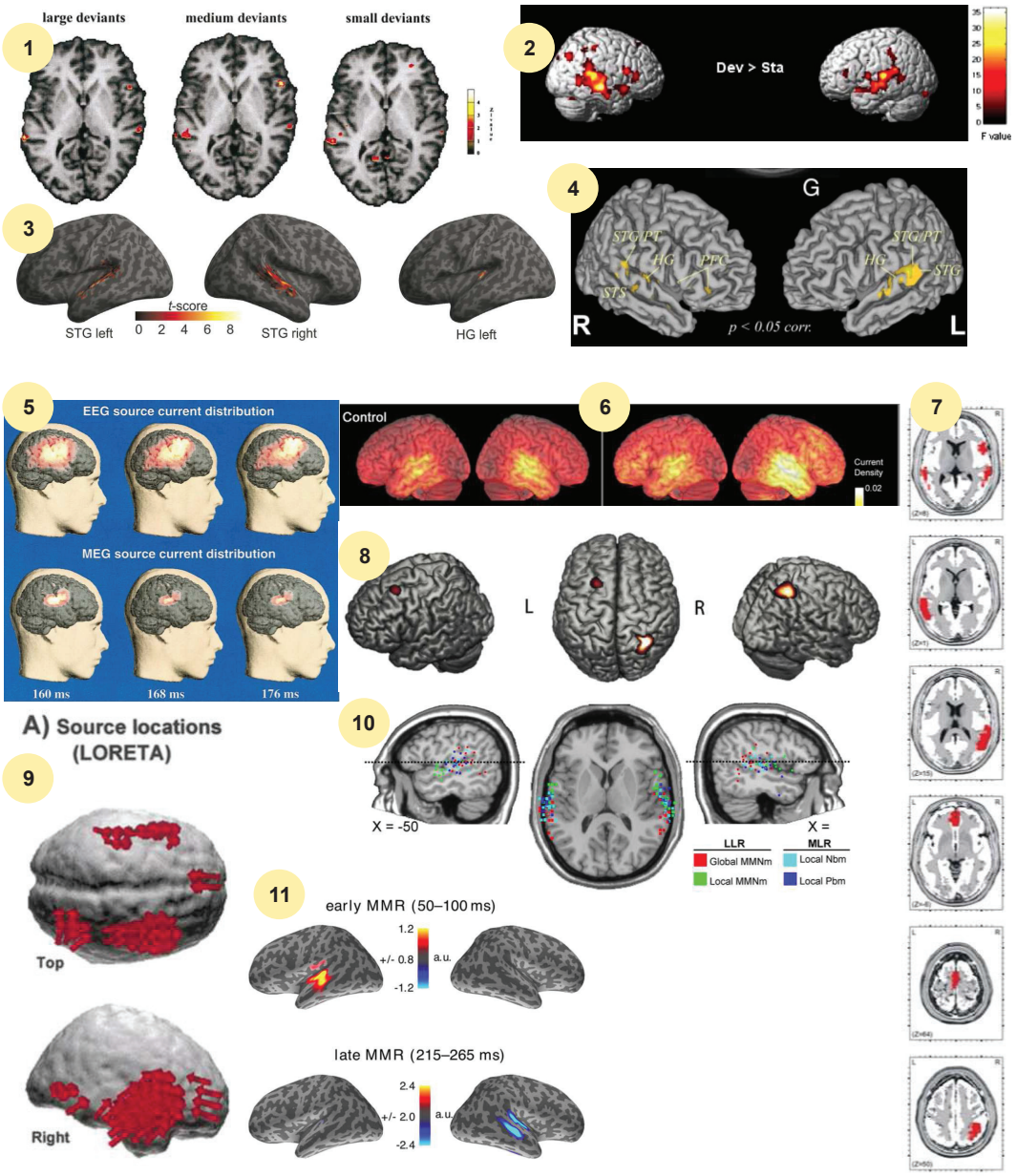

Figure 3.4 – Overview of experimental findings regarding MMN generators. From 1 to 4 : fMRI studies with 1=Opitz et al., 2002; 2=Gomot et al., 2006 (children findings); 3=Cacciaglia et al., 2015; 4=Schŏnwiesner et al., 2007). From 5 to 11: Distributed sources using EEG or MEG with 5= Rinne et al., 2000; 6=Fulham et al., 2014; 7=Marco-Pallares et al., 2005; 8=Chakalov et al., 2014; 9=Waberski et al., 2001; 10=Recasens et al., 2014; 11=Ruhnau et al., 2013. We refer the reader to these articles for more details.

- Using MEG, temporal activations were reconstructed in right STG (Lappe et al., 2013a; Paraskevopoulos et al., 2014; Recasens, Grimm, Capilla, et al., 2014), left STG (Lappe et al., 2013b) or bilateral STG (Ruhnau et al., 2013; Recasens, Grimm, Wollbrink, et al., 2014) with both studies also showing clusters in HG. Clusters in right MTG, PT and HG (Recasens, Grimm, Capilla, et al., 2014), right insula (Lappe et al., 2013b) and bilateral STS (Ruhnau et al., 2013) were also reported. Frontal activations were found in right SFG (Paraskevopoulos et al., 2014), left MFG (Paraskevopoulos et al., 2014; Chakalov et al., 2014) and bilateral IFG and SFG (Lappe et al., 2013a). Parietal sources were also found (Lappe et al., 2013b; Chakalov et al., 2014).

This review confirms that both EEG and MEG signals contain sufficient information that can be captured with distributed methods to disentangle frontal, temporal and even parietal participations to the generation of deviance responses. However, it also reveals a critical lack of robustness in the characterization of the cortical network recruited during deviance processing. Not only this issue may result from differences in experimental designs (including different physical properties of stimuli) but also from the variability of methods used for data preprocessing, event-related potentials/fields (ERP/ERF) computations (including the selection of standard trials entering the difference deviant-standard) and of course source reconstruction (including forward modeling). A common point shared by the majority of the EEG, MEG and fMRI studies cited here consists in the fact that their respective experimental design entails a modulation of oddball sequences by a factor of interest (most often deviance magnitude or deviance feature is used). Differences in the sources of the resulting MMNs thus constitute a proxy to infer how deviance processing (and more generally auditory processing) is achieved over the corresponding brain areas. For instance, several fMRI studies manipulated different ranges of deviance magnitude to investigate where in the brain ensuing modulation could be measured, that could to possibly clarify the respective role of frontal and temporal regions (Opitz et al., 2002; Rinne et al., 2005; Schönwiesner et al., 2007).

3.3.2 Characterization of cortical dynamics (premises)

Recent modeling methods such as DCM now allow characterizing finely the dynamics of cortical interactions during deviance processing. However, spatio-temporal results obtained before such methods became available appear to be very informative with regard to this connectivity issue.

Initial ECD studies investigated the N1 and MMN generators within the supratemporal plane (based on standard and difference responses respectively) with the aim to disentangle the neuronal adaptation of N1 and *genuine* deviance processing (Scherg et al., 1989; Sams et al., 1991). Their findings indicated the MMN dipole being anterior to the N1 dipole, and were confirmed by several subsequent source reconstruction studies (Näätänen & Alho, 1995; Rosburg et al., 2004; Recasens, Grimm, Capilla, et al., 2014) (but see Maess et al., 2007), as well as lesion-based studies (reviewed in Alho, 1995). Despite the acknowledged limitation of ECD approach (imposing the number of activated sources), these results could suggest auditory processing resting on a spatial progression within the supratemporal plane, from posterior (at N1 latency) to anterior (at MMN latency).

In the same vein, findings from studies that investigated the relative role of the frontal and temporal generators of the MMN provide some premises that could help refining our understanding about the fronto-temporal connectivity. Actually, several findings reported a temporal activation occurring before the frontal one (Rinne et al., 2000; Waberski et al., 2001; Opitz et al., 2002; Doeller et al., 2003; Rinne et al., 2005; Fulham et al., 2014) but the reverse hypothesis (frontal regions being activated before temporal ones at the latency of the MMN) was also introduced by Gomot et al.(2000), with subsequent findings in line with this view (Yago et al., 2001; Tse & Penney, 2008; Lappe et al., 2013a; Tse et al., 2013).

These initial findings, based on estimated source activity, thus appear useful but somehow limited in the sense that the different methodological approaches employed were not equipped to assess

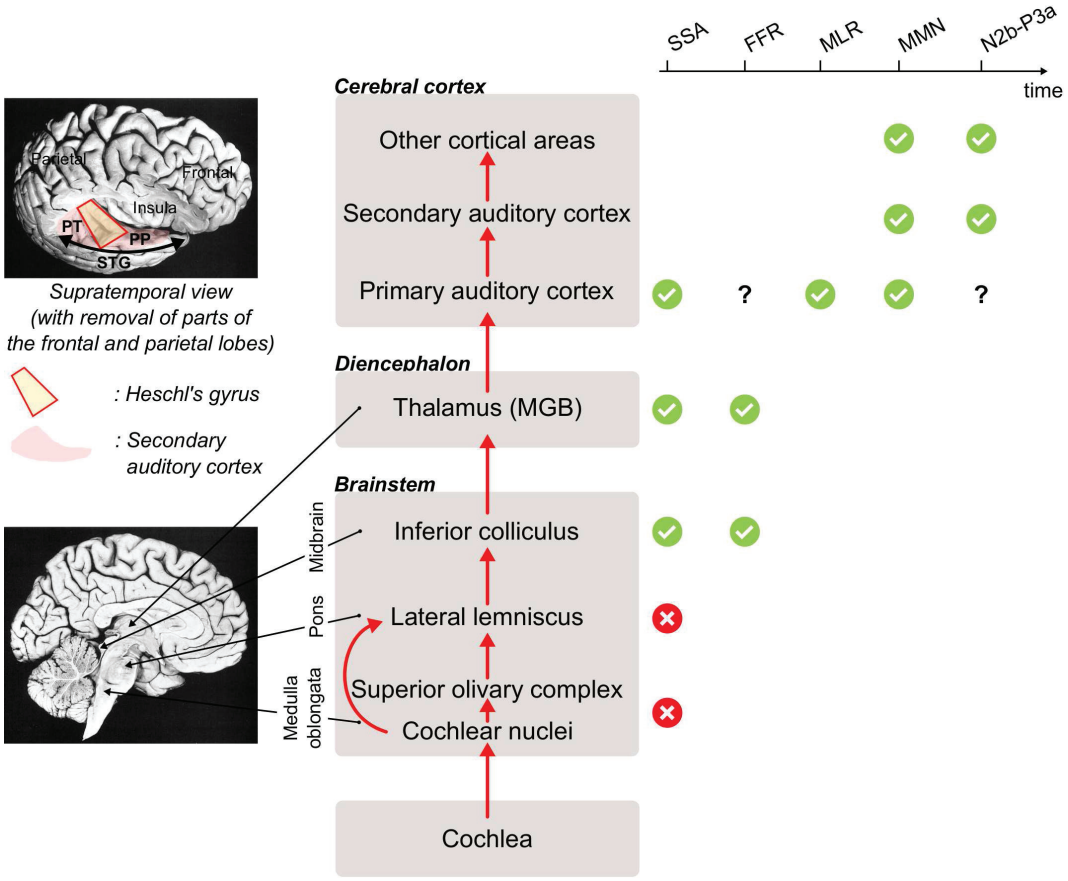

Figure 3.5 – Auditory hierarchy with associated deviance related findings. The neurobiological description of the auditory hierarchy is adapted from A.Caclin (thesis); only the ascendant connections have been depicted (red); MGB=Medial Geniculate Body, PT=Planum Temporale, PP=Planum Polare, STG=Superior Temporal Gyrus. For each level of the hierarchy, green ticks indicate that empirical support for deviance processing could be provided, red crosses indicate an absence of evidence and interrogation points suggest ambiguous findings. The absence of symbol correspond to an absence of findings (at the present day).

the cortical connectivity *per se* (they prevent from estimating the feedback influence of targeted sources for instance). Such analysis can precisely be achieved with DCM. Using such models, Garrido et al.(2007) did not test the hypothesis of a frontal-to-temporal dynamics, but using a temporal-to-frontal model, they succeeded in revealing that feedback connections were needed to fit deviance response from the latency of around 220 ms. Because DCM of the MMN fall under a predictive coding account of deviance processing, we will present them in details in chapter 4, §4.2.3.

3.3.3 Auditory hierarchy for deviance processing

In addition to the numerous (human and animal) studies that have reported MMN, other deviance-related responses could be measured at different latencies and at different levels of the auditory hierarchy that were reviewed in (Escera & Malmierca, 2014). Corresponding findings are listed below (see also Figure 3.5):

- *Stimulus-specific adaptation (SSA).* First evidence for neurons showing SSA were reported in the seminal study of Ulanovsky et al. (2003), conducted with single-unit recordings in the cat's primary auditory cortex. Subsequent (numerous) studies also indicated evidence

for SSA in the inferior colliculus (IC) and the thalamus (recorded with rat, guinea pig or monkey) but investigation in the higher-cortical levels appears to be lacking. SSA can take various forms (depending on experimental manipulation), with the more sophisticated one located in the primary auditory cortex (Pérez-González & Malmierca, 2014).

- *Frequency following responses (FFR).* These evoked responses measured with scalp electrodes are part of the auditory brainstem responses (ABR) and are assumed to reflect activity within the brainstem (where the waves IV and V of the early latency responses originate). A deviance-related modulation of the FFR could be measured in recent studies using speech stimuli (Slabu et al., 2012) and amplitude-modulated sounds (Shiga et al., 2015). The brainstem origin of such responses could be challenged by recent findings supporting a cortical contribution (Bellier et al., 2014; Coffey et al., 2016), but the implication of the IC and the medial geniculate body of the thalamus (MGB) in deviance processing was confirmed by a fMRI study (Cacciaglia et al., 2015).

- *Middle latency responses (MLR).* In a recent series of studies (main findings described in Escera et al., 2014), different components of the MLR (Na, Pa, Nb, Pb) were found to be modulated by the occurrence of deviants using physical changes such as frequency, intensity and location. Deviant responses were found to exhibit either a decrease or an increase[1] of peak amplitude (in comparison to standard response). A cortical contribution from the primary auditory cortex could be found using source reconstruction. It should be noted that only a few studies addressed the localization of these deviance responses. The works by Recasens and collaborators (Recasens, Grimm, Capilla, et al., 2014; Recasens, Grimm, Wollbrink, et al., 2014) suggested a temporal activation centered in HG extending to the medial part of the STG. In Ruhnau et al. (2012), generators of early mismatch response (from 50 ms to 100 ms after deviant onset) were identified in bilateral STG and STS, including the primary auditory cortex (PAC).

- *Late latency responses (LLR).* Beyond the MMN, deviant tones can also elicit the N2b-P3a complex (Morlet et al., 2014). The characterization of their cortical generators remain unclear, notably with regards to the auditory cortex contribution. Higher-level sources, in the anterior cingulate cortex (ACC) for instance could be evidenced (Crottaz-Herbette & Menon, 2006).

In their review, Escera and Malmierca proposed that MLR deviance responses could bridge the gap between the SSA, occurring at the spike time-scale and measured with animal recordings, and the MMN, peaking from around 130 ms and which could entail more sophisticated perceptual processing. Beside, recent studies reported that MLR modulation by deviants was not exhibited in the particular case of *complex* rules (resting on larger statistical dependencies than the simple repetition pattern of typical oddball sequences) (Cornella et al., 2015; Althen et al., 2013; Recasens, Grimm, Wollbrink, et al., 2014). Based on these findings, Escera and collaborators support the view that deviance processing, relying upon change detection and regularity extraction mechanisms, would be grounded in the auditory hierarchy, where the neurobiological levels of this

[1] One should be aware that an increase (or decrease) of brain responses measured for deviants with EEG or MEG does not necessarily implies a larger (or smaller) brain activity related to deviant processing (due to the non-linear biophysical mapping of neuronal activity to external sensors). Such increase (or decrease) just informs about a different underlying processing.

hierarchy could be related to the *"ascending levels of complexity"* involved in deviance processing.

To sum up, despite an extensive literature, the neurophysiological description of the MMN (including its contributing generators) still remains an unresolved issue, and thereby continues to be investigated. All the studies cited here suggest that a finer characterization could be achieved by no longer considering this component in isolation from the other deviance-related responses generated along the auditory hierarchy. The next (challenging) step pertains to the description of the connectivity within this hierarchical structure to go beyond speculative interpretations, and relies upon advanced methods such as DCM.

3.4 What functional role for the MMN?

So far, what we know about the MMN from a cognitive point of view, is that it is elicited by a change in the sensory environment that has been *detected* by the auditory system. All MMN studies investigating its generators or its modulation by experimental factors, have contributed to give insight into its cognitive function. Several mechanisms have been proposed to explain the MMN, and we present below the dominant models with the latest being predictive coding. As we will see, predictive coding appears as the only model (in the present day) that provides a detailed mechanistic description of the MMN while reconciling other (exclusive) existing models.

3.4.1 The sensory memory account

The (early) sensory memory model[1] of the MMN has emerged from the observation that a minimum of two standard repetitions was needed to elicit a MMN, hence suggesting the idea of a *sensory memory* that would allow the standard representation to be stored. This view, described in Näätänen et al. (2005), assumes that exposure to repetitive standards generates a *memory trace* of the standards and that the MMN reflects the outcome of a comparison between this trace and the incoming sound. Specific neuronal populations would thus be in charge of computing this comparison, at the core of change detection.

The large number of studies that have explored the generators of the MMN (in particular the early ECD studies) attempted to reveal these neuronal populations, as well as clarifying their (still debated) feature specificity at both the temporal and frontal levels. A commonly shared view is that temporal regions could entail the *pre-attentive* change detection with comparison output conveyed to frontal regions in order to trigger, if necessary, the orientation of attention (as observed for novel - or salient or alerting - deviant stimuli) (Giard et al., 1990). The latter aspect is often referred to as the *involuntary switch of attention* preceding the *orienting* responses (N2b, P3a).

As it became clearer that the MMN could index the detection of contextual changes, the sensory memory model was refined to extend the mechanisms behind this component: in addition to change detection, the MMN would also rest on rule extraction mechanisms. This variant, sometimes referred to as the *model adjustment hypothesis* (Garrido, Kilner, Stephan, & Friston, 2009), considers deviants violating the extracted rule instead of mismatching the memory trace. The differences between these two versions of the memory-based model are detailed in Winkler et al.

[1] also referred to as *memory-mismatch hypothesis*, or *memory-trace explanation*

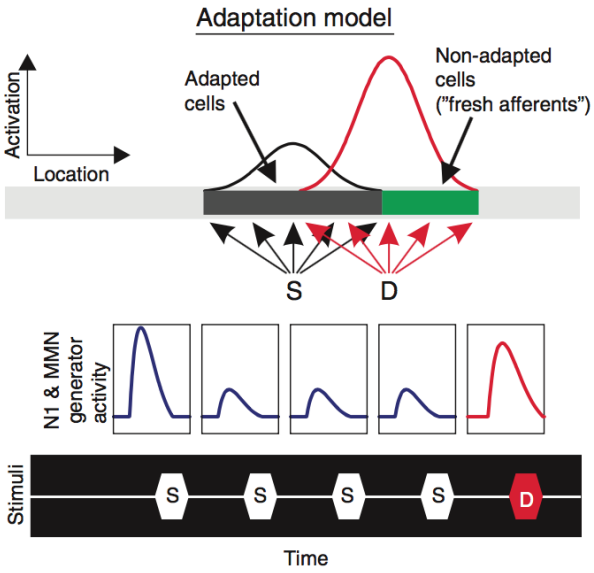

Figure 3.6 – The adaptation model proposed by May and collaborators (adapted from May et al. (2010).

(2007, see their Table 1, p.157).

Such memory-based model strongly opposes to the adaptation model (see below) with the MMN elicited by abstract rules or omitted standards being the most convincing empirical evidence in favor of memory-based models (Näätänen et al., 2005).

3.4.2 The adaptation model

The adaptation model proposed by May and colleagues (Jääskeläinen et al., 2004; May & Tiitinen, 2010) challenges the underlying cognitive mechanisms of the MMN. According to these authors, the MMN is not a brain component *per se*, following the N1 and having its own generators, but it rather derives from a differential adaptation affecting the standard and the deviant N1. In other words, the MMN would only exist by the subtractive effect on the N1 responses. Importantly, in May and Tiitinen (2010) the adaptation model is also termed as fresh-afferents model, and adaptation is defined as refractoriness - *"the differential responding to the standard and deviant is due to adaptation (refractoriness) of the neuronal population"* - or more evasively as *"the short-term modification of the responsiveness of neurons by sensory stimulation"*. The key feature of this model is that it requires only two neuronal populations: one for the standards and one for the deviants (see Figure 3.6). Using simulations, May and Tiitinen (2010) demonstrated how adaptation and lateral inhibition can account for the difference in N1 deviant and standard responses leading to the (delayed) MMN.

In a recent review, Fishman (2014) reported empirical findings supporting the adaptation model. A large piece of the argumentation is based on inconsistent predictions that the sensory memory and the adaptation models would generate in the case of $deviant_{oddball}$ stimuli, embedded in a typical oddball sequence, and $deviant_{silenced}$ embedded in the same sequence but with standards being silenced. According to Fishman, memory-based model would predict $deviant_{oddball} > deviant_{silenced}$ because of the absence of comparison outcome in the silenced condition (but see

Alcaini et al., 1994), whereas the adaptation would predict $deviant_{oddball} \leq deviant_{silenced}$ because $deviant_{silenced}$ would have equal or eventually more fresh afferent activity. Several human and animal findings supporting these adaptation predictions are cited. Besides, supporting evidence would also result from the unresolved issue of identifying cortical generators being activated for deviants only, that would be the signature of the sensory memory model. However, the author recognized the difficulty of the adaptation model to account for the MMN observed with abstract rules, unless considering the (unlikely) differential adaptation over neurons encoding specific abstract features.

Critically, the type of adaptation involved in such model of the MMN differs from the more sophisticated SSA that could play a key role in the temporal integration of sensory information (a mechanism also likely to be behind the MMN). Interestingly, the model proposed by May et al. (2015), based on a time-varying synaptic connectivity between cortical columns within the supratemporal plane, could provide successful simulations of such columns eliciting SSA. Precisely, connection strength between two columns was set to depend on the synaptic adaptation (neuronal fatigue) of the column targeting the other one. Synaptic adaptation rested on a specific time constant whose tuning could generate SSA with unexpected sounds violating a typical frequency oddball sequence but also with a within-pair frequency change sequence. This model thus demonstrates how specific cortical columns can be activated only when a deviant pattern occurs (a property required for SSA), and highlights the role of varying temporal integration windows to endow the auditory system with a change detection mechanism. However, this model entails several limitations including notably the lack of result generalizability for non-frequency based rules. Crucially, the (essential) tuning of time constants, operated by the experimenter for the simulations reported in the study, has to adapt to the temporal structure of sound sequence to induce deviance responses, an issue not accounted for by the model.

3.4.3 The predictive coding model

The predictive coding view of the MMN, with one of its first description to be found in Friston (2005), assumes the ability of the brain to represent the generative model of sound sequence (namely the statistical dependencies that govern the relationship within sequence items) as it is exposed to its acoustic environment. Each sound is treated as a new observation by the brain and as such, induces updates of predictions and precision-weighted prediction errors along the auditory hierarchy (Figure 2.3). Hence, *expected* standards contribute to enhance the precision of predictions and to reduce prediction errors within state units, whereas *unexpected* deviants trigger prediction errors within error units propagating through the hierarchy. Under this view, the MMN would thus reflect these precision-weighted prediction errors and their associated updates.

Some authors have drawn a close parallel with the (early) sensory memory model, as it also involves a comparator between experienced and incoming sounds. *Memory trace* could thus be equated to model *predictions* and *mismatch* (or comparison outcome) to *prediction error* (Winkler & Czigler, 2012). In Fishman (2014), these two models are even explicitly presented as being equivalent. This parallel appears somehow limited due to the absence of mechanistic assumptions in the memory-based model (describing the MMN conceptually). In fact, the predictive coding account of the MMN reconciles the mutually exclusive memory-based and adaptation models.

From a theoretical perspective, both models contradict each other regarding the existence of specific deviance populations, since the adaptation model refutes any deviance processing. Under predictive coding, such deviance populations exist (the error units) but *belong* explicitly to the auditory system. As already mentioned, such error units could even take place within the cortical layers of each hierarchical levels. Using DCM, empirical evidence for such reconciliation could be provided (Garrido, Kilner, Kiebel, & Friston, 2009). Another argument of predictive coding accounting for both model predictions pertains to the $deviant_{oddball}$ and $deviant_{silenced}$ findings reviewed in (Fishman, 2014), with $deviant_{silenced}$ interrupting the short timescale rule of silence, hence triggering prediction errors.

Numerous findings of MMN modulation by experimental factors can be successfully interpreted under the predictive coding hypothesis (Garrido, Kilner, Stephan, & Friston, 2009; Winkler & Czigler, 2012). A persuasive illustration is given by the widely cited increase of MMN amplitude measured with larger deviance magnitude: the predictive coding here accounts for larger prediction errors, hence larger MMN. The fact that the MMN was found to not habituate after long exposure to the sound sequence (Morlet et al., 2014) could at first sight disproves this hypothesis. However, it could on the contrary points to the fact predictive coding entails several prediction errors generated at different levels of the hierarchy, with ascending timescale processing (Kiebel et al., 2008). Hence, lower levels having a short temporal integration window would always predict a standard.

As already mentioned in Chapter 2, §2.3.3, one of the strength of predictive coding concerns the novel (mechanistic) hypothesis that it raises. Applied to the MMN in particular, predictive coding allows investigating the processing of repetitive standards, an issue that had never be addressed finely under the sensory memory model (the so-called memory-trace formation). The succession of standards induces the (perceptual) learning of the regularities and should accordingly lead to reduced prediction errors and to larger precision of predictions. In Auksztulewicz et al. (2016), findings that addressed the *"dual role of descending predictions"*, namely the suppression of prediction errors and the changes in prediction precision, are reviewed in the context of repetition suppression (RS), an effect introduced in chapter 2, §2.3.3. One plausible candidate for its underlying mechanism could be adaptation, with for instance SSA being its correlate in the primary auditory cortex. Todorovic and de Lange (2012) aimed at dissociating RS from the expectation suppression (ES), another cause of reduced activity (inherited from predictive coding) that could be evidenced in Summerfield et al. (2008). Using an orthogonal design in MEG, they measured main effects of RS and ES at the latency of the P50 and around 100 ms respectively. They concluded in favor of a cascade of updates revealed by these separate activities, where RS would correspond to local regularity treated at a low level and ES to larger timescale processing in a higher level. In the same vein, other attempts to isolate spatially and temporally different ERP components observed after repeating standards have been made (Recasens et al., 2015), that could suggest hierarchical activities at play during perceptual inference.

3.5 Summary

This chapter aimed at presenting the (auditory) MMN, an auditory evoked response essential to cognitive neuroscience and clinical research that remains poorly understood. Neurophysiological

findings reported here support the hypothesis of deviance processing achieved within the auditory hierarchy. This thereby calls for a better characterization of the dynamic interactions between MMN generators that would give insight (new hypothesis) into the functional processes involved at each level of the hierarchy. Several psychological models of the MMN have been proposed for years that have not yet succeeded in fully describing this component. Recent predictive coding account appears very much relevant to investigate its role within auditory information processing. Remarkably, this Bayesian framework provides a plausible account for numerous MMN experimental findings and appears reconciling inconsistent former models. Importantly, this mechanistic framework allows describing the MMN at both the neurophysiological and psychological levels. More specifically, findings from each perspective could reveal the importance to consider all deviance responses as a whole within the hierarchical auditory system. Hence, two timescales of analysis are required for a comprehensive characterization of deviance processing: the *ERP timescale* of perceptual inference, to clarify the role of the different components with regard to the predictive coding message-passing scheme along the auditory hierarchy, and the *experimental timescale* of perceptual learning, to assess the learning of regularities emerging from successive standards and deviant processing. As was explained in previous chapter, evidencing predictive coding in the brain remains a great challenge and the relevance of oddball sequence processing to attain that aim was described here. Recent advances in Bayesian modeling have just started exploring such hypothesis.

Chapter 4

Advanced Bayesian modeling of mismatch responses

This final introduction chapter aims at presenting recent advanced Bayesian modeling tools that now allow testing in a principled fashion the hypothesis of predictive coding for deviance processing. Predictive coding here refers specifically to the framework proposed by Friston (2005), namely *generalized predictive coding*, with the minimization of the free energy being the mechanism for the suppression of prediction errors. Within this framework, neural activity reflects the dynamics of Bayesian information processing at different levels, which could presumably contribute to the (observable) evoked responses. We start with a brief recall of predictive coding predictions in the context of oddball sequences. Next, we present two approaches that allow investigating empirical Bayesian inference within the auditory hierarchy: the neurobiologically informed dynamic causal model (DCM) and the computational learning model. Finally, we report the very few recent studies which have proposed to combine these two (neurophysiological and functional) perspectives.

4.1 Expected physiological and functional dynamics under predictive coding

In this section, we summarize the predictions regarding standard and deviant processing that predictive coding entails at both the functional and physiological levels. These predictions constitute a guidance for subsequent modeling.

Standard tone repetitions. From the observation of successive standards, the brain learns the generative model of sound sequences (for instance, the sequence has a two-tone structure and tone category, rare or frequent, follows a Bernoulli distribution with specific parameters to be inferred). As the number of repetition increases, predictions elaborated under this model becomes more and more precise, hence giving lesser importance to new sensory inputs (Figure 1.1). From a physiological perspective, standard tone repetitions yield reduced activity (possibly mediated by synaptic adaptation) and reduced bottom-up and top-down cortical interactions (Friston, 2005; Auksztulewicz & Friston, 2016).

Rule violation by deviants. Unexpected deviants induce prediction errors which trigger message passing along the auditory hierarchy until all errors have been explained away. The amplitude of prediction errors varies according to the deviance magnitude but is furthermore weighted by its precision, *i.e.* the relative precision of the observations and the predictions. From a physiological perspective, deviants should cause enhanced forward connectivity (in comparison to standard processing) but also enhanced backward connectivity as well as within-level activity (Friston, 2005).

For both tone types, the dynamics of functional and physiological updates elicited after stimulus delivery can be characterized within trial at the ERP timescale (successive ERP components should reflect the chronology of updates with possibly the latter indexing higher level activity) but also over trials at the experimental timescale (sound processing should be modulated through learning, hence depending on the history of perceived tones). Furthermore, higher levels in the auditory hierarchy are expected to have larger temporal integration windows than lower ones (Kiebel et al., 2008).

4.2 Dynamic causal modeling (DCM)

This section describes the method of Dynamic Causal Modeling (DCM) that we used to model EEG and MEG mismatch responses. We start by a short introduction of the method that applies to several types of neuroimaging data, and then provide a detailed description of its main features in the case of electrophysiological evoked responses. Finally, we overview the insights that DCM has provided into the specific context of auditory deviance responses.

4.2.1 General presentation

DCM allows describing the dynamics of cortical interactions between different regions composing a hierarchy in response to a stimulation and generating the corresponding neuroimaging data (Figure 4.1). These interactions are defined in terms of *effective connectivity*, accounting for the causal (directed) influence that a brain region exerts on another[1]. DCM represents the brain as a dynamical system receiving inputs (the stimuli, u), treating them using a specific evolution function f (resting on hidden states x and parameters θ) and producing observable outputs (the measurable brain activity y) according to a specific observation function g (also resting on hidden states x and parameters ψ). A dynamic causal model is thus a generative spatio-temporal model of observable brain activity embedding biophysical assumptions about the cortical implementation of information processing. Importantly, DCM can be used to test how the cortical coupling within the hierarchy is modulated by experimental manipulation, which is precisely needed to better understand how standard and deviant processing differ along the auditory hierarchy (at least, the cortical part of this hierarchy).

Practically speaking, conducting a DCM analysis starts with specifying one or several generative models accounting for specific brain responses. Bayesian inference then provides the (approximate) log-evidence for each alternative hypothesis, which is first used to compare models or families of models. Finally, the winning model or model family is used to make inference on

[1]Effective connectivity differs from functional connectivity: the latter provides a description of the *undirected* statistical relations between two signals.

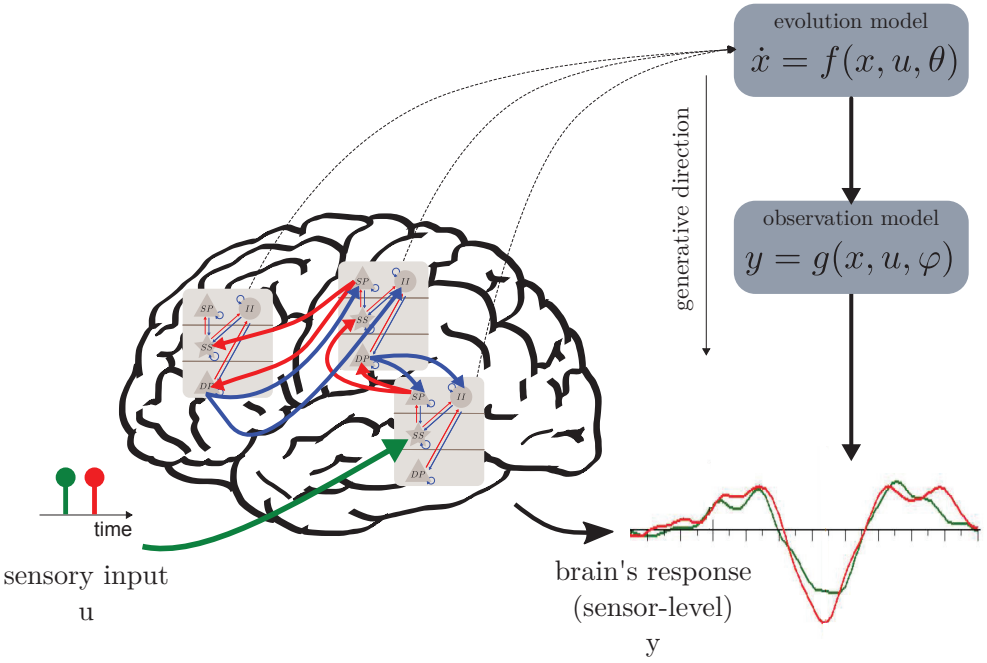

Figure 4.1 – Schematic view of a DCM, a generative model of neuroimaging data (depicted here in the specific case of electrophysiological data). The brain is modeled as a cortical network that is perturbed by some inputs (green arrow) which are processed along the hierarchical organization made of the interconnected nodes or sources (light brown rectangles). Extrinsic connections are connections between sources, they are forward (red arrows) or backward (blue arrows) connections. Intrinsic connections appeal to the internal circuitry within each source, for which an example is given here (based on the CMC model used in this work, see also Figure 4.2). The dynamics of hidden states x in response to input u are described using an evolution model f and can be mapped onto sensors using an observation model g to generate the data y.

(hidden) states and parameters characterizing the effective connectivity. DCM has been originally proposed for fMRI data (Friston et al., 2003) and was then extended to model electrophysiological responses (David et al., 2006; Kiebel et al., 2006). In what follows, we focus on DCM for EEG and MEG evoked responses.

4.2.2 DCM for EEG and MEG evoked responses

In this particular case, DCM generates the evoked responses that are modeled as the output of a dynamical system (the cortical brain) in response to an experimental input. There is a crucial distinction to be made between the *extrinsic* connectivity that refers to the coupling between the different sources composing the DCM architecture, and the *intrinsic* connectivity at play between the different neuronal populations composing a source. Each connection (extrinsic or intrinsic) is defined by its origin and termination but also by its strength, a time-independent variable which quantifies the influence that the *source* population exerts on the *target* population. The different features composing DCM for EEG and MEG evoked responses are detailed below.

Neuronal microcircuit. Each source (or cortical area) composing a DCM comprises interconnected neuronal populations defined following the laminar structure of the cortex, and can thereby be assimilated to a cortical macrocolumn. Several models describing the local circuitry within a source (and associated neural activity) have been proposed (for a review, see Moran, Pinotsis, & Friston, 2013), each embedding approximations informed by animal findings and widely accepted in the

field of neural circuits. The original version of DCM for evoked responses rests on the Jansen and Rit *neural mass model* composed of three neuronal populations (David et al., 2005; Kiebel et al., 2006) (we will refer to this model as the *"ERP"* model, following the terminology used in the SPM software). A neural mass model provides a simplified description of the activity within a cortical area, resting on a small number of state variables like the mean membrane potential and the mean current of each population (*mean* refers to the *mean-field approximation*, where the neuron-level characteristics are being replaced by their averaged value over the whole population). DCM was recently augmented with another neural mass model based on a *canonical microcircuit* (CMC, Bastos et al., 2012). This model, that we used in this work, entails neurophysiological assumptions strongly inherited from predictive coding and thus make DCM even more convenient to test the biological correlates of this hypothesis. Precisely, the CMC implemented in DCM is a reduced form of the model proposed by Bastos et al. (2012) which was conceived as a possible implementation of predictive coding in the brain. The CMC for DCM is composed of four neuronal populations (Brown & Friston, 2012; Moran, Campo, et al., 2013; Auksztulewicz & Friston, 2016) (Figure 4.2):

- The superficial pyramidal cells (SP) predominantly located in cortical layers 2 and 3. The outputs of these neurons have been evidenced to be involved in *extrinsic* excitatory forward connexions (denoted A^{f} in Figure 4.2), and could thereby encode precision-weighted prediction errors on cause ξ_ν, leading population SP to represent the cause error unit (using the terminology introduced in chapter 2, §2.3.2).

- The deep pyramidal cells (DP) predominantly located in cortical layers 5 and 6. The outputs of these neurons are known to be involved in *extrinsic* inhibitory backward connexions (A^{b}), and could thereby encode predictions on states μ_x to be conveyed to lower levels. This would lead population DP to represent the state unit that update state conditional expectations.

- The spiny stellate cells (SS), predominantly located in cortical layer 4. Since these neurons have been evidenced to receive *extrinsic* excitatory forward input ($A_{SS}{}^{\mathrm{f}}$), population SS could represent the cause unit updating conditional expectations of cause μ_ν with regard to prediction errors received from lower levels.

- The inhibitory interneurons (II) of cortical layers 2, 3. Population II could be associated to the state error unit ξ_x, whose activity remains confined to the cortical area.

The intrinsic coupling between these four populations is presented in Figure 4.2: it rests on six excitatory and inhibitory *recurrent* connections (linking two neuronal populations together) as well as four inhibitory *self* connections (connecting a population to itself). The strengths (or gain) of intrinsic connections (denoted $\gamma_1, ..., \gamma_{10}$) vary over sources. Of particular importance is the strength of self inhibitory connection of population SP (γ_7). Indeed, as was mentioned in chapter 1, Bayesian processing of information rests on the precision weighting of information in order to give greater importance to more reliable beliefs. In a predictive coding scheme, *precision-weighted* prediction errors conveyed to higher levels in the hierarchy enable *"filtering out"* irrelevant (or poorly plausible) updates that should not be performed. This key computational aspect is transposed in CMC into parameter γ_7 which reflects the negative log-precision of prediction errors: $\gamma_7 = -\ln(\Pi_\nu)$ (Brown & Friston, 2013). Strictly speaking, Π_ν represents a weighting term, and could possibly reflect a precision ratio between (bottom-up) sensory and (top-down) predictions

as expected under predictive coding (Mathys et al., 2014). High value of Π_ν suppresses the self-inhibition of SP and enables the forward passing of prediction errors, hence triggers belief updating.

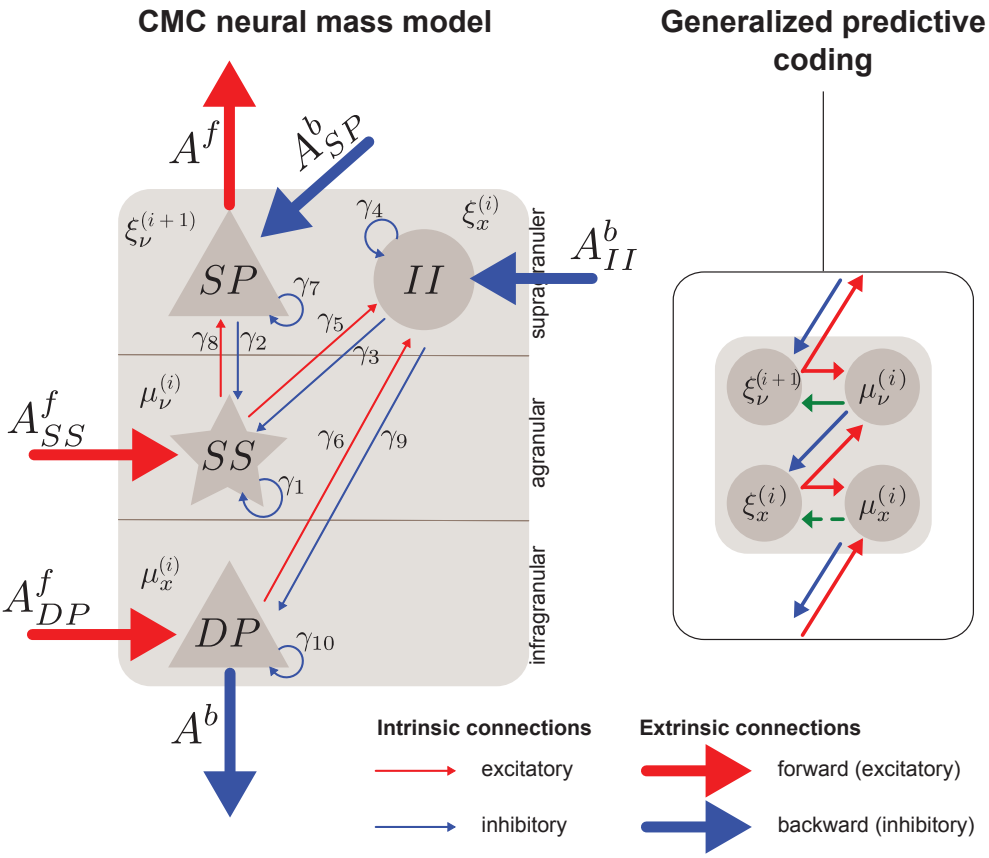

Figure 4.2 – Canonical Micro-Circuit model (adapted from Auksztulewicz et al. (2016). Left: schematic view of the circuitry within a source. Light brown rectangle represent a level within the hierarchy being a cortical macrocolumn; dark brown areas correspond to the different neuronal population composing this source, with triangle SP=supra-pyramidal cells, triangle DP=deep pyramidal cells, star SS= spiny stellate cells and circle II=inhibitory interneurons. $A_{SS}{}^{\mathrm{f}}$, $A_{DP}{}^{\mathrm{f}}$,$A_{SP}{}^{\mathrm{b}}$,$A_{II}{}^{\mathrm{b}}$: gain of the extrinsic forward (from SP to SS), forward (from SP to DP), backward (from DP to SP) and backward (from DP to II) connections targeting this level respectively. A^{f}, A^{b}: gain of the extrinsic forward and backward connection originating from this level respectively. γ_i: intrinsic coupling parameter. Right: typical organization of a hierarchical level of generalized predictive coding, as illustrated in Figure 2.3, is shown on the right to highlight the correspondence between both approaches.

Intercortical connexions. In DCM, the connections between the different cortical areas conform to the rules derived by Felleman and Van Essen (1991a) while adopting simplifying assumptions (Bastos et al., 2012) leading to the following scheme (Figure 4.2):

- Feedforward (or bottom-up) connections originate in superficial layers (population SP) and targets the agranular layers predominantly (population SS in layer 4) but also the infragranular layers (population DP)

- Feedback (or top-down) connections originate in infragranular layers (population DP) and target supragranular layers (predominantly population SP but also II).

Extrinsic connection strengths can be specified in specific matrices noted $A_{SS}{}^{\mathrm{f}}$, $A_{DP}{}^{\mathrm{f}}$,$A_{SP}{}^{\mathrm{b}}$,$A_{II}{}^{\mathrm{b}}$, which correspond to the extrinsic forward (from SP to SS), forward (from SP to DP), backward (from DP to SP) and backward (from DP to II) connections, respectively. For instance, the strength of the forward connection (from SP to SS) linking the third to the fifth source of a DCM is encoded in $A_{SS}{}^{\mathrm{f}}(5,3)$.

As can be seen in Figure 4.2, CMC has been designed so that the input signal that each cortical area receives is segregated into two signals: a forward message originating from population SP and a backward message originating from population DP. By dissociating the pyramidal cells of supra- and infragranular layers, CMC exploits the suitability of the hierarchical organization of the cortex for predictive coding, with prediction errors associated with excitatory connexions and predictions with inhibitory connections enabling the suppression of prediction errors.

The evolution model. The evolution model of DCM describes the dynamics of each neuronal population of each source in response to a stimulation, given the specified intrinsic and extrinsic interactions among these populations. It takes the following form:

$$\dot{x} = f(x, u, \theta)$$

where x, u and θ stand for the hidden states, the (exogenous) input or stimulation and the evolution parameters, respectively. More precisely, hidden states x represent the mean post-synaptic membrane potential and the mean current of each neuronal population (hence x for the four-population CMC is a vector of 8 elements). The evolution function f entails a set of ordinary differential equations adapted from the Jansen and Rit model (Jansen & Rit, 1995). Under this model, each population receives a synaptic input (or more exactly the mean density of pre-synaptic inputs) and transforms it into a post-synaptic membrane potential by means of the convolution with an impulse response p ; in turn, this membrane potential is converted into a firing rate with a (non-linear) sigmoidal function S and thus becomes a synaptic input for every connected (intrinsic and extrinsic) population (Figure 4.3). The expressions of p and S operators are given below:

$$p(t) = \begin{cases} \dfrac{H}{\tau} t \exp(-\dfrac{t}{\tau}) & if \quad t \geq 0 \\ 0 & if \quad t < 0 \end{cases} \tag{4.1}$$

$$S(x) = \frac{1}{1 + \exp(-rx)} - \frac{1}{2} \tag{4.2}$$

where H is a synaptic parameter controlling the maximum postsynaptic responses, τ is a synaptic time constant and r the sigmoid parameter. Parameters H and τ are specified for every neuronal populations in each source (contrary to r whose value is common to all populations of the DCM). In CMC, H is also equivalent to the strength of connections. Based on these variables and the Jansen and Rit operators, the eight evolution equations of DCM describing the rate of changes of voltage x_{volt} and currents x_{curr} for each population can be formulated; we provide those corresponding to population SP (the remaining six equations having a similar form can be found in Auksztulewicz

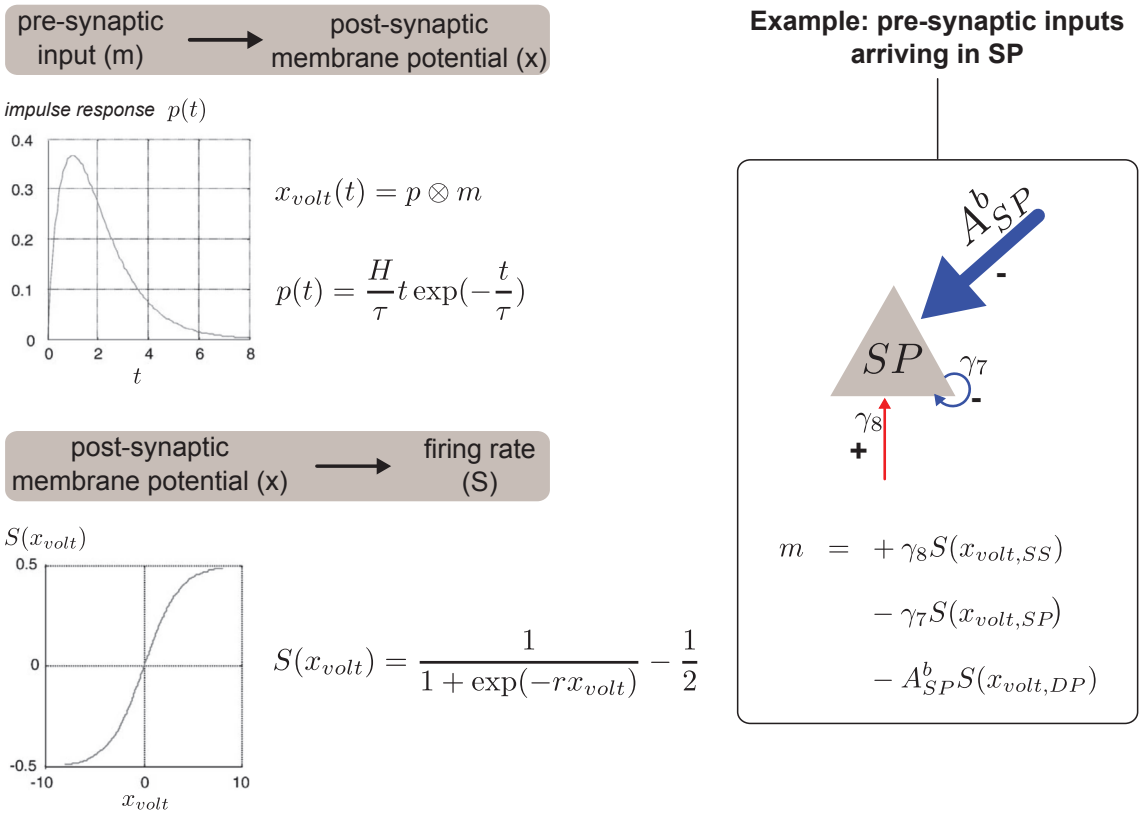

Figure 4.3 – Operators of the Jansen and Rit neural mass model (adapted from David et al., 2006). H, τ and r represent the synaptic parameter controlling the maximum postsynaptic responses, the synaptic time constant and the sigmoid parameter respectively (see main text). Example of synaptic inputs in population SP is provided (right) using the notations and color codes employed in Figure 4.2.

et al. (2015)):

$$\begin{cases} \dot{x}_{volt,SP} = x_{curr,SP} \\ \dot{x}_{curr,SP} = \dfrac{\gamma_8 S(x_{volt,SS}) - \gamma_7 S(x_{volt,SP}) - A_{SP}{}^{b} S(x_{volt,DP})}{\tau_{SP}} - \dfrac{2x_{curr,SP}}{\tau_{SP}} - \dfrac{x_{volt,SP}}{\tau_{SP}{}^2} \end{cases} \tag{4.3}$$

where γ_8, γ_7 represent the intrinsic gains as described in Figure 4.2. Variables $\gamma_{\cdot}$ and $x_{\cdot,\cdot}$ correspond to vectors containing the values relative to each source of the DCM structure. These equations show how the voltage is updated as a function of the current, and how the current evolves with both current and voltage. Inhibitory synaptic inputs are integrated as negative contributions in order to model the decrease of neuronal responsivity that they induce. Importantly, stimulus input is modeled as a thalamic excitatory input arriving at population SS of the sources that have been declared to receive exogenous inputs. Finally, it should be pointed out that the evolution parameter θ comprises the synaptic parameters (r, τ), the intrinsic coupling parameters (γ), the extrinsic coupling parameters (A matrices), as well as trial-specific effect parameters (see below) and input and conduction delay parameters (David et al., 2006). Simulations of evoked responses constitute an efficient way to experience how such parameters affect the recurrent dynamics taking place in the network (see for instance, David et al., 2005; Kiebel et al., 2007).

Modulatory connections. DCM with CMC also comprises an additional (and optional) extrin-

sic *modulatory* connection between population SP of a cortical area and population DP of its first parent (Brown & Friston, 2013). Applying this connexion makes the self-inhibitory gain of population SP $\gamma_7{}^{(i)}$ dependent of the firing rate of the higher-level population DP $S(x_{volt,DP}{}^{(i+1)})$ in such a way that larger value increases the suppression of self-inhibition in SP. This connection aims at modeling the top-down influence of predictions, assuming that a larger activity in DP reflects larger prediction updates that have to trigger prediction errors until those have been explained away. Using this connection lends $\gamma_7{}^{(i)}$ be no longer time-independent.

Trial-specific effects. Another key feature of DCM is the trial-specific effect (embodied in the B matrix, following the same notation as in the SPM implementation) which enables modulating the connection strength relatively to the type of exogenous input (a standard or a deviant for instance). It should be noted that *trial-specific* here refers to the modulation of evoked responses by experimental manipulation. Non-diagonal elements of matrix B encode the modulation of extrinsic connections. For instance, the strength of the forward connection (from SP to SS) linking the third to the fifth source of a DCM is equal to $A_{SS}{}^{f}(5,3)$ for standard inputs, and to $B(5,3)A_{SS}{}^{f}(5,3)$ for deviant inputs. Diagonal elements pertain to intrinsic modulations and apply to the self inhibitory gain of population SP γ_7. This trial-specific modulation assumes differential synaptic changes induced by the different input types. Modulation of backward connections entailing a modulatory connection (see paragraph above) applies to the intrinsic gain γ_7 instead of the backward extrinsic gain.

The observation model. The observation model of DCM rests on a typical forward model as used for static source reconstructions, mapping neuronal activity to EEG or MEG sensors (Kiebel et al., 2006). It writes:

$$y = g(x, u, \psi) + \varepsilon_n$$

where y denotes the EEG or MEG data, x the hidden states (being the evolution states $x_1, ..., x_8$), u the exogenous input, ψ the observation parameters and ε_n the residuals. Function g corresponds to the forward model, transforming neuronal activity into electrophysiological responses, and includes simplifications about this biophysical mapping. Measurement noise is assumed to be zero mean Gaussian and is defined relatively to the *temporal* covariance of data Σ_n, parameterized by hyperparameter λ. In DCM with CMC neural mass model, cortical areas are modeled as ECD, whose position and orientation are represented by parameter ψ (note that ECD parameters are thus time-independent). As already seen, each source comprises four neuronal populations; only the post-synaptic activity (depolarization) of the excitatory populations SP, DP and SS are allowed to project on sensors, with greater (prior) importance assigned to the superficial pyramidal cells (SP).

Model inversion. Previous paragraphs have described the main features of DCM, being a generative model of brain responses. This paragraph is about inverting this generative model (from observed evoked responses) to infer the effective connectivity at play during the experimental manipulations that have caused the observed EEG or MEG data. DCM inversion provides conditional expectations (or posteriors) of both evolution and observation parameters. Interestingly, regarding the latter parameters, DCM inversion can be seen as a source reconstruction approach augmented with temporal constraints. Practically speaking, DCM inversion aims at inferring the

conditional distribution of θ and ψ given data y and a model m using Bayes's rule:

$$p(\theta, \psi|y, m) = \frac{p(y|\theta, \psi, m)p(\theta, \psi|m)}{p(y|m)}$$

Inversion rests on Variational Bayes (VB) to estimate $q(\theta)$ and $q(\psi)$, the mean-field approximations to $p(\theta, \psi|y, m)$, using the Laplace approximation. The variational free energy $F(q, \lambda, m)$, with λ the error variance, is maximized using a EM scheme (VB-EM). Step E provides the conditional distribution of θ and ψ (by means of a descent on $F(q, \lambda, m)$) and step M estimates λ. The iterative scheme can be found in Kiebel et al. (2006).

Summary. DCM is a Bayesian-based method aiming at inferring the (hidden) effective connectivity taking place during an experimental manipulation. Despite simplifying assumptions, strong efforts have been put into DCM to endow this procedure with a biological plausibility: DCM rests on both a dynamic system physiologically informed by widely accepted animal findings and a biophysical forward model mapping neuronal activity to sensors. DCM with CMC has been proposed to address the present need of investigating predictive coding in the brain. Model inversion providing both the posterior estimates of unknown parameters and the (approximate) of model evidence, allows for bayesian model selection. DCM thus enables the formal comparison of alternative hypothesis about the cortical implementation of predictive coding in the context of mismatch responses.

4.2.3 DCM of mismatch responses

This section aims at presenting the key results of DCM analyses of deviance responses obtained in less than ten years. We first describe the early studies whose aim was twofold: validating the DCM method to model evoked responses and characterizing at the same time the change in effective connectivity induced by deviance processing. Latest DCM studies include those that have continued to investigate the neural mechanisms of the MMN guided by predictive coding predictions (reported here), and other works that used DCM of MMN as a proxy to reveal different perceptual processing between groups or between experimental conditions.

