UniGe | MaLGa

SLING

Kernel methods for new physics searches

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 (under revision): https://github.com/FalkonHEP, https://github.com/mletizia/FalkonNPLM_1D.

COMETA

The problem

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

against a reference model *R* to test the hypothesis

H_0: $p_{\text{true}}(x) = p(x|R)$.

Model-independence: avoid alternatives to R (e.g. BSM) or signal hypotheses.

H_1: $p_{\text{true}}(x) \neq p(x|R)$.

Goodness-of-fit

The problem

Theoretical expectation as a reference sample

$$\mathcal{R} = \{x_i\}_{i=1}^{N_\mathcal{R}}, \qquad x_i \overset{\text{iid}}{\sim} p(x|R), \qquad N(R), \qquad \text{with} \quad N_\mathcal{R} \gg N_\mathcal{D}.$$

$$\Rightarrow \quad \text{Test } H_0 \text{ by comparing } \mathcal{D} \text{ against } \mathcal{R}.$$

- Flexible.
- Efficient in multivariate and large-scale regimes.
- Unbinned.
- Interpretable.

The New Physics Learning Machine

$$n(x|R) = N(R)p(x|R)$$

Model data as local deformation of the reference

$$n(x|w) = e^{f_w(x)} n(x|R) \quad \Rightarrow \quad f_w(x) = \log \frac{n(x|w)}{n(x|R)} \quad \left(f(x) = \log \frac{n(x|BSM)}{n(x|R)} \right)$$

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

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

$N(w) - N(R)$

The New Physics Learning Machine

Choose $\hat{w}$ from the data as a supervised learning 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 $f_{\hat{w}} \approx f^* = \log \frac{n(x|1)}{n(x|0)} = \log \frac{n_{\text{true}}(x)}{n(x|R)}$

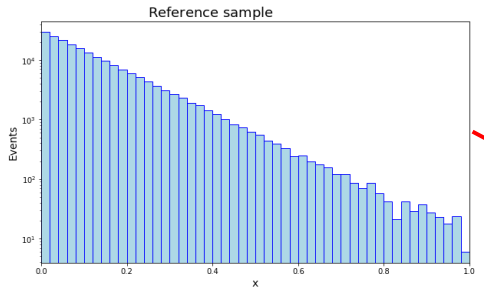
$$\Rightarrow \quad t_{\hat{w}}(\mathcal{D}) = -2\left[\frac{N(R)}{N_{\mathcal{R}}} \sum_{x \in \mathcal{R}} \left(e^{f_{\hat{w}}(x)} - 1\right) - \sum_{x \in \mathcal{D}} f_{\hat{w}}(x)\right]$$

The New Physics Learning Machine

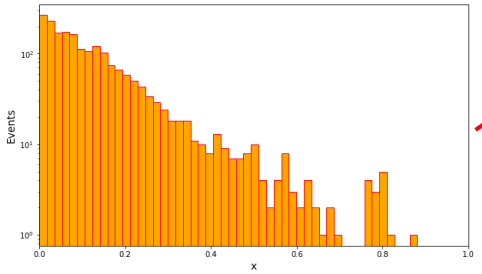
INPUT (unbinned)

