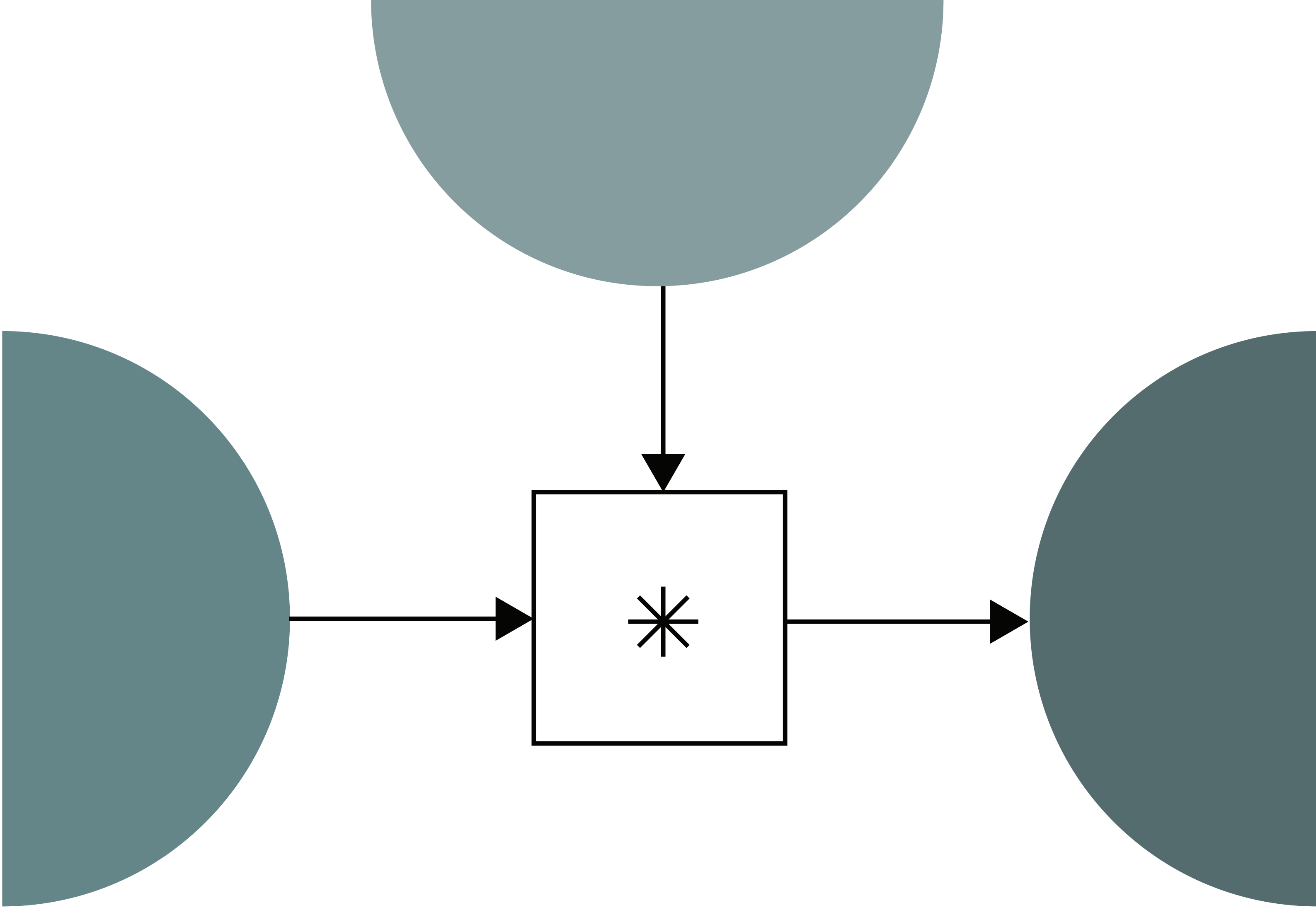

DYNAMIC CAUSAL MODELS FOR INFERENCE ON NEUROMODULATORY PROCESSES IN NEURAL CIRCUITS

Dario Schöbi

Diss. ETH No. 26727

DISS. ETH NO. 26727

DYNAMIC CAUSAL MODELS FOR INFERENCE ON NEUROMODULATORY PROCESSES IN NEURAL CIRCUITS

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

DARIO SCHÖBI

MSc ETH Physics, ETH Zurich

born on *22.08.1988*

citizen of Berneck, SG

accepted on the recommendation of

Prof. Dr. Klaas Enno Stephan (examiner)

Prof. Dr. Olivier David (co-examiner)

PD Dr. Jakob Heinzle (co-examiner)

2020

ABSTRACT

Computational psychiatry aims at solving clinically relevant problems using computational methods. One such problem in psychiatry is the lack of quantitative stratification of patients with respect to treatment. The theoretical foundation of how the problem could potentially be solved has been laid out in the recent years. Dynamic system models were developed which allow to model effective connectivity between brain regions, in principle down to the resolution of synaptic properties, based on non-invasive brain imaging data. In combination with a Bayesian optimization framework, they allow for the definition and testing of hypotheses about hidden pathological mechanisms that could be underlying psychiatric disorders – a model aided way of *differential diagnosis*. These methods have found multiple applications in studies investigating different disorders and concepts of cognitive function in recent years. Still, computational psychiatry has yet to prove its utility in robustly answering clinical questions.

The success of the approach hinges on the framework to return accurate, robust and construct valid measures. In this dissertation we aim to assess these points for the measurement device: "Dynamic Causal Models". First, we investigate the standard integration scheme for a system of non-linear delay differential equations. We show that the standard scheme violates construct validity with respect to conductance-delays between cortical regions. We propose an alternative scheme which respects changes in the dynamics due to delays while still being orders of magnitude faster than the considered benchmark. Second, we implement and test an augmentation of the optimization method to overcome local optima. We show that local optima are clearly present, hence strict optimality of estimates and correct model identification cannot be guaranteed. Third, we derive an analytical relationship between the theoretical Bayesian criteria of model goodness and a frequentist measure (residual sum of squares). These relationship allows for the analytical

assessment of the robustness of results and their sensitivity to slight changes. Additionally, we explain a tendency of bias towards complex models often observed in empirical studies. We then employ these developments in two empirical datasets. First, we infer on graded, muscarinic changes to synaptic function in rodents based on electrophysiological recordings. We show that we can significantly distinguish muscarinic agents with agonistic and antagonistic effects. Additionally, we illustrate an approach how parameters can be constrained across multiple conditions different in their nature. Constraining improves classification accuracy and supports the theoretically discussed findings in the second methodological part. Second, we infer on effective connectivity in healthy controls performing in a working memory paradigm, using functional magnet-resonance imaging. The task is novel and explicitly designed to alleviate performance related confounds in patient studies. We also show how to estimate the amount of irreducible noise efficiently to generally inform prior specification of noise parameters for this modality of the models. This reduces the bias in favor of complex models, thus providing more reliable identification of hidden networks.

ZUSAMMENFASSUNG

Das Ziel von Komputationaler Psychiatrie besteht im Lösen von klinisch relevanten Problemen mithilfe komputationaler Methoden. Ein konkretes Problem der Psychiatrie ist der Mangel an quantitativen Methoden, welche die Stratifizierung von Patienten bezüglich Behandlungserfolgs erlauben würden. Die theoretischen Grundlagen, wie dieses Problem gelöst werden könnte, wurde in den vergangenen Jahren formuliert. Es wurden Modelle von dynamischen Systemen entwickelt, welche anhand nicht-invasiver bildgebenden Massnahmen die Modellierung von effektiver Konnektivität zwischen Hirnregionen, im Prinzip bis zur Auflösung von synaptischen Eigenschaften, erlauben. In Kombination mit Bayesianischen Optimierungsverfahren erlaubt dies das Aufstellen und Testen von Hypothesen zu latenten pathophysiologischen Mechanismen, welche psychiatrischen Erkrankungen zugrunde liegen könnten – ein modellbasierter Ansatz der *Differentialdiagnose*. Diese Methoden haben in den vergangenen Jahren diverse Anwendungen in der Untersuchung von verschiedenen Erkrankungen und Konzepten der kognitiven Funktionen gefunden. Nichtsdestotrotz, komputationale Psychiatrie bleibt noch immer eine konkrete Anwendung in der robusten Beantwortung einer klinisch relevanten Fragestellung schuldig.

Der Erfolg des Ansatzes hängt davon ab, dass die Methoden genaue, robuste und konstrukt-valide Messungen liefern. In dieser Dissertation untersuchen wir diese Punkte für die „Messgeräte“ vom Typ: “Dynamisch Kausale Modelle“. Zuerst untersuchen wir das Standard-Schema zur Integration von nicht-linearen, retardierten Differenzialgleichungen. Wir zeigen, dass das Standard-Schema, zumindest in Bezug auf Verzögerungen aufgrund Leitfähigkeit zwischen kortikalen Regionen, Konstruktvalidität verletzt. Wir definieren ein alternatives Schema, welches die durch Verzögerungen induzierten dynamischen Veränderungen respektiert, wobei es um ein Vielfaches schneller ist als ein Referenzschema. Zweitens implementieren und testen wir eine Erweiterung des Optimierungsverfahrens zur Überwindung lokaler Optima. Wir zeigen, dass lokale Optima klar vorhanden sind, und daher strikte Optimalität von Schätzwerten und korrekte Modellidentifikation nicht garantiert sind. Drittens leiten wir analytische Verhältnisse zwischen dem theoretischen, bayesianischen Kriterium von Modellgüte und einem frequentistischen Mass her (der erklärten Quadratsumme). Diese Beziehungen erlauben eine analytische Untersuchung der Robustheit der Resultate und ihrer Sensitivität unter kleinen Änderungen. Zusätzlich

erklären wir eine Neigung zu komplexeren Modellen, welche in empirischen Studien oft beobachtet worden ist.

Diese Entwicklungen wenden wir in zwei empirischen Datensätzen an. Zuerst inferieren wir, basierend auf elektrophysiologischen Messungen, abgestufte, muskarinerge Veränderungen von synaptischen Funktionen in Nagetieren. Wir zeigen signifikante Unterschiede zwischen muskarinergen Stoffen mit agonistischer und antagonistischer Wirkung. Zusätzlich illustrieren wir eine Möglichkeit, wie Modellparameter über mehrere Konditionen unterschiedlicher Art, eingeschränkt werden können. Diese Einschränkung verbessert die Klassifikationsgenauigkeit und verstärkt die theoretisch diskutierten Funde im zweiten methodischen Teil. Zweitens inferieren wir, basierend auf funktionellen Magnetresonanzbildern, effektive Konnektivität während einer Arbeitsgedächtnisaufgabe in einer gesunden Kontrollgruppe. Die Aufgabe selbst bietet eine, speziell für Patientenstudien motivierte Neuentwicklung, um Konfunde aufgrund von Leistungsunterschieden zu reduzieren. Zusätzlich zeigen wir eine Möglichkeit die nicht-reduzierbare Störung der Messung effizient zu schätzen, was die informierte Spezifikation der a priori Verteilung von Störungsparametern für diese Modalität der Modelle erlaubt. Dies reduziert die Neigung in der Bevorzugung komplexer Modelle und führt zu einer zuverlässigen Identifikation von Netzwerken.

ACKNOWLEDGEMENTS

First and foremost, I want to thank my supervisors Klaas Enno Stephan and Jakob Heinzle.

Klaas, thank you for your courage, persistence and will, to build up the TNU. Even after all these years, you put your heart and time on the line to create an environment that allows us young scientists to follow our research undisturbed. I think it is a tough endeavor and I truly hope you will be successful in finding the one (or many more) clinical application until you retire.

Jakob, suffice to say that I learned the most from you. You have an exceptional eye for consistency and logic. Thank you for your always open door and always open mind.

Many thanks to Olivier, for your time and effort to review my dissertation. It was an honor to be reviewed by the scientist who has defined the model I have worked with for so long.

Stefan and Eduardo, the two of you were probably my second most important sources of knowledge. I don't know how many 'quick questions' turned into long discussions, but you were always forthcoming in entering them.

I was blessed to not only have met co-workers, but also fantastic friends along the way. At times, I have spent more time with you than anyone else, Cao and Stefan. Also outside work, I experienced a never ending supply of support. All of you made me always feel welcome, no matter how long we had not seen each other – Mario, Simon, Stef, Sebastian, Alexander, Muriel, Nadine and my football team.

To the place I called home, whether it is in the rhine valley, Bern, Basel, St.Gallen or Liverpool - My family.

To Matthias; For a roof, a talk, a beer or controller. For a friendship over a decade. For being like a brother.

To Sara; My first and last line of defense during this PhD and many lines in between.

CONTENTS

1 | INTRODUCTION

1.1 THE MOTIVATION

Computational Psychiatry (CP) is a young field that emerged in the early 2000s, bringing the application of computational theories and mathematics into the realm of psychiatry. It is part of a larger family of disciplines embedded under the umbrella term of 'computational neuroscience' (Frässle, Yao et al. 2018). While 'computational neuroscience' generally aims at describing how the brain processes information, CP's goal is to develop and test mathematical models of the brain or behavior, to address and solve explicit clinical problems. One of the clinical problems in psychiatry is that the current symptom based nosology is successful in broadly stratifying patients into subgroups, yet these subgroups contain little predictive information about treatment success or disease trajectory. In a nutshell, psychiatry lacks a predictive 'biomarker' (Kapur, Phillips et al. 2012). CP tries to bridge this gap. The term 'computational' here has a double meaning. On one hand, CP sees the brain as an organ that performs computations in order to process information. As a biological system however, it is bounded by certain constraints such as availability of pathways and messengers that allow for the communication between brain areas. These areas are thought to have important functions for the interpretation of the information. This has been formalized by seeing the brain as an inference machine that creates a predictive, learning model of the world in order to minimize surprise about new, sensory stimulation (Friston, Kilner et al. 2006). However, if any of the previously mentioned (biological) resources are 'malfunctioning', this could lead to aberrant models of the world, aberrant model-adjustments and in turn constant "surprise"[1]. As a consequence, constant surprise causes a lot of stress (as it would on any organ). (Mal-) Adaptation to this stress might result in pathological beliefs and/or behavior.

On the other hand, 'computational' also refers to the tools CP uses in order to understand this problem. It tries to build and formalize simplified, mathematical models of the mechanisms underlying cognition, behavior or physiology and their possible perturbation in psychiatric disorders (Frässle, Yao et al. 2018). In other words, its main goal is to use computational tools to gain an aetiological understanding of psychiatric diseases, moving away from a purely symptom based description. This new level of description of disease processes holds promise to yield more credible predictions of clinical trajectory and allow for differential treatment (e.g. (Montague, Dolan et al. 2012, Stephan and Mathys 2014, Wang and Krystal 2014, Huys, Maia et al. 2016)).

One of the most prominent examples of such a description of putative disease processes is the 'dysconnection hypothesis'[2] that has been formulated in the context of schizophrenia. It postulates aberrant interaction between NMDA[3] receptor (NMDAR) dependent plasticity and neuromodulators (NNI) (Friston 1998, Stephan, Baldeweg et al. 2006, Stephan, Friston et al. 2009). According to this theory, the spectrum nature of schizophrenia is a consequence of the variable ways in which genetic-environmental influences can alter NNI, resulting in tremendous variability of clinical trajectories and treatment responses across patients.

Despite clearly formulated hypotheses, CP has yet to deliver a proof of its clinical utility. There are multiple reasons. First, as shown in the example of schizophrenia, the stated hypotheses are inherently a highly multivariate problem. Second, medication affecting different neurotransmitter systems have shown efficacy for certain psychiatric diseases. However, it is unclear why or how exactly they work in the brain as medication is typically applied systemically. Third, connected to this, positron emission tomography (PET) is the only non-invasive brain imaging technique to investigate neurotransmitter systems directly; yet, requires the injection of a positron-emitting radioligand.

---

[1] Here, surprise is meant metaphorically. However, 'surprise' also has a mathematical meaning which is part of many models

[2] The hypothesis was originally named 'disconnection' before being refined into 'dysconnection'.

[3] N-methyl-D-aspartate receptor (NDMAR)

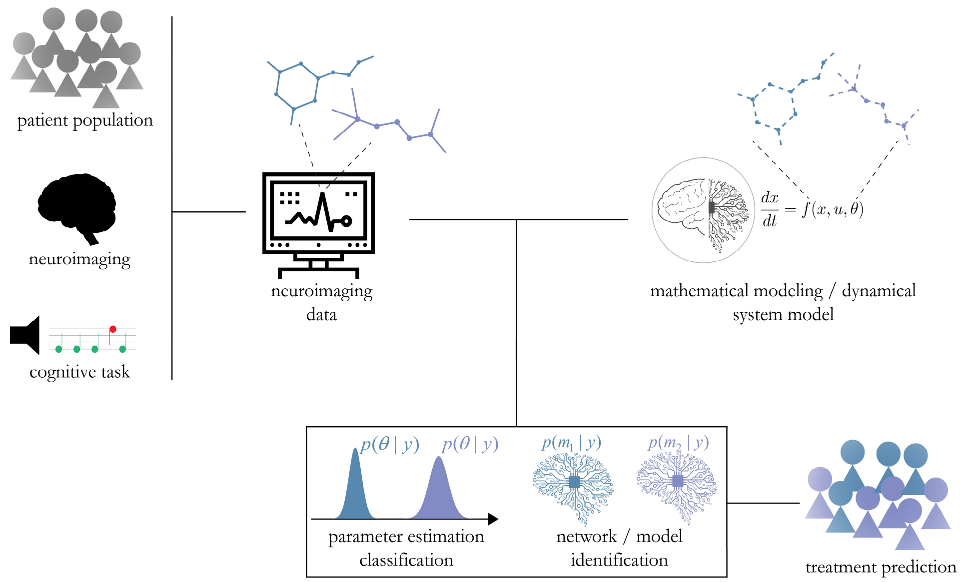

Figure 1| Schematic overview of the generative embedding approach; Mathematical models used to infer on hidden neuronal processes. Patient stratification based on inferred parameter estimates or model selection. Overview in line with (Brodersen, Deserno et al. 2014, Frässle, Yao et al. 2018). Parts of the Figure courtesy of Sara Tomiello.

A possible way to address the investigation of neuromodulatory action using other methods was suggested in the context of modeling brain physiology and connectivity. Here, specific cognitive tasks[4] are used, where evidence for involvement of particular neurotransmitters has been demonstrated. These tasks are combined with non-invasive neuroimaging techniques (electroencephalography (EEG), functional magnetic resonance imaging (fMRI)), eye-tracking or other physiological readouts. The measures of neural activity (or other measures) are then thought to convey information about the neuromodulatory processes (e.g. (Aponte, Schöbi et al. 2019), for review, see (Iglesias, Tomiello et al. 2017, Heinzle and Stephan 2018)). The last step to gain an aetiological understanding of disease mechanisms, is to model this task-based (or unconstrained) activity by means of a dynamical system model. CP then argues that it is at the level of models (or their parameters) where pathophysiologies are encoded. The hypothesis is that patients separate into subgroups based on these models and/or model parameters, for example with respect to treatment. This differentiation can be formulated as a statistical problem of comparing parameter estimates or distributions, or by encoding different hypotheses in different model architectures (Frässle, Yao et al. 2018). The procedures described here have been termed

[4] or unconstrained cognition

'Generative Embedding' (Brodersen, Deserno et al. 2014) and show a clear analogy with what is known in medicine as *differential diagnosis*.

Two tasks that have robustly shown sensitivity and specificity to neuromodulatory influences are the mismatch-negativity (MMN) and working memory (WM) paradigm (Iglesias, Tomiello et al. 2017). Furthermore, both paradigms have been shown to be clinically relevant for schizophrenia as elicited neural responses differ between healthy controls and patients (e.g. (Roth, Pfefferbaum et al. 1981, Park and Gooding 2014)). However, the two tasks are assumed to involve very different neurotransmitter systems. In brief, the MMN refers to characteristic differences in the evoked response potentials (ERPs) to attended and un-attended stimuli. It was suggested that it reflects a prediction error signal updating an internal model of the world. Hence it could resemble learning. While it has been shown that the MMN can be manipulated with drugs affecting the cholinergic transmitter system (Moran, Campo et al. 2013), it seems that dopamine is less involved (Leung, Croft et al. 2009). Dopamine on the other hand has been associated with WM (Sawaguchi and Goldman-Rakic 1991). WM denotes a special type of memory storage with ability to retain and manipulate limited amounts of information over short periods of time (D'Esposito and Postle 2015). Importantly, NMDAR function is expected to be crucial in the generation of persistent activation over time due to their longer time constants. Therefore, WM is an interesting testbed to infer on NMDAR function and potential modulation thereof by dopamine (Durstewitz and Seamans 2002). The empirical studies presented in this thesis focus in more detail on these two tasks.

The mathematical modeling of these task-based data asks for a very special type of models that respect both the nature of the latent (hidden) processes on the neuronal level and the resolution of the imaging technique. One such method, dynamic causal modeling (DCM), was formulated in the early 2000s for fMRI (Friston, Harrison et al. 2003). It models the effective (directed) connectivity between brain regions. The way neuromodulatory action here can be understood is in the enhancement or reduction of connection strengths mediated by neuromodulators. Not much later, DCM were adapted to be used for electrophysiological data (David, Kiebel et al. 2006). Because of the much higher temporal resolution of EEG, synaptic action can be directly incorporated in the models.

Critically, in order for DCMs (and, in fact, any model) to become truly useful for CP, the robustness and reproducibility of these tools is paramount. This is because in CP, reproducibility, generalizability and stability of model based approaches have direct consequences on acceptance (in order to achieve actual, clinical implementation) of the new methods into everyday clinical practice and, ultimately, on patient well-being. This is even more critical given the remarkable complexity of these models and the computational challenges that are associated with that. Despite the importance of the robustness of the approach, so far only little work has been done on systematically testing the robustness of DCMs in a standard setting where it is typically applied to data. Notably, this scenario is by no means exclusive to DCM, but can be observed for a wide range of computational modeling approaches in all fields of science. For instance, according to Henderson et al. (2018), over 10'000 reinforcement learning (RL) related papers have been published yearly between 2012 and 2016 (Henderson, Islam et al. 2018), with those papers displaying a large variability in the exact (deep) RL methods used and the way results are reported. The authors conclude that as a consequence, results and indicated improvements by the respective methods do not easily generalize and that there is an overall lack of homogeneity. On a broader scheme, this critique is ubiquitous across all scientific disciplines as revealed by a current search on google scholar using the terms '*reproducibility*, *crisis*, *science*' yielding over 10'000 results since 2018.

To address this shortcoming in the context of DCM, the goal of this thesis is to conduct a systematic investigation of the robustness of the method and provide a more thorough understanding of potential challenges, as well as to develop practical solutions and recommendations that address these issues when applying these models to real-world data.

In this context, it is useful to briefly emphasize some important characteristics of the models utilized here (i.e., DCMs) – and to point out some important differences compared to classical machine learning (ML) models. First, they are biophysically interpretable models as they grounded in our current knowledge about fundamental physiological processes. Consequently, their parameter estimates promise to convey information about neurophysiology and the data generating processes. Second, the models used in CP are applied to reduce the complexity of the problem and to, potentially, facilitate a mechanistic understanding. While classical ML algorithms can be very good at prediction, they typically operate in orders of magnitude higher dimensional parameter spaces and would provide little to no systematic understanding of a disease. Third, all models we will present are so called generative models which follow Bayesian statistics. That is, the parameters of the models are considered random variables, while data is fixed. Importantly, parameters (θ)

are assigned a priori probabilities (prior distributions $p(\theta)$) and the posterior belief about parameters, conditioned on the data (y), follows Bayes' Rule

$$p(\theta \mid y) = \frac{p(y \mid \theta) \cdot p(\theta)}{p(y)}. \tag{1.1}$$

There is a lot of debate on the function and specification of prior distributions, and generally between the Bayesian approach and the frequentist counterpart. When I took a course on Bayesian statistics in 2017, the lecturer's response to the dispute was (and I am paraphrasing):

> "… it is not about superiority of Bayesian over frequentist statistics. It is about understanding their relationship, making sensible choices and trying to get the best out of both worlds. "

I think this view is very useful and much of this thesis is devoted to getting a quantitative understanding of the overarching theoretical framework that is paving the way towards using models as computational assays for differential diagnosis, and their explicit implications for DCMs. Regardless of whether the models are Bayesian or not, the idea for them to provide physiologically interpretable meaning comes at a cost. Face, construct and predictive validity pose additional demands on them to be truly able to achieve their goal. All models are subject to decisions of the modeler that might affect all three previous demands (e.g. prior distributions). As we will see, there are other crucial factors that challenge construct validity, robustness and reproducibility. These include assumptions about the independence of datapoints, correlations between parameters, assumptions about the irreducible variance in the data and, very often, small sample sizes.

1.2 THE STORYLINE

In this dissertation, we address the problems laid out above by first considering theoretical and computational aspects of the modeling and model inversion and then applying these new developments in the context of two experimental datasets.

In Chapter 2, we look at the integration of delay differential equations underlying DCM for EEG. Including delays between different neuronal populations assures that the model respects the physiologically known conductance delays between cortical regions, which cannot be neglected given the high sampling rate of EEG. The delays can change a system's dynamic repertoire. The cost of this strive for realism is that delay differential equations are much more difficult to integrate compared to their non-delayed counterpart. The first part of the methodological section benchmarks the current integration scheme for these models against two, in-house developed variants and a state of the art scheme (based on a Runge-Kutta scheme) in a number of simple systems. We show that the default implementation violates construct validity both in terms of causality and dynamics, whereas they appear to be respected in our alternative procedures. While accuracy is arguably the most important criterion for assessing the performance of the integration schemes, computational efficiency is a critical factor as well.

In Chapter 3, we investigate the problem of optimization in DCMs. Already in 2011, Daunizeau and colleagues mentioned some critical aspects of the statistical inference techniques (Daunizeau, David et al. 2011). Some of these points have been investigated, but often at the cost of much more computationally costly methods (Penny and Sengupta 2016, Sengupta, Friston et al. 2016)[5]. Most crucially, network identification and parameter estimations hinge on an optimal (in a Bayesian sense) computation of the posterior distribution. It itself however, depends on the optimization landscape (we will derive the dependency in the chapter). For non-convex optimization problems, the local optimization routine based on a gradient ascent routinely implemented in DCM is not guaranteed to find the globally optimal solution. If inversions end in a local optimum, objective model assessment criteria derived from Bayesian theory therefore might not be applicable. Local extrema also pose replicability problems, as results will depend on particular starting values of the optimization and local gradients. In this second part, we demonstrate that the objective function (the negative free energy) for the case of our simulations does present with critical structures which, from the algorithm's perspective are identified as local optima. To account for this problem, we augment the standard inversion approach by starting multiple inversions in parallel from different starting values. They allow us to diagnose and quantify the problem of local optima both in terms of resulting posterior distributions and the score of model goodness. We assess model and parameter recovery and compare as to what degree conclusions drawn are dependent on the starting values. The presented augmentation of the inversion helps to overcome some of the caveats, while it remains easy to implement, building on the same (default) optimization framework. But it also illustrates that further constraints are needed, if meaningful homogeneity in

[5] Unfortunately, to our best of knowledge, these works have not been followed up.

parameter estimates and comparability of effects need to be achieved across studies in the future. Our results also provide a possible explanation for why studies may find seemingly conflicting results if standard procedures are used.

In Chapter 4 we focus on the criteria of model goodness. We derive approximate relationships between a measure of model goodness that can be interpreted in absolute values, i.e. residual sum of squares and an objective Bayesian measure of model goodness, the negative free energy. These relationships provide an explanation for many surprising quantitative findings in the literature. First, they explain why empirical evidence favoring one model over another can vastly exceed the standard thresholds. Second, they illustrate how false assumptions about the (in-) dependence of residuals can cause tendencies to overfit and bias towards complex models. Third, they illustrate that hyperprior specification deserves special care, as misspecification can lead to biased inference on models simply driven by these a priori assumptions. The conclusion from these findings will be the following: While one can argue that inference is always dependent on modelling assumptions, our results demonstrate that the field should be aware of the severe consequences some of them might have in practice and, if possible, agree on standards in terms of data-preprocessing, model assumptions etc. in order to make results comparable across studies and robust for single patient predictions.

These three chapters conclude the theoretical and model development part of this thesis. The software for the integration scheme, the augmentation and diagnostics of the inversion procedure and the diagnostics of hyperparameter optimization will be made publicly available in the in-house developed, open source software collection TAPAS (https://www.tnu.ethz.ch/en/software/tapas.html).

The second contribution of this thesis is the application of these models to empirical data, both in EEG and fMRI, capitalizing on the lessons we learned in the first part of this thesis and the developed improvements of the DCM machinery. As highlighted above, the empirical analyses will concern MMN and working memory paradigms, and thus have a strong link to neuromodulatory processes and clinical questions.

In Chapter 5, we first show the application of DCM for electrophysiological measures. We use DCM for evoked response potentials (ERPs) to infer on graded, muscarinic changes of effective connectivity in rodents in an MMN paradigm. Here, we make explicit use of the first two methodological advances and show that DCM can successfully reduce the number of features by a factor of 50. The ensuing parameter estimates not only have physiological interpretation, but also selectively convey information about the pharmacological setting. Throughout this thesis, we will refer to this study as *RATMPI*.

The second empirical contribution (Chapter 6) is the application of DCM for fMRI in a working memory paradigm. The paradigm was developed during this PhD and allows for online titration of task difficulty – a crucial confound when it comes to bringing WM to a clinical setting. Here, we will make use of the methodological developments two and three, and propose a way how informed hyperpriors can be specified in DCM for fMRI. We will refer to this study as *PRSSI*.

The final Chapter 7 of this thesis will provide a discussion that connects the findings of the methodological and data analysis chapters and provides an outlook.

Notably, the work presented in this thesis (and briefly outlined above) represent only a selection of the projects I contributed to over the course of my PhD. In particular, I was further involved in two additional studies. First, I was the main contributor to an EEG study with 162 healthy controls, where participants performed the mentioned Working Memory task under pharmacological intervention. In a double blind, placebo-controlled, between subject design, participants underwent a cholinergic or dopaminergic intervention with either agonistic or antagonistic effect. Second, I was main contributor in a longitudinal patient study with patients suffering from psychotic symptoms. Here, the goal will eventually be to predict the success of a treatment switch from a non-cholinergic to a cholinergic anti-psychotic, based on WM EEG data. The analysis of these datasets is subject to future work, and lies beyond the scope of this dissertation.

In the rest of this introduction, we will provide a brief introduction the generative model of DCM and how hypotheses can be tested in a Bayesian framework. They will later be used for reference. We here focus on DCM for ERP since these represent the main workhorse utilized in this thesis. The present summary is thought as a concise summary of the mathematical literature associated to DCM for ERP and we refer to the original work (David, Kiebel et al. 2006) or this review paper on the mathematical machinery (Ostwald and Starke 2016) for an in-depth treatment of DCM for ERPs. Furthermore, a comprehensive treatment of model comparison in DCMs can be found in (Penny 2012).

1.3 INTRODUCTION TO DYNAMIC CAUSAL MODELING FOR ERP

1.3.1 ARCHITECTURE

DCM for ERP assumes that the data measured in EEG is generated by neural activity in one or multiple connected cortical columns. Each column typically comprises three different kinds of cell populations: *excitatory interneurons (stellate cells), inhibitory interneurons (inhibitory cells)* and *(excitatory) pyramidal neurons (pyramidal cells)*[6]. Populations can be understood as modeling population dynamics, i.e. postsynaptic potential (probability-) distributions over a population of single cells. If the distribution activity is simplified and only described by its mean, one is operating in the *Neural Mass* approximation. In the following, we will always use this assumption. Populations within a cortical column are connected via *intrinsic* connections. Between columns, populations are *extrinsically* connected. Extrinsic connections show a well-defined layer specificity which distinguishes forward and backward connections in the cortical hierarchy based on the seminal work by (Van Essen, Anderson et al. 1992). **Figure 2** illustrates a simple neural mass model which dates back to the work of (Jansen and Rit 1995). This allows for modeling processes along cortical hierarchies, which is fundamental to test hypotheses about information-processing and learning.

[6] Over the years, multiple variants of DCM for EEG have been formulated. In light of the relevance for this thesis, we limit our discussion to the original convolutions based version, including three subpopulations David, O., S. J. Kiebel, L. M. Harrison, J. Mattout, J. M. Kilner and K. J. Friston (2006). "Dynamic causal modeling of evoked responses in EEG and MEG." Neuroimage **30**(4): 1255-1272.. We used this variant for most of the methodological developments. In the empirical RATMPI study, we used a different architecture of the single, cortical column; the so called Canonical Microcircuit (CMC, Bastos, A. M., W. M. Usrey, R. A. Adams, G. R. Mangun, P. Fries and K. J. Friston (2012). "Canonical microcircuits for predictive coding." Neuron **76**(4): 695-711.). The details of this circuit will be explained in the corresponding chapter.

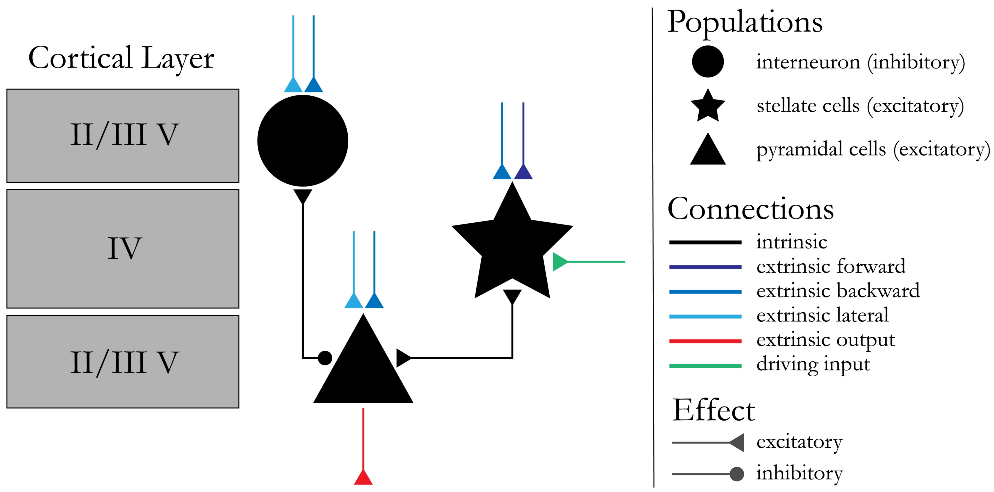

Figure 2 | Schematic overview of a single cortical column. Architecture based on the original, three population model. Outgoing connections (to other columns) always originate from the pyramidal cell. Both, interneurons and pyramidal cells are thought to populate supra-granular layer II/III and infra-granular layer V.

1.3.2 Neuronal Model

There are two fundamental mathematical operations that give rise to the generic equations underlying the convolution based DCM formalism and hence the neuronal dynamics. First, average post-synaptic population potentials are being modeled. They are thought to arise from average incoming activity of other populations. This average is formulated as a convolution ($\otimes$) between a parametrized convolution kernel (H) and the incoming firing of another populations (σ). The convolution kernel here describes a cell-specific property of how unit-firing changes the post-synaptic potential, parametrized by a gain h and a (inverse) decay parameter κ. Hence, one could think of it as a summary of all channel properties onto one population. Second, average post-synaptic potentials lead to average cell-firing at time t, i.e. $\sigma = \sigma(v(t))$. This is formulated via a sigmoid-transformation (Jansen and Rit 1995). For a single cell subject to incoming firing, the post-synaptic potential is described by

$$\begin{aligned} v(t) &= H(t) \otimes \sigma(t) = \int_{-\infty}^{t} H(t-\tau)\sigma(\tau)d\tau \\ H(t) &= h \cdot \kappa \cdot t \cdot \exp(-\kappa t), \quad \forall t > 0 \end{aligned} \tag{1.2}$$

We now derive the general form of the equations by taking the derivative wrt. t:

$$\begin{aligned}\dot{v}(t) &= h(0)\sigma(t) + \int_{-\infty}^{t} H\kappa e^{-\kappa(t-\tau)}\sigma(\tau)d\tau - \kappa \cdot v(t) \\ &= \int_{-\infty}^{t} H\kappa e^{-\kappa(t-\tau)}\sigma(\tau)d\tau - \kappa \cdot v(t)\end{aligned} \tag{1.3}$$

where we made use of the *Leibniz rule* in the first step, and the fact that $h(0) = 0$.

Taking the second order derivative yields

$$\begin{aligned}\ddot{v}(t) &= H\kappa e^{-\kappa(t-t)}\sigma(t) - \int_{-\infty}^{t} H\kappa^2 e^{-\kappa(t-\tau)}\sigma(\tau)d\tau - \kappa\dot{v}(t)d\tau \\ &= H\sigma(t) - \int_{-\infty}^{t} H\kappa^2 e^{-\kappa(t-\tau)}\sigma(\tau)d\tau - \kappa\dot{v}(t)d\tau.\end{aligned} \tag{1.4}$$

Multiplying Eq. (1.3) with κ yields

$$\int_{-\infty}^{t} H\kappa^2 e^{-\kappa(t-\tau)}\sigma(\tau)d\tau = \kappa\dot{v}(t) + \kappa^2 v(t). \tag{1.5}$$

which can be inserted in Eq. (1.4) to result in the 2nd order differential equation underlying the convolution based DCMs:

$$\ddot{v}(t) = H\kappa\sigma(t) - 2\kappa\dot{v}(t) - \kappa^2 v(t). \tag{1.6}$$

Note that this equation is equivalent to the dynamic equations of a driven, critically dampened, harmonic oscillator. Eq. (1.6) describes the average, post-synaptic potential of a single population of cells for a single, arbitrary input $\sigma(t)$. For multiple inputs, this can be trivially extended to

$$\ddot{v}_l(t) = \left[\sum_{j=1}^{n} \Gamma_{lj} H_{lj} \kappa_{lj} \sigma_j(t)\right] - 2\kappa_{lj}\dot{v}_l(t) - \kappa^2{}_{lj} v_i(t). \tag{1.7}$$

Here, we already index that we have selected an arbitrary population (l) that is connected to multiple other populations (j). The connections here are encoded by an indicator matrix Γ_{lj}, which is $\Gamma_{lj} = 1$ if two populations are connected and $\Gamma_{lj} = 0$ otherwise. This matrix can be directly read out from the connectivity pattern in **Figure 2**. Here, we basically state that each connection is modeled through its own kernel (i.e. channel properties), parametrized by the kernel gain H_{lj} and inverse kernel decay κ_{lj}. However, a general problem of all models is to somehow evoke parameter constraints. Otherwise, there are simply too many inherent correlations and parameters cannot be estimated without large uncertainties in the estimates. For example, for the *intrisic* connectivity pattern of a single cortical column shown in **Figure 2**, without constraint, the connection specific kernels would already amount to eight parameters per cortical column.

Therefore, usually three simplifying assumptions are made:

- (inverse) kernel decays κ are region and effect specific (whether a connection is excitatory (e) or inhibitory (i)):

$$\kappa_{lj} = \kappa_r^{e/i}, \quad r = \{1, 2, ..., R\}$$

- kernel gains H are region, effect and connection specific

$$H_{lj} = \begin{cases} H_r^{e/i} \cdot A_{lj}, & \text{for extrinsic connections} \\ H_r^{e/i} \cdot \gamma_{lj}, & \text{for intrinsic connections} \end{cases}$$

This brings the notion of *connections strengths* A_{ij}, γ_{ij} which are factors enhancing or depleting the region specific baseline gain.

- The intrinsic connection strengths γ_{ij} are fixed across regions, i.e. specific intrinsic connections between the same populations have the same strength across regions

$$\gamma_{ij} = \gamma_p, \quad p = \{1, 2, 3, 4\}.$$

Together, these simplifications result in the postsynaptic potential population i in region r to obey the following differential equation:

$$\ddot{v}_{rl}(t) = \left[\sum_{j=1}^{n} H_r^{e/i} \kappa_r^{e/i} (A_{lj} + \gamma_{lj}) \sigma_j(t) \right] - 2\kappa_r^{e/i} \dot{v}_l(t) - \kappa^2{}_r^{e/i} v_l(t). \tag{1.8}$$

Here, the indicator-matrix Γ is absorbed directly in the variables A and γ. Under these assumptions and considering for example four cortical columns, the number of synaptic parameters over the two cortical columns are 20 instead of 32 (not taking into account extrinsic connections).

For the system described in Eq. (1.8), initial conditions of $v(t \le 0) = 0$ is a fixed point, hence the system will always stay at rest. An external, driving stimulus then perturbs the system away from this equilibrium position. Typically, one assumes that it is some pulse shaped, thalamic input u caused by some external stimulation (e.g. a brief tone). With the arguments so far, this can be directly added in Eq. (1.8).

$$\ddot{v}_{rl}(t) = \left[\sum_{j=1}^{n} H_r^{\,e/i} \kappa_r^{\,e/i} \left[(A_{lj} + \gamma_{lj}) \sigma_j(t) + C_l u_l(t) \right] \right] - 2\kappa_r^{\,e/i} \dot{v}_l(t) - \kappa^2{}_r^{\,e/i} v_l(t). \tag{1.9}$$

External stimulations are always thought to excite the populations in the granular layer IV as can be read out from the connectivity pattern in **Figure 2**. Therefore, $C_l = 0$ if l does not correspond to the index of a stellate cell. Moreover, there might be multiple different inputs driving the system, and we therefore also index u. An alternative interpretation would be to consider u as an additional, non-interacting state whose output does not undergo a sigmoid transformation. This would be congruent with Eq. (1.9) and gives it the same notion of some un-observable (thalamic) state, with C enhancing/depleting the baseline synaptic gain, equal to the interpretation of A and γ.

Finally, we allow that one population might exert its effect over another population with some temporal delay τ. In other words, the post-synaptic potential changes with respect to some earlier, average firing, hence

$$\ddot{v}_{rl}(t) = \left[\sum_{j=1}^{n} H_r^{\,e/i} \kappa_r^{\,e/i} \left[(A_{lj} + \gamma_{lj}) \sigma_j(t - \tau_{lj}) + C_l u_l(t) \right] \right] - 2\kappa_r^{\,e/i} \dot{v}_l(t) - \kappa^2{}_r^{\,e/i} v_l(t). \tag{1.10}$$

These delays render the ordinary differential equations in Eq. (1.9) to become delay differential equations (Eq. (1.10)), which has some important implications on how the system needs to be integrated. We will leave this point for the respective Chapter 2. As mentioned in the footnote, there are other assumptions or constraints that one could make on the system, which will give rise to other variants of the model. Here, we merely aimed at outlining the origin of the equations for the convolution based DCM, based on the ERP architecture with three subpopulations, as implemented in Statistical Parametric Mapping (SPM12, ver. 6906) software package (www.fil.ion.ucl.ac.uk/spm).

In the literature, state variables (here $v(t)$) are often referred to as x and Eq. (1.10) is expanded into a system of 1st order differential equations. If we denote

$$\begin{aligned} &\dot{x}_l(t) = x_{l-1}(t) \\ &\ddot{x}_l(t) = \dot{x}_{l-1}(t) = \left[\sum_{j=1}^{2n} H_r^{\,e/i} \kappa_r^{\,e/i} \left[(A_{lj} + \gamma_{lj}) \sigma_j(t - \tau_{lj}) + C_l u_l(t) \right] \right] - 2\kappa_r^{\,e/i} x_{l-1}(t) - \kappa^2{}_r^{\,e/i} x_l(t) \\ &\quad l = 1, \dots, 2n \end{aligned} \tag{1.11}$$

the set of differential equations in Eq. (1.11) is of order 1, but models the same dynamics as Eq. (1.10). Here, n denotes the total number of populations (across all cortical columns).

From the perspective of electrical circuits, the auxiliary states can be interpreted as currents flowing through the cell membrane (the change of voltage of a capacitor equals the current). It should be noted that the actual implementation extends the states associated to pyramidal cells into hyper- and depolarizing states. However, for our brief review, we will not go into this level of detail.

1.3.3 FORWARD MODEL

In order to derive a full generative model that can generate EEG data, a forward mapping from latent states described by the dynamical system above to observed variables is necessary. This is needed to assign the likelihood of the data, given the prediction and parameters. We will see in Chapter 3, how this can then be turned into an objective function that trades off accuracy and complexity with respect to which parameters are optimized. For electrophysiological data, the observed data lives either in the space of principal components of scalp potentials, or epidural/local field potentials (LFP). For simplicity, we will focus on the latter, also because in this thesis, no scalp data was analyzed but only LFPs. In this case, the number of channels (n_c) matches the number of cortical columns. We assume that the measured LFP potentials (measured in the unit of volts [V]) come from a linear mixture of the average post-synaptic population potentials. Put simply

$$\begin{gathered} y_p(t) = G \cdot x(t) \\ t \in pst \\ y_p \in M^{n_c \times pst} \\ x \in M^{2n \times pst} \end{gathered} \tag{1.12}$$

where we explicitly indicate that both, the predicted data y_p and the hidden states x are matrices. Here, $x(t)$ are the integrated states over peri-stimulus time (pst) of Eq. (1.11). Hence, G is a matrix, mapping from the latent states to the predicted observation: $G \in M^{nc \times 2n}$. There are some constraints on the entries of this so-called lead field matrix G:

- Only voltage states (x_i in Eq. (1.11)) contribute to the LFP potentials
- There is no mixture between the voltage of a single cortical columns and other channels
- Each population exerts equal gain J_P (across cortical columns)

- Each cortical column (i) exerts a region (i) specific gain L_i (equal across populations of this column)

Collectively, this amounts to

$$G_{ij} = \begin{cases} 0 & \text{for } j \text{ not being a voltage state} \\ 0 & \text{for } j \text{ not being a voltage state not associated with channel } i \\ L_i \cdot J_p & \text{for } j \text{ being the post-synaptic potential of a population } p \text{ in column } i \end{cases} \tag{1.13}$$

In this mathematical formulation (depending on how the states are exactly concatenated), G has a block-diagonal form.

1.3.4 Interactive Simulations

In order to make DCM more accessible, we have created a didactic, interactive visualization tool (**Figure 3**). It allows to dynamically change parameters and visualizes changes in the predicted responses at the hidden neuronal level and changes in the observed responses. It is MATLAB based, ties into the existing functionality of SPM, and will be released in one of the upcoming TAPAS releases in 2020.

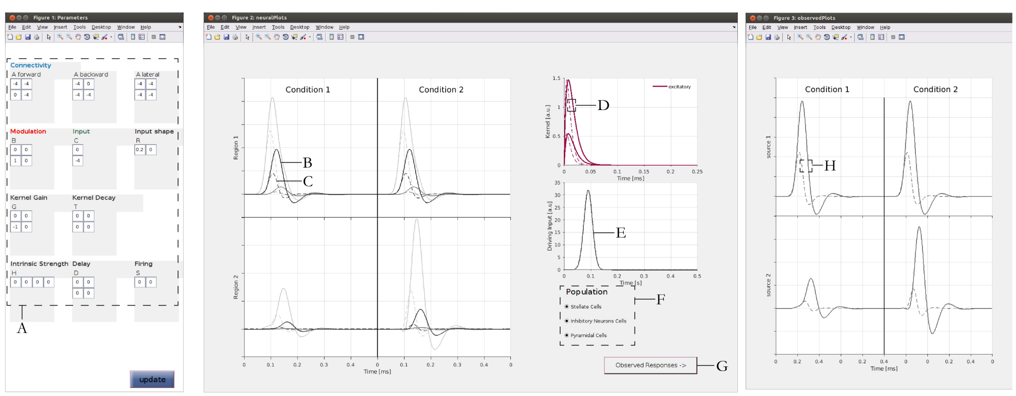

Figure 3| Interactive user-interface for the simulation of signals in the convolution based DCM for ERP framework. A) Interactively change parameter setting to visualize changes in the signal. B) Responses of the neuronal populations for the current parameter set. C) Responses of the neuronal populations for the previous parameter set. D) Convolution kernel for the current and previous parameter set. E) Driving Input for the current and previous parameter set. F) Toggle to display only a subset of populations. G) Toggle to activate the 'observed-responses-panel' (right panel). H) observed responses for the current and previous parameter set.

1.3.5 Generative Model

The generative model then entails the probabilistic information about the data-generating process under assumption of noise. We assign a Gaussian likelihood to datapoints that is, the probability of a time-series y given a prediction y_p plus some error, which we assume to be distributed according to a Gaussian with some parametrized error covariance $\Sigma(\theta_h)$, is given by

$$y \mid yp \sim N(0, \Sigma(\theta_h)). \tag{1.14}$$

We will go in detail into this part in Chapter 3. Hence, the full joint probability of data and parameters can be written as

$$\begin{aligned} p(y, \theta_h, \theta_g, \theta_p) &= p(y \mid \theta_h, \theta_g, \theta_p) \cdot p(\theta_h) \cdot p(\theta_g) \cdot p(\theta_p) \\ &= N(y - G(\theta_g) \cdot x(\theta_p) \mid \Sigma(\theta_h)) \cdot p(\theta_h) \cdot p(\theta_g) \cdot p(\theta_p). \end{aligned} \tag{1.15}$$

For completeness, we have written down all dependencies of the variables in Eq. (1.15). N denotes a multivariate normal distribution. We assume a priori independence between the parameter classes, i.e. *neuronal* parameters ($\theta_p = \{H, \kappa, \gamma, C, A, \tau\}$), parameters of the *forward* model ($\theta_g = \{L, J\}$) and *hyperparameters* of the noise model (θ_h). This notion of the parameter classes and the notation will be used throughout the methodological part of this thesis.

1.4 MODEL EVIDENCE AND MODEL COMPARISON

In a Bayesian setting, the goodness of a model m can be understood in terms of the model evidence or the marginal likelihood of the data y under the prior $p(\theta \mid m)$[7]:

$$p(y \mid m) = \int p(y \mid \theta, m) p(\theta \mid m) d\theta. \tag{1.16}$$

Put simply, it denotes the likelihood of the observed data under the prior. Comparing two competing models in a Bayesian setting amounts to computing the ratio over model evidences (or the difference in log-model evidences):

---

[7] Note that in our notation, we indicate explicitly that the parameters θ come from model m. Alternatively, one could write $p(y \mid \theta_m)$ or just omit the index and keep the dependency in mind.

$$B_{12} = \frac{p(y \mid m_1)}{p(y \mid m_2)}$$
$$\log B_{12} = \log p(y \mid m_1) - \log p(y \mid m_2). \tag{1.17}$$

These quantities are known as *Bayes' Factor* and *Log Bayes' Factor* (Kass and Raftery 1995). This can be seen as a Bayesian way of hypothesis testing. For example, in the context of DCM, m_1 and m_2 could encode different hypotheses about how the network changes under some perturbing stimulation. To draw the link to the empirical studies presented in this thesis, these 'perturbations' or 'condition specific effects' could either be the presentation of an unexpected tone (MMN) or the need to retain information in working memory (WM). Importantly, Kass and Raftery (based on (Jeffreys 1961)) also provided guidelines on interpreting the Bayes' Factor in a quantitative manner. According to those, a Bayes' Factor of

$$B_{12} > 20$$
$$\log B_{12} > 3$$

speak for *strong evidence* in favor of m_1 over m_2 (Kass and Raftery 1995). The log of the quantity in Eq. (1.16) (the log model evidence, LME) has an interesting property:

$$\begin{aligned}
\log p(y) &= \int \log p(y) p(\theta \mid y) d\theta \\
&= \int \log \frac{p(y \mid \theta) p(\theta)}{p(\theta \mid y)} p(\theta \mid y) d\theta \\
&= \underbrace{\int \log p(y \mid \theta) p(\theta \mid y) d\theta}_{Accuracy} - \underbrace{\int \log \frac{p(\theta \mid y)}{p(\theta)} p(\theta \mid y) d\theta}_{Complexity}.
\end{aligned} \tag{1.18}$$

Accuracy is the expected log-likelihood of the data under the posterior. *Complexity* is the Kullback-Leibler-Divergence between the posterior and prior distribution, and a measure of similarity. By definition, it is positive (semi-) definite and only equal to zero if the two distributions are identical. Therefore, it can be seen as penalizing overly complex models and formalizes the principle of Occam's razor (Beal 2003, Penny 2012). This property becomes important when comparing two models. Specifically, the decision which of two models provides a better explanation for the data depends not only on how well it predicts the data, but rather on a full accuracy-complexity-tradeoff.

Obviously, the tradeoff hinges on the correct computation of the model evidence. As we will see, this is a computationally challenging problem, as computing the integral in Eq.

(1.16) is usually not feasible. In Chapter 3, we will show that one can approximate the LME and turn the problem of computing it into a problem of optimization. This optimization not only returns a value for the model goodness, but also an approximation to the posterior distribution of the parameters.

This concludes the introduction to Dynamic Causal Modeling for ERPs and the introduction to this dissertation.

2| INTEGRATION OF DELAY DIFFERENTIAL EQUATIONS

2.1 DECLARATION

These methodological developments and analyses were done under the supervision of Jakob Heinzle and Klaas Enno Stephan. Data collection for the empirical analysis was done by the Max-Planck Institute in Cologne by Fabienne Jung as part of the doctoral thesis. This dataset was previously used in a publication (Jung, Stephan et al. 2013). For more information on the dataset, we refer to Chapter 5.

2.2 INTRODUCTION

In DCM for EEG, neural dynamics are described in terms of delay differential equations (DDEs, see Eq. (1.11)). Delays refer to temporally delayed effects one population exerts over another. In the convolution based DCM framework, delays come in two flavors – extrinsic and intrinsic delays, which however are merely associated with differences in the a priori expectation of their magnitude. They both describe delays between the output of a population, i.e. the sigmoid transformed post-synaptic potentials and the receiving (input) population. These directed dependencies are easily accessible when looking at the full connectivity pattern of a network (see **Figure 2**). The biophysical reason for delays is simple – action potentials travel at a finite velocity through axons, with large variation in speed depending on multiple factors (e.g. myelination or axon diameter). Reported velocities range from 0.1 up to 120 m per second (Swadlow and Waxman 2012). For cortico-cortical connections (which are most relevant for EEG), animal studies have reported delays between 0.5 ms and 42 ms[8] (Miller 1975, Swadlow 1990, Ferraina, Paré et al. 2002), but considering relays of activity through non-modelled sources, one could also think of longer delays, especially, since axonal connections in humans tend to be longer.

Delays render ordinary differential equations (ODEs) to become delay differential equations (DDEs). We will not go into too much detail about the general application of delay differential equations, and reduce the general overview to the relevant features for the discussion of this chapter. The following overview constitutes a brief summary of chapters 1-3 of a basic text book on "Numerical methods for delay differential equations" (Bellen and Zennaro 2013).

In general, 1[st] order ODEs are characterized by the following set of equations

$$\begin{aligned} \dot{x}(t) &= f(t, x(t)), t_0 < t < t_f \\ x(t_0) &= x_0 \end{aligned}$$

where x denotes the state variable or state, t denotes time, t_0 and t_f the initial and end timepoint and x_0 the initial value (hence, also the commonly used term *initial value problem*). The step to DDEs then includes dependencies of the previous equation on past values of x

$$\begin{aligned} \dot{x}(t) &= f(t, x(t-\tau)), t_0 <= t < t_f \\ x(t) &= \phi(t), t < t_0 \end{aligned} \tag{2.1}$$

with $\tau \in [-r, 0], r \in [0, \infty)$ now being semi-positive delays, and ϕ an initial state function. This initial state function renders the solution un-smooth at the transition point, as there (generally) is a jump in the derivative. Such a discontinuity propagates throughout the integration interval, and as a consequence, $x(t)$ is only C^1-continuous on the interval $[t_0, t_f]$ (Bellen and Zennaro 2013).

In the first chapter of their book, (Bellen and Zennaro 2013) provide a number of examples illustrating implications for the behavior of a system when introducing delays. Most prominently, delays can stabilize or destabilize a system, can lead to non-uniqueness of solutions, exhibit oscillatory or chaotic behavior, etc. when compared to the sibling ODE

[8] Delays up to 42 ms is arguably long, even for humans. Miller et. al (1975) do report such a latency in cat cortex, but most of the mass in the histogram is around 10-20 ms.

system. Collectively, these difficulties increase the demand on the integration scheme. Particular integration schemes can be ill-defined to capture the nature of the DDE, and theoretical error bounds might not apply anymore. They conclude the chapter with the following statement:

> "…integration of DDEs cannot be based on the mere adaptation of some standard ODE code to the presence of delayed terms. Integration of DDEs actually requires the use of specifically designed methods, according to the nature of the equation and the behavior of the solution."

Early methods to deal with simple, state independent delay problems, used a fine-grained integration mesh, such that all delayed states would fall on grid-points. An integration step can then be computed via a classical ODE step, e.g. the forward Euler method (Elsgolts 1964). (Feldstein and Goodman 1973) introduced a method, where the grid-points are independent of the delays. He combined an Euler step, with a linear interpolation between grid-points to approximate delayed states (Equation 3.1.3 in (Bellen and Zennaro 2013):

$$\begin{aligned} x_{n+1} &= x_n + hf(t_n, x_n, z_n) \\ x_0 &= x(t_0), \\ z_n &= (1-\tilde{\tau})x_{q(n)} + \tilde{\tau}x_{q(n)+1}. \end{aligned} \tag{2.2}$$

Here $q(n)$ and $q(n)+1$ code for the timepoints immediately succeeding and preceding the delayed time and $\tilde{\tau}$ the delayed time normalized to this interval. Hence, z_n is simply the piecewise linear approximation to $y(t-\tau)$. (We simplified the notation for consistency with the notation in out chapter and **Figure 4** provides a visualization of the same concept). Because of the freedom in the choice of interpolation and the discrete ODE step, the previous equations are merely an example of a whole class of DDE integration schemes, the so called *continuous extension* of ODE methods. Another method that circumvents the problem of interpolations has been proposed by the *method of steps* (Bellman 1961). Here, the DDE in Eq. (2.1) can, over some time intervals defined by τ, be reformulated as a system of ODEs and solved step-wise. As one progresses over the integration interval, the dimensionality of the ODE systems increases, but since the global error bounds of ODE methods apply, the method also performs of that order (Bellen and Zennaro 2013). These three example are of course not a comprehensive list of DDE integration methods, but more discussion would go beyond the scope of the very specific problems of this thesis, and we refer to the literature for a deeper background (e.g. (Balachandran, Kalmár-Nagy et al. 2009, Erneux 2009, Smith 2011, Bellen and Zennaro 2013)).

In this chapter, we first briefly summarize the current default integration scheme for DCM for ERPs used in SPM12, 6906, by sketching the derivation of the integration scheme and discussing potential shortcomings. Because of their intriguing simplicity, we implemented two variants of the previously introduced continuous extension of ODE methods and compared them to the SPM standard scheme. Conscious of the potential non-applicability of the proposed schemes, we benchmarked all integration schemes to a standard, MATLAB based scheme for DDEs in a number of simple systems. Finally, we compared the performance of these integration schemes in an empirical rodent dataset, where we applied DCM for ERPs.

The reason why we give the integration scheme particular attention, is that the qualitative nature of the effects of delays can have (potentially strong) implications for parameter estimation. Put simply – if one tried to model the dynamics of a true, delayed physical system, and the model does not adequately represent said delays (through the integration procedure), different parameters might be falsely attributed to fit certain features of the system, which are actually effects of the delays, or vice versa. Hence, one could, in a broader sense, think of the integration scheme as part of the choice of model, which the inferred parameters ultimately depend on. And as with every model, construct validity is vital. In a clinical setting, where we would like to propose that the inferred parameters have some meaning about underlying pathophysiologies and additionally represent biophysical processes, an inaccurate integration procedure could lead to a serious confound.

2.3 METHODS

2.3.1 LOCAL LINEARIZATION DELAYED INTEGRATION

In SPM12 (ver. 6906), the commonly used integration scheme (spm_int_L) for the integration of dynamical systems is based on developments by (Ozaki 1992). The scheme assumes that the system is locally (in time) linear, and the integration can be efficiently and robustly performed through an adaptive step size incorporated by the Jacobian

$$x(t+dt) = x(t) + [\exp(J(x(t))dt) - I]J(x(t))^{-1} f(x(t)) \tag{2.3}$$

(The derivation of this update is provided in Ozaki's 1992 paper, but it's easy to see that it follows from Eq. (2.5) and Eq. (2.7) with $D = 0$, see below). This equation in itself does not incorporate delays yet. The proposed solution of spm_int_L implements delays efficiently but approximately in the Jacobian, which is motivated from a second linearization of the following form:

$$
\begin{aligned}
&\dot{x}(t) = f(x(t-\tau)) \approx f(x(t)) - \frac{df}{dx}\dot{x}(t)\tau \\
&\dot{x}(t) \approx (I + J\tau)^{-1} f(x(t))
\end{aligned}
\tag{2.4}
$$

We will refer to this integrator in the following as local linearization delayed integration *LLDI-scheme* or spm_int_L (the name of the function implementing this scheme in SPM12). For higher dimensional cases, the constant delay τ turns into a delay matrix D. We will give a quick sketch of how the final update equation is then derived (and drop the arguments in the following way for readability: $J(x(t)) = J, f(x(t)) = f(x)$.

Define,

$$
Q = (1 + D \cdot J) \tag{2.5}
$$

and use the Taylor-expansion of $x(t + \Delta t)$:

$$
x(t+\Delta t) = \sum_{k=0} \frac{\Delta t^k}{k!} \frac{d^{(k)} x(t)}{dt^{(k)}} \tag{2.6}
$$

Using Eq. (2.4) and Eq. (2.5) we can replace

$$
\frac{dx}{dt} \approx Q^{-1} f(x)
$$

and by neglecting all higher order derivatives and using the chain rule, we find

$$
\begin{aligned}
&\frac{d^k x}{dt^k} = \frac{d^{k-1}}{dt^{k-1}}(Q^{-1} f(x)) = \frac{d^{k-2}}{dt^{k-2}}(Q^{-1} \frac{dx}{dt}\frac{df}{dx}) \\
&= \frac{d^{k-2}}{dt^{k-2}}(Q^{-1} Q^{-1} J \cdot f(x)) \\
&= \ldots \\
&= Q^{-k} J^{k-1} f(x), \quad \forall k > 0
\end{aligned}
$$

Inserting this into Eq. (2.6) yields

$$
\begin{aligned}
x(t+\Delta t) &= x(t) + \sum_{k=1} \frac{(\Delta t)^k}{k!} Q^{-k} J^{k-1} f(x) \\
&= x(t) + \left[\sum_{k=1} \frac{(\Delta t)^k}{k!} Q^{-k} J^{k} \right] J^{-1} f(x) \\
&= x(t) + \left[\sum_{k=0} \frac{(\Delta t)^k}{k!} Q^{-k} J^{k} - I \right] J^{-1} f(x) \\
&= x(t) + \left[\exp(\Delta t Q^{-1} J) - I \right] J^{-1} f(x)
\end{aligned}
$$

This is the final updating algorithm of the integration, or in a slightly different notation:

$$
x(t_{n+1}) = x(t_n) + [\exp(dt \cdot Q^{-1} J) - I] J^{-1} f(x(t_n)). \tag{2.7}
$$

One big advantage of this integration scheme is that one does not need to keep track of the history of the states, as $f(x(t))$ is only ever evaluated at the current timepoint. This makes the approach very efficient (although there are matrix inversions and matrix exponentials involved). Also, if the system is linear in the states, one only needs to compute the Jacobian once, since df/dx is constant. For the equations of the convolution based neural mass models, this is in principle not the case due to the sigmoid transformation of the states, converting voltage of a population to a firing rate. Yet, the standard implementation in SPM does make that additional approximation, evaluating J only once. There are schemes in place which would allow for the continuous evaluation of the Jacobian (see spm_int_J.m). However, they make integration one order of magnitude slower. One crucial question is when the linear approximation of delays (in time) becomes invalid. This is a priori unknown and will differ for different systems. We will touch upon this topic in the simulations. Finally, based on the update equation in Eq (2.7) it is also not clear how/whether causal structures can be preserved, since it basically is an ODE ozaki update with adapted step-size. In fact, the simulations will show violation of causal dependencies imposed by delays in systems even simpler as the equations for DCM for EEG.

2.3.2 Continuous Extensions of ODE Methods

Euler (forward) integration schemes have been around for two centuries and are based on the following idea. If the dynamics of a system are prescribed by

$$
\dot{x}(t) = f(x)
$$

then for a small timestep dt, the system will evolve in the direction of its gradient (in time), i.e.

$$x(t+dt) = x(t) + dt \cdot f(x). \tag{2.8}$$

Obviously, delays are not incorporated in Eq. (2.8). In the spirit of continuous extension of ODEs (Feldstein and Goodman 1973), we used an approximation to the states $x(t-\tau) \approx \tilde{x}(t-\tau)$, by linearly interpolating between the two neighbouring, evaluated timesteps. Thus, for

$$\begin{gathered} t_i < t - \tau < t_{i+1} \\ t_j - t_{j-1} = dt \\ \tilde{x}(t-\tau) = x(t_i) + \frac{x(t_{i+1}) - x(t_i)}{dt} \cdot (t - \tau - t_i) \end{gathered} \tag{2.9}$$

A graphical illustration is provided in **Figure 4**.

In full, the update (φ) of the linearized delayed Euler scheme (LDE) is given by

$$\vec{x}(t_{t_{n+1}}) = \vec{x}(t_n) + \varphi(\vec{x}, dt, t_n, \tau) = \vec{x}(t_n) + dt \cdot f(\vec{\tilde{x}}(t_n - \tau)) \tag{2.10}$$

Please note that we used a short hand notation here. States $\vec{x}$ denote state vectors and likewise are the delays $\tau = \tau_{ij}$ state-dependent. Therefore, Eq. (2.10) is an extension of Eq. (2.9) for a system of DDEs with state-dependent delays. There are two major discussion-points for the interpolation method – one is the tracking of the already mentioned propagation of discontinuities, the other so called *overlapping*, i.e. when delays become smaller than the step size of integration. The second problem can be easily solved by reducing the step size. For the first problem, there is the hope that in the context of DCM, it might be less relevant. In DCM, the solution can be expected to be relatively smooth at transition points, because for many real world scenarios, the input lag of the driving input is longer than the delays, and the initial conditions are otherwise a steady state of the system. Therefore, propagation of jumps in the derivative at the transition point might be less of a problem. One intriguing property here is that the continuous extension is independent of the discrete ODE method, and different updates can be combined.

Hence, we also introduced a hybrid of the default integration scheme for DCMs (spm_int_L), and the continuous extension. This will allow us to draw an extended comparison and to assess the impact of the linearization proposed in LLDI. We will refer to the scheme as *linearized delayed ozaki scheme (LDO/ozaki).* The linearization on delayed states combined with an Ozaki update (Eq. (2.3)) results in the following update:

$$x(t_{n+1}) = x(t_n) + [\exp(dt \cdot J - I)]J^{-1}f(\tilde{x}(t_n - \tau)).$$

In line with LLDI, we also opt to compute the Jacobian only once, i.e. assume a linear approximation to the system (in x).

