

RESEARCH ARTICLE

WILEY

Spatiotemporally resolved multivariate pattern analysis for M/EEG

Cameron Higgins^{1,2}  | Diego Vidaurre^{2,3}  | Nils Kolling¹  | Yunzhe Liu^{4,5,6}  |
Tim Behrens^{1,7}  | Mark Woolrich^{1,2} 

¹Wellcome Centre for Integrative Neuroimaging, University of Oxford, Oxford, UK

²Department of Psychiatry, University of Oxford, Oxford, UK

³Center of Functionally Integrative Neuroscience, Department of Clinical Medicine, Aarhus University, Aarhus, Denmark

⁴State Key Laboratory of Cognitive Neuroscience and Learning, IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing, China

⁵Chinese Institute for Brain Research, Beijing, China

⁶Max Planck University College London Centre for Computational Psychiatry and Ageing Research, University College London, London, UK

⁷Wellcome Trust Centre for Neuroimaging, University College London, London, UK

Correspondence

Cameron Higgins, Wellcome Centre for Integrative Neuroimaging, University of Oxford, Oxford, UK.
Email: cameron.higgins@ohba.ox.ac.uk

Funding information

Biotechnology and Biological Sciences Research Council, Grant/Award Number: BB/R010803/1; H2020 European Research Council, Grant/Award Number: ERC-StG-2019-850404; James S. McDonnell Foundation, Grant/Award Number: JSMF220020372; Medical Research Council, Grant/Award Numbers: RG89702, RG94383; NIHR Oxford Health Biomedical Research Centre; Novo Nordisk, Grant/Award Number: NNF19OC-0054895; Wellcome Trust, Grant/Award Numbers: 104765/Z/14/Z, 106183/Z/14/Z, 203139/Z/16/Z, 214314/Z/18/Z, 215573/Z/19/Z, 219525/Z/19/Z; EU-Project euSNN, Grant/Award Number: MSCA-ITN H2020-860563

Abstract

An emerging goal in neuroscience is tracking what information is represented in brain activity over time as a participant completes some task. While electroencephalography (EEG) and magnetoencephalography (MEG) offer millisecond temporal resolution of how activity patterns emerge and evolve, standard decoding methods present significant barriers to interpretability as they obscure the underlying spatial and temporal activity patterns. We instead propose the use of a generative encoding model framework that simultaneously infers the multivariate spatial patterns of activity and the variable timing at which these patterns emerge on individual trials. An encoding model inversion maps from these parameters to the equivalent decoding model, allowing predictions to be made about unseen test data in the same way as in standard decoding methodology. These SpatioTemporally Resolved MVPA (STRM) models can be flexibly applied to a wide variety of experimental paradigms, including classification and regression tasks. We show that these models provide insightful maps of the activity driving predictive accuracy metrics; demonstrate behaviourally meaningful variation in the timing of pattern emergence on individual trials; and achieve predictive accuracies that are either equivalent or surpass those achieved by more widely used methods. This provides a new avenue for investigating the brain's representational dynamics and could ultimately support more flexible experimental designs in the future.

KEYWORDS

decoding, EEG, encoding, MEG, single trial task dynamics

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Authors. *Human Brain Mapping* published by Wiley Periodicals LLC.

1 | INTRODUCTION

The use of decoding models for the analysis of magnetoencephalography (MEG) and electroencephalography (EEG) data has substantially grown in recent years (Grootswagers, Wardle, & Carlson, 2017). Increasingly researchers are turning to these methods—collectively referred to as Multivariate pattern analysis (MVPA)—for their increased sensitivity to distributed patterns of variation that can be attributed to a stimulus (Haynes & Rees, 2006). While projecting high dimensional neural data down to a single metric of classification accuracy affords researchers greater statistical sensitivity, it generally comes at a cost to interpretability: the precise nature of the link between significant changes in decoding accuracy and the underlying neuroscience driving those changes can often be indirect or opaque (Haufe et al., 2014; Kriegeskorte & Douglas, 2019; Naselaris & Kay, 2015; Valentin, Harkotte, & Popov, 2020). Furthermore, commonly used methods that align all trials in time and proceed in a timepoint-by-timepoint fashion offer no sensitivity to patterns that are not perfectly synchronised in time across different trials (Borst & Anderson, 2015; Vidaurre, Myers, Stokes, Nobre, & Woolrich, 2019).

The main focus in the use of MVPA for M/EEG is to leverage these modalities' high temporal resolution to investigate the brain's representational dynamics over time (King & Dehaene, 2014). A convention has emerged for analysing trial data whereby successive spatial filters—each a vector with one linear coefficient applied to each sensor—are trained on data at different timestamps from the

presentation of some stimulus (Carlson, Hogendoorn, Kanai, Mesik, & Turret, 2011; Carlson, Tovar, Alink, & Kriegeskorte, 2013; Cichy, Pantazis, & Oliva, 2014; Grootswagers et al., 2017; Haxby, Connolly, & Guntupalli, 2014; van de Nieuwenhuijzen et al., 2013). While this convention has certainly been useful, it is not without its flaws, as illustrated in Figure 1. Firstly, spatial filters are blind to any temporal structure in the data. These filters cannot detect patterns of temporal autocorrelation that are subtle but consistent over multiple timepoints. Secondly, spatial filters cannot be interpreted as reflecting the brain activity associated with stimuli; this has been widely characterised in the neuroscience literature (Haufe et al., 2014; Kriegeskorte & Douglas, 2019; Valentin et al., 2020), but is somewhat counter intuitive. As shown in Figure 1, given an evoked response from two different stimuli in the presence of correlated noise, the coefficients of a linear spatial filter need not visually resemble or reveal these underlying patterns. Post hoc methods have been proposed to map backwards from a spatial filter to recover a forward model of the data (Haufe et al., 2014), however these come with a number of caveats and are not guaranteed to provide the most accurate forward model parameter estimates. We instead propose to turn this approach on its head, and ask: why not first learn a generative model of the data, and then use that model to make predictions of unseen stimuli?

In neuroscience terminology, this amounts to fitting an *encoding* model, then inverting that encoding model to make predictions through an equivalent *decoding* model (Friston et al., 2008; Haxby

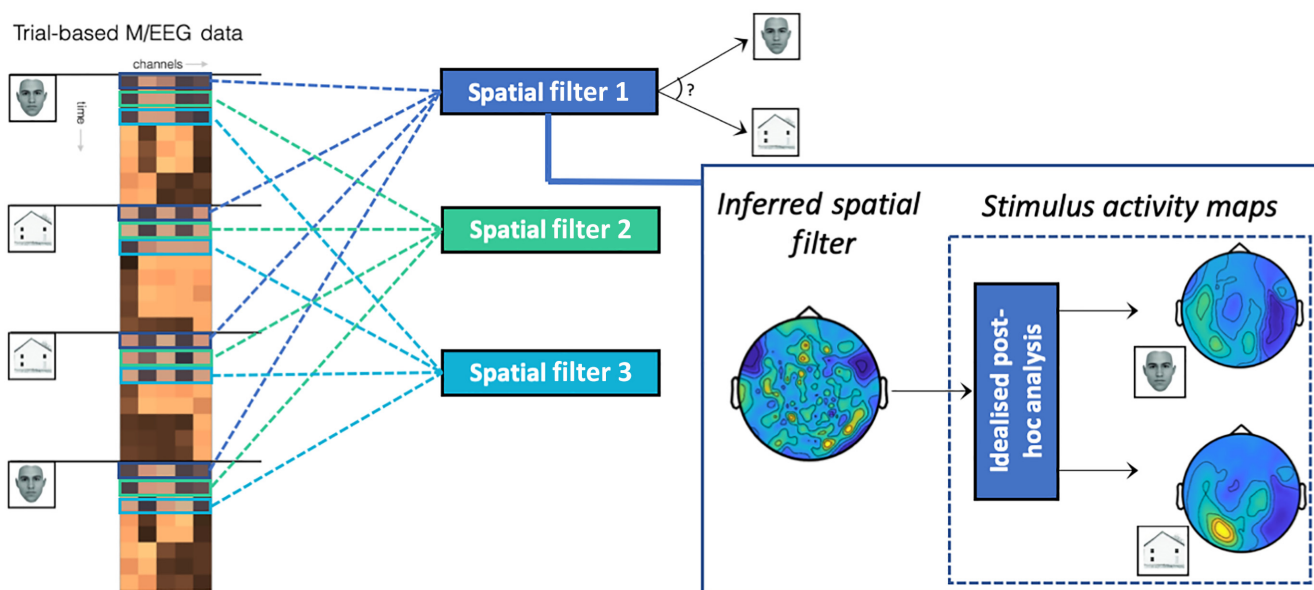


FIGURE 1 Conventional approaches to decoding in M/EEG are mass univariate through time and difficult to interpret in space without further post hoc analysis. In a typical M/EEG experiment decoding different types of stimuli—in this example, images of faces and houses adapted from Negrini, Brkic, Pizzamiglio, Premoli, and Rivolta 2017; (the exemplary maps are based on real MEG data from the dataset presented below)—the conventional approach extracts all data at one timestep from the stimulus onset time and trains a spatial filter on that data to distinguish the conditions. This process is repeated independently at all timesteps. This approach ignores the time series nature of this data, and the spatial filters are neither able to detect patterns that are consistent over multiple timesteps, nor patterns that are consistent over trials but not perfectly aligned in time. Furthermore, the spatial filter coefficients are not directly interpretable (Haufe et al., 2014; Kriegeskorte & Douglas, 2019; Valentin et al., 2020) without a further post hoc analysis step (inset) such as the seminal method proposed by Haufe et al. (2014)

et al., 2014; Naselaris, Kay, Nishimoto, & Gallant, 2011). This approach has been successful in fMRI (Casey, Thompson, Kang, Raizada, & Wheatley, 2011; Friston et al., 2008; Kay, Naselaris, Prenger, & Gallant, 2008; Mitchell et al., 2008; Naselaris, Olman, Stansbury, Ugurbil, & Gallant, 2015; Naselaris, Prenger, Kay, Oliver, & Gallant, 2009; Nishimoto et al., 2011; Schoenmakers, Barth, Heskes, & van Gerven, 2013) but has only seen quite limited adoption for M/EEG (di Liberto, O'Sullivan, & Lalor, 2015; Kupers, Benson, & Winawer, 2020). In a similar vein, we therefore propose a linear generative model of stimulus evoked activity based upon the popular General Linear Model (GLM) framework for M/EEG data (Trujillo-Barreto, Aubert-Vázquez, & Penny, 2008). While the GLM in neuroimaging is traditionally associated with mass univariate analysis, it can be extended to be multivariate across space, allowing multivariate predictions to be made on unseen test data through a simple application of Bayes rule (Friston et al., 2008; Haxby et al., 2014). Notably, the inversion of an encoding model in M/EEG can be framed as a solution to the inverse problem, where it has been successfully applied to obtain estimates of source-localised activity (Trujillo-Barreto et al., 2008)—we apply the same general approach to instead make predictions about unseen test-data in a decoding-style manner. The widespread utility of the GLM approach in neuroimaging suggests it may have a broad applicability across a range of different experimental paradigms; in support of this claim, we demonstrate its use across two very different datasets.

Further, we show how this encoding framework can be extended to take advantage of approaches that adapt to timing differences across different trials (Anderson, Fincham, Schneider, & Yang, 2012; Borst & Anderson, 2015; Obermaier, Guger, Neuper, & Pfurtscheller, 2001; Vidaurre et al., 2019; Williams, Daly, & Nasuto, 2018), thereby more fully utilising the high temporal resolution that is the main benefit of M/EEG as a recording paradigm. Such temporal variability over trials can elucidate key aspects of cognitive processing (Anderson et al., 2012; Anderson, Betts, Ferris, & Fincham, 2010; Borst & Anderson, 2015; Borst, Ghuman, & Anderson, 2016), however has generally only been investigated using methods tailored to specific task paradigms (Anderson et al., 2012; Anderson & Fincham, 2014a, 2014b). In contrast, we show that the same hidden Markov modelling (HMM) approach can be formulated using the popular GLM framework, potentially allowing such questions to be pursued across a broader range of neuroimaging experiments (Vidaurre et al., 2019). With this approach, we are able to characterise the emergence of distinct representational states on individual trials and better model patterns that endure over multiple timepoints but may not be perfectly aligned in time. Previous requirements for patterns to be perfectly aligned over multiple trials limited experimental designs to paradigms that ensure maximal inter-trial reproducibility (Light et al., 2010), or alternatively designed to elucidate large changes in activity that could be more easily identified (Anderson & Fincham, 2014a; Borst & Anderson, 2015). Our more flexible modelling approach overcomes this limitation, potentially allowing more ambitious experimental designs in which subtle patterns of inter-trial variability are anticipated and can be quantified. We

believe this opens a new door to investigating representational dynamics, by not only characterising when certain aspects of a stimulus emerge in data, but also asking how these representational dynamics are modulated across different trials.

2 | METHODS

Our approach extends the traditional general linear model (GLM) framework by (a) incorporating a latent Markov variable to explain time-varying dynamics within trials, and (b) modelling the multivariate spatial distribution simultaneously over all recorded channels. In so doing, we maintain the benefits of a generative model from the GLM approach, namely interpretable maps of linear activation strengths over each channel; alongside the benefits of multivariate methods, namely an increased sensitivity to distributed patterns of variation over multiple timesteps and amenability to hierarchical modelling (i.e., latent variable modelling of differences in dynamics over trials). These models are fit independently to each subject, resulting in distinct model parameters for each subject upon which standard group statistical tests can be run.

2.1 | Standard general linear model setup

We begin with the standard formulation of a GLM for evoked response analysis. For a specific subject, with a total of N samples of M/EEG data recorded across P sensors in a paradigm with Q regressors, we denote by Y the neural data of dimension $[N \times P]$, with an associated design matrix of regressors X of dimension $[N \times Q]$. This standard GLM setup is not explicit about how the N samples relate to time or trials. The convention in M/EEG analysis is to fit independent GLMs over successive timepoints within a trial in the same manner as in Figure 1, such that N would equal the number of trials and the effects of each stimulus are modelled independently at each timepoint within the trial. Note that our use of “neural data” here is generic; in this article we use raw sensor data, but this could equivalently be commonly used features such as bandlimited power. The GLM is formulated as follows:

$$P(Y|X, W, \Sigma) = N(XW, \Sigma) \quad (1)$$

The matrix W , of dimension $[Q \times P]$, is directly interpretable as each row is an estimate of the corresponding stimulus' activation pattern (i.e., its effects on each sensor's measured signal) while the confidence in that estimate depends on the entries of the $[P \times P]$ residual covariance matrix Σ .

