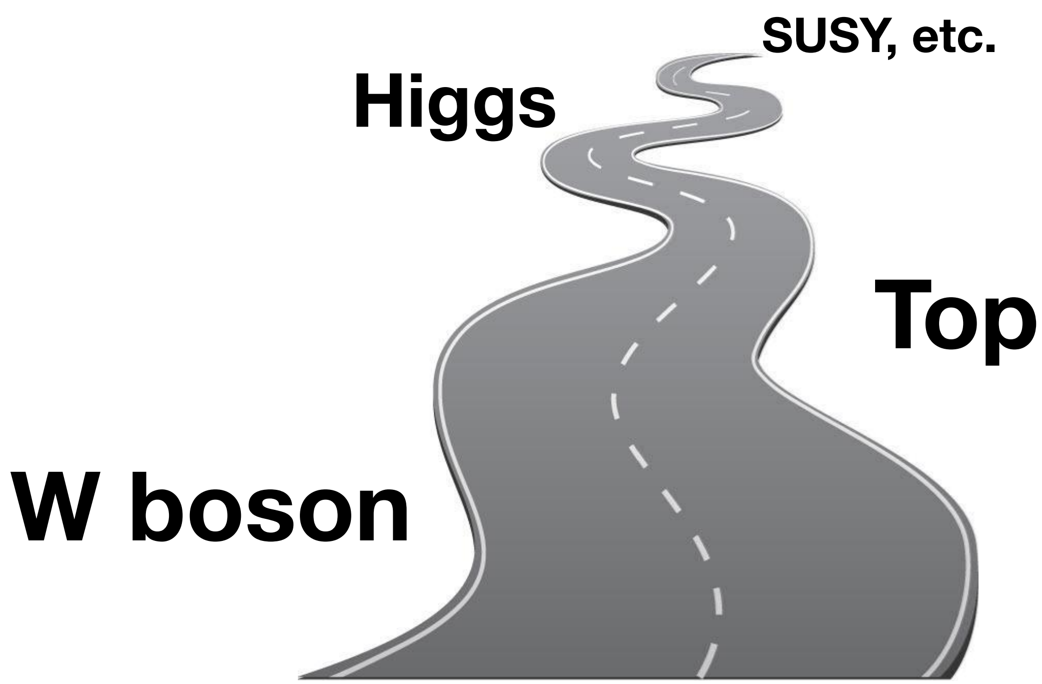
Learning New Physics from a Machine

Andrea Wulzer

based on 1806.02350,
with R.T. D'Agnolo

The Challenge

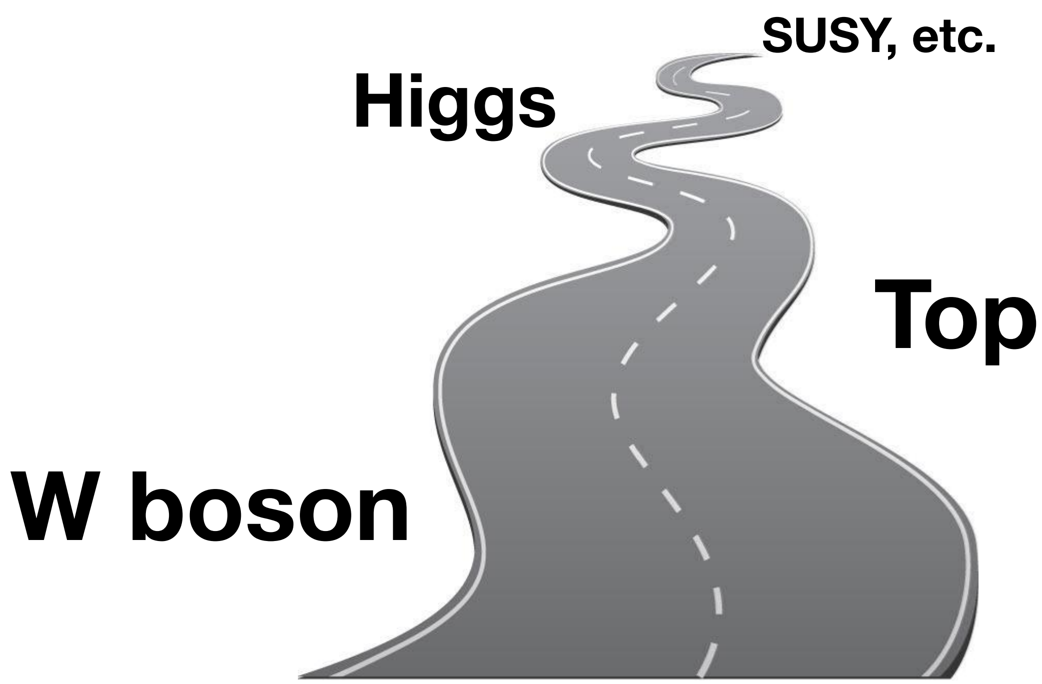
HEP yesterday



HEP today

The Challenge

HEP yesterday

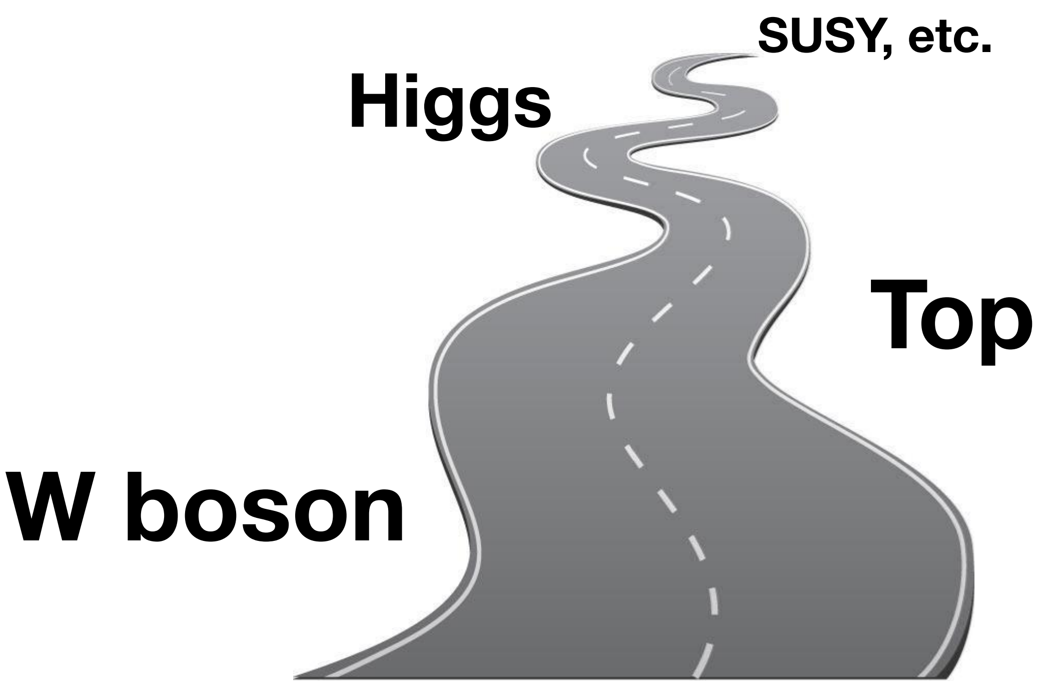


HEP today

Standard (Reference) Model **not guaranteed to fail** in any specific process

The Challenge

HEP yesterday



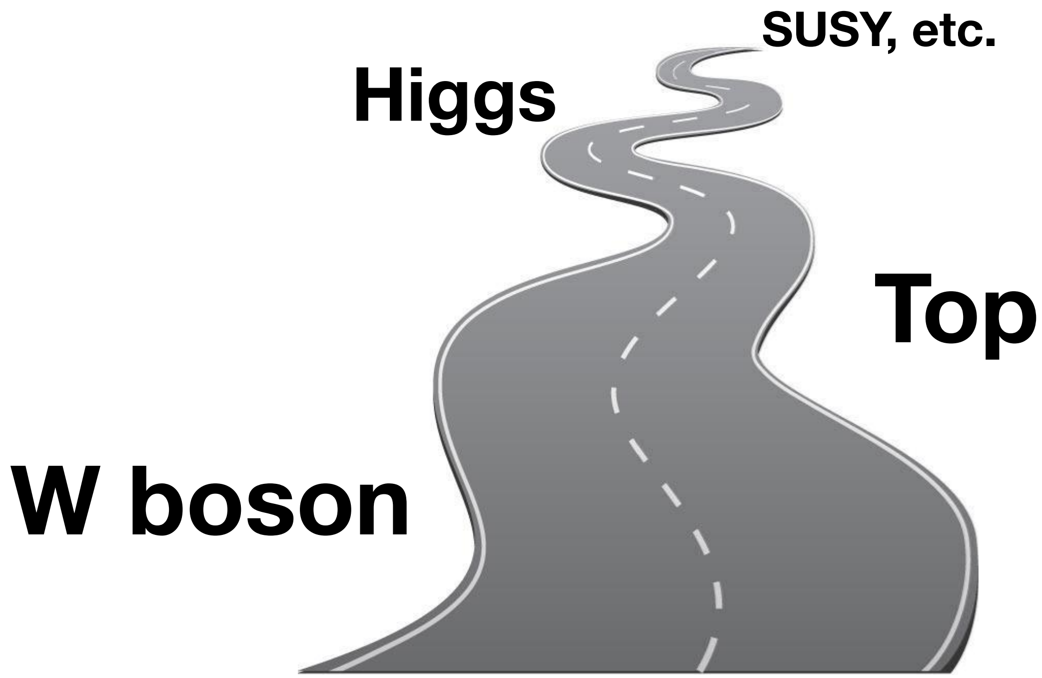
HEP today

Standard (Reference) Model **not guaranteed to fail** in any specific process

BSM guidance has never been that weak!

The Challenge

HEP yesterday



HEP today

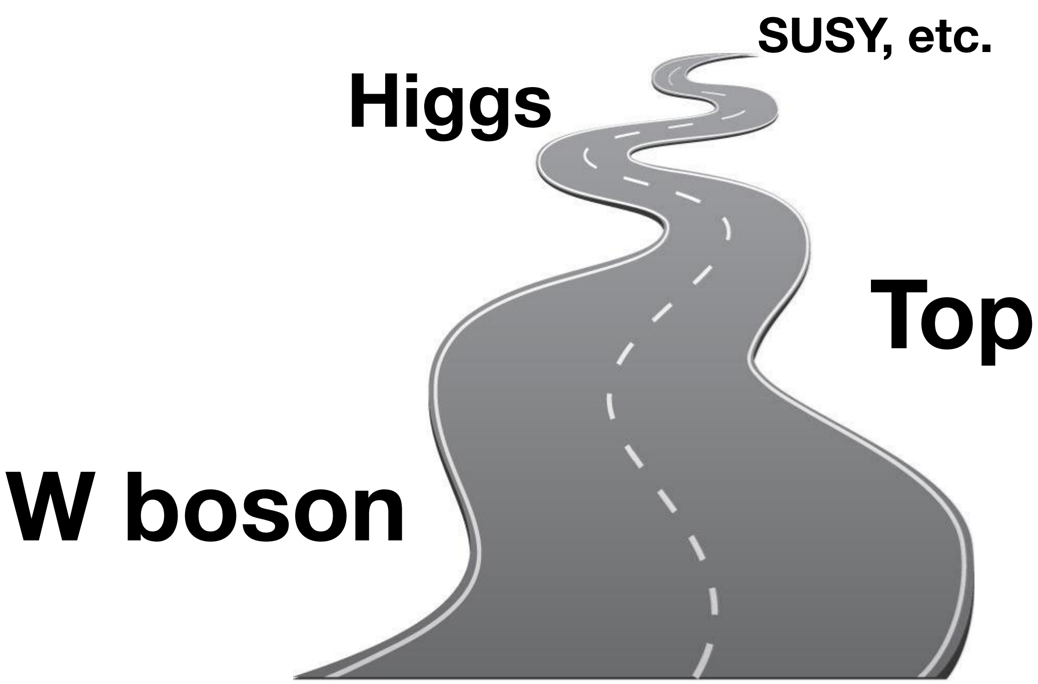
Standard (Reference) Model **not guaranteed to fail** in any specific process

BSM guidance has never been that weak!

We should not only be searching for specific BSM signals
[also because model builders' fantasy is large but finite]

The Challenge

HEP yesterday



HEP today

Standard (Reference) Model **not guaranteed to fail** in any specific process

BSM guidance has never been that weak!

We should not only be searching for specific BSM signals
[also because model builders' fantasy is large but finite]

Similar considerations apply to fund. int. phys. in general

The Challenge

“We haven’t seen anything” $\overset{?}{\Rightarrow}$ “There is nothing to see”

The Challenge

"We haven't seen anything" $\stackrel{?}{\Rightarrow}$ "There is nothing to see"

Sharp answer is **NO!**

We knew **a priori** that **NP is elusive** [because of previous exp.]

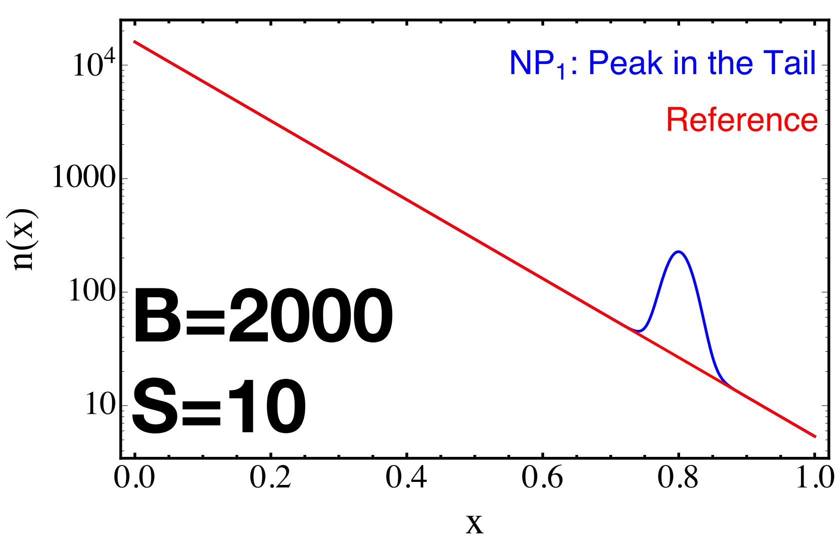
The Challenge

"We haven't seen anything" $\overset{?}{\Rightarrow}$ "There is nothing to see"

Sharp answer is **NO!**

We knew **a priori** that **NP is elusive** [because of previous exp.]

- NP can be **large** only in **low-probability** phase space regions



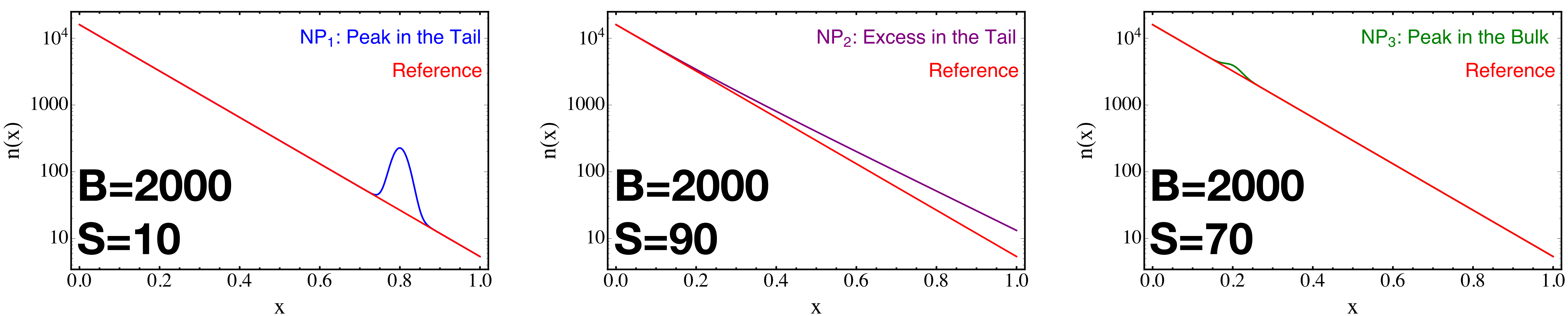
The Challenge

“We haven’t seen anything” $\overset{?}{\Rightarrow}$ “There is nothing to see”

Sharp answer is **NO!**

We knew **a priori** that **NP is elusive** [because of previous exp.]