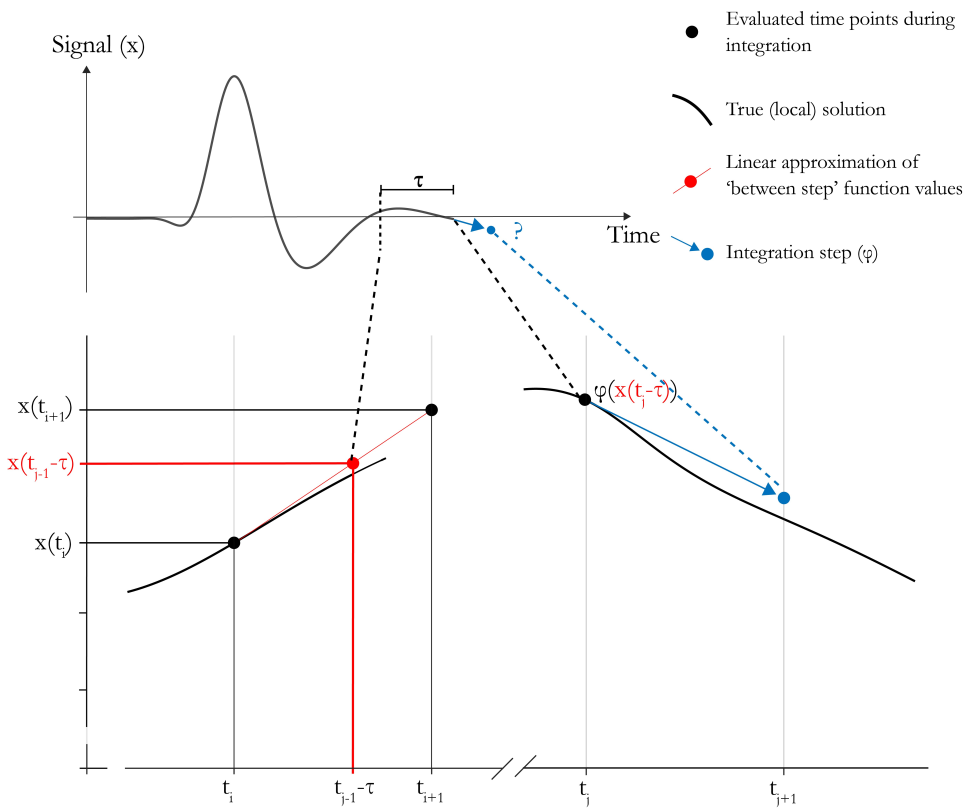

Figure 4| Schematic overview of the implementation of delays. The update function φ requires function values at previous timepoints. The function at times between sampled timepoints is linearly approximated for efficiency. The update step is then either done through an Euler or Ozaki step.

2.4 RESULTS

In this section we test the two-fold linearization integration scheme against the proposed history tracking integration schemes, which interpolate linearly between consecutive time points in order to evaluate the states at time $t - \tau$. In particular, we are interested in the following questions:

- What effects do the different integration schemes have on the qualitative behavior of the considered systems?
- Are causal dependencies preserved between coupled states in high dimensional dynamical systems? Are they consistent with initial conditions?
- At what length of delays do the approximations start to fail?

These three points will be addressed by two simple, illustrative examples: A one-dimensional decaying exponential and two coupled harmonic oscillators. We benchmarked the performance of the three integration schemes in comparison to a Runge-Kutta (RK) based scheme implemented in MATLAB 2017b (dde_23) for the integration of delay differential equations (Shampine, Thompson et al. 2000, Shampine and Thompson 2001). All simulations involving DCM for ERPs were performed using SPM12, ver. 6906.

2.4.1 The One-Dimensional Decaying Exponential

Consider the following dynamical system

$$\begin{aligned} \dot{x}(t) = f(x(t)) = -x(t-\tau), \\ x(t<0) = x_0 \end{aligned}$$

While simple, this system has an interesting property. For $\tau = 0$ it is simply a decaying exponential $x(t) = x_0 \exp(-at)$ with time constant $a = 1$. For $\tau = \pi/2$ the solution is a superposition of a sine and cosine function, i.e. an oscillating system. The proof is simple:

If

$$x(t) = A \cdot \sin(t) + B \cdot \cos(t)$$

then

$$\begin{aligned} \dot{x}(t) &= A \cdot \cos(t) - B \cdot \sin(t) \\ &= -A \cdot \sin(t - \frac{\pi}{2}) - B \cdot \cos(t - \frac{\pi}{2}) \\ &= -x(t - \frac{\pi}{2}), \end{aligned}$$

where we used trigonometric identities in the second step. Thus, we would assume that as the delays are increased, the system will start to show oscillatory behavior. We simulated the system with the following specification (**Table 1**).

$$\dot{x}(t) = f\big(x(t)\big) = -ax(t-\tau),$$

Integration time	Integration steps	Initial condition	Parameters	Delays
$t \in [0, 0.5]$	$dt = 0.001$	$x(t < 0) = 10$	$a = 10$	$\tau \in [0, 0.1]$

Table 1| Parameter setting for simulation of one dimensional, decaying exponential.

We considered delays up to the decay time of the system. The results are displayed in **Figure 5**. For small delays, all integrators perform reasonably well, with LDO being slightly superior to LDE for the same step size. As the delay increases, all integrators who use the delayed states show the onset of the expected oscillatory behavior and perform very similar to the benchmark (dde_23).

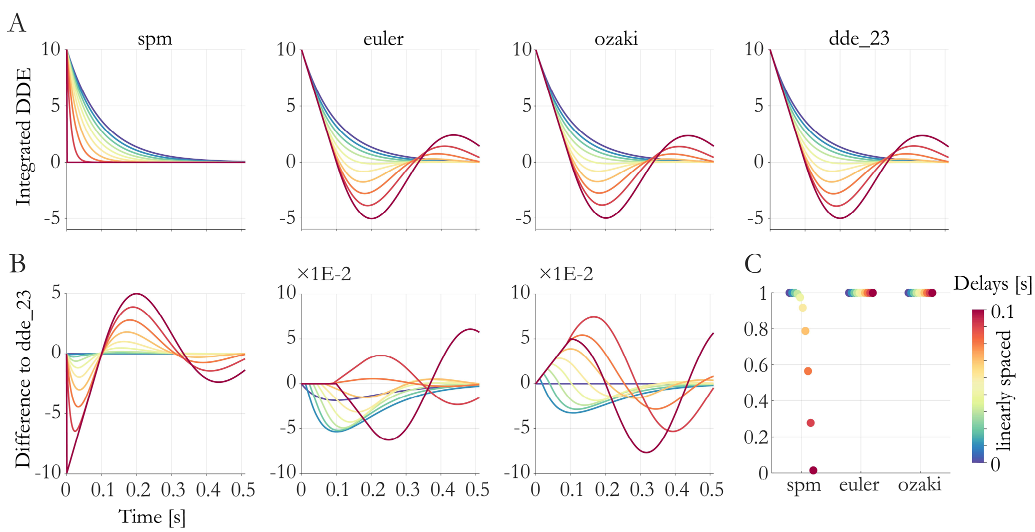

Figure 5| Signals computed for different levels of delays, for a 1 dimensional, delayed exponential decay. MATLABs integration scheme (dde_23) taken as a reference. A) Integrated states with the four integration schemes. B) Deviation of spm (LLDI), euler (LDE) and ozaki (LDO) compared to dde_23. C) Explained Variance of the three former integration schemes (compared to dde_23) for all delays.

In contrast, LLDI completely fails to perform oscillations and diverges for $\tau^* > 0.1$ s. For this simple system, we can understand this behavior analytically.

The full updating scheme for this system is given:

$$f\big(x(t)\big) = -x(t), \quad f\big(x(t-\tau)\big) = -ax(t-\tau)$$

$$J\big(x(t)\big) = \frac{df}{dx} = -a$$

$$\tau = D$$

And therefore, the LLDI update according to Eq. (2.7) is given by

$$x(t_{n+1}) = x(t_n) - \frac{1}{a}\exp(-\frac{1 + a(dt - \tau)}{1 - a \cdot \tau}) f(x(t_n))$$

There are two important observations to be made here: First, in the one-dimensional case Q is a scalar, and in this setup, the LLDI update is equivalent to an Euler update of the non-delayed system (with an adjusted stepsize of $-\frac{1}{a}\exp(-\frac{1+a(dt-\tau)}{1-a\cdot\tau})$). Obviously, the non-delayed system does not oscillate. Second, at $\tau = 1/a$, the update has a singularity for finite dt, which is what we observe in the extreme behavior for the highest delay (**Figure 5**A).

In conclusion, the linear interpolation to compute delayed states leads to similar, qualitative behaviors as the more computationally expensive dde_23, for the whole range of considered delays. LLDI only performs well for very small delays but is unable to reproduce the qualitative behavior (oscillations) at larger delays. The qualitative change of the system into oscillatory behavior starts around delays of $\tau \approx 60$ ms, i.e. 60% of the systems intrinsic time constant, where LLDI explains about 80% of the variance (vE) of the benchmarking signal (dde_23, **Figure 5**C).

2.4.2 Coupled Harmonic Oscillators

For this second example, we considered two coupled harmonic oscillators, which is a system that, dynamically, closely resembles the equation underlying convolution based DCMs.

The dynamics of a single harmonic oscillator (HO) is represented by the following differential equation,

$$\ddot{x} + f\dot{x} + \omega^2 x = 0, \quad x(t < 0) = \phi(t) = x_0$$

where the initial conditions are encoded in $\phi(t)$. The variable f represents some force proportional to the velocity, e.g. friction, and ω^2 some retractive force due to some displacement from the equilibrium position.

One can introduce auxiliary variables x_1 and x_2, to transform this second order ordinary differential equation (ODE) into two first order ODEs

$$\dot{x}_1 = x_2$$

$$\dot{x}_2 = -f x_2 - \omega^2 x_1$$

Or in matrix form

$$\dot{\vec{x}} = A\vec{x}$$

with

$$A = \begin{pmatrix} 0 & 1 \\ -\omega^2 & -f \end{pmatrix}$$

Hence, for coupled harmonic oscillators, one can use the same trick, and write the equation as

$$\begin{aligned}
\dot{x}_1 &= x_2 \\
\dot{x}_2 &= -f_1 x_2 - \omega_1^2 x_1 \\
\dot{x}_3 &= x_4 \\
\dot{x}_4 &= \kappa x_2 - f_2 x_4 - \omega_2^2 x_3
\end{aligned}$$

where κ then denotes the coupling from the first to the second HO. We now consider a system, where only the first oscillator is driven by some external driving input $u(t)$ and is coupled to the second harmonic oscillator. In matrix form, this yields

$$\dot{\vec{x}}(t) = f(x(t)) = A\vec{x}(t) + \vec{C}u(t),$$

with

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -\omega_1^2 & -f_1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & \kappa & -\omega_2^2 & -f_2 \end{pmatrix}, C = \begin{pmatrix} 0 \\ c \\ 0 \\ 0 \end{pmatrix}$$

Finally, let's consider some delay from the first to the second HO, i.e.

$$\dot{x}_i(t) = f(x(t)), i = 1, 2, 3$$

and

$$\dot{x}_4(t) = f(x_j(t), x_2(t-\tau)), j = 1, 3, 4.$$

We simulated the two harmonic oscillators with the following configuration (**Table 2**)

Integration time	Integration steps	Initial condition	Parameters	Delays
$t \in [0, 0.5]$	$dt = 0.001$	$\vec{x}(t < 0)$ $= [0\ 0\ 0\ 0]'$;	$w_1 = 10 \cdot \pi$ $w_2 = 10 \cdot \pi$ $f_1 = 20$ $f_2 = 20$ $\kappa = 6 \cdot \pi$	$\tau \in [0, 0.03]$

Table 2| Parameter setting for the simulation of a two dimensional harmonic oscillator.

A gamma pulse was used to drive the system $u(t) = gamma(\theta)$ with mean at 0.2 s. The integrated states are displayed in **Figure 6**.

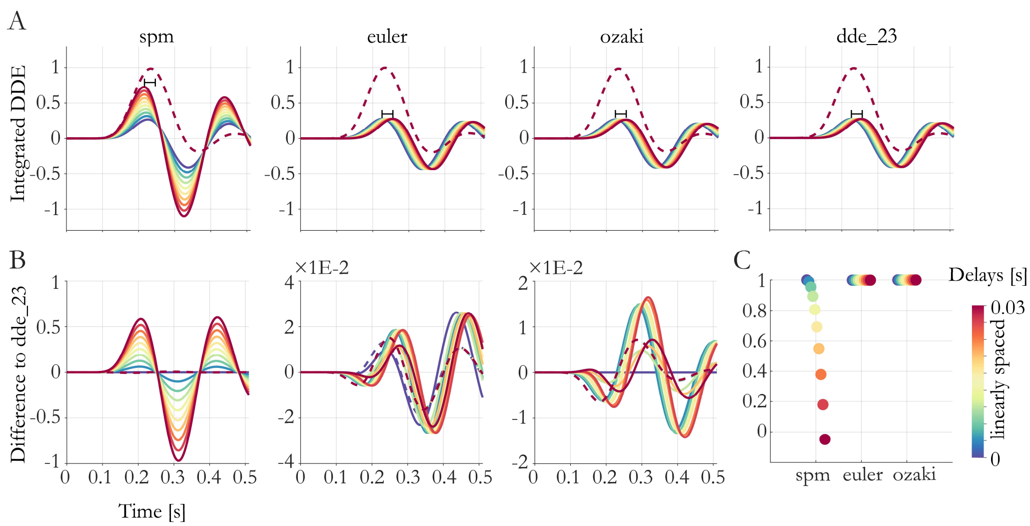

Figure 6| Signals computed for different levels of delays for coupled harmonic oscillators. MATLABs integration scheme (dde_23) taken as a reference. A) Integrated signals with the four integration schemes. The dashed line shows the activity of the first harmonic oscillator. Solid lines indicated activity of the second oscillator for a series of delays indicated by color (c.f. Panel C). Black line indicates 0.03 ms (for illustration). B) Difference of spm (LLDI), euler (LDE) and ozaki (LDO) scheme to dde_23. C) Explained variance of the three former integration schemes (compared to dde_23) for all delays.

Again, for very small delays, all integration schemes perform reasonably well. As the delay increases, LLDI starts to change drastically in the qualitative behavior. Delays are not at all taken into account, in fact, the peak amplitude for the second HO is reached even earlier for increasing delays. As a consequence, the underlying frequencies in the system change (note that the amplitude of the residuals are of the same magnitude as the true signal). vE of LLDI drops to 80% for delays around $\tau = 15$ ms, i.e. about 50% of the time constants of the un-damped, individual oscillators ($\frac{1}{w_1} \approx 30$ ms)

2.4.3 Convolution based DCM for ERP

Finally, we turn to the case of a simple convolution based DCM with three populations. We connected two regions, with a forward connection from region 1 to region 2, and a backward connection from region 2 to region 1. Region 1 received a standard, gaussian-like, driving input (the default input in the DCM for ERP framework (see *spm_erp_u.m*)). We linearly changed extrinsic delays only from region 1 to region 2. Most parameters were set to their default prior means (as per SPM12, ver. 6906). All changed parameters are provided in **Table 3**:

Integration time	Integration steps	Initial condition	Parameters	Delays
$t \in [0, 0.5]$	$dt = 0.001$	$M.x = zeros(2, 9);$	$A\{1\}(2,1) = 1$ $A\{2\}(1,2) = 0$ $C(1,1) = 0$ $M.ons = 64$ $M.dur = 16$	$D(2,1) = [0, 1.5]$

Table 3| Parameter changes. All remaining parameters set to their default prior means as specified in SPM12, ver. 6906).

It is important to note that delays are also specified in log-space and extrinsic delays are scaled by a factor of 16 ms. Hence, we effectively increased the extrinsic forward delays from 16 up to 71.7 ms, while the backward and intrinsic delays were fixed at 16 ms and 2 ms, respectively. This choice was made to remain close to the default settings. The performance of the four integration schemes is shown in **Figure 7**.

For these simulations, we need to consider that there is no case without delays, hence all integrated systems are dependent on the specific integration of DDEs. This explains why there is no case, where LLDI performs at the same level as ode_23. Otherwise, and as one would intuit due to the close familiarity with the simulations from the coupled HOs, we see a similar behavior. While both integration schemes with the linear approximation perform qualitatively similar to dde_23, LLDI fails to capture the delayed response in the second region. Again, it predicts oscillations at different frequencies and delays artefactually increase the amplitude in the second region. Performance drops to $vE \approx 80\%$ for delays around $\tau \approx 35\, ms$. Under the default prior variance of $pC.D = \frac{1}{16}$, delays of this magnitude are about 3 standard deviation away from the prior mean.

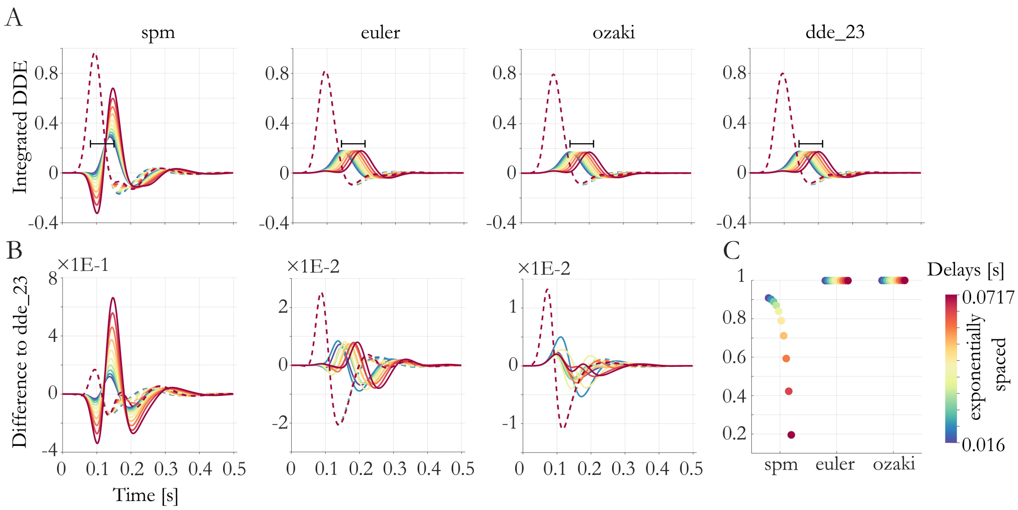

Figure 7| Signals computed for different levels of delays, for a two-source ERP model, for increasing delays on the forward connection. Only pyramidal voltage is shown. The dashed line shows the activity of the first harmonic oscillator. Solid lines indicated activity of the second oscillator for a series of delays indicated by color (c.f. Panel C). MATLABs integration scheme (dde_23) taken as a reference. A) Integrated signals with the four integration schemes. Black Line indicates 0.0717 s (for illustration). B) Difference of spm (LLDI), euler (LDE) and ozaki (LDO) scheme to dde_23. C) Explained variance of the three former integration schemes (compared to dde_23) for all delays.

2.4.4 Computational Efficiency

We tested the runtime for all four integrations in two settings (**Table 4**). Clearly, LLDI is very efficient, since the only added complexity compared to the non-delayed Ozaki (ODE) integration is the computation of a single matrix inversion. The Ozaki integrator with linear extension is in the same order of speed as the delayed Euler scheme. That is not surprising since the increased complexity is (again) only a single matrix inversion. On the other hand, dde_23 requires much more computational resources. The reason is two-fold. First, dde_23 performs multiple RK steps, i.e. multiple evaluations of the dynamic equation. Second, it uses subsampling to control the error, thus more integration steps. The latter disadvantage of course is reduced, when one considers subsampling for the other integrators (LDE, LDO), or more evaluations of the Jacobian (LLDI) for non-linear problems. A default inversion of ERPs takes roughly up to 80 optimization steps (in practice, it tends to be more for larger systems). A single optimization step, with 20 parameters, requires 21 integrations (one for the signal, 20 for the gradients). If we assume an additional 1s per integration required for dde_23 compared to the other integrators, this would amount to a difference of roughly 20 mins per model, per subject, per condition. It has to be noted that none of the integrators, including dde_23 was optimized for speed.

2.5.5 EMPIRICAL DATASET

In order to investigate the effects of the choice of integrator on an empirical dataset, we modelled LFP recordings of rodents, who underwent an auditory MMN paradigm. The rodents were chosen from the non-pharmacological part of the RATMPI study. All motivations for the model space and multistart are provided in Chapter 3 and 5 of this thesis.

In brief, the dataset consisted of eight hemispheres (four rats for each hemisphere, both hemispheres were available for three rats). The data consisted of epidural recordings from primary auditory cortex (A1) and posterior auditory field (PAF) in both hemispheres. The two regions were connected through a forward and backward connection, respectively, driving input only entered A1 (**Figure 38**). We inverted 16 DCMs (based on the canonical microcircuit) with different modulation (by *deviant*) of forward, backward and intrinsic connections, starting from 100 starting values for each model and rat/hemisphere.

	2 Sources				**4 Sources**			
average integration time	**LLDI**	**LDE**	**LDO**	**dde_23**	**LLDI**	**LDE**	**LDO**	**dde_23**
(in ms)	42	112	112	1251	47	239	238	4714
(in % of dde_23)	3.4	9	9	100	1	5.1	5	100

Table 4| Average integration time for the four types of DDE integrators, and a two and four region DCM (18 and 36 states respectively). Signals are integrated over 500 ms (Integration step size 1ms). Averages were computed over 1000 random initialization of the delays.

Our primary interest concerns, whether there is an effect of integration scheme on model selection and parameter estimation, in particular for modulatory effects. The results are summarized in **Figure 8**.

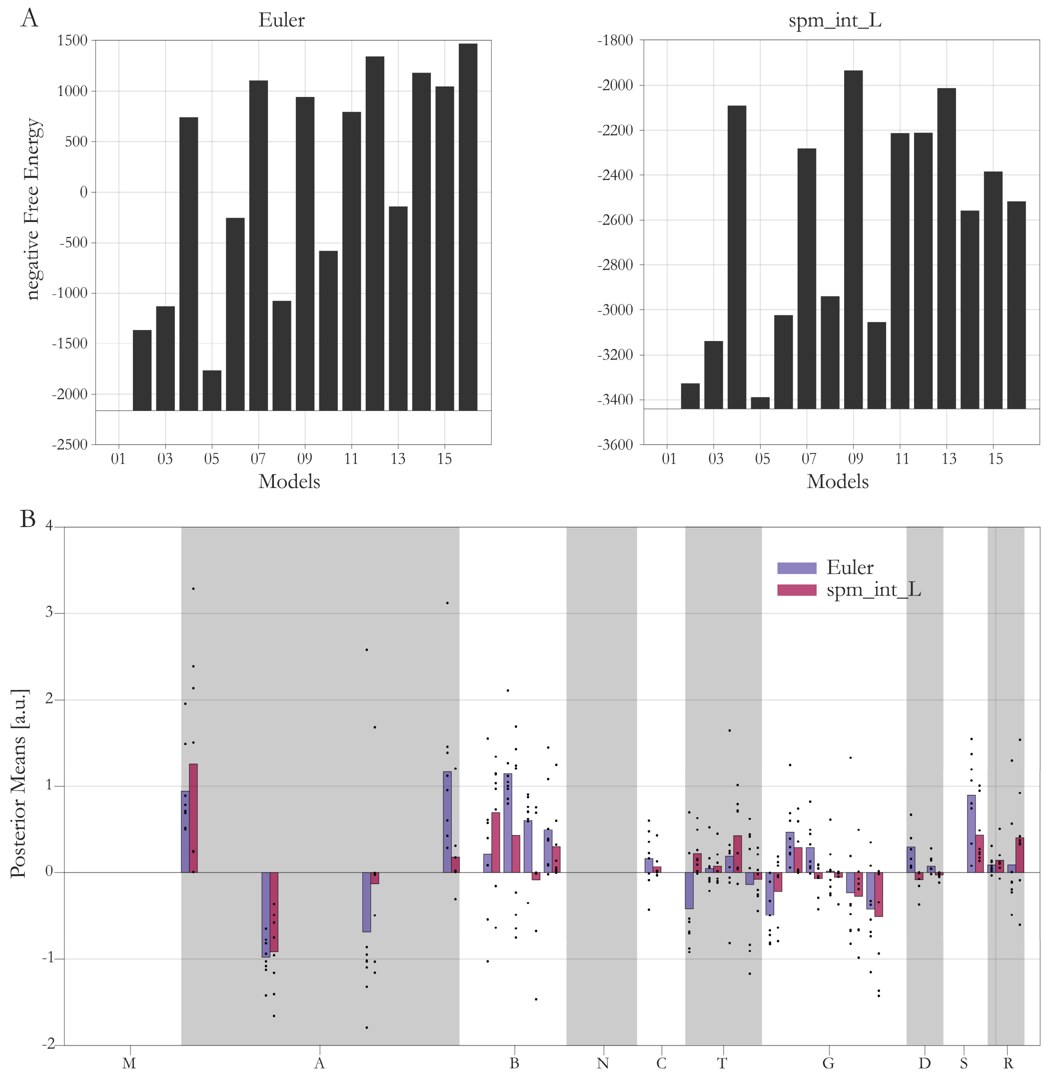

Figure 8| Comparison between integration schemes for inversion of the non-pharmacological part of RATMPI. Data consisted of four rodents with two hemispheres each (N=8). Hemispheres were treated as independent rodents. A) Model comparison (fixed effects BMS) for the two integration schemes. B) Posterior estimates of the most complex model (Model 16) across eight hemispheres. Bars depict average parameter estimates over hemispheres, single dots depict posterior means for individual hemispheres.

In terms of model selection, we can observe a difference in conclusion about the winning model (**Figure 8**A). The Euler-based integration scheme deems model 16 (i.e. including all modulations) the most likely model to have generated the data. LLDI considers model 9 the winning model. In this model, the self-inhibition of the superficial pyramidal cell is not modulated by *deviant*. In principle, one could also formally draw a family comparison between the two integration schemes, where the Euler-based scheme clearly outperforms LLDI (it achieves overall higher negative free energies). Arguably, negative free energy might not be the most natural quantity to assess the goodness of an integration scheme.

However, we can observe that the differences are certainly in a numerical regime that is relevant for model comparison.

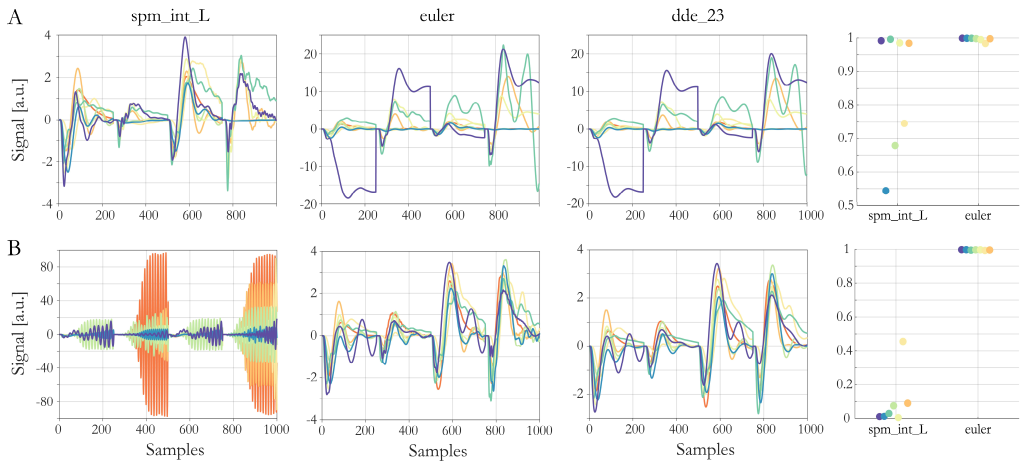

Figure 9| Comparison between integration results when generating signals from posterior means based on one integration scheme. A) Posterior means stem from inversion using spm_int_L (N=8 hemispheres). Correlation between integrated signals with spm_int_L / Euler and dde_23 (right-most plot) for all signals. B) Posterior means stem from inversion using Euler scheme with linear interpolation for delayed states. Correlation between integrated signals with spm_int_L / Euler and dde_23 (right-most plot) for all hemispheres.

In terms of posterior estimates, we found sign-flips (of the group-average parameters) for the backward modulation, the kernel decay of the stellate cells, two kernel gains and both extrinsic delays (**Figure 8**B). Additionally, fairly big differences could be observed in many other parameters, including modulation parameters (*B*). Two points are worth some attention here: First, there is overall a fairly high variance in posterior means across rats and second, it is not clear what *high* exactly means. One would have to test in a sensitivity analysis like fashion, whether these parameter differences (conditioned at some point in parameter space) actually make a difference in terms of predicted signal. What we could test however, is whether integration of the system conditioned on the parameters derived with one scheme, is stable under the other scheme. For this, we took the posterior means derived with LLDI, and integrated the system with LDE (and vice versa). For reference, we again used dde_23. The results are summarized in **Figure 9**. Clearly, at the respective posterior means, the two integration schemes generate very different signals. While LDE performs very similarly to dde_23, LLDI diverges at the posterior estimates acquired with the LDE scheme, for all rats. That could be due to the fact that under the Euler scheme, posterior delays are estimated to be higher than with LLDI (see **Figure 8**). For the posterior means acquired with LLDI, the system is slightly better behaved, but still with three out of eight parameter sets leading to different results. Again, the likely reason is that under LLDI, delays are estimated to be shorter, hence in a regime where also the simulations have shown that the three integrators perform similarly.

2.5 DISCUSSION

In this chapter, we only provided a simplistic introduction to the problem of delay differential integrations. Of course, the whole development did not stop in 1964 and goes beyond continuous extensions with a discrete ODE step (e.g. Method of Steps (Bellman 1961)). However, the simplicity made continuous extensions to ODEs an interesting vantage point to start, and with the MATLAB implementation of dde_23, we had a benchmark for comparison.

We formulated three questions about the effect delays have on dynamical systems convergent on DCMs, all of them with the final goal of understanding, whether parameter estimation might be impaired due to the particular choice of integration scheme. (i) Does the implementation of delays in the integration scheme violate the causal structure? (ii) Does the integration scheme affect the qualitative behavior of the delayed system? (iii) What are the regions of stability for the delays?

The first two questions can, in all generality be answered with a straight *yes*. In the second and third simulation of the coupled HOs and the simple DCM, we observed that delays violated the causal structure in LLDI, as we observed an oscillatory response in the second HO at times 'forbidden' under the imposed delays. In all three simulated systems, we saw a qualitative change in behavior with increasing delays in LLDI – once it was a failure to oscillate (1D), twice it was a change in the characteristic frequencies of the system, and visible effects on peak amplitude (HO and DCM). Both effects are expected to impact estimation of other parameters in empirical data. In the convolution based DCM framework, the convolution kernel can be understood as the impulse response function of a filter. Therefore, its parameters define the filter properties. If changes in delays artefactually change the underlying frequencies of the system, it will in turn also change the filter properties. The same holds for peak amplitudes. In principle, all kernel-related parameters can lead to a change in the amplitude – local excitability, connection strengths or input gain, as well as parameters of the forward model and delays. However, the LLDI scheme predicts delay related amplitude effects that are much more severe than expected (under the alternative schemes and the benchmark). Apart from increasing the peak amplitude in the second region in the DCM simulations, LLDI predicts the presence of an additional, early negative peak. In contrast, the linear approximation to delayed states worked qualitatively well, for the given time constants of the system and the integration step. There is not much difference between the two updates, i.e. Euler vs Ozaki step in terms of accuracy. They also perform similarly in terms of their computational efficiency. That is not surprising, since the matrix inversion and exponential needed in the Ozaki scheme are done only once, hence the most time consuming step would be the interpolation of delayed states, which is equivalent in both schemes. Also in terms of their error, both schemes perform very similar, which intuitively hints towards the majority of the error coming from the interpolation. The clear advantage dde_23 has over the competitors is the fact that the time step of integration is subsampled and error bounds can be defined. Ultimately, it will be a question of trading off computational efficiency and accuracy. The speed benefits of the linear interpolation over dde_23 might not be relevant in rigorous clinical applications, but most certainly in the use for the scientific community. Because of the multiple RK steps and the subsampling, dde_23 is easily four times slower than the other schemes. Whether an inversion of a model takes 30 mins or 2 hours or more, is definitely a consideration for the community, especially without access to high performance clusters.

The last question might be the most important – are the ranges of delays for real, physiological systems relevant in terms of qualitative changes to the dynamics of the system. First of all, we would like to point out that our simulations only cover a set of particular choices, from which we cannot easily generalize. What we could observe however, was that typically at delays around 25%-50% of the intrinsic time constants of the system, LLDI performed qualitatively differently[9]. Auditory ERPs show frequencies in the range of 10 – 20 Hz, i.e. time constants of 50-100 ms (Haenschel, Baldeweg et al. 2000). Therefore, based on the animal literature mentioned in the introduction, the fact that relays via other, un-modeled states need to be taken into account and our empirical findings, delays can be of relevant magnitude.

The final discussion point regards, whether the integration scheme can affect conclusions drawn in terms of BMS or parameter estimates. Importantly, in many applications of DCM, one is not interested in absolute parameter values, but rather in changes of parameters

[9] The current prior in SPM expects conductance delays between cortical columns on the order of 9-25 ms (i.e approximately 95% of the prior mass).

across conditions (e.g. experimental manipulations, stimulation, pharmacological interventions, etc.). Since delays are estimates to be the same across conditions, differential effects on other parameters might be less impacted. Our empirical analysis showed that conclusions were quite different. Average modulatory parameters displayed fairly large differences in magnitude and even a sign flip for the backward modulation could be observed. A fixed effects Bayesian model comparison resulted in evidence for the most complex model (*m16*) when inverting the models with LDE, while model comparison with LLDI spoke in favour of (*m09*), which only allows for modulation of extrinsic connections. This is not surprising, as the two types of integration schemes generated very different signal at the respective posterior means.

A previous study has also looked in detail at the effect of integration scheme on parameter estimation in simulations and on empirical data (Lemarechal, George et al. 2018). Our results are very much in agreement with their finding. They found the same qualitative differences how delays change the influence one region exerts over another region, and severe changes in data features. Also, they directly compared spm_int_L to dde_23 in a synthetic DCM setting, which augments our simulation work. They simulated data for a two and six region DCM (with dde_23), with modulation of both forward and backward connections for different levels of delays. Over the two network sizes, their finding indicated that spm_int_L had trouble identifying the correct model structure based on the expected posterior model probability (EPP). For a system with modulation of forward and backward connection, they found a non-negligible model probability that the data was generated either by **only** forward, or **only** backward modulation (EPP = 42 % and EPP = 47% for two and six regions, respectively). Also, they found that modulation strength under spm_int_L was systematically underestimated and delays poorly recovered. Both results of course might have depended on the exact choice of simulated networks. Correlations between parameters can lead to similar combinations of parameters resulting in similar behavior of the system. Additionally, non-linearities in the likelihood function render the objective function non-unimodal – a challenge for an optimization scheme. Both of these points will come up in the chapter on the use of a multistart approach for optimization. With this in mind, we could not confirm that spm_int_L *generally* underestimates modulation strength as presented in (Lemarechal, George et al. 2018) (see **Figure 8**). Of course, for empirical datasets, we never know ground truth. But reasonable doubts can be cast about the validity of the approach of spm_int_L for the typical frequencies and delays in the empirical data, as discussed earlier. Also, it is surprising that in terms of model selection, based on spm_int_L, we would have concluded that no modulation of intrinsic connections exists, while self-modulation of A1 is estimated (on average) to be of highest magnitude (**Figure 8**B). Granted, Bayesian Model Selection does not always concur with the intuition one would have from classical t-tests on the posterior means, but it is nonetheless surprising. Especially, since we observe a fairly large MMN effect around the timepoint of the first peak in A1 in the empirical data. Simply because of timing reasons, this effect must almost exclusively come from a modulation of local excitability, because all other effects would need to be mediated via PAF, and therefore arrive delayed. But as the simulations have shown, causality can be violated in spm_int_L and intuitions might not hold anymore.

In conclusion, for robust parameter inference, where construct validity is of concern, in this thesis all forthcoming analyses use the Euler-based integration method with linear interpolation of delayed states. The time penalty involved seems acceptable (unlike for dde_23), given the accuracy tradeoff.

3| LOCAL EXTREMA IN VBL OPTIMIZATION OF DCM

3.1 DISCLAIMER

These methodological developments and analyses were done under the supervision of Jakob Heinzle and Klaas Enno Stephan. Stefan Frässle and Eduardo Aponte consulted in constructive discussions.

3.2 INTRODUCTION

As we motivated in the introduction to this thesis, the call for reproducibility, replicability and robustness of scientific studies has grown louder. Of course, this call has also reached the computational psychiatry community (as an example, see the work regarding DCMs (Frässle, Stephan et al. 2015, Litvak, Garrido et al. 2015)). From a translational perspective, it is particularly of relevance, as the model universally connects patient-data (e.g. electrophysiologial measures) and potential pathophysiological mechanisms (for review, see (Frässle, Yao et al. 2018)). This connection makes it evident, why robustness of the computational methods is so crucial; The step to clinical utility requires robust measures which depends on both, the model and its inversion.

The standard optimization procedure used in DCM is based on a Variational Bayes approach under the Laplace approximation (VBL, (Friston, Mattout et al. 2007)). VBL is known to work well for well-behaved objective functions (e.g. linear models). However, the dynamical equations underlying the convolution based DCM for ERPs are non-linear, and the objective function is likely to present with multiple local maxima. This could greatly impact the robustness of parameter estimates, and sub-optimally identified maxima could lead to erroneous optimal values for the objective function and hence inference on model structure. Therefore, both measures proposed for subgroup-identification, as proposed by the idea of generative embedding (Brodersen, Schofield et al. 2011), are crucially affected by the inference machinery.

This problem of multiple local maxima was already raised in a response to (justified) challenges regarding the plausibility of the biophysical models in general, but also the robustness of the inference techniques in DCM (Daunizeau, David et al. 2011). The authors discussed both DCM for fMRI and EEG, and agreed that a potential local maxima issue "could manifest in inconsistent parameter estimations and model comparisons" across experiments. Complementing the points raised so far is also the fact that the approximate posterior densities are can be too tight (i.e. an overconfidence in posterior precision) (Beal 2003). Since the posterior precision directly affects the objective function in DCM (as will later be shown in this chapter), a bias could pose a problem (especially) for model comparison. However, by comparing the VB approach to sampling algorithms which do not depend on the Laplace approximation, it has been shown in simulations that the Laplace approximation was appropriate for at least the fMRI variant of DCM (Chumbley, Friston et al. 2007). However, please also note that Aponte et. al. 2018 did find differences between a sampling based approach and VBL in an empirical attention to motion and face perception dataset. Also they illustrated a dependency of VBL on starting values (Aponte, Raman et al. 2018). Another study then followed up on the issue in DCM for EEG. Again, using sampling as a benchmark, it was shown that while the values of the objective functions were different, the methods agreed in terms of model comparison (Penny and Sengupta 2016).

Still, to our best of knowledge, DCM lacks a quantitative and qualitative investigation of the local maxima problem. Potential confounds induced by local maxima are important, especially since multiple studies have generally shown the predictive potential of DCMs for EEG (Moran, Jung et al. 2011, Moran, Symmonds et al. 2011, Moran, Jones et al. 2015). In this chapter, we try to bridge that gap. For that, we first sketch the derivation of the objective function used in DCM for ERPs and point to the seminal literature that has introduced the concepts. We then simulated data from a convolution based DCM for ERP with two regions and five different modulation structures with increasing complexity. We inverted all dataset under all models (i.e. hypothesis about how connections are changed under an experimental manipulations) using a multistart approach. Here, we specify multiple starting values for the VBL optimization. In the presence of local maxima, one would then expect that different starting values lead to different solutions and hence inference on the levels of parameters and possibly also models. We quantitatively compared the conclusions drawn about network structure and parameter values between the default starting value (which corresponds to the prior mean) and the starting value resulting in the highest negative free energy (i.e. proxy of a 'global' solution). Additionally, we (sparsely) mapped extended structures of the optimization landscape by combining the results from all starting values.

Overall, our simulations indicate the need for the general use of a multistart-augmentation of the VBL inversion approach to reduce the effect of local critical points on parameter and model inference. Even then, dependencies between model parameters create a tough optimization problem, where the objective function seems to present with slowly ascending ridges and steep ravines; a challenge for any optimization scheme. Hence the goal of this chapter is to show that at this level of difficulty, model interpretation goes beyond simple reports of negative free energies, that one should be cautious in over-interpreting model selection as well as parameter estimation results and that proper diagnostics are vital. All diagnostics presented in this chapter will become part of a software toolbox for MATLAB (https://www.tnu.ethz.ch/en/software/tapas.html), which is openly available for the community.

3.2.1 Bayes' Rule, Posterior Distributions and Log Model Evidence

Generative models are fully described by two ingredients - A *prior* $p(\theta)$ and a *likelihood* $p(y|\theta)$. The prior defines some a priori probability density over model *parameters* θ. The likelihood defines the likelihood of *data* y given (conditioned on) a set of parameters. Together, the two terms build the joint probability of both parameters and data,

$$p(y,\theta) = p(y|\theta) \cdot p(\theta). \tag{3.1}$$

This equation defines the full generative model. The name derives from the fact that given a prior and a likelihood function, one can (in a probabilistic sense) generate data by simply sampling parameters from the prior $\theta_i \sim p(\theta)$, and plugging them into the likelihood

$y_i \sim p(y|\theta_i)$. This is also known as forward modelling. The beauty of Eq. (3.1) is the fact that it also prescribes the reverse direction through the so called Bayes' Rule:

$$p(\theta \mid y) = \frac{p(y \mid \theta) \cdot p(\theta)}{p(y)}. \tag{3.2}$$

This quantity, the so called *posterior density,* defines the probability of the parameters after having seen the data. It is directly related to the joint probability of parameters and data, normalized by the marginal probability of the data. *Estimation* of model parameters or *model inversion* then refers to the process of computing these posterior densities. In practice, computing the posterior density can be computationally expensive, as for many problems (all problems where the prior is not conjugate to the likelihood), one needs to evaluate the normalization constant, which is an integral of the dimensionality of the parameter space. For DCM for ERPs, there can be tens (sometimes more than a hundred) of parameters depending on the size of the network, rendering numerical integration infeasible. Additionally, $p(\theta|y)$ has, in general, not a simple parametric form of a known distribution (again, assuming non-conjugate priors).

On the other hand, the normalization constant in Eq. (3.2) provides a measure for model goodness. One can rewrite $p(y)$ as the marginal likelihood of y, integrating out uncertainty about the parameters,

$$p(y) = \int p(y \mid \theta) p(\theta) d\theta \tag{3.3}$$

and interpret it as the expected likelihood under the prior $p(y) = E[p(y|\theta)]_{p(\theta)}$. It has also been termed Model Evidence (Bishop 2006).

This quantity (specifically the log model evidence) can be split into two opposing terms:

$$\begin{aligned}
&\log p(y) = \log p(y) \int p(\theta \mid y) d\theta \\
&= \int \log p(y) p(\theta \mid y) d\theta \\
&= \int \log \frac{p(y \mid \theta) p(\theta)}{p(\theta \mid y)} p(\theta \mid y) d\theta \\
&= \int \log p(y \mid \theta) p(\theta \mid y) d\theta - \int \log \frac{p(\theta)}{p(\theta \mid y)} p(\theta) d\theta \\
&= E[\log p(y \mid \theta)]_{p(\theta|y)} - KL[p(\theta \mid y) \,\|\, p(\theta)]
\end{aligned} \tag{3.4}$$

The first term, $E[\log(p(y|\theta)]_{p(\theta|y)}$ encodes the accuracy of a model in terms of an expected log-likelihood under the posterior. The second term, $KL[p(\theta|y)||p(\theta)]$, is positive semi-definite, hence always acts against accuracy and encodes complexity in terms of a Kullback-

Leibler (KL) divergence. As a consequence, one can interpret the log model evidence as a tradeoff between accuracy and model complexity.

Clearly, computing the log-model evidence is of the same complexity as computing the posterior (in fact, it depends explicitly on computing the posterior). So the two computational challenges are inherently linked – computing the posterior distributions to compare and interpret parameter estimates and computing the log model evidence as an objective function to compare models and interpret network architectures. This is where variational calculus comes into play.

3.2.2 Approximate Variational negative Free Energy

Parameter inference refers to computing multivariate posterior distributions for the whole set of parameters of a model. As mentioned previously, this endeavor is analytically not tractable. One way around computing the integral is to define a lower-bound approximation $L(q)$ to the log model evidence (Eq. (3.4)) through an *approximate posterior distribution* $q(\theta)$, known as Variational Bayes (VB) (Bishop 2006).

It can be shown that for any distribution $q(\theta)$, the log model evidence can be decomposed in the following way (Bishop 2006)

$$\log p(y) = L(q) + KL(q \,||\, p)$$

with

$$L(q) = \int q(\theta)\log\frac{p(y,\theta)}{q(\theta)}d\theta = \int q(\theta)\log p(y\,|\,\theta)d\theta - \int q(\theta)\log\frac{q(\theta)}{p(\theta)}d\theta \tag{3.5}$$

$$KL(q \,||\, p) = -\int q(\theta)\log\frac{p(\theta\,|\,y)}{q(\theta)}d\theta \tag{3.6}$$

Again, since a KL divergence is always positive semi-definite,

$$L(q) = \log p(y) - KL(q \,||\, p)$$

will always be a lower bound approximation to the log model evidence $logp(y)$, with equality if, and only if $q(\theta) = p(\theta|y)$ (then, the log-ratio in Eq. (3.6) vanishes, and so does the KL term). Since the log model evidence is independent of the parameters, computing the posterior therefore turns into the problem of maximizing $L(q)$ (with respect to $q(\theta)$), the so called *variational negative free energy*, which, in turn directly acts as an approximation to the log model evidence. In other words, the variational negative free energy is maximized, if $q(\theta)$ is as close as possible to the true posterior $p(\theta|y)$ (where similarity is measured in terms of KL-divergence), which also provides the best lower bound approximation to the log model evidence under the assumed approximate densities $q(\theta)$. Similarly to Eq. (3.4), the term in Eq. (3.5) is a tradeoff between the accuracy of a model (expected log-likelihood under the approximate posterior) and the complexity of the model (KL-divergence between the approximate posterior and the prior). Please note that, given $q(\theta)$, all quantities in Eq. (3.5) are known (unlike the true posterior). Unfortunately, the integral is again of the same complexity as before. This is addressed through a *Laplace Approximation.*

The Laplace Approximation approximates an expected value (formulated in terms of the integral in Eq. (3.5)), through a 2nd-order Taylor Approximation. In brief, let $f(\theta)$ be any smooth function and $h(\theta)$ be any probability density, with mode θ_0. Then

$$\begin{aligned} E[f(\theta)]_{h(\theta)} &= \int f(\theta)h(\theta)d\theta \\ &\approx (2\pi)^{p/2} f(\theta_0)h(\theta_0)\det(\Sigma_{\theta_0}). \end{aligned} \tag{3.7}$$

Here, p denotes the dimensionality of the space ($\theta \in M^{p\times 1}$) and $\hat{\Sigma}_{\theta}^{-1} = J(\theta_0)_{ij} = -\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log q(\theta)$. If we combine the negative free energy (Eq. (3.5)) with Eq. (3.7) and apply the Laplace approximation twice, where we define $f_1(\theta) = logp(y|\theta)$, $f_2(\theta) = log\frac{p(\theta)}{q(\theta)}$, and $h(\theta) = q(\theta)$, we end up at the *approximate variational negative free energy* $\mathrm{F}(\theta)$(Friston, Mattout et al. 2007)

$$F(q) = (2\pi)^{p/2} det(\Sigma_{\theta_0}) q(\theta_0)[\log p(y\,|\,\theta_0) + \log p(\theta_0) - \log q(\theta_0)] \tag{3.8}$$

So far, we have been describing a general framework for variational inference. Turning to the case of DCM for ERPs, we need to give the terms in Eq. (3.8) some parametric form to end up at the final objective function for DCM for ERPs. From now on, whenever we refer to the *negative free energy*, we will be referring to the *negative free energy under the Laplace Approximation* as defined in Eq. (3.8) to increase the readability of the text.

3.2.3 Objective Function in DCM for ERPs

In the (convolution based) DCM for ERP framework, parameters come in three flavours – neuronal, leadfield and noise parameters. The set of neuronal parameters includes all parameters influencing the dynamic equation of hidden neuronal populations (θ_p), i.e. *kernel gain* and *decay, connection strength, condition specific modulation* and *delay* parameters, as well as parameters of *population firing* and *driving input*. Leadfield parameters (θ_g) define the mapping from hidden neuronal sources to measured data and include *population specific* and *region specific gains*. Finally, noise parameters (θ_h) quantify the amount of unexplainable variance in the data (i.e. noise). We will focus on the noise model later in this thesis.

We have already devoted a small chapter to the generative model in DCM in the introduction to this thesis. We stated that the likelihood of the data y ($n_s =$ *number of datapoints*) is assumed to be Gaussian distributed around some prediction y_p, with some correlation structure Σ_y in the unexplainable noise. Additionally, we assume a multivariate Gaussian prior distribution $(q(\theta))$ over the parameters with mean μ_θ and covariance Σ_θ. For the approximate posterior distributions $(q(\theta))$, we choose again a multivariate Gaussian distribution with mean θ_0 and covariance Σ_{θ_0} ($p =$ *number of parameters*):

$$\begin{aligned} p(y\,|\,\Sigma_y) &= (2\pi)^{n_s/2}\det(\Sigma_y)^{-1/2}\exp[-\frac{1}{2}(y-y_p)^T\Sigma_y^{-1}(y-y_p)] \\ p(\theta) &= (2\pi)^{p/2}\det(\Sigma_\theta)^{-1/2}\exp[-\frac{1}{2}(\theta-\mu_\theta)^T\Sigma_\theta^{-1}(\theta-\mu_\theta)] \\ q(\theta) &= (2\pi)^{p/2}\det(\Sigma_{\theta_0})^{-1/2}\exp[-\frac{1}{2}(\theta-\theta_0)^T\Sigma_{\theta_0}^{-1}(\theta-\theta_0)] \end{aligned} \tag{3.9}$$

It is important to note that θ_0 and Σ_{θ_0} are the same vector and matrix as in Eq. (3.8), since for a Gaussian distribution, the mean equals the mode. This simplifies Eq. (3.8) even further. Since the exponential term vanishes in $q(\theta_0)$ the first term in $F(q)$ cancels, and one is left with (David, Kiebel et al. 2006)

$$F(q) = \log p(y\,|\,\theta_0) + \log p(\theta_0) - \log q(\theta_0). \tag{3.10}$$

As mentioned earlier, optimization of the negative free energy refers to finding $\theta_0 = \hat{\theta}$ and $\Sigma_{\theta_0} = \hat{\Sigma}_\theta$, such that the KL-divergence between the approximate posterior $q(\theta)$ and the true posterior $p(\theta|y)$ is minimized. Note that the 'hat-notation' is used to indicate the sufficient statistics of the approximate posterior. If we plug in the parametric forms from Eq. (3.9), we end up with

$$\begin{aligned}F(q(\hat{\theta},\hat{\Sigma}_{\theta}) = &-\frac{n_s}{2}\log(2\pi)\\ &-\frac{1}{2}\log(\det(\Sigma_y))\\ &-\frac{1}{2}(y-y_p)^T\Sigma_y^{-1}(y-y_p)\\ &-\frac{1}{2}\log(\det(\Sigma_{\theta}))\\ &+\frac{1}{2}\log(\det(\hat{\Sigma}_{\theta}))\\ &-\frac{1}{2}(\hat{\theta}-\mu_{\theta})^T\Sigma_{\theta}^{-1}(\hat{\theta}-\mu_{\theta}).\end{aligned} \tag{3.11}$$

The first three terms in Eq. (3.11) approximate the accuracy part of the negative free energy, the last three parts complexity. It is quite simple to get an intuition for the terms:

- The negative free energy increases, if
  - the residual error decreases (3^{rd} term)
- The negative free energy decreases, if
  - a parameter is added to a model (e.g. modulation, 4^{th} term)
  - the posterior mean becomes unlikely under the prior (6^{th} term)
  - the determinant of the posterior covariance decreases (5^{th} term)

This last condition can be understood in the following way: The determinant can be seen as a measure of volume spanned by the vectors of the matrix. This volume is maximized for orthogonal vectors, i.e. uncorrelated parameters and decreases for pairwise correlations. Therefore, it is a useful measure of complexity of a model, since over-parametrization generally leads to correlations between the parameters, and in turn to decreases in the negative free energy. Also, model comparison can be understood in the same line of thought. Whenever a parameter is added to a model (for example *modulation*), prior variance increases (i.e. $\log(\det(\Sigma_{\theta}))$, 4^{th} term). It is then the balance of the terms two, three, five and six to determine, whether the introduction of the parameter is beneficial in terms of free energy.

3.2.4 Gradient Ascent and Multistart

In the previous paragraph, we have derived the parametric form of the objective function, with respect to which we are trying to optimize the parameters of a multivariate normal distribution. At the maximum, this multivariate gaussian then approximates the posterior distribution over parameters $p(\theta|y)$, and the value of the negative free energy can be used as an approximation to the log-model evidence.

How the parameters are optimized, has not yet been discussed. DCMs for ERPs are, by default, inverted using a gradient ascent based scheme on the negative free energy (Friston, Mattout et al. 2007). Here, optimization of the parameters can be understood as a walk through the negative free energy landscape, where the trajectory is dependent on the local gradient

$$\dot{\theta} = f(\theta, \theta_0, \nabla_\theta F(\theta), \alpha) \tag{3.12}$$

The trajectory of θ can therefore be understood as an integration problem, and follows an Ozaki update with a Levenberg-Marquardt adjustment of the step size (through α), if updates are rejected (Levenberg 1944, Ozaki 1992, Friston, Mattout et al. 2007). Importantly, the trajectory depends on initial conditions θ_0.

In the presented variational framework, the gradients can be computed analytically, which allow for very efficient optimization (Friston, Mattout et al. 2007). The downside is that for non-convex objective functions, the algorithm can get stuck in local optima. Unfortunately, the objective function in DCM for ERPs is non-convex because of a non-linear mapping from parameters to the predicted signal.

In the case of DCM for ERPs with neuronal parameters (θ_p), parameters of the leadfield model (θ_g) and parameters of specifying the noise (θ_h), i.e. $\theta = \{\theta_p, \theta_g, \theta_h\}$, this results in the following mapping:

$$\begin{aligned}
&\theta_p \xrightarrow[\text{dynamic equations}]{} \dot{x}(t, \theta_p) = f(x, s(x(t-\tau)), \theta_p) \\
&\dot{x}(t, \theta_p) \xrightarrow[\int dt]{} x(\theta_p) \\
&x(\theta_p) \xrightarrow[\text{forward model}]{} y_p(\theta_p, \theta_g) = G(\theta_g) x(\theta_p)
\end{aligned} \tag{3.13}$$

While the integration is a linear operation, the sigmoid transform s, the delays τ and explicitly the kernel decay, which appears as $\kappa, \kappa^2 \in \theta_p$ in the dynamic equations, are not. Whether prior constraints allow the parameters to enter regimes of potential extreme values is of course unknown, but generally local critical points can be expected to be a problem for a gradient ascent based optimization scheme.

One way to alleviate the problem of local extreme values is the use of a multistart scheme. Here, one defines multiple starting points (θ_0^i) for the gradient ascent/descent with the goal to increase the chance that at least some of the trajectories through the free energy landscape (as prescribed by Eq. (3.12)), do not get stuck in local optima. Since the parametric form of F is independent of the starting value, one can then post-hoc compare the results from the different starting values, and compare the heights of the found peaks in order to find globally better solution. Obviously, there is no way to know a suitable number of starting points in general, and the optimal number will depend strongly on the given problem. Additionally, if one would like to equally vary the starting values of all parameters one becomes a victim of the curse of dimensionality – the dimensionality of the problem increases exponentially with each parameter.

3.3 METHODS

The introduction ended on the proposition of a multistart in order to diagnose and, at least partially, overcome the expected non-convex shape of the free energy landscape. In order to assess to what extent it affects parameter estimation and model selection, we ran a simple convolution based, two region DCM (based on the three population model), with different modulation structures (David, Kiebel et al. 2006). Models were chosen with increasing complexity, starting with a null model (*m01*), a modulation of a forward connection (*m04*), modulation of the forward connection and the excitability of the second region (*m11*), additional modulation of the backward connection (*m15*) and a model where all connections are modulated by condition (*m16*) (**Figure 10**). Please note that we kept the model coding consistent with the empirical rodent study (RATMPI) in this thesis. We created N=20 synthetic datasets by sampling from some pre-set distribution (see **Table 5**), where many parameters were sampled form the priors, except modulation parameters (B), the baseline-connectivity (A) and the leadfield parameters. The motivation behind the exact choice of posterior distributions was five-fold:

1. Creating strong baseline connections between the regions such that the effect of the driving input would propagate well through the system.
2. Inducing fairly (strong positive) forward modulatory parameters to avoid sampling low effects that might be misidentified as noise. The directionality of effects was also partly motivated by the results of the RATMPI study (Chapter 5).
3. Inducing a downregulation of local excitability to avoid entering regimes of over-excitability

4. Increasing the leadfield gain of the second region to allow for a similar signal-to-noise ratio across the two regions.
5. Having models of increasing complexity to investigate both, the extreme ends of complexity, and models of average complexity.

Parameter	**Effect**	**Sampling**
T	Kernel Decay	
G	Kernel Gain	
S	Sigmoid Transform	$\frac{1}{2}\mathcal{N}(0,\frac{1}{4})$
H	Intrinsic connectivity	
D	Delay	
R	(Driving) Input Shape	
C	(Driving) Input Gain	$\frac{1}{2}\mathcal{N}(0,\sqrt{\frac{1}{32}})$
J	Population Gain[10]	
L	Leadfield Gain	$\begin{bmatrix}2\\10\end{bmatrix} + \mathcal{N}(0,\sqrt{\frac{1}{10}})$
A forward	Forward connection (extrinsic)	$1 + \mathcal{N}(0,\frac{1}{20})$
A backward	Backward connection (extrinsic)	
B	Modulation	$\begin{pmatrix}-0.5 & 0.75\\0.75 & -0.5\end{pmatrix} + 0.1 \cdot \mathcal{N}\left(0,\sqrt{\frac{1}{8}}\right)$
Noise	Noise (Hyperprior)	7

Table 5| Setting for simulation of synthetic data. Only parameters with non-zero prior variance sampled. Depending on the model, certain modulations were excluded when generating synthetic data. $\mathcal{N}(\mu,\sigma)$ denotes a normal distribution with mean μ and standard deviation σ.

For each of these synthetic basis-datasets, we then simulated noisy data for all five model structures. Hence, for each simulation model, we had 20 simulations under different parameter settings, where the parameters (except the additional modulation) were equal across models. Synthetic noise was generated under the assumptions of an autoregressive (AR(1)) process (according to the standard noise model in SPM12) at a specified SNR of 7. All models were then inverted using a multistart approach. Here, we sampled 100 starting values from the priors (except modulatory parameters)

[10] Only *stellate cell* and *inhibitory cell* voltage state related values are sampled. *Pyramidal cell* gain is fixed to 1.

$$\theta^i_0 \sim p(\theta), \quad i = 1, \ldots, N \tag{3.14}$$

and inverted each dataset under each model (also the ones not generating the true data), resulting in $N \cdot nModels^2 \cdot nStartingValues = 50000$ inversions. Apart from the multistart, the inversion was performed under default settings using the LDE-integration scheme (see chapter on Delay Differential Integrators).

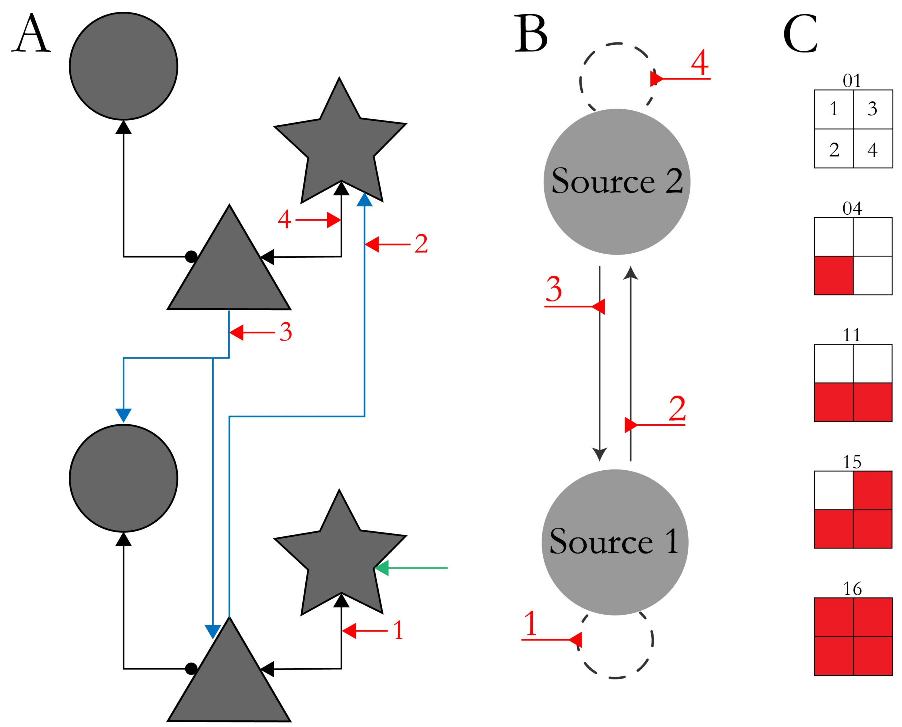

Figure 10| Connectivity structure and model space for simulations. A) Microscopic connectivity pattern. Stellate cells depicted as stars, pyramidal cells as triangles and inhibitory cells as circles. Arrowheads depict excitatory, circles inhibitory connections. Driving input shown in green, modulatory input in red (numbers correspond to the entry in the modulatory matrix (C). B) macroscopic view of the connectivity. C) Model space. Model code chosen in accordance to empirical study (Chapter 5).

3.4 RESULTS

Synthetic data were generated according to **Table 5** and inverted using a multistart procedure, starting from 100 starting values randomly sampled from the prior (starting values for modulatory parameters were kept at the prior mean). This is consistent with the way the empirical data was analysed. However, not sampling modulatory parameters could be a potential limitation and will be discussed in the *Discussion*. All inversion results were collected and compared between default starting (referred to as 'default') values and the starting values resulting in the highest negative free energy ('best') in terms of:

- Model Recovery (Confusion Matrix)
    - Each parameter-set and simulation model was inverted using **all** inversion models. The confusion matrix then displays the frequency of how often (for each simulating model) a particular inversion model scored highest in terms of negative free energy. Results are reported as Balanced Accuracies (BA), calculated as the average recovery of the true model[11].
- Model Recovery (fixed effects BMS)
    - Summed negative free energy (over simulations) as used in a fixed effects group level Bayesian model comparison. In the literature, $F_A - F_B = \Delta F > 3$ (Kass and Raftery 1995) indicates strong evidence of model A being more likely to have generated the data than model B (assuming all *subjects* have been using the same model).
- Parameter Recovery
    - Recovery of the true, generating modulatory parameter values (MAP estimates used as point estimates).
- Parameter correlations
    - Average posterior correlations of the parameters to visualize pairwise dependencies between parameter. Posterior covariances are transformed into correlations, Fishers' z-transformed, averaged (across simulations) and transformed back.
- Visualization of a multimodal landscape
    - Negative free energies over all starting values for a given simulation and model
    - Trajectories of the modulatory parameters over all multistarts
    - MAP estimates over multistarts

We first briefly present the results based on the default starting values and the results of the starting values that resulted in the highest free energy before formally comparing the differences and discussing their implications.

---

[11] Please note that as all groups are of the same size, the "accuracy" is inherently balanced.

The results for the default starting values are summarized in **Figure 11** in terms of recovery of modulatory parameters, model recovery and average posterior parameter correlations. Modulatory parameters are generally underestimated for all models with more than one modulation. There are at least two reasons for this. First, it could be a general effect of shrinkage priors, pulling posterior estimates towards the prior mean (of zero). This is also desired, as enough evidence in the data is needed (manifested by the ability to fit the data better and thus increase model accuracy) for a parameter to deviate strongly from the prior. Hence, it is an intended constraint. Second, the chosen modulation structure of the simulated data has two forces acting against each other – an increase in activity in the second region mediated via an increase in forward connection, and a decrease mediated via decreased excitability in the second region. Since they are counteracting, it is possible that any ratio of the two effects lead to a similar differential effects (raw amplitude could be taken up by other parameters, e.g. leadfield). This is what seems to be the case throughout models *m11, m15* and *m16*. Importantly, the forward and backward modulations almost vanish in *m16*, indicating that the features of the data can be taken into account by only local excitability. The strong dependence between extrinsic and intrinsic (gain) modulations[12] is quantitatively shown by their strong (anti)-correlation in **Figure 11**C. **Figure 11**B shows model recovery in terms of a confusion matrix. In total, 50% of the modulation structures were correctly identified, i.e. 10/20, 9/20, 5/20, 6/20 and 20/20 simulations for the five models. Please note that the models are arranged in order of increasing complexity. Therefore, the upper-triangular shape indicates a tendency to select more complex models (overall, 39% were identified as more complex, 11% as less complex models than the true model). This topic will be revisited in the next chapter (Chapter 4) of this thesis.

[12] We use the notion of *intrinsic* and *gain* modulation synonymously. Generally, intrinsic modulations act as a modulation on the excitatory gain of all populations in a region.

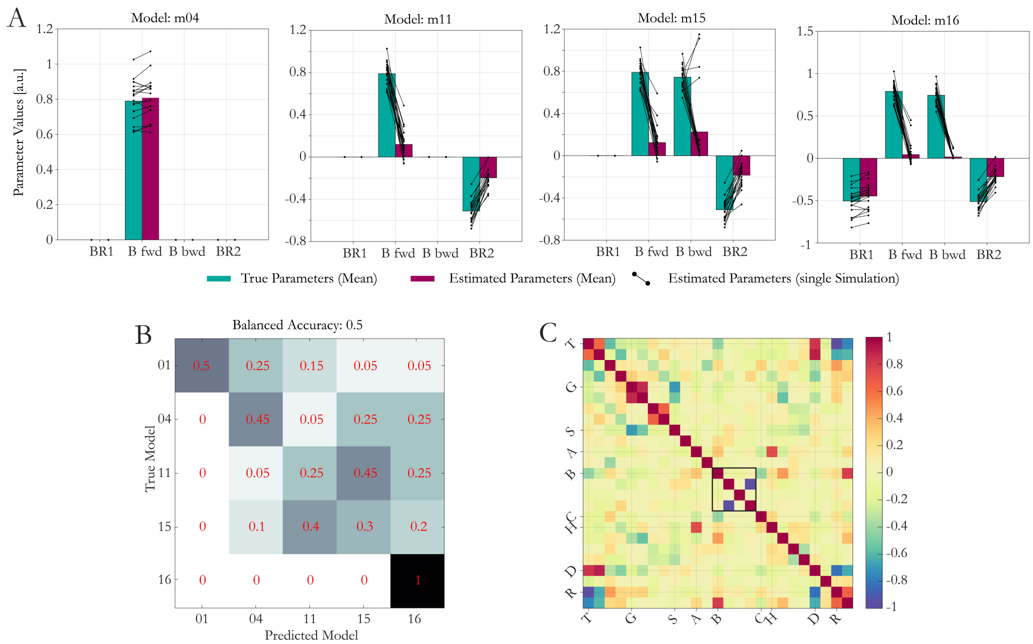

Figure 11| Model and parameter recovery for the inversion with default starting values (prior mean). A) Modulation parameter recovery for inversions with the true models (m01, m04, m11, m15, m16). Bars depict average parameter values over simulations. Black lines depict single simulation recovery. Color coding according to legend. B) Confusion matrix (winner takes all). Probabilities of 0.05 correspond to classification of a single simulation. C) Average (over simulations) posterior correlation between parameters for m16. Black Box highlights correlations between modulation parameters.

The inversion results when starting from the starting value resulting in the highest negative free energy are depicted in **Figure 12**. While showing the same underestimation of modulation strengths, parameters are – on average – recovered much better for all models, despite a similar average correlation between intrinsic and extrinsic modulations. The confusion matrix shows a model recovery of 60% (12/20, 9/20, 6/20, 13/20 and 20/20) for the five models. Put simply, model recovery was at least as good as with the default starting value. Again, we observed the same trend to favour more complex models; only 3% of all classifications were in favour of simpler, 37% of more complex models.

