

Learning to validate generative models

Humberto Reyes-González

with P. Cappelli G. Grosso, M.
Letizia, M. Zanetti.

[arXiv:2511.09118](https://arxiv.org/abs/2511.09118)

Motivation

In the Gen AI era in particle physics...

How *good* are our ML surrogates with respect to the truth?

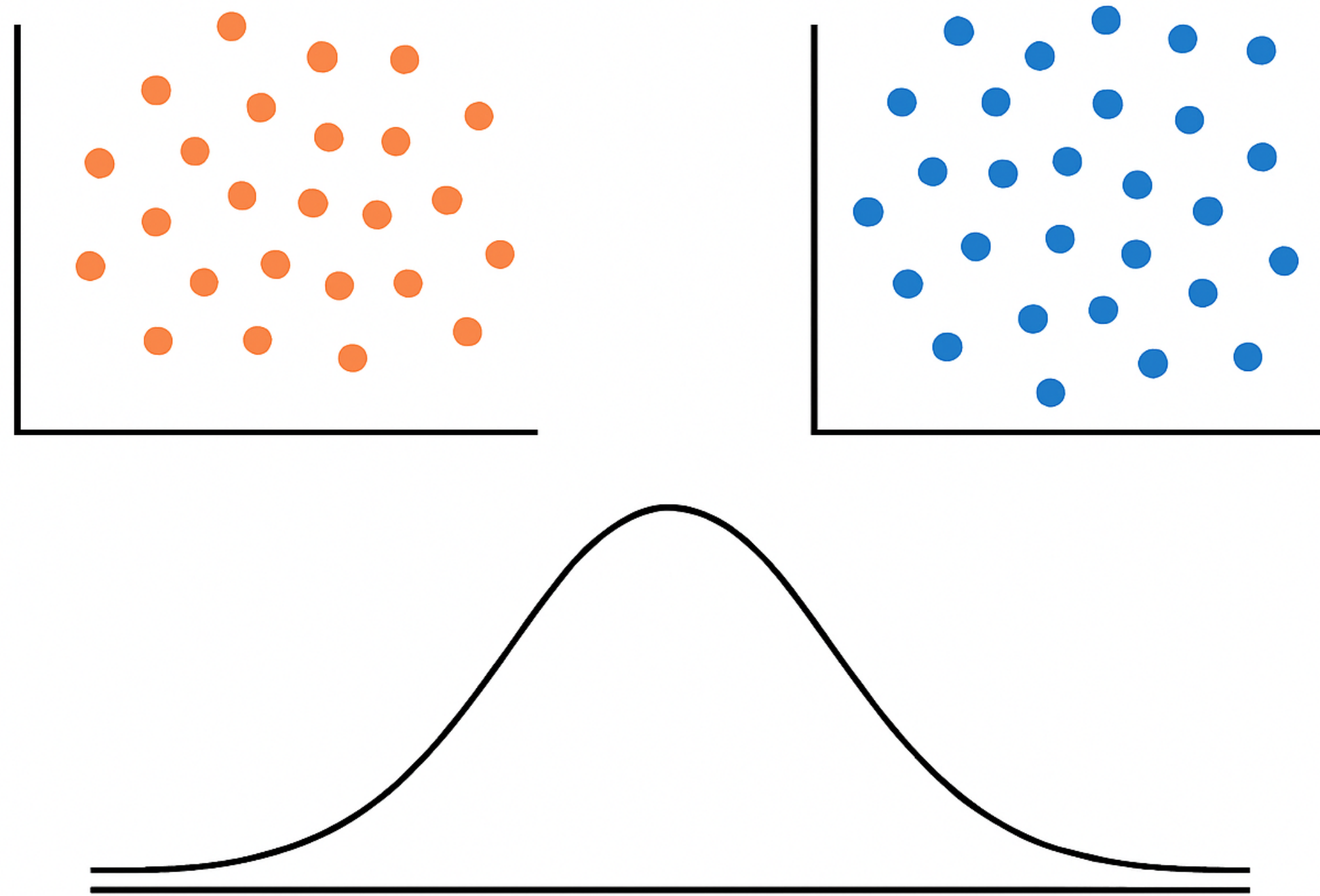
how far are we from *fully* learning HEP data?

Validation of generative networks with hypothesis testing.



Introduction

How likely is that the *true* and *generated* samples come from the same distribution?



How do I determine this if data is high dimensional?

Two-sample hypothesis testing: 3 Simple steps

Choose a multivariate **test statistic**. $t : \mathcal{X}^{N_{\mathcal{R}}} \times \mathcal{X}^{N_{\mathcal{D}}} \rightarrow \mathbb{R}$.

A large t means a tension with the **null hypothesis**. But how *significant* it it?

Construct the distribution under the null hypothesis $p(t|H_0)$ using one of the datasets as a reference.

