Received: 15 August 2017 | Revised: 19 October 2017 | Accepted: 6 November 2017

DOI: 10.1002/wcs.1460

WILEY WIREs COGNITIVE SCIENCE

ADVANCED REVIEW

Generative models for clinical applications in computational psychiatry

Stefan Frässle[1] | Yu Yao[1] | Dario Schöbi[1] | Eduardo A. Aponte[1] | Jakob Heinzle[1] | Klaas E. Stephan[1,2]

[1]Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich & ETH Zurich, Zurich, Switzerland

[2]Wellcome Trust Centre for Neuroimaging, University College London, London, UK

Correspondence
Stefan Frässle, Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich & ETH Zurich, 8032 Zurich, Switzerland.
Email: stefanf@biomed.ee.ethz.ch

Funding information
University of Zurich; René and Susanne Braginsky Foundation; Marie Curie Actions for People COFUND Program; ETH Zurich Postdoctoral Fellowship Program

Despite the success of modern neuroimaging techniques in furthering our understanding of cognitive and pathophysiological processes, translation of these advances into clinically relevant tools has been virtually absent until now. Neuromodeling represents a powerful framework for overcoming this translational deadlock, and the development of computational models to solve clinical problems has become a major scientific goal over the last decade, as reflected by the emergence of clinically oriented neuromodeling fields like Computational Psychiatry, Computational Neurology, and Computational Psychosomatics. Generative models of brain physiology and connectivity in the human brain play a key role in this endeavor, striving for computational assays that can be applied to neuroimaging data from individual patients for differential diagnosis and treatment prediction. In this review, we focus on dynamic causal modeling (DCM) and its use for Computational Psychiatry. DCM is a widely used generative modeling framework for functional magnetic resonance imaging (fMRI) and magneto-/electroencephalography (M/EEG) data. This article reviews the basic concepts of DCM, revisits examples where it has proven valuable for addressing clinically relevant questions, and critically discusses methodological challenges and recent methodological advances. We conclude this review with a more general discussion of the promises and pitfalls of generative models in Computational Psychiatry and highlight the path that lies ahead of us.

This article is categorized under:
Neuroscience > Computation
Neuroscience > Clinical Neuroscience

1 | INTRODUCTION

Psychiatry faces fundamental conceptual and practical challenges with regard to reliable differential diagnosis, as well as prediction of clinical trajectories and treatment success in individual patients (Kapur, Phillips, & Insel, 2012; Krystal & State, 2014; Owen, 2014; Stephan, Bach, et al., 2016). At the moment, psychiatric diagnostics is informed by a syndromatic nosology as defined by the Diagnostic and Statistical Manual of Mental Disorders (DSM; American Psychiatric Association, 2013) or the International Classification of Diseases (ICD; World Health Organization, 1990). These schemes do not rest on pathophysiological or aetiological concepts, but suggest a descriptive taxonomy based on symptoms and signs. More importantly, however, the clinical categories proposed by these schemes (e.g., schizophrenia or depression) lack predictive validity with regard to clinical trajectories and do not provide treatment predictions for individual patients (Cuthbert & Insel, 2010, 2013; Kapur et al., 2012). Hence, physicians select therapies with respect to symptoms and side effects, typically engaging in prolonged trial-and-error treatment until eventually an effective medication is found. This has been illustrated by studies like the Sequential Treatment Alternatives to Relieve Depression (STAR*D) study which suggested that—in a sample of roughly 3,000 depressed patients—even after four sequential treatment adjustments, only two-thirds of the patients showed a

therapeutic benefit (Rush et al., 2006). This is not only frustrating for patients and physicians alike, but also time-consuming and expensive (Insel, 2008).

This unsatisfactory state of affairs in psychiatry is widely recognized, and there is a strong drive towards exploiting novel approaches that could not only furnish a deeper understanding of the pathophysiological processes underlying mental disorders (Stephan, Binder, et al., 2016) but also enable individual treatment predictions. While prominent efforts in this regard have been made by (epi)genetics and neuroimaging, theses promises are yet to be fulfilled (Braff & Freedman, 2008; Kapur et al., 2012).

In neuroimaging, techniques such as functional magnetic resonance imaging (fMRI) and magneto-/electroencephalography (M/EEG) enable non-invasive measures of human brain function and thus offer functional readouts from symptom-producing neural circuits in psychiatric conditions. These techniques have been maturing over decades (Berger, 1929; Ogawa, Lee, Kay, & Tank, 1990) and have considerably advanced our understanding of the physiology of cognitive processes. However, these advances have not yet been translated into routine clinical practice (Filiou & Turck, 2011; Kapur et al., 2012). While there are several possible reasons for this lack in clinical utility (Kapur et al., 2012; Stephan, Iglesias, Heinzle, & Diaconescu, 2015), one fundamental aspect is the descriptive nature of most clinical neuroimaging studies: on their own, neither localized changes in brain anatomy or activity nor aberrations of functional connectivity provide a mechanistic understanding of pathophysiology and do not easily inform the development of biologically grounded clinical tests (Stephan et al., 2015).

A promising alternative is a computational approach to neuroimaging, with mathematical models that capture hypothesized physiological and computational mechanisms. This is at the heart of clinical neuromodeling (Figure 1), with different specialized fields that are currently emerging, including Computational Psychiatry (Friston, Stephan, Montague, & Dolan, 2014; Huys, Maia, & Frank, 2016; Maia & Frank, 2011; Montague, Dolan, Friston, & Dayan, 2012; Stephan & Mathys, 2014; Wang & Krystal, 2014), Computational Neurology (Jirsa et al., 2016; Maia & Frank, 2011), and Computational Psychosomatics (Petzschner, Weber, Gard, & Stephan, 2017). While various computational approaches exist (for review, see Stephan et al., 2015), we restrict our discussion to so-called generative models (Box 1). These are described by the likelihood function, which is the probability of the data given a set of model parameters, and the prior distribution, which encodes the *a priori* plausible regime of parameter values. Generally, generative (Bayesian) models have a number of advantages compared to frequentist approaches and have thus gained increasing popularity over the last years. In all brevity, these advantages include: First, generative models force us to think carefully about the mechanisms underlying measurements and alternative hypotheses about the data-generating process. Furthermore, when writing down the exact form of the generative model, inherent assumptions have to be made explicit. Second, having specified a generative model, one can easily generate synthetic (simulated) data by sampling parameter values from the prior and inserting them into the likelihood function. This allows for testing the utility of a given model for explaining certain phenomena before acquiring data. Third, generative models allow for inference on model structure, based on the model evidence, which encodes the probability of the data given a model. This provides a formal way to compare competing hypotheses about the mechanisms that have generated the observed data (e.g., neuroimaging data, clinical symptoms). Simultaneously, generative models enable inference on model parameters which ideally afford some degree of mechanistic interpretability on the putative processes underlying the studied phenomenon (e.g., cognitive functions in health, or symptom-producing abnormalities in psychiatric and neurological conditions).

Their quest for mechanistic interpretability renders generative models naturally relevant for clinical applications. For instance, as is described in detail below, model comparison could provide a formal basis for differential diagnosis, and model parameter estimates provide compact, quantitative summaries of pathophysiological mechanisms. The latter can be used as input for machine learning techniques to stratify heterogeneous spectrum disorders or predict outcomes. This "generative

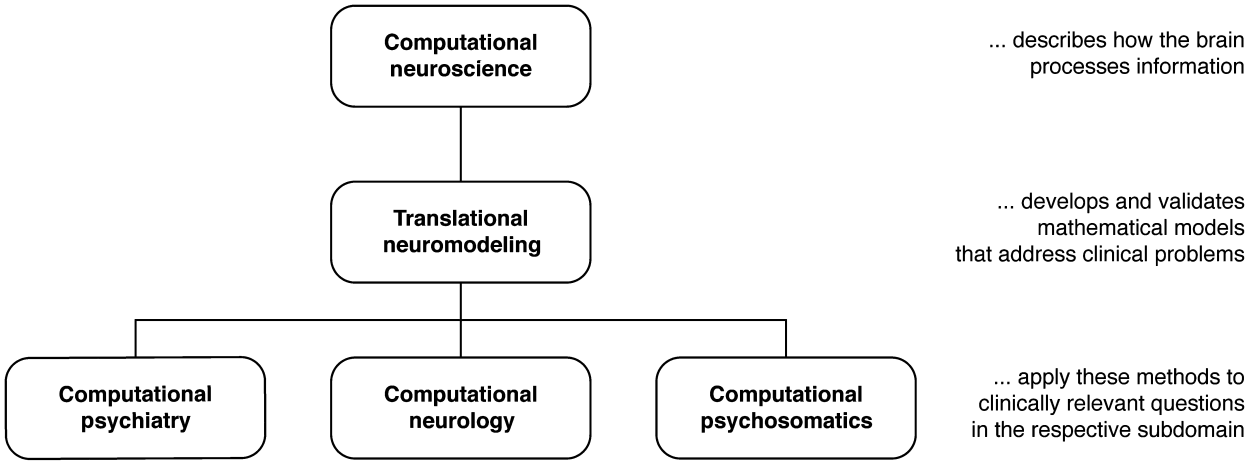


FIGURE 1 Taxonomy for different disciplines in the computational neurosciences and their relation to clinical questions. (Reprinted with permission from Stephan, Siemerkus, Bishop, & Haker, 2017. Copyright 2017 Hogrefe AG)

BOX 1

DEFINITION OF GENERATIVE MODELS

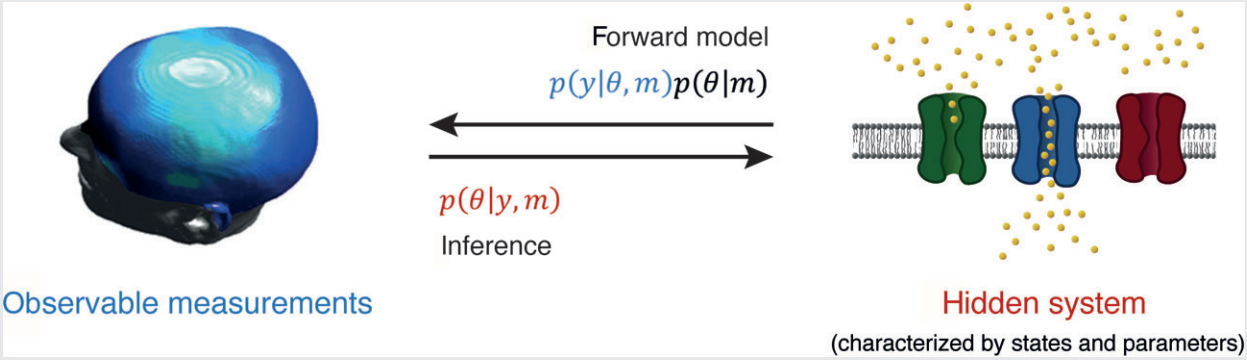
GENERATIVE MODELS

A generative model provides the joint probability $p(y,\theta|m)$ over measured data y and model parameters θ, given the model m. This requires specifying the likelihood function $p(y|\theta,m)$, which describes the probability of the data given a set of model parameters, and the prior distribution $p(\theta|m)$, which encodes the *a priori* plausible regime of parameter values. Together, the likelihood function and the prior density represent a full probabilistic forward mapping from the latent (hidden) model parameters to the measured data. Having specified such a forward mapping, one can generate synthetic (simulated) data by sampling parameter values from the prior and inserting them into the likelihood function.

More importantly, generative models allow one to infer the latent (hidden) parameter values of the system from the measured data. This is known as "model inversion" (or simply "inference") and essentially corresponds to computing the posterior distribution of the model parameters according to Bayes theorem:

$$p(\theta|y,m) = \frac{p(y|\theta,m)p(\theta|m)}{p(y|m)} \quad (2)$$

where $p(y|m)$ is the model evidence or marginal likelihood, which encodes how likely it is to obtain the measured data under the model m when randomly sampling from the prior. Since evaluating Equation (2) directly is often computationally infeasible, model inversion almost always proceeds using approximate Bayesian estimation techniques, for example, variational Bayes or Markov chain Monte Carlo sampling.



BOX 1. Schematic illustration of generative models, which provide a full probabilistic forward mapping from the latent (hidden) model parameters to the measured data in terms of the likelihood function and the prior distribution. Using Bayes theorem, the latent (hidden) parameter values of the system can be inferred from the measured data—a process that is known as "model inversion" (or simply "inference") and essentially corresponds to computing the posterior distribution of the model parameters. (Reprinted with permission from Stephan, Manjaly, et al. (2016). Copyright 2016 Frontiers)

embedding" approach (Brodersen et al., 2011, 2014) views a generative model as a theory-driven dimensionality reduction device that projects high-dimensional data onto a mechanistically interpretable feature space.

This review focuses on dynamic causal models (DCMs), a frequently used generative modeling framework that is used both for inferring physiological processes in local neuronal circuits and for inferring effective connectivity in distributed networks from neuroimaging data. Effective connectivity refers to the directed influences that neuronal populations exert over another. This requires a generative model that provides a forward mapping from hidden (latent) neuronal circuit dynamics to observable signals (Friston, Moran, & Seth, 2013). This is in contrast to functional connectivity which represents statistical dependencies between regional measurements—and is thus essentially descriptive and undirected (Friston, 2011). Reviews on other methods for estimating effective connectivity can be found elsewhere (Roebroeck, Formisano, & Goebel, 2011; Valdes-Sosa, Roebroeck, Daunizeau, & Friston, 2011). Initially introduced for fMRI data (Friston, Harrison, & Penny, 2003), DCM was later extended to electrophysiological data (MEG/EEG; David et al., 2006; David, Harrison, & Friston, 2005). Regardless of the exact data modality, DCM rests on a hierarchically structured likelihood function or forward model that distinguishes between (a) state or evolution equations that describe the dynamics of hidden neuronal (and, for fMRI, hemodynamic) states and (b) observation equations that map the system's states onto experimental measurements, such as fMRI or M/EEG signals.

Models of effective connectivity are particularly promising for studying pathophysiological mechanisms in the human brain because aberrant functional integration of large-scale brain networks has been suggested to play a central role in disease concepts of various psychiatric (and neurological) disorders (Deco & Kringelbach, 2014; Menon, 2011). At a smaller scale, impairments in synaptic plasticity have been proposed as putative mechanisms for circuit dysfunction and the pathophysiology of brain disorders (Klassen et al., 2011; Lau & Zukin, 2007). One theory that explicitly refers to interactions between these different scales (i.e., network and synaptic abnormalities) is the "dysconnection hypothesis" of schizophrenia. This posits that impairments in dopaminergic and/or cholinergic regulation of NMDA receptor dependent synaptic plasticity lead to dysconnectivity in distributed circuits for perception and learning (Friston, 1998; Friston, Brown, Siemerkus, & Stephan, 2016; Stephan, Baldeweg, & Friston, 2006; Stephan, Friston, & Frith, 2009). DCM is a useful framework to study interactions across network scales since its different variants cover a wide range of levels of description, ranging from relatively coarse, phenomenological measures of effective connectivity between large neuronal populations (DCM for fMRI; Friston et al., 2003) to estimates of the conductance of ion channels at specific synapses (conductance-based DCM for M/EEG; Gilbert et al., 2016; Moran, Symmonds, Stephan, Friston, & Dolan, 2011).

