Theoretical topics on Anomaly Detection

Marco Letizia

University of Genova and INFN Genova

About me

Theoretical physicist based at MaLGa@University of Genova & INFN Genova

- Statistical tests for model-agnostic searches

- Generative models in HEP and their validation

Frequent collaborators:

G. Grosso (MIT), M. Pierini (CERN), L. Rosasco (U. of Genova) R. Torre (INFN Genova), A. Wulzer (IFAE), M. Zanetti (U. of Padova).

New physics beyond the Standard Model

Dark matter; Strong CP problem; Matter/anti-matter asymmetry; Neutrino masses; Hierarchy problem; Flavor puzzle; ...

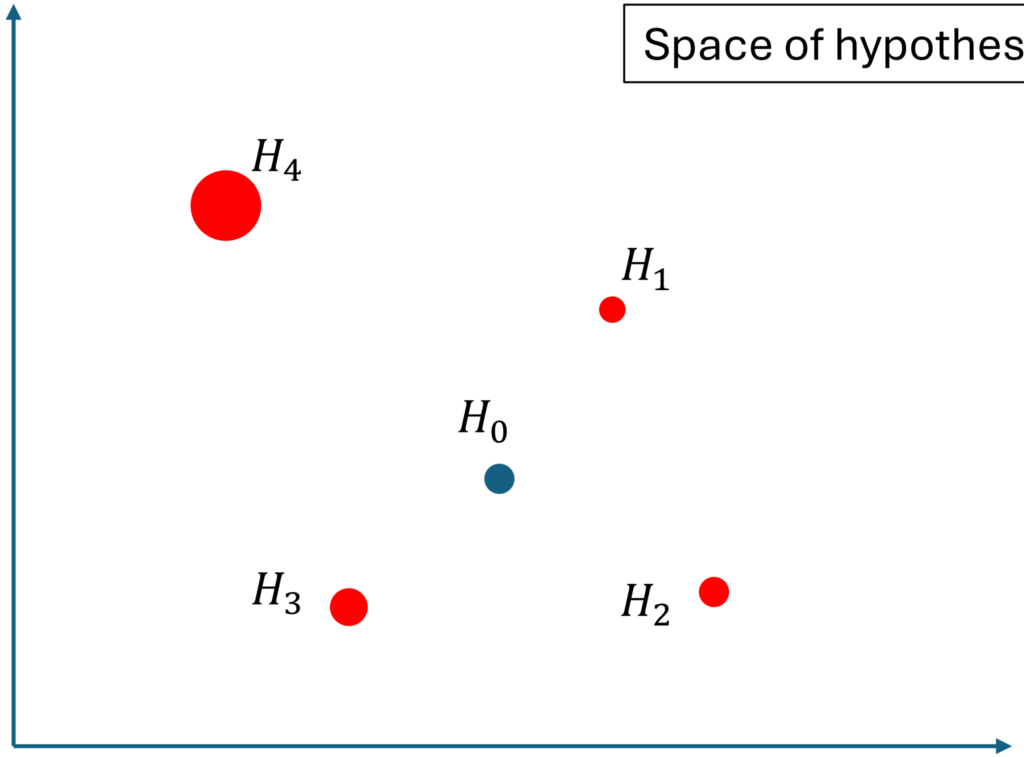
Thousands of LHC searches mostly targeting **specific models**

→ no observation of physics beyond the SM

What if the correct BSM scenario has not been formulated?