- NP can be **large** only in **low-probability** phase space regions
- or it can be a **tiny correction** to Reference [or both]



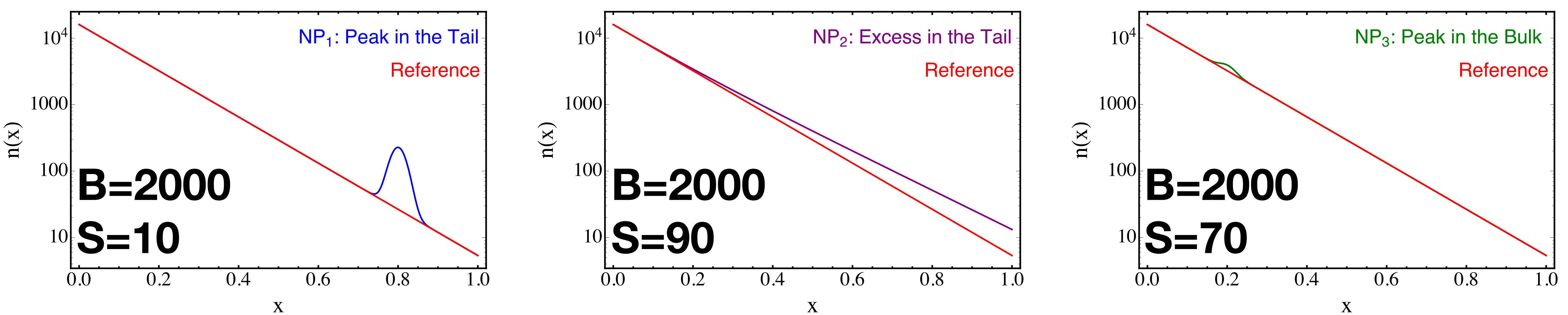
The Challenge

“We haven’t seen anything” $\overset{?}{\Rightarrow}$ “There is nothing to see”

Sharp answer is **NO!**

We knew **a priori** that **NP is elusive** [because of previous exp.]

- NP can be **large** only in **low-probability** phase space regions
- or it can be a **tiny correction** to Reference [or both]
- Likely shows up in **peculiar kin. var.s** [try search Z’ w/o knowing m_{ll}]



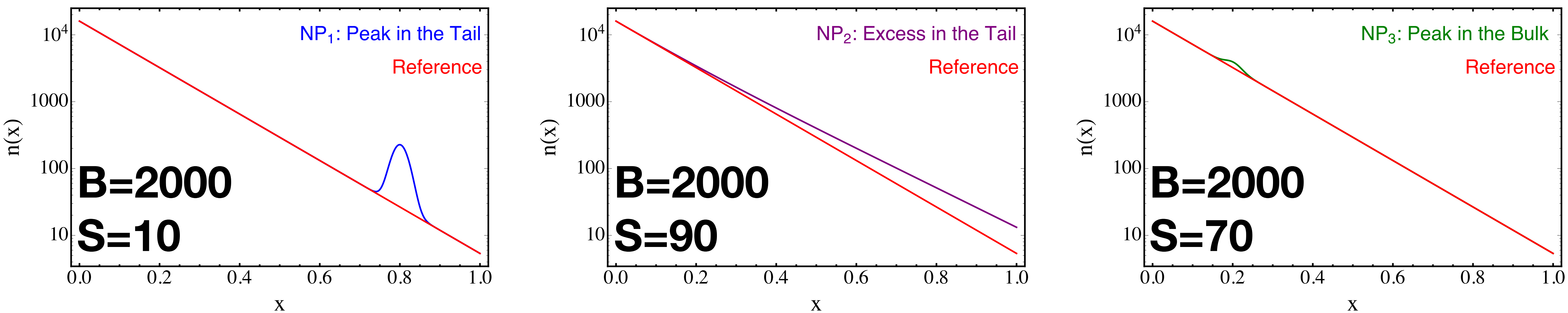
The Challenge

“We haven’t seen anything” $\overset{?}{\Rightarrow}$ “There is nothing to see”

Sharp answer is **NO!**

We knew **a priori** that **NP is elusive** [because of previous exp.]

- NP can be **large** only in **low-probability** phase space regions
- or it can be a **tiny correction** to Reference [or both]
- Likely shows up in **peculiar kin. var.s** [try search Z’ w/o knowing m_{ll}]



These peculiarities make most standard techniques to assess data compatibility with Reference fated to fail

Searching for the Unknown

New Physics Search:

algorithm aimed at discovering data **departures** from a given **Reference Model**

Searching for the Unknown

New Physics Search:

algorithm aimed at discovering data **departures** from a given **Reference Model**

Model-Independent NP Search:

ideally sensitive to “any” NP model, rather than to specific “BSM” alternatives

Searching for the Unknown

New Physics Search:
algorithm aimed at discovering data **departures** from a given **Reference Model**

Model-Independent NP Search:
ideally sensitive to “any” NP model, rather than to specific “BSM” alternatives

Disadvantage: negative M-I searches are **not informative**

Advantage: might **discover** model we had not thought of

Searching for the Unknown

New Physics Search:
algorithm aimed at discovering data **departures** from a given **Reference Model**

Model-Independent NP Search:
ideally sensitive to “any” NP model, rather than to specific “BSM” alternatives

Disadvantage: negative M-I searches are **not informative**

Advantage: might **discover** model we had not thought of

Important Remark: [not only to please statisticians]

- hypothesis test **unavoidably requires alternative** hypothesis, or **probability model**, to compare with
- **M-I physically** means that the alternative distribution is not selected as the one predicted by known alternative physics model

Maximum Likelihood

[J. Neyman and E. S. Pearson, 1933]

Data: $\mathcal{D} = \{x_i\}, \; i = 1, \ldots, \mathcal{N}_{\mathcal{D}}$

Reference Distribution: $n(x|\mathrm{R})$

Alternative Distribution: $n(x|\mathbf{w})$

$$n(x) = N\,P(x)$$
$$N = \int dx\, n(x)$$

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

Maximum Likelihood

[J. Neyman and E. S. Pearson, 1933]

Data: $\mathcal{D} = \{x_i\},\ i = 1, \ldots, \mathcal{N}_{\mathcal{D}}$

Reference Distribution: $n(x|\mathrm{R})$

Alternative Distribution: $n(x|\mathbf{w})$

$$n(x) = N\,P(x)$$
$$N = \int dx\, n(x)$$

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

Fully Model-Dependent (ideal):

$$2\log\left[\frac{e^{-N(\mathrm{NP})}}{e^{-N(\mathrm{R})}}\prod_{i=1}^{\mathcal{N}_{\mathcal{D}}}\frac{n(x_i|\mathrm{NP})}{n(x_i|\mathrm{R})}\right] = t_{\mathrm{id}}$$

Maximum Likelihood

[J. Neyman and E. S. Pearson, 1933]

Data: $\mathcal{D} = \{x_i\},\ i = 1, \ldots, \mathcal{N}_\mathcal{D}$

Reference Distribution: $n(x|\mathrm{R})$

Alternative Distribution: $n(x|\mathbf{w})$

$$n(x) = N\,P(x)$$
$$N = \int dx\, n(x)$$

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

Fully Model-Dependent (ideal):

$$2\log\left[\frac{e^{-N(\mathrm{NP})}}{e^{-N(\mathrm{R})}}\prod_{i=1}^{\mathcal{N}_\mathcal{D}}\frac{n(x_i|\mathrm{NP})}{n(x_i|\mathrm{R})}\right] = t_{\mathrm{id}}$$

Model-Independent:

$$-2\,\underset{\mathbf{w}}{\mathrm{Min}}\left[N(\mathbf{w}) - N(\mathrm{R}) - \sum_{i=1}^{\mathcal{N}_\mathcal{D}} f(x_i;\mathbf{w})\right]$$

Alternative in parametrised form: $n(x|\mathbf{w}) = n(x|\mathrm{R})\,e^{f(x;\mathbf{w})}$

Maximum Likelihood

[J. Neyman and E. S. Pearson, 1933]

Data: $\mathcal{D} = \{x_i\},\ i = 1, \ldots, \mathcal{N}_{\mathcal{D}}$

Reference Distribution: $n(x|\mathrm{R})$

Alternative Distribution: $n(x|\mathbf{w})$

$$n(x) = N\,P(x)$$
$$N = \int dx\, n(x)$$

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

Fully Model-Dependent (ideal):

$$2\log\left[\frac{e^{-N(\mathrm{NP})}}{e^{-N(\mathrm{R})}}\prod_{i=1}^{\mathcal{N}_{\mathcal{D}}}\frac{n(x_i|\mathrm{NP})}{n(x_i|\mathrm{R})}\right] = t_{\mathrm{id}}$$

Model-Independent:

$$-2\,\underset{\mathbf{w}}{\mathrm{Min}}\left[N(\mathbf{w}) - N(\mathrm{R}) - \sum_{i=1}^{\mathcal{N}_{\mathcal{D}}} f(x_i;\mathbf{w})\right]$$

Alternative in parametrised form: $n(x|\mathbf{w}) = n(x|\mathrm{R})\,e^{f(x;\mathbf{w})}$

If f piece-wise constant in bins: $t(\mathcal{D}) = 2\sum_{\alpha=1}^{N_{\mathrm{bin}}}\left[N_\alpha(\mathrm{R}) - O_\alpha + O_\alpha\log\frac{O_\alpha}{N_\alpha(\mathrm{R})}\right]$

Maximum Likelihood

[J. Neyman and E. S. Pearson, 1933]

Data: $\mathcal{D} = \{x_i\},\ i = 1, \ldots, \mathcal{N}_{\mathcal{D}}$

Reference Distribution: $n(x|\mathrm{R})$

Alternative Distribution: $n(x|\mathbf{w})$

$$n(x) = N\,P(x)$$
$$N = \int dx\, n(x)$$

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

Fully Model-Dependent (ideal):

$$2\log\left[\frac{e^{-N(\mathrm{NP})}}{e^{-N(\mathrm{R})}}\prod_{i=1}^{\mathcal{N}_{\mathcal{D}}}\frac{n(x_i|\mathrm{NP})}{n(x_i|\mathrm{R})}\right] = t_{\mathrm{id}}$$

Model-Independent:

$$-2\,\underset{\mathbf{w}}{\mathrm{Min}}\left[N(\mathbf{w}) - N(\mathrm{R}) - \sum_{i=1}^{\mathcal{N}_{\mathcal{D}}} f(x_i;\mathbf{w})\right]$$

Alternative in parametrised form: $n(x|\mathbf{w}) = n(x|\mathrm{R})\,e^{f(x;\mathbf{w})}$

If f piece-wise constant in bins: $t(\mathcal{D}) = 2\sum_{\alpha=1}^{N_{\mathrm{bin}}}\left[N_\alpha(\mathrm{R}) - O_\alpha + O_\alpha \log\frac{O_\alpha}{N_\alpha(\mathrm{R})}\right]$

recover **binned histogram test !**

Maximum Likelihood

[J. Neyman and E. S. Pearson, 1933]

Basic idea: $f(x; \mathbf{w}) = \mathrm{NN}$

replace histograms with NN, literally!

$t_{\rm id}$

$$-2 \underset{\mathbf{w}}{\mathrm{Min}} \left[N(\mathbf{w}) - N(\mathrm{R}) - \sum_{i=1}^{N_{\mathcal{D}}} f(x_i; \mathbf{w}) \right]$$

Alternative in parametrised form: $n(x|\mathbf{w}) = n(x|\mathrm{R})\, e^{f(x;\mathbf{w})}$

If f piece-wise constant in bins: $t(\mathcal{D}) = 2 \sum_{\alpha=1}^{N_{\rm bin}} \left[N_\alpha(\mathrm{R}) - O_\alpha + O_\alpha \log \frac{O_\alpha}{N_\alpha(\mathrm{R})} \right]$

recover **binned histogram test !**

Maximum Likelihood

[J. Neyman and E. S. Pearson, 1933]

Basic idea: $f(x;\mathbf{w}) = \mathrm{NN}$

replace histograms with NN, literally!

Highly motivated attempt:

- NN "effective" unbiased function approximants
- **NNPDF** exploits this virtue since 2002
- Often introduced as **alternative to histograms** to fit distributions
- Better dimensionality scaling

$$-2\,\underset{\mathbf{w}}{\mathrm{Min}}\left[N(\mathbf{w}) - N(\mathrm{R}) - \sum_{i=1}^{N_{\mathcal{D}}} f(x_i;\mathbf{w})\right]$$

Alternative in parametrised form: $n(x|\mathbf{w}) = n(x|\mathrm{R})\,e^{f(x;\mathbf{w})}$

If f piece-wise constant in bins: $t(\mathcal{D}) = 2\sum_{\alpha=1}^{N_{\mathrm{bin}}}\left[N_\alpha(\mathrm{R}) - O_\alpha + O_\alpha \log\frac{O_\alpha}{N_\alpha(\mathrm{R})}\right]$

recover **binned histogram test !**

Maximum Likelihood Loss

Easy to turn the evaluation of “t” into supervised training problem:

$$t(\mathcal{D}) = -2\,\underset{\mathbf{w}}{\mathrm{Min}}\left[N(\mathbf{w}) - N(\mathrm{R}) - \sum_{i=1}^{\mathcal{N}_{\mathcal{D}}} f(x_i;\mathbf{w})\right] \qquad N(\mathbf{w}) = \int dx\, n(x|\mathrm{R})\, e^{f(x;\mathbf{w})}$$

Maximum Likelihood Loss

Easy to turn the evaluation of "t" into supervised training problem:

$$t(\mathcal{D}) = -2\,\underset{\mathbf{w}}{\mathrm{Min}}\left[N(\mathbf{w}) - N(\mathrm{R}) - \sum_{i=1}^{\mathcal{N}_{\mathcal{D}}} f(x_i;\mathbf{w})\right] \qquad N(\mathbf{w}) = \int dx\, n(x|\mathrm{R})\, e^{f(x;\mathbf{w})}$$

Use **Reference Sample**, distributed according to the Reference Model

$$\mathcal{R} = \{x_i\},\ i = 1,\ldots,\mathcal{N}_{\mathcal{R}}$$

Maximum Likelihood Loss

Easy to turn the evaluation of “t” into supervised training problem:

$$t(\mathcal{D}) = -2\,\underset{\mathbf{w}}{\text{Min}}\left[N(\mathbf{w}) - N(\mathrm{R}) - \sum_{i=1}^{\mathcal{N}_{\mathcal{D}}} f(x_i;\mathbf{w})\right] \qquad N(\mathbf{w}) = \int dx\, n(x|\mathrm{R})\, e^{f(x;\mathbf{w})}$$

