Predicting individual traits from models of brain dynamics accurately and reliably using the Fisher kernel

Ahrends C[1] & Vidaurre D[1,2]

[1] Center of Functionally Integrative Neuroscience, Department of Clinical Medicine, Aarhus University, Aarhus, Denmark

[2] Department of Psychiatry, University of Oxford, Oxford, United Kingdom

Abstract

How specifically brain activity unfolds across time, namely the nature of brain dynamics, can sometimes be more predictive of behavioural and cognitive subject traits than both brain structure and summary measures of brain activity that average across time. Brain dynamics can be described by models of varying complexity, but what is the best way to use these models of brain dynamics for characterising subject differences and predicting individual traits is unclear. While most studies aiming at predicting subjects' traits focus on having accurate predictions, for many practical applications, it is critical for the predictions not just to be accurate but also reliable. Kernel methods are a robust and computationally efficient way of expressing differences in brain dynamics between subjects for the sake of predicting individual traits, such as clinical or psychological variables. Using the Hidden Markov model (HMM) as a model of brain dynamics, we here propose the use of the Fisher kernel, a mathematically principled approach to predict phenotypes from subject-specific HMMs. The Fisher kernel is computed in a way that preserves the mathematical structure of the HMM. This results in benefits in terms of both accuracy and reliability when compared with kernels that ignore the structure of the underlying model of brain dynamics.

1 Introduction

Predicting subject traits —for example clinical or psychological— from brain data is an important endeavour for practical purposes such as the developing of biomarkers, and also as a tool for scientific interpretation. In both cases, it is important that the predictions are accurate as well as reliable. With both criteria in mind, this paper addresses the unresolved question of how to predict subject traits from brain dynamics, i.e., using information of how brain activity fluctuates over time during unconstrained cognition —that, is during wakeful rest. These dynamic patterns can be estimated from brain recordings such as fMRI or MEG using models of varying complexity (Lurie et al., 2019). We here focus on models of time-varying parameters that capture within-session changes in functional connectivity and amplitude, and how these can be used to predict behavioural variables or traits. In particular, we use the Hidden Markov Model (HMM) as a model of time-varying amplitude and functional connectivity (FC) dynamics (Vidaurre et al., 2017), which was shown previously to be able to predict some subjects' traits, such as intelligence, more accurately than structural or static (time-averaged) FC representations (Vidaurre et al., 2021).

For many prediction problems, simple linear predictive models are preferred because of their computational efficiency and good performance. However, models of brain dynamics in unconstrained cognition do not directly lend themselves to being used as linear predictors. This is because, in order to describe the complex space-time patterns present in the data, models of brain dynamics are usually parametrised by a high number of parameters, and these parameters have complex relationships between them reflecting the model's assumptions. For instance, a state-based model of time-varying FC may be described, among others, by transition probabilities between states and a FC matrix per state, in which each entry represents the correlation between two brain regions. Obviously, a change of 0.1 in the transition probability matrix is not equal to a change of 0.1 in a state's FC matrix. Additionally, these two different parameters are interdependent in a way that is specified by the model. Using these parameters in a linear way for a prediction problem would therefore ignore the complex relationships between parameters.

Kernels (Shawe-Taylor & Cristianini, 2004), on the other hand, are a convenient approach to work with high-dimensional, complex features, such as parameters from a model of brain dynamics. In general, kernels are similarity functions between subjects, and they can be used straightforwardly in a prediction algorithm. While feature matrices can be very high dimensional, a kernel is represented by a (no. of subjects by no. of subjects) matrix. How to construct a kernel is however not always straightforward, and choosing an inappropriate kernel can lead to poor performance (Azim & Ahmed, 2018; Shawe-Taylor & Cristianini,

2004). We here propose a method that allows using the entire set of parameters of a generative probabilistic model of brain dynamics in a way that takes the complex relationships between the parameters into account. This is achieved by constructing the Fisher kernel (Jaakkola et al., 1999; Jaakkola & Haussler, 1998), which preserves the structure of the underlying model of brain dynamics (here, the HMM), including the complex relationships between parameters encoded by the model. In this paper, we compare the Fisher kernel to kernels that use the same features, but do not take the structure of the model of brain dynamics into account.

In evaluating the kernels, we consider two criteria that are important in both scientific and more practical applications. First, predictions should be as accurate as possible; i.e., the correlation between predicted and actual values should be high and the mean error should be low. Second, they should be reliable, in the sense that a predictive model should not produce excessively large errors, and the outcome should be robust to the choice of which subjects are included in the training set. The latter criterion is especially important if we want to be able to meaningfully interpret prediction errors, e.g., in assessing brain age (Cole & Franke, 2017; Denissen et al., 2022; Smith et al., 2019). Despite this crucial role in interpreting model errors, error variance is rarely taken into account in models predicting phenotypes from neuroimaging features in the literature.

A central point of this paper is that taking the complex relationships between parameters into account is important for obtaining good predictions according to the two above criteria. To investigate this, we will compare the Fisher kernel to two other kernels that ignore the model structure and do not take these relationships into account. We will also assess the Fisher kernel versus a kernel based on the Kullback-Leibler divergence that was previously used for prediction (Vidaurre et al., 2021), showing that the Fisher kernel is substantially more reliable.

2 Methods

We here aim to predict behavioural and demographic variables from a model of brain dynamics using different kernel functions. The general workflow is illustrated in **Figure 1**. We start with the concatenated fMRI time-series of a group of subjects (**Figure 1a**, step 1). We estimate a model of brain dynamics at the group level, where the state descriptions are shared across subjects (**Figure 1a**, step 2). Next, we estimate subject-specific versions of this group-level brain dynamics model by dual estimation (Vidaurre et al., 2021) (**Figure 1a**, step 3). Next, we use this description of the individuals' brain dynamics to predict their phenotypes. For this, we first project the parameters from the model of brain dynamics into

an embedding space (**Figure 1b**, step 4). We then estimate the similarity between each pair of subjects of the subject-level brain dynamics models in the embedding space using different kernel functions (**Figure 1b**, step 5). Finally, we use these different kernels to predict the behavioural variables using kernel ridge regression (**Figure 1b**, step 6). The first three steps are identical for all kernels and therefore carried out only once. The fourth and fifth steps (projecting the parameters and constructing the kernels) is carried out once for each of the different kernels. The last step is repeated 3,500 times for each kernel to predict a set of 35 different behavioural variables using 100 randomised iterations of 10-fold cross validation. We thus here evaluate a total of 24,500 predictive models in their ability to predict phenotypes, using different kernels constructed from the same model of brain dynamics.