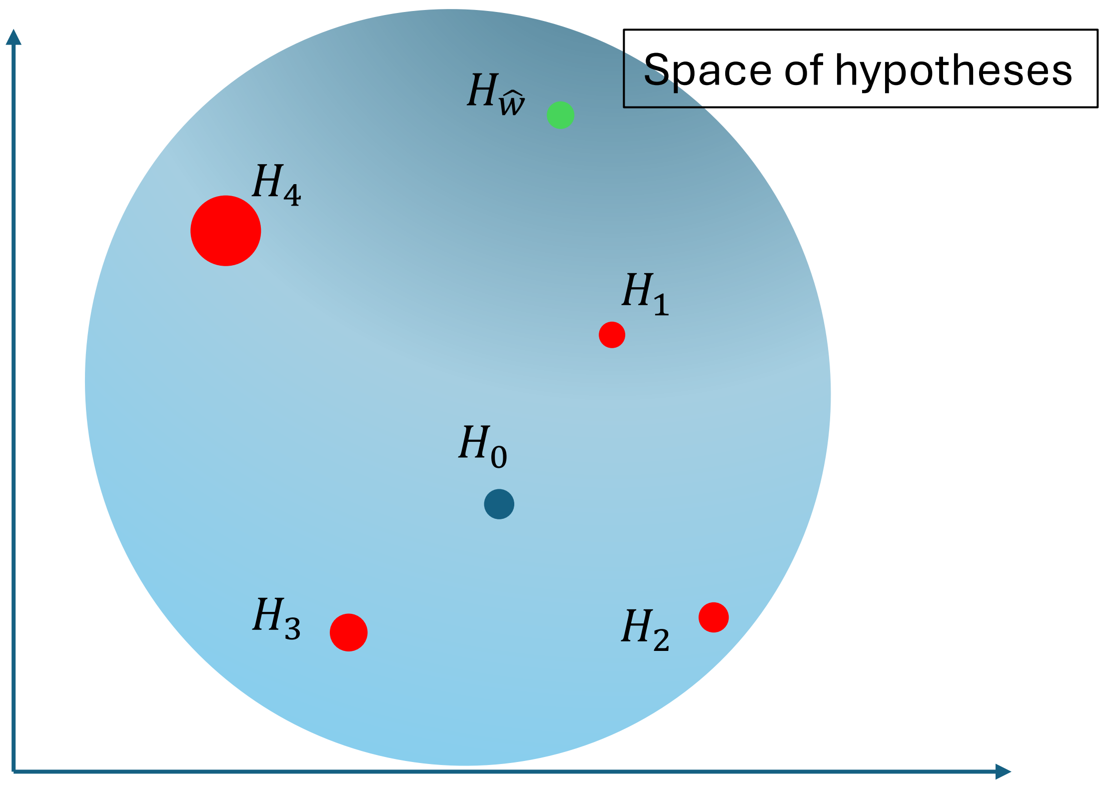
New physics **beyond the Standard Model**

