NPLM: Learning New Physics aware of systematic uncertainties

R.T. d'Agnolo[1], ***G. Grosso***[2,3], *M. Pierini*[3], *A. Wulzer*[2], *M. Zanetti*[2]

[1] Université Paris-Saclay and CEA, [2] Università degli Studi di Padova and INFN, [3] CERN experimental department

May 11, 2022

In this talk

- How to include systematic uncertainties in NPLM
- NPLM application to HEP:
  5D analysis of a di-body final state at the LHC

"Learning New Physics from an Imperfect Machine" Eur. Phys. J. C

More about NPLM:

- "Learning New Physics from a Machine" Phys. Rev. D
- "Learning Multivariate New Physics" Eur. Phys. J. C
- Previous IML talk 27/04/21

NPLM algorithm

New Physics Learning Machine (NPLM)

Including systematic uncertainties

- Goal: performing a **log-likelihood-ratio hypothesis test**
  (End-to-end strategy, from the data to a *p*-value for the discovery)

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

- Exploiting a Neural Network (NN) to **parametrize** the data distribution in terms of a Reference distribution ($\mathrm{R}_{\mathbf{0}}$)

$$\boxed{n(x\,|\,\mathrm{H}_{\hat{\mathbf{w}},\,\hat{\nu}})} = \boxed{n(x\,|\,\mathrm{R}_0)}\;\boxed{\frac{n(x\,|\,\mathrm{R}_{\hat{\nu}})}{n(x\,|\,\mathrm{R}_0)}}\; e^{\boxed{f(x;\,\hat{\mathbf{w}})}}$$

Data distribution learnt by the NN

Alternative hypothesis

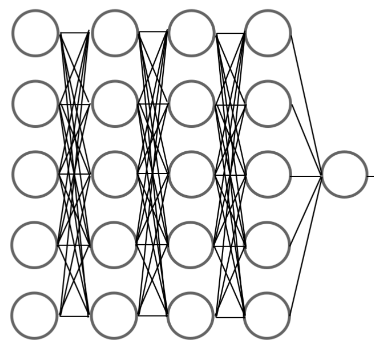
Reference distribution

Null hypothesis

New term containing the dependence on ν

$r(x;\nu)$

NN model



- **Signal-model-independent:** reduced assumptions on the signal hypothesis

$\mathrm{R}_{\boldsymbol{\nu}}$: reference (null) hypothesis
$\mathrm{H}_{\mathbf{w},\,\boldsymbol{\nu}}$: alternative hypothesis
$\mathbf{w}$: trainable parameters on the NN model
$\boldsymbol{\nu}$: set of nuisance parameters modelling the uncertainties effects
$\mathcal{D}$: data sample
$\mathcal{A}$: auxiliary sample (to constrain $\boldsymbol{\nu}$)

"Learning New Physics from an Imperfect Machine" Eur. Phys. J. C

New Physics Learning Machine (NPLM)

Including systematic uncertainties

Maximum Likelihood from minimal loss:

Test statistic

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

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

$\mathbf{w}$: trainable parameters on the NN model

$\boldsymbol{\nu}$: set of nuisance parameters modelling the uncertainties effects

$\mathcal{D}$: data sample

$\mathcal{A}$: auxiliary sample (used to constrain $\boldsymbol{\nu}$)

Tau term:

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

Contains the dependence on a NN model (universal approximator)

Delta term:

$$\Delta(\mathcal{D},\,\mathcal{A}) = 2\max_{\boldsymbol{\nu}}\log\left[\frac{\mathcal{L}(\mathrm{R}_{\boldsymbol{\nu}}|\mathcal{D})\,\mathcal{L}(\boldsymbol{\nu}|\mathcal{A})}{\mathcal{L}(\mathrm{R}_{\mathbf{0}}|\mathcal{D})\,\mathcal{L}(\mathbf{0}|\mathcal{A})}\right] = -2\min_{\boldsymbol{\nu}} L\left[\boldsymbol{\nu};\widehat{\delta}(x)\right]$$

Built on the knowledge of the Reference model (purely SM term)

$$r(x;\nu) = \frac{n(x\,|\,\mathrm{R}_\nu)}{n(x\,|\,\mathrm{R}_0)}$$

Taylor's expansion learning:

$$\widehat{r}(x;\nu) = \exp\left[\underbrace{\widehat{\delta}_1(x)}_{\text{NN 1}}\,\nu + \underbrace{\widehat{\delta}_2(x)}_{\text{NN2}}\,\nu^2 + \underbrace{\ldots}_{\ldots}\right]$$

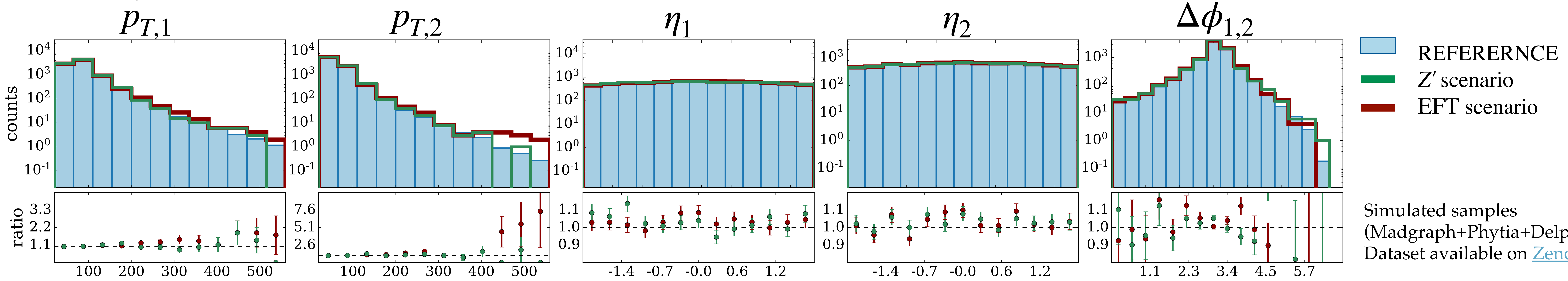
"Learning New Physics from an Imperfect Machine" Eur. Phys. J. C

Application:
5D analysis of a di-body final state at the LHC

Di-body final state at the LHC

Dataset

5D analysis — Input variables:



Simulated samples
(Madgraph+Phytia+Delphes)
Dataset available on [Zenodo]

Uncertainties on the reference sample (SM):

- Global normalization effect: $\sigma_{\rm N} = 2.5\,\%$
- Momentum scale effect:

$$p_{T\,1,2}^{\rm (b,e)} = \exp\left[\nu_{\rm s}\sigma_{\rm s}^{\rm (b,e)}/\sigma_{\rm s}^{\rm (b)}\right] p_{T\,1,2}^{\rm (b,e)}$$

(b) barrel region $|\eta| < 1.2$, (e) endcaps region $|\eta| \geq 1.2$

- Muon-like regime: $\sigma_{\rm S}^{\rm (b)} = 0.05\,\%$, $\sigma_{\rm S}^{\rm (e)} = 0.15\,\%$
- Electron-like regime: $\sigma_{\rm S}^{\rm (b)} = 0.3\,\%$, $\sigma_{\rm S}^{\rm (e)} = 0.9\,\%$
- Tau-like regime: $\sigma_{\rm S}^{\rm (b)} = \sigma_{\rm S}^{\rm (e)} = 3\,\%$

