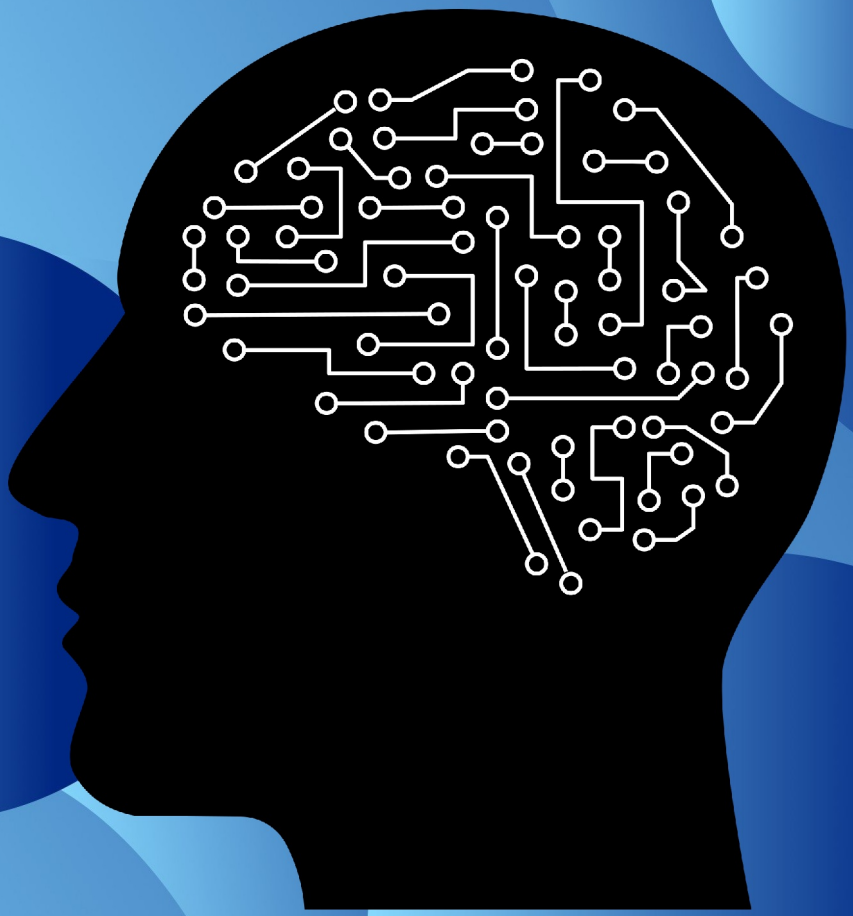
Data-Driven Modelling and Analysis of Attention-Working Memory Interplay

S.M. Ohkawa

Master of Science Thesis



TU Delft Delft University of Technology

Delft Center for Systems and Control

Data-Driven Modelling and Analysis of Attention-Working Memory Interplay

Master of Science Thesis

For the degree of Master of Science in Systems and Control at Delft University of Technology

S.M. Ohkawa

April 7, 2025

Faculty of Mechanical Engineering (ME) · Delft University of Technology

Abstract

In this project, a data-driven modelling algorithm for learning new dynamical models from experimental magnetoencephalography (MEG) data is introduced. This algorithm provides a contrast to existing hypothesis-driven modelling techniques in neuronal dynamics, and is useful for generating new insights from data when hypotheses for the neural mechanisms underlying a process are not readily available. The algorithm utilises universal differential equations (UDEs), combining white-box modelling with machine learning techniques. The algorithm is applied to a single-subject human MEG dataset to produce an oscillator network model. The model captures the frequency-domain behaviour of and interaction between several brain regions of interest during completion of a working memory (WM) task. The machine learning techniques are used to identify the role of attention mechanisms in these interaction dynamics, providing neuroscientists with data-driven insights into the brain dynamics underlying the attention-WM interplay.

Table of Contents

Chapter 1

Introduction

Both working memory (WM) and attention are fundamental concepts in human cognition that are essential for the completion of everyday tasks, with attention key in prioritising the processing of the wealth of sensory input that humans receive and WM linking the sensory input with past experiences. Dynamical modelling of interactions between neuronal populations has been instrumental in building understanding of both processes, as these models describe potential neural mechanisms which produce the behaviours and neural recordings that neuroscientists observe.

Techniques for large-scale modelling of neuronal population interactions vary tremendously and models are often constructed in a hypothesis-driven manner to represent known neuronal processes [17]. With advancements in neuroimaging techniques producing larger datasets, data-driven methods have the opportunity to bring new insights into the neuroscience domain. This project aims to design a data-driven modelling algorithm to investigate the interplay between attention and WM processes.

In this chapter, background concepts and studies related to attention and WM are presented, culminating in the formulation of the project goals. In Section 1-1, motivation for data-driven modelling of attention and WM is presented. In Section 1-2, relevant background knowledge from the neuroscience domain is summarised. In Section 1-3, related work in neuronal dynamics is introduced. In Section 1-4, the project goals are described, and an outline for the remaining chapters is provided in Section 1-5.

1-1 Motivation

Fundamental understanding of WM and attention is an important step to understanding general human cognition and behaviours, especially the linking of low-level neuronal functions with the complex emergent properties of the brain. Both concepts have seen decades of study [39, 94], first through the lens of psychology and cognitive science, and more recently through a biophysical lens in neuroscience and neuronal dynamics.

In a more practical perspective, the fundamental roles attention and WM play in cognition mean that improved understanding of their mechanisms can contribute to the understanding of several neurological disorders. Studies have shown that individuals with major depressive disorder tend to have reduced visual WM capabilities, and that those with bipolar disorder have reduced verbal WM [21]. Both WM and attention networks are compromised in individuals diagnosed with attention deficit hyperactivity disorder (ADHD) [21, 94], and it has been found that focused training of WM can improve symptoms associated with ADHD [94]. Although biophysical understanding of these disorders is growing, clinical diagnosis often still occurs based on patient behaviours or expert analysis of neuroimaging data [101].

In the research domain, studies show that techniques from computational sciences and dynamics have the potential to improve practices in the medical domain. The authors in [101] have developed a computational toolset which successfully diagnoses ADHD and schizophrenia based on neuroimaging data such as functional magnetic resonance imaging (fMRI), showing how a direct data-driven approach can streamline the processing of highly complex data into meaningful results. Hemodynamic models have also been used to improve the resolution of transcranial doppler measurements, giving clinical experts a more complete basis on which to base their diagnoses. Researchers have demonstrated how such a technique can improve diagnosis of cerebral vasospasm [99].

Thus, a data-driven modelling approach to study attention and WM has both scientific and medical value. Dynamic modelling allows these abstract concepts to be described in terms of concrete neuronal activity, connecting the biophysical with the behavioural. Informing the modelling through data has the potential for the emergence of new models and for characterisation of differences between individuals. These models can inform experts in the medical community, and some have already been shown to improve clinical diagnoses.

1-2 Background

This section outlines some preliminary knowledge from the neuroscience domain which informs the methods introduced in subsequent chapters.

1-2-1 Magnetoencephalography (MEG)

This section describes the working principles of MEG, the mode of data collection for this project. MEG records fluctuations in the magnetic field outside the head due to neuronal activity in the brain. Capturing these magnetic fields requires sensitive equipment, as the fields from the brain are generally less than 10^{-15} T, as compared to Earth's magnetic fields at about 10^{-4} T [102]. The bulk of this signal comes from excitatory cortical pyramidal cells, which have long dendrites and are often oriented normal to the cortical surface. When they fire in synchrony, the magnetic field resulting from the current flow is strong enough to be recorded by magnetometers outside the head.

The small magnitude of the desired signals means that MEG data is subject to large disturbances, due to e.g. inherent noise, cardiac artifacts, subject body movements [55], and often some manual pre-processing and filtering is necessary to isolate the desired signals [102]. The FieldTrip MATLAB toolbox [88] is a popular toolbox for completing this preprocessing, as well as for subsequent time-domain and spectral analysis.

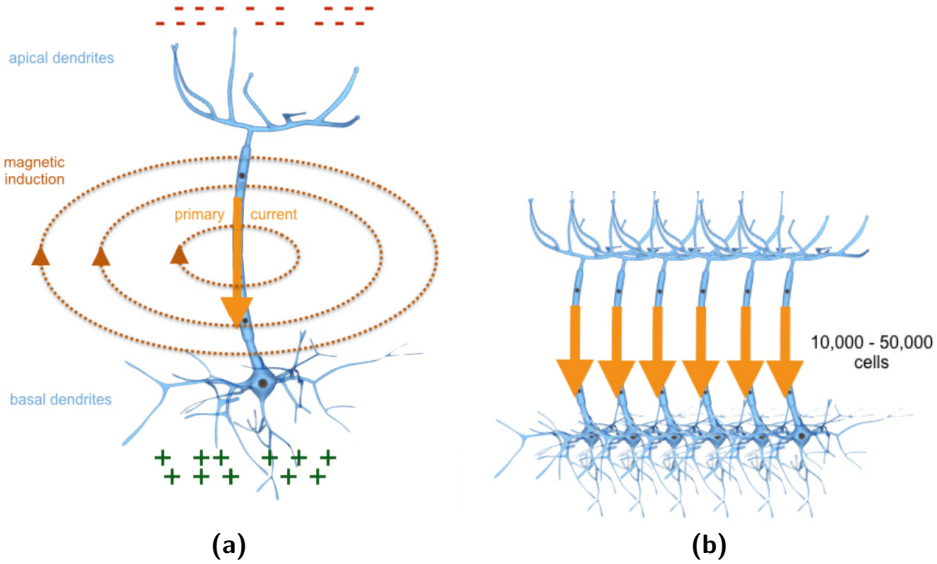


Figure 1-1: Neuronal origin of MEG signals [7]. (a) - A single pyramidal cell, with long apical dendrite. The flow of current down the dendrite produces a magnetic field. (b) - A collection of pyramidal cells. These are generally oriented in parallel with each other and normal to the cortical surface, so that the magnetic fields from each cell are constructive and the resulting magnetic field is large enough that it can be measured outside the head. It is estimated that 10,000-50,000 active cells are required to produce fields that are strong enough to detect with MEG.

1-2-2 Visual Attention and Working Memory in Neuroscience

This section outlines some key results and existing theories surrounding visual attention and WM. Theories in cognitive science show a clear link between the two processes, with attention directing WM processes to align with internally stored goals. In particular, this project focuses on attentional protection of WM in the presence of distracting inputs, i.e. mechanisms which maintain a target memory and prevent its replacement or corruption due to other inputs. From a mechanistic perspective, modulation of oscillatory behaviours in neuronal populations has been identified as key to both WM and attention processes, and so is the primary focus of subsequent chapters.

Neuroimaging Results in Attention Studies

Attention has been a subject of interest in human studies for decades, with theories appearing as early as the 1860s (see [94] for a timeline of attention-related studies). It is generally seen as a mechanism through which the brain's limited capacity for sensory processing is managed, by allocating cognitive resources based on some rules. Visual attention relates exclusively to the visual sensory modality, with attention allocated on the basis of spatial location or visual feature. This project investigates covert, top-down visual attention, that is internally directed visual attention without physical movement of the eyes.

Early studies of attention focused primarily on the firing rates of neurons as the mechanism of attention. Neural recordings have demonstrated that attending to specific visual features increases the firing rate of neuronal populations that are selective to those features [96, 82, 83].

These observations lead to the development of rate-based modulation theories of attention, notably the *biased-competition principle* [95], which posits that signals from early visual cortex compete for representation through overlapping receptive fields in higher visual cortex. In this context, attention works by biasing the response of the higher-level populations towards signals from lower neuronal populations that are selective for the attended feature.

More recent studies have identified modulation of oscillatory power, frequency and phase as another key mechanism in visual attention [20]. Neuronal recordings in the visual cortex of macaque monkeys have identified that increased gamma-band (35-90 Hz) synchronisation and reduced low-frequency (<20 Hz) synchronisation can be seen in neuronal populations that are selective to the attended feature [43, 44]. The increase in gamma-band synchronisation and power received notable attention has been observed in other macaque visual cortex studies [118, 13]. However, researchers have reviewed these results and confirmed with additional human electroencephalography (EEG) studies that only certain visual stimuli result in reliable narrowband gamma oscillations in the visual cortex, especially when considering natural stimuli [56]. If the gamma oscillations are largely dependent on the type of visual stimulus, then they are not likely to be key in more general functions such as visual attention, making theories based on these gamma-band observations [42] controversial.

Recently, alpha rhythms have been linked with distractor suppression, with neural recordings showing increased alpha power and synchronisation in human brain regions associated with task-irrelevant stimuli [120, 16]. However, contradictory experimental results have made the details of this mechanism unclear [53]. For example, whether alpha modulation occurs in anticipation of or in response to a distractor stimulus is up for debate, with studies reporting differing conclusions [120, 16].

Despite the contradictory results, there have been several theories put forward to describe how alpha-band modulation could suppress the processing of certain neuronal population signals. Some researchers have suggested that synchronised alpha oscillations enhance already existing differences in neuronal population excitability, further biasing the response of a network towards the preferred population [69]. Others have tackled the role of alpha rhythms in distractors more directly. In an extension to an existing theory on how gamma-band modulation could impact neuronal population excitability, researchers have posited that the relatively slow build up in excitation in the longer alpha period (compared to the short gamma period) means that the overall excitability of the neuronal population does not reach the short peak that can be seen in gamma [42].

Working Memory and Attention Interplay

In the cognitive science domain, many theories of WM include attention, often in the role of a central executive which directs WM maintenance and storage depending on the current goals [20, 77].

A multicomponent model in [6] includes attention as one of the primary mechanisms through which a component termed the *central executive* influences the other components. This *central executive* maintains the task-relevant contents of the episodic buffer via attentional refreshing. The *central executive* also utilises perceptual selective attention to distinguish between distractors and task relevant inputs to the modality specific workspaces (visuo-spatial sketchpad or phenomenological loop). Attention, then, works both as a filter for items into WM,

and as a mechanism to avoid forgetting previously stored items. In the embedded-processes model [24], a concept termed the *focus of attention* denotes the subset of activated long-term memory which is currently utilised in achieving some set goal. This focus can be shifted by the central executive or due to the appearance of a salient stimulus input. Items which are brought into activated long-term memory can decay over time, but are refreshed when brought into the *focus of attention*, in a mechanism similar to the attentional refreshing of the multicomponent model.

Although the specific components of the models differ, the roles of attention in them are complementary, involved with sensory filtering and memory maintenance. In this project, the maintenance role of attention, and how this is activated in the presence of distracting input, is the focus.

Looking then at insight from neuroimaging studies, oscillatory modulation of neuronal activity has also been observed during WM tasks. For example, researchers have shown that alpha oscillation desynchronisation recorded using EEG in humans can predict recall accuracy during a WM task [86]. Given the role of alpha synchronisation in distractor suppression, this result also supports the role of attention in enhancing and protecting task-related WM items, as is suggested some cognitive science WM theories [6]. In addition to alpha-band activity, there is evidence that lower frequency beta and theta synchronisation is linked to WM management [11, 33]. Increased activity and synchronisation in these frequencies are present during the execution of WM tasks in the prefrontal cortex (PFC) regions that have been linked with WM, while alpha modulation is most prominent in the sensory cortices. These findings suggest that low frequency synchronisation could reflect top-down executive control and high-frequency synchronisation the enhancement or suppression of sensory inputs into WM storage.

1-2-3 Relevant Brain Regions

In this section, brain regions hypothesised to play a role in visual attention and WM are presented. It should be noted that this discussion is not exhaustive, as the human brain is incredibly complex and both processes likely involve the cooperation of many tens of regions. Instead, this section focuses on key brain regions whose involvement is supported by substantial literature.

Visual Attention

Because visual attention is closely related to the processing of visual sensory information, the visual cortex is often studied and modelled in this context. A more comprehensive description of the human visual cortex can be found in [112], with key points summarised here. The visual cortex can be broken down into serial components, with primary visual cortex (V1) receiving sensory input through the retina, optic nerve, and lateral geniculate nucleus (LGN). At this level, different neuronal populations generally correspond to distinct receptive fields. At higher levels of the visual cortex, pooling of the signals means that populations often have overlapping receptive fields, but are instead selective for higher-order visual features, e.g. the direction of motion of a stimulus.

Determining the origin of top-down signalling to the visual cortex is more challenging, especially because the nature of this signalling is not well understood. However, there is evidence that several regions of the PFC are activated early on in tasks which involve endogenous or top-down attention [20, 36]. The frontal eye field (FEF) has been especially implicated, with studies involving microstimulation of the FEF resulting in neuronal recordings similar to those seen in attention studies [83], and simultaneous recordings of FEF and V4 supporting the hypothesis that signals in FEF initialise and synchronise gamma oscillations in V4 [52].

The lateral intraparietal area (LIP), a part of the intraparietal suculus (IPS) in the posterior parietal cortex, is also activated during allocation of visual attention [20]. Some researchers hypothesise that it stores a saliency map for visual stimuli, informed both by bottom-up sensory information and top-down goal-oriented signals [15]. Studies in monkeys have supported a causal relationship between FEF and LIP, with salient stimuli leading first to LIP activity and then FEF and visual search leading first to FEF and then LIP activity [40]. This suggests that the FEF and LIP may be a serial part of a network which links the PFC to the visual cortex in a mechanism for visual attention. A recent review [84] supports the theory that the right and left FEF and IPS (including the LIP) correspond to the contralateral visual fields, although some experimental results suggest that the right hemisphere responses are more general and dominant [9].

Working Memory

There is evidence that the PFC is a source of top-down signals for WM processes, likely linked to the central executive present in many of the conceptual theories of WM [39]. Early reviews of neuroimaging data recorded during execution of WM tasks implicate the dorsolateral (DL) and ventrolateral (VL) PFC as important parts of the WM network [89, 110]. A more recent review [21] supports these findings, with additional evidence for the anterior cingulate cortex involvement with the selection of task relevant stimulus input.

Although the importance of the DLPFC in the WM network is well supported [21], the details of the lateralisation of this region are under contention. Studies have demonstrated through transcranial direct current stimulation that activation of the right DLPFC increases WM capacity in the presence of distractors [76], supporting a role for the right in both WM and attention. There exists evidence that the left DLPFC is critical for WM tasks while the right is involved with higher-order decision making, e.g. attention direction [8]. Some researchers argue that the split is domain-specific [41, 114], i.e. verbal versus spatial memory. However, some studies, using repetitive transcranial magnetic stimulation, have suggested that right stimulation improves verbal WM and reduces spatial WM accuracy [41], while others have produced evidence for the opposite [114]. These results highlight that, similar to the attention network, the brain regions which are involved in the WM network remains an open question.

1-3 Related Work

This section summarises the existing body of literature surrounding dynamical modelling of attention and WM. White-box models based on structural understanding of the brain form a large proportion of these models. As neuroimaging data has become more widely available,

data-driven techniques have also been developed, with dynamic causal modelling (DCM) remaining one of the most popular.

1-3-1 White-Box Models

Much of the early dynamical modelling work in neuronal dynamics consists of white-box models. These models are typically constructed with the aim of replicating some behaviour observed in neuroimaging studies and rely heavily on physical understanding of neurons and organisation of neuronal populations to justify their structure. They have increased dramatically in complexity as understanding of brain structures grows.

White-box neurodynamic models exist at several resolutions, from single-neurons to populations of tens of thousands of neurons to full cortical models. At the single-neuron level, spiking models such as the Izhikevich [59, 60] and Hodgkin-Huxeley [57] are popular. Neuronal population models sometimes include large collections of these individual spiking neurons [37] but often make some simplifying assumptions to represent the collective activity directly. The Wilson-Cowan [116] and Jansen-Rit [61] models are two examples of population convolution models that have seen great success [85]. Models of attention tend to lie at the cortical level, as neuroimaging studies suggest that the process involves interaction between different populations within the visual cortex and with other distant brain regions (see Section 1-2-2).

One of the earliest mathematical models of attention, first proposed in [96], is based on the *biased-competition principle* discussed in Section 1-2-2. This model and later extensions [34, 35, 36, 37] describe interactions between excitatory and inhibitory neuron populations in the visual cortex, with external inputs representing attention signals biasing competition between excitatory populations that are selective for different features. See Figure 1-2 for a schematic.

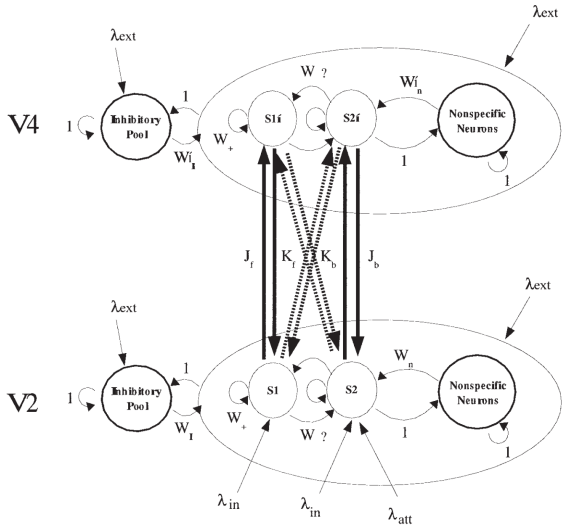


Figure 1-2: [37]. Spiking biased competition model of the visual cortex. Excitatory populations Si respond selectively to different features, and have strong connections $J_i > K_i$ to *corresponding* populations in later visual cortex. Competition is implemented through shared inhibitory pools of neurons.

New development of hypotheses for the neural mechanisms of attention have lead to a similar growth in the diversity of attention models. For every hypothesis, a model is developed which describes in further detail the biological processes which are hypothesised to play a role. For example, three general areas which have seen quite some attention in recent literature are:

1. Macro-Structure. These models increase the detail of the cortical models from two populations (one excitatory and one inhibitory) to many different populations. Generally they are organised into layers, which include populations with different characteristics, and columns, which include populations with similar stimulus selectivity. Some of the models also expand to include interaction with brain regions outside the visual cortex.

2. Biological Neuron Type. These models increase the detail of the different neuron types present in the brain, further than the typical excitatory versus inhibitory characterisation. Often focus is placed on the role of different inhibitory interneuron types.

3. Neuromodulator. These models increase the detail of the synaptic connections, whose action is affected by the concentration of neuromodulators. They include separate dynamics for these concentrations, or alter the structure of the models in ways that reflect the presence or absence of some neuromodulators.

Figure 1-3 gives an overview of some of the extensions and associated publications.

Although these models describe existing hypotheses well and offer great insight into potential mechanisms for behaviours observed in data, they are challenging to apply to data directly. The level of complexity makes optimisation of parameters with these models challenging and highly subject to initial guesses, which are difficult to justify biologically despite the biological

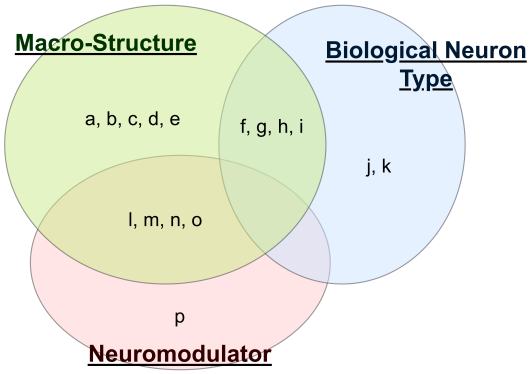


Figure 1-3: Modern white-box attention models, organised by the biological focus of the model. *Macrostructure* refers to cortical level structures, e.g. columns and layers. *Biological neuron type* refers to the explicit modelling of different cell types more detailed than excitatory/inhibitory, e.g. types of inhibitory interneurons. *Neuromodulator* refers to the modelling of neuromodulators in synaptic connections, e.g. acetylcholine, dopamine. The alphabetic references are as follows: a - [108], b - [109], c - [91], d - [12], e - [54], f - [19], g - [105], h - [72], i - [103], j - [71], k - [107], l - [2], m - [3], n - [4], o - [62], p - [38]

motivation for the structure. Detailed data with respect to the different parts of the model is also not widely available.

For example, consider the model in [4], which is built on the experimental and biological support that neuromodulators, especially acetylcholine (ACh), play a role in the attentional modulation of neuronal activity [5, 104]. This computational model investigates the interaction between top-down attentional signals, visual stimuli and the basal forebrain (BF), where ACh is produced. The schematic of the model is given in Figure 1-4, including a visual stimulus pathway through the lateral geniculate nucleus (LGN) and the thalamic reticulate nucleus (TRN) in addition to the typical microcircuit neural model. Simulations of this model show that top-down attention signals can produce the rate-based modulation behaviours observed in the visual cortex via ACh release.

Although these models can be related to data on a coarse level, it is difficult to justify the detailed structure with data. Retrieving measurements from two different populations in early visual cortex with different receptive fields often requires invasive techniques which are not often applied to humans, and high-time-resolution measurements of neuromodulators are likewise difficult to obtain. However, these models can be used to constrain or inform data-driven modelling algorithms, which are introduced in the following sections.

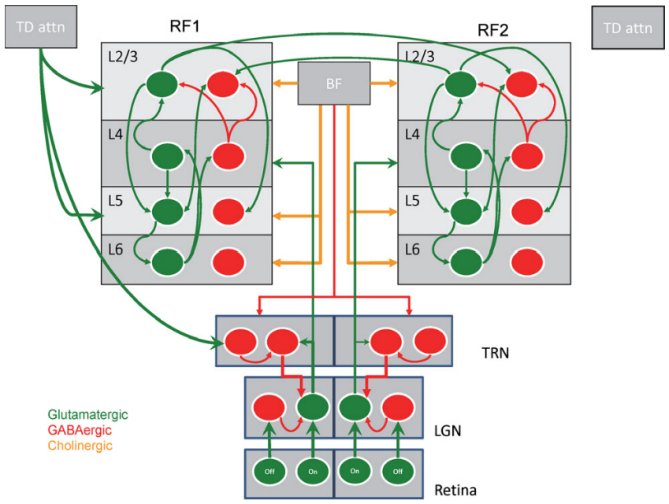


Figure 1-4: Cortical model used in [4] to investigate the role of ACh in a neural mechanism for visual attention. The BF releases ACh when stimulated, leading to cholinergic stimulation of the general visual cortex. Top down attention signals are also directed to certain layers in the visual cortex and to the TRN, where ACh is released locally. Red circles are inhibitory populations, green are excitatory.

1-3-2 Dynamic Causal Modelling

DCM is one of the most popular data-driven modelling techniques in neuroscience, with the original paper [46] now with over 5,000 citations. Although proponents of DCM generally frame it as a tool to test competing hypotheses about effective connectivity in the brain, it is at its core a nonlinear system parameter estimation technique which leverages Bayesian inference, which can be applied to a broad range of problems.

The goal in this framework is to determine a posterior distribution for the parameters θ of a model m based on observations y. Considering Bayes' rule, this can be written as,

$$p(\theta|y,m) = \frac{p(y|\theta,m)p(\theta|m)}{p(y|m)}. \tag{1-1}$$

The likelihood $p(y|\theta,m)$ is given by the relationship between the system output and the model with specified parameters, all of which are known. The priors $p(\theta|m)$ are informed by expert knowledge. However, the model evidence

$$p(y|m) = \int p(y|\theta,m)p(\theta|m)d\theta \tag{1-2}$$

generally cannot be calculated analytically. Therefore, an approximate method is needed to calculate both the model evidence and the posterior distribution. One such method, popular in DCM, is Variational Laplace [119].

Although DCM was originally developed for modelling fMRI data, researchers have extended the framework to MEG and EEG [30, 65]. The model structure relies heavily on previously developed white-box models, demonstrating how they can contribute to data-driven modelling. In this case, the observation equation is a forward source reconstruction model (see Section 3-2), and the process equation is typically a neuronal mass model reflecting the cross-membrane potential of a neuronal population. A neural mass model specifically for MEG and EEG data based on the Jansen-Rit population model was developed in [29] and is the standard model applied to DCM for MEG and EEG [30, 65]. Each brain region is modelled as a cortical column with three different neuronal populations, each in a separate layer. A schematic of one of these regions is shown in Figure 1-5.

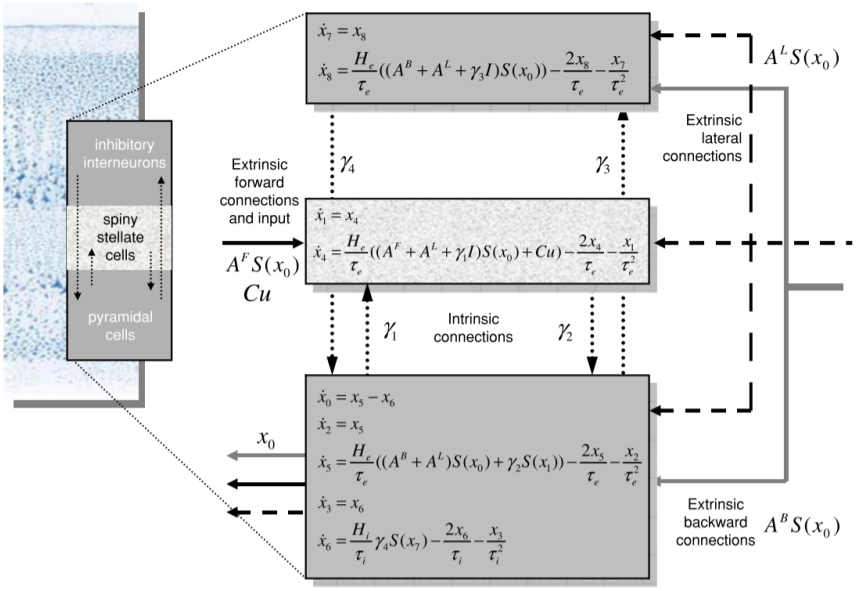


Figure 1-5: [65]. Multi-layer cortical model based on the Jansen-Rit neuronal population model for the replication of MEG and EEG signals. Deep pyramidal cells are the primary source of these signals and are driven through intrinsic connections with excitatory spiny stellate cells and inhibitory populations and extrinsic connections to other cortical columns.

