Uncovering interactions with Random Forests

Jake Michaelson
Marit Ackermann
Andreas Beyer

TECHNISCHE UNIVERSITÄT DRESDEN

biotec
Biotechnology Center TU Dresden

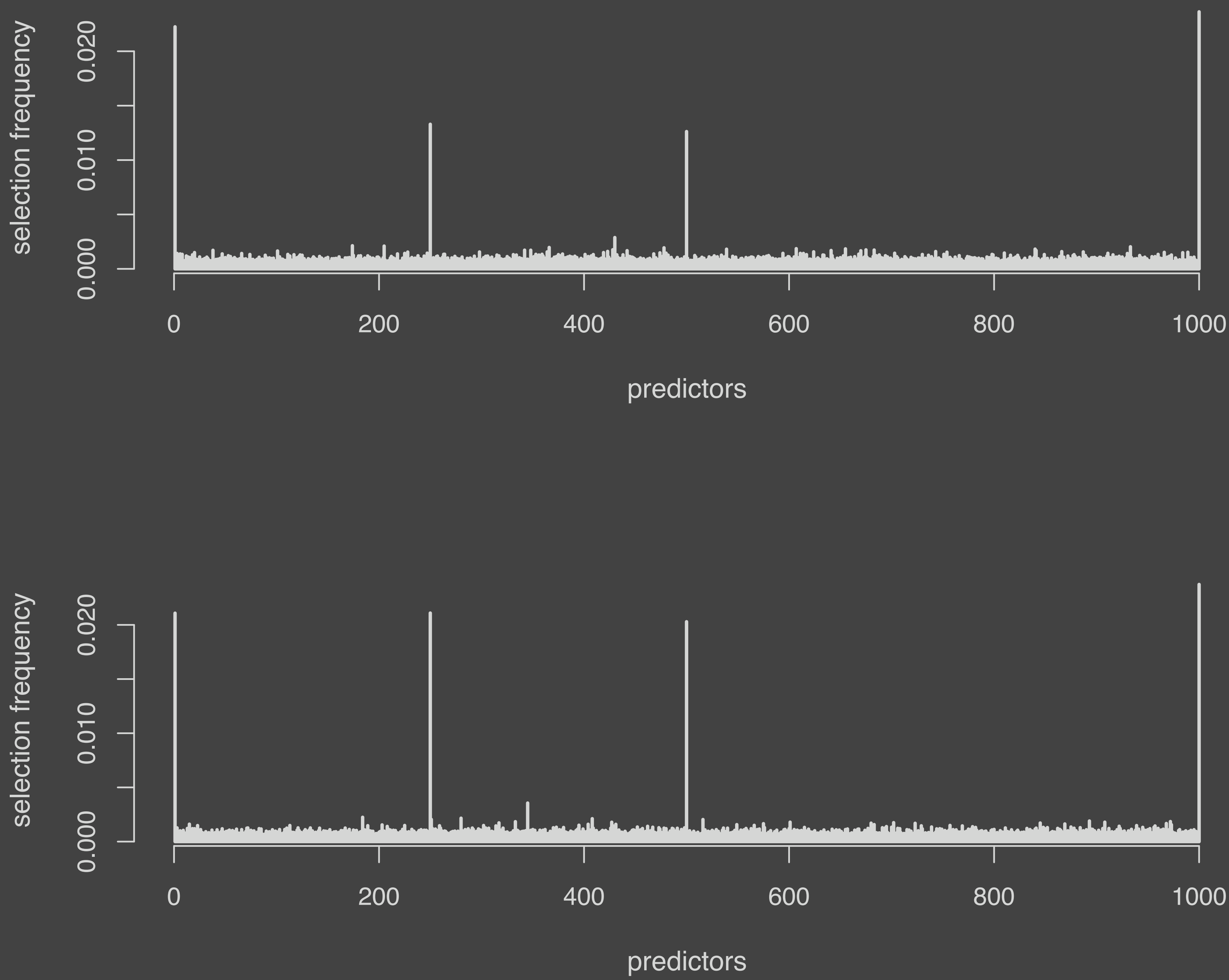
Random Forests

- >> ensembles of decision trees
  - >> diverse trees trying to solve the same problem
- >> used frequently for:
  - >> prediction (knowledge of model less important)
  - >> feature selection (prediction less important)

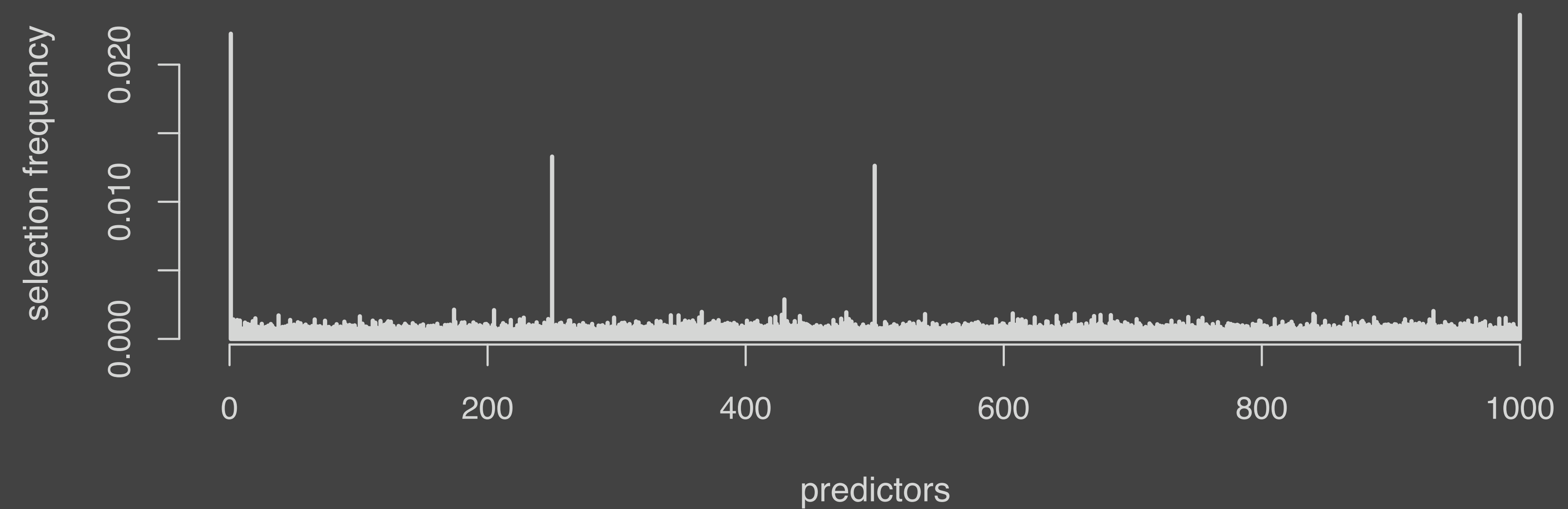
RF interactions: prior art

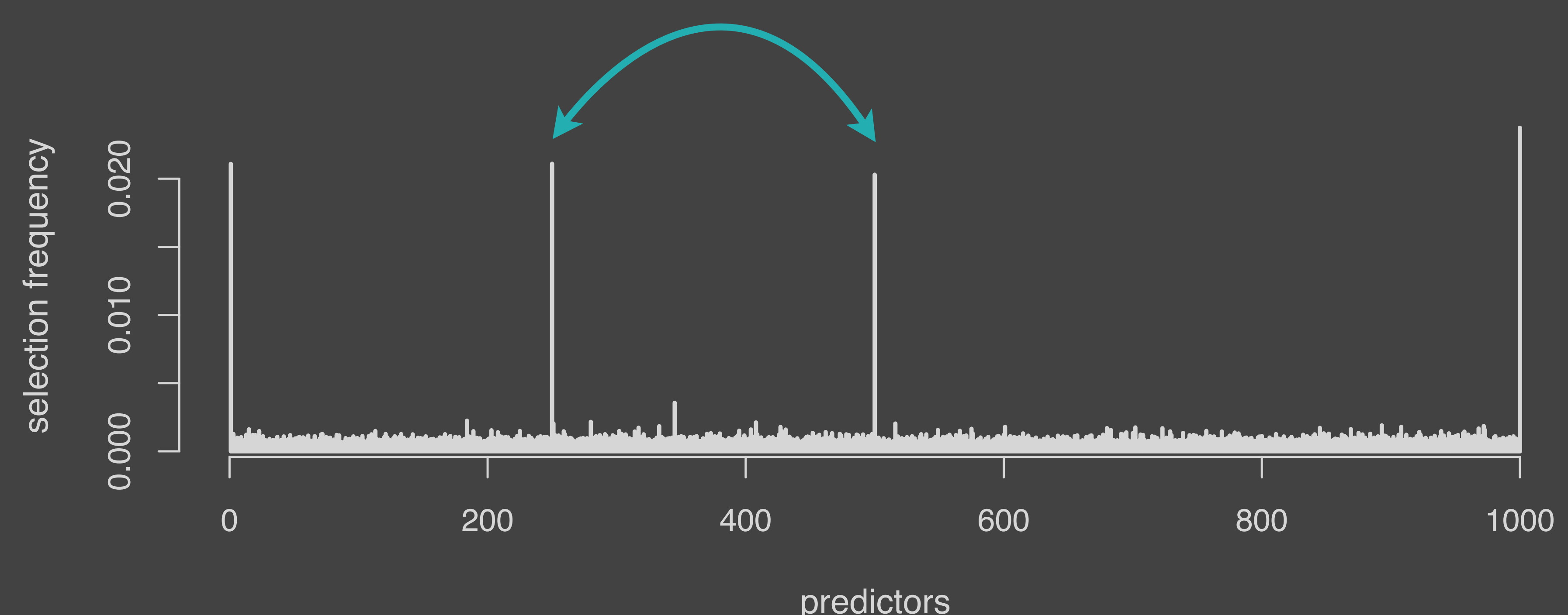
- online official RF manual
- Lunetta, et al. (2004)
- Bureau, et al. (2005)
  - pairwise permutation importance
- Mao and Mao (2008)
- Jiang, et al. (2009)
  - selection with RF Gini importance, conventional (LM-based) interaction test (up to 3-way)