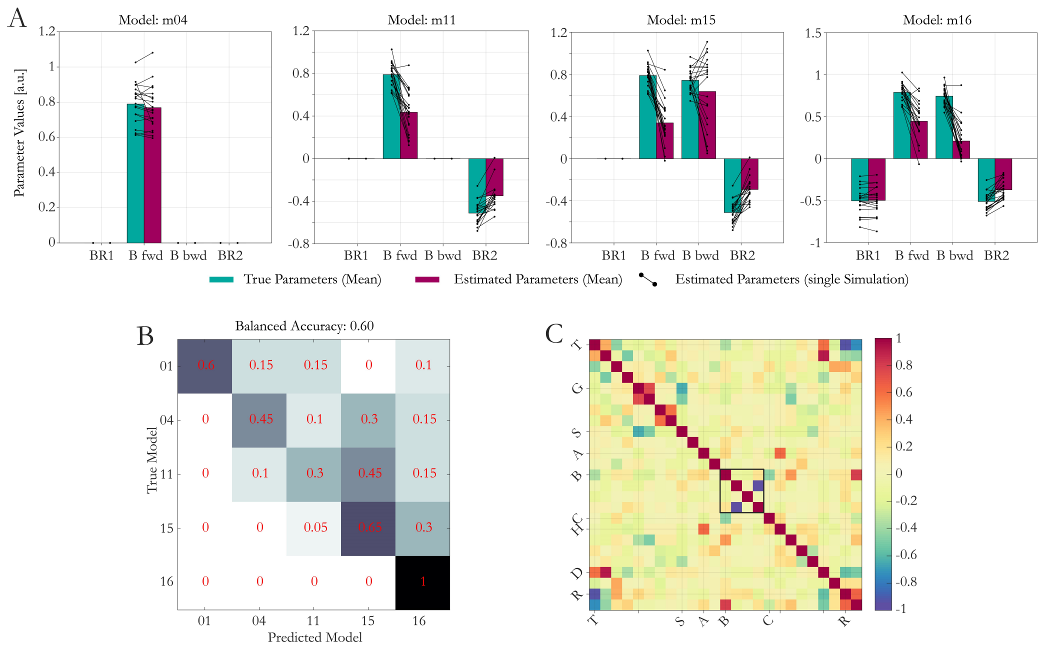

Figure 12| Model and parameter recovery for the inversion with starting values, which resulted in the highest negative free energy. A) Modulation parameter recovery for inversions with the true models (m01, m04, m11, m15, m16). Bars depict average parameter values over simulations. Black lines depict single simulation recovery. Color coding according to legend. B) Confusion matrix (winner takes all). Probabilities of 0.05 correspond to classification of a single simulation. C) Average (over simulations) posterior correlation between parameters for m16. Black Box highlights correlations between modulation parameters.

We also directly compared the values of the negative free energies in the spirit of a fixed effects, Bayesian model comparison. Fixed effects BMS is a commonly used measure to infer on model structure by addressing the question, which model is the most likely to have generated the (group) data, when assuming all subjects come from the same model (**Figure 13**, **Table 6**, FFX) (Stephan, Penny et al. 2009). Hence, if one had used fixed effects BMS to infer on model structure for the default VB starting values, the correct conclusion would have been drawn for only the two most complex models. If one had used the highest negative free energies as identified by the multistart, the simulating model would have been identified in four out of five cases (highest negative free energies shown in bold). By definition, the free energies of the default starting values are lower (or equal) than for the best starting values. Note that for these simulations, there was always a set of parameters leading to a better optimum than was reached with the default starting values. This finding does not automatically generalize to all settings, but since we are dealing with fairly simple models here, it is very likely that this will be generally the case.

We also considered a value that is independent of the exact shape of the likelihood function, the explained variance (vE), to assess model goodness. Variance explained through the ratio between residual variance and data variance ($vE = 1 - \frac{var(e_y)}{var(y)}$) measures purely the goodness of model fit. For our simulations, the true explained variances were fixed at 94%. **Table 6** shows that the best starting values predict the same vE as the true vE, for all models with the same or higher complexity as the true model (with one exception). This indicates that on average, the models with enough flexibility were able to correctly distinguish noise from signal. The results for the default starting values do not lag too much behind. It always performs slightly worse, but we would argue that the fit is definitely in a range, where one could not intuit a suboptimal solution.

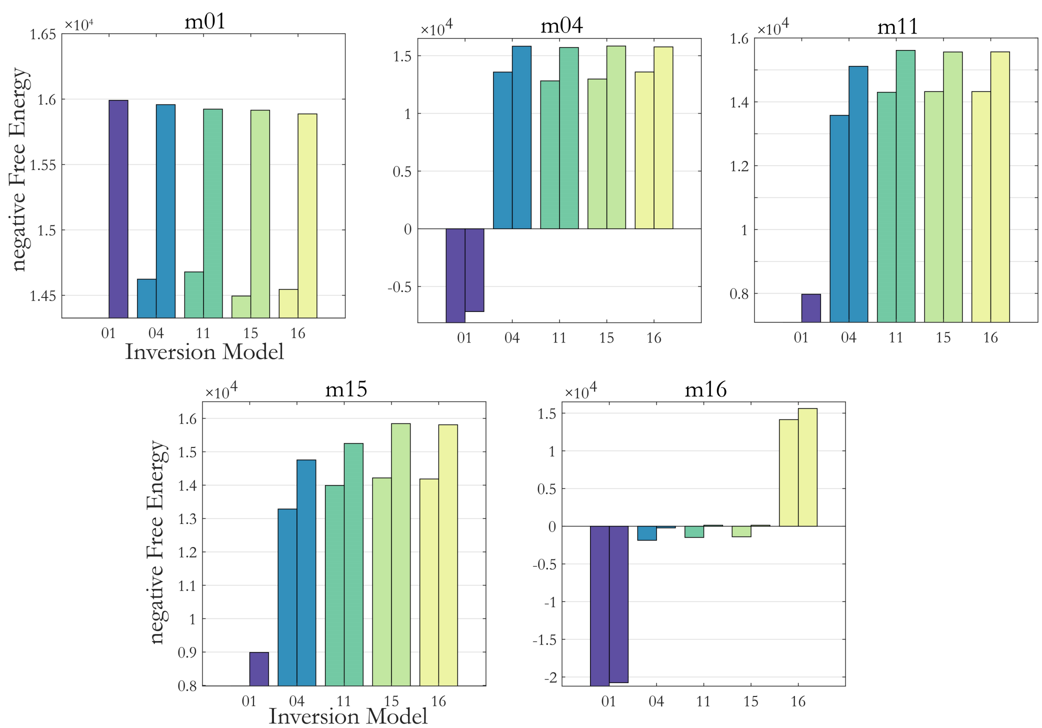

Figure 13| Comparison of fixed effects BMS for all synthetic datasets, simulation (shown in the five panels) and inversion (shown on the Y-axis of each panel) models. Left bars (of each pair of bars) correspond to result from default, right bars to results from the starting values resulting in the highest negative free Energy. Colour coding indicates inversion model. The lower limit of the Y-Axes are set to the lowest value of the negative free energy for each simulating model.

		DEFAULT STARTING VALUES					BEST STARTING VALUES				
		inversion model					**inversion model**				
		m01	**m04**	**m11**	**m15**	**m16**	**m01**	**m04**	**m11**	**m15**	**m16**
FFX	**simulation model** **m01**	14327	14623	**14679**	14495	14545	**15991**	15958	15924	15915	15887
	m04	-8136	13583	12819	12979	**13589**	-7173	15820	15710	**15837**	15764
	m11	7093	13579	14299	**14322**	**14322**	7972	15113	**15613**	15563	15566
	m15	7982	13285	13990	**14218**	14185	8988	14755	15250	**15844**	15809
	m16	-21213	-1869	-1487	-1401	**14144**	-20756	-221	139	135	**15617**
vE	**simulation model** **m01**	0.969	0.971	0.971	0.970	0.971	0.974	0.974	0.974	0.974	0.974
	m04	0.852	0.967	0.964	0.965	0.967	0.857	0.974	0.973	0.974	0.974
	m11	0.939	0.967	0.970	0.970	0.970	0.942	0.972	0.974	0.974	0.974
	m15	0.943	0.966	0.968	0.969	0.969	0.946	0.970	0.972	0.974	0.974
	m16	0.726	0.892	0.894	0.894	0.970	0.729	0.900	0.902	0.903	0.974

Table 6| Summary of inversion results in terms of fixed effect BMS (FFX) and average explained variance as predicted by the models.

In an attempt to visualize important aspects of the free energy landscape, we ran a series of additional diagnostics. First, in **Figure 14**A, one can see the large amount of variation in posterior means over the 100 starting values. This is worrisome if one is keen on robust estimation of all parameters but we would expect that usually, the focus lies on the modulatory parameters. There is arguably also large variation found for condition specific effects (i.e. modulatory parameters *B*) but there appears to be a trend that the sign is estimated correctly. **Figure 14**B specifically illustrates the walk of the modulatory parameters through the free energy landscape. There, the size of the circle at the endpoint is in relation to the free energy values for a particular solution. Of course, this is merely a projection of all parameter onto the small subspace spanned by two modulatory parameters, hence trajectories that end close to each other can be very different in terms of negative free energy. We hand-picked an example that nicely illustrates the utility of the multistart, where one of the trajectories ends very close to the simulated values (marked as 'x'). This, of course, is not necessarily the case (e.g. correlations, shrinkage priors, local extrema) and in fact, as **Figure 12**A indicated, at least extrinsic modulation were generally underestimated. In summary, looking at the diagnostics in **Figure 12** illustrates the following: While **Figure 12**A showed convincingly that the multistart is highly beneficial in terms of both model and parameter recovery, the free energy landscape appears to be very rough. This intuition about roughness can also be seen in **Figure 15**. Here we illustrate the changes in negative free energy as we changed the two modulatory parameters at the posterior mean (dataset 22, generating model $m11$, inversion model $m11$). All other parameters were kept fixed. There are two things we would like to point at: (i) The plateau like structure indicating a correlation between the two parameters; (ii) A very sharp decrease in negative free energy of up to $\Delta F = 1500$ when moving away from the ridge. These arguably large differences in negative free energy will be the topic of the next chapter.

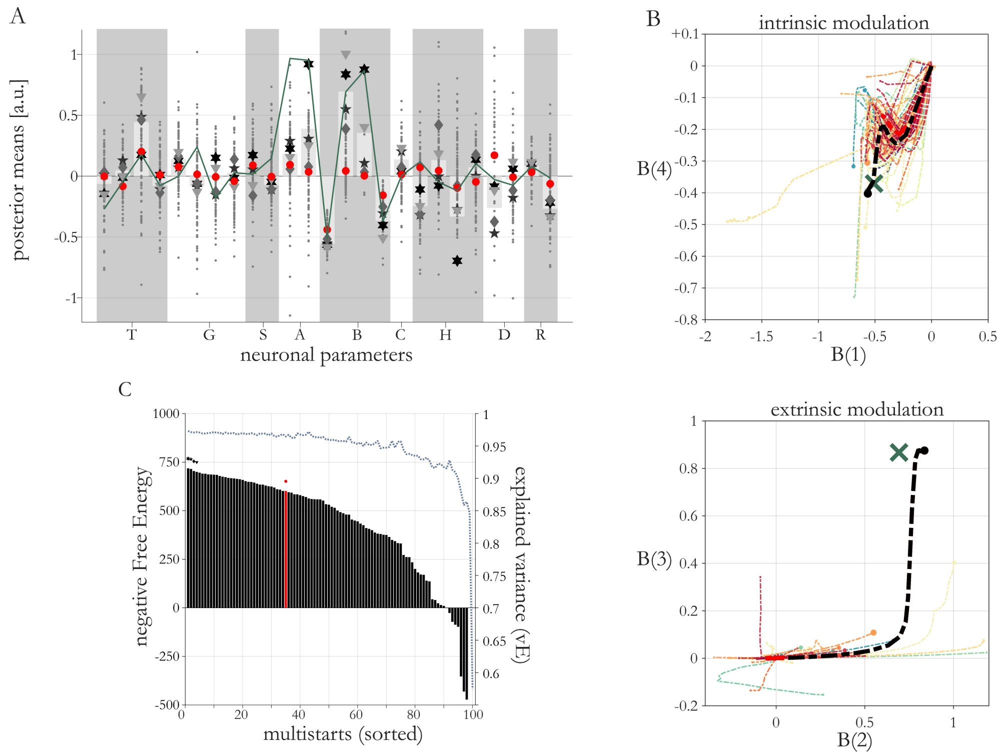

Figure 14 | Multistart diagnostics for a single inversion (dataset 32, simulation model 16, inversion model 16). A) Posterior means of all neuronal parameters over the 100 starting values. The four starting values resulting in the highest negative free energy marked as gray stars, the default starting value as a red circle. All means resulting from other starting values indicated by gray dots (some dots could be outside of the plot-range). The green line indicates the true parameters. B) Trajectory of the modulatory parameters over the course of the optimization. Numbers of B parameters (in brackets) according to **Figure 10**. Black line indicated best inversion, red line inversion from the default starting value. Size of end-circles relate to the negative free energy values. C) Negative free energy (black bars) and explained variance (dotted line) over all starting values of the multistart. Results are sorted in order of decreasing negative free energy. Red bar indicates default starting value.

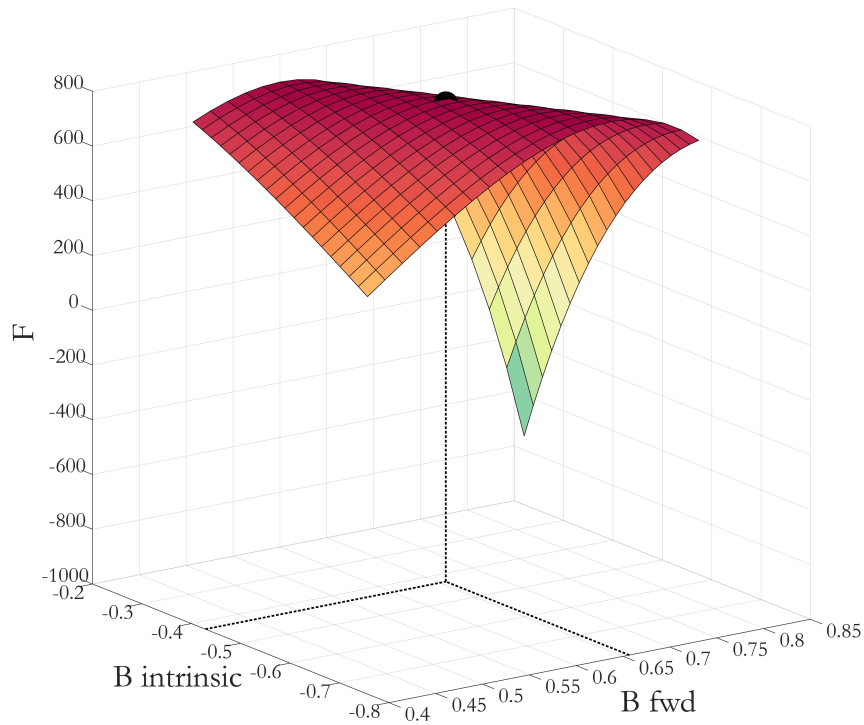

Figure 15| Sensitivity analysis with respect to changes in modulatory parameters around the posterior mean (maximum of the negative free energy). Parameters were changed around the posterior mean $\pm$ 1/16 (i.e. 50% of the standard deviation of the prior). Dotted lines show the posterior mean. All other parameters were kept fixed.

3.5 DISCUSSION

In this chapter, we have investigated the performance of a multistart VBL scheme, by comparing results acquired from the default starting value and the starting value leading to the highest negative free energy ('best'). The focus of investigation was whether conclusions drawn from an inversion (either in terms of network identification or parameter estimates) are dependent on the starting values. From a theoretical perspective, there could be at least three reasons for this to occur. First, the landscape is truly multimodal. Then a gradient ascent based optimization scheme is known to get stuck in local optima. Second, the stopping criteria of the optimization is prematurely satisfied. This could occur, if the landscape presented with very slowly ascending ridges and the optimization does not finish within the maximum number of pre-defined steps, or updates get too small in magnitude and an optimum is erroneously assumed. Third, the Laplace approximation underlying the negative free energy is invalid, i.e. local properties of the objective functions are badly approximated.

As a disclaimer, we want to point out that we have specified a very tough optimization problem. We deliberately set fairly large effect sizes on modulation and connection strength. That of course forces those parameters to deviate relatively strongly from the prior mean, which in the presence of local maxima, can be challenging for the optimization algorithm. Additionally, we wanted to create a real world test-bed, which included inferring on almost all parameters of the network (and which is in line with the default prior setting in SPM12, ver. 6906).

Overall, our results indicated the presence of at least the first two problems. The multistart diagnostics shows the presence of multiple posterior neighborhoods. Whether they occurred due to satisfaction of the convergence criteria, or correspond to true local maxima is difficult to assess. But a previous study, independent of DCM, has made a point of true local maxima being rare in high dimensional optimization problems (Dauphin, Pascanu et al. 2014, Pascanu, Dauphin et al. 2014). There, the argument is that a high dimensional random matrix is very unlikely to have all negative eigenvalues (which would be needed for a local maximum). We would agree with that intuition in the case of DCM, simply because of the pattern of scattering of the found maxima in **Figure 14**. There are known limitations of some gradient descent based methods not being able to overcome saddle-points (as they can act as attractors). We explicitly checked the behavior of our method around saddle points and found that it is able to escape them (**Figure 16**). This suggests that the used gradient ascent scheme is at least in principle able to find its way out of saddle points. However, stopping criteria might be critical here and could potentially cause the algorithm to stop optimization too early around a saddle point. Another critical aspect is the speed (adjustments of the stepsize) for very flat landscapes. There have been multiple advances in *momentum based* gradient ascent methods that utilize history of updates or local hessian information with the explicit goal of increasing update speed in flat regimes (e.g. (Jin, Netrapalli et al. 2017)). However, also note that all optimizations in **Figure 14** did converge for all starting values within 300 steps. We did not explicitly investigate the validity of the Laplace approximation, but two other studies have critically looked at it (Daunizeau, David et al. 2011, Penny and Sengupta 2016). Daunizeau et al (2011) illustrated nicely the potential effect of the Laplace approximation to non-gaussian posteriors (see Fig. 3 in their paper). Our intuition would commend that these situations will occur, simply because of the way parameter dependencies shape the landscape. Both **Figure 14**A and **Figure 14**B do not indicate clustering of solutions, neither in terms of negative free energy nor posterior means one might have hoped for. For example, this could have become visible by two distinct clusters of posterior means that would correspond to more clearly separated negative free energy regimes. However, solutions seem to lie on an extended, connected submanifold. Penny et al. (2016) formally compared the a multistart VB against a sampling method (MCMC), where the latter is independent of the Laplace approximation, and MCMC is guaranteed to converge to the true posterior (at least in the limit of infinite samples). Their results show that MCMC indeed outperforms VB (even under a multistart routine), as both, log-model evidence values increased and parameters were recovered better. However, MCMC is computationally much more demanding, especially if approximations to the log model evidence are needed. Then it requires samples from multiple chains at different temperatures (thermodynamic integration, see (Aponte, Raman et al. 2018))

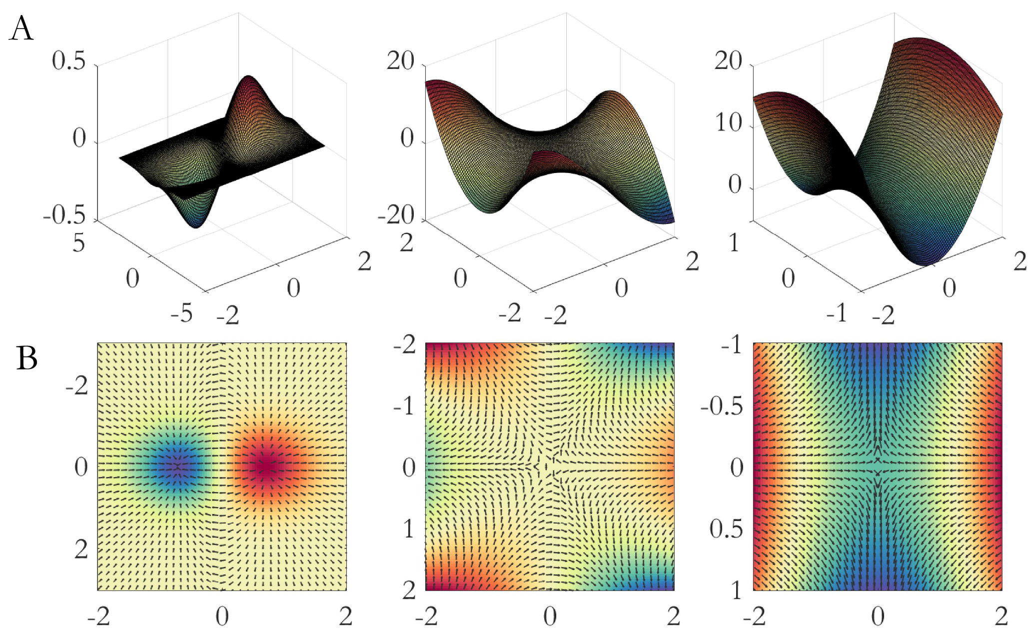

Figure 16| Behavior of the gradient descent method around different sorts of critical points in two dimensions. A) Landscape plotted in 3D. B) Heatplot of the landscape. Arrows indicate the direction of the vector-field induced by the gradient step (according to the Ozaki-update). Blue areas act as attractors.

Of course, optimization schemes do not end with VBL and MCMC. One further alternative, also from the family of global optimization schemes, are Gaussian processes (GPs) (Rasmussen 2003). GPs are very efficient due to their mathematical formulation. Here, one would consider the landscape as an instantiation of a random, multivariate Gaussian process. From sampled points (i.e. evaluated free energies for some parameters), the GP prescribes where to sample next in the search for the highest point, which is usually a balance between already sampled high points, and points of the landscape, where there is high uncertainty (Snoek, Larochelle et al. 2012). It has been shown that GPs can be very useful also in the context of DCMs for fMRI (Lomakina, Paliwal et al. 2015). However, in this particular context of DCM for ERPs, GPs might not be the most suitable method. They tend to scale badly to high dimensions (Moriconi, Kumar et al. 2019) and would likely also involve the Laplace approximation. But it is definitely a worthy consideration for the future.

We have used arguably very few multistarts in comparison to the dimensionality of the parameter space (about 0.01%) which by no means can be considered exhaustive. Hence, diagnostics and results only refer to a scarce representation of the full optimization space. Despite the small number, the multistart proved useful in model recovery. This should not suggest that the results from the default starting value are at all bad. Models were recovered with a respectable balanced accuracy of 50% (default) and 60% (best) respectively. From a fixed effects BMS perspective, 2/5 (default) and 4/5 (best) models would have been recovered. However, it is important to note that 'truth' is a difficult concept, even for simulations. We would not have expected a perfect recovery of models to begin with. In the presence of correlations between the parameters, the contribution of the determinant of the posterior covariance (Eq. (3.11)) is not easy to intuit. In fact, it is possible that there are cases, where a larger set of parameters achieve a similar fit while scoring lower in KL-divergence, simply because either the overall correlation is lower, or because the posterior means are closer to the prior mean. We think this partly explains the tendency in favour of more complex models. However, the whole next chapter will be devoted to this investigation.

In order to get some benchmark for the model recovery results, we ran an additional inversion for the true models (diagonal in the confusion matrix), by setting the simulating parameters as the starting values. This gave us an indication, whether the true parameters were in a stable neighborhood (i.e. local/global maximum) and an estimate of the free energy at these points. Obviously, this can only be done for synthetic data where ground truth is known. The results showed that true parameters were locally stable, and when using the resulting free energy estimates (for the diagonal of the confusion matrix and all the 'best' starting values for all other models as in **Figure 12**), balanced accuracy increased to 73% (as opposed to 60%). Hence, the true parameters do correspond to higher points in the negative free energy landscape and the multistart as used here, performed around 13% below benchmark. Fixed effects BMS recovered all models, i.e. the sum of negative free energies was always highest for the true, simulating model. These discrepancies could potentially be resolved by increasing the number of multistarts or by also sampling modulatory parameters for the starting values.

For parameter recovery, the benefit of the multistart was similar. For both, default and best starting values, parameters tended to be underestimated relative to the true parameters, which is expected for a local optimum reached with a trajectory starting from the prior mean, and given the effect of shrinkage priors. Nonetheless, all optimizations resulted in very good fits of the data (see, **Table 6,** vE). Combined, this indicates that dependencies between parameters allowed for very accurate predictions, even with a reduced magnitude in modulatory parameter values. Luckily, with only few exceptions, the signs of the modulatory parameters were estimated correctly, independently of whether the best or the default starting values were used. In terms of absolute connection strengths, the multistart outperformed the default starting value, which is most likely due to its ability to overcome local critical points introduced by the strong correlation between modulation in the *forward* and *intrinsic* connection (**Figure 11**C and **Figure 12**C). As to why there seems to be a tendency to attribute the modulation rather to intrinsic/gain than to extrinsic modulation, we can only speculate (at least for the default starting value). But it could be due to the gradients of the system to parameter changes around the prior mean; If the system's response to a change in intrinsic modulation is higher than to extrinsic, the closest critical point might simply lie along that sub-dimension. Nonetheless, as mentioned earlier, the true values do constitute a stable optimum, hence by increasing the number of starting values and/or sample modulatory parameters, there is a chance that one could improve parameter recovery.

There is one potentially interesting side remark to be made here. There is a possibility that the 'closest local optimum' to the prior mean can act as an artificial constraint. In other words, parameter estimates might be more comparable across subjects, which could prove useful for classification or statistics (see Supplementary of empirical RATMPI study in Chapter 5). In fact, other optimization settings use a similar concept called 'early-stopping', where it is used to prevent overfitting explicitly in combination with gradient based methods (Caruana, Lawrence et al. 2001). While it is being investigated in neural-net optimization where one is arguably facing a much larger set of parameters it shows that such a property might be useful for regularization and might become an important point to consider for DCM, especially when taking the results from the next chapter into account.

To summarize, the beauty of the multistart is manifold. First, it is very easily parallelizable. Second, it allows for diagnostics of the free energy landscape, always keeping in mind that the diagnostics only refer to a very scarce part of the optimization space. Third, it stays in the optimization framework of the variational free energy under Laplace approximation, which is very fast and efficient. Fourth, because of the assumption that $q(\theta)$ is normal, the posterior means are given directly by maximum a posteriori (MAP) estimates and can serve as point estimates which can be used for subsequent statistical tests.

Therefore, while it might not be the most elegant solution, the multistart clearly helps for inference on model structure and parameter estimation for an optimization scheme relying on the Laplace approximation. We do however see a need to say a word of caution about the use of the traditional benchmarks for model comparison (e.g. $\Delta F > 3$ (Raftery 1995)). We observed differences on the order of 10^3 between 'default' and 'best' starting value, which makes the superiority of the multistart evident (**Figure 13**). However, there is one obvious question that arises: Are free energy differences slightly above the classical threshold a trustworthy statistical result, if free energy differences between different starting values can be much higher. Differences in the observed orders are – for lack of a better word – dramatic. This is surprising since there is no large difference in terms of model fit measured in variance explained which is independent of the parametric form of the free energy (**Table 6**). However, the ranking of the models in **Figure 14**C seems to follow the ranking of vE with an, at this point unknown quantitative relationship. These results beg the question, where this remarkable difference comes from, whether it is stable wrt. the values of the hyperparameters and/or the parametrization of the accuracy. The next chapter will be devoted to exactly these points.

And we also see a need to communicate that interpretation of non-modulatory parameters deserves caution. At least, when posterior means are taken as point estimates. There are simply too many model inherent correlations between parameters when running under a standard inference setup (where almost all parameters are inferred on simultaneously). Ways around the correlations is probably an even tougher problem and we can only suggest some thoughts. One option is to constrain parameters further, either through highly informed priors, or analytical methods (e.g. Bayesian Model Reduction (Friston, Litvak et al. 2016)). Another option could be the use of clever designs, where all parameters of interest can be cast as a problem of condition specific effects (see Supplementary material of the RATMPI study as an example). Its generalization would be formally constraining parameters across multiple datasets (i.e. hierarchical approaches (Friston, Zeidman et al. 2015, Friston, Litvak et al. 2016, Yao, Raman et al. 2018)). Or finally, comparing full posterior distributions in a Bayesian fashion, by taking the correlation structure into account. However, it is questionable whether the approach is valid for very peculiar correlation structures only poorly approximated under the Laplace approximation. None of these ideas are novel in the sense that they have not been used in other fields. But given the empirical and in silico evidence (empirical evidence is provided in all empirical DCM studies in this thesis), we think that the methods and careful diagnostics as presented above should be included as standard routines when hypotheses want to be rigorously tested. This will also refer to the Figures shown in the next chapter, but generally, the diagnostics suggested are:

- negative Free energy over multistart (e.g. **Figure 14**C)
- posterior means over multistart (e.g. **Figure 14**A)
- comparison between the best and the default starting value (e.g. **Figure 13**)
- relative contribution of the terms resulting in the negative free energy (see next chapter)

4| HYPERPRIOR SELECTION IN DCM

4.1 DISCLAIMER

These methodological developments and analyses were done under the supervision of Jakob Heinzle and Klaas Enno Stephan. Stefan Frässle, Yu Yao and Eduardo Aponte consulted in constructive discussions

4.2 GLOSSARY/TERMINOLOGY

We provide a brief glossary to reduce the dependency on the previous chapter.

multistart:

Running a single model inversion from multiple starting values. (Here, we used 100 starting values sampled from the prior. Modulatory parameters were kept at their default starting values, i.e. absence of modulation).

BEST (starting value / inversion):

Inversion starting from the value resulting in the highest negative free energy. We will use the all-caps notation to make it clear that BEST in the given context refers to the results based on these starting values.

DEFAULT (starting value / inversion):

Inversion starting from the default starting value (prior mean). We will use the all-caps notation to make it clear that DEFAULT in the given context refers to the results based on these starting values.

posterior noise precision (Eh):

Either referring to the full posterior distribution or the posterior mean, depending on the context. We use noise- and hyperparameter interchangeably.

prior noise precision (hE, hC):

Prior noise distribution with mean (hE) and variance (hC) as sufficient statistics. The distribution is defined in log-space.

modality	**specification (SPM12, ver. 6906)**	**prior mean**	**prior variance**	**approximate SNR**
fMRI	spm_dcm_estimate.m	6	1/128	20[13]
EEG	spm_dcm_erp.m	6	1/128	20

SNR: (Penny 2012)

Signal to noise ratio. Related to variance explained as $vE = \frac{1}{1+\left(\frac{1}{SNR}\right)^2}$.

Proportion of explained variance (vE):

Equivalent to R^2. Calculated as $vE = \frac{var(y_p)}{var(y_p)+var(e_y)}$. Here, y_p denotes the true signal, e_y denotes the residuals. Alternatively, $vE = 1 - \frac{var(e_y)}{var(y_p)+var(e_y)}$.

4.3 INTRODUCTION

In the previous chapter, we have shown the utility of the multistart in simulations. Starting the gradient descent based inversion from multiple starting values clearly showed benefits in terms of model identification and parameter recovery. We have shown model

[13] The data in fMRI is not normalized to unit variance (i.e. $var(y) \neq 1$). Therefore, this is a rough approximation.

identification in terms of a standard fixed effects Bayesian Model Comparison (FFX BMS) and a confusion matrix with the associated balanced accuracy (BA).

However, there were two striking and potentially worrisome observations that we made. First, there was a clear tendency in favor of more complex models in the model recovery analysis. Second, free energy differences between different inversions, presumably ending in different local minima, vastly exceeded the conventional thresholds. (Typically, log-Bayes factor differences of three indicate strong evidence in favor of a model compared to another model). In our simulations, log-Bayes factors often amounted to values in the double or triple digits on the single-subject and group level, respectively. Observation of these large differences is not restricted to our analyses. We checked five recent publications (PubMed search; keywords: *effective connectivity, DCM, ERP*) where DCM for ERP was used and fixed effects BMS results were reported (Youssofzadeh, Prasad et al. 2015, Ranlund, Adams et al. 2016, Díez, Ranlund et al. 2017, Penny, Iglesias-Fuster et al. 2018, Chen, Kuo et al. 2019). Average (across subjects) log-Bayes factors were on the order of $\Delta F(best - runner\ up) = 1 - 100$ (two studies looked at two groups)[14]. It is easy to see that for any moderately large group of participants, the evidence in favor of the winning model is very strong. Obviously, the studies differed in research question and model specification. While such large differences may or may not represent veridical extremely strong evidence for one of the models, they are difficult to interpret given our previous findings suggesting that such large differences may simply be caused by different starting values of the gradient ascent. If that is the case, it also begs the question how representative model comparisons are.

Interestingly, these large numbers seemed disproportionate to a measure of pure model fit, the explained variance (vE), which is independent of the parametric form of the negative free energy (F). Explained variance is typically not the primary quantity of interest in Bayesian models, but it conveys information about model fit that can be interpreted in absolute values. In the Bayesian framework, model fit has to be understood in a probabilistic sense through the accuracy term in the negative free energy. Here, observed datapoints are more or less probable, given a prediction and the assumptions about noise (or irreducible variance). These assumptions are incorporated in the noise model, which is parametrized by the hyper- or noise parameters.

[14] We averaged differences in group-level log-Bayes factors to make them comparable.

Hyperparameter optimization is an omnipresent challenge in machine learning, as they tend to have a large impact on the performance of an algorithm and have been identified as one of the confounds that affect the reproducibility of results (Henderson, Islam et al. 2018). They scale and shape the overall optimization landscape, which might prove to be pivotal when interpreting model goodness by a quantity that consists of multiple terms with different dependency on parameters/data. Finding a suitable balance between parameters and data is not something exclusive to the free energy framework. As we have shown in the previous chapter, F prescribes a very natural balance for this tradeoff. Here, the weight for the data enters in the form of the accuracy, whereas the weight of parameters enters in the form of the Kullback-Leibler-divergence (KL-divergence) between the approximate posterior and the prior. The KL-divergence (under the Laplace approximation) consists of three parts; The probability of the posterior mean under the prior distribution, the determinant of the prior (co)-variance and the determinant of the posterior covariance. We have discussed the behavior of these three terms in the previous chapter and highlighted that they essentially serve to regularize model inversion. This is of course not the only way to exert regularization. Other methods, for example R*idge-* or *Lasso-Regression* also include penalty terms for estimates of parameters taking extreme values (Bishop 2006). These approaches are often used for inferring on point estimates (means) and for both aforementioned regularization methods, one can show that they correspond to a Bayesian case for a certain choice of prior. The motivation to use regularization is to control for overfitting. According to the Oxford dictionary (Dictionary), overfitting is defined as follows:

> "The production of an analysis which corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably."

The idea of the model generalizing to unseen data is thus an inherent property of the log model evidence (LME, or its approximation F). Other measures with the same purpose are the Akaike Information Criteria (AIC) and the Bayesian Information Criteria (BIC) (Akaike 1974, Schwarz 1978). The three measures are of the following forms:

$$\begin{aligned} LME &= \langle \log p(y\,|\,\theta) \rangle - KL[p(\theta\,|\,y) \,\|\, p(\theta)] \\ AIC &= \log p(y\,|\,\theta) - n_p \\ BIC &= \underbrace{\log p(y\,|\,\theta)}_{Accuracy} \underbrace{- \log(N) \cdot n_p / 2}_{Complexity} \end{aligned} \tag{4.1}$$

Here, $\langle \cdot \rangle$ denotes the expectation with respect to the posterior $p(\theta|y)$, n_p the number of parameters and N the total number of datapoints. One can see that all three measures incorporate some form of regularization. Unlike AIC and BIC, LME penalizes explicitly large covariances in the posterior distribution. Large posterior covariances (or correlations) are generally associated with some level of redundancy in one of the parameters, which are often one of the causes for overfitting. Unfortunately, it is difficult to have a quantitative understanding about the KL-divergence. Hence, while we will focus in this chapter on the LME (or better its lower bound approximation F), we will sometimes draw comparisons to AIC and BIC because there, quantitative changes to complexities are known (e.g. when introducing an additional parameter).

Obtaining a more quantitative understanding of this 'generalization' aspect of the LME is critical and has been motivated by the findings described in the previous chapter of this thesis. This is arguably even more relevant if LMEs should be used as a measurement device to delineate disease processes (e.g., via BMS as a formalization of differential diagnosis) where generalization across measurements and/or samples of patients is paramount. Additionally, as we have seen in the previous chapter, F (i.e. the approximation to the LME) and the posterior distribution over parameters are inherently interlinked. Hence, any (mal-) regularization in F could potentially translate to biased inference on parameters, which might affect the (biophysical) interpretation of those estimates.

In order to develop a quantitative understanding of these aspects, a fundamental aspect to note is that, while the LME in Eq. (4.1) describes a clear tradeoff between accuracy and complexity, this tradeoff ultimately depends on prior assumptions. In this chapter, we will show that the assumption and role of the noise in DCM can play a large role in the balance between these two components. We will derive three expressions; (i) the functional form of an exponential and a linear function of the noise precision, whose intersection point approximates the mean estimate of the posterior noise precision. Both functions are parametrized by the a priori assumptions about the noise and the empirical residuals. This first expression also results in two additional approximations: (ii) a direct relationship between the posterior noise precision and the negative free energy, which is linear when prior effects can be neglected and (iii) an approximation to the relationship between explained variance and posterior noise precision. These relationships allow us to explain, in a quantitative manner a range of phenomena observed in synthetic and real data, for DCM for both modalities. We will provide a brief summary of the implications already at this point:

As will be derived throughout this chapter, our approximation will result in the following (explicit) dependency between the posterior noise precision mean (Eh) and F:[15]

$$\begin{aligned} F(E_h) \; &= \left(\frac{N}{2}+\frac{1}{hC}\right)\cdot Eh - \frac{(Eh-hE)^2}{2\cdot hC} \\ &-\mathrm{KL}[q(\theta_{p,g})\,|\,p(\theta_{p,g})] \\ &+\text{const.} \end{aligned} \tag{4.2}$$

Here, we denote the prior noise precision mean (hE), the prior noise precision variance (hC). The distributions $q(\theta)$ and $p(\theta)$ denote the approximate posterior and the prior. Hence, the remaining KL-divergence is only with regard to neuronal parameters (θ_p) and parameters of the leadfield (θ_g).[16] We will also denote the second term in Eq. (4.2) the *precision weighted prediction error* (pwPE) term associated to the noise (see Eq. (4.18)). It is part of the noise KL-term. Importantly, everything in the 'const'-term in Eq. (4.2) is constant across models (or was ignored as part of the approximation).

From the expression in Eq. (1.2), one can derive another important quantity, namely, the expected improvements to F for changes in the posterior noise precision. This is interesting as it conveys information about the balance between accuracy and complexity within a single optimization, but also across models:

$$\frac{\partial F}{\partial E_h} = \left(\frac{N}{2}+\frac{1}{hC}\right) - \frac{(Eh-hE)}{hC} \tag{4.3}$$

Based on Eq. (4.2) and Eq. (4.3), we will outline in this chapter that the following intuitions can be gained:

1. The relationship between the posterior noise precision Eh and F contains a linear and a quadratic term (Eq. (4.2)).
2. These two terms are functions of the number of datapoints and the hyperprior mean and variance. If the (a priori) expected noise precision is too optimistic $(Eh < hE)$, all three contributions act by increasing the slope (Eq. (4.3)). This also implies that the

[15] Please note that Eq. (4.2) denotes the same tradeoff as in Eq. (4.1). We only made use of the parametrization of the accuracy in DCM and wrote the explicit dependence with respect to the mean of the posterior distribution of the noise parameter. The KL-divergence term here is the KL under Laplace approximation.
[16] We can only split the KL-divergence into the KL associated with neuronal/forward parameters and noise, because a mean-field approximation is used, rendering the noise parameter independent of the others.

noise pwPE term acts "against" parameter complexity.[17] The current values of the hyperpriors (SPM12, ver. 6906) for both fMRI and EEG assume little noise with high precision. These settings can cause bias for more complex models and overfitting, if the true SNR is below 20 (see Glossary)[18]

3. For high number of datapoints and high pwPE terms, it becomes crucial that Eh is accurately (and optimally) inferred (small inaccuracies in the estimate lead to relevant changes in F).

We will see that these five points have substantial implications for the optimization scheme when it comes to inferring model structure robustly and avoiding overfitting. To make a few illustrative examples outlining potential challenges (neglecting the noise pwPE term for now):

- Considering the situation of a multistart for a single model and dataset. In our current setup with N=2000 datapoints, differences in the estimates of Eh on the order of $\Delta Eh = 0.001$ result in differences in F on the order of $\Delta F = 1$.
- Similarly, considering a comparison between two models (m_0 vs m_1) with N=2000 datapoints. Adding a single parameter (e.g. modulation) would result in additional complexity. From the perspective of AIC and BIC, this added complexity would be on the order of $n_p = 1$, and $\frac{\log(n)p}{2} = 3.8$ respectively. (It could potentially be very different in terms of KL-divergence). Hence, if the more complex model (m_1) results in a higher estimate of Eh on the order of $\Delta Eh = 0.001$ or $\Delta Eh = 0.0038$, respectively, the respective model comparison would begin to speak in favor of the more complex model.

Therefore, the remaining question will be how the posterior noise precision relates to 'goodness of fit' (vE). For this, we will derive the following dependency:

$$\mathrm{var}(vE) = 1 - \frac{2}{[(1+\varphi^2)(N-1) - 2\varphi\beta(N-2)]\exp(Eh)}\left[\frac{N}{2} - \frac{Eh - hE}{hC}\right]. \tag{4.4}$$

[17] We use 'against' here sloppily. It is obviously *less* complex with respect to the noise prior, but implies that 'fitting better, results in less complexity'.

[18] We distinguish between 'bias' and 'overfitting'. Bias refers to a tendency towards selecting certain models inherent in the equations (e.g. better fitting through added flexibility of more complex models). Overfitting is meant in a stronger sense and means the actual fitting of true noise. In our simulation setting, we can assess, whether models predict variance of the data beyond the true, generated signal.

Here, φ and β denote the *assumed* and *true* autocorrelation of the residuals[19]. Again, we will show how this leads to the following implications:

4. Posterior noise precision Eh increases strongly with vE for high values of vE.
5. Therefore, especially when vE is close to 1, small differences in vE can lead to large differences in Eh, which in turn lead to large differences in F (a scenario that is particularly relevant for DCM for EEG).

We want to point out that there is nothing fundamentally problematic or new with the above relationships. They concur with the general intuition about Bayesian models and the effect of priors. Similarly, it is known that Bayesian estimates converge towards Maximum Likelihood estimates in the presence of many independent datapoints. Having said this, the present chapter does provide, to the best of our knowledge, the first a full quantitative analysis of these relationships in the context of DCM and the practical implications that this might have when fitting empirical datasets.

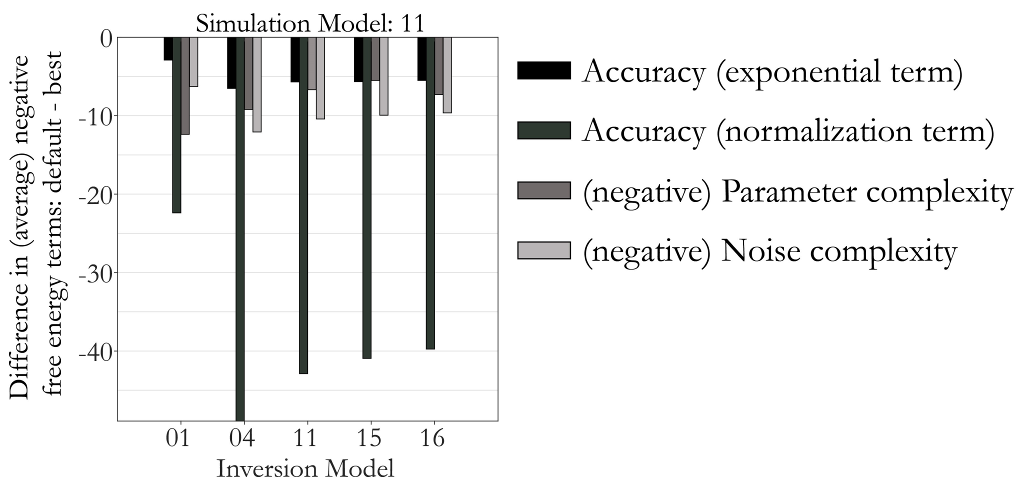

Figure 17 | Difference in negative free energy separated into accuracy and complexity terms for default vs. best starting values. Results are averaged over the synthetic datasets presented in Chapter 3 (N=20). For the accuracy terms, negative averages indicate higher accuracy for the best starting value, negative values for the complexity terms indicate less complexity for the best starting value.

[19] In the limit case of a gaussian likelihood with a diagonal covariance matrix, $\varphi = 0$. Therefore, the equation would simplify to $vE = 1 - \frac{2}{(N-1)\exp(Eh)}\left[\frac{N}{2} - \frac{Eh-hE}{hC}\right]$. Thus, if one neglects the effect of the prior (pwPE term), this implies the natural interpretation that the estimated (log-) noise precision equals the inverse residual variance:

$$Eh = -\log\left(\frac{N-1}{N}(1 - vE)\right).$$

Please note that the term $(N-1)/N$ cancels if one used the biased estimator for the variance.

Along those lines of obtaining a more quantitative understanding, the derivations obtained in this chapter will then allow for a more thorough analysis. We started this chapter off by mentioning the large differences in F we have observed within a single multistart (same inversion, same data, different starting values). Specifically, **Figure 17** illustrates how the different terms of the negative free energy contribute to this difference, for simulations under the model of average complexity (*m11*). These contributions underline our intuition that the hyperparameters deserve special attention; The two terms that contribute the most to the difference in F across starting values are both explicitly parametrized by the hyperparameters (normalization term and noise complexity).

We have structured this chapter in the following way: In the *Methods* section, we derive the relationships between vE, Eh and the terms contributing to the negative free energy (see Eq. (4.2), (4.3) and (4.4)). We then provide a rational why structure in the noise is likely to show similarities to the underlying true signal. This could pose a challenge if optimization is strongly accuracy driven. We will hypothesize a potential solution to reduce the risk of inferring the wrong model structure in this scenario. In the *Results* section, we first diagnose the hypothesized relationships for the simulations shown in the previous chapter. These analyses will make it apparent that our a priori assumptions about the noise, specifically about prior noise variance, created a setting where KL-constraints on parameters were (almost) completely outweighed by a penalty for less accurate fitting (i.e. KL-constraints on the noise). We then show that if we adapt these assumptions, the ability to recover models markedly increases. However, additional simulations will show that the situation is different for structured noise. Here, mere adaption of the prior noise assumptions does not exert sufficient regularization. We propose rigorous down-sampling to accommodate the noise-model assumptions. We will also show the results acquired under the default starting values as they provide additional insights into the multistart and show that sometimes, the default starting values can exert beneficial constraints.

Note that we will look at a reduced model space compared to the previous chapter, and focus on the model with no modulation (*m01*), forward and intrinsic modulation (*m11*) in the second region, and the most complex model (*m16*). This allows for a more focused analysis as we do not focus on whether modulatory effects can generally be distinguished. In the *Discussion*, we comment on the shortcomings of our simulations and possible future steps. We also provide a supplement, which illustrates the diagnostic plots for all empirical studies presented in this thesis (RATMPI and PRSSI), covering both EEG and fMRI. We provide a simple solution for specifying informed hyperpriors, which is, as of now, unfortunately only possible in fMRI. We decided to keep these two aspects separate in order not to complicate the reading of this chapter. Overall, this chapter illustrates that hyperprior selection is crucial in both EEG and fMRI, and model comparison of any kind should be interpreted with caution[20].

4.4 METHODS

4.4.1 APPROXIMATE RELATIONSHIPS

The noise model in DCM defines the assumptions about variance and the correlation structure of the residuals, which directly affect the accuracy term in the negative free energy. It is parametrized by a single or multiple parameters if region specific SNRs are assumed. Based on Eq. (3.11), the dependency of the negative free energy on the hyperparameters θ_h is as follows (we will use the same notion of *neuronal parameters* θ_p, *forward parameters* θ_g and *noise/hyperparameters* θ_h as in the previous chapter):

$$
\begin{aligned}
F(\theta_h) = &-\frac{1}{2}\log(\det(\Sigma_y(\theta_h))\\
&-\frac{1}{2}(y-y_p)^T\Sigma_y(\theta_h)^{-1}(y-y_p)\\
&+\frac{1}{2}\log(\det(\hat{\Sigma}_{\theta_h}(\theta_h)))\\
&-\frac{1}{2}(\theta_h-\mu_{\theta_h})^T\Sigma_{\theta_h}{}^{-1}(\theta_h-\mu_{\theta_h})\\
&+\text{const wrt. } \theta_h.
\end{aligned}
\tag{4.5}
$$

with

$$
\begin{aligned}
(\Sigma_y(\theta))^{-1} &= \Pi_y(\theta_h) = \exp(\theta_h)\cdot Q.\\
e_y &= (y-y_p)
\end{aligned}
\tag{4.6}
$$

[20] We use here 'caution' to refer to multiple critical aspects over the last two chapters. This also includes all aspects, from local maxima, hyperprior selection, sampling rate to structure in the noise.

where Q is a fixed noise matrix based on the (inverse) of an autoregressive process (AR(1)), parametrized by a fixed coefficient φ. e_y denotes the vector of residuals[21].

Since θ_h is defined in log-space, Eq. (4.5) can be simplified into

$$\begin{aligned} F(\theta_h) = &\frac{N\cdot\theta_h}{2} - \frac{1}{2}\exp(\theta_h)e_y{}^T Q e_y \\ &+\frac{1}{2}\log(\det(\hat{\Sigma}_{\theta_h}(\theta_h))) \\ &-\frac{1}{2}(\theta_h-\mu_{\theta_h})^T \Sigma_{\theta_h}{}^{-1}(\theta_h-\mu_{\theta_h}) \\ &+\text{const wrt. } \theta_h. \end{aligned} \tag{4.7}$$

We approximate the posterior noise precision mean by setting the first order derivative $\partial F / \partial\theta_h |_{\theta_h = Eh} = 0$. A crucial point here is that we assume conditional independence between the hyperparameters and all other parameters of a DCM, which is also incorporated in a mean-field approximation (Friston, Mattout et al. 2007). It is also important to note that the posterior covariance $\hat{\Sigma}_{\theta_h}$ only implicitly depends on the posterior mean: It is estimated by the curvature of the log-joint at the mean (Friston, Mattout et al. 2007). Hence, we will ignore this term for the moment, and show the validity when comparing the approximation to true results.

For simplicity, we will consider the one dimensional case and denote $hE = \mu_{\theta_h}$ and $hC = \Sigma_{\theta_h}$. Then

$$\frac{\partial F}{\partial\theta_h} = \frac{N}{2} - \frac{1}{2}\exp(\theta_h)e_y{}^T Q e_y - \left.\frac{(\theta_h - hE)}{hC}\right|_{\theta_h = Eh} = 0 \tag{4.8}$$

and therefore

$$\begin{aligned} \frac{1}{2}\exp(Eh)e_y{}^T Q e_y &= \frac{N}{2} + \frac{hE}{hC} - \frac{1}{hC}Eh \\ \exp(Eh)\cdot Z &= \frac{N}{2} + \frac{hE}{hC} - \frac{1}{hC}Eh \end{aligned} \tag{4.9}$$

with $Z = \frac{1}{2}e_y^T Q e_y$ depending on the residuals and the correlation structure. Hence, the approximate posterior noise precision estimate is defined as the intersection between an exponential function (scaled by a variable depending on the structure of the residuals) and

---

[21] The basis matrix Q is defined in the function *spm_Q*. By default, the coefficient for Q is set to $\varphi = 1/2$.

a linear function (depending on the prior and the number of datapoints). Since z and hC are semi-positive by definition, there can be only one (or zero) intersection point, making the estimate well defined[22]. Interestingly, Eq. (4.9) also provides a relationship between Eh and the exponential part of the accuracy in F. It predicts to scale linearly, with a slope of approximately $m = 1/hC$.

If we now plug Eq. (4.9) into Eq. (4.7), we see that

$$F(E_h) = \frac{N \cdot E_h}{2} - \left[\frac{N}{2} + \frac{hE}{hC} - \frac{E_h}{hC}\right]$$
$$-\frac{1}{2}(Eh - hE)^T hC^{-1}(Eh - hE) \tag{4.10}$$
$$+\text{const wrt. } Eh.$$

which makes the linear relationship between F and posterior noise precision mean (Eh) for an accurate prior (the third term (prediction error) is small) obvious. It not only scales with the number of datapoints, but also with the inverse prior variance. Their net effect is additive. This relationship allows us to understand the relationship between F and Eh, independently of the accuracy term and the residuals.

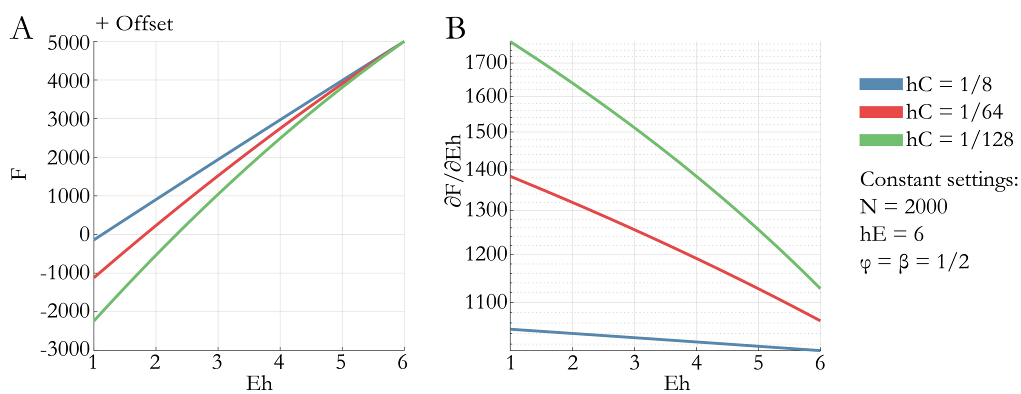

Figure 18 | Relationship between posterior noise precision mean and negative free energy (A) and the partial derivative (B). Note that the absolute values of F in (A) gain an additional offset from the terms not explicitly depending on Eh.

Figure 18 illustrates the linear and quadratic dependency of F on Eh for different prior variances of the noise. The current default setting for DCM in both EEG and fMRI assume

[22] Note that we will illustrate the effect of the different factors on the intersection point in the Supplement, to keep the main part of this chapter at reasonable length.

very little noise ($hE = 6$) with very high precision $hC = \frac{1}{128}$ (in EEG, this corresponds to an SNR of approximately 20 or to an explained variance of approximately 99.75%, see Glossary). The effect of these settings is two-fold. First, the linear part in Eq. (4.10) scales with $m = \frac{N}{2} + 128$. Second, the quadratic part will contribute positively to the scaling, if the posterior noise precision is *lower* than the prior noise precision mean (see. Eq. (4.3)). Put simply, if a model cannot predict 99.75% of the variance of the signal, it will get penalized through this precision weighted prediction error term (pwPE). The less it predicts, the stronger the penalty.

In the following, we will approximate the expression for Z as a function of vE, which will provide an approximate relationship between vE and Eh. This does not contribute additional information as it basically is just a change of variables from the posterior noise precision mean to explained variance. However, this dependency between vE and Eh allows us to bridge the gap between vE (which can be understood in absolute values) and the negative free energy.

As mentioned previously, Q (in Eq. (4.9)) encodes an autoregressive process of order one (AR(1)). We will briefly outline some facts about AR(1) processes that will be used later, but refer to common textbooks on time series analysis for further information (e.g. (Shumway and Stoffer 2017)). For any AR(1) process X_t with coefficient φ and Gaussian white noise (W_t) the following dependencies hold:

$$\begin{aligned} X_t &= \varphi X_{t-1} + W_t \\ \mathrm{var}(X_t X_{t-1}) &= \varphi\, \mathrm{var}(X_t) \end{aligned} \tag{4.11}$$

This AR(1) process is encoded in the matrix Q in Eq. (4.7)[23]. Importantly the entries of the matrix are defined by the coefficient φ through the Yule-Walker equation:

[23] Note that the presented derivation with respect to Q only holds for the case of DCM for ERPs. DCM for fMRI uses a different normalization of the data and assumption about the noise correlation structure (Q is diagonal). The equations could be easily adapted, as the data variance is known, and Q being diagonal is just a limit case of $\varphi = 0$.

$$Q = \begin{pmatrix} 1+\varphi^2 & -\varphi & 0 & 0 & \cdots \\ -\varphi & 1+\varphi^2 & -\varphi & 0 & \cdots \\ 0 & -\varphi & 1+\varphi^2 & -\varphi & \cdots \\ 0 & 0 & -\varphi & 1+\varphi^2 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{4.12}$$

$$Q_{i,j} = \begin{cases} 1+\varphi^2 & i=j \\ -\varphi & i=j\pm 1 \\ 0 & \text{otherwise} \end{cases}$$

This allows us rewrite Z as

$$\begin{aligned} Z &= (1+\varphi^2)e_y^T e_y - \varphi e_{y,t}{}^T e_{y,t-1} - \varphi e_{y,t-1}{}^T e_{y,t} \\ &\approx (1+\varphi^2)\cdot(N-1)\cdot(1-vE) - 2\varphi e_{y,t}{}^T e_{y,t-1} \end{aligned} \tag{4.13}$$

where we assumed that the expectation is $E[e_y] = 0$, and used the fact that the time series in DCM for EEG are normalized, i.e. $var(y) = 1$. Otherwise, Eq. (4.13) would additionally depend on the variance of the data.

Eq. (4.13) and Eq. (4.9) now provide an approximation to the relationship between the explained variance (vE) and the posterior noise precision Eh:

$$\operatorname{var}(vE) = 1 - \frac{2}{(1+\varphi^2)\exp(Eh)(N-1)}\left[\frac{N}{2} + \varphi e_{y,t}{}^T e_{y,t-1}\exp(Eh) - \frac{Eh-hE}{hC}\right]. \tag{4.14}$$

Eq. (4.14) still contains the covariance between the residual time series. If the residuals follow a true AR(1) process (with coefficient β), then the covariance, according to Eq. (4.11) can be approximated as

$$e_{y,t-1}{}^T e_y = (N-2)\cdot \operatorname{var}(e_{y,t-1}{}^T e_y) = (N-2)\cdot\beta\cdot(1-vE) \tag{4.15}$$

Note that, for the sake of generality, we allow that the true (empirical) noise AR(1) process could have a different coefficient, i.e. $\varphi \neq \beta$. Plugging this expression into Eq. (4.14) and solving again for vE yields the second approximation for the relationship between vE and Eh:

$$\operatorname{var}(vE) = 1 - \frac{2}{[(1+\varphi^2)(N-1) - 2\varphi\beta(N-2)]\exp(Eh)}\left[\frac{N}{2} - \frac{Eh-hE}{hC}\right]. \tag{4.16}$$

that only depends on the assumption of the noise correlation (φ) and the true noise correlation (β), as well as the prior mean hE and variance hC. This second approximation allows us to investigate the direct dependency of vE and F(**Figure 19**). Overall, we observe that for high regimes of vE, the relationship between vE and Eh gets very steep. In turn, the relationship between vE and F gets very steep. While for the simulations presented in the previous chapter, we arguably operated in an exceptionally high vE scenario, 90% of explained variance is not uncommon in DCM for EEG. In fact, we have observed similar values in the analysis of the empirical RATMPI study (also see Supplementary Material). Hence, improving model fit (i.e. accuracy) generally improves F, unless the parameter KL-divergence increases on the same order of magnitude. This holds for both, the gradient ascent during a single optimization and for comparisons between models.

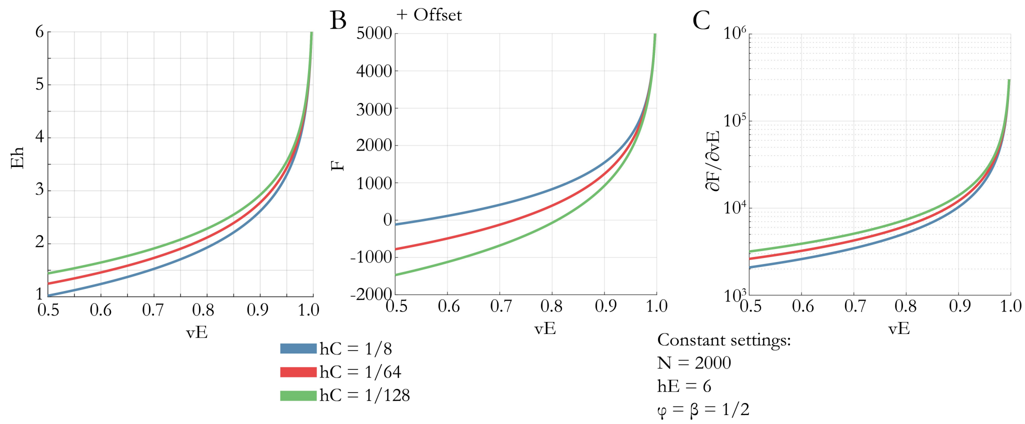

Figure 19| Relationship between explained variance and posterior noise precision mean (A), the negative free energy (B) and the partial derivative (C). Note that the absolute values of F in (B) gain an additional offset from the terms not explicitly depending on Eh.

The validity of approximations (4.14) and (4.16) can be seen in **Figure 20**. Note that for approximation 1 (Eq. (4.14)), we used the true covariance between the residuals (and the axes are flipped – in fact, we are predicting vE based on Eh). The approximation error therefore results from the fact that we did not take the posterior noise covariance into account and ignoring a residual mean (see Eq. (4.13)). The second approximation, Eq. (4.16) tends shows a larger difference. This difference is however most likely mediated by differences in the empirical error covariance (and hence $\beta \neq \frac{1}{2}$[24]).

[24] We will discuss the dependency between the true and expected autocorrelations of the residuals and their effect on the approximation in Supplement B.

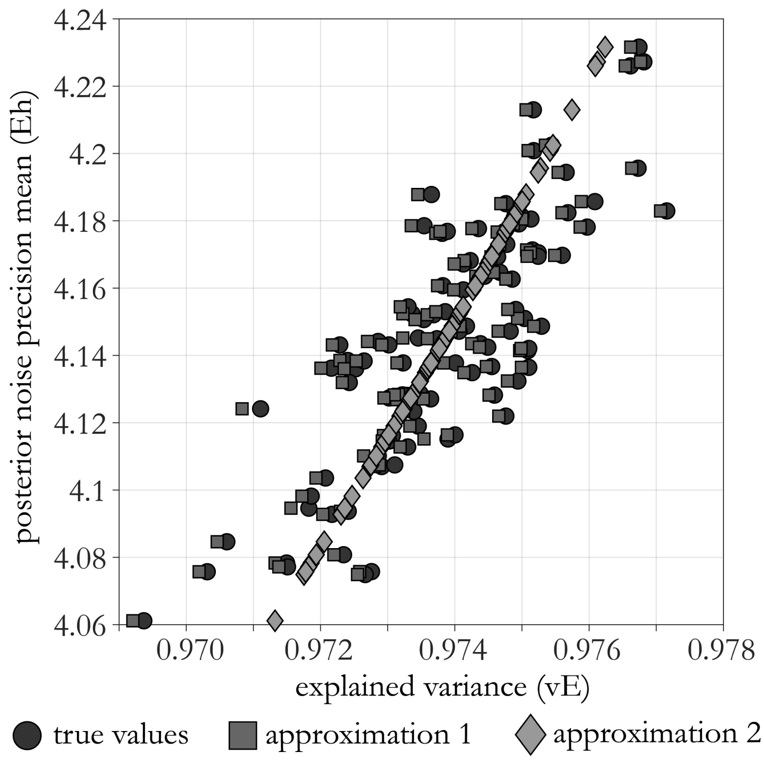

Figure 20| Relationship between posterior noise precision estimate and explained variance. A single marker refers to the inversion of a single dataset under the true model. Circles refer to the values as returned by the inversion. Squares refer to the first approximation of vE as a function of Eh. Diamonds refer to the second approximation of vE as a function of Eh.

In conclusion, **Figure 18** and **Figure 19** illustrate that under the current default hyperprior, there will rarely be an explicit constraint acting against maximizing Eh (or vE). Of course, a models dynamic repertoire is usually limited and more accurate fitting typically is accompanied by an increase in parameter complexity. In both figures, this term is not included. However, one can intuit the numeric range of constraints needed, in order for complexity to outweigh the accuracy-complexity tradeoff. Even more so, since better fitting results in less noise-complexity under the default priors. This begs the question what happens, if the noise contains structure that is in a similar frequency range as the underlying signal, i.e. if noise is explainable by the model. In the following, we investigate this scenario.

4.4.2 Filtered Noise

The current default noise model in DCM for EEG is based on an autoregressive process of order 1 (AR(1)), with a coefficient $\varphi = 1/2$ (Eq. (4.11)). This defines an exponentially decaying correlation between residuals. However, there are three reasons why this assumption could be violated in the case of DCM for ERPs. First, every model is a simplification of the true processes. If there are un-modelled processes, such as non-modelled regions, dynamics, etc. those processes will always lead to some structure in the residuals. Second, and more importantly, it is known that the power spectrum in the EEG noise follows a *1/f*-shape (so called 'colored noise'). That is, slow frequencies are expressed with more power, fast frequencies with less power (Ward and Greenwood 2007). Third, the ERPs that are being modeled are usually filtered. The filtering changes the structure of the noise. While one can debate in what sense the first point should affect the inference process, the two other points are based on a first principle argument and might need to be taken into account. Both filtered white noise and filtered *1/f*-noise exhibit correlation structures that are different from an AR(1) process. **Figure 21**A shows the different predicted correlation matrices for the AR(1) process, a filtered white noise process and a filtered *1/f*-noise process. One can observe that filtered noise results in a much slower decay in correlation over increasing time lags.

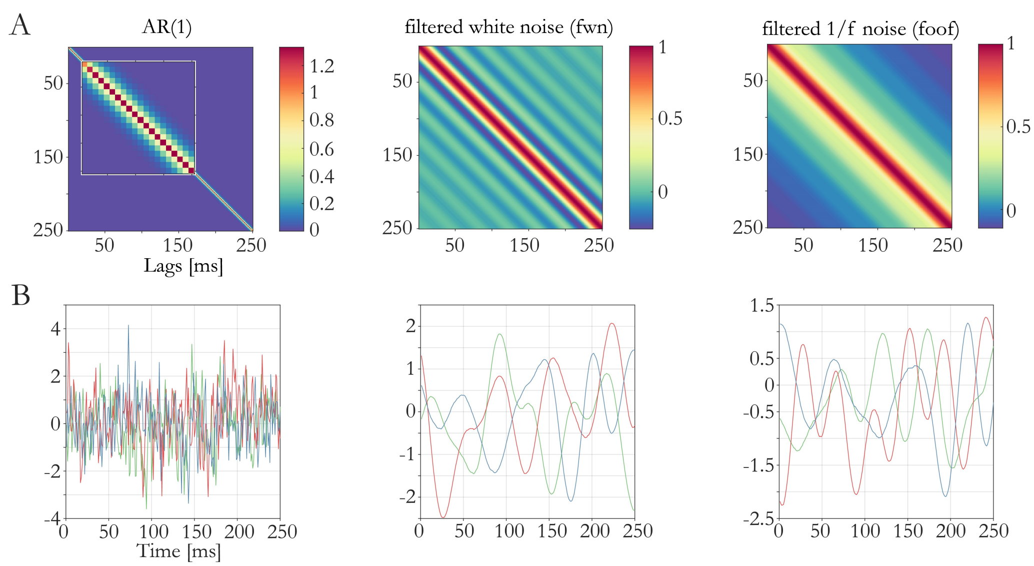

Figure 21| A) Basis matrices of noise Model (Q^{-1}) based on a AR(1) process (spm default), a filtered white noise process and a filtered 1/f process. The filter is chosen in accordance to typical digital filter used on the data (0.1 Hz highpass, 30 Hz-lowpass filter). For the AR(1) basis matrix, please note the zoomed in box (lags 0-20) B) Three samples of a noise process for the choice of covariance structure.

