UniGe | MaLGa

SLING

The New Physics Learning Machine

ML for goodness-of-fit via Neyman—Pearson testing

Marco Letizia – Machine Learning Genoa Center

In collaboration with:

G. Grosso (IAIFI), M. Pierini (CERN), L. Rosasco (MaLGa), A. Wulzer (IFAE), M. Zanetti (UniPd).

Based on: arXiv:2204.02317, arXiv:2303.05413, arXiv:2305.14137.

Code: https://github.com/FalkonHEP (under revision)

Overview

NPLM is a test designed to compare data with a statistical model (GoF):

- Developed for new physics searches at the LHC
- Data-driven
- Signal-agnostic
- Unbinned and multivariate
- No data splitting
- Flexible
- Interpretable

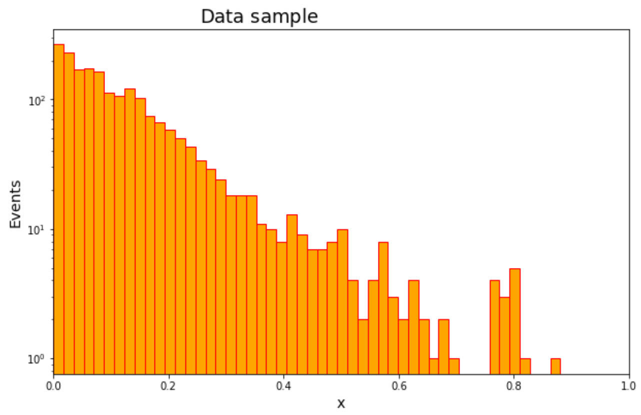
Similarities with two-sample testing:

can it be used to evaluate simulators and surrogates?

Goodness-of-fit as a two-sample test

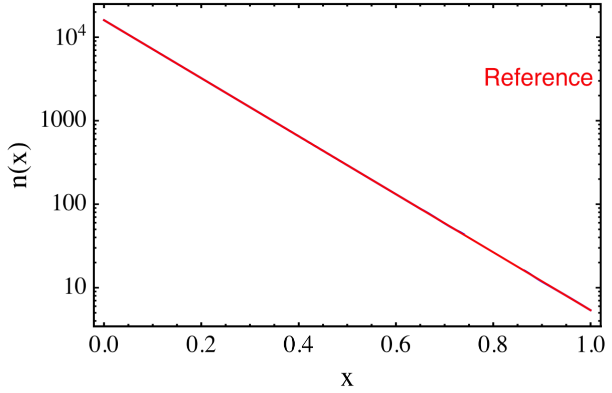
Data

$$\mathcal{D} = \{x_i\}_{i=1}^{N_\mathcal{D}}, \qquad x_i \sim p_{\text{true}}(x)$$



Theoretical expectation (R)

$$p(x|R), \qquad N(R)$$

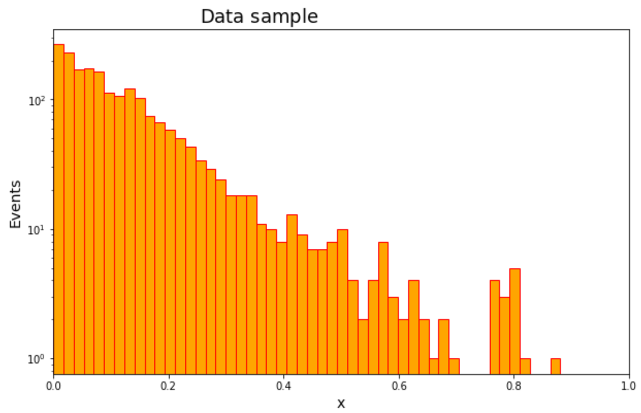


$$\boxed{p_{\text{true}}(x) \neq p(x|R)}$$

Goodness-of-fit as a two-sample test

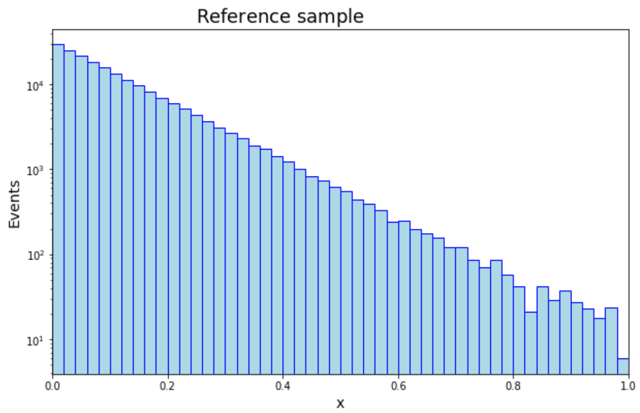
Data

$$\mathcal{D} = \{x_i\}_{i=1}^{N_\mathcal{D}}, \qquad x_i \sim p_{\mathrm{true}}(x)$$



