

Systematics in model-independent ML-based searches

Gaia Grosso

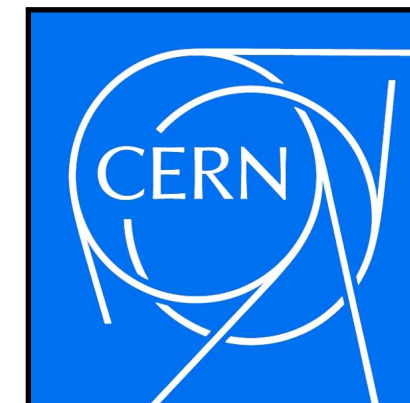
CERN, University and INFN of Padova

Mainly based on:

[Phys. Rev. D 99, 015014](#) (d'Agnolo, Wulzer)

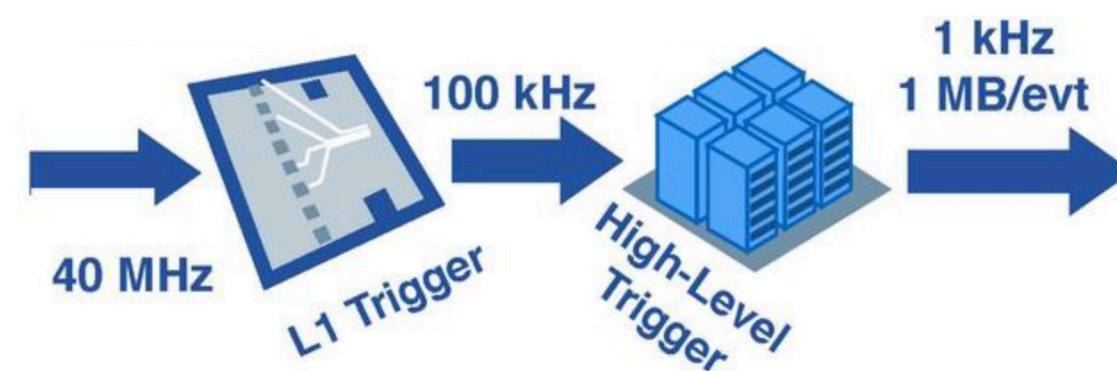
[Eur. Phys. J. C 81, 89 \(2021\)](#) (d'Agnolo, Grosso, Pierini, Wulzer, Zanetti)

[Eur. Phys. J. C 82, 275 \(2022\)](#) (d'Agnolo, Grosso, Pierini, Wulzer, Zanetti)

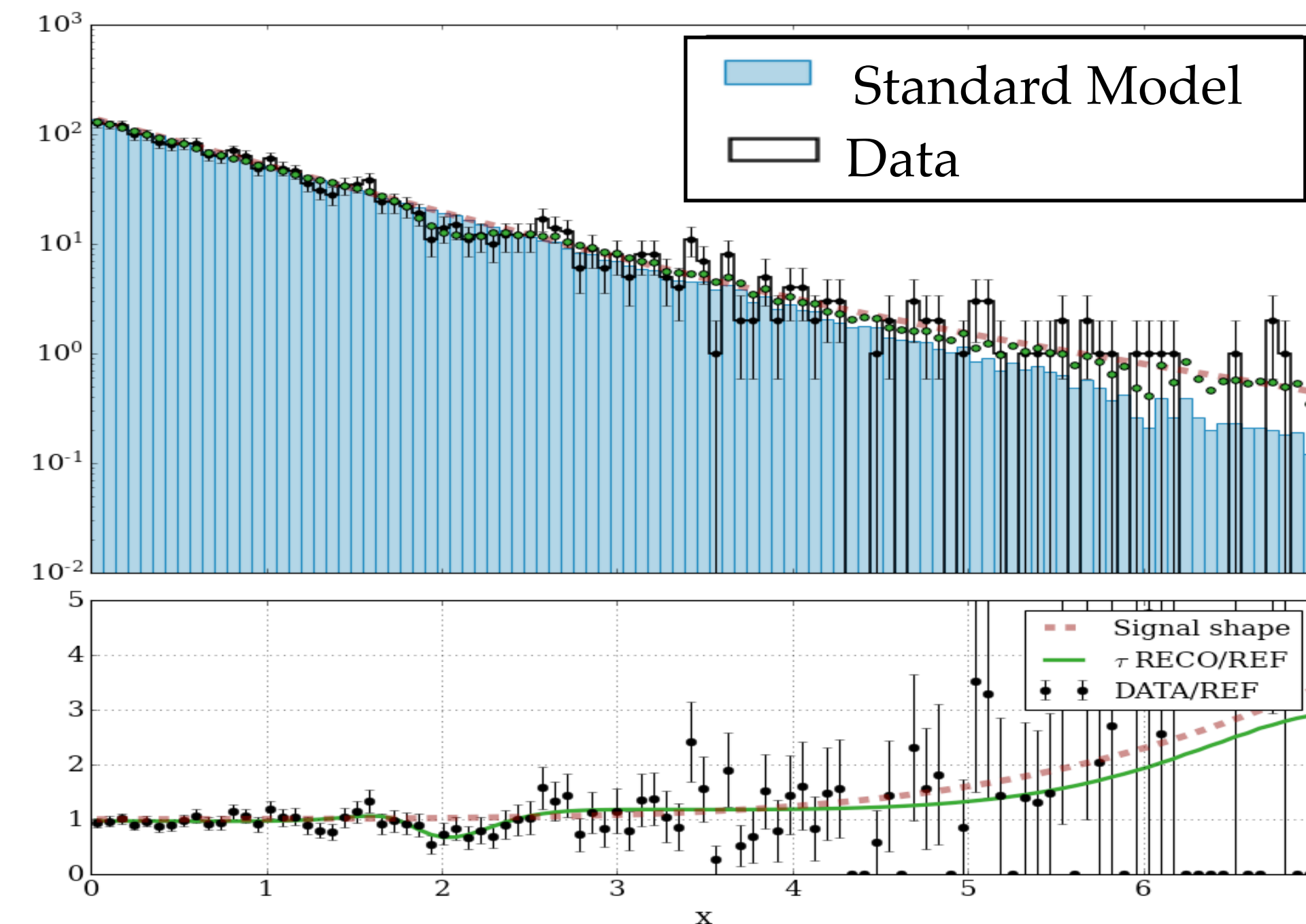


How good is the SM description of the data?

experiment



Summary statistics



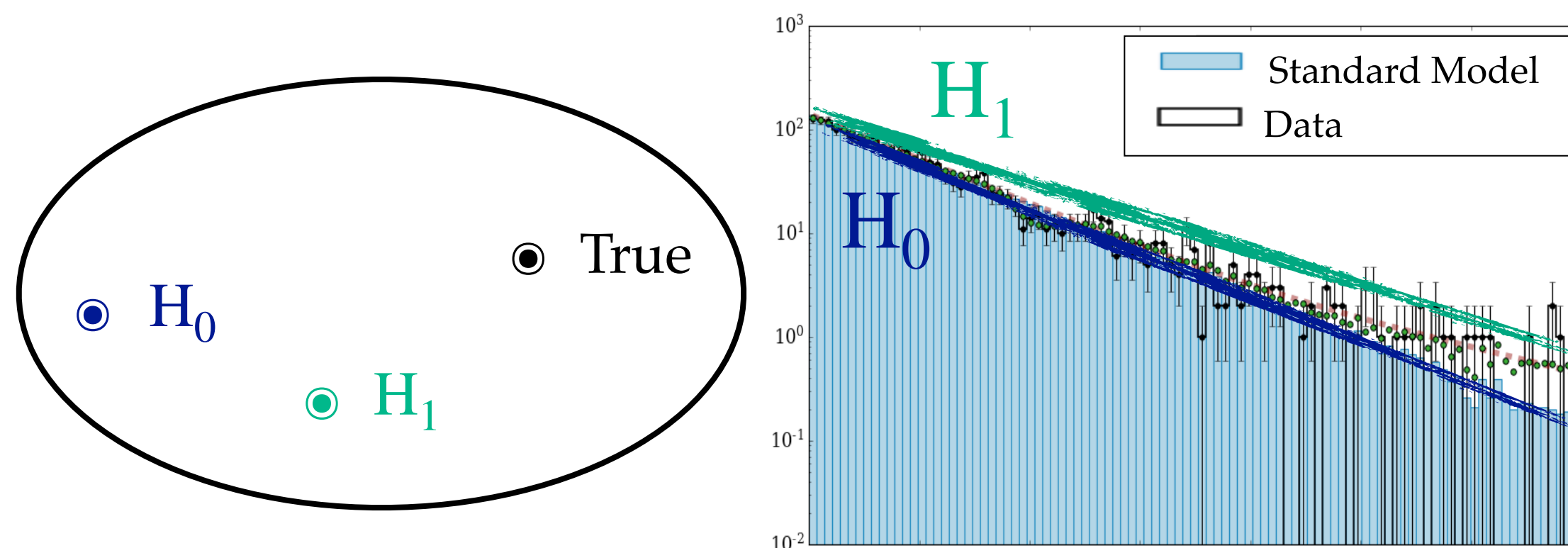
This is a problem of Goodness of Fit of the SM.

(The need for a “Reference” sample makes it a 2-samples test.)

Likelihood-ratio test

traditionally solved in HEP as a hypothesis test based on likelihoods ratio

→ Model-dependent approaches



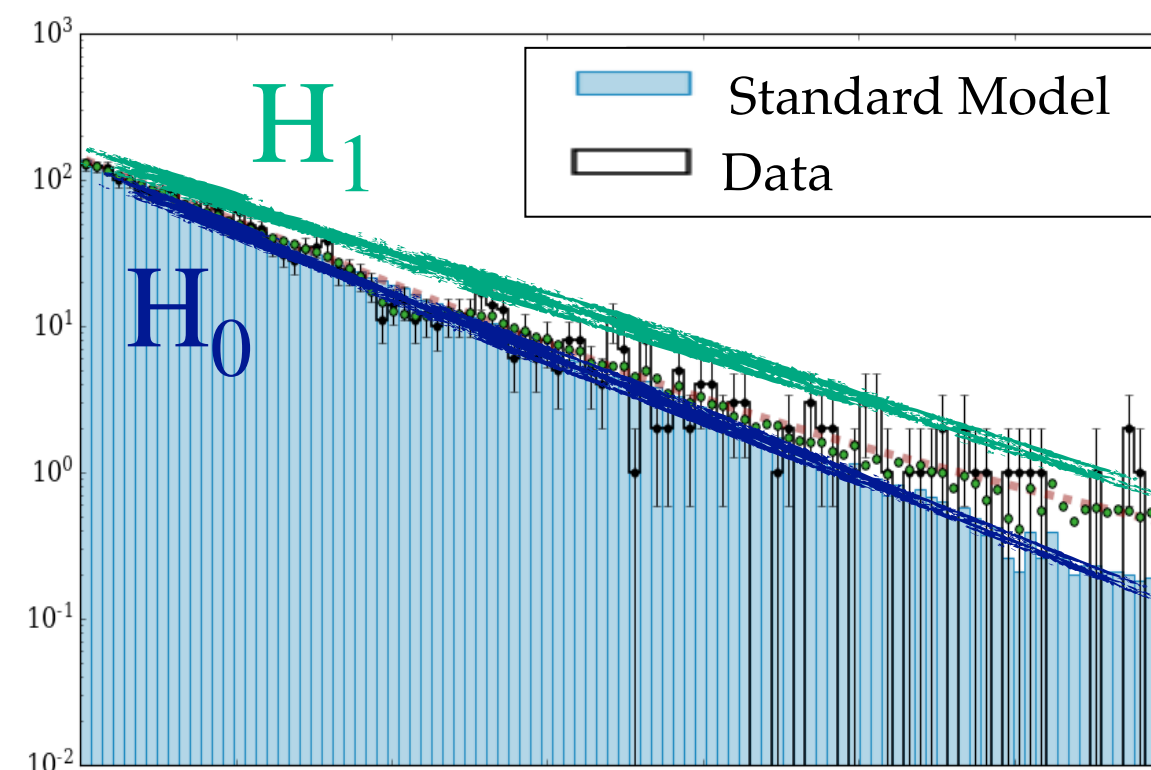
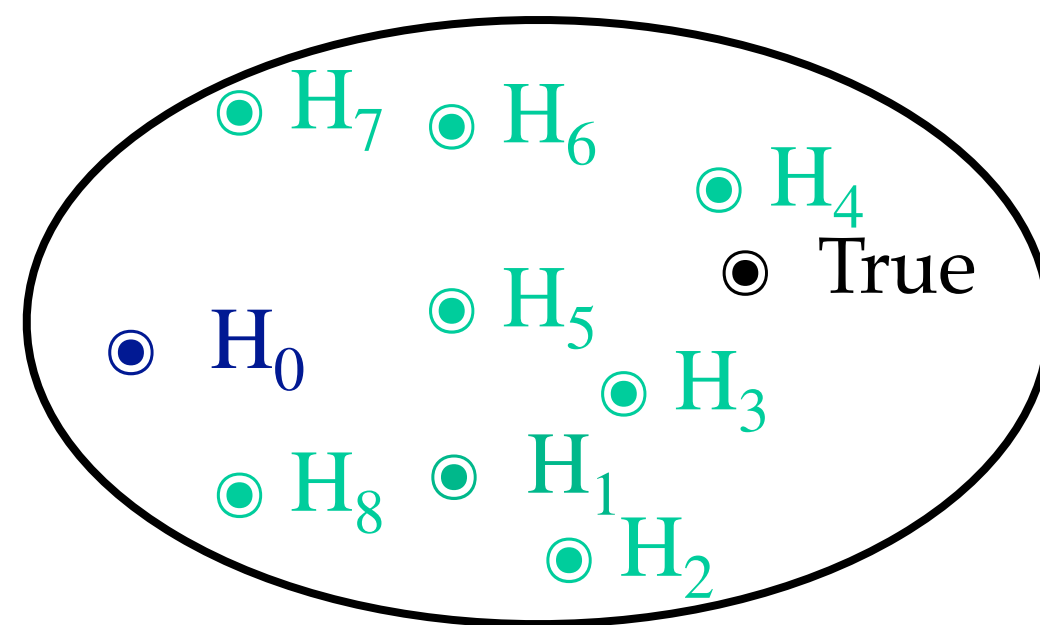
$$t(\mathcal{D}) = 2 \log \frac{\mathcal{L}(\mathcal{D} | H_1)}{\mathcal{L}(\mathcal{D} | H_0)}$$

Sensitivity (and **optimality**) guaranteed (according to Neyman and Pearson)
only if the the data do follow the chosen alternative

Likelihood-ratio test

traditionally solved in HEP as a hypothesis test based on likelihoods ratio

→ Model-dependent approaches

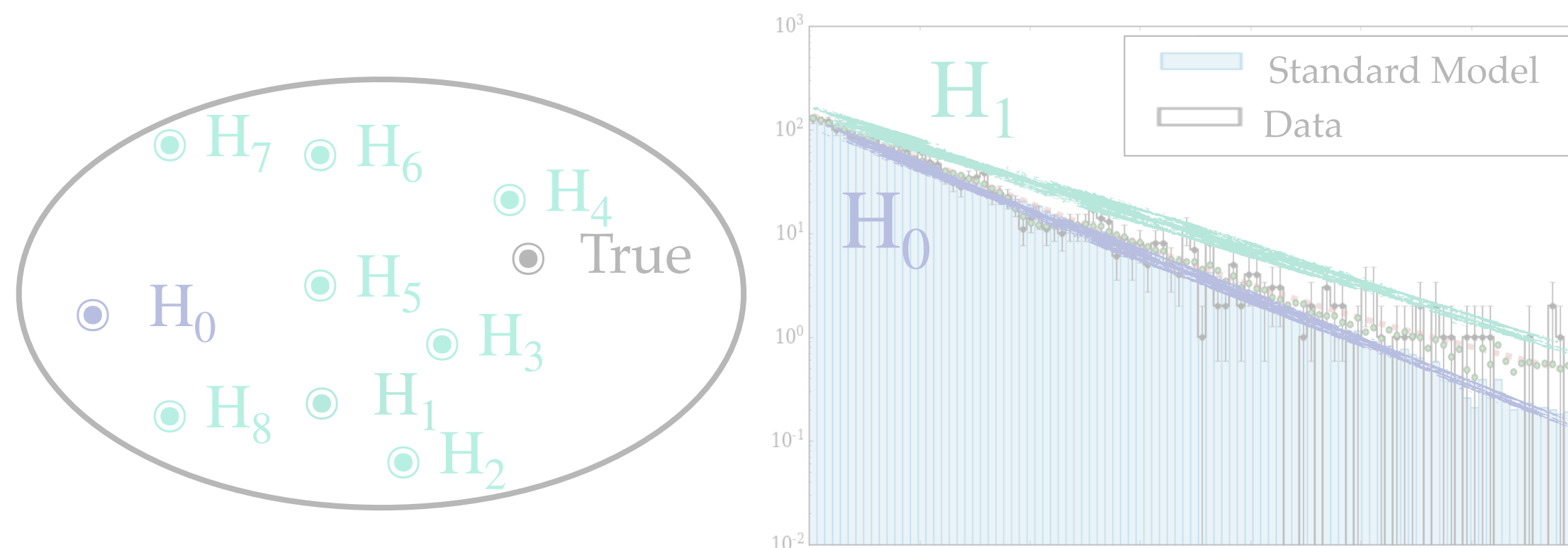


$$t(\mathcal{D}) = 2 \log \frac{\mathcal{L}(\mathcal{D} | H_1)}{\mathcal{L}(\mathcal{D} | H_0)}$$

