Notip: Non-parametric True Discovery Proportion control for brain imaging

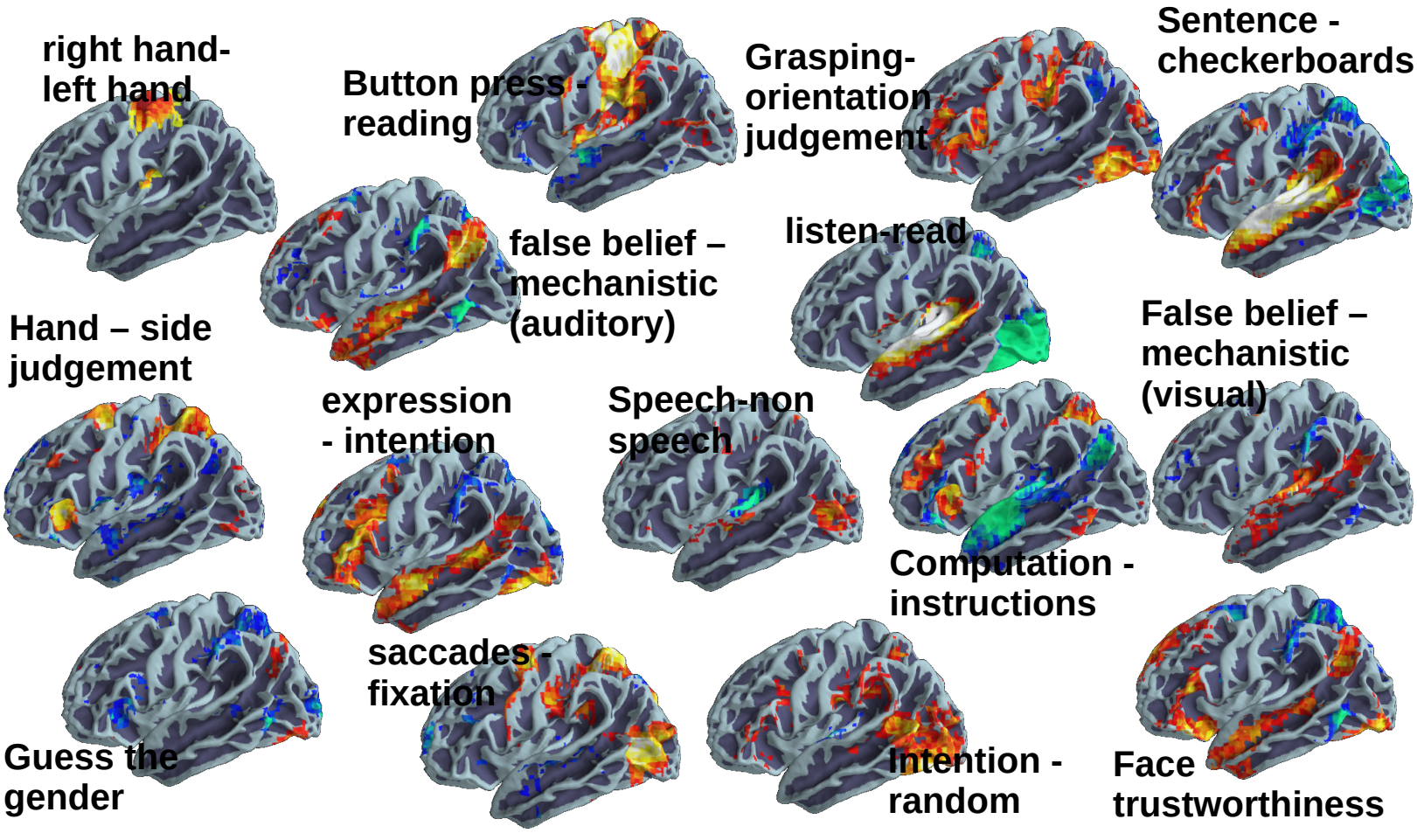
OHBM 2023: Beyond Blobology course

NeuroImage 2022

https://doi.org/10.1016/j.neuroimage.2022.119492

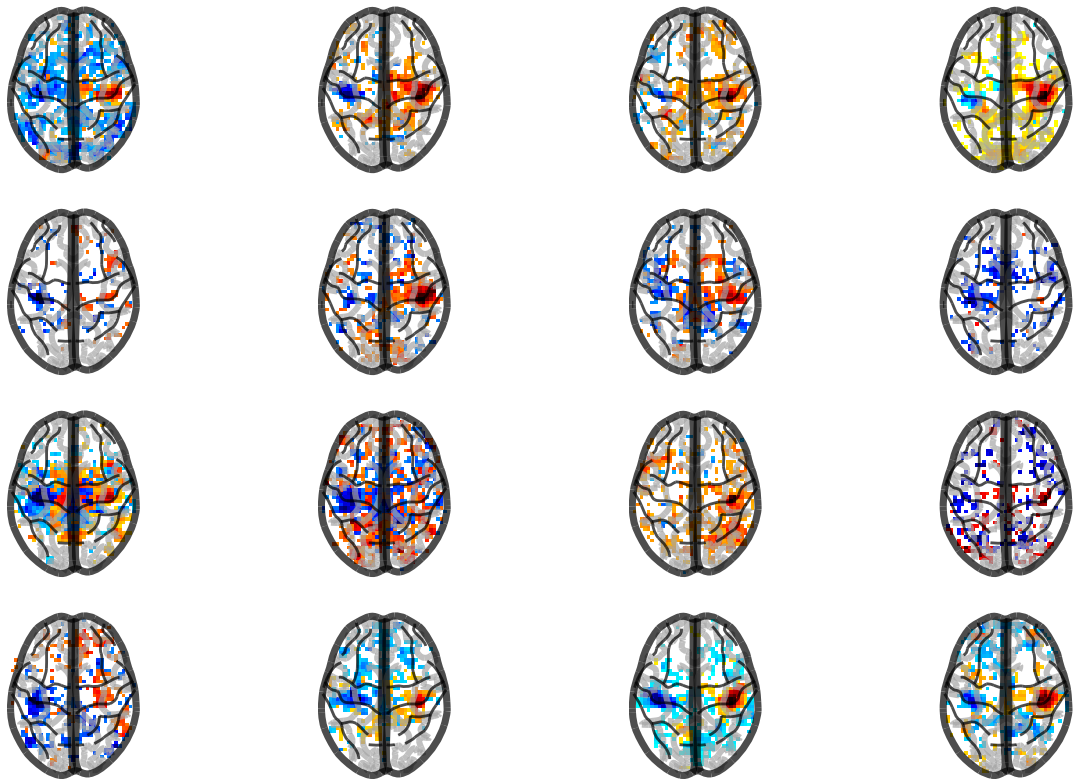
Alexandre Blain
Joint work with Bertrand Thirion, Pierre Neuvial

Mapping the human brain: an important and challenging problem



Detection of the activation of a brain area

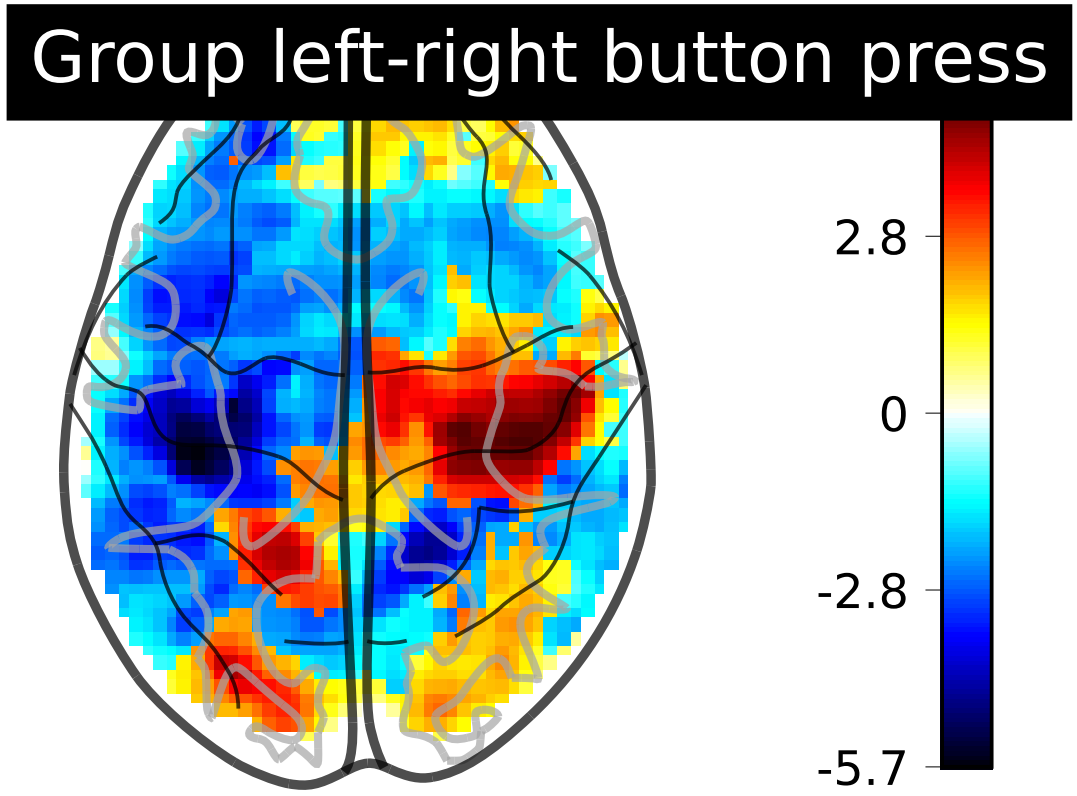
$\hookrightarrow$ Each voxel i: testing $H_{0,i}$: "inactive voxel" vs $H_{1,i}$: "active voxel"



$\Longleftrightarrow n_{samples} = 16, p \simeq 50000$ tests

Detection of the activation of a brain area

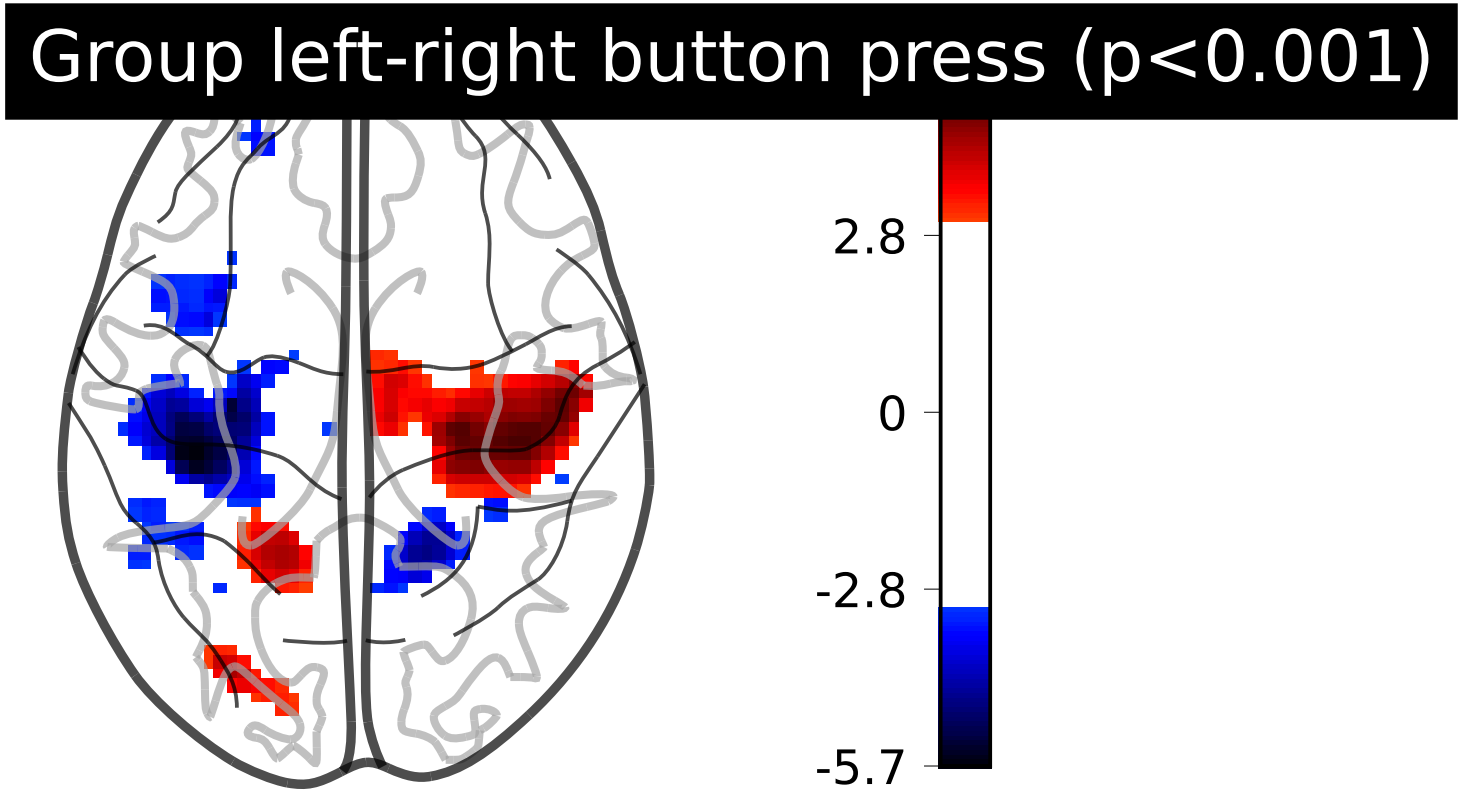
$\hookrightarrow$ Each voxel i: testing $H_{0,i}$: "inactive voxel" vs $H_{1,i}$: "active voxel"



$\Longleftrightarrow$ vector of p-values $[p_1, \ldots, p_m]$

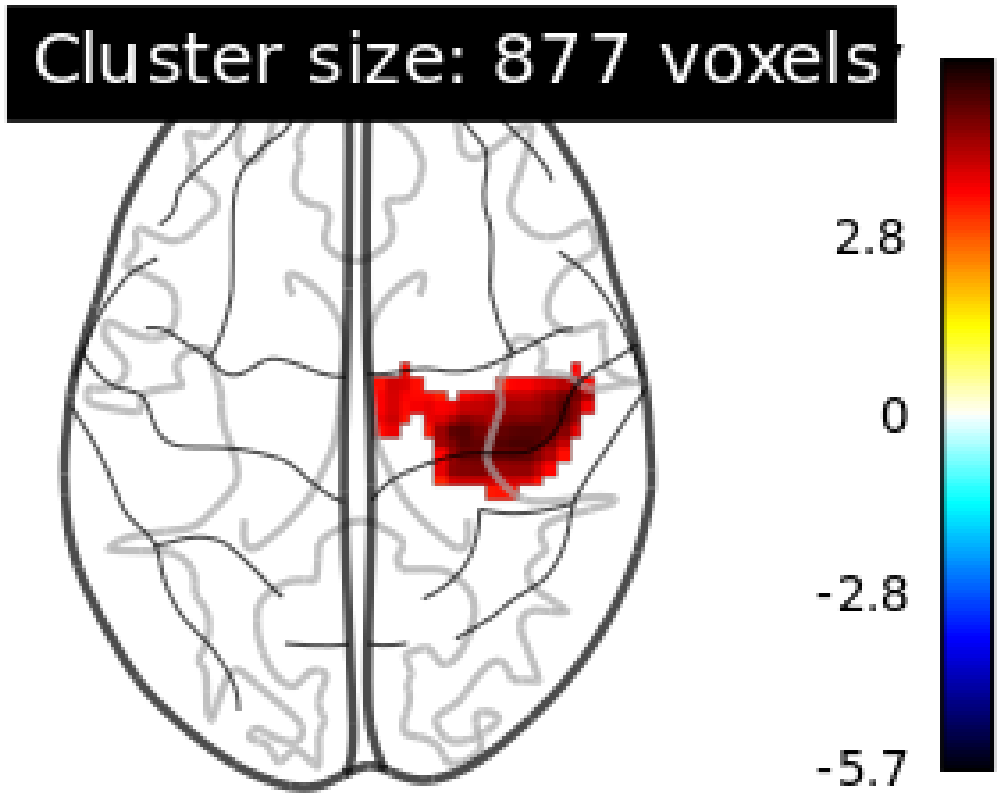
Detection of the activation of a brain area

↪ Each voxel i: testing $H_{0,i}$: "inactive voxel" vs $H_{1,i}$: "active voxel"



Relevant regions extracted by **thresholding** statistics maps

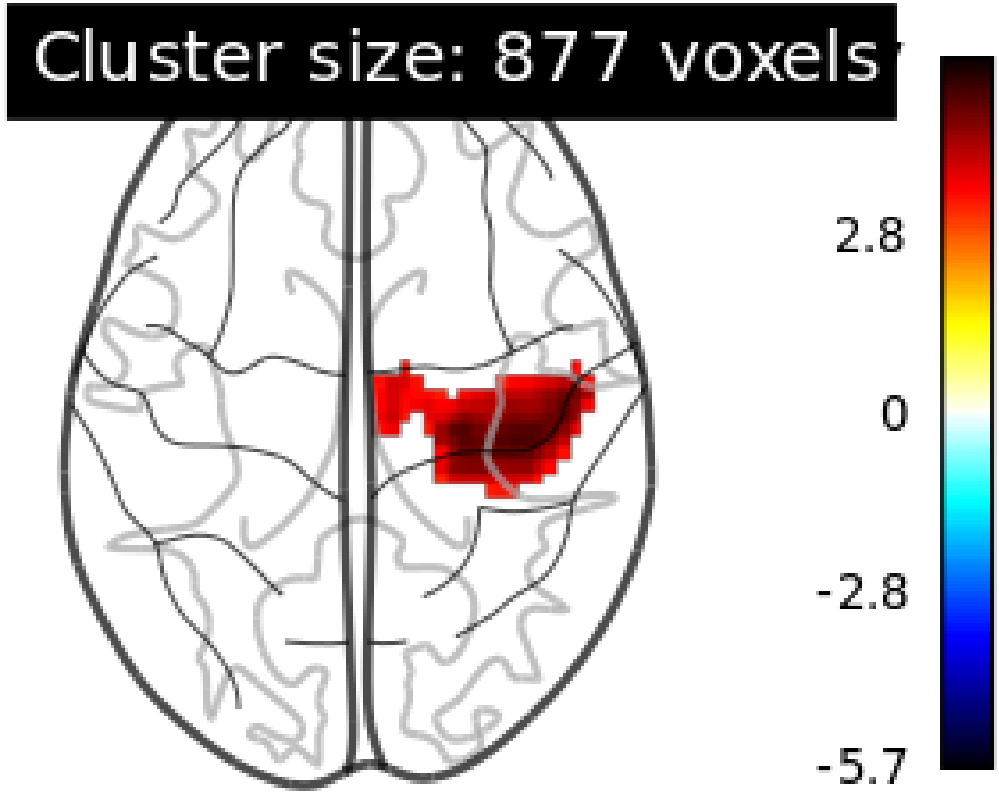
Typical inference = inference on **cluster size** after extracting clusters by **thresholding** [Poline et al. JCBFM 1993]



If **cluster size** > size given by null distribution:

$\hookrightarrow$ Declare cluster **active**

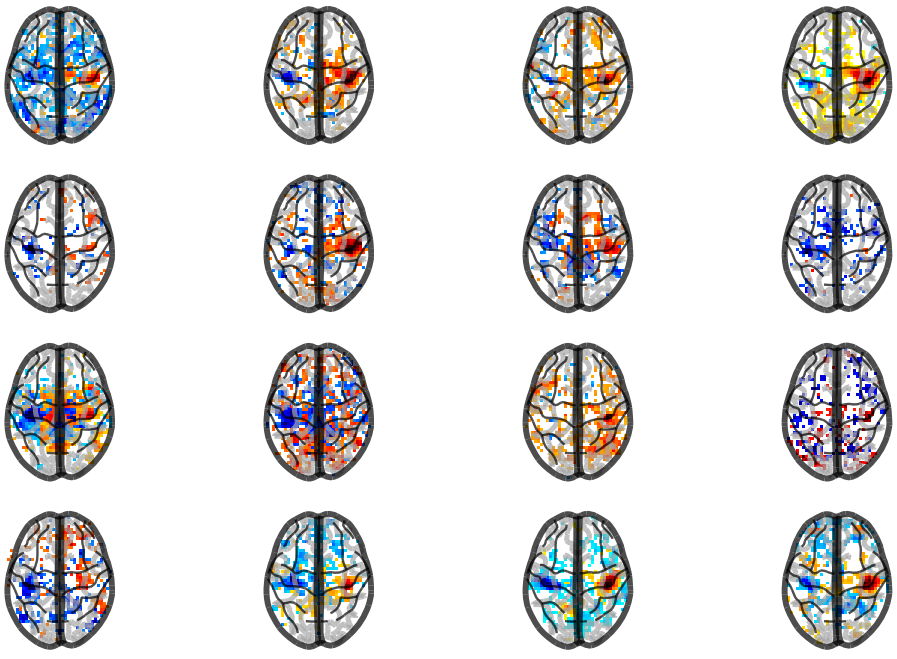
Typical inference = inference on **cluster size** after extracting clusters by **thresholding** [Poline et al. JCBFM 1993]



TDP (=**true proportion of active voxels**) in cluster?