The common use of the GLM in mass univariate analysis may lead some readers to mistakenly believe it is a univariate model. This is because, putting aside spatial priors that could be used on W , this model stores any multivariate information in the residual covariance matrix Σ ; and this matrix is simply not used in mass univariate

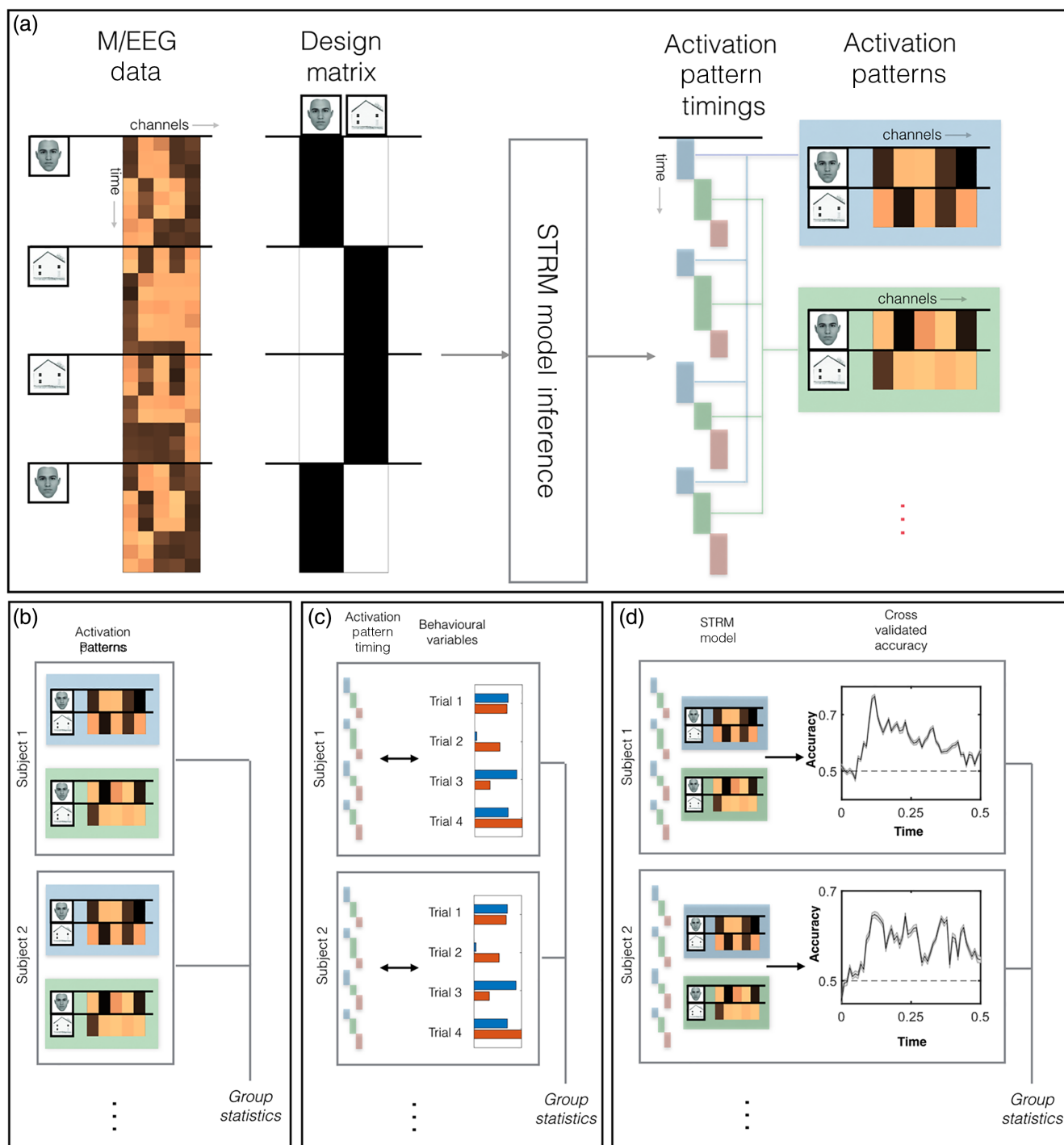


FIGURE 2 SpatioTemporally Resolved MVPA (STRM) for M/EEG. (a) The STRM Model receives as inputs the M/EEG data and corresponding design matrix, and outputs a set of activation patterns and their corresponding timing on each trial. Our analysis pipeline fits this model to data at the subject level, then extracts three different summary statistics (Panel b–d) to analyse at the group level. (b) The consistency of state activation patterns can be summarised by taking the group mean activation patterns (as we do for the STRM-Regression model); or of a subject level summary measure such as an ANOVA *F*-statistic (as we do for the STRM-Classification model). (c) Behaviourally meaningful variation in the timing of the activation patterns can be identified by regressing behavioural readouts on individual trials against state timecourses, and fitting group statistics to the regression parameters. (d) The STRM model can be inverted to make multivariate predictions on unseen test data; we can then run standard group statistics on the decoding accuracies obtained

hypothesis tests that, for example, do group level inference on W . However as has been established (Friston et al., 2008), the residual covariance matrix is essential when the model is inverted to make

multivariate predictions about unseen test data. Thus, our focus is on the full, multivariate GLM where all terms of the covariance matrix are inferred and used to generate predictions.

2.2 | SpatioTemporally Resolved MVPA (STRM)

We propose to extend this model hierarchically using the HMM framework. This assumes some level of correlation of activation patterns over multiple timepoints that is modelled by a latent state variable, at the same time as having the ability to capture differences in dynamics over trials. Specifically, we model the data vector observed at any point in time t on trial n as conditional upon a latent, discrete and mutually exclusive state $Z_{n,t} \in [1, 2, \dots, K]$, as follows:

$$P(Y_{n,t} | Z_{n,t} = k, X_{n,t}, W_{\{1:K\}}, \Sigma_{\{1:K\}}) = N(X_{n,t} W_k, \Sigma_k) \quad (2)$$

where K is the number of states, and the k th state is associated with a distinct set of residual covariance patterns Σ_k and stimulus activation patterns W_k that linearly map between the trial stimulus X and the data Y . The latent states $Z_{n,t}$ can then be inferred (with suitable constraints, as outlined below) to explain exactly when each distinct pattern emerges on each different trial, allowing us to infer the combination of activity patterns and corresponding latent state timecourses (by which term we refer specifically to the posterior state probabilities) that best explain the data at each time point (Figure 2).

2.3 | Bayesian model hierarchy

We wish to estimate the unknown model parameters $[Z_{\{1:N,1:T\}}, W_{\{1:K\}}, \Sigma_{\{1:K\}}]$, as well as the hidden transition probabilities Φ and hierarchical prior parameters $\Lambda_{\{1:K\}}$ over the regression model

weights (defined fully below). We do this by inferring the full posterior distribution as follows (see also Figure 3):

$$P(Z_{\{1:N,1:T\}}, W_{\{1:K\}}, \Sigma_{\{1:K\}}, \Lambda_{\{1:K\}}, \Phi | X_{\{1:N,1:T\}}, Y_{\{1:N,1:T\}}) \\ = \frac{P(Y_{\{1:N,1:T\}} | X_{\{1:N,1:T\}}, Z_{\{1:N,1:T\}}, W_{\{1:K\}}, \Sigma_{\{1:K\}}) P(Z_{\{1:N,1:T\}}, W_{\{1:K\}}, \Sigma_{\{1:K\}}, \Lambda_{\{1:K\}}, \Phi)}{P(Y_{\{1:N,1:T\}} | X_{\{1:N,1:T\}})} \quad (3)$$

Note we have used the notation $X_{\{1:N,1:T\}}$ to denote the entire set of design matrix entries over trials 1 to N and over all timepoints within those trials 1 to T . While these design matrix entries could vary as a function of time within each trial, they do not in either of the datasets we analyse here (i.e., each trial has a single value for each column of the matrix X that does not change until the next trial). In the below analysis, we expand upon these terms, with modelling decisions explained and justified in turn. We omit the model evidence term (i.e., the denominator in the above equation) as it shall be methodologically sufficient to compute this posterior up to proportion.

2.3.1 | Data likelihood

The likelihood of each observation is conditionally independent of every other observation given the current value of the latent state variable. Thus, we can write the full likelihood of all observations over N trials each of T timepoints as a product over time and trials:

$$P(Y_{\{1:N,1:T\}} | X_{\{1:N,1:T\}}, Z_{\{1:N,1:T\}}, W_{\{1:K\}}, \Sigma_{\{1:K\}}) \\ = \prod_{n=1}^N \prod_{t=1}^T P(Y_{n,t} | X_{n,t}, Z_{n,t}, W_{\{1:K\}}, \Sigma_{\{1:K\}}) \quad (4)$$

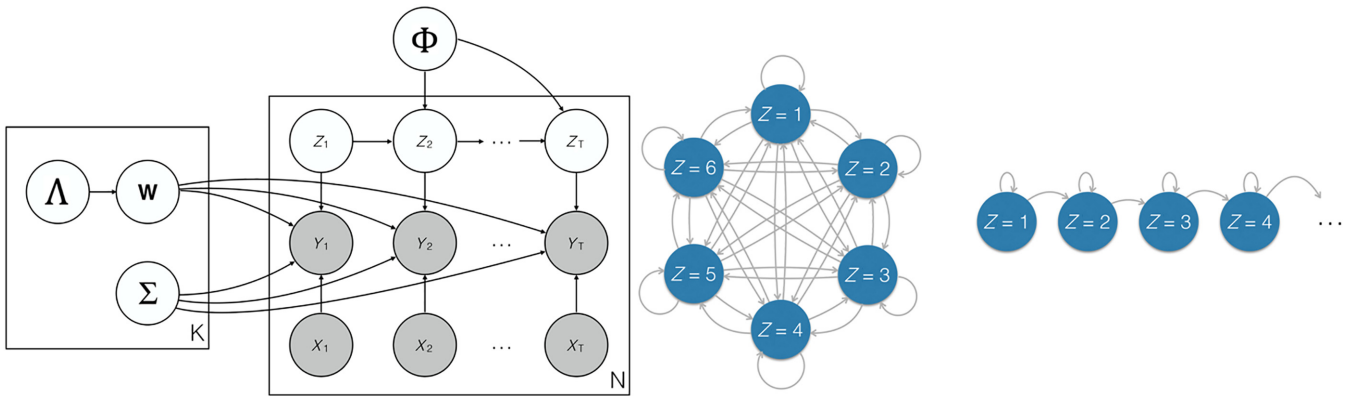


FIGURE 3 Bayesian model outline and left-to-right sequential state dynamics. Left Panel: the full model outline in Bayesian plate notation. For each of N trials of length T , we have data observations $Y_{n,t}$ conditioned upon the corresponding design matrix entries $X_{n,t}$. These data observations $Y_{n,t}$ are also conditioned upon a latent Markov variable $Z_{n,t}$ which models the state sequence unique to each trial, and upon the associated state parameters W_k and Σ_k , which are modelled separately for each of the K total states. The latent state variables are themselves conditioned upon the transition matrix Φ , while the activation patterns in each state are conditioned upon an automatic relevance determination prior parameter Λ . Right panel: We depart from conventional HMM modelling, which freely permits any state to transition to any other state as in the diagram on the left, by instead imposing a left-to-right sequential HMM. As shown on the right, this restricts the permissible state transitions to a consecutive sequence, such that state 1 can either persist or transition to state 2 at each timestep; similarly, if state 2 is active it can either persist or transition to state 3 at each timestep. This structure imposes more aggressive regularisation to overcome the overfitting issues often associated with supervised HMM models

where each individual observation is modelled by a GLM, as in Equation (2).

2.3.2 | Left-to-right latent state prior

We assume that $P(Z_{\{1:N,1:T\}}, W_{\{1:K\}}, \Sigma_{\{1:K\}}, \Lambda_{\{1:K\}}, \Phi)$ factorises over parameters, that is:

$$P(Z_{\{1:N,1:T\}}, W_{\{1:K\}}, \Sigma_{\{1:K\}}, \Lambda_{\{1:K\}}, \Phi) = P(Z_{\{1:N,1:T\}} | \Phi) P(\Phi) \prod_{k=1}^K P(W_k | \Lambda_k) P(\Lambda_k) P(\Sigma_k) \quad (5)$$

and model the latent state variable $Z_{n,t}$ using a Markovian prior:

$$P(Z_{\{1:N,1:T\}} | \Phi) P(\Phi) = P(\Phi) \prod_{n=1}^N P(Z_{n,1}) \prod_{t=2}^T P(Z_{n,t} | Z_{n,t-1}, \Phi) \quad (6)$$

However, we make one important departure from the far more ubiquitous unsupervised HMM model. In the cases where HMMs are used for unsupervised analyses of M/EEG data, for example, to find resting state networks (Higgins et al., 2021; Vidaurre et al., 2018), they are typically allowed to transition freely between states. In supervised data analysis, this can lead to severe overfitting (Ghahramani, 2001), so we instead constrain the model to follow a common trajectory over each trial: we assume that the state order is a fixed sequence, where only the timing of state transitions is allowed to vary. Every trial begins in state 1 and subsequently progresses consecutively through the states, with state transition times freely inferred a trial specific basis (see Figure 3). This has the advantage of enforcing an interpretable sequence of activation over trials, while reducing the number of free parameters governing state transitions. Where unconstrained HMMs must consider a full $K \times K$ transition matrix, this left-to-right HMM constraint means we need only model a $K \times 1$ vector Φ , the k th entry of which (denoted by p_k) captures the probability of state k transitioning to state $k+1$. This structure is enforced by the following prior over the state transitions:

$$P(Z_{n,1} = k) = \begin{cases} 1 & \text{if } k = 1 \\ 0 & \text{if } k \neq 1 \end{cases}$$

$$P(Z_{n,t} | Z_{n,t-1} = k, p_k) = \begin{cases} k+1 & \text{with prob } p_k \\ k & \text{with prob } 1 - p_k \end{cases} \quad (7)$$

We then set the following conjugate priors over the transition matrix entries (a Dirichlet distribution—as the multivariate extension of a Beta distribution—is the conjugate prior to HMM transition probability matrices in general, however as our left-to-right HMM constraint limits the dimensionality of each transition matrix row to 2, this is reduced to a standard Beta distribution):

$$P(\Phi) = \prod_{k=1}^K \text{Beta}(p_k | \alpha, \beta) \quad (8)$$

We set the hyperparameters $\alpha = 1$ and $\beta = 1$ (corresponding to a uniform distribution).

2.3.3 | Observation model parameter prior

For the observation model parameters $W_{\{1:K\}}$ and $\Sigma_{\{1:K\}}$, we apply conjugate priors; this is an inverse Wishart distribution for the covariance matrix and an automatic relevance determination (ARD) prior as specified below for the stimulus activation patterns. The use of an ARD prior prunes away inferred stimulus activation patterns on sensors that are less consistent over trials, in a manner that performs favourably in neuroimaging (Woolrich et al., 2009; Yamashita, Sato, Yoshioka, Tong, & Kamitani, 2008). Denoting by $w_{k,ij}$ the i,j th entry of the matrix W_k , and similarly by $\lambda_{k,ij}$ the i,j th entry of the matrix Λ_k , this is implemented as follows:

$$P(W_{\{1:K\}}, \Sigma_{\{1:K\}}, \Lambda_{\{1:K\}}) = \prod_{k=1}^K P(\Sigma_k) \prod_{i=1}^P \prod_{j=1}^Q P(w_{k,ij} | \lambda_{k,ij}) P(\lambda_{k,ij}) \quad (9)$$

$$P(\Sigma_k) = W^{-1}(A, \nu)$$

$$P(w_{k,ij} | \lambda_{k,ij}) = N(0, \lambda_{k,ij}^{-1})$$

$$P(\lambda_{k,ij}) = G(a, b)$$

We then set the values of the hyperparameters to $A = \frac{1}{P} I_P$ and $\nu = P$, where I_P denotes the identity matrix of dimension P . A result of this is that the expected value of the covariance matrix inverse Σ_k^{-1} under this prior is the identity matrix I_P , corresponding to a prior specifying normalised and uncorrelated data (which itself is a justified assumption given the model is fit to the principal component space as outlined below). We furthermore set hyperparameter values $a = 1$ and $b = 1$. This completes the Bayesian hierarchy.

By inferring a full covariance matrix and matrix of regression weights for each state, we have a potentially quite large parameter space that grows as a function of the number of states K . While the above priors ensure that the model parameters are always well-defined, their accuracy may nonetheless suffer when the number of states is large. Our weakly informative priors can be interpreted as adding a small number of virtual data points to the analysis that encode weakly held assumptions about the data—while this regularises the problem and ensures parameters are well-defined, if there are very few data points assigned to a state then these assumptions may still become overly dominant in the parameter estimation. This is one of the reasons we use cross validation to optimise the hyperparameter K . In other datasets where the number of datapoints is much more limited than those characterised here, our toolbox also supports further limiting the parameter space by restricting the form of Σ_k to be diagonal and/or to be uniform over different states.

2.3.4 | Variational Bayesian inference

As a hierarchical extension of a linear model with conjugate priors, this model is amenable to classic variational Bayesian inference methods. These methods are efficient and scale well to large datasets (Beal, 2003); in the context of the specified model, they are an extension of the popular expectation maximisation algorithm to full posterior inference (i.e., rather than point estimates) for each of the specified model parameters, which are generally more robust than point estimate methods (Johnson, 2007).