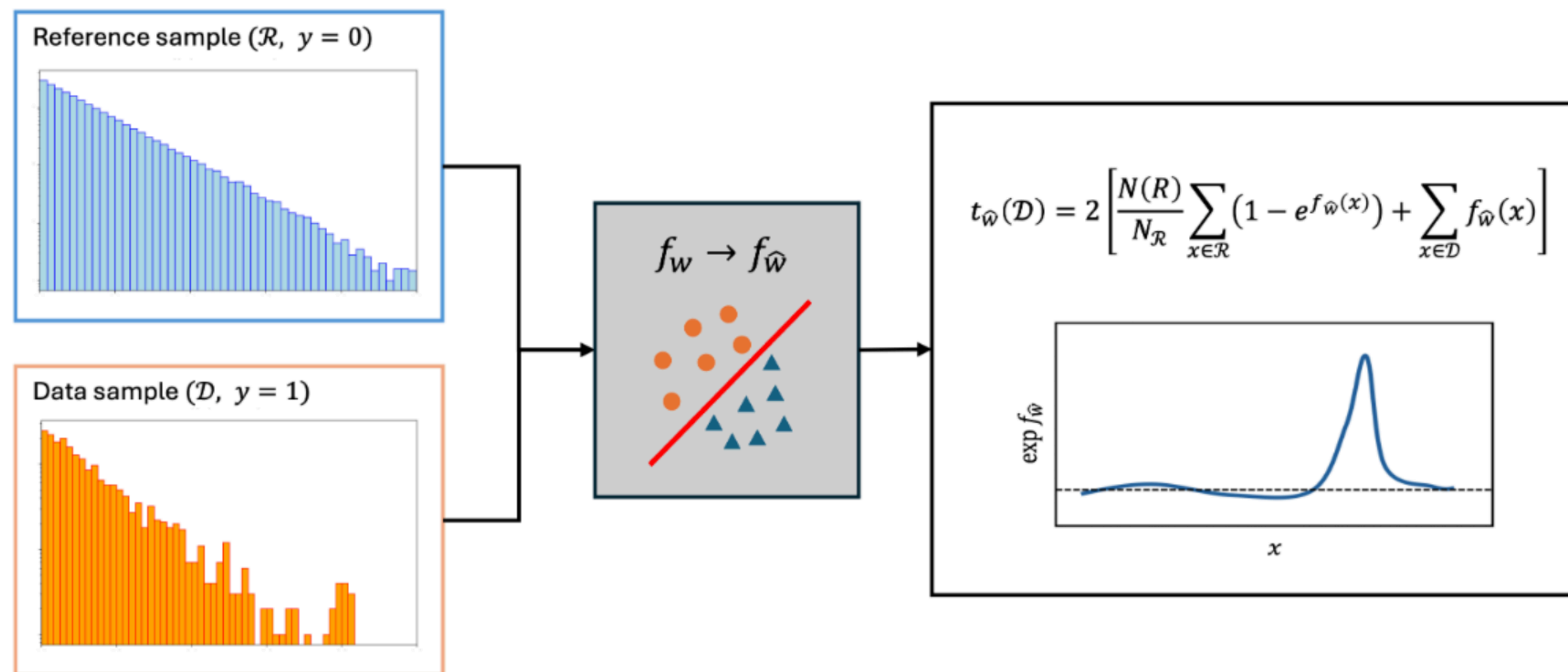
Compute the **p-value** and the corresponding **Z-score** of the **alternative hypothesis**.

$$p_{\text{value}} = P(t \geq t_{\text{obs}}|H_0) = \int_{t_{\text{obs}}}^{\infty} p(t|H_0)dt, \quad Z_{\text{score}} = \Phi^{-1}(1 - p_{\text{value}})$$

We need efficient and sensitive metrics -> We have the NPLM

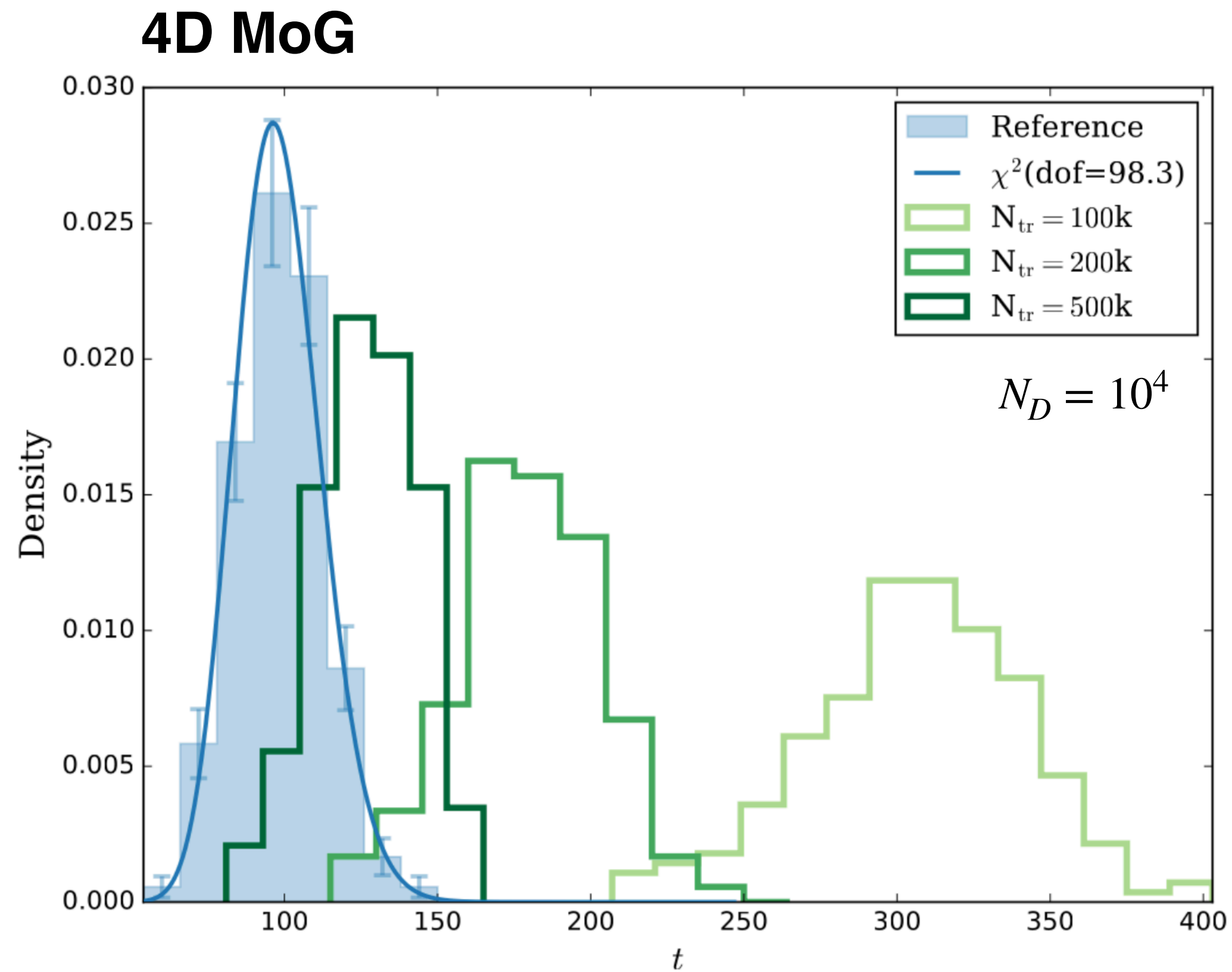
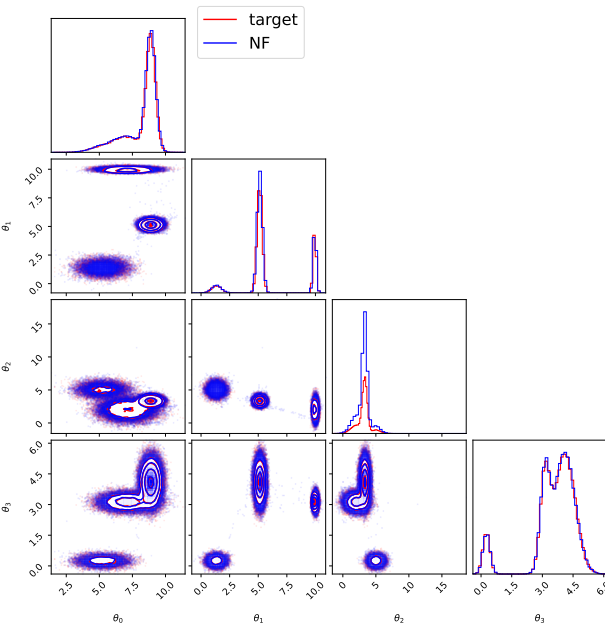
The NPLM framework

