

Bayesian model selection (BMS)

Klaas Enno Stephan



Translational Neuromodeling Unit

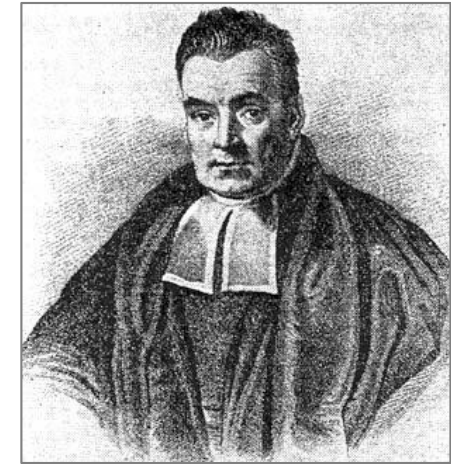
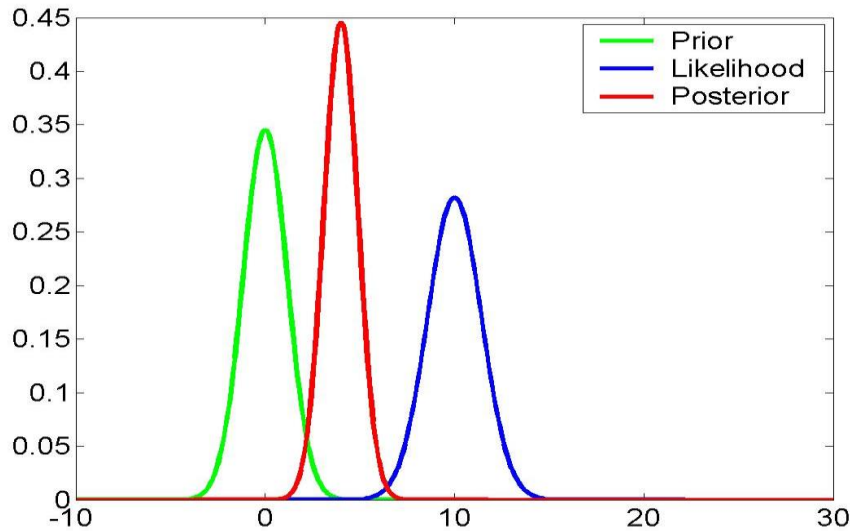


Universität
Zürich^{UZH}



Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Reminder: Bayes' rule



The Reverend Thomas Bayes
(1702-1761)

$$\text{Posterior (inference)} \quad p(\theta | y) = \frac{\text{Likelihood (data)} \quad p(y | \theta) \quad \text{Prior (prediction)} \quad p(\theta)}{\text{Evidence (normalisation term)} \quad p(y)}$$

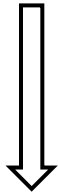
θ : parameters
 y : data

"... the theorem expresses how a ... degree of belief should rationally change to account for availability of related evidence."

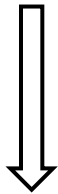
Wikipedia

Model comparison and selection

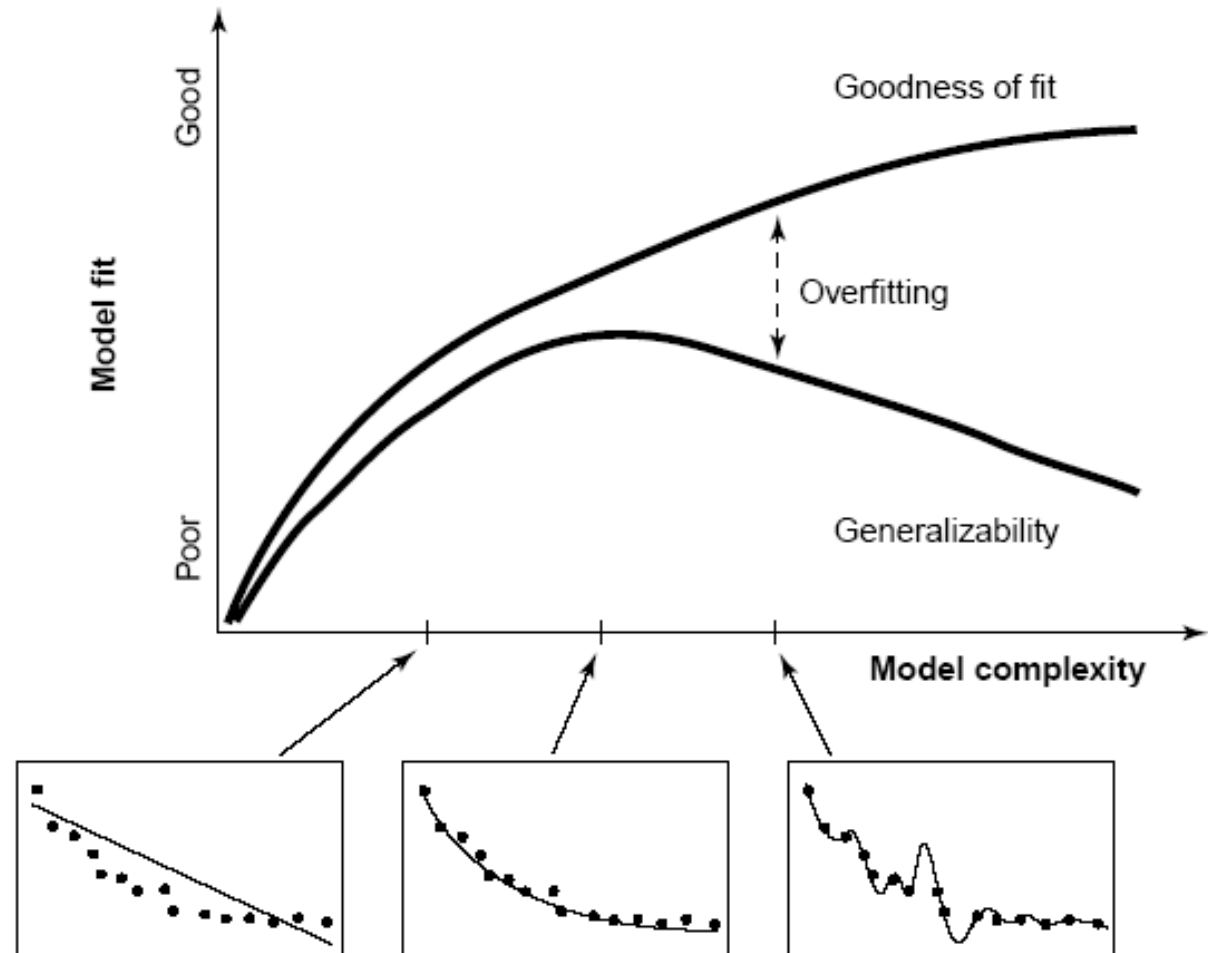
Given competing hypotheses on structure & functional mechanisms of a system, which model is the best?



Which model represents the best balance between model fit and model complexity?



For which model m does $p(y|m)$ become maximal?

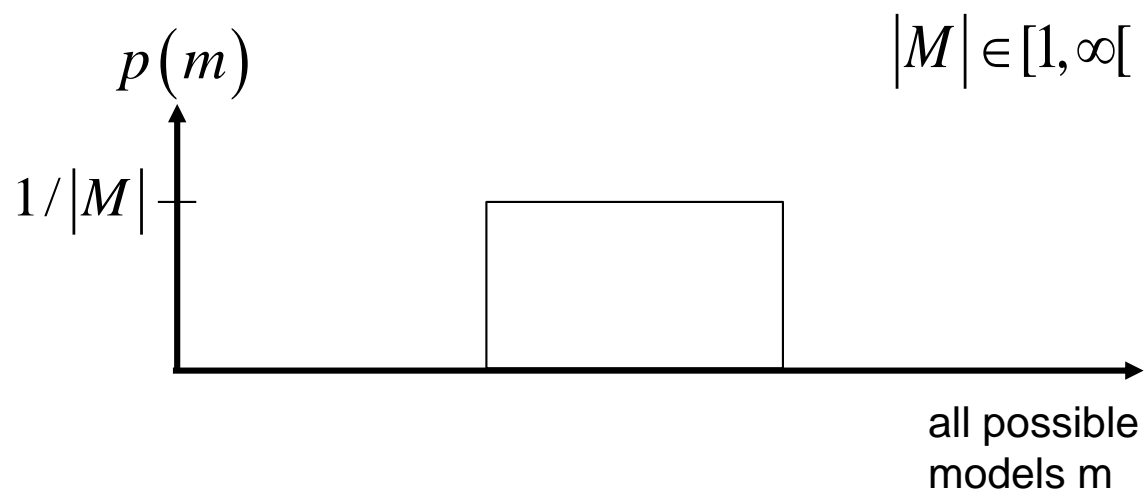


Bayesian model selection (BMS)

- A major goal of generative modeling: comparing competing hypotheses about how observed data were generated
 - a priori definition of hypothesis set (model space) is crucial
- BMS = selecting the model (within the pre-specified model space) that has the highest probability of having generated the data
- Model selection is not equal to model validation!
 - only compares the relative goodness of competing hypotheses within the pre-specified model space
- Model validation requires external criteria (external to the measured data)
 - e.g. predictive validity: can model predict new (unseen) data?

BMS: Defining the model space

- Model space M : all models m that are deemed plausible *a priori*. This is the researcher's hypothesis space.
- Defining M is equivalent to defining a prior over models, $p(m)$, such that a finite set of models has non-zero prior probability.
- Usually, all models in M are assigned equal prior probability:



BMS: Applying Bayes' rule to models

- Goal: select the model that has the highest posterior probability given the data:

$$p(m_i | y) = \frac{p(y | m_i) p(m_i)}{p(y)} = \frac{p(y | m_i) p(m_i)}{\sum_{j=1}^{|M|} p(y | m_j) p(m_j)}$$

- Under uniform priors over models, all prior model probabilities are identical:

$$p(m_i | y) = \frac{p(y | m_i)}{\sum_{j=1}^{|M|} p(y | m_j)} \text{ identical for all models } i$$

- We can thus use the model evidence to compare and select models.

