Learning New Physics from an (Imperfect) Machine

Andrea Wulzer

UNIVERSITÀ DEGLI STUDI DI PADOVA

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Based on:

D'Agnolo, AW, 2018

D'Agnolo, Grosso, Pierini, AW, Zanetti, 2019

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

The Challenge

What if* the **RIGHT BSM model** has not been formulated?

***very likely**

The Challenge

What if* the **RIGHT BSM model** has not been formulated?

***very likely**

Would we **still see the SM fail to describe data?**

The Challenge

What if* the **RIGHT BSM model** has not been formulated?

***very likely**

Would we **still see the SM fail to describe data?**

Most likely not !

BSM is tiny departure from SM, or large in tiny prob. region

Affecting few (unknown) observables over ∞ many we can measure

The Challenge

What if* the **RIGHT BSM model** has not been formulated?

***very likely**

Would we **still see the SM fail to describe data?**

Most likely not !

BSM is tiny departure from SM, or large in tiny prob. region

Affecting few (unknown) observables over ∞ many we can measure

Regular New Physics searches are **Model Dependent**

Choose observables sensitive to **one BSM model**

This observable in general **not** sensitive to **another BSM model**

The Challenge

What if* the **RIGHT BSM model** has not been formulated?

***very likely**

Would we **still see the SM fail to describe data?**

Most likely not !

BSM is tiny departure from SM, or large in tiny prob. region

Affecting few (unknown) observables over ∞ many we can measure

Regular New Physics searches are **Model Dependent**

Choose observables sensitive to **one BSM model**

This observable in general **not** sensitive to **another BSM model**

We must design **Model Independent** searches

aimed at detecting “generic” data departures from SM

The Challenge

What if* the **RIGHT BSM model** has not been formulated?

***very likely**

Would we **still see the SM fail to describe data?**

Most likely not !

BSM is tiny departure from SM, or large in tiny prob. region

Affecting few (unknown) observables over ∞ many we can measure

Regular New Physics searches are **Model Dependent**

Choose observables sensitive to **one BSM model**

This observable in general **not** sensitive to **another BSM model**

We must design **Model Independent** searches

aimed at detecting “generic” data departures from SM

SM = “Reference Model”, to be compared with data
without reference to alternative physics model

Maximum Likelihood

(Foundation of entire LHC statistical practice)

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

I.i.d. measurements of, e.g., reconstructed particle momenta in a region of interest

Maximum Likelihood

(Foundation of entire LHC statistical practice)

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

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

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

depending on **parameters** (composite)

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

Maximum Likelihood

(Foundation of entire LHC statistical practice)

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

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

Alternative Distribution: $n(x|\mathbf{w})$
depending on **parameters** (composite)

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

Test statistic:

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

Model Dependent Strategy

$$n(x|\mathbf{w}) = n(x|\mathrm{NP})$$

Alternative as predicted by “NP” model.
Few, or no, free parameters

Maximum Likelihood

(Foundation of entire LHC statistical practice)

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

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

Alternative Distribution: $n(x|\mathbf{w})$
depending on **parameters** (composite)

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

Test statistic:

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

Model Dependent Strategy

$$n(x|\mathbf{w}) = n(x|\mathrm{NP})$$

Alternative as predicted by “NP” model.
Few, or no, free parameters

Model Independent Strategy

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

Alternative in parametrised form.
$f(x;\mathbf{w})$ is flexible function approximant

Maximum Likelihood

(Foundation of entire LHC statistical practice)

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

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

Alternative Distribution: $n(x|\mathbf{w})$
depending on **parameters** (composite)

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

Test statistic:

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

Model Dependent Strategy

$$n(x|\mathbf{w}) = n(x|\mathrm{NP})$$

Alternative as predicted by “NP” model.
Few, or no, free parameters

Model Independent Strategy

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

Alternative in parametrised form.
$f(x;\mathbf{w})$ is flexible function approximant

If $f(x;\mathbf{w})$ is **piece-wise constant**

Binned Histogram Test
(e.g., MUSiC at CMS)

Maximum Likelihood

(Foundation of entire LHC statistical practice)

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

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

Alternative Distribution: $n(x|\mathbf{w})$
depending on **parameters** (composite)

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

Test statistic:

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

Model Dependent Strategy

$$n(x|\mathbf{w}) = n(x|\mathrm{NP})$$

Alternative as predicted by “NP” model.
Few, or no, free parameters

Model Independent Strategy

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

Alternative in parametrised form.
$f(x;\mathbf{w})$ is flexible function approximant

If $f(x;\mathbf{w})$ is **a neural network**

Our Proposal

Maximum Likelihood

(Foundation of entire LHC statistical practice)

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

$N\,P(x)$

$dx\,n(x)$

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

replace histograms with NN, literally!

Highly motivated attempt:

- NN "effective" flexible but smooth function approx.
- Often used as **alternative to hist.** to fit distributions
- Better dimensionality scaling

Model Dependent Strategy

$$n(x|\mathbf{w}) = n(x|\mathrm{NP})$$

Alternative as predicted by "NP" model
Few, or no, free parameters

Model Independent Strategy

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

Alternative in parametrised form.
$f(x; \mathbf{w})$ is flexible function approximant

If $f(x; \mathbf{w})$ is **a neural network**

Our Proposal

Maximum Likelihood Loss

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

We evaluate “t” by supervised training using “ML-Loss”

- Observed (or Toy) **Data are class “1”**
- Alternatives to use ML-Loss exist, but nothing to gain

Maximum Likelihood Loss

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

We evaluate “t” by supervised training using “ML-Loss”

- Observed (or Toy) **Data are class “1”**
- Alternatives to use ML-Loss exist, but nothing to gain

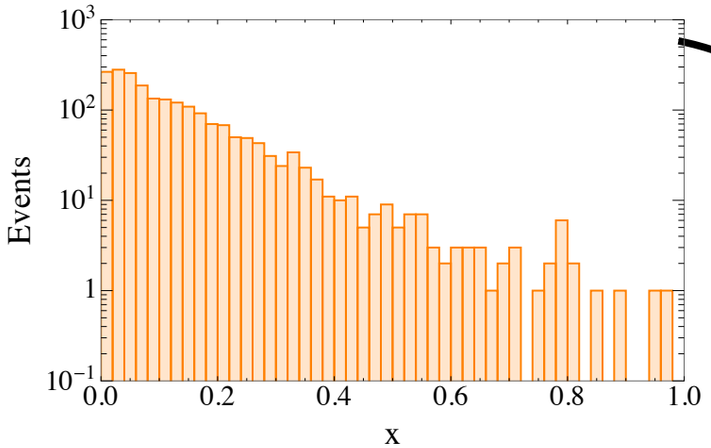
Class “0” is a **Reference Sample** $\mathcal{R} = \{x_i\},\ i = 1, \ldots, \mathcal{N}_{\mathcal{R}}$

- SM-distributed synthetic instances of the features “x”
- Can come from **Monte Carlo**, or **Data Driven**
- Nothing different from “**background sample**” in regular searches
- Preferably, more abundant than the data: $\mathcal{N}_{\mathcal{R}} \gg N(\mathrm{R})$

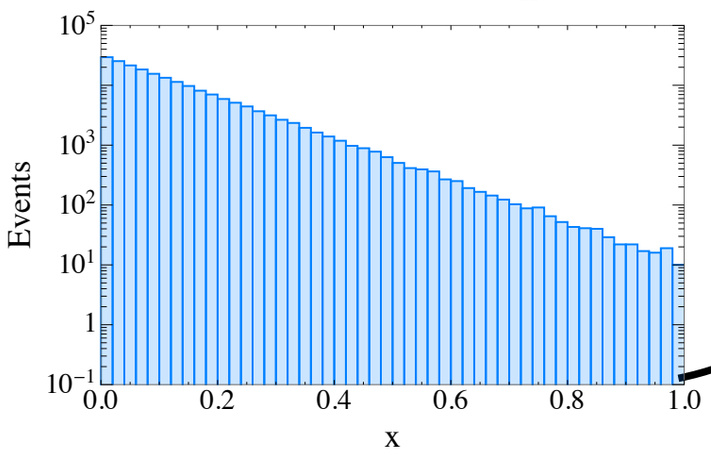
The Algorithm

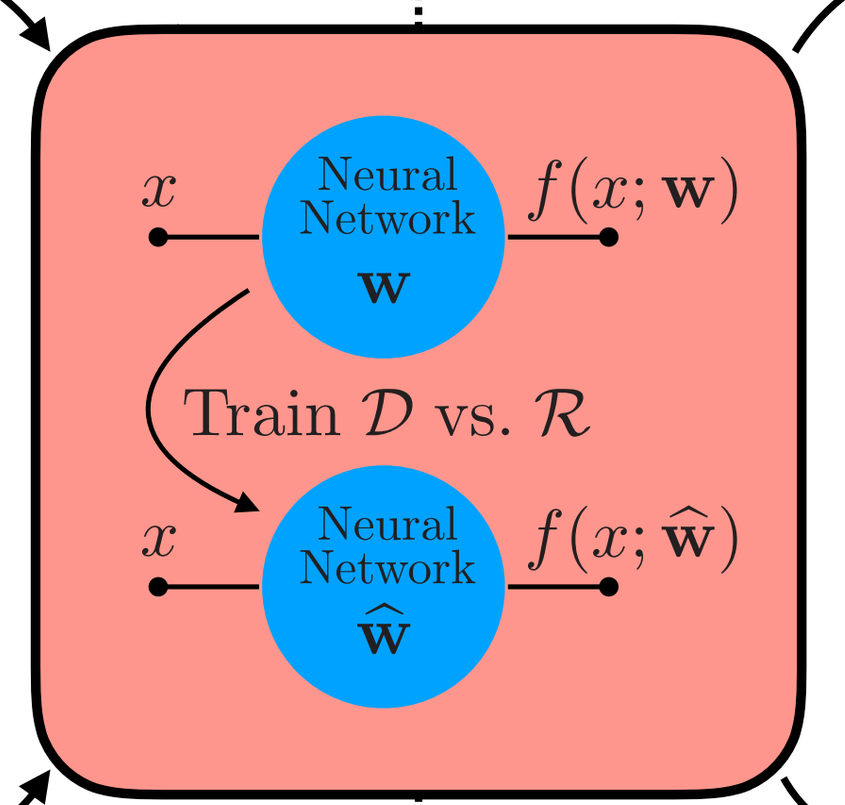
INPUT

Data sample $\mathcal{D}$



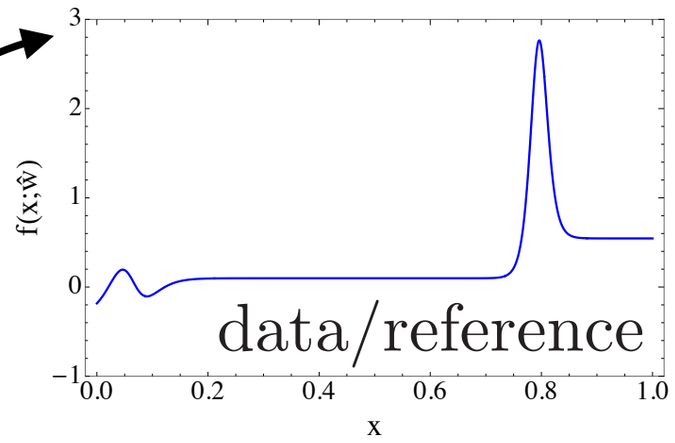
Reference sample $\mathcal{R}$





OUTPUT

Dist. log ratio



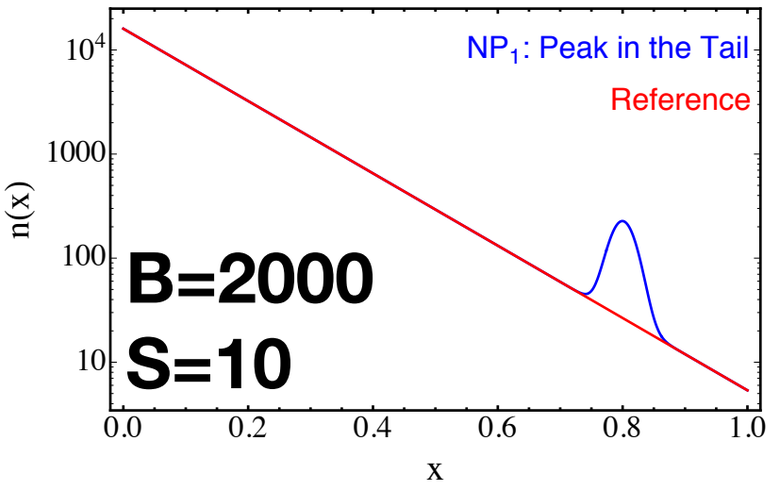
$$f(x;\widehat{\mathbf{w}}) \simeq \log\left[\frac{n(x|\mathrm{T})}{n(x|\mathrm{R})}\right]$$

