Efficient machine learning for statistical hypothesis testing

Marco Letizia – Machine Learning Genoa Center and INFN

In collaboration with: G. Grosso (IAIFI-MIT), M. Pierini (CERN), L. Rosasco (UniGe-MaLGa), A. Wulzer (IFAE), M. Zanetti (UniPd).

Foundations [1,2]

The ***New Physics Learning Machine*** is a methodology powered by machine learning to perform a likelihood-ratio goodness-of-fit test with data-driven hypotheses. The goal is to assess how well a reference model R, seen as a generator, fits a set of observations. A supervised classifier is trained on a *reference sample*

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

and a *data sample*

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

to approximate the true density ratio

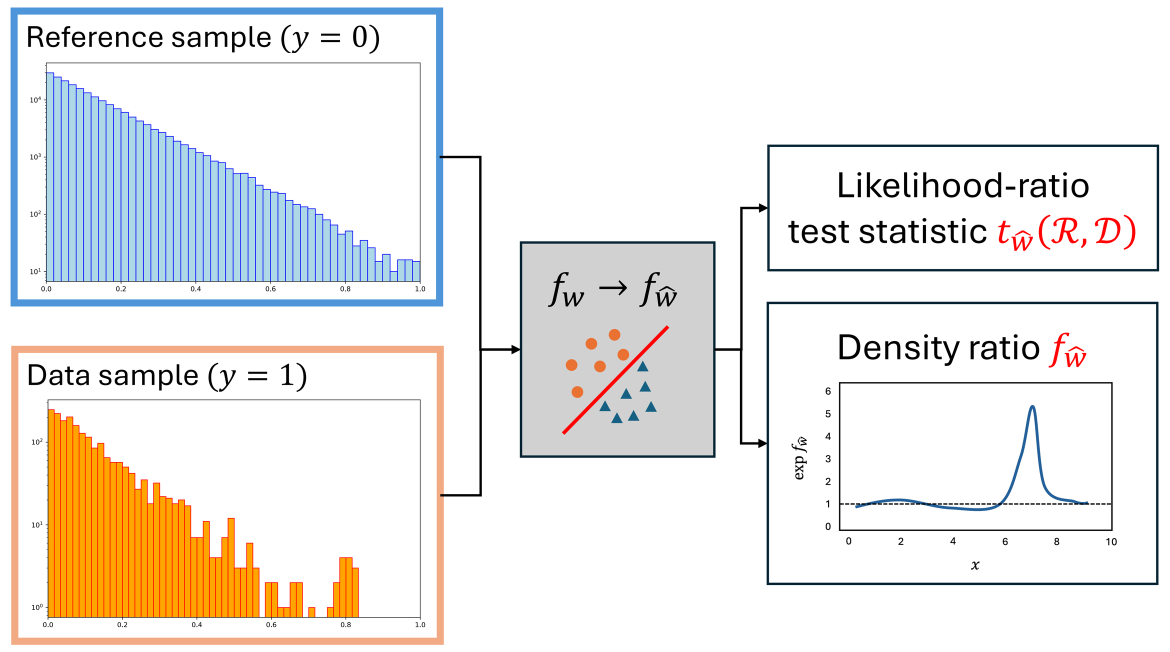
$$f_{\widehat{w}}(x) \approx f^*(x) = \log \frac{p_{\text{true}}(x)}{p(x|R)}, \qquad \widehat{w} \quad \text{learned optimal parameters.}$$

Rather than using standard metrics such as accuracy, the model is evaluated in-sample with a metric derived from the (extended) likelihood-ratio

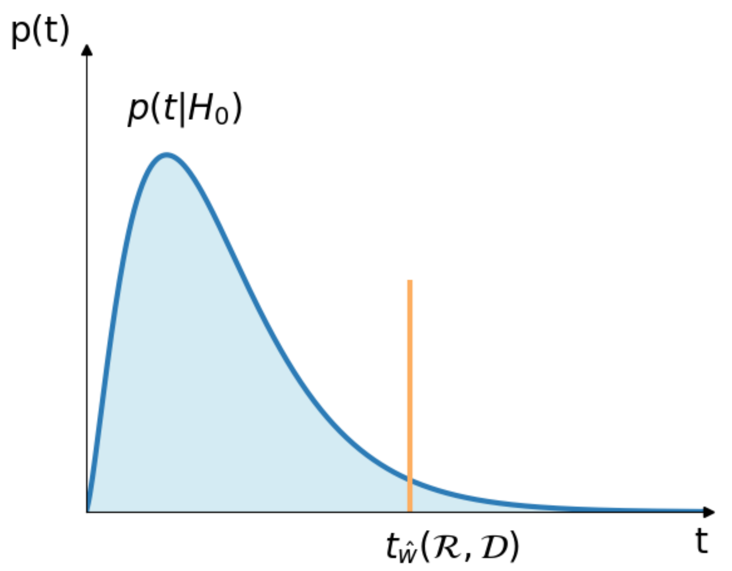
$$t_{\widehat{w}}(\mathcal{R},\mathcal{D}) = -2\left[\sum_{x\in\mathcal{R}} \frac{N(R)}{N_\mathcal{R}}\left(e^{f_{\widehat{w}}(x)} - 1\right) - \sum_{x\in\mathcal{D}} f_{\widehat{w}}(x)\right],$$