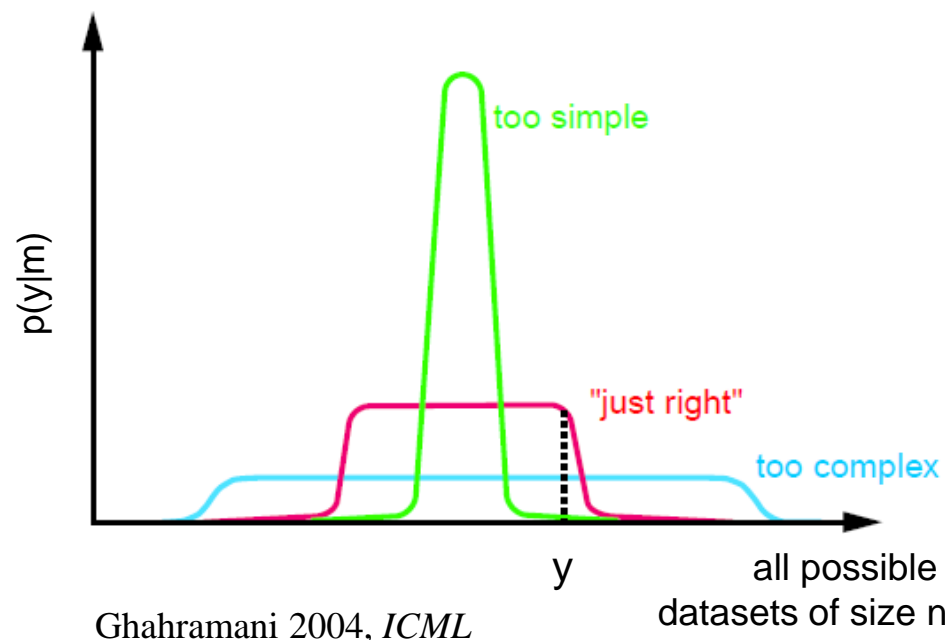
Model evidence: How can it be interpreted?

Model evidence = marginal likelihood:

$$p(y | m) = \int p(y | \theta, m) p(\theta | m) d\theta$$

⇒ expected probability that randomly selected parameters from the prior would generate the observed data

⇒ measure of model goodness: log evidence accounts for both accuracy and complexity of the model



Overly simple models are unlikely to generate the given data.

Overly complex models can generate many possible datasets and are thus unlikely to have generated the specific dataset that is given.

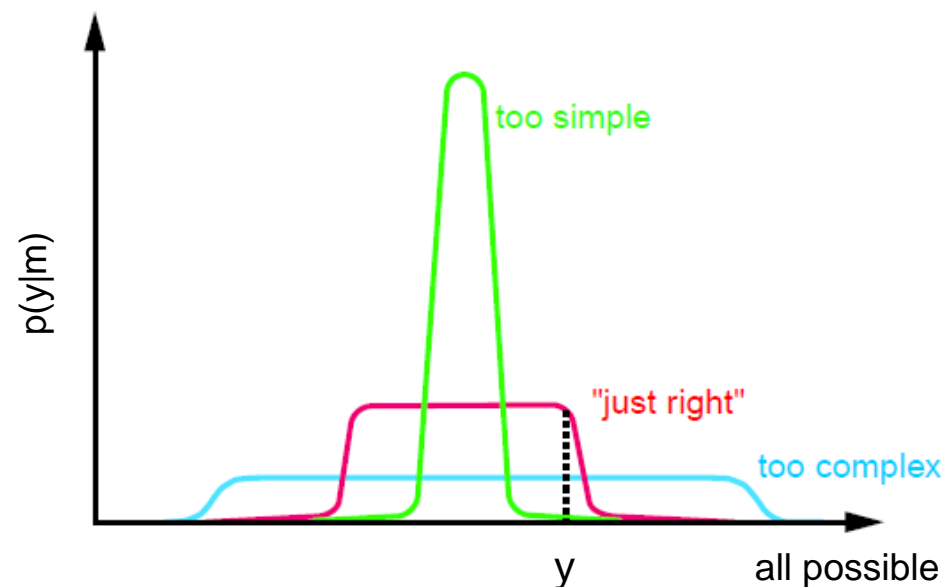
Model evidence: How can it be interpreted?

Model evidence = marginal likelihood:

$$p(y | m) = \int p(y | \theta, m) p(\theta | m) d\theta$$

⇒ “If I randomly sampled from my prior and plugged the resulting value into the likelihood function, how close would the predicted data be – on average – to my observed data?”

⇒ measure of model goodness: log evidence accounts for both accuracy and complexity of the model



Ghahramani 2004, *ICML*

Overly simple models are unlikely to generate the given data.

Overly complex models can generate many possible datasets and are thus unlikely to have generated the specific dataset that is given.

Model evidence: How can it be computed?

- **Analytically:**
 - only possible in rare cases (e.g. linear Gaussian models)
- **By approximation:**
 - Akaike Information Criterion (AIC)
 - Bayesian Information Criterion (BIC)
 - Variational Bayes (VB)
 - negative free energy as lower bound approximation to the log evidence
 - Sampling-based methods
 - prior arithmetic mean estimator
 - posterior harmonic mean estimator
 - thermodynamic integration (TI)

AIC/BIC approximations to the model evidence

- For numerical reasons, we typically approximate the log evidence (the logarithm is a strictly monotonic function).
- The log evidence (and its approximations) can be decomposed into accuracy and complexity terms:

$$\begin{aligned}\log p(y | m) &= \textit{accuracy}(m) - \textit{complexity}(m) \\ &= \log p(y | \theta, m) - \textit{complexity}(m)\end{aligned}$$

Akaike Information Criterion: $AIC = \log p(y | \theta, m) - p$

No. of
parameters

Bayesian Information Criterion: $BIC = \log p(y | \theta, m) - \frac{p}{2} \log N$

No. of
data points

Two expressions of AIC/BIC in the literature

- definition of AIC/BIC shown on the previous slide:
AIC/BIC serve as (asymptotic) approximations of the log model evidence
 - for details, see Raftery (1995) *Sociological Methodology* 25: 111-163, or Penny et al. (2004) *NeuroImage* 22: 1157-1172
- in the literature, often the negative of these expressions is used
 - essentially an approximation to surprise
 - not in line with original BIC paper by Schwarz (1978) *Annals of Statistics*, 6: 461-464
- in principle, either version is fine for model selection
 - but important for methods assuming that AIC/BIC approximate log evidence
 - e.g. random effects BMS (described below):
only AIC/BIC expressions from previous slides lead to correct results

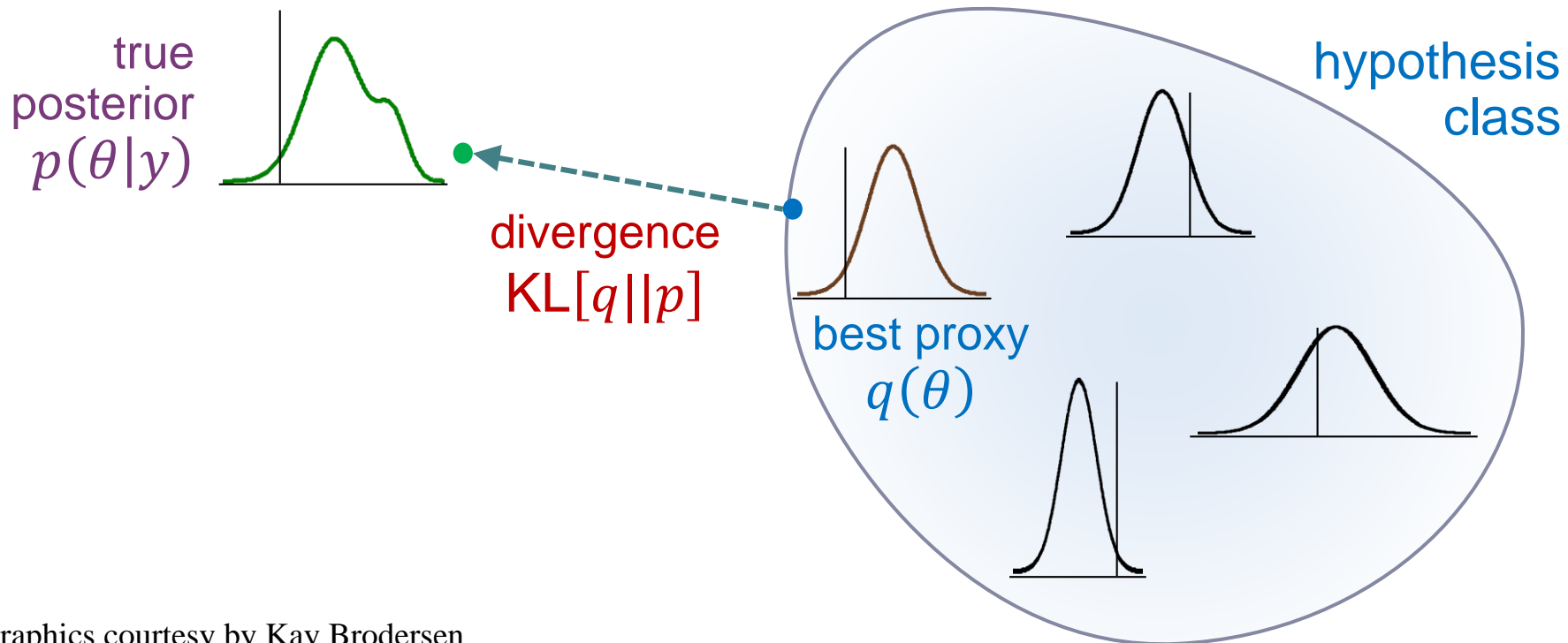
Pros and cons of AIC/BIC

- **easy and fast to compute**
 - only the MLE of the parameters and counting (parameters and data points) is needed.
- but: a **limited concept of model complexity**
 - all parameters contribute equally to complexity, regardless of how they can affect the data
 - e.g. do not take into account dependencies amongst parameters
- a richer concept of model complexity is provided by the **negative free energy** (aka "**evidence lower bound**", **ELBO**)

Recap: Variational Bayes (VB)

Basic idea of VB: find an approximate density $q(\theta)$ that is maximally similar to the true posterior $p(\theta|y)$.

This is often done by assuming a particular form for q (fixed form VB) and then optimising its sufficient statistics.