Theoretical expectation (R)

$$\mathcal{R} = \{x_i\}_{i=1}^{N_\mathcal{R}}, \qquad x_i \sim p(x|R), N(R)$$



$\mathcal{N}_\mathcal{R} \gg \mathcal{N}_\mathcal{D}$

$$\boxed{p_{\mathrm{true}}(x) \neq p(x|R)}$$

The New Physics Learning Machine

Model data as local deformation of reference

$$\boxed{n(x|\cdot) = N(\cdot)p(x|\cdot)}$$

$$n(x|w) = e^{f_w(x)} n(x|R) \quad \Rightarrow \quad f_w(x) = \log \frac{n(x|w)}{n(x|R)} \approx \log \frac{n_{\text{true}}(x)}{n(x|R)}$$

Likelihood: $$L(\mathcal{D}|\cdot) = \frac{e^{-N(\cdot)}}{N_{\mathcal{D}}!} \prod_{x=1}^{\mathcal{N}_{\mathcal{D}}} n(x|\cdot)$$

Likelihood ratio test: $$t_w(\mathcal{D}) = -2\left[\frac{N(R)}{N_{\mathcal{R}}} \sum_{x\in\mathcal{R}} \left(e^{f_w(x)} - 1\right) - \sum_{x\in\mathcal{D}} f_w(x)\right]$$

The New Physics Learning Machine

Choose $\hat{w}$ from the data: turn it into a supervised problem

Data: $\{(x_i, y_i)\}_{i=1}^{N_{\mathcal{D}}+N_{\mathcal{R}}}$, with $\begin{cases} y_i = 0 \text{ if } x_i \in \mathcal{R} \\ y_i = 1 \text{ if } x_i \in \mathcal{D} \end{cases}$

Loss $\ell(f_w(x), y)$: minimum $f_{\hat{w}} \approx f^* = \log \dfrac{n(x|1)}{n(x|0)} = \log \dfrac{n_{\text{true}}(x)}{n(x|R)}$

$$\Rightarrow \quad t_{\hat{w}}(\mathcal{D}) = -2\left[\frac{N(R)}{N_{\mathcal{R}}} \sum_{x \in \mathcal{R}} \left(e^{f_{\hat{w}}(x)} - 1\right) - \sum_{x \in \mathcal{D}} f_{\hat{w}}(x)\right]$$

The New Physics Learning Machine

- Maximum likelihood by minimum loss with neural networks

D'Agnolo et al (2018), arXiv:1806.02350; D'Agnolo et al (2019), arXiv:1912.12155.

- Fast kernel-based logistic regression ML et al (2022), arXiv:2204.02317

Logistic loss: $$\ell(f(x), y) = (1-y)\frac{N(R)}{N_{\mathcal{R}}}\log\left(1+e^{f}\right) + y\log\left(1+e^{-f}\right)$$

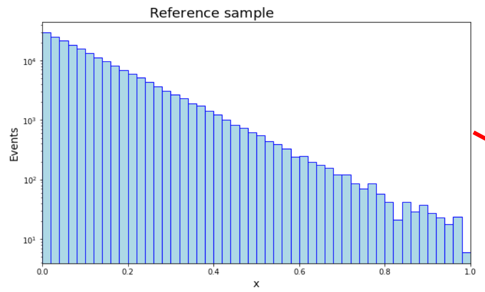
Kernels: $$f_w = \sum_{i=1}^{n} w_i\, k_\sigma(x, x_i)\,, \qquad k_\sigma(x, x_i) = \exp - \frac{\|x - x_i\|^2}{2\sigma^2}$$

Falkon library: G. Meanti et al, arXiv:2006.10350

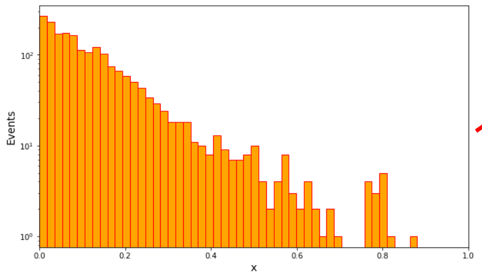
The New Physics Learning Machine

INPUT (unbinned)

