PHYSTAT-2samples 2023
June 1, 2023

Goodness of fit by Neyman-Pearson testing

G. Grosso[1], M. Letizia, A. Wulzer, M. Pierini
[1]University and INFN of Padova

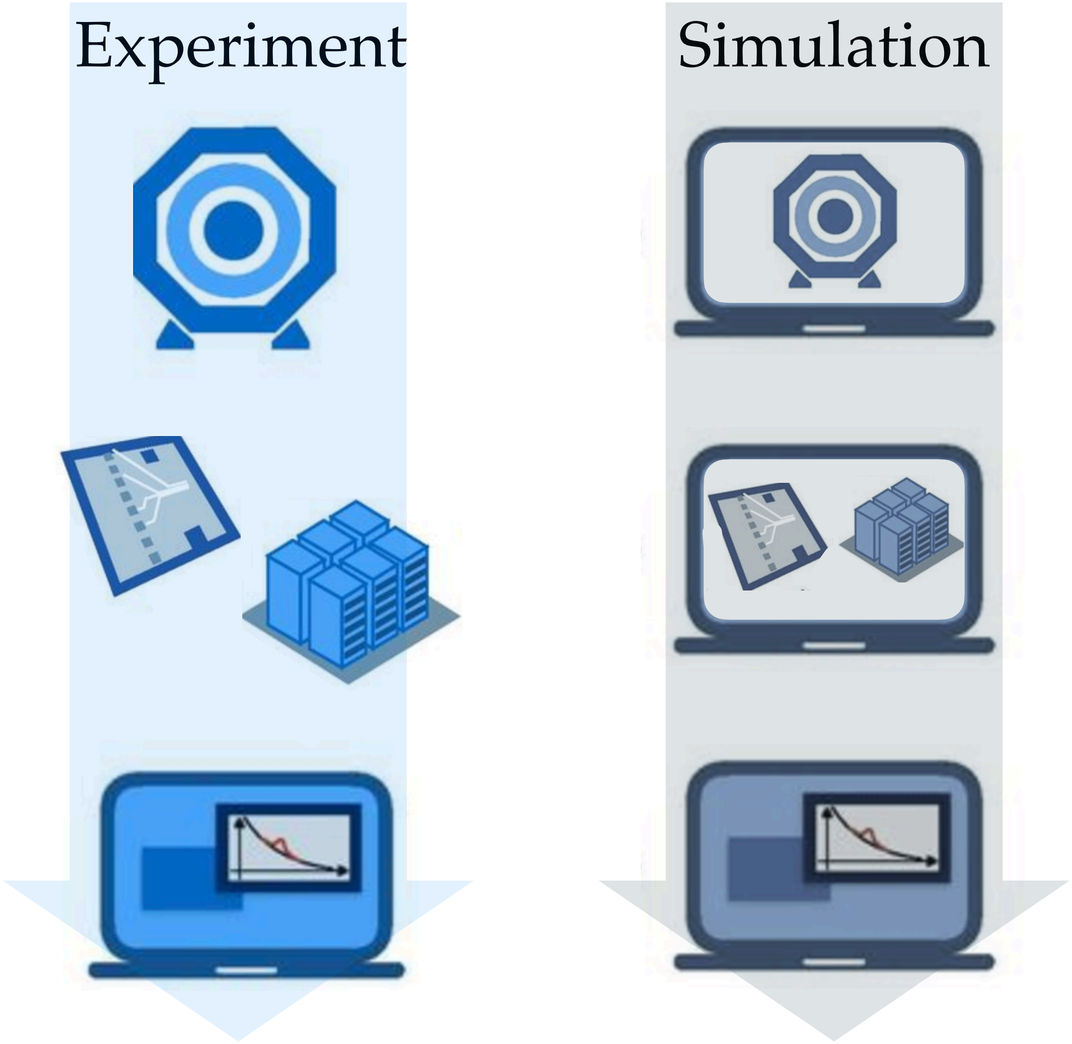
Mainly based on:
!! NEW !! 2305.14137 (Grosso, Letizia, Pierini, Wulzer)
!! NEW !! arXiv:2303.05413 (Lai, Grosso,, Letizia et al.)
Eur. Phys. J. C 82, 879 (2022) (Letizia, Grosso, Pierini Wulzer et al.)
Eur. Phys. J. C 81, 89 (2021) (d'Agnolo, Grosso, Pierini, Wulzer, Zanetti)
Phys. Rev. D 99, 015014 (d'Agnolo, Wulzer)

How well do we understand the LHC data?

Detector malfunctioning ?

Reconstruction Bugs ?

Unexpected Physics Beyond the Standard Model ?



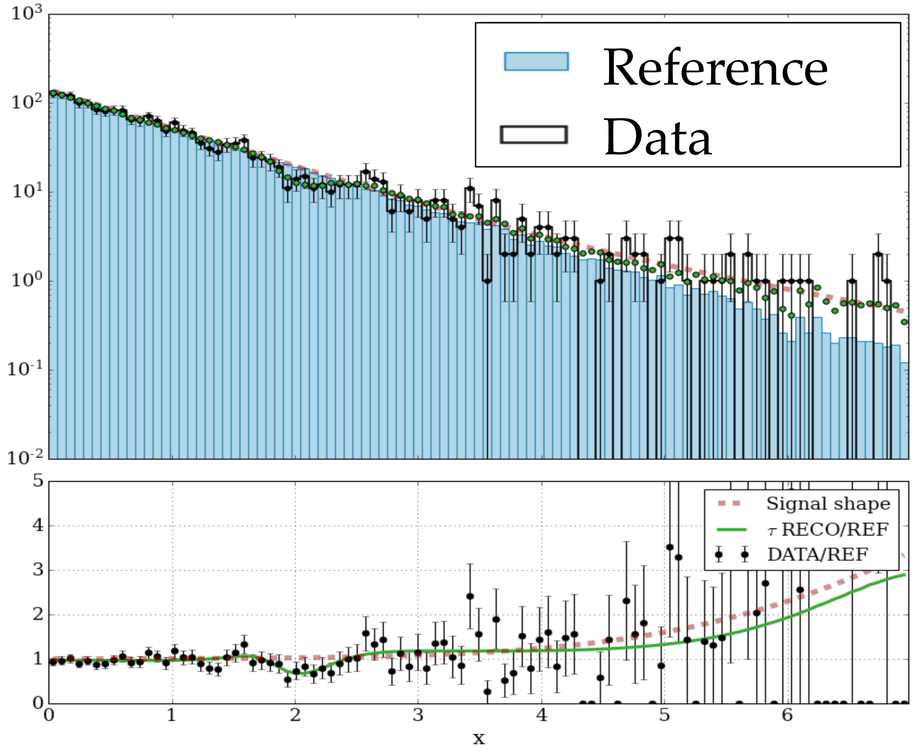
Can we trust our simulations ?

See Raghav's talk

Are two simulators equivalent ?

How well do we understand the LHC data?

These are problems of **Goodness of Fit**:



- **Data**: experimental measurements of the natural process $\{x_i\}_{i=1}^{N_D}$

- **Reference model**: expected nominal behaviour of the data

  (Standard Model, normal operating condition of a detector…).

  Most of the times not known in close form: Reference sample $\{y_i\}_{i=1}^{N_R}$ → **2sample test**

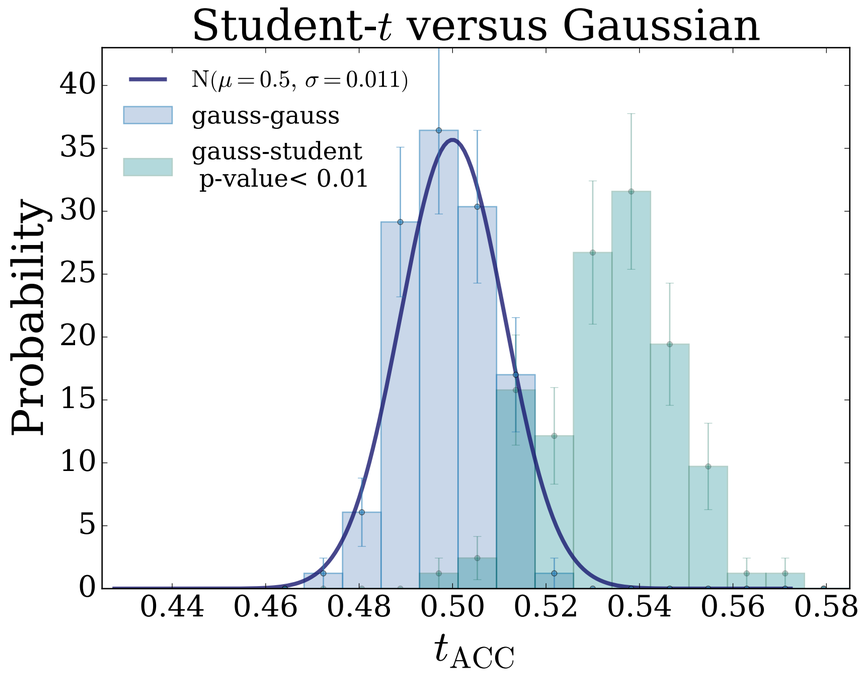
At LHC: multi-dimensional, large statistics samples → **Machine Learning approaches**

Classifier-based GOF approaches

Friedman (2003): Training a classifier to tell apart two samples (data and reference samples).