a typical problem

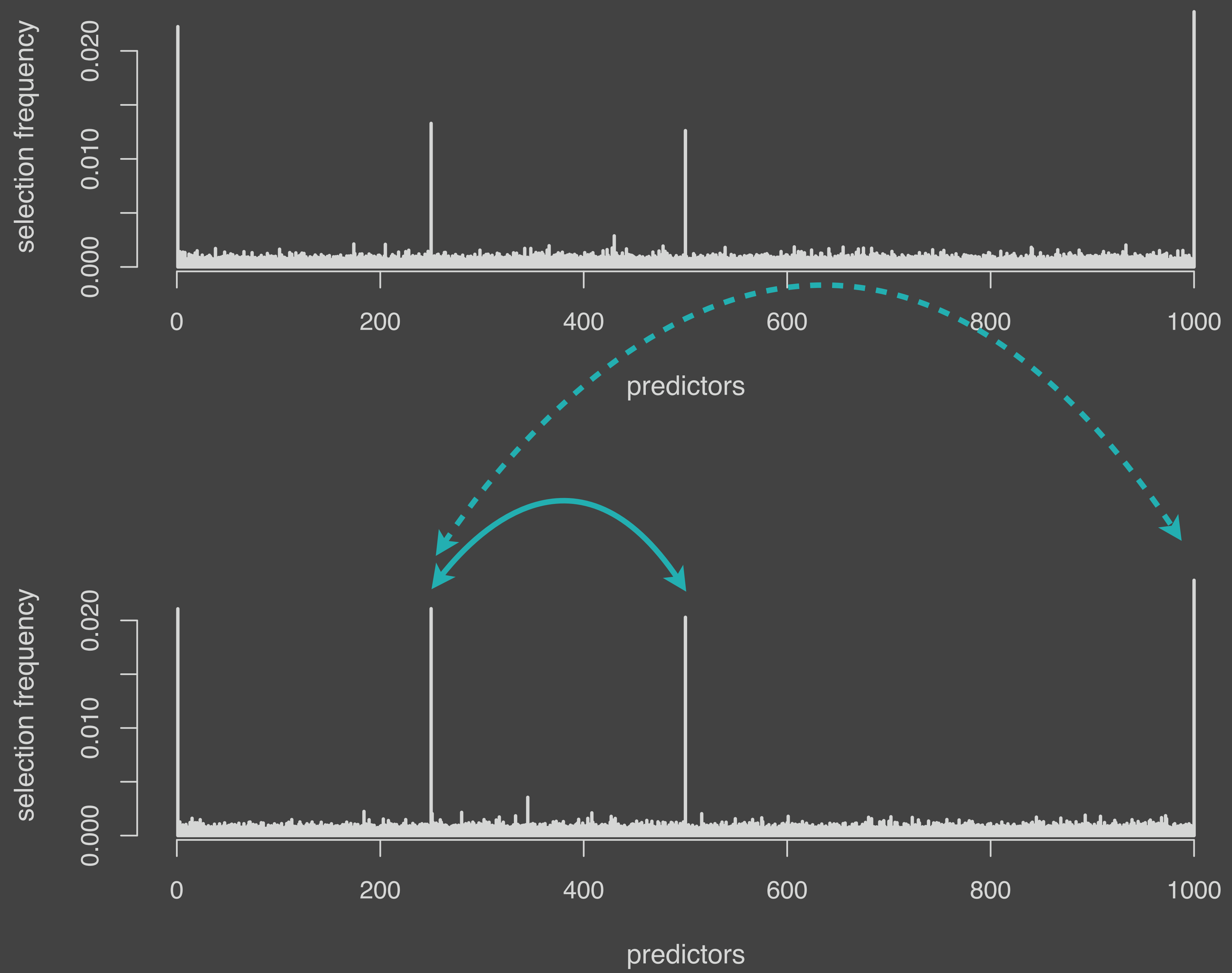


a typical problem

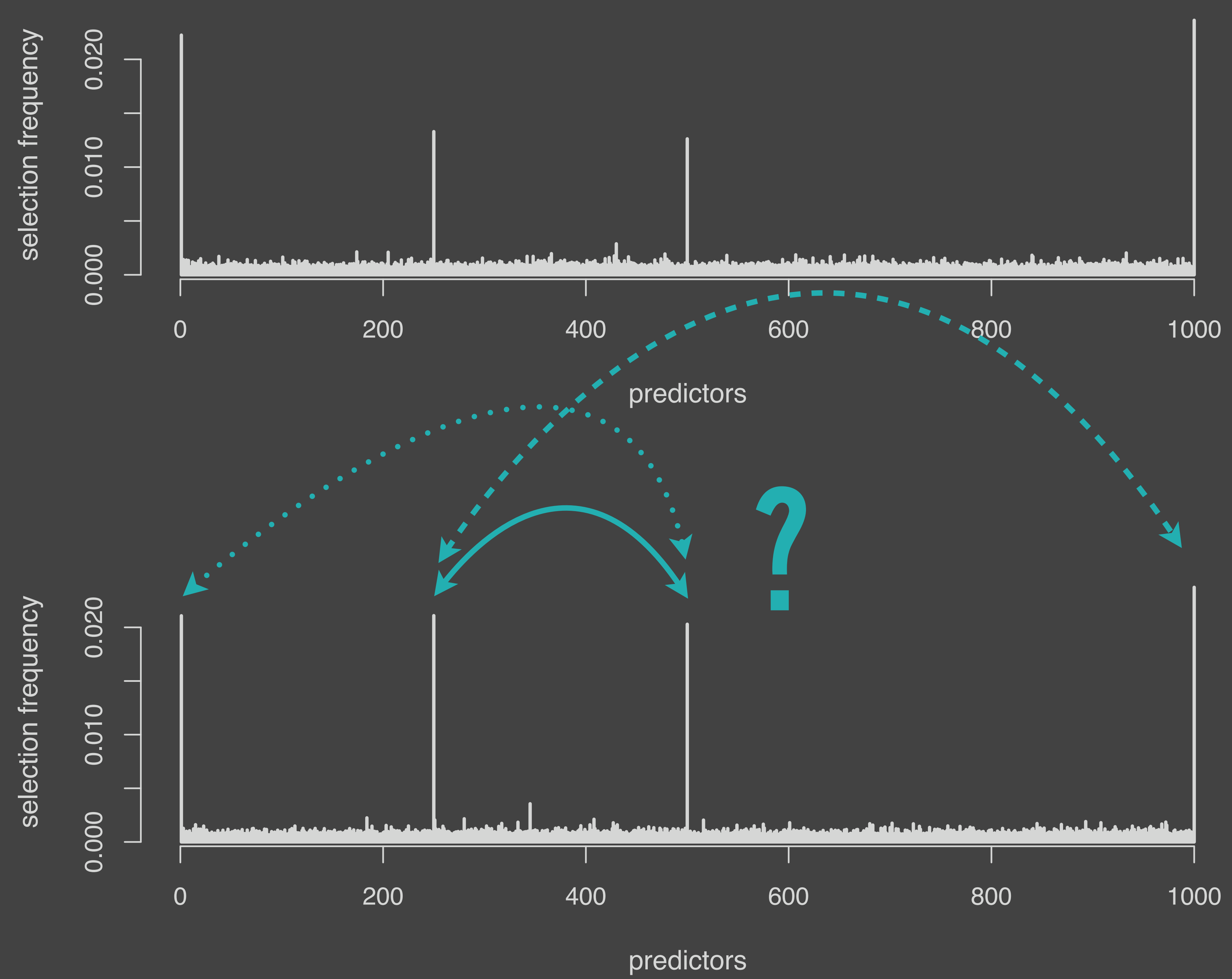




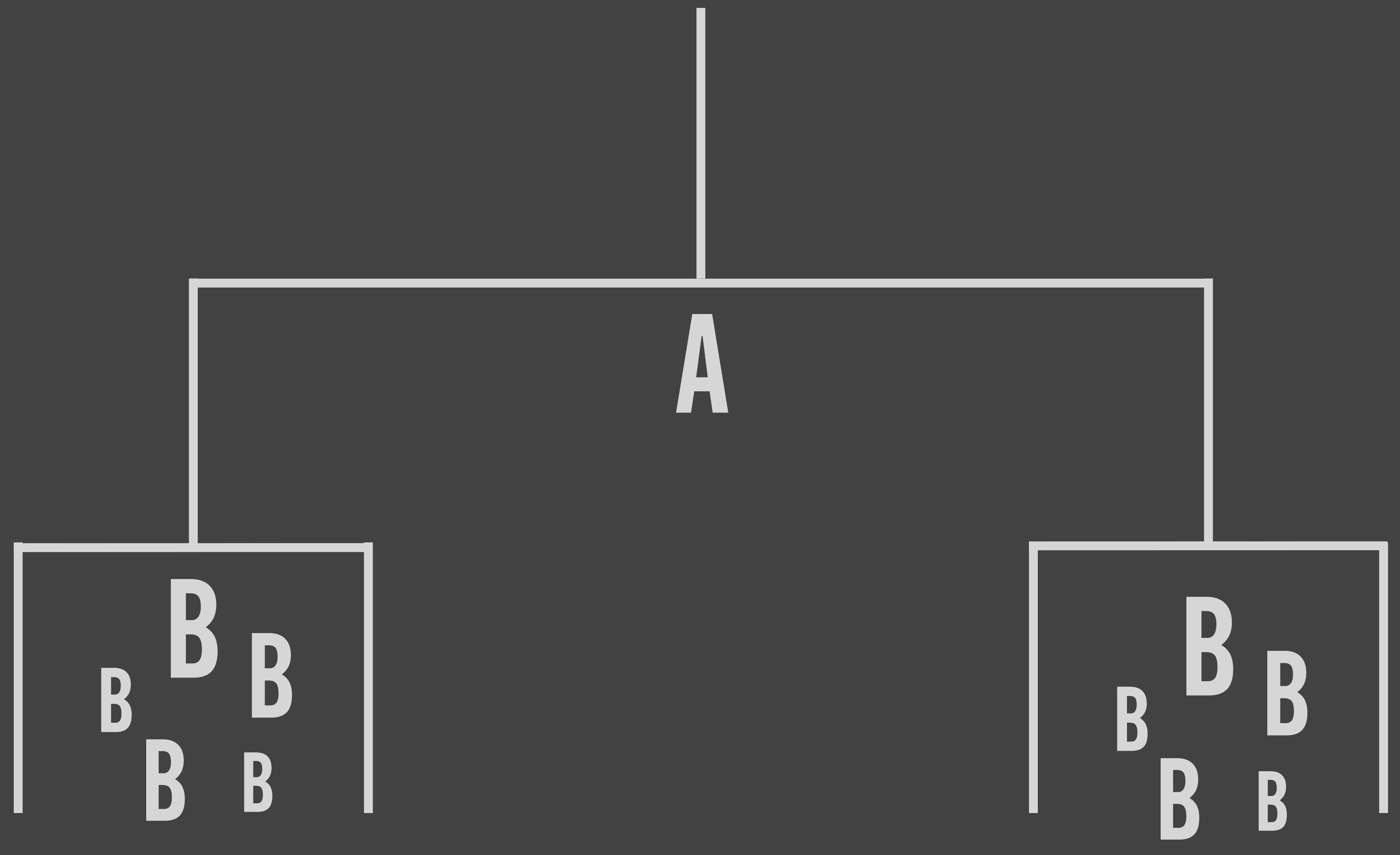
a typical problem



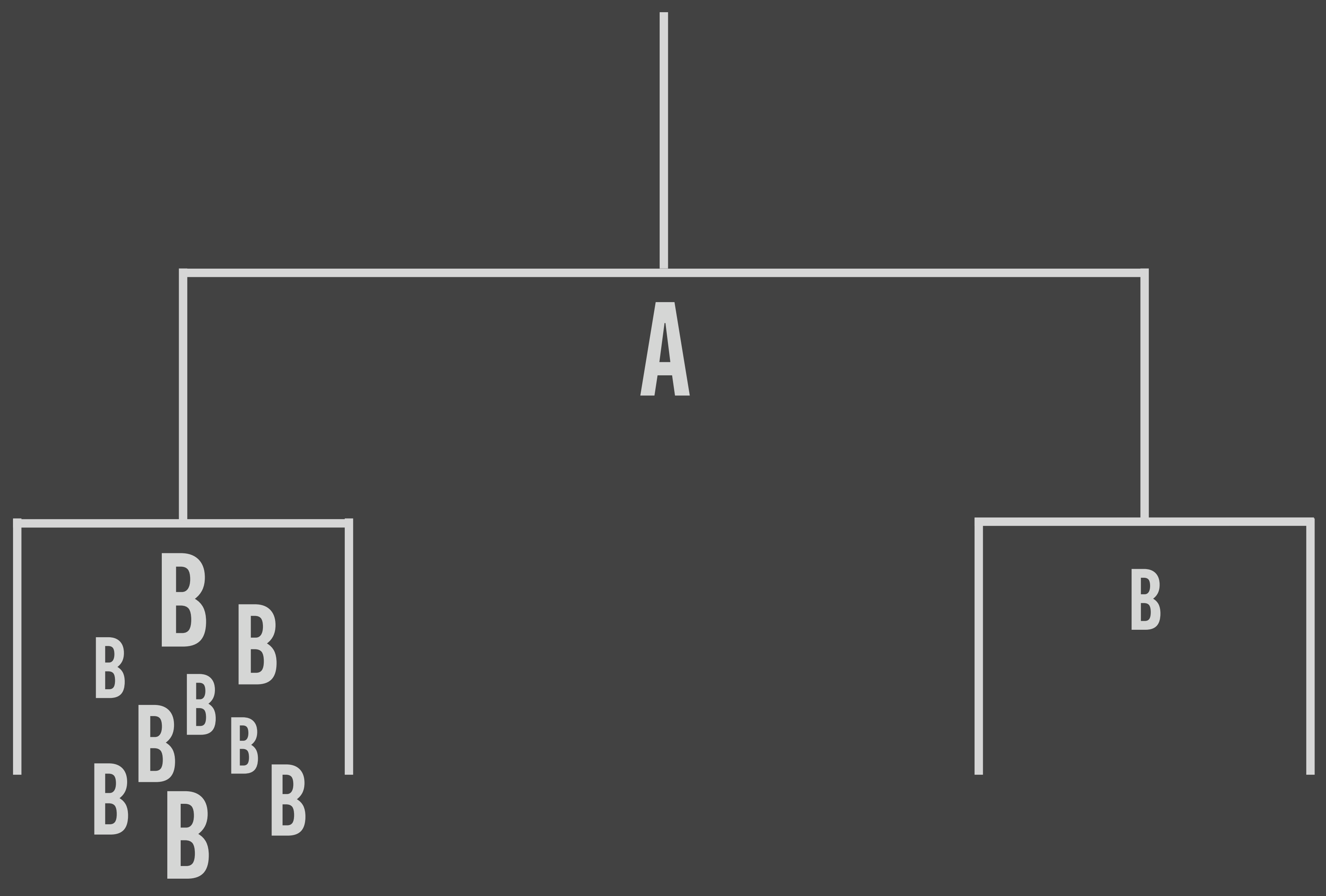
a typical problem



split symmetry



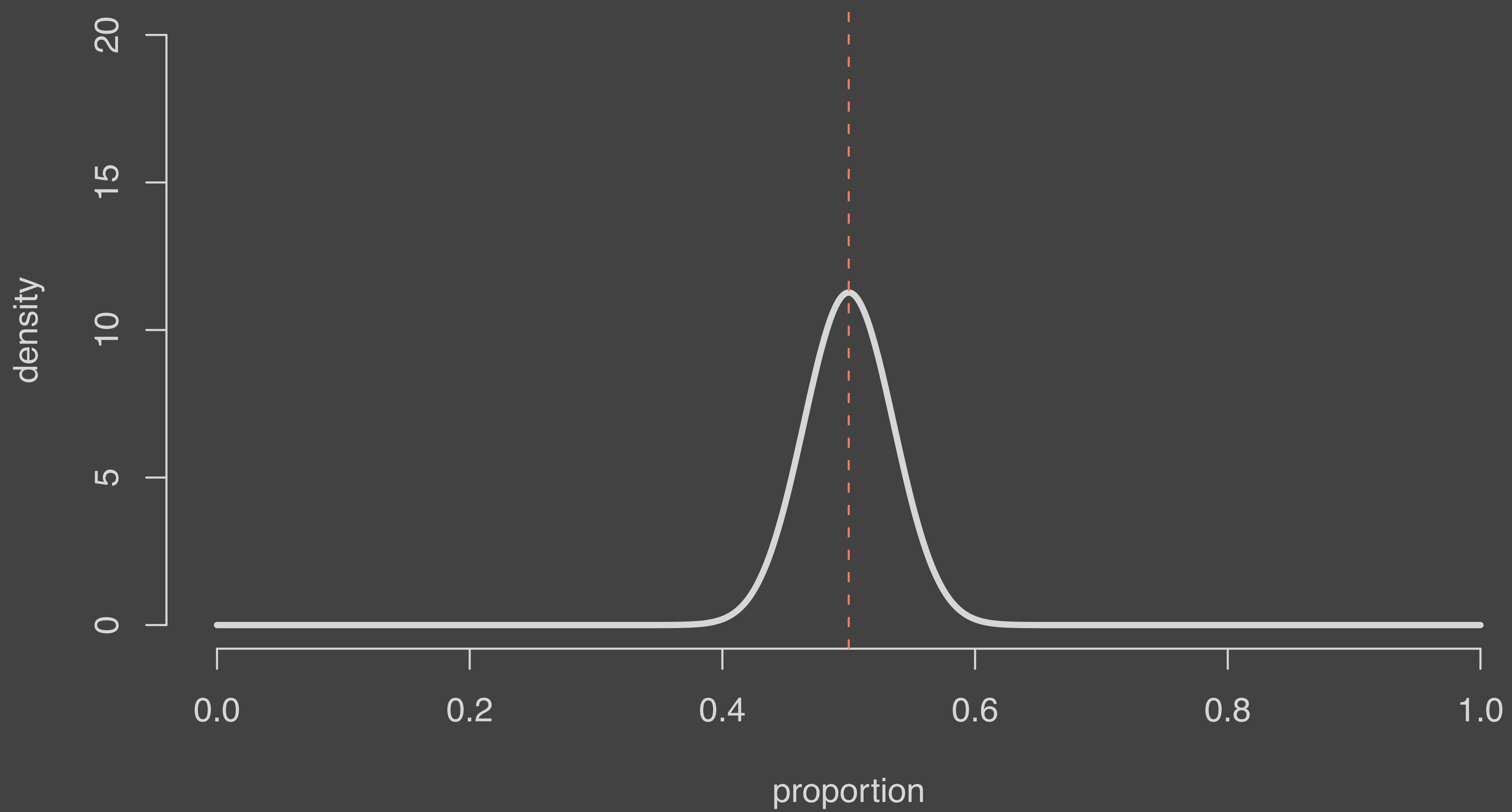
split asymmetry



testing split symmetry

- independence of predictors A and B:
  - expect B as left daughter 50% of the time
  - expect B as right daughter 50% of the time
  - the prior (a beta density) is centered around 0.5

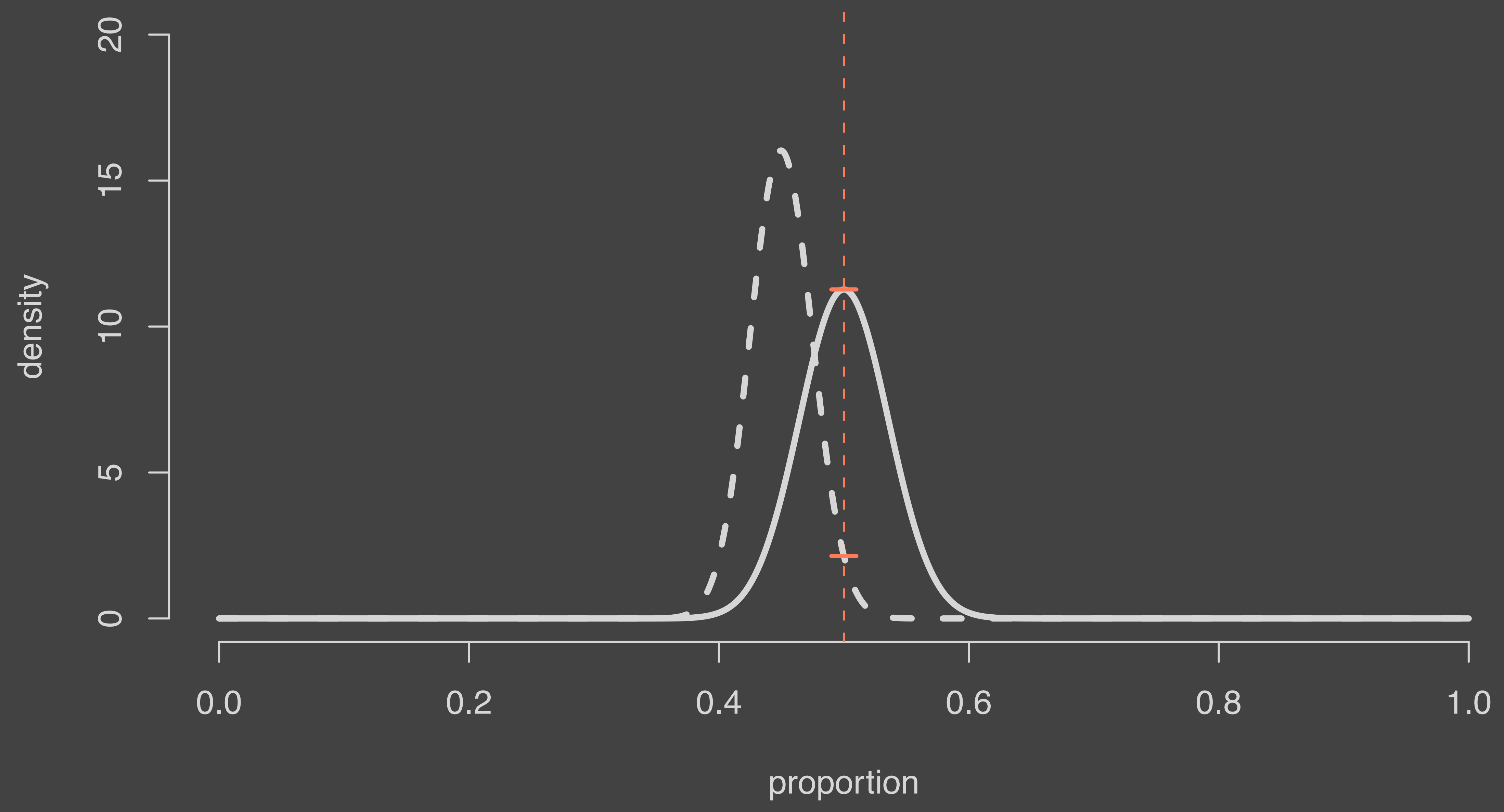
testing split symmetry



testing split symmetry

- we update the prior density parameters with the observed left/right daughter counts:
  - $a_{posterior} = a_{prior} + AB_{left}$
  - $b_{posterior} = b_{prior} + AB_{right}$
- ... and take the posterior/prior density ratio at 0.5
- this is the Bayes factor

testing split symmetry