Sensitivity (and **optimality**) guaranteed (according to Neyman and Pearson)
only if the the data do follow the chosen alternative

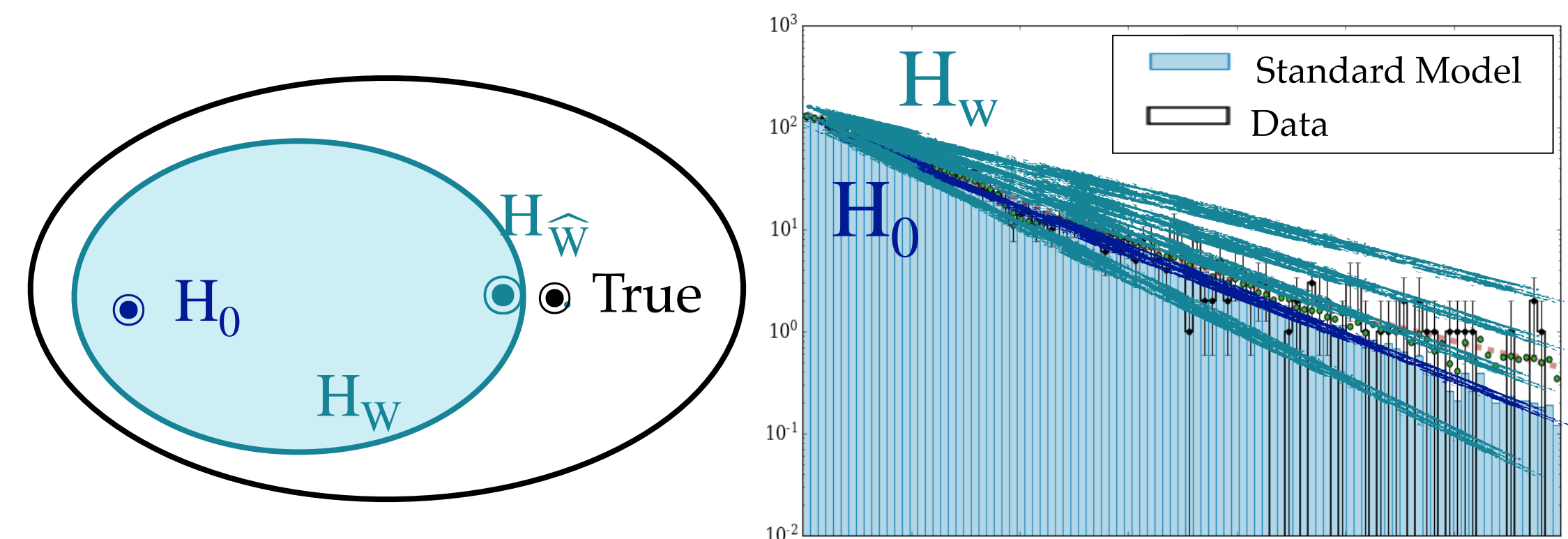
Likelihood-ratio test

→ Model-dependent approaches



$$t(\mathcal{D}) = 2 \log \frac{\mathcal{L}(\mathcal{D} | H_1)}{\mathcal{L}(\mathcal{D} | H_0)}$$

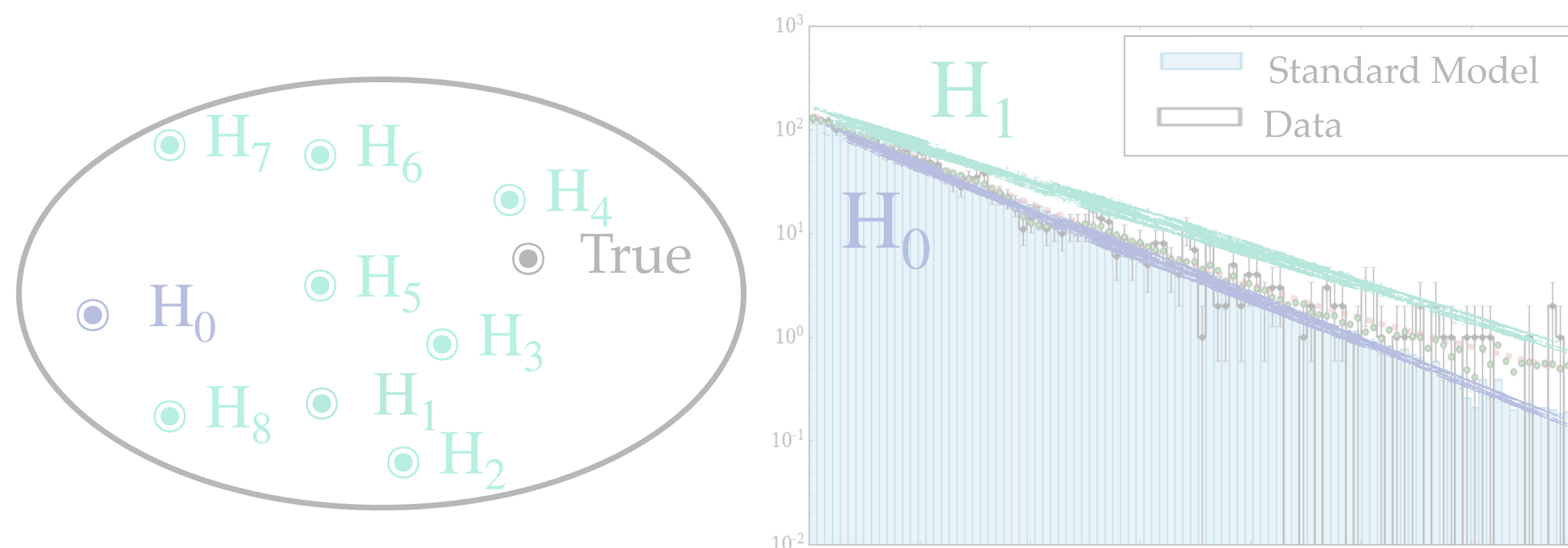
→ Model-*in*dependent approaches



$$t(\mathcal{D}) = \max_{\mathbf{w}} \left[2 \log \frac{\mathcal{L}(\mathcal{D} | H_{\mathbf{w}})}{\mathcal{L}(\mathcal{D} | H_0)} \right]$$

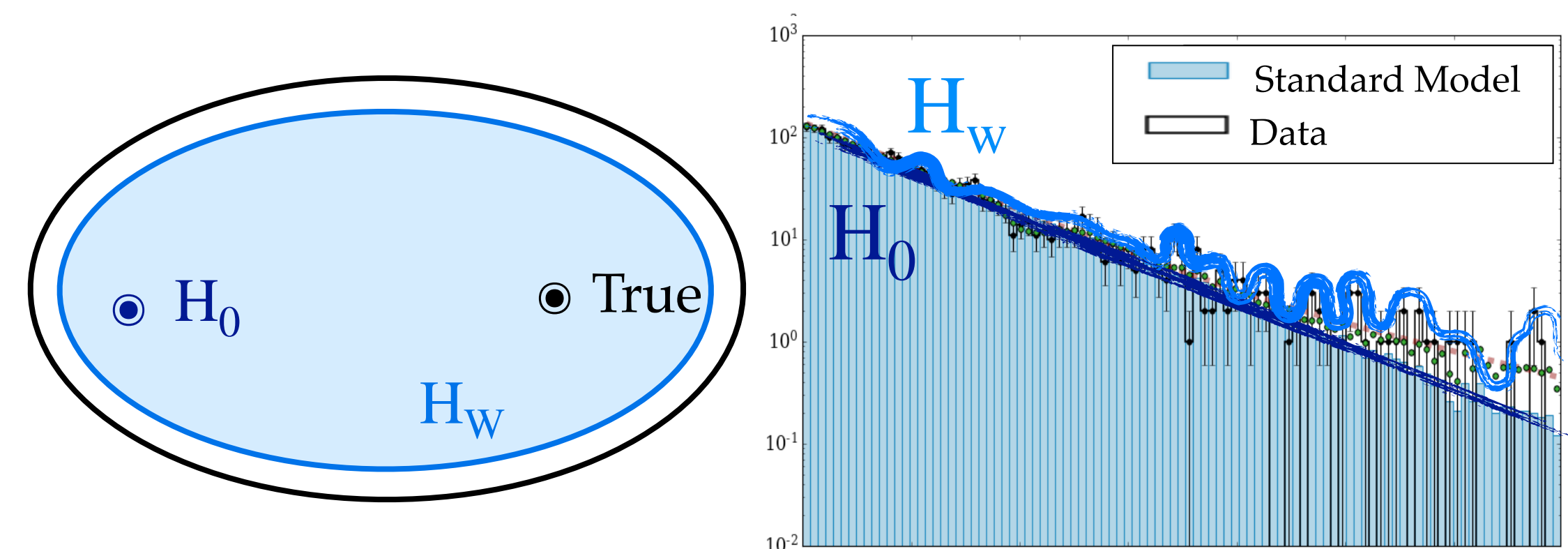
Likelihood-ratio test

→ Model-dependent approaches



$$t(\mathcal{D}) = 2 \log \frac{\mathcal{L}(\mathcal{D} | H_1)}{\mathcal{L}(\mathcal{D} | H_0)}$$

→ Model-*in*dependent approaches



$$t(\mathcal{D}) = \max_{\mathbf{w}} \left[2 \log \frac{\mathcal{L}(\mathcal{D} | H_{\mathbf{w}})}{\mathcal{L}(\mathcal{D} | H_0)} \right]$$

Machine learning based

Model-independent approaches

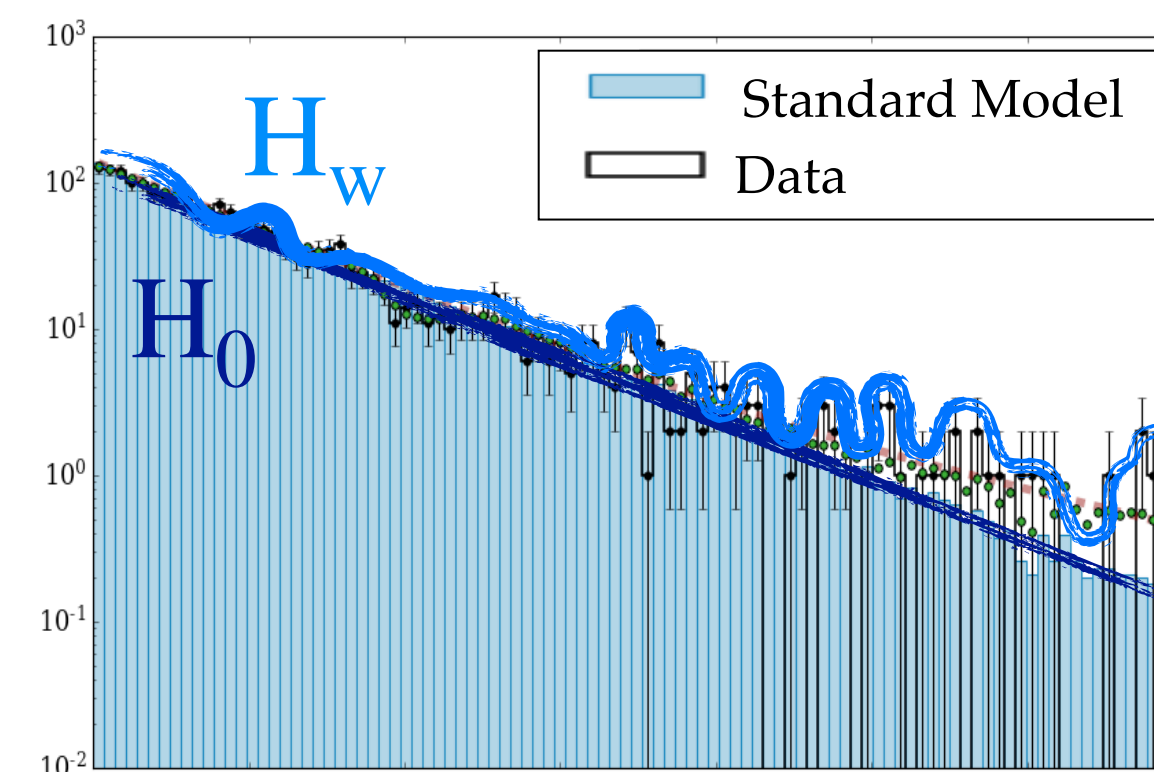
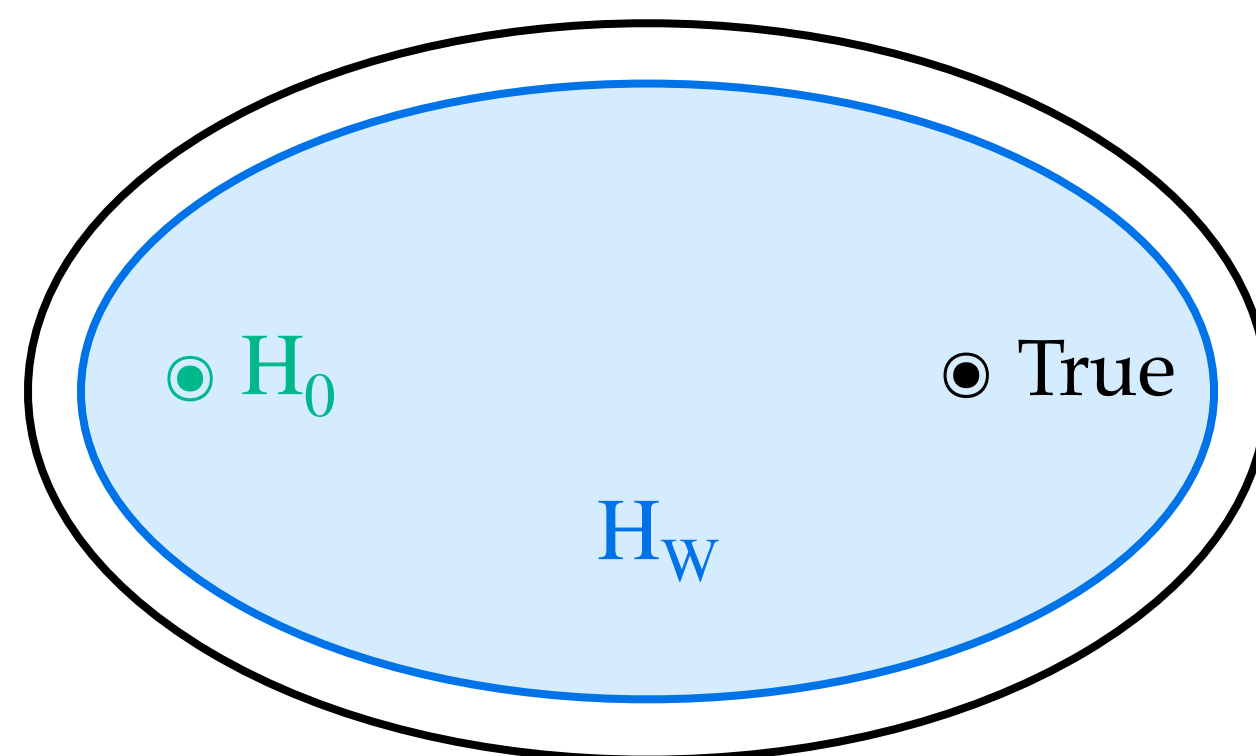
New Physics Learning Machine (NPLM)

Expand the family of alternatives
to increase the chance of containing
the True data distribution

“Statistical” anomaly detector

Universal approximator
(NN, kernel methods, ...)

$$t(\mathcal{D}) = \max_{\mathbf{w}} \left[2 \log \frac{\mathcal{L}(\mathcal{D} | H_{\mathbf{w}})}{\mathcal{L}(\mathcal{D} | H_0)} \right]$$



H_0 : null hypothesis (SM)
 H_w : alternative hypothesis (NP)

New Physics Learning Machine (NPLM)

Maximum-Likelihood-ratio test statistic:

$$t(\mathcal{D}, \mathcal{A}) = 2 \log \left[\frac{\max_{\mathbf{w}, \nu} \mathcal{L}(H_{\mathbf{w}, \nu} | \mathcal{D}, \mathcal{A})}{\max_{\nu} \mathcal{L}(R_{\nu} | \mathcal{D}, \mathcal{A})} \right] = 2 \log \left[\frac{\max_{\mathbf{w}, \nu} \mathcal{L}(H_{\mathbf{w}, \nu} | \mathcal{D}) \mathcal{L}(\nu | \mathcal{A})}{\max_{\nu} \mathcal{L}(R_{\nu} | \mathcal{D}) \mathcal{L}(\nu | \mathcal{A})} \right]$$

R_{ν} : reference hypothesis (null)