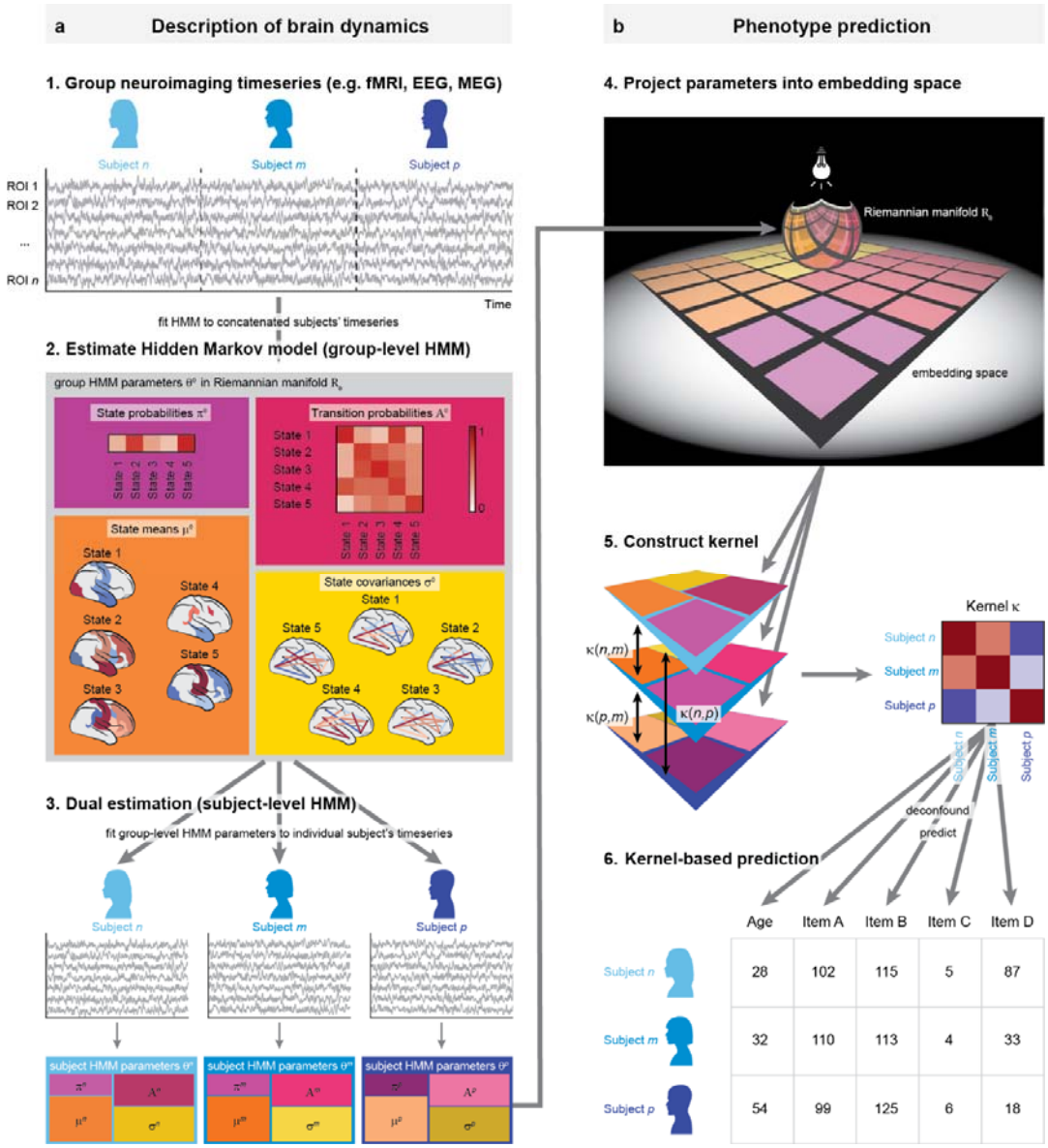

Figure 1 *Workflow of our prediction approach.* ***a)*** *To generate a description of brain dynamics, we (1) concatenate all subjects' individual timeseries; then (2) estimate a Hidden Markov Model (HMM) on these timeseries to generate a group-level description of brain dynamics; then (3) dual-estimate into subject-level HMM models. Steps 1-3 are the same for all kernels.* ***b)*** *In order to then use this description of all subjects' brain dynamics, we project each set of subject-level HMM parameters into an embedding space (4). This projection can be done in different ways: the Fisher kernel reflects the structure of the parameters in their original space while other projections (such as the naïve kernels) do not. We then construct a kernel matrix κ (5), where each pair of subjects is assigned a similarity value given their parameters in the embedding space. Finally, we fed κ to kernel ridge regression to predict a variety of demographic and behavioural traits in a cross-validated fashion.*

2.1 HCP imaging and behavioural data

We used data from the open-access Human Connectome Project (HCP) S1200 release (Smith et al., 2013; Van Essen et al., 2013), which contains MR imaging data and various demographic and behavioural data from 1200 healthy, young adults (age 22-35). All data

described below, i.e., timecourses of the resting-state fMRI data and demographic and behavioural variables, are publicly available at https://db.humanconnectome.org.

Specifically, we used resting state fMRI data of 1,001 subjects, for whom any of the behavioural variables of interest were available. Each participant completed four resting state scanning sessions of 14 mins and 33 seconds duration each. This resulted in 4,800 timepoints per subject. For the main results, we used all four resting state scanning sessions of each participant to fit the model of brain dynamics (but see **Supplementary Figure 1** for results with just one session). The acquisition parameters are described in the HCP acquisition protocols and in Van Essen et al. (2013); Van Essen et al. (2012). Briefly, structural and functional MRI data were acquired on a 3T MRI scanner. The resting state fMRI data was acquired using multiband echo planar imaging sequences with an acceleration factor of 8 at 2 mm isotropic spatial resolution and a repetition time (TR) of 0.72 seconds. The preprocessing and timecourse extraction pipeline is described in detail in Smith et al. (2013). For the resting state fMRI scans, preprocessing consisted of minimal spatial preprocessing and surface projection (Glasser et al., 2013), followed by temporal preprocessing. Temporal preprocessing consisted of single-session Independent Component Analysis (ICA) (Beckmann, 2012) and removal of noise components (Griffanti et al., 2014; Salimi-Khorshidi et al., 2014). Data were high-pass-filtered with a cut-off at 2,000 seconds to remove linear trends.

The parcellation was estimated from the data using multi-session spatial ICA on the temporally concatenated data from all subjects. Using this approach, a data-driven functional parcellation with 50 parcels was estimated, where all voxels are weighted according to their activity in each parcel, resulting in a weighted, overlapping parcellation. While other parcellations are available for the resting-state fMRI HCP dataset, we chose this parcellation because dynamic changes in FC can be better detected in this parcellation compared to functional or anatomical parcellations or more fine-grained parcellations (Ahrends et al., 2022). Timecourses were extracted using dual regression (Beckmann et al., 2009), where group-level components are regressed onto each subject's fMRI data to obtain subject-specific versions of the parcels and their timecourses. We normalised the timecourses of each subject to ensure that the model of brain dynamics and, crucially, the kernels were not driven by (averaged) amplitude and variance differences between subjects.

Subjects in the HCP study completed a range of demographic and behavioural variables. Following Vidaurre et al. (2021), we here focus on a subset of those variables, including age and various cognitive variables. The cognitive variables span items assessing memory, executive function, fluid intelligence, language, processing speed, spatial orientation, and

attention. The full list of the 35 behavioural variables used here, as well as their categorisation within the HCP dataset can be found in **Supplementary Table 1**.

2.2 The Hidden Markov Model as a model of brain dynamics

To estimate brain dynamics, we here use the Hidden Markov Model (Vidaurre et al., 2016; Vidaurre et al., 2017). However, the kernels, which are explained in detail in the following section, can be constructed from any generative probabilistic model.

The Hidden Markov model (HMM) is a generative probabilistic model, which assumes that an observed time-series, such as BOLD signal in a given parcellation, was generated by a sequence of "hidden states" (Baum & Eagon, 1967; Baum & Petrie, 1966). We here model the states as Gaussian distributions, defined both in terms of mean and covariance —which can be interpreted as distinct patterns of amplitude and FC (Vidaurre et al., 2017). Since we are often interested in modelling time-varying FC specifically, we repeated all computations for a variety of the HMM where state means were pinned to zero (given that the data was demeaned and the global average is zero) and only the covariance was allowed to vary across states (Vidaurre et al., 2017), which is shown in **Supplementary Figure 2**.

The HMM is described by a set of parameters θ, containing the state probabilities π, the transition probabilities A, the mean vectors μ of all states, and the covariance matrices σ of all states:

$$\theta = [\pi, A, \mu, \sigma] \qquad \pi \in \mathbb{R}^{1\mathrm{x}K}, A \in \mathbb{R}^{K\mathrm{x}K}, \mu \in \mathbb{R}^{K\mathrm{x}M}, \sigma \in \mathbb{R}^{K\mathrm{x}MxM}$$

where K is the number of states and M is the number of parcels in the parcellation. The entire set of parameters θ is estimated from the data. Here, specifically, we used $K = 6$ states.

The space of parameters θ forms a Riemannian manifold R_θ, where the relationships between the different parameters of the HMM are acknowledged by construction. The Fisher kernel, as described below, is built upon a projection on this manifold, so predictions based on this kernel account for the mathematical structure of the HMM.

Here, we fit the HMM to the concatenated timeseries of all N subjects (see **Figure 1a**, step 1), so states are comparable across the group of subjects. We refer to the group-level estimate as HMM^0, which is defined by the parameters θ^0 (see **Figure 1a**, step 2):

$$\theta^0 = [\pi^0, A^0, \mu^0, \sigma^0]$$

In order to use the information from the HMM to predict subjects' phenotypes, we estimate subject-specific versions of the group-level HMM (see **Figure 1a**, step 3) through dual estimation (Vidaurre et al., 2017), where the parameters from the group-level HMM^0 are adapted to a particular subject's timeseries; we will refer to the subject-specific estimate for subject n as HMM^n, with parameters θ^n.

2.3 Kernels from Hidden Markov Models

Kernel methods can readily be adapted to deal with nonlinear decision boundaries in prediction, by projecting the data into a high-dimensional (possibly infinite-dimensional) space through an embedding $x \rightarrow \phi_x$; then, by estimating a linear separating hyperplane on this space, we can effectively have a nonlinear estimator on the original space (Shawe-Taylor & Cristianini, 2004). In practice, instead of working explicitly in a higher-dimensional embedding space, the so-called kernel trick uses a kernel function $\kappa(n, m)$ containing the similarity between data points n and m (here, subjects) in the higher-dimensional embedding space (Schölkopf et al., 2002; Shawe-Taylor & Cristianini, 2004), which can be simpler to calculate. Once $\kappa(\cdot,\cdot)$ is computed for each pair of subjects, this is all that is needed for the prediction. This makes kernels computationally very efficient. The downside is that finding the right kernel can be a challenge because there are many available alternatives for the embedding.

Here, in combination with a linear predictive model, we apply a kernel that is specifically conceived to be used to compare instances of a generative model like the HMM. We expected that this will result in better predictions than existing methods. Using the same HMM estimate, we compare three different kernels, which embed the HMM parameters into three distinct spaces, corresponding to different embeddings (see **Figure 1b**, step 4): the naïve kernel, the naïve normalised kernel, and the Fisher kernel. While the first two kernels (naïve and naïve normalised kernel) do not take the relationships between the HMM's parameters into account, the Fisher kernel preserves the structure of the HMM and weights the parameters according to how much they contribute to generating a particular subject's timeseries. We then construct linear and Gaussian versions of the different kernels (see **Figure 1b**, step 5), which take the general form $\kappa(n, m) = {\phi_{x^n}}^T \phi_{x^m}$ for the linear kernel and $\kappa_g(n, m) = exp\ (-\frac{\|\phi_{x^n} - \phi_{x^m}\|^2}{2\tau^2})$ for the Gaussian kernel. We compare these to a kernel constructed using the Kullback-Leibler divergence, previously used for predicting behavioural phenotypes (Vidaurre et al., 2021).