Use **Reference Sample**, distributed according to the Reference Model

$$\mathcal{R} = \{x_i\},\ i = 1,\ldots,\mathcal{N}_{\mathcal{R}}$$

Approximate integral as Monte Carlo sum:

$$N(\mathbf{w}) = \int dx\, n(x|\mathrm{R})\, e^{f(x;\mathbf{w})} = \frac{N(\mathrm{R})}{\mathcal{N}_{\mathcal{R}}} \sum_{x\in\mathcal{R}} e^{f(x;\mathbf{w})}$$

Maximum Likelihood Loss

Easy to turn the evaluation of "t" into supervised training problem:

$$t(\mathcal{D}) = -2\,\underset{\mathbf{w}}{\mathrm{Min}}\left[N(\mathbf{w}) - N(\mathrm{R}) - \sum_{i=1}^{\mathcal{N}_{\mathcal{D}}} f(x_i;\mathbf{w})\right] \qquad N(\mathbf{w}) = \int dx\, n(x|\mathrm{R})\, e^{f(x;\mathbf{w})}$$

Use **Reference Sample**, distributed according to the Reference Model

$$\mathcal{R} = \{x_i\},\ i = 1,\ldots,\mathcal{N}_{\mathcal{R}}$$

Approximate integral as Monte Carlo sum:

$$N(\mathbf{w}) = \int dx\, n(x|\mathrm{R})\, e^{f(x;\mathbf{w})} = \frac{N(\mathrm{R})}{\mathcal{N}_{\mathcal{R}}} \sum_{x\in\mathcal{R}} e^{f(x;\mathbf{w})}$$

Get **t = -2 * minimal loss.** The trained **net is distribution log ratio**

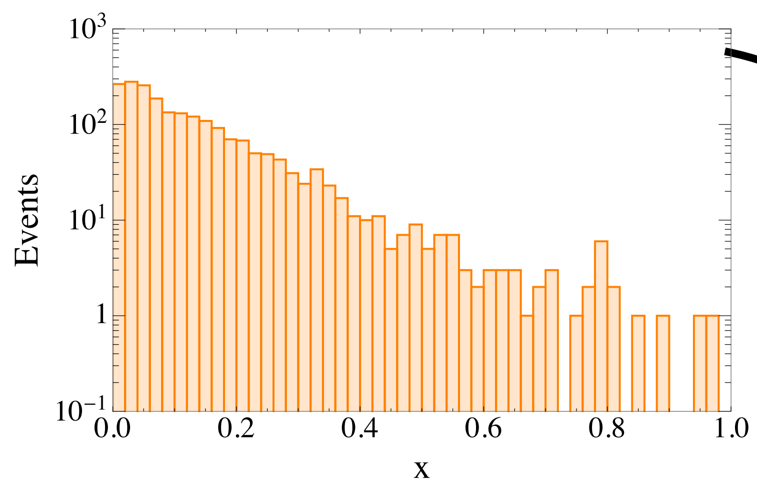
$$t(\mathcal{D}) = -2\,\underset{\{\mathbf{w}\}}{\mathrm{Min}}\left[\frac{N(\mathrm{R})}{\mathcal{N}_{\mathcal{R}}}\sum_{x\in\mathcal{R}}(e^{f(x;\mathbf{w})}-1) - \sum_{x\in\mathcal{D}} f(x;\mathbf{w})\right] \equiv -2\,\underset{\{\mathbf{w}\}}{\mathrm{Min}}\, L[f(\,\cdot\,,\mathbf{w})]$$

$$L[f] = \sum_{(x,y)}\left[(1-y)\frac{N(\mathrm{R})}{\mathcal{N}_{\mathcal{R}}}(e^{f(x)}-1) - y\, f(x)\right]$$

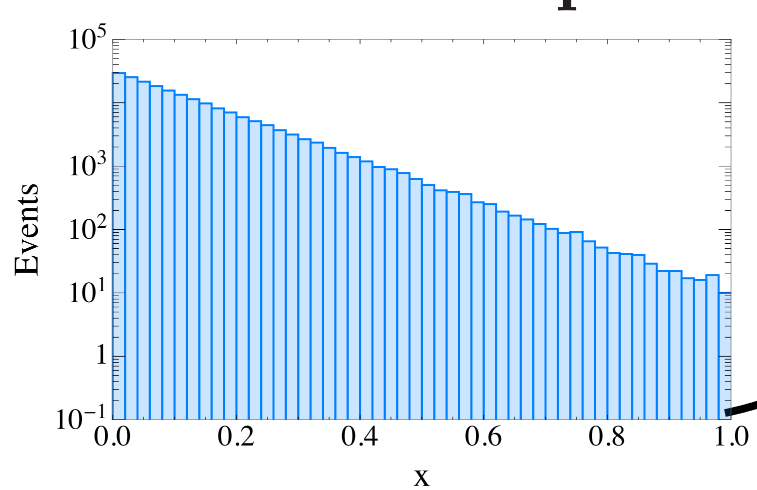
The Algorithm

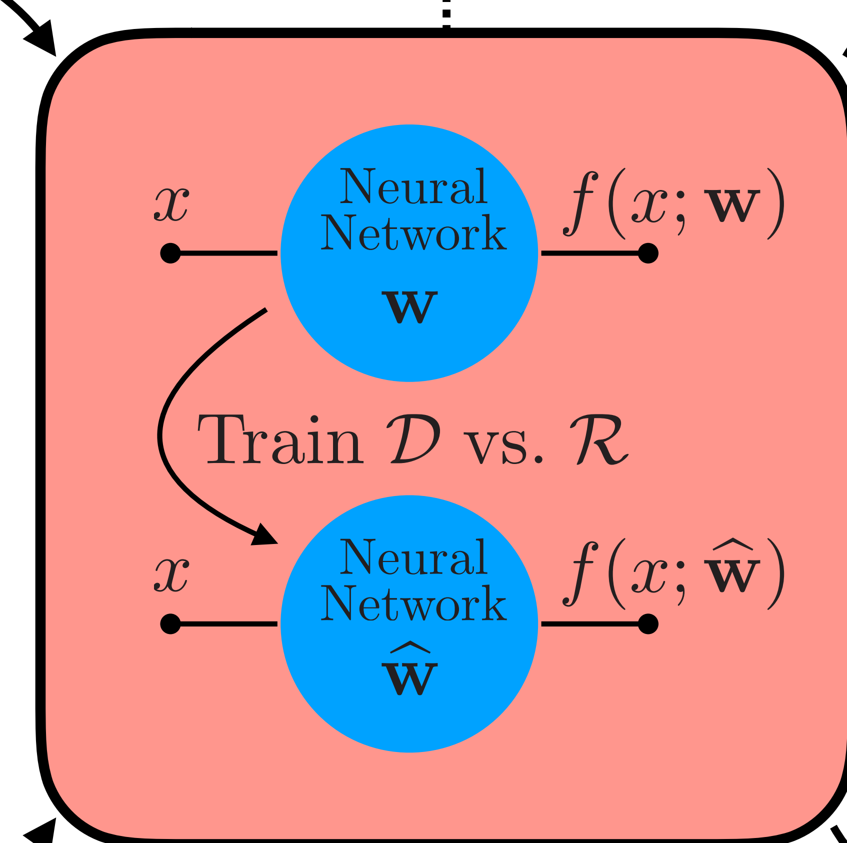
INPUT

Data sample $\mathcal{D}$



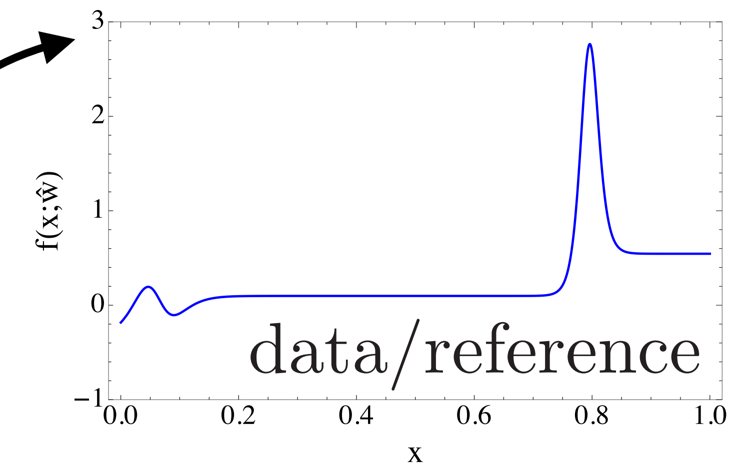
Reference sample $\mathcal{R}$





OUTPUT

Dist. log ratio



$$f(x;\widehat{\mathbf{w}}) \simeq \log\left[\frac{n(x|\mathrm{T})}{n(x|\mathrm{R})}\right]$$

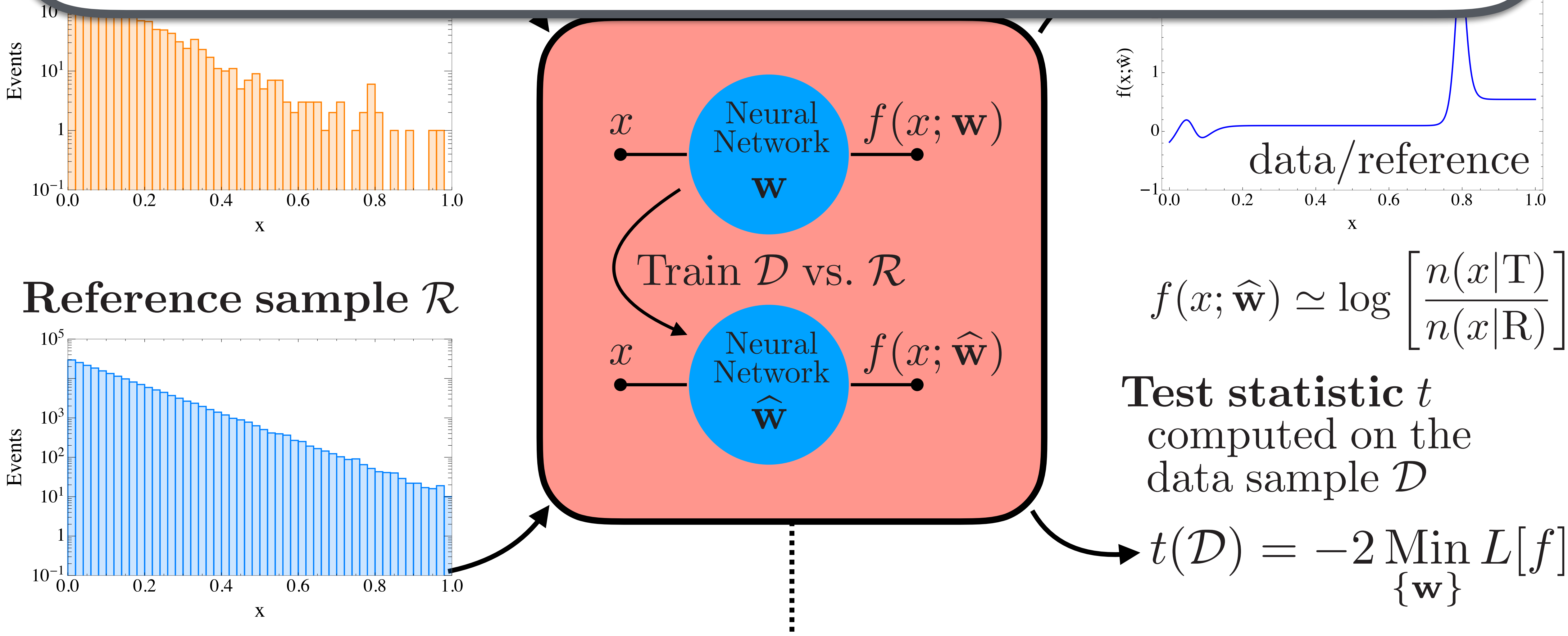
Test statistic t computed on the data sample $\mathcal{D}$

$$t(\mathcal{D}) = -2\,\underset{\{\mathbf{w}\}}{\mathrm{Min}}\,L[f]$$

The Algorithm

Non-Neymann—Pearson formulation:

- **learn likelihood ratio**, use it for test
- other loss functions can be used, connection with lik.free inf.
- Neymann—Pearson loss performs [a bit] better in our examples

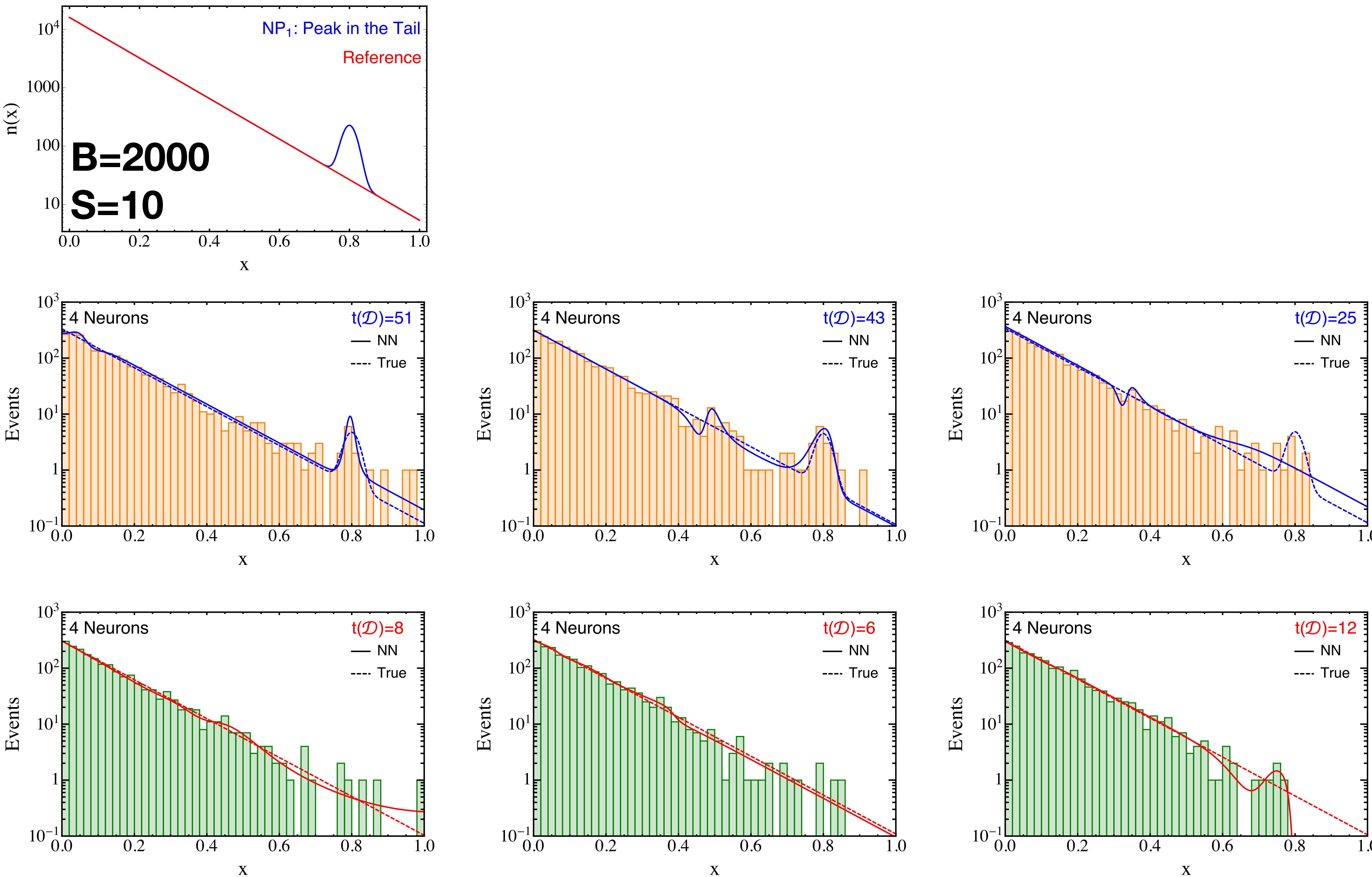


$$f(x;\widehat{\mathbf{w}}) \simeq \log\left[\frac{n(x|\mathrm{T})}{n(x|\mathrm{R})}\right]$$