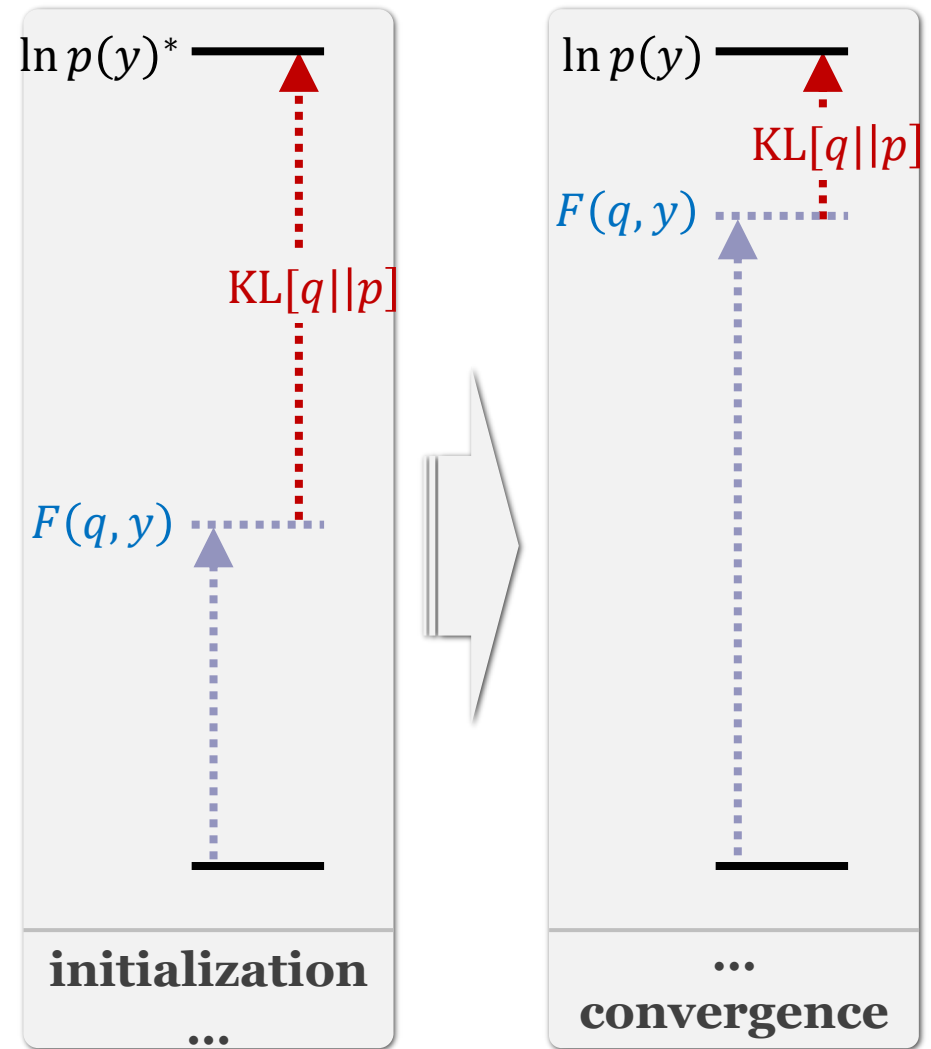
Recap: Variational Bayes

$$\ln p(y) = \underbrace{\text{KL}[q||p]}_{\substack{\text{divergence} \\ \geq 0 \\ \text{(unknown)}}} + \underbrace{F(q, y)}_{\substack{\text{neg. free} \\ \text{energy} \\ \text{(easy to evaluate} \\ \text{for a given } q\text{)}}}$$

$F(q, y)$ is a functional wrt. the approximate posterior $q(\theta)$.

Maximizing $F(q, y)$ is equivalent to:

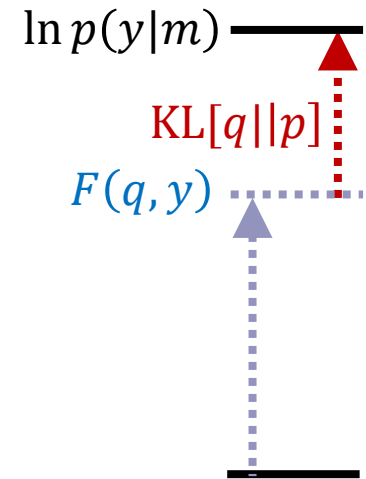
- minimizing $\text{KL}[q||p]$,
- making $q(\theta)$ the best possible estimate of the posterior,
- tightening $F(q, y)$ as a lower bound to the log model evidence.



The negative free energy approximation F

F is a lower bound on the log model evidence:

$$\log p(y | m) = F + KL[q(\theta), p(\theta | y, m)]$$



Like AIC/BIC, F is an accuracy/complexity tradeoff:

$$F = \underbrace{\langle \log p(y | \theta, m) \rangle}_{\text{accuracy}} - \underbrace{KL[q(\theta), p(\theta | m)]}_{\text{complexity}}$$

The derivations can be found in many papers; e.g. see appendix to Stephan et al. 2007, *NeuroImage* (not the first!)

The complexity term in F

- In contrast to AIC & BIC, the complexity term of the negative free energy F accounts for parameter interdependencies.

Under Gaussian assumptions about the posterior (Laplace approximation):

$$KL[q(\theta), p(\theta | m)] \\ = \frac{1}{2} \ln |C_{\theta}| - \frac{1}{2} \ln |C_{\theta|y}| + \frac{1}{2} (\mu_{\theta|y} - \mu_{\theta})^T C_{\theta}^{-1} (\mu_{\theta|y} - \mu_{\theta})$$

- The complexity term of F is higher
 - the more independent the prior parameters (\uparrow effective DFs)
 - the more dependent the posterior parameters
 - the more the posterior mean deviates from the prior mean

Bayes factors

To compare two models, we could just compare their log evidences.

But the log evidence is just some number – not very intuitive!

A more intuitive interpretation of model comparisons is made possible by Bayes factors:

$$B_{12} = \frac{p(y | m_1)}{p(y | m_2)}$$

positive value, $[0; \infty[$

Kass & Raftery classification:

B_{12}	$p(m_1 y)$	Evidence
1 to 3	50-75%	weak
3 to 20	75-95%	positive
20 to 150	95-99%	strong
≥ 150	$\geq 99\%$	Very strong

Fixed effects BMS at group level

Group Bayes factor (GBF) for $1 \dots K$ subjects:

$$GBF_{ij} = \prod_k BF_{ij}^{(k)}$$

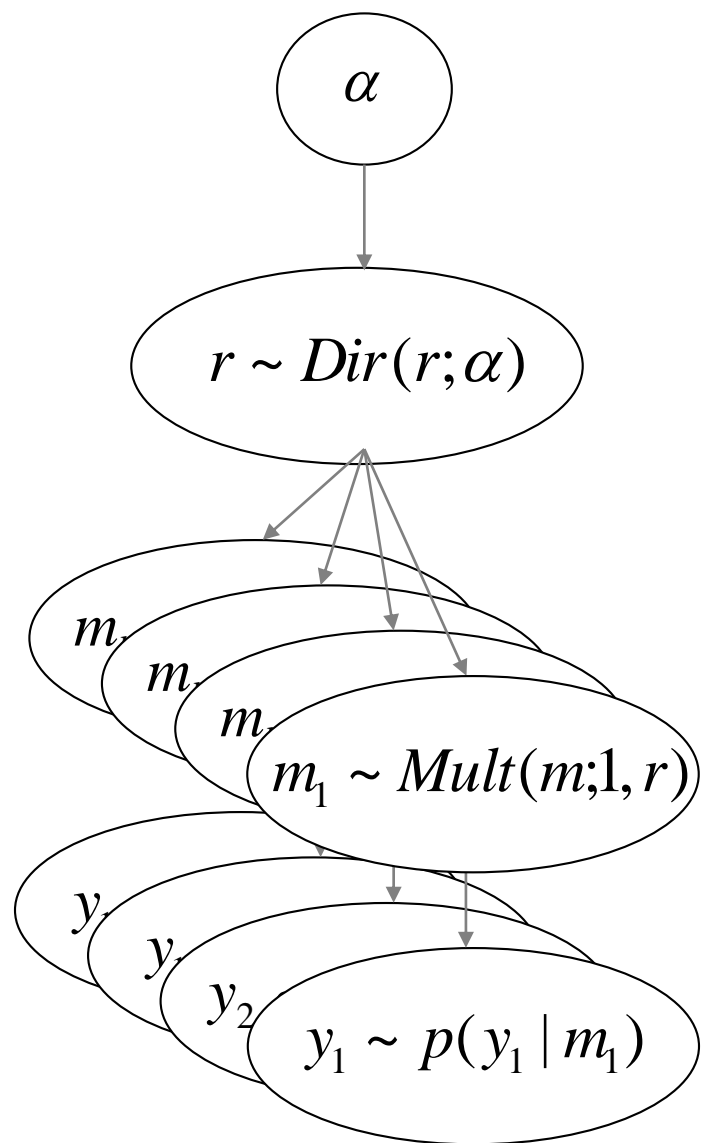
Average Bayes factor (ABF):

$$ABF_{ij} = \sqrt[K]{\prod_k BF_{ij}^{(k)}}$$

Problems:

- blind with regard to group heterogeneity
- sensitive to outliers

Random effects BMS for heterogeneous groups



Dirichlet parameters α
= “occurrences” of models in the population

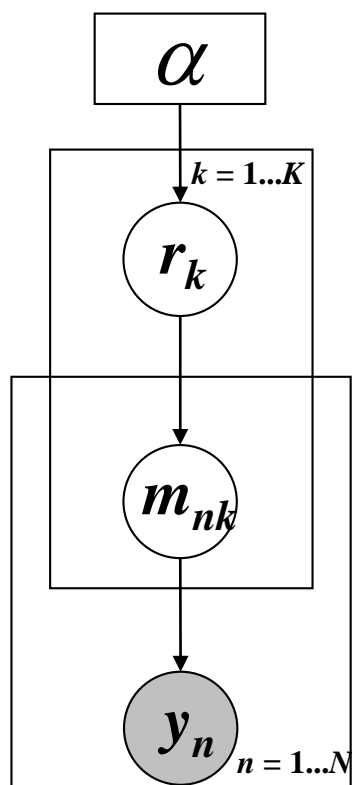
Dirichlet distribution of model probabilities r

