Generative embedding enables model-based classification in fMRI

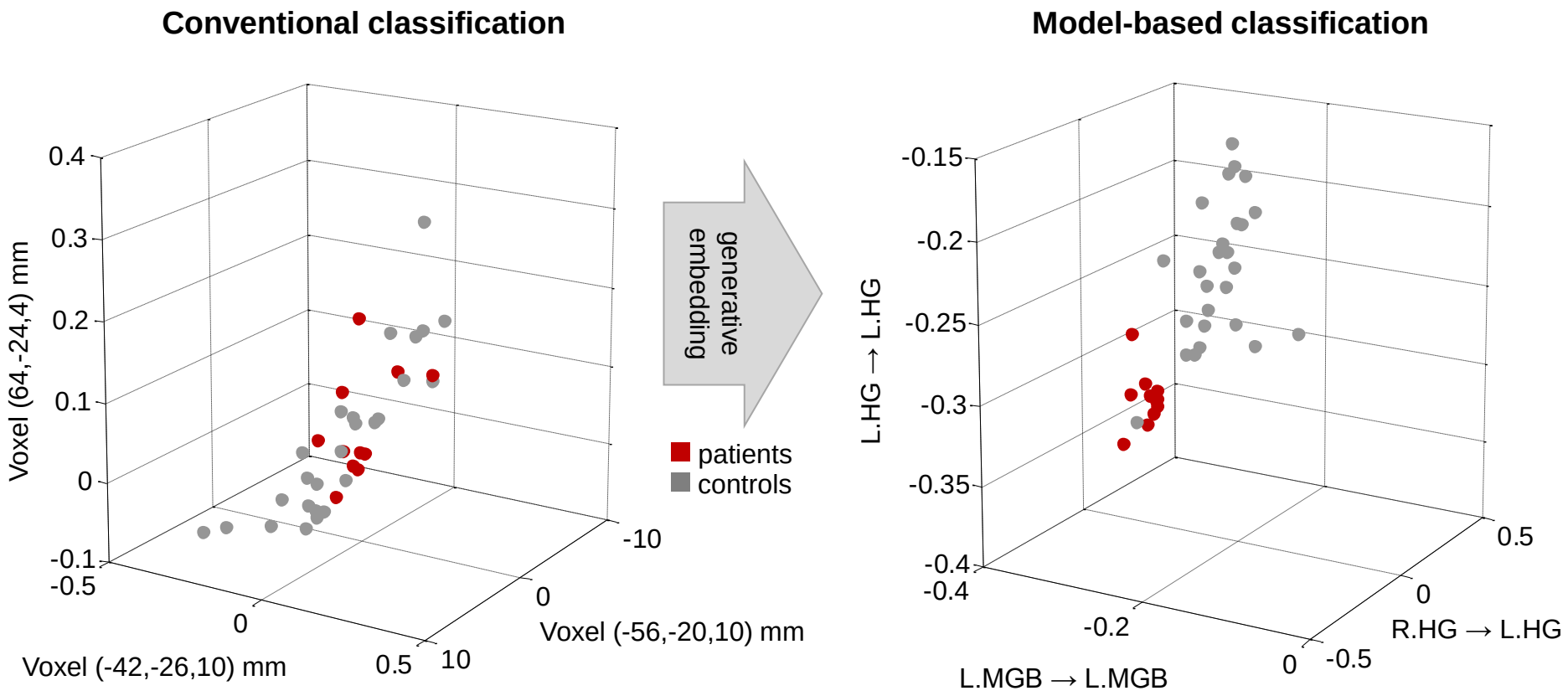
Kay Henning Brodersen

Computational Neuroeconomics Group
Department of Economics, University of Zurich

Machine Learning and Pattern Recognition Group
Department of Computer Science, ETH Zurich

http://people.inf.ethz.ch/bkay/

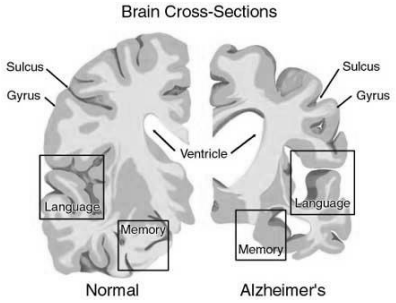
Conventional vs. model-based classification



Prediction & inference

The goal of **prediction** is to find a highly accurate encoding or decoding function.

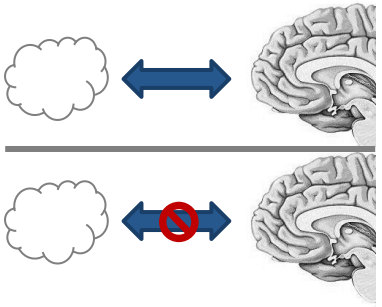
predicting a cognitive state using a brain-machine interface



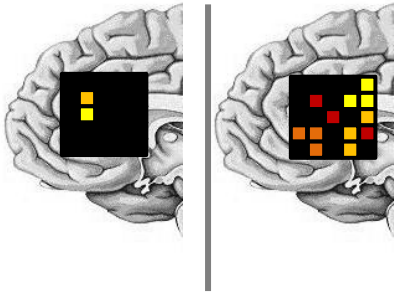
predicting a subject-specific diagnostic status

⇨ powerful discriminative algorithms for classification

The goal of **inference** is to decide between competing hypotheses about mechanisms or representations in the brain.



comparing a model that links distributed neuronal activity to a cognitive state with a model that does not



weighing the evidence for sparse coding vs. dense coding

⇨ mechanistically interpretable generative models of brain function

Dissecting disorders that are hard to dissect

Neurological and psychiatric spectrum disorders are typically defined in terms of particular symptom sets, despite increasing evidence that the same symptom may be caused by very different pathologies.

Can we learn what distinguishes different subgroups, and design an accurate prediction algorithm?

1. Due to the high data dimensionality, algorithms struggle to separate informative from uninformative features, resulting in poor generalization performance.

2. Popular off-the-shelf classifiers may allow for inference on voxel weights. But they are typically based on activity and do not afford connectivity-based mechanistic interpretability.

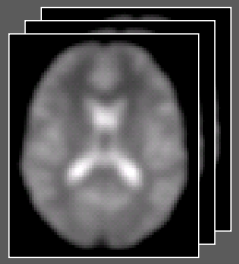
Data representations in classification analyses

Model-based classification

Can we exploit the rich discriminative information encoded in individual patterns of connection strengths?

Activation-based classification

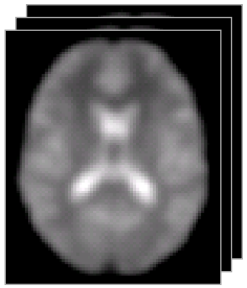
- depression
- schizophrenia
- mild cognitive impairment



Structure-based classification

- mild traumatic brain injury
- Alzheimer's disease
- autistic spectrum disorder
- frontotemporal dementia
- mild cognitive impairment
- schizophrenia
- aphasia