Reference sample $\mathcal{R}$



Data sample $\mathcal{D}$



Falkon

$f_w \rightarrow f_{\widehat{w}}$

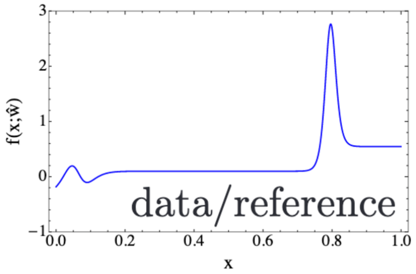
OUTPUT

Likelihood ratio test statistic

$t_{\widehat{w}}(\mathcal{D})$

Density ratio

$f_{\widehat{w}}(x) \approx \log \dfrac{n_{\text{true}}(x)}{n(x|R)}$

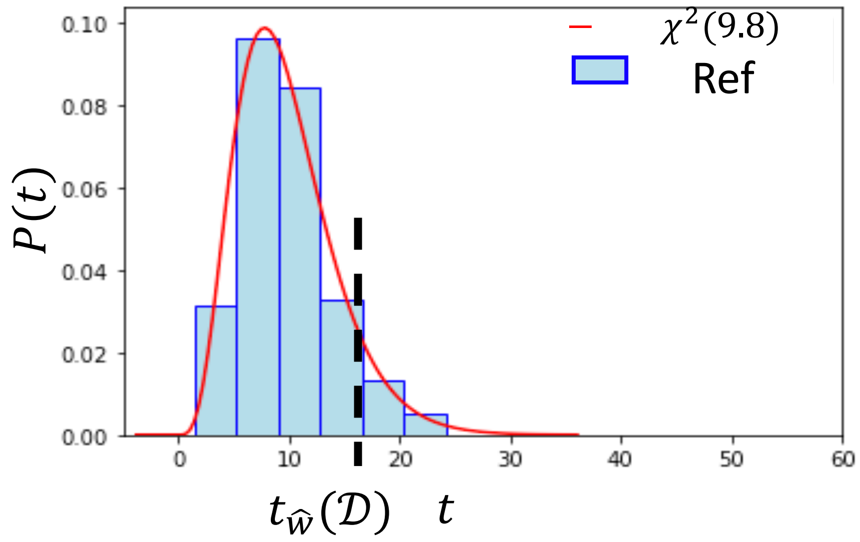


The New Physics Learning Machine

Large $t_{\hat{w}}(\mathcal{D}) \rightarrow$ disagreement with the reference model.

How large? We need to calibrate.

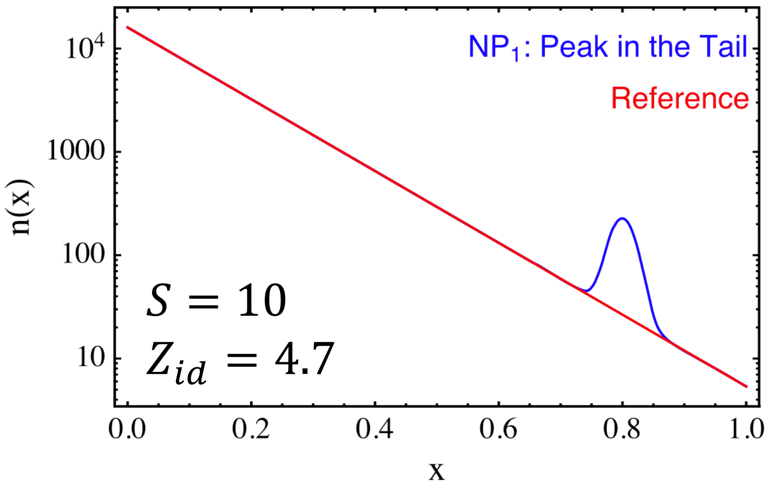
- Train the model on $\mathcal{R}$ against multiple R-distributed *toys.*
- Permutations: train on random permutations of the dataset.
- Exact.