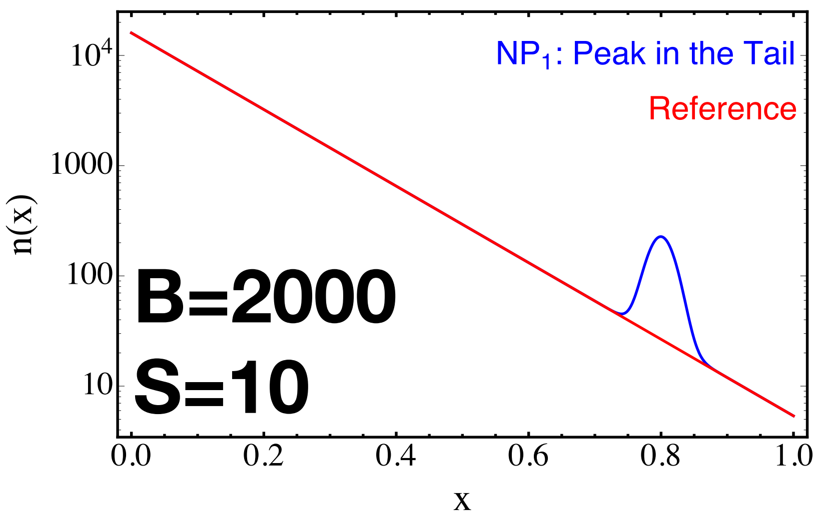
Test statistic t computed on the data sample $\mathcal{D}$

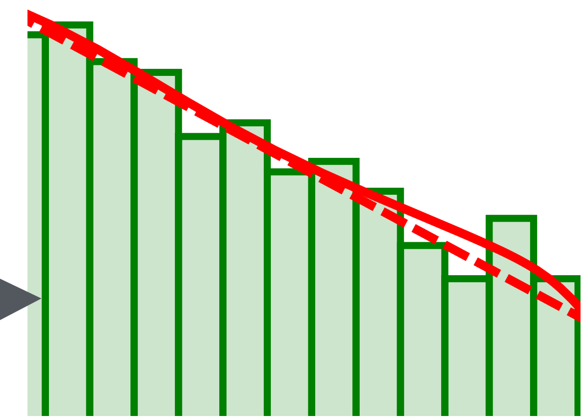
$$t(\mathcal{D}) = -2\,\underset{\{\mathbf{w}\}}{\mathrm{Min}}\, L[f]$$

Illustration



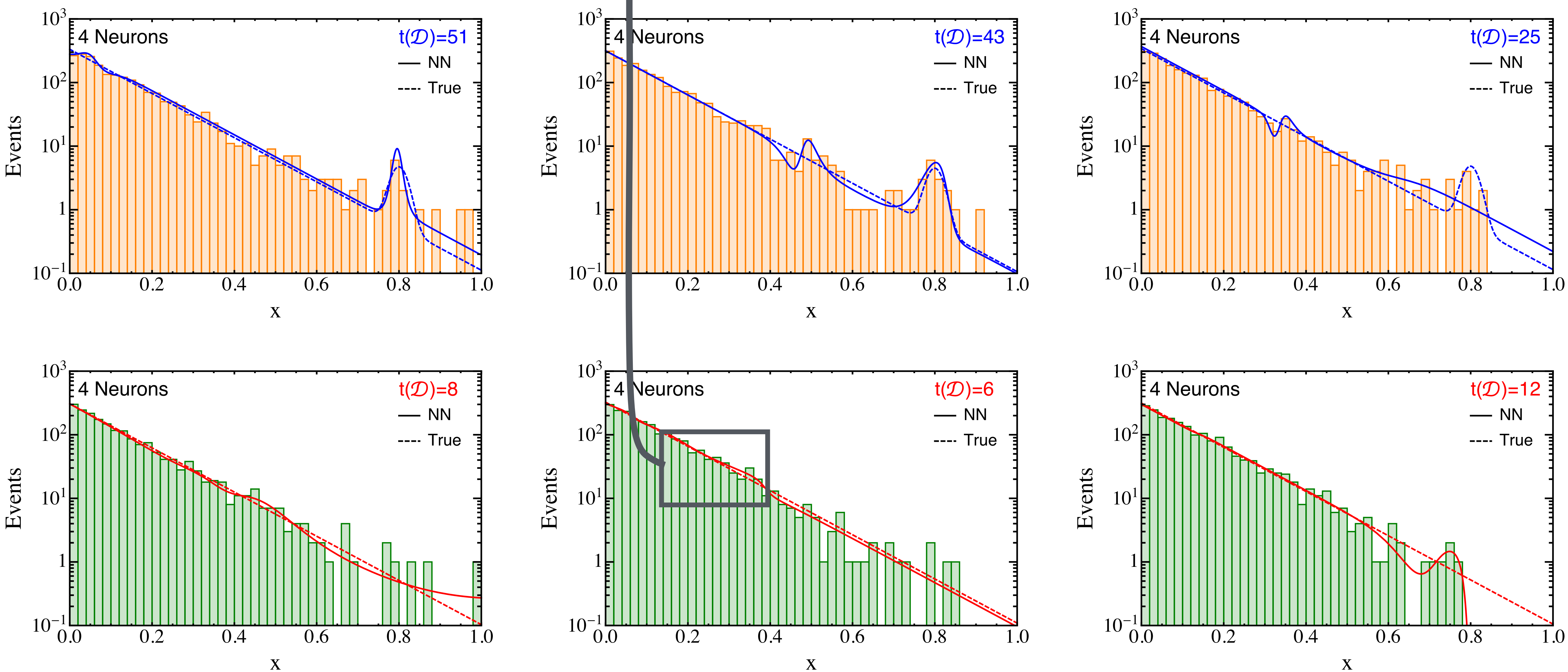
Illustration





Bins: Non-discrepant data fluctuations wash out reach

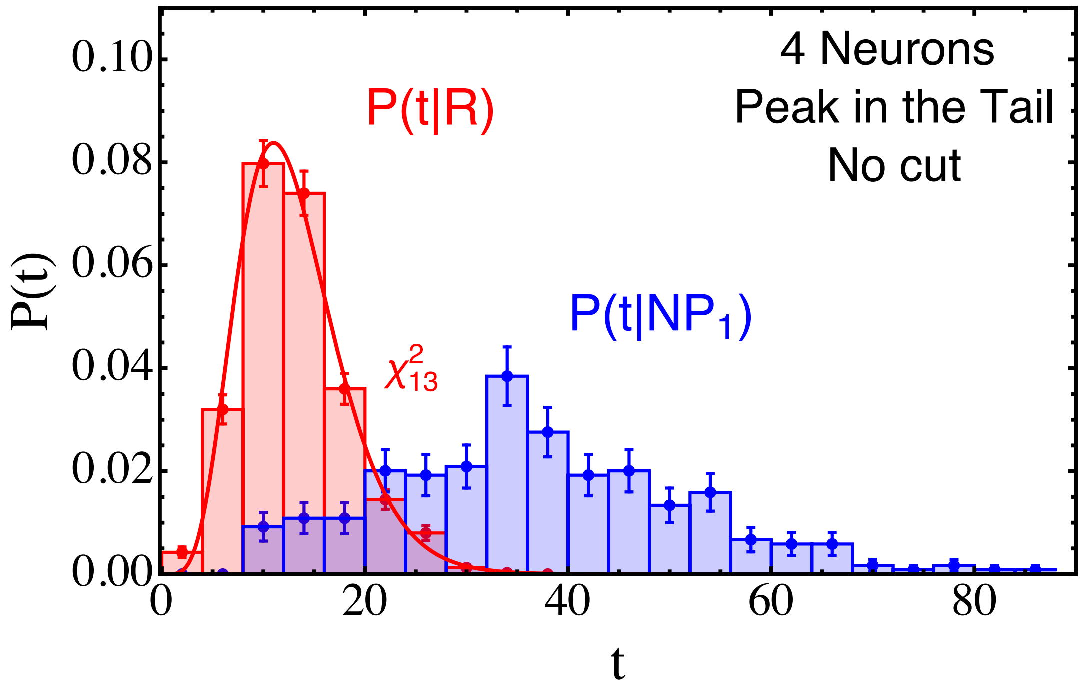
NN: Smooth curve. Can handle non-discrepant data



Quantifying Performances

Our Z-score vs **Ideal** Z-score

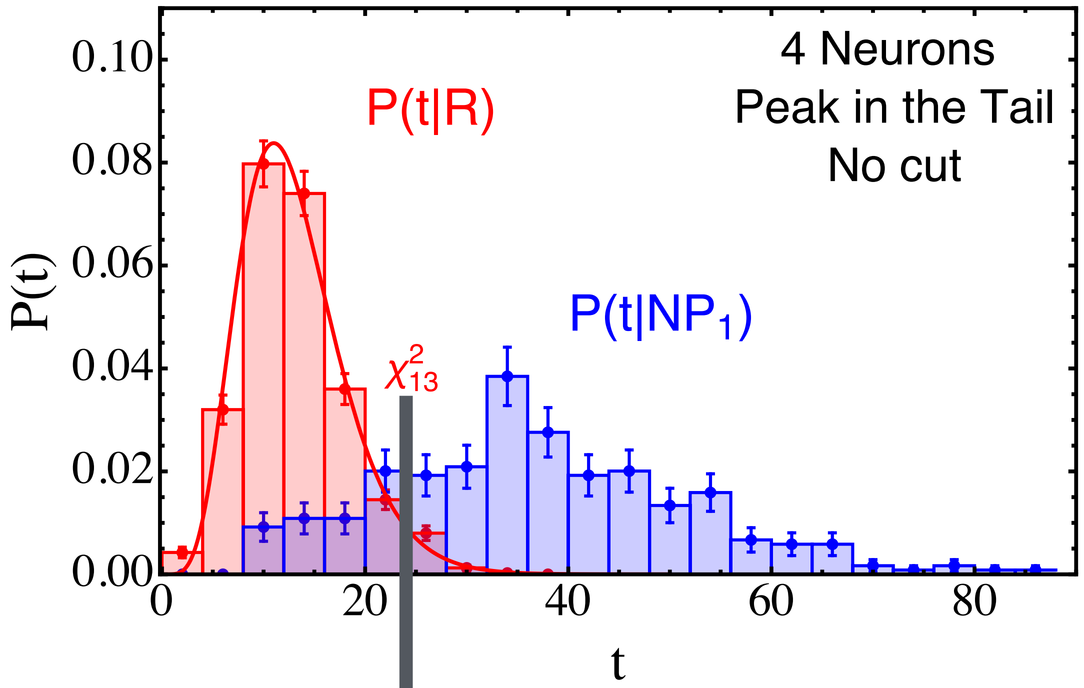
Run over R and NP toys [repeat train.]
Compute NP p-value distribution



Quantifying Performances

Our Z-score vs **Ideal** Z-score

Run over R and NP toys [repeat train.]
Compute NP p-value distribution

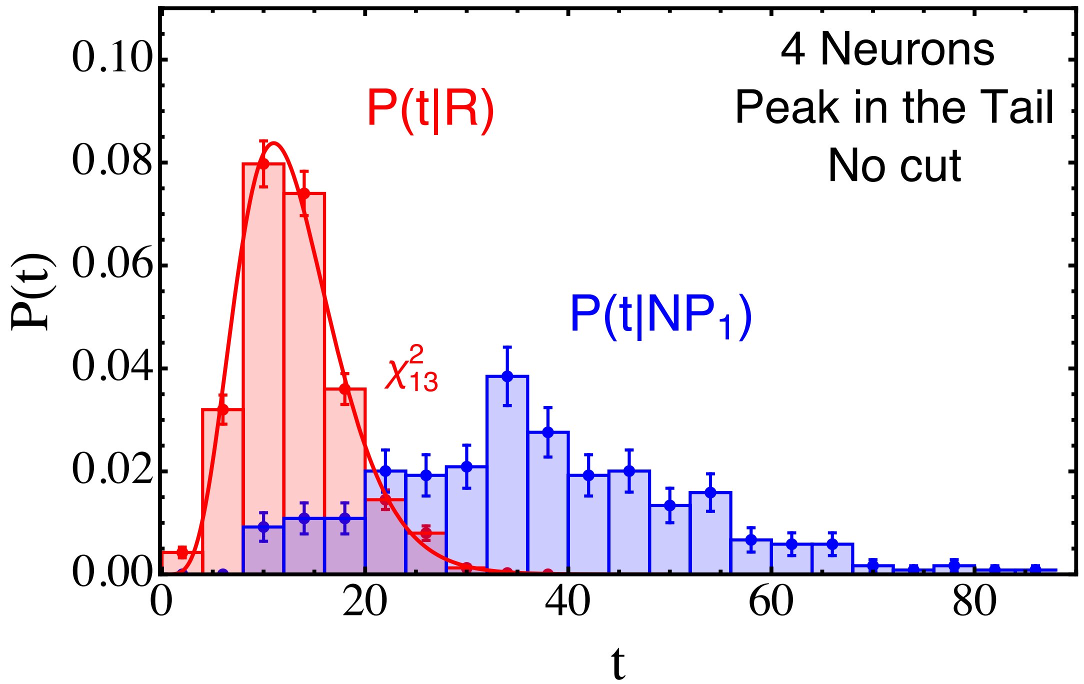


Agreement with χ^2 dof=#NNpar.
Expected for ML in As.Lim.

Quantifying Performances

Our Z-score vs **Ideal** Z-score

Run over R and NP toys [repeat train.]
Compute NP p-value distribution

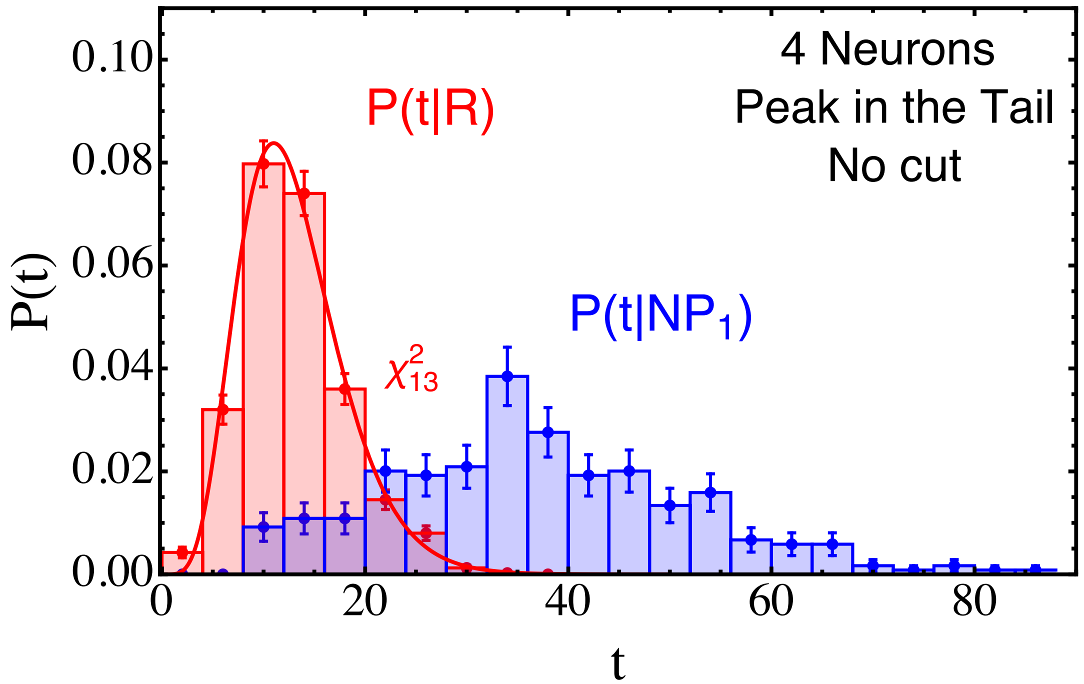


How difficult is to see “NP”?