Multinomial distribution of model labels m

Measured data y

**Model inversion
by Variational
Bayes or MCMC**

Random effects BMS: plate notation



Dirichlet parameters α
= “occurrences” of models in the population

Dirichlet distribution of model probabilities r

Multinomial distribution of model labels m

Measured data y

**Model inversion
by Variational
Bayes or MCMC**

Four equivalent options for reporting model ranking by random effects BMS

1. Dirichlet parameter estimates

$$\alpha$$

2. **expected posterior probability** of obtaining the k -th model for any randomly selected subject

$$\langle r_k \rangle_q = \alpha_k / (\alpha_1 + \dots + \alpha_K)$$

3. **exceedance probability (XP)** that a particular model k is more likely than any other model (of the K models tested), given the group data

$$\exists k \in \{1 \dots K\}, \forall j \in \{1 \dots K \mid j \neq k\} :$$

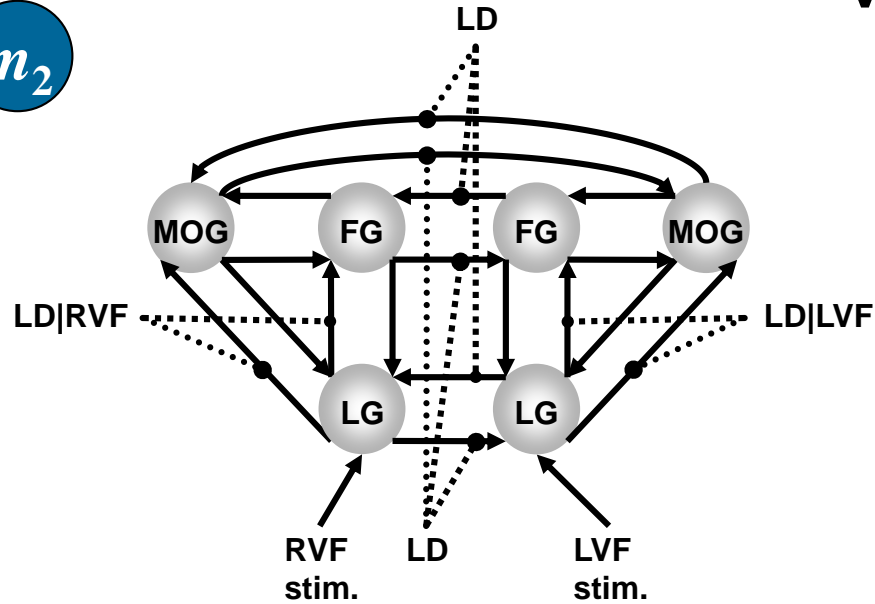
$$\varphi_k = p(r_k > r_j \mid y; \alpha)$$

4. **protected exceedance probability (PXP)**: see Rigoux et al. (2014) and below

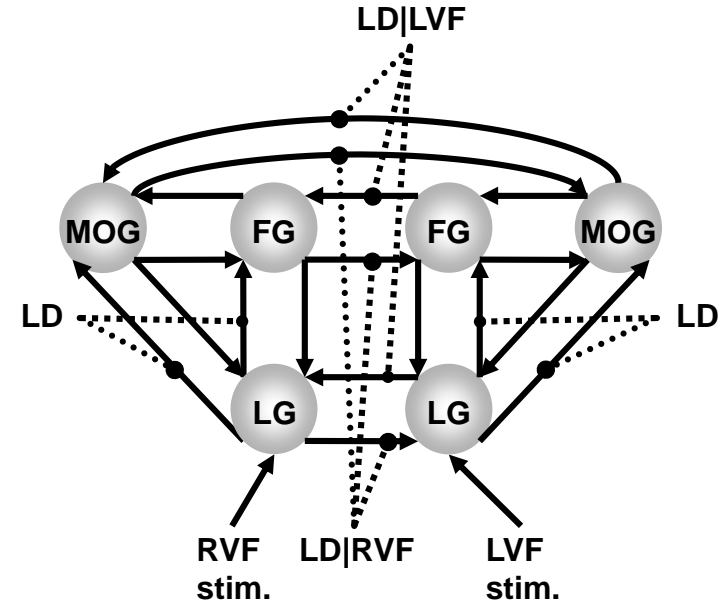
$$\begin{aligned} PXP &= P(r_k \geq r_{k' \neq k} \mid y) \\ &= P(r_k \geq r_{k' \neq k} \mid y, H_1)P(H_1 \mid y) + P(r_k \geq r_{k' \neq k} \mid y, H_0)P(H_0 \mid y) \end{aligned}$$

Example: Hemispheric interactions during vision

m_2



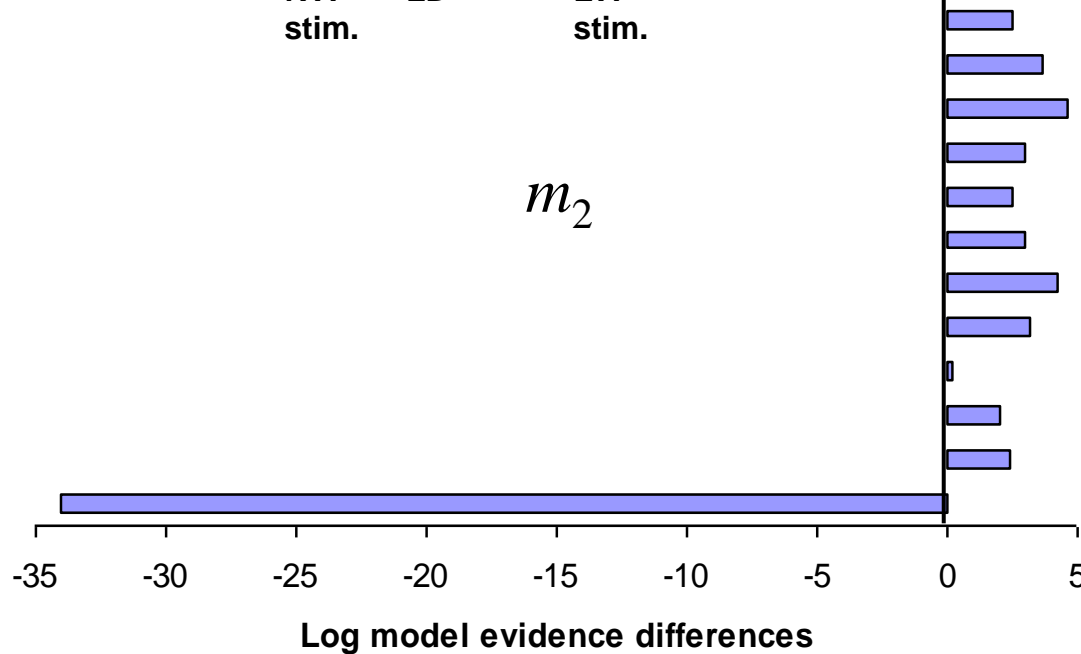
m_1



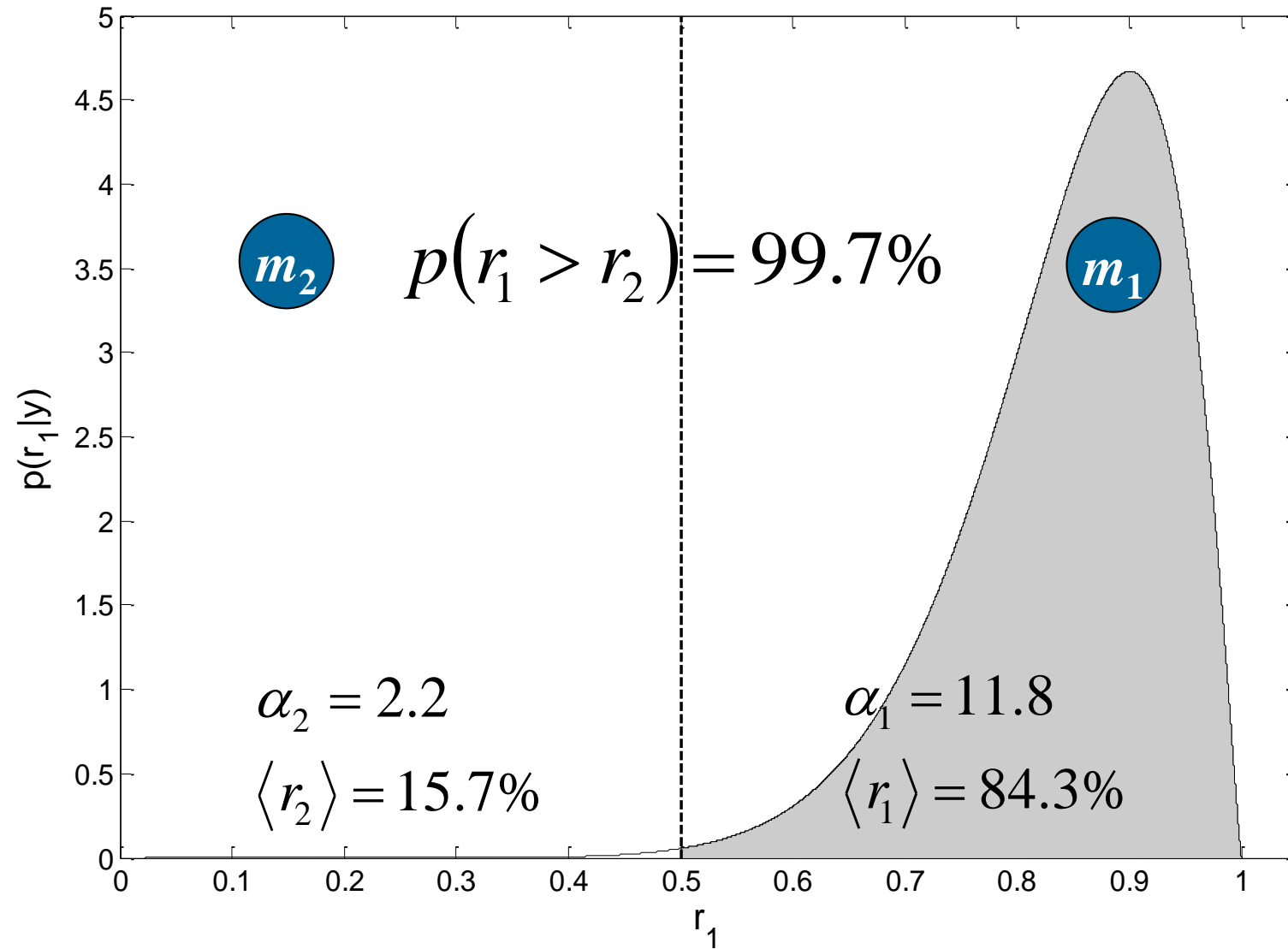
Subjects

m_2

m_1



Data: Stephan et al. 2003, *Science*
Models: Stephan et al. 2007, *J. Neurosci.*



Protected exceedance probability (PXP): Using BMA to protect against chance findings