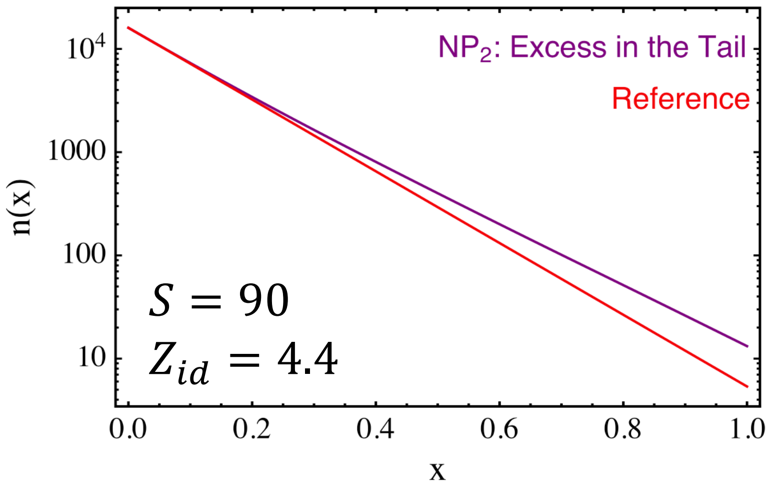


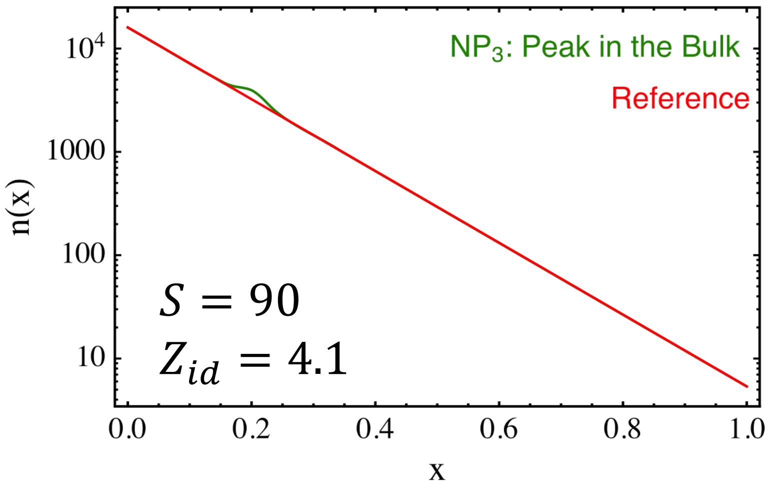
$$\rightarrow p_{\text{value}} = \int_{t_{\hat{w}}(\mathcal{D})}^{\infty} dt\, p(t)\,, \qquad Z = \Phi^{-1}(1 - p_{\text{value}})$$

Univariate example

$N_{\mathcal{R}} = 2 \times 10^5, \qquad N(R) = 2000, \qquad N_{\mathcal{D}} = N(R) + S$