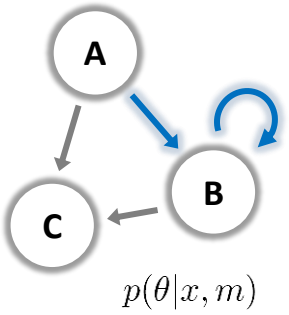
Generative embedding for fMRI



measurements from an individual subject

step 1 — model inversion

$\mathcal{X} \to \mathcal{M}_\Theta$



$p(\theta|x,m)$

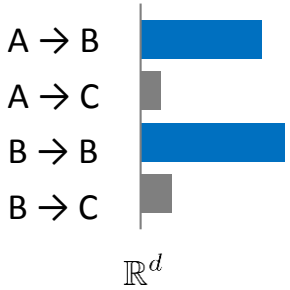
subject-specific inverted generative model

step 2 — kernel construction

$\mathcal{M}_\Theta \to \mathbb{R}^d$

$k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$

$k_{\mathcal{M}} : \mathcal{M}_\Theta \times \mathcal{M}_\Theta \to \mathbb{R}$

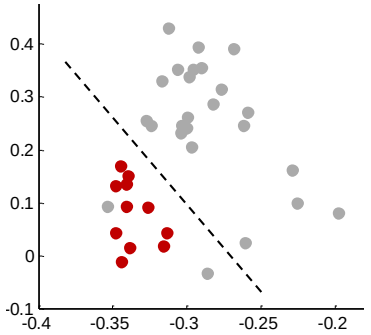


$\mathbb{R}^d$

subject representation in the generative score space

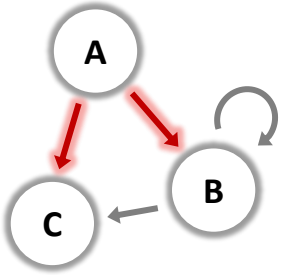
step 3 — classification

$$\hat{c} = \mathrm{sgn}\left(\sum_i^n \alpha_i^* k(x_i, x) + b^*\right)$$



separating hyperplane to discriminate between groups

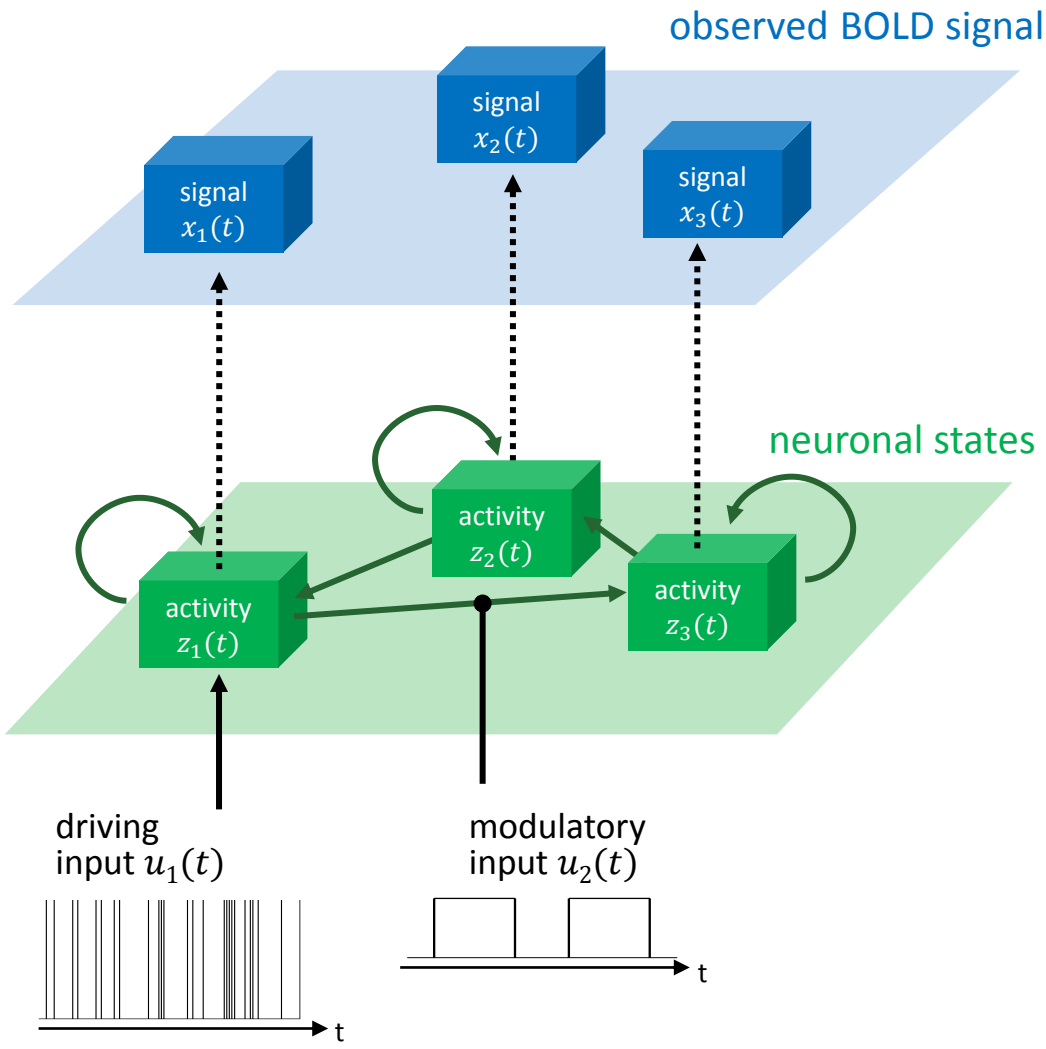
step 4 — interpretation



jointly discriminative connection strengths

Brodersen, Haiss, Ong, Jung, Tittgemeyer, Buhmann, Weber, Stephan (2010) *NeuroImage*
Brodersen, Schofield, Leff, Ong, Lomakina, Buhmann, Stephan (*under review*)

The generative model can be a dynamic causal model



haemodynamic forward model

$$x = g(z, \theta_h)$$

neural state equation

$$\dot{z} = (A + \sum u_j B^{(j)})z + Cu$$

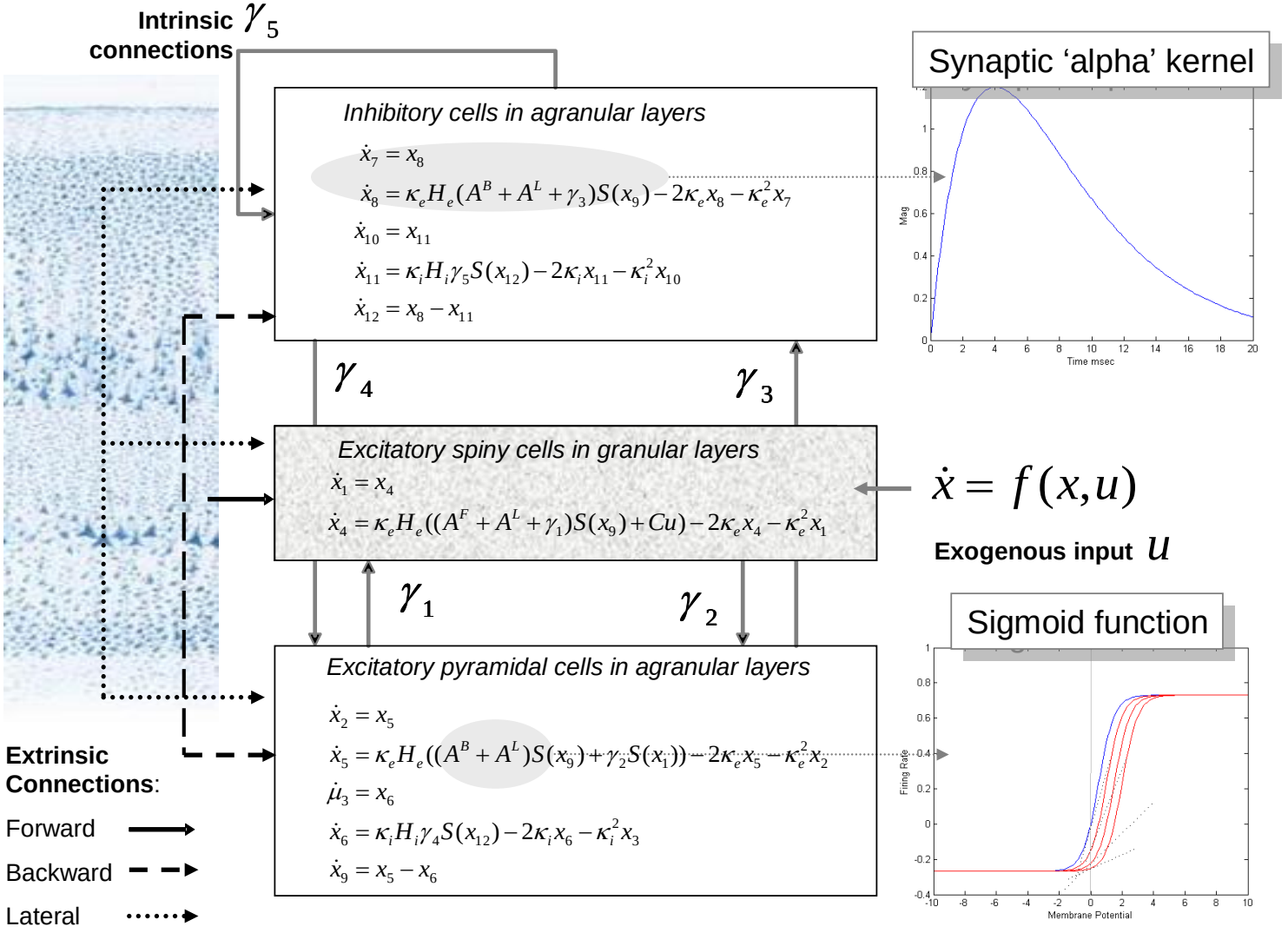
intrinsic connectivity $\rightarrow$ $A = \frac{\partial \dot{z}}{\partial z}$