Design **model-agnostic strategies** to maximize the LHC discovery potential



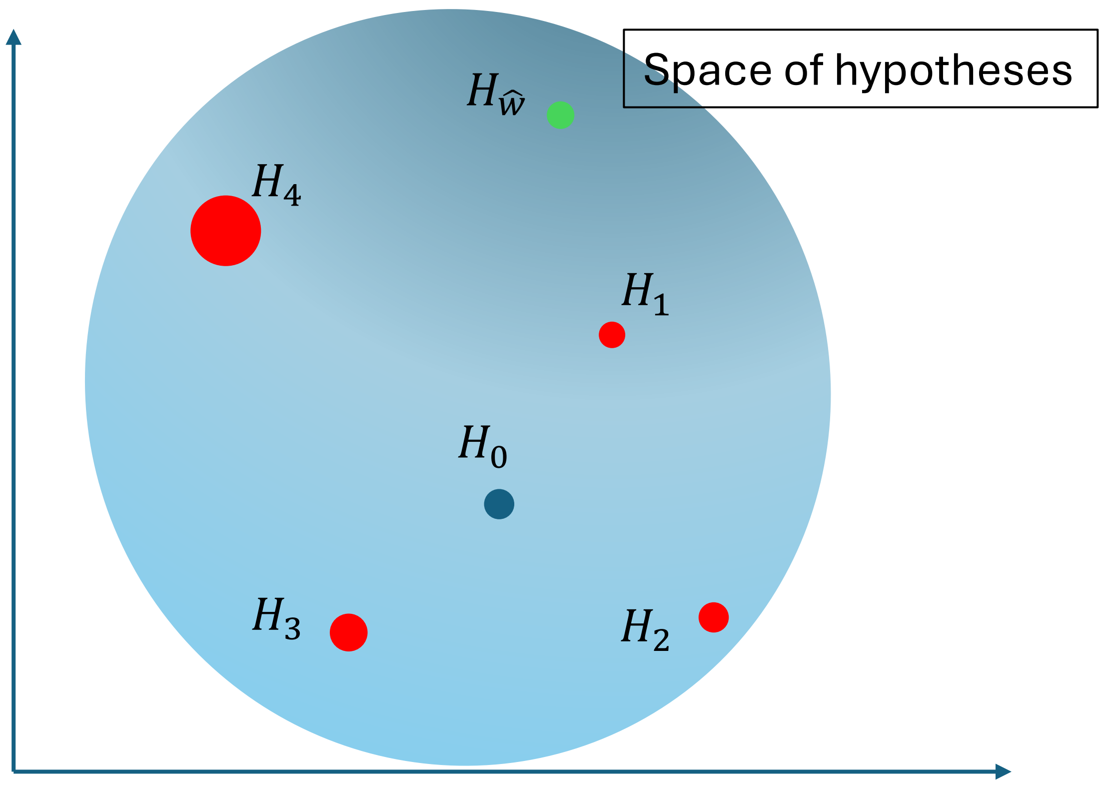
New physics beyond the Standard Model

Design **model-agnostic strategies** to maximize the LHC discovery potential



New physics beyond the Standard Model

Design **model-agnostic strategies** to maximize the LHC discovery potential



Hard problem:
- NP is a tiny and/or hidden effect
- Affecting few (unknown) observables over ∞ many
- High dimensional problem
- Hard to simulate

Partial model independence:
- Simplified models
- Effective field theories
- Bump hunts

ML-driven model independence:
- Interesting phase space regions
- ~~Phenomenological modeling~~

Model-agnostic strategies

Anomaly detection

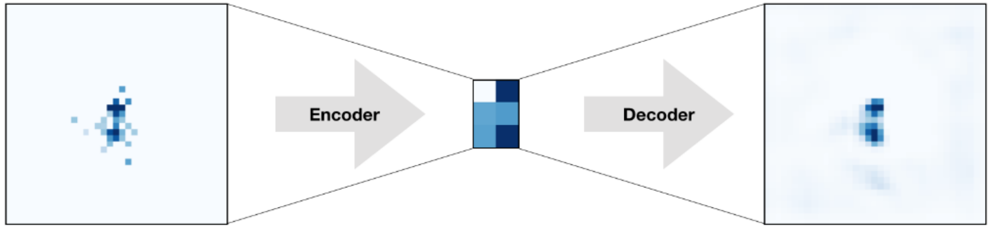
Goodness-of-fit

Anomaly detection

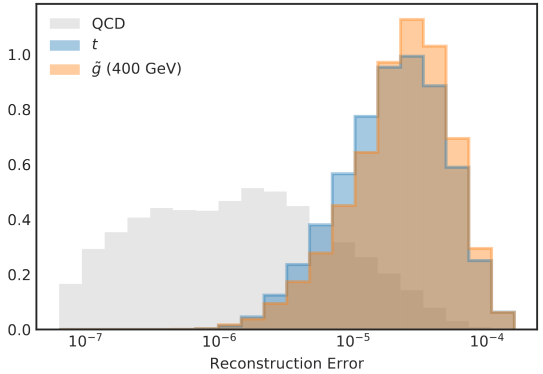
Identify interesting feature space regions by cutting on an anomaly score