In what follows, we first briefly revisit the basic concepts of DCM. Second, we describe proof-of-concept studies that illustrate the utility of DCM for Computational Psychiatry. Third, we examine methodological challenges that need to be addressed in order to advance the clinical applicability of DCM, and summarize recent methodological extensions to the original framework that may offer solutions to these problems. We conclude this article by outlining important future steps for the field of computational psychiatry.

2 | DYNAMIC CAUSAL MODELING

DCM represents a Bayesian framework for inferring effective (directed) connectivity among latent (hidden) neuronal states from measured neuroimaging data (Friston et al., 2003). This rests on modeling a neuronal circuit as a multiple-input-multiple-output (MIMO) system, using a likelihood function with two hierarchical layers: a model of hidden neuronal (and, for fMRI, hemodynamic) states and an observation model that links hidden states to measured data. Jointly, this provides a probabilistic forward mapping from the parameters of the system (e.g., synaptic connection strengths) to changes in fMRI (Friston et al., 2003) or M/EEG (David et al., 2006; Kiebel, David, & Friston, 2006; Moran, Stephan, Dolan, & Friston, 2011) signals. Augmenting this forward mapping with plausible prior distributions over parameters turns the model into a full generative model (Box 1), for which the exact form depends on the modality of the acquired neuroimaging data as well as the scientific question of interest.

2.1 | DCM for fMRI

DCM was initially introduced for fMRI data (Friston et al., 2003). In the original article, the dynamics of interacting neuronal populations were described using a bilinear differential equation

$$\frac{dx}{dt} = \left(A + \sum_j B^{(j)} u_j\right) x + Cu \tag{1}$$

This derives from a Taylor approximation to the evolution function of an arbitrary deterministic dynamical system (Stephan et al., 2008). Equation (1) captures how the dynamics of the neuronal states x unfold as a function of the synaptic coupling between network nodes or brain regions (endogenous connectivity A) and experimentally controlled manipulations u that perturb the system. Experimental manipulations either directly affect the neuronal states (driving inputs C) or modulate the endogenous connections between the different nodes (modulatory inputs B). Over the last decade, various extensions to this bilinear neuronal state equation have been introduced. For instance, nonlinear DCM accounts for how endogenous connections can be altered dynamically by inputs from other brain regions, thus modeling processes related to short-term synaptic plasticity and synaptic gain control (Stephan et al., 2008). Similarly, neuronal fluctuations have been embedded into the framework, yielding both stochastic DCM (Daunizeau, Friston, & Kiebel, 2009; Li et al., 2011) and spectral DCM (Friston, Kahan, Biswal, & Razi, 2014), two variants that enable the analysis of the "resting state" (i.e., unconstrained cognition in the absence of external perturbations).

Different clinical questions might be addressed more naturally by the DCM variants described above. If pathophysiological mechanisms are expected to relate to aberrant synaptic plasticity due to regionally specific abnormal modulatory influences as, for instance, in network models of bipolar disorder (Breakspear et al., 2015) or the dysconnection hypothesis of schizophrenia (Friston, Brown, et al., 2016), nonlinear DCM might represent a natural choice. Conversely, stochastic DCM and spectral DCM are useful candidates for testing network abnormalities during the "resting state" (e.g., Bastos-Leite et al.,

2015; Hyett, Breakspear, Friston, Guo, & Parker, 2015). The choice of the optimal DCM variant should therefore be tailored to the specific hypothesis about disease-relevant processes.

Regardless of the exact form of the neuronal state equations, they are coupled to a hemodynamic model that translates the predicted neuronal dynamics into region-wise blood oxygen level dependent (BOLD) signals via a cascade of differential equations. This rests on the Balloon-Windkessel model (Buxton, Wong, & Frank, 1998), which was augmented to account for neurovascular coupling (Friston, Mechelli, Turner, & Price, 2000). In brief, the hemodynamic model describes how changes in the neuronal states induce changes in cerebral blood flow, which, in turn, affect venous blood volume and deoxyhemoglobin content (for recent extensions, see Havlicek et al., 2015). These two quantities then enter a static BOLD signal observation equation that yields a prediction of BOLD signal time courses (Stephan, Weiskopf, Drysdale, Robinson, & Friston, 2007) (for a graphical summary of DCM for fMRI, see Figure 2). While the hemodynamic parameters are typically of little interest in effective connectivity analyses, they account for regional variations in the shape of hemodynamic responses and thus help avoid erroneous interpretations (David et al., 2008). More comprehensive reviews on DCM for fMRI can be found elsewhere (Daunizeau et al., 2011; Friston et al., 2013; Kahan & Foltynie, 2013; Stephan et al., 2010).

2.2 | DCM for M/EEG

The original neuronal model in DCM for fMRI (see Equation (1)) contains a rather abstract description of neuronal population dynamics and thus cannot provide a detailed account of synaptic processes underlying brain function. The motivation for its relatively coarse nature is that fMRI data represent a low-pass filtered transformation of synaptic activity, and this places a limit on system identifiability (but see Friston et al., 2017).

On the contrary, electrophysiological measurements support more sophisticated models of neuronal dynamics as they contain far richer temporal information. In its original description for event-related responses (ERPs), DCMs of electromagnetic responses were cast in terms of a neural mass model; this assumes that the dynamics of an ensemble of neurons can be represented by its first moment (mean; David et al., 2005; David et al., 2006). In these models, the neural masses for each source were based on the Jansen-Rit model (Jansen & Rit, 1995), which comprises three interacting neuronal subpopulations. In DCM for ERPs, these three subpopulations represent excitatory spiny stellate cells in granular layer IV, whereas both inhibitory interneurons and excitatory pyramidal cells are assigned to supragranular and infragranular cortical layers[1] (see Figure 3; David et al., 2006). They are interconnected via intrinsic (within-source) connections, while different sources are connected via extrinsic (between-source) connections according to established anatomical connectivity rules (Felleman & Van Essen, 1991). The neural mass model essentially predicts the depolarization of pyramidal cells, which are assumed to represent the main source of measured M/EEG signals due to the spatial alignment of their dendritic trees.

In DCM for ERPs, the neuronal dynamics of each neuronal subpopulation rests on two operators: First, a convolution operator which transforms the average density of presynaptic inputs into an average postsynaptic membrane potential. Second, the output operator which converts the average postsynaptic membrane potential into an average firing rate. The predicted potential of pyramidal cells then enters an electromagnetic forward model (essentially a linear mapping specified by a lead field matrix that describes the conduction of electromagnetic fields; Mosher, Leahy, & Lewis, 1999).

Subsequently, refined variants of DCM for M/EEG have been developed. These extensions include conductance-based DCMs that consider the dynamics of specific ion channels and thus potentially allow for physiologically more elaborate assessments (Moran, Jung, et al., 2011). Additionally, neural field models have been introduced that treat the neuronal subpopulations as manifolds on the cortical surface (instead of point sources) by simultaneously modeling temporal and spatial variations in cortical activity using partial differential equations (Pinotsis, Moran, & Friston, 2012). Later developments finessed the model structure guided by the idea of the "canonical microcircuit" (Pinotsis et al., 2013), based on previous work by Douglas and Martin (1991) in visual cortex. Specifically, four instead of three neuronal subpopulations were introduced to explicitly accommodate sources of forward and backward connections in cortical hierarchies by distinguishing superficial and deep pyramidal cells, with distinct spectral outputs (Bastos et al., 2012; Moran, Pinotsis, & Friston, 2013). Again, comprehensive reviews of the different variants of DCM for M/EEG can be found elsewhere (Daunizeau et al., 2011; Moran, Pinotsis, et al., 2013).

2.3 | Variational Bayes

Given a generative model, one can exploit approximate Bayesian estimation techniques (Bishop, 2006) to infer the model's parameters from data. This is known as model inversion or simply inference, and yields an estimate of the posterior density over model parameters, which describes the probability density of each parameter given the measured data (see Box 1). Model inversion within the DCM framework rests on variational Bayes under the Laplace approximation (VBL; Friston, Mattout, Trujillo-Barreto, Ashburner, & Penny, 2007), which is computationally highly efficient. In brief, VBL for DCM assumes a mean-field split between parameters and hyperparameters and applies a Laplace or Gaussian approximation to the variational

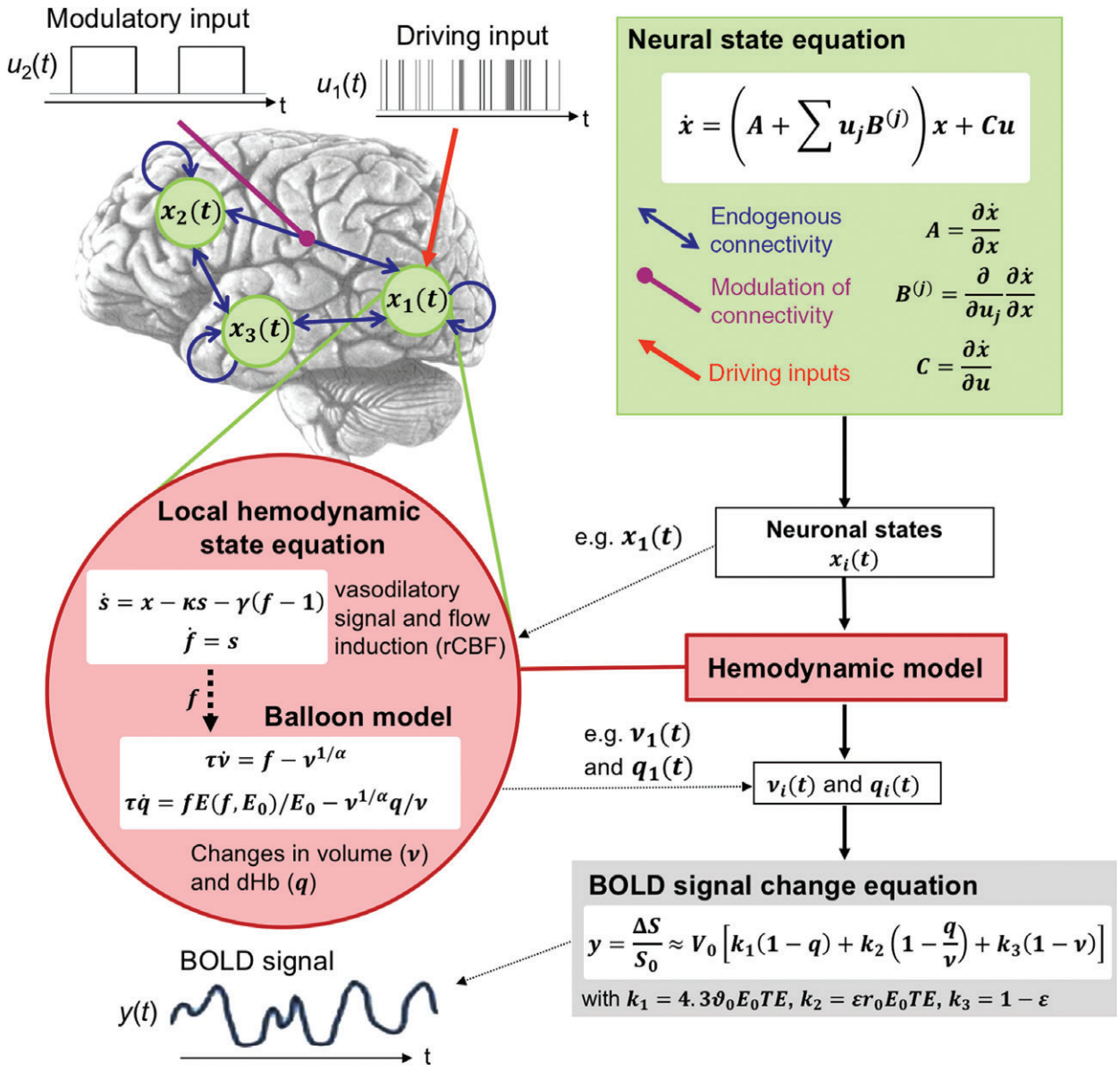


FIGURE 2 Graphical summary of the generative model of DCM for fMRI, comprising the neuronal and hemodynamic model, as well as the (static) nonlinear BOLD signal change equation. The neuronal state equation is cast as a bilinear differential equation, describing the dynamics of neuronal states as a function of the endogenous connectivity (A matrix), the modulatory influences (B matrix) and driving inputs (C matrix). The neuronal states then enter a cascade of differential equations, which make up the hemodynamic model and describe how neuronal dynamics lead to changes in cerebral blood flow, which, in turn, affect venous blood volume and deoxyhemoglobin content. These two quantities then enter a static BOLD signal observation equation that yields a prediction of BOLD signal time courses. A more comprehensive description is provided elsewhere (Daunizeau, David, & Stephan, 2011; Friston et al., 2013; Kahan & Foltynie, 2013; Stephan et al., 2010). (Reprinted with permission from Stephan et al. (2015). Copyright 2015 Elsevier)

densities (approximate posteriors). Under these assumptions, one not only obtains an approximation to the true posterior density $p(\theta|y,m)$, but also an estimate of the negative free energy. The negative free energy is a lower bound to the logarithm of the model evidence or marginal likelihood $p(y|m)$ (but see below) which represents a measure of model "goodness", taking into account both accuracy and complexity of a model (Bishop, 2006). It thus serves as a principled metric for selecting the most plausible amongst alternative hypotheses (models) of how the data were generated (Bayesian model selection, BMS; Penny, 2012; Stephan, Penny, Daunizeau, Moran, & Friston, 2009). As we shall see later, inference both at the level of model structure and model parameters can be informative for clinical applications.

