# Goodness-of-fit by Neyman-Pearson Testing: The NPLM Method

Andrea Wulzer

ICREA @ IFAE
Institut de Física d'Altes Energies

Based on:

D'Agnolo, AW, 2018
D'Agnolo, Grosso, Pierini, AW, Zanetti, 2019
D'Agnolo, Grosso, Pierini, AW, Zanetti, 2021
Grosso, Letizia, AW, et. al., 2022
Grosso, Letizia, AW, Zanetti, et. al., 2023
Grosso, Letizia, Pierini, AW, 2023
Grosso, 2024
Grosso, Letizia, 2024
Cappelli, Grosso, Letizia, Reyes-González, Zanetti, to appear

# Neyman-Pearson Testing

An **Hypothesis** H in Statistics is a p.d.f. according to which data might be distributed:

$$\mathrm{H} \;\leftrightarrow\; P_{\mathrm{H}}(\mathrm{data})$$

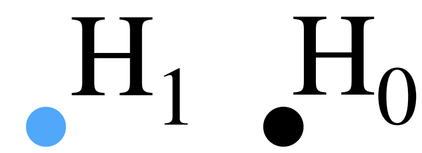
The **Likelihood** is probability seen as function of the hypothesis, and not of the data:

$$\mathscr{L}(\mathrm{H}) = P_{\mathrm{H}}(\mathrm{data})$$

A **Test of Hypothesis** is a **comparative statement** on the relative plausibility of **two Hypotheses** as distribution of a data instance.

# Neyman-Pearson Testing

For two **simple** Hypotheses $H_0$ vs $H_1$:



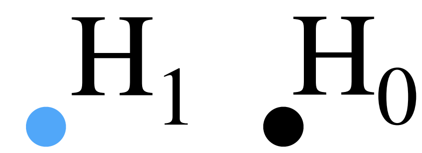

N&P found the **best test**, the one with highest chance to falsify $H_0$ if $H_1$ is true, and viceversa. The **Neyman–Pearson lemma:**

“ *The* ***best test*** *employs as test statistics the variable t:* “

$$t = 2 \log \frac{\mathscr{L}(H_1)}{\mathscr{L}(H_0)}$$

# Neyman-Pearson Testing

For two **simple** Hypotheses $H_0$ vs $H_1$:

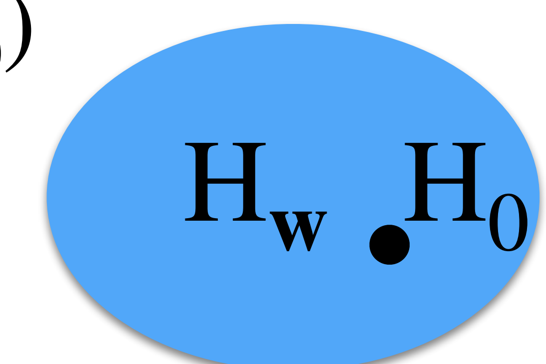



N&P found the **best test**, the one with highest chance to falsify $H_0$ if $H_1$ is true, and viceversa. The **Neyman–Pearson lemma:**

“ *The* ***best test*** *employs as test statistics the variable t:* “

$$t = 2\log\frac{\mathscr{L}(H_1)}{\mathscr{L}(H_0)}$$

For simple vs **composite** Hypothesis:




“Best” test unknown, but **good test** is **Maximum Likelihood**:

“ *The* ***ML test*** *employs as test statistics the variable* $t_{ML}$*:* “

$$t_{ML} = 2\max_{\mathbf{w}}\log\frac{\mathscr{L}(H_{\mathbf{w}})}{\mathscr{L}(H_0)} = 2\log\frac{\mathscr{L}(H_{\hat{\mathbf{w}}})}{\mathscr{L}(H_0)}$$

# Goodness of Fit

Statisticians formulate an interesting problem: **g.o.f.***

Be $\mathscr{D}$ some data, and R **one hypothesis** for their distribution

Does R provide the **right description** of $\mathscr{D}$ ?

**Not** a problem of Hypothesis testing, as only one hyp. involved.

**But, it can be addressed** by performing an HT, with $H_0 = R$ .

*often question emerges after optimising distribution free parameters on the data, as a way to assess fit quality. But the problem is more general

# Goodness of Fit

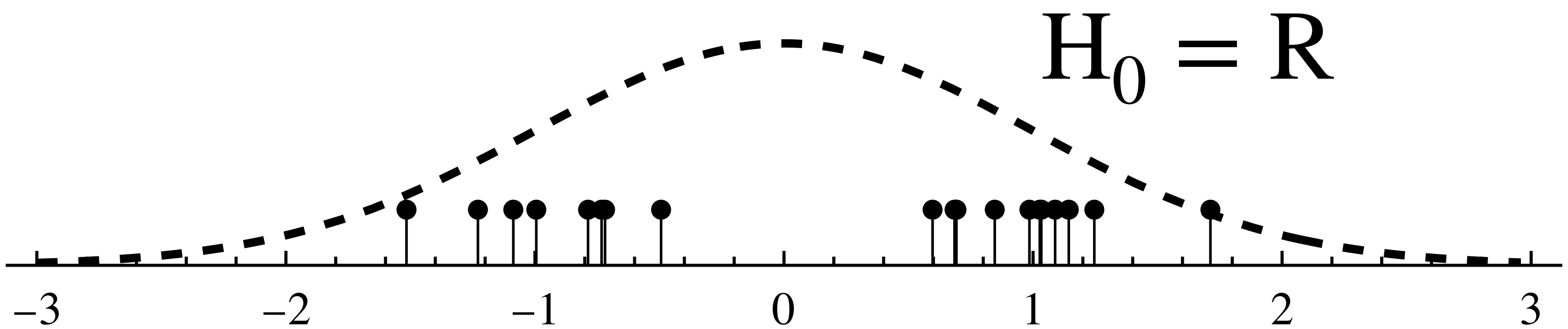
Statisticians formulate an interesting problem: **g.o.f.**

Be $\mathscr{D}$ some data, and R **one hypothesis** for their distribution

Does R provide the **right description** of $\mathscr{D}$ ?

**Example:** are these data described by a Standard Gaussian?

We try to answer by comparing the SG with some **Alternative** Hypothesis $H_1$. If $H_1$ works much better, R is in trouble.



# Goodness of Fit

Statisticians formulate an interesting problem: **g.o.f.**

Be $\mathscr{D}$ some data, and R **one hypothesis** for their distribution

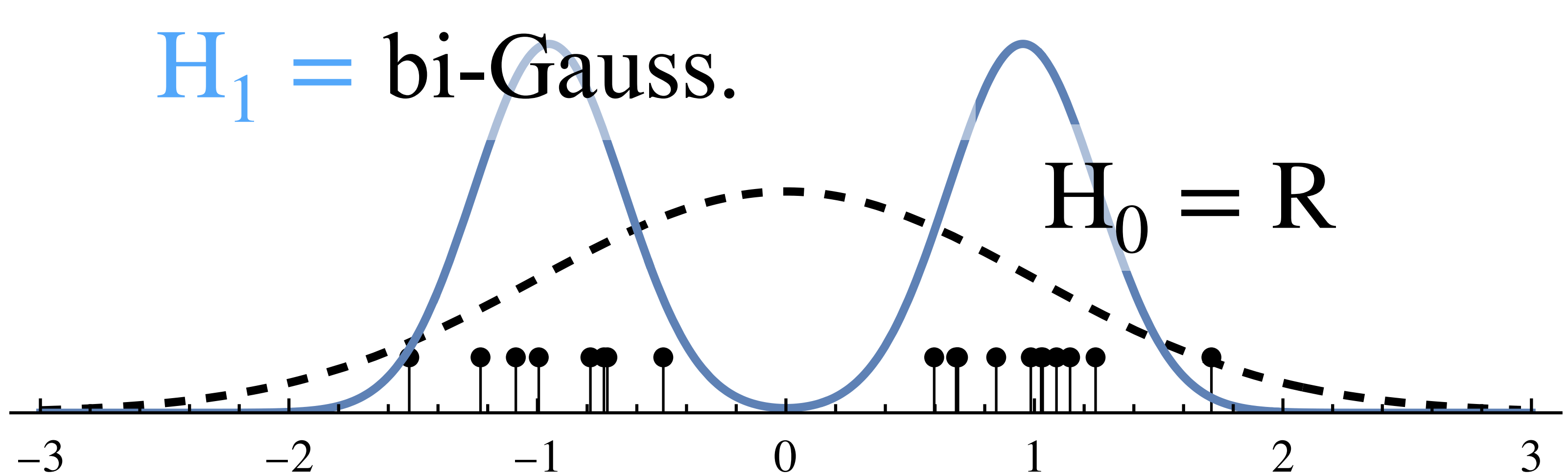
Does R provide the **right description** of $\mathscr{D}$ ?

**Example:** are these data described by a Standard Gaussian?

We try to answer by comparing the SG with some **Alternative** Hypothesis $H_1$. If $H_1$ works much better, R is in trouble.

Conclusion strongly depends on which $H_1$ we try:

- If $H_1 = H_T$ is **true** distribution, very likely we see **tension** of R (low p-value)



# Goodness of Fit

Statisticians formulate an interesting problem: **g.o.f.**

Be $\mathscr{D}$ some data, and R **one hypothesis** for their distribution

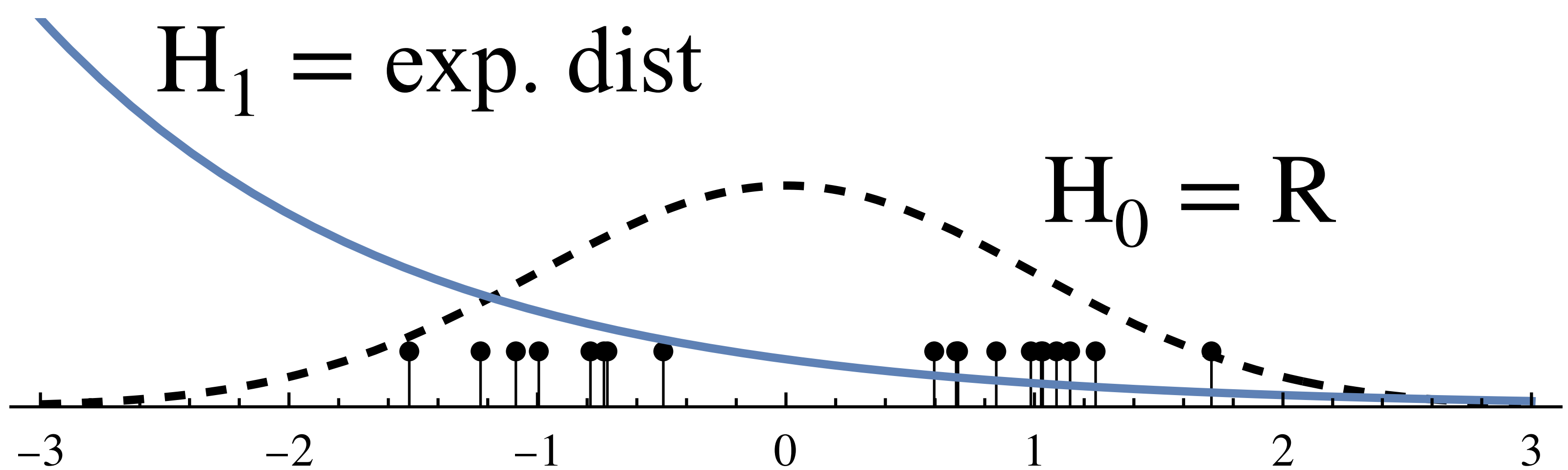
Does R provide the **right description** of $\mathscr{D}$ ?

**Example:** are these data described by a Standard Gaussian?

We try to answer by comparing the SG with some **Alternative** Hypothesis $H_1$. If $H_1$ works much better, R is in trouble.

Conclusion strongly depends on which $H_1$ we try:

- If $H_1 = H_T$ is **true** distribution, very likely we see **tension** of R (low p-value)
- If $H_1 \neq H_T$, we are likely to conclude that R is “good” (high p-value)



# Goodness of Fit

Statisticians formulate an interesting problem: **g.o.f.**

Be $\mathscr{D}$ some data, and R **one hypothesis** for their distribution

Does R provide the **right description** of $\mathscr{D}$ ?

Answering is more **easy** the more **restrictive** assumptions we make on how the true distribution, if not R, can look like.

But, more **partial** as well.

# Goodness of Fit

Statisticians formulate an interesting problem: **g.o.f.**

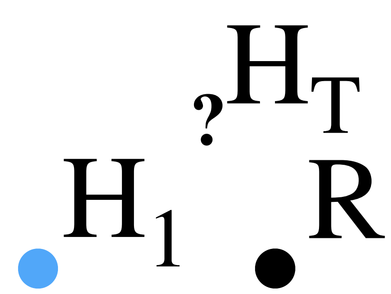
Be $\mathscr{D}$ some data, and R **one hypothesis** for their distribution

Does R provide the **right description** of $\mathscr{D}$ ?

Answering is more **easy** the more **restrictive** assumptions we make on how the true distribution, if not R, can look like.

But, more **partial** as well.

Simple vs Simple hypothesis test




- Optimal approach provided by **Neyman–Pearson Lemma**
- Optimal answer to very specific question: **test has no or very limited power if truth ≠ $H_1$**

# Goodness of Fit

Statisticians formulate an interesting problem: **g.o.f.**

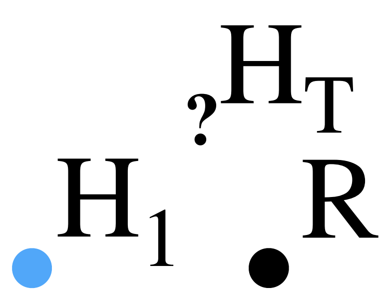
Be $\mathscr{D}$ some data, and R **one hypothesis** for their distribution

Does R provide the **right description** of $\mathscr{D}$ ?

Answering is more **easy** the more **restrictive** assumptions we make on how the true distribution, if not R, can look like.

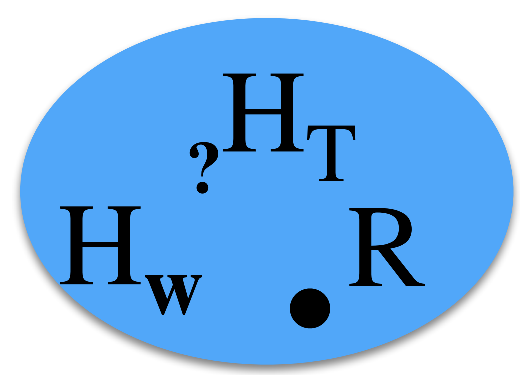
But, more **partial** as well.