2.3.1 Naïve kernel

The naïve kernel is based on a simple vectorisation of the subject-specific version of the HMM's parameters, each on their own scale. This means that the kernel does not take relationships between the parameters into account and the parameters are here on different scales. This procedure can be thought of as computing Euclidean distances between two sets of HMM parameters, ignoring the actual geometry of the space of parameters. For each subject n, we vectorise parameters θ^n obtained through dual estimation of the group-level parameters θ^0 to map the example $x^n \rightarrow \phi_{x^n}$ to:

$$\theta^n = (\pi^n, A^n, \mu^n, \sigma^n) \qquad \theta^n \in \mathbb{R}^{1x(K+K*K+K*M+K*M*M)}$$

We will refer to this vectorised version of the subject-specific HMM parameters as "naïve θ". The naïve θ are the features used in the naïve kernel κ_N. We first construct a linear kernel from the naïve θ features using the inner product of the feature vectors. The linear naïve kernel κ_{Nl} between subjects n and m is thus defined as:

$$\kappa_{Nl}(n, m) = \langle \theta^n, \theta^m \rangle \qquad \kappa_{Nl} \in \mathbb{R}^{NxN}$$

where $\langle \theta^n, \theta^m \rangle$ denotes the inner product between θ^n and θ^m. Using the same feature vectors, we can also construct a Gaussian kernel from the naïve θ. The Gaussian naïve kernel κ_{Ng} for subjects n and m is defined as:

$$\kappa_{Ng}(n, m) = \exp\left(-\frac{\|\theta^n - \theta^m\|^2}{2\tau^2}\right) \qquad \kappa_{Ng} \in \mathbb{R}^{NxN}$$

where τ is the radius of the radial basis function, and $\|\theta^n - \theta^m\|$ is the L_2-norm of the difference of the feature vectors (naïve θ for subjects n and m). Compared to a linear kernel, a Gaussian kernel embeds the features into a more complex space, which can potentially improve the accuracy. However, this kernel has an additional parameter τ that needs to be chosen, typically through cross-validation. This makes a Gaussian kernel computationally more expensive and, if the additional parameter τ is poorly estimated, more error prone. The effect of the hyperparameters on errors is shown in **Supplementary Figure 3**.

While the naïve kernel takes all the information from the HMM into account by using all parameters from a subject-specific version of the model, it uses these parameters in a way that ignores the structure of the model that these parameters come from. In this way, the different parameters in the feature vector are difficult to compare, since e.g., a change of 0.1 in the transition probabilities between two states is not of the same magnitude as a change of 0.1 in one entry of the covariance matrix of a specific state. In the naïve kernel, these two very different types of changes would be treated indistinctly.

2.3.2 Naïve normalised kernel

To address the problem of parameters being on different scales, the naïve normalised kernel makes the scale of the subject-specific vectorised parameters (i.e., the naïve θ) comparable across parameters. Here, the mapping $x \rightarrow \phi_x$ consists of a vectorisation and normalisation across subjects of the subject-specific HMM parameters, by subtracting the mean from each parameter and dividing by the standard deviation. This kernel does not respect the geometry of the space of parameters either.

As for the naïve kernel, we can then construct a linear kernel from these vectorised, normalised parameters $\tilde{\theta}^n$ by computing the inner product for all pairs of subjects n and m to obtain the linear naïve normalised kernel κ_{NNl}:

$$\kappa_{NNl}(n, m) = \langle \tilde{\theta}^n, \tilde{\theta}^m \rangle \qquad \kappa_{NNl} \in \mathbb{R}^{NxN}$$

We can also compute a Gaussian kernel from the naïve normalised feature vectors to obtain the Gaussian version of the naïve normalised kernel κ_{NNg}:

$$\kappa_{NNg}(n, m) = \exp\left(-\frac{\|\tilde{\theta}^n - \tilde{\theta}^m\|^2}{2\tau^2}\right) \qquad \kappa_{NNg} \in \mathbb{R}^{NxN}$$

In this way, we have constructed a kernel in which parameters are all on the same scale, but which still ignores the complex relationships between parameters originally encoded by the underlying model of brain dynamics.

2.3.3 Fisher kernel

The Fisher kernel (Jaakkola et al., 1999; Jaakkola & Haussler, 1998) is specifically designed to preserve the structure of a generative probabilistic model (here, the HMM). This can be thought of as a “proper” projection on the manifold illustrated in **Figure 1b**, step 4, represented by the light creating a shadow grid on a new surface. Similarity between subjects is here defined in reference to a group-level model of brain dynamics. The mapping $x \rightarrow \phi_x$ is given by the “Fisher score”, which indicates how much each parameter of the time-varying function and FC model contributes to generating the timeseries of a particular subject, when compared to the group of subjects. In other words, the Fisher score defines how (in which direction in the Riemannian parameter space) we would have to change the group-level model to best explain a particular subject’s timeseries. The similarity between subjects can then be described based on this score, so that two subjects are defined as similar if the group-level model would have to be changed in a similar direction for both, and dissimilar otherwise.

More precisely, the Fisher score is given by the natural gradient of the log-likelihood with respect to each model parameter:

$$g(\theta^0, x^n) = \left(\frac{\partial \log \mathcal{L}_{\theta^0}(x^n)}{\partial \theta^0}\right) \qquad g \in \mathbb{R}^{1x(K+K*K+K*M+K*M*M)}$$

where x^n is the timeseries of subject n, and $\mathcal{L}_\theta(x) = P(x|\theta)$ represents the likelihood of the timeseries x given the model parameters θ. This way, the Fisher score maps an example (i.e., a subject's timeseries) x^n into a point in the gradient space of the Riemannian manifold R_θ defined by the HMM parameters.

The Fisher kernel κ_{F-} is the inner product of the Fisher score g, scaled by the Fisher information matrix F, which gives a local metric on the Riemannian manifold R_θ:

$$\kappa_{F-}(n,m) = g(\theta^0, x^n)^{\mathrm{T}} F_{R_\theta}^{-1} g(\theta^0, x^m) \qquad \kappa_{F-} \in \mathbb{R}^{NxN}$$

for subjects n and m. F_{R_θ} is the Fisher information matrix, defined as

$$F_{R_\theta} = \mathbb{E}_x[g(\theta^0, x) g(\theta^0, x)^{\mathrm{T}}]$$

where the expectation is with respect to x under the distribution $P(x|\theta)$.

In practice, the Fisher information matrix F_{R_θ} can be disregarded because, apart from being computationally intensive to calculate, it is in practice immaterial (Jaakkola et al., 1999; Jaakkola & Haussler, 1998; van der Maaten, 2011). Therefore, we can reduce the Fisher kernel to the so-called practical Fisher kernel, for which the linear version κ_{Fl} takes the form:

$$\kappa_{Fl}(n,m) = \langle g(\theta^0, x^n), g(\theta^0, x^m) \rangle \qquad \kappa_{Fl} \in \mathbb{R}^{NxN}$$

In this study, we will use the practical Fisher kernel for all computations.

One issue when working with the linear Fisher kernel is that the gradients are typically all close to zero, so that the norms used for the prediction are very small. To mitigate this, we can plug the gradient features (i.e., the Fisher scores g) into a Gaussian kernel, such that, for subjects n and m, where x^n is the timeseries of subject n and x^m is the timeseries of subject m, the Gaussian Fisher kernel κ_{Fg} is defined as

$$\kappa_{Fg}(n,m) = \exp\left(-\frac{\|g(\theta^0, x^n) - g(\theta^0, x^m)\|^2}{2\tau^2}\right) \quad \kappa_{Fg} \in \mathbb{R}^{NxN}$$

2.3.4 Kullback-Leibler divergence

The Kullback-Leibler (KL) divergence is an information-theoretic distance measure which estimates divergence between probability distributions —in this case between subject-specific versions of the HMM. Here, the KL divergence of subject n from subject m, $\mathrm{KL}(\mathrm{HMM}^n||\mathrm{HMM}^m)$ can be interpreted as how much new information the HMM of subject n contains if the true distribution was the HMM of subject m. The KL divergence is not symmetric, i.e., $\mathrm{KL}(\mathrm{HMM}^n||\mathrm{HMM}^m)$ is different than $\mathrm{KL}(\mathrm{HMM}^m||\mathrm{HMM}^n)$. We use an approximation of KL divergence which is described in detail in Do (2003), as in Vidaurre et al. (2021). That is, given two models HMM^n from HMM^m for subject n and subject m, we have

$$\mathrm{KL}(\mathrm{HMM}^n||\mathrm{HMM}^m) = \sum_k v_k KL(A_k^n, A_k^m) + v_k KL(G_k^n, G_k^m)$$

Where A_k^n are the transition probabilities from state k into any other state according to HMM^n and G_k^n are the state Gaussian distributions for state k and HMM^n (respectively A_k^m and G_k^m for HMM^m) (see MacKay et al. (2003)). Since the transition probabilities are Dirichlet-distributed and the state distributions are Gaussian distributed, the KL divergence for those has a closed-form solution. Variables v_k can be computed numerically such that

$$vA^n = v,$$

$$\lim_{x\to\infty} \pi^n A^{n^x} = v$$

To be able to use the KL divergence as a kernel, we symmetrise the KL divergence matrix as

$$\mathcal{D}_{KL}(n,m) = 0.5\ \mathrm{KL}(\mathrm{HMM}^n||\mathrm{HMM}^m) + 0.5\ \mathrm{KL}(\mathrm{HMM}^m||\mathrm{HMM}^n) \qquad \mathcal{D}_{KL} \in \mathbb{R}^{NxN}$$

This symmetrised KL divergence can be plugged into a radial basis function, analogous to the Gaussian kernels to obtain a similarity matrix κ_{KL}

$$\kappa_{KL}(n,m) = \exp\left(-\frac{(\mathcal{D}_{KL}(n,m))^2}{2\tau^2}\right) \qquad \kappa_{KL} \in \mathbb{R}^{NxN}$$

The resulting KL similarity matrix can be used in the predictive model in a similar way as the kernels described above.