building a graph

$$p_{post} = p_h \frac{BF}{(p_h \cdot BF + 1 - p_h)}$$

- using the Bayes factor from each pair of predictors, we calculate the posterior probability of symmetry
  - i.e. that the true proportion is 0.5
  - we use a high prior probability of the hypothesis (e.g. $p_h = 0.999999$)

building a graph

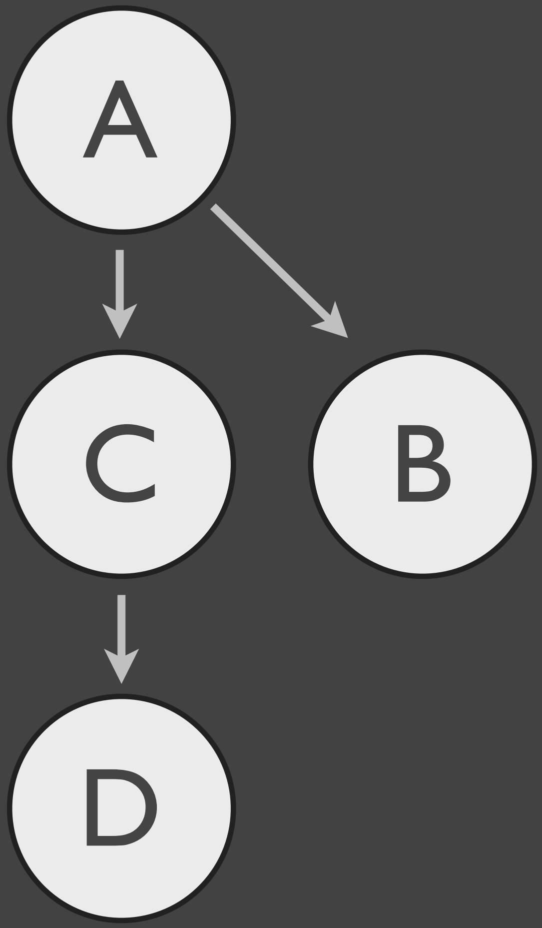
posterior probabilities

	A	B	C	D
A	1	0.001	0.001	0.3
B	0.8	1	0.99	0.2
C	0.99	0.3	1	0.003
D	1	0.89	0.99	1

adjacency matrix

	A	B	C	D
A	0	1	1	0
B	0	0	0	0
C	0	0	0	1
D	0	0	0	0

graph

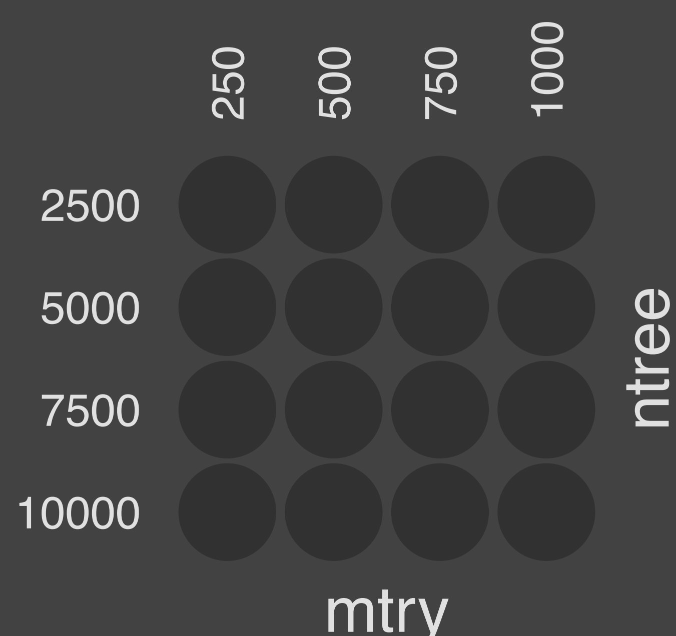


simulations

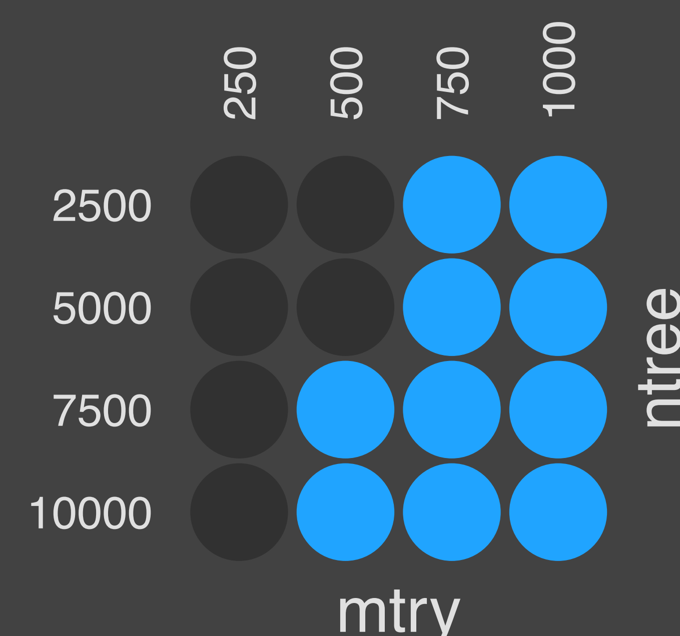
- 1000 binary predictor variables, 200 observations
  - 3 - 4 predictors participate in true model
- tested ability of the method to recover the true topology of the simulated model
- recorded TP, FP while varying mtry and ntree

test models

3 independent effects (i.e. no edges)