Simple vs Simple hypothesis test




- Optimal approach provided by **Neyman–Pearson Lemma**
- Optimal answer to very specific question: **test has no or very limited power if truth ≠ $H_1$**

Simple vs Composite hypothesis test

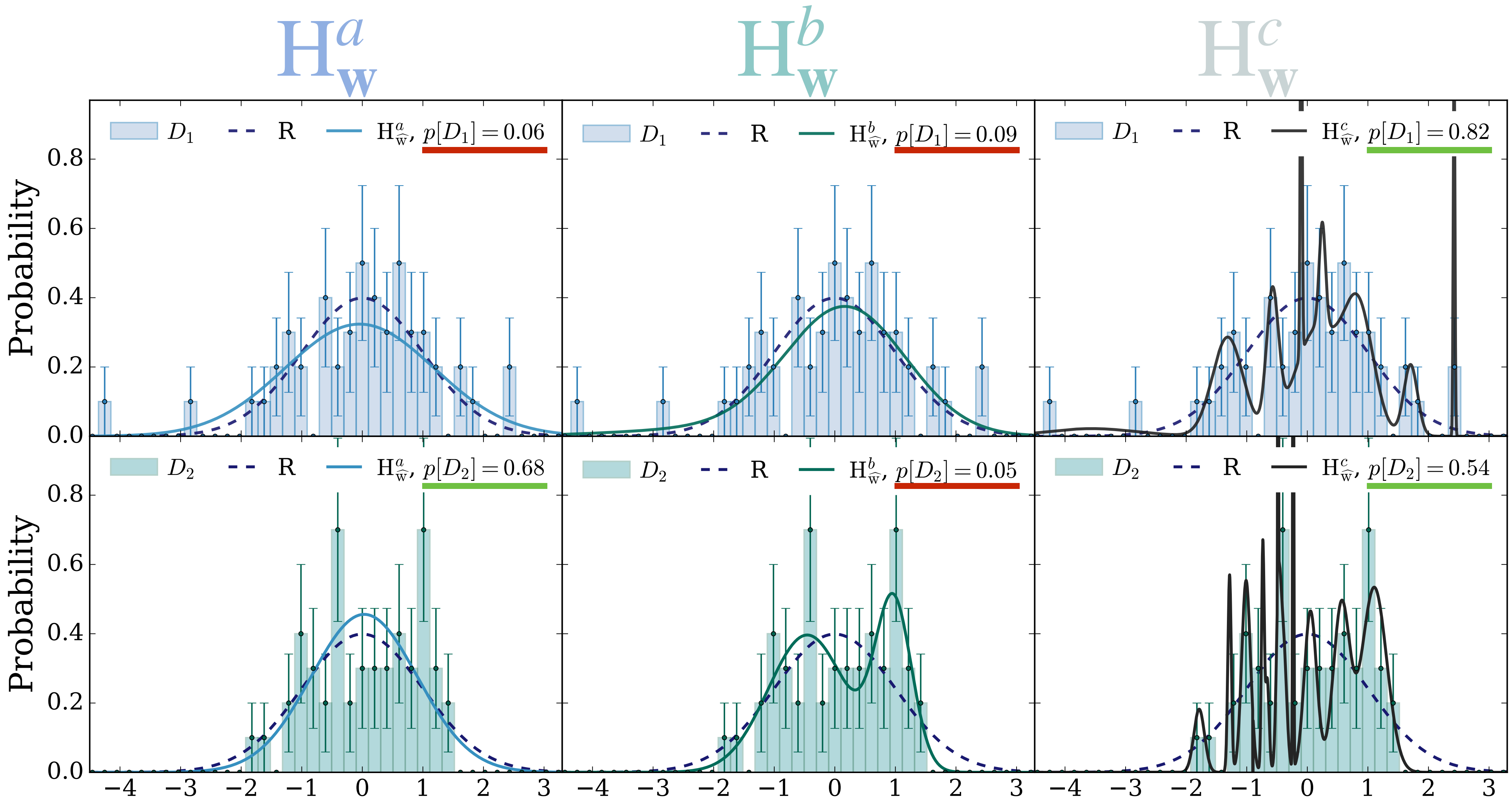


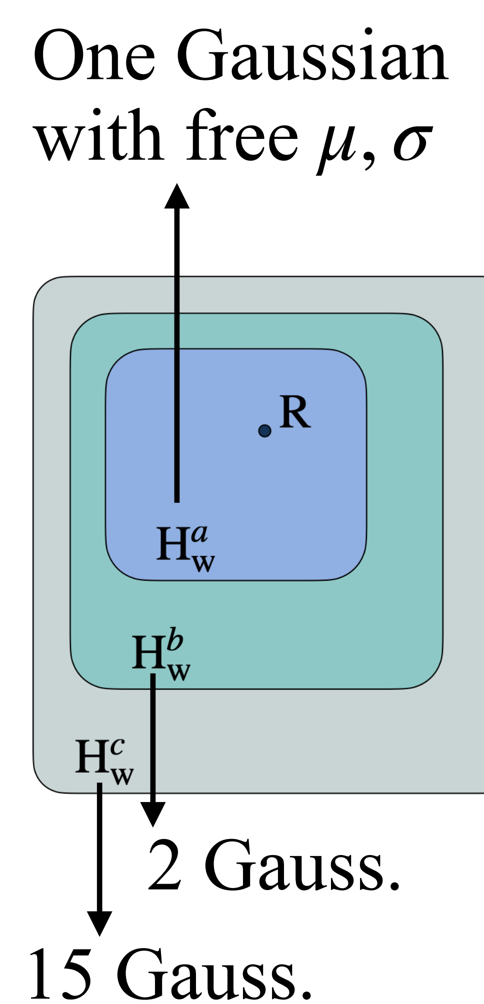

- No Optimal solution. But, **Maximum Likelihood Ratio** is **Good solution**
- Answers a more general question. It has **some power if truth is in $H_w$. But, larger $H_w$ = less power**

# Goodness of Fit

Toy example: 2 datasets, not from R, tested with 3 different $H_w$'s.

Red is good: means R in trouble — Green is bad: means that R looks OK

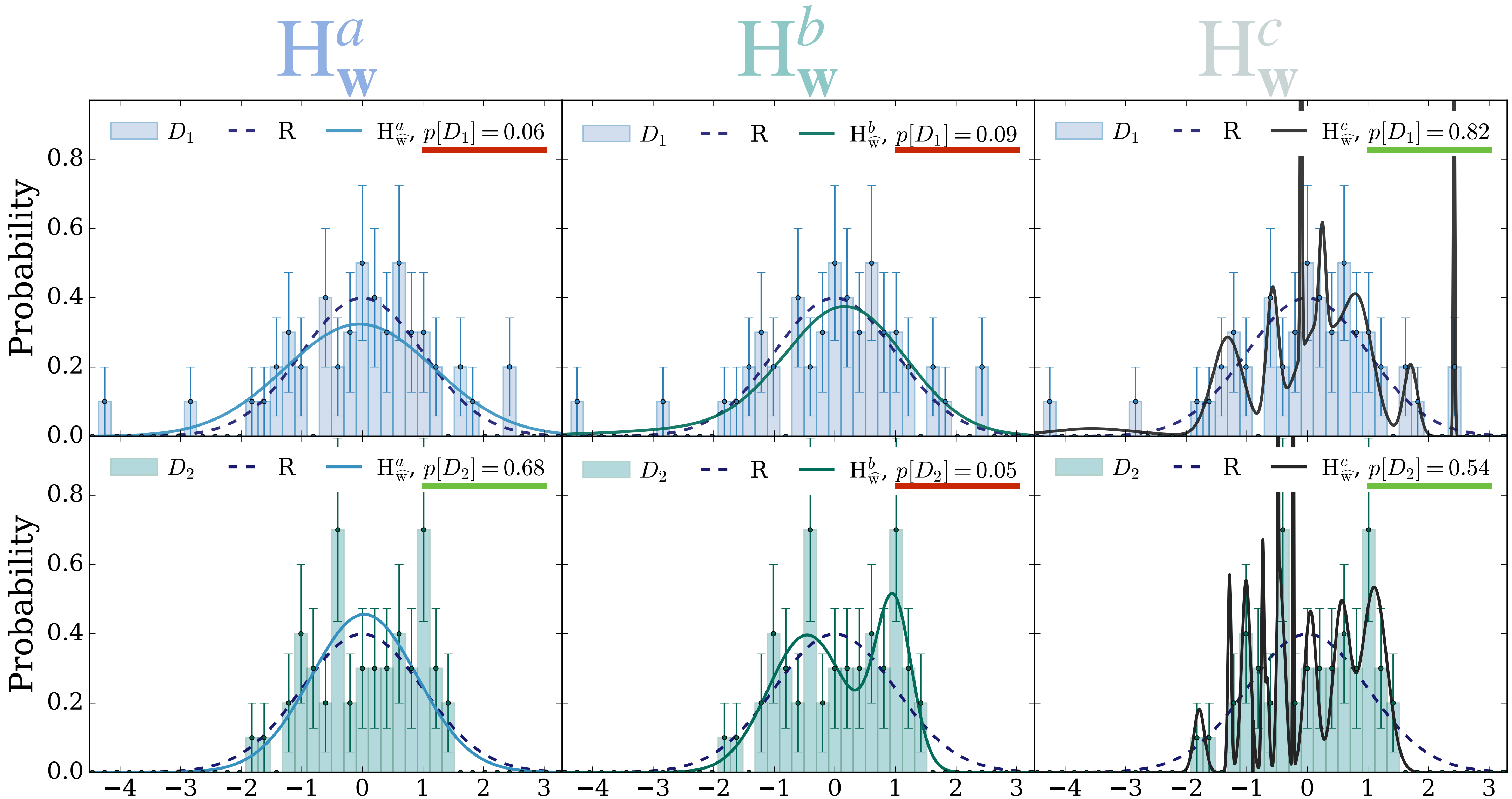


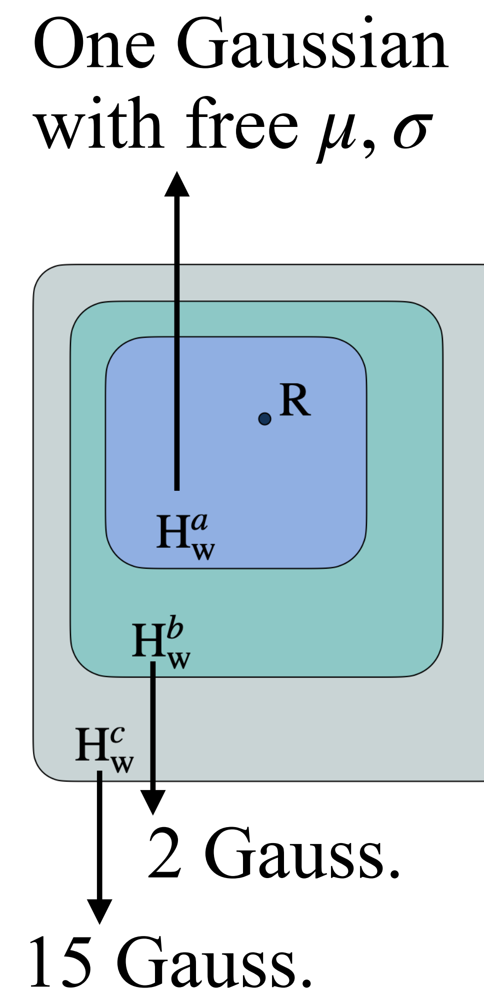



# Goodness of Fit

Toy example: 2 datasets, not from R, tested with 3 different $H_w$'s.

Red is good: means R in trouble — Green is bad: means that R looks OK




**We need large $H_w$ but avoid overfitting**

# New Physics Learning Machine (NPLM)

Data: i.i.d. measurements of feature vector $x$ (e.g., particle mom.)

$$\mathscr{D} = \{x_i\}_{i=1}^{\mathscr{N}}$$

In LHC, number of points is Poisson variable with expected $N$

Hypotheses: number density in $x$ space (in LHC, $d\sigma \times$ lumi . )

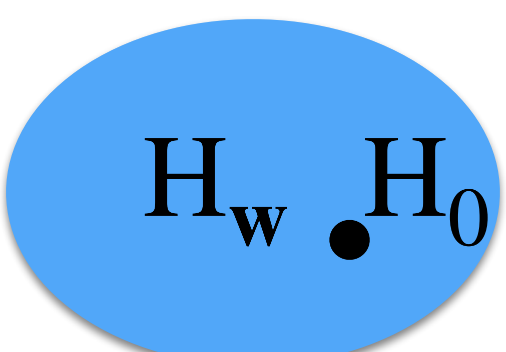
$$n(x) = N \cdot P(x)\,, \qquad N = \int dx n(x)$$

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

In LHC, the **SM prediction**

Alternative Hypothesis:

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



# New Physics Learning Machine (NPLM)

Data: i.i.d. measurements of feature vector $x$ (e.g., particle mom.)

$$\mathscr{D} = \{x_i\}_{i=1}^{\mathcal{N}}$$

In LHC, number of points is Poisson variable with expected $N$

Hypotheses: number density in $x$ space (in LHC, $d\sigma \times$ lumi . )

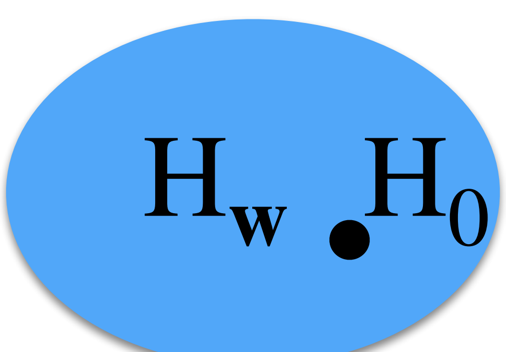
$$n(x) = N \cdot P(x)\,, \qquad N = \int dx n(x)$$

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

In LHC, the **SM prediction**

Alternative Hypothesis:

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




In NPLM, set of functions $f(x;\,\mathbf{w})$ that defines the Alternatives is **Neural Network** or other approximant good in many dimensions, like **kernels**

# New Physics Learning Machine (NPLM)

NPLM computes the Maximum Likelihood test statistic

$$t_{\mathrm{ML}}(\mathscr{D}) = 2\log\frac{\mathscr{L}(\mathrm{H}_{\hat{\mathbf{w}}})}{\mathscr{L}(\mathrm{R})} = 2\log\frac{e^{-N(\hat{\mathbf{w}})}}{e^{-N(\mathrm{R})}}\prod_{x\in\mathscr{D}}\frac{n(x\,|\,\mathrm{H}_{\hat{\mathbf{w}}})}{n(x\,|\,\mathrm{R})}$$

Using (since $n(x\,|\,\mathrm{R})$ not available) a **Reference Sample**

$$\mathscr{R} = \{x_i\}_{i=1}^{N_R}$$

$\mathscr{R}$ is made of instances of $x$ that follow the R distribution

If possible, $N_R \gg N(\mathrm{R})$, but this is not a strict requirement

# New Physics Learning Machine (NPLM)

NPLM computes the Maximum Likelihood test statistic

$$t_{\mathrm{ML}}(\mathscr{D}) = 2\log\frac{\mathscr{L}(\mathrm{H}_{\hat{\mathbf{w}}})}{\mathscr{L}(\mathrm{R})} = 2\log\frac{e^{-N(\hat{\mathbf{w}})}}{e^{-N(\mathrm{R})}}\prod_{x\in\mathscr{D}}\frac{n(x\,|\,\mathrm{H}_{\hat{\mathbf{w}}})}{n(x\,|\,\mathrm{R})}$$

Using (since $n(x\,|\,\mathrm{R})$ not available) a **Reference Sample**

$$\mathscr{R} = \{x_i\}_{i=1}^{N_R}$$

$\mathscr{R}$ is made of instances of $x$ that follow the R distribution

If possible, $N_R \gg N(\mathrm{R})$, but this is not a strict requirement

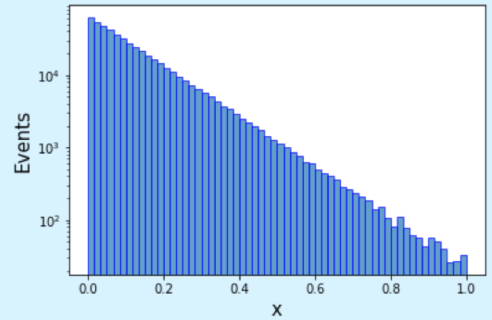
Computation of $t$ by **supervised training** $\mathscr{D}$ vs $\mathscr{R}$

In **NN** implementation, using special loss function that gives $t = -2\min[\mathrm{loss}]$

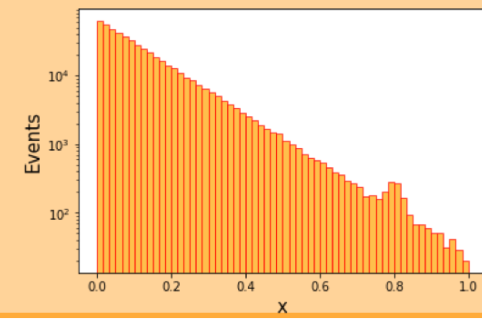
In **kernel** implementation, by learning "$\hat{\mathbf{w}}$" and plugging in

## INPUT

Reference sample ($R$)
label=0

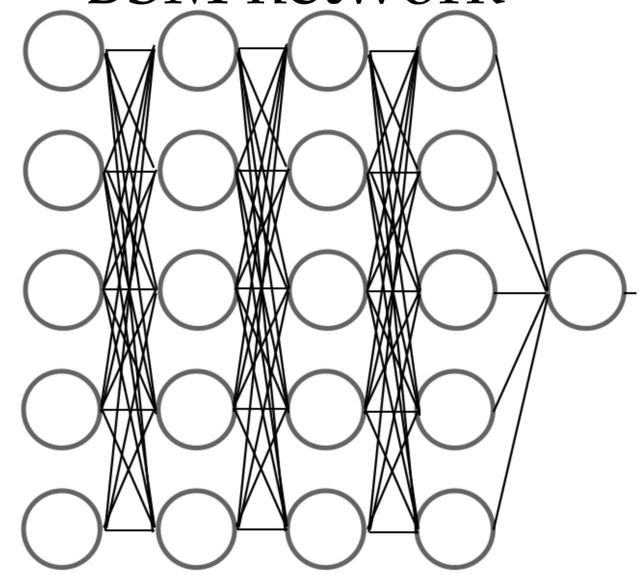



Data sample ($D$)
label=1




Unbinned training samples!