Main features:

- Balanced problem ($N_D = N_R$)
- Train-test split (out of sample evaluation)
- Test:
  - Classification metrics: accuracy, AUC (Charkavarti et al. (2021), Lopez et al. (2017))
  - Standard 1D GOF on the classifier output: classifier for dimensionality reduction (Friedman (2003))
- Calibration and p-value:
  - Toy experiments (training a new model each time)
  - The single value of test statistic is not sufficient without a fair comparison to the test statistic distribution under the null

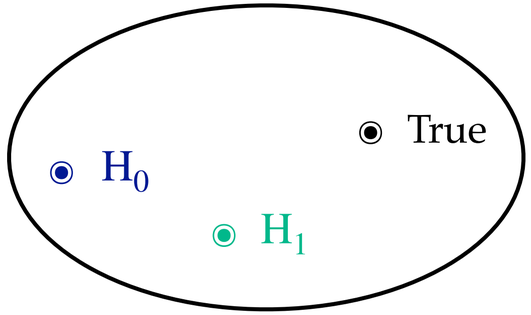


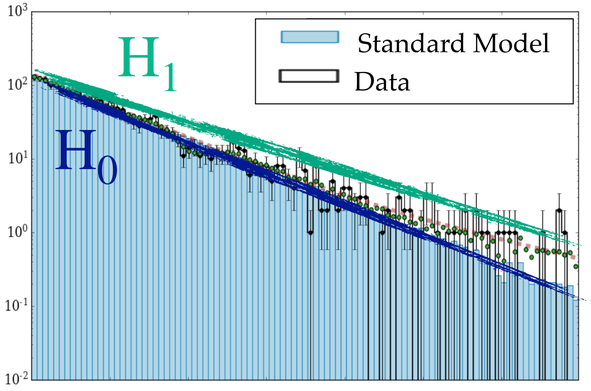
Neyman-Pearson testing

Comparing the null hypothesis (Reference) with an alternative.
Traditionally employed in BSM searches at the LHC.

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

- Choosing an alternative is mandatory
- The alternative defines the landscape signals that the test is sensitive to.
- Sensitivity (and optimality) are guaranteed (according to Neyman and Pearson) only if the data do follow the chosen alternative



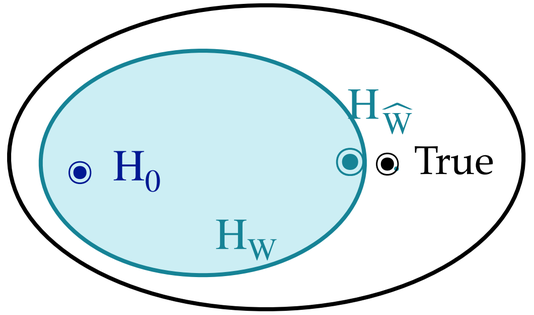


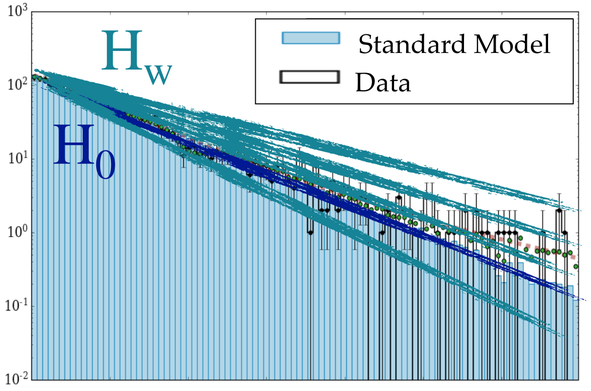
Neyman-Pearson testing

Comparing the null hypothesis (Reference) with an alternative.
Release the assumptions on the alternative

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

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



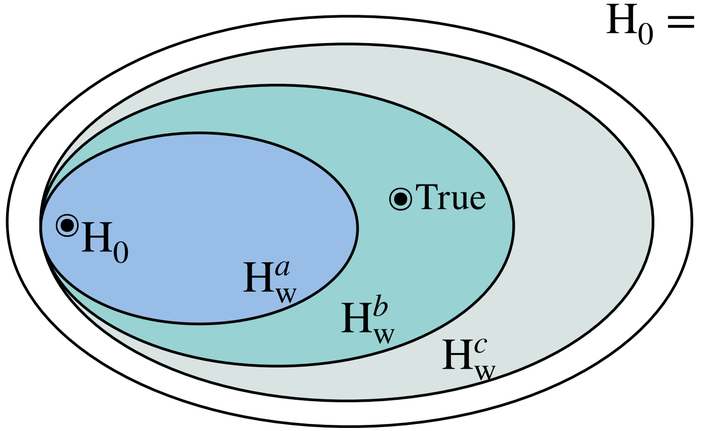


Neyman-Pearson testing

The problem of sensitivity

$$D \sim \frac{1}{2}\text{Gauss}(\mu = -1, \sigma = 1) + \frac{1}{2}\text{Gauss}(\mu = 1, \sigma = 1)$$

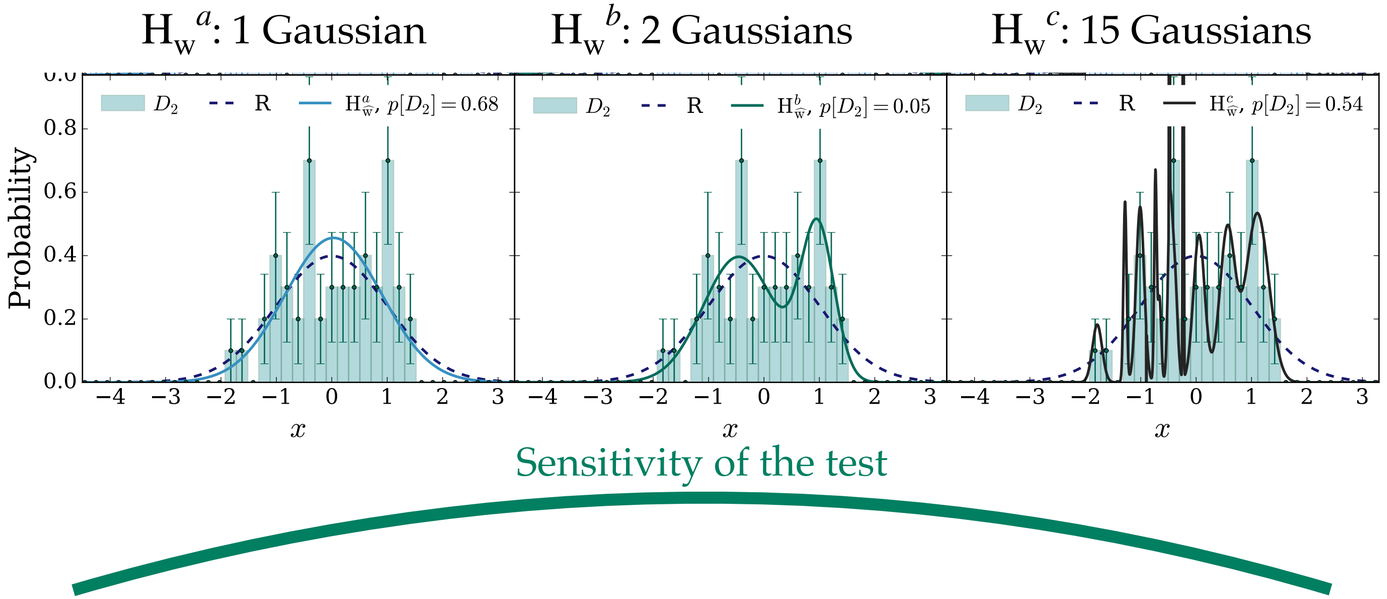
$$\text{H}_0 = \text{Gauss}(\mu = 0, \sigma = 1)$$



The specificity of the alternative determines the sensitivity of the test.

- Too complex models are over-sensitive to statistical fluctuations, independently on the nature of the data
- Too simple models may not contain an element that approximate the True hypothesis well enough

A successful GoF strategy based on Neyman–Pearson testing should balance between flexibility and regularisation.



Neyman-Pearson testing

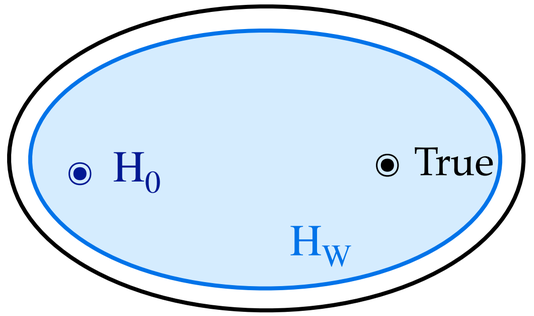
Comparing the null hypothesis (Reference) with an alternative.
Release the assumptions on the alternative

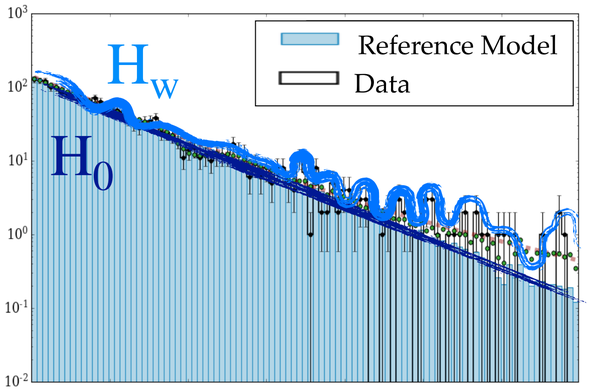
$$t(\mathcal{D}) = \max_{\mathbf{w}} \left[2 \log \frac{\mathcal{L}(\mathcal{D}\,|\mathrm{H}_{\mathbf{w}})}{\mathcal{L}(\mathcal{D}\,|\mathrm{H}_{\mathbf{0}})} \right]$$

NPLM algorithm

Universal approximator (NN, kernel methods, …)

$$n(x|\mathrm{H}_{\mathbf{w}}) = e^{f(x;\mathbf{w})}\, n(x|\mathrm{R}_{\mathbf{0}})$$





Alternative ≡ Data

(The *saturated model* proposed by Baker and Cousins — Nucl.Instrum.Meth. 221 (1984) — is a binned version of this approach)

New Physics Learning Machine (NPLM)

Maximum Likelihood from Minimal Loss

Test statistic

$$\bar{t}(\mathcal{D}) = 2 \max_{\mathbf{w}} \log \left[\frac{\mathcal{L}(\mathcal{D} \mid \mathrm{H}_{\mathbf{w}})}{\mathcal{L}(\mathcal{D} \mid \mathrm{R}_{\mathbf{0}})} \right] = 2 \max_{\mathbf{w}} \left\{ \log \left[\frac{e^{-N(\mathbf{w})}}{e^{-N(\mathrm{R})}} \prod_{i=1}^{\mathrm{N_D}} \frac{n(x_i|\mathbf{w})}{n(x_i|\mathrm{R})} \right] \right\}$$

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

Loss function

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

w: trainable parameters on the NN model

D: data sample

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

Assumptions:

- $\mathrm{N_R} \gg \mathrm{N_D}$ 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

