Bayesian model selection & averaging

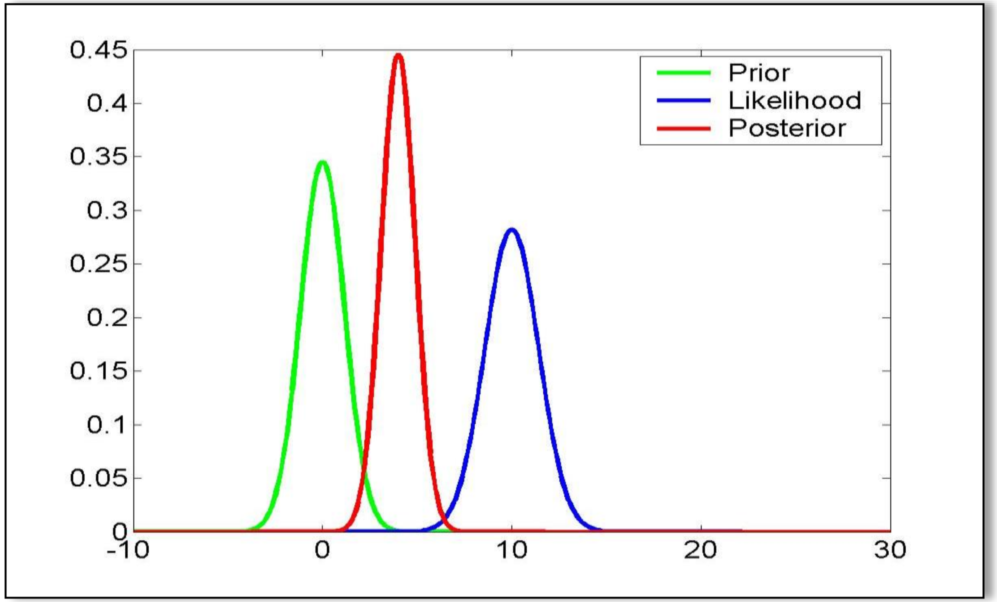
Klaas Enno Stephan

Translational Neuromodeling Unit

Universität Zürich UZH

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Bayes' theorem



The Reverend Thomas Bayes
(1702-1761)

$$p(\theta \mid y, m) = \frac{p(y \mid \theta, m)\, p(\theta \mid m)}{p(y \mid m)}$$

posterior = likelihood • prior / evidence

Posterior mean & variance of univariate Gaussians

$$y = \theta + \varepsilon$$

Likelihood & Prior

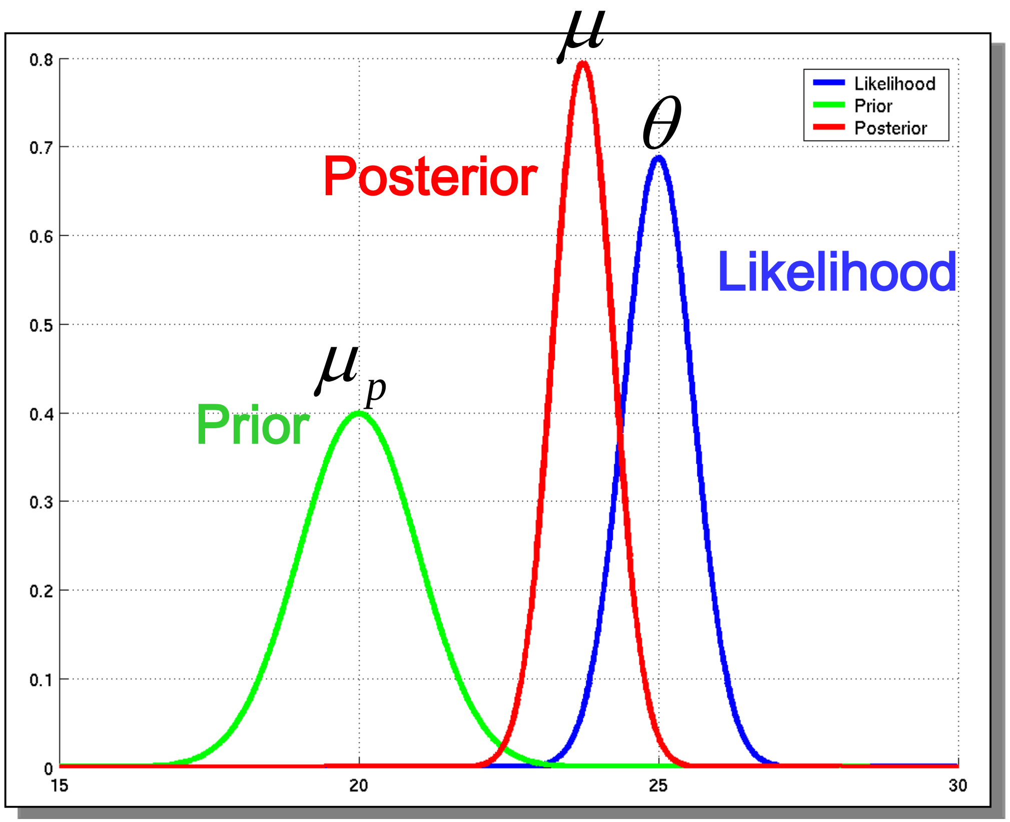
$$p(y \mid \theta) = N(\theta, \sigma_e^2)$$

$$p(\theta) = N(\mu_p, \sigma_p^2)$$

Posterior: $p(\theta \mid y) = N(\mu, \sigma^2)$

$$\frac{1}{\sigma^2} = \frac{1}{\sigma_e^2} + \frac{1}{\sigma_p^2}$$

$$\mu = \sigma^2 \left(\frac{1}{\sigma_e^2} \theta + \frac{1}{\sigma_p^2} \mu_p \right)$$



Posterior mean = variance-weighted combination of prior mean and data mean

Same thing – but expressed as precision weighting

Likelihood & prior

$$p(y\,|\,\theta) = N(\theta, \lambda_e^{-1})$$

$$p(\theta) = N(\mu_p, \lambda_p^{-1})$$

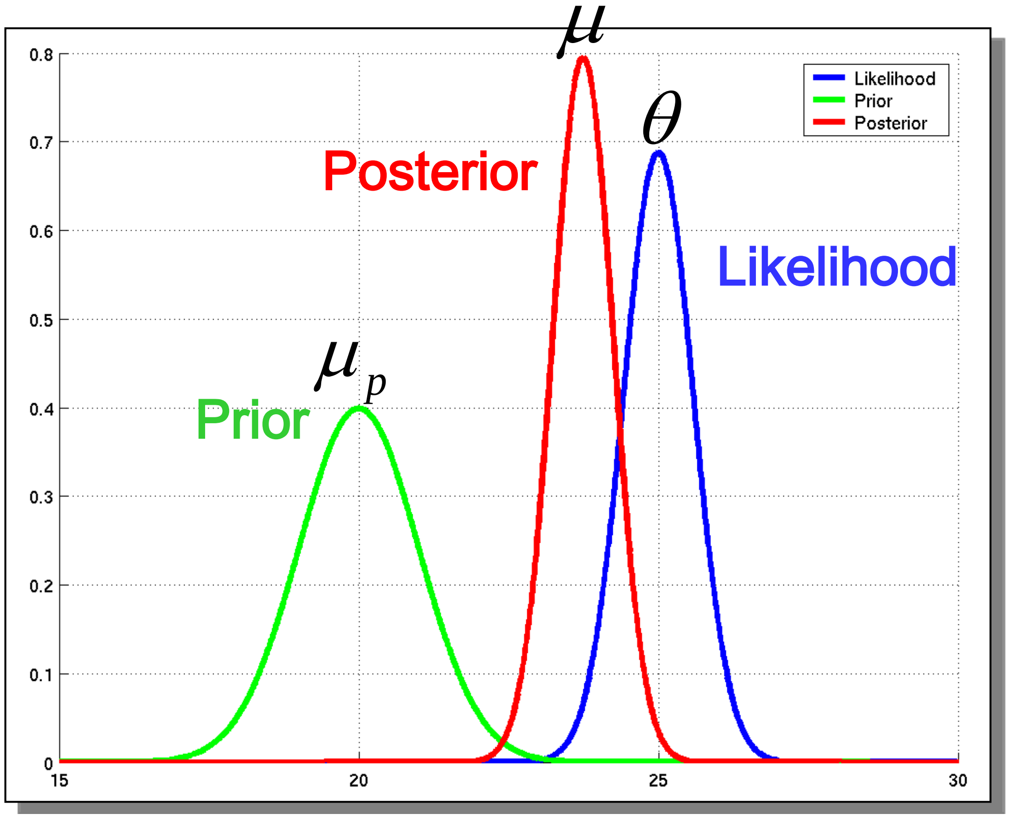
$$y = \theta + \varepsilon$$

Posterior: $p(\theta\,|\,y) = N(\mu, \lambda^{-1})$

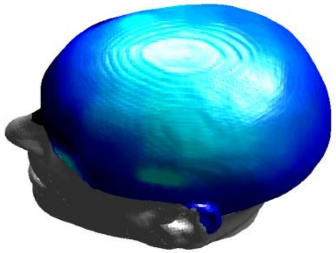
$$\lambda = \lambda_e + \lambda_p$$

$$\mu = \frac{\lambda_e}{\lambda}\theta + \frac{\lambda_p}{\lambda}\mu_p$$

Relative precision weighting



Generative model

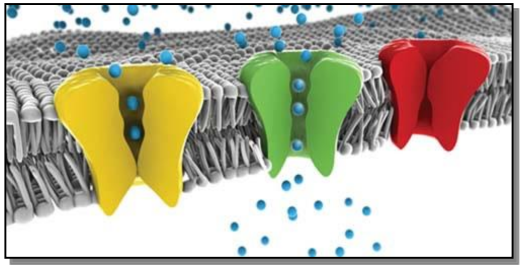


$$p(y \mid \theta, m) \cdot p(\theta \mid m)$$

$$\longleftarrow$$

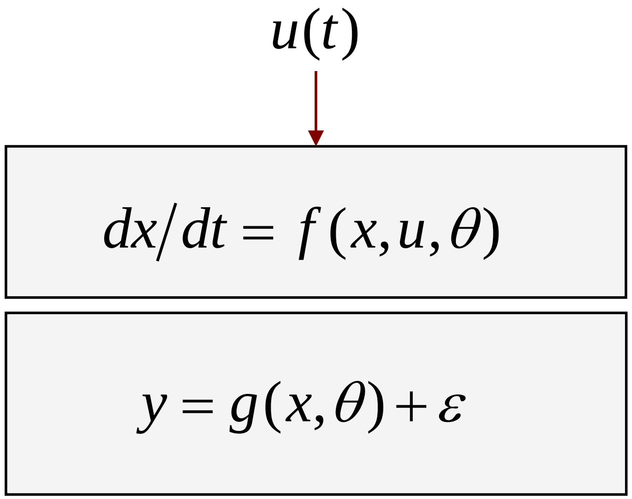
$$\longrightarrow$$

$$p(\theta \mid y, m)$$



1. enforces mechanistic thinking: how could the data have been caused?
2. generate synthetic data (observations) by sampling from the prior – can model explain certain phenomena at all?
3. inference about parameters → p(θ|y)
4. inference about model structure: formal approach to disambiguating mechanisms → p(y|m) or p(m|y)

Model inversion



$u(t)$

Neural dynamics

$$dx/dt = f(x,u,\theta)$$

Observer function

$$y = g(x,\theta) + \varepsilon$$

$$p(y \mid \theta, m) = N(g(\theta), \Sigma(\theta))$$

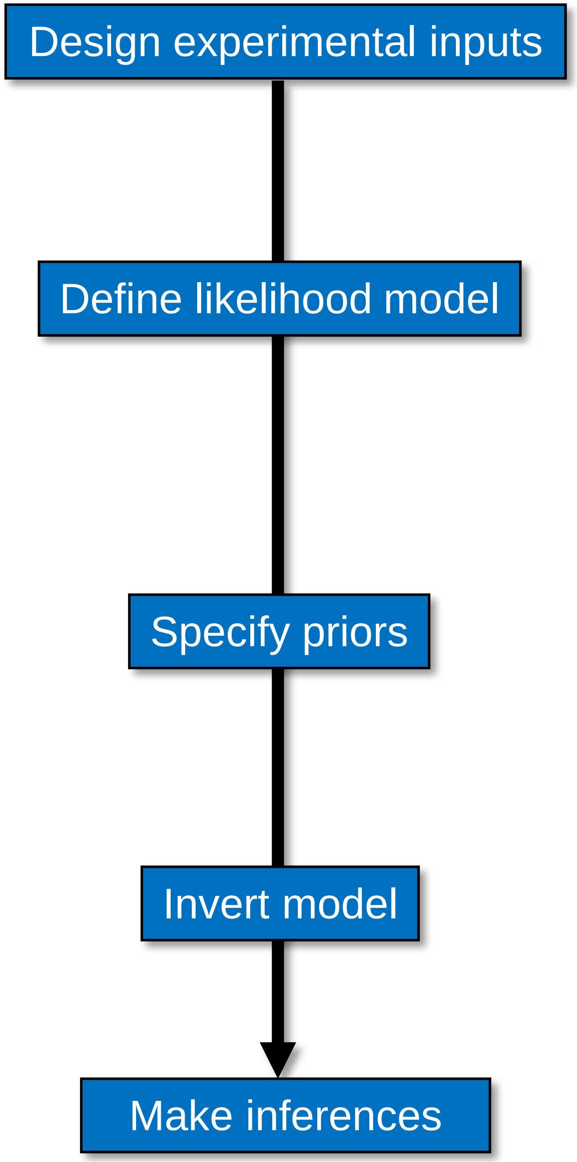
$$p(\theta, m) = N(\mu_\theta, \Sigma_\theta)$$