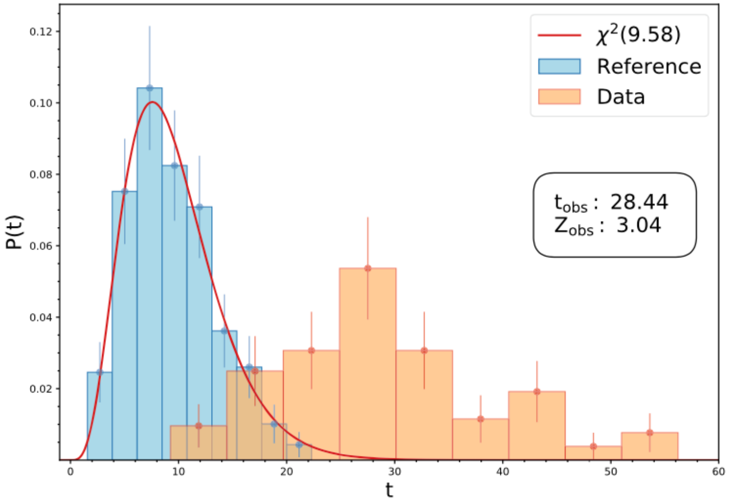


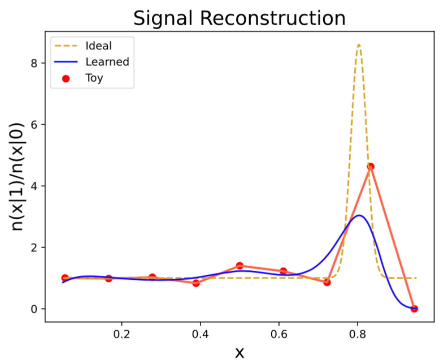


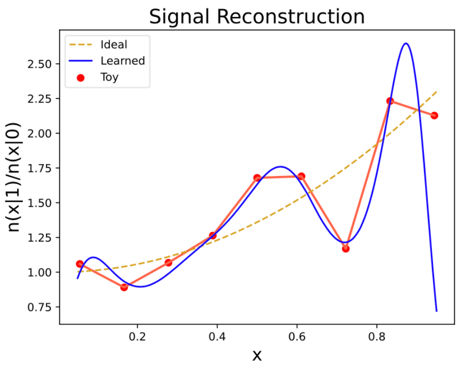
300 R-toys
100 D-toys

$Z_{obs} = (2.43, 3.04, 2.82)$

$\bar{t}_{tr} = 2.11$ sec



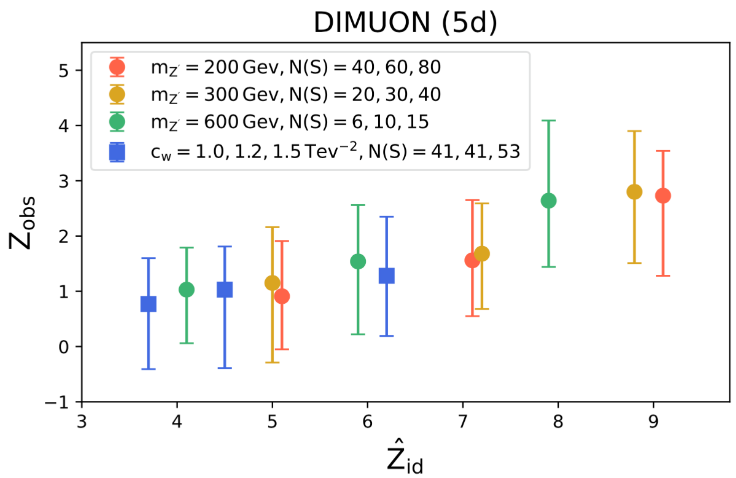


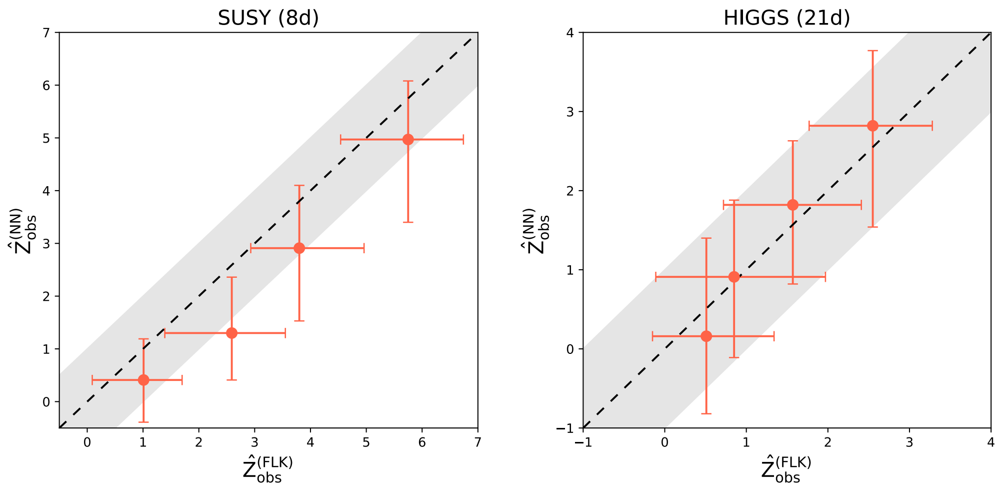


Multivariate

$pp \to \mu^+\mu^-$: SM vs SM+Z'/EFT $[p_{T1}, p_{T2}, \eta_1, \eta_2, \Delta\phi]$, SUSY (8d), HIGGS (21d)

$N(R) = 2\times10^4$, $N_R = 10^5$ | $N(R) = 10^5$, $N_R = 5\times10^5$





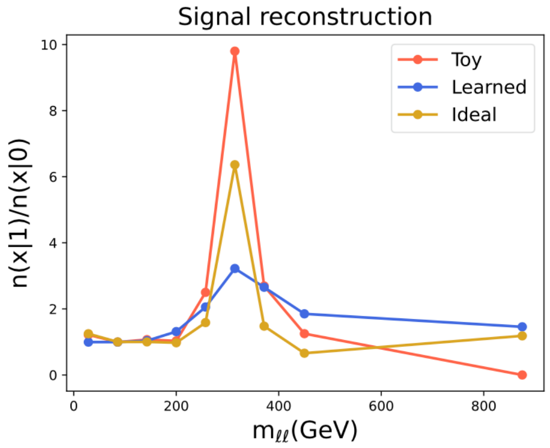


Table 1 Average training times per single run with standard deviations (low level features and reference toys). Note that time measured in hours (for NN) and seconds (for Falkon)