3 | APPLICATION OF DCM TO CLINICAL QUESTIONS

This section presents a selective overview of studies that provide initial examples of the potential utility of generative models for clinically relevant questions, such as differential diagnosis and the dissection of spectrum disorders (Figure 4). Notably, this

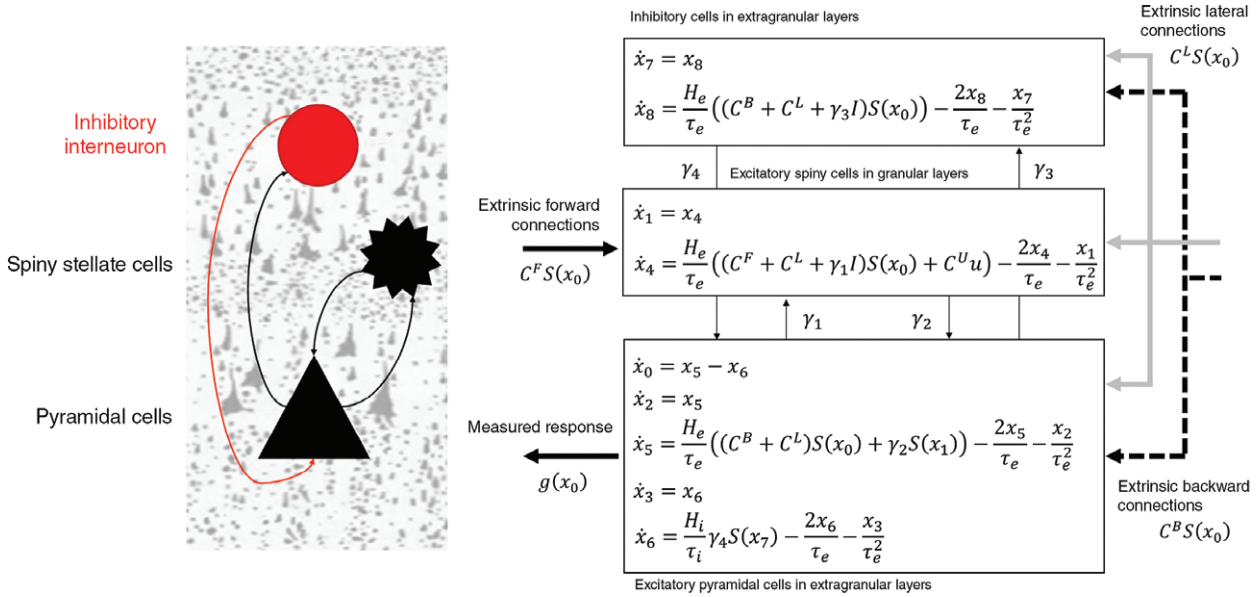


FIGURE 3 Graphical summary of the generative model of DCM for EEG/MEG, representing a single source by a neural mass model based on the Jansen-and-Rit model (1995). The neural mass model comprises three interacting neuronal subpopulations (*left*). In DCM for EEG/MEG, these subpopulations are taken to mimic excitatory spiny stellate cells in granular layer IV, inhibitory interneurons in supragranular layers II and III, and excitatory deep pyramidal cells in infra-granular layers V and VI. Subpopulations are interconnected via intrinsic (i.e., within-source) connections $\gamma_{1,2,3,4}$. Dynamics of the neuronal states are described by a set of differential equations (*right*). The model effectively yields a prediction of the depolarization of pyramidal cells (which is assumed to underlie the measured M/EEG signals) by first transforming average density of presynaptic inputs into an average postsynaptic membrane potential (i.e., convolution), which is then converted into an estimate of the average rate of action potentials fired by each neuronal subpopulation. (Reprinted with permission from David et al. (2006). Copyright 2006 Elsevier and Moran, Pinotsis, and Friston (2013). Copyright 2013 Frontiers)

is not meant to represent a comprehensive list and many more studies have applied DCM to psychiatric and neurological disorders than can possibly be covered here, including studies on schizophrenia (Deserno, Sterzer, Wüstenberg, Heinz, & Schlagenhauf, 2012; Dima et al., 2009; Lefebvre et al., 2016; Li et al., 2017), depression (Almeida et al., 2009; Schlösser et al., 2008; Vai et al., 2016), autism (Grèzes, Wicker, Berthoz, & de Gelder, 2009; Radulescu et al., 2013), Parkinson's disease (Dirkx et al., 2016; Marreiros, Cagnan, Moran, Friston, & Brown, 2013), or epilepsy (Papadopoulou et al., 2017). Here, we selectively focus on a few studies that not only illustrate how DCM might contribute to the understanding of pathophysiology, but also provide an intuition of how a generative modeling approach could, eventually, improve clinical care in psychiatry.

3.1 | Differential diagnosis

3.1.1 | Bayesian model selection

A fundamental challenge for psychiatry is the problem of differential diagnosis. van Leeuwen, den Ouden, and Hagoort (2011) presented a compelling example how BMS and DCM could enable a formal approach to differential diagnosis. Strictly speaking, this study did not address a clinical condition, but a rare cognitive peculiarity in the healthy population: synesthesia (Hochel & Milan, 2008), or more specifically, grapheme–color synesthesia (the simultaneous experience of color when seeing written letters). In this condition, enhanced activation of the color-sensitive area V4 and the superior parietal lobe (SPL) in synesthetes had been reported by several studies (Hubbard, Arman, Ramachandran, & Boynton, 2005; Sperling, Prvulovic, Linden, Singer, & Stirn, 2006; Weiss & Fink, 2009). Two competing hypotheses of this finding had been proposed: Enhanced V4 activity during synesthesia might either arise from direct bottom-up cross-activation from grapheme processing areas in the fusiform gyrus (i.e., cross-wiring; Brang, Hubbard, Coulson, Huang, & Ramachandran, 2010; Ramachandran & Hubbard, 2001), or from indirect top-down effects originating in higher-order parietal areas (i.e., disinhibition feedback; Grossenbacher & Lovelace, 2001). van Leeuwen, den Ouden, and Hagoort (2011) constructed two competing DCMs which captured these opposing mechanisms. Random-effects BMS (Stephan, Penny, et al., 2009) was then used to test which of the two models provided the most accurate description of the measured fMRI data. Interestingly, across all synesthetes, there was no strong preference for one or the other model. However, when dividing subjects according to their subjective reports into "projectors" (who experience the physical colocalization of color and letters) and "associators" (who experience an internal association of color induced by letters), the two competing models mapped almost perfectly onto the different subgroups (Figure 5).

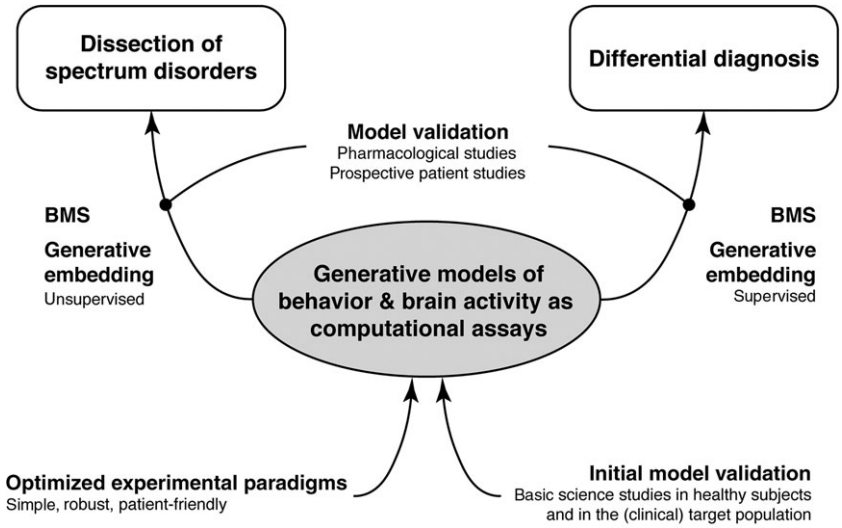


FIGURE 4 Schematic summary of key prospective endeavors in Computational Psychiatry and the necessary methodological building blocks. Ultimately, Computational Psychiatry strives to enable generative models of brain activity (and behavior) as computational assays for differential diagnosis and dissection of spectrum disorders in routine clinical practice. (Reprinted with permission from Stephan and Mathys (2014). Copyright 2014 Elsevier)

While not describing a clinical case, these results speak to the potential of BMS for distinguishing individuals with similar phenotypes (symptoms) based on differences in the underlying physiological mechanisms (here, effective connectivity). Generally, BMS provides a principled framework for differential diagnosis where the plausibility of different explanations (models of disease mechanisms) for a given set of clinical observations can be evaluated formally, in terms of the posterior probability of a model (Figure 6).

3.1.2 | Generative embedding

An alternative computational approach to differential diagnosis is generative embedding (Brodersen et al., 2011). In brief, generative embedding combines generative models of (neuroimaging or behavioral) data with (un)supervised machine learning techniques, such as classification or clustering (Bishop, 2006). Specifically, in a first step, DCM is used to infer the posterior densities over model parameters (e.g., neuronal connectivity) from measured data. In a second step, features of these posterior densities (e.g., maximum a posteriori estimates, Bishop, 2006) enter a supervised or unsupervised learning technique. This provides a simple solution to some key challenges of machine learning approaches to neuroimaging data (Brodersen et al., 2011): First, classifying/clustering subjects directly in "raw data" space (e.g., voxel-wise fMRI time series) is typically difficult because the dimensionality of the feature space (i.e., the number of voxels) is very high compared to the number of available subjects. Second, even when a sparse set of meaningful features can be extracted, the results of machine learning techniques in voxel space can be difficult to interpret and do not allow for mechanistic interpretations. In other words, the strength of generative embedding is that a generative model like DCM essentially acts as a theory-driven data compression method that reduces the high-dimensional fMRI data into a small set of neurobiologically interpretable parameter estimates that then serve as mechanistically interpretable features for (un)supervised learning.

Brodersen et al. (2011) introduced the generative embedding framework to neuroimaging, illustrating the potential utility in two clinical datasets. In a first paper (Brodersen et al., 2011), the authors asked whether aphasic patients (with a lesion in the left frontal and/or temporal cortex) could be differentiated from healthy controls based on fMRI data acquired during a simple speech recognition task (Schofield et al., 2012). Importantly, the study only modeled activity in parts of the auditory cortex that were unaffected by the lesion, thus asking whether the presence or absence of a "hidden" lesion could be predicted based on DCM parameter estimates obtained from the healthy part of the brain. The authors used a previously established linear DCM (i.e., only A and C matrix in Equation (1)) of the auditory system; this model comprised the medial geniculate body, Heschl's gyrus, and planum temporale, each in both hemispheres (Schofield et al., 2012). DCMs were then inverted for each individual separately and the inferred neuronal connectivity parameters entered a support vector machine (SVM). Using this approach, aphasic patients and healthy controls could be classified almost perfectly (balanced accuracy of 98% under leave-one-out cross-validation; Figure 7a,b). Importantly, the generative embedding approach significantly outperformed classification approaches that operated directly on regional BOLD signals or measures of functional connectivity.

Using a second, distinct fMRI dataset (Deserno et al., 2012), the same method was used to differentiate schizophrenic patients from healthy controls with relatively high accuracy (78%, leave-one-out cross-validation) using linear SVMs. Again, prediction accuracy of the DCM-based estimates was significantly higher than for other features derived from regional BOLD activity or functional connectivity.

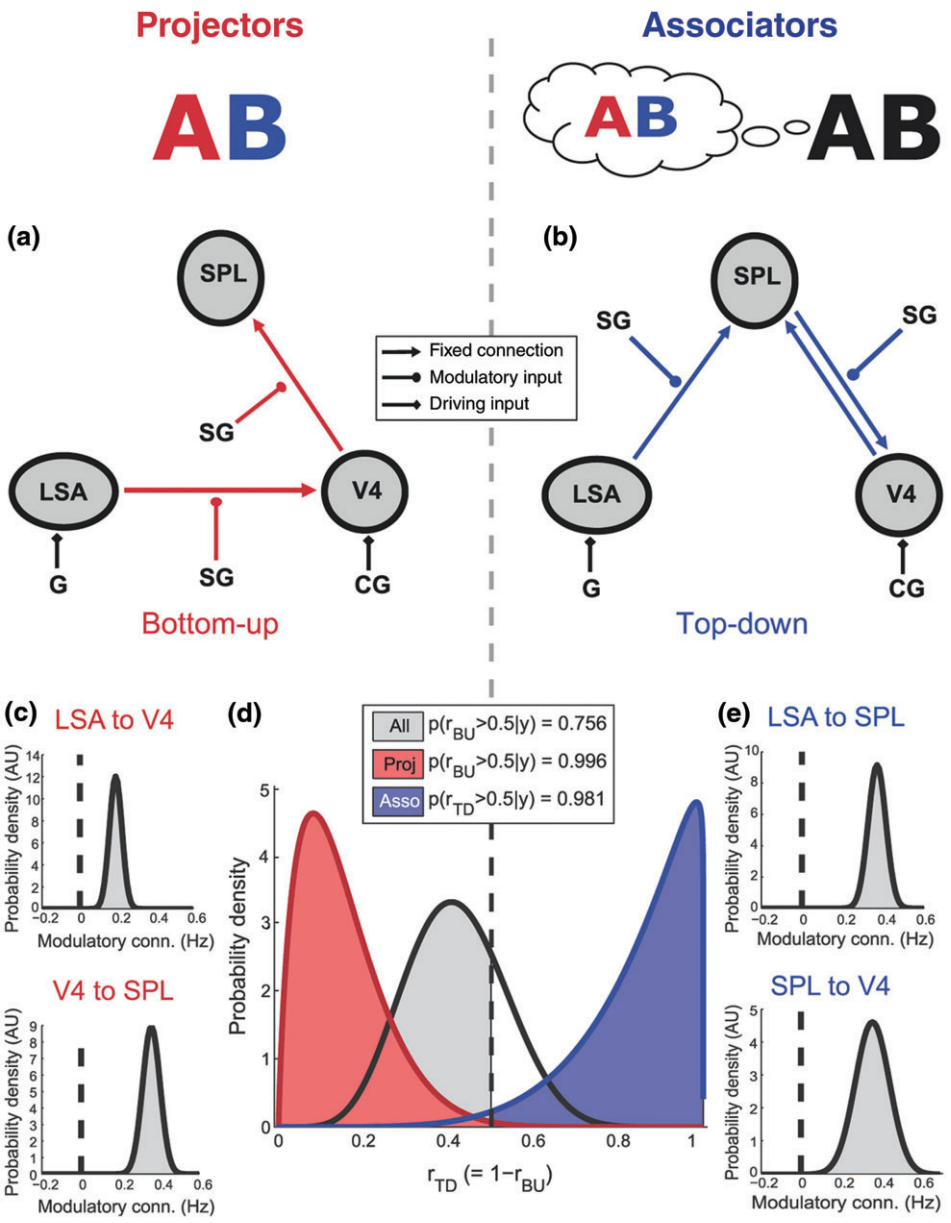


FIGURE 5 Example demonstrating Bayesian model selection (BMS) in DCM as a formal tool for differential diagnosis. Subjects with different forms of grapheme–color synesthesia were analyzed, namely projector synesthesia (*left*) and associator synesthesia (*right*). (a) Two alternative hypotheses of the putative effective connectivity underlying synesthesia, formulated as a bottom-up DCM and (b) top-down DCM. (d) BMS results with shaded areas representing the posterior probability distribution of the winning model. Results suggest that no (strong) evidence was found for either model when comparing DCMs across the entire populations (*grey*). However, dividing subjects into projectors (*red*) and associators (*blue*) based on their synesthetic experience, the two competing models mapped almost perfectly onto the different subgroups. (c) Posterior densities of modulatory parameters for projectors and (e) associators. (Reprinted with permission from van Leeuwen et al. (2011). Copyright 2011 Society for Neuroscience)

These studies highlight the potential benefit of DCM (or other generative models) as a mechanistically interpretable feature extraction or dimensionality reduction method. We emphasize that the studies described above do not yet achieve anything that is truly useful for clinical practice: diagnosing patients with aphasia or schizophrenia, respectively, does not represent a burning clinical problem. The former is a clinically straightforward diagnosis, and the latter is defined by DSM/ICD criteria; any classifier trained with respect to these criteria simply replaces clinical interviews with a more expensive technology that is calibrated identically and does not change clinical practice (compare the discussion in Stephan, Schlagenhauf, et al., 2017). By contrast, a generative embedding approach would have potential clinical utility if it managed to predict clinical outcomes (Harle et al., 2015) or distinguished diagnoses that have predictive validity (e.g., distinguishing between different forms of movement disorders, such as progressive supranuclear palsy and Parkinson's disease; Zhang et al., 2016).