Inference on model structure

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

Inference on parameters

$$p(\theta \mid y, m) = \frac{p(y \mid \theta, m)\, p(\theta, m)}{p(y \mid m)}$$



Design experimental inputs

Define likelihood model

Specify priors

Invert model

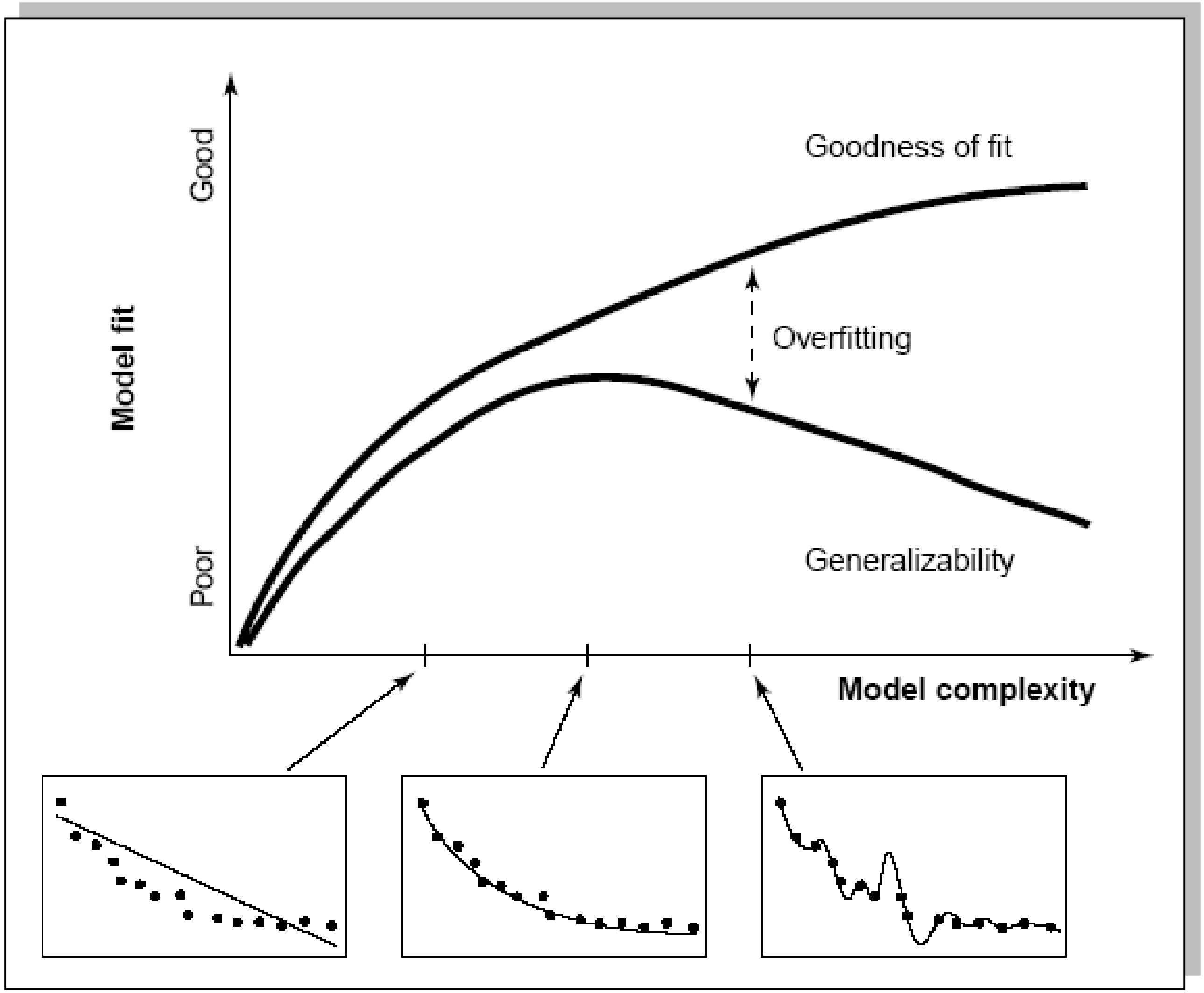
Make inferences

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?

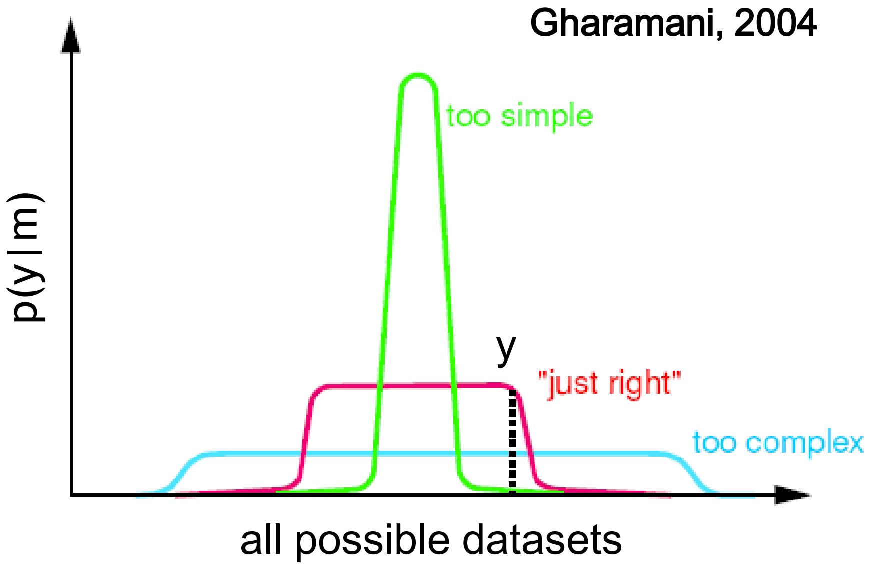


Pitt & Miyung (2002) *TICS*

Bayesian model selection (BMS)

Model evidence (marginal likelihood):

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



Gharamani, 2004

- accounts for both accuracy and complexity of the model

- “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?”

Various approximations, e.g.:

- negative free energy, AIC, BIC

McKay 1992, *Neural Comput.*
Penny et al. 2004a, *NeuroImage*

Model space (hypothesis set) M

Model space M is defined by prior on models.

Usual choice: flat prior over a small set of models.

$$p(m) = \begin{cases} 1/\left|M\right| \text{ if } m \in M \\ 0 \text{ if } m \notin M \end{cases}$$

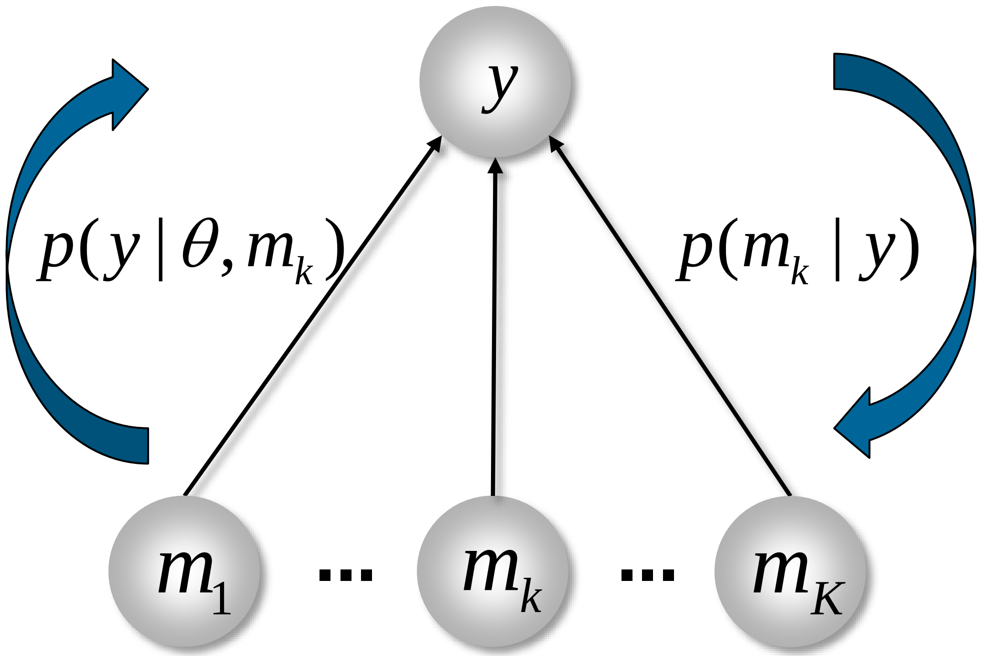
In this case, the posterior probability of model i is:

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

Long-term goal: Differential diagnosis based on generative models of disease symptoms

SYMPTOM
(behaviour
or physiology)

HYPOTHETICAL
MECHANISM



$$p(m_k \mid \mathrm{y}) = \frac{p(y \mid m_k)\, p(m_k)}{\sum_k p(y \mid m_k)\, p(m_k)}$$

Approximations to the log evidence

Logarithm is a monotonic function → Maximizing log model evidence = Maximizing model evidence

Log model evidence = balance between fit and complexity

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

No. of parameters

Akaike Information Criterion: $AIC = \log p(y \mid \theta, m) - \boxed{p}$

No. of data points

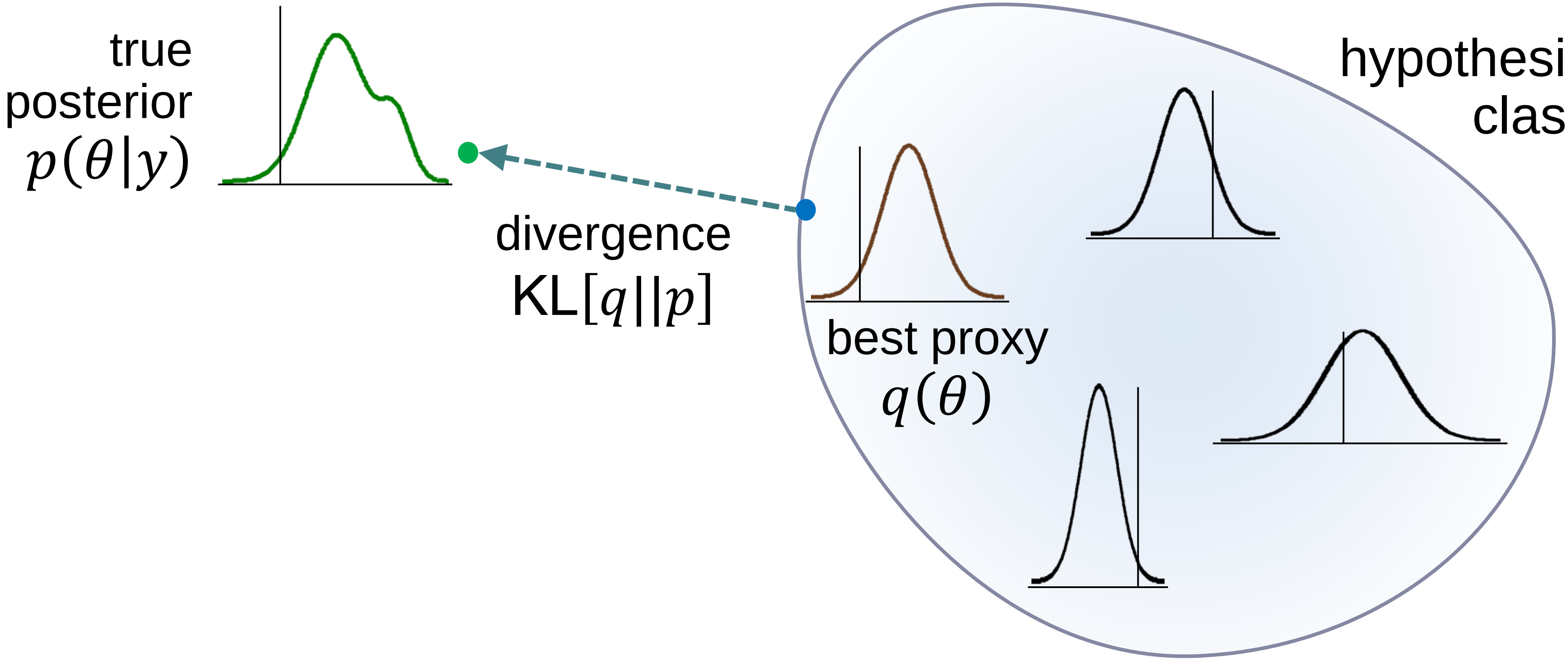
Bayesian Information Criterion: $BIC = \log p(y \mid \theta, m) - \frac{p}{2} \log \boxed{N}$

Penny et al. 2004a, *NeuroImage*

Variational Bayes (VB)

Idea: 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 optimizing its sufficient statistics.



The (negative) free energy approximation F

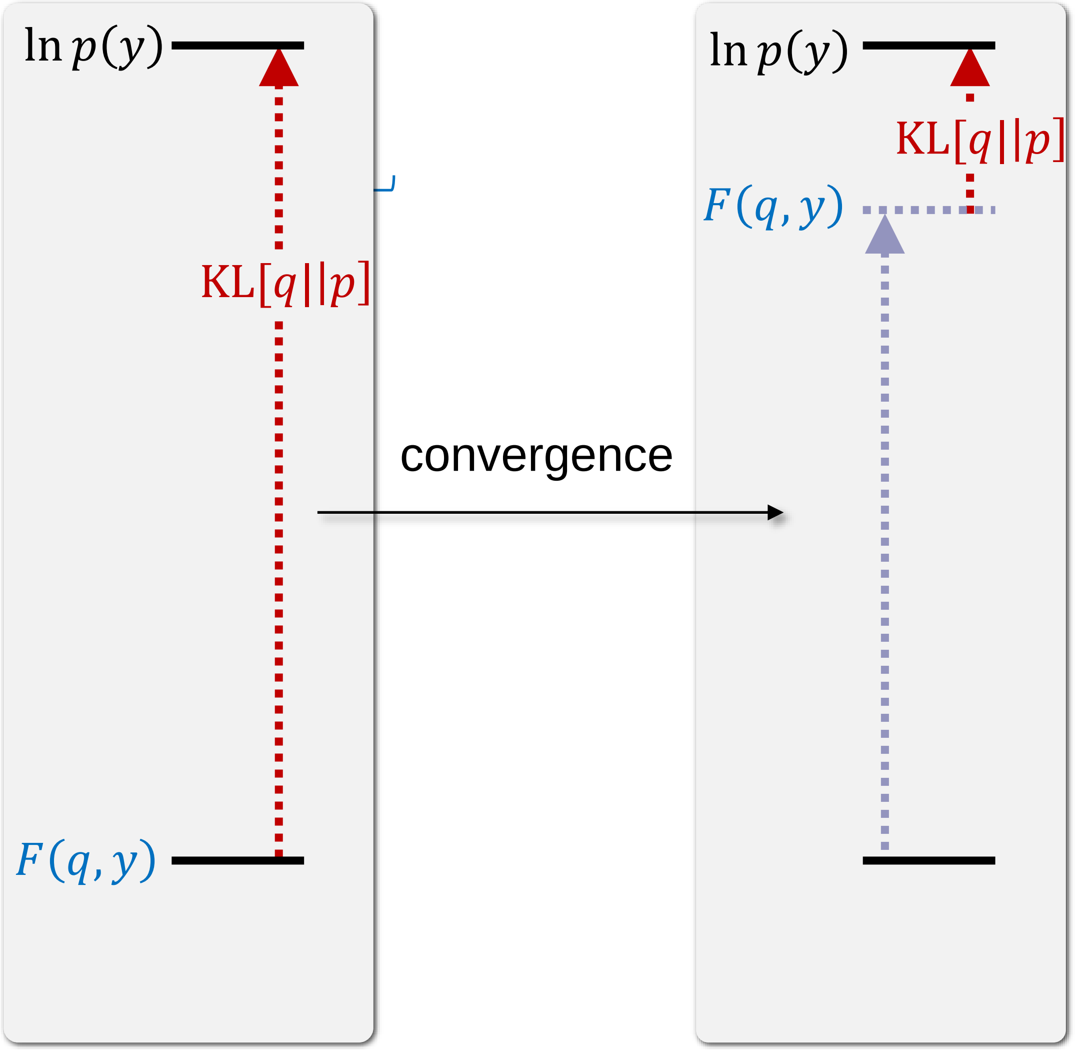
$$\ln p(y) = \underbrace{\mathrm{KL}[q||p]}_{} + F(q,y)$$

divergence ≥ 0 (unknown)

neg. free energy (easy to evaluate for a given q)