New Physics Learning Machine (NPLM)

Likelihood ratio from Binary Cross Entropy

Test statistic

$$\bar{t}(\mathcal{D}) = 2 \max_{\mathbf{w}} \log \left[\frac{\mathcal{L}(\mathcal{D} \mid \mathrm{H}_{\mathbf{w}})}{\mathcal{L}(\mathcal{D} \mid \mathrm{R}_{\mathbf{0}})} \right] = 2 \max_{\mathbf{w}} \left\{ \log \left[\frac{e^{-N(\mathbf{w})}}{e^{-N(\mathrm{R})}} \prod_{i=1}^{\mathrm{N_D}} \frac{n(x_i|\mathbf{w})}{n(x_i|\mathrm{R})} \right] \right\}$$

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

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

Loss function

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

w: trainable parameters on the NN model

D: data sample

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

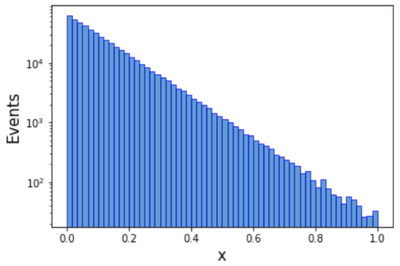
Assumptions:

- $\mathrm{N_R} \gg \mathrm{N_D}$ 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

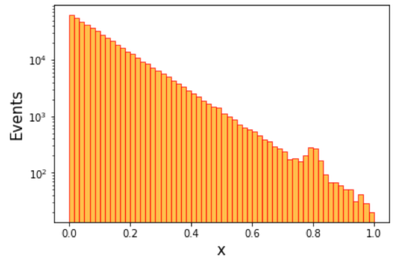
New Physics Learning Machine (NPLM)

INPUT

Reference sample (R)
label=0



Data sample (D)
label=1



MODEL $f(\cdot, \mathbf{w})$

Kernel method (FALKON)

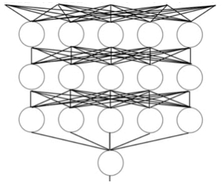
$$f_w(x) = \sum_{i=1}^{M} w_i k_\sigma(x, \tilde{x}_i)$$

$$k_\sigma(x, x') = e^{-\|x-x'\|^2/2\sigma^2}$$

OR

Neural Network

$$f(x, w) = \mathrm{NN}(x, w)$$

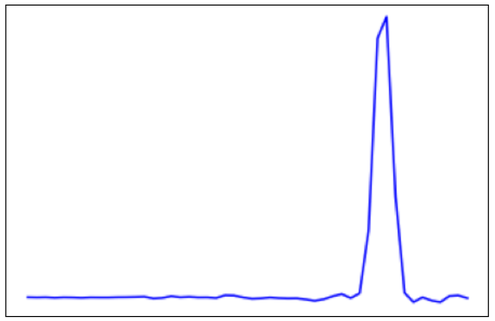


$\mathbf{w}$ —training→ $\hat{\mathbf{w}}$

OUTPUT

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

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



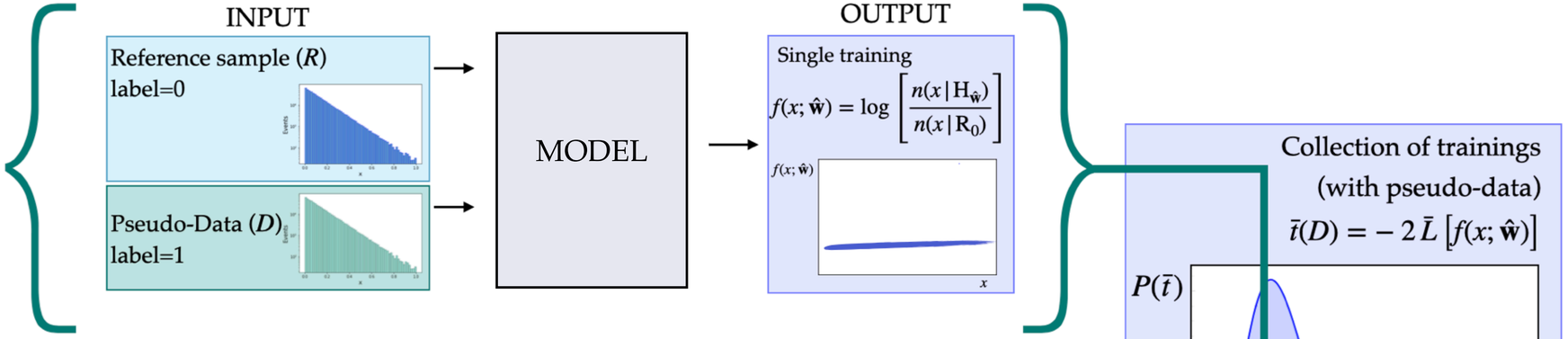
$$\bar{t}(D) = 2\sum_{x\in\mathcal{D}} f_{\mathbf{w}}(x) - 2\sum_{x\in\mathcal{R}} \frac{\mathrm{N(R)}}{\mathrm{N}_{\mathcal{R}}}\left[e^{f(x;\mathbf{w})} - 1\right]$$

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

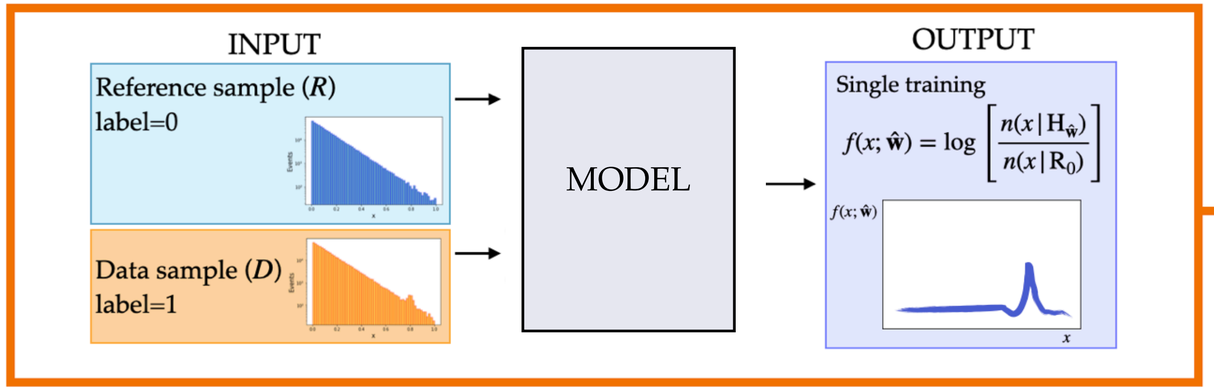
New Physics Learning Machine (NPLM)

Frequentist *p*-value (aka calibration):

1. Run NPLM on toy experiments to simulated the response under the null hypothesis



2. Run NPLM the data of interest and check where the test outcome falls to compute an exclusion *p*-value



New Physics Learning Machine (NPLM)

Controlling type I errors: NN model regularisation

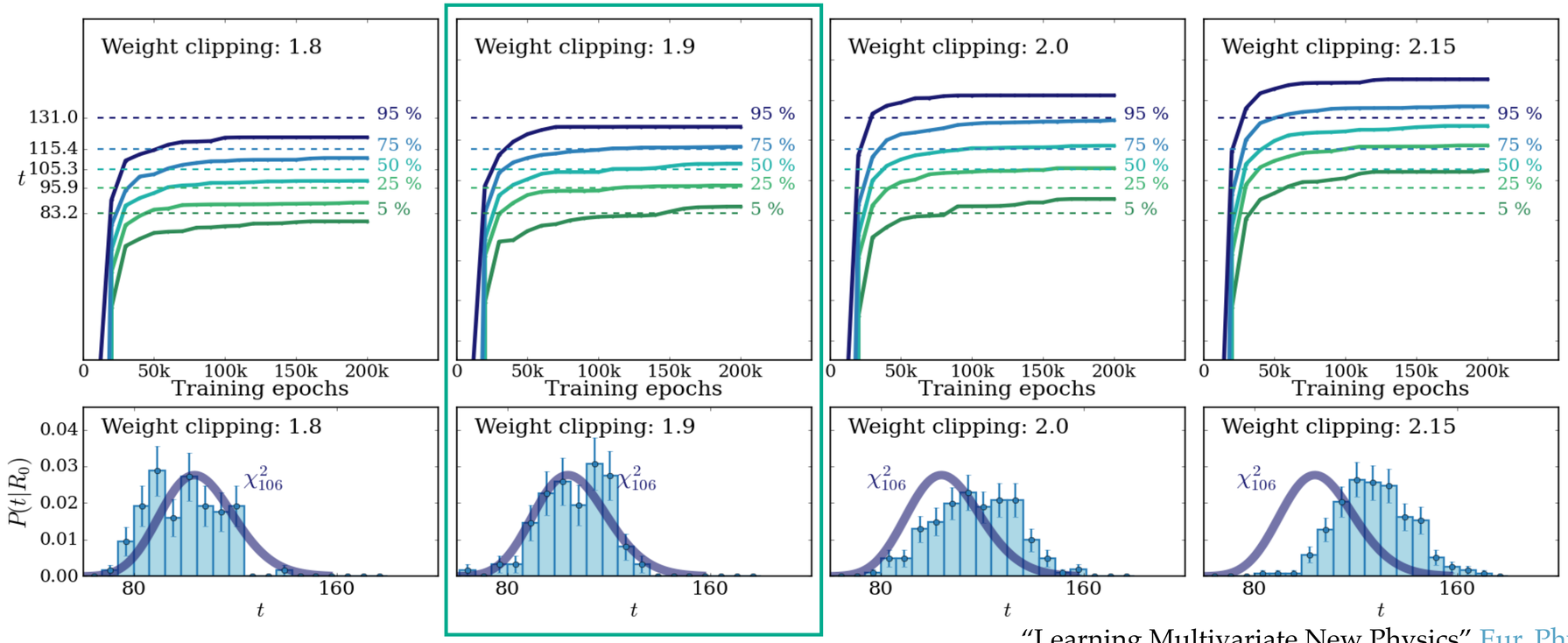
Weight clipping parameter:
Upper boundary to the magnitude that each trainable parameter can assume during the training.