One of the key advantages of DCM is that, in addition to parameter fitting, the framework allows for different models to be compared in terms of the approximate model evidence $p(y|m)$. This model evidence does not depend on the specific parameters, but rather gives the likelihood that the observed data was generated by the model m. The DCM framework, then, can be used to compare how well different models explain data using an approximation of model evidence and estimate the parameters of a given model based on data. In the neuroscience community, these tools are typically used to compare competing hypotheses with respect to effective connectivity between brain regions, with different models representing different hypotheses. In many DCM applications, the parameters are the weighting of directional synaptic connections between neuronal populations and the model structures differ in the presence and absence of these connections.

Although the uptake of DCM has been large, it has not been without criticism. One of the most compelling arguments (brought forth in [78, 28, 79]) highlights the technique's heavy

reliance on the initial choice of model set, which can be considered a type of implicit prior on the Bayesian inference. One group has shown in particular that previously accepted DCM results can be challenged if the considered model set is expanded, and that biologically implausible models can in this case exhibit higher model evidence than the previously accepted model [78]. The reply to this criticism [45] has acknowledged this reliance, conceding that DCM methods implicitly assume that all considered models have the same prior plausibility. The authors emphasise that the purpose of DCM is not to determine the best possible model, but to compare between a set of carefully chosen models which represent equally plausible hypotheses.

This criticism highlights the reliance of Bayesian inference and model comparison on both explicit and implicit priors, whose validity must be well justified. The quality of this justification then limits the strength of the claim that can be made from the inference. DCM is fundamentally a hypothesis-driven modelling framework which requires an apriori selection of hypotheses or models to compare. Thus, DCM is well suited to answering scientific questions where a complete set of competing hypotheses is known, but less so in more exploratory domains where mechanisms are unclear and the model set is difficult to justify.

1-3-3 Data-Driven Models

In this section, data-driven models outside the Bayesian framework of DCM are discussed. These models tend to rely much less on established biophysical structures and more on behaviours seen in data, providing a more abstract representation of neuronal processes. The move away from biophysical processes means that these models tend to avoid the common criticisms of DCM and white-box modelling, where the model structures are assumed apriori. Avoiding these assumptions on model structure, however, also makes it more challenging to relate the trained model to biophysical processes, often making them more useful for prediction but less for scientific insight.

Some models make little to no attempt at reflecting underlying neuronal mechanisms, and instead aim primarily to predict future time-series or classify different dynamical states. They make minimal apriori assumptions on model structure. A common framework is to assume a certain number of states and a linear model, giving for example a general auto-regressive (AR) [100] or auto-regressive moving average (ARMA) [68] model.

Because oscillating behaviours are so ubiquitous in neuroimaging studies, especially MEG and EEG (see Section 1-2-2), a body of literature surrounding oscillator-based data-driven models has also been developed. In this case, the biological origin of the oscillating behaviour via excitatory-inhibitory couplings or feedback loops between neuronal layers is ignored. Instead, this behaviour is represented through an abstract self-sustaining oscillator, and interactions between the oscillators model interactions between neuronal populations.

Coupled linear oscillators have been shown to replicate time-domain eight-channel EEG recordings [74, 73]. Provided the time-domain recordings are band-pass filtered to include a restricted frequency domain, researchers have shown that one oscillator per EEG channel is sufficient to replicate the recordings. In another study focusing on wide-band frequency characteristics of EEG, researchers have suggested that a coupled (stochastic) system of 10 linear oscillators is required to reproduce the frequency behaviour of a single EEG channel [49]. Thus, although these linear models lend themselves well to parameter optimisation, they

can be limiting in terms of the behaviours they can replicate (or in other words limiting in the high structural complexity required to exhibit more complex behaviours).

Nonlinear oscillators can often exhibit far more complex behaviours with simpler parameterisation than their linear counterparts, and have similarly been used to replicate MEG and EEG recordings. For example, researchers in [49] have shown that a system of two coupled nonlinear oscillators can reproduce the wide-band frequency behaviour of a single EEG channel as well as the aforementioned 10 linear oscillator system. The oscillators designed for this purpose merge dynamics from two popular nonlinear oscillators: each node exhibits self-sustained oscillation through Van der Pol dynamics, and coupling then occurs via Duffing oscillator terms. A schematic is given in Figure 1-6, with the accompanying dynamics,

$$\ddot{x}_1 + (k_1 + k_2)x_1 - k_2x_2 = -b_1x_1^3 - b_2(x_1 - x_2)^3 + \epsilon_1\dot{x}_1(1 - x_1^2) \tag{1-3}$$

$$\ddot{x}_2 - k_2x_1 + k_2x_2 = b_2(x_1 - x_2)^3 + \epsilon_2\dot{x}_2(1 - x_2^2) + \mu dW. \tag{1-4}$$

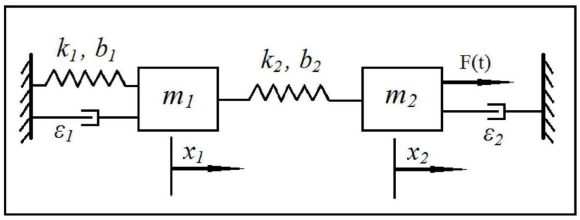


Figure 1-6: [48]. Schematic of the two-state Duffing-Van der Pol oscillator.

Researchers have shown that this model performs well when replicating power spectra of EEG signals as well as other informative metrics such as Shannon and sample entropies [50, 48]. Duffing coupling has also been shown to replicate time-domain cross-frequency interactions hypothesised to occur in the brain [73], where behaviour in one frequency range can modulate behaviour in a different range.

It should be noted that there is significant crossover between the neuronal population models often used in white-box modelling and the nonlinear oscillators mentioned in this section. Both the Wilson-Cowan [116] and Jansen-Rit [61] models have biophysical roots and describe the interactions between different types of neuronal populations. From a behavioural point of view, both models are also nonlinear oscillators. Both models have been successful in replicating some behaviours seen in neuroimaging data like MEG and EEG [117, 111, 10, 29]. However, because they have biological roots, they have large numbers of parameters with highly nonlinear coupling when compared to, for example, the phenomenological Van der Pol or Duffing oscillators mentioned in this section. Thus, in data-driven modelling contexts, phenomenological nonlinear oscillators are often chosen in favour of these neuronal population models, which can exhibit similar behaviours without the heavy parameterisation.

1-4 Project Goals

The general goal of this project is to use dynamical modelling techniques to understand the interplay between visual attention and WM. Many studies suggest that modulation of

oscillatory dynamics are important to this interaction, and there exists a substantial body of literature surrounding the development of oscillatory dynamical models to replicate observed neuronal behaviours. White-box models in this domain tend to be very complex and focus on describing one neural mechanism in detail. Although they are typically well biophysically justified, data resolution is often not high enough to verify the detailed structures with data. DCM, a popular data-driven technique in neuroscience, brings these models closer to data by maintaining a biophysically relevant model structure while leveraging Bayesian techniques for parameter fitting and model comparison. Although this framework is well suited to hypothesis testing, it does not lend itself to exploratory modelling where well-formed hypotheses are not readily available. For attention and WM, there exists a wide range of hypotheses, many of which have been captured by the wealth of existing white-box models and many others which are not easy to translate to dynamical models, so a more flexible modelling framework which relies more on data is preferred.

One of the key interactions between WM and attention occurs when attention is internally directed in order to maintain and protect an item stored in WM from distracting inputs. The model developed in this project focuses on this aspect of WM-attention interplay, and is built on human MEG data collected in [120]. Researchers in [120] have investigated how visual presentation of distractors during a WM task, and the resulting activation of attentional protection mechanisms, affects activity in the human brain. This project has two goals: first, to develop a phenomenological data-driven oscillatory network model using MEG data; second, to perform model-based analysis to contribute to a mechanistic understanding of the attention-WM interplay. This method should leverage well-supported knowledge in the neuroscience domain while remaining flexible enough to discover new neural mechanisms from data.

1-5 Thesis Outline

The remainder of the report is structured as follows. Chapter 2 provides theoretical background for several techniques in the scientific machine learning domain, which are utilised in the modelling algorithm. Chapter 3 presents the experimental framework and results from the researchers who produced the dataset used as the basis for the model. The modelling algorithm is presented in detail in Chapter 4. Results from application of the method to the dataset and subsequent analysis are summarised in Chapter 5. Finally, concluding remarks and potential directions for future research are outlined in Chapter 6.

Chapter 2

Theoretical Background

In the previous chapter, existing work in dynamical modelling in neuroscience as well as hypothesised neural mechanisms for the attention-working memory (WM) interplay were discussed. Although many dynamical models have evolved to describe different biophysical processes in detail, there is not currently a consensus on which of these many processes are important in attention mechanisms.

Many data-driven modelling methods in neuroscience are fundamentally *parameter identification* methods, in which the model structure is fixed and parameters are determined based on data. Selection of model structure typically fixes the neural mechanisms that are being investigated. In this case where the potential mechanisms are so varied, it is difficult to justify a single model structure. If a single model structure cannot be selected, *function identification* techniques are required.

In this chapter, background on the function identification techniques used in the project are presented.

2-1 Function Identification

Work in function identification has increased dramatically with the general increase in available data, as seen especially in the explosion of machine learning techniques and applications. In some machine learning techniques, the function identification problem is recast as a parameter identification problem by using a universal function approximator, often a neural network. In this project, only simple feedforward neural networks are considered, and any subsequent mention of neural networks refers to this class. There exist universal approximation theorems which essentially show that there exists a set of parameters for a neural network such that the network can approximate a wide range of functions, provided the network contains sufficient depth and width [58]. Thus, provided the network is large enough, it can represent many different functions depending on the parameter values.

The field of scientific machine learning aims to merge the use of these universal function approximators with expert knowledge in an effort to better restrict the function identification

problem [93]. This can occur through the introduction of learning biases, structural limitations on the network, or a combination of both [113]. One development of particular interest in dynamical modelling is the neural ordinary differential equation (NODE) and subsequent universal differential equation (UDE) extention, which are introduced in this section.

2-1-1 Neural Ordinary Differential Equations

The seminal NODE paper was published in 2018 [23], offering a new approach for training hybrid physics-based and deep learning models [113]. For some context, readers with a background in machine learning are pointed to [23] for a comparison to the residual neural network, which can be thought of as a discrete counterpart to the continuous transformations represented by NODEs.

NODEs are a class of ordinary differential equation (ODE) in which the dynamics of a state x are defined by a neural network N, with parameters λ, giving for example,

$$\frac{dx(t)}{dt} = N(x(t), \lambda). \tag{2-1}$$

This form means that the NODE replicates time-series data not through the classical input-output (time-state) relationship, but instead by capturing the dynamic relationship between the state and its derivative. One of the key challenges in training the network is the need to simulate the NODE, usually through a numerical solver, in order to produce a trajectory of the state which can then be compared to data. Numerical integration typically involves recursive use of the differential equation function, making calculation of the derivatives which are necessary for optimisation of the parameters challenging. A solution to this problem is the primary contribution of [23], and a simple example is presented in the following section for explanatory purposes.

Training

This section outlines a training algorithm for a basic time-invariant NODE, defined in (2-1). A schematic representation of the training philosophy is given in Figure 2-1. Denote the measurements of the real state as $X_m = [x_m(t_1)^T, x_m(t_2)^T, \dots, x_m(t_N)^T]^T$ and similarly denote the solution to the NODE at the corresponding times as X. The goal of the network training is to find network parameters λ which minimise the distance between the measurement and NODE solution vectors, giving the optimisation problem

$$\min_{\lambda, x_0} \mathcal{L}, \tag{2-2}$$

$$\mathcal{L} = ||X_m - X||_2^2, \tag{2-3}$$

assuming that the initial condition of the NODE $x(t_0)$ is also a parameter.

There is substantial research surrounding the calculation of the vector X from the ODE in (2-1), popular numerical integration techniques for solving ODEs include the Euler and

higher-order Runge-Kutta algorithms. Typically, the algorithms involve calculation of the state derivative using the dynamic function (in this case N) and application of this derivative to the current state to produce a future state in a forward step. This occurs in a loop until the state at all measurement times $t_i, i = 1, \ldots, N$ is known.

Minimisation of the loss function $\mathcal{L}$ typically occurs through numerical optimisation, with most algorithms requiring the gradient of the loss with respect to the parameters. In a traditional neural network where the state solution x is modelled directly as a function of time t, this derivative is relatively easy to compute as it involves only a single pass of the operations in the network. Reverse-mode autodifferentiation, or backpropagation, is a typical algorithm to use in this case. It calculates the derivatives by stepping backwards through the network operations from the loss function value to the parameters. In contrast, the state solution x from a NODE involves looping through an ODE solver, which uses recursive evaluations of the network. Backpropagation in this case involves passing back through all of the operations of the ODE solver, making it very memory and computationally intensive. Note that the ODE solver loop typically occurs at a much finer timescale than the measurements are taken, so there are many intermediate calculations to be tracked.

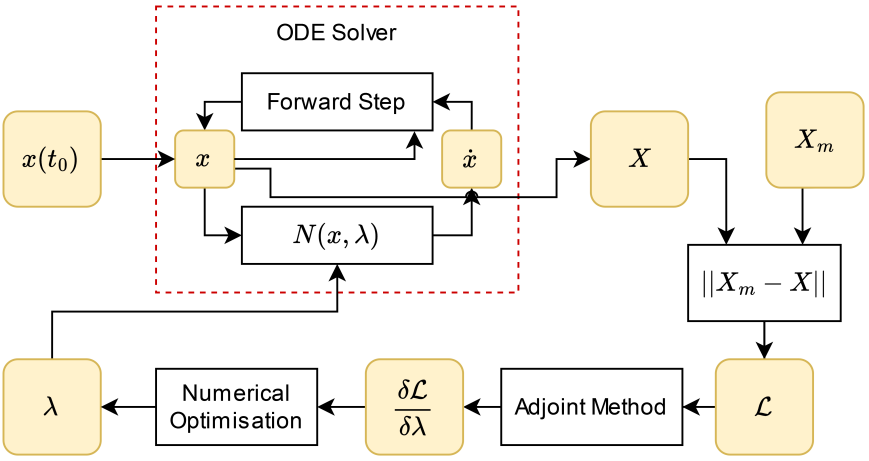


Figure 2-1: Basic NODE training schematic. An ODE solver is used to estimate the solution to the ODE, which is compared to measured data. Numerical optimisation of the network parameters requires the gradient of the loss function with respect to the parameters, which is difficult to achieve due to the ODE solver operations. Adjoint sensitivity methods can (theoretically) achieve this irrespective of the ODE solver used.

This leads to the primary contribution of [23]: the use of an adjoint sensitivity method to compute the gradient of the loss function with respect to the parameters, regardless of the ODE solver used. In brief, this method calculates the gradients by defining an adjoint state as the derivative of the loss function with respect to the parameters of interest. This adjoint state is governed by a secondary ODE which depends only on the ODE function, and not the solver used to propagate the solution. Numerically solving this secondary ODE backwards in time then gives the adjoint state solution, which is by definition the gradients of interest. With these gradients calculated, training of the network can proceed through numerical optimisation as usual. Details of the adjoint sensitivity algorithm and proof of its validity are provided in Appendix A.

This particular method for differentiation is often referred to as the backsolve adjoint method, but there exist many more and it is nontrivial to determine which method performs the best in which scenario. For example, researchers have identified that the backsolve adjoint method, although theoretically correct, can have large error in gradient estimation due to numerical errors in the ODE solver [66]. This issue is exaggerated in the backsolve adjoint method because the ODE state is solved for in both the forward and reverse directions. Errors in numerical methods mean that the forward and backward solutions may not be the same, and this issue is especially present for stiff ODEs. The researchers have offered several improvements, culminating in what is referred to as the quadrature adjoint method, but at high memory cost. Since then, many variations of the adjoint method have been developed, trading off efficiency, memory cost and robustness.

A theoretical review of the available methods can be found in [98], along with recommendations for selecting a method based on problem requirements. Figure 2-2 gives a summary of the performance metrics of the adjoint methods, with the forward sensitivity metric given as a comparison.

	Method	Stability	Non-Stiff Performance	Stiff Performance	Memory
Discrete	ReverseDiffAdjoint	Good	$\mathcal{O}(n+p)$	$\mathcal{O}(n^3+p)$	High
	TrackerAdjoint	Good	$\mathcal{O}(n+p)$	$\mathcal{O}(n^3+p)$	High
Continuous	Forward sensitivity eq.	Good	$\mathcal{O}(np)$	$\mathcal{O}(n^3p^3)$	$\mathcal{O}(1)$
	Backsolve adjoint	Poor	$\mathcal{O}(n+p)$	$\mathcal{O}((n+p)^3)$	$\mathcal{O}(1)$
	Backsolve adjoint◀	Medium	$\mathcal{O}(n+p)$	$\mathcal{O}((n+p)^3)$	$\mathcal{O}(nK)$
	Interpolating adjoint	Good	$\mathcal{O}(n+p)$	$\mathcal{O}((n+p)^3)$	High
	Interpolating adjoint◀	Good	$\mathcal{O}(n+p)$	$\mathcal{O}((n+p)^3)$	$\mathcal{O}(nK)$
	Quadrature adjoint	Good	$\mathcal{O}(n+p)$	$\mathcal{O}(n^3+p)$	High
	Gauss adjoint	Good	$\mathcal{O}(n+p)$	$\mathcal{O}(n^3+p)$	High
	Gauss adjoint◀	Good	$\mathcal{O}(n+p)$	$\mathcal{O}(n^3+p)$	$\mathcal{O}(nK)$

Figure 2-2: Comparison between different adjoint methods [98]. The black triangles indicate methods where checkpointing is used. n is the number of ODEs and p is the number of parameters.

In general, forward mode automatic differentiation is recommended for small systems (less than 50 parameters and ODEs), while reverse mode methods are recommended for larger systems. For applications which involve numerical solvers, direct autodifferentiation methods are not recommended due to the high memory requirements. Within adjoint methods, discrete methods tend to be more robust and continuous methods more efficient, although NODEs specifically have shown decreased training time using discrete methods. A separate study which used practical results from differentiation packages written in the Julia programming language [81] largely agrees with these recommendations.

Practically, the SciMLSensitivty package in the SciML Julia environment [93] is open source and implements most modern differentiation algorithms.

2-1-2 Universal Differential Equations

A promising extension to NODEs, published in [93], adds additional flexibility to the NODE framework through the integration of known dynamics. These models are known as UDEs. In this framework, the differential equation is not assumed completely unknown and represented with a universal function approximator, as is the case in NODEs. Instead, known differential equations that scientists have previously developed are utilised in the model, and the neural network function represents only unknown parts of the model. This has the advantage of reducing the complexity of the relationship the neural network must learn.

As with NODEs, the primary challenge in training UDEs is determining the derivative of the state trajectory with respect to the parameters in an efficient manner. Luckily, the adjoint sensitivity methods introduced in Section 2-1-1 are also applicable to UDEs. The proof of the backsolve adjoint method provided in Appendix A does not rely on the fact that the ODE is defined by a neural network function $N(x(t))$, except for the existence of derivatives. Therefore, provided the known dynamics are differentiable with respect to the inputs and parameters, the adjoint sensitivity algorithm holds.

A simple example which demonstrates the capability of this method for nonlinear system identification is the rediscovery of the well-known Lotka-Volterra dynamics [93]

$$\dot{x} = \alpha x - \beta xy, \tag{2-4}$$
$$\dot{y} = \gamma xy - \delta y. \tag{2-5}$$

In this example, the Lotka-Volterra dynamics have yet to be discovered and a scientist is interested in modelling the dynamics of predator-prey interactions. The birth rate α of the prey x is known. Based on existing observations, the scientist assumes that the predator population decays linearly in the absence of prey. The interaction dynamics are completely unknown. This combination of existing knowledge and unknown dynamics can be represented by a UDE,

$$\dot{x} = \alpha x - U_1(\theta, x, y), \tag{2-6}$$
$$\dot{y} = U_2(\theta, x, y) - \theta_1 y, \tag{2-7}$$

where U_1 and U_2 are feedforward neural networks parameterised by θ and θ_1 is the predator decay rate, both unknown.

To learn both the unknown parameters and the unknown dynamics, the model parameters can be trained on observations of the real system. For this example, time-series data for x and y are generated using the true Lotka-Volterra dynamics, plus noise. Note that these measurements need not be sampled regularly in time. The parameters are then fit to the data using a method similar to that described in Section 2-1-1. Results are shown in Figure 2-3, with good agreement between the measurements and model output, demonstrating that the UDE does capture the unknown interaction dynamics well.

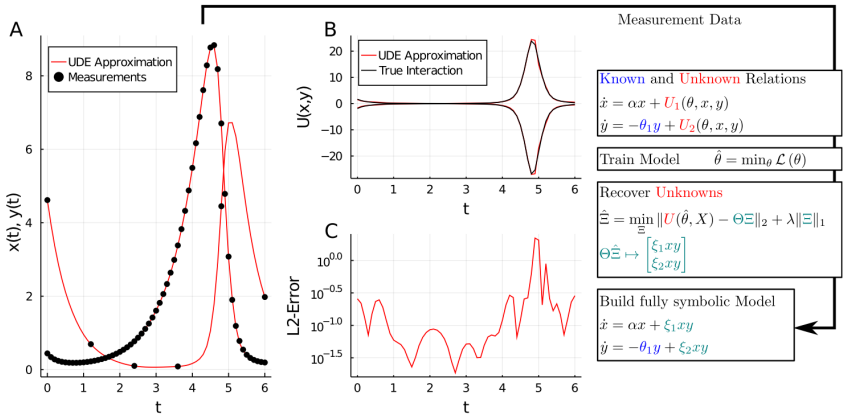


Figure 2-3: Comparison of measurements and the true underlying model with the neural network and symbolic models recovered using a UDE approach [93].

Figure 2-3 also shows a symbolic model built on the learned neural networks, which agrees with the true underlying Lotka-Volterra dynamics. This is learned using a symbolic regression technique, which is discussed in the following section.

UDEs have been applied in a wide range of fields. In physics, they have successfully recovered known dynamics of black hole systems, as well as discovered new equations of motion for black hole mergers based on gravitational wave measurements [64]. In electrochemical engineering, UDEs have been shown to outperform traditional models for lithium-ion battery performance and degradation [14]. In the medical domain, UDEs have been applied in pandemic modelling, most recently on a COVID-19 dataset [26], to discover new dynamics for the influence of quarantine policies on infection spread. To the best of the author's knowledge, there has yet to be an application in neuronal dynamics.

2-2 Symbolic Regression

In the previous section, the training of ODE models including neural networks is discussed. The resulting systems are powerful predictors, but have limited use for scientific insight into the structure of the dynamics as the neural network function acts as a black box. Ideally, the neural network function could be distilled into an arithmetic expression, which would then offer greater scientific insight into the dynamics. It has also been shown that distillation of neural networks to arithmetic expressions can improve their generalisability, showing improved prediction performance outside the training domain [25, 92]. This simplification of the neural network function can be achieved through symbolic regression.

Symbolic regression encompasses any technique which determines an arithmetic expression for the relationship between two datasets. These techniques are powerful tools in the data-driven modelling domain for automatic function discovery. The sparse identification of nonlinear dynamics (SINDy) method [18] is one such technique that frames the arithmetic search as an optimisation problem with sparse constraints. When applied to neural networks, the method requires,

1. a library of candidate functions from which the expression for the network will be built and
2. a large set of input/output data from the network.

As the name suggests, researchers originally developed this method to directly address the learning of nonlinear dynamics from data, but it can generally be used for any nonlinear function identification. For simplicity, consider the problem of learning an arithmetic expression for the function

$$y = f(x). \tag{2-8}$$

Because the function is known, a dataset can be built by choosing values of x within the domain of interest and calculating the corresponding y. This gives the dataset,

$$X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{bmatrix}, Y = \begin{bmatrix} y_1^T \\ y_2^T \\ \vdots \\ y_N^T \end{bmatrix}. \tag{2-9}$$

A library of candidate functions can be constructed from the dataset X,

$$\Theta(X) = \begin{bmatrix} \mathbf{1} & X & X^2 & \dots & \sin(X) & e^X & \dots \end{bmatrix}. \tag{2-10}$$

Provided the library is rich enough, it can be assumed that the function $f(x)$ consists of a small selection of these functions. Selection of the library is a challenging step in this process. If possible it should be informed by the target application. If this is not possible, monomial or polynomial expansions of the input X are a common choice, as these can capture complex nonlinear functions through their Taylor expansion representation.

The task is to find a sparse weighted summation of the library functions such that this sum matches the measured Y to within some tolerance ϵ. This can be framed as a sparse identification problem

$$\min_{\Xi} ||\Xi||_0, \tag{2-11}$$

$$||Y - \Theta(X)\Xi||_2 < \epsilon, \tag{2-12}$$

where Ξ are the sparse coefficients of the candidate functions. Optimising the zero-norm directly is a challenging task, so there exists a large body of literature dedicated to solving problems of this form. One method introduced in the original SINDy paper [18] is the sequential thresholding least-squares (STLSQ) method, which reforms the zero-norm into a two-norm plus a magnitude constraint,

$$\min_{\Xi} \frac{1}{2}||Y - \Theta(X)\Xi||_2 + \rho||\Xi||_2 \tag{2-13}$$

$$|\Xi_i| > \lambda. \tag{2-14}$$

The hyperparameters λ and ρ must be tuned to provide a reasonable balance between sparsity of the solution and error between the measurements and sparse summation. The SciML Julia environment includes a DataDrivenDiffEq package which readily solves problems of this form.

Chapter 3

Experimental Background and Data Analysis

Data-driven modelling algorithms are heavily influenced by the data they aim to capture and the context in which that data is produced. The experimental context is important for deciding the inputs and outputs of the system, as well as providing a basis for interpretation of the modelling results. The data utilised in this project is taken from a human magnetoencephalography (MEG) study [120], which is discussed in Section 3-1.

Equally important as the experimental context is the identification of data characteristics which are important to the research goal. These characteristics inform the behaviours which the model must capture in order to describe the target process well. In the context of this project, this means defining the data characteristics which are key to attention and working memory (WM) processes. Data analysis performed in this project to define these characteristics is presented in Section 3-2.

3-1 Experiment: Setup and Findings

This section outlines the experimental setup used to obtain the data used in this project. It also summarises the relevant findings from the researchers who designed and executed the study. Further detail can be found in the original paper [120].

3-1-1 Experimental Design

The original study focused on alpha-band power in the human visual and auditory cortices and how this relates to distractor suppression during WM tasks. Because the current study focuses on visual attention, only the results relevant to visual WM and visual distractors are discussed. A summary of the experimental setup follows.

The experiment includes two relevant trial types, 'no distractor' and 'distractor'. There are roughly eighty trials per type per subject. Multiple subjects participated in the experiment and all of these results inform the conclusions presented in the original study. However, for simplicity, only a single subject dataset is used in this study. For the complete duration of each trial, the subject's neural activity is recorded via MEG. Each trial consists of a single working memory task which consists of three phases,

1. Encoding. The target stimulus is presented and the brain state changes to store this information.
2. Maintenance. The target stimulus is no longer presented, but the target information is stored. During 'no distractor' trials, nothing is visually presented to the subject, and the brain simply continues to store the information. During the 'distractor' trials, visual distractors are presented to the subject. These distractors interfere with the maintenance of the target stimulus memory.
3. Recall. The subject memory is probed through a visual cue to recall the stored target stimulus.

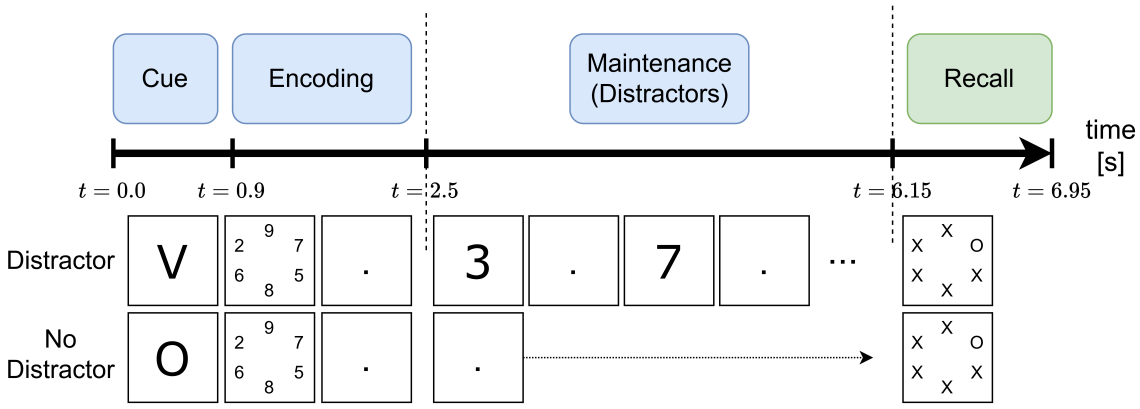


Figure 3-1: Schematic of the 'distractor' and 'no distractor' experimental trials with phases of WM. The Recall phase of interest is highlighted in green.

During the 'distractor' trials increased attention is required from the subject to prevent the distractors from corrupting the target memory. It is therefore assumed that some attentional mechanism is active during the 'distractor' trials which is inactive during the 'no distractor' trials.

3-1-2 Findings

The authors of the original study [120] have found that alpha-band (8-13 Hz) activity in the visual cortex is increased in the 'distractor' trials as compared to the 'no distractor' trials, supporting the hypothesised link between alpha-band activity in sensory cortices and distractor suppression (see Section 1-2-2). However, this difference occurs not during the maintenance phase when the distractors are presented, but in the following recall phase.