Maximizing $F(q,y)$

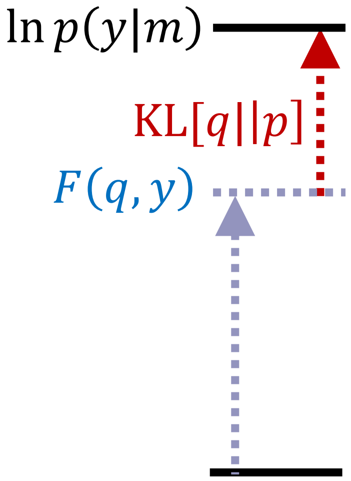
- minimises $\mathrm{KL}[q||p]$
- obtains a lower bound approximation to the log evidence
- obtains $q(\theta)$ as our best estimate of the posterior



The (negative) free energy approximation *F*

F is a lower bound on the log model evidence:

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



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

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

The complexity term in *F*

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

$$KL\left[q(\theta), p(\theta \mid m)\right]$$

$$= \frac{1}{2}\ln\left|C_\theta\right| - \frac{1}{2}\ln\left|C_{\theta|y}\right| + \frac{1}{2}\left(\mu_{\theta|y} - \mu_\theta\right)^T C_\theta^{-1}\left(\mu_{\theta|y} - \mu_\theta\right)$$

- determinant = measure of “volume” (space spanned by the eigenvectors of the matrix)

- The complexity term of *F* is higher
  - the more independent the prior parameters (↑ effective DFs)
  - the more dependent the posterior parameters (i.e., poor identifiability is penalised!)
  - 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:

positive value, [0; ∞[

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

Kass & Raftery classification:

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

Kass & Raftery 1995, *J. Am. Stat. Assoc.*

Fixed effects BMS at group level

Group Bayes factor (GBF) for 1...*K* subjects:

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

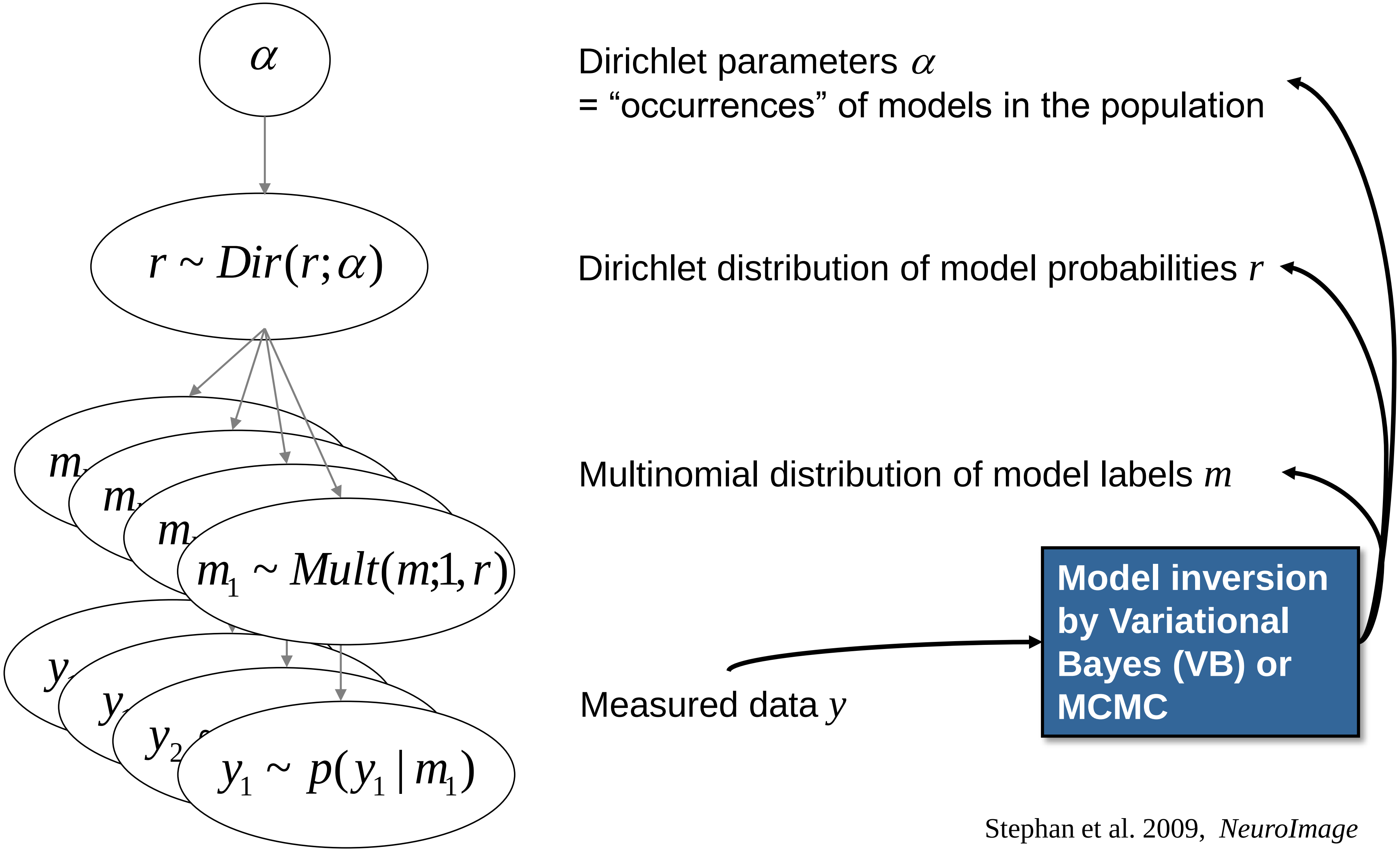
Average Bayes factor (ABF):

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

Problems:

- blind with regard to group heterogeneity
- sensitive to outliers

Random effects BMS for heterogeneous groups



α

$r \sim Dir(r;\alpha)$

$m_1 \sim Mult(m;1,r)$

$y_1 \sim p(y_1 \mid m_1)$

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

Dirichlet distribution of model probabilities r

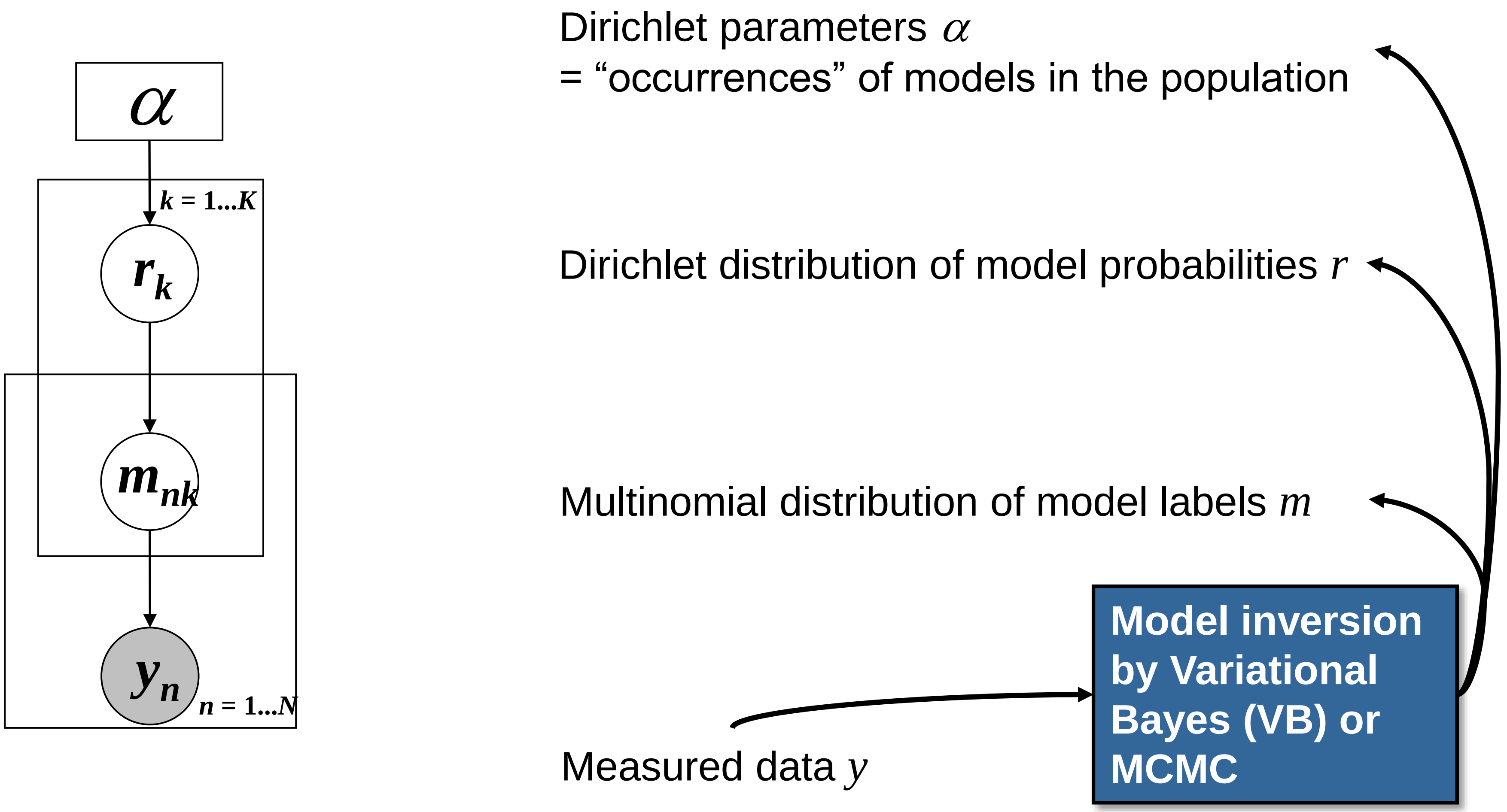
Multinomial distribution of model labels m

Measured data y

Model inversion by Variational Bayes (VB) or MCMC

Stephan et al. 2009, *NeuroImage*

Random effects BMS for heterogeneous groups



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

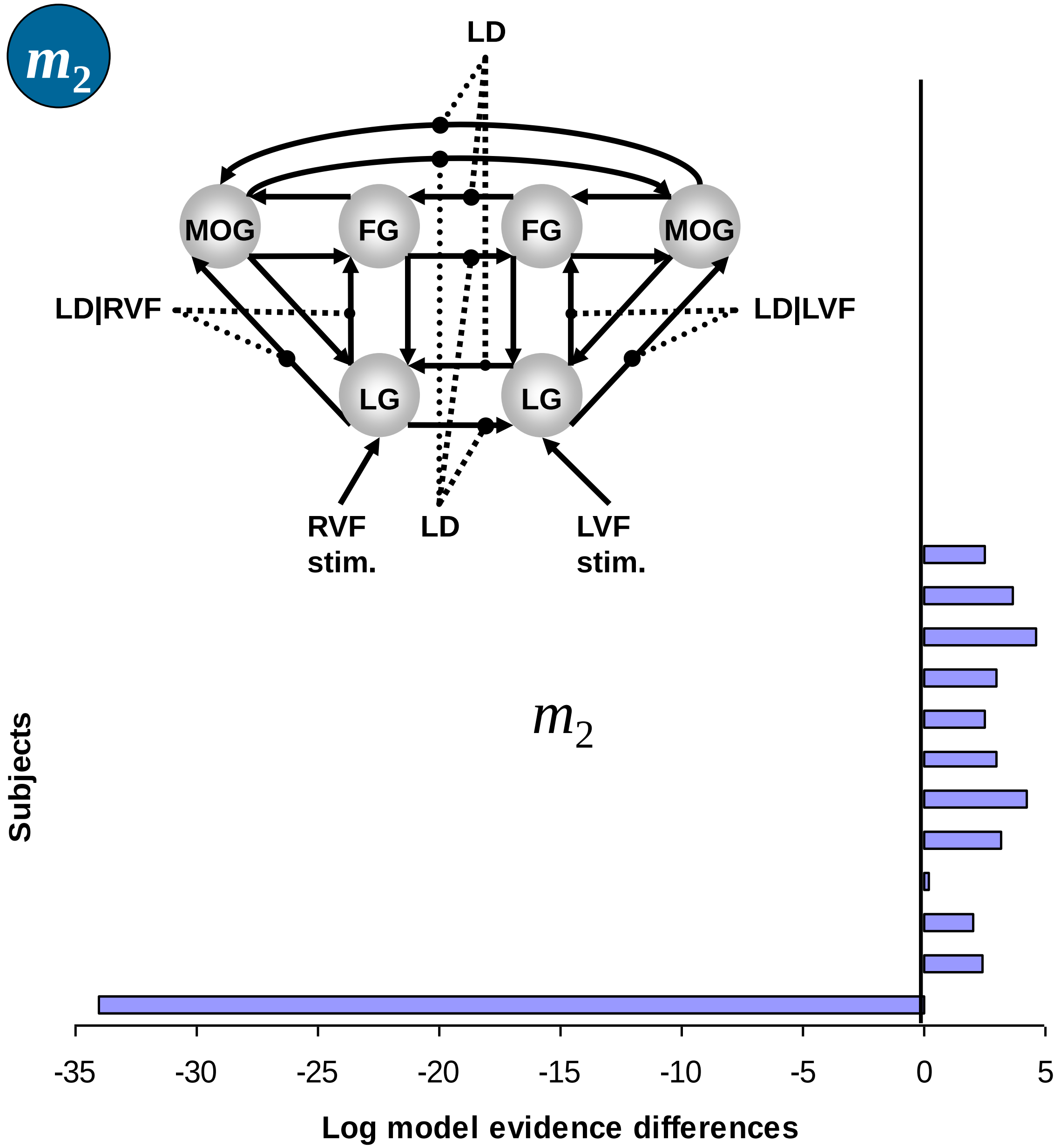
Dirichlet distribution of model probabilities r

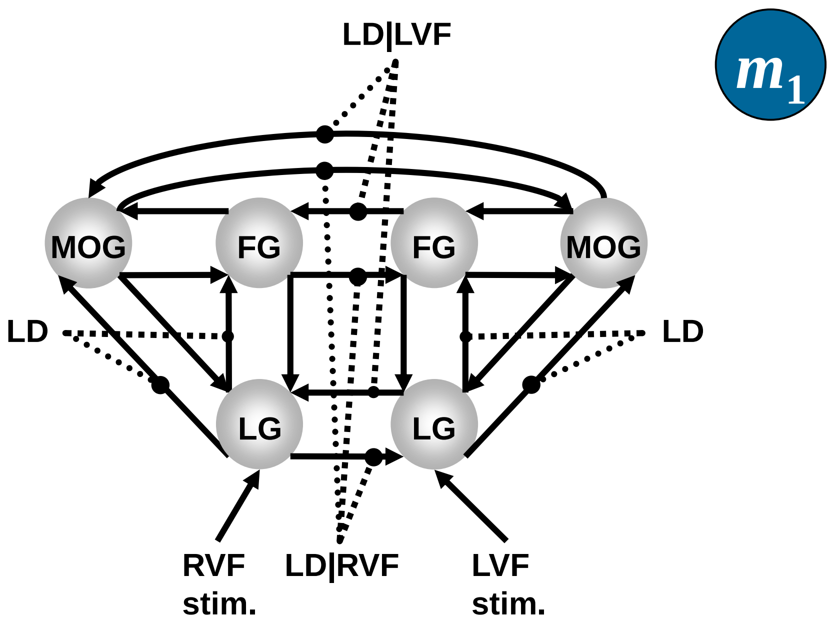
Multinomial distribution of model labels m