We then compared the predictions made from these three processes to the autocorrelation of empirical data. In brief, we computed the average residuals from the non-pharmacological part of the RATMPI dataset (averaged over rodents, **Figure 22**A). Note that these residuals stem from an inversion using the AR(1) model, with the most complex modulation structure (m16). We then computed the predicted autocorrelation from these three processes and compared them to the empirical autocorrelation (**Figure 22**B). One can see that both noise models based on a filtered process respect the empirical residual autocorrelation much better. Again, what we observe in those residuals is most likely a contribution of all three factors, possibly also true dynamics. The problem is that there is no way to truly differentiate the components, as the filtered noise process can also exhibit oscillations in the frequencies generally underlying the ERP. Nonetheless, the results in **Figure 22**C are derived purely from a first principle argument and are in no way fitted to the data.[25]

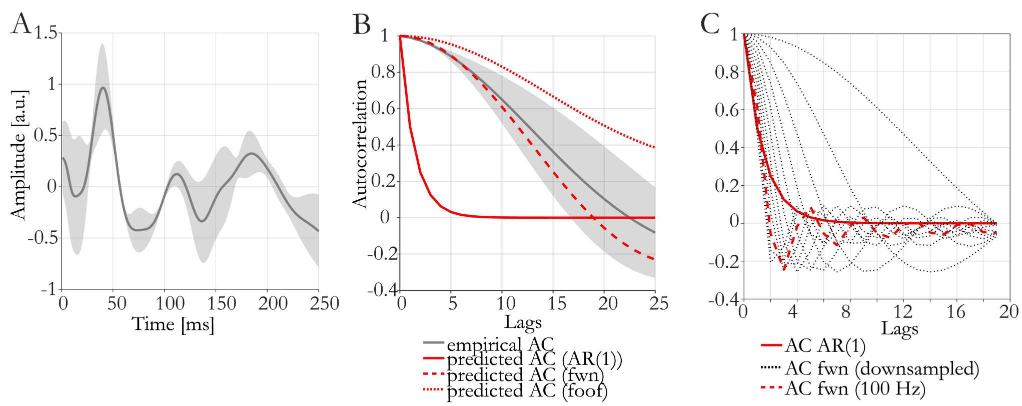

Figure 22| A) Average (over rodents) residuals of the non-pharmacological part of the RATMPI study (one condition, one region). B) True and predicted autocorrelation of the residuals according to the different noise models in **Figure 21**. C) Autocorrelation of AR(1) process and filtered white noise process for different lags (i.e. down-sampling).

The fact that the frequency spectrum of the ERP model overlaps with the noise poses additional problems when it comes to BMS. Together with the theoretical intuition that the negative free energy can be very much driven by fit, this could further bias model selection towards more complex models. In order to investigate this, we simulated new data where we added noise from a filtered *1/f*-process onto our simulations. The filter was chosen according to a typical low- and highpass filter for ERPs, effectively creating a bandpass filter between 0.1 and 30 Hz. Otherwise, the simulations and inversion routine (multistart) were exactly analogous to the settings in the previous chapter (**Table 5**, **Figure 10**).The resulting average (noisy data) and average true signal are illustrated in **Figure 23**.

[25] Arguably, the residuals do come from an inversion with the AR(1) model.

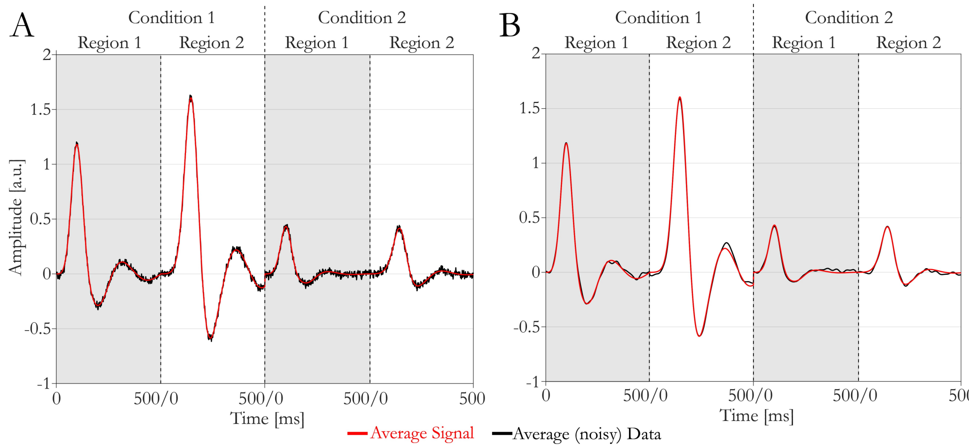

Figure 23| Average Data (model 16) for the different noise structures. A) True signal (red) with AR(1) noise ($\varphi = 1/2$). B) True signal (red) with filtered 1/f noise (black). Both simulations were done for a SNR =7.

4.4.3 Matching of the Residual Autocorrelation[26]

These new simulations exhibited two major differences to the previous ones. First, they were done in a slightly higher regime of explained variance[27], making the (diagnostically observed) relationship between fit and negative free energy even steeper. Second, the noise model assumptions (AR(1)) did not match the empirical noise anymore. From the last chapter, we knew that model recovery performs well (despite bias), even when there is a high number of datapoints, if the assumptions about the residual correlations is correct. Therefore, an ideal way would be to use the knowledge about the correlation structure in the noise to inform the residual forming correlation matrix. Unfortunately, the matrices based on filtered noise in **Figure 21**A are close to being non-invertible because of the strong correlations and can therefore not directly be used.

[26] Please note that distinguishing (particularly structured) noise from data is a notoriously difficult problem in the analysis of time series and signal processing. We are well aware that there might be sophisticated methods to deal with these problems. We have only with a very naïve understanding, especially compared to the informed reader. As this was the last problem identified and tackled for this thesis, the proposed solution should be taken with a grain of salt and goodwill. Anything beyond this very rudimentary and preliminary analysis would have extended vastly beyond the scope of these last weeks.
[27] Due to the correlation structure in the residuals, the mapping from SNR to vE is not exactly the same. The equation in the Glossary is only an approximation in the presence of correlations between noise and data.

Therefore, we considered a different strategy where the down-sampled, hypothesized colored noise matched approximately the modeled AR(1) process. In brief, there is a mathematical relationship between the coefficients of an autoregressive process and the power spectrum through the Wiener-Khinchin-Theorem, but here, we generated multiple instantiations of random signals and used a digital filter. We then increased the lag for the autocorrelation for the structured process, but keeping the lags for the AR(1) fixed. Based on this, we opted for a down-sampling by a factor of 10 (i.e. 100 Hz, **Figure 22**C). It is important to note that this was done independently of the data based purely on assumptions about the noise. However, the down-sampled residuals did match the AR(1) process nicely (**Figure 28**B). Hence, we recomputed the full inversion for the down-sampled data and compared the results.

4.5 RESULTS

4.5.1 DIAGNOSTICS

Based on the derivation in the methods section and the assumed relationships, we confirmed our intuition in the synthetic data of the last chapter by running diagnostics. We here illustrate the results for the DEFAULT starting values[28]. As a brief reminder, we only consider the reduced model space consisting of *m01, m11* and *m16* (no modulation, forward and gain modulation in the second region, full modulation structure). We simulated N=20 datasets for each model, and computed inversions with all models. We will discuss the results by considering the negative free energy as a sum of three terms.

$$
\begin{aligned}
F = &-\frac{n_s}{2}\log(2\pi) - \frac{1}{2}(y-y_p)^T \Sigma_y(Eh)^{-1}(y-y_p) - \frac{1}{2}\log(\det(\Sigma_y(Eh)) \\
&-\frac{1}{2}(Eh-hE)^T hC^{-1}(Eh-hE) + \frac{1}{2}\log(\det(Ch^{-1}hC)) \\
&-\frac{1}{2}(\theta-\mu_\theta)^T \Sigma_\theta^{-1}(\theta-\mu_\theta) + \frac{1}{2}\log(\det(\hat{\Sigma}_\theta^{-1}\Sigma_\theta)).
\end{aligned}
\tag{4.17}
$$

[28] We chose the DEFAULT starting value here, because there, the fixed effects BMS only recovered 1/3 models (the most complex model) correctly (see **Table 10**). This makes it easier to illustrate the numerical dependencies that can lead to erroneous network identification from a BMS perspective.

Here, the first term is the *accuracy*, the second and third term are the KL-complexity terms associated with the *noise* and all other parameters (*neuronal* and *forward*).

Figure 24 depicts the different contributing terms. One can see that when simulating with $m01$ and $m11$, FFX BMS indicated strong evidence for a model of higher complexity ($m11$ and $m16$ respectively, **Figure 24**A). These differences are mostly driven by accuracy and the choice of hyperprior. In brief; more complex models result in higher complexity in terms of the parameters **Figure 24**C. The (average) additional complexity (3.94, 3.14) when adding two (modulatory) parameters is comparable to AIC/BIC complexities (4 and 15, respectively). This complexity acts against the slight increases in accuracy (18, 3, **Figure 24**B). However, for the given noise prior (hE, hC =6, $1/128$), an *increase* in accuracy generally implies a *decrease* in complexity. In that sense, noise complexity acts against parameter complexity and favours better accuracy. As a result, at least in the situations with $m11$, this a priori assumption tilts the accuracy-complexity tradeoff in favour of the more complex model 16 (3 – 3.14 + 0.7=0.86)[29]. In the situation when simulating with $m01$, the noise-KL term alone does not cause the bias in favour of $m11$. Even when neglecting the noise-KL, the better accuracy of $m11$ outweighs the parameter complexity (18-3.94).

[29] Please note that this accuracy-complexity tradeoff regards the *average* across 20 datasets. Hence, to compute the FFX BMS differences, one needs to multiply this difference with 20. The resulting difference comes from a rounding error.

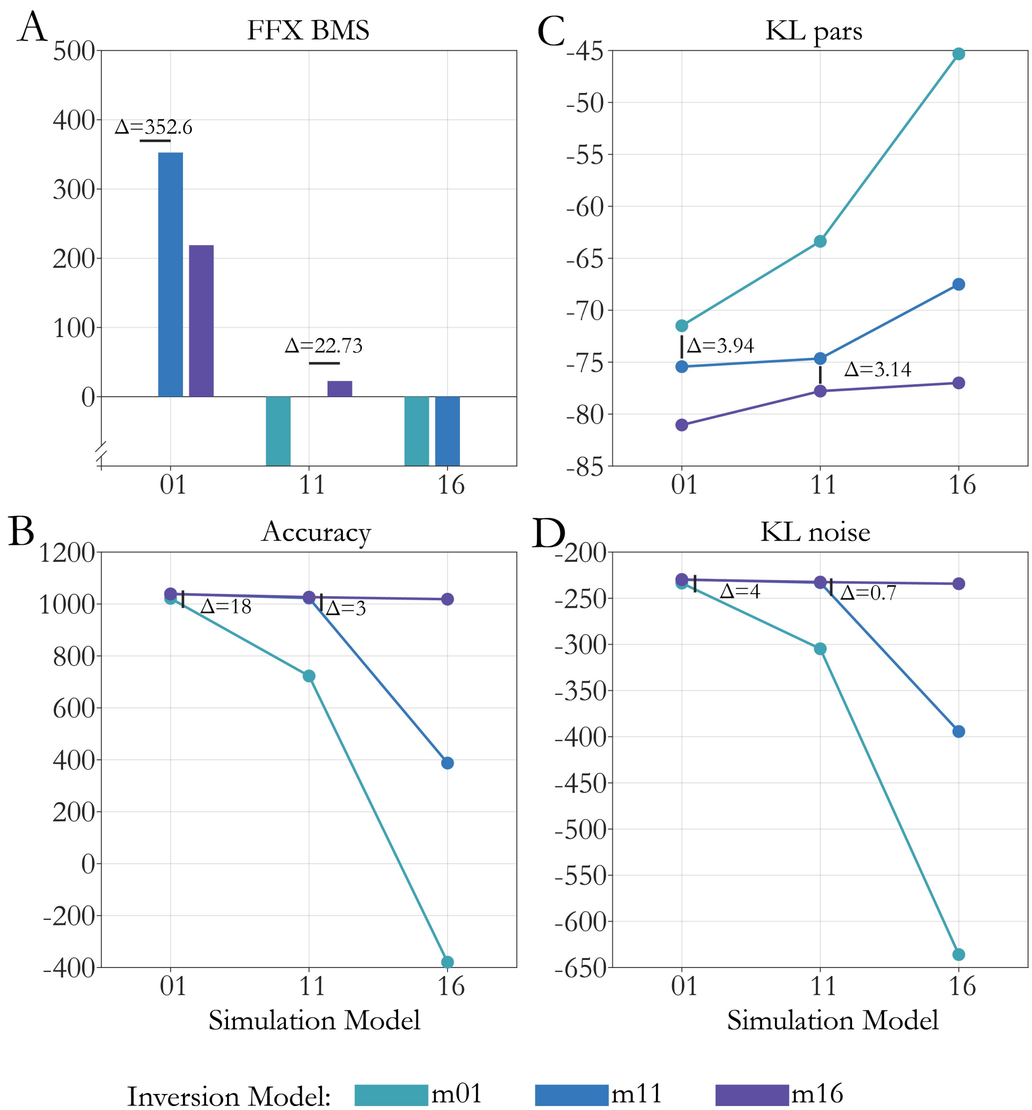

Figure 24| Fixed effects BMS and comparison between the quantities of the negative free energy for the synthetic dataset with AR(1) and tight prior noise precision. Results shown for DEFAULT starting values. A) Barplot is normalized to the true, data-generating models (FFX BMS). Barplot is cut-off for values below -100. Only the most complex model is recovered. In line plots (B-D), dots represent average values over all 20 synthetic datasets. Y-Axis in [a.u.] (for the respective quantity shown in the title. Accuracy (B), parameter-KL (C) and noise-KL (D) depict averages (N=20).

In terms of goodness of fit, this arguably delicate balance is emphasized (**Table 7**). When comparing the explained variances for the simulation with *m01*, the difference in average explained variance (compared to m11) amounts to 0.2%. Based on **Figure 20**C, we can visually estimate the slope of F wrt. vE (for N=2000) to approximately $\frac{\partial F}{\partial vE} \approx 30000$. Hence, differences of $\Delta vE = 0.002$ are expected to result in free energy differences in the double digits (on average).

		inversion Model		
		m01	**m11**	**m16**
simulation Model	**m01**	0.969	0.971	0.971
	m11	0.939	0.970	0.970
	m16	0.726	0.894	0.970

Table 7| Average explained variance for the different simulation and inversion models. Data was simulated with AR(1) noise. Inversion performed under the same noise model with tight prior noise precision.

We want to reiterate again that this discussion was based on the results from the DEFAULT starting values[30]. Therefore, it is clear that they (might) correspond to a suboptimal solution (in terms of negative free energy). In fact, it is true that the BEST starting values managed to recover all three models correctly (from the FFX BMS perspective). Therefore, the multistart most likely managed to find better (and after this discussion most likely more accurate solutions) for the simpler models. Nonetheless, it illustrated that the success of tradeoff does depend on the hyperprior specification see Eq. (4.10) and also the starting values.

4.5.2 HYPERPRIOR INDUCED BIAS

In our simulations, we do know ground truth. While we simulated 'true' noise under the assumed noise process, the simulated SNR is below the assumed SNR. We therefore re-analyzed the results, incorporating the fact that we might have less knowledge about the amount of noise.

We can formally include higher uncertainty about the amount of noise, by increasing the width of the prior noise precision (i.e. hC). The two prior settings are depicted in **Table 8**. For consistency, we will always refer to these two settings as 'tight', or 'wide'. Note that the actual value of $hC = 1/8$ was a somewhat arbitrary choice. We wanted to choose a setting,

[30] See glossary for the definition of the DEFAULT vs BEST starting values.

where differences become clear (yet it is closer to the width the prior distributions for other parameters, e.g. modulatory parameters).

Hyperprior (prior noise precision)	**mean (hE)**	**variance (hC)**
tight (default in SPM12)	6	1/128
wide	6	1/8

Table 8 | Two inversion settings specifying the a priori noise assumptions. Both expect the same magnitude of noise, but with different certainty. Note that lower certainty corresponds to higher variance, i.e. a wider distribution.

Since the prior noise variance directly scales the noise complexity term, i.e.

$$pwpE_{noise} = -\frac{(Eh - hE)^2}{2hC} \tag{4.18}$$

we would assume that setting a 'wider' prior noise precision reduces the overwriting effect the noise complexity can exert on the parameter complexity. This in turn could improve model recovery, as more complex models are more appropriately penalized. (Note that 'appropriately' here simply means in accordance with our (known) simulations). The BMS results and the contributing terms are illustrated in **Figure 25** and **Table 9**.

		inversion model			
		m01	**m11**	**m16**	
simulation model	**m01**	1099.3	1100.2	1102.3	**Accuracy**
	m04	778.18	1096.8	1093.4	
	m16	-307.34	486.62	1094.3	
	m01	-51.78	-56.8	-62.62	**KL pars**
	m04	-48.54	-64.87	-66.26	
	m16	-33.41	-61.01	-64.71	
	m01	-19.29	-19.29	-19.27	**KL noise**
	m04	-25.11	-19.35	-19.42	
	m16	-50.1	-31.1	-19.4	

Table 9| Comparison between the quantities of the negative free energy for the synthetic data with AR(1) noise, and wide prior noise precision (also see Figure 25). Quantities are averaged[31].

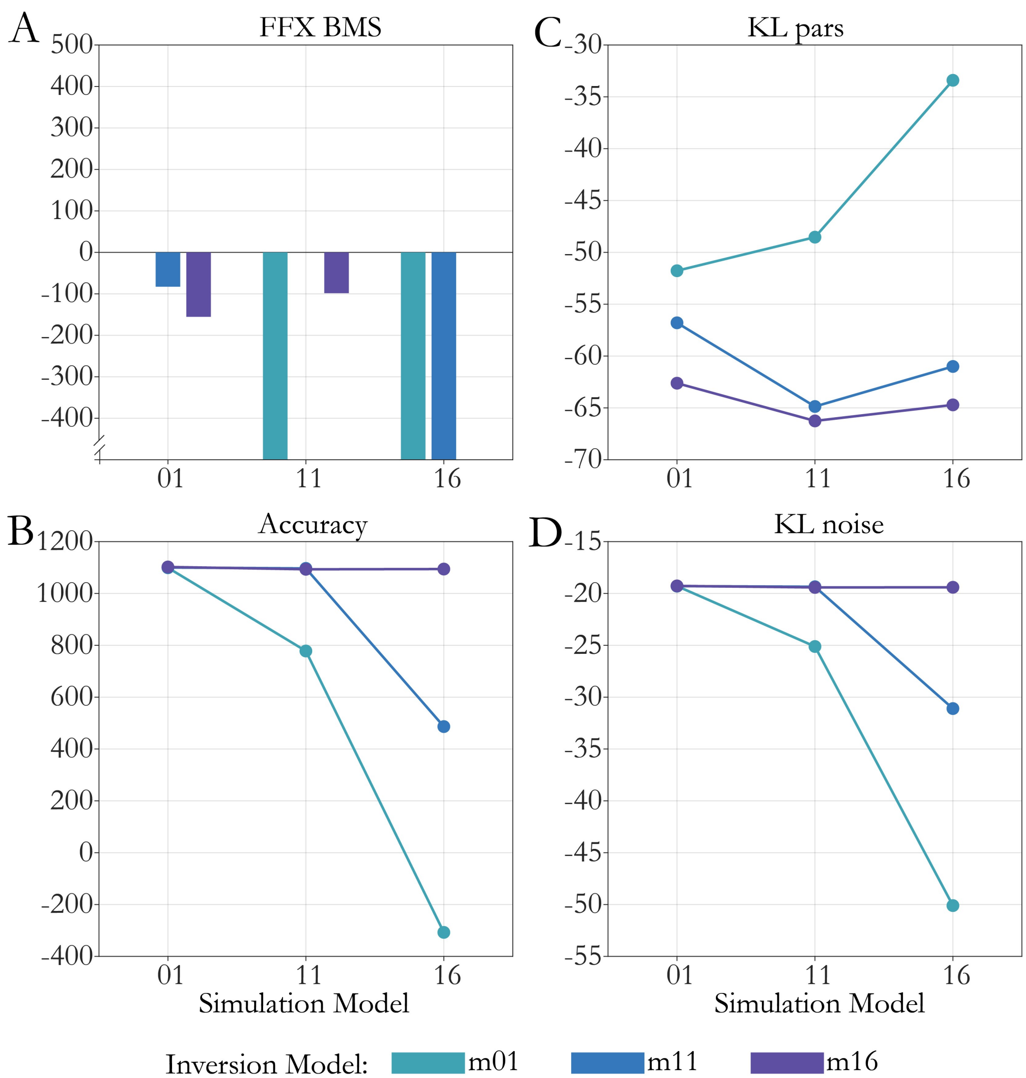

Figure 25| Fixed effects BMS and comparison between the quantities of the negative free energy for the synthetic dataset with AR(1) and wide prior noise precision. Results shown for BEST starting values. A) Barplot is normalized to the true, data-generating models (FFX BMS). Barplot is cut-off for values below -500. All three models are correctly recovered. In line plots, dots represent average values over all 20 synthetic datasets. Y-Axis in [a.u.] (for the respective quantity shown in the title. Accuracy (B), parameter-KL (C) and noise-KL (D) depict averages (N=20).

[31] To avoid confusion when comparing the results to the FFX BMS results in **Figure 25**A. The negative free energies in **Figure 25**A are normalized to the true model and the sum of the single-subjects negative free energies (Hence, differences in the table increase by a factor of 20 compared to **Figure 25**A).

We observe that all three models are correctly recovered in terms of a fixed effects BMS (**Figure 25**A). Importantly, the precision weighted prediction error term is numerically in a similar range as the complexity term for the parameters. Hence, the wider noise prior on the precision prevents the noise complexity to overwrite parameter complexity. Arguably, the model is better constrained. 'Better' is to be understood as 'in closer agreement with the true, synthetic data', where we know that we have simulated more noise than assumed under the prior mean. As we add two additional parameters, complexity increases in the range of $\Delta KL = 2 - 6$ (for models of equal or higher complexity than the simulating model, **Table 9**). This is similar to an AIC constraint ($2\,p = 4$) but only slightly lower than a BIC constraint ($\frac{\log(n)p}{2} = 7.6$). The fact that our model assumptions satisfy ground truth more closely improves model recovery also in terms of the confusion matrix (**Table 10**). Using the wider noise prior increased balanced accuracy (BA) from 76.67% to 91.67%.

		DEFAULT starting value		**BEST starting value**	
simulated noise	prior noise variance	Balanced Accuracy	BMS (number of models recovered)	Balanced Accuracy	BMS (number of models recovered)
AR(1)	tight	75	1/3	75	3/3
	wide	76.67	3/3	91.67	3/3

Table 10| Model recovery analysis for the simulations with simulated AR(1) noise. Comparison between two prior settings, and DEFAULT and BEST starting value. Models included in the analysis are m01, m11 and m16.

These results are not at all counterintuitive. If anything, they adhere to the expectation one has about Bayesian models; Results become 'better', if the priors reflect the underlying truth more closely.

An illustrative case is the simulation with *m01* and inversion with *m01* and *m11*. Here the (too complex) *m11* results in slightly higher accuracy than the true *m01* (1099.3 vs. 1100.2), but is also of higher parameter complexity (-51.78 vs. -56.8). Importantly, due to the wider noise prior, this increase in parameter complexity is not overwhelmed by a decrease in noise complexity for the more complex model (equal to the second order) as illustrated in the previous case (**Figure 24**D). In summary, the true model *m01* results in a higher negative free energy, and is correctly recovered by FFX BMS[32].

Nonetheless, the explicit dependency of the negative free energy on the hyperparameters show that given a default setting, the model will strive for (almost) maximum fit. This is only counteracted by parameter complexity, which showed numerically similar magnitude as AIC/BIC in this setting. For a very tight hyperprior, parameter complexity quickly gets overwritten if the assumed noise precision is not met. It begs the question what happens if the underlying assumptions about the noise are wrong. More specifically, what happens if true noise is expressed in a similar frequency bandwidth as the signal.

4.5.3 Noise induced Bias

We first briefly illustrate the model recovery results when using the same, initial setting of the hyperpriors (tight) in terms of model recovery. We expected a clear decrease in our ability to recover models, as all the critical points elaborated in the previous simulation setting also apply here. In addition, models can now become prone to overfitting. If noise shares frequencies with the signal that can be accounted for by the model, then it is to be expected that given the tendencies of the model to optimize fit, it would likely result in the fitting of noise. (The average data is visualized in **Figure 23**B).

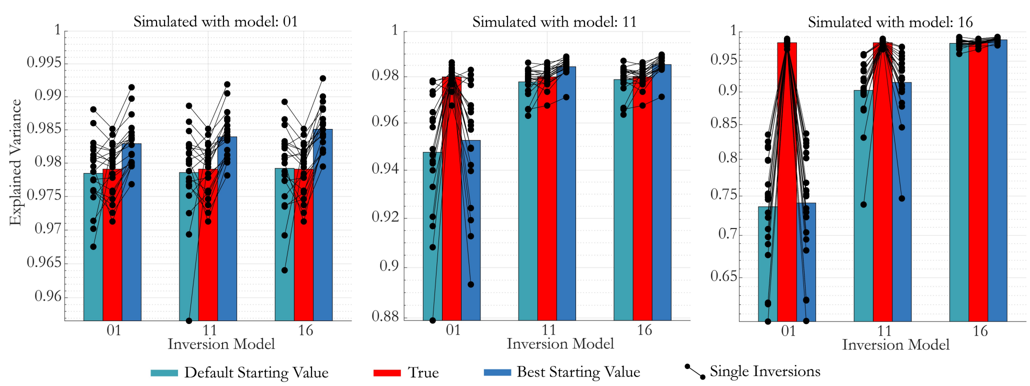

Figure 26| Explained variance (log-scale) across the three simulation and inversion models. Simulations performed with filtered 1/f noise. Inversions under tight hyperpriors without down-sampling. Bars depict

[32] To be fair, the differences in **Figure 24**D where higher, because the DEFAULT starting value led to a 'worse' local maximum. When using the multistart for the tight noise prior, also all models are recovered correctly form a BMS perspective!

average (across 20 simulations), single dots depict single inversions. Red bars depict the true explained variance (as simulated).

Figure 26 depicts the predicted and true explained variance across all simulations, for BEST and DEFAULT starting value. We clearly see the tendency to overfit for all models of equal or higher complexity than the simulating model. This is not surprising given the prior and the assumption that datapoints are almost independent. More striking is the fact that the DEFAULT starting value exhibits an artificial[33] constraint on overfitting. While restricted solutions (in terms of fit) are found from the DEFAULT starting value even for the true model, it is much less prone to explain noise for more complex models. The intuition behind this is the following: If the explanation (fitting) of noise results in higher negative free energy, it is much more likely that some of the starting values will find such solutions. On the other hand, an optimization starting from the prior mean (DEFAULT) will inevitably end up in a 'closer' local optimum (to the prior mean), which may or may not involve the fitting of noise. This also shows in a superior performance in terms of model recovery (**Figure 27**). Under the DEFAULT starting values, more models are correctly recovered (50% vs. 40%), and there is lower tendency to favor the most complex model (27/40 vs. 33/40 assignments).

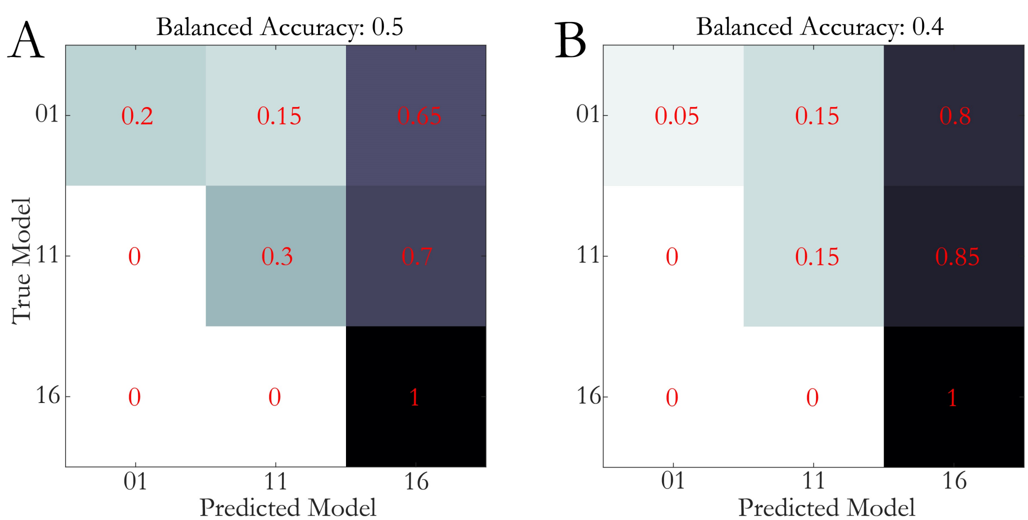

Figure 27| Confusion matrix of the inversion under tight prior noise precision. Data simulated with filtered 1/f noise. A) Results for the DEFAULT starting value. B) Results for the BEST starting value. Probabilities of 0.05 correspond to the assignment of a single dataset (N=20).

[33] Artificial here refers to a 'non-intended' constraint in the sense that the optimization does not intend to find only a local optimum.

While these results were again acquired with a non-exhaustive sampling of starting values (N=100), we would expect that the recovery rather gets worse than better under this inversion setting[34]. Also from a fixed effects BMS perspective, only the most complex model was recovered (**Table 11**).

4.5.4 Avoiding Noise induced Bias

In the methods section we have already outlined the strategy of down-sampling. Down-sampling comes with two benefits. Parameter constraints contribute (relatively) more strongly to the negative free energy, because the ratio between the number of datapoints and number of parameters gets lower. Secondly, the correlation of the empirical noise is expected to match the correlation structure of the assumed noise better. For completeness, we augment the down-sampling approach again with the adjustment of the width of the prior noise precision in a factorial fashion. That is, we ran four settings, with and without down-sampling (to 100 Hz) under tight and wide prior noise variance (no-downsampling with tight hyperpriors was already presented in the previous section).

			DEFAULT starting value		**BEST starting value**	
simulated noise	prior noise variance	down-sampling	Balanced Accuracy	BMS (number of models recovered)	Balanced Accuracy	BMS (number of models recovered)
colored (structured)	tight		50	1/3	40	1/3
	wide		48.33	1/3	45	1/3
	tight	yes	78.33	3/3	71.67	3/3
	wide	yes	83.33	3/3	78.33	3/3

Table 11| Model recovery analysis for the simulations with simulated filtered 1/f noise. Comparison between two prior settings, and DEFAULT and BEST starting value. Models included in the analysis are m01, m11 and m16.

The results are provided in **Table 11**. We can see that merely adjusting the hyperprior variance alone is not enough to clearly improve model recovery. This is clear from a mathematical perspective as well. Considering Eq. (4.10), the negative free energy still scales

[34] If the fitting of true noise results in higher negative free energy, more complex models would most likely be more flexible to do so.

with the number of datapoints (i.e. N/2), even if the contribution of the hyperprior mean and variance is reduced. In order to make the hyperprior act against trying to fit the data perfectly, one would need to explicitly overestimate the noise a priori. Then, the pwpE term of the noise would eventually act against the slope induced by the number of datapoints, as the predicted precision exceeds the prior mean.

On the other hand, down-sampling shows clear benefits overall, especially in combination with making the hyperprior wider. Independent of the starting value, fixed effects BMS recovers all three models, and BA improves to 83.33% and 78.33% percent for DEFAULT and BEST starting value, respectively. **Figure 28** shows that the predicted and true autocorrelation match nicely, however, there still appears to be structure in the residuals that could potentially be explained by the model. (In fact, the residual oscillations look very much like the mismatch between true and predicted autocorrelation, **Figure 28**B). Figure 28C then shows the confusion matrix for the BEST starting values. If we compare the number of misclassifications in favor of the most complex model with the previous figure (**Figure 27**B), it is greatly reduced (7/40 compared to 33/40).

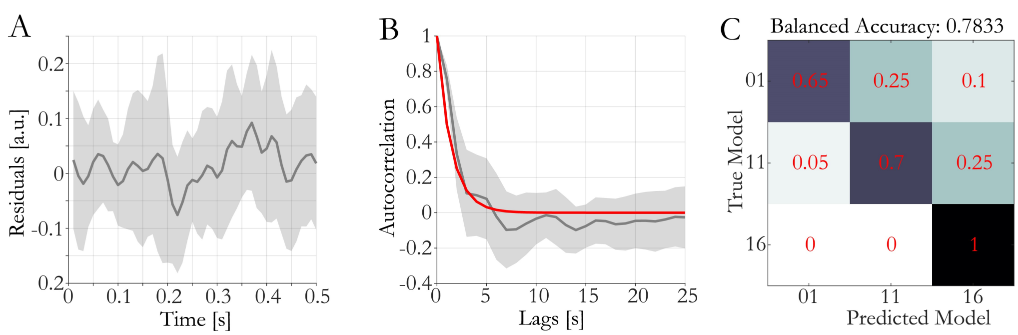

Figure 28 | A) Average residuals of the inversion under m16 with down-sampling and wide hyperpriors. Grey line depicts average, shaded area SD over 20 inversions. Data was simulated with filtered 1/f noise. B) Residual autocorrelation. Grey line depicts average, shaded area SD over 20 inversions. Red line depicts expected autocorrelation by the model. C) Model recovery for the BEST starting values.

Still, the DEFAULT starting value performs better in terms of balanced accuracy, indicating that there is still a tendency to overfit. Unfortunately, we can confirm this intuition. Even for the 'best' (in terms of model recovery) inversion assumptions (down-sampling + wide hyperpriors), the diagnostic figure looks virtually the same as in **Figure 26**. Hence, while overfitting is still present, the superior performance of the model recovery is most likely due to a better balance between the accuracy and complexity terms of the negative free energy (**Table 12**). Note that the fact that we still, a priori, overestimate noise precision, the noise-pwPE term still decreases with increasing fit. Nonetheless, the reduction in datapoints reduces the overall contribution of accuracy (mediated by the log-determinant term) ultimately also leading to smaller differences in terms of free energy.

		inversion model			
		m01	**m11**	**m16**	
simulation model	**m01**	201.73	203.84	205.87	**Accuracy**
	m04	133.12	203.93	206.7	
	m16	-30.8	77.2	210.96	
	m01	-41.13	-45.47	-48.75	**KL pars**
	m04	-35.14	-48.84	-52.35	
	m16	-22.99	-37.63	-53.08	
	m01	-6.41	-6.31	-6.21	**KL noise**
	m04	-14.19	-6.3	-6.14	
	m16	-43.96	-22.76	-5.18	

Table 12| Comparison between the quantities of the negative free energy for the synthetic data with AR(1) noise, and wide prior noise precision (also see **Figure 25**). Quantities are averaged.

We end this chapter on the recovery of the parameters. **Figure 30** illustrates the parameter recovery over all data and inversion settings discussed in this chapter, for both DEFAULT and BEST starting values. Interestingly, parameters are consistently better recovered when using the multistart, independent of the noise and inversion setting. In fact, parameter estimation seems much less affected from the structure in the noise, and is even better at times for down-sampled data. This might be contradictory to what the model recovery results with the structured noise would have led to intuit. However, it is important to keep in mind that the complexity term in the negative free energy is elegant, but difficult to intuit. Therefore, it is possible that the true parameter set does not necessarily result in the highest negative free energy. In fact, similar to the previous chapter, we have benchmarked the model recovery performance by running a single inversion starting from the true values. We do not show it explicitly in this chapter, but for structured noise, the benchmark performs worse than the multistart in terms of balanced accuracy, which supports this idea. There are trivial reasons why this happens; The true parameters would only be recovered if the posterior and the prior would match exactly. Otherwise, parameters will always be subject to some level of constraint, as outlined in the introduction. Also, correlations can create other parameter sets resulting in higher negative free energy than the true set. But here, it can most likely also be attributed to the ability of the model to explain noise, which is beneficial in terms of the negative free energy. The better parameter recovery of the multistart comes however at a cost; the variance in the estimates is larger across datasets than for the DEFAULT starting values.

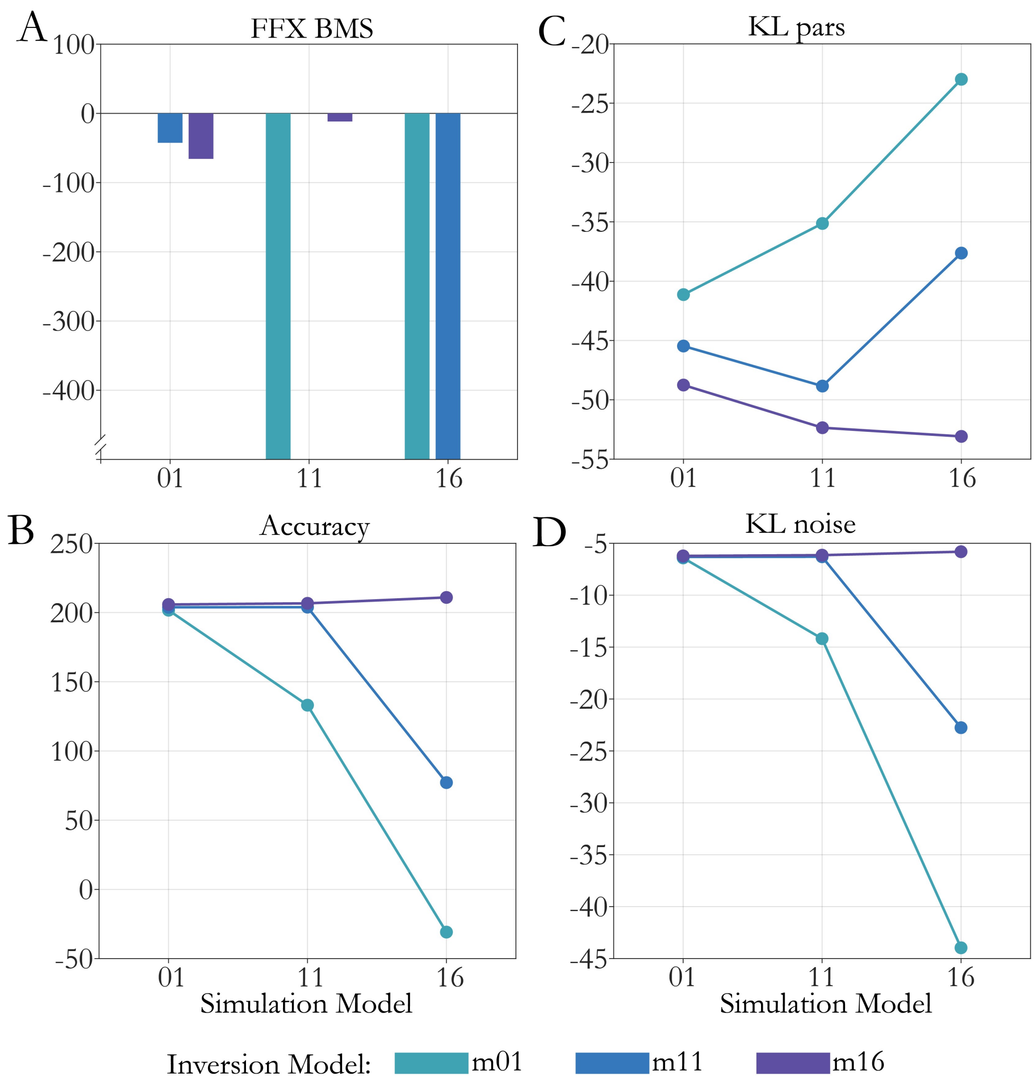

Figure 29| Fixed effects BMS and comparison between the quantities of the negative free energy for the synthetic dataset with filtered 1/f noise and wide prior noise precision. Results shown for BEST starting values. A) Barplot is normalized to the true, data-generating models (FFX BMS). Barplot is cut-off for values below -500. All three models are correctly recovered. In line plots, dots represent average values over all 20 synthetic datasets. Y-Axis in [a.u.] (for the respective quantity shown in the title. Accuracy (B), parameter-KL (C) and noise-KL (D) depict averages (N=20).

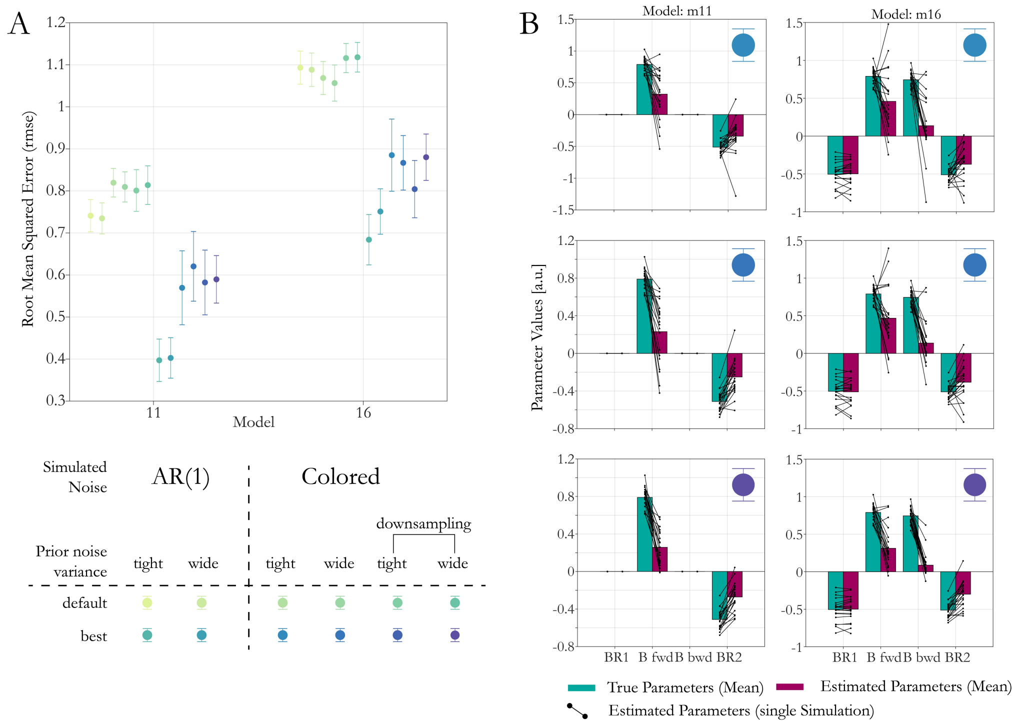

Figure 30| Recovery of modulation parameters. A) Recovery across all inversion settings for DEFAULT and BEST starting value, measured in RMSE. Errorbars depict SEM. B) Specific recovery for data with colored noise as in A, depicted by the respective symbol. All results shown for the BEST starting value for inversion under tight prior noise variance (top), wide variance (middle) and wide with down-sampling (bottom).

4.6 DISCUSSION

In this chapter, we have discussed the impact of the noise model on the network/parameter inference problem. The noise model differentiates noise from data. These assumptions are formalized in three aspects of the noise model:

- The (absolute) amount of noise (prior mean, hE)
- The certainty about the amount of noise (prior variance, hC)
- The covariance structure of the noise (in the present terminology the matrix Q). An alternative view on Q is that it incorporates the independence assumption on the data/residuals. This is important because as we have seen, the negative free energy depends explicitly on the number of datapoints.

Also, from the modelers perspective, those three things are to his/her disposal to control the inference behavior (wrt. the noise model).

In a nutshell, one could argue that, whenever these assumptions were more in line with ground truth, model recovery was better. On one hand, this holds with respect to the prior noise variance. Since our data were not exactly simulated in the same regime of SNR as assumed under the prior mean, model recovery (in terms of BA) increased from 75% to above 91%, once we made the true setting more likely under the prior. On the other hand, in the simulations with filtered noise, BA increased from 40% to 78% when the expected and true correlation structure matched better. This behavior is as expected from a Bayesian perspective, which we have pointed at multiple times throughout this thesis.

Despite this, it is important to see all the other aspects that these simulations illustrate:

First, the default prior noise assumptions inevitably create a scenario, where there is bias in favor of more complex models if the true signal does not explain close to vE=99.7% of the data. If the model cannot predict close to this amount, then the precision weighted prediction error term of the noise penalizes strongly. In other words, from a negative free energy perspective, the models that fail to explain this amount, tend to be less plausible. At the same time, the estimated noise precision (Eh) scales strongly with vE in the scenarios, where models do explain a lot of variance. Finally, the relationship between F and Eh has an additional linear contribution which scales with the number of datapoints (N/2). In combination, even for minor differences in vE, large differences in Eh and hence in F can be observed. This explains the observation of large negative free energy differences across starting values, which motivated this chapter. This is not necessarily a bad thing! But it becomes a challenging question for a modeler to which degree he/she can be certain about the level of noise in the data. As we have shown, even in this synthetic setup with arguably little noise, too high a priori certainty clearly impaired recovery. We have only observed more plausible result as we increased the prior noise variance. But arguably, whenever in doubt, one can cast it as a question of model comparison and run two inversions with a wide (e.g. $hC=1/8$ or even wider) and a tight prior noise precision.

Second, considering a scenario where noise shares frequencies with the explainable signal, the arguably strong assumptions the default noise model makes become more important. All numerical arguments from the first point still hold, creating a scenario prone to overfitting and bias. We believe that the filtering of the data as a standard preprocessing step for ERPs makes a strong argument that this is actually relevant for empirical data.

From a Bayesian perspective, there are at least two alternative approaches one could take to prevent overfitting; i) Giving the prior more weight; ii) giving the data less weight. The first approach would generally involve a more conservative assumption about the amount of noise (i.e. lower prior noise precision mean). This would need to be combined with a tight prior. One could then see it as the data needing to act 'against' this prior. In other words, there needs to be evidence in the data such that the posterior distribution of the hyperparameters 'moves away' from the prior distribution. In mathematical terms, the quadratic pwpE of the noise then exerts a constraint on the relationship between Eh and F, eventually overcoming the linear term in Eq. (4.10), making lower residual variance less likely (in a free energy sense). We did not choose this approach because it seemed particularly difficult to find relationships that would hold in general. Also, it would involve again a strong assumption on the certainty of noise, which we explicitly wanted to omit. At least, not unless a good approximation to the level of SNR can be provided. We will illustrate in the supplementary material that one can actually approximate SNR in DCM for fMRI. Instead and based on the impressive performance in the first simulations, we opted to try to match the expected residual noise structure and the true noise structure. From an independent argument, we predicted that a sampling rate of 100 Hz should satisfy the assumptions. From a different perspective, it is a challenge on the independence assumption of the datapoints. In our results, the reduction was clearly beneficial in terms of model recovery (increase up to 38%). However, the benefit rather arose from a better balance between datapoints and prior constraints and we could not prevent overfitting. This is definitely something to consider in the future. One alternative could be to find a way to invert the almost ill-conditioned covariance matrix, as explained in the methods. Another way could be, to regress out expected noise structure (as in confound regression), for example with the help of a Gaussian process with an appropriate kernel.

Third, multimodality is clearly present in these models. We looked at the convergence of one of the optimizations and it appears as if the models actually find a good solution after fairly few iterations. **Figure 31** illustrates the true steps in negative free energy over a couple of optimizations. It seems, as if a good solution is found very quickly, followed by a long period of only marginal updates and many rejected steps. This speaks again for a rather complex landscape. Also, the solution can be very different across starting values. This can be a challenge and possible shortcoming of the multistart. While it consistently did result in parameter estimates closer to ground truth, it also resulted in larger between-'subject' variance of the estimates and showed to be more affected by overfitting. Overfitting was even an issue after down-sampling, which creates arguably a better balance between accuracy-complexity. However, to truly avoid overfitting, one would have to consider a prior setting which assumes a more conservative SNR. This does, by no means, make the multistart not useful. From a strict Bayesian perspective, it will inevitably lead to a solution that is as good or better in terms of the negative free energy. But given the complexity of the optimization problem, it seems unavoidable to find additional ways to constrain the models. One example is hierarchical models. Interestingly, the DEFAULT starting values also seemed to exhibit an artificial way of constraining by forcing solutions to be closer to the prior mean and might tie into the literature on 'early stopping' (Caruana, Lawrence et al. 2001). It performed consistently better in terms of model recovery, resulted in lower variance across subjects, but at the cost of higher RMSE. But given the fact that a non-exhaustive search of the starting value space cannot guarantee a truly globally optimal solution, it might be worth considering to exploit when classifications or parameter statistics are the focus of interest. However, one might need to reconsider the interpretation of parameter estimates. But given the fact that the true set of parameters does not necessarily correspond to the optimum in free energy terms, and considering the effects of correlations between parameters, it does question to which extent single point estimates can truly be attributed uniquely to effects. This does not mean that they do not resemble meaningful features, but rather have to be analyzed in a combined manner (e.g. classification).

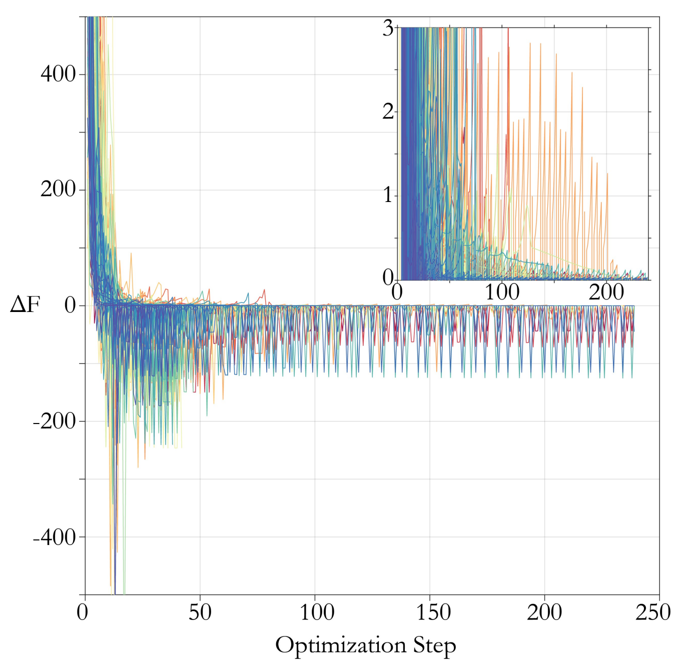

Figure 31| Updates (ΔF) of the negative free energy over the course of the optimization. Shown for all datasets (20), simulation (3) and inversion models (3), i.e. 180 optimizations. Negative steps are rejected. Zoomed in box for visualization of smaller steps.

4.6.1 General Remarks

We have created a relatively extreme example, where the models were able to predict a large amount of variance. From a mathematical view, the relationships are less steep in more moderate regions of vE. Generally, it illustrates that even generating a synthetic setup is not straightforward without creating scenarios that do not resemble empirical data at all. Please note that we did not plan on creating any bias towards either side. This includes the simulations with structure in the data that goes beyond an AR(1) process with $\varphi = 1/2$. The noise covariance structure implemented in SPM12 (ver. 6906) would in principle allow for different values of the residual covariance. However, estimating φ is not provided as an option in the current implementation. Theoretically, the autocorrelations should, at least, depend on the used sampling rate. Maybe it was originally motivated by the tutorial dataset which is sampled at 200 Hz, hence much closer to what we have considered here as an optimal scenario (Garrido, Kilner et al. 2007). But to our knowledge this dependency between the decay rate of the noise, sampling frequency and filtering has not been discussed in the literature. However, an AR(1) process will always lead to an exponentially decaying covariance matrix which will never allow for strong, 'short-temporal' correlations as we would expect from the filtered models. Thus, the frequencies for AR(1) process with any time constant would most likely never be in a range that can be explained by the model. For all these scenarios, we would hypothesize that model recovery would be similar to our AR(1) simulations (provided adequate priors are chosen).

We also did not plan on biasing the results by choosing such a high sampling rate of 1 kHz. Its explicit quantitative implications only became clear once we diagnosed the results. Again, to our knowledge, it is not something that the DCM community is made aware of. In the end, it is a known fact about Bayesian models to tend to converge towards a Maximum Likelihood setting in the presence of many datapoints. But it does make the number of datapoints a decision of impact for the user. This includes choices about down-sampling during preprocessing, the length of the modeled time window or how many features (PCA-components) should be chosen, when one deals with scalp data.

To conclude: For high (and tight) expectations about prior noise precision and in the presence of many datapoints the relationship between goodness of fit and negative free energy gets very steep. This causes strong demands on the optimization in order for inference to be 'correct'. (Here, 'correct' is meant 'correct' in a strict sense, i.e. independent of the starting value of the optimization and respecting the numerical implications). First, the true, global optimum must have been found. Secondly, the posterior noise precision must be estimated to the respective precision. The first point is very challenging and questionable if we take the results from the multistart chapter into account. Ultimately, this does cast some doubt about the applicability of the standards threshold for the negative free energy, at least in EEG. However, one needs to keep in mind that this is something that also occurs in other fields of statistics as well. Even in simple linear regression, very low effect sizes can be significant with enough datapoints. This does not mean that anything would be wrong with the mathematical principles behind the statistics, but the interpretation of the findings might need to be put into perspective.

We want it to be clear that neither this nor any previously mentioned points should be understood as a critique on the developers, or on any other study. All results are valid given the assumptions and choices, and it is the modelers responsibility to defend them. If anything, it shows that we ourselves sometimes failed to make sensible choices, or were not aware of all implications. Ultimately, the goal of this chapter was to illustrate exactly those.

4.7 Supplement A: Noise Modeling in the empirical studies

In this section we present the diagnostic figures introduced for synthetic data, for all analyses performed on real data in this thesis. This includes the RATMPI and PRSSI study, hence cases for both EEG and fMRI. Importantly, these two projects exhibited completely different scales of SNR.

4.7.1 RATMPI

For study specific information, we refer to the corresponding chapter in this thesis. For simplicity, we will only illustrate the free energy diagnostics for the vehicle condition in the pharmacological dataset of this study. Briefly, we modeled two sources in primary and secondary auditory regions. Sampling rate was set to 1 kHz, the time window of interest lasted 250 ms in two conditions. Collectively, this amounts to $N = 1000$ datapoints. We used a standard ERP model based on the canonical microcircuit. In total, seven rodents with data from two hemispheres each were included. The model space consisted of 16 models where extrinsic and intrinsic connections were modulated by condition in a 2×2 factorial fashion. We also employed a multistart approach. The noise hyperpriors were adjusted to $hE = 4,\ \ hC = 1/2$, informed by inversions performed on an independent dataset. **Figure 32** shows negative free energy diagnostics for all 16 models (averaged across rats/hemispheres).

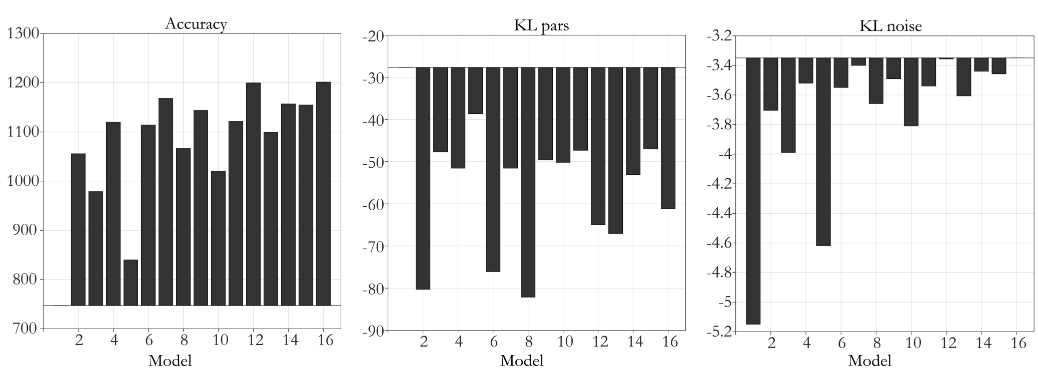

Figure 32| Diagnostics of the terms contributing to the negative free energy for the RATMPI study. Figure illustrates averages over two hemispheres and seven rodents. Only the vehicle condition of the pharmacological part of the dataset is displayed.

From a quantitative perspective, there are two aspects we want to draw attention two. (i) We do observe a decrease (trend) in noise-KL complexity with increasing model complexity. However, since we 'widened' the noise prior (hC =$1/2$), the numerical impact is much less severe than observed in the simulations. Overall, the values are reassuring, as there are about 20 times more neuronal/forward parameters than noise parameters, the respective KL-complexities show an adequate balance. The model with the highest complexity (*m16*) includes four additional modulatory parameters compared to the simplest model (*m01*). The increase in complexity is about $KL(m_{16}) - KL(m_1) \approx 30$, which is higher than an AIC / BIC increase (8 and 13.8 respectively). Of course, we do find ourselves in a setting where the number of datapoints greatly exceed the number of parameters (however less severe than in the synthetic setting), and also here, slight changes in fit can lead to considerable changes in ***F***. Additionally, we did observe residual autocorrelation structures that do not match the model (similar to **Figure 22**B). Unfortunately, some of the implications of this chapter became only apparent after the analysis of the dataset. However, as we will show in this study, model selection was not of primary interest. Instead, we focused on the ability to predict the pharmacological conditions based on parameter estimates. As we have seen in the synthetic data, parameter recovery seems less effected by some strong relationships and we used permutation testing in the classifications, which generally protects against overfitting.

4.7.2 PRSSI

In this study, we used a classical, bilinear DCM for fMRI to investigate working memory (WM) connectivity changes in a fronto-parietal network, using four sources in the left and right prefrontal and parietal cortex (PFC and PAR), respectively. The regressors of interest consisted of three phases, encoding (SAMPLE[35]), retention (DELAY) and retrieval (PROBE). The task also consisted of a control (CONTROL) condition, where during the DELAY phase, no retention of the SAMPLE stimulus was needed. During time series extraction, we regressed out all PROBE related variance (and the mean), keeping only DELAY and SAMPLE related variance and using the two as independent driving inputs into parietal regions. Additionally, DELAY during the memory (MEMORY) condition

[35] We use the all caps notation to refer to factors of the design e.g. experimental manipulations.

acted as modulator of intrinsic and extrinsic connections. In summary, the model space consisted of 16 models differing in how DELAY MEMORY could modulate the connectivity between PAR and PFC or act via local self-inhibition in a factorial fashion, i.e. $2^4 = 16$ models. Each condition consisted of 304 measurements per voxel (TR=2.5s, 12.5 min scanning time), resulting in ($N = 2432$ datapoints) over the two conditions and four regions. In total, 46 participants were included in the analysis.

In a first stage, we used the default settings of SPM12, ver. 6906, which makes the a priori assumption of $hE = 6, \quad hC = 1/128$. Despite the different normalization of the data, this corresponds to very high SNR, with very high precision.

Figure 33 illustrates the decomposition of the negative free energy. The values alone of the noise-KL indicate that the prior is badly specified (noise-KL is over 15 times larger than parameter-KL). Under the default noise prior, the estimated posterior noise precisions were very unlikely. In turn, this implies that the quadratic noise-pwPE term is very steep. In other words, slight increases in Eh can lead to large improvements in F, which in turn can again outweigh the parameter-KL term. In other words, models are not properly constrained anymore, which can lead to a bias in favor of more complex models. Unsurprisingly, under this prior, we found clear evidence in favor the most complex model.

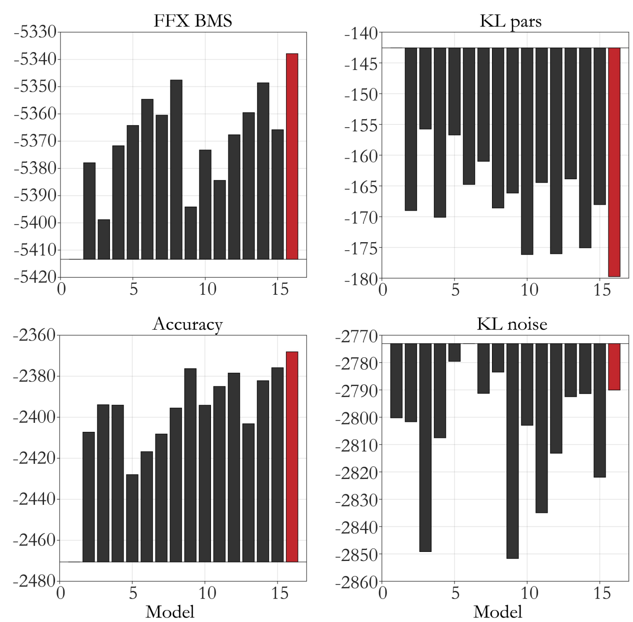

Figure 33| Diagnostic of the terms contributing to the negative free energy for the PRSSI study. Results shown for the default hyperparameters. All quantities shown are averages over N=46 participants (averages also shown for FFX BMS). Red bar indicates winning model from a FFX BMS perspective.

Of course, in empirical data ground truth is not known. However, in the case of DCM for fMRI, one can compute an a priori approximation to the expected range of hyperparameters. We outline the procedure in detail in the chapter devoted to the study. In brief, based on the GLM used for time series extraction, one can compute the amount of variance explained by this multivariate linear model. Of course, the GLM is a simplified model when compared to the DCM and does not account for local differences in the BOLD response (unless temporal derivatives are included), but it can still serve as a proxy. From this analysis, we could already intuit the much lower regime of SNR than expected by default (see **Figure 58**B). We therefore used the average (and variance across participants) explained variance from the GLM to inform prior mean and variance for the hyperpriors. Based on this selection, we re-ran all inversion.

The same diagnostic plots show now a much reduced effect of the prior (**Figure 34**). Importantly, FFX BMS is not only accuracy driven anymore, as the winning model *m08* is actually lower in accuracy than the most complex model. But due to a negligible difference in the noise-KL and a meaningful difference in parameter-KL, the *m08* is deemed superior in terms of the accuracy-complexity tradeoff. In summary, we would argue that such a balance is much more reassuring that the conclusions drawn are not strongly driven by a priori assumptions.

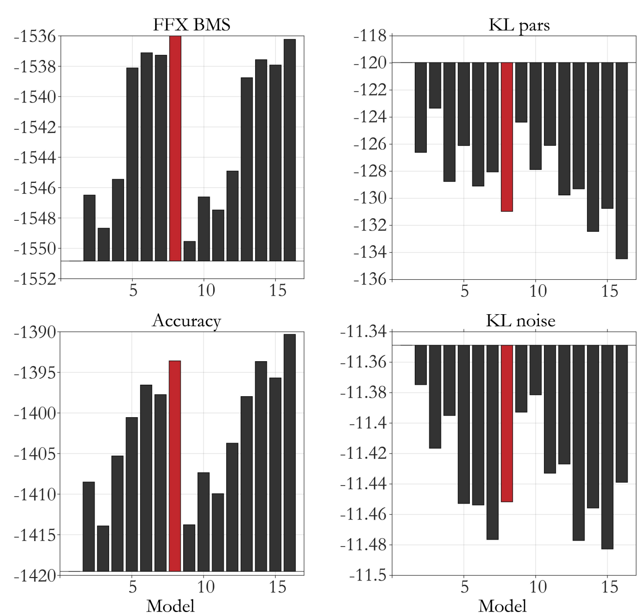

Figure 34| Diagnostic of the terms contributing to the negative free energy for the PRSSI study. Results shown for the adjusted hyperparameters. All quantities shown are averages over N=46 participants (averages also shown for FFX BMS). Red bar indicates winning model from a FFX BMS perspective.

4.7.3 Discussion

With this supplement, we illustrate that the theoretically discussed points for consideration translate to empirical data. Importantly, the severity of the effects discussed in this chapter on noise hyperpriors will depend greatly on the SNR of a particular dataset, on the prior settings and on the structure of the noise. It is highly recommended to check in a given inversion whether changes in free energy are completely dominated by one of the components, e.g. by fit. It is clear that the range of reasonable hyperpriors differs dramatically between different modalities (EEG and fMRI) which lead to completely different SNR scenarios. Hence, a default setting in a software, cannot provide a one size fits all solution. While we cannot provide a simple recipe for all cases, we do think that diagnostics of the sort presented, could help understand some results often encountered, for example when the most complex model is consistently selected. Clearly, the hyperparameters do deserve some special attention. This is the responsibility of a modeler. In retrospective, one could without doubt question whether in the analyses of experimental we always made the most sensible choices. Having said this, we would once more like to emphasize that the results derived by any inversion are correct in the sense that they provide the optimal solution under a given prior and under the noise assumptions (and starting value). If one or both of these components are off, model inversion might be biased. This bias led, in all cases discussed here, to more complex models being selected more likely.

4.8 Supplement B: Technical aspects

We end this chapter on hyperpriors by illustrating one final time the meaning of the dependencies from a different perspective.

Based on Eq. (4.9), we can discuss the parameters that influence the estimate of Eh directly, by looking at their effects on the intersection point. **Figure 36**, B-D illustrate changes to the intersection point when changing different variables. To give a brief intuition:

- A-priori expectation about the noise precision acts as an offset of the linear part. For example, higher prior noise precision then also pulls the estimate towards higher values.
- A-priori variance of the noise precision influences both, the offset and the slope of the linear function. Hence, its influence depends on the values of N and hE.
- The prior generally acts more in regimes where Eh is low.
- Z depends on the residuals and scales the slope of the exponential function. Low Z means lower slope and for a given linear function, Eh increases.

All of these intuitions are as expected from a simple Bayesian perspective.

In **Figure 35**, we have plotted different scenarios of the relationship between vE and Eh under the assumption of the true residuals following an AR(1) process. In comparison to **Figure 36**, they have to be understood as being mediated by changes of Z, scaling the exponential part of (4.9). Generally, we observe that if expected covariance between residuals is *higher* than the true correlation (warm areas), the exponential function gets steeper, i.e. the noise precision *decreases* (see blue curve in **Figure 35**C). If the expected covariance between residuals is *lower* than the true correlation (warm areas), the noise precision *increases* (see green curve in **Figure 35**C).

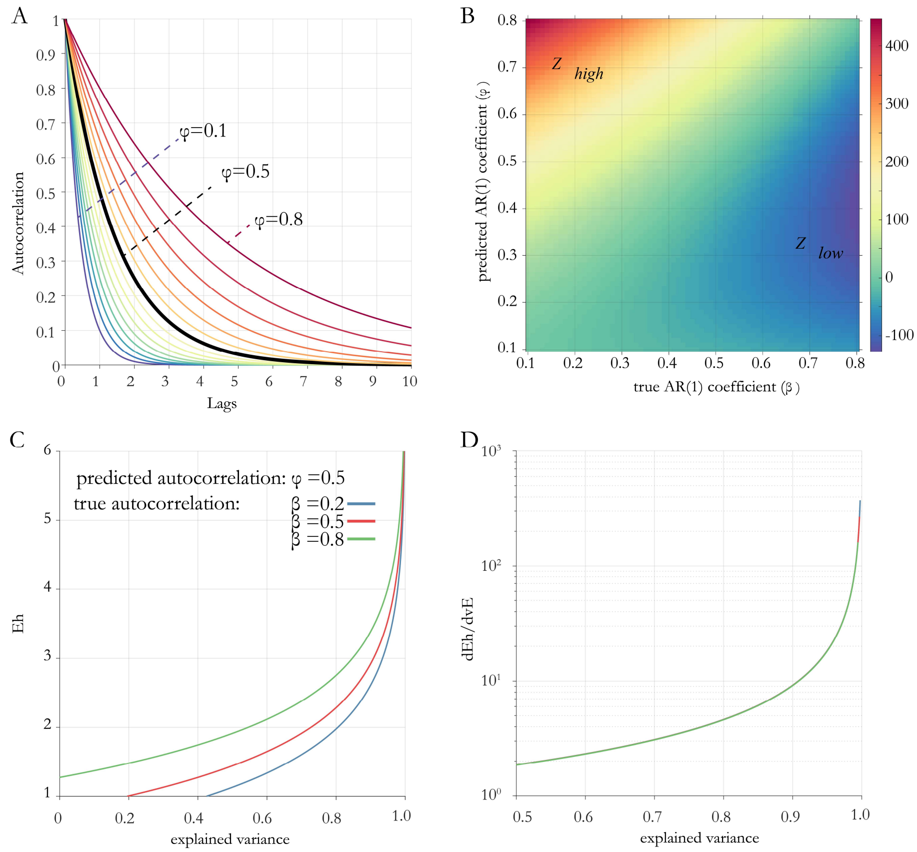

Figure 35| Impact of autoregressive coefficients on the estimation of vE and the ensuing Eh. A) Decay of autocorrelation for AR(1) process with different coefficients. Black line indicates current default in SPM12. B) Changes of Z for overestimating (warm area) and underestimating correlation between residuals (cold area). The heatmap is normalized to the diagonal (predicted correlation = true correlation). C) Changes in relationship between vE and Eh for over/and underestimation of residual correlations (according to Eq. (4.16)). D) Derivative of C.