## BSM network



NN training

$\mathbf{w} \longrightarrow \hat{\mathbf{w}}$

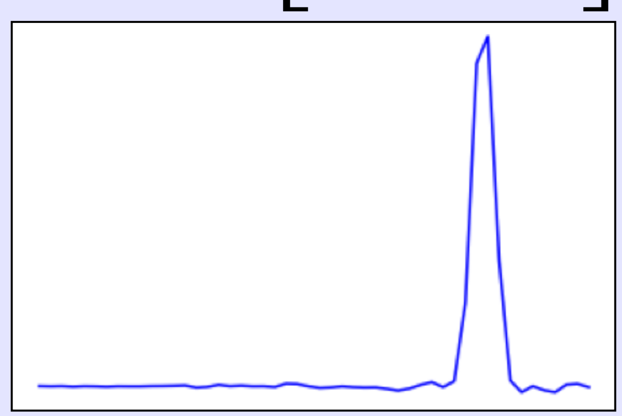
## OUTPUT

Single training

$$t(D) = -2\,L\left[f(x;\hat{\mathbf{w}})\right]$$

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

$f(x;\hat{\mathbf{w}})$

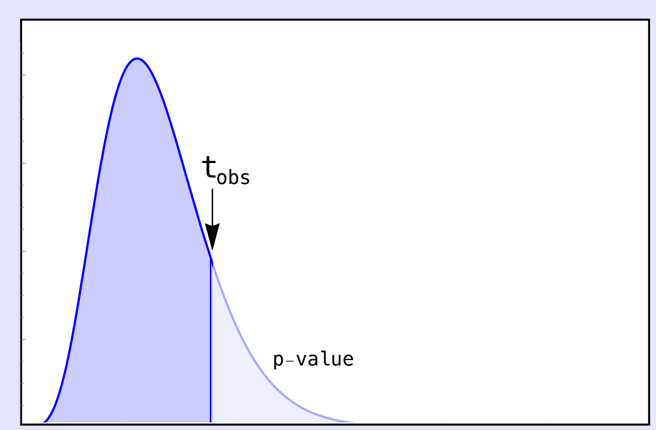


$x$

Many trainings
(with pseudo-data)

Empirical distribution of t
→ p-value for new datasets

$P(\bar{t}\,)$

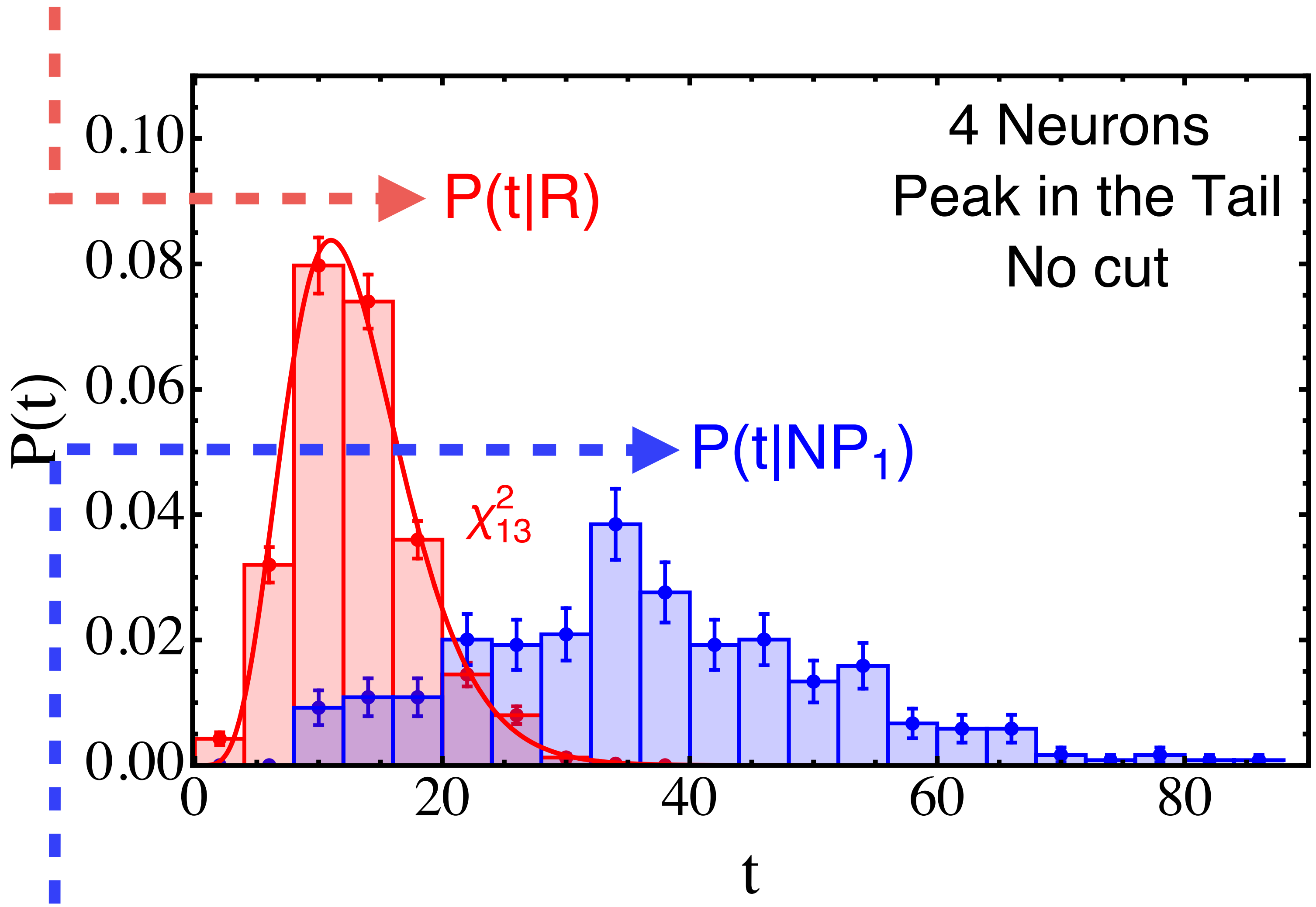



$\bar{t}$

# Illustrating Performances

(Simple 1d example with exponential Reference)

Distribution of the test statistic "t" in Reference Hypothesis



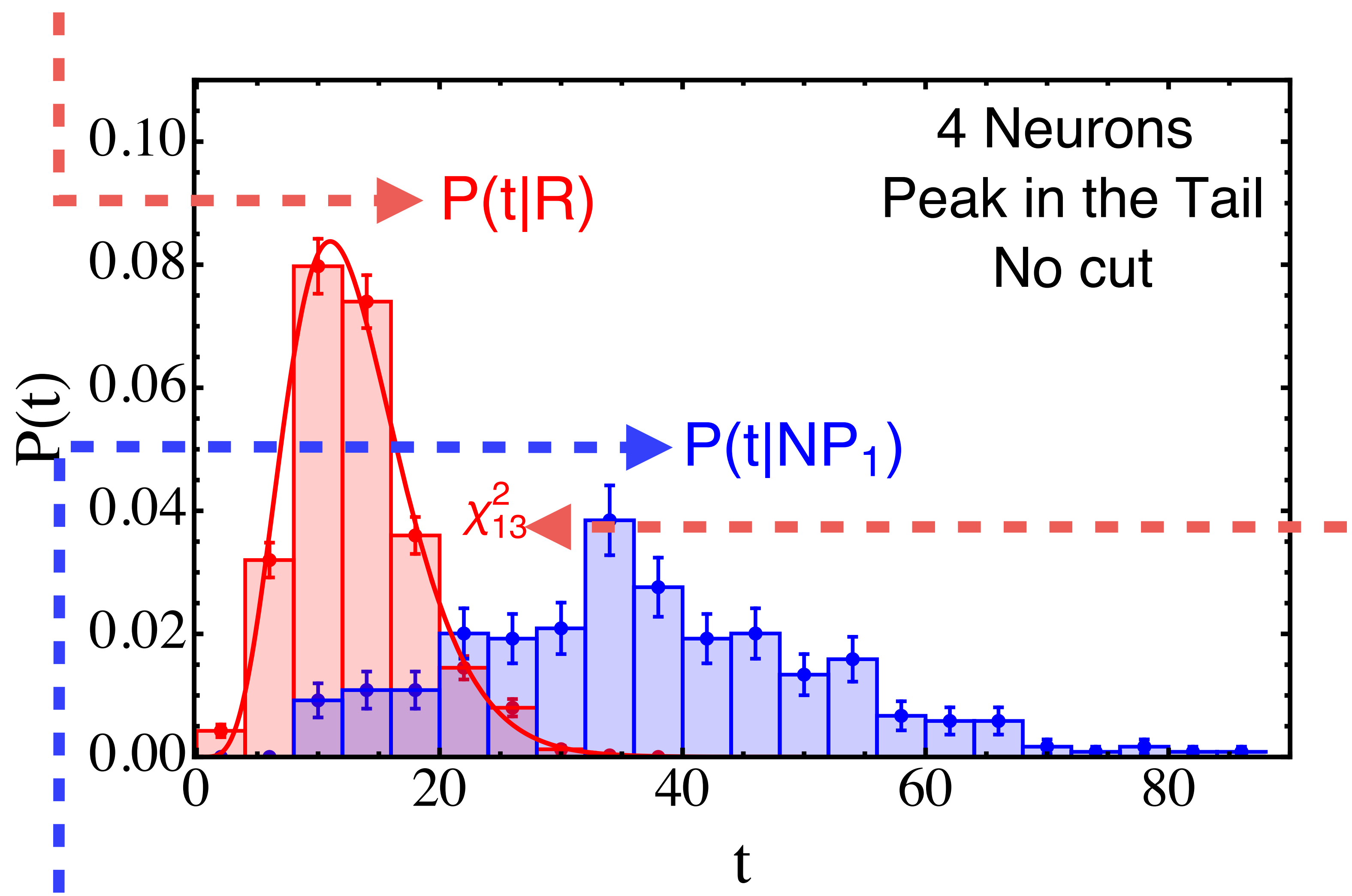

Distribution of "t" in one New Physics Model Hypothesis

t ⟶ p ⟶ Z-score (we use $Z = \Phi^{-1}(1-p)$)

# Illustrating Performances

(Simple 1d example with exponential Reference)

Distribution of the test statistic "t" in Reference Hypothesis




Notice agreement with **Wilks' Formula:**

Sufficiently **regularised networks** found to **behave as if their number of d.o.f. was equal to number of parameters.**

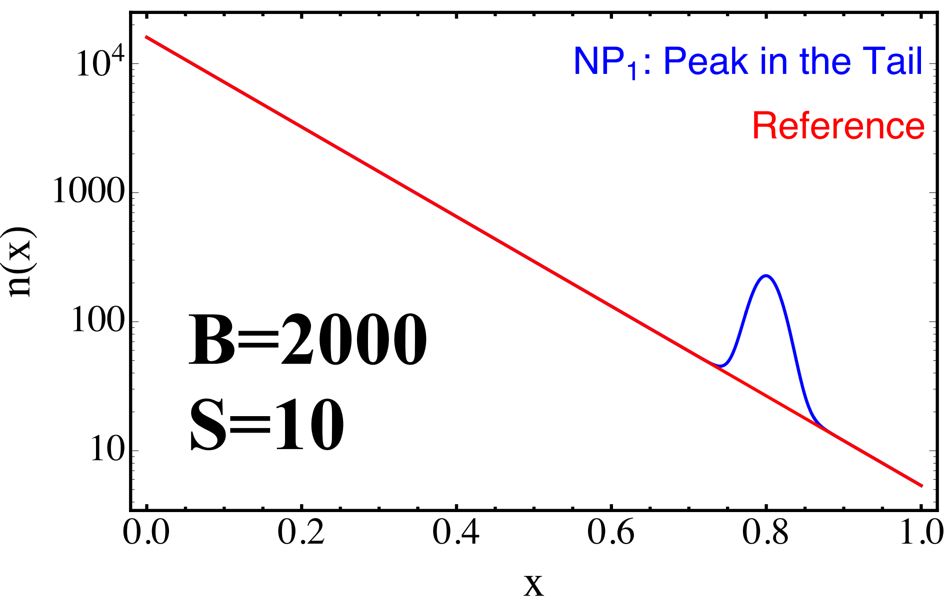
**Theoretical reason mysterious**

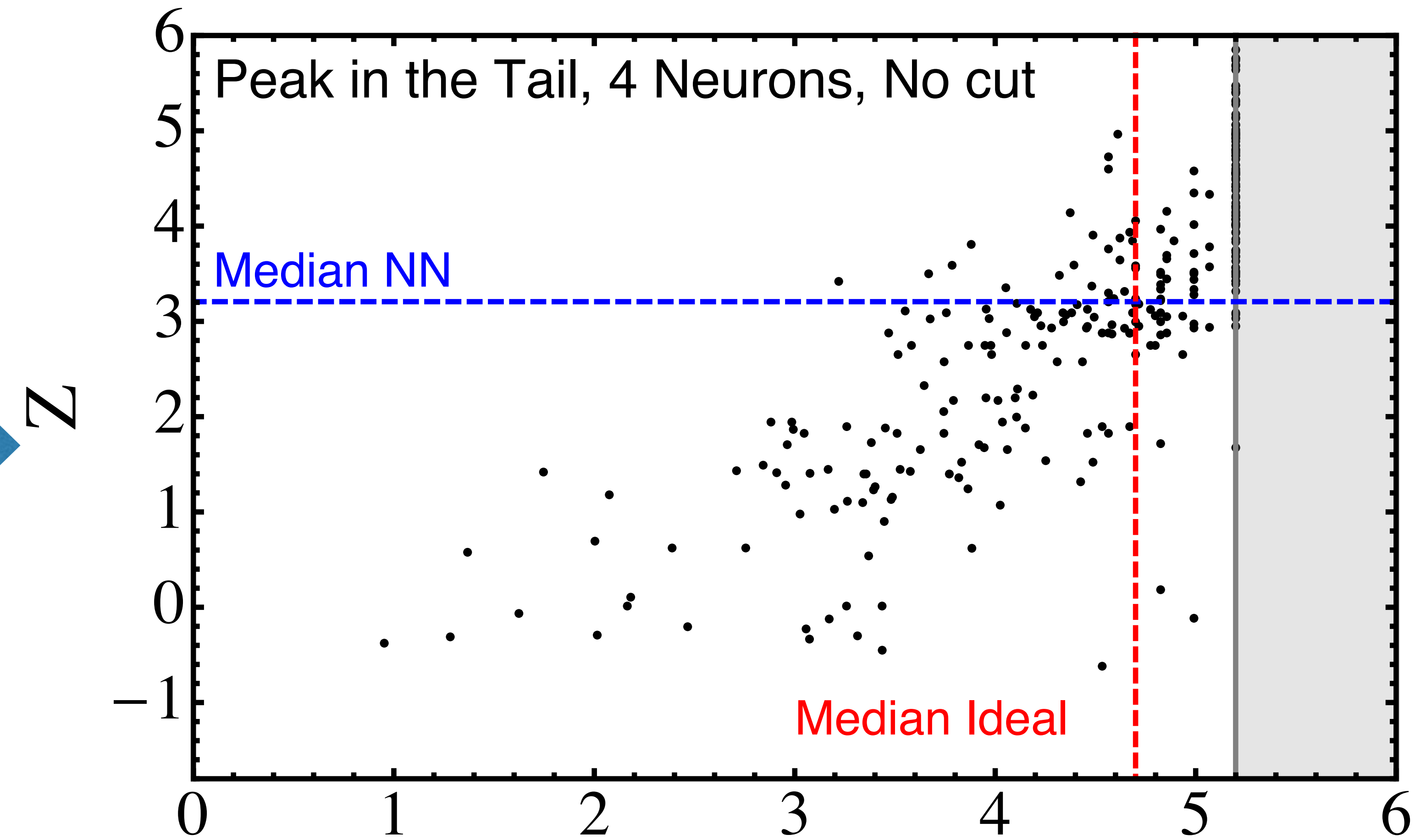
Distribution of "t" in one New Physics Model Hypothesis

t → p → Z-score (we use $Z = \Phi^{-1}(1 - p)$)

# Illustrating Performances

(Simple 1d example with exponential Reference)