modulation of connectivity $\rightarrow$ $B^{(j)} = \frac{\partial}{\partial u_j}\frac{\partial \dot{z}}{\partial z}$

direct inputs $\rightarrow$ $C = \frac{\partial \dot{z}}{\partial u}$

Jansen & Rit (1995) *Biological Cybernetics*
Friston, Harrison & Penny (2003) *NeuroImage*
Stephan & Friston (2007), *Handbook of Brain Connectivity*

The generative model can be a dynamic causal model



Moran et al. 2009 *NeuroImage*

Training and testing a model-based classifier

Training a kernel-based discriminant classifier:

$$\max_\alpha \mathcal{L}(\alpha) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j c_i c_j \, k(x_i, x_j) + \sum_{i=1}^{n}\alpha_i$$

$$s.t. \quad \sum_{i=1}^{n} c_i\alpha_i = 0$$

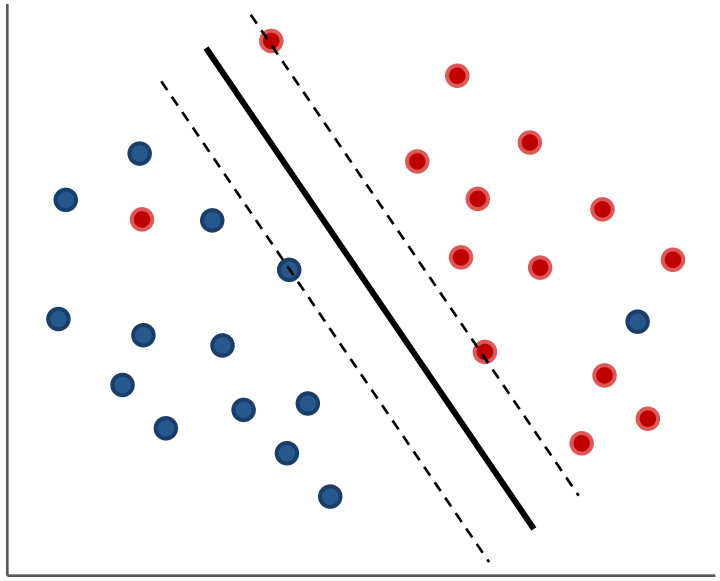
$$0 \le \alpha_i \le C \quad \forall i = 1,\ldots,n$$

Using the model to make predictions:

$$f(x_{n+1}) = \sum_{i=1}^{n}\alpha_i^* \, k(x_i, x_{n+1}) + b^*$$

$$\hat{c}_{n+1} := \mathrm{sgn}(f(x_{n+1}))$$

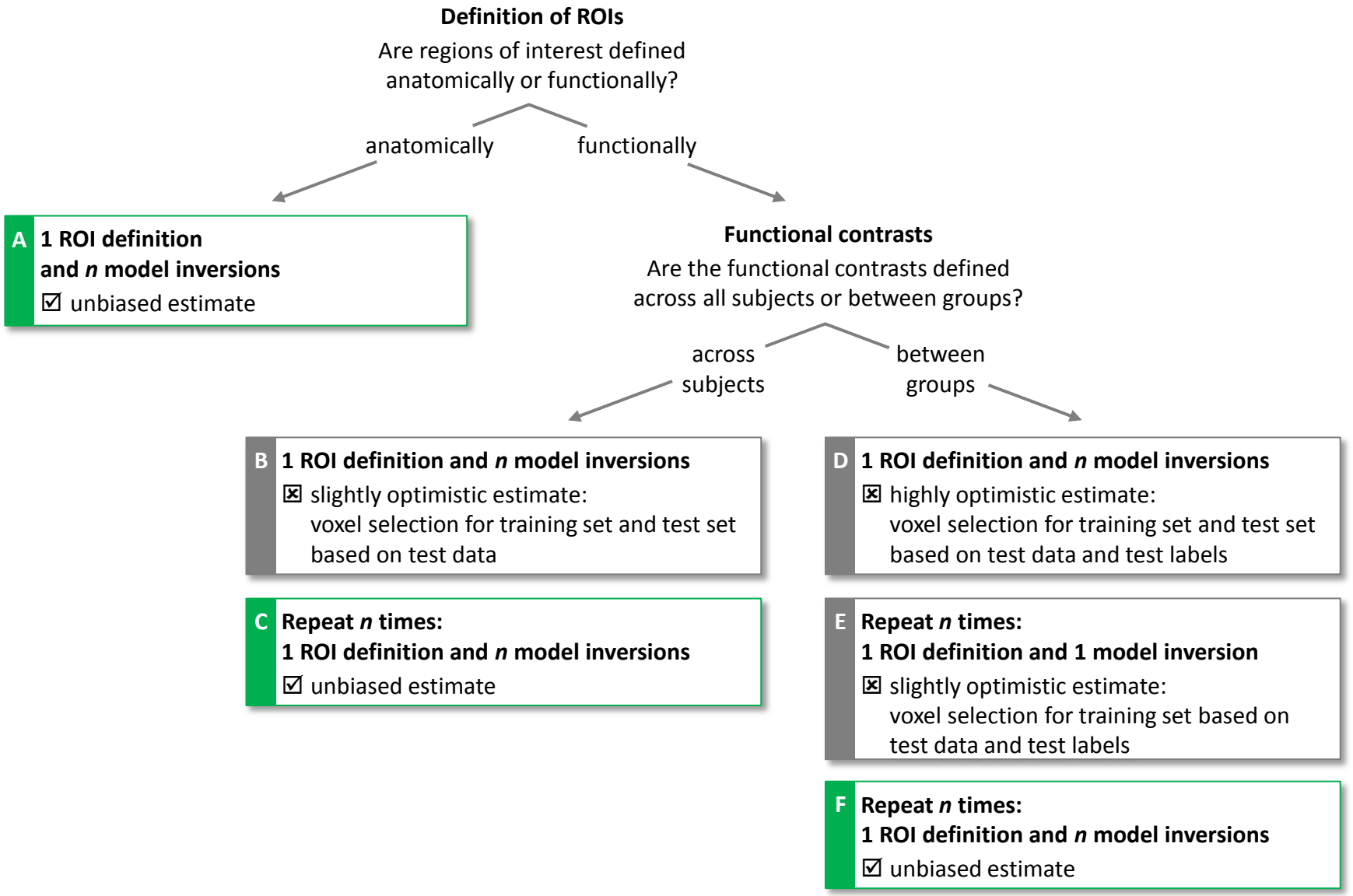
Linear SVM



In the case of generative embedding:

$k(x_i, x_j) = x_i^T x_j$

Specifying and inverting the model – how?