- The New Physics Learning Machine (NPLM) was originally developed for discovering new physics.
[arXiv:1806.02350](#)
- **Idea:** Model q as a deformation of $q(x) = e^{f_w(x)} p_R(x) \rightarrow$ **machine learn** $f_w(x)$.
- **Fast:** learn $f_w(x)$ with a *kernel classifier using the Nyström approximation*. [arXiv:2204.02317](#), [arXiv:2303.05413](#)

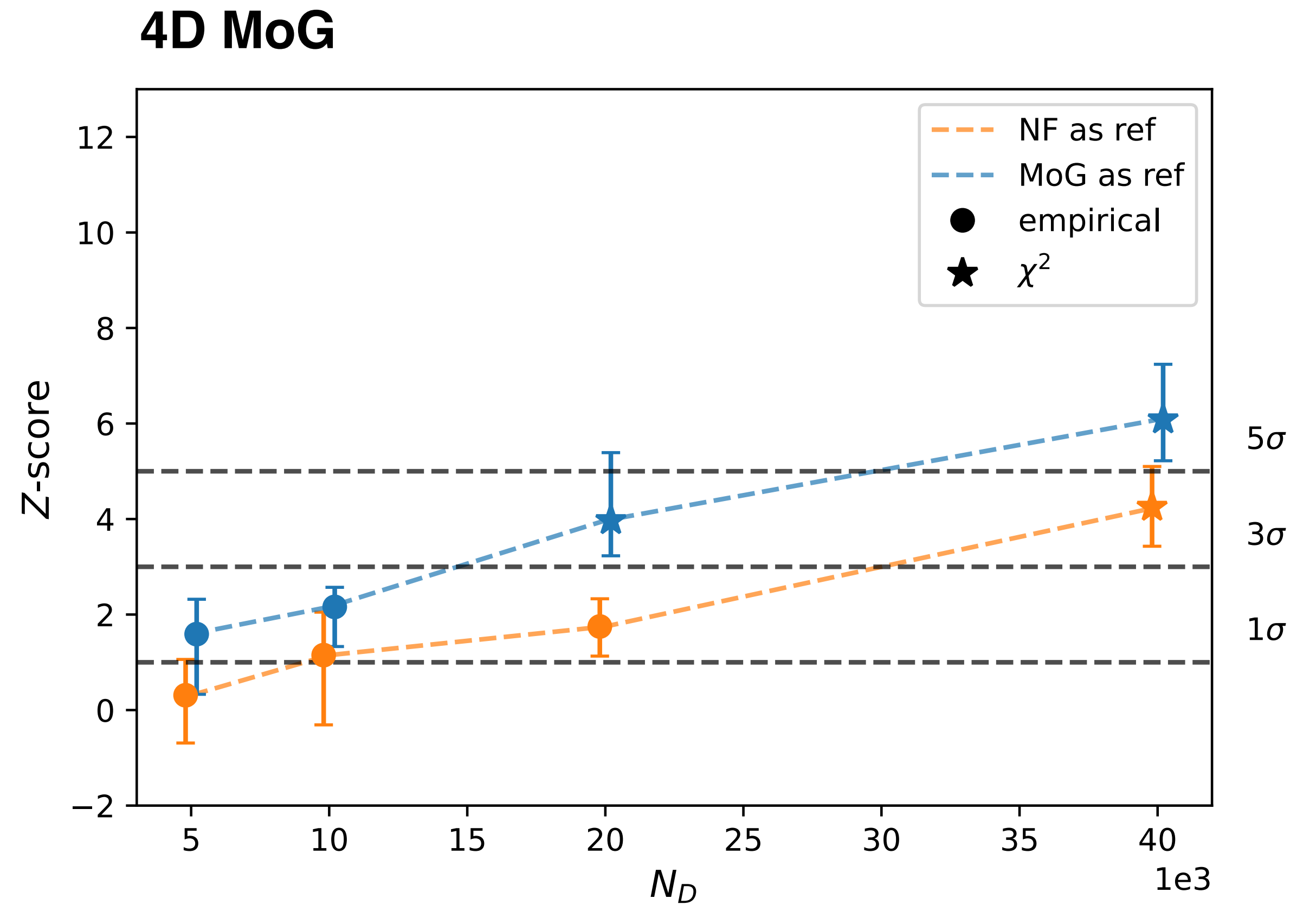


Mixture of Gaussians vs Normalising Flows

- Ref sample size: 100k



More data used in training, more precise generative networks.