Di-body final state at the LHC

Dataset

New Physics benchmarks:

Resonance in the two-body invariant mass

- *Z′* **scenario:** new vector boson with the same SM coupling as the Z boson and mass of 300 GeV.
  - Muon-like, electron-like regimes:
    $M_{12} > 100\,\mathrm{GeV}, L = 0.35\,\mathrm{fb}^{-1}, \mathrm{N(S)} = 120$
  - Tau-like regime:
    $M_{12} > 120\,\mathrm{GeV}, L = 1.1\,\mathrm{fb}^{-1}, \mathrm{N(S)} = 210$

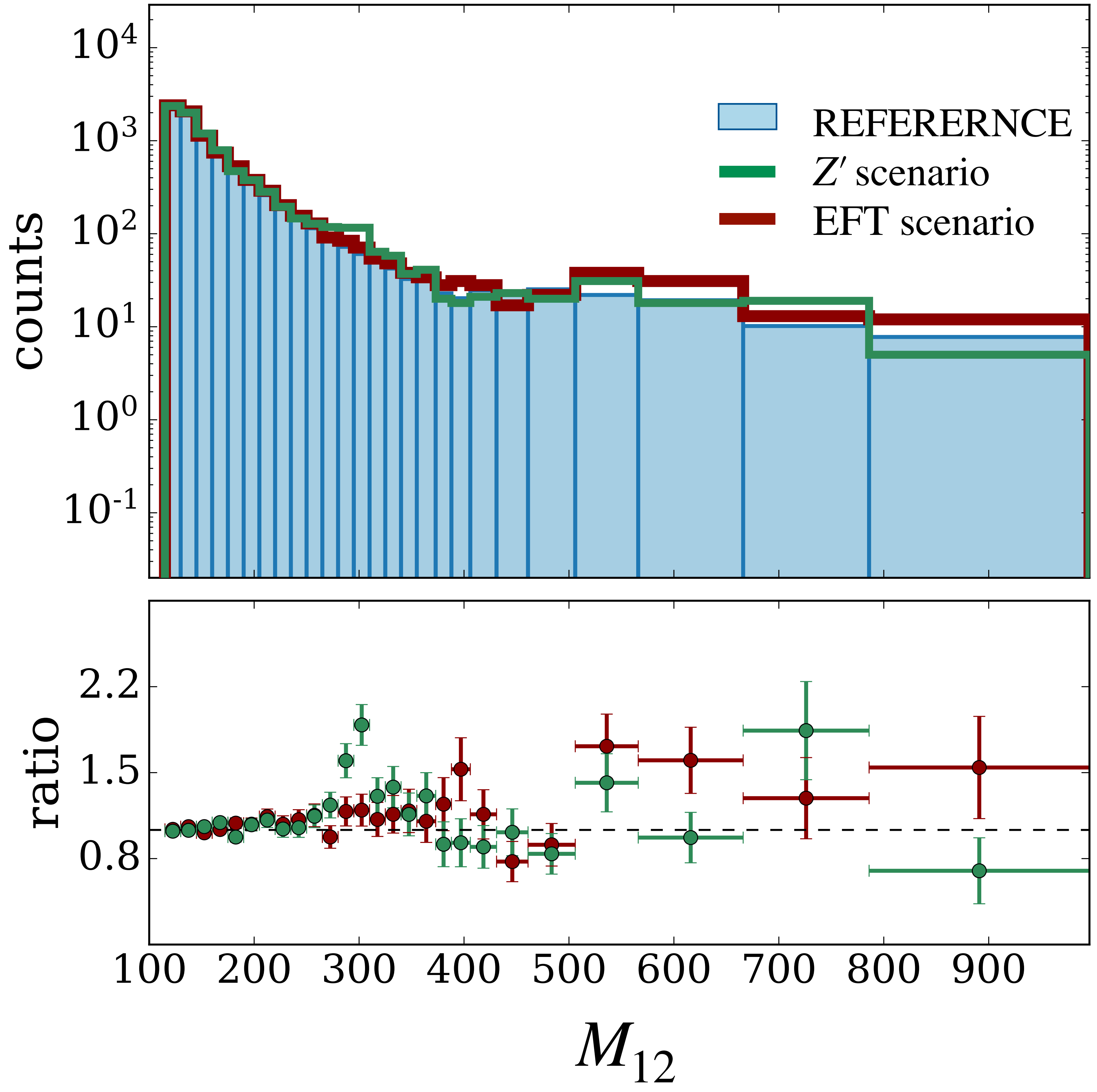
Non resonant excess in the tail of the two-body invariant mass

- EFT **scenario:** dimension-6 4-fermions contact operator:

$$\frac{c_W}{\Lambda} J^a_{L\mu} J^\mu_{La}.$$

  - Muon-like, electron-like regimes:
    $M_{12} > 100\,\mathrm{GeV}, L = 0.35\,\mathrm{fb}^{-1}, c_W = 1.0\,\mathrm{TeV}^{-2}$
  - Tau-like regime:
    $M_{12} > 120\,\mathrm{GeV}, L = 1.1\,\mathrm{fb}^{-1}, c_W = 0.25\,\mathrm{TeV}^{-2}$