Cluster-level inference is not very informative

TDP lower bound on **data-driven** cluster?



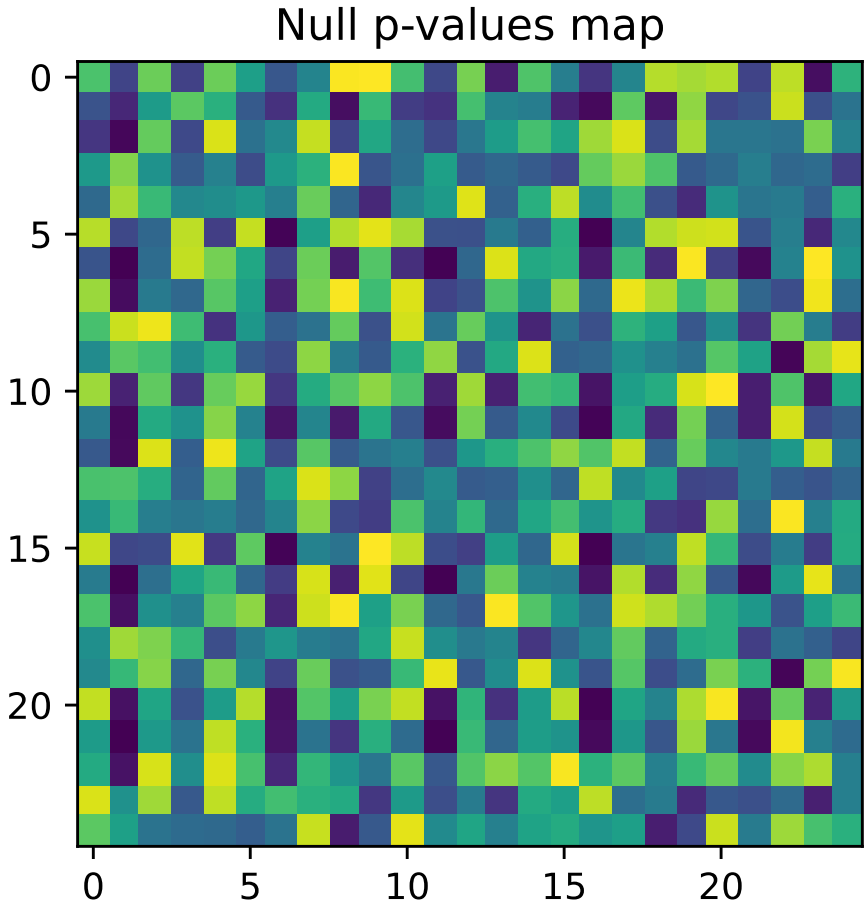
Selection

Inference

→ Bias

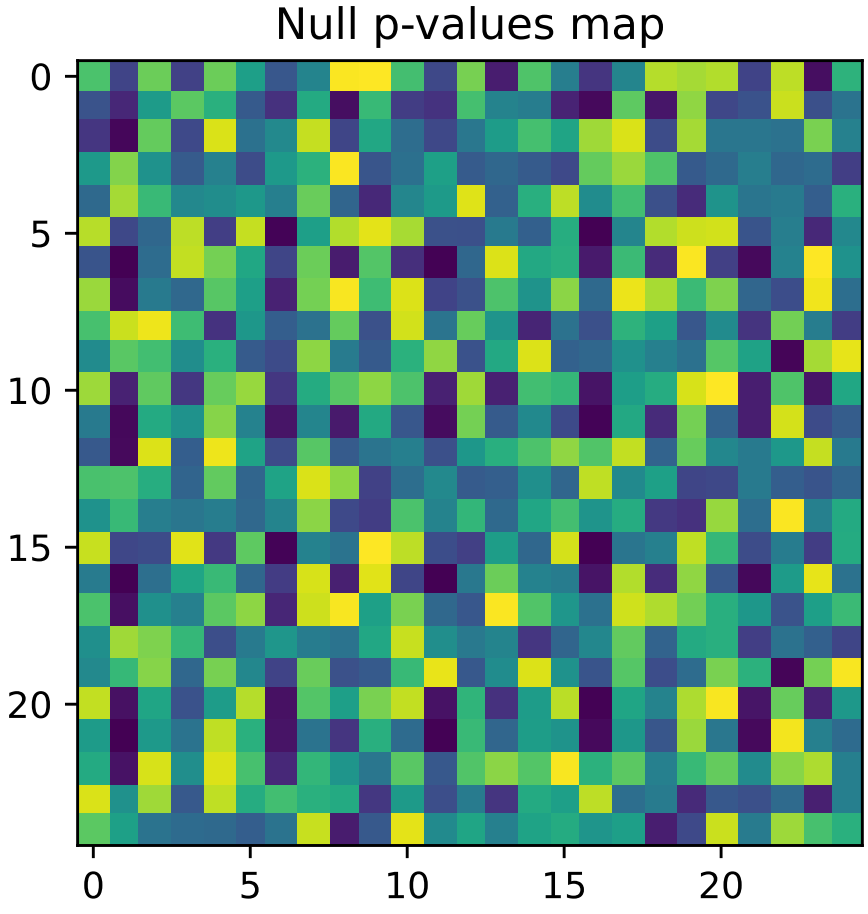
↪ Selective inference

Selective inference can lead to biased discoveries



Canonical approach: inference using [Benjamini-Hochberg 1995]

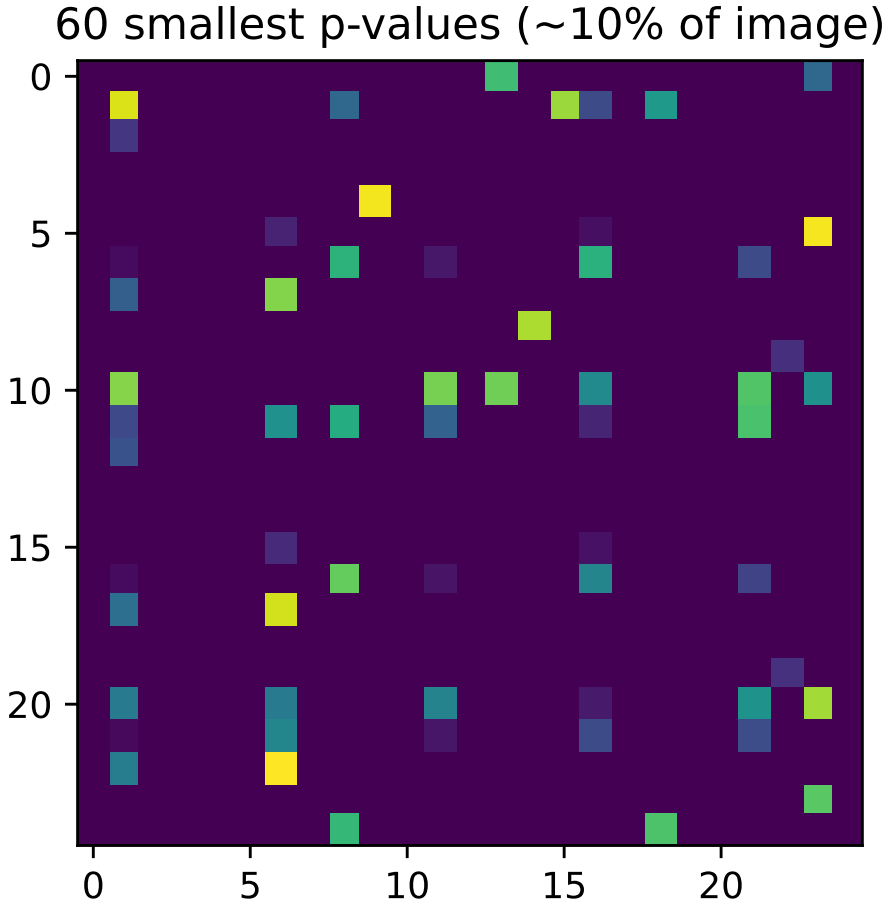
Selective inference can lead to biased discoveries



0 rejections; FDP $(BH_{0.1}) = 0$ %

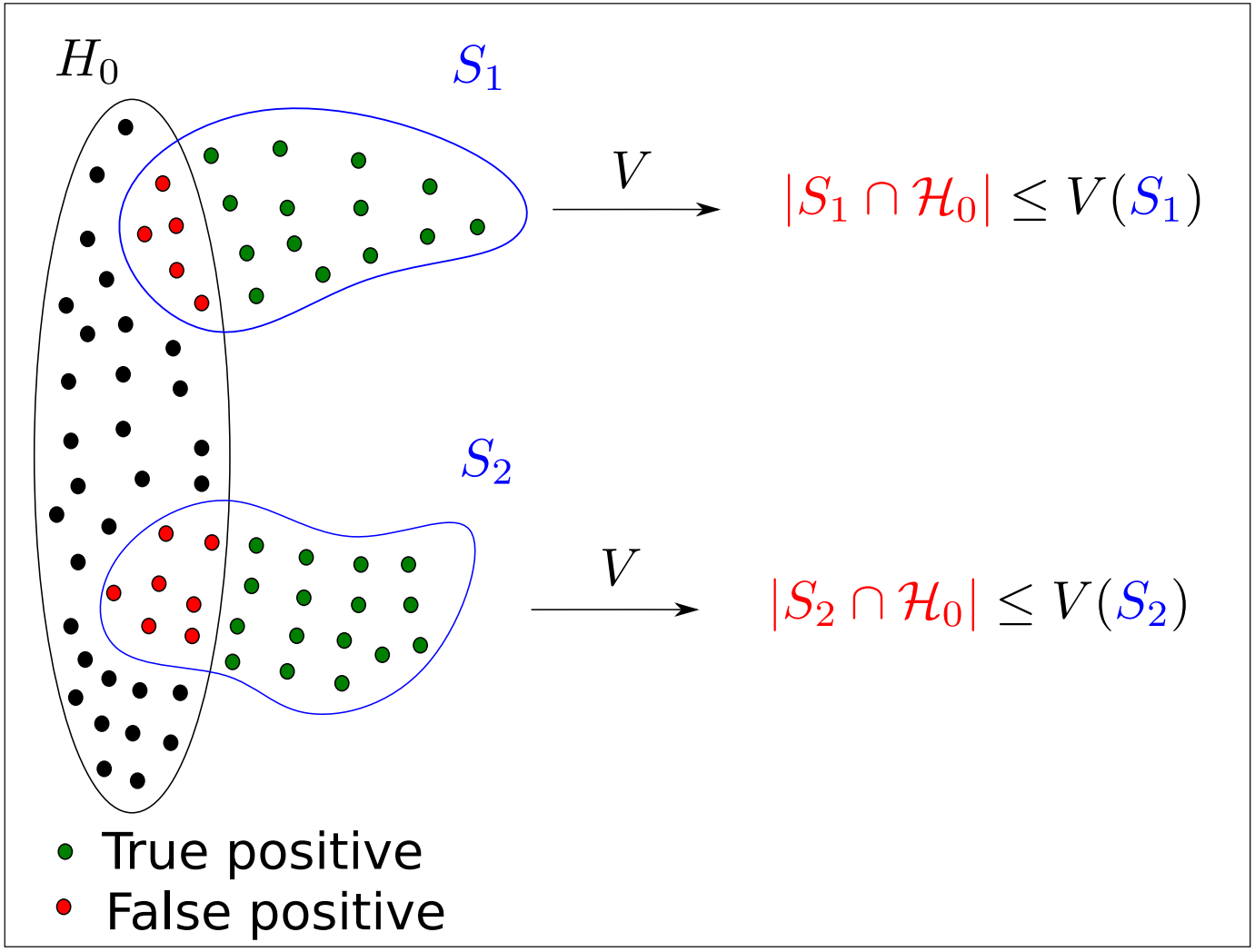
Double dipping

Selective inference can lead to biased discoveries



62% of selected p-values rejected; FDP ($BH_{0.1}$) = 100%!!

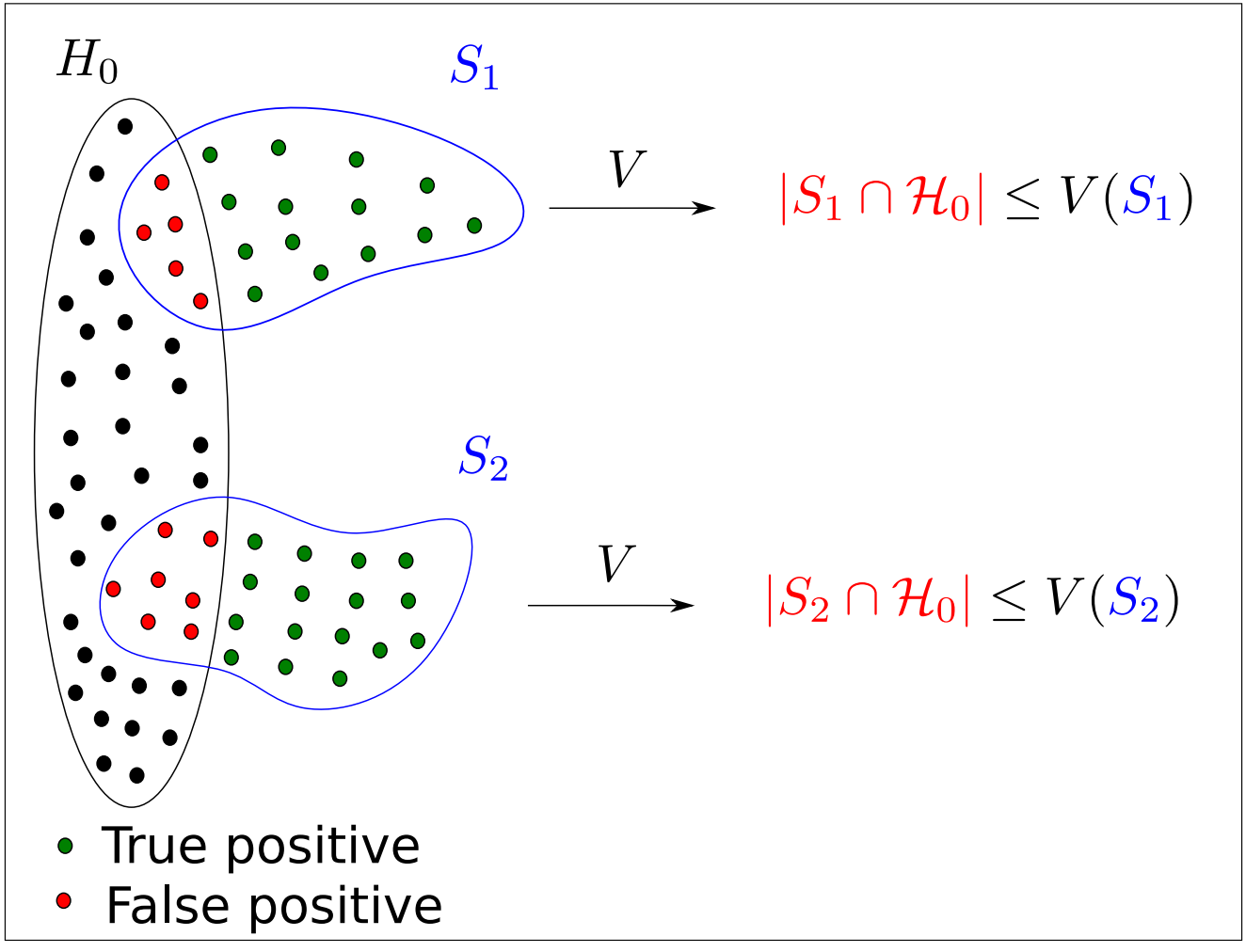
False Discovery Proportion control



$\hookrightarrow$ We want to build V such that:

$$\mathbb{P}\left(\forall S, |S \cap \mathcal{H}_0| \leq V(S)\right) \geq 1 - \alpha$$

False Discovery Proportion control



↪ We want to build V such that:

$$\mathbb{P}\left(\forall S, \mathit{FDP}(S) \leq \frac{V(S)}{|S|}\right) \geq 1 - \alpha$$

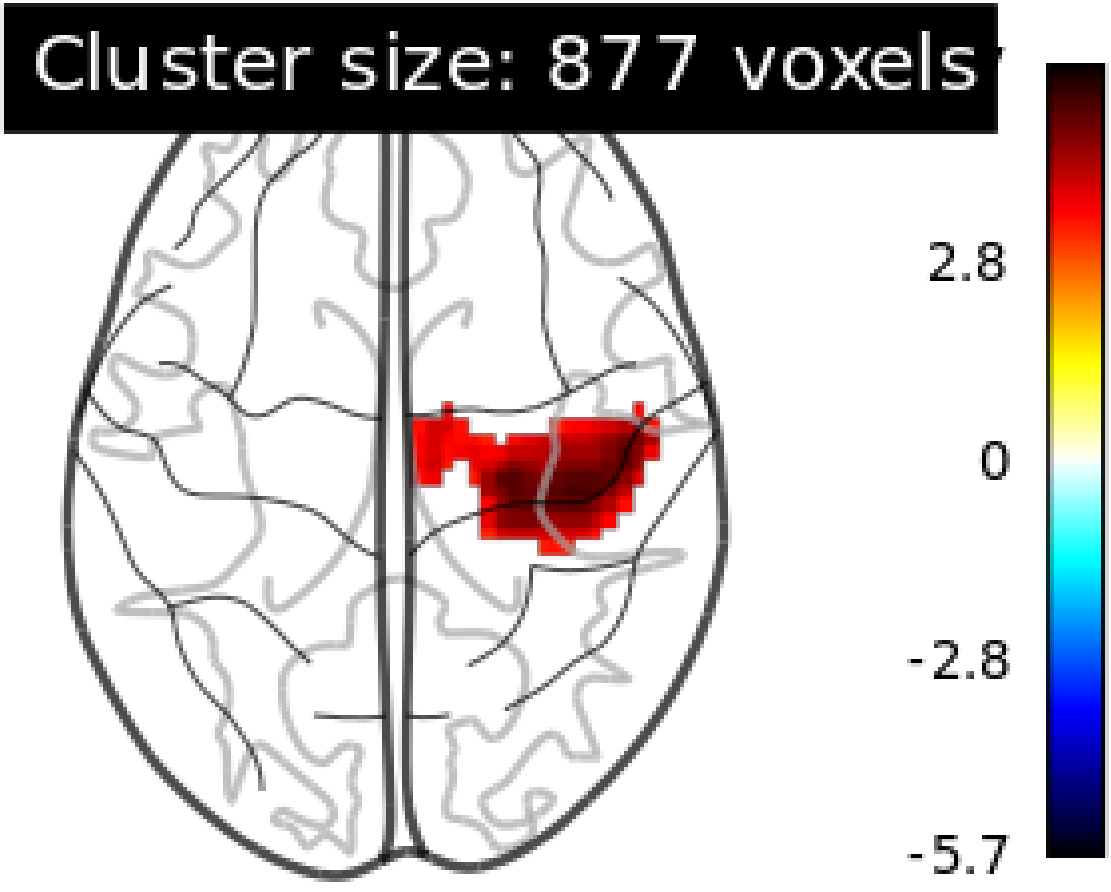
Solution to the selective inference bias: controlling the number of False Positives in **data-driven subsets**

$\hookrightarrow$ **Post-hoc bound**, we can select data-driven subsets [Goeman Solari, 2011] [Genovese Wasserman, 2006]