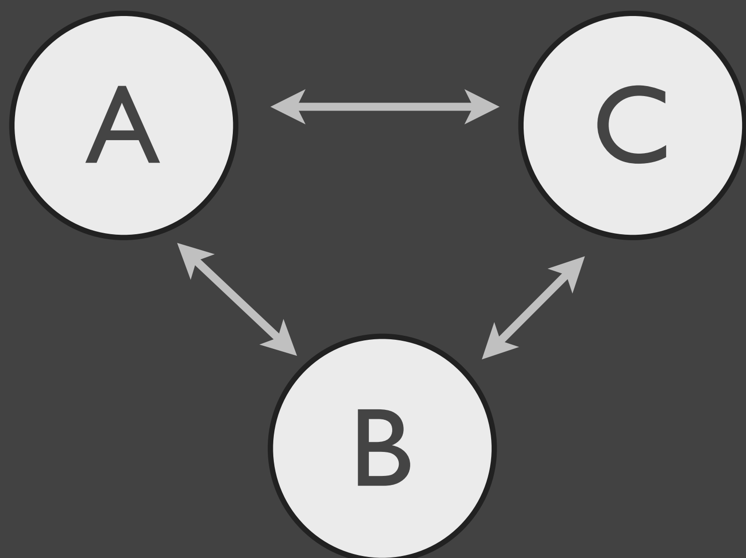


TP

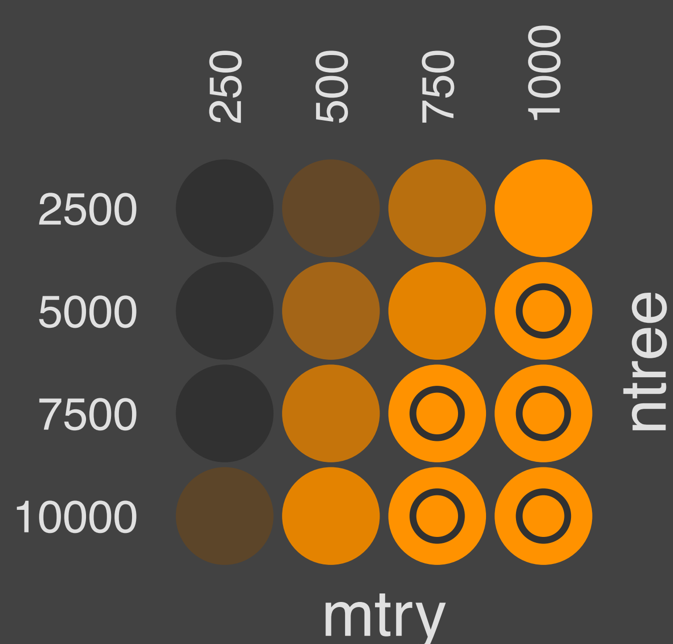


FP

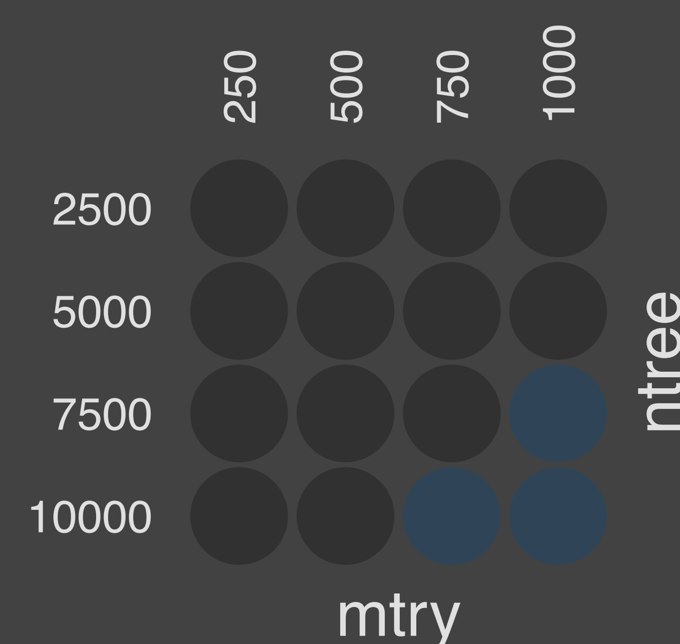
test models



3-way unordered interaction

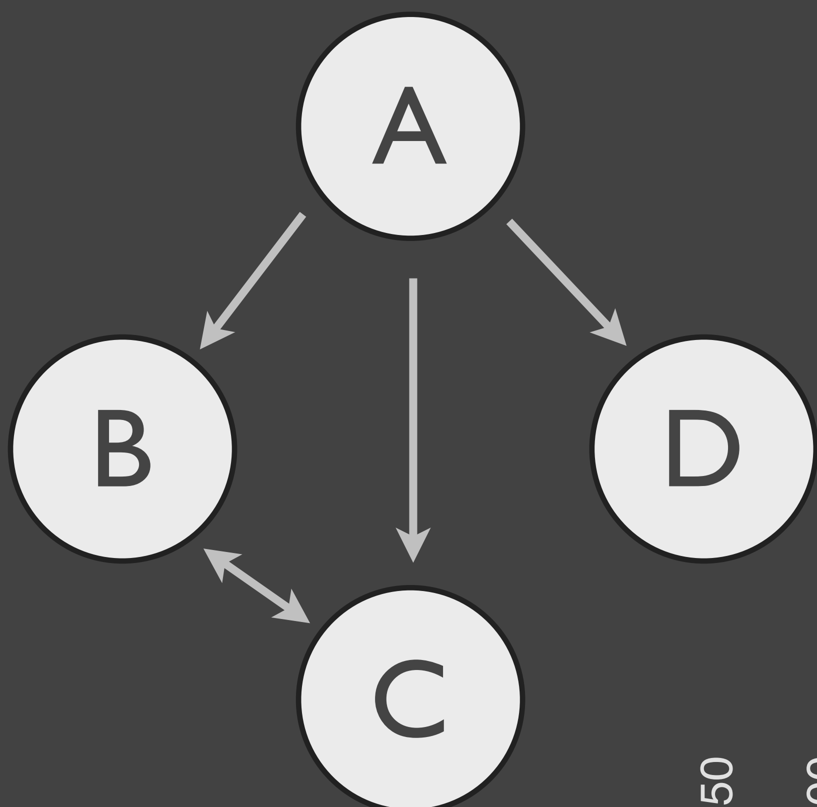


TP

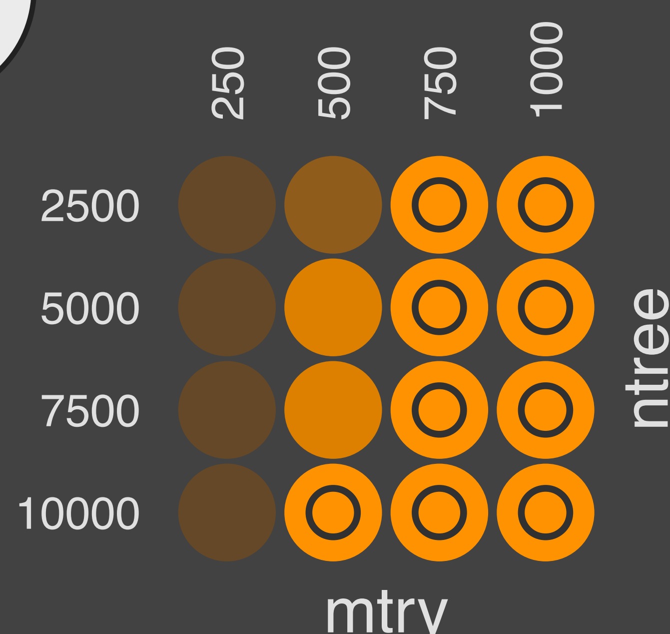


FP

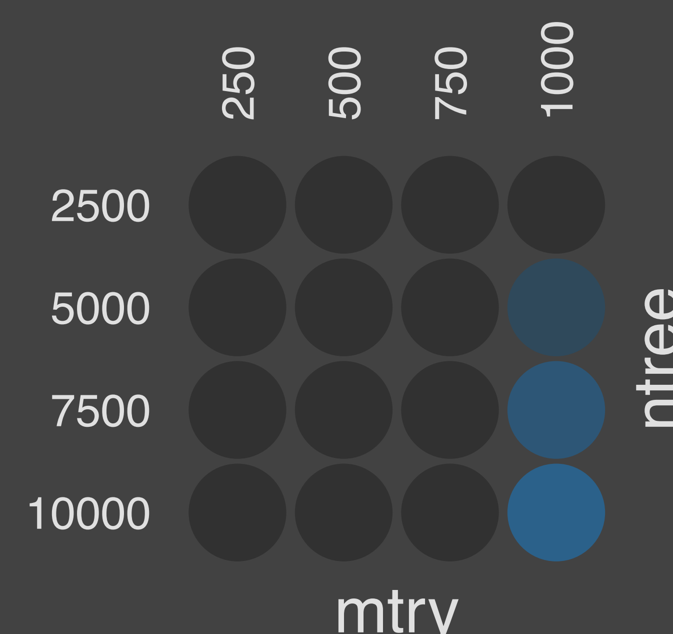
test models



one main effect,
one ordered 3-way interaction,
one ordered 2-way interaction

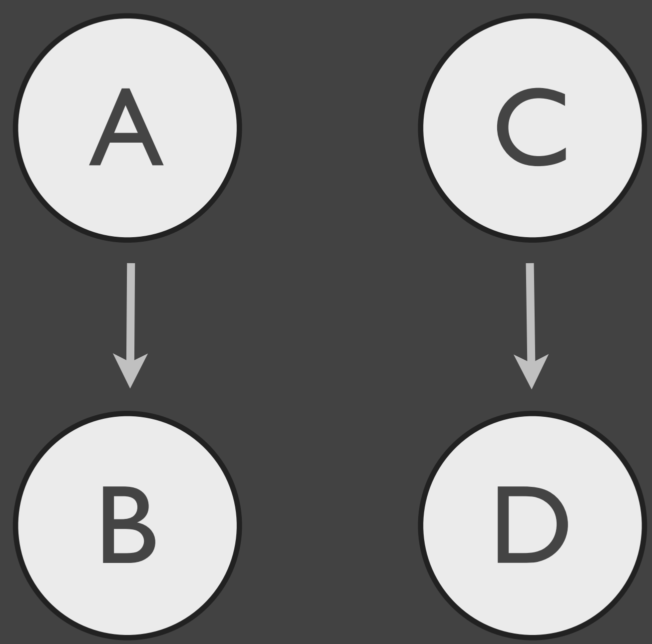