So far, we have discussed supervised applications of generative embedding that are useful for predicting known clinical entities of interest. A different approach is required when the goal is to establish procedure for differential diagnosis and delineate (hitherto unknown) subgroups in heterogeneous spectrum diseases, as is almost universally the case in psychiatry (Owen, 2014; Stephan, Binder, et al., 2016). This is the scenario we turn to now.

3.2 | Dissection of spectrum disorders

Clinical categories based on syndromatic classifications such as DSM or ICD have limited predictive validity with regard to clinical trajectories and treatment prediction for individual patients (Cuthbert & Insel, 2010, 2013; Kapur et al., 2012). This is because these descriptive categories refer to groups of patients with similar phenotypes that are likely caused by different

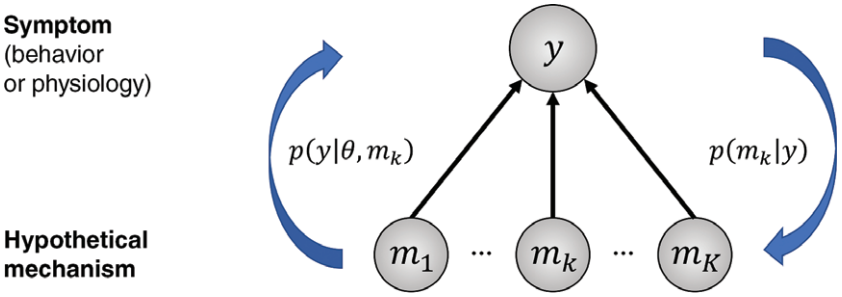


FIGURE 6 Bayesian model selection (BMS) as a principled framework for differential diagnosis. Given measurements (e.g., clinical observations), the relative plausibility of a set of competing hypotheses (models) of how the observations have been generated, can be evaluated in terms of the posterior model probability. (Reprinted with permission from Stephan, Schlagenhauf, et al. (2017). Copyright 2017 Elsevier)

pathophysiological mechanisms (Cuthbert & Insel, 2013; Kapur et al., 2012). Similarly, there are no clear-cut boundaries between DSM-defined clinical categories as is indicated by the significant comorbidity structure of psychiatric diseases (Borsboom, Cramer, Schmittmann, Epskamp, & Waldorp, 2011). Establishing computational tools that replicate DSM diagnoses is therefore simply a more complex way of replicating diagnoses whose lack of predictive utility is known (Cuthbert & Insel, 2010, 2013; Kapur et al., 2012; Stephan et al., 2015). Instead, tools are needed that dissect heterogeneous spectrum diseases into subgroups that share underlying pathophysiological mechanisms and enable more reliable predictions of clinically relevant outcomes.

This can be addressed by unsupervised machine learning techniques, such as clustering, which can carve out subgroups in a population by identifying structure in the data (Hastie, Tibshirani, & Friedman, 2009). Again, embedding this into the

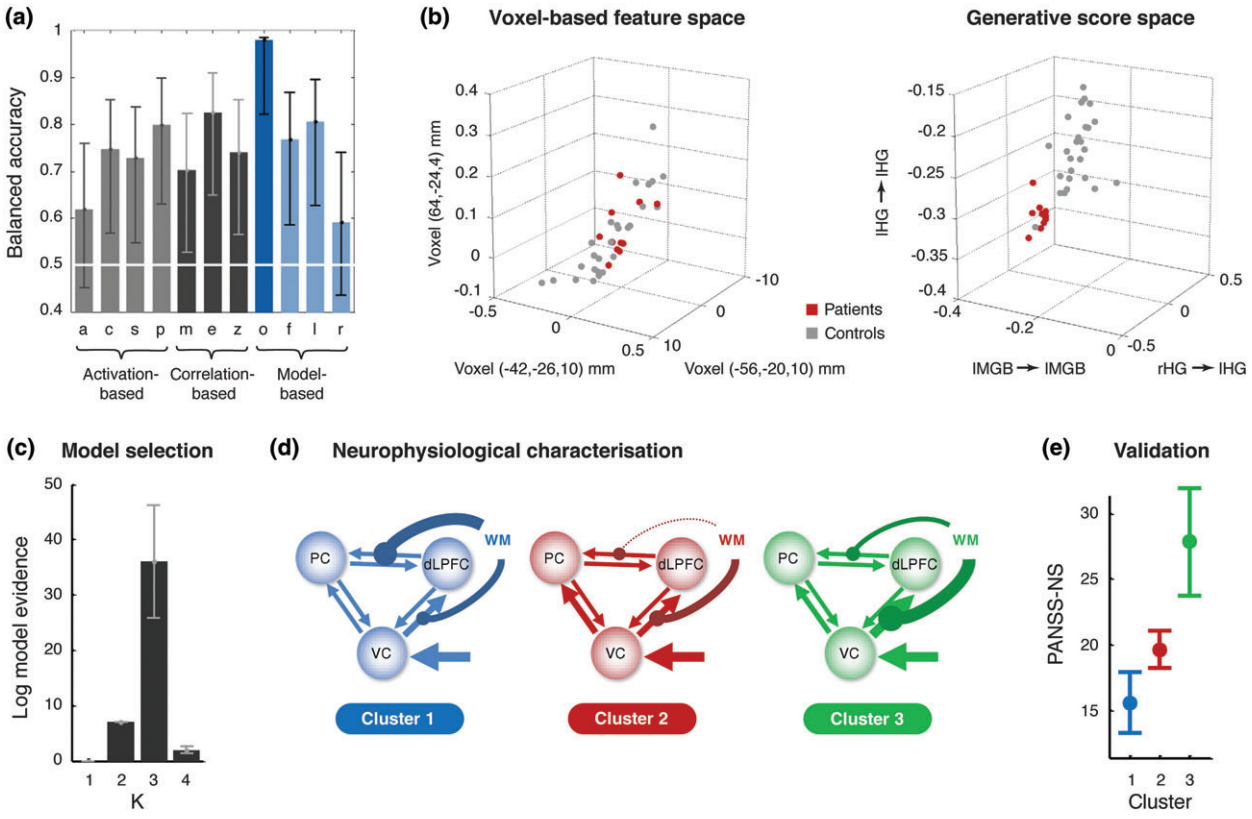


FIGURE 7 Examples demonstrating generative embedding based on DCM as a formal tool for differential diagnosis and dissection of spectrum disorders. (a) Classification accuracy of the supervised generative embedding approach for various measures. Input features were either based on measures of BOLD activity (*light grey*), functional connectivity (*dark grey*), or effective connectivity (*blue*). For all measures, the balanced accuracy and its 95% posterior probability interval is shown, as well as chance level (50%). Generative embedding based on the posterior means of the model parameters of a plausible DCM significantly outperformed more conventional classification approaches that operated on regional BOLD activity or measures of functional connectivity. Furthermore, balanced accuracy was markedly reduced for biologically unlikely models. (b) Representation of aphasic patients (*red*) and healthy controls (*grey*) in the reduced voxel space—that is, the space spanned by the BOLD activity in the three peaks of the most discriminative activation clusters (*left*), as well as in the reduced generative score space—that is, the space spanned by the three individually most discriminative effective connectivity parameters (*right*). (c) Results of the unsupervised generative embedding approach based on the variational Bayesian inversion of a Gaussian mixture model, operating on the posterior parameter estimates of a three-region DCM. Results suggested highest model evidence for the number of clusters being equal to three. (d) Different effective connectivity profiles for the three distinct subgroups. (e) Clusters of the schizophrenic patients differed significantly in the negative symptom severity scores on the Positive and Negative Syndrome Scale (PANSS). (Reprinted with permission from Brodersen et al. (2011). Copyright 2011 PLOS and Brodersen et al. (2014). Copyright 2014 Elsevier)

inversion of a generative model can greatly enhance both performance and interpretability. Brodersen et al. (2014) demonstrated the potential utility of unsupervised generative embedding for dissecting spectrum diseases using an fMRI dataset of 41 patients with schizophrenia that performed a working memory task (Deserno et al., 2012). Using the posterior parameter estimates of a three-region DCM comprising visual (VC), parietal (PC) and dorsolateral prefrontal cortex (dlPFC), the authors showed that variational Gaussian mixture models (Bishop, 2006) detected the presence of three distinct patient subgroups that were characterized by distinct effective connectivity profiles (Figure 7c,d). Importantly, these three clusters mapped onto clinically distinct subgroups (Figure 7e), in the sense that schizophrenic patients from different clusters showed significant differences in their negative symptom severity scores on the Positive and Negative Syndrome Scale (PANSS). In other words, a purely physiologically informed and connectivity-based demarcation of subgroups showed a remarkable correspondence to a specific clinical symptom dimension. While this result is not of any clinical utility (given that the symptoms were known), it illustrates the potential of generative embedding and motivates prospective validation studies that test whether future clinical outcomes are related to distinct clinical subgroups. In other words, it remains to be tested whether a re-definition of spectrum diseases by generative models provides more accurate predictions of treatment response and disease trajectories than DSM/ICD diagnoses.

3.3 | Development of computational assays

So far, our discussion focused on macroscopic measures of brain connectivity obtained from DCM for fMRI which contains a rather abstract description of neuronal population dynamics (due to the low-pass filter properties of the hemodynamic response) and therefore only provides coarse representations of the underlying synaptic processes. On the contrary, DCM for electrophysiological responses supports much more fine-grained models of neuronal dynamics and enables inference on the relative strength of synaptic transmission at different cell types and via specific neurotransmitters. The feasibility of inferring synaptic processes from epidural local field potential recordings was demonstrated by Moran, Jung, et al. (2011) using an anesthetic agent (isoflurane) in rodents. By administering different doses of the anesthetic while recording local field potentials (LFPs) from auditory cortex, the authors demonstrated that a neural mass DCM could track changes in the excitatory-inhibitory balance of synaptic transmission across different levels of anesthesia. More precisely, consistent with established neurophysiological findings (Berg-Johnsen & Langmoen, 1992; Detsch, Vahle-Hinz, Kochs, Siemers, & Bromm, 1999; Larsen, Haugstad, Berg-Johnsen, & Langmoen, 1998), DCM parameter estimates representing the amplitude of excitatory postsynaptic potentials linearly decreased with increasing levels of anesthesia, whereas parameters related to inhibitory postsynaptic potentials displayed a nonlinear (saturating) increase.

In a second study, Moran and colleagues utilized conductance-based DCM, a more refined variant of DCM for M/EEG which distinguishes ionotropic receptors with sufficiently different time constants (e.g., α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA), *N*-methyl-D-aspartate (NMDA), γ-aminobutyric acid ($GABA_A$)), to assess dopaminergic modulation of NMDA and AMPA receptor conductances (Moran, Symmonds, et al., 2011). Recording MEG data during a working memory task in a within-subject, placebo-controlled, pharmacological intervention study, the inferred conductances of AMPA and NMDA receptors matched the profile established in previous electrophysiological studies in primates (Gao & Goldman-Rakic, 2003; Goldman-Rakic, 1996; Robbins, 2000): AMPA receptor conductance was reduced under L-Dopa, while the model parameter representing NMDA receptor sensitivity (nonlinearity) was enhanced. Importantly, the AMPA and NMDA receptor related parameter estimates significantly predicted drug-induced performance changes during working memory.

Furthermore, DCMs for M/EEG demonstrated changes in synaptic plasticity in the auditory cortex during a mismatch negativity (MMN) paradigm under ketamine administration (Schmidt et al., 2013), and explained increases in MMN amplitude under the acetylcholinesterase inhibitor galantamine in terms of increased postsynaptic gain of supragranular pyramidal cells in auditory cortex (Moran, Campo, et al., 2013). Similarly, a DCM of the effective connectivity between dorsal hippocampus and medial prefrontal cortex during ketamine administration in rats showed a decrease of reciprocal connectivity mediated via NMDA receptors that exhibited a monotonic dose–response relationship (Moran et al., 2015).

These studies demonstrate the utility of DCM for providing detailed estimates of transmitter- and receptor-specific transmission and highlight the potential of generative models as *in vivo* computational assays of pathophysiologically relevant synaptic processes. A possible way of how such computational assays could be used in clinical settings was demonstrated in a recent study by Gilbert et al. (2016). The authors constructed a physiologically detailed conductance-based DCM with ligand-gated sodium, calcium, and chloride channels, as well as with voltage-gated potassium and calcium channels. This model was applied to MEG data from a large cohort of 94 healthy controls. The ensuing parameter estimates served to construct a reference distribution against which two patients with monogenic channelopathies (i.e., diseases caused by the mutation of a single gene encoding a specific ion channel) were compared. Specifically, the two patients had mutations affecting the potassium channel gene KCNJ2 and the calcium channel gene CACNA1A, respectively. The conductance-based DCM

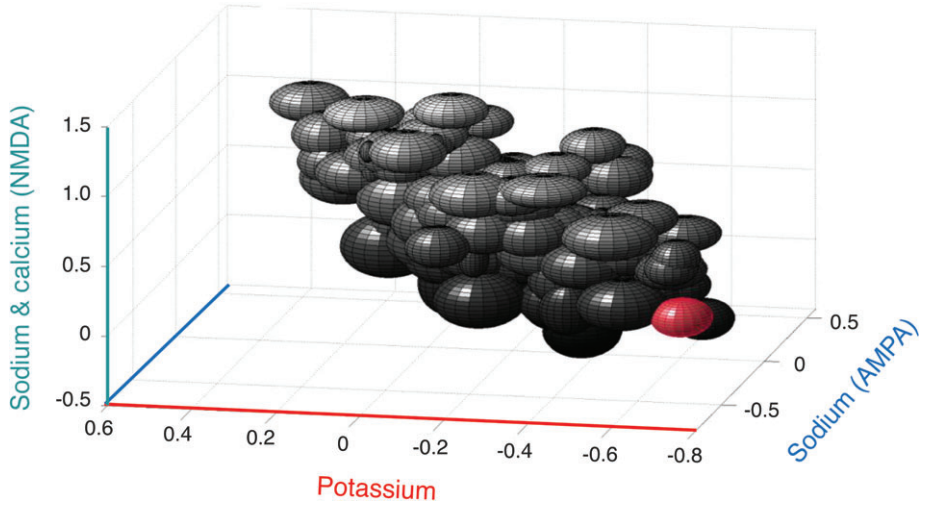


FIGURE 8 Development of computational assays based on DCM for M/EEG for model-based pathophysiological phenotyping. A conductance-based DCM with ligand-gated sodium, calcium, and chloride channels, as well as voltage-gated potassium and calcium channels was constructed. Shown are the posterior estimates of two ionotropic (AMPA, NMDA) and one potassium channel for a large cohort of 94 healthy controls (*dark grey ellipsoids*). These serve as a multivariate reference distribution against which a single patient (*red ellipsoid*), suffering from a mutation affecting the potassium channel gene KCNJ2, could be compared. This patient is placed at the edge of the multivariate distribution, suggesting that DCM could identify the synaptic channel abnormality with high sensitivity and specificity. (Reprinted with permission from Gilbert et al. (2016). Copyright 2016 Elsevier)

inferred ion channel abnormalities that were consistent with the known channelopathies in both patients and distinguished patients and controls with high sensitivity and specificity (Figure 8). This result illustrates that generative models of electromagnetic responses can infer sub-synaptic properties of neuronal circuits, including ion channel conductances and their mutations from non-invasively acquired M/EEG data. This is of considerable clinical relevance: an assay of dysfunctional synaptic signaling could not only guide the search for potential targets of novel treatments in heterogeneous disorders, but also serve as predictors of treatment response in individual patients.