Parametric bound [Rosenblatt et al., NeuroImage 2018] (ARI):

$$V^{ARI}(S) = \min_{1 \leq k \leq |S|} \left\{ \sum_{i \in S} 1\left\{ p_i(X) \geq \frac{\alpha k}{m} \right\} + k - 1 \right\}$$

ARI: TDP ($\geq 40\%$) with high probability



See [Rosenblatt et al., NeuroImage 2018]

Solution to the selective inference bias: controlling the number of False Positives in **data-driven subsets**

↪ **Post-hoc bound**, we can select data-driven subsets [Goeman and Solari, 2011] [Genovese and Wasserman, 2006]

Parametric bound [Rosenblatt et al., NeuroImage 2018] (ARI):

$$V^{ARI}(S) = \min_{1 \leq k \leq |S|} \left\{ \sum_{i \in S} 1 \left\{ p_i \geq \frac{\alpha k}{m} \right\} + k - 1 \right\}$$

↪ Not adaptive to data characteristics: conservativeness

Solution to the selective inference bias: controlling the number of False Positives in **data-driven subsets**

$\hookrightarrow$ **Post-hoc bound**, we can select data-driven subsets [Goeman and Solari, 2011] [Genovese and Wasserman, 2006]

Semi-parametric bound [Blanchard et al., Annals of Stats 2020], [Andreella et al. 2023]:

$$V(S) = \min_{1 \leq k \leq |S|} \left\{ \sum_{i \in S} 1 \left\{ p_i \geq \frac{\lambda k}{m} \right\} + k - 1 \right\}$$

With λ to choose using randomization on inference data

Solution to the selective inference bias: controlling the number of False Positives in **data-driven subsets**

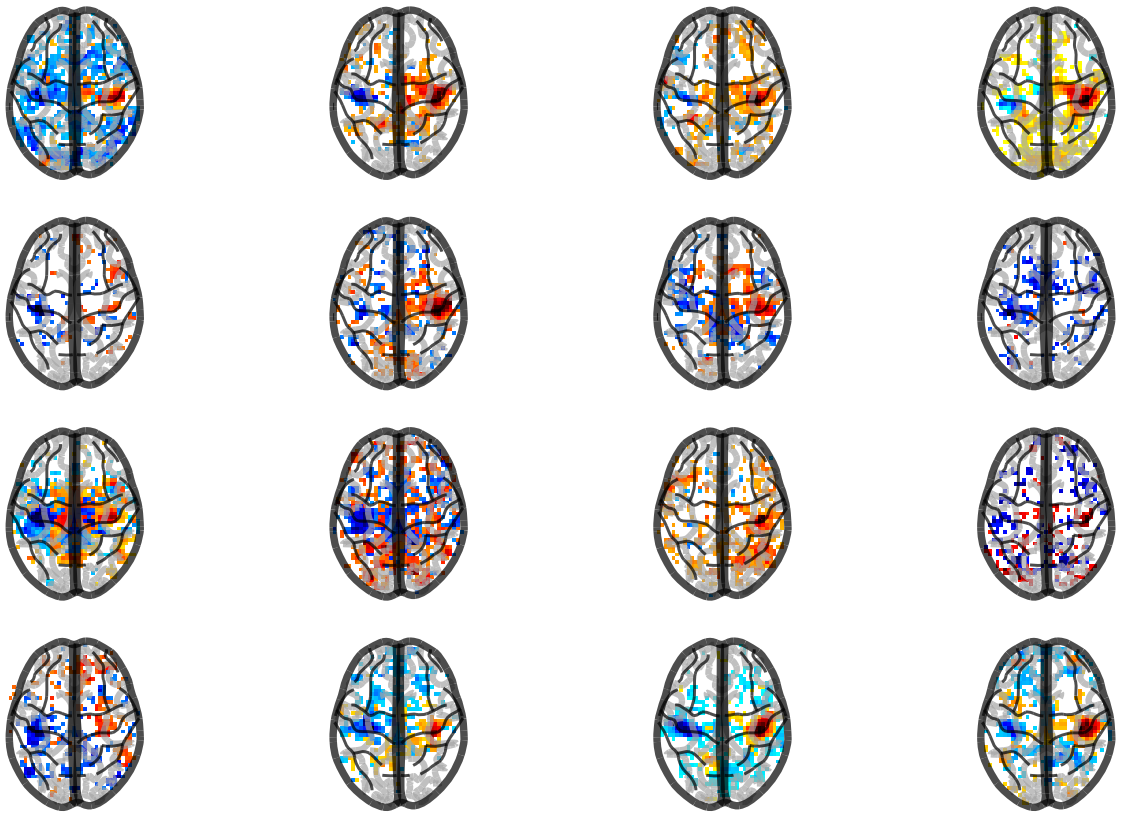
$\hookrightarrow$ **Post-hoc bound**, we can select data-driven subsets [Goeman and Solari, 2011] [Genovese and Wasserman, 2006]

Nonparametric bound [Blain et al., NeuroImage 2022]:

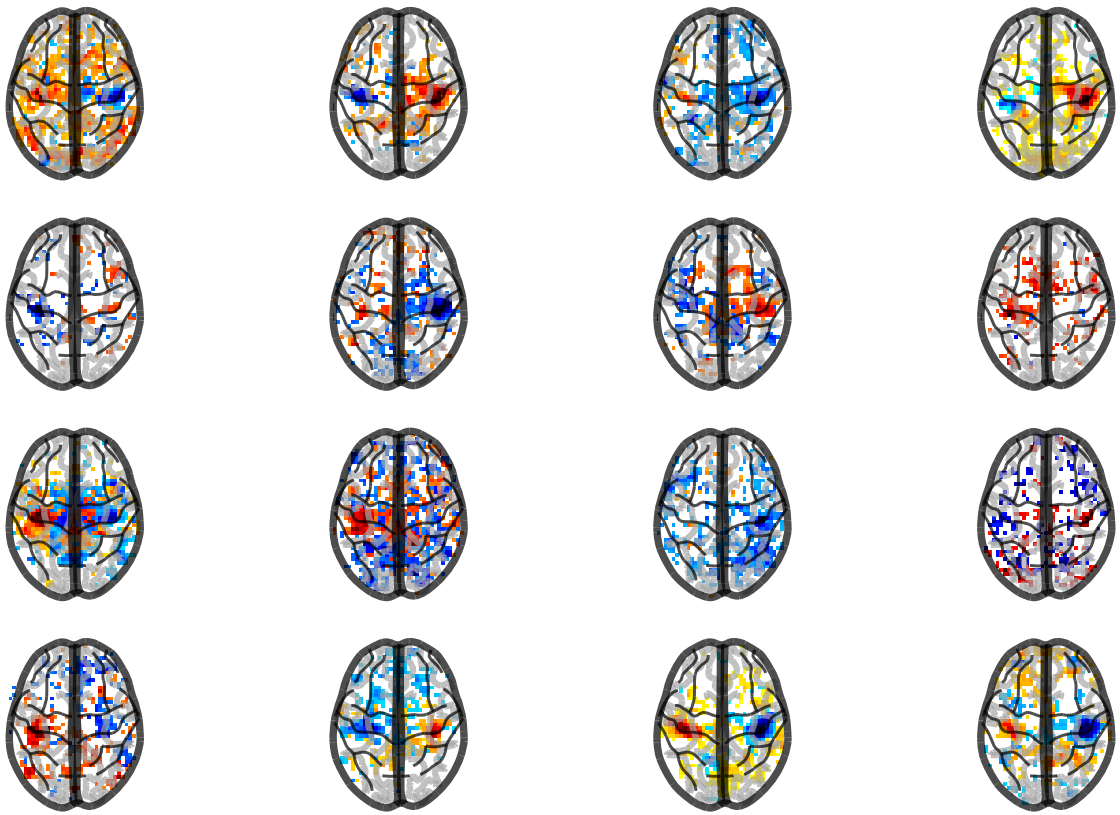
$$V^{Notip}(S) = \min_{1 \leq k \leq |S|} \left\{ \sum_{i \in S} 1\left\{p_i \geq t_k\right\} + k - 1 \right\}$$

With t_k learned from the data

Original data:

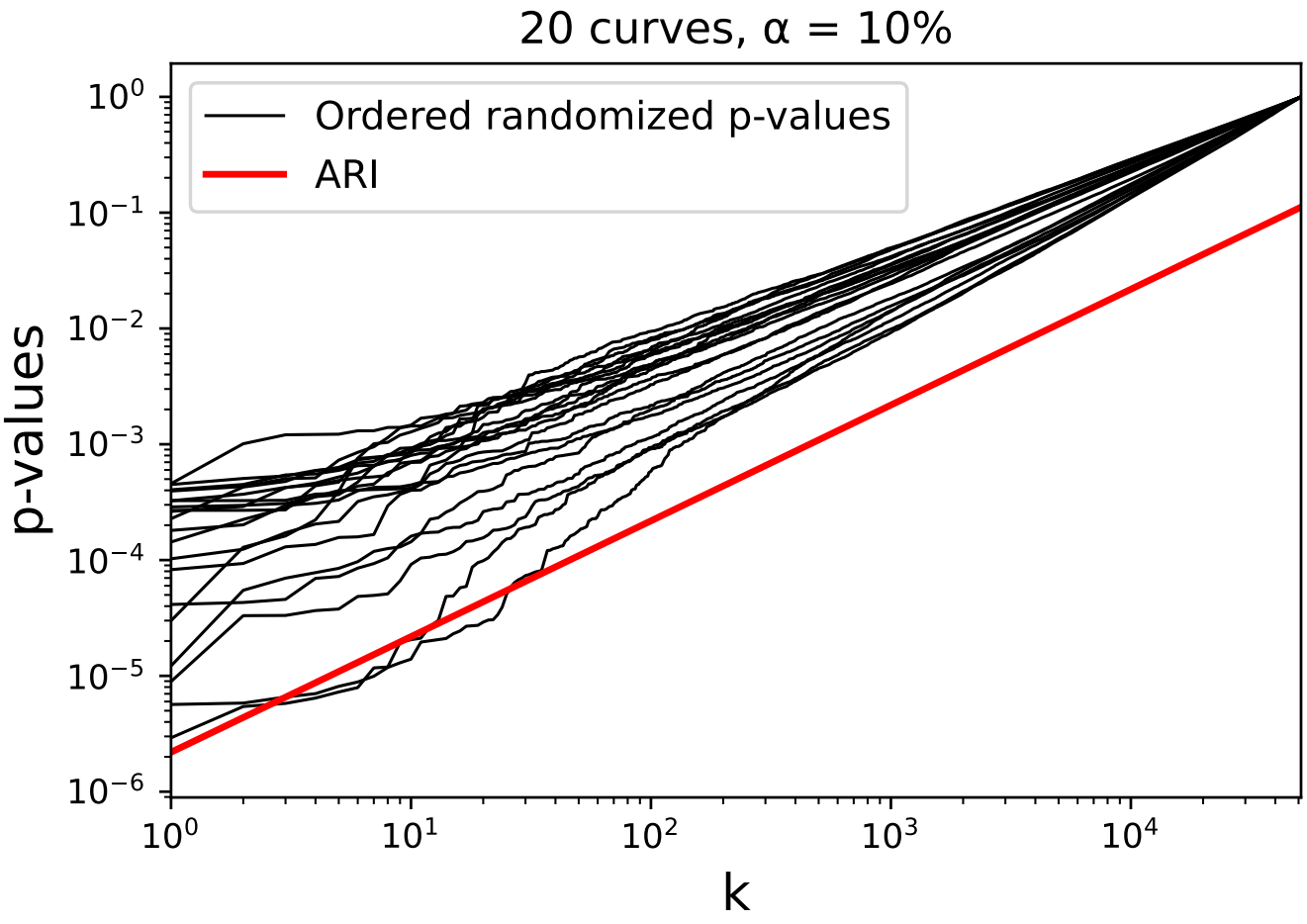


Sign-flipped (i.e. null) data:



↪ Allows **estimating** $p_{(k:\mathcal{H}_0)}$

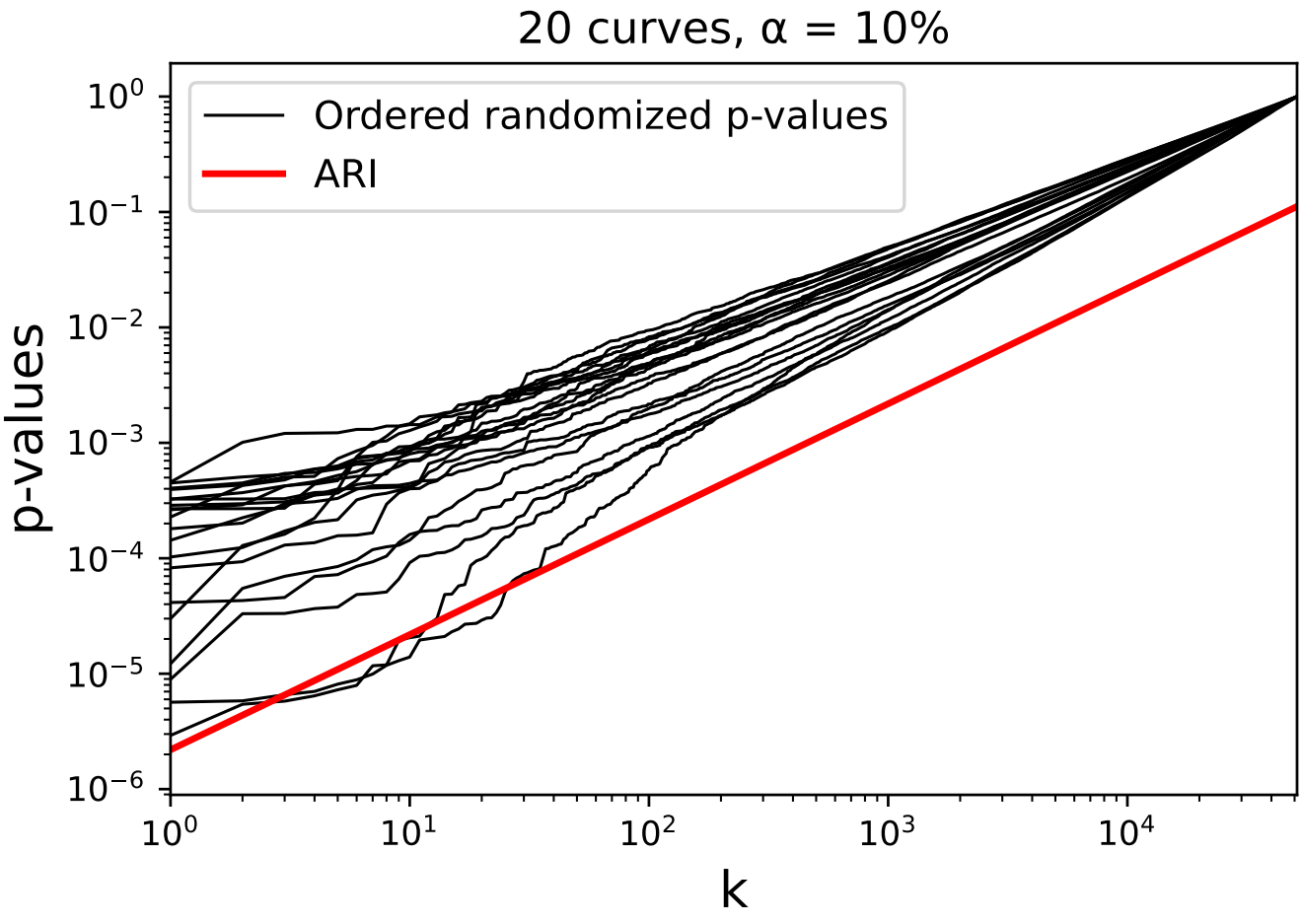
Simes inequality: conservativeness



$$\mathbb{P}\left(\exists k \in \{1, \ldots, m_0\} : p_{(k:\mathcal{H}_0)} < \alpha k/m\right) \leq \alpha$$

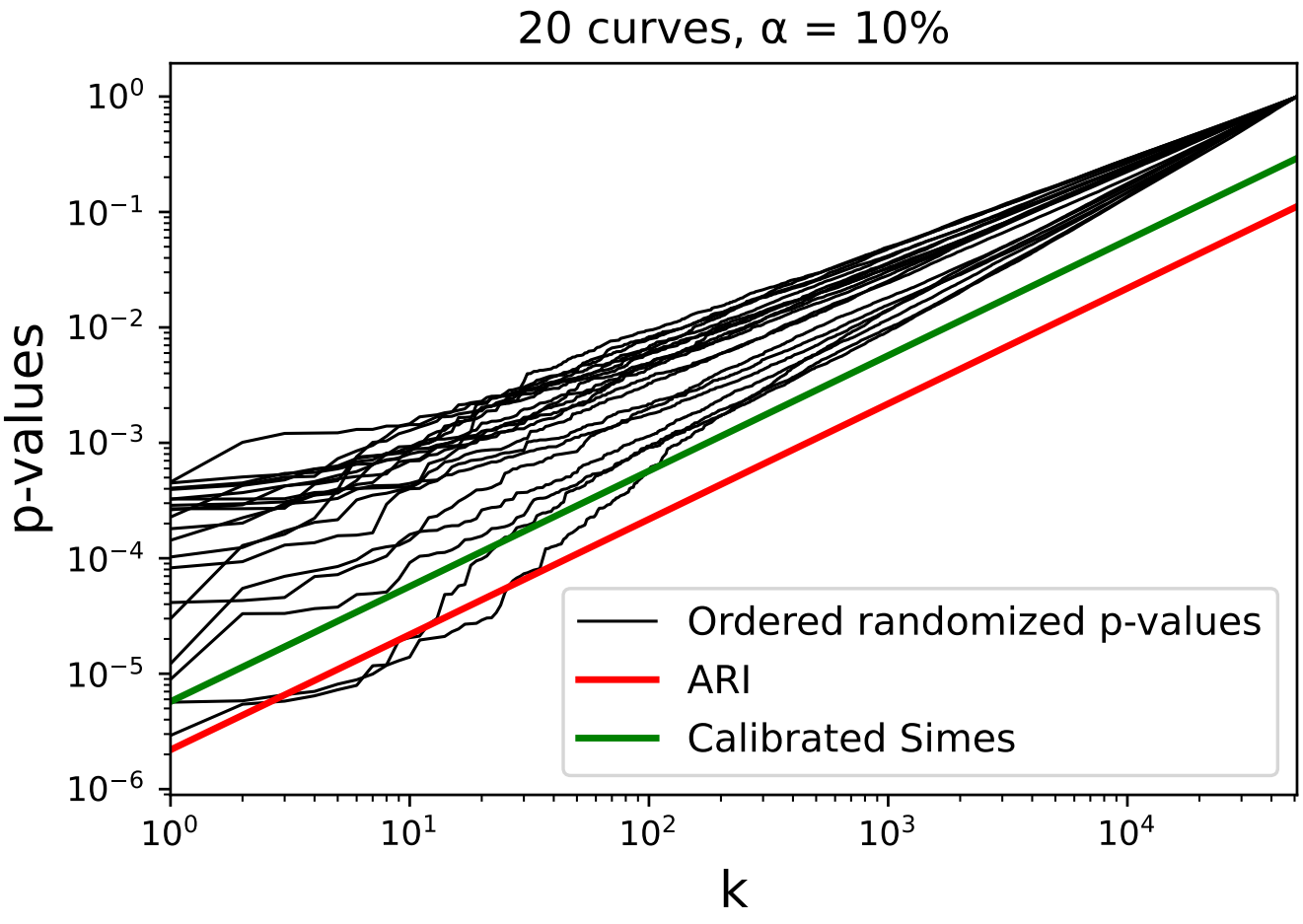
$\Leftrightarrow$ ARI crosses ≤ 2 curves ($= \alpha\%$ of 20 curves)