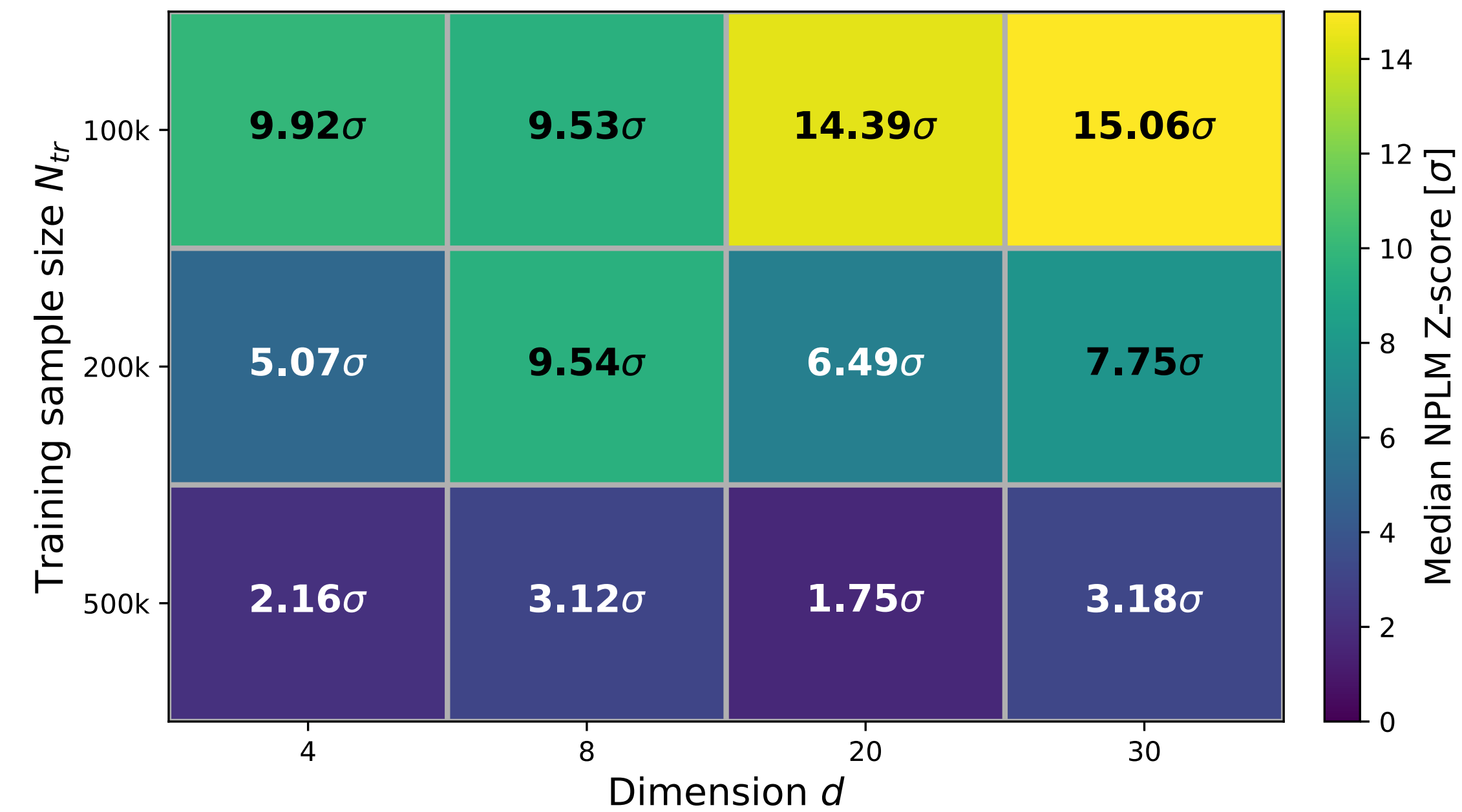


More data used in testing, more sensitive tests.

