Boosting Statistical Anomaly Detection via *multiple test* with NPLM

Gaia Grosso [1, 2, 3, *], Marco Letizia [4,5], Phil Harris [1, 2]

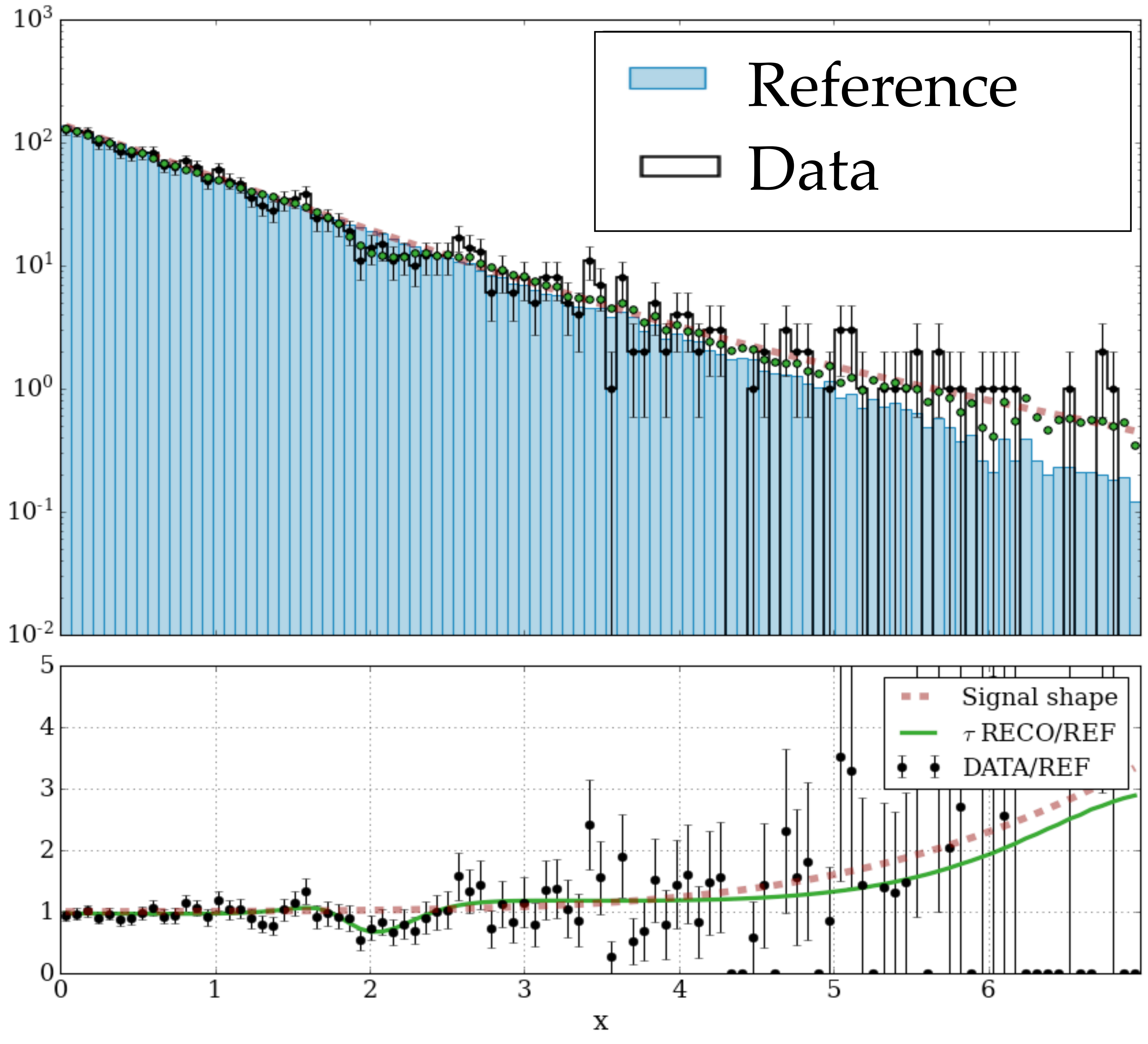
[1] NSF Institute for Artificial Intelligence and Fundamental Interactions (IAIFI)
[2]Massachusetts Inst. of Technology, [3]Harvard University
[4]MaLGa Center - DIBRIS, University di Genova, [5]INFN, Sezione di Genova

*gaia.grosso@cern.ch

March 13th, 2023

ACAT 2024

Statistical anomaly detection as a goodness of fit

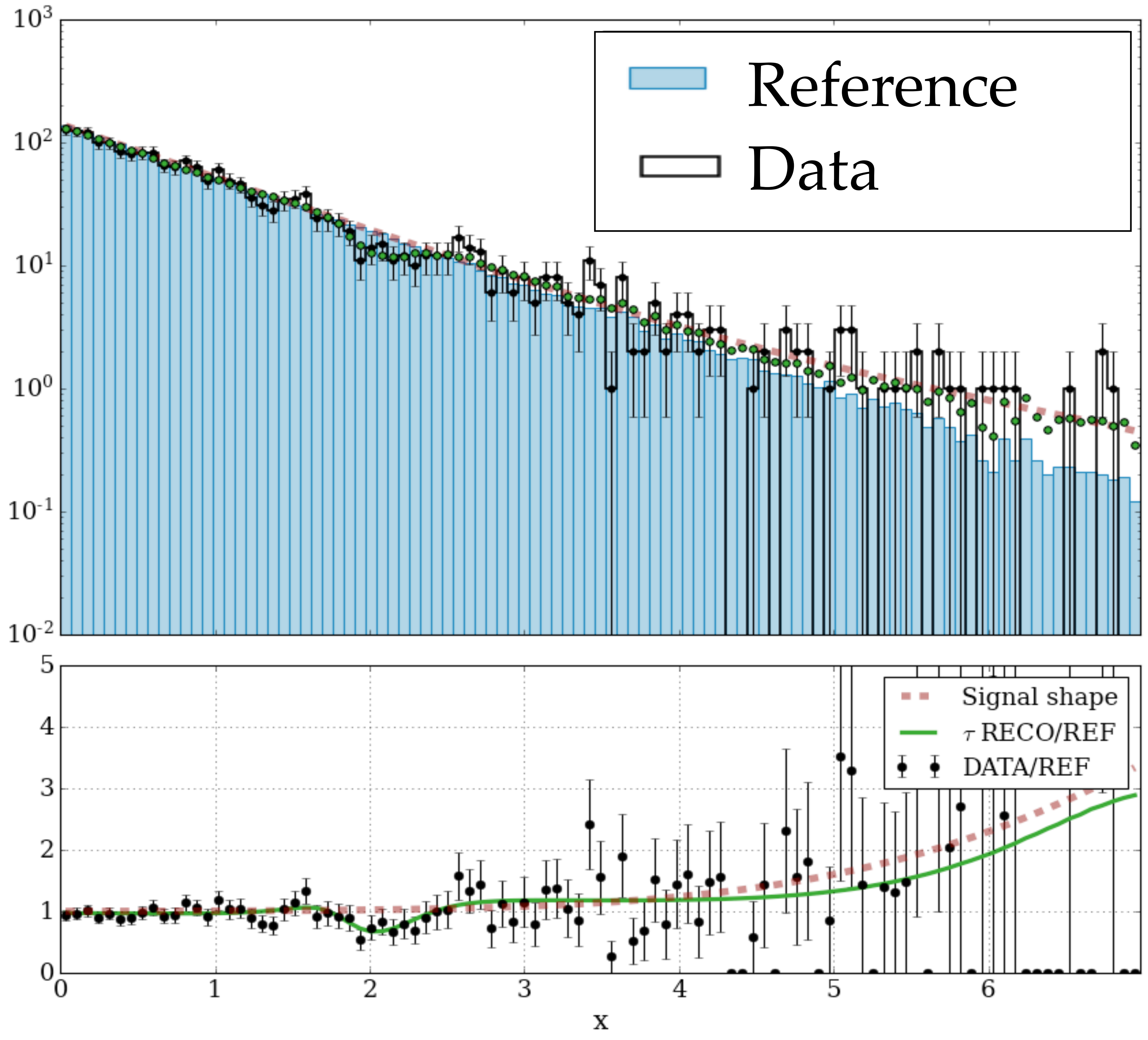


Problems defined by:

- **Data**: experimental measurements of the natural process
- **Reference model**: expected nominal behavior of the data

Is the Reference model a *good* description of the data?