Simes inequality: conservativeness



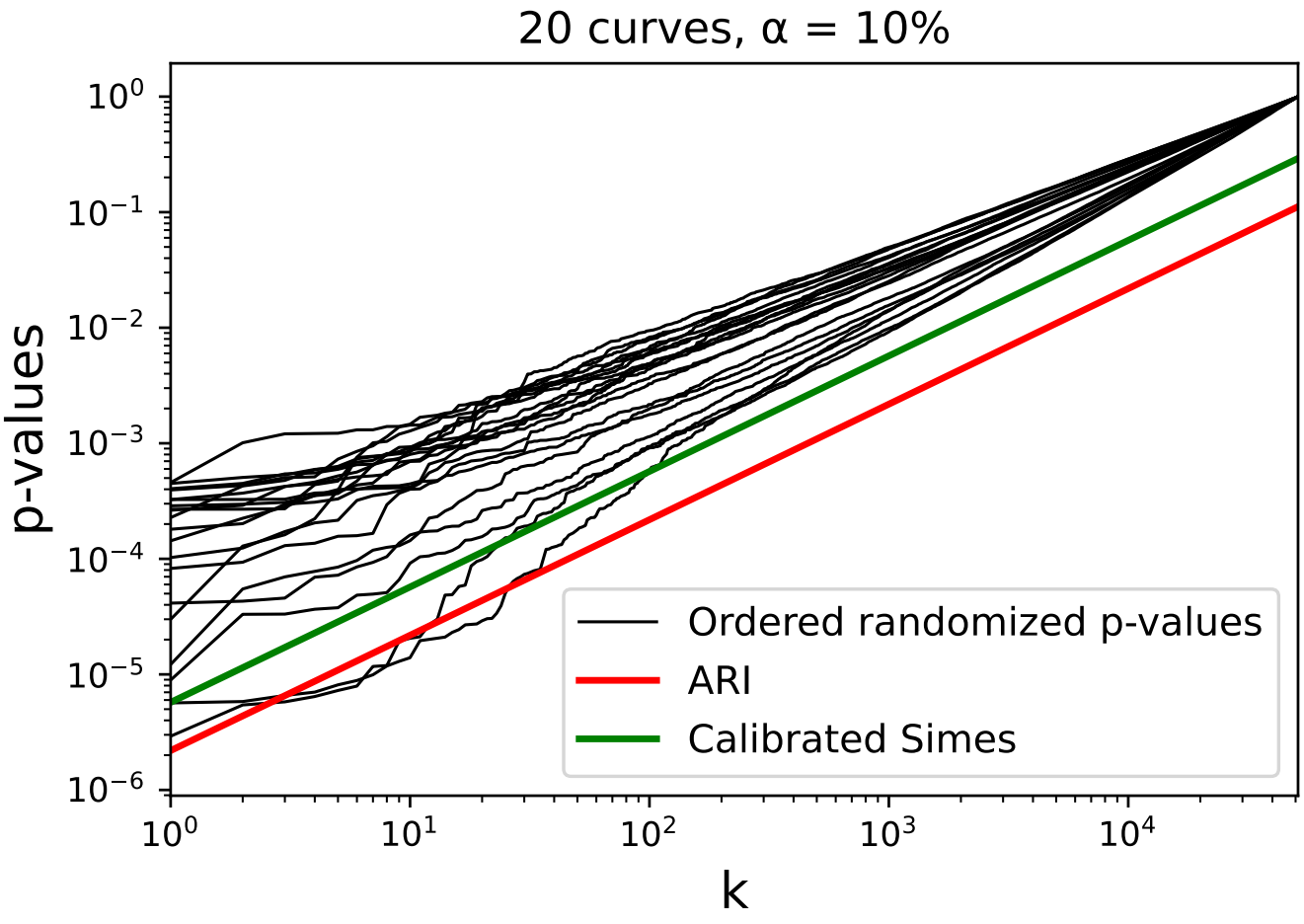
$$\mathbb{P}\left(\exists k \in \{1, \ldots, m_0\} : p_{(k:\mathcal{H}_0)} < \alpha k/m\right) \leq \alpha$$

$\Leftrightarrow$ ARI crosses ≤ 2 curves ($= \alpha\%$ of 20 curves)



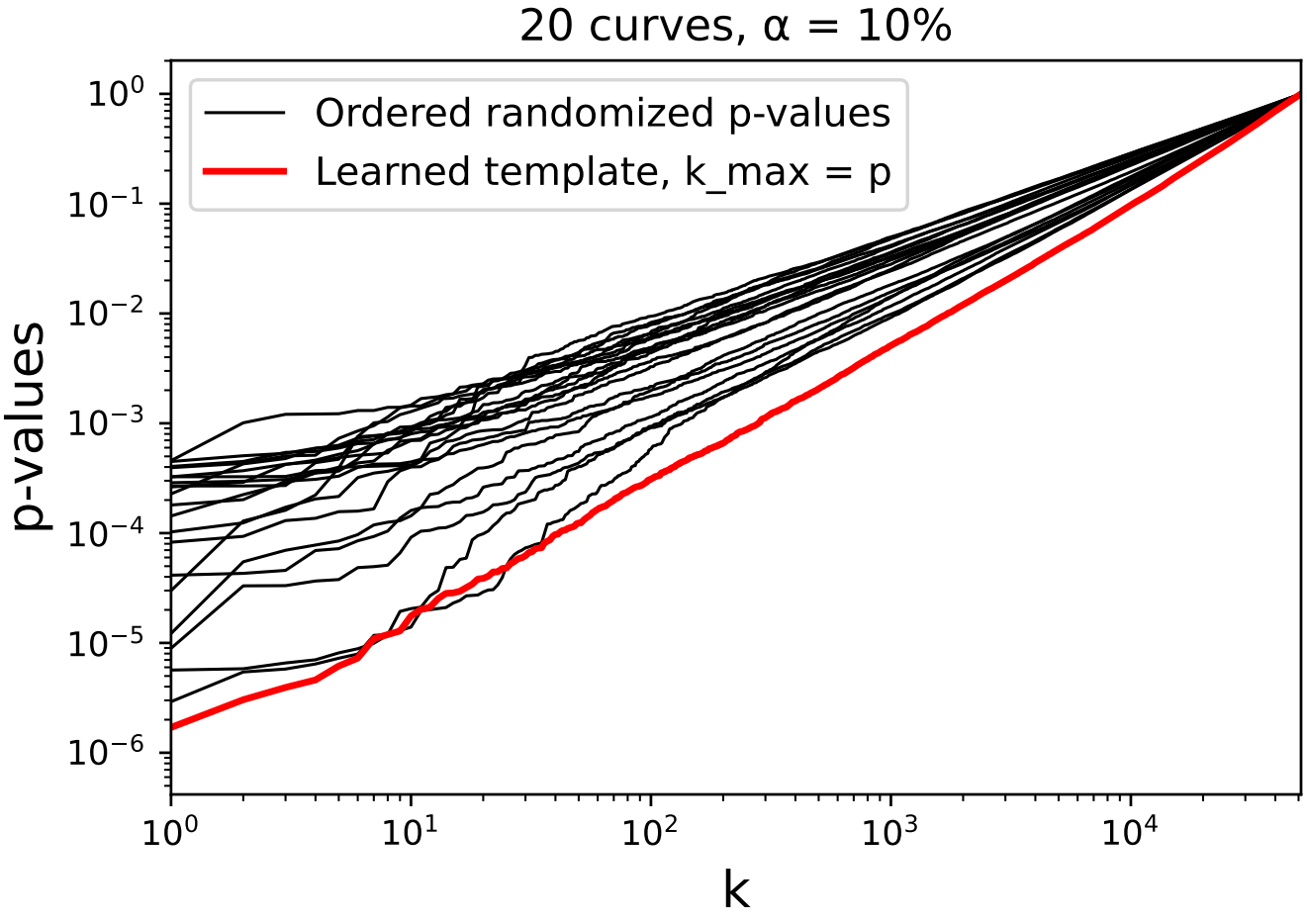
$$\mathbb{P}\left(\exists k \in \{1, \ldots, m_0\} : p_{(k:\mathcal{H}_0)} < \lambda k/m\right) \leq \alpha$$

$\Leftrightarrow$ Calibrated Simes crosses ≤ 2 curves ($= \alpha\%$ of 20 curves)



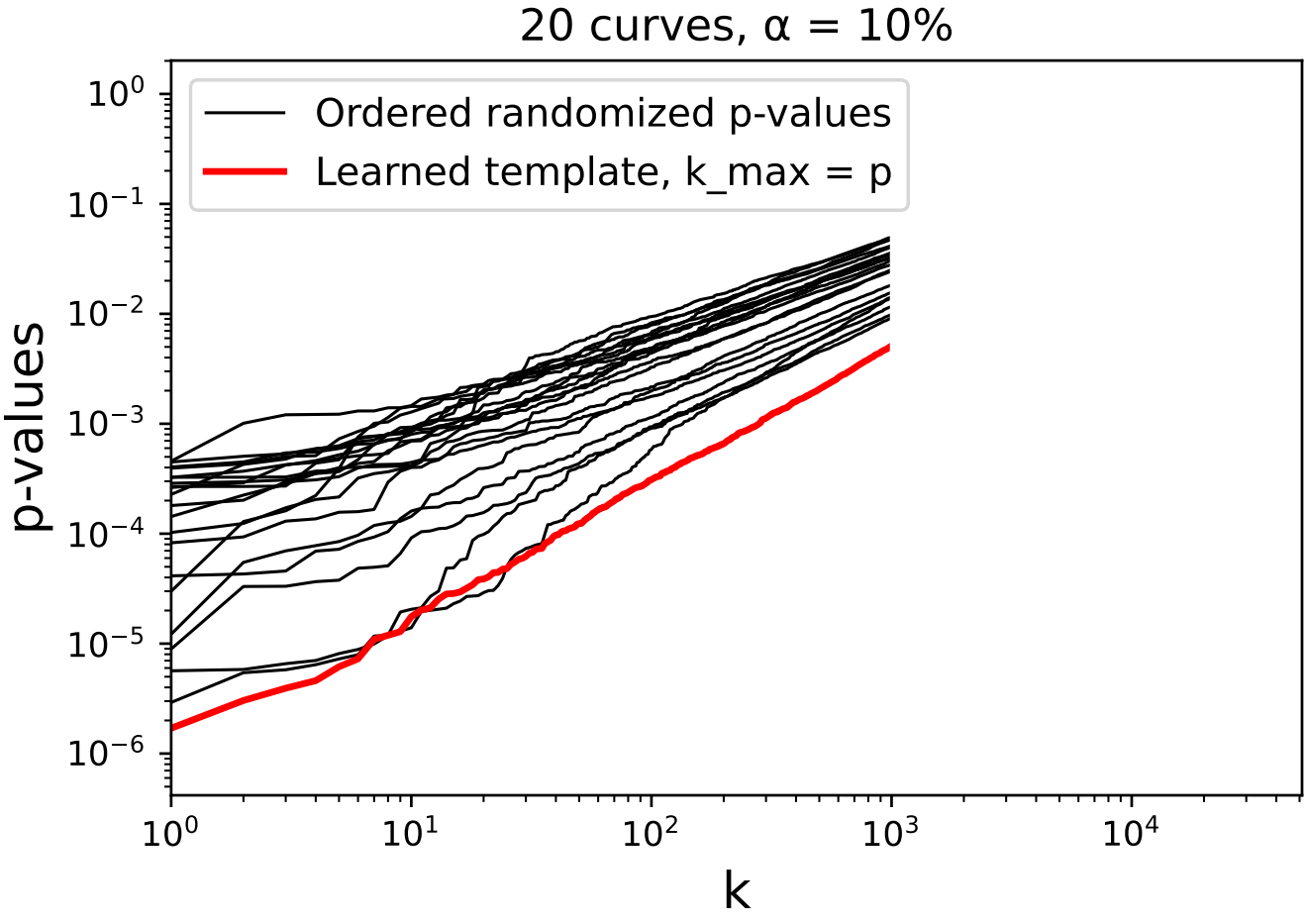
$$\mathbb{P}\left(\exists k \in \{1, \ldots, m_0\} : p_{(k:\mathcal{H}_0)} < t_k\right) \leq \alpha \qquad \text{(JER control)}$$

$\Leftrightarrow$ t_k crosses ≤ 2 curves ($= \alpha\%$ of 20 curves)



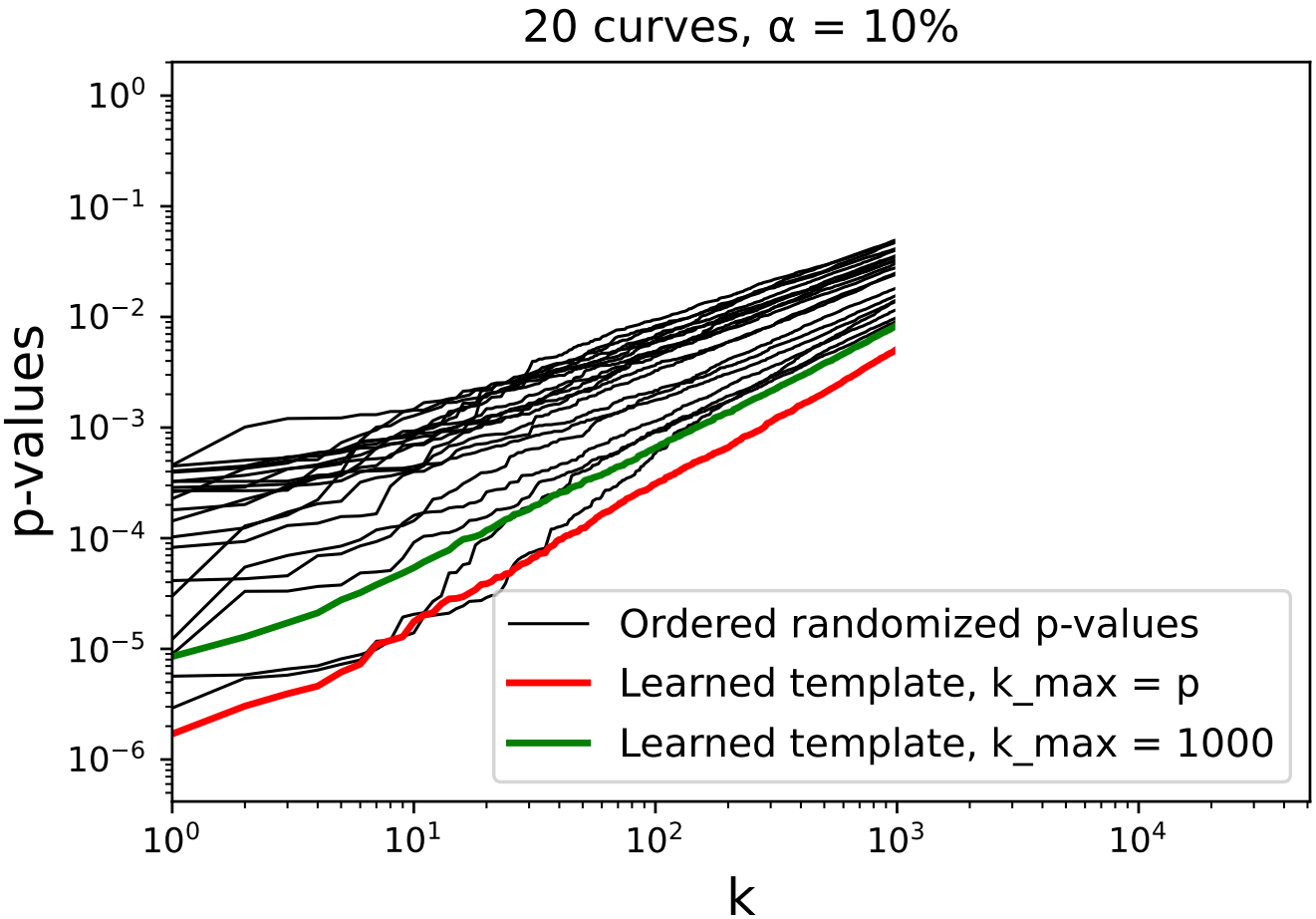
$$\mathbb{P}\left(\exists k \in \{1, \ldots, m_0\} : p_{(k:\mathcal{H}_0)} < t_k\right) \leq \alpha \qquad \text{(JER control)}$$

$\Leftrightarrow$ t_k crosses ≤ 2 curves ($= \alpha\%$ of 20 curves)



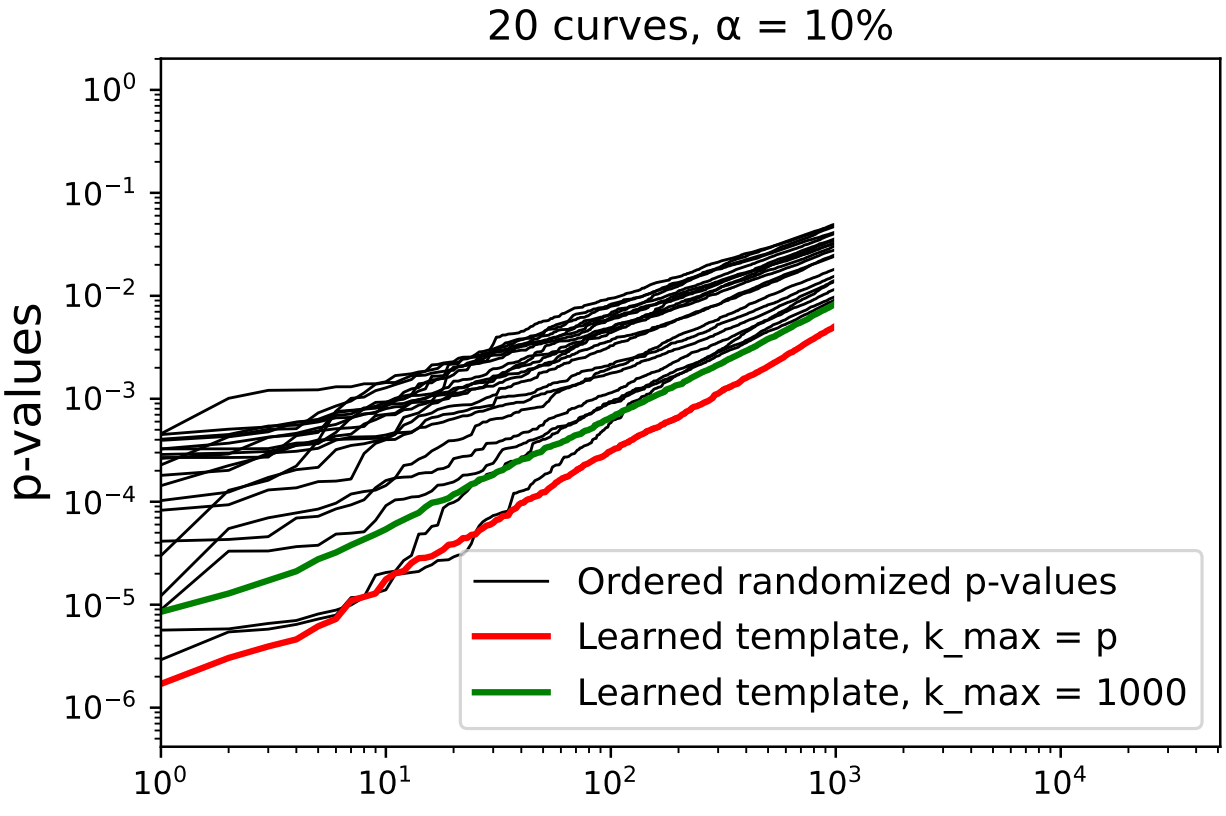
$$\mathbb{P}\left(\exists k \in \{1, \ldots, k_{max} \wedge m_0\} : p_{(k:\mathcal{H}_0)} < t_k\right) \leq \alpha \qquad \textbf{(JER co}$$

$\Leftrightarrow$ t_k crosses ≤ 2 curves before k_{max} ($= \alpha\%$ of 20 curves)