The initial work of Garrido and colleagues. First studies aiming at characterizing the DCM of deviance processing were performed by Garrido and colleagues in the late 2000's using EEG data and DCM-ERP. A series of five papers addressed the following questions:

- The architecture of the deviance processing (Figure 4.4.a). Two passive frequency deviance studies compared a three-level asymmetric architecture (including five sources inherited from fMRI MMN studies: bilateral primary auditory cortex, A1; bilateral posterior STG and right IFG) to bilateral networks comprising one, two or three levels (Garrido et al., 2008; Garrido, Kilner, Kiebel, & Friston, 2009). Model comparison based on real EEG data supported the five-source DCM.

- The extrinsic connectivity (Figure 4.4.b). Based on deviant responses elicited during an active oddball frequency paradigm where subjects had to count deviants, forward extrinsic connectivity was found more likely than reciprocal extrinsic connectivity to accommodate

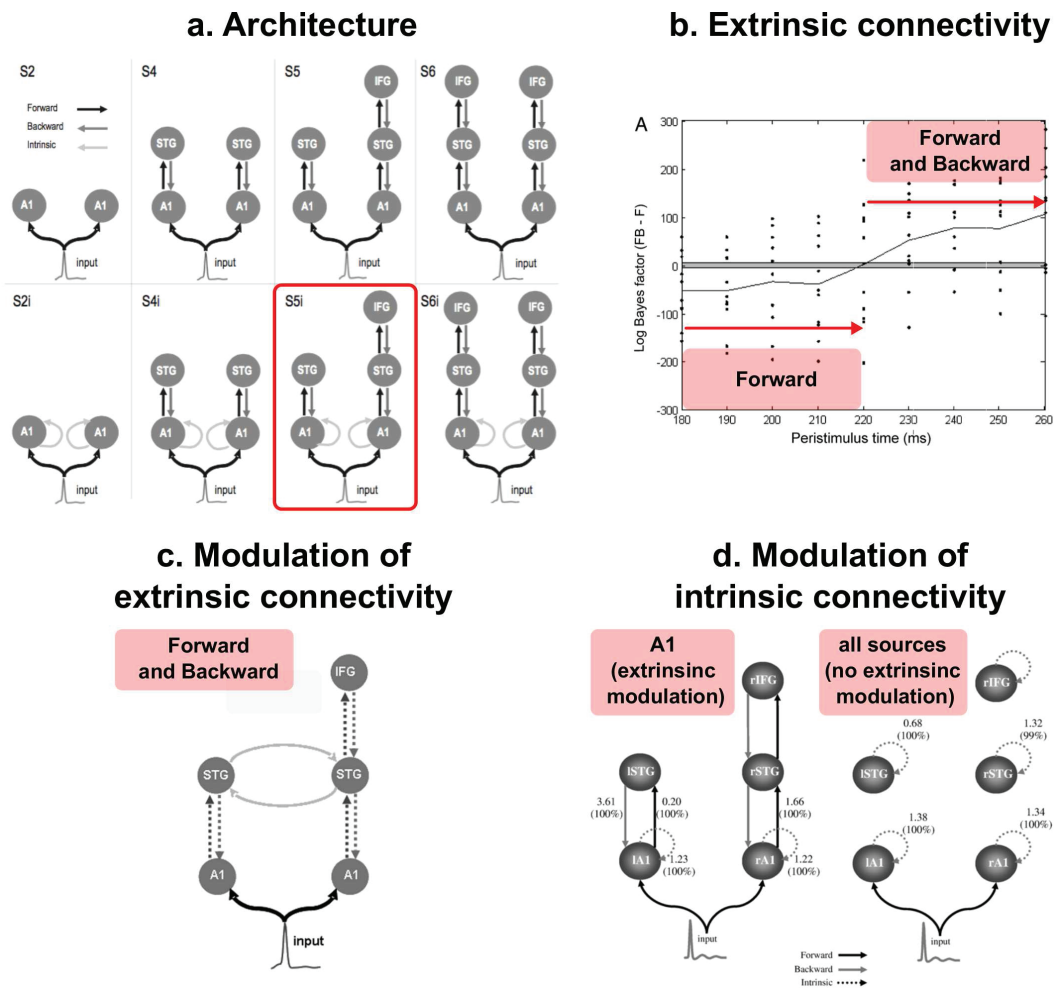

Figure 4.4 – Initial findings in DCM of mismatch responses. a) Model space used to test the network architecture, with winning model (red rectangle), adapted from Garrido et al. (2009). b) Bayesian comparison of model F (forward connectivity) and model FB (forward and backward connectivity), with individual log-Bayes factor values (black dots) measured for each time-interval DCM inversion. Winning models are specified in red rectangles, with a shift around 220 ms; adapted from Garrido et al. (2007). c) Winning model selected in a trial-specific modulation study, adapted from Garrido et al. (2007). Note that lateral connections between STG sources were specified (these are specific to the ERP model, David et al., 2006) but they were not used in subsequent studies. d) The two winning models selected in an intrinsic modulation study, adapted from Kiebel et al. (2007).

deviant ERPs from stimulus onset to the end of the MMN (around 220 ms), but the reverse conclusion was obtained for larger time interval including the P3 component (Garrido, Kilner, Kiebel, & Friston, 2007) (No trial-specific modulation was considered in this study). From a dynamical system point of view, this result suggests that backward connections are required at later latency than the MMN to induce the generation of late component at lower levels.

- The modulation of extrinsic connectivity (Figure 4.4.c). Using the same data (including standard and deviant responses) and a model space adapted to the extrinsic modulation issue, Bayesian model comparison supported synaptic changes in both forward and backward connections in comparison to forward only or backward only to accommodate the difference between standard and deviant ERPs (Garrido, Kilner, Kiebel, Stephan, & Friston, 2007).

Consistent results in favor of this reciprocal modulation were also reported in (Kiebel et al., 2007; Garrido et al., 2008; Garrido, Kilner, Kiebel, & Friston, 2009). However, the results of subsequent statistical analysis of the modulation gain per connection (the non-diagonal elements of matrix B) show different patterns over the studies of Kiebel et al. (2007) and Garrido et al. (2009): a significant increase of backward connection strength with deviants could be observed only between STG and A1 in the left hemisphere for the former study and bilaterally in the latter study. Regarding the modulation of forward connection, it was found significant between A1 and STG in the study of Kiebel (with opposite effect between hemispheres) and in the right hemisphere in the study of Garrido (with opposite effect between levels, namely an increase of connection strength between A1 and STG and a decrease between STG and IFG).

- The modulation of intrinsic connectivity. It was first assessed in (Kiebel et al., 2007) with a model space composed of DCM with no intrinsic modulation, intrinsic modulation within A1 only or within all sources (in combination with changes in extrinsic modulation). Results strongly supported intrinsic modulation for A1 but not for the higher levels where extrinsic modulation was found equally plausible to account for the deviant effect (Figure 4.4.d). Garrido and collaborators also support intrinsic modulation within A1 but did not test its possible involvement higher up, in STG and IFG (Garrido et al., 2008; Garrido, Kilner, Kiebel, & Friston, 2009). Analysis of the modulation gain in each source (the diagonal elements of matrix B) consistently revealed an A1 gain increase with deviants.

To sum up, these pioneering studies validated DCM as an efficient procedure to test model-driven hypothesis regarding the generation of mismatch responses. In particular, predictive coding was found outperforming (and reconciling) other MMN models (the adaptation and the sensory memory models). BMS analysis concluded in favor of a three-level architecture with reciprocal connectivity and trial-specific modulation of this connectivity. However, inference on extrinsic and intrinsic parameters (conducted in the latest studies) rather failed to exhibit consistent deviant modulation among studies.

Recent findings. We report here the (few) DCM MMN studies dedicated to further improve the characterization of effective connectivity during deviance processing. Two EEG studies employed neuropharmacological manipulations, which highlight a noticeable aspect of DCM: it allows investigating in a precise and formal manner the effect of such manipulation on (synaptic-based) connectivity, that could not be captured by traditional ERP comparison. Using DCM-ERP, deviance responses elicited under placebo and under ketamine (an antagonist to NMDA receptors known to reduce MMN amplitude) were compared. This comparison revealed a significant decrease of the forward connection strength between left A1 and left STG with ketamine (Schmidt et al., 2013). This result demonstrated the potential of this DCM procedure but the choices used for the analysis may have prevent from a finer characterization of synaptic changes (for instance, the window used for inversion suits for the MMN but sensor-level analysis revealed a significant ketamine effect at the transition between the MMN and the P3a). The second study constitutes a significant step towards the characterization of predictive coding for deviance processing as it addressed neuromodulation as the correlate of precision of predictions (Moran, Campo, et al., 2013). Using DCM with CMC, mismatch responses under placebo and galantamine (cholinergic modulation by galantamine is a plausible candidate for precision encoding) were compared by means

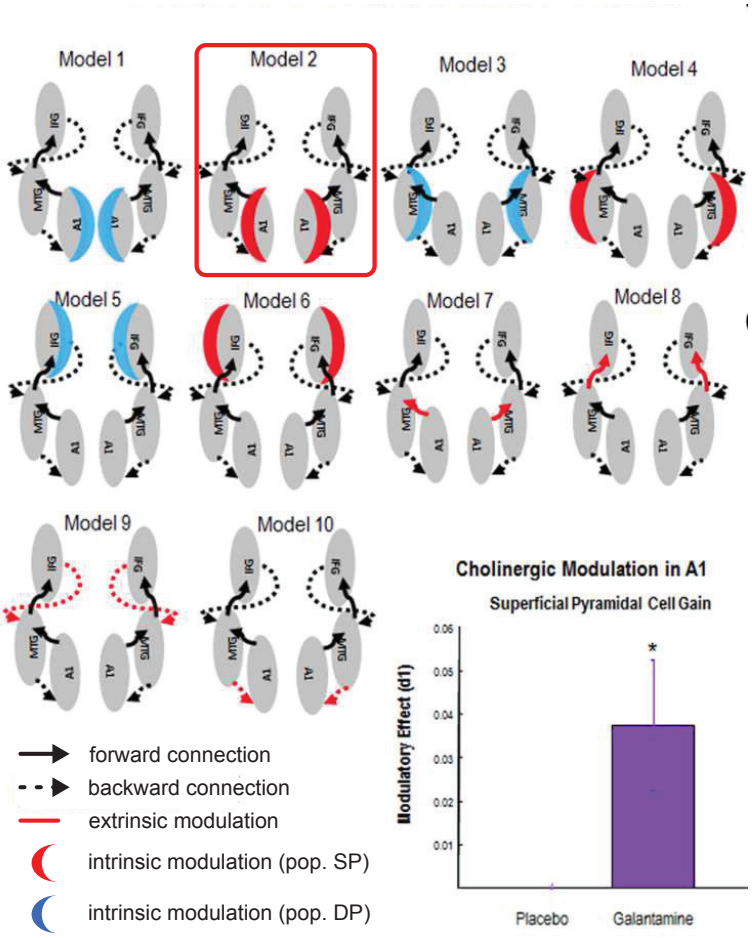

Figure 4.5 – Neuromodulation as the correlate for precision of prediction encoding (adapted from Moran et al., 2013). a) Model space embedding the different (and exclusive) connectivity changes that could reflect galantamine modulation, with the winning model (red rectangle) corresponding to modulation in population SP in bilateral A1. b) Comparison of self-inhibitory gain in both conditions.

of BMS applied to an appropriate model space (Figure 4.5). Each DCM embedded a specific hypothesis regarding the influence of galantamine (for instance, a decrease of the self-inhibitory gain in superficial pyramidal cells would support larger precision weighting of prediction errors). Their findings showed such cholinergic modulation in the lower level of the cortical auditory hierarchy. Finally, an MEG study using CMC attempted to measure the difference in synaptic changes induced by the opposite effects of attention and expectation on evoked responses (an increase and a decrease respectively) (Auksztulewicz & Friston, 2015). Such characterization would inform on the cortical implementation of predictive coding as both effects should affect differentially the precision of predictions. To that aim, they manipulated both effects orthogonally using oddball-like sequences and conducted a factorial-design DCM analysis. Most importantly, this study illustrates the great potential of DCM (with CMC in particular) to disentangle competing model-driven hypothesis to explain sound processing.

Summary. This overview emphasizes the utility of DCM to investigate the neural underpinnings of mismatch responses. Latest studies involved a two-step analysis where BMS first enables selecting the more plausible generative model among competing DCMs. Inference on parameters (derived

from the winning model[1]) further allows examining more finely changes in connectivity induced by experimental manipulations (or difference between groups, Schofield et al., 2009; Boly et al., 2011; Cooray et al., 2014). Both steps can be employed to challenge predictive coding predictions relative to standard and deviant processing. The CMC neural mass model in particular makes it possible to test the hypothesis of self-inhibitory connection encoding precision of information by means of pharmacological but also experimental manipulations (see for instance the effect of attention on precision assessed in a visual cue-target task, Brown & Friston, 2013). Besides, the cited studies reported findings that give insights into MMN generation, notably with regard to the architecture and the reciprocal connectivity behind the auditory hierarchy. As noted above, many aspects remain to be clarified, including the direction of change in connectivity induced by deviants and its confrontation with predictive coding expectations. In (2009), Garrido and colleagues pointed out the difficulty to interpret brain mechanisms (from a functional perspective) from DCM estimates. All together, this suggests that it is worthwhile spending additional efforts using DCM to better characterize the dynamics of neuronal interactions during auditory processing, and it emphasizes the importance of the predictive coding message-passing scheme to guide this investigation.

4.3 Computational learning models

We present here different (and recent) studies that have proposed computational (or cognitive) models of the MMN based on the Bayesian learning of environmental sensory regularities. Contrary to DCM that operates at the ERP timescale to provide the timecourse of neural responses over peristimulus time, computational learning models consider trial-by-trial changes (thus at the experimental timescale) in order to relate brain activity to the Bayesian processing of each stimulation. The purpose of this section is twofold: it aims at illustrating how such models can be designed (in order to frame specific functional hypothesis into a mathematical formulation) and tested against real data. And, it describes the findings regarding the processing of oddball sequences obtained with such modeling approaches. Two emblematic and pioneering studies in particular are reviewed.

Throughout this section, we have used the terminology of *perceptual* and *response* model, as well as the notations introduced in chapter 2, §2.4.1. Consequently, notations adopted in some of the original papers may have been replaced by ours for sake of consistency throughout the document.

4.3.1 The MMN as a Bayesian surprise (Ostwald et al., 2012)

This study tested the hypothesis that neural activity, as recorded with EEG, reflects prediction error and hence Bayesian inference in the brain. They did so using a somatosensory roving[2] (oddball) paradigm (Ostwald et al., 2012). Roving sequences were composed of two electrical stimuli with either a high or low intensity (Figure 4.6-a). The attention of the subjects was engaged to

[1] or the winning family in the case of large model space. Parameters are thus obtained with BMA performed over the different models composing the winning family.

[2] The roving paradigm constitutes a variant of the typical oddball paradigm (Cowan et al., 1993), where the sound sequence is composed of alternating trains of repeating items hence enabling the different physical properties of standards and deviants to be mirrored.

the stimulations as they were asked to count the number of trains in each session. Description of both the perceptual models designed in this study and the procedure (based on the inversion of a response model) used to confront EEG data to perceptual model predictions are provided below as they have guided one of the approaches that we used in this work. Each perceptual model corresponds to a specific hypothesis of how the brain may treat the sequence of sensory stimulations.

Bayesian perceptual model. The hypothesis behind this model is that the brain instantiates a generative model of the stimulus sequence, and updates its parameters through (Bayesian) learning. As the sequence rests on two different stimuli, the authors have considered a beta-Bernoulli model as follows:

- at each trial t, the probability to observe a stimulus u being of high intensity is given by μ (and the probability to observe low intensity is thus equal to $1-\mu$). The likelihood of the generative model is thus a Bernoulli function:

$$\begin{cases} u|\mu \sim \text{Bern}\,(\mu) \\ \\ p(u|\mu) = \mu^{\text{u}}(1-\mu)^{\text{1-u}} \end{cases} \tag{4.4}$$

  with $u = 1$ for high intensity (and $u = 0$ for low intensity).

- μ is treated as a parameter to be learned and is subject to uncertainty. To model the prior knowledge on μ (before having observed trial t), the common choice is the beta distribution of parameters a_t and b_t, where a_t (b_t) can be interpreted as the number of high (low) intensity stimuli already observed before trial t. The prior distribution of the generative model for trial t thus expresses as:

$$\begin{cases} \mu \sim \text{beta}\,(a,b) \\ \\ p(\mu) = \dfrac{\Gamma\,(a_{t-1}+b_{t-1})}{\Gamma\,(a_{t-1})\,\Gamma\,(b_{t-1})}\mu^{\text{a}_{\text{t-1}}}(1-\mu^{\text{b}_{\text{t-1}}}) \end{cases} \tag{4.5}$$

  where Γ denotes the gamma function.

Once trial t has been delivered, Bayesian inference consists in inverting this generative model to infer the updated belief on μ, namely $p(\mu|S)$. The conditional expectation of this parameter evolves over trials according to update equations that specifically form the evolution part of the response model (see below).

The response model. This model encompasses the evolution function derived from the inversion of the perceptual model inversion, as well as an observation model mapping internal states (the Bayesian surprise, see below) to the observed brain data. Since the beta prior is conjugate to the binomial (or Bernoulli) likelihood, Bayesian inference for perceptual model inversion is analytically tractable and the update equation for μ derives from the expression of the posterior distribution of μ:

$$p(\mu|a_t,b_t) = \frac{\Gamma\,(a_t+b_t)}{\Gamma\,(a_t)\,\Gamma\,(b_t)}\mu^{\text{a}_\text{t}}(1-\mu^{\text{b}_\text{t}}) \tag{4.6}$$

Interestingly, a forgetting parameter τ has been introduced in the model to account for different temporal integration windows that the brain may entertain. This is achieved by means of an exponential function that weights the different stimulus counts a_t and b_t, yielding to a_{w_t} and b_{w_t}. Each trial of the sequence is associated with a prior and a posterior distribution of μ, parameterized with the (weighted) stimulus counts at that trial. The following expression:

$$p(\mu|a_{w_t}, b_{w_t}) = \frac{\Gamma\left(a_{w_t} + b_{w_t}\right)}{\Gamma\left(a_{w_t}\right)\Gamma\left(b_{w_t}\right)}\mu^{\mathrm{a_{w_t}}}(1-\mu)^{\mathrm{b_{w_t}}} \tag{4.7}$$

thus characterizes the posterior distribution of μ at trial t, and its prior at trial $t+1$. The observation model rests on the *Bayesian surprise* which is computed at each trial to furnish a quantitative measure of belief updating or surprise, or equivalently prediction error. The Bayesian surprise is defined as the KL divergence (see chapter 1, §1.3.2) between the prior and the posterior distribution of μ, and is thus informed by the conditional expectation and the precision (or inverse variance) of each distribution. We call the *trajectory* of Bayesian surprise the vector X containing the N values of this measure computed for each of the N stimuli of the sequence u. The observation model is a two-level linear model of the form:

$$\begin{cases} y = X\theta_1 + \varepsilon_1 \\ \\ \theta_1 = 0 + \varepsilon_2 \end{cases} \tag{4.8}$$

where y indicates the data feature to be fitted, ε_1 and ε_2 the first and second level gaussian noise, respectively. Different trajectories obtained with different forgetting values are showed in Figure 4.6. Five models of this kind were considered in this study, resting on different values for τ, including infinity value (corresponding to full memory with no forgetting). These models are denoted BS0 (infinite τ), BS1, BS2, BS3 and BS4.

Non-Bayesian models. Three additional models were used in this study, that did not involve Bayesian learning of the statistical regularities of the stimulus sequence. The first model is called the simple change model (model SC) and assigns to trial t a value of 0 if the current stimulus is equal to the preceding one, and 1 otherwise. The second model is called the linear-change model (LIN), which is similar to model SC but assigns, at each deviant, a prediction error proportional to the number of preceding standards. This aimed at modeling the larger MMN amplitude observed as the number of repeated standards increases. The last model, model M0, corresponds to the null hypothesis and assigns a constant value of 1 to each trial. Models SC and LIN are classical models of the MMN. Each of these models also provides a regressor X composed of N values to be fitted to the EEG data y.

Data feature to be fitted. This paragraph describes the data features y entering Eq. (4.8). One specificity of this study pertains to the procedure by which the data y were obtained. First, source reconstruction was performed on individual evoked responses that resulted at the group-level in 6 dipoles with fixed locations and orientations. Secondly, for each subject, single-trial responses were projected onto these sources. Then, for each subject and for each source, y_k denoted the vector of data at peri-stimulus time k and was thus composed of N values (for N trials).

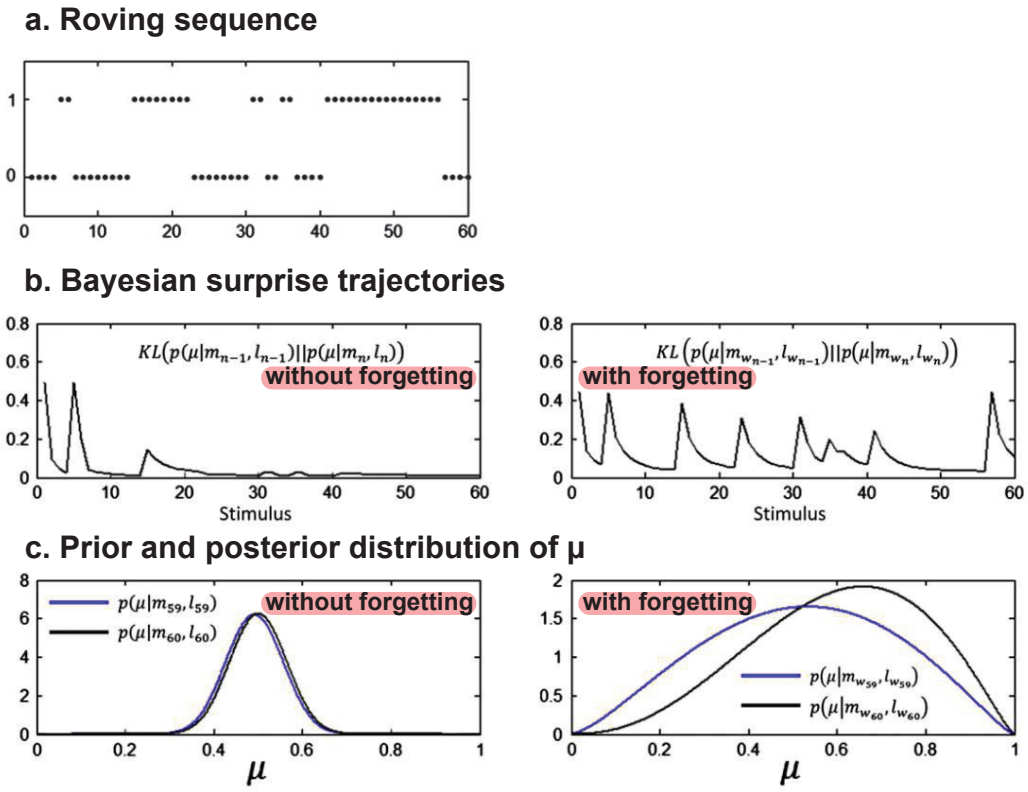

Figure 4.6 – Simulations of Bayesian surprise, adapted from Ostwald et al. (2012). a) Schematic view of the two-stimuli roving sequence, composed of 60 items. b) Trajectories of Bayesian surprise obtained with two forgetting values; (left) without forgetting, the model rapidly learns the real probability to have a deviant, leading to accurate predictions for μ precluding update belief: the Bayesian surprise rapidly reaches near zero values. (right) with forgetting, the inferred value of μ at each trial depends on the number of standards and deviants delivered within the sliding-window for temporal integration, leading to prediction errors reflected by the peaks of the Bayesian surprise trajectory. c) (left) without forgetting, prior (blue) and posterior (black) distributions of μ at the end of the sequence (trial 60) are almost identical; (right) with forgetting, both distributions differ that generate non-null Bayesian surprise.

Model inversion. Model inversion aims at inferring the hidden parameters of the perceptual and the response models given observed data y_k and stimulus sequence u delivered to the subject. In this study, none of the perceptual models involved any free hidden parameters. Strictly speaking, model inversion does not imply a meta-Bayesian scheme (as described in in chapter 2, §2.4.1) since it rests on a two-step procedure: first, trajectory X_i is computed for each model m_i. Second, the following model:

$$\begin{cases} y_k = X_i\theta_{1,k} + \varepsilon_{1,k} \\ \\ \theta_{1,k} = 0 + \varepsilon_{2,k} \end{cases}$$

is inverted for each sample k using EM (performed with the parametric empirical Bayes procedure implemented in SPM, Friston et al., 2007). A total of M inversions were thus performed, with M being equal to the product of the number of models, the number of subjects, the number of sources and the number of samples. In this way, a *relative free-energy map* could be obtained for each source, each subject, where each pixel (i, j) represents the value of the log-Bayes factor for model m_i (namely the relative free energy $F_{m_i} - F_{M0}$), at sample j. Using a FFX model, these individual maps were summed over the group of subjects, as reported in Figure 4.7. Besides, BMS

was conducted based on pair-wise free energy comparisons for relevant time-windows, indicating the more plausible models over each window.

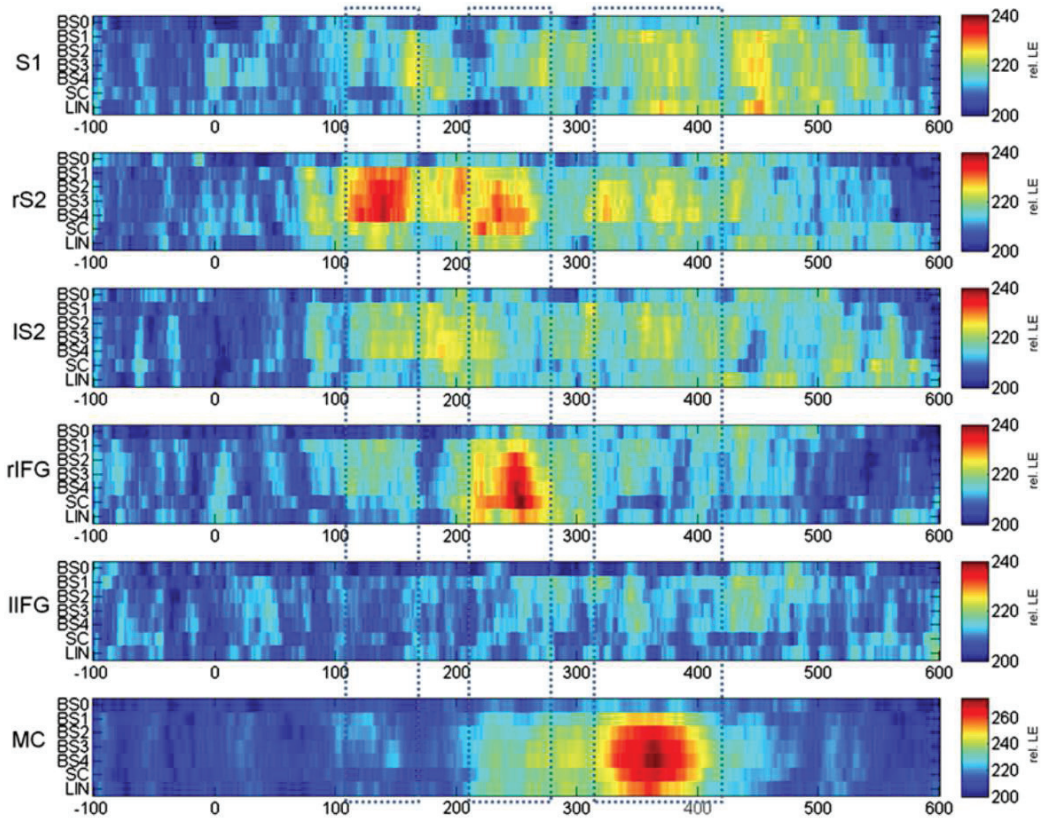

Figure 4.7 – Findings from Ostwald et al. (2012). The six relative free energy maps obtained for each source are depicted (S1=right primary sensory cortex, rS2=right secondary somatosensory cortex, lS2=left secondary somatosensory cortex, rIFG=right inferior frontal gyrus, lIFG=left inferior frontal gyrus, MC=medial cingulate cortex). For each map, peri-stimulus time (from -100 ms to 600 ms) is represented on x-axis, and models BS0-BS4, SC and LIN are represented on y-axis; the color at each pixel encodes the relative free energy difference $F_{m_i} - F_{M0}$, *ie* the log-Bayes Factor of model m_i and the null model.

Results. Using this approach, a spatio-temporal characterization of the functional mechanisms at play during this active oddball task could be assessed. Relative free energy maps indicated only positive values (> 200), meaning that the null model could be rejected for any latency, any source and any model (chapter 1, §1.4.3). Three time-windows were found to be discriminative between models: around 150 ms after stimulus onset, the activity of the right secondary somatosensory ECD was associated with the Bayesian surprise model with forgetting (BS1–BS4); around 250 ms, model SC and model BS4 outperformed others in the right inferior frontal cortex; finally around 350 ms, the medial cingulate ECD supported Bayesian learning models with different forgetting values (BS2–BS4). Learning model without forgetting (BS0) was always found with the lowest relative free energy, suggesting that infinite temporal integration was not a plausible account for the observed data. These results are encouraging, as they rest on single-trial modeling and yet succeeded in revealing that Bayesian learning models outperform classic (non learning) other models at relevant time intervals.

4.3.2 Free energy principle models of the MMN (Lieder et al., 2013)

This study rests on the EEG data described in (Garrido et al., 2008), with auditory mismatch responses elicited during a roving frequency paradigm made of seven different frequencies (Lieder, Daunizeau, et al., 2013). It aimed at testing whether single-trial MMN amplitudes could reflect computational quantities related to the Bayesian processing of sound sequence and specifically expected under the free energy principle. We describe the key features of this study using the framework of previous section.

Bayesian perceptual model. The brain's generative model of the present roving sequence is assumed to rest on hidden causes ν (related to the mapping of the true frequency of u_t to its percept or pitch) and hidden parameter $\vartheta^{(\mathrm{p})}$ (comprising notably the expected length of trains and a frequency transition probability matrix).

Response model. The evolution function f of the response model describes perceptual model inversion achieved after each sound presentation, leading to the conditional expectation of ν and $\vartheta^{(\mathrm{p})}$ that are represented as the hidden state $x = \{x_\nu, x_{\vartheta^{(\mathrm{p})}}\}$. Their dynamics over the experimental timescale follows:

$$x_{t+1} = f(x_t, u_t, \vartheta^{(\mathrm{r})}) \tag{4.9}$$

where $\vartheta^{(\mathrm{r})}$ denotes the hidden parameters of the response model (including subject-specific parameters). Assuming the free energy principle, posterior distribution $p(\nu, \vartheta^{(\mathrm{p})}|u_t)$ is approximated by $q(\nu, \vartheta^{(\mathrm{p})})$ with q being a delta distribution. This assumption lends to analytical one-step update of x that minimizes the free energy of the perceptual model $m_{(p)}$, hence conforming to the free energy principle[1]:

$$x_{t+1} = \underset{x_{t+1}}{\mathrm{argmin}} F(x_{t+1}, u_t, x_t, m^{(\mathrm{p})}) \tag{4.10}$$

This evolution model was combined with different observation models g_i to test specific hypothesis regarding the mapping of internal quantities to the observed EEG data. These different mappings appeal for instance to a precision-weighted prediction error relative to the sensory input u_t or to the adjustment of parameters $\vartheta^{(\mathrm{p})}$ (resting upon the difference between expected values at trial $t-1$ and trial t). Similarly to the study of Ostwald et al. (2012), each model g_i has a general linear form (Eq. (4.8)) with regressor X expressing as follows:

$$X = g_i(x, u, \vartheta^{(\mathrm{r})}) \tag{4.11}$$

Non-Bayesian models. As in Ostwald et al. (2012), alternative response models that did not involve Bayesian information processing were considered. They aimed at modeling change detection mechanisms and synaptic adaptation. These *"phenomenological"* models rest on a linear observation model g that provided a regressor X to be confronted to the data (Eq. (4.8)).

Data feature to be fitted. This study dealt with single-trial MMN amplitudes measured at selected fronto-central electrodes, leading to a vector y of N values in Eq. (4.8) obtained for each

[1]The *argmin* operator applied to a function f returns the values of the domain of f at which minima are attained.

electrode (generalization to the multivariate linear model required to account for multiple electrodes is ignored here for the sake of clarity).

Model inversion. Individual inversion of the multivariate Bayesian linear regression model was performed for each perceptual model using a Monte-Carlo procedure, leading to posterior distributions of $\vartheta^{(p)}$ and $\vartheta^{(r)}$, as well as the free energy approximating model log-evidence. Bayesian model family comparison were conducted over the group of subjects using a RFX model.

Results. Model comparison failed to reveal a particular model having significant larger model evidence. However, family model comparison based on the *"phenomenological"* and the *"free energy principle"* families was clearly in favor of the latter ($p(fam_{fep}|y) = 0.87$). This suggests that Bayesian learning models were more equipped to capture the single-trial changes in the MMN amplitude. These changes depend on the sequence of preceding sounds delivered before each deviant (past observations). Such short timescale context of deviant occurrence was further evidenced in this study to affect MMN amplitude. These results highlight the importance of trial-by-trial analysis to assess the learning of regularities that shape the MMN and that is ignored in typical evoked response studies. However, the authors also discussed some limitations of their study, including the lack of neurophysiological model mapping computational quantities to EEG responses.

To conclude, the studies reported in this section illustrate the typical procedure one should adopt when attempting to model cognitive functions (perceptual inference and learning in our case) based on neuroimaging data. It starts with designing a perceptual model, then update equations handling model inversion should be specified (possibly involving relevant assumptions to derive one-step and/or exact updates). These equations form the evolution function that enter the response model in combination with an observation function providing the biophysical mapping. As with DCM, competing hypothesis (alternative perceptual models, response models or both) can be formally compared with Bayesian model comparison resting on model log-evidence derived for each response model. Besides, this overview points to the small number of studies that have addressed the MMN as a Bayesian inference process. However, and encouragingly, empirical findings detailed here were fairly in favor of learning models and thereby call for further trial-by-trial modeling analysis to provide insights into how the brain learns the statistical regularities of sound sequences.

4.4 Attempts of computationally-informed dynamic causal models

In (2013), Lieder and collaborators pointed out the simplification used in their computational modeling approach - also used in the other studies cited above - where they have related learning-based quantities (such as prediction error) to neuroimaging data without having explicitly characterized the intermediate relationship between these quantities and neurophysiological data (such as synaptic gain). This could be theoretically resolved by a *"meta-response"* model that would involve two levels of response: one handling the computational to biophysical mapping, and the one mapping those biophysical variables onto observations. In Figure 4.8, we attempted to represent this (complex) issue with the different levels it entails and we also added the different modeling

work with regard to each level.

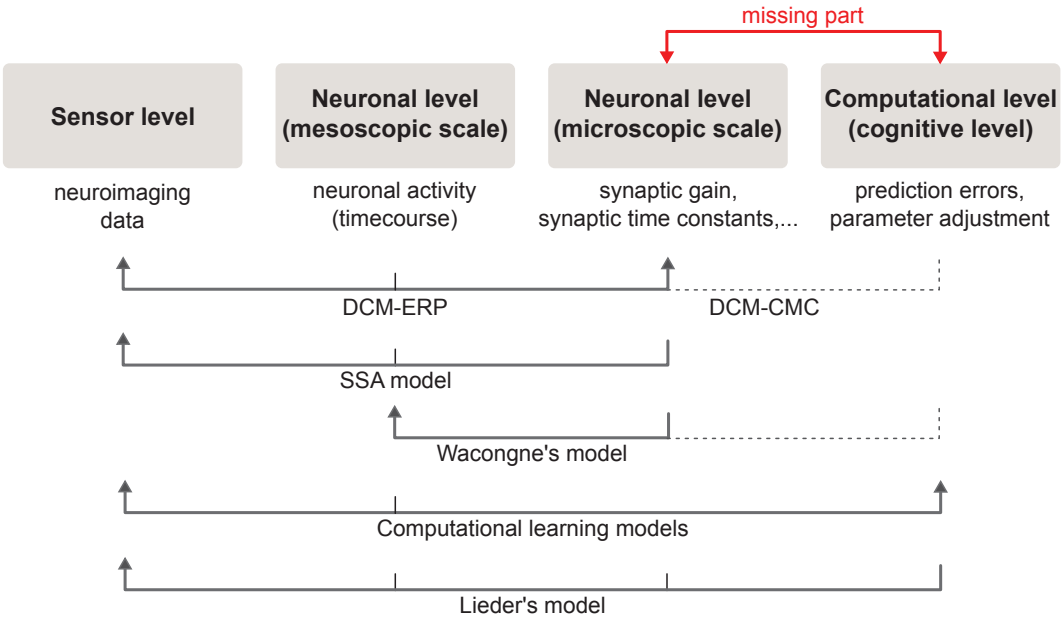

Figure 4.8 – Schematic view of the different modeling levels. SSA model refers to (May et al., 2015), Wacongne's model refers to (Wacongne et al., 2012), computational learning models refer to (Ostwald et al., 2012; Lieder et al., 2013a; Mathys et al., 2014), Lieder's model refers to (Lieder et al., 2013b). Single arrow indicates model for which the generative direction has been reported but not model inversion, double arrow indicates that both directions were documented. Dashed line indicates that predictions from learning models were considered. Tick indicates a level accounted for by the model.

DCM with CMC appears as a first attempt to bridge the gap between the computational and neurophysiological levels. Indeed, specific neuronal populations subjected to synaptic changes have been attributed to the computational units of a predictive coding scheme in agreement with current neurobiological knowledge. It should be recalled that DCM operates at the ERP timescale, and the modulation of synaptic gains by experimental manipulation (reflected in DCM's B matrix) is time-independent (only synaptic changes between two conditions only can be assessed, but see Garrido, Kilner, Kiebel, Stephan, et al., 2009). The SSA model of May and collaborators (2015) presented in chapter 3, §3.4.2 is a generative model of neural responses (and related electrophysiological responses) within the auditory cortex that rests on purely physiological (bottom-up) mechanisms of synaptic adaptation induced by the flow of incoming sounds (at the experimental timescale), which can only account for frequency-based associations between sounds. A more sophisticated (model-driven) approach can be found in Wacongne et al. (2012), where the authors proposed a generative model of neural activity based on the dynamics of synaptic currents within a network, with additional assumptions inherited from predictive coding. Indeed, this model involves reciprocal interactions within the hierarchy enabling a message passing conforming to a simple predictive coding implementation (precision of information is not accounted for). In addition, this model entails a top-down modulation of synaptic gains controlled by a memory unit at the top of a hierarchical structure. Critically, synaptic plasticity, modeled using time-dependent rules, is at the core of the model, that could underlie sequence learning, and the detection of unexpected sounds enabling updates within the memory unit. Likewise DCM with CMC, the computation of prediction errors is not accounted for by the model but these messages are as-

signed to the activity of specific neuronal populations. These are neurons receiving excitatory bottom-up and inhibitory top-down inputs. Using such approach, MMN-like neuronal activity could be successfully simulated, with modulation by experimental manipulations consistent with the experimental MMN literature. In a recent study, this model was augmented with synaptic adaptation (Wacongne, 2016) that could reveal the difficulty to disentangle top-down and bottom-up effects on synaptic dynamics as both were found to contribute to MMN-like responses. To date, these evolution models have been applied in the generative direction to provide simulated responses; their integration in a meta-Bayesian response model to enable inversion with real data has not been proposed yet.

Finally, the last study reported here is an attempt toward a full response model for mismatch responses (Lieder, Stephan, et al., 2013). The perceptual model employed in this study was designed with a three-level hierarchical structure with each level attributed to a particular cortical region (including two sources within the primary auditory cortex, and the IFG). Model inversion (resting on a VB scheme) was performed for each item of sound sequence, that led to trajectories of learning quantities. In particular, hierarchical prediction errors were assumed to contribute to the MMN responses: they were thus transformed into firing rates by means of a simple computational-to-neurophysiological mapping and the ensuing activities were subsequently projected onto a single fronto-central electrode according to the observation model specifications. The timecourse of the signal simulated at this electrode revealed a MMN-like component. Besides, the modulation of the MMN amplitude and latency with experimental factors (deviance magnitude and deviance probability) was consistent with the MMN literature. As discussed by the authors, this framework constitutes a proof of validity that synthetic response resembling the MMN and showing similar phenomenological properties can be obtained from the dynamics of Bayesian learning quantities. Although this study rests on predictions of a response model (not confronted to real electrophysiological responses), these results support the plausibility of predictive coding.

To sum up, it appears that despite convincing (but rare) efforts, there is a need for further advanced *"meta-response"* models, allowing to characterize precisely the neural correlates of Bayesian information processing in the brain from neuroimaging data.

4.5 Summary

This chapter focused on the existing methodological approaches that enable addressing the predictive coding account of mismatch responses from both a neurophysiological and a functional perspective. The reviewed studies confirm that the corresponding tools are reasonably ready to be employed. They rest on sophisticated Bayesian methodologies detailed here for a better understanding of the modeling analysis conducted in this thesis. Both DCM and learning model empirical findings remain scarce at the present day but they all appear consistent with predictive coding at their respective level (namely deviant-based neuronal dynamics resting on reciprocal cortical interactions for the former and learning-based trajectories providing better correlations with electrophysiological data than non-learning ones for the latter). Hence, they contribute to motivate further investigations, as it is clear that more empirical evidence is needed to improve our understanding of mismatch processing in the light of predictive coding. A relevant aspect indicated by different studies (for instance, Moran, Campo, et al., 2013) pertains to the use of

experimental factors to modulate deviance responses. Indeed, estimating the neurophysiological or the cognitive mechanisms related to such modulations can provide deep insights into the characterization of standard and deviant processing *per se*. Finally, although the fusion of both perspectives into a full response model is not yet available, conducting separate analyses using the two approaches could still be very informative to bridge the gap between them, provided that mechanistic hypothesis could be formulated for both approaches.

Part II

Experimental work

The first chapters aim at characterizing the current knowledge about the predictive coding view of mismatch responses. **Chapter 1** provided the basics of Bayesian inference, and recalled the relevance of a Bayesian framework to deal with information and uncertainty, typically encountered by both the observer (the subject or patient being studied) and the experimenter, in their respective observation process. **Chapter 2** introduced the Bayesian brain hypothesis and in particular its application to perception through the proposal of the free energy principle and generalized predictive coding. Generalized predictive coding considers the brain as approximating optimal Bayesian inference to infer the causes of its sensory inputs and therefore appears as a principled formulation of Helmholtz's view of perception. Bringing empirical evidence to this compelling framework was shown to be a great challenge and **Chapter 3** indicated how mismatch responses constitute a privileged way to that aim. Indeed, both the neural mechanisms and the functional role behind the MMN (and more generally deviance responses) remain unclear and predictive coding furnishes new hypothesis that could improve their understanding. This requires the Bayesian modeling approaches presented in **Chapter 4** to provide mechanistic insights that cannot be captured by traditional sensor-level analysis. To date, model-based findings corroborate this theoretical framework and despite the small number of studies published so far, existing ones have highlighted the relevance of both the *ERP-timescale* (the MMN in relationship with other deviance responses) and *experimental-timescale* (trial-by-trial) analysis to fully explore the neural and functional underpinnings of perceptual inference and learning.

Based on these reviews, it appears that much experimental and modeling work remain to be done to link deviance processing to a Bayesian inference scheme. Building upon previous findings, the objective of the present thesis was thus to further investigate the predictive coding account of mismatch responses, using advanced modeling approaches to refine both the neurophysiological correlates and the functional characterization of these brain components.

This methodological work was based on a single and original experimental study which we analyzed from different perspectives in order to test different hypotheses regarding the effect of contextual manipulation (predictability) on the implicit treatment of sound sequences. **Chapter 5** describes our oddball paradigm and the experimental conditions we used to investigate the predictive coding view of the MMN. This study was conducted using simultaneous EEG and MEG recordings as their complementarity is now widely acknowledged. This chapter reports sensor-level findings confirming that deviance responses at various latencies are attenuated by an increased predictability, following an expected reduction in prediction error. **Chapter 6** is dedicated to the characterization of deviance cortical generators that we addressed using recent advanced distributed procedures entailing group-level inference and EEG-MEG fusion. In addition to bringing empirical support to fused source reconstruction (in comparison to unimodal inversion), this analysis provided the architecture of the hierarchical network entering our subsequent modeling analysis. **Chapter 7** describes the two step DCM approach we performed to first establish the underlying neuronal network at play during deviance processing (under the twofold guidance of EEG and MEG), and then to identify the modulations of effective connectivity underlying the effect of predictability. We hoped here to elucidate the neural correlates of modulations of precision-weighted prediction error by predictability. Finally, **Chapter 8** describes our complementary computational approach aiming at identifying the ongoing implicit learning processes. This approach tries to explain trial-by-trial variations of evoked responses in the above identified

cortical network and enables comparing alternative models of the computation and the updating of precision-weighted prediction error. Importantly, the obtained results did corroborate our DCM findings.

Chapter 5

Effect of deviant predictability on mismatch responses

5.1 Objectives

As was explained in chapter 2, §2.3.3, predictive coding brings new hypothesis regarding how brain activity should be modulated by experimental manipulations. The contextual expectancy of a stimulus in particular deserves to be considered under the light of this framework which formally predicts reduced brain activity induced by fulfilling top-down predictions. Consistent fMRI findings were indeed reported in Summerfield et al. (2008) using predictable and unpredictable repetitions of visual stimuli. Regarding deviance processing, expected deviant should be associated to reduced surprise hence smaller MMN if this component were to reflect prediction errors. In a study addressing the conscious processing of auditory regularities, deviance was designed to induce local (simple repetition) and global (five-tone pattern repetition) violations using the so-called Local-Global paradigm (Bekinschtein et al., 2009). Such sound sequence embedded expected and unexpected local deviants but the modulation of mismatch responses by deviant expectancy was not specifically assessed because it was out of the scope of the study. The manipulation of the predictability of deviants appears relevant to test the validity of mismatch responses being such prediction errors. It was the main question of the present study. To that aim, we designed two frequency oddball sequences entailing different contextual manipulation of deviance expectancy. Our study was conducted with simultaneous EEG and MEG recordings. This chapter is dedicated to the analysis of sensor-level data and is organized as follows: first, we report the article that described the EEG analysis of deviance responses and their modulation by deviance predictability. We then present the results obtained with the sensor-level MEG data, and finally we report some attempts that we made to characterize the perceptual learning of the regularities taking place over the oddball sequences used in the present study.

5.2 EEG analysis - Article

Implicit learning of predictable sound sequences modulates human brain responses at different levels of the auditory hierarchy.

Front Hum Neurosci, 2015, 9:505

frontiers in Human Neuroscience

ORIGINAL RESEARCH
published: 16 September 2015
doi: 10.3389/fnhum.2015.00505

Implicit learning of predictable sound sequences modulates human brain responses at different levels of the auditory hierarchy

Françoise Lecaignard[1,2,3], Olivier Bertrand[1,2], Gérard Gimenez[3], Jérémie Mattout[1,2†] and Anne Caclin[1,2†]*

[1] Lyon Neuroscience Research Center, CRNL, INSERM, U1028 – CNRS, UMR5292, Brain Dynamics and Cognition Team, Lyon, France, [2] University Lyon 1, Lyon, France, [3] MEG Department, CERMEP Imaging Center, Lyon, France

Deviant stimuli, violating regularities in a sensory environment, elicit the mismatch negativity (MMN), largely described in the Event-Related Potential literature. While it is widely accepted that the MMN reflects more than basic change detection, a comprehensive description of mental processes modulating this response is still lacking. Within the framework of predictive coding, deviance processing is part of an inference process where prediction errors (the mismatch between incoming sensations and predictions established through experience) are minimized. In this view, the MMN is a measure of prediction error, which yields specific expectations regarding its modulations by various experimental factors. In particular, it predicts that the MMN should decrease as the occurrence of a deviance becomes more predictable. We conducted a passive oddball EEG study and manipulated the predictability of sound sequences by means of different temporal structures. Importantly, our design allows comparing mismatch responses elicited by predictable and unpredictable violations of a simple repetition rule and therefore departs from previous studies that investigate violations of different time-scale regularities. We observed a decrease of the MMN with predictability and interestingly, a similar effect at earlier latencies, within 70 ms after deviance onset. Following these pre-attentive responses, a reduced P3a was measured in the case of predictable deviants. We conclude that early and late deviance responses reflect prediction errors, triggering belief updating within the auditory hierarchy. Beside, in this passive study, such perceptual inference appears to be modulated by higher-level implicit learning of sequence statistical structures. Our findings argue for a hierarchical model of auditory processing where predictive coding enables implicit extraction of environmental regularities.

Keywords: mismatch negativity, auditory regularity, predictive coding, early deviance response, EEG, P3a

OPEN ACCESS

Edited by:
Klaus Gramann, Berlin Institute of Technology, Germany

Reviewed by:
Jeroen Stekelenburg, Tilburg University, Netherlands
Matthew G. Wisniewski, Air Force Research Laboratory, USA

***Correspondence:**
Françoise Lecaignard, Lyon Neuroscience Research Center, CRNL, INSERM, U1028 – CNRS, UMR5292, Brain Dynamics and Cognition Team, Centre Hospitalier Le Vinatier (Bâtiment 452) 95, Boulevard Pinel, Lyon, 69500 Bron, France
francoise.lecaignard@inserm.fr

† These authors have contributed equally to this work.

Received: *28 May 2015*
Accepted: *31 August 2015*
Published: *16 September 2015*

Citation:
Lecaignard F, Bertrand O, Gimenez G, Mattout J and Caclin A (2015) Implicit learning of predictable sound sequences modulates human brain responses at different levels of the auditory hierarchy. Front. Hum. Neurosci. 9:505. doi: 10.3389/fnhum.2015.00505

Introduction

Oddball paradigms involve sequences of a repeating (standard) pattern that sets up a regular environment, and infrequent (deviant) stimuli, which violate this regularity and subsequently elicit mismatch responses in the brain. They have been extensively employed in humans using non-invasive electrophysiology recordings, because of their ease of recording, their unique ability to

reveal mechanisms of perceptual inference and learning (Kujala and Näätänen, 2010), as well as their clinical relevance (Näätänen et al., 2012; Morlet and Fischer, 2014). The well-known mismatch negativity (MMN), first described in Näätänen et al. (1978), is observed in such paradigms and has been described in several sensory modalities although mostly studied in audition (for review, see Näätänen et al., 2007). A large literature is dedicated to the functional interpretation of the MMN and several models, resting either on psychological concepts, on computational frameworks or even on neural adaptation processes have been proposed [for review, see Näätänen et al. (2007) and Garrido et al. (2009b)]. Adaptation refers to a decrease of neural responsiveness after several repetitions of a stimulus, and is widely acknowledged to contribute to the difference in responses to standards and deviants. A considerable number of MMN findings argue against the adaptation model (that implies a full account of the MMN by adaptation effects) and suggest that this component reflects an automatic detection of change in the acoustic environment, with strong support to the MMN as the output of a comparator between observed and expected sensory inputs (Näätänen et al., 2007). In the current study, we were interested in recent theories based on a predictive coding scheme that have been proposed to account for the generation of the MMN (Friston, 2005) [see also Winkler and Czigler (2012) for a review of findings compatible with this account]. These theories rest upon a hierarchical organization of the brain, wherein predictions regarding incoming inputs are conveyed to lower levels by top-down messages, while bottom-up prediction errors reflecting mismatch between observations and predictions are sent back to higher levels. In this view, the MMN reflects a prediction error that triggers the update of predictions by means of message-passing between the different levels of the auditory hierarchy (Friston, 2005).

Importantly, predictive coding models of mismatch responses do not entail a single prediction regarding incoming inputs but multiple ones, generated at different levels of the hierarchy (Friston, 2005). Precisely, these predictions pertain to the physical attributes of sound and to the statistical dependencies within the sound sequence. Accordingly, prediction errors, hence likely the MMN, should be affected by at least three factors: (1) the acoustic separation between the predicted and observed stimuli (also referred to as the deviance magnitude), (2) the variability of the acoustic features, and (3) the sequence predictability, deriving from statistical regularities. Factor (1), deviance magnitude, has already been proved to modulate the MMN. For instance, Tiitinen et al. (1994) showed that for frequency deviation spanning above 2% of the standard frequency, the larger the deviation, the larger the MMN amplitude. The last two factors affect prediction error through modulations of sound predictability, by influencing either the predictability of the sound's acoustic features [factor (2)], or the predictability of the stimulus category [standard or deviant, factor (3)]. Importantly, predictability may influence both the content of the prediction and its precision or confidence. The two evolve with learning and could modulate the MMN amplitude, provided that the MMN reflects a precision-weighted prediction error (Friston and Kiebel, 2009). Consequently, we hypothesized that the MMN amplitude should be reduced as the occurrence of the deviant stimulus becomes more predictable.

In the two following sections, we review the findings describing effects of above-defined factors 2 and 3 on the MMN amplitude. It reveals that they have been rarely studied so far, probably because of the methodological difficulties to disentangle those effects from those of deviance magnitude. Yet, validating the above hypothesis is required in order to assess the predictive coding perspective on the MMN and to refine our functional understanding of this widely used electrophysiological marker. The present study was carefully designed to overcome methodological caveats and specifically observe the effect of sequence predictability on the MMN.

Effect of Acoustic Feature Variability on the MMN

Among the few studies that investigated the effect of predictability on the MMN, the majority manipulated the variability of the acoustic features of standard stimuli. In Daikhin and Ahissar (2012), the authors used a frequency oddball sequence with variable standard frequency belonging to a uniform distribution with a 2% deviation. Compared to a fixed standard condition, the authors found no significant difference in average responses to standards but a reduced MMN. This suggests that conditions with jittered standards yield a blurred representation of the standard stimulus, producing a less precise prediction and hence weaker responses to deviance. More recently, larger deviations were used (Garrido et al., 2013), with sequences of sounds whose frequencies were drawn from either a narrow or a broad Gaussian distribution (mean frequency of 500 Hz with standard deviations of 250 and 1500 Hz, respectively). Outlier sounds elicited an MMN-like response, which was reduced in the case of the broad distribution. This confirms the ability of the brain to extract statistical rules from sound sequences and gives strong support to the existence of predictions of future events that would be weighted by their inferred precision.

However, since these studies manipulated the predictability of the standards in ways that inherently involve changes in the acoustic parameters, the observed results might be confounded with deviance magnitude and adaptation effects (induced by refractoriness) that are likely to differ between conditions.

Effect of Sequence Predictability on the MMN

Sequence (or sound category) predictability refers to rules that define the statistical dependencies of items within the sequence. Rules are usually categorized into simple (local) ones resting on short time-scale dependencies and complex (abstract or global) ones generating larger time-scale regularities or contingent relations. The violation of the latter also elicits a MMN (in both cases of passive and active paradigms) and has largely been described in the literature (for review, see Näätänen et al., 2010). Passive studies used the MMN as a marker of rule violation in order to reveal fairly high-level implicit learning processes (see for instance Bendixen et al., 2008; Todd et al., 2013). They were, however, not designed to test the effect of sequence predictability on the MMN *per se*.

Deviant predictability should be distinguished from deviant probability. The latter refers to the ratio of deviant events within the sequence, irrespective of its temporal structure, while the former refers to the statistical nature of the temporal sequence, irrespective of deviant occurrence frequency. Some studies have manipulated the deviant probability in order to measure its effect on the MMN (Sams et al., 1983; Sato et al., 2000). In our study, we manipulated deviant predictability only, which avoids the confounding effect of refractoriness inherent to the manipulation of deviant probability (i.e., varying the number of standards preceding a deviant).

