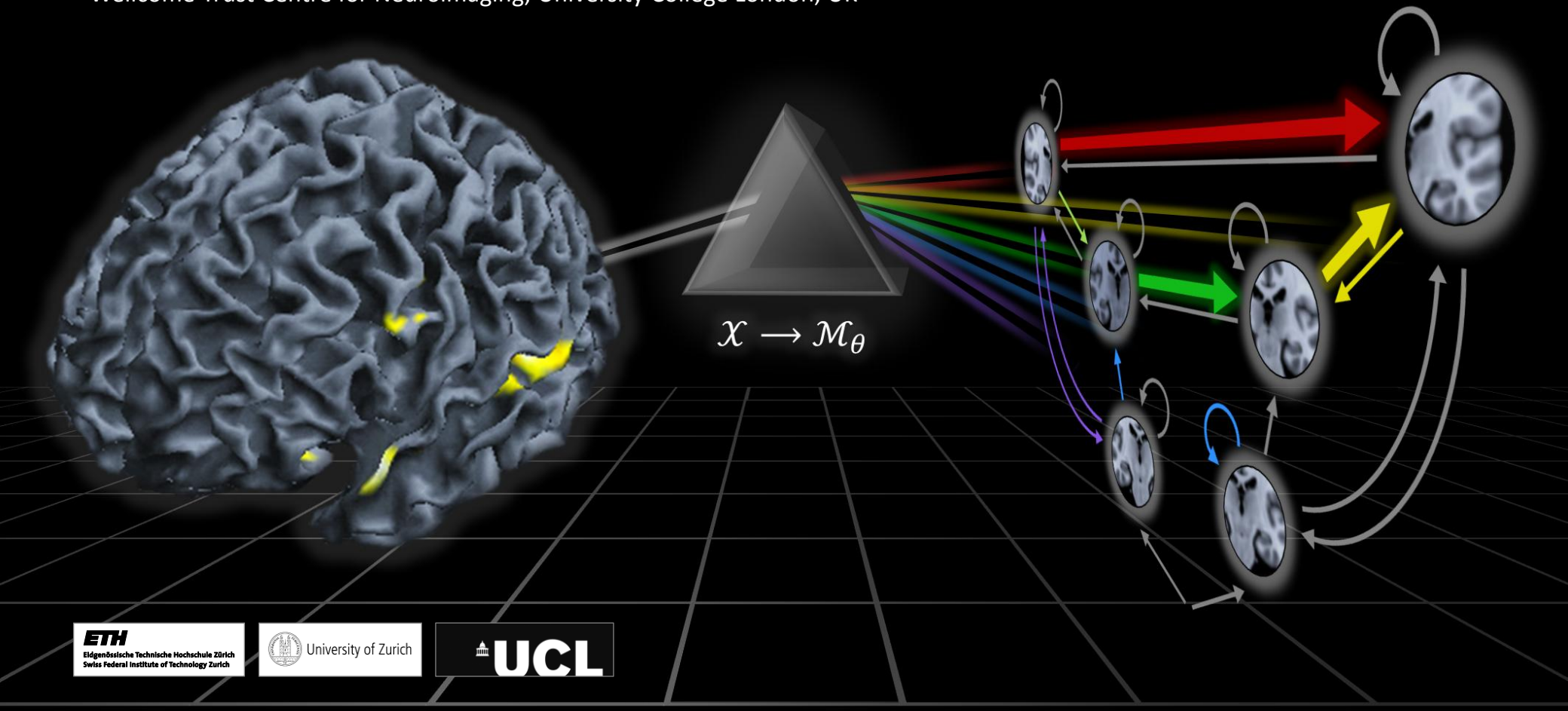
Generative embedding for model-based classification

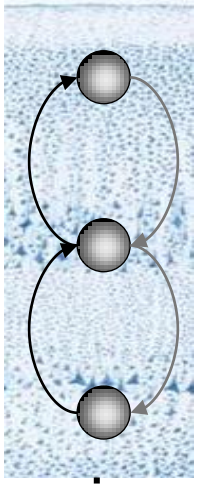
Kay H. Brodersen[1,2], Thomas M. Schofield[3], Alexander P. Leff[3], Cheng Soon Ong[1], Ekaterina I. Lomakina[1,2], Joachim M. Buhmann[1], Klaas E. Stephan[2,3]

[1] Machine Learning Laboratory, Department of Computer Science, ETH Zurich, Switzerland
[2] Laboratory for Social and Neural Systems Research, Department of Economics, University of Zurich, Switzerland
[3] Wellcome Trust Centre for Neuroimaging, University College London, UK

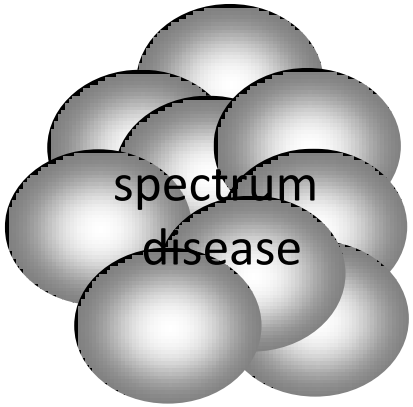


Model-based inference on individual pathophysiology

❶ model of neuronal (patho)physiology

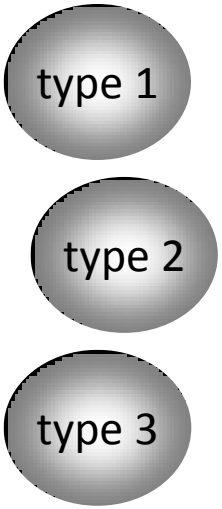


❷ application to brain activity data from individual patients



model-based
diagnostic tests

❸ diagnostic classification

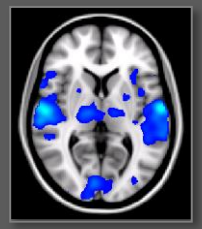


Classification approaches by data representation

Model-based classification

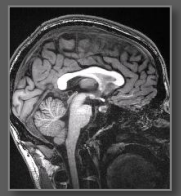
How do patterns of hidden quantities (e.g., connectivity among brain regions) differ between groups?