Full Bayesian approach to performance evaluation

Model

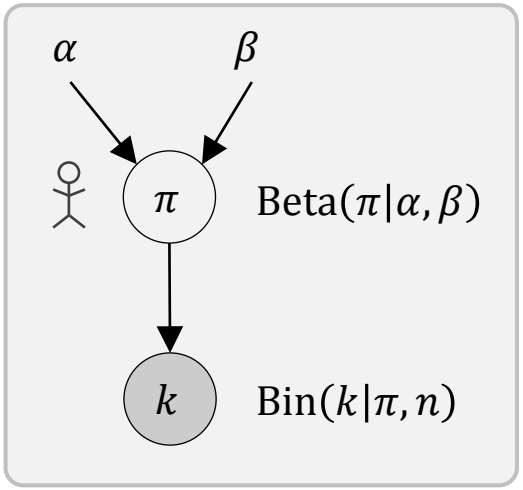
We model the likelihood function for k correct predictions as:

$$p(k|\pi, n) = \mathrm{Bin}(k|\pi, n)$$

The accuracy π can be modelled as a latent random variable with a conjugate Beta prior:

$$p(\pi|\alpha, \beta) = \mathrm{Beta}(\pi|\alpha, \beta)$$

This prior is uninformative when using the hyperparameters $\alpha = \beta = 1$.



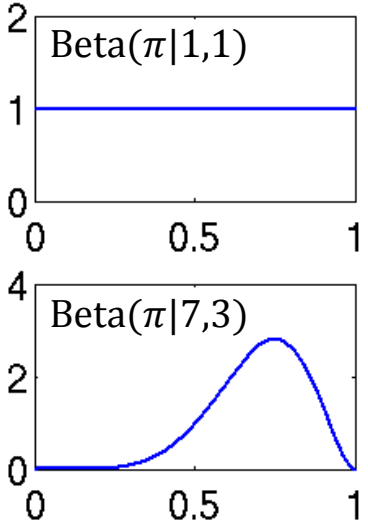
Inference

Inverting the model yields the posterior classification accuracy,

$$p(\pi|k, n, \alpha, \beta) = \mathrm{Beta}(\pi|\alpha + k, \beta + n - k),$$

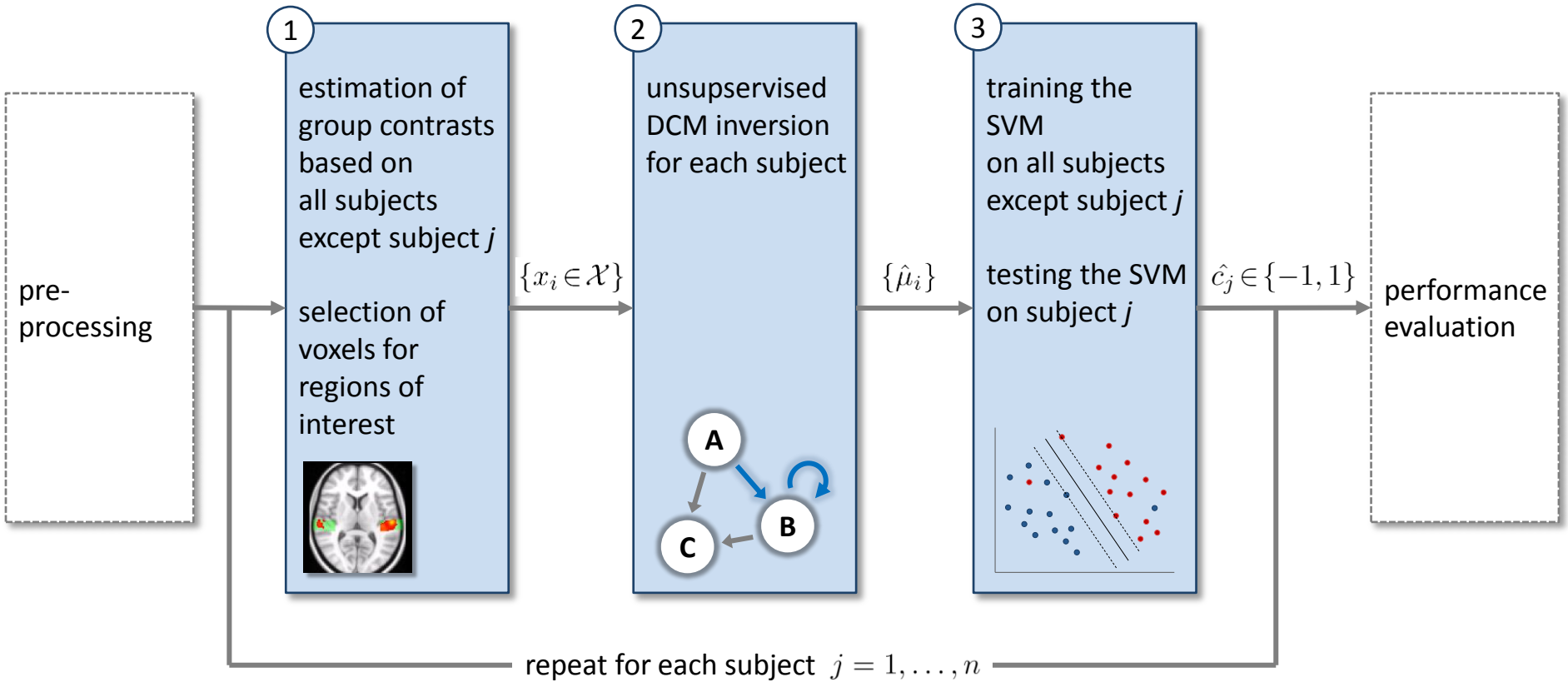
which we can summarize in various ways:

- expected accuracy: $\frac{k+1}{n+2}$
- MAP accuracy: $\frac{k}{n}$
- posterior interval: $[B^{-1}_{0.025}(k+1,n-k+1);\ B^{-1}_{0.975}(k+1,n-k+1)]$

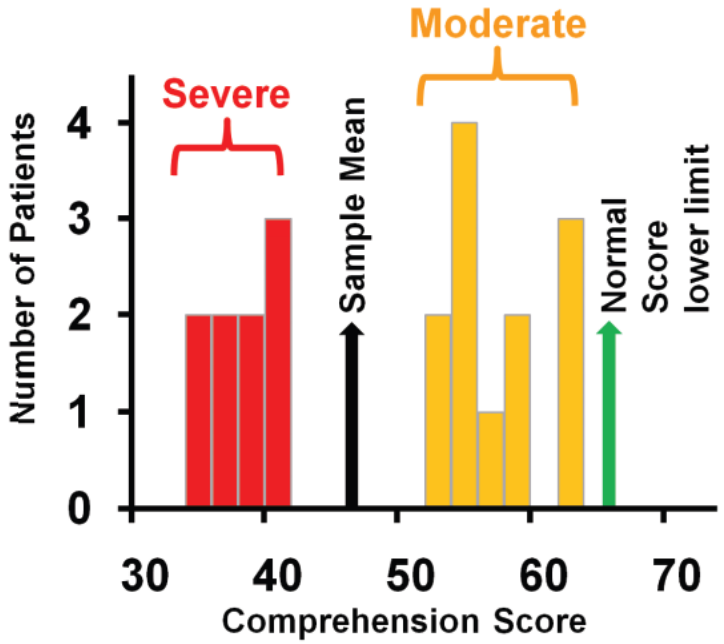


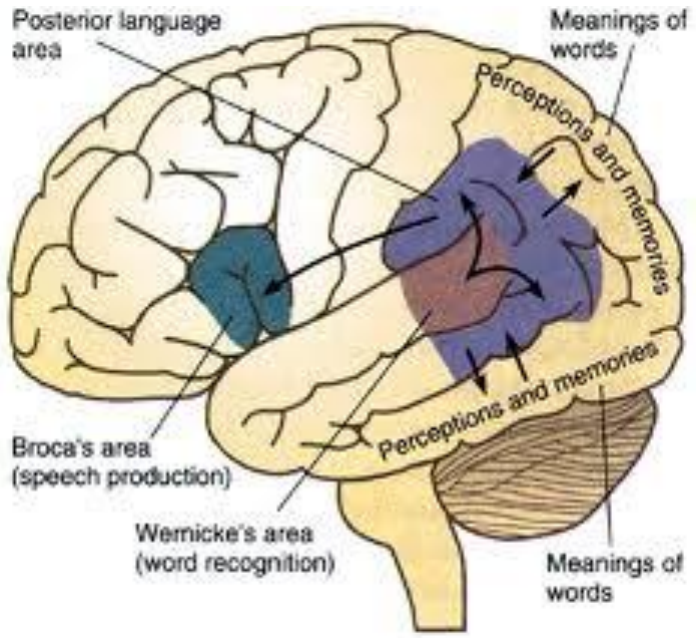
Brodersen, Chumbley, Mathys, Daunizeau, Ong, Buhmann & Stephan *(in preparation)*