Variational Bayesian methods apply the mean-field approximation, whereby they assume the full posterior specified in Equation (3) can be approximated by the following factorised form:

$$P(Z_{\{1:N1:T\}}, W_{\{1:K\}}, \Sigma_{\{1:K\}}, \Lambda_{\{1:K\}}, \Phi | X_{\{1:N1:T\}}, Y_{\{1:N1:T\}}) \approx Q(\Phi) \prod_{k=1}^K Q(W_k) Q(\Sigma_k) Q(\Lambda_k) \prod_{n=1}^N \prod_{t=1}^T Q(Z_{n,t}) \quad (10)$$

This factorisation assumption approximates the full posterior as a product of independent univariate distributions. The advantage of such an approach is that it allows the inference of the approximate posterior—which is otherwise analytically intractable—to be framed as an optimisation problem, where minimisation of the free energy equates to minimising the Kullback–Leibler divergence between the true posterior and the approximate factorised posterior. Crucially, this optimisation problem naturally breaks down to series of analytically tractable sub-routines by virtue of the factorisation applied (Beal, 2003). In the context of HMMs, these are referred to as the VB-M step, which optimises the approximate factorised posterior distribution over observation model parameters $Q(W_k)$, $Q(\Sigma_k)$, and $Q(\Lambda_k)$ and hidden state transition probabilities $Q(\Phi)$; and the VB-E step, which optimises the approximate factorised posterior distribution over hidden state variables $\prod_{n=1}^N \prod_{t=1}^T Q(Z_{n,t})$. These two steps are iterated until convergence.

This procedure requires an initialisation step; we initialise the model using uniform (over trials), state timecourse parameters where each state is visited for an interval T/K seconds long (T being the length of the trial and K being the number of states). This corresponds to an initialisation assumption of no variation in the state timecourses or state duration over trials. We then proceed with the variational inference procedure as outlined above. For further technical details including variational update equations (see Higgins, 2019); for evidence of variational inference converging to known ground truth in simulations see Figure S1.

2.4 | Inverting the encoding model

What we have outlined above is a generative *encoding* model, mapping the spatiotemporal activity patterns associated with each stimulus. Such a model can be inverted to obtain an equivalent *decoding* model by Bayes rule (Friston et al., 2008; Haxby et al., 2014; Naselaris

et al., 2011). Specifically, if we define the generative encoding model parameters learned on some training set as $\theta = \{W_{\{1:K\}}, \Sigma_{\{1:K\}}\}$, for a held-out test set of data Y_{test} we can make predictions about the associated stimulus X_{test} as follows:

$$P(X_{\text{test}} | Y_{\text{test}}, \tilde{Z}_{\text{test}}, \tilde{\theta}) = \frac{P(Y_{\text{test}} | X_{\text{test}}, \tilde{Z}_{\text{test}}, \tilde{\theta}) P(X_{\text{test}})}{P(Y_{\text{test}} | \tilde{Z}_{\text{test}}, \tilde{\theta})} \quad (11)$$

where \tilde{Z}_{test} and $\tilde{\theta}$ are parameter estimates defined fully below. Bayes rule used in this manner allows us to conveniently map between an encoding model and its equivalent decoding model, allowing us to make predictions of the associated stimuli for any unseen test data. Note that $P(Y_{\text{test}} | X_{\text{test}}, \tilde{\theta}, \tilde{Z})$ refers to the GLM observation model introduced above (Equation (2)). When fitting this model to unseen test data, we substitute point estimates $\tilde{\theta} = \{\tilde{W}_{\{1:K\}}, \tilde{\Sigma}_{\{1:K\}}\}$ using the expected values from the posterior distributions inferred on the training dataset (Equation (10)).

We similarly need to estimate the expected values of the latent state variables \tilde{Z}_{test} for each timepoint and trial in the test data. However, as we are using a supervised HMM, these variables cannot be computed by the same inferential step used in training, as this would require previous knowledge of the held-out labels we are trying to predict (Vidaurre et al., 2019). Furthermore, existing Bayesian approaches that compute a joint posterior over both design matrix entries and latent state timecourses and then marginalise out over the latent state variables (Beal, 2003) are not tractable without imposing additional assumptions that seemed unsuitable to us here (see section 2 of Supporting Information). Emphasising that this problem arises only in the computation of held-out accuracy metrics and not the model-fitting pipeline (for which state timecourses are estimated using the variational Bayesian methods outlined in Section 2.3.4), we instead estimate the cross-validated state timecourse variables using a conservative post hoc procedure that is entirely independent of the true label values (shown schematically in Figure S3). To simplify notation let us define Γ_t as the $[1 \times K]$ vector of posterior state probabilities at time t , where the k th entry $\gamma_{t,k} = P(Z_t = k | Y_{\text{test}}, \tilde{\theta})$, such that we wish to obtain an estimate of $\tilde{\Gamma}_t$ for the held-out trials. Taking the training data Y_{train} and the inferred posterior state timecourse probability Γ_t , we train a linear regression model at each timestep within a trial to estimate the inferred state timecourses from the training data itself: $\Gamma_t = Y_{t,\text{train}} B_t + \epsilon_t$. We then use these linear weights to estimate the test set timecourses for all states from the data itself at each timepoint within a trial:

$$\tilde{\Gamma}_t = \sigma(Y_{t,\text{test}} B_t)$$

where σ denotes the softmax function that ensures these are probability estimates that sum to one. We discuss the implications of this step later in the article and expand upon alternative choices in section 2 of Supporting Information (see also Figure S4). We also emphasise that this step is only used when computing cross-validated accuracy metrics, and as such it does not feature in the initial model

fit (Figure 2a), the encoding model analysis (Figure 2b), or the analysis of pattern timing variation (Figure 2c).

The other relevant term for our model inversion is the prior over the structure of the design matrix, $P(X)$. We consider two cases, the first for where the design matrix consists of mutually exclusive classes (i.e., a classification paradigm); and a second where the design matrix may contain continuous valued regressors (i.e., a regression paradigm). We refer to these as SpatioTemporally Resolved MVPA (STRM)-Classification and STRM-Regression, respectively.

2.4.1 | STRM-classification

When the design matrix consists of a total of Q mutually exclusive classes, the form of $P(X)$ is categorical. Assuming the vectors X_t follow a one-hot vector encoding scheme, and writing the prior probability of each class as a $[Q \times 1]$ vector C , we have the following:

$$P(X_t) = X_t C \quad (12)$$

Assuming all classes are equally probable we have all entries of $C_i = \frac{1}{Q}$. This leads us to an analytical solution to Equation (11) (see Appendix A for details):

$$P(X_t|Y_t) = X_t \sigma(L) \quad (13)$$

where L is a $[Q \times 1]$ vector of unnormalized class likelihoods with each entry $L_i = \sum_{k=1}^K \frac{\tilde{\gamma}_{t,k}}{2} (Y_t - l_i \tilde{W}_k) \tilde{\Sigma}_k^{-1} (Y_t - l_i \tilde{W}_k)^T$; where l_i denotes the $[1 \times Q]$ vector obtained by taking the i th row of the identity matrix of dimension Q , and $\sigma(L)$ is the softmax function, which outputs a $[Q \times 1]$ vector with each entry $\sigma_i(L) = \frac{e^{L_i}}{\sum_{j=1}^Q e^{L_j}}$. Note that $\tilde{\gamma}_{t,k}$ is the estimate for the probability of each state being activated at time t (i.e., the terms discussed above that are estimated by a regression model).

This solution is equivalent to classification by Linear Discriminant Analysis (LDA); so, with the inferred latent state dynamics $\tilde{\gamma}_{t,k}$ we now have a dynamic form of LDA. Note that in different applications, users may also choose to model the covariance matrix Σ as strictly diagonal; in which case this model is equivalent to a dynamic form of Gaussian Naive Bayes classification.

2.4.2 | STRM-Regression

Similarly, given a model that relates a set of continuous valued regressors X_t to observed data Y_t , we can make new predictions given some assumption of the overall distribution of the regressors; we will assume they are standardised and uncorrelated, such that:

$$P(X_t) = N(0, I_Q) \quad (14)$$

where I_Q is the identity matrix of dimension Q . As a conjugate prior, this ensures for any new observation Y_{test} that the posterior of Equation (11) has a Gaussian distribution (see Appendix B):

$$P(X_t|Y_t) = N(\mu_{x|y}, \Sigma_{x|y})$$

$$\Sigma_{x|y} = \left[I_Q + \sum_{k=1}^K \tilde{\gamma}_{t,k} \tilde{W}_k \tilde{\Sigma}_k^{-1} \tilde{W}_k^T \right]^{-1}$$

$$\mu_{x|y} = \Sigma_{x|y} \left[\sum_{k=1}^K \tilde{\gamma}_{t,k} \tilde{W}_k \tilde{\Sigma}_k^{-1} Y_{\text{test}}^T \right] \quad (15)$$

In the absence of the latent state variable $\tilde{\gamma}_{t,k}$, this is a case of linear Gaussian systems (LGS) model (Murphy, 2012; Roweis & Ghahramani, 1999); thus, with the inclusion of the latent state variable, we have a dynamic LGS model.

2.5 | MEG visual data analysis

We test the model on two datasets (for additional verification on synthetic data simulated from the generative model see Supporting Information). The first dataset we analyse comprises a visual stimulus decoding task previously published as part of a larger study (Liu, Dolan, Kurth-Nelson, & Behrens, 2019).

2.5.1 | Task outline

All participants signed written consent in advance; ethical approval for the experiment was obtained from the Research Ethics Committee at University College London under ethics number 9929/002. A total of 22 participants fixated on a cross onscreen and were presented with visual stimuli in a randomised order. There was a total of eight distinct visual stimuli (for details see Liu et al., 2019). To ensure continuous engagement with the task, on 20% of trials the stimuli were inverted; the participant was required to push a button to indicate if the stimuli was inverted. The below analysis uses only the data on the non-inverted trials.

2.5.2 | Data preprocessing

MEG data was acquired at a rate of 600 samples per second on a 275 channel CTF scanner. Data was filtered within a passband of 0.1–49 Hz, downsampled to 100 samples per second using a polyphase low-pass filter with cutoff 25 Hz (Higgins, van Es, Quinn, Vidaurre, &

Woolrich, 2022), then epoched to extract periods in time from the moment of stimulus presentation to 500 ms later. To ensure training data was balanced across classes, for each subject we determined the number of trials in each class and, if n is the number of trials of the least sampled class, only kept the first n trials of all other classes. This resulted in balanced class sets for each subject, with different subjects having between 26 and 33 total samples of each class. At this point the data was separated into two streams. First, the sensor space data to be used for learning the decoding models underwent dimensionality reduction using PCA, which was applied to each subject's epoched data (i.e., the matrix of all trials for a subject concatenated together) keeping the top 50 components (which accounted for 98.2% of the total variance) (Grootswagers et al., 2017). We then normalised each principal component by the variance over all that subject's concatenated data, such that each component had unit variance. Second, and separately, the same data was projected into source space using a LCMV beamformer (van Veen, van Drongelen, & Suzuki, 1997; Woolrich, Hunt, Groves, & Barnes, 2011) projecting onto an 8 mm MNI grid. Beamformer weights were computed using the data covariance computed over each subjects' epoched data, concatenated over all trials—notably our models have the potential to infer latent-state specific data covariances, which may in future work further enhance the accuracy of beamformer projections but which was not pursued in the present study. Source data was then parcellated into 38 anatomically defined regions of interest (ROIs) derived from an independent component analysis of fMRI resting state data from the Human Connectome Project (Colclough et al., 2016). This parcellation has been proven to be effective in a number of applications to MEG data (Colclough et al., 2017; Higgins et al., 2021; Vidaurre et al., 2018). Source leakage was then corrected for by orthogonalization as outlined in (Colclough, Brookes, Smith, & Woolrich, 2015). Ultimately the reconstructed source data only accounted for 46% of the total variance in the original sensor data—thus, this data must be interpreted with caution as it only reflects a partial view of the total information available to the classifiers, but can nonetheless be informative as to the neural origins of this activity.

2.5.3 | STRM-Classification model

As the stimulus set was categorical, we established a design matrix with nine regressors; these corresponded to one regressor for each distinct visual stimulus, along with an intercept term. Thus, the inferred model coefficients for each latent state corresponded to a generic activation pattern over all stimuli for that state (i.e., the intercept term), and an effect specific to each visual stimulus for that state. We initially fixed the parameter for the number of states to $K=8$ and sought to explore the activation parameters and state timecourses for this fixed number of states. Only later in the pipeline, when determining classification accuracy, do we seek to optimise this parameter. Where the parameter optimisation procedure could not be used, we can demonstrate that our results and conclusions are not sensitive to the specific choice of hyperparameter (see section 3 in Supporting Information and Figure S5).

We fit our STRM-Classification model to the principal component space derived from the sensor level data as in Grootswagers et al. (2017). Models were fit independently to the data from each subject. When evaluating spatial activity maps and correlates of posterior state probabilities, we fit a single model to all data for each subject; when evaluating classification accuracy we trained and tested an ensemble of models for each subject in a cross-validation procedure outlined below.

2.5.4 | Characterising spatial activity maps

Having determined the state timecourses from fitting our STRM model, we projected these state timecourses back onto the original sensor data to obtain sensor activity maps per stimulus per state, and then for additional interpretability onto the equivalent source localised data to obtain source space activity maps per stimulus, per state and per subject (Figure 2b). Importantly, the interpretation of this source space projection is subject to the usual caveats of the inverse problem, which is ill-posed and does not have a unique solution without imposing additional modelling assumptions. Our use of LCMV beamformers finds the lowest power solution to the inverse problem subject to a unit gain constraint (van Veen et al., 1997; Woolrich et al., 2011). Further caution is warranted given the source data represent only a partial view of the information content available to the sensor space model. The resulting activity maps can nonetheless be informative as to the neural origins of the signal, despite these caveats. As we are interested in the regions that *differentiate* the different stimuli, we computed the f statistic for a within subject ANOVA at each ROI, effectively asking how different the representations of the eight stimuli were at that particular ROI, for that subject. Maps in Figure 5 plot the average F -statistic over subjects per voxel, thresholded at the top 75th percentile for visualisation.

2.5.5 | Characterising state timecourse variability

The state timecourses inferred from fitting our STRM model characterise the timing at which specific activity patterns emerge on individual trials. We wished to explore what variables might influence these timings, and so we fit a post hoc multiple regression model asking whether specific behavioural and physiological variables allowed us to predict the timing of specific states on individual trials (as illustrated schematically in Figure 2c).

For each subject, given a STRM fit with K states, we fit $(K-1)$ multiple regression models, with the k th model predicting the timing of the transition from the k th to the $(k+1)$ th state (i.e., using the state transition time as the dependent variable). In each model we used the same four independent variables; the first two regressors were the inter-stimulus interval (ISI) and participant reaction time (see Figure 6), capturing variation in the overall timed structure of the trial. Note that the inter stimulus interval was randomly generated for each trial from a uniform distribution between 0.5 and 2 s, whereas the reaction time

was determined by the participant and thus reflects a behavioural readout.

The remaining two independent regressors were designed to determine whether spontaneous changes in baseline electrophysiological patterns immediately prior to image presentation influenced the speed at which different patterns emerged. For this, we computed the power spectral density (PSD) in each ROI of the source space data over the 200 ms preceding stimulus presentation. This data was considerably high dimensional, so we applied a non-negative matrix factorisation across the data from all subjects to extract two primary modes of spatio-spectral variation. Specifically, we arranged the baseline PSD estimates for each subject across N ROIs in F frequency bands and M trials into a single matrix of dimension $[M \times NF]$. We then applied a non-negative matrix factorisation (NNMF) decomposition to the row-wise concatenation of all subjects' PSD matrices as in Vidaurre et al. (2018), obtaining two main modes of spatio-spectral variation that corresponded to a broadband mode and a visual alpha mode (see Figure 6). We used the expression strengths of each of these two modes on each trial as the two independent variables in the multiple regression model, thus asking the degree to which baseline broadband power and baseline visual alpha power influenced the timing of subsequent visual processing states.

Finally, to eliminate collinearity we decorrelated all four regressors using symmetric orthogonalization. These multiple regression models were fit at the subject level and comprised a total of $4(K - 1)$ multiple comparisons. We evaluated significance of effects at the group level through two tailed t -tests on the distribution of model coefficients over subjects, using Bonferroni correction of p values.

2.5.6 | STRM-Classification predictive accuracy

To assess the performance of STRM-Classification model, we used 10-fold cross-validation. This entailed a 10-fold partitioning of each subject's data, followed by an iterative procedure of holding out one fold for testing while training the model on the remaining data for that subject, then testing the classification performance using the procedure outlined in Equation (11) on the held-out partition. This was repeated iteratively until all trials had been classified; we defined classification accuracy as the proportion of trials upon which the correct label was predicted by the classifier.