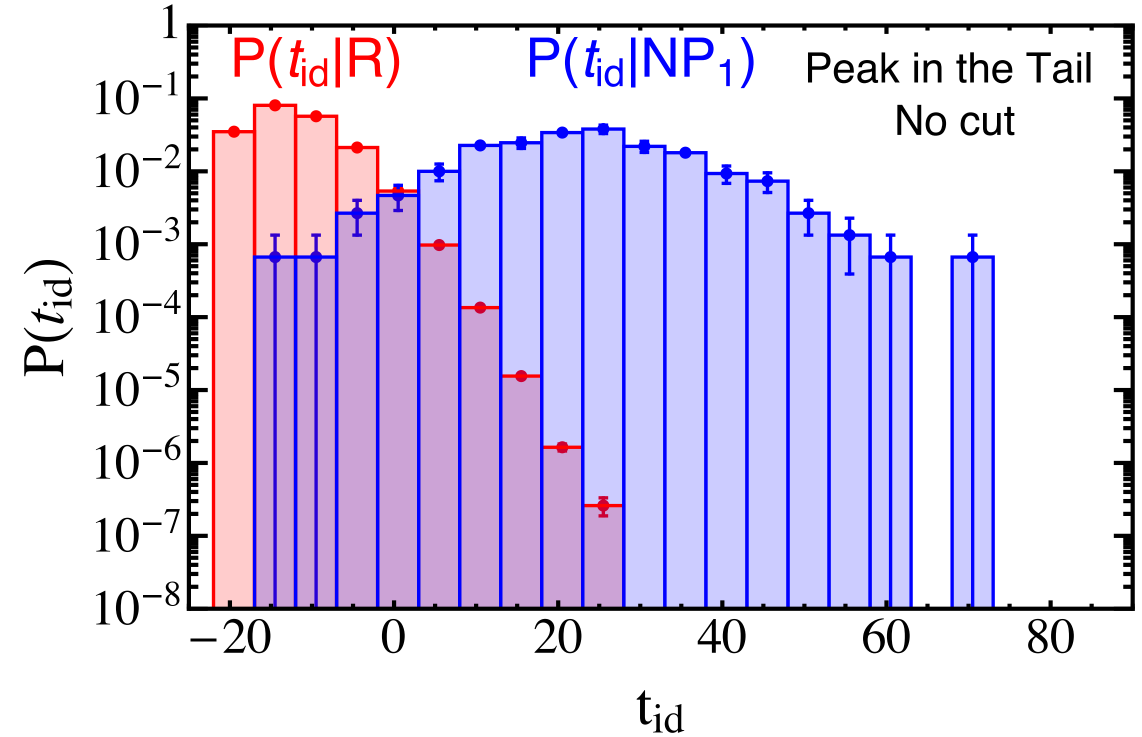
Quantifying Performances

Our Z-score vs **Ideal** Z-score

Run over R and NP toys [repeat train.]
Compute NP p-value distribution



How difficult is to see “NP”?
Compute ideal p-value [Ney.-Pea.]

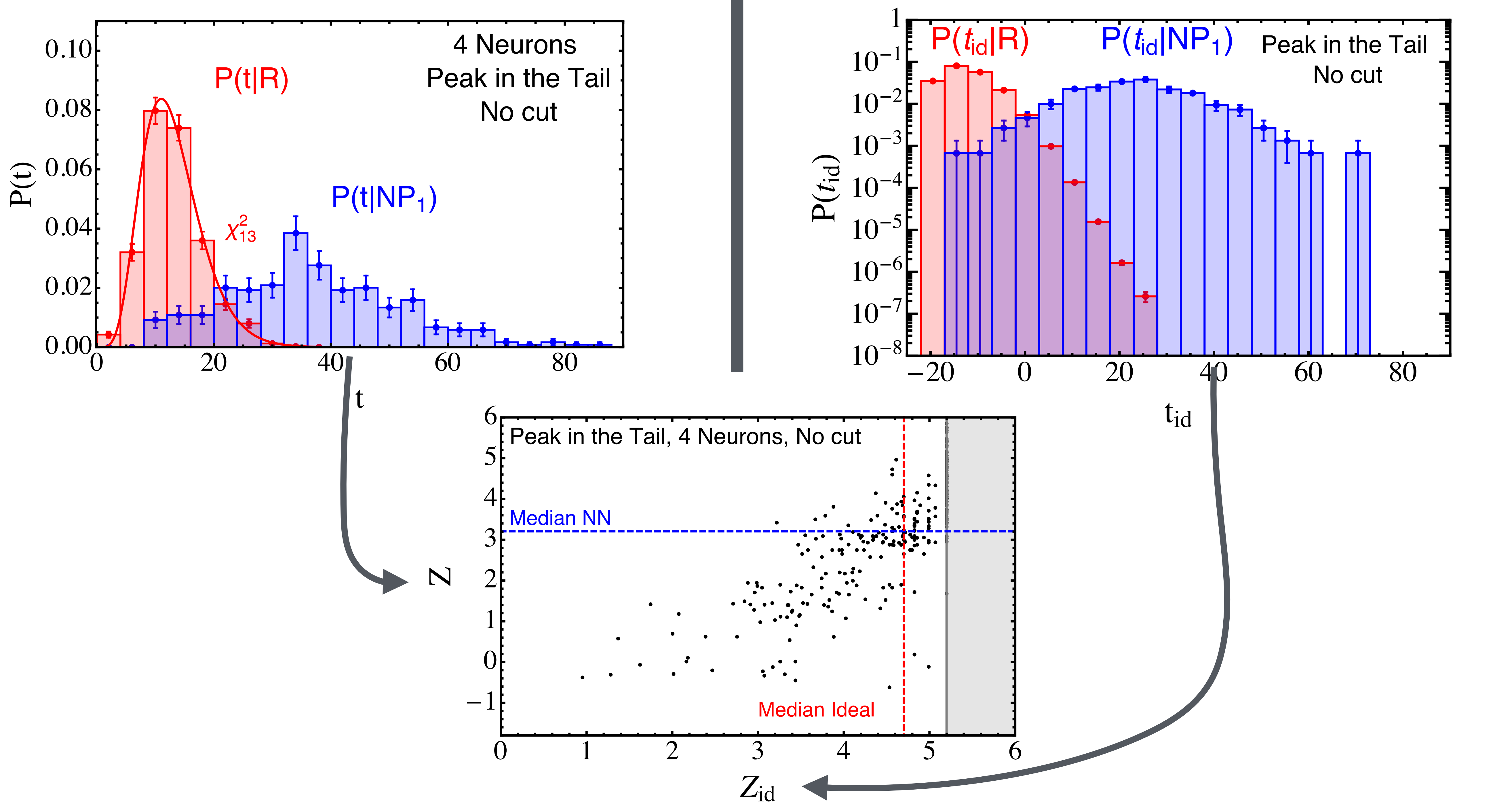


Quantifying Performances

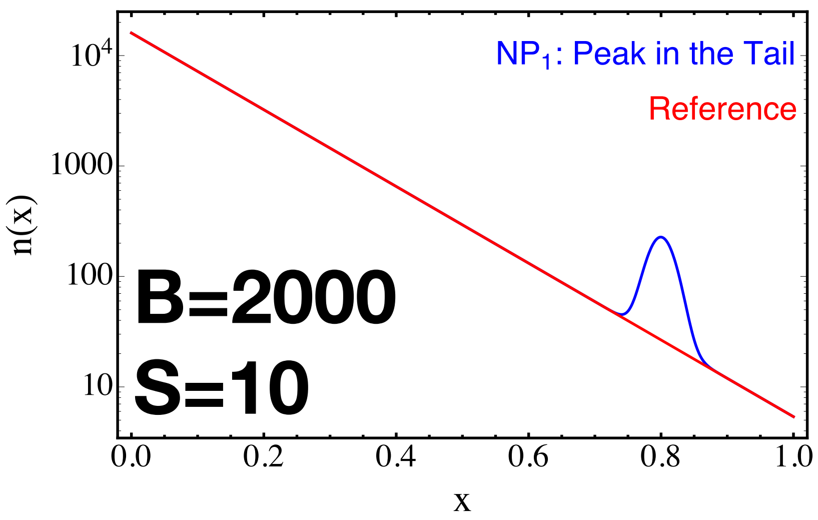
Our Z-score vs **Ideal** Z-score

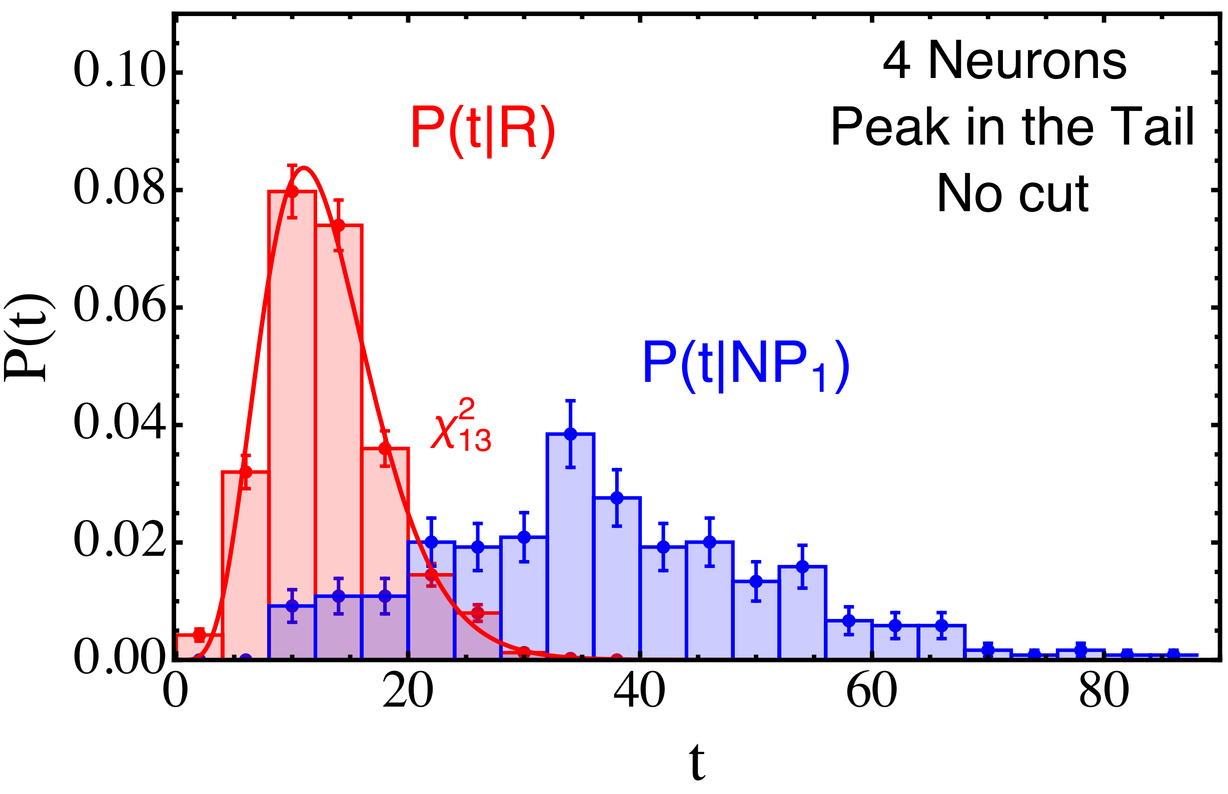
Run over R and NP toys [repeat train.]
Compute NP p-value distribution

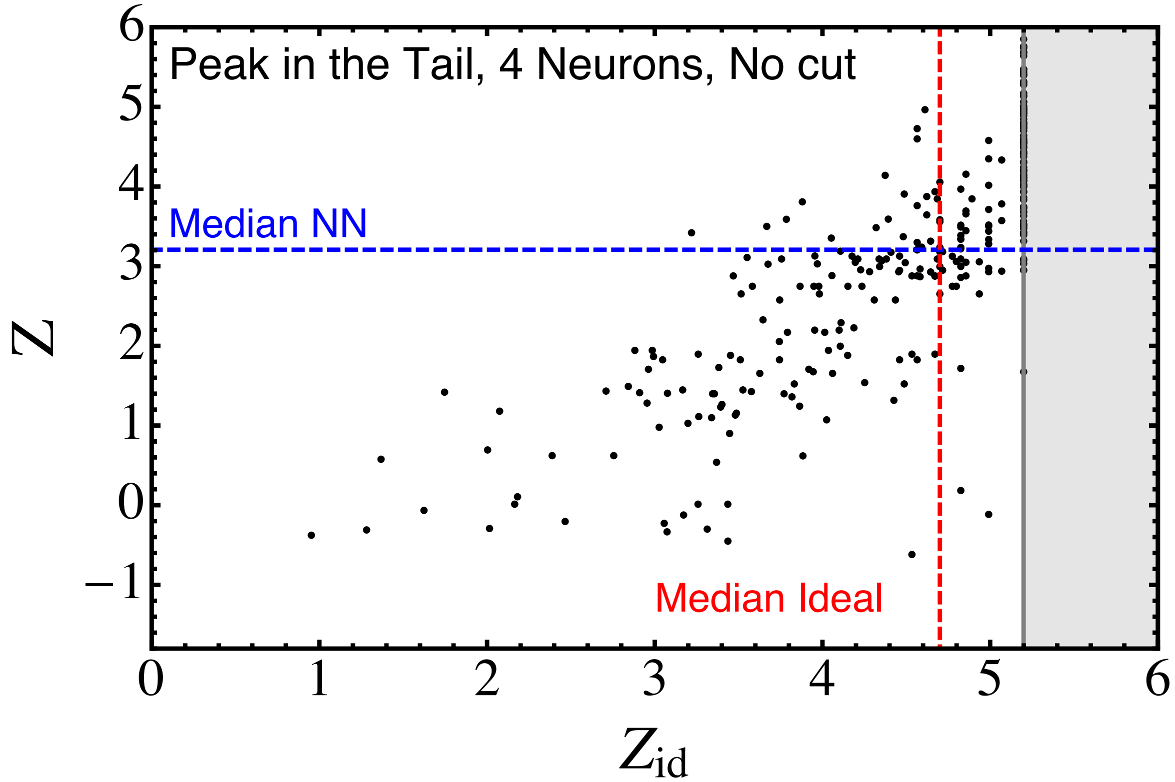
How difficult is to see “NP”?
Compute ideal p-value [Ney.-Pea.]