Statistical anomaly detection as a goodness of fit

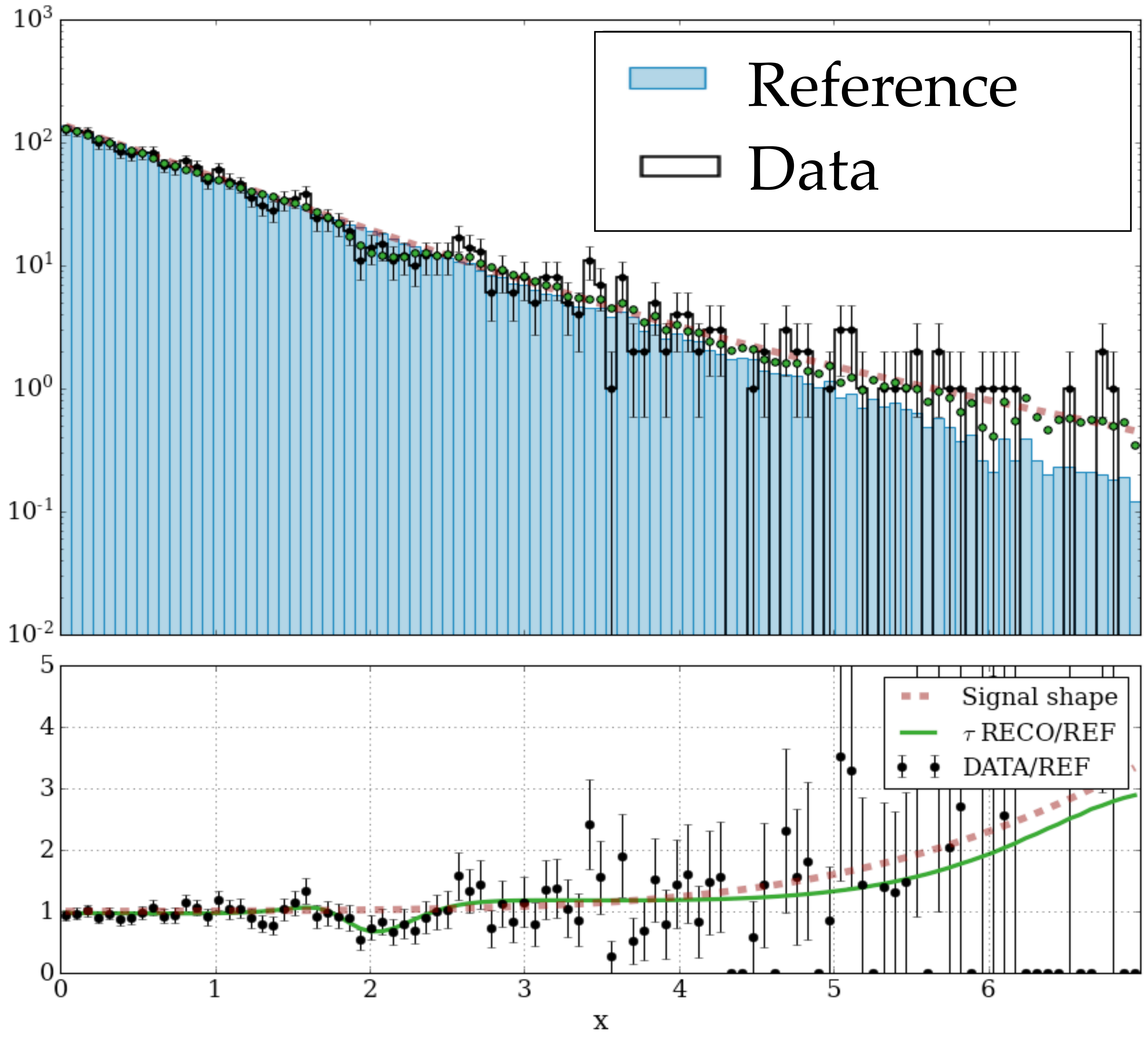


Many use cases:

- Experimental system monitoring (DQM)
- Signal agnostic New Physics searches
- Data validation

→ assessing the *goodness* of **Generative models**

Statistical anomaly detection as a goodness of fit



Challenges:

- Anomalies are *rare*!
- Anomalies are *unexpected*!

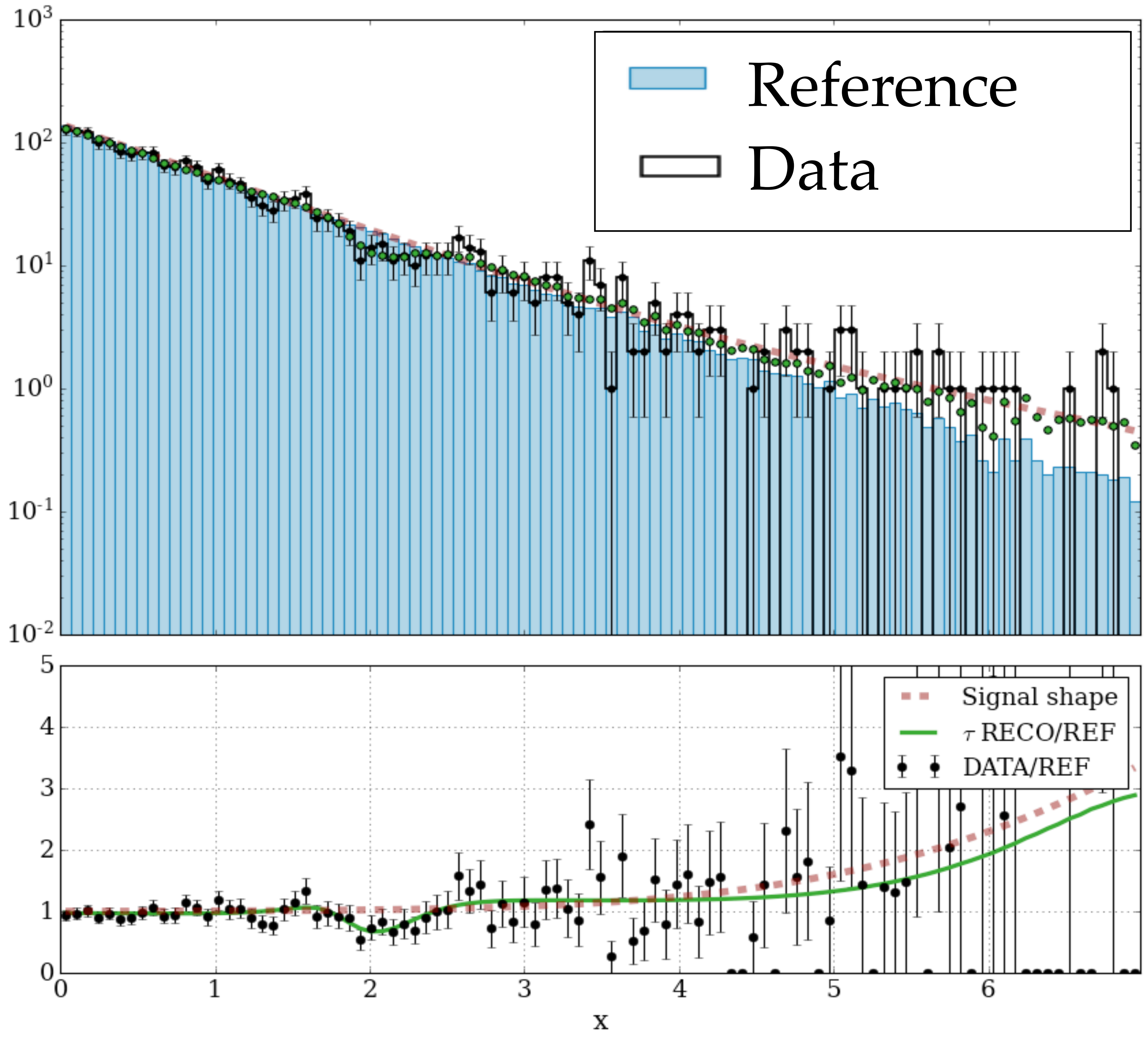
→ *large statistics samples* to reach sensitivity

→ *high dimensional raw data* for inclusive test

→ the ideal statistical model of the data (aka *inductive bias*) is unknown

Motivation for ML based solutions

Statistical anomaly detection as a goodness of fit



How to exploit large datasets?

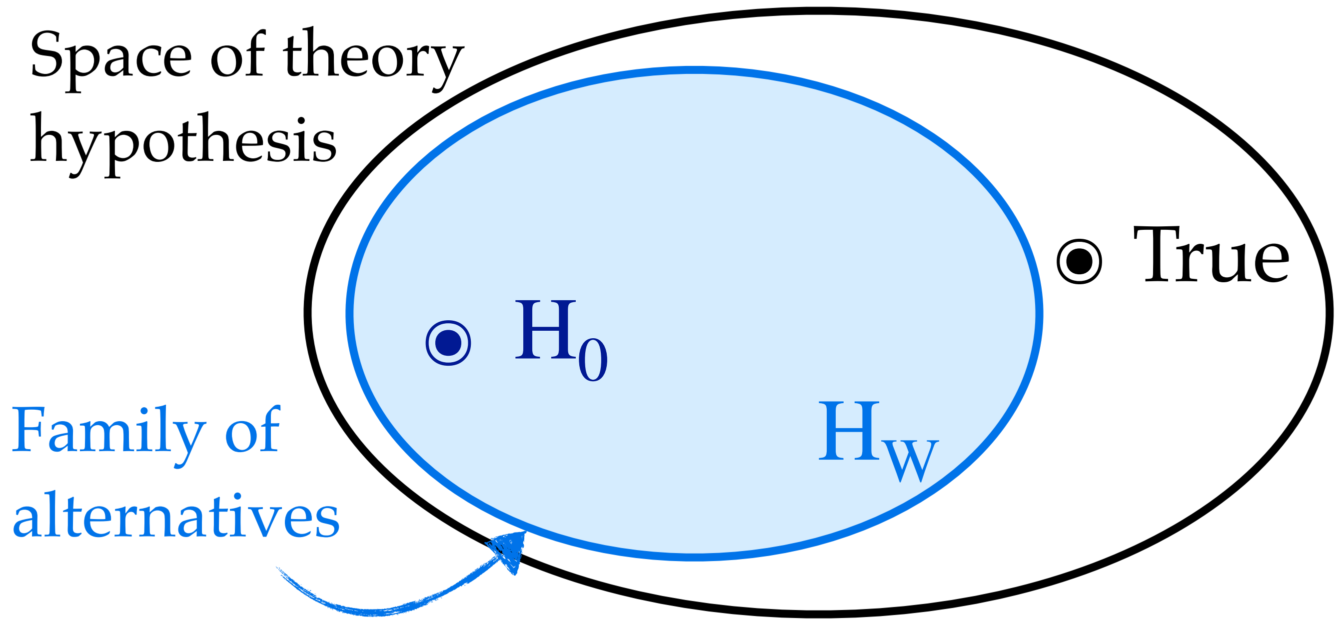
How to work around inductive biases?