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

$$t(\mathcal{D}) = -2 \operatorname*{Min}_{\{\mathbf{w}\}} L[f]$$

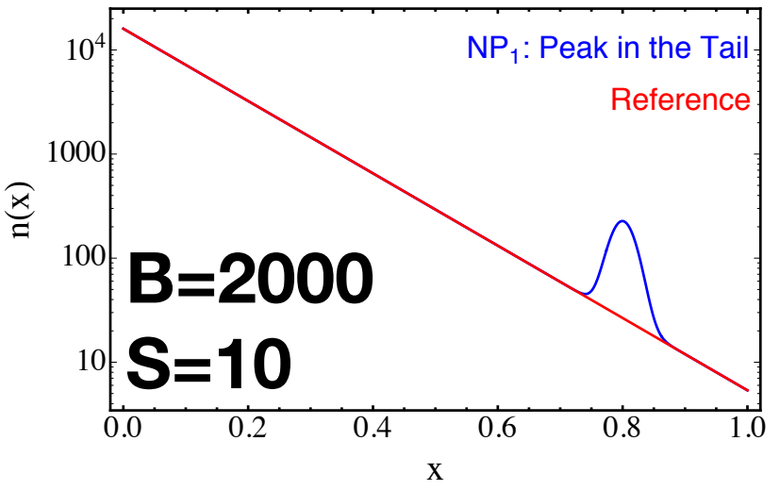
Illustrating Performances

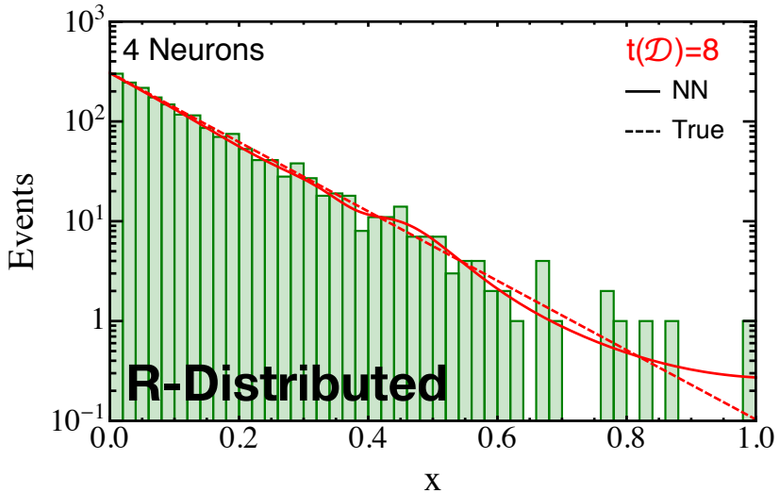
(Simple 1d example with exponential Reference)

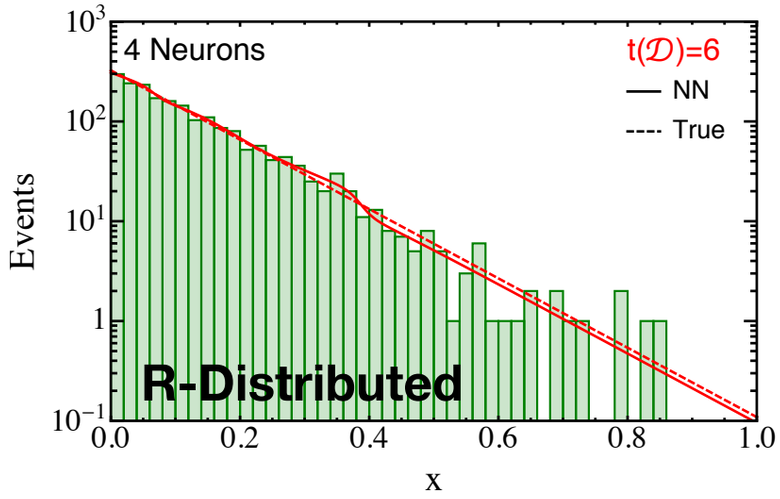


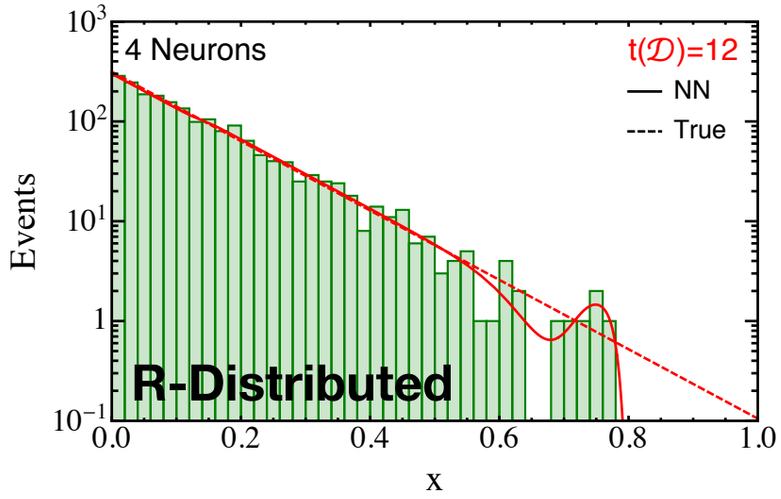
Illustrating Performances

(Simple 1d example with exponential Reference)



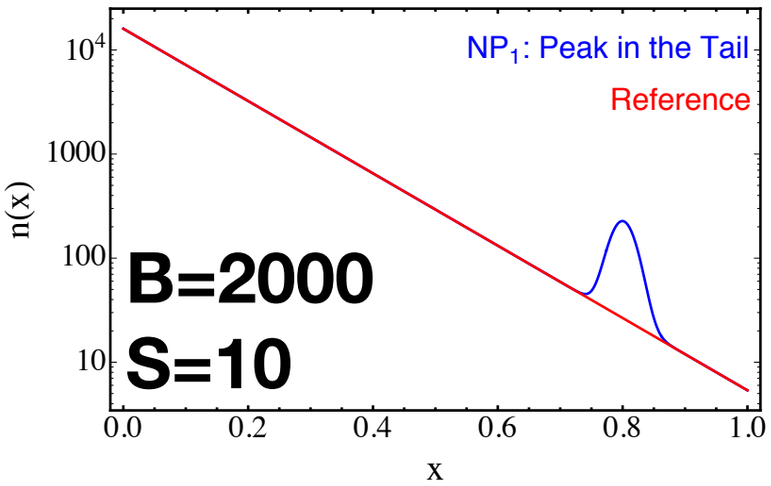


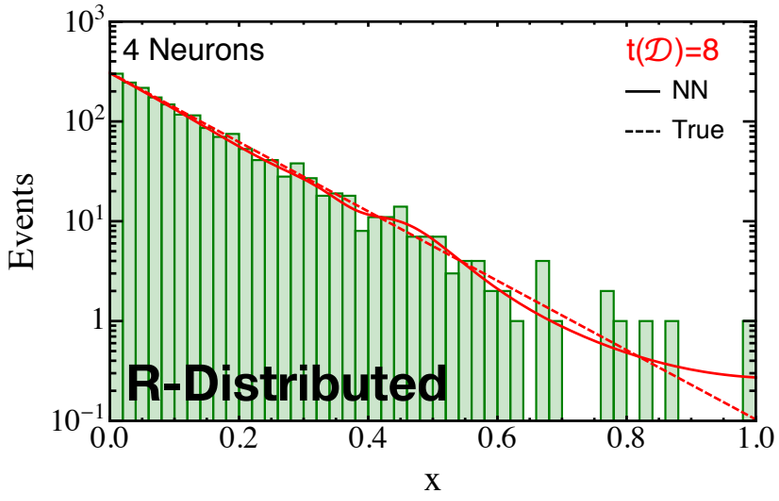


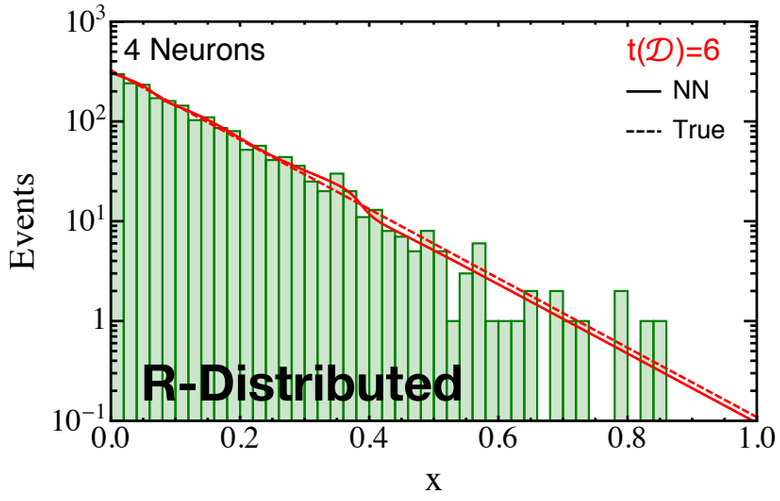


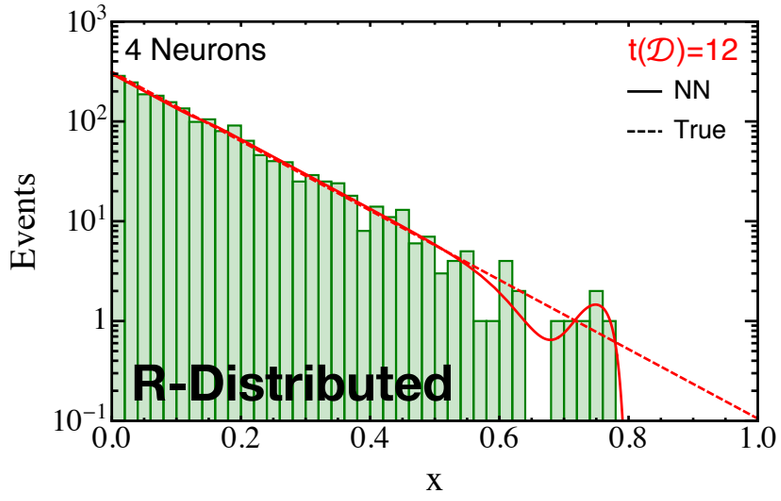
Illustrating Performances

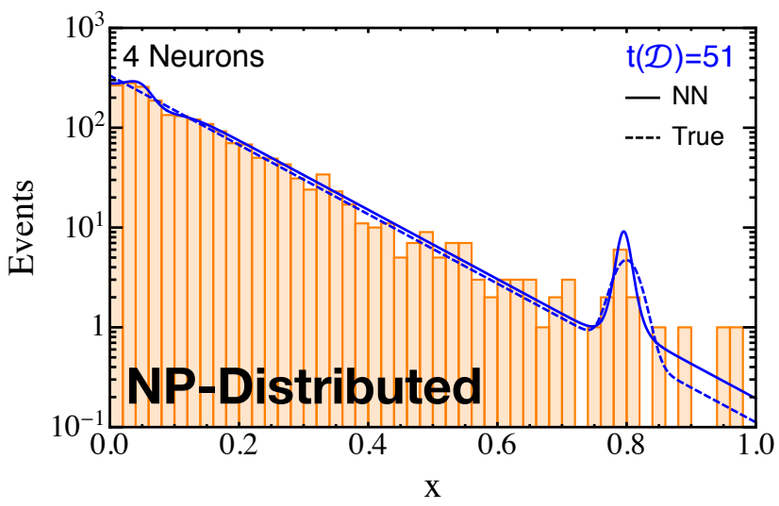
(Simple 1d example with exponential Reference)