Measured data y

Model inversion by Variational Bayes (VB) or MCMC

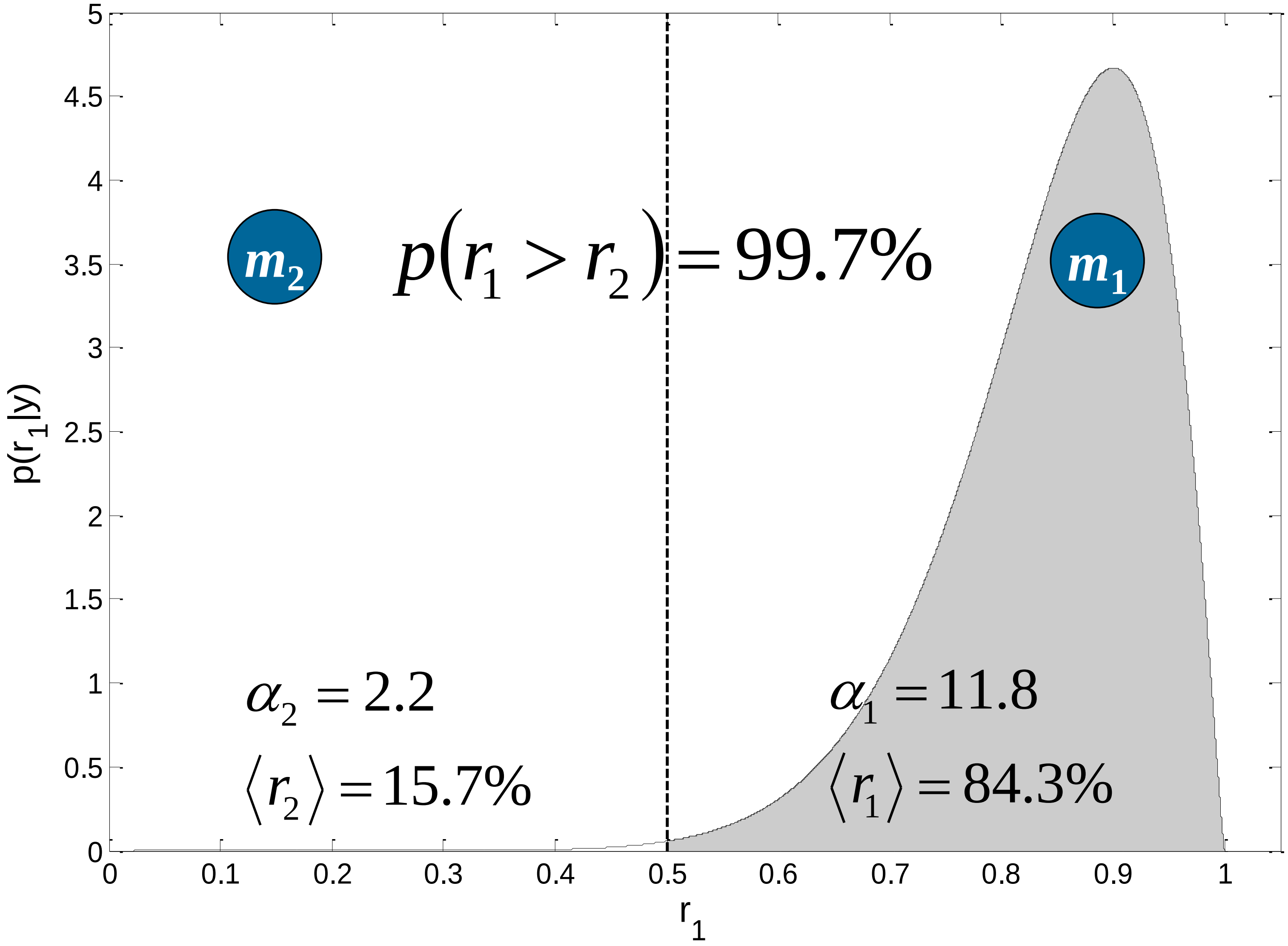
Stephan et al. 2009, *NeuroImage*





m_1

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



Stephan et al. 2009a, *NeuroImage*

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

$$\left\langle r_k \right\rangle_q = \alpha_k / (\alpha_1 + \ldots + \alpha_K)$$

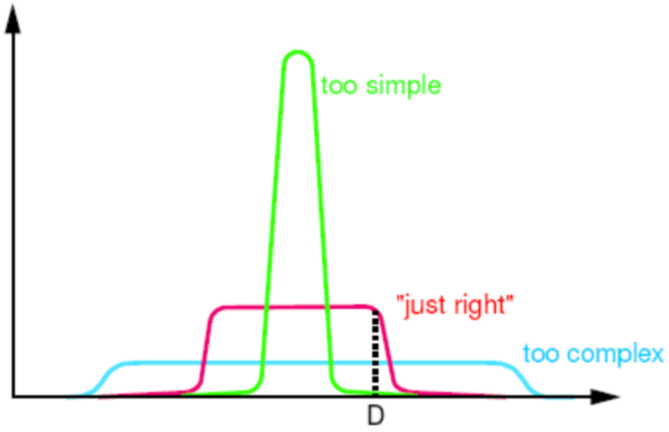
3. **exceedance probability** that a particular model *k* is more likely than any other model (of the *K* models tested), given the group data

$$\exists k \in \{1...K\}, \forall j \in \{1...K \mid j \neq k\}:$$
$$\varphi_k = p(r_k > r_j \mid y; \alpha)$$

4. **protected exceedance probability**: see below

Overfitting at the level of models

- ↑ #models ⇒ ↑ risk of overfitting
- solutions:
  - regularisation: definition of model space = choosing priors p(m)
  - family-level BMS
  - Bayesian model averaging (BMA)



posterior model probability:

$$p\left(m \mid y\right) = \frac{p\left(y \mid m\right) p(m)}{\sum_m p\left(y \mid m\right) p(m)}$$

BMA:

$$p\left(\theta \mid y\right) = \sum_m p\left(\theta \mid y, m\right) p\left(m \mid y\right)$$

Model space partitioning: comparing model families

- partitioning model space into K subsets or families:

$$M = \left\{ f_1, \ldots, f_K \right\}$$

- pooling information over all models in these subsets allows one to compute the probability of a model family, given the data

$$p\left(f_k \right)$$

- effectively removes uncertainty about any aspect of model structure, other than the attribute of interest (which defines the partition)

Stephan et al. 2009, *NeuroImage*
Penny et al. 2010, *PLoS Comput. Biol.*

Family-level inference: fixed effects

- We wish to have a uniform prior at the family level:

$$p\left(f_k \right) = \frac{1}{K}$$

- This is related to the model level via the sum of the priors on models:

$$p\left(f_k \right) = \sum_{m \in f_k} p(m)$$

- Hence the uniform prior at the family level is:

$$\forall m \in f_k : p\left(m\right) = \frac{1}{K\left| f_k \right|}$$

- The probability of each family is then obtained by summing the posterior probabilities of the models it includes:

$$p\left(f_k \mid y_{1..N} \right) = \sum_{m \in f_k} p\left(m \mid y_{1..N} \right)$$

Penny et al. 2010, *PLoS Comput. Biol.*

Family-level inference: random effects

- The frequency of a family in the population is given by:

$$s_k = \sum_{m \in f_k} r_m$$

- In RFX-BMS, this follows a Dirichlet distribution, with a uniform prior on the parameters α (see above).

$$p(s) = Dir\left(\alpha\right)$$

- A uniform prior over family probabilities can be obtained by setting:

$$\forall m \in f_k : \alpha_{prior}(m) = \frac{1}{|f_k|}$$

Stephan et al. 2009, *NeuroImage*
Penny et al. 2010, *PLoS Comput. Biol.*

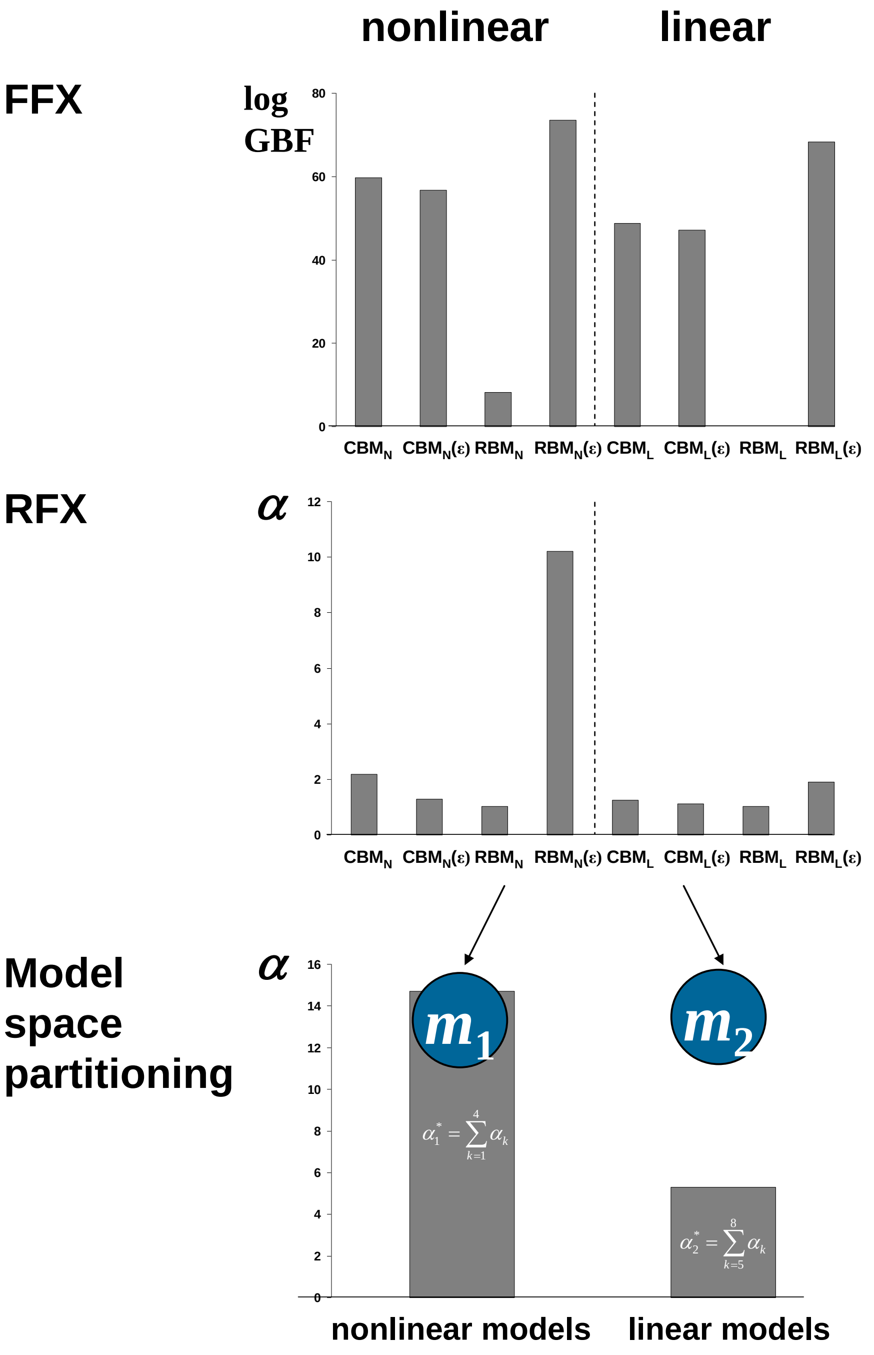
Family-level inference: random effects – a special case

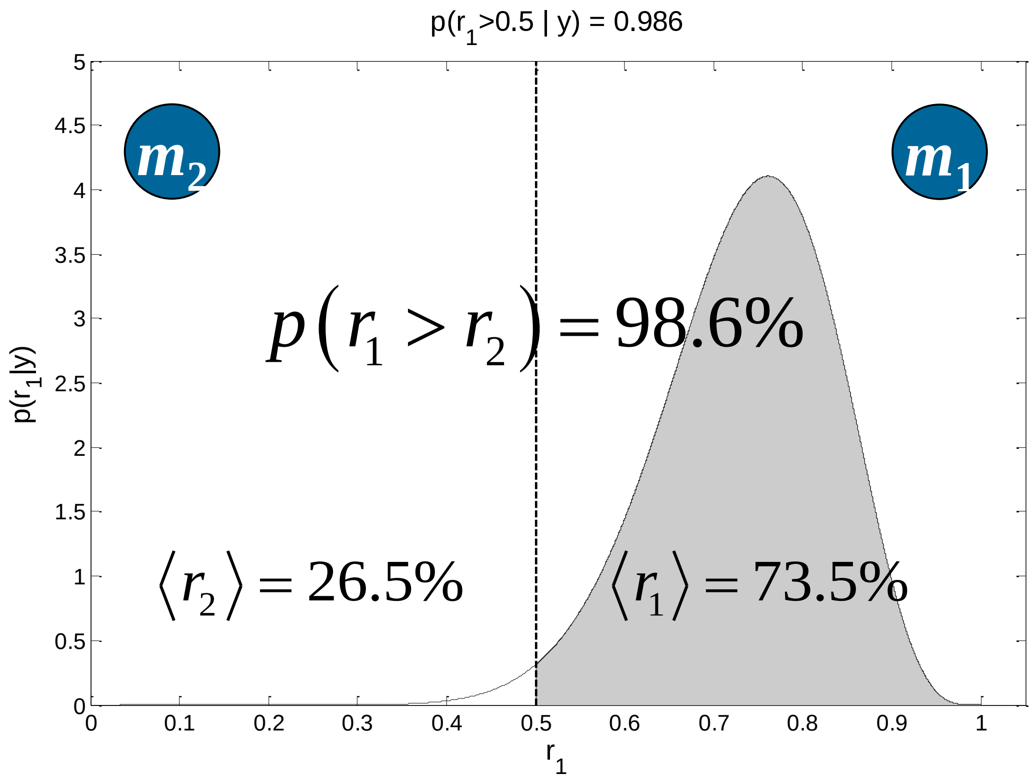
- When the families are of equal size, one can simply sum the posterior model probabilities within families by exploiting the agglomerative property of the Dirichlet distribution:

$$
\begin{aligned}
\left(r_1, r_2, \ldots, r_K\right) &\sim Dir\left(\alpha_1, \alpha_2, \ldots, \alpha_K\right) \\
&\Rightarrow r_1^* = \sum_{k \in N_1} r_k, r_2^* = \sum_{k \in N_2} r_k, \ldots, r_J^* = \sum_{k \in N_J} r_k \\
&\sim Dir\left(\alpha_1^* = \sum_{k \in N_1} \alpha_k, \alpha_2^* = \sum_{k \in N_2} \alpha_k, \ldots, \alpha_J^* = \sum_{k \in N_J} \alpha_k\right)
\end{aligned}
$$