Activation-based classification



Which functional differences allow us to separate groups?

Structure-based classification



Which anatomical structures allow us to separate patients and healthy controls?

Colleagues & collaborators

Thomas Schofield
University College London

Joachim M Buhmann
ETH Zurich

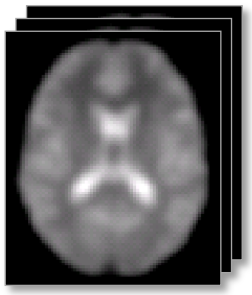
Cheng Soon Ong
ETH Zurich

Klaas Enno Stephan
University of Zurich · University College London

Kate Lomakina
University of Zurich · ETH Zurich

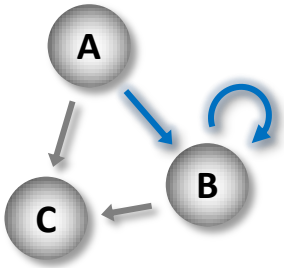
Alexander Leff
University College London

Model-based classification through generative embedding



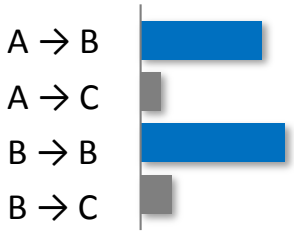
measurements from
an individual subject

step 1 —
model inversion



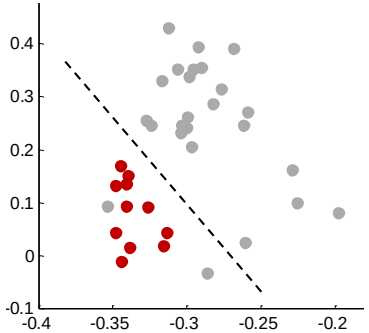
subject-specific
inverted generative model

step 2 —
kernel construction



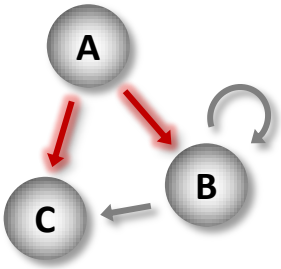
subject representation in the
generative score space