As a general benchmark for model robustness, we also construct a KL divergence matrix of static FC, as described in Vidaurre et al. (2021), shown in **Supplementary Figure 4**.

2.4 Predictive model: Kernel ridge regression

Similarly to Vidaurre et al. (2021), we use kernel ridge regression (KRR) to predict demographic and behavioural variables from the different kernels, although these could be used in any kernel-based prediction model or classifier, such as a support vector machine. KRR is the kernelized version of ridge regression (Saunders et al., 1998):

$$\hat{y} = h\alpha \qquad \alpha \in \mathbb{R}^{S_{train}x1}, h \in \mathbb{R}^{S_{test}xS_{train}}$$

where α are the regression weights; h is the (number of subjects in test set by number of subjects in training set) kernel matrix between the subjects in the training set and the subjects in the test set; $\hat{y}$ are the predictions in the (out-of-sample) test set; S_{train} are the number of subjects in the training set; and S_{test} are the number of subjects in the test set. The regression weights α can be estimated using the kernels specified above as

$$\alpha = (\kappa + \lambda I)^{-1} * y \qquad \kappa \in \mathbb{R}^{S_{train}xS_{train}}$$

where λ is a regularisation parameter that we can choose through cross-validation; I is the identity matrix; κ is the (S_{train} by S_{train}) kernel matrix of the subjects in the training set; and y are the training examples.

We use KRR to separately predict each of the 35 demographic and behavioural variables from each of the kernels, removing subjects with missing entries from the prediction. The hyperparameters λ (and τ in the case of Gaussian kernels) were selected in a nested cross-validated (CV) fashion using 10 folds for both inner and outer loops assigning 90% of the data to the training set and 10% to the test set for each fold. We used a CV scheme to account for family structure in the HCP dataset so that subjects from the same family were never split across folds (Winkler et al., 2015). We repeated the nested 10-fold CV 100 times, so that different combinations of subjects were randomly assigned to the folds at each new CV iteration to obtain a distribution of model performance values for each variable. This is to explicitly show how susceptible each model was to changes in the training folds, which we can take as a measure of the robustness of the estimators, as described below.

2.4.1 Evaluation criteria

We evaluate the models in terms of two outcome criteria: prediction accuracy and reliability. The first criterion, prediction accuracy, concerns how close the model's predictions were to the true values on average. We here quantify prediction accuracy as Pearson's correlation coefficient $r(\hat{y}, y)$ between the model-predicted values and the actual values of each variable, as well as the average errors of the models. Average errors are here defined as normalised root mean square errors (NRMSE). The root mean square error (RMSE) is

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}}$$

To make the results comparable across the different traits, the RMSE is then divided by the range of the respective variable to obtain the NRMSE

$$NRMSE = \frac{RMSE}{y_{max} - y_{min}}$$

The second criterion, reliability, concerns two aspects: i) that the model will not show too large errors for some subjects that could harm interpretation; and ii) that the model's accuracy will be consistent across random variations of the training set —in this case by using different (random) partitions for the CV folds. As mentioned, this is important if we want to interpret prediction errors for example in clinical contexts, which assumes that the error size of a model in a specific subject reflects something biologically meaningful, e.g., to what extent this subject defies the stereotypical population (Greene et al., 2022) or whether a certain disease causes the brain to "look" older to a model than the actual age of the subject (Denissen et al., 2022). While average errors can be informative to understand the general model performance, maximum errors inform us about single cases where a model (that may typically perform well) fails. The maximum absolute error (MAXAE) is the single largest error made by the model in each run, i.e.,

$$MAXAE = \max_{i \in N}(|y_i - \hat{y}_i|)$$

We here look at the normalised maximum absolute error (NMAXAE), which is, analogous to the NRMSE, defined as

$$NMAXAE = \frac{ME}{y_{max} - y_{min}}$$

We then consider the risk of large errors (NMAXAE > 10), very large errors (NMAXAE > 100), and extreme errors (NMAXAE > 1,000) as the percentage of runs (across variables and CV iterations) where the model's NMAXAE exceeds the given threshold. Since NMAXAE is normalised by the range of the actual variable, these thresholds correspond to one, two, and three orders of magnitude of the actual variable's range. If we are predicting age, for instance, and the true ages of the subjects range from 25 years to 35 years, an NMAXAE of 1 would mean that the model's least accurate prediction is off by 10 years – which would be acceptable – while an NMAXAE of 10 would mean that the least accurate prediction is off by 100 years, an NMAXAE of 100 would be off by 1,000 years, and an NMAXAE of 1,000 would be off by 10,000 years. A model that makes such large errors,

even in single cases, would be unusable for interpretation. Our reliability criterion in terms of maximum errors is therefore that the risk of large errors (NMAXAE > 10) should be 0%.

For a model to be reliable, it should also be robust in the sense of susceptibility to changes in the training examples, e.g., how much the model's estimation suffers when encountering outliers during training. Robustness is an important consideration in prediction studies, as it determines how reproducible a study or method is —which is often a shortcoming in neuroimaging-based prediction studies (Varoquaux et al., 2017). We evaluated the robustness by iterating nested 10-fold CV 100 times for each variable, randomising the subjects in the folds at each iteration, so that the models would encounter a different combination of subjects in the training and test sets each time they are run. Looking at this range of 100 predictions for each variable, we can assess whether the model's performance changes drastically depending on which combinations of subjects it encountered in the training phase, or whether the performance is the same regardless of the subjects encountered in the training phase. The former would be an example of a model that is susceptible to changes in training examples and the latter an example of a model that is robust. We here quantify robustness as the mean across variables of the standard deviation (S.D.) of the prediction accuracy, *r* and NRMSE, across the 100 permutations, where a small mean S.D. indicates higher robustness.

We test for significant differences in prediction accuracy and robustness using permutation tests with 5,000 permutations. P-Values are corrected across multiple comparisons using the Benjamini-Hochberg's False Discovery Rate (FDR) procedure (Benjamini & Hochberg, 1995). To compare the Fisher kernel to the naïve and the naïve normalised kernels, we first evaluate the linear and Gaussian versions of the kernels together. We then also compare the linear and Gaussian versions of each kernel with each other. We also compare each kernel to the KL divergence model.

2.5 Implementation

The code used in this study was implemented in Matlab and is publicly available within the repository of the HMM-MAR toolbox at https://github.com/OHBA-analysis/HMM-MAR. Code for constructing kernels from an HMM can be found at https://github.com/OHBA-analysis/HMM-MAR/utils/prediction/hmm_kernel.m. We construct the Fisher kernel as described in Jaakkola et al. (1999); Jaakkola and Haussler (1998). The HMM gradient computation on which the Fisher kernel is based is adapted from van der Maaten (2011). The Gaussian version of the Fisher kernel is described in Shawe-Taylor and Cristianini (2004). Code for the approximation of KL divergence from an HMM can be found at

https://github.com/OHBA- analysis/HMM- MAR/blob/master/utils/math/hmm_kl.m, as described in Vidaurre et al. (2021). Code for KRR using randomly permutable CV that accounts for family structure is available at https://github.com/OHBA-analysis/HMM-MAR/utils/prediction/predictPhenotype_kernels.m. All code used in this paper, including scripts to reproduce the figures and additional application examples of the Fisher Kernel can be found in https://github.com/ahrends/FisherKernel.

3 Results

We compare kernels derived from a model of brain dynamics in their ability to predict subject traits. We here show the superiority of the Fisher kernel, which preserves the underlying structure of the brain dynamics model. We assess the performance of the predictive models that use these different kernels in terms of two outcome criteria: prediction accuracy and reliability.

3.1 The Fisher kernel outperforms other kernels in terms of prediction accuracy

We found that the Fisher kernel has the highest prediction accuracy on average across the range of variables and CV iterations. In terms of correlation between model-predicted and actual values, the Fisher kernel significantly outperformed both the naïve kernel (mean *r* κ_F: 0.196, mean *r* κ_N: 0.127; p_{BH}=0.002) and the naïve normalised kernel (mean *r* κ_{NN}: 0.159; p_{BH}=0.009). The Fisher kernel also had the smallest average errors (mean *NRMSE* κ_F: 0.0013), although the comparisons to the naïve kernels were not significant (mean *NRMSE* κ_N: 0.003, p_{BH}=0.199; mean *NRMSE* κ_{NN}: 0.0019, p_{BH}=0.260). Comparing the two naïve kernels, the normalised version improved upon the non-normalised naïve kernel, but not significantly so (*r* κ_N vs. κ_{NN}: p_{BH}=0.171, *NRMSE* κ_N vs. κ_{NN}: p_{BH}=0.236).