→ *multiple testing* in the context of the **Neyman-Pearson** GoF test
[GG, Letizia, Wulzer, Pierini 2305.14137]

ML-based Neyman-Pearson GoF test

Compare the Reference hypothesis with an *alternative*.
Inductive bias: definition of the family of alternatives

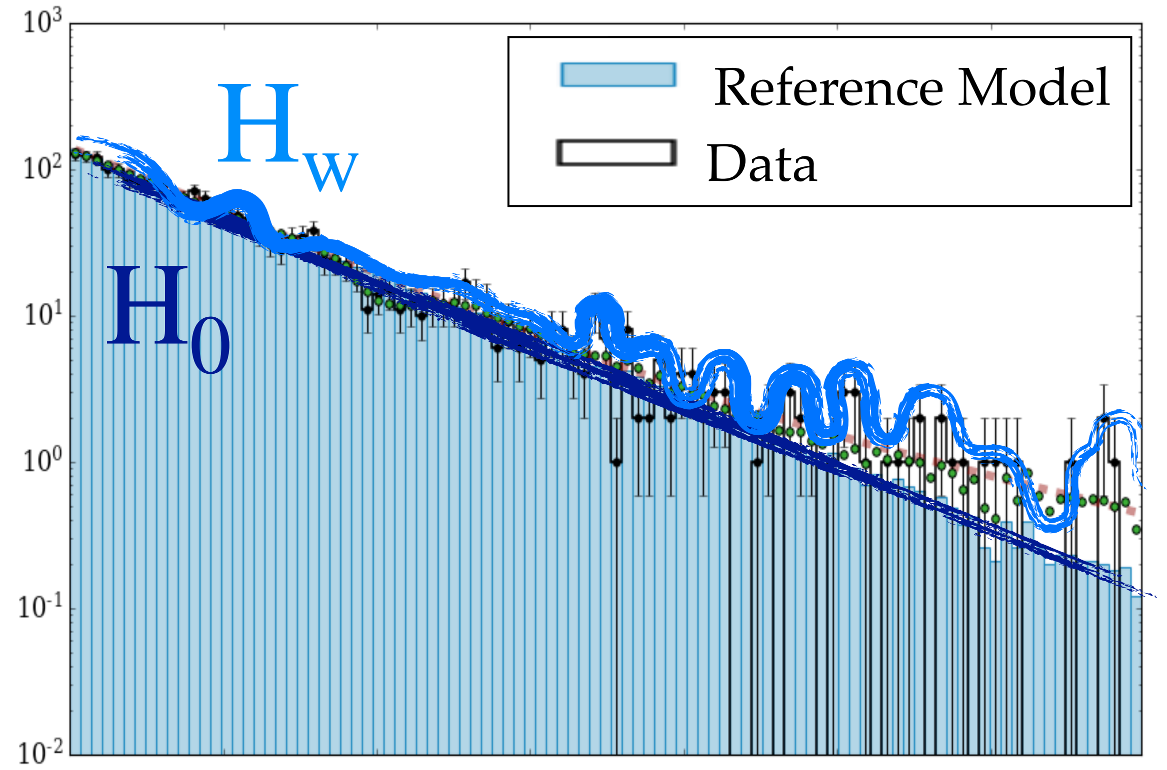
$$t(\mathcal{D}) = \max_{\mathbf{w}} \left[2 \log \frac{\mathcal{L}(\mathcal{D} \,|\mathrm{H}_{\mathbf{w}})}{\mathcal{L}(\mathcal{D} \,|\mathrm{H}_{\mathbf{0}})} \right]$$



New physics Learning Machine (NPLM)

Universal approximator
(NN, kernel methods, …)

$$n(x|\mathrm{H}_{\mathbf{w}}) = e^{f(x;\mathbf{w})}\, n(x|\mathrm{R}_{\mathbf{0}})$$



"Learning New Physics from a Machine" Phys. Rev. D

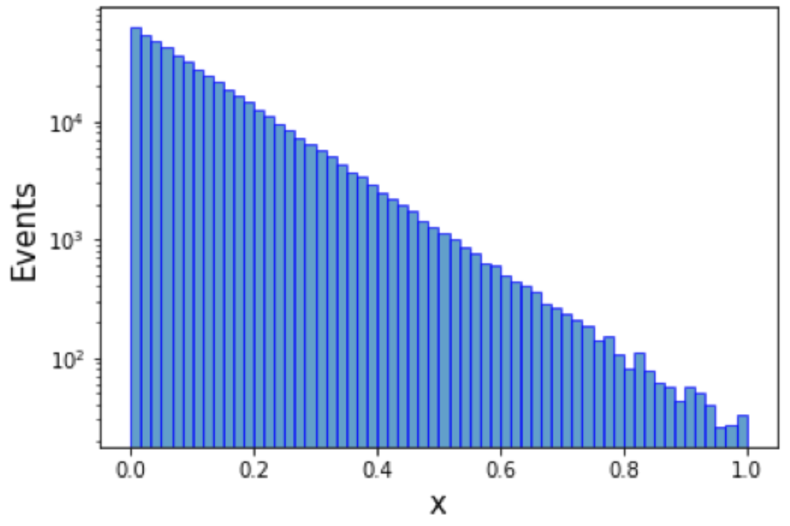
ML-based Neyman-Pearson GoF test

NPLM: implementation

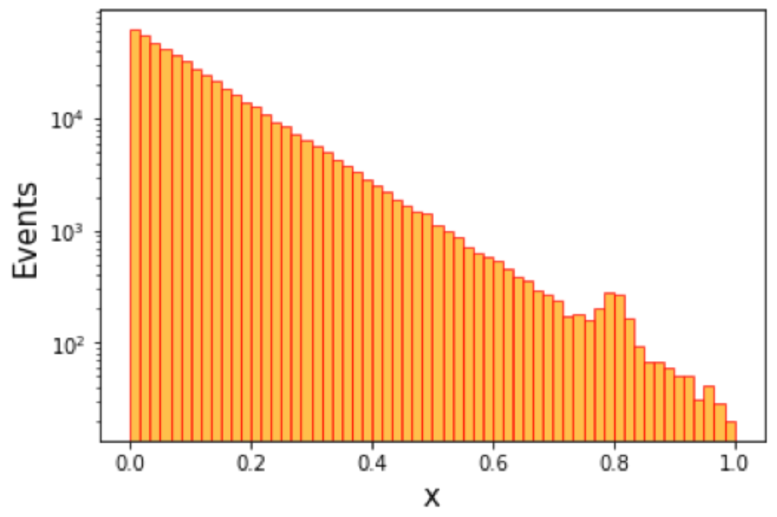
INPUT (n-dimensional)

Reference sample (R)
label=0

(1D example)



Data sample (D)
label=1



MODEL $f(\cdot, \mathbf{w})$

Kernel-based model

$$f_w(x) = \sum_{i=1}^{M} w_i k_\sigma(x, \tilde{x}_i)$$

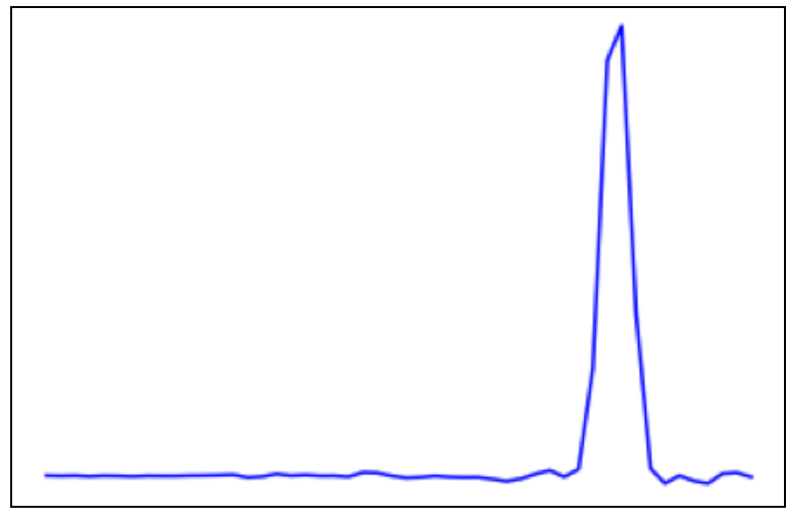
$$k_\sigma(x, x') = e^{-\|x-x'\|^2/2\sigma^2}$$