OUTPUT

Reference sample $\mathcal{R}$



Data sample $\mathcal{D}$



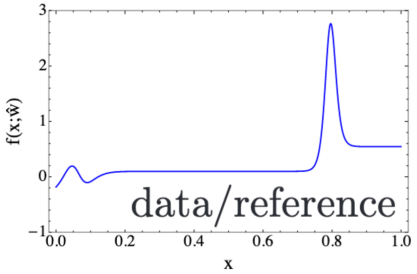
$f_w \to f_{\hat{w}}$

Likelihood ratio test statistic

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

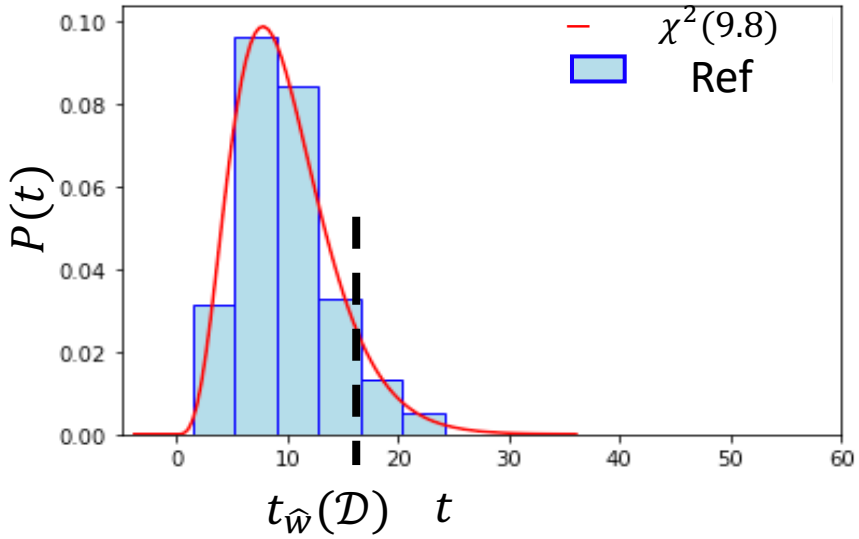
Density ratio

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



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.*

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

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 logistic regression

  ML et al (2022), arXiv:2204.02317.

Learning new physics with a kernel machine

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

Kernel methods: $$f_w(x) = \sum_{i=1}^{N} w_i\, k(x, x_i)\,, \qquad k_\sigma(x, x') = \exp -\frac{\|x - x'\|^2}{2\sigma^2}.$$

- Universal approximators.
- Convex optimization with guarantees.

→ Falkon: a SOTA solver for kernel methods G. Meanti et al, arXiv:2006.10350

Learning new physics with a kernel machine

Kernel methods are expensive:

$\mathcal{O}(n^2)$ in space and $\mathcal{O}(n^3)$ in time $-$ store and invert $K_{nn} \in \mathbb{R}^{n \times n}$.

Some approximation is needed.

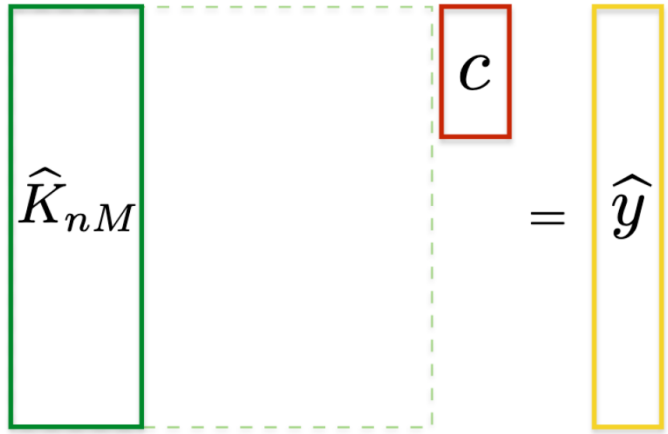
Falkon makes use of:

- Random projections (Nyström)
  - To reduce the size of the problem – $\mathcal{O}(n)$ in space
  - Efficient preconditioned conjugate gradient – $\mathcal{O}(n\sqrt{n}\log n)$ in time
- Efficient (multi-)GPU implementation

Learning new physics with a kernel machine

- Random projections (Nyström)

$$f_w(x) = \sum_{i=1}^{N} w_i \, k(x, x_i) \quad \rightarrow \quad \sum_{i=1}^{M} w_i \, k(x, x_i) \,,$$



$\{x_1, \dots, x_M\} \subset \{x_1, \dots, x_n\}$ sampled uniformly at random (centers)

Optimal statistical bounds can be obtained with $M = \mathcal{O}(\sqrt{n})$

Theorem (Rudi, Camoriano, R. '15)

Let $(\tilde{x}_i)_{i=1}^{M} \subseteq (x_i)_{i=1}^{n}$ *picked uniformly at random, if* $\lambda = 1/\sqrt{n}$ *and* $M \geqslant \sqrt{n}$ *then*