Moreover, the magnitude of this alpha-band increase has been shown to predict reaction times in the recall phase, forming a clear link with behavioural outcomes.

3-2 Data Analysis

The goal of this study is to build a data-driven model which provides insights into the neural mechanisms underpinning the attention-WM interplay. To do so, both the 'distractor' and 'no distractor' datasets are analysed to determine a suitable metric for model fitting. Attentional protection mechanisms are assumed to be active in the 'distractor' dataset and inactive in the 'no distractor' dataset. Therefore, the chosen data characteristic should capture observable differences between the two datasets for which a mechanistic explanation is desirable. The authors in [120] have identified alpha-band activity during the recall phase as an important behaviour in this process. Therefore, the dataset for the current study is refined to include only the 0.8 s recall phase, starting from the recall probe and ending at the subject response (6.15 - 6.95 s), and frequency-domain analysis is the focus. For simplicity, only a single subject dataset is considered.

The following sections outline the data analysis pipeline running from raw MEG data to the data characteristic chosen as the loss function for model fitting: the median power spectrum from 5-25 Hz.

3-2-1 Preprocessing and Source Reconstruction

This project uses the data from the experiment in [120] with preprocessing completed. The preprocessing steps included down-sampling from 1200 to 300 Hz, notch filtering for removal of line noise and harmonics, and independent component analysis for removal of artefacts from heart and eye movements [120]. This was completed using the FieldTrip MATLAB toolbox [88].

The following step for data processing is source reconstruction, which maps MEG data, collected by an array of sensors outside the brain, to the activity of certain regions within the brain, to determine how experimental manipulation changes the behaviour of these regions. Based on the discussion in Section 1-2-3, four brain regions of interest can be identified as relevant to visual attention or WM: primary visual cortex (V1), L-intraparietal suculus (IPS), L-dorsolateral (DL) prefrontal cortex (PFC), R-DLPFC. For ease of notation, these are sometimes referred to as Regions 1, 2, 3, 4, respectively. As a reminder, both left and right DLPFC are thought to be involved in WM processes, and the roles of the two regions are hypothesised to be different, preventing combining them into a single node. The sensory cortex V1 is a key focus point of many visual attention studies, as in the original study for this dataset [120]. Because the visual information is presented centrally in the visual field, left and right V1 are combined into a single node. IPS is a candidate for the origin of top-down attention signals to V1, and the left lobe in particular was identified in the original study as having a notable increase in alpha power just before the recall phase [120].

To complete the source reconstruction process, first a *forward model* that relates electrical activity inside the brain to measurements outside the brain is necessary. Then, the *inverse problem* of reconstructing internal electrical activity based on externally recorded signals can

be addressed. For both the current and original studies, the inverse problem is solved using the MATLAB FieldTrip toolbox [88], and a theoretical description of the solution can be found in Appendix B.

After this process is completed for each node in the cortical mesh, a parcellation map is used to map the vertices in the mesh to different regions based on their location. A parcellation map provided by the original researchers is used in this project. The time-varying mean taken across all of the contained vertices is used to represent each parcel. For some larger regions, times-series of different parcels are also averaged (e.g. left and right V1 are averaged to create the V1 time-series). The data is then low-pass filtered to below 30 Hz to isolate the lower-frequency bands of interest.

The result is a measure of time-varying activity for each trial for the four brain regions that were identified as relevant to visual attention or WM: V1, L-IPS, L-DLPFC, R-DLPFC.

3-2-2 Trial-to-Trial Variability

Theoretically, each trial within a dataset captures the same process, e.g. recall of an encoded memory which was undisturbed during the maintenance period. However, many sources of variation between the trials exist which are not experimentally controlled. For example, stray thoughts unrelated to the task, recall of other memories, and regulation of other body processes can all vary between trials, and it is unreasonable to assume that the specified brain regions are not involved in these processes. Figure 3-2, showing six trials within the 'no distractor' dataset, illustrates how severe this variability is in time-domain analysis.

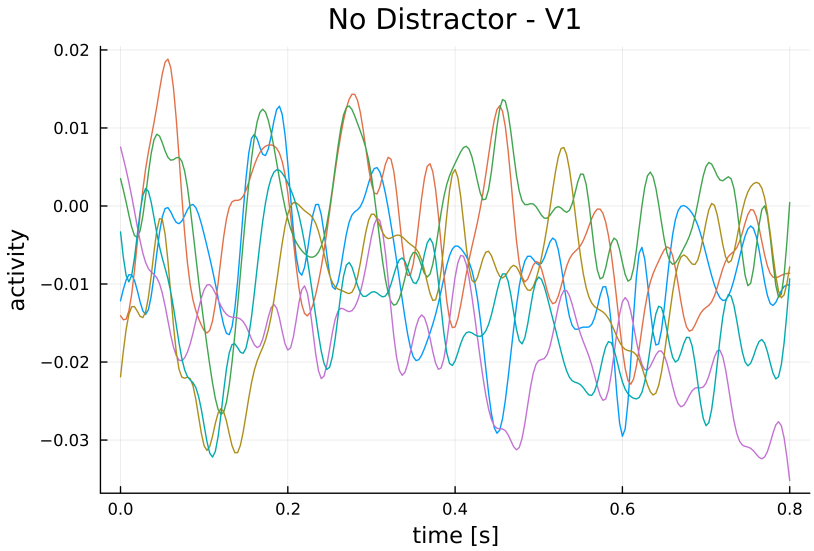


Figure 3-2: Six trials in the 'no distractor' dataset for brain region V1. Variation in the time-domain is very high, and a consistent signal is difficult to identify through visual inspection.

Trial-to-trial variability must therefore be accounted for in some way. There are several ways this can be achieved. First, if the potential sources of variability are well understood, they can be explicitly modelled. For example, it is common in MEG studies to allow for subject-to-subject variability during analysis and when using underlying models [47, 90]. However,

in this study the sources of variability are not well understood as only a single subject is considered and all environmental variables are controlled.

Second, the variability can be captured by a noise term [68, 32]. This method can be used when the sources of variability are not well understood. However, learning an appropriate noise term from the datasets can be very challenging, especially when the distribution of the data is non-normal (e.g. multi-modal, skewed). Because this method is challenging, it is preferred when capturing the variability is important to the research question at hand.

Finally, an appropriate measure of central tendency can be used as a representative value. This method can be used when the characteristics of the data distributions are not important. Instead, a calculation such as the mean or median can be used to determine a typical value based on the data distribution, removing the trial-to-trial variations [80]. Many model building studies for MEG and electroencephalography (EEG) take this approach, condensing repeated trials to a single representative datapoint and fitting a model to it [31, 70, 22]. In contrast to the second method, this should be used when data variability is an undesired consequence of experimental limitations and only characteristics common to all trials are of interest. In the following section, we explain and elaborate on the approach taken in this project.

3-2-3 Frequency-domain Analysis

Frequency-domain analysis of the source-reconstructed activity is the focus of this study, as both the original study on this dataset [120] and a large body of literature (see Section 1-2-2) support that modulation of oscillatory dynamics are important to visual attention processes.

It is known that frequency characteristics of neuroimaging measurements such as EEG and MEG are generally non-stationary [1]. This non-stationarity means that time-frequency analysis such as Morlet wavelet transforms [51] and sliding-window fourier transforms [120, 50] are popular methods for studying induced neuronal responses. Although it is similarly likely that the frequency characteristics of this recall phase dataset are not stationary, the short length of the phase (0.8 s) and the low frequencies which are of interest in neuronal dynamics prevent the use of any time-varying frequency analysis. Instead, the full 0.8 s are used to determine one frequency power spectrum which reflects the entire recall phase. The spectrum from 5 to 25 Hz is investigated. This frequency range is suitably large to comfortably capture the alpha-band modulation identified in the original study, as well as lower frequency modulation associated with WM in existing literature. The lower bound is restricted to allow for at least four cycles in the time-domain window. The upper bound is restricted to avoid analysis of gamma-range modulation, given its controversial role in attention dynamics (see Section 1-2-2).

A Hanning window is first applied to mitigate the effects of spectral leakage. A fast Fourier transform (FFT) is then applied to each trial. The trial duration of 0.8 s and the sampling rate of 300 Hz gives a frequency resolution of about 1.2 Hz. The amplitudes of the frequency components are determined by taking the magnitude of the FFT output and this is squared to give a representation of the power at each frequency. The full dataset is then normalised by the maximum power over all frequencies and trial types.

Visual inspection of this frequency data shows substantial trial-to-trial variability, most of the points fitting a unimodal skewed distribution. For example, Figure 3-3 shows the distribution

of the power at 10 Hz in V1 in the 'distractor' dataset. The distribution is heavily skewed toward lower values, with some higher outliers, and a peak at slightly under 0.1. As previously mentioned, this variability could be caused by a host of factors which are impossible to experimentally control and are not important to the 'distractor'/'no distractor' distinction. However, a shift in the distributions of the two datasets would suggest that despite the variability, there exists a meaningful difference in the power spectrum under the two conditions. This difference would be well captured by a measure of central tendency of the two distributions. In this case of skewed unimodal distributions, the median is more appropriate than the more commonly used mean, as the mean is disproportionately impacted by the outliers.

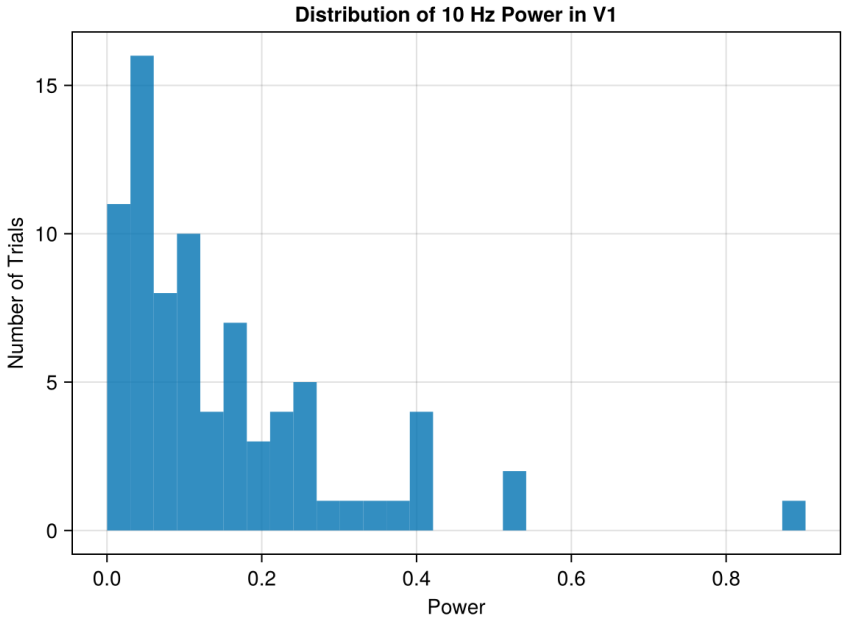


Figure 3-3: Histogram of the power at 10 Hz in V1 'distractor' trials. The distribution is heavily skewed towards lower powers, making the median the best measure of central tendency. Created using Makie [27].

Figure 3-4 below shows the median power over frequencies for Region 1 - V1. The red line denotes the 'no distractor' dataset median and the blue line the 'distractor' dataset median. In addition to the median, the inter-quartile range (IQR) which contains the middle 50% of the data, is denoted by the shaded regions, giving an indication of the full distribution. Note the increase in the 8-11 Hz band between the 'no distractor' and "distractor cases", circled in green, clear in both the median and the shift in IQR. This suggests that there is a significant increase in the lower alpha-band power in the 'distractor' compared to the 'no distractor' dataset, and that the median is representative of this increase. This observation is consistent with the previous results from the original study [120].

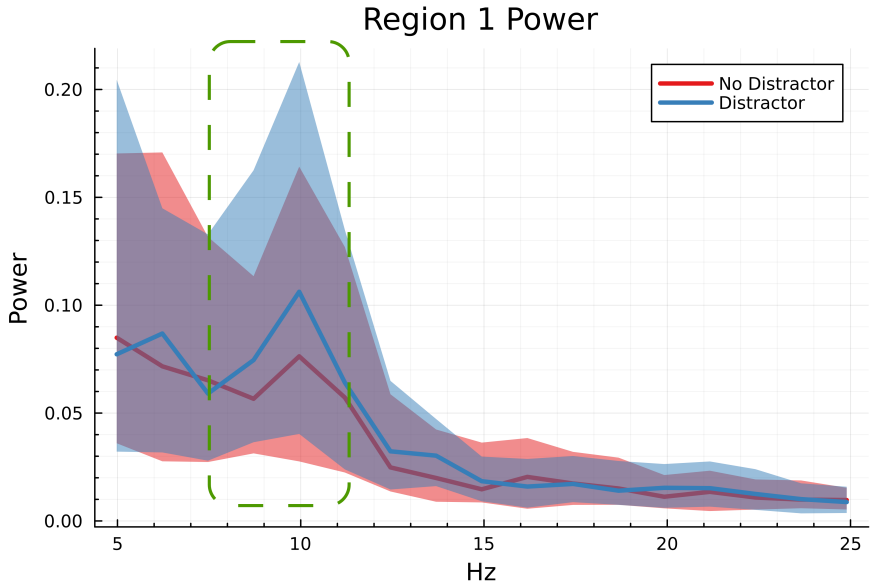


Figure 3-4: Region 1 - V1, IQR and median of the power distribution for the two datasets. Circled in green is the increased alpha-band power (here 8-11 Hz) in the 'distractor'.

Results from the other three regions are provided in Figures 3-5, 3-6, 3-7. Regions 2 and 3 (L-IPS and L-DLPFC) show increases in power similar to that seen in Region 1, but in different frequency ranges. Some less pronounced high frequency modulation can also be observed. Region 4 (R-DLPFC) shows very little change between the two cases.

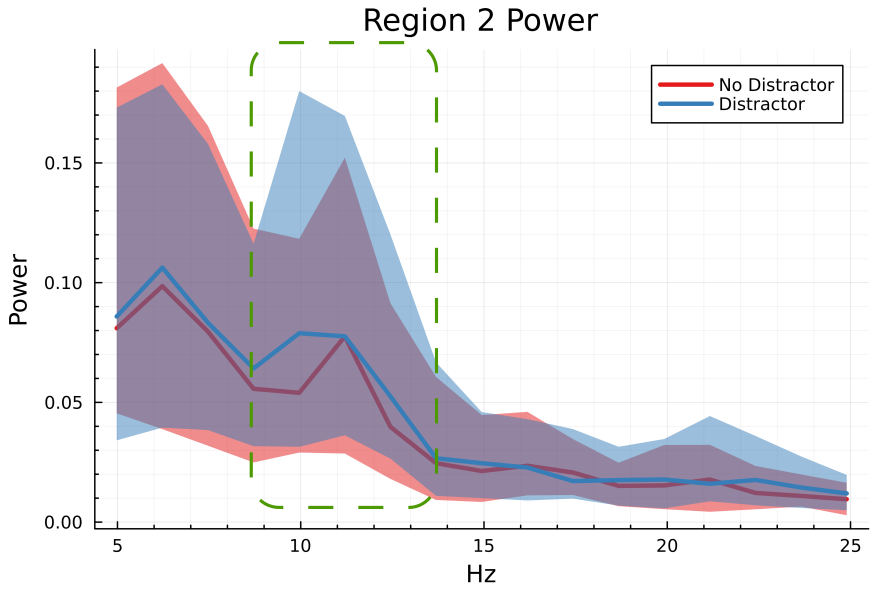


Figure 3-5: Region 2 - L-IPS, IQR and median of the power distribution for the two datasets. Circled in green is a clear increase in the 9-14 Hz power in the 'distractor' case.

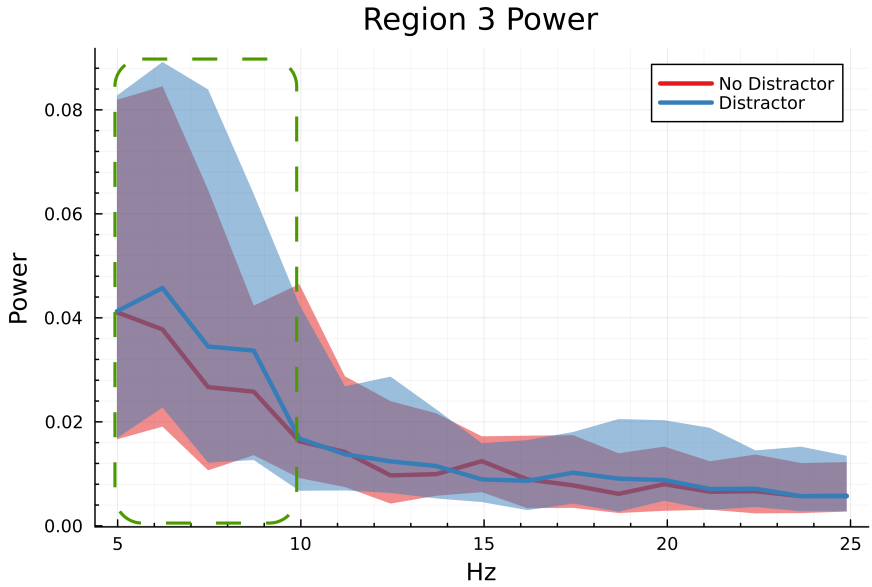


Figure 3-6: Region 3 - L-DLPFC, IQR and median of the power distribution for the two datasets. Circled in green is a clear increase in 5-10 Hz power in the 'distractor' case.

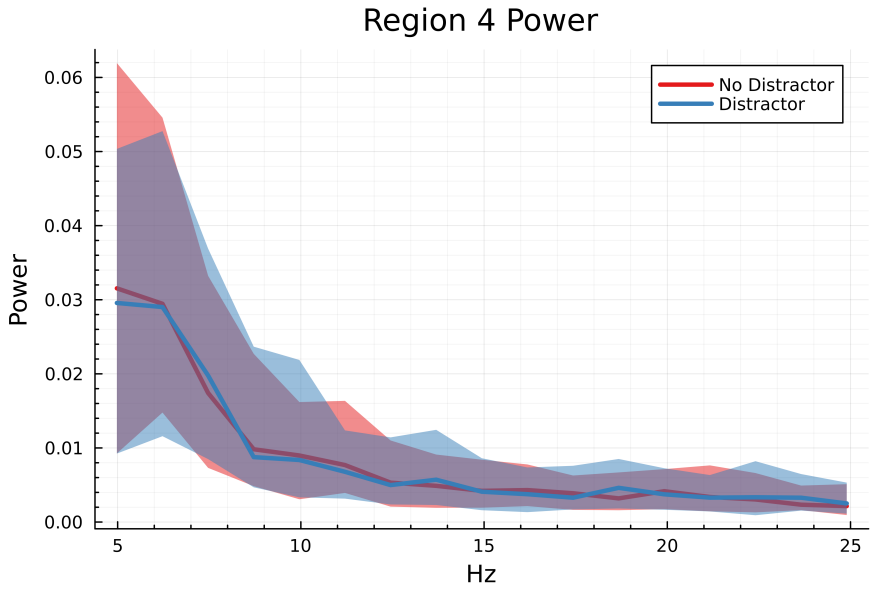


Figure 3-7: Region 4 - R-DLPFC, IQR and median of the power distribution for the two datasets. Behaviour is largely similar between the two cases.

Overall, this analysis demonstrates that there are significant differences between the power distributions of the 'distractor' and 'no distractor' datasets in most regions of interest (with the exception of Region 4). This is shown clearly by the shift in IQR in certain frequency ranges. The experimental design means that these differences are likely caused by the activation of attention mechanisms in the 'distractor' trials. It is therefore of interest to neuroscientists to determine which neural mechanisms cause this change in behaviour, as this would suggest how the attention mechanisms are implemented in the brain.

Here, the issue of trial-to-trial variability, as discussed in Section 3-2-2, becomes relevant. The large range of the IQRs show that this variability is significant, as is expected for repeated MEG recordings. Three methods for handling this variability have been presented in Section 3-2-2, two of which are appropriate for this experimental design. First, the variability can be captured by a stochastic term which must be learned from data. As discussed previously, this increases the difficulty of the modelling problem significantly. Second, an appropriate measure of central tendency can be used as a representative of the distribution. For this dataset, the median is identified as the appropriate measure. Comparison of the medians to the IQRs in Figures 3-4 to 3-7 reveals that the medians are, in general, a good representation of the difference in power spectra distribution in the two cases. Therefore, the medians of both datasets can be considered a representation of the typical behaviour of the single subject under the 'distractor' and 'no distractor' conditions. This allows the use of a deterministic model to represent the dynamics under the two conditions, greatly simplifying the modelling task.

Chapter 4

Oscillatory Network Model for Attention-Working Memory Interplay

This chapter outlines the data-driven modelling algorithm designed in this project. The algorithm includes both the structure and parameterisation of the model and the methods used to train the parameters. A general model which captures the full experiment is described in Section 4-1. This general model is broken into two modules, one baseline model and one model for capturing the attention mechanism. These models are described in Sections 4-2 and 4-3, respectively.

4-1 Data-Driven Model Design

This model is designed to replicate the power spectra of the four brain regions analysed in Chapter 3 under the 'distractor' and 'no distractor' conditions. The change in dynamics between these two conditions is assumed to be caused by activation of attention mechanisms in the 'distractor' condition. Analysis of these dynamics can then provide neuroscientists with an indication of how these mechanisms may be implemented in the brain, with respect to interaction between these four regions.

4-1-1 Model Structure

To capture the interaction dynamics between the four regions of interest, the state s is defined to represent the activity in these regions,

$$s = \begin{bmatrix} s_1 \\ s_2 \\ s_3 \\ s_4 \end{bmatrix} = \begin{bmatrix} \text{V1} \\ \text{L-IPS} \\ \text{R-DLPFC} \\ \text{L-DLPFC} \end{bmatrix}. \tag{4-1}$$

The output of the system is then directly related to these states, as data for the activity of these brain regions is available from the source reconstruction process described in Chapter 3, that is

$$y_i = \lambda(s_i). \tag{4-2}$$

As a reminder, the working memory (WM) task can be broken into three phases: encoding, maintenance and recall. All experimental manipulation between the 'distractor' and 'no distractor' conditions occurs in the first two periods. In the beginning of the trial, during the initial encoding phase, the subject is given a cue which indicates whether to expect distractors in the trial. During the maintenance phase, distractors appear (or not) as expected. During the recall phase, however, the subject is probed to recall the target memory in exactly the same manner across the two conditions. Thus, the change in power spectra discussed in Chapter 3 must be due to an internal change in dynamics which is triggered by earlier experimental changes. This internal change is assumed to be the activation of attention mechanisms.

The recall period of the experiment is the focus of this study. There is no stimulus input during this phase, with the exception of the recall probe at the beginning of the period which is treated as an initial condition perturbation. To capture the internal change in dynamics in the model, the dynamics are split into a baseline model and an attention model,

$$\dot{s} = \overbrace{l(s)}^{\text{baseline}} + \nu \overbrace{g(s)}^{\text{attention}}, \qquad \nu \in \{0, 1\}. \tag{4-3}$$

The attention dynamics are (de)activated by the binary parameter ν.

During the 'no distractor' trials, the attention mechanisms are not active ($\nu = 0$) because no distracting input is expected. Only the baseline dynamics are active, giving,

$$\dot{s} = l(s). \tag{4-4}$$

This baseline model represents the brain under baseline recall conditions, and existing literature can inform the structure of this function. Selection and parameterisation of the baseline model is discussed in Section 4-2.

During the 'distractor' trials, the attention mechanisms are active ($\nu = 1$) because a distracting input is expected. This gives the dynamics

$$\dot{s} = l(s) + g(s). \tag{4-5}$$

The function $g(s)$ represents the change in dynamics which occurs when attention is internally activated. The mechanism for attention is currently unknown, so there exists significantly less literature to inform the structure of this function. Selection and parameterisation of the attention model is discussed in Section 4-3.

4-1-2 Loss Function for Model Parameter Fitting

Specific parametersations of the model defined in (4-3) are provided in the following sections, but no matter the specifics the goal of the modelling algorithm is to find parameters of the model such that the model output matches characteristics observed in the datasets. This is achieved by defining a loss function which captures the distance between the model output and the data and minimising this loss with respect to the parameters.

Denote the power spectrum calculated from the 'no distractor' dataset as described in Chapter 3 P_{mn}. Denote P_{md} as the 'distractor' dataset equivalent. These are the data characteristics the model must capture. Denote $Y_{\nu=0}, Y_{\nu=1}$ as the time-domain outputs of the model under attention inactive and active conditions, simulated from 0-0.8 s at 300 Hz just as the time-domain data is formatted. These outputs are produced from a numerical ordinary differential equation (ODE) solver,

$$Y_{\nu=\{1,0\}} = \text{ODESolver}(\dot{s} = l(s) + \nu g(s), y_i = \lambda(s_i), s(0), t). \tag{4-6}$$

Let the function $P(Y) = |FFT(Y)|^2$ denote the power spectrum of Y calculated as the squared magnitude of the FFT from 5-25 Hz. This function transforms the time-domain output of the model to a measure equivalent to the data metrics P_{md}, P_{mn}. The distance between the data and the model output is then the loss function,

$$\mathcal{L} = ||P(Y_{\nu=1}) - P_{md}||_2^2 + ||P(Y_{\nu=0}) - P_{mn}||_2^2. \tag{4-7}$$

4-2 Baseline Model

There is no generally accepted model for neuronal population activity as recorded by magnetoencephalography (MEG). The baseline model, then, must be selected from a candidate set which encompasses neuronal population and MEG models which exist in the literature.

Based on the general model described in Section 4-1, the baseline model dynamics are active under both the 'distractor' ($\nu = 1$) and 'no distractor' ($\nu = 0$) conditions, and are the *only* dynamics active in the 'no distractor' condition. These dynamics should, then, be able to replicate the behaviour of the 'no distractor' dataset very well. Thus, the candidate function which achieves the smallest loss with respect to the 'no distractor' dataset is selected as the baseline model.

4-2-1 Candidate Functions

Many dynamical models exist to represent the activity of neuronal populations or to replicate behaviours seen in neuroimaging recordings. In a data-driven modelling context, these models trade off expressive power, i.e. flexibility to be able to represent the data well, simplicity, i.e. ease of training due to the loss landscape shape, and biological intepretability, i.e. biophysical

versus phenomenological structures. Models based on populations of spiking neurons tend to be more representative of biological structures, but their highly nonlinear nature makes them difficult to fit to data without a reasonable initial guess for the parameters. On the other hand, more abstract oscillator models can fit neuroimaging data well, but can be difficult to map to biological processes.

To capture this range, three candidate models are selected. The linear harmonic oscillator is one of the simplest systems that exhibits oscillatory behaviours and has been successfully used to replicate neuroimaging recordings [73]. The Duffing-Van der Pol oscillator is a combination of two well-known nonlinear oscillators, and was developed to better replicate the frequency spectra seen in electroencephalography (EEG) recordings [50]. The Wilson-Cowan model is a staple of neuronal population dynamics, reflecting known interactions between different types of neuronal populations which give rise to oscillatory behaviours [115]. Table 4-1 below summarises the characteristics of the chosen models.

Oscillator Model	Expressive Power	Simplicity	Biological Interpretability
Linear Harmonic	Medium	High	Low
Duffing-Van der Pol	High	Medium	Low
Wilson-Cowan	High	Low	Medium

Table 4-1: Characteristics of the three candidate baseline models.

Note that this is not meant to be an exhaustive list, but rather gives a reasonable coverage of the existing literature where the three metrics are concerned. For example, models with higher biological interpretability than the Wilson-Cowan models do exist, but are not considered because their very high complexity makes fitting them to data infeasible.

Of these candidates, the linear harmonic oscillator model achieves the lowest loss. However, it does so by capturing only the dynamics of a limited frequency range, see Section 5-1 for more detail. To improve the overall fit, two extensions to the model, which aim to increase the expressive power without too much cost to complexity, are considered. The double node harmonic oscillator and double network harmonic oscillator both introduce additional states to each region, but retain the basic linear oscillator structure. They differ primarily in the connectivity between the oscillators.

Detailed descriptions of the original three candidate functions and the two extensions of the linear harmonic oscillator are given below.

Linear Harmonic Oscillator

One of the most basic models for replicating oscillatory behaviours is the linear harmonic oscillator. As discussed in Section 1-2-2, this model has been used in data-driven neuroimaging algorithms but studies have shown that the linearity assumption limits the expressive power of the model, and that a complex network of oscillators is often required to capture complex behaviours.

Each region $i \in 1, \ldots, 4$ contains a self-sustaining linear harmonic oscillator with a unique frequency ω. The network is fully connected via linear diffusive coupling with strength k. This gives the dynamical model,

$$\dot{s}_i = \overbrace{\begin{bmatrix} 0 & 1 \\ -\omega_i^2 & 0 \end{bmatrix} s_i}^{\text{harmonic oscillator}} + \overbrace{\begin{bmatrix} 0 \\ \sum_{\substack{j=1 \\ j\neq i}}^{4} k_{ij}(s_{j1} - s_{i1}) \end{bmatrix}}^{\text{linear diffusive coupling}}, \tag{4-8}$$

$$s_i = [s_{i1}, s_{i2}]^T \in \mathbb{R}^2, s = [s_1, s_2, s_3, s_4]^T. \tag{4-9}$$

The model outputs are the first state of each oscillator,

$$y = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} s. \tag{4-10}$$

The model has four natural frequencies and 12 coupling parameters, giving a total of 16 parameters. In addition to this, there are eight states requiring eight initial conditions, meaning the full optimisation problem has 24 decision variables.