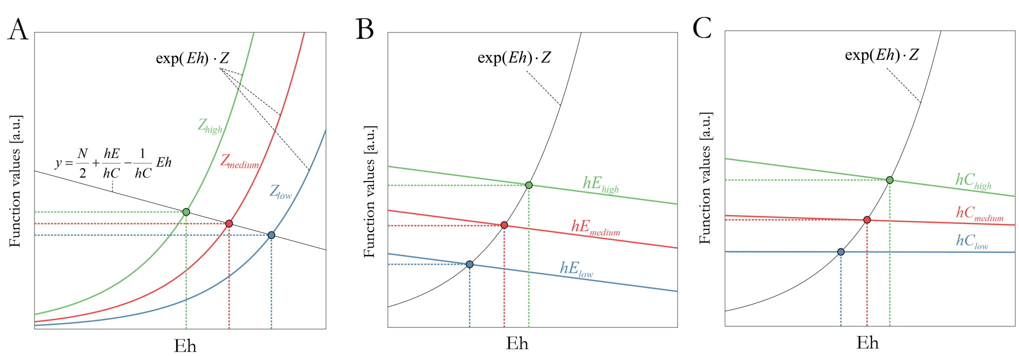

Figure 36| Approximation to the expected posterior precision, and dependency on different parameters. Relationship only shown as an illustration, and do not correspond to a specific setting in DCM. Therefore, no axes values are displayed. A)-C) X-Coordinate of intersection point correspond to approximate posterior noise precision (mean).

4.8.1 DEPENDENCY BASED DIAGNOSTICS

We will briefly revisit the results presented for the simulations with AR(1) noise and tight noise prior from a different perspective. This does not contain more information that the results we already discussed **Figure 24**, but it provides an additional view that focuses more on the dependencies between the quantities. In line with Eq. (4.9) and Eq. (4.10), we would approximately expect the following quantitative dependencies. (Please note that we have simulated with an SNR of 7 (true $vE \approx 0.974$))

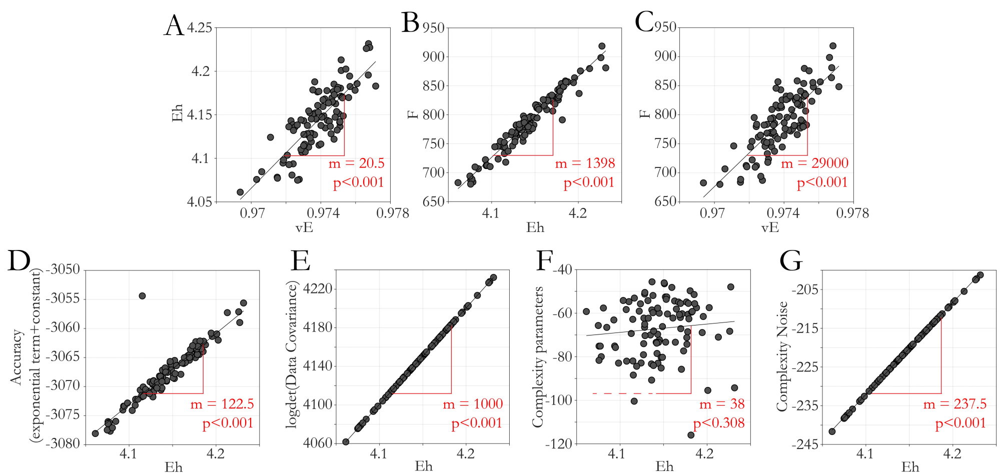

Figure 37 | Relationship between explained variance (vE), the posterior noise estimates (Eh) and the quantities of the negative free energy. A single dot represents the inversion of a single dataset with the true, data generating model (100 points). Equations indicate the slope of the regression line, p-Values the significance of the slope (uncorrected)

There are six relationships that we focus on in **Figure 37**: (i) vE vs Eh (A); (ii) Eh vs the exponential part of the accuracy; (iii) Eh vs the log-determinant of the data covariance (E); (iv) Eh vs noise complexity (G); (v) Eh vs F (B); (vi) vE vs F (C):

(i) ***vE* vs *Eh* (Figure 37A):**

Expected Relationship: visual readout of **Figure 35**C,D for $vE = 0.974, \varphi = \beta = 1/2$)):

$$Eh = 4.15; \frac{\partial Eh}{\partial vE} = 34.75$$

Empirical Relationship:

$$m = 20.5$$

Comment: There is some level of approximation error, since the derivative for such a high SNR scenario gets really steep, and a mismatch between the true covariance structure of the residuals and the predicted (mediated trough β) can lead to errors (also see **Figure 36**, A).Quantitatively, we are below a factor of ~30-40 as predicted by the approximation in **Figure 35D**, nonetheless, the regression line shows a slope in the double digits.

(ii) **Eh vs. exponential part of the accuracy (Figure 37D):**

Expected Relationship: (Eq. (4.9))

$$\frac{\partial\left(-\frac{1}{2}\cdot \exp(Eh)\, e_y^T Q e_y\right)}{\partial Eh} = \frac{1}{hC} = 128$$

Empirical Relationship:

$$m = 122.5$$

Comment: This matches really well and differences come most likely from an approximation error.

(iii) **Eh vs. log-determinant of accuracy (Figure 37E):**

Expected Relationship: (Eq. (1.10):

$$\frac{\partial\left(\frac{N}{2}Eh\right)}{\partial Eh} = \frac{N}{2} = 1000$$

Empirical Relationship:

$$m = 1000$$

Comment: This relationship is analytically clear.

(iv) **Eh vs. noise complexity (Figure 37G) :**

Expected Relationship: Eq. (4.10)

$$\frac{\partial\left(-\frac{(Eh-hE)^2}{2hC}\right)}{\partial Eh} = -\frac{Eh-hE}{hC}\bigg|_{Eh=4.15} = 235.52$$

Empirical Relationship:

$$m = 237.5$$

Comment: This relationship is also analytically clear. Error comes from not considering the posterior noise covariance term of the noise KL-divergence.

(v) ***Eh* vs. *F* (Figure 37B):**

Expected Relationship: (sum of partial derivatives in (ii)-(iv)

$$\frac{\partial F}{\partial Eh} = 1363.5$$

Empirical relationship:

$$m = 1398$$

Comment: The slope is overall dominated by the number of datapoints, the prior noise precision and the quadratic noise pwpE term. The approximation error comes from the unknown contribution of the parameter KL-divergence.

(vi) ***vE* vs *F* (Figure 37C):**

Expected relationship:

$$\frac{\partial F}{\partial vE} = \frac{\partial F}{\partial Eh} \cdot \frac{\partial Eh}{\partial vE} = 47382$$

Empirical relationship:

$$m = 29000$$

Comment: This relationship obviously yields the largest approximation error. It is mainly an amplification of the approximation error discussed in (i).

Our approximations predict the empirical relationships very well. As mentioned, we lack a good approximation to how parameter complexity changes by producing a more accurate fit.

Across all models, there clearly is a steep relationship between vE and F, and a simple linear regression estimated a slope of $m = 29000$. Put simply, for the given range of SNR, improving the fit by only 0.01% would have led to an increase in F of the common threshold of $\Delta F = 3$. Or, as we observed between DEFAULT and BEST starting values, an increase in fit of 0.5% increases the negative free energy by approximately 150. With 20 datasets, the differences in the fixed effects BMS are easily in the observed order. Under this premise, the bias towards more complex models is not at all surprising, it is simply easier for them to fit better due to added degrees of freedom. At this level, results can also become very sensitive to the exact value of the posterior noise precision. With a slope of $m \approx 1400$ estimates must be accurate up to 10E-3.

5| MODEL-BASED PREDICTION OF MUSCARINIC RECEPTOR FUNCTION FROM AUDITORY MISMATCH NEGATIVITY RESPONSES

5.1 DISCLAIMER

This chapter contains a manuscript that is currently in preparation for publishing. The empirical data for this chapter were acquired at the Max-Planck Institute of Cologne as part of the Doctoral Thesis of Fabienne Jung. Analysis of the data was done under the supervision of Klaas Enno Stephan and Jakob Heinzle. Analyses were performed according to an analysis plan which is provided in Appendix A.

The second part of this chapter contains additional analysis that are, as of now, not planned for publication.

MODEL-BASED PREDICTION OF MUSCARINIC RECEPTOR FUNCTION FROM AUDITORY MISMATCH NEGATIVITY RESPONSES

Dario Schöbi[1], Fabienne Jung[3], Stefan Frässle[1], Rudolf Graf[3], Heike Endepols[3], Rosalyn J. Moran[4], Marc Tittgemeyer[3], Jakob Heinzle[1,*] & Klaas Enno Stephan[1,2,3,*]

[1] Translational Neuromodeling Unit, Inst. for Biomedical Engineering, Univ. of Zurich & Swiss Institute of Technology (ETH Zurich), Wilfriedstrasse 6, 8002, Zurich, Switzerland.

[2] Wellcome Centre for Human Neuroimaging, University College London. 12 Queen Square, London, WC1N, 3AR, UK.

[3] Max Planck Institute for Metabolism Research, Gleueler Strasse 50, 50931 Cologne, Germany.

[4] Department of Neuroimaging, Institute for Psychiatry, Psychology & Neuroscience, King's College London, De Crespigny Park, London Se5 8AF, UK.

* These authors contributed equally to this work

Corresponding authors:

Dario Schöbi
University of Zurich & ETH Zurich
Translational Neuromodeling Unit (TNU)
Institute for Biomedical Engineering
Wilfriedstrasse 6
8032 Zurich, Switzerland
Phone: +41 44 634 91 12
E-mail: dschoebi@biomed.ee.ethz.ch

5.2 ABSTRACT

Drugs affecting neuromodulatory transmitters, such as dopamine or acetylcholine, take centre stage among therapeutic strategies in psychiatry. Such drugs are known to change both neuronal gain and synaptic plasticity and therefore affect electrophysiological measures. An important goal for clinical diagnostics is to exploit this effect in the reverse direction, i.e., to infer the status of specific neuromodulatory systems from electrophysiological measures.

In this study, we provide proof-of-concept that the functional status of cholinergic (specifically muscarinic) receptors can be inferred from electrophysiological data using generative models. To this end, we used epidural EEG recordings over two auditory cortical regions during a mismatch negativity (MMN) paradigm in rats. All animals were treated, across sessions, with muscarinic receptor agonists and antagonists at different doses. Together with a placebo condition, this resulted in five levels of muscarinic receptor status. Using a generative model embodying a small network of coupled cortical microcircuits, we estimated synaptic parameters and their change across pharmacological conditions. The ensuing parameter estimates showed both, graded muscarinic effects and distinguishability between agonistic and antagonistic pharmacological conditions.

This finding illustrates the potential utility of generative models of electrophysiological data as computational assays of muscarinic function. In application to EEG data of patients from heterogeneous spectrum diseases, e.g. schizophrenia, such models might help identify subgroups of patients that respond differentially to cholinergic treatments.

5.3 INTRODUCTION

Many pathophysiological theories of psychiatric diseases emphasize abnormalities of neuromodulatory transmitters, such as dopamine or acetylcholine (Tandon and Greden 1989, Cohen and Servan-Schreiber 1992, Stephan, Baldeweg et al. 2006, Howes and Kapur 2009, Higley and Picciotto 2014). Indeed, the large majority of drugs used in clinical psychiatry affect synthesis, reuptake, or postsynaptic action of neuromodulatory transmitters. However, patients with the same diagnosis according to ICD/DSM often show great variability in their response to the same treatment, a likely consequence of pathophysiological heterogeneity under contemporary diagnostic classification schemes (Kapur, Phillips et al. 2012, Krystal and State 2014, Stephan, Bach et al. 2016) and highlights the need for clinical tests that pinpoint specific abnormalities of neuromodulation in individual patients.

The present study is motivated by pathophysiological theories of cholinergic (Stephan, Friston et al. 2009) and more specifically muscarinic, abnormalities in schizophrenia (Raedler, Bymaster et al. 2007, Scarr and Dean 2008). Empirically, both post-mortem and in vivo studies have provided evidence for abnormalities in muscarinic receptor availability (Raedler, Knable et al. 2003, Scarr, Cowie et al. 2009, Scarr, Craig et al. 2013). Importantly, a 'muscarinic receptor-deficit schizophrenia' (MRDS) subgroup was identified that is unrelated to treatment success, illness duration, gender or age (Scarr, Cowie et al. 2009). MRDS is defined by substantially decreased numbers of muscarinic receptors in dorsolateral prefrontal cortex and associated differences in gene expression and synaptic properties (Scarr, Cowie et al. 2009, Gibbons, Scarr et al. 2013, Scarr, Craig et al. 2013, Dean, Thomas et al. 2015, Scarr, Hopper et al. 2018).The existence of marked differences in muscarinic receptors across the schizophrenia spectrum has implications for treatment – not least because clozapine and olanzapine, two antipsychotics with particular efficacy but also side effects, have distinctive antagonistic activity at muscarinic receptors (Weiner, Meltzer et al. 2004, Raedler 2007) (for a comparative overview of antipsychotics, see (Kapur and Remington 2001)). Therefore, if muscarinic receptor status could be determined non-invasively and cost-efficiently in individual patients, this might guide personalized treatment selection.

Unfortunately, with the exception of specialized positron emission tomography procedures (that are not widely available, expensive, and expose patients to radioactivity), there currently do not exist non-invasive in vivo measures of neuromodulatory processes in humans. A recent proposal concerns the combination of biophysical modeling and electrophysiology/neuroimaging in order to construct computational assays of neuromodulation (Stephan, Baldeweg et al. 2006, Friston, Moran et al. 2013, Stephan and Mathys 2014). This approach rests on so-called “generative” models that describe how latent (hidden) neuronal population processes generate measured activity, e.g. electrophysiological measures obtained with magneto-/electroencephalography (MEG/EEG) (David, Kiebel et al. 2006, Moran, Pinotsis et al. 2013). As demonstrated previously, under suitably chosen experimental perturbations and using Bayesian techniques, these models can infer neuronal processes from measurements. If such model-based inference enabled one to differentiate between distinct abnormalities of neuromodulatory function, computational assays might support differential diagnosis and personalized treatment predictions (Stephan and Mathys 2014).

Generative models have been used in previous studies for inferring changes in neuromodulatory processes from electrophysiological data (Moran, Symmonds et al. 2011, Moran, Campo et al. 2013). These studies used dynamic causal modeling (DCM; (David, Kiebel et al. 2006)), a generative framework for inferring circuit-level processes from MEG/EEG data, and pharmacologically manipulated the synthesis of dopamine (Moran, Symmonds et al. 2011) and metabolism of acetylcholine (Moran, Campo et al. 2013), respectively. While demonstrating that the models identified biologically plausible drug-induced changes in synaptic parameters, the interpretability of these studies with regard to clinical utility is limited. This is for three reasons. First, experimental control over neuromodulatory status was limited: drugs were orally administered to human volunteers, with no control over individual differences in drug uptake and metabolism. Second, the pharmacological intervention was restricted to a single dosage and direction of perturbation. Third, and most importantly, none of these studies examined the model’s ability to predict neuromodulatory status of an individual and out-of-sample.

Here, we tested the feasibility of computational assays of neuromodulation, focusing on muscarinic receptor function. In this proof-of-concept study we strove to overcome the limitations of our previous work, using a rodent model where pharmacological interventions are better controlled and can be repeated in the same animal with different doses and drugs. For the experimental paradigm, we chose the auditory mismatch negativity (MMN) which is reliably impaired in schizophrenia (Baldeweg, Klugman et al. 2004, Umbricht and Krljes 2005, Erickson, Ruffle et al. 2016) and is sensitive to cholinergic manipulations (for review on the MMN, see (Garrido, Kilner et al. 2009)). Epidural EEG recordings were obtained bilaterally from primary and secondary auditory areas. Measurements were obtained telemetrically in awake rats, thus avoiding any confounds by anesthesia. Importantly, all animals underwent five pharmacological conditions: (i) two dosages of the muscarinic antagonist scopolamine, (ii) vehicle, and (iii) two dosages of the muscarinic agonist pilocarpine. The measured EEG activity was modeled as arising from the neuronal dynamics within a set of connected cortical microcircuits. The animal-specific parameter estimates of this generative circuit model served as features for subsequent out-of-sample predictions (“generative embedding”; (Brodersen, Schofield et al. 2011)). This approach allowed us to test whether dose-dependent changes in muscarinic receptor function could be predicted, based on estimates of neuronal processes in cortical circuits, from EEG measurements of individual animals. Rightful concerns have been raised about severe underestimations of confidence intervals for classification using small sample sizes (Varoquaux, Raamana et al. 2017, Varoquaux 2018). We are doubtlessly in such a scenario. Therefore, we resorted to rigorous permutation testing to compute p-Values, which are considered valid and should protect against overfitting even for small datasets (Varoquaux, Raamana et al. 2017, Varoquaux 2018).

5.4 Methods

5.4.1 Data Acquisition

The data for this study were acquired at the Max-Planck-Institute for Metabolism Research at Cologne, Germany. They have been described in a PhD thesis (Jung 2013). (For a detailed explanation of the acquisition protocol, see (Jung 2013)). In brief, electrodes were implanted over the primary auditory cortex (A1) and posterior auditory field (PAF) (secondary auditory cortex) in both hemispheres of ten black hooded rats. Following surgery, animals recovered for ten days. In five sessions, rats received different intraperitoneal injections: 1 or 2 mg/kg of the non-selective, muscarinic antagonist scopolamine, 3 or 6 mg/kg of the non-selective, muscarinic agonist pilocarpine, or NaCl (vehicle). In order to avoid interactions between treatments, drug injections were administered only every third day, in counterbalanced order across rats. An additional set of six animals received only the placebo treatment (non-pharmacological dataset). These additional data served to optimize the settings for the subsequent analysis of the pharmacological data. Importantly, they did not enter the main analysis.

Acoustic stimuli were delivered in a sound-attenuated cage using a Tucker Davies Technologies® (TDT, Alachua, USA) System 3 and two free-field magnetic speakers (FF1, TDT). Stimuli consisted of short sine tone-bursts, with bandwidths between 7-9 kHz and 16-18 kHz. In total, 1000 tones were presented at a frequency of 2 Hz with 10% deviant probability. Both bandwidths were used as the standard tone once, in two individual sessions per drug condition. All electrophysiological measures were pre-amplified and transmitted (wirelessly) to a high frequency receiver (TSE Systems GmbH, Bad Homburg, Germany). This setup allowed the rats to move inside the cage without constraint.

5.4.2 Analysis Plan

After analysis of the non-pharmacological data but prior to the analysis of the pharmacological data, a version-controlled and time-stamped analysis plan was created. This plan detailed the analysis pipeline ex ante (see Methods section). The analysis plan is provided in Appendix A.

5.4.3 Preprocessing

Preprocessing was implemented using Statistical Parametric Mapping SPM12 (ver. 6906) (Litvak, Mattout et al. 2011). Electrophysiological data were down-sampled to 1000 Hz (including an anti-aliasing filter), and band-pass filtered between 1 Hz and 30 Hz. Trials exceeding an amplitude of 500 μV were considered artefactual and excluded from the analysis. This (liberal) threshold was chosen based on a visual inspection of the single trial ERPs and comparing the average ERPs before and after artefact rejection showed negligible effects on the averaged waveforms. Finally, standard and deviant tone responses were averaged in a time window of 0 - 250 ms. As standard for MMN, we averaged over all corresponding trials from both sessions, thus removing any potential confounds due to frequency differences in standards and deviants.

All analyses were done individually for each hemisphere. Data from a given hemisphere were excluded if the recording in one of the channels (A1 or PAF) was considered faulty (assessed through visual inspection of the average ERPs prior to any statistical and model-based analysis). This led to exclusion of one left hemisphere and three right hemisphere recordings. We excluded a hemisphere for all pharmacological conditions, even if the recording was of poor quality in only one pharmacological condition.

5.4.4 Classical Analysis

The MMN paradigm followed a classical oddball design, where the definition of the 'Standard' tone frequency did not change throughout a session. We compared the full ERP time series (0 – 250) ms for each of the four electrodes in a fully factorial 2×5×N mixed effects ANOVA (N=9 for left, N=7 for the right hemisphere) , with fixed effects factors TONE[36]=[Standard, Deviant], PHARMA=[2mg scopolamine, 1mg scopolamine, vehicle, 3mg pilocarpine, 6mg pilocarpine] and their interaction, and ANIMAL as a random effect (indicated by the notation '1 | '):

$$Y_{\{tone,pharm\ \ ,animal\}} = \beta_0 + \beta_1 * TONE + \beta_2 * PHARMA + \beta_3 * 1|ANIMAL + \beta_4 * PHARMA \times TONE$$

In brief, we fitted the mixed effects ANOVA to every time point in the time window [0, 250] ms post stimulus, where we expected the MMN and thus potential drug effects. The selection of the time window was based on the initial analysis of the non-pharmacological dataset. We corrected for multiple comparisons using an FDR correction (over time but not electrodes) (Benjamini and Yekutieli 2001). Electrodes were analyzed separately. We only included rats, where all recordings (in all pharmacological conditions) were valid to have a balanced dataset.

All statistical analyses were computed using the open source statistical software R (ver. 3.5.2) and the packages *lme4, R.matlab* and *lmerTest.*

[36] The all-caps notation is used to explicitly refer to the factor in the statistical model.

5.4.5 Generative Modeling

We modeled the data using a convolution based DCM for electrophysiological data (David, Kiebel et al. 2006, Kiebel, Garrido et al. 2009). In this neural mass model, the average presynaptic firing rate (of a neural population) is transformed into a postsynaptic potential by a convolution operator, while the average potential is converted into average firing rate via a sigmoid activation function. Anatomically, we used a canonical microcircuit (CMC) model (Bastos, Usrey et al. 2012), where each cortical column (source) comprises two types of pyramidal cell populations, an inhibitory interneuron and an excitatory (spiny stellate) population (**Figure 38**A). The CMC naturally maps onto computations required for predictive coding (Bastos, Usrey et al. 2012), which provides a unifying idea about the computations underlying the MMN and supports cortical hierarchies such as our two-level model (A1 and PAF) (Lieder, Daunizeau et al. 2013).

In our setting, the DCMs consisted of two reciprocally connected sources, A1 and PAF. Driving input encoding auditory stimulation (by any tone) targeted region A1. Based on this basic structure, we explored a full factorial model space comprising all possible combinations of modulation by TONE (deviant vs. standard) on the forward connection, the backward connection and the intrinsic connection in both regions. This resulted in $2^4 = 16$ models (**Figure 38**B).

As described in the Supplementary Material in detail, we made a number of changes to the default implementation of the CMC in SPM12, motivated by extensive prior testing of the framework on the non-pharmacological dataset. These changes included the use of a custom-written integration scheme for delay differential equations based on Euler's method. Furthermore, the priors for the main analysis of the drug data were informed by the inversion of the non-pharmacological data. Finally, in order to increase the chance of finding the global optimum, we ran the Variational Bayes inversion routine under a multi-start approach.

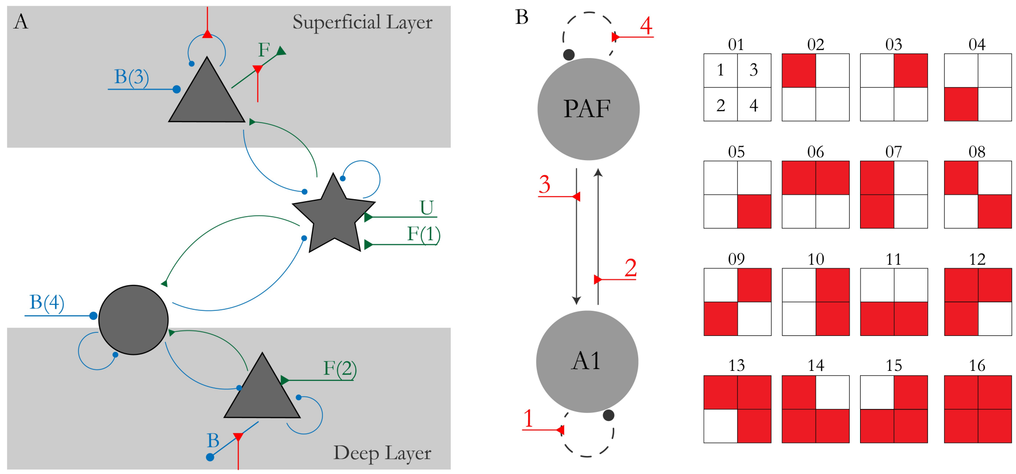

Figure 38| A) Connectivity pattern of the canonical microcircuit. Curved arrows indicate intrinsic (within region), straight arrows extrinsic (between region) or driving connections. Green color indicates excitatory (triangular arrowheads), blue color inhibitory connections (round arrowheads). Labelling indicates forward (F), backward (B) connections, or driving input (U) which entered A1. The index in the bracket refers to the corresponding parameter in the A-Matrix. Arrowheads show the direction of the connection. Red connections depict putative modulation by TONE (shown on the outgoing connections). Pyramidal cell populations are depicted by triangles, stellate cells by a star and inhibitory neurons by circles. B) Definition of model space. We consider 16 models, where connection strength (or excitability) can change by TONE. Red boxes indicate modulation by TONE, boxes 1 - 4 correspond to the connections on the left. The full modulation structure (*m16*) is not displayed.

5.4.6 Model Selection and Averaging

Model goodness was assessed in terms of the negative free energy, which provides a lower-bound approximation to the log model evidence (Friston, Mattout et al. 2007). We used random effects Bayesian Model Selection (BMS) (Stephan, Penny et al. 2009) to compute the posterior probability that a specific model generated the data of a randomly selected subject from the group. This quantity is of particular interest, when one wants to infer on model structure.

Our primary interest, however, concerned the potential representation of drug effects in the estimated model parameters. Bayesian Model Averages (BMA) were calculated on the individual animal level (Penny, Stephan et al. 2010) to marginalize out model uncertainty. In other words, for a given model parameter, BMA computes its average posterior distribution over all models considered, where this average is weighted by the posterior model probabilities. We used BMA estimates in all subsequent statistical tests.

5.4.7 Statistics and Classification

Statistical analyses of the drug effects focused on estimates of DCM parameters that have a biological interpretation in terms of synaptic properties. These include the connectivity (4), the kernel gain (6) and decay (4), and the modulatory parameters (4) (18 parameter estimates in total). For these parameter estimates, three different approaches were considered to test for pharmacological effects.

First, we computed a generic 1×5 ANOVA with a fixed factor DRUG and a random effect ANIMAL. We performed this test separately for each BMA estimate as dependent variable and used Bonferroni correction to correct for multiple comparisons.

Second, we investigated the drug-effect relationship, where we use the notion of a 'drug-effect' as the change in parameters estimates, as we move from the drug with the most antagonistic effect (2mg/kg scopolamine) to the drug with the most agonistic effect (6mg/kg pilocarpine). For this, we computed a mixed effects model for the same parameters used in the previous ANOVA, assuming a linear fixed effect of DRUG, $X_{drug} = [\dots, -2, -1, 0\ 1, 2, \dots]^T$ corresponding to 1 and 2 mg/kg of scopolamine, vehicle, and 3 and 6 mg/kg of pilocarpine respectively and a random effect of ANIMAL (hemisphere specific). The dots indicate different rats/hemispheres.

$$\theta_{\{animal,hemi,drug\}} = \beta_0 + \beta_{drug} \cdot X_{drug} + \beta_{animal} \cdot 1\,|X_{(hemi,animal)}.$$

For a single parameter vector θ (e.g. modulation of the forward connection), the values are ordered according to the subscript, i.e. *animal, hemisphere* and *drug*. Hence, every five consecutive entries in θ correspond to the five different drug conditions, from the most antagonistic to the most agonistic effect. Hence, X_{drug} codes for a linear effect over pharmacological interventions.

Third, we used a linear support vector machine (SVM) with leave-one-out cross-validation (LOOCV) to test whether one could predict the drug label from the model parameter estimates (Allwein, Schapire et al. 2000). Specifically, this was based on the same 18 BMA estimates used in the statistical tests described above. Hyperparameters of the SVM were optimized within each cross-validation set (nested cross-validation). In order to characterize the information that could be gained from the DCM parameters, we tested five different classifications: In four binary classifications, we compared the two extreme drug conditions against each other and individually against the vehicle condition, and the two antagonists vs. the two agonists. Finally, we applied a multiclass classification for all levels of the pharmacological factor. In order to be able to perform LOOCV in a balanced way, rats with data from only one hemisphere were omitted during classification. For more details on the classification procedure, see Supplementary Material.

5.5 RESULTS

5.5.1 CLASSICAL ANALYSIS

From the 20 recorded hemispheres (10 animals), the data of three right and one left hemisphere had to be excluded because of poor recording quality as diagnosed by visual inspection. For the remaining 14 hemispheres, all 5 pharmacological conditions were included in the analysis, resulting in 70 data points for the statistical analyses.

First, as described in the Methods, we ran a mixed effects ANOVA with fixed effects TONE and PHARMA, their interaction and a random effect of ANIMAL. We found prolonged effects of TONE, PHARMA, and their interaction (see **Figure 39**). These effects are consistent over the time window of interest, electrodes and deviant probability. The effect of TONE, visualized by the difference wave in **Figure 39**, exhibits two main peaks – an early negative peak around 25-50 ms and a "late" positive peak around 100-150 ms. It is also these two peaks that showed consistent interaction effects in all regions. Interestingly, the earlier peak is very dominant in the raw ERPs of both standard and deviant tones, most notably visible in the right hemisphere electrodes, with the two pilocarpine conditions exhibiting an additional dip right after 50 ms (see the arrow in **Figure 39**).

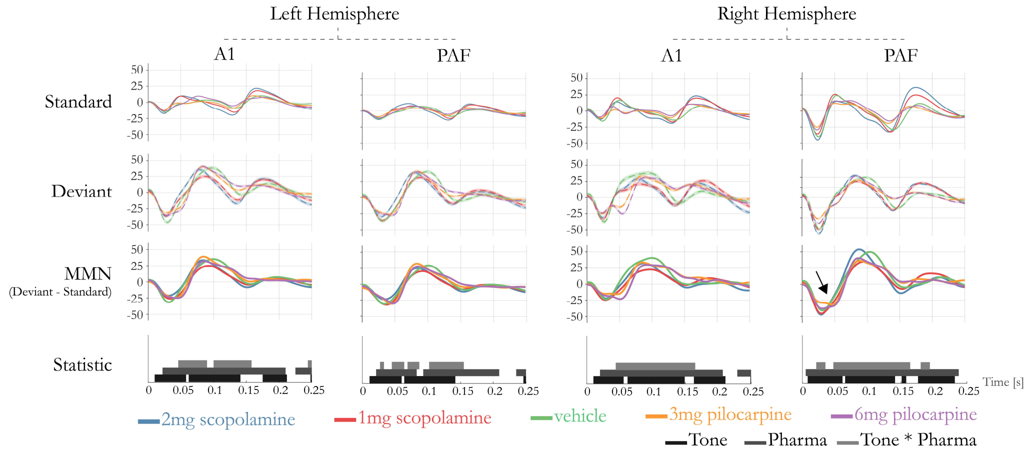

Figure 39| Grand Average Evoked Responses and results from the mixed effects ANOVA. Average (over animals and trials) Standard and Deviant tones are shown together with average difference waves for all drugs and both hemispheres (A and B). Statistical results are indicated by grey bars, whenever the effect (main or interaction effect) was significant at $p < 0.05$, FDR corrected.

5.5.2 Dynamic Causal Modeling

The animal-specific ERPs were modeled using DCM. Notably, for consistency with the classification results described below, we use only those rats where both hemispheres were included in the data analysis (N=14, i.e. seven rats, two hemispheres each).

Using the multi-start VB approach described in the Supplementary Material, we inverted each of the 16 models shown in **Figure 38** from 100 different starting values, for each rat, pharmacological condition, and hemisphere.

In terms of the primary measure of model goodness – the (negative) free energy – the multi-start approach was clearly beneficial (see Supplementary Material). This illustrates the multi-modal nature of the objective function and the necessity of our multi-start procedure.

Random effects model selection between the 16 competing DCMs did not yield a conclusive result (**Figure 40**A), although there was a tendency for more complex models to perform better, especially for the agonist conditions where the (protected) exceedance probability was approaching 0.95 (Rigoux, Stephan et al. 2014). The most complex model (model 16) performed consistently well across all pharmacological conditions. Runner ups were also models of increased complexity, such as models 7, 9, 11, 12, 15. Common to all these models is the presence of a modulation of the forward modulation.

The overall fit for the winning (most complex) model is illustrated in **Figure 40**B by comparing average prediction of the model (averaged over both hemispheres and rats) and empirical data. For reference, the average prediction would explain 88% (2mg scopolamine), 88% (1mg scopolamine), 93% (vehicle), 93% (3mg pilocarpine), 93% (6mg pilocarpine) of the average signal variance.

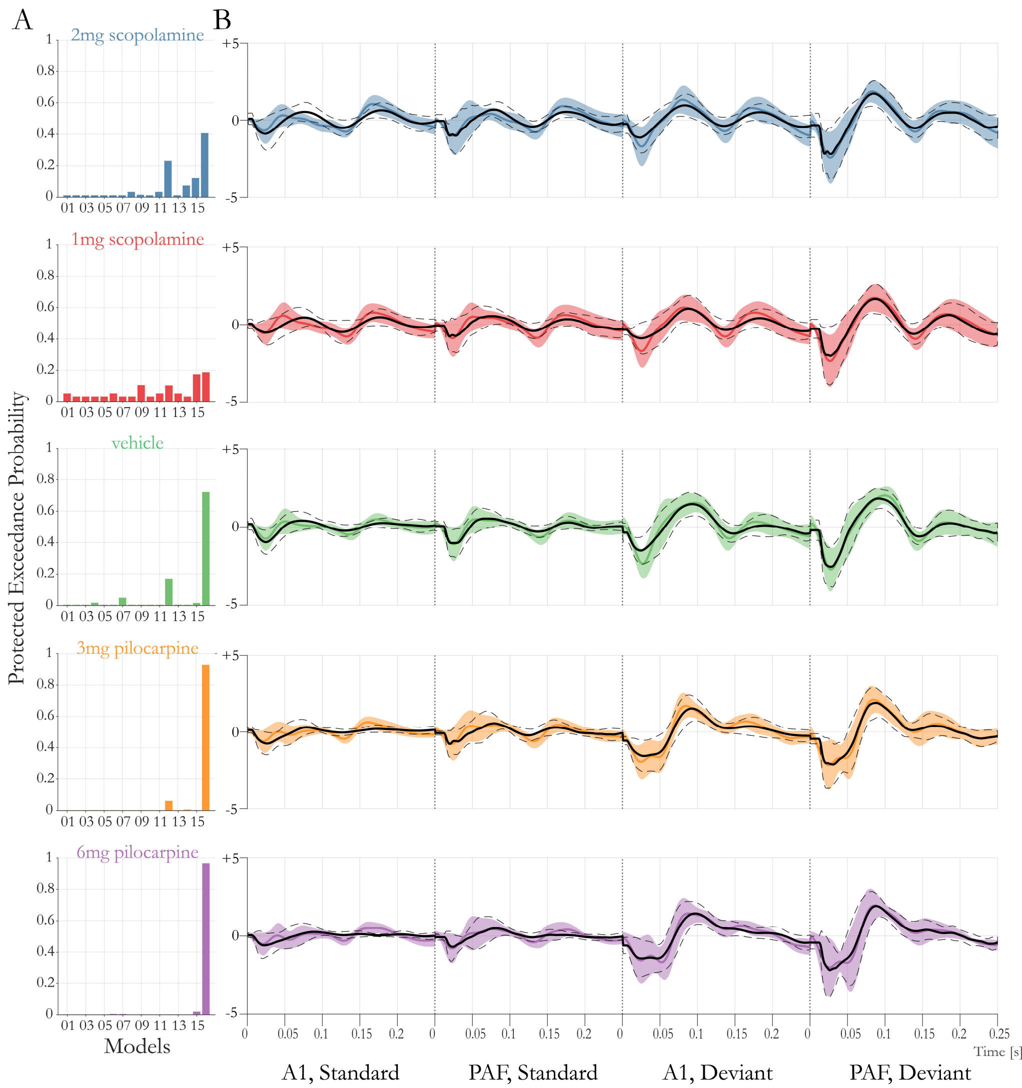

Figure 40| A) Bayesian Model Selection (BMS). Protected Exceedance probabilities reported for all sixteen models and drugs. B) Average (over animals) data (colored line) and prediction (black solid line) for model 16. Shaded area depicts standard deviation of the data (over animals and hemispheres), dotted lines depict standard deviation of prediction (over animal and hemispheres).

5.5.3 Parameter Estimation and Statistics

Since there was not a clearly winning model in all pharmacological conditions, we computed BMAs on the individual animal level, effectively marginalizing out the model from the posterior distributions. Our primary interest were parameters with a biological interpretation in terms of synaptic processes, i.e. extrinsic connection strengths, modulatory influences, kernel gain and decay (in total 18). We used these BMA estimates in two separate ANOVAs. First, we tested for any effect of DRUG, while correcting for the random effect of ANIMAL. Second, we tested for a linear effect of drug (i.e. across the different levels of muscarinic effects, from the highest antagonistic via vehicle to the highest agonistic dose). ANOVAs were computed for each parameter of interest and Bonferroni corrected for the 18 tests. The results are summarized in **Table 13** and **Figure 41**. For the one-way ANOVA with random effect ANIMAL, there was a significant effect on the kernel gain of self-inhibition of the superficial pyramidal cell in PAF, and on the kernel decay of the inhibitory cell, $p < 0.05$ (corrected). The latter parameter is set to be the same for both regions. When testing for a linear effect of drug, we observed a significant linear relationship in five parameters: The forward connection to the deep pyramidal layer (see F(2) in **Figure 38**), the modulation of the forward connections, the modulation of the backward connections, and the same two kernel parameters found in the previous ANOVA.

			classical		linear	
Class	**Connection**	**Parameter**	**F-Values**	**p-Values (uncorr.)**	**F-Values**	**p-Values (uncorr.)**
A Matrix (A)	SPC→SC	F(1)	1.3705	0.2551	5.3372	0.0242
	SPC→DPC	F(2)	2.5341	0.0496	10.1793	**0.0022 (-)**
	DPC→SPC	B(1)	1.0172	0.406	0.4711	0.4950
	DPC→IC	B(2)	0.6058	0.6598	0.9121	0.3429
Modulation (B)	A1→A1	1	2.8086	0.0334	5.8079	0.0189
	A1→PAF	2	3.8664	0.0074	12.455	**0.0008 (+)**
	PAF→A1	3	4.2405	0.0044	16.8168	**0.0001 (+)**
	PAF→PAF	4	0.5758	0.6813	0.0043	0.9479
Kernel Gain (G)	SPC→SPC (A1)	G1	1.1165	0.3564	1.9527	0.1668
	SPC→SPC (PAF)	G2	4.6626	**0.0023**	17.9828	**0.0001 (+)**
	SPC→SC (A1)	G3	0.8903	0.4755	1.5026	0.2249
	SPC→SC (PAF)	G4	1.3646	0.2572	0.6301	0.4303
	IC→SC (A1)	G5	1.5008	0.2123	4.8325	0.0313
	IC→SC (PAF)	G6	2.9784	0.0262	8.6147	0.0047

Kernel Decay (T)	SC	T1	0.7326	0.5734	0.793	0.3766
	SPC	T2	0.7238	0.5792	1.3634	0.2474
	IC	T3	4.62	**0.0026**	16.2289	**0.0002 (-)**
	DPC	T4	1.6404	0.1761	4.0477	0.0486

Table 13| ANOVA statistics on the BMA estimates for the classical ANOVA, and the ANOVA where DRUG was treated as a factor with a linear effect from the most antagonistic to the most agonistic drug condition. All parameters are ordered as in **Figure 41**. The connection is explicitly provided with the following abbreviations: Superficial Pyramidal Cell (SPC), Deep Pyramidal Cell (DPC), Inhibitory Cell (IC), Stellate Cell (SC). Modulatory effects act on connections as illustrated in **Figure 38**. Bonferroni corrected, significant results ($p<0.05$) in bold. Sign in bracket indicates direction of the linear effect.

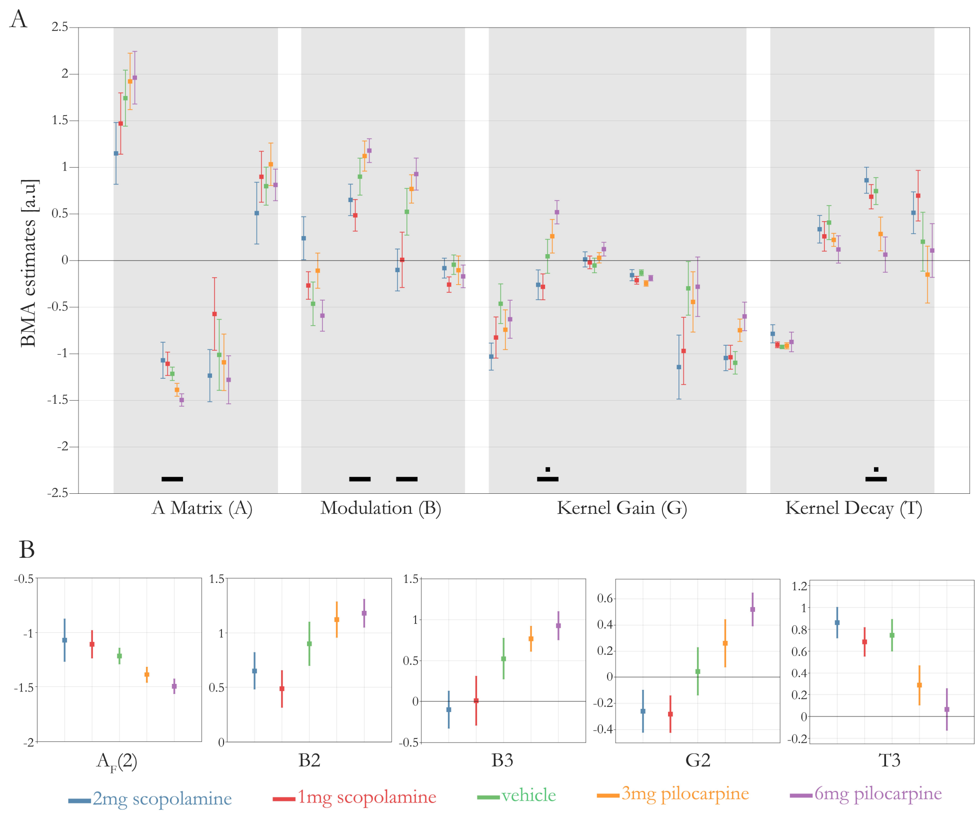

Figure 41| A) BMA estimates for all animals (n=7). BMAs are computed on the first level and pooled over both hemispheres. Errorbars depict SEM. Mixed effects (MFX) ANOVA on the BMA parameters are displayed. We considered two MFX models. First, a model with fixed factor DRUG (5 levels) and random effect ANIMAL. Black squares indicate significant results at $p < 0.05$ (Bonferroni corrected). Second, a MFX ANOVA with a linear, fixed effect of DRUG and random effect of ANIMAL. Black horizontal bars indicate significant results at $p < 0.05$ (Bonferroni corrected). B) zoomed in boxes of parameters showing a significant linear effect. Labeling according to **Table 13**.

5.5.4 Classification

Finally, we tested whether it was possible to predict the drug label (or even level) from the model parameter estimates. We used the BMA estimates as features for a linear SVM with

LOOCV. Here, in each fold, the classifier was trained on the drug labels of all but one rat and then the drug label of the left-out rat was predicted. We computed the balanced accuracy (BA) as performance score of classification and considered the five classifications described in the Methods. In conclusion, we were able to predict the individual drug levels in a multiclass classification with 31.4% BA (p = 0.024, chance level: 20%). Also, we could distinguish between the most extreme antagonistic and agonistic effects with 92.9% BA (p < 0.001, chance: 50%), between the highest dosage of pilocarpine and vehicle with 71.4% BA (p = 0.032, chance: 50%), and between both drugs with antagonistic and agonistic effects with 73.2% BA (p = 0.001, chance: 50%). Classification between the highest dosage of scopolamine and vehicle was not significantly different from chance, with 39.29% BA (p > 0.10. Classification results are summarized in **Figure 42**. All p-values reported here are based on permutation test on the drug labels and were not corrected for multiple comparisons. However, all classifications with p < 0.01 are significant when Bonferroni corrected for the five test under the chosen alpha level (indicated by two stars in **Figure 42**F).

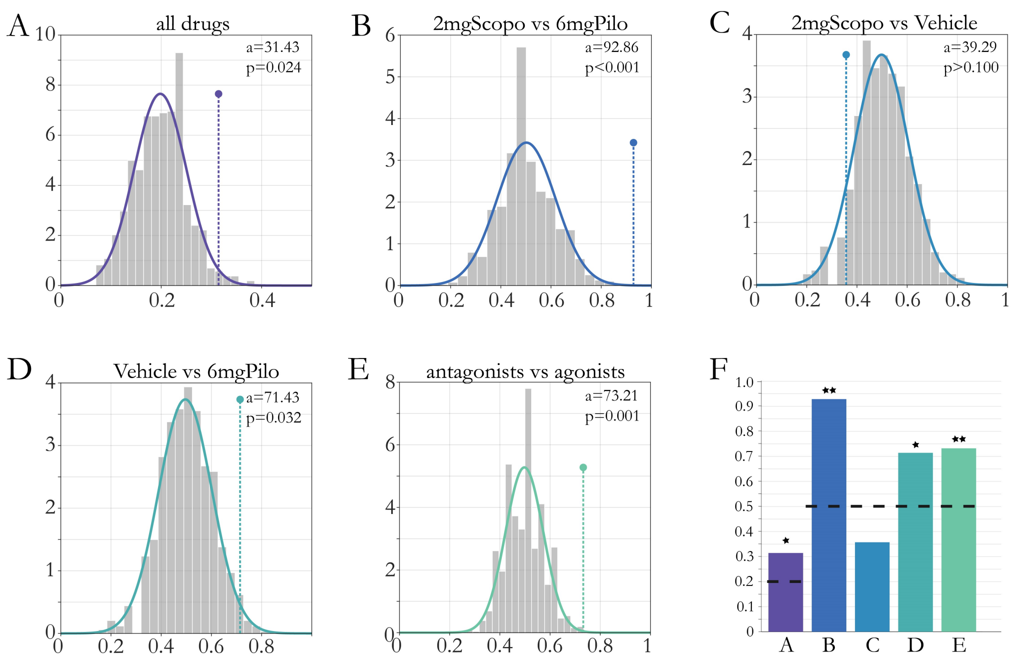

Figure 42| Permutation statistics for multiclass (A) and binary (B - E) classifications on the BMA results. Grey bars depict crossvalidation (CV) accuracies of permuted labels, solid line a gaussian fit on the histogram. Dotted line depicts CV accuracy for the true labels. Numbers refer to the CV accuracy of the true labels (a) and the percentage of a permutations leading to a higher accuracy (p). F) Balanced Accuracies for all classifications in (A-E). Stars indicate significance at p<0.05 (1 Star) and p < 0.01 (2 Stars) based on permutation statistics. The black dotted line indicates chance level for the specific classifications.

5.6 DISCUSSION

In this study, we investigated changes in epidural EEG recordings induced by graded pharmacological manipulations of muscarinic receptors during the auditory MMN. Using physiologically interpretable DCMs of auditory circuits, we were able to explain ERP changes across different levels of muscarinic receptor function in terms of underlying synaptic mechanisms likely affected by the drugs. We then identified several model parameter estimates that exhibited significant linear drug-effect relationships. Finally, we demonstrated that the estimated synaptic parameters allowed us to predict the type of drug (antagonist versus agonist) with nearly 93% accuracy and, less precisely, the dose under which a given dataset had been recorded.

Our model comparison results suggested that both forward and backward connections within a small auditory circuit consisting of primary (A1) and secondary (PAF) areas were modulated by the occurrence of a surprising tone (deviant). This result fits well with predictive coding accounts of the MMN, where surprising events lead to (precision-weighted) prediction error updates of an internal model, in order to minimize surprise (Baldeweg 2007, Garrido, Friston et al. 2008, Garrido, Kilner et al. 2009). More specifically, previous modelling studies of the MMN suggested that the occurrence of deviants modulate long-range glutamatergic connections as well as local gain adaption (Garrido, Friston et al. 2008, Moran, Campo et al. 2013). Our results are consistent with these findings, with slightly reduced emphasis on local gain modulation.

In addition, the modulation of the forward connection from A1 and PAF exhibited a linear drug-effect relationship. The backward modulation also showed a significant positive linear relation to drug level. In other words, the more strongly muscarinic receptors were activated, the stronger the increase in forward and backward connection for deviant tones. This finding is in contrast with a previous study in humans which investigated the effect of galantamine on the auditory MMN reported mainly local gain increases in A1 (Moran, Campo et al. 2013). It is possible that this difference is due to the different action of galantamine which not only increases the level of available ACh in general, but may also allosterically potentiate nicotinic receptors ((Samochocki, Höffle et al. 2003) but see (Kowal, Ahring et al. 2018). Physiologically, the drug-induced changes in long-range connections between auditory areas (which are glutamatergic and presumably draw on both AMPA and NMDA receptors; see discussion in (Schmidt, Diaconescu et al. 2012) could be mediated by short-term changes in synaptic transmission. Specifically, muscarinic agents are known to change AMPA and NMDA receptor function by various mechanisms, including rapid processes such as phosphorylation or changes in subunit composition (Marino, Rouse et al. 1998, Grishin, Benquet et al. 2005, Shinoe, Matsui et al. 2005, Di Maio, Mastroberardino et al. 2011, Lopes, Soares et al. 2013, Zhao, Ge et al. 2018, Zhao, Ge et al. 2019), for review, also see (Butcher, Torrecilla et al. 2009).

Linear but not deviant-specific pharmacological effects were found in two of the kernel parameters. To discuss these results in more detail, we consider their effects on the two pyramidal cell (PC) populations (see **Figure 40**), since those directly contribute to the measured EEG signal. On the one hand, we observed an increase in the kernel gain of inhibitory self-connections of the superficial PC in the PAF. This results in a smaller (in absolute values) initial peak of the ERP. On the other hand, there was a decrease in the kernel time constant of the inhibitory cell, i.e. faster decay. The inhibitory cell directly drives the deep PC but has no direct influence on the superficial PC. Since the deep PC is (intrinsically) driven only by the IC, a faster decay of inhibition will result in less deactivation in the deep PC. This, in turn causes the overall signal to decay less quickly back to zero after the first peak. Both of these results – a grading of the (absolute) amplitude and a graded decay back to zero can be observed in the ERPs in **Figure 39**, around 25-50 ms. A similar dichotomy of muscarinic action into a fast - net inhibitory - and a slower depolarizing effect was observed in vitro (McCormick and Prince 1985).

A central aim of the present study was to test the feasibility of predicting the muscarinic receptor status underlying a given dataset, out of sample and from the parameter estimates of a physiologically interpretable circuit model. The strategy of using parameter estimates from a generative model for subsequent (un)supervised learning is known as "generative embedding" (Brodersen, Schofield et al. 2011) and plays a central role in attempts to establish computational assays for psychiatry (Stephan, Schlagenhauf et al. 2017). This approach has two main advantages: it offers a theory-led dimensionality reduction (from high dimensional noisy data to a small set of model parameter estimates), and it enables the interpretation of machine learning results in terms of biological mechanisms represented by a model.

In this study, generative embedding suggested that a relatively simple model of a small cortical circuit can be used to predict muscarinic receptor status from EEG data. Interestingly, when considering the different pharmacological conditions separately, the most robust discrimination was obtained under the muscarinic agonist pilocarpine. That is, all classifications involving pilocarpine resulted in balanced accuracies significantly above chance, and the higher the difference in dosage, the better the classification. By contrast, distinguishing the muscarinic antagonist scopolamine from placebo proved more challenging. There could be several reasons for this, including drug differences of neuronally effective dosage regimes or strong non-linearities in drug-effect relationships.

While the classification accuracies for different dose levels are not yet close to clinically required levels of precision, the more general question whether muscarinic receptor function had been diminished or enhanced (antagonist vs. agonist) could be answered more decisively, with balanced accuracy above 90%. If this result could be replicated in a human EEG study with sufficiently large sample size, a computational assay for distinguishing hyper- vs. hypo-activity of muscarinic receptors might become plausible. As described in the Introduction, given the general importance of individual neuromodulatory differences in schizophrenia (Stephan, Friston et al. 2009), the likely existence of schizophrenia subgroups with differences in muscarinic receptors (Scarr, Cowie et al. 2009, Scarr, Hopper et al. 2018), and the distinctive anti-muscarinic properties of clozapine and olanzapine as two of the most potent antipsychotics (for a comparative overview of antipsychotics, see (Kapur and Remington 2001)), such an assay could eventually find important clinical applications for differential diagnosis and treatment selection in schizophrenia.

Our study has both strengths and limitations. Concerning strengths, it represents, to our knowledge, the first study using a graded manipulation of a neuromodulatory transmitter during the auditory MMN – from strong/weak inhibition via placebo to weak/strong enhancement. In terms of experimental approach, it uses highly selective drugs, obtains multiple recordings from both hemispheres, and avoids the common confound of anesthesia.

With regard to limitations, our sample size deserves consideration. While relatively large for rodent studies with in vivo recordings, the sample size is not sufficiently large for out-of-sample predictions with decisive robustness. Our statistical results should thus be taken with a grain of salt and need to be replicated in (human) studies of larger size. Another limitation is that the particular generative model used in this study does not allow one to directly map synaptic parameters onto a particular neurotransmitter system. In other words, there is no single parameter in our model that explicitly represents muscarinic function. Instead, it is likely that we are observing a net effect of pharmacologically altered muscarinic receptor function on several mechanisms represented in the model. For example, it is known that muscarinic receptors change glutamatergic synaptic transmission through influencing both NMDA and AMPA receptors (Marino, Rouse et al. 1998, Grishin, Benquet et al. 2005, Shinoe, Matsui et al. 2005, Di Maio, Mastroberardino et al. 2011, Lopes, Soares et al. 2013, Zhao, Ge et al. 2018, Zhao, Ge et al. 2019); an effect that can be (and was) observed in the estimates of model parameters encoding glutamatergic long-range connections. Similarly, muscarinic receptor activation strongly affects neuronal excitability and gain (McCormick, Wang et al. 1993, Shimegi, Kimura et al. 2016); his effect is captured by estimates of parameters representing the gain of postsynaptic kernels. To date, no generative models exist that represent neuromodulatory transmitter action directly, through distinct parameters, within the dynamical system that describes neural population activity. This represents an area of active ongoing research.

5.7 FUNDING AND DISCLOSURE

This work was supported by the René and Susanne Braginsky Foundation (KES) and the University of Zurich (KES).

5.8 SUPPLEMENTARY MATERIAL

5.8.1 Classical Analysis

For the classical analysis, all trials were separated into 'Standard' or 'Deviant' tones. While one might consider balancing the number of trials by, for example, using only every 'Deviant' preceding trial in the definition of a 'Standard' tone, we decided to keep all 'Standard' trials as this improves the calculation of the standard ERP. In addition, preliminary analysis on the non-pharmacological dataset suggested that both definitions led to very similar averaged ERPs, which were subsequently used for modeling.

5.8.2 Dynamic Causal Modeling

Dynamic Causal Modeling (DCM) was originally introduced in the domain of fMRI (Friston, Harrison et al. 2003), but has subsequently been adapted for the modeling of electrophysiological data (David, Kiebel et al. 2006), (Kiebel, Garrido et al. 2009). DCM provides a generative modelling framework for inferring neurophysiological processes from measured brain activity (imaging or electrophysiology) data. Model inversion provides parameter estimates that convey, to some degree, a physiological interpretation of the neuronal population processes underlying the generation of the data.

Because of the millisecond temporal resolution of electrophysiological measures, DCM for EEG allows for the modeling of networks of simplified cortical columns. Each cortical column consists of *intrinsically* connected neuronal populations, while columns are *extrinsically* connected to each other. The populations differ in their influences on other populations (effective connectivity), the strengths of excitatory and inhibitory synapses, their receiving of driving inputs, and in their contribution to the measured signal, (David, Kiebel et al. 2006), (Kiebel, Garrido et al. 2009).

In order to facilitate model inversion given the pharmacological data, we initially conducted an analysis of a separate non-pharmacological dataset (N=6) in order to constrain the subsequent model-based inference on pharmacologically induced circuit changes (compare **Figure 43** for an overview of the analysis strategy). Specifically, we informed the choice of priors for the analysis of the pharmacological data by model parameter estimates from the non-pharmacological dataset. To this end, we inverted all models for all rats of the non-pharmacological group (individually for the valid hemispheres). We then defined the prior mean $E[\pi(\theta)]$ for the pharmacological group as the average of the posterior means $E[p(\theta|y)]$ of the non-pharmacological group (averaged over models, rats and hemispheres). (As an exception, the prior for modulatory influences was kept at its default values. This served to maintain a conservative attitude by keeping a shrinkage prior, i.e., assuming the absence of modulation with small prior variance). Notably, there are two interpretations for this procedure. One can think of it as a Bayesian model average with equal posterior model probabilities.

$$E[\pi(\theta)] = E[p(\theta|y)] = E[\sum_{m=1}^{M} p(\theta|y,m)p(m|y)] = \frac{1}{M}\sum_{m=1}^{M} E[p(\theta|y,m)]\ .$$

Note that we use y as a shorthand notation for the data over all animals and hemispheres.

Alternatively, the expression above corresponds to a parameter average in the space spanned by the parameters that are common to all models. As for the prior variance, we used the variance of posterior means over rats, models and hemispheres

$$var[\pi(\theta)] = var\big[E[p(\theta|y,m)]\big].$$

This means that the prior acknowledges the expected between-rat variability.

Additionally, we reparametrized the leadfield gain matrix, to assume an equal forward gain from both regions or sources. This was as well motivated by the mean ERP signals in the non-pharmacological dataset.

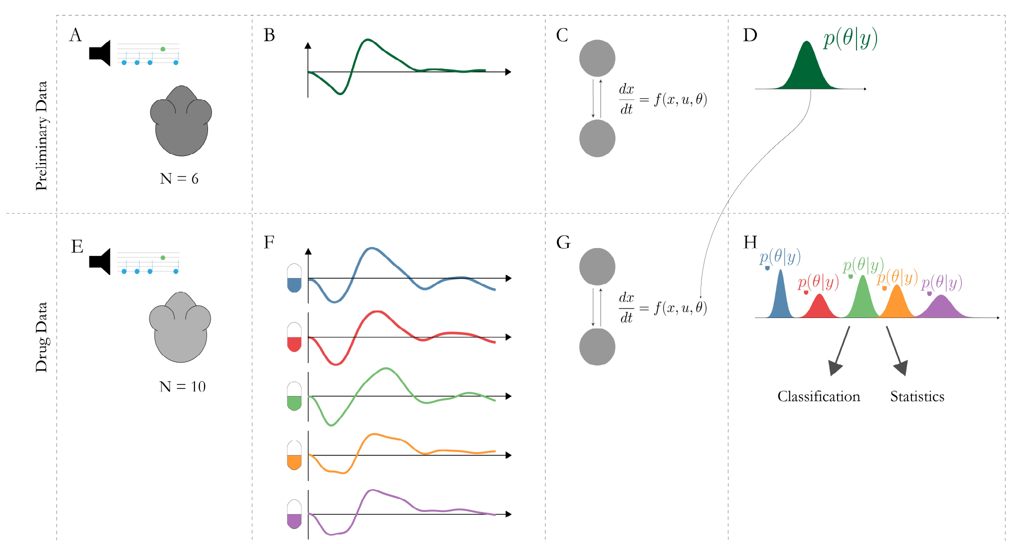

Figure 43| Study Design. A) Preliminary data of N=6 rats. Auditory oddball MMN paradigm with 10% and 20% deviant probability. B) Epidural recordings bilaterally from A1 and PAF under no

pharmacological manipulation. C) Dynamic Causal Modeling of the preliminary dataset. D) Inferred posterior distributions of the preliminary dataset. E) Drug data of N=10 (different) rats. Same auditory MMN paradigm as in (A). F) EEG recordings under 5-fold muscarinic manipulation. G) Dynamic Causal Modeling of drug dataset, with priors informed from the preliminary modeling. H) Inferred parameters for all pharmacological conditions. MAPs of posterior distributions used for classification and statistical testing for interactions with the drug.

The neural model of the DCM is given by a set of delay differential equations (DDEs) that describe the dynamics of neural activity in cortical column at synaptic level, i.e. *states*, and give rise to the particular family of DCM for ERPs – here convolution based DCMs. Computing the expected postsynaptic potential of a neuronal population that will ultimately give rise to the measured signal through a linear transformation therefore requires the integration of a system of DDEs. Note that delays are explicitly parametrized and induce a dependency of a neuronal state with another state in the past, i.e.

$$\dot{x}_J(t) = f\big(x_i\big(t - \tau_{iJ}\big), \theta\big),$$
$$i = 1, \dots, n, \tau_{ii} = 0.$$

In the convolution based DCM formulation, these states x_i correspond to the postsynaptic voltage and transmembrane currents of a neuronal population. A recent publication also questions whether the default integration methods implemented in SPM12 are ideally suited for rigorous parameter estimation (Lemarechal, George et al. 2018). Our own analyses are in accordance with this report. We therefore implemented and tested an alternative integration scheme tailored to this particular integration problem (similar to (Lemarechal, George et al. 2018)). This scheme keeps the whole history of the system in memory and explicitly samples the states at the desired delayed time. The states are integrated at the sampling step size of 1 kHz (1 ms). Delayed states are then approximated by linearly interpolating between the two neighboring, evaluated time steps. The update is then performed according to a simple Euler step, i.e.

$$x_{n+1} = x_n + dt \cdot f(\tilde{x}(t - \tau), \theta, u),$$

where $\tilde{x}$ denotes the interpolated (delayed) state, dt the integration step size, u, θ inputs and parameters, and n the integration step. Simulation have shown that this method provided more plausible results in a number of better understood systems of DDEs

Estimating model parameters in a Bayesian framework rests on finding the posterior distribution of said parameters, as described by Bayes Rule

$$p(\theta|Y) = \frac{p(Y|\,\theta)p(\theta)}{p(Y)}.$$

Usually, this expression is not analytically tractable, and one has to obtain approximate inference by optimization algorithms. In the context of DCM, this is typically done by using a gradient aescent scheme on the negative free energy. In brief, by maximizing the negative free energy (or minimizing the free energy), one converges to a conditional distribution over parameters that, under a chosen distributional form, provides an optimal approximation to the posterior distribution (Friston, Mattout et al. 2007).

Gradient aescent schemes, as used here, are prone to converge to local optima for non-convex problems. A multistart procedure may find better solutions (in terms of the negative free energy) when the objective function landscape is multimodal (Penny and Sengupta 2016). Since in our study, parameter estimation was of primary interest, we ran the optimization under a multistart routine, with 100 starting values per model inversion. This is arguably still a limited number for a high dimensional parameter space. Nevertheless, the multistart procedure not only provided better results than standard VB and suggested the presence of local optima, but the distribution over possible solution also gave a hint at how well behaved the optimization problem is for our particular problem. **Figure 44** illustrates the increase in negative free energy that could be obtained by the multistart approach, demonstrating its usefulness.

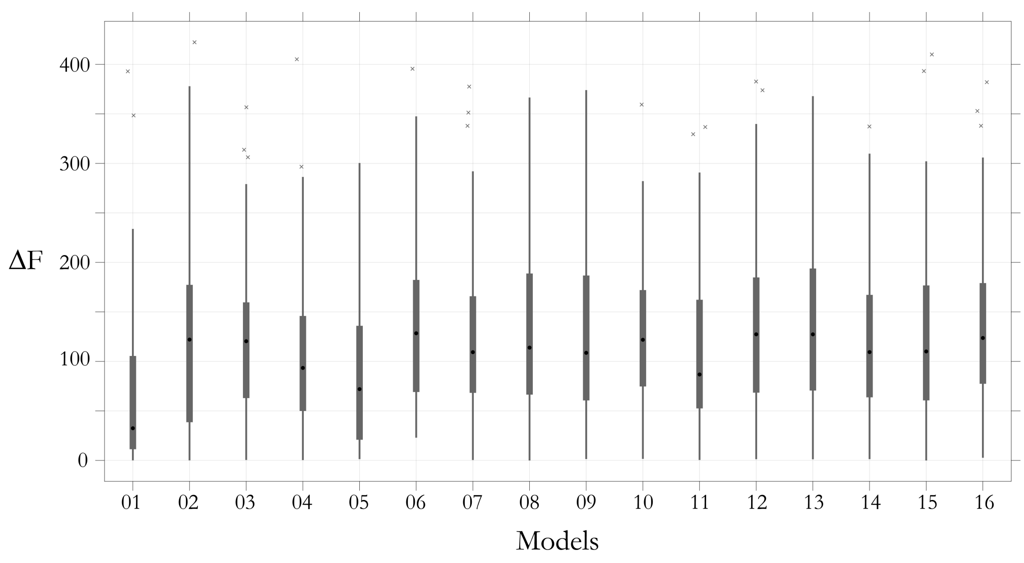

Figure 44| Boxplot of the effect of the Multistart Approach. Difference in (negative) Free Energy between 'best' and default starting value of the optimization. 'Best' refers to the starting value resulting in the highest negative free energy. Results are pooled over rats, hemispheres and drug levels.

5.8.3 Classification

To assess classification accuracy, we used a nested, leave one subject (animal) out (LOOCV) cross-validation approach, where, in an outer loop, we left out the set of parameters for one animal completely, i.e. all pharmacological conditions and both hemispheres. In an inner loop, we then performed another LOOCV cross-validation run, where we trained the model on the pharmacological conditions for n-2 animals, and predicted the drug labels of the left out animal (of the inner loop). During this inner loop, we optimized the hyperparameters of a support vector machine (SVM) classifier (Kernel Scale and Box Constraint) with respect to the cross-validation result. Put simply, the result from the inner loop was a SVM optimized for generalizability to new data. Finally, we then used this SVM on the left out animal of the outer loop and predicted its drug labels. Note that the left out animal of the outer loop was not used in the inner loop, so no leaking of information was possible between test data and any parameter optimized within the inner loop. Importantly, in the LOOCV approach we took, it is also not possible that the classifier learns animal specific parameters based on the other hemisphere, which could be interpreted as a leak of information from the parameters of the other hemisphere of the same animal, or any other pharmacological condition.

All classifications where implemented using the MATLAB function *fitcecoc.m* using SVM learners with linear kernels and stardardization. Coding of the multiclass problem was done in a fashion, where all binary combinations for class assignments are compared to each other (*binarycomplete*) (Allwein, Schapire et al. 2000).

For estimation of the confidence intervals, we computed a permutation test with 1000 permutations of the drug labels (within-animal and hemisphere), and re-estimated the full classification (including the optimization of the hyperparameters). The p-values then correspond to the proportion of drug-label configurations that would have allowed for a better classification than the actual one.