To date, only a couple of studies have compared MMN responses elicited by unpredictable sequences (embedding *unpredictable* deviants) and predictable ones (embedding *predictable* deviants). In Scherg et al. (1989), a fully predictable sequence (*one frequency deviant every fifth tone*) was compared with an unpredictable one with the same global deviant probability ($p = 0.2$). The authors found no significant effect of the predictability manipulation on the MMN amplitude. They hypothesized that this result was compatible with initial findings (and widely confirmed since) suggesting that the MMN derives from an automatic process independent of participant's attention (Näätänen et al., 1978, 2010). However, using the same paradigm but with different temporal characteristics, Sussman et al. (1998) and Sussman and Gumenyuk (2005) found a disappearance of the MMN in the predictable condition, which the authors interpreted as an automatic perceptual effect of tone grouping that could only occur in the predictable condition. However, as judiciously pointed by Fishman (2014), this effect could also be attributable to predictability. Importantly though, none of these studies rigorously controlled for adaptation effects as the number of standard preceding a deviant differed between the regular and irregular conditions. Others studies proposed oddball sequences embedding predictable deviants (Jankowiak and Berti, 2007; Bekinschtein et al., 2009) but their aim was not to measure the effect of predictability on mismatch responses. In some respect, although using a very different setting, a few studies already reported MMN-like responses that were modulated by the predictability of musical sequences. For instance, in Brattico et al. (2006), out-of-key tone responses suggest that less probable transitions are processed like deviants. In Vuust et al. (2009), subtle rhythmic violations were shown to induce larger magnetic MMN-like responses in musical experts compared to novices, whereas large violations induced responses in both groups. In line with those studies, the current experiment aims at generalizing those findings by testing the effect of predictability in isolation of deviance magnitude and independently of acquired skills over the lifespan.

From the existing literature briefly reviewed here, it is clear that empirical findings are compatible with the predictive coding view of the MMN. Nevertheless, direct evidence is missing and finely controlled sequence predictability appears as a good candidate to resolve this issue. As reported above, little is known on the effect of sequence predictability on the MMN, since it has never been studied genuinely. The widely acknowledged automaticity of the MMN has possibly inclined to the worthlessness of searching for any predictability-driven modulation. Today, recent (computational) theories of brain function (Friston, 2005; Winkler and Czigler, 2012) rather suggest that sequence predictability should affect deviance responses as follows: the more predictable the occurrence of a deviant sound, the finer the prediction, hence the smaller the prediction error and the smaller the MMN amplitude. Therefore, we used a passive oddball paradigm with unpredictable and predictable sound sequences differing by the transitional probabilities between sounds within each sequence type. The strict conservation of the acoustic properties of the sequence between conditions was achieved by means of a statistical structure determined over a relatively long time range in the predictable condition. Our design also includes the appropriate control for adaptation effects. Furthermore, we used small deviance magnitudes in a passive oddball paradigm, in order to limit automatic attention-orienting processes. These processes are typically reflected by the N2b-P3a complex (brain orienting response) following the MMN under specific condition of attention (Näätänen et al., 1982; Morlet et al., 2014). As mentioned above, the ability of the brain to encode implicitly large time-scale regularities has been indirectly demonstrated in several MMN studies, therefore we expected that participants would learn the statistical rule in the predictable condition. We hypothesized that predictable deviants would elicit reduced deviance responses. Conversely, in the absence of any implicit (or explicit) learning of the rule, no difference between conditions would emerge. Additionally, as recent studies point to earlier deviance responses than the MMN (Escera et al., 2014), we used an analysis strategy that did not make any assumptions regarding the temporal specificity of predictability effects.

Materials and Methods

Participants

Twenty-seven adults (14 female, mean age 25 $\pm$ 4 years, ranging from 18 to 35) participated in this experiment. All participants were free from neurological or psychiatric disorder, and reported normal hearing. One participant had professional musical education and has been excluded from the analysis for he did not respect the instruction to ignore the sounds. 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).

Stimuli and Sound Sequences

The large use of frequency deviance in MMN studies encouraged us to choose this acoustic feature to test the prediction error model of the MMN. However, undesirable adaptation effects are of particular importance in this particular case because of the tonotopic organization of the auditory pathways. They would in particular impact the amplitude of exogenous event-related potentials (ERPs) in the P50 and N1 wave latency range. We therefore introduced a supplementary condition in order to control for such adaptation effects, using intensity deviance

(see below). Overall, three kinds of sequences were used: (1) an unpredictable sequence with frequency deviance: UF, (2) a predictable sequence with frequency deviance: PF, and (3) an unpredictable sequence with intensity deviance: UI. Note that we did not considered a predictable sequence with intensity deviance for the sake of experiment length and also because the feature specificity of the prediction error model of the MMN is beyond the scope of the current study. All the sequences shared the same deviant probability ($p = 0.17$).

Sound duration was 70 ms (including 5 ms rise-time and 5 ms fall-time) and the stimulus onset asynchrony (SOA) was fixed to 610 ms. Two different frequencies ($f_1 = 500$ Hz and $f_2 = 550$ Hz) and two different intensities ($i_1 = 50$ dB SL (*sensation level)* and $i_2 = 60$ dB SL) were combined to define the four different stimuli that were used across conditions. In this (passive) study, we carefully chose the deviance magnitude in the frequency sequences in order to satisfy a trade-off between eliciting a deviance response, on the one hand, and both minimizing refractoriness effects and avoiding to attract the subject's attention, on the other hand. Therefore, although even smaller deviance have been previously used (Sams et al., 1985), we used a 10% deviance which falls in the lower range of recently implemented deviance magnitudes [e.g., 8% in (Daikhin and Ahissar, 2012), 10% in (Schwartze et al., 2013), 23% in (Recasens et al., 2014), 30% in (Grimm et al., 2011), and 50% in (Todd et al., 2014)].

To design the predictable sequences (**Figure 1**), we did not use a fixed number of standards between two deviants as in Scherg et al. (1989), because this cannot be mirrored in the unpredictable sequence without inducing different refractoriness effects. This issue could be avoided by the construction of a statistical structure unfolding over a larger time-scale. Precisely, the rule that we designed increments the number of standards progressively within a cycle: it starts with one deviant after two standards, followed by one deviant after three standards and so on until one deviant after eight standards. From now on, a chunk with *n* standards will refer to a series of *n* standard sounds ending with a deviant stimulus (*n* ranging from 2 to 8). The 42-tone cycle, composed of seven incrementing chunks, was repeated 16 times in the sequence, thus leading to a total of 560 standards and 112 deviants. For the unpredictable sequences, each cycle was shuffled so as to permute the order of the seven chunks with the constraint that no chunk with *n* standard was preceded or followed by a chunk with either *n−1* or *n+1* standards. Additionally, the transition between two cycles was such that no successive chunks with *n* standards could occur. Altogether this randomization allowed to (1) avoid any global rule to emerge in the unpredictable sequence and (2) have exactly the same number of chunks with *n* standards in predictable and unpredictable conditions. Note that the number of deviants presented at a 2–3 chunk timescale may differ between UF and PF (for instance, the set of 16 sounds that precede a "chunk of 8 standards" deviant comprises exactly one deviant in PF and two deviants on average in UF) but the fact that adaptation saturates rapidly [2–3 standard repetitions,(Demarquay et al., 2011)] led us to assume that this particularity did not introduce any significant adaptation effect difference between PF and UF, in the current analysis that we conducted with standards just preceding deviants.

Each sequence type (UF, PF, UI) was delivered twice in separate blocks resulting in 224 deviants in each condition. For each type of deviance (frequency or intensity), the sound property used as the standard (e.g., for frequency deviance, f_1) for the first block was used as the deviant for the second (reverse) block. The irrelevant feature was constant within a block but changed between the two reverse blocks [e.g., for frequency deviance, first block with properties (f_1,i_1) for standards and (f_2,i_1) for deviants, and reverse block with properties (f_2,i_2) for standards and (f_1,i_2) for deviants]. The order of the six resulting blocks was counterbalanced between participants with the constraint that no successive sound sequences of the same kind could be delivered. Additionally, in order to avoid any bias of perceptive association between frequency and intensity, half of the participants received the associated properties (f_1,i_1) and (f_2,i_2) as standards whereas the other half received the pairs (f_1,i_2) and (f_1,i_2). Altogether these acoustical matching constraints on stimuli and sequences were applied to ensure comparisons between conditions with an optimal control for undesirable effects of specific acoustic properties.

All stimuli were delivered using Presentation software (Neurobehavioral Systems, Albany, CA, USA).

Procedure

The present study was conducted using simultaneous EEG and MEG recordings, although the MEG data will not be analyzed here. Participants were seated upright in a comfortable armchair in a sound-attenuated, magnetically shielded recording room, at a 1 m distance from the screen. Sounds were presented binaurally through air-conducting tubes using Etymotic ER-3A foam earplugs (Etymotic Research, Inc., USA). Participants were instructed to ignore the sounds and watch a silent movie of their choice with subtitles. Before recordings, participants' sound detection thresholds using the sound with (f_1,i_1) characteristics were determined for each ear, and the level was adjusted so that the sounds were presented at 50 dB SL (i_1) or 60 dB SL (i_2) with a central position (stereo) with respect to the participant's head. Each of the six blocks lasted 7 min resulting in a total recording time of ∼50 min, including short breaks between sequences. At the end of the experiment, participants were asked to report to which extent they had been following the instruction to ignore the sounds and whether they had noticed the different sound attributes (e.g., "*Did you notice anything in particular about the sounds?"*) and sequence temporal regularities (e.g., "*Did you notice that some sounds were less frequent than others?", "Did you notice any regularities in sound presentation?"*).

EEG Recordings

EEG recordings were carried simultaneously to MEG ones using the EEG recording system provided with the MEG equipment (275-channel whole head system, CTF-275 by VSM Medtech Inc.). EEG data were collected from 63 electrodes (including the two mastoids) whose locations were defined by the 10–15 extension of the international 10–20 system. Reference electrode and ground electrode were placed on the tip of the nose and

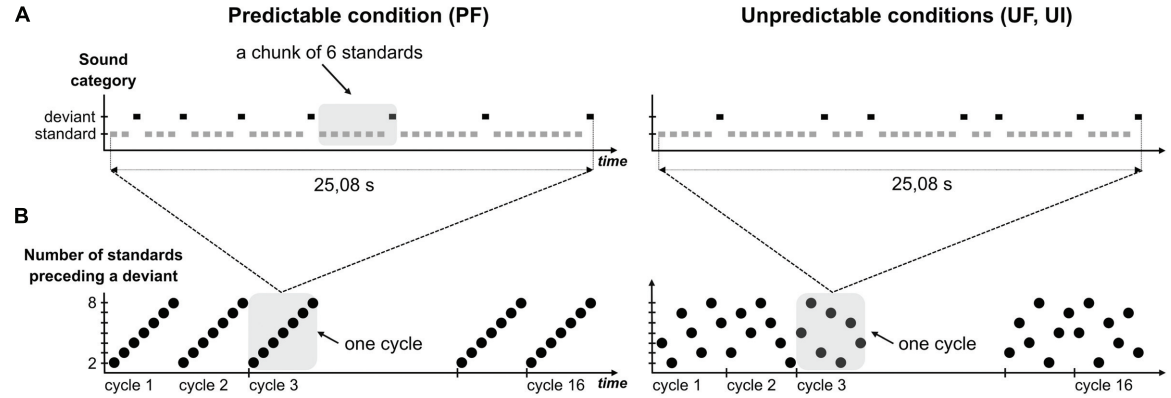

FIGURE 1 | Experimental design. (A) Schematic view of a complete cycle in predictable (left) and unpredictable conditions (right). Rectangles symbolize single tones with standards and deviants colored in gray and black, respectively. Sound duration was 70 ms with stimulus onset asynchrony (SOA) set to 610 ms. In every condition, each cycle entails seven deviants, each of them being preceded by a number of standards ranging from 2 to 8. A chunk of *n* standards corresponds to n+1 tones (*n* consecutive standards and the following deviant), as illustrated by the shaded area in condition PF. Chunks are sorted by their size in predictable condition, whereas these are shuffled in unpredictable ones. **(B)** Variation of the size of chunks (black circle) within cycles, over sound sequence in predictable (left) and unpredictable conditions (right). Each sequence is composed of 16 cycles and examples of shuffled cycles are presented for unpredictable conditions. Shaded areas delineating one cycle in both sequence types highlight their difference with respect to sound predictability.

left shoulder, respectively. One bipolar EOG derivation was recorded from two electrodes placed on the supra-orbital and infra-orbital ridges of the left eye. Throughout the recordings, impedances were below 15 kΩ. Signal was amplified, band-pass filtered (0.016–150 Hz), digitized (sampling frequency 600 Hz) and stored for off-line analysis. Head position relative to the MEG sensors was acquired continuously (continuous sampling at a rate of 150 Hz) using coils placed at three fiducial points (nasion, left and right preauricular points).

Data Preprocessing

The software package for electrophysiological analysis (ELAN[1]) developed at the Lyon Neuroscience Research Center (Aguera et al., 2011) was used for ERP computation and statistical analysis.

EEG and MEG data were preprocessed independently but for the sake of a combined analysis, which will be reported in a further study, we only used time epochs that survived the procedures applied for artifact rejection for both techniques. A total of 5 participants out of 27 had to be excluded from the group. For two participants, raw MEG recordings were contaminated by ferromagnetic artifacts caused by metallic elements, which created a temporally stationary artifact at the participant's respiratory frequency. One participant's EEG data had a very bad SNR. One participant had individual MR images that disclosed a ventriculomegaly. Finally, as mentioned above, one participant did not ignore the sounds as instructed but counted them leading to an explicit detection of the predictable rule in PF sequences. Preprocessing of raw data for the remaining 22 participants comprised the following successive steps: (1) an initial rejection of data segments corrupted by head movements above 15 mm within each sequence was automatically performed (in prevision of future MEG data analysis), (2) three stop-band filters centered on 50, 100, and 150 Hz (with bandwidth of ±2 Hz) were applied to get rid of the power line artifact in the EEG data, (3) using EEGlab routines[2], an independent component analysis (ICA) correction for ocular artifacts was achieved (largest possible time windows – free from artifacts from all origin but ocular – were selected from continuous stop-band filtered data to derive ICA components) for all participants but one for whom ICA correction failed to improve the SNR of EEG and MEG data, (4) individual recordings were automatically inspected from −200 ms to 410 ms with respect to the onset of each sound; trials with signal amplitude range exceeding 2000 fT for MEG data and 150 μV for EEG data over the 610 ms time-window at any sensor were excluded from the analysis (for the participant whose data did not receive any ICA correction, a threshold of 100 μV was used for the EOG signal range), (5) a 2–45 Hz band-pass digital filter (bidirectional Butterworth, fourth order) was applied to EEG and MEG data. It should be noted here that most MMN studies rely on filtered data with lowpass cutoff frequency lower than 45 Hz (20 or 30 Hz are commonly used), leading to *smoother* baselines and ERPs.

Event-Related Potential (ERP) Computation

Data collected within the first 20 s of each block was excluded from averaging to ensure that no transitory effect could bias the ERPs. Responses to standards just preceding a deviant and to deviants were considered for averaging within an epoch of 610 ms including a pre-stimulus period of 200 ms. Baseline correction was achieved by subtracting the mean value of the signal during the pre-stimulus period. ERPs for each stimulus type (standard and deviant) were first computed per block. The two reverse blocks for each condition (UF, PF, and UI) were then pooled by averaging corresponding ERPs. Difference response (also referred

[1]http://elan.lyon.inserm.fr

[2]http://sccn.ucsd.edu/eeglab/index.html

to as deviance response) was obtained by subtracting the standard ERP from the deviant one.

Statistical Analysis

We applied permutation tests based on a *t*-statistic at the group-level at each sample of each electrode of the ERP time series in bandwidth 2–45 Hz, correcting for multiple comparison in the temporal dimension (Blair and Karniski, 1993; Besle et al., 2008). For each test, we ran 100,000 permutations by randomly redistributing the ERPs of the two conditions to be compared. We tested for (1) an effect of deviance in the three conditions (i.e., standard vs. deviant in UF, UI, and PF), (2) an effect of predictability (i.e., PF vs. UF) in difference, deviant and standard responses, (3) an effect of acoustic features (i.e., UF vs. UI) in the difference, deviant, and standard responses. Finally, since the first analysis above revealed a significant effect of deviance at both early and late latencies as well as a smaller effect at the P3a latency, we also conducted further analysis in tests (2) and (3) in three local time windows [0, 80] ms, [100, 210] ms and [250, 350] ms. Hence, permutation tests were run both on the entire time series [−200, 410] ms for each effect of interest (1, 2, 3) and on specific local time windows for (2, 3).

Adaptation Effect Characterization

To isolate the effect of predictability on genuine mismatch responses in conditions UF and PF, we had to characterize the effect of adaptation. Our experiment was designed to minimize this effect and we hypothesized that, if present, it would be the same in the UF and PF conditions. To this aim, we used a small deviance magnitude to reduce refractoriness effect as much as possible and imposed strong acoustical constraints on sound sequences such as a strict balancing of the number of standards preceding a deviant across conditions. Moreover, we introduced a third condition using an intensity deviance (condition UI) as a control condition for these possible adaptation effects. Adaptation effects for intensity deviance cannot be ruled out, although their existence remains rather controversial [but see Bilecen et al. (2002)]. We assumed that the MMN to intensity would not be contaminated by refractoriness, or at least to a far smaller extent than the MMN to frequency. Furthermore, we carefully matched the intensity and frequency deviance magnitude thanks to a prior behavioral deviance detection task so that frequency and intensity MMN would have similar amplitudes. Consequently, comparison between UI and UF *difference* responses should help characterizing (in the temporal and spatial dimensions) the undesired adaptation effects possibly entering UF and PF *difference* responses.

Control for Possible Filtering Confounds in Early Effects

As early effects were revealed by statistical tests in both the deviant vs. standard and the predictable vs. unpredictable comparisons, additional analysis were needed to control for their validity. As explained in Acunzo et al. (2012), the bidirectional low-pass filter that we applied on our data may have generated artifactual responses preceding the sharper deflections of the ERPs, namely the N1 and MMN components. In order to test whether our early effects were of such artifactual origin, we repeated the whole ERP analysis (using the statistical analysis described above) on unfiltered data to control for any bias induced by filtering (particularly low-pass filtering). These unfiltered data correspond to the data recorded by the acquisition system (0.016–150 Hz acquisition bandwidth) with further application of three stop-band filters and ICA correction as described in the Data processing section. Trials averaged for both ERP types (standard and deviant) were those retained for the analysis in the 2–45 Hz bandwidth. Note that this complementary analysis also allows to check that the 2 Hz high-pass filter that we used for the main analysis did not obscure some differences between conditions, e.g., in the very low frequencies.

Results

Post-experimental debriefing with the 22 participants whose data were retained for statistical analysis (11 female, mean age: 25 ± 5 years, ranging from 18 to 35) revealed that 15 of them noticed that sounds could take different intensities, 12 noticed that sounds could take different frequencies and nine noticed that some sounds were less frequent than others. Critically, none of them reported to have inferred the global rule of the PF sequence. Given our design, this implies that any difference between deviance responses in UF and PF reflects implicit learning of a global rule in PF.

On average per subject, the number of retained standard trials (standard sounds just preceding a deviant sound) was 177 ± 16 for the UF sequence, 174 ± 18 for the UI sequence and 172 ± 17 for the PF sequence. Similarly for deviants, the number of retained trials was 174 ± 17 for the UF sequence, 172 ± 22 for the UI sequence and 172 ± 21 for the PF sequence.

Multiple Deviance-Specific Responses

Figure 2 displays ERPs (with bandwidth 2–45 Hz) at electrodes Fz and TP9, for the standard, deviant, and difference responses, in each experimental condition. It also shows the statistically significant patterns in the deviance responses and the corresponding scalp topographies at relevant latencies. In every condition, the standards just preceding a deviant elicited a N1 component peaking around 95 ms, associated with a negativity distributed over fronto-central electrodes and followed by a fronto-central P2 component peaking around 155 ms. As shown on **Figure 2**, testing for deviance effects revealed three significant time-windows for the unpredictable sequences and two for the predictable one: an early time-window (within 70 ms after stimulus onset) for conditions UF and UI, and for the three conditions, we could detect a MMN and a P3a.

At early latencies, larger responses were elicited with deviants in condition UF compared to standards, leading to a positive difference response spanning from about 10 to 90 ms over the frontal and central areas. It was confirmed statistically significant from 11 to 28 ms at six adjacent electrodes located in left fronto-central area (−0.2 and 0.6 μV at Fz at 20 ms for standards and deviants, respectively). In condition UI, the deviant response

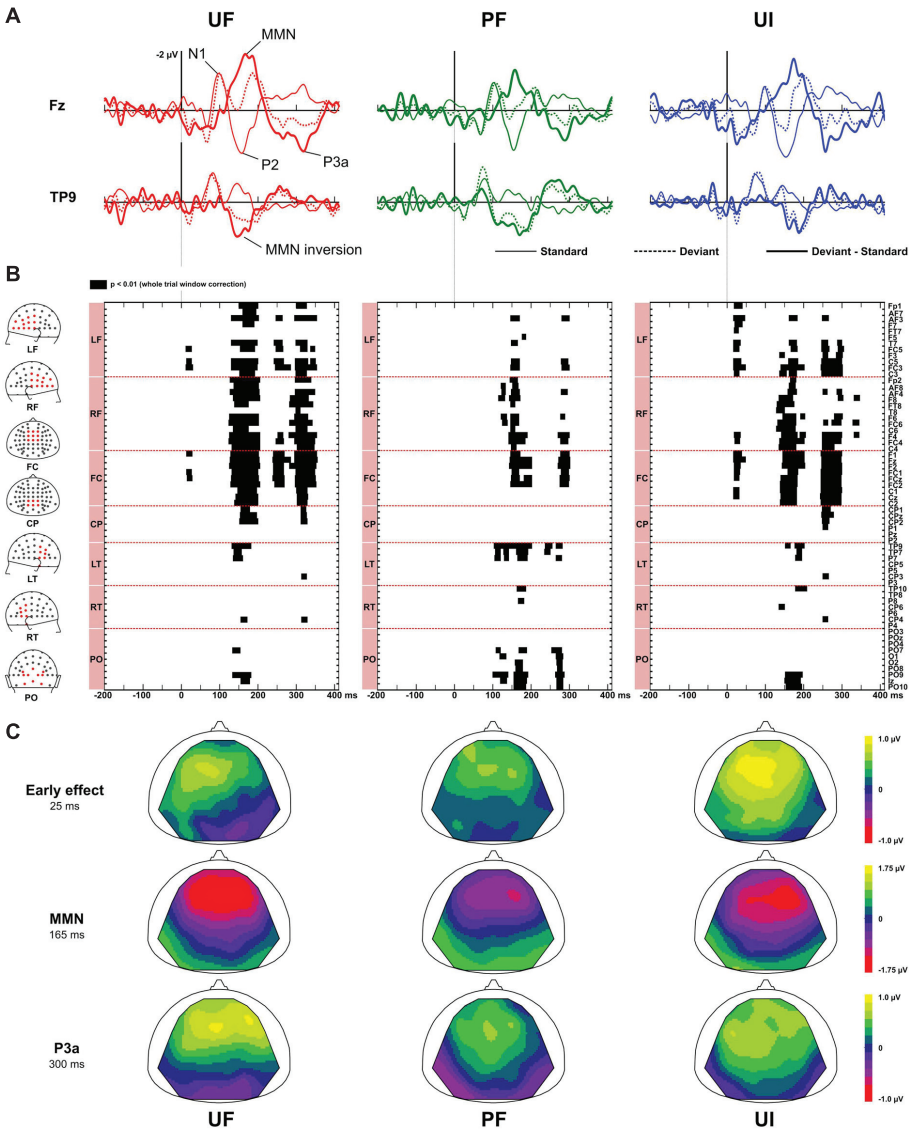

FIGURE 2 | Deviance effects. (A) Grand-average ERPs ($n = 22$ participants) elicited by standards just preceding a deviant (solid line), deviants (dotted line) and difference responses (bold solid line) at electrode Fz and TP9 in bandwidth 2–45 Hz for condition UF (left column), PF (middle column), and UI (right column). Main components in standard (N1, P2) and difference responses [mismatch negativity (MMN), P3a] are shown for condition UF. **(B)** Statistical maps obtained with non-parametric tests ($n = 100{,}000$ permutations) when comparing standard and deviant responses, at each electrode and each latency of the whole trial. Three intervals of significance were revealed for unpredictable sequences (UF, UI) at early latencies, and at the latency of the MMN and the P3a whereas only two were observed for condition PF at the latency of the MMN and P3a. Electrodes are sorted by spatial clusters (left column, from top to bottom: LF, left frontal, RF, right frontal, FC, fronto-central, CP, centro-parietal, LT, left temporal, RT, right temporal, PO, parieto-occipital). **(C)** Scalp topographies of the grand-average difference ERP, at the early effect (left column), the MMN (middle column) and the P3a (right column) latencies, for each condition. The MMN significant (positive) inversion is visible in each condition. Similarly, early deviance effect in condition UF and UI also entail a (negative) inversion but this does not reach significance.

was very similar to the one in UF, thus leading to very similar difference responses (deviant – standard) in those two conditions. Statistical analysis for UI revealed a significant interval occurring from 16 to 38 ms on left frontal and fronto-central areas. On the contrary, in condition PF, no significant effect was found at this early latency range. Because at this early latency there is an overlap of slow components (such as the P50) and fast Middle Latency Responses (MLR), we ran a complementary analysis with two different filtering (2–15 and 15–45 Hz) to further characterize this deviance effect. As shown on **Figure 3**, statistical analysis in the bandwidth 2–15 Hz confirmed the significant early deviance effect measured in UF (from 13 to 58 ms) whereas statistical tests in the bandwidth 15–45 Hz did not reveal any significant effect. A similar pattern was observed for condition UI (data not shown). Altogether, these results suggest that early deviance effects measured here in UF and UI pertain to a slow component at the latency of the P50 and do not concern the peaks of the MLR *per se*.

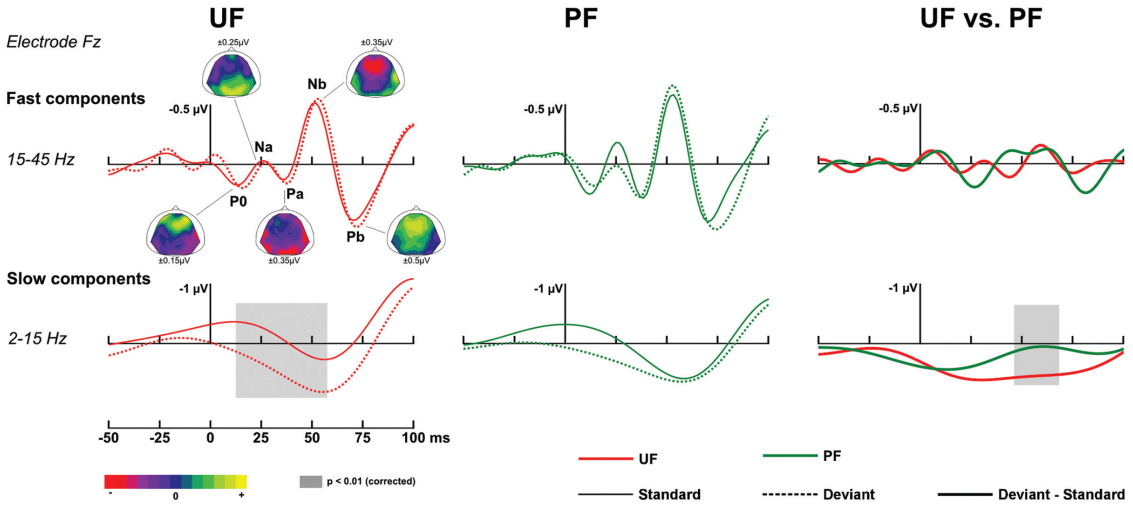

FIGURE 3 | Early responses. Traces at electrode Fz in the time interval [−50, 100] ms, with original [−200, 0] ms baseline correction. ***Fast components, bandwidth 15–45 Hz*** (top row). Grand-average ERPs elicited by standards just preceding a deviant (solid line) and deviants (dotted line) for condition UF (left column) and PF (middle column). Data were re-referenced to the average of both mastoids to facilitate the identification of Middle Latency Responses (MLR) components. Fast MLR components are indicated for condition UF with corresponding scalp topographies (from standard ERPs, with original nose reference allowing for the visibility of temporal polarity inversion) at the latencies 13, 26, 36, 50, and 68 ms for P0, Na, Pa, Nb, and Pb, respectively. Right column: Grand-average ERPs corresponding to difference responses (bold lines) for condition UF and PF. ***Slow components, bandwidth 2–15 Hz*** (bottom row).

In the MMN latency range, difference response in condition UF showed a typical MMN peaking around 165 ms, with large negativity over the frontal electrodes (−1.9 µV at Fz) combined with a positivity at the mastoids (the MMN inversion), with both deflections ending at the same latency. A similar difference response was observed in condition UI. The emergence of the MMN was statistically significant from 125 to 205 ms over 33 fronto-central electrodes and mastoids for UF, and from 128 to 205 ms over fronto-central electrodes, mastoids and occipital electrodes for UI. In condition PF, the difference response revealed the MMN inversion starting around 100 ms over the parieto-occipital areas, followed by the MMN *per se* (−1.4 µV at Fz), peaking at about 156 ms with a large negativity over frontal electrodes. Statistical tests confirmed the emergence of the MMN inversion (from 105 to 200 ms over mastoid and occipital electrodes) and of the MMN proper (from 120 to 200 ms over fronto-central electrodes and parieto-occipital electrodes). In all three conditions, the MMN inversion ended at the same latency than the frontal negativity deflection, suggesting that the N2b component, which does not invert in polarity at the mastoids, was negligible if any.

Finally, in the P3a latency range, a large positive deflection at fronto-central electrodes could be seen for difference responses of all conditions. These typical P3a components were maximal at around 316, 295, and 290 ms for UF, UI, and PF, respectively (with corresponding peak amplitude at Fz: 1.4, 0.8, and 1.0 µV for UF, UI, and PF, respectively). For condition UF, the emergence of the P3a was statistically significant from 238 to 270 ms over 12 frontal and fronto-central electrodes, and from 295 to 355 ms over 31 fronto-central and centro-parietal electrodes (including Fz, FCz, Cz, and CPz). Similarly, for condition UI, emergence was significant from 245 to 303 ms over 26 fronto-central and centro-parietal electrodes (including Fz, FCz, Cz, and CPz). For condition PF, statistical significance was measured from 265 to 281 ms over nine temporal and parieto-occipital electrodes (including TP9, P0z, and Iz), and from 280 to 303 ms over 13 frontal and fronto-central electrodes.

Predictability Modulates the Early Deviance Response, the MMN and the P3a

Figure 4 displays difference responses for conditions UF and PF at electrode Fz, as well as scalp topographies of the double difference waveforms (UF difference response – PF difference response). The effect of predictability was first assessed by comparing the difference responses obtained with the predictable and unpredictable sequences (PF vs. UF). Second, in order to disentangle the relative contribution of standard and deviant stimuli, we further assessed the effect of predictability on those two responses, separately.

Difference responses (**Figure 4**) differ as early as around 35 ms due to a weak (non-significant) deviance response measured in PF whereas a large significant fronto-central positivity was measured in UF (see above). It was confirmed significant from 55 to 65 ms on 13 electrodes, with more positive potentials in UF compared to PF (at 60 ms, 0.6 and −0.04 µV at Fz for UF and PF, respectively). Moreover, statistical analysis in the bandwidth 2–15 Hz revealed a significant effect from 46 to 68 ms (14 electrodes). No significant effect was found in the 15–45 Hz frequency band (**Figure 3**). Following this early effect, the scalp topography of the double difference (**Figure 4**) shows that the MMN peak is larger in the UF condition than in the PF one (from 163 to 190 ms over 15 fronto-central electrodes). We also observed a tendency for the MMN inversion in the PF condition to start earlier than in the UF condition (from about

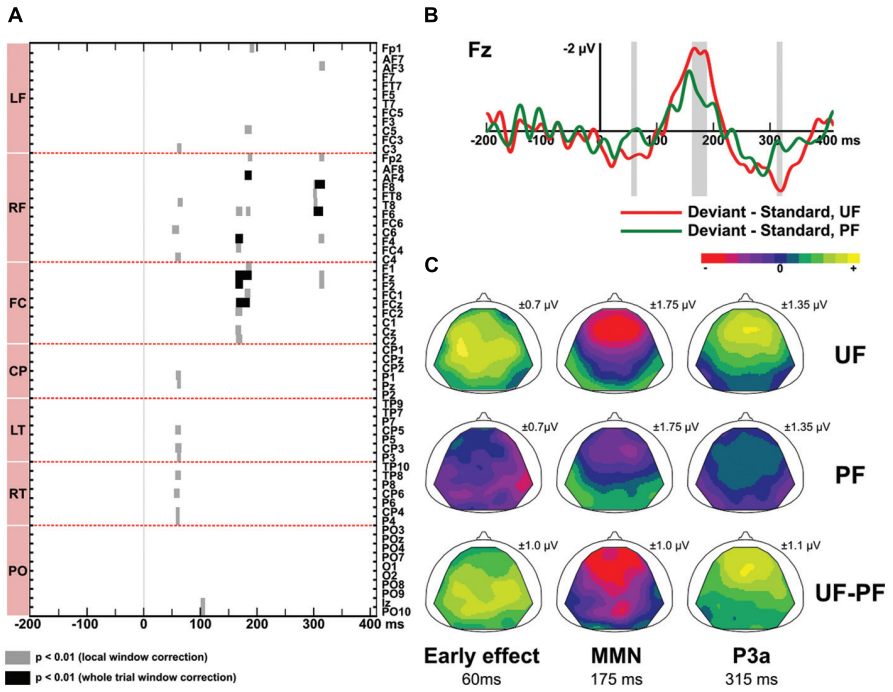

FIGURE 4 | Predictability effect (UF vs. PF). (A) Statistical maps of the permutation tests comparing difference responses between condition UF and PF, at each electrode and each latency of the whole trial. Black and gray areas indicate significant differences ($p < 0.01$) resulting from whole trial [−200, 410] ms and local tests, respectively. Results revealed three intervals of significant difference: at early latencies (13 electrodes), at the latency of the MMN (15 electrodes) and at the latency of the P3a (7 electrodes). **(B)** Grand-average ERPs elicited by difference responses at electrode Fz in bandwidth 2–45 Hz for condition UF (red) and PF (green). Shaded areas display the windows of statistical significance (at any electrode). **(C)** Scalp topographies of the difference responses in bandwidth 2–45 Hz, at the latency of the predictability effect, for the early effect (left column), the MMN (middle column) and the P3a (right column), in conditions UF and PF. The range of voltage values used for the color scale is mentioned for each map.

100 to 130 ms) and to be enhanced at parieto-occipital electrodes from about 150 to 210 ms, but these effects were not statistically significant. Finally, the statistical analysis also revealed a larger P3a component in UF compared to PF (at 315 ms, 1,4 and 0.3 μV for UF and PF, respectively at Fz), with significance spanning from 310 to 320 ms over seven electrodes (Fp2, AF3, Fz, F2, F4, F6, F8).

In response to deviants, permutation tests confirmed that more positive potentials were recorded in UF compared to PF in the early latency range (at 65 ms, 1.1 and 0.7 μV for UF and PF, respectively at Fz), with significance spanning from 58 to 72 ms over 18 fronto-central and left centro-parietal electrodes. Moreover, the negative deflection following the N1 was significantly larger in UF from 178 to 190 ms at electrodes F1, F3, Fz, FC1, FCz, and FC3 (at Fz: −1.2 μV at 185 ms for UF, and −0.8 μV at 205 ms for PF). These two effects observed for the deviant thus mirrored those observed in the difference wave tests. At the latency of the P3a, no significant difference between UF and PF could be measured.

In response to standards, no significant effect of sequence predictability could be observed in the ERPs of standards just preceding a deviant. Larger N1 and P2 components were observed in the UF compared to the PF condition (see **Figure 2**, standard traces at electrode Fz for UF and PF) but this tendency did not reach significance.

To sum up, an effect of predictability was observed, not only at the latency of the MMN but also earlier, within 70 ms after deviant onset. These two effects go as expected: the more predictable the sequence, the smaller the deviance response. The P3a component was also modulated by the sequence predictability, with larger amplitude observed in UF. The first two effects seem to derive mostly from a deviant response contribution, the P3a one could not be statistically attributed to either standard or deviant responses only.

Controls for Non–Predictability based Biases in UF and PF Responses

First, characterization of undesirable adaptation effects in frequency deviance sequences (UF and PF) was achieved by the comparison between UF and UI conditions. Statistical tests did not reveal any significant effect neither on the difference response (with the exception of TP9 and TP7, from 136 to 148 ms), nor on the deviant and standard responses taken separately, suggesting that the deviance effects observed in UF are, at least to a large extent, not resulting from undesirable refractoriness effects on exogenous ERPs (P50, N1 in particular).

Second, statistical analysis of unfiltered ERPs confirmed every significant effect reported above in bandwidth 2–45 Hz. However, it should be noted that the spatial and temporal extents of those effects were reduced with unfiltered data, which is perfectly

sensible at lower SNR. In the Supplementary material, we provide the unfiltered difference responses for conditions UF and PF at electrode Fz, as well as the corresponding statistical maps obtained from the permutation tests.

Discussion

In this study, we measured different deviance responses elicited by oddball sequences only differing by their statistical temporal structure, referred to as predictability. Our results indicate that sequence predictability modulates deviance responses such that the more predictable the deviant stimulus, the smaller the deviance response. This modulation affects not only the MMN but also earlier slow responses, at the latency of P50 and the auditory MLR components, thereby arguing in favor of various mismatch responses reflecting prediction errors and updates at different levels of the auditory hierarchy. In addition, the measured modulation of the P3a is consistent with unpredictable deviants inducing a larger attentional capture effect. Importantly, these effects were elicited while participants were unaware of the sequence structure. This substantiates the ability of the brain to implicitly monitor statistical properties of the environment such as sequence predictability.

Deviance Effects are not Confounded with Adaptation Effects

Regarding deviance responses, refractoriness state difference between UF and PF should be minimized by the sequence design, which involves the same number of stimulus chunks of each size for both conditions. Moreover, UI and UF deviance responses did not significantly differ, suggesting that not only these responses are similar for both features but also, and more importantly, that frequency deviance of a small magnitude (50 Hz) did not elicit any refractoriness effect detectable in the EEG with our analysis strategy. These findings ensure that observed significant differences between deviant and standard responses are genuine deviance effects. We can thus assume that the significant difference between deviance responses observed in condition UF and PF is not confounded with adaptation effects.

Sequence Predictability Reduces MMN Amplitude

Contrary to Scherg et al. (1989), we measured a significant modulation of the MMN amplitude by sequence predictability, which we interpret as reflecting a smaller prediction error due to a more predictable deviance occurrence. In Scherg et al. (1989), the absence of effect has been interpreted as a result of the automaticity of the MMN, which would prevent this component from being modulated by high-level cognitive processes such as rule extraction. It should be noted that their result derived from a preliminary study conducted with only five participants and relied on a statistical analysis focusing on the MMN amplitude at electrode Fz. Visual inspection of deviance responses for a deviance magnitude of 50 Hz (see Figure 3 in Scherg et al., 1989) shows a difference between regular and irregular sequences which is compatible with our findings. It then appears plausible that a more comprehensive analysis, over all sensors and time bins, would reveal a significant modulation by predictability. However, their experimental design was not adapted to characterize the effect of predictability in isolation from any possible refractoriness confound.

The reduction of the MMN amplitude when predictability of deviance occurrence increases is in line with predictive coding or the Bayesian brain hypothesis (Knill and Pouget, 2004; Friston, 2005). It allows formulating interpretations regarding the underlying mechanisms of prediction updating. UF and PF sequences only differ by their statistical regularities (brought by the global rule). In condition PF, exposure to at least two or three incrementing chunks is required in order to start inferring the regularity of the sequence; with the more chunks, the stronger the confidence in that rule. Perceptual learning - here defined as the process by which the brain encodes over trials the statistical structure of a sensory environment (Friston and Stephan, 2007) -by contrast with the process of learning of new perceptual skills [like in Alain et al. (2007) for instance]- could thus explain the observed modulation of the MMN in the PF compared to UF condition. Predictions, which are updated dynamically through sequential exposure to the stimuli, could indeed be refined in PF through the learning, although approximate, of sequence statistical dependencies. Importantly, none of the participants did report being aware of the differences between experimental conditions. As instructed, they obviously paid little attention to the sounds. This interpretation is consistent with the small amplitude measured for the N2b and P3a components, as we know that they typically follow the MMN under specific condition of stimulus salience or attention orienting toward the stimulus. Altogether, these findings strongly suggest that those perceptual learning processes are implicit. A large number of studies have proposed that the MMN elicited by the violation of complex rules indirectly evidence the implicit learning capacities of the brain. Beside oddball paradigms, the brain ability to track and learn abstract rules without awareness has been straightforwardly evidenced by a large number of studies in the fields of implicit and statistical learning (Perruchet and Pacton, 2006). In line with these accounts, our data argue for a unified implicit learning process that optimizes predictions at different levels. Hence the brain would be constantly tracking the regularities of the environment by means of statistical and implicit learning so as to infer the hidden causal rule(s) governing incoming sensations. Throughout this inference process, mismatch responses would reflect the dynamics of prediction updating, which is guided by the minimization of prediction errors (Friston, 2005). The decrease of mismatch responses observed for the predictable sequence gives support to the idea that the brain optimizes its predictions, even independently of awareness. The MMN has already been proposed to be weighted by the confidence about predictions established through stimulus exposure (Winkler et al., 2009; Todd et al., 2014). Interestingly, the presence of an MMN in condition PF suggests that prediction errors were not abolished for the fully predictable sequence. This could be due to the predictions derived from the approximate learning of the global rule but also to the fact that the local (repetition) rule in UF is still valid in PF

sequences. Despite the existence of high-level predictions derived from the learned global rule, low-level predictions integrating incoming information on a short time-scale might still generate prediction error signals. This is in line with Horváth et al. (2001) who demonstrated the simultaneous integration of different rules at different time-scales, and with Kiebel et al. (2009) pointing to different time-scale prediction errors, corresponding to different levels of an internal hierarchical model.

Under the predictive coding view of the MMN, one could expect the predictability effect to affect both responses to deviants and standards. However, for the latter we only observed a tendency of smaller N1 and P2 responses to predictable standards but no statistically reliable difference. One possible explanation for this lack of significance relates to the passive nature of this paradigm that induces rather small responses to standards, thus yielding a poor signal-to-noise ratio when comparing PF and UF.

Note that in the current study, we manipulated simple perceptual stimuli and observed a modulation of automatic sensory processes by temporal predictability. It would be interesting to replicate our paradigm with conceptual stimuli to test whether this contextual modulation also operates on higher-level processes. Our prediction is that the same effects would be observed and likely express on later components related to more abstract processes like those pertaining to semantic information for instance.

Early Markers of Deviance Detection and Deviance Predictability

Contrary to the majority of MMN studies, we conducted our statistical analysis on entire epochs (from −200 to 400 ms) and this strategy revealed earlier markers of mismatch than the MMN for the unpredictable sequences (UF, UI), within 70 ms after deviant onset. We could identify a statistically significant deviance effect at low frequencies (below 15 Hz). It is worth noting that our set-up and experimental design was not adapted for a fine characterization of fast MLR components, which can also be modulated in oddball paradigm (see below), as there were only ∼175 trials retained on average per stimulus type (typically over 1000 for MLR studies), and an upper bound of bandwidth limited to 45 Hz (typically 150 Hz or 200 Hz for MLR studies). Critically, the genuineness of these early responses had to be controlled with regard to adaptation effects and high-pass filtering bias. Results of these tests, namely an absence of significant difference between UF and UI responses and all effects measured in the bandwidth 2–45 Hz retrieved significantly with unfiltered data, allow us to conclude with high confidence in favor of *genuine* deviance responses for every early effect reported in this study.

Recent findings have already confirmed deviance processing within 50 ms after stimulus onset (for review see Grimm and Escera, 2012; Escera et al., 2014). Contrary to the current results, these findings pertain to the rapid components of the MLR with for instance, an enhancement of the Nb component elicited with pure tone frequency deviants measured with EEG (Grimm et al., 2011) and MEG (Recasens et al., 2014) recordings. Such early mismatch responses complement single-neuron recordings (in animal studies) showing novelty detection responses within midbrain, thalamus and primary auditory cortex (Ulanovsky et al., 2003; Ayala and Malmierca, 2012). Interestingly, Escera and Malmierca (2014) proposed a model of the auditory system dedicated to deviance detection processing at the latency of the MLR that unifies scalp and neuron level findings. Together with the current results, these findings suggest that deviance processing expresses very early and affects both the fast and slow components of the deviant response at early latencies.

Predictable and unpredictable deviance responses were also measured significantly different from about 60 ms over temporo-parietal electrodes. As for the MMN modulation by sequence predictability, we propose that implicit learning is the key mechanism that explains how such early components can be shaped by a global rule. The predictability effect at both early and late latencies could reflect a modulation of high-level predictions on low-level ones within the deviance processing hierarchy. Besides, our results confirm sequence predictability as a suitable tool to characterize the different components of deviance response properly.

Interestingly, previous studies of early deviance effects failed to measure such early ERPs after a global rule violation (Cornella et al., 2012; Althen et al., 2013; Recasens et al., 2014). Escera et al. (2014) and Escera and Malmierca (2014) suggest that these findings corroborate the hierarchical organization of the auditory system, where the different time-scales defining the regularities of the environment would be processed in a forward direction. This model is totally in accordance with a predictive coding implementation (Kiebel et al., 2009), where early deviance responses and the MMN would reflect prediction errors and updates at different levels. However, this view cannot explain the reduced (and thus non-significant) early deviance response in PF as no global rule violation occurs in this condition: mismatch responses are elicited by local rule violation just as they are in the unpredictable sequences. Hence, perceptual learning of the global context may be a plausible explanation to account for the results in PF, with high-level predictions controlling lower level ones. Hence our study provides a new (complementary) contribution to the characterization of the hierarchical auditory system, highlighting top-down (backward) modulations within this hierarchy.

Modulation of the P3a by Sequence Predictability

Following the MMN, the P3a is widely acknowledged as reflecting attention-orienting processes (Polich, 2007). Despite the small frequency and intensity deviance magnitudes that were used, a small but significant P3a component was observed in each of the three experimental conditions. However, its small amplitude, smaller than the MMN deflections, (see **Figure 2**), suggests that the automatic orientation toward the deviants remained rather limited. Note that since the presence of a P3a cannot be interpreted as the signature of an explicit engagement of attention [for instance, it was measured during sleep (Ruby et al., 2008) and with patients with disorders of consciousness (Morlet and Fischer, 2014)], this finding remains compatible with the absence of awareness of the sequence structure as inferred via verbal report in every participant. Interestingly, sequence predictability

also induced a significant modulation of the P3a with larger responses to unpredictable deviants. This further suggests that the P3a also reflect a (third) prediction error. This is definitely in keeping with the predictive coding model of deviance processing, where unexpected stimuli trigger a cascade of prediction errors (conveyed from lower levels to higher ones) that induce in turn adjustments of predictions within each level of the hierarchy. The Dynamic Causal Modeling (DCM) study of Garrido et al. (2007) supports this view, as the authors showed that frontal-to-temporal connections become necessary to explain auditory deviance responses up to the latency of the P3a. An alternative (but compatible) interpretation is that the smaller P3a in the case of predictable deviants reflects a smaller automatic shift of attention.

Conclusion

The recent prediction error model of the MMN yields new expectations regarding its modulations by specific experimental factors, and one of them, sequence predictability, was employed here to refine our understanding of deviance processing. Indeed, we proposed a passive auditory oddball paradigm allowing for the measurement of this effect on genuine deviance responses. We observe a decrease of deviance responses induced by sequence predictability, which directly relates these ERPs to prediction errors and thereby substantiates the predictive coding scheme. Moreover, the threefold predictability effect observed at early and late latencies gives strong support to an auditory hierarchy computing prediction errors at different levels. The statistical structure of sound sequence could be encoded implicitly, possibly through a bayesian inference and learning process implemented within the hierarchy (Kiebel et al., 2009), and large time-scale regularities could induce high-level predictions that modulate both the content and the precision of lower-level ones. These new findings thus raise questions regarding the neural implementation of the predictive coding scheme and the dynamics of deviance processing within the dedicated hierarchy. Hence, further use of our paradigm, in conjunction with generative modeling approaches (Garrido et al., 2009a; Wacongne et al., 2012; Lieder et al., 2013) as well as suitable design optimization methods to compare such models (Sanchez et al., 2014) should help shedding light onto the neurocomputational mechanisms underlying rule learning and deviance processing.

Acknowledgments

We thank Sébastien Daligault for his help in the construction of the experimental material and for programming support in interfacing with the CC IN2P3. We also thank Claude Delpuech and Patrick Bouchet for their participation in the setting of simultaneous MEG-EEG recordings, and Emmanuel Maby for his help on data preprocessing. We acknowledge CC-IN2P3 for providing computing resources and services needed for this work. 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).

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fnhum.2015.00505

References

Acunzo, D. J., Mackenzie, G., and Van Rossum, M. C. (2012). Systematic biases in early ERP and ERF components as a result of high-pass filtering. *J. Neurosci. Methods* 209, 212–218. doi: 10.1016/j.jneumeth.2012.06.011

Aguera, P.-E., Jerbi, K., Caclin, A., and Bertrand, O. (2011). ELAN: a software package for analysis and visualization of meg, eeg, and lfp signals. *Comput. Intell. Neurosci.* 2011, 1–11. doi: 10.1155/2011/158970

Alain, C., Snyder, J. S., He, Y., and Reinke, K. S. (2007). Changes in auditory cortex parallel rapid perceptual learning. *Cereb. Cortex* 17, 1074–1084. doi: 10.1093/cercor/bhl018

Althen, H., Grimm, S., and Escera, C. (2013). Simple and complex acoustic regularities are encoded at different levels of the auditory hierarchy. *Eur. J. Neurosci.* 38, 3448–3455. doi: 10.1111/ejn.12346

Ayala, Y. A., and Malmierca, M. S. (2012). Stimulus-specific adaptation and deviance detection in the inferior colliculus. *Front. Neural Circuits* 6:89. doi: 10.3389/fncir.2012.00089

Bekinschtein, T. A., Naccache, L., Dehaene, S., Rohaut, B., Tadel, F., and Cohen, L. (2009). Neural signature of the conscious processing of auditory regularities. *Proc. Natl. Acad. Sci. U.S.A.* 106, 1672–1677. doi: 10.1073/pnas.0809667106

Bendixen, A., Prinz, W., Horváth, J., Trujillo-Barreto, N. J., and Schröger, E. (2008). Rapid extraction of auditory feature contingencies. *Neuroimage* 41, 1111–1119. doi: 10.1016/j.neuroimage.2008.03.040

Besle, J., Fischer, C., Bidet-Caulet, A., Lecaignard, F., Bertrand, O., and Giard, M. H. (2008). Visual activation and audiovisual interactions in the auditory cortex during speech perception: intracranial recordings in humans. *J. Neurosci.* 28, 14301–14310. doi: 10.1523/JNEUROSCI.2875-08.2008

Bilecen, D., Seifritz, E., Scheffler, K., Henning, J., and Schulte, A.-C. (2002). Amplitopicity of the human auditory cortex: an fMRI study. *Neuroimage* 17, 710–718. doi: 10.1006/nimg.2002.1133

Blair, R. C., and Karniski, W. (1993). An alternative method for significance testing of waveform difference potentials. *Psychophysiology* 30, 518–524. doi: 10.1111/j.1469-8986.1993.tb02075.x

Brattico, E., Tervaniemi, M., Näätänen, R., and Peretz, I. (2006). Musical scale properties are automatically processed in the human auditory cortex. *Brain Res.* 1117, 162–174. doi: 10.1016/j.brainres.2006.08.023

Cornella, M., Leung, S., Grimm, S., and Escera, C. (2012). Detection of simple and pattern regularity violations occurs at different levels of the auditory hierarchy. *PLoS ONE* 7:e43604. doi: 10.1371/journal.pone.0043604

Daikhin, L., and Ahissar, M. (2012). Responses to deviants are modulated by subthreshold variability of the standard. *Psychophysiology* 49, 31–42. doi: 10.1111/j.1469-8986.2011.01274.x

Demarquay, G. V., Caclin, A., Brudon, F., Fischer, C., and Morlet, D. (2011). Exacerbated attention orienting to auditory stimulation in migraine patients. *Clin. Neurophysiol.* 122, 1755–1763. doi: 10.1016/j.clinph.2011.02.013

Escera, C., Leung, S., and Grimm, S. (2014). Deviance detection based on regularity encoding along the auditory hierarchy: electrophysiological evidence in humans. *Brain Topogr.* 27, 527–538. doi: 10.1007/s10548-013-0328-4

Escera, C., and Malmierca, M. S. (2014). The auditory novelty system: an attempt to integrate human and animal research. *Psychophysiology* 51, 111–123. doi: 10.1111/psyp.12156

Fishman, Y. I. (2014). The mechanisms and meaning of the mismatch negativity. *Brain Topogr.* 27, 500–526. doi: 10.1007/s10548-013-0337-3

Friston, K. (2005). A theory of cortical responses. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 360, 815–836. doi: 10.1098/rstb.2005.1622

Friston, K., and Kiebel, S. J. (2009). Predictive coding under the free-energy principle. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 364, 1211–1221. doi: 10.1098/rstb.2008.0300

Friston, K., and Stephan, K. E. (2007). Free-energy and the brain. *Synthese* 159, 417–458. doi: 10.1007/s11229-007-9237-y

Garrido, M. I., Kilner, J. M., Kiebel, S. J., and Friston, K. (2007). Evoked brain responses are generated by feedback loops. *Proc. Natl. Acad. Sci. U.S.A.* 104, 20961–20966. doi: 10.1073/pnas.0706274105

Garrido, M. I., Kilner, J. M., Kiebel, S. J., and Friston, K. (2009a). Dynamic causal modeling of the response to frequency deviants. *J. Neurophysiol.* 101, 2620–2631. doi: 10.1152/jn.90291.2008

Garrido, M. I., Kilner, J. M., Stephan, K. E., and Friston, K. (2009b). The mismatch negativity: a review of underlying mechanisms. *Clin. Neurophysiol.* 120, 453–463. doi: 10.1016/j.clinph.2008.11.029

Garrido, M. I., Sahani, M., and Dolan, R. J. (2013). Outlier responses reflect sensitivity to statistical structure in the human brain. *PLoS Comput. Biol.* 9:e1002999. doi: 10.1371/journal.pcbi.1002999.s002

Grimm, S., and Escera, C. (2012). Auditory deviance detection revisited: evidence for a hierarchical novelty system. *Int. J. Psychophysiol.* 85, 88–92. doi: 10.1016/j.ijpsycho.2011.05.012

Grimm, S., Escera, C., Slabu, L., and Costa-Faidella, J. (2011). Electrophysiological evidence for the hierarchical organization of auditory change detection in the human brain. *Psychophysiology* 48, 377–384. doi: 10.1111/j.1469-8986.2010.01073.x

Horváth, J., Czigler, I., Sussman, E. S., and Winkler, I. (2001). Simultaneously active pre-attentive representations of local and global rules for sound sequences in the human brain. *Brain Res. Cogn. Brain Res.* 12, 131–144. doi: 10.1016/S0926-6410(01)00038-6

Jankowiak, S., and Berti, S. (2007). Behavioral and event-related potential distraction effects with regularly occurring auditory deviants. *Psychophysiology* 44, 79–85. doi: 10.1111/j.1469-8986.2006.00479.x

Kiebel, S. J., Daunizeau, J., and Friston, K. (2009). Perception and hierarchical dynamics. *Front. Neuroinform.* 3:20. doi: 10.3389/neuro.11.020.2009

Knill, D. C., and Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* 27, 712–719. doi: 10.1016/j.tins.2004.10.007

Kujala, T., and Näätänen, R. (2010). The adaptive brain: a neurophysiological perspective. *Prog. Neurobiol.* 91, 55–67. doi: 10.1016/j.pneurobio.2010.01.006

Lieder, F., Stephan, K. E., Daunizeau, J., Garrido, M. I., and Friston, K. (2013). A neurocomputational model of the mismatch negativity. *PLoS Comput. Biol.* 9:e1003288. doi: 10.1371/journal.pcbi.1003288

Morlet, D., Demarquay, G., Brudon, F., Fischer, C., and Caclin, A. (2014). Attention orienting dysfunction with preserved automatic auditory change detection in migraine. *Clin. Neurophysiol.* 125, 500–511. doi: 10.1016/j.clinph.2013.05.032

Morlet, D., and Fischer, C. (2014). MMN and novelty P3 in coma and other altered states of consciousness: a review. *Brain Topogr.* 27, 467–479. doi: 10.1007/s10548-013-0335-5

Näätänen, R., Astikainen, P., Ruusuvirta, T., and Huotilainen, M. (2010). Automatic auditory intelligence: an expression of the sensory. *Brain Res. Rev.* 64, 123–136. doi: 10.1016/j.brainresrev.2010.03.001

Näätänen, R., Gaillard, A. W., and Mäntysalo, S. (1978). Early selective-attention effect on evoked potential reinterpreted. *Acta Psychol. (Amst.)* 42, 313–329. doi: 10.1016/0001-6918(78)90006-9

Näätänen, R., Kujala, T., Escera, C., Baldeweg, T., Kreegipuu, K., Carlson, S., et al. (2012). The mismatch negativity (MMN)–a unique window to disturbed central auditory processing in ageing and different clinical conditions. *Clin. Neurophysiol.* 123, 424–458. doi: 10.1016/j.clinph.2011.09.020

Näätänen, R., Paavilainen, P., Rinne, T., and Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: a review. *Clin. Neurophysiol.* 118, 2544–2590. doi: 10.1016/j.clinph.2007.04.026

Näätänen, R., Simpson, M., and Loveless, N. E. (1982). Stimulus deviance and evoked potentials. *Biol. Psychol.* 14, 53–98. doi: 10.1016/0301-0511(82)90017-5

Perruchet, P., and Pacton, S. (2006). Implicit learning and statistical learning: one phenomenon, two approaches. *Trends Cogn. Sci.* 10, 233–238. doi: 10.1016/j.tics.2006.03.006

Polich, J. (2007). Updating P300: an integrative theory of P3a and P3b. *Clin. Neurophysiol.* 118, 2128–2148. doi: 10.1016/j.clinph.2007.04.019

Recasens, M., Grimm, S., Capilla, A., Nowak, R., and Escera, C. (2014). Two sequential processes of change detection in hierarchically ordered areas of the human auditory cortex. *Cereb. Cortex* 24, 143–153. doi: 10.1093/cercor/bhs295

Ruby, P., Caclin, A., Boulet, S., Delpuech, C., and Morlet, D. (2008). Odd sound processing in the sleeping brain. *J. Cogn. Neurosci.* 20, 296–311. doi: 10.1162/jocn.2008.20023

Sams, M., Alho, K., and Näätänen, R. (1983). Sequential effects on the ERP in discriminating two stimuli. *Biol. Psychol.* 17, 41–58. doi: 10.1016/0301-0511(83)90065-0

Sams, M., Paavilainen, P., Alho, K., and Näätänen, R. (1985). Auditory frequency discrimination and event-related potentials. *Electroencephalogr. Clin. Neurophysiol.* 62, 437–448. doi: 10.1016/0168-5597(85)90054-1

Sanchez, G., Daunizeau, J., Maby, E., Bertrand, O., Bompas, A., and Mattout, J. (2014). Toward a new application of real-time electrophysiology: online optimization of cognitive neurosciences hypothesis testing. *Brain Sci.* 4, 49–72. doi: 10.3390/brainsci4010049

Sato, Y., Yabe, H., Hiruma, T., Sutoh, T., Shinozaki, N., Nashida, T., et al. (2000). The effect of deviant stimulus probability on the human mismatch process. *Neuroreport* 11, 3703–3708. doi: 10.1097/00001756-200011270-00023

Scherg, M., Vajsar, J., and Picton, T. W. (1989). A source analysis of the late human auditory evoked potentials. *J. Cogn. Neurosci.* 1, 336–355. doi: 10.1162/jocn.1989.1.4.336

Schwartze, M., Farrugia, N., and Kotz, S. A. (2013). Dissociation of formal and temporal predictability in early auditory evoked potentials. *Neuropsychologia* 51, 320–325. doi: 10.1016/j.neuropsychologia.2012.09.037

Sussman, E. S., and Gumenyuk, V. (2005). Organization of sequential sounds in auditory memory. *Neuroreport* 16, 1519–1523. doi: 10.1097/01.wnr.0000177002.35193.4c

Sussman, E. S., Ritter, W., and Vaughan, H. G. (1998). Predictability of stimulus deviance and the mismatch negativity. *Neuroreport* 9, 4167–4170. doi: 10.1097/00001756-199812210-00031

Tiitinen, H., May, P. J. C., Reinikainen, K., and Näätänen, R. (1994). Attentive novelty detection in humans is governed by pre-attentive sensory memory. *Nature* 372, 90–92. doi: 10.1038/372090a0

Todd, J., Heathcote, A., Mullens, D., Whitson, L. R., Provost, A., and Winkler, I. (2014). What controls gain in gain control? Mismatch negativity (MMN), priors and system biases. *Brain Topogr.* 27, 578–589. doi: 10.1007/s10548-013-0344-4

Todd, J., Provost, A., Whitson, L. R., Cooper, G., and Heathcote, A. (2013). Not so primitive: context-sensitive meta-learning about unattended sound sequences. *J. Neurophysiol.* 109, 99–105. doi: 10.1152/jn.00581.2012

Ulanovsky, N., Las, L., and Nelken, I. (2003). Processing of low-probability sounds by cortical neurons. *Nat. Neurosci.* 6, 391–398. doi: 10.1038/nn1032

Vuust, P., Ostergaard, L., Pallesen, K. J., Bailey, C., and Roepstorff, A. (2009). Predictive coding of music - Brain responses to rhythmic incongruity. *Cortex* 45, 80–92. doi: 10.1016/j.cortex.2008.05.014

Wacongne, C., Changeux, J.-P., and Dehaene, S. (2012). A neuronal model of predictive coding accounting for the mismatch negativity. *J. Neurosci.* 32, 3665–3678. doi: 10.1523/JNEUROSCI.5003-11.2012

Winkler, I., and Czigler, I. (2012). Evidence from auditory and visual event-related potential (ERP) studies of deviance detection (MMN and vMMN) linking predictive coding theories and perceptual object representations. *Int. J. Psychophysiol.* 83, 132–143. doi: 10.1016/j.ijpsycho.2011.10.001

Winkler, I., Denham, S. L., and Nelken, I. (2009). Modeling the auditory scene: predictive regularity representations and perceptual objects. *Trends Cogn. Sci.* 13, 532–540. doi: 10.1016/j.tics.2009.09.003

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

5.3 MEG analysis

We now present the analysis of sensor-level MEG data acquired simultaneously to EEG data during this study.