where $N(R)$ is the expected number of events under the reference model.

The test is sensitive to both *distribution and normalisation shifts*.



The null hypothesis is estimated empirically by training and evaluating NPLM on the reference model against iteself, to perform a goodness-of-fit test.



$$p_{\text{value}} = \int_{t_{\widehat{w}}(\mathcal{R},\mathcal{D})}^{\infty} dt\, p(t|H_0)$$

$$Z = \Phi^{-1}(1 - p_{\text{value}})$$

We focus on the implementation based on kernel methods and the Falkon library, highly performant while extremely efficient.[3] It is based on the (regularised) weighted logistic loss

$$\ell(y, f_w(x)) = \frac{N(R)}{N_\mathcal{R}}(1-y)\log(1+e^{f_w(x)}) + y\log(1+e^{-f_w(x)}),$$

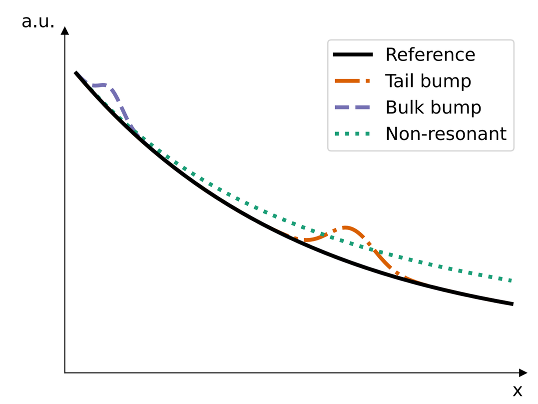
and considers functions of the following form

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

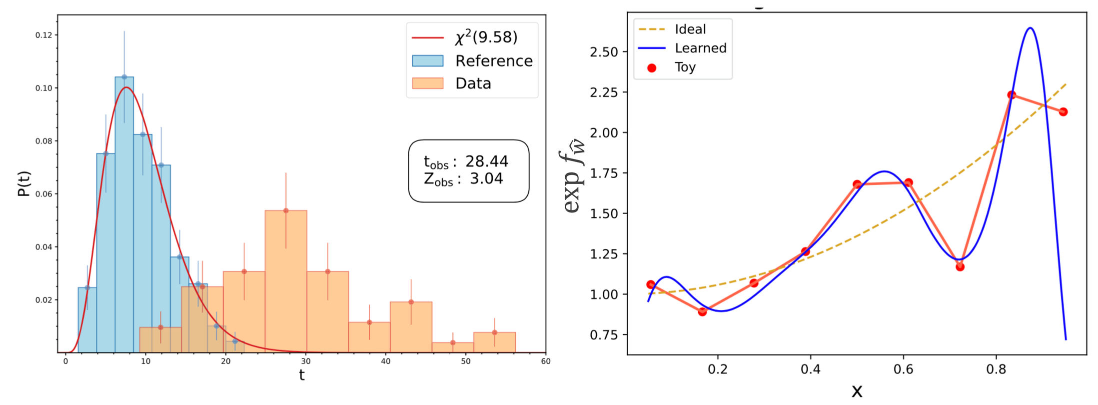
where σ is the width of the gaussian kernel, a hyperparameter.