NPLM vs MMD

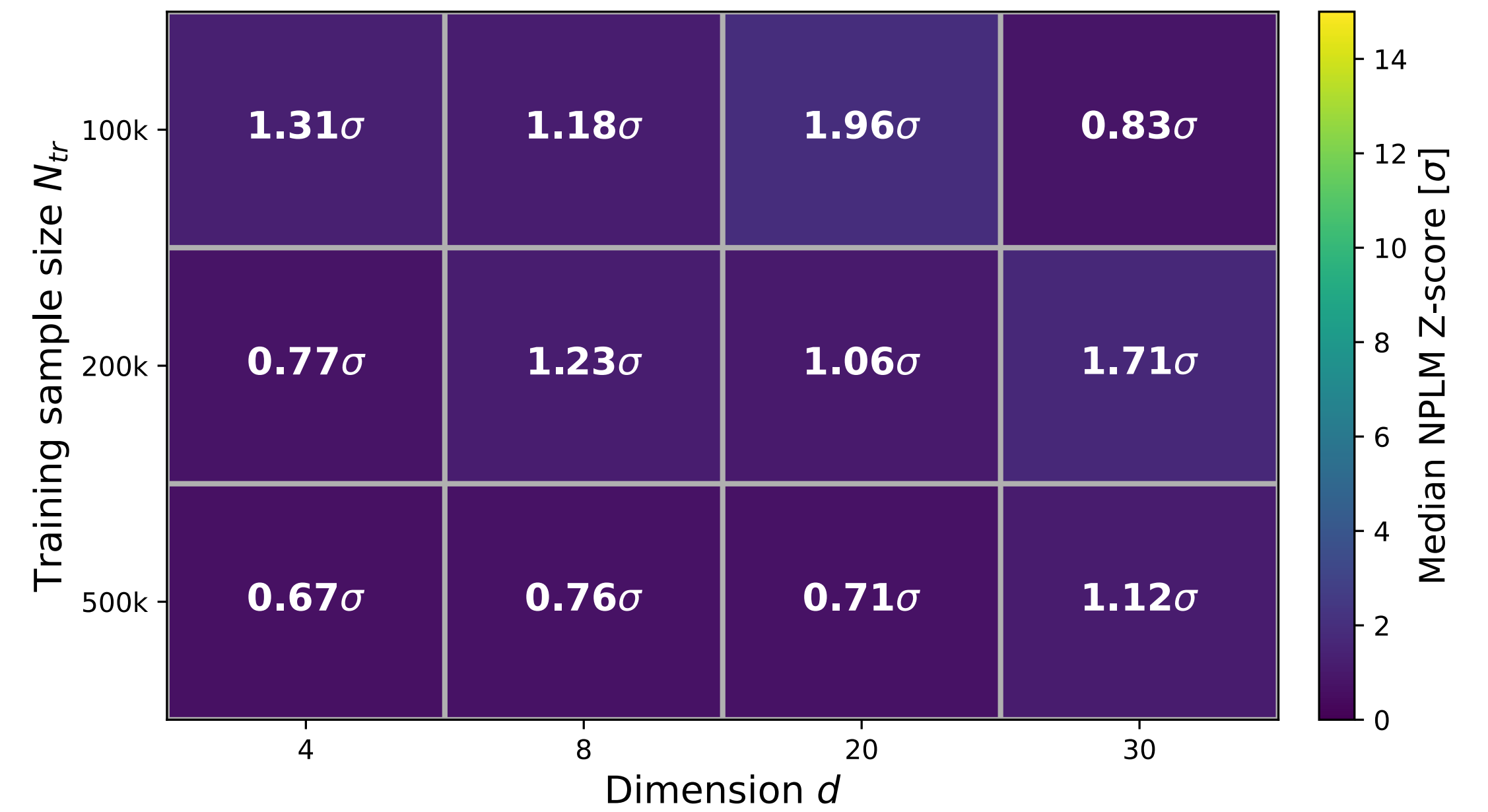
NPLM

True distribution as reference



MMD

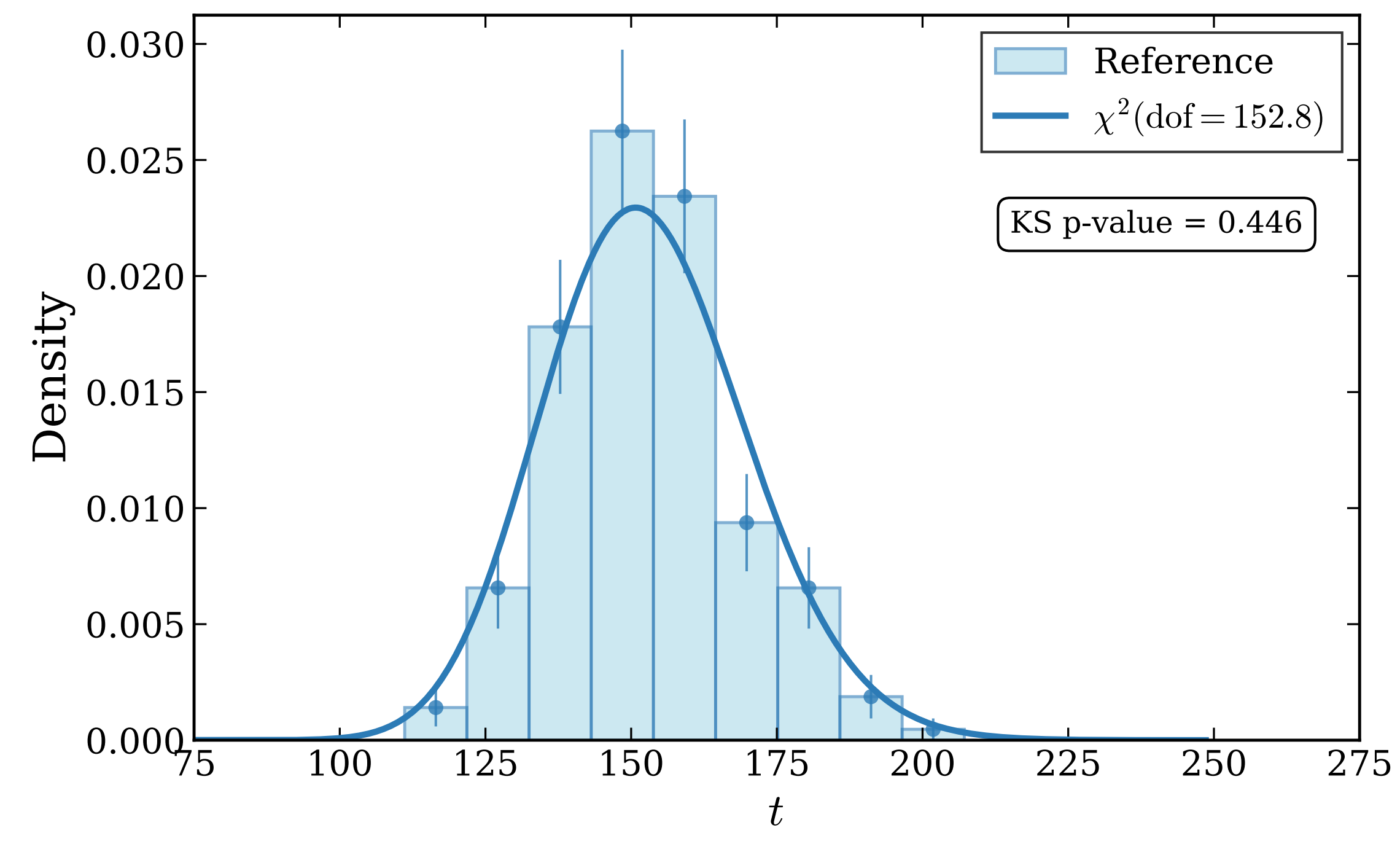
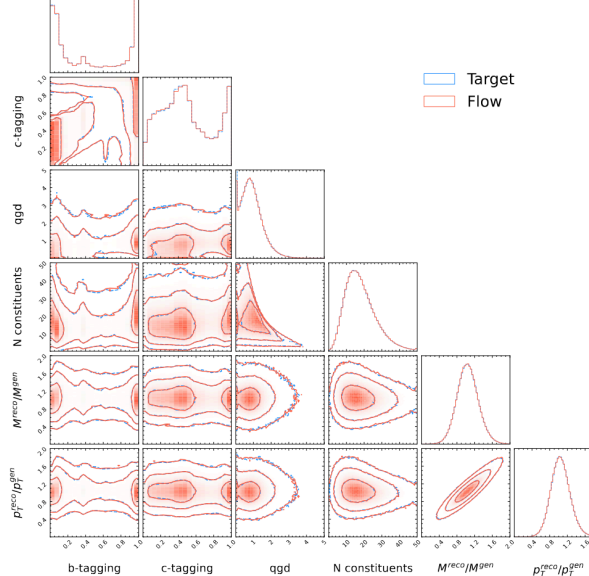
Comparison to MMD (ref= true)