Stephan et al. 2009, *NeuroImage*

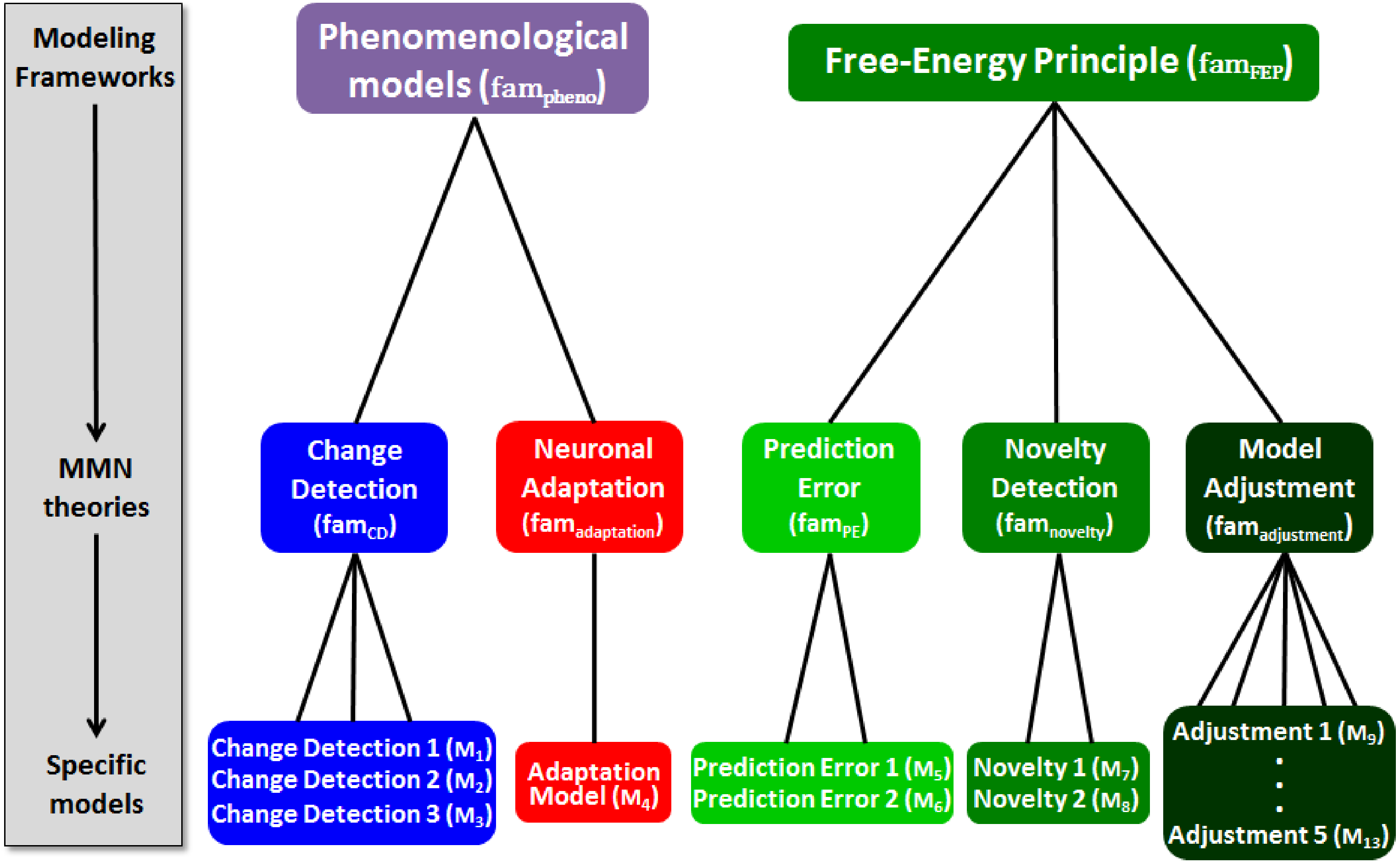
Model space partitioning: comparing model families





Stephan et al. 2009, *NeuroImage*

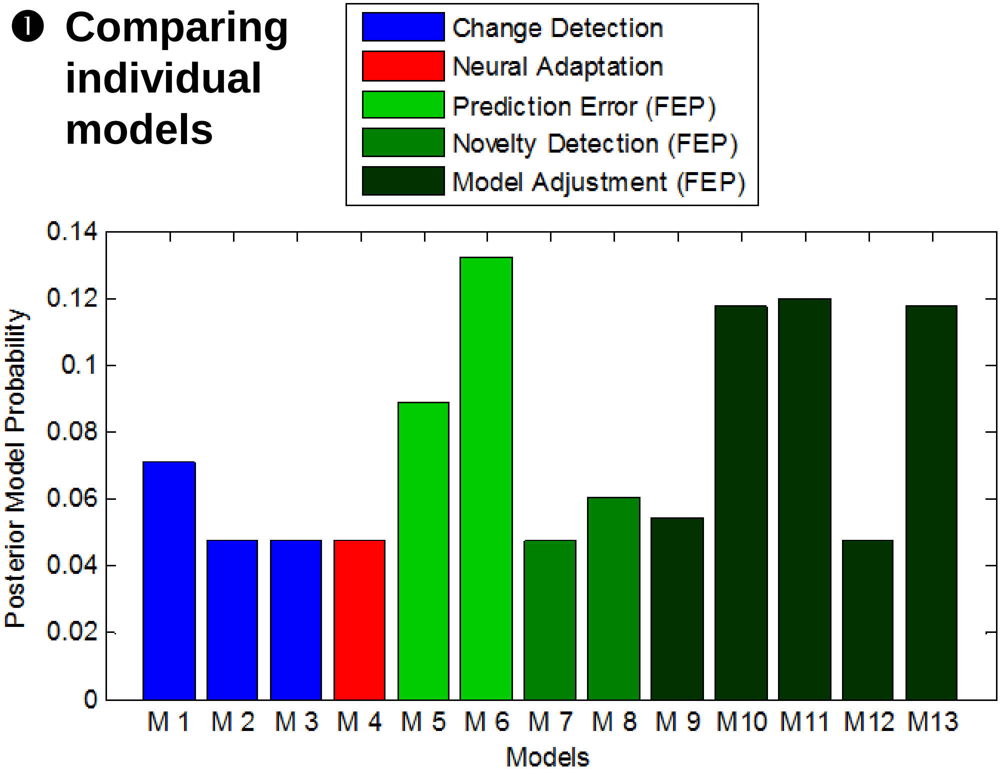
Modelling Trial-by-Trial Changes of the Mismatch Negativity (MMN)


Modeling Frameworks
MMN theories
Specific models
Phenomenological models (fam_{pheno})
Free-Energy Principle (fam_{FEP})
Change Detection (fam_{CD})
Neuronal Adaptation ($fam_{adaptation}$)
Prediction Error (fam_{PE})
Novelty Detection ($fam_{novelty}$)
Model Adjustment ($fam_{adjustment}$)
Change Detection 1 (M_1)
Change Detection 2 (M_2)
Change Detection 3 (M_3)
Adaptation Model (M_4)
Prediction Error 1 (M_5)
Prediction Error 2 (M_6)
Novelty 1 (M_7)
Novelty 2 (M_8)
Adjustment 1 (M_9)
:
Adjustment 5 (M_{13})

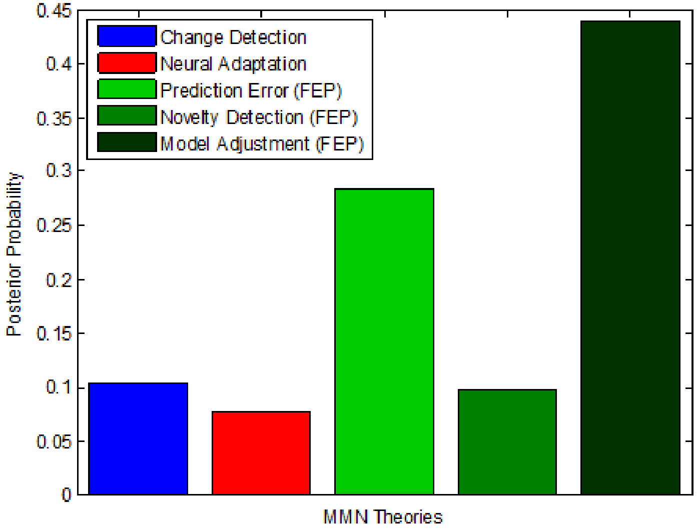
Lieder et al. 2013, *PLoS Comput. Biol.*

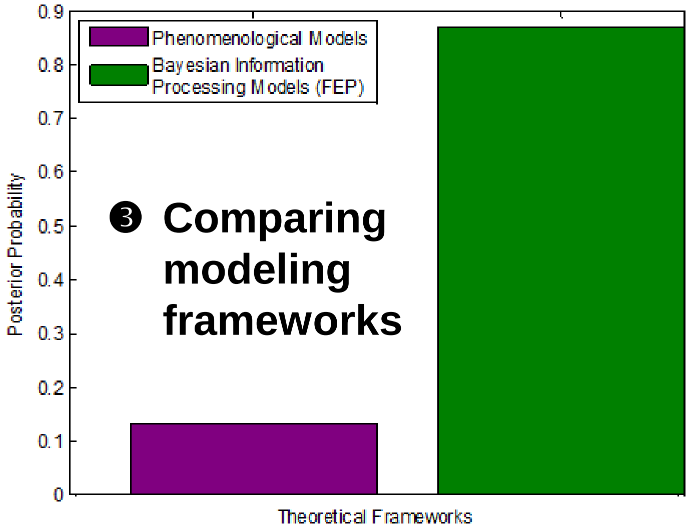
MMN model comparison at multiple levels

❶ **Comparing individual models**



❷ **Comparing MMN theories**





❸ **Comparing modeling frameworks**

Lieder et al. 2013, *PLoS Comput. Biol.*

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:

$$p\left(\theta \mid y\right) = \sum_{m} p\left(\theta \mid y, m\right) p\left(m \mid y\right)$$

group-level BMA:

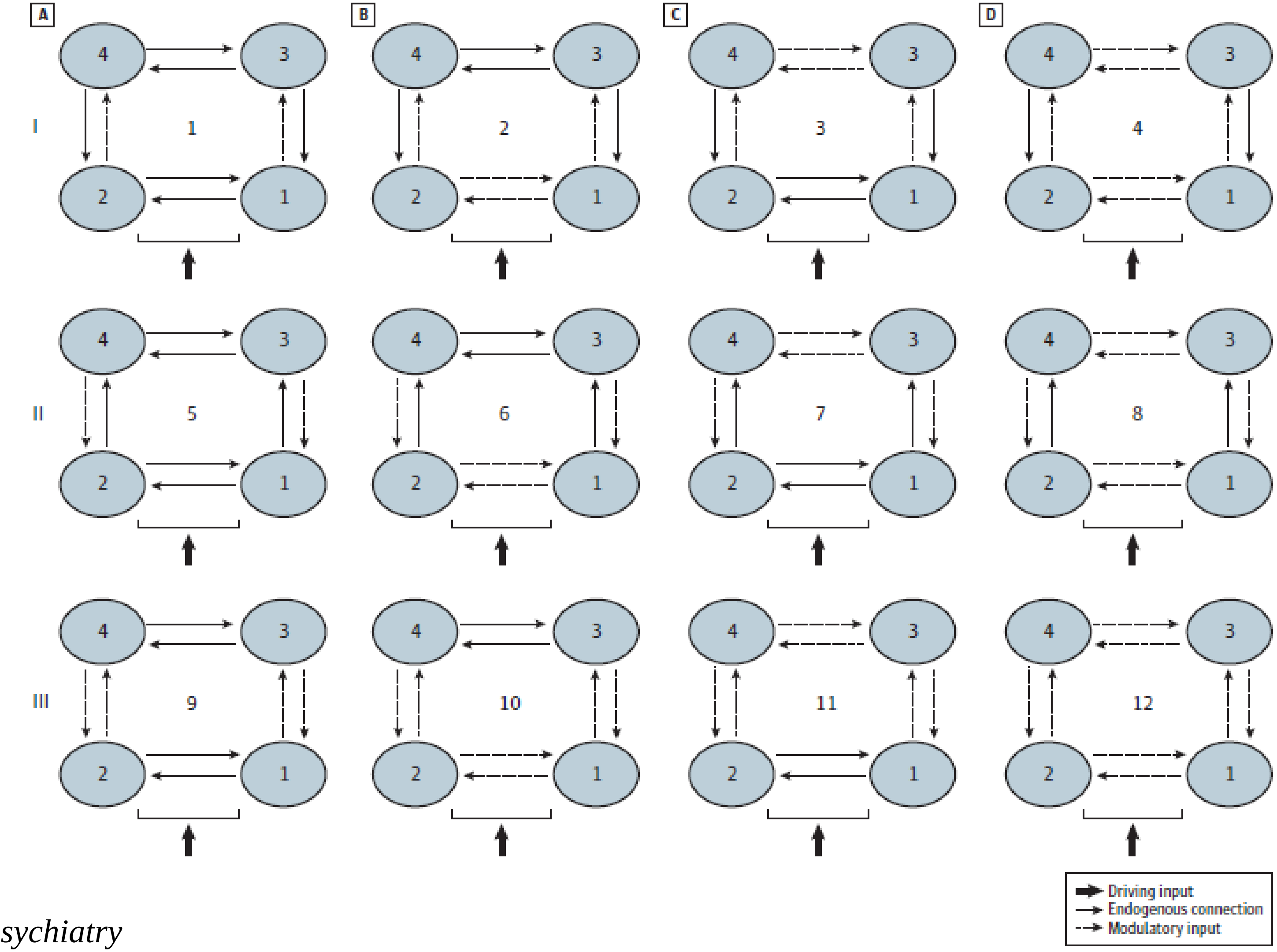
$$p\left(\theta_n \mid y_{1..N}\right) = \sum_{m} p\left(\theta_n \mid y_n, m\right) p\left(m \mid y_{1..N}\right)$$