We next compared the linear with the Gaussian versions of each kernel. The linear versions performed significantly better than the Gaussian versions for the Fisher kernel (mean *r* κ_{Fl}: 0.204, mean *r* κ_{Fg}: 0.187; p_{BH}=0.009; mean *NRMSE* κ_{Fl}: 0.0005, mean *NRMSE* κ_{Fg}: 0.002; p_{BH}=0.048) and the naïve normalised kernel (mean *r* κ_{NNl} 0.187:, mean *r* κ_{NNg}: 0.131; p_{BH}=0.0002; mean *NRMSE* κ_{NNl}: 0.0008, mean *NRMSE* κ_{NNg}: 0.003; p_{BH}=0.032), both in terms of correlation between model-predicted and actual values and in terms of average errors. The opposite was the case for the naïve kernel, where the Gaussian version significantly outperformed the linear version (mean *r* κ_{Nl}: 0.103, mean *r* κ_{Ng}: 0.150; p_{BH}=0.004; mean *NRMSE* κ_{Nl}: 0.001, mean *NRMSE* κ_{Ng}: 0.004; p_{BH}=0.0003).

Finally, we compared all models to the baseline KL divergence model. Only the linear versions of the Fisher kernel and the naïve normalised kernel significantly outperformed the KL divergence model, both in terms of correlation between model-predicted and actual values (mean *r* κ_{KL}:0.154, compared to κ_{Fl}: p_{BH}=0.002; compared to κ_{NNl}: p_{BH}=0.008) and in terms of average errors (mean *NRMSE* κ_{KL}: 0.003, compared to κ_{Fl}: p_{BH}=0.003; compared to κ_{NNl}: p_{BH}=0.006). While the Gaussian version of the Fisher kernel had a higher average correlation between model-predicted and actual values and smaller average errors than the KL divergence model, this comparison was not significant (*r* p_{BH}=0.07, *NRMSE* p_{BH}=0.53). The Gaussian version of the naïve normalised kernel performed worse than the KL divergence model (*r* p_{BH}=0.553, *NRMSE* p_{BH}>1). Both versions of the naïve kernel were less accurate than the KL divergence model in terms of correlation between model-predicted and actual values and in terms of average errors: This comparison was significant for the linear version of the naïve kernel (*r*: p_{BH}=0.017; *NRMSE*: p_{BH}=0.05), but not for the Gaussian version (*r*: p_{BH}>1; *NRMSE*: p_{BH}=0.27). The distributions of correlation coefficients and average errors for the different kernels are shown in **Figure 2a** and **Figure 2b,** respectively. Also, **Table 1** summarises the performance of all kernels. We found the same pattern of higher accuracy of the Fisher kernel compared to other kernels when fitting the HMM to only the first resting-state session of each subject, as shown in **Supplementary Figure 1**. The overall accuracies were lower, however, indicating that predicting traits from an HMM benefits from a large amount of available data per subject. Similarly, the Fisher kernel outperformed the other kernels when constructing the kernels from an HMM where states are only defined in terms of covariance (not mean) at comparable accuracy, as shown in **Supplementary Figure 2**.

In summary, with regard to prediction accuracy we found that the linear Fisher kernel has the highest prediction accuracy, outperforming both the naïve kernels, its Gaussian version, and the baseline KL divergence model.

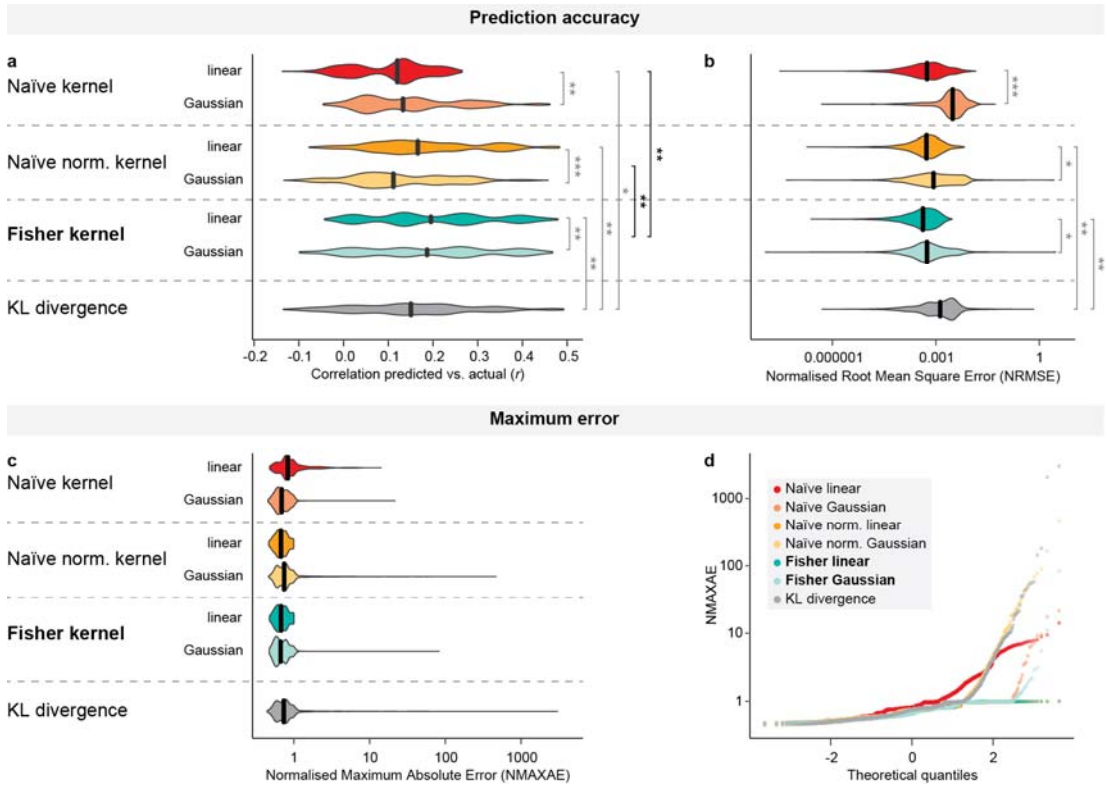

Figure 2 *Distributions across behavioural variables of model performance by kernel.* ***a)*** *Correlation between model-predicted and actual values. The violin plots show the distribution of Pearson's correlation coefficients (r) between predicted and actual variable values as a measure of prediction accuracy (x-axis) of each kernel (y-axis). The Fisher kernel has the highest average prediction accuracy.* ***b)*** *Average errors. The violin plots show the distribution of the normalised root mean square errors (NRMSE) on the x-axis for each kernel (y-axis).* ***c)*** *Maximum error size. The plots show the distribution of normalised maximum errors (NMAXAE) as a measure of maximum error size (x-axis) by kernel (y-axis). Differences between the kernels mainly lie in the tails of the distributions, where the Gaussian naïve normalised kernel and the KL divergence model produce excessive maximum errors in some runs. The x-axis is plotted on the log-scale.* ***d)*** *The quantile plot highlights how the differences in maximum errors lie in the right tails of the distribution.The y-axis is plotted on the log-scale. a)-c) Each violin plot pools 3,500 runs (100 iterations of 10-fold CV for all 35 variables) that were predicted from each kernel. Asterisks indicate significant Benjamini-Hochberg-corrected p-values for permutation using 5,000 permutations: *: $p_{BH} < 0.05$, **: $p_{BH} < 0.01$, ***: $p_{BH} < 0.001$.*

Table 1 *Summary of model performance. The table shows the average performance and the range of the prediction accuracy measures, correlation between model-predicted and actual values and average errors (NRMSE), and of maximum errors (NMAXAE). The best performing models are highlighted in bold.*

Kernel		**Naïve**		**Naïve norm.**		**Fisher**		**KL divergence**
		Linear	Gaussian	Linear	Gaussian	Linear	Gaussian	
Correlation coefficient ®	**min.**	-0.137	-0.046	-0.077	-0.134	**-0.042**	-0.099	-0.135
	mean	0.103	0.15	0.187	0.131	**0.204**	0.187	0.154
	median	0.124	0.136	0.167	0.109	**0.195**	0.189	0.149
	max.	0.265	0.462	0.483	0.458	0.48	0.467	**0.493**
NRMSE	**min.**	$2.983*e^{-8}$	$4.863*e^{-7}$	$1.820*e^{-7}$	$4.578*e^{-8}$	$2\text{-}379*e^{-7}$	**$1.145*e^{-8}$**	$5.002*e^{-7}$
	mean	0.001	0.004	0.0008	0.003	**0.0005**	0.002	0.003
	median	0.0006	0.003	0.0005	0.0009	**0.0004**	0.0005	0.001
	max.	0.015	0.053	0.007	2.665	**0.003**	3.007	0.686
NMAXAE	**min.**	0.476	0.455	0.464	0.462	0.474	0.472	**0.442**
	mean	1.08	0.74	**0.702**	1.354	0.707	0.727	2.617

median	0.812	0.678	0.686	0.758	0.675	**0.653**	0.756
max.	14.193	21.668	**0.996**	468.628	1.003	83.097	3022.214

3.2 The Fisher kernel has a lower risk of large errors and is more robust than other kernels

We now show empirically that that the Fisher kernel is more reliable than other kernels, both in terms of risk of large errors and in terms of robustness over CV iterations.