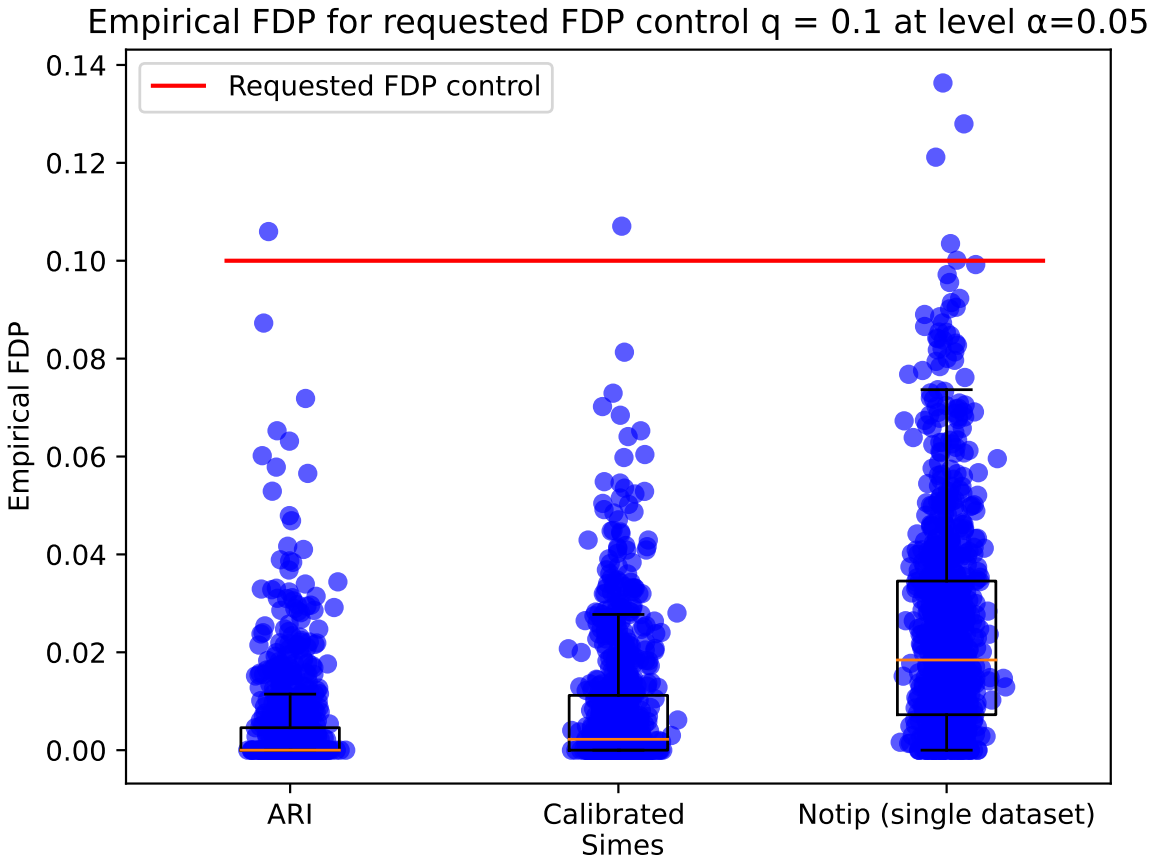
$$\min_{1 \leq k \leq k_{max} \wedge |S|} \left\{ \sum_{i \in S} 1\{p_i(X) \geq t_k\} + k - 1 \right\}$$

k_{max}: bound tradeoff



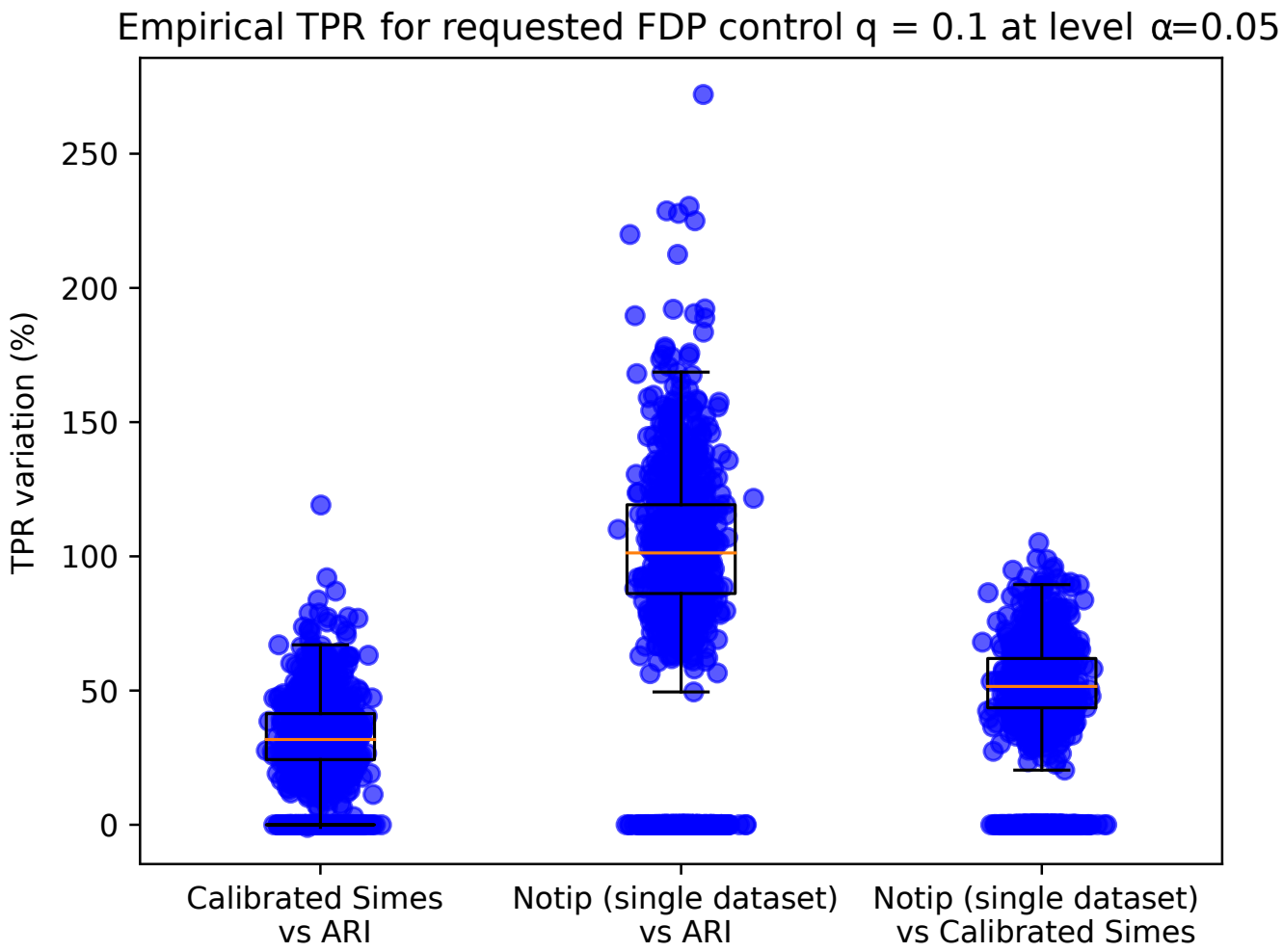
$$\min_{1 \le k \le k_{max} \wedge |S|} \left\{ \sum_{i \in S} 1\{p_i(X) \ge t_k\} + k - 1 \right\}$$

Simulated data



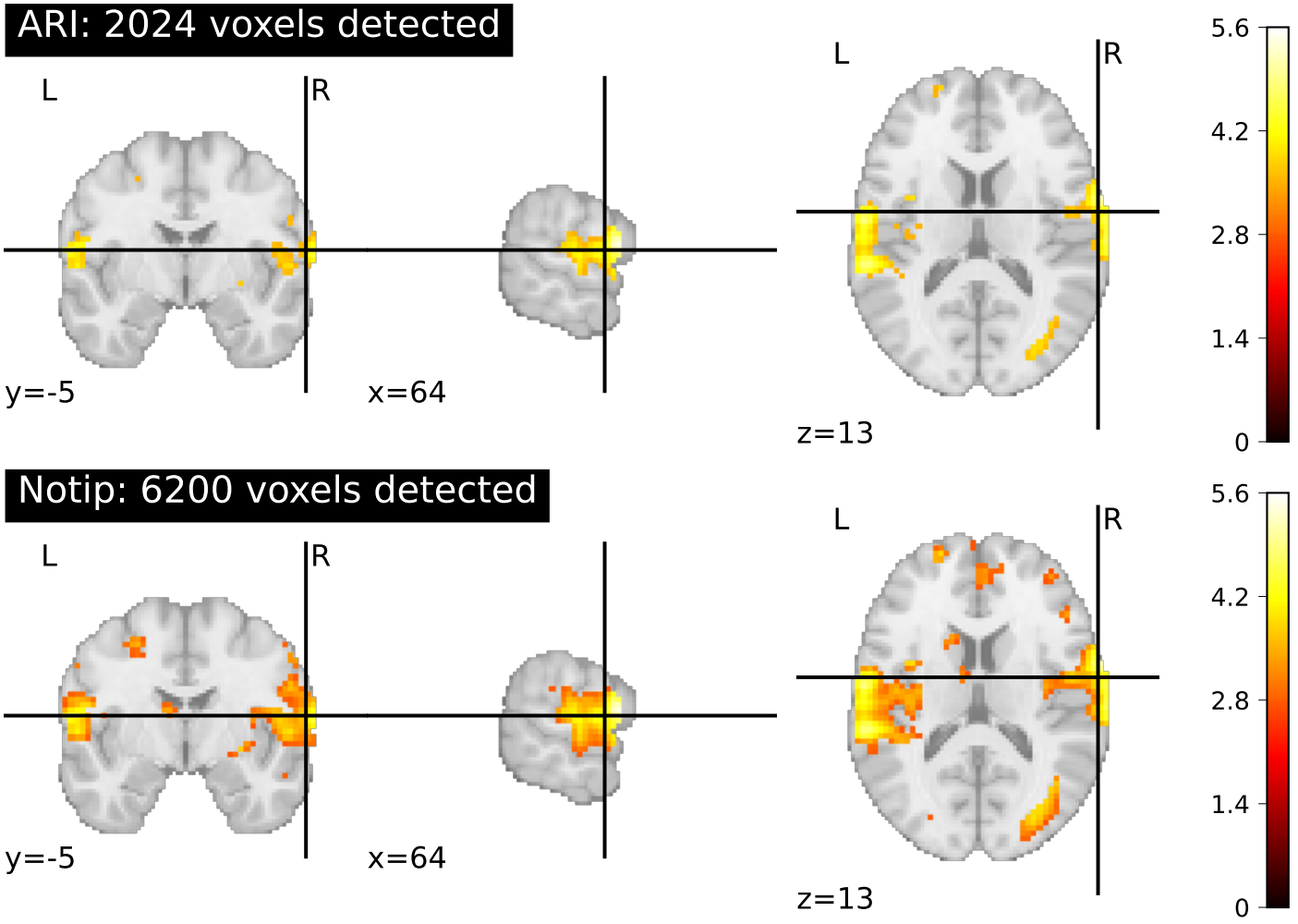
↪ FDP control preserved on simulated data

Simulated data



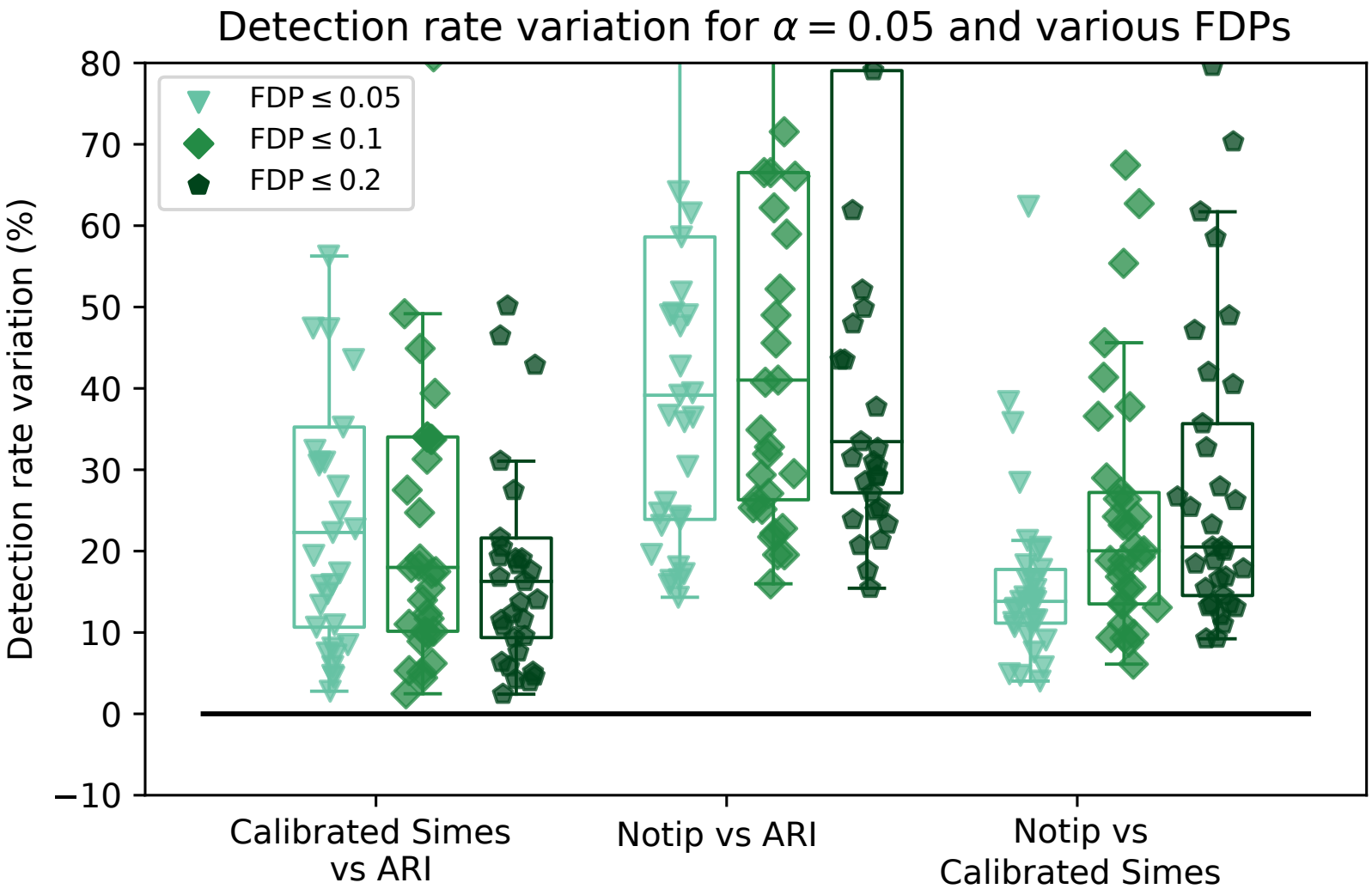
↪ Notip exhibits superior power on simulated data

Power gain using Notip



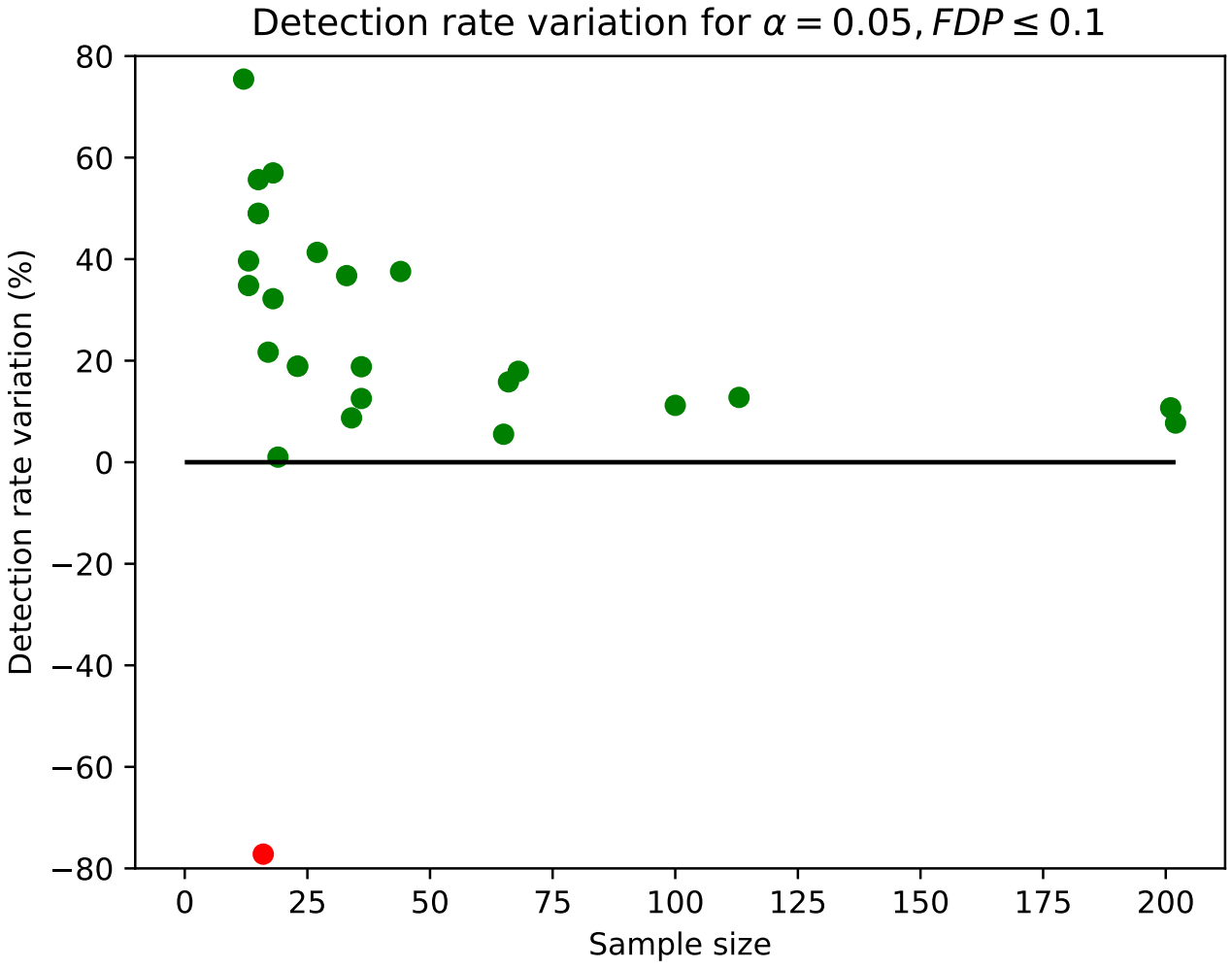
On this example : $\simeq 50$ subjects and $\simeq 52000$ voxels. FDP $< 10\%$

Power gain using Notip



↪ Detection rate comparison on 36 fMRI datasets

Robustness w.r.t low sample size



↪ $n_{subjects}$ varying from 8 to 200

Conclusion and perspective

- Relevance of multiple testing for brain imaging
- Need for post-hoc bounds (ARI and Notip)
- Learning templates to gain power ($\simeq 40\%$ compared to ARI, $\simeq 20\%$ compared to calibrated Simes)
- Avoid using extremely low sample sizes (instability)
- Perspective: beyond one-sample and two-sample designs? See [Davenport et al., arXiv:2208.13724]

Python code available at: `https://github.com/alexblnn/Notip`
Details and more in [Blain Thirion Neuvial, NeuroImage 2022]
at `https://doi.org/10.1016/j.neuroimage.2022.119492`

Additional material

Simes inequality:

$$\mathbb{P}\left(\exists k \in \{1, \ldots, m_0\} : p_{(k:\mathcal{H}_0)} < \alpha k/m\right) \leq \alpha$$

Interpolation or closed testing lead to valid post hoc FDP upper bounds:

$$V^{ARI}(S) = \min_{1 \leq k \leq |S|} \left\{ \sum_{i \in S} 1\left\{ p_i(X) \geq \frac{\alpha k}{m} \right\} + k - 1 \right\}$$

See [Blanchard et al., Annals of Stats 2020] [Rosenblatt et al., NeuroImage 2018]

Problem: Smooth data

$\hookrightarrow$ **conservativeness** of Simes inequality

Simes inequality:

$$\mathbb{P}\left(\exists k \in \{1, \dots, m_0\} : p_{(k:\mathcal{H}_0)} < \alpha k/m\right) \leq \alpha$$

$p_{(k:\mathcal{H}_0)}$ is **unknown**!