Outlier detection

Autoencoders

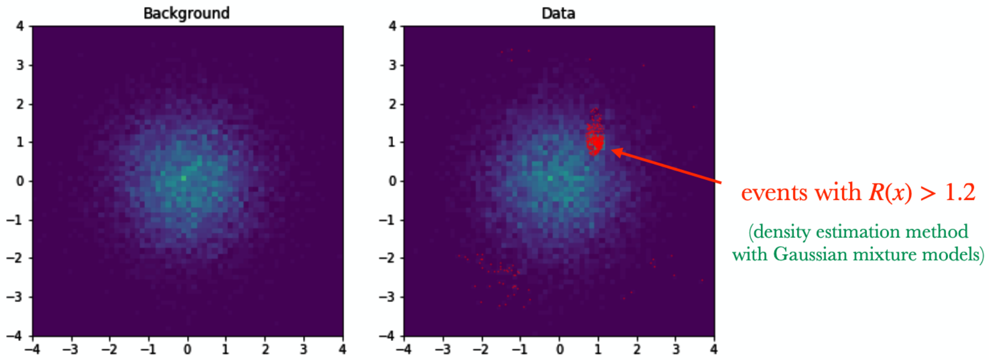


Farina, Nakai & **DS** 1808.08992



$$L = \frac{1}{N}\sum_i (x_i - x_i^{reco})^2$$

Overdensities



[figure by David Shih]

$$R(x) = \frac{p_{data}(x)}{p_{bkg}(x)}$$

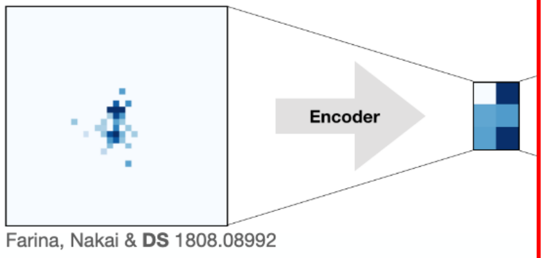
Learned with binary classifiers or neural density estimators.