step 3 —
classification



separating hyperplane to
discriminate between groups

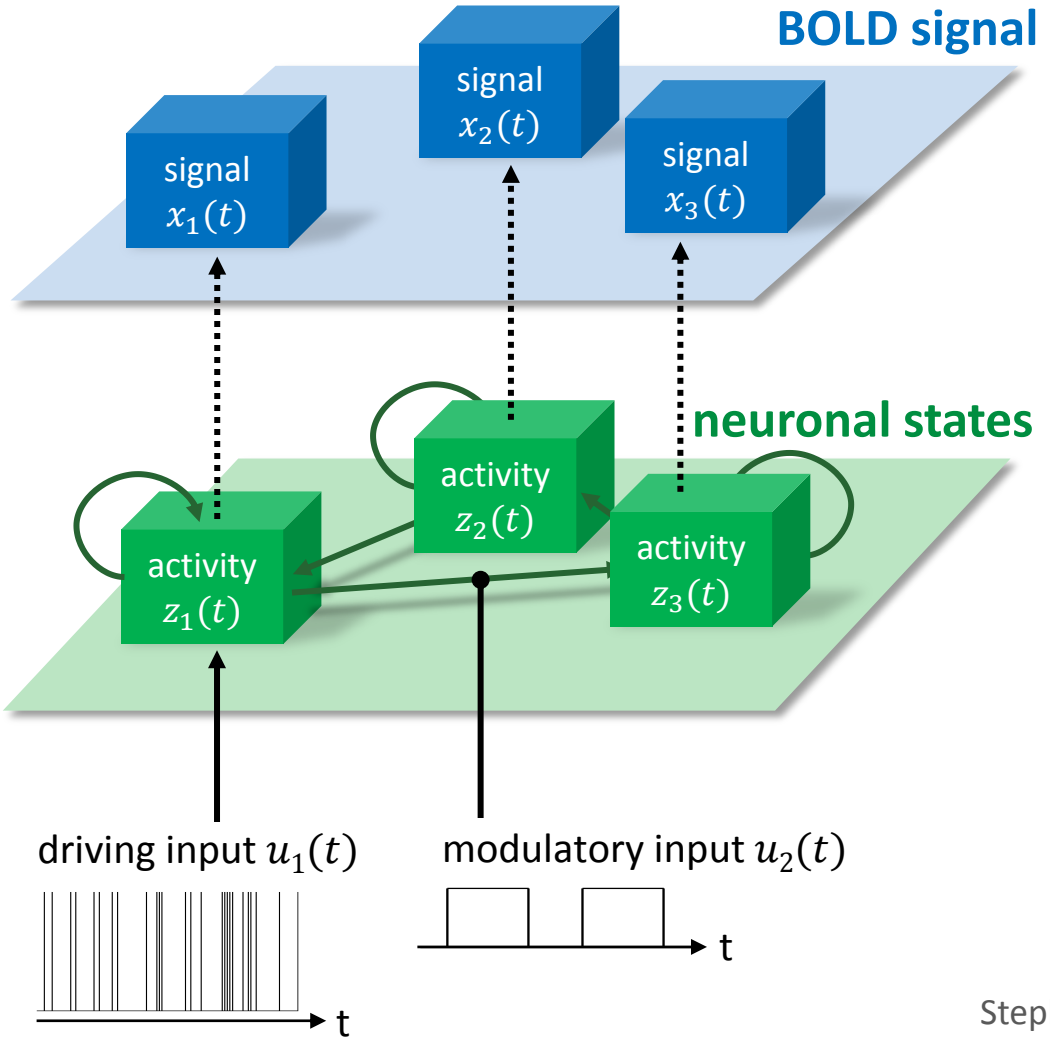
step 4 —
interpretation



jointly discriminative
connection strengths

Brodersen et al. (2011) *NeuroImage;* Brodersen et al. (2011) *PLoS Comp Biol*

Choosing a generative model: DCM for fMRI



hemodynamic forward model

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

neural state equation

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

intrinsic connectivity

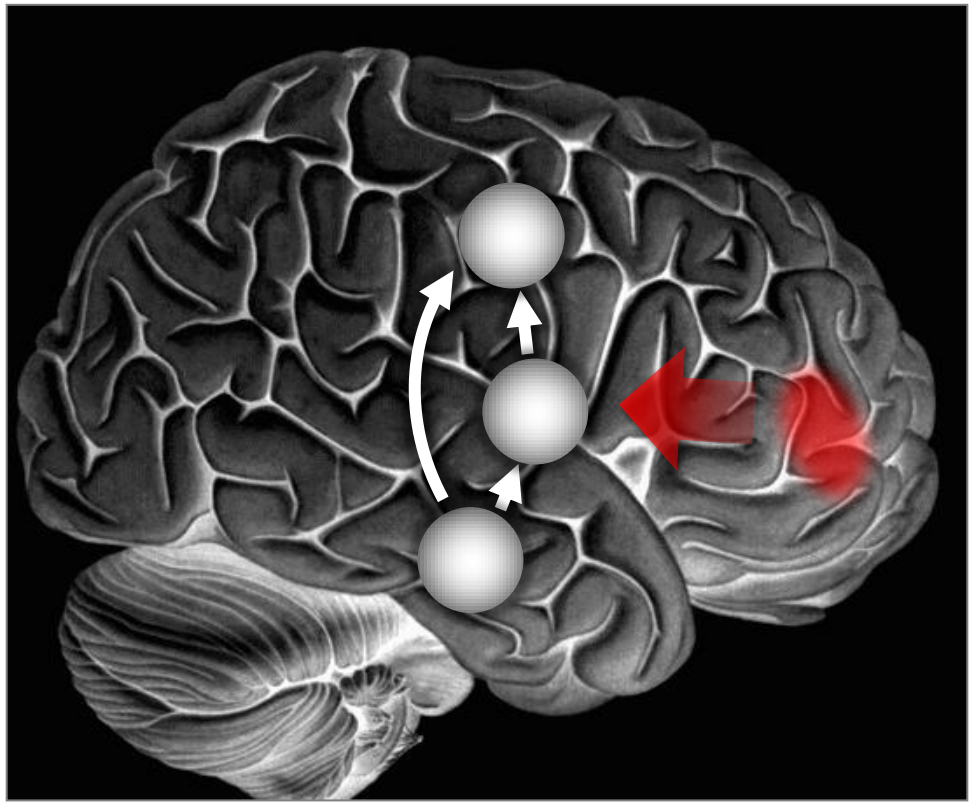
modulation of connectivity

direct inputs

Friston, Harrison & Penny (2003) *NeuroImage*
Stephan & Friston (2007) *Handbook of Brain Connectivity*

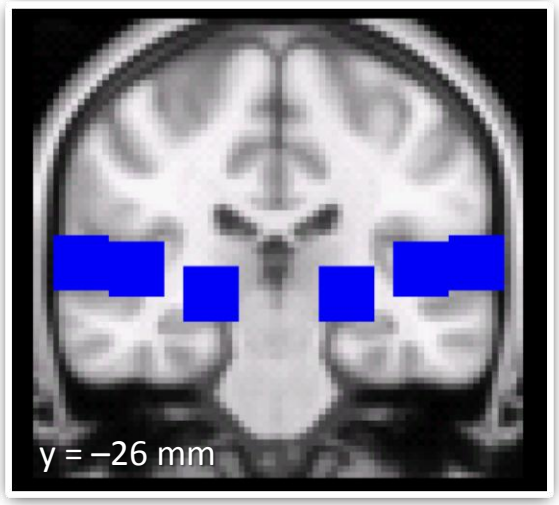
Example: diagnosing stroke patients

To illustrate our approach, we aimed to distinguish between stroke patients and healthy controls, based on non-lesioned regions involved in speech processing.