5.3.1 Material and methods

MEG analysis rested on the data obtained from exactly the same subjects (N=22) and using the same material as for the EEG study. Preprocessing of data included independent component analysis (ICA) correction for ocular artifacts and filtering (2-45 Hz band-pass digital filter, bidirectional Butterworth, 4th order) (as detailed in Lecaignard et al., 2015). In addition, specific MEG treatments were applied: data segments corresponding to head movements larger than 15 mm relative to the average position within a session and to SQUID jumps were rejected. Critically, the stimuli entering EEG and MEG analysis throughout every analysis presented in this thesis were exactly the same. Filtered data were epoched from -200 ms to 410 ms post-stimulus (mean signal during pre-stimulus time-interval was removed for baseline correction). Standard sounds entering standard evoked response refer to every sound preceding a deviant.

5.3.2 Results

We first present the deviance responses measured in each condition (UF, PF and UI). We then report the findings regarding the predictability effect (UF vs. PF). Mismatch responses obtained with frequency and intensity (condition UF and UI) will be briefly compared (more details will be provided in next chapter dedicated to the characterization of frequency and intensity deviance cortical generators).

Deviance responses. Deviance results obtained at different latency and for each condition are presented in Figure 5.1. The spatial distribution of peak amplitudes for each response was associated to a typical MEG auditory pattern composed of two anterior and two posterior poles with diametrically opposed sign (Figure 5.1-B). Early deviance response could be measured in all conditions with a significant early effect in the first 100 ms ($[-5, 90]$ ms, $[-15, 90]$ ms, $[-5, 95]$ ms for condition UF, PF and UI respectively). MMN was observed in all conditions, peaking around 175 ms and with significance over $[110, 215]$ ms, $[105, 225]$ ms, $[135, 225]$ ms for UF, PF and UI respectively. It should be noted that visual inspection could reveal similar peak amplitudes at anterior poles for every condition (175 ms) combined with posterior poles in UF and PF peaking earlier (around 160 ms for left posterior pole, and around 170 ms for right posterior pole). In condition UI, both posterior poles were found peaking later than anterior ones (around 185 ms). Finally, following the MMN, a late component peaking around 275 ms could be also observed, with significant emergence from 225 ms to 325 ms, 230 to 345 ms and 235 to 315 ms for condition UF, PF and UI respectively. Findings from EEG data led us to label this component as a P3a.

Predictability modulation of deviance responses. Contrary to the EEG findings, the comparison of difference responses obtained in UF and PF at the sensor-level revealed that there was no consistent evidence of predictability effect on deviance responses, as can be seen in Figure 5.1-c and Figure 5.2. An absence of modulation was also observed between UF and PF standard responses (it was also the case with EEG data). Finally, comparison of UF and PF deviant

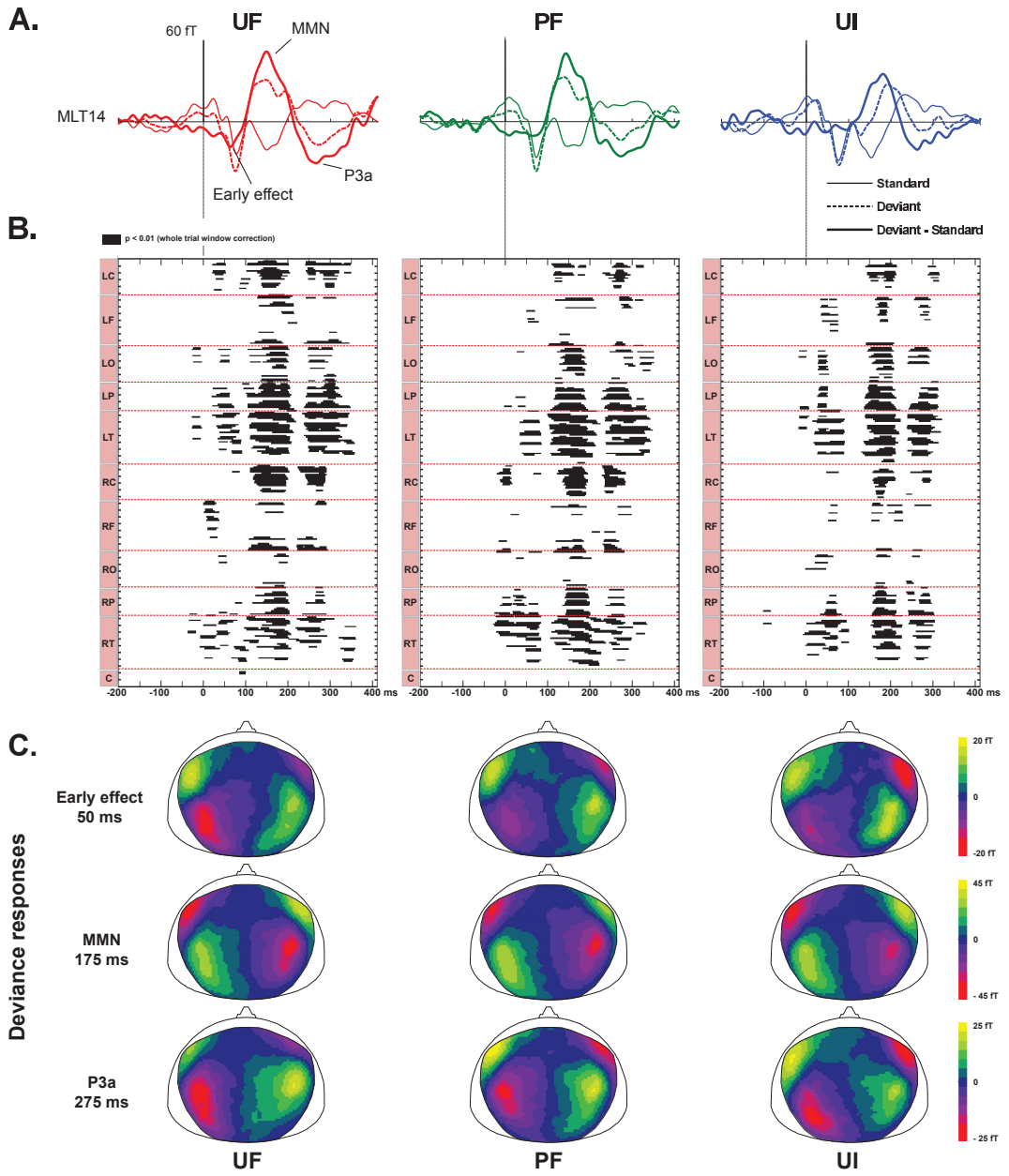

Figure 5.1 – Deviance responses measured with MEG. (a) Standard, deviant and difference evoked responses measured at left temporal sensor MLT14, for frequency (UF, red; PF, green) and intensity (UI, blue) deviance. (b) Statistical maps revealing the different periods of significant deviance responses with a corrected for multiple comparison threshold of $p < 0.01$. (c) Scalp topographies obtained for each condition, at three relevant latencies corresponding to the early deviance, the MMN and the P3a responses. Color scales are indicated per map latency. Black dot indicates sensor MLT14.

responses could reveal two time intervals with a significant difference resting on more than 10 adjacent sensors: an early effect at 16 sensors within the right anterior pole with significance spanning from 15 ms to 25 ms; and a later effect, at the latency of the MMN, involving 16 sensors at the left posterior pole, with significance from 190 ms to 210 ms. Both effects exhibit larger event-related field (ERF) amplitude for condition UF.

Frequency and intensity deviance responses. The MMN elicited by (unpredictable) frequency deviants (UF) was found to start earlier and to have larger amplitude than the MMN elicited with

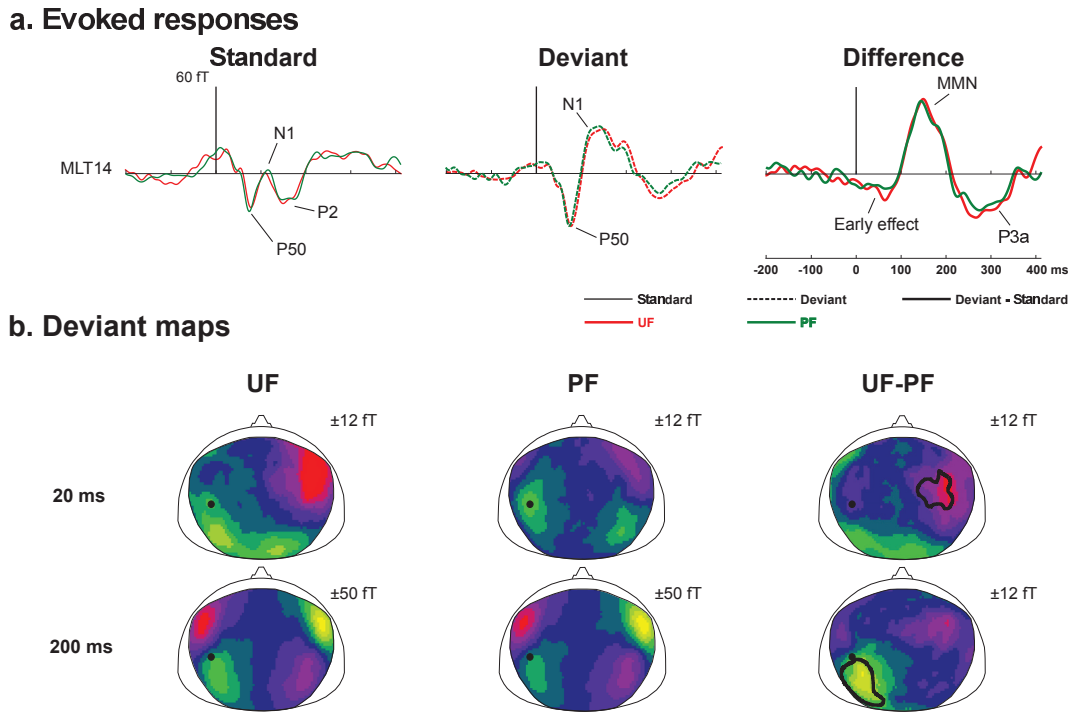

Figure 5.2 – Predictability effect assessed with MEG. (a) Standard, deviant and difference evoked responses measured at left temporal sensor MLT14, for UF (red) and PF (green) conditions. (b) Scalp topographies obtained for deviant responses for UF (left column), PF (middle) and UF-PF (right) at early latency (upper row) and late latency (lower row). Color scales are indicated for each map. Black dot indicates sensor MLT14 and sensors showing a significant predictability effect have been surrounded by a black line.

intensity deviants (UI). For instance, the peak at sensor MLT14 was measured around 150 ms with an amplitude of 51 fT in condition UF, and around 185 ms with an amplitude of 34 fT in condition UI (Figure 5.1-a). Statistical comparison of UF and UI deviance responses showed a significant interval from 90 ms to 160 ms. Such difference between conditions was found to originate from deviant contribution as no difference between standard traces could be measured. Regarding deviant responses, the three deflections (with alternating sign) following the N1 exhibited significant larger amplitude in UF compared to UI. These differences affected posterior bilateral poles for the first two deflections (with a larger effect on the right side), spanning from 95 ms to 165 ms for the first one (the rising slope of the MMN), and from 210 ms to 250 ms for the second peak during the P3a. The last peak was measured larger in UF only for right sensors from 330 ms to 360 ms.

5.3.3 Conclusion

The three deviance responses reported in our EEG analysis could be also observed with MEG recordings. The early effect was found to start around stimulus onset and to last until the N1 component. Importantly, it was consistent on MEG traces in condition PF while EEG analysis failed to reach statistical significance. Following this effect, an MMN could be measured in all conditions, with anterior poles peaking around 175 ms, consistent with EEG findings; posterior poles showed a more complex pattern over conditions that was not found with EEG. Finally, the P3a was found to peak earlier than in EEG (around 275 ms and around 300 ms for MEG and EEG respectively). These findings support the plausibility of the three different mismatch responses

elicited by our oddball sequences. Furthermore, the observed differences between modalities can be explained by the acknowledged different (theoretical) sensitivity of EEG and MEG regarding the orientation and the depth of sources (Crouzeix-Cheylus, 2001). Different spatio-temporal patterns were observed in the two modalities and suggest multiple spatially distinct areas involved during auditory deviance processing. Their respective activity could indeed be captured (or not) by a modality according to specific biophysical properties. The complex pattern of posterior MEG responses is consistent with the high sensitivity of MEG to measure temporal activity with large SNR, hence enabling the dissociation of different activation distributed along the superior temporal gyrus (Yvert et al., 2001).

Surprisingly, the EEG modulation of deviance responses by sequence predictability could not be retrieved with MEG, but two effects on deviant responses (with however consistent latencies, and larger amplitude with unpredictable deviants, as expected). One should consider here a practical limitation encountered with MEG when performing group-level sensor-level averages of evoked responses: the variability of the relative position of the head relative to the sensors may induce a spatial blurring, hence degrade the statistical sensitivity of the group-level analysis (not to mention within-subject head movements over time). However, the current lack of evidence may potentially provides some indications regarding the underlying neural activity behind this contextual manipulation. Indeed, our EEG analysis led us to consider the contribution of a frontal area for the implicit learning of the statistical structure in PF. Different arguments supported this interpretation: first, the existing literature in the field of implicit learning reports such frontal implication (Conway & Pisoni, 2008). Second, from a predictive coding perspective, the predictability modulation should have the form of a descending prediction influencing information processing at lower levels. Finally, a more speculative argument pertains to our sensor-level EEG analysis as the predictability effect on the MMN was found at fronto-central electrodes but not at mastoid sites. This latter aspect should be considered cautiously as it gets over the so-called inverse problem of inferring neural mechanisms from external recordings. However, the absence of MEG modulation could also support a fronto-medial contribution since these regions (having a quasi-radial orientation) may poorly express on MEG gradiometers.

Contrary to the EEG findings, frequency and intensity deviance responses measured with MEG exhibited different patterns at the latency of the MMN and the P3a. This result suggests similar early processing of deviants until 100 ms, that could refer to low-level processing of sound attributes. Then, both processing could differ that could engage different areas within the auditory cortex. This was already suggested by equivalent current dipole (ECD) source reconstruction studies with EEG data (Giard et al., 1995) and MEG data (Rosburg et al., 2004). This result also supports the larger spatial resolution of MEG for temporal activations.

5.4 Attempts to identify electrophysiological markers of perceptual learning

Following the EEG and MEG analysis of deviant predictability effect on mismatch responses, we attempted to characterize the learning of the statistical structure in condition PF using scalp-level data, although the present study had not been designed to that aim. Such analysis would help

at identifying electrophysiological markers reflecting how the brain extracts the regularities of its environment, that could inform subsequent computational learning modeling. In particular, one could hypothesized that the brain has the capacity to learn the incrementing structure of a PF sequence, or at least to figure out that deviants get more and more spaced in time (possibly in a cyclic manner). Alternatively, it could be that the brain learns that the size of chunks (or equivalently the number of standards in between two deviants) evolves slowly or remains stable in PF whereas strong transitions occurs in condition UF. As already seen, rule extraction could rest on the perceptual learning of sequence regularities and its investigation requires a trial-by-trial model-driven analysis to capture the dynamics of such learning. In the following, we present some attempts to start to tackle this issue, that were conducted with EEG event-related responses over the experimental timescale; none of them succeeded to reveal consistent effect, due to the inadequate number of stimuli required for such averaging analysis.

Dynamics of rule learning. We first tried to assess the dynamics of rule learning by examining the predictability effect on mismatch responses throughout the experimental sessions. Precisely, each subject received two sessions of each condition UF and PF, each of them comprising 16 cycles of 42 sounds (see Figure 1 in the EEG article). For each of these four sessions, we computed the average standard (preceding a deviant) and deviant responses over cycles 1 to 6 and over cycles 11 to 16 to derive two mismatch responses whose comparison could reflect the effect of rule exposure if any. Each of these 8 evoked responses entailed a maximum of 42 events per subject; at the group level, a MMN could clearly be observed but visual inspection failed to reveal a learning effect (*ie* an MMN difference between the beginning and the end of a session), neither an interaction of this learning with the predictability effect. This observation suggests a rapid learning of the rule in PF, taking place within the first 6 cycles of the regular sequence. No difference could neither be observed between the first session and the second session for both conditions UF and PF.

Predictability effect on standard repetition. As already described in previous chapters, the decrease of activity observed for repeated stimuli could derive from bottom-up adaptation and from top-down predictions, that were referred to as repetition suppression (RS) and expectation suppression in (Todorovic & de Lange, 2012). We assumed that bottom-up effects should be the same in both conditions (as discussed in the EEG article) but expected a different modulation of top-down predictions over the sequential presentation of standards induced by the PF rule learning. To assess such modulation on the MMN amplitude, we averaged standard stimuli according to their position relative to deviants in the sequence. We restricted the analysis to position 2 to 6 to ensure a minimum of 75 events per average. A two-way anova with repeated measures (with condition and standard position as factors) conducted on MMN amplitudes could reveal significant main effects for both factors but failed to reveal an interaction, that was necessary to conclude. A possible interpretation for such failure concerns the interaction of standard position and the size of chunks that could affect the expectancy of incoming sounds (for instance, in condition PF, a standard at position 3 may not be treated equivalently depending if it belongs to a small or a large chunk) . The number of stimulus (per position and per chunk size) was insufficient to consider any further investigation.

Evolution of deviant expectation over sound sequence. Finally, we aimed at refining the predictability effect measured with all deviant responses by sorting deviants according to the number

of preceding standards (we denoted deviants of rank 2 to 8). For both UF and PF sequences, the brain could have learnt that *on average* a deviant could occur every five standards, leading deviant of rank 5 to be more expected than extreme ranks (2 or 8), hence eliciting a smaller MMN (visual inspection of corresponding traces could indeed show a trend). In addition, the higher-level incrementing rule in PF should affect such prediction in the sense that deviant expectation should remain stable (and high) over deviant ranks. This analysis rested on averaged over 25 events per evoked response, which hampered the statistical analysis. In the same vein, in PF, deviants of rank 2 should elicit a MMN as they violate the incrementing rule, but the small number of related stimuli prevented from such analysis.

The failure of these analysis could be due to the inadequation of our experimental design for the event-related investigations cited here (in particular with regards to the number of stimuli per condition). These unfruitful attempts confirmed that further modeling analysis should only address the predictability *modulation* of deviance responses (as was initially planned before starting the study) and could not resolve the (unknown) underlying mechanisms by which the rule in PF was learnt. Most of all, they could reveal the limitation of evoked response analysis to capture electrophysiological biomarkers of the learning dynamics, thereby calling for more sophisticated methodologies such as computational trial-by-trial analysis to provide a more fine-grained investigation of brain activity reflecting such processes.

5.5 Conclusion

As expected under predictive coding, the predictability of deviant occurence was shown to decrease the amplitude of the MMN measured with EEG. Two additional deviance responses could be measured that also appeared modulated by this effect in the same manner. This modulation led us to interpret these three deviance responses as reflecting different prediction errors elicited along the auditory hierarchy. We believe that this modulation derived from the implicit learning of the statistical structure of PF sequence, that could involve higher cognitive levels possibly implicating frontal regions. Critically, MEG analysis confirmed these three responses but their modulation by sequence predictability could not be visible on difference responses (but on deviant responses). None of the two modalities could reveal an effect of predictability on standard responses using sensor-level data (unless considering all standards but the one following a deviant, as described in the EEG article), although one could reasonably consider that it should affect both standard and deviant processing. Difference between EEG and MEG was also encountered with the intensity deviance response. Taken together, these results highlight the complementary information about neural processes provided by both modalities. In the present study, they both helped at characterizing deviance processing and its modulation by sequence predictability. Advanced modeling analysis appears necessary at this stage to further characterize this modulation under the perspective of predictive coding, hence possibly providing new insights into the mechanisms behind mismatch responses. As such inversion-based methods require a large amount of information contained in measured data, the EEG and MEG complementarity observed so far therefore encouraged us to fully exploit these data in every subsequent analysis presented in this thesis.

Chapter 6

Spatial characterization of the cortical network for auditory deviance processing

6.1 Objectives

Both DCM and computational learning models of the MMN (chapter 4) build on a hierarchy accounting for specific neurophysiological or functional hypothesis that should remain biologically plausible. For instance, the DCM of MMN proposed by Garrido and collaborators (2009) is organized as a three-level hierarchy composed of equivalent current dipoles whose locations were estimated in a former fMRI study (Opitz et al., 2002). This brings to the critical importance to inform such models with accurate description of the cortical network associated with deviance processing. Despite a large literature dedicated to the characterization of the auditory MMN generators (see chapter 3, §3.3.1) limitations in spatial and temporal resolutions have prevented from a precise description that could now be attained with recent methodological advances in source reconstruction using EEG and MEG data. Moreover, there is a lack of findings regarding the localization of early mismatch sources, which is likely due to the recency of related studies. Hence, it thus appears essential to use up-to-date methodologies to describe accurately early and late deviance response sources, as such characterization could contribute to improve subsequent deviance modeling analysis. Using an empirical Bayesian distributed approach integrating EEG-MEG fusion, we performed a source reconstruction analysis to locate mismatch cortical generators elicited by frequency and intensity deviances. Furthermore, the potential of Bayesian model selection that this statistical framework offers was exploited to assess quantitatively the performance of fused inversion to resolve source prior models in comparison to unimodal one. All these findings are reported in the article in preparation presented below.

6.2 Article

Empirical evaluation of fused MEG-EEG source reconstruction applied to auditory mismatch generators

(in preparation)

Authors: Françoise Lecaignard, Olivier Bertrand, Anne Caclin, Jérémie Mattout

ABSTRACT

Combining complementary EEG and MEG information for source reconstruction has been consistently evidenced to enhance localization performances using simulated data. Such fusion has been integrated in a Bayesian scheme (Henson et al., 2009), thereby providing an advanced reconstruction approach exploiting both modalities (EEG, MEG) in a Bayesian optimal way. Bayesian framework for model inversion allows estimating posterior estimates of unknown quantities, and also enables model comparison that appears perfectly adapted to test quantitatively the added value of fusion in the case of real data, as we propose here. Fused EEG-MEG source reconstruction was applied to the Mismatch Negativity (MMN), a well-known brain component elicited by a deviant stimuli violating a regular auditory stream. Despite a great number of studies about the underlying generators of the MMN, this issue is still controversial. Furthermore, recent findings in auditory deviance studies revealed earlier mismatch responses than the MMN (Escera & Malmierca, 2014) whose cortical sources have not been fully explored yet. Fused localization methods combining high spatial and temporal resolutions thus appears relevant to refine the characterization of mismatch generators. In this study, we used Bayesian source reconstruction with EEG-MEG fusion to locate early and late mismatch cortical generators elicited by frequency and intensity deviances. Bilateral sources in supratemporal cortex and inferior frontal gyrus were found for both features, and interestingly, fusion could reveal an accurate spatio-temporal dissociation between conditions within the supratemporal plane. Using Bayesian model comparison, we could confirm empirically that fused inversion provides an increased spatial resolution compared to unimodal ones. Our findings provide empirical support for fused inversion using simultaneous EEG and MEG recordings. The fine-grained spatial description of the auditory cortical hierarchy achieved here represents a crucial step prior to further address the outstanding issue of characterizing the neurophysiological and computational mechanisms behind mismatch responses.

Introduction

Imaging Human brain function is acknowledged to require high spatial and temporal resolutions. A concrete illustration underlining this twofold necessity pertains to the large literature dedicated to the generators of the auditory Mismatch Negativity (MMN), a brain response elicited by a change (deviant) in a regular acoustic environment that plays a central role in cognitive and clinical neuroscience (Morlet & Fischer, 2014; Sussman & Shafer, 2014). Spatial characterization of MMN sources is needed to improve our understanding of this response (known to peak within 230 ms after deviant onset), and strong efforts using different neuroimaging techniques have been made to that aim for about three decades. Functional Magnetic Resonance Imaging (fMRI) and electrophysiological techniques such as Electro- and Magneto-encephalography (EEG and MEG respectively) were mostly employed, that favored spatial or temporal precision respectively. Both fMRI (see for review Deouell, 2007) and electrophysiological studies (Giard et al., 1995; Waberski et al., 2001; Marco-Pallarés et al., 2005; Lappe et al., 2013b; Ruhnau et al., 2013; Recasens, Grimm, Wollbrink, et al., 2014; Fulham et al., 2014) suggested dominant temporal and frontal contributions. However, these findings also strongly reflect a lack of robustness in the characterization of deviance sources that could result from the fact that none of these modalities is efficiently informed in both spatial and temporal dimensions. In addition, recent EEG and MEG studies now suggest earlier deviance response than the MMN (for review, see Escera & Malmierca, 2014), hence further demonstrating the need to combine temporal and spatial information for a comprehensive description of brain responses.

Spatial description of brain activity using EEG and MEG recordings requires solving an underdetermined inverse problem when using distributed source reconstruction methods. With recent advances producing more informed modeling, distributed inversion now appears promising to combine in a straightforward fashion a fine degree of spatial and temporal precision required to describe brain functioning. Two methods in particular were of interest in the current study, that could help incorporating additional information to constrain model inversion, thereby reducing source localization uncertainty. First, the formulation of the inverse problem within a Bayesian framework allows confronting initial assumptions (priors) to sensor measurements in a principled fashion (Friston, Henson, et al., 2006a; Mattout et al., 2006). In particular, the unknown spatial covariance of sources, embodying experimenter's prior knowledge about the spatial properties of activated sources, can be estimated through the Bayesian inversion scheme integrated into the Multiple Sparse Priors (MSP) method (Friston, Harrison, et al., 2008). Importantly, MSP has recently been enriched with group-level inversion (Litvak & Friston, 2008) to further refine these priors. Second, integration of EEG and MEG data for solving the inverse problem augments the quantity of information introduced in source modeling. MEG-EEG fusion not only allows accounting for information missed by one modality and captured by the other one (Dale & Sereno, 1993; Fuchs et al., 1998), but also crucially provides complementary information: under the quasi-static approximation of the Maxwell's Equations, scalp-level recordings in each modality result from the same neuronal activity inducing decoupled (hence independent) electric and the magnetic fields (Plonsey & Heppner, 1967). There is a broad literature dedicated to fused MEG-EEG source reconstruction that suggests (in spite of the various modeling assumptions employed) greater performances for this approach compared to separate inversions. In short, reduced localization errors could be reported with fused inversion for both superficial and deep sources (Fuchs et al.,

1998), as well as for different signal-to-noise ratio (SNR) and sensor montages (Babiloni et al., 2004). Decrease of the undesirable sensitivity of inversion methods to source orientation (Baillet et al., 1999) and enhanced precision of source estimates (Henson et al., 2009) were also reported. Most of these studies employed simulated data (for which the true source distribution is known) and comparative evaluation then relied on several metrics accounting for differences between the reference and reconstructed distributions. In the case of real data, the ill-posed property of the inverse problem (absence of unique solution) prevents from similar quantitative analysis. Recent attempts considered specific cases for which fMRI results (Sharon et al., 2007), brain response widely described in the literature (Molins et al., 2008) or intracranial recordings with epileptic patients (Chowdhury et al., 2015) approximated true solutions to be compared with; all these studies, resting on data acquired from 2 to 6 subjects, were in favor of reduced mislocalizations with fused modalities.

In this context, the aim of the current study was twofold: to model brain activity spatiotemporally using advanced electrophysiological Bayesian methods including fused MEG-EEG inversion (Henson et al., 2009) and to propose a general method to evaluate quantitatively the performance of unimodal and fused source reconstruction with empirical data. Our approach investigates the ability of each modality (EEG, MEG and MEG-EEG) to separate different source distributions (being spatial models) and relies on Bayesian model comparison (Stephan et al., 2009) to provide a quantitative measure of this spatial model resolution.

We applied these advanced methods and our evaluation approach to the investigation of deviance generators, including early deviance effect and the MMN per se. We considered data originating from a previous passive auditory oddball study (Lecaignard et al., 2015) with two deviance features (frequency and intensity, separately manipulated) and conducted with simultaneous EEG and MEG recordings. Despite the large literature in the field of deviance sources, to date no study has been conducted using fused inversion (simultaneous recordings but separate source modeling were achieved in Huotilainen et al., 1998; Rinne et al., 2000). Furthermore, only a few MEG studies addressed the localization of early deviance responses (Recasens, Grimm, Capilla, et al., 2014; Recasens, Grimm, Wollbrink, et al., 2014; Ruhnau et al., 2013), with activity circumscribed in the primary auditory cortex. Prior to these deviance-related inversions, we controlled the performance of the overall inversion scheme (from forward model computation to individual source estimates) on auditory P50 component elicited by standard (regular) sounds. This response has been previously associated with primary auditory cortex activity using intracranial recordings (Pantev et al., 1995; Yvert et al., 2002) and was therefore considered fairly appropriate for validation purposes.

Our study presents estimates of cortical activity obtained with advanced methods comprising group-level inference and fused MEG-EEG source inversion. For the first time, separate (EEG, MEG) and fused MEG-EEG inversions were evaluated empirically by means of a model-comparison method. Applied to the reconstruction of early and late auditory deviance generators, our results demonstrate the usefulness of fused inversion that produced a fine-grained description of a fronto-temporal network.

Material and Methods

This section is divided into three parts. In the first section, we describe the methods employed for source reconstruction, comprising forward model computation and model inversion with MSP, group-level inference and fused MEG-EEG. In the second section, we detail the approach that we propose for the quantitative evaluation of EEG, MEG and MEG-EEG inversions. Finally, the third section presents the empirical data used to validate our approach. These data corresponds to auditory evoked responses recorded using frequency and intensity oddball sequences (conditions FRQ and INT respectively, in separate sessions; these conditions correspond to the UF and UI conditions in Lecaignard et al. (2015).

Methods for source reconstruction

Forward model computation

For both MEG and EEG modalities, realistic Boundary Element Model (BEM) (Hämäläinen & Sarvas, 1989) was employed here to account for head geometry and avoid oversimplification induced by spherical models. Individual head-models were composed of three layers (scalp, skull and brain) with homogenous and isotropic conductivities set to 0.33, 0.0041 and 0.33 Sm respectively (Rush & Driscoll, 1968). Layer boundaries were defined with individual meshes (scalp, outer skull and inner skull) and composed of 5120 faces and 2562 nodes each. Source domain included Ns=20484 sources (mean average distance = 3.4 mm) distributed on the cortical mesh (grey-white matter interface) and dipole orientation was constrained to be normal to the surface. All meshes derived from canonical uniformly tessellated templates that had been warped to account for subject-specific anatomy using a spatial non-linear transformation (inverse normalization of MRIs, see Mattout et al., 2007). Coregistration of functional data (MEG and EEG sensor locations) and anatomical information (head-model including source domain) in a common framework was achieved for both modalities separately using a rigid spatial transformation matching coordinates of fiducials and head-shapes specified relative to both sensor space and MRI space. For MEG data, head position was averaged across FRQ and INT sessions to allow for a common forward model between conditions. For each participant and each modality, computation of accurate BEM was performed with the software Openmeeg (`http://openmeeg.github.io`) as it was shown to outperform other traditional BEM methods (Gramfort et al., 2010). Re-referencing to the average mastoids was applied to EEG BEM. The resulting lead-field operator or gain-matrix $L \in \mathbb{R}^{N_c \times N_s}$ (with N_c sensors and N_s sources) embodying the pre-cited anatomical and biophysical assumptions, enters the following linear generative model M of data Y:

$$Y = LJ + \varepsilon_n \tag{6.1}$$

where J represents source distribution, i.e. the magnitude of dipole at each node of the cortical mesh, and ε_n represents the residual or error term, accounting for the fact that Y may provide partial and noisy information about J and that approximations enter L. The linearity of this generative model derives from the normal constraint on dipole orientation.

Model inversion using Multiple Sparse Priors (MSP)

Within a hierarchical Bayesian framework, we defined J as a multivariate Gaussian distribution of the form $J \sim \mathcal{N}(0, C_s)$ with $C_s \in \mathbb{R}^{N_s \times N_s}$ the (unknown) spatial covariance of sources. We assumed a multivariate Gaussian measurement noise $\varepsilon_n \sim \mathcal{N}(0, C_n)$ with $C_n \in \mathbb{R}^{N_m \times N_m}$ the (unknown) spatial covariance of measurement noise (relatively to a normalized spatial space composed of N_m modes that will be defined in the following section). We used Multiple Sparse Priors (Friston, Harrison, et al., 2008) to estimate the distribution J that satisfies the general equation of linear model with Gaussian errors:

$$\tilde{J} = C_s L^T (C_n + L C_s L^T)^{-1} Y \tag{6.2}$$

MSP also allows estimating C_s and C_n, the spatial source and noise covariances. Precisely, regarding source-level covariance, we defined C_s as a linear combination of N_p variance components $Q_i \in \mathbb{R}^{N_s \times N_s}$ (the Sparse Priors) weighted by hyperparameters λ_i:

$$C_s = \sum_{i=1}^{N_p} \lambda_i^s Q_i^s \tag{6.3}$$

We used SPM8 default Sparse Priors including 256 components in each hemisphere, each defined as a patch of nodes on the cortical mesh, whose spatial extent from a seed point derives from a Green's Function. In addition, we used a bilaterality constraint leading to a total of $N_p = 712$ variance components (hence 712 source hyperparameters to estimate). At the sensor level (relatively to normalized space), we considered one variance component per modality (each equal to identity matrix) weighted by its related hyperparameter:

$$\begin{cases} C_n = \lambda_{EEG} Q_{EEG} & \text{for EEG inversion} \\ C_n = \lambda_{MEG} Q_{MEG} & \text{for MEG inversion} \\ C_n = \lambda_{EEG} Q_{EEG} + \lambda_{MEG} Q_{MEG} & \text{for fused EEG-MEG inversion} \end{cases} \tag{6.4}$$

Initial source-level hyperparameters were set to the default values specified by the software. MSP rests upon expectation maximization (EM), a widely used variational inversion scheme and provides (Restricted Maximum Likelihood, ReML) estimates of posterior hyperparameters $\lambda = \left\{\lambda_1^s, ..., \lambda_{N_p}^s; \lambda_{modality}\right\}$ and Maximum A Priori (MAP) estimate of J using Eq. (6.2) (Friston et al., 2007). Posterior estimates of hyperparameters quantify the contribution of each variance component (Mattout et al., 2006). EM is an iterative process guided by the maximization of the free energy F, an approximation of the log-evidence of the model (the log-value of $p(Y|M)$, the probability of observing the data Y given the generative model M defined in Eq. (6.1)). Critically, F is composed of two counterbalancing terms: the accuracy indexing the quality of the fit, and the complexity which reflects the propension of the model to overfit the data and thereby its lack of generalizability.The higher the model evidence (hence the higher F), the better the model. EM stops when convergence on F is reached, furnishing posterior estimates of J,C_s and C_n and an approximation of model evidence.

Group-level inference

Group-level inference (Litvak & Friston, 2008) is a recent advance of MSP that aims at refining

prior source covariance C_s by accounting for the assumption that distribution J should be common to all participants. This is a two-step procedure (Figure 6.1.a) where Step 1 performs group-level data inversion using SPM8 default Sparse Priors. Iterative updates follow the principle of source sparsity implemented in the Greedy-Search (GS) approach (Friston, Chu, et al., 2008) that we used to select relevant variance components. Posterior estimates of source hyperparameters, referred to as *group priors* because they are informed by the variance of data at the group-level, furnish posterior C_s using Eq. (6.3), that becomes a (group-informed) prior entering Step 2. Precisely, Step 2 proceeds to individual-level MSP inversions with two/three hyperparameters to estimate: $\{\lambda_s, \lambda_{EEG}\}$, $\{\lambda_s, \lambda_{MEG}\}$ and $\{\lambda_s, \lambda_{EEG}, \lambda_{MEG}\}$ for EEG, MEG and fused MEG-EEG inversions respectively. We assigned all prior hyperparameters to the same default values as in Step 1. Prior to data inversion, group-level inference involves the normalization of individual sensor data in a common spatial-mode space (Friston, Harrison, et al., 2008). In short, this space is composed of N_m orthogonal virtual sensors (referred to as spatial modes) resulting from the singular value decomposition (SVD) of a group-informed gain matrix. Projection of individual data on these N_m spatial modes thus implies the rejection of sensor-level signals that could not be generated by the cortical sources involved in the selected SVD components (data reduction). Data reduction is also achieved using a subsequent projection of data on temporal modes (Friston, Henson, et al., 2006b). For each subject, spatially and temporally projected data $\tilde{Y}_i \in \mathbb{R}^{(N_m \times N_t)}$ is rescaled (using the trace of $\tilde{Y}_i \tilde{Y}_i^T$) to accommodate signal amplitude differences over spatial modes.

Fused MEG-EEG inversion

The fused MEG-EEG inversion approach proposed by Henson et al. (2009) was employed in the current study. This method entails the necessary rescaling of data and gain matrix over modalities to accommodate different physical nature of signals (hence different measurement units). This rescaling leads to two crucial aspects: (1) projected data on MEG and EEG spatial modes become homogeneous and (2) sensor-level hyperparameters λ_{EEG} and λ_{MEG} can be quantitatively compared to assess the relative contribution of each modality to account for the variance of data to be inverted. This second point was conducted using paired Student's t-tests in the case of MMN inversion ($[150, 200]$ ms) for condition FRQ and INT.

Comparative evaluation for separate and fused inversions

In this section, we propose a method based on Bayesian Model Comparison (BMC, Penny et al., 2010) for the quantitative comparison of separate (EEG, MEG) and fused (MEG-EEG) source reconstructions. Precisely, our approach evaluates the ability of each modality to separate different hypothesis (or models) represented by different source distributions defined over the source domain. Figure 6.1.b depicts the framework of our approach, which exploits the two-step characteristic of group-level inference (described in Figure 6.1.a). Group priors (resulting from Step 1 of group-level inference scheme) obtained with EEG, MEG and MEG-EEG inversions were considered of utmost importance (among the infinity of possible source distribution that could be addressed) for multimodal comparison as they entail the spatial information that could be inferred by each modality over the group of subjects. Hence, for each modality $mod \in \{EEG, MEG, MEG-EEG\}$, we defined three generative models $M_{mod,EEG}$, $M_{mod,MEG}$ and $M_{mod,MEG-EEG}$ that each embedded source covariance priors C_s resulting from Step 1 with modality EEG, MEG and MEG-EEG respectively. At the individual level (Step 2 of group-level inference scheme), a total of 9 cross-modal inversions were computed for each subject:

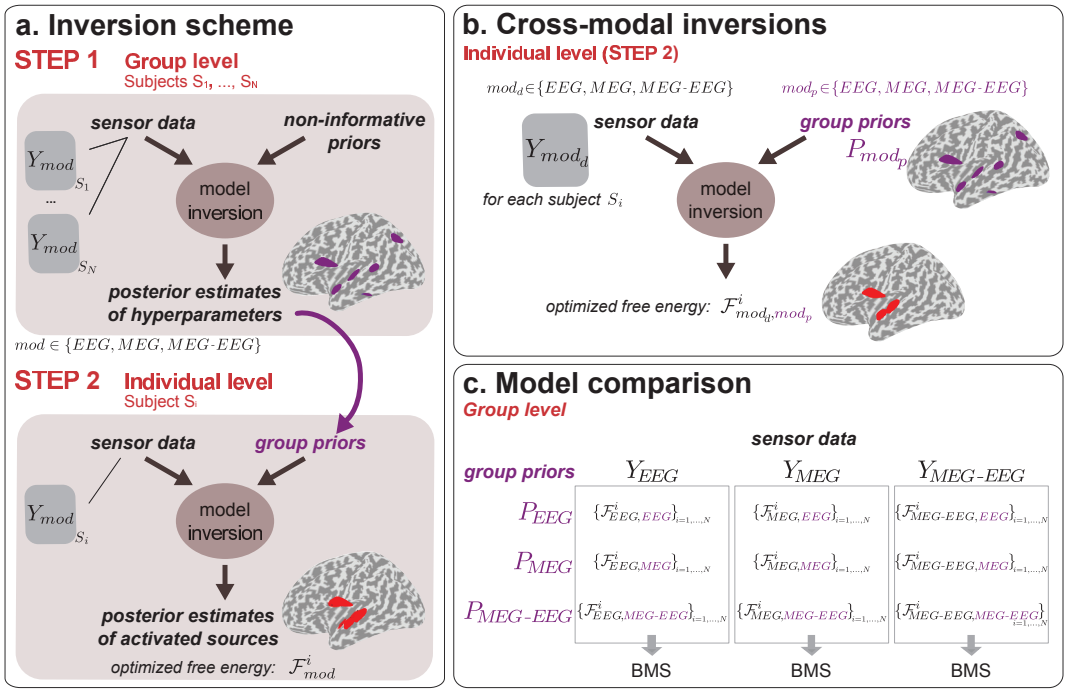

Figure 6.1 – Schematic view of the inversion scheme and the fusion evaluation procedure. (a) The two-step procedure of group-level inference (Litvak et al., 2008). *Y_mod* denotes observed EEG, MEG or EEG-MEG data. Step 1 derives the group priors illustrated as purple clusters on inflated cortical surface, Step 2 derives the individual source distribution J represented by red clusters on the cortical surface. (b) Procedure for cross-modal inversions: individual inversions (for each imaging modality) are conducted with group-priors inherited from the different group-level inversions achieved with EEG, MEG and EEG-MEG data. (c) Bayesian model selection (BMS) is conducted per modality (column) and rests on the free energy obtained for each individual inversion achieved with EEG (first row), MEG (second row) and MEG-EEG (third row) group priors.

three modalities for data ($mod_d \in \{EEG, MEG, MEG-EEG\}$) combined with three modalities for group priors ($mod_p \in \{EEG, MEG, MEG-EEG\}$). Thereafter, for each modality mod_d, the resulting free energy approximating evidences $p(Y_{mod_d}|M_{mod_d,EEG})$, $p(Y_{mod_d}|M_{mod_d,MEG})$ and $p(Y_{mod_d}|M_{mod_d,MEG-EEG})$ were compared across subjects using BMC with a random effect (RFX) model. BMC aimed at evaluating whether the three models M_{mod_d,mod_p} with $mod_p \in \{EEG, MEG, MEG-EEG\}$, could be separated by modality mod_d. To account for inter-individual variability, we computed the following free energy difference between models for each subject:

$$F_{mod_d,mod_{p=d}} - F_{mod_d,mod_{p,p\neq d}} \approx \frac{p(Y_{mod_d}|M_{mod_d,mod_{p=d}})}{p(Y_{mod_d}|M_{mod_d,mod_{p,p\neq d}})} \tag{6.5}$$

for each of the 9 cross-modal inversions. Following the usual principles of Kass and Raftery (1995), we interpreted the free energy differences as follows: a value comprised between -3 and 3 would indicate that both generative models (both group priors) have comparable evidence (we would then assume that modality used for inversion is not informed enough to disentangle these two models); a value larger (lower) than 3 corresponds to first (second) model having a greater evidence (for these two cases, we would then assume that modality mod_d used for inversion contains enough information to resolve corresponding models). For each BMC applied to modality mod_d (Figure 6.1.c), we expected model M_{mod_d,mod_p} with $d = p$ to be the best model as this would reflect that data Y_{mod_d} is sufficiently informed at the group-level to recognize (and prefer) the

group priors that it has generated. We also expected specific patterns across modalities by considering that EEG may have a lower spatial resolution than MEG (hence would be less informed) and that MEG and EEG contain complementary information (Lopes da Silva, 2013). Namely, we hypothesized that *i)* EEG inversion would also perform well with MEG and MEG-EEG group priors (EEG would have a rather low ability to resolve models), *ii)* MEG inversion would also perform well with MEG-EEG group priors but not with EEG ones, and *iii)* MEG-EEG inversion with its own group priors would outperform EEG and MEG group priors (fused inversion would have the larger model resolution). An original aspect of the proposed method pertains to the fact that it allows comparing quantitatively EEG, MEG and fused MEG-EEG source reconstructions applied to real (not simulated) data. We carried out this empirical evaluation for the FRQ and INT MMN as described below.

Empirical data for source reconstruction and multimodal evaluation

Data originate from a passive auditory oddball study with simultaneous MEG-EEG recordings where EEG analysis revealed two deviance responses: an early effect occurring within 70 ms after stimulus onset and a late effect (MMN) peaking at 170 ms post-stimulus (Lecaignard et al., 2015). We refer the reader to this study for a more detailed description of material and methods.

Participants
27 adults (14 female, mean age 25±4 years, ranging from 18 to 35) participated in this experiment. 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). Two participants have been added to the 5 ones excluded for EEG analysis (Lecaignard et al., 2015), because they had too low SNR for MEG data, leading the current analysis based on a total of 20 participants.

Experimental design
The oddball sequences embedding unpredictable occurrence of deviants (UF, UI) employed in the original study were retained for the current analysis, leading to two conditions for the present work that we rename here as FRQ (frequency deviance) and INT (intensity deviance). Both sequence types had the same deviant probability ($p = 0.17$). Two different frequencies ($f1$=500 Hz and $f2$=550 Hz) and two different intensities ($i1$=50 dB SL (sensation level) and $i2$=60 dB SL) were combined to define the four different stimuli that were used across conditions, with each condition (FRQ and INT) delivered twice in reverse sessions (standard and deviant physical properties were exchanged between sessions). Further details about stimuli and sequences can be found in Lecaignard et al. (2015). Participants were instructed to ignore the sounds and watch a silent movie of their choice with subtitles.

Data acquisition
The original study was conducted using simultaneous MEG and EEG recordings. Participants were seated upright in a comfortable armchair in a sound-attenuated, magnetically shielded recording room, at a 1 m distance from the screen. Sounds were presented binaurally through air-conducting tubes using Etymotic ER-3A foam earplugs (Etymotic Research, Inc. United States

of America). Sound level was adjusted individually according to participants' detection thresholds (performed before recordings using the sound with 500 Hz). Electrode positions relative to three anatomical fiducials (landmarks positioned at nasion, left and right pre-auricular points) were localized using a digitization stylus (Fastrak, Polhemus, Colchester, VT, USA). Special care was taken to minimize head position drifts between sessions. Finally, T1-weighted magnetic resonance imaging images (MRIs) of the head were obtained for each subject (Magnetom Sonata 1.5 T, Siemens, Erlangen, Germany). High MRI contrast markers were placed at fiducial locations to facilitate their pointing on MRIs hence minimizing anatomical and functional coregistration errors.

MEG recordings were carried out using a 275-channel whole-head MEG system (CTF-275 by VSM Medtech Inc.) with continuous sampling at a rate of 600Hz, a 0.016–150Hz filter bandwidth, and first-order spatial gradient noise cancellation. EEG recordings were carried simultaneously to MEG ones using the EEG recording system provided with the MEG equipment (same sampling rate and filter bandwidth). EEG data were collected from 63 electrodes whose locations were defined by the 10–5 extension of the international 10–20 system. Reference electrode and ground electrode were placed on the tip of the nose and left shoulder respectively. Throughout the recordings, impedances were below 15 $k\Omega$. Head position relative to the MEG sensors was acquired continuously (continuous sampling at a rate of 150 Hz) using head localization coils placed at fiducial points.

Data preprocessing and event-related field/potential (ERF/ERP) computation

The software package for electrophysiological analysis (ELAN, `http://elan.lyon.inserm.fr`) was used for ERF/ERP computation and statistical analysis. Continuous measures of fiducial position were averaged within each session to account for participant's head movement. Data segments corresponding to head movements larger than 15 mm relative to the average position and to SQUID jumps (for MEG data) were rejected. Further preprocessing included independent component analysis (ICA) correction for ocular artifacts and filtering (2-45 Hz band-pass digital filter, bidirectional Butterworth, 4th order), as detailed in Lecaignard et al. (2015). Filtered data were epoched from -200 ms to 410 ms post-stimulus (mean signal during pre-stimulus time-interval was removed for baseline correction). Importantly, we only used time epochs that survived the procedures applied for artifact rejection for both modalities. Responses to standards just preceding a deviant and to deviants were considered for averaging and difference responses (also referred to as deviance response) were obtained by subtracting the standard ERF/ERP from the deviant one. Importantly, EEG evoked responses were re-referenced to the averaged mastoid electrodes for the current study for compatibility with the forward model. The effect of deviance (deviant vs. standard) was tested with MEG and EEG responses (with averaged mastoid reference) for both conditions FRQ and INT. Statistical analysis included permutation tests (100,000 permutations) at each time sample with correction for multiple testing in the temporal dimension (see initial EEG study for a detailed description).

Data for source reconstruction

We used SPM8 software (Wellcome Department of Imaging Neuroscience, `http://www.fil.ion.ucl.ac.uk/spm`). Standard and deviant ERFs and ERPs (with averaged mastoid reference) were imported in SPM8 for both FRQ and INT conditions, and down-sampled (200Hz) for data reduction. We started with reconstruction of the cortical sources of the P50 component elicited by stan-

dards just preceding a deviant (from 60 to 70 ms) for each modality (EEG, MEG, MEG-EEG). This preliminary step constitutes a control for the validity of our inversion scheme. Following this, deviance-related reconstructions were estimated for difference responses (deviant-standard) for each condition separately (frequency, intensity), for each modality in three post-stimulus time windows: from 15 to 75 ms (early deviance effect), from 110 to 150 ms (MMN rising edge), and from 150 to 200 ms (MMN peak). Overall, a total of 21 separate inversions were computed for each of the 20 participants. In addition, comparative evaluation of separate (EEG, MEG) and fused (MEG-EEG) inversions was applied to the MMN source reconstruction ($[150, 200]$ms) for both conditions (FRQ, INT). Regarding data normalization, 7 and 21 spatial modes (explaining 99.0% and 99.9% of the group-informed gain matrix variance) were retained for EEG and MEG respectively. Data reduction using temporal modes was achieved for all inversions but the P50-standard ones (number of samples in time interval $[60, 70]$ ms was too low). For deviance-related inversions, the number of temporal modes allowing for 100.0% of the variance of the spatially projected data to be explained was equal to 6, 4 and 5 for $[15, 75]$ ms, $[110, 150]$ ms and $[150, 200]$ ms time intervals respectively.