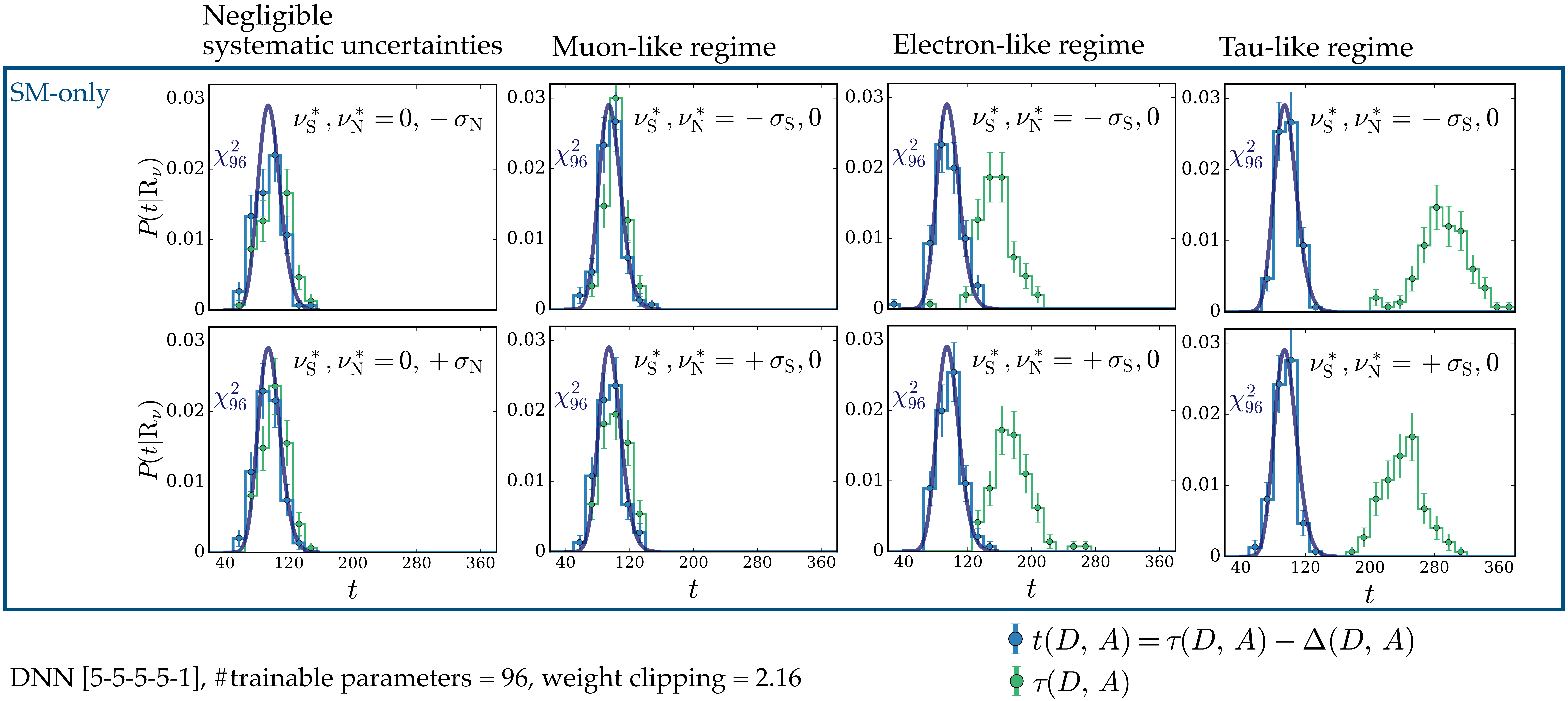
Example:
Tau-like regime



NOTE:
M_{12} is **not** given as an input to the algorithm!

Di-body final state at the LHC

$\tau - \Delta$ validation

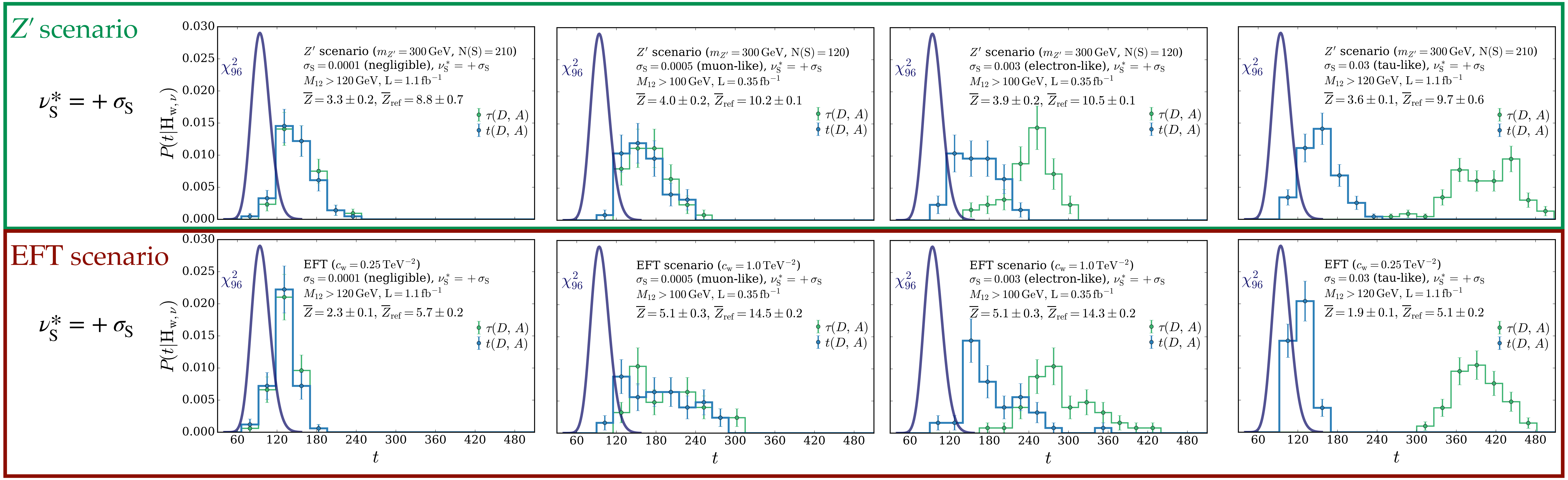


DNN [5-5-5-5-1], #trainable parameters = 96, weight clipping = 2.16

Di-body final state at the LHC

Sensitivity to New Physics scenarios

Negligible systematic uncertainties	Muon-like regime	Electron-like regime	Tau-like regime



Z-score: $Z = \Phi^{-1}\left[1 - p\right]$

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

$\tau(D,\, A)$

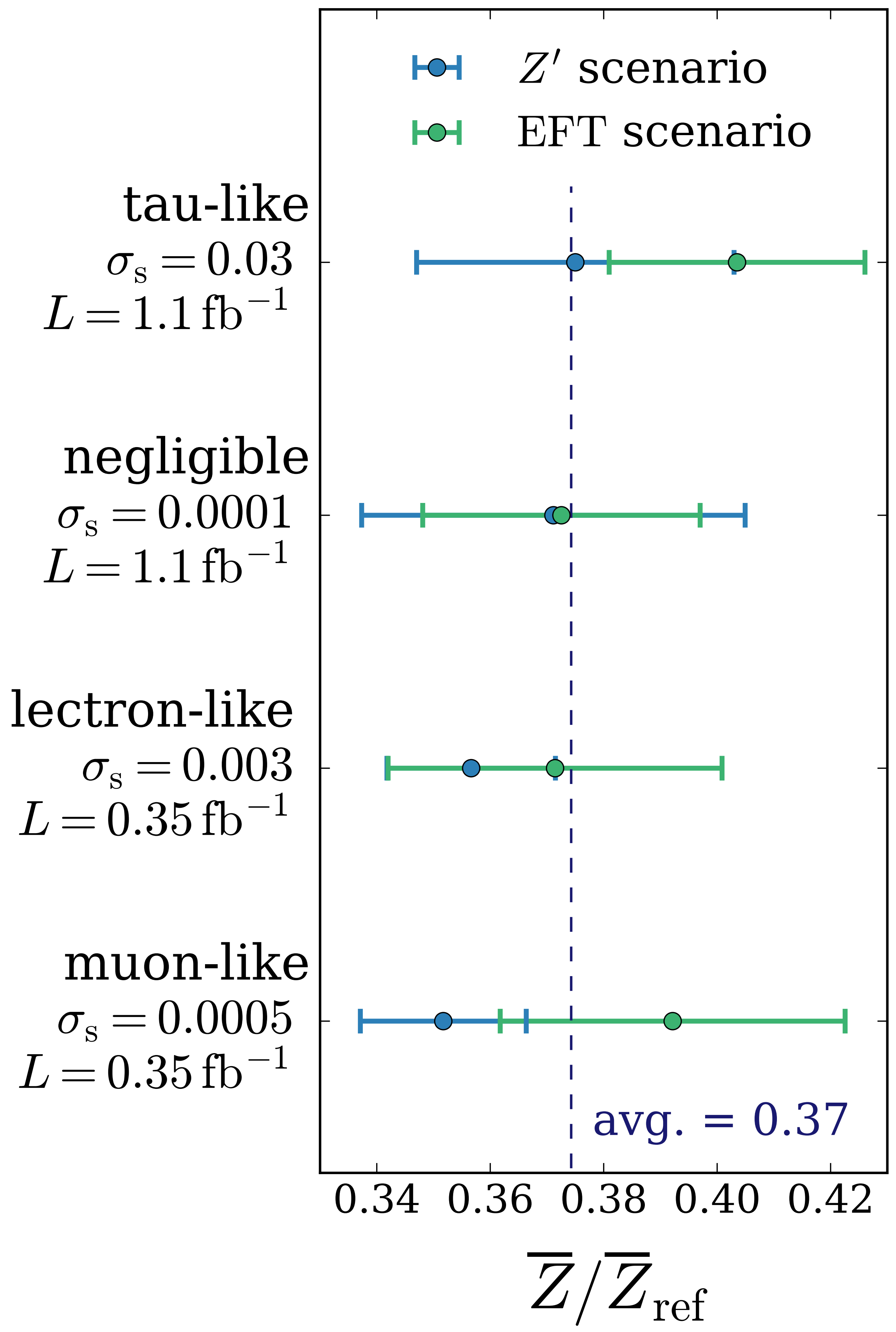
DNN [5-5-5-5-1], #trainable parameters = 96, weight clipping = 2.16