Example: diagnosing stroke patients

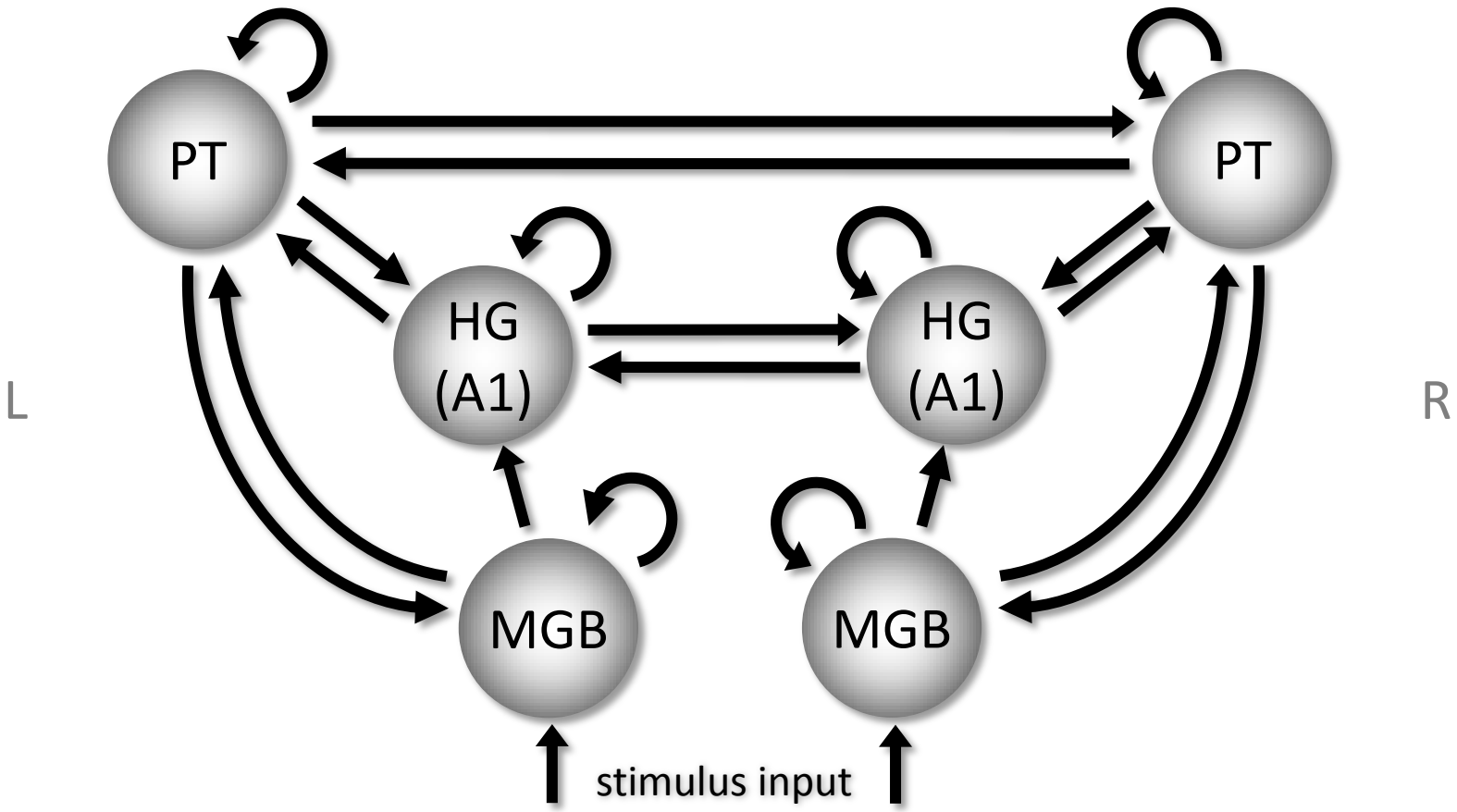
L



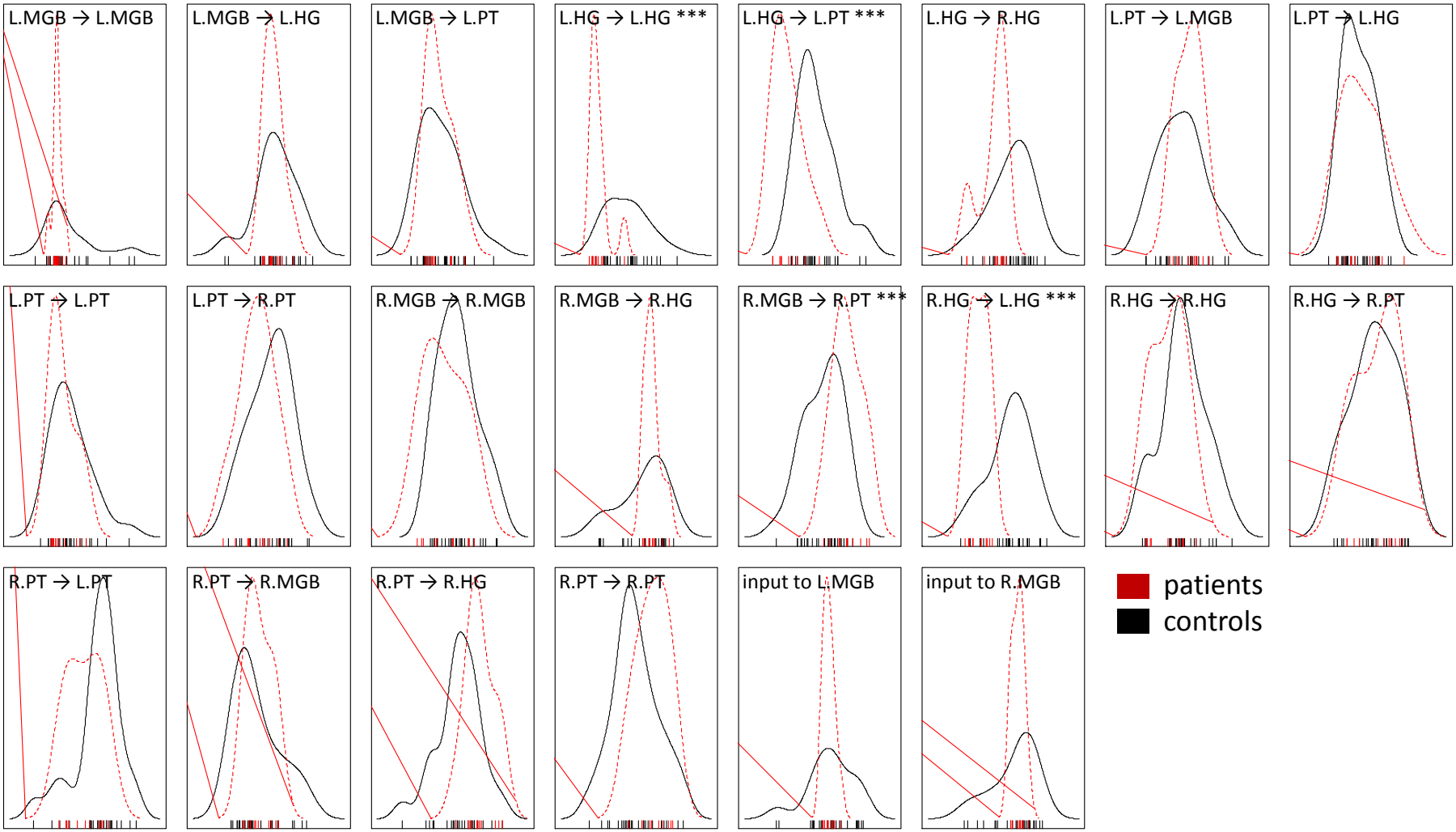
anatomical
regions of interest

R

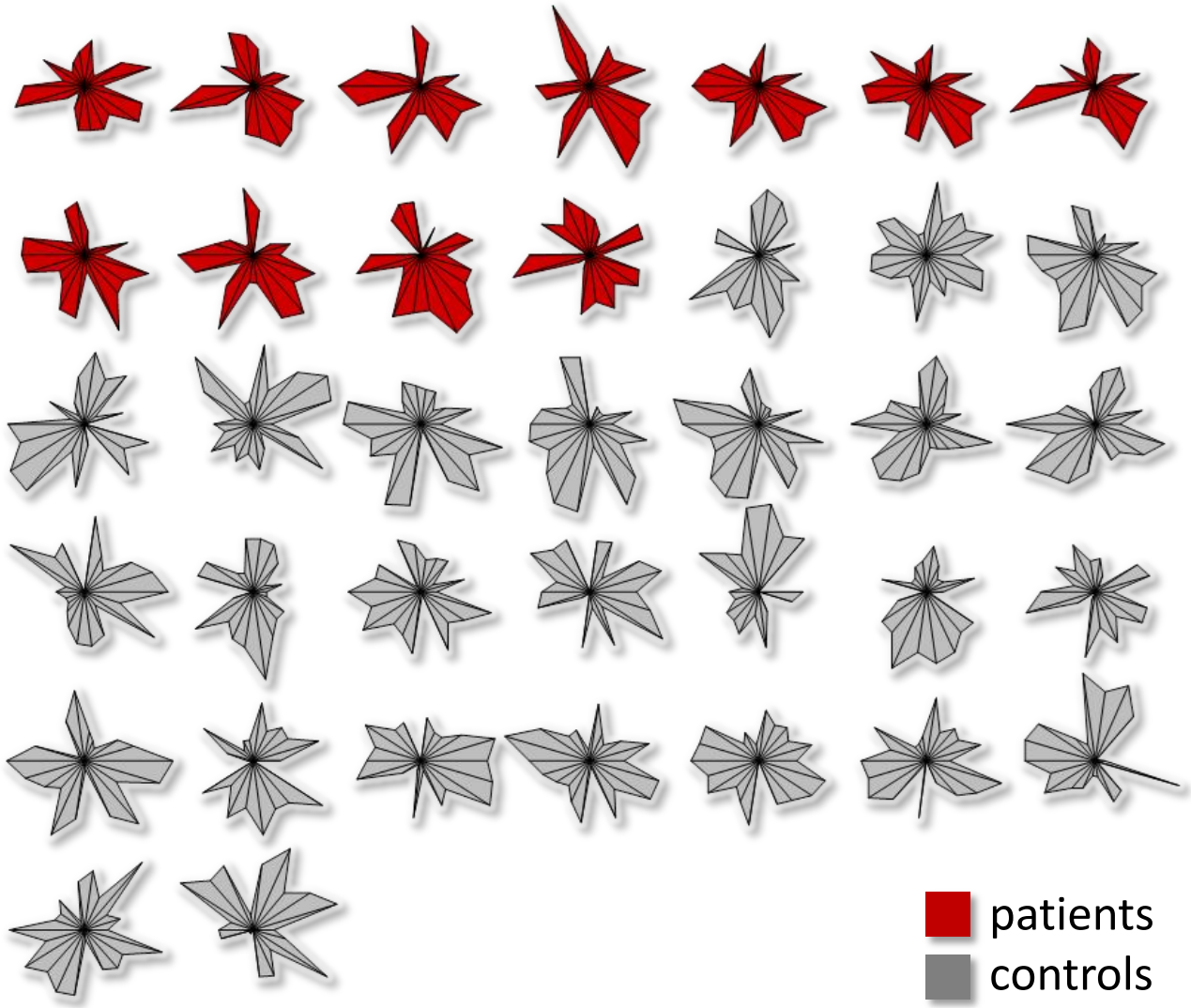
Example: diagnosing stroke patients



Univariate analysis: parameter densities



Multivariate analysis: connectional fingerprints



Full Bayesian approach to performance evaluation

Model

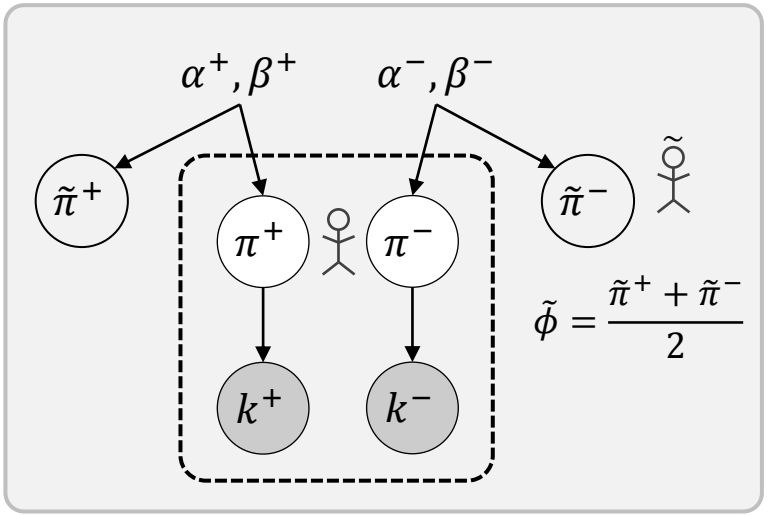
We model the likelihood functions for k^+ positive and k^- negative correct predictions as:

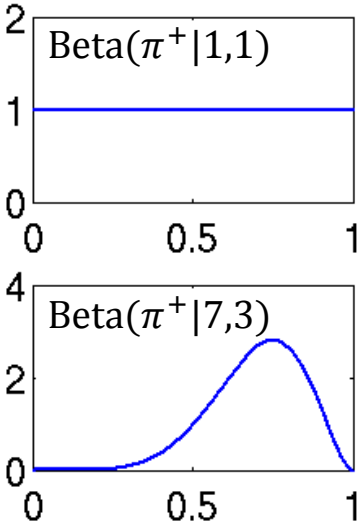
$$p(k^+|\pi^+, n^+) = \mathrm{Bin}(k^+|\pi^+, n^+)$$
$$p(k^-|\pi^-, n^-) = \mathrm{Bin}(k^-|\pi^-, n^-)$$

The class-specific accuracies π^+ and π^- can be modelled as latent random variables with conjugate Beta priors:

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

This prior is uninformative when using the hyperparameters $\alpha^+ = \beta^+ = \alpha^- = \beta^- = 1$. The balanced accuracy is given by $\phi := \frac{1}{2}(\pi^+ + \pi^-)$.