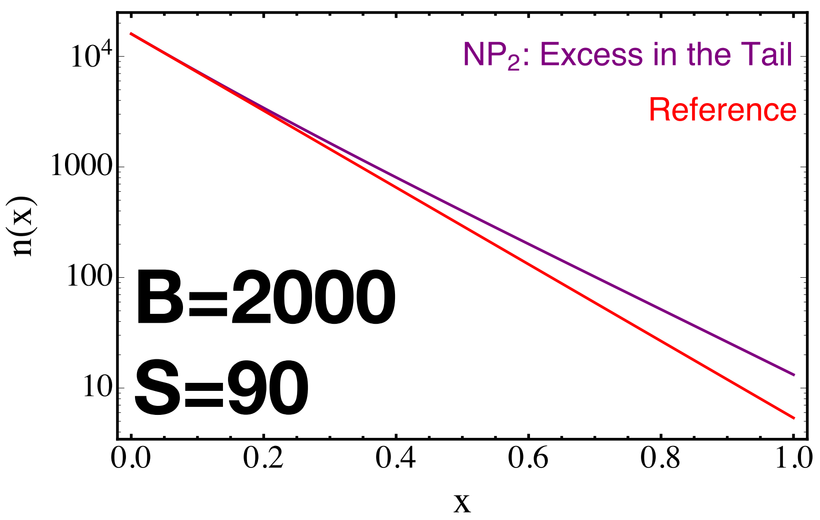
Quantifying Performances: NP_1

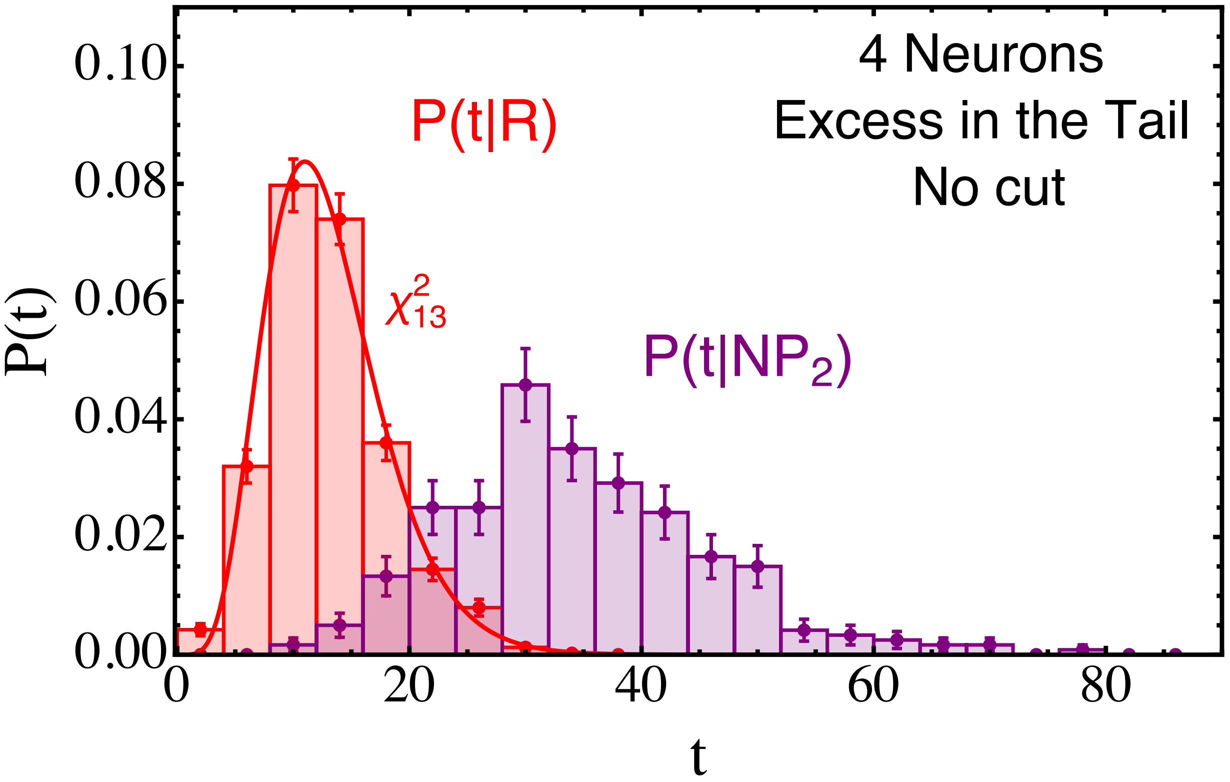


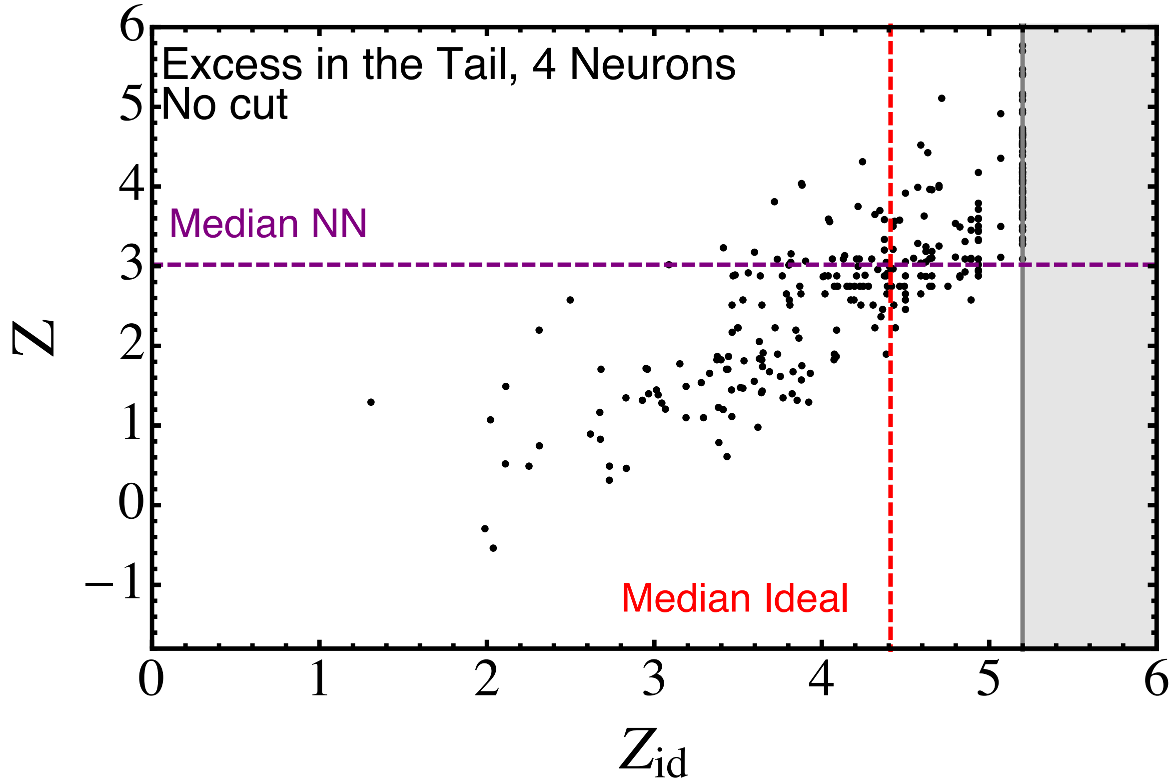




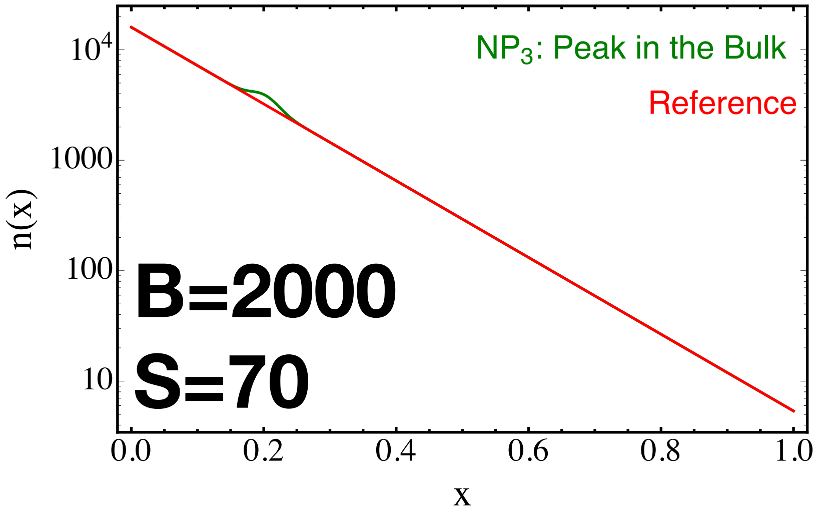
Quantifying Performances: NP_2

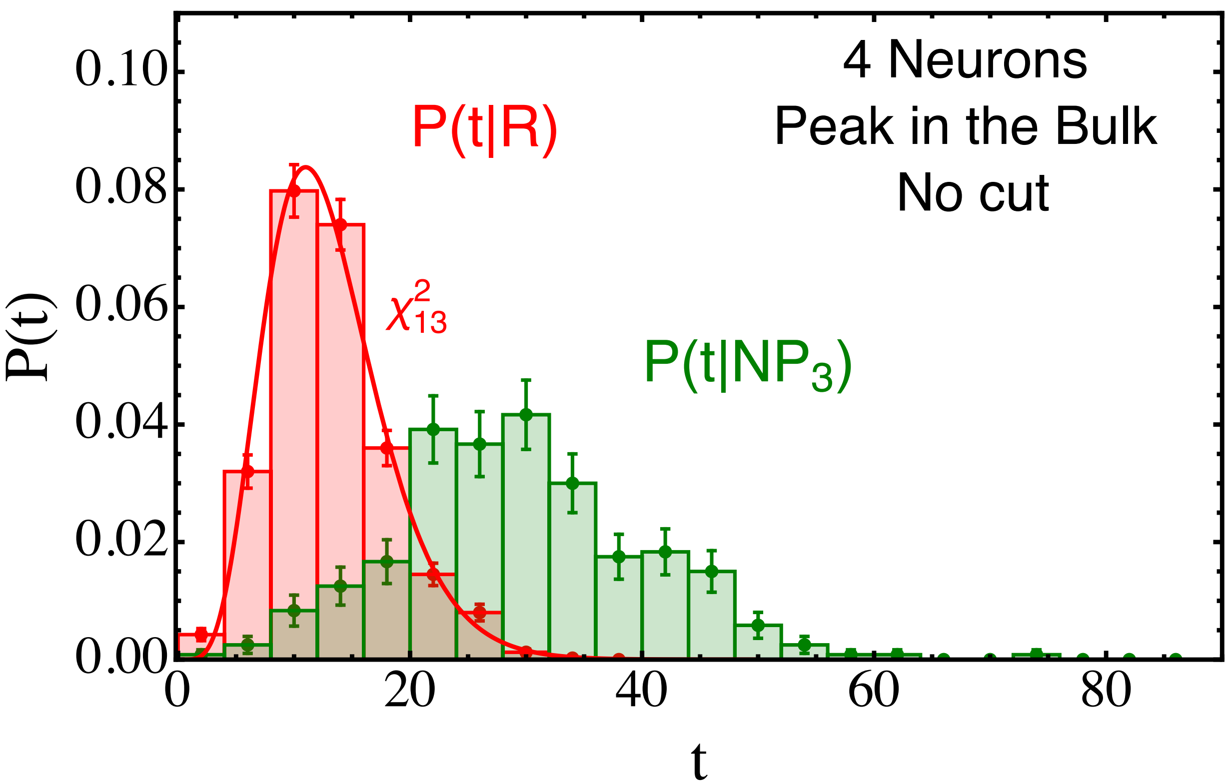


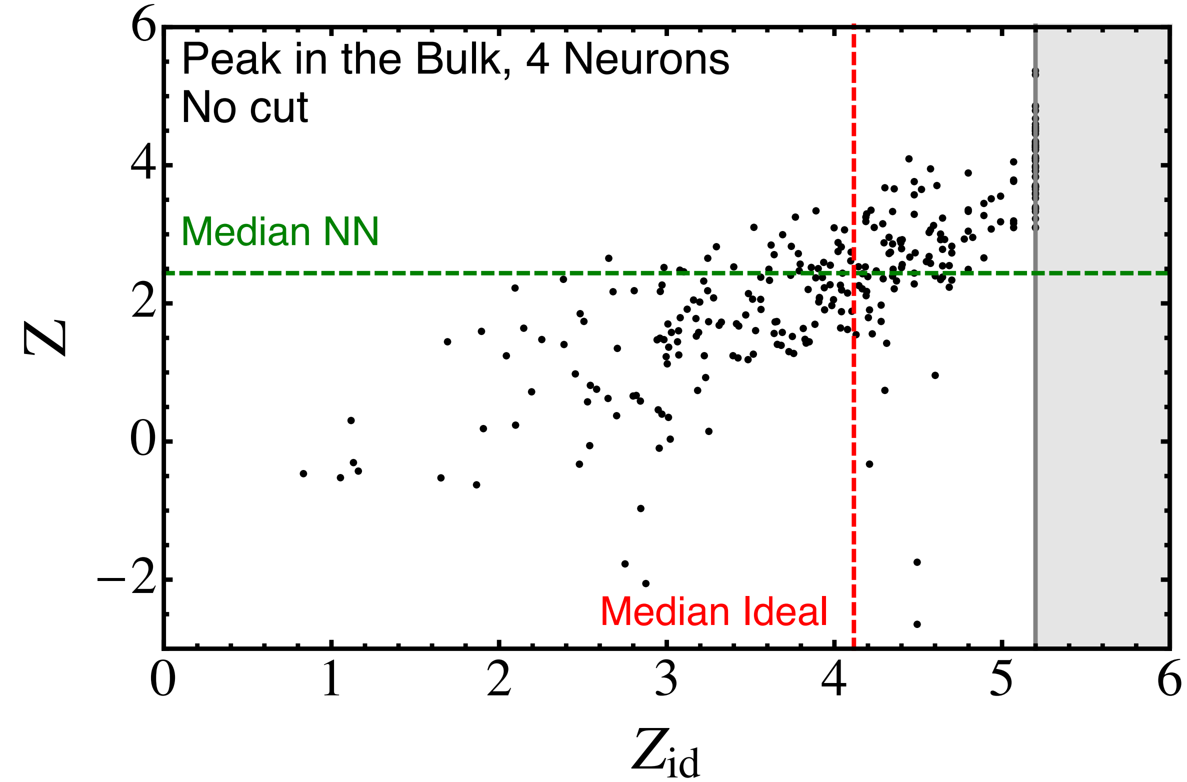




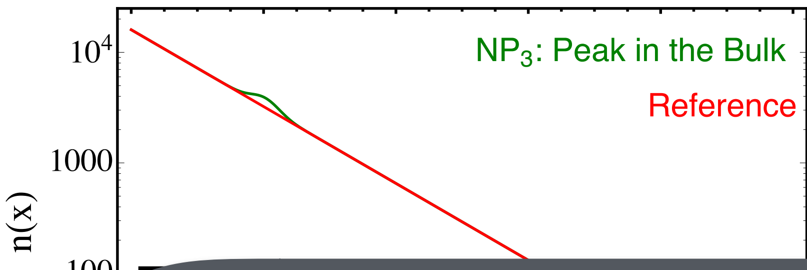
Quantifying Performances: NP_3