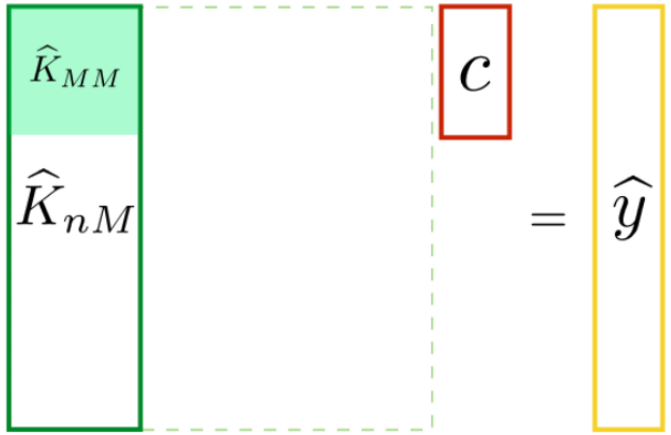
$$\mathbb{E}L(\widehat{f}_{\lambda,M}) - \min_{f \in \mathcal{H}} L(f) \lesssim \frac{1}{\sqrt{n}}$$

Learning new physics with a kernel machine

- Conjugate gradient with efficient preconditioning

Preconditioner: $\mathrm{P}^{\mathrm{T}}\mathrm{P} = (K_{nM}K_{nM} + \lambda n K_{MM})^{-1}$

Nyström → $P^T P \approx \left(\frac{n}{M} K_{MM}^2 + \lambda n K_{MM}\right)^{-1}$



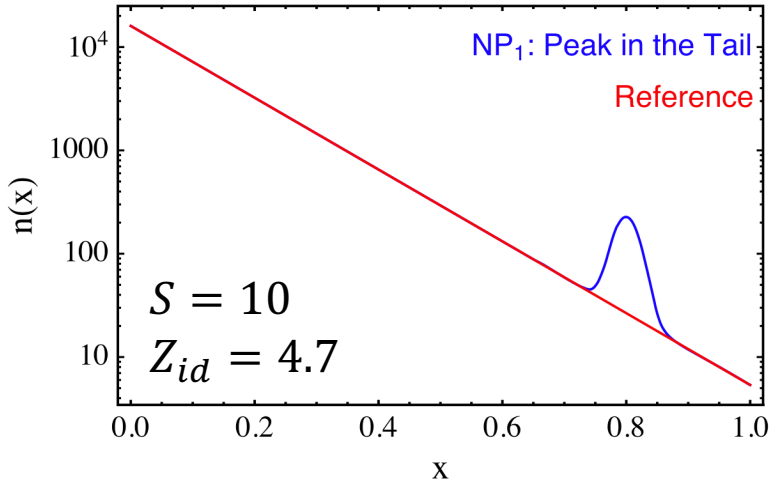
Theorem (Rudi, Carratino, Rosasco '17)

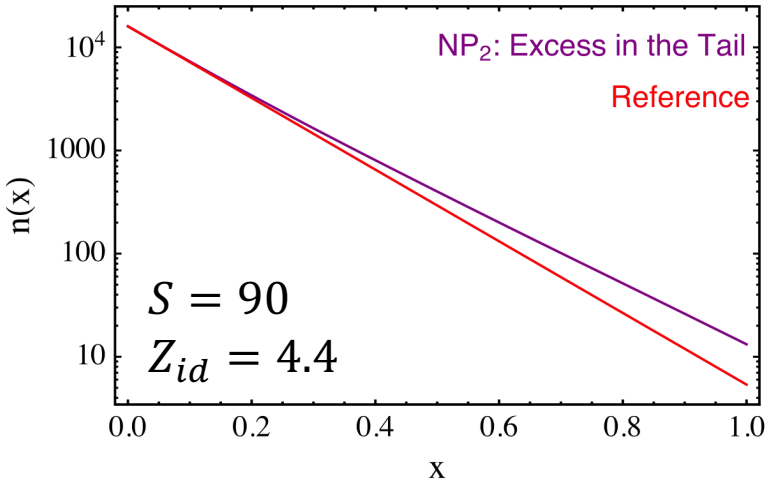
Let $(\tilde{x}_i)_{i=1}^{M} \subseteq (x_i)_{i=1}^{n}$ *uniformly at random, then if* $\lambda = 1/\sqrt{n}$, $M \geqslant \sqrt{n}$ *and* $t \geqslant \log(n)$