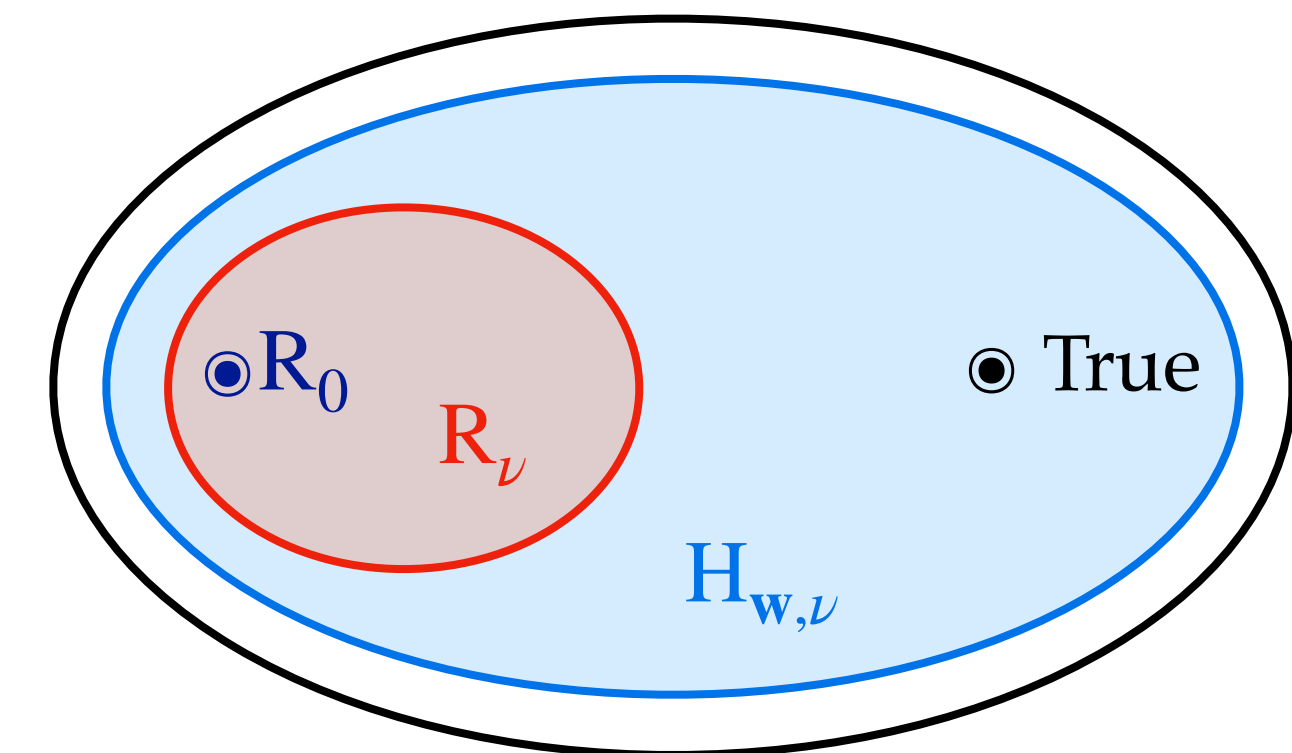
$H_{\mathbf{w}, \nu}$: alternative hypothesis

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

ν : set of nuisance parameters modelling the uncertainties effects

\mathcal{D} : data sample

\mathcal{A} : auxiliary sample (used to constrain ν)



New Physics Learning Machine (NPLM)

Maximum-Likelihood-ratio test statistic:

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

\mathbf{R}_{ν} : reference hypothesis (null)

$\mathbf{H}_{\mathbf{w}, \nu}$: alternative hypothesis

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

ν : set of nuisance parameters modelling the uncertainties effects

\mathcal{D} : data sample

\mathcal{A} : auxiliary sample (used to constrain ν)

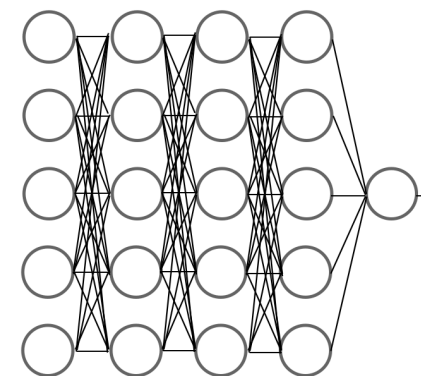
Parametrisation of the alternative hypothesis:

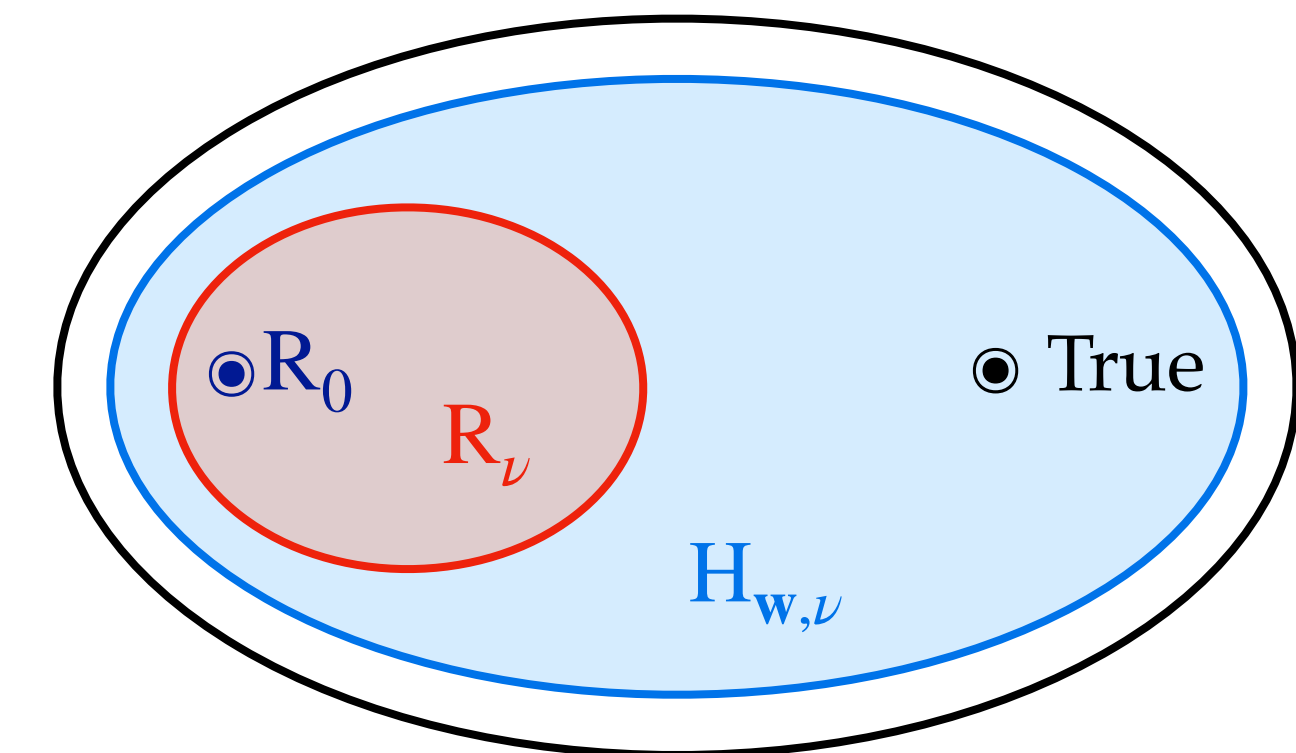
$$n(x | \mathbf{H}_{\hat{\mathbf{w}}, \hat{\nu}}) = \underbrace{n(x | \mathbf{R}_0)}_{\text{central value SM } (\nu=0)} \frac{n(x | \mathbf{R}_{\hat{\nu}})}{n(x | \mathbf{R}_0)} e^{f(x; \hat{\mathbf{w}})}$$

NN model

Model of systematic uncertainties effects

$$\hat{r}(x; \nu) = \frac{n(x | \mathbf{R}_{\nu})}{n(x | \mathbf{R}_0)} = \exp \left[\hat{\delta}_1(x) \nu + \hat{\delta}_2(x) \nu^2 + \dots \right]$$





Note:

This parametrisation choice guarantees $\mathbf{R}_{\nu} \subseteq \mathbf{H}_{\mathbf{w}, \nu}$
 ($\mathbf{R}_{\nu} = \mathbf{H}_{\mathbf{w}, \nu}$ for $f(\cdot; \mathbf{w}) \equiv 0$)

“Learning New Physics from an Imperfect Machine” [Eur. Phys. J. C](#)

New Physics Learning Machine (NPLM)

Maximum-Likelihood-ratio test statistic:

$$t(\mathcal{D}, \mathcal{A}) = 2 \log \left[\frac{\max_{\mathbf{w}, \nu} \mathcal{L}(H_{\mathbf{w}, \nu} | \mathcal{D}) \mathcal{L}(\nu | \mathcal{A})}{\max_{\nu} \mathcal{L}(R_{\nu} | \mathcal{D}) \mathcal{L}(\nu | \mathcal{A})} \right] \cdot \frac{\mathcal{L}(R_0 | \mathcal{D}) \mathcal{L}(\mathbf{0} | \mathcal{A})}{\mathcal{L}(R_0 | \mathcal{D}) \mathcal{L}(\mathbf{0} | \mathcal{A})}$$

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

R_{ν} : reference hypothesis (null)

$H_{\mathbf{w}, \nu}$: alternative hypothesis

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

ν : set of nuisance parameters modelling the uncertainties effects

\mathcal{D} : data sample

\mathcal{A} : auxiliary sample (used to constrain ν)

Tau term:

$$\tau(\mathcal{D}, \mathcal{A}) = 2 \max_{\mathbf{w}, \nu} \log \left[\frac{\mathcal{L}(H_{\mathbf{w}, \nu} | \mathcal{D}) \mathcal{L}(\nu | \mathcal{A})}{\mathcal{L}(R_0 | \mathcal{D}) \mathcal{L}(\mathbf{0} | \mathcal{A})} \right] = -2 \min_{\mathbf{w}, \nu} L \left[f(x, \mathbf{w}), \nu; \hat{\delta}(x) \right]$$

Depends on the NN model,
sensitive to New Physics

Delta term:

$$\Delta(\mathcal{D}, \mathcal{A}) = 2 \max_{\nu} \log \left[\frac{\mathcal{L}(R_{\nu} | \mathcal{D}) \mathcal{L}(\nu | \mathcal{A})}{\mathcal{L}(R_0 | \mathcal{D}) \mathcal{L}(\mathbf{0} | \mathcal{A})} \right] = -2 \min_{\nu} L \left[\nu; \hat{\delta}(x) \right]$$

Purely SM term, sensitive only to
uncertainties related discrepancies

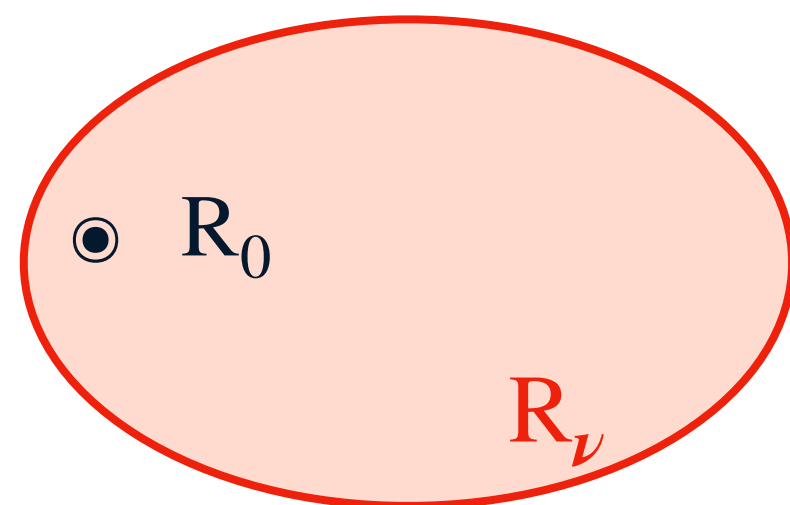
New Physics Learning Machine (NPLM)

Modelling the family of Reference hypotheses

The dependences on nuisance parameters, both in τ and Δ , are modelled via $r(x; \nu)$

r modelling

$$r(x; \nu) \equiv \frac{n(x|\mathbf{R}_\nu)}{n(x|\mathbf{R}_0)}$$



Normalization uncertainties:

Analytic description

$$r(x; \nu) \equiv \frac{n(x|\mathbf{R}_\nu)}{n(x|\mathbf{R}_0)} = \exp \left[\sum_{i=1}^{N_\nu} \nu_i \right]$$

Shape uncertainties:

Local Taylor's expansion around the nuisance central value

$$\hat{r}(x; \nu) = \exp \left[\hat{\delta}_1(x) \nu + \hat{\delta}_2(x) \nu^2 + \dots \right]$$

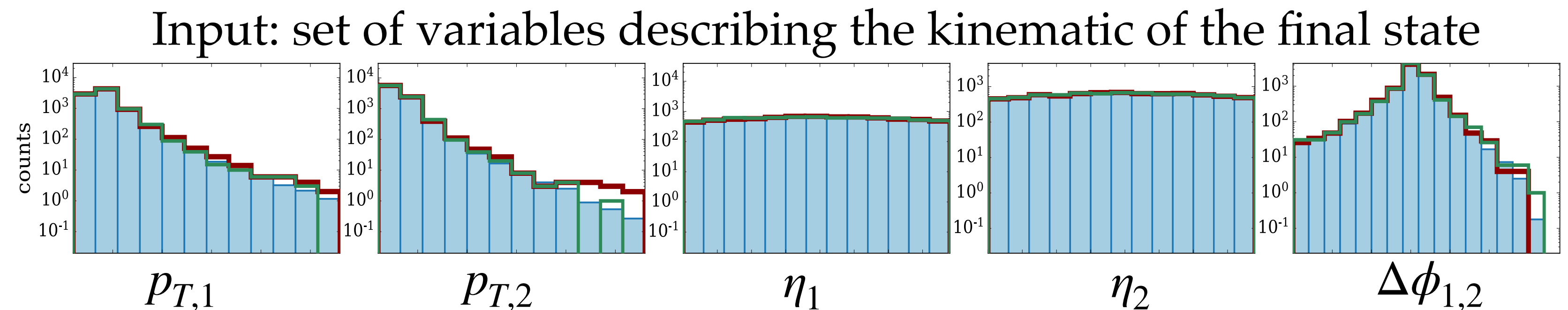
New Physics Learning Machine (NPLM)

Modelling the family of Reference hypotheses: shape effects

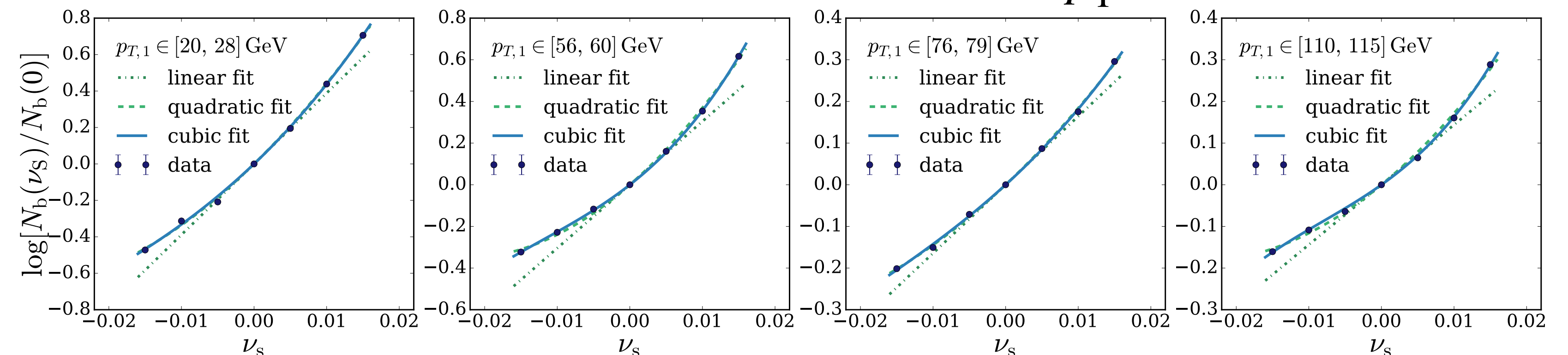
1) Preliminary study: Binned analysis to determine the proper order for the Taylor's expansion

Example:

Two-body final state
(5D analysis)



Bin-wise scale effects on the electrons p_T distribution



- Right order of approximation
- Where and how many instances are needed

New Physics Learning Machine (NPLM)