Whereas we previously held the parameter K for the number of states fixed to a single value, the classification accuracy provides a clear metric by which this parameter could be optimised. Thus, the above cross-validation procedure was run using a total of 10 different values of K ranging from 4 to 22 in steps of 2. These can either be evaluated separately or optimised by nested cross-validation, with the data of the different subjects forming the outer cross-validation loop. This latter procedure entails holding out one subject, determining the value of the parameter K that maximises the classification accuracy for the remaining subjects, and selecting that value when determining the accuracy for the held-out subject.

To assess statistical significance, we used two tests. To test whether the STRM-Classifier accuracy was significantly greater than the equivalent timepoint-by-timepoint classifier accuracy, we used non-parametric cluster permutation tests with a t threshold of 1 (Nichols & Holmes, 2003). To test whether the overall accuracy (averaged over all timepoints) exceeded other classifiers, we first applied a group level ANOVA followed by pairwise t -tests.

2.6 | EEG data analysis

To demonstrate the general applicability of this method, we also applied it to a set of EEG data collected during a decision making and reward learning task. This dataset was selected as it used stimuli (i.e., reward outcomes) that varied continuously rather than categorically, corresponding to the STRM-Regression model. Furthermore, the more complex task design involved additional variables that we believed may modulate neural processes and thus state timecourse patterns.

2.6.1 | Task outline

A total of 30 participants completed the task while undergoing EEG scanning. Twenty-three were of sufficient data quality to be included in the analysis (the other participants were excluded either because of not understanding the task or excessive noise). The task consisted of navigating between different foraging patches in an effort to maximise the total reward accumulated over the course of the recording session. Specifically, there were two types of decisions participants had to make. First, for the "site selection" phase, they chose between one of two patches; they then entered their chosen patch (after a variable waiting period), and accrued reward (the "Reward accrual" phase). During the reward accrual phase, participants were shown their current reward level which changed every second (i.e., decaying from different starting points and different decay rates). Participants were continually prompted: they could either do nothing or continue to accrue reward at a depleting rate, or press the space bar to collect their accumulated reward during this patch visit. They would then again be asked to choose between the two patches. Decay rates and starting points of both patches were learnable but changed periodically throughout the experiment.

For the purposes of our analysis, the only task event analysed is the period where participants receive the overall reward (the "Reward receipt" phase). At that moment, participants had to estimate the average reward rate they had achieved in the current trial in order to decide to leave earlier or later in the next trial or indeed pick the other patch next time. The average reward rate is computed by combining the accumulated reward at a patch and the time invested to receive it.

All participants signed written consent in advance; ethical approval for the experiment was obtained from the Medical Sciences

Inter-Divisional Research Ethics Committee at the University of Oxford under project number R28535/RE001.

2.6.2 | Data preprocessing

EEG signals were recorded with active electrodes from 59 scalp electrodes mounted equidistantly on the standard 10–20 system elastic map (EasyCap). All electrodes were referenced to the right mastoid electrode and re-referenced offline to the average. Continuous EEG was recorded using a SynAmps RT 64-channel Amplifier (1,000 Hz sampling rate). The data were epoched from -1 to 2 s around the reward events. The data were then band-pass-filtered at $[1-35]$ Hz. We denoised our EEG data in a multiple step procedure. First, we removed trials containing large artefacts semi-manually using “ft rejectvisual” function in FieldTrip. This function visualises trial variances and allows the user to remove trials and electrodes directly from the GUI. We did this so that the following ICA would not be dominated by very few excessive noise trials. We then ran an ICA and correlated the component timecourses with the computed vertical and horizontal EOG's. If an ICA component showed high correlations (defined as >0.3) with either EOG and had a characteristic topography for eye movements or blinks, we regressed out that component. In the next step we removed trials with excessive muscle artefacts by taking eight electrodes (“PO7,” “PO8,” “POZ,” “PO4,” “PO3,” “O1,” “OZ,” “O2”) near the neck which were most affected by muscle movements, z-scoring them and removing trials with a conservative z-score cutoff at 20. After muscle filtering a second visual inspection was applied to the data in case some very noisy trials remained. We downsampled the data to 100 samples per second using a polyphase filter with 25 Hz cutoff. Additionally, to further remove ocular artefacts, we regressed the EOG channel signals from each EEG electrode signal. We applied dimensionality reduction as previously described for the MEG data, using PCA with eigenvalue normalisation and keeping the top 20 components (out of 59 channels), which explained a total of 93.6% of the variance.

2.6.3 | STRM-Regression model

Each trial consisted of a presentation on screen of the reward value that had just been accumulated. For our STRM-Regression model, we used a design matrix, X , with two continuous regressors: the first being a mean activation value to model the mean change in activity for a state over all trials, and the second being the value of reward presented on screen in that trial. This second regressor was normalised over trials.

We fit the STRM-Regression model to the principal component space derived from the sensor level data as in (Grootswagers et al., 2017). This is done independently to the data from each session for each subject (note that for any group statistics we first averaged model parameters for a given subject over both of their recording sessions, then ran group statistics on these subject average values). As

outlined in Section 2.5.3 we use a total of $K=8$ states (as previously, we initially hold this parameter fixed while characterising the model output, and later show how it can be optimised for decode accuracy). Where the parameter optimisation procedure could not be used, we can demonstrate that our results and conclusions are not sensitive to the specific choice of hyperparameter (see section 3 of Supporting Information and Figure S6).

2.6.4 | Characterising spatial activity maps

Having determined the state timecourses from initially fitting STRM to the principal component space derived from the sensor level data, we then projected the state timecourse information back onto the original sensor data to obtain sensor activity maps per regressor per state. Maps in Figure 8 plot the group average value of model coefficients for each state; that is, the mean activity pattern associated with each state and the value evoked effect within each state.

2.6.5 | Characterising state timecourse variability

In exactly the same way as conducted for the MEG data, having determined the state timecourses from fitting our STRM-Regression model, we then asked whether the timings of these stimulus processing states were significantly modulated by other variables within the task - specifically, variables that the participant is required to track in order to optimally complete the task.

For each subject, we fit $(K-1)$ multiple regression models, with the k th model predicting the timing of the transition from the k th to the $(k+1)$ th state (i.e., using the state transition time as the dependent variable). In each model we used three independent variables. These consisted of the value of the stimulus itself; the total time invested at the site (i.e., how long the participant had spent in order to accumulate the value shown in the stimulus); and the recent reward history (i.e., what the participant might expect to gain from leaving the site in search of another). To eliminate collinearity these variables were decorrelated by symmetric orthogonalization and normalised (Colclough et al., 2015). After fitting these models to the data for each subject, we conducted a group analysis testing whether any of the regressor coefficients were significantly different from zero. This involved $3(K-1)$ multiple comparisons; p values were evaluated for significance after Bonferroni correction.

2.6.6 | STRM-Regression predictive accuracy

In the same manner as outlined for the MEG data, we then assessed the performance of the STRM-Regression model to predict continuous stimuli using 10-fold cross-validation and showed how the parameter K controlling the number of states could be optimised by cross-

validation over subjects. We used as accuracy metric the Pearson correlation between predictions and the true values. To assess statistical significance, when comparing accuracy versus time of two classifiers we applied a cluster permutation test with *t*-statistic threshold of 1 (Nichols & Holmes, 2003); when comparing the overall accuracy (averaged over all timepoints) of a group of models we applied a group ANOVA.

3 | RESULTS

We present here the results obtained from fitting the same model to two separate datasets. The first is a set of MEG recordings of categorical visual stimulus presentations, the second a set of EEG recordings of participants completing a cognitive task.

3.1 | Decoding visual stream representations using MEG

Visual stimuli evoke a cascade of feedforward and feedback activity through the dorsal and ventral visual streams (Goodale & Milner, 1992; Hochstein & Ahissar, 2002). Existing methods to identify the spatiotemporal evolution of these representational dynamics have relied upon methods that fuse MEG and fMRI recordings (Cichy et al., 2014) or require invasive recording types (Goodale, 1993). We asked whether these results could be corroborated using MEG as the sole recording modality, by utilising our STRM-Classification model. We fit our STRM-Classification model to a previously published dataset comprising MEG recordings of 22 participants viewing randomly presented visual images from a set of eight stimuli (Liu et al., 2019).

3.1.1 | Classification accuracy

An advantage of MVPA is that model fit can be straightforwardly assessed using the metric of classification accuracy. We therefore asked how STRM-Classification compared with conventional analyses. Figure 4 plots the classification accuracy obtained by STRM-Classification versus that of the equivalent conventional approach, namely timepoint-by-timepoint LDA classification. This model does lead to enhanced classification accuracy during later timepoints (where activity patterns might be expected to be less well aligned); however, this gain is relatively modest. It appears that although our model learns state timecourse information that we will show correlates with meaningful behaviour and physiological patterns, these do not result in dramatic gains to classifier performance. Performance is reasonably robust to the STRM model order; as shown in the second plot of Figure 4, performance is equivalent for $K=10$ or higher. This plot furthermore identifies a performance tuning curve, making this amenable to parameter optimisation and motivating the

cross-validation approach we use to generate the upper and lower panels of Figure 4 (see Section 2).

There are two innovations that the difference in classification accuracy shown in Figure 4 (top panel) could be ascribed to; it could be due to the forward modelling procedure and the assumptions it imposes (which in this case are equivalent to the assumptions of LDA); and/or it could be due to the time-varying dynamics that our model permits. To test the effect of the forward modelling procedure, we compared the accuracies achieved by a range of different classification methods. First, we compared the classification accuracy obtained using generative classifiers (LDA and Naive Bayes) with those achieved by a range of discriminative machine learning classifiers (Support Vector Machines—linear and with radial basis function kernels; logistic regression in binomial and multinomial configurations; and K-nearest neighbour classifiers). All classifiers were implemented on a timepoint-by-timepoint basis. As shown in Figure 4, generative model-based classifiers offer significant improvements in classification accuracy over these methods. Thus, we conclude that—at least for this dataset—the assumptions imposed by our generative model accurately represent the data and afford greater classification accuracy as a result.

3.1.2 | STRM visual stream activation patterns

Given STRM-Classification models fit independently to each subject's data, we asked how consistent the inferred spatial patterns of activity were over subjects. Specifically, we asked whether ROIs emerged that were particularly informative of class differences, and whether these were the same across subjects. The interpretation of these source-reconstructed maps is subject to the usual caveats of the inverse problem, which is ill-posed without additional modelling assumptions (see Section 2.5.2)—these can nonetheless be informative if these modelling assumptions turn out to be a good fit for the data. Nothing in the model expressly enforces common patterns of activation for a given state across subjects, these are only unified by their common position in the temporal sequence of states (see Figure 3). Figure 5 plots the mean state sequence timings across all subjects for a STRM-Classification model fit with $K=8$ states. As demonstrated in the central panel of the figure, which shows the inferred state timecourses on individual trials for an example subject, all trials follow the enforced left-to-right HMM state sequence while allowing for variability in precise activation pattern timing on individual trials.

The STRM-Classification model was fit to the principal component space computed from the sensor data (see Section 2.5.2); to assess which ROIs were most informative as to the different class, we projected the state timecourses onto the source space data and conducted subject-level ANOVAs to determine which ROIs displayed the greatest variation over conditions. Plotting the group-mean *f* statistic for each ROI identifies a clear progression through the visual hierarchy. Applying a percentage threshold across all states (see Section 2.5.4), state 1 identifies no significant sources of information,

consistent with its interpretation as preceding the arrival of visual information to cortex. State 2, which accounts for periods around 100 ms after stimulus onset that show elevated decoding accuracy, similarly shows no single ROI above thresholding, suggesting that these early visual representations may have limited spatial consistency over subjects. States 3 and 4 show the propagation of this information from occipital visual cortex to outer visual areas; these states encompass periods in time 120–250 ms after stimulus onset. Throughout all subsequent states, the visual cortex remains the dominant source of information relevant to the decoding, however from states 5–8 significant information additionally emerges in inferior temporal and then

lateral temporal areas. Thus, information first appears in temporal areas in state 5 which corresponds to roughly 250 ms after stimulus presentation, then endures to the end of the recording 500 ms after stimulus presentation (Figure 5).

3.1.3 | Activation pattern timings are modulated by behaviour and physiology

Our STRM model resolves the specific timings on each individual trial when patterns of stimulus-associated activity emerge. We then asked whether these timings were meaningful—specifically, how they related to other measures that varied over individual trials. We investigated this by fitting a multiple regression model to predict the timings of the transitions between different states using the state timecourses inferred by STRM. This multiple regression (fit independently for the transition times between each pair of states) had two regressors indicating event timing on each trial (reaction times and ISIs), and two regressors indicating spontaneous variations in broadband power and visual alpha power during a baseline period 200 ms prior to stimulus presentation (Sederberg, Pala, Zheng, He, & Stanley, 2019). This tested whether behaviour and/or changes in the baseline distribution of power affected the timing of visual information processing (Figure 6).

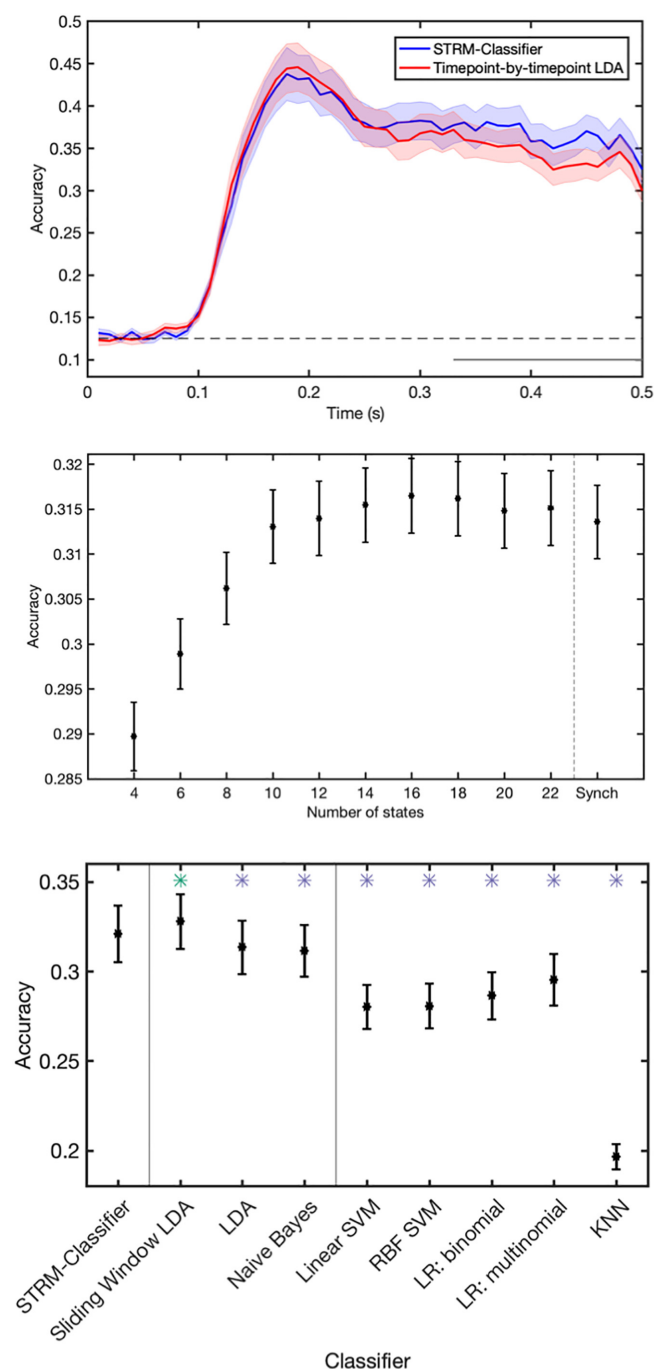


FIGURE 4 Classification accuracy achieved by different MVPA methods. Top panel: Plotting the accuracy (mean over subjects \pm SE) versus time for STRM-Classifer (with K optimised through cross-validation—see Section 2) versus timepoint-by-timepoint spatially resolved decoding identify marginal improvements in classification accuracy over later timepoints when temporal patterns are identified. Middle panel: Plots of mean accuracy over all timepoints between 0 and 0.5 s as a function of the number of states K ; plot shows mean over subjects \pm SE. This shows this relationship is robust for values of the parameter K above a sufficient level; equivalent decoding accuracy is achieved when using $K = 10$ states or higher. Lower panel: the STRM-Classifer performs favourably when compared with discriminative classification methods. We here compare the STRM classifier with eight other classification methods—three other spatially resolved classifiers (LDA with optimised sliding window length—see Section 2; LDA using the conventional timepoint-by-timepoint decoding approach, and Naive Bayes using the conventional timepoint-by-timepoint decoding approach); and additionally five different discriminative classifiers fit using timepoint-by-timepoint methods (Linear SVMs, non-linear SVMs using a radial basis function (RBF) kernel; binomial and multinomial logistic regression (LR), and K -nearest neighbour (KNN) classifiers with K optimised through cross-validation). We find in general that generative encoding model based classifiers (STRM, LDA and Naive Bayes) outperform discriminative classifiers (SVM, LR and KNN), and furthermore that STRM decoding outperforms equivalent methods when used with the conventional timepoint-by-timepoint decoding approach; however, we similarly find that these gains are slightly surpassed by optimised sliding window methods. Asterisks denote significantly different from STRM-Classifer accuracy at Bonferroni corrected levels; green asterisks show significantly higher accuracy (Sliding Window LDA); blue asterisks show significantly lower (all other classifiers)