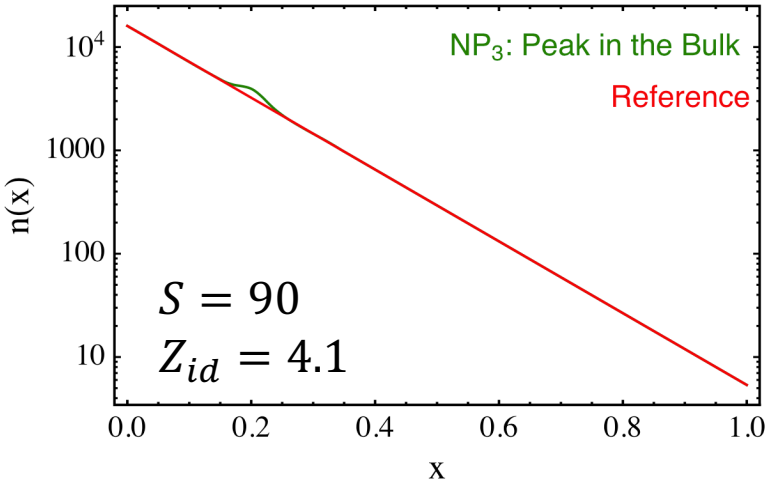
$$\mathbb{E}L(\hat{f}_{\lambda,M,t}) - \min_{f\in\mathcal{H}} L(f) \lesssim \frac{1}{\sqrt{n}}$$

Univariate example

$N_{\mathcal{R}} = 2\times10^5, \quad N(R) = 2000, \quad N_{\mathcal{D}} = N(R) + S$