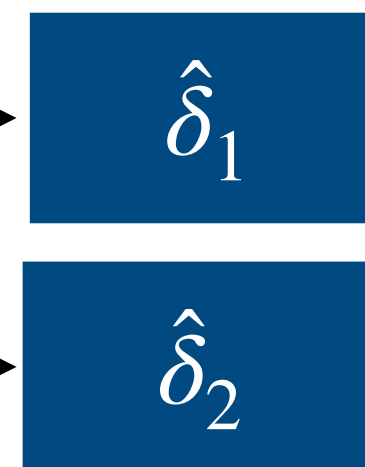
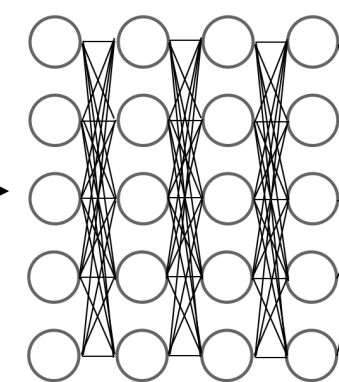
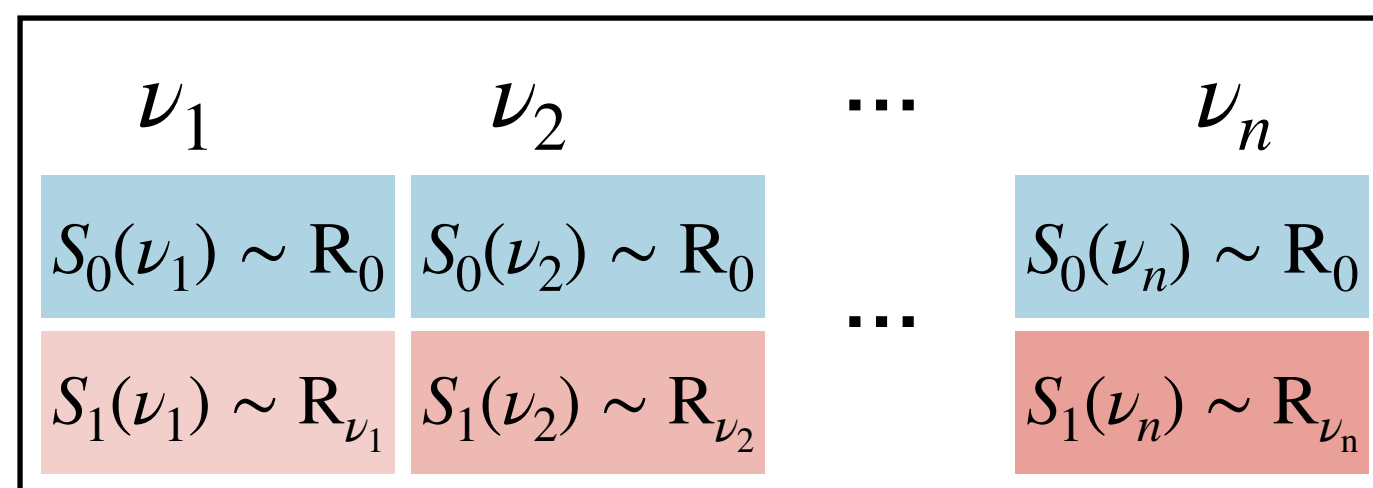
Modelling the family of Reference hypotheses: shape effects

2) Taylor's expansion learning: Training a neural network model to learn each coefficient of the Taylor's expansion of

$$r(x; \nu) = \frac{n(x | R_\nu)}{n(x | R_0)}$$

Parametrised classifier

Input samples



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

Loss function*

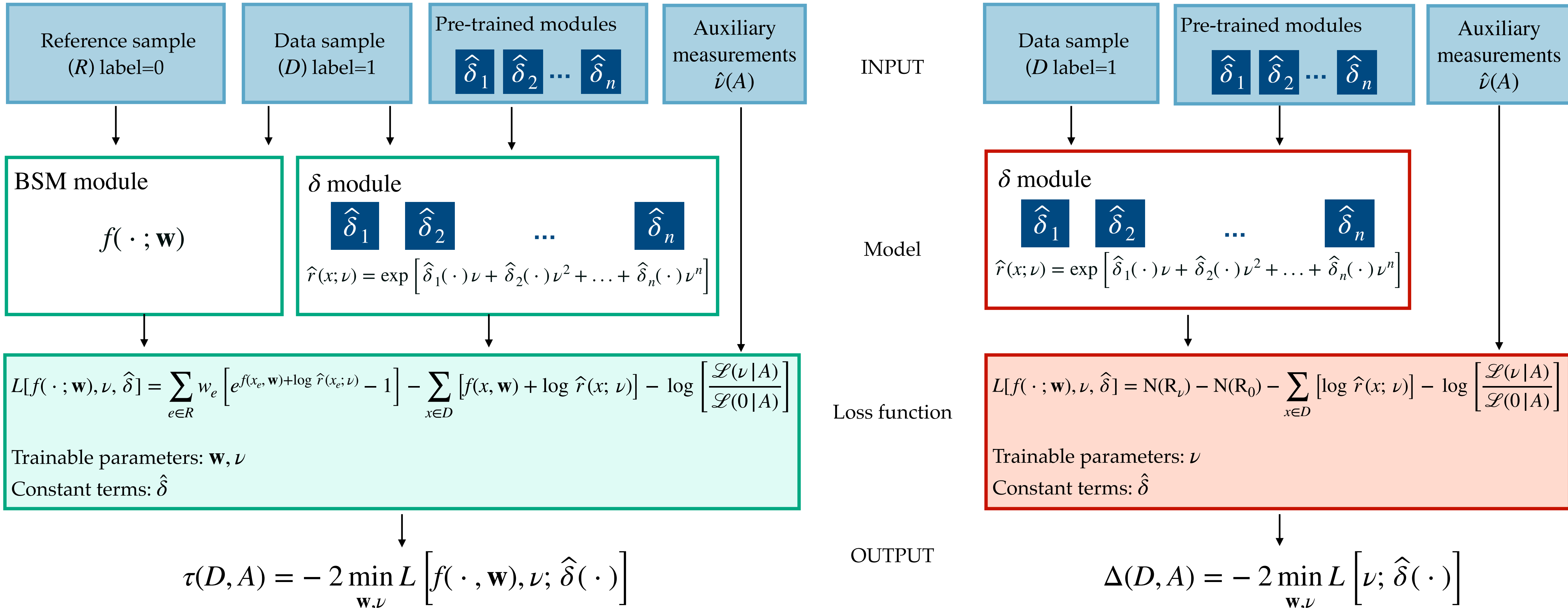
$$L[\hat{\delta}(\cdot)] = \sum_{\nu_i} \left[\sum_{e \in S_0(\nu_i)} w_e c(x_e)^2 + \sum_{e \in S_1(\nu_i)} w_e [1 - c(x_e)]^2 \right], \quad c(x) = \frac{1}{1 + \hat{r}(x; \nu)}$$

* Parametrized classifiers for optimal EFT sensitivity [arXiv:2007.10356](https://arxiv.org/abs/2007.10356)

New Physics Learning Machine (NPLM)

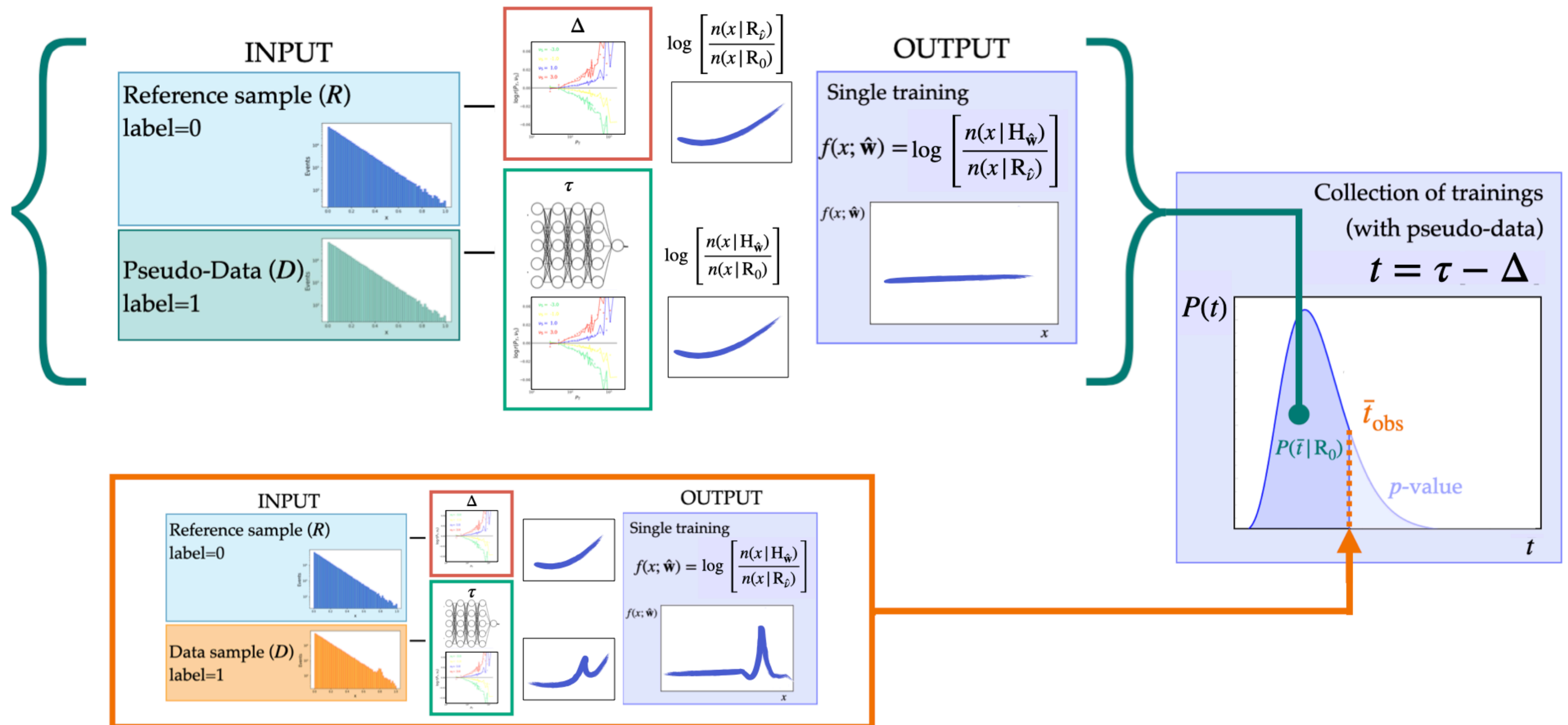
τ term

Δ term



New Physics Learning Machine (NPLM)

Calibration



New Physics Learning Machine (NPLM)

Controlling type I errors: NN model regularisation

Due to the **finite size** of the training samples, the **sparsity** of the data (especially in multivariate problems) and the **ill-definition of the loss** (unbounded from below), the distribution of $t(D)$ under R_0 doesn't converge to a stable configuration naturally (hope for the emergence of a χ^2).

→ a (NN) MODEL **REGULARIZATION** procedure can solve this problem!

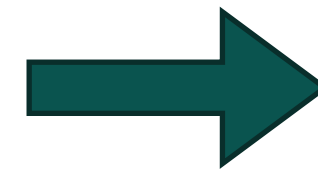
The regularisation enforces a level of *smoothness* in the model, preventing from “overfitting” sparse data points in the sample.

New Physics Learning Machine (NPLM)

Controlling type I errors: NN model regularisation

Weight clipping parameter:

Upper boundary to the magnitude that each trainable parameter can assume during the training.







For a chosen NN architecture, tuning the weight clipping allows to recover a good agreement of the empirical distribution of t under R_0 with a target $\chi^2_{|w|}$ distribution.

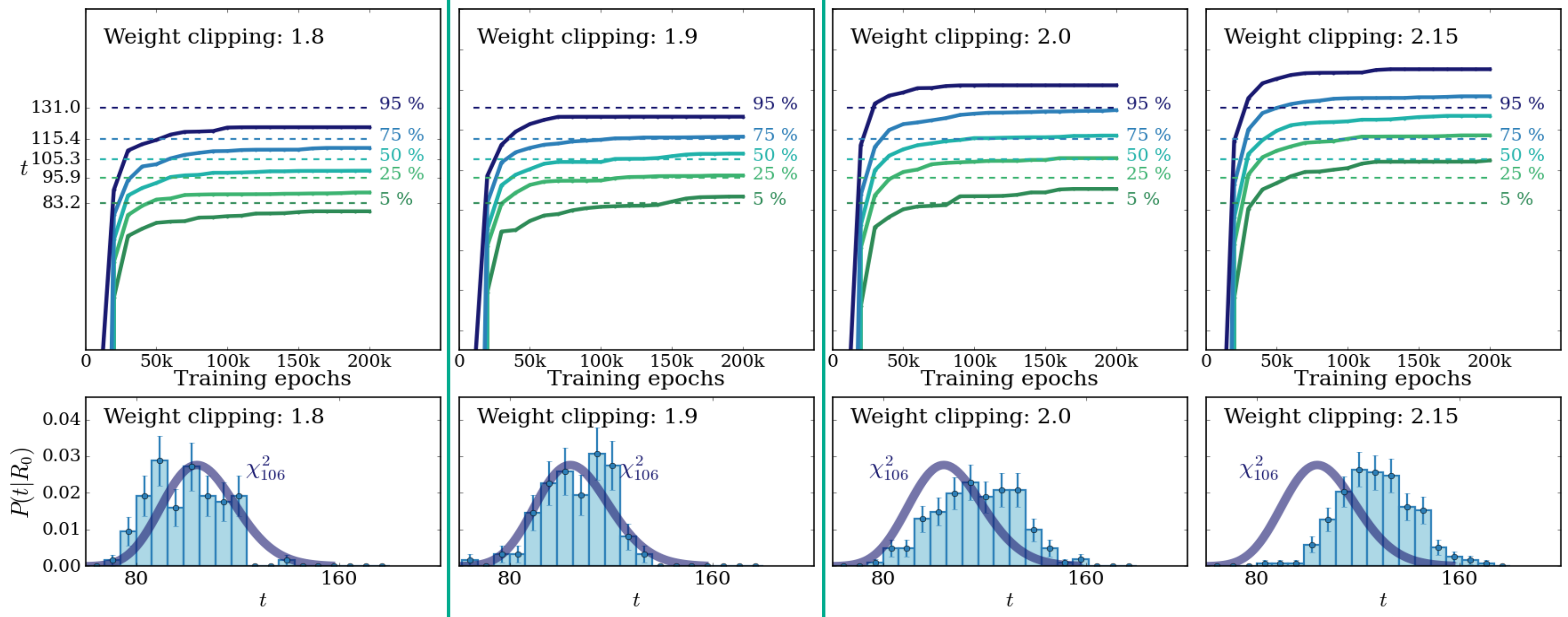
Example:

NN model: 5-7-7-1,

Number of parameters: 106

Legend:

-  Percentiles of the empirical \bar{t} distribution under R_0
-  Percentiles of the target $\chi^2_{|w|}$
-  Empirical \bar{t} distribution under R_0
-  Target $\chi^2_{|w|}$



“Learning Multivariate New Physics” [Eur. Phys. J. C](#)

New Physics Learning Machine (NPLM)

Controlling type I errors: 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 $\nu^* = \pm\sigma_\nu$:

$$\mathcal{D} \sim R_{\nu^*}, \quad \nu^* = \pm\sigma_\nu$$

The $t(D)$ distribution under the reference hypothesis R_{ν^*} is **compatible with the $\chi^2_{|w|}$** (found by regularizing) for values of the true nuisance parameters within the uncertainty ($\nu^* = \pm\sigma_\nu$).

t does not depend on the true value of the nuisance parameters!

We can build a *frequentist* test statistic targeting the $\chi^2_{|w|}$.