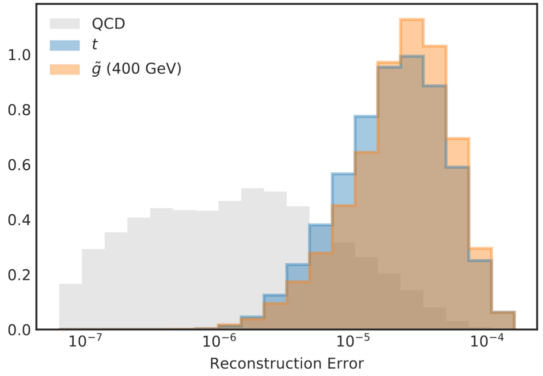
Anomaly detection

Identify interesting feature space regions by cutting on an anomaly score

<u>Outlier detection</u>

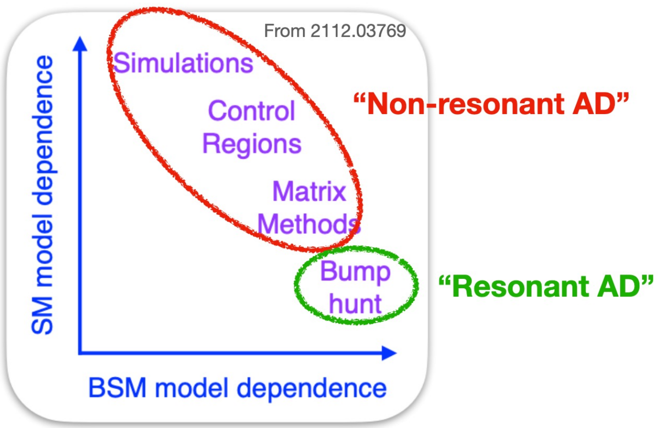
Autoencoders

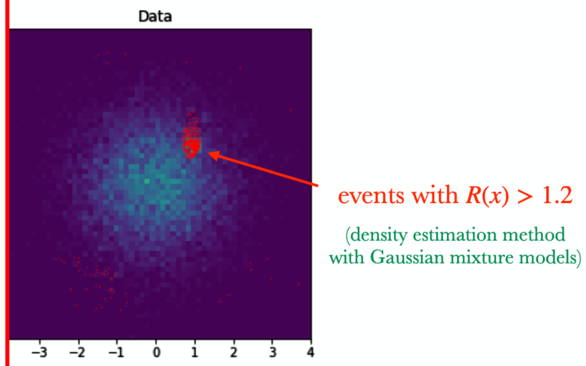




$$L = \frac{1}{N}\sum_i (x_i - x_i^{reco})^2$$

Other ingredient is background estimation:





$$R(x) = \frac{p_{data}(x)}{p_{bkg}(x)}$$

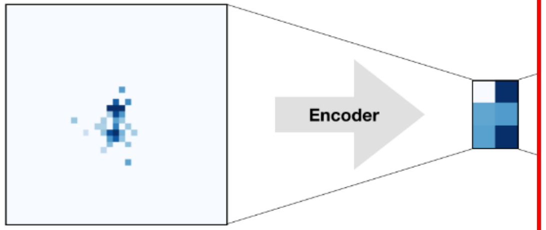
Learned with binary classifiers or neural density estimators.

Anomaly detection

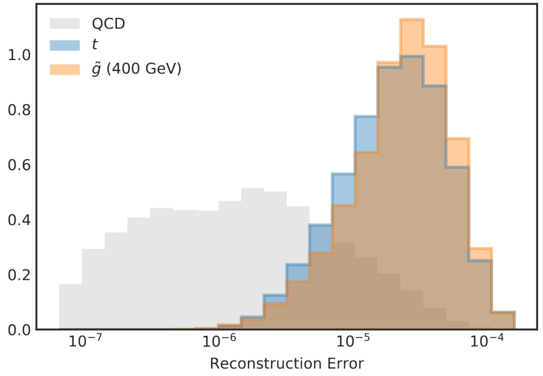
Identify interesting feature space regions by cutting on an anomaly score

Outlier detection

Autoencoders

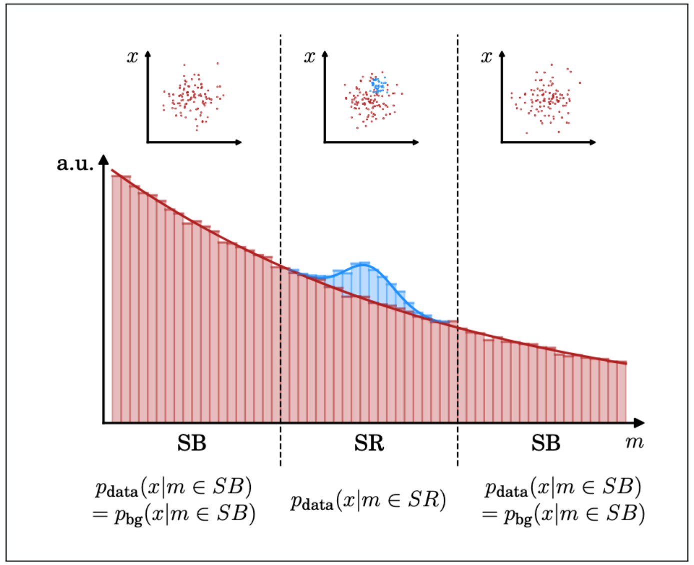


Farina, Nakai & **DS** 1808.08992



$L =$

Generative models for background estimation



CWoLa, ANODE, SALAD, CATHODE, CURTAINS, FETA, R-ANODE, …

David Shih's talk @ AD Topical Meeting

Roberto Seidita's talk!

events with $R(x) > 1.2$

(density estimation method with Gaussian mixture models)

