Exploring evaluation methods for generative models in HEP

Marco Letizia

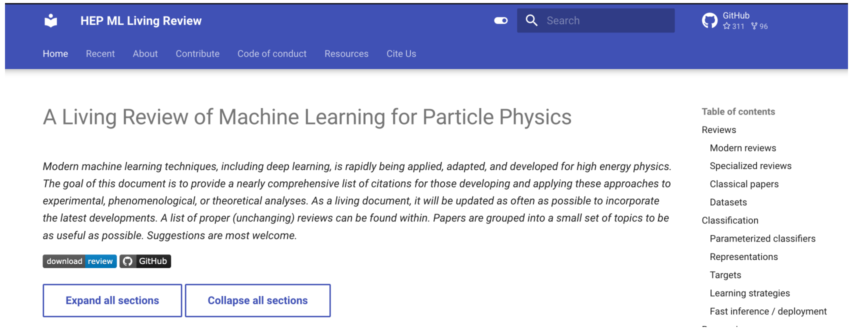
Machine Learning Genoa Center and INFN

In collaboration with:

Andrea Coccaro (INFN), Gaia Grosso (IAIFI Boston),

Riccardo Torre (INFN), Humberto Reyes-Gonzalez (RWTH Aachen)

Generative models in HEP



https://iml-wg.github.io/HEPML-LivingReview/

Generative models: new examples from estimated pdf

(GANs, normalizing flows, diffusion models)

$$x_{new} \sim p_{\text{gen}}(x) \approx p_{\text{true}}(x)$$

- Fast simulations
- Data augmentation
- Anomaly detection
- Background estimation

Standardized and robust model evaluation is crucial!

- R. Kansal, A. Li, J. Duarte, N. Chernyavskaya, M. Pierini, B. Orzari, T. Tomei arXiv:2211.10295
- A. Coccaro, ML, H. Reyes-Gonzalez, R. Torre arXiv:2302.12024
- R. Das, L. Favaro, T. Heimel, C. Krause, T. Plehn, D. Shih arXiv:2305.16774
- J. Gavranovič, B. P. Kerševan arXiv:2310.08994

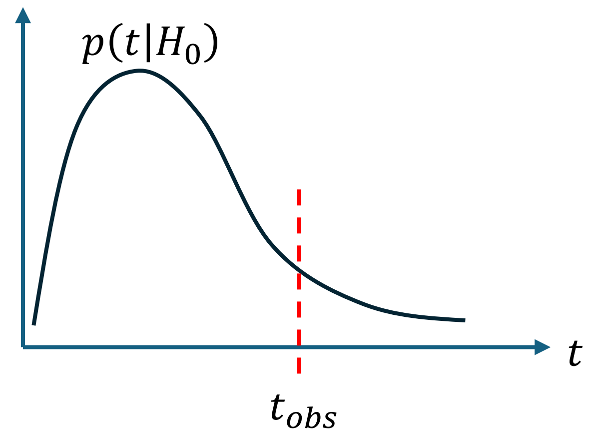
UniGe | MaLGa

Evaluation of generative models

Address the problem as a two-sample test:

reject H_0: $p_{\text{gen}} = p_{\text{true}}$ from data $X = \{x_i\} \sim p^n_{\text{true}},\ Y = \{y_i\} \sim p^m_{\text{gen}}$

- Define a test statistic $t: \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$
- Compute observed test $t_{obs} = t(X, Y)$
- Assess $p(t)$ under the null hypothesis H_0
  - Analytic
  - Toys $t(X, X')$
  - Permutations
  - Bootstrap



- $p(t_{obs}) = \int_{t_{obs}}^{\infty} p(t|H_0)dt \to Z = \Phi^{-1}(1 - p(t_{obs}))$

Evaluation of generative models

Large scale and multivariate regime: prioritise sensitivity or efficiency.