$\mathbf{w}$ —training→ $\hat{\mathbf{w}}$

OUTPUT

Log-ratio of densities:

$$f(x; \hat{\mathbf{w}}) = \log\left[\frac{n(x\,|\,\mathrm{H}_{\hat{\mathbf{w}}})}{n(x\,|\,\mathrm{R}_0)}\right]$$



NPLM statistic (scalar):

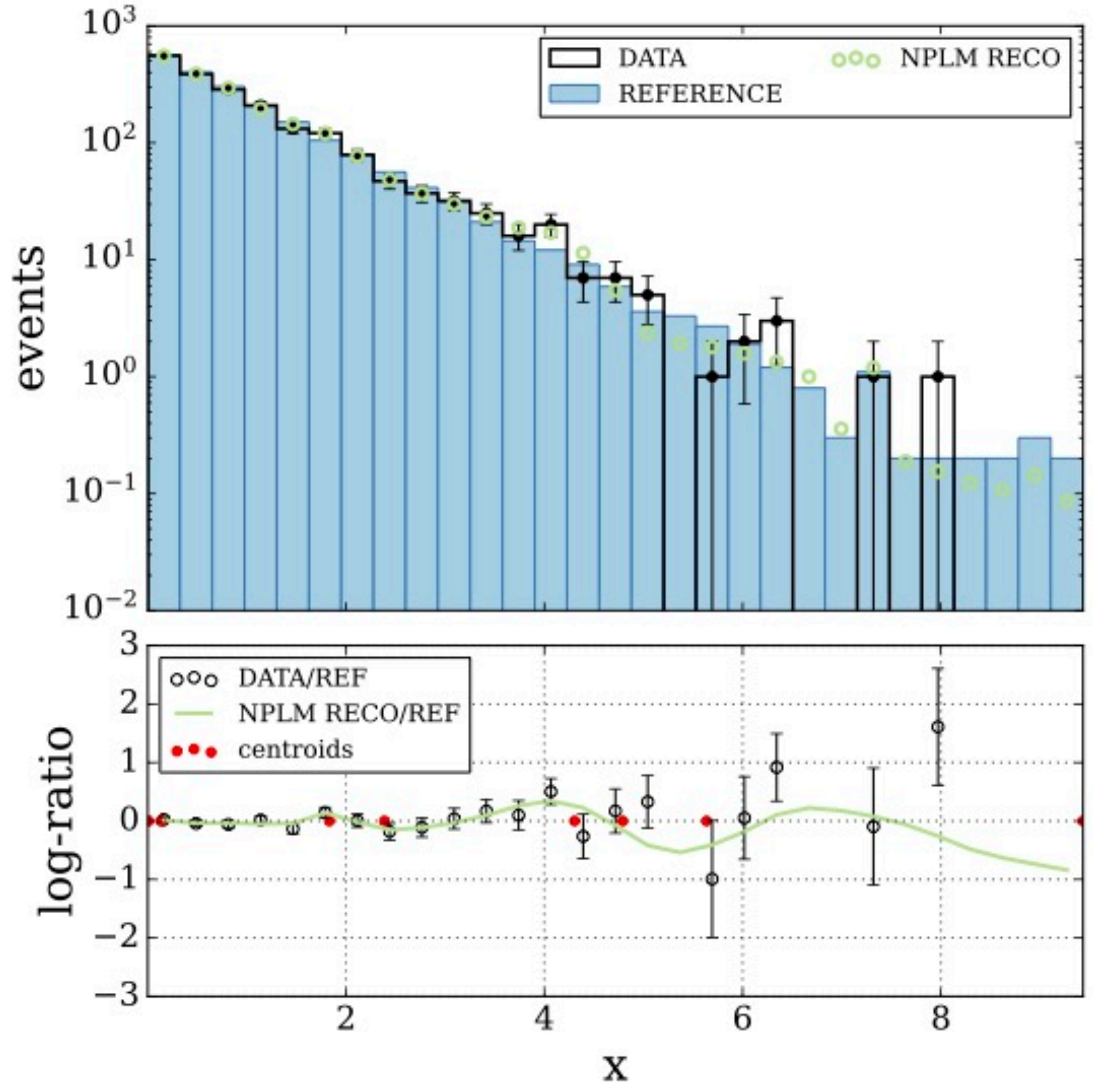
$$\bar{t}(\mathcal{D}) = 2\sum_{x\in\mathcal{D}} f_{\mathbf{w}}(x) - 2\sum_{x\in\mathcal{R}} \frac{\mathrm{N(R)}}{\mathrm{N}_{\mathcal{R}}}\left[e^{f(x;\mathbf{w})} - 1\right]$$

"Learning New Physics from a Machine" Phys. Rev. D

How to mitigate wrong inductive biases?

Inductive bias from model selection

Kernel Methods model hyper-parameter choice



Kernel-based model: $f_w(x) = \sum_{i=1}^{M} w_i k_\sigma(x, \tilde{x}_i)$

$k_\sigma(x, x') = e^{-\|x-x'\|^2/2\sigma^2}$ (gaussian kernel)

Loss:

$\hat{L}(f_w) + \lambda R(f_w)$

Weighted binary cross entropy:

$$\hat{L}(f_w) = \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} a_0(1-y)\log\left(1+e^{f(x)}\right) + a_1 y \log\left(1+e^{-f(x)}\right)$$

Regularization term:

$$R(f_w) = \sum_{ij} w_i w_j k_\sigma(x_i, x_j)$$

Hyperparameters:

- M: number of kernels
- σ: kernel width
- λ: L2 regularization

The hyper parameters M, σ, λ define the family of alternatives

Aggregation of multiple tests

Aggregation rule for the p-value:

$$p_{\text{aggreg}} = \min_{\sigma \in [\sigma_1, \ldots, \sigma_n]} [p_\sigma]$$

$[\sigma_1, \ldots, \sigma_n] = [5\%, 25\%, 50\%, 75\%, 95\%]$ quantiles of the pairwise distance between reference-distributed data points (after standardization).

Strategy:

1. **Test toys under the null hypothesis** by sampling background events: $\{t_\sigma(D_{\text{pseudo}}), \sigma \in [\sigma_1, \ldots, \sigma_n]\}_{i=0}^{N_{\text{toys}}}$.
2. **Test the data of interest** $\forall \sigma : \{t_\sigma(D), \sigma \in [\sigma_1, \ldots, \sigma_n]\}$
3. Compute the empirical **p-values** $\forall \sigma : \{p_\sigma(D), \sigma \in [\sigma_1, \ldots, \sigma_n]\}$
4. Select the **minimum p-value:** $p_{\text{aggreg}} = \min_{\sigma \in [\sigma_1, \ldots, \sigma_n]} [p_\sigma]$

Aggregation of multiple tests

1D proof-of-concept

Signal benchmarks: gaussian resonances with various width (σ_{NP}) and locations ($\bar{x}_{NP}$).
[N(S) signal injection over N(B) = 2000 events]

NPLM simple tests with different kernel widths (σ)

Aggregation of simple tests

N(S)	7	18	13	10	90
$\bar{x}_{NP}$	4	4	4	6.4	1.6
σ_{NP}	0.01	0.16	0.64	0.16	0.16
$\sigma = 0.1$	**0.008 ± 0.003**	0.032 ± 0.006	0.002 ± 0.001	0.026 ± 0.005	0.30 ± 0.02
$\sigma = 0.3$	0.001 ± 0.001	0.056 ± 0.007	0.001 ± 0.001	0.14 ± 0.01	0.49 ± 0.02
$\sigma = 0.7$	0	**0.059 ± 0.008**	0.003 ± 0.002	**0.21 ± 0.01**	**0.53 ± 0.02**
$\sigma = 1.4$	0	0.045 ± 0.007	0.005 ± 0.002	0.19 ± 0.01	0.41 ± 0.02
$\sigma = 3.0$	0	0.020 ± 0.004	**0.008 ± 0.003**	0.11 ± 0.01	0.23 ± 0.02
aggregation	**0.009 ± 0.003**	**0.11 ± 0.01**	**0.013 ± 0.004**	**0.27 ± 0.02**	**0.62 ± 0.02**