Di-body final state at the LHC

Sensitivity to New Physics scenarios

Summary of the results:

- Comparable performances in the resonant and non-resonant scenarios:
  - NPLM is simultaneously sensitive to multiple sources of New Physics;
- Comparable performances at different systematic uncertainties regimes:
  - NPLM is robust against the presence of systematic uncertainties;
  - the presence of systematic uncertainties affects NPLM in the same measure as any other hypothesis test;
- No information about the New Physics signal has been provided to the algorithm at any step of its implementation:
  - The performances of NPLM are lower than any model-dependent strategy by construction ($\overline{Z}/\overline{Z}_{\rm ref} = 0.37$);



Z-score: $Z = \Phi^{-1}\left[1-p\right]$

- $\overline{Z}$: Z-score from NPLM
- $\overline{Z}_{\rm ref}$: Z-score from a model-dependent (optimized) test statistics

Conclusions

Outlook on future perspectives

Current limitations and future developments:

- Accuracy and size of the Reference sample
- Accuracy in the (multivariate) modelling of the nuisance effects
- Training time
  A solution from Kernel Methods ("Learning new physics efficiently with nonparametric methods" 2204.02317)
- Optimisation of NPLM sensitivity performances:
  how do we choose the NN architecture? Is the regularization heuristic optimal? (ongoing work)

Set a limit on the actual **luminosity** that we are allowed to inspect, but do not obstacle the applicability of NPLM.

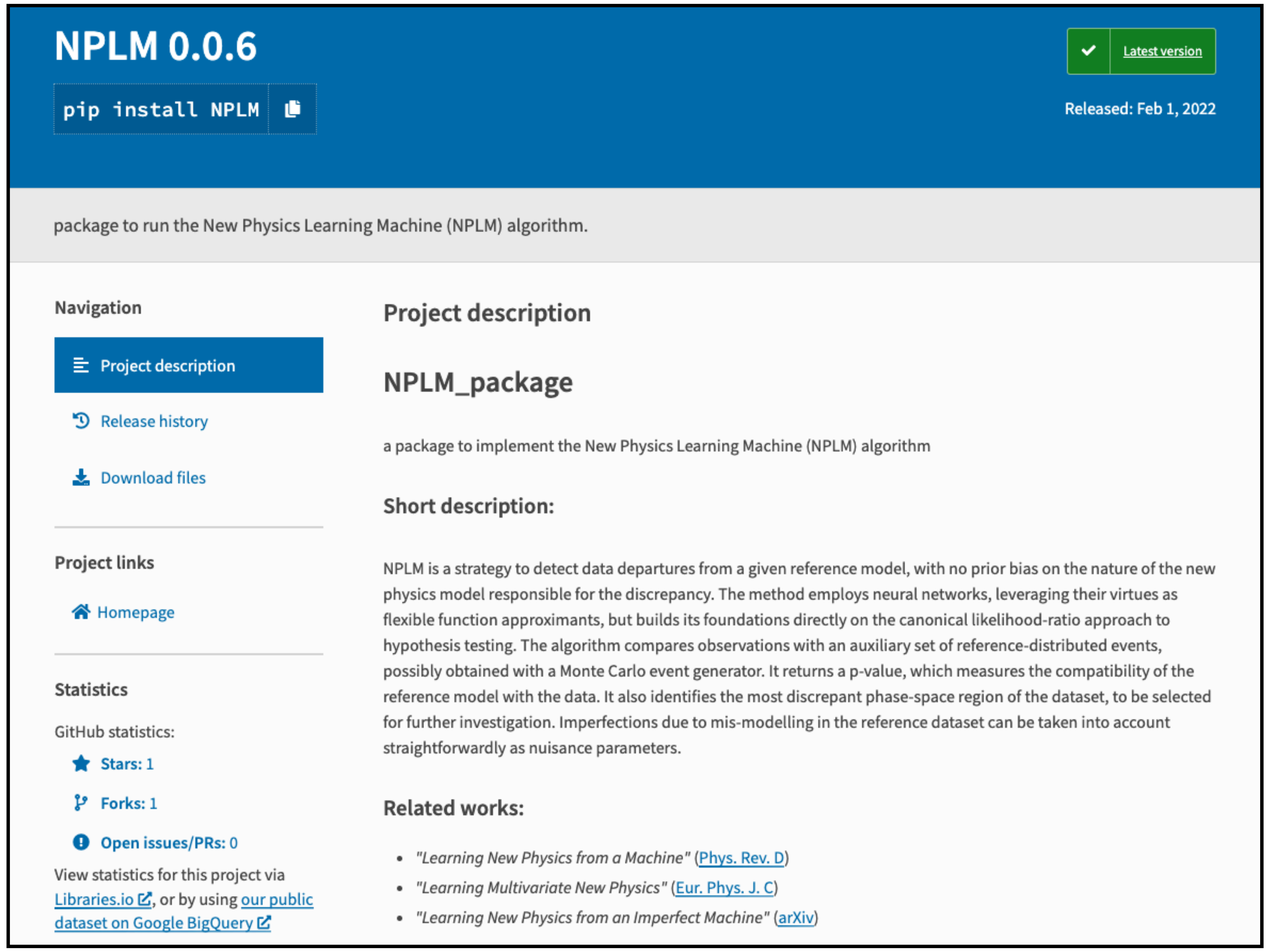
NPLM is ready to be performed on a real analysis at the LHC!

✓ Heuristic method to setup **multivariate** analysis
✓ Strategy to account for **systematic uncertainties**

Outlook on future perspectives

Getting started with NPLM

- NPLM package: python-based package to run the NPLM analysis strategy
- Tutorial on 1D toy model for getting started



NPLM 0.0.6

pip install NPLM

Latest version

Released: Feb 1, 2022

package to run the New Physics Learning Machine (NPLM) algorithm.