- From univariate to multivariate tests
  A. Coccaro, ML, H. Reyes-Gonzalez, R. Torre arXiv:2302.12024
  S. Grossi, R. Torre, *to appear soon*

- The New Physics Learning Machine
  R. Tito D'Agnolo, A. Wulzer arXiv:1806.02350
  ML, G. Losapio, M. Rando, G. Grosso, A. Wulzer, M. Pierini, M. Zanetti, L. Rosasco arXiv:2204.02317
  G. Grosso, ML, M. Pierini, A. Wulzer arXiv:2305.14137
  P. Cappelli, G. Grosso, ML, H. Reyes-Gonzalez, *work in progress*

From univariate to multivariate tests

A. Coccaro, ML, H. Reyes-Gonzalez, R. Torre arXiv:2302.12024
S. Grossi, R. Torre, *to appear soon*

Some metrics:

- Dimension-avg KS test

$$D_{x,y} = \sup_x \left| F_p(x) - F_q(x) \right|, \qquad \bar{D} = \frac{1}{d} \sum_i D^{(i)}$$

- Sliced 1-Wasserstein distance N. Bonneel, J. Rabin, G. Peyré, and H. Pfiste, *JMIV* (2015)

$$W_1(p,q) = \int \left| F_p(x) - F_q(x) \right| dx\,, \qquad SWD = \frac{1}{k} \sum_i W_1^{(i)} \quad \text{(randomly on a hypersphere)}$$

- Maximum mean discrepancy (a.k.a. KPD) *A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, A. Smola, JMLR (2012)*

$$\mathrm{MMD}^2 = \sup_{f\in\mathcal{F}} \left(\mathbb{E}_p[f(x)] - \mathbb{E}_q[f(x)] \right)$$

From univariate to multivariate tests

S. Grossi, R. Torre, *to appear soon*

Benchmark: mixtures of q Gaussians in N dimensions

$$\begin{cases} q=3,\ N=5 \\ q=5,\ N=20 \\ q=10,\ N=100 \end{cases}$$

Deformations:

- ϵ shift in the means
- ϵ shift in the standard deviations
- Both

For each case and test, find ϵ:

reject H_0 at level $\alpha = 0.95, 0.99.$

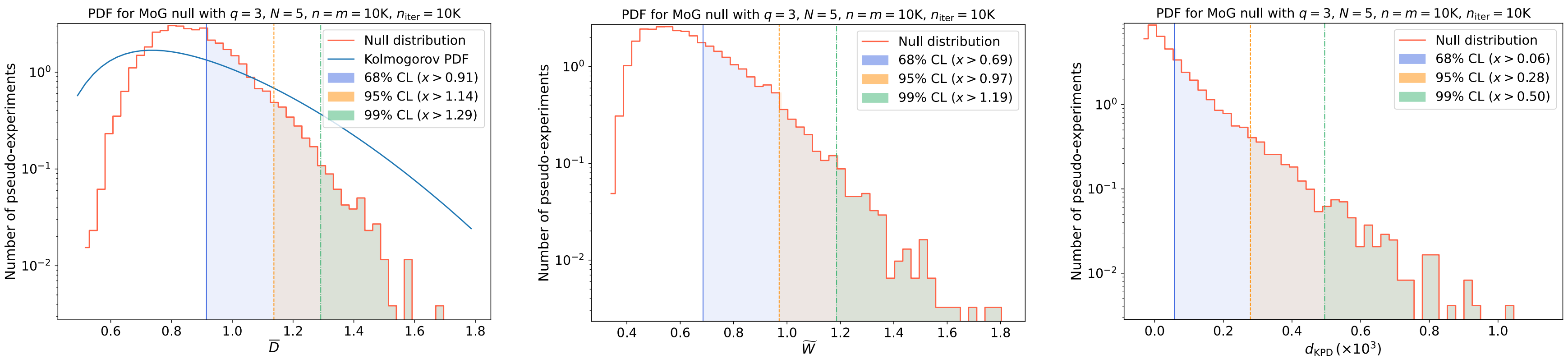
In this synthetic case, we can use the exact likelihood-ratio test to provide a notion of best performance (Neyman-Pearson lemma)