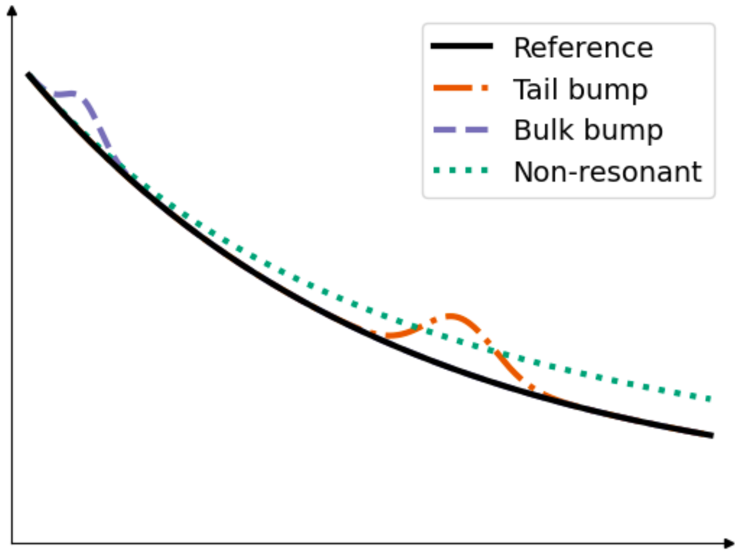
of neural density estimators.

Goodness-of-fit

How well a reference model R (the SM) describes the measured data $\mathcal{D}$.

- Without referencing an alternative hypothesis
- Full statistical test: it ouputs a global significance
- SM distribution is not known in closed form

  → Assume SM reference data are available $\mathcal{R}$
- “Interpretable” to some extent



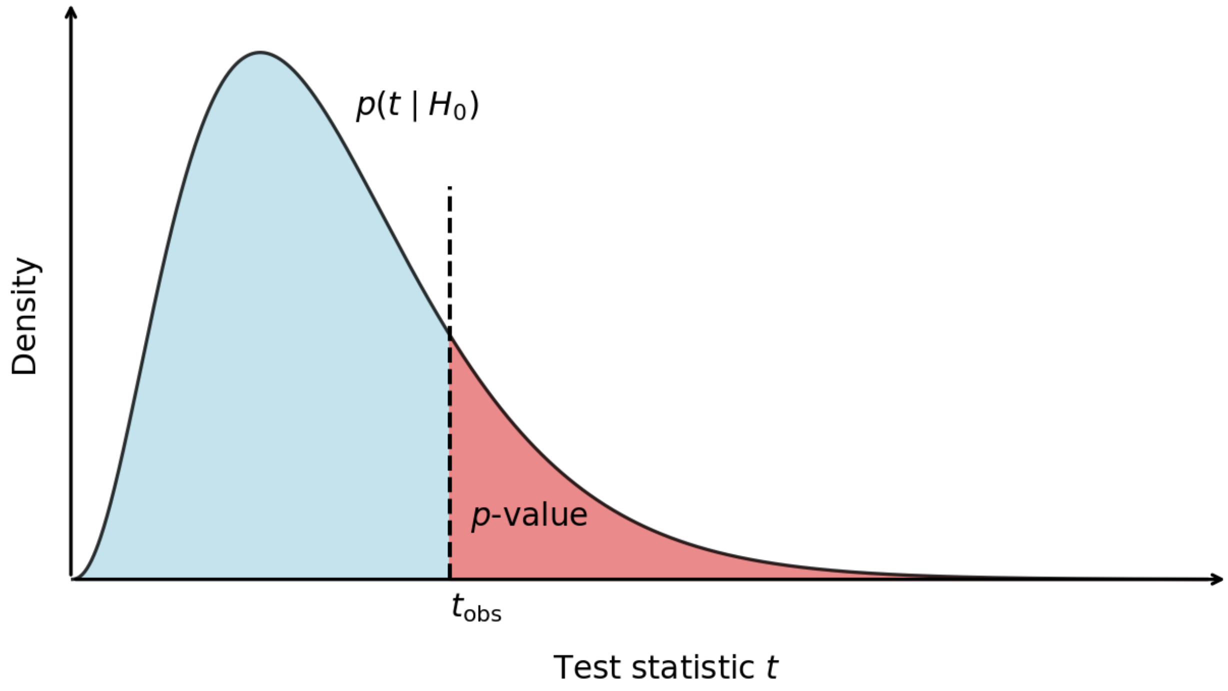
→ **two-sample hypothesis testing**

Assess if the null hypothesis $H_0 : p_R = p_{\text{true}}$ can be rejected from finite data.