The Fisher kernel has the overall lowest risk of large errors. We assessed the risk of large errors (NMAXAE > 10), very large errors (NMAXAE > 100), and extreme errors (NMAXAE > 1,000), corresponding to one, two, and three orders of magnitude of the range of the actual variable. For the Fisher kernel, the risk of large errors is low at 0.057% only in the Gaussian version κ_{Fg} and 0% in the linear version κ_{Fl}. In the naïve kernel, both the linear κ_{Nl} and the Gaussian version κ_{Ng} have a low risk of large errors at 0.029% for the linear version and 0.057% for the Gaussian version. The Gaussian naïve normalised kernel κ_{NNg} on the other hand has a slightly higher risk of large errors at 1.140% and a risk of very large errors at 0.057%, while its linear version κ_{NNl} has a 0% risk of large errors. The risk of large errors was highest in the KL divergence model κ_{KL} at 0.890% risk of large errors, 0.110% risk of very large errors, and even 0.057% risk of extreme errors. These peak errors were smallest in the linear naïve normalised kernel κ_{NNl} at 0.996, closely followed by the linear Fisher kernel κ_{Fl} at 1.003. The Gaussian versions of all kernels produced a larger range of errors, reaching up to 21.668, 83.097, and 468.628 for the Gaussian naïve kernel κ_{Ng}, the Gaussian Fisher kernel κ_{Fg}, and the Gaussian naïve normalised kernel κ_{NNg}, respectively. In the case of the naïve kernel, even the linear kernel κ_{Nl} can produce considerable errors, reaching up to 14.193. We observed the by far largest errors in the KL divergence model, where normalised maximum errors in some cases went up to 3,022.214. The summary of maximum errors is reported in **Table 1** and the risk of large errors for the different kernels in **Table 2**. The maximum error distributions are shown in **Figure 2c-d**.

A reason for the higher risk of large errors in the Gaussian versions of the kernels may be that an additional hyperparameter, the radius τ of the radial basis function, had to be estimated during CV for the Gaussian versions of the kernels, leaving more room for error. **Supplementary Figure 3** shows the relation between the estimated hyperparameters (the regularisation parameter λ and the radius τ of the radial basis function) and how large errors in the predictions may be related to poor estimation of these parameters.

We found that the linear Fisher kernel κ_{Fl} and the linear naïve normalised kernel κ_{NNl} had the most robust performance on average across the range of variables tested. **Table 2** shows the standard deviation of the measures of prediction accuracy (correlation between model-predicted and actual values and average errors) over 100 iterations of 10-fold CV, averaged across the 35 variables tested. Comparing the Fisher kernel κ_F to the naïve kernels κ_N and κ_{NN} was not significant (κ_F S.D. *r*: 0.015, S.D. *NRMSE*: 0.005, compared to naïve kernel κ_N S.D. *r*: 0.019, S.D. *NRMSE*: 0.0008, p_{BH}=0.452; compared to naïve normalised kernel κ_{NN} S.D. *r*: 0.024, S.D. *NRMSE*: 0.006, p_{BH}=0.500), similar to the comparison between the naïve κ_N and the naïve normalised kernel κ_{NN} (p_{BH}=0.113). We did, however, find that the linear versions of the Fisher kernel and the naïve normalised kernel were significantly more robust than their respective Gaussian versions (κ_{Fl} S.D. *r*: 0.011, S.D. *NRMSE*: 0.0004, compared to κ_{Fg} S.D. *r*: 0.018, S.D. *NRMSE*: 0.010, p_{BH}=0.0002; κ_{NNl} S.D. *r*: 0.011, S.D. *NRMSE*: 0.0004, compared to κ_{NNg} S.D. *r*: 0.036, S.D. *NRMSE*: 0.012, p_{BH}=0.0002). All but the Gaussian naïve normalised κ_{NNg} and the Gaussian Fisher kernel κ_{Fg} significantly outperformed the KL divergence model κ_{KL} in terms of robustness (κ_{KL} S.D. *r*: 0.042, S.D. *NRMSE*: 0.008, compared to κ_{Nl} p_{BH}=0.0002; compared to κ_{Ng} p_{BH}=0.0002, compared to κ_{NNl} p_{BH}=0.0002; compared to κ_{NNg} p_{BH}>1, compared to κ_{Fl} p_{BH}=0.002; compared to κ_{Fg} p_{BH}=0.550). This large variation in model performance depending on the CV fold structure in the KL divergence model is problematic. As shown in **Supplementary Figure 4**, the same problem occurs when using static FC features in a KL divergence model, indicating robustness issues of the KL divergence measure in this type of predictive model independent of the underlying features.

The ranges in model performance across CV iterations for each variable of the different kernels are shown in **Figure 3**. **Figure 3a** shows the range in correlation between model-predicted and actual values and **Figure 3b** the range in average errors. Robustness of all kernels is also summarised in **Table 2**.

Overall, we found that the linear Fisher kernel and the linear naïve normalised kernel were the most reliable. For both kernels, the risk of large errors was zero and the variability over CV iterations was the smallest. The Gaussian versions of all kernels had higher risks of large errors and a larger range in model performance over CV iterations, indicating that their performance was less reliable. The KL divergence model was the most problematic in terms of reliability with a risk of extreme errors ranging up to three orders of magnitude of the actual variable's range and considerable susceptibility to changes in CV folds.

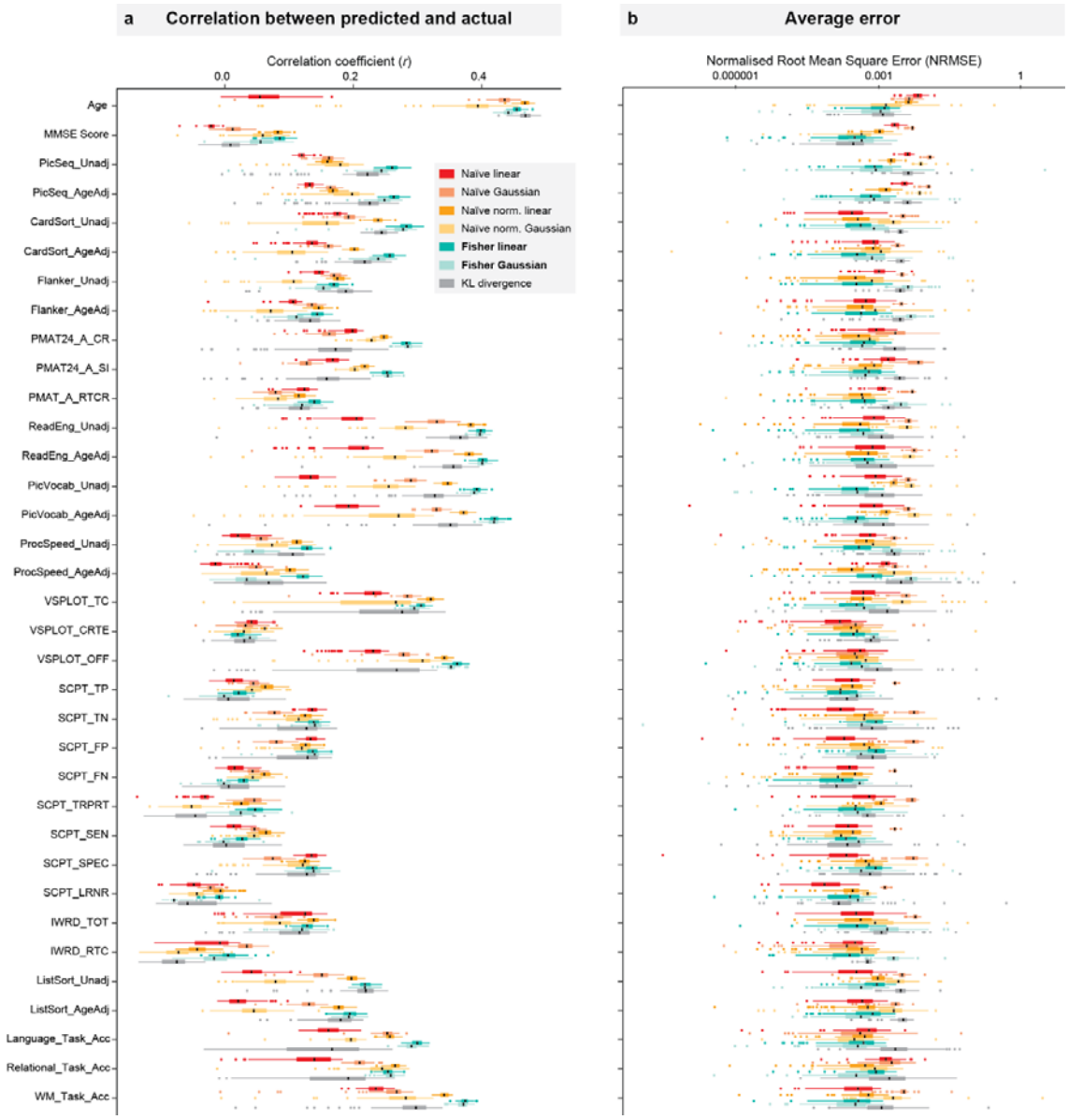

Figure 3 *Robustness of model performance estimates over CV iterations by behavioural variable and model type.* ***a)*** *Correlation between predicted and actual values. The boxplots show the distribution over 100 iterations of 10-fold CV of correlation coefficient values (x-axis) of each kernel, separately for each of the 35 predicted variables (y-axis). The Fisher kernel shows the narrowest range, indicating high robustness.* ***b)*** *Normalised root mean square errors (NRMSE) (x-axis) over 100 iterations of 10-fold CV for each kernel, separately for each of the 35 predicted variables (y-axis). (a-b) Black lines within each boxplot represent the median.*