$$t = \log \frac{\mathcal{L}(\epsilon = 0)}{\mathcal{L}(\epsilon)}$$

From univariate to multivariate tests

S. Grossi, R. Torre, *to appear soon*

$$n = 10000, n_{\text{toys}} = 10000, q = 3, N = 5.$$

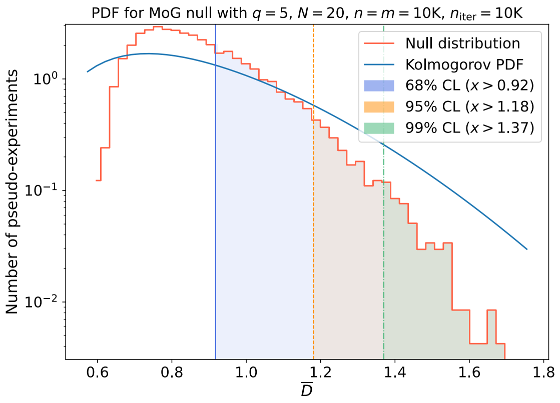


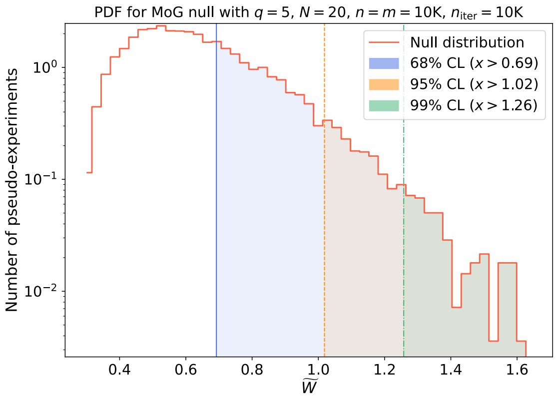
Statistic	$\epsilon^{\mu}_{95\%\text{CL}}$	$\epsilon^{\mu}_{99\%\text{CL}}$	t^{μ} (s)	$\epsilon^{\sigma}_{95\%\text{CL}}$	$\epsilon^{\sigma}_{99\%\text{CL}}$	t^{σ} (s)	$\epsilon^{\mu-\sigma}_{95\%\text{CL}}$	$\epsilon^{\mu-\sigma}_{99\%\text{CL}}$	$t^{\mu-\sigma}$ (s)	t^{null} (s)
t_{LLR}	0.0011	0.0016	1422	0.0014	0.0021	1178	0.00091	0.0013	1316	-
$\overline{D}$	0.009	0.013	779	0.02	0.028	714	0.0076	0.011	689	59
$\widetilde{D}$	0.027	0.037	739	0.056	0.078	701	0.023	0.031	727	826
$\widetilde{W}$	0.046	0.064	521	0.093	0.13	479	0.041	0.058	561	422
$\overline{\Vert \cdot \Vert_F}$	0.059	0.083	630	0.15	0.24	569	0.053	0.075	541	30
d_{FPD}	0.077	0.1	605	0.16	0.22	510	0.069	0.094	569	439
d_{KPD}	0.12	0.16	518	2.5	2.8	357	0.12	0.16	525	2197