These studies on DCM for M/EEG data illustrate how fine-grained physiological inference can be obtained when exploiting the rich temporal information of electrophysiological data. However, even the much coarser estimates of glutamatergic long-range connections in DCM for fMRI can prove useful for clinical applications, as demonstrated by the generative embedding examples discussed above (Brodersen et al., 2011, 2014). An alternative fMRI approach to computational assays utilizes trial-wise computational quantities with putative neurochemical interpretability that are obtained from generative models of behavior. For example, certain types of prediction errors or precision (inverse uncertainty) weights may reflect the release of dopamine or acetylcholine (for review, see Iglesias, Tomiello, Schneebeli, & Stephan, 2016). Using quantities like prediction errors to define regressors in standard general linear model analyses of fMRI ("model-based fMRI"; Glascher & O'Doherty, 2010) could enable one to assay individual differences in neuromodulatory systems (for possible caveats with regard to interpreting prediction error signals in BOLD data, see Cevora & Henson, 2017). This model-based fMRI approach is increasingly finding application in pathophysiological studies of mental disorders (for a review in the context of schizophrenia, see Stephan et al., 2015). An alternative approach, less widely used so far, is to include computational quantities as modulatory inputs in DCMs for fMRI (den Ouden, Daunizeau, Roiser, Friston, & Stephan, 2010; den Ouden, Friston, Daw, McIntosh, & Stephan, 2009; Roy et al., 2014). Provided these quantities can be interpreted neurochemically, this modeling approach could serve to infer the influence of neuromodulatory transmitters on short-term plasticity at glutamatergic synapses (Stephan et al., 2008).

4 | LIMITATIONS AND METHODOLOGICAL ADVANCES

Despite the potential of generative models, a number of methodological challenges represent limiting factors for the clinical applicability of these methods at the moment. Here, we restrict our discussion to two key limitations (further issues are addressed in Section 5). First, the variational Bayesian framework for model inversion in DCM has several potential weaknesses. Second, DCM is currently limited to small network models (on the order of 10 regions). In the following sections, we discuss these limitations and highlight recent methodological advances that may further enhance the utility of DCM for Computational Psychiatry.

4.1 | Robustness of statistical inference technique

As highlighted above, one central goal of Computational Psychiatry is the development of computational assays for predicting clinical trajectories and treatment responses in individual patients. For this endeavor, the stability of the inference procedure and the reliability of the ensuing posterior parameter estimates become paramount (Woolrich & Stephan, 2013). In other words, if generative models cannot reliably inform clinical decisions, they will be of no practical use for psychiatry because treatment recommendations or diagnoses might randomly change across multiple measurements.

As described above, the VBL scheme in DCM uses (distributional) assumptions that can render model inversion vulnerable (Daunizeau et al., 2011). One issue is that VBL rests on gradient ascent and is thus inherently susceptible to local maxima. Furthermore, even when the global maximum is found, inference might still be affected by the approximations currently used in DCM (Friston et al., 2007). For example, when the Laplace approximation to the negative free energy (i.e., the second order Taylor series expansion of the log joint around the approximate (variational) posterior means) is violated, the free energy is no longer guaranteed to represent a lower bound on the log model evidence (Wipf & Nagarajan, 2009).

While previous experimental studies reported good reproducibility of DCM across multiple sessions and different subjects, suggesting robust model inversion (Rowe, Hughes, Barker, & Owen, 2010; Schuyler, Ollinger, Oakes, Johnstone, & Davidson, 2010), recent work on the test–retest reliability of DCM provided a more mixed picture (Frässle et al., 2015). Here, test–retest reliability refers to the stability of model parameter estimates obtained when applying the method to multiple datasets acquired under the same condition in the same subject over time. Comparing the reliability of deterministic DCM for two different software versions—classical DCM (cDCM) and DCM10 as implemented in SPM5 and SPM8, respectively—showed that reliability was indeed acceptable for cDCM. However, a marked reduction in the stability of model selection and model parameter estimation was observed for the more recent DCM10 version—a finding that was attributed to differences in the prior distributions across the two software versions. Specifically, the study concluded that the stronger regularization afforded by the tighter cDCM priors rendered the objective function landscape smoother by dampening local extrema that are far away from the *a priori* plausible regime and thus made it easier for the gradient ascent to reach the same maximum over multiple measurements. On the contrary, the probability of local extrema in the objective function appeared to be increased under the more flexible DCM10 priors, leading to a reduction in reliability because the algorithm got trapped in different local extrema in each session.

In summary, these results suggest that local extrema in the objective function and the choice of prior distributions can become limiting factors for the stability of model inversion and thus the clinical applicability of DCM. Next, we discuss two methodological advances that address these limitations: (a) sampling-based global optimization schemes, and (b) empirical Bayesian procedures.

4.1.1 | Markov chain Monte Carlo

Sampling-based inversion schemes, typically based on Markov chain Monte Carlo (MCMC), represent an appealing alternative to VBL. This is because sampling-based schemes do not require distributional assumptions about the posterior density and are guaranteed to converge to the exact posterior in the limit of infinite samples. Hence, MCMC is, in principle, capable of finding the global maximum even for complicated multimodal objective functions and dealing with singularities in the posterior more gracefully. However, this comes at the cost of high computational demands (i.e., run-times), which have prohibited the application of sampling-based schemes for inverting generative models of neuroimaging data until recently. This is aggravated by the fact that, unlike VB, sampling-based inversion techniques do not offer an estimate of the (log) evidence for free. While several MCMC strategies have been devised to provide log evidence estimates, with thermodynamic integration (TI; Calderhead & Girolami, 2009; Kirkwood, 1935; Lartillot & Philippe, 2006) as a current gold standard, these typically pose non-negligible additional computational demands. It is only recently that methodological advances have turned MCMC into a feasible alternative for the inversion of DCMs (Aponte et al., 2016; Chumbley, Friston, Fearn, & Kiebel, 2007; Penny & Sengupta, 2016; Raman, Deserno, Schlagenhauf, & Stephan, 2016; Sengupta, Friston, & Penny, 2015; Sengupta, Friston, & Penny, 2016).

For instance, recent advances have exploited the power of graphics processing units (GPUs) for speeding up sampling-based inversion schemes, as implemented in the "*massively parallel dynamic causal modeling*" (mpdcm) toolbox (Aponte et al., 2016). In *mpdcm*, which presently focuses on DCM for fMRI, the evaluation of the likelihood function as the computationally most expensive operation during sampling (because it requires integrating differential equations in the neuronal and hemodynamic models) is delegated to highly efficient GPUs. Similarly, different gradient-free and gradient-based MCMC sampling schemes have been introduced for electrophysiological DCMs, where nonlinearities are more pronounced and thus problems with local extrema more likely (Sengupta et al., 2015; Sengupta et al., 2016). Finally, significant advances are presently being made in developing TI implementations with acceptable computational requirements (Aponte et al., in preparation). This will facilitate obtaining robust estimates of log evidences for DCMs (and other generative models), regardless of data modality.

4.1.2 | Empirical Bayes

Empirical Bayes (EB) provides a principled way for "estimating" prior distributions by exploiting measurements from multiple subjects in the context of a hierarchical Bayesian model (Efron & Morris, 1973; Kass & Steffey, 1989). In hierarchical models, constraints on the posterior density over model parameters at any given level are provided by the level above. In other words, under the hierarchical structure of a multi-subject random or mixed-effects model, single-subject inference is informed by information from the entire population. These constraints are so-called empirical priors because they are informed by empirical data (of the entire group). For the standard parametric empirical Bayesian (PEB) framework, this essentially means that single-subject data are generated by adding random (Gaussian) variations to the group means (Friston, Litvak, et al., 2016).

Friston, Litvak, et al. (2016) have proposed a PEB model for DCM that includes Bayesian model reduction (BMR). BMR refers to the inversion of reduced models based on the posterior densities of the full model, and was originally introduced in the context of post hoc model optimization and model discovery (Friston, Li, Daunizeau, & Stephan, 2011; Rosa, Friston, & Penny, 2012). BMR is a highly efficient way to invert large number of models because the posterior of all reduced models can be evaluated analytically after a single (computationally expensive) inversion of the full model. BMR is, however, restricted to nested models and cannot be used to compare models with structurally different likelihood functions.

The PEB framework has been used in several empirical and methodological studies. For example, it served to examine the reproducibility of DCM for ERPs across independent datasets, distinct models, and different inversion schemes (Litvak, Garrido, Zeidman, & Friston, 2015). PEB was also used to study inter-subject variability of DCM for MEG, using visually triggered gamma oscillations, and to demonstrate the use of Bayesian cross-validation for assessing the predictive validity of DCM (Pinotsis, Perry, Litvak, Singh, & Friston, 2016).

An alternative approach that unifies DCM, mixture models and EB within a single hierarchical model was introduced by Raman et al. (2016). Their model combines the inference of subject-specific connectivity parameters with unsupervised learning of the population structure, that is, the detection of subgroups in the sample. This allows for empirical Bayesian inference, where subgroup-specific prior distributions inform the subjects' parameter estimates; and conversely, the definition of subgroups (clustering) is informed by the parameter estimates across subjects. Dissecting a heterogeneous spectrum of patients into more homogeneous subgroups while at the same time harvesting group-level information to finesse the local extrema problem inherent in the (first-level) inversion of DCMs has promising potential for future clinical applications. While the original model operated under an MCMC-based inversion scheme, recent work has introduced complementary variational Bayesian procedures for the inversion of this hierarchical DCM (Yao et al., in preparation).

While these are important methodological developments, the practical utility of both sampling-based global optimization schemes and empirical Bayesian techniques for overcoming DCM's current limitations with regard to local extrema in the objective function and the choice of prior distributions remains to be tested.

4.2 | Small network models

Apart from the statistical and computational limitations highlighted above, a more conceptual concern has also been raised regarding the restriction of DCM to relatively small networks (on the order of 10 regions). This restriction is necessary to keep model inversion numerically feasible (e.g., to avoid intractably large error covariance matrices). One potentially problematic consequence is, however, that it introduces the "missing region" problem: the possibility that ignoring interactions with a region outside the modeled system could affect inference on connectivity (Daunizeau et al., 2011; Roebroeck et al., 2011). This is less of a problem when one has clear-cut hypotheses about specific circuits that can be activated by carefully designed experimental manipulations. However, it might become a limiting factor for capturing pathophysiological processes of relevance for Computational Psychiatry. For example, in various mental disorders, such as schizophrenia (Bullmore, Frangou, & Murray, 1997; Friston, Brown, et al., 2016; Friston & Frith, 1995; Stephan et al., 2006; Stephan, Friston, et al., 2009), autism (Courchesne et al., 2007; Kennedy, Redcay, & Courchesne, 2006; Muller, 2007) or depression (Greicius et al., 2007; Mayberg, 1997; Wang, Hermens, Hickie, & Lagopoulos, 2012), global dysconnectivity has been postulated as a hallmark of the disease and a possible cause of symptoms; this points to the clinical utility of whole-brain models of functional integration (Menon, 2011). Consequently, a key endeavor in Computational Psychiatry is the development of large-scale network models with biophysically interpretable state equations and parameters that encode (patho)physiological mechanisms of neuronal population dynamics (Deco & Kringelbach, 2014; Stephan et al., 2015).

At present, these efforts are visible in two main development streams. First, whole-brain biophysical network models can be constructed by combining mean-field models of population activity with diffusion-weighted imaging data (Deco & Kringelbach, 2014). While biologically detailed, these models are not proper generative models and, due to their complexity, have very limited scope for parameter estimation; typically, only a single global scaling parameter can be estimated (Deco et al., 2013). Second, an established variant of DCM for "resting state" fMRI data (spectral DCM; Friston, Kahan, Biswal, &

Razi, 2014) was recently combined with BMR and a procedure to exploit functional connectivity estimates for defining shrinkage priors on effective connectivity (Seghier & Friston, 2013). This has made it possible to apply DCM to a network consisting of 36 brain regions (Razi et al., 2017). While representing a significant advance, it is presently not clear how far this approach can be pushed. Generally, extending DCM to whole-brain networks may require an approach that scales gracefully across several orders of brain network cardinality. One possible candidate for such an approach is regression DCM.

4.2.1 | Regression DCM

Regression DCM (rDCM) was recently introduced as a novel variant of DCM for fMRI that is specifically designed to deliver estimates of whole-brain effective connectivity (Frässle et al., 2017). In brief, rDCM rests on translating a linear DCM from the time into the frequency domain and reformulating model inversion as a special case of Bayesian linear regression. Drawing upon a mean-field approximation across regions, one can derive analytic variational update equations for the model parameters that enable extremely efficient inference—three to four orders of magnitude faster than in classical DCM. Given that run-time scales gracefully with the number of regions, rDCM can deal with very large networks, potentially with hundred regions.

A simple example is provided in Figure 9. This shows simulation results where rDCM adequately recovered effective connection strengths in a whole-brain network consisting of 66 regions, with a realistic human structural connectome and 300 free parameters to be estimated. Notably, this computation only took 3 seconds on a standard computer. More recent work with empirical fMRI data demonstrated the feasibility of rDCM for whole-brain networks with more than 100 regions. This augmented rDCM with sparsity constraints to enable automatic "pruning" of fully connected graphs (sparse rDCM; Frässle et al., in preparation).

These developments bring whole-brain physiological phenotyping of individual patients within reach and open up exciting possibilities for advancing the utility of generative models for clinical diagnosis and prognosis. Having said this, rDCM is only in its infancy and many limitations of the current implementation (e.g., fixed hemodynamic response function, lack of bilinear effects) need to be addressed in forthcoming extensions.

5 | FUTURE STEPS

In this article, we have reviewed DCM as a generative modeling framework for the development of computational assays that could improve diagnosis and treatment prediction for individual patients. What are the practical next steps that are needed to translate currently available generative models into clinically applicable tools?

As already hinted at in the previous section, one important step is to evaluate the success of ongoing methodological developments (e.g., sampling-based global optimization schemes and empirical Bayesian techniques) for improving test–retest reliability of DCM. So far, these developments concern DCM for fMRI, for which local extrema in the objective function and the choice of prior distributions have been identified as limiting factors for reliability (Frässle et al., 2015). On the contrary, systematic analyses of test–retest reliability of the more complex models in DCM for M/EEG are absent so far (but see above and Garrido, Kilner, Kiebel, Stephan, & Friston, 2007; Litvak et al., 2015, for an analysis of the related concept of reproducibility). Given the important role of DCMs of electrophysiological data for inferring pathophysiologically relevant quantities, evaluating the reliability of these variants as well represents an important step towards establishing their clinical utility.