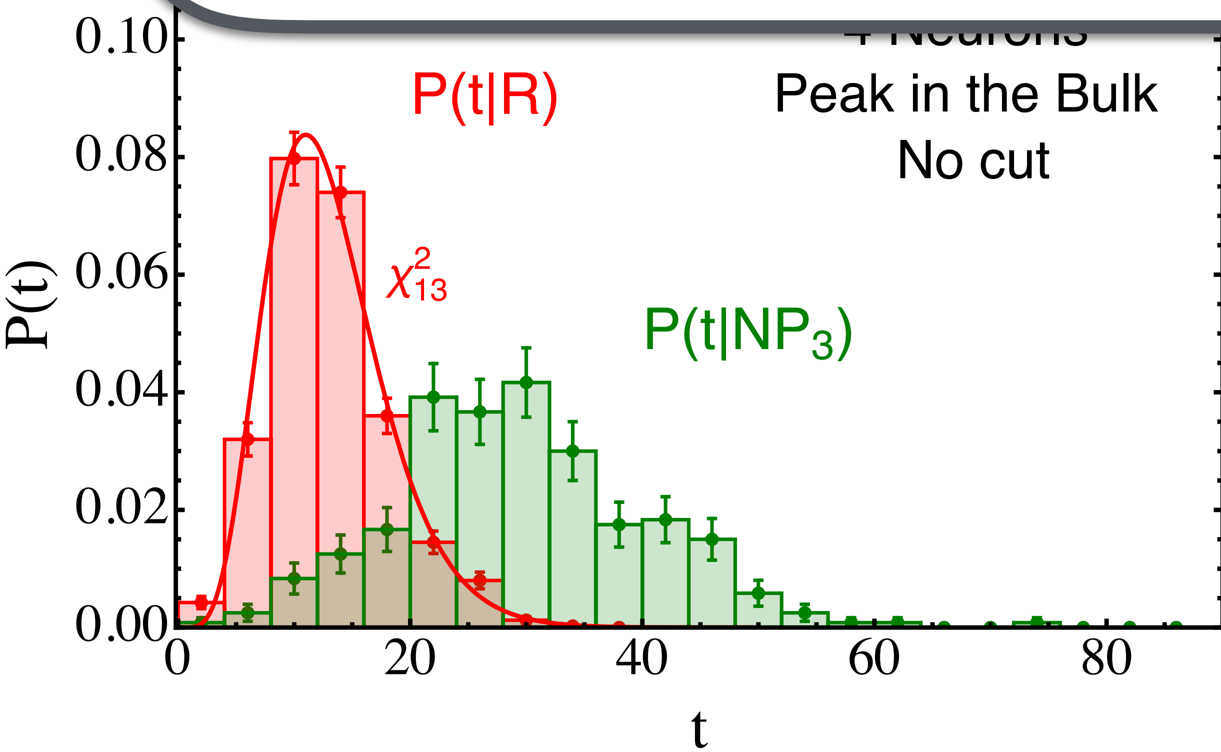


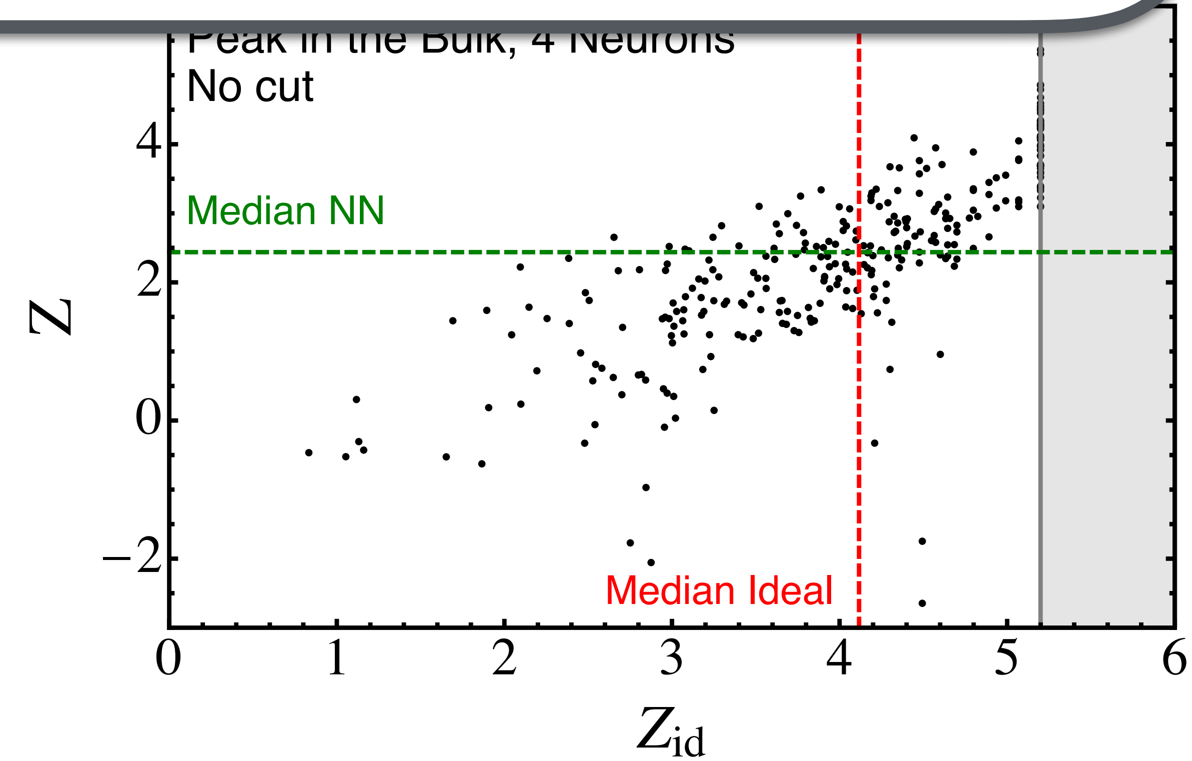


Quantifying Performances: NP_3



Very Model-Independent !!

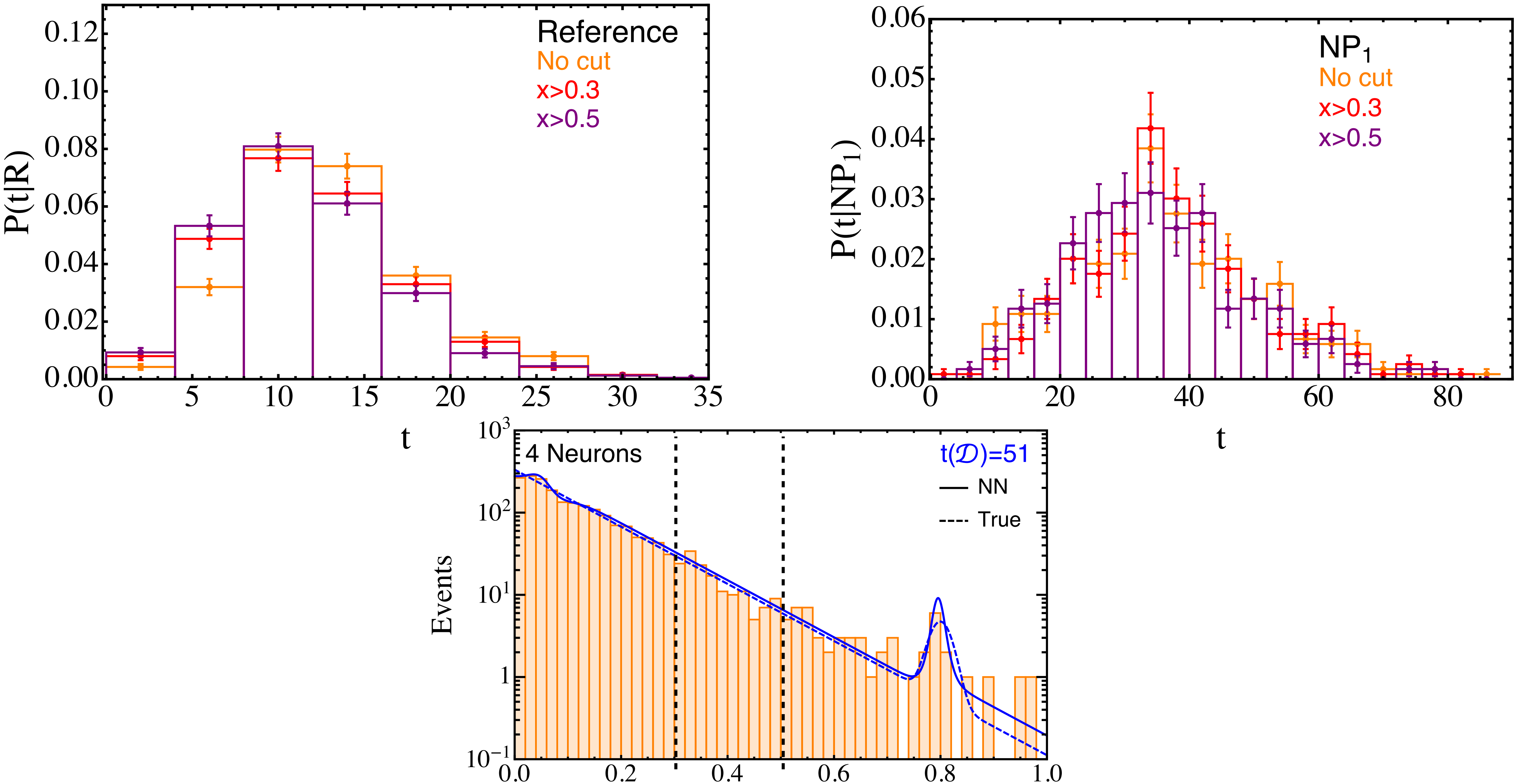




Other features

(In)-Sensitivity to Cuts:

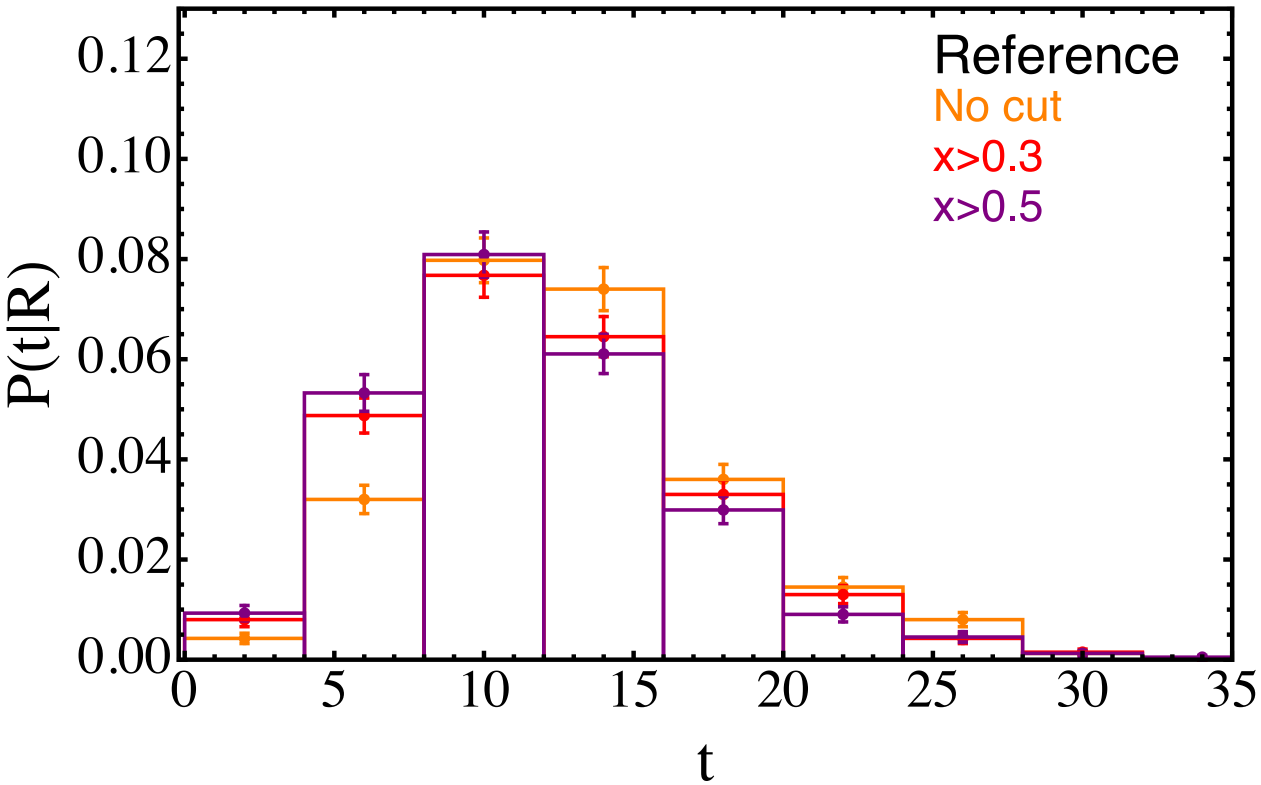
Unlike binned histogram, **NN reach not affected by signal-free data**

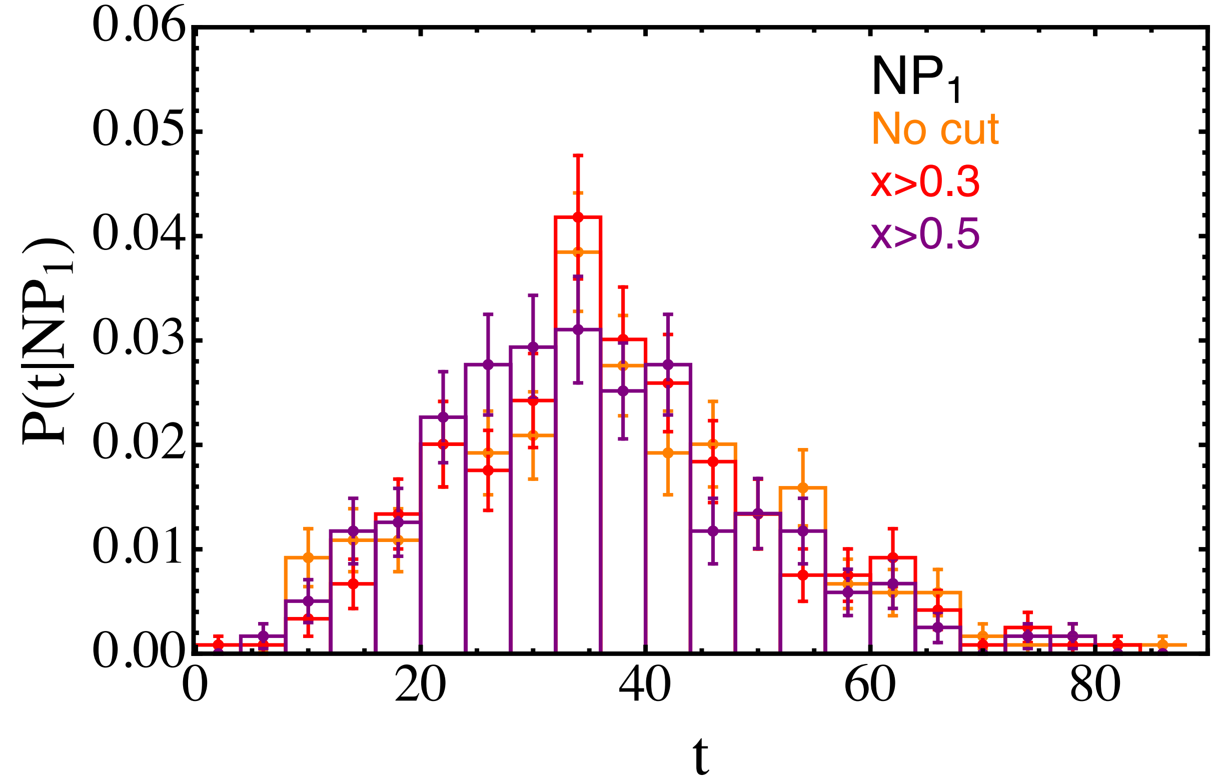


Other features

(In)-Sensitivity to Cuts:

Unlike binned histogram, **NN reach not affected by signal-free data**





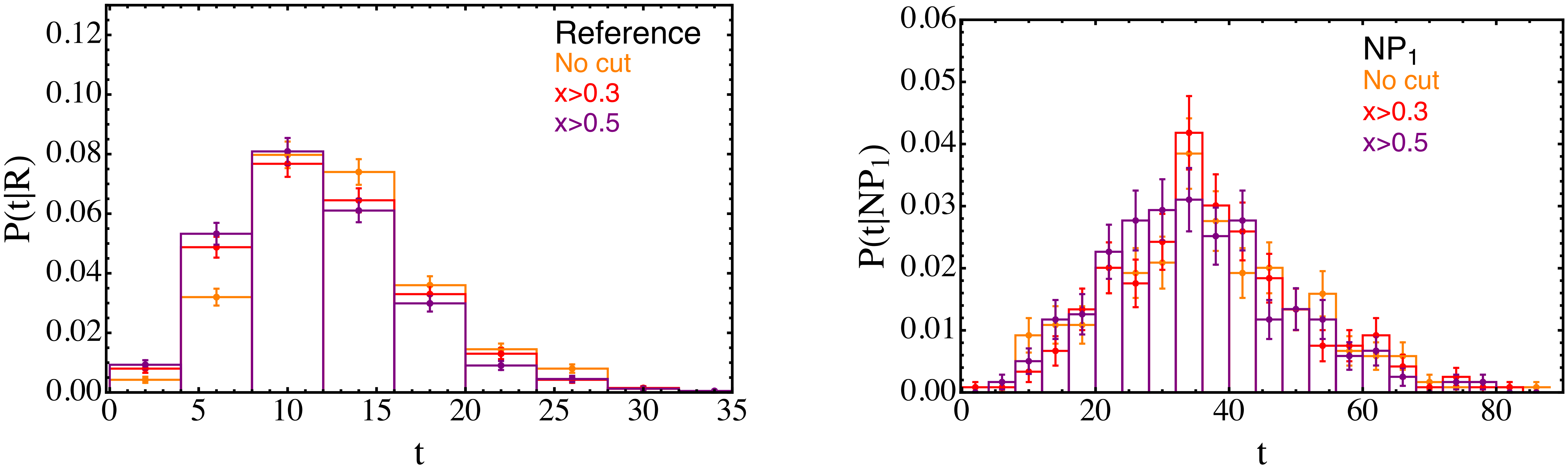
Mild Sensitivity to Hyperparameters

But larger Networks much more difficult to train

Other features

(In)-Sensitivity to Cuts:

Unlike binned histogram, **NN reach not affected by signal-free data**



Mild Sensitivity to Hyperparameters

But larger Networks much more difficult to train

Significant degradation with dimensionality

To be expected, but how it scales with “d”?

Pending Issues

"Easy" ones:

- **Include systematics** in Reference data (MC or from control region). Seemingly straightforward to treat them as nuisance parameters
- Reduce Reference sample size by weighting.

Pending Issues

"Easy" ones:

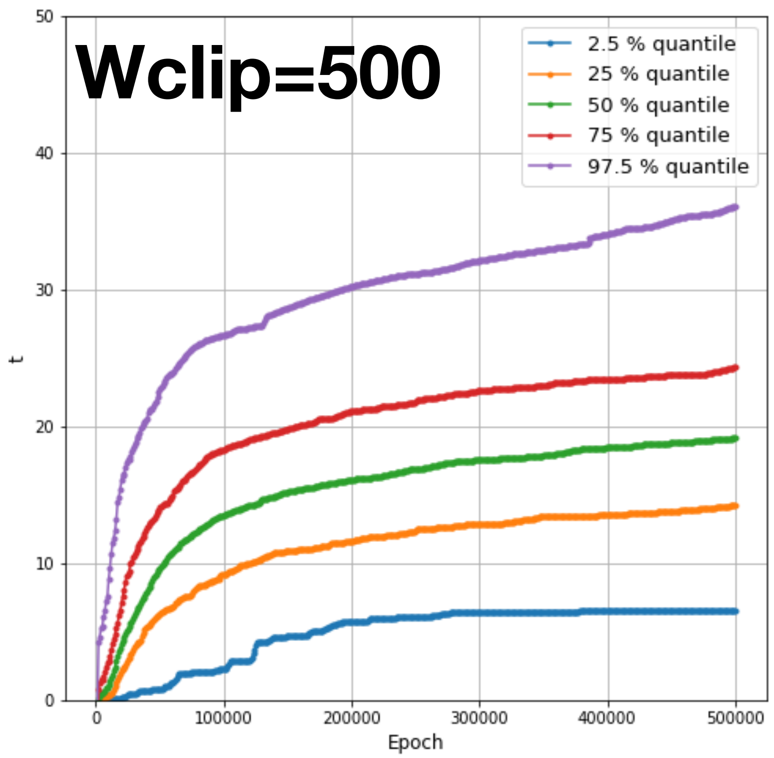
- **Include systematics** in Reference data (MC or from control region). Seemingly straightforward to treat them as nuisance parameters
- Reduce Reference sample size by weighting.

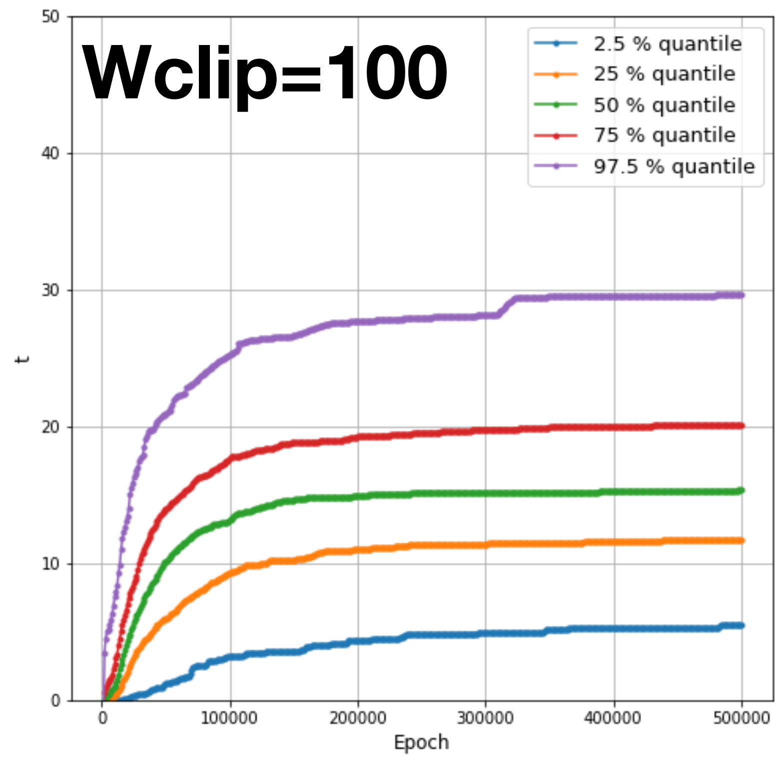
Hard one: **Model Selection**

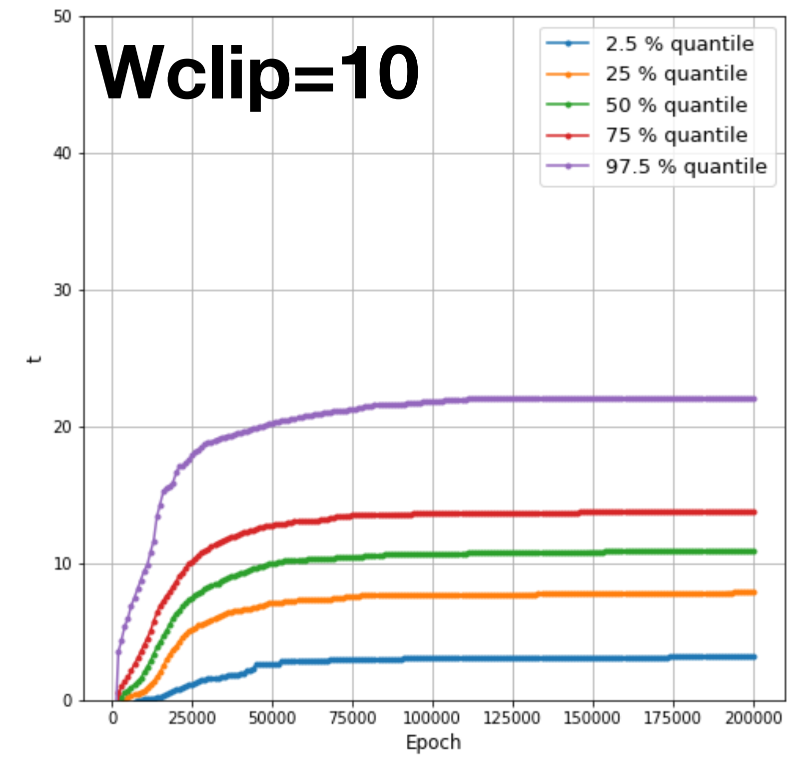
- Even if sensitivity was mild in our examples, how to choose NN architecture/WeightClipping [or reg.]?
- **Heuristic approach:** more capacity is better, bound from training convergence in finite time and from **agreement with** χ^2
- **Mathematical approach(?):** apply NN convergence theorems?

Towards Model Selection

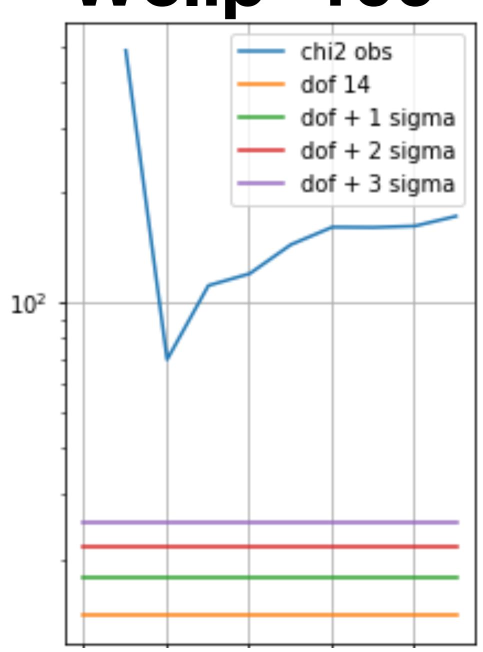
[Cerri, **Gaia Grosso**, D'Agnolo, Pierini, AW, Zanetti in progress]



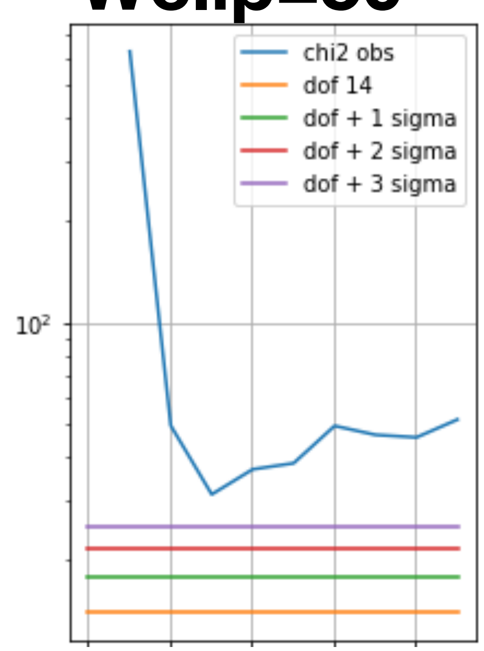




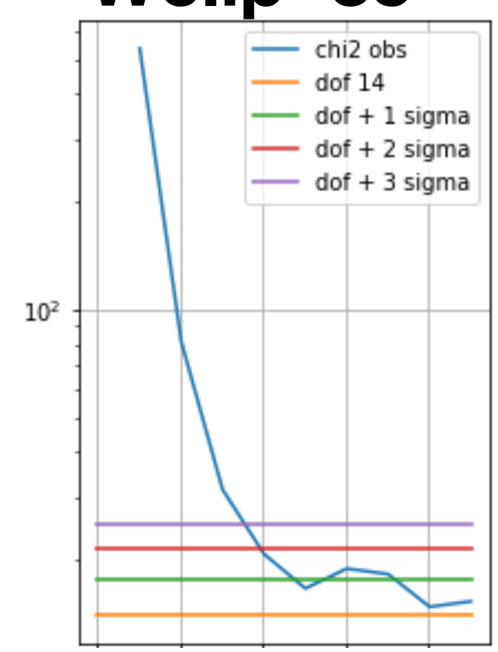
Wclip=100



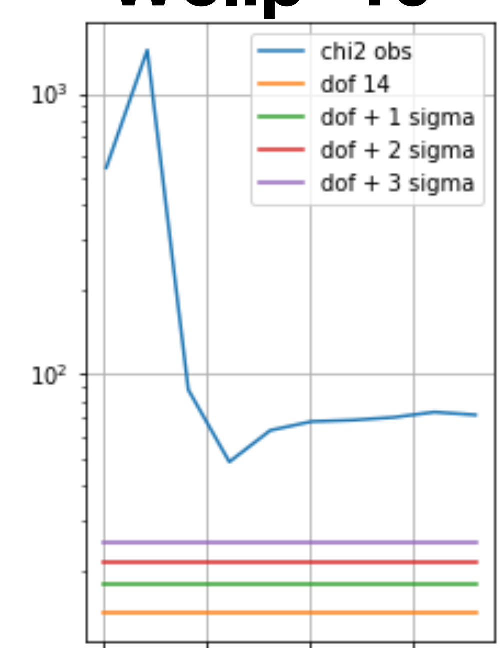
Wclip=50



Wclip=35



Wclip=10



Other Approaches

CWoLa Hunting: **[Collins, Howe, Nachman: arXiv:1805.02664]**

Data/Reference regions selected by mass-window (like BumpHunter)

NN learns Data/Reference distribution ratio of additional variables

Ratio provides additional discriminant and improves BumpHunter reach

Novelty Detection: **[Hajer et al.: arXiv:1807.10261; Pierini et al., in progress]**

Slightly different: we don't necessarily care of "rare" SM events

Non-QCD jets: **[Aguilar–Saavedra et al.: arXiv:1709.01087,Heimel et al.: arXiv:1808.08979]**

Gaussian Mixture pdf: **[Kuusela et al.: arXiv:1112.3329]**

Use Gaussian Mixture pdf estimate for Data and for Reference

Nearest-Neighbours pdf: **[De Simone, Jacques: arXiv:1807.06038]**

Use Nearest-Neighbours pdf estimate for Data and for Reference

Few More Comments

Model-Independent search algorithms also good for:

- Comparison between different Monte Carlo Generators
- Data Validation

Few More Comments

Model-Independent search algorithms also good for:

- Comparison between different Monte Carlo Generators
- Data Validation

When and if these techniques make it to real analyses, I suspect we will find plenty of wrong Monte Carlos …

Few More Comments

Model-Independent search algorithms also good for:

- Comparison between different Monte Carlo Generators
- Data Validation

When and if these techniques make it to real analyses, I suspect we will find plenty of wrong Monte Carlos …

But maybe we will find New Physics as well !!

Few More Comments

Model-Independent search algorithms also good for:

- Comparison between different Monte Carlo Generators
- Data Validation

When and if these techniques make it to real analyses,
I suspect we will find plenty of wrong Monte Carlos …

But maybe we will find New Physics as well !!

Thank You