Table 2 *Reliability of different kernels. Risk of large (NMAXAE > 10), very large (NMAXAE > 100), and extreme (NMAXAE > 1,000) errors for each kernel (in %). Average robustness across randomised CV iterations in the different kernels as measured by range in prediction accuracy estimates. For each measure, we computed the standard deviation across 100 iterations of 10-fold CV of each kernel in each of the 35 variables and report the mean across the 35 variables. The ranges on which standard deviations were computed are plotted separately for each variable in Figure 3. The best performing models are highlighted in bold.*

Kernel	**Naïve**	**Naïve norm.**	**Fisher**	**KL divergence**

	Linear	Gaussian	Linear	Gaussian	Linear	Gaussian	
Risk of NMAXAE > 10	**0.029**	0.057	**0**	1.140	**0**	0.057	0.890
Risk of NMAXAE > 100	**0**	**0**	**0**	0.057	**0**	**0**	0.110
Risk of NMAXAE > 1,000	**0**	**0**	**0**	**0**	**0**	**0**	0.057
Mean S.D. correlation coefficient *r*	0.02	0.018	**0.011**	0.036	**0.011**	0.018	0.042
Mean S.D. *NRMSE*	0.0005	0.001	**0.0004**	0.012	**0.0004**	0.010	0.008

3.3 The Fisher kernel can predict traits like age, language skills, and spatial orientation at relatively high accuracy

We found that using the Fisher kernel, we can predict certain traits at relatively high accuracy. To illustrate the kernels' potential to predict traits from models of brain dynamics, we here consider the best predictions from all kernels. So far, we have focussed on model performance averaged over all variables that were to be predicted. There are, however, differences in how well these demographic or behavioural variables *can* be predicted from a model of brain dynamics. While there are some variables which may show a strong link to brain dynamics, other variables may be more related to static or structural measures, or just be difficult to predict in general. To understand the potential of different kernels to predict phenotypes from a model of brain dynamics, we now focus only on variables that can be predicted from any kernel at relatively high accuracy.

We here consider a mean Pearson's *r* across CV iterations greater than 0.3 as "relatively high" accuracy. **Table 3** shows the variables that can be predicted at relatively high accuracy by at least one kernel. All these variables can be predicted at relatively high accuracy by more than one kernel. On average, six variables (variables #12: ReadEng_Unadj, #13: ReadEng_AgeAdj, #14: PicVocab_Unadj, #15: PicVocab_AgeAdj, #20: VSPLOT_OFF, and #35: WM_Task_Acc) are best predicted by the Fisher kernel, one variable is best predicted by the naïve normalised kernel (variable #18: VSPLOT_TC), and one variable is best predicted by the KL divergence model (variable #1: Age). In all these variables, the linear naïve kernel produces the least accurate predictions on average. Examples of best-predicted variables for all kernels can be found in **Supplementary Figure 5**.

The differences between the linear and the Gaussian versions of the kernels depend on the kernel used: in the case of the naïve kernel, the Gaussian version improves the prediction accuracy, while it worsens the prediction accuracy of the naïve normalised kernel and the Fisher kernel in these variables.

The linear Fisher kernel, closely followed by its Gaussian version, is thus the kernel that produces the best predictions in the traits that can be predicted well by any one kernel.

Table 3 *Variables predicted at relatively high accuracy by at least one kernel. The table shows mean correlation coefficients for all kernels in those variables that could be predicted at relatively high accuracy (correlation coefficient > 0.3) by at least one kernel. The best performing model is highlighted in bold.*

Kernel		**Naïve**		**Naïve norm.**		**Fisher**		**KL divergence**
		Linear	Gaussian	Linear	Gaussian	Linear	Gaussian	
Mean correlation coefficient *r* (Variable #)	**1: Age**	0.062	0.430	0.467	0.371	0.455	0.442	**0.468**
	12: ReadEng_Unadj	0.193	0.329	0.383	0.267	**0.399**	0.396	0.357
	13: ReadEng_AgeAdj	0.200	0.318	0.38	0.252	**0.401**	0.400	0.351
	14: PicVocab_Unadj	0.131	0.288	0.346	0.242	**0.392**	0.388	0.314
	15: PicVocab_AgeAdj	0.189	0.330	0.372	0.245	**0.420**	0.419	0.338
	18: VSPLOT_TC	0.227	0.282	**0.320**	0.228	0.305	0.295	0.248
	20: VSPLOT_OFF	0.223	0.277	0.342	0.296	**0.362**	0.352	0.235
	35: WM_Task_Acc	0.234	0.254	0.341	0.269	**0.375**	0.370	0.283

4 Discussion

We here proposed the use of the Fisher kernel to predict traits from a neuroimaging model of brain dynamics, showing that it outperforms other methods in terms of two outcome criteria: prediction accuracy and reliability. With this approach, certain traits, such as language, and spatial orientation skills, could be predicted at relatively high accuracy using a kernel based on a model of brain dynamics. We argue that, overall, the Fisher kernel is a suitable method to combine a generative model of brain dynamics with a predictive model, because it preserves the structure of the underlying brain dynamics model. We have compared the

Fisher kernel to kernels which ignore the structure of the brain dynamics model as well as to a previously used model based on Kullback-Leibler divergence. Although we here focussed on fMRI, the method is similarly applicable to other modalities like MEG or EEG.

We also compared linear and Gaussian versions of all kernels. Only the naïve kernel was on average improved by using its Gaussian version. In the other kernels, we found that although their Gaussian versions may in some runs predict more accurately, they are also more error-prone and susceptible to changes in the training set. Implementing Gaussian kernels in a predictive model is also computationally more expensive, making them less practical. This makes the linear Fisher kernel the recommended choice for producing accurate, robust, and reliable results. The Fisher kernel also outperformed the baseline Kullback-Leibler divergence model in our outcome criteria, highlighting the potential of well-defined kernel methods when combining generative models of brain dynamics with prediction models. Our results were consistent when using less data per subject and when modelling brain dynamics only in terms of covariance, which shows that the Fisher kernel also outperforms other kernels in different settings. Importantly, the overall lower accuracies when using only one scanning session consisting of 1,200 timepoints per subject indicate, however, that predicting phenotypes from models of brain dynamics with relatively high accuracy requires a large amount of data for each individual. While we here only focussed on linear predictions using a kernel regression model, the Fisher kernel can straightforwardly be implemented in any kernel-based prediction model or classifier, such as support vector machines (SVM), kernel fisher discriminant analysis (k-FDA), kernel logistic regression (KLR), or nonlinear PCA.

In summary, we have demonstrated that constructing a “proper” kernel, like the Fisher kernel, can achieve better, more reliable performance than other models when using models of brain dynamics to predict subject traits. Such a kernel can be useful in various applications. For instance, there is growing interest in combining different data types or modalities, such as structural, static, and dynamic measures to predict phenotypes (Engemann et al., 2020; Schouten et al., 2016). While directly combining the features from each modality can be problematic, kernels that are properly derived for each modality can be easily combined using approaches such as Multi Kernel Learning (MKL) (Gönen & Alpaydın, 2011), which can improve prediction accuracy of multimodal studies (Vaghari et al., 2022). In a clinical context, being able to describe individual variability for diagnosis and individual patients’ outcome prediction is a major goal (Marquand et al., 2016; Stephan et al., 2017; Wen et al., 2022; Wolfers et al., 2015). In order to be able to use these predictive models in a clinical context, we have to ensure that models are reliable, particularly if we want to interpret model errors, as in models of “brain age”. We here showed that, despite estimating

model performance in a standard cross-validated fashion, there can be extreme errors and large variation in prediction accuracies in some predictive models, while models using the Fisher kernel produce reliable results. Developing such reliable models should therefore be a key focus of neuroimaging-based prediction studies.

Acknowledgements

DV is supported by a Novo Nordisk Foundation Emerging Investigator Fellowship (NNF19OC-0054895) and an ERC Starting Grant (ERC-StG-2019-850404). We thank Ben Griffin for useful discussions and technical collaboration.

References

Ahrends, C., Stevner, A., Pervaiz, U., Kringelbach, M. L., Vuust, P., Woolrich, M. W., & Vidaurre, D. (2022). Data and model considerations for estimating time-varying functional connectivity in fMRI. *Neuroimage*, *252*, 119026. https://doi.org/10.1016/j.neuroimage.2022.119026

Azim, T., & Ahmed, S. (2018). *Composing Fisher Kernels from Deep Neural Models*. Springer.