It is worth noting that, since we split the dataset into the two hemispheres and fitted independent DCMs to both hemispheres, we are left with a solution space of dimensionality animals x hemispheres x pharmacology x models. For practicality, we typically report the results pooled over a subset of the factors. For some of the analyses, this makes assumptions about the independence of data or parameters. Specifically, for the specification of priors, we average over hemispheres and models to avoid overfitting and bias. In the random effects model selection, we treat the two hemispheres of the same rat as independent animals, which seems reasonable given we do not make strong claims about the results of the model selection (but average over models instead). Finally, in the statistical tests (including classification), we include the knowledge about the two hemispheres by means of the LOOCV classification and by the animal specific mean as a single random effect in both hemispheres.

5.9 ADDITIONAL ANALYSES

5.9.1 INTRODUCTION

In this chapter, we present additional analyses performed on the RATMPI dataset introduced in the previous chapter. It will be split into three parts. These analyses are not meant to enter the publication at this time.

First, all analysis presented in the previous chapter, including supplement, were according to an analysis plan specified a priori (Appendix A of this thesis). However, the whole dataset consisted of an additional condition, where recordings were acquired in a 20% deviant probability condition (MMN_0.2). We exactly repeated the analyses performed on the 10% deviant probability (MMN_0.1) condition on this second half of the data. For more information on details, we refer to the previous chapter. Importantly, for the DCM analyses, we also informed the specification of the priors from an inversion of the non-pharmacological, 20% deviant probability dataset.

Second, in the chapter on the multistart scheme, we have shown in simulations that the multistart was beneficial in both model and parameter recovery. However, for a finite number of starts, it does not guarantee to find the true global maximum. Additionally, in the presence of strong correlations between parameters, the multistart might actually be prone to find different local solutions for different datasets, which could lead to larger between subject variance in terms of the posterior estimates. This could result in parameters being more difficult to compare across subjects, a challenge for both statistical analyses of posterior means or BMAs and classification (the latter might be a bit less affected, if the estimates are still linearly separable). Hence, if all inversions would end up in the same (or a similar) local optimum, it might improve classification. This could in principle be imposed by an early-stopping criteria induced by starting at the prior mean as discussed in Chapter 4. We therefore repeated the statistical analyses of parameters based on the results obtained when starting from the default starting values. Indeed, we will show that between subject variance of BMA estimates decreases, which seems to be beneficial for the ANOVAs to detect effects, and the classifier to predict the pharmacological labels.

Third, a formal way of constraining parameter has already been outlined in the original analysis plan as an additional analysis. There, we proposed to invert all pharmacological conditions for a single animal simultaneously, splitting potential drug effects directly into

(TONE) -specific and -unspecific effects. This greatly reduces the numbers of parameters for a single animal (from up to 135[37] to 73 synaptic parameters for the two most complex models). We applied this procedure in the 10% deviant probability dataset. When classification is performed on those results (using the best starting values), distinguishability between pharmacological conditions improves wrt. the original pipeline, and performs at a similar level as the results in part two, where we considered the local minimum closest to the default value.

Finally, we then provide a discussion over all results obtained on the RATMPI dataset. We will discuss the consistency of results over deviant probabilities, starting values and inversion protocol. In the end, this empirical dataset suggests that the correlations between parameters ask for some level of constraints, when parameter comparisons or classifications are the main goal.

Because of the analogy in the analyses, we omit the *Methods* section for parts 1 and 2 and continue directly to the *Results*.

5.9.2 Results Part 1: 20% Deviant Probability (MMN_0.2)

CLASSICAL ANALYSIS OF EVOKED RESPONSE POTENTIALS (ERPS)

All preprocessing steps were equal to the 10% deviant condition. In particular, we again excluded datasets where aberrant recording was evident. This resulted in the same (same animals) datasets as for the MMN_0.1 dataset (N=9 rats for the left, N=7 rats for the right hemisphere). We then repeated the mixed effects ANOVA on the group level, considering TONE and PHARMA as fixed effects, and ANIMAL as a random effect, including the interaction between TONE and PHARMA. The results are illustrated in **Figure 45**. We found highly similar effects, including the additional dip in the ERPs for the two agonistic pharmacological settings (yellow and purple curve) around 25-50 ms post stimulus, and late positivity 150-200 ms for the two antagonistic pharmacological settings (see arrows in **Figure 45**). Both effects were more pronounced as main effects and much less prominent in the difference wave (MMN). However, there was a significant (FDR-corrected) TONE x PHARMA interaction around 50 ms and a prolonged effect around 100 ms.

---

[37] 135 amount from individual inversions of the five pharmacological levels, with 27 independent parameters each.

We would like to point out here that for the given input in DCM (i.e. a brief pulse of excitation early in the trial), it is likely that DCM has more trouble explaining the late components appropriately. Due to the overall self-inhibiting dynamics of the system, the predicted signal will decay over time. Also, the MMN is attributed to a very particular component (N1) of the mismatch signal and later components are rather associated to other cognitive processes (Garrido, Kilner et al. 2009). But this could at least partially explain, why for the classification results of the MMN_0.1 datasets, distinguishing antagonistic pharmacological conditions was less decisive.

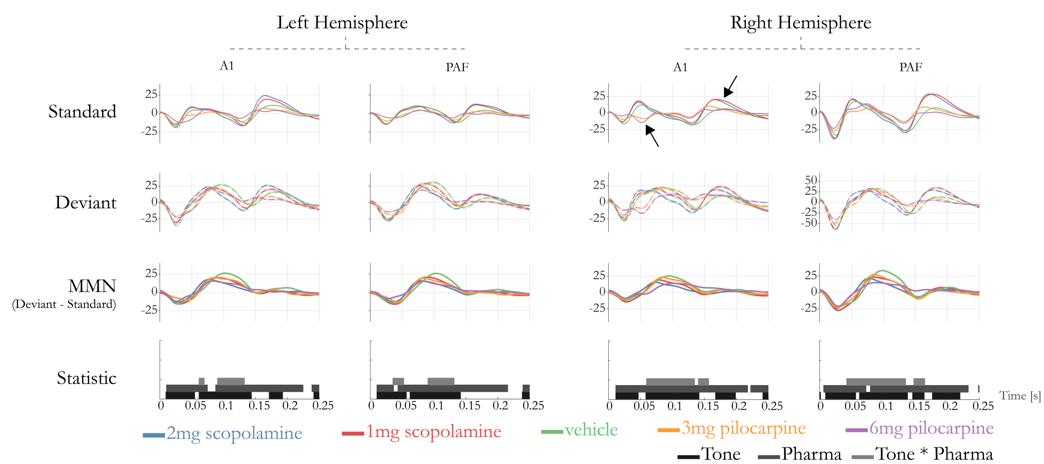

Figure 45 | 20% deviant probability condition. Grand Average Evoked Responses and results from the mixed effects ANOVA. Average (over animals and trials) Standard and Deviant tones are shown together with average difference waves for all drugs and both hemispheres. Statistical results are indicated by grey bars, whenever the effect (main or interaction effect) was significant at p < 0.05, FDR corrected. Arrows indicate peak mentioned in the paragraph.

DYNAMIC CAUSAL MODELING

We applied the same convolution based DCM framework (based on the canonical microcircuit (CMC) structure) and inverted all datasets (all pharmacological conditions, hemispheres and animals) using the model space of 16 models. Only animals were considered, where the recordings for all pharmacological conditions and both hemispheres were valid (N=14). We again used a multistart approach with 100 starting values and identified the 'best' solution as the one acquired with the starting value resulting in the highest negative free energy. The results from the random effects Bayesian model selection (RFX BMS) as well as the average model predictions (model 16) are shown in **Figure 46**.

The results are pooled over hemispheres, effectively treating the different hemispheres as independent animals. The tendency toward more complex models with increasing agonistic effect of the drugs is less clear compared to the MMN_0.1 setting. However, RFX BMS was only decisive in the 3mg Pilocarpine setting. Qualitatively, well performing models were models 12, 13, 15 and 16, hence models with high complexity (at least three connections modulated by TONE). Interestingly, the only model with three modulated connections which resulted in little protected exceedance probability, was model 14, which does not allow for modulation of the backward connection (from PAF to A1). This is in slight contrast to the MMN_0.1 results, where modulation of the forward connection seemed more crucial. **Figure 46**B provides an illustration of the average model prediction (model 16) vs. the average data. Overall, the average prediction (of model 16) fits the average data very well, which of course does not necessarily imply that the same holds equally for single animal predictions.

Figure 46| 20% deviant probability condition, best starting values. A) Bayesian Model Selection (BMS). Protected Exceedance probabilities reported for all sixteen models and drugs. B) Average (over animals) data (colored line) and prediction (black solid line) for model 16. Shaded area depicts standard deviation of the data (over animals and hemispheres), dotted lines depict standard deviation of prediction (over animal and hemispheres).

PARAMETER ESTIMATION AND STATISTICS

As in the 10% probability data, there was no clear winning model across all pharmacological conditions. Hence, we performed BMA on the single animal level, marginalizing out uncertainty about the model. The resulting BMA estimates were then used in two subsequent ANOVAs. One ANOVA tested for a general effect, the second ANOVA tested for a linear effect of PHARMA. For more information, we refer to the previous chapter on the MMN_0.1 dataset (*Methods* and *Supplementary Material*).

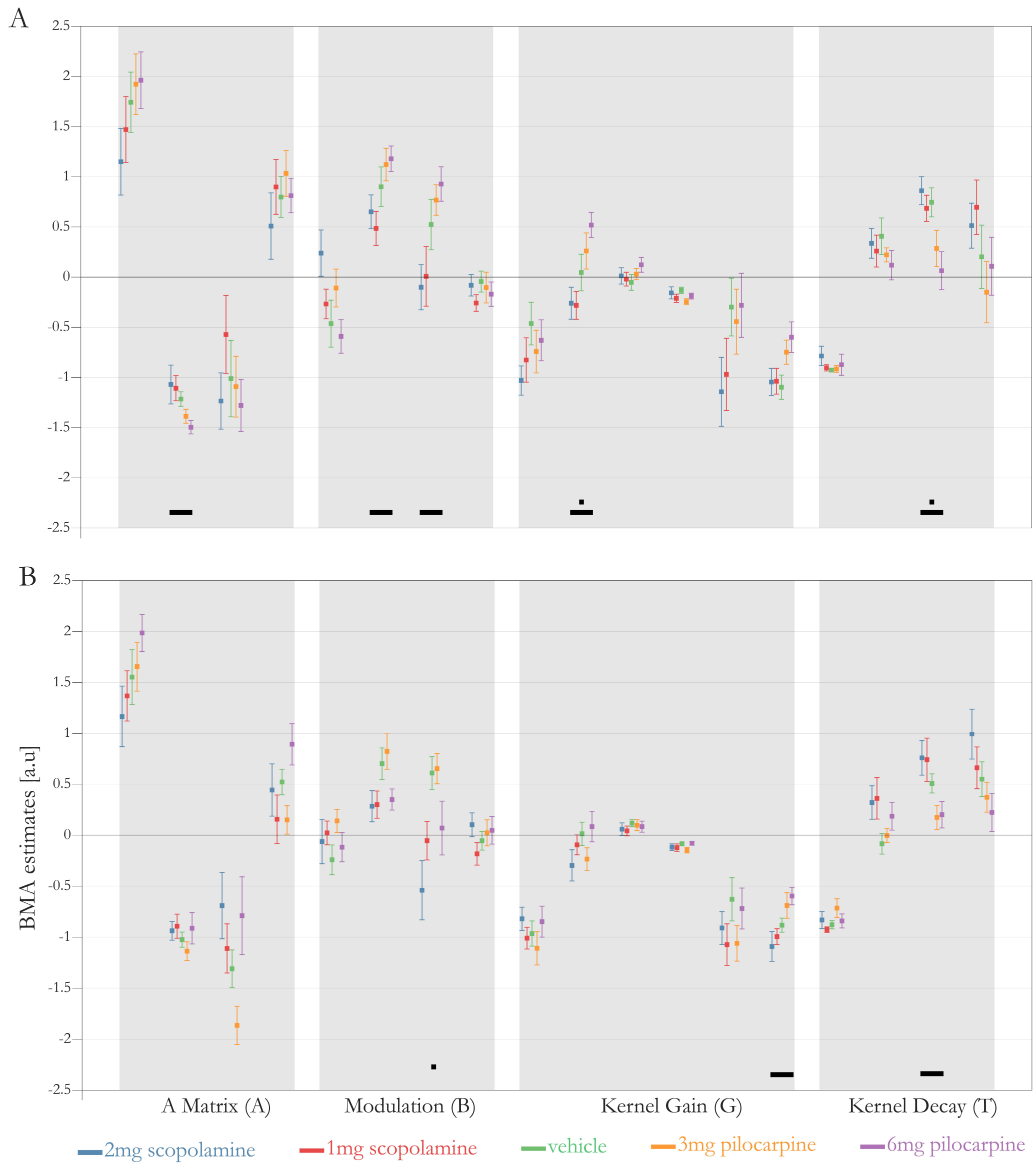

Figure 47| A) 10% deviant probability, best starting values. B) 20% best probability, best starting values. BMA estimates for all animals and hemispheres (N=14). BMAs are computed on the first level and pooled over both hemispheres. Errorbars depict SEM. Mixed effects (MFX) ANOVA on the BMA parameters are displayed. We considered two MFX models. First, a model with fixed factor DRUG (5 levels) and random effect ANIMAL. Black squares indicate significant results at $p<0.05$ (Bonferroni corrected). Second, a MFX ANOVA with a linear, fixed effect of DRUG and random effect of ANIMAL. Black horizontal bars indicate significant results at $p < 0.05$ (Bonferroni corrected).

On average (across drug conditions), the parameters all show the same trends when compared to the MMN_0.1 setting (**Figure 47**). We show the parameters from the last chapter again for comparison. In particular, we observe a positive forward modulation across the two datasets. Qualitatively, the clearest linear trends over pharmacological conditions can be seen in the 6th kernel gain parameter, the 3rd and 4th kernel decay and the first forward connection from A1 to PAF (consistent with the MMN_0.1 dataset), but only the first two show Bonferroni corrected significance over the 18 tests ($p<0.05$). Additionally, the backward modulation shows significant overall difference across drugs, which is in line with the importance of backward connections predicted by the RFX BMS.

CLASSIFICATION

We then proceeded to use the 18 parameters related to synaptic functioning (all parameters displayed in **Figure 47**B) as features for classification. We focused again on the same five comparisons: 1) the strongest antagonist (2mg Scopolamine) vs. the strongest agonist (6mg Pilocarpine), 2) 2mg Scopolamine vs. Vehicle, 3) Vehicle vs 6mg Pilocarpine, 4) the two antagonists vs. the two agonists and 5) a multiclass classification involving all four pharmacological interventions and Vehicle. As discussed in the main discussion, this classification is based on very few samples. Hence, we again used permutations of the labels to benchmark the performance of the classifiers, and to compute the statistical significance, i.e. the probability of a particular accuracy under permutations of the labels. The classification results are summarized in **Figure 48**. As one might expect from the fact that we found only two (as opposed to five (MMN_0.1)) significant linear effects in the ANOVA results, only antagonists and agonists could be significantly distinguished (BA=0.643, p=0.047), which however would not survive Bonferroni correction. Interestingly, distinguishing between 2mg Scopolamine and 6mg Pilocarpine only performed at chance level, but showed the highest distinguishability in the MMN_0.1 dataset.

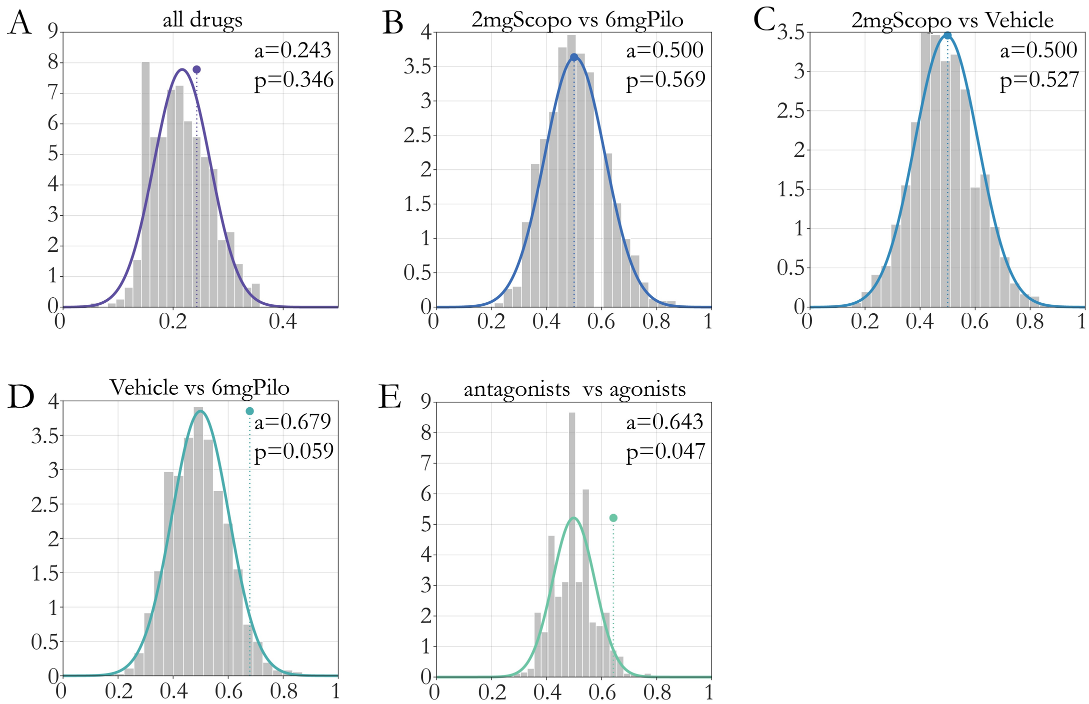

Figure 48| Permutation statistics for multiclass (A) and binary (B - E) classifications on the BMA results. Grey bars depict crossvalidation (CV) accuracies of permuted labels, solid line a gaussian fit on the histogram. Dotted line depicts CV accuracy for the true labels. Numbers refer to the CV accuracy of the true labels (a) and the percentage of a permutations leading to a higher accuracy (p). classifications.

5.9.3 Results Part 2: Parameter Statistics and Classification revisited

As mentioned in the introduction to this supplementary chapter, we have motivated the possibility that the default starting values might exert artificial constraints over the parameters. The reason would lie in the update scheme of the gradient ascent based optimization. If the gradients at the prior mean point in similar directions (across subjects), one might find a similar local posterior neighborhood, where parameters are more comparable. In other words, in a free energy landscape with many local minima, starting gradient ascent from the posterior mean results in the parameter estimates associated to the local minimum closest to the prior mean, i.e. which is reached by always moving up in the landscape from the prior mean. This can provide some "regularization" by selecting more similar local minima across subjects. It is important to note that this is a heuristic motivated by the fact that in such high-dimensional spaces, even the multistart is likely to fail finding the true global maximum for a non-exhaustive search of the starting value space (which is computationally infeasible). Also, as seen in the multistart chapter, it is likely that these parameters estimates will not correspond to the 'true parameters', which however could be an acceptable confound, if classifications are valued over assigning effects to single parameters. And finally, for these analyses we were using very dense sampling of the EEG signal resulting in a high number of datapoints, which could lead to bias in favor of complex models as illustrated in the chapter on hyperpriors.

The BMA results are illustrated in **Figure 49** for both, the 10% and 20% deviant probability dataset. We can see very consistent estimation of parameters across the two datasets. A similar observation was also made for the 'best' starting values, which is probably due to the priors being similar, despite being optimized based on the individual deviant probability non-pharmacological datasets. Secondly, variance across animals (and pharmacological conditions) is clearly reduced, indicated by the shorter errorbars depicting the SEM and the reduced variance across drugs (obviously, the number of datapoints is the same as for the 'best' starting value). Finally, both statistical analyses yield the same significant results in two and four parameters for the classical and linear ANOVA respectively. We could identify consistent (over MMN_0.1 and MMN_0.2) linear trends in the backward connection (terminating at the inhibitory population), the backward modulation and two kernel decays. We will draw the comparison to the counterpart for the 'best' starting values in the discussion. In conclusion, there appears to be qualitative evidence for a benefit through the artificial constraint of the starting value. Of course, p-values should not be the measure of superiority, but the qualitative reduction in variance and in turn increased comparability of parameters might be. While the findings shown here are based on a heuristic, they speak to the necessity of further constraining the parameter estimates (in particular of parameters of no interest) across participants and drug conditions in order to obtain more consistent results.

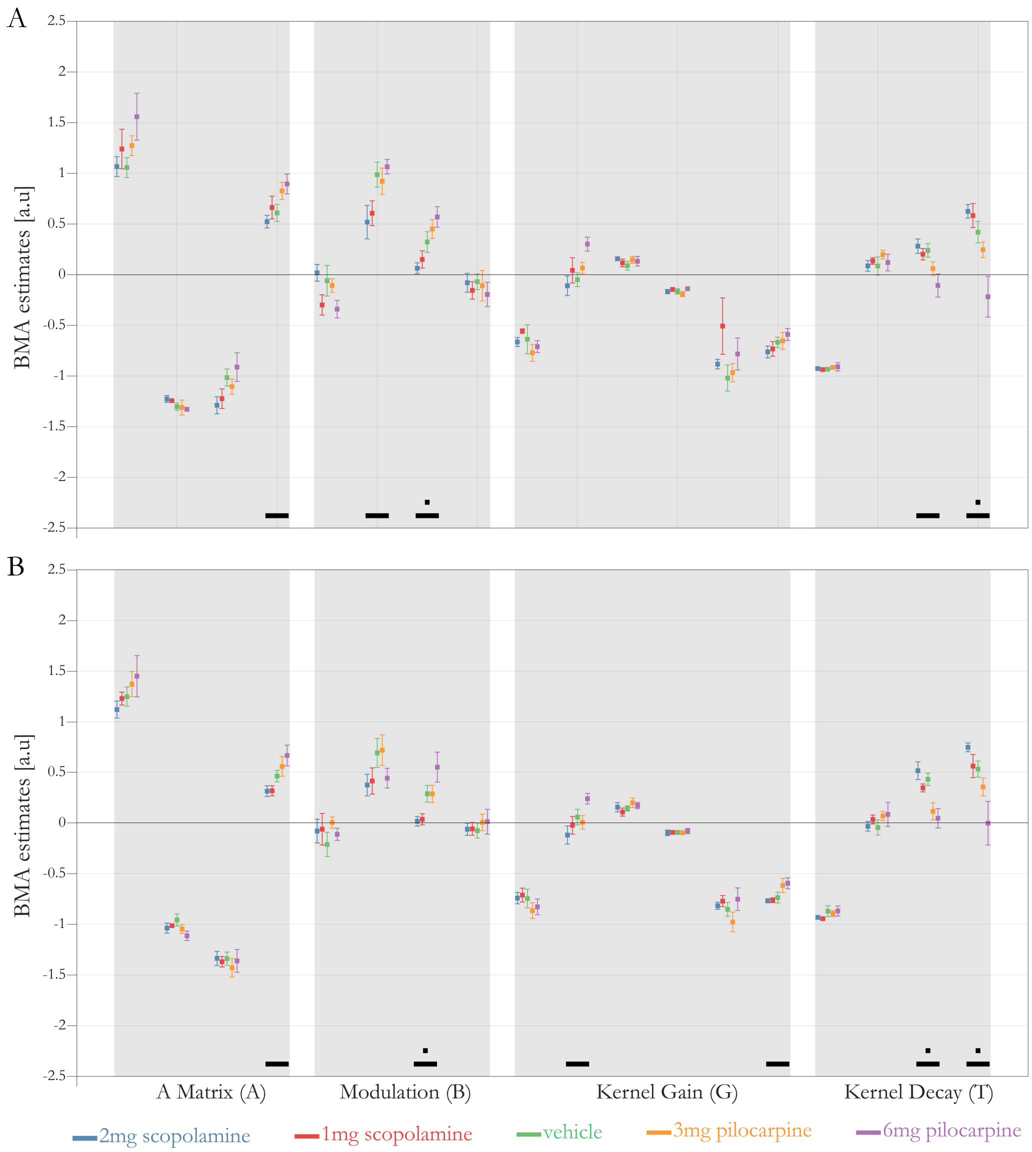

Figure 49| A) 10% deviant probability, default starting values. B) 20% deviant probability, default starting values. BMA estimates for all animals (n=7). BMAs are computed on the first level and pooled over both hemispheres. Errorbars depict SEM. Mixed effects (MFX) ANOVA on the BMA parameters are displayed. We considered two MFX models. First, a model with fixed factor DRUG (5 levels) and random effect ANIMAL. Black squares indicate significant results at $p < 0.05$ (Bonferroni corrected). Second, a MFX ANOVA with a linear, fixed effect of DRUG and random effect of ANIMAL. Black horizontal bars indicate significant results at $p < 0.05$ (Bonferroni corrected).

We then used these BMA estimates for the same LOOCV classification procedure. Balanced accuracies (BA) and p-Values based on the permutation test are provided in **Table 14**. Interestingly, the ability to classify increased clearly, particularly for the MMN_0.2 dataset. In line with the previously found results for the best starting values, we failed to distinguish antagonist (2mg Scopolamine) and Vehicle, but performed very well in predicting drug effects (agonist vs. antagonists).

Deviant probability	Comparison	N (per drug)	Balanced Accuracy (BA)	p-Values (uncorrected)
10%	**all**	**14**	**0.371**	**0.001**
	2mg Scopolamine vs 6mg Pilocarpine	**14**	**1**	**<0.001**
	2mg Scopolamine vs Vehicle	14	0.679	0.065
	Vehicle vs 6mg Pilocarpine	**14**	**0.857**	**0.002**
	Antagonist vs Agonist	**28**	**0.946**	**<0.001**
20%	**all**	**14**	**0.414**	**<0.001**
	2mg Scopolamine vs 6mg Pilocarpine	**14**	**0.786**	**0.006**
	2mg Scopolamine vs Vehicle	14	0.75	0.02
	Vehicle vs 6mg Pilocarpine	14	0.75	0.02
	Antagonist vs Agonist	**28**	**0.804**	**<0.001**

Table 14| Summary of classification results for the default starting values. All p-Values are based on a permutation of the drug label, and correspond to the ratio of permutation resulting in higher BA than the true label assignment. Bold text refers to significant results under Bonferroni correction (5 tests) at an alpha level of 5%.

5.9.4 Results Part 3: Constraining Parameters across Pharmacological Conditions

In the original analysis plan, we outlined a strategy to estimate all pharmacological conditions for a single animal (and hemisphere) simultaneously. The central idea was to distinguish between TONE-independent effects, i.e. connectivity changes under the pharmacological interventions affecting both, standard and deviant equally, and TONE-dependent effects, which modulate the connectivity only in the DEVIANT condition. Both effects can additionally be drug dependent or independent. In order to reduce the model space, TONE-independent effects were always expected to affect the full connectivity structure, while TONE-dependent effects could affect single or multiple connections, according to the model space defined in the original analysis (16 models). This specifies a 2x2 factorial design naturally over four model families, where each family is comprised of 16 submodels (see **Figure 50**).

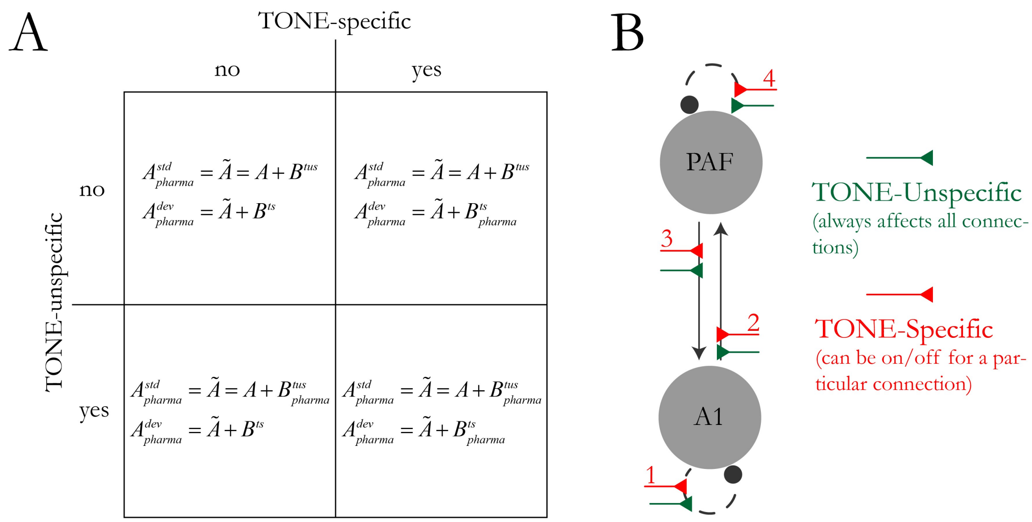

Figure 50| A) Factorial Model space, of whether (yes/no) PHARMA alters TONE-specific or TONE-unspecific effects. Definition of A-Matrix for each pharmacological condition (pharma) and standard (std) / deviant (dev). TONE-Specific (ts) affect only the deviant condition, TONE-Unspecific (tus) affect standard and deviant condition. The subscript PHARMA specifies a pharma-specific modulatory matrix (as opposed to a single matrix that is constant across drugs). B) Visual description of the TONE-specific and –unspecific effects. All TONE-specific modulations refer to one of the 16 models how connections can be affected by DEVIANT defined in the original analysis. TONE-Unspecific modulations always affect all connections.

As an example, we consider the case of submodel 8, i.e. TONE-specific modulation of both intrinsic connections and family 3 (i.e. pharma-dependent TONE-unspecific and pharma-independent TONE-specific effects). For a single animal, where all five pharmacological conditions are concatenated, this would result in the following design specification:

$$A_i = A + \sum_{j=1}^{6} X_{i,j} B\{j\}$$

with

$$X = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$B\{1\} = \begin{pmatrix} b_{A1} & 0 \\ 0 & b_{PAF} \end{pmatrix}$$

$$B\{2\} = \begin{pmatrix} b^{2mgScopo}{}_{11} & b^{2mgScopo}{}_{12} \\ b^{2mgScopo}{}_{21} & b^{2mgScopo}{}_{22} \end{pmatrix}$$

$$B\{3\} = \begin{pmatrix} b^{1mgScopo}{}_{11} & b^{1mgScopo}{}_{12} \\ b^{1mgScopo}{}_{21} & b^{1mgScopo}{}_{22} \end{pmatrix}$$

$$B\{4\} = \begin{pmatrix} b^{3mgPilo}{}_{11} & b^{3mgPilo}{}_{12} \\ b^{3mgPilo}{}_{21} & b^{3mgPilo}{}_{22} \end{pmatrix}$$

$$B\{5\} = \begin{pmatrix} b^{6mgPilo}{}_{11} & b^{6mgPilo}{}_{12} \\ b^{6mgPilo}{}_{21} & b^{6mgPilo}{}_{22} \end{pmatrix}$$

or in another notation

$$\begin{pmatrix} A_{2mgScopo}^{std} \\ A_{2mgScopo}^{dev} \\ A_{1mgScopo}^{std} \\ A_{1mgScopo}^{dev} \\ A_{Vehicle}^{std} \\ A_{Vehicle}^{dev} \\ A_{3mgPilo}^{std} \\ A_{3mgPilo}^{dev} \\ A_{6mgPilo}^{std} \\ A_{6mgPilo}^{dev} \end{pmatrix} = \begin{pmatrix} A + B\{2\} \\ A + B\{1\} + B\{2\} \\ A + B\{3\} \\ A + B\{1\} + B\{3\} \\ A \\ A + B\{1\} \\ A + B\{4\} \\ A + B\{1\} + B\{4\} \\ A + B\{5\} \\ A + B\{1\} + B\{5\} \end{pmatrix}.$$

Put simply, this estimates a baseline connectivity matrix for all pharmacological and TONE conditions (A), models changes in baseline connectivity induced by the drugs ($B\{2\} - B\{5\}$) and a fixed (across drugs) change induced by deviant ($B\{1\}$) as modulations. See that in this formulation, the standard condition in the Vehicle setting acts as the baseline connectivity. Therefore, kernel decay parameters (4), most of the kernel gains (5), delays (2), driving input parameters (2) and parameters of the population firing (2) and population gain (as part of the leadfield (2)) are kept fixed across all pharmacological conditions. Numbers in brackets correspond to the number of parameters in the non-reduced models. Note that we allowed for pharma-wise changes in the source gain (leadfield) to account for changes merely related to conductivity of the electrodes between different sessions.

This framework of distinguishing between TONE-specific and TONE-unspecific effects also allowed us to cast questions of drug effects in terms of Bayesian model/family comparison. The four cells in the factorial design (**Figure 50**A) code for the following families:

- No drug effect on average connectivity; no drug effect on TONE-specific connectivity (top left)
- No drug effect on average connectivity; drug effect on TONE-specific connectivity (top right)
- Drug effect on average connectivity; no drug effect on TONE-specific connectivity (bottom left)
- Drug effect on average connectivity; Drug effect on TONE-specific connectivity (bottom right)

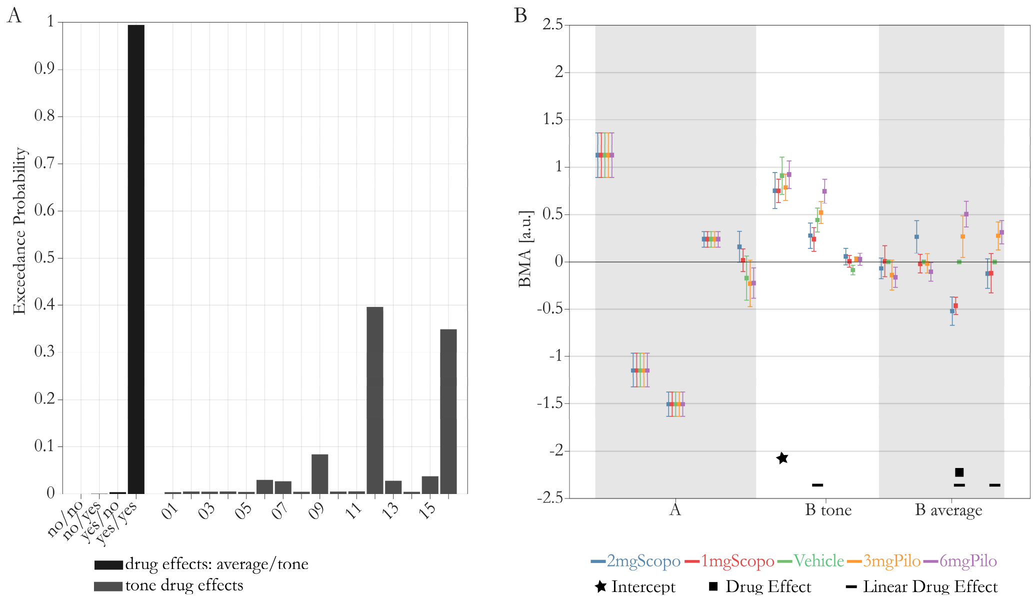

Figure 51| A) Bayesian Family Comparison results; with and without drug effect on TONE-specific (tone) / -unspecific (average) modulation (black bars) and for the different submodels, marginalized over the factorial structure of drug effects (grey bars). B) BMA connectivity estimates over animals and hemispheres (hemispheres are treated as independent animals, n=14). BMAs are computed over all models on the subject-level, and pooled over both hemispheres. Errorbars depict SEM. Baseline Connectivity (A) shown for completeness. Significant mixed effects (MFX) ANOVA estimates on the BMA parameters are displayed as black symbols. Significant Intercepts only considered, when significant in both ANOVAs. All results are Bonferroni corrected for the B parameters (eight tests).

We again ran the a full multistart inversion, using 100 starting values from the prior and inverted all 64 models; 4 families with 16 submodels, each. We used the results from the starting values resulting in the highest negative free energy, computed the BMS and used the BMA estimates in subsequent statistical tests. The results are summarized in **Figure 51**. The results confirm once more the previously found effects of the drugs. The family comparison indicates drug effects on both, average connectivity (TONE-unspecific) and TONE-specific modulations. When marginalizing over the drug effects, exceedance probabilities for the submodels indicate TONE-specific modulation of extrinsic connections (models 09, 12 and 16). This is confirmed in the ANOVA over BMA parameters and can be understood in the following way. The TONE-specific forward modulation (see second BMA estimates in *B tone* in **Figure 51**B) shows an average, yet drug independent effect. The backward modulation shows a linear effect over drugs, i.e. TONE-specific backward modulation increases from antagonistic to agonistic effects of the drug. Additionally, linear TONE-unspecific (average connectivity) changes can be observed on the backward connection and the intrinsic connection of PAF. It is important to note that latter refers to a change in kernel function (kernel gain of the self-inhibition of the superficial pyramidal cell in PAF), which was also observed in the original analyses.

Of course, this modeling approach reduces the flexibility of the model to fit the data, but the compromise it not too dramatic. See **Figure 52** for an illustration of the model fits and the explained variance in the two conditions. One notable difference is that the additional dip in the first peak of the ERPs in the agonistic conditions is less accurately fitted than in the more flexible, standard analysis. This dip was discussed in the original analysis of the MMN_0.1 dataset. We investigated whether this affected the ability to classify by running the classification procedure on only the eight modulatory parameters (see **Figure 51**B), which are the only drug dependent neuronal parameters in this inversion. Please note that this feature selection does not allow for direct comparison to the results in previous parts. There, other parameters could change with drug as well, and were therefore included in the classification. Again, we used permutation testing to obtain p-values for the Balanced Accuracies. The results are shown in **Table 15**.

Deviant probability	**Comparison**	**N (per drug)**	**Balanced Accuracy (BA)**	**p-Values (uncorrected)**
10%	**all**	**14**	**0.357**	**<0.001**
	2mg Scopolamine vs 6mg Pilocarpine	**14**	**0.929**	**<0.001**
	2mg Scopolamine vs Vehicle	**14**	**0.821**	**0.001**
	Vehicle vs 6mg Pilocarpine	14	0.75	0.013
	Antagonis vs Agonist	**28**	**0.788**	**<0.001**

Table 15| Summary of classification results for new analysis approach, where all pharmacological conditions were inverted simultaneously. All p-Values are based on a permutation of the drug label, and correspond to the ratio of permutation resulting in higher BA than the true label assignment. Bold text refers to significant results under Bonferroni correction (5 tests) at an alpha level of 5%.

The classification results indicate that for this new modeling approach, the parameter estimates are most discriminative for the pharmacological conditions. In four out of five classifications, there was only one or no other permutation of drug labels resulting in a higher BA results. The one classification not significant under rigorous Bonferroni correction (Vehicle vs 6mg Pilocarpine) just failed to reach significance. However, it is worth pointing out that the different classifications are not independent tests, hence Bonferroni correction is very conservative. There are multiple reasons as to why the classification might be superior in this new design. We will put these results in perspective with all other presented results of the RATMPI study in the final, overarching discussion.

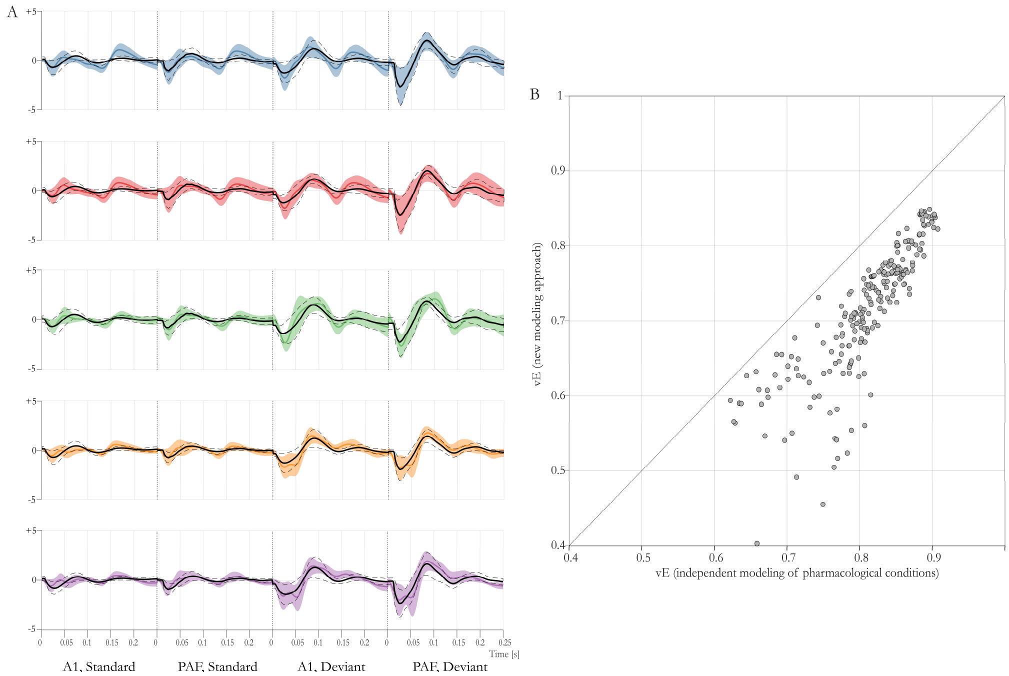

Figure 52| Average (over animals and hemispheres) data (colored line) and prediction (black solid line) for model 16. Shaded area depicts standard deviation of the data, dotted lines depict standard deviation of prediction. B) Comparison between explained variance (goodness of fit) over all models, hemispheres and animals.

5.9.5 Discussion

In this chapter, we have presented the analyses of two datasets, where electrophysiological measures were obtained in rodents under a graded muscarinic intervention. The two datasets comprised of a 10% and 20% deviant probability condition of an auditory mismatch negativity paradigm. We have presented three DCM modeling strategies: (i) We used a multistart approach to invert the datasets of all rodents, hemispheres and pharmacological conditions independently; (ii) We compared the statistical and classification results on parameter estimates obtained with the starting values resulting in the highest negative free energy, to the results obtained for the default starting values of the VB-optimization (i.e. starting from the prior mean); (iii) We used a different modeling strategy inverting all pharmacological conditions for a single animal and hemisphere simultaneously, which in turn constrained multiple parameters across drug condition and cast questions about pharmacological effects as a question of model selection.

We now discuss (in)-consistencies in results across these different approaches and deviant probability from a technical perspective. This discussion should in no way invalidate the original results. In contrary, it illustrates that although quite different approaches were taken, many of the findings were similar. It is not surprising that not all significance statements match if quite different analysis strategies are compared. For the interpretation in terms of physiology, cognitive functions and translational implications, we refer to the discussion of the original analysis.

The statistical results of the parameter estimates are summarized in **Table 16** and **Table 18**. Overall, the effects are fairly consistent, independent of the specific routine. Out of 18 possible parameters, only eight showed effects (between two and six depending on the routine and dataset. Drugs exhibit both, main effects and interaction effects with TONE, which is in line with the family comparison of the alternative modeling approach. The most likely reason for the small differences in the exact subgroup of parameters that show significant effects are correlations of the posterior parameter estimates. As we have already pointed out in the discussion to the multistart chapter, drawing strong conclusions from statistics on single parameter estimates should be done with caution.

			significant linear drug effects							
modeling	dataset	starting values	Af(2)	Ab(2)	B(2)	B(3)	G(2)	G(6)	T(3)	T(4)
classical	MMN_0.1	default								
		best								
	MMN_0.2	default								
		best								
alternative	MMN_0.1	best								

Table 16| Summary over significant linear drug effects over inversion strategies and datasets. Green bars indicate significant results for a particular parameter (Bonferroni corrected at an alpha level of 5%). Light grey areas depict that an effect has a slightly different interpretation for the given inversion approach. Black bars indicate parameters that are not part of the model.

			significant classification results				
modeling	dataset	starting values	M-- vs M++	M-- vs V	V vs M++	M-/-- vs M+/++	all
classical	MMN_0.1	default					
		best					
	MMN_0.2	default					
		best					
alternative	MMN_0.1	best					

Table 17| Summary over significant classification results. Green bars indicate significant results for a particular parameter (Bonferroni corrected at an alpha level of 5%). M--: 2mg Scopolamine; M-: 1mg Scopolamine; V: Vehicle; M+: 3mg Pilocarpine; M++: 6mg Pilocarpine. Please note that within each model and approach the classification was based on all parameters related to synaptic action that could vary with drug. This differs between the classical and alternative modeling approaches.

There is a simple explanation for the superior classification performance when constraining parameter estimates across drug conditions; Classification then simply relied on fewer features (8 vs 18). This in turn could protect against overfitting, which is a known problem especially for such few number of samples per class. Sometimes in classification procedures, people optimize over an alpha level of a classical result in order to identify the features needed. Here, we were very liberal and included all parameters associated to synaptic action. By doing so, we might have been creating a more challenging setup when it came to classification.

Another option for the large between animal variance could also lie in the way the BMA is computed. Here, we performed BMAs on the subject level throughout the analyses, and hence subject level posterior model probabilities are used. This might cause larger variation in comparison to the counterpart, where the group-level posterior model probabilities weigh the parameters. However, if BMAs are used for subsequent classification one needs to protect the estimated values against potential leakage of information and thus would need to re-compute BMAs for every crossfold. This would change the features in every crossfold, which is not desired. We have thus refrained from further pursuing this approach.

Along a different line of thought, the default starting values also showed some benefits in constraining variance across posterior estimates. This confirms our intuition gained in Chapter 4. While it seems an artificial method at the moment and does not quite adhere to the 'optimiality' principles, it is related to early-stopping which has been drawing attention in different optimization settings.

At this point, knowing the limitations and problems outline so far, one could ask the question: Which parameters are truly affected by drug? Which pharmacological settings are truly distinguishable? To put this into context, one needs to keep in mind the previous results from the multistart and hyperparameter chapter. From there we know that parameters are correlated and that even with the multistart, we are unlikely to find the 'true' parameters. Additionally, the results there would indicate an identifiability issue, maybe not in a mathematical but at least in a naïve meaning of the word, where it simply means that different sets of parameters lead to – in a probabilistic sense – (practically) identical accuracies. And finally, the dataset is not large enough to generalize. In conclusion, it would be irresponsible to claim to have found 'truth'.

What we do dare to say, however is that DCM allowed us to substantially reduce the amount of features (i.e datapoints) in an informed manner, from 1000 to 18 in the original approach.

These 18 parameters (or even only 8/18) contained information about drug effects that survived both rigorous permutation testing and Bonferroni correction. The confounds of the optimization and the model (e.g. correlations) could simply mean that one needs to take a more pragmatic view at the moment, accept that the utility lies not in single parameter estimates but rather in finding a manifold/subspace in parameter space that allows for classification. Depending on where that manifold is probed, the estimates of individual parameters can differ substantially. The way forward does seem to lie in constraining parameters, always knowing that the inference drawn then depends on the constraints cast on the network.

6| EFFECTIVE CONNECTIVITY DURING A SELF-CALIBRATING VISUO-SPATIAL WORKING MEMORY PARADIGM

6.1 DISCLAIMER

This chapter contains a manuscript that is currently in preparation for publishing. The data for this chapter were acquired by Jolanda Malamud, Sara Tomiello, Katharina Wellstein and Gabrielle Zbaeren. J. Malamud used the data to do an initial analysis on the paradigm as part of her Master Thesis under the supervision of Sandra Iglesias and me. The strategy for the model based analysis has been greatly extended since, and will result in a shared first-authorship. The roles are as follows:

- **Katharina Wellstein**: Data acquisition
- **Gabrielle Zbaeren**: Data acquisition and analysis of an MMN task as part of her master thesis.
- **Sara Tomiello**: Data acquisition and analysis of a reward learning task as part of her dissertation.
- **Jolanda Malamud**: Data acquisition and initial analysis of the working memory task as part of her master thesis
- **Dario Schöbi**: Derivation and programming of the task, supervision of J. Malamud, re-analysis of the WM data with a different strategy for the DCM analysis (see Changelog in Appendix B), writing of the paper.

These final analyses were done under the supervision of Klaas Enno Stephan, Jakob Heinzle and Sandra Iglesias. All analyses were performed according to an analysis plan which is provided in Appendix B.

MANUSCRIPT: EFFECTIVE CONNECTIVITY DURING A SELF-CALIBRATING VISUO-SPATIAL WORKING MEMORY PARADIGM

Dario Schöbi[1*], Jolanda Malamud[2*], Sara Tomiello[1], Stefan Frässle[1], Jakob Heinzle[1], Klaas Enno Stephan[1,3,4], Sandra Iglesias[1]

[1] Translational Neuromodeling Unit, Inst. for Biomedical Engineering, Univ. of Zurich & Swiss Institute of Technology (ETH Zurich), Wilfriedstrasse 6, 8002, Zurich, Switzerland.

[2] Max Planck University College London Centre for Computational Psychiatry and Ageing Research, 10-12 Russell Square, London, WC1B 5EH, UK.

[3] Wellcome Trust Centre for Neuroimaging, University College London, 12 Queen Square, London, WC1N 3AR, UK.

[4] Max Planck Institute for Metabolism Research, Gleueler Strasse 50, 50931 Cologne, Germany.

* These authors contributed equally to this work

Corresponding authors:

Dario Schöbi
University of Zurich & ETH Zurich
Translational Neuromodeling Unit (TNU)
Institute for Biomedical Engineering
Wilfriedstrasse 6
8032 Zurich, Switzerland
Phone: +41 44 634 91 12
E-mail: dschoebi@biomed.ee.ethz.ch

6.2 ABSTRACT

Working Memory (WM) has been proposed to provide a potential readout of dopaminergic neuromodulation, which could be crucial in the process of understanding the heterogeneity in treatment response in many psychiatric diseases. One reason why this approach has failed so far, could lie in performance confounding the results between patients and healthy controls, or within a clinical population. Here, we present a novel, spatial, delayed match-to-sample paradigm, in which difficulty is dynamically changed over the course of the experiment to keep performance similar across participants. In doing so we circumvent the need for prolonged pre-testing and can adapt to performance changes during the experiment. The aim of this study was threefold: (i) to investigate in healthy volunteers whether by keeping performance constant we would still find the hypothesized WM-related activations in frontal and parietal regions, (ii) to identify the regions involved in this specific WM implementation to inform the source localization and modeling in two sibling EEG studies, one measured in healthy volunteers with a pharmacological intervention and a second study with patients suffering from schizophrenia, (iii) to investigate the working memory induced effective connectivity in the fronto-parietal network when controlling for task difficulty.

Our fMRI results showed that the hypothesized fronto-parietal network involved during working memory was indeed not confounded by difficulty. Further, dynamic causal modeling indicated that working memory decreased the parietal self-inhibition and increased the connection strength from parietal to frontal regions.

6.3 INTRODUCTION

The first mention of the term 'working memory' (WM) goes back to the 1960s, where it has been described as a 'quick access memory' supporting daily tasks (Miller, Galanter et al. 1960). Since then, it has been refined in multiple ways and today WM terms a general system for the retention and manipulation of information over short periods of time (Baddeley 1992). As such, it is also understood that WM provides a system for an internal representation of the world. While the exact form is neither functionally nor cognitively fully understood, WM is evidently crucial in planning, believe building, goal directed behavior and learning (D'Esposito and Postle 2015).

Starting in the early 90s, a number of studies investigated the influence of dopamine (DA) on WM. Animal studies showed impaired WM when altering prefrontal (PFC) D1 receptor function (Sawaguchi and Goldman-Rakic 1991, Williams and Goldman-Rakic 1995, Murphy, Arnsten et al. 1996, Zahrt, Taylor et al. 1997). The results, however, were somewhat contradictory, as both D1 agonist and antagonist interventions resulted in reduced WM function. This alleged dichotomy was addressed by the proposition of an inverted U-shape between D1 receptor function and WM performance (Vijayraghavan, Wang et al. 2007, Cools and D'Esposito 2011). From a network perspective, a more recent study showed a strong correlation between mean cortical D1 density and the decoupling between the default mode (task-unspecific) network and the fronto-parietal (task-specific) network when participants engaged in a Sternberg-Item Recognition paradigm ($R^2 = 0.98$) (Roffman, Tanner et al. 2016).

In parallel, the so-called dopamine hypothesis of schizophrenia was formulated (Carlsson 1988, Davis, Kahn et al. 1991, Howes and Kapur 2009). Here, psychotic symptoms were discussed as a consequence of dopaminergic hypo- and hyperfunction in sub- and prefrontal cortex, motivated by the effectiveness of dopaminergic drugs in psychotic disorders (Seeman 1987). Although the average efficacy of dopaminergic interventions in schizophrenia is undebated, there is a vast heterogeneity in terms of single patient outcomes for a specific drug (Leucht, Komossa et al. 2009, Kapur, Phillips et al. 2012, Leucht, Cipriani et al. 2013).

Collectively, these findings motivate the need for a readout of DA function in patients to tackle this problem of single patient treatment predictions. This readout could be provided by WM (not the least because impaired WM seems omnipresent in psychiatric disorders

(Gold, Barch et al. 2018)). The central idea is to use a combination of non-invasive brain imaging and mathematical modeling of hidden neural mechanisms (i.e. DA function) that underlie neural activity and might give rise to the pathophysiology. Hence, the model, in this approach, can be understood as a 'measurement device' to differentiate patients along a dimension spanned by the model parameters (Stephan and Mathys 2014). The hope is that this dimension provides an etiological understanding of spectrum diseases, allowing for a much more fine-grained differentiation between patients and for individual treatment prediction, relapse risk, etc.

There have already been a number of proof of concept studies on this idea. A promising study by (Brodersen, Deserno et al. 2014) showed that combining functional magnetic resonance imaging (fMRI), mathematical modeling and machine learning enabled to predict scores on the positive and negative syndrome scale (PANSS) (Kay, Fiszbein et al. 1987) solely based on model parameters modelling the effective connectivity of a brain network involved in WM. In another study using electrophysiology and WM, (Moran, Symmonds et al. 2011) showed the connection between the dopamine precursor levodopa, the measured theta-band power, and NMDA function. Here, the model parameters modeling NMDA receptor function was predictive of performance. While these results are not of direct clinical relevance yet, they showed the potential of this approach to connect hidden physiological processes to symptoms or behavioral scores by means of a model.

Overall, many studies have investigated WM in relation to DA and/or schizophrenia e.g. (Goldman-Rakic 1994, Abi-Dargham, Mawlawi et al. 2002, Park and Gooding 2014). However, there is large variation in the results across task designs, subpopulations, and activation patterns in neuroimaging (Manoach 2003, Lee and Park 2005, Forbes, Carrick et al. 2009, Rottschy, Langner et al. 2012). This is a likely consequence of the fact that particular designs involve multiple cognitive processes that might not be associated to WM per se (e.g. perception, attention). A common, possible major confound, particularly regarding the involvement of prefrontal regions, lies in performance accuracy (Lencz, Bilder et al. 2003, Manoach 2003, Van Snellenberg, Torres et al. 2006). Avoiding this confound is crucial for the aforementioned approach, where we hypothesize that the WM-relevant brain-networks and dopaminergic modulation thereof, could be the dimension along which patients separate in terms of treatment outcome. A previous study has illustrated exactly this problem in an fMRI study with schizophrenic patients and healthy controls, by modeling the neural activity elicited during WM in a fronto-parietal network using dynamic causal modeling (DCM) (Deserno, Sterzer et al. 2012). Among other findings they showed a significant group x task x performance interaction in dorsolateral PFC (dlPFC), which is in line with previously reported dopaminergic hypofunction in dlPFC (Okubo, Suhara et al. 1997).

There are a number of studies which tackled the problem in different ways. For example, (Barch, Braver et al. 1997) tried to dissociate task difficulty and working memory by augmenting a continuous performance test (CPT) WM design with two experimental manipulations; Changing the duration of delay period, and changing the task-difficulty by making the stimuli more ambiguous. Their results support the hypothesis that (among other areas) DLPFC activation is changed under changes in WM demands (delay duration) but not task-difficulty (stimulus ambiguity). Another approach is titration of difficulty. Verhaeghen and collegues showed that age-related differences in WM vanished in an 2-back paradigm, if participants were titrated to a similar performance level in the 1-back variant beforehand (Verhaeghen, Geigerman et al. 2019).

In summary, in order to bring WM to a prospective patient setting where neuroimaging and computational modeling can be used, it is crucial to reduce the sources for between patient variation that could cloud the sensitivity of the model parameters for treatment prediction (Callicott, Ramsey et al. 1998). In this line of thought we present a novel, spatial, delayed match-to-sample WM paradigm, where the difficulty of the task can be parametrically manipulated over trials to titrate the accuracy of an individual to a desired level, similar to (Lencz, Bilder et al. 2003). There are three major motivations. First, in prospect of a clinical setting, tasks need to be kept short and feasible for patients, which forbids prolonged pre-testing. Additionally, based on the presented results, there is a need to keep the task difficulty similar across patient, reducing between patient variation solely due to performance accuracy. Second, the delayed match-to-sample nature of the task is preferred over n-back founded on substantial animal literature of DA effects on WM, e.g. (Sawaguchi and Goldman-Rakic 1991, Murphy, Arnsten et al. 1996). Also, (Miller, Price et al. 2009) couldn't find convergent findings between performance in an n-back and a digit-span-backward task, suggesting that n-back involves additional other cognitive processes (for example because of the processing speed demands) and hence might not be a clean readout for WM ability in a clinical setting. Third, this task has also been measured in two (not yet published) electrophysiological studies and serves for the definition of source priors. One is a double-blind, placebo controlled, electroencephalography (EEG) study in healthy participants, with agonistic and antagonistic cholinergic and dopaminergic pharmacological manipulations. The other is an EEG study with schizophrenic patients, who underwent a medication switch.

Hence, this study can be understood as a precursor study with four major goals: (i) Does the titration of difficulty work; (ii) can we identify a similar fronto-parietal involvement for this particular paradigm, where task difficulty is controlled, thus alleviating the confound of performance; (iii) Definition of source priors for the sibling studies; (iv) can we apply DCM to investigate how MEMORY[38] alters the connectivity in the network, putting the results in perspective with comparable studies, and further informing the modeling in the sibling studies.

6.4 METHODS

General Information

Forty-eight (48), healthy, right-handed male participants with a mean age of 23 (SD[39] = 2.8) participated in the study. Exclusion criteria included alcohol consumption 24 hours and analgesics within three days prior to the participation, current psychiatric disorders, medication or nicotine consumption and past or current neurological conditions or regular drug use. All participants gave written informed consent before the study. All experimental procedures were approved by the local ethics board.

Task Design

We probed WM using a visuo-spatial delayed match to sample paradigm and a control condition using the same visual stimulation without the working memory component. The task was programmed with Psychophysics Toolbox Version 3 (PTB-3, http://psychtoolbox.org) and each condition lasted for 82 trials. The sample phase (also referred to as cue or encoding phase) consisted of a pattern of six grey dots presented for 0.5 *s*, followed by a delay period of 4 (or 8 *s*) only showing a red fixation cross (**Figure**

[38] All caps notation refers to a factor of the analysis, i.e. a condition or factor in a statistical analysis.
[39] Standard Deviation

53A). After the delay period, the fixation cross vanished and the probe stimuli (i.e. the target or decoding phase) appeared in the left and right half of the screen. One of the probe stimuli had the exact same spatial arrangement as the sample stimulus, the other one was a slightly noisy version of the sample stimulus. The nature of what is called 'noise' will be explained below. Additionally, both probe stimuli were of different greyscale, one being darker, one lighter than the sample stimulus. The probe stimuli were presented for a 1.8 *s* or until a response was given (through a right hand button press).

In the MEMORY condition, participants were instructed to choose, which of the probe stimuli had the same spatial arrangement as the sample stimulus, irrespective of the greyscale (**Figure 53**B3). In the CONTROL condition, participants were to choose the probe stimulus that was darker, irrespective of the sample stimulus (note that in the control condition, neither of the two stimuli had the same spatial arrangement as the sample stimulus, to reduce the risk of the participants to unnecessarily remembering the sample stimulus, **Figure 53**B2).

The nature of the task allowed for a parametric change of the task-difficulty, such that the participants would overall perform at a pre-set accuracy level of 70 %, i.e. an average of seven correct responses for the ten preceding trials. We will also refer to this process as titration.

In the MEMORY condition, the difficulty level was adjusted by changing the amount of noise given to the alternative probe stimulus. In detail, all the points were placed on concentric circles with different (fixed) radii. 'Noise' then referred to a change in the angular coordinate in random direction, as depicted in **Figure 53**B3. Hence, a decrease in the change of the angular coordinate means that the spatial arrangement of the noisy probe stimulus resembled the one of the sample stimulus more closely, making the task more difficult. In the CONTROL condition, the greyscale difference between the two probe stimuli was changed to ensure the desired accuracy in response.

For both conditions, we interleaved 30 trials with a delay period of 8 seconds. The particular order of these 4 and 8 second trials was defined in an efficiency analysis before the start of the measurement. In brief, while all three phases of working memory might convey crucial information about the underlying neural mechanisms, our primary research question was devoted to the delay period. By construction of the experiment, there are non-negligible correlations between the delay period regressor and the en- and decoding regressors, making the design less sensitive to detect effects during the delay periods. The correlations come from the immediate temporal succession of the stimuli presentation, delay period and retrieval in combination with the temporal smoothing effect of the hemodynamic response. In order to overcome this problem and de-correlate the delay period from the de-coding (PROBE*)* regressor, only the first 4 seconds of all delay periods were modeled (see Supplementary Material). Unfortunately, we were not able to de-correlate encoding (SAMPLE) from DELAY. Implications of this will be discussed later in the paper.

Behavioral Data Analysis

Analysis of the behavioral data was limited to computing summary statistics in terms of average accuracies across subjects (over trials) to validate the titration procedure (**Figure 53**C). One participant hat to be excluded for not reaching an average of above 60% correct responses in the MEMORY condition, despite the titration. Hence, we couldn't be sure that the participant properly understood the task.

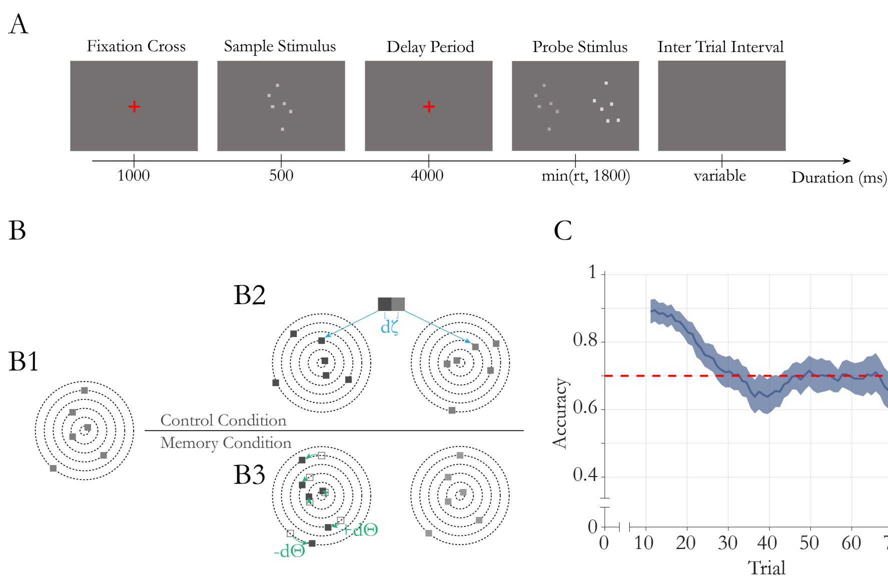

Figure 53| A) Design of a single trial. B) Design for the two conditions. B1) Sample stimulus. The six squares are placed on six concentric circles with increasing radius. To prevent overlap, squares are alternatingly placed above or below the center. B2) Probe stimulus in control condition. In both patterns, the squares are placed randomly, independent of the sample stimulus. The greyscale of the squares of the two patterns differed by dζ, also independent of the sample stimulus. B3) Probe stimulus in memory condition. For one of the patterns, the squares were moved by an angle dΘ with respect to the sample stimulus, with equal probability in either direction. The greyscale of both patterns was chosen randomly. C) Behavioral results over all participants. Solid line indicates average accuracy (over participants) during memory, shaded area depicts 95% confidence interval. Red line indicates the desired accuracy level of 70%.

Functional Data Acquisition & Preprocessing

Whole brain T1-weighted structural and a T2*-weighted echo-planar volumes (EPI) were acquired using a 3 T scanner (Philips Ingenia) and a 32-channel headcoil. For each condition of the paradigm, the parameters of the fMRI sequences were as follows: 304 volumes with 32 slices per volume, voxel size 2 x 2 x 3 mm, TR 2.5 s, TE 36 ms. Finally, we recorded cardiac and respiratory signals during the scan using a respiration belt and a finger plethysmograph. These physiological measures were later used to correct for potential physiological noise.

The fMRI data was preprocessed using a standard preprocessing pipeline, with additional segmentation of the functional images to improve coregistration. In summary, preprocessing consisted of the following steps: slice timing correction to correct for temporal differences in the slice acquisition, realignment to correct for head motion, segmentation of the structural and functional images, co-registration of the functional images onto the structural image, normalisation to warp structural and functional images to the Montreal Neurological Institute (MNI) template, and smoothing using a 6 x 6 x 6 smoothing kernel.

All co-registered images were visually inspected, and one subject had to be excluded due to errors in the data acquisition (a substantial signal cut-out from occipital regions due to an error in the planning of the acquisition).

All analyses were performed on an external cluster from the ETH, using the open source software Statistical Parametric Mapping (SPM, ver. 6906, Wellcome Trust Centre for Neuroimaging, London, UK, http://www.fil.ion.ucl.ac.uk), and MATLAB R2017b (Mathworks, Natick, MA, USA).

Subject Level Modeling

All preprocessed, subject-specific, whole brain fMRI data was modelled using a (first level) general linear model (GLM). Since the two conditions were measured in blocks with a small break in between, the data consisted of two sessions, one for each condition. The design matrix for each session consisted of three task regressors (sample, delay, and probe) and the session mean. Additionally we included and a variable number of regressors related to physiological noise (three regressors for cardiac, four regressors for respiratory phases and 12 for their interaction) and movement (12 movement regressors and a variable number of stick-regressors whenever between-slice translation or rotation exceeded 1mm and 1 deg respectively, c.f. *TAPAS PhysIO Toolbox* (Kasper, Bollmann et al. 2017)). Based on (Sladky, Friston et al. 2011), we opted not to include temporal derivatives of the hemodynamic response (HRF), as we already used slice timing correction during preprocessing. Finally, a high-pass filter with a cutoff of 128 *s* entered the GLM to remove slow signal drifts.

This standard design allowed for the investigation of the main research question, i.e. the differential activation during the delay period between the memory and control condition. Additionally, we implemented a design, where we split the delay period regressor into periods preceding correct and incorrect responses. Hereby, we were able to investigate the main and interaction effects of CONDITION and CORRECTNESS on the second level.

Group Level Modeling

To investigate the differential group effect, we specified a GLM on the group level. Here, the first level (subject-specific) contrast images entered as data. To determine, whether neural activity was consistently enhanced or depleted across participants, we ran two one-tailed-t-tests (testing for positive and negative effects) on the contrast images. Depending on the first level design, we tested for the following effects on the group level:

For the standard design, i.e. without modeling CORRECTNESS, we tested for the differential effect of the delay period for MEMORY > CONTROL and vice versa. To investigate whether ACCURACY or DIFFICULTY was linearly related to this differential activation, we additionally entered the average, subject-specific ACCURACY and DIFFICULTY values as covariates (z-scored). For the design including CORRECTNESS, we tested for the interaction CONDITION × CORRECTNESS and MEMORY_CORRECT vs. MEMORY_FALSE, in order to investigate the activation preceding correct and incorrect MEMORY trials. Significance was assessed at $p < 0.05$ FWE cluster level correction with a cluster defining threshold of $p<0.001$ (uncorrected). We did not additionally correct for the individual t-tests.

Dynamic Causal Modeling

While the GLM based, subject- and group level analyses answered questions regarding consistent (across participants) in- or decreases of BOLD signal under the experimental manipulation, they give little mechanistic understanding about the nature of potential communication between involved areas. Effective connectivity was assessed using dynamic causal modeling (DCM). DCM enables to identify the directed dependencies of a functional network that ultimately gave rise to the effects found in the GLM analyses (Friston, Harrison et al. 2003). DCMs are built from the two building blocks: 1) a neuronal equation that describes how the activity of unobservable neuronal states changes over time, and 2) a forward equation that describes in a probabilistic manner, how the hidden neuronal activity is translated into the measured signal or feature, here the BOLD response in a set of regions. Inference then refers to the process of estimating the most likely parameter values (and their uncertainty) of the system that gave rise to the observed data. For instance, the parameter values resembling the connection strength between two regions can then give meaning, why the observed neural activity was different between two conditions.