Summary of the analysis

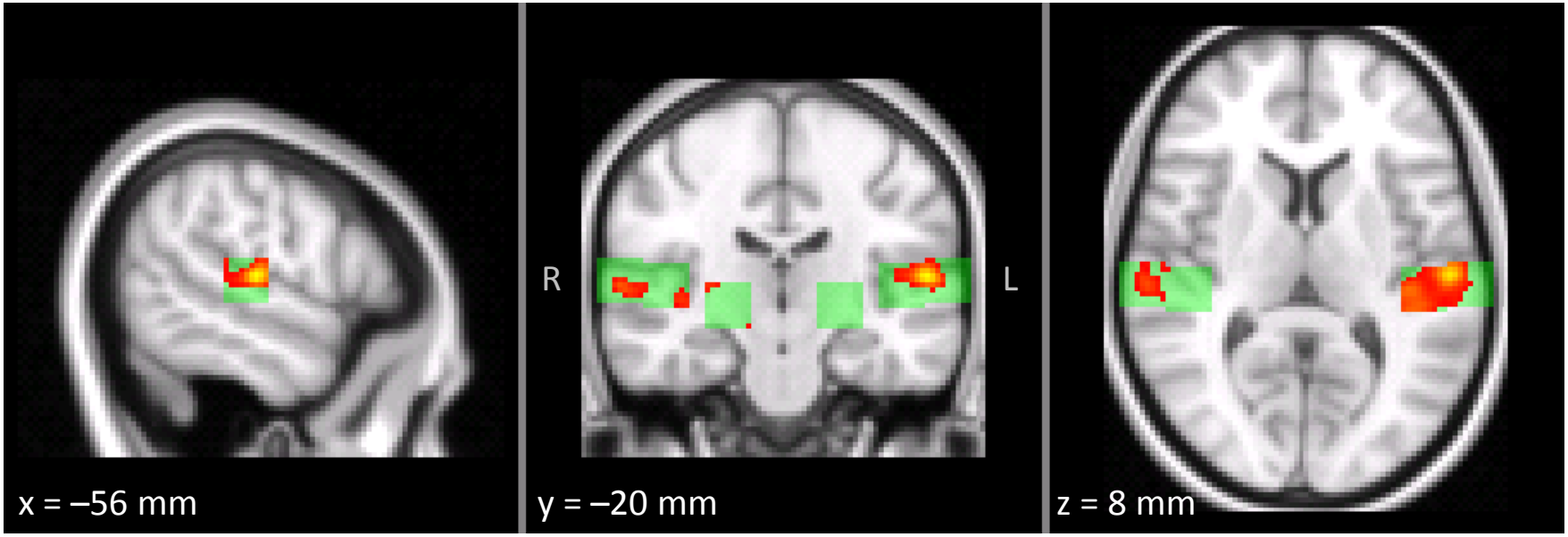


Example: diagnosis of moderate aphasia

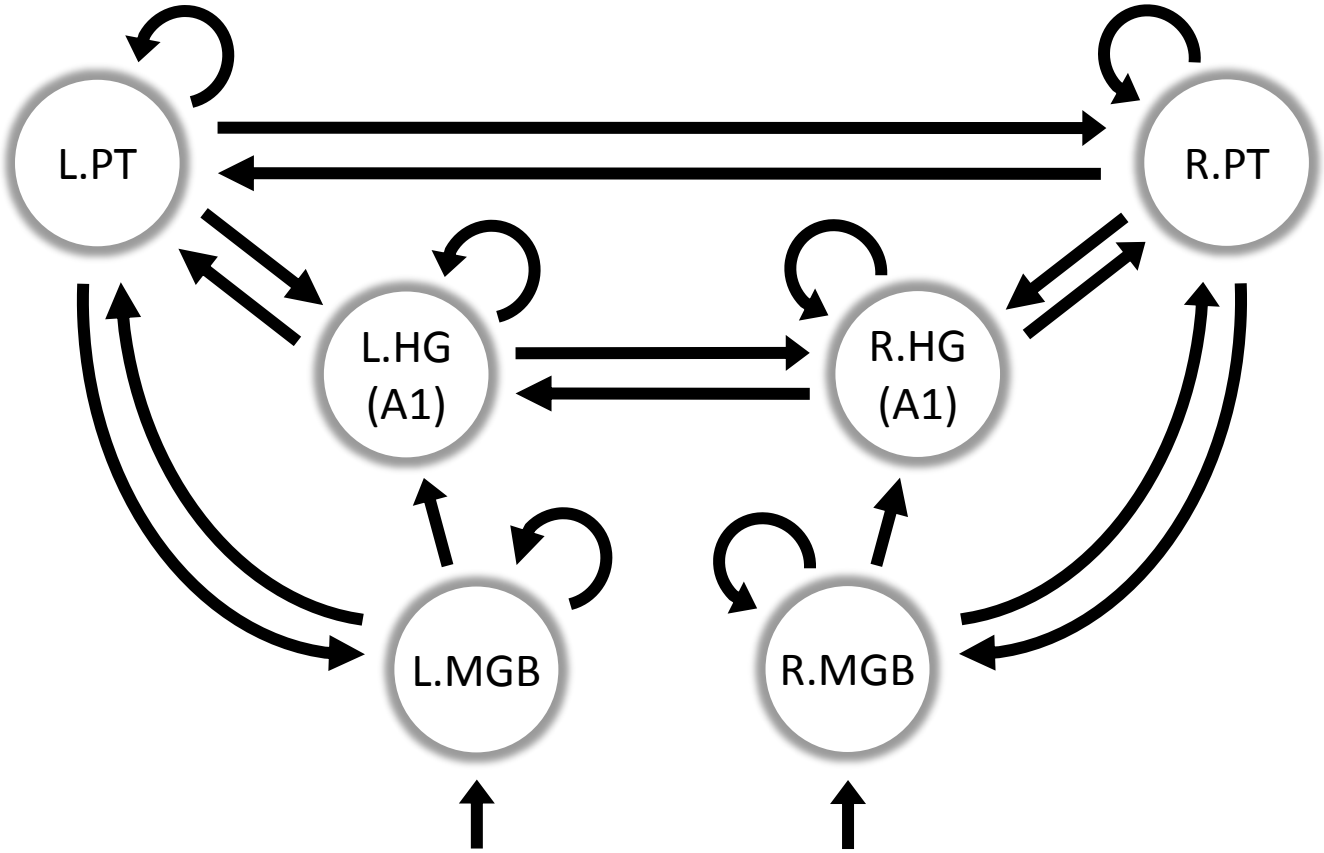




Regions of interest

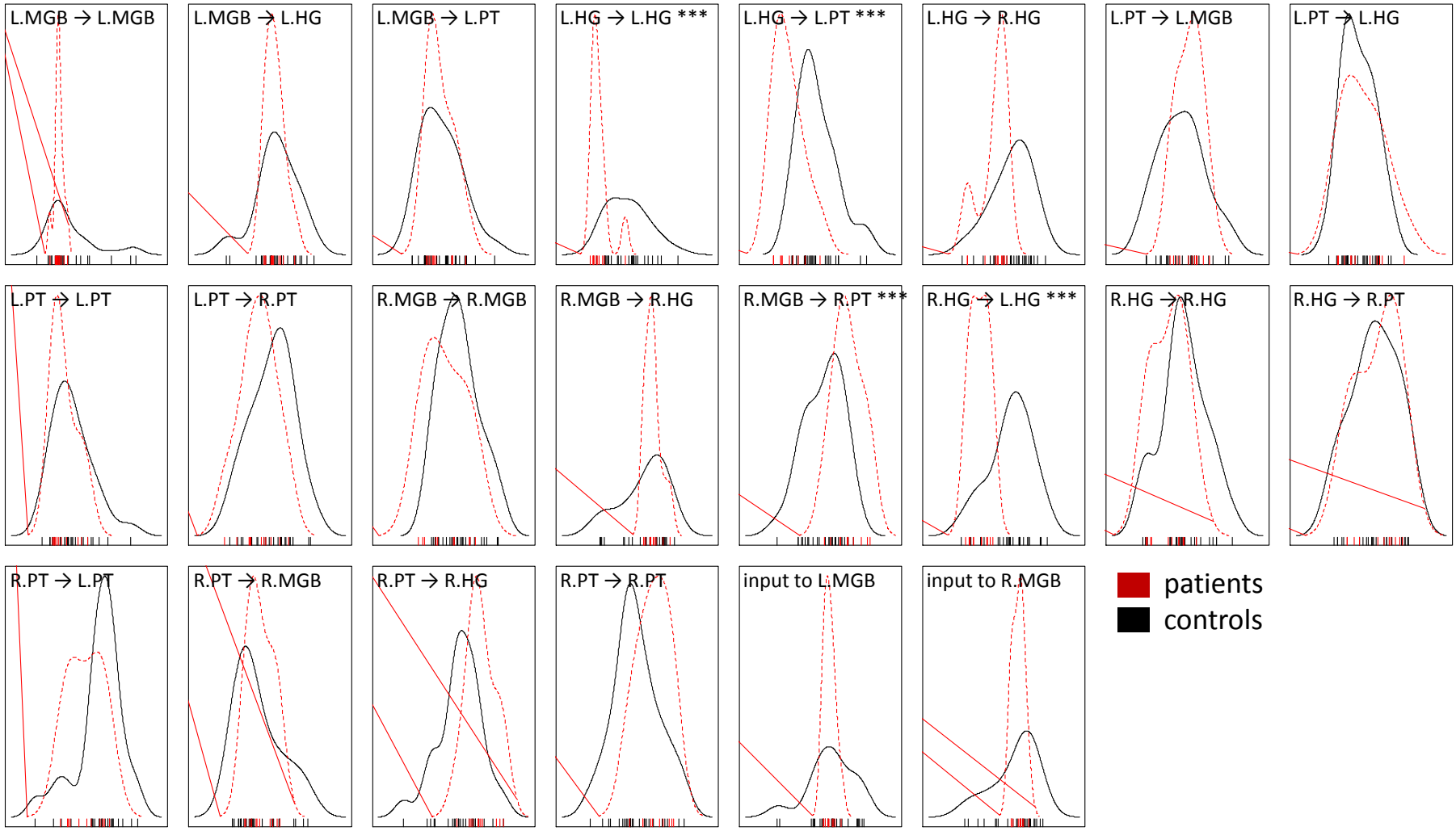