Signal-agnostic searches [2]

1D benchmark



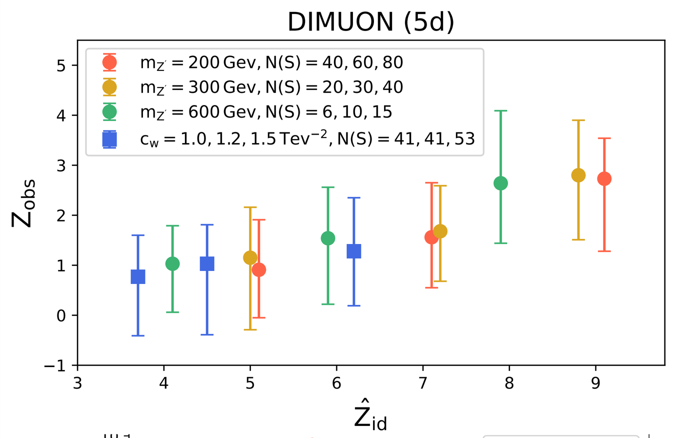
Distribution of the test statistic and signal reconstruction for the non-resonant signal.

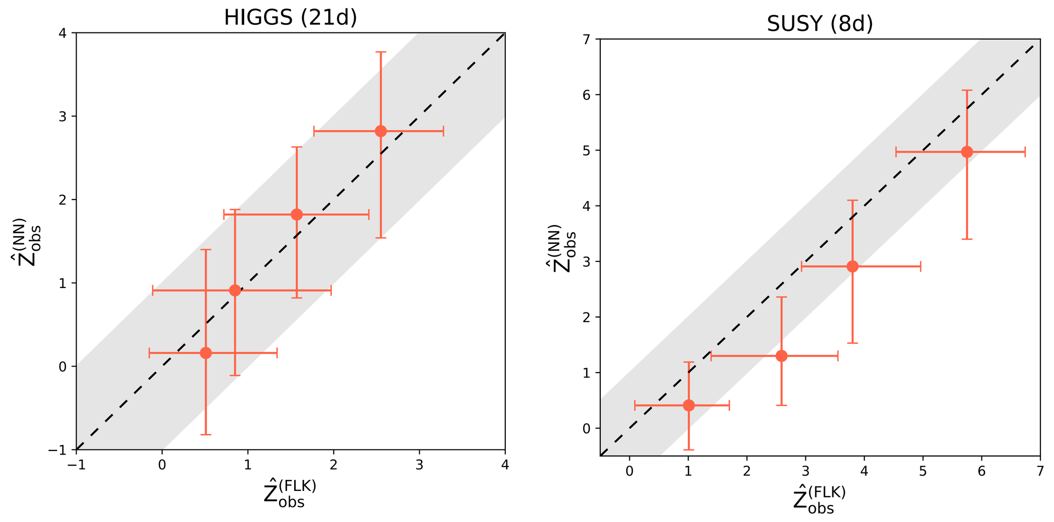


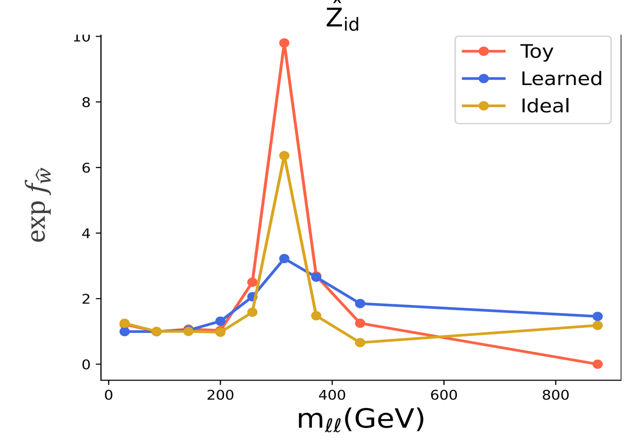
- $N_\mathcal{R} = 200k$, $N(R) = 2000$
- $N(S) = 10$ (tail), 90 (bulk/non-res)
- $\bar{t}_{training} \approx 2$ sec

Median Z	Tail	Non-res	Bulk
Z_{id}	4.7	4.4	4.1
Z_{obs}	2.4	3.0	2.8

Multivariate benchmarks







Average training times per single run (low-level features, null hypothesis).