Here, we use the simplest variant of DCM for fMRI, the bilinear DCM, as we wanted to stay close to existing literature (Deserno, Sterzer et al. 2012), where however slightly different task designs and networks were used. In the bilinear DCM framework, the neural activity of state x is described by

$$\frac{dx}{dt} = (A + \sum_j u^j B^j)x + Cu \tag{6.1}$$

where A describes the fixed connectivity, B^j the additive change to the fixed connectivity under a modulatory input u^j, and C the strength of the driving input u. The (integrated) neural activity is then transformed into a predicted BOLD signal using a hemodynamic forward (Balloon) model (Friston, Mechelli et al. 2000). In the framework of DCM for fMRI, estimating the parameters means optimizing the variational free energy under Laplace approximation (Friston, Mattout et al. 2007).

Regions of Interest and Time Series Extraction

In order to extract the time series, an extra GLM was generated, where the two sessions (MEMORY and CONTROL conditions) were concatenated.

We used a three-step procedure to extract the time series for the single subjects. In a first step, we set up an additional GLM to regress out all confounds on the subject level (i.e. physiological noise and noise due to movement) and used the resulting residual images, containing only task related variance. This approach is conservative in terms of the task regressors only explaining the part of the variance orthogonal to the confounds. In a second step, we re-computed the second level statistics on these confound-cleaned data. Preliminary analyses had shown a strong correlation between the SAMPLE and DELAY regressor. Therefore, we decided to look at the combined effects of the two to define the regions of interest (ROI) and to model SAMPLE and DELAY explicitly as individual inputs to the DCMs. In brief, we defined the ROI on the group level from the differential effect SAMPLE MEMORY + DELAY MEMORY > SAMPLE CONTROL + DELAY CONTROL, within the mask defined by the main effect SAMPLE MEMORY + DELAY MEMORY + SAMPLE CONTROL + DELAY CONTROL > 0 (note that the main effect is orthogonal to the differential effect). Next, we identified the MNI coordinates of the highest T value in bilateral parietal cortex (PAR; more specifically superior parietal lobule) and bilateral prefrontal cortex (PFC) from the group level contrast images. The masking was motivated by the fact that the main effect of MEMORY and CONTROL served as the driving input to the DCM. In a third step, we used these MNI coordinates to extract the time series on the single subject level. For that, we created a mask of 8 mm radius around these MNI coordinates and extracted the time series at the subject specific maxima for the differential contrast. The time series then corresponded to the eigenvariates of all voxels within 4-mm spheres around the subject specific maxima for all four regions.

Model Space

The group level contrast SAMPLE MEMORY + DELAY MEMORY > SAMPLE CONTROL + DELAY CONTROL (masked by the respective main effect) showed significant activity in parietal and medial frontal regions, which were then defined as sources for the DCM. We only considered fully connected networks, with bidirectional intrahemispheric connections between PAR and PFC, and interhemispheric (lateral) connections between left PAR and right PAR and respectively for PFC (A-Matrix, Eq. 1). The sample period, as well as the delay period acted as driving inputs into PAR (u, 1), respectively (u, 2), and the delay period during memory as modulatory input (u^j, Eq. 1). Further, we tested all different combinations of modulatory input on the connections between PAR and PFC, always in a symmetric manner across hemispheres (B-Matrix, Eq. 1). In particular, we allowed for a modulation of condition on the connection for PAR to PFC, PFC to PAR, both, or none, leading to four submodels. Additionally, we considered local adaptation due to condition, i.e. modulation of the self-connections in PAR, PFC, both, or none, leading to four additional submodels. Together, this specified a 4 x 4 factorial design, i.e. 16 models and two potential model families naturally specified over the factors (see **Figure 55**).

BAYESIAN MODEL SELECTION AND BAYESIAN MODEL AVERAGING

For each participant (N=46) all 16 models were inverted independently. We employed a multistart procedure for the optimization and informed the hyperpriors of the model from a simple regression model (see Supplementary Material). In a Bayesian setting, the goodness of a model can be assessed through its model evidence, i.e. the marginal likelihood of the data under the prior. In practice, we often only have a lower bound approximation to the log model evidence called the negative free energy. In brief, Bayesian model selection (BMS) in terms of the highest negative free energy looks for a model with the best balance between model fit and complexity. Here, we employed a Random-Effects (RFX) BMS scheme over both, single models and families, considering that different participants might use different models or different model families (Stephan, Penny et al. 2009). As a final step, we used Bayesian model averaging (BMA) on the subject level over all models. BMA accounts for the uncertainty about the exact structure of the model and computes a weighted average of the parameters over multiple models/within a family (Penny, Stephan et al. 2010). These estimates were then used for the subsequent statistical tests.

6.5 RESULTS

6.5.1 BEHAVIORAL RESULTS

We investigated the effect of the on-line calibration of task difficulty. Overall, an adjustment phase of about 20-30 trials was needed until participants' accuracy was titrated to the desired level of 70 %, (*N*=46; average accuracy over participants = 69.5 %, SEM[40] = 2.4 %, **Figure 53**C).

6.5.2 GROUP-LEVEL GLM RESULTS

As explained in the Methods section, our primary research question addressed the differential activation between the MEMORY and CONTROL condition during the delay period. Additionally, we investigated the relation between average DIFFICULTY and the amount of differential activation and the effects of CORRECTNESS. Anatomical identification was done using the *Anatomical Toolbox* implemented in SPM12 (ver. 6906) (Eickhoff, Paus et al. 2007). All presented results are cluster level significant at $p < 0.05$ with a cluster defining threshold of $p < 0.001$ uncorrected. A list of the statistically significant clusters is provided in **Table 18**.

First, we modeled the two conditions as two sessions and looked at regions with consistently enhanced BOLD signal during MEMORY > CONTROL across participants. We found significant activation distributed bilaterally over parietal regions (e.g. the superior parietal lobule (SPL) and pre- and postcentral gyrus (PCG)) and the caudate. Additionally, we found a significant group level activation of the left middle frontal gyrus (MFG). The contralateral prefrontal region showed significant activation as part of a larger fronto-parietal cluster (**Figure 54**A). For the inverse contrast MEMORY < CONTROL we found significant depletion of activity during MEMORY in left and right posterior cingulate cortex (PCC, **Figure 54**B).

Secondly, we computed the correlation between the average, participant-specific difficulty and the differential activation, i.e. the contrast images MEMORY > CONTROL. We found

[40] Standard Error of the Mean

significant positive correlations between average DIFFICULTY and MEMORY > CONTROL in the right middle temporal gyrus (MTG) and left superior temporal gyrus (STG, **Figure 54**D).

Thirdly, we took the CORRECTNESS of the responses into account. In the MEMORY condition activity was significantly higher in MFG for trials preceding correct responses versus incorrect responses (**Figure 54**C). Neither the inverse contrast, nor the interaction between CONDITION and CORRECTNESS yielded any significant result.

Contrast	Region	x	y	z	Cluster Size	t-Value	p-Value
Memory > Control	PCG	-46	-38	56	7470	8.1202	<0.0001
	MFG	-30	2	56	584	5.8592	<0.0001
	Thalamus	-14	-12	20	150	5.1279	0.0105
	Precentral Gyrus	-54	12	32	132	5.0311	0.0187
Memory < Control	PCC	8	-50	28	533	5.0156	< 0.0001
Difficulty	MTG	58	-46	6	118	5.007	0.0285
	STG	-52	-6	-6	129	4.5559	0.0196
Memory: correct > incorrect	MFG	38	38	18	115	4.2252	0.0272

Table 18| Significant gray matter activations during working memory. Cluster level significant brain activations (FWE $p<0.05$ corrected; cluster defining threshold: $p<0.001$, uncorrected, N=46). Coordinates in MNI space. Labeling according to Anatomical Toolbox (SPM12, (Eickhoff, Paus et al. 2007)). T-Values refer to the peak coordinate of the cluster. P-Values are cluster level corrected. PCG: postcentral gyrus; MFG: middle frontal gyrus; PCC: posterior cingulate cortex; MTG: middle temporal gyrus; STG: superior temporal gyrus.

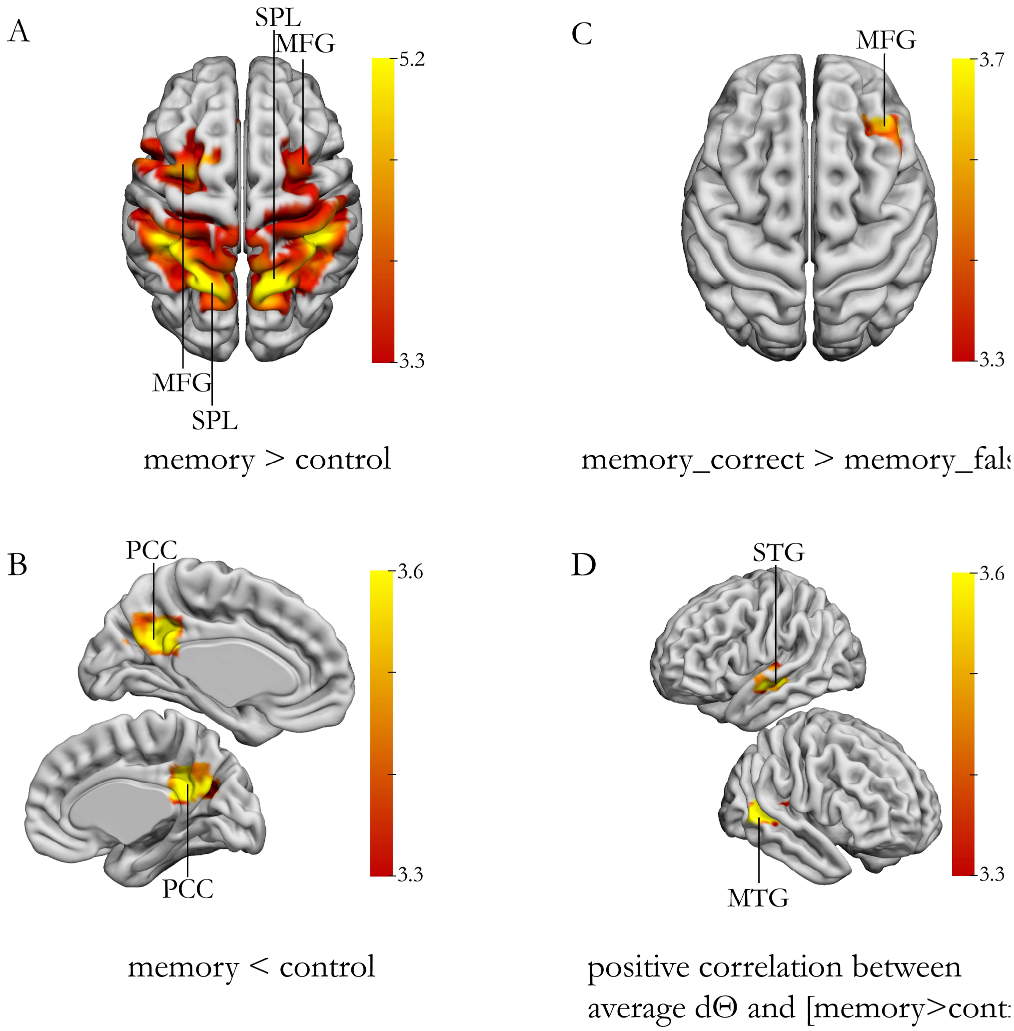

Figure 54| Group level statistics (N=46). All results are cluster level corrected (red to yellow, FWE $p<0.05$ corrected, cluster defining threshold $p<0.001$ uncorrected). Cluster sizes (kE) of biggest clusters are given in the brackets. A) Differential activation (memory > control) for standard design, without modeling correctness (kE=7470). B) Differential activation (memory < control) for standard design (kE=533). C) Differential activation (memory_correct > memory_false) for first level design including response correctness (kE = 115). D) (Positive) correlation between subject specific average dΘ and differential activation (memory > control) for standard design (kE=118).

6.5.3 Dynamic Causal Modeling

The regions of interest (ROIs) were defined from the largest differential activation SAMPLE MEMORY + DELAY MEMORY > SAMPLE CONTROL + DELAY CONTROL within the activations of the main effect. This defined the group level coordinates (in MNI x, y, z coordinates): *left PAR* (-18 -62 58), *right PAR* (20 -66 54), *left*

PFC (-24 -2 52), *right PFC* (24 -2 50). The motivation for adjusting for both, SAMPLE and DELAY during time series extraction, can be seen in **Figure 58**B. For a detailed motivation of the approach we refer to the Supplementary Material.

We considered a model space consisting of 16 models in a 4 x 4 factorial fashion. The factors correspond to the four ways how extrinsic forward and backward connections and the intrinsic self-connections could be modulated by the condition of interest, i.e. DELAY MEMORY. Parietal regions were driven through SAMPLE and DELAY (as two individual driving inputs) and connections were modulated by DELAY MEMORY (**Figure 55**A).

Each model for each subject ($N = 46$) was optimized independently starting from 100 different starting values. We diagnosed the utility of the multistart approach by comparing the result, from the starting value resulting in the highest negative free energy, to the result, from the default starting value (prior mean), in terms of free energy and explained variance. As shown in (**Figure 58**A), the multistart approach was beneficial in terms of free energy.

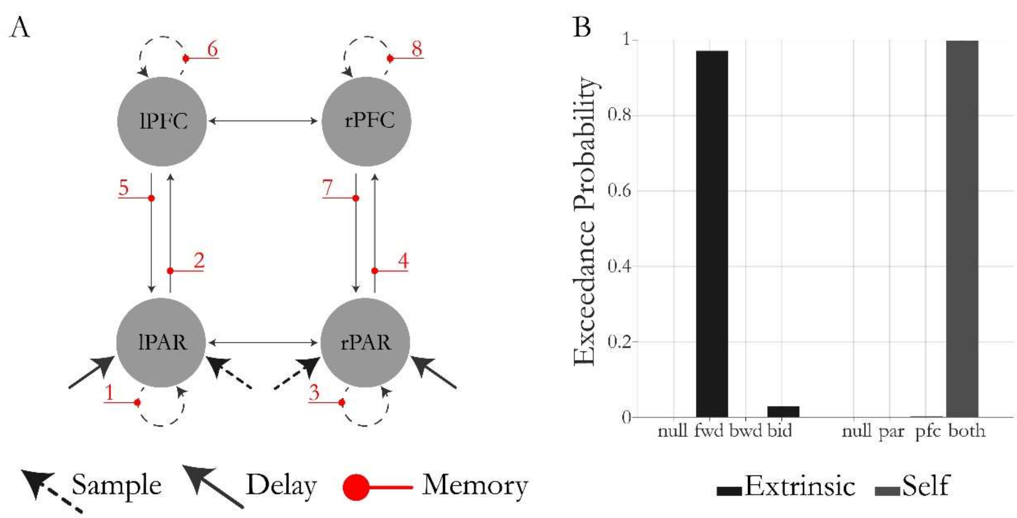

Figure 55| A) Model space. Bidirectional intrahemispheric connections and connections between homotopic interhemispheric regions. Black arrows indicate driving input (sample, delay), red arrows indicate modulatory input (delay memory). All modulatory inputs can be either on or off (for both hemispheres), leading to $2^4 = 16$ models. Numbers indicate parameter numbering in **Figure 56**) Two independent Random-Effects Bayesian (family) comparisons. All models are separated into one of four families depending on their extrinsic modulation (black family comparison) or their intrinsic modulation (dark grey family comparison).

We then ran a RFX BMS considering two model families over the two factors of the design – extrinsic and intrinsic modulation (**Figure 55**B). The family comparisons clearly speak for the presence of a modulation (by DELAY MEMORY) of the forward connection from PAR to PFC and the intrinsic connections in all four regions. This family comparison is mostly driven by the outperforming single model (*fwd_both)* with modulations of the connections *1, 2, 3, 4, 6, 8,* achieving a protected exceedance probability of about 85% (**Figure 55**B, **Figure 57**A).

To test for effects of single parameters we computed BMAs on the first level (**Figure 56**A) and ran a two-tailed t-test on the BMA means. While there is a substantial amount of variation of the modulation parameter estimates (*B*-parameters) across participants, the directionality of effect is very consistent (**Figure 56**A). In summary, DELAY MEMORY significantly increased the connection strength of the forward connections, decreased the self-inhibition in PAR but increased self-inhibition in PFC. There was no significant change in the backward connections from PFC to PAR. All results were tested at $p<0.05$, Bonferroni corrected (eight tests, i.e. $p<0.00625$, **Table 19**).

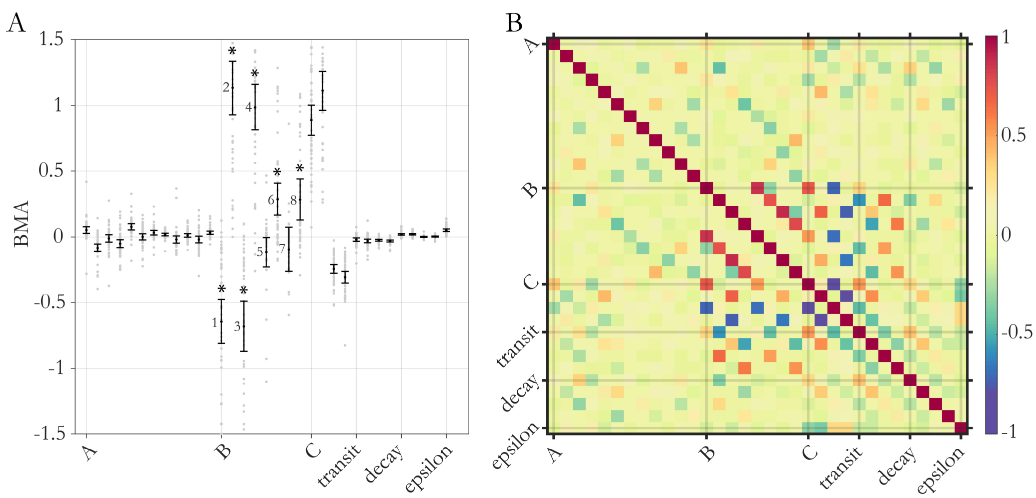

Figure 56| A) Bayesian Model Average (BMA) estimates computed on the subject level (dots). Some dots might not be visible because of the plot-range. Errorbars depict mean (over subjects), whiskers depict 95% confidence intervals. Numbers correspond to the labeling of the connections in **Figure 55**. Stars indicate significant modulatory effects ($p < 0.05$, Bonferroni corrected). B) Average (over subjects) posterior correlation matrix for the most complex model (all modulations). Modulation parameters exhibit some strong correlations with each other and input parameters.

Index	Connection	Mean	CI		p-Value (uncorrected)
1	lPAR → lPAR	-0.6451	-0.8124	-0.4777	< 0.0001
2	lPAR → lPFC	1.1334	0.9311	1.3357	< 0.0001
3	rPAR → rPAR	-0.6814	-0.872	-0.4908	< 0.0001
4	rPAR → rPFC	0.987	0.8152	1.1588	< 0.0001
5	lPFC → lPAR	-0.116	-0.228	-0.004	0.0426
6	lPFC → lPFC	0.2859	0.1639	0.4078	< 0.0001
7	rPFC → rPAR	-0.0941	-0.2614	0.0731	0.2631
8	rPFC → rPFC	0.2837	0.1274	0.4399	0.0007

Table 19| Summary of BMA estimates across participants

Despite the consistency of the estimates across subjects, we also looked at the average posterior correlation (across subjects) for the most complex model. In brief, we converted posterior covariance matrices to correlations, computed the Fishers z-transform, averaged over subjects, and converted back (inverse Fisher z-transform). These average correlations are shown in **Figure 56**, B for all parameters. We want to point specifically to the *B* and *C* (input) parameters. Strong correlations are present between all 'incoming' connections into a region: Parietal self-modulation is strongly correlated with both input and backward connections, PFC self-modulation with forward modulation. Additionally, there is a strong anti-correlation between the two inputs. This was to be expected due to the strong correlation between the two regressors. However, we would like to point out that the negative *C* parameters scaling the DELAY-input are consistent with the results from the GLM analyses. Here, the negative t-test on the main effect DELAY showed the expected activation pattern (this contrast is not shown in this paper). This is most likely because the average DELAY activity is lower than the average BOLD (enhanced by the salient visual stimulation from SAMPLE and PROBE).

We also benchmarked the quality of the inversion by comparing the amount of explained variance between a simple GLM, regressing the extracted time series against the task regressors, and the best DCM (**Figure 57**B; For information on the regression approach, we refer to the Supplementary Material – Hyperprior Settings). Based on the fact that the regression results, at least approximately, are connected to the group level effect that the DCM was trying to explain, we could see that there was a high consistency in terms of how much variance the DCMs explained. This was reassuring in two ways; i) it indicated no absurd overfitting of the data, and ii) that the features leading to the group level activity were modelled reasonably well (**Figure 57**C). Note that the 'flat-lining' of the prediction in the prefrontal regions during the control condition will be discussed in the Discussion.

All the aforementioned results were acquired by informing the hyperpriors from the results from a simple multiple regression between the extracted time series and the task regressors on the first level (SAMPLE, DELAY, and PROBE timings convolved with a canonical HRF and filtered). We compared these results under adjusted hyperpriors with results using the default hyperpriors of SPM12. From a free energy perspective, the adjusted hyperpriors lead to a clear benefit in terms of free energy (**Figure 58**C). Also, when looking at the contribution of the noise complexity to the free energy, we observed a much better behavior under the adjustment (**Figure 58**D). While for the default prior, the contribution of the noise complexity alone is about one order of magnitude higher than all other parameters combined, the contribution is much more sensible under the adjustment (Note that in pure numbers there are up to five times more non-noise related parameters than noise related ones).

6.6 DISCUSSION

The general working memory system is thought to comprise of a number of subcomponents (Baddeley and Hitch 1974). While the exact division and function of these subcomponents is debated (Courtney, Ungerleider et al. 1996, Oberauer 2002, Cowan 2016), one overarching theme is that different stimulus modalities activate different subsystems and lead to different characteristic activations of brain areas. If we pick up the notion of the slave systems related to WM, we would expect that our paradigm would fall into the so called visuo-spatial sketchpad (Baddeley 1992). Even here, previous studies have indicated differences depending of the exact nature of what needs to be retained in WM – spatial or object information (Courtney, Ungerleider et al. 1996). For retention of spatial information, the involvement of parietal and prefrontal (typically dlPFC) regions are often reported (D'Esposito, Postle et al. 2000, Curtis and D'Esposito 2003, Daniel, Katz et al. 2016), including a study which showed that training of WM can lead to overall stronger activations in those regions (Olesen, Westerberg et al. 2004). The general fronto-parietal network was identified from a group level analysis on the standard design (**Figure 54**A).

We observed cluster-level significant increases in BOLD signal during memory, bilaterally in parietal regions (SPL) and frontal regions (MFG).

In the literature PFC seems to take variable roles. A number of monkey (and human) studies point to PFC involvement depending on task difficulty (usually measured by 'memory load') for instance (D'Esposito, Postle et al. 2000, Leung, Seelig et al. 2004). At this point, we would like to point out that we don't equate our interpretation of difficulty to the traditional definition of 'memory load'. Independent of the exact definition, clamping of performance was motivated by the need to diminish a confounding source of variance in the task related neural activity across subjects. We indeed found no evidence for (linear) accuracy related differences in neural activity. Difficulty related differences were only found in temporal regions, which are not part of the general fronto-parietal network (**Figure 54**D). Note that the difficulty is inversely related to $d\Theta$ used as covariate in **Figure 54**D; the lower $d\Theta$, the higher the difficulty. Latter results were somewhat counterintuitive, as it would suggest that participants who achieved better resolution in pattern differentiation, show less differential activity in STG and MTG. It could resemble less efficient coding or different tactics in stimulus representation. However, we acknowledge that the general argument is a bit of a stretch, to assume (and test for) linear relationship between an abstract parameter modeling difficulty and the differential activation. Therefore, we refrain from drawing strong conclusion, and simply rely on the lack of evidence that the fronto-parietal network was confounded by DIFFICULTY in this self-calibrating paradigm.

Behaviorally, titration to the desired difficulty level of 70% correct responses happened around trial 20, which could certainly be improved with a more adaptive titration protocol (e.g. changing step sizes or window size over the course of the experiment). However, the benefit is that it allows the participant to get accommodated to the experiment by reducing frustration early on, which could potentially be an issue for a clinical population.

The negative contrast, i.e. MEMORY < CONTROL, yielded a significant increase of BOLD signal during control in PCC. This could very well be a (partly) activated default mode network, as cognition is basically unconstrained in the delay period during CONTROL.

By only looking at trials of the memory condition divided by the correctness of response, there is a significant difference in the right MFG between MEMORY trials preceding correct and incorrect responses. While MFG is generally part of the fronto-parietal network, there is no direct overlap between the effect of correctness and MEMORY >

CONTROL. Based on the previously hypothesized role of PFC, the decrease in activity could be interpreted as a failure in focusing attention stored in PAR and thus leading to incorrect responses. Interestingly, other studies have also reported right hemisphere lateralization in WM paradigms, a finding which has also been focus of investigation in schizophrenia (Walter, Wunderlich et al. 2003, Van Snellenberg, Torres et al. 2006, Nagel, Herting et al. 2013).

Using DCM, we could then successfully model the activation in the bilateral fronto-parietal network, achieving similar (or higher) explained variance than by using regression. Model selection strongly indicated memory-induced local changes in connectivity and increases in the connection strength from PAR to PFC. This could resemble a recruitment of PFC during memory, in line with the 'focus of attention on the internal representation' role (Curtis and D'Esposito 2003, Postle 2006, Lara and Wallis 2015). Alternatively, (D'Ardenne, Eshel et al. 2012) formulated it as a gating mechanism mediated by dopaminergic neurons regulating the encoding of the stimulus representation in PFC. Our results support either interpretation (and distinguishing between them was not the goal of these analyses), as the model predicts no consistent (across participants) delay period related activity in PFC in the control condition (**Figure 57**C). The facilitation of this increase of connection strength could be due to long range glutamatergic connections – something that is definitely not accessible with fMRI, but will be one of the research questions for the sibling EEG studies. The EEG counterpart of DCM directly models inhibitory and excitatory processes of neuronal population, or in the more advanced models, even channel dynamics directly for different types of ion channels (NMDA, AMPA or GABA) (Moran, Pinotsis et al. 2013).

There are a number of studies where DCM has been applied to fMRI working memory data but the results tend to be mixed. In a verbal n-back WM task, (Nielsen, Madsen et al. 2017) compared healthy controls to first episode schizophrenia patients (FES), using a network of three regions between V1, inferior parietal lobule (IPL) and inferior frontal gyrus (IFG). Here their model selection results pointed towards modulation of forward and bidirectional connection between IPL and IFG, with less emphasis on purely backward connections. Similarly, (Dima, Jogia et al. 2014) used DCM in a verbal n-back WM paradigm. Among other things, they report a strong evidence for MEMORY modulating the forward connection (posterior to anterior) in the 2- and 3-back condition (predominantly for the right hemisphere!).

In contrast, other studies have reported more emphasis on modulation of backward (frontal-to-parietal) connections. The primary results of Deserno et al. (2012) indicates evidence for a change in backward connection strength during memory (Deserno, Sterzer et al. 2012). They used the same network as (Nielsen, Madsen et al. 2017), but employed a numerical n-back task. In another study, (Heinzel, Lorenz et al. 2017) looked at WM related effective connectivity in a younger and older population during a n-back WM task ($n = 0,1,2,3$). They also observed overall higher evidence for backward modulation (PFC to PAR). Interestingly, the BMA estimates were load-dependent, and increased with increasing *n*. Also more emphasis on backward connections (for the control group) were reported by (Schmidt, Smieskova et al. 2014). They used a similar network to ours in a letter n-back task, but with a slightly different model space. Interestingly, they also observed a right hemisphere lateralization in terms of modulation strength.

Of course, this is not a comprehensive list (also see (Ma, Steinberg et al. 2012, Nielsen, Madsen et al. 2017, Jung, Friston et al. 2018)). But these studies used bigger networks, stochastic DCM (Jung et al. 2018) and direct comparison, especially regarding to the front-parietal interaction, is difficult due to the models ability to mediate modulatory effects through additional regions.

There are a number of reasons that could explain the discrepant findings, both in terms of task design (e.g. letter vs. numerical n-back tasks) and modeling (e.g. considered regions, modulation and connectivity structure etc.). For example, we intentionally abstained from using a variant where PFC receives direct driving input (or mediated through V1), as it leads to many symmetries in the model, which can be a problem for estimation (something that we also observe in terms of the posterior correlation between modulatory parameters that effect the same source (**Figure 56**B)). Hence, it could very well be that given direct input to PFC the dynamics could equally well be explained through a modulation of the backward connection. Generally, in this kind of Bayesian inference, the inference drawn is always dependent on the set of models tested.

In conclusion, we presented a novel spatial delayed match-to-sample working memory paradigm measured in fMRI. Novelty and usefulness of the paradigm were successfully assessed in terms of three criteria stated in the Introduction. First – the titration of the difficulty allowed for an equal level of performance accuracy across participants. Titration could be done on-line, removing the need for prolonged pre-testing, which would mean another burden for a clinical population. Additionally, the trial-by-trial changes in difficulty take effects into account that might have arisen during the experiment, like tiredness or attention. Second – we did find a delay period related enhancement of neural activity during memory over distributed parietal regions and middle frontal regions. This is essential for our task to be a suitable paradigm to investigate prefrontal dopaminergic neuromodulation, the dimension along which a heterogeneous psychiatric population is hypothesized to differ. Finally, we could successfully apply generative models of directed connectivity (i.e. DCM), plausibly suggesting a network where MEMORY recruits resources in PFC to retain the stimulus related information. Collectively, these findings give us substantial information for the modeling in the sibling EEG studies, acquired in patients and healthy controls in a pharmacological setting.

6.7 SUPPLEMENTARY MATERIAL

6.7.1 ASSESSMENT OF TIME SERIES EXTRACTION

First, we identified the group level coordinates from the largest effect found for the delay period, i.e. DELAY MEMORY > DELAY CONTROL. For time series extraction we then adjusted for the F-test removing all unwanted variance but the effects of DELAY (Extraction 1). Second, we identified the group level coordinates from the largest effect of SAMPLE MEMORY + DELAY MEMORY > SAMPLE CONTROL + DELAY CONTROL. On the subject level, we adjusted for the respective F-test, i.e. keeping the effects of delay and sample and removing all other unwanted variance from the time series (Extraction 2).

We then computed a regression of the task related regressors on the respective extracted time series and compared the amount of variance explained between the three options: (i) regressing DELAY on extraction 1; (ii) regressing SAMPLE+DELAY as a single regressor of 4.5s on the data from extraction 2; (iii) regressing SAMPLE and DELAY as two individual regressors on the data from extraction 2.

In summary, we observed that keeping SAMPLE and DELAY as two individual inputs allowed us to keep substantially more task-related variance (**Figure 58**B). While the data from extraction 1 and 2 is not the same, it appears as if a lot of task-related variance was lost due to the correlation between SAMPLE and DELAY, when only adjusting for DELAY. Therefore, we decided to extract all the variance and model the effect induced by SAMPLE and DELAY directly as two individual inputs in the DCM.

6.7.2 DYNAMIC CAUSAL MODELING

MULTISTART

As mentioned earlier, inversion of the model refers to trying to optimize an objective function, with respect to some parameters, given data. The standard inversion method in SPM12 employs a gradient descent based optimization scheme on the approximate free energy, which, although computationally efficient, has the downside that it can’t overcome local maxima arising in non-convex optimization problems ((Penny and Sengupta 2016) illustrated the issues for DCM for EEG). To overcome this caveat, we inverted the models from 100 starting values randomly sampled from the prior. Hence, all results presented in this paper correspond to the results acquired from the starting value that led to the highest negative free energy.

HYPERPRIOR SETTINGS

The hyperprior mean and variance define a priori expectation about the amount of irreducible noise in the data and the certainty about this expectation. The default setting in SPM12, ver. 6906, sets this expectation at a mean precision (inverse variance) of $\mu_H = 6$ with a variance of $\sigma_H^2 = \frac{1}{128}$, i.e. *very* low noise, with extremely high certainty. While this setting might be suitable for very salient paradigms with high signal to noise ratio (SNR), we were skeptical about the validity of this assumption for this particular paradigm. Furthermore, the combination of very low noise with high certainty leads to two downsides: first, the noise term (mathematically) affects the objective function implausibly strongly, leading to a very rough landscape making it hard to find the optimum. Second, this in turn seems to affect model selection in favor of more complex model, as the penalty for model complexity is overwhelmed by fitting single datapoints better, effectively leading to overfitting.

After a first analysis under the standard prior and observation of the described pitfalls, we opted to inform the setting of the hyperpriors by computing a simple multiple regression on the extracted time series, using the task regressors of the first level. The hyperpriors were then estimated as the average residual precision (averaged across subjects) and the variance (across subjects) thereof.

Formally, we validated this approach also in terms of free energy, as illustrated in **Figure 58**C.

MODELING DIAGNOSTICS

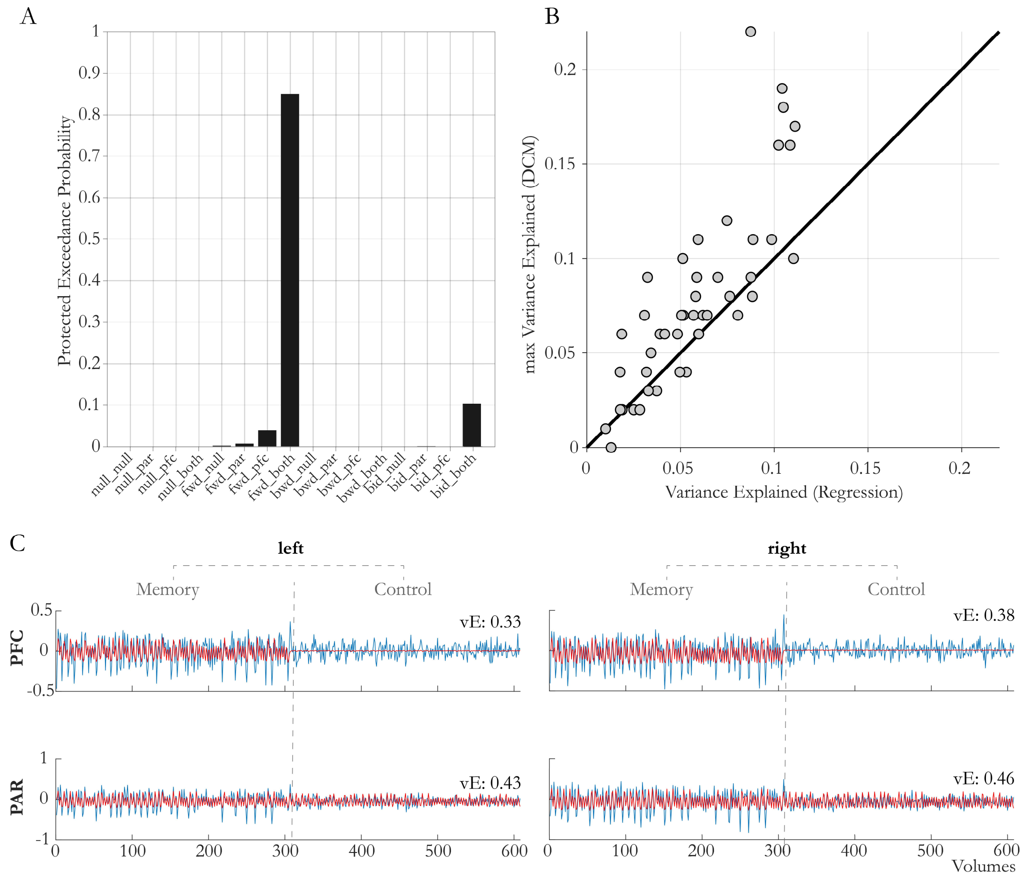

Figure 57| A) Random effects model comparison for all individual models (all subjects). B) Comparison between the maximal variance explained with DCM and with a simple linear regression, where we used the first level GLM (task regressors) on the extracted time series. C) Average fit (red) vs average data (blue). Averages are computed over subjects, for the model with the highest protected exceedance probability (fwd_both). vE indicates explained variance by the average.

MULTISTART DIAGNOSTICS

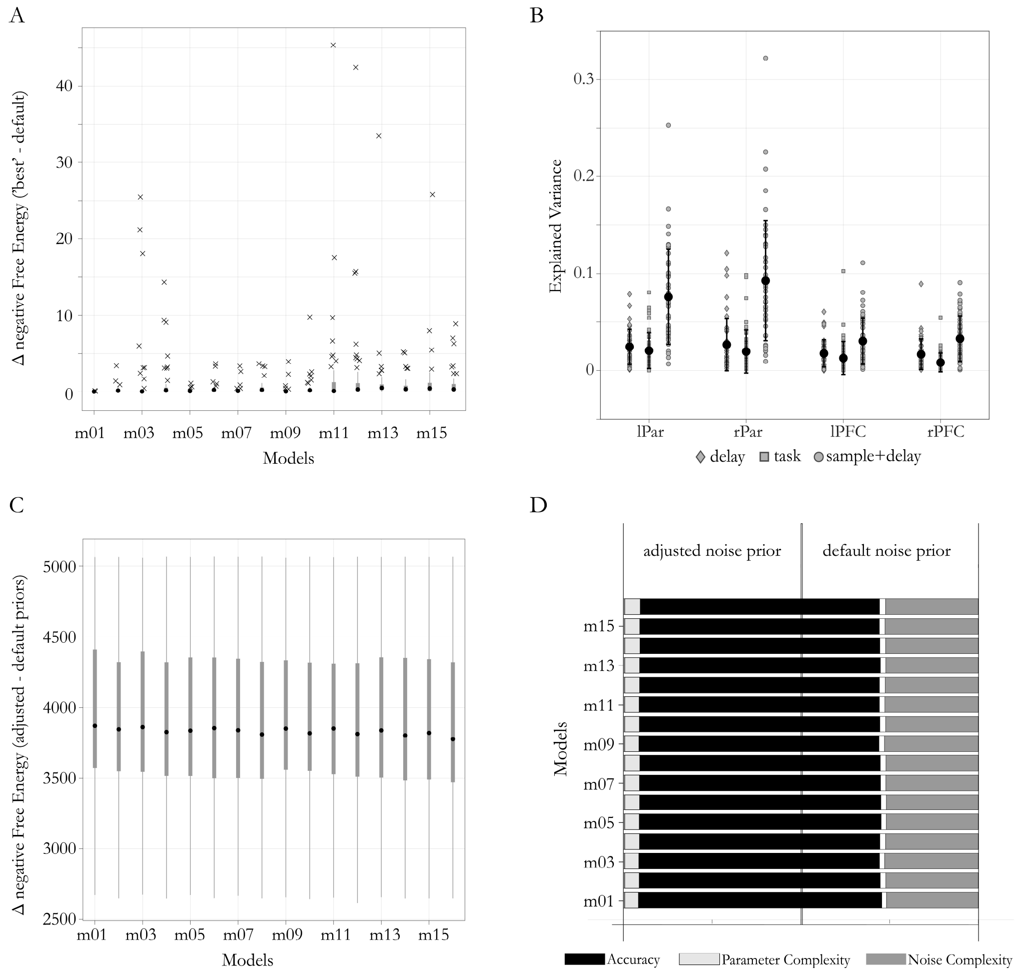

Figure 58| A) Multistart diagnostics: Difference in negative free energy between inversion starting from default starting value and the starting value leading to the highest negative free energy. Circles represent median, crosses represent outliers. B) Regression on the extracted time series (extracted with the respective F-test). Delay; Delay Regressor (diamond), Task; Sample and Delay as one regressor with 4.5s duration (square); Sample+Delay; Sample and Delay as two independent regressors (circle). C) Noise modeling diagnostics. Difference in negative free energy between best inversion for the default noise prior and the adjusted noise prior. D) Contribution to the free energy of the accuracy and complexity terms. Note the implausibly high contribution of the noise complexity for the default noise prior.

7 | DISCUSSION AND OUTLOOK

7.1 THE STORY SO FAR

This dissertation has made the following five contributions to move the application of DCM further towards clinical utility.

In Chapter 2, we have presented different variants of delay differential integration schemes for DCM for ERP. The most important reason was to assess whether the delay-inherent mechanics are accurately mapped by the integration scheme. Considering the expected delays in biophysical systems, our results indicate that delays contribute to the dynamics of the system which cannot be appropriately taken into account by the default integration scheme. This has an effect on both, parameter estimation and model selection, as estimation of directed, condition specific effects might not respect their temporal succession. Our simulations focused on simple systems and augment the work by Lemaréchal and colleagues (Lemarechal, George et al. 2018), who nicely illustrated relevant integration-scheme related differences in simulations and empirical data. While the authors made their code publicly available, this has, to the best of our knowledge, not (yet) found its way into the standard implementations of SPM or the application to empirical data. A current search on Google scholar yielded three citations of the paper (Jafarian, Zeidman et al. 2019, Kashyap, Bhattacharjee et al. 2019, West, Berthouze et al. 2019). However, none of them have actually employed their scheme for the analysis of data (one is a review). There are many possible reasons why this could be the case, most importantly perhaps that the publication is still relatively novel and changes of these sort take time. However, apart from that, we see two other critical issues that may contribute to the slow uptake: speed and visibility. We hope to have been able to address these two points by (i) demonstrating that the continuous extension of ODE methods was only about a factor of two slower than the standard scheme, and (ii) increasing visibility and usability of our tools by making them part of TAPAS which has a solid user base.

The second (Chapter 3) and third (Chapter 4) methodological contributions have the same underlying premise: How representative and robust are the measures we propose to put forward as computational assays of pathophysiology? There is evidence based on other fields of modeling that the estimation of complex dynamical systems is challenging. For example, a study on models used in systems biology investigated models with around 20-30 kinetic parameters (Gutenkunst, Waterfall et al. 2007). The authors showed extremely large confidence intervals associated with estimated parameters, which still yield tight predictive densities. They use the notion of 'sloppiness', referring to large eigenvalues of a sensitivity matrix composed by the Hessian – not un-similar to the computation of the posterior covariance matrix used in DCM.

We argued that it is crucial to understand the landscape over which parameters are optimized. From a theoretical perspective, there have always been known uncertainties as to how much optimization of DCM is affected by potentially multi-modal posterior distributions and how this might confound BMS (Daunizeau, David et al. 2011). But it is only with the computational power now that these uncertainties can be rigorously tested. Apart from the work presented in this thesis, there have been other efforts in testing robustness in DCM for ERPs. There, potential multi-modality was directly addressed by using a sampling based method (Penny and Sengupta 2016, Sengupta, Friston et al. 2016) which has the advantage of not relying on the Laplace approximation as described in Chapter 3 and converging to the true posterior in the limit of infinite samples. They illustrated nicely that multiple local optima are present in the sampling distributions and an overall superior performance in parameter estimates for sampling-based methods when compared to VB. Unfortunately, to the best of our knowledge, this work has not been followed up so far. One potential reason that might have prevented a more enthusiastic uptake of the approach is that MCMC is computationally very expensive (although recent advances have been made to speed up sampling-based inversion (Aponte, Raman et al. 2016, Aponte, Raman et al. 2018)). As a matter of fact, staying in the VB framework and running multistart inversions (as done here) for a moderately sized dataset and model space also requires the optimization of 10000+ models. Realistically, this is only feasible with access to high performance clusters. However, if access to such a cluster is available, our approach may have advantages over sampling-based approaches as it is much easier parallelizable and stays within a framework that is established in the community for years. On a more general level, this also raises the non-trivial questions for an optimal tradeoff between the accuracy of an inversion scheme and the run-time of that algorithm. While it may or may not be acceptable to obtain less accurate results for the sake of high computational efficiency to address basic neuroscience questions (e.g., by only looking at group effects), the situation is arguably different for what CP tries to achieve. As highlighted before, to render DCMs into computational assays that are truly meaningful for everyday clinical practice, robustness is paramount and hence, physicians may happily accept long runtimes of the (computational) "test" if results are truly informative and robust (Frässle, Yao et al. 2018).

We have discussed the technical aspects of our findings at length in Chapters 3 and 4, thus we will focus on the aspect of clinical utility. We argue that three things are to be considered. First, we are unlikely to find the 'true', global optimum with a finite number of starting values. This implies that theoretical properties of the negative free energy (superiority of models, generalizability, etc.) cannot be guaranteed in a strict sense. Second, the negative free energy is, for the case of many datapoints, very much accuracy driven and prone to bias induced by the hyperpriors and noise structure. We have provided solutions for both problems. Both are intuitive from a Bayesian perspective and address the balance between accuracy and complexity in DCMs directly. This was done either in terms of giving the hyperprior less weight (wider noise prior / informed hyperpriors) or giving the data less weight and bringing the residual structure in closer agreement with the model assumptions (down-sampling). For empirical data we would expect that a combination of both is needed to avoid entering numerically critical regimes between accuracy and regularization. This might still imply that other measures have to be considered to assess generalizability (e.g. posterior predictive densities, cross-validation, etc.). Third, ground truth does not necessarily correspond to the solution with the highest negative free energy, neither in terms of best model nor in terms of parameter estimates. This is not surprising given the correlations between some parameters, but consequences for parameter estimates are that (true) effects can get diluted and spread over multiple parameters[41]. Overall, the multistart really showed its utility in recovering true parameters better.

[41] In our simulations, modulatory effects were always underestimated (on average).

From a translational perspective, it is probably critical what the dimensionality is that ultimately allows for patient stratification. Despite the aforementioned points on the negative free energy, we would still consider it an interesting quantity for subgroup identification. But based on the evidence provided in this thesis, it will crucially depend on a suitable balance of accuracy and complexity, which seems to go beyond a one-size-fits all solution in prior specification. On the other hand, if parameter estimates span the dimension for patient stratification, it seems critical whether true, pathophysiolocial effect sizes do get distributed over multiple parameters. Then, statistics on point estimates could be less meaningful than classification approaches. For the typically small sample sizes in patient studies, cross-validation with permutation testing should then be the goal standard (Varoquaux, Raamana et al. 2017, Varoquaux 2018).

While this may shed a rather critical light on DCM, it should be emphasized that all analyses and arguments have to be understood under the premise of using DCM and the free energy framework as a clinically relevant measurement device that lives up to the highest standards. At the moment, we would argue that the measures this 'device' produces go beyond the report of negative free energies and point estimates. Nevertheless, our findings do suggest that DCM can serve as a meaningful feature reduction tool. These features might not *uniquely* resemble extrinsic or gain modulation. But they still can be used for classification and jointly allow for predicting – as we have, for instance, demonstrated with regard to pharmacological manipulations in the *RATMPI* study. This arguably puts DCM in a spot between completely agnostic approaches such as deep-neural-nets and the unique attribution of inferred effects onto single, biophysical properties. There is still an information gain both in terms of network architectures and condition specific effects that can be obtained from biophysically interpretable result conditioned on confounding variables that might explain similar effects than with a purely data-driven approach.

In the long run, we think that the goal should be to enforce constraints and find common, good solution for all individual dataset. While it would require a much more formal investigation than started in this thesis, the early stopping criterion induced by the default starting values could be an interesting consideration for the future (Caruana, Lawrence et al. 2001). Another formal way of constraining are the use of hierarchical models, where subject level parameters are informed by a hierarchically higher level representing a population distribution (Yao, Raman et al. 2018). We have presented a way how this can be done in the supplementary material of the RATMPI chapter for multiple pharmacological conditions. Our approach can, in principle, be seen as a way of hierarchical modeling with infinite precision on some population priors (e.g. delays, kernel gains, etc.). This approach could also be used to infer on multiple subjects simultaneously. Another possible application could be to invert multiple models simultaneously. In this context, open questions would remain regarding model comparison, but correlations of modulation parameters could potentially be reduced.

In the empirical chapters, which are in preparation for publication, we have applied DCM to both EEG and fMRI data, using the advances we made in the methodological chapters. In the first study, we looked at evoked response potentials elicited during an auditory mismatch negativity paradigm. Here, we showed the utility of DCM to act as an informed way to reduce the complexity of the LFP data to 18 parameters effectively modeling synaptic action. Out of these 18 parameters, an even smaller subset showed linear effects of drug. Additionally, we could distinguish multiple pharmacological conditions based on a linear classifier. Aware of the problems of small sample sizes in classification, we assessed significance through permutation testing. Two things make the results even more convincing. First, we did not hand-pick the parameters used in the classification based on the classical statistics. Hence, there was an increased risk of overfitting when training the classifier due to the higher number of input features (i.e., model parameters), but still, the permutation test statistics yielded the true labels to carry significant information. Secondly, the supplementary material to this chapter showed the same analysis for the 20% deviant probability condition, and the alternative modeling approach. All strategies consistently showed the overall ability to differentiate the drug conditions based on BMA estimates. Sometimes even better than with the a priori specified modeling pipeline. Importantly, all classifications that were subject to some level of constraint - either by fixing parameters across drug conditions or by fixing the starting value – performed better than under the classical multistart.

In the fMRI studies where participants performed in the novel Working Memory (WM) paradigm, the model predicted nicely the involvement of prefrontal regions during WM. Interestingly, the model predictions indicate that there was virtually no systematic activation (across participants) in PFC during the control condition delay period. Despite the much more linear equations underlying DCM for fMRI, the multistart also proved its utility for this dataset. Similarly, as for the case of DCM for EEG, we can again only stress the necessity of carefully diagnosing results. The explicit modeling of the sample stimulus in the DCM and informing the noise hyperpriors based on the GLM results would not have been detectable by only looking at BMS results. Importantly, the informed setting of hyperpriors is generally applicable in DCM for fMRI and should be used if one cannot strongly defend the default settings.

Overall, this thesis has aimed at providing a deeper understanding of model inversion in DCM by conducting thorough and systematic assessments of the implications of particular choices and assumptions (e.g., integration scheme, hyperprior specifications, noise settings) for the robustness and interpretability of modelling results. Based on our observations, we have provided practical solutions that will hopeful prove beneficial for future studies aiming to apply those tools to EEG or fMRI data. In particular, we demonstrate how this could be achieved in the context of our two empirical datasets.

7.2 THE STORY TO COME

In the empirical EEG study, we have shown the ability of DCM for ERP to detect muscarinic effects. Obviously, this should only be understood as a proof of concept of the translational relevance for psychiatry, as the data was acquired in rodents where electrodes could be placed very close to the neuronal sources. Nevertheless, the results hold promise that these models are indeed sensitive to neuromodulatory processes and thus clinically relevant mechanisms in the brain. However, to move these observations closer to an application in humans, I was involved in the planning and acquisition of EEG data in healthy volunteers, who performed essentially the same working memory paradigm as presented in the fMRI study. In total, 162 participants where measured in a between-subject, double-blind study design, with pharmacological interventions affecting the cholinergic or dopaminergic system (placebo controlled). The study was conducted in two separate arms with 81 participants each. The two arms consisted of drugs with antagonistic and agonistic action for the two neuromodulatory systems[42]. In the first study arm, we plan to examine the effects of amisulpride and biperiden, which act antagonistically on presynaptic D2/D3-autoreceptors and muscarinic (M1) receptors, respectively. In the agonist arm of the study, we investigate the effects of the acetylcholinesterase inhibitor galantamine and the dopamine precursor L-Dopa.

[42] The pharmacological action was either with respect to re-uptake, availability or receptor function.

Here, the goal will be to identify drug-induced changes in the oscillations (power spectrum) during the delay period. Based on an interesting finding by Moran et. al., 2011 we would expect particularly changes in low-frequencies under Levodopa (Moran, Symmonds et al. 2011). DCM for EEG most naturally maps these frequencies to synaptic action as the convolution kernel can be understood as a filter. On the other hand, the conductance based formalism allows to explicitly model NMDA-channel action. This is particularly intriguing, because simulation studies have shown that the longer time-constants of NMDA channels are needed to generate prolonged, sustained activation (Durstewitz and Seamans 2002). Also in the mentioned empirical study, (Moran, Symmonds et al. 2011) showed that the DCM parameters associated with the NMDA-nonlinearity showed a significant interaction between drug and memory. Additionally, the change in these estimated parameters was predictive for performance – a purely behavioral measure. While we overall cannot comment on the computational challenges for spectral DCM based on our methodological assessments, these results are very promising for a translational purpose. It shows the potential of measuring the interactions between NMDARs and neurotransmitters (NNI) from peripheral readouts by means of mathematical modeling.

In the same line of thought, this study will inform the analyses of a prospective patient study. Here, we are in the process of measuring patients suffering from schizophrenia as they undergo a medication switch from an antipsychotic treatment without cholinergic to an antipsychotic with cholinergic effects. Although multiple meta-analyses show large variation of treatment outcomes to different antipsychotics (e.g. (Leucht, Komossa et al. 2009, Leucht, Cipriani et al. 2013)) the lack of tools for clinicians to predict both, efficacy and side-effects represent major clinical problems. Olanzapine and clozapine, both so called atypical (or second generation) antipsychotics have proven much potential with regard to their efficacy. Both antipsychotics bind to muscarinic receptors, with low affinity towards dopamine receptors (Bymaster, Felder et al. 2003). This potency has led to an augmentation of the dopamine hypothesis of schizophrenia (Raedler, Bymaster et al. 2007). Although both medications are superior to many other antipsychotics, prescription is hindered by potentially harmful side-effects (Leucht, Cipriani et al. 2013). This has led to clinical guidelines (American Psychiatric Association, 2019; National Collaborating Centre for Mental Health (UK), 2014) recommending switching to clozapine only after two antipsychotics have shown to be ineffective. Patients would largely benefit if one could predict the success of a treatment with clozapine or olanzapine in order to enable an early remission. However, there are still no clinical properties or measures established that would help to decide if an early prescription or a switch to olanzapine and clozapine is useful in an individual (Kahn, van Rossum et al. 2018). Our patient study is investigating exactly this question. Among other tasks, patients perform the working memory task (measured with EEG) up to four days prior to or after the medication switch, i.e., right before or after a perturbation of the system by either olanzapine or clozapine. They are clinically followed up two and eight weeks after the switch to assess treatment success. This study therefore directly addresses the question of clinical utility. If it were possible to predict success of the medication switch before (or directly after first administration) based on a set of model parameters, this could prevent possible failure of prolonged trial and error of different medication or at least allow a risk stratification of potential efficacy and side-effects. Very similar to the rodent study, it could be formulated as a classification problem, using DCM estimates as features and treatment success or clinical symptom scores as classes/output. Because of the very difficult recruitment of patients, the study was augmented with a matched, healthy control group, allowing for cross-sectional comparisons.

These two studies consequently bring the clinical utility of DCM in humans one step further. To achieve aspired goals, the findings presented in this thesis will be paramount as they raise awareness for potential practical limitations of DCM, highlight the need for careful diagnostics of inferred results, and provide the machinery and procedures to address some of these limitations.

APPENDIX A| ANALYSIS PLAN FOR THE RATMPI STUDY

Internally, the project is referred to as 'RATMPI'.

Dario Schöbi, Jakob Heinzle

FOREWORD

This is the analysis plan concerning the analysis of the pharmacological 'black hooded rat' dataset acquired at the Max Planck Institute for Metabolism Research in Cologne.

It concerns a paradigm, where epidural (electrophysiological) signals were recorded in awake rats during an auditory oddball mismatch negativity paradigm (MMN). For each rat, multiple sessions were recorded manipulating two experimental factors: The probability of a deviant tone occurring, and, in a subset of the rats, a five-fold pharmacological intervention.

Importantly, the subset of the rats that did not undergo the pharmacological intervention, is used as a training data set (later referred to as *noPharma*). This training set is used to inform the preprocessing pipeline, definition of the time window for statistical analyses but most importantly, the definition of the priors and scaling parameters for Dynamic Causal Modeling. Crucially, this allows us to constrain the analysis without having to look at the pharmacological data and thus excludes any risk for double dipping.

The goal of this analysis plan is to formulate the research question, the distinction between planned and potential post-hoc analyses, the rational for certain analysis steps and generally the information needed to reproduce the results. All code is version controlled and documented on GITLAB (https://tnurepository.ethz.ch/dschoebi/ratmpi).

RESEARCH QUESTION

There are three major questions that we want to address with this study. For their immediate impact on the modeling decisions, please refer to the section on *'Dynamic Causal Modeling'*.

1. Can we identify a drug dependent change in connection and/or gain parameters
2. Can we identify characteristics in the dose response curve
3. Can we predict the pharmacological condition for left out rats

All questions might be phrased in a comparison between

1. Vehicle vs. Drug
2. Vehicle vs Agonists vs. Antagonists
3. Individual dosages

DATASET

- 16 black hooded rats split into 2 major groups: without (6/16, called *noPharma* or *non-pharmacological data* throughout) and with (10/16, called *pharma* or *pharmacological data* throughout this analysis plan) pharmacological intervention (listed below)
- epidural recordings bilaterally from A1 and PAF
- auditory oddball MMN paradigm
- Two datasets: dataset1 and dataset2. They refer to the frequency (tone height) of the standard tone.
- Two designs: MMN_0.1 and MMN_0.2. They refer to the probability of a deviant tone being 10% and 20% respectively.
- approximately 1000 trials. (Note: The sequence of the auditory stimulation was not fixed, and tones presented with given probability. The experiment was stopped, when 100 and 200 *Deviant* tones (for 10% and 20% condition) were presented (with one exception in one rat, where a slightly different protocol was used.). Hence, the number of *Standard* tones varies.

Experimental factors (within) for non-pharmacological group:

- standard coding: s16-18_d7-9, s7-9_d16-18
- deviant probability: 10%, 20%

Experimental factors (within) for pharmacological group:

- standard coding: s16-18_d7-9, s7-9_d16-18
- deviant probability: 10%, 20%
- pharmacology: 2mg Scopolamine, 1mg Scopolamine, Vehicle, 3mg Pilocarpine, 6mg Pilocarpine (Note: The placebo condition in the *pharma data* will be called "Vehicle" throughout this analysis plan in order to distinguish it from the *noPharma* test data.)

This analysis plan concerns primarily the analysis of the pharmacological group, where the non-pharmacological group is used to inform the overall pipeline. The MMN_0.1 and MMN_0.2 dataset will be treated independently.

PREPROCESSING

Preprocessing has been made fully compatible with SPM12, and includes the following steps:

RAW-DATA

The raw data is stored in two separate data files in ASCII format. Each of these data files corresponds to one of the two experimental frequencies to be the standard frequency, and are called *MMN_0.1_s7-9_d16_18* and *MMN_0.1_s16-18_d7-9* respectively. They will be generically referred to as *dataset1* and *dataset2*.

The format of the data is slightly different for two sets of rats and summarized below:

Raw data for animals with code: 27905, 27907, 27908, 27909

File name: rawDataNew.ASC

Matrix Columns:

1. Discontinuous Time (zeros, when deviant potential was inserted)
2. Posterior Auditory Field electrode right (rPAF)
3. Posterior Auditory Field electrode left (lPAF)
4. Anterior electrode right (rA1)
5. Anterior electrode left (lA1)
6. Potential Trigger Standard Tone
7. Continuous Time (no zeros)
8. Potential Trigger Standard Tone (equivalent to column 6)
9. Potential Trigger Deviant Tone

Raw data for animals with code: 29985, 27986, 27987, 27988, 27989, 27990

File Name: raw Data.ASC

Matrix Columns

1. Time
2. Posterior Auditory Field electrode right (rPAF)
3. Posterior Auditory Field electrode left (lPAF)
4. Anterior electrode right (rA1)
5. Anterior electrode left (lA1)
6. Potential Trigger Standard Tone
7. Potential Trigger Deviant Tone

Special Remarks.

- Control Rats (27120, 27121, 27123, 27124, 27125, 27126), where no pharmacology was applied (*noPharma dataset)* are not explicitly listed but follow the *raw Data.ASC* format.
- 27906 died after the first treatment.

The following steps will be performed on the individual datasets (*dataset1, dataset2)* until the *merging* step.

Convert

The ASCII files are loaded and read into a new matrix *data* only consisting of the columns of interest.

data = Time, lA1, rA1, lPAF, rPAF, standard Trigger, deviant Trigger.

Time of events are detected using *Kai Brodersens* peak finder routine from the original analysis (by Fabienne Jung). A trigger is defined as the event, where the amplitude in columns 6 and 7 of *data* exceeds 3. The next 100 ms are then ignored to avoid finding the same peaks multiple times. The times are then stored in a trigger file.

The data was sampled at 2000 Hz. All activity columns (data(:, 2:7)) are converted into an SPM compatible file format.

Event Definition

Events time are loaded from the trigger file, and marked in the SPM dataset as Stimuli of type *'devTrigger'* and '*stdTrigger',* respectively.

Chantype

Channel type (as returned by the SPM methods *D.chantype*) are set as 'LFP' for 'lA1', 'rA1'; 'lPAF', 'rPAF', and 'Other', for *'stdTrigger'* and *'devTrigger'*.

Filtering and downsampling

The data is **highpass** filtered (1 Hz, 5th order butterworth filter), downsampled (**1000 Hz**) and **lowpass** filtered (30 Hz, 5th order butterworth filter). Note that the downsampling routine applies an additional lowpass filter to prevent aliasing.

Epoch

The data is epoched [-100, 500] around stimulus onset and baseline corrected. Time windows for statistical analysis are defined below.

Merge

Dataset 1 and dataset 2 (both preprocessed up to and including epoching) are merged into a single dataset. This is done to eliminate frequency specific effects [citation Fabienne Jung, Dissertation].

Artefact Rejection

All trials exceeding an amplitude of 500 μV are excluded from the analysis. Note that other methods, such as z-score (within and between trials) were tested (on the *noPharma* data), but there was no indication for them being superior (assessed by visual inspection). The chosen threshold was deliberately high due to the large number of trials.

Average

Standard Averaging over all non-excluded trials (due to artefact rejection) is performed for both conditions.

Crop

The final preprocessed dataset is split into the two hemispheres. This allows us to compute the classical and model based analyses individually for both hemispheres. Hence, rats with only a single region in one hemisphere missing or with poor recording do not need to be excluded completely from the analysis.

CLASSICAL ANALYSIS

EXCLUSION OF RATS

Averaged (over trials) evoked responses were visually inspected for both hemispheres, all rats and conditions. The only exclusion criterion was a clearly aberrant recording, judged through visual inspection, prior to the model based analysis.

ANALYSIS WINDOW

All grand average / difference wave figures and analyses generally include either the epoch [-100, 500] ms (for display purposes) or [0, 250] ms (for statistical analysis) around stimulus onset. The latter time window is motivated by three arguments. First, auditory activity should not occur prior to stimulation. Second, time constants of the rodent brain are expected to be shorter than in humans (partly due to much shorter axonal connections). Empirical evidence seems to indicate that typical auditory evoked potentials are roughly 2 times slower in humans, where one typically expects MMN related processes until up to *500* ms post stimulus. Third, [0, 250] ms roughly corresponds to the time window where the *noPharma* group shows a clear ERP activity.

QUALITATIVE ANALYSIS

Visual inspection of the ERP and difference curves across rats and pharmacological conditions. A preliminary inspection will be used to test, whether the assumed time window between 0 and 250 ms captures most of the ERPs in the Pharma data as well.

QUANTITATIVE ANALYSIS

All statistical analyses concern the time window [0, 250] ms (post stimulus presentation).

Summaries of parameters describing the data such as *number of trials, number of artefacts, signal range* are computed.

First Level Statistics

Two Way ANOVA on the single trial ERPs. Factors are

tone = {'Standard', 'Deviant'}

drug = {'2mgScopo', '1mgScopo', 'Vehicle', '3mgPilo', '6mgPilo'}

Individual ANOVAs are calculated for each time point and electrode, and false discovery rate (FDR) corrected.

Second Level Statistics

- Mixed effects model on the second level using a fixed effect of *tone* and *drug* and a random effect of *subject (rat)*.

Possible Extension

- Conjunction analysis (logical AND on the First-Level statistics). Note: The true conjunction might be overly conservative as it requires statistical significance in every single rat.
- Global null (conjunction) analysis
- Permutation based correction of statistics on the time series

DYNAMIC CAUSAL MODELING (DCM)

Exclusion of Rats

See CLASSICAL ANALYSIS -> EXCLUSION OF RATS

Neuronal Model

We use the standard structure of a canonical microcircuit model (CMC). All changes to the default model in the form of changed scaling parameters / priors are informed from the analysis of the *noPharma* dataset only.

Changes to the default model

Input Design

Due to the much shorter time constants, including the time between stimulus and start of the stimulus related neural response, we reparametrized the input as a gamma pulse, where the parameters are also estimated in the model inversion.

Integrator

We use the in house developed delayed Euler integration scheme. We found that it models a system described by delay differential equations more realistically for a wider range of delays than the (default) delayed Ozaki integration scheme implemented in SPM12.

Alternatively, we could use a delayed RK4 integration scheme. Preliminary analyses have indicated that the delayed Euler integration scheme performs stably for this system.

The step size is set to the sampling step size (1 ms). In case of any evidence of integrator failure, we will opt to reduce the integration step size.

Condition specific Effects (Modulation)

Condition specific effects (entries in the B matrix) are programmed to only change connections that are explicitly present in the system.

MULTI-START APPROACH

We use a multi-start approach for the standard gradient based variational Bayes (VB) optimization of the model, to overcome at least some of the expected (and shown in preliminary analyses) local maxima. We sampled 99 starting values (these were kept identical across rats) from the prior. Additionally, one of the starting values will correspond to prior means (This corresponds to the default starting value for the gradient ascent in SPM12). The parameters shaping the input and the modulation parameters are not included in the multi-start procedure. The same starting values are used for all models.

Note: We use the terminology *'best solution'* for the set of starting values that led to the highest negative free energy after inversion.

PRIORS

The priors on the parameters are informed from the multi-start analysis performed on the *noPharma* dataset. Whenever we refer to the *default priors*, we mean their *default values* as implemented in *spm_dcm_neural_priors.m,* and *spm_L_priors.m (SPM12, 6906).* Importantly, all data used to define the priors for the *pharmacological datasets* stems from the *noPharma* dataset assuring that no circular inference was possible.

The procedure to define the prior was as follows:

1. Inversion of grand average of the *noPharma* dataset using the multi-start approach
2. Define a new prior mean pE as the average over posterior means (over models and hemispheres) of the best inversions. The modulation prior mean is kept at the default value of 0.
   Formally, this is equivalent to the expectation of the Bayesian Model Averaging (BMA) posterior under equal posterior model probability:

$$pE = E[p(\theta|y)] = E[\sum_{m=1}^{M} p(\theta|y,m)p(m|y)] = \frac{1}{M}\sum_{m=1}^{M} E[p(\theta|y,m)]$$

Here, we use y as a short hand notation for the data in both hemispheres. We use the constraint of equal posterior model probability to avoid biasing the average toward the results from the winning model(s).

For the prior variance pC, we use the same values as in step 1.

1. Inversion of the single rats from the *noPharma* group for both hemispheres, using the prior defined in step 2.
2. Computing the prior means for the pharmacological dataset: For the prior means, we use the same argument as in step 2, but compute the BMA expectation also over subjects, hence y represents the data of all subjects and hemispheres. For the prior variance, we now use the variance over MAP estimates, i.e.

$$pC = var[E[p(\theta|y, m)]]$$

where the variance is computed over subjects, models and hemispheres. Naturally, the prior covariance for the modulation parameters is set according to the model structure, and to default values.

Remark 1: We considered the options of using model specific prior means in step 4. The results on the *noPharma* data however indicate that the posterior estimates are very consistent across models, and the model specific means almost always fell within one standard deviation of the model averaged means.

Remark 2: We also considered using the same prior variance as in step 1. However, both option would lead to values of similar scale, so we don't expect either of the options to perform differently.