TP

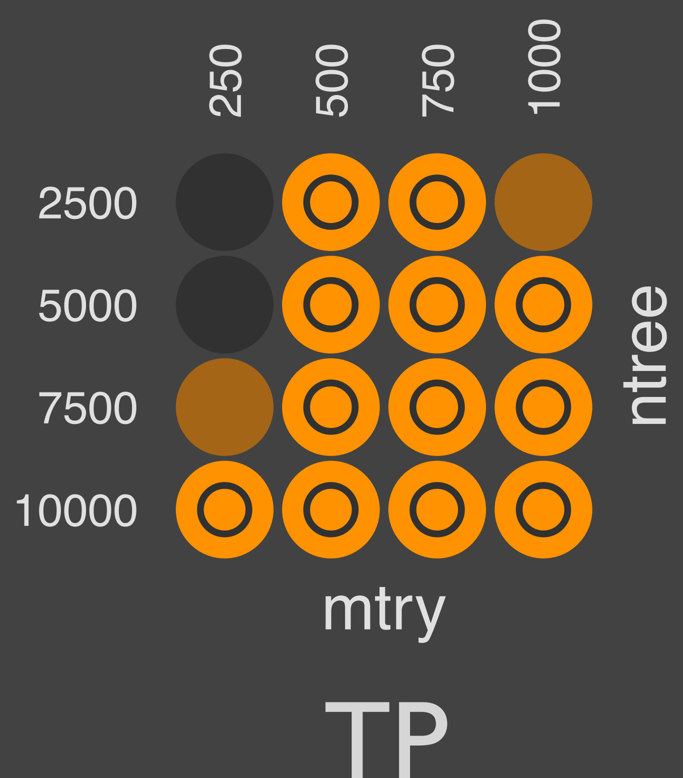


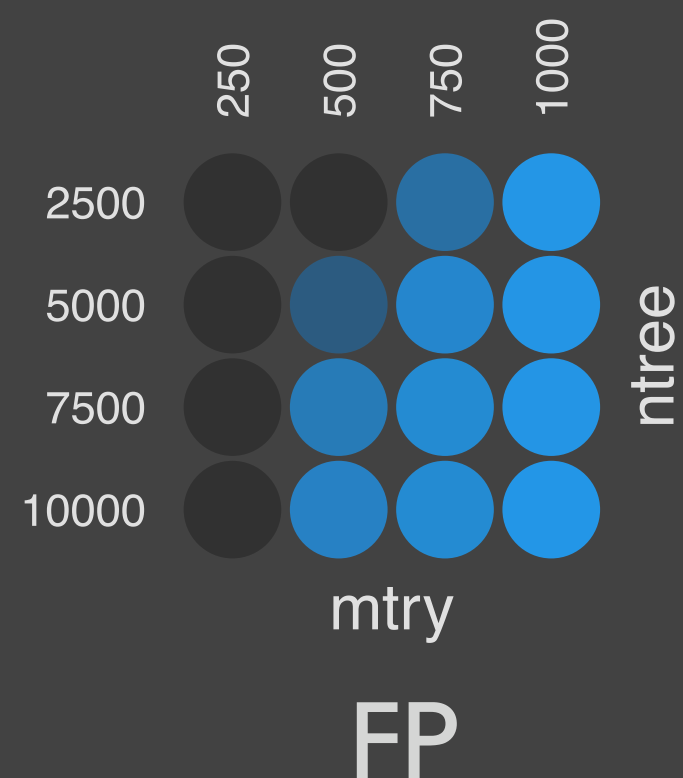
FP

test models



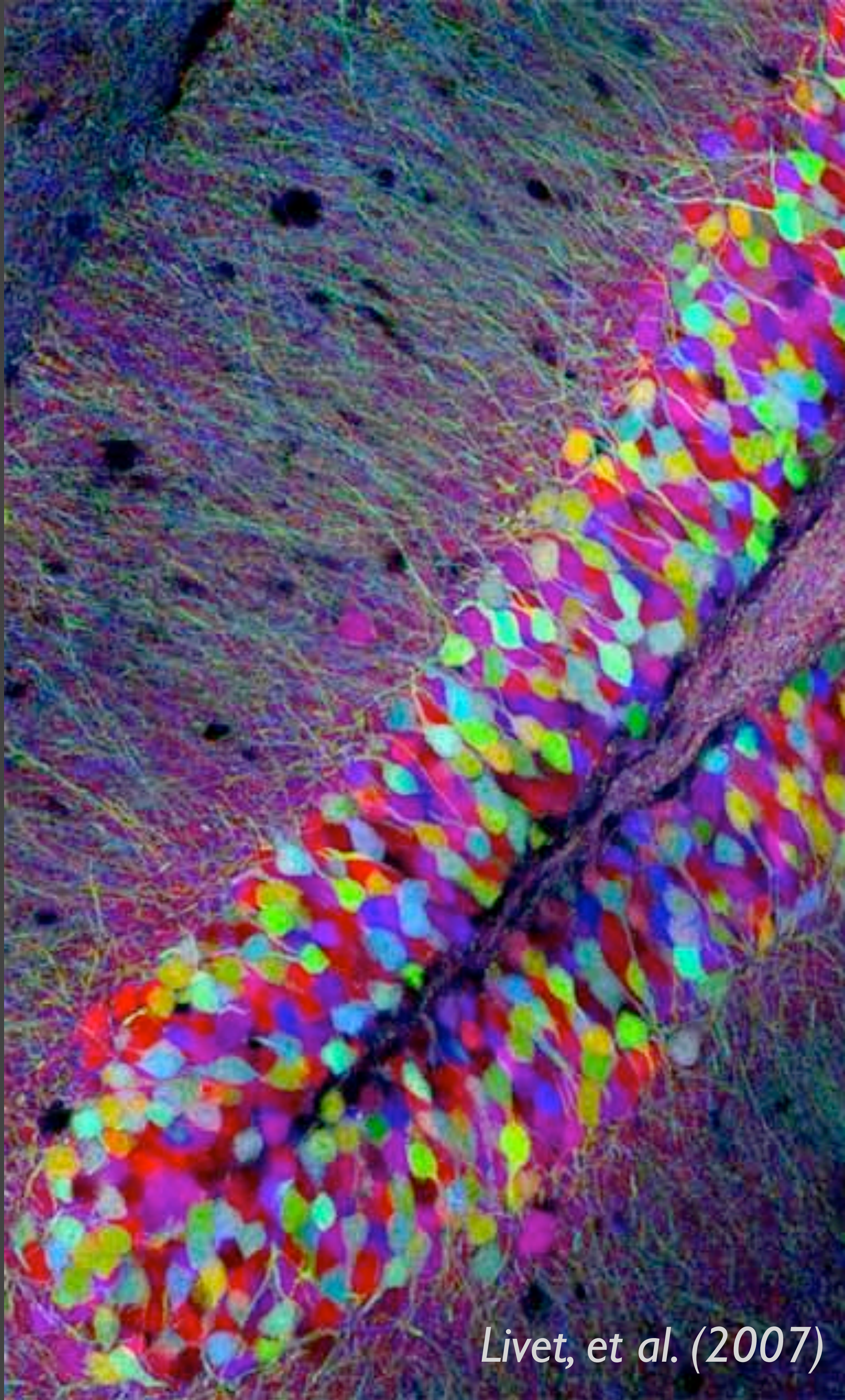
two independent, ordered two-way interactions





real world

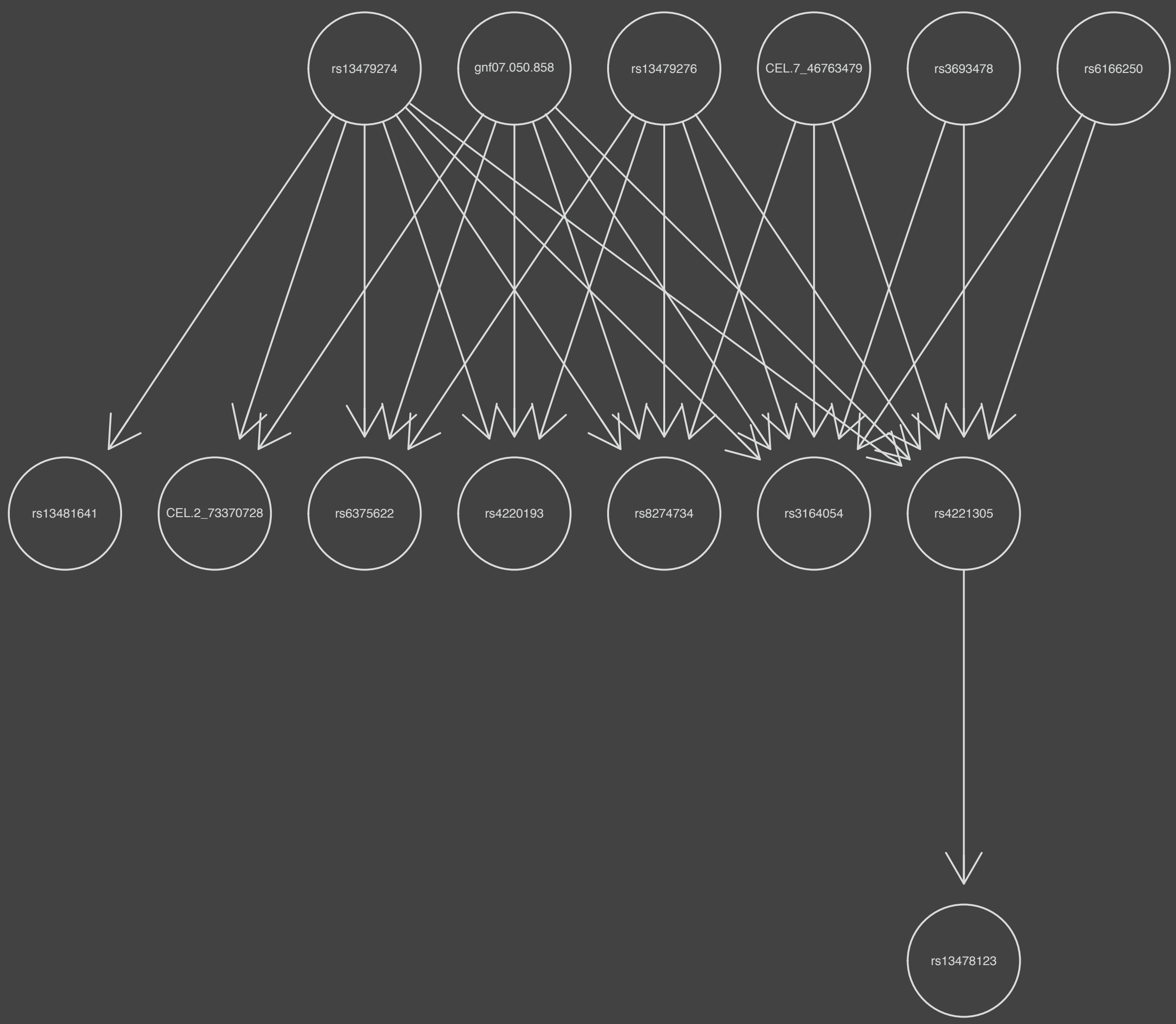
- Gabrb3
  - neurotransmitter receptor subunit
  - absence (or misexpression) yields autism-like behavior
  - what mechanisms influence Gabrb3 expression?