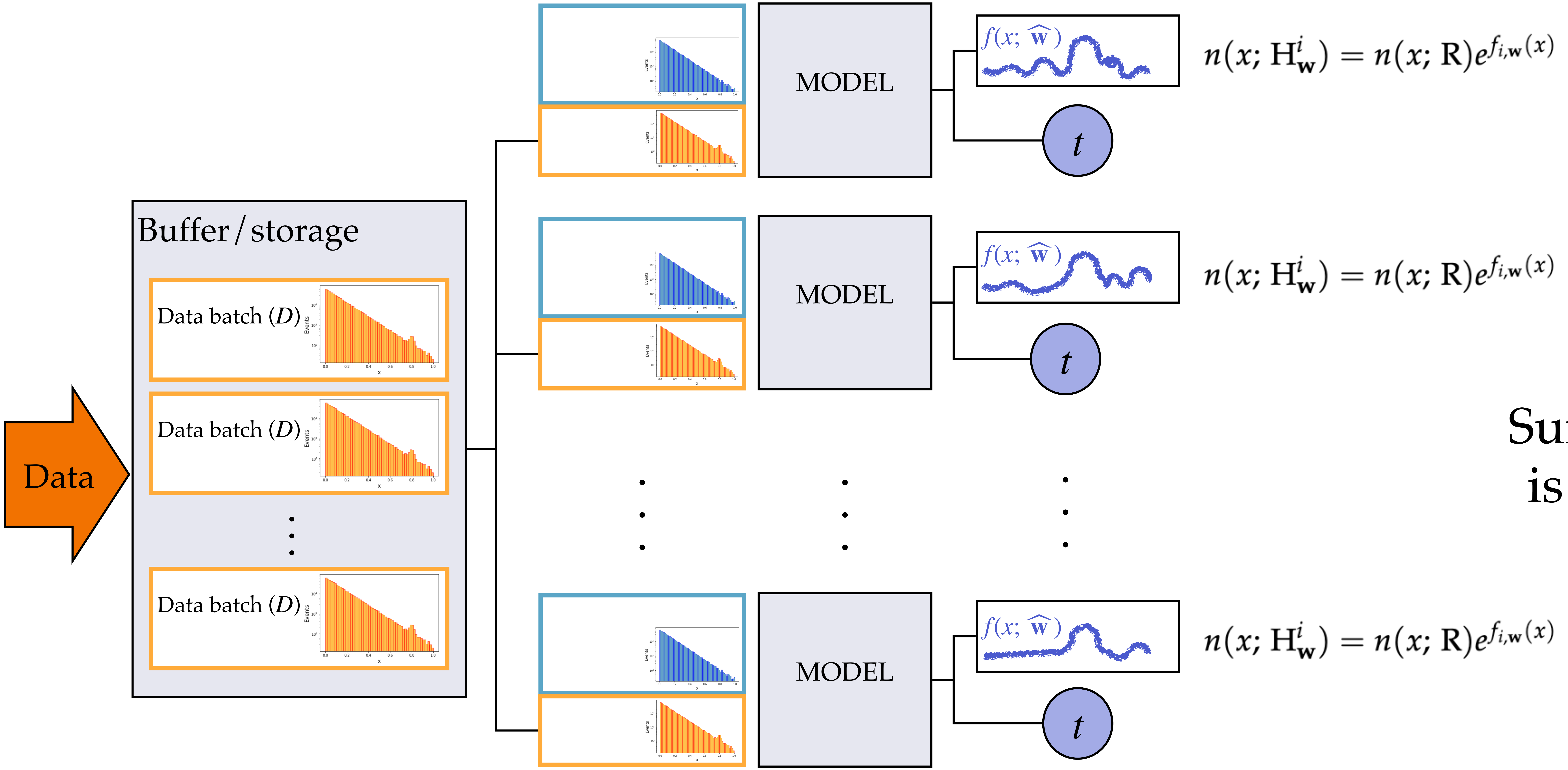
Table 1: 1D experiments: probability of observing $Z \geq 3$

Single tests have different power over the signal benchmarks.
The aggregation shows **uniform enhanced power** over the range of benchmarks.

Preliminary multi-dimensional tests confirm these findings

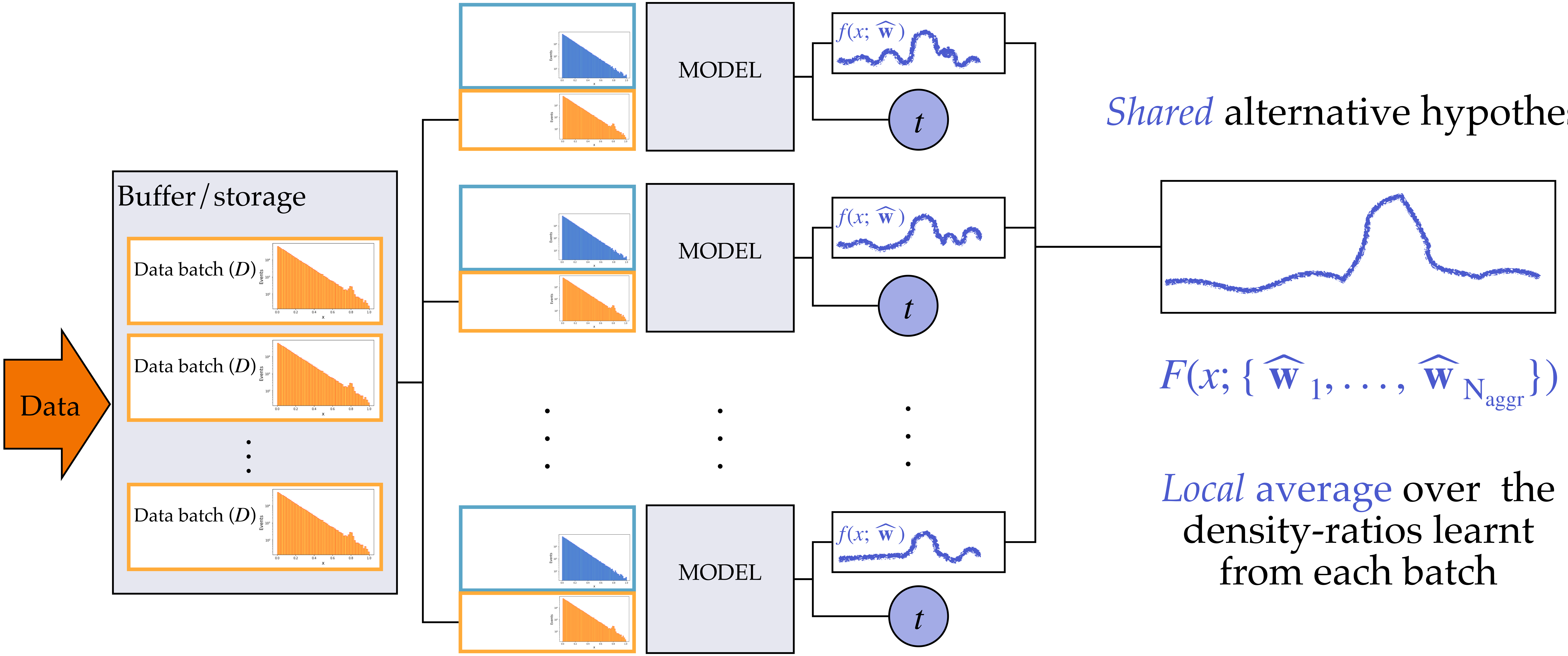
How to deal with large samples?

Combining NPLM over *multiple batches*



Summing over t is not optimal!

Combining NPLM over *multiple batches*



Shared alternative hypothesis

$$F(x; \{\widehat{\mathbf{w}}_1, \ldots, \widehat{\mathbf{w}}_{N_{aggr}}\})$$

Local average over the density-ratios learnt from each batch

Combining NPLM over *multiple batches*

Shared alternative hypothesis

$$F_{\mathbf{w}}^{N_{\text{aggr}}}(x) = \log \frac{n(x; \mathrm{H}_{\mathbf{w}}^{N_{\text{aggr}}})}{n(x; \mathrm{R})}$$

$$= \log \left[\frac{1}{\mathrm{N}_{\text{aggr}}} \sum_{i=1}^{N_{\text{aggr}}} e^{f_{i,\mathbf{w}}(x)} \right]$$

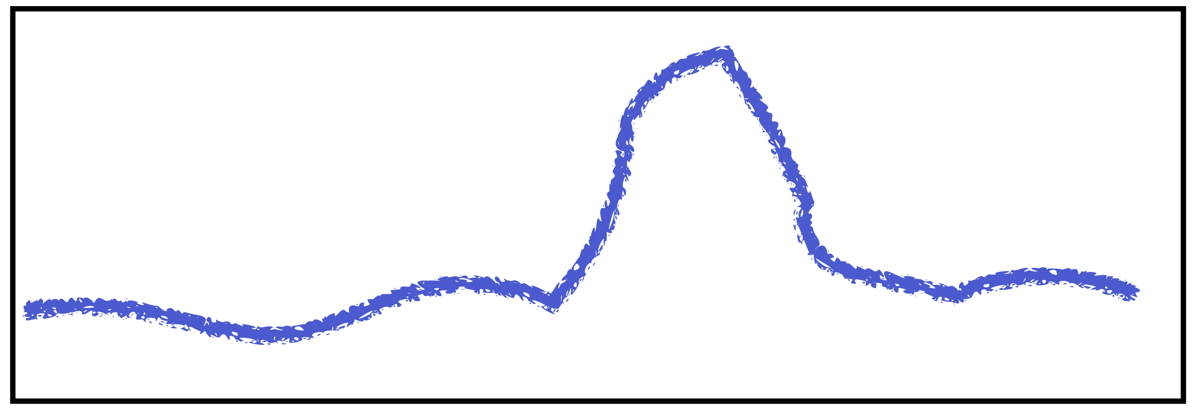
N_{aggr}: # of aggregated batches

$n(x; \mathrm{H}_{\mathbf{w}}^{i}) = n(x; \mathrm{R}) e^{f_{i,\mathbf{w}}(x)}$

$$t_{\text{AGGR}}^{N_{\text{aggr}},N_{\text{test}}}(\mathcal{D}) = 2 \sum_{i=1}^{N_{\text{test}}} \log \frac{\mathcal{L}(\mathcal{D}_i | \mathrm{H}_{\mathbf{w}}^{N_{\text{aggr}}})}{\mathcal{L}(\mathcal{D}_i | \mathrm{R})}$$

$$= 2 \sum_{i=1}^{N_{\text{test}}} \left[\sum_{x \in \mathcal{R}} w_{\mathcal{R}}(x) (1 - e^{F_{\mathbf{w}}^{N_{\text{aggr}}}(x)}) + \sum_{x \in \mathcal{D}_i} F_{\mathbf{w}}^{N_{\text{aggr}}}(x) \right]$$