Neuronal model

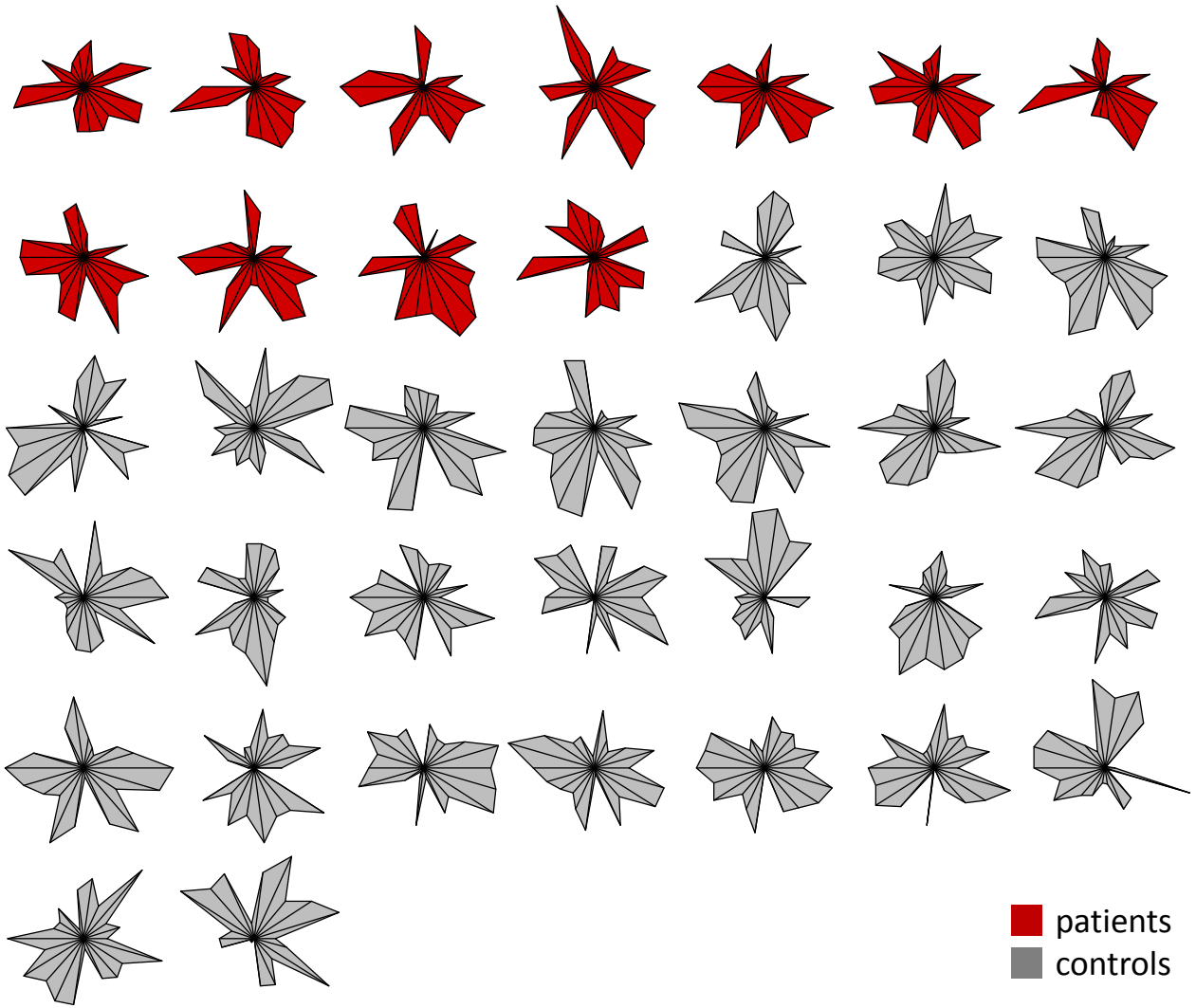


Schofield, Penny, Stephan, Crinion, Thompson, Price & Leff *(under review)*
Brodersen, Schofield, Leff, Ong, Lomakina, Buhmann & Stephan *(under review)*