Statistical analysis on source distributions
We conducted our statistical analyses at the group-level using the recent surface-based approach proposed in SPM12. Posterior estimates of source activity and associated variance at each node of the cortical mesh (the source domain) resulted from posteriors of $\tilde{J}$ and C_s. The energy of posterior mean was considered for statistical analysis. One-sample t-tests were performed at each node, thresholded at $p < 0.05$ with Family Wise Error (FWE) whole-brain correction. In addition, we imposed the size of subsequent significant clusters to be greater than 20 nodes. Distance between two local maxima within a cluster was constrained to be larger than 5 nodes.

Results

This section presents our results as follows: first we briefly report for both MEG and EEG the grand-average responses. Second, we describe the results obtained for the localization of the standard P50 generators using EEG, MEG and MEG-EEG (condition FRQ). We then present the comparative evaluation for EEG, MEG and fused inversions that we conducted with FRQ and INT difference responses, at the MMN peak ($[150, 200]$ ms). Finally, as multimodal comparison was in favor of fused MEG-EEG inversion, we report the corresponding sources obtained for the time intervals $[15, 75]$ ms, $[110, 150]$ ms and $[150, 200]$ ms in the difference responses, for both conditions FRQ and INT.

Sensor-level analysis

Grand-average responses at gradiometer MLP56 and electrode FCz for standard, deviant and difference responses for condition FRQ and INT are shown in Figure 6.2. Regarding EEG responses, the early deviance effect and the MMN described in Lecaignard et al. (2015) were recovered for both FRQ and INT conditions with the current group of 20 participants and average-mastoid reference. The emergence of the early effect was statistically significant from 15 to 55 ms and from 5 to 65 ms for FRQ and INT respectively. For the MMN, emergence was statistically significant from 115 to 210 ms and from 113 to 211 ms for FRQ and INT respectively. Regarding MEG

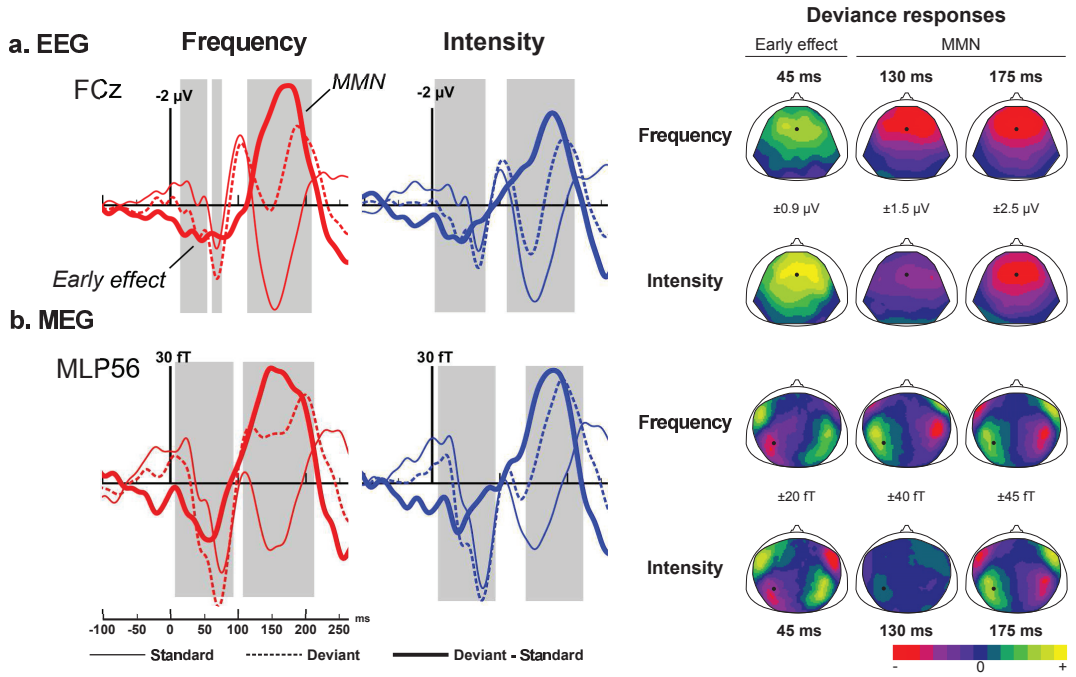

Figure 6.2 – Deviance responses measured with EEG and MEG. (a) EEG analysis with average mastoid reference. Left: Standard, deviant and difference evoked responses measured at frontal electrode FCz for frequency (red) and intensity (blue) deviance. Grey areas indicate significant deviance time-intervals. Right: Scalp topographies at relevant latencies for the early effect, the rising edge and the peak of the MMN, for frequency (upper row) and intensity (lower row) deviance. Voltage color scale is indicated for each time-window. (b) MEG analysis. Traces measured at left temporo-parietal gradiometer MLP56 are represented using the same color code as for EEG; Scalp maps at relevant latencies are represented on the left. Black dots on EEG and MEG scalp maps indicate electrode FCz and gradiometer MLP56 respectively.

responses, these two components could be also observed, with significant emergence from 5 to 90 ms, and 105 to 210 ms respectively for condition FRQ, and from 3 to 90 ms and 140 to 225 ms respectively for condition INT.

Both EEG and MEG traces show a tendency for the MMN in INT to start later than in condition FRQ, we therefore distinguished the rising edge of this component from the peak per se to increase the spatial sensitivity of reconstructions. The time windows were defined as follows: $[15, 75]$ ms for the early deviance effect, $[110, 150]$ ms for the rising edge of the MMN and $[150, 200]$ ms for the MMN peak. For the sources of the standard P50 component (condition FRQ) visual inspection of EEG and MEG standard responses led us to select the $[60, 70]$ ms time window.

Source reconstructions for the standard P50

For each modality, R the percentage of variance of data Y explained by the estimated source distribution $\tilde{J}$ (comparable to a goodness-of-fit measure) was equal on average across subjects to 82.0%(±27.5), 91.1%(±3.4) and 85.9%(±10.9) for EEG, MEG and MEG-EEG inversions respectively. Two subjects presented very low values of R for EEG inversion (but not for MEG), that could be explained by the small (adapted) amplitude of EEG standard responses within the short window of $[60, 70]$ ms (this was not observed for MEG data). Value of R without these two subjects was equal to 90.1%(±7.3). EEG source distribution comprised 18 significant clusters including posterior STG (expanding through STS), the superior frontal gyrus (SFG), the

inferior temporal gyrus (ITG), the posterior central gyrus and the intraparietal sulcus (in both hemispheres). Results with MEG inversions showed significant activity in 12 clusters distributed bilaterally in Heschl's gyrus (HG), posterior inferior frontal gyrus (IFG), posterior STG (including its inferior bank), middle temporal sulcus (MTS) and orbitofrontal regions. Activity in right temporo-parietal junction (TPJ) expanding through posterior central gyrus was also observed. Finally, fused MEG-EEG inversions revealed 8 clusters located in HG, posterior IFG, posterior STG (expanding through STS in left hemisphere, and TPJ in right one) and orbitofrontal regions bilaterally. Contribution from bilateral supratemporal planes with in particular HG (for MEG and MEG-EEG) is consistent with literature and led us assume that the framework employed in the current study (with pre-cited assumptions, BEM forward model and MSP group-level inference with default Sparse Priors), although limited by the ill-posed nature of the inverse problem, provide plausible estimation of auditory cortical activity.

Multimodal evaluation

For each modality, source reconstructions were computed for each subject using difference responses at the time interval $[150, 200]$ ms. The value of R for condition FRQ was equal on average to 95.1%($\pm$2.1), 94.2%($\pm$2.3) and 93.6%($\pm$2.6) for EEG, MEG and MEG-EEG inversions respectively. In condition INT, it was equal on average to 94.7%($\pm$2.5), 93.8%($\pm$2.3) and 93.1%($\pm$2.7) for EEG, MEG and MEG-EEG inversions respectively. Regarding the contribution of each modality (EEG, MEG) in the case of fused inversion, paired Student's t-tests were used to compare the mean values of hyperparameters $\lambda_{EEG}$and λ_{MEG}. In both condition FRQ and INT, inversions across subjects led to no significant difference between modalities (t(19)=1.30, p=0.21 for FRQ; t(19)=1.98, p=0.06 for INT).

Separate and fused MMN source distributions (qualitative comparison)

Figure 6.3 shows the results of the statistical analysis projected on the inflated cortical surface for each modality (EEG, MEG and MEG-EEG) and each condition (FRQ, INT). In condition FRQ, EEG inversion revealed activity in the supratemporal plane expanding anteriorly from the Planum Polare (PP) and in the lower bank of the posterior part of the STG in both hemispheres. No frontal area was found significant. MEG inversion indicated a large cluster in the supratemporal plane (number of nodes in cluster $k > 120$) expanding from the lateral part of HG through PP in both hemispheres. A bilateral frontal area was retrieved in the posterior part of the IFG. Smaller supratemporal clusters were found with fused MEG-EEG inversion: in the right hemisphere, one cluster (k=92) including the lateral part of HG and PP could be measured, and in the left hemisphere HG and PP were found in separate clusters (k=55 and 25 respectively). Bilateral clusters similar to MEG ones were found in the frontal lobe. In condition INT, the MMN with EEG inversion was associated with bilateral activity in the posterior part of the STG and the intraparietal sulcus. MEG inversion revealed activity in the supratemporal plane and the IFG in both hemispheres, similarly to FRQ condition. With fused inversion, largest clusters were found in the lateral part of HG in both hemispheres. Other activity in the right hemisphere was found significant in the posterior part of IFG, the posterior part of STG and in the inferior temporal gyrus (ITG).

Comparative evaluation for condition FRQ

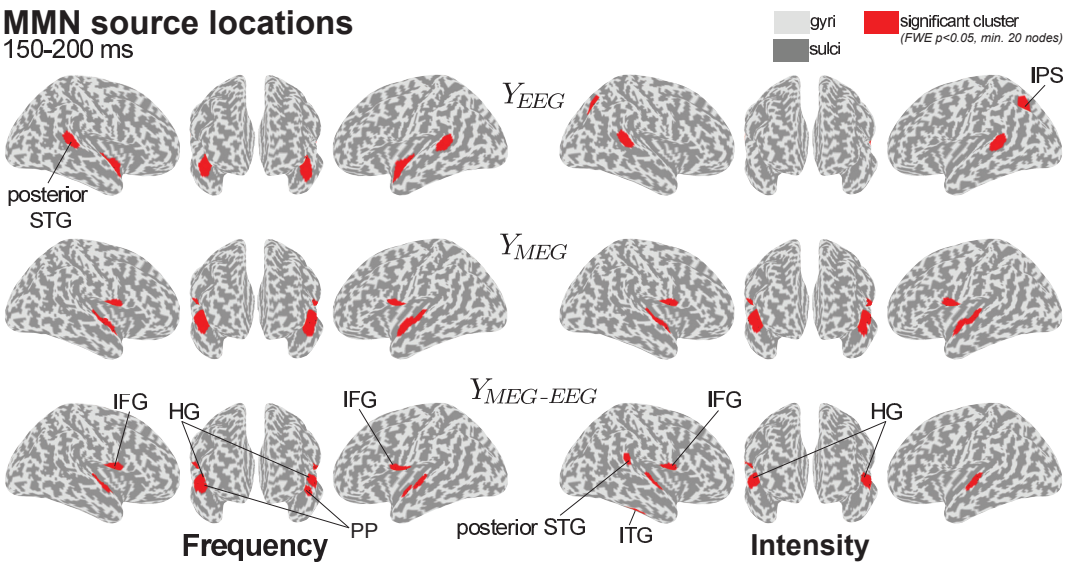

Figure 6.3 – Reconstructed sources of the MMN (150-200 ms). Significant clusters are represented in red on the canonical inflated cortical surface. First three columns from the left correspond to right, front and left views revealing frequency deviance generators, obtained with EEG (upper row), MEG (middle row) and MEG-EEG (lower row) data. The same three columns on the right show the results for intensity deviance.

Group priors obtained with each modality for condition FRQ are displayed on inflated cortical

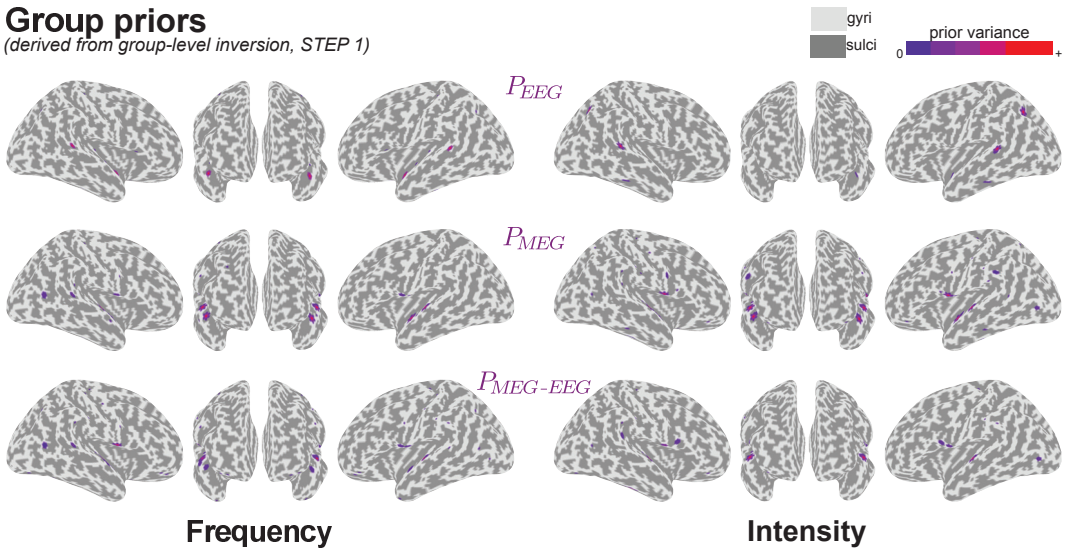

Figure 6.4 – Group priors for the MMN (150-200 ms). Purple areas represent the activated clusters resulting from group-level inversion that will enter as priors in subsequent individual inversions. Results are presented using the same organization as in Figure 6.3.

surfaces in Figure 6.4. Several differences can be observed between modalities from a visual inspection. First, priors in the supratemporal planes are mostly located more anteriorly in EEG than in MEG and MEG-EEG (although small spots in the vicinity of HG and PP are also found in EEG). Priors in bilateral posterior STG were found only for EEG, but priors located in right posterior MTG could be revealed with MEG and MEG-EEG. Regarding frontal priors, they were found in the same area (the posterior part of IFG) but with a far smaller extent in EEG. Priors in the intraparietal sulcus were identified with EEG only. In summary, MEG and MEG-EEG group priors appears similar (favoring posterior IFG and supratemporal planes including HG and PP

bilaterally), whereas EEG differs by strengthening priors in posterior STG and anterior temporal lobe.

Separate and fused inversion evaluation using these group priors were then conducted, with results shown in Figure 6.5. At the group level, RFX BMC (Figure 6.5.b) indicated that for all modalities EEG, MEG and MEG-EEG, the generative model embedding group priors derived from the same modality ($M_{mod_d,mod_{p=d}}$) had the greatest posterior probability. Precisely, the following model exceedance probabilities were found: $p(M_{EEG,EEG}|Y_{EEG}) = 1.00$, $p(M_{MEG,MEG}|Y_{MEG}) = 0.97$ and $p(M_{MEG-EEG,MEG-EEG}|Y_{MEG-EEG}) = 0.88$. In the latter case, model $M_{MEG-EEG,MEG-EEG}$ was followed by $M_{MEG-EEG,MEG}$ having a posterior probability of 0.11. Group priors obtained with EEG inversion were associated with posterior probability close to zero in all modalities but EEG.

At the individual level, free energy differences (see Eq. (6.5)) calculated for each of the 9

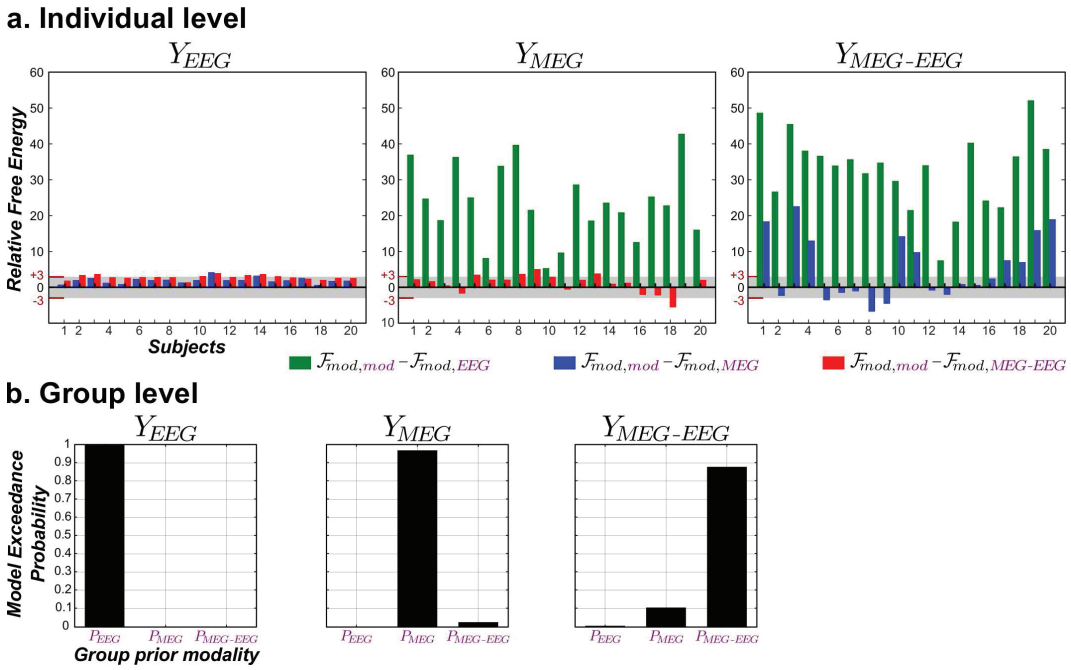

Figure 6.5 – Cross-modal evaluation the frequency MMN (150-200 ms). (a) Individual free energy differences obtained for EEG (left), MEG (middle) and fused (right) inversions; Grey area indicates the (-3 +3) interval corresponding to an absence of evidence between group-prior models. For each graphic, the color of the two bars assigned for each subject indicate which relative difference it represents following the color code presented below the three graphs. (b) BMS results using a RFX model for EEG (left), MEG (middle) and fused (right) inversions.

cross-modal inversions are displayed in Figure 6.5.a. For EEG inversion, MEG group priors compared to EEG ones were found equivalent ($-3 \leq F_{EEG,EEG} - F_{EEG,MEG} \leq 3$) across 18 subjects and different ($F_{EEG,EEG} - F_{EEG,MEG} > 3$) for 2 subjects; MEG-EEG group priors were found equivalent across 13 subjects and different ($F_{EEG,EEG} - F_{EEG,MEG-EEG} > 3$) for 7 subjects. For MEG inversion, EEG group priors induced lower free energy than MEG ones ($F_{MEG,MEG} - F_{MEG,EEG} > 3$) for all subjects; MEG-EEG group priors were found equivalent to MEG ones for 14 subjects, led to a positive difference ($F_{MEG,MEG} - F_{MEG,MEG-EEG} > 3$) for 5 subjects and negative difference ($F_{MEG,MEG} - F_{MEG,MEG-EEG} < -3$) for 1 subject. For MEG-EEG inversion, all subjects obtained lower free energy with EEG group priors than with

MEG-EEG ones; MEG group priors were found equivalent for 8 subjects, led to a positive difference ($F_{MEG-EEG,MEG-EEG} - F_{MEG-EEG,MEG} > 3$) for 9 subjects and a negative difference ($F_{MEG-EEG,MEG-EEG} - F_{MEG-EEG,MEG} < -3$) for 3 subjects.

Comparative evaluation for condition INT

Group priors obtained for condition INT are shown in Figure 6.4. From visual inspection we noticed that EEG group priors were located in bilateral posterior STG, ITG and intraparietal sulcus. Anterior part of the right supratemporal plane was also involved in this model. MEG group priors were predominantly distributed in lateral HG, PP and posterior IFG in both hemispheres. Smaller clusters were also located in the vicinity of temporo-parietal junction, the central sulcus, the occipital lobe, the orbitofrontal region and the ITG. Finally for MEG-EEG, group priors were located mostly in lateral HG, posterior IFG (two distinct clusters) in both hemispheres. Bilateral priors in posterior MTG and in the central sulcus, as well as in right posterior STG expanding through the temporo-parietal junction and in right ITG could also be identified. To sum up, similarly to condition FRQ, EEG inversion led to group priors that strongly differ from MEG and MEG-EEG (with namely the absence of priors in the IFG and in the area embedding HG and PP). Regarding MEG and MEG-EEG, contrary to condition FRQ, two noticeable differences can be reported: the absence of priors in PP for fused inversion, and the numerous clusters with small extent for MEG inversion.

Results of our comparative evaluation for condition INT are shown in Figure 6.6. Group-level model comparison using RFX BMC (Figure 6.6.b) showed greatest posterior probability for model $M_{mod_d,mod_{p=d}}$ in the case of EEG and MEG-EEG inversions, with $p(M_{EEG,EEG}|Y_{EEG}) = 1.00$ and $p(M_{MEG-EEG,MEG-EEG}|Y_{MEG-EEG}) = 0.97$. Regarding MEG inversion, BMC indicated comparable posterior probabilities evidences for MEG and MEG-EEG group-prior models (0.52 and 0.48 respectively). Group priors obtained with EEG inversion were associated with posterior probability equal to zero in all modalities but EEG.

At the individual level (Figure 6.6.a), for EEG inversion, MEG group priors were found equivalent to EEG ones ($-3 \leq F_{EEG,EEG} - F_{EEG,MEG} \leq 3$) for 15 subjects and induced lower free energy ($F_{EEG,EEG} - F_{EEG,MEG} > 3$) for 5 subjects; MEG-EEG group priors were found equivalent across 19 subjects and different ($F_{EEG,EEG} - F_{EEG,MEG-EEG} > 3$) for 1 subject. For MEG inversion, results are similar to those obtained with condition FRQ. In particular, EEG group priors also induced lower free energy than MEG ones for all subjects and MEG-EEG group priors were found equivalent to MEG ones for 14 subjects, led to a positive difference ($F_{MEG,MEG} - F_{MEG,MEG-EEG} > 3$) for 3 subjects and a negative difference ($F_{MEG,MEG} - F_{MEG,MEG-EEG} < -3$) for 3 subjects. For MEG-EEG inversion, EEG group priors induced lower free energy than MEG-EEG ones for all subjects; MEG group priors were found equivalent for 6 subjects, led to a positive difference ($F_{MEG-EEG,MEG-EEG} - F_{MEG-EEG,MEG} > 3$) for 12 subjects and a negative difference ($F_{MEG-EEG,MEG-EEG} - F_{MEG-EEG,MEG} < -3$) for 2 subjects.

Summary

Reconstructions of the sources of frequency and intensity MMN were performed using EEG, MEG and MEG-EEG inversions. Source distributions all provided a very good fit of data and for fused inversion, EEG and MEG contributed equally to the inversion process. Using MEG

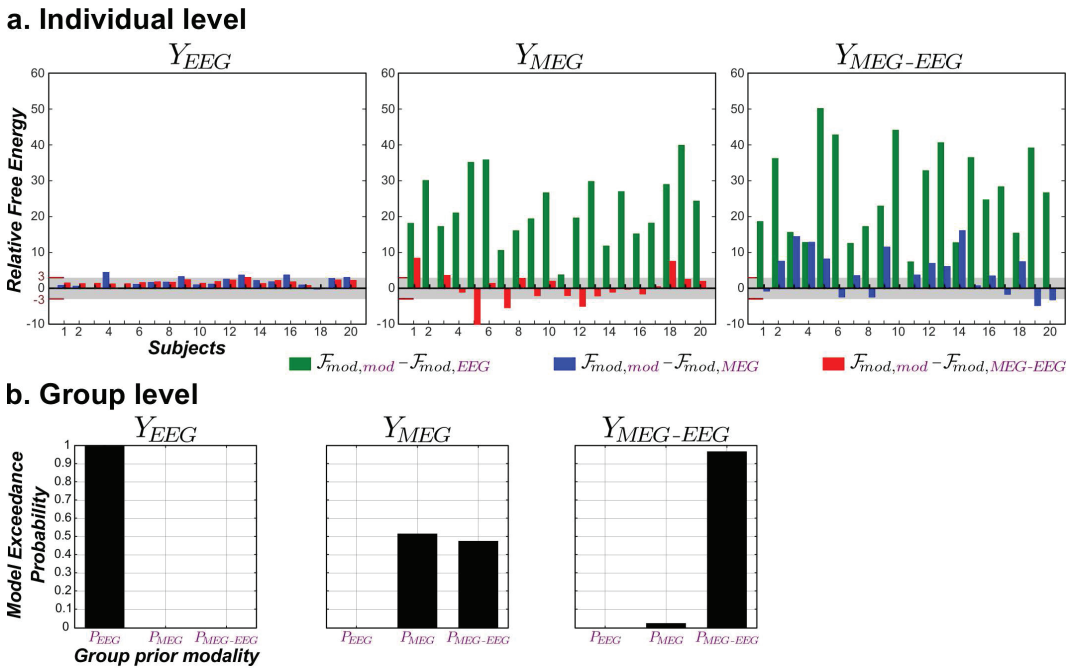

Figure 6.6 – Cross-modal evaluation the intensity MMN (150-200 ms).

and MEG-EEG, a bilateral fronto-temporal network could be identified, with a finer dissociation of supratemporal clusters obtained with fused inversion. EEG results differed from MEG and MEG-EEG ones, with notably an absence of significant frontal contribution Our model-based method for multimodal comparison indicated that on average across subjects, inversions with every modality perform better with their respective group priors (with the exception of MEG in condition INT). Considering within-subject effects:

- For EEG inversion, no significant difference could be measured for the majority of subjects ($N \geq 13$) between EEG and MEG or MEG-EEG group priors. Beside, EEG group priors always ($N = 20$) perform less successfully when used in MEG and MEG-EEG inversions.
- For MEG inversion, the use of MEG-EEG group priors led to non-different free energy for the majority of subjects ($N = 14$ in all conditions).
- For MEG-EEG inversion, the use of MEG-EEG group priors led to equal or larger free energy for the majority of subjects ($N \geq 17$). In condition INT (where visual inspection of MEG and MEG-EEG group-prior maps revealed larger difference compared to FRQ), relative free energy for MEG-EEG inversion showed a larger number of participants ($N = 12$) with MEG-EEG group priors outperforming MEG ones compared to condition FRQ ($N = 9$).

Our evaluation method thus revealed a better ability for fused MEG-EEG inversion to resolve spatial models (group-prior distributions). In the following section, we therefore present the deviance-related source reconstructions obtained only with this modality.

Fused MEG-EEG sources for auditory mismatch responses

Figure 6.7 shows the results obtained for each deviance type, each time interval with fused inversion.

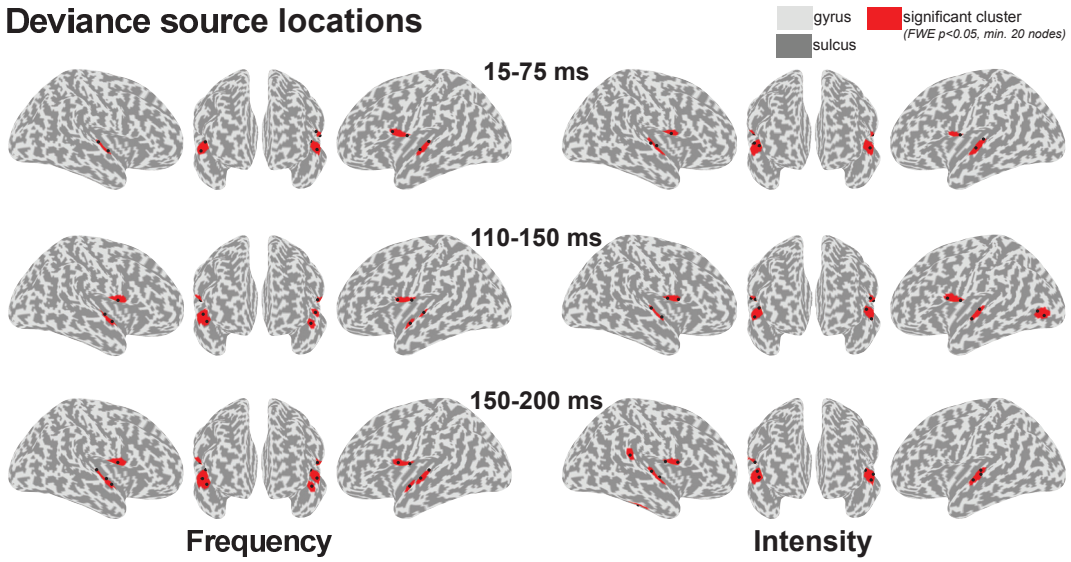

Figure 6.7 – Deviance generators obtained with fused inversion. First thee columns present right, front and left views for frequency deviance; following three columns present results for intensity deviance. For each deviance type, significant distributed network for early deviance (upper), the rising edge (middle) and the peak (lower) of the MMN are represented by red clusters. Black dots within clusters indicate local maxima.

Frequency-deviance sources

Reconstructions of difference responses within time windows $[15, 75]$ ms, $[110, 150]$ ms and $[150, 200]$ ms were associated with R equal on average to 90.7%($\pm$4.8), 92.3%($\pm$4.4) and 93.6%($\pm$2.6) respectively.

Early-deviance effect ($[15, 75]$) was found to involve HG in both hemispheres and left posterior IFG. Following this, reconstruction of the rising edge of the MMN ($[110, 150]$ ms) indicated supratemporal activity in HG and PP, within a large cluster in the right hemisphere (comprising two local maxima), and separated in two distinct clusters in the left hemisphere (with HG cluster being smaller). Significant activity was also found in bilateral posterior IFG. Finally, as described in previous section, the peak of the MMN ($[150, 200]$ ms) was associated with activity in both hemispheres peaking in HG, PP and posterior frontal IFG. The total number of significant sources within bilateral supratemporal planes was larger for the peak than for the rising edge of the MMN (178 and 108 respectively), while it remained constant within IFG (116 and 112 respectively).

Intensity-deviance sources

Reconstructions of difference responses within time windows $[15, 75]$ ms, $[110, 150]$ ms and $[150, 200]$ ms were associated with R-value equal on average to 91.4%($\pm$5.2), 90.5%($\pm$5.4) and 93.1%($\pm$2.7) respectively.

Within the early-deviance window ($[15, 75]$ ms), activity was more widespread in bilateral HG but was also found in posterior IFG. Reconstructions within $[110, 150]$ ms produced significant clusters in bilateral HG and posterior IFG. In addition, there was a contribution from left middle occipital gyrus (MOG). Finally, sources in HG and posterior IFG were observed in both hemispheres for the MMN peak reconstruction ($[150, 200]$ ms). Smaller clusters were found in ITG and posterior STG in the right hemisphere.

Summary
For both conditions, a fronto-temporal network could be retrieved for the three time windows corresponding to the deviance-related responses observed in ERP/ERF using fused MEG-EEG inversions. Frontal contributions could be measured as soon as the early deviance response window. Regarding temporal activity, in the frequency condition fused inversion allowed to separate HG and PP clusters spatially, but also temporally, as PP was not found associated with the rising edge of the component. For the intensity condition, supratemporal activity was found circumscribed to HG.

Discussion

In the present study, we conducted a source reconstruction analysis of auditory mismatch generators using state-of-the-art methodologies at every stage of the scheme, including a particular care for data coregistration, an accurate realistic forward model, a Bayesian-based group-level inversion procedure coupled with advanced surface-based statistical tools, and most of all the integration of simultaneous EEG and MEG recordings. Importantly, we also proposed a simple and efficient procedure resting on Bayesian model comparison for the quantitative evaluation of fused inversion with empirical data. The whole framework was validated with the reconstruction of the P50 generators. Applied to the mismatch responses, it enabled to locate bilateral sources in the supratemporal cortex and the IFG for both frequency and intensity deviances with subtle spatio-temporal difference between conditions. Beside, using our comparison approach, we showed that fused inversion provided an increased spatial resolution compared to unimodal ones.

Our findings reflect the different sensitivity of EEG and MEG measurements that both capture the (different) observable effects of the same hidden distributed neuronal activation. Contribution of the IFG could be revealed by MEG inversion for both deviance features but not by EEG. In fact, very few EEG studies reported such contribution like in Rinne et al. (2000), and most of them involved constrained ECD modeling (Jemel et al., 2002; MacLean et al., 2015). It should also be noted that very few MEG studies succeeded in observing such inferior frontal activation (Lappe et al., 2013a; Recasens et al., 2015). Regarding the activity within the supratemporal plane, modalities differed with EEG providing large clusters in posterior STG whereas MEG could reveal activity spanning over the expected primary auditory cortex. These findings thus confirm the long time acknowledged high potential of MEG to resolve temporal lobe activity (see for instance early MMN dipole studies, Alho, 1995). The different source distribution obtained with EEG and MEG should derive from the fact that they don't capture the same aspects of the underlying biophysical activity (originating from the same neural activation) (Lopes da Silva, 2013). Furthermore, the two imaging techniques do not have the same sensibility to various assumptions embedded in their respective forward models (*eg,* the conducting properties of the head, sensor positions and orientations). For instance EEG is more sensitive than MEG to white matter anisotropy changes within brain tissues (Güllmar et al., 2010), a property not accounted for by the forward models used here. Taken together, these findings support fused inversion to exploit these complementary EEG and MEG information, as was observed in the present study: only fused inversion could provide subtle dissociation within the supratemporal plane induced by

frequency and intensity deviances. This highlights the importance to include EEG information to improve MEG spatial resolution in the particular case of temporal activations.

This augmented performance of fused inversion could be measured quantitatively using a novel procedure that we proposed and that rests on the Bayesian framework employed in source reconstruction. Bayesian model comparison is a powerful tool to select which model is the more likely to have generated the observed data, taking into account for not only the fit of data but also the relative complexity of models that cause (undesired) data overfitting. By essence, BMC is not to be used to compare source inversion obtained with different data (in our case, EEG, MEG and MEG-EEG data). To circumvent this issue, we compared for each data type the inversions conducted with the different group priors obtained by each modality, that reflect the amount of information captured by each modality at the group-level. We thus exploited both the Bayesian scheme and group-level inference procedure to derive an easy-to-achieve comparison tool, estimating the capacity of each modality to resolve EEG, MEG and fused group priors. Application at the group level to auditory mismatch reconstruction suggest that each modality recognized its proper priors, but individual inspection could suggest the lack of information in EEG data that prevented to disentangle modality models, whereas MEG and MEG-EEG appeared sufficiently informed to do so. Moreover, fused inversion was found to outperform MEG as suggested by intensity deviance results where a larger resolution capacity could be measured. Importantly, no significant difference between the measurement noise estimate (hyperparameter λ_n) obtained for each modality in the case of fused inversion led us assume that both modalities equally contributed to the inversion scheme. The larger accuracy for fused inversion obtained here is definitely in accordance with expectations from simulation-based literature and could constitute the first robust empirical evidence resting on a sizeable group of subjects (N=20) and a quantitative procedure. Generalizability of the present results to other brain activations should be evaluated using the efficient scheme that we propose, that could further help to characterize the spatial complementarity of EEG and MEG recordings. In particular, this could help deciding whether or not simultaneous recordings remain advised (and under some technical adjustments, such as estimating the number of required sensors per modality), which is an important practical aspect to consider when designing a new experiment.

The fronto-temporal distributed network identified with fused inversion for frequency deviance processing, including the rising edge and the peak of the MMN component, is perfectly consistent with previous findings. However, the fine-grained description that could be established within the supra-temporal plane implies a spatial resolution that, to our knowledge, had rarely been attained before using either EEG or MEG distributed methods. In fact, the present findings are more comparable to those reported in the fMRI study of Schönwiesner and collaborators (2007), despite some differences (notably regarding the implication of posterior STG) that could be attributed to the duration deviance that they used. We failed to reveal other contributing generators often (but not systematically) reported, like sources in the parietal lobe or the ACC. This could be due to the MSP approach guiding the iterative optimization procedure behind model inversion: MSP implements the principle of sparsity (regarding the number of activated sources) and consequently progressively cancels out sources that are less likely than others to fit the observed data. Parietal or ACC sources could have been treated as less plausible candidates to explain the MMN than IFG and supra-temporal regions. Another noticeable aspect regarding our findings pertains to the

posterior to anterior progression that we measured between the rising edge and the MMN peak, that appears in keeping with several studies that had explored the N1 and the MMN generators (Scherg et al., 1989; Recasens, Grimm, Capilla, et al., 2014). These results support the (complex) multiple distributed activated area at play during auditory deviance processing.

Regarding the generators or the early response elicited with both deviance features, temporal activity was clearly circumscribed within bilateral Heschl's gyrus, that supports the adequate spatial resolution reached with fused inversion. These HG sources, as well as those measured for the P50 response in our face validity procedure are consistent with MLR findings from intracranial recordings studies (Liégeois-Chauvel et al., 1994), but they slightly differ from recent MEG findings in Recasens et al. (2014), where the Nb source could be located in right hemisphere, involving HG but also additional temporal areas. Bilateral contributions were however reported using a similar procedure in (Recasens, Grimm, Wollbrink, et al., 2014), with generators found in HG but also anterior part of the STG. Crucially, a major difference between these studies and the present findings pertains to frontal sources that we were able to recover. Under the assumption of a hierarchical organization for deviance processing that could unfold from subcortical areas to higher cognitive cortical levels (Escera & Malmierca, 2014), such frontal contribution as soon as these early latency becomes highly expected.

Finally, although investigating the difference between frequency and intensity deviance processing was beyond the scope of the study, distinct spatio-temporal patterns reconstructed in our study suggested similar early underlying mechanisms, followed by different neural processes taking place within the supratemporal plane and frontal regions. The anterior progression (from HG to PP) observed with frequency deviance could rest on the specificity of frequency, a more complex property of stimulus whose treatment could recrute additional auditory areas. Such different processing is largely accepted, that had been supported by early ECD MMN studies conducted with EEG (Giard et al., 1995) and MEG (Levänen et al., 1996) data. The observed interhemispheric frontal asymmetry for intensity deviance could conform several MMN studies that have reported a right hemisphere dominance (see for instance Paavilainen et al., 1991). Intensity deviance also indicated unexpected contributions such as the middle occipital gyrus. These results are not consistent with previous findings but it is worth recalling that (at least for this contribution in particular) the intensity MMN was not significant over the time interval used for this reconstruction (from 100 to 150 ms), as can be seen in Figure 6.2; we therefore assumed that these findings reflect local minima into which the iterative algorithm had converged. Such false positive results should also constitute a reminder of the ill-posed nature of the source reconstruction problem, that will always remain whatever the advanced methodologies integrated in the inversion framework.

Conclusion

Using advanced MEG-EEG localization methods, a bilateral fronto-temporal network conforming previous findings was identified for both frequency and intensity deviance ERPs. Interestingly, a high degree of spatial resolution could be attained with fused inversion, allowing an accurate spatio-temporal description that could reveal differences within the supratemporal plane between deviance types. Evaluation of spatial model resolution for each modality speaks clearly in favor of using fused MEG-EEG data as it proved best to disentangle spatial models, even compared

to MEG. Despite the fact that simultaneous acquisitions may appear less straightforward that unimodal ones in terms of experimental procedure, the present findings suggest that they should be considered as an attractive and powerful option that we recommend, particularly in the case of auditory studies. The refined spatial description of the auditory cortical hierarchy achieved here represents a crucial step prior to further hypothesis testing regarding the neurophysiological and computational mechanisms behind mismatch responses.

Chapter 7

Neurophysiological modeling of deviance responses: insights from predictability manipulation

7.1 Introduction

The predictive coding account of auditory processing rests on precision-weighted prediction errors generated by unexpected sounds along the auditory hierarchy. Characterizing the neural correlates of such errors could shed light onto the neurophysiological mechanisms behind this predictive framework in auditory processing. DCM and CMC (Kiebel et al., 2006; Bastos et al., 2012) has been proposed to tackle this challenging issue. These dynamical causal models aim at describing the effective connectivity at play during a mental process, and CMC relates explicitly the precision weighting to the intrinsic connectivity, and the prediction error to the forward extrinsic connectivity. As briefly reviewed in chapter 4, §4.2.3, some studies have addressed the effective connectivity behind deviance processing using DCM, with findings validating the ability of these models to predict electrophysiological mismatch responses on the one hand, and supporting the predictive coding message-passing scheme induced by unexpected sounds, the deviants, on the other hand.

Using a manipulation of deviance predictability in an EEG-MEG passive study, we could relate different mismatch responses (including the MMN) to prediction errors (Lecaignard et al., 2015). Precisely, we could observe that the more predictable the sound sequence, the smaller the mismatch responses, as expected under predictive coding. As discussed in chapter 5, this predictability effect involved the implicit learning of auditory regularities, and such (high-level) learning was hypothesized to shape the lower-level precision-weighted prediction errors. Importantly, an unresolved issue pertains to the fact that predictability could affect either the precision weighting or the prediction error or both. Using DCM with CMC, the present study aimed at characterizing the neurobiological underpinnings of auditory evoked responses during a passive oddball paradigm, and at assessing the influence of contextual manipulations on the deviance-related effective connectivity, with regard in particular to the ensuing adjustment of precision-weighted prediction errors. The present analysis, if it were to reveal specific neurophysiological changes in both the extrinsic and the intrinsic connectivity induced by regularity learning, would

thus establish a relation between electrophysiological data and Bayesian computation in the brain.

As a second aim, the acknowledged better performance obtained for source reconstruction using fused inversion (Dale & Sereno, 1993; Fuchs et al., 1998; Babiloni et al., 2004; Henson et al., 2009) and the consistent findings obtained with our empirical evaluation conducted in chapter 6 encouraged us to combine EEG and MEG DCM. Indeed, the information captured by each modality is complementary over the spatial dimension (as indicated with successful fused source inversions) and could arguably be so over the temporal one (DCM inversion is driven by the fit of data over this dimension). In the present study, we propose to combine posterior DCM estimates obtained with each modality (under the assumption of conditional independence of the data), to derive a *"posterior fusion" (p-MEEG)*. In addition, we also further attempted to fuse modalities within the generative model of DCM, that would arguably increase inversion performances as this would provide additional constraints in the model. The approach that we propose and the tests that we performed (with simulated and real data) to face its validity are presented in the Annex section of this chapter (§7.7). Basically, following the scheme for fused (static) source reconstruction proposed in Henson et al.(2009), we could extend the observation model of DCM to account for fused data. Despite convincing simulation-based results, our approach was found unsuccessful when applied to real data, that calls for further methodological improvements.

This work addresses the underlying neurophysiological mechanisms of deviance processing at the latency of the MMN, and their modulation by the contextual predictability. The overall study was conducted using the EEG and MEG mismatch responses obtained in the oddball study presented in chapter 5 and rested on the cortical network identified with fused source reconstruction described in previous chapter. This chapter is organized as follows: we first describe the general methods for the different DCM studies conducted here. We then present the first one (*Study 1*) addressing the characterization of the network structure for deviance processing (in terms of architecture and system inputs), followed by a presentation of the second one (*Study 2*), aiming at assessing the synaptic changes induced by deviant stimuli (namely the trial-specific effect between standard and deviant responses, see §4.2.2). The analysis of the predictability effect on deviance processing DCMs (UF vs. PF) is then described. Finally, in the last section, we discuss our findings.

7.2 Material and methods for all DCM studies

DCM studies rested on data originating from a passive auditory oddball study conducted with simultaneous MEG-EEG recordings (Lecaignard et al., 2015). This study comprised two types of frequency oddball sequences having different temporal structure: an unpredictable frequency sequence (condition UF) with deviant stimuli occurring pseudo-randomly, and a predictable frequency sequence (condition PF) with deviant stimuli occurring in a deterministic fashion. *Study 1* was dedicated to the characterization of the network structure for deviance processing and involved standard and deviant responses of a typical frequency oddball sequence (condition UF). *Study 2* was dedicated to the changes in connectivity induced by deviants within the structure selected in *Study 1*. It was first carried out with unpredictable data (UF) to refine the DCM of the MMN, and then conducted with predictable data (PF) to address the effect of the perceptual learning of statistical regularities on DCM connectivity.

Participants. 27 adults (14 female, mean age 25 ± 4 years, ranging from 18 to 35) participated in this experiment. 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). Five subjects were excluded from the analysis due to EEG or MEG artefacts, another due to individual MR images disclosing a ventriculomegaly, and another due to a failure to respect task instructions. The current analysis was thus based on a total of 20 participants.

Experimental design. Both unpredictable and predictable sequence types had the same deviant probability ($p = 0.17$). The original study also included an intensity oddball sequence (data not used here), that led us to manipulate the intensity and frequency attributes of sounds. Consequently, two different frequencies ($f1$=500 Hz and $f2$=550 Hz) and two different intensities ($i1$=50 dB SL (sensation level) and $i2$=60 dB SL) were combined to define the four different stimuli that were used across conditions, with each condition delivered twice in reverse sessions (standard and deviant physical properties were exchanged between sessions). Further details about stimuli and sequences can be found in Lecaignard et al. (2015). Participants were instructed to ignore the sounds and watch a silent movie of their choice with subtitles.

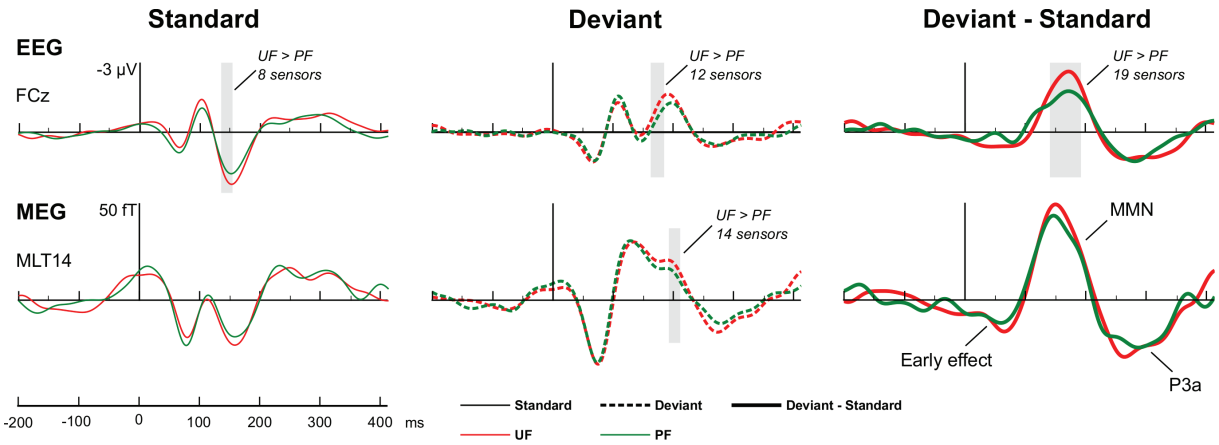

Figure 7.1 – EEG and MEG data for DCM analysis. Grand-average ERP (n = 20 participants) elicited by standards just preceding a deviant (solid line), deviants (dotted line) and difference responses (bold solid line) at electrode FCz (upper row) and gradiometer MLT14 (lower row) in bandwidth 2–20 Hz for condition UF (red) and PF (green). EEG responses with average mastoid reference. Shadowed area correspond to significant time interval for the comparison of UF and PF traces (p>0.01, corrected for multiple comparisons).

Data for model inversion. Simultaneous MEG and EEG recordings were collected using a 275-channel whole-head MEG system (CTF-275 by VSM Medtech Inc.) and the associated EEG recording system (63 electrodes) provided with the MEG equipment. We refer the reader to Lecaignard et al. (2015) and to chapter 6 for a detailed description of data collection and preprocessing. Responses to standards just preceding a deviant and to deviants were considered for averaging and difference responses were obtained by subtracting the standard ERF/ERP from the deviant one. EEG evoked responses were re-referenced to the averaged mastoid electrodes for the current study. Evoked responses epoched from -200 ms to 410 ms post-stimulus were imported in

SPM12 (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). Resulting data entering the current DCM analysis are shown in Figure 7.1. Statistical analysis described in Lecaignard et al. (2015) was replicated with the 2-20 Hz data, that confirmed 2-45 Hz findings on MMN amplitude. The present study focused on the MMN component and visual inspection of group-level traces led us to select the time interval of 0 ms to 220 ms for DCM inversion. Within DCM procedure, every data point was time-weighted by a Hanning window to ensure that system's dynamics was set to zero before being excited by the (thalamic) inputs. Finally, before starting data inversion, DCM procedure performs the projection of sensor data onto virtual sensors (referred to as spatial modes) for data reduction. Basically, this projection rests on the singular value decomposition (SVD) of the gain matrix mapping the N_s sources composing the DCM to the external sensors. Loosely speaking, this step corresponds to the rejection of sensor-level information that could not have been generated by the N_s sources of the DCM. A total of $N_s = 8$ sources were considered for the analysis presented here (see below), leading to an average of $N_m = 8(\pm 2.7)$ spatial modes with EEG data, and $N_m = 13$ spatial modes for every subject with MEG data.

Source locations. DCM architecture for both analysis was informed by the fine cortical source reconstructions performed with the UF difference responses using fused EEG-MEG inversion (see chapter 6). These inversions could reveal 6 bilateral clusters over Heschl's gyrus (HG), the planum polare (PP), and the inferior frontal gyrus (IFG). Each cluster was represented here by a central point being the spatial average of all the local maxima over the different time intervals. In addition, we assumed a bilateral superior frontal contribution (SF) that was motivated by our sensor-level findings but also by previous electrophysiological studies that had already reported a double contribution from frontal areas (see for instance, Marco-Pallarés et al., 2005; Fulham et al., 2014). More precisely, the predictability effect that could be captured with EEG (at fronto-central sites) but not with MEG (see chapter 5, §5.3.2) led us consider a further contribution to deviance processing that would express reliably on frontal electrodes and poorly on gradiometers. We attempted to locate such bilateral frontal contribution by reconstructing the sources of the

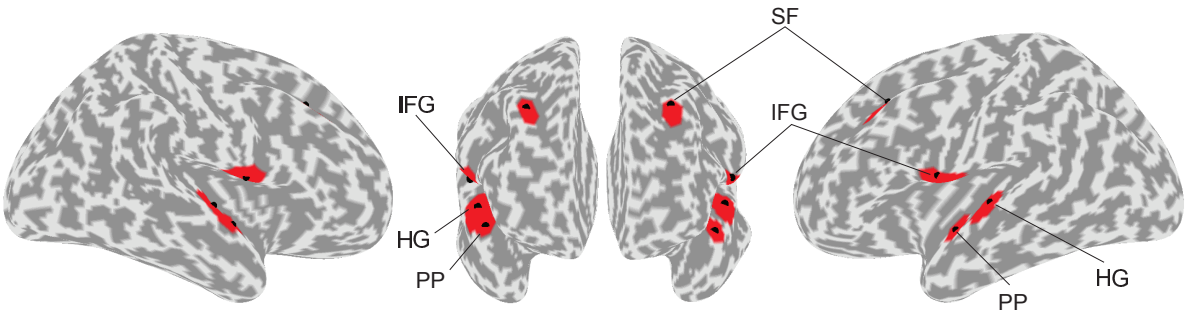

Figure 7.2 – Sources for DCM analysis. Projection of the eight ECD used for DCM architecture on a canonical surface in MNI space. Black dots indicate ECD locations, red clusters derive from the fused source reconstruction on deviance responses. HG=Heschl's gyrus; PP=planum polare; IFG=inferior frontal gyrus; SF=superior frontal. MNI coordinates: HG left (-51, -15, 4), right (52, -10, 5); PP left (-49, -8, -10), right (49, -7, -8). IFG left (-53, 3, 7), right (57, 2, 6). SF left (-28, 24, 44), right (30, 25, 41).

difference between UF and PF deviance responses over the significant time intervals reported in Lecaignard et al. (2015), namely 55-65 ms and 160-190 ms. Fused inversion revealed left

and right frontal clusters (36 and 29 nodes respectively, thresholded at $p < 0.05$ with Family Wise Error (FWE) whole-brain correction) for the former interval, and a left contribution of 34 nodes for the MMN interval (with $p < 0.001$ not corrected). Using the same procedure as for other clusters, these frontal contributions were summarized into a central point to define ECD locations. The eight resulting equivalent current dipoles (ECD) are represented on a canonical inflated cortical surface in Figure 7.2 (with MNI coordinates provided). Contrary to orientation, dipole locations were not estimated during (individual) DCM inversion in order to exploit the fine spatial information gathered at the group level from fused EEG and MEG source reconstruction. For each modality, the forward model used for DCM inversion was a realistic Boundary Element Model (BEM) (Hämäläinen & Sarvas, 1989) computed with the software Openmeeg (`http://openmeeg.github.io`) as it was shown to outperform other traditional BEM methods (Gramfort et al., 2010), as in chapter 6.

General DCM specifications. DCM analysis were performed with SPM12 (Wellcome Department of Imaging Neuroscience, `http://www.fil.ion.ucl.ac.ik/spm`). We used the CMC neural mass model (Bastos et al., 2012; Brown & Friston, 2013) described in chapter 4, §4.2 to exploit its relevance to test predictive coding predictions. For each evolution, observation and measurement noise parameters to be estimated through DCM inversion, we used default values of SPM12 as prior expectations and prior variance. Each DCM inversion involved standard response as the initial state of the system and deviant response resulting from the experimental perturbation (deviants possibly inducing changes in synaptic connectivity).

Posterior fusion of EEG and MEG DCM. The approach that we propose to combine EEG and MEG DCM rests on the assumption of the conditional independence of EEG and MEG data under the quasi-static approximation of Maxwell equations, which is largely admitted for signals below 1 kHz (as is the case here). This leads to:

$$p(y_{EEG}, y_{MEG}) = p(y_{EEG})p(y_{MEG}) \tag{7.1}$$

and to:

$$p(y_{EEG}, y_{MEG}|\theta) = p(y_{EEG}|\theta)p(y_{MEG}|\theta) \tag{7.2}$$

with θ the evolution parameters. The posterior distribution of θ given y_{EEG} and y_{MEG} writes:

$$p(\theta|y_{EEG}, y_{MEG}) = \frac{p(y_{EEG}, y_{MEG}|\theta)p(\theta)}{p(y_{EEG}, y_{MEG})} \tag{7.3}$$

from which we derive (with Eq.(7.1) and Eq.(7.2)):

$$p(\theta|y_{EEG}, y_{MEG}) = \frac{p(\theta|y_{EEG})p(\theta|y_{MEG})}{p(\theta)} \tag{7.4}$$

with $p(\theta|y_{EEG})$ and $p(\theta|y_{MEG})$ the posterior distributions of θ obtained with unimodal inversion of EEG and MEG data, respectively. 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 $q(\theta, y_{EEG}) \sim \mathcal{N}(\mu_e, \sigma_e)$, $q(\theta, y_{MEG}) \sim \mathcal{N}(\mu_m, \sigma_m)$, $q(\theta, y_{EEG}, y_{MEG}) \sim \mathcal{N}(\mu_p, \sigma_p)$ 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 $q(\theta, y_{MEG})$ (whose expressions can be found in most statistic books). Substituting these expressions in Eq. (7.4) gives:

$$\begin{cases} \sigma_p = \dfrac{\sigma_{em}\sigma_o}{\sigma_{em} + \sigma_o} \\ \mu_p = \sigma_p(\dfrac{\mu_e}{\sigma_e} + \dfrac{\mu_m}{\sigma_m}) \end{cases} \tag{7.5}$$

p-MEEG model evidence could be approximated using EEG and MEG model evidences as follows:

$$p(y_{EEG}, y_{MEG}|\theta) \approx p(y_{EEG}|\theta)p(y_{MEG}|\theta) \tag{7.6}$$

Consequently, F_p the variational free energy approximation to p-MEEG model log-evidence could be approximated by:

$$F_p \approx F_e + F_m \tag{7.7}$$

with F_e and F_m the free energy values for EEG and MEG respectively. In practice, for each DCM analysis, every model of the model space was inverted with EEG and MEG data separately, and resulting posterior EEG and MEG estimates of θ were subsequently combined using the expressions in Eq. (7.5) to derive the p-MEEG posterior distribution of θ, and the p-MEEG free energy was obtained by summing unimodal free energies.