Two-sample testing

Test statistic $t_{obs} = t(\mathcal{D}, \mathcal{R})$, $t(\cdot,\cdot)$ must be a agnostic

Large t_{obs} $\Rightarrow$ tension with $H_0 \colon p_R = p_{\text{true}}$. How do we quantify?



$$p_{\text{value}} = \int_{t_{obs}}^{\infty} p(t|H_0)\, dt$$

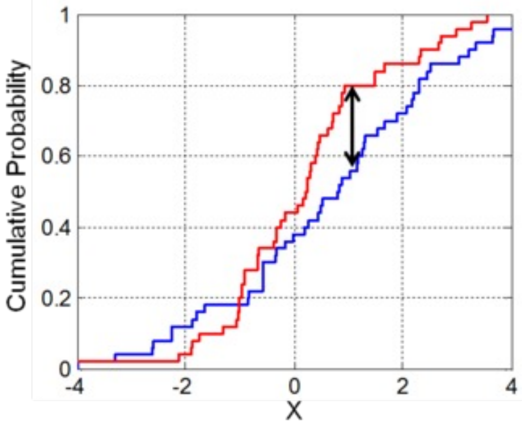
$p(t|H_0)$ estimated:
empirical distribution of possible outcomes under H_0 with SM toy data

$$t_{H_0} = t\left(\mathcal{D}^{(R)}, \mathcal{R}\right)$$

Goodness-of-fit testing

Classical methods:

- Univariate (e.g. KS)
- Binned (e.g. χ^2)



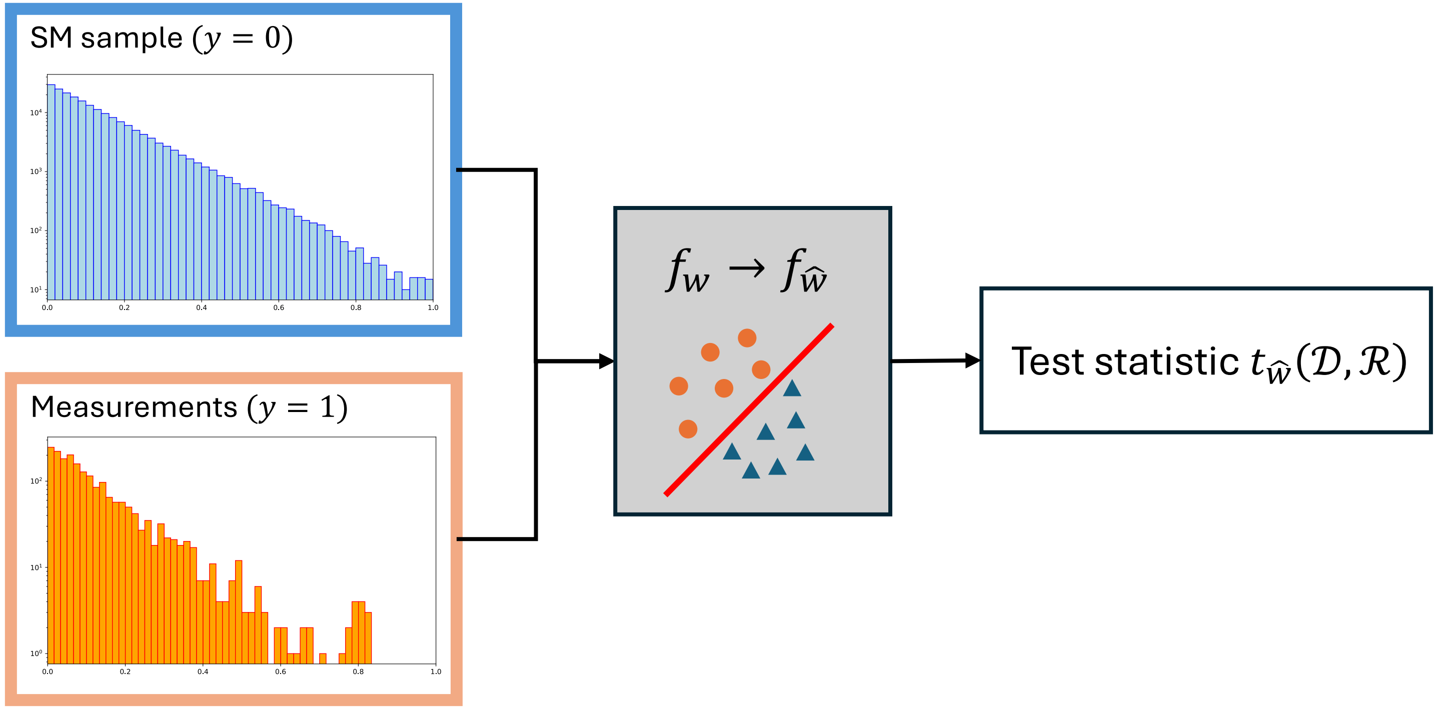
$$\chi^2 = \sum_{i=1}^{n_{\text{bins}}} \frac{(O_i - E_i)^2}{E_i}$$