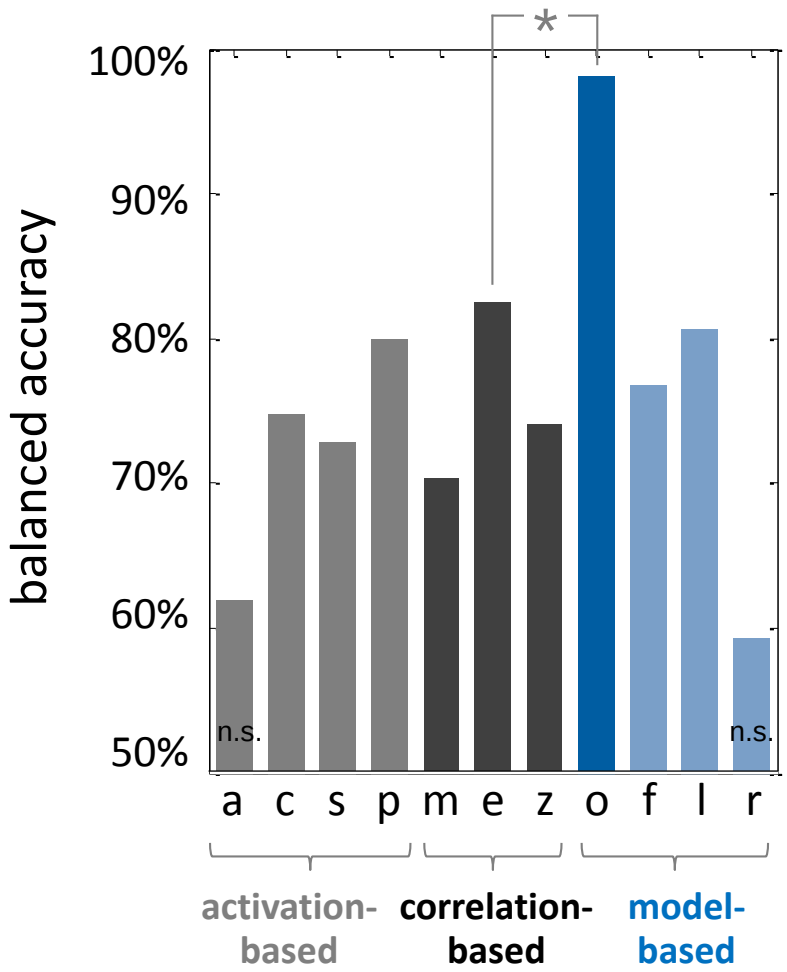
Inference

Inverting the model yields the posterior balanced classification accuracy,

$$p(\phi|k^+, k^-, n^+, n^-, \alpha^+, \beta^+, \alpha^-, \beta^-)$$
$$= \int_0^1 \mathrm{Beta}(2(\phi - z)|\alpha_n^+, \beta_n^+)\, \mathrm{Beta}(2z|\alpha_n^-, \beta_n^-)\, dz\,.$$

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

Classification performance



Activation-based analyses

- **a** anatomical feature selection
- **c** mass-univariate contrast feature selection
- **s** locally univariate searchlight feature selection
- **p** PCA-based dimensionality reduction

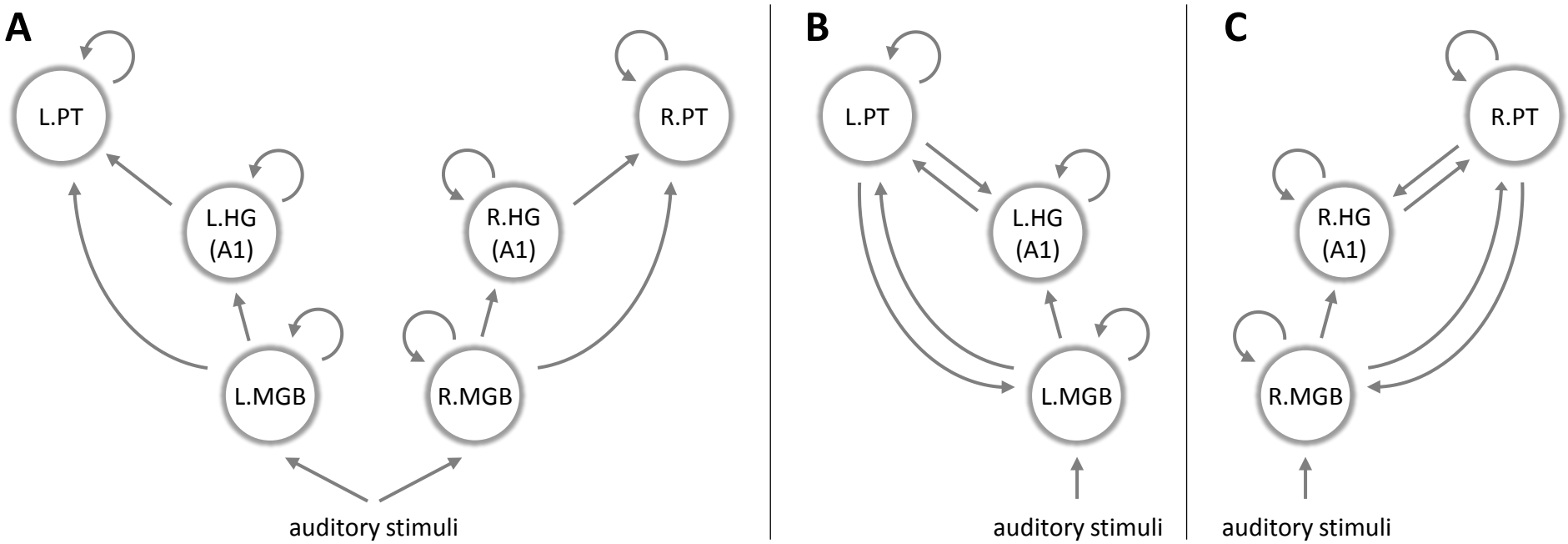
Correlation-based analyses

- **m** correlations of regional means
- **e** correlations of regional eigenvariates
- **z** Fisher-transformed eigenvariates correlations