It is worth emphasizing that test–retest reliability does not only depend on the stability of the computational modeling framework, but also on the measured data itself. The robustness of the data can be affect by various factors including scanner-related noise, physiological noise from the subject, head motion, task-unrelated cognitive processes, and changes in cognition over time (e.g., learning; Bennett & Miller, 2010). While test–retest reliability of fMRI has been studied frequently for various cognitive processes, there are only few tasks for which high reliability has been established. These typically involve motor or sensory processes (Aron, Gluck, & Poldrack, 2006; Maldjian, Laurienti, Driskill, & Burdette, 2002; Raemaekers et al., 2007), whereas tasks probing higher cognitive functions yield less stable results (Caceres, Hall, Zelaya, Williams, & Mehta, 2009; Fliessbach et al., 2010; Nord, Gray, Charpentier, Robinson, & Roiser, 2017; Schunck et al., 2008). Similarly, findings on the reliability of activation patterns and functional connectivity measures obtained during the "resting state" have been inconclusive so far. Initial studies reported high reliability of resting state data (Braun et al., 2012; Shehzad et al., 2009; Zuo et al., 2010); more recently, this has been called into question (Anderson, Ferguson, Lopez-Larson, & Yurgelun-Todd, 2011; Laumann et al., 2015; Noble et al., 2017), in particular for the short scan times commonly used in resting state studies. For example, Nobel and colleagues found poor reliability of resting state functional connectivity measures in a multi-site study and identified within-subject variance across sessions as the main source of variability in the connectivity estimates, outweighing other factors related to site, scanner, or day of scan (Noble et al., 2017). Hence, establishing experimental paradigms and fMRI protocols that provide robust activation of disease-relevant neural circuits and enable computational modeling of the pathophysiological mechanisms represents a key goal for forthcoming studies (Frässle, Paulus, Krach, & Jansen, 2016).

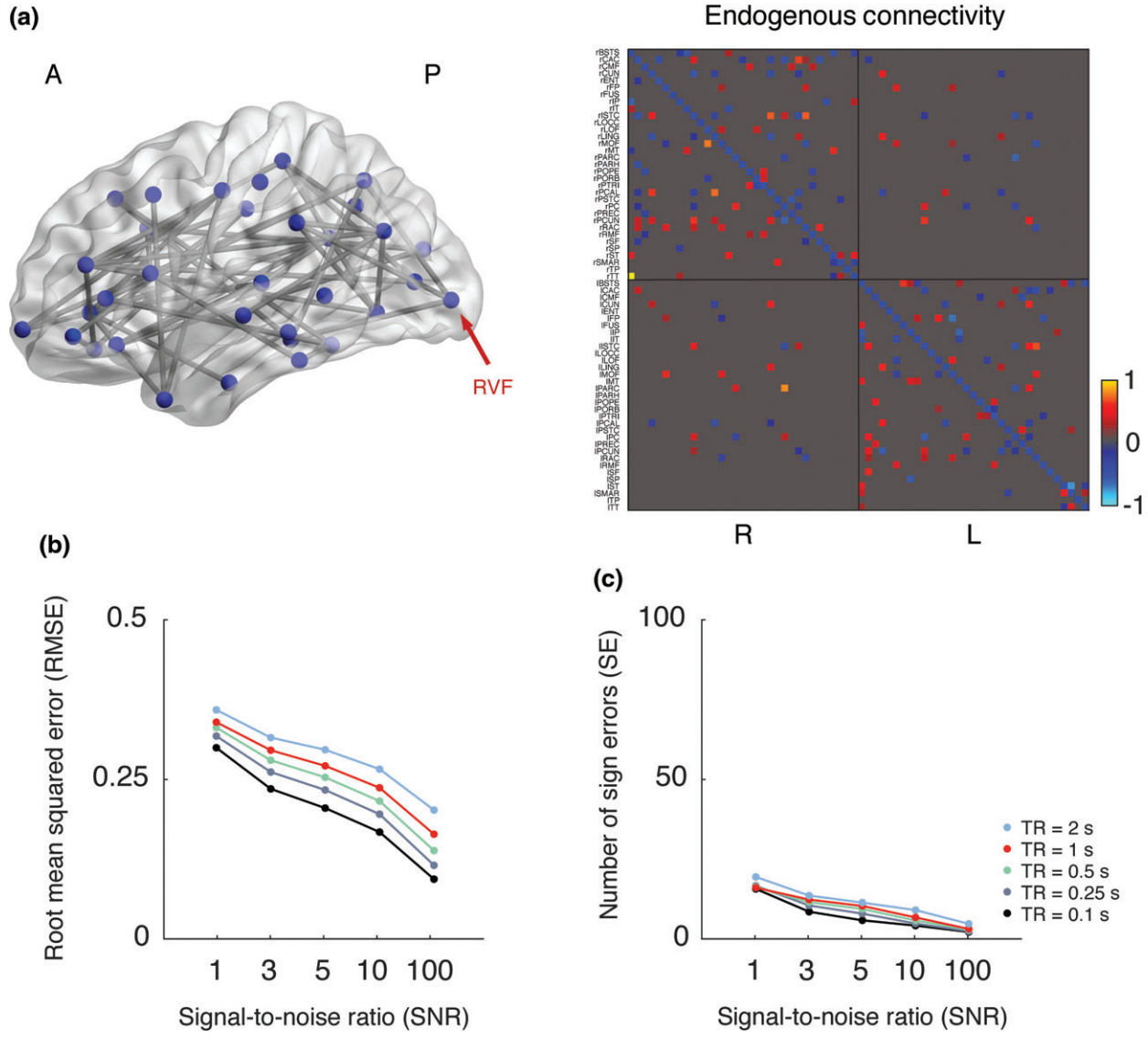


FIGURE 9 Regression DCM (rDCM) as a novel variant of DCM for inferring whole-brain effective connectivity patterns from fMRI data. (a) Endogenous connectivity architecture (A matrix) among the 66 brain regions from the parcellation reported by Hagmann et al. (2008) as well as the driving inputs, mimicking the effects of visual stimulation in the right and the left visual field (*left*), as well as an actual "observation" of the endogenous connectivity (*right*). L = left hemisphere; R = right hemisphere; A = anterior; P = posterior; LVF = left visual field; RVF = right visual field. (b) Parameter recovery of rDCM in terms of the root mean squared error (RMSE) and (c) the number of sign errors (SE) for various combinations of the signal-to-noise ratio (SNR) and the repetition time (TR) of the synthetic fMRI data. Results are shown when restricting the analysis to parameter estimates with a non-negligible effect size (i.e., the 95% Bayesian credible interval of the posterior not containing zero). (Reprinted with permission from Frässle et al. (2017). Copyright 2017 Elsevier)

Separately from reliability, another challenge is to evaluate the predictive validity of generative models with respect to major clinical questions. Using data from studies with controlled perturbations of pathophysiologically relevant processes in animals and humans (e.g., pharmacology), we need to challenge generative models to infer which perturbation was administered to the individual subject. As reviewed in the present paper, first attempts have been made in this direction (Gilbert et al., 2016; Moran, Symmonds, et al., 2011; Moran, Jung, et al., 2011; Moran et al., 2015); however, the sensitivity of currently available generative models for inferring pathophysiologically relevant quantities (e.g., status of ion channels and neurotransmitters) has not been investigated comprehensively yet.

Apart from perturbation studies, evaluation of the predictive validity of generative models like DCM also necessitates prospective patient studies with clinically relevant outcomes (e.g., treatment response). These could be either observational studies or clinical trials, and will be indispensable for demonstrating the efficacy and utility of computational assays for routine clinical practice (Paulus, Huys, & Maia, 2016; Stephan et al., 2015). For instance, such prospective studies could assess the utility of generative models for predicting whether an individual patient will benefit from an intervention, such as first-line treatment in first-episode patients or medication switch in chronic patients, based on neuroimaging data acquired prior to that intervention. This necessitates following up patients after the intervention to record clinical (symptom) trajectories

against which generative models can be challenged. The utility of generative models would then be evaluated by computing the accuracy of a model-based prediction (based on generative embedding) of who will respond to treatment and who will not. That is, do parameter estimates identify responders and non-responders with clinically useful levels of sensitivity and specificity? Only if model-based estimates improve prediction of treatment outcomes and clinical trajectories for individual patients, a generative modeling approach will be of value for the clinician. Unfortunately, only few studies with prospective designs exist to date—and those that do exist, tend to suffer from relatively small sample sizes (Kapur et al., 2012). While underpowered studies are a general concern in neuroscience, this is aggravated when high-dimensional data sets are investigated using machine learning tools like classification and/or clustering (Arbabshirani, Plis, Sui, & Calhoun, 2017). Acquiring large datasets that address clinically relevant questions, such as the prediction of disease trajectories and treatment success, is therefore important for translating generative models into clinical tools. Additionally, to make the most out of these valuable datasets, they should be shared amongst researchers. Unless such large shared datasets from prospective patient studies become available, rapid progress of Computational Psychiatry, Computational Neurology, and Computational Psychosomatics is unlikely. Fortunately, efforts are moving in this direction, as reflected by clinical studies like the Netherlands Study of Depression and Anxiety (NESDA; Penninx et al., 2008), as well as large-scale epidemiological (observational) studies like the UK Biobank (Sudlow et al., 2015), Rhineland study, and German National Cohort Study (Consortium German National Cohort, 2014). Close communication and exchange between experimentalists and modelers will be needed to ensure that these datasets are suitable for computational analyses.

6 | CONCLUSION

Generative models of neuroimaging and electrophysiological data have great potential for Computational Psychiatry and may help establish concrete solutions to some of the most urgent clinical problems in psychiatry. Their capacity for inferring pathophysiological mechanisms from non-invasively obtained measurements could guide differential diagnosis and treatment prediction in individual patients—and, ultimately, result in the development of standardized computational assays for routine clinical practice. In this article, we have reviewed DCM as a generative modeling framework for Computational Psychiatry, which, in principle, can elucidate the neurophysiological states of disease-relevant circuits. Clearly, major challenges related to statistical, conceptual and practical issues need to be tackled before DCMs can support clinical practice. In this respect, we have highlighted recent developments that address current methodological problems. Further methodological advances and careful validation studies, including prospective patient studies, may enable computational assays that usefully inform clinical decision making in psychiatry.

ACKNOWLEDGMENTS

We acknowledge support by the ETH Zurich Postdoctoral Fellowship Program and the Marie Curie Actions for People COFUND Program (to S.F.), as well as the René and Susanne Braginsky Foundation and the University of Zurich (to K.E.S.).

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

NOTE

[1]The figures in the original papers for DCM for ERPs (David et al., 2006) are sometimes misinterpreted as suggesting that the model assigns inhibitory interneurons exclusively to supragranular layers and pyramidal cells exclusively to infragranular layers. Inspection of the model's equations of inter-areal interactions reveals that both types of neurons exist in both layers.

REFERENCES

Almeida, J. R., Versace, A., Mechelli, A., Hassel, S., Quevedo, K., Kupfer, D. J., & Phillips, M. L. (2009). Abnormal amygdala-prefrontal effective connectivity to happy faces differentiates bipolar from major depression. *Biological Psychiatry*, *66*, 451–459.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5 R)*. Washington, DC: American Psychiatric, Association Publishing.

Anderson, J. S., Ferguson, M. A., Lopez-Larson, M., & Yurgelun-Todd, D. (2011). Reproducibility of single-subject functional connectivity measurements. *American Journal of Neuroradiology*, *32*, 548–555.

Aponte, E. A., Raman, S., Sengupta, B., Penny, W. D., Stephan, K. E., & Heinzle, J. (2016). mpdcm: A toolbox for massively parallel dynamic causal modeling. *Journal of Neuroscience Methods*, *257*, 7–16.

Arbabshirani, M. R., Plis, S., Sui, J., & Calhoun, V. D. (2017). Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *NeuroImage*, *145*, 137–165.

Aron, A. R., Gluck, M. A., & Poldrack, R. A. (2006). Long-term test-retest reliability of functional MRI in a classification learning task. *NeuroImage*, *29*, 1000–1006.

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, *76*, 695–711.

Bastos-Leite, A. J., Ridgway, G. R., Silveira, C., Norton, A., Reis, S., & Friston, K. J. (2015). Dysconnectivity within the default mode in first-episode schizophrenia: A stochastic dynamic causal modeling study with functional magnetic resonance imaging. *Schizophrenia Bulletin*, *41*, 144–153.

Bennett, C. M., & Miller, M. B. (2010). How reliable are the results from functional magnetic resonance imaging? *Annals of the New York Academy of Sciences*, *1191*, 133–155.

Berger, H. (1929). Über das Elektrenkephalogramm des Menschen. *Archiv für Psychiatrie und Nervenkrankheiten*, *87*, 527–570.

Berg-Johnsen, J., & Langmoen, I. A. (1992). The effect of isoflurane on excitatory synaptic transmission in the rat hippocampus. *Acta Anaesthesiologica Scandinavica*, *36*, 350–355.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York. 12, 13, 47, 105: Springer.

Borsboom, D., Cramer, A. O., Schmittmann, V. D., Epskamp, S., & Waldorp, L. J. (2011). The small world of psychopathology. *PLoS ONE*, *6*, e27407.

Braff, D. L., & Freedman, R. (2008). Clinically responsible genetic testing in neuropsychiatric patients: A bridge too far and too soon. *The American Journal of Psychiatry*, *165*, 952–955.

Brang, D., Hubbard, E. M., Coulson, S., Huang, M., & Ramachandran, V. S. (2010). Magnetoencephalography reveals early activation of V4 in grapheme-color synesthesia. *NeuroImage*, *53*, 268–274.

Braun, U., Plichta, M. M., Esslinger, C., Sauer, C., Haddad, L., Grimm, O., … Meyer-Lindenberg, A. (2012). Test-retest reliability of resting-state connectivity network characteristics using fMRI and graph theoretical measures. *NeuroImage*, *59*, 1404–1412.

Breakspear, M., Roberts, G., Green, M. J., Nguyen, V. T., Frankland, A., Levy, F., … Mitchell, P. B. (2015). Network dysfunction of emotional and cognitive processes in those at genetic risk of bipolar disorder. *Brain*, *138*, 3427–3439.

Brodersen, K. H., Deserno, L., Schlagenhauf, F., Lin, Z., Penny, W. D., Buhmann, J. M., & Stephan, K. E. (2014). Dissecting psychiatric spectrum disorders by generative embedding. *Neuroimage Clin*, *4*, 98–111.

Brodersen, K. H., Schofield, T. M., Leff, A. P., Ong, C. S., Lomakina, E. I., Buhmann, J. M., & Stephan, K. E. (2011). Generative embedding for model-based classification of fMRI data. *PLoS Computational Biology*, *7*, e1002079.

Bullmore, E. T., Frangou, S., & Murray, R. M. (1997). The dysplastic net hypothesis: An integration of developmental and dysconnectivity theories of schizophrenia. *Schizophrenia Research*, *28*, 143–156.

Buxton, R., Wong, E., & Frank, L. (1998). Dynamics of blood flow and oxygenation changes during brain activation: The balloon model. *Magnetic Resonance in Medicine*, *39*, 855–864.

Caceres, A., Hall, D., Zelaya, F., Williams, S., & Mehta, M. (2009). Measuring fMRI reliability with the intra-class correlation coefficient. *NeuroImage*, *45*, 758–768.

Calderhead, B., & Girolami, M. (2009). Estimating Bayes factors via thermodynamic integration and population MCMC. *Computational Statistics and Data Analysis*, *53*, 4028–4045.