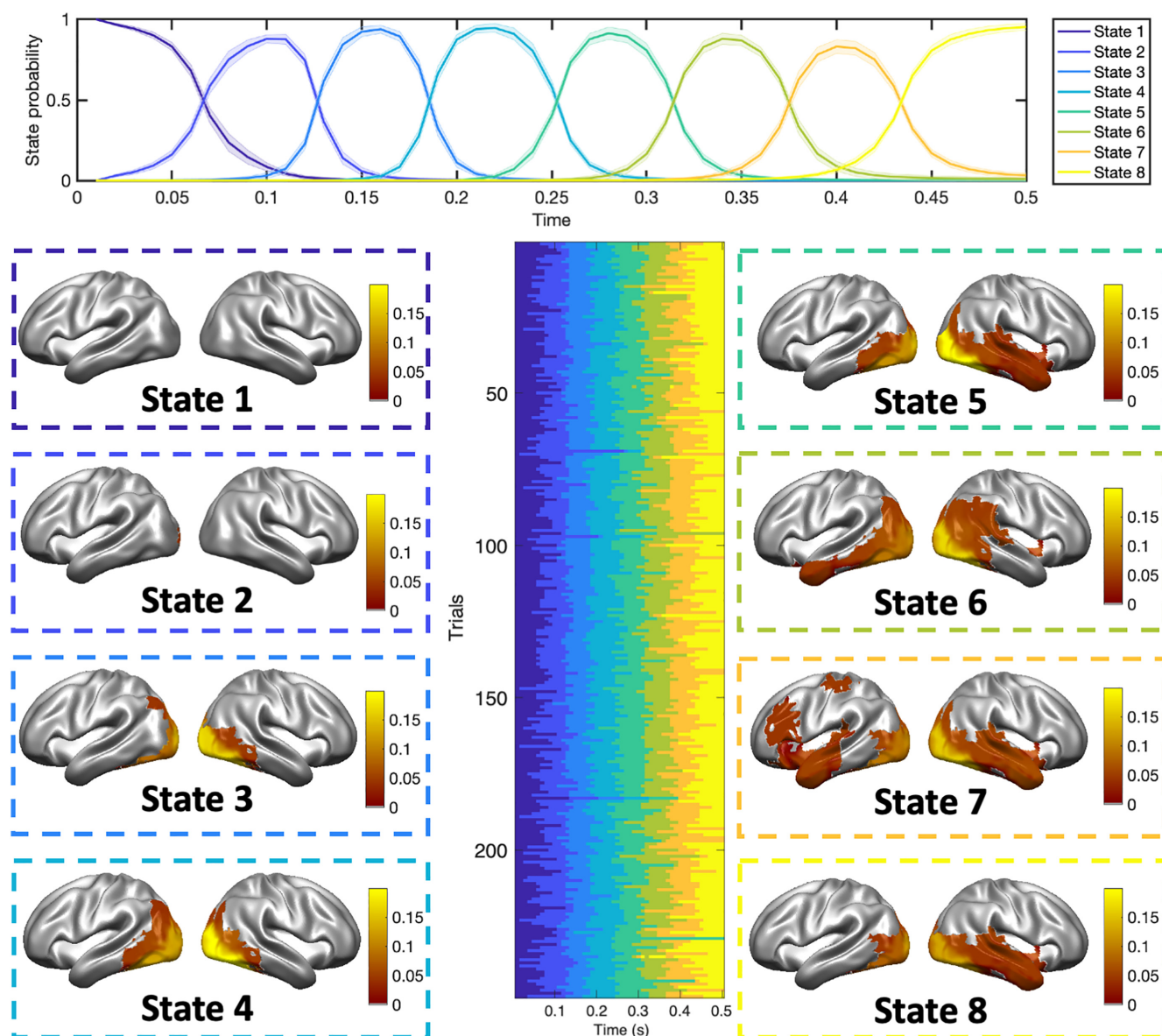


FIGURE 5 Resolving the successive stages of visual stimulus processing in space and time. Fitting the STRM-Classification model independently to each subject's MEG data recorded during visual stimulus presentation, using $K = 8$ states allows us to investigate the stages of visual stimulus processing and the times on individual trials at which they emerge, subject to the usual interpretational caveats associated with the inverse problem and loss of information content in our source reconstruction methods. Top panel: Average timing of each state over a trial (mean \pm SE over subjects), demonstrating the mean time after stimulus presentation that each state emerges. Lower central panel: a raster plot of state timecourses inferred for a sample subject over 248 trials, demonstrating the variability in timings over successive trials within the common left-to-right HMM pattern progressing from state 1 to 8. Lower outer panels: the thresholded, group-mean f statistics, per ROI, as a result of multiple subject-level ANOVAs; this displays the amount of information contributed by that ROI to discriminate the different visual stimuli (see Section 2). Statistics are thresholded at the 75th percentile of all test statistics obtained

We found a strong relationship between response times and state transition times. This is significant for all states from 3 onwards, with an increasing effect size toward later states. ISIs, on the other hand, did not exhibit a significant relationship; there is an apparent trend whereby longer ISIs were associated with faster state sequences, but this was not significant after multiple comparison correction.

Spontaneous changes in the baseline power similarly modulate the subsequent propagation of visual processing states in a manner that appears selective to specific stages in the visual hierarchy. High

levels of broadband power are associated with earlier transitions out of state 1 into state 2, whereas high levels of visual alpha band power are associated selectively with later transitions out of state 4 into state 5—the state we associate with emergence of information outside visual cortex. This is broadly consistent with proposed roles for alpha governing top-down control of attention and gating of information as it moves downstream (Jensen & Mazaheri, 2010); the relationship between broadband activity and earlier visual states is perhaps more surprising, however such patterns have been shown to be strong

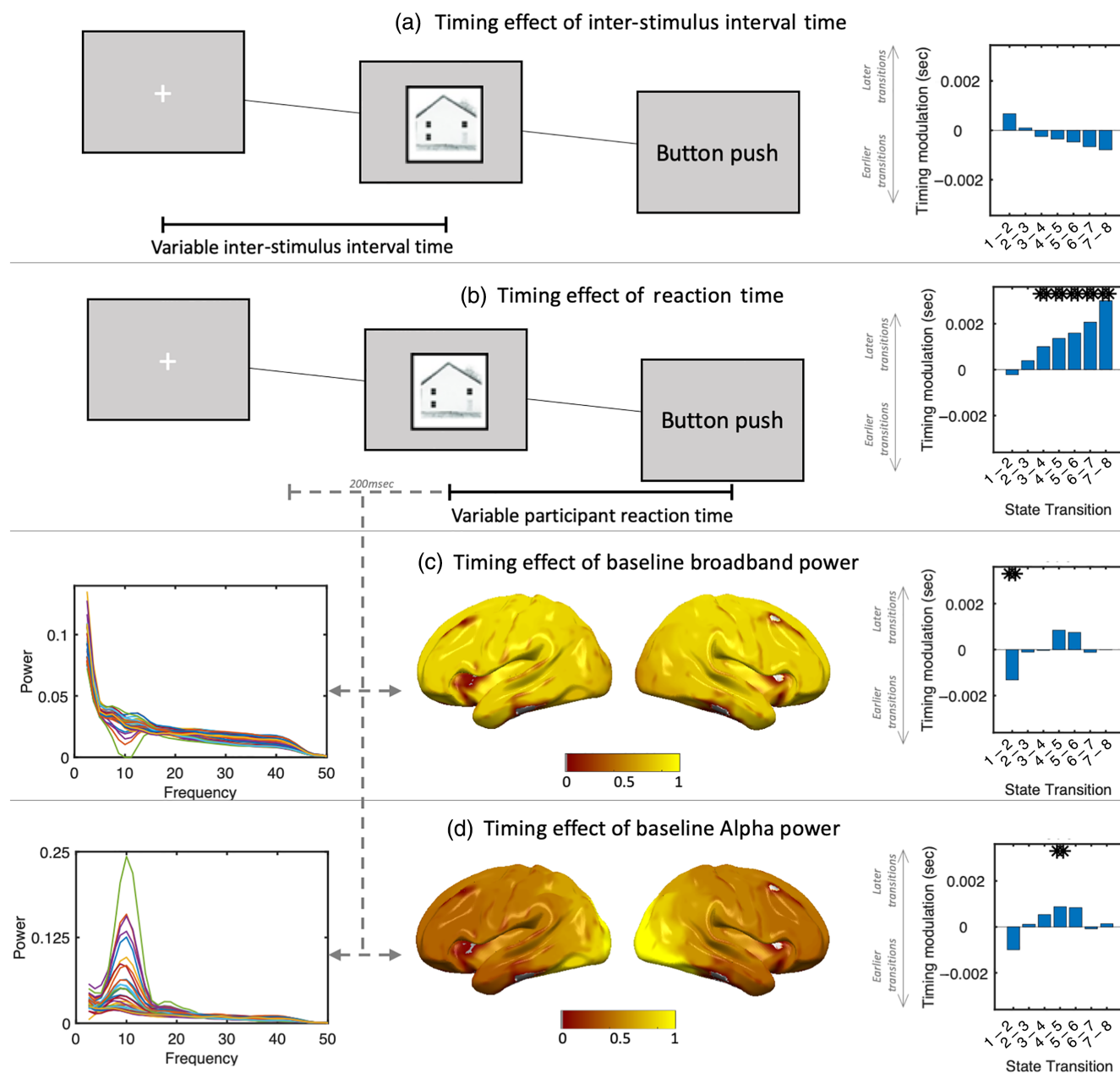


FIGURE 6 The timing of visual processing is modulated by behaviour and physiology. Panel a: inter-stimulus intervals do not significantly modulate state transition times. Panel b: Longer participant reaction times are predictive of delayed transitions into states 3–8, with increasing effect size toward later states. Panel c: increases in baseline broadband power are associated with more rapid transitions into state 2, an early visual processing state. Panel d: increases in baseline alpha power over visual areas is associated with delayed transitions from state 4 into state 5. In all bar plots, asterisks denote significance at Bonferroni corrected levels ($p = 2.1e - 3$)

modulators of neural activity in a wide variety of ways (Miller et al., 2014; Podvalny, Flounders, King, Holroyd, & He, 2019).

These examples serve primarily as a model validation, demonstrating that the observed variation in activity pattern timings does not arise randomly but rather is reflective of underlying neural processes that are behaviourally and physiologically relevant. They furthermore suggest that activation timings in different stages of the visual cortex are not modulated uniformly, but rather that different variables may selectively affect propagation timings in different parts of the visual hierarchy.

3.2 | Decoding stages of cognitive processing in EEG

The ability to discern and analyse the timing of activity patterns emerging across different trials is potentially more salient in the case of non-sensory representations. Recently, decoding has emerged as a popular paradigm for analysing higher cognitive functions in complex tasks at high temporal resolution (Holdgraf et al., 2017; Kriegeskorte & Kievit, 2013), however these methods still mostly assume these higher cognitive functions are perfectly aligned across trials. We thus asked

whether the model we had proposed generalised to a very different EEG dataset, that involved prediction of a continuous variable (value) over multiple trials in a complex foraging task, where contextual variables differed substantially across trials and could potentially determine activity pattern timings.

3.2.1 | Accuracy of dynamic LGS decoding

We first asked whether STRM-Regression resulted in benefits to decoding accuracy. We measured accuracy by the cross-validated Pearson correlation between test set model predictions and their true values. We found no significant differences in performance between STRM-Regression and other metrics. We also asked how sensitive this

behaviour was to the parameter controlling the number of states; as shown in Figure 7 (middle panel) it is very consistent across all values of this parameter tested (no significant differences were observed; one way ANOVA $p = .98$).

We then investigated what drove this performance and how it compared with similar methods. An equivalent encoding model trained using optimised sliding window techniques (i.e., sliding window LGS; see Section 2.6.6) achieved equivalent performance. There remained a small trend, where the sliding window method performed marginally better than STRM-Regression, mirroring the relationship obtained for STRM-Classification, but these differences were not significant. Overall, they suggest that the model's sensitivity to trial specific differences in pattern onset timing—which we can show below have strong correlations with behavioural variables—is nonetheless not a major determinant of predictive accuracy. Finally, the question remains whether the overall forward model-based approach itself performs favourably compared with other discriminative methods. To assess this, we compared model performance against that obtained by three decoding methods: linear regression and SVM regression using either linear or radial basis function kernels. Comparing all models on the basis of their correlation coefficient identified no significant group variation (one way ANOVA: $p = .19$), despite an apparent trend for the decoding models to achieve lower correlations. Overall, these results suggest that the generative forward model approach to decoding performs no worse than more commonly used regression techniques; and that their use alongside more targeted time series methods (i.e., the STRM and sliding window models) may offer very slight performance improvements.

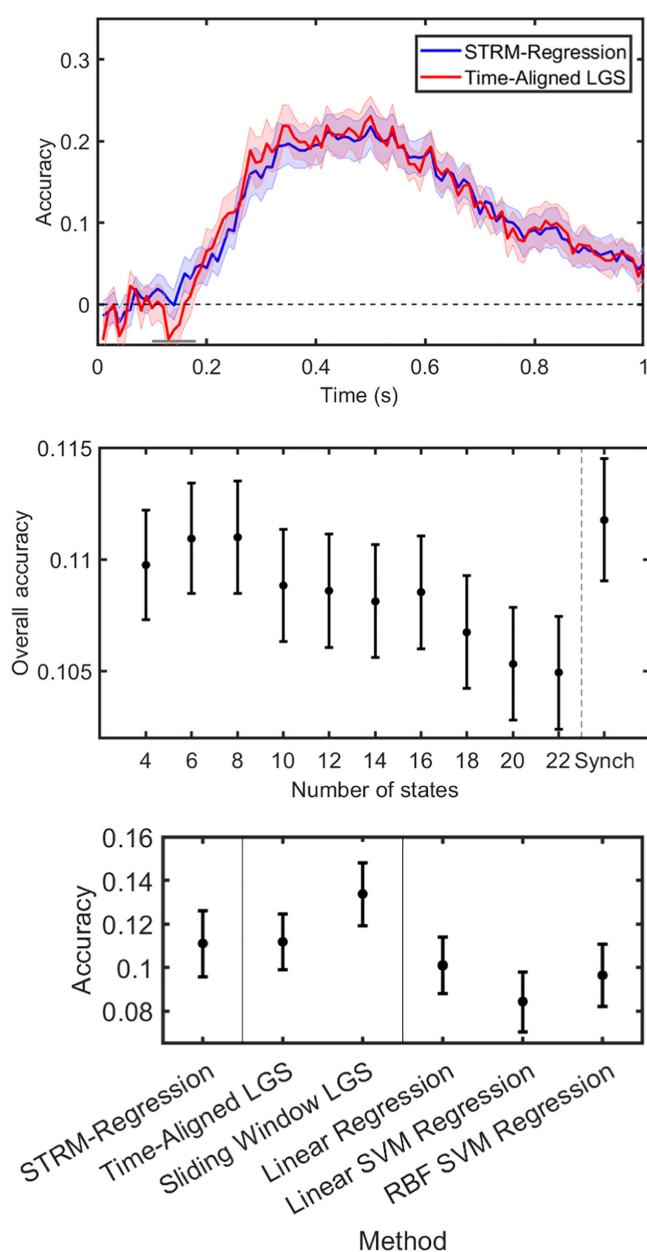


FIGURE 7 Predictive accuracy of STRM-Regression decoding. Top panel: plotting the Pearson correlation between model predictions and true regressor values over time, we see equivalent performance by both metrics. The STRM-Regression output shown here was obtained by optimising the value of K (the number of states) by subject-level cross-validation (see Section 2). Middle panel: this performance is robust over a range of values for the parameter K controlling the number of states, with STRM-Regression displaying no significant difference from synchronous models for all values of K tested (ANOVA, $p = .98$); this plot does further identify a performance tuning curve that justifies the use of optimisation through cross-validation. Lower panel: Performance of STRM-Regression against a range of different decoding models, both generative (LGS—in both synchronous and optimised sliding window modes) and discriminative (linear regression and Support vector machine regression, using linear and radial basis function kernels). The Pearson correlation shows no significant difference between groups (ANOVA; $p = .19$). While not significant, the trend is the same as obtained for STRM-classification: STRM-Regression is broadly consistent in performance with its timepoint-by-timepoint LGS equivalent, however optimised sliding window methods are trending toward superior performance than STRM-Regression. There is no evidence that discriminative models (Linear Regression and SVM regression) in general outperform generative models (LGS and STRM), with the results here trending in the other direction.