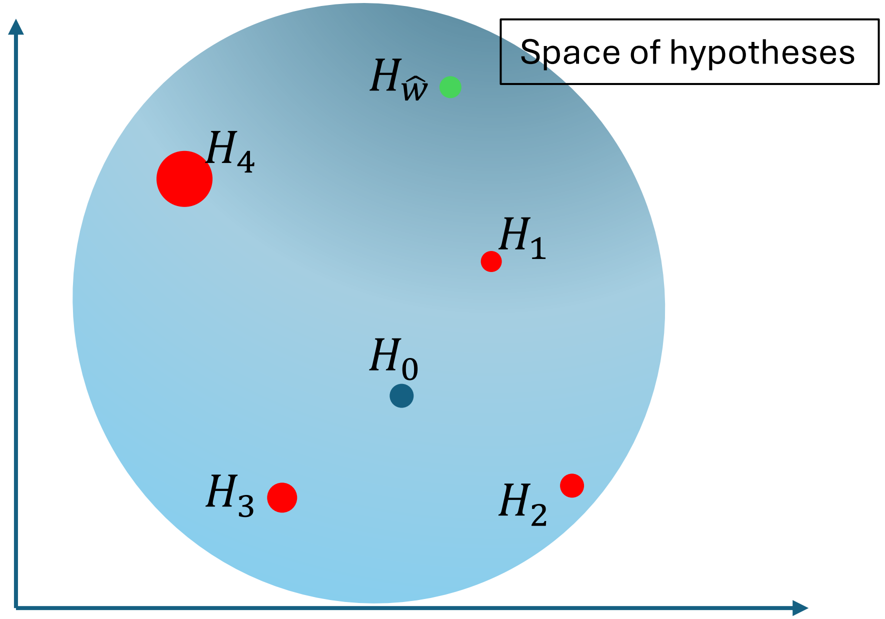
Limited options for $d > 3$ and large *N*:

- *Maximum Mean Discrepancy*
- *Wasserstein Distance*
- *Classifier two-sample tests*
- *The New Physics Learning Machine*

D'agnolo, Grosso, ML, Pierini, Wulzer, Zanetti
NPLM, Multivariate NPLM,Imperfect NPLM,
Fast NPLM, NPLM as a GoF, ...

The New Physics Learning Machine





(unbinned!)