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** $\chi^2_{|\mathbf{w}|}$ 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 $\chi^2_{|\mathbf{w}|}$
- Empirical $\bar{t}$ distribution under R_0
- Target $\chi^2_{|\mathbf{w}|}$



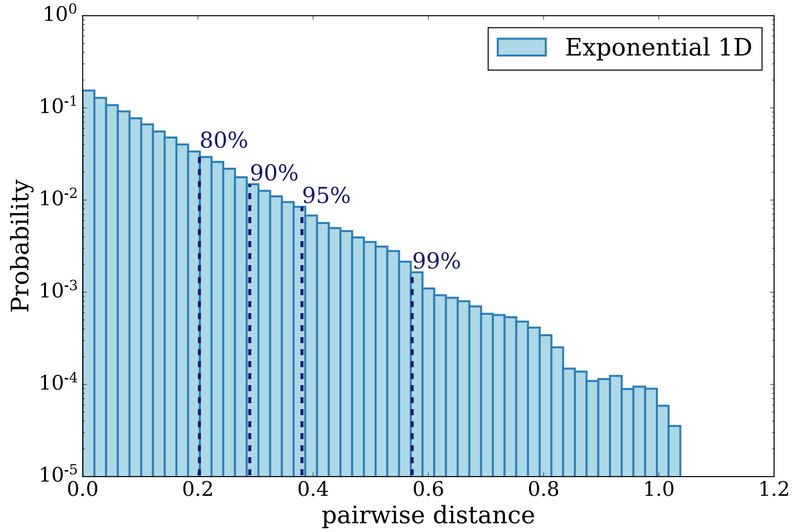
"Learning Multivariate New Physics" Eur. Phys. J. C

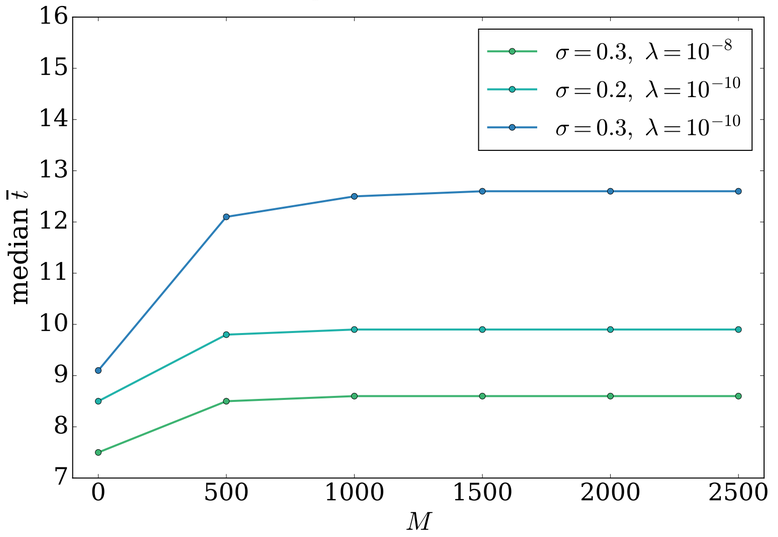
New Physics Learning Machine (NPLM)

Controlling type I errors: Kernel Method regularisation

- χ^2 is found for each choice of (M, σ, λ), provided that $N_R \gg N_D$.

- Number of centres M: controls the expressive power of the model and therefore it should be as large as possible (at least as large as $\sqrt{N}$ to achieve statistically optimal bounds of the training convergence).
- Gaussian width σ: is selected as the 90th percentile of the pairwise distance between reference-distributed data points (after standardisation).
- Regularisation parameter λ: is kept as small as possible while keeping training stable, i.e. avoiding large training times or non-numerical outputs.



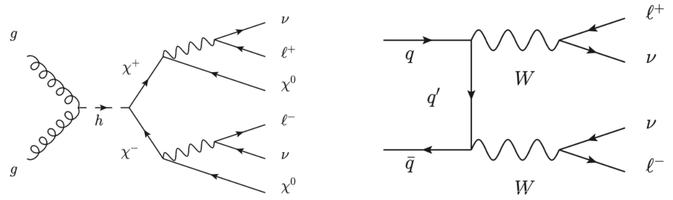


"Learning New Physics Efficiently with non-parametric models" Eur. Phys. J. C

NPLM: NN vs. Kernel methods

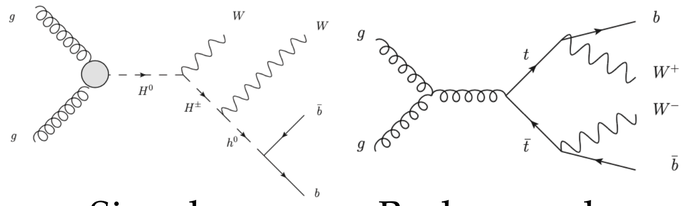
Benchmarks:

- 8D problem: (SUSY dataset*)



Signal Background

- 21D problem: (HIGGS dataset**)

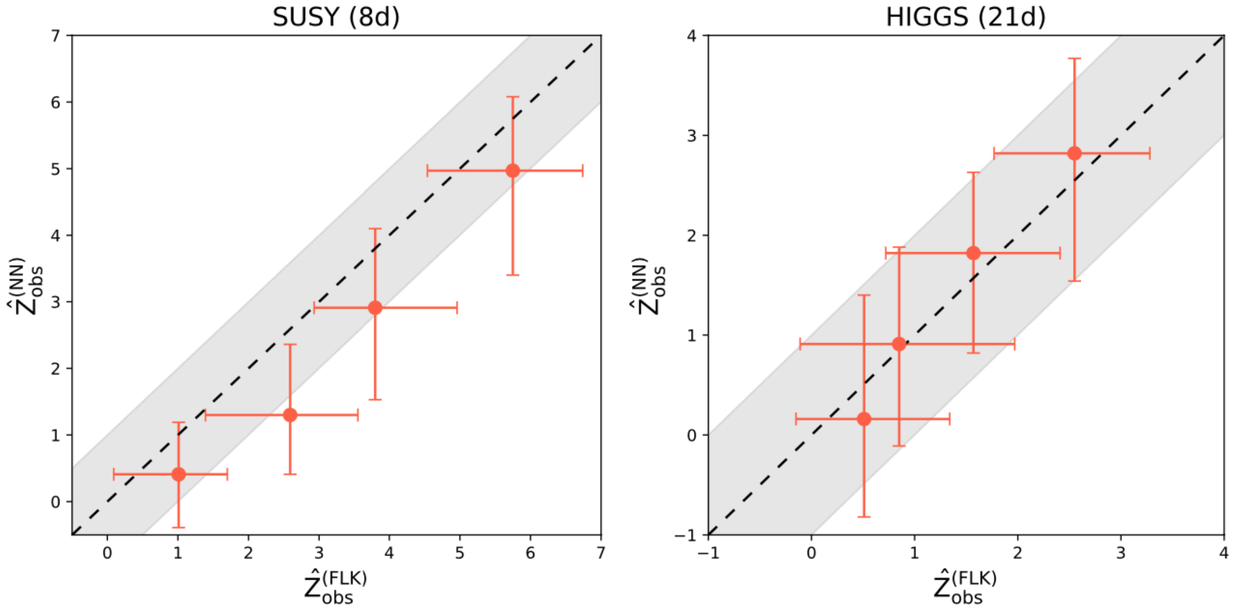


Signal Background

* SUSY dataset: https://archive.ics.uci.edu/ml/datasets/SUSY

** HIGGS dataset: https://archive.ics.uci.edu/ml/datasets/HIGGS

Falkon vs. NN: sensitivity performance



Falkon vs. NN: execution time

Model	SUSY	HIGGS
FLK	**(18.2 ± 1.2) s**	**(22.7 ± 0.4) s**
NN	(73.1 ± 10) h	(112 ± 9) h

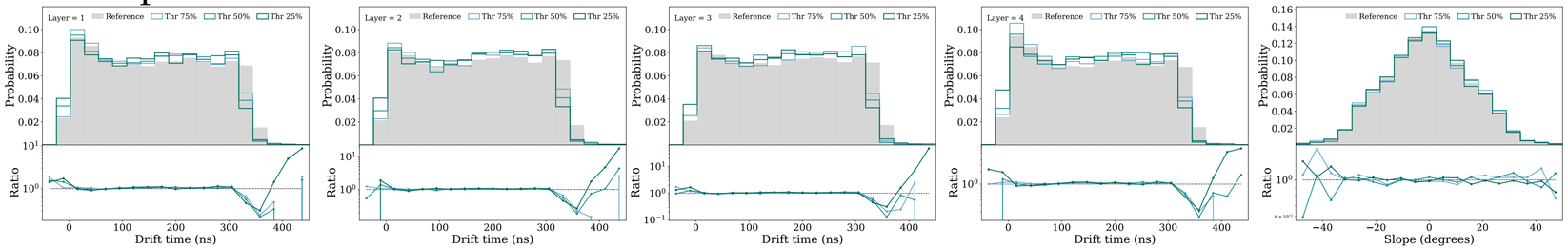
Learning new physics efficiently with nonparametric methods. [Eur. Phys. J. C, 82(10)]

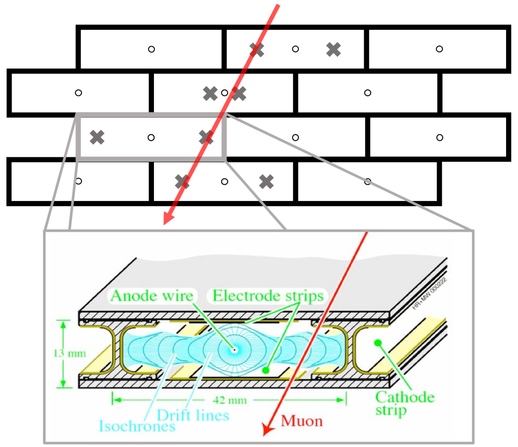
NPLM with Kernel methods

Efficient GOF computation on GPUs with Falkon

Example: Online Data Quality monitoring of a Drift Tube Chamber

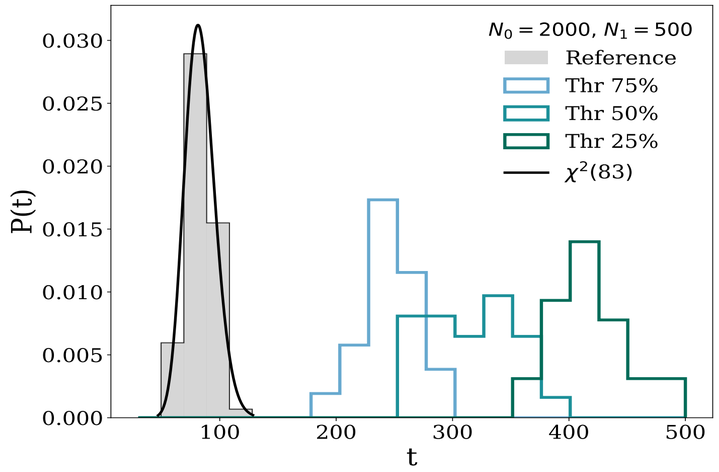
5D input data





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

$M = 2000, \sigma = 4.5, \lambda = 10^{-7}$
$N(D) = 500, N_{ref} = 2000$
Execution time: ~ 0.5 s



More about this in our recent preprint arXiv:2303.05413

Neyman-Pearson testing with NPLM for GOF

Comparing the null hypothesis (Reference) with an alternative.
Release the assumptions on the alternative (**Alternative ≡ Data**)

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

Peculiar aspects due to the nature of the test:

- Unbalanced problem (ideally $N_D \ll N_R$)
- Regularisation scheme to control type I errors (no need for train-test splitting)
- In-sample evaluation of the test

How do we compare to other GOF methods?