Baum, L. E., & Eagon, J. A. (1967). An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bulletin of the American Mathematical Society*, *73*(3), 360-363, 364. https://doi.org/

Baum, L. E., & Petrie, T. (1966). Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *The Annals of Mathematical Statistics*, *37*(6), 1554-1563, 1510. https://doi.org/10.1214/aoms/1177699147

Beckmann, C. F. (2012). Modelling with independent components. *Neuroimage*, *62*(2), 891-901. https://doi.org/10.1016/j.neuroimage.2012.02.020

Beckmann, C. F., Mackay, C. E., Filippini, N., & Smith, S. M. (2009). Group comparison of resting-state FMRI data using multi-subject ICA and dual regression. *Neuroimage*, *47*, S148. https://doi.org/10.1016/S1053-8119(09)71511-3

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289-300. http://www.jstor.org/stable/2346101

Cole, J. H., & Franke, K. (2017). Predicting Age Using Neuroimaging: Innovative Brain Ageing Biomarkers. *Trends Neurosci*, *40*(12), 681-690. https://doi.org/10.1016/j.tins.2017.10.001

Denissen, S., Engemann, D. A., De Cock, A., Costers, L., Baijot, J., Laton, J., Penner, I. K., Grothe, M., Kirsch, M., D'Hooghe M, B., D'Haeseleer, M., Dive, D., De Mey, J., Van Schependom, J., Sima, D. M., & Nagels, G. (2022). Brain age as a surrogate marker for cognitive performance in multiple sclerosis. *Eur J Neurol*, *29*(10), 3039-3049. https://doi.org/10.1111/ene.15473

Do, M. N. (2003). Fast approximation of Kullback-Leibler distance for dependence trees and hidden Markov models [Article]. *IEEE Signal Processing Letters*, *10*(4), 115-118. https://doi.org/10.1109/LSP.2003.809034

Engemann, D. A., Kozynets, O., Sabbagh, D., Lemaître, G., Varoquaux, G., Liem, F., & Gramfort, A. (2020). Combining magnetoencephalography with magnetic resonance

imaging enhances learning of surrogate-biomarkers. *eLife*, *9*, e54055. https://doi.org/10.7554/eLife.54055

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., Van Essen, D. C., Jenkinson, M., & Consortium, W. U.-M. H. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage*, *80*, 105-124. https://doi.org/10.1016/j.neuroimage.2013.04.127

Greene, A. S., Shen, X., Noble, S., Horien, C., Hahn, C. A., Arora, J., Tokoglu, F., Spann, M. N., Carrión, C. I., Barron, D. S., Sanacora, G., Srihari, V. H., Woods, S. W., Scheinost, D., & Constable, R. T. (2022). Brain–phenotype models fail for individuals who defy sample stereotypes. *Nature*, *609*(7925), 109-118. https://doi.org/10.1038/s41586-022-05118-w

Griffanti, L., Salimi-Khorshidi, G., Beckmann, C. F., Auerbach, E. J., Douaud, G., Sexton, C. E., Zsoldos, E., Ebmeier, K. P., Filippini, N., Mackay, C. E., Moeller, S., Xu, J., Yacoub, E., Baselli, G., Ugurbil, K., Miller, K. L., & Smith, S. M. (2014). ICA-based artefact removal and accelerated fMRI acquisition for improved resting state network imaging. *Neuroimage*, *95*, 232-247. https://doi.org/10.1016/j.neuroimage.2014.03.034

Gönen, M., & Alpaydın, E. (2011). Multiple kernel learning algorithms. *The Journal of Machine Learning Research*, *12*, 2211-2268. http://jmlr.org/papers/v12/gonen11a.html

Jaakkola, T., Diekhans, M., & Haussler, D. (1999). Using the Fisher kernel method to detect remote protein homologies. *Proc Int Conf Intell Syst Mol Biol*, 149-158.

Jaakkola, T., & Haussler, D. (1998). Exploiting Generative Models in Discriminative Classifiers. NIPS,

Lurie, D. J., Kessler, D., Bassett, D. S., Betzel, R. F., Breakspear, M., Kheilholz, S., Kucyi, A., Liégeois, R., Lindquist, M. A., McIntosh, A. R., Poldrack, R. A., Shine, J. M., Thompson, W. H., Bielczyk, N. Z., Douw, L., Kraft, D., Miller, R. L., Muthuraman, M., Pasquini, L., . . . Calhoun, V. D. (2019). Questions and controversies in the study of time-varying functional connectivity in resting fMRI. *Network Neuroscience*, *4*(1), 30-69. https://doi.org/10.1162/netn_a_00116

MacKay, D. J. C., Kay, D. J. C. M., & Press, C. U. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press. https://books.google.dk/books?id=AKuMj4PN_EMC

Marquand, A. F., Rezek, I., Buitelaar, J., & Beckmann, C. F. (2016). Understanding Heterogeneity in Clinical Cohorts Using Normative Models: Beyond Case-Control Studies. *Biological psychiatry*, *80*(7), 552-561. https://doi.org/10.1016/j.biopsych.2015.12.023

Salimi-Khorshidi, G., Douaud, G., Beckmann, C. F., Glasser, M. F., Griffanti, L., & Smith, S. M. (2014). Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers. *Neuroimage*, *90*, 449-468. https://doi.org/10.1016/j.neuroimage.2013.11.046

Saunders, C., Gammerman, A., & Vovk, V. (1998). Ridge regression learning algorithm in dual variables.

Schouten, T. M., Koini, M., de Vos, F., Seiler, S., van der Grond, J., Lechner, A., Hafkemeijer, A., Möller, C., Schmidt, R., de Rooij, M., & Rombouts, S. A. R. B. (2016). Combining anatomical, diffusion, and resting state functional magnetic resonance imaging for individual classification of mild and moderate Alzheimer's disease. *NeuroImage: Clinical*, *11*, 46-51. https://doi.org/10.1016/j.nicl.2016.01.002

Schölkopf, B., Smola, A. J., & Bach, F. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press. https://doi.org/10.1017/CBO9780511809682

Smith, S. M., Beckmann, C. F., Andersson, J., Auerbach, E. J., Bijsterbosch, J., Douaud, G., Duff, E., Feinberg, D. A., Griffanti, L., Harms, M. P., Kelly, M., Laumann, T., Miller, K.

L., Moeller, S., Petersen, S., Power, J., Salimi-Khorshidi, G., Snyder, A. Z., Vu, A. T., . . . Consortium, W. U.-M. H. (2013). Resting-state fMRI in the Human Connectome Project. *Neuroimage*, *80*, 144-168. https://doi.org/10.1016/j.neuroimage.2013.05.039

Smith, S. M., Vidaurre, D., Alfaro-Almagro, F., Nichols, T. E., & Miller, K. L. (2019). Estimation of brain age delta from brain imaging. *Neuroimage*, *200*, 528-539. https://doi.org/https://doi.org/10.1016/j.neuroimage.2019.06.017

Stephan, K. E., Schlagenhauf, F., Huys, Q. J. M., Raman, S., Aponte, E. A., Brodersen, K. H., Rigoux, L., Moran, R. J., Daunizeau, J., Dolan, R. J., Friston, K. J., & Heinz, A. (2017). Computational neuroimaging strategies for single patient predictions. *Neuroimage*, *145*, 180-199. https://doi.org/https://doi.org/10.1016/j.neuroimage.2016.06.038

Vaghari, D., Kabir, E., & Henson, R. N. (2022). Late combination shows that MEG adds to MRI in classifying MCI versus controls. *Neuroimage*, *252*, 119054. https://doi.org/10.1016/j.neuroimage.2022.119054

van der Maaten, L. (2011). Learning Discriminative Fisher Kernels. Proceedings of the 28th International Conference on Machine Learning, Bellevue, WA, USA.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., & Ugurbil, K. (2013). The WU-Minn Human Connectome Project: an overview. *Neuroimage*, *80*, 62-79. https://doi.org/10.1016/j.neuroimage.2013.05.041

Van Essen, D. C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T. E. J., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S. W., Della Penna, S., Feinberg, D., Glasser, M. F., Harel, N., Heath, A. C., Larson-Prior, L., Marcus, D., Michalareas, G., Moeller, S., . . . Consortium, W. U.-M. H. (2012). The Human Connectome Project: a data acquisition perspective. *Neuroimage*, *62*(4), 2222-2231. https://doi.org/10.1016/j.neuroimage.2012.02.018

Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., & Thirion, B. (2017). Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *Neuroimage*, *145*, 166-179. https://doi.org/https://doi.org/10.1016/j.neuroimage.2016.10.038

Vidaurre, D., Llera, A., Smith, S. M., & Woolrich, M. W. (2021). Behavioural relevance of spontaneous, transient brain network interactions in fMRI. *Neuroimage*, *229*, 117713. https://doi.org/10.1016/j.neuroimage.2020.117713

Vidaurre, D., Quinn, A. J., Baker, A. P., Dupret, D., Tejero-Cantero, A., & Woolrich, M. W. (2016). Spectrally resolved fast transient brain states in electrophysiological data. *Neuroimage*, *126*, 81-95. https://doi.org/10.1016/j.neuroimage.2015.11.047

Vidaurre, D., Smith, S. M., & Woolrich, M. W. (2017). Brain network dynamics are hierarchically organized in time. *Proc Natl Acad Sci U S A*, *114*(48), 12827-12832. https://doi.org/10.1073/pnas.1705120114

Wen, J., Varol, E., Sotiras, A., Yang, Z., Chand, G. B., Erus, G., Shou, H., Abdulkadir, A., Hwang, G., Dwyer, D. B., Pigoni, A., Dazzan, P., Kahn, R. S., Schnack, H. G., Zanetti, M. V., Meisenzahl, E., Busatto, G. F., Crespo-Facorro, B., Rafael, R. G., . . . Alzheimer's Disease Neuroimaging, I. (2022). Multi-scale semi-supervised clustering of brain images: Deriving disease subtypes. *Med Image Anal*, *75*, 102304. https://doi.org/10.1016/j.media.2021.102304

Winkler, A. M., Webster, M. A., Vidaurre, D., Nichols, T. E., & Smith, S. M. (2015). Multi-level block permutation. *Neuroimage*, *123*, 253-268. https://doi.org/https://doi.org/10.1016/j.neuroimage.2015.05.092

Wolfers, T., Buitelaar, J. K., Beckmann, C. F., Franke, B., & Marquand, A. F. (2015). From estimating activation locality to predicting disorder: A review of pattern recognition for neuroimaging-based psychiatric diagnostics. *Neuroscience & Biobehavioral Reviews*, *57*, 328-349. https://doi.org/10.1016/j.neubiorev.2015.08.001