Cevora, J., & Henson, R. N. (2017). Reconsidering the imaging evidence used to implicate prediction error as the driving force behind learning. *Frontiers in Psychology*, *8*, 1380.

Chumbley, J. R., Friston, K. J., Fearn, T., & Kiebel, S. J. (2007). A metropolis-hastings algorithm for dynamic causal models. *NeuroImage*, *38*, 478–487.

Consortium German National Cohort (2014). The German National Cohort: Aims, study design and organization. *European Journal of Epidemiology*, *29*, 371–382.

Courchesne, E., Pierce, K., Schumann, C. M., Redcay, E., Buckwalter, J. A., Kennedy, D. P., & Morgan, J. (2007). Mapping early brain development in autism. *Neuron*, *56*, 399–413.

Cuthbert, B., & Insel, T. (2010). Toward new approaches to psychotic disorders: The NIMH research domain criteria project. *Schizophrenia Bulletin*, *36*, 1061–1062.

Cuthbert, B. N., & Insel, T. R. (2013). Toward the future of psychiatric diagnosis: The seven pillars of RDoC. *BMC Medicine*, *11*, 126.

Daunizeau, J., David, O., & Stephan, K. (2011). Dynamic causal modelling: A critical review of the biophysical and statistical foundations. *NeuroImage*, *58*, 312–322.

Daunizeau, J., Friston, K., & Kiebel, S. (2009). Variational Bayesian identification and prediction of stochastic nonlinear dynamic causal models. *Physica D: Nonlinear Phenomena*, *238*, 2089–2118.

David, O., Guillemain, I., Saillet, S., Reyt, S., Deransart, C., Segebarth, C., & Depaulis, A. (2008). Identifying neural drivers with functional MRI: An electrophysiological validation. *PLoS Biology*, *6*, 2683–2697.

David, O., Harrison, L., & Friston, K. J. (2005). Modelling event-related responses in the brain. *NeuroImage*, *25*, 756–770.

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

Deco, G., & Kringelbach, M. L. (2014). Great expectations: Using whole-brain computational connectomics for understanding neuropsychiatric disorders. *Neuron*, *84*, 892–905.

Deco, G., Ponce-Alvarez, A., Mantini, D., Romani, G. L., Hagmann, P., & Corbetta, M. (2013). Resting-state functional connectivity emerges from structurally and dynamically shaped slow linear fluctuations. *The Journal of Neuroscience*, *33*, 11239–11252.

den Ouden, H. E., Daunizeau, J., Roiser, J., Friston, K. J., & Stephan, K. E. (2010). Striatal prediction error modulates cortical coupling. *The Journal of Neuroscience*, *30*, 3210–3219.

den Ouden, H. E., Friston, K. J., Daw, N. D., McIntosh, A. R., & Stephan, K. E. (2009). A dual role for prediction error in associative learning. *Cerebral Cortex*, *19*, 1175–1185.

Deserno, L., Sterzer, P., Wüstenberg, T., Heinz, A., & Schlagenhauf, F. (2012). Reduced prefrontal-parietal effective connectivity and working memory deficits in schizophrenia. *The Journal of Neuroscience*, *32*, 12–20.

Detsch, O., Vahle-Hinz, C., Kochs, E., Siemers, M., & Bromm, B. (1999). Isoflurane induces dose-dependent changes of thalamic somatosensory information transfer. *Brain Research*, *829*, 77–89.

Dima, D., Roiser, J. P., Dietrich, D. E., Bonnemann, C., Lanfermann, H., Emrich, H. M., & Dillo, W. (2009). Understanding why patients with schizophrenia do not perceive the hollow-mask illusion using dynamic causal modelling. *NeuroImage*, *46*, 1180–1186.

Dirkx, M. F., den Ouden, H., Aarts, E., Timmer, M., Bloem, B. R., Toni, I., & Helmich, R. C. (2016). The cerebral network of Parkinson's tremor: An effective connectivity fMRI study. *The Journal of Neuroscience*, *36*, 5362–5372.

Douglas, R. J., & Martin, K. A. (1991). A functional microcircuit for cat visual cortex. *The Journal of Physiology*, *440*, 735–769.

Efron, B., & Morris, C. (1973). Stein's estimation rule and its competitors—Empirical Bayes approach. *Journal of the American Statistical Association*, *68*, 117–130.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, *1*, 1–47.

Filiou, M. D., & Turck, C. W. (2011). General overview: Biomarkers in neuroscience research. *International Review of Neurobiology*, *101*, 1–17.

Fliessbach, K., Rohe, T., Linder, N., Trautner, P., Elger, C., & Weber, B. (2010). Retest reliability of reward-related BOLD signals. *NeuroImage*, *50*, 1168–1176.

Frässle, S., Lomakina, E. I., Razi, A., Friston, K. J., Buhmann, J. M., & Stephan, K. E. (2017). Regression DCM for fMRI. *NeuroImage*, *155*, 406–421.

Frässle, S., Paulus, F. M., Krach, S., & Jansen, A. (2016). Test-retest reliability of effective connectivity in the face perception network. *Human Brain Mapping*, *37*, 730–744.

Frässle, S., Stephan, K. E., Friston, K. J., Steup, M., Krach, S., Paulus, F. M., & Jansen, A. (2015). Test-retest reliability of dynamic causal modeling for fMRI. *NeuroImage*, *117*, 56–66.

Friston, K., Brown, H. R., Siemerkus, J., & Stephan, K. E. (2016). The dysconnection hypothesis. *Schizophrenia Research*, *176*, 83–94.

Friston, K., Harrison, L., & Penny, W. (2003). Dynamic causal modelling. *NeuroImage*, *19*, 1273–1302.

Friston, K., Li, B., Daunizeau, J., & Stephan, K. (2011). Network discovery with DCM. *NeuroImage*, *56*, 1202–1221.

Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., & Penny, W. (2007). Variational free energy and the Laplace approximation. *NeuroImage*, *34*, 220–234.

Friston, K., Moran, R., & Seth, A. (2013). Analysing connectivity with granger causality and dynamic causal modelling. *Current Opinion in Neurobiology*, *23*, 172–178.

Friston, K. J. (1998). The disconnection hypothesis. *Schizophrenia Research*, *30*, 115–125.

Friston, K. J. (2011). Functional and effective connectivity: A review. *Brain Connectivity*, *1*, 13–36.

Friston, K. J., & Frith, C. D. (1995). Schizophrenia: A disconnection syndrome? *Clinical Neuroscience*, *3*, 89–97.

Friston, K. J., Kahan, J., Biswal, B., & Razi, A. (2014). A DCM for resting state fMRI. *NeuroImage*, *94*, 396–407.

Friston, K. J., Litvak, V., Oswal, A., Razi, A., Stephan, K. E., van Wijk, B. C., … Zeidman, P. (2016). Bayesian model reduction and empirical Bayes for group (DCM) studies. *NeuroImage*, *128*, 413–431.

Friston, K. J., Mechelli, A., Turner, R., & Price, C. J. (2000). Nonlinear responses in fMRI: The balloon model, Volterra kernels, and other hemodynamics. *NeuroImage*, *12*, 466–477.

Friston, K. J., Preller, K. H., Mathys, C., Cagnan, H., Heinzle, J., Razi, A., & Zeidman, P. (2017). Dynamic causal modelling revisited. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2017.02.045

Friston, K. J., Stephan, K. E., Montague, R., & Dolan, R. J. (2014). Computational psychiatry: The brain as a phantastic organ. *Lancet Psychiatry*, *1*, 148–158.

Gao, W. J., & Goldman-Rakic, P. S. (2003). Selective modulation of excitatory and inhibitory microcircuits by dopamine. *Proceedings of the National Academy of Sciences of the United States of America*, *100*, 2836–2841.

Garrido, M. I., Kilner, J. M., Kiebel, S. J., Stephan, K. E., & Friston, K. J. (2007). Dynamic causal modelling of evoked potentials: A reproducibility study. *NeuroImage*, *36*, 571–580.

Gilbert, J. R., Symmonds, M., Hanna, M. G., Dolan, R. J., Friston, K. J., & Moran, R. J. (2016). Profiling neuronal ion channelopathies with non-invasive brain imaging and dynamic causal models: Case studies of single gene mutations. *NeuroImage*, *124*, 43–53.

Glascher, J. P., & O'Doherty, J. P. (2010). Model-based approaches to neuroimaging: Combining reinforcement learning theory with fMRI data. *WIREs Cognitive Science*, *1*, 501–510.

Goldman-Rakic, P. S. (1996). Regional and cellular fractionation of working memory. *Proceedings of the National Academy of Sciences of the United States of America*, *93*, 13473–13480.

Greicius, M. D., Flores, B. H., Menon, V., Glover, G. H., Solvason, H. B., Kenna, H., … Schatzberg, A. F. (2007). Resting-state functional connectivity in major depression: Abnormally increased contributions from subgenual cingulate cortex and thalamus. *Biological Psychiatry*, *62*, 429–437.

Grèzes, J., Wicker, B., Berthoz, S., & de Gelder, B. (2009). A failure to grasp the affective meaning of actions in autism spectrum disorder subjects. *Neuropsychologia*, *47*, 1816–1825.

Grossenbacher, P. G., & Lovelace, C. T. (2001). Mechanisms of synesthesia: Cognitive and physiological constraints. *Trends in Cognitive Sciences*, *5*, 36–41.

Hagmann, P., Cammoun, L., Gigandet, X., Meuli, R., Honey, C. J., Wedeen, V. J., & Sporns, O. (2008). Mapping the structural core of human cerebral cortex. *PLoS Biology*, *6*, e159.

Harle, K. M., Stewart, J. L., Zhang, S., Tapert, S. F., AJ, Y., & Paulus, M. P. (2015). Bayesian neural adjustment of inhibitory control predicts emergence of problem stimulant use. *Brain*, *138*, 3413–3426.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. New York, NY: Springer.

Havlicek, M., Roebroeck, A., Friston, K., Gardumi, A., Ivanov, D., & Uludag, K. (2015). Physiologically informed dynamic causal modeling of fMRI data. *NeuroImage*, *122*, 355–372.

Hochel, M., & Milan, E. G. (2008). Synaesthesia: The existing state of affairs. *Cognitive Neuropsychology*, *25*, 93–117.

Hubbard, E. M., Arman, A. C., Ramachandran, V. S., & Boynton, G. M. (2005). Individual differences among grapheme-color synesthetes: Brain-behavior correlations. *Neuron*, *45*, 975–985.

Huys, Q. J., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, *19*, 404–413.

Hyett, M. P., Breakspear, M. J., Friston, K. J., Guo, C. C., & Parker, G. B. (2015). Disrupted effective connectivity of cortical systems supporting attention and interoception in melancholia. *JAMA Psychiatry*, *72*, 350–358.

Iglesias, S., Tomiello, S., Schneebeli, M., & Stephan, K. E. (2016). Models of neuromodulation for computational psychiatry. *WIREs Cognitive Science*, *8*, e1420.

Insel, T. R. (2008). Assessing the economic costs of serious mental illness. *The American Journal of Psychiatry*, *165*, 663–665.

Jansen, B. H., & Rit, V. G. (1995). Electroencephalogram and visual evoked potential generation in a mathematical model of coupled cortical columns. *Biological Cybernetics*, *73*, 357–366.

Jirsa, V. K., Proix, T., Perdikis, D., Woodman, M. M., Wang, H., Gonzalez-Martinez, J., … Bartolomei, F. (2016). The virtual epileptic patient: Individualized whole-brain models of epilepsy spread. *NeuroImage*, *145*, 377–388.

Kahan, J., & Foltynie, T. (2013). Understanding DCM: Ten simple rules for the clinician. *NeuroImage*, *83*, 542–549.

Kapur, S., Phillips, A. G., & Insel, T. R. (2012). Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Molecular Psychiatry*, *17*, 1174–1179.

Kass, R., & Steffey, D. (1989). Aproximate Bayesian inference in conditionally indepedent hierarchical models (parametric empirical Bayes models). *Journal of the American Statistical Association*, *84*, 717–726.

Kennedy, D. P., Redcay, E., & Courchesne, E. (2006). Failing to deactivate: Resting functional abnormalities in autism. *Proceedings of the National Academy of Sciences of the United States of America*, *103*, 8275–8280.

Kiebel, S. J., David, O., & Friston, K. J. (2006). Dynamic causal modelling of evoked responses in EEG/MEG with lead field parameterization. *NeuroImage*, *30*, 1273–1284.

Kirkwood, J. G. (1935). Statistical mechanics of fluid mixtures. *Journal of Chemical Physics*, *3*, 300–313.

Klassen, T., Davis, C., Goldman, A., Burgess, D., Chen, T., Wheeler, D., … Noebels, J. (2011). Exome sequencing of ion channel genes reveals complex profiles confounding personal risk assessment in epilepsy. *Cell*, *145*, 1036–1048.

Krystal, J. H., & State, M. W. (2014). Psychiatric disorders: Diagnosis to therapy. *Cell*, *157*, 201–214.

Larsen, M., Haugstad, T. S., Berg-Johnsen, J., & Langmoen, I. A. (1998). Effect of isoflurane on release and uptake of gamma-aminobutyric acid from rat cortical synaptosomes. *British Journal of Anaesthesia*, *80*, 634–638.

Lartillot, N., & Philippe, H. (2006). Computing Bayes factors using thermodynamic integration. *Systematic Biology*, *55*, 195–207.

Lau, C. G., & Zukin, R. S. (2007). NMDA receptor trafficking in synaptic plasticity and neuropsychiatric disorders. *Nature Reviews. Neuroscience*, *8*, 413–426.

Laumann TO, Gordon, E. M., Adeyemo, B., Snyder, A. Z., Joo, S. J., Chen, M. Y., … Petersen, S. E. (2015). Functional system and areal organization of a highly sampled individual human brain. *Neuron*, *87*, 657–670.

Lefebvre, S., Demeulemeester, M., Leroy, A., Delmaire, C., Lopes, R., Pins, D., … Jardri, R. (2016). Network dynamics during the different stages of hallucinations in schizophrenia. *Human Brain Mapping*, *37*, 2571–2586.

Li, B., Cui, L. B., Xi, Y. B., Friston, K. J., Guo, F., Wang, H. N., … Lu, H. (2017). Abnormal effective connectivity in the brain is involved in auditory verbal hallucinations in schizophrenia. *Neuroscience Bulletin*, *33*, 281–291.

Li, B., Daunizeau, J., Stephan, K. E., Penny, W., Hu, D., & Friston, K. (2011). Generalised filtering and stochastic DCM for fMRI. *NeuroImage*, *58*, 442–457.

Litvak, V., Garrido, M., Zeidman, P., & Friston, K. (2015). Empirical Bayes for group (DCM) studies: A reproducibility study. *Frontiers in Human Neuroscience*, *9*, 670.

Maia, T., & Frank, M. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nature Neuroscience*, *14*, 154–162.

Maldjian, J. A., Laurienti, P. J., Driskill, L., & Burdette, J. H. (2002). Multiple reproducibility indices for evaluation of cognitive functional MR imaging paradigms. *American Journal of Neuroradiology*, *23*, 1030–1037.

Marreiros, A. C., Cagnan, H., Moran, R. J., Friston, K. J., & Brown, P. (2013). Basal ganglia-cortical interactions in Parkinsonian patients. *NeuroImage*, *66*, 301–310.