Livet, et al. (2007)

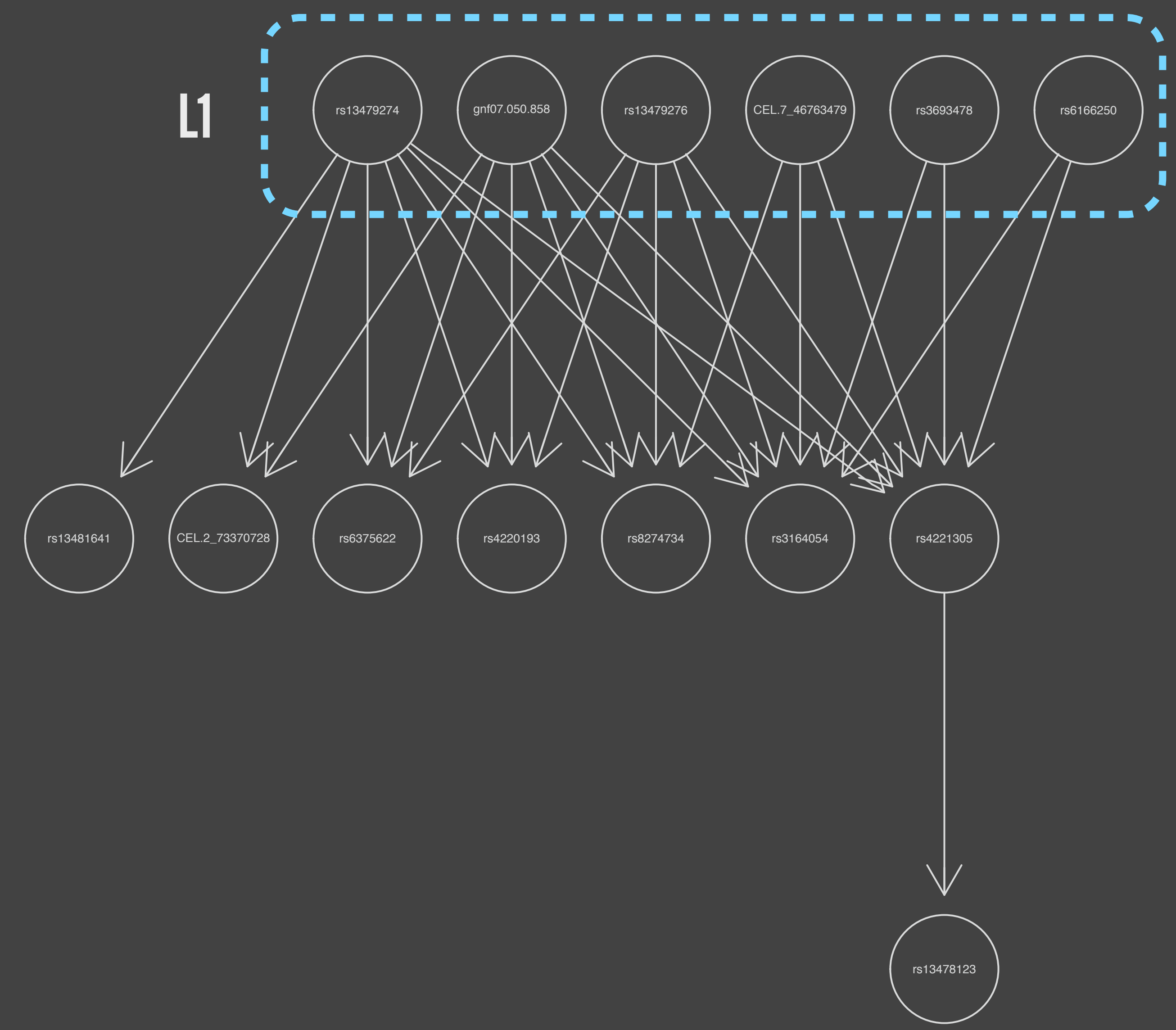
regulation of Gabrb3

grow an RF that regresses hippocampal Gabrb3 expression on the genotypes (m = 3,794) of the same population of mice, then extract the interaction graph



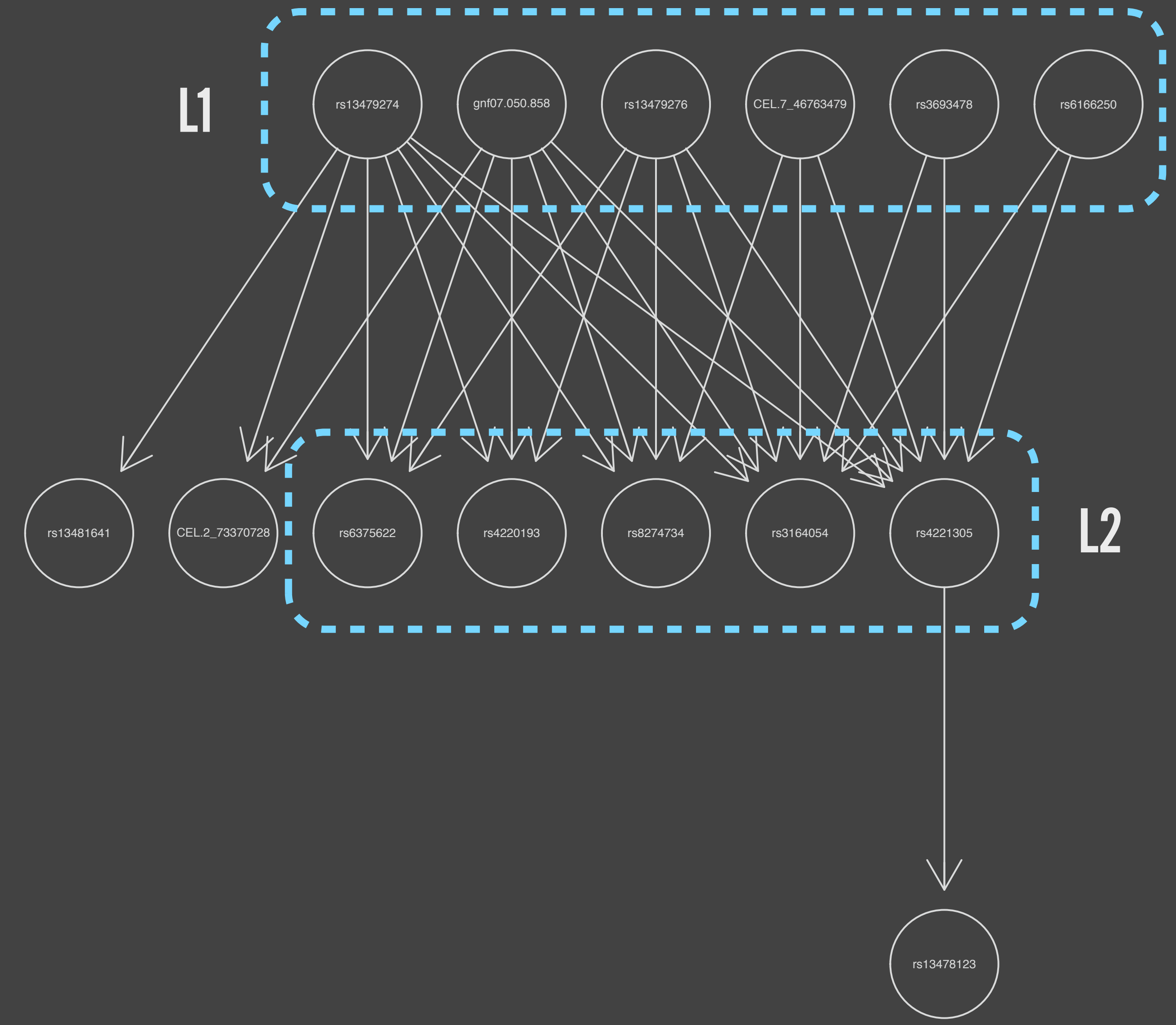
regulation of Gabrb3

grow an RF that regresses hippocampal Gabrb3 expression on the genotypes (m = 3,794) of the same population of mice, then extract the interaction graph



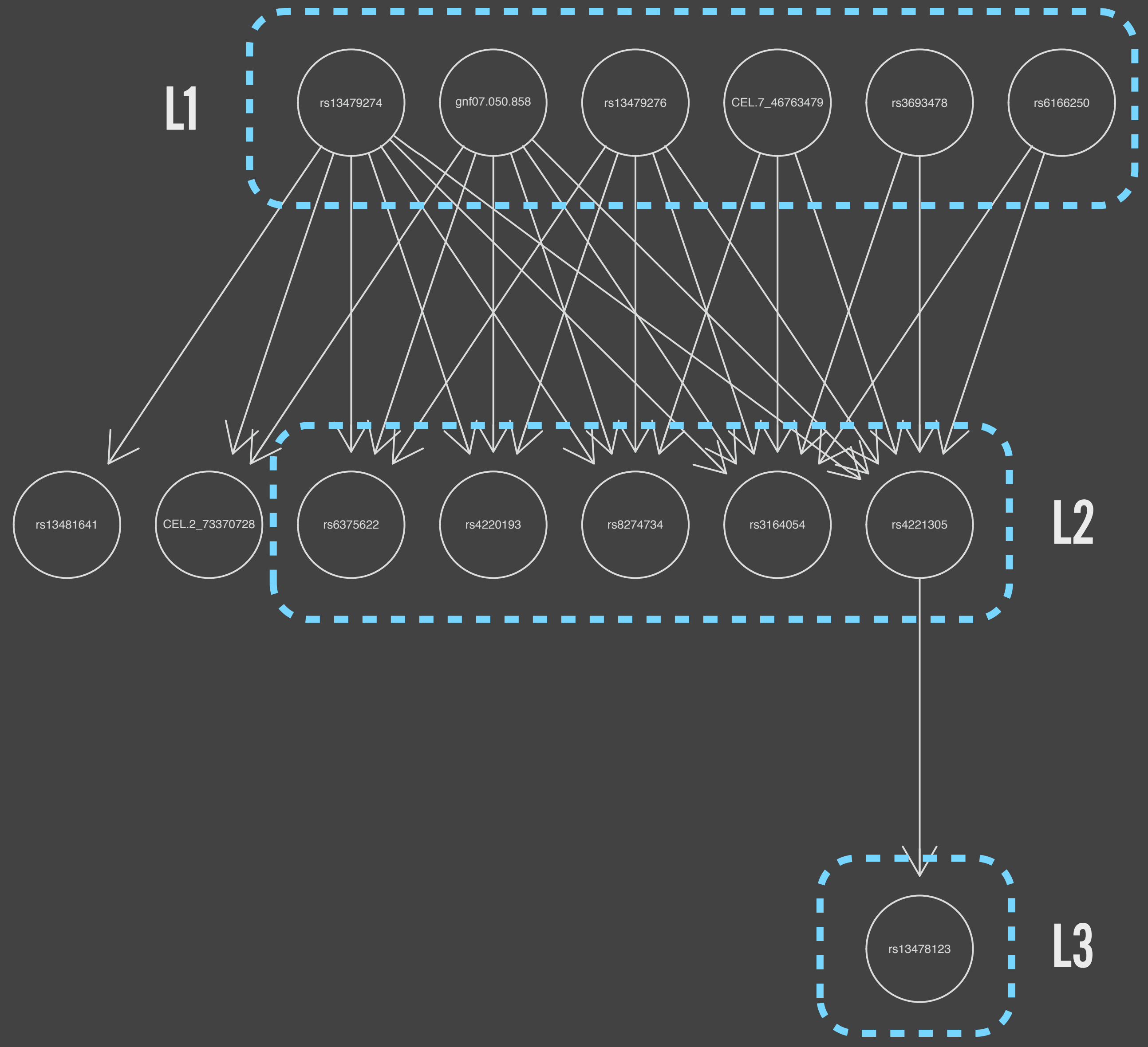
regulation of Gabrb3

grow an RF that regresses hippocampal Gabrb3 expression on the genotypes ($m = 3{,}794$) of the same population of mice, then extract the interaction graph