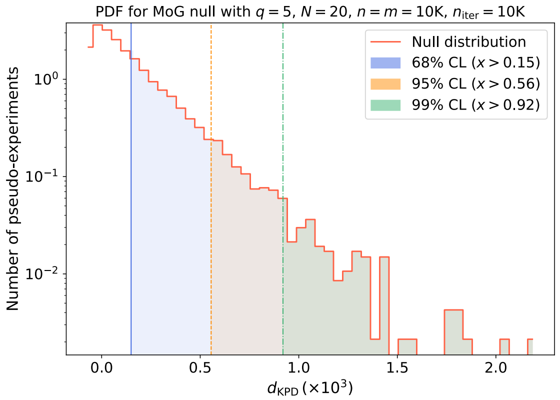
From univariate to multivariate tests

S. Grossi, R. Torre, *to appear soon*

$$n = 10000, n_{\text{toys}} = 10000, q = 5, N = 20.$$





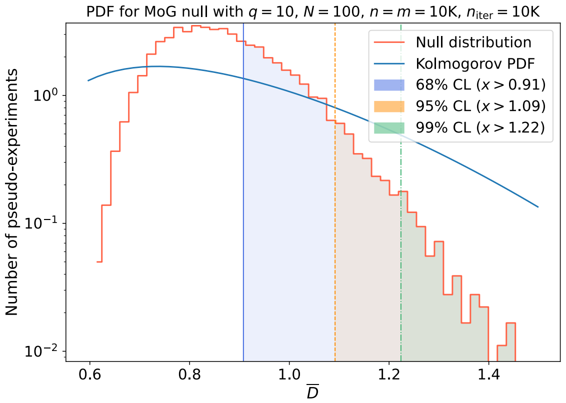


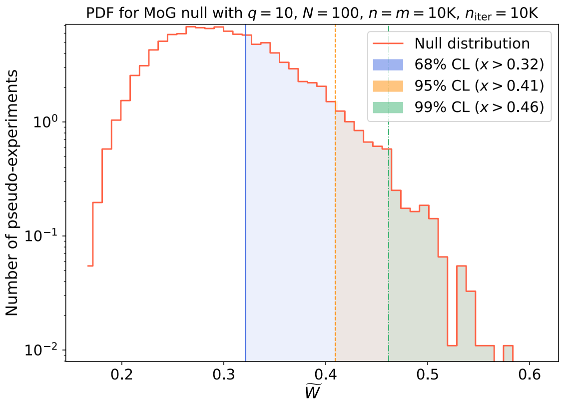
Statistic	$\epsilon^{\mu}_{95\%\text{CL}}$	$\epsilon^{\mu}_{99\%\text{CL}}$	t^{μ} (s)	$\epsilon^{\sigma}_{95\%\text{CL}}$	$\epsilon^{\sigma}_{99\%\text{CL}}$	t^{σ} (s)	$\epsilon^{\mu-\sigma}_{95\%\text{CL}}$	$\epsilon^{\mu-\sigma}_{99\%\text{CL}}$	$t^{\mu-\sigma}$ (s)	t^{null} (s)
t_{LLR}	0.00058	0.0008	5447	0.00079	0.0011	5368	0.00049	0.00065	6230	-
$\overline{D}$	0.01	0.015	2148	0.027	0.039	1971	0.0097	0.014	2170	446
$\lVert \cdot \rVert_F$	0.064	0.092	1678	0.54	0.69	1338	0.064	0.092	1773	90
$\widetilde{D}$	0.074	0.1	1887	0.22	0.33	1659	0.067	0.095	1853	1125
d_{FPD}	0.12	0.16	1694	0.36	0.47	1427	0.11	0.15	1642	512
$\widetilde{W}$	0.12	0.17	1791	0.39	0.58	1658	0.11	0.16	2036	692
d_{KPD}	0.24	0.32	1588	2.9	3.3	1136	0.24	0.32	1598	2027

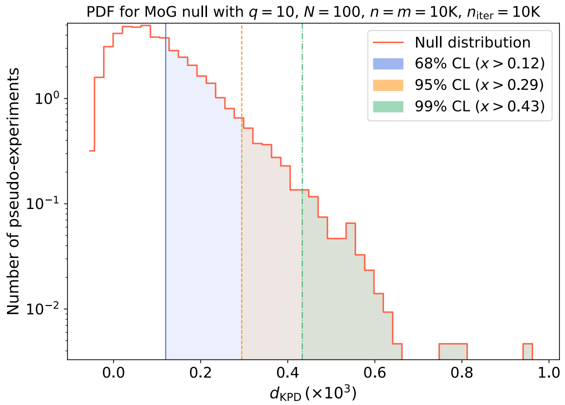
From univariate to multivariate tests

S. Grossi, R. Torre, *to appear soon*

$$n = 10000, n_{\text{toys}} = 10000, q = 10, N = 100.$$







Statistic	$\epsilon^{\mu}_{95\%\text{CL}}$	$\epsilon^{\mu}_{99\%\text{CL}}$	t^{μ} (s)	$\epsilon^{\sigma}_{95\%\text{CL}}$	$\epsilon^{\sigma}_{99\%\text{CL}}$	t^{σ} (s)	$\epsilon^{\mu-\sigma}_{95\%\text{CL}}$	$\epsilon^{\mu-\sigma}_{99\%\text{CL}}$	$t^{\mu-\sigma}$ (s)	t^{null} (s)
t_{LLR}	1e-05	1e-05	4213	1e-05	2e-05	4049	0	1e-05	4309	-
$\overline{D}$	0.0041	0.0076	617	0.0095	0.023	578	0.0035	0.006	678	651
$\widetilde{D}$	0.061	0.082	834	0.45	0.55	712	0.06	0.08	1041	8442
$\widetilde{W}$	0.087	0.12	603	0.53	0.63	481	0.086	0.12	618	5075
$\overline{\Vert\cdot\Vert_F}$	0.094	0.13	356	0.75	0.89	254	0.094	0.13	330	57
d_{FPD}	0.1	0.13	486	0.36	0.45	405	0.099	0.13	468	2185
d_{KPD}	0.21	0.28	458	3.7	4.4	284	0.2	0.28	462	2971