Mayberg, H. S. (1997). Limbic-cortical dysregulation: A proposed model of depression. *The Journal of Neuropsychiatry and Clinical Neurosciences*, *9*, 471–481.

Menon, V. (2011). Large-scale brain networks and psychopathology: A unifying triple network model. *Trends in Cognitive Sciences*, *15*, 483–506.

Montague, P., Dolan, R., Friston, K., & Dayan, P. (2012). Computational psychiatry. *Trends in Cognitive Sciences*, *16*, 72–80.

Moran, R., Pinotsis, D. A., & Friston, K. (2013). Neural masses and fields in dynamic causal modeling. *Frontiers in Computational Neuroscience*, *7*, 57.

Moran, R. J., Campo, P., Symmonds, M., Stephan, K. E., Dolan, R. J., & Friston, K. J. (2013). Free energy, precision and learning: The role of cholinergic neuromodulation. *The Journal of Neuroscience*, *33*, 8227–8236.

Moran, R. J., Jones, M. W., Blockeel, A. J., Adams, R. A., Stephan, K. E., & Friston, K. J. (2015). Losing control under ketamine: Suppressed cortico-hippocampal drive following acute ketamine in rats. *Neuropsychopharmacology*, *40*, 268–277.

Moran, R. J., Jung, F., Kumagai, T., Endepols, H., Graf, R., Dolan, R. J., … Tittgemeyer, M. (2011). Dynamic causal models and physiological inference: A validation study using isoflurane anaesthesia in rodents. *PLoS ONE*, *6*, e22790.

Moran, R. J., Stephan, K. E., Dolan, R. J., & Friston, K. J. (2011). Consistent spectral predictors for dynamic causal models of steady-state responses. *NeuroImage*, *55*, 1694–1708.

Moran, R. J., Symmonds, M., Stephan, K. E., Friston, K. J., & Dolan, R. J. (2011). An in vivo assay of synaptic function mediating human cognition. *Current Biology*, *21*, 1320–1325.

Mosher, J. C., Leahy, R. M., & Lewis, P. S. (1999). EEG and MEG: Forward solutions for inverse methods. *IEEE Transactions on Biomedical Engineering*, *46*, 245–259.

Muller, R. A. (2007). The study of autism as a distributed disorder. *Mental Retardation and Developmental Disabilities Research Reviews*, *13*, 85–95.

Noble, S., Scheinost, D., Finn, E. S., Shen, X., Papademetris, X., McEwen, S. C., … Constable, R. T. (2017). Multisite reliability of MR-based functional connectivity. *NeuroImage*, *146*, 959–970.

Nord, C. L., Gray, A., Charpentier, C. J., Robinson, O. J., & Roiser, J. P. (2017). Unreliability of putative fMRI biomarkers during emotional face processing. *NeuroImage*, *156*, 119–127.

Ogawa, S., Lee, T. M., Kay, A. R., & Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences of the United States of America*, *87*, 9868–9872.

Owen, M. J. (2014). New approaches to psychiatric diagnostic classification. *Neuron*, *84*, 564–571.

Papadopoulou, M., Cooray, G., Rosch, R., Moran, R., Marinazzo, D., & Friston, K. (2017). Dynamic causal modelling of seizure activity in a rat model. *NeuroImage*, *146*, 518–532.

Paulus, M. P., Huys, Q. J., & Maia, T. V. (2016). A roadmap for the development of applied computational psychiatry. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *1*, 386–392.

Penninx, B. W., Beekman, A. T., Smit, J. H., Zitman, F. G., Nolen, W. A., Spinhoven, P., … NESDA Research Consortium. (2008). The Netherlands Study of Depression and Anxiety (NESDA): Rationale, objectives and methods. *International Journal of Methods in Psychiatric Research*, *17*, 121–140.

Penny, W. (2012). Comparing dynamic causal models using AIC, BIC and free energy. *NeuroImage*, *59*, 319–330.

Penny, W., & Sengupta, B. (2016). Annealed importance sampling for neural mass models. *PLoS Computational Biology*, *12*, e1004797.

Petzschner, F. H., Weber, L. A. E., Gard, T., & Stephan, K. E. (2017). Computational psychosomatics and computational psychiatry: Toward a joint framework for differential diagnosis. *Biological Psychiatry*, *82*, 421–430.

Pinotsis, D. A., Moran, R. J., & Friston, K. J. (2012). Dynamic causal modeling with neural fields. *NeuroImage*, *59*, 1261–1274.

Pinotsis, D. A., Perry, G., Litvak, V., Singh, K. D., & Friston, K. J. (2016). Intersubject variability and induced gamma in the visual cortex: DCM with empirical Bayes and neural fields. *Human Brain Mapping*, *37*, 4597–4614.

Pinotsis, D. A., Schwarzkopf, D. S., Litvak, V., Rees, G., Barnes, G., & Friston, K. J. (2013). Dynamic causal modelling of lateral interactions in the visual cortex. *NeuroImage*, *66*, 563–576.

Radulescu, E., Minati, L., Ganeshan, B., Harrison, N. A., Gray, M. A., Beacher, F. D., … Critchley, H. D. (2013). Abnormalities in fronto-striatal connectivity within language networks relate to differences in grey-matter heterogeneity in Asperger syndrome. *Neuroimage Clin*, *2*, 716–726.

Raemaekers, M., Vink, M., Zandbelt, B., van Wezel, R. J., Kahn, R. S., & Ramsey, N. F. (2007). Test-retest reliability of fMRI activation during prosaccades and antisaccades. *NeuroImage*, *36*, 532–542.

Ramachandran, V. S., & Hubbard, E. M. (2001). Psychophysical investigations into the neural basis of synaesthesia. *Proceedings of the Biological Sciences*, *268*, 979–983.

Raman, S., Deserno, L., Schlagenhauf, F., & Stephan, K. E. A. (2016). Hierarchical model for integrating unsupervised generative embedding and empirical Bayes. *Journal of Neuroscience Methods*, *269*, 6–20.

Razi, A., Seghier, M. L., Zhou, Y., McColgan, P., Zeidman, P., Park, H.-J., … Friston, K. J. (2017). Large-scale DCMs for resting state fMRI. *Network Neuroscience*, *1*, 222–241.

Robbins, T. W. (2000). Chemical neuromodulation of frontal-executive functions in humans and other animals. *Experimental Brain Research*, *133*, 130–138.

Roebroeck, A., Formisano, E., & Goebel, R. (2011). The identification of interacting networks in the brain using fMRI: Model selection, causality and deconvolution. *NeuroImage*, *58*, 296–302.

Rosa, M., Friston, K., & Penny, W. (2012). Post-hoc selection of dynamic causal models. *Journal of Neuroscience Methods*, *208*, 66–78.

Rowe, J., Hughes, L., Barker, R., & Owen, A. (2010). Dynamic causal modelling of effective connectivity from fMRI: Are results reproducible and sensitive to Parkinson's disease and its treatment? *NeuroImage*, *52*, 1015–1026.

Roy, M., Shohamy, D., Daw, N., Jepma, M., Wimmer, G. E., & Wager, T. D. (2014). Representation of aversive prediction errors in the human periaqueductal gray. *Nature Neuroscience*, *17*, 1607–1612.

Rush, A., Trivedi, M., Wisniewski, S., Nierenberg, A., Stewart, J., Warden, D., … Fava, M. (2006). Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: A STAR*D report. *American Journal of Psychiatry*, *163*, 1905–1917.

Schlösser, R. G., Wagner, G., Koch, K., Dahnke, R., Reichenbach, J. R., & Sauer, H. (2008). Fronto-cingulate effective connectivity in major depression: A study with fMRI and dynamic causal modeling. *NeuroImage*, *43*, 645–655.

Schmidt, A., Diaconescu, A. O., Kometer, M., Friston, K. J., Stephan, K. E., & Vollenweider, F. X. (2013). Modeling ketamine effects on synaptic plasticity during the mismatch negativity. *Cerebral Cortex*, *23*, 2394–2406.

Schofield, T. M., Penny, W. D., Stephan, K. E., Crinion, J. T., Thompson, A. J., Price, C. J., & Leff, A. P. (2012). Changes in auditory feedback connections determine the severity of speech processing deficits after stroke. *The Journal of Neuroscience*, *32*, 4260–4270.

Schunck, T., Erb, G., Mathis, A., Jacob, N., Gilles, C., Namer, I. J., … Luthringer, R. (2008). Test-retest reliability of a functional MRI anticipatory anxiety paradigm in healthy volunteers. *Journal of Magnetic Resonance Imaging*, *27*, 459–468.

Schuyler, B., Ollinger, J., Oakes, T., Johnstone, T., & Davidson, R. (2010). Dynamic causal modeling applied to fMRI data shows high reliability. *NeuroImage*, *49*, 603–611.

Seghier, M. L., & Friston, K. J. (2013). Network discovery with large DCMs. *NeuroImage*, *68*, 181–191.

Sengupta, B., Friston, K. J., & Penny, W. D. (2015). Gradient-free MCMC methods for dynamic causal modelling. *NeuroImage*, *112*, 375–381.

Sengupta, B., Friston, K. J., & Penny, W. D. (2016). Gradient-based MCMC samplers for dynamic causal modelling. *NeuroImage*, *125*, 1107–1118.

Shehzad, Z., Kelly, A. M., Reiss, P. T., Gee, D. G., Gotimer, K., Uddin, L. Q., … Milham, M. P. (2009). The resting brain: Unconstrained yet reliable. *Cerebral Cortex*, *19*, 2209–2229.

Sperling, J. M., Prvulovic, D., Linden, D. E., Singer, W., & Stirn, A. (2006). Neuronal correlates of colour-graphemic synaesthesia: A fMRI study. *Cortex*, *42*, 295–303.

Stephan, K., Baldeweg, T., & Friston, K. (2006). Synaptic plasticity and dysconnection in schizophrenia. *Biological Psychiatry*, *59*, 929–939.

Stephan, K., Friston, K., & Frith, C. (2009). Dysconnection in schizophrenia: From abnormal synaptic plasticity to failures of self-monitoring. *Schizophrenia Bulletin*, *35*, 509–527.

Stephan, K., Kasper, L., Harrison, L., Daunizeau, J., den Ouden, H., Breakspear, M., & Friston, K. (2008). Nonlinear dynamic causal models for fMRI. *NeuroImage*, *42*, 649–662.

Stephan, K., & Mathys, C. (2014). Computational approaches to psychiatry. *Current Opinion in Neurobiology*, *25*, 85–92.

Stephan, K., Penny, W., Daunizeau, J., Moran, R., & Friston, K. (2009). Bayesian model selection for group studies. *NeuroImage*, *46*, 1004–1017.

Stephan, K., Penny, W., Moran, R., den Ouden, H., Daunizeau, J., & Friston, K. (2010). Ten simple rules for dynamic causal modeling. *NeuroImage*, *49*, 3099–3109.

Stephan, K. E., Bach, D. R., Fletcher, P. C., Flint, J., Frank, M. J., Friston, K. J., … Breakspear, M. (2016). Charting the landscape of priority problems in psychiatry, part 1: Classification and diagnosis. *Lancet Psychiatry*, *3*, 77–83.

Stephan, K. E., Binder, E. B., Breakspear, M., Dayan, P., Johnstone, E. C., Meyer-Lindenberg, A., … Friston, K. J. (2016). Charting the landscape of priority problems in psychiatry, part 2: Pathogenesis and aetiology. *Lancet Psychiatry*, *3*, 84–90.

Stephan, K. E., Iglesias, S., Heinzle, J., & Diaconescu, A. O. (2015). Translational perspectives for computational neuroimaging. *Neuron*, *87*, 716–732.

Stephan, K. E., Manjaly, Z. M., Mathys, C. D., Weber, L. A., Paliwal, S., Gard, T., … Petzschner, F. H. (2016). Allostatic self-efficacy: A metacognitive theory of dyshomeostasis-induced fatigue and depression. *Frontiers in Human Neuroscience*, *10*, 550.

Stephan, K. E., Schlagenhauf, F., Huys, Q. J., Raman, S., Aponte, E. A., Brodersen, K. H., … Heinz, A. (2017). Computational neuroimaging strategies for single patient predictions. *NeuroImage*, *145*, 180–199.

Stephan, K. E., Siemerkus, J., Bishop, M., & Haker, H. (2017). Hat computational psychiatry Relevanz für die klinische praxis der Psychiatrie? *Zeitschrift für Psychiatrie, Psychologie und Psychotherapie*, *65*, 9–19.

Stephan, K. E., Weiskopf, N., Drysdale, P. M., Robinson, P. A., & Friston, K. J. (2007). Comparing hemodynamic models with DCM. *NeuroImage*, *38*, 387–401.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., … Collins, R. (2015). UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine*, *12*, e1001779.

Vai, B., Bulgarelli, C., Godlewska, B. R., Cowen, P. J., Benedetti, F., & Harmer, C. J. (2016). Fronto-limbic effective connectivity as possible predictor of antidepressant response to SSRI administration. *European Neuropsychopharmacology*, *26*, 2000–2010.

Valdes-Sosa, P. A., Roebroeck, A., Daunizeau, J., & Friston, K. (2011). Effective connectivity: Influence, causality and biophysical modeling. *NeuroImage*, *58*, 339–361.

van Leeuwen, T. M., den Ouden, H. E., & Hagoort, P. (2011). Effective connectivity determines the nature of subjective experience in grapheme-color synesthesia. *The Journal of Neuroscience*, *31*, 9879–9884.

Wang, L., Hermens, D. F., Hickie, I. B., & Lagopoulos, J. (2012). A systematic review of resting-state functional-MRI studies in major depression. *Journal of Affective Disorders*, *142*, 6–12.

Wang, X. J., & Krystal, J. H. (2014). Computational psychiatry. *Neuron*, *84*, 638–654.

Weiss, P. H., & Fink, G. R. (2009). Grapheme-colour synaesthetes show increased grey matter volumes of parietal and fusiform cortex. *Brain*, *132*, 65–70.

Wipf, D., & Nagarajan, S. (2009). A unified Bayesian framework for MEG/EEG source imaging. *NeuroImage*, *44*, 947–966.

Woolrich, M. W., & Stephan, K. E. (2013). Biophysical network models and the human connectome. *NeuroImage*, *80*, 330–338.

World Health Organization. (1990). *International classification of diseases*. Geneva, Switzerland: World Health Organization Press.

Zhang, J., Rittman, T., Nombela, C., Fois, A., Coyle-Gilchrist, I., Barker, R. A., … Rowe, J. B. (2016). Different decision deficits impair response inhibition in progressive supranuclear palsy and Parkinson's disease. *Brain*, *139*, 161–173.

Zuo, X. N., Kelly, C., Adelstein, J. S., Klein, D. F., Castellanos, F. X., & Milham, M. P. (2010). Reliable intrinsic connectivity networks: Test-retest evaluation using ICA and dual regression approach. *NeuroImage*, *49*, 2163–2177.

How to cite this article: Frässle S, Yao Y, Schöbi D, Aponte EA, Heinzle J, Stephan KE. Generative models for clinical applications in computational psychiatry. *WIREs Cogn Sci*. 2018;9:e1460. https://doi.org/10.1002/wcs.1460