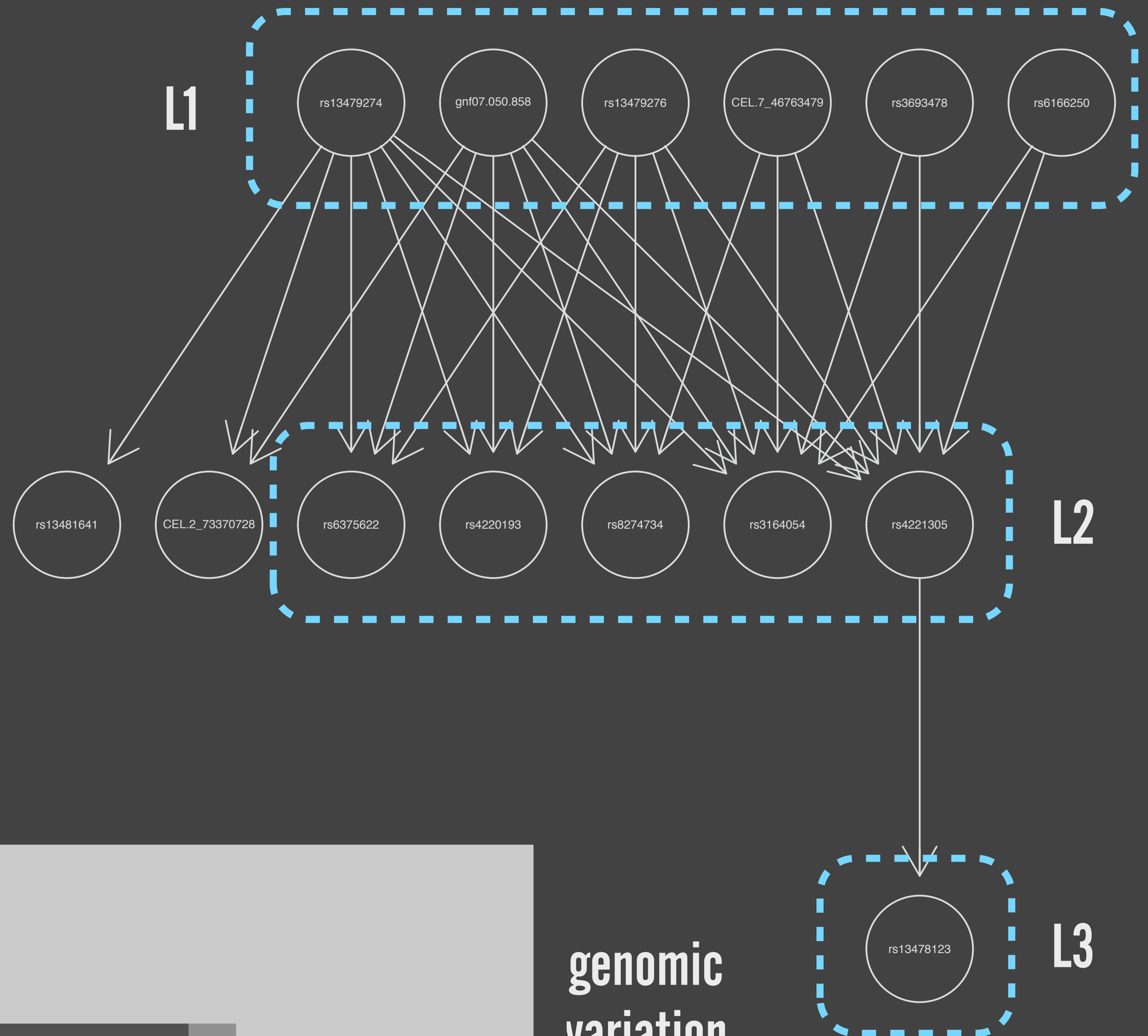
regulation of Gabrb3

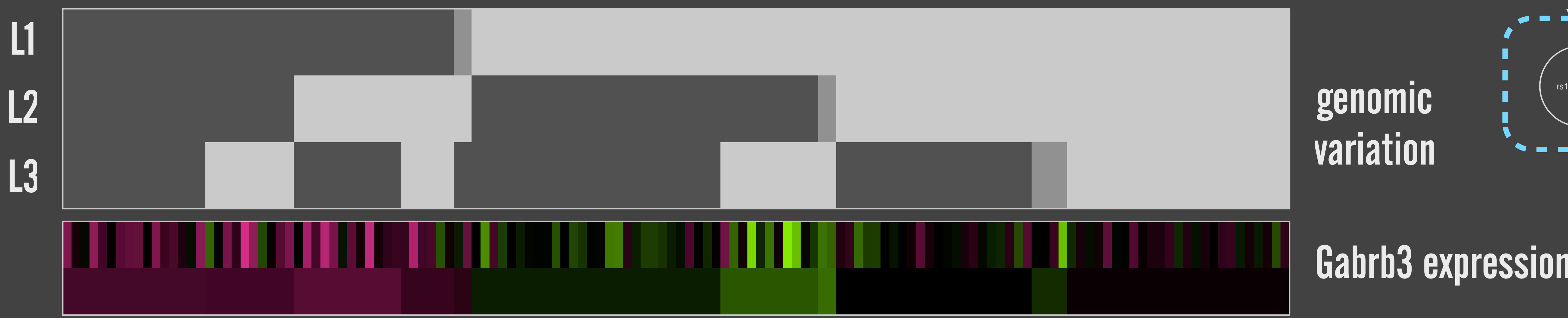
grow an RF that regresses hippocampal Gabrb3 expression on the genotypes (m = 3,794) of the same population of mice, then extract the interaction graph



regulation of Gabrb3

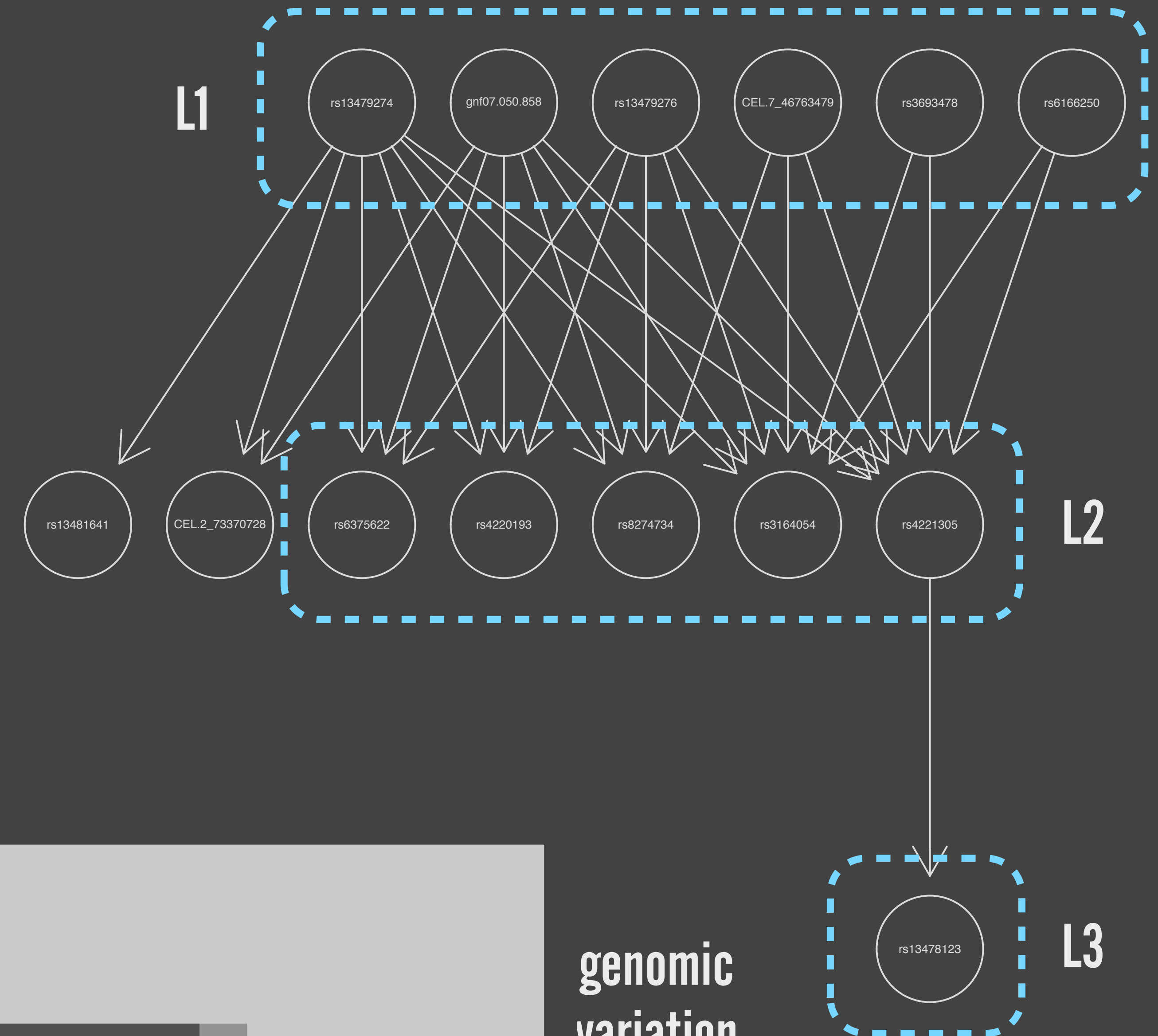
grow an RF that regresses hippocampal Gabrb3 expression on the genotypes (m = 3,794) of the same population of mice, then extract the interaction graph

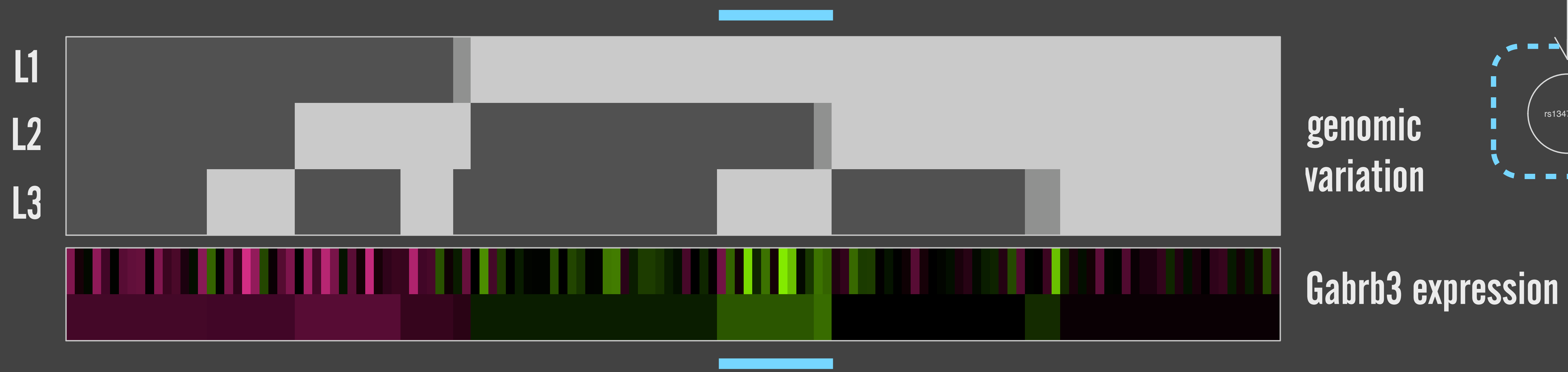




regulation of Gabrb3

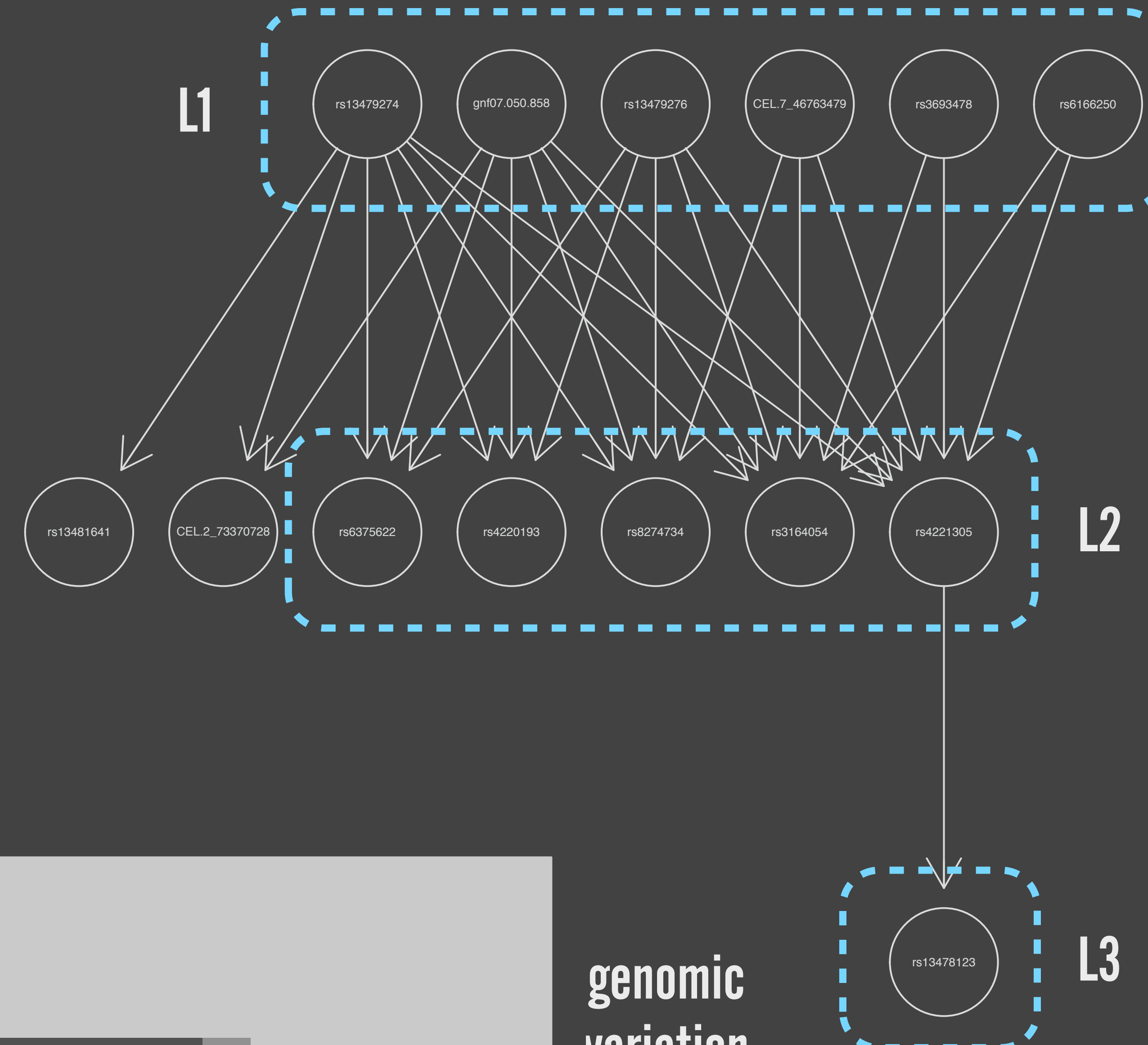
grow an RF that regresses hippocampal Gabrb3 expression on the genotypes (m = 3,794) of the same population of mice, then extract the interaction graph