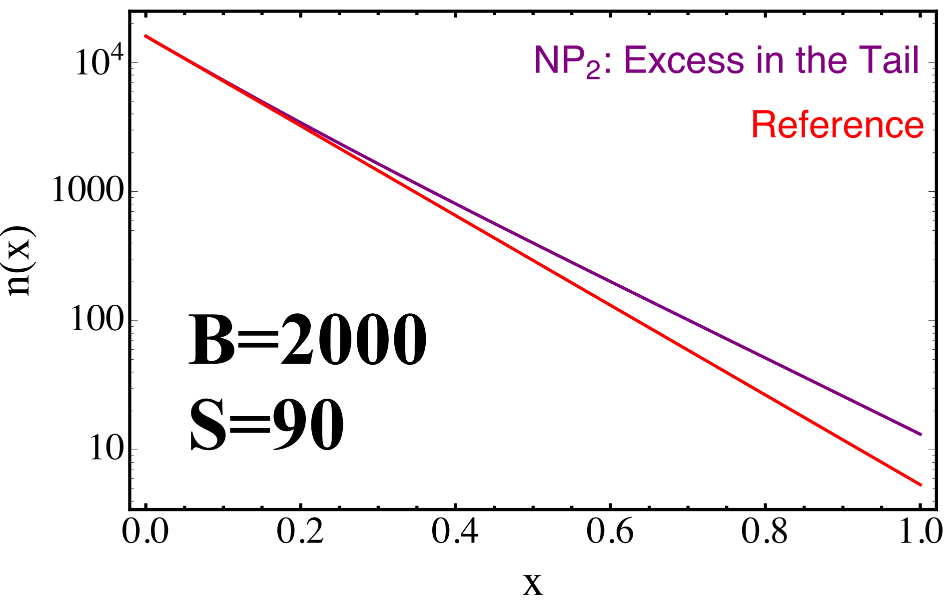

"Ideal Z-score": $Z_{id}$

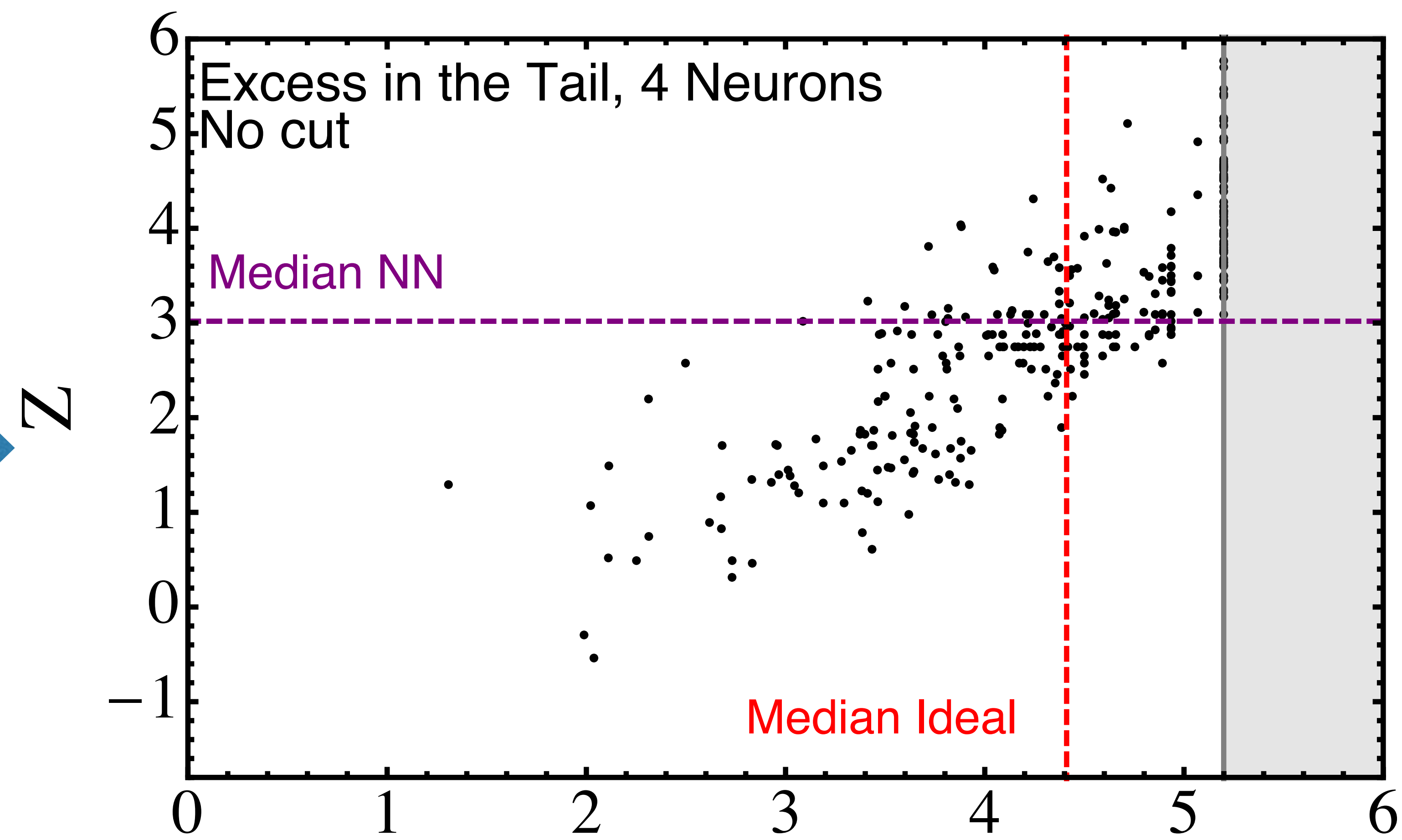
A "measure of dataset discrepancy":

Z-score of optimal test for **NP1** model

# Illustrating Performances

(Simple 1d example with exponential Reference)



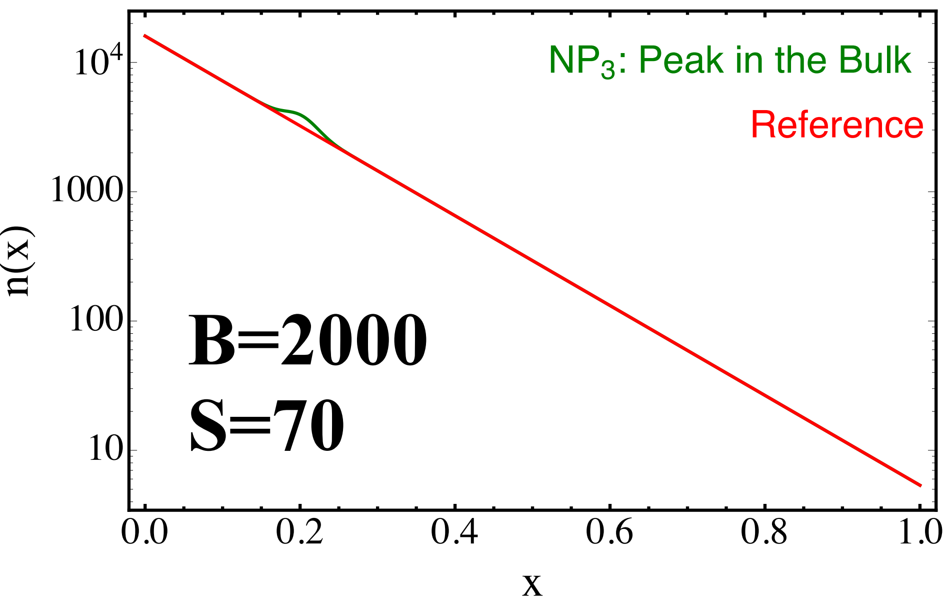

"Ideal Z-score": $Z_{\rm id}$

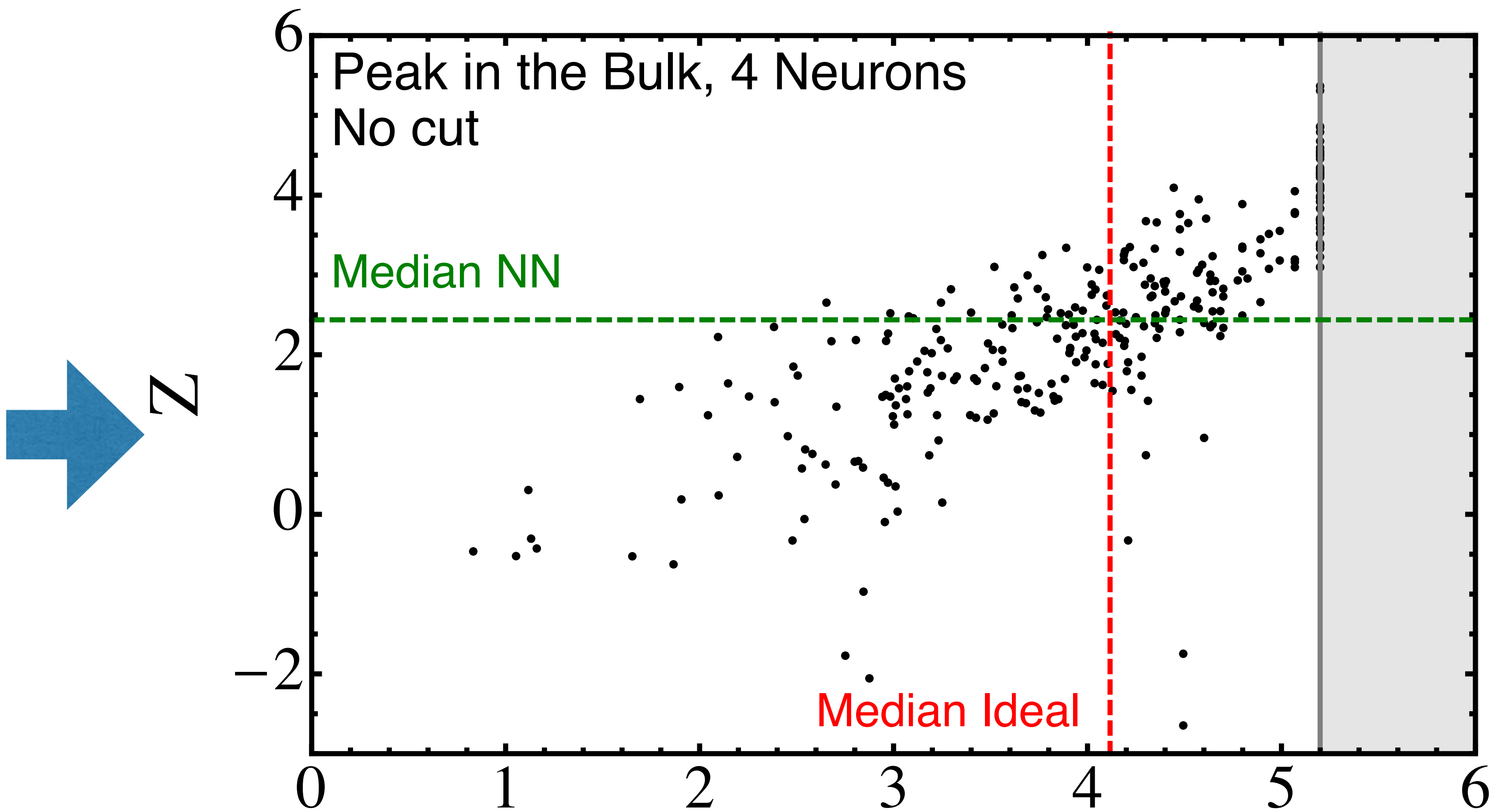
A "measure of dataset discrepancy":

Z-score of optimal test for **NP2** model

# Illustrating Performances

(Simple 1d example with exponential Reference)

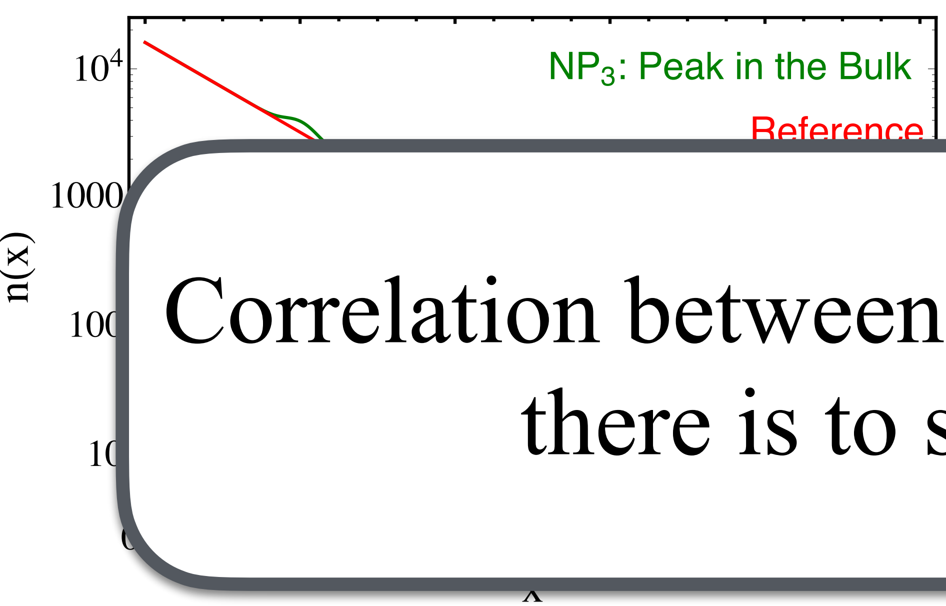


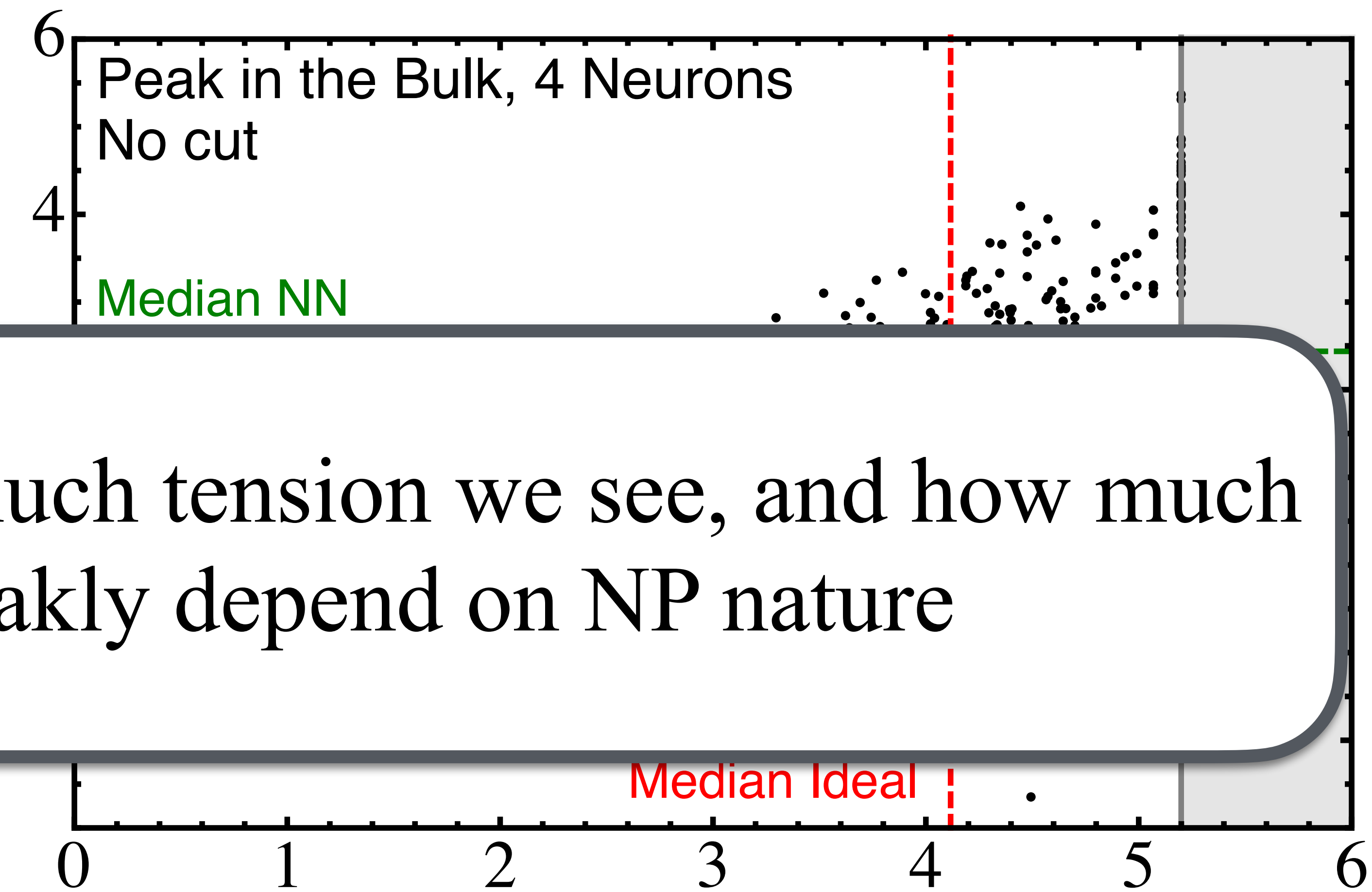




"Ideal Z-score": $Z_{\text{id}}$
A "measure of dataset discrepancy":
Z-score of optimal test for **NP3** model

# Illustrating Performances

(Simple 1d example with exponential Reference)




Correlation between how much tension we see, and how much there is to see. Weakly depend on NP nature

"Ideal Z-score": $Z_{\text{id}}$

A "measure of dataset discrepancy":

Z-score of optimal test for **NP3** model

# Comparing Performances

[Grosso, Letizia, Pierini, AW, 2023]

Many classical methods for g.o.f. with one-dimensional data:

- **$\chi^2$:** Bin data and compare with expected in each bin
- **EDF tests:** Compare EDF with CDF. Variants are KS, CvM, AD.
- **Spacing tests:** Spacings of CDF(points). Variants are Moran, RPS