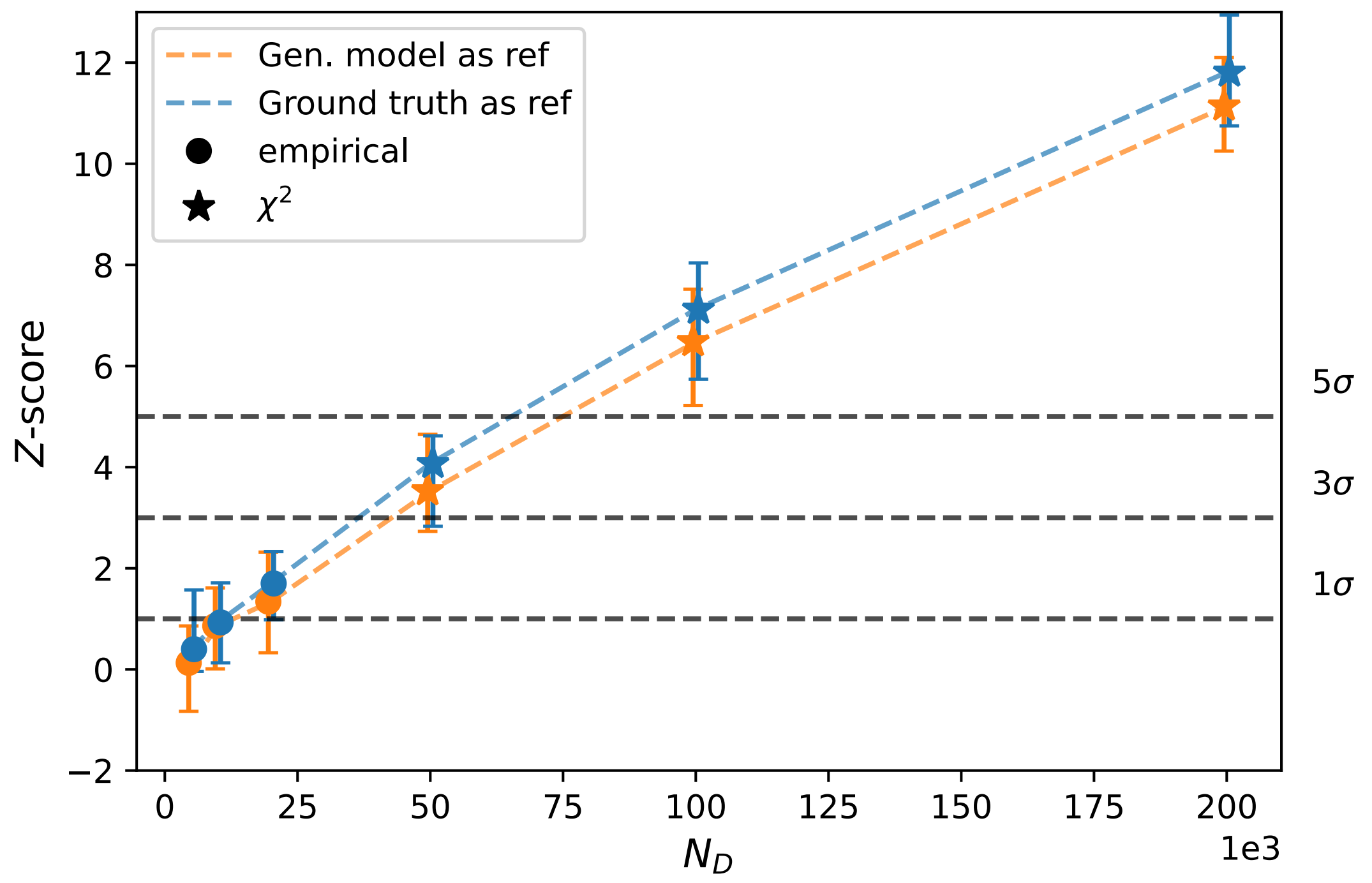
The NPLM is clearly more sensitive.