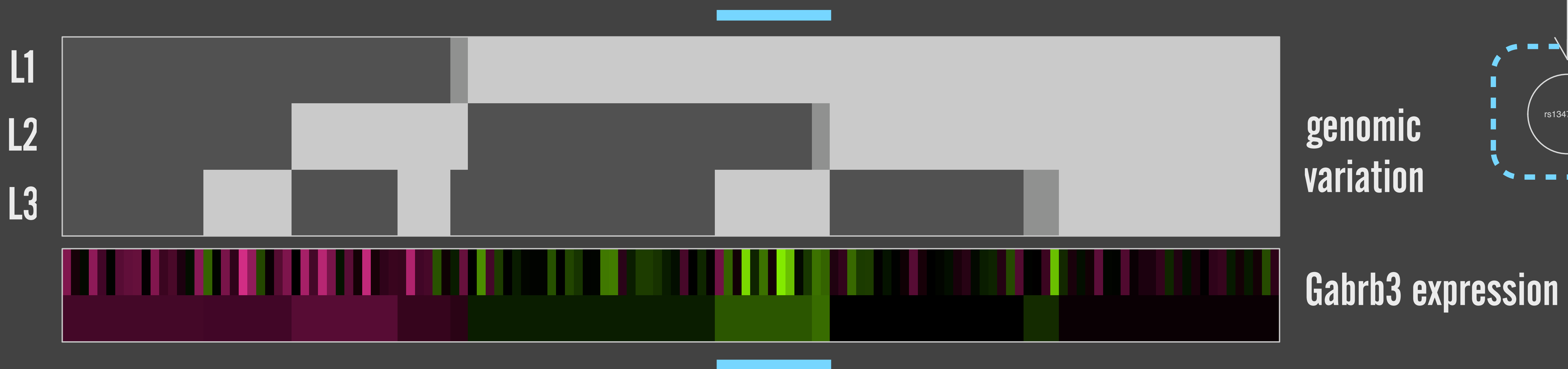




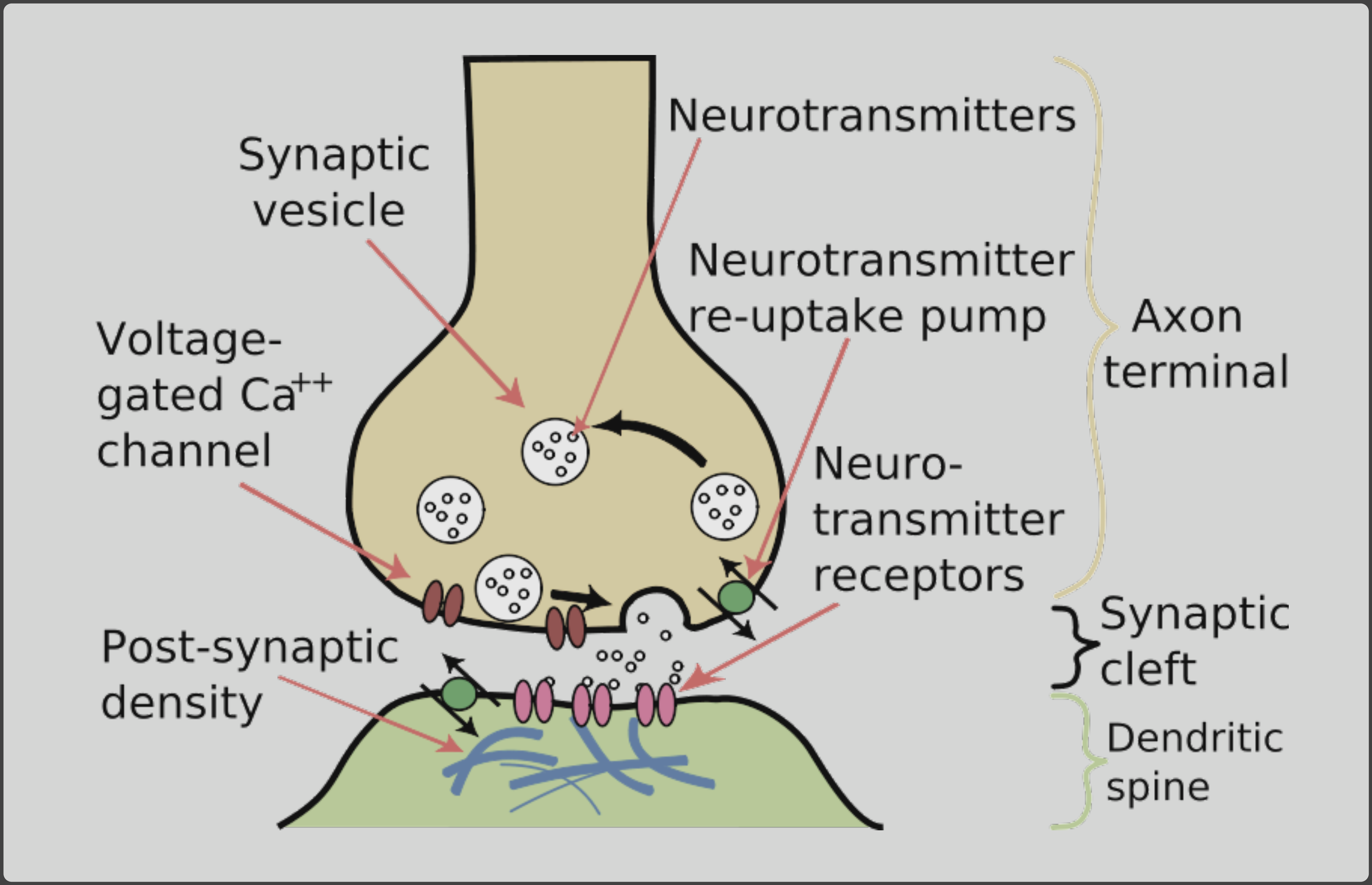
regulation of Gabrb3

L1 - Gabrb3 (cis effect)
L2 - Dscam (axon guidance)
L3 - Magi2 (synaptic scaffolding)

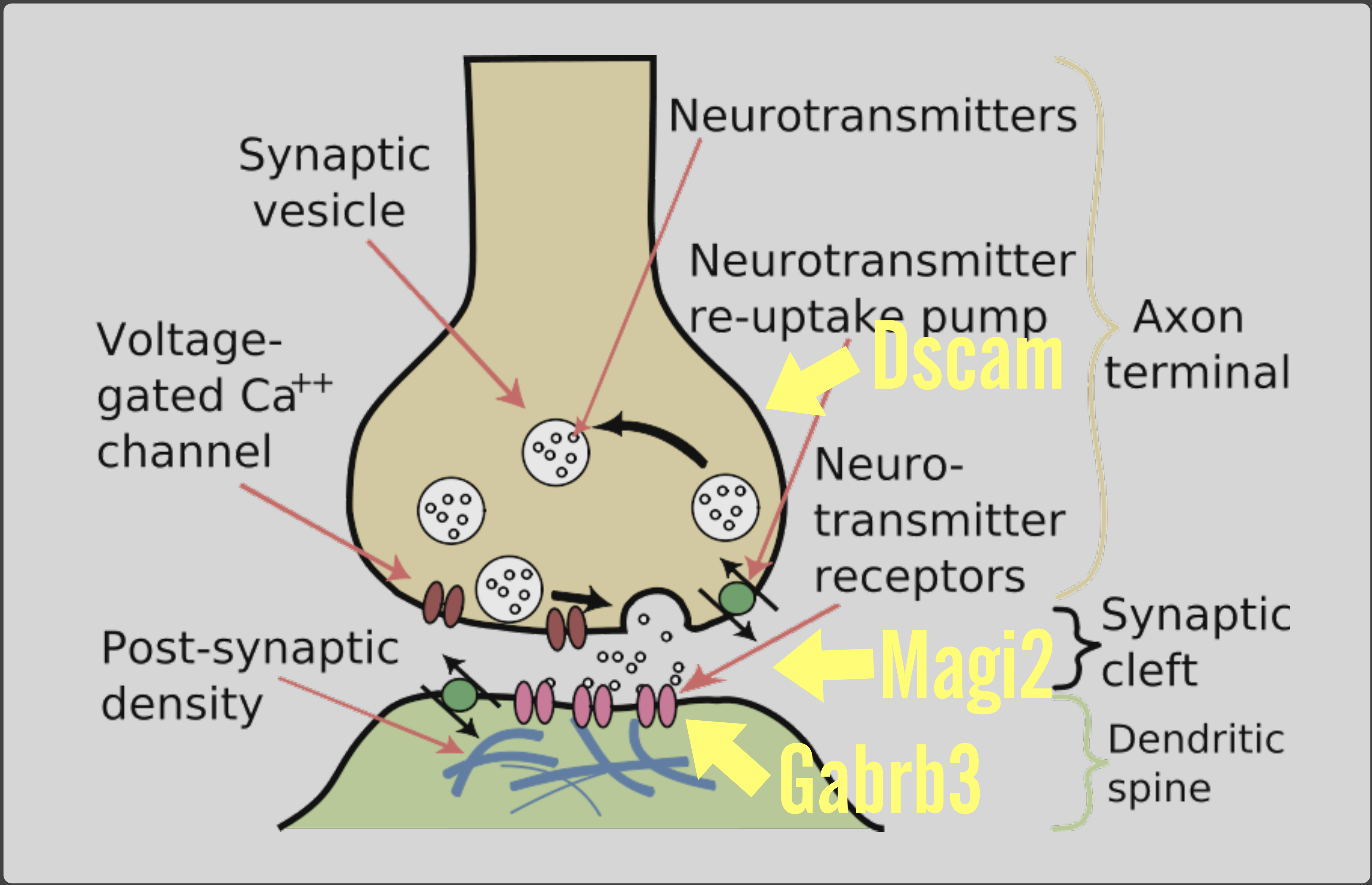




the context


Synaptic
vesicle
Neurotransmitters
Neurotransmitter
re-uptake pump
Voltage-
gated Ca++
channel
Axon
terminal
Neuro-
transmitter
receptors
Post-synaptic
density
Synaptic
cleft
Dendritic
spine

the context


Synaptic vesicle
Neurotransmitters
Neurotransmitter re-uptake pump
Dscam
Axon terminal
Voltage-gated Ca++ channel
Neuro-transmitter receptors
Synaptic cleft
Magi2
Post-synaptic density
Gabrb3
Dendritic spine

conclusion

>> (a)symmetry of transitions between subsequently selected variables can give us clues about the degree of dependence between them

>> constructing a graph of these dependencies can illustrate the emergent dependency structure of the predictors in light of the response

forthcoming...

- does this work for continuous and categorical predictors?
- what about correlated predictors?
- strategy for choosing optimal mtry and ntree?

RF is an example of a tool that is useful in doing analyses of scientific data.

But the cleverest algorithms are no substitute for human intelligence and knowledge of the data in the problem.

Take the output of random forests not as absolute truth, but as smart computer generated guesses that may be helpful in leading to a deeper understanding of the problem.

- Breiman & Cutler

Thanks!

HELMHOLTZ
ASSOCIATION

Alliance on Systems Biology

KTF

THE KLAUS TSCHIRA
FOUNDATION GGMBH

jacob.michaelson@biotec.tu-dresden.de