Testing normalizing flows in high-dimensions

A. Coccaro, ML, H. Reyes-Gonzalez, Riccardo Torre [arXiv:2302.12024]

Correlated mixtures of Gaussians

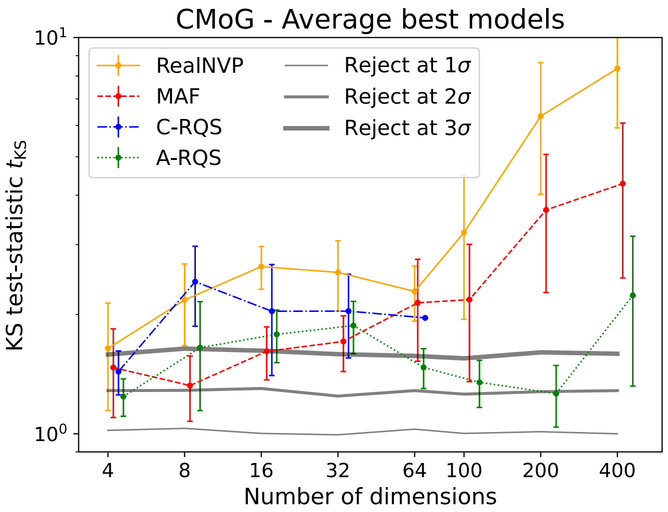
$q = 3$

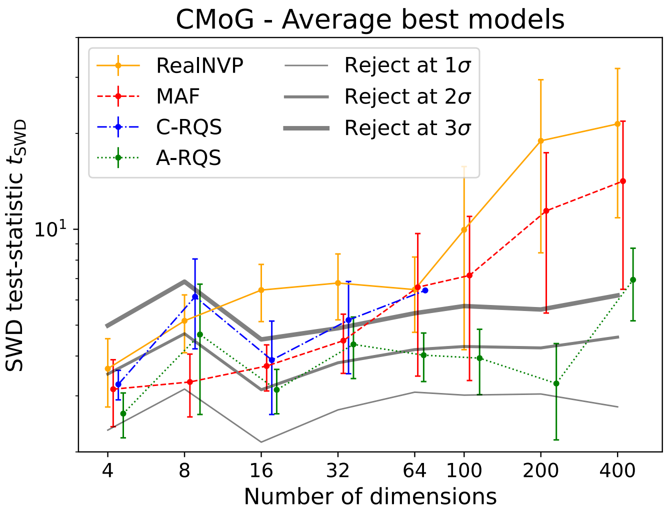
$N = 4 - 400$

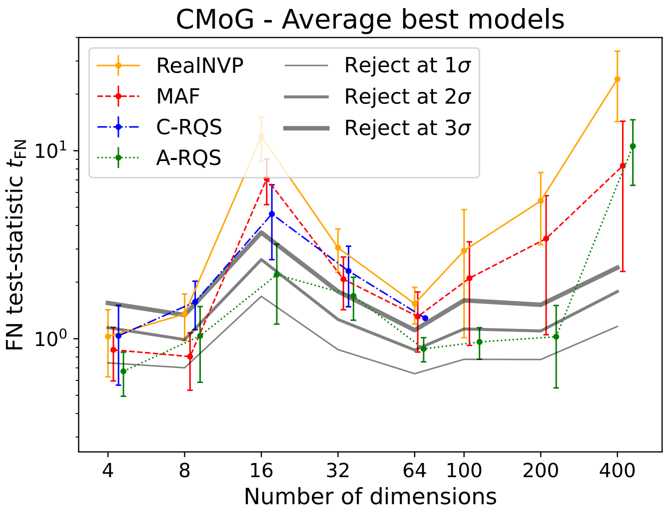
Coupling and autoregressive flows:

- RealNVP
- MAF
- Rational quadratic splines

$N_{\text{test}} = N_{\text{train}} = 10^5$



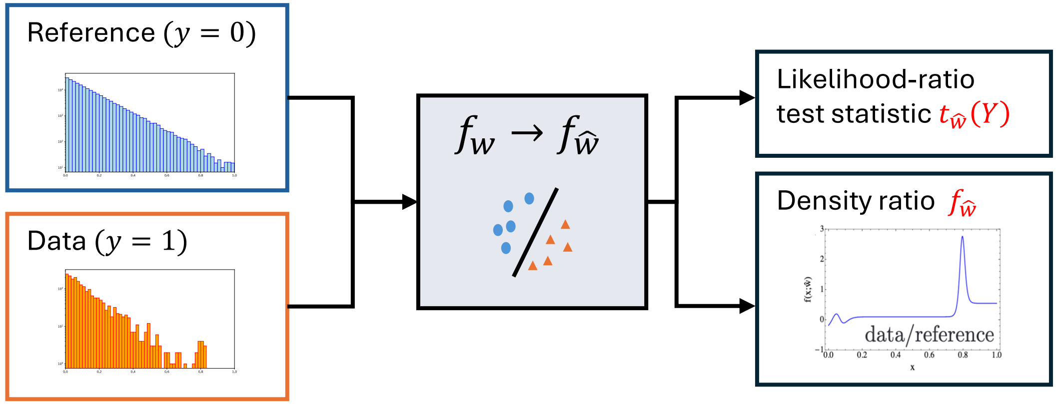




The New Physics Learning Machine

R. Tito D'Agnolo, A. Wulzer arXiv:1806.02350
ML, G. Losapio, M. Rando, G, Grosso, A. Wulzer, M. Pierini, M. Zanetti, L. Rosasco arXiv:2204.02317
G. Grosso, ML, M. Pierini, A. Wulzer arXiv:2305.14137

Likelihood ratio goodness-of-fit test with with supervised learning



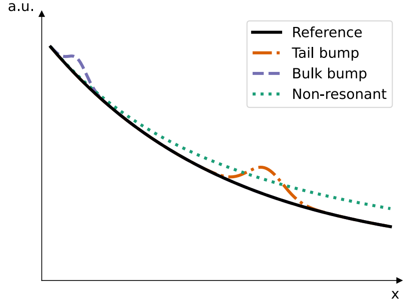
$$f_{\hat{w}} \approx \log \frac{p(x|1)}{p(x|0)}$$

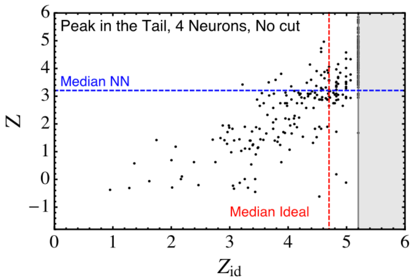
$$t_{\hat{w}}(\mathrm{Data}) = -2 \sum_{x \in \mathrm{Data}} \log f_{\hat{w}}(x)$$

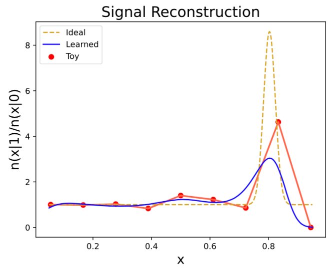
Multivariate; unbinned; efficient; no data splitting.