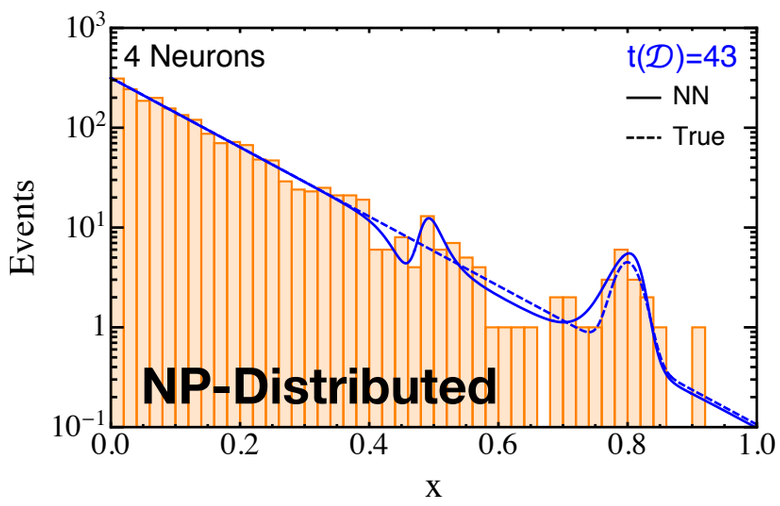


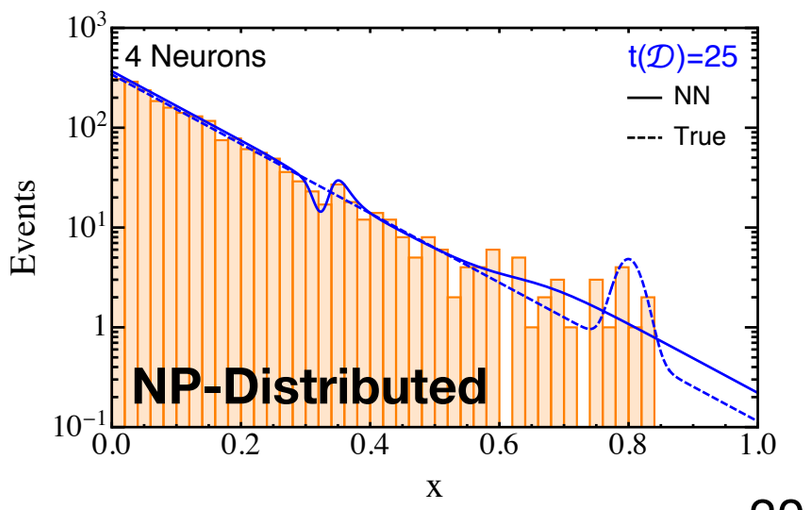










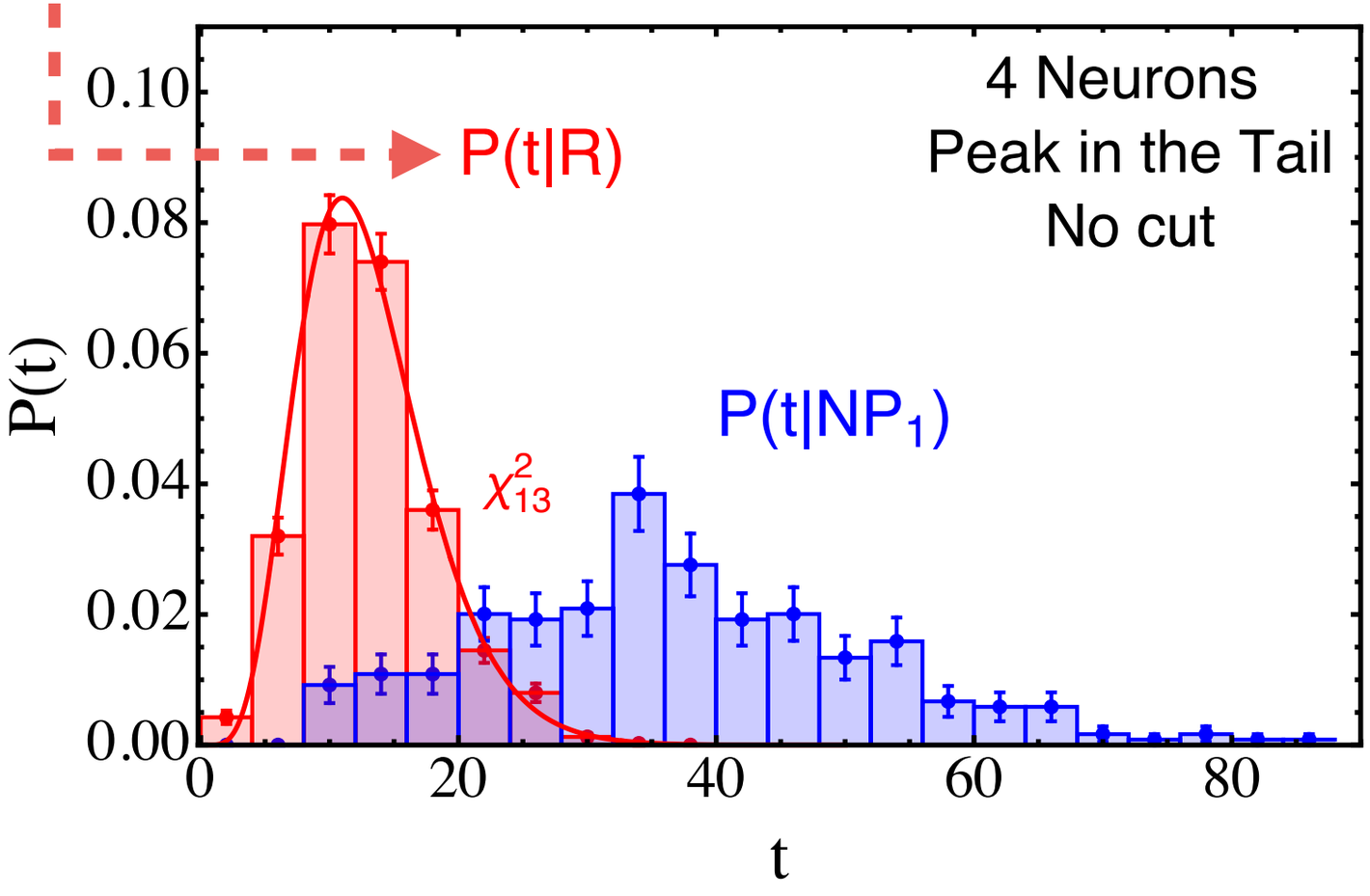


Quantifying Performances

(Simple 1d example with exponential Reference)

Distribution of the test statistic "t" in Reference Hypothesis

- Will give us observed p-value: $p = \int_{t_{\mathrm{obs}}} P(t|\mathrm{R})$
- Computed by repeating training on Reference-distributed Toy data

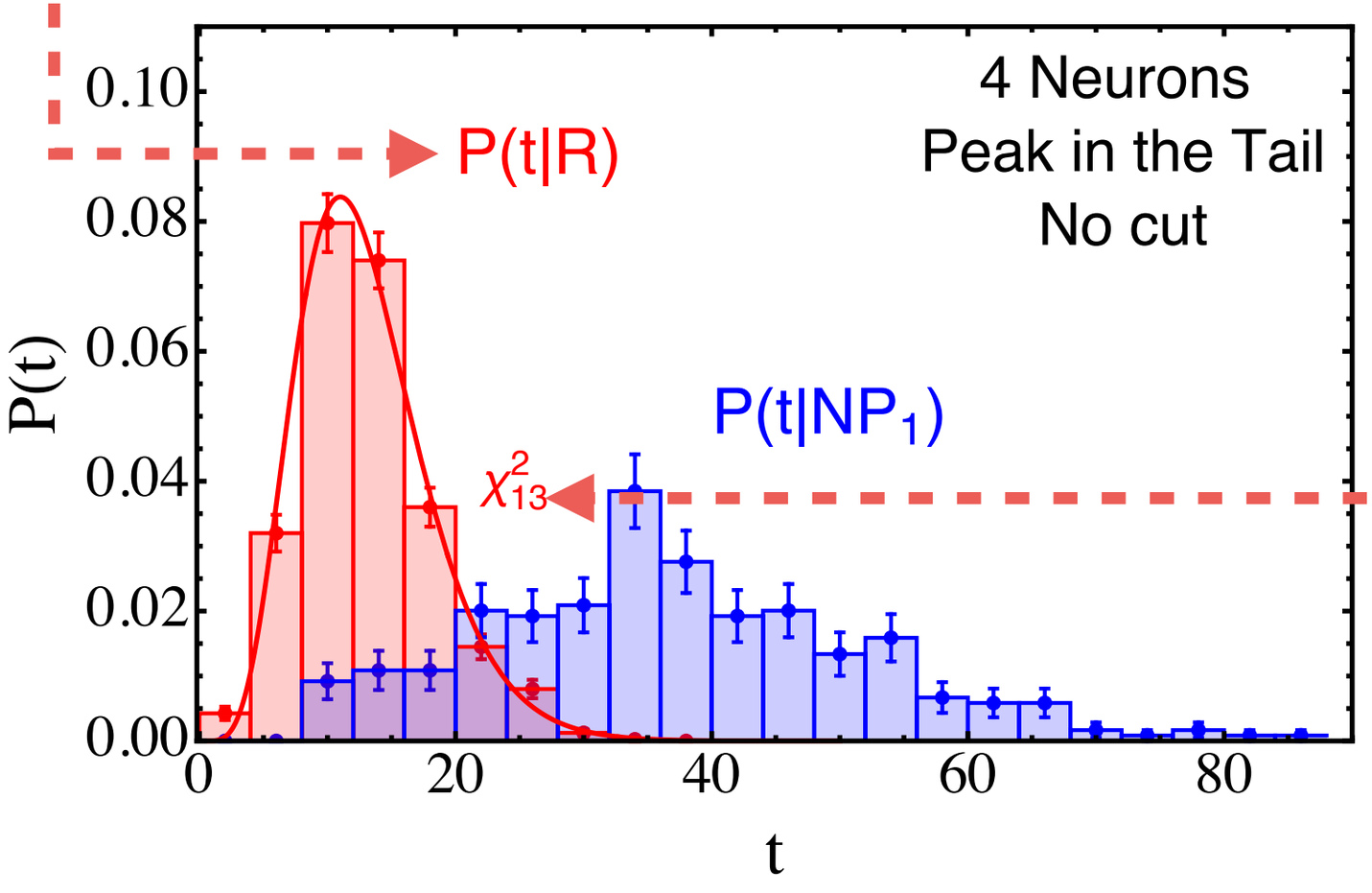


Quantifying Performances

(Simple 1d example with exponential Reference)

Distribution of the test statistic "t" in Reference Hypothesis

- Will give us observed p-value: $p = \int_{t_{\mathrm{obs}}} P(t|\mathrm{R})$
- Computed by repeating training on Reference-distributed Toy data



Notice agreement with "Asymptotic Formula"

(more on this later)

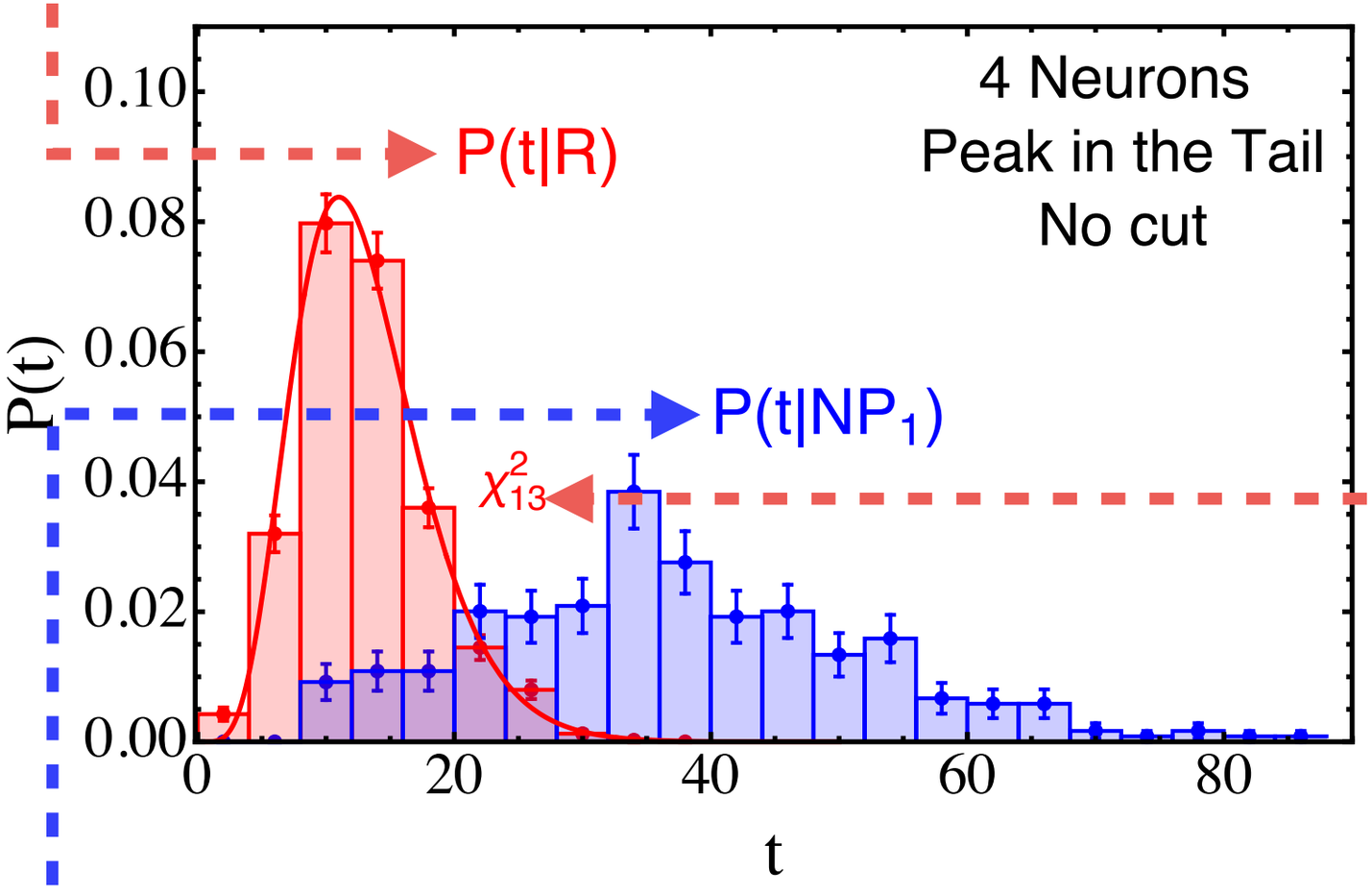
Quantifying Performances

(Simple 1d example with exponential Reference)

Distribution of the test statistic "t" in Reference Hypothesis

Will give us observed p-value: $p = \int_{t_{\rm obs}} P(t|\mathrm{R})$

Computed by repeating training on Reference-distributed Toy data



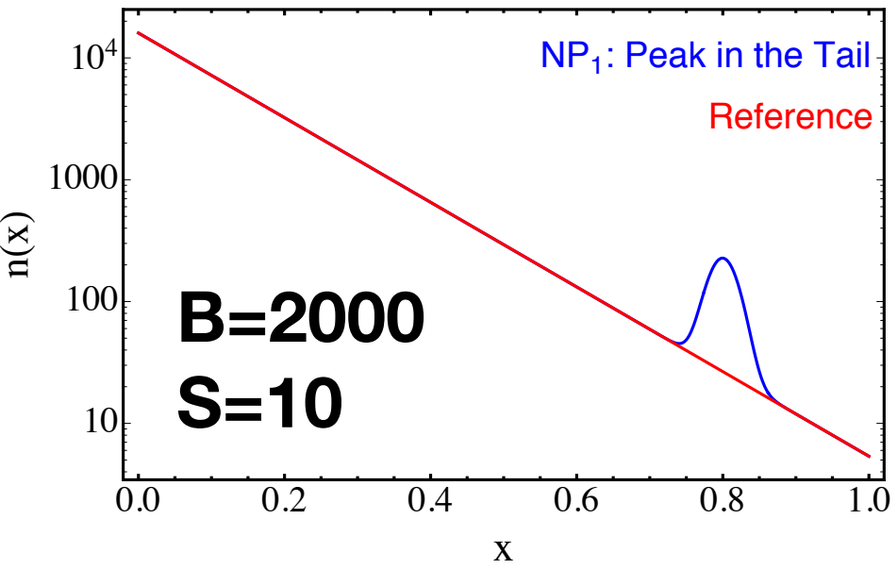
Notice agreement with "Asymptotic Formula"
(more on this later)