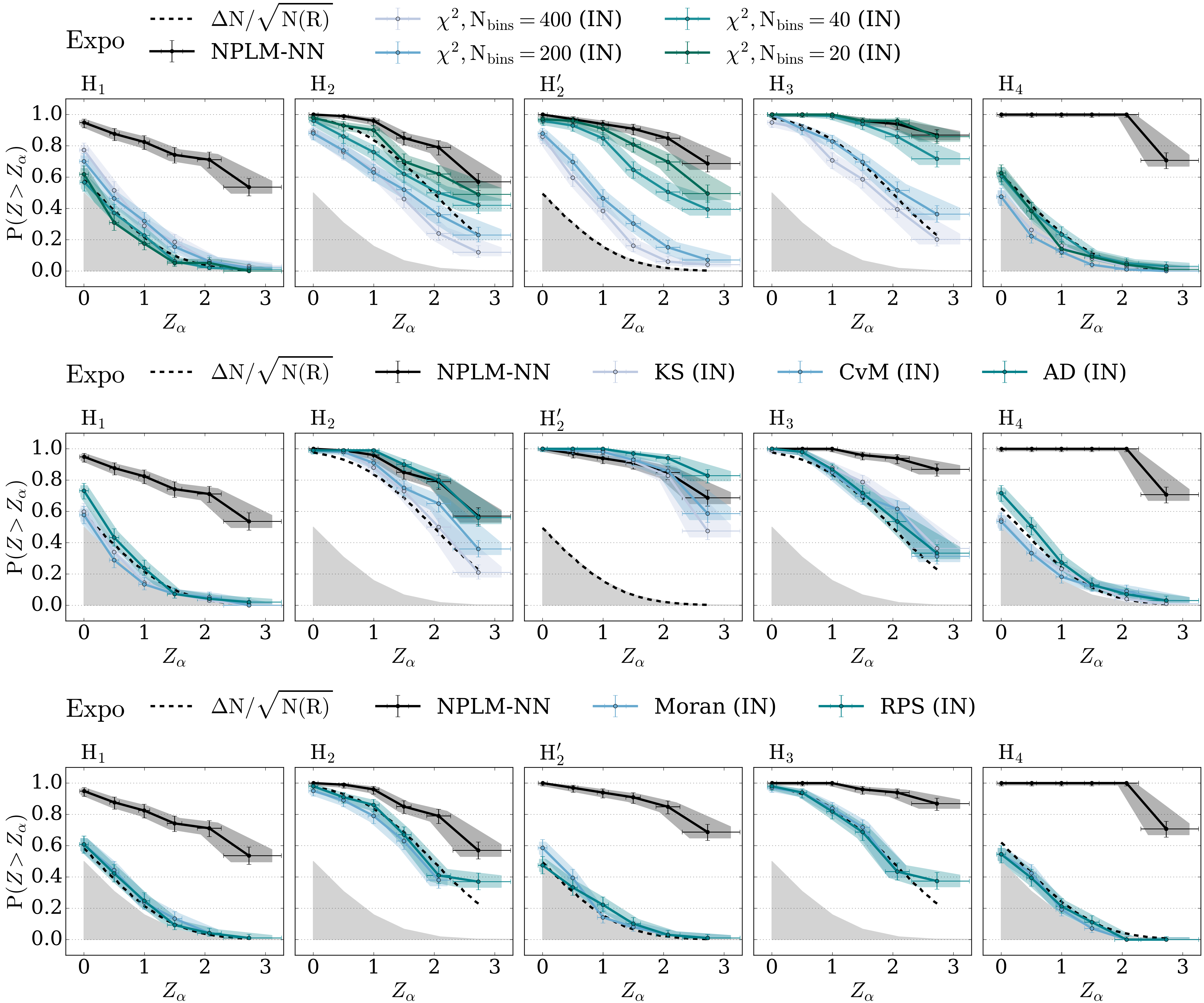
# Comparing Performances

[Grosso, Letizia, Pierini, AW, 2023]

Many classical methods for g.o.f. with one-dimensional data:

- **$\chi^2$:** Bin data and compare with expected in each bin
- **EDF tests:** Compare EDF with CDF. Variants are KS, CvM, AD.
- **Spacing tests:** Spacings of CDF(points). Variants are Moran, RPS

While $d = 1$ g.o.f. is considered a "solved problem", and $d > 1$ is what we care, interesting that **NPLM works better**.



Expo
$\Delta N/\sqrt{N(R)}$
NPLM-NN
$\chi^2, N_{bins}=400$ (IN)
$\chi^2, N_{bins}=200$ (IN)
$\chi^2, N_{bins}=40$ (IN)
$\chi^2, N_{bins}=20$ (IN)
KS (IN)
CvM (IN)
AD (IN)
Moran (IN)
RPS (IN)
$H_1$
$H_2$
$H'_2$
$H_3$
$H_4$
$P(Z>Z_\alpha)$
$Z_\alpha$

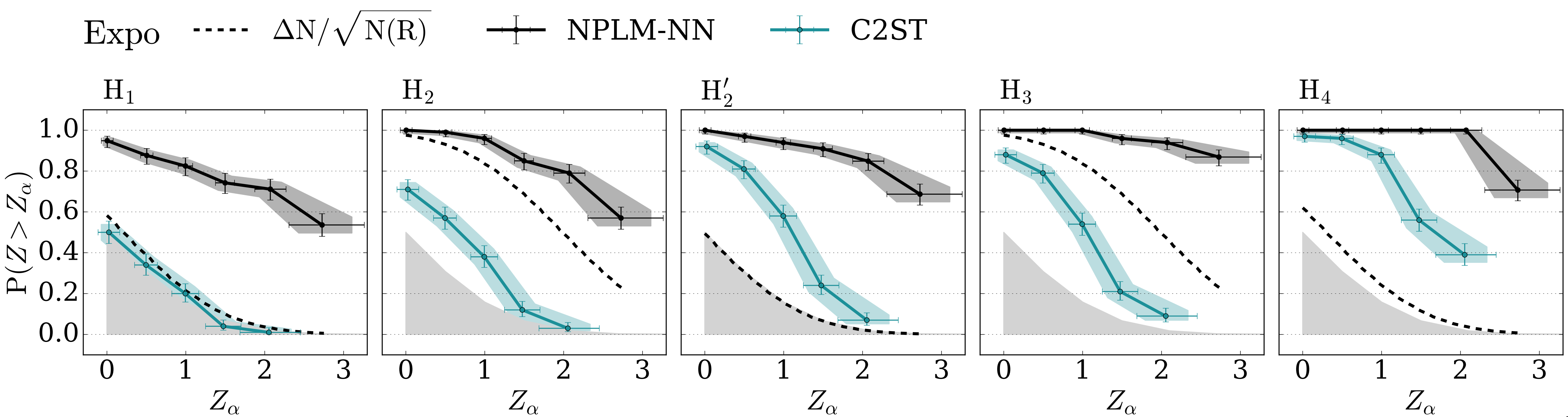
# Comparing Performances

[Grosso, Letizia, Pierini, AW, 2023]

For $d > 1$, most established solution are **Classifier-Based Tests**

- **General idea:** Train $\mathscr{D}$ vs $\mathscr{R}$. Get more decisive classifier if $\mathscr{D} \nsim$ R
  Use **some metric** evaluated on trained classifier output for Hypothesis Test.
  [Friedman, 2003]
- **C2ST:** Most natural implementation. Uses classification accuracy metric.
  [Lopez-Paz, Oquab, 2016]
  Employed for generative models validation
- **Variants:** We studied different metric and compared in/out evaluation.

## NPLM vs C2ST: $d = 1$



# Comparing Performances

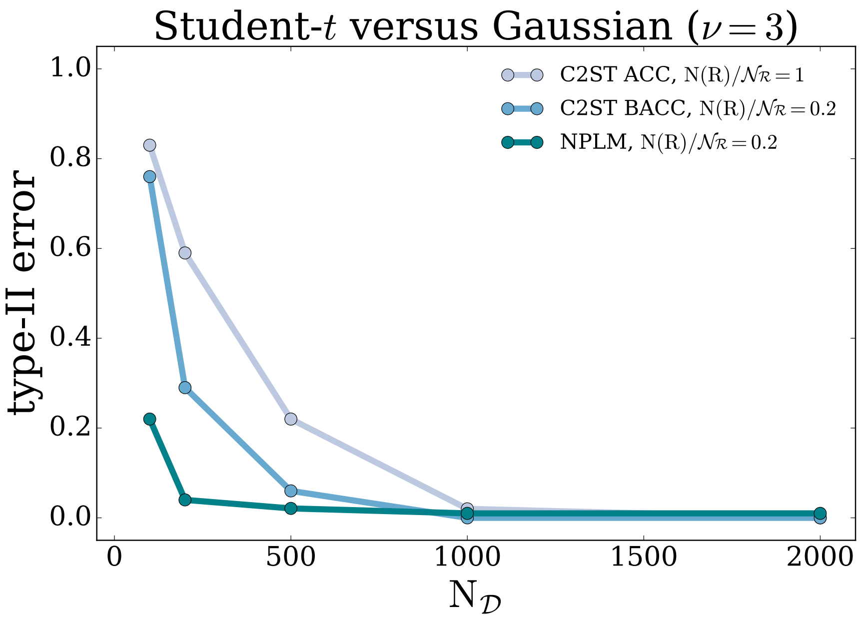
[Grosso, Letizia, Pierini, AW, 2023]

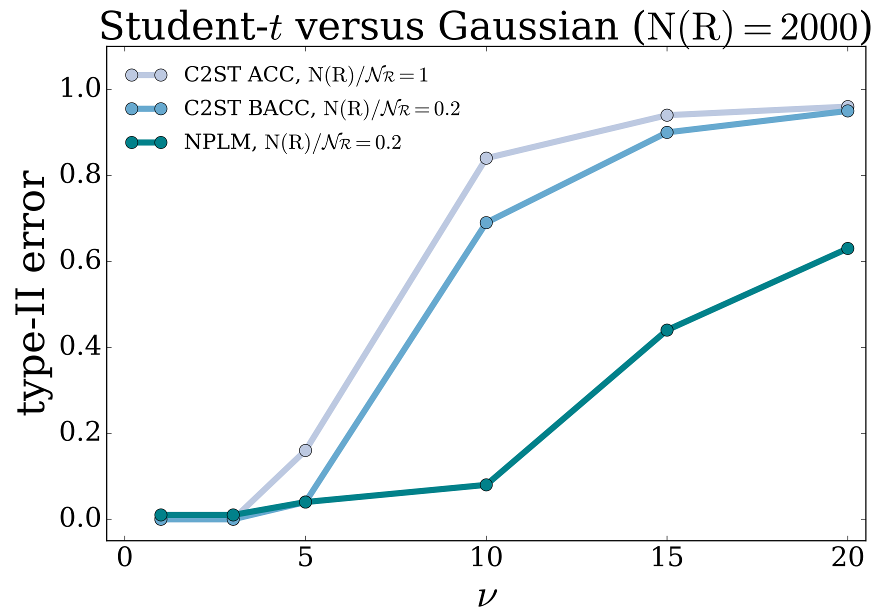
For $d > 1$, most established solution are **Classifier-Based Tests**

- **General idea:** Train $\mathscr{D}$ vs $\mathscr{R}$. Get more decisive classifier if $\mathscr{D} \nsim$ R
  Use **some metric** evaluated on trained classifier output for Hypothesis Test.
  [Friedman, 2003]
- **C2ST:** Most natural implementation. Uses classification accuracy metric.
  [Lopez-Paz, Oquab, 2016]
  Employed for generative models validation
- **Variants:** We studied different metric and compared in/out evaluation.

NPLM vs C2ST: $d = 1$

Low is Good



# Comparing Performances

[Grosso, Letizia, Pierini, AW, 2023]

For $d > 1$, most established solution are **Classifier-Based Tests**

- **General idea:** Train $\mathscr{D}$ vs $\mathscr{R}$. Get more decisive classifier if $\mathscr{D} \nsim$ R

  Use **some metric** evaluated on trained classifier output for Hypothesis Test.

  [Friedman, 2003]

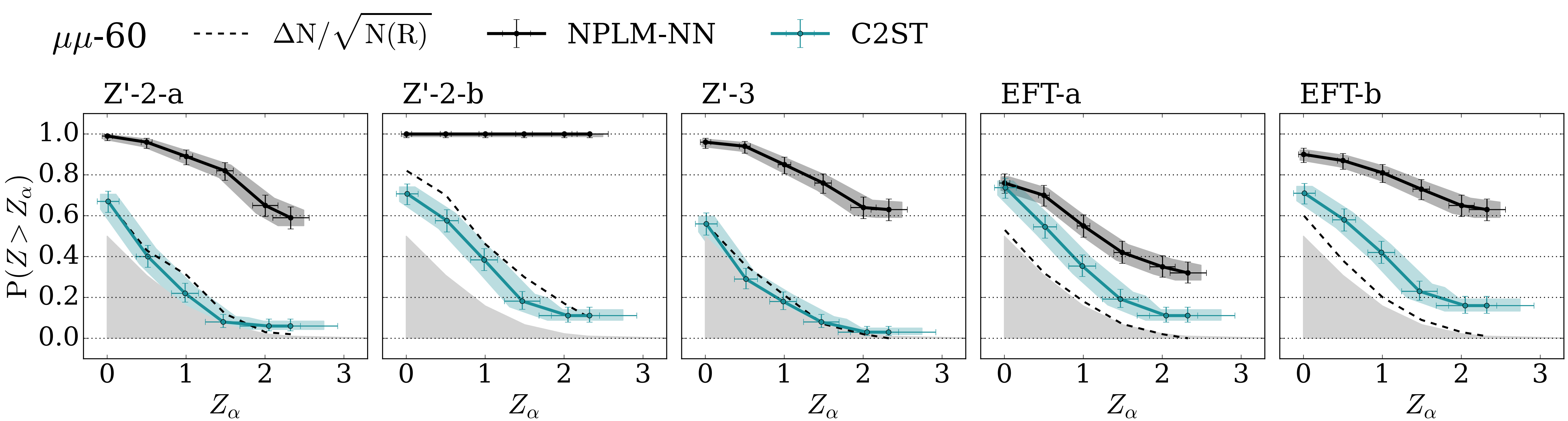
- **C2ST:** Most natural implementation. Uses classification accuracy metric.

  [Lopez-Paz, Oquab, 2016]

  Employed for generative models validation

- **Variants:** We studied different metric and compared in/out evaluation.

NPLM vs C2ST: $d = 5$



# Comparing Performances

[Grosso, Letizia, Pierini, AW, 2023]

For $d > 1$, most established solution are **Classifier-Based Tests**

- **General idea:** Train $\mathscr{D}$ vs $\mathscr{R}$. Get more decisive classifier if $\mathscr{D} \nsim$ R

  Use **some metric** evaluated on trained classifier output for Hypothesis Test.

  [Friedman, 2003]
- **C2ST:** Most natural implementation. Uses classification accuracy metric.

  [Lopez-Paz, Oquab, 2016]

  Employed for generative models validation
- **Variants:** We studied different metric and compared in/out evaluation.

NPLM **is** a Classifier-Based Test. Why so much better?

After comparison of many CBT variants, we conclude that the key is using Maximum Likelihood Ratio as metric, and in-sample eval.