$\hookrightarrow$ **Randomization** to estimate $p_{(k:\mathcal{H}_0)}$

See [Hemerik et al., Biometrika 2019], [Blanchard et al., Annals of Stats 2020]

If the law of X under the null is invariant by sign-flipping [Romano Wolf, 2005] we denote the empirical JER using B sign-flips:

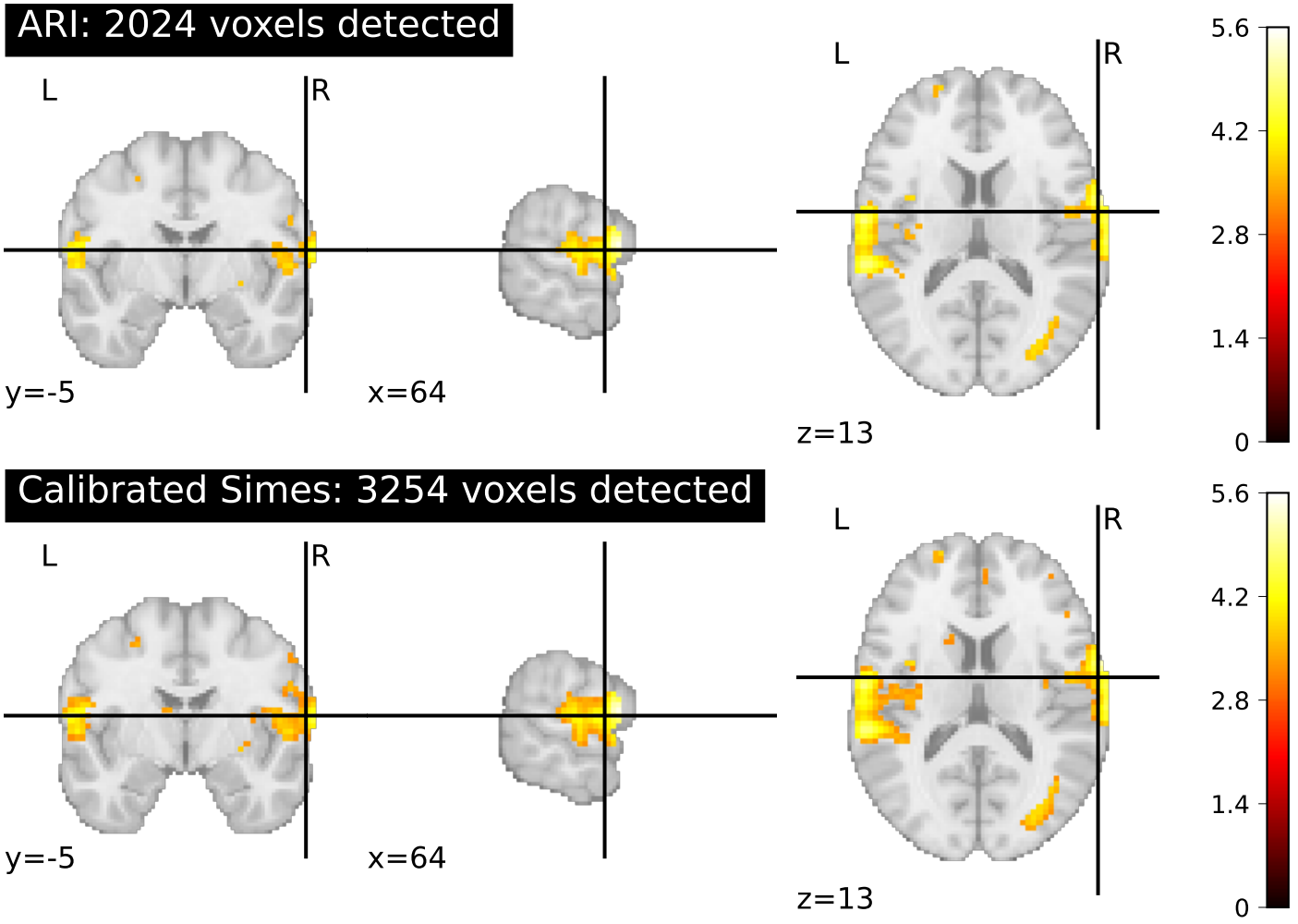
$$\widehat{JER}(\lambda) = \frac{1}{B}\sum_{b=1}^{B} 1\left\{\exists k \in \{1,\ldots,k_{max} \wedge m_0\} : p^b_{(k:\mathcal{H}_0)}(X) < t_k(\lambda)\right\}$$

Then we have **Theorem 4.8** of [Blanchard et al., 2020]:

$$\lambda(\alpha) = \max\left\{\lambda \geq 0 : \widehat{JER}(\lambda) \leq \alpha\right\}$$

is a valid λ-calibration; i.e., $t_k(\lambda(\alpha))$ **controls the JER at level** α

Power gain by calibration

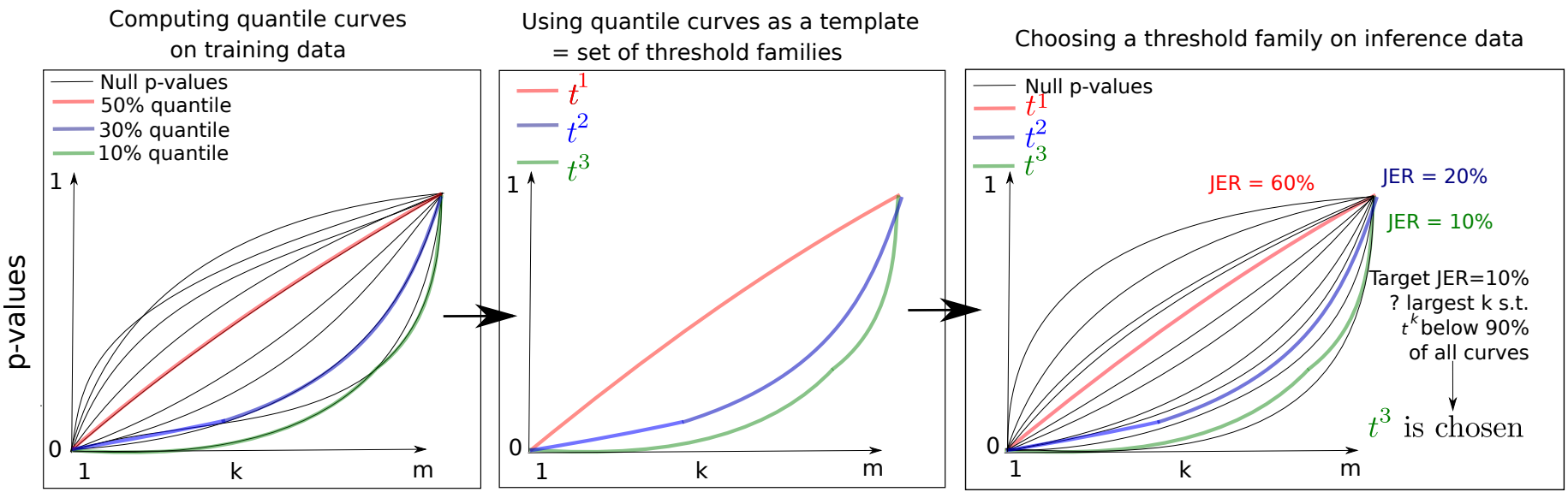


On this example : $\simeq 50$ subjects and $\simeq 52000$ voxels. FDP $< 10\%$

Can another t_k outperform the Simes template?

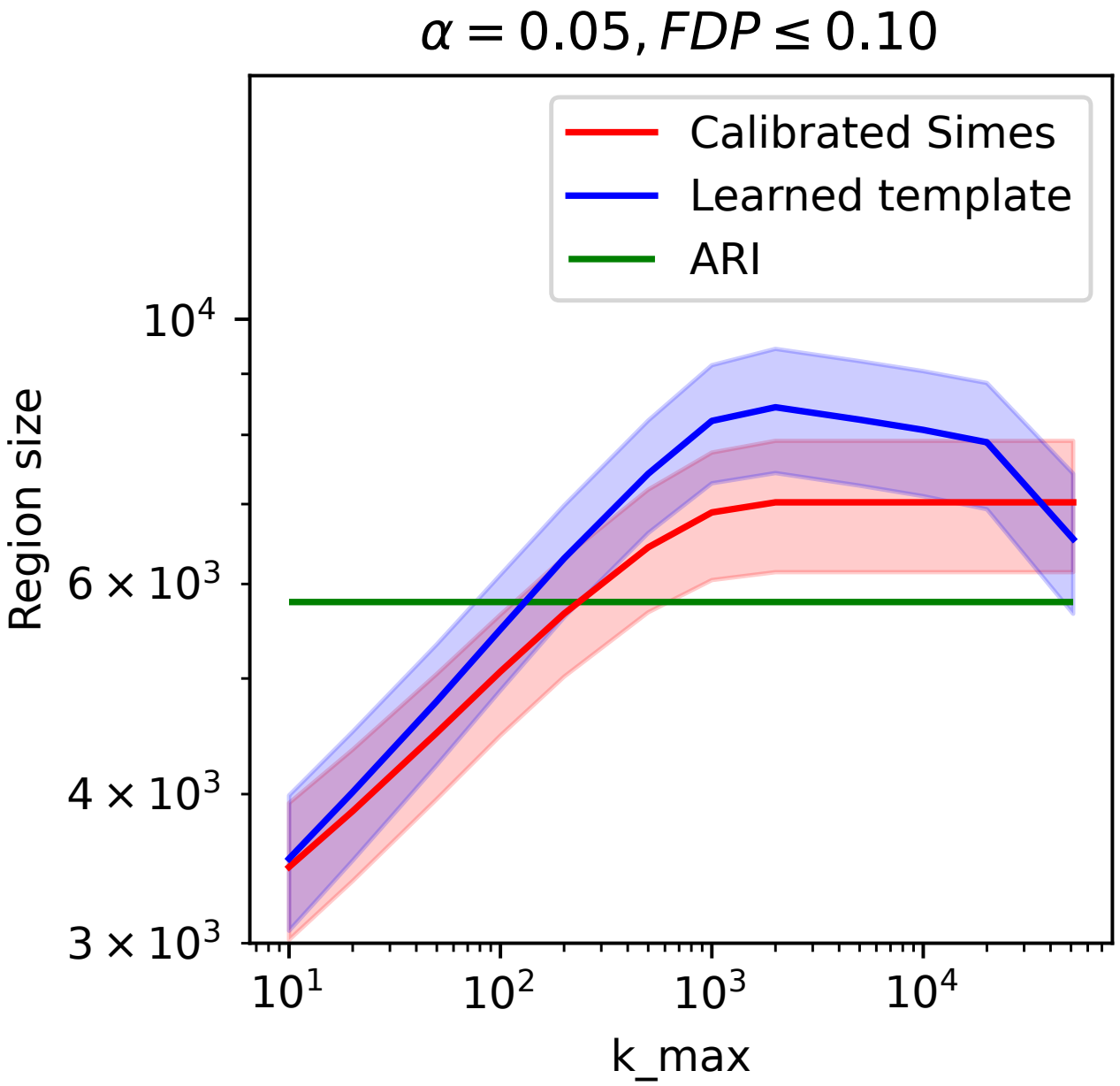
Challenging question; [Andreella et al., arXiv:2012.00368]

Idea: **learning** templates from the data

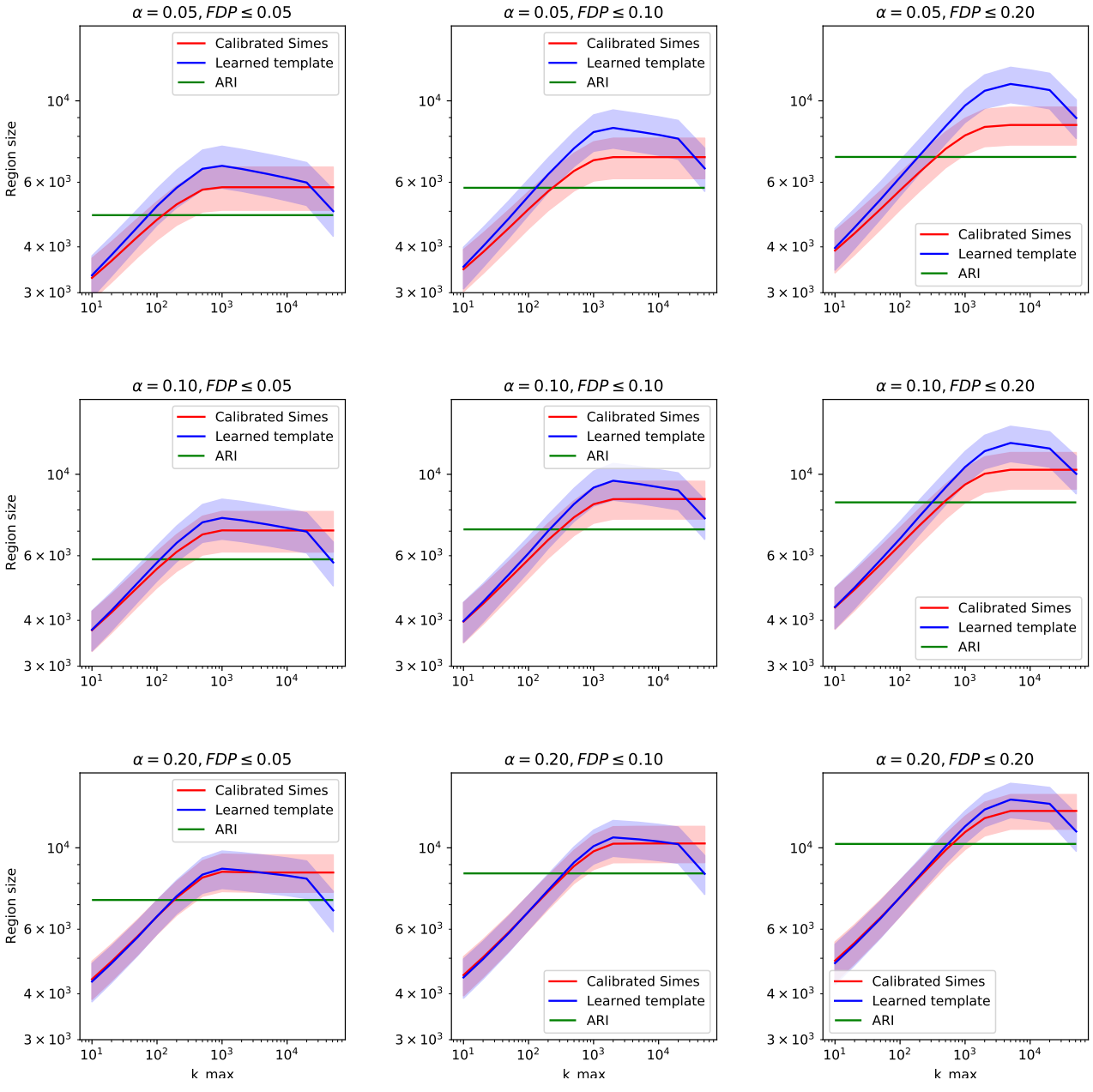


Similar idea: [Meinshausen, Scandinavian Journal of Stats 2006]