Navigation

Project description

Release history

Download files

Project links

Homepage

Statistics

GitHub statistics:

Stars: 1

Forks: 1

Open issues/PRs: 0

View statistics for this project via Libraries.io, or by using our public dataset on Google BigQuery

Project description

NPLM_package

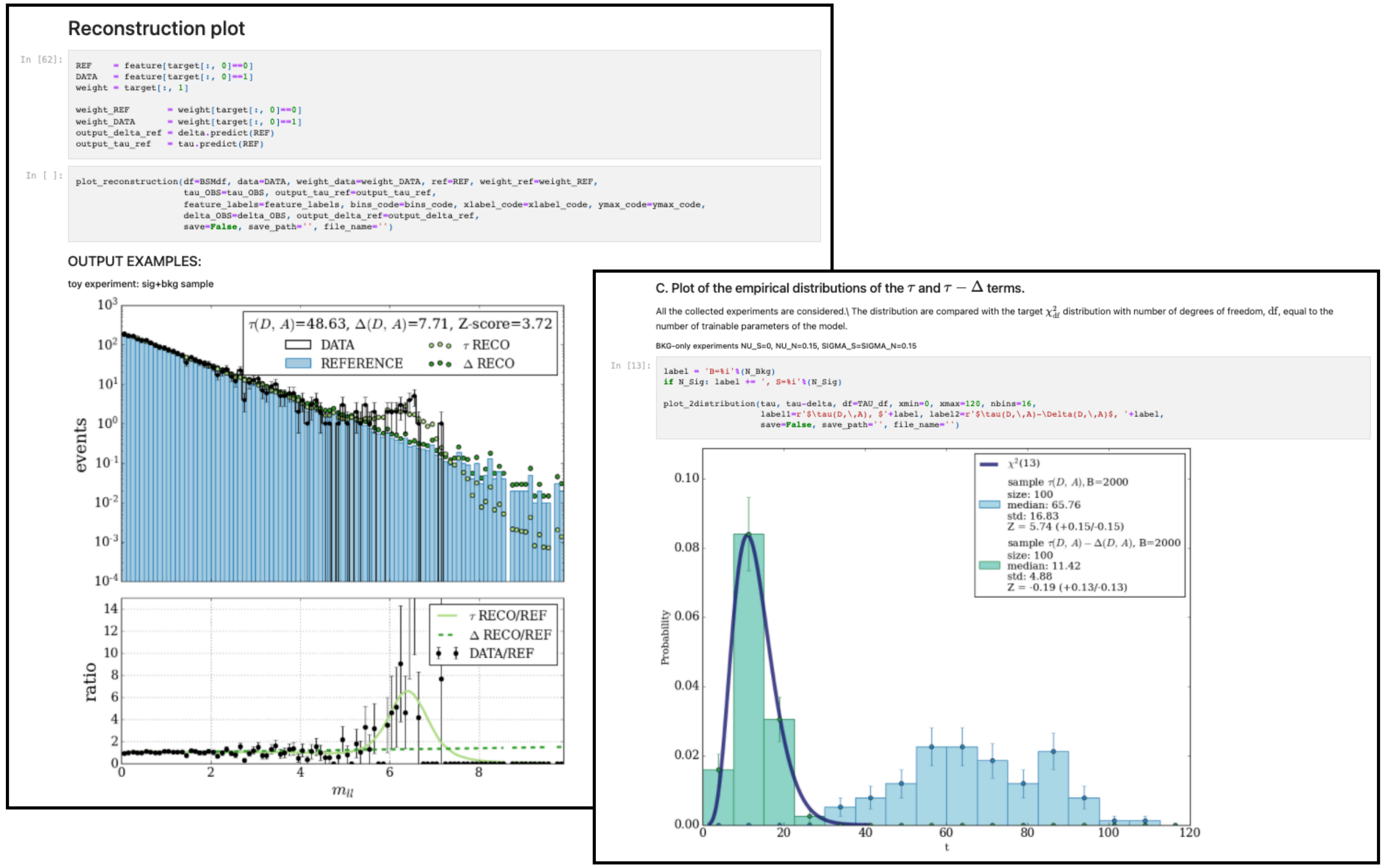
a package to implement the New Physics Learning Machine (NPLM) algorithm

Short description:

NPLM is a strategy to detect data departures from a given reference model, with no prior bias on the nature of the new physics model responsible for the discrepancy. The method employs neural networks, leveraging their virtues as flexible function approximants, but builds its foundations directly on the canonical likelihood-ratio approach to hypothesis testing. The algorithm compares observations with an auxiliary set of reference-distributed events, possibly obtained with a Monte Carlo event generator. It returns a p-value, which measures the compatibility of the reference model with the data. It also identifies the most discrepant phase-space region of the dataset, to be selected for further investigation. Imperfections due to mis-modelling in the reference dataset can be taken into account straightforwardly as nuisance parameters.

Related works:

- "Learning New Physics from a Machine" (Phys. Rev. D)
- "Learning Multivariate New Physics" (Eur. Phys. J. C)
- "Learning New Physics from an Imperfect Machine" (arXiv)



Backup

New Physics Learning Machine (NPLM)

Main Concepts (negligible uncertainties)

- Goal: performing a **log-likelihood-ratio hypothesis test**
  (End-to-end strategy, from the data to a *p*-value for the discovery)

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

$\mathrm{R}_{\mathbf{0}}$: reference (null) hypothesis
$\mathrm{H}_{\mathbf{w}}$: alternative hypothesis

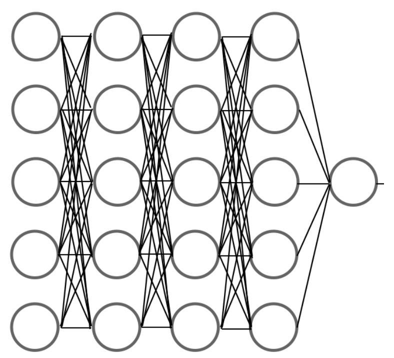
- Exploiting a Neural Network (NN) to **parametrize** the data distribution in terms of a Reference distribution ($\mathrm{R}_{\mathbf{0}}$)

$$\boxed{n(x\,|\,\mathrm{T})} \approx \boxed{n(x\,|\,\mathrm{H}_{\hat{\mathbf{w}}})} = \boxed{n(x\,|\,\mathrm{R}_0)}\; e^{\boxed{f(x,\hat{\mathbf{w}})}}$$

True (T) data distribution — Unknown

Data distribution learnt by the NN — Alternative hypothesis

Reference distribution — Null hypothesis (SM)



NN model

- **Signal-model-independent**: reduced assumptions on the signal hypothesis

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

New Physics Learning Machine (NPLM)

Main Concepts (negligible uncertainties)

Maximum Likelihood from minimal loss:

Test statistic

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

Loss function

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

$\mathbf{w}$: trainable parameters on the NN model

D: data sample

R: reference sample (built according to the R_0 hypothesis); could be weighted (w_x)

Assumptions:

- $N_R \gg N_D$ the statistical fluctuations of the reference sample are negligible.
- the weights of the reference sample (w) are such that the reference sample is normalised to match the data sample luminosity $\sum_{x\in R} w_x = N(\mathrm{R}_0)$