NPLM vs. C2ST

Classifier 2 Sample Test (C2ST, Lopez et al. (2017)):

- Train-test split: $R = R_{tr} \cup R_{te}$ and $D = D_{tr} \cup D_{te}$
- Metric:
  - The original C2ST uses the accuracy (ACC)
  - We replace it with a modified version of the balanced accuracy (BAL-ACC) which is sensitive to normalization effects and accounts for unbalanced samples

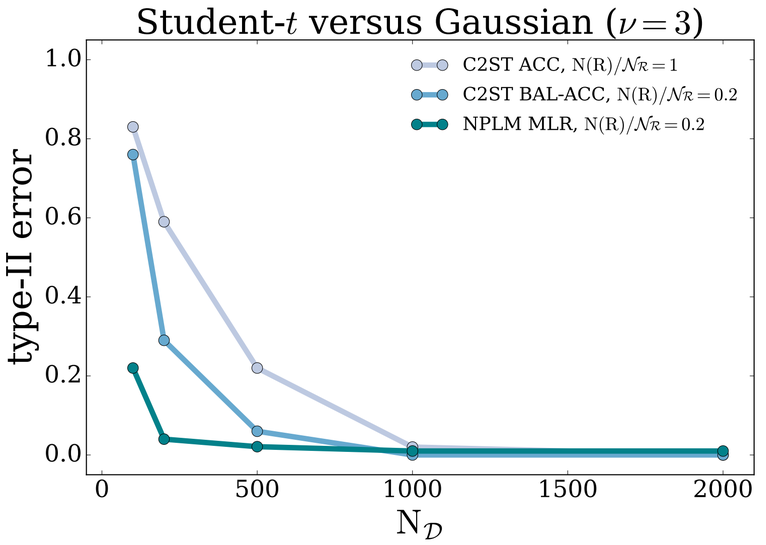
$$t'_{\mathrm{BACC}} = \frac{2}{\mathrm{N(R)} + \mathrm{N}_{\mathcal{D}}} \left[\frac{\mathrm{N(R)}}{\mathrm{N}_{\mathcal{R}}} \sum_{x \in \mathcal{R}_{te}} \mathbb{I}[c_{\hat{\mathbf{w}}}(x) < 1/2] + \sum_{x \in \mathcal{D}_{te}} \mathbb{I}[c_{\hat{\mathbf{w}}}(x) > 1/2] \right]$$

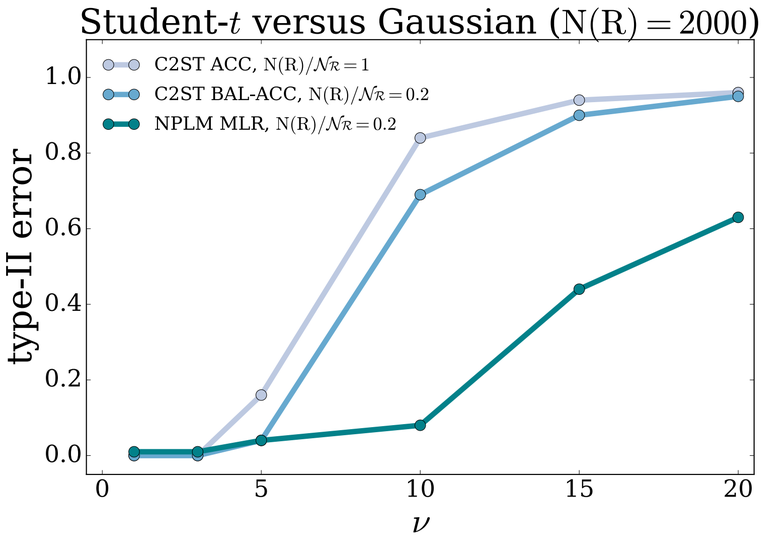
- Out of sample test statistic
- p-value from empirical test statistic distribution under the null

---

Comments:

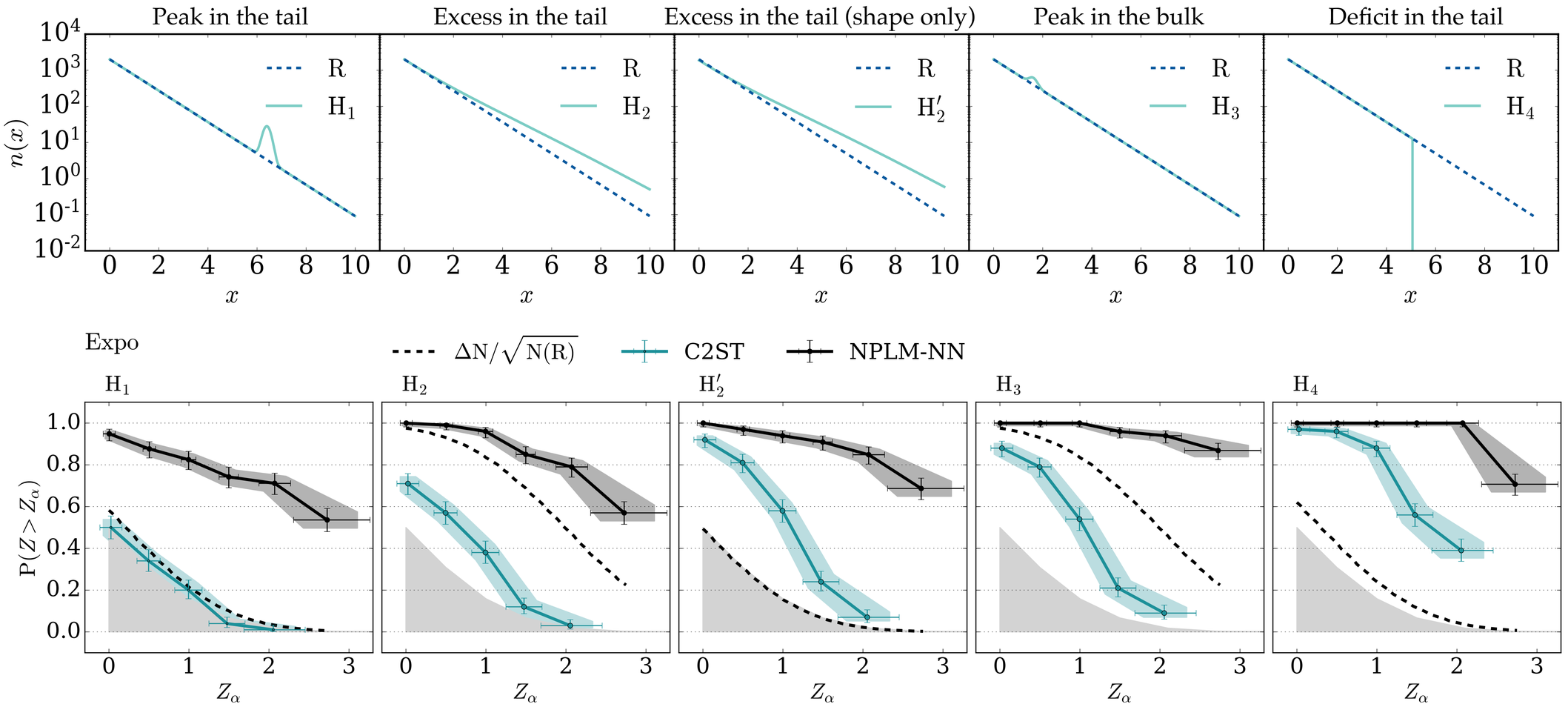
- Unbalanced problem ($\mathrm{N_R} > \mathrm{N_D}$): performance improvement
- additional improvement NPLM wrt C2ST, why?
  - train-test splitting vs. in-sample ?
  - ACC vs. LRT ?