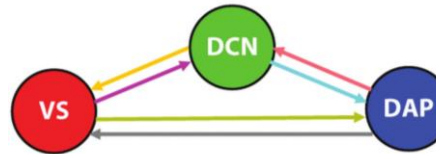
- XPs express our confidence that the posterior probabilities of models are different – under the hypothesis H_1 that models differ in probability: $r_k \neq 1/K$
- does not account for the possibility of the "null hypothesis" H_0 : $r_k = 1/K$
- **protected exceedance probability:**
Bayesian model averaging (BMA, see below) over H_0 and H_1 :

$$\begin{aligned} PXP &= P(r_k \geq r_{k' \neq k} \mid y) \\ &= P(r_k \geq r_{k' \neq k} \mid y, H_1) P(H_1 \mid y) \\ &\quad + P(r_k \geq r_{k' \neq k} \mid y, H_0) P(H_0 \mid y) \end{aligned}$$

Example: DA effects on striatal connectivity

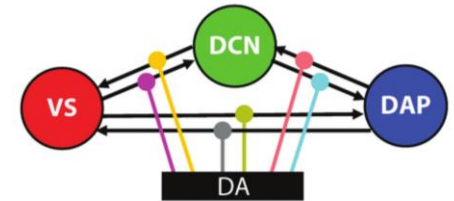
- task-free fMRI with 2x2 pharmacological design (sulpiride, bromocriptin, ON/OFF)
- stochastic DCM
- DA drug effects on the striatum vary widely across individuals and might average to zero
- BMS: identifies where DA drugs alter intrastriatal connectivity, even when the sign of this effect is not consistent across participants

intra-striatal connectivity



A1						
A2						
A3						
A4						
A5						

dopaminergic modulation



B0						
B1						
B2						
B3						
B4						
B5						

winning model

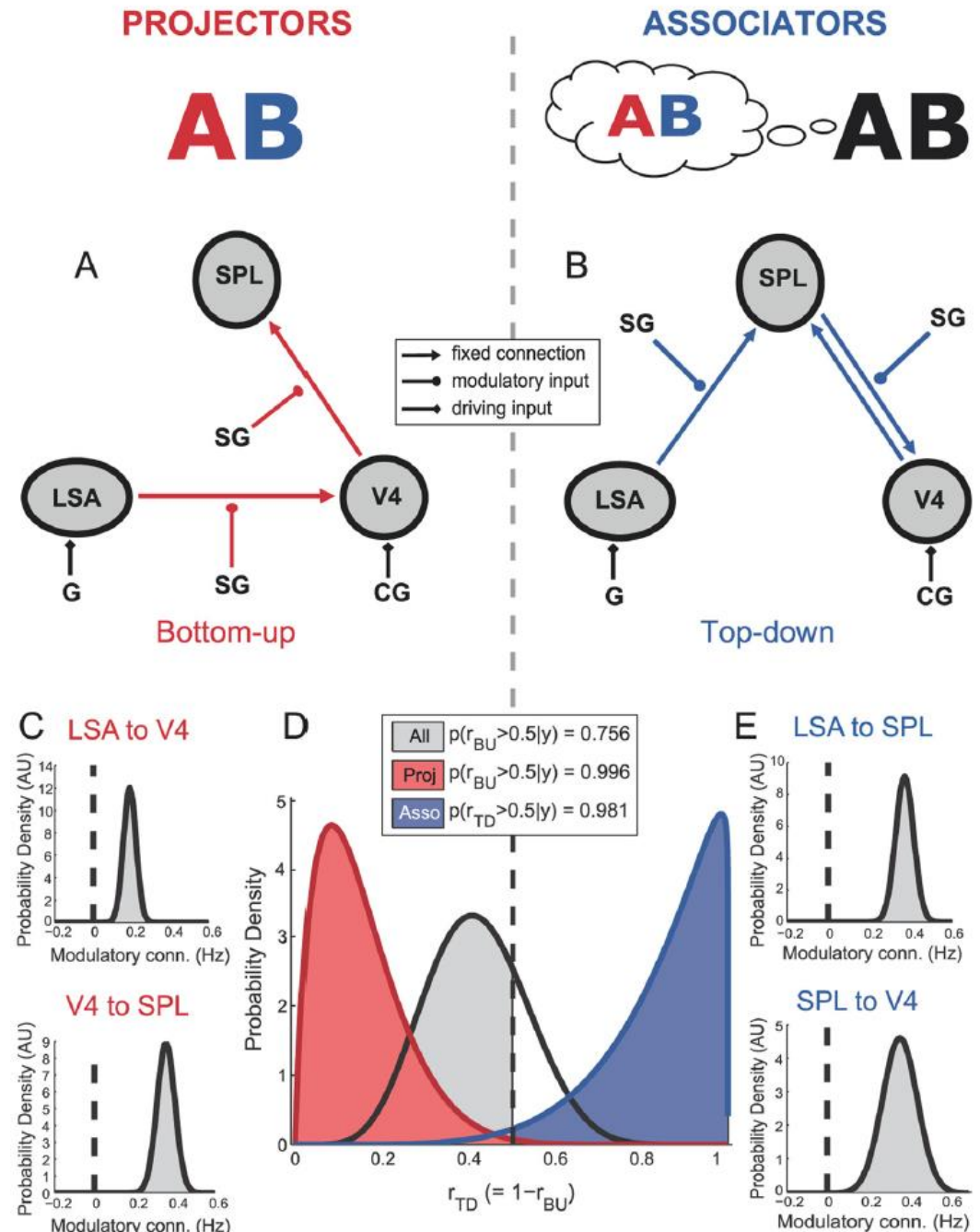


VS: ventral striatum
DCN: dorsal caudate nucleus
DAP: dorsal anterior putamen

Example: Synaesthesia

- “projectors” experience color externally colocalized with a presented grapheme
- “associators” report an internally evoked association
- across all subjects: no evidence for either model
- but BMS results map precisely onto projectors (bottom-up mechanisms) and associators (top-down)

van Leeuwen et al. 2011, *J. Neurosci.*



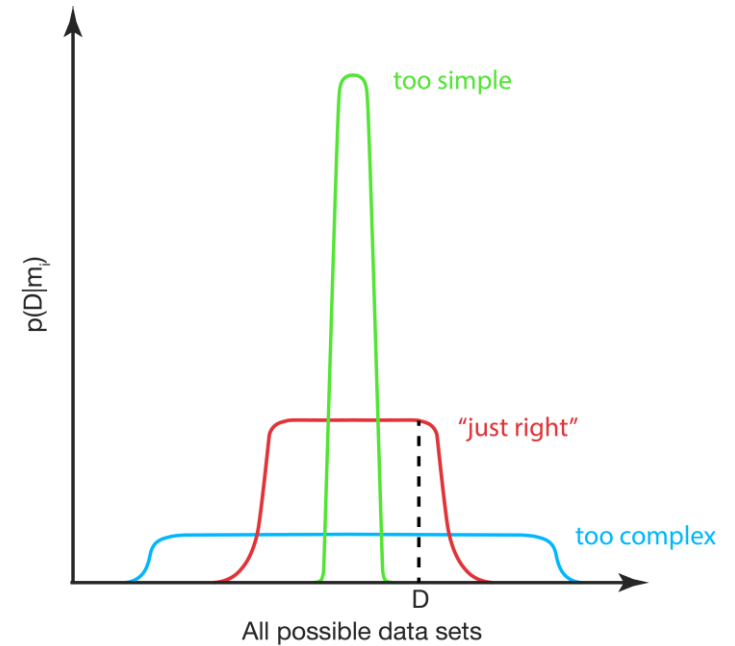
Overfitting at the level of models

Common question:

There is an infinite number of possible models for a given dataset. Should we not strive for a "brute force" search and compare as many models as possible?

Not necessarily.

The more models are included in the model space, the risk of overfitting (at the level of models) increases, too.



Animation courtesy of Stefan Frässle

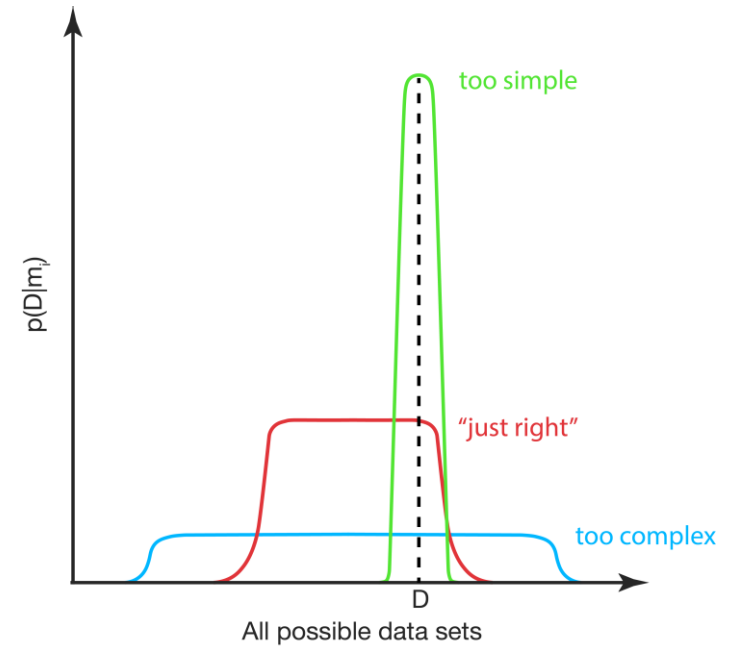
Overfitting at the level of models

Common question:

There is an infinite number of possible models for a given dataset. Should we not strive for a "brute force" search and compare as many models as possible?

Not necessarily.