“Learning New Physics from a Machine” Phys. Rev. D

New Physics Learning Machine (NPLM)

Main Concepts (negligible uncertainties)

Maximum Likelihood from minimal loss:

Test statistic

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

Loss function

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

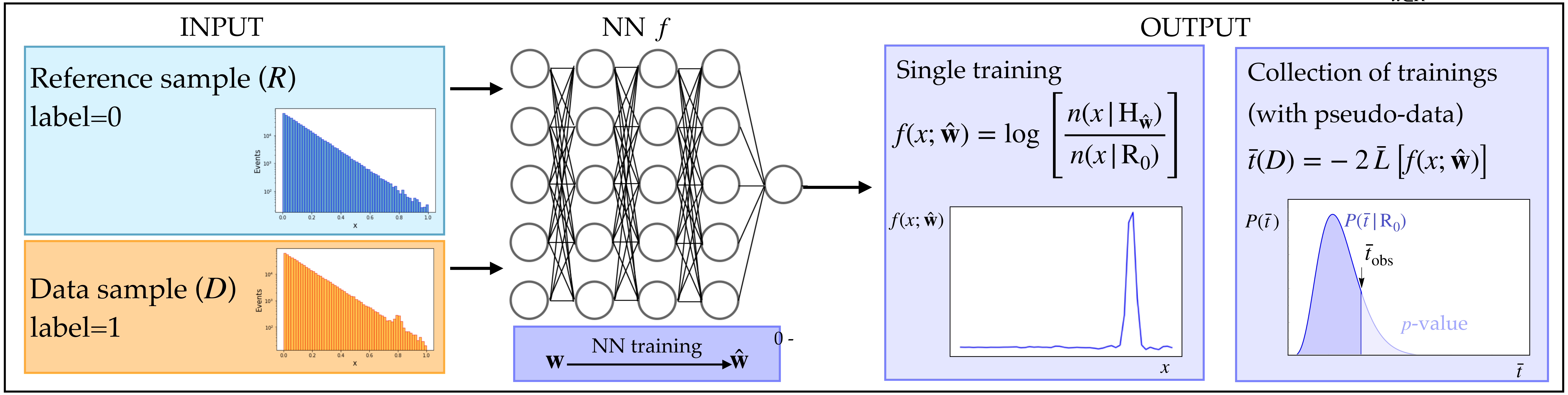
$\mathbf{w}$: trainable parameters on the NN model

D: data sample

R: reference sample (built according to the R_0 hypothesis); could be weighted (w_x)

Assumptions:

- $N_R \gg N_D$ the statistical fluctuations of the reference sample are negligible.
- the weights of the reference sample (w) are such that the reference sample is normalised to match the data sample luminosity $\sum_{x\in R} w_x = N(\mathrm{R}_0)$



“Learning New Physics from a Machine” Phys. Rev. D

New Physics Learning Machine (NPLM)

Including systematic uncertainties

New Physics Learning Machine (NPLM)

Including systematic uncertainties

Validation of the $(\tau - \Delta)$ procedure

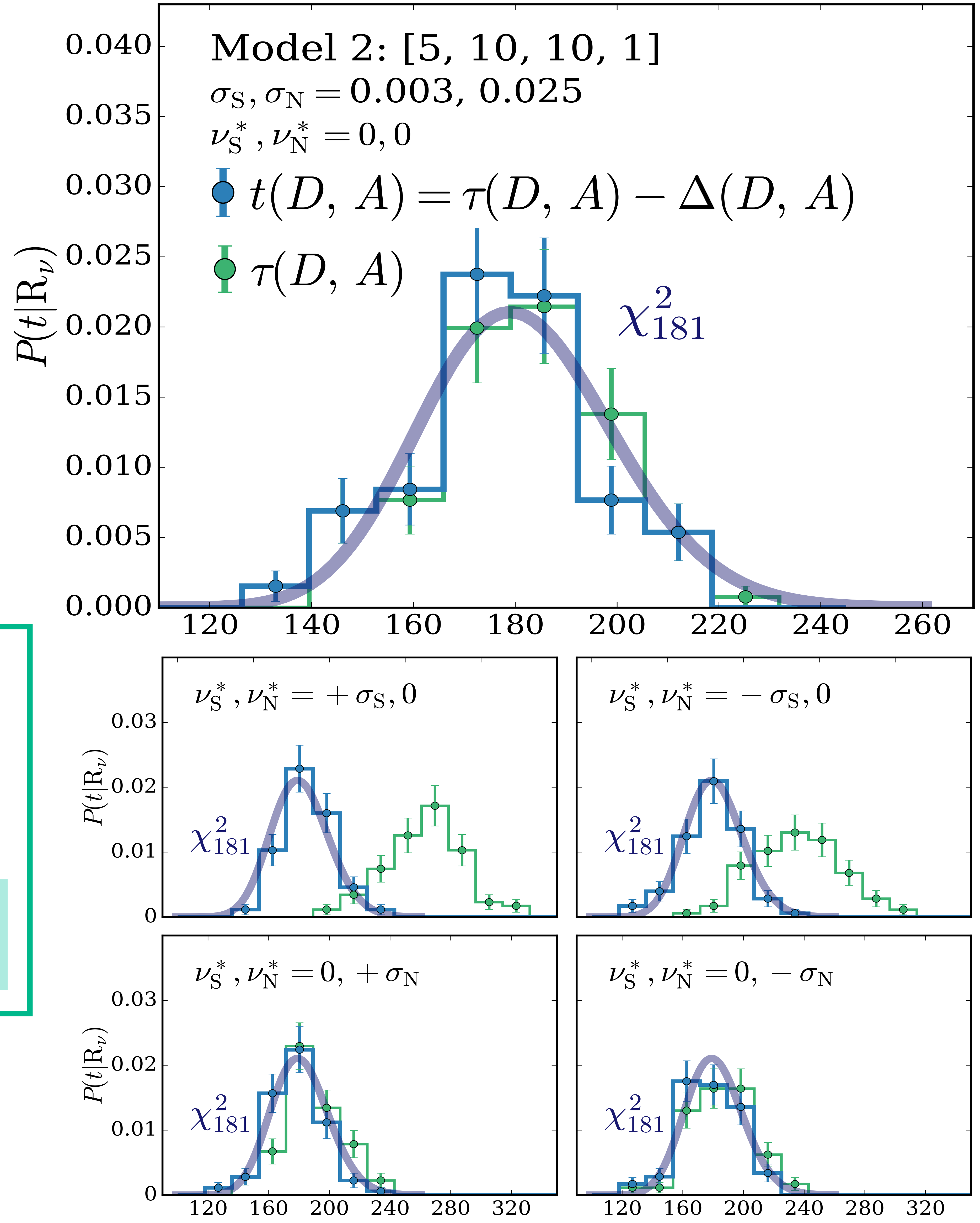
"Toy Data" : test the procedure on simulated toys following the Reference (SM) hypothesis with generation value for the nuisance parameters $\nu^* = \pm\sigma_\nu$:

$$\mathcal{D} \sim \mathrm{R}_{\boldsymbol{\nu}^*}, \quad \boldsymbol{\nu}^* = \pm\sigma_{\boldsymbol{\nu}}$$

The $\bar{t}$ distribution under the reference hypothesis $\mathrm{R}_{\boldsymbol{\nu}^*}$ is **compatible with the target** $\chi^2_{|\mathbf{w}|}$ for values of the true nuisance parameters within the uncertainty ($\boldsymbol{\nu}^* = \pm\sigma_{\boldsymbol{\nu}}$).