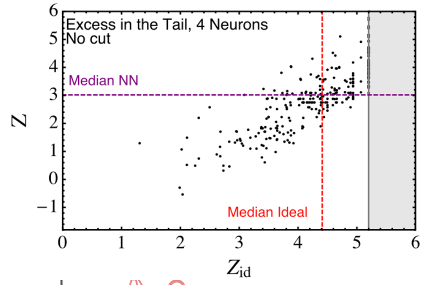
The New Physics Learning Machine

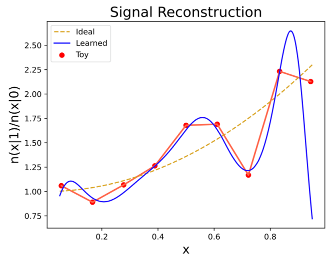
Univariate



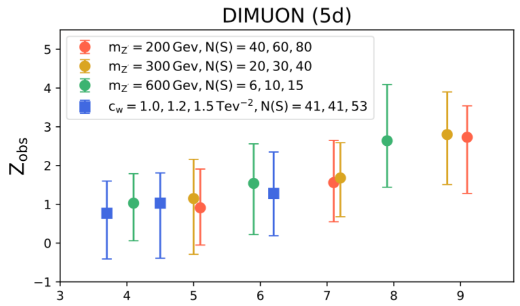


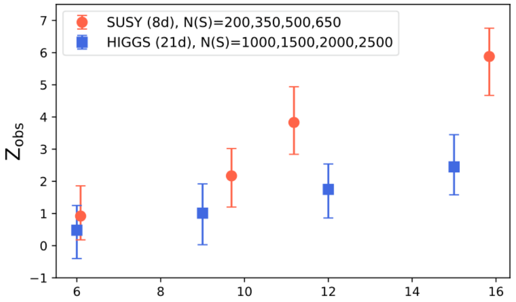


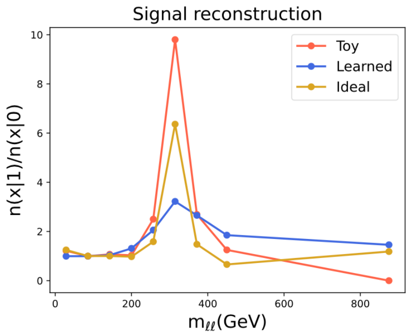




Multivariate







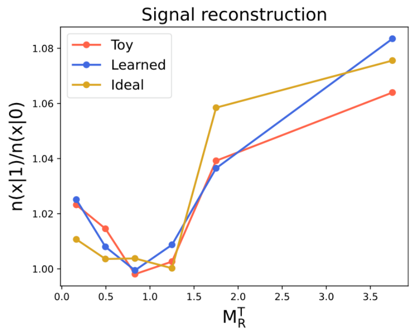


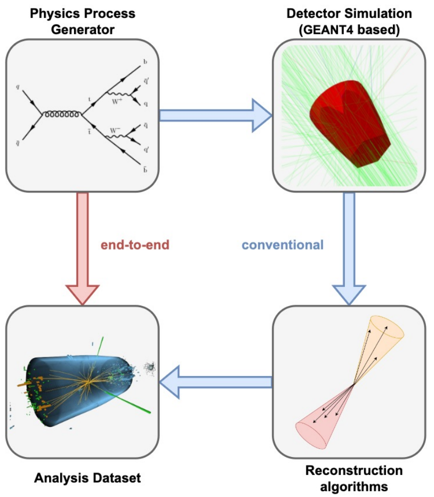
Table 1 Average training times per single run with standard deviations (low level features and reference toys). Note that time measured in hours (for NN) and seconds (for Falkon)

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)

The New Physics Learning Machine

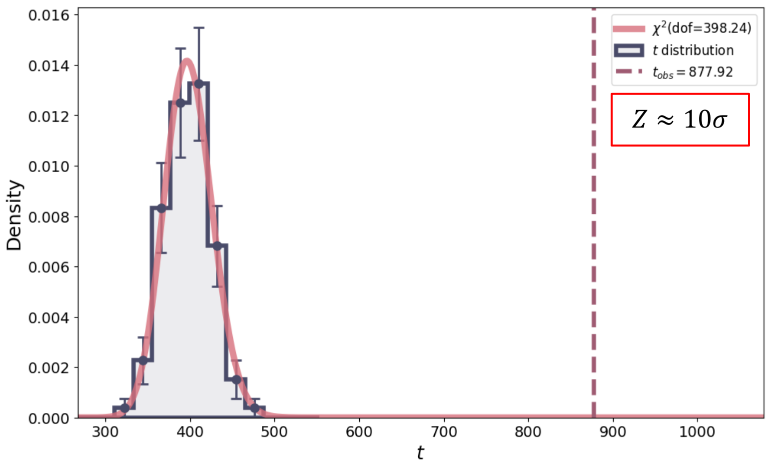
P. Cappelli, G. Grosso, ML, H. Reyes-Gonzalez, *work in progress*