The more models are included in the model space, the risk of overfitting (at the level of models) increases, too.



Animation courtesy of Stefan Frässle

Overfitting at the level of models

In brief: $\uparrow \text{\#models} \Rightarrow \uparrow \text{risk of overfitting}$

Three solutions in the context of BMS:

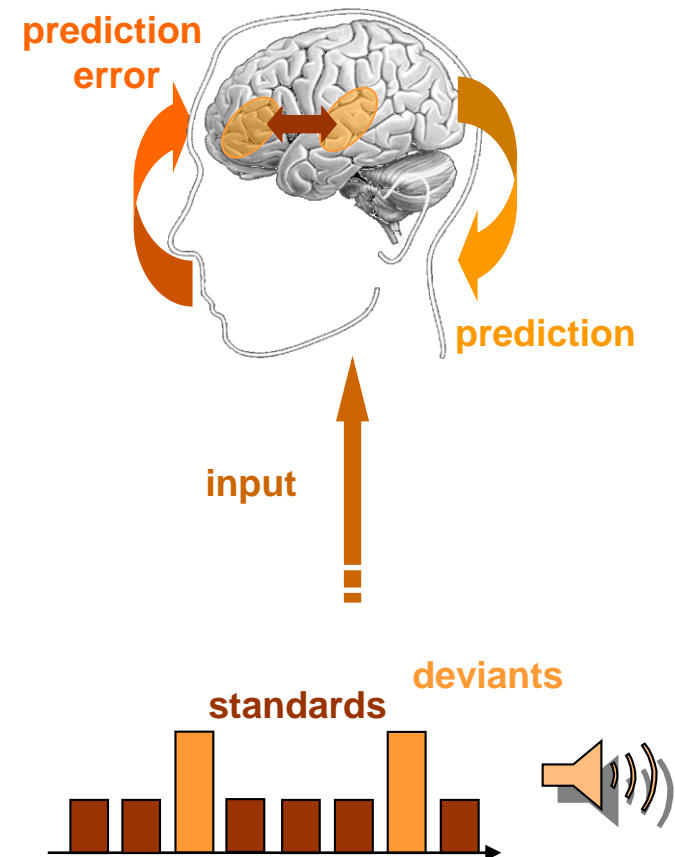
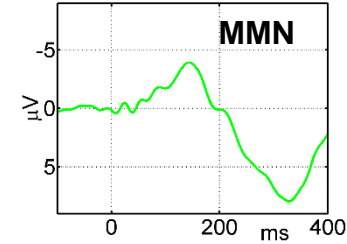
- ❶ Regularisation: choosing priors over models $p(m)$ such that model space is small
- ❷ Family-level BMS
- ❸ Bayesian model averaging (BMA)

② Family-level BMS

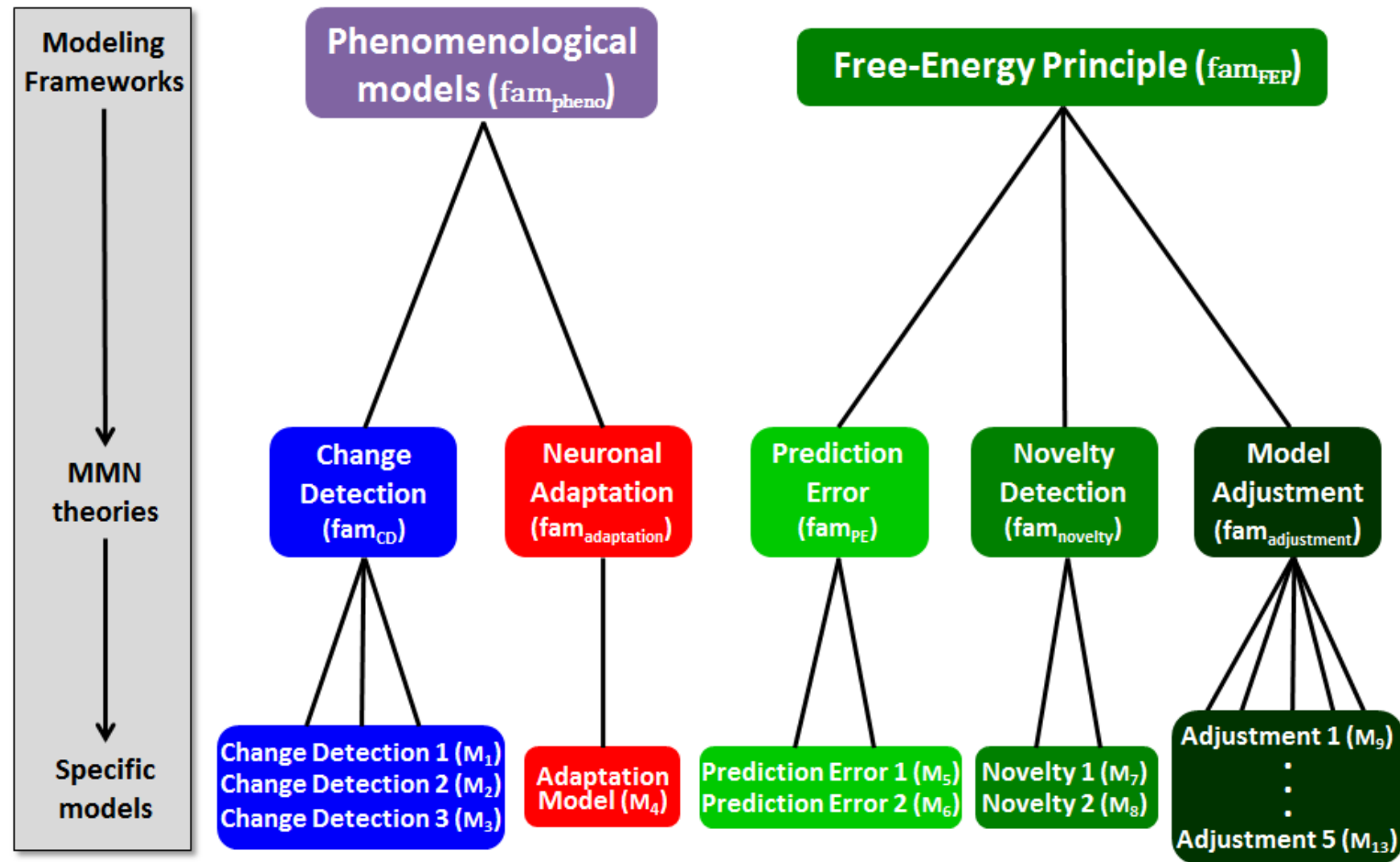
- rests on partitioning model space into K subsets or families: $M = \{f_1, \dots, f_K\}$
- pooling information over all models in these subsets allows one to compute the probability of a model family, given the data
- effectively removes uncertainty about any aspect of model structure, other than the attribute of interest (which defines the partition)
- the uniform prior over models is replaced with a uniform prior over families:
 - For **FFX**, this means family prior $p(f) = \frac{1}{K}$ and model prior $p(m) = \frac{1}{KN_k}$.
 - For **RFX**, fair priors over families are achieved by setting $\alpha_{prior}(m) = \frac{1}{N_k}$.

Mismatch negativity (MMN)

- elicited by surprising stimuli (scales with unpredictability)
- ↓ in schizophrenic patients
- classical interpretations:
 - pre-attentive change detection
 - neuronal adaptation
- more recent theories:
 - reflection of (hierarchical) Bayesian inference



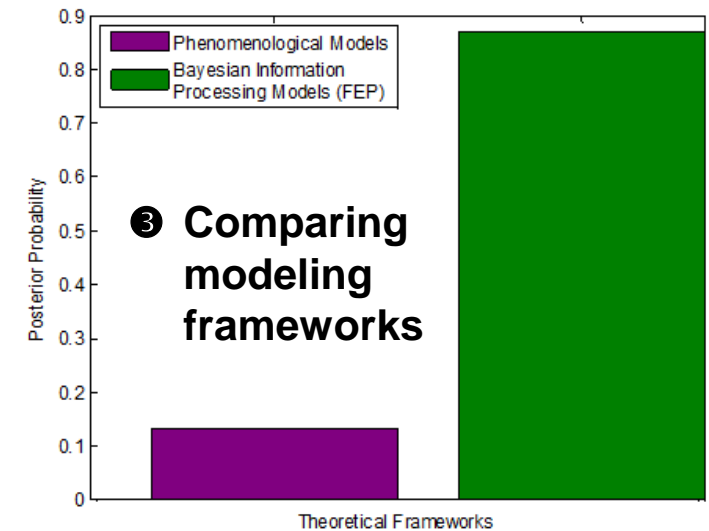
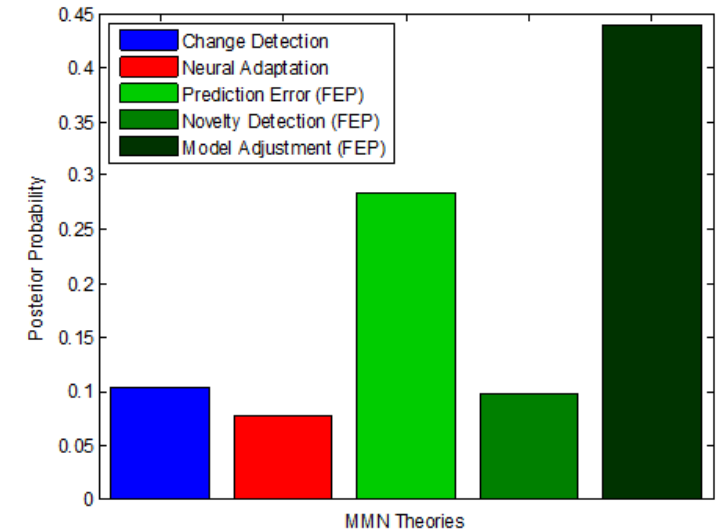
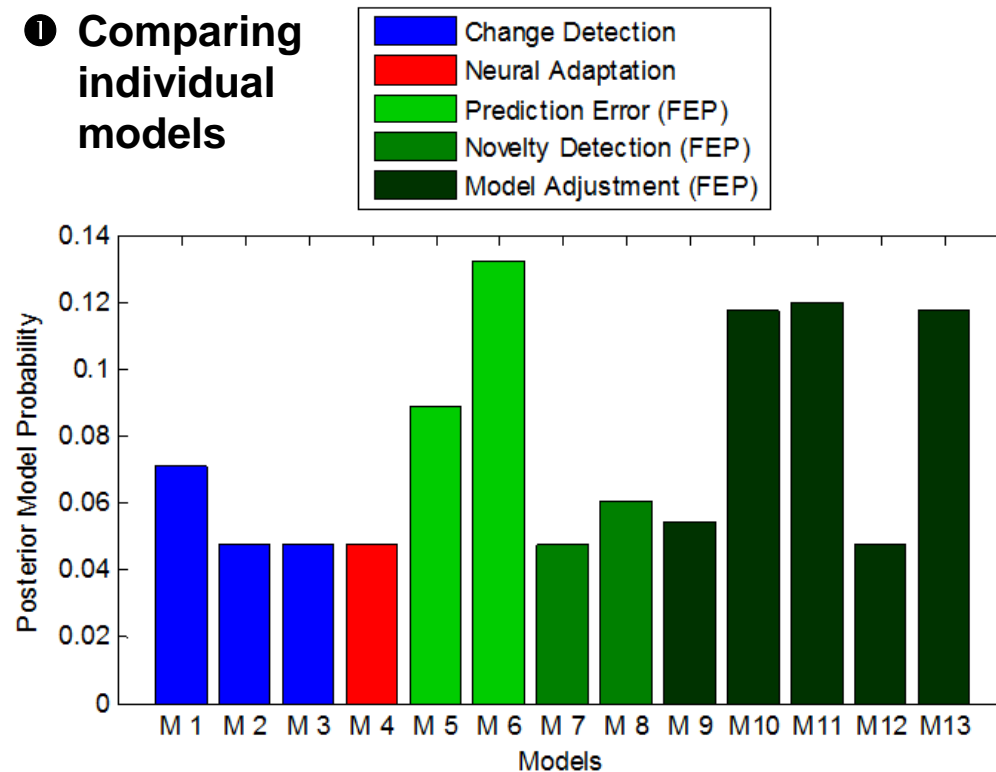
Modelling Trial-by-Trial Changes of the Mismatch Negativity (MMN)