Model	DIMUON	SUSY	HIGGS
FLK	**(44.9 ± 3.4) s**	**(18.2 ±1.2) s**	**(22.7 ± 0.4) s**
NN	(4.23 ± 0.73) h	(73.1 ± 10) h	(112 ± 9) h

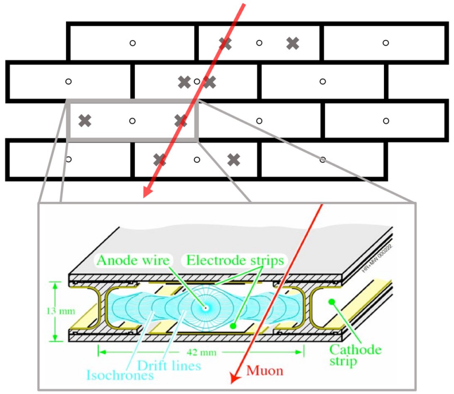
Bold values indicate the lowest for each column (lower is better)

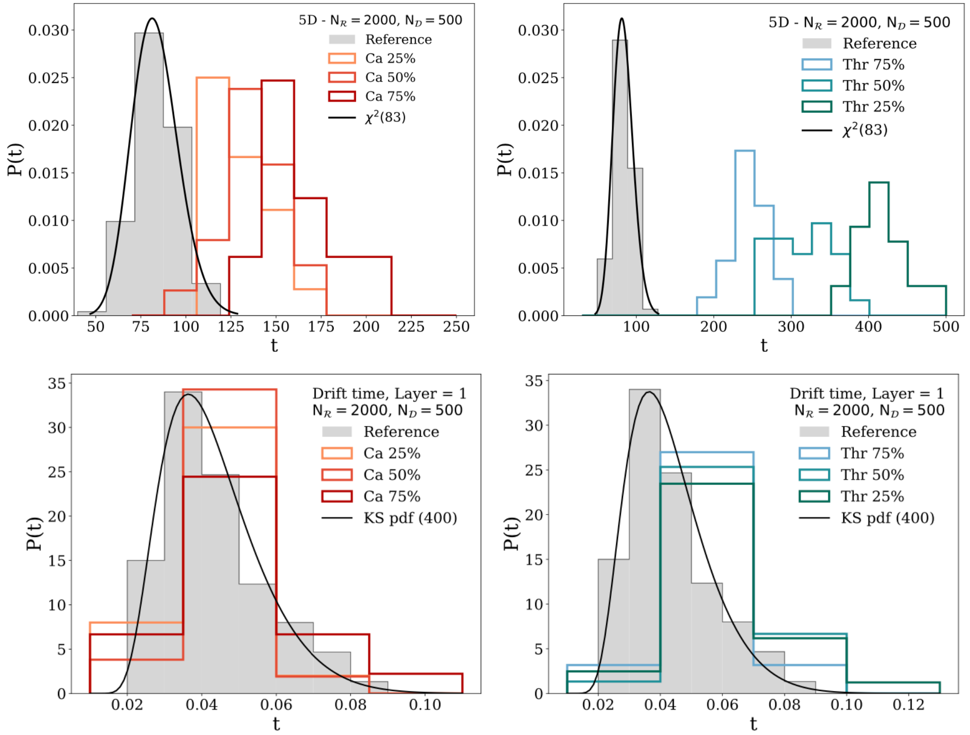
Data: https://zenodo.org/records/4442665

Data Quality Monitoring

G. Grosso et al, [arXiv:2303.05413](https://arxiv.org/abs/2303.05413)

Drift tube chambers from Legnaro INFN National Laboratory.