Duffing-Van der Pol Oscillator

This Duffing-Van der Pol oscillator is an extension of a nonlinear oscillator designed to replicate the wide-band power spectra of EEG signals [50] (see Section 1-3 for details).

Each region $i \in 1, \ldots, 4$ contains a self-sustaining Van der Pol oscillator with unique parameters l, μ, which together determine the uncoupled frequency of the regions. The network is fully connected via Duffing oscillator coupling. This coupling includes a diffusive linear term with strength k plus a diffusive cubic term with strength b. This gives the dynamical model,

$$\dot{s}_i = \begin{bmatrix} \dot{s}_{i1} \\ \dot{s}_{i2} \end{bmatrix} = \overbrace{\begin{bmatrix} s_{i2} \\ \mu_i s_{i2}(1 - s_{i1}^2) - l_i s_{i1} \end{bmatrix}}^{\text{Van der Pol oscillator}} + \overbrace{\begin{bmatrix} 0 \\ \sum_{\substack{j=1 \\ j\neq i}}^{4} k_{ij}(s_{j1} - s_{i1}) + b_{ij}(s_{j1} - s_{i1})^3 \end{bmatrix}}^{\text{Duffing coupling}}, \tag{4-11}$$

$$s_i = [s_{i1}, s_{i2}]^T \in \mathbb{R}^2, s = [s_1, s_2, s_3, s_4]^T. \tag{4-12}$$

The model outputs are the first state of each oscillator, as in (4-10). When fitting this model to the data, it was observed that in general, the magnitude of power this model can generate is significantly smaller than the data. To address this, a single scaling parameter is introduced. This additional parameter does not compromise the validity of the model because the data power being fit to has been normalised to the maximum power observed in the dataset, and thus has no physical significance. The primary interest is the relative power across regions and frequencies, which is preserved after introduction of the scalar.

The model has eight Van der Pol parameters and 24 coupling parameters, giving a total of 32 parameters. In addition to this, there are eight states requiring eight initial conditions, plus the power scaling parameter, meaning the full optimisation problem has 41 decision variables.

Wilson-Cowan Oscillator

The Wilson-Cowan oscillator is a classical neuronal population model [115, 116], which also acts as a nonlinear oscillator. The original formulation of the model includes many parameters, making optimisation challenging. Instead, a simplified model based on the deterministic part of the model used in [111], is utilised here.

Each region $i \in 1, \ldots, 4$ includes an excitatory population, represented by s_{i1}, and an inhibitory population, represented by s_{i2}. The oscillation of each region emerges from the interaction between the two populations. As the terms excitatory and inhibitory suggest, excitatory populations can only excite or increase the activity of connected populations, while inhibitory populations can only inhibit or decrease the activity. The inhibitory populations connect only to themselves and to the excitatory population within the same region; this occurs through a non-diffusive linear term with strength b. The excitatory populations also connect to the local inhibitory populations through a non-diffusive linear term with strength l. These external inputs are summed and pass through a logistic sigmoid function $\mathcal{S}$. To form the greater network, the excitatory populations are fully connected through a non-diffusive linear term with strength k. This gives the dynamical model,

$$\dot{s}_{i1} = -\alpha_{i1} s_{i1} + (1 - s_{i1})\beta_{i1}\mathcal{S}\Big(\overbrace{\sum_{j=1}^{4} k_{ij} s_{j1}}^{\text{network excitatory}} - \overbrace{b_{i1} s_{i2}}^{\text{local inhibitory}} + h_{i1} \Big), \tag{4-13}$$

$$\dot{s}_{i2} = -\alpha_{i2} s_{i2} + (1 - s_{i2})\beta_{i2}\mathcal{S}\Big(\underbrace{l_i s_{i1}}_{\text{local excitatory}} - \underbrace{b_{i2} s_{i2}}_{\text{local inhibitory}} + h_{i2} \Big), \tag{4-14}$$

$$\mathcal{S}(s) = \frac{1}{1 + e^{-a(s-\theta)}}, \tag{4-15}$$

$$s_i = [s_{i1}, s_{i2}]^T \in \mathbb{R}^2, s = [s_1, s_2, s_3, s_4]^T, \tag{4-16}$$

where $\alpha_i, \beta_i, k_{i,j}$ are positive constant parameters denoting the coefficients and coupling weights, respectively. Each population also receives a constant input h_i, representing background input from unmodelled regions. The parameters of the sigmoid function are set at $a = 1, \theta = 4$.

MEG records the activity of excitatory populations, so the model outputs are the first state of each region, as in (4-10). As in the Duffing-Van der Pol oscillator, a scalar parameter for the power is introduced.

This model has 16 internal dynamics parameters and 36 coupling parameters, giving a total of 52 parameters. In addition to this, there are eight states requiring eight initial conditions, plus the power scaling parameter, meaning the full optimisation problem has 61 decision variables.

Double Node Harmonic Oscillator

A schematic of the double node harmonic oscillator is given in Figure 4-1, where each region is modelled as a set of two interconnected harmonic oscillators, each with a distinct natural frequency. This extension preserves the linearity of the model, while increasing its expressive power by increasing the number of autonomous oscillators in the network, allowing coverage of a wider frequency range.

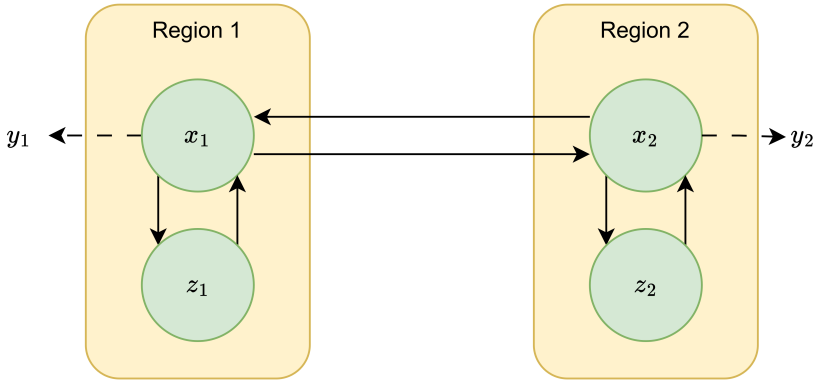


Figure 4-1: A two-region schematic of the connectivity of the double node oscillator network. An additional oscillator exists within each region, but connects only to the primary oscillator within the region. The model output is taken from the primary oscillator only.

Each region $i \in 1, \ldots, 4$ consists of two oscillators, represented by states x_i and z_i. The oscillator x_i is fully connected to other x oscillators via linear diffusive coupling, as in the linear oscillator network. An additional coupling of strength b to the second oscillator within the region is also present. The oscillator z_i is connected only to the second oscillator within the region via a coupling of strength l. This gives the dynamical model,

$$\dot{x}_i = \overbrace{\begin{bmatrix} 0 & 1 \\ -\omega_i^2 & 0 \end{bmatrix} x_i}^{\text{harmonic oscillator}} + \overbrace{\begin{bmatrix} 0 \\ \sum_{\substack{j=1 \\ j \neq i}}^{4} k_{ij}(x_{j1} - x_{i1}) \end{bmatrix}}^{\text{cross-region diffusive coupling}} + \overbrace{\begin{bmatrix} 0 \\ b_i(z_{i1} - x_{i1}) \end{bmatrix}}^{\text{within-region coupling}}, \tag{4-17}$$

$$\dot{z}_i = \underbrace{\begin{bmatrix} 0 & 1 \\ -\phi_i^2 & 0 \end{bmatrix} z_i}_{\text{harmonic oscillator}} + \underbrace{\begin{bmatrix} 0 \\ l_i(x_{i1} - z_{i1}) \end{bmatrix}}_{\text{within-region coupling}}, \tag{4-18}$$

$$x_i = [x_{i1}, x_{i2}]^T \in \mathbb{R}^2, x = [x_1, x_2, x_3, x_4]^T, \tag{4-19}$$

$$z_i = [z_{i1}, z_{i2}]^T \in \mathbb{R}^2, z = [z_1, z_2, z_3, z_4]^T \tag{4-20}$$

$$s = [x, z]^T. \tag{4-21}$$

The model outputs are the first state of the fully connected oscillator of each region, as in (4-10).

The model has eight natural frequencies, eight within-region coupling parameters, 12 network

coupling parameters, giving a total of 28 parameters. In addition to this, there are 16 states requiring 16 initial conditions, meaning the full optimisation problem has 44 decision variables.

Double Network Harmonic Oscillator

The double network harmonic oscillator, like the double node, has an additional linear oscillator in each region. The architecture, as shown in Figure 4-2, differs from the previous one as the oscillators in each region are not directly connected. In fact, each linear oscillator in the region is fully connected to corresponding, i.e. low or high frequency, linear oscillators in the other regions, forming two fully connected networks. The training algorithm is designed such that one of the networks captures low-frequency behaviours and the other captures high-frequency behaviours. There also exists cross-coupling between the networks, but not within the regions. This extension preserves the linearity of the model, while increasing its expressive power by increasing the number of autonomous oscillators in the network and allowing for more complex connectivity. Although the number of parameters is very large, this model allows for the design of a multi-stage training algorithm which helps to narrow down the parameters search space. Details of this algorithm are outlined in Section 4-2-2.

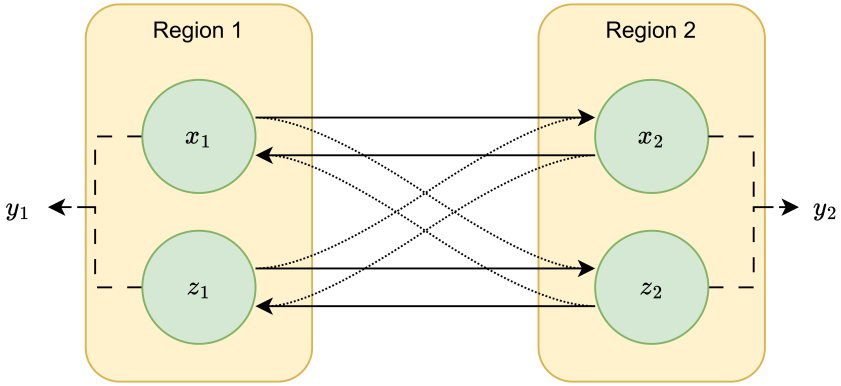


Figure 4-2: A two-region schematic of the connectivity of the double network oscillator network. The solid black lines denote coupling for the two fully connected networks. There also exists cross coupling between the networks, denoted by the fine dotted lines. The model output is the sum of the primary and secondary oscillators.

Each region $i \in 1, \ldots, 4$ consists of two oscillators, represented by states x_i and z_i. Linear diffusive coupling within the networks x and z have strengths k and c, respectively. Influence of the z network on the x network occurs via linear diffusive coupling of strength b, and vice-versa with strength l. This gives the dynamical model,

$$\dot{x}_i = \overbrace{\begin{bmatrix} 0 & 1 \\ -\omega_i^2 & 0 \end{bmatrix} x_i}^{\text{harmonic oscillator}} + \overbrace{\begin{bmatrix} 0 \\ \sum_{\substack{j=1 \\ j\neq i}}^{4} k_{ij}(x_{j1} - x_{i1}) \end{bmatrix}}^{\text{within-network coupling}} + \overbrace{\begin{bmatrix} 0 \\ \sum_{\substack{j=1 \\ j\neq i}}^{4} b_{ij}(z_{j1} - x_{i1}) \end{bmatrix}}^{\text{cross coupling}}, \tag{4-22}$$

$$\dot{z}_i = \underbrace{\begin{bmatrix} 0 & 1 \\ -\phi_i^2 & 0 \end{bmatrix} z_i}_{\text{harmonic oscillator}} + \underbrace{\begin{bmatrix} 0 \\ \sum_{\substack{j=1 \\ j\neq i}}^{4} c_{ij}(z_{j1} - z_{i1}) \end{bmatrix}}_{\text{within-network coupling}} + \underbrace{\begin{bmatrix} 0 \\ \sum_{\substack{j=1 \\ j\neq i}}^{4} l_{ij}(x_{j1} - z_{i1}) \end{bmatrix}}_{\text{cross coupling}}, \tag{4-23}$$

$$x_i = [x_{i1}, x_{i2}]^T \in \mathbb{R}^2, \qquad x = [x_1, x_2, x_3, x_4]^T, \tag{4-24}$$

$$z_i = [z_{i1}, z_{i2}]^T \in \mathbb{R}^2, \qquad z = [z_1, z_2, z_3, z_4]^T \tag{4-25}$$

$$s = [x, z]^T. \tag{4-26}$$

The model outputs are the sum of the first states of the two oscillators in each region,

$$y = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} (x + z). \tag{4-27}$$

The model has eight natural frequencies, 24 within-network coupling parameters, and 24 cross coupling parameters, giving a total of 56 parameters. In addition to this, there are 16 states requiring 16 initial conditions, meaning the full optimisation problem has 72 decision variables.

4-2-2 Training

For selection of the baseline model $\dot{s} = l(s)$ from the set of candidate functions, only the 'no distractor' dataset is used. The parameters of each candidate function are trained in the optimisation problem,

$$\min_{s(0),\theta} \mathcal{L}_{\nu=0} \tag{4-28}$$

$$\mathcal{L}_{\nu=0} = ||P(Y_{\nu=0}) - P_{mn}||_2^2 \tag{4-29}$$

$$Y_{\nu=0} = \text{ODESolver}(\dot{s} = l(\theta, s), y_i = \lambda(s_i), s(0), t). \tag{4-30}$$

The loss function is adapted from (4-7) to include only the 'no distractor' case and l, λ, θ are the process and output functions, and parameters, of each of the candidate functions listed above.

For the original three candidate functions (linear harmonic, Van der Pol and Duffing oscillators) as well as the double node harmonic oscillator model, this problem is solved in two stages. First, a global optimisation algorithm, the multi-level single linkage (MLSL) method

[97], is employed to search a broad area in the parameter space. This method requires upper and lower bounds on all parameters. It is challenging to determine these bounds such that they encompass all reasonable operating modes of the models but are not so large that the global search becomes infeasible. This is decided based on existing studies which utilise these models, an understanding of the role the parameters play in the models (e.g. the explicit frequency parameters in the oscillators), and model simulations. The MLSL method works by completing many local optimisations from a number of random starting points within the parameters search space, using a clustering technique to make the global search more efficient. The sequential least squares quadratic programming (SLSQP) algorithm is used for local optimisation. Although some theoretical guarantees for optimality do exist for this algorithm, time restrictions prevented the optimisation from running to conclusion, so the best parameter set from a time-limited search is selected. This parameter set is then further refined using the Adam optimiser [67].

The large number of states and parameters in the double network harmonic oscillator makes global optimisation in this case quite challenging. Therefore, several additional optimisation stages are developed to fit this model. First, the high- and low-frequency networks are independently fit to the high- and low-frequency portions of the power spectrum. For the low-(high-)frequency fit, the 15-25 (5-15) Hz portion of the data is zeroed, and parameter fitting proceeds with the aforementioned MLSL followed by Adam optimisation. To fit the parameters of the full network, including both networks and cross-coupling, the results from the previous optimisations are used as initial conditions, with cross-coupling initialised at zero. This allows a local optimisation algorithm, here Adam, to be used, as the parameters are already nearby a reasonable solution.

For the optimisation of all the candidate functions, algorithms are used which require the gradient of the loss function with respect to the parameters. When ODE solver operations are included in the loss function, calculating these gradients can be challenging, as discussed in Section 2-1-1. However, for the baseline case all candidate functions have a sufficiently small number of states and parameters so that forward-mode autodifferentiation is sufficient to compute these gradients.

The double network harmonic oscillator achieves the smallest loss, and is therefore selected as the baseline model. Results and comparison of the candidate model outputs are provided in the following chapter.

4-3 Attention Model

As discussed in more detail in Chapter 1, hypotheses for the neural mechanisms of attention vary widely, making selection of a model structure for the attention model challenging. Rather than choosing a subset of these hypotheses and comparing them to each other, a function identification technique can be employed to learn the attention model directly from data. A neural network N is used to represent the attention model, making the complete model a universal differential equation (UDE) as described in Section 2-1. Note that the double network harmonic oscillator is selected as the baseline model l, which gives the dyanmics

$$\begin{bmatrix} \dot{x} \\ \dot{z} \end{bmatrix} = l(x,z) + \nu N(x,z) \tag{4-31}$$

$$x_i = [x_{i1}, x_{i2}]^T \in \mathbb{R}^2, \qquad x = [x_1, x_2, x_3, x_4]^T, \tag{4-32}$$

$$z_i = [z_{i1}, z_{i2}]^T \in \mathbb{R}^2, \qquad z = [z_1, z_2, z_3, z_4]^T. \tag{4-33}$$

To remain consistent with the structure of the baseline model, the structure of the neural network must be restricted. In the baseline model, each neural population is captured by a two-state oscillator, with the first state representing the oscillator position and the second the velocity. The oscillators are second-order, with no damping terms, meaning the dynamics of the velocity states include only terms from the position states. All populations are fully connected, with the exception of populations within the same region.

To maintain this pattern in the neural network, it is decomposed into eight separate networks, each with a single output corresponding to the dynamics of a different velocity term x_{i2} or z_{i2}. Each of these networks takes as input all of the position oscillator terms of the populations, except for the term corresponding to the oscillator in the same region. For example, the neural network describing the attention dynamics for low-frequency Region 1, N_{x1}, has as input $x_{11}, x_{21}, x_{31}, x_{41}, z_{21}, z_{31}, z_{41}$. In total, this gives the dynamics,

$$\dot{x}_i = l_{xi}(x,z) + \nu \begin{bmatrix} 0 \\ N_{xi}(x_{j1}, z_{k1}) \end{bmatrix}, j \in \{1,2,3,4\}, k \in \{1,2,3,4\} \setminus i, \tag{4-34}$$

$$\dot{z}_i = l_{zi}(x,z) + \nu \begin{bmatrix} 0 \\ N_{zi}(x_{j1}, z_{k1}) \end{bmatrix}, j \in \{1,2,3,4\} \setminus i, k \in \{1,2,3,4\}. \tag{4-35}$$

Each neural network has input dimension seven and output dimension one. There are four hidden layers of dimension 15, each with a ReLU activation function. Interestingly, during tuning of the network it was observed that the ReLU activation function lead to significantly faster convergence of the network compared to logistic sigmoid or tanh functions.

4-3-1 Training

With the model fully parameterised, training proceeds by solving the full optimisation problem,

$$\min_{s(0),\theta,\phi} \mathcal{L} \tag{4-36}$$

$$\mathcal{L} = ||P(Y_{\nu=0}) - P_{mn}||_2^2 + ||P(Y_{\nu=1}) - P_{md}||_2^2 \tag{4-37}$$

$$Y_{\nu=0} = \text{ODESolver}(\dot{s} = l(\theta, s), y_i = \lambda(s_i), s(0), t) \tag{4-38}$$

$$Y_{\nu=1} = \text{ODESolver}(\dot{s} = l(\theta, s) + N(\phi, s), y_i = \lambda(s_i), s(0), t). \tag{4-39}$$

This includes both the 'distractor' and 'no distractor' datasets, the parameters of the baseline model θ and the parameters of the neural network ϕ. The baseline model parameters are

initialised at the results from the previous baseline fit, ensuring that the 'no distractor' dataset is well represented. The neural network parameters are initialised near zero. Because the parameters are initialised well, the optimisation algorithm Adam is used to determine the optimal parameters.

The Adam optimiser requires the gradient of the loss function with respect to the parameters. As discussed in Section 2-1, calculation of these gradients is challenging for UDEs. This one in particular contains 5000 parameters, making direct autodifferentiation infeasible. An improved version of the backsolve adjoint sensitivity method discussed in Section 2-1-1, called the Gauss adjoint method, is used. This method improves both computational and memory efficiency compared to autodifferentiation and backsolve adjoint sensitivity for this problem.

The result of the optimisation is a baseline model $l(\theta^*, s) = l^*(s)$ which captures the nominal WM recall dynamics, and an neural network $N(\phi^*, s) = N^*(s)$ which captures the change in these dynamics due to activation of attention mechanisms.

4-3-2 Symbolic Regression

To determine what neural mechanisms the function $N^*(s)$ is describing, the symbolic regression technique sparse identification of nonlinear dynamics (SINDy), as described in Section 2-2, is used to construct an arithmetic expression which captures the input-output behaviour of $N^*(s)$.

The first step in this process is the creation of the input-output dataset. This is generated by using the trajectory of the 'distractor' ($\nu = 1$) condition as the input set, and passing these points through the network function,

$$S_{in} = \text{ODESolver}(\dot{s} = l^*(s) + N^*(s), s^*(0), t), \tag{4-40}$$
$$= [X_{in}, Z_{in}], \tag{4-41}$$
$$Y_{out} = N^*(S_{in}). \tag{4-42}$$

This selection of the input set guarantees that the full input-output set captures the most important relationship learned by the neural network, as this is the trajectory that network is trained on.

The next step is the selection of the library of candidate functions. This library includes a monomial expansion of the network inputs from degree two to degree five, plus a collection of diffusive coupling terms as is used in the baseline model. The network function N^* is a collection of eight separate networks, each with a different set of inputs. This structure means that the library of candidate functions is also different for each of the separate networks, and that the sparse optimisation must be completed separately for each network. As an example, the library for the network N^*_{x1} is

$$\Theta_{x1}(S) = \Theta_{x1}(X, Z) \tag{4-43}$$
$$= \begin{bmatrix} Z_{2,3,4} - X_1 & X_{2,3,4} - X_1 & X^2_{1,2,3,4} & \dots & X^5_{1,2,3,4} & Z^2_{2,3,4} & \dots & Z^5_{2,3,4} \end{bmatrix}, \tag{4-44}$$

including transformations of the same inputs as the function N_{x1}. The sparse identification problem is then

$$\min_{\Xi} ||\Xi||_0, \tag{4-45}$$

$$||Y_{out} - \Theta(S_{in})\Xi||_2 < \epsilon, \tag{4-46}$$

which is solved using the sequential thresholding least-squares (STLSQ) algorithm as discussed in Section 2-2. The result of solving all eight sparse identification problems is a new function which is roughly equivalent to the network function but with significantly fewer terms and parameters,

$$N^*(x, z) \approx SR(p, x, z), \tag{4-47}$$

where p are the nonzero elements of the sparse vector Ξ^*. This gives the final model structure,

$$\begin{bmatrix} \dot{x} \\ \dot{z} \end{bmatrix} = l(\theta, x, z) + \nu SR(p, x, z) \tag{4-48}$$

$$x_i = [x_{i1}, x_{i2}]^T \in \mathbb{R}^2, \qquad x = [x_1, x_2, x_3, x_4]^T, \tag{4-49}$$

$$z_i = [z_{i1}, z_{i2}]^T \in \mathbb{R}^2, \qquad z = [z_1, z_2, z_3, z_4]^T, \tag{4-50}$$

where p are the parameters of the symbolic regression model and θ are the parameters of the baseline model. To ensure optimality of the parameters, the optimisation problem for the full dataset defined in (4-39) is repeated with this model structure,

$$\min_{s(0),\theta,p} \mathcal{L} \tag{4-51}$$

$$\mathcal{L} = ||P(Y_{\nu=0}) - P_{mn}||_2^2 + ||P(Y_{\nu=1}) - P_{md}||_2^2 \tag{4-52}$$

$$Y_{\nu=0} = \text{ODESolver}(\dot{s} = l(\theta, s), y_i = \lambda(s_i), s(0), t) \tag{4-53}$$

$$Y_{\nu=1} = \text{ODESolver}(\dot{s} = l(\theta, s) + SR(p, s), y_i = \lambda(s_i), s(0), t). \tag{4-54}$$

Chapter 5

Results and Analysis

In the previous chapter, the data-driven modelling algorithm designed to learn dynamics describing the attention-working memory (WM) interplay was presented. Sections 5-1 and 5-2 outline the results of the application of this algorithm to the single-subject experimental data described in Chapter 3.

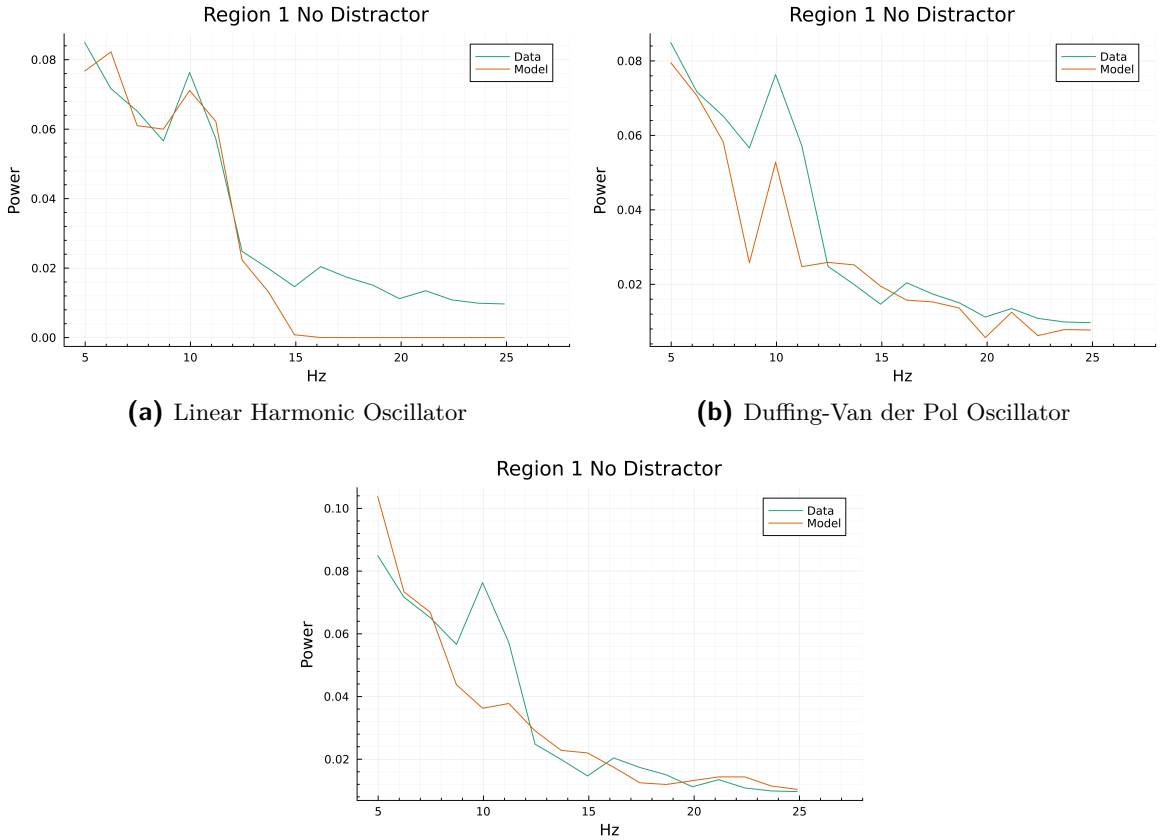
To put these results into a larger context, results from a hypothesis-driven approach to modelling this dataset are presented in Section 5-3. The application of two different algorithms to the same dataset allows for a discussion of the relative merits of the two approaches, demonstrating the scenarios in which the new method may be preferred.

Finally, to verify the results of the single-subject model, the data-driven modelling algorithm is applied to a second subject dataset. Analysis of the second subject model and a comparison to the first subject model are presented in Section 5-4.

5-1 Baseline Model

In this section, results from the fitting of the candidate baseline model functions to the 'no distractor' dataset are presented. Based on these results, the double network harmonic oscillator model is selected as the baseline model, representing undisturbed recall dynamics.

A comparison of the data and model outputs for the original three candidate models, Linear Harmonic, Duffing-Van der Pol, Wilson-Cowan, is shown in Figure 5-1. See Appendix Figures C-1 to C-3 for Regions 2, 3 and 4. Of these three candidate models, the Linear Harmonic achieves the lowest loss. Visual inspection of the fit shows that the Linear Harmonic model captures the low-frequency dynamics well, but the high-frequency behaviour (above 13 Hz) is nearly absent. In contrast, both nonlinear models capture the high-frequency behaviour more accurately, but fail to capture the low-frequency. This suggests that the nonlinearities help to capture wide-spectrum behaviours, but can also make optimisation of parameters challenging.