3.2.2 | Spatially resolved stimulus activation maps

Each trial within the foraging task consisted of participants being presented with an amount of reward they had just accumulated (see Section 2.6.1 and Figure 9 for full task outline). We first set out to decode the amount of reward presented on each trial and determine the origins of the activation patterns that predict it.

We fit STRM-Regression models, independently for each subject and session, to one-second epochs of EEG data following presentation of the amount of reward the participant had received. This allowed us to identify clear sensor space spatial topographies associated with each regressor in the STRM-Regression design matrix, with

the first regressor representing a common mean pattern for each state over all trials and a second representing the value of the reward received. As shown in Figure 8, these maps identify patterns emerging initially over medial parietal regions in sensor space and propagating forward over successive states. The spatial location of the value signal is consistent with the literature on value encoding, which is associated with encoding initially in the superior parietal cortex and later in the medial prefrontal cortex (Hunt et al., 2012; Kolling, Behrens, Mars, & Rushworth, 2012). While the sensor space maps in Figure 8 do not afford the same precision in terms of anatomical interpretation as achieved in the visual MEG source space maps of Figure 5 (note that we did not perform source reconstruction on the EEG data due to the

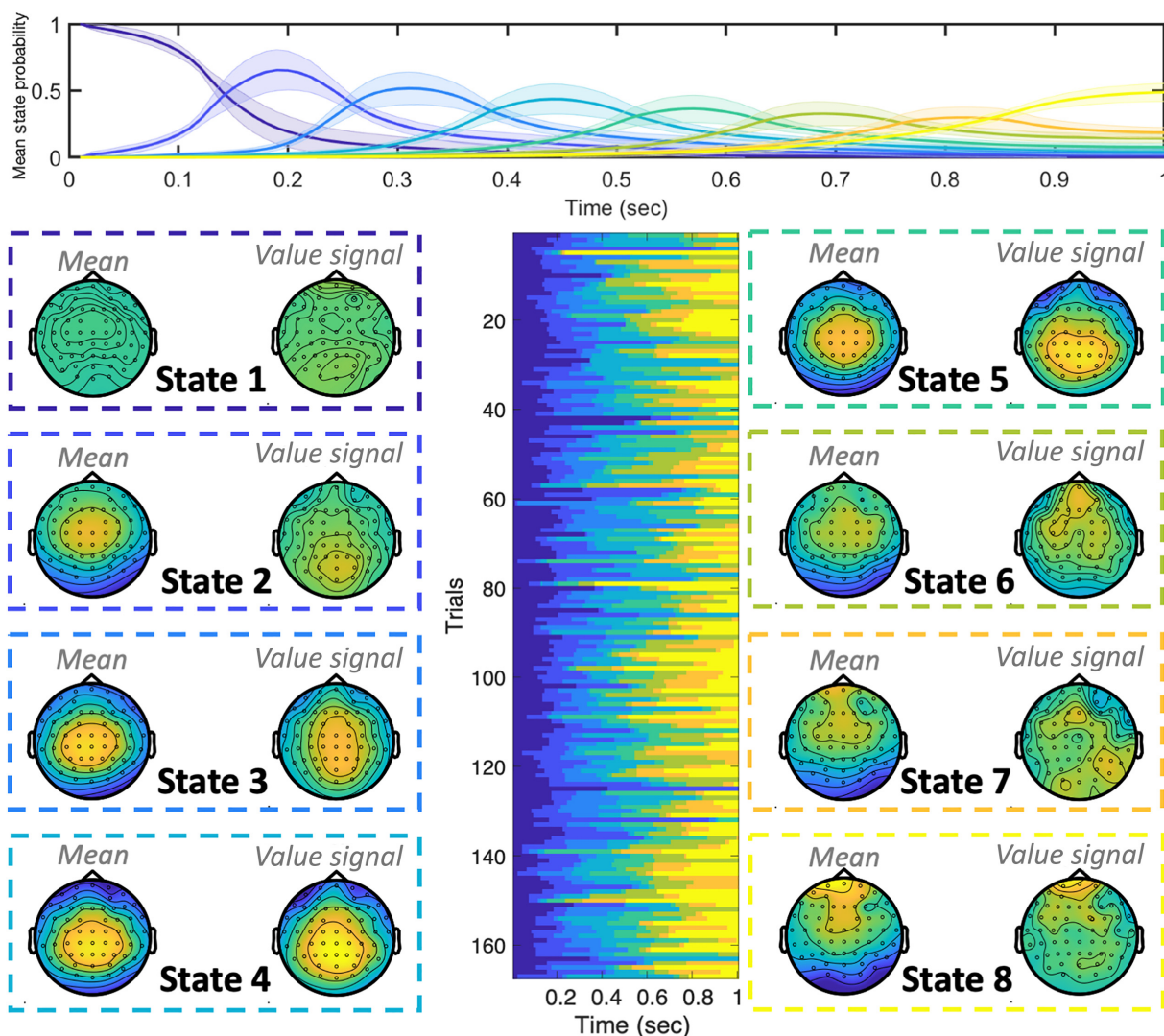


FIGURE 8 The stages of EEG Value Processing. Fitting the STRM-Regression model independently to each subject's EEG data with $K = 8$ states identified a consistent pattern of sequential activity on each trial, but with significant variation in the timing of events on individual trials. Top panel: mean state timecourse \pm SE over subjects. Lower centre panel: example state timecourses for one subject exhibiting significant variation over trials. Lower outer panels: Mean (over subjects) activation patterns for each state and each regressor. Each trial is associated with consistent patterns of activity, comprising a mean pattern of activation common to that stage of cognitive processing on all trials and a separate value-specific component. Both components are characterised by medial parietal activation; in the case of the value signal this appears to emerge initially in parietal areas (e.g., state 2) and later move to more frontal regions (state 6). In response to concerns about eye movement artefacts accounting for significant decode-ability, none of these topographies appears eye movement related

moderate variability in cap and electrode positioning across subjects), they nonetheless suggest an anatomically derived interpretation: we can loosely interpret states 1–4 as representing the stages of value processing concentrated on parietal areas, whereas states 5–8 represent the emergence of a value processing state more concentrated on medial frontal regions.

One common concern using decoding in EEG is whether the prominent artefacts associated with eye movements may confound the decoding accuracy. To make neuroscientific statements, experimentalists must be assured that the underlying source of their decoders' performance is neural and not some muscle or ocular confound. Common decoding techniques are unable to indicate the role that such artefacts may have played in the predictions made by the decoder; while it is common to provide a post hoc analysis of ERPs to refute such claims, this evidence can only ever be indirect as it does not explain how the classifiers made the predictions. In contrast, the predictions made by this method are a measure of how closely the spatial patterns match the parameters shown in Figure 8; the clear midline activity patterns that these maps show are therefore direct

evidence to refute claims of muscle confounds and eye motion artefacts driving the decoding result.

3.2.3 | Timing of value processing is modulated by cognitive variables

Effective completion of the foraging task requires participants to reconcile the amount of reward they receive on each trial with (a) the “accrual time,” that is, the amount of time they had invested at the site in order to accrue that much reward, and (b) what they would currently expect to receive if they moved to a different patch. For example, a given stimulus has a different intrinsic value if the participant had spent a particularly long time to receive it; or if the participant expected to accrue reward more quickly elsewhere. These two variables we will refer to as cognitive variables, as they are not explicitly presented to the participant on screen as a stimulus but must be estimated and tracked alongside the completion of the task.

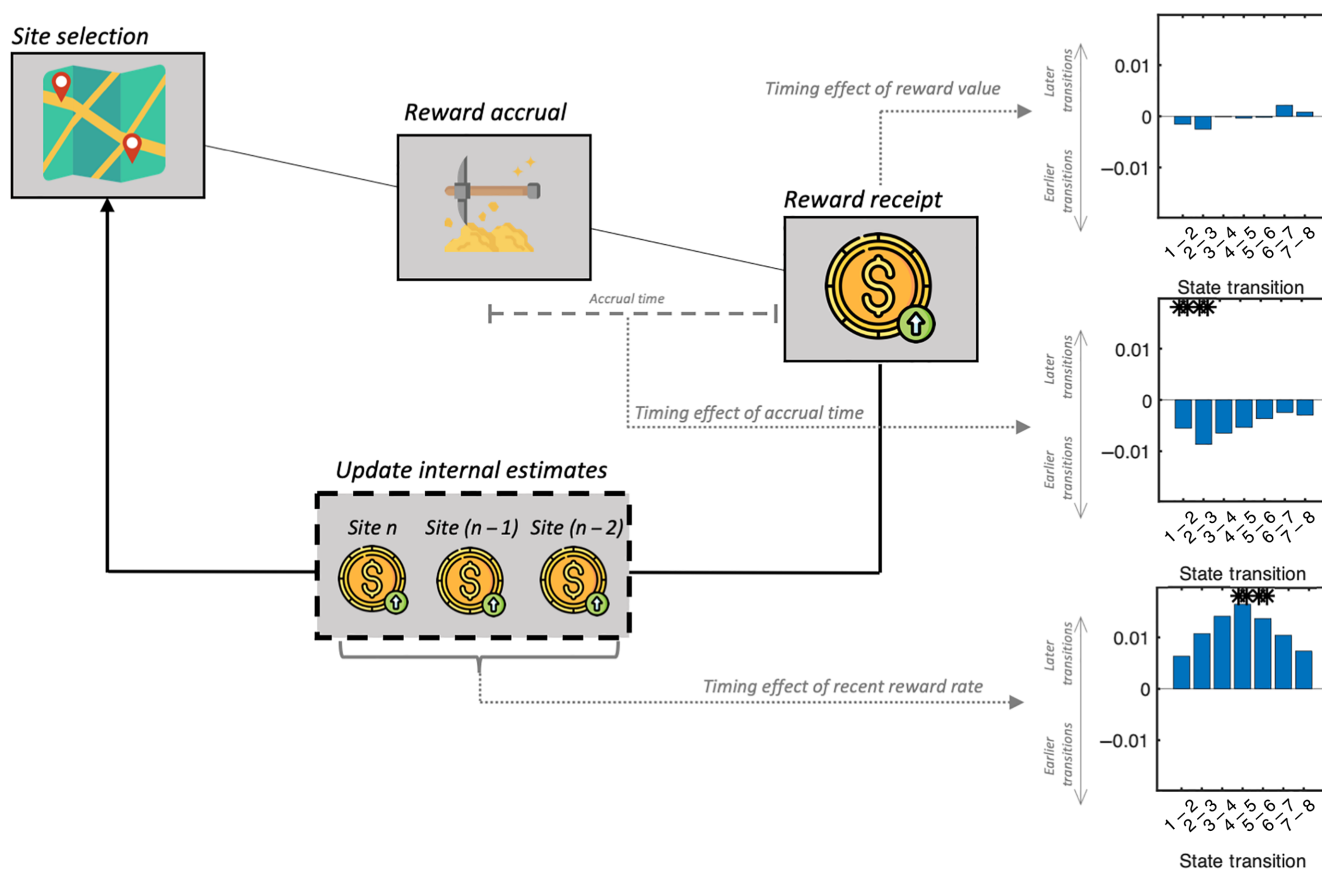


FIGURE 9 The timing of value processing is modulated by cognitive variables. We investigated whether latent cognitive variables within the overall task structure significantly affected the timing at which different stages of value processing emerged on different trials. Specifically, we asked whether three regressors—the overall reward accumulated, the accrual time, and the recent reward rate—could predict the transition times between states on individual trials. We found no significant effect of reward accumulated, the variable being decoded, however we found significant effects of accrual times (reflecting how much time the participant invested to obtain the reward) and recent reward history (reflecting the expected value of alternative sites). Longer accrual times were associated with more rapid transitions through early stimulus processing states, whereas high rewards in recent history were associated with delayed transitions into intermediate stimulus processing states

We asked whether these cognitive variables influence the timing at which value is processed on individual trials. We investigated this by fitting a multiple regression model to predict the timings of the transitions between different states using the state timecourses inferred by STRM. This multiple regression had three regressors: accumulated reward, accrual time and recent reward rate, as shown in Figure 9. We found no significant relationship between value (the regressor being decoded) and state transition times at Bonferroni corrected levels. We did find a strong relationship between accrual time and state progression, with longer times invested at the site associated with more rapid transitions into states 2 and 3, the very earliest stages of stimulus processing in parietal cortical areas. We similarly found a relationship with the recent reward history: on trials where much more reward had been received recently, the onset of intermediate stages of stimulus processing (states 5 and 6 specifically) were delayed. Successful completion of this task requires participants to discount the value of a stimulus against how much time they invested to receive it and how it compares to recent trials; these results suggest these two comparisons are undertaken at different times and specifically modulate the timing of distinct stimulus processing states.

Once again, these results serve fundamentally as model validation; they demonstrate that the inferred timing of value processing on individual trials varies systematically with cognitive variables that reflect key parameters within the task. Notably, again this timing is modulated differentially by the different regressors—that is, early stages of stimulus processing are influenced by accrual times whereas intermediate stages of stimulus processing are modulated by the recent reward rate. Viewed alongside the anatomical maps of each state, these differential effects could suggest distinct variables modulating the parietal and prefrontal components of value processing.

4 | DISCUSSION

We have introduced a method for multivariate analysis of task evoked responses that identifies spatially and temporally resolved patterns associated with trial stimuli. We claim that such an approach supports a more direct interpretation of decoding accuracy metrics by linking the classifier predictions to an interpretable linear model of the underlying activity patterns. By resolving these in time, we are able to utilise the high temporal resolution of M/EEG to investigate how the precise timing of activity patterns is modulated over different trials.

Conventional methods that assume the same process occurs at the same millisecond on every trial are not just limiting data analysis, but also constraining experimental design. Researchers are currently limited to repetitive paradigms designed to have maximally reproducible timing over trials. We have shown that relaxing this assumption with STRM models reveals meaningful patterns of temporal variation over trials. Ultimately, this new modelling approach could open the door to more flexible experimental designs, allowing tasks more dependent on higher order cognitive functions that have greater variability in their onset timing.

In a closely related line of work, Anderson & Fincham (2014a) have shown that such timing variability can be highly informative about how participants actually complete a task (Anderson et al., 2010, 2012; Anderson & Fincham, 2014b; Borst et al., 2016; Borst & Anderson, 2015). Their methods overlap with our own, with the key difference that they use an HMM with an *unsupervised* observation model. This means a model where the likelihood of the data is independent of any stimulus information (i.e., the design matrix X_t). A range of similar methods exist outside the HMM framework which have all demonstrated that—in certain tasks—unsupervised patterns can be informative of the task structure itself (Sakoğlu et al., 2010; Sporns, Faskowitz, Teixeira, Cutts, & Betzel, 2021; Wu, Caprihan, & Calhoun, 2021). In contrast, our *supervised* model applies the GLM framework such that the likelihood of the data is conditioned upon this stimulus information (as specified in Equation (2)). This is a very significant difference in practice; unsupervised models may detect large-scale fluctuations in distributed activity—such as that between active perception and motor response phases of a task (Borst & Anderson, 2015)—but will have limited sensitivity to more subtle differences, such as those that are evoked by different visual images. Thus, we argue that our methods build upon these by generalising the study of inter-trial timing differences to a much broader range of experimental paradigms.