Model	DIMUON	SUSY	HIGGS
Kernels	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

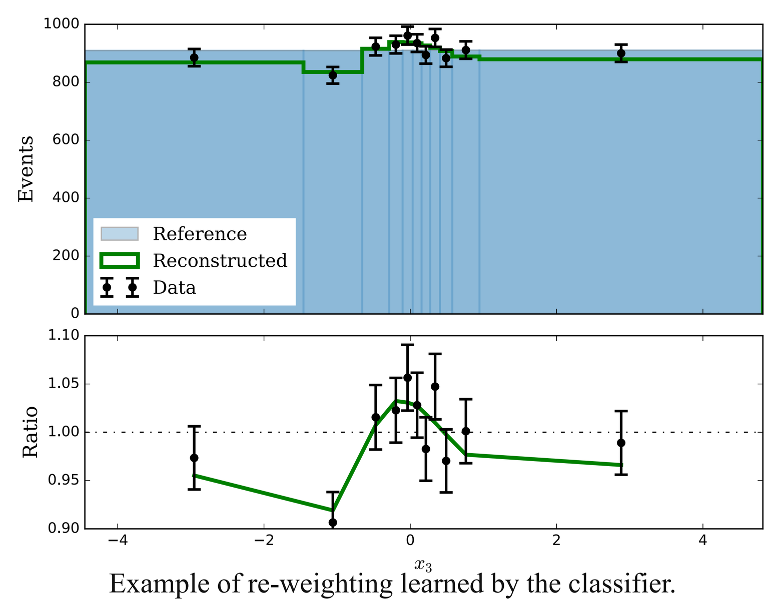
Evaluation of generative models [4]

The efficiency of the kernel-based model opens the door to several applications.

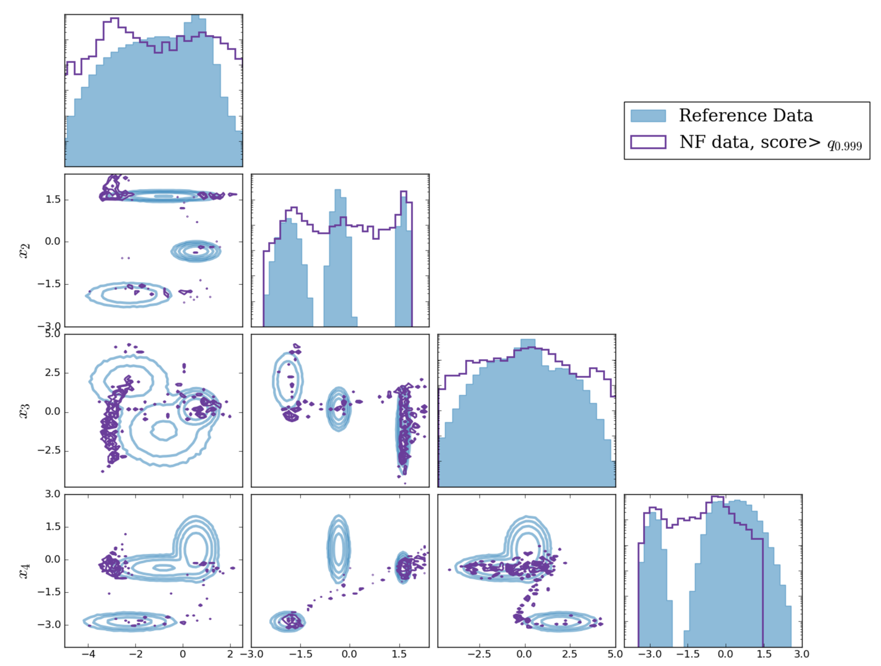
• RealNVP on correlated mixtures of Gaussians. • NPLM test : $N_\mathcal{R}$=100k; $N_\mathcal{D}$=10k.

N_{tr} \ d	4	8	12	16	20	30
100k	$9.88^{+1.22}_{-1.29}$	$8.88^{+1.12}_{-1.19}$	$14.73^{+1.23}_{-0.94}$	$16.81^{+1.04}_{-1.06}$	$14.46^{+1.09}_{-0.84}$	$14.97^{+1.09}_{-0.84}$
200k	$4.79^{+1.00}_{-1.07}$	$9.90^{+0.94}_{-1.05}$	$9.56^{+1.04}_{-1.04}$	$8.34^{+0.96}_{-1.09}$	$6.45^{+0.97}_{-1.07}$	$7.32^{+0.90}_{-0.81}$
500k	$1.93^{+1.02}_{-0.99}$	$3.01^{+0.74}_{-1.13}$	$3.16^{+1.10}_{-1.02}$	$5.05^{+1.02}_{-0.99}$	$2.07^{+0.81}_{-0.97}$	$3.06^{+1.13}_{-0.86}$

Median significance at varying dimensionality and number of training examples.



Example of re-weighting learned by the classifier.