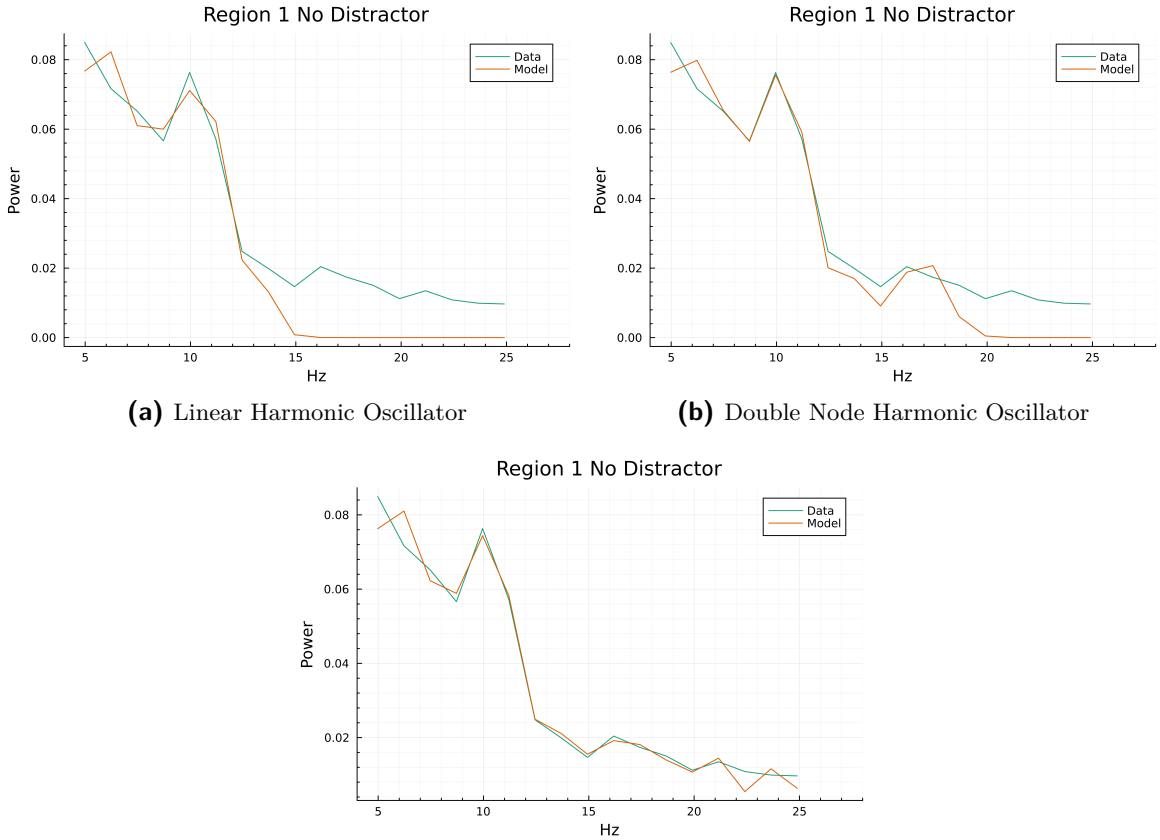
(a) Linear Harmonic Oscillator

(b) Duffing-Van der Pol Oscillator

(c) Wilson-Cowan Oscillator

Figure 5-1: Fit of the original three candidate baseline models to to the 'no distractor' dataset for Region 1 - primary visual cortex (V1).

To maintain linearity while increasing expressive power, an additional oscillator is added to each node of the Linear Harmonic oscillator. The results are the Double Node and Double Network Harmonic oscillator models discussed in Section 4-2, which differ in their connectivity. A comparison of the data and model outputs for the Linear Harmonic model and its extensions is shown in Figure 5-2. See Appendix Figures C-1 to C-3 for Regions 2, 3 and 4. The Double Node model succeeds in improving the high-frequency fit, but only up to about 18 Hz. The Double Network model, with more complex connectivity, achieves the lowest loss of all the candidate functions, replicating the data well for the full 5-25 Hz range.



(a) Linear Harmonic Oscillator

(b) Double Node Harmonic Oscillator

(c) Double Network Harmonic Oscillator

Figure 5-2: Fit of the extended linear oscillator models to the 'no distractor' dataset for Region 1 - V1.

This selection of baseline model has interesting implications for the development of modelling algorithms to replicate spectral magnetoencephalography (MEG) data. It suggests that linear models can be preferred in data-driven modelling scenarios despite the fact that most evidence suggests that neuronal population dynamics are nonlinear. This may be due to the simpler loss landscape resulting from the use of more linear models, which simplifies solving of the optimisation problem. However, the high dimensionality and number of parameters required in the linear model necessitates the use of a multi-stage fitting algorithm to determine the optimal parameters.

This result also confirms previous studies which suggest that a network of linear oscillators can be used to replicate frequency-domain MEG and electroencephalography (EEG) data but that a large number of oscillators are required to replicate wide spectra [49]. It would be interesting to investigate whether the introduction of more data characteristics to the loss function would impact this result. It could be possible, for example, that capturing the time-domain characteristics of MEG signals in addition to the frequency-domain requires a prohibitive number of linear oscillators, forcing the move to nonlinear models.

5-2 Attention Model

In this section, results from the fitting of the complete model, including both baseline and attention dynamics, to both the 'no distractor' and 'distractor' datasets are presented. This fitting process occurs in two stages. First, a neural network is used to represent the attention model, and a baseline plus neural network model is fit to the full dataset. These results are presented in Section 5-2-1. Then, symbolic regression techniques are used to convert the neural network to arithmetic expressions, and this model is again fit to the full dataset. These results are presented in Section 5-2-2.

5-2-1 Neural Network

The first step in the function identification of the attention model is the training of the neural network, with the model

$$\begin{bmatrix} \dot{x} \\ \dot{z} \end{bmatrix} = l(\theta, x, z) + \nu N(\phi, x, z), \tag{5-1}$$

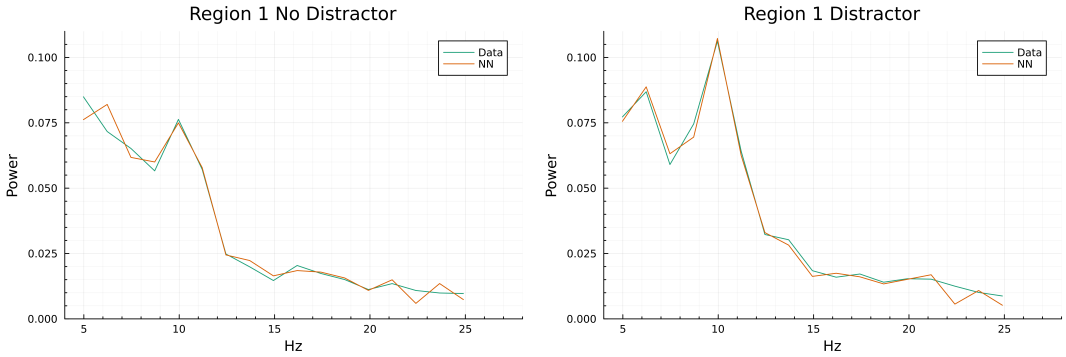
where l is the double network harmonic oscillator with parameters θ and N is a feedforward neural network with parameters ϕ. A reminder that the state is defined as

$$x_i = [x_{i1}, x_{i2}]^T \in \mathbb{R}^2, \qquad x = [x_1, x_2, x_3, x_4]^T, \tag{5-2}$$

$$z_i = [z_{i1}, z_{i2}]^T \in \mathbb{R}^2, \qquad z = [z_1, z_2, z_3, z_4]^T \tag{5-3}$$

where x are the low-frequency oscillators and z are the high-frequency oscillators. The regions correspond to the state indices as 1 - V1, 2 - L-intraparietal suculus (IPS), 3 - L-dorsolateral (DL)prefrontal cortex (PFC), 4 - R-DLPFC.

A comparison of the data and model output for Region 1 (V1) is shown in Figure 5-3. See the Appendix Figures C-4 to C-6 for the results for Region 2, 3, and 4. Visual inspection of the fit demonstrates that the baseline model captures the 'no distractor' behaviour well, as is expected from the baseline model selection process. Together, the baseline plus the neural network also capture the 'distractor' behaviour well, suggesting that the network can replicate the attention dynamics.



(a) No Distractor, $\nu = 0$ **(b)** Distractor, $\nu = 1$

Figure 5-3: Fit of the universal differential equation (UDE), with baseline double extended linear oscillator and attention neural network models, to the full dataset for Region 1 - V1.

5-2-2 Symbolic Regression

The next step in the function identification of the attention model is the distillation of the neural network into an arithmetic expression through symbolic regression. The resulting expression is given in (5-4),

$$N(\phi, x, z) \approx SR(p, x, z)$$

$$= \begin{bmatrix} 0 \\ p_1(z_{21} - x_{11}) + p_2(z_{41} - x_{11}) + p_3 x_{41}^2 + p_4 z_{21}^2 + p_5 z_{31}^2 + p_6 x_{41}^3 \\ 0 \\ p_7 z_{41}^2 \\ 0 \\ p_8 x_{41}^2 + p_9 x_{41}^3 \\ 0 \\ p_{10}(z_{11} - x_{41}) + p_{11}(z_{31} - x_{41}) + p_{12} z_{31}^2 \\ 0 \\ p_{13}(z_{41} - z_{11}) + p_{14} x_{41}^2 + p_{15} z_{11}^2 + p_{16} z_{21}^2 + p_{17} z_{31}^2 + p_{18} z_{41}^2 + p_{19} z_{41}^3 \\ 0 \\ p_{20}(z_{31} - z_{21}) + p_{21} x_{41}^2 + p_{22} z_{41}^2 + p_{23} x_{41}^3 + p_{24} z_{41}^5 \\ 0 \\ p_{25} z_{41}^2 \\ 0 \\ p_{26} z_{31}^2 \end{bmatrix}. \quad (5\text{-}4)$$

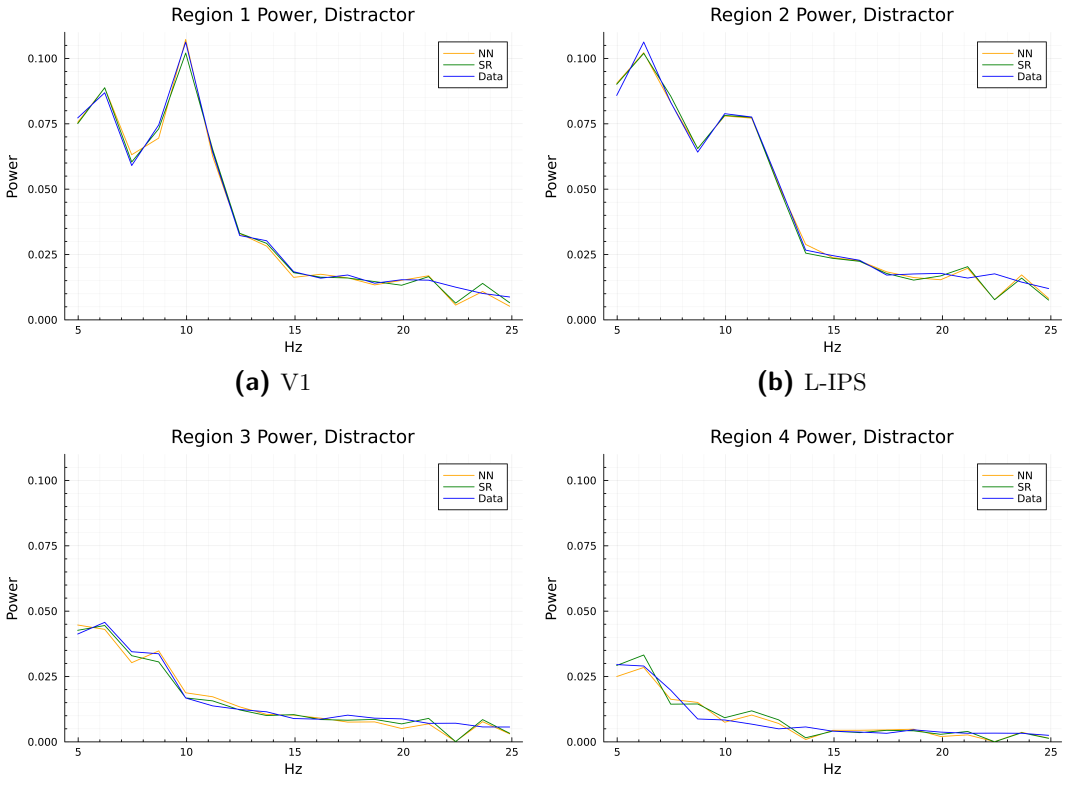
Even without considering the specific parameter values, analysis of which terms appear in each expression is an effective way to determine the significant relationships between states, as the sparse optimisation algorithm used for symbolic regression prevents insignificant terms from appearing in the final solution. This analysis is provided with some context later in the section.

Finally, the full model including the symbolic regression attention model,

$$\begin{bmatrix} \dot{x} \\ \dot{z} \end{bmatrix} = l(\theta, x, z) + \nu SR(p, x, z), \tag{5-5}$$

is optimised with respect to the full dataset to produce a final model with optimal parameters θ^*, p^*.

A comparison of the data, neural network model, and symbolic regression model for all four regions is shown in Figure 5-4. The close agreement between the neural network and symbolic regression models demonstrates that the symbolic regression technique is able to reduce the highly parametersed network into a simpler expression which captures the same behaviours.

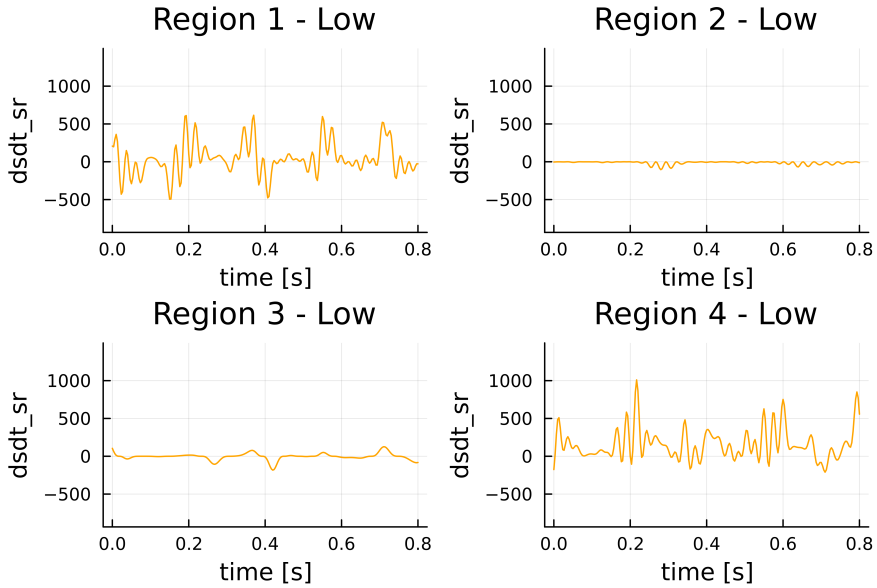


(a) V1 **(b)** L-IPS

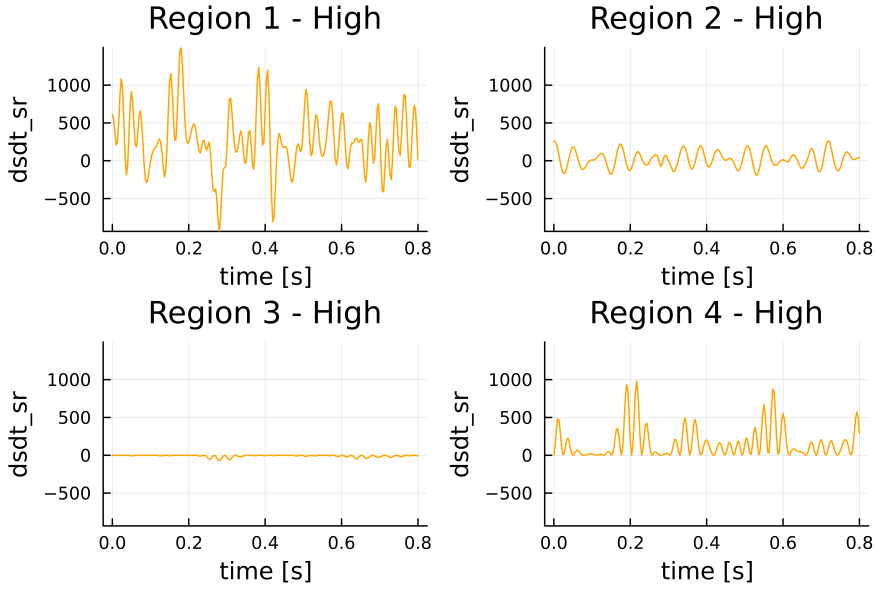
(c) L-DLPFC **(d)** R-DLPFC

Figure 5-4: Comparison between the symbolic regression and neural network models in capturing the 'distractor' dataset.

To aid interpretation of the symbolic regression model to provide some insight into the underlying attention mechanisms, Figure 5-5 shows the value of the nonzero terms of $SR(x(t), z(t))$ evaluated over the solutions $x(t), z(t)$ of the full model. These terms indicate the impact of the attention mechanism, here represented by the function $SR(x, z)$, on the dynamics of the oscillators over the course of the solution trajectory. In particular, visualisation of these terms allows for an assessment of which states in the model are most impacted by the activation of the attention model.



(a) Terms influencing low-frequency dynamics ($SR_{xi}, i = 1, 2, 3, 4$).



(b) Terms influencing high-frequency dynamics ($SR_{zi}, i = 1, 2, 3, 4$).

Figure 5-5: $SR(x(t), z(t))$ evaluated over the solutions $x(t), z(t)$ of the full model.

A first observation made clear from this visualisation is that Region 1 (V1) experiences

significant changes in dynamics with the activation of the attention mechanisms, as both the high- and low-frequency SR function terms for this region are large in magnitude over the course of the solution trajectory. This aligns well with existing literature in visual attention, much of which focuses on changes in behaviour of the visual cortex [120, 109, 36, 63]. The expressions for this region,

$$SR_{x_1}(x,z) = p_1(z_{21} - x_{11}) + p_2(z_{41} - x_{11}) + p_3x_{41}^2 + p_4z_{21}^2 + p_5z_{31}^2 + p_6x_{41}^3, \tag{5-6}$$
$$SR_{z_1}(x,z) = p_{13}(z_{41} - z_{11}) + p_{14}x_{41}^2 + p_{15}z_{11}^2 + p_{16}z_{21}^2 + p_{17}z_{31}^2 + p_{18}z_{41}^2 + p_{19}z_{41}^3, \tag{5-7}$$

confirm this observation, containing many terms representing complex inputs from other regions. Terms are coloured to indicate which regions they correspond to, with V1: orange, L-IPS: olive, L-DLPFC: pink and R-DLPFC: blue. The majority of terms in both the high- and low-frequency expressions correspond to the R-DLPFC (blue), suggesting that input from this region is important to the change in behaviour of V1 when distractors are anticipated. This observation is especially interesting considering that the data analysis of R-DLPFC reveals very minimal changes in power spectrum of this region between the 'distractor' and 'no distractor' cases (see Figure 3-7). Thus, although data analysis suggests that the R-DLPFC is not active in attention mechanisms, this dynamical analysis suggests that the region plays an important role in controlling other regions.

In addition to the R-DLPFC, terms relating to high-frequency L-IPS (olive) and L-DLPFC (pink) also appear, suggesting that the inputs from these regions to V1 also change with activation of attention mechanisms. Parameters p_5, p_{16}, p_{17} in particular are substantial in magnitude compared to the other parameters, in addition to p_{19}, p_{14} which link with Region 4.

The high-frequency node of Region 1 is also the only node to show an internal change in dynamics, with the relevant term highlighted in orange. This suggests that in addition to a change in influence from the other three regions in the network, there are internal changes that occur within V1 when attention mechanisms are activated.

A second observation from Figure 5-5 is that Regions 2 and 3 (L-IPS and L-DLPFC) have relatively little change in dynamics, as both the high- and low-frequency SR function terms for this region are relatively small in magnitude over the course of the solution trajectory, with the exception of high-frequency Region 2. The expressions for these regions,

$$SR_{x_2}(x,z) = p_7z_{41}^2, \tag{5-8}$$
$$SR_{x_3}(x,z) = p_8x_{41}^2 + p_9x_{41}^3, \tag{5-9}$$
$$SR_{z_2}(x,z) = p_{20}(z_{31} - z_{21}) + p_{21}x_{41}^2 + p_{22}z_{41}^2 + p_{23}x_{41}^3 + p_{24}z_{41}^5, \tag{5-10}$$
$$SR_{z_3}(x,z) = p_{25}z_{41}^2, \tag{5-11}$$

once again confirm these observations, containing relatively few terms. The colour-coded terms again demonstrate the importance of input from Region 4 (R-DLPFC) to Regions 2 and 3 in the attention model, especially for the high-frequency Region 2 expression (z_2). However, this result for expressions x_2, x_3, z_3 should be considered with some caution, as

these outputs of the SR function are very small. These small outputs suggest that the input-output relationship for these terms is fairly weak, making the symbolic regression techniques less reliable.

Finally, although data analysis has shown that the behaviour of Region 4 (R-DLPFC) does not change much with activation of attention mechanisms, Figure 5-5 demonstrates that there are significant changes in the underlying dynamics. The expressions for Region 4,

$$SR_{x_4}(x,z) = p_{10}(z_{11} - x_{41}) + p_{11}(z_{31} - x_{41}) + p_{12}z_{31}^2, \tag{5-12}$$
$$SR_{z_4}(x,z) = p_{26}z_{31}^2, \tag{5-13}$$

demonstrate that these are largely due to changes in influence from Region 3 (L-DLPFC), highlighted in pink, plus a change in the diffusive coupling with V1. Note that although the expression for z_4 has only a single term, the parameter p_{26} has significant magnitude, meaning the dynamics of z_4 are still substantially altered.

In summary, the symbolic regression model suggests that activation of attention mechanisms has a large impact on the dynamics of V1 and R-DLPFC and a small impact on the dynamics of L-DLPFC and L-IPS. Many of the changes in dynamics are due to input from the R-DLPFC, which is itself impacted through input from the L-DLPFC.

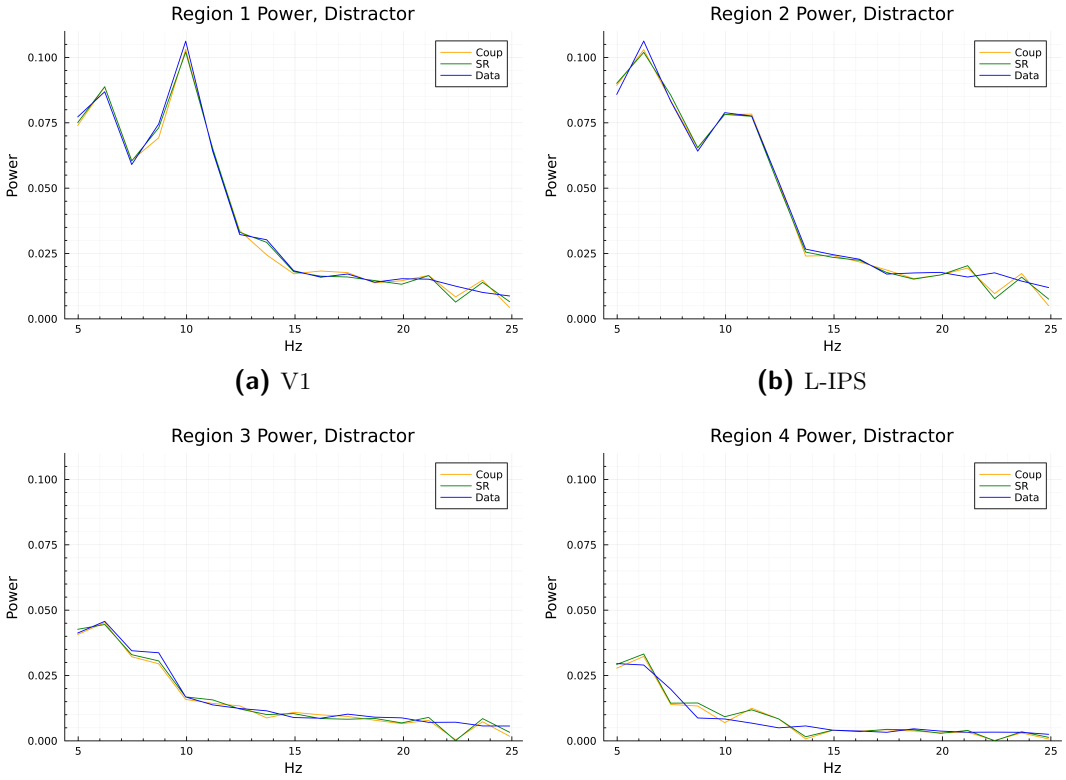
5-3 Comparison with Connectivity-Based Attention Model

In this section, results from a second modelling strategy are presented as a comparison to the algorithm developed in this project. This second strategy is hypothesis-driven, meaning apriori assumptions on the attention mechanism are used to define the model structure, rather than letting the model structure be learned from data. A common strategy in data-driven modelling for MEG, often used in dynamic causal modelling (DCM), is to account for changes in experimental conditions by allowing a change in the coupling between neuronal populations (see Section 1-3). In this experimental context, this equates to describing the activation of the attention mechanism as a change in the coupling strength between the populations. This gives the dynamical model,

$$\begin{bmatrix}\dot{x}\\ \dot{z}\end{bmatrix} = l(\theta, x, z) + \nu CO(\Delta k, x, z), \tag{5-14}$$
$$= l([\omega, k + \nu\Delta k], x, z), \tag{5-15}$$

where the function $CO(\Delta k, x, z)$ contains the coupling terms from the baseline function l with the change in coupling parameters Δk. This can be equivalently written as in (5-15), with the baseline parameters θ split into the natural frequencies ω and the coupling parameters k.

A comparison of the data, data-driven symbolic regression model, and hypothesis-driven coupling change model, is shown in Figure 5-6. The hypothesis-driven model is able to capture the behaviour of the dataset just as well as the symbolic regression model, achieving very similar loss function values.



(a) V1 **(b)** L-IPS

(c) L-DLPFC **(d)** R-DLPFC

Figure 5-6: Comparison between the symbolic regression and coupling change models in capturing the 'distractor' dataset.

Because the structure and parameters in the hypothesis-driven model are motivated by a biological process, i.e. strength of synaptic connections between brain regions, it is easier to analyse the optimal parameters here than in the symbolic regression model. A schematic of the most significant changes in coupling identified in the model is given in Figure 5-7. A change in coupling is deemed significant if the coupling value in the baseline model is above a given threshold and the change in that value in the attention model is greater than 5%.

Although this schematic alone offers insight into the neural mechanisms being captured by this model, differing model structures makes comparison of the symbolic regression model to these results challenging. To aid in the comparison, Figure 5-8 shows the value of the nonzero terms of $CO(x(t), z(t))$ evaluated over the solutions $x(t), z(t)$ of the full model. This is analogous to Figure 5-5 for the symbolic regression model.

Together, Figures 5-7 and 5-8 suggest that the coupling change model is describing a neural mechanism which is completely different from the one described in the symbolic regression model. Figure 5-8 demonstrates that the change in dynamics is spread evenly across the four regions, with roughly equal magnitude output from the attention model for all populations. This is contrary to Figure 5-5, which suggests that changes in the dynamics are concentrated to specific populations. Figure 5-7 demonstrates that all oscillators experience a significant incoming or outgoing change in coupling strength. This supports the observation that all regions experience a change in dynamics, while also suggesting that these changes are due to

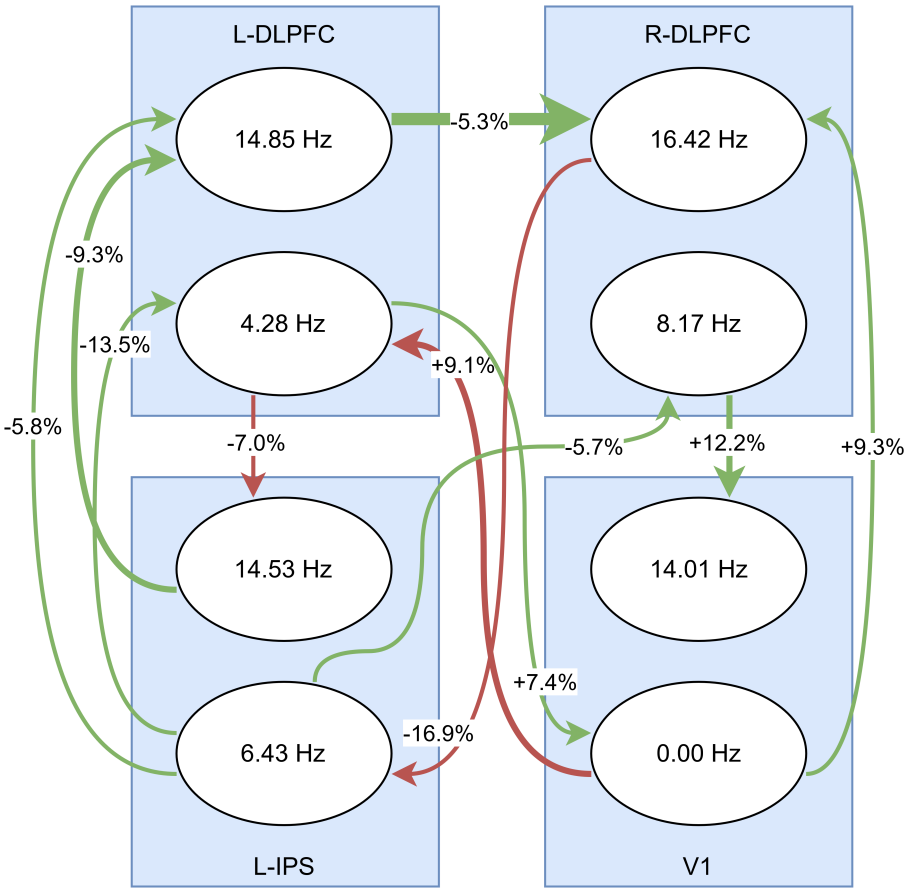
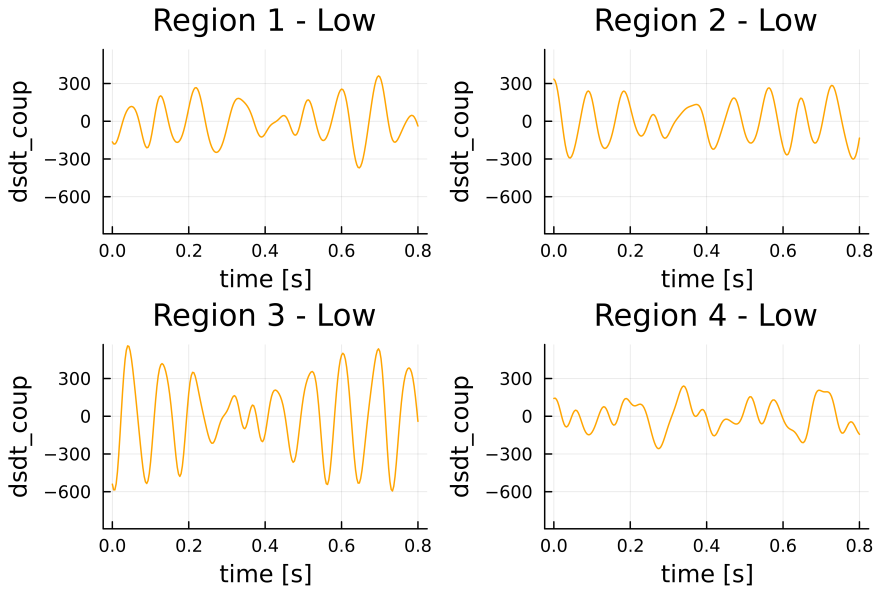


Figure 5-7: Coupling between neural populations which change significantly with activation of attention mechanisms. Green arrows indicate positive (excitatory) coupling, red arrow indicate negative (inhibitory) coupling. The weight of the arrow reflects the strength of the coupling in the baseline model, and the percentage change in the coupling strength is written on each arrow.