The approach of fitting an interpretable encoding model to training data and then performing a model inversion to make predictions about unseen data is well established in the fMRI literature (Casey et al., 2011; Friston et al., 2008; Kay et al., 2008; Mitchell et al., 2008; Naselaris et al., 2009, 2015; Nishimoto et al., 2011; Schoenmakers et al., 2013), but has only seen limited adoption so far in M/EEG (di Liberto et al., 2015; Kupers et al., 2020). Our focus in this article has been to emphasise the interpretability benefits of this approach—which are often overlooked—and demonstrate how it can be readily extended to time series analysis for data at high temporal resolution in ways that we believe offer significant benefits to conventional timepoint-by-timepoint decoding.

Trujillo-Barreto et al. (2008) used a very similar GLM model inversion to source localise M/EEG activity. Their model applies two hierarchical GLMs, with the first modelling the source-level data as an explicit function of an experimentally defined design matrix, and the second modelling the propagation of activity from source to sensor space. Broadly speaking, we anticipate the main benefit of their approach to be greater spatial resolution of source related activity, whereas ours by comparison offers enhanced temporal resolution through the explicit modelling of HMM states. This highlights a potential avenue for future research whereby these models could be combined to take advantage of both these benefits, with the HMM-state specific covariance modelling contributing to greater accuracy of source-localised activity estimates (Woolrich et al., 2013).

Furthermore, we demonstrate that this generative model approach to decoding—with or without the HMM component—achieves equivalent or better accuracies than discriminative classifiers on two different datasets. This corroborates limited evidence from similar surveys of classifiers in M/EEG data (Grootswagers

et al., 2017; Guggenmos, Sterzer, & Cichy, 2018), suggesting this may be a common feature of this data and the model we propose may generalise well.

Nonetheless, there are a few trade-offs to consider when using the model we have presented. While the observed predictive accuracy was fairly robust across a range of different parameter values, in both datasets any performance gains could be recovered or surpassed by using optimised sliding window techniques, for example, sliding window LGS in Figure 7 (see section 2 in Supporting Information). There are several possible reasons for this. On the one hand, when computing cross-validated accuracy our methods do not allow a direct way to obtain corresponding state timecourses for the held-out test set, instead relying on an estimation procedure that may introduce additional error (see Section 2.4). Nonetheless, we did attempt a number of different approaches to estimating state timecourses for held-out data, none of which consistently outperformed the optimised sliding window techniques (see section 2 in Supporting Information), suggesting that the subtle changes in activity pattern timing are not a major determinant of accuracy. Another potential issue with our model is the assumption of mutual exclusivity between states, which produces sharp jumps in time between inferred activity patterns. If the underlying activity patterns are in fact smooth, then their arbitrary discretization introduces larger errors for timepoints adjacent to state boundaries. On the one hand, this approach is highly interpretable and supports relatively straightforward analyses of the impacts of different behavioural variables on state transition times; on the other hand, this may come at a cost to predictive accuracy around state boundaries. We furthermore expect, given the challenge of overfitting with these datasets that motivated our left-right state constraint, that any comparable dynamic models with a *continuous* state space (e.g., Penny & Roberts, 1999) would be very difficult to suitably regularise.

The STRM model requires a hyperparameter specifying the number of states. As we have shown, this hyperparameter can be optimised through cross-validation, finding values that maximise predictive accuracy. Nonetheless, the different parameter values for each subject make group comparison difficult, such that in some cases this parameter must be assigned arbitrarily. Researchers concerned about such an arbitrary selection could perhaps base their decision upon the temporal generalisation profile of their data (King & Dehaene, 2014)—as a rule of thumb we might suggest matching the expected state duration (i.e., $\frac{T}{K}$) to the minimum width of the diagonal pattern observed in such temporal generalisation matrices.

Our work is very closely related to a number of others; notably a seminal work which showed how any decoding model could be inverted to recover an interpretable forward model of the original data (Haufe et al., 2014). We instead propose the opposite approach; to fit an encoding model of the data and the invert that to make predictions. How are these two approaches different in practice? The first point we would make pertains to the classification accuracies we have reported; when using generative classifiers these are generally either better or at least not significantly different than a range of discriminative models tested, a finding that appears consistent with other

examples in the literature (Grootswagers et al., 2017; Guggenmos et al., 2018). Note that all of these discriminative classifiers should, given enough datapoints, converge to the same boundary; generative classifiers however can learn faster if their modelling assumptions are accurate (Ng & Jordan, 2002). This ultimately equates to greater estimation error in the inverted model parameters; that is, the error in forward model parameters obtained via a model inversion is likely to be greater than the error in a directly fitted forward model. We can demonstrate this using simulations (see section 4 in Supporting Information). While the improvement obtained is modest when using a directly fitted forward model compared with model inversions, we can also demonstrate this improvement is not exclusively limited to scenarios where generative classifiers achieve greater classification accuracy than discriminative classifiers (see section 4 in Supporting Information). Furthermore, the interpretation of an inverted decoding model can be problematic wherever regularisation is used (Haufe et al., 2014). Given the ubiquity of regularisation in decoding applications this is not a niche problem. Thus, where interpretation of forward model parameters is the goal, we would argue that one should just fit a forward model directly.

Finally, our work was originally motivated by that of (Vidaurre et al., 2019), which introduced the HMM architecture in the context of decoding models. This demonstrated the potential of time series models for multivariate analysis, but left open the question of model interpretability. Similar to the question we posed in Figure 1, if each state is a different spatial filter, then how should one interpret observed differences in timing on individual trials? Given the lack of direct interpretability of spatial filters themselves, this question does not present an obvious answer. In contrast, by setting up our work around a forward encoding model based on the widely used GLM framework, each state has a clear correspondence to a set of stimulus activation patterns. This affords a straightforward interpretation of each state as a successive stage of stimulus processing, allowing us to then explore its spatial and temporal properties and provide a richer picture of the overall patterns of activation.

5 | CONCLUSION

Neuroscientists want to understand what information is represented in brain activity, as well as how and when it is expressed. Conventional methods limit what researchers can investigate in several ways. By obscuring the spatial patterns from which decoding accuracy metrics are derived, researchers are often left to interpret a result without clear knowledge of its spatial origins. By assuming the same process occurs at the same millisecond on every trial, researchers are unable to investigate meaningful patterns of temporal variation over trials and are limited in the experimental designs they can pursue.

The STRM model addresses these points with two main innovations; firstly, through the use of an interpretable encoding model to reveal how activity patterns are spatially distributed, and secondly through the use of an HMM to reveal how activity patterns are

temporally distributed. M/EEG recordings offer the potential to reveal how the brain represents information at millisecond resolution; the methods we have developed seek to leverage this resolution to simultaneously answer the question of *when* and *where* these patterns emerge. By characterising both the spatial and temporal characteristics of neural representations, we may obtain a more holistic understanding of brain function, offer a new perspective on the role of timing in cognitive processes, and support more flexible experimental designs in the future.

ACKNOWLEDGMENTS

The Wellcome Centre for Integrative Neuroimaging is supported by core funding from the Wellcome Trust (203139/Z/16/Z). MW's research is supported by the NIHR Oxford Health Biomedical Research Centre, the Wellcome Trust (106183/Z/14/Z, 215573/Z/19/Z), the New Therapeutics in Alzheimer's Diseases (NTAD) study supported by UK MRC and the Dementia Platform UK (RG94383/RG89702) and the EU-Project euSNN (MSCA-ITN H2020-860563). TB was supported by Wellcome Trust senior (104765/Z/14/Z) and principal (219525/Z/19/Z) research fellowships, a Wellcome collaborator award (214314/Z/18/Z) and by the James S. McDonnell Foundation (JSMF220020372) during the course of the research. YL was supported by the Max Planck Society. NK's research is funded by the BBSRC fellowship, BBSRCAFL Fellowship (BB/R010803/1). DV is supported by a Novo Nordisk Emerging Investigator Award (NNF19OC-0054895) and by the European Research Council (ERC-StG-2019-850404). CH is supported by the Wellcome Trust (215573/Z/19/Z).

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

AUTHOR CONTRIBUTIONS

Cameron Higgins: Formal analysis, investigation, methodology, software, validation, visualisation, writing – original draft preparation. **Diego Vidaurre:** Conceptualization, methodology, software, supervision, writing – review and editing. **Nils Kolling:** Data curation, investigation, writing – review and editing. **Yunzhe Liu:** Data curation, writing – review and editing. **Tim Behrens:** Conceptualization, funding acquisition, supervision, writing – review and editing. **Mark Woolrich:** Conceptualization, funding acquisition, methodology, project administration, supervision, writing – review and editing.

DATA AVAILABILITY STATEMENT

All software code needed to fit the model and run the analyses in this paper is available online at <https://github.com/OHBA-analysis/HMM-MAR>. This is a large repository; the script to complete the specific analyses in this paper is located at: <https://github.com/OHBA-analysis/HMM-MAR/blob/master/examples/STRM.m>. The two datasets presented in this paper (MEG Visual Stimuli and EEG value based decision making) are publicly available on Mendeley Data at the doi: 10.17632/jwjkzsg4dx.1.

ORCID

Cameron Higgins  <https://orcid.org/0000-0002-0051-5325>
 Diego Vidaurre  <https://orcid.org/0000-0002-9650-2229>
 Nils Kolling  <https://orcid.org/0000-0002-0283-6966>
 Yunzhe Liu  <https://orcid.org/0000-0003-0836-9403>
 Tim Behrens  <https://orcid.org/0000-0003-0048-1177>
 Mark Woolrich  <https://orcid.org/0000-0001-8460-8854>

REFERENCES

- Anderson, J. R., Betts, S., Ferris, J. L., & Fincham, J. M. (2010). Neural imaging to track mental states while using an intelligent tutoring system. *Proceedings of the National Academy of Sciences of the United States of America*, 107(15), 7018–7023. <https://doi.org/10.1073/pnas.1000942107>
- Anderson, J. R., & Fincham, J. M. (2014a). Discovering the sequential structure of thought. *Cognitive Science*, 38(2), 322–352. <https://doi.org/10.1111/cogs.12068>
- Anderson, J. R., & Fincham, J. M. (2014b). Extending problem-solving procedures through reflection. *Cognitive Psychology*, 74, 1–34. <https://doi.org/10.1016/j.cogpsych.2014.06.002>
- Anderson, J. R., Fincham, J. M., Schneider, D. W., & Yang, J. (2012). Using brain imaging to track problem solving in a complex state space. *NeuroImage*, 60(1), 633–643. <https://doi.org/10.1016/j.neuroimage.2011.12.025>
- Beal, M. J. (2003). Variational algorithms for approximate bayesian inference (PhD Thesis [Issue May]). Gatsby Computational Neuroscience Unit, University College London.
- Borst, J. P., & Anderson, J. R. (2015). The discovery of processing stages: Analyzing EEG data with hidden semi-Markov models. *NeuroImage*, 108(1), 60–73. <https://doi.org/10.1016/j.neuroimage.2014.12.029>
- Borst, J. P., Ghuman, A. S., & Anderson, J. R. (2016). Tracking cognitive processing stages with MEG: A spatio-temporal model of associative recognition in the brain. *NeuroImage*, 141, 416–430. <https://doi.org/10.1016/j.neuroimage.2016.08.002>
- Carlson, T., Tovar, D. A., Alink, A., & Kriegeskorte, N. (2013). Representational dynamics of object vision: The first 1000 ms. *Journal of Vision*, 13(10), 1. <https://doi.org/10.1167/13.10.1>
- Carlson, T. A., Hogendoorn, H., Kanai, R., Mesik, J., & Turret, J. (2011). High temporal resolution decoding of object position and category. *Journal of Vision*, 11(10), 1–17. <https://doi.org/10.1167/11.10.9.Introduction>
- Casey, M., Thompson, J., Kang, O., Raizada, R., & Wheatley, T. (2011). Population codes representing musical timbre for high-level fMRI categorization of music genres. In: Proceedings of the 1st international conference on machine learning and interpretation in neuroimaging, pp. 34–41. https://doi.org/10.1007/978-3-642-34713-9_5
- Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature Neuroscience*, 17(3), 455–462. <https://doi.org/10.1038/nn.3635>
- Colclough, G. L., Brookes, M. J., Smith, S. M., & Woolrich, M. W. (2015). A symmetric multivariate leakage correction for MEG connectomes. *NeuroImage*, 117, 439–448. <https://doi.org/10.1016/j.neuroimage.2015.03.071>
- Colclough, G. L., Smith, S. M., Nichols, T. E., Winkler, A. M., Sotiropoulos, S. N., Glasser, M. F., ... Woolrich, M. W. (2017). The heritability of multi-modal connectivity in human brain activity. *eLife*, 6, 1–19. <https://doi.org/10.7554/eLife.20178>
- Colclough, G. L., Woolrich, M. W., Tewarie, P. K., Brookes, M. J., Quinn, A. J., & Smith, S. M. (2016). How reliable are MEG resting-state connectivity metrics? *NeuroImage*, 138, 284–293. <https://doi.org/10.1016/j.neuroimage.2016.05.070>
- di Liberto, G. M., O'Sullivan, J. A., & Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current*