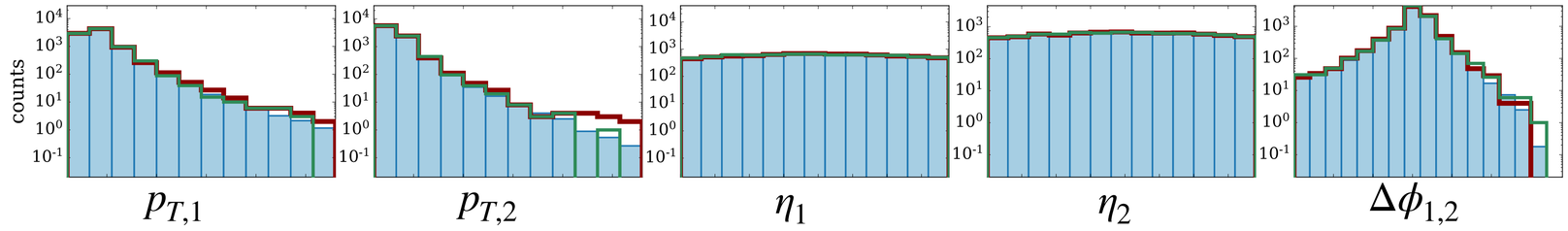
NPLM vs. C2ST

1D toy model: Reference model $n(x|\mathrm{R}) = \mathrm{N}(\mathrm{R})e^{-x}$



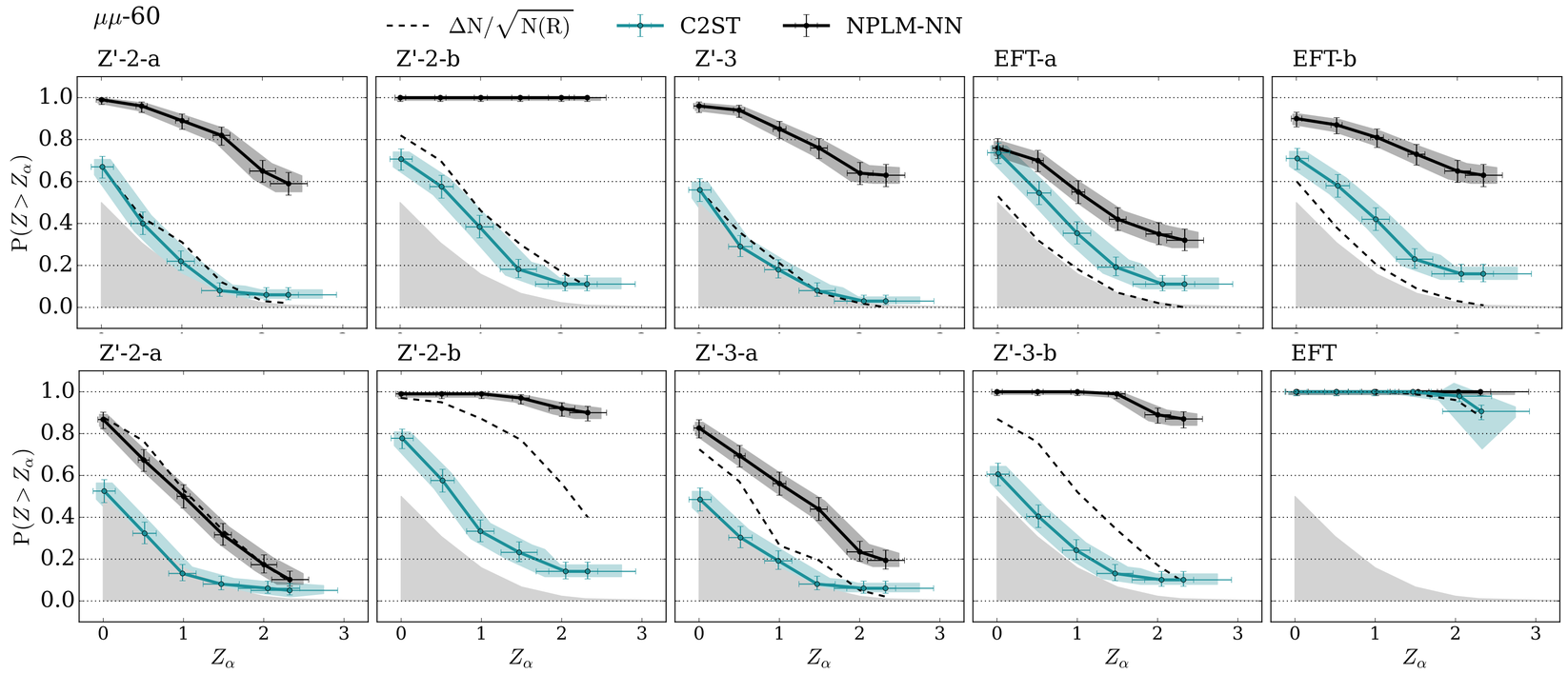
NPLM vs. C2ST

5D $\mu\mu$ final state:



REFERERNCE
Z' scenario
EFT scenario

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



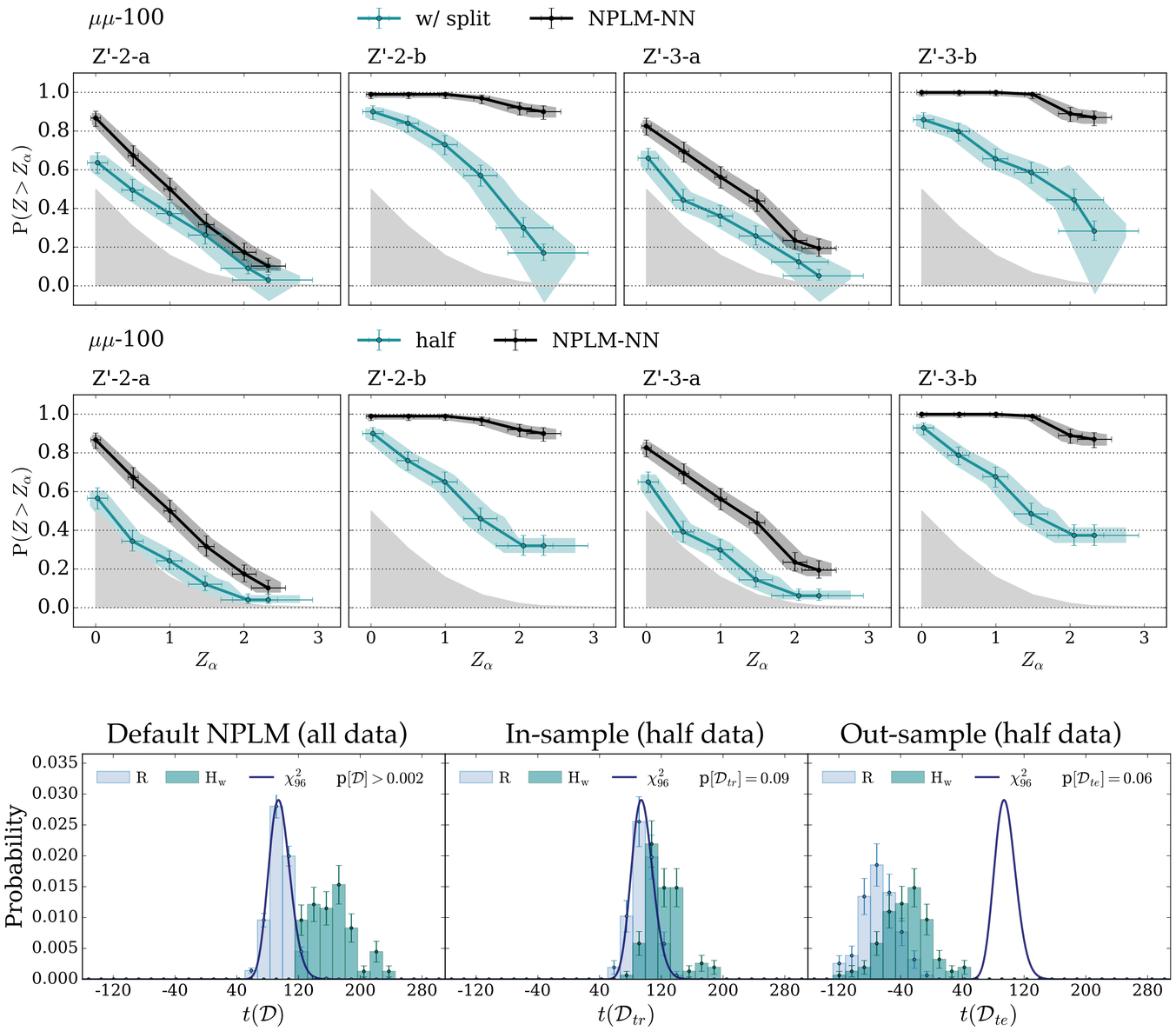
$M_{12} > 60$ GeV (**$\mu\mu$-60**).

$M_{12} > 100$ GeV (**$\mu\mu$-100**)

Train-test splitting

- With the NPLM method, out of sample evaluation is almost equivalent to using half of the data sample.
- NPLM is regularised to "overfit" the specific fluctuations of the training data. By construction, the learnt model does not generalise well to a new data sample.
- $t(D_{te})$ is negative: the alternative learnt from D_{tr} is worst than the Reference