DATASET:

- Drift times (t_i): the four drift times of the muon track.
- Slope (ϕ): the angle with respect to the vertical axis.
- Reference data is collected in a controlled regime.
- Anomalies:
  - reduced voltage of cathodic strips to 75%, 50%, and 25% of their nominal value (-1.2 kV)
  - lowered front-end thresholds to 75%, 50%, and 25% of nominal value (100 mV)

$$\boxed{\bar{t}_{tr} \approx 0.5 \text{ sec}}$$

Data: https://zenodo.org/records/7128223

Outlook

- Ongoing effort to apply NPLM to real analysis (CMS).
- Compare/combine with alternative approaches to AD and GoF.

Challenges:

- Inexact simulations – nuisance parameters (worked out for NN).
- Reliance on reference toys.
- Model selection for signal-agnostic approaches.

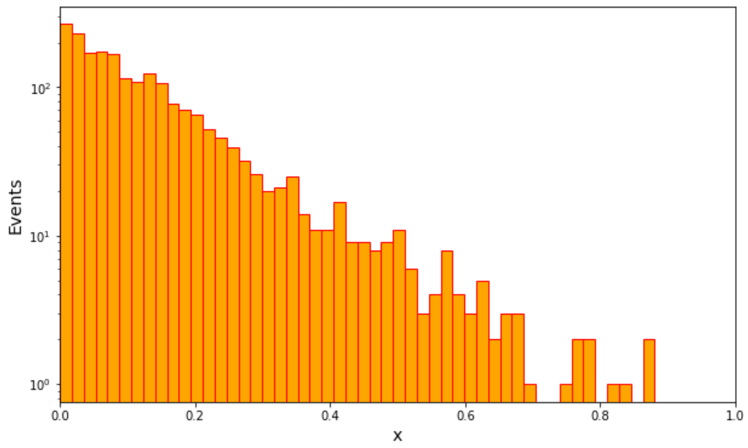
Opportunities (mostly driven by efficiency):

- Indications that NPLM could be SOTA for two-sample testing.
- Compare MC generators.
- Evaluation of generative models.

Backup

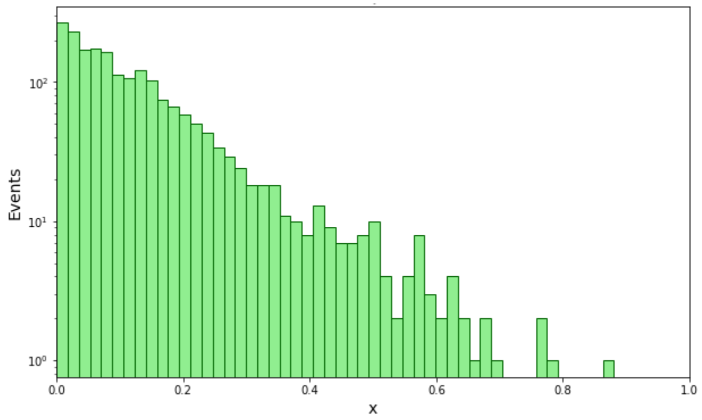
MC test data

$$\mathcal{D} = \{x_i\}_{i=1}^{N_\mathcal{D}}, \qquad x_i \sim p_{\mathrm{MC}}(x)$$



Generative model

$$\mathcal{R} = \{x_i\}_{i=1}^{N_\mathcal{R}}, \qquad x_i \sim p_{gen}(x)$$



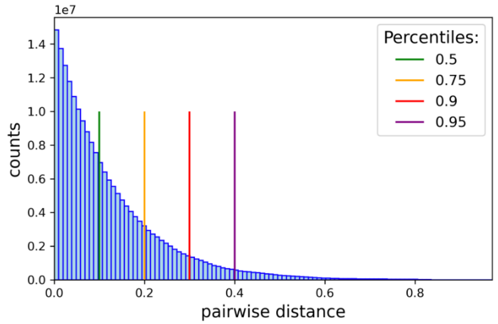
$$\boxed{p_{\mathrm{MC}}(x) \neq p_{gen}(x)}$$

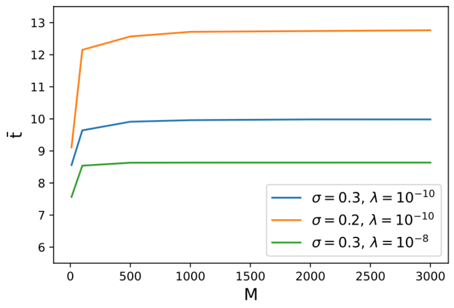
Backup

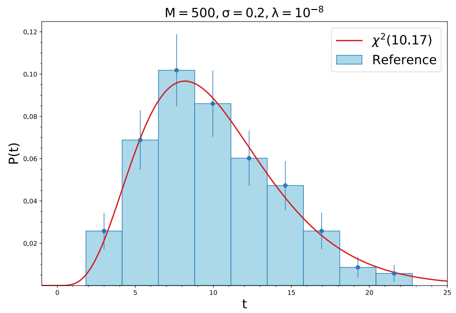
Falkon has three main hyperparameters (M, σ, λ)