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

Penny et al. 2010, *PLoS Comput. Biol.*

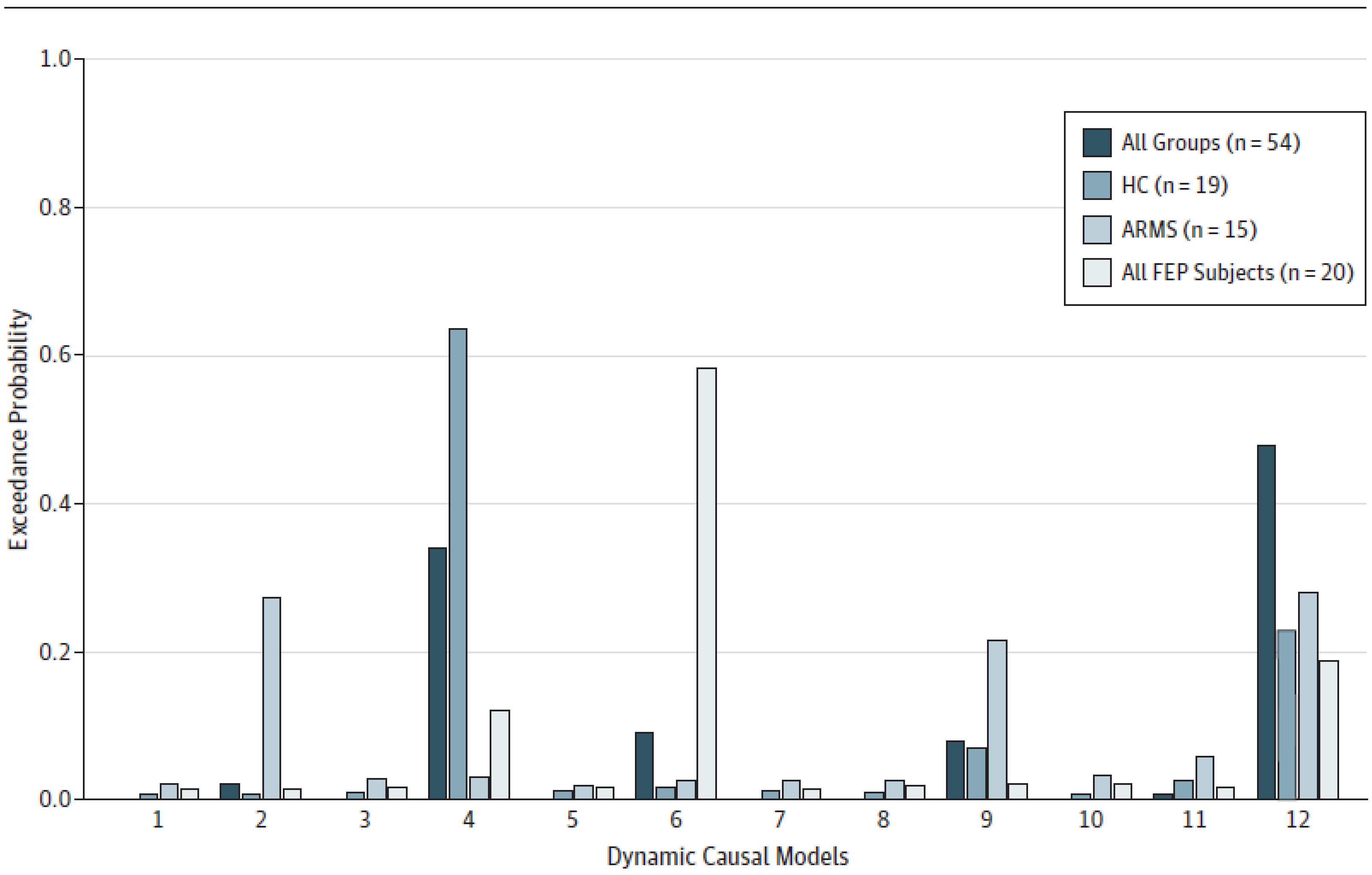
Prefrontal-parietal connectivity during working memory in schizophrenia

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



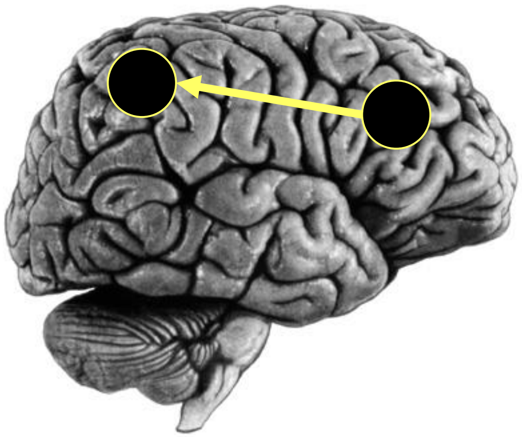
Schmidt et al. 2013, *JAMA Psychiatry*

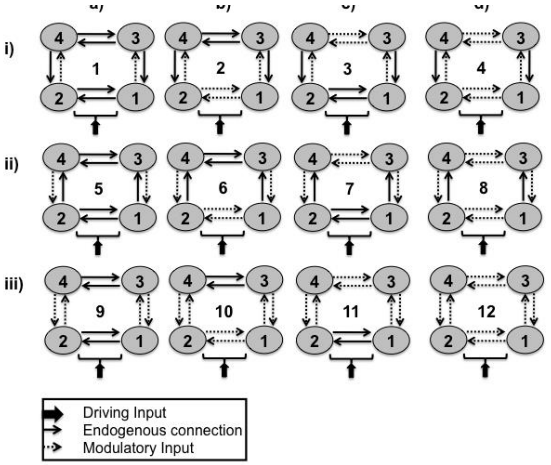
BMS results for all groups

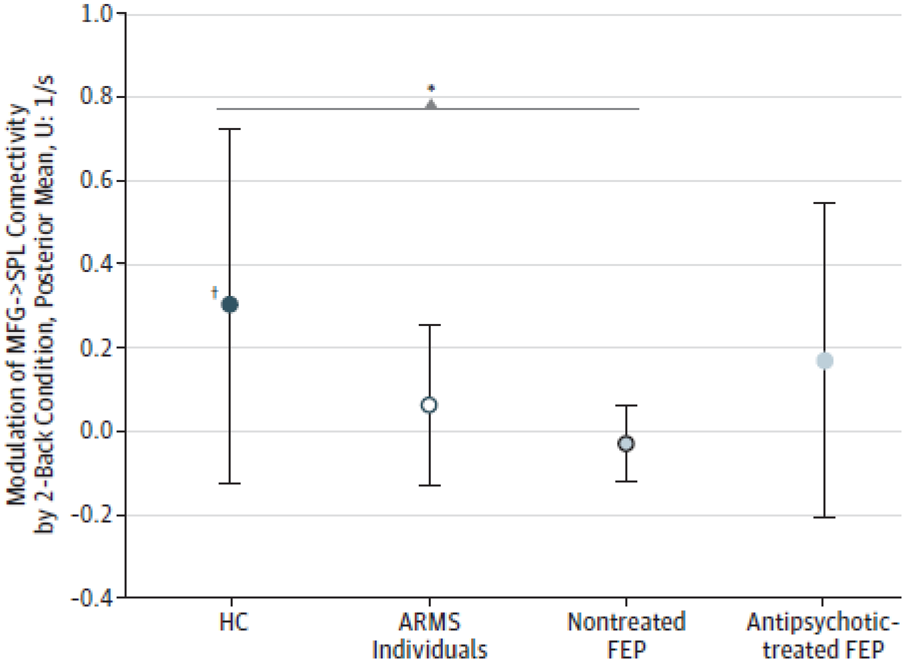


Schmidt et al. 2013, *JAMA Psychiatry*

BMA results: PFC → PPC connectivity







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

Schmidt et al. 2013, *JAMA Psychiatry*

Protected exceedance probability: Using BMA to protect against chance findings

- EPs 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 possibility "null hypothesis" H_0: $r_k = 1/K$
- **Bayesian omnibus risk (BOR)** of wrongly accepting H_1 over H_0:

$$P_o = \frac{1}{1 + \frac{p(m|H_1)}{p(m|H_0).}}$$

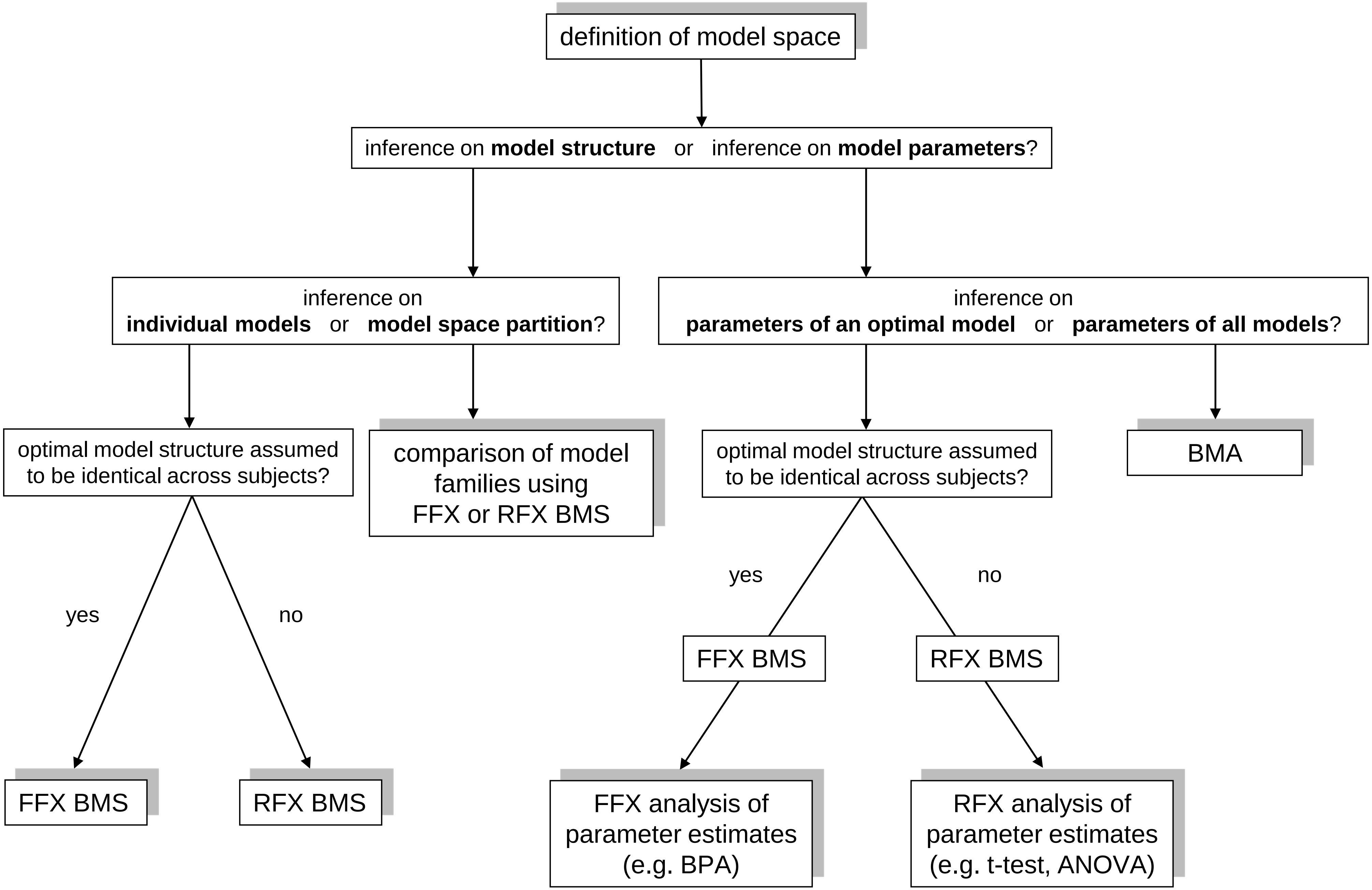
- **protected EP**: Bayesian model averaging over H_0 and H_1:

$$\begin{aligned} \tilde{\varphi}_k &= P(r_k \geq r_{k' \neq k}|y) \\ &= P(r_k \geq r_{k' \neq k}|y, H_1)P(H_1|y) + P(r_k \geq r_{k' \neq k}|y, H_0)P(H_0|y) \\ &= \varphi_k(1 - P_0) + \frac{1}{K}P_0 \end{aligned}$$

Rigoux et al. 2014, *NeuroImage*

Random effects BMS software

- **SPM**
  - function `spm_bms`
  - simple to use: only needs a log evidence matrix (subjects × models)
  - works with any log evidence approximation (F, AIC, BIC)
  - model inversion by VB or MCMC
  - http://www.fil.ion.ucl.ac.uk/spm/
- **VBA Toolbox**
  - VB only (not MCMC), but additional tests for group differences in model structure
  - http://mbb-team.github.io/VBA-toolbox/



definition of model space

inference on **model structure** or inference on **model parameters**?

inference on
individual models or **model space partition**?

inference on
parameters of an optimal model or **parameters of all models**?

optimal model structure assumed to be identical across subjects?

comparison of model families using FFX or RFX BMS

optimal model structure assumed to be identical across subjects?

BMA

yes

no

yes

no

FFX BMS

RFX BMS

FFX BMS

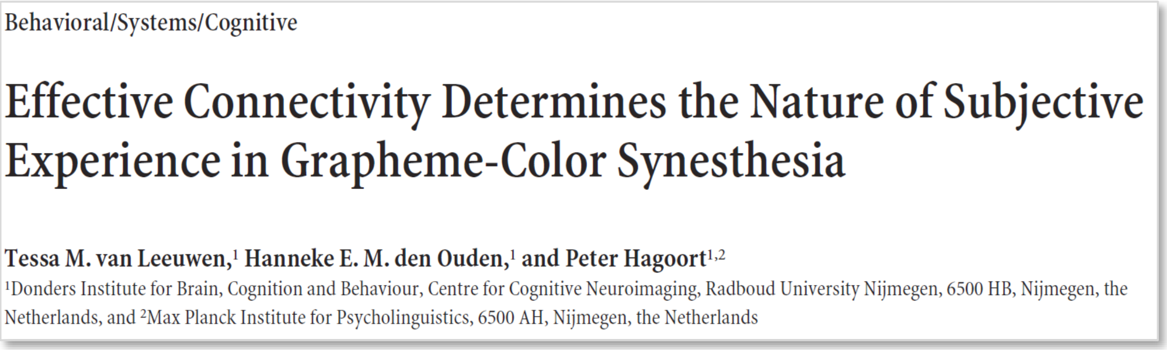
RFX BMS

FFX analysis of parameter estimates (e.g. BPA)

RFX analysis of parameter estimates (e.g. t-test, ANOVA)

Stephan et al. 2010, *NeuroImage*

Some examples of empirical BMS/BMA applications

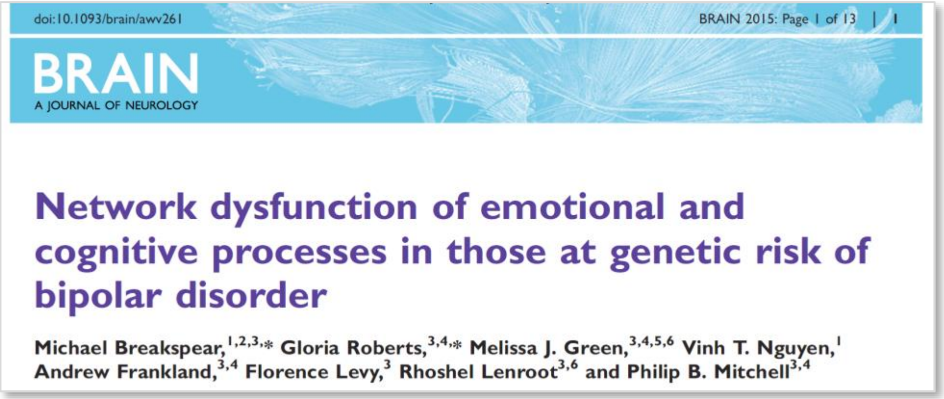


Behavioral/Systems/Cognitive

Effective Connectivity Determines the Nature of Subjective Experience in Grapheme-Color Synesthesia

Tessa M. van Leeuwen,[1] Hanneke E. M. den Ouden,[1] and Peter Hagoort[1,2]
[1]Donders Institute for Brain, Cognition and Behaviour, Centre for Cognitive Neuroimaging, Radboud University Nijmegen, 6500 HB, Nijmegen, the Netherlands, and [2]Max Planck Institute for Psycholinguistics, 6500 AH, Nijmegen, the Netherlands

van Leeuwen et al. 2011,
J. Neurosci.