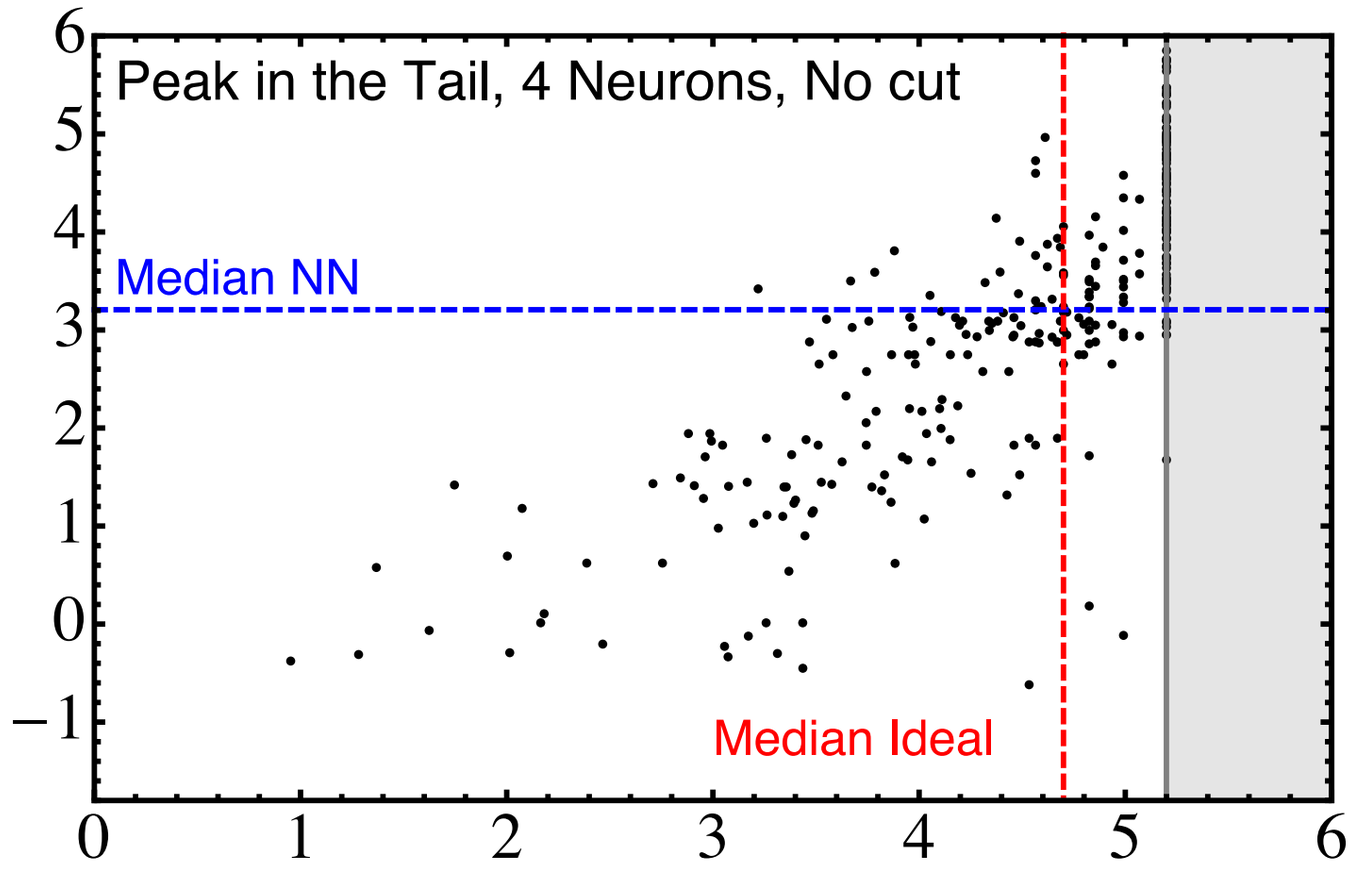
Distribution of "t" in one New Physics Model Hypothesis

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

Quantifying Performances

(Simple 1d example with exponential Reference)





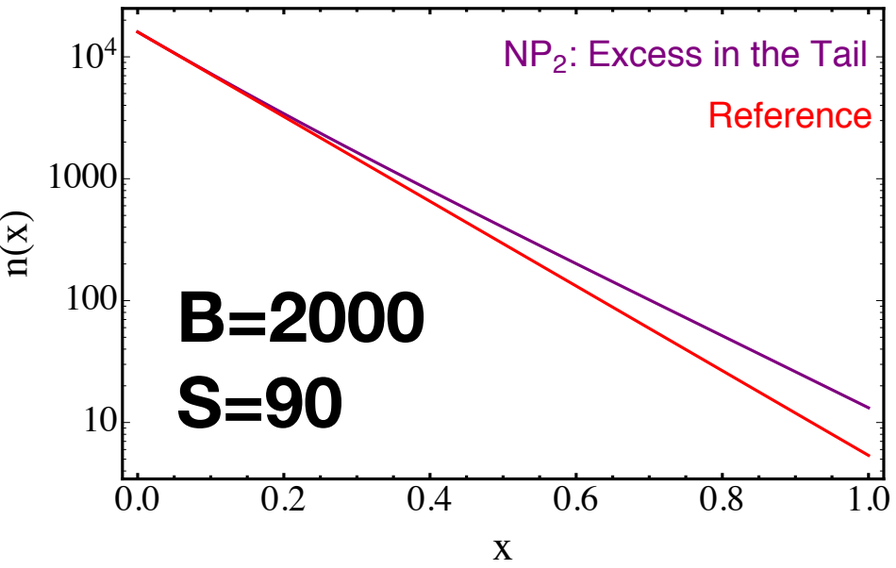
"Ideal Z-score": Z_{id}

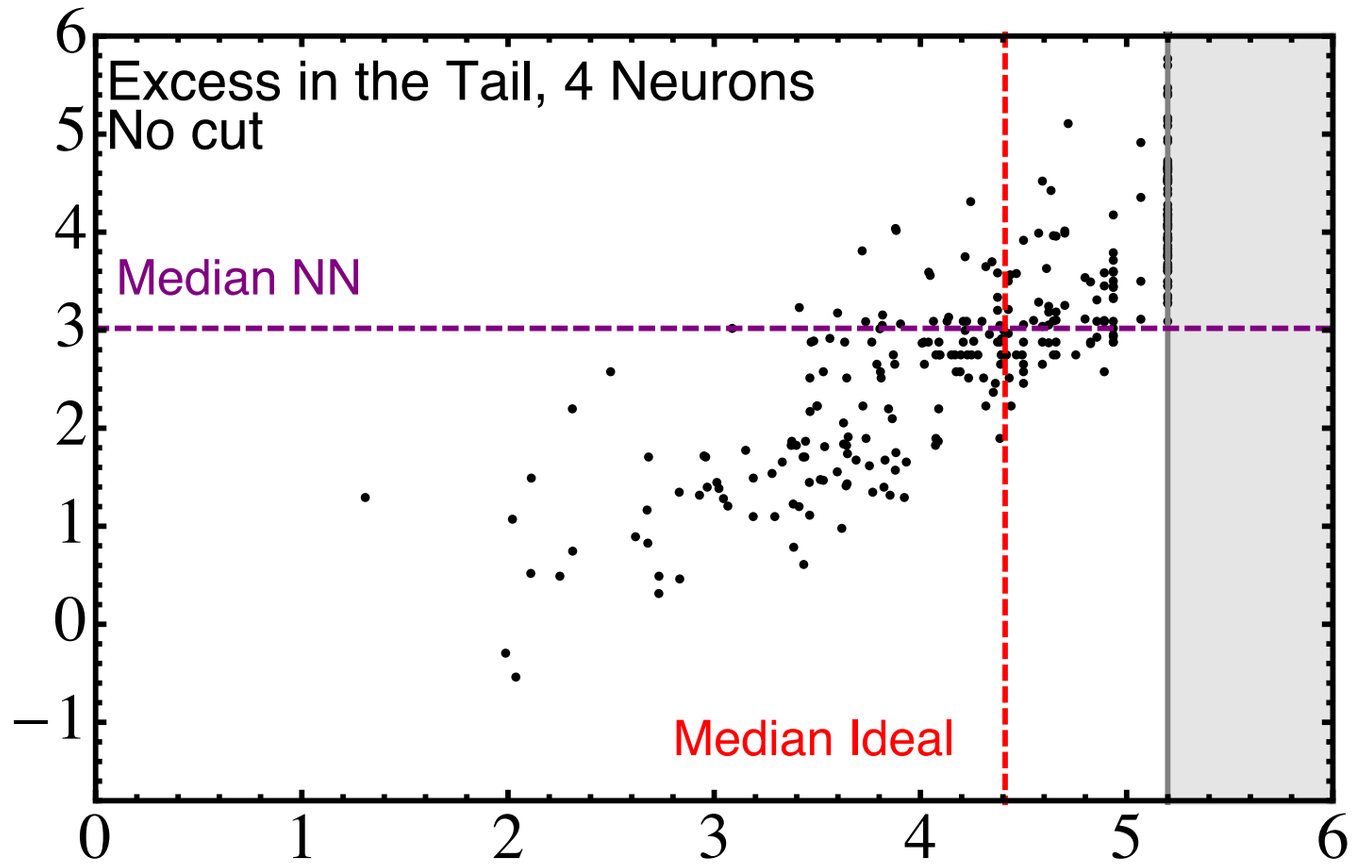
A "measure of dataset discrepancy"

(the Z-score of optimal test for NP1 model)

Quantifying Performances

(Simple 1d example with exponential Reference)





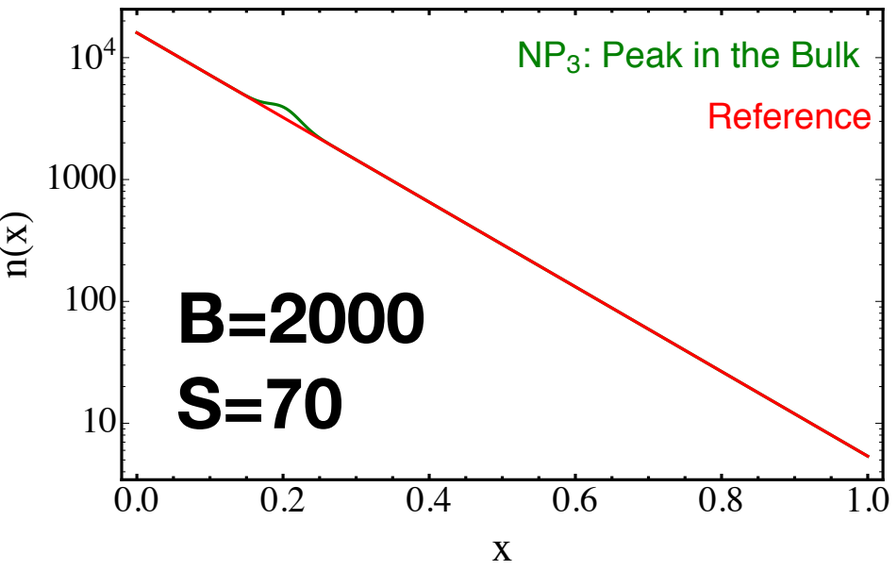
"Ideal Z-score": Z_{id}

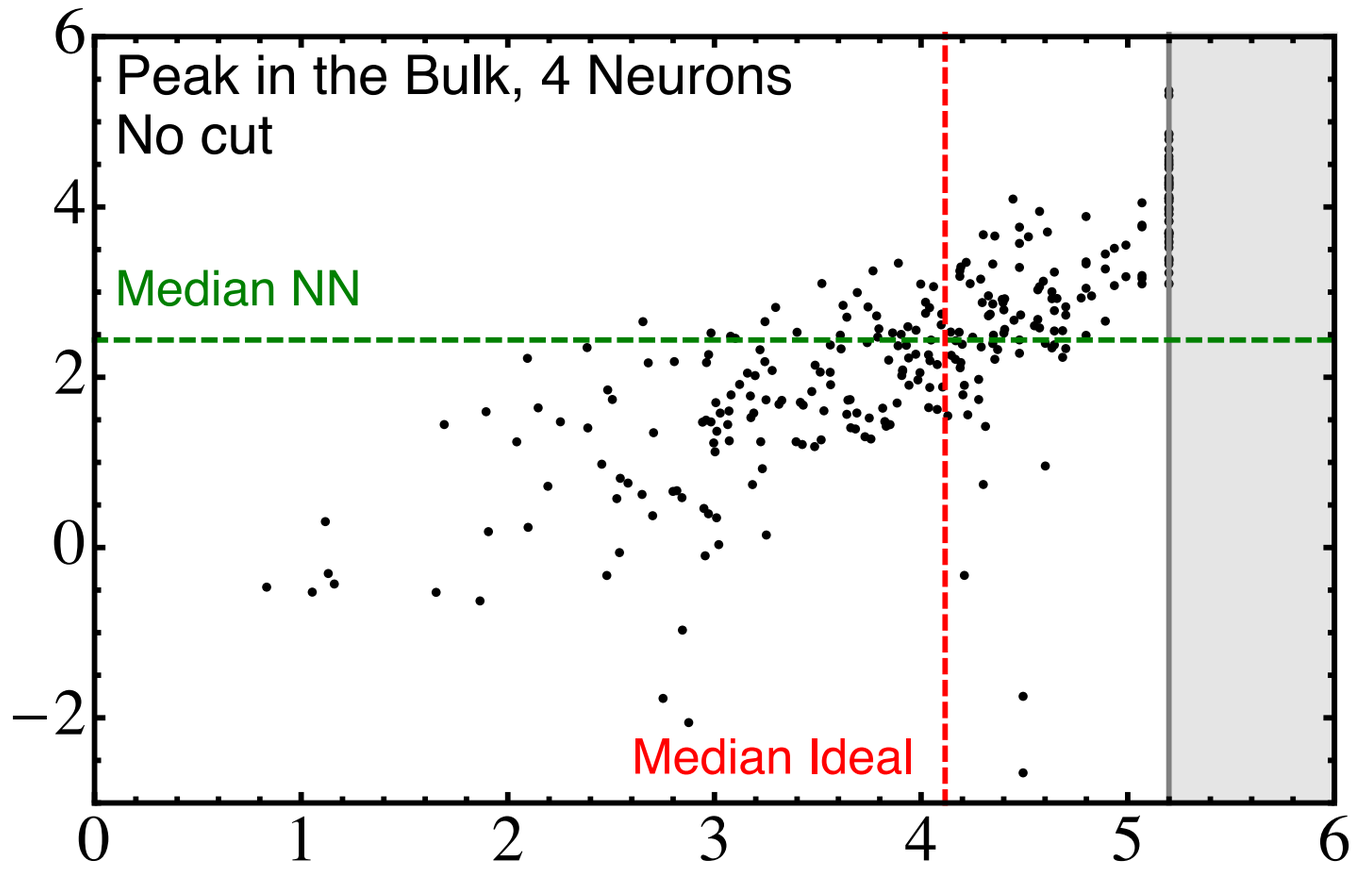
A "measure of dataset discrepancy"