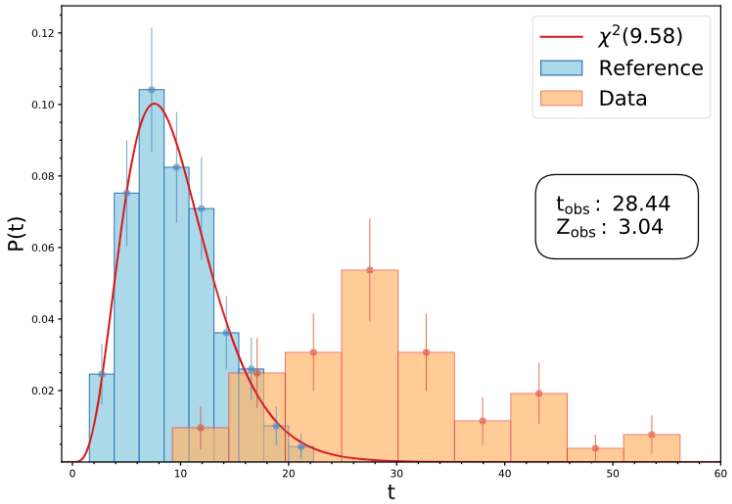


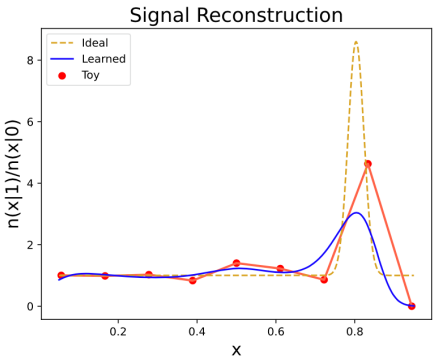


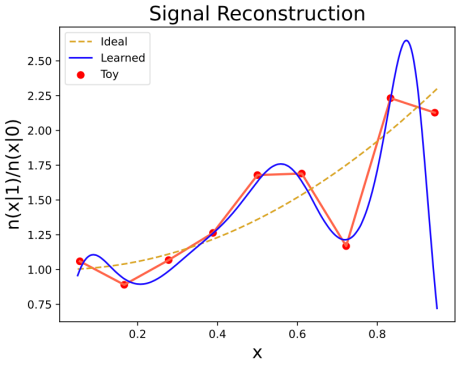
300 R-toys
100 D-toys

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

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



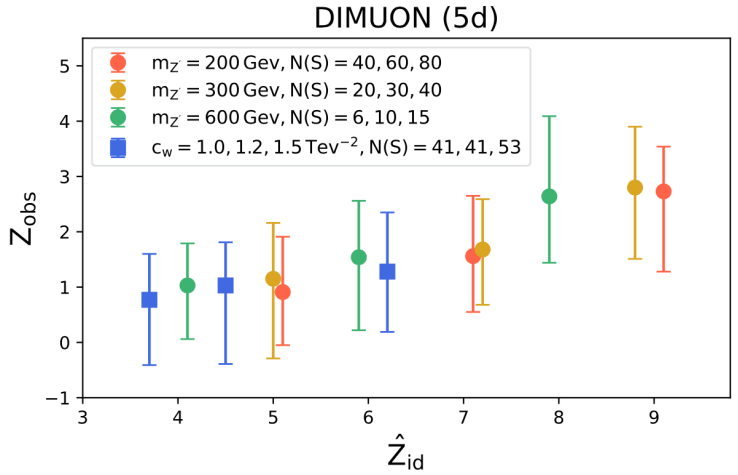


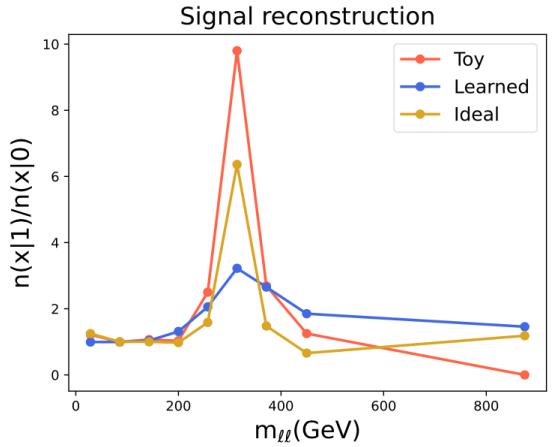


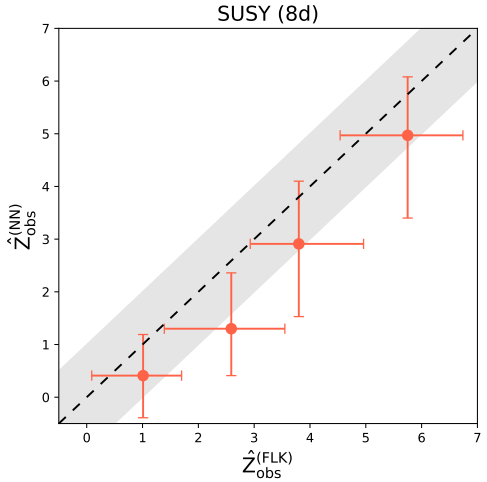
Multivariate

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

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







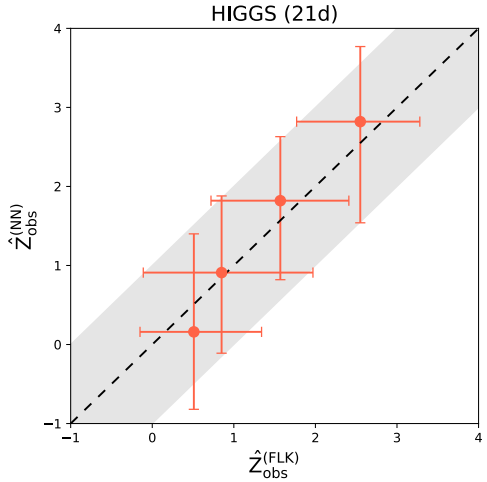


Table 1 Average training times per single run with standard deviations (low level features and reference toys). Note that time measured in hours (for NN) and seconds (for Falkon)

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

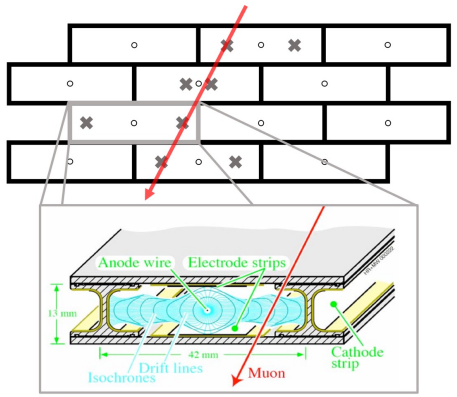
Bold values indicate the lowest for each column (lower is better)