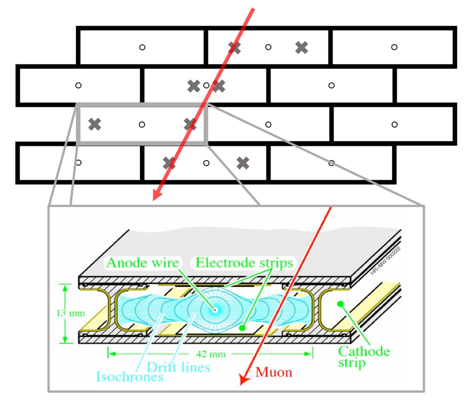
Using the classifier score to identify mismodeled correlations.

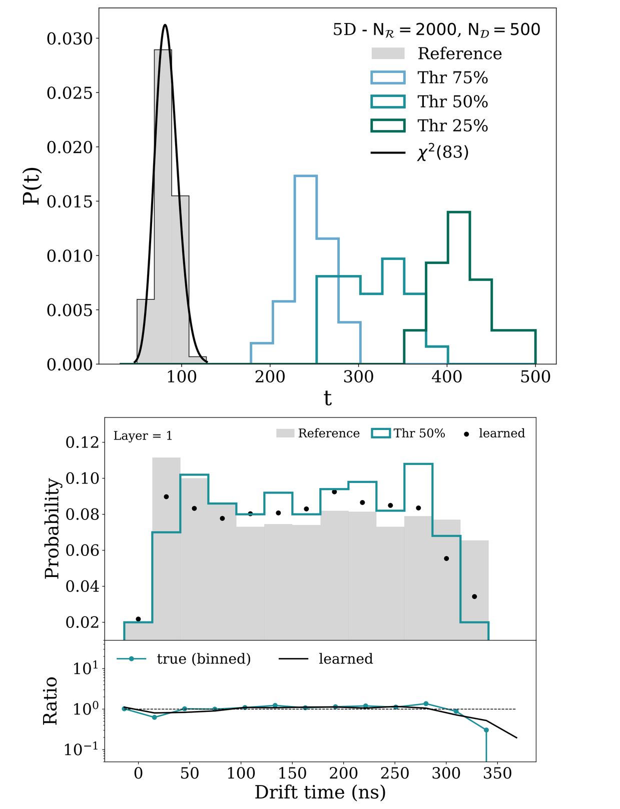
Data Quality Monitoring [5]

NPLM for monitoring particle detectors in real-time.

Reduced scale CSM drift tubes; features: 4 drift times + crossing angle; $\bar{t}_{training} \approx 0.5$ sec;

anomalies: lowered cathodic voltages and front-end thresholds.





Anomaly	NPLM (5D)	KS (t_1)	KS (t_2)	KS (t_3)	KS (t_4)	KS (ϕ)
Cathode 75%	1.1×10^{-6}	0.50	0.41	0.43	0.40	0.42
Cathode 50%	3.4×10^{-4}	0.47	0.27	0.47	0.37	0.41
Cathode 25%	0.0019	0.45	0.44	0.13	0.45	0.50
Threshold 75%	$<10^{-7}$	0.23	0.14	0.16	0.14	0.48
Threshold 50%	$<10^{-7}$	0.09	0.10	0.06	0.17	0.42
Threshold 25%	$<10^{-7}$	0.11	0.07	0.04	0.11	0.66

Median *p*-values.

Multiple testing [6]

Model selection can bias the test towards certain signal hypotheses. Multiple testing strategies can tame this effect for a more uniform response.

- min–p: $\quad p_{\min} = -\log \min_{1\le i\le n} p_i$
- avg–p: $\quad p_{\text{avg}} = -\frac{1}{n}\sum_{1\le i\le n} p_i$
- prod–p: $\quad p_{\text{prod}} = -\sum_{1\le i\le n} \log p_i$
- smax–t: $\quad t_{\text{smax}} = \log \frac{1}{n}\sum_{1\le i\le n} e^{t_i}$

Different tests are characterized by different kernel widths.