Simulated jet observables (16D) vs Flow matching

Data and ML model from DOI 10.1088/2632-2153/ad563c



Well fitted χ^2 approximation.



Great sensitivity to jet observables.

Conclusions.

- ✓ Multivariate hypothesis testing can be done for generative network validation in high dimensions.
- ✓ ML-based classifiers can be used for efficient and sensitive multivariate hypothesis testing.
- ✓ The framework is generalizable to basically any type of data and test statistic.
- ✓ Stay tuned for GENOVA: A general framework for statistical validation of generative networks in HEP.



THANK YOU!



See you at **122** !

BACK UP

Metrics timing comparison

Data	t_{SW}	$t_{\overline{\text{KS}}}$	t_{SKS}	t_{FGD}	t_{MMD}	t_{NPLM}
CG ($d = 20, n = 20k$)	0.028	0.011	0.040	0.12	0.23	4.4
MoG ($d = 20, n = 50k$)	0.034	0.016	0.051	0.28	1.4	6.2
JetNet (jet features, $n = 20k$)	0.27	0.32	0.87	0.50	0.95	4.5
JetNet (particle features, $n = 20k$)	0.27	0.35	0.86	1.3	0.86	6.2

Table 4: Mean evaluation time in seconds for a single test (null hypothesis).

The kernel classifier

$$f_w(x) = \sum_{i=1}^N w_i k_\sigma(x, x_i)$$

$$k_\sigma(x, x') = \exp \left(-\|x - x'\|^2 / 2\sigma^2 \right)$$

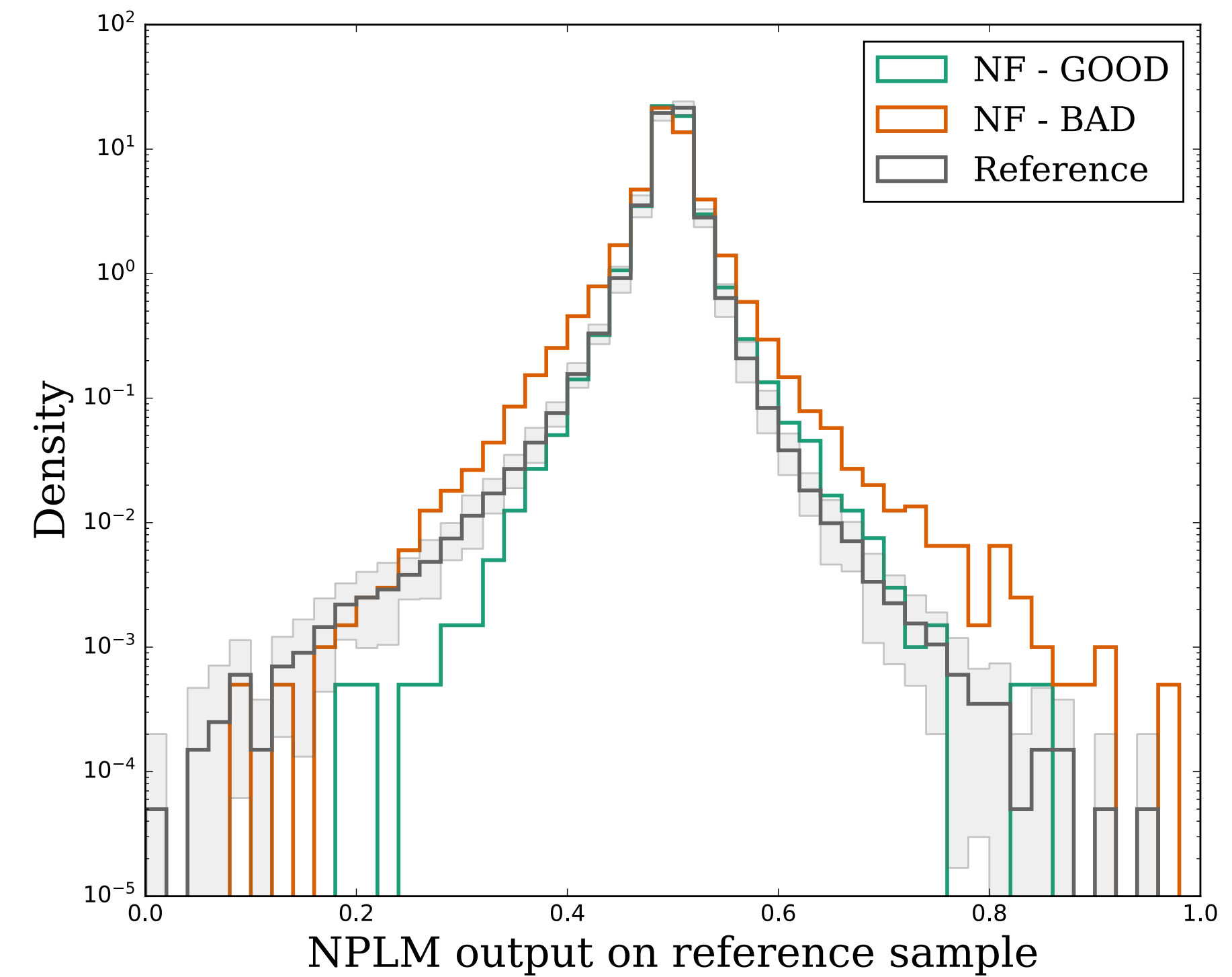
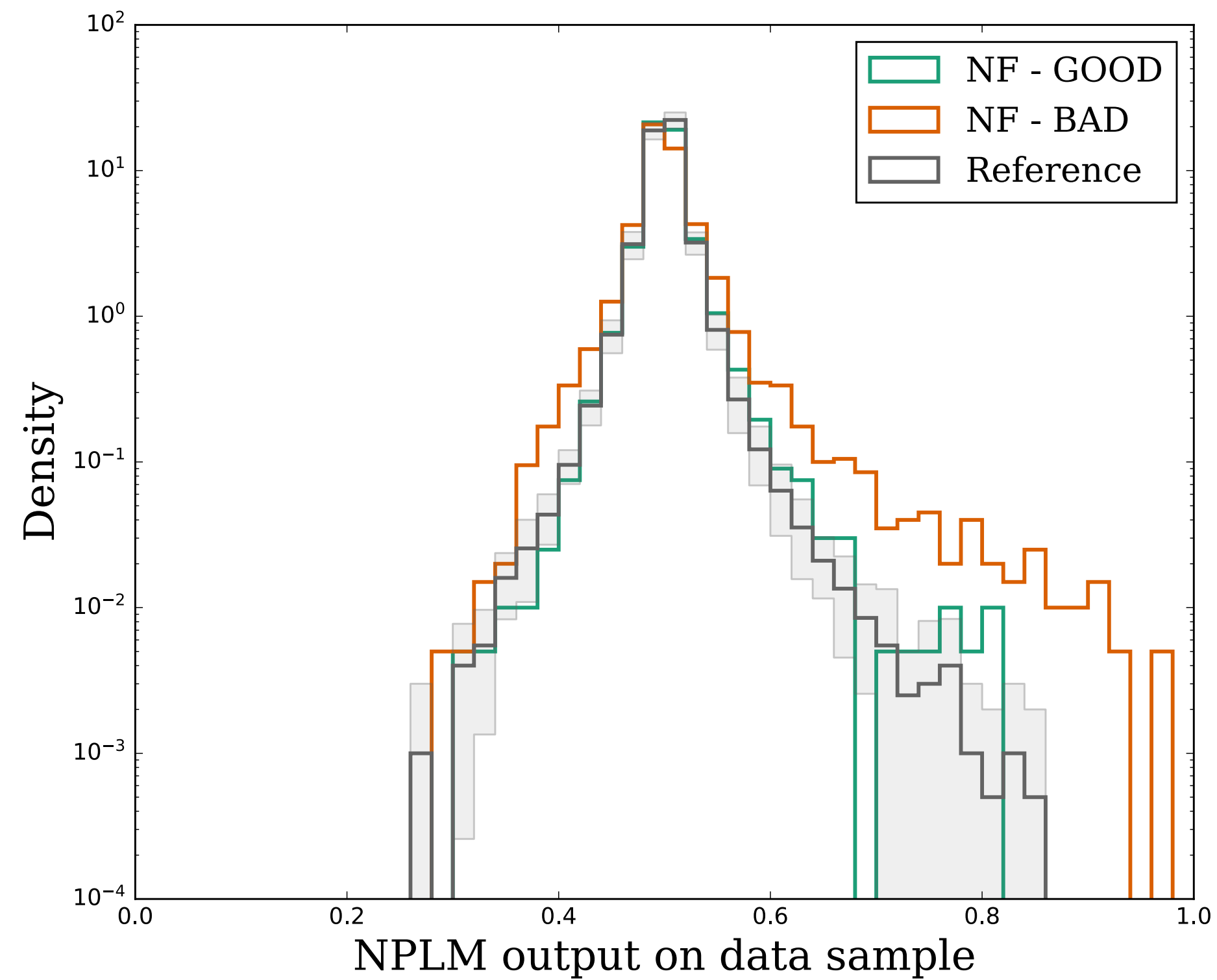
Metrics timing comparison