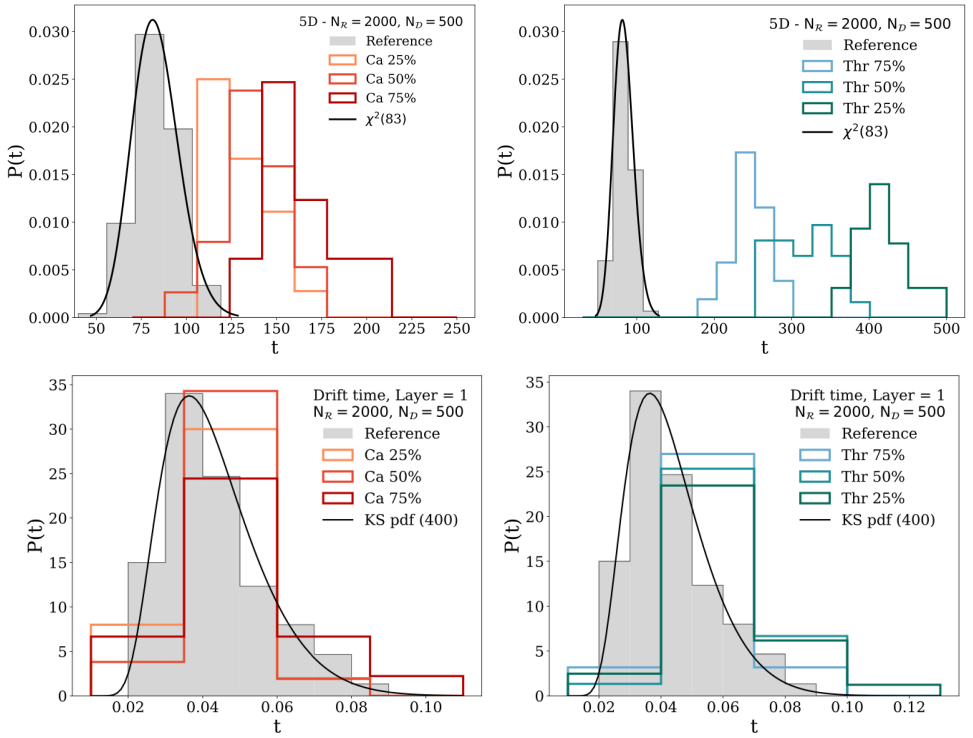
Data: https://zenodo.org/records/4442665

Data Quality Monitoring

G. Grosso et al, arXiv:2303.05413

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)

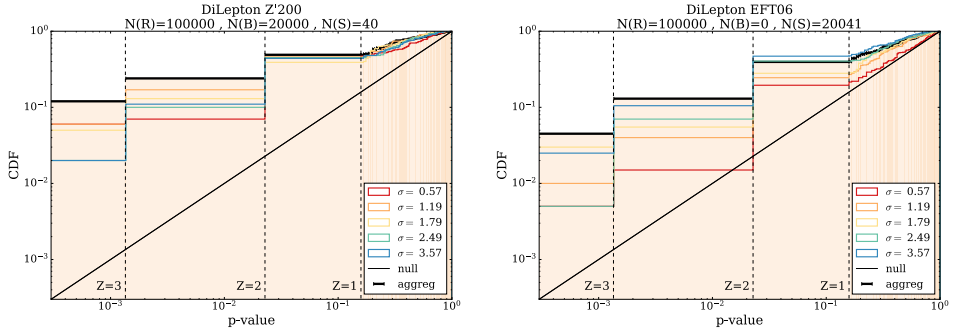
$\bar{t}_{tr} \approx 0.5$ sec

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

Current developments

- Test aggregation

Sensitivity to any given signal depends on model hyperparameters → combine multiple tests



Inspired by *Schrab et al,* arXiv:2110.15073

- Systematic uncertainties

Extension to profile likelihood formalism D'Agnolo et al arXiv:1912.12155

nuisance parameters $f_w(x) \to f_w(x) + \log r_\nu(x)\,, \qquad r_\nu(x) = \exp\left[\nu\,\delta_1(x) + \frac{\nu^2}{2}\,\delta_2(x) + \cdots\right]$