$\bar{t}$ is **independent** of the true value of the nuisance parameters!

We can build a *frequentistic* test statistic relying on the asymptotic $\chi^2_{|\mathbf{w}|}$.



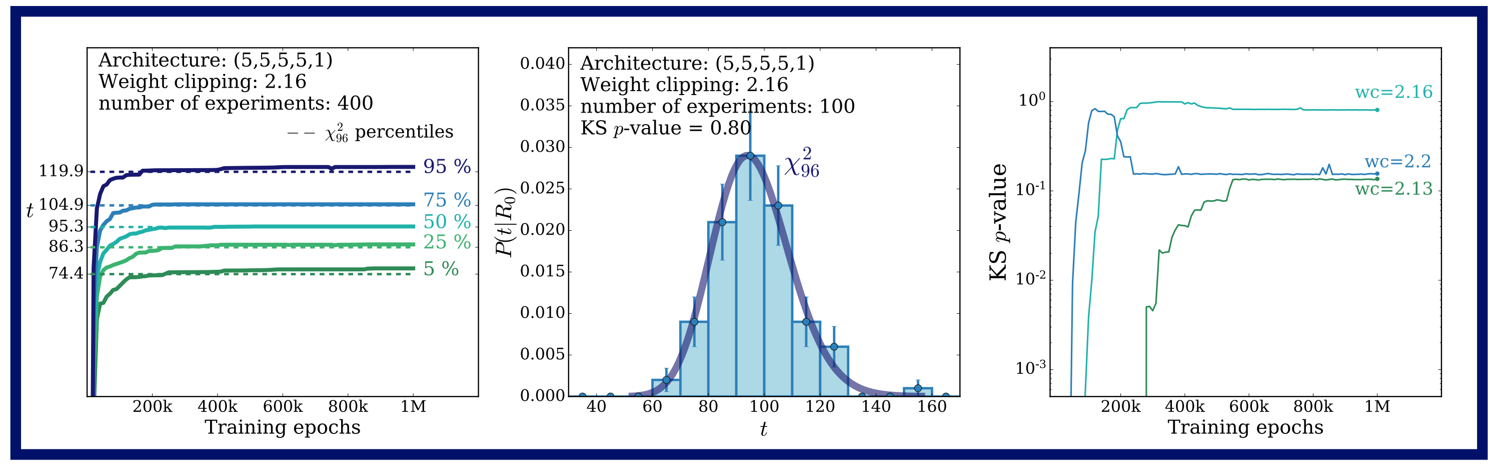
New Physics Learning Machine (NPLM)

Including systematic uncertainties

Final procedure in steps:

1. **NN (f) REGULARIZATION**:

   weight clipping tuning → target $\chi^2_{|\mathbf{w}|}$;



2. **NUISANCE TAYLOR'S EXPANSION LEARNING**:

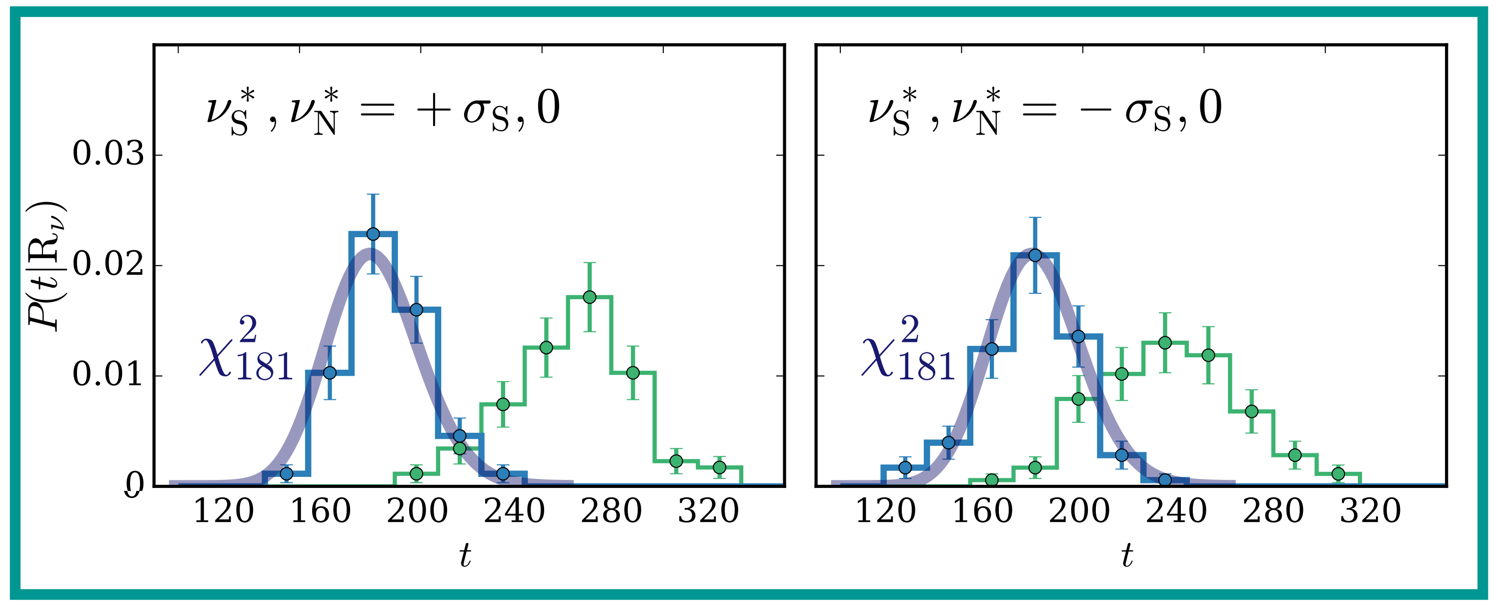
   modelling $\widehat{r}(x;\nu) = \exp\left[\widehat{\delta}_1(x)\,\nu + \widehat{\delta}_2(x)\,\nu^2 + \ldots\right]$;

$$\widehat{r}(x;\nu) = \exp\Big[\underbrace{\widehat{\delta}_1(x)}_{\text{NN 1}}\,\nu + \underbrace{\widehat{\delta}_2(x)}_{\text{NN2}}\,\nu^2 + \underbrace{\ldots}_{\ldots}\Big]$$

3. **$\tau - \Delta$ VALIDATION**:

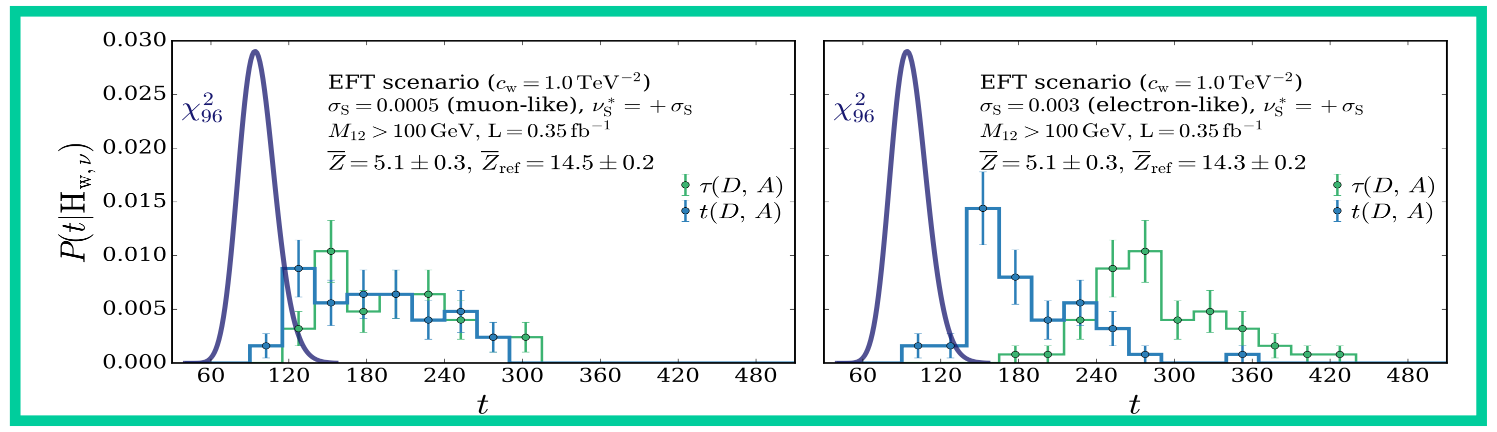
   $$\mathcal{D} \sim \mathrm{R}_{\boldsymbol{\nu}^*}, \quad \boldsymbol{\nu}^* = \pm\sigma_{\boldsymbol{\nu}}$$

   Verifying that the target $\chi^2_{|\mathbf{w}|}$ is always recovered;



4. **TESTING THE DATA**:

   running the procedure on real data.



Di-body final state at the LHC

Sensitivity to New Physics scenarios

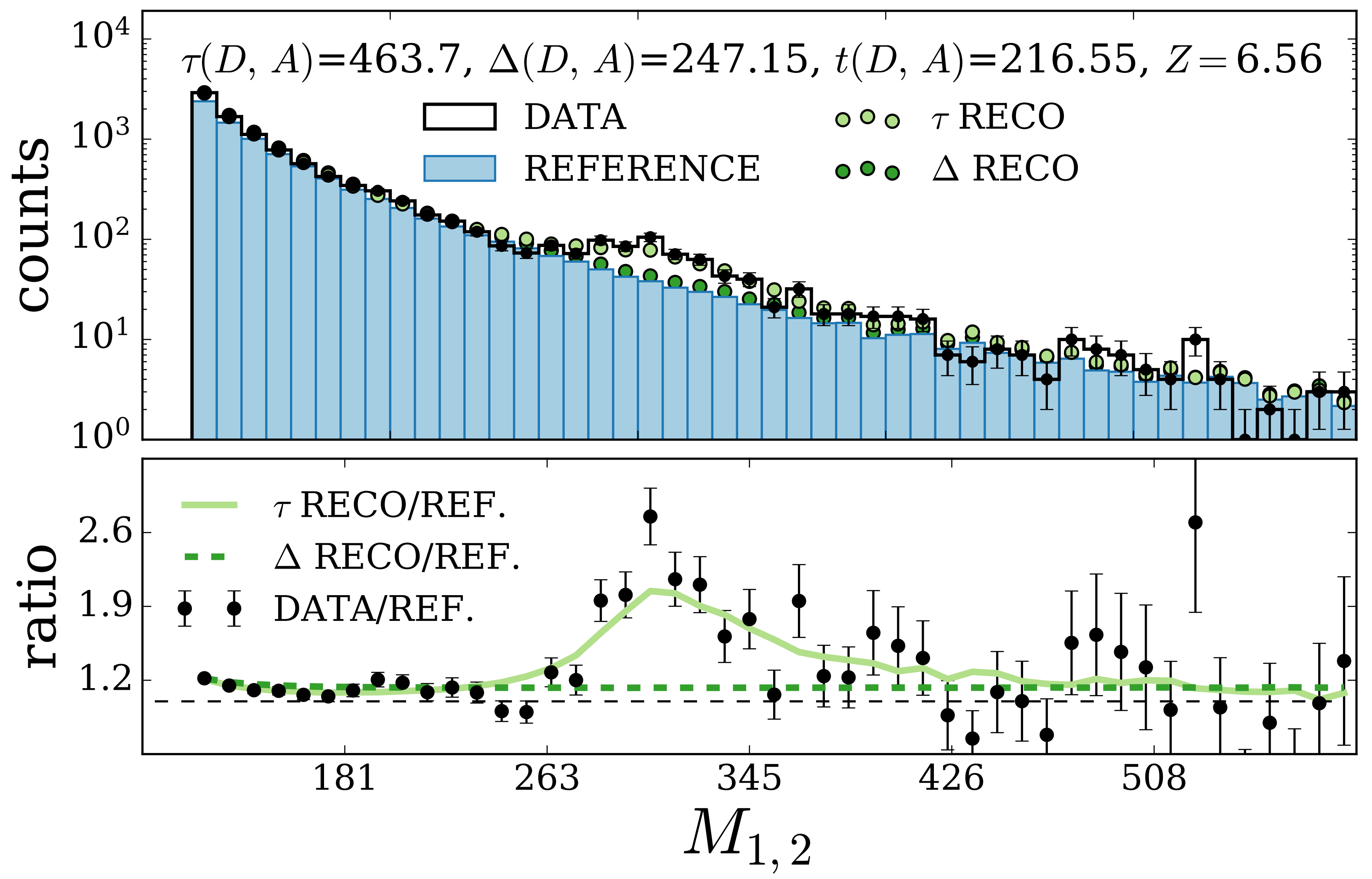
Summary of the results:

- Comparable performances in the resonant and non-resonant scenarios:
  - NPLM is simultaneously sensitive to multiple sources of New Physics;
- Comparable performances at different systematic uncertainties regimes:
  - NPLM is robust against the presence of systematic uncertainties;
  - the presence of systematic uncertainties affects NPLM in the same measure as any other hypothesis test;
- No information about the New Physics signal has been provided to the algorithm at any step of its implementation:
  - The performances of NPLM are lower than any model-dependent strategy by construction ($\overline{Z}/\overline{Z}_{\text{ref}} = 0.37$);
- NPLM is able to *learn* non trivial combinations of the input variables and point to the source of the significant discrepancy.

τ reconstruction: $n(x\,|\,\mathrm{H}_{\hat{\mathbf{w}},\hat{\nu}}) = n(x\,|\,\mathrm{R}_0)\dfrac{n(x\,|\,\mathrm{R}_{\hat{\nu}})}{n(x\,|\,\mathrm{R}_0)}\,e^{f(x;\,\hat{\mathbf{w}})}$

Δ reconstruction: $n(x\,|\,\mathrm{R}_{\hat{\nu}})$

Z' scenario (tau-like regime), m = 300 GeV, N(S) = 210



EFT scenario (tau-like regime), c_W = 0.25 TeV^{-2}