N_{aggr}: # of aggregated batches
N_{test}: # of tested batches



$F(x; \{\widehat{\mathbf{w}}_1, \ldots, \widehat{\mathbf{w}}_{\mathrm{N}_{\text{aggr}}}\})$

Local average over the density-ratios learnt from each batch

Combining NPLM over *multiple batches*

1D proof-of-concept

Signal benchmarks:
- Broad peak: $\bar{x} = 4, \sigma = 0.64$
- Narrow peak: $\bar{x} = 4, \sigma = 0.01$

TESTS:

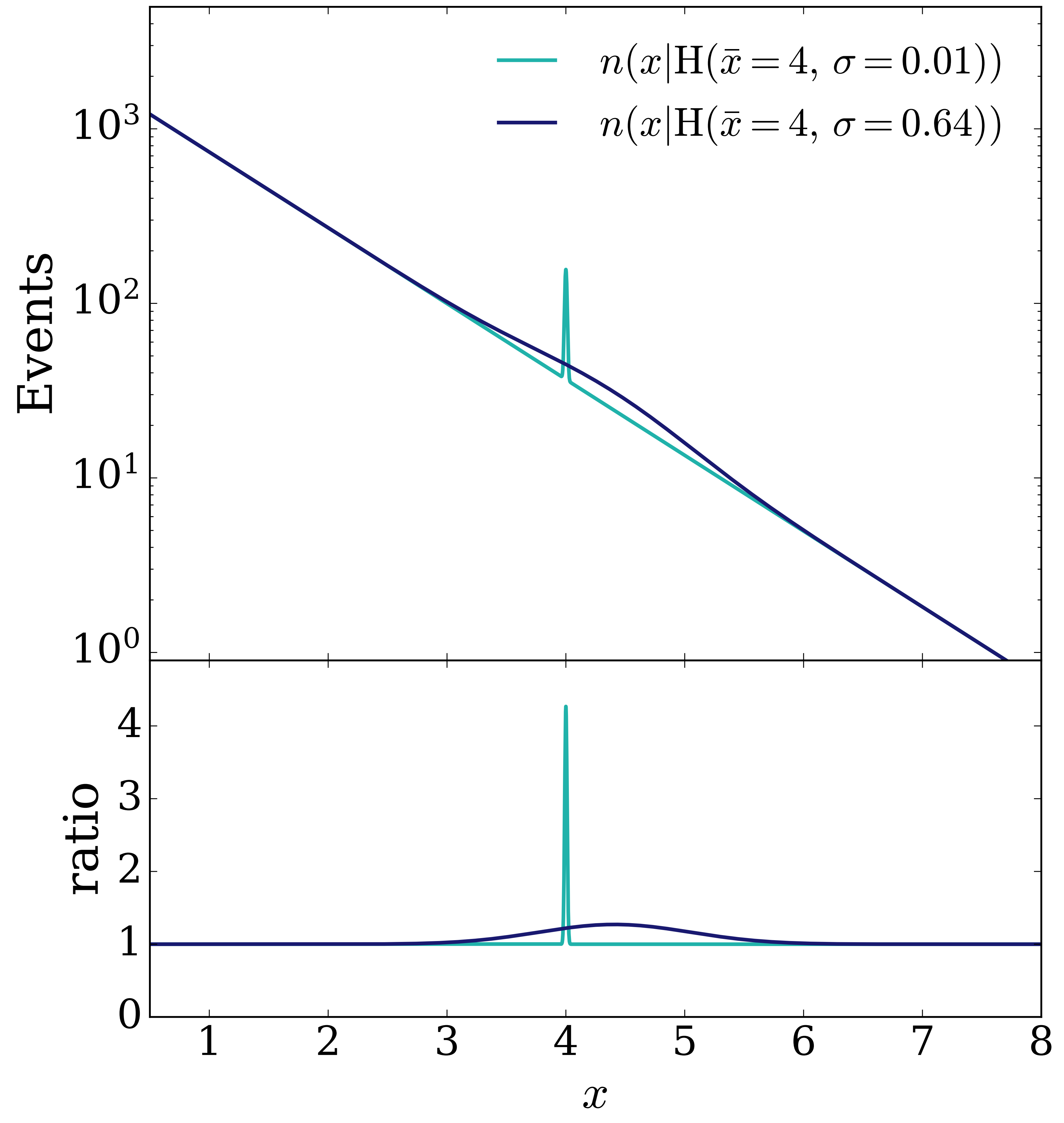
Single batch jobs:

1) N(B) = 2 000
   - Narrow peak: N(S) = 3 ($Z_{ideal} = 1.7$)
   - Broad peak: N(S) = 13 ($Z_{ideal} = 1.5$)

2) N(B) = 16 000
   - Narrow peak: N(S) = 24 ($Z_{ideal} = 4.8$)
   - Broad peak: N(S) = 104 ($Z_{ideal} = 4.2$)

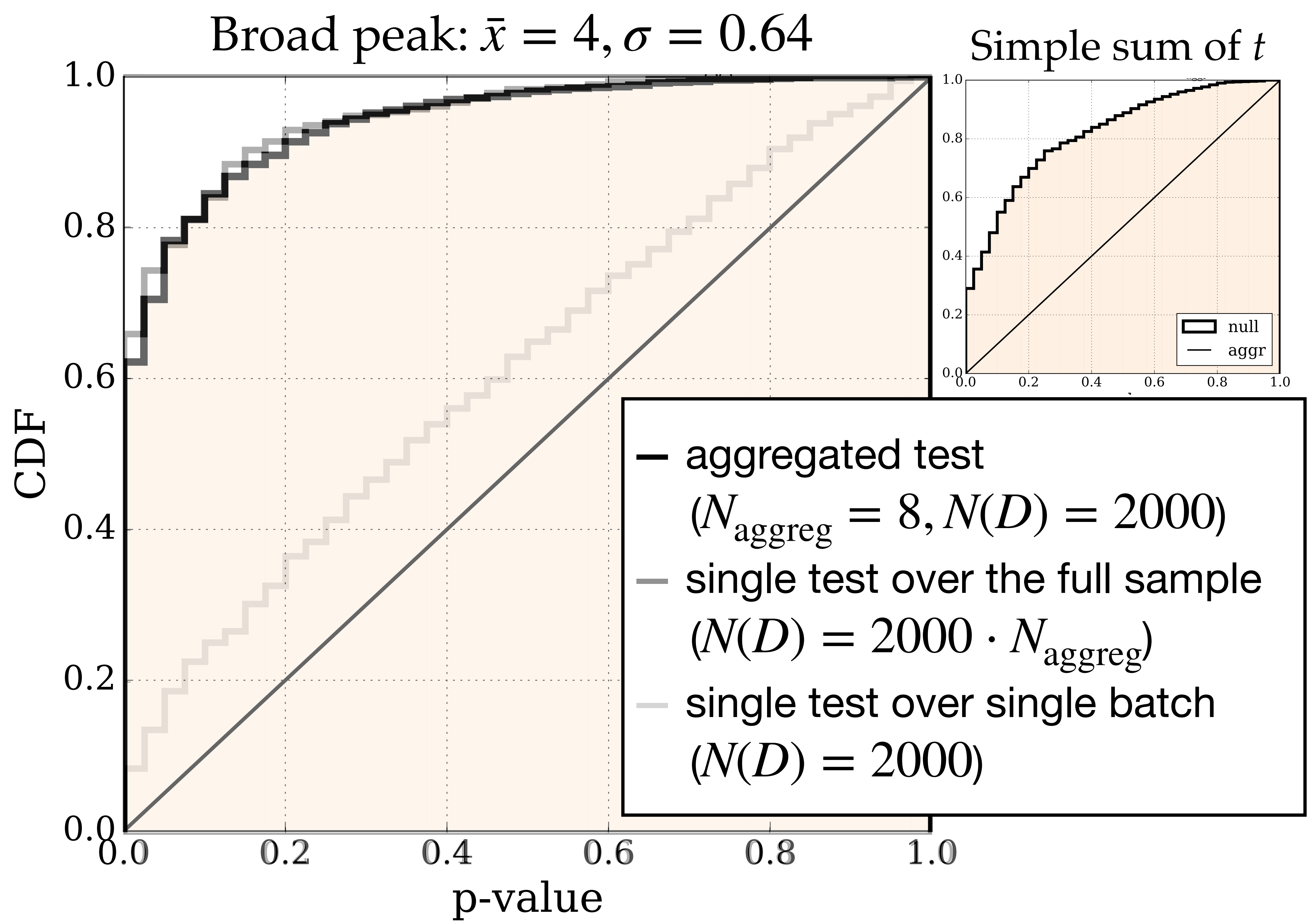
Aggregation over 8 batches:

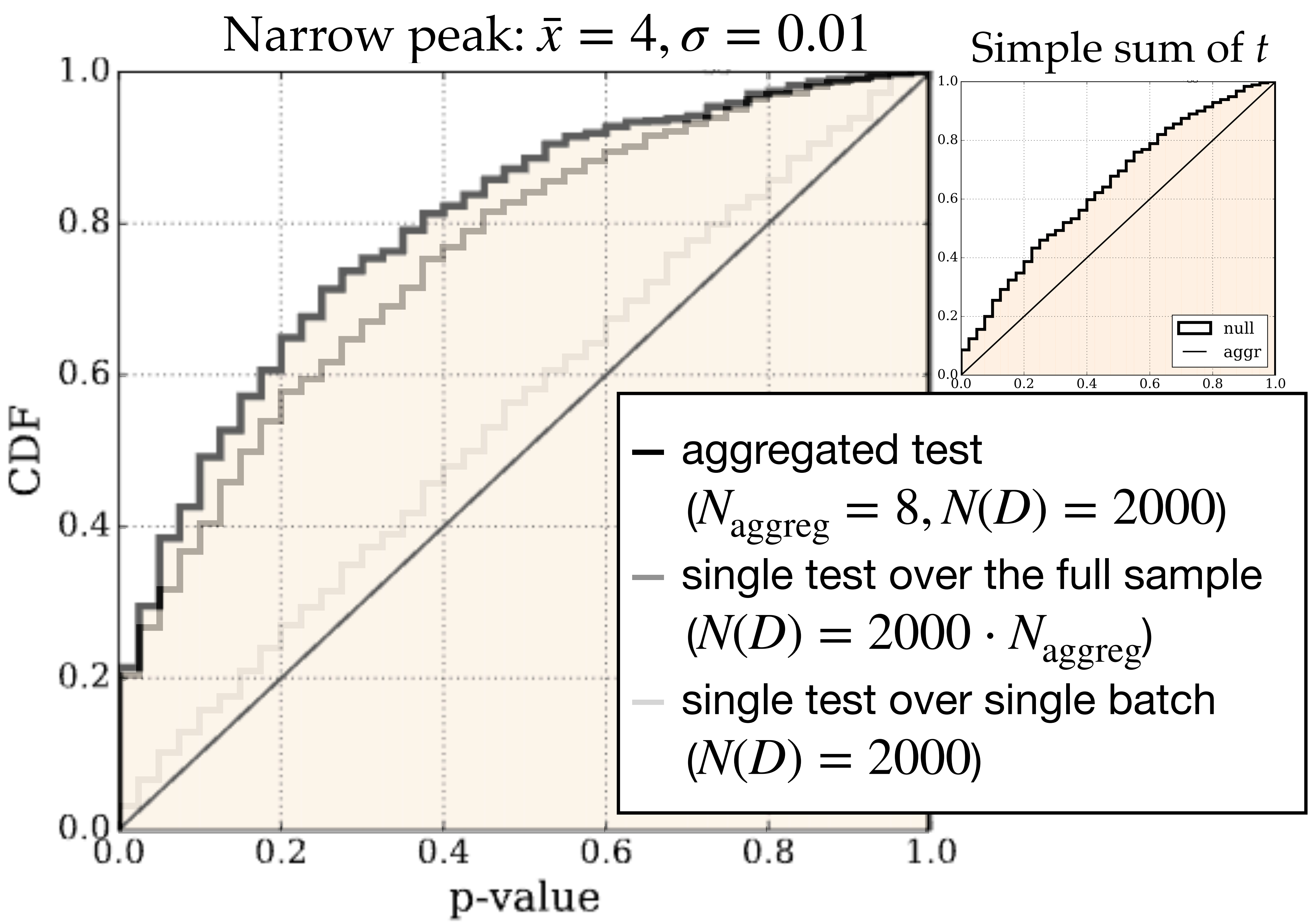
N(B) = 2 000, $N_{aggr} = 8$
- $N_{test} = 1$
- $N_{test} = 8$



Combining NPLM over *multiple batches*

1D proof-of-concept





The proposed combination lead to performances that are comparable to the ones obtained with the full statistics!
Clear gain with respect to simple sum of tests

Physics benchmark	**Pr(p-value<0.001) ($Z > 3$) [%]** single train (N(R)=16000)	8 splits (N(R)=2000)	**Pr(p-value<0.02) ($Z > 2$) [%]** single train (N(R)=16000)	8 splits (N(R)=2000)
narrow resonance	8.0 ± 0.9	$\mathbf{7.3 \pm 0.9}$	19 ± 1	$\mathbf{20 \pm 1}$
broad resonance	33 ± 2	$\mathbf{54 \pm 2}$	64 ± 3	$\mathbf{78 \pm 3}$

Summary and next steps

We addressed two outstanding questions of agnostic goodness of fit with multiple testing:

- How to mitigate induced biases due to model selection
- How to exploit large statistics in a fast and efficient way

Relevant to:

- Quasi-online monitoring
- New Physics searches
- Data validation

Towards a resource *efficient*, *automatized*, and *powerful* GoF tool

Ongoing efforts:

- Multiple testing in presence of systematic uncertainties
- Comparison with state-of-the-art GoF approaches

Backup slides

Inductive bias from model selection

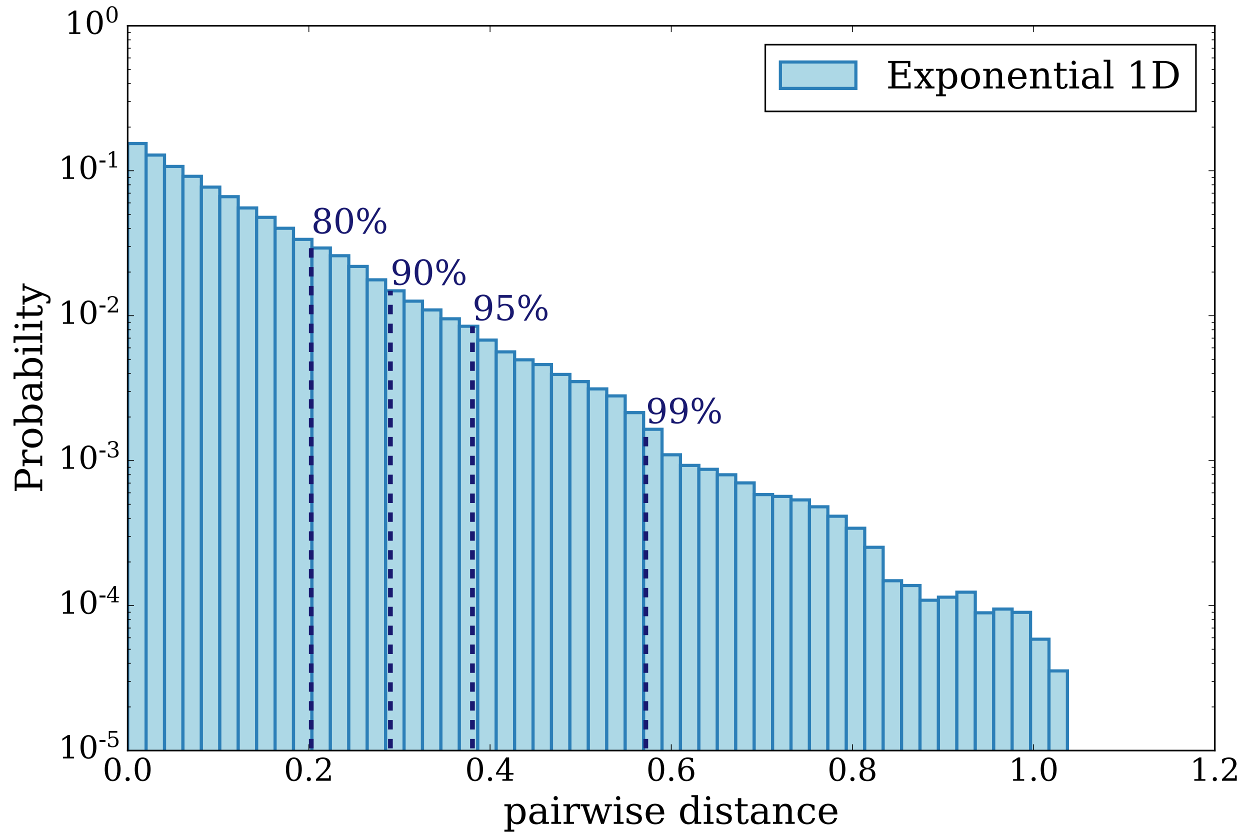
Kernel Methods model hyper-parameter choice