Train-test splitting is disadvantageous for NPLM

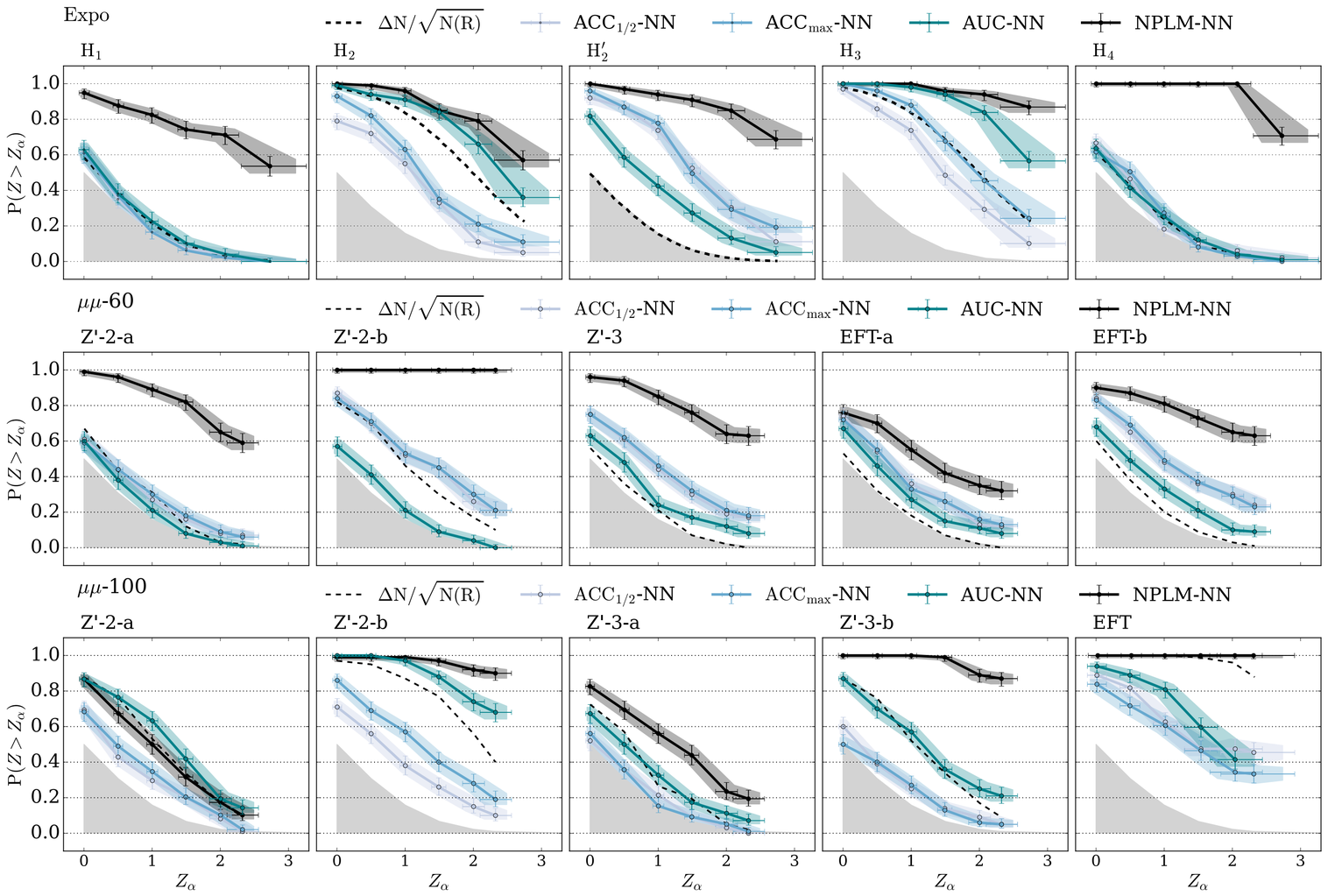


Classifier-inspired variants of the test statistics in NPLM

Replace LRT with other 1D GOF tests

- Classifer-based tests
  - Modified* Balanced Accuracy (ACC)
  - Modified* Area Under the ROC Curve (AUC)

* sensitive to n normalisation effects

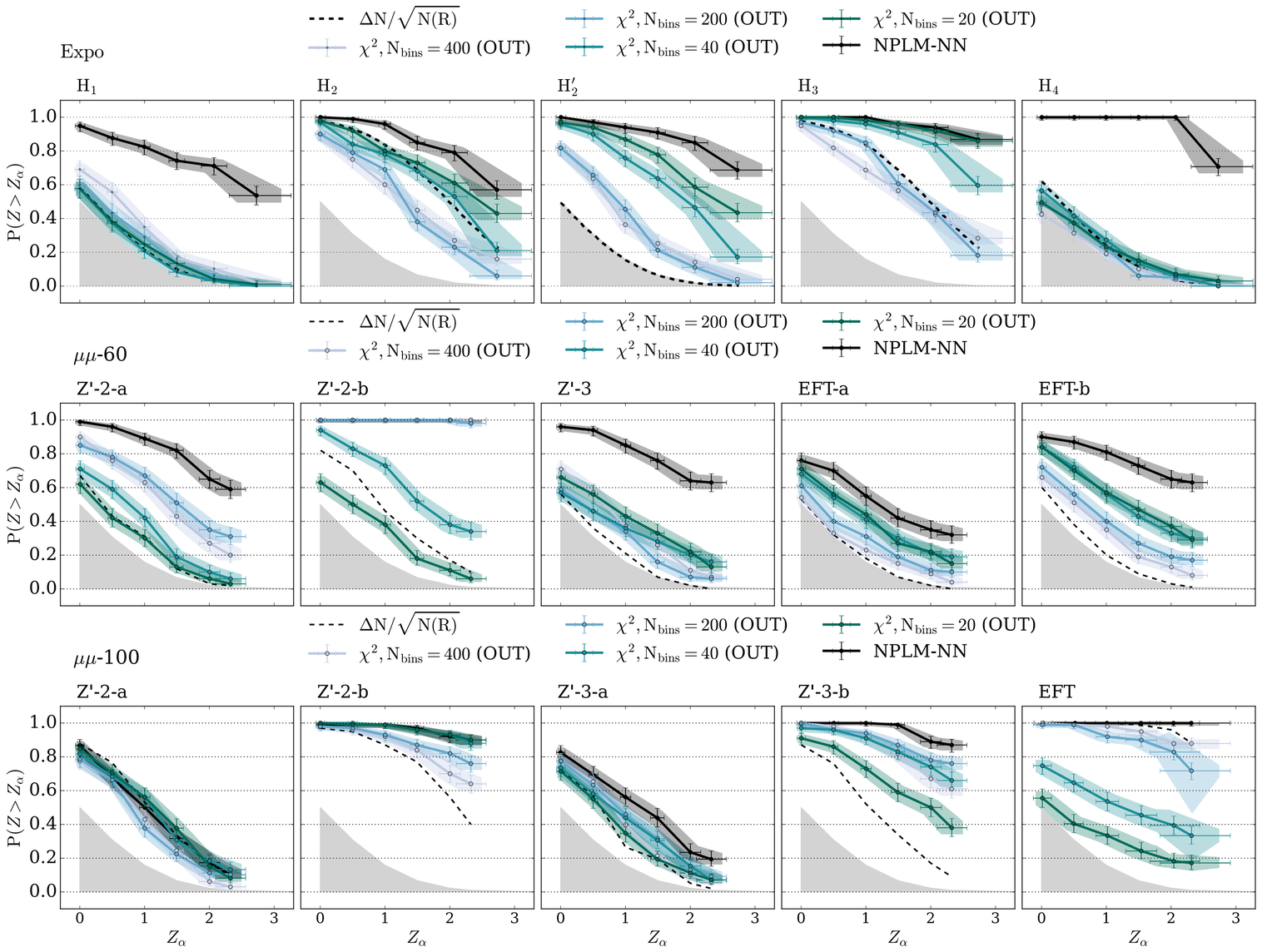


Classifier-inspired variants of the test statistics in NPLM

Replace LRT with other 1D GOF tests

- Classifer-based tests:
  - Modified Balanced Accuracy (ACC)
  - Modified Area Under the ROC Curve (AUC)

- Standard 1D GoF tests:
  - χ^2 tests
  - EDF tests
    - Kolmogorov-Smirnov (KS)
    - Cramer-von Mises (CvM)
    - Anderson-Darling (AD)
  - Spacing statistics:
    - Moran (M)
    - Recursive Product Spacing (RPS)

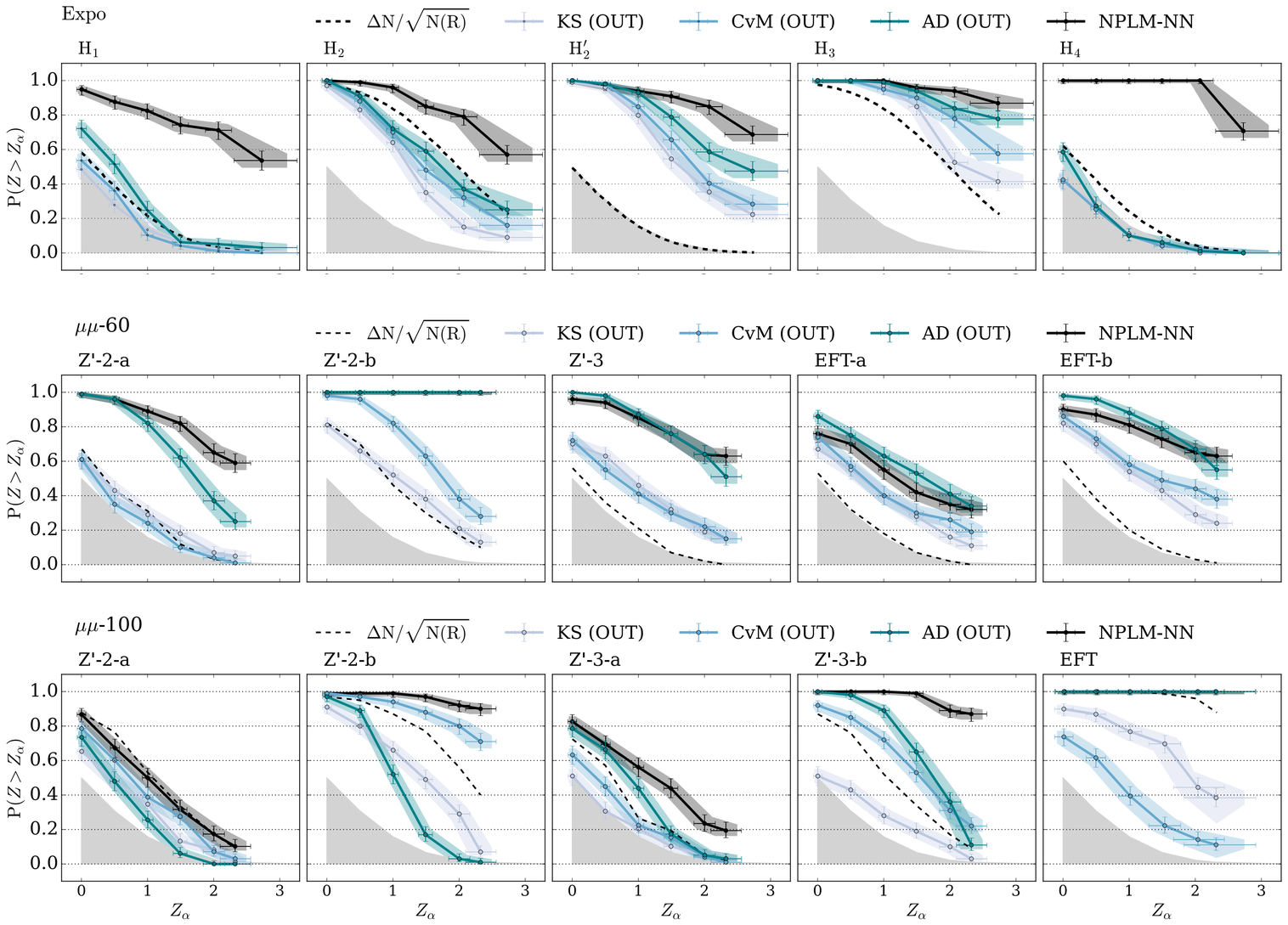


Classifier-inspired variants of the test statistics in NPLM

Replace LRT with other 1D GOF tests

- Classifer-based tests:
  - Modified Balanced Accuracy (ACC)
  - Modified Area Under the ROC Curve (AUC)