(the Z-score of optimal test for NP2 model)

Quantifying Performances

(Simple 1d example with exponential Reference)





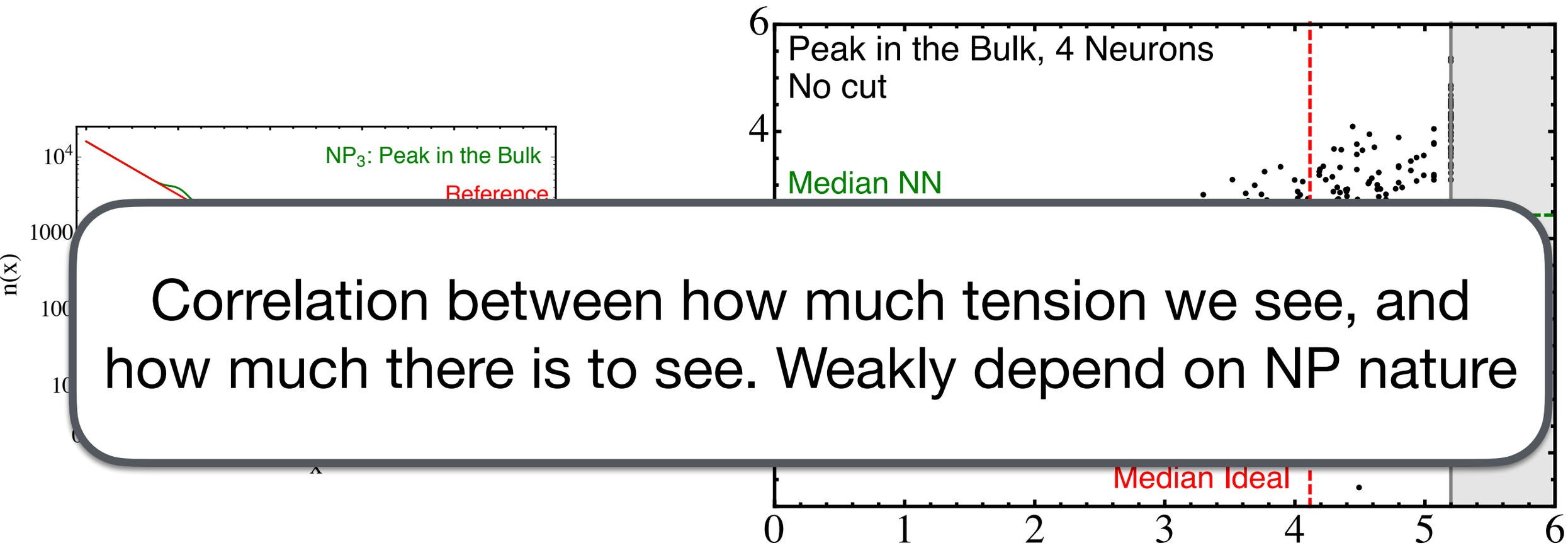
"Ideal Z-score": Z_{id}

A "measure of dataset discrepancy"

(the Z-score of optimal test for NP3 model)

Quantifying 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_{id}

A "measure of dataset discrepancy"

(the Z-score of optimal test for NP3 model)

Asymptotic Formulae

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

By Wilks-Wald Theorem, expect P(t|R) a χ^2, with as many d.o.f. as fit parameters (for us, number of NN par.s)

Asymptotic Formulae

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

By Wilks-Wald Theorem, expect P(t|R) a χ^2, with as many d.o.f. as fit parameters (for us, number of NN par.s)

Provided statistics is large relative to “complexity” of model being fitted

or, which is the same

Provided fit model “simple enough”, for given data statistics

Asymptotic Formulae

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

By Wilks-Wald Theorem, expect P(t|R) a χ^2, with as many d.o.f. as fit parameters (for us, number of NN par.s)

Provided statistics is large relative to “complexity” of model being fitted

or, which is the same

Provided fit model “simple enough”, for given data statistics

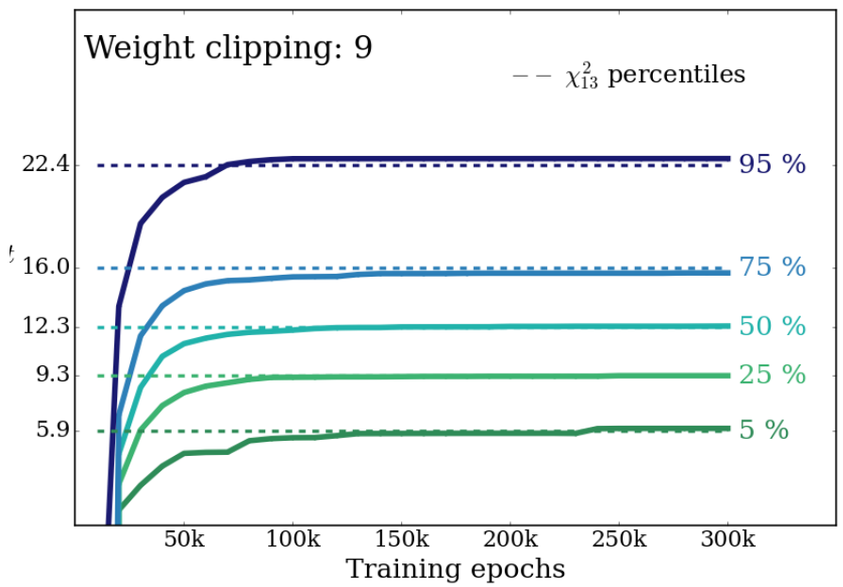
We use **χ^2-compatibility** as **Model Selection criterion**

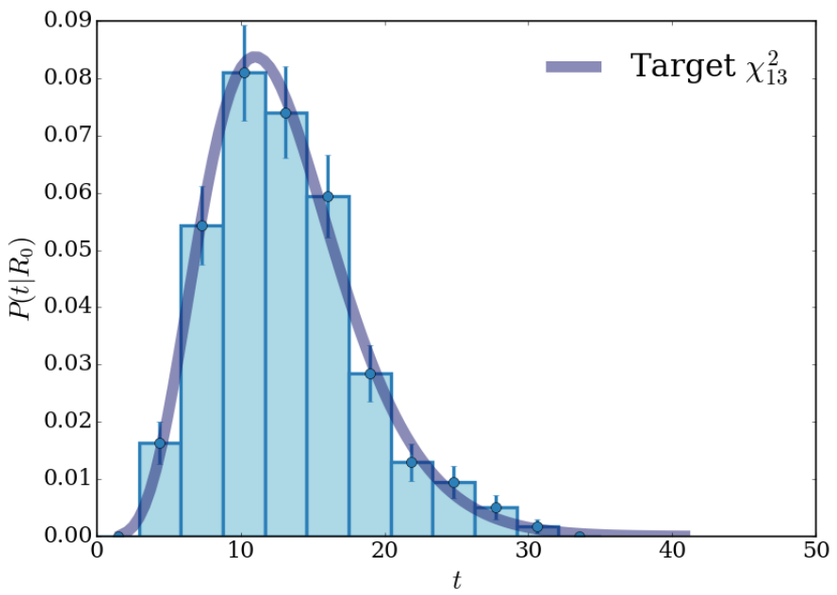
Asy.For. violation = sensitivity to low-statistics portion of dataset = overfitting

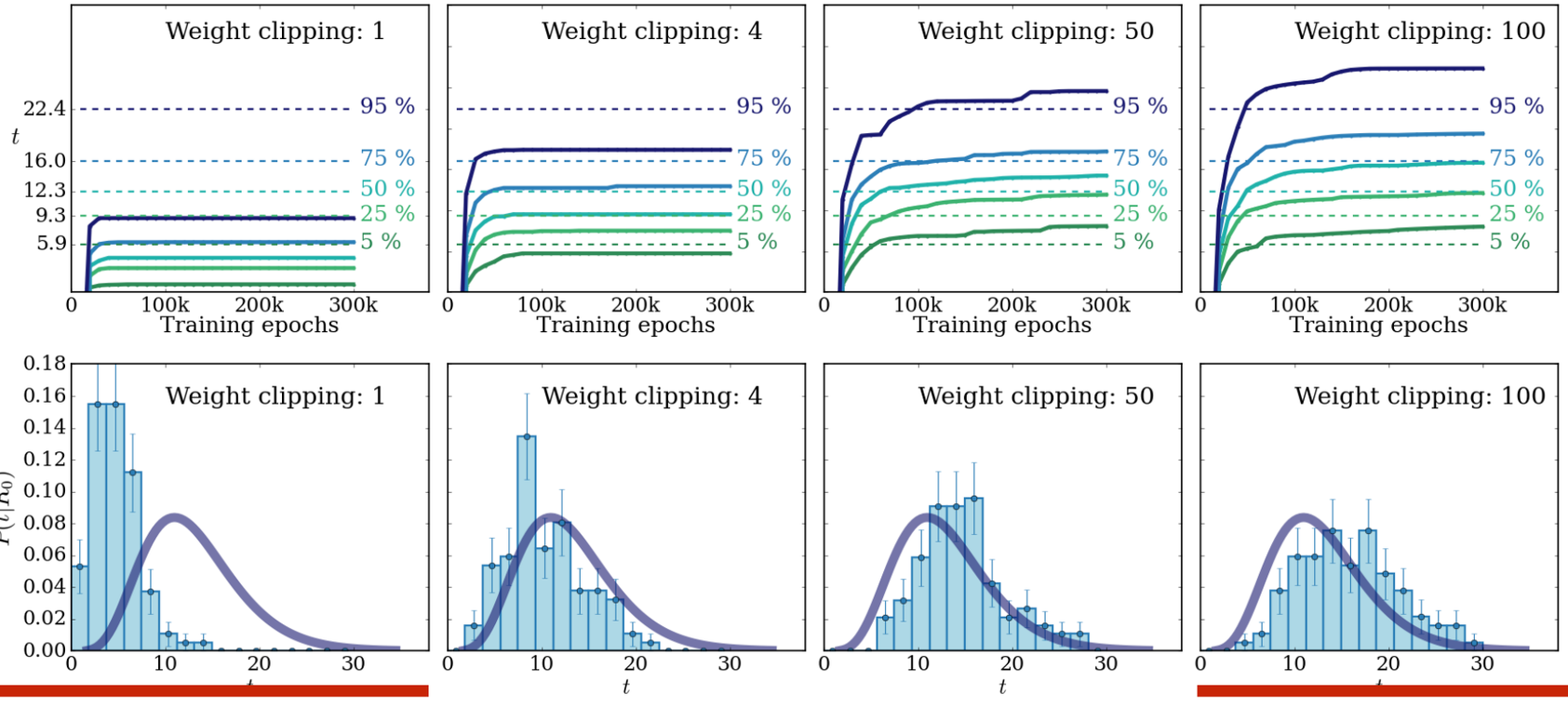
Criterion used in particular to select **Weight Clipping** regularisation par.

Weight Clipping Selection

(Simple 1d example with exponential Reference)







Asy.For. violation by fit parameters boundary

Asy.For. violation by sensitivity to sparse data points

An **Imperfect** Machine

Reference Model Predictions are unavoidably imperfect

e.g., PDF/Lumi/Detector Modeling …

Imperfections are **Nuisance Parameters**

Constrained by **Auxiliary Measurements**

Define a **composite** Reference hypothesis

An **Imperfect** Machine

Reference Model Predictions are unavoidably imperfect

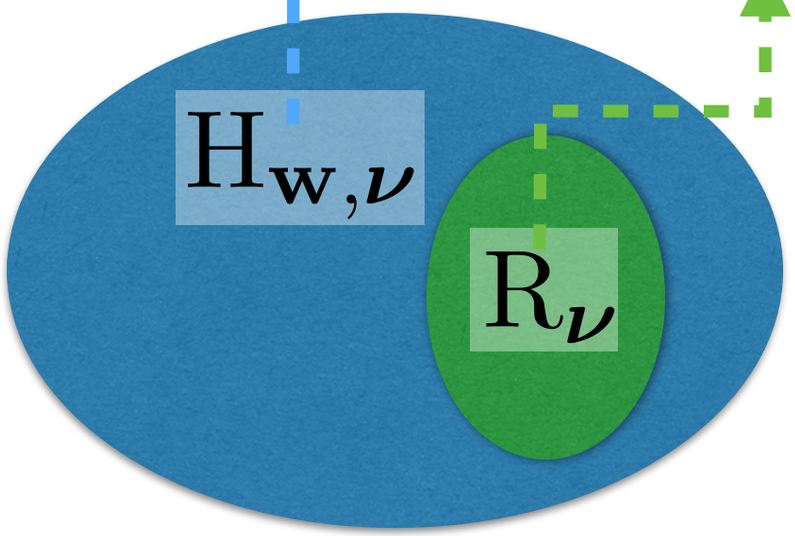
e.g., PDF/Lumi/Detector Modeling …

Imperfections are **Nuisance Parameters**

Constrained by **Auxiliary Measurements**

Define a **composite** Reference hypothesis

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



Just like in no-nuisance case:

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

Beyond-Reference effects parametrised by NN

An **Imperfect** Machine

Reference Model Predictions are unavoidably imperfect

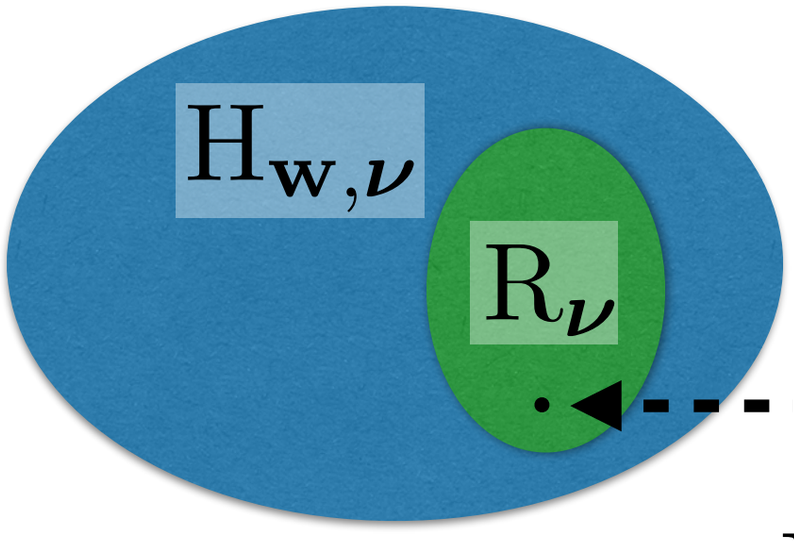
e.g., PDF/Lumi/Detector Modeling …

Imperfections are **Nuisance Parameters**

Constrained by **Auxiliary Measurements**

Define a **composite** Reference hypothesis

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



Central-Value Reference: $\mathrm{R}_{\mathbf{0}}$

Nuisance set to their C-V

$$t(\mathcal{D},\mathcal{A}) = \tau(\mathcal{D},\mathcal{A}) - \Delta(\mathcal{D},\mathcal{A})$$

An **Imperfect** Machine

Reference

Imperfect

$t(\mathcal{D}, \mathcal{A}) = 2\ \mathrm{m}$
$\mathbf{w}$

"Delta" term by direct likelihood maximisation

After **learning the effect of nuisance** locally on distribution

$$r(x;\boldsymbol{\nu}) \equiv \frac{n(x|\mathrm{R}_{\boldsymbol{\nu}})}{n(x|\mathrm{R}_{\mathbf{0}})} = \exp\left[\nu\,\delta_1(x) + \frac{1}{2}\nu^2\,\delta_2(x) + \ldots\right]$$

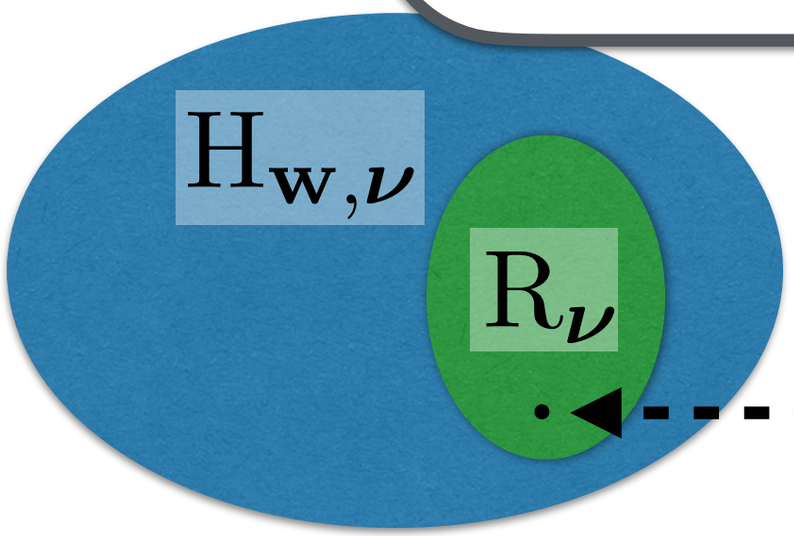
Adaptation of **likelihood-free inference** techniques

Would require dedicated seminar. [See e.g. 1907.10621, 2007.10356, ...]

Just be aware that:

i) learning requires (enough) R data with non-C-V nuisance

ii) the **quality** of the reconstruction can play crucial role



$$t(\mathcal{D}, \mathcal{A}) = \tau(\mathcal{D}, \mathcal{A}) - \Delta(\mathcal{D}, \mathcal{A})$$

Central-Value Reference: $\mathrm{R}_{\mathbf{0}}$
Nuisance set to their C-V

An **Imperfect** Machine

"Tau" term by training on Data

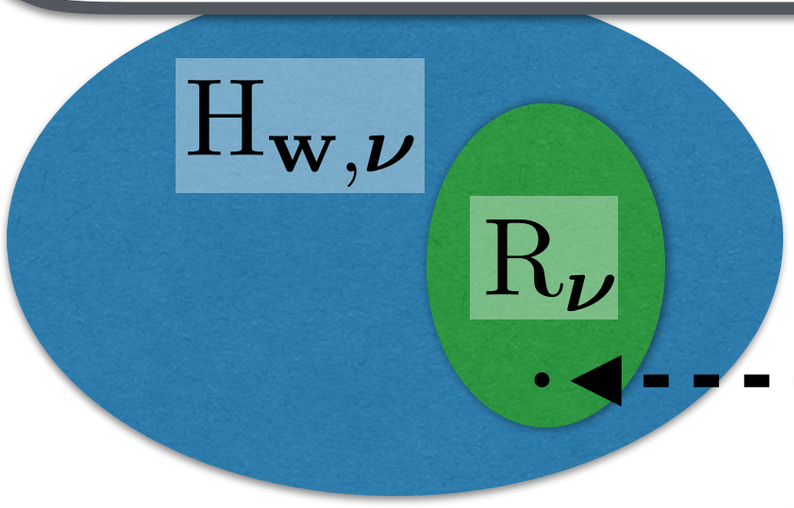
Almost like for no nuisance, but with modified ML-Loss:

$$L\left[f(\cdot;\mathbf{w}),\,\boldsymbol{\nu};\,\widehat{\delta}(\cdot)\right] = -\sum_{x_i\in\mathcal{D}}\left[f(x_i;\mathbf{w}) + \log(r(x_i;\boldsymbol{\nu}))\right] + \sum_{\mathrm{e}\in\mathcal{R}} w_{\mathrm{e}}\left[e^{f(x_{\mathrm{e}};\mathbf{w})+\log(r(x_{\mathrm{e}};\boldsymbol{\nu}))} - 1\right]$$

$$+\log\left[\frac{\mathcal{L}(\boldsymbol{\nu}|\mathcal{A})}{\mathcal{L}(\mathbf{0}|\mathcal{A})}\right]$$

t

And, with simultaneous **training over the nuisance** parameters
Data trained against **Central-Value Reference** sample **only**



$$t(\mathcal{D},\mathcal{A}) = \tau(\mathcal{D},\mathcal{A}) - \Delta(\mathcal{D},\mathcal{A})$$

Central-Value Reference: $\mathrm{R}_{\mathbf{0}}$
Nuisance set to their C-V

An **Imperfect** Machine

Reference Model Predictions are unavoidably imperfect

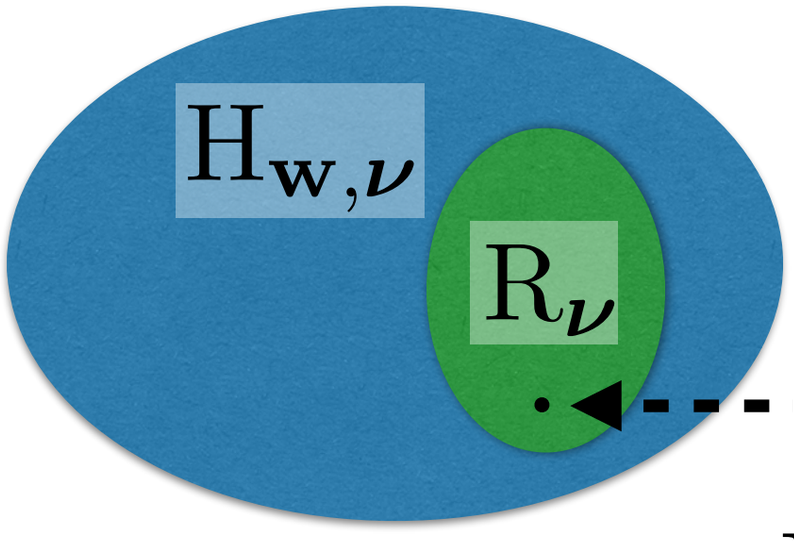
e.g., PDF/Lumi/Detector Modeling …

Imperfections are **Nuisance Parameters**

Constrained by **Auxiliary Measurements**

Define a **composite** Reference hypothesis

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



Central-Value Reference: $\mathrm{R}_{\mathbf{0}}$

Nuisance set to their C-V

$$t(\mathcal{D},\mathcal{A}) = \tau(\mathcal{D},\mathcal{A}) - \Delta(\mathcal{D},\mathcal{A})$$

If we do all right, by Wilks-Wald we get:

(**without weight clipping re-optimisation**)

$$P(t|\mathrm{R}_{\boldsymbol{\nu}}) = P(t|\mathrm{R}_{\mathbf{0}}) = \chi^2_d$$

An **Imperfect** Machine

Reference Model Predictions are unavoidably imperfect

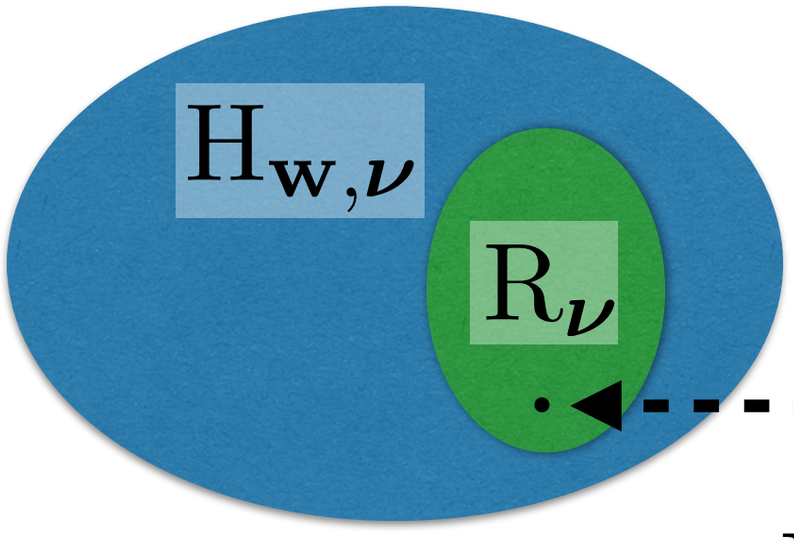
e.g., PDF/Lumi/Detector Modeling …

Imperfections are **Nuisance Parameters**

Constrained by **Auxiliary Measurements**

Define a **composite** Reference hypothesis

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



Central-Value Reference: Nuisance set to their C-V

$$t(\mathcal{D},\mathcal{A}) = \tau(\mathcal{D},\mathcal{A}) - \Delta(\mathcal{D},\mathcal{A})$$

If we do all right, by Wilks-Wald we get:

(**without weight clipping re-optimisation**)

$$P(t|\mathrm{R}_{\boldsymbol{\nu}}) = P(t|\mathrm{R}_{\mathbf{0}}) = \chi^2_d$$

Independence of t distribution on the **true value** of **nuisance** is **essential** for frequentist test