7.3 DCM structure for deviance processing

Model space underlying this study (*Study 1*) was designed to assess the structure of the DCM for deviance processing, namely its architecture and its inputs. This study was conducted with condition UF.

7.3.1 Methods

DCM specifications. The scope of this analysis pertains to the characterization of the DCM structure for deviance processing, with regard to its architecture (*i.e.* the sources composing the network and the number of hierarchical levels) and the sources targeted by thalamic inputs (Figure 7.3). For DCM architecture, we hypothesized a four-level hierarchy composed of the eight ECD presented above, connected with intra-hemispherical bidirectional (forward and backward) connections. This model supports the contribution of every sources (HG, PP and IFG) identified with our source analysis (chapter 6) to the generation of the MMN. In addition, the inclusion of the superior frontal level models the plausible involvement of SF sources in the perceptual learning of the acoustic regularities. Alternative hypothesis entailed two- and three-level networks allowing to test the contribution of PP and SF sources. A total of five model families (*A1*, *A2*, *A3*, *A4* and *A5*) could thus be designed for the architecture issue, that are presented in Figure 7.3.a. Regarding DCM inputs, all models were designed with HG receiving the thalamic inputs. In addition, we also considered the possibility that they could arrive in IFG sources (as we know that the inferior frontal cortex receives thalamic afferents). This hypothesis was motivated by electrophysiological findings that could report frontal regions being activated before temporal ones

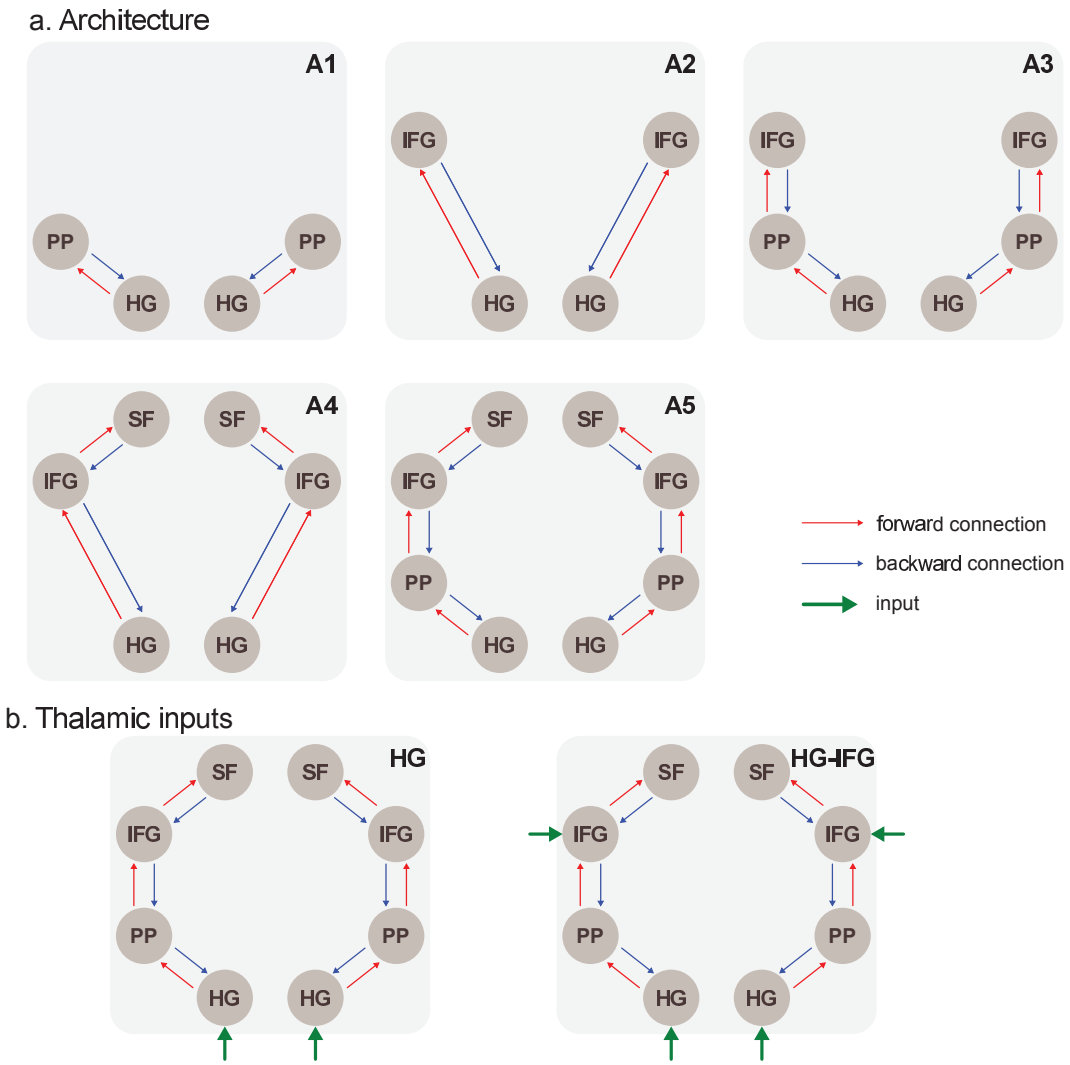

Figure 7.3 – Model space for *Study 1*. a) Architectures families. Five families of models (denoted A1, A2, A3, A4 and A5) were constructed with varying number of levels (from 2 to 4). b) Input families. Two families were considered (HG and HG-IFG). Red and blue arrows represent forward and backward connections respectively; Green arrows indicate where thalamic inputs enter the DCM. For every model of the model space, forward and extrinsic backward connections were allowed to be modulated with deviants. The combination of architecture and input families led to 9 models, as the HG-IFG input was not applicable to the architecture A1. The subsequent combination with factor $ModI$ (equal to 0 or 1) and with factor M (equal to 0 or 1) lend to 36 models for Study 1. HG= Heschl's gyrus; PP=planum polare; IFG=inferior frontal gyrus; SF=superior frontal.

(as mentioned in chapter 3, §3.3.2). Furthermore, the fact that we could measure a significant modulation of deviance responses by deviant predictability as soon as the P50 latency (Lecaignard et al., 2015) could also support the early involvement of these frontal regions. The input factor thereby included two model families (*HG* and *HG-IFG*) depicted in Figure 7.3.b. Each model constructed for *Study 1* included trial-specific modulation (between standard and deviant) of the forward ($ModF$) and backward ($ModB$) connections, that we denoted $ModF = 1$ and $ModB = 1$ respectively. Regarding the trial-specific modulation of intrinsic connections ($ModI$), we considered the case of an absence of such intrinsic modulations ($ModI = 0$) and the modulation of all sources ($ModI = 1$). Finally, we considered models having a modulatory connection (M, defined in chapter 4) for none or all sources ($M = 0$ and $M = 1$ respectively). A total of 36

models composed the model space for *Study 1*, partitioned over the architecture, input, intrinsic modulation and modulatory connection families. Importantly, only the architecture and input model subsets were considered for model family comparison (we assumed that the other two were not relevant to address the DCM structure issue, but related models allowed integrating over all possible parameter values). For each subject, these models were inverted separately for EEG and MEG data in condition UF.

Statistical analysis. The different model families specified above were quantitatively evaluated using family level inference (Penny et al., 2010) with a RFX model. Precisely, we performed such comparison over the architecture and the input families, for EEG, MEG and p-MEEG in condition UF.

7.3.2 Results

The 36 models designed for *Study 1* were inverted for each of the 20 subjects and for each modality (EEG, MEG) separately. The percentage of variance of data explained on average across subjects (n=20) and across models (n=36) was equal to 92.5%(±10.5) for EEG, and to 78.1%(±11.6) for MEG.

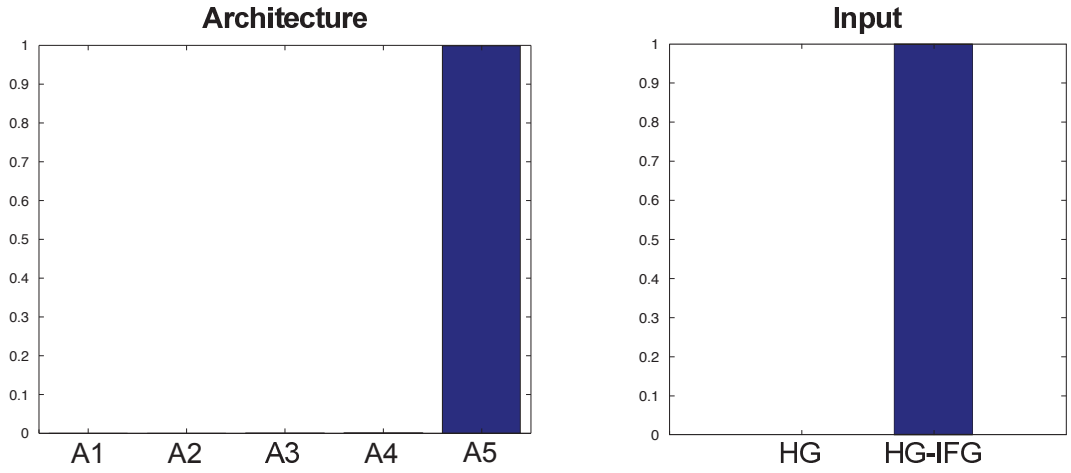

Figure 7.4 – Family level inference for Study 1 with p-MEEG DCM. Family exceedance probabilities are represented with bar diagrams for factors Architecture (left) and Input (right).

DCM architecture. Family level inference revealed that family $A5$, having four hierarchical levels with intra-hemispherical reciprocal connections, outperformed other families with both EEG and MEG data: posterior confidence and posterior exceedance probabilities for this family were equal to 0.48 and 0.81 for EEG, and to 0.60 and 0.98 for MEG. Using p-MEEG fusion, this statistical test was also in favor of $A5$, with corresponding posterior confidence and posterior exceedance probabilities equal to 0.68 and >0.99 respectively (Figure 7.4, left).

DCM inputs. Family level inference was clearly in favor of models with inputs arriving in both HG and IFG sources for both EEG and MEG inversion. Posterior confidence and posterior exceedance probabilities for family $HG-IFG$ were equal to 0.74 and 0.98 for EEG, and to 0.77 and >0.99 for MEG. Here again, family level inference performed with p-MEEG DCM strengthened unimodal

findings, with posterior confidence and posterior exceedance probabilities for $HG - IFG$ equal to 0.81 and >0.99 (Figure 7.4, right).

As every modality (EEG, MEG and p-MEEG) provided consistent and strong evidence for architecture $A5$ and $HG - IFG$ inputs, subsequent analysis presented in this study were conducted with this DCM structure. Besides, as p-MEEG findings proved consistent here, we will focus for the following reports on results derived from Bayesian comparison with this fused approach. Unimodal findings will also be reported, that could reveal the different (and complementary) sensibility of each modality.

7.4 Neural connectivity for deviance processing

Model space here enabled addressing the changes in connectivity induced by deviants. This study, Study 2, was applied to condition UF with the aim to further characterize the underlying effective connectivity behind the MMN.

7.4.1 Methods

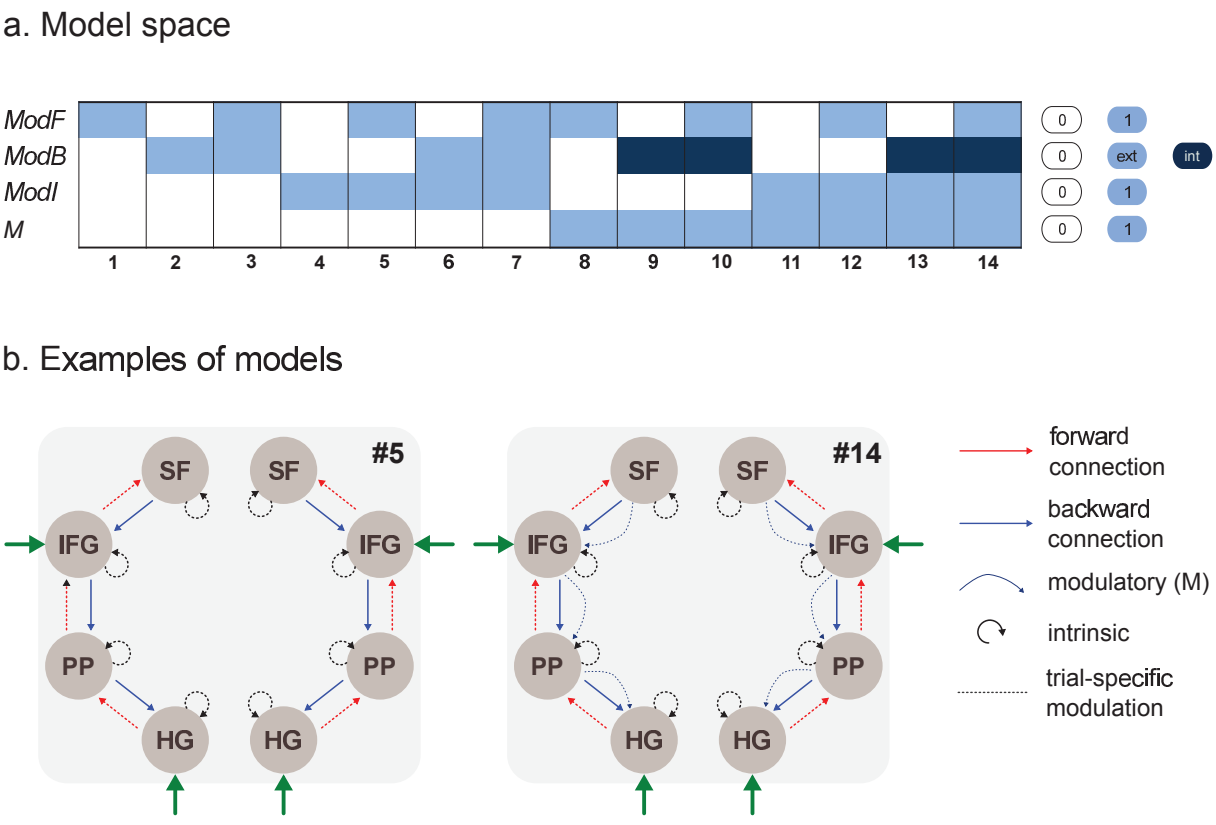

Figure 7.5 – Model space for *Study 2*. a) Schematic view of the 14 models (columns) resulting from the combination of factors ModF, ModB, ModI and M (rows). Note that the combination of factors M=1 and ModB=1 turns into a backward trial-specific modulation affecting the intrinsic connectivity, and gives ModB=Bint (dark blue), whereas the combination of M=1 and ModB=1 gives ModB=Bext (light blue). b) Two models from model space are illustrated. Red, blue, dark blue and black arrows represent forward, extrinsic backward, modulatory and intrinsic connections respectively; green arrows indicate where thalamic inputs enter the DCM. Dotted arrows indicate connections allowed to be modulated by deviants.

DCM specifications. This analysis aimed at characterizing how and where in the hierarchy deviant stimuli modify the effective connectivity observed during standard processing. This was done by considering the network structure selected by the family comparison conducted in *Study 1*, namely the four-level network with HG and IFG receiving inputs ($A5$). Model space was constructed as follows: for each connection types (forward, backward and intrinsic), trial-specific modulation could be allowed for none or all sources. Besides, we also tested the existence of a modulatory connection (M) at none or all sources, that could reflect a top-down influence on sensory precision. As was described in chapter 4, in the specific case of $M = 1$ in combination with $ModB = 1$, the backward modulation no longer affects the extrinsic backward connection but the intrinsic self-inhibitory connection of population SP (γ_7). We will refer to B_{ext} in the case of extrinsic backward modulation ($M = 0$ and $ModB = 1$) and to B_{int} in the case of intrinsic backward modulation ($M = 1$ and $ModB = 1$). A total of 14 models composed the model space for *Study 2* (Figure 7.5), partitioned over families corresponding to forward modulation ($ModF = 0$, $ModF = 1$), backward modulation ($ModB = 0$, $ModB = B_{int}$, $ModB = B_{ext}$), intrinsic modulation ($ModI = 0$, $ModI = 1$) and modulatory connection ($M = 0$, $M = 1$). Under predictive coding, we expected deviants to induce an increase of forward and backward extrinsic modulation, that would reflect larger ascending prediction errors and descending predictions elicited by deviants in comparison to standards. We also expected a decrease of intrinsic modulation that implies a decrease of self-inhibition within population SP assigned to the prediction error units: this would thus allow error units to send forward errors through the hierarchy when a deviant occurs. For each subject, these models were inverted separately for EEG and MEG data in condition UF and PF.

Statistical analysis. We performed family level inference with a RFX model over the factors forward modulation, backward modulation, intrinsic modulation and modulatory connection, for EEG, MEG and p-MEEG inversion of data with condition UF. For each of these factors, using Bayesian model averaging (BMA), we computed the group-level and the individual posterior estimates under the models composing the corresponding winning family. Group estimates were examined to assess the direction of changes (increase or decrease) for each trial-specific modulation (forward, backward and intrinsic) for condition UF.

7.4.2 Results

The percentage of variance of data explained on average across subjects (n=20) and across models (n=14) was equal to 96.5%(±5.1) for EEG, and to 87.1%(±7.4) for MEG.

Deviant effect on connectivity (p-MEEG analysis). The winning model indicated by the four family level inferences conducted on factors $ModF$, $ModB$, $ModI$ and M included a modulation of the gain of forward, backward (extrinsic) and intrinsic connections but rejected the modulatory connections (Figure 7.6.a). Posterior confidence and posterior exceedance probabilities were equal to 0.68 and 0.96 for models with $ModF = 1$, to 0.52 and 0.87 for $ModB = B_{ext}$, to 0.82 and $>$ 0.99 for $ModI = 1$ and to 0.68 and 0.96 for $M = 0$. More precisely, the examination of BMA posterior estimates (at the group-level) showed that over hemispheres and levels, an increase of the forward gain ($B^{\mathrm{f}} = 1.15 \pm 0.14$) as well as an increase of the backward gain ($B^{\mathrm{b}} = 1.18 \pm 0.18$) was measured with deviants (Figure 7.6.c). Regarding the intrinsic modulation, large variability

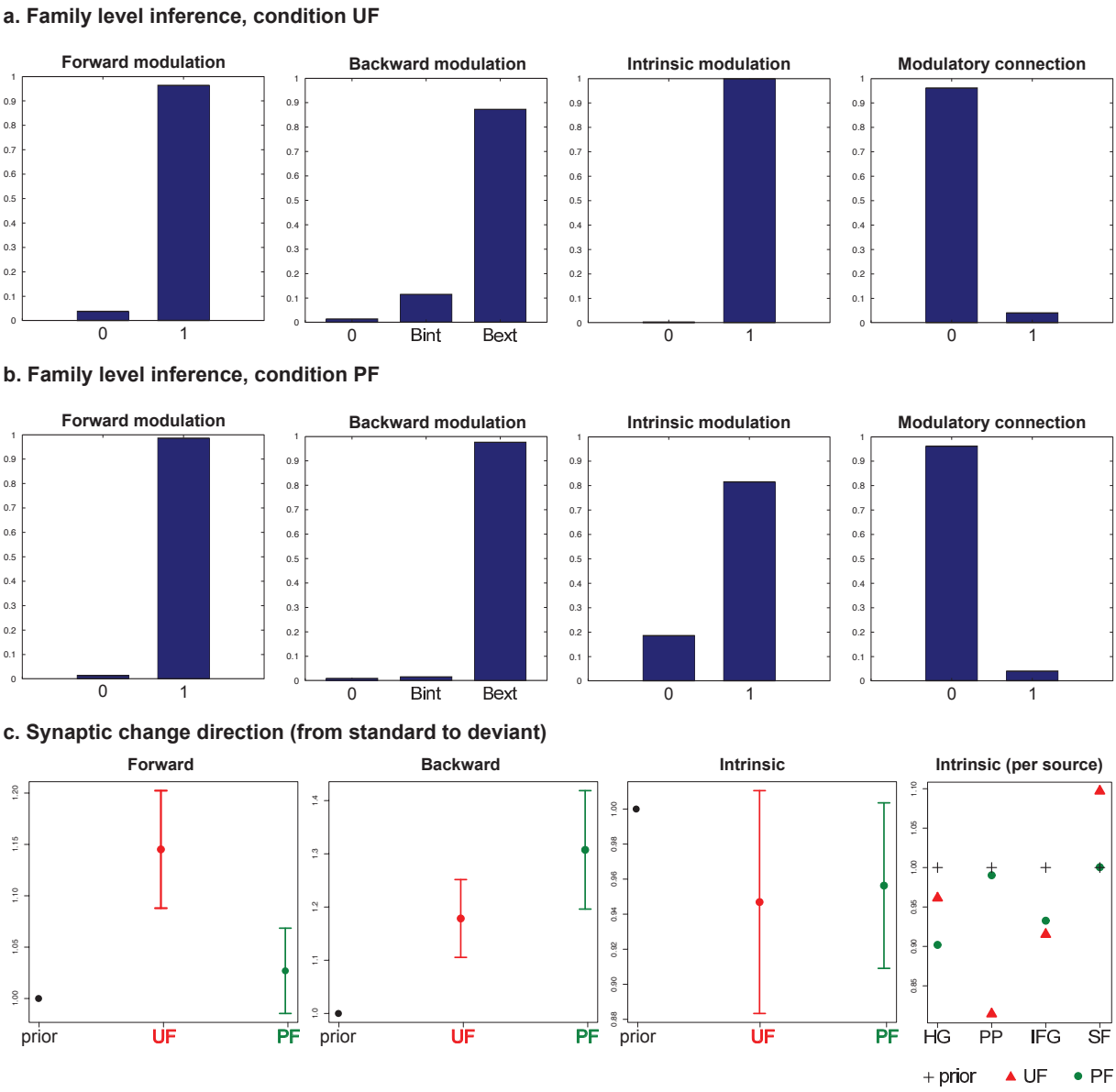

Figure 7.6 – Family level inference for *Study 2* with p-MEEG. a) Condition UF. Family exceedance probabilities are represented for the following factors (from left to right): forward modulation ($ModF$), backward modulation ($ModB$), intrinsic modulation ($ModI$) and modulatory connection (M). b) Condition PF. c) Direction of synaptic gain modulation induced by deviants for forward (first column), backward (second col.) and intrinsic (third col.) modulations. For each graph, black dots correspond to the prior expectation of trial-specific gain (equal to 1; left) and to BMA group posterior estimates (averaged over the network) for UF (middle) and PF (right); corresponding standard error bars are indicated. For intrinsic modulation, values per sources (average over hemispheres) are also provided (fourth col.)

between sources led us to consider BMA estimates per sources, that revealed a decrease of intrinsic gain with deviants (hence a decrease of self-inhibition in population SP) for all sources but SF ($B^{\mathrm{i}} = 0.96 \pm 0.25$, 0.92 ± 0.21, 0.81 ± 0.04 and 1.09 ± 0.18 for HG, PP, IFG and SF respectively).

Unimodal analysis. With EEG data, family level inferences were in favor of $ModF = 1$ (posterior confidence : posterior exceedance probabilities $= 0.64 : 0.91$), $ModB = B_{ext}$ $(0.43 : 0.69)$, $ModI = 1$ $(0.55 : 0.67)$ and $M = 0$ $(0.64 : 0.91)$. BMA posterior estimates indicated larger forward gain over every connections ($B^{\mathrm{f}} = 1.20 \pm 0.12$) , larger backward gain ($B^{\mathrm{b}} = 1.10 \pm 0.06$)

and larger intrinsic gain ($B^{\mathrm{i}} = 1.18 \pm 0.98$). This latter measure is inconsistent with MEG (see next) and p-MEEG findings and is associated to large variability across sources. Regarding MEG analysis, family level inferences revealed $ModF = 1$ $(0.74 : 0.99)$, $ModB = B_{ext}$ $(0.36 : 0.41)$, $ModI = 1$ $(0.86 : > 0.99)$ and $M = 0$ $(0.51 : 0.55)$. Similar change direction were measured with BMA posterior estimates compared to p-MEEG: larger forward gain over every connections ($B^{\mathrm{f}} = 1.20 \pm 0.16$), larger backward gain ($B^{\mathrm{b}} = 1.08 \pm 0.24$) and smaller intrinsic gain ($B^{\mathrm{i}} = 0.81 \pm 0.25$).

7.5 Predictability effect on deviance processing

This section concerns the analysis of the predictability effect on the effective connectivity behind the MMN. *Study 2* was replicated with condition PF. We then performed a statistical analysis (repeated-measures ANOVA) of posterior estimates obtained with conditions UF and PF.

7.5.1 Methods

Statistical analysis. Study 2 family level inference (described in previous section) was applied to the DCMs in condition PF. As in condition UF, we then computed the group-level and the individual BMA posterior estimates under the models composing the corresponding winning family. The evaluation of the synaptic changes induced by the predictability effect was achieved by comparing the BMA posterior estimates of specific DCM parameters obtained with UF and PF inversions. Precisely these parameters were:

- the gain of extrinsic forward connection (${A_{SS}}^{\mathrm{f}}$ and ${A_{DP}}^{\mathrm{f}}$ matrices, see chapter 4, §4.2.2); BMA was computed under the winning family regarding factor $ModF$,
- the gain of extrinsic backward connection (${A_{SP}}^{\mathrm{b}}$ and ${A_{II}}^{\mathrm{b}}$); BMA computed under the winning family for factor $ModB$,
- the gain of intrinsic self-inhibitory connection in population SP (γ_7); BMA computed under the winning family for factor $ModI$,
- the trial-specific modulation estimates between conditions (B); BMA computed under the winning family for factor $ModF$ for the forward-related terms, for factor $ModB$ for the backward-related terms and for factor $ModI$ for the intrinsic-related terms.

For each of these parameters, we conducted a repeated-measures ANOVA (over individual BMA estimates) with factors *Condition* (UF, PF), *Hemisphere* (Left, Right) and *Levels* (HG-PP, PP-IFG, IFG-SF); in the case of intrinsic-related parameters, this latter factor was replaced by factor *Sources* (HG, PP, IFG, SF). We only report main effects involving the *Condition* factor. Regarding connection strength parameters, a significant difference between UF and PF would indicate that the learning in PF had induced a change affecting the processing of both standard and deviant sounds. For trial-specific modulation parameters, different UF and PF estimates would suggest an influence of this learning on the deviant processing only. Differences affecting the self-inhibition in population SP (γ_7) were expected as they would reflect different precision-weighting of prediction errors between conditions.

7.5.2 Results

The percentage of variance of data explained by each of the 280 models inverted with PF data was equal on average to 97.5%(±3.0) for EEG, and to 86.1%(±8.5) for MEG.

Predictable-deviant effect on connectivity. Using PF data, family level inferences based on p-MEEG free energy were in favor of $ModF = 1$ $(0.73 : 0.99)$, $ModB = B_{ext}$ $(0.61 : 0.98)$, $ModI = 1$ $(0.59 : 0.82)$ and $M = 0$ $(0.68 : 0.96)$ (Figure 7.6.b). Direction of synaptic changes were similar than those observed with condition UF (Figure 7.6.c): larger forward gain over every connections (but to a lesser extent than UF; $B^{\mathrm{f}} = 1.02 \pm 0.10$), larger backward gain ($B^{\mathrm{b}} = 1.31 \pm 0.27$) and smaller intrinsic gain ($B^{\mathrm{i}} = 0.96 \pm 0.13$). Unimodal EEG analysis indicated a winning model with $ModF = 1$ $(0.68 : 0.96)$, $ModB = B_{ext}$ $(0.39 : 0.55)$, $ModI = 1$ $(0.58 : 0.78)$ and $M = 0$ $(0.55 : 0.68)$, which was consistent with MEG analysis where the following winning family were selected: $ModF = 1$ $(0.55 : 0.67)$, $ModB = B_{ext}$ $(0.61 : 0.97)$, $ModI = 1$ $(0.82 : >0.99)$ and $M = 0$ $(0.68 : 0.96)$.

Effect of deviant predictability on MMN connectivity (Figure 7.7). Using p-MEEG estimates, there was reduced backward connection strength for PF compared with UF over the right hemisphere (${A_{II}}^{\mathrm{b}}$;$F_{(1,19)} = 4.71$; $p = 0.03$). Interestingly, there was also reduced self-inhibition in population SP for PF compared with UF over every sources (γ_7; $F_{(1,19)} = 10.52$; $p = 0.001$). Over the left hemisphere, a tendency for smaller backward extrinsic modulation for UF compared to PF could be observed (B^{b}; $F_{(1,19)} = 3.47$; $p = 0.06$). No other difference between condition could be measured. Unimodal EEG analysis indicated larger forward connection strength over left hemisphere for UF (${A_{SS}}^{\mathrm{f}}$; $F_{(1,19)} = 7.09$; $p = 0.008$); MEG analysis also revealed reduced self-inhibition in population SP for PF compared with UF (γ_7; $F_{(1,19)} = 8.83$; $p = 0.003$).

7.6 Conclusion

This study aimed at improving our understanding of the neural mechanisms underlying auditory deviance processing in the light of predictive coding. Dynamic causal models with CMC allowed us to test mechanistic hypothesis regarding how MMN sources communicate in terms of effective connectivity within the auditory hierarchy and how and where deviant processing change this connectivity. Based on EEG and MEG evoked responses and a fused reconstruction of deviance generators, our DCM findings are in favor of a bilateral temporo-frontal structure composed of four levels with a twofold thalamic input arriving in bilateral HG and IFG. Interestingly, the changes of connectivity induced by deviants relative to standards conform to predictive coding expectations with larger forward and backward connections (on average over the whole structure) and reduced self-inhibition of neuronal population assigned to the error units of the predictive coding scheme. Moreover, the implicit learning of the temporal structure of oddball sequences that could be evidenced in Lecaignard et al. (2015) was found to influence the synaptic gain of specific DCM connections hosting strong assumptions regarding a Bayesian processing of information. EEG and MEG data provided consistent results but still revealed different sensitivity; their respective DCM could be fused *a posteriori* to fully exploit their complementary information.

To investigate the DCM of the MMN, the current approach was inspired from the initial studies

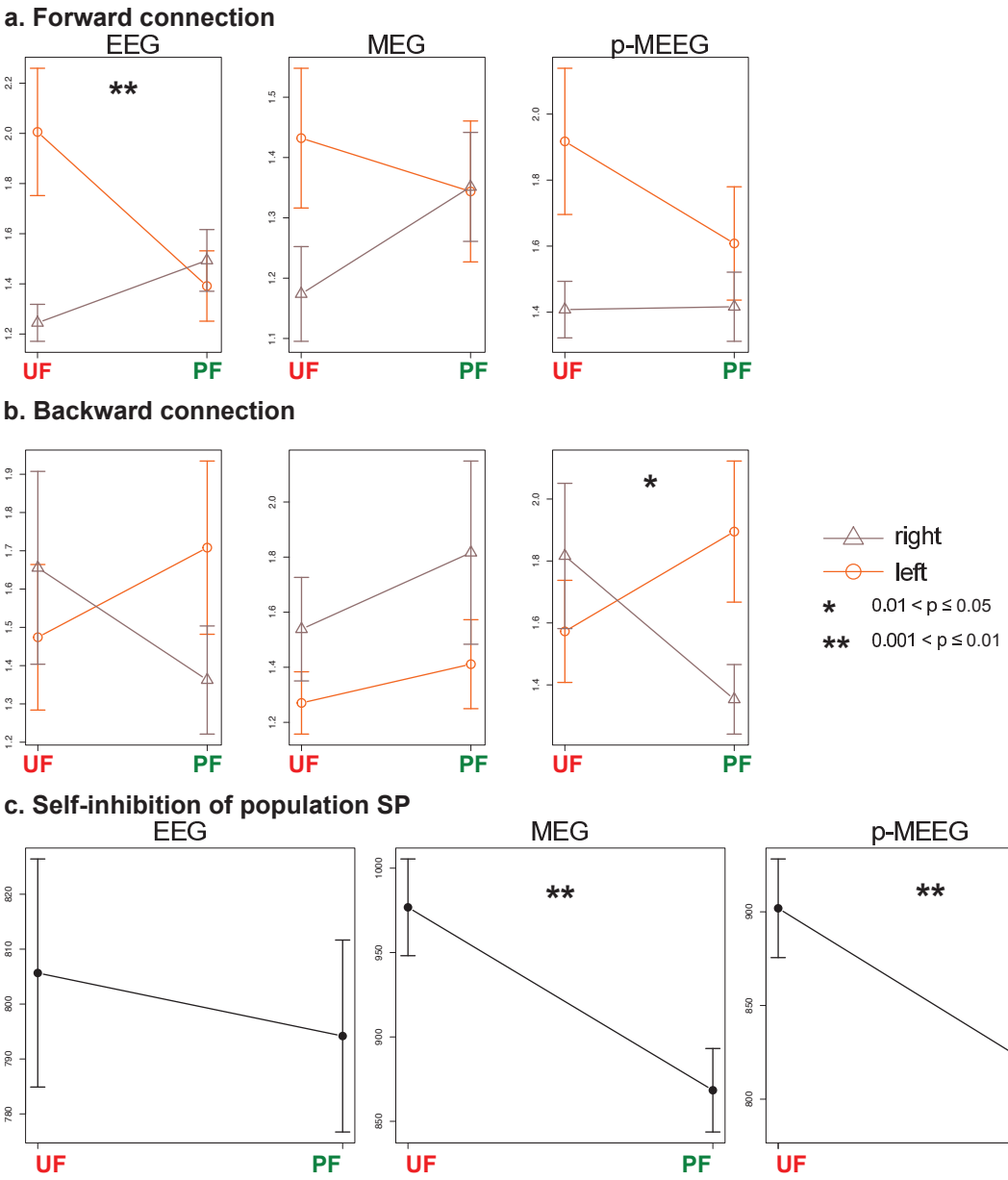

Figure 7.7 – Effect of deviant predictability on connection strength. a) Forward connections ($A_{SS}{}^{f}$). EEG, MEG and p-MEEG BMA group values of connections strength over levels are represented per hemisphere for UF (red) and PF (green) conditions. Standard error bars indicate variability between levels. b) Backward connections ($A_{II}{}^{b}$). c) Intrinsic connections (γ_7). BMA group values are represented over hemisphere and over sources, with resulting variability indicated by standard error bars. Red stars (if any) indicate significant main effects or interactions involving factor *Condition* (UF, PF) in the repeated-measure ANOVA.

of Garrido and collaborators (Garrido, Kilner, Stephan, & Friston, 2009) that exploited the suitability of DCM to evaluate quantitatively competing hypothesis about MMN generation. Most of subsequent MMN DCM studies (Boly et al., 2011; Moran, Campo, et al., 2013) employed the three-level structure proposed in Garrido et al. (2009), that was spatially informed by fMRI mismatch studies. In the present study, we constructed a model space with two- and three-level structures (with different supratemporal contributions compared to the studies of Garrido and colleagues) as well as a four-level structure (HG, PP, IFG and SF) motivated by previous electrophysiological source studies and our predictability analysis at the sensor- and source levels. The winning family (selected by BMS with RFX) suggested the contribution of both HG and PP,

consistent with early ECD findings reporting a posterior (at N1 latency) to anterior (at MMN latency) progression within this region (Alho, 1995). The four-level architecture outperformed other structures as expected, that could suggest the implication of superior frontal regions in the perceptual learning of sequence regularities (even in the case of a typical unpredictable oddball sequence, where such learning should also be achieved). Another notable result pertains to system input targeting two levels, namely HG and the IFG, for which family level inference provided strong evidence with both EEG and MEG data. This could suggest an efficient cortical scheme with parallel deviance processing in HG and IFG. This scheme would be triggered by thalamic deviant-related inputs, as we know from animal findings that unexpected sounds are already signaled within subcortical regions (Escera & Malmierca, 2014). Note that such processing would not challenge the hierarchical nature of deviance processing (required by predictive coding) as it is not incompatible with the distinct feedforward and feedback pathways. Regarding the trial-specific modulation induced by deviants, our results are clearly compatible with the predictive coding message-passing scheme as larger strength for forward and backward connections could reflect larger ascending prediction errors and descending predictions. This result was found consistently across conditions (UF, PF) and across modalities (EEG, MEG) and pursues the work by Garrido and colleagues by indicating the direction of synaptic change that could be assessed over the entire network. A decrease in intrinsic connectivity was also found more likely to explain the deviant effect (the MMN), an effect mostly seen with MEG data (across UF and PF). With CMC, such modulation corresponds to a reduced self-inhibition in population SP in charge of prediction error computations; we will discuss this interpretation in the following. Finally, the modulatory connection was not selected by the BMS (in both conditions), possibly due to the larger penalty term in the free energy associated to the related increase of DCM parameters, not counterbalanced by a better fit of the data.

Regarding the effect of sequence predictability on the effective connectivity for the MMN, it is first noteworthy that with MEG data, despite the lack of significant reduced MMN amplitude (but a small effect on deviant responses), DCM with MEG (and p-MEEG) succeeded in revealing a reliable effect between condition UF and PF. As models in *Study 2* could fit MEG data adequately leading to reliable MMN DCM analysis in both UF and PF conditions, this aspect could be regarded as the ability of a sophisticated generative model, biologically informed, to capture relevant information that could not reach significance (but was arguably present) with traditional ERP/ERF analysis.

As discussed in Lecaignard et al. (2015), we interpreted the decrease of MMN amplitude with predictability as deriving from the high-level implicit learning of sequence regularities that could influence information processing at low levels. In particular, such influence could affect both the prediction error and its precision within these levels, in a way that their respective characterization may be ambiguous. However, CMC has been explicitly conceived to resolve this issue (while ensuring the biological plausibility of this model Bastos et al., 2012). The current comparison of UF and PF DCM could involve two types of synaptic parameters: those representing the strength of connections, and those related to trial-specific modulation reflecting the deviant perturbation of the initial standard state. No effect on the latter could be observed (but a tendency with p-MEEG, not significant with unimodal analysis). Interestingly, for connection strength, a tendency of larger forward connection for UF compared to PF was visible over the left hemisphere (Figure

7.7.a) that reached significance with EEG. This could suggest synaptic changes induced by the regularity learning and affecting every sounds. From a computational perspective, it is interpreted as larger prediction errors (elicited by each item of the sequence) in the case of unpredictable sequence. The laterality of the effect could involve the left hemisphere specialization for temporal processing observed for speech but also non-speech stimuli (Zatorre & Belin, 2001). Reduced backward connections over the right hemisphere were revealed with p-MEEG for PF condition. Since this effect could not be seen with unimodal analysis, further investigations are required to clarify this point. Finally, the most notable result of this statistical analysis could pertain to the reduced self-inhibition for condition PF observed with MEG and p-MEEG. Combined with the effect of contextual predictability on the forward connectivity, these findings confirm the potential of CMC to disambiguate the two terms entering the *precision-weighted prediction errors*. In addition, reduced self-inhibition was observed in PF compared to UF, but also with deviant compared to standard. We further discuss the implications of these findings in the following paragraph.

Generalized predictive coding proposed in Friston et al. (2005) entails the weighting of ascending prediction errors by their precisions (inverse of variance) with larger confidence in these errors leading to larger forward messages. This endows the hierarchy with a filter to trigger updates only for reliable information. This computational scheme has been mapped onto neurophysiology with CMC: as explained in Brown and Friston (2013), DCM parameter γ_7 is the gain of the intrinsic inhibitory connection within population SP (assigned to the error units) and represents the negative log precision weighting. Low values of γ_7 thus correspond to *i)* at the neurophysiological level, low self-inhibition enabling the signal generated in SP to be send in a forward direction (or in other words, larger excitability within SP) and *ii)* at the computational level, large sensory precision up-weighting the ascending prediction errors. In Moran et al. (2013), the effect of cholinergic neuromodulation on mismatch responses was found more likely to affect self-inhibition in SP compared to extrinsic connection. In the current study, we observed a decrease of γ_7 for deviants (with UF and PF, mostly seen with MEG), hence a larger precision weighting that could reflect the fact that deviants are more informative than standards. Regarding the predictability effect, reduced γ_7 was observed with predictable sequences (PF), corresponding to larger up-weighting of prediction errors (if any), that could thus facilitate the processing of auditory inputs (for both stimulus types). It appears that a more efficient processing with contextual expectancy has already been reported for words but also music stimuli (Tillmann et al., 1998), as suggested by better task performances. Similar effect was also found in a recent fMRI visual study where fewer voxels being more informed were found involved in the processing of expected stimuli (Kok et al., 2012). Taken together, these findings conform to predictive coding predictions with the learning of environmental regularities influencing the precision weighting in order to improve the capacity for prediction updates when prediction errors occur (if they do).

Conclusion. Using DCM with CMC applied to EEG and MEG mismatch responses, we addressed the issue of characterizing the neural mechanisms behind the MMN, guided by the predictive coding account of sensory processing. Consistently over modalities, our results conform predictive coding predictions as they suggest larger ascending prediction errors, larger descending predictions and larger weighting of prediction errors observed along a hierarchical network for unexpected events. This analysis thus corroborates the biological plausibility of the predictive coding message-passing scheme that would implement the perceptual learning of environmental

regularities. Furthermore, the experimental manipulation of sequence predictability could suggest reduced prediction errors (over the left hemisphere) and larger precision weighting of predictions errors in the case of predictable sequences. Of course, the outstanding question of the functional role associated to the effective connectivity remains to be clarified in a more straightforward fashion, that could be addressed with computational modeling.

7.7 Annex: Fused EEG-MEG DCM attempts

This section presents our attempts to fuse EEG and MEG data within the inversion scheme of DCM. As already mentioned, this was motivated by the acknowledged complementarity of these modalities in the spatial dimension (evidenced by source studies based on simulations and empirical studies, including ours in chapter 6). Loosely speaking, DCM can be seen as a method for solving the non-linear inverse problem of source reconstruction, augmented with the temporal dimension. As such, one can see that it is limited by its ill-posed nature, calling for introducing additional independent information that EEG and MEG can provide. In the following, we start by describing our approach, that simply involved the adjustment of DCM observation model, then we present the material used for simulation and real data inversion. As will be described, ensuing results exhibited poor data fits with real data; we thus discuss the possible reasons for this failure and possible future improvements in the last paragraph. It should be noted that facing the validity of such approach should rest on a rigorous evaluation strategy aiming at quantifying fused inversion performances. This could not be done so far and in the current report, we mostly focused on the percentage of explained variance (R), quantifying the error between observed and reconstructed signals.

7.7.1 Methods

As mentioned in chapter 2, §2.4.2, the generative model of DCM is of the form (we omitted time t for clarity):

$$\begin{cases} \dot{x} = f(x, u, \theta) \\ \\ y = g(x, u, \psi) + \varepsilon_n \end{cases} \tag{7.8}$$

with x the hidden states describing neuronal activity, u the thalamic inputs entering the system, θ the evolution parameters and ψ the observation parameters. The residual term ε_n is defined as gaussian:

$$\varepsilon_n \sim \mathcal{N}(0\,; \Sigma_n) \tag{7.9}$$

with Σ_n the temporal covariance. In DCM, Σ_n is modeled explicitly as an autoregressive process of order 1:

$$\Sigma_n = h_e \times AR(1) \tag{7.10}$$

with h_e the hidden noise parameter to be inferred. Likewise fused source reconstruction, we aimed at inferring the neuronal activity that is exactly the same for both modalities but expresses

differently over electrodes and gradiometers. Precisely, we sought for θ and ψ that satisfy:

$$\begin{cases} \dot{x} = f(x, u, \theta) \\ \\ y_{EEG} = g_{EEG}(x, u, \psi) + \varepsilon_{n,EEG} \\ \\ y_{MEG} = g_{MEG}(x, u, \psi) + \varepsilon_{n,MEG} \end{cases} \tag{7.11}$$

As explained in Henson et al. (2009), combining modalities having different dimensionalities requires the generative model (precisely the observation model) to put them in the same framework. Current practice is to scale both the data and the forward model to accommodate these inhomogeneities. In fact, the data reduction framework employed in SPM, involving data projection to spatial and temporal modes (see chapter 6) nicely allows such scaling and a method was proposed in Henson et al. (2009) for source reconstruction. Projection of data is also performed in DCM (transposing data y to y^*), that allows the fit to be selectively driven by the activity that can only be generated within the DCM network (specified by the experimenter). In addition, SPM12 also includes the cited scaling of data. Importantly, at each iteration of the inversion scheme:

- Predicted data y_p are generated *at the sensor-level* for each modality using equation Eq. (7.8) applied to θ and ψ from previous iteration.
- Spatial projection (including the necessary scaling) is applied subsequently, leading to ${y_p}^*$ for each modality, with all time series originating from EEG and MEG being homogeneous data.
- These data are then subtracted to the observed projected data y^* to derive the prediction errors guiding the optimization scheme (VB-EM procedure) in order to infer θ, ψ and h_e, that will serve as priors for the following iteration (until convergence of the free energy is reached).

Regarding the residual term, contrary to source reconstruction where it is informed along the spatial dimension, it represents here time-lag relationships that we assumed to be similar between EEG and MEG signals. Prior expectation on h_e was defined to represent 1% of the variance of the projected data y^* (hence informed by both modalities).

Tests with simulated data. These tests aimed at checking the validity of the proposed scheme. This analysis rested on data from four participants (denoted $S1$, $S2$, $S3$ and $S4$) and for each of them, we performed as follows: we first inverted a specific DCM (three-level network with bidirectional connections, trial-specific modulation enabled for forward, backward and intrinsic connections) with real data using EEG inversion (DCMe), MEG inversion (DCMm) and fused inversion (DCMf). We used DCM with the neural mass model microcircuitry (ERP). Each inversion provided posterior estimates of θ and ψ that we used to generate synthetic EEG and MEG data, with additional gaussian noise (we used 10% of the total variance). We then inverted these data with DCMe, DCMm and DCMf (leading to 9 inversions). For the sake of time, since the use of openmeeg within DCM is very time-consuming, we conducted this analysis using the spherical models available in SPM (three-shell sphere for EEG, and local spheres for MEG). We enabled dipole locations to be inferred and likewise previous analysis, we applied a Hanning filter on data.

Application to real data. This was achieved using the same four participants data. We considered the model space of *Study 1* (36 models) described in section §7.2 and the forward model of openmeeg (with fixed dipoles). Here, we used DCM with CMC neural model. Three inversions were conducted per subjects (DCMe, DCMm, DCMf) for each of the 36 models, on the time interval 0 to 220 ms.

7.7.2 Results with simulated data

Inversions were performed successfully for every subject. Regarding spatial modes, 6 modes were retained for EEG data for all subjects, 11 modes for MEG data for $S1$ and $S3$, and 12 modes for $S2$ and $S4$. Figure 7.8 indicates the percentage of explained variance (R) obtained for each inversion. These values derived from the difference between the observed and the predicted data both projected on the spatial modes. These values could reasonably suggest a good quality of fit as they were always larger than 96%, with the exception of $S4$ where DCMf inversion of MEG data gave 85% and DCMm inversion of EEG data gave 80%. In addition, visual inspection of observed and reconstructed neuronal time series (within each source) as well as source location errors provided a reliable indication that the fused scheme had correctly fitted the data, and that it had performed equivalently to unimodal inversions.

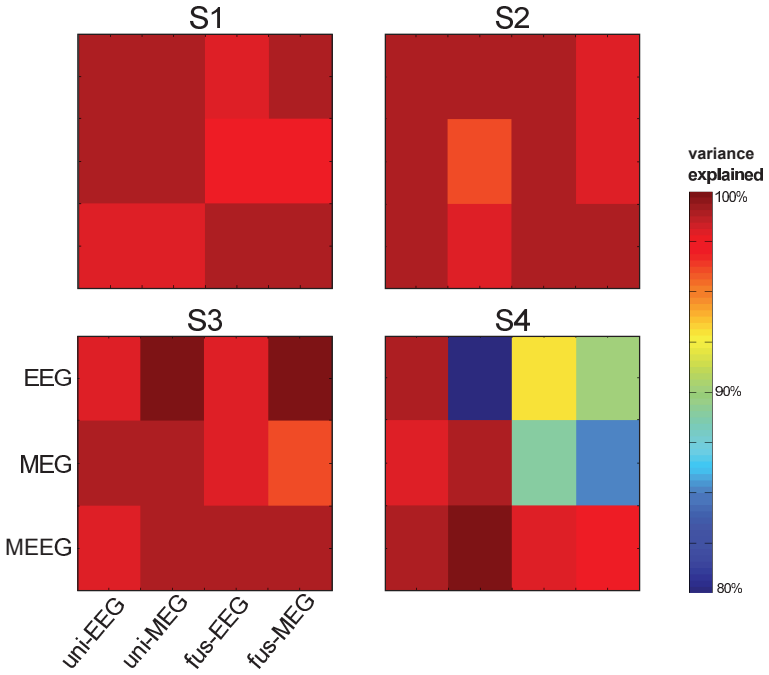

Figure 7.8 – Quality of data fit with simulation tests. For each subject, the values of R for each of the 9 inversions are represented by a colored matrix, where the rows represent the modality that has provided the evolution parameters required to generate the synthetic data, and the columns represent the modality used to compute R values: uni-EEG (uni-MEG) corresponds to EEG (MEG) data inverted with DCMe (DCMm). Fus-EEG and fus-MEG correspond to EEG and MEG data respectively, both estimated with DCMf inversion. R color scale ranges from 80% to 100% as indicated.

7.7.3 Results with real data

Using DCM with CMC and openmeeg BEM, we could perform DCM inversion of standard and deviant responses from 0 to 220 ms. Regarding data projection, 8, 11, 7 and 7 spatial modes were retained for EEG for $S1$, $S2$, $S3$ and $S4$ respectively, and 13 modes for MEG for all subjects. We first analyzed the quality of fit obtained using DCMe, DCMm and multimodal DCMf over the model space. Table 7.1 provides the averaged R values across the 36 models, for each data set

Subject	uni-EEG	uni-MEG	fus-EEG	fus-MEG
S1	85.6	68.5	46.3	61.5
S2	96.7	83.1	74.5	78.5
S3	86.6	72.6	57.6	74.9
S4	78.8	56.2	33.4	41.1

Table 7.1 – Averaged R values (%) across model space for real data. Labels uni-EEG, uni-MEG, fus-EEG and fus-MEG are referred to in Figure 7.8.

type (EEG or MEG) and for each DCM type. Fused values for each data set were consistently reduced with respect to those for modalities considered separately. It should be noticed that unimodal R values, for MEG in particular, indicated rather low data fit (see for instance subject $S4$, $R = 56.2\%$) but were associated to a large variability over model space. For instance, for subject $S4$, the model having the greater evidence with DCMm (model m_{34}) had a R value equal to 96.6% and 85.5% for EEG and MEG respectively in unimodal inversion, and to 46.0% and 53.9% for EEG and MEG respectively in fused inversion. Visual inspection of observed and reconstructed traces on spatial modes suggested different patterns behind these reduced R values with DCMf. Precisely, we observed that data fit could be inhomogeneous across spatial modes, or across trials (for instance standard responses were reliably reconstructed but not deviant ones), or both. An example is given in Figure 7.9, showing the traces obtained for the first spatial mode for subject $S4$ and for model m_{34}. In this particular case, for EEG, unimodal inversion perfectly modeled the observed data, whereas fused inversion failed to generate the deflexion around 150 ms. Regarding MEG data, fused inversion provided a better fit compared to unimodal inversion. Having examined the individual traces over the different spatial modes for each subject, we could see that EEG data was under-estimated in most cases. We checked that the scaling applied on projected data would not be involved, which was not the case as suggested in Figure 7.10. Finally, we examined the posterior estimates of h_e to assess whether residuals would be of the same order of magnitude between modalities (Table 7.2). On average across time samples (n=45) and models, we measured that for each subject EEG always presented the lowest value, followed by MEG and followed by fused inversion.

Subject	DCMe	DCMm	fDCMf
S1	0.0016	0.0051	0.0061
S2	0.0007	0.0023	0.0034
S3	0.0014	0.0035	0.0039
S4	0.0040	0.0054	0.0094

Table 7.2 – Average h_e values across model space across time samples measured with DCMe, DCMm and DCMf.

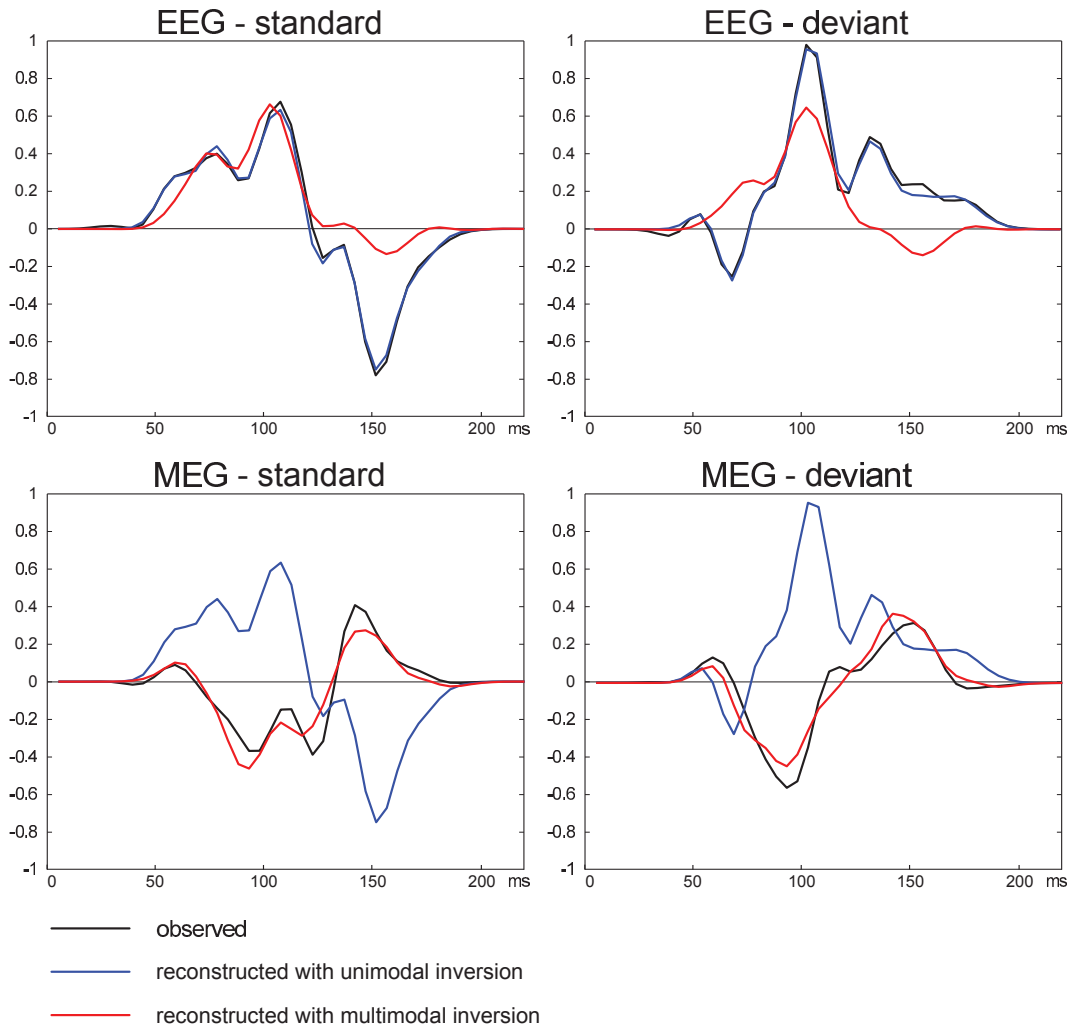

Figure 7.9 – Example of reconstructed data obtained with one subject ($S4$). For each modality (EEG, MEG, rows) and each experimental condition (standard, deviant, columns), observed data at first spatial mode (black) is represented to be compared with its estimates obtained with unimodal (blue) and fused (red) inversions.