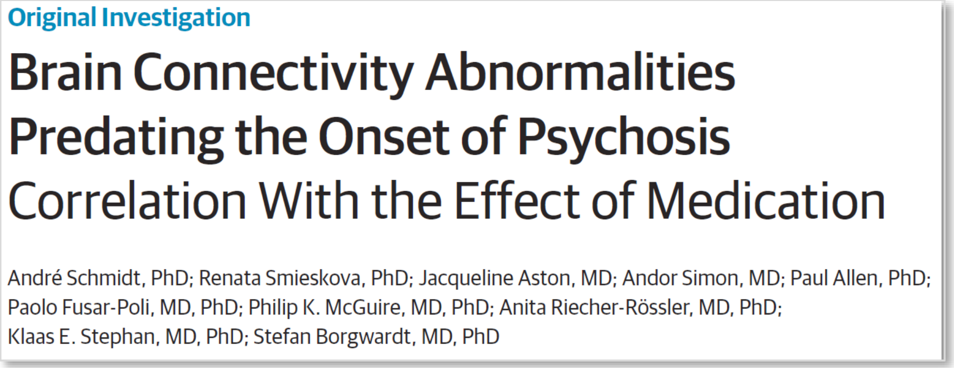
doi:10.1093/brain/awv261 BRAIN 2015: Page 1 of 13 | 1

BRAIN
A JOURNAL OF NEUROLOGY

Network dysfunction of emotional and cognitive processes in those at genetic risk of bipolar disorder

Michael Breakspear,[1,2,3,*] Gloria Roberts,[3,4,*] Melissa J. Green,[3,4,5,6] Vinh T. Nguyen,[1]
Andrew Frankland,[3,4] Florence Levy,[3] Rhoshel Lenroot[3,6] and Philip B. Mitchell[3,4]

Breakspear et al. 2015,
Brain



Original Investigation

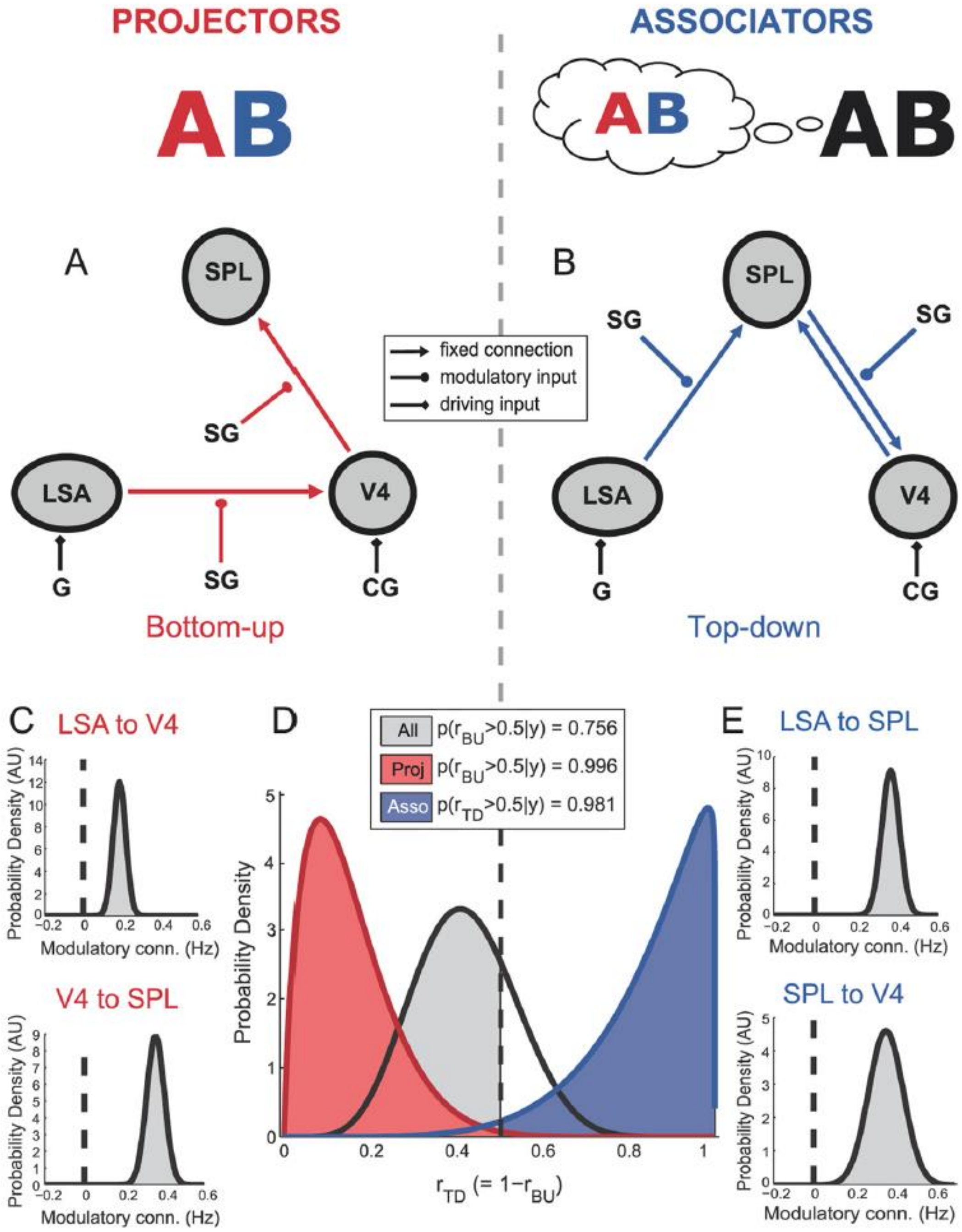
Brain Connectivity Abnormalities Predating the Onset of Psychosis
Correlation With the Effect of Medication

André Schmidt, PhD; Renata Smieskova, PhD; Jacqueline Aston, MD; Andor Simon, MD; Paul Allen, PhD;
Paolo Fusar-Poli, MD, PhD; Philip K. McGuire, MD, PhD; Anita Riecher-Rössler, MD, PhD;
Klaas E. Stephan, MD, PhD; Stefan Borgwardt, MD, PhD

Schmidt et al. 2013,
JAMA Psychiatry

Application: 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)



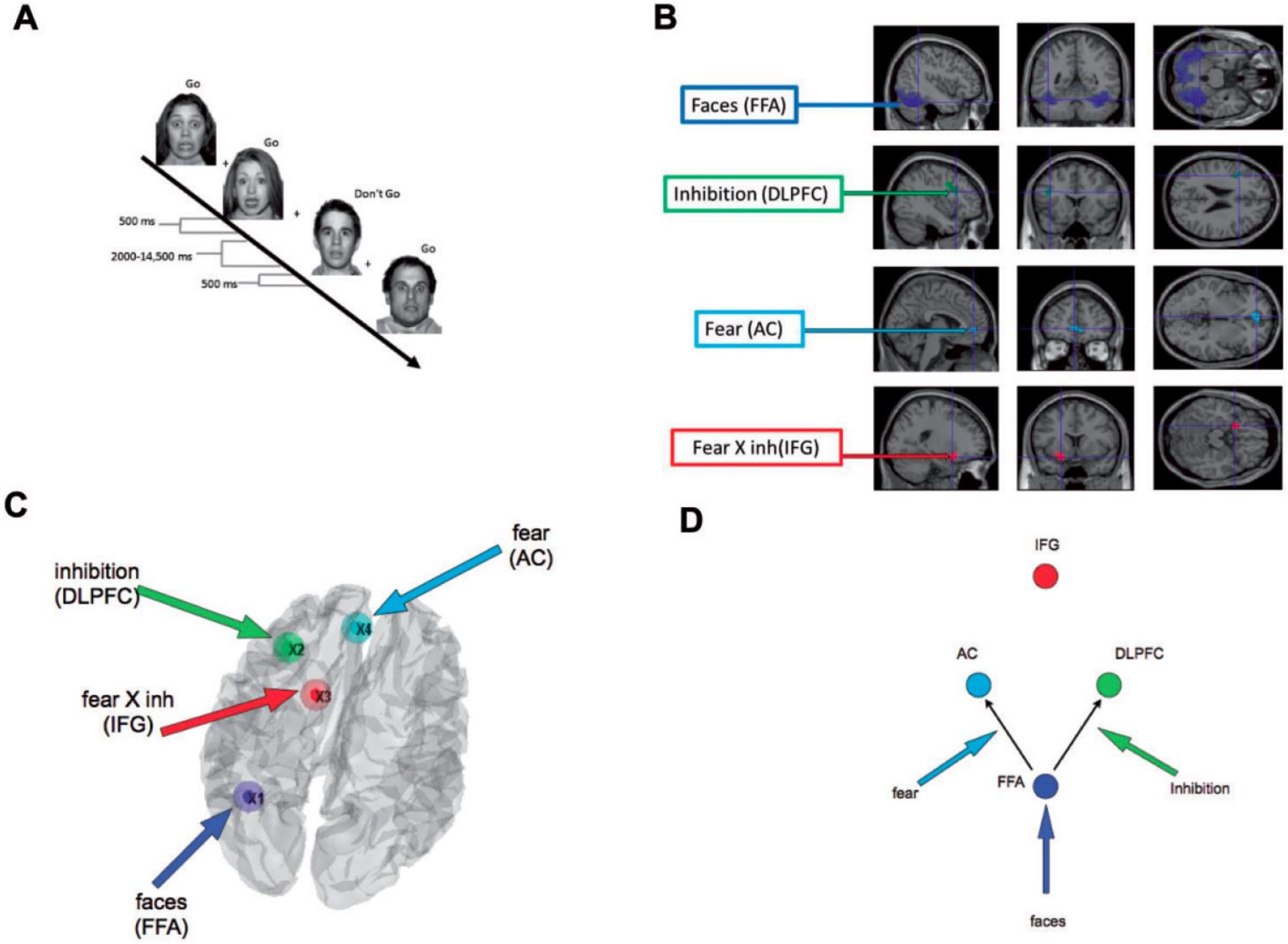
van Leeuwen et al. 2011, *J. Neurosci.*

Go/No-Go task to emotional faces
(bipolar patients, at-risk individuals, controls)

- interaction of motor inhibition and fear perception

- hypoactivation of left IFG in the at-risk group during fearful distractor trials

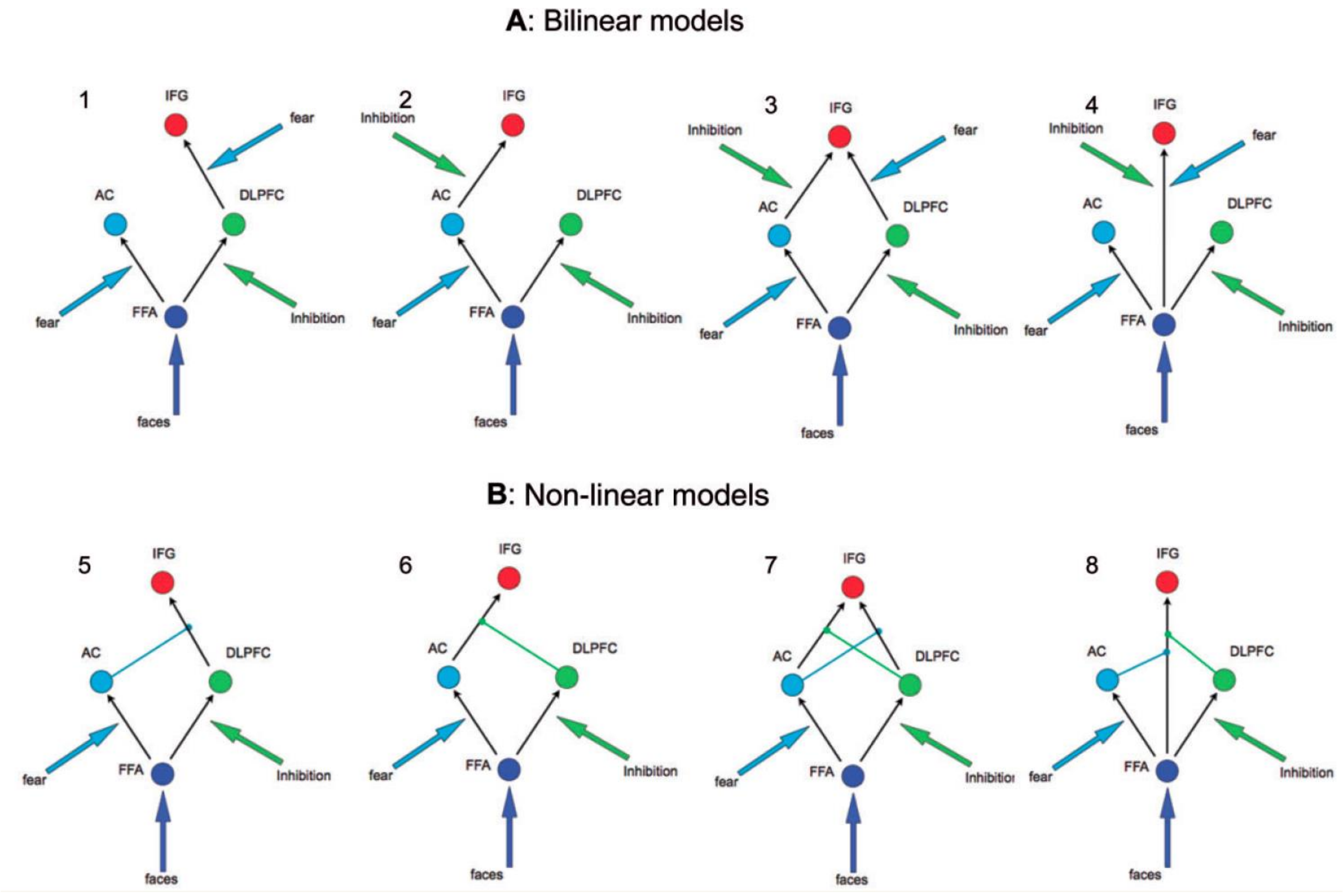
- What is the most likely circuit mechanism explaining the fear x inhibition interaction in IFG?