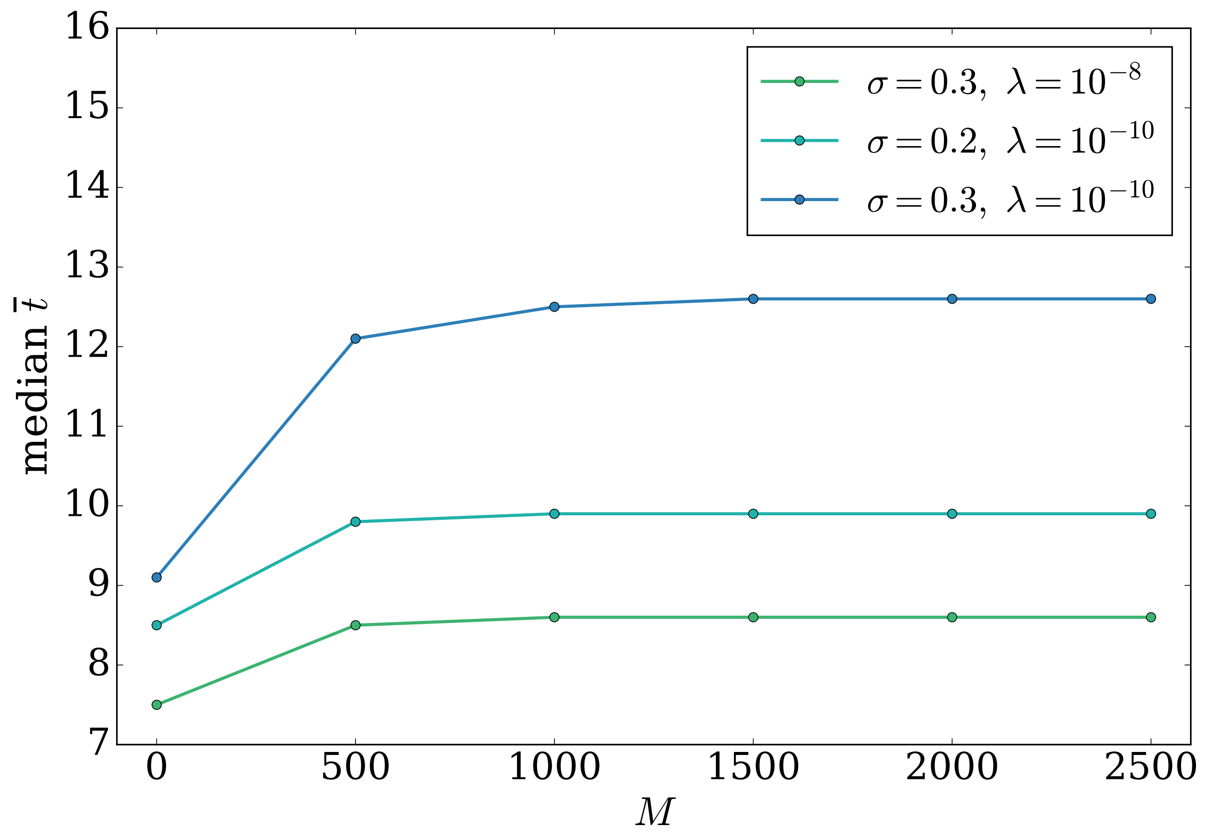
Asymptotic χ^2 is behavior is observed for each choice of (M, σ, λ), provided that $N_R \gg N_D$.

How to choose (M, σ, λ)??

Heuristics:

- Number of centers M: at least as large as $\sqrt{N}$ to achieve statistically optimal bounds of the training convergence).
- Gaussian width σ: we select it as the 90th percentile of the pairwise distance between reference-distributed data points (after standardisation).
- Regularisation parameter λ: is kept as small as possible while keeping training stable





The hyper parameter choice impact the sensitivity to different signal patterns. How to mitigate this effect?

ML-based Neyman-Pearson GoF test

NPLM: Likelihood ratio from weighted Binary Cross Entropy

Test statistic (unbinned extended likelihood ratio)

$$\bar{t}(\mathcal{D}) = 2\max_{\mathbf{w}} \log\left[\frac{\mathcal{L}(\mathcal{D}\,|\,\mathrm{H}_{\mathbf{w}})}{\mathcal{L}(\mathcal{D}\,|\,\mathrm{R}_{0})}\right] = 2\max_{\mathbf{w}}\left\{\log\left[\frac{e^{-N(\mathbf{w})}}{e^{-N(\mathrm{R})}}\prod_{i=1}^{\mathrm{N_D}}\frac{n(x_i|\mathbf{w})}{n(x_i|\mathrm{R})}\right]\right\}$$

$$= 2\sum_{x\in\mathcal{D}} f_{\mathbf{w}}(x) - 2\sum_{x\in\mathcal{R}}\frac{\mathrm{N(R)}}{\mathrm{N}_{\mathcal{R}}}\left[e^{f(x;\,\mathbf{w})} - 1\right]$$

$$f(x;\,\widehat{\mathbf{w}}) = \log\left[\frac{n(x\,|\,\mathrm{H}_{\widehat{\mathbf{w}}})}{n(x\,|\,\mathrm{R})}\right]$$

Loss function

$$\bar{L}\left[f(x;\,\mathbf{w})\right] = -\sum_{x\in\mathcal{D}}\log\left[1+e^{-f_{\mathbf{w}}(x)}\right] + \sum_{x\in\mathcal{R}}\frac{\mathrm{N(R)}}{\mathrm{N}_{\mathcal{R}}}\log\left[1+e^{f_{\mathbf{w}}(x)}\right]$$

w: trainable parameters on the NN model
D: data sample
R: reference sample (built according to the R_0 hypothesis); could be weighted (*w*)

Assumptions:

- $N_R \gg N_D$ the statistical fluctuations of the reference sample are negligible.
- the weights of the reference sample (*w*) are such that the reference sample is normalised to match the data sample luminosity

Aggregation of multiple tests

Preliminary results: 1D, 5D, 21D benchmarks

N(S)	7	18	13	10	90
$\bar{x}_{NP}$	4	4	4	6.4	1.6
σ_{NP}	0.01	0.16	0.64	0.16	0.16
$\sigma = 0.1$	$\mathbf{0.008 \pm 0.003}$	0.032 ± 0.006	0.002 ± 0.001	0.026 ± 0.005	0.30 ± 0.02
$\sigma = 0.3$	0.001 ± 0.001	0.056 ± 0.007	0.001 ± 0.001	0.14 ± 0.01	0.49 ± 0.02
$\sigma = 0.7$	0	$\mathbf{0.059 \pm 0.008}$	0.003 ± 0.002	$\mathbf{0.21 \pm 0.01}$	$\mathbf{0.53 \pm 0.02}$
$\sigma = 1.4$	0	0.045 ± 0.007	0.005 ± 0.002	0.19 ± 0.01	0.41 ± 0.02
$\sigma = 3.0$	0	0.020 ± 0.004	$\mathbf{0.008 \pm 0.003}$	0.11 ± 0.01	0.23 ± 0.02
aggregation	$\mathbf{0.009 \pm 0.003}$	$\mathbf{0.11 \pm 0.01}$	$\mathbf{0.013 \pm 0.004}$	$\mathbf{0.27 \pm 0.02}$	$\mathbf{0.62 \pm 0.02}$

Table 1: 1D experiments: probability of observing $Z \geq 3$

N(S)	1000	2500
$\sigma = 4.3$	0.003 ± 0.002	0.11 ± 0.01
$\sigma = 5.3$	0.006 ± 0.002	0.19 ± 0.01
$\sigma = 6.0$	0.007 ± 0.003	0.25 ± 0.02
$\sigma = 6.6$	0.007 ± 0.003	0.36 ± 0.02
$\sigma = 7.5$	$\mathbf{0.008 \pm 0.003}$	$\mathbf{0.49 \pm 0.02}$
aggregation	$\mathbf{0.009 \pm 0.003}$	$\mathbf{0.49 \pm 0.02}$

Table 3: **HIGGS dataset:** probability of observing $Z \geq 3$

test	Z′ M = 180 GeV, W = 0.02 GeV	Z′ M = 180 GeV, W = 2 GeV	Z′ M = 200 GeV, W = 10 GeV	Z′ M = 300 GeV, W = 15 GeV	Z′ M = 600 GeV, W = 30 GeV	EFT $c_w = 10^{-6}$
$\sigma = 0.57$	$\mathbf{0.04 \pm 0.02}$	0.04 ± 0.02	$\mathbf{0.06 \pm 0.02}$	0.01 ± 0.01	0.01 ± 0.01	0.005 ± 0.005
$\sigma = 1.19$	0.03 ± 0.02	$\mathbf{0.05 \pm 0.02}$	$\mathbf{0.06 \pm 0.02}$	0.01 ± 0.01	0.02 ± 0.01	0.02 ± 0.007
$\sigma = 1.79$	0.03 ± 0.02	0.04 ± 0.02	0.05 ± 0.01	$\mathbf{0.05 \pm 0.02}$	$\mathbf{0.04 \pm 0.02}$	$\mathbf{0.03 \pm 0.01}$
$\sigma = 2.49$	0.01 ± 0.01	0	0.02 ± 0.01	0.02 ± 0.01	0.02 ± 0.01	0.005 ± 0.005
$\sigma = 3.57$	0.01 ± 0.01	0.02 ± 0.01	0.02 ± 0.01	0.02 ± 0.01	0.02 ± 0.01	0.02 ± 0.01
aggregation	$\mathbf{0.08 \pm 0.03}$	$\mathbf{0.11 \pm 0.03}$	$\mathbf{0.12 \pm 0.02}$	$\mathbf{0.08 \pm 0.03}$	$\mathbf{0.08 \pm 0.02}$	$\mathbf{0.05 \pm 0.02}$

Table 2: 5D MUMU experiments: probability of observing $Z \geq 3$