7.7.4 Conclusion

Combining EEG and MEG within DCM rests on the adjustment of the observation model to accommodate the different properties of these independent signals. This could be achieved here by incorporating the scaling scheme proposed by Henson et al. (2009) for source reconstruction, that had been validated to improve inversion performances in the same study, and also in the present thesis (chapter 6). Despite persuasive simulation-based findings, going from theory to real data was found to fail in the current DCM study, with data fit being strongly reduced in comparison to those from unimodal inversions. Clearly, the work presented in this section constitutes a first attempt for fused DCM, resting on crude quantitative evaluation, and would require deeper investigations, that for a matter of time, could not be achieved during this PhD.

Among the different tests that could be further conducted, one in particular concerns the examination of the evolution and the observation posterior estimates. Indeed, in the current chapter, results obtained in *Study 2* with regard to the predictability effect suggested differences between modalities, with EEG possibly being more sensitive to extrinsic connection and MEG to intrinsic connections. It thus could be that such differences prevented the optimization procedure to

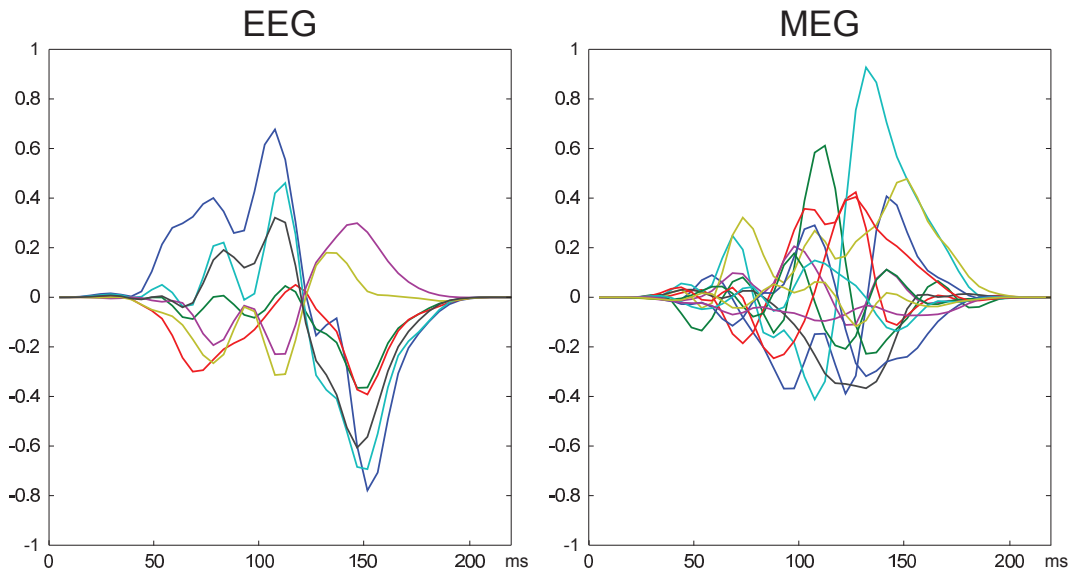

Figure 7.10 – Projected data y^* for EEG and MEG modalities (subject $S4$). The different modes retained for EEG (left) and MEG (right) are superimposed, that all correspond to homogeneous data due to application of scaling procedure inspired from Henson et al. (2009).

converge towards values being optimal for both modalities. This calls for a critical point that deserves a comment: it could be that our approach failed because of the complex nature of EEG and MEG signals which prevents present day's generative model to provide accurate output. Such issue could have been emphasized here as fusion would thus impose to accommodate model errors (in addition to *true* neuronal information) from both modalities. We know that EEG and MEG data comprise a mixture of neuronal contributions from various origins. DCM allows to model other contributions than the commonly accepted pyramidal cell ones (Nunez & Harth, 1982) but still ignores subcortical signals for instance. In addition, despite active research in the field, finely modeling the effect of the conducting properties of the head on sensor data still remains a great challenge, leading to forward models embedding acknowledged and admitted simplifications, whose influence on reconstructed data varies between EEG and MEG (Gencer & Acar, 2004; Vallaghé & Clerc, 2009; Güllmar et al., 2010; Acar & Makeig, 2013). Hence, the current failure could illustrate the fact that as the degree of sophistication of the (non-linear) model increases, the influence of the assumptions (and simplifications) entering the model increases.

The goal of fused DCM was to exploit the complementary information collected by EEG and MEG. Alternatives still allowing for such benefit could rest on sequential inversions, with posterior estimates from one modality inversion entering as priors for the subsequent inversion with the other modality. One could also consider to fit data reconstructed at the cortical level using fused source reconstruction, instead of fitting sensor-level data. This would require a simpler observation model, mapping for instance the CMC supra-pyramidal time series to these reconstructed cortical activities.

Chapter 8

Computational single-trial analysis of auditory responses: evidence for Bayesian learning at the MMN latency

8.1 Introduction

This chapter describes the computational modeling approach we used to characterize the cognitive mechanisms at play during deviance processing. Many ERP studies support the MMN generation to involve *i)* rule extraction mechanisms, that would endow the brain the capacity to represent the regularities of its environment (Bendixen et al., 2007), and *ii)* change detection mechanisms (at the origin of the MMN *per se*) signaling a violation of these regularities, with possibly further orienting of attention (under specific conditions of change saliency). Under the predictive coding hypothesis, these processes would be achieved within a more general (and versatile) scheme. Precisely, predictive coding assumes that the brain performs Bayesian inference to infer the causes of its sensory inputs in order to maintain an internal generative model of its environment (the *perceptual* model, see chapter 2 §2.4.1). This subsumes both perceptual inference to recognize the cause of the auditory input (for instance, a standard or a deviant), and perceptual learning to update model parameters (for instance, the probability to have a standard). Hence, each standard sound will contribute to increase the confidence in the model (or its precision), whereas unexpected deviants could be signaled by the prediction errors they induce, propagating upward along the hierarchy.

Bayesian learning in the brain for auditory processing can be investigated using computational models applied to oddball sequences. Such studies imply analyzing trial-by-trial responses to assess the trajectory of precision-weighted prediction errors and updates over the course of the experiment, provided that those are indeed computed in the brain. As indicated by the literature review in chapter 4 §4.3, only a few computational account of the MMN have been proposed so far. The approach proposed in the study of Ostwald and colleagues (2012) aimed at comparing learning and non-learning mechanisms with somatosensory deviance responses and revealed Bayesian

learning processes at relevant mismatch latencies. Both the efficiency of the approach and their results encouraged us to perform a similar analysis for auditory processing. Hence, the first goal of this study was to characterize (or to refute) perceptual learning underlying the processing of standards and deviants in an auditory oddball sequence. In addition, we were also interested in examining the effect of contextual manipulations. Precisely, using deviance responses embedded with unpredictable and predictable frequency oddball sequences (Lecaignard et al., 2015), we expected that the implicit learning of the predictable structure (brought out by the sensor-level analysis) would reduce prediction errors but would also increase their weighting by their estimated reliability. These predictions derive straightforwardly from a predictive coding scheme for auditory processing and were already corroborated by consistent changes in the effective connectivity underlying deviance processing revealed by our DCM analysis (chapter 7).

Importantly, we restricted our computational analysis to the previously identified cortical network using both EEG and MEG recordings, as described in chapter 6 and 7. In what follows, we first detail the material and methods of the study, with a description of the model space constructed to address our twofold aim. The results section then reports our findings, namely revealing the most likely implicit learning mechanisms at play during passive listening of the auditory oddball sequences, but also its modulations depending on the context. Finally, the implications of these findings are discussed.

8.2 Material and methods

As mentioned in the previous section, this study was based on the EEG and MEG data originating from our passive auditory oddball study conducted with unpredictable and predictable frequency oddball sequences (conditions UF and PF, respectively). The current analysis is twofold. *Analysis 1* was performed similarly on all time series from both conditions and aimed at testing whether any learning model could account for trial-by-trial variations in the ERPs. Subsequent *Analysis 2* involved separate UF and PF data inversion to address the predictability effect on perceptual learning. This latter analysis was conducted over the MMN time interval only.

Experimental design and participants. Data used from the current study derive from the experimental design described earlier (Lecaignard et al., 2015) and rested on the acquisitions from the 20 participants included in the source reconstruction and DCM analysis (chapter 6, chapter 7).

Preprocessing and data feature extraction. Computational model inversion aimed at fitting the source-level trial-by-trial data within each cortical cluster identified by the fused reconstruction analysis on difference responses (deviant - standard ERP). We denoted the four left hemisphere clusters as lHG, lPP, lIFG, and lSF, and the four right ones as rHG, rPP, rIFG and rSF (with HG: Heschl's gyrus, PP: planum polare, IFG: inferior frontal gyrus and SF: superior frontal). Trial-by-trial time series were obtained using the distributed source inversion procedure employed in chapter 6, resting on BEM forward models for both modalities (computed with openmeeg) and the Bayesian framework with Multiple Sparse Priors (MSP) available in SPM8. Sensor-level single-trial data were epoched from -200 ms to 410 ms post-stimulus, imported in SPM8, down-sampled (200 Hz) and low-pass filtered (20 Hz low-pass digital filter, bidirectional Butterworth, 5th order). Source inversion involved the epochs for each of the 674 events composing

an oddball sequence (trial-rejection obtained with sensor-level analysis will be applied within the meta-Bayesian scheme, see below), leading to a total of 2696 trial data inversions carried out per subject (two reverse sessions were delivered per condition). Crucially, we could spatially constrain these distributed inversions with the results of our deviance generator analysis (chapter 6) by means of the two-step group-level inference scheme (Figure 6.1). Practically speaking, group priors entering step 2 individual inversions corresponded to the clusters resulting from the deviance source analysis (chapter 6) and the additional bilateral superior frontal sources revealed by the UF vs. PF contrast (chapter 7), represented in Figure 8.1.a. After each inversion, for each cluster, the activities reconstructed at all nodes composing the cluster were averaged to derive a single trace informed by both EEG and MEG data. We controlled the validity of the entire procedure by computing the grand-average standard (preceding a deviant), deviant and difference responses for each cluster, using exactly the same events as for the sensor-level analysis (with regard to artifact rejection), for both condition UF and PF. The resulting traces, shown in Figure 8.1.b.

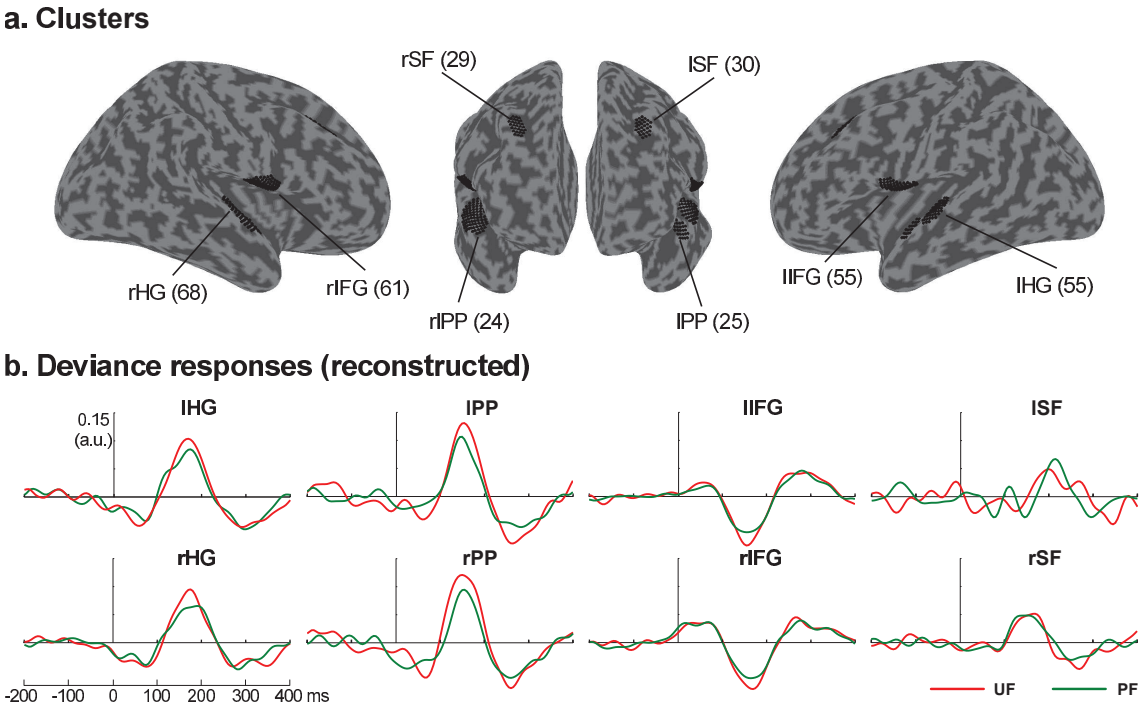

Figure 8.1 – Fitted data. a) Clusters used for trial-by-trial source reconstruction represented on a canonical inflated cortical surface (light grey=gyri, dark grey=sulci), with right view (left), front view (middle) and left view (right). Cluster labelling is referred to in the main text. Black dots indicate nodes within each cluster, with corresponding total number of nodes indicated between parenthesis. b) Group-average difference responses reconstructed at each cluster using fused inversion, for condition UF (red) and PF (green).

8.2.1 Analysis 1: Modeling auditory mismatch responses

Model space. We constructed a model space partitioned into five families of response models which were inspired by models used in Ostwald et al. (2012) and Lieder et al. (2013) (see chapter 4).

The first family (fam_{null}) is made of a single model, the null model M0 as follows:

$$\begin{cases} y = \theta_0 + \varepsilon_n \\ \\ \theta_0 = 0 + \varepsilon_0 \end{cases} \tag{8.1}$$

The second family (fam_{noL}) contains the non-learning models, namely models SC and LIN as in the Ostwald's study:

$$\begin{cases} y = \theta_0 + X_t\theta_t + \varepsilon_n \\ \\ \theta_0 = 0 + \varepsilon_0 \\ \\ \theta_t = 0 + \varepsilon_t \end{cases} \tag{8.2}$$

with X_t the vector of N values defined in chapter 4. Subscript notation t refers to perceptual processes operating at the *tone* level, that relate single items of the sound sequence to each other using non-learning or learning-based mechanisms. It is to be distinguished from the *chunk* level also considered in this study (see below), involving regularities established over a group of items.

The third family (fam_{L_t}) included the Bayesian learning models described in Ostwald's study, reflecting the perceptual learning that the brain would perform at the tone level to estimate the probability μ to hear a standard (under a Bernoulli distribution). Likewise the original study, we considered different values for the forgetting parameter τ_t, precisely 2, 6, 10 and 100, which down-weights the influence of past observations . Models in family fam_{L_t} assume that the brain encodes the Bayesian surprise pertaining to belief updating about μ. We recall here that this measure is informed by both the expectation and the variance (or inverse precision) of the prior and posterior distributions. The observation model for each model of fam_{L_t} has the form of Eq. (8.2), with X being the trajectory of the Bayesian surprise over the experimental session.

The fourth family ($fam_{L_{tc}}$) is equivalent to fam_{L_t} augmented with a chunk-level learning, that was inspired from the study of Lieder et al. (2013) where they designed a model enabling the learning of the expected size of roving trains. Indeed, as was noticed in chapter 5, §5.4, it could be that the brain has the cognitive capacity to learn and predict implicitly the size of chunks, which is here always equal to 5 on average over an entire sequence, but presents differences between conditions at a more local timescale: the size of the chunk varies very smoothly in the condition PF whereas it is subjected to high variability in the UF condition. We thus considered worthy to include chunk-level learning models that could possibly capture the observed difference between UF and PF.

These perceptual (generative) models are based on the following rationale: let U_c be the size of a chunk (or equivalently, the number of standards in between two deviants) and Z_c the prior knowledge about U_c. As in Lieder et al. (2013), the likelihood of U_c given Z_c was modeled using a Poisson distribution and the prior distribution of Z_c was chosen to be a Gamma distribution of parameters g and d, the number of observed tones and observed chunks, respectively. These two

distributions forms the following generative model (expressed at chunk k):

$$\begin{cases} U_c \sim Poisson(Z_c) \\ \\ p(U_c = N|Z_c) = Z_c^N \dfrac{exp(-Z_c)}{N!} \\ \\ Z_c \sim \Gamma(g_k, d_k) \end{cases} \tag{8.3}$$

with N in the range of 2 to 8 in the current study. Each deviant provides a new observation U_c which allows updating Z_c. This update is easily tractable since Gamma prior distribution is conjugate to the Poisson likelihood. Hence, after having observed chunk k: $Z_c \sim \Gamma(g_{k+N}, d_{k+1})$ with N the number of standards in chunk k. As for the tone-level learning models, we considered a forgetting value τ_c, with values equal to 2, 6, 10 or 25. The resulting observation model writes:

$$\begin{cases} y = \theta_0 + X_t\theta_t + X_c\theta_c + \varepsilon_n \\ \\ \theta_0 = 0 + \varepsilon_0 \\ \\ \theta_t = 0 + \varepsilon_t \\ \\ \theta_c = 0 + \varepsilon_c \end{cases} \tag{8.4}$$

where X_c indicates the trajectory of the chunk-level Bayesian surprise computed as the KL divergence between the prior and the posterior Gamma distributions. Because of the four possible values for τ_t and τ_c, family $fam_{L_{tc}}$ is made of 16 response models.

The fifth family that we considered, $fam_{L_{tci}}$, was equivalent to $fam_{L_{tc}}$ but with observation model comprising an additional interaction term: $y = \theta_0 + X_t\theta_t + X_c\theta_c + X_{tc}\theta_{tc} + \varepsilon_n$ with vector X_{tc} being equal to the product of X_t and X_c. This family also comprised 16 models, which led our whole model space to include 39 models.

Model inversion. Model inversions were performed with the VBA toolbox (Daunizeau et al., 2014) presented in §2.4.3. Contrary to Ostwald's study, models were tested against the data by means of a meta-Bayesian analysis (we did not perform a regression analysis on vectors X). These were achieved at each time sample. 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. Hence, given the 39 models and 8 clusters, 12792 meta-Bayesian inversions were carried out per subject. Importantly, we used multi-session based inversions , meaning that individual model evidence and parameter posterior estimates were informed by all the data from each subject. In other words, individual UF and PF data (4 sessions) were fitted all at once. Furthermore, VBA allows rejecting specific trials: this can be done by taking them into account within the evolution model but not within the observation model, to avoid parameter optimization to be affected by noisy data. Exactly the same events considered for our sensor-level analysis thus entered the current study. 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. The prior distribution over θ_0 was defined with zero mean and a variance of 100; other parameters (θ_t, θ_c, θ_{tc}) were defined with prior mean and prior variance equal to 1 and 100. Finally, we set the following initial condition for each model type. For model SC and LIN, model inversion started with previous observation defined as a standard, and the number of observed standards was set to 0 for model LIN. For every tone-level learning model, model inversion started with the number of observed standards and deviants both equal to 1. For chunk-level learning models, model inversion started with the number of observed tones equal to 5, the number of observed chunk equal to 1, and the number of standards in the current chunk equal to 0.

Statistical analysis. For each cluster, we constructed the *relative free energy maps* as in Ostwald's study (Figure 4.7) which allow figuring out which models outperform the static null model M0, and if any, at which latency. For each model m_i, at each time sample, we summed the free energy values obtained across subjects from which we subtracted the summed free-energy values obtained for M0. This is equivalent to the group log-Bayes factor between model m_i and M0. Then we considered that a value greater (lower) than 20 reflect strong evidence in favor of model m_i (the null model). We conducted a family-level with a RFX model comparison for each cluster and each time sample. We expected Bayesian learning models to outperform the null and the non-learning models at the latency of mismatch responses.

8.2.2 Analysis 2: Assessing the effect of predictability

As family level inference conducted in *Analysis 1* concluded in favor of family fam_{L_t}, we aimed at testing the hypothesis that the learning behind the predictability effect could induce some changes in the value of τ_t (which can be interpreted as the size of temporal integration window). Indeed, practically speaking, it takes a minimum of 3 chunks to capture the incrementing structure of PF sequence. Hence it could be that PF data would be better fitted with large values of τ_t whereas UF data would require smaller values. In this subsequent analysis restricted to the winning model family in *Analysis 1*, we thus fitted UF and PF sessions separately in order to reveal a possible difference in this crucial integration parameter.

Simulations. We first simulated pseudo-MMN amplitudes obtained for both conditions with several values of τ_t to assess whether the exact same parameter value could still lead to the reduced amplitude in PF observed with EEG by the mere virtue of the sequences themselves. This was done with τ_t equal to 6, 8, 10, 12, 14, 16, 18, 20, 25, 30, 40, 50, 75 and 100. Precisely, for each τ_t and for each subject (*i.e.* using the sound sequence delivered to each subject), we used the VBA routine to simulate the Bayesian surprise trajectory elicited by the different sound sequences delivered to the subject. We then selected the Bayesian surprise values obtained for each standard preceding a deviant and for each deviant, and using exactly the same procedure that had been used to compute the event-related difference response at the sensor level, we could calculate the simulated group-average pseudo-MMN amplitude in conditions UF and PF.

Learning model inversion. Model inversions were conducted separately on UF and PF trial-by-trial source data at samples exhibiting statistical significance in *Analysis 1*. SF clusters were not considered in this analysis as results in *Analysis 1* failed to reveal greater evidence for the learning model (at any latencies). τ_t was expressed as $\tau_t = exp(\theta)$ and θ was specified as an

evolution parameter to be estimated (defined with a prior mean of 10 and prior variance of 10). Likewise *Analysis 1*, model inversions were performed with the VBA toolbox; 96 inversions were conducted per subject (1 model, 2 conditions, 8 samples, 6 clusters). Multi-session inversion was employed to fit the two reverse sessions collected per condition at once. The same other model specifications described for *Analysis 1* were used for *Analysis 2*.

Statistical analysis. We assumed that the value of τ_t could not change within the time interval used for model inversion (spanning the MMN), we therefore averaged τ_t estimates across samples for each cluster and each condition. Predictability effect could thus be analyzed by conducting a repeated-measures ANOVA on these posterior estimates with factors *Condition* (UF, PF), *Hemisphere* (Left, Right) and *Sources* (HG, PP, IFG).

8.3 Results

We first present the results of *Analysis 1* addressing the cognitive processes elicited by oddball sequence, and performed with UF and PF simultaneous data inversion. We then report findings for *Analysis 2* conducted with Bayesian tone-level learning models inverted on UF and PF data separately to investigate the predictability effect on perceptual learning.

8.3.1 Implicit perceptual treatment of the oddball sequence

For each time sample (from -50 ms to 350 ms, 41 samples), each cluster, individual source activity for UF and PF sessions was modeled using the 39 different models presented in previous section (model space for *Analysis 1*). Relative free energy maps are provided in Figure 8.2. For each cluster, the null model (M0) was found with greater evidence over pre-stimulus time. Two time intervals indicated non-null models outperforming M0: at the latency of the MMN and at the latency of the P3a. Regarding models of family fam_{noL}, model SC exhibited smaller free energy values at all sources, all time samples (but three around 150 ms in rIFG). Model LIN was found better than M0 for all sources but SF ones, at the latency of the MMN (from around 140 ms to 190 ms). Regarding learning models, every model involving a tone-level learning with $\tau_t = 2$ failed to provide larger model log-evidence compared to M0 at any source location, any time samples. Models of family fam_{L_t} had greater evidence than M0 at the latency of the MMN (from 120 ms to 220 ms) for all sources but left SF. Model with $\tau_t = 10$ also exhibited a peak at the latency of the P3a in rHG, bilateral PP and bilateral IFG. Models of family $fam_{L_{tc}}$ outperformed M0 at the latency of the MMN at all sources but lSF and rSF; this effect was mostly observable with models with $\tau_t = 10$. Finally, regarding models of family $fam_{L_{tci}}$, only rHG, rIFG and lIFG could reveal a peak at the latency of the MMN, for models with $\tau_t = 10$. Model families were subsequently compared to each other using family level inference with a RFX model (Figure 8.3). Posterior exceedance probability of family fam_{L_t} was found greater than those for other families at the latency of the MMN for all sources but superior frontal ones. Precisely, the interval showing this effect was equal to 150 ms to 200 ms for lHG (6 samples), 130 ms to 180 ms for lPP (6 samples), 130 ms to 200 ms for lIFG (8 samples), 150 ms to 200 ms for rHG (6 samples), 150 ms for rPP (1 sample) and 140 ms to 190 ms for rIFG (6 samples). Every other time samples (across sources) were associated to M0 having the larger posterior exceedance probabilities. No

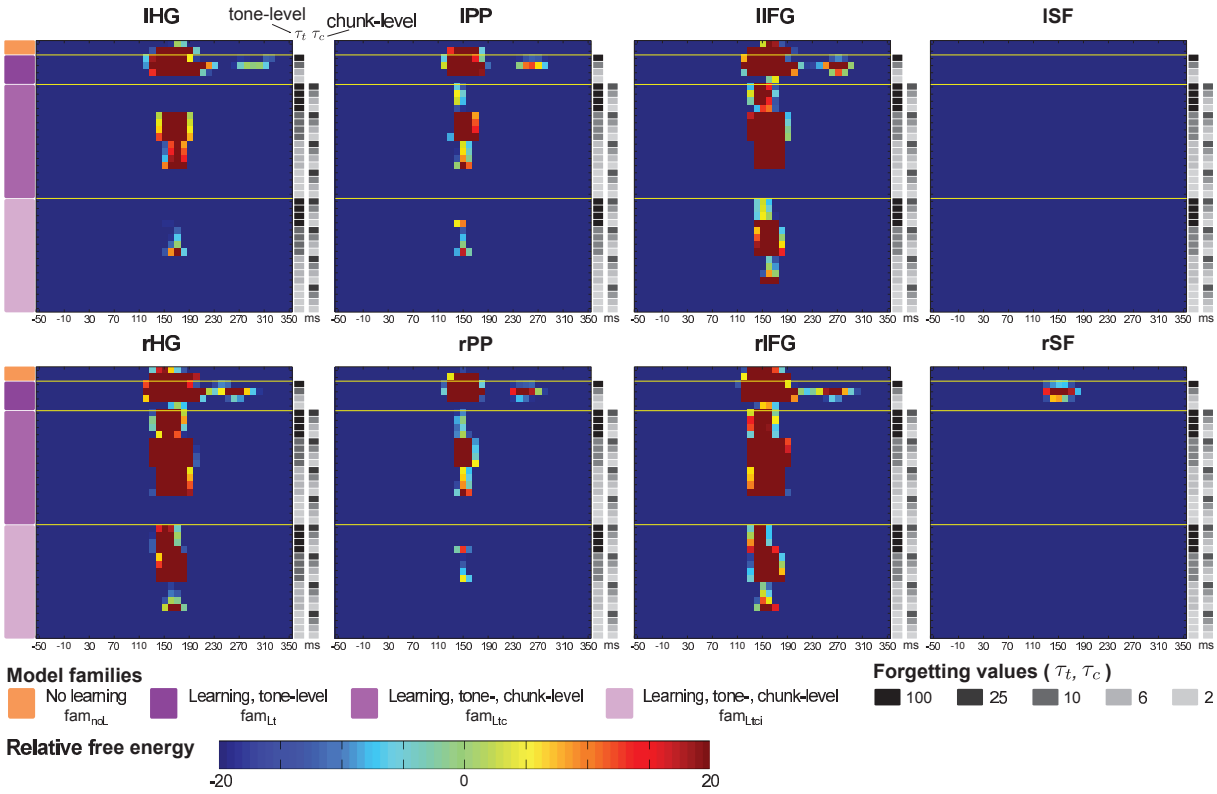

Figure 8.2 – Relative free energy maps. Group free-energy values obtained for sources in the left (upper row), and right hemisphere (lower row). For each map, the 39 models (rows) are sorted by family with color code indicated on the left. First and second rows (fam_{noL}) correspond to model SC and model LIN, respectively. For learning models, grey squares on the right side indicate the value(s) of τ_t (right col.) and τ_c (left col.). The value of the relative free energy obtained for each model m_i (row) and each time sample (column) is represented using the $[-20 + 20]$ color scale so that a dark red (blue) pixel indicates strong evidence in favor of model m_i (the null model).

spatial effect revealing a cognitive specialization could be measured during this analysis. We thus retained fam_{L_t} models for subsequent analysis.

8.3.2 Predictability effect on the forgetting value τ_t

Bayesian surprise simulations. The simulated group-average MMN-like amplitudes obtained from the Bayesian surprises generated with UF and PF sequences and the different values of τ_t value are shown in Figure 8.4. Interestingly, one can see that *i)* the larger τ_t, the smaller the amplitude, and that *ii)* UF and PF amplitudes are very similar over the whole range of τ_t values that we considered (relative differences all below 6%). In other words, a difference in τ_t is mandatory in order to explain the observed difference in ERPs between UF and PF. This observation thus justifies and motivates the following analysis.

Predictability effect on τ_t posterior estimates. As expected, larger τ_t values were estimated with condition PF compared to condition UF ($F_{(1,19)} = 10.89$; $p = 0.001$). On average, τ_t was equal to 16.4 and to 21.2 with UF and PF, respectively (Figure 8.5.a). These values yield different forgetting kinetics between conditions, as illustrated in Figure 8.5.b at the 200$^{\text{th}}$ trial. Likewise Ostwald's study, we could compute the weighted stimulus counts a_{w_n} and b_{w_n} (see §4.3.1, with n indicating the current trial number) resulting from PF and UF group-averaged τ_t estimates,

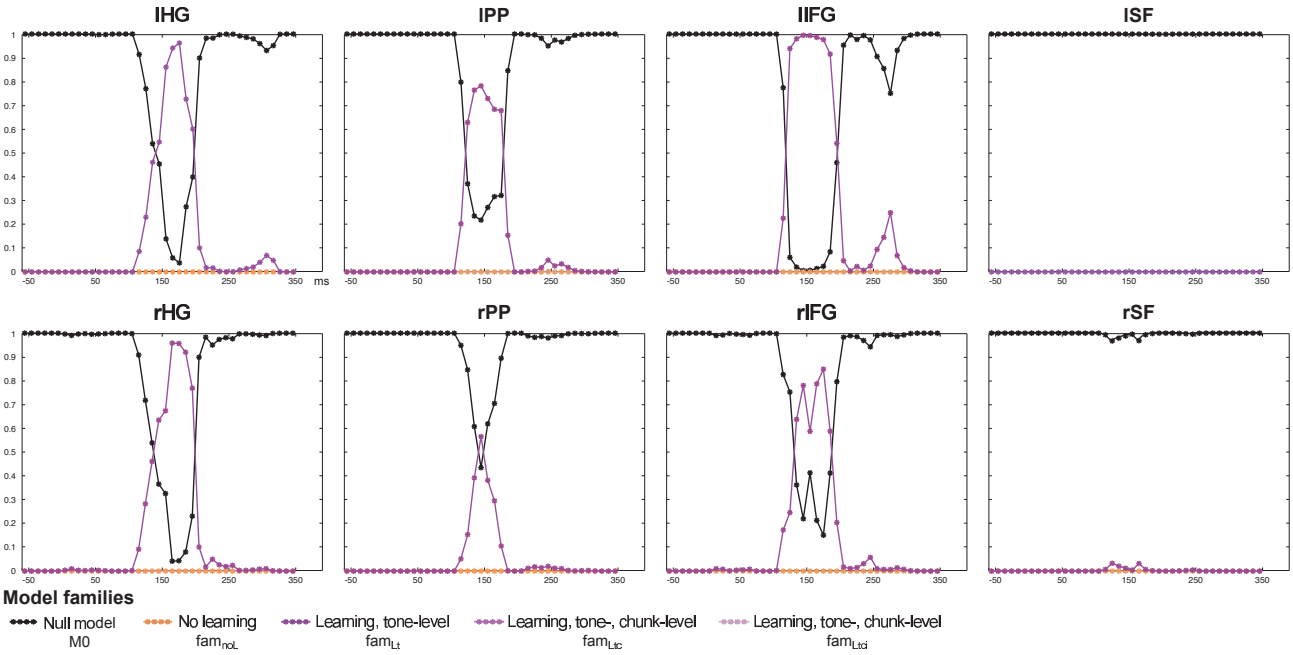

Figure 8.3 – Family level inference (RFX) for *Analysis 1*. Posterior exceedance probabilities of each family estimated at each time sample for each source (upper row: left hemisphere; lower row: right hemisphere).

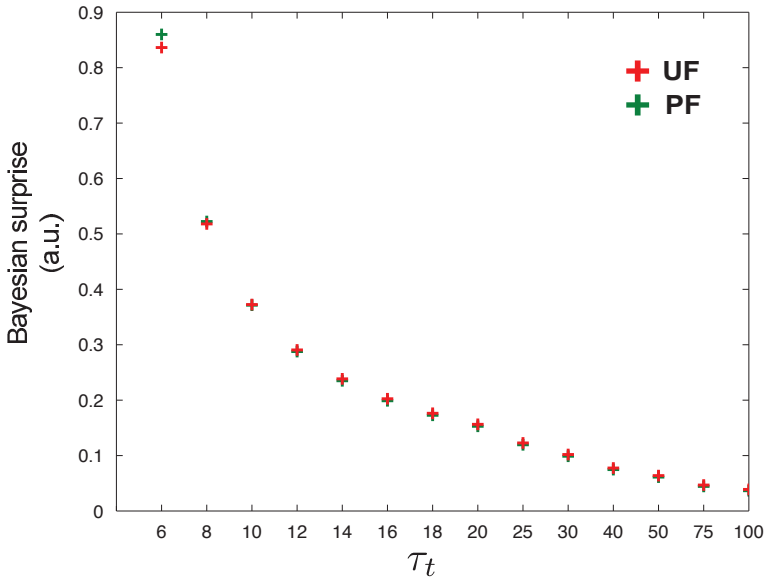

Figure 8.4 – Simulations of Bayesian surprise (group-average) for different integration values (τ_t) for condition UF (red) and PF (green); amplitudes are expressed in arbitrary units (a.u.).

which approximately correspond to a size of temporal integration window equal of around 17 sounds and 22 sounds, respectively (see the example given in Figure 8.5.c). Considering the fact that sequences were built with a fixed SOA of 610 ms, this translates into around 10 s and 13 s, respectively. Besides, the examination of τ_t posterior estimates for each sources could suggest that in condition PF no differential spatial effect could be exhibited whereas in condition UF, HG clusters showed a tendency for smaller values that PP and IFG (Figure 8.6); however this result did not reach significance.

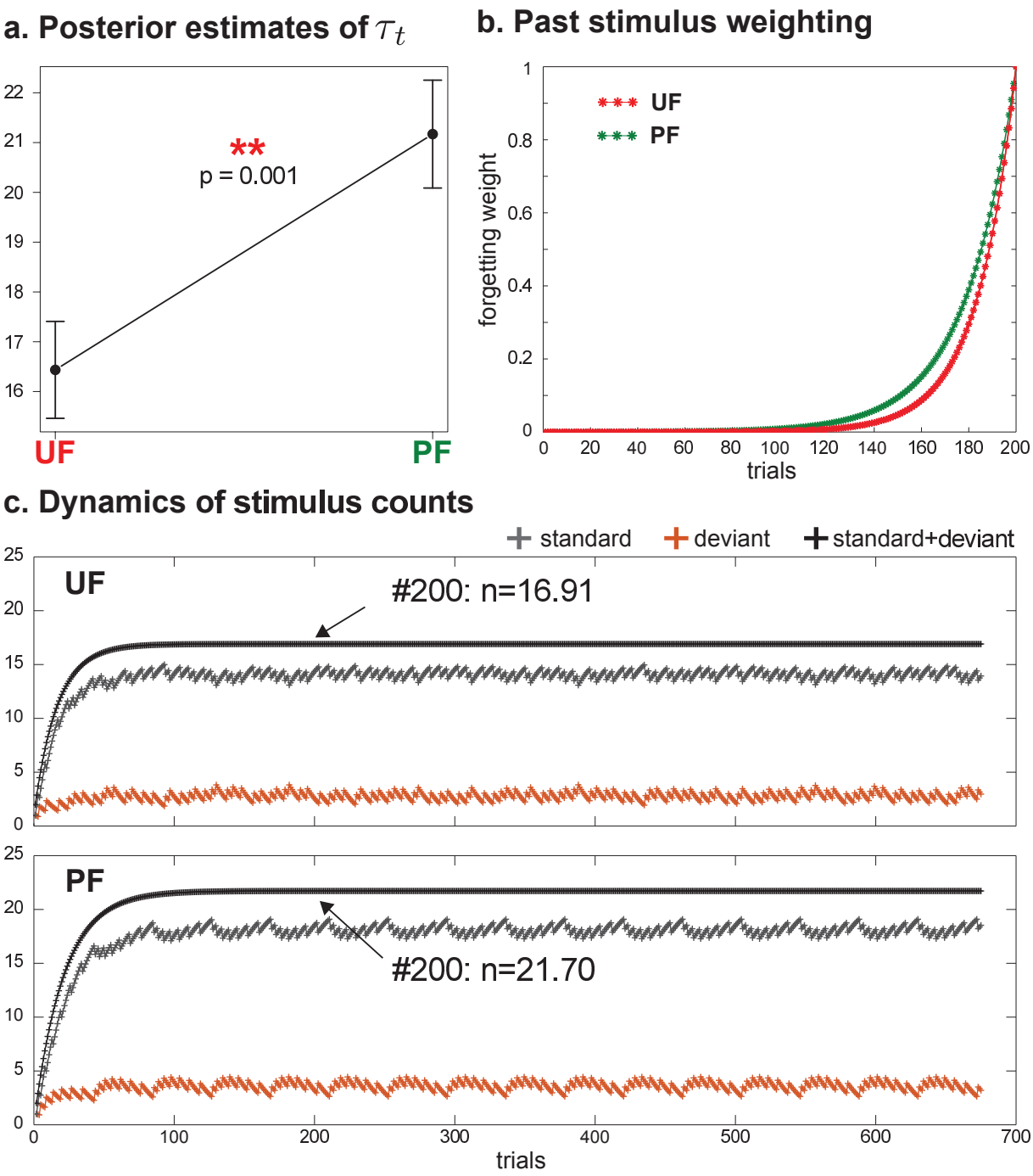

Figure 8.5 – Predictability effect on temporal integration window (*Analysis 2*). a) Group average of τ_t posterior estimates (across sources and hemispheres) for condition UF and PF. b) Example of the forgetting weighting applied at the 200$^{\text{th}}$ trial, using the average τ_t value for condition UF (red) and PF (green). c) Examples of the dynamics of stimulus counts over oddball sessions. Upper and lower graphs represent these dynamics for deviant (a_{w_n}, orange) and standard (b_{w_n}, grey) counts, as well as the total count (black) reflecting the size of the temporal integration window, for one particular UF sequence and PF sequence, respectively. For each graph, the arrow indicates the total number of stimuli at the 200$^{\text{th}}$ trial computed for these particular sequences.

8.4 Conclusion

This study investigated the cognitive processes engaged during the passive listening of a frequency oddball sequence, particularly with regard to changes related to the manipulation of the contextual

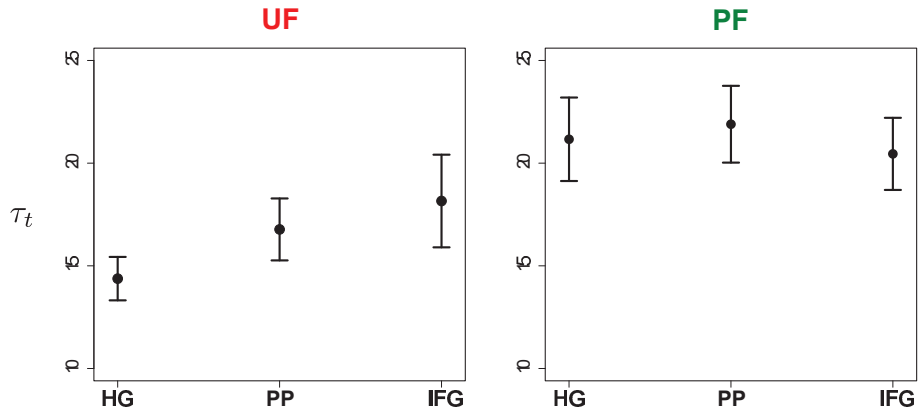

Figure 8.6 – Posterior estimates of τ_t per sources (group average across hemispheres) for condition UF (left) and PF (right).

predictability. First analysis performed with unpredictable and predictable sequences altogether could reveal Bayesian learning at the specific latency of the MMN at all sources but the superior frontal ones. No other time interval - and in particular at the latency of the early deviance effect and the P3a - provided strong evidence in favor of any non-null (*e.g.* learning) model. The subsequent predictability analysis provided a strong insight into the temporal integration windows underlying perceptual learning. Specifically, larger integration values were found to better fit the trial-by-trial updates at the latency of the MMN in the case of predictable deviance.

It is worth noticing that throughout this study, the inversion scheme could benefit from the combination of information captured by both EEG and MEG modalities as it relied on the source time series obtained with fused source reconstruction. Despite the fact that MEG sensor-level analysis failed to reveal a reliable predictability effect, source-level difference responses (informed by both modalities) did exhibit this effect. This suggests that this effect is present in MEG data but was hardly visible at the sensor level (in fact, a small effect on deviant responses did reach significance) and continues to highlight the usefulness of EEG-MEG fusion to increase the signal-to-noise ratio (SNR). Another remarkable result of this study pertains to the trial-by-trial approach which was necessary to investigate the dynamics of belief updating. The feasibility of such not-so-common analysis has already been demonstrated by several neuroimaging studies (Iglesias et al., 2013; Schwartenbeck, Fitzgerald, Mathys, Dolan, & Friston, 2015; Tomiello et al., 2016) and also, to a lesser extent, electrophysiological ones as in Ostwald's study (2012) or more recently (Weber et al., 2015, 2016). This methodological aspect of our study thus contributes to establishing the potential of exploiting the trial-by-trial update information, which is undoubtedly reduced using averaging procedures performed in more traditional event-related studies. The flexibility offered by the VBA toolbox (notably the rejection of artifactual trials and multi-session inversions) combined with the efficiency of its VB inversion scheme made this analysis possible.

Considering evoked responses reflecting hierarchical prediction errors, one could have expected significant effect of non-null models at the latency of the early deviance effect and the P3a, but also within the highest level of the hierarchy indicated by the DCM analysis. Regarding the former point, we believe that the absence of such evidence is to be related to the small amplitude of these components. In Ostwald's study, using a counting task, a spatio-temporal pattern involving Bayesian learning could be revealed, with in particular a significant effect at the medial cingu-

late source from around 300 ms to 400 ms. Our study addressed the predictive coding account of passive listening, and the small deviance magnitude that we employed (to avoid attentional effects) may arguably have led to a small P3a. Of course, the possibility of an inadequate model space to describe early and P3a effects should also be considered. Regarding the second point, it is important to recall that SF clusters did not emerge robustly with the statistical source analysis (see chapter 7), and reconstructed SF difference responses shown in Figure 8.1 suggest a low SNR. Yet, these sources were found necessary to fit deviance responses (over the spatial and the temporal dimension) with DCM, providing support for the existence of a fourth level in the current auditory hierarchy but whose locations might be very approximate here due to its low SNR activity in this task. This point could be improved in further analysis by reconstructing the sources of the grand-average of all sounds composing the oddball sequence: indeed, in the current study, we aimed at modeling the learning reflected in every time series and we no longer restricted the analysis to the processes behind deviant and standard (preceding a deviant) ones.

Noticeably, chunk-level models failed compared to tone-level models in *Analysis 1*. Several explanations can be put forward: first, it could be that such learning occurred but related brain activity was not measurable with our approach (such learning could for instance recrute higher-level frontal sources, poorly estimated in our case). Second, modeling the expected size of chunks by a Poisson distribution may also be inadequate or should have entered a (more realistic) hierarchical generative model of sounds, with higher levels imposing constraints on lower ones. This aspect will be discussed in the last paragraph. Another possibility pertains to the larger number of parameters to be inferred with chunk-level models compared to tone-level ones, that have increased the complexity of these models, not counterbalanced by a larger accuracy. This aspect, resulting from the BMS that we performed, also calls for the speculative interpretation of the brain being able to conduct such model comparison (an issue already suggested in Summerfield et al., 2011). Indeed, one could envisage that during passive listening, the brain would select in an optimal fashion the cognitive process that best enable to adapt to a varying environment. During the current oddball sequence exposure, a simple tone-level learning could represent a sufficient mechanism ensuring this goal, for both unpredictable and predictable sequences.

In *Analysis 2*, we could further refine the characterization of the perceptual learning indicated in *Analysis 1*, by estimating how contextual predictability influences the temporal integration window on which this learning relies. First, posterior estimates of τ_t correspond to a size of the temporal-integration window of about 10 ms and 13 ms for UF and PF, respectively. This is definitely in line with the sensory-memory duration usually reported in typical MMN studies (Näätänen et al., 2007). In addition, the current results confirmed our predictions of the widening of this window with predictability. It appears necessary to clarify the implications of a larger τ_t value for the (computational) precision-weighted prediction errors, a key feature of Bayesian information processing. Likewise Ostwald's study, we assumed the prediction errors to be the Bayesian surprise reflecting the updates of deviant and standard weighted counts, a_{w_n} and b_{w_n}, respectively. In fact, as τ_t increases, prediction errors decrease and their weighting increase. Precisely, the larger τ_t, the larger the number of past stimuli entering the weighted counts, and the smaller the contribution of the current observation (a standard or a deviant) to the count updates. This reduction amounts to a reduction of prediction error. In addition, the effect of τ_t on the precision weighting of prediction errors can be intuitively predicted by considering the fact that

the deviant:standard ratio is estimated with a larger confidence or precision when information is computed from a larger number of events (for instance, 10:90 versus 1:9). Such modulations can be formally demonstrated using the Bayesian surprise expression at trial n:

$$\begin{aligned} BS_n = & \log(\frac{\Gamma(a_{n-1}+b_{n-1})}{\Gamma(a_n+b_n)}) \\ & + \log(\frac{\Gamma(a_n)}{\Gamma(a_{n-1})}) \\ & + \log(\frac{\Gamma(b_n)}{\Gamma(b_{n-1})}) \\ & + (a_{n-1}-a_n) \times [\psi(a_{n-1}) - \psi(a_{n-1}+b_{n-1})] \\ & + (b_{n-1}-b_n) \times [\psi(b_{n-1}) - \psi(a_{n-1}+b_{n-1})] \end{aligned}$$

with Γ and ψ indicating the Gamma and Euler function, respectively. Subscript w in deviant and standard weighted count notations has been omitted for sake of clarity. Assuming that the first three terms can be neglected (this was confirmed with simulations), we focused on the fourth and fifth terms that can be interpreted as precision-weighted prediction errors regarding deviant and standard counts, respectively. In the case of deviants, $(a_{n-1}-a_n)$ reflects the prediction error, and $[\psi(a_{n-1}) - \psi(a_{n-1}+b_{n-1})]$ its weighting (same definition for standards). Applying the forgetting weight τ to past events gives the count updates at trial n:

$$\begin{cases} a_n = u_n + \exp(\frac{-1}{\tau})a_{n-1} \\ \\ b_n = (1-u_n) + \exp(\frac{-1}{\tau})b_{n-1} \end{cases}$$

with u_n being equal to 0 in the case of a standard, and 1 for a deviant. One can see that τ enters both the prediction error and the weighting terms. In the following, we will focus on deviant updates (a similar demonstration for standards can easily be derived). Regarding the prediction error term, it can be expressed as:

$$a_{n-1} - a_n = (\exp(\frac{-1}{\tau}) - 1)a_{n-1} - u_n$$

Let τ_1 and τ_2 be such that $\tau_2 > \tau_1$, we have in absolute terms $(\exp(\frac{-1}{\tau_2}) - 1) < (\exp(\frac{-1}{\tau_1}) - 1)$ hence smaller updates $(a_{n-1}-a_n)$ with τ_2, hence smaller prediction errors. Regarding the weighting term, since ψ is monotonically increasing for positive real numbers (like weighted counts), the behavior of $[\psi(a_{n-1}) - \psi(a_{n-1}+b_{n-1})]$ with τ depends on the corresponding behavior of b_{n-1}. As $\tau_2 > \tau_1$ gives $\exp(\frac{-1}{\tau_2}) > \exp(\frac{-1}{\tau_1})$, one can see that $b_n(\tau_2) > b_n(\tau_1)$, leading to a larger weighting term with τ_2. Having examined how these terms vary with τ, we can consequently conclude that the predictable structure of PF sequences yielded smaller predictions errors and larger precision weights compared to unpredictable oddball sequences.

Interestingly, consistent effects were observed with DCM as reported in chapter 7. Precisely, DCM findings were that higher predictability was associated with a reduction of the forward extrinsic connections (visible with EEG), that could relate to reduced prediction errors, and a reduction of self-inhibition in the supra-pyramidal population associated to the error units, that

could reflect an increase of precision weighting. Taken together, it is interesting to point out converging results measured with the same data sets but obtained from completely different schemes (DCM at the neurophysiological level, fitting evoked responses; functional modeling at the computational level, fitting trial-by-trial responses).
Several findings from different research fields support the hypothesis of a temporal hierarchy within which higher levels would pertain to larger temporal integration windows. (Escera & Malmierca, 2014; Kiebel et al., 2008). However, our result could not subsume such nuances along our hierarchy. In other words, we did not observe a gradient of τ values from auditory to frontal regions. Only a non-significant tendency could be observed in the UF condition.

Our results reveal a perceptual learning mechanism at play during auditory processing, reflected by single-trial variations of responses at the MMN latency. They also suggest the ability of the brain to adapt the temporal integration window to the statistical structure of the environment. Together, these findings speak to the predictive coding account of auditory evoked responses, enabling an efficient processing of information under environmental predictability. These computational findings are nicely compatible with the ones we obtain with DCM, both bringing empirical support to the top-down weighting of prediction errors induced by high-level processes.

Part III

General Discussion

Chapter 9

Discussion and perspectives

9.1 Summary of the main results

The aim of this PhD work was to refine the predictive coding account of (passive) auditory processing through measuring and interpreting EEG and MEG mismatch responses. Predictive coding, and more generally the Bayesian brain theory, assume that the brain entertains a generative model of the environment and adapts to its changes using Bayesian computation. This includes the learning of environmental regularities, which makes the oddball paradigm, involving sequence of expected and unexpected sounds, particularly well-suited to test formal hypothesis about the underlying computational and neurobiological processes. In particular, we aimed at examining the influence of contextual predictability on deviance responses to characterize its effects in terms of message passing and computation of precision-weighted prediction errors. Using simultaneous EEG-MEG recordings and advanced neurophysiological and computational Bayesian modeling, we obtained the following findings:

- Three sensor-level deviance responses (the expected MMN, but also an earlier effect and the P3a) exhibited reduced amplitude with increasing predictability (significant on EEG data), corroborating the computation of a cascade of prediction errors along the auditory hierarchy.

- A fronto-temporal network for deviance processing could be reconstructed at the cortical level using fused source reconstruction (augmented with group-level inference), revealing fine-grained patterns within the supratemporal plane conditioned to the type of deviance (frequency or intensity) but also the latency of deviance responses (early mismatch effect and the MMN).

- Within this cortical network (augmented with an additional frontal contribution), implicit perceptual learning proved at play at the latency of the MMN using the computational modeling of trial-by-trial responses. Using Dynamic Causal Modeling based on a Canonical Micro-Circuit representation of cortical sources, we could show that this learning process was grounded in a four-level cortical hierarchy underlying the generation of mismatch evoked responses.

- Predictability was found to influence the size of the temporal integration window, on the one hand, and the strength of forward extrinsic and self-inhibitory intrinsic connections,

on the other hand. From both perspectives, a higher predictability was found to dampen prediction errors while boosting their precision weights.

It is worth noticing that those results were obtained using EEG and MEG data in combination. Importantly, the two modalities did not contradict each other but showed different sensitivity hence proved highly complementary. In other words, those precise and important findings could not have been derived if we had followed a unimodal approach, be it with EEG or MEG alone.

9.2 Implications for future research

At the core of the predictive coding principle is the hierarchical message passing of precision-weighted prediction errors, whose suppression drives the learning of environmental regularities. In the current work, we manipulated the regularity of auditory oddball sequences so as to test the biological and cognitive plausibility of this inference framework. Our findings did shed light on such processes but also highlighted the usefulness of probabilistic modeling. In what follows, we discuss those different aspects and their possible implications for future work.

9.2.1 The auditory hierarchy serving Bayesian inference

Our findings, in the context of auditory evoked responses, neatly support the key architectural principle of predictive coding, namely a hierarchical organization, from both a neurophysiological and cognitive perspective. However, assessing Bayesian learning with a computational modeling approach showed fragile evidence in favor of a hierarchical cognitive sophistication (the widening of the temporal integration window with ascending levels did not reach significance). We detail here further investigations that could refine the description of the suggested predictive hierarchy.

Characterizing auditory components as hierarchical errors. Our results at the sensor-level revealed deviance responses at three different latencies with both EEG and MEG, and their modulation by deviance predictability was interpreted as a signature of prediction error. From a dynamical system perspective, the hierarchical architecture provides constraints, that shape the dynamics of electrophysiological responses. According to Friston, the different components of evoked responses represent hierarchical prediction errors that are suppressed by means of their respective top-down adjustments (Friston, 2005). This view provides a mechanistic interpretation of the temporal delays between components, and relates late components to higher-level prediction errors, hence with higher-level learning processes. Seeking for empirical evidences in favor of such a spatio-temporal hierarchy of prediction errors is thus an important challenge. Recent studies reporting early mismatch responses elicited by simple repetition rules but not by more structured ones (Cornella et al., 2012; Recasens, Grimm, Wollbrink, et al., 2014) are remarkably in line with this view. Given our data, it would be worth investigating the early effect and the P3a using our twofold modeling approach. Notably, using DCM and an approach inspired from the work of Garrido et al. (2007), we could replicate our analysis with shorter and longer time intervals, to assess the ensuing modulation of forward and backward connections. It should be recalled that learning models failed to reveal a significant effect at other latencies than the MMN, possibly due to the lack of high SNR. This speaks in favor of intracranial trial-by-trial analysis of intracranial data acquired during the same protocol in implanted epileptic patients. Considering larger deviance magnitude (inducing larger component amplitude due to larger adaptation effects) may also be

a valuable option to increase the SNR. Indeed, an interesting idea brought by predictive coding pertains to adaptation effects, long considered as masking the *genuine* MMN, but which could be re-considered as a useful phenomenon to facilitate the investigation of early to late deviance responses. Precisely, under this perspective, difference responses (*deviant − standard*) are no longer the most relevant contrast to investigate perceptual learning compared to the trial-by-trial modulations. Augmenting computational models with an account of deviance magnitude (see for instance Lieder, Daunizeau, et al., 2013; May et al., 2015)) could help quantifying the contribution of adaptation effects in shaping prediction error signals. In that respect, SSA as modeled by May and collaborators is very inspiring. SSA is shown to possibly emerge from synaptic depression which induces time-varying synaptic connectivity. This can explain dynamical frequency-based associations (hence temporal integration) and events disrupting such associations could generate MMN-like responses. Hence, large deviance magnitude could trigger learning at higher levels in the hierarchy, just like large stimulus saliency can induce late (high-level) attention orienting components.