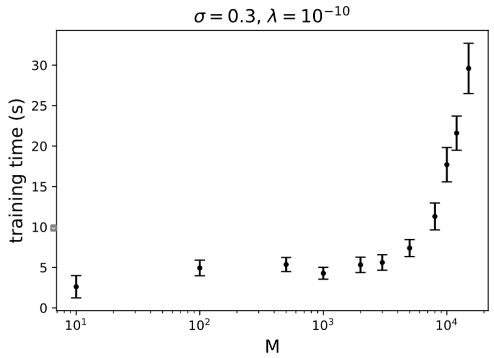
No cross-validation to preserve model-independence.

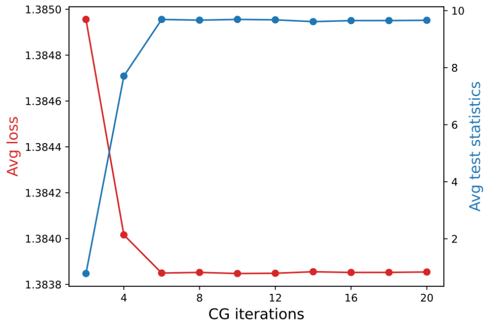
→ mix of heuristics, statistical considerations and efficency

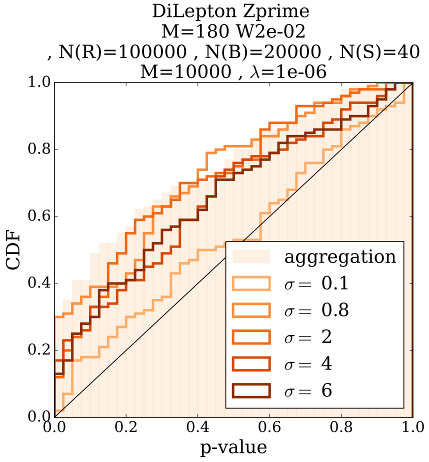


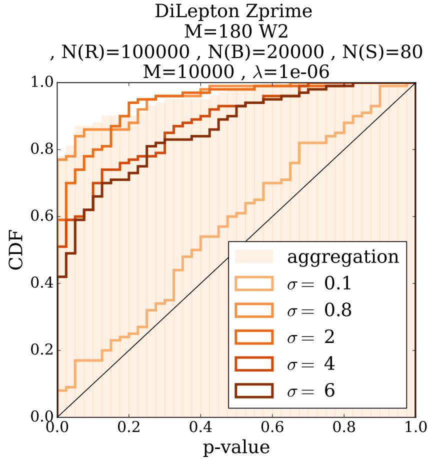








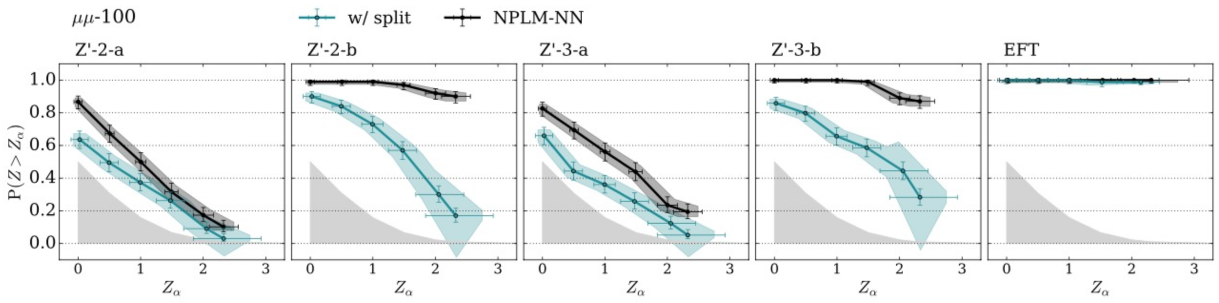




PRELIMINARY

Backup G. Grosso, ML, M. Pierini, A. Wulzer [arXiv:2305.14137](https://arxiv.org/abs/2305.14137)

Train-test split



Different metrics