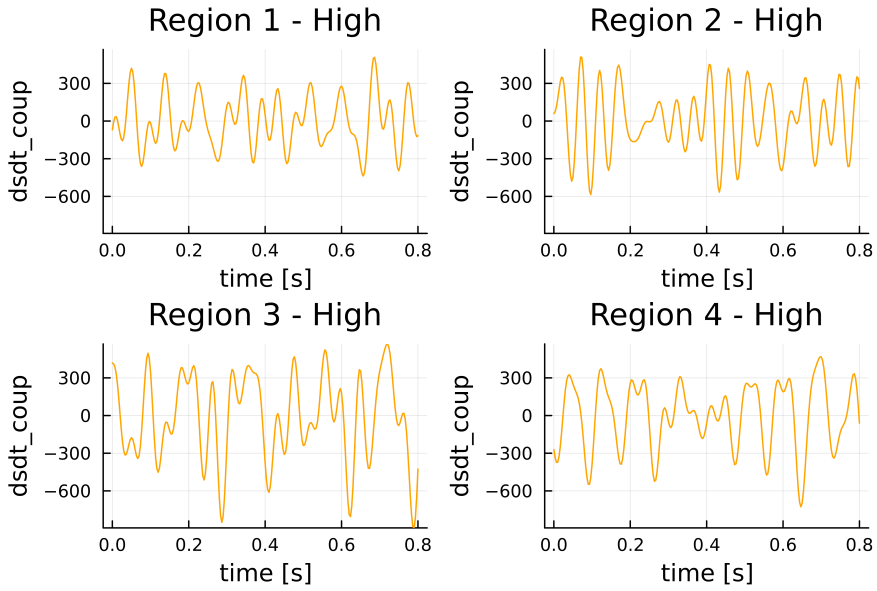
inputs from many different regions. Again, this is contrary to the analysis of the symbolic regression model, which suggests that the changes in dynamics can be connected to inputs from specific regions. In short, this model suggests that these attention mechanisms are not implemented through focused changes in the relationships within and between targeted regions, but rather through a more general change in connectivity in the network.

The aim of this comparison is to raise awareness for the caution required when drawing conclusions based on studies which utilise a single modelling framework. Even with the same baseline model, these two modelling approaches suggest completely different neural mechanisms to neuroscientists looking to utilise the modelling results. With the currently selected data metric of median power spectrum, there is no way to differentiate between these two models, as they achieve similar loss function values. Thus, from a data point of view, it is unclear which model better represents the underlying neural mechanism.

However, this comparison also highlights the usefulness of the developed modelling algorithm for exploratory analysis. By allowing for more flexibility in the learning of the attention model, the symbolic regression model suggests a new mechanism for attention dynamics which en-



(a) Terms influencing low-frequency dynamics ($CO_{xi}, i = 1, 2, 3, 4$)



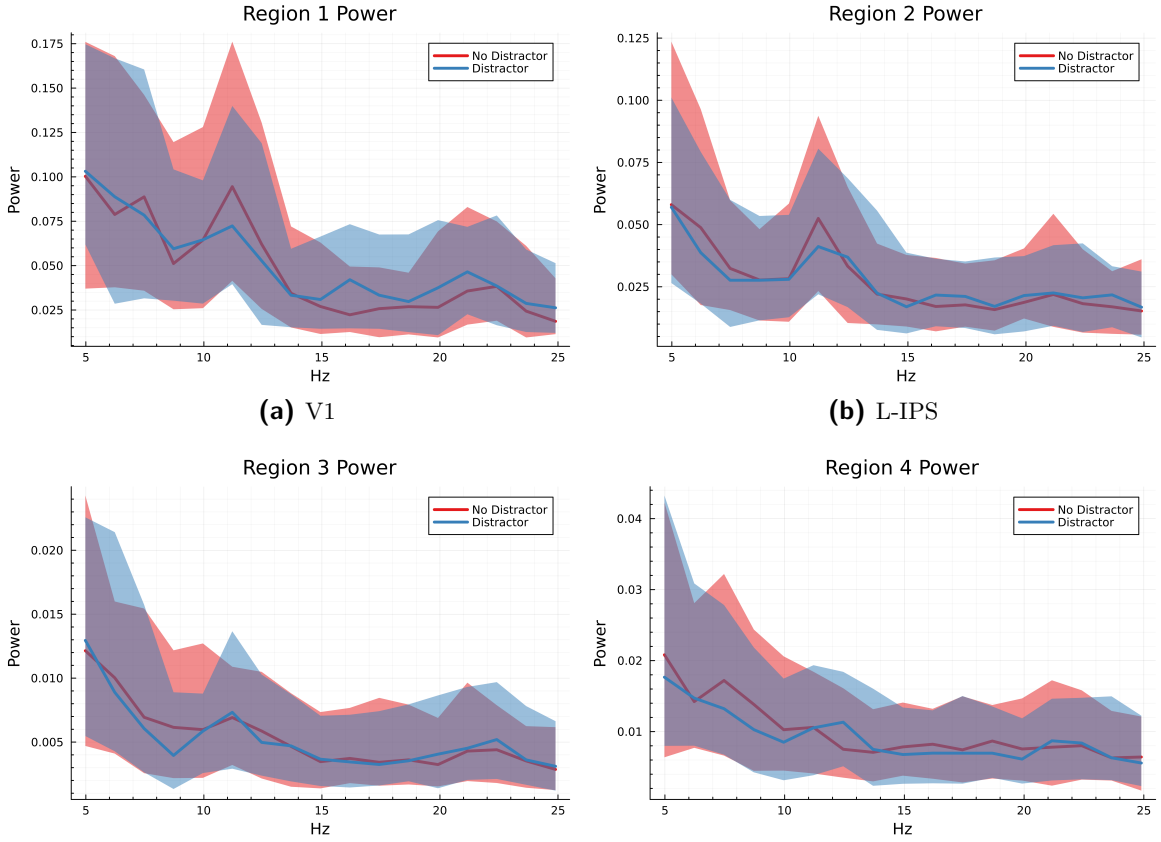
(b) Terms influencing high-frequency dynamics ($CO_{zi}, i = 1, 2, 3, 4$)

Figure 5-8: $CO(x(t), z(t))$ evaluated over the solutions $x(t), z(t)$ of the full model.

compasses both changes in coupling strength and more general changes in dynamic behaviour. This generality allows the model to capture mechanisms which may not be understood yet in the scientific community, and thus would not otherwise be captured by a hypothesis-driven model.

5-4 Comparison With Second Subject Model

To test some of the observations made based on the single-subject model, a second model is built using the same data-driven fitting algorithm on a dataset from a second subject in the same study. A visualisation of this dataset (see Figure 5-9) shows very different behaviour from that observed in the original subject dataset. For example, alpha-band power in V1 of this subject decreases rather than increases under the 'distractor' condition, contrary to the original subject and to most observations in the literature. The increase in power is instead seen at higher frequencies, from about 16-20 Hz.



(a) V1 **(b)** L-IPS

(c) L-DLPFC **(d)** R-DLPFC

Figure 5-9: Frequency-domain analysis for second subject.

The algorithm could achieve similar loss levels to those seen for the original dataset (supporting figures are given in Appendix Figure C-10). Despite the differences in data behaviour, analysis of the second subject model confirms the majority of observations made for the first model.

The symbolic regression model which represents the changes in dynamics when the attention module is activated is given in (5-17). Terms are coloured to indicate which regions they correspond to, with V1: orange, L-IPS: olive, L-DLPFC: pink and R-DLPFC: blue. Also provided in Figure 5-10 is a plot of the attention module function $SR(x, z)$ as a function of the solution, analogous to Figure 5-5.

$$N2(\phi, x, z) \approx SR2(p, x, z) \tag{5-16}$$

$$= \begin{bmatrix} 0 \\ p_1(z_{31} - x_{11}) + p_2(z_{41} - x_{11}) + p_3 x_{11}^2 + p_4 z_{41}^2 + p_5 z_{41}^3 + p_6 x_{41}^4 + p_7 z_{41}^5 \\ 0 \\ p_8 x_{31}^2 \\ 0 \\ p_9 z_{41}^2 + p_{10} z_{41}^3 + p_{11} z_{41}^4 + p_{12} z_{41}^5 \\ 0 \\ p_{13}(z_{11} - x_{41}) + p_{14}(z_{31} - x_{41}) + p_{15} z_{31}^2 \\ 0 \\ p_{16} x_{41}^2 + p_{17} z_{21}^2 \\ 0 \\ p_{18} x_{11}^2 + p_{19} x_{41}^2 + p_{20} z_{11}^2 + p_{21} z_{21}^2 + p_{22} z_{41}^2 \\ 0 \\ p_{23} z_{31}^2, \\ 0 \\ p_{24} x_{31}^2 + p_{25} z_{41}^2 \end{bmatrix} \tag{5-17}$$

Visual inspection of Figure 5-10, confirms that most of the change in dynamics occurs in Regions 1 and 4 (V1 and R-DLPFC), with relatively little change in Region 3 (L-DLPFC) and low-frequency Region 2 (L-IPS).

Inspection of (5-17) confirms that the impact of R-DLPFC (blue) is significant for nearly all regions in either the high- or low-frequency network. The expressions for the V1 terms

$$SR2_{x1}(x, z) = p_1(z_{31} - x_{11}) + p_2(z_{41} - x_{11}) + p_3 x_{11}^2 + p_4 z_{41}^2 + p_5 z_{41}^3 + p_6 x_{41}^4 + p_7 z_{41}^5, \tag{5-18}$$

$$SR2_{z1}(x, z) = p_{16} x_{41}^2 + p_{17} z_{21}^2, \tag{5-19}$$

suggest that the inputs to V1 change significantly when attention mechanisms are activated, especially those from R-DLPFC to the low-frequency node.

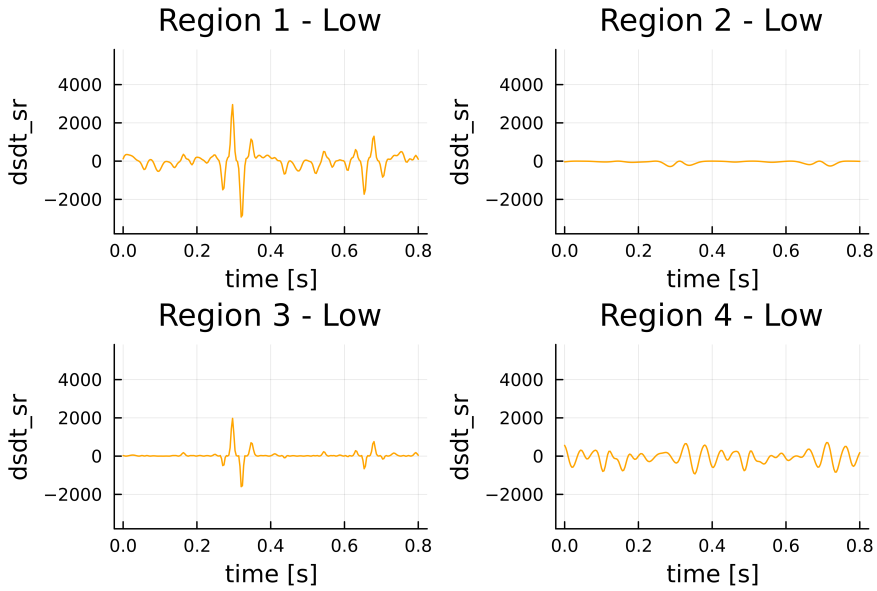
The expressions for the R-DLPFC terms

$$SR2_{x4}(x, z) = p_{13}(z_{11} - x_{41}) + p_{14}(z_{31} - x_{41}) + p_{15} z_{31}^2, \tag{5-20}$$

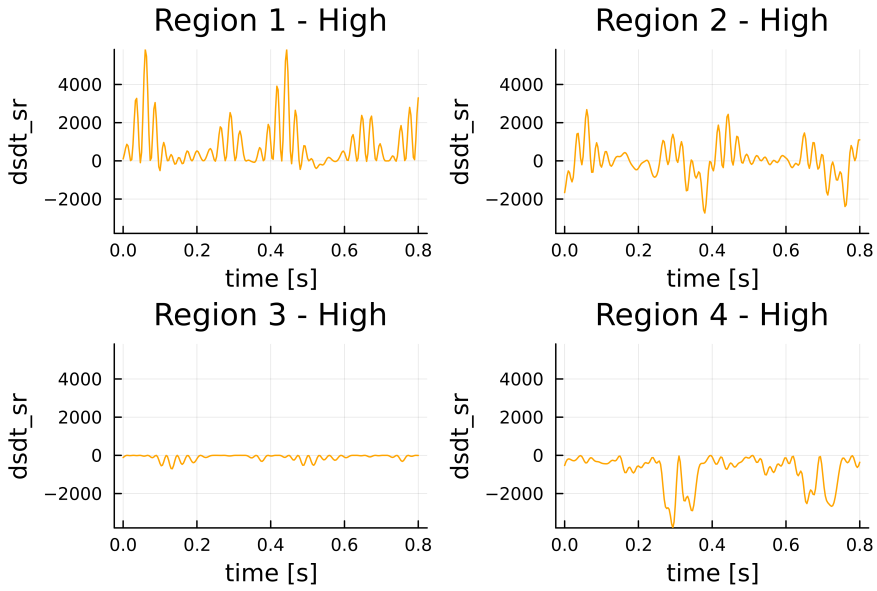
$$SR2_{z4}(x, z) = p_{24} x_{31}^2 + p_{25} z_{41}^2, \tag{5-21}$$

suggest that the flow from the L- to R-DLPFC is significant to the attention mechanisms. In fact, the terms for the low-frequency expressions for the two subjects are identical ($SR_{x4} = SR2_{x4}$).

Also of note is the greater number of changes in internal dynamics present in this model. Expressions for low-frequency V1, high-frequency L- and R-DLPFC, and high-frequency L-IPS all contain terms which relate to their own dynamics, suggesting that the autonomous



(a) Terms influencing low-frequency dynamics ($SR2_{xi}, i = 1, 2, 3, 4$).



(b) Terms influencing high-frequency dynamics ($SR2_{zi}, i = 1, 2, 3, 4$).

Figure 5-10: $SR2(x(t), z(t))$ evaluated over the solutions $x(t), z(t)$ of the full model.

characteristics of these populations change when attention mechanisms are active. This lends additional support to the conclusion that attention mechanisms influence more than only the coupling between regions.

Chapter 6

Conclusions and Recommendations

This chapter presents the key results from this project and suggestions for future work based on these results. In particular, the modelling algorithm is considered with respect to its utility for neuroscientists, both in understanding of the attention-working memory (WM) interplay and for more general use in magnetoencephalography (MEG) studies.

6-1 Summary

In Section 5-1, several baseline models are compared in terms of their ability to replicate the power spectrum of an MEG signal in a data-driven modelling context. It is found that a higher-dimension network of linear oscillators outperforms a lower-dimension network of nonlinear oscillators for this task. This result highlights the importance of a well-formed optimisation problem in data-driven modelling tasks, meaning model selection should be motivated by both physical understanding of the system and simplification of the loss function. In addition, a multi-stage parameter fitting procedure is introduced which allows this network to capture wide-frequency behaviour, an important part of modelling MEG studies.

In Sections 5-2 and 5-4, the developed modelling algorithm is applied to two single-subject datasets and the resulting expressions representing activation of attentional mechanisms are analysed. This analysis reveals some mechanisms which are common to both models. These common elements suggest that activation of attention mechanisms:

1. significantly changes the dynamics of Regions 1 and 4 (primary visual cortex (V1) and R-dorsolateral (DL)prefrontal cortex (PFC)),
2. does not significantly change the dynamics of Region 3 (L-DLPFC) and low-frequency Region 2 (L-intraparietal suculus (IPS)),
3. changes the influence of Region 4 (R-DLPFC) on all other nodes, in a nonlinear way,

4. changes the influence of Region 3 (L-DLPFC) on Region 4 (R-DLPFC), in a nonlinear way,
5. causes an internal change in dynamics in some regions.

Points 1 to 4 suggest that there exists a pathway from the L-DLPFC to the R-DLPFC and then through to other nodes, especially V1, which is significantly altered when attention mechanisms are activated during the WM task. The appearance of inputs from the R-DLPFC in nearly all the nodes of the attention model suggests that this region plays an important role in the attention-WM interplay.

The L-IPS is included as a region of interest in this model because it is hypothesised as an origin of top-down attention signalling to the visual cortex. The results support this hypothesis, as output from this region appears only in the dynamics of V1 in the attention model. However, in both single-subject models, the V1 attention dynamics are dominated by terms from the R-DLPFC, suggesting that a direct connection to R-DLPFC may be just as significant as the L-IPS to changing the behaviour of V1 when attention mechanisms are activated.

Point 5 suggests that the attention-WM interplay influences not only the dynamical relationship between different regions, but also the internal dynamics of certain regions. Both single-subject models include such internal changes, although they disagree as to which populations experience them.

In Section 5-3, a hypothesis-driven model is fit to the same dataset to allow for comparison of the new algorithm to a more traditional modelling approach. This comparison demonstrates that although both models replicate the data well, the flexibility of the new algorithm allows for the data-driven discovery of models which suggest new mechanisms for attention dynamics. For example, the symbolic regression model reveals a L-DLPFC $\rightarrow$ R-DLPFC $\rightarrow$ V1 pathway which may be important to attention dynamics, which is not at all present in the hypothesis-driven model. This insight can serve as a starting point for further investigation by neuroscientists, for data analysis or experimental design.

In addition to demonstrating the neuroscientific insights that can arise from data-driven dynamical modelling, this project extends work in the scientific machine learning domain with a novel application to MEG datasets. A linear oscillator network model and training algorithm are introduced which allow the baseline model to capture the wide-spectrum behaviour of several brain regions of interest. The results then demonstrate how neural networks, in combination with the baseline network model, can be used to capture changes in the frequency-response of the system, validating the existing universal differential equation (UDE) framework. In addition, symbolic regression techniques which have previously been used in conjunction with UDEs are shown to replicate the behaviour captured by the neural network well, validating the use of these techniques for improving the interpretability of neural networks.

6-2 Future Work

In this section, application of the current model to other problems in neuroscience, as well as potential improvements to the current algorithm, are discussed.

In the context of learning the attention-WM interplay, the current algorithm must be extended to capture the full multi-subject dataset to allow for any significant conclusions. This requires some consideration of how to model the subject-to-subject variability. Despite this challenge, this extension would improve the training of the neural network considerably, as it would lead to a fuller coverage of the network input domain and force the network to capture dynamics which apply to all subjects. This would likely lead to more generalisable insights, rather than the single-subject observations discussed in this report.

This extension touches on another limitation of this algorithm with respect to the generalisability of the model analysis. The neural network is currently trained on a single trajectory, reflecting the 'distractor' case behaviour. Similarly, the symbolic regression step is completed using only input-output data from a single trajectory. This means there is limited coverage of the network's input domain during training, leading to overfitting of the network to data and restricting the generalisability of the resulting model and subsequent analysis. Unfortunately, this issue is difficult to avoid in the current setting, as there is only a single case in which the network is active in the dynamics. However, a slightly different application of the method, for example where the activation variable ν is taken from a larger set or is continuous, could mitigate these issues. In this case, the network would be trained over several model solutions, improving the input domain coverage. An example of such a case is provided at the end of this section.

The data metrics selected for training of the model also leave some room for improvement. Although frequency-domain characteristics of MEG data such as the power spectrum are likely important to attention dynamics during the recall period, it is unlikely that these are the only important metrics, especially given the transient nature of the time-domain response. It is therefore recommended that some time-domain characteristics are added to the loss function for this study. Selection of these characteristics should be completed in consultation with neuroscientists, to ensure that they are relevant to attention dynamics. This will likely impact the selection of the baseline model, requiring further improvements.

Currently, confidence that this algorithm reveals mechanisms which are relevant to the underlying brain dynamics relies solely on the fact that the model replicates the data behaviours well. However, the hypothesis-driven model demonstrates that many different models can achieve this. One way to aid in further validation of this algorithm is to purposely include an irrelevant brain region in model. For example, the trials used in this study focus only on visual stimuli, so the auditory cortex should have little to no impact on the attention dynamics. If such an irrelevant node is included in the model, the attention model identified by this algorithm should not include any terms related to this node.

Although this framework is presented as a method for learning attention dynamics in the 'distractor'/'no distractor' conditions, it is extendable to other areas of interest to neuroscientists. To apply this method, it is only required that the MEG data be separated into different conditions and that there is an expected change in dynamics between these conditions. For example, the current dataset could be split based on behavioural outcomes such as the reaction time after the memory probe, e.g. slowest 50 percent versus fastest 50 percent of responses. The baseline model could then represent the "slow" dynamics, and the learned model the difference between the "slow" and "fast" dynamics. In this context, the algorithm would learn from data the dynamics which allows for faster reaction times. If the activation variable ν is made continuous rather than binary even more than two conditions could be

captured by the model, e.g. ν inversely proportional to the reaction time. This modelling algorithm, then, offers neuroscientists a method which suggests new neural mechanisms for a process of interest in a data-driven way, aiding in the formulation of new hypotheses.

Appendix A

Mathematical Background for Adjoint Sensitivity Method

This appendix outlines the mathematical proof for use of the backsolve adjoint sensitivity method for calculating derivatives. The Gauss adjoint sensitivity method used in this project is an extension of this method, with improvements in stability of the solution.

Theorem 1 gives the foundation of the adjoint method, and both the theorem and proof are heavily inspired by Appendix B in [23] in addition to the adjoint method described in [75].

First, it is necessary to clarify two different derivative operators:

$$\text{Momentary: } \frac{\delta \mathcal{L}}{\delta x(t)}, \text{Lasting: } \frac{\mathcal{D}\mathcal{L}}{\mathcal{D}x(t)}. \tag{A-1}$$

The loss $\mathcal{L}$ depends on the trajectory of the state $x(t)$. The momentary derivative describes the change of the loss if the state x is incrementally perturbed *only* at time t. In contrast, the lasting derivative describes the change of the loss if the state x is incrementally perturbed at time t and this change carries forward to affect all future states that the loss depends upon.

Theorem 1. *Define a scalar function $\mathcal{L}$ which depends on $z(t)$ at times $t_i, i = 0, \ldots, N$. If $z(t)$ follows $\frac{dz(t)}{dt} = f(z(t))$ and $a(t) = \frac{\mathcal{D}\mathcal{L}}{\mathcal{D}z(t)}$ then $\frac{da(t)}{dt} = -a(t)\frac{\delta f(z(t))}{\delta z(t)}$ for $t_i < t < t_{i+1}$.*

Proof. Define a scalar function $\mathcal{L}$ which depends on $z(t)$ at times $t_i, i = 0, \ldots, N$. Let $z(t)$ follow $\frac{dz(t)}{dt} = f(z(t))$, $a(t) := \frac{\mathcal{D}\mathcal{L}}{\mathcal{D}z(t)}$.

The evolution of the state z after some time Δt is given by

$$z(t + \Delta t) = z(t) + \int_t^{t+\Delta t} f(z(\tau))d\tau = G(z(t), \Delta t, t). \tag{A-2}$$

The Taylor series of $G(z(t), \Delta t, t)$ around $z(t)$ is

$$G(z(t), \Delta t, t) = z(t) + \Delta t f(z(t)) + O(\Delta t^2). \tag{A-3}$$

Consider a Δt small enough such that $t_i < t < t + \Delta t < t_{i+1}$. Applying the chain rule to $a(t)$ gives

$$a(t) = \frac{\mathcal{D}\mathcal{L}}{\mathcal{D}z(t)} = \frac{\mathcal{D}\mathcal{L}}{\mathcal{D}z(t+\Delta t)} \frac{\mathcal{D}z(t+\Delta t)}{\mathcal{D}z(t)} = a(t+\Delta t)\frac{\mathcal{D}z(t+\Delta t)}{\mathcal{D}z(t)}. \tag{A-4}$$

The differential equation governing the evolution of $a(t)$ can be derived through the definition of the derivative as

$$\begin{aligned}
\frac{da(t)}{dt} &= \lim_{\Delta t \to 0} \frac{a(t+\Delta t) - a(t)}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{a(t+\Delta t) - a(t+\Delta t)\dfrac{\mathcal{D}z(t+\Delta t)}{\mathcal{D}z(t)}}{\Delta t} && \text{(by Equation A-4)} \\
&= \lim_{\Delta t \to 0} \frac{a(t+\Delta t)(I - \dfrac{\mathcal{D}}{\mathcal{D}z(t)}(z(t) + \Delta t f(z(t)) + O(\Delta t^2))}{\Delta t} && \text{(by Equation A-3)} \\
&= \lim_{\Delta t \to 0} \frac{a(t+\Delta t)(-\dfrac{\mathcal{D}}{\mathcal{D}z(t)}\Delta t f(z(t)) - O(\Delta t^2))}{\Delta t} \\
&= \lim_{\Delta t \to 0} -a(t+\Delta t)\frac{\delta}{\delta z(t)} f(z(t)) - a(t+\Delta t)O(\Delta t) \\
&= -a(t)\frac{\delta f(z(t))}{\delta z(t)} && \text{(A-5)}
\end{aligned}$$

□

This theorem gives the evolution of the adjoint state a for times $t_i < t < t_{i+1}$, achieved through numerical integration of Equation A-5. However, this numerical integration requires knowledge of the state $z(t)$ at all time steps taken by the ordinary differential equation (ODE) solver, in order to calculate $\dfrac{\delta f(z(t))}{\delta z(t)}$. This can be achieved through simultaneous numerical integration of $\dfrac{dz(t)}{dt} = f(z(t))$.

To find the evolution of the adjoint at the missing times t_i, $a(t)$ can be expanded using the chain rule as

$$\begin{aligned}
a(t) &= \frac{\mathcal{D}\mathcal{L}}{\mathcal{D}z(t)} \\
&= \sum_{i=0}^{N} \frac{\delta \mathcal{L}}{\delta z(t_i)} \frac{\mathcal{D}z(t_i)}{\mathcal{D}z(t)}.
\end{aligned} \tag{A-6}$$

The lasting derivative in this expression can be simplified as

$$\frac{\mathcal{D}z(t_i)}{\mathcal{D}z(t)} = \begin{cases} 0, & t > t_i \\ 1, & t = t_i \end{cases}. \tag{A-7}$$

Equations A-6,A-7 prove that the full adjoint trajectory will consist of segments and discontinuities. For example, the expressions for $a(t)$ at $t_{N-1} \leq t \leq t_N$ are

$$\begin{aligned} a(t_N) &= \frac{\delta \mathcal{L}}{\delta z(t_N)}, \\ a(t) &= \frac{\delta \mathcal{L}}{\delta z(t_N)} \frac{\mathcal{D}z(t_N)}{\mathcal{D}z(t)} && \text{(for } t_{N-1} < t < t_N\text{)} \\ a(t_{N-1}) &= \frac{\delta \mathcal{L}}{\delta z(t_N)} \frac{\mathcal{D}z(t_N)}{\mathcal{D}z(t)} + \frac{\delta \mathcal{L}}{\delta z(t_{N-1})}. \end{aligned}$$

The appearance of the $\frac{\delta \mathcal{L}}{\delta z(t_{N-1})}$ term demonstrates the discontinuity. Moreover, because the loss $\mathcal{L}$ depends explicitly on z at times t_i and the value of $z(t_i)$ are known through forward numerical integration of $\frac{dz(t)}{dt} = f(z(t))$, $\frac{\delta \mathcal{L}}{\delta z(t_i)}$ are known.

In summary, consider a scalar function $\mathcal{L}(z(t_0), \ldots, z(t_N))$. If $z(t)$ follows $\frac{dz(t)}{dt} = f(z(t))$ and $a(t) = \frac{\mathcal{D}\mathcal{L}}{\mathcal{D}z(t)}$ then the loss gradient $\frac{\mathcal{D}\mathcal{L}}{\mathcal{D}z(t_0)}$ can be calculated using the following algorithm.

Algorithm 1 Basic Adjoint Differentiation

Initialise $z(t_0)$
$[\hat{z}(t_0), \hat{z}(t_1), \ldots, \hat{z}(t_N)] = \text{ODESolve}(z(t_0), f(z(t)), [t_0, \ldots, t_N])$ ▷ Solve forward time
$a(t_N) = \frac{\mathcal{D}\mathcal{L}(\hat{z}(t_0), \ldots, \hat{z}(t_N))}{\mathcal{D}z(t_N)}$
$z(t_N) = \hat{z}(t_N)$
$i \leftarrow N$
while $i > 0$ **do**
  $[z(t_{i-1}), a(t_{i-1})] = \text{ODESolve}([z(t_i), a(t_i)], [f(z(t)), -a(t)\frac{\delta f(z(t))}{\delta z(t)}], [t_i, t_{i-1}])$ ▷ Solve reverse time
  $a(t_{i-1}) = a(t_{i-1}) + \frac{\delta \mathcal{L}(\hat{z}(t_0), \ldots, \hat{z}(t_N))}{\delta z(t_i)}$ ▷ Adjoint Discontinuity at t_is
  $i = i - 1$
end while
return $a(t_0) = \frac{\mathcal{D}\mathcal{L}}{\mathcal{D}z(t_0)}$

A schematic of this procedure is given in Figure A-1.