Insights from oscillations. An important aspect that we did not discuss yet is the putative functional role of oscillations in a predictive coding scheme. Indeed, Several studies reported different oscillatory signatures exhibited by the feedforward and feedbackward pathways, as reviewed in (Bastos et al., 2012; Markov et al., 2014; Bastos et al., 2015). Importantly, those findings substantiate the asymmetry between those two pathways and hence further emphasize the hierarchical organization of the cortex. In particular, supra-granular layers (mostly involved in ascending signals) were related to theta- and gamma-band oscillations whereas infra-granular layers (mostly involved in descending signals) presented neurons showing beta-band oscillations. As explained in (Friston, Bastos, et al., 2015), the generation of synchronous oscillatory activities could modulate the synaptic gain, hence the weighting of prediction errors. Assessing the effect of predictability on (non-phase locked) oscillatory activity in various frequency bands could possibly further substantiate the hierarchical organization and the weighting of information by their relative precision. As a matter of fact, visual inspection of EEG traces during pre-stimulus interval could suggest differences in the alpha and beta bands between UF and PF (see Figure 4 in Lecaignard et al., 2015).

9.2.2 Precision-weighted prediction errors

Our results revealed an increase of the precision-weighting of prediction errors with contextual expectancy, that would facilitate the processing of auditory information. In our point of view, this is an important step towards establishing the ability of the brain to manipulate information by taking into account their estimated reliability.

Neural correlates of error weighting. The dominant hypothesis for a possible neural implementation of precision-weighting prediction error rests on synaptic signaling, and neuromodulation (possibly involving dopamine, acetylcholine) (Friston, 2010).The latter controls the synaptic gain of prediction error units, hence their precision weighting. DCM with CMC explicitly accounts for such signaling, which is modeled by the self-inhibitory gain of supra-pyramidal cells (presumably reflecting the error units) . Pharmacological manipulations were proved relevant to explore the underlying mechanisms behind precision-weighting of prediction errors at a neurobiological

level (Moran, Campo, et al., 2013) and more recently, at a computational level (Tomiello et al., 2016). In our work, the use of both levels of modeling demonstrated the possibility to precisely and quantitatively investigate this issue. And our findings speak in favor of the modulation of synaptic signaling, resulting from the perceptual learning of environmental regularities.

Deepening our understanding of precision-weights. Our results corroborates that less weight is attributed to less reliable information. Computationally, this weight is a trade-off between the likelihood precision and the prior precision, and formally represents a means for the brain to appropriately balance the two types of information, namely bottom-up sensory inputs and top-down predictions. Each precision estimate evolves with the update dynamics behind perceptual learning. Elucidating the processes and neural mechanisms behind the computation, the adjustment and the combination of those precisions is becoming a central question which not only call for appropriate and dedicated models but also for new and finely tuned experimental manipulations. For instance, regarding the likelihood precision, we could consider the reasoning applied in many psychophysical studies (Ernst & Banks, 2002) that aimed at testing whether the brain integrates multiple information in a Bayesian way. Using dichotic listening of oddball sequences with different intensities or noise levels between the two ears, as well as conflicting trials (a deviant at one ear and a standard at the other), we might be able to isolate the optimal adjustment of sensory precision and its contribution to perceptual learning and decision making. Regarding prior precision, one limitation of the current study pertained to the fact that the learning of regularities in PF sequences was very rapid, preventing from examining how it could have gradually influence the weighting of prediction errors. We could thus envisage a more complex pattern of sound association as in (Bendixen et al., 2008; Furl et al., 2011), where the duration of exposure to the sound sequence was shown to modulate deviance responses. However, this calls for more complex perceptual models involving high-order transition probabilities and for which the inversion of the response model may not be straightforwardly tractable (but see Lieder, Daunizeau, et al., 2013).

9.3 Towards model-driven clinical applications of the MMN

As previously emphasized and also further supported by our work, the free energy principle and both its neurobiological and computational implications has already provided great insights into several cognitive neuroscience domains. In this section, we consider how this framework can also promisingly benefit to clinical applications, in particular to neurology and psychiatry, and we discuss the possibility of our study and modeling approach to contribute to that aim.

Classical neuroimaging studies for clinical applications contribute to improve the accuracy of diagnostic procedures at the individual patient level (in terms of predictive outcome value), in order to refine the selection for specific treatments. Critically, this involves the long-standed issue of classifying patients into meaningful groups based on key criteria, that must be achieved with routine clinical procedure using single-subject measurements. In (Stephan et al., 2015b), the authors propose computationally informed biomarkers to enter (and refine) the diagnostic scheme, in order to infer the (possibly impaired) mechanisms that have generated the neuroimaging data. Typically, a relevant maker would be a neural correlate of precision-weighted prediction error

(DCM connectivity parameters for instance). Regarding deviance responses and the MMN in particular, oddball paradigms have been extremely used for various clinical domains for decades (Näätänen, 2003; Sussman et al., 2014). The MMN is indeed a window into cognitive processing as it allows assessing dysfunctioning in regularity learning or in perceptual discrimination for instance. From a technical point of view, it is easy and rapid to measure, without requiring the attention of the patient. Existing methodologies for patient evaluation are bound to the typical measure of the MMN amplitude, duration or latency. They sometimes but rarely involve its underlying generators using source reconstruction. To go further, some DCM studies have already attempted to promote the added value of quantitative neurophysiological generative models to better exploit oddball paradigms in clinical settings, namely in the case of patients with altered states of consciousness (Boly et al., 2011), or in schizophrenia (Dima et al., 2012). In the following, we present two examples (one in neurology, one in psychiatry) of possible clinical applications of our paradigm, in close relation to side projects I have contributed during this PhD (see §9.4.1).

MMN and computational neurology. We consider here patients in altered states of consciousness, for whom a cognitive state evaluation in the absence of explicit communication is challenging. Such patients are typically grouped into the categories of comatose state, vegetative state (VS) or minimally conscious state (MCS) (Giacino et al., 2004). The use of MMN paradigms with comatose patients has proved efficient to predict coma outcome (Fischer et al., 2004). Using DCM, Boly et al. (2011) tested the hypothesis of impaired effective connectivity in VS and MCS patients during an oddball paradigm, with findings suggesting the implication of backward (fronto-temporal) connectivity in VS patients compared to healthy subjects. Our oddball study focused on the implicit processing of environmental regularities and indicated specific computational and biological markers characterizing the ensuing perceptual learning. Based on such mechanistic markers, conducting our paradigm with VS and MCS patients could help examining their cognitive ability to process auditory information compared to healthy subjects. The predictability effect could further inform us about the degree of preserved cognitive function, as it was found to involve high level processing in the current work. This would require some adjustments such as increasing the magnitude of the deviance, to allow measuring the predictability effect at the individual level (namely the detection of a reduced MMN amplitude with predictable sequence), as indicated by our preliminary attempts. In addition, further investigating the usefulness of the P3a appears of significant relevance using our modeling approach, as this later component presumably reflects the transition from automatic to attention-orienting and voluntary processes.

MMN and computational psychiatry. The case of autism is naturally addressed by the emerging field of computational psychiatry (Friston et al., 2014) as Bayesian brain theories succeeded in framing the phenomenology of this mental disorder, which would be characterized by impaired perceptual inferences. Precisely, as proposed in a recent review (Haker et al., 2016), an altered weighting of prediction errors favoring bottom-up sensory input information at the expense of top-down predictions would be involved, compromising the learning processes required for adaptive behavior to emerge, especially in complex environment such as in social interactions. Clarifying these impairments rests on investigating the computation of sensory and prediction precision as well as assessing their physiological underlying correlates and should involve the manipulation of environmental stability to test how patients with autism adapt (and learn) in a volatile context (Behrens et al., 2007; Robic et al., 2014). Impaired sensory processing in autism has long been

explored with the MMN (Näätänen et al., 2014), with findings showing various amplitude and latency modulation patterns depending on the experimental material (speech vs. non-speech for instance). Hence, computational models of the MMN provide new perspectives for investigating autism (in particular regarding how the brain learns in a changing environment) while resting on easy-to-run paradigms that are convenient for patients. This could efficiently be achieved using the twofold modeling strategy employed in this work, first using our paradigm to examine the dynamics of precision-weighted prediction errors, and second using model-driven adjustments of oddball sequences to enable a finer examination of the relative adjustment of precisions (this could possibly rest on the suggestions developed in §9.2.2).

9.4 Concluding remarks

To conclude, this thesis aimed at establishing the validity of the predictive coding account of auditory processing during a passive oddball paradigm, in terms of its cognitive and biological plausibility. This framework provides precise predictions regarding how deviance responses, or equivalently hierarchical precision-weighted prediction errors, should be affected by environmental regularities. We thus manipulated the predictability of frequency oddball sequences, leading to unpredictable and predictable deviants, and deployed a twofold modeling analysis informed by simultaneous EEG and MEG recordings, to address on one hand the functional significance (using learning models of trial-by-trial responses) and on the other hand the neurophysiological underpinnings (using models of effective connectivity) of deviance responses. Our findings, at both levels, support the key principles of predictive coding, namely a hierarchical organization for perceptual inference and learning and the weighting of prediction errors which is crucial to trigger learning processes in a changing environment according to the reliability of information. These results speak neatly in favor of computational and mechanistic modeling, being efficient approaches to better understand perceptual learning.

9.4.1 Related works performed during this PhD

During this PhD, I took part to different projects sharing methodological and cognitive concerns with the questions addressed in the current thesis:

- with Marie Gomot (Inserm U930, Tours) and Jérémie Mattout, using DCM, we attempted to characterize the differences in deviance responses measured with children with autism and neurotypical children in terms of effective connectivity. Precisely, using EEG data in a passive frequency oddball sequence (Gomot et al., 2002), the aim was to test whether thalamic afferences would arrive in the ACC for children with autism, an hypothesis corroborated by a subsequent fMRI study (Gomot et al., 2006).

- with Dominique Morlet (CRNL, Dycog, Lyon) and Jérémie Mattout, we attempted to model deviance responses measured in patients having altered states of consciousness using DCM and EEG recordings in a duration deviance paradigm. This work was motivated by the study of Boly et al. (2011), and aimed at establishing a relationship (at the group-level) between the effective connectivity at play during oddball sequence, patients' score obtained using the commonly used Coma Recovery Scale (CRS) and the statistical emergence of brain components (the N1, the MMN and the P3a) at relevant EEG sensors.

- with Ludovic Bellier (CRNL, Dycog, Lyon) , Anne Caclin and Jérémie Mattout, we aimed at testing the hypothesis of cortical contribution to speech auditory brainstem responses (ABR), long established as originating from brainstem. Using DCM, we considered a model space comprising cortical, subcortico-cortical, and subcortical networks and expected model comparison to provide empirical evidence for subcortico-cortical contributions, as suggested by intracranial recordings collected with a comparable experimental setup. Unfortunately, the inadequacy of DCM evolution parameters to model the rapid component of these responses prevented us to pursue this work.
- with Antoine Lutz (CRNL, Dycog, Lyon) , Anne Caclin and Jérémie Mattout, we discussed the predictive coding account of the modulation of deviance responses by different meditative state (relative to different meditation practice styles). Preliminary sensor-level (EEG) analysis suggest that the level of expertise and the state of mind do interfere with auditory mismatch processing. These early results now call for quantitative hypothesis to be formally tested using Bayesian model comparison.

9.4.2 List of publications related to this PhD work

Articles in peer-reviewed journals

Lecaignard, F., Bertrand, O., Gimenez, G., Mattout, J., Caclin, A. (2015). Implicit learning of predictable sound sequences modulates human brain responses at different levels of the auditory hierarchy. Frontiers in Human Neuroscience, 9:505.

Lecaignard, F., Bertrand, O., Caclin, A.,Mattout, J. Empirical evaluation of fused MEG-EEG source reconstruction applied to auditory mismatch generators. *In preparation.*

Sanchez, G., **Lecaignard, F.**, Otman, A., Maby, E., Mattout, J. (2016). Active SAmpling Protocol (ASAP) to Optimize Individual Neurocognitive Hypothesis Testing: A BCI-Inspired Dynamic Experimental Design. Frontiers in Human Neuroscience, 10:347.

Book chapters

Lecaignard, F., Mattout, J. (2015) Forward models for EEG-MEG in "Brain Mapping: An Encyclopedic Reference", Arthur W. Toga, Editor, vol. 1, pp. 549-555. Academic Press

References

Acar, Z. A., & Makeig, S. (2013, July). Effects of Forward Model Errors on EEG Source Localization. *Brain Topography*, *26*(3), 378–396.

Adams, R. A., Bauer, M., Pinotsis, D., & Friston, K. (2016, February). Dynamic causal modelling of eye movements during pursuit: Confirming precision-encoding in V1 using MEG. *NeuroImage*, *132*, 175–189.

Alain, C., Woods, D. L., & Knight, R. T. (1998, November). A distributed cortical network for auditory sensory memory in humans. *Brain Research*, *812*(1-2), 23–37.

Alain, C., Woods, D. L., & Ogawa, K. H. (1994, December). Brain indices of automatic pattern processing. *Neuroreport*, *6*(1), 140–144.

Alcaini, M., Giard, M. H., Thévenet, M., & Pernier, J. (1994, November). Two separate frontal components in the N1 wave of the human auditory evoked response. *Psychophysiology*, *31*(6), 611–615.

Alho, K. (1995, February). Cerebral generators of mismatch negativity (MMN) and its magnetic counterpart (MMNm) elicited by sound changes. *Ear and hearing*, *16*(1), 38–51.

Alho, K., Rinne, T., Herron, T. J., & Woods, D. L. (2014, January). Stimulus-dependent activations and attention-related modulations in the auditory cortex: a meta-analysis of fMRI studies. *Hearing research*, *307*, 29–41.

Althen, H., Grimm, S., & Escera, C. (2013, November). Simple and complex acoustic regularities are encoded at different levels of the auditory hierarchy. *The European journal of neuroscience*, *38*(10), 3448–3455.

Anderson, J. C., & Martin, K. A. C. (2015, January). Interareal Connections of the Macaque Cortex: How Neocortex Talks to Itself. In K. S. Rockland (Ed.), *Axons and brain architecture* (pp. 117–131). San Diego.

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

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

Babiloni, F., Babiloni, C., Carducci, F., Romani, G. L., Rossini, P. M., Angelone, L. M., & Cincotti, F. (2004, May). Multimodal integration of EEG and MEG data: A simulation study with variable signal-to-noise ratio and number of sensors. *Human Brain Mapping*, *22*(1), 52–62.

Baillet, S., Garnero, L., Marin, G., & Hugonin, J. P. (1999, May). Combined MEG and EEG source imaging by minimization of mutual information. *IEEE transactions on bio-medical engineering*, *46*(5), 522–534.

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. (2012, November). Canonical microcircuits for predictive coding. *Neuron*, *76*(4), 695–711.

Bastos, A. M., Vezoli, J., & Fries, P. (2015, April). Communication through coherence with inter-areal delays. *Current opinion in neurobiology*, *31*, 173–180.

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007, September). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221.

Bekinschtein, T. A., Dehaene, S., Rohaut, B., Tadel, F., Cohen, L., & Naccache, L. (2009, February). Neural signature of the conscious processing of auditory regularities. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(5), 1672–1677.

Bellier, L., Bidet-Caulet, A., Bertrand, O., Thai-Van, H., & Caclin, A. (2014). **Auditory Brainstem Responses in the Human Auditory Cortex?! Evidence from sEEG**. In *Organization for human brain mapping.* Hamburg.

Bendixen, A., Prinz, W., Horváth, J., Trujillo-Barreto, N. J., & Schröger, E. (2008, July). Rapid extraction of auditory feature contingencies. , *41*(3), 1111–1119.

Bendixen, A., Roeber, U., & Schröger, E. (2007, October). Regularity extraction and application in dynamic auditory stimulus sequences. *Journal of Cognitive Neuroscience*, *19*(10), 1664–1677.

Boly, M., Garrido, M. I., Gosseries, O., Bruno, M.-A., Boveroux, P., Schoenberg, C., ... Friston, K. (2011, May). Preserved feedforward but impaired top-down processes in the vegetative state. *Science (New York, NY)*, *332*(6031), 858–862.

Brown, H. R., Adams, R. A., Parees, I., Edwards, M., & Friston, K. (2013, November). Active inference, sensory attenuation and illusions. *Cognitive processing*, *14*(4), 411–427.

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

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

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

Chakalov, I., Paraskevopoulos, E., Wollbrink, A., & Pantev, C. (2014, October). Mismatch negativity to acoustical illusion of beat: how and where the change detection takes place? , *100*, 337–346.

Chennu, S., & Bekinschtein, T. A. (2012). Arousal modulates auditory attention and awareness: insights from sleep, sedation, and disorders of consciousness. *Frontiers in psychology*, *3*, 65.

Chowdhury, R. A., Zerouali, Y., Hedrich, T., Heers, M., Kobayashi, E., Lina, J.-M., & Grova, C. (2015, November). MEG-EEG Information Fusion and Electromagnetic Source Imaging: From Theory to Clinical Application in Epilepsy. *Brain Topography*, *28*(6), 785–812.

Coffey, E. B. J., Herholz, S. C., Chepesiuk, A. M. P., Baillet, S., & Zatorre, R. J. (2016). Cortical contributions to the auditory frequency-following response revealed by MEG. *Nature Communications*, *7*, 11070.

Conway, C. M., & Pisoni, D. B. (2008, December). Neurocognitive Basis of Implicit Learning of Sequential Structure and Its Relation to Language Processing. *Annals of the New York Academy of Sciences*, *1145*(1), 113–131.

Cooray, G., Garrido, M. I., Hyllienmark, L., & Brismar, T. (2014, September). A mechanistic model of mismatch negativity in the ageing brain. *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, *125*(9), 1774–1782.

Cornella, M., Bendixen, A., Grimm, S., Leung, S., Schröger, E., & Escera, C. (2015, November). Spatial auditory regularity encoding and prediction: Human middle-latency and long-latency auditory evoked potentials. *Brain Research*, *1626*, 21–30.

Cornella, M., Leung, S., Grimm, S., & Escera, C. (2012, August). Detection of Simple and Pattern Regularity Violations Occurs at Different Levels of the Auditory Hierarchy. *PLoS ONE*, *7*(8), e43604.

Cowan, N., Winkler, I., Teder, W., & Näätänen, R. (1993, July). Memory prerequisites of mismatch negativity in the auditory event-related potential (ERP). *Journal of experimental psychology. Learning, memory, and cognition*, *19*(4), 909–921.

Crottaz-Herbette, S., & Menon, V. (2006, May). Where and when the anterior cingulate cortex modulates attentional response: combined fMRI and ERP evidence. *Journal of Cognitive Neuroscience*, *18*(5), 766–780.

Crouzeix-Cheylus, A. (2001). *Méthodes de localisation des générateurs de l'activité électrique cérébrale à partir de signaux électro-et magnéto-encéphalographiques* (Unpublished doctoral dissertation).

Dale, A. M., & Sereno, M. I. (1993, September). Improved Localization of Cortical Activity bu Combining EEG and MEG with MRI Cortical Surface Reconstruction : A Linear Approach. *Journal of Cognitive Neuroscience*, *5*(2), 162–176.

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

Daunizeau, J., den Ouden, H. E. M., Pessiglione, M., Kiebel, S., Friston, K., & Stephan, K. E. (2010). Observing the observer (II): deciding when to decide. *PLoS ONE*, *5*(12), e15555.

Daunizeau, J., den Ouden, H. E. M., Pessiglione, M., Kiebel, S., Stephan, K. E., & Friston, K. (2010). Observing the observer (I): meta-bayesian models of learning and decision-making. *PLoS ONE*, *5*(12), e15554.

David, O., Harrison, L., & Friston, K. (2005, April). Modelling event-related responses in the brain. *NeuroImage*, *25*(3), 756–770.

David, O., Kiebel, S., Harrison, L. M., Mattout, J., Kilner, J. M., & Friston, K. (2006,

May). Dynamic causal modeling of evoked responses in EEG and MEG. *NeuroImage*, *30*(4), 1255–1272.

Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995, September). The Helmholtz machine. *Neural computation*, *7*(5), 889–904.

Demarquay, G. v., Caclin, A., Brudon, F., Fischer, C., & Morlet, D. (2011, September). Exacerbated attention orienting to auditory stimulation in migraine patients. *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, *122*(9), 1755–1763.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977, April). **Maximum Likelihood from Incomplete Data via the EM Algorithm** . *Journal of the Royal Statistical Society. Series B Methodological*, *39*(1), 1–38.

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

Deouell, L. Y., Bentin, S., & Giard, M. H. (1998, July). Mismatch negativity in dichotic listening: evidence for interhemispheric differences and multiple generators. *Psychophysiology*, *35*(4), 355–365.

Devaine, M., Hollard, G., & Daunizeau, J. (2014, December). The social Bayesian brain: does mentalizing make a difference when we learn? *PLoS Computational Biology*, *10*(12), e1003992.

Dima, D., Frangou, S., Burge, L., Braeutigam, S., & James, A. C. (2012, March). Abnormal intrinsic and extrinsic connectivity within the magnetic mismatch negativity brain network in schizophrenia: A preliminary study. *Schizophrenia Research*, *135*(1-3), 23–27.

Doeller, C. F., Opitz, B., Mecklinger, A., Krick, C., Reith, W., & Schröger, E. (2003, October). Prefrontal cortex involvement in preattentive auditory deviance detection: neuroimaging and electrophysiological evidence. *NeuroImage*, *20*(2), 1270–1282.

Ernst, M. O., & Banks, M. S. (2002, January). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433.

Escera, C., Alho, K., Schröger, E., & Winkler, I. (2000, May). Involuntary attention and distractibility as evaluated with event-related brain potentials. *Audiology & neuro-otology*, *5*(3-4), 151–166.

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

Escera, C., & Malmierca, M. S. (2014, February). The auditory novelty system: an attempt to integrate human and animal research. *Psychophysiology*, *51*(2), 111–123.

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

Felleman, D. J., & Van Essen, D. C. (1991a). Distributed hierarchical processing in the

primate cerebral cortex. *Cerebral cortex (New York, NY : 1991)*, *1*(1), 1–47.

Felleman, D. J., & Van Essen, D. C. (1991b, January). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, *1*(1), 1–47.

Fischer, C., Luaute, J., Adeleine, P., & Morlet, D. (2004, August). Predictive value of sensory and cognitive evoked potentials for awakening from coma. *Neurology*, *63*(4), 669–673.

Fishman, Y. I. (2014, July). The mechanisms and meaning of the mismatch negativity. *Brain Topography*, *27*(4), 500–526.

Flandin, G., & Penny, W. D. (2007, February). Bayesian fMRI data analysis with sparse spatial basis function priors. *NeuroImage*, *34*(3), 1108–1125.

Friston, K. (2005, April). A theory of cortical responses. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*, *360*(1456), 815–836.

Friston, K. (2008, November). Hierarchical Models in the Brain. , *4*(11), e1000211.

Friston, K. (2010, February). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138.

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

Friston, K., Adams, R. A., Perrinet, L., & Breakspear, M. (2012). Perceptions as hypotheses: saccades as experiments. *Frontiers in psychology*, *3*, 151.

Friston, K., Bastos, A. M., Pinotsis, D., & Litvak, V. (2015, April). LFP and oscillations-what do they tell us? *Current opinion in neurobiology*, *31*, 1–6.

Friston, K., Chu, C., Mourão-Miranda, J., Hulme, O., Rees, G., Penny, W. D., & Ashburner, J. (2008, January). Bayesian decoding of brain images. *NeuroImage*, *39*(1), 181–205.

Friston, K., Daunizeau, J., Kilner, J., & Kiebel, S. (2010, March). Action and behavior: a free-energy formulation. *Biological cybernetics*, *102*(3), 227–260.

Friston, K., Harrison, L., Daunizeau, J., Kiebel, S., Phillips, C., Trujillo-Barreto, N., ... Mattout, J. (2008, February). Multiple sparse priors for the M/EEG inverse problem. *NeuroImage*, *39*(3), 1104–1120.

Friston, K., Harrison, L., & Penny, W. D. (2003, August). Dynamic causal modelling. *NeuroImage*, *19*(4), 1273–1302.

Friston, K., Henson, R., Phillips, C., & Mattout, J. (2006a, September). Bayesian estimation of evoked and induced responses. *Human Brain Mapping*, *27*(9), 722–735.

Friston, K., Henson, R., Phillips, C., & Mattout, J. (2006b, September). Bayesian estimation of evoked and induced responses. *Human Brain Mapping*, *27*(9), 722–735.

Friston, K., & Kiebel, S. (2009, May). Predictive coding under the free-energy principle. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*, *364*(1521), 1211–1221.

Friston, K., Kilner, J., & Harrison, L. (2006, July). A free energy principle for the brain. *Journal of physiology, Paris*, *100*(1-3), 70–87.

Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., & Penny, W. D. (2007, January). Variational free energy and the Laplace approximation. *NeuroImage*, *34*(1), 220–234.

Friston, K., Rigoli, F., Ognibene, D., Mathys, C. D., Fitzgerald, T. H. B., & Pezzulo, G. (2015, March). Active inference and epistemic value. *Cognitive Neuroscience*, 1–28.

Friston, K., & Stephan, K. E. (2007a, September). Free-energy and the brain. *Synthese*, *159*(3), 417–458.

Friston, K., & Stephan, K. E. (2007b, December). Free-energy and the brain. *Synthese*, *159*(3), 417–458.

Friston, K., Stephan, K. E., Montague, R., & Dolan, R. J. (2014, July). Computational psychiatry: the brain as a phantastic organ. *The lancet. Psychiatry*, *1*(2), 148–158.

Fuchs, M., Wagner, M., Wischmann, H. A., Köhler, T., Theissen, A., Drenckhahn, R., & Buchner, H. (1998, August). Improving source reconstructions by combining bioelectric and biomagnetic data. *Electroencephalography and clinical neurophysiology*, *107*(2), 93–111.

Fulham, W. R., Michie, P. T., Ward, P. B., Rasser, P. E., Todd, J., Johnston, P. J., ... Schall, U. (2014, June). Mismatch Negativity in Recent-Onset and Chronic Schizophrenia: A Current Source Density Analysis. *PLoS ONE*, *9*(6), e100221.

Furl, N., Kumar, S., Alter, K., Durrant, S., Shawe-Taylor, J., & Griffiths, T. D. (2011, February). Neural prediction of higher-order auditory sequence statistics. , *54*(3), 2267–2277.

Garrido, M. I., Friston, K., Kiebel, S., Stephan, K. E., Baldeweg, T., & Kilner, J. M. (2008, August). The functional anatomy of the MMN: a DCM study of the roving paradigm. *NeuroImage*, *42*(2), 936–944.

Garrido, M. I., Kilner, J. M., Kiebel, S., & Friston, K. (2007, December). Evoked brain responses are generated by feedback loops. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(52), 20961–20966.

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

Garrido, M. I., Kilner, J. M., Kiebel, S., Stephan, K. E., Baldeweg, T., & Friston, K. (2009, October). Repetition suppression and plasticity in the human brain. *NeuroImage*, *48*(1), 269–279.

Garrido, M. I., Kilner, J. M., Kiebel, S., Stephan, K. E., & Friston, K. (2007, July). Dynamic causal modelling of evoked potentials: A reproducibility study. , *36*(3), 571–580.

Garrido, M. I., Kilner, J. M., Stephan, K. E., & Friston, K. (2009, March). The mismatch negativity: a review of underlying mechanisms. *Clinical neurophysiology : official*

journal of the International Federation of Clinical Neurophysiology, *120*(3), 453–463.

Gencer, N. G., & Acar, C. E. (2004, February). Sensitivity of EEG and MEG measurements to tissue conductivity. *Physics in Medicine and Biology*, *49*(5), 701–717.

Giacino, J. T., Kalmar, K., & Whyte, J. (2004, December). The JFK Coma Recovery Scale-Revised: measurement characteristics and diagnostic utility. *Archives of physical medicine and rehabilitation*, *85*(12), 2020–2029.

Giard, M. H., Lavikahen, J., Reinikainen, K., Pessiglione, M., Bertrand, O., Pernier, J., & Näätänen, R. (1995). Separate representation of stimulus frequency, intensity, and duration in auditory sensory memory: an event-related potential and dipole-model analysis. *Journal of Cognitive Neuroscience*, *7*(2), 133–143.

Giard, M. H., Pessiglione, M., Pernier, J., & Bouchet, P. (1990, November). Brain generators implicated in the processing of auditory stimulus deviance: a topographic event-related potential study. *Psychophysiology*, *27*(6), 627–640.

Gomot, M., Bernard, F. A., Davis, M. H., Belmonte, M. K., Ashwin, C., Bullmore, E. T., & Baron-Cohen, S. (2006, January). Change detection in children with autism: An auditory event-related fMRI study. , *29*(2), 475–484.

Gomot, M., Giard, M. H., Adrien, J.-L., Barthelemy, C., & Bruneau, N. (2002, September). Hypersensitivity to acoustic change in children with autism: electrophysiological evidence of left frontal cortex dysfunctioning. *Psychophysiology*, *39*(5), 577–584.

Gomot, M., Giard, M. H., Roux, S., Barthélémy, C., & Bruneau, N. (2000, September). Maturation of frontal and temporal components of mismatch negativity (MMN) in children. *Neuroreport*, *11*(14), 3109–3112.

Gramfort, A., Papadopoulo, T., Olivi, E., & Clerc, M. (2010). OpenMEEG: opensource software for quasistatic bioelectromagnetics. *Biomedical engineering online*, *9*(1), 45.

Grill-Spector, K., Henson, R., & Martin, A. (2006, January). Repetition and the brain: neural models of stimulus-specific effects. *Trends in Cognitive Sciences*, *10*(1), 14–23.

Grotheer, M., & Kovács, G. (2016, January). Can predictive coding explain repetition suppression? *CORTEX*.

Güllmar, D., Haueisen, J., & Reichenbach, J. R. (2010, May). Influence of anisotropic electrical conductivity in white matter tissue on the EEG/MEG forward and inverse solution. A high-resolution whole head simulation study. *NeuroImage*, *51*(1), 145–163.

Haker, H., Schneebeli, M., & Stephan, K. E. (2016, June). Can Bayesian Theories of Autism Spectrum Disorder Help Improve Clinical Practice? *Frontiers in psychiatry*, *7*(3), 1174–17.

Hämäläinen, M. S., & Sarvas, J. (1989, February). Realistic conductivity geometry model

of the human head for interpretation of neuromagnetic data. *IEEE transactions on bio-medical engineering*, *36*(2), 165–171.

Hanna, J. (2014, November). Neurophysiological evidence for whole form retrieval of complex derived words: a mismatch negativity study. , 1–13.

Heinzle, J., Hepp, K., & Martin, K. A. C. (2007, August). A Microcircuit Model of the Frontal Eye Fields. *Journal of Neuroscience*, *27*(35), 9341–9353.

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

Huotilainen, M., Winkler, I., Alho, K., Escera, C., Virtanen, J., Ilmoniemi, R. J., ... Näätänen, R. (1998, July). Combined mapping of human auditory EEG and MEG responses. *Electroencephalography and clinical neurophysiology*, *108*(4), 370–379.

Iglesias, S., Mathys, C. D., Brodersen, K. H., Kasper, L., Piccirelli, M., den Ouden, H. E. M., & Stephan, K. E. (2013, October). Hierarchical Prediction Errors in Midbrain and Basal Forebrain during Sensory Learning. *Neuron*, *80*(2), 519–530.

Jääskeläinen, I. P., Ahveninen, J., Bonmassar, G., Dale, A. M., Ilmoniemi, R. J., Levänen, S., ... Belliveau, J. W. (2004, April). Human posterior auditory cortex gates novel sounds to consciousness. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(17), 6809–6814.

Jansen, B. H., & Rit, V. G. (1995, September). Electroencephalogram and visual evoked potential generation in a mathematical model of coupled cortical columns. *Biological cybernetics*, *73*(4), 357–366.

Jemel, B., Achenbach, C., Müller, B. W., Röpcke, B., & Oades, R. D. (2002). Mismatch negativity results from bilateral asymmetric dipole sources in the frontal and temporal lobes. *Brain Topography*, *15*(1), 13–27.

Joffily, M., & Coricelli, G. (2013). Emotional valence and the free-energy principle. *PLoS Computational Biology*, *9*(6), e1003094.

Kass, R., & Raftery, A. E. (1995). Bayes Factor. *Journal of the American Statistical Association*, *90*(430), 773–795.

Kiebel, S., Daunizeau, J., & Friston, K. (2008, November). A Hierarchy of Time-Scales and the Brain. *PLoS Computational Biology*, *4*(11), e1000209–12.

Kiebel, S., David, O., & Friston, K. (2006, May). Dynamic causal modelling of evoked responses in EEG/MEG with lead field parameterization. , *30*(4), 1273–1284.

Kiebel, S., Garrido, M. I., & Friston, K. (2007, June). Dynamic causal modelling of evoked responses: The role of intrinsic connections. , *36*(2), 332–345.

Kiebel, S., Garrido, M. I., Moran, R., Chen, C. C., & Friston, K. (2009, June). Dynamic causal modeling for EEG and MEG. *Human Brain Mapping*, *30*(6), 1866–1876.

Knill, D. C., & Pouget, A. (2004, December). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in neurosciences*, *27*(12), 712–719.

Kok, P., & de Lange, F. P. (2015, April). Predictive Coding in Sensory Cortex. In E.-J. Wagenmakers & B. U. Forstmann (Eds.), *An introduction to model-based cognitive*

neuroscience (pp. 221–244). New-York.

Kok, P., Jehee, J. F. M., & de Lange, F. P. (2012, July). Less is more: expectation sharpens representations in the primary visual cortex. *Neuron*, *75*(2), 265–270.

Lappe, C., Steinsträter, O., & Pantev, C. (2013a). A beamformer analysis of MEG data reveals frontal generators of the musically elicited mismatch negativity. *PLoS ONE*, *8*(4), e61296.

Lappe, C., Steinsträter, O., & Pantev, C. (2013b). Rhythmic and melodic deviations in musical sequences recruit different cortical areas for mismatch detection. *Frontiers in Human Neuroscience*, *7*, 260.

Lecaignard, F., Bertrand, O., Gimenez, G., Mattout, J., & Caclin, A. (2015). Implicit learning of predictable sound sequences modulates human brain responses at different levels of the auditory hierarchy. *Frontiers in Human Neuroscience*, *9*, 505.

Levänen, S., Ahonen, A., Hari, R., McEvoy, L., & Sams, M. (1996, March). Deviant auditory stimuli activate human left and right auditory cortex differently. *Cerebral Cortex*, *6*(2), 288–296.

Lieder, F., Daunizeau, J., Garrido, M. I., Friston, K., & Stephan, K. E. (2013). Modelling trial-by-trial changes in the mismatch negativity. *PLoS Computational Biology*, *9*(2), 1–16.

Lieder, F., Stephan, K. E., Daunizeau, J., Garrido, M. I., & Friston, K. (2013). A neurocomputational model of the mismatch negativity. *PLoS Computational Biology*, *9*(11), 1–14.

Liégeois-Chauvel, C., Musolino, A., Badier, J.-M., Marquis, P., & Chauvel, P. (1994, May). Evoked potentials recorded from the auditory cortex in man: evaluation and topography of the middle latency components. *Electroencephalography and clinical neurophysiology*, *92*(3), 204–214.

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

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

MacLean, S. E., Blundon, E. G., & Ward, L. M. (2015, August). Brain regional networks active during the mismatch negativity vary with paradigm. *Neuropsychologia*, *75*, 242–251.

Maess, B., Jacobsen, T., Schröger, E., & Friederici, A. D. (2007, August). Localizing pre-attentive auditory memory-based comparison: magnetic mismatch negativity to pitch change. , *37*(2), 561–571.

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

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

Markov, N. T., Vezoli, J., Chameau, P., Falchier, A., Quilodran, R., Huissoud, C., ...

Kennedy, H. (2014, January). Anatomy of hierarchy: feedforward and feedback pathways in macaque visual cortex. *The Journal of comparative neurology*, *522*(1), 225–259.

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

Mathys, C. D., Daunizeau, J., Friston, K., & Stephan, K. E. (2011). A bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience*, *5*, 39.

Mathys, C. D., Lomakina, E. I., Daunizeau, J., Iglesias, S., Brodersen, K. H., Friston, K., & Stephan, K. E. (2014). Uncertainty in perception and the Hierarchical Gaussian Filter. *Frontiers in Human Neuroscience*, *8*(47), 825.

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

Mattout, J., Phillips, C., Penny, W. D., Rugg, M. D., & Friston, K. (2006, April). MEG source localization under multiple constraints: an extended Bayesian framework. *NeuroImage*, *30*(3), 753–767.

Maunsell, J. H., & Van Essen, D. C. (1983, December). The connections of the middle temporal visual area (MT) and their relationship to a cortical hierarchy in the macaque monkey. *Journal of Neuroscience*, *3*(12), 2563–2586.

May, P. J. C., & Tiitinen, H. (2010, January). Mismatch negativity (MMN), the deviance-elicited auditory deflection, explained. *Psychophysiology*, *47*(1), 66–122.

May, P. J. C., Westö, J., & Tiitinen, H. (2015, March). Computational modelling suggests that temporal integration results from synaptic adaptation in auditory cortex. *The European journal of neuroscience*, *41*(5), 615–630.

Molholm, S., Martinez, A., Ritter, W., Javitt, D. C., & Foxe, J. J. (2005, May). The neural circuitry of pre-attentive auditory change-detection: an fMRI study of pitch and duration mismatch negativity generators. *Cerebral Cortex*, *15*(5), 545–551.

Molins, A., Hämäläinen, M. S., Stufflebeam, S., & Brown, E. N. (2008, September). Quantification of the benefit from integrating MEG and EEG data in minimum l2-norm estimation. , *42*(3), 1069–1077.

Moran, R., Campo, P., Symmonds, M., Stephan, K. E., Dolan, R. J., & Friston, K. (2013, May). Free energy, precision and learning: the role of cholinergic neuromodulation. *Journal of Neuroscience*, *33*(19), 8227–8236.

Moran, R., Pinotsis, D. A., & Friston, K. (2013). Neural masses and fields in dynamic causal modeling. *Frontiers in computational neuroscience*, *7*, 57.

Morlet, D., Demarquay, G. v., Brudon, F., Fischer, C., & Caclin, A. (2014, March). Attention orienting dysfunction with preserved automatic auditory change detection in migraine. *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, *125*(3), 500–511.

Morlet, D., & Fischer, C. (2014, July). MMN and novelty P3 in coma and other altered states of consciousness: a review. *Brain Topography*, *27*(4), 467–479.

Mumford, D. (1992). On the computational architecture of the neocortex. *Biological cybernetics*, *66*(3), 241–251.

Näätänen, R. (1992, April). Attention and Event-Related Potentials. In Hove & London (Eds.), *Attention and brain function* (pp. 236–352). Hillsdale, New Jersey.

Näätänen, R. (2003, May). Mismatch negativity: clinical research and possible applications. *International journal of psychophysiology : official journal of the International Organization of Psychophysiology*, *48*(2), 179–188.

Näätänen, R., & Alho, K. (1995). Generators of electrical and magnetic mismatch responses in humans. *Brain Topography*, *7*(4), 315–320.

Näätänen, R., Astikainen, P., Ruusuvirta, T., & Huotilainen, M. (2010, September). Automatic auditory intelligence: An expression of the sensory. *Brain Research Reviews*, *64*(1), 123–136.

Näätänen, R., Gaillard, A. W., & Mäntysalo, S. (1978, July). Early selective-attention effect on evoked potential reinterpreted. *Acta psychologica*, *42*(4), 313–329.

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

Näätänen, R., Paavilainen, P., Rinne, T., & Alho, K. (2007, December). The mismatch negativity (MMN) in basic research of central auditory processing: a review. *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, *118*(12), 2544–2590.

Näätänen, R., Sussman, E. S., Salisbury, D., & Shafer, V. L. (2014, July). Mismatch negativity (MMN) as an index of cognitive dysfunction. *Brain Topography*, *27*(4), 451–466.

Nunez, P. L., & Harth, E. (1982). Electric Fields of the Brain: The Neurophysics of EEG. *Physics Today*, *35*(6), 59.

Opitz, B., Rinne, T., Mecklinger, A., von Cramon, D. Y., & Schröger, E. (2002, January). Differential contribution of frontal and temporal cortices to auditory change detection: fMRI and ERP results. , *15*(1), 167–174.

O'Reilly, J. X., Jbabdi, S., & Behrens, T. E. J. (2012, April). How can a Bayesian approach inform neuroscience? *The European journal of neuroscience*, *35*(7), 1169–1179.

Ostwald, D., Spitzer, B., Guggenmos, M., Schmidt, T. T., Kiebel, S., & Blankenburg, F. (2012, August). Evidence for neural encoding of Bayesian surprise in human somatosensation. *NeuroImage*, *62*(1), 177–188.

Paavilainen, P., Alho, K., Reinikainen, K., Sams, M., & Näätänen, R. (1991, June). Right hemisphere dominance of different mismatch negativities. *Electroencephalography and clinical neurophysiology*, *78*(6), 466–479.

Pantev, C., Bertrand, O., Eulitz, C., Verkindt, C., Hampson, S., Schuierer, G., & Elbert, T. (1995, January). Specific tonotopic organizations of different areas of the human auditory cortex revealed by simultaneous magnetic and electric recordings.

Electroencephalography and clinical neurophysiology, *94*(1), 26–40.

Paraskevopoulos, E., Kuchenbuch, A., Herholz, S. C., Foroglou, N., Bamidis, P., & Pantev, C. (2014, June). Tones and numbers: A combined EEG-MEG study on the effects of musical expertise in magnitude comparisons of audiovisual stimuli. *Human Brain Mapping*, *35*(11), 5389–5400.

Penny, W. D., Stephan, K. E., Daunizeau, J., Rosa, M. J., Friston, K., Schofield, T. M., & Leff, A. P. (2010, March). Comparing families of dynamic causal models. *PLoS Computational Biology*, *6*(3), e1000709.

Pérez-González, D., & Malmierca, M. S. (2014). Adaptation in the auditory system: an overview. *Frontiers in integrative neuroscience*, *8*, 19.

Picton, T. W. (1980). The use of human event-related potentials in psychology. *Techniques in psychophysiology*, 357–395.

Plonsey, R., & Heppner, D. B. (1967, December). Considerations of quasi-stationarity in electrophysiological systems. *The Bulletin of mathematical biophysics*, *29*(4), 657–664.

Pouget, A., Beck, J. M., Ma, W. J., & Latham, P. E. (2013, September). Probabilistic brains: knowns and unknowns. *Nature Neuroscience*, *16*(9), 1170–1178.

Rao, R. P., & Ballard, D. H. (1999, January). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87.

Recasens, M., Grimm, S., Capilla, A., Nowak, R., & Escera, C. (2014, January). Two sequential processes of change detection in hierarchically ordered areas of the human auditory cortex. *Cerebral cortex (New York, NY : 1991)*, *24*(1), 143–153.

Recasens, M., Grimm, S., Wollbrink, A., Pantev, C., & Escera, C. (2014, November). Encoding of nested levels of acoustic regularity in hierarchically organized areas of the human auditory cortex. *Human Brain Mapping*, *35*(11), 5701–5716.

Recasens, M., Leung, S., Grimm, S., Nowak, R., & Escera, C. (2015, March). Repetition suppression and repetition enhancement underlie auditory memory-trace formation in the human brain: an MEG study. *NeuroImage*, *108*, 75–86.

Restuccia, D., Della Marca, G., Valeriani, M., Leggio, M. G., & Molinari, M. (2007, January). Cerebellar damage impairs detection of somatosensory input changes. A somatosensory mismatch-negativity study. *Brain*, *130*(Pt 1), 276–287.

Rinne, T., Alho, K., Ilmoniemi, R. J., Virtanen, J., & Näätänen, R. (2000, July). Separate time behaviors of the temporal and frontal mismatch negativity sources. *NeuroImage*, *12*(1), 14–19.

Rinne, T., Degerman, A., & Alho, K. (2005, May). Superior temporal and inferior frontal cortices are activated by infrequent sound duration decrements: an fMRI study. , *26*(1), 66–72.

Rissling, A. J., Miyakoshi, M., Sugar, C. A., Braff, D. L., Makeig, S., & Light, G. A. (2014). Cortical substrates and functional correlates of auditory deviance processing

deficits in schizophrenia. *YNICL*, *6*(C), 424–437.

Robic, S., Sonié, S., Fonlupt, P., Henaff, M. A., Touil, N., Coricelli, G., ... Schmitz, C. (2014, November). Decision-Making in a Changing World: A Study in Autism Spectrum Disorders. *Journal of Autism and Developmental Disorders*, *45*(6), 1603–1613.

Rockland, K. S., & Pandya, D. N. (1979, December). Laminar origins and terminations of cortical connections of the occipital lobe in the rhesus monkey. *Brain Research*, *179*(1), 3–20.

Rosburg, T., Haueisen, J., & Kreitschmann-Andermahr, I. (2004, April). The dipole location shift within the auditory evoked neuromagnetic field components N100m and mismatch negativity (MMNm). *Clinical Neurophysiology*, *115*(4), 906–913.

Ruby, P., Caclin, A., Boulet, S., Delpuech, C., & Morlet, D. (2008, February). Odd sound processing in the sleeping brain. *Journal of Cognitive Neuroscience*, *20*(2), 296–311.

Ruhnau, P., Herrmann, B., Maess, B., Brauer, J., Friederici, A. D., & Schröger, E. (2013). Processing of complex distracting sounds in school-aged children and adults: evidence from EEG and MEG data. *Frontiers in psychology*, *4*, 717.

Ruhnau, P., Herrmann, B., & Schröger, E. (2012, March). Finding the right control: the mismatch negativity under investigation. *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, *123*(3), 507–512.

Rush, S., & Driscoll, D. A. (1968, November). Current distribution in the brain from surface electrodes. *Anesthesia and analgesia*, *47*(6), 717–723.

Sams, M., Alho, K., & Näätänen, R. (1984, July). Short-term habituation and dishabituation of the mismatch negativity of the ERP. *Psychophysiology*, *21*(4), 434–441.

Sams, M., Kaukoranta, E., Hämäläinen, M. S., & Näätänen, R. (1991, January). Cortical activity elicited by changes in auditory stimuli: different sources for the magnetic N100m and mismatch responses. *Psychophysiology*, *28*(1), 21–29.

Sanchez, G., Daunizeau, J., Maby, E., Bertrand, O., Bompas, A., & Mattout, J. (2014, March). Toward a New Application of Real-Time Electrophysiology: Online Optimization of Cognitive Neurosciences Hypothesis Testing. *Brain Sciences*, *4*(1), 49–72.

Schairer, K. S., Gould, H. J., & Pousson, M. A. (2001). Source generators of mismatch negativity to multiple deviant stimulus types. *Brain Topography*, *14*(2), 117–130.

Scherg, M., Vajsar, J., & Picton, T. W. (1989). A Source Analysis of the Late Human Auditory Evoked Potentials. *Journal of Cognitive Neuroscience*, *1*(4), 336–355.

Schmidt, A., Diaconescu, A. O., Kometer, M., Friston, K., Stephan, K. E., & Vollenweider, F. X. (2013, September). Modeling Ketamine Effects on Synaptic Plasticity During the Mismatch Negativity. *Cerebral Cortex*, *23*(10), 2394–2406.

Schofield, T. M., Iverson, P., Kiebel, S., Stephan, K. E., Kilner, J. M., Friston, K., ... Leff, A. P. (2009, July). Changing meaning causes coupling changes within higher

levels of the cortical hierarchy. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(28), 11765–11770.

Schönwiesner, M., Novitski, N., Pakarinen, S., Carlson, S., Tervaniemi, M., & Näätänen, R. (2007, March). Heschl's gyrus, posterior superior temporal gyrus, and mid-ventrolateral prefrontal cortex have different roles in the detection of acoustic changes. *Journal of Neurophysiology*, *97*(3), 2075–2082.

Schröger, E., & Wolff, C. (1996, November). Mismatch response of the human brain to changes in sound location. *Neuroreport*, *7*(18), 3005–3008.

Schwartenbeck, P., Fitzgerald, T. H. B., Mathys, C. D., Dolan, R., & Friston, K. (2015, October). The Dopaminergic Midbrain Encodes the Expected Certainty about Desired Outcomes. *Cerebral cortex (New York, NY : 1991)*, *25*(10), 3434–3445.

Schwartenbeck, P., Fitzgerald, T. H. B., Mathys, C. D., Dolan, R., Kronbichler, M., & Friston, K. (2015, November). Evidence for surprise minimization over value maximization in choice behavior. *Nature Publishing Group*, 1–14.

Sharon, D., Hämäläinen, M. S., Tootell, R. B. H., Halgren, E., & Belliveau, J. W. (2007, July). The advantage of combining MEG and EEG: comparison to fMRI in focally stimulated visual cortex. , *36*(4), 1225–1235.

Shiga, T., Althen, H., Cornella, M., Zarnowiec, K., Yabe, H., & Escera, C. (2015). Deviance-Related Responses along the Auditory Hierarchy: Combined FFR, MLR and MMN Evidence. *PLoS ONE*, *10*(9), e0136794.

Slabu, L., Grimm, S., & Escera, C. (2012, January). Novelty detection in the human auditory brainstem. *Journal of Neuroscience*, *32*(4), 1447–1452.

Sohoglu, E., & Davis, M. H. (2016, March). Perceptual learning of degraded speech by minimizing prediction error. *Proceedings of the National Academy of Sciences of the United States of America*, 201523266–17.

Spratling, M. W. (2016, January). A review of predictive coding algorithms. *Brain and cognition*.

Stephan, K. E., Iglesias, S., Heinzle, J., & Diaconescu, A. O. (2015a, August). Translational Perspectives for Computational Neuroimaging. *Neuron*, *87*(4), 716–732.

Stephan, K. E., Iglesias, S., Heinzle, J., & Diaconescu, A. O. (2015b, August). Translational Perspectives for Computational Neuroimaging. *Neuron*, *87*(4), 716–732.

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R., & Friston, K. (2009, July). Bayesian model selection for group studies. *NeuroImage*, *46*(4), 1004–1017.

Summerfield, C., Behrens, T. E. J., & Koechlin, E. (2011, August). Perceptual classification in a rapidly changing environment. *Neuron*, *71*(4), 725–736.

Summerfield, C., & de Lange, F. P. (2014, October). Expectation in perceptual decision making: neural and computational mechanisms. *Nature Publishing Group*, 1–12.

Summerfield, C., Trittschuh, E. H., Monti, J. M., Mesulam, M.-M., & Egner, T. (2008, September). Neural repetition suppression reflects fulfilled perceptual expectations. *Nature Neuroscience*, *11*(9), 1004–1006.

Sussman, E. S., Chen, S., Sussman-Fort, J., & Dinces, E. (2014, July). The five myths of MMN: redefining how to use MMN in basic and clinical research. *Brain Topography*, *27*(4), 553–564.

Sussman, E. S., & Shafer, V. L. (2014, June). New Perspectives on the Mismatch Negativity (MMN) Component: An Evolving Tool in Cognitive Neuroscience. *Brain Topography*, *27*(4), 425–427.

Téglás, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B., & Bonatti, L. L. (2011, May). Pure reasoning in 12-month-old infants as probabilistic inference. *Science (New York, NY)*, *332*(6033), 1054–1059.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011, March). How to grow a mind: statistics, structure, and abstraction. *Science (New York, NY)*, *331*(6022), 1279–1285.

Tillmann, B., Bigand, E., & Pineau, M. (1998, October). Effects of Global and Local Contexts on Harmonic Expectancy. *Music Perception: An Interdisciplinary Journal*, *16*(1), 99–117.

Todd, J., & Mullens, D. (2011, April). Implementing conditional inference in the auditory system: What matters? *Psychophysiology*.

Todd, J., Provost, A., Whitson, L. R., Cooper, G., & Heathcote, A. (2013, January). Not so primitive: context-sensitive meta-learning about unattended sound sequences. *Journal of Neurophysiology*, *109*(1), 99–105.

Todorovic, A., & de Lange, F. P. (2012, September). Repetition suppression and expectation suppression are dissociable in time in early auditory evoked fields. *Journal of Neuroscience*, *32*(39), 13389–13395.

Tomiello, S., Schöbi, D., Weber, L. A. E., Wellstein, K., Stefanics, G., Haker, H., ... Stephan, K. E. (2016, June). **Timing of prediction error signaling in reward learning: A computational trial-by-trial EEG analysis**. In *Organization for human brain mapping* (pp. 1–1). Geneva.

Trujillo-Barreto, N. J. (2015). Bayesian Model Inference. In A. Toga (Ed.), *Brain mapping an encyclopedic reference* (pp. 535–539). Lisa Tickner.

Tse, C.-Y., & Penney, T. B. (2008, July). On the functional role of temporal and frontal cortex activation in passive detection of auditory deviance. *NeuroImage*, *41*(4), 1462–1470.

Tse, C.-Y., Rinne, T., Ng, K. K., & Penney, T. B. (2013, December). The functional role of the frontal cortex in pre-attentive auditory change detection. *NeuroImage*, *83*, 870–879.

Ulanovsky, N., Las, L., & Nelken, I. (2003, March). Processing of low-probability sounds by cortical neurons. *Nature Neuroscience*, *6*(4), 391–398.

Vallaghé, S., & Clerc, M. (2009, April). A global sensitivity analysis of three- and four-layer EEG conductivity models. *IEEE transactions on bio-medical engineering*, *56*(4), 988–995.

Waberski, T. D., Kreitschmann-Andermahr, I., Kawohl, W., Darvas, F., Ryang, Y., Rodewald, M., ... Buchner, H. (2001, August). Spatio-temporal source imaging reveals subcomponents of the human auditory mismatch negativity in the cingulum and right inferior temporal gyrus. *Neuroscience letters*, *308*(2), 107–110.

Wacongne, C. (2016, April). A predictive coding account of MMN reduction in schizophrenia. *Biological psychology*, *116*, 68–74.

Wacongne, C., Changeux, J.-P., & Dehaene, S. (2012, March). A neuronal model of predictive coding accounting for the mismatch negativity. *Journal of Neuroscience*, *32*(11), 3665–3678.

Weber, L. A. E., Diaconescu, A. O., Mathys, C. D., Schmidt, A., Kometer, M., Vollenweider, F., & Stephan, K. E. (2015, May). A computational single-trial analysis of MMN under ketamine. In *Mmn conference* (pp. 1–1). Leipzig.

Weber, L. A. E., Tomiello, S., Schöbi, D., Iglesias, S., Diaconescu, A. O., Stefanics, G., ... Stephan, K. E. (2016, June). Hierarchical Prediction Errors during Auditory Mismatch: A Computational Single-Trial EEG Analysis. In *Organization for human brain mapping.* Geneva.

Westheimer, G. (2008). Was Helmholtz a Bayesian? *Perception*, *37*(5), 642–650.

Winkler, I. (2007, January). Interpreting the Mismatch Negativity. *Journal of Psychophysiology*, *21*(3-4), 147–163.

Winkler, I., & Czigler, I. (2012, February). Evidence from auditory and visual event-related potential (ERP) studies of deviance detection (MMN and vMMN) linking predictive coding theories and perceptual object representations. *International journal of psychophysiology : official journal of the International Organization of Psychophysiology*, *83*(2), 132–143.

Winkler, I., Karmos, G., & Näätänen, R. (1996, December). Adaptive modeling of the unattended acoustic environment reflected in the mismatch negativity event-related potential. *Brain Research*, *742*(1-2), 239–252.

Yago, E., Escera, C., Alho, K., & Giard, M. H. (2001, August). Cerebral mechanisms underlying orienting of attention towards auditory frequency changes. *Neuroreport*, *12*(11), 2583–2587.

Yvert, B., Crouzeix, A., Bertrand, O., Seither-Preisler, A., & Pantev, C. (2001, May). Multiple supratemporal sources of magnetic and electric auditory evoked middle latency components in humans. *Cerebral cortex (New York, NY : 1991)*, *11*(5), 411–423.

Yvert, B., Fischer, C., Guénot, M., Krolak-Salmon, P., Isnard, J., & Pernier, J. (2002, September). Simultaneous intracerebral EEG recordings of early auditory thalamic and cortical activity in human. *The European journal of neuroscience*, *16*(6), 1146–1150.

Zatorre, R. J., & Belin, P. (2001, October). Spectral and temporal processing in human auditory cortex. *Cerebral Cortex*, *11*(10), 946–953.