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

Summary

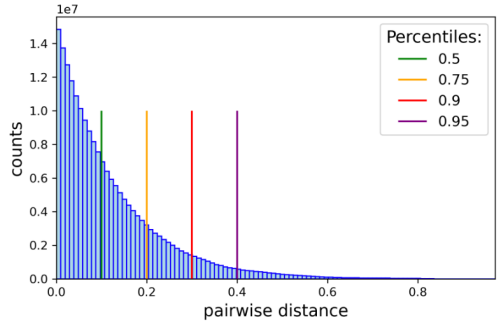
- *New Physics Learning Machine*: methodology to compare a model to the data.
- Developed for new physics searches, it can be used for two-sample testing.
- Implementation based on SOTA large-scale kernel methods.
- Many developments/applications
  - Systematics
  - Algorithmic ideas combining optimization and statistics
  - Data quality monitoring
  - Evaluation of generative models
  - Comparison of MC simulators
  - Connections with foundation models for HEP

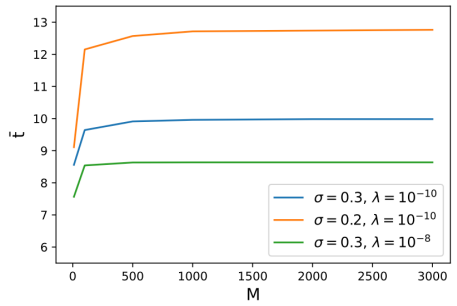
Backup

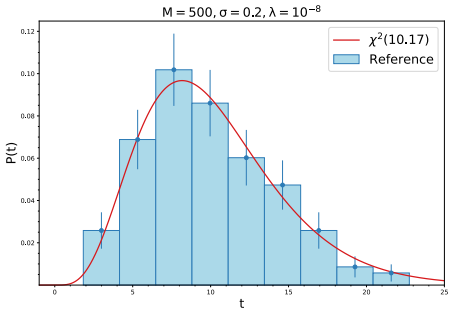
Falkon has three main hyperparameters (M, σ, λ)

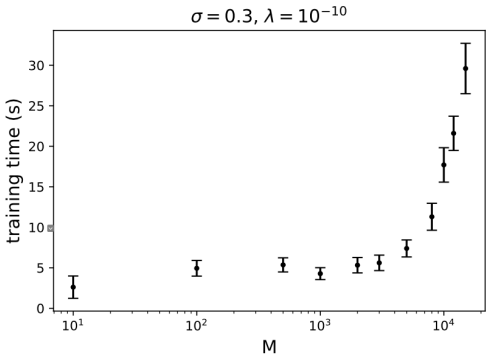
No cross-validation to preserve model-independence.

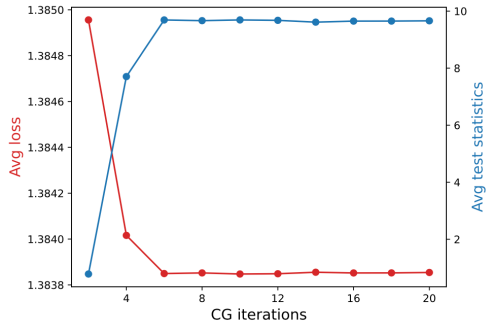
→ mix of heuristics, statistical considerations and efficency





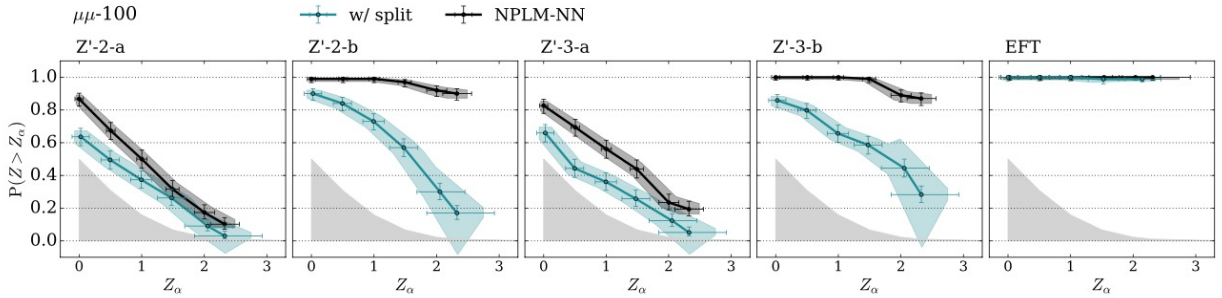






Backup G. Grosso, ML, M. Pierini, A. Wulzer arXiv:2305.14137

Train-test split



Different metrics