New Physics Learning Machine (NPLM)

Controlling type I errors: validation of the $(\tau - \Delta)$ procedure

Reference sample: $R \sim R_0$

Data sample: $D \sim R_{\nu^*}$

Example:

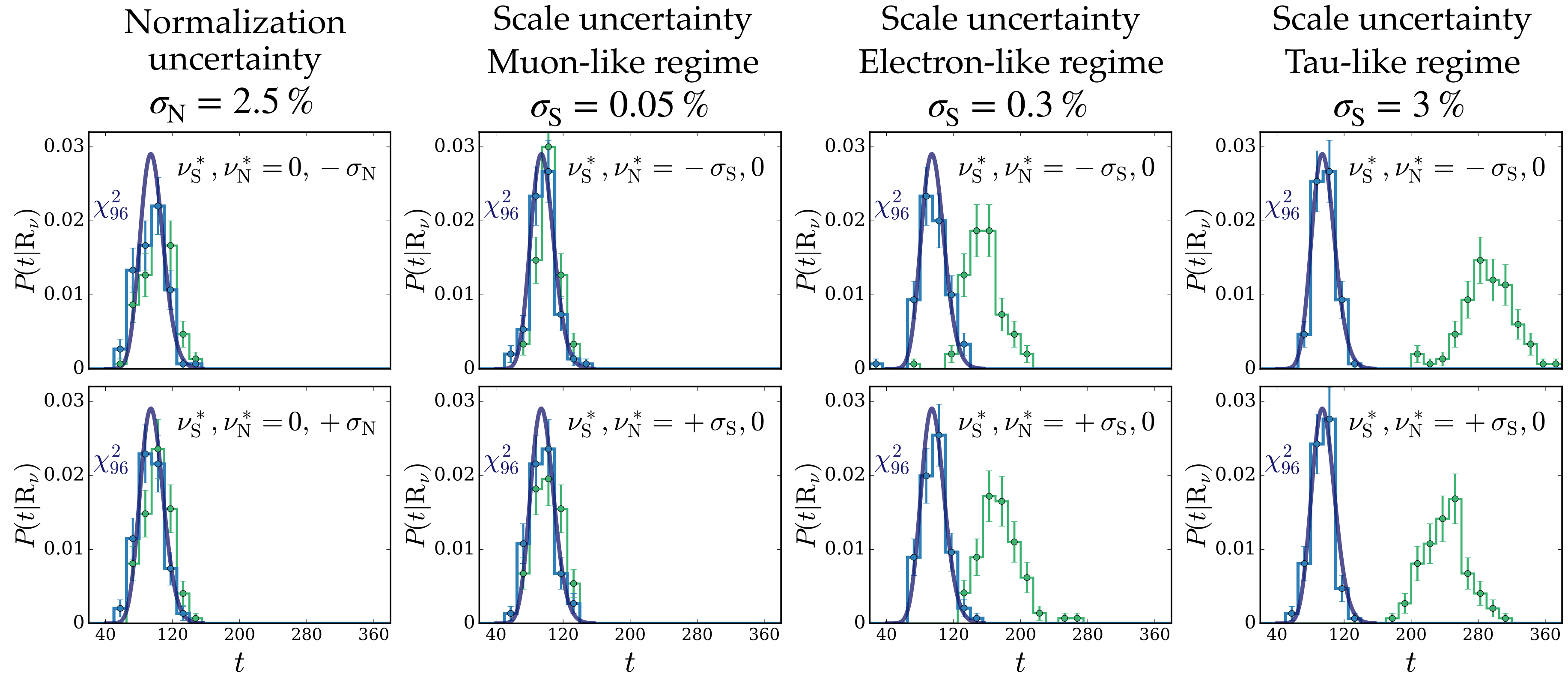
NN [5-5-5-5-1]

trainable parameters = 96

weight clipping = 2.16

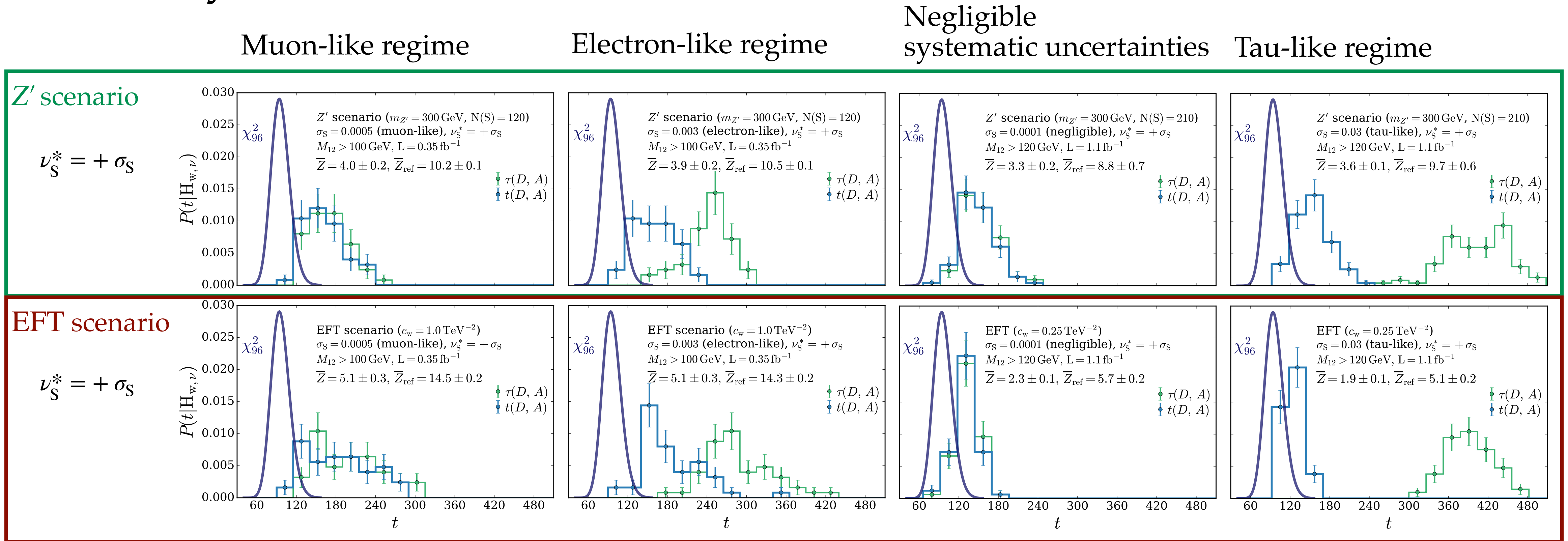
$\tau(D, A)$

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



New Physics Learning Machine (NPLM)

Sensitivity to BSM scenarios



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

\bar{Z} : Z-score from the median of the empirical $t(D, A)$ distribution

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

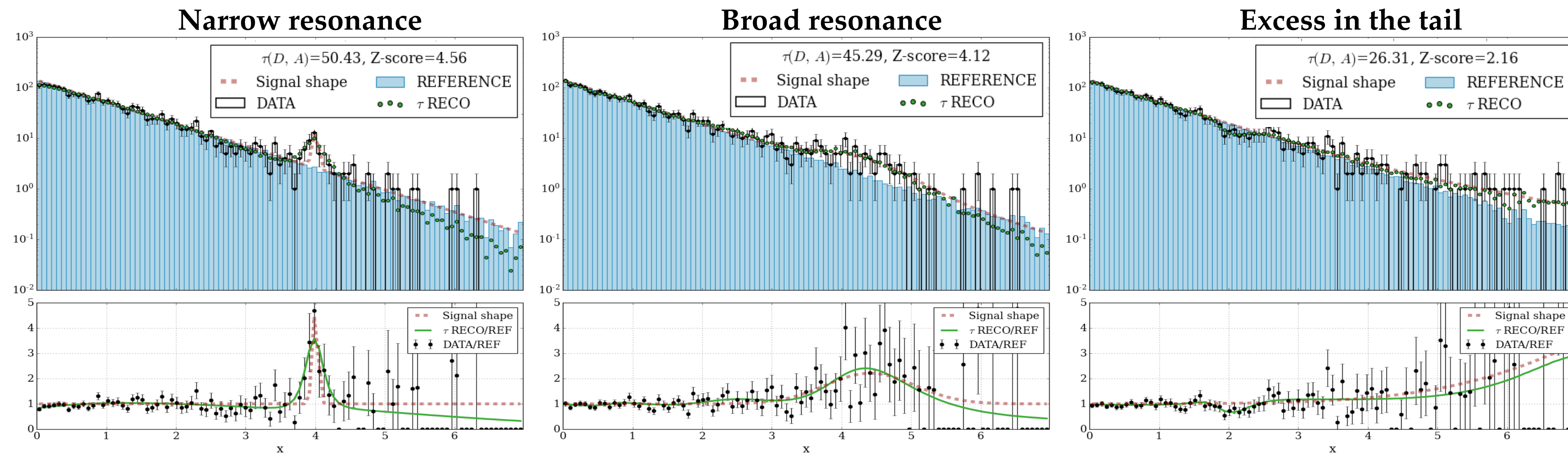
$\tau(D, A)$

New Physics Learning Machine (NPLM)

Summary and outlook:

NPLM features:

- End-to-end analysis strategy: from data to a **frequentist** p-value for discovery
- **Multivariate** (GoF) test
- Dealing with systematic uncertainties
- Sensitive to multiple signal patterns at once (**global** p-value)



Signal reconstruction with the NN: $n(x | H_{\hat{w}}) = n(x | R_0) e^{f(x; \hat{w})}$

Architecture: [1-4-1] (13 dof), weigh clipping 9, $N_{\text{bkg}} = 2000$

New Physics Learning Machine (NPLM)

Summary and outlook:

NPLM features:

- End-to-end analysis strategy: from data to a **frequentist** p-value for discovery
- **Multivariate** (GoF) test
- Dealing with systematic uncertainties
- Sensitive to multiple signal patterns at once (**global** p-value)

NPLM Challenges:

- Execution time (working on a fast **kernel methods** based implementation [Eur. Phys. J. C, 82\(10\)](#))
- Accurate multidimensional modelling of systematics (accuracy needed grows with the experimental luminosity)

New Physics Learning Machine (NPLM)

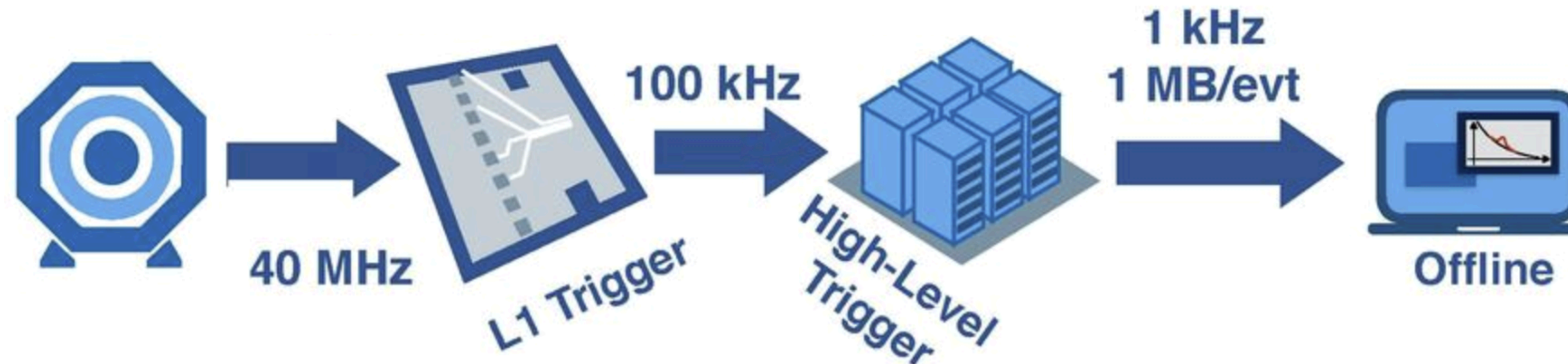
Summary and outlook:

Theoretical aspects to be further investigated:

- The model regularisation selects a portion of the space of the hypotheses, how to avoid performances losses? (exploring smarter regularisation alternatives)
- Emergence of a χ^2 for the test statistics distribution in the null hypothesis
- How to choose the (NN) model architecture? The χ^2 criterion offers an indication of good candidates, how to select one over them?

Model-independence beyond statistical tests

Data acquisition and experiment design



- Reducing uncertainties vs maximising power:
Can we be signal agnostic?

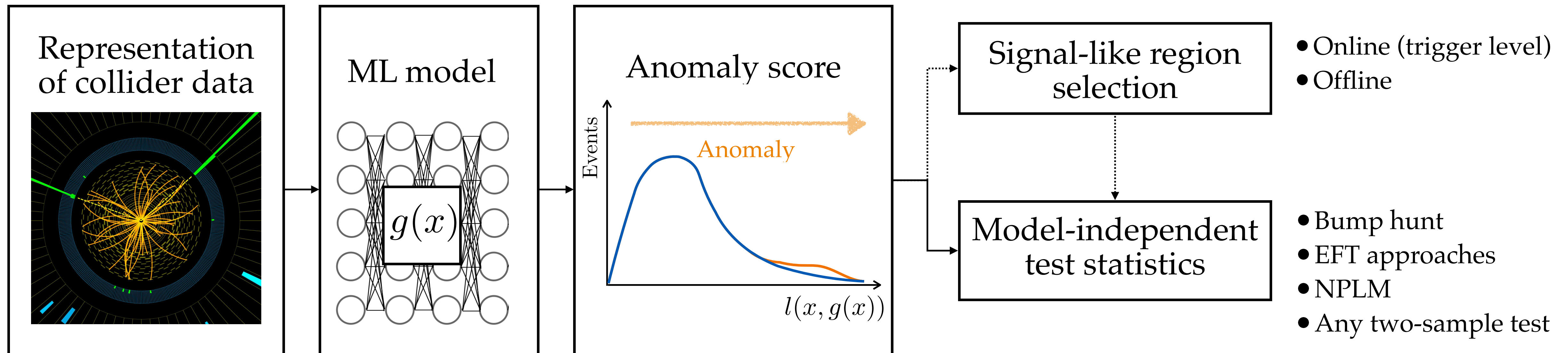
- Loose selection requirements
Only if more resources are available! (Hard)
- Smart selections
(ML-based) anomaly detection
- Online analysis/DQM

- Multivariate analysis
- Loose selections
- Model-independent test of the data (GoF)

NPLM

ML Tools for model-independent analyses

Anomaly detectors:



Some comments to be further discussed:

- Trained on data (no sys. unc. but need for control regions → less inclusive test)
- Trained on MC (study the response to input domain shifts, like those produced by systematic uncertainties)
- Which model-independent test to run on top? **Is there an “optimal” GoF test for model-independent searches?**

Backup slides

New Physics Learning Machine (NPLM)

Main Idea: Maximum Likelihood from minimal loss

Parametrizing the alternative distribution

$$n(x|\mathbf{H}_{\mathbf{w}}) = e^{f(x;\mathbf{w})} n(x|\mathbf{R}_0)$$

\mathbf{R}_0 : Reference (null) hypothesis (SM)

$\mathbf{H}_{\mathbf{w}}$: Alternative hypothesis (BSM)

Test statistic

$$\bar{t}(\mathcal{D}) = 2 \max_{\mathbf{w}} \log \left[\frac{\mathcal{L}(\mathcal{D} | \mathbf{H}_{\mathbf{w}})}{\mathcal{L}(\mathcal{D} | \mathbf{R}_0)} \right] = 2 \max_{\mathbf{w}} \left\{ \log \left[\frac{e^{-N(\mathbf{w})}}{e^{-N(\mathbf{R})}} \prod_{i=1}^{N_{\mathcal{D}}} \frac{n(x_i|\mathbf{w})}{n(x_i|\mathbf{R})} \right] \right\}$$

$$= -2 \min_{\mathbf{w}} \{ \bar{L}[f(\cdot; \mathbf{w})] \}$$

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

\mathcal{D} : data sample

\mathcal{R} : reference sample (built according to the \mathbf{R}_0 hypothesis); could be weighted (w)

Loss function

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

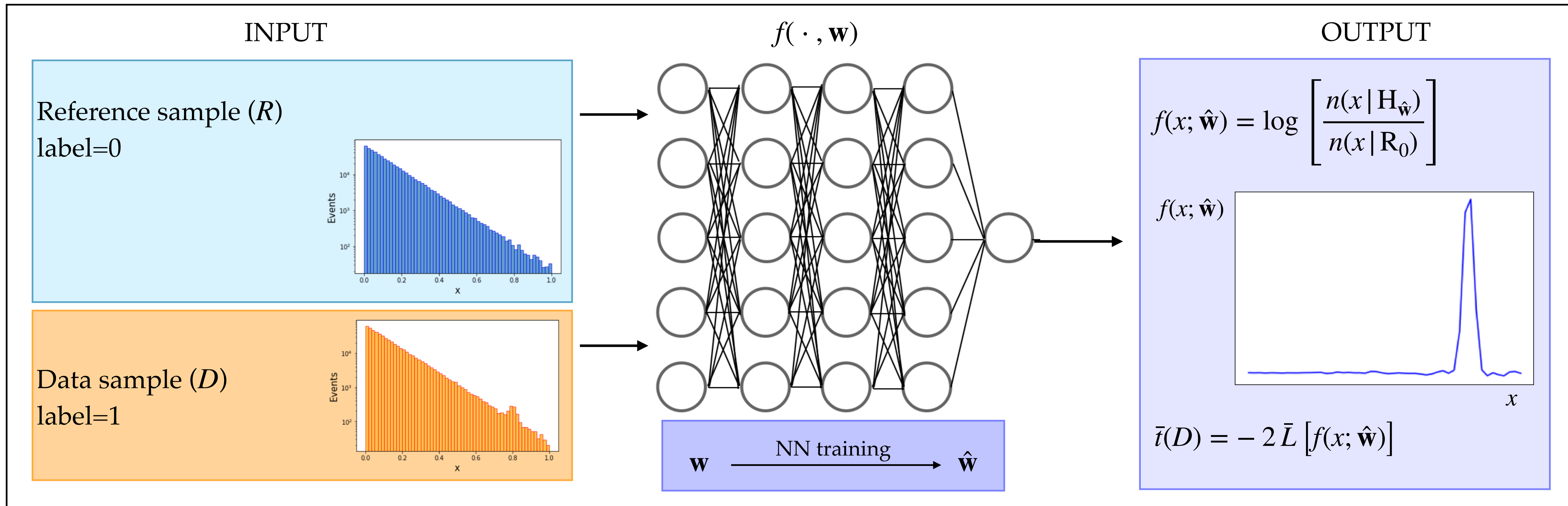
Assumptions:

- $N_R \gg N_D$ 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 $\sum_{x \in \mathcal{R}} w_x = N(\mathbf{R}_0)$

"Learning New Physics from a Machine" [Phys. Rev. D](#)

New Physics Learning Machine (NPLM)

Main Idea: Maximum Likelihood from minimal loss



New Physics Learning Machine (NPLM)

The fit on nuisance parameters, both in τ and Δ , is constrained by an auxiliary likelihood term

We postulate that the new physics is absent in the auxiliary data*

$$\mathcal{L}(H_{\mathbf{w},\nu}|\mathcal{A}) = \mathcal{L}(R_{\nu}|\mathcal{A}) = \mathcal{L}(\nu|\mathcal{A})$$

Auxiliary term

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

We suppose that the nuisance parameters are measured from the auxiliary dataset, independently from each others