MMN model comparison at multiple levels

② Comparing MMN theories

① Comparing individual models



③ Bayesian Model Averaging (BMA)

- abandons dependence of parameter inference on a single model and takes into account model uncertainty
- uses the entire model space considered (or an optimal family of models)
- averages parameter estimates, weighted by posterior model probabilities
- represents a particularly useful alternative
 - when none of the models (or model subspaces) considered clearly outperforms all others
 - when comparing groups for which the optimal model differs

single-subject BMA:

$$p(\theta | y) \\ = \sum_m p(\theta | y, m) p(m | y)$$

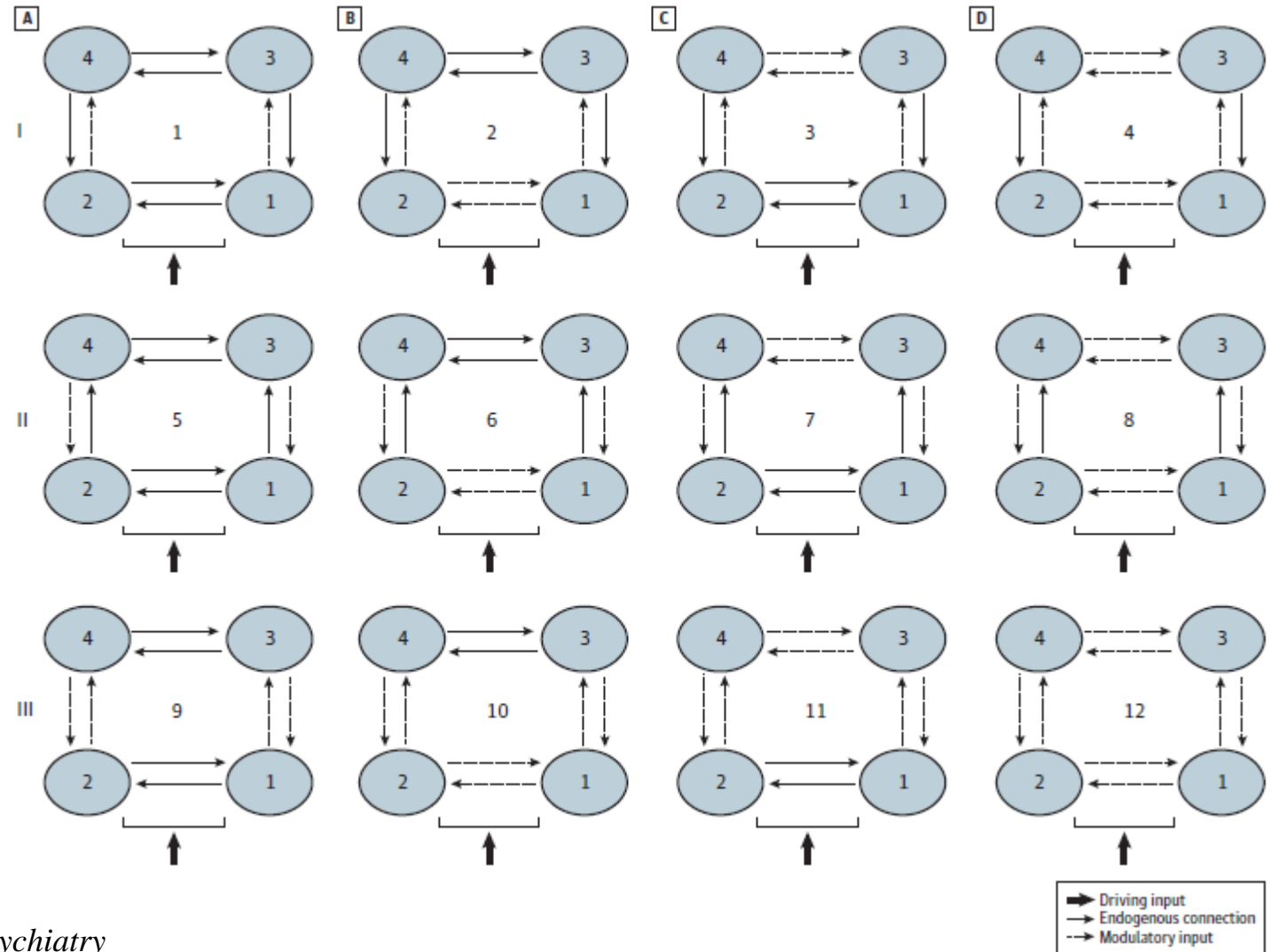
group-level BMA:

$$p(\theta_n | y_{1..N}) \\ = \sum_m p(\theta_n | y_n, m) p(m | y_{1..N})$$

NB: $p(m|y_{1..N})$ can be obtained by either FFX or RFX BMS

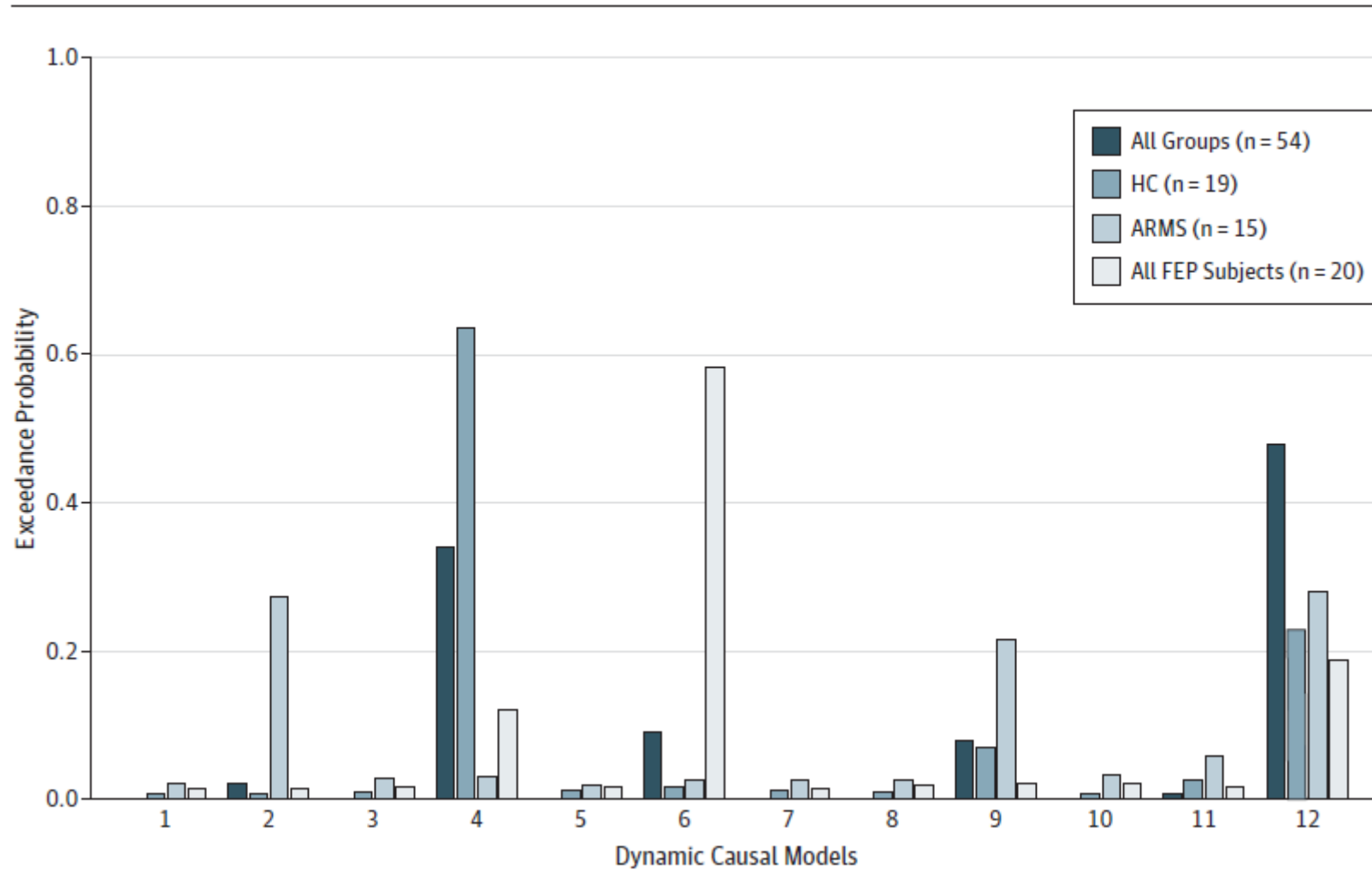


Prefrontal-parietal connectivity during working memory in schizophrenia

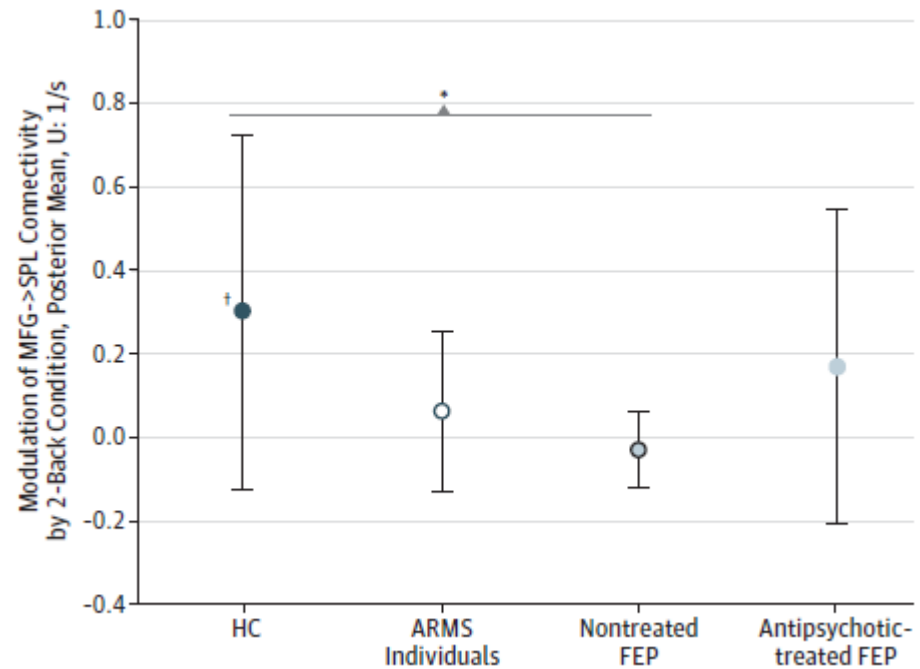
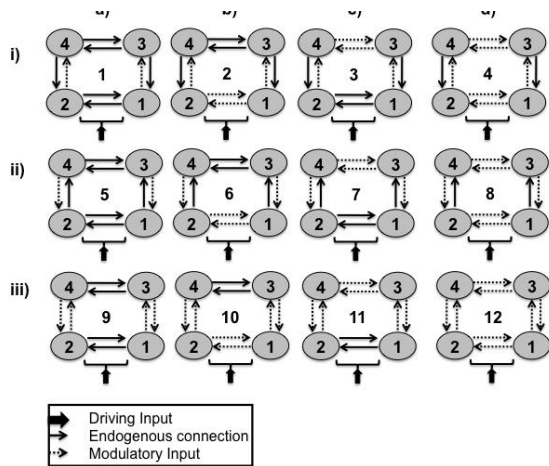
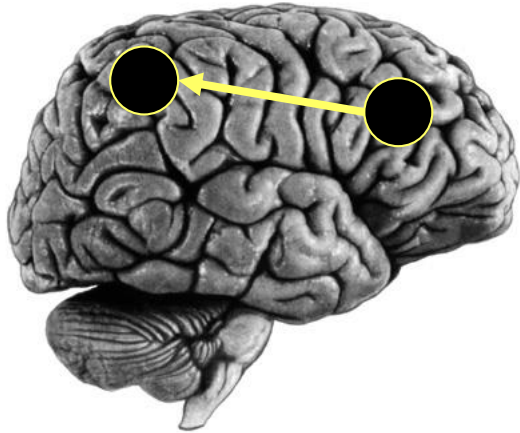


- 17 at-risk mental state (ARMS) individuals
- 21 first-episode patients (13 non-treated)
- 20 controls

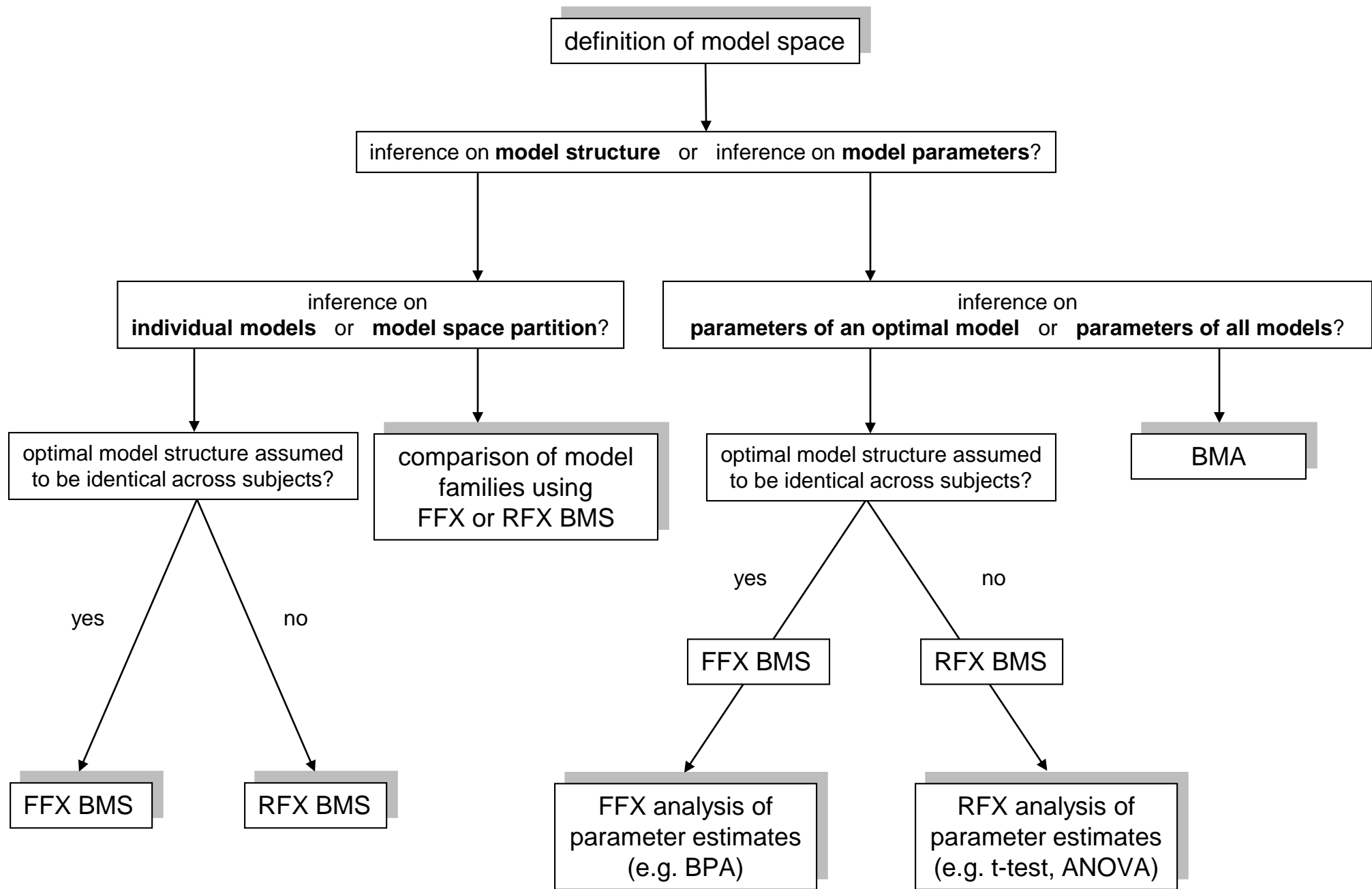
BMS results for all groups



BMA results: PFC → PPC connectivity



17 ARMS, 21 first-episode (13 non-treated),
20 controls



Important topics I have not covered in this talk

- **Bayesian model reduction**
 - an efficient alternative to RFX-BMS (applicable to nested models only)
 - see:
Friston KJ et al. (2016) NeuroImage 128:413-431
<https://www.sciencedirect.com/science/article/pii/S105381191501037X>
- estimating the model evidence with **sampling-based methods**, such as **thermodynamic integration**
 - for a tutorial review, see:
Aponte, Yao et al. (2022) Cognitive Neurodynamics
<https://link.springer.com/article/10.1007/s11571-021-09696-9>

Further reading

- Penny WD, Stephan KE, Mechelli A, Friston KJ (2004) Comparing dynamic causal models. *NeuroImage* 22:1157-1172.
- Penny WD, Stephan KE, Daunizeau J, Joao M, Friston K, Schofield T, Leff AP (2010) Comparing Families of Dynamic Causal Models. *PLoS Computational Biology* 6: e1000709.
- Penny WD (2012) Comparing dynamic causal models using AIC, BIC and free energy. *Neuroimage* 59: 319-330.
- Rigoux L, Stephan KE, Friston KJ, Daunizeau J (2014) Bayesian model selection for group studies – revisited. *NeuroImage* 84: 971-985.
- Stephan KE, Weiskopf N, Drysdale PM, Robinson PA, Friston KJ (2007) Comparing hemodynamic models with DCM. *NeuroImage* 38:387-401.
- Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ (2009) Bayesian model selection for group studies. *NeuroImage* 46:1004-1017.
- Stephan KE, Penny WD, Moran RJ, den Ouden HEM, Daunizeau J, Friston KJ (2010) Ten simple rules for Dynamic Causal Modelling. *NeuroImage* 49: 3099-3109.
- Stephan KE, Iglesias S, Heinzle J, Diaconescu AO (2015) Translational Perspectives for Computational Neuroimaging. *Neuron* 87: 716-732.
- Stephan KE, Schlagenhauf F, Huys QJM, Raman S, Aponte EA, Brodersen KH, Rigoux L, Moran RJ, Daunizeau J, Dolan RJ, Friston KJ, Heinz A (2017) Computational Neuroimaging Strategies for Single Patient Predictions. *NeuroImage* 145: 180-199.

Thank you