Breakspear et al. 2015, *Brain*

Model space

- models of serial (1-3), parallel (4) and hierarchical (5-8) processes


A: Bilinear models

B: Non-linear models

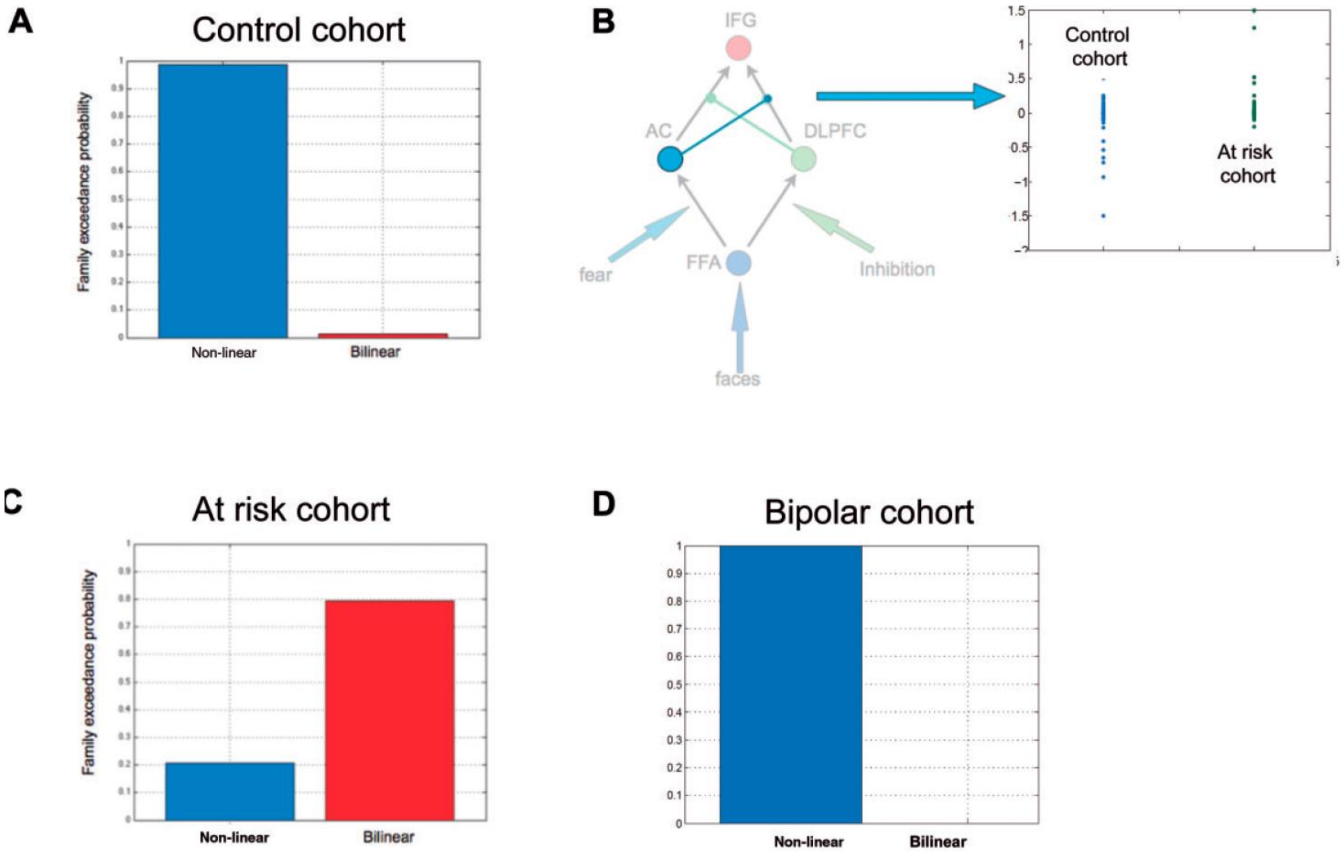
Breakspear et al. 2015, *Brain*

Family-level BMS

- family-level comparison: nonlinear models more likely than bilinear ones in both healthy controls and bipolar patients

- at-risk group: bilinear models more likely

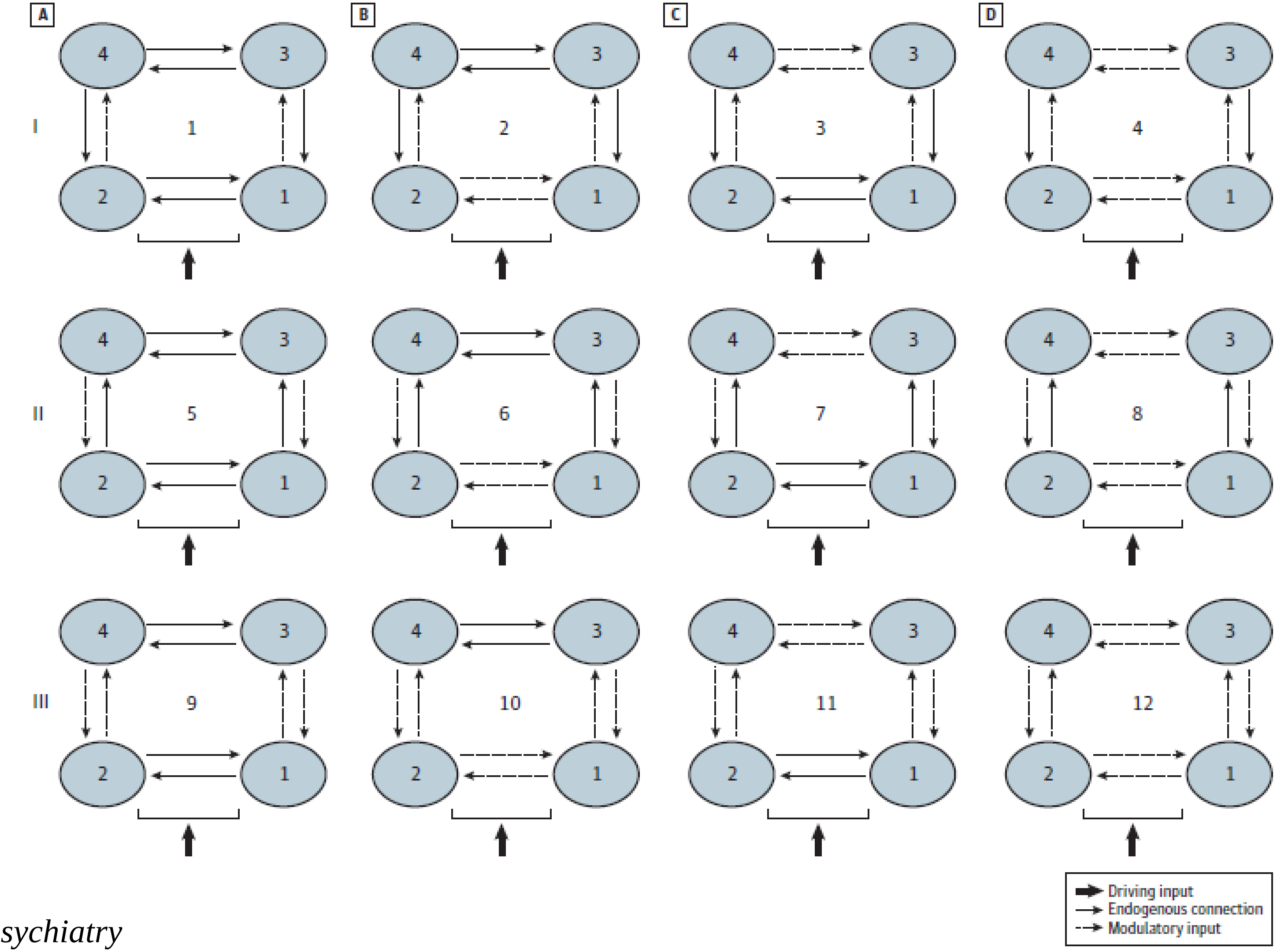
- significant group difference in ACC modulation of DLPFC→IFG interaction



Breakspear et al. 2015, *Brain*

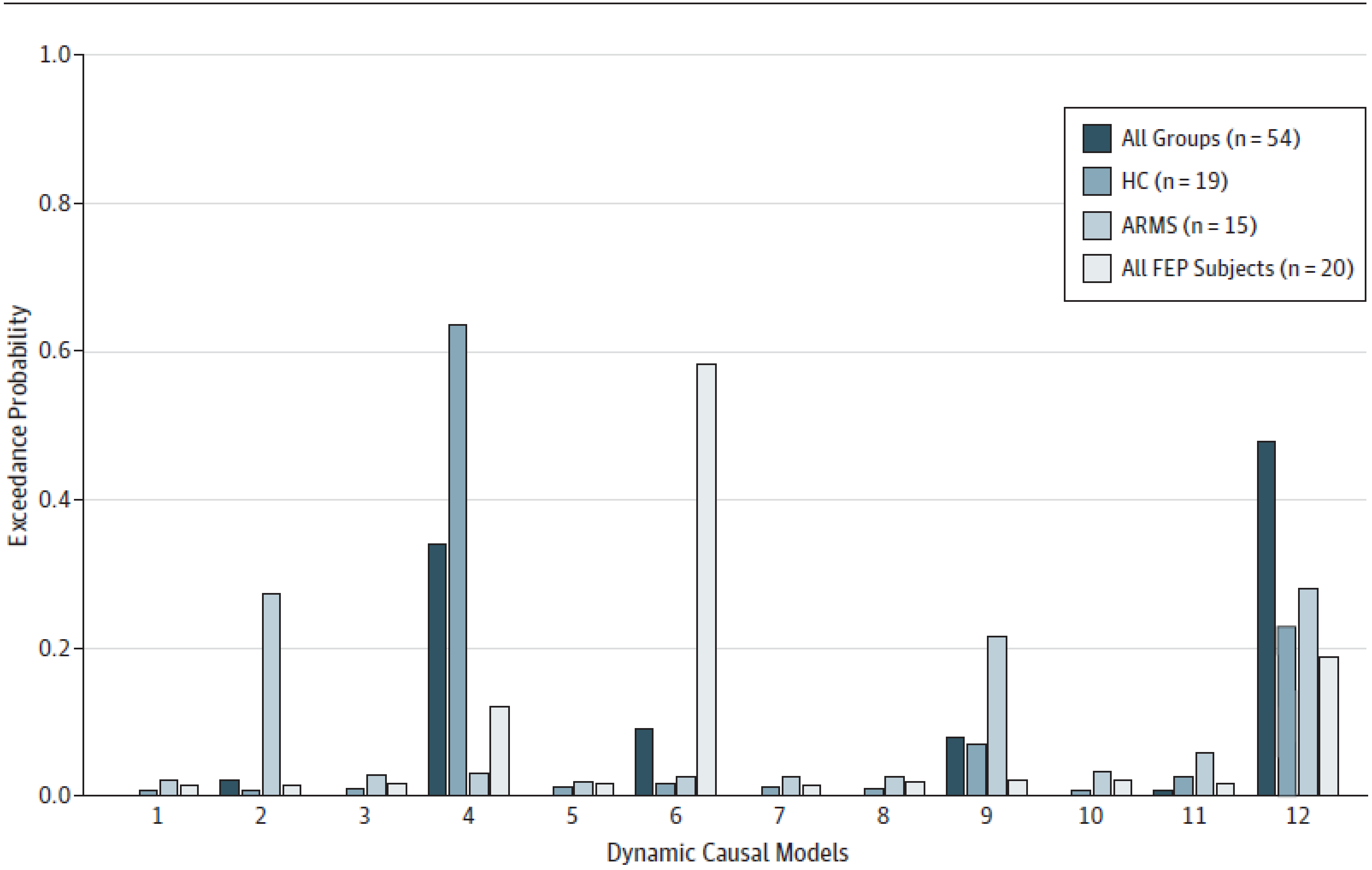
Prefrontal-parietal connectivity during working memory in schizophrenia

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



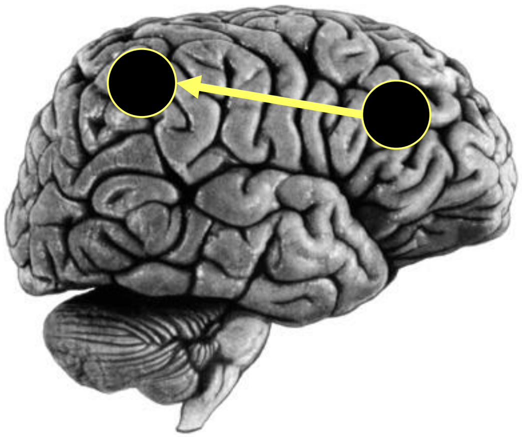
Schmidt et al. 2013, *JAMA Psychiatry*

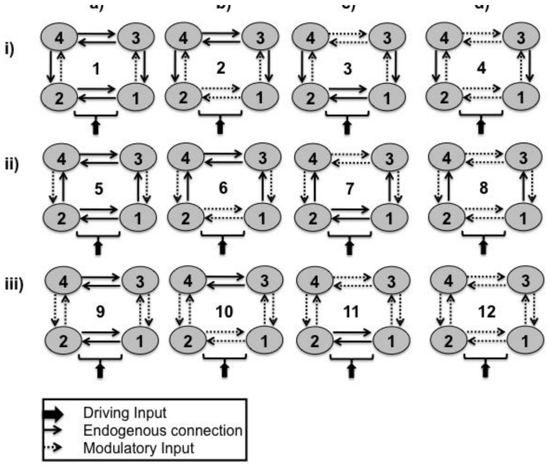
BMS results for all groups

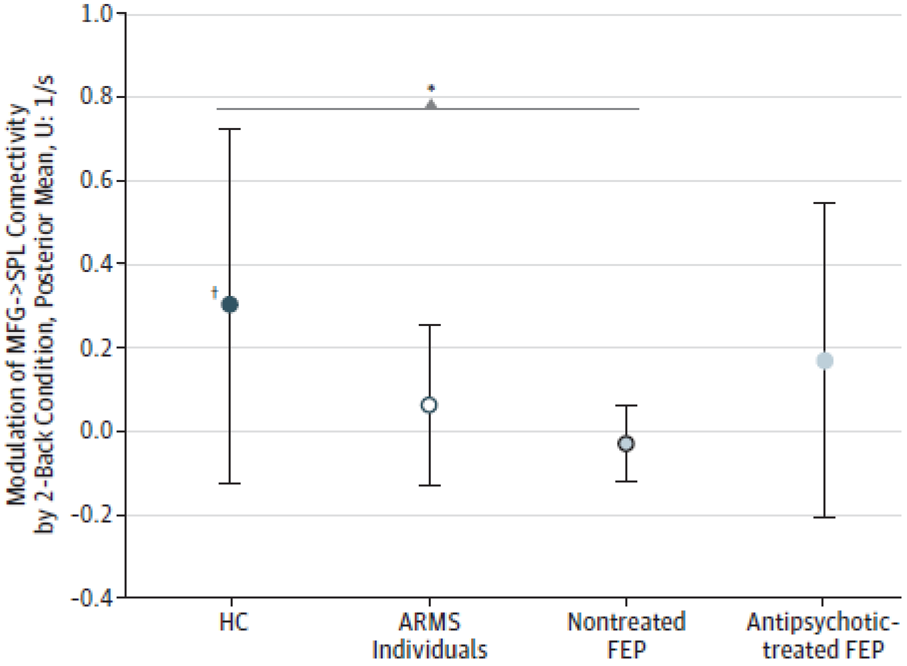


Schmidt et al. 2013, *JAMA Psychiatry*

BMA results: PFC → PPC connectivity







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

Schmidt et al. 2013, *JAMA Psychiatry*

Further reading on BMS

- 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.

Thank you