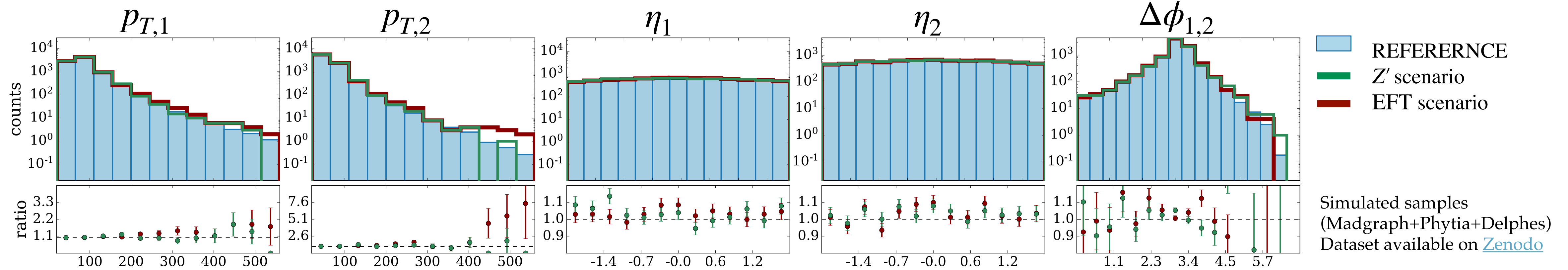
$$\mathcal{A} \rightarrow \hat{\nu}, \text{ random variable } \hat{\nu} \sim \mathbf{N}(\nu^*, \sigma_{\nu})$$

$$2 \log \left[\frac{\mathcal{L}(\nu|\mathcal{A})}{\mathcal{L}(\mathbf{0}|\mathcal{A})} \right] = \sum_{\nu_i} - \left(\frac{\hat{\nu}_i - \nu_i}{\sigma_i} \right)^2 + \left(\frac{\hat{\nu}_i}{\sigma_i} \right)^2$$

* note: we do see the effects of new physics in the auxiliary data nonetheless, if present (see discussion in section 2.6 of [Eur. Phys. J. C](#))

Two-body final state

5D analysis — Input: set of variables fully describing the kinematic of the final state



Uncertainties on the Reference hypothesis (SM):

- Global normalization effect: $\sigma_N = 2.5\%$
- Momentum scale effect:

$$p_{T1,2}^{(b,e)} = \exp \left[\nu_s \sigma_s^{(b,e)} / \sigma_s^{(b)} \right] p_{T1,2}^{(b)} \quad \text{(b) barrel region } |\eta| < 1.2, \quad \text{(e) endcaps region } |\eta| \geq 1.2$$

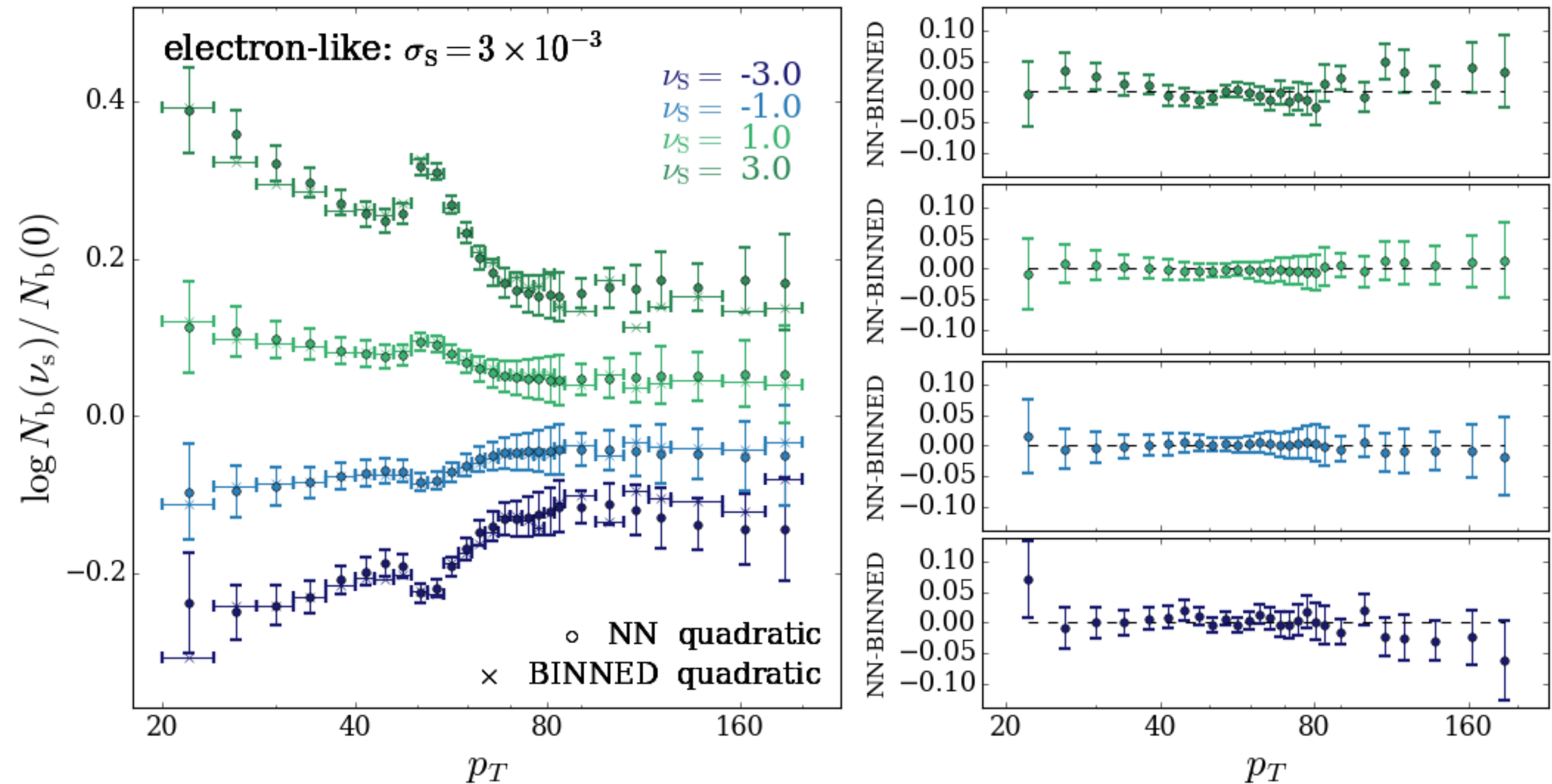
- Muon-like regime: $\sigma_S^{(b)} = 0.05\%$, $\sigma_S^{(e)} = 0.15\%$
- Electron-like regime: $\sigma_S^{(b)} = 0.3\%$, $\sigma_S^{(e)} = 0.9\%$
- Tau-like regime: $\sigma_S^{(b)} = \sigma_S^{(e)} = 3\%$

New Physics Learning Machine (NPLM)

Modelling the family of Reference hypotheses: shape effects

Taylor's expansion learning

Example: result of the NN training vs. Binned approach

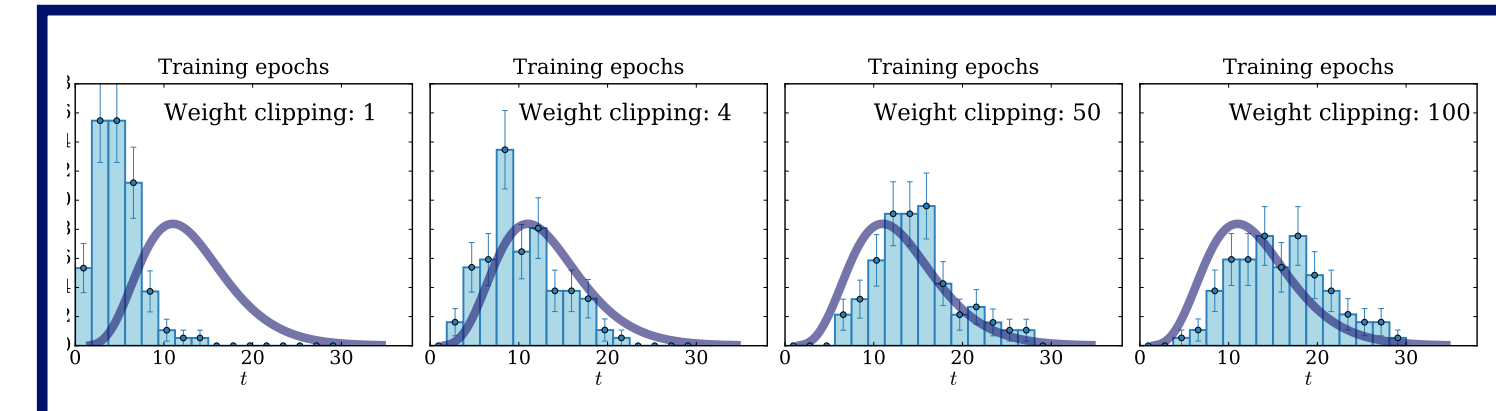


New Physics Learning Machine (NPLM)

Final procedure in steps:

1. NN MODEL REGULARIZATION:

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



2. NUISANCE TAYLOR'S EXPANSION LEARNING:

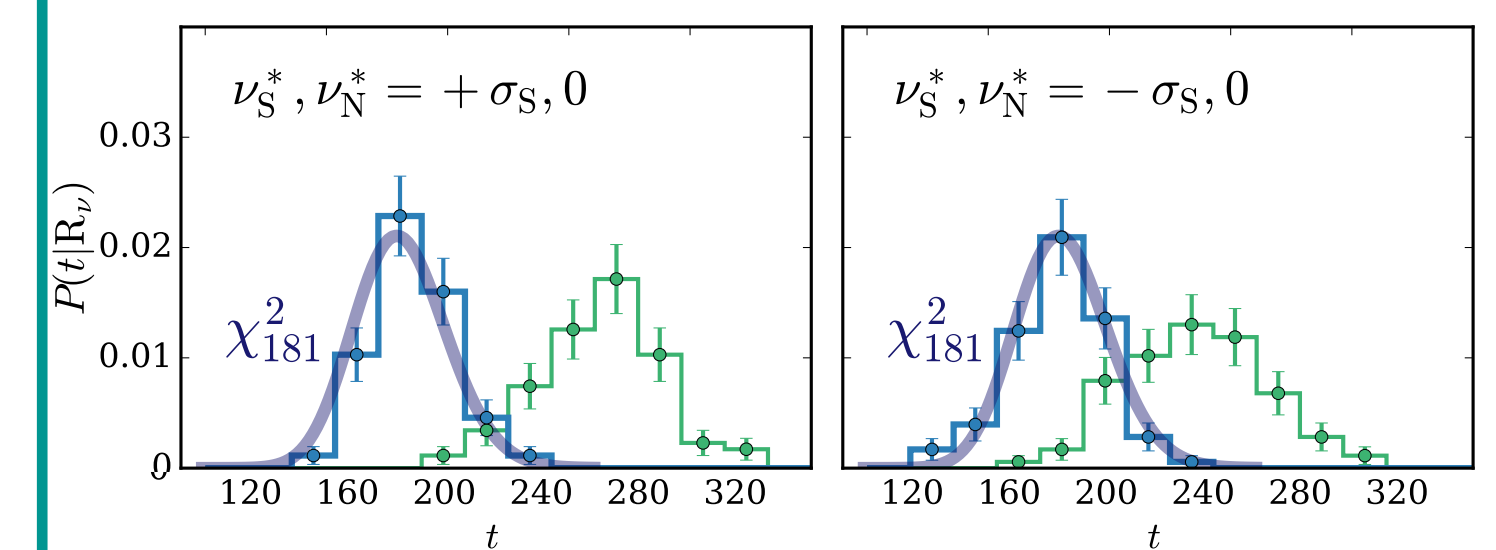
modelling $\hat{r}(x; \nu) = \exp \left[\hat{\delta}_1(x) \nu + \hat{\delta}_2(x) \nu^2 + \dots \right]$;

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

3. VALIDATION:

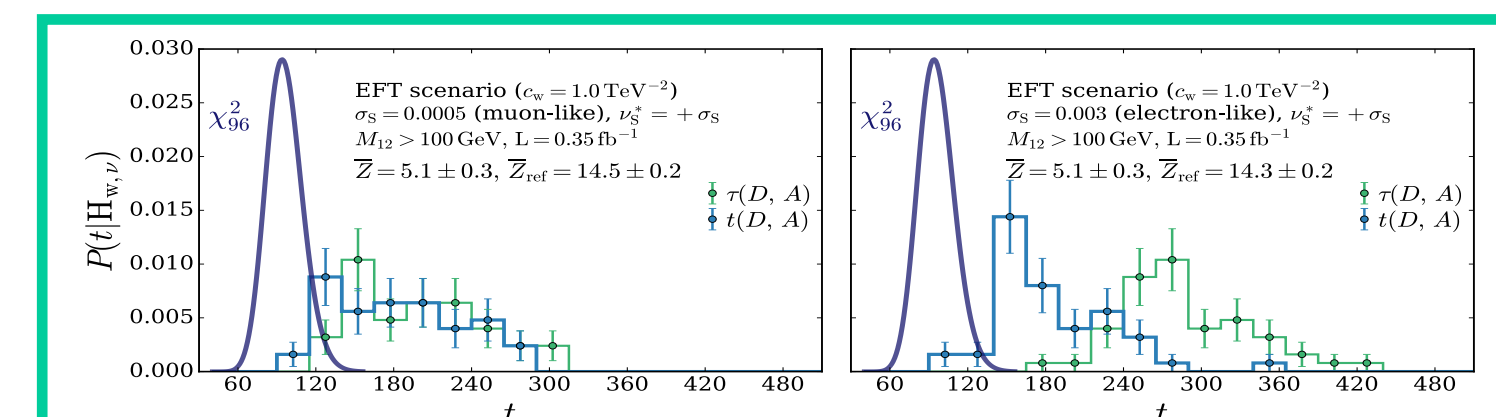
$$\mathcal{D} \sim R_{\nu^*}, \quad \nu^* = \pm \sigma_\nu$$

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



4. TESTING THE DATA:

running the procedure on real data.



Two-body final state

BSM scenarios

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:
 $M_{12} > 100 \text{ GeV}, L = 0.35 \text{ fb}^{-1}, N(S) = 120$
 - Tau-like regime:
 $M_{12} > 120 \text{ GeV}, L = 1.1 \text{ fb}^{-1}, N(S) = 210$

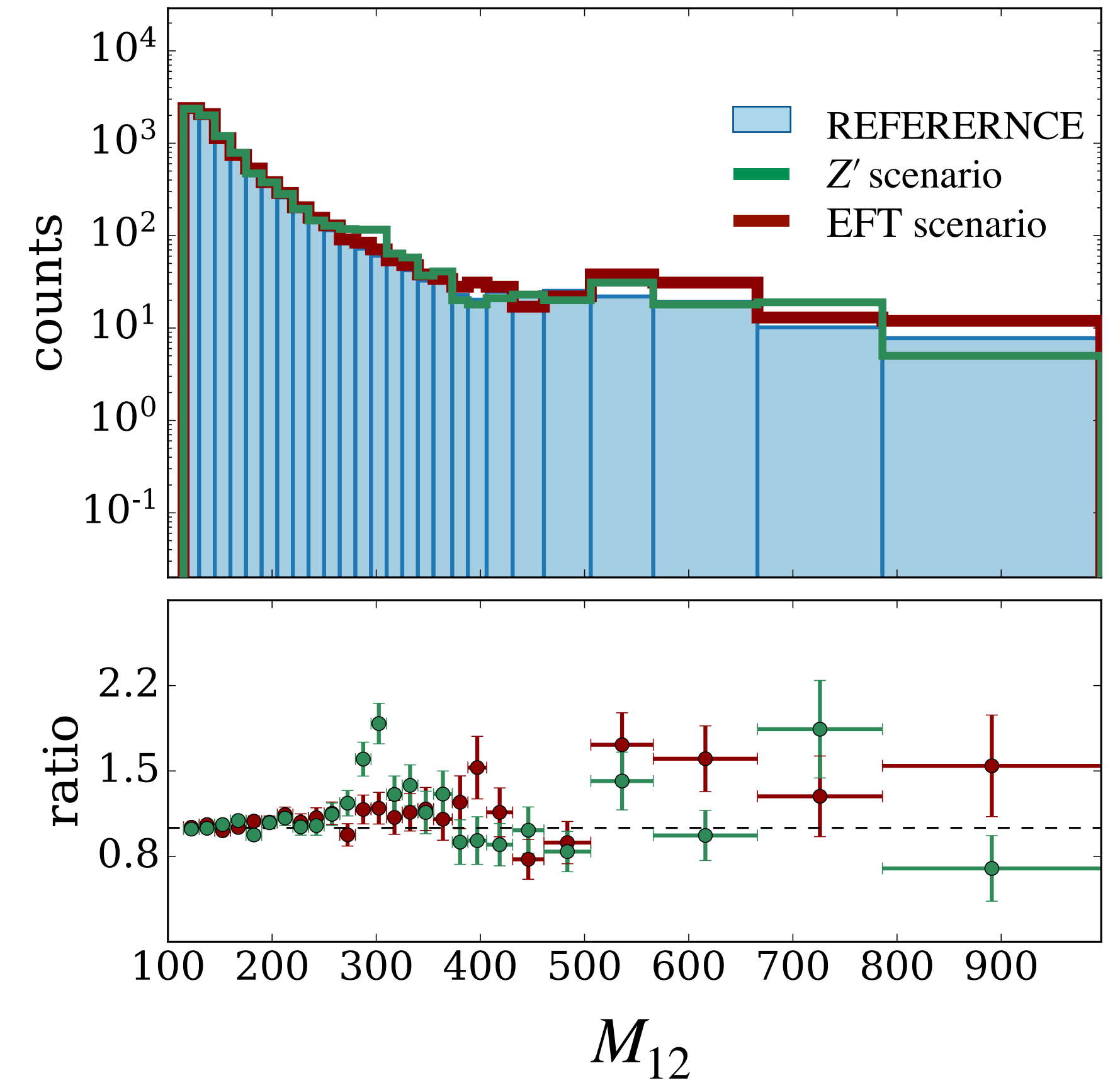
Non resonant excess in the tail of the two-body invariant mass