- Biology, 25(19), 2457–2465. <https://doi.org/10.1016/j.cub.2015.08.030>
- Friston, K., Chu, C., Mourão-Miranda, J., Hulme, O., Rees, G., Penny, W., & Ashburner, J. (2008). Bayesian decoding of brain images. *NeuroImage*, 39(1), 181–205. <https://doi.org/10.1016/j.neuroimage.2007.08.013>
- Ghahramani, Z. (2001). An introduction to hidden Markov models and Bayesian networks. In H. Bunke & T. Caelli (Eds.), *Hidden Markov models* (Vol. 45, pp. 9–41). River Edge, NJ: World Scientific.
- Goodale, M. A. (1993). Visual pathways supporting perception and action in the primate cerebral cortex. *Current Opinion in Neurobiology*, 3(4), 578–585. [https://doi.org/10.1016/0959-4388\(93\)90059-8](https://doi.org/10.1016/0959-4388(93)90059-8)
- Goodale, M. A., & Milner, D. A. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1), 20–25. <https://doi.org/10.1093/litthe/11.1.80>
- Grootswagers, T., Wardle, S. G., & Carlson, T. A. (2017). Decoding dynamic brain patterns from evoked responses: A tutorial on multivariate pattern analysis applied to time series neuroimaging data. *Journal of Cognitive Neuroscience*, 29(4), 677–697. <https://doi.org/10.1162/jocn>
- Guggenmos, M., Sterzer, P., & Cichy, R. M. (2018). Multivariate pattern analysis for MEG: A comparison of dissimilarity measures. *NeuroImage*, 173, 434–447. <https://doi.org/10.1016/j.neuroimage.2018.02.044>
- Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J. D., Blankertz, B., & Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, 87, 96–110. <https://doi.org/10.1016/j.neuroimage.2013.10.067>
- Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding neural representational spaces using multivariate pattern analysis. *Annual Review of Neuroscience*, 37, 435–456. <https://doi.org/10.1146/annurev-neuro-062012-170325>
- Haynes, J.-D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7(7), 523–534. <https://doi.org/10.1038/nrn1931>
- Higgins, C. (2019). Uncovering temporal structure in neural data with statistical machine learning models (Doctoral Thesis). University of Oxford.
- Higgins, C., Liu, Y., Vidaurre, D., Kurth-Nelson, Z., Dolan, R., Behrens, T., & Woolrich, M. (2021). Replay bursts in humans coincide with activation of the default mode and parietal alpha networks. *Neuron*, 109(5), 882–893. <https://doi.org/10.1016/j.neuron.2020.12.007>
- Higgins, C., van Es, M. W. J., Quinn, A., Vidaurre, D., & Woolrich, M. (2022). The relationship between frequency content and representational dynamics in the decoding of neurophysiological data. *BioRxiv Preprints*, 1–49. <https://doi.org/10.1101/2022.02.07.479399>
- Hochstein, S., & Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*, 36(5), 791–804. [https://doi.org/10.1016/S0896-6273\(02\)01091-7](https://doi.org/10.1016/S0896-6273(02)01091-7)
- Holdgraf, C. R., Rieger, J. W., Micheli, C., Martin, S., Knight, R. T., & Theunissen, F. E. (2017). Encoding and decoding models in cognitive electrophysiology. *Frontiers in Systems Neuroscience*, 11, 1–24. <https://doi.org/10.3389/fnsys.2017.00061>
- Hunt, L. T., Kolling, N., Soltani, A., Woolrich, M. W., Rushworth, M. F. S., & Behrens, T. E. J. (2012). Mechanisms underlying cortical activity during value-guided choice. *Nature Neuroscience*, 15(3), 470–476. <https://doi.org/10.1038/nn.3017>
- Jensen, O., & Mazaheri, A. (2010). Shaping functional architecture by oscillatory alpha activity: Gating by inhibition. *Frontiers in Human Neuroscience*, 4, 186. <https://doi.org/10.3389/fnhum.2010.00186>
- Johnson, M. (2007). Why doesn't EM find good HMM POS-taggers? In: Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning, June, pp. 296–305.
- Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, 452(7185), 352–355. <https://doi.org/10.1038/nature06713>
- King, J. R., & Dehaene, S. (2014). Characterizing the dynamics of mental representations: The temporal generalization method. *Trends in Cognitive Sciences*, 18(4), 203–210. <https://doi.org/10.1016/j.tics.2014.01.002>
- Kolling, N., Behrens, T. E. J., Mars, R. B., & Rushworth, M. F. S. (2012). Neural mechanisms of foraging. *Science*, 335(6077), 95–98. <https://doi.org/10.1126/science.1216930>
- Kriegeskorte, N., & Douglas, P. K. (2019). Interpreting encoding and decoding models. *Current Opinion in Neurobiology*, 55, 167–179. <https://doi.org/10.1016/j.conb.2019.04.002>
- Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, 17(8), 401–412. <https://doi.org/10.1016/j.tics.2013.06.007>
- Kupers, E. R., Benson, N. C., & Winawer, J. (2020). A visual encoding model links magnetoencephalography signals to neural synchrony in human cortex. *BioRxiv Preprints*, 1–40. <https://doi.org/10.1101/2020.04.19.049197>
- Light, G. A., Williams, L. E., Minow, F., Sprock, J., Rissling, A., Sharp, R., ... Bruff, D. L. (2010). Electroencephalography (EEG) and event-related potentials (ERPs) with human participants. *Current Protocols in Neuroscience*, 52(1), 6–25.
- Liu, Y., Dolan, R. J., Kurth-Nelson, Z., & Behrens, T. E. J. (2019). Human replay spontaneously reorganizes experience. *Cell*, 178, 640–652. <https://doi.org/10.1016/j.cell.2019.06.012>
- Miller, K. J., Honey, C. J., Hermes, D., Rao, R. P. N., Den Nijs, M., & Ojemann, J. G. (2014). Broadband changes in the cortical surface potential track activation of functionally diverse neuronal populations. *NeuroImage*, 85, 711–720. <https://doi.org/10.1016/j.neuroimage.2013.08.070>
- Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, 320(May), 1191–1195.
- Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. Cambridge, MA: MIT Press. https://doi.org/10.1007/978-94-011-3532-0_2
- Naselaris, T., & Kay, K. N. (2015). Resolving ambiguities of MVPA using explicit models of representation. *Trends in Cognitive Sciences*, 19(10), 551–554. <https://doi.org/10.1016/j.tics.2015.07.005>
- Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage*, 56(2), 400–410. <https://doi.org/10.1016/j.neuroimage.2010.07.073>
- Naselaris, T., Olman, C. A., Stansbury, D. E., Ugurbil, K., & Gallant, J. L. (2015). A voxel-wise encoding model for early visual areas decodes mental images of remembered scenes. *NeuroImage*, 105, 215–228. <https://doi.org/10.1016/j.neuroimage.2014.10.018>
- Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., & Gallant, J. L. (2009). Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63(6), 902–915. <https://doi.org/10.1016/j.neuron.2009.09.006>
- Negrini, M., Brkic, D., Pizzamiglio, S., Premoli, I., & Rivolta, D. (2017). Neurophysiological correlates of featural and spacing processing for face and non-face stimuli. *Frontiers in Psychology*, 8, 1–9. <https://doi.org/10.3389/fpsyg.2017.00333>
- Ng, A. Y., & Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in Neural Information Processing Systems*, 14(841), 1–8. <https://doi.org/10.1007/s11063-008-9088-7>
- Nichols, T., & Holmes, A. (2003). Nonparametric permutation tests for functional neuroimaging. *Human Brain Mapping*, 15, 1–25. <https://doi.org/10.1016/B978-012264841-0/50048-2>
- Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21(19), 1641–1646. <https://doi.org/10.1016/j.cub.2011.08.031>

- Obermaier, B., Guger, C., Neuper, C., & Pfurtscheller, G. (2001). Hidden Markov models for online classification of single trial EEG data. *Pattern Recognition Letters*, 22(12), 1299–1309. [https://doi.org/10.1016/S0167-8655\(01\)00075-7](https://doi.org/10.1016/S0167-8655(01)00075-7)
- Penny, W. D., & Roberts, S. J. (1999). Dynamic logistic regression. *Proceedings of the International Joint Conference on Neural Networks*, 3(5), 1562–1567. <https://doi.org/10.1109/ijcnn.1999.832603>
- Podvalny, E., Flounders, M. W., King, L. E., Holroyd, T., & He, B. J. (2019). A dual role of prestimulus spontaneous neural activity in visual object recognition. *Nature Communications*, 10(1), 1–13. <https://doi.org/10.1038/s41467-019-11877-4>
- Roweis, S., & Ghahramani, Z. (1999). A unifying review of linear Gaussian models. *Neural Computation*, 11(2), 305–345. <https://doi.org/10.1162/089976699300016674>
- Sakoğlu, Ü., Pearlson, G. D., Kiehl, K. A., Wang, Y. M., Michael, A. M., & Calhoun, V. D. (2010). A method for evaluating dynamic functional network connectivity and task-modulation: Application to schizophrenia. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 23(5–6), 351–366. <https://doi.org/10.1007/s10334-010-0197-8>
- Schoenmakers, S., Barth, M., Heskes, T., & van Gerven, M. (2013). Linear reconstruction of perceived images from human brain activity. *NeuroImage*, 83, 951–961. <https://doi.org/10.1016/j.neuroimage.2013.07.043>
- Sederberg, A. J., Pala, A., Zheng, H. J. V., He, B. J., & Stanley, G. B. (2019). State-aware detection of sensory stimuli in the cortex of the awake mouse. *PLoS Computational Biology*, 15(5), e1006716. <https://doi.org/10.1101/499269>
- Sporns, O., Faskowitz, J., Teixeira, A. S., Cutts, S. A., & Betzel, R. F. (2021). Dynamic expression of brain functional systems disclosed by fine-scale analysis of edge time series. *Network Neuroscience*, 5(2), 405–433. <https://doi.org/10.1162/NETN>
- Trujillo-Barreto, N. J., Aubert-Vázquez, E., & Penny, W. D. (2008). Bayesian M/EEG source reconstruction with spatio-temporal priors. *NeuroImage*, 39(1), 318–335. <https://doi.org/10.1016/j.neuroimage.2007.07.062>
- Valentin, S., Harkotte, M., & Popov, T. (2020). Interpreting neural decoding models using grouped model reliance. *PLoS Computational Biology*, 16(1), 1–17. <https://doi.org/10.1371/journal.pcbi.1007148>
- van de Nieuwenhuijzen, M. E., Backus, A. R., Bahramisharif, A., Doeller, C. F., Jensen, O., & van Gerven, M. A. J. (2013). MEG-based decoding of the spatiotemporal dynamics of visual category perception. *NeuroImage*, 83, 1063–1073. <https://doi.org/10.1016/j.neuroimage.2013.07.075>
- van Veen, B. D., van Drongelen, W., & Suzuki, A. (1997). Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. *IEEE Transactions on Biomedical Engineering*, 44(9), 867–880. <https://doi.org/10.1109/10.623056>
- Vidaurre, D., Hunt, L. T., Quinn, A. J., Hunt, B. A. E., Brookes, M. J., Nobre, A. C., & Woolrich, M. W. (2018). Spontaneous cortical activity transiently organises into frequency specific phase-coupling networks. *Nature Communications*, 9(2987), 1–13. <https://doi.org/10.1038/s41467-018-05316-z>
- Vidaurre, D., Myers, N. E., Stokes, M., Nobre, A. C., & Woolrich, M. W. (2019). Temporally unconstrained decoding reveals consistent but time-varying stages of stimulus processing. *Cerebral Cortex*, 29(2), 863–874. <https://doi.org/10.1093/cercor/bhy290>
- Williams, N. J., Daly, I., & Nasuto, S. J. (2018). Markov model-based method to analyse time-varying networks in EEG task-related data. *Frontiers in Computational Neuroscience*, 12, 1–18. <https://doi.org/10.3389/fncom.2018.00076>
- Woolrich, M., Baker, A., Luckhoo, H., Mohseni, H., Barnes, G., Brookes, M., & Rezek, L. (2013). Dynamic state allocation for MEG source reconstruction. *NeuroImage*, 77, 77–92. <https://doi.org/10.1016/j.neuroimage.2013.03.036>
- Woolrich, M., Hunt, L., Groves, A., & Barnes, G. (2011). MEG beamforming using Bayesian PCA for adaptive data covariance matrix regularization. *NeuroImage*, 57(4), 1466–1479. <https://doi.org/10.1016/j.neuroimage.2011.04.041>
- Woolrich, M. W., Jbabdi, S., Patenaude, B., Chappell, M., Makni, S., Behrens, T., ... Smith, S. M. (2009). Bayesian analysis of neuroimaging data in FSL. *NeuroImage*, 45(1), S173–S186. <https://doi.org/10.1016/j.neuroimage.2008.10.055>
- Wu, L., Caprihan, A., & Calhoun, V. (2021). Tracking spatial dynamics of functional connectivity during a task. *NeuroImage*, 239, 118310. <https://doi.org/10.1016/j.neuroimage.2021.118310>
- Yamashita, O., Sato, M. A., Yoshioka, T., Tong, F., & Kamitani, Y. (2008). Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns. *NeuroImage*, 42(4), 1414–1429. <https://doi.org/10.1016/j.neuroimage.2008.05.050>

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Higgins, C., Vidaurre, D., Kolling, N., Liu, Y., Behrens, T., & Woolrich, M. (2022). Spatiotemporally resolved multivariate pattern analysis for M/EEG. *Human Brain Mapping*, 43(10), 3062–3085. <https://doi.org/10.1002/hbm.25835>

APPENDIX A

A.1 | STRM-Classification Bayesian model inversion

We wish to infer the posterior distribution given by:

$$P(X_t|Y_t, \tilde{\theta}, \tilde{Z}) = \frac{P(Y_t, \tilde{\theta}, \tilde{Z}|X_t)P(X_t)}{P(Y_t, \tilde{\theta}, \tilde{Z})}$$

where $\tilde{\theta} = \{\tilde{W}_{1:Q,1:K}, \tilde{\Sigma}_{1:K}\}$ are the maximum a posteriori point estimates learned during training, and \tilde{Z} are the state timecourse probability estimates as outlined in the text. We now introduce the lower case notation $x_t = i$ to denote class i being active, such that the $[1 \times Q]$ vector $X_t = [I(x_t = 1), I(x_t = 2), \dots, I(x_t = K)]$ where I is the Iverson bracket returning 1 if the statement is true and 0 if false. We then have:

$$\begin{aligned} \log P(x_t = i|Y_t, \tilde{\theta}, \tilde{Z}) &= \sum_{k=1}^K \frac{-\tilde{\gamma}_{t,k}}{2} \log(2\pi|\tilde{\Sigma}_k|) \\ &\quad + \sum_{k=1}^K \frac{-\tilde{\gamma}_{t,k}}{2} (Y_t - I_i \tilde{W}_k)^T \tilde{\Sigma}_k^{-1} (Y_t - I_i \tilde{W}_k) - \log Q \\ &\quad - \log P(Y_t, \tilde{\theta}, \tilde{Z}) \\ &= \sum_{k=1}^K \frac{-\tilde{\gamma}_{t,k}}{2} (Y_t - I_i \tilde{W}_k)^T \tilde{\Sigma}_k^{-1} (Y_t - I_i \tilde{W}_k) + C \end{aligned}$$

where the term C is constant with respect to the class i , and I_i denotes the $[1 \times Q]$ vector obtained by taking the i th row of the identity matrix of dimension Q . As the classes are mutually exclusive and exhaustive, we have

$$\begin{aligned} \sum_{i=1}^Q P(x_t = i|Y_t, \tilde{\theta}, \tilde{Z}) &= 1 = e^C \sum_{i=1}^Q e^{L_i} \\ e^C &= \frac{1}{\sum_{i=1}^Q e^{L_i}} \end{aligned}$$

where L is a $[Q \times 1]$ vector of unnormalized class log-likelihoods with each entry $L_i = \sum_{k=1}^K \frac{-\tilde{\gamma}_{t,k}}{2} (Y_t - I_i \tilde{W}_k)^T \tilde{\Sigma}_k^{-1} (Y_t - I_i \tilde{W}_k)$. It, therefore, follows that

$$P(x_t = i|Y_t, \tilde{\theta}, \tilde{Z}) = \sigma_i(L).$$

Or in vector notation:

$$P(X_t|Y_t, \tilde{\theta}, \tilde{Z}) = X_t \sigma(L)$$

where $\sigma(L)$ is the softmax function which outputs a $[Q \times 1]$ vector with each entry $\sigma_i(L) = \frac{e^{L_i}}{\sum_{j=1}^Q e^{L_j}}$.

APPENDIX B

B.1 | STRM-Regression Bayesian model inversion

We wish to infer the posterior distribution given by:

$$P(X_t|Y_t, \tilde{\theta}, \tilde{Z}) = \frac{P(Y_t, \tilde{\theta}, \tilde{Z}|X_t)P(X_t)}{P(Y_t, \tilde{\theta}, \tilde{Z})}$$

where $P(X_t) = N(0, I_Q)$ and $\tilde{\theta} = \{\tilde{W}_{1:Q,1:K}, \tilde{\Sigma}_{1:K}\}$ are the maximum a posteriori point estimates learned during training, and $\tilde{Z}_t = [\tilde{z}_{t,1}, \tilde{z}_{t,2}, \dots, \tilde{z}_{t,K}]$ are the state probabilities estimated as outlined in the text. We then have:

$$\begin{aligned} \log P(X_t|Y_t, \tilde{\theta}, \tilde{Z}) &= \sum_{k=1}^K \frac{-\tilde{\gamma}_{t,k}}{2} \log(2\pi|\tilde{\Sigma}_k|) \\ &\quad + \sum_{k=1}^K \frac{-\tilde{\gamma}_{t,k}}{2} (Y_t - X_t \tilde{W}_k)^T \tilde{\Sigma}_k^{-1} (Y_t - X_t \tilde{W}_k) - \frac{Q}{2} \log 2\pi \\ &\quad - \frac{1}{2} X_t X_t^T - \log P(Y_t, \tilde{\theta}, \tilde{Z}) \end{aligned}$$

If we make the below substitution:

$$\begin{aligned} \Sigma_{x|y} &= \left[I_Q + \sum_{k=1}^K \tilde{\gamma}_{t,k} \tilde{W}_k \tilde{\Sigma}_k^{-1} \tilde{W}_k^T \right]^{-1} \\ \mu_{x|y} &= \Sigma_{x|y} \left[\sum_{k=1}^K \tilde{z}_{t,k} \tilde{W}_k \tilde{\Sigma}_k^{-1} Y_t^T \right] \end{aligned}$$

Then the posterior can equivalently be written:

$$\log P(X_t|Y_t, \tilde{\theta}, \tilde{Z}) = -\frac{1}{2} (X_t - \mu_{x|y})^T \Sigma_{x|y}^{-1} (X_t - \mu_{x|y}) + C$$

where the term C is constant with respect to X_t . This form can be recognised as a normal distribution, equivalently written:

$$P(X_t|Y_t, \tilde{\theta}, \tilde{Z}) = N(\mu_{x|y}, \Sigma_{x|y})$$