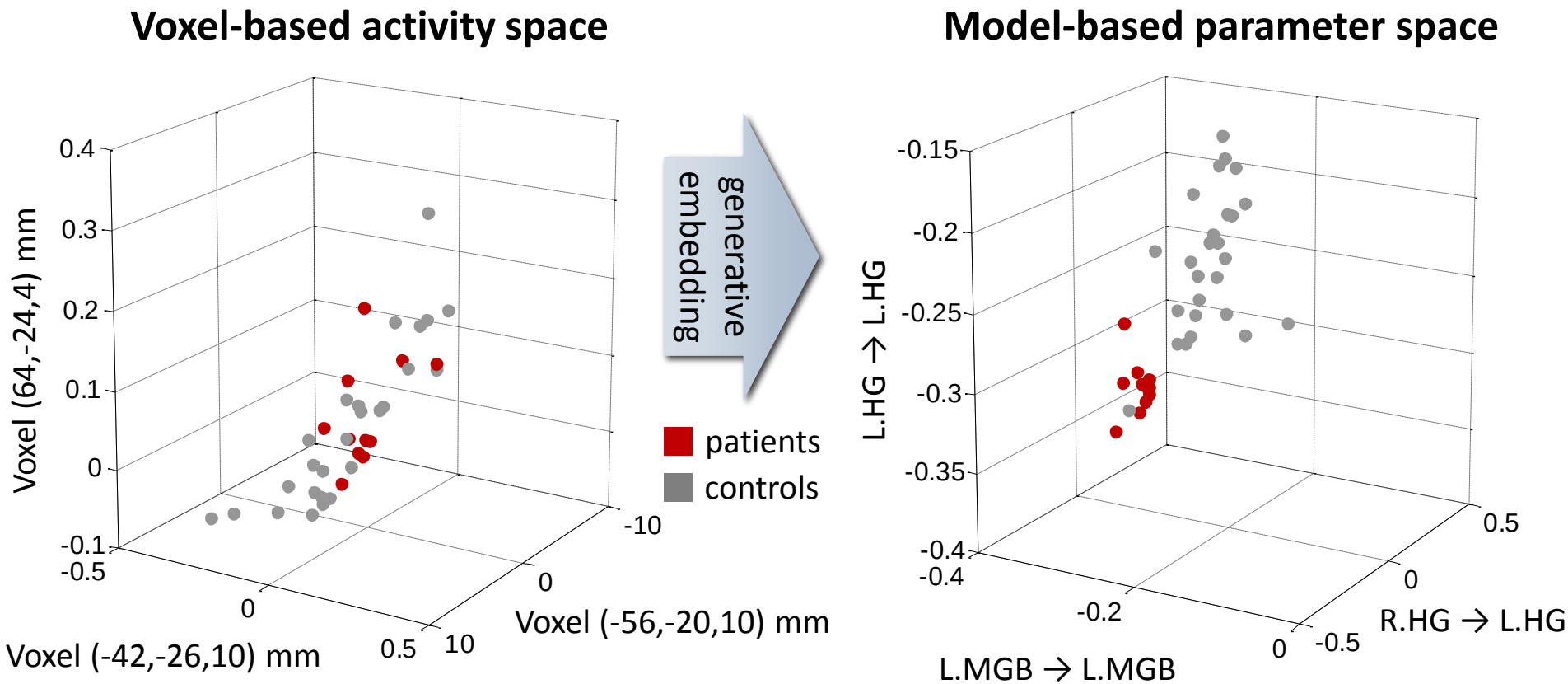
Model-based analyses

- **o** gen.embed., original full model
- **f** gen.embed., less plausible feedforward model
- **l** gen.embed., left hemisphere only
- **r** gen.embed., right hemisphere only

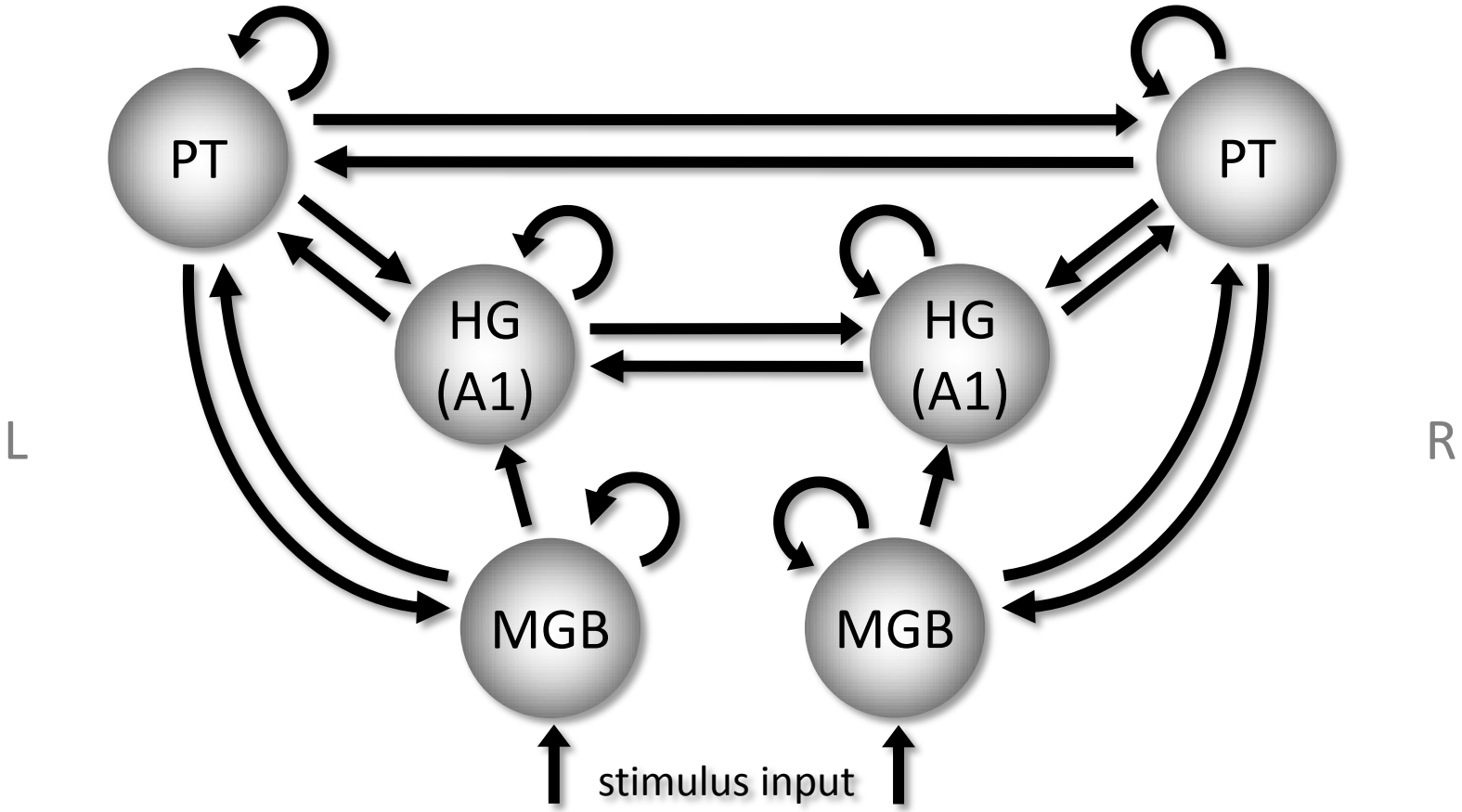
Biologically less plausible models perform poorly