- Standard 1D GoF tests:
  - χ^2 tests
  - EDF tests
    - Kolmogorov-Smirnov (KS)
    - Cramer-von Mises (CvM)
    - Anderson-Darling (AD)
  - Spacing statistics:
    - Moran (M)
    - Recursive Product Spacing (RPS)

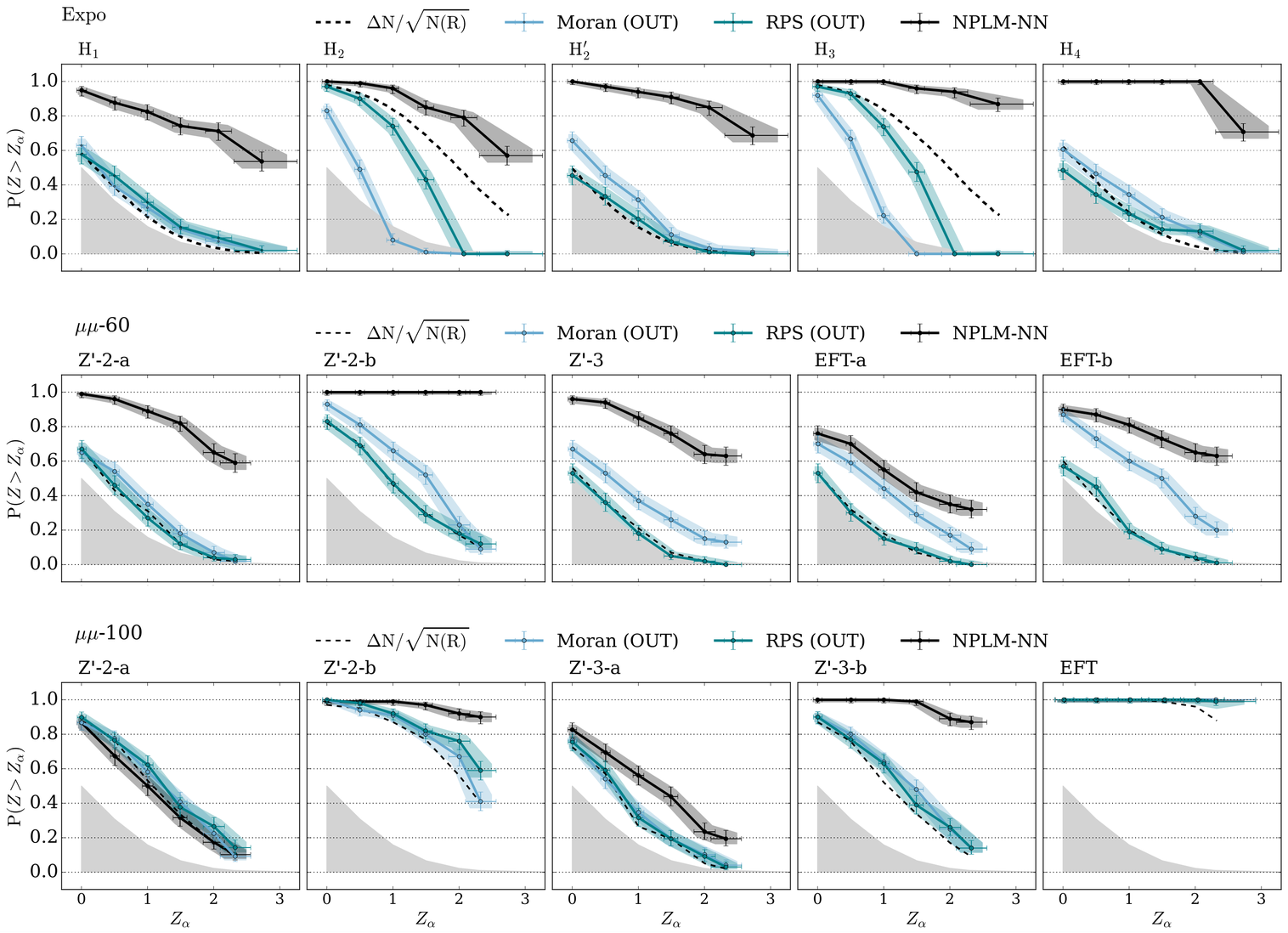


Classifier-inspired variants of the test statistics in NPLM

Replace LRT with other 1D GOF tests

- Classifer-based tests
  - Modified Balanced Accuracy (ACC)
  - Modified Area Under the ROC Curve (AUC)

- Standard 1D GoF tests:
  - χ^2 tests
  - EDF tests
    - Kolmogorov-Smirnov (KS)
    - Cramer-von Mises (CvM)
    - Anderson-Darling (AD)
  - Spacing statistics
    - Moran (M)
    - Recursive Product Spacing (RPS)

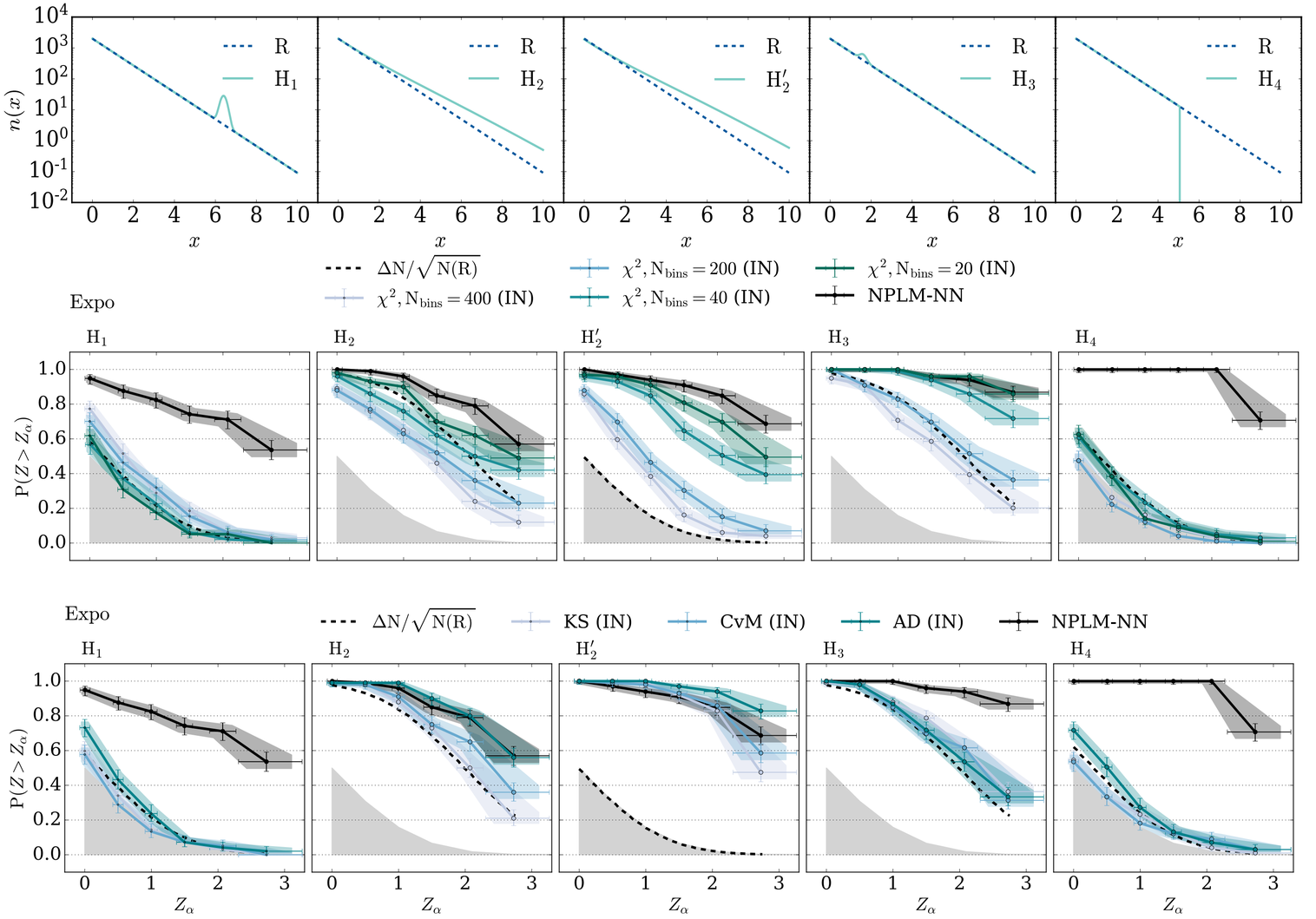


NPLM vs. standard 1D GOF tests

Comparison with standard 1D GoF tests

- χ^2 tests
- EDF tests
  - Kolmogorov-Smirnov (KS)
  - Cramer-von Mises (CvM)
  - Anderson-Darling (AD)

The likelihood-ratio test has, on average, the highest power against the list of considered anomalies



Summary

NPLM as a general tool for GOF, beyond New Physics searches.

- Signal-agnostic test
- Global p-value: one test to detect them all
- Multivariate: allow to look at the data in a more inclusive way, no need to compress the information in one variable and hence to make assumptions on which observables are relevant
- Systematic uncertainties for "imperfect" Reference models Eur. Phys. J. C 82, 275 (2022) (crucial for LHC analysis)
- Fast execution: kernel methods on GPUs

Summary

NPLM as a general tool for GOF, beyond New Physics searches.

- Comparisons with a classifier-based approach (C2ST) show on average better performances of NPLM.
  - Strong advantages in NPLM come from:
    - in-sample evaluation of the test
    - LRT as the test statistic
  - Comparisons to other ML-based approaches to be done
- Comparisons with standard 1D GOF show on average better performances of NPLM.

- Choice of benchmarks:
  - defines the figure of merit for the comparison.
  - never going to be exhaustive, but we should aim at the bast possible representation of the landscape of possible anomalies
  - Our selected benchmarks are biased by the initial problem of BSM searches and should be extended (DQM use case is a first step in that direction)