**Distinctive feature of NPLM is implementing N&P Testing!**

# Applications

Some of the many applications of g.o.f. are:

- **Model-Agnostic BSM Searches**
- **Data Quality Monitoring:** Tell if apparatus operates "normally"
- **Generative Models:** GM validation and selection

# The LHC g.o.f. challenge

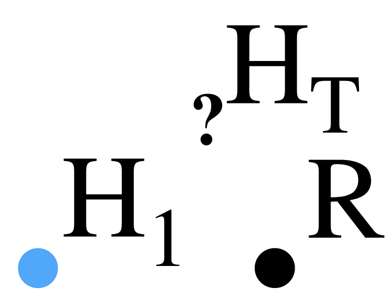
By analysing the LHC data, we would like to find evidence of **failure of the SM theory**, suggesting need of **BSM.**

**This is a tremendously hard gof problem!**

BSM is tiny departure from SM, or large in tiny prob. region
Affecting few (unknown) observables over ∞ many we can measure

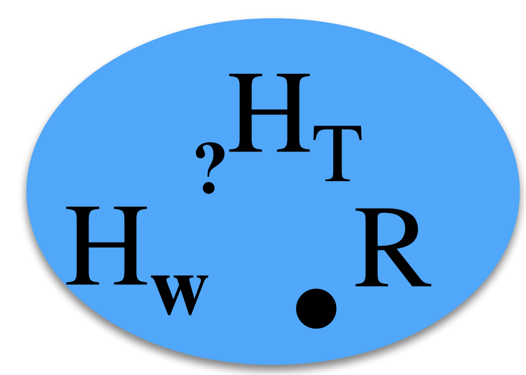
Our generic discussion …

Simple vs Simple hypothesis test




- Optimal approach provided by **Neyman–Pearson Lemma**
- Optimal answer to very specific question: **test has no or very limited power if truth $\neq$ $H_1$**

Simple vs Composite hypothesis test




- No Optimal solution. But, **Maximum Likelihood Ratio** is **Good solution**
- Answers a more general question. It has **some power if truth is in $H_w$. But, larger $H_w$ = less power**

# The LHC g.o.f. challenge

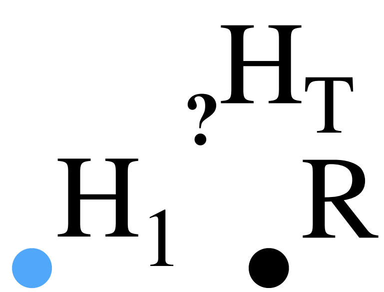
By analysing the LHC data, we would like to find evidence of **failure of the SM theory**, suggesting need of **BSM.**

**This is a tremendously hard gof problem!**

BSM is tiny departure from SM, or large in tiny prob. region
Affecting few (unknown) observables over ∞ many we can measure

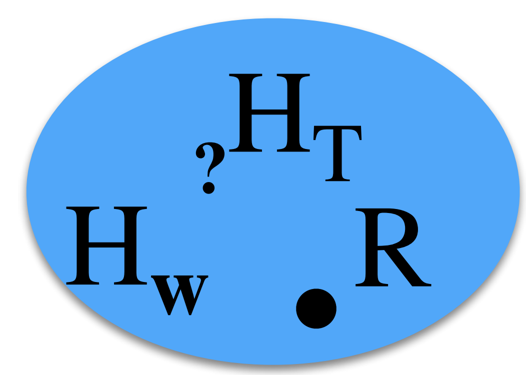
Our generic discussion … perfectly matches LHC practice:

**Model-dependent**
BSM searches




- Optimise sensitivity to **one specific BSM model**
- Fail to discover other models. **What if the right theoretical model is not yet formulated?**

**Model-independent**
searches




- Could reveal **truly unexpected** new physical laws.
- No hopes to find Optimal strategy. But we must aim at a Good strategy

# Key Challenge: Uncertainties

[D'Agnolo, Grosso, Pierini, AW, Zanetti, 2021]

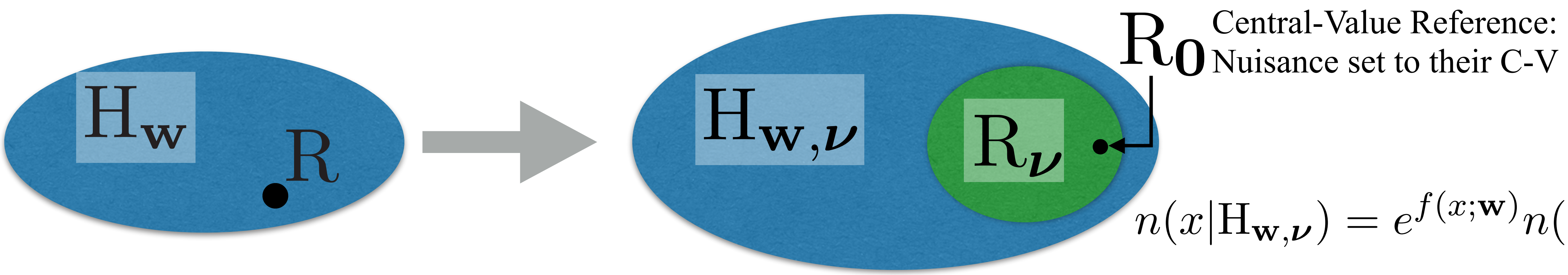
Reference Sample is an **imperfect** representation of SM

e.g., PDF/Lumi/Detector Modeling …

Imperfections are **Nuisance Parameters**

Constrained by **Auxiliary Measurements**

Define a **composite** Reference hypothesis




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

Strategy conceptually unchanged.

$$t(\mathcal{D},\mathcal{A}) = 2\,\log \frac{\max\limits_{\mathbf{w},\boldsymbol{\nu}} \left[\mathcal{L}(\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}}|\mathcal{D}) \cdot \mathcal{L}(\boldsymbol{\nu}|\mathcal{A})\right]}{\max\limits_{\boldsymbol{\nu}} \left[\mathcal{L}(\mathrm{R}_{\boldsymbol{\nu}}|\mathcal{D}) \cdot \mathcal{L}(\boldsymbol{\nu}|\mathcal{A})\right]}$$

$$= 2\,\max_{\mathbf{w},\boldsymbol{\nu}} \log\left[\frac{\mathcal{L}(\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}}|\mathcal{D})}{\mathcal{L}(\mathrm{R}_{\mathbf{0}}|\mathcal{D})} \cdot \frac{\mathcal{L}(\boldsymbol{\nu}|\mathcal{A})}{\mathcal{L}(\mathbf{0}|\mathcal{A})}\right] - 2\,\max_{\boldsymbol{\nu}} \log\left[\frac{\mathcal{L}(\mathrm{R}_{\boldsymbol{\nu}}|\mathcal{D})}{\mathcal{L}(\mathrm{R}_{\mathbf{0}}|\mathcal{D})} \cdot \frac{\mathcal{L}(\boldsymbol{\nu}|\mathcal{A})}{\mathcal{L}(\mathbf{0}|\mathcal{A})}\right] = \tau(\mathcal{D},\mathcal{A}) - \Delta(\mathcal{D},\mathcal{A})$$

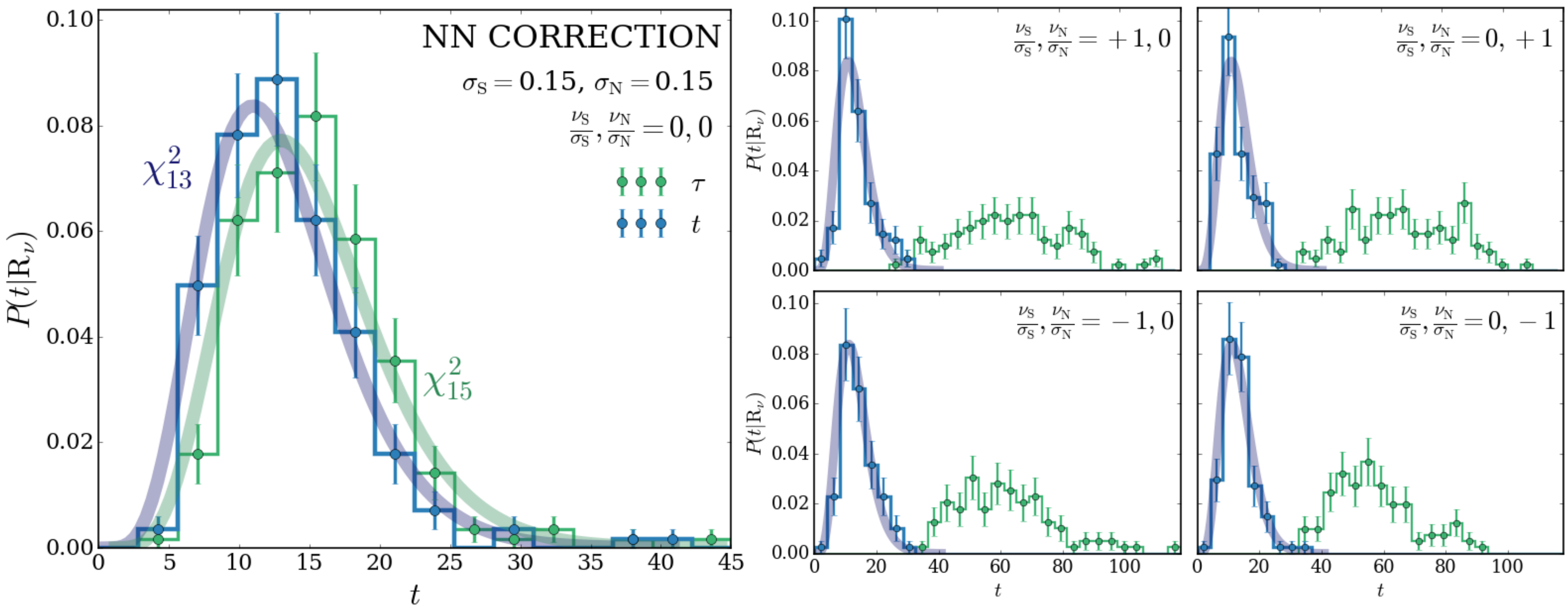
Implementation slightly more complex

# An **Imperfect** Machine at Work

[D'Agnolo, Grosso, Pierini, AW, Zanetti, 2021]

Tau distribution distorted by non-central value nuisance

if not corrected, produces false positives




t = Tau-Delta independent of true nuisance value

**this is essential for a feasible test**

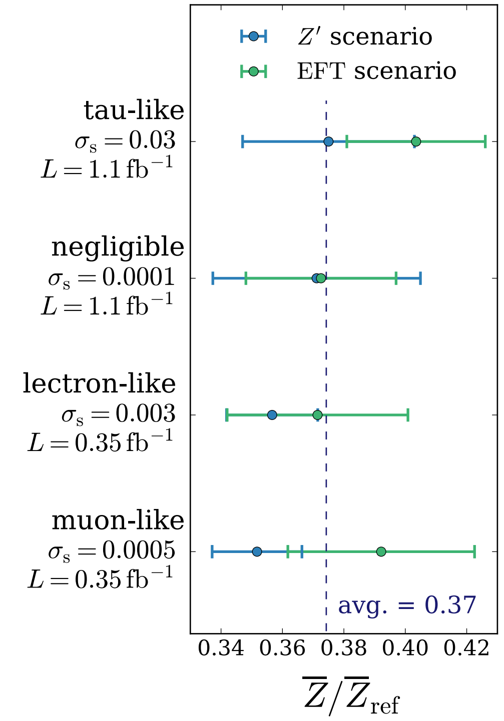
# Towards LHC

## Our proposed strategy is fully defined, including:

- Hyperparameters and regularisation selection
- Systematic approach to Reference mis-modelling

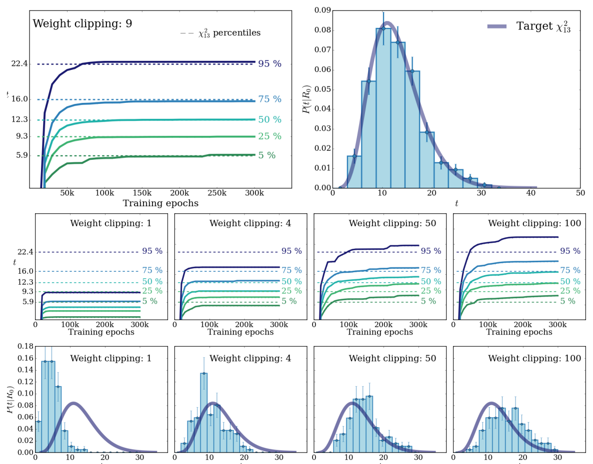
## Validated on problems of realistic scale of complexity:

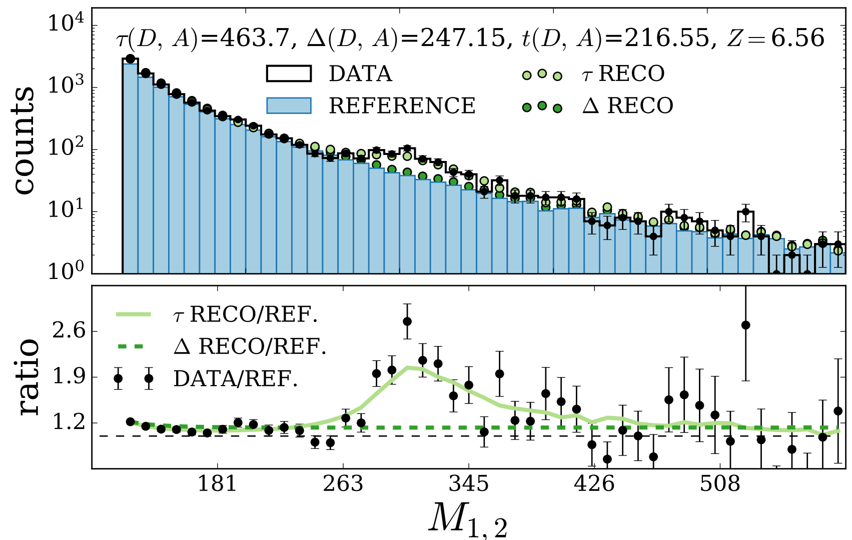
- 2-body final state with uncertainties ($d$ = 5)
- ll+MET “SUSY” ($d$ = 8)
- Heavy Higgs to WWbb ($d$ = 21)




## Results in summary:

- model-selection strategy converges
- sensitivity to resonant or non-resonant NP
- “uniform” response to NP of different nature
- trained network reconstruct NP

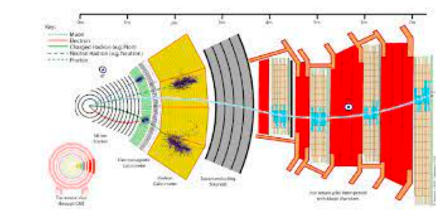




# Data Quality Monitoring

[Grosso, Letizia, AW, Zanetti, et. al., 2023]

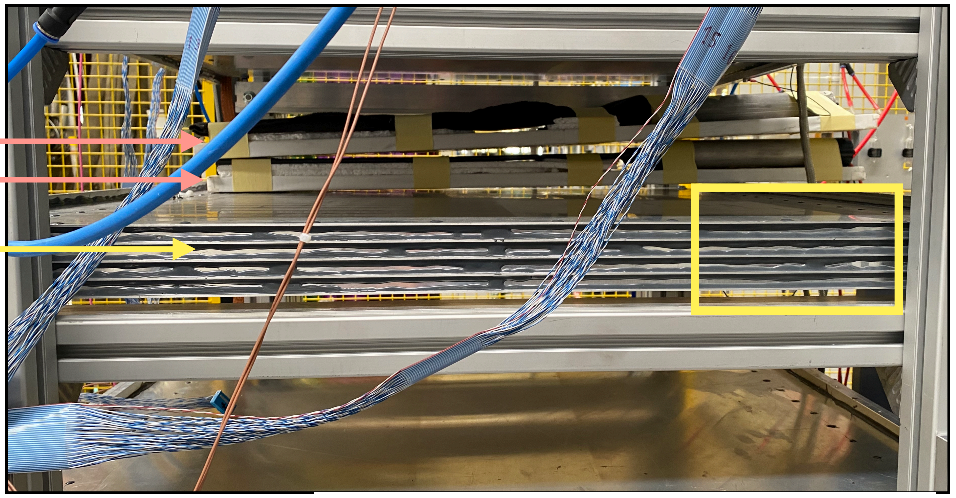
## No Reference uncertainties: $\mathscr{R}$ is data in good operation condition



### $n$D DQM
Online monitoring of a DT chamber:

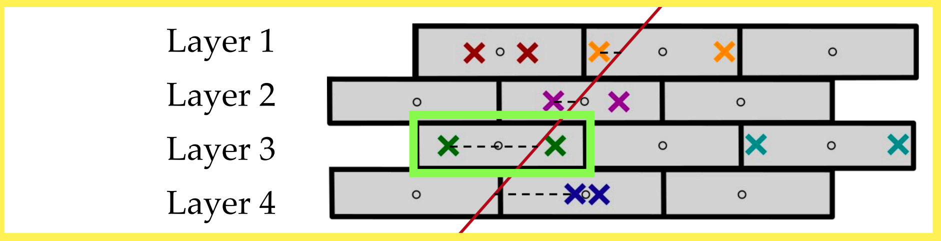
**Setup (Legnaro INFN national laboratory):**

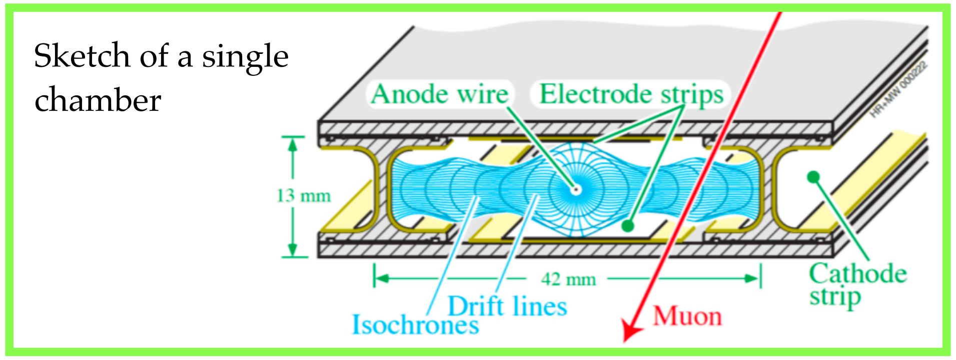
- 2 scintillators as signal trigger
- 1 drift tube chamber: 4 layers 16 wires each (16x4=64 wires)
- Source of signals: cosmic muons (triggered rate ~3 MHz)
- **Event**: muon track reconstructed interpolating 3/4 hits (one per layer)



Observables (6D problem):

- 4 drift times [$t_{\text{drift},1}$, $t_{\text{drift},2}$, $t_{\text{drift},3}$, $t_{\text{drift},4}$]: time for the ionised electrons to reach the wire from the interaction point ($v_{\text{drift}} =$ cm/s) .
- $\theta$: reconstructed track angle
- $N_{\text{hits}}$: average number of hits per time window ("orbit")




Dipartimento di Fisica e Astronomia Galileo Galilei
UNIVERSITÀ DEGLI STUDI DI PADOVA
UniGe | MaLGa

# Data Quality Monitoring

[Grosso, Letizia, AW, Zanetti, et. al., 2023]

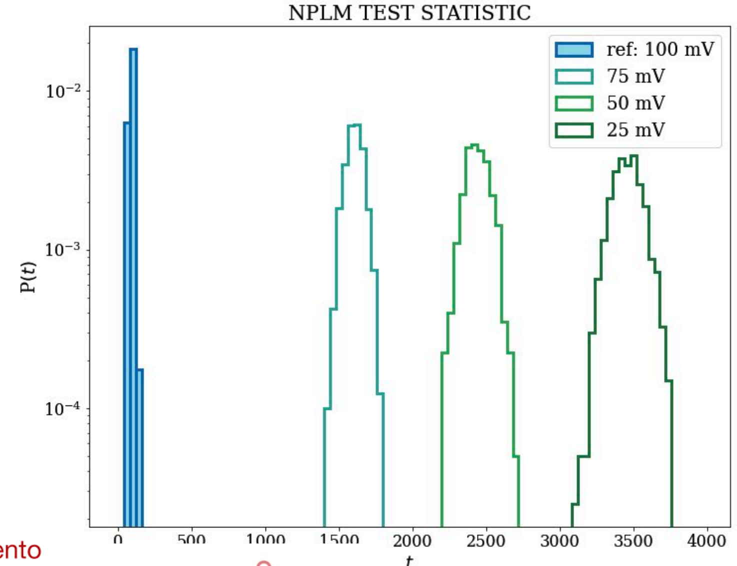
## Much better than standard methods, and **fast enough**

### *n*D DQM

Online monitoring of a DT chamber:

- **Reference sample:** long run in optimal conditions
- **Anomalous samples**: short runs acquired in presence of a controlled anomaly in the value of the **threshold tension** of the DT chamber
- Result of the test statistics

Complete separation of the distributions!