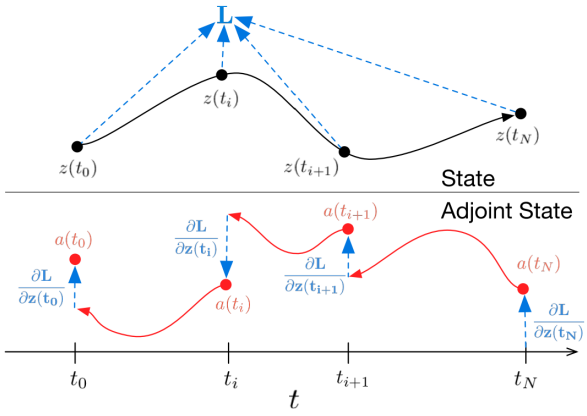


Figure A-1: A schematic of the backward adjoint method for differentiation of a state-dependent loss function through an ODE solver [23].

Theorem 1 and Algorithm 1 above can be applied to find the gradient of the loss function with respect to the parameters of the neural network λ by considering an augmented state which follows an augmented ODE

$$z(t) := \begin{bmatrix} x(t) \\ \lambda(t) \end{bmatrix}, \frac{dz(t)}{dt} = \begin{bmatrix} N(x(t), \lambda) \\ 0 \end{bmatrix} = f_z(z(t)). \tag{A-8}$$

Then define the adjoint state as

$$a_z(t) = \begin{bmatrix} a(t) & a_\lambda(t) \end{bmatrix} = \frac{\mathcal{D}\mathcal{L}}{\mathcal{D}z(t)} = \begin{bmatrix} \frac{\mathcal{D}\mathcal{L}}{\mathcal{D}x(t)} & \frac{\mathcal{D}\mathcal{L}}{\mathcal{D}\lambda(t)} \end{bmatrix}. \tag{A-9}$$

Note that numerical optimisation of the loss function requires $a_\lambda(t_0)$, which describes *lasting* derivative of the loss with respect to the parameters at the initial condition. By Theorem 1, the evolution of the adjoint state between times t_i is governed by

$$\begin{aligned} \frac{da_z(t)}{dt} &= -a_z(t)\frac{\delta f_z(z(t))}{\delta z(t)} \\ &= -\begin{bmatrix} a(t) & a_\lambda(t) \end{bmatrix} \begin{bmatrix} \frac{\delta N(x(t),\lambda)}{\delta x(t)} & \frac{\delta N(x(t),\lambda)}{\delta \lambda} \\ 0 & 0 \end{bmatrix} \\ &= -a(t) \begin{bmatrix} \frac{\delta N(x(t),\lambda)}{\delta x(t)} & \frac{\delta N(x(t),\lambda)}{\delta \lambda} \end{bmatrix}. \end{aligned} \tag{A-10}$$

Algorithm 1 can then be rewritten with the augmented system as

Algorithm 2 Augmented Adjoint Differentiation

Initialise $x(t_0), \lambda$

$[\hat{x}(t_0), \hat{x}(t_1), \ldots, \hat{x}(t_N)] = \text{ODESolve}(x(t_0), N(x(t), \lambda), [t_0, \ldots, t_N])$ ▷ Solve forward time

$a(t_N) = \dfrac{\mathcal{D}\mathcal{L}(\hat{x}(t_0), \ldots, \hat{x}(t_N))}{\mathcal{D}x(t_N)}$

$a_\lambda(t_N) = 0$

$x(t_N) = \hat{x}(t_N)$

$i \leftarrow N$

while $i > 0$ **do**

$[x(t_{i-1}), a(t_{i-1}), a_\lambda(t_{i-1})] = \text{ODESolve}([x(t_i), a(t_i), a_\lambda(t_i)], [N(z(t), \lambda), -a(t)[\dfrac{\delta N(z(t), \lambda)}{\delta x(t)}, \dfrac{\delta N(z(t), \lambda)}{\delta \lambda}]],$

▷ Solve reverse time

$a(t_{i-1}) = a(t_{i-1}) + \dfrac{\delta\mathcal{L}(\hat{z}(t_0), \ldots, \hat{z}(t_N))}{\delta z(t_i)}$ ▷ Adjoint Discontinuity at t_is

$i = i - 1$

end while

return $a_\lambda(t_0) = \dfrac{\mathcal{D}\mathcal{L}}{\mathcal{D}\lambda(t_0)}$, optionally $a(t_0) = \dfrac{\mathcal{D}\mathcal{L}}{\mathcal{D}x(t_0)}$

Note that returning $a(t_0)$ is optional depending on whether the initial condition $x(t_0)$ is assumed known or is a parameter to be optimised along with λ.

Appendix B

Mathematical Background for Source Reconstruction

This appendix outlines the process for source reconstruction of time-domain magnetoencephalography (MEG) signals through the linearly constrained minimum variance (LCMV) beamformer method.

First a *forward model* that relates electrical activity inside the brain to measurements outside the brain is necessary. Then, the *inverse problem* of reconstructing internal electrical activity based on externally recorded signals can be addressed.

Forward Model

Researchers have used biologically detailed simulations of pyramidal neurons to determine the approximations which can be used to map neuronal activity to MEG and electroencephalography (EEG) recordings [87]. Detailed compartmental models have been compared with multi-current-dipole and single current-dipole approximations, and researchers have found that the single current-dipole represents individual neurons and even entire neuronal populations well.

The activity of the current-dipole moments or polarities (x) can then be mapped to the MEG recordings (y) through a lead-field matrix (L), as follows,

$$y = Lx + n, \tag{B-1}$$

where n represents some noise. Although this is an appealing linear model, generation of the lead-field matrix is non-trivial. It takes into account

1. placement of the MEG sensors,

2. geometry of the head, also called a head-model, and
3. density (location) and orientation of the current-dipoles, also called the cortical mesh.

Placement of the MEG sensors is known and known calibration points are used to map these locations to the head geometry. For the head-model, functional magnetic resonance imaging (fMRI) scans for each subject are available, and a single-shell volume conduction model is used. A standard cortical mesh and parcellation map are used, see [120] for further details.

Inverse Problem

The forward model can be used to generate MEG data from neuronal population models, but it is more often desired to determine neuronal population behaviour from MEG data. This is termed the inverse problem. This problem is ill-posed, i.e. there does not exist a unique solution in the absence of additional constraints. Put formally, the inverse problem is to find the best lead-field parameters q (often the locations of the dipoles) and polarities (x) such that the forward model holds ($x^*, q^* = \arg\min_{x,q} ||y - L(q)x||$).

A popular method to solve this problem for known location q_0 is the LCMV beamformer method, introduced in [106]. Consider an alternative formulation of the forward model, where the lead-field matrix and current-dipole moments are broken into a sum

$$y = \sum_i L(q_i)x(q_i) + n \tag{B-2}$$

and q_i is the location of current-dipole i. Consider a spatial filtering of the MEG data at location q_0, as follows,

$$z = W^T(q_0)y. \tag{B-3}$$

If the filter is designed such that

$$W^T(q_0)L(q) = \begin{cases} I, q = q_0 \\ 0, q \neq q_0 \end{cases}, \tag{B-4}$$

the filtered MEG data would give the current-dipole moment at the location q_0 (plus noise), as follows,

$$z = \sum_{i \neq 0} W^T(q_0)L(q_i)x(q_i) + W^T(q_0)L(q_0)x(q_0) + n, \tag{B-5}$$

$$= x(q_0) + n. \tag{B-6}$$

However, the filter defined in Equation B-4 is not feasible. Instead, the unit response requirement at the desired location q_0 can be retained, and the complete attenuation at $q \neq q_0$ replaced with a softer minimum variance requirement, giving an optimisation problem for the design of the filter

$$\min_{W(q_0)} \text{tr}(W^T(q_0)C(n)W(q_0)) \text{ s.t. } W^T(q_0)L(q_0) = I, \tag{B-7}$$

where $C(n)$ is the covariance of the noise term n. This is termed the linearly constrained minimum variance (LCMV) problem [106] and has a closed-form solution

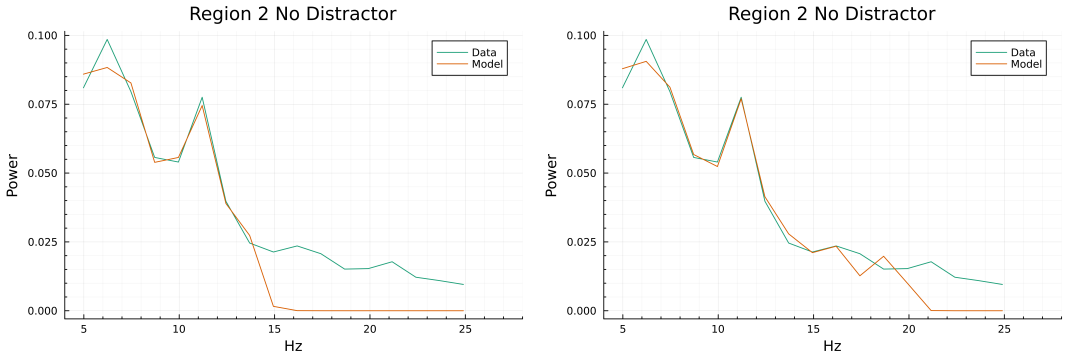
$$W^*(q_0) = (L^T(q_0)C^{-1}(n)H(q_0))^{-1}H^T(q_0)C^{-1}(n). \tag{B-8}$$

This enables one to calculate the polarisation of a current-dipole based on MEG data at any location in the brain, which can then be translated to a neuronal population model.

Appendix C

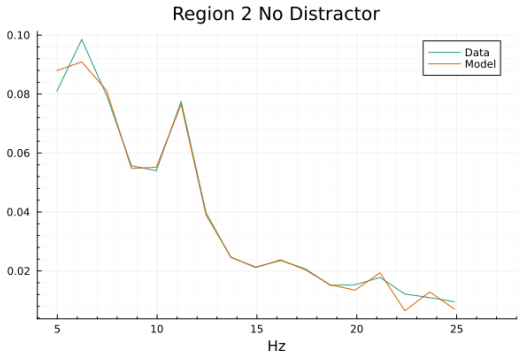
Supplementary Figures for Results and Analysis

C-1 Supplementary Figures for Baseline Model Fitting



(a) Linear Harmonic Oscillator

(b) Double Node Harmonic Oscillator

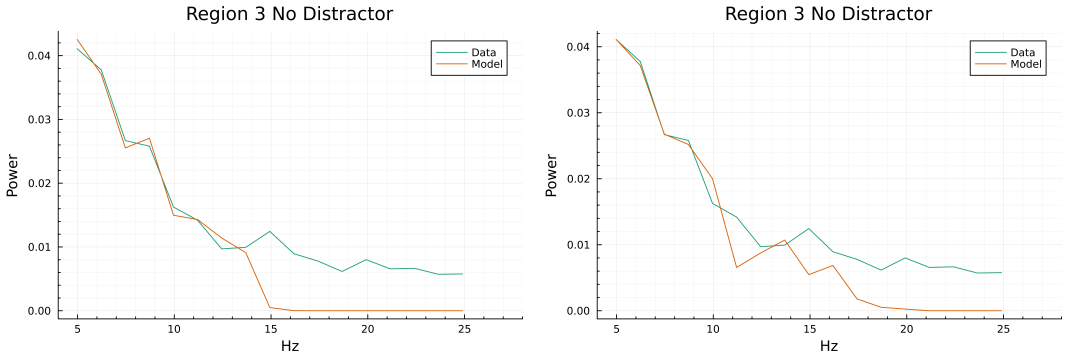


(c) Double Network Harmonic Oscillator

(d) Duffing-Van der Pol Oscillator

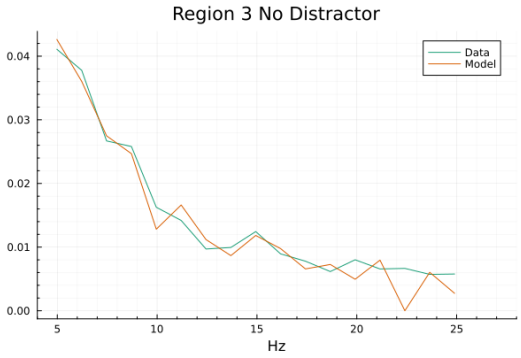
(e) Wilson-Cowan Oscillator

Figure C-1: Fit of the candidate baseline models to the "no distractor" data for Region 2 - L-intraparietal suculus (IPS).



(a) Linear Harmonic Oscillator

(b) Double Node Harmonic Oscillator

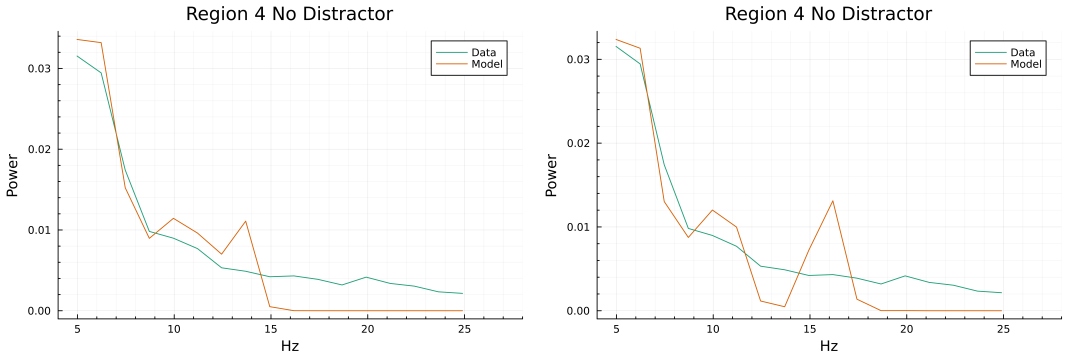


(c) Double Network Harmonic Oscillator

(d) Duffing-Van der Pol Oscillator

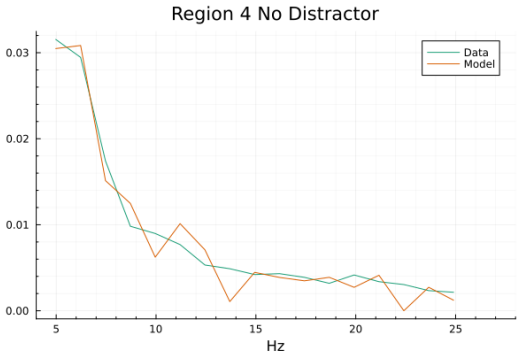
(e) Wilson-Cowan Oscillator

Figure C-2: Fit of the candidate baseline models to the "no distractor" data for Region 3 - R-dorsolateral (DL)prefrontal cortex (PFC).

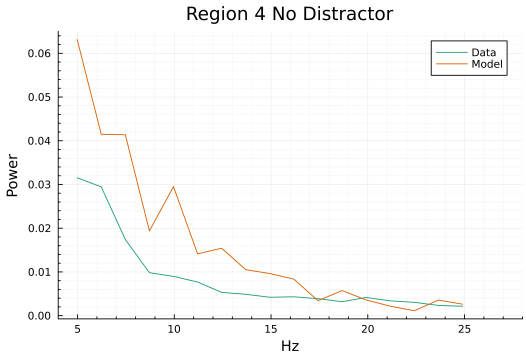


(a) Linear Harmonic Oscillator

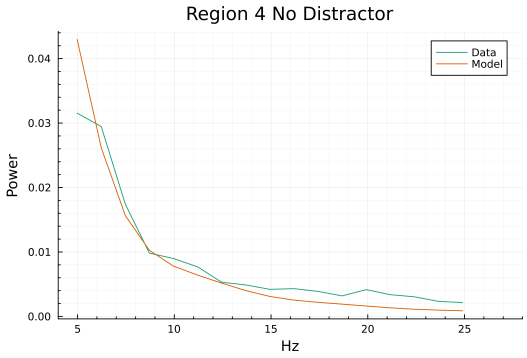
(b) Double Node Harmonic Oscillator



(c) Double Network Harmonic Oscillator



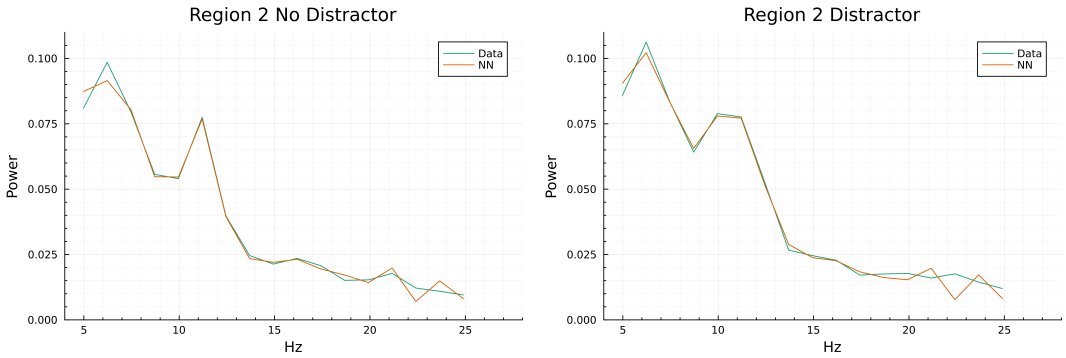
(d) Duffing-Van der Pol Oscillator



(e) Wilson-Cowan Oscillator

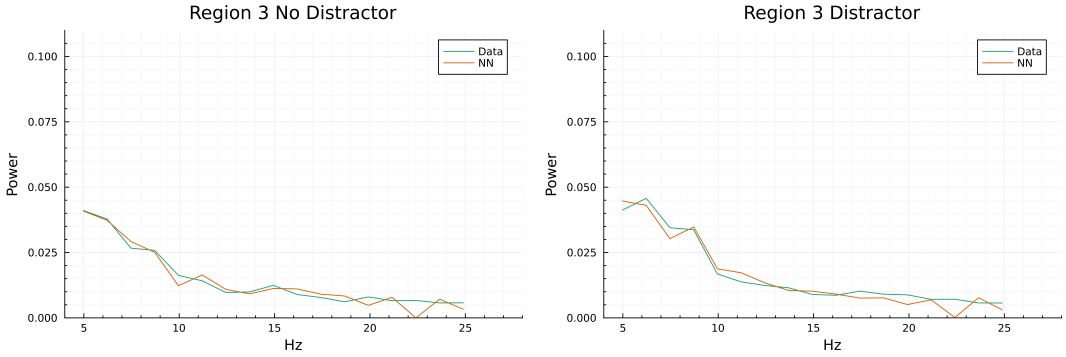
Figure C-3: Fit of the candidate baseline models to the "no distractor" data for Region 4 - R-DLPFC.

C-2 Supplementary Figures for Attention Model Fitting



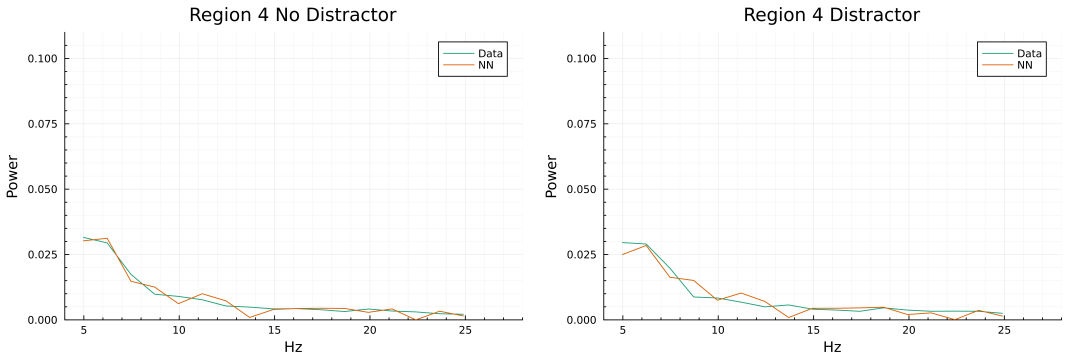
(a) No Distractor, $\nu = 0$ **(b)** Distractor, $\nu = 1$

Figure C-4: Fit of the universal differential equation (UDE), with baseline double extended linear oscillator and attention neural network models, to the full dataset for Region 2 - L-IPS.



(a) No Distractor, $\nu = 0$ **(b)** Distractor, $\nu = 1$

Figure C-5: Fit of the UDE, with baseline double extended linear oscillator and attention neural network models, to the full dataset for Region 3 - L-DLPFC.



(a) No Distractor, $\nu = 0$ **(b)** Distractor, $\nu = 1$

Figure C-6: Fit of the UDE, with baseline double extended linear oscillator and attention neural network models, to the full dataset for Region 4 - R-DLPFC.

C-3 Supplementary Figures for Coupling Change Model Fitting

Figures displaying the coupling strength between populations in the baseline model, plus significant changes in coupling with activation of attention mechanisms. The weight of the arrow reflects the strength of the coupling in the baseline. Green arrows indicate positive (excitatory) coupling, red arrows indicate negative (inhibitory) coupling. The percentage change in the coupling strength, if significant, is written on each arrow. The change is deemed significant if the coupling value in the baseline model is above a given threshold (500) and the change in that value in the attention model is greater than 5%.

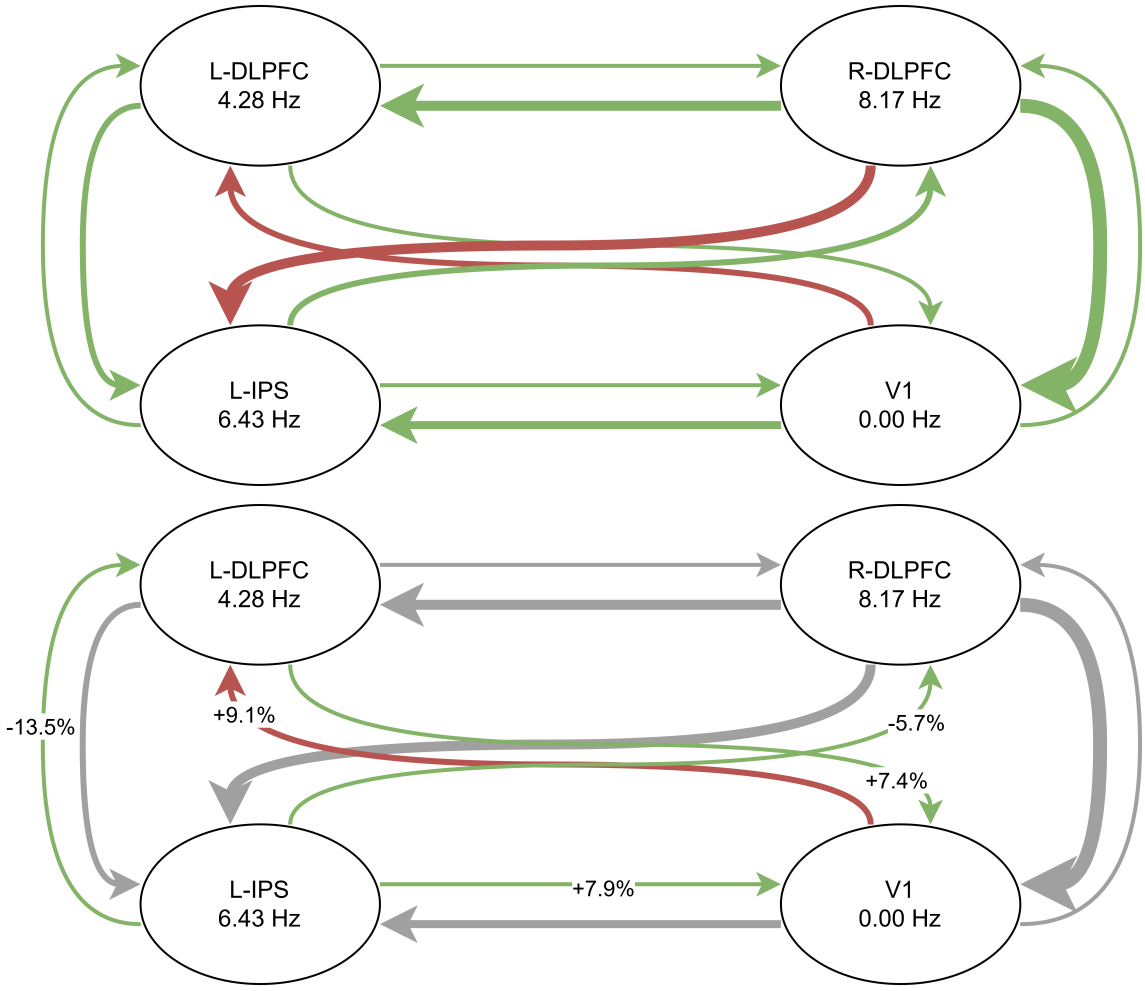


Figure C-7: Coupling between low-frequency neural populations which change significantly with activation of attention mechanisms.

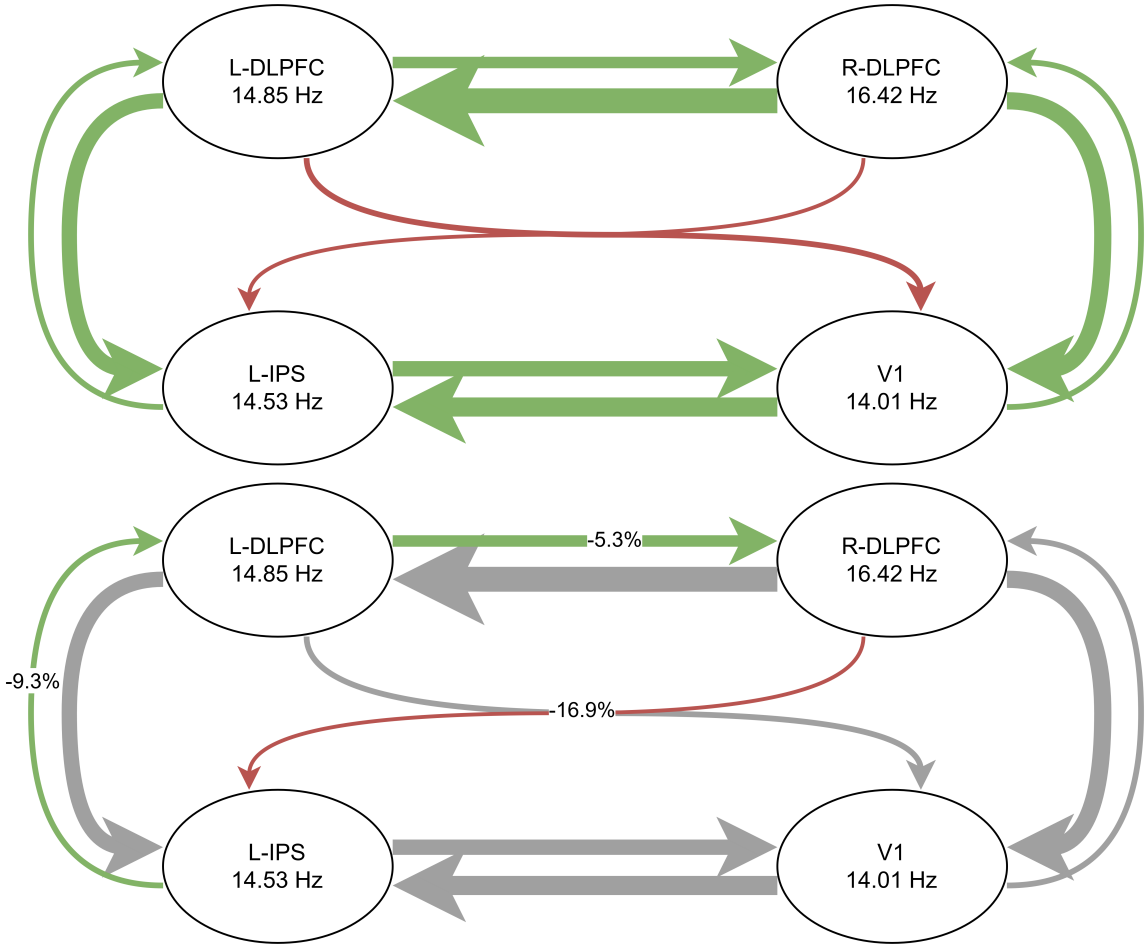


Figure C-8: Coupling between high-frequency neural populations which change significantly with activation of attention mechanisms.