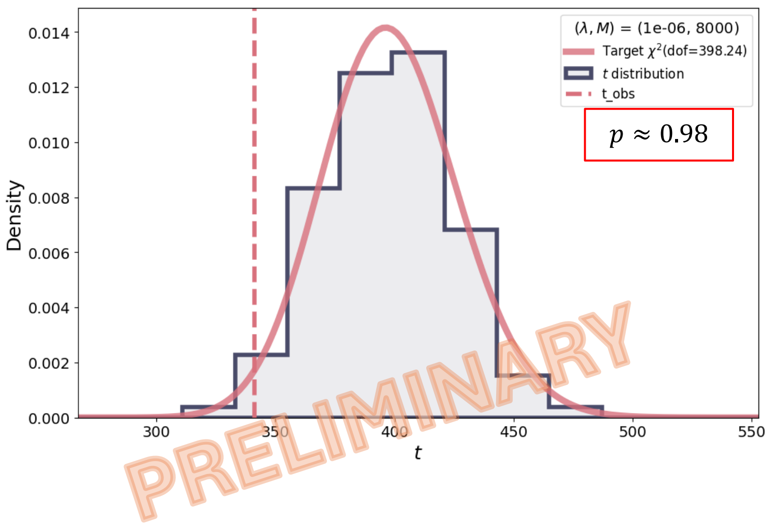


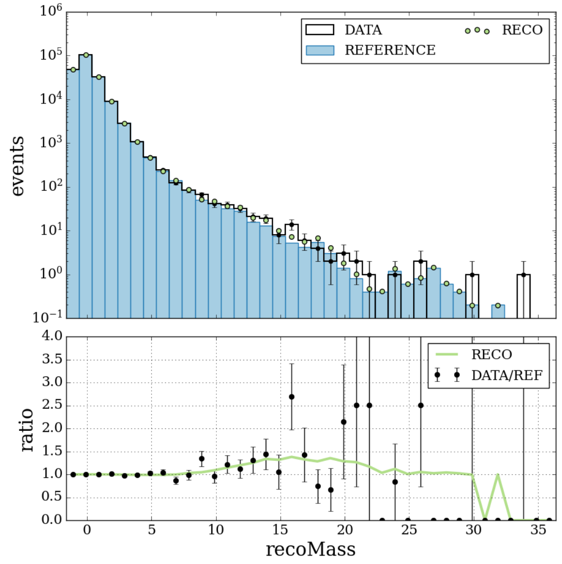
End-to-end simulations with normalizing flows: F. Vaselli, F. Cattafesta, P. Asenov, A. Rizzi [arXiv:2402.13684]

Particle jets dataset with PYTHIA + Delphes-like smearing ($pp \rightarrow t\bar{t}$).

16 features, n=1M, m=200k, $\bar{t}_{\text{training}} \approx 25$ s.







Conclusion

- Modern machine learning provides powerful methods to accelerate HEP research.
- Robust evaluation methods and uncertainty quantification is crucial for applications in precision sciences.

- We discussed nonparametric methods to evaluate generative models based on the framework of two-sample testing.
- We explored techniques that prioritise either efficiency or sensitivity, the latter based on machine learning.

https://github.com/NF4HEP

https://github.com/GaiaGrosso/NPLM_package

https://github.com/FalkonHEP

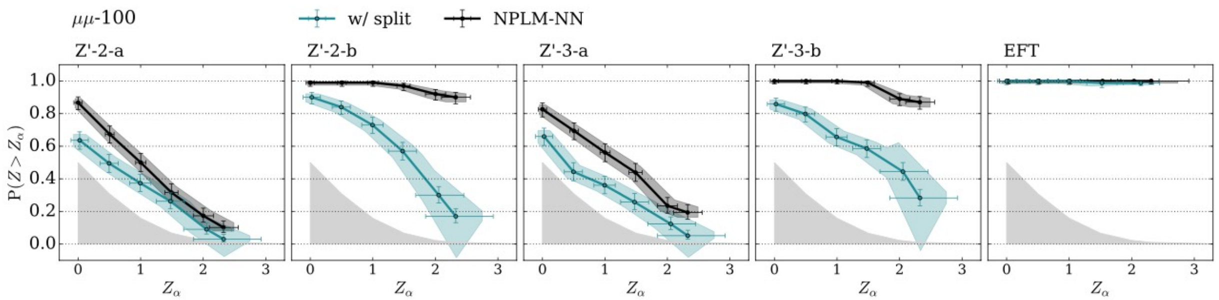
https://github.com/mletizia/FalkonNPLM_1D

THANK YOU

The New Physics Learning Machine

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

Train-test split



Different metrics