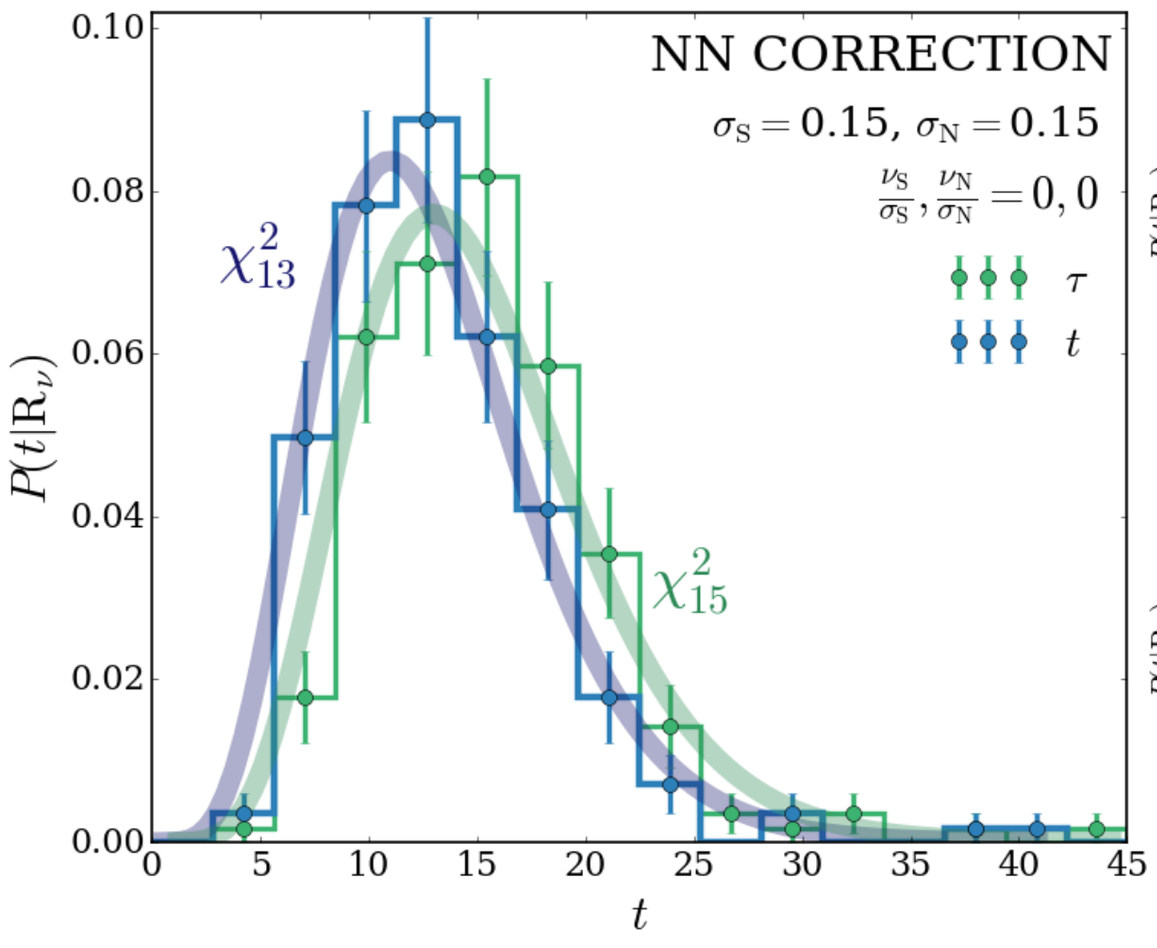
An **Imperfect** Machine

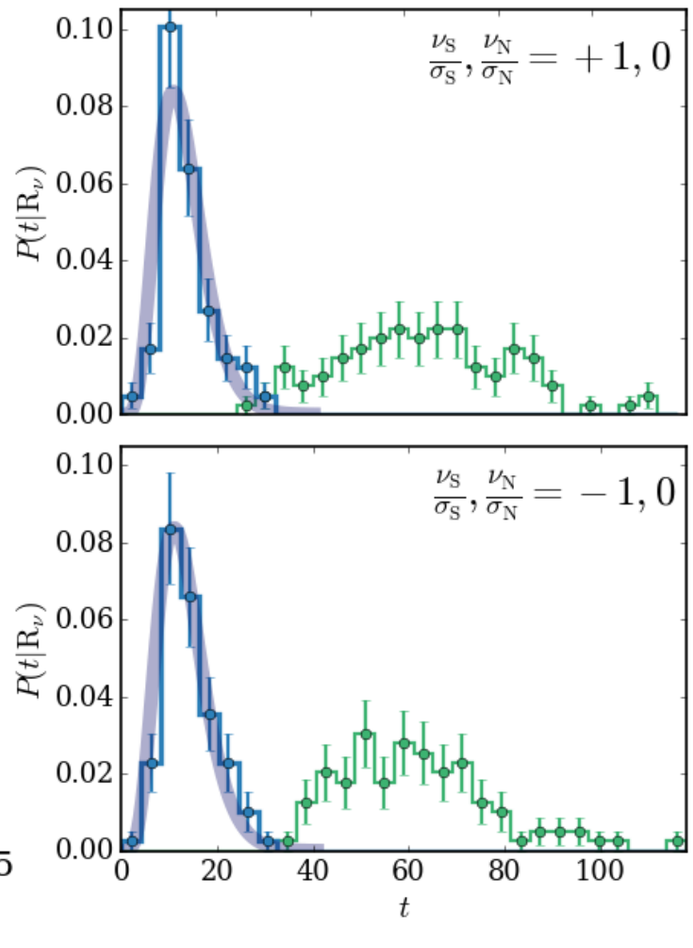
(Simple 1d example with exponential Reference)

Tau distribution distorted by non-central value nuisance

if not corrected, produces false positives

t = Tau-Delta independent of nuisance



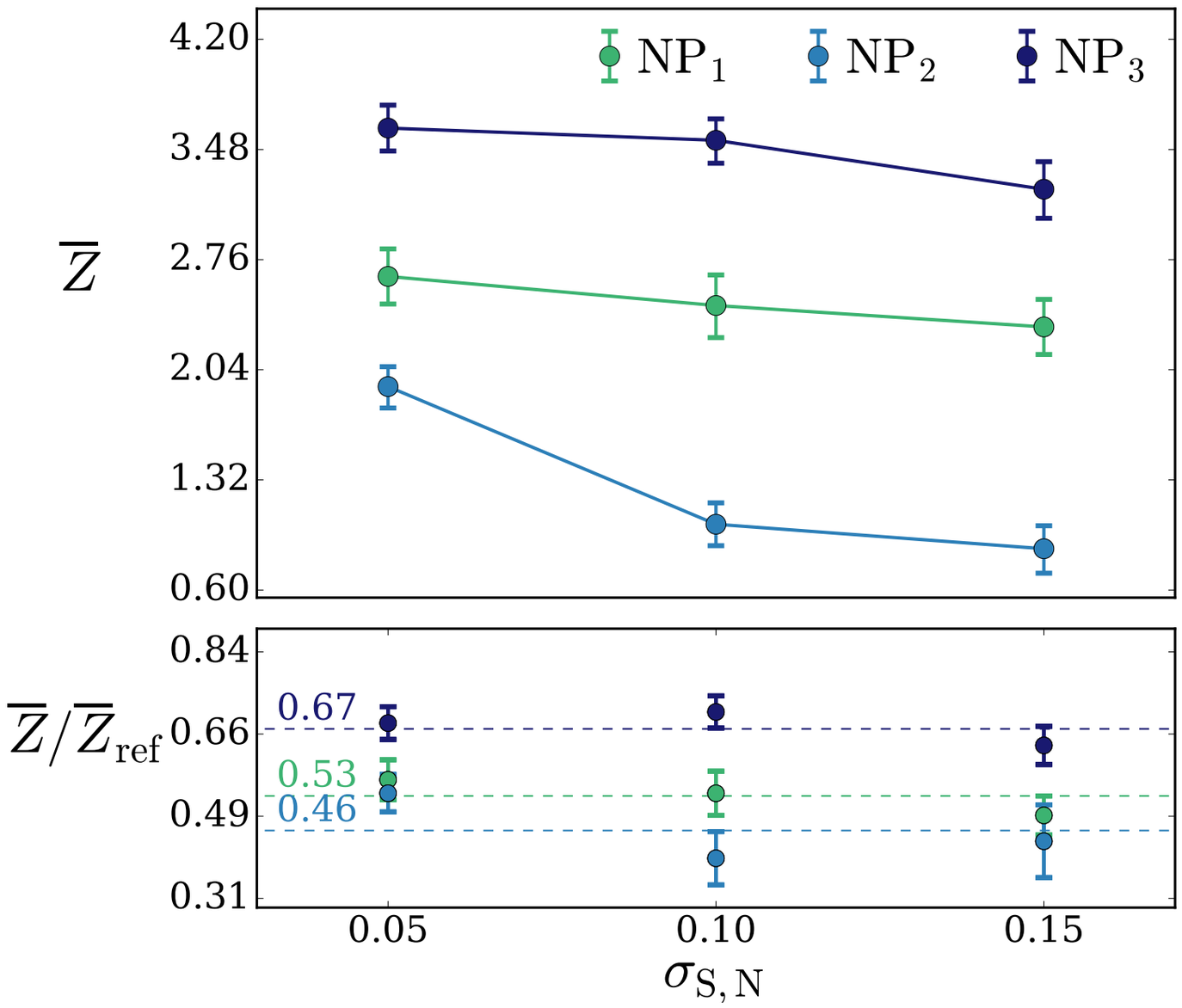


An **Imperfect** Machine

(Simple 1d example with exponential Reference)

Sensitivity to NP lowers as much as the reference one

→ Strategy is **effective**, not only **robust** to nuisance



A “~ Realistic” Problem

An emulation of two-particle final state at the LHC

- Reasonably high feature space dimensionality (= 5)

A “~ Realistic” Problem

An emulation of two-particle final state at the LHC

- Reasonably high feature space dimensionality (= 5)
- Systematic uncertainties (on scale and normalisation) at **realistic** level

(for **muon**, for **electrons**, and for **tau leptons**)

A “~ Realistic” Problem

An emulation of two-particle final state at the LHC

- Reasonably high feature space dimensionality (= 5)
- Systematic uncertainties (on scale and normalisation) at **realistic** level
  (for **muon**, for **electrons**, and for **tau leptons**)
- **Conservative limit** on **number of available Reference data points**
  Because detector sim. are demanding, or because stat. is limited for data-driven backgrounds

A “~ Realistic” Problem

An emulation of two-particle final state at the LHC

- Reasonably high feature space dimensionality (= 5)
- Systematic uncertainties (on scale and normalisation) at **realistic** level
  (for **muon**, for **electrons**, and for **tau leptons**)
- **Conservative limit** on **number of available Reference data points**
  Because detector sim. are demanding, or because stat. is limited for data-driven backgrounds

  → Reference Sample cannot be taken much larger that Data
  enhanced sensitivity to weight clipping

  → Learning nuisance effect on distribution more challenging

A “~ Realistic” Problem

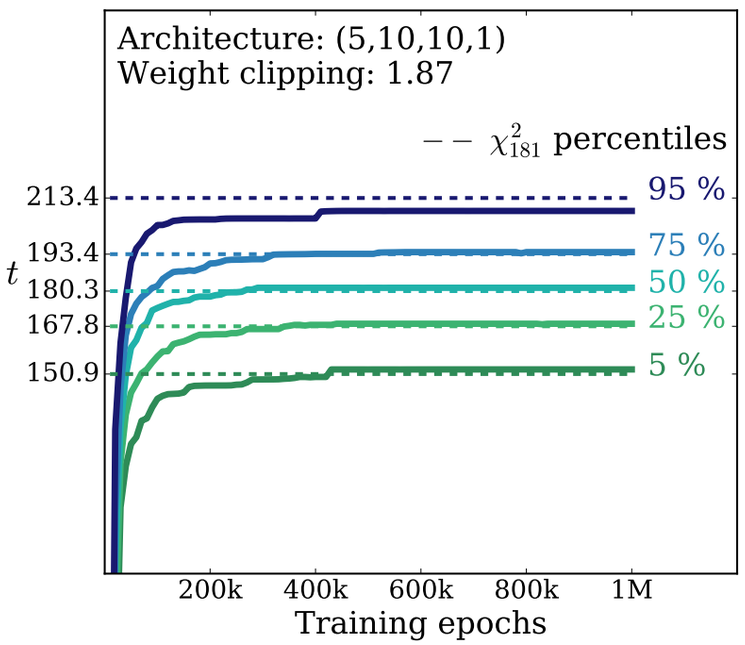
An emulation of two-particle final state at the LHC

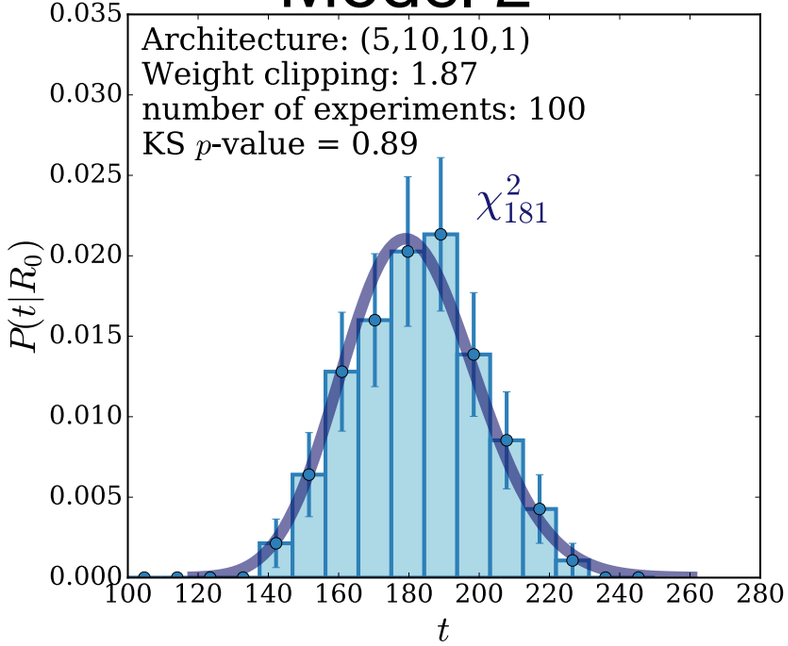
- Reasonably high feature space dimensionality (= 5)
- Systematic uncertainties (on scale and normalisation) at **realistic** level
  (for **muon**, for **electrons**, and for **tau leptons**)
- **Conservative limit** on **number of available Reference data points**
  Because detector sim. are demanding, or because stat. is limited for data-driven backgrounds

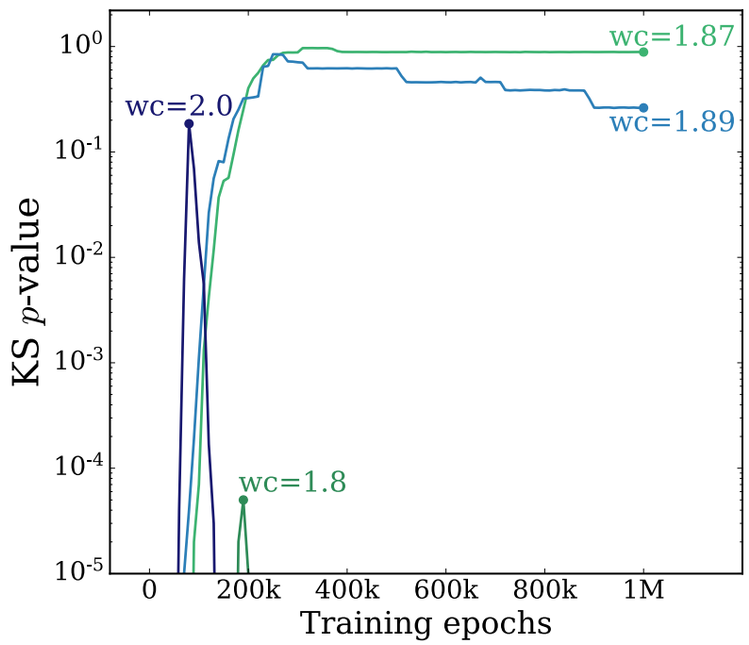
Results:

- Model selection gives 3 “maximally complex” viable architectures

Model 2







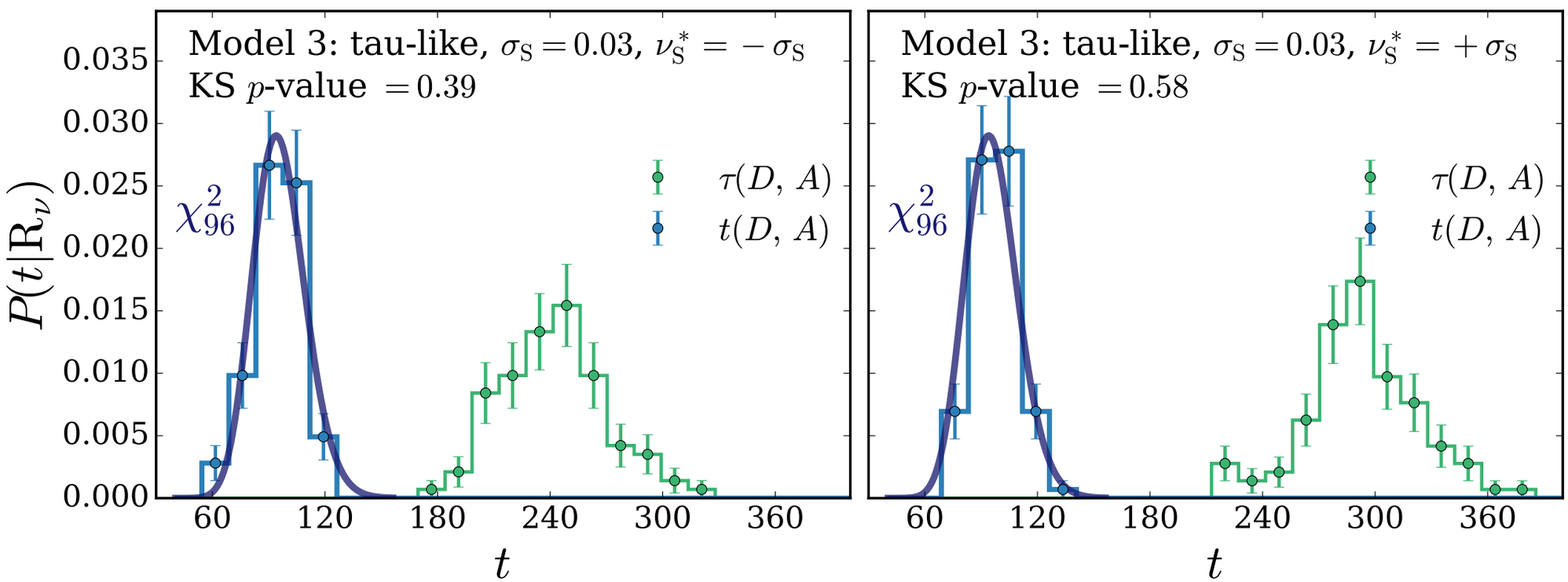
A “~ Realistic” Problem

An emulation of two-particle final state at the LHC

- Reasonably high feature space dimensionality (= 5)
- Systematic uncertainties (on scale and normalisation) at **realistic** level
  (for **muon**, for **electrons**, and for **tau leptons**)
- **Conservative limit** on **number of available Reference data points**
  Because detector sim. are demanding, or because stat. is limited for data-driven backgrounds

Results:

- Model selection gives 3 “maximally complex” viable architectures
- All successfully validated, for all uncertainties scenarios



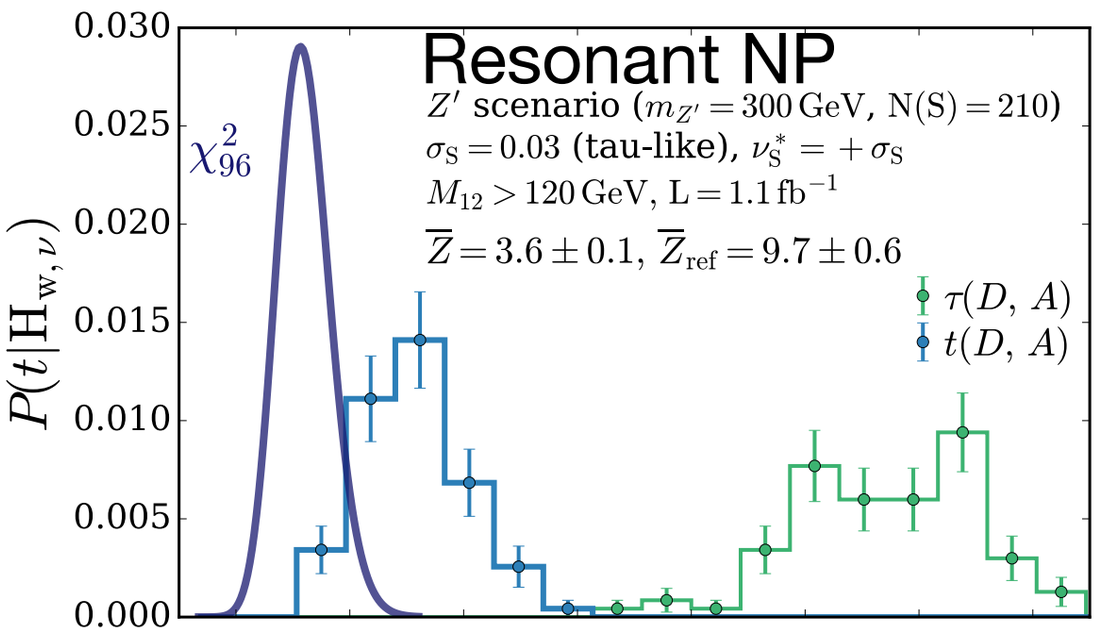
A “~ Realistic” Problem

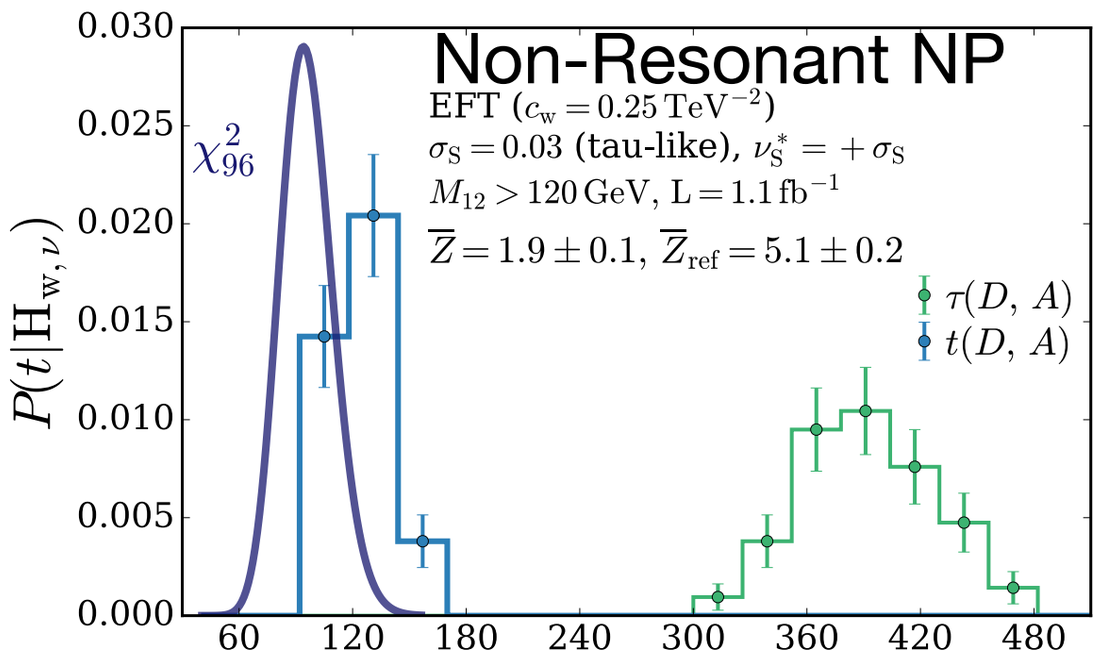
An emulation of two-particle final state at the LHC

- Reasonably high feature space dimensionality (= 5)
- Systematic uncertainties (on scale and normalisation) at **realistic** level
  (for **muon**, for **electrons**, and for **tau leptons**)
- **Conservative limit** on **number of available Reference data points**
  Because detector sim. are demanding, or because stat. is limited for data-driven backgrounds

Results:

- Model selection gives 3 “maximally complex” viable architectures
- All successfully validated, for all uncertainties scenarios
- Sensitivity to NP lowers “just as much” as it should





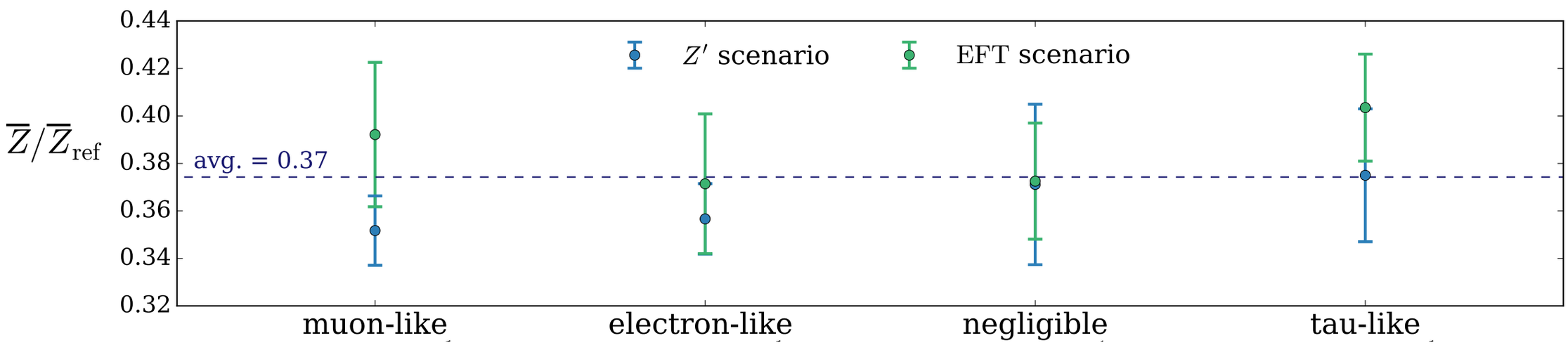
A “~ Realistic” Problem

An emulation of two-particle final state at the LHC

- Reasonably high feature space dimensionality (= 5)
- Systematic uncertainties (on scale and normalisation) at **realistic** level
  (for **muon**, for **electrons**, and for **tau leptons**)
- **Conservative limit** on **number of available Reference data points**
  Because detector sim. are demanding, or because stat. is limited for data-driven backgrounds

Results:

- Model selection gives 3 “maximally complex” viable architectures
- All successfully validated, for all uncertainties scenarios
- Sensitivity to NP lowers “just as much” as it should



Outlook

Strategy is fully defined, conceptually and methodologically

Further progress requires **full-fledged implementation** in **realistic LHC final state** (2 leptons?, 4 leptons?, more exotic?)

Outlook

Strategy is fully defined, conceptually and methodologically

Further progress requires **full-fledged implementation** in **realistic LHC final state** (2 leptons?, 4 leptons?, more exotic?)

Expected implementation challenges (limit on **lumi.** we can handle)

- Statistically accurate enough (large or smart) Reference Sample
- Generation of Reference-distributed Toys
- Accurate learning of nuisance Likelihood
- Training execution time

Outlook

Strategy is fully defined, conceptually and methodologically

Further progress requires **full-fledged implementation** in **realistic LHC final state** (2 leptons?, 4 leptons?, more exotic?)

Expected implementation challenges (limit on **lumi.** we can handle)

- Statistically accurate enough (large or smart) Reference Sample
- Generation of Reference-distributed Toys
- Accurate learning of nuisance Likelihood
- Training execution time

Faster/Smarter Monte Carlo
weighted samples
generative models
fast (but accurate) detector sim.
Toys at NLO

Generic need for the whole **HL-LHC** analysis program!

Outlook

Strategy is fully defined, conceptually and methodologically

Further progress requires **full-fledged implementation** in **realistic LHC final state** (2 leptons?, 4 leptons?, more exotic?)

Expected implementation challenges (limit on **lumi.** we can handle)

- Statistically accurate enough (large or smart) Reference Sample
- Generation of Reference-distributed Toys
- Accurate learning of nuisance Likelihood
- Training execution time

Faster/Smarter Monte Carlo
- weighted samples
- generative models
- fast (but accurate) detector sim.
- Toys at NLO

Generic need for the whole **HL-LHC** analysis program!

Likelihood-free Inference Techniques
- being worked out for EFT (MadMiner)

Stimulate and **exploit** these developments

Outlook

Strategy is fully defined, conceptually and methodologically

Further progress requires **full-fledged implementation** in **realistic LHC final state** (2 leptons?, 4 leptons?, more exotic?)

Expected implementation challenges (limit on **lumi.** we can handle)

- Statistically accurate enough (large or smart) Reference Sample
- Generation of Reference-distributed Toys
- Accurate learning of nuisance Likelihood
- Training execution time

Non-NN Models
Kernel Method “Falkon”
[Letizia, …, Rosasco, … to appear]

Faster/Smarter Monte Carlo
weighted samples
generative models
fast (but accurate) detector sim.
Toys at NLO
Generic need for the whole **HL-LHC** analysis program!

Likelihood-free Inference Techniques
being worked out for EFT (MadMiner)
Stimulate and **exploit** these developments

Outlook

Model-Independent search algorithms also good for:

- Comparison between Monte Carlo Generators
- Data Validation

Outlook

Model-Independent search algorithms also good for:

- Comparison between 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 …

Outlook

Model-Independent search algorithms also good for:

- Comparison between 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 !!

Outlook

Model-Independent search algorithms also good for:

- Comparison between 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

Backup

Maximum Likelihood Loss

Turn the evaluation of "t" into supervised training problem:

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

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

We need a **Reference Sample**, distributed according to 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

Turn the evaluation of "t" into supervised training problem:

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

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

We need a **Reference Sample**, distributed according to Reference Model

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

Approximate integral as Monte Carlo sum:

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

In order to read this as "equal", we need

$$\mathcal{N}_\mathcal{R} \gg N(\mathrm{R})$$

Might not be trivial to get such large sample for a real LHC analysis

Maximum Likelihood Loss

Turn the evaluation of “t” into supervised training problem:

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

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

We need a **Reference Sample**, distributed according to 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. Trained net is fit to distribution log ratio**

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

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