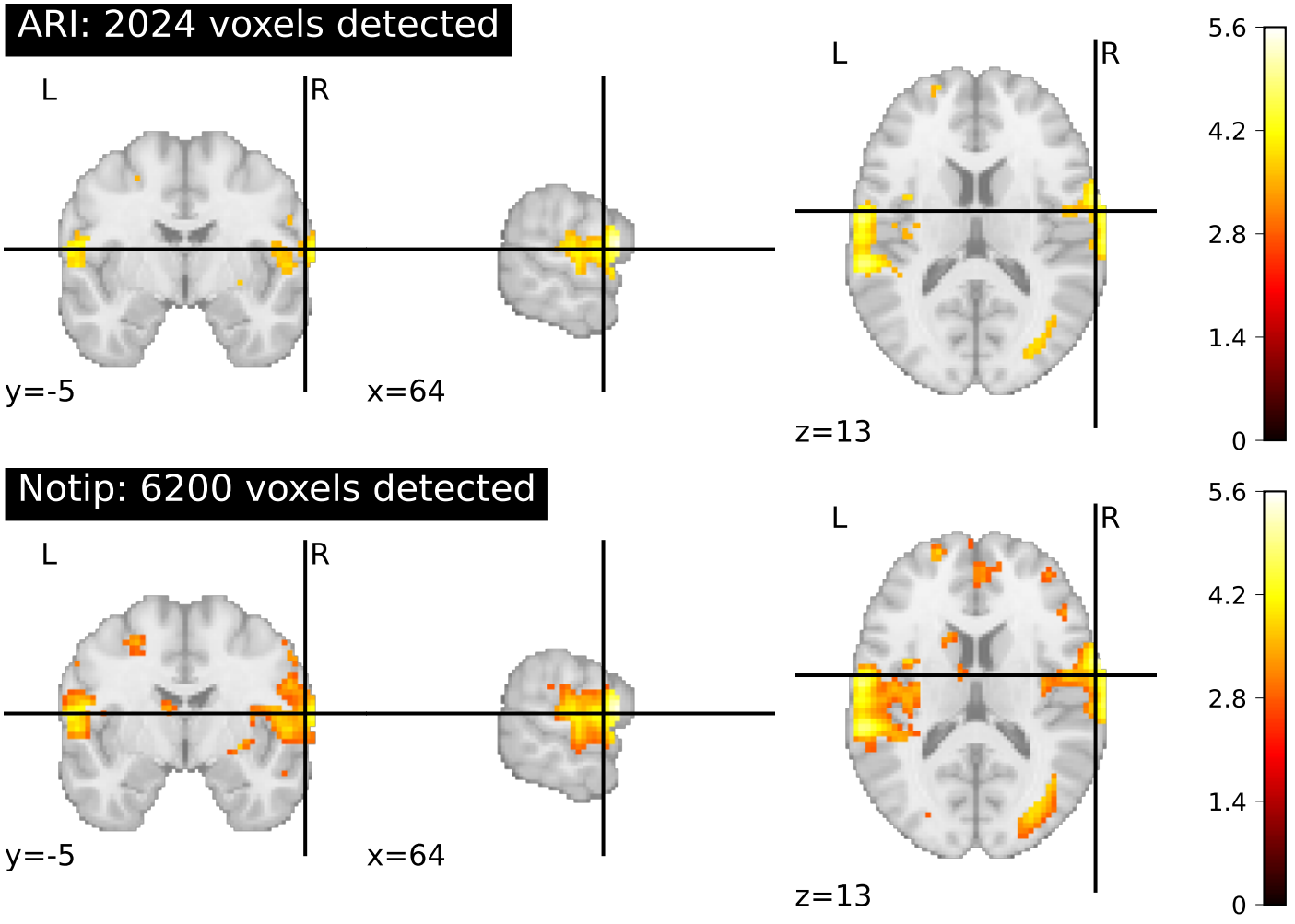
k_{max}: sensitivity study



k_{max}: sensitivity study

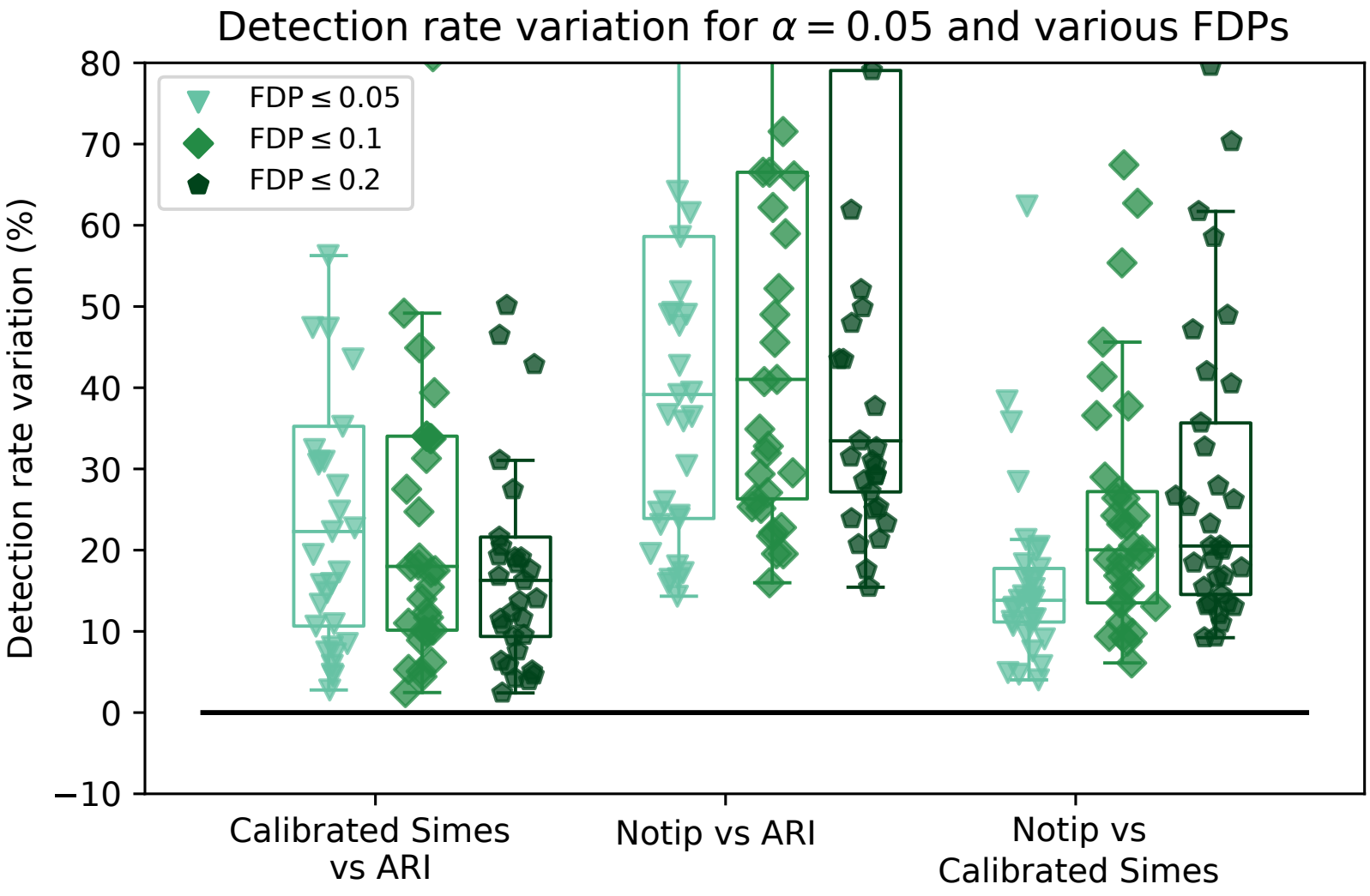


Power gain by template learning



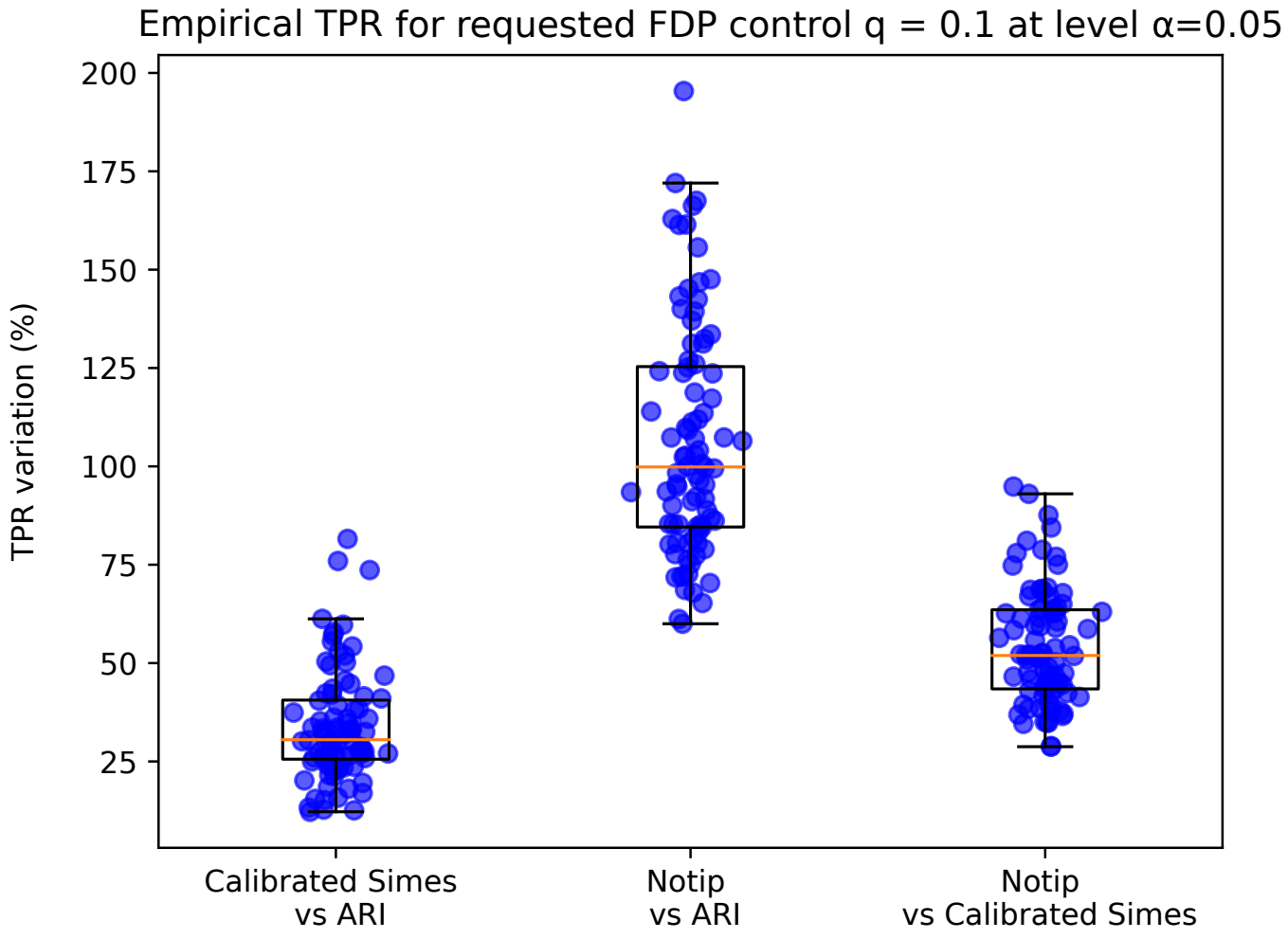
On this example : $\simeq 50$ subjects and $\simeq 52000$ voxels. FDP $< 10\%$

Power gain by template learning



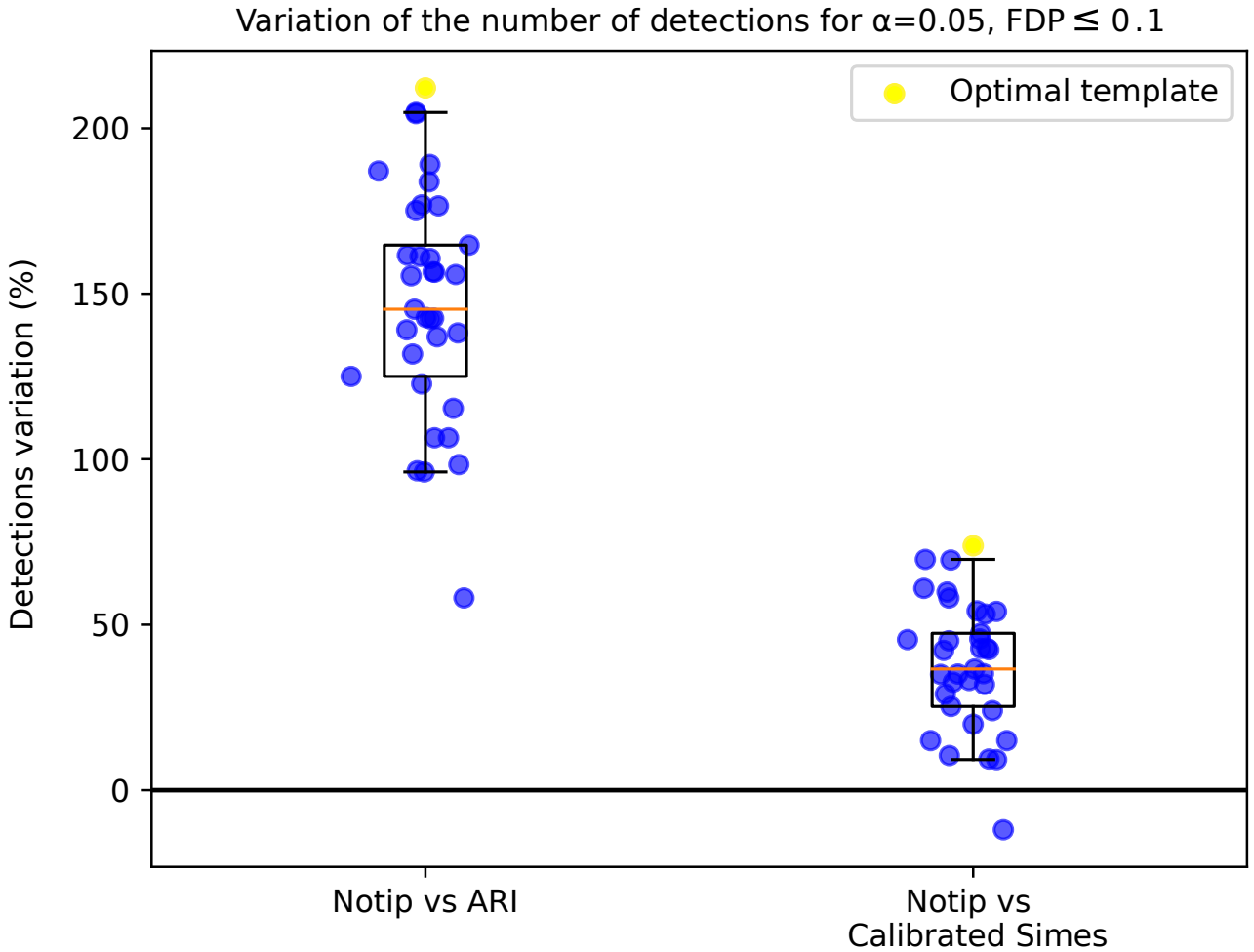
↪ Detection rate comparison on 36 fMRI datasets

Power gain by template learning



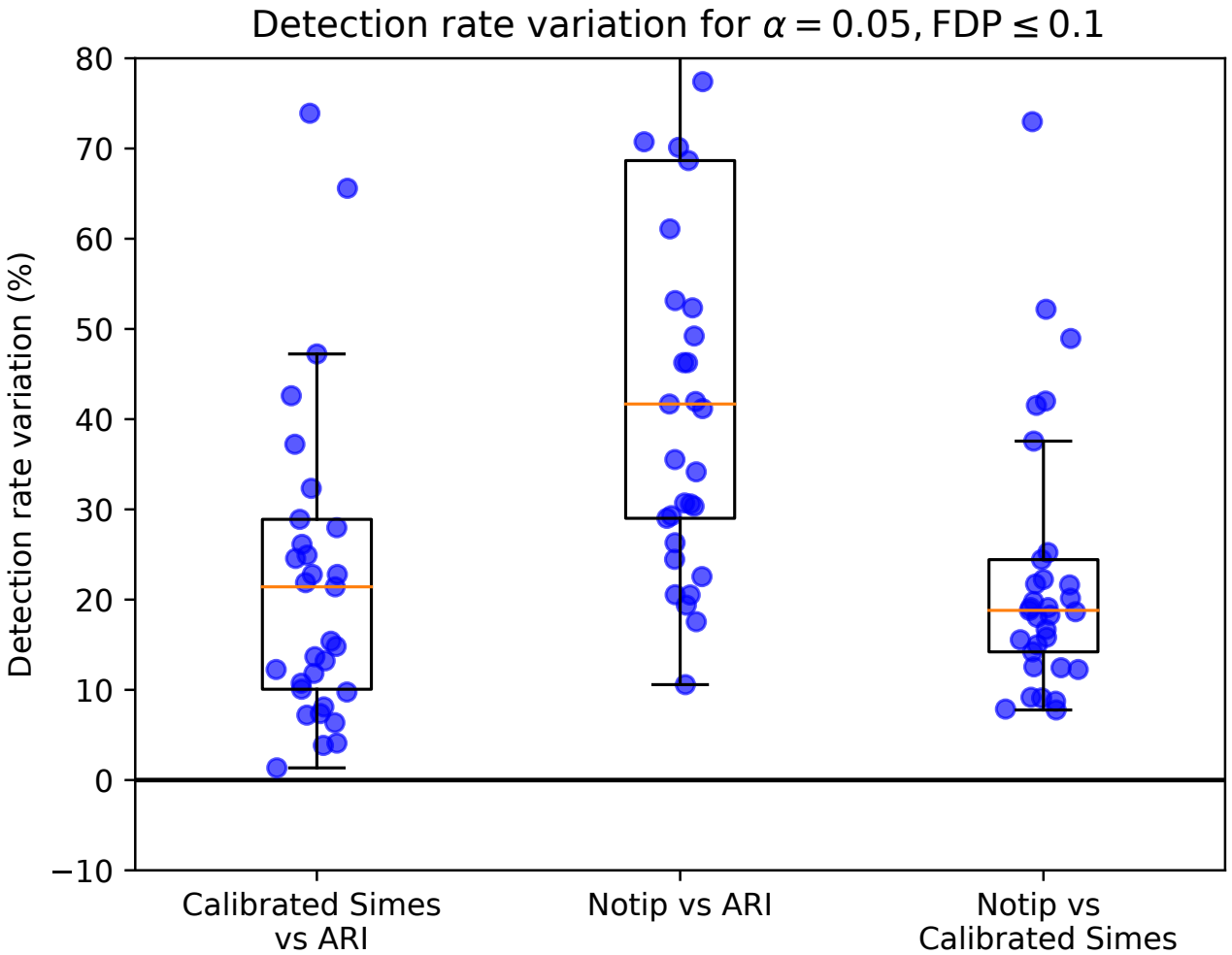
↪ **Detection rate comparison** on simulated data

Template learning: Sensitivity to training set choice



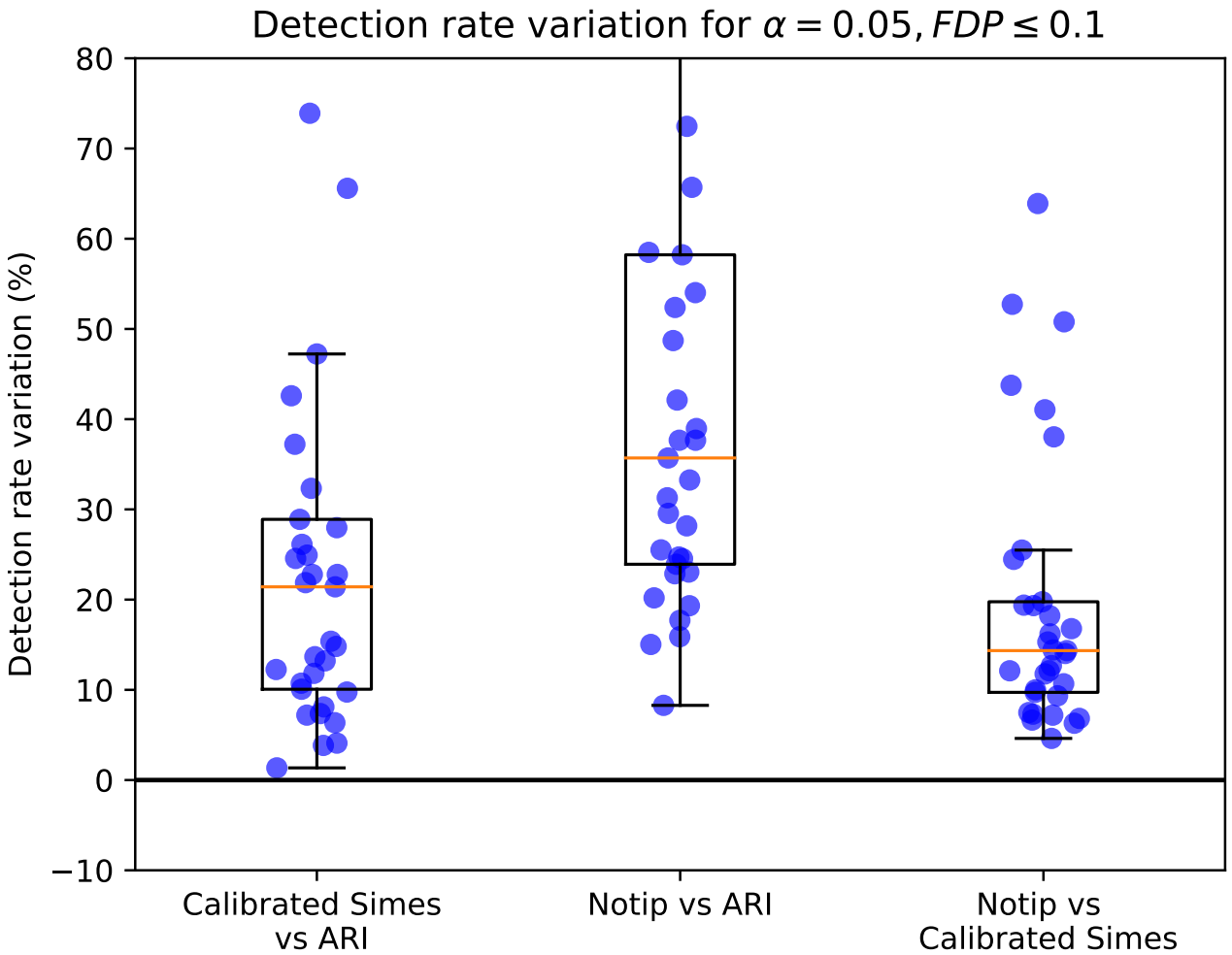
↪ **35 different** training sets

Template learning: low training sample size



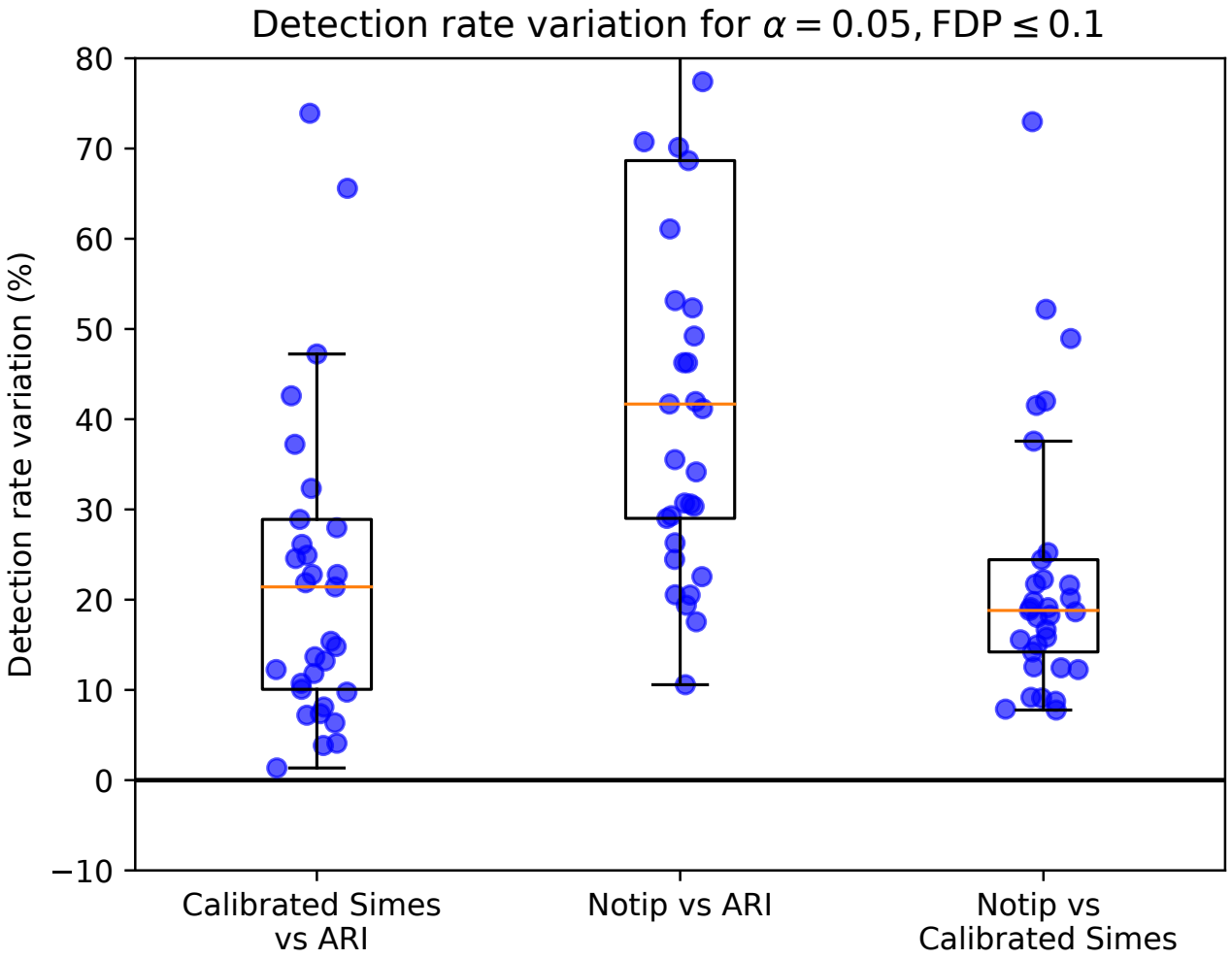
$\hookrightarrow$ $n_{train} = 113$ subjects