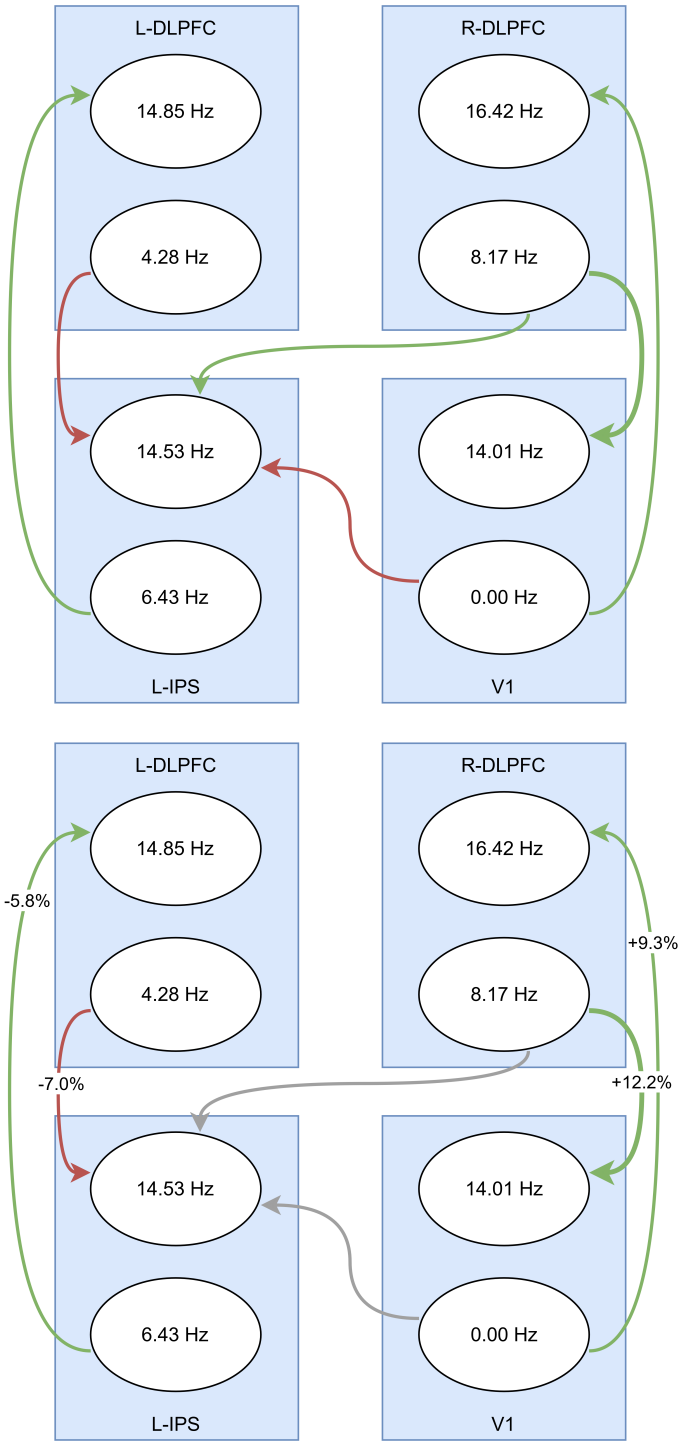
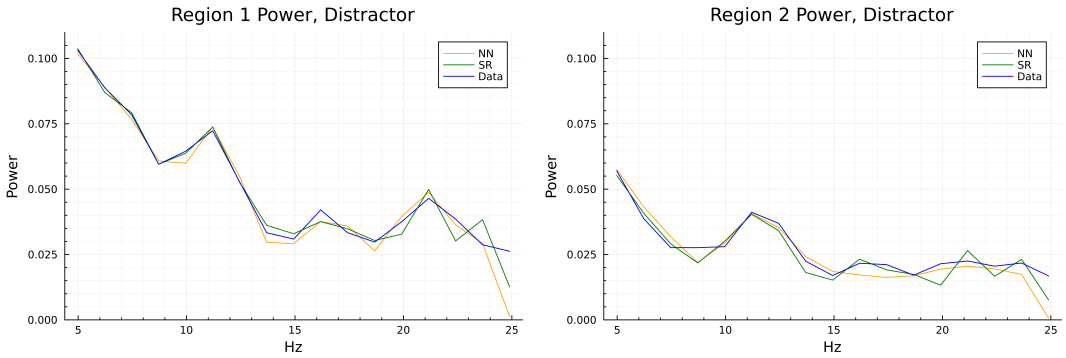
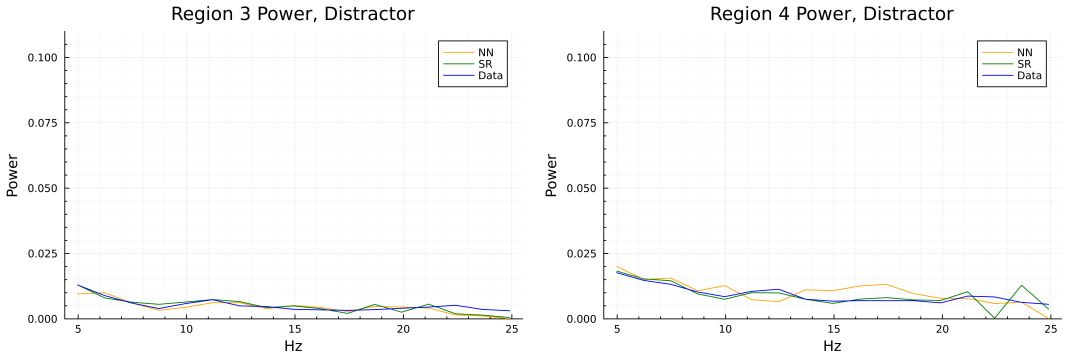


Figure C-9: Cross-coupling between the high- and low-frequency neural populations which change significantly with activation of attention mechanisms.

C-4 Supplementary Figures for Second Subject Model Fitting



(a) primary visual cortex (V1) **(b)** L-IPS



(c) L-DLPFC **(d)** R-DLPFC

Figure C-10: Comparison between the symbolic regression and neural network models in capturing the "distractor" dataset.

Bibliography

[1] M. Akin. Comparison of Wavelet Transform and FFT Methods in the Analysis of EEG Signals. Technical Report 3, 2002.

[2] M. Avery, J. L. Krichmar, and N. Dutt. Spiking Neuron Model of Basal Forebrain Enhancement of Visual Attention. In *IEEE World Congress on Computational Intelligence*, Brisbane, 6 2012.

[3] M. C. Avery, N. Dutt, and J. L. Krichmar. A large-scale neural network model of the influence of neuromodulatory levels on working memory and behavior. *Frontiers in Computational Neuroscience*, 7(OCT):58994, 10 2013.

[4] M. C. Avery, N. Dutt, and J. L. Krichmar. Mechanisms underlying the basal forebrain enhancement of top-down and bottom-up attention. *European Journal of Neuroscience*, 39(5):852–865, 3 2014.

[5] M. C. Avery and J. L. Krichmar. Neuromodulatory systems and their interactions: A review of models, theories, and experiments. *Frontiers in Neural Circuits*, 11:300066, 12 2017.

[6] A. Baddeley, G. Hitch, and R. Allen. A Multicomponent Model of Working Memory. In R. H. Logie, V. Camos, and N. Cowan, editors, *Working Memory: State of the Science*, pages 10–43. Oxford University Press, 11 2020.

[7] S. Baillet. Magnetoencephalography for brain electrophysiology and imaging. *Nature Neuroscience 2017 20:3*, 20(3):327–339, 2 2017.

[8] A. K. Barbey, M. Koenigs, and J. Grafman. Dorsolateral prefrontal contributions to human working memory. *Cortex*, 49(5):1195–1205, 5 2013.

[9] P. Bartolomeo and T. Seidel Malkinson. Hemispheric lateralization of attention processes in the human brain. *Current Opinion in Psychology*, 29:90–96, 10 2019.

[10] M. Benayoun, J. D. Cowan, W. van Drongelen, and E. Wallace. Avalanches in a stochastic model of spiking neurons. *PLoS Computational Biology*, 6(7):21, 2010.

[11] K. Benchenane, P. H. Tiesinga, and F. P. Battaglia. Oscillations in the prefrontal cortex: a gateway to memory and attention. *Current Opinion in Neurobiology*, 21(3):475–485, 6 2011.

[12] F. Beuth and F. H. Hamker. A mechanistic cortical microcircuit of attention for amplification, normalization and suppression. *Vision Research*, 116:241–257, 11 2015.

[13] N. P. Bichot, A. F. Rossi, and R. Desimone. Parallel and serial neural mechanisms for visual search in macaque area V4. *Science*, 308(5721):529–534, 4 2005.

[14] A. Bills, S. Sripad, W. L. Fredericks, M. Guttenberg, D. Charles, E. Frank, and V. Viswanathan. Universal Battery Performance and Degradation Model for Electric Aircraft. 7 2020.

[15] J. W. Bisley, K. Mirpour, F. Arcizet, and W. S. Ong. The Role of the Lateral Intraparietal Area in Orienting Attention and its Implications for Visual Search. *The European journal of neuroscience*, 33(11):1982, 6 2011.

[16] M. Bonnefond and O. Jensen. Report Alpha Oscillations Serve to Protect Working Memory Maintenance against Anticipated Distracters. *Current Biology*, 22, 2012.

[17] M. Breakspear. Dynamic models of large-scale brain activity. *Nature Neuroscience 2017 20:3*, 20(3):340–352, 2 2017.

[18] S. L. Brunton, J. L. Proctor, and J. N. Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences of the United States of America*, 113(15):3932–3937, 2016.

[19] C. I. Buia and P. H. Tiesinga. Role of interneuron diversity in the cortical microcircuit for attention. *Journal of Neurophysiology*, 99(5):2158–2182, 5 2008.

[20] T. J. Buschman and S. Kastner. From Behavior to Neural Dynamics: An Integrated Theory of Attention, 10 2015.

[21] W. J. Chai, A. I. Abd Hamid, and J. M. Abdullah. Working memory from the psychological and neurosciences perspectives: A review. *Frontiers in Psychology*, 9(MAR):327922, 3 2018.

[22] C. C. Chen, S. J. Kiebel, and K. J. Friston. Dynamic causal modelling of induced responses. *NeuroImage*, 41(4):1293–1312, 7 2008.

[23] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud. Neural Ordinary Differential Equations. In *Conference on Neural Information Processing Systems*, Montreal, 2018.

[24] N. Cowan, C. C. Morey, and M. Naveh-Benjamin. An Embedded-Processes Approach to Working Memory. In R. H. Logie, V. Camos, and N. Cowan, editors, *Working Memory*, pages 44–84. Oxford University Press, 11 2020.

[25] M. Cranmer, A. Sanchez-Gonzalez, P. Battaglia, R. Xu, K. Cranmer, D. Spergel, and S. Ho. Discovering Symbolic Models from Deep Learning with Inductive Biases. *Advances in Neural Information Processing System*, 33, 6 2020.

[26] R. Dandekar, C. Rackauckas, and G. Barbastathis. A Machine Learning-Aided Global Diagnostic and Comparative Tool to Assess Effect of Quarantine Control in COVID-19 Spread. *Patterns*, 1(9), 12 2020.

[27] S. Danisch and J. Krumbiegel. Makie.jl: Flexible high-performance data visualization for Julia. *Journal of Open Source Software*, 6(65):3349, 9 2021.

[28] J. Daunizeau, O. David, and K. E. Stephan. Dynamic causal modelling: A critical review of the biophysical and statistical foundations, 9 2011.

[29] O. David and K. J. Friston. A neural mass model for MEG/EEG: Coupling and neuronal dynamics. *NeuroImage*, 20(3):1743–1755, 2003.

[30] O. David, S. J. Kiebel, L. M. Harrison, J. Mattout, J. M. Kilner, and K. J. Friston. Dynamic causal modeling of evoked responses in EEG and MEG. *NeuroImage*, 30(4):1255–1272, 5 2006.

[31] O. David, J. M. Kilner, and K. J. Friston. Mechanisms of evoked and induced responses in MEG/EEG. *NeuroImage*, 31(4):1580–1591, 7 2006.

[32] J. C. De Munck, F. Bijma, P. Gaura, C. A. Sieluzycki, M. I. Branco, and R. M. Heethaar. A maximum-likelihood estimator for trial-to-trial variations in noisy MEG/EEG data sets. *IEEE Transactions on Biomedical Engineering*, 51(12):2123–2128, 12 2004.

[33] I. E. de Vries, H. A. Slagter, and C. N. Olivers. Oscillatory Control over Representational States in Working Memory. *Trends in Cognitive Sciences*, 24(2):150–162, 2 2020.

[34] G. Deco and E. T. Rolls. Attention and working memory: a dynamical model of neuronal activity in the prefrontal cortex. *European Journal of Neuroscience*, 18:2374–2390, 2003.

[35] G. Deco and E. T. Rolls. A Neurodynamical cortical model of visual attention and invariant object recognition. *Vision Research*, 44(6):621–642, 3 2004.

[36] G. Deco and E. T. Rolls. Attention, short-term memory, and action selection: A unifying theory. *Progress in Neurobiology*, 76(4):236–256, 7 2005.

[37] G. Deco and E. T. Rolls. Neurodynamics of biased competition and cooperation for attention: A model with spiking neurons. *Journal of Neurophysiology*, 94(1):295–313, 7 2005.

[38] G. Deco and A. Thiele. Cholinergic control of cortical network interactions enables feedback-mediated attentional modulation. *European Journal of Neuroscience*, 34(1):146–157, 7 2011.

[39] M. D. . Esposito and B. R. Postle. The Cognitive Neuroscience of Working Memory. *Annu. Rev. Psychol*, 66:115–142, 2015.

[40] I. C. Fiebelkorn and S. Kastner. Functional specialization in the attention network. *Annual Review of Psychology*, 71(Volume 71, 2020):221–249, 1 2020.

[41] P. J. Fried, R. J. Rushmore, M. B. Moss, A. Valero-Cabré, and A. Pascual-Leone. Causal evidence supporting functional dissociation of verbal and spatial working memory in the human dorsolateral prefrontal cortex. *European Journal of Neuroscience*, 39(11):1973–1981, 2014.

[42] P. Fries. Rhythms For Cognition: Communication Through Coherence. *Neuron*, 88(1):220, 10 2015.

[43] P. Fries, J. H. Reynolds, A. E. Rorie, and R. Desimone. Modulation of oscillatory neuronal synchronization by selective visual attention. *Science*, 291(5508):1560–1563, 2 2001.

[44] P. Fries, T. Womelsdorf, R. Oostenveld, and R. Desimone. The Effects of Visual Stimulation and Selective Visual Attention on Rhythmic Neuronal Synchronization in Macaque Area V4. *Journal of Neuroscience*, 28(18):4823–4835, 4 2008.

[45] K. Friston, J. Daunizeau, and K. E. Stephan. Model selection and gobbledygook: Response to Lohmann et al., 7 2013.

[46] K. J. Friston, L. Harrison, and W. Penny. Dynamic causal modelling. *NeuroImage*, 19(4):1273–1302, 8 2003.

[47] K. J. Friston, V. Litvak, A. Oswal, A. Razi, K. E. Stephan, B. C. Van Wijk, G. Ziegler, and P. Zeidman. Bayesian model reduction and empirical Bayes for group (DCM) studies. *NeuroImage*, 128:413–431, 3 2016.

[48] P. Ghorbanian, S. Ramakrishnan, and H. Ashrafiuon. Stochastic non-linear oscillator models of EEG: The alzheimer's disease case. *Frontiers in Computational Neuroscience*, 9(APR), 4 2015.

[49] P. Ghorbanian, S. Ramakrishnan, A. J. Simon, and H. Ashrafiuon. Stochastic Dynamic Modeling of the Human Brain EEG Signal. In *Dynamic Systems and Control Conference*. American Society of Mechanical Engineers, 10 2013.

[50] P. Ghorbanian, S. Ramakrishnan, A. Whitman, and H. Ashrafiuon. A phenomenological model of EEG based on the dynamics of a stochastic Duffing-van der Pol oscillator network. *Biomedical Signal Processing and Control*, 15:1–10, 1 2015.

[51] A. S. Ghuman, J. R. McDaniel, and A. Martin. A wavelet-based method for measuring the oscillatory dynamics of resting-state functional connectivity in MEG. *NeuroImage*, 56(1):69–77, 5 2011.

[52] G. G. Gregoriou, S. J. Gotts, H. Zhou, and R. Desimone. High-Frequency, long-range coupling between prefrontal and visual cortex during attention. *Science*, 324(5931):1207–1210, 5 2009.

[53] S. Haegens and E. Zion Golumbic. Rhythmic facilitation of sensory processing: A critical review, 3 2018.

[54] C. Han, R. Shapley, and D. Xing. Gamma rhythms in the visual cortex: functions and mechanisms. *Cognitive Neurodynamics 2021 16:4*, 16(4):745–756, 12 2021.

[55] R. Hari and R. Salmelin. Magnetoencephalography: From SQUIDs to neuroscience. Neuroimage 20th Anniversary Special Edition., 6 2012.

[56] D. Hermes, K. J. Miller, B. A. Wandell, and J. Winawer. Stimulus dependence of gamma oscillations in human visual cortex. *Cerebral Cortex*, 25(9):2951–2959, 9 2015.

[57] A. L. Hodgkin and A. F. Huxley. A Quantitative Description of Membrane Current and Its Application to Conduction and Excitation in Nerve. *J. Physiol*, 117:500–544, 3 1952.

[58] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1 1989.

[59] E. M. Izhikevich. Weakly pulse-coupled oscillators, FM interactions, synchronization, and oscillatory associative memory. *IEEE Transactions on Neural Networks*, 10(3):508–526, 1999.

[60] E. M. Izhikevich. *Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting*. MIT Press, 2007.

[61] B. H. Jansen and V. G. Rit. Biological Cybernetics Electroencephalogram and visual evoked potential generation in a mathematical model of coupled cortical columns. *Biological Cybernetics*, 73:357–366, 5 1995.

[62] T. Kanamaru and K. Aihara. Acetylcholine-mediated top-down attention improves the response to bottom-up inputs by deformation of the attractor landscape. *PLOS ONE*, 14(10):e0223592, 10 2019.

[63] C. Katsanevaki, A. M. Bastos, H. Cagnan, C. A. Bosman, K. J. Friston, and P. Fries. Attentional effects on local V1 microcircuits explain selective V1-V4 communication. *NeuroImage*, 281, 11 2023.

[64] B. Keith, A. Khadse, and S. E. Field. Learning orbital dynamics of binary black hole systems from gravitational wave measurements. *Physical Review Research*, 3(4), 12 2021.

[65] S. J. Kiebel, M. I. Garrido, R. J. Moran, and K. J. Friston. Dynamic causal modelling for EEG and MEG. *Cognitive Neurodynamics*, 2(2):121–136, 6 2008.

[66] S. Kim, W. Ji, S. Deng, Y. Ma, and C. Rackauckas. Stiff neural ordinary differential equations. *Chaos*, 31(9), 9 2021.

[67] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. 12 2014.

[68] L. Kipiński and W. Kordecki. Time-series analysis of trial-to-trial variability of MEG power spectrum during rest state, unattended listening, and frequency-modulated tones classification. *Journal of Neuroscience Methods*, 363, 11 2021.

[69] W. Klimesch. Alpha-band oscillations, attention, and controlled access to stored information. *Trends in Cognitive Sciences*, 16(12):606–617, 12 2012.

[70] S. Laxminarayan, G. Tadmor, S. G. Diamond, E. Miller, M. A. Franceschini, and D. H. Brooks. Modeling habituation in rat EEG-evoked responses via a neural mass model with feedback. *Biological Cybernetics*, 105(5-6):371–397, 12 2011.

[71] J. H. Lee, C. Koch, and S. Mihalas. A computational analysis of the function of three inhibitory cell types in contextual visual processing. *Frontiers in Computational Neuroscience*, 11:238376, 4 2017.

[72] J. H. Lee, M. A. Whittington, and N. J. Kopell. Top-Down Beta Rhythms Support Selective Attention via Interlaminar Interaction: A Model. *PLOS Computational Biology*, 9(8):e1003164, 2013.

[73] L. Leistritz, P. Putsche, K. Schwab, W. Hesse, T. Süße, J. Haueisen, and H. Witte. Coupled oscillators for modeling and analysis of EEG/MEG oscillations. *Biomedizinische Technik*, 52(1):83–89, 2 2007.

[74] L. Leistritz, T. Suesse, J. Haueisen, B. Hilgenfeld, and H. Witte. Methods for parameter identification in oscillatory networks and application to cortical and thalamic 600 Hz activity. In *Journal of Physiology Paris*, volume 99, pages 58–65, 1 2006.

[75] L. Lettermann, A. Jurado, T. Betz, F. Wörgötter, and S. Herzog. Tutorial: a beginner's guide to building a representative model of dynamical systems using the adjoint method. *Communications Physics*, 7(1), 12 2024.

[76] S. Li, Y. Cai, J. Liu, D. Li, Z. Feng, C. Chen, and G. Xue. Dissociated roles of the parietal and frontal cortices in the scope and control of attention during visual working memory. *NeuroImage*, 149:210–219, 4 2017.

[77] R. H. Logie, V. Camos, and N. Cowan, editors. *Working Memory: State of the Science*. Oxford University Press, 1 edition, 2021.

[78] G. Lohmann, K. Erfurth, K. Müller, and R. Turner. Critical comments on dynamic causal modelling. *NeuroImage*, 59(3):2322–2329, 2 2012.

[79] G. Lohmann, K. Müller, and R. Turner. Response to commentaries on our paper: Critical comments on dynamic causal modelling, 7 2013.

[80] S. J. Luck. Event-Related Potentials. In H. Cooper, P. Camic, D. Long, A. Panter, D. Rindskopf, and K. Sher, editors, *APA handbook of research methods in psychology*, volume 1, pages 523–546. American Psychological Association, 2012.

[81] Y. Ma, V. Dixit, M. Innes, X. Guo, and C. Rackauckas. A Comparison of Automatic Differentiation and Continuous Sensitivity Analysis for Derivatives of Differential Equation Solutions. In *IEEE High Performance Extreme Computing Conference*, pages 1–9, 12 2021.

[82] J. C. Martinez-Trujillo and S. Treue. Feature-based attention increases the selectivity of population responses in primate visual cortex. *Current Biology*, 14(9):744–751, 5 2004.

[83] J. H. Maunsell and S. Treue. Feature-based attention in visual cortex. *Trends in Neurosciences*, 29(6):317–322, 6 2006.

[84] P. Mengotti, A. S. Käsbauer, G. R. Fink, and S. Vossel. Lateralization, functional specialization, and dysfunction of attentional networks. *Cortex*, 132:206–222, 11 2020.

[85] R. Moran, D. A. Pinotsis, and K. Friston. Neural masses and fields in dynamic causal modelling. *Frontiers in Computational Neuroscience*, 7(APR 2013):44756, 4 2013.

[86] N. E. Myers, M. G. Stokes, L. Walther, and A. C. Nobre. Oscillatory Brain State Predicts Variability in Working Memory. *Journal of Neuroscience*, 34(23):7735–7743, 6 2014.

[87] S. Næss, G. Halnes, E. Hagen, D. J. Hagler, A. M. Dale, G. T. Einevoll, and T. V. Ness. Biophysically detailed forward modeling of the neural origin of EEG and MEG signals. *NeuroImage*, 225:117467, 1 2021.

[88] R. Oostenveld, P. Fries, E. Maris, and J.-M. Schoffelen. FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational intelligence and neuroscience*, 2011:156869, 2011.

[89] A. M. Owen, K. M. McMillan, A. R. Laird, and E. Bullmore. N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human Brain Mapping*, 25(1):46–59, 5 2005.

[90] D. A. Pinotsis, R. Loonis, A. M. Bastos, E. K. Miller, and K. J. Friston. Bayesian Modelling of Induced Responses and Neuronal Rhythms. *Brain Topography*, 32(4):569–582, 7 2019.

[91] T. C. Potjans and M. Diesmann. The Cell-Type Specific Cortical Microcircuit: Relating Structure and Activity in a Full-Scale Spiking Network Model. *Cerebral Cortex*, 24(3):785–806, 3 2014.

[92] C. Rackauckas. Universal Differential Equations for Scientific Machine Learning, 8 2020.

[93] C. Rackauckas, Y. Ma, J. Martensen, C. Warner, K. Zubov, R. Supekar, D. Skinner, A. Ramadhan, and A. Edelman. Universal Differential Equations for Scientific Machine Learning. 1 2020.

[94] A. Raz and J. Buhle. Typologies of attentional networks. *Nature Reviews Neuroscience*, 7:367–379, 2006.

[95] J. H. Reynolds, L. Chelazzi, and R. Desimone. Competitive Mechanisms Subserve Attention in Macaque Areas V2 and V4. *Journal of Neuroscience*, 19(5):1736–1753, 3 1999.

[96] J. H. Reynolds and R. Desimone. The Role of Neural Mechanisms of Attention in Solving the Binding Problem. *Neuron*, 24(1):19–29, 9 1999.

[97] A. H. G. Rinnooy Kan and G. T. Timmer. Stochastic global optimization methods part I: Clustering methods. *Mathematical Programming*, 39(1):27–56, 9 1987.

[98] F. Sapienza, J. Bolibar, F. Schäfer, B. Groenke, A. Pal, V. Boussange, P. Heimbach, G. Hooker, F. Pérez, P.-O. Persson, and C. Rackauckas. Differentiable Programming for Differential Equations: A Review. 6 2024.

[99] M. Sarabian, H. Babaee, and K. Laksari. Physics-Informed Neural Networks for Brain Hemodynamic Predictions Using Medical Imaging. *IEEE Transactions on Medical Imaging*, 41(9):2285–2303, 9 2022.

[100] A. Shakeel, T. Tanaka, and K. Kitajo. Time-series prediction of the oscillatory phase of eeg signals using the least mean square algorithm-based ar model. *Applied Sciences (Switzerland)*, 10(10), 5 2020.

[101] A. Shoeibi, N. Ghassemi, M. Khodatars, P. Moridian, A. Khosravi, A. Zare, J. M. Gorriz, A. H. Chale-Chale, A. Khadem, and U. Rajendra Acharya. Automatic diagnosis of schizophrenia and attention deficit hyperactivity disorder in rs-fMRI modality using convolutional autoencoder model and interval type-2 fuzzy regression. *Cognitive Neurodynamics 2022 17:6*, 17(6):1501–1523, 11 2022.

[102] S. P. Singh. Magnetoencephalography: Basic principles. *Annals of Indian Academy of Neurology*, 17(Suppl 1):S107, 2014.

[103] R. Tani and Y. Kashimori. Coordination of top-down influence on V1 responses by interneurons and brain rhythms. *BioSystems*, 207, 9 2021.

[104] A. Thiele and M. A. Bellgrove. Neuromodulation of Attention, 2 2018.

[105] P. H. Tiesinga and C. I. Buia. Spatial attention in area V4 is mediated by circuits in primary visual cortex. *Neural Networks*, 22:1039–1054, 2009.

[106] B. D. Van Veen, W. Van Drongelen, M. Yuchtman, and A. Suzuki. Localization of Brain Electrical Activity via Linearly Constrained Minimum Variance Spatial Filtering. Technical Report 9, 1997.

[107] N. Wagatsuma, S. Nobukawa, and T. Fukai. A microcircuit model involving parvalbumin, somatostatin, and vasoactive intestinal polypeptide inhibitory interneurons for the modulation of neuronal oscillation during visual processing. *Cerebral Cortex*, 33(8):4459–4477, 4 2023.

[108] N. Wagatsuma, T. C. Potjans, M. Diesmann, and T. Fukai. Layer-dependent attentional processing by top-down signals in a visual cortical microcircuit model. *Frontiers in Computational Neuroscience*, 5:10958, 7 2011.

[109] N. Wagatsuma, T. C. Potjans, M. Diesmann, K. Sakai, and T. Fukai. Spatial and Feature-Based Attention in a Layered Cortical Microcircuit Model. *PLOS ONE*, 8(12):e80788, 12 2013.

[110] T. D. Wager and E. E. Smith. Neuroimaging studies of working memory: A meta-analysis. *Cognitive, Affective and Behavioral Neuroscience*, 3(4):255–274, 2003.

[111] E. Wallace, M. Benayoun, W. van Drongelen, and J. D. Cowan. Emergent oscillations in networks of stochastic spiking neurons. *PLoS ONE*, 6(5), 2011.

[112] B. Wandell, S. Dumoulin, and A. Brewer. Visual Cortex in Humans. In *Encyclopedia of Neuroscience*, volume 10, pages 251–257. Elsevier, 2009.

[113] R. Wang and R. Yu. Physics-Guided Deep Learning for Dynamical Systems: A Survey. 7 2021.

[114] L. K. White, W. Makhoul, M. Teferi, Y. I. Sheline, and N. L. Balderston. The role of dlPFC laterality in the expression and regulation of anxiety, 2 2023.

[115] H. R. Wilson and J. D. Cowan. EXCITATORY AND INHIBITORY INTERACTIONS IN LOCALIZED POPULATIONS OF MODEL NEURONS. *Biophysical Journal*, 12, 1972.

[116] H. R. Wilson and J. D. Cowan. A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Kybernetik*, 13(2):55–80, 9 1973.

[117] H. R. Wilson and J. D. Cowan. Evolution of the Wilson–Cowan equations, 12 2021.

[118] T. Womelsdorf, P. Fries, P. P. Mitra, and R. Desimone. Gamma-band synchronization in visual cortex predicts speed of change detection. *Nature*, 439(7077):733–736, 12 2006.

[119] P. Zeidman, K. Friston, and T. Parr. A primer on Variational Laplace (VL). *NeuroImage*, 279:120310, 10 2023.

[120] Y. J. Zhou, A. Ramchandran, and S. Haegens. Alpha oscillations protect working memory against distracters in a modality-specific way. *NeuroImage*, 278, 9 2023.

Glossary

List of Acronyms

MEG	magnetoencephalography
WM	working memory
DCM	dynamic causal modelling
ADHD	attention deficit hyperactivity disorder
fMRI	functional magnetic resonance imaging
V1	primary visual cortex
PFC	prefrontal cortex
DL	dorsolateral
FEF	frontal eye field
LIP	lateral intraparietal area
IPS	intraparietal suculus
ACh	acetylcholine
EEG	electroencephalography
BF	basal forebrain
TRN	thalamic reticulate nucleus
ODE	ordinary differential equation
NODE	neural ordinary differential equation
SINDy	sparse identification of nonlinear dynamics
UDE	universal differential equation
LCMV	linearly constrained minimum variance
FFT	fast Fourier transform
IQR	inter-quartile range
AR	auto-regressive

ARMA auto-regressive moving average

STLSQ sequential thresholding least-squares

MLSL multi-level single linkage

SLSQP sequential least squares quadratic programming