N(S)	7	18	13	10	90
$\bar{x}_{\text{NP}}$	4	4	4	6.4	1.6
σ_{NP}	0.01	0.16	0.64	0.16	0.16
$\sigma = 0.01$	0.0028 ± 0.0008	0.0010 ± 0.0006	0.0005 ± 0.0004	0.0001 ± 0.0001	0.029 ± 0.004
$\sigma = 0.3$	$\mathbf{0.012 \pm 0.002}$	0.107 ± 0.007	0.008 ± 0.002	0.246 ± 0.009	0.65 ± 0.01
$\sigma = 0.7$	0.006 ± 0.001	$\mathbf{0.123 \pm 0.007}$	$\mathbf{0.011 \pm 0.002}$	$\mathbf{0.36 \pm 0.01}$	$\mathbf{0.70 \pm 0.01}$
$\sigma = 1.4$	0.004 ± 0.001	0.078 ± 0.006	$\mathbf{0.012 \pm 0.002}$	0.29 ± 0.01	0.54 ± 0.01
$\sigma = 4.5$	0.0023 ± 0.0007	0.020 ± 0.003	$\mathbf{0.011 \pm 0.002}$	0.098 ± 0.007	0.28 ± 0.01
$\sigma = 9.0$	0.0028 ± 0.0008	0.018 ± 0.003	$\mathbf{0.012 \pm 0.002}$	0.075 ± 0.006	0.24 ± 0.01
$\sigma = 2.3$	0.0023 ± 0.0007	0.044 ± 0.005	0.013 ± 0.002	0.028 ± 0.004	0.36 ± 0.01
min-p	$\mathbf{0.008 \pm 0.001}$	$\mathbf{0.103 \pm 0.007}$	0.007 ± 0.002	$\mathbf{0.32 \pm 0.01}$	$\mathbf{0.66 \pm 0.01}$
prod-p	0.005 ± 0.001	0.083 ± 0.006	$\mathbf{0.012 \pm 0.002}$	0.26 ± 0.01	0.65 ± 0.01
avg-p	0.006 ± 0.001	0.049 ± 0.005	0.011 ± 0.002	0.068 ± 0.006	0.50 ± 0.01
smax-t	0.0028 ± 0.0008	0.0010 ± 0.0006	0.0005 ± 0.0004	0.0001 ± 0.0001	0.029 ± 0.004

Table 1: **EXPO 1D –** probability of observing $Z \geq 3$.

test	Z′ M = 180 GeV W = 0.02 GeV	Z′ M = 300 GeV W = 15 GeV	Z′ M = 600 GeV W = 30 GeV	EFT $c_w = 1.5 \times 10^{-6}$
$\sigma = 0.31$	0.007 ± 0.003	0.004 ± 0.002	0.0010 ± 0.0008	0.0010 ± 0.0008
$\sigma = 1.19$	$\mathbf{0.096 \pm 0.009}$	0.10 ± 0.01	0.006 ± 0.002	0.017 ± 0.004
$\sigma = 1.79$	0.065 ± 0.008	0.11 ± 0.01	0.012 ± 0.003	0.026 ± 0.005
$\sigma = 2.49$	0.036 ± 0.006	0.11 ± 0.01	0.027 ± 0.005	0.053 ± 0.007
$\sigma = 4.23$	0.037 ± 0.006	$\mathbf{0.13 \pm 0.01}$	$\mathbf{0.066 \pm 0.008}$	0.13 ± 0.01
$\sigma = 8.0$	0.023 ± 0.004	0.068 ± 0.008	0.056 ± 0.007	$\mathbf{0.22 \pm 0.01}$
$\sigma = 3.0$	0.031 ± 0.005	0.13 ± 0.01	0.044 ± 0.006	0.092 ± 0.009
min-p	0.065 ± 0.008	0.16 ± 0.01	$\mathbf{0.057 \pm 0.007}$	$\mathbf{0.23 \pm 0.01}$
prod-p	0.089 ± 0.009	$\mathbf{0.18 \pm 0.01}$	0.028 ± 0.005	0.083 ± 0.009
avg-p	$\mathbf{0.14 \pm 0.01}$	0.15 ± 0.01	0.035 ± 0.006	0.098 ± 0.009
smax-t	0.007 ± 0.003	0.004 ± 0.002	0.0010 ± 0.0008	0.0010 ± 0.0008

Table 3: **MUMU 5D –** Probability of observing $Z \geq 3$.

[1] G. Grosso, M. Letizia, M. Pierini, A. Wulzer, SciPost Physics 2024, 2305.14137; [2] M. Letizia, G. Losapio, M. Rando, G. Grosso, A. Wulzer, M. Pierini, M. Zanetti, L. Rosasco, EPJC 2022, 2204.02317; [3] G. Meanti, L. Carratino, L. Rosasco, A. Rudi, NeurIPS 2020, 2006.10350;
[4] P. Cappelli, G. Grosso, M. Letizia, M. Zanetti, *in preparation*; [5] G. Grosso, N. Lai, M. Letizia, J. Pazzini, M. Rando, L. Rosasco, A. Wulzer, M. Zanetti, MLST 2023, 2303.05413; [6] G. Grosso, M. Letizia, 2408.12296.