Data	t_{SW}	$t_{\overline{\text{KS}}}$	t_{SKS}	t_{FGD}	t_{MMD}	t_{NPLM}
CG ($d = 20, n = 20k$)	0.028	0.011	0.040	0.12	0.23	4.4
MoG ($d = 20, n = 50k$)	0.034	0.016	0.051	0.28	1.4	6.2
JetNet (jet featues, $n = 20k$)	0.27	0.32	0.87	0.50	0.95	4.5
JetNet (particle featues, $n = 20k$)	0.27	0.35	0.86	1.3	0.86	6.2

Table 4: Mean evaluation time in seconds for a single test (null hypothesis).

NPLM Diagnostics (MoGs)

Probing anomalous points



NPLM hyperparameters (MoGs).

Three hyper parameters needs to be chosen.

M = the number of centres

λ = regularisation factor

σ = Gaussian kernel width

NPLM hyperparameters (MoGs).

The chosen three hyperparameters are: $M = 10^4$ and $\lambda = 10^{-10}$, while σ depends on the MoG dimension d and is given by the following values $\sigma = 4.96, 6.08, 9.69, 10.9$ for $d = 4, 8, 20, 30$. respectively. The test is performed with $N_R = 10^5$ reference points.

NPLM hyperparameters (Jet features).

The size of the reference sample is fixed at 1 million events. The size of the generated data varies between 5,000 to 200,000 events. The hyperparameters of the NPLM are $M = 8000$, $\lambda = 10^{-6}$, and $\sigma = 7.4$. Each estimation is performed with $N_{\text{toys}} = 200$ toy datasets.

Having a limited number of reference data, each training is performed on a reference sample and a toy dataset bootstrapped from the reference data.

Fullsim parameters.

- p_T , η , ϕ ,
- invariant mass
- b-tagging discriminator
- number of constituents
- neutral hadron fraction (nhf),
- charged hadron fraction (chf)
- neutral electromagnetic fraction (nef)
- charged electromagnetic fraction (cef)
- quark-gluon discriminator (qgd)
- jet identification (JetId)
- number of charged and neutral particles
- c-tagging discriminator and number of secondary. vertices (nSV).