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

$$\frac{c_W}{\Lambda} J_{L\mu}^a J_{La}^\mu$$

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

Example:
Tau-like regime



NOTE:

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

Two-body final state

BSM scenarios

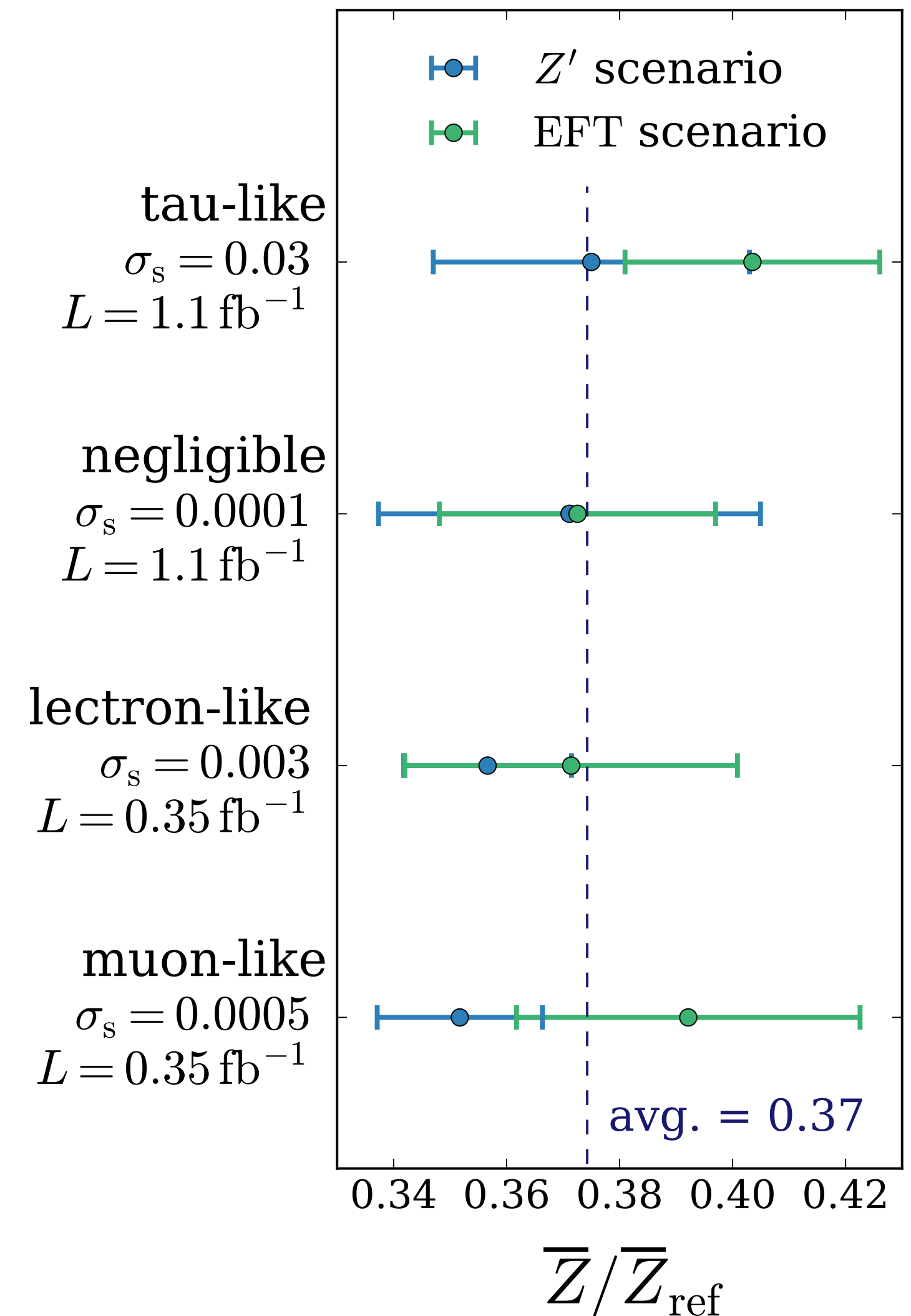
Summary of the results:

- Comparable performances in the resonant and non-resonant scenarios:
 - NPLM is **simultaneously sensitive to any source 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 ($\bar{Z}/\bar{Z}_{\text{ref}} = 0.37$);

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

- \bar{Z} : Z-score from NPLM

- \bar{Z}_{ref} : Z-score from a model-dependent (optimized) test statistics



Two-body final state

Signal reconstruction with the NN:

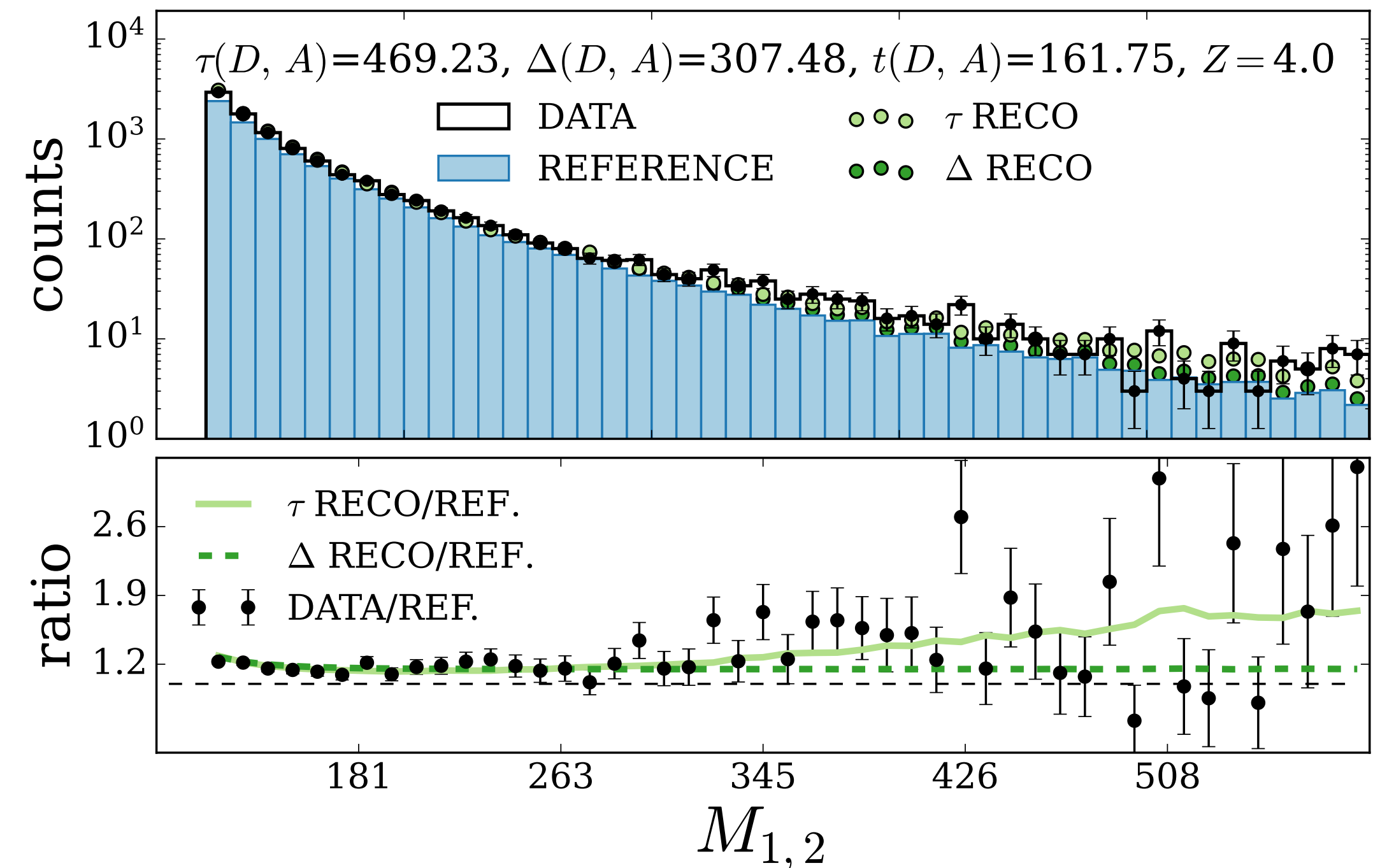
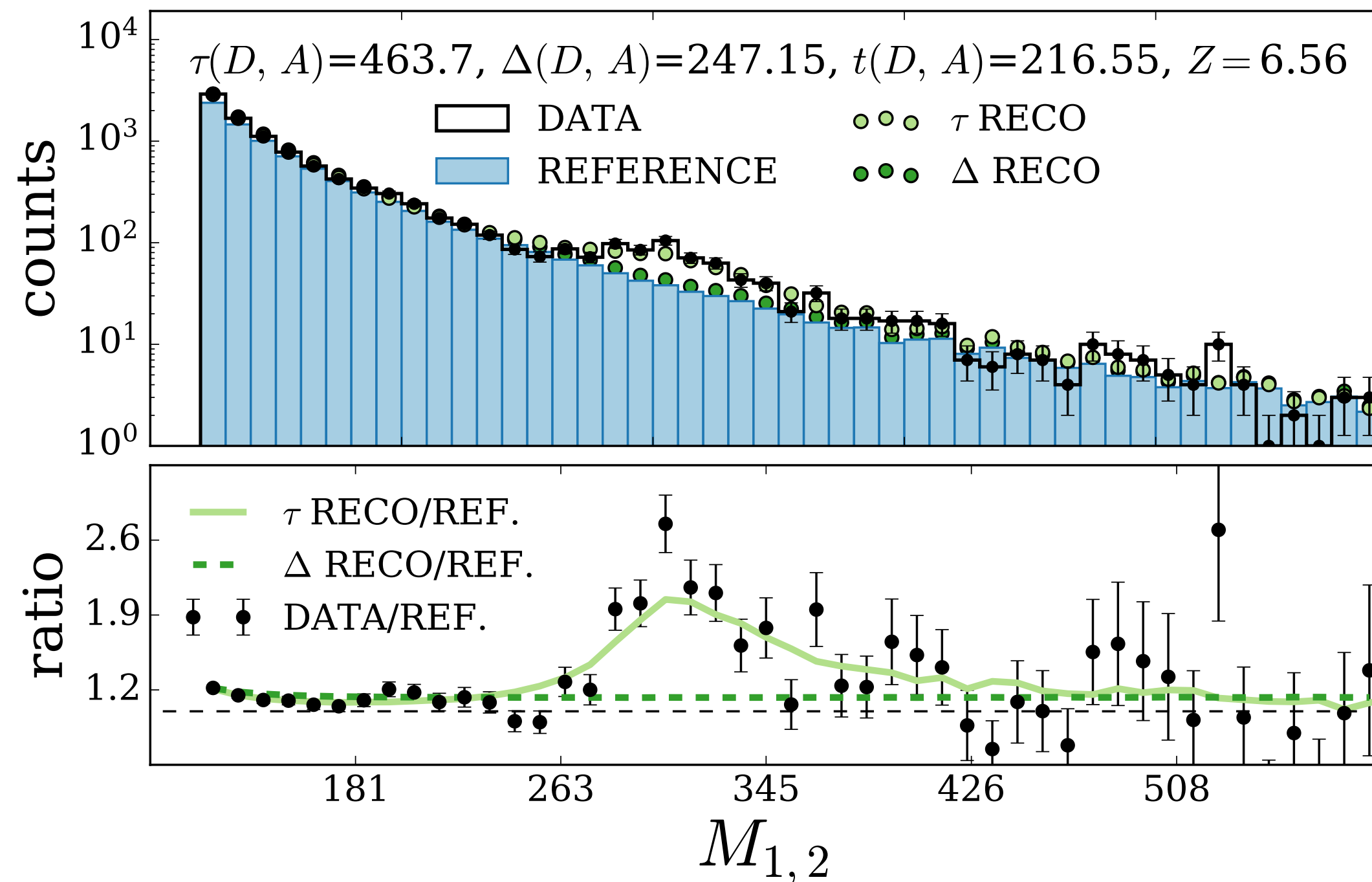
Architecture: [5-5-5-5-1] (96 dof), weigh clipping 2.15, L = 240 fb⁻¹

$$\tau \text{ reconstruction: } n(x | H_{\hat{w}, \hat{v}}) = n(x | R_0) \frac{n(x | R_{\hat{v}})}{n(x | R_0)} e^{f(x; \hat{w})}$$

$$\Delta \text{ reconstruction: } n(x | R_{\hat{v}})$$

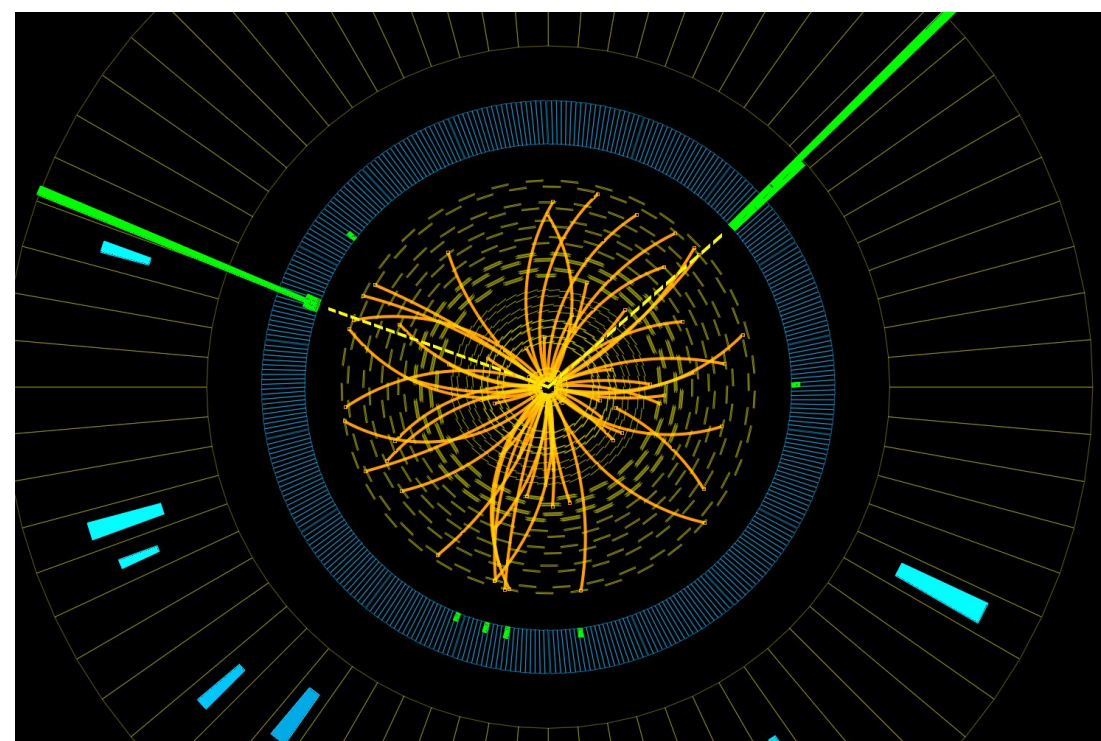
NOTE:

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

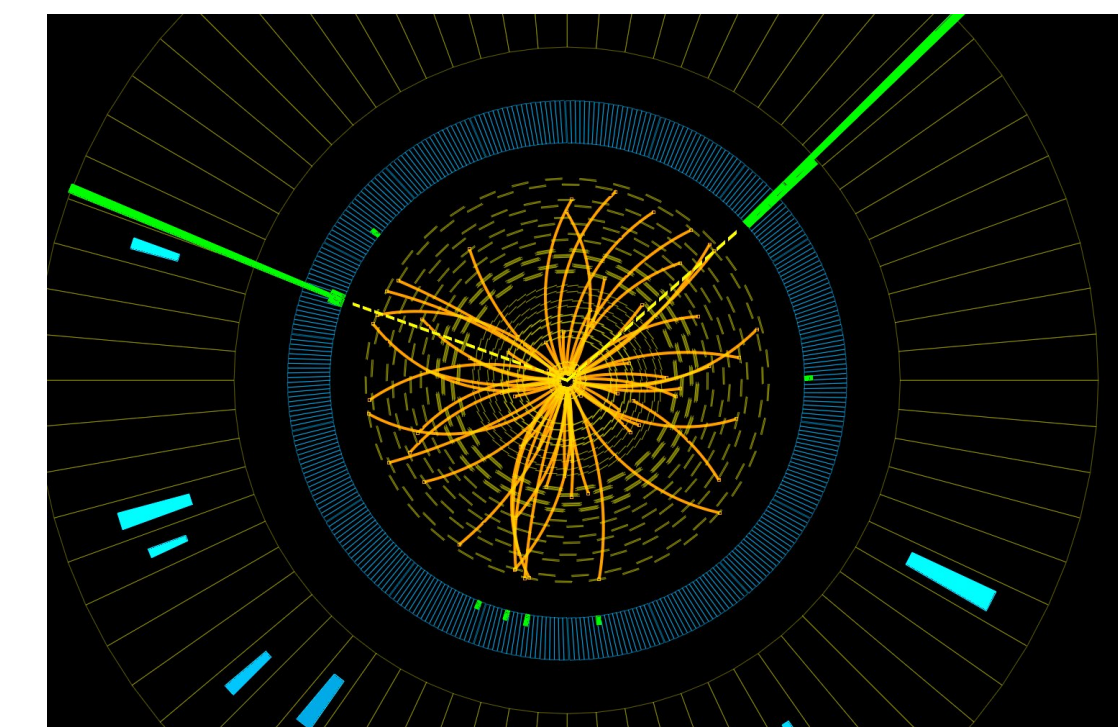
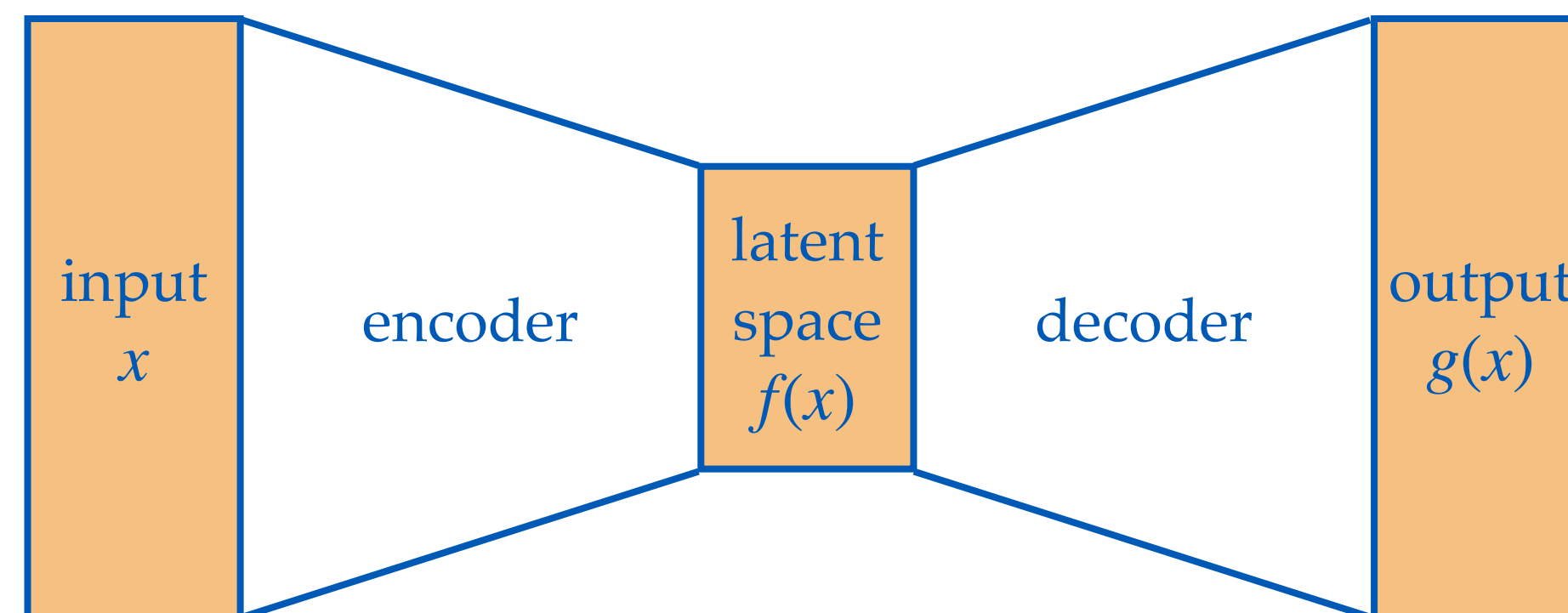


Model-independent approaches

Anomaly detectors: autoencoders



Collider data representation

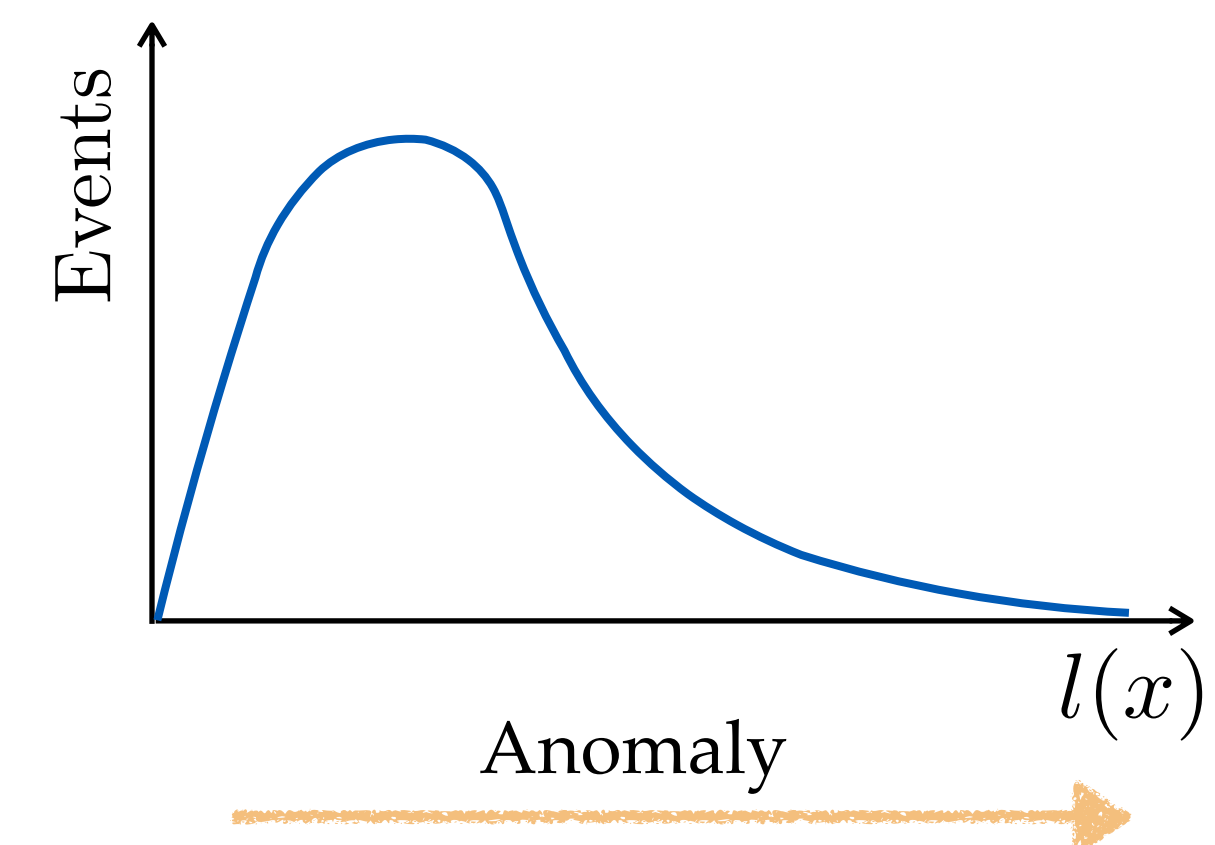


Collider data representation

Reconstruction loss: $L_{\text{reco}} = \sum_x l(x, g(x))$

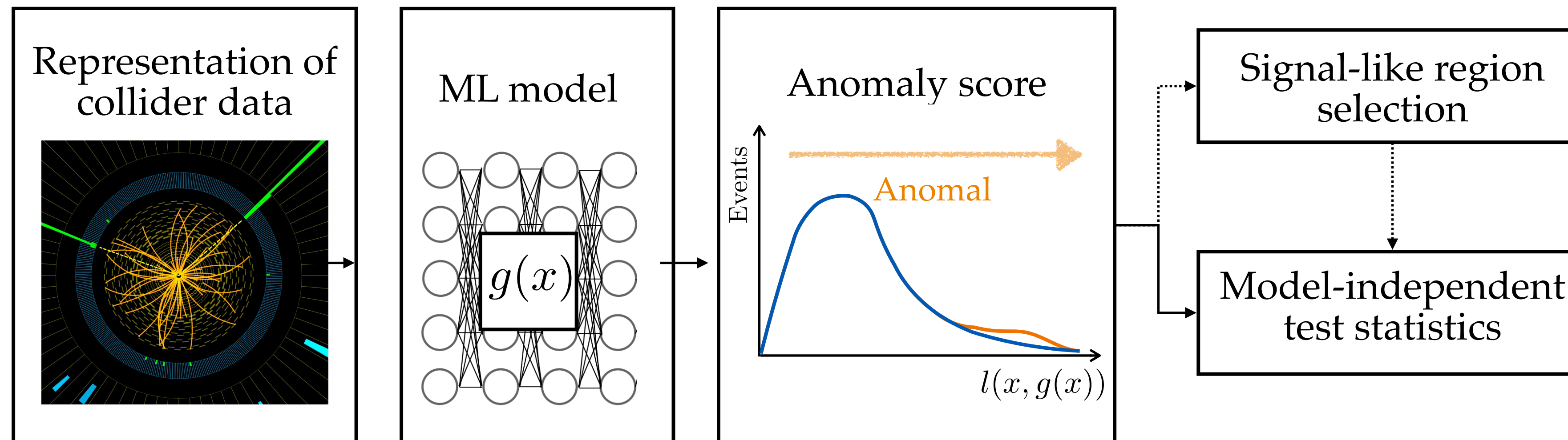
Summary statistics: $l(x, g(x))$

Test statistics



ML Tools for model-independent analyses

Anomaly detectors:

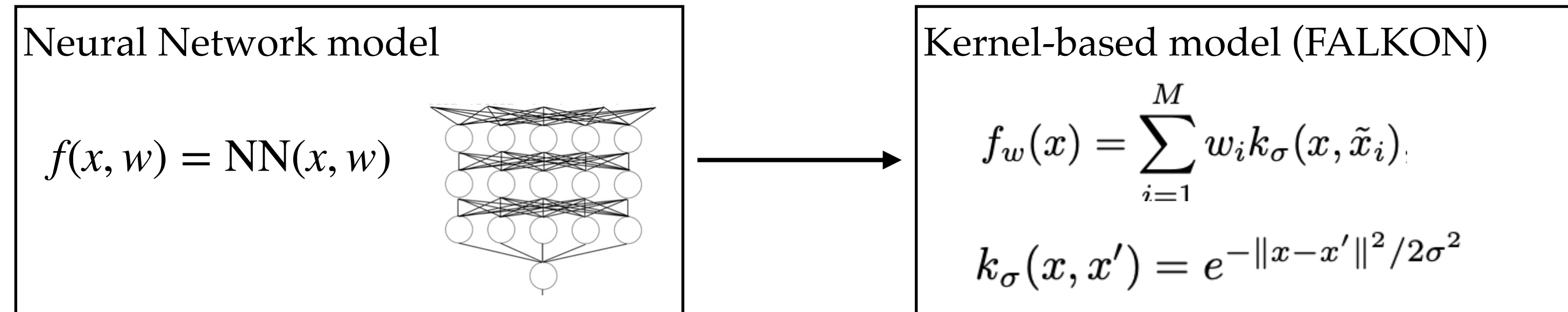


Some examples:

- CWoLa Hunting [[Phys. Rev. Lett. 121 \(2018\), no. 24, 241803](#) , [Phys. Rev. D 99 \(2019\), no. 1, 014038](#)]
- Tag N' Train (TNT) [[JHEP 01 \(2021\) 153](#)]
- Classifying Anomalies THrough Outer Density Estimation (CATHODE) [[arXiv:2109.00546](#)]
- Variational Autoencoders [[arXiv:2110.08508](#)]

Training speed up: kernel methods

- NN model replaced with kernels:



- NPLM loss replaced with a weighted cross entropy loss

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

Weighted cross entropy:

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

Regularization term:

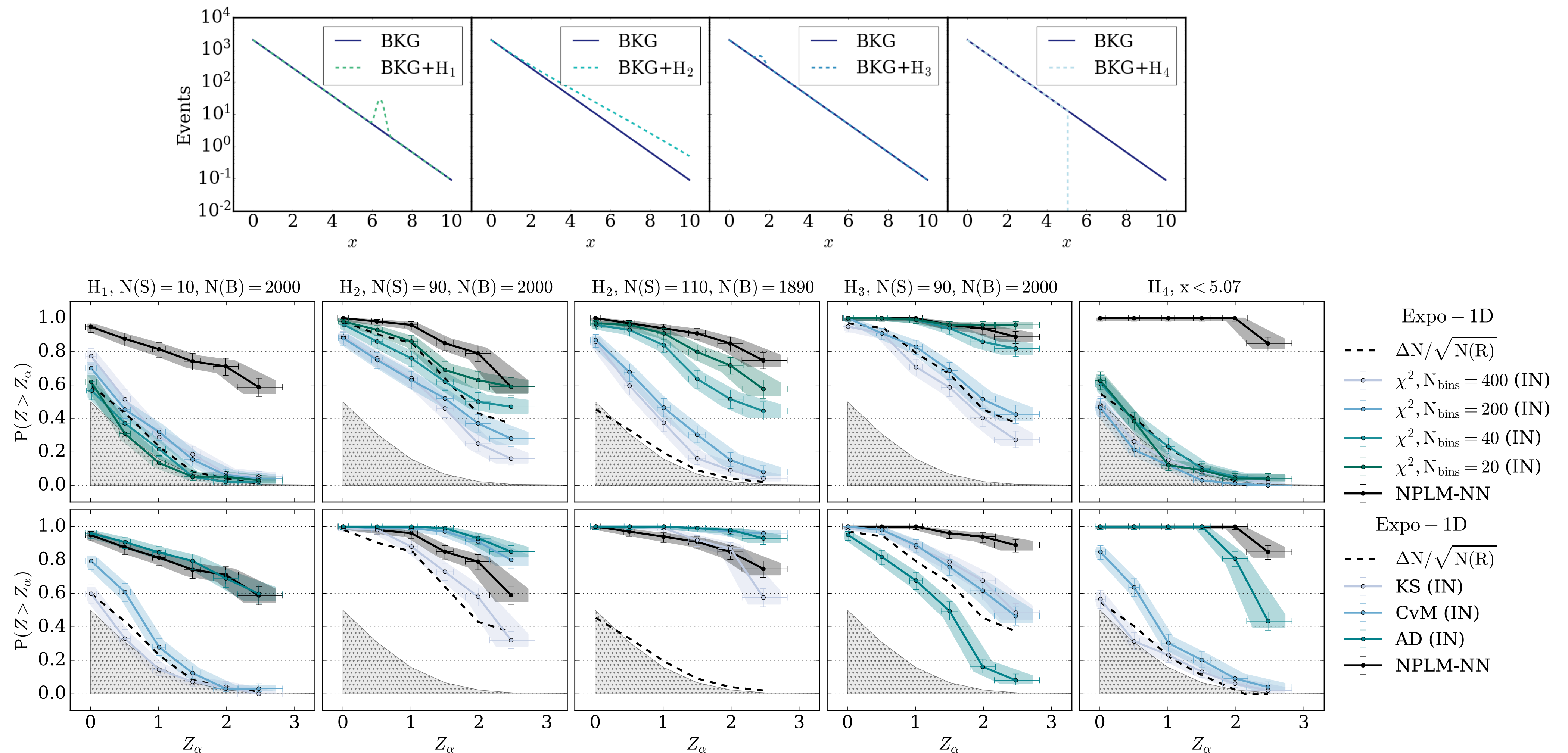
$$R(f_w) = \sum_{ij} w_i w_j k_\sigma(x_i, x_j)$$

This is also a smoothness requirement!

- Different heuristic to select the model hyperparameters (M, σ, λ)
- Online training on **GPU**: drastic drop of training time!
- (Systematic uncertainties not yet implemented)

New Physics Learning Machine (NPLM)

Sensitivity and comparison



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

Latest version
Released: Feb 1, 2022

pip install NPLM

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](#))

```

Reconstruction plot

In [62]:
REF = feature[target[:, 0]==0]
DATA = feature[target[:, 0]==1]
weight = target[:, 1]

weight_REF = weight[target[:, 0]==0]
weight_DATA = weight[target[:, 0]==1]
output_delta_ref = delta.predict(REF)
output_tau_ref = tau.predict(REF)

In [ ]:
plot_reconstruction(df=BSMdf, data=DATA, weight_data=weight_DATA, ref=REF, weight_ref=weight_REF,
tau_OBS=tau_OBS, output_tau_ref=output_tau_ref,
feature_labels=feature_labels, bins_code=bins_code, xlabel_code=xlabel_code, ymax_code=ymax_code,
delta_OBS=delta_OBS, output_delta_ref=output_delta_ref,
save=False, save_path='', file_name='')
    
```