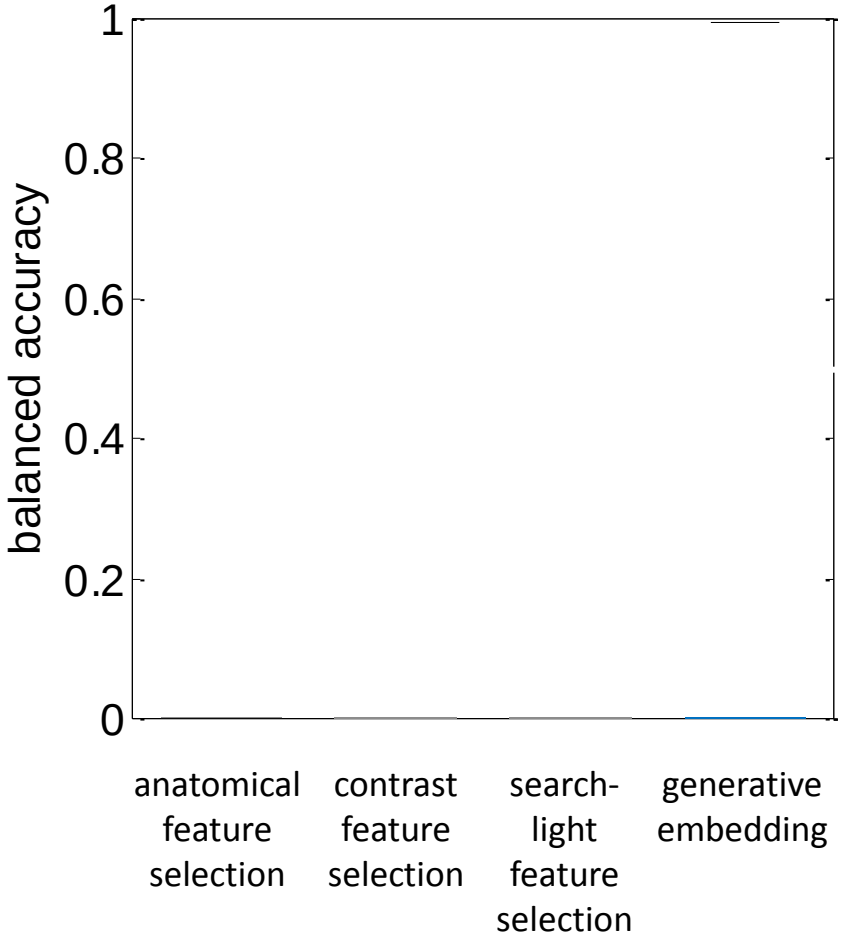
Univariate analysis

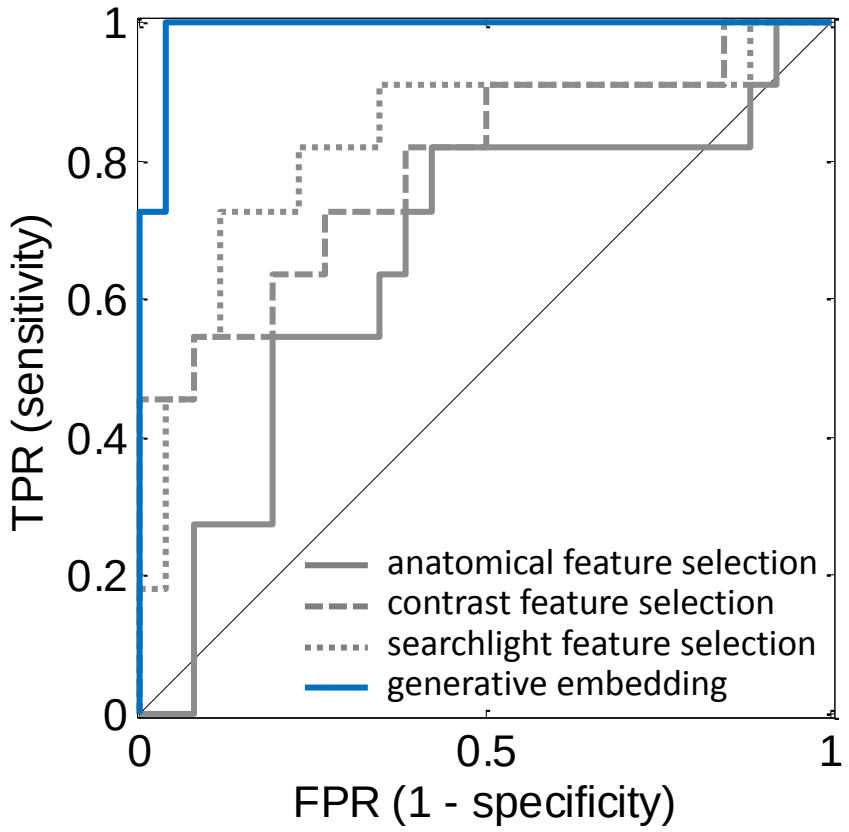


Connectional fingerprints



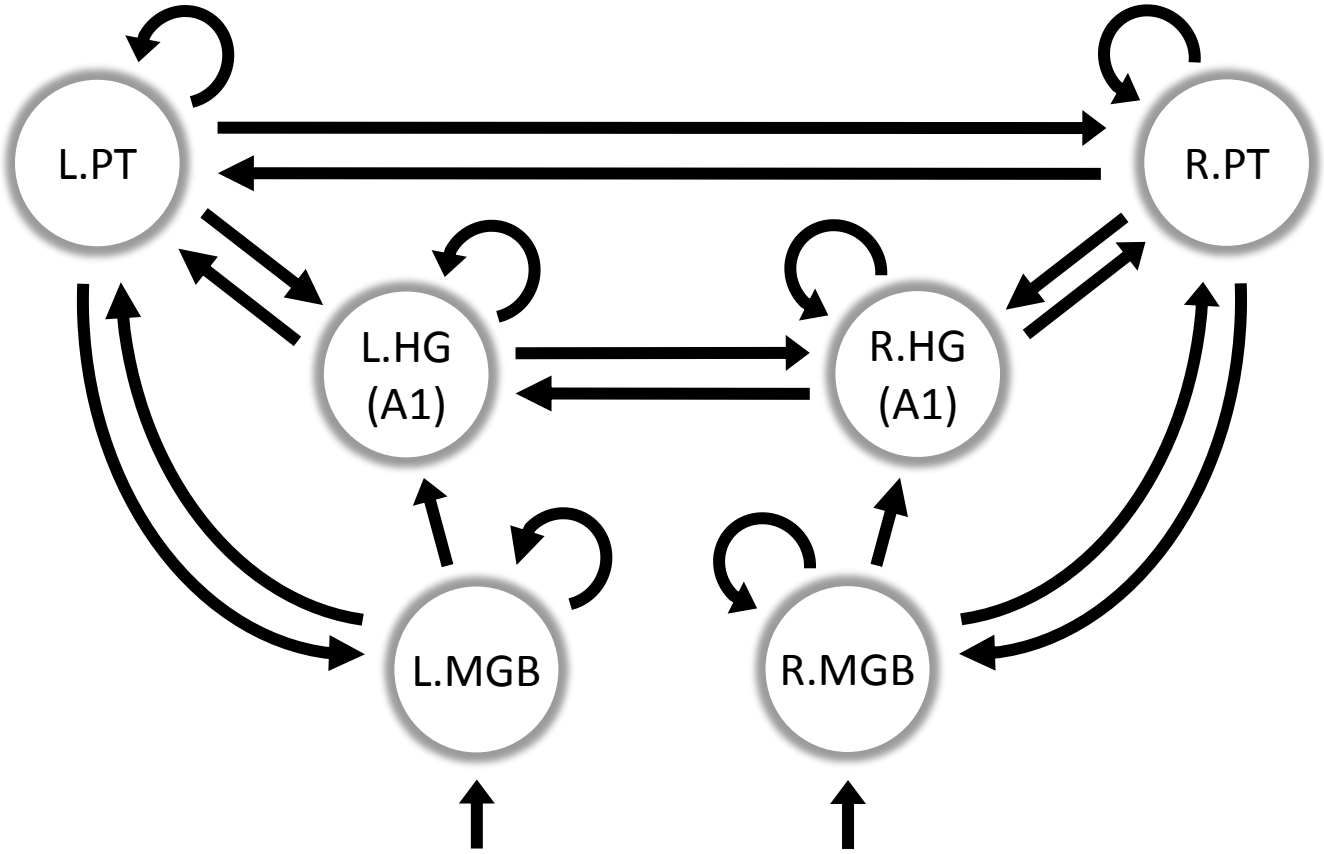
Classification performance





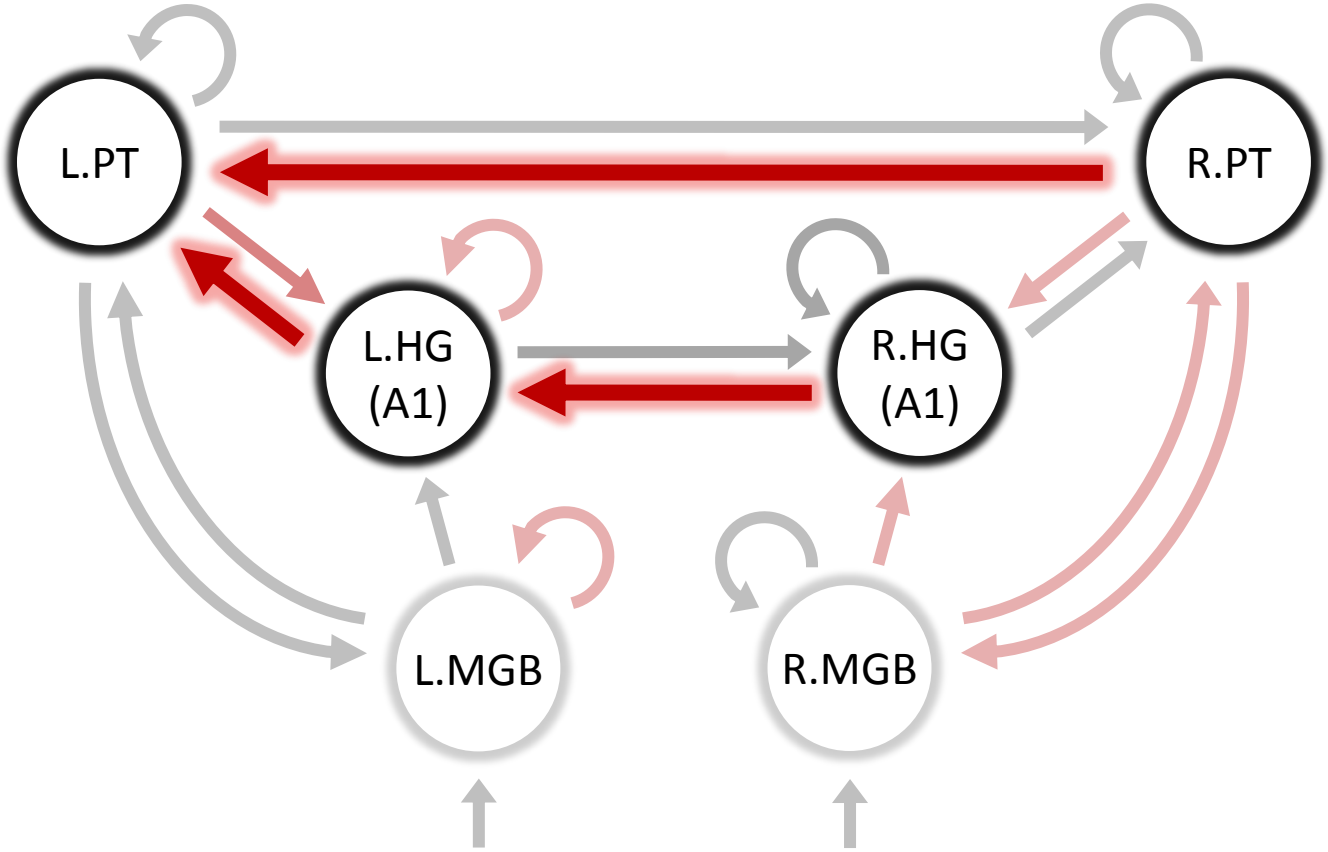
Brodersen, Schofield, Leff, Ong, Lomakina, Buhmann & Stephan *(under review)*

Discriminative features in model space



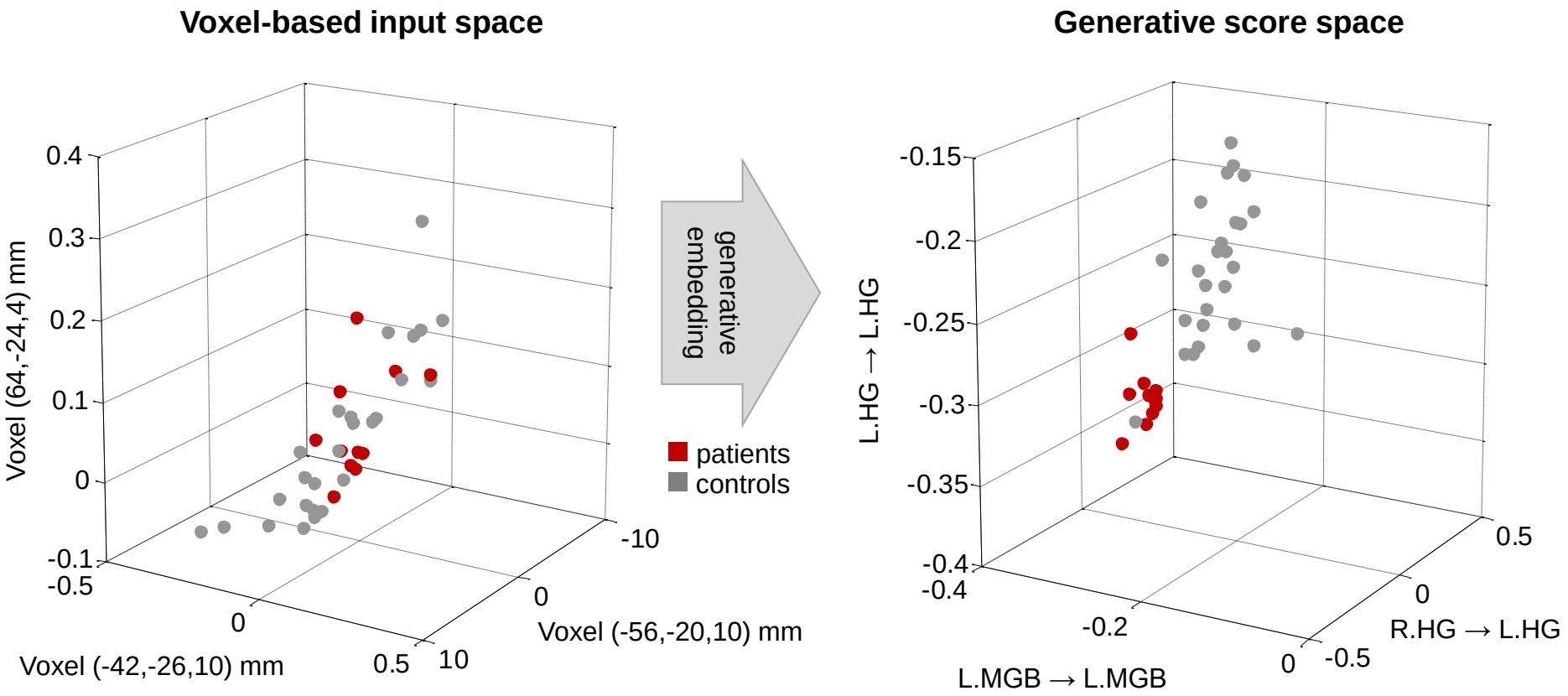
Brodersen, Schofield, Leff, Ong, Lomakina, Buhmann & Stephan *(under review)*

Discriminative features in model space



Brodersen, Schofield, Leff, Ong, Lomakina, Buhmann & Stephan *(under review)*

Illustration of the generative score space



Summary

1. **Strong classification performance**

   Generative embedding exploits the rich discriminative information encoded in 'hidden' quantities, such as coupling parameters. It may therefore outperform conventional schemes.

2. **Creation of a low-dimensional, interpretable feature space**

   The approach replaces high-dimensional fMRI data by a low-dimensional subject-specific fingerprint, where each dimension has a specific biological interpretation.

3. **Broad applicability**

   Generative embedding can be used both for trial-by-trial decoding (EEG, MEG, or LFP data) and for subject-by-subject classification analyses (fMRI data).

Thanks to …

Thomas Schofield
University College London

Klaas Enno Stephan
University of Zurich · University College London

Alexander Leff
University College London

Joachim M Buhmann
ETH Zurich

Cheng Soon Ong
ETH Zurich

Kate Lomakina
University of Zurich · ETH Zurich