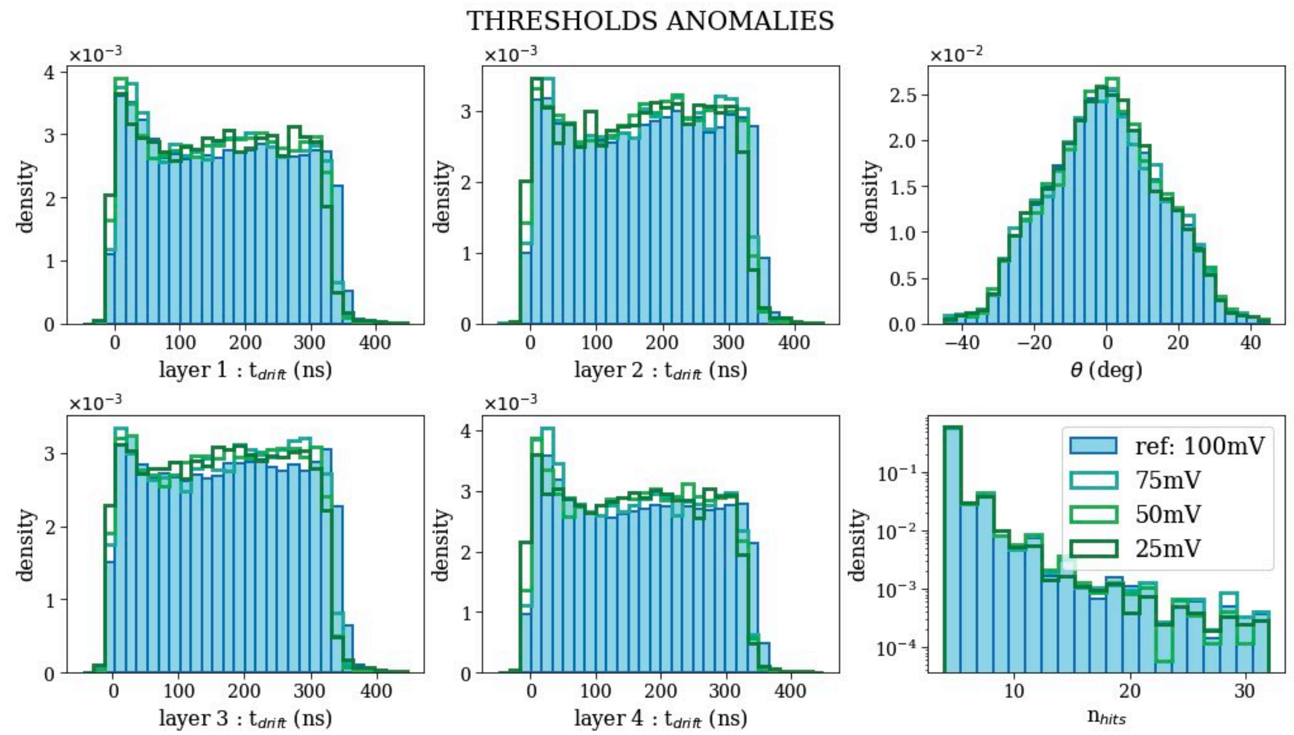

**NPLM with Falkon**

M = 50, $\sigma = 4.84$, $\lambda = 10^{-7}$

N(D) = 5000

$N_{\text{ref}} = 200\,000$

Execution time: $\sim 1.5$ s




Distribution of the observables at different values of the threshold tension

→ more about this in Marco's talk tomorrow!

Dipartimento di Fisica e Astronomia Galileo Galilei
UNIVERSITÀ DEGLI STUDI DI PADOVA
UniGe | MaLGa

# Generative Models Validation

[**To Appear:** Cappelli, Grosso, Letizia, Reyes-González, Zanetti]

A mixture of Gaussians in *d* dimension, vs a Normalising Flow

Tested with NPLM using 10K points, $\ll$ NF training sample size

| $N_{tr}$ \ $d$ | 4 | 8 | 12 | 16 | 20 | 30 |
|---|---|---|---|---|---|---|
| 100k | $9.88^{+1.22}_{-1.29}$ | $8.88^{+1.12}_{-1.19}$ | $14.73^{+1.23}_{-0.94}$ | $16.81^{+1.04}_{-1.06}$ | $14.46^{+1.09}_{-0.84}$ | $14.97^{+1.09}_{-0.84}$ |
| 200k | $4.79^{+1.00}_{-1.07}$ | $9.90^{+0.94}_{-1.05}$ | $9.56^{+1.04}_{-1.04}$ | $8.34^{+0.96}_{-1.09}$ | $6.45^{+0.97}_{-1.07}$ | $7.32^{+0.90}_{-0.81}$ |
| 500k | $1.93^{+1.02}_{-0.99}$ | $3.01^{+0.74}_{-1.13}$ | $3.16^{+1.10}_{-1.02}$ | $5.05^{+1.02}_{-0.99}$ | $2.07^{+0.81}_{-0.97}$ | $3.06^{+1.13}_{-0.86}$ |

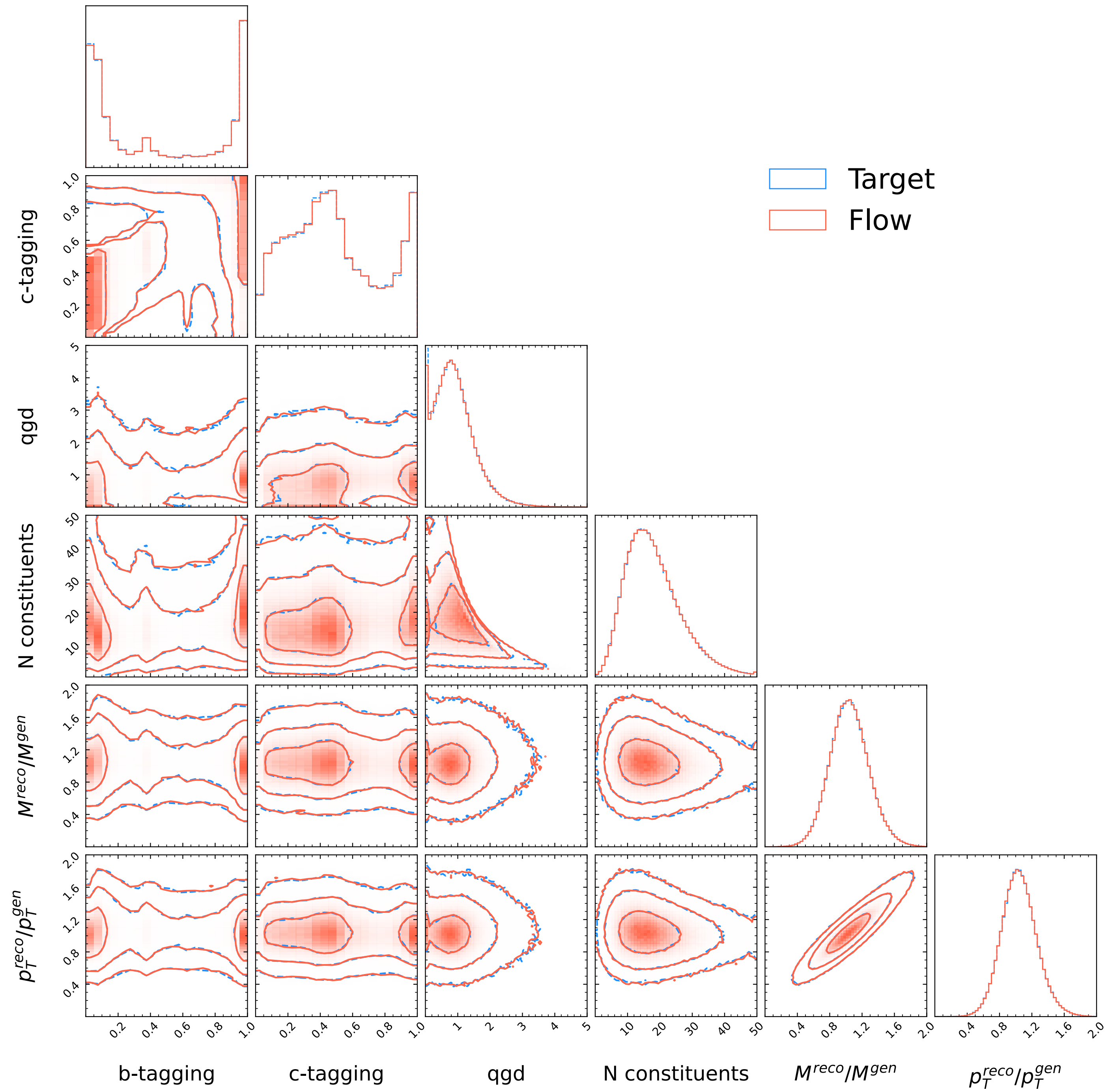
Table 1: Table of median Z-scores obtained with the NPLM method for various NFs models, characterised by training samples of different size ($N_{tr}$) and different number of dimensions ($d$). We report errors estimated as the 68% confidence interval, defined symmetrically around the median value.

Very high Z-scores. Consistently go down as $N_{tr}$ increases

# Generative Models Validation

[**To Appear:** Cappelli, Grosso, Letizia, Reyes-González, Zanetti]

Surrogate detector simulator [Vaselli, Cattafesta, Asenov, Rizzi; 2402.13684].
With realistic-looking 2d marginals:



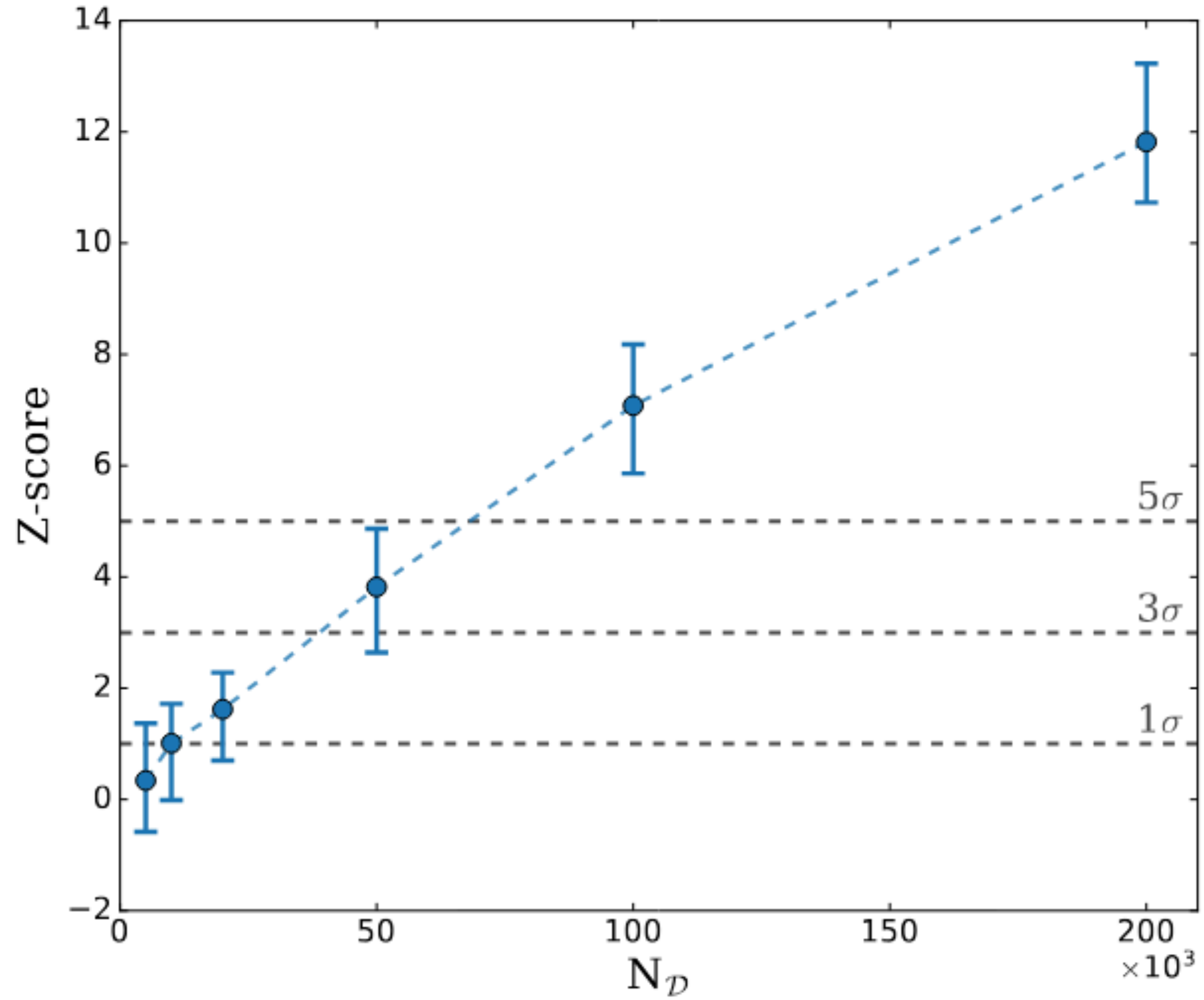
# Generative Models Validation

[**To Appear:** Cappelli, Grosso, Letizia, Reyes-González, Zanetti]

Surrogate detector simulator [Vaselli, Cattafesta, Asenov, Rizzi; 2402.13684].
With realistic-looking 2d marginals:

Tested with NPLM using less data than training size 500K

| $N_{\mathcal{D}}$ | Z-score |
|---|---|
| 5 k | $0.34^{+1.03}_{-0.92}$ |
| 10 k | $1.01^{+0.71}_{-1.02}$ |
| 20 k | $1.62^{+0.66}_{-0.92}$ |
| 50 k | $3.82^{+1.05}_{-1.18}$ |
| 100 k | $7.08^{+1.10}_{-1.22}$ |
| 200 k | $11.82^{+1.41}_{-1.09}$ |



# Generative Models Validation

[**To Appear:** Cappelli, Grosso, Letizia, Reyes-González, Zanetti]

Surrogate detector simulator [Vaselli, Cattafesta, Asenov, Rizzi; 2402.13684].
With realistic-looking 2d marginals:

Tested with NPLM using less data than training size 500K

Personal Conclusions:

- Data augmentation with Generative Models is a **mirage.**
  Because NPLM distinguishes small generated sample from true
- Maybe we can augment some marginal. Maybe we need finite accuracy because of systematics mis-modeling.
  But please explain/demonstrate why and how

# Generative Models Validation

[**To Appear:** Cappelli, Grosso, Letizia, Reyes-González, Zanetti]

Surrogate detector simulator [Vaselli, Cattafesta, Asenov, Rizzi; 2402.13684].
With realistic-looking 2d marginals:

Tested with NPLM using less data than training size 500K

Personal Conclusions:

- Data augmentation with Generative Models is a **mirage.**
  Because NPLM distinguishes small generated sample from true
- Maybe we can augment some marginal. Maybe we need finite accuracy because of systematics mis-modeling.
  But please explain/demonstrate why and how

Objective Conclusion:

- NPLM is very sensitive to mis-modelling
- Could be the best metric for generative models selection

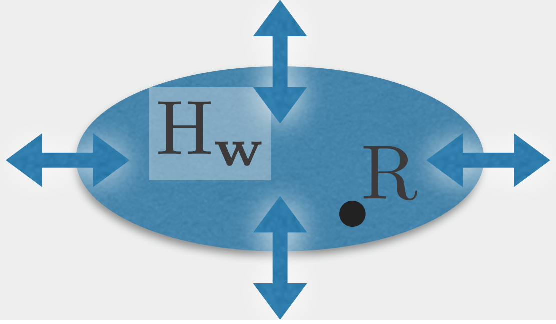
# Take-home messages

## Goodness-of-fit

- A truly profound problem of Science!
- Could serve for model-agnostic BSM searches.
- But also for Data Validation, for DQM, validation of generators including Generative Models
- NPLM in our studies is found better than other methods

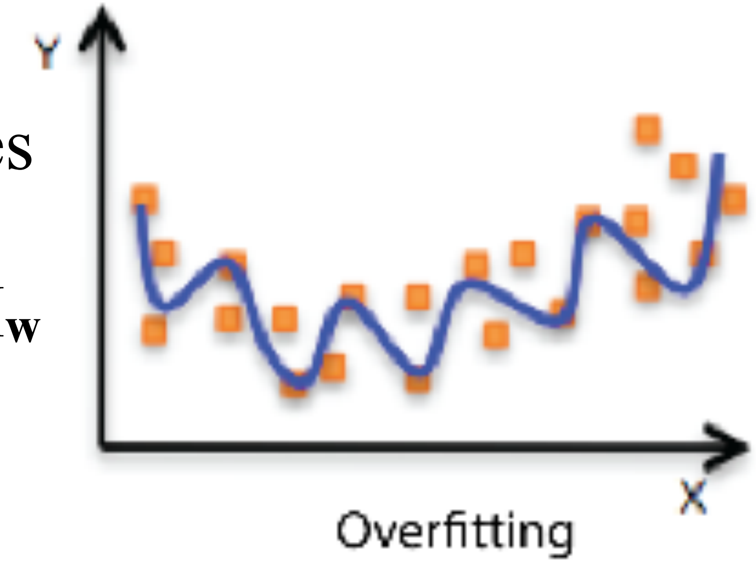
Thank You

# Model Selection




## Which hypotheses (distributions) our (statistical) model contains?

- Not "all of them", otherwise it would fail (overfitting)
- It should contain approximations of all the reasonable ones
- No Statistical Learning notion of model capacity seems reasonable physics measure of volume or boundaries of $H_w$
- Minimal allowed variation scale would sound reasonable, but no theory developed



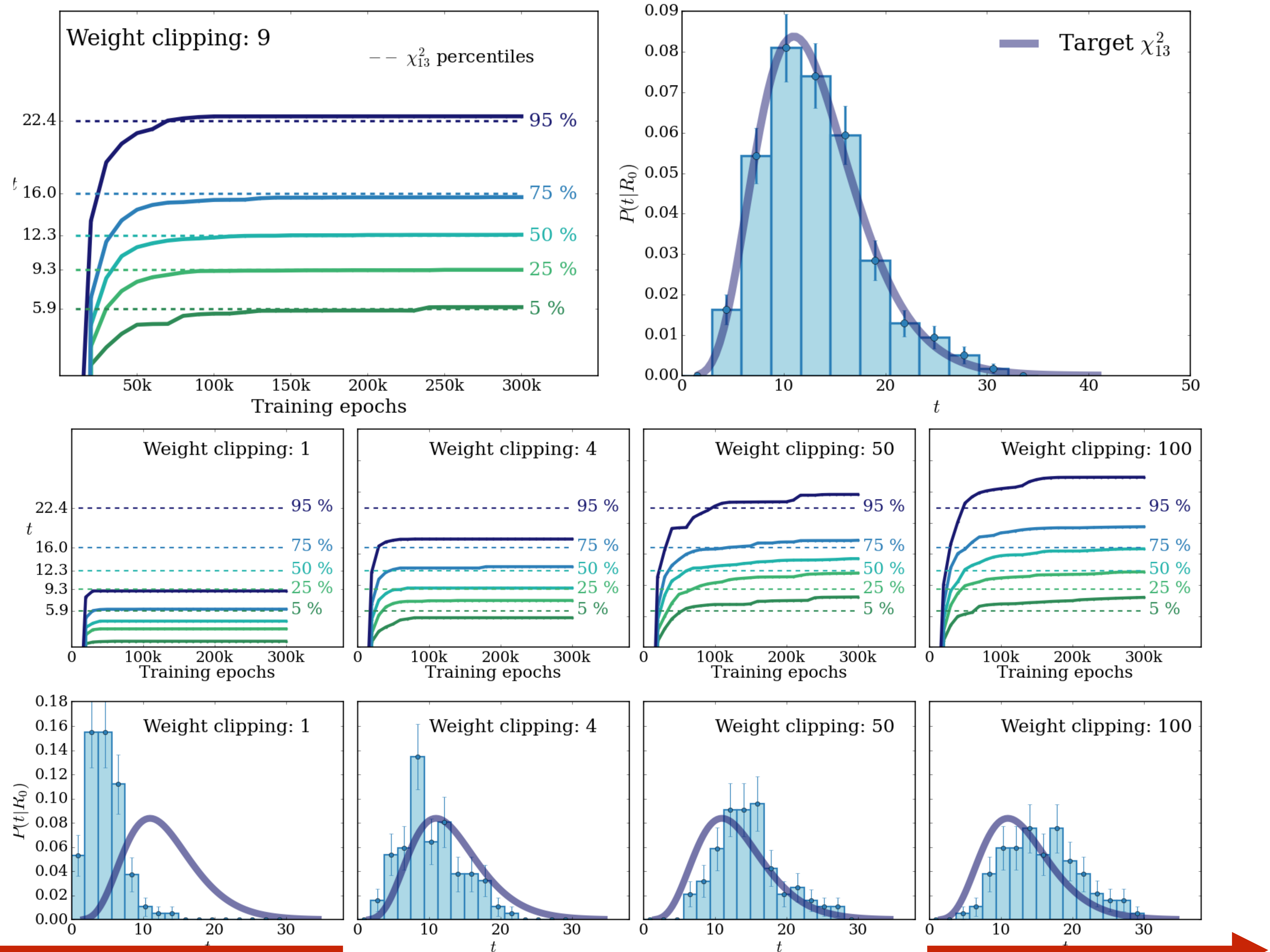

## Waiting for principled approach, solution is **$\chi^2$-compatibility:**

- **Naive** Wilks Theorem application:

P(t|R) is $\chi^2$, with as many d.o.f. as fit parameters (for us, num. of NN par.s)
Provided statistics is large relative to fitted model "complexity"
… or, which is the same …
Provided model is "simple enough", for given data statistics

- Asy. For. violation = sensitivity to low-statistics portion of dataset = overfitting
- Regularisation by Weight Clipping, that forbids sharp variations
- NN with too many parameters cannot be made $\chi^2$-compatible. Take largest allowed

# Weight Clipping Selection

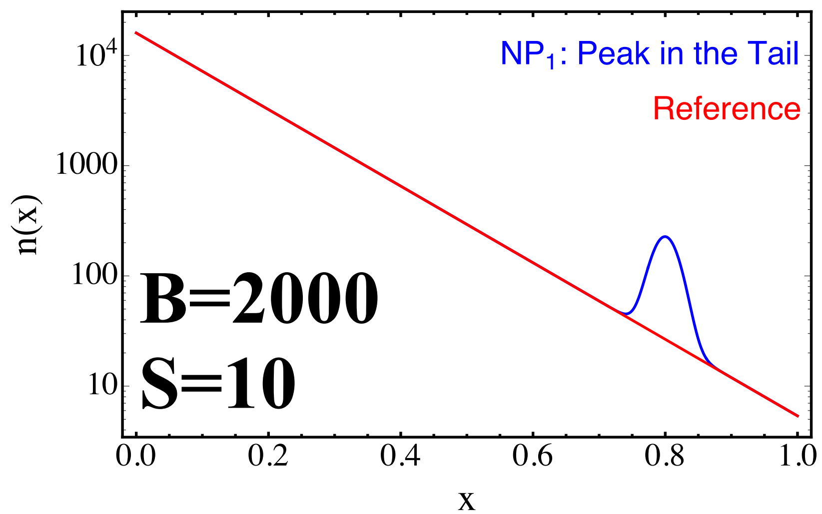


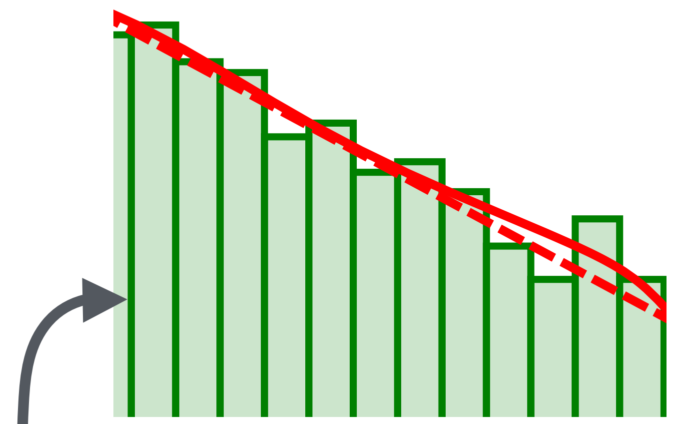

Asy. For. violation by fit parameters boundary

Asy. For. violation by sensitivity to sparse data points

# Illustrating Performances

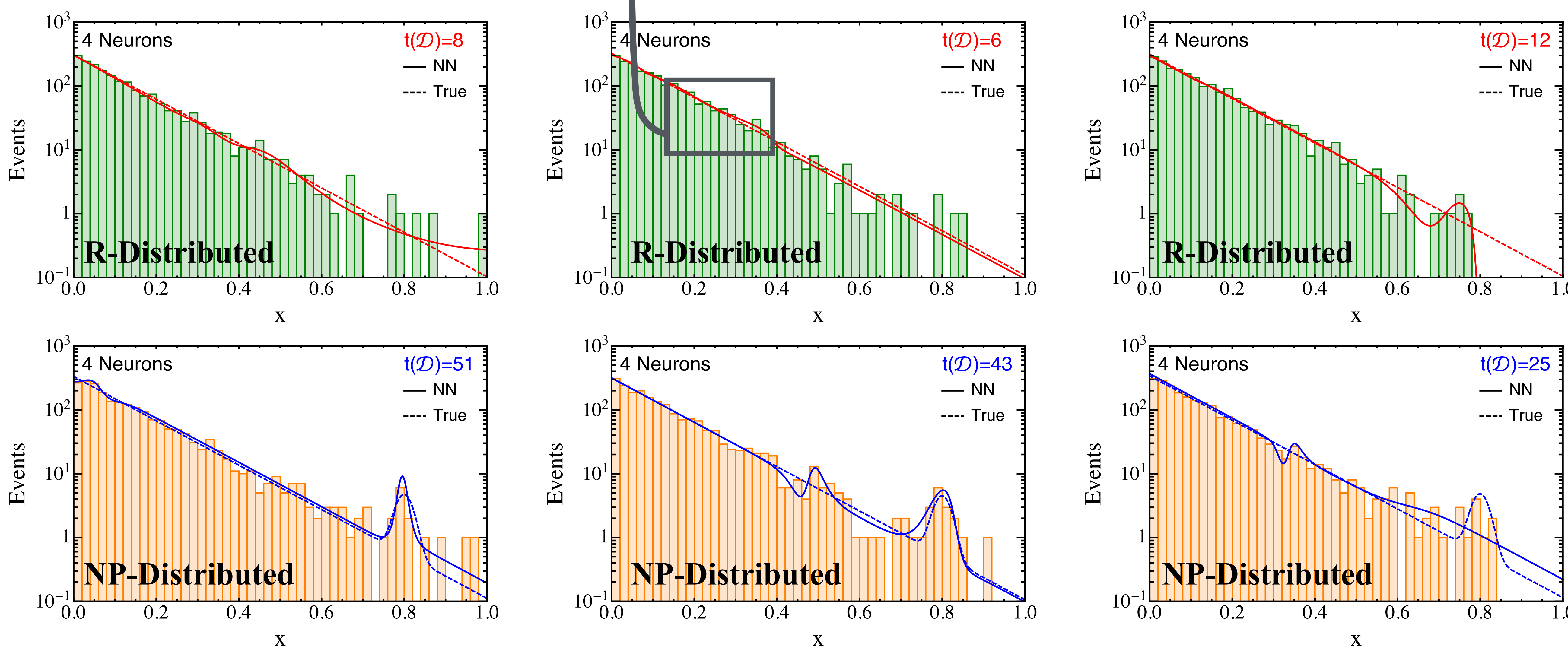
(Simple 1d example with exponential Reference)



**Bins:** Non-discrepant data fluctuations wash out reach

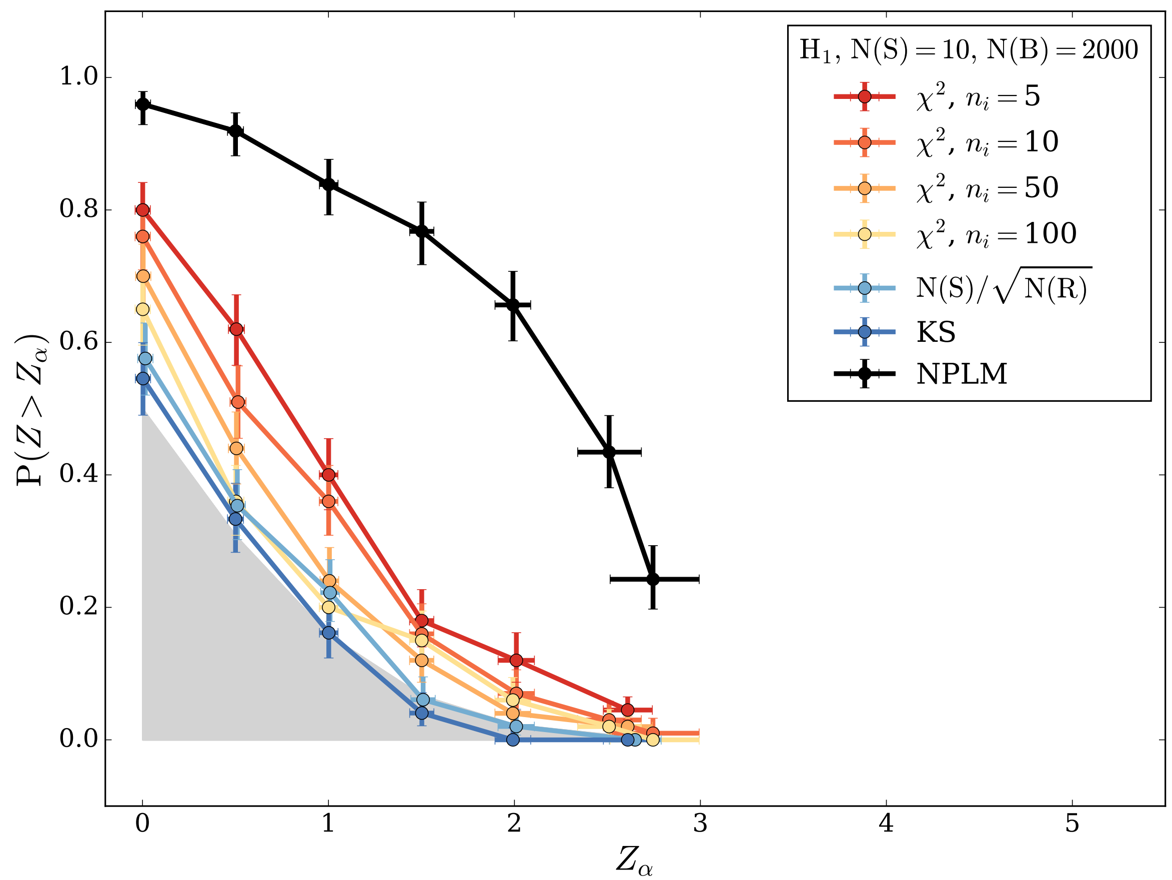
**NN:** Smooth curve. Can handle non-discrepant data



# Illustrating Performances

(Simple 1d example with exponential Reference)

Probability to find evidence of $\mathrm{R}$ being wrong at some level of confidence.




We are better than binned $\chi^2$ because our model has less parameters but same effective expressive power.

Same reason why bins are outdated as statistical models.

Gap to bins grows (exponentially) with (the curse of) dimensionality.