Template learning: low training sample size



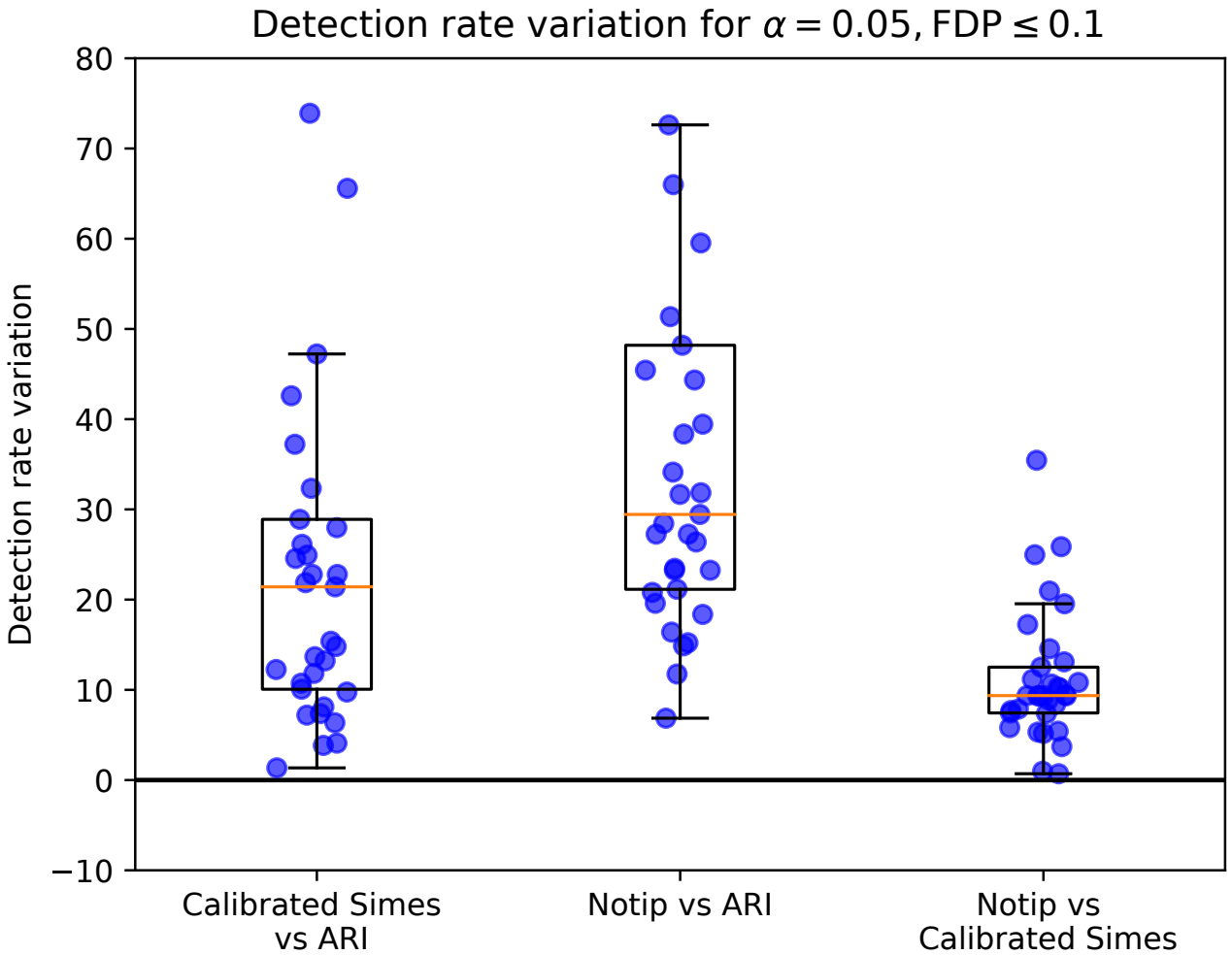
$\hookrightarrow$ $n_{train} = 20$ subjects

Template learning: influence of smoothing



$\hookrightarrow$ Training: $FWHM = 4$mm, inference: $FWHM = 4mm$

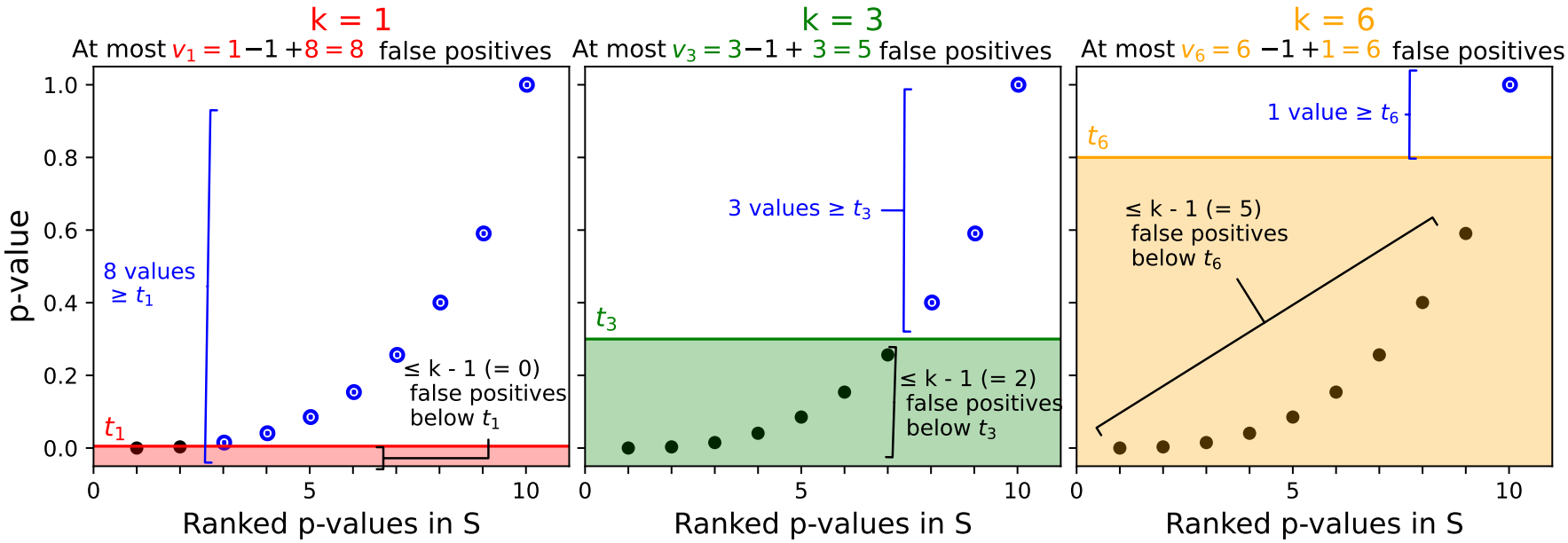
Template learning: influence of smoothing

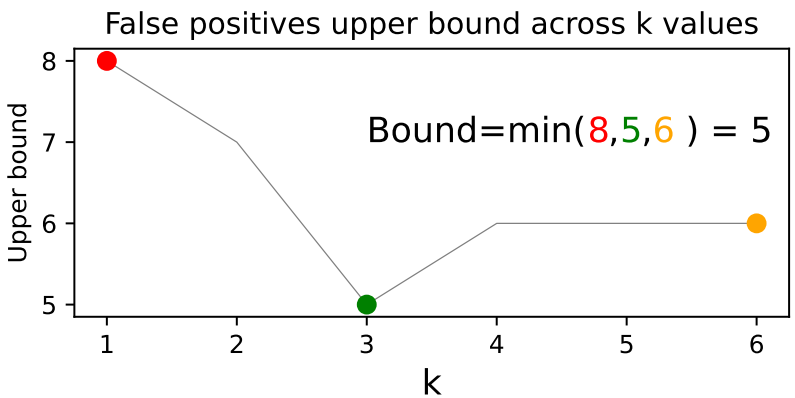


↪ Training: $FWHM = 4$mm, inference: $FWHM = 8mm$

Appendix : Bound computation

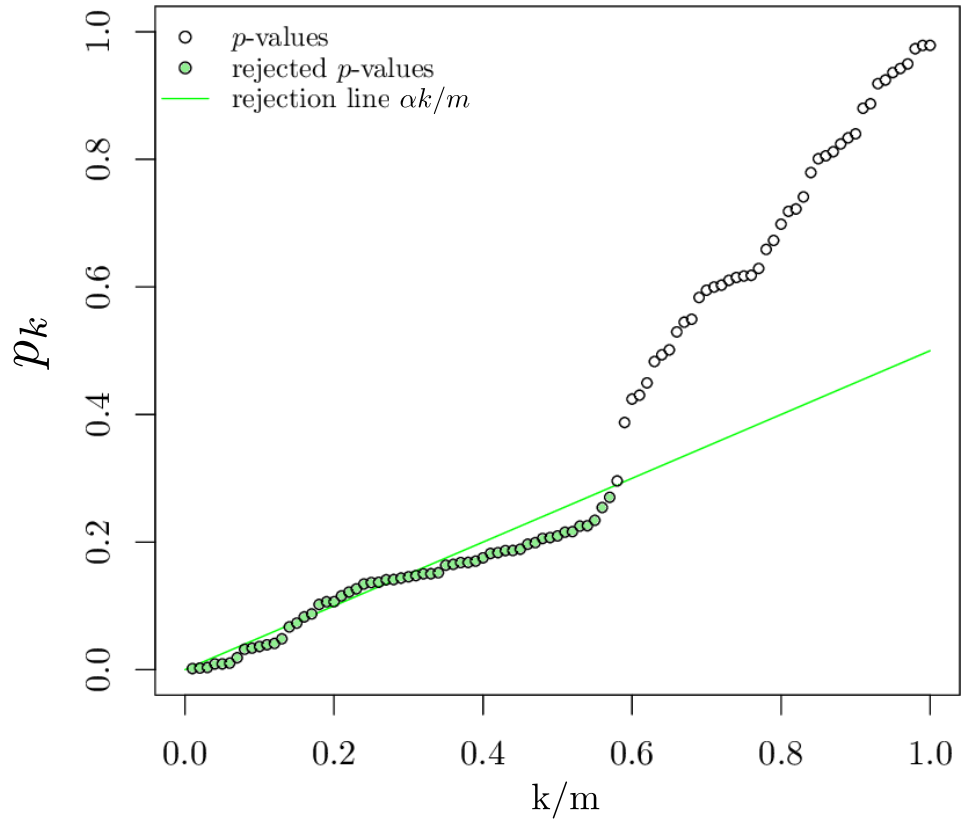
$$V(S) = \min_{1 \leq k \leq |S|} \left\{ \sum_{i \in S} \mathbf{1}\left\{ p_i(X) \geq t_k \right\} + k - 1 \right\} \tag{1}$$





Appendix: Benjamini-Hochberg (BH) procedure

Sorted vector $p_{(k:m)}$, find largest k s.t. $p_{(k:m)} \leq \alpha k/m$



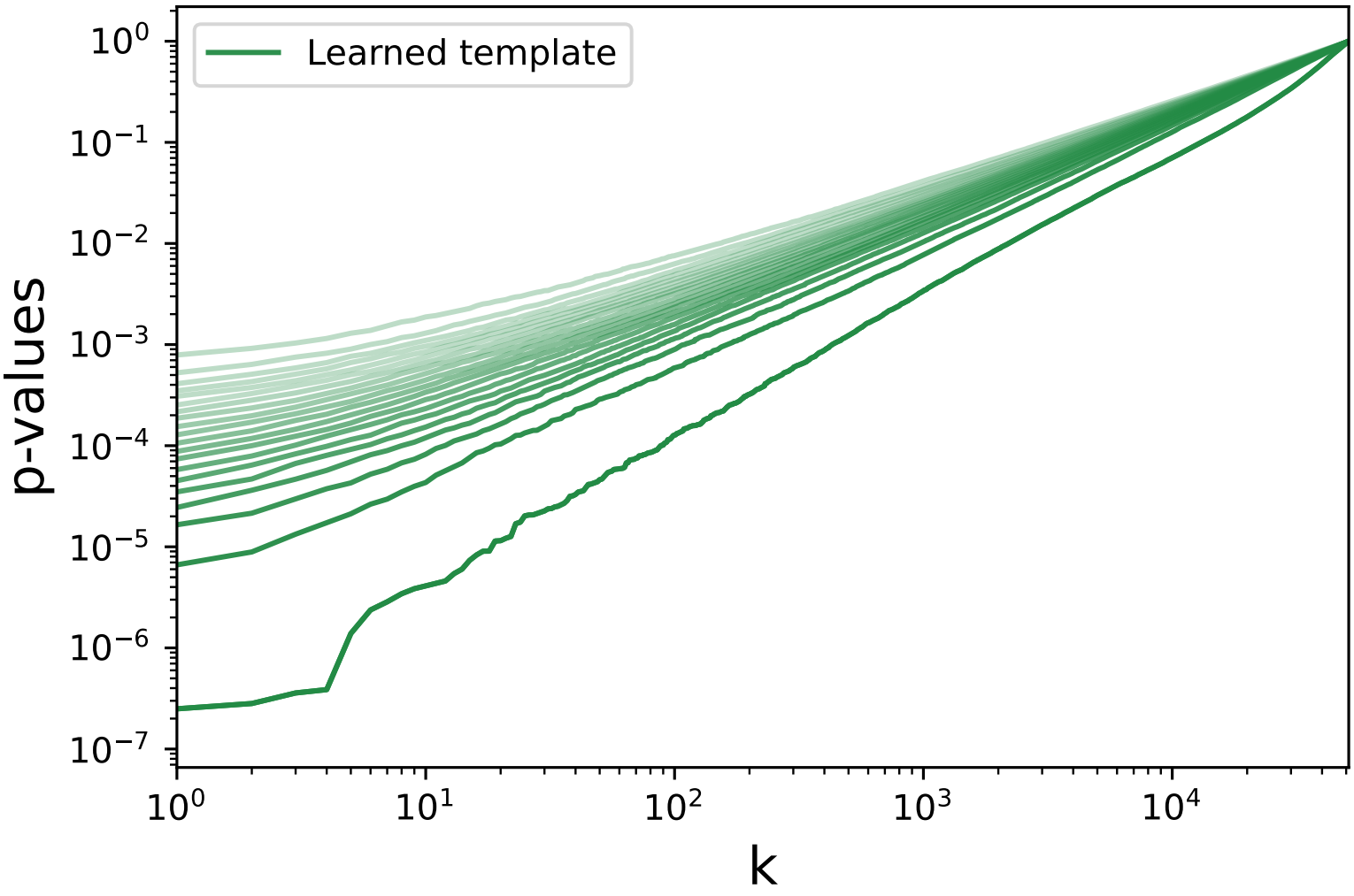
False Discovery Proportion (FDP): $\mathrm{FDP}(S) = \frac{|S \cap \mathcal{H}_0|}{|S|}$

BH: $\mathrm{FDR}(S) = \mathbb{E}[\mathrm{FDP}(S)] = \mathbb{E}\left[\frac{|S \cap \mathcal{H}_0|}{|S|}\right] \leq \alpha$

Algorithm Computing one-sample randomized p-values

function get_randomized_p_values(X, B)
    $n, p \leftarrow$ shape(X)
    pval0 $\leftarrow$ zeros(B, p)
    For $b \in [1, B]$:
        flip $\leftarrow$ diag(draw_random_vector($\{-1, 1\}^n$))
                                        ▷ matrix of shape (n, n)
        $X_{flipped} = flip\ X$
        pval0[b] $\leftarrow$ one_sample_t_test($X_{flipped}$, 0)
                                        ▷ 0 = null hypothesis
    end For
return pval0
end function

Appendix: learned template



$V_k(S) \geq k - 1$, therefore if $k > q|S|$ then $V_k(S)/|S| \geq q$ for any S. These values of k are useless for obtaining a FDP bound less than q. This motivates a choice of k_{max} of the form

$$k_{max} = q_{max}|S_{max}|\,, \tag{2}$$

where q_{max} is the maximum proportion of false positives that can be tolerated by users and $|S_{max}|$ is the size of the largest set of voxels of interest. We set $q_{max} = 0.5$.
Most contrasts studied in the literature lead to less of 5% of the image domain to be declared active. We set $|S_{max}| = 0.05m$.
Then $k_{max} = 0.5 * 0.05m = 0.025m$. If $m \simeq 50,000$, we settle on using $k_{max} = 0.02m = 1,000$.

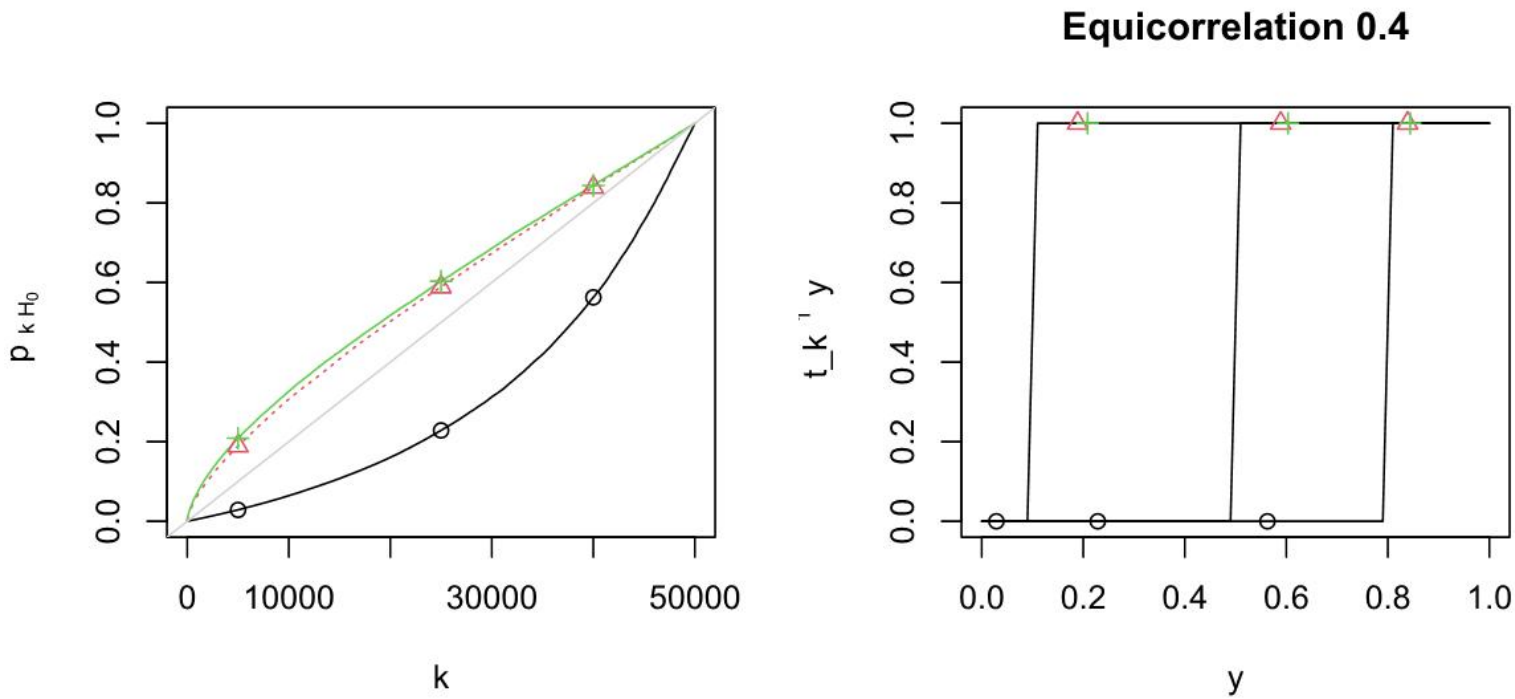


Figure: Template Beta, equicorrelation

Appendix : JER Budget of Simes template