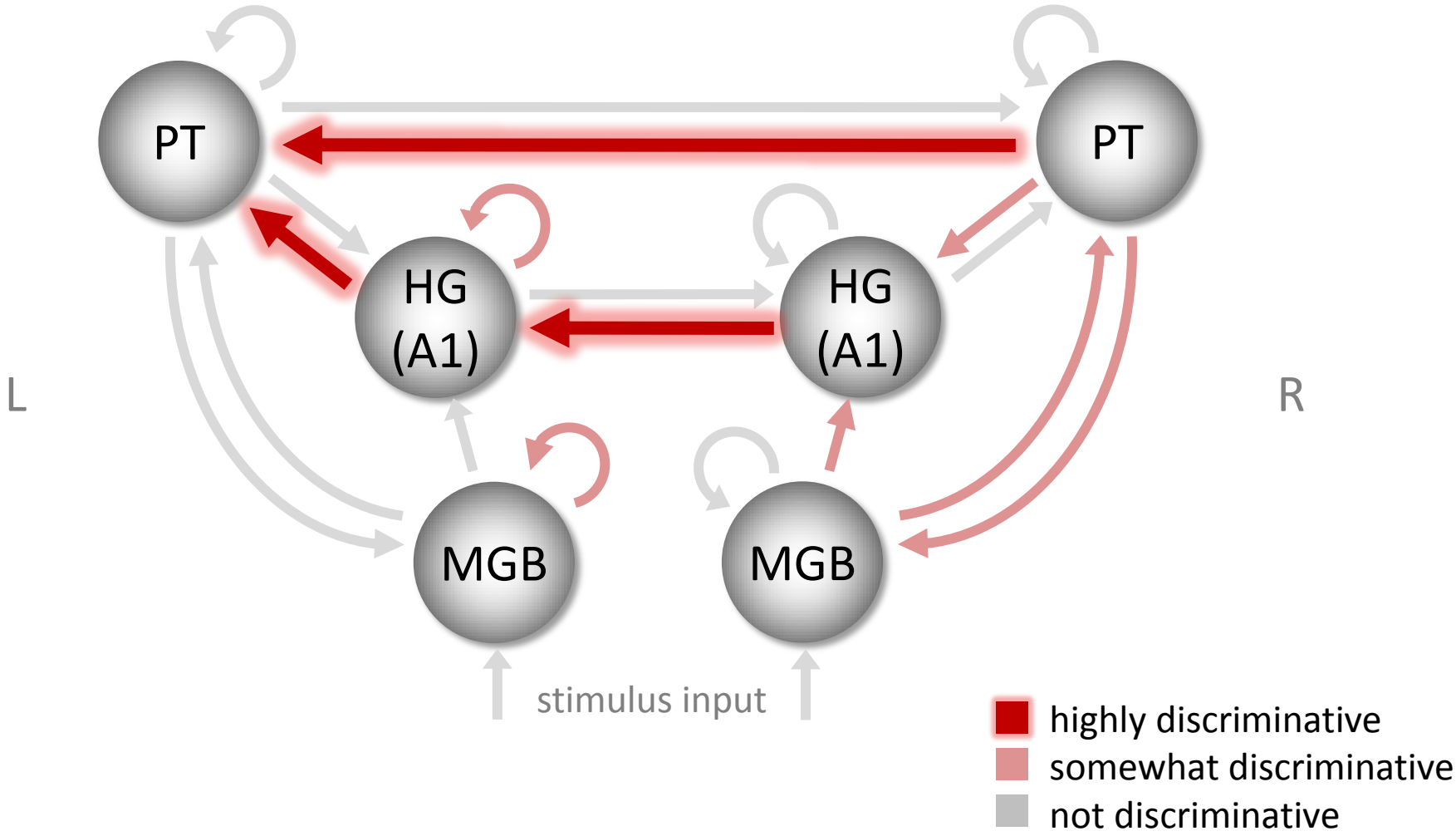
The generative projection



Discriminative features in model space



Discriminative features in model space



Summary: generative embedding for fMRI

❶ **Strong classification performance.** Generative embedding exploits the rich discriminative information encoded in 'hidden' quantities, such as coupling parameters.

❷ **Creation of an interpretable feature space.** High-dimensional fMRI data are replaced by low-dimensional subject-specific fingerprints with biologically interpretable axes.

❸ **Future applications.** Generative embedding could help dissect spectrum disorders into physiologically defined subgroups.