LEADFIELD

We only consider the two pyramidal cell populations of the CMC to contribute to the signal. The Leadfield is reparametrized, to assume an equal contribution of both sources to the signal, and has an additional parameter modelling an additional gain for one of the regions. In principle, we would assume the two sources having the same gain factor. However, qualitative analyses show evidence that in some situations, one electrode shows higher amplitude, most likely due to better conductance of the electrode. Therefore, we fit both parameters to account for this variation.

NOISE

Hyperpriors for the noise estimate are set very tightly in SPM, assuming that the model will be able to fit the data highly accurately and thus forcing good fits. We relaxed this assumption and changed the noise prior (precision of observation noise) to accommodate more uncertainty in the precision estimate. In particular, we decreased the mean (to 4) and increased the variance (to 0.5) of the prior noise precision.

Scaling Parameters

Whenever we talk about an in- or decrease of a scaling parameter, it's meant in comparison with their default value as defined in *spm_fx_cmc (SPM12, 6906).* Note that changing the scaling factors leads to an effective change in prior mean and variance.

We decreased the scaling factor of the between region conductance delay to 1 *ms*. The between region conductance delays are still optimized in the inversion.

We increased the scaling factor of the forward connections in order to assure sufficient activation in the second region (lPAF and rPAF). The forward parameter of the the A matrix is still optimized.

Design Specification

We will now discuss three possible implementations (design matrices), to answer the three research question (see section *Research Questions).*

A general comment on how designs are specified in DCM for EEG.

Specifying the design matrix in DCM for EEG is similar to specifying a 2nd level GLM design matrix, where the coefficients β_j are replaced by the *trial effect matrices* $B\{j\} \in M(n_r, n_r)$,where n_r are the number of regions, and M stands for a matrix of the respective dimensions. The design matrix X together with $B = [B\{1\}, \dots, B\{n_{effects}\}]$then specify, how the connectivity matrix A_i in condition i comes about, i.e.

$$A_i = A + \sum_{j=1}^{n_{effects}} X_{ij} B\{j\}.$$

Important: This is bound for confusion, because the design matrix in the GUI will be specified with conditions in columns, and effect in rows, but the matrix will be transposed in the code. **Here, we will always use the notation with effects in columns, and conditions in rows**.

DESIGN 1: CLASSICAL APPROACH

Here, we propose inverting five individual DCMs per rat, for the five different pharmacological conditions. Each of the inversions will include the two conditions standard and deviant, where we allow for all gains / connections to be modulated by condition individually. This will amount to $2^4 = 16$ models to be inverted per rat **and** pharmacological conditions. As a consequence, **all parameters may vary across drug condition**.

The design matrix will be coded as

$$X = [0,1]^T.$$

The three questions may be answered as follows for the different ways of pooling[43]

1. Can we identify a drug dependent change in connection and/or gain parameters:
    a. 5x2 ANOVA on all connectivity parameters from winning model or BMA
    b. Difference in winning model for pooling
2. Can we identify characteristics in the dose response curve
    a. Taking the parameter estimates from winning model or BMA, and testing dose response curve for linear / quadratic shapes
3. Can we predict the pharmacological condition for left out rats
    a. classification on the parameter estimates (SVM)

Advantages of this approach:

- Simple and used in similar studies
- No assumptions about which parameters are fixed across drug conditions

Disadvantages of this approach:

- ANOVA has been indicated to be much less sensitive than BMS
- Design not optimized to answer the very particular questions

[43] Pooling refers to the different ways how one could pool over drugs: *dosage, antagonists vs vehicle vs agonists, vehicle vs drug*

- Stability wrt to the possible local minima and resulting posterior estimate

DESIGN 2: ESTIMATION OF ALL PHARMACOLOGICAL LEVELS IN A SINGLE DESIGN

Here, we concatenate the time series of all pharmacological levels (5) into a single time series with 10 conditions, coded in the following order

[std_2mg_Scopo, dev_2mg_Scopo, std_1mg_Scopo, dev_1mg_Scopo, std_Vehicle, dev_Vehicle, std_3mg_Pilo, dev_3mg_Pilo, std_6mg_Pilo, dev_6mg_Pilo]

We then set up 10 modulation matrices

$$B_{tone}^{drug},$$

with

$$tone = \{std, dev\}$$

and

$$drug = \{2mgScopo, 1mgScopo, Vehicle, 3mgPilo, 6mgPilo\},$$

where B will also be indexed by which connections are allowed to be modulated with the same $2^4 = 16$ models. Thus 16 models will be inverted per rat **but not anymore pharmacological levels**.

The design matrix will be coded as

$$X = eye(10).$$

Put simply, the connectivity matrix for drug d and tone T will be defined as

$$A_T^d = A + B_T^d.$$

The three questions may be answered as follows for different ways of pooling:

1. Can we identify a drug dependent change in connection and/or gain parameters:
   a. 5 x 2 ANOVA on connectivity parameters from winning model or BMA. Effective parameter estimates will need to be reconstructed properly.
2. Can we identify characteristics in the dose response curve
   a. Taking the parameter estimates from winning model or BMA, and testing dose response curve for linear / quadratic shapes
3. Can we predict the pharmacological condition for left out rats

a. classification on the parameter estimates (SVM)

Advantages of this approach:

- Parameters can be fixed across different drug conditions
- Fewer models need to be estimated

Disadvantages of this approach:

- ANOVA has been indicated to be much less sensitive than BMS
- Design not optimized to answer the very particular questions
- Assumptions about fixing certain model parameters across drug levels, i.e. Baseline A matrix is the same for all drug levels. Thus certain models are not part of the model space.
- Dealing with leadfield changes over time
- Dealing with noise level over pharmacological conditions

DESIGN 3: SEPARATING MMN-SPECIFIC AND MMN-UNSPECIFIC EFFECTS

Here, we concatenate the time series of all pharmacological levels (5) into a single time series with 10 conditions, coded in the following order

[std_2mg_Scopo, dev_2mg_Scopo, std_1mg_Scopo, dev_1mg_Scopo, std_Vehicle, dev_Vehicle, std_3mg_Pilo, dev_3mg_Pilo, std_6mg_Pilo, dev_6mg_Pilo]

We distinguish between two kind of modulation

- **MMN specific Modulation:** Here, we implement the same $2^4 = 16$ of models as previously introduced.
- **MMN unspecific Modulation:** Alternatively, we implement one effect, where the presence of a drug changes the overall connectivity in the network, i.e. the baseline A matrix (as opposed to design 2).

$$B_{u.s.} = \begin{pmatrix} b_{A1} & b_B \\ b_F & b_{PAF} \end{pmatrix}$$

We consider only the full effect matrix, where all parameters can be changed.

Both modulations can be:

- on / off

- fixed across pharmacological levels or individual across pharmacological level

Giving rise to **4 Families** encoded by different design matrices, leading to

$$16 \times 4 = 64$$

models.

Put simply, the connectivity matrix for drug d and family F will be defined as

$$A_{std}^{d} = A + B_{u.s.}^{F}$$

$$A_{dev}^{d} = A + B_{u.s.}^{F} + B_{MMN}^{F}$$

where the superscript F codes for the family, and thus whether the parameters are allowed to change across pharmacological levels, and whether it is even present.

The three questions may be answered as follows for different ways of pooling:

1. Can we identify a drug dependent change in connection and/or gain parameters:
    a. 5 x 2 ANOVA on connectivity parameters from winning model or BMA. Effective parameter estimates will need to be reconstructed properly.
    b. Model comparison on the model families described above testing for:
        i. an overall change in the ERP with no drug induced change in the MMN (Family 1)
        ii. an overall change in the ERP with drug induced change in the MMN (Family 2)
        iii. no overall change in the ERP with drug induced change in the MMN (Family 3)
        iv. no overall change in the ERP with no drug induced change in the MMN (Family 4)
2. Can we identify characteristics in the dose response curve
    a. Taking the parameter estimates from winning model or BMA, and testing dose response curve for linear / quadratic shapes
3. Can we predict the pharmacological condition for left out rats
    a. classification on the parameter estimates (SVM)

Advantages of this approach:

- Parameters can be fixed across different drug conditions
- Fewer models need to be estimated

- Important questions in 1 can be answered through model selection

Disadvantages of this approach:

- Dealing with leadfield changes over time
- Dealing with noise level over pharmacological conditions
- Difficult to communicate

CONCLUSION OF THE DESIGN DISCUSSION

In a first approach, Design 1 will be implemented in standard fashion. However, for scientific interest, we also consider Design 3, to get a better understanding of the capability of the inversion procedures for such an approach. There, we will allow the leadfield parameters to change over sessions, to account for effect of overall electrode conductance change, etc.

POSSIBLE EXTENSIONS

Some things that might be considered in the process of the analysis.

- Improved noise model (assuming a noise model that considers already filtered noise).
- Normalizing all ERPs by a range normalization to circumvent the problem with better or worse conducting electrodes.
- A priori fixing certain parameters (or reparametrization) after the completed preliminary analysis in order to reduce complexity of the model.

APPENDIX B| ANALYSIS PLAN FOR THE PRSSI STUDY

Dario Schöbi, Jolanda Malamud, Sara Tomiello, Stefan Frässle, Jakob Heinzle, Klaas Enno Stephan, Sandra Iglesias

Study Arm: fMRI during Working Memory (WM) Task

AMENDMENT:

This amendment was written as a revision of the original analysis plans

prssi_analysisPlan_ver1.0.docx: Original analysis plan written prior to data acquisition.

prssi_analysisPlan_ver1.1.docx: Analysis plan written after a preliminary data analysis in the context of the master thesis by Jolanda Malamud (JM).

prssi_analysisPlan_ver2.0.docx: Analysis plan written as a starting point for the re-analysis of the Master Thesis (by JM). It outlines the plan for the classical analysis, and a following DCM analysis.

prssi_analysisPlan_ver3.0.docx: Build on the analysis plan (ver. 2.0), and discuss alternative ways of the DCM analysis in order to answer additional questions about the effect of difficulty, and/or correctness of the response on the fronto-parietal network.

prssi_analysisPlan_ver3.1.docx: Discusses the choice of an informed hyperprior for the DCM analysis based on the residuals of a simple GLM analysis of the task regressors on the extracted time series.

In this version of the analysis plan, we introduce an improvement for the DCM analysis.

When inverting the DCMs according to the original plan, the DCMs failed to explain a reasonable amount of variance even under the multistart approach. They flat-lined in approx. half of the subjects even for the 'best models'. At this point, we suspected that this might be due to the strong correlation between the *sample* and *delay* regressor (71.6 %), which affected the extraction of time courses such that part of variance which could be attributed to *delay* regressors was removed when adjusting for the *sample* regressor (i.e. when the *sample* regressor was not included in the F-contrast for data extraction).

In brief: Strong correlations make it difficult to attribute variance uniquely to either regressor. This, in turn, could affect the shape and strength of the delay period related activity after adjusting for effects of interest in the time series extraction, rendering it difficult to accurately model the driving input (*delay*) in the DCM. Additionally, the correlation between *sample* and *delay* regressors likely also affected our GLM analyses, given that we did not find the expected differential effect (*memory > control)* in PFC regions on the group level after regressing out the physiological confounds.

Therefore, we opt to increase the design related variance by extracting time series that contain both *sample* and *delay* regressors and model the effects of *sample* and *delay* explicitly in the DCM. For this, we also consider the combined effects of the two in the GLMs. To motivate this, we regress the GLMs containing *sample* and *delay* onto the extracted time series (see section on time series extraction for the definition of the ROI). These preliminary analyses have shown that using both regressors greatly increases the amount of general WM related activity in all regions of the DCM (compared to TSE procedure used in ver3.1). Note that including both *sample* and *delay* as inputs results in neural activity in the DCM that depends on both inputs. While modulatory effects will still be temporally confined to the *delay* time window this period might still contain residual activation from *sample* activity. **Table 20** summarizes explained variance (by a simple GLM) of the new and the previous design. There is a strong increase in the explained variance. It is to be expected that an increase in design related variance should also increase the stability of parameter estimation.

	left PAR	**right PAR**	**left PFC**	**right PFC**
previous design	2.45 ± 0.27 %	2.67 ± 0.40 %	1.77 ± 0.21 %	1.68 ± 0.23%
new design	7.59 ± 0.72 %	9.26 ± 0.91 %	3.03 ± 0.35 %	3.28 ± 0.35 %

Table 20| Average (over subjects) and standard deviation of explained variance of a GLM fitted onto the extracted time series

The modified strategy for extracting the data and DCM is as follows. First, we identify regions of interest based on the confound-cleaned data, as this is the data used for time series extraction. Second, we formally identify the ROI on the second level by using only regions that show main effect ***and*** differential effect activity, which we consider crucial to have 'enough' driving input into the DCMs. Third, due to the high correlation between the regressors *sample* and *delay*, we consider the effect of both in the second level statistics, adjust for both during the time series extraction and model *sample* and *delay* explicitly as independent driving inputs to the DCMs.

CHANGELOG

1. Use the physiological confound cleaned data for the second level statistics and ROI identification.
2. Look at the combined effects of *sample* and *delay,* and model both as independent driving input in the DCM analysis.
3. Some reformatting: We have integrated additional GLMs directly in the analysis plan. These were introduced in revision 3.1 of the analysis plan, but were mentioned at the end. This change improves readability of the analysis plan.

PROJECT IN BRIEF

The goals of this project are twofold. First, we have recently designed a novel delayed match to sample working memory task that is being/has been used in a number of EEG studies. In this fMRI study we would like to determine the areas that are active during this new working memory (WM) paradigm. This information can then be used as spatial priors for source reconstruction analyses in those EEG studies which used the same WM paradigm.

Furthermore, we are interested in the effective connectivity between regions involved in this specific WM task, in particular in the fronto-parietal network. This is of particular interest for our paradigm, which is designed to minimize between-subject variance in performance. Therefore, we model the activity in the brain regions involved in WM with a Dynamic Causal Model (DCM). This then allows us to answer the question, whether changes in WM related activity is a purely local effect, driven by extrinsic connections between regions, or both.

As previously mentioned, the identification of these source priors is the main purpose of this study. There are two additional questions that can be addressed, although the design is not optimized for them. First, is there a difference in delay related activity preceding correct and incorrect responses? Second, is the performance (i.e. average difficulty) score related to the delay related differential activation?

48 healthy, male volunteers will be scanned using a 3T Philipps Ingenia Scanner at the IBT lab.

This analysis plan is established in order to specify the different steps of the fMRI analysis in advance.

BEHAVIORAL DATA:

Criteria for exclusion of subjects (or / and):

- 25% missing trials
- 40% wrong responses

FMRI PREPROCESSING:

1. Transforming .rec files into .nii files using rec2nifti.pl
2. Reorienting the origin of the functional and the structural images to the anterior commissure (AC)
3. Pre-processing of fMRI data
    a) Slice-timing correction (STC) to align all voxel time series to acquisition time of the 16th slice
    b) Realignment and Unwarping
    c) Segmentation of structural and functional images
    d) Image Calculator calculate mean image of structural segmented images
    e) Co-registration
        - Interpolate functional images into structural image

f) Normalization of functional images

g) Smoothing

- Gaussian Kernel: 6x6x6 mm

h) Normalization of structural images

4. PhysIO Toolbox to generate physiological noise regressors for the memory and control condition. Note that optional PhysIO pipelines have been set up in close collaboration with Lars Kasper (LK), if there were errors in the acquisition of the physiological measures. This is documented in *prssi_physIO.m* and the *Wiki* accompanying code documentation.

CHANGES TO THE DEFAULT FMRI PREPROCESSING:

After a first analysis by JM (see *prssi_analysisPlan_ver1.0*), some changes were made to the default settings for preprocessing, to improve the quality of the functional images. The changes are as follows, but can also be seen in the documented code accompanying the analysis:

1. Realignment and Unwarping
    a. Estimation Options
        i. Quality: 1
        ii. Num Passes: Register to mean
        iii. Interpolation: 7th Degree B-Spline
        iv. Wrapping: Wrap Y
    b. Unwarp Estimation Options
        i. Number of interactions: 2
    c. Unwarp Reslicing Options
        i. Interpolation: 7th Degree B-Spline
        ii. Wrapping: Wrap Y
2. Segmentation of structural images
    a. Data
        i. Save Bias Corrected
    b. Tissues
        i. Native Tissue for TPM file 6: Native Space
    c. Warping & MRF

        i. Deformation Fields: Forward

3. Image Calculator calculate mean image of structural segmented images
    a. Options
        i. Interpolation: 7th Degree Sinc
        ii. Data Type: FLOAT32 – single prec. Float
4. Segmentation of functional images
    a. Data
        i. Save Field and Corrected
    b. Tissues
        i. For all TPMs: Native Tissue: Native + Dartel Imported
        ii. For all TPMs: Warped Tissue: Modulated + Unmodulated
    c. Warping & MRF
        i. Deformation Fields: Forward
5. Image Calculator calculate mean image of functional segmented images
    a. Options
        i. Interpolation: 7th Degree Sinc
6. Co-registration: interpolation and smoothing of functional images into structural images (Coregister (Estimate))
7. Normalisation of structural images in MNI space (Normalise Write)
    a. Writing Options
        i. Voxel sizes: 1 1 1
        ii. Interpolation: 7th Degree B-Spline
8. Normalisation of functional images in MNI space (Normalise Write)
    a. Writing Options
        i. Interpolation: 7th Degree B-Spline
9. Normalisation of structural scalp stripped images in MNI space (Normalise Write)
    a. Writing Options
        i. Voxel sizes: 1 1 1
        ii. Interpolation: 7th Degree B-Spline
10. Smoothing
    a. FWHM: 6 6 6

GLM ANALYSES

We consider two designs (design matrices) in GLM analyses to investigate the effects the delay period related activity, the trial by trial difficulty changes and the effect of correctness. Designs concerns analyses, where we address additional research questions, but not primary research goals. Therefore, all primary research questions and the basis for the DCM modeling are based on Design 1.

DESIGN 1: FIRST LEVEL

REGRESSORS

Session Memory:

- Delay Memory (duration 4.0s)
- Sample Memory (duration: 0s)
- Probe Memory (duration: 0s)
- multiple regressors (physiological confounds)

Session Control:

- Delay Control (duration 4.0s)
- Sample Control (duration: 0s)
- Probe Control (duration: 0s)
- multiple regressors (physiological confounds)

CONTRASTS:

Note: zeros represent adjustment for variable matrix dimensions due to a variable number of confounds.

- con_0001
    - Delay Memory > Delay Control
    - c = [1, 0, ..., 0, -1]
- con_0002
    - Delay Memory < Delay Control
    - c = [-1, 0, ..., 0, 1]

- con_0003
  - Average Delay
  - c = [1, 0, ..., 0, 1]
- con_0004
  - Negative Average Delay
  - c = [-1, 0, ..., 0, -1]
- con_0005
  - Memory: Sample + Delay > Control: Sample + Delay
  - c = [1, -1, 0, ..., 0, 1, -1]
- con_0006
  - Average: Sample + Delay
  - c = [1, 1, 0, ..., 0, 1, 1]

DESIGN 1: SECOND LEVEL

CONTRASTS OF INTEREST:

- con_0001
  - con_0005 (Masked by con_0006)
  - correlation between con_0001 and subject specific average difficulty (in the memory condition)
  - correlation between con_0001 and subject specific average accuracy (in the memory condition)

  Significance is tested at $p<0.05$ family-wise error (FWE) corrected at cluster level (cluster defining threshold: $p < 0.001$ uncorrected)

DESIGN 3: FIRST LEVEL

General Comment: Although it is the second design which we consider, we name this 'Design 3' to be consistent with the code.

REGRESSORS

Session Memory:

- Delay Memory preceding correct responses (duration 4.0s)
- Delay Memory preceding incorrect responses (duration 4.0s)
- Sample Memory (duration: 0s)
- Probe Memory (duration: 0s)
- multiple regressors (physiological confounds)

Session Control:

- Delay Control preceding correct responses (duration 4.0s)
- Delay Control preceding incorrect responses (duration 4.0s)
- Sample Control (duration: 0s)
- Probe Control (duration: 0s)
- multiple regressors (physiological confounds)

CONTRAST:

Note: zeros represent adjustment for variable matrix dimensions due to a variable number of confounds.

- con_0001
  - Delay: Memory > Delay Control
  - c = [1, 1, 0, …, 0, -1, -1]
- con_0002
  - Delay Memory correct > Delay Control correct
  - c = [1, 0, ..., 0, -1, 0]
- con_0003
  - Correct > Incorrect
  - c = [1, -1, 0, ..., 0, 1, -1]
- con_0004
  - Delay: Memory correct > Delay: Memory incorrect
  - c = [1, -1]
- con_0005
  - Interaction Memory * Correctness
  - c = [1, -1, 0, ..., 0, -1, 1]

DESIGN 3: SECOND LEVEL

CONTRASTS OF INTEREST:

- con_0001
- con_0002
- con_0003
- con_0004
- con_0005

Significance is tested at $p < 0.05$ family-wise error (FWE) corrected at cluster level (cluster defining threshold: $p < 0.001$ uncorrected)

TIME SERIES EXTRACTION (TSE):

For the concatenation of the time series, we will have to regress out all nuisance regressors on the first level, and take the resulting residual images to extract the data to be modeled. This is conservative in the sense that the noise regressors can explain away potential variance also explained by the task.

DESIGN CONCATENATE: FIRST LEVEL

REGRESSORS

- Delay Memory (duration 4.0s)
- Delay Control (duration: 4.0s)
- Sample Memory (duration: 0s)
- Sample Control (duration: 0s)
- Probe Memory (duration: 0s)
- Probe Control (duration: 0s)

CONTRASTS:

- con_0001
  - Delay Memory > Delay Control

    - c = [1, -1]
- con_0002
    - Delay Memory < Delay Control
    - c = [-1, 1]
- con_0003
    - Average Delay
    - c = [1, 1]
- con_0004
    - Negative Average Delay
    - c = [-1, -1]
- con_0005
    - Memory: Sample + Delay > Control: Sample + Delay
    - c = [1, -1, 1, -1]
- con_0006
    - Average: Sample + Delay
    - c = [1, 1, 1, 1]
- con_0007 (F – contrast)
    - Delay
    - eye(2)
- con_0008 (F – contrast)
    - Sample+Delay
    - eye(4)

- Identification of the group level maxima from con_0005 > 0 (masked with con_0006 > 0, $p<0.001$ uncorrected)
- Identification of subject specific maxima within an 8mm sphere around the group maximum.
- Extraction of the single subject time series as the first eigenvariate in a sphere of 4mm around the subject specific maximum of con_0005 (without thresholding).

DYNAMIC CAUSAL MODELING (DCM)

NETWORK

Fronto-parietal Network (bilateral, 4 Regions)

MODEL SPACE:

- 4 x 4 factorial (16 models)
- Input Structure:
    - Sample (Driving Input)
    - Delay Period (Driving Input)
    - Working Memory (Modulatory Input)
- Driving Input only in parietal regions
- Factors:
    - Extrinsic Modulation:
        - None
        - Forward
        - Backward
        - Bidirectional
    - Self-Modulation
        - None
        - Parietal
        - Frontal
        - Both

MODEL INVERSION:

We consider two possible inversion schemes.

Primarily, we consider a hierarchical inversion approach, where we make use of the in-house developed toolbox HUGE, assuming one cluster. This is the empirical Bayes variant of HUGE.

If the default version of HUGE fails to provide a sufficiently good fit, we will opt for two alternatives:

Alternative 1: HUGE can be run using a multistart scheme. If there is evidence that the inversion suffers from local minima, we opt to run HUGE with a multistart.

Alternative 2: If the hierarchical approach itself seems to be not suitable for the dataset, we will go back to the classical approach, estimating the model for each subject individually, potentially as well under a multistart approach.

INFORMED HYPERPRIORS

The default hyperpriors in SPM12(6909) assume very high SNR (prior expected log(precision) = 6 corresponding to a noise variance of = exp(-6) = 0.0025), with high certainty (variance of prior = 1/128 = 0.0078). Reviewing the DCM inversions under this prior indicated problems induced by the extreme settings for this prior: First, the Free Energy Landscape becomes highly multimodal with very high differences in Free Energy. Second, the contribution of the noise term in the KL divergence becomes unreasonably high, potentially biasing the Model Selection in favour of more complex models. Third, the assumed SNR seems very far off the actual SNR (estimated as below).

In order to circumvent these problems we chose to inform the setting of the prior by the GLM. We performed a simple, linear regression between the (rescaled) time series modelled in the DCM and the 1st level GLM (task regressors) to inform the noise prior. Based on this regression, we set the prior variance of the noise to be the average (over subjects) residual variance of the regression and its variance to be the variance across subjects.

MODEL SELECTION

For HUGE, model selection is done at the group level and quantified in terms of Bayes Factors. For the classical approach of inverting DCMs subject per subject, we will employ a random effects (RFX) BMS procedure.

TESTING OF PARAMETERS

One primary question is, whether a connection is (consistently across subjects) enhanced / reduced during Memory. This effect is represented in the B-Matrix.

In the classical approach (without the use of HUGE), we test the posterior means (across subjects) in a classical t-test at $p < 0.05$ significance level. This could also be done in terms of a BMA mean, if there is no model that is clearly superior to the others.

In the HUGE framework, we can then test the significance of a parameter in a Bayesian way, i.e. by computing the mass of the posterior distribution (of the group parameter) below and above a threshold corresponding to the absence of an effect.

For all frequentists statistical tests, we use Bonferroni or false discovery rate (FDR) correction to correct for multiple comparisons.

SOFTWARE INFORMATION

All analyses were conducted using the following software versions

- SPM12, ver. 6906
- MATLAB 2017b (local platform)
- MATLAB 2017b (Euler)
- R, v3.5.2

REFERENCES

Abi-Dargham, A., O. Mawlawi, I. Lombardo, R. Gil, D. Martinez, Y. Huang, D.-R. Hwang, J. Keilp, L. Kochan and R. Van Heertum (2002). "Prefrontal dopamine D1 receptors and working memory in schizophrenia." Journal of Neuroscience **22**(9): 3708-3719.

Akaike, H. (1974). "A new look at the statistical model identification." IEEE Transactions on Automatic Control **19**(6): 716-723.

Allwein, E. L., R. E. Schapire and Y. Singer (2000). "Reducing multiclass to binary: A unifying approach for margin classifiers." Journal of machine learning research **1**(Dec): 113-141.

Aponte, E. A., S. Raman, S. Frässle, J. Heinzle, W. D. Penny and K. E. Stephan (2018). "Thermodynamic integration for dynamic causal models." bioRxiv: 471417.

Aponte, E. A., S. Raman, B. Sengupta, W. D. Penny, K. E. Stephan and J. Heinzle (2016). "mpdcm: A toolbox for massively parallel dynamic causal modeling." Journal of Neuroscience Methods **257**: 7-16.

Aponte, E. A., D. Schöbi, K. E. Stephan and J. Heinzle (2019). "Computational Dissociation of Dopaminergic and Cholinergic Effects on Action Selection and Inhibitory Control." Biological Psychiatry: Cognitive Neuroscience and Neuroimaging.

Baddeley, A. (1992). "Working memory." Science **255**(5044): 556-559.

Baddeley, A. D. and G. Hitch (1974). Working memory. Psychology of learning and motivation, Elsevier. **8:** 47-89.

Balachandran, B., T. Kalmár-Nagy and D. E. Gilsinn (2009). Delay differential equations, Springer.

Baldeweg, T. (2007). "ERP repetition effects and mismatch negativity generation: a predictive coding perspective." Journal of Psychophysiology **21**(3-4): 204-213.

Baldeweg, T., A. Klugman, J. Gruzelier and S. R. Hirsch (2004). "Mismatch negativity potentials and cognitive impairment in schizophrenia." Schizophrenia research **69**(2-3): 203-217.

Barch, D. M., T. S. Braver, L. E. Nystrom, S. D. Forman, D. C. Noll and J. D. Cohen (1997). "Dissociating working memory from task difficulty in human prefrontal cortex." Neuropsychologia **35**(10): 1373-1380.

Bastos, A. M., W. M. Usrey, R. A. Adams, G. R. Mangun, P. Fries and K. J. Friston (2012). "Canonical microcircuits for predictive coding." Neuron **76**(4): 695-711.

Beal, M. J. (2003). Variational algorithms for approximate Bayesian inference, university of London London.

Bellen, A. and M. Zennaro (2013). Numerical methods for delay differential equations, Oxford university press.

Bellman, R. (1961). "On the computational solution of differential-difference equations." Journal of Mathematical Analysis and Applications **2**(1): 108-110.

Benjamini, Y. and D. Yekutieli (2001). "The control of the false discovery rate in multiple testing under dependency." The annals of statistics **29**(4): 1165-1188.

Bishop, C. M. (2006). Pattern recognition and machine learning, Springer Science+ Business Media.

Brodersen, K. H., L. Deserno, F. Schlagenhauf, Z. Lin, W. D. Penny, J. M. Buhmann and K. E. Stephan (2014). "Dissecting psychiatric spectrum disorders by generative embedding." NeuroImage: Clinical **4**: 98-111.

Brodersen, K. H., T. M. Schofield, A. P. Leff, C. S. Ong, E. I. Lomakina, J. M. Buhmann and K. E. Stephan (2011). "Generative Embedding for Model-Based Classification of fMRI Data." PLOS Computational Biology **7**(6): e1002079.

Butcher, A. J., I. Torrecilla, K. W. Young, K. C. Kong, S. C. Mistry, A. R. Bottrill and A. B. Tobin (2009). "N-methyl-D-aspartate receptors mediate the phosphorylation and desensitization of muscarinic receptors in cerebellar granule neurons." Journal of Biological Chemistry **284**(25): 17147-17156.

Callicott, J. H., N. F. Ramsey, K. Tallent, A. Bertolino, M. B. Knable, R. Coppola, T. Goldberg, P. Van Gelderen, V. S. Mattay and J. A. Frank (1998). "Functional magnetic resonance imaging brain mapping in psychiatry: methodological issues illustrated in a study of working memory in schizophrenia." Neuropsychopharmacology **18**(3): 186-196.

Carlsson, A. (1988). "The current status of the dopamine hypothesis of schizophrenia." Neuropsychopharmacology.

Caruana, R., S. Lawrence and C. L. Giles (2001). Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. Advances in neural information processing systems.

Chen, C.-C., J.-C. Kuo and W.-J. Wang (2019). "Distinguishing the Visual Working Memory Training and Practice Effects by the Effective Connectivity During n-back Tasks: A DCM of ERP Study." Frontiers in Behavioral Neuroscience **13**(84).

Chumbley, J. R., K. J. Friston, T. Fearn and S. J. Kiebel (2007). "A Metropolis–Hastings algorithm for dynamic causal models." NeuroImage **38**(3): 478-487.

Cohen, J. D. and D. Servan-Schreiber (1992). "Context, cortex, and dopamine: a connectionist approach to behavior and biology in schizophrenia." Psychological review **99**(1): 45.

Cools, R. and M. D'Esposito (2011). "Inverted-U–shaped dopamine actions on human working memory and cognitive control." Biological psychiatry **69**(12): e113-e125.

Courtney, S. M., L. G. Ungerleider, K. Keil and J. V. Haxby (1996). "Object and spatial visual working memory activate separate neural systems in human cortex." Cerebral cortex **6**(1): 39-49.

Cowan, N. (2016). Working Memory Capacity: Classic Edition, Routledge.

Curtis, C. E. and M. D'Esposito (2003). "Persistent activity in the prefrontal cortex during working memory." Trends in cognitive sciences **7**(9): 415-423.

D'Esposito, M. and B. R. Postle (2015). "The cognitive neuroscience of working memory." Annual review of psychology **66**: 115-142.

D'Ardenne, K., N. Eshel, J. Luka, A. Lenartowicz, L. E. Nystrom and J. D. Cohen (2012). "Role of prefrontal cortex and the midbrain dopamine system in working memory updating." Proceedings of the National Academy of Sciences **109**(49): 19900-19909.

D'Esposito, M., B. R. Postle and B. Rypma (2000). Prefrontal cortical contributions to working memory: evidence from event-related fMRI studies. Executive control and the frontal lobe: Current issues, Springer**:** 3-11.

Daniel, T. A., J. S. Katz and J. L. Robinson (2016). "Delayed match-to-sample in working memory: A BrainMap meta-analysis." Biological psychology **120**: 10-20.

Daunizeau, J., O. David and K. E. Stephan (2011). "Dynamic causal modelling: A critical review of the biophysical and statistical foundations." NeuroImage **58**(2): 312-322.

Dauphin, Y. N., R. Pascanu, Ç. Gülçehre, K. Cho, S. Ganguli and Y. Bengio (2014). Identifying and attacking the saddle point problem in high-dimensional non-convex optimization**:** 2933-2941.

David, O., S. J. Kiebel, L. M. Harrison, J. Mattout, J. M. Kilner and K. J. Friston (2006). "Dynamic causal modeling of evoked responses in EEG and MEG." Neuroimage **30**(4): 1255-1272.

Davis, K. L., R. S. Kahn, G. Ko and M. Davidson (1991). "Dopamine in schizophrenia: a review and reconceptualization." The American journal of psychiatry.

Dean, B., N. Thomas, C.-Y. Lai, W. J. Chen and E. Scarr (2015). "Changes in cholinergic and glutamatergic markers in the striatum from a sub-set of subjects with schizophrenia." Schizophrenia research **169**(1-3): 83-88.

Deserno, L., P. Sterzer, T. Wüstenberg, A. Heinz and F. Schlagenhauf (2012). "Reduced prefrontal-parietal effective connectivity and working memory deficits in schizophrenia." Journal of Neuroscience **32**(1): 12-20.

Di Maio, R., P. G. Mastroberardino, X. Hu, L. Montero and J. T. Greenamyre (2011). "Pilocapine alters NMDA receptor expression and function in hippocampal neurons: NADPH oxidase and ERK1/2 mechanisms." Neurobiology of Disease **42**(3): 482-495.

Dictionary, O. E. "overfitting, n.", Oxford University Press.

Díez, Á., S. Ranlund, D. Pinotsis, S. Calafato, M. Shaikh, M.-H. Hall, M. Walshe, Á. Nevado, K. J. Friston, R. A. Adams and E. Bramon (2017). "Abnormal frontoparietal synaptic gain mediating the P300 in patients with psychotic disorder and their unaffected relatives." Human brain mapping **38**(6): 3262-3276.

Dima, D., J. Jogia and S. Frangou (2014). "Dynamic causal modeling of load‐dependent modulation of effective connectivity within the verbal working memory network." Human brain mapping **35**(7): 3025-3035.

Durstewitz, D. and J. K. Seamans (2002). "The computational role of dopamine D1 receptors in working memory." Neural Networks **15**(4): 561-572.

Eickhoff, S. B., T. Paus, S. Caspers, M.-H. Grosbras, A. C. Evans, K. Zilles and K. Amunts (2007). "Assignment of functional activations to probabilistic cytoarchitectonic areas revisited." Neuroimage **36**(3): 511-521.

Elsgolts, L. E. (1964). Qualitative methods in mathematical analysis, American Mathematical Soc.

Erickson, M. A., A. Ruffle and J. M. Gold (2016). "A Meta-Analysis of Mismatch Negativity in Schizophrenia: From Clinical Risk to Disease Specificity and Progression." Biol Psychiatry **79**(12): 980-987.

Erneux, T. (2009). Applied delay differential equations, Springer Science & Business Media.

Feldstein, A. and R. Goodman (1973). "Numerical solution of ordinary and retarded differntial equations with discontinuou derivatives." Numerische Mathematik **21**(1): 1-13.

Ferraina, S., M. Paré and R. H. Wurtz (2002). "Comparison of cortico-cortical and cortico-collicular signals for the generation of saccadic eye movements." Journal of neurophysiology **87**(2): 845-858.

Forbes, N., L. Carrick, A. McIntosh and S. Lawrie (2009). "Working memory in schizophrenia: a meta-analysis." Psychological medicine **39**(6): 889-905.

Frässle, S., K. E. Stephan, K. J. Friston, M. Steup, S. Krach, F. M. Paulus and A. Jansen (2015). "Test-retest reliability of dynamic causal modeling for fMRI." Neuroimage **117**: 56-66.

Frässle, S., Y. Yao, D. Schöbi, E. A. Aponte, J. Heinzle and K. E. Stephan (2018). "Generative models for clinical applications in computational psychiatry." Wiley Interdisciplinary Reviews: Cognitive Science **9**(3): e1460.

Friston, K., J. Kilner and L. Harrison (2006). "A free energy principle for the brain." Journal of Physiology-Paris **100**(1-3): 70-87.

Friston, K., J. Mattout, N. Trujillo-Barreto, J. Ashburner and W. Penny (2007). "Variational free energy and the Laplace approximation." Neuroimage **34**(1): 220-234.

Friston, K., R. Moran and A. K. Seth (2013). "Analysing connectivity with Granger causality and dynamic causal modelling." Current opinion in neurobiology **23**(2): 172-178.

Friston, K., P. Zeidman and V. Litvak (2015). "Empirical Bayes for DCM: a group inversion scheme." Frontiers in systems neuroscience **9**: 164.

Friston, K. J. (1998). "The disconnection hypothesis." Schizophrenia research **30**(2): 115-125.

Friston, K. J., L. Harrison and W. Penny (2003). "Dynamic causal modelling." NeuroImage **19**(4): 1273-1302.

Friston, K. J., V. Litvak, A. Oswal, A. Razi, K. E. Stephan, B. C. Van Wijk, G. Ziegler and P. Zeidman (2016). "Bayesian model reduction and empirical Bayes for group (DCM) studies." Neuroimage **128**: 413-431.

Friston, K. J., A. Mechelli, R. Turner and C. J. Price (2000). "Nonlinear responses in fMRI: the Balloon model, Volterra kernels, and other hemodynamics." Neuroimage **12**(4): 466-477.

Garrido, M. I., K. J. Friston, S. J. Kiebel, K. E. Stephan, T. Baldeweg and J. M. Kilner (2008). "The functional anatomy of the MMN: a DCM study of the roving paradigm." Neuroimage **42**(2): 936-944.

Garrido, M. I., J. M. Kilner, S. J. Kiebel, K. E. Stephan and K. J. Friston (2007). "Dynamic causal modelling of evoked potentials: A reproducibility study." NeuroImage **36**(3): 571-580.

Garrido, M. I., J. M. Kilner, K. E. Stephan and K. J. Friston (2009). "The mismatch negativity: a review of underlying mechanisms." Clin Neurophysiol **120**(3): 453-463.

Gibbons, A. S., E. Scarr, S. Boer, T. Money, W.-J. Jeon, C. Felder and B. Dean (2013). "Widespread decreases in cortical muscarinic receptors in a subset of people with schizophrenia." International Journal of Neuropsychopharmacology **16**(1): 37-46.

Gold, J. M., D. M. Barch, L. M. Feuerstahler, C. S. Carter, A. W. MacDonald III, J. D. Ragland, S. M. Silverstein, M. E. Strauss and S. J. Luck (2018). "Working memory impairment across psychotic disorders." Schizophrenia bulletin **45**(4): 804-812.

Goldman-Rakic, P. S. (1994). "Working memory dysfunction in schizophrenia." The Frontal Lobes and Neuropsychiatric Illness. Washington, DC: 71-82.

Grishin, A. A., P. Benquet and U. Gerber (2005). "Muscarinic receptor stimulation reduces NMDA responses in CA3 hippocampal pyramidal cells via Ca2+-dependent activation of tyrosine phosphatase." Neuropharmacology **49**(3): 328-337.

Gutenkunst, R. N., J. J. Waterfall, F. P. Casey, K. S. Brown, C. R. Myers and J. P. Sethna (2007). "Universally sloppy parameter sensitivities in systems biology models." PLoS computational biology **3**(10).

Haenschel, C., T. Baldeweg, R. J. Croft, M. Whittington and J. Gruzelier (2000). "Gamma and beta frequency oscillations in response to novel auditory stimuli: a comparison of human electroencephalogram (EEG) data with in vitro models." Proceedings of the National Academy of Sciences **97**(13): 7645-7650.

Heinzel, S., R. C. Lorenz, Q.-L. Duong, M. A. Rapp and L. Deserno (2017). "Prefrontal-parietal effective connectivity during working memory in older adults." Neurobiology of aging **57**: 18-27.

Heinzle, J. and K. E. Stephan (2018). Chapter 5 - Dynamic Causal Modeling and Its Application to Psychiatric Disorders. Computational Psychiatry. A. Anticevic and J. D. Murray, Academic Press: 117-144.

Henderson, P., R. Islam, P. Bachman, J. Pineau, D. Precup and D. Meger (2018). "Deep Reinforcement Learning that Matters."

Henderson, P., R. Islam, P. Bachman, J. Pineau, D. Precup and D. Meger (2018). Deep reinforcement learning that matters. Thirty-Second AAAI Conference on Artificial Intelligence.

Higley, M. J. and M. R. Picciotto (2014). "Neuromodulation by acetylcholine: examples from schizophrenia and depression." Current opinion in neurobiology **29**: 88-95.

Howes, O. D. and S. Kapur (2009). "The dopamine hypothesis of schizophrenia: version III—the final common pathway." Schizophrenia bulletin **35**(3): 549-562.

Huys, Q. J. M., T. V. Maia and M. J. Frank (2016). "Computational psychiatry as a bridge from neuroscience to clinical applications." Nature Neuroscience **19**(3): 404-413.

Iglesias, S., S. Tomiello, M. Schneebeli and K. E. Stephan (2017). "Models of neuromodulation for computational psychiatry." Wiley Interdisciplinary Reviews: Cognitive Science **8**(3): e1420.

Jafarian, A., P. Zeidman, V. Litvak and K. Friston (2019). "Structure learning in coupled dynamical systems and dynamic causal modelling." Philosophical Transactions of the Royal Society A **377**(2160): 20190048.

Jansen, B. H. and V. G. Rit (1995). "Electroencephalogram and visual evoked potential generation in a mathematical model of coupled cortical columns." Biological Cybernetics **73**(4): 357-366.

Jeffreys, H. (1961). Theory of probability, Clarendon, Oxford.

Jin, C., P. Netrapalli and M. I. Jordan (2017). "Accelerated gradient descent escapes saddle points faster than gradient descent." arXiv preprint arXiv:1711.10456.

Jung, F. (2013). "Mismatch responses in the awake rat: Evidence from epidural recordings of auditory cortical fields." Universität zu Köln **Dissertation/Thesis**.

Jung, F., K. E. Stephan, H. Backes, R. Moran, M. Gramer, T. Kumagai, R. Graf, H. Endepols and M. Tittgemeyer (2013). "Mismatch responses in the awake rat: evidence from epidural recordings of auditory cortical fields." PLoS One **8**(4): e63203.

Jung, K., K. J. Friston, C. Pae, H. H. Choi, S. Tak, Y. K. Choi, B. Park, C. A. Park, C. Cheong and H. J. Park (2018). "Effective connectivity during working memory and resting states: A DCM study." Neuroimage **169**: 485-495.

Kahn, R. S., I. W. van Rossum, S. Leucht, P. McGuire, S. W. Lewis, M. Leboyer, C. Arango, P. Dazzan, R. Drake and S. Heres (2018). "Amisulpride and olanzapine followed by open-label treatment with clozapine in first-episode schizophrenia and schizophreniform disorder (OPTiMiSE): a three-phase switching study." The Lancet Psychiatry **5**(10): 797-807.

Kapur, S., A. G. Phillips and T. R. Insel (2012). "Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it?" Molecular psychiatry **17**(12): 1174.

Kapur, S. and G. Remington (2001). "Atypical Antipsychotics: New Directions and New Challenges in the Treatment of Schizophrenia." Annual Review of Medicine **52**(1): 503-517.

Kashyap, R., S. Bhattacharjee, W. Sommer and C. Zhou (2019). "Repetition priming effects for famous faces through dynamic causal modelling of latency-corrected event-related brain potentials." European Journal of Neuroscience **49**(10): 1330-1347.

Kasper, L., S. Bollmann, A. O. Diaconescu, C. Hutton, J. Heinzle, S. Iglesias, T. U. Hauser, M. Sebold, Z.-M. Manjaly and K. P. Pruessmann (2017). "The PhysIO toolbox for modeling physiological noise in fMRI data." Journal of neuroscience methods **276**: 56-72.

Kass, R. E. and A. E. Raftery (1995). "Bayes factors." Journal of the american statistical association **90**(430): 773-795.

Kay, S. R., A. Fiszbein and L. A. Opler (1987). "The positive and negative syndrome scale (PANSS) for schizophrenia." Schizophrenia bulletin **13**(2): 261-276.

Kiebel, S. J., M. I. Garrido, R. Moran, C. C. Chen and K. J. Friston (2009). "Dynamic causal modeling for EEG and MEG." Hum Brain Mapp **30**(6): 1866-1876.

Kowal, N. M., P. K. Ahring, V. W. Liao, D. C. Indurti, B. S. Harvey, S. M. O'Connor, M. Chebib, E. S. Olafsdottir and T. Balle (2018). "Galantamine is not a positive allosteric modulator of human α4β2 or α7 nicotinic acetylcholine receptors." British journal of pharmacology **175**(14): 2911-2925.

Krystal, John H. and Matthew W. State (2014). "Psychiatric Disorders: Diagnosis to Therapy." Cell **157**(1): 201-214.

Lara, A. H. and J. D. Wallis (2015). "The role of prefrontal cortex in working memory: a mini review." Frontiers in systems neuroscience **9**: 173.

Lee, J. and S. Park (2005). "Working memory impairments in schizophrenia: a meta-analysis." Journal of abnormal psychology **114**(4): 599.

Lemarechal, J. D., N. George and O. David (2018). "Comparison of two integration methods for dynamic causal modeling of electrophysiological data." Neuroimage **173**: 623-631.

Lencz, T., R. M. Bilder, E. Turkel, R. S. Goldman, D. Robinson, J. M. Kane and J. A. Lieberman (2003). "Impairments in perceptual competency and maintenance on a visual delayed match-to-sample test in first-episode schizophrenia." Archives of general psychiatry **60**(3): 238-243.

Leucht, S., A. Cipriani, L. Spineli, D. Mavridis, D. Örey, F. Richter, M. Samara, C. Barbui, R. R. Engel and J. R. Geddes (2013). "Comparative efficacy and tolerability of 15 antipsychotic drugs in schizophrenia: a multiple-treatments meta-analysis." The Lancet **382**(9896): 951-962.

Leucht, S., K. Komossa, C. Rummel-Kluge, C. Corves, H. Hunger, F. Schmid, C. A. Lobos, S. Schwarz and J. M. Davis (2009). "A Meta-Analysis of Head-to-Head Comparisons of Second-Generation Antipsychotics in the Treatment of Schizophrenia." Am J Psychiatry **166**(2): 152-163.

Leung, H.-C., D. Seelig and J. C. Gore (2004). "The effect of memory load on cortical activity in the spatial working memory circuit." Cognitive, Affective, & Behavioral Neuroscience **4**(4): 553-563.

Leung, S., R. Croft, V. Guille, K. Scholes-Balog, B. O'Neill, K. L. Phan and P. Nathan (2009). "Acute dopamine and/or serotonin depletion does not modulate mismatch negativity (MMN) in healthy human participants." Psychopharmacology **208**: 233-244.

Levenberg, K. (1944). "A method for the solution of certain non-linear problems in least squares." Quarterly of applied mathematics **2**(2): 164-168.

Lieder, F., J. Daunizeau, M. I. Garrido, K. J. Friston and K. E. Stephan (2013). "Modelling trial-by-trial changes in the mismatch negativity." PLoS computational biology **9**(2): e1002911.

Litvak, V., M. Garrido, P. Zeidman and K. Friston (2015). "Empirical Bayes for group (DCM) studies: a reproducibility study." Frontiers in human neuroscience **9**: 670.

Litvak, V., J. Mattout, S. Kiebel, C. Phillips, R. Henson, J. Kilner, G. Barnes, R. Oostenveld, J. Daunizeau, G. Flandin, W. Penny and K. Friston (2011). "EEG and MEG data analysis in SPM8." Comput Intell Neurosci **2011**: 852961.

Lomakina, E. I., S. Paliwal, A. O. Diaconescu, K. H. Brodersen, E. A. Aponte, J. M. Buhmann and K. E. Stephan (2015). "Inversion of hierarchical Bayesian models using Gaussian processes." Neuroimage **118**: 133-145.

Lopes, M. W., F. M. S. Soares, N. de Mello, J. C. Nunes, A. G. Cajado, D. de Brito, F. M. de Cordova, R. M. S. da Cunha, R. Walz and R. B. Leal (2013). "Time-dependent modulation of AMPA receptor phosphorylation and mRNA expression of NMDA receptors and glial glutamate transporters in the rat hippocampus and cerebral cortex in a pilocarpine model of epilepsy." Experimental Brain Research **226**(2): 153-163.

Ma, L., J. L. Steinberg, K. M. Hasan, P. A. Narayana, L. A. Kramer and F. G. Moeller (2012). "Working memory load modulation of parieto-frontal connections: Evidence from dynamic causal modeling." Human brain mapping **33**(8): 1850-1867.

Manoach, D. S. (2003). "Prefrontal cortex dysfunction during working memory performance in schizophrenia: reconciling discrepant findings." Schizophrenia research **60**(2-3): 285-298.

Marino, M. J., S. T. Rouse, A. I. Levey, L. T. Potter and P. J. Conn (1998). "Activation of the genetically defined m1 muscarinic receptor potentiates N-methyl-D-aspartate (NMDA) receptor currents in hippocampal pyramidal cells." Proceedings of the National Academy of Sciences **95**(19): 11465-11470.

McCormick, D. A. and D. A. Prince (1985). "Two types of muscarinic response to acetylcholine in mammalian cortical neurons." Proceedings of the National Academy of Sciences **82**(18): 6344-6348.

McCormick, D. A., Z. Wang and J. Huguenard (1993). "Neurotransmitter Control of Neocortical Neuronal Activity and Excitability." Cerebral Cortex **3**(5): 387-398.

Miller, G. A., E. Galanter and K. H. Pribram (1960). "Plans and the structure of behavior."

Miller, K., C. Price, M. Okun, H. Montijo and D. Bowers (2009). "Is the n-back task a valid neuropsychological measure for assessing working memory?" Archives of Clinical Neuropsychology **24**(7): 711-717.

Miller, R. (1975). "Distribution and properties of commissural and other neurons in cat sensorimotor cortex." Journal of Comparative Neurology **164**(3): 361-373.

Montague, P. R., R. J. Dolan, K. J. Friston and P. Dayan (2012). "Computational psychiatry." Trends in cognitive sciences **16**(1): 72-80.

Moran, R. J., P. Campo, M. Symmonds, K. E. Stephan, R. J. Dolan and K. J. Friston (2013). "Free energy, precision and learning: the role of cholinergic neuromodulation." Journal of Neuroscience **33**(19): 8227-8236.

Moran, R. J., M. W. Jones, A. J. Blockeel, R. A. Adams, K. E. Stephan and K. J. Friston (2015). "Losing control under ketamine: suppressed cortico-hippocampal drive following acute ketamine in rats." Neuropsychopharmacology **40**(2): 268-277.

Moran, R. J., F. Jung, T. Kumagai, H. Endepols, R. Graf, R. J. Dolan, K. J. Friston, K. E. Stephan and M. Tittgemeyer (2011). "Dynamic causal models and physiological inference: a validation study using isoflurane anaesthesia in rodents." PLoS One **6**(8): e22790.

Moran, R. J., D. A. Pinotsis and K. J. Friston (2013). "Neural masses and fields in dynamic causal modeling." Frontiers in computational neuroscience **7**: 57.

Moran, R. J., M. Symmonds, K. E. Stephan, K. J. Friston and R. J. Dolan (2011). "An in vivo assay of synaptic function mediating human cognition." Curr Biol **21**(15): 1320-1325.

Moran, R. J., M. Symmonds, K. E. Stephan, K. J. Friston and R. J. Dolan (2011). "An in vivo assay of synaptic function mediating human cognition." Current Biology **21**(15): 1320-1325.

Moriconi, R., K. Kumar and M. P. Deisenroth (2019). "High-Dimensional Bayesian Optimization with Manifold Gaussian Processes." arXiv preprint arXiv:1902.10675.

Murphy, B., A. Arnsten, P. Goldman-Rakic and R. Roth (1996). "Increased dopamine turnover in the prefrontal cortex impairs spatial working memory performance in rats and monkeys." Proceedings of the National Academy of Sciences **93**(3): 1325-1329.

Nagel, B. J., M. M. Herting, E. C. Maxwell, R. Bruno and D. Fair (2013). "Hemispheric lateralization of verbal and spatial working memory during adolescence." Brain and cognition **82**(1): 58-68.

Nielsen, J. D., K. H. Madsen, Z. Wang, Z. Liu, K. J. Friston and Y. Zhou (2017). "Working memory modulation of frontoparietal network connectivity in first-episode schizophrenia." Cerebral Cortex **27**(7): 3832-3841.

Oberauer, K. (2002). "Access to Information in Working Memory: Exploring the Focus of Attention." Learning, Memory **28**(3): 411-421.

Okubo, Y., T. Suhara, K. Suzuki, K. Kobayashi, O. Inoue, O. Terasaki, Y. Someya, T. Sassa, Y. Sudo and E. Matsushima (1997). "Decreased prefrontal dopamine D1 receptors in schizophrenia revealed by PET." Nature **385**(6617): 634.

Olesen, P. J., H. Westerberg and T. Klingberg (2004). "Increased prefrontal and parietal activity after training of working memory." Nature neuroscience **7**(1): 75.

Ostwald, D. and L. Starke (2016). "Probabilistic delay differential equation modeling of event-related potentials." Neuroimage **136**: 227-257.

Ozaki, T. (1992). "A bridge between nonlinear time series models and nonlinear stochastic dynamical systems: a local linearization approach." Statistica Sinica: 113-135.

Park, S. and D. C. Gooding (2014). "Working memory impairment as an endophenotypic marker of a schizophrenia diathesis." Schizophrenia Research: Cognition **1**(3): 127-136.

Pascanu, R., Y. N. Dauphin, S. Ganguli and Y. Bengio (2014). "On the saddle point problem for non-convex optimization." arXiv preprint arXiv:1405.4604.

Penny, W., J. Iglesias-Fuster, Y. T. Quiroz, F. J. Lopera and M. A. Bobes (2018). "Dynamic Causal Modeling of Preclinical Autosomal-Dominant Alzheimer's Disease." Journal of Alzheimer's disease : JAD **65**(3): 697-711.

Penny, W. and B. Sengupta (2016). "Annealed Importance Sampling for Neural Mass Models." PLoS Comput Biol **12**(3): e1004797.

Penny, W. D. (2012). "Comparing Dynamic Causal Models using AIC, BIC and Free Energy." NeuroImage **59**(1): 319-330.

Penny, W. D., K. E. Stephan, J. Daunizeau, M. J. Rosa, K. J. Friston, T. M. Schofield and A. P. Leff (2010). "Comparing families of dynamic causal models." PLoS Comput Biol **6**(3): e1000709.

Postle, B. R. (2006). "Working memory as an emergent property of the mind and brain." Neuroscience **139**(1): 23-38.

Raedler, T. J. (2007). "Comparison of the in-vivo muscarinic cholinergic receptor availability in patients treated with clozapine and olanzapine." International Journal of Neuropsychopharmacology **10**(2): 275-280.

Raedler, T. J., F. P. Bymaster, R. Tandon, D. Copolov and B. Dean (2007). "Towards a muscarinic hypothesis of schizophrenia." Mol Psychiatry **12**(3): 232-246.

Raedler, T. J., M. B. Knable, D. W. Jones, R. A. Urbina, J. G. Gorey, K. S. Lee, M. F. Egan, R. Coppola and D. R. Weinberger (2003). "In vivo determination of muscarinic acetylcholine receptor availability in schizophrenia." American Journal of Psychiatry **160**(1): 118-127.

Raftery, A. E. (1995). "Bayesian model selection in social research." Sociological methodology **25**: 111-164.

Ranlund, S., R. A. Adams, Á. Díez, M. Constante, A. Dutt, M. H. Hall, A. Maestro Carbayo, C. McDonald, S. Petrella and K. Schulze (2016). "Impaired prefrontal synaptic gain in people with psychosis and their relatives during the mismatch negativity." Human brain mapping **37**(1): 351-365.

Rasmussen, C. E. (2003). Gaussian processes in machine learning. Summer School on Machine Learning, Springer.

Rigoux, L., K. E. Stephan, K. J. Friston and J. Daunizeau (2014). "Bayesian model selection for group studies - revisited." Neuroimage **84**: 971-985.

Roffman, J. L., A. S. Tanner, H. Eryilmaz, A. Rodriguez-Thompson, N. J. Silverstein, N. F. Ho, A. Z. Nitenson, D. B. Chonde, D. N. Greve, A. Abi-Dargham, R. L. Buckner, D. S. Manoach, B. R. Rosen, J. M. Hooker and C. Catana (2016). "Dopamine D₁ signaling organizes network dynamics underlying working memory." Science Advances **2**(6): e1501672.

Roth, W. T., A. Pfefferbaum, A. F. Kelly, P. A. Berger and B. S. Kopell (1981). "Auditory event-related potentials in schizophrenia and depression." Psychiatry Research **4**(2): 199-212.

Rottschy, C., R. Langner, I. Dogan, K. Reetz, A. R. Laird, J. B. Schulz, P. T. Fox and S. B. Eickhoff (2012). "Modelling neural correlates of working memory: a coordinate-based meta-analysis." Neuroimage **60**(1): 830-846.

Samochocki, M., A. Höffle, A. Fehrenbacher, R. Jostock, J. Ludwig, C. Christner, M. Radina, M. Zerlin, C. Ullmer, E. F. R. Pereira, H. Lübbert, E. X. Albuquerque and A. Maelicke (2003). "Galantamine Is an Allosterically Potentiating Ligand of Neuronal Nicotinic but Not of Muscarinic Acetylcholine Receptors." Journal of Pharmacology and Experimental Therapeutics **305**(3): 1024-1036.

Sawaguchi, T. and P. S. Goldman-Rakic (1991). "D1 Dopamine Receptors in Prefrontal Cortex: Involvement in Working Memory." Science: 947-950.

Scarr, E., T. Cowie, S. Kanellakis, S. Sundram, C. Pantelis and B. Dean (2009). "Decreased cortical muscarinic receptors define a subgroup of subjects with schizophrenia." Molecular psychiatry **14**(11): 1017.

Scarr, E., J. Craig, M. Cairns, M. Seo, J. Galati, N. Beveridge, A. Gibbons, S. Juzva, B. Weinrich and M. Parkinson-Bates (2013). "Decreased cortical muscarinic M1 receptors in schizophrenia are associated with changes in gene promoter methylation, mRNA and gene targeting microRNA." Translational psychiatry **3**(2): e230.

Scarr, E. and B. Dean (2008). "Muscarinic receptors: do they have a role in the pathology and treatment of schizophrenia?" J Neurochem **107**(5): 1188-1195.

Scarr, E., S. Hopper, V. Vos, M. S. Seo, I. P. Everall, T. D. Aumann, G. Chana and B. Dean (2018). "Low levels of muscarinic M1 receptor–positive neurons in cortical layers III and V in Brodmann areas 9 and 17 from individuals with schizophrenia." Journal of psychiatry & neuroscience: JPN **43**(5): 338.

Schmidt, A., A. O. Diaconescu, M. Kometer, K. J. Friston, K. E. Stephan and F. X. Vollenweider (2012). "Modeling ketamine effects on synaptic plasticity during the mismatch negativity." Cerebral Cortex **23**(10): 2394-2406.

Schmidt, A., R. Smieskova, A. Simon, P. Allen, P. Fusar-Poli, P. K. McGuire, K. Bendfeldt, J. Aston, U. E. Lang and M. Walter (2014). "Abnormal effective connectivity and psychopathological symptoms in the psychosis high-risk state." Journal of psychiatry & neuroscience: JPN **39**(4): 239.

Schwarz, G. (1978). "Estimating the Dimension of a Model." Ann. Statist. **6**(2): 461-464.

Seeman, P. (1987). "Dopamine receptors and the dopamine hypothesis of schizophrenia." Synapse **1**(2): 133-152.

Sengupta, B., K. J. Friston and W. D. Penny (2016). "Gradient-based MCMC samplers for dynamic causal modelling." NeuroImage **125**: 1107-1118.

Shampine, L. F. and S. Thompson (2001). "Solving DDEs in Matlab." Applied Numerical Mathematics **37**(4): 441-458.

Shampine, L. F., S. Thompson and J. Kierzenka (2000). "Solving delay differential equations with dde23." URL http://www.runet.edu/~thompson/webddes/tutorial.pdf.

Shimegi, S., A. Kimura, A. Sato, C. Aoyama, R. Mizuyama, K. Tsunoda, F. Ueda, S. Araki, R. Goya and H. Sato (2016). "Cholinergic and serotonergic modulation of visual information processing in monkey V1." Journal of Physiology-Paris **110**(1): 44-51.

Shinoe, T., M. Matsui, M. M. Taketo and T. Manabe (2005). "Modulation of Synaptic Plasticity by Physiological Activation of M₁ Muscarinic Acetylcholine Receptors in the Mouse Hippocampus." The Journal of Neuroscience **25**(48): 11194-11200.

Shumway, R. H. and D. S. Stoffer (2017). Time series analysis and its applications: with R examples, Springer.

Sladky, R., K. J. Friston, J. Trostl, R. Cunnington, E. Moser and C. Windischberger (2011). "Slice-timing effects and their correction in functional MRI." Neuroimage **58**(2): 588-594.

Smith, H. L. (2011). An introduction to delay differential equations with applications to the life sciences, Springer New York.

Snoek, J., H. Larochelle and R. P. Adams (2012). Practical bayesian optimization of machine learning algorithms. Advances in neural information processing systems.

Stephan, K. E., D. R. Bach, P. C. Fletcher, J. Flint, M. J. Frank, K. J. Friston, A. Heinz, Q. J. Huys, M. J. Owen and E. B. Binder (2016). "Charting the landscape of priority problems in psychiatry, part 1: classification and diagnosis." The Lancet Psychiatry **3**(1): 77-83.

Stephan, K. E., T. Baldeweg and K. J. Friston (2006). "Synaptic plasticity and dysconnection in schizophrenia." Biological psychiatry **59**(10): 929-939.

Stephan, K. E., K. J. Friston and C. D. Frith (2009). "Dysconnection in schizophrenia: from abnormal synaptic plasticity to failures of self-monitoring." Schizophrenia bulletin **35**(3): 509-527.

Stephan, K. E. and C. Mathys (2014). "Computational approaches to psychiatry." Current opinion in neurobiology **25**: 85-92.

Stephan, K. E. and C. Mathys (2014). "Computational approaches to psychiatry." Curr Opin Neurobiol **25**: 85-92.

Stephan, K. E., W. D. Penny, J. Daunizeau, R. J. Moran and K. J. Friston (2009). "Bayesian model selection for group studies." Neuroimage **46**(4): 1004-1017.

Stephan, K. E., F. Schlagenhauf, Q. J. Huys, S. Raman, E. A. Aponte, K. H. Brodersen, L. Rigoux, R. J. Moran, J. Daunizeau and R. J. Dolan (2017). "Computational neuroimaging strategies for single patient predictions." Neuroimage **145**: 180-199.

Swadlow, H. A. (1990). "Efferent neurons and suspected interneurons in S-1 forelimb representation of the awake rabbit: receptive fields and axonal properties." Journal of Neurophysiology **63**(6): 1477-1498.

Swadlow, H. A. and S. G. Waxman (2012). "Axonal conduction delays." Scholarpedia **7**(6): 1451.

Tandon, R. and J. F. Greden (1989). "Cholinergic hyperactivity and negative schizophrenic symptoms: a model of cholinergic/dopaminergic interactions in schizophrenia." Archives of General Psychiatry **46**(8): 745-753.

Umbricht, D. and S. Krljes (2005). "Mismatch negativity in schizophrenia: a meta-analysis." Schizophr Res **76**(1): 1-23.

Van Essen, D. C., C. H. Anderson and D. J. Felleman (1992). "Information processing in the primate visual system: an integrated systems perspective." Science **255**(5043): 419-423.

Van Snellenberg, J. X., I. J. Torres and A. E. Thornton (2006). "Functional neuroimaging of working memory in schizophrenia: task performance as a moderating variable." Neuropsychology **20**(5): 497.

Varoquaux, G. (2018). "Cross-validation failure: small sample sizes lead to large error bars." Neuroimage **180**: 68-77.

Varoquaux, G., P. R. Raamana, D. A. Engemann, A. Hoyos-Idrobo, Y. Schwartz and B. Thirion (2017). "Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines." NeuroImage **145**: 166-179.

Verhaeghen, P., S. Geigerman, H. Yang, A. C. Montoya and D. Rahnev (2019). "Resolving Age-Related Differences in Working Memory: Equating Perception and Attention Makes Older Adults Remember as Well as Younger Adults." Experimental Aging Research **45**(2): 120-134.

Vijayraghavan, S., M. Wang, S. G. Birnbaum, G. V. Williams and A. F. Arnsten (2007). "Inverted-U dopamine D1 receptor actions on prefrontal neurons engaged in working memory." Nature neuroscience **10**(3): 376.

Walter, H., A. P. Wunderlich, M. Blankenhorn, S. Schäfer, R. Tomczak, M. Spitzer and G. Grön (2003). "No hypofrontality, but absence of prefrontal lateralization comparing verbal and spatial working memory in schizophrenia." Schizophrenia Research **61**(2-3): 175-184.

Wang, X.-J. and J. H. Krystal (2014). "Computational psychiatry." Neuron **84**(3): 638-654.

Ward, L. M. and P. E. Greenwood (2007). "1/f noise." Scholarpedia **2**(12): 1537.

Weiner, D. M., H. Y. Meltzer, I. Veinbergs, E. M. Donohue, T. A. Spalding, T. T. Smith, N. Mohell, S. C. Harvey, J. Lameh, N. Nash, K. E. Vanover, R. Olsson, K. Jayathilake, M. Lee, A. I. Levey, U. Hacksell, E. S. Burstein, R. E. Davis and M. R. Brann (2004). "The role of M1 muscarinic receptor agonism of N-desmethylclozapine in the unique clinical effects of clozapine." Psychopharmacology **177**(1): 207-216.

West, T. O., L. O. Berthouze, S. F. Farmer, H. Cagnan and V. Litvak (2019). "Mechanistic Inference of Brain Network Dynamics with Approximate Bayesian Computation." bioRxiv: 785568.

Williams, G. V. and P. S. Goldman-Rakic (1995). "Modulation of memory fields by dopamine Dl receptors in prefrontal cortex." Nature **376**(6541): 572.

Yao, Y., S. S. Raman, M. Schiek, A. Leff, S. Frässle and K. E. Stephan (2018). "Variational Bayesian inversion for hierarchical unsupervised generative embedding (HUGE)." NeuroImage **179**: 604-619.

Youssofzadeh, V., G. Prasad and K. Wong-Lin (2015). "On self-feedback connectivity in neural mass models applied to event-related potentials." NeuroImage **108**: 364-376.

Zahrt, J., J. R. Taylor, R. G. Mathew and A. F. Arnsten (1997). "Supranormal stimulation of D1 dopamine receptors in the rodent prefrontal cortex impairs spatial working memory performance." Journal of neuroscience **17**(21): 8528-8535.

Zhao, L.-X., Y.-H. Ge, J.-B. Li, C.-H. Xiong, P.-Y. Law, J.-R. Xu, Y. Qiu and H.-Z. Chen (2019). "M1 muscarinic receptors regulate the phosphorylation of AMPA receptor subunit GluA1 via a signaling pathway linking cAMP-PKA and PI3K-Akt." The FASEB Journal **33**(5): 6622-6631.

Zhao, L.-X., Y.-H. Ge, C.-H. Xiong, L. Tang, Y.-H. Yan, P.-Y. Law, Y. Qiu and H.-Z. Chen (2018). "M1 muscarinic receptor facilitates cognitive function by interplay with AMPA receptor GluA1 subunit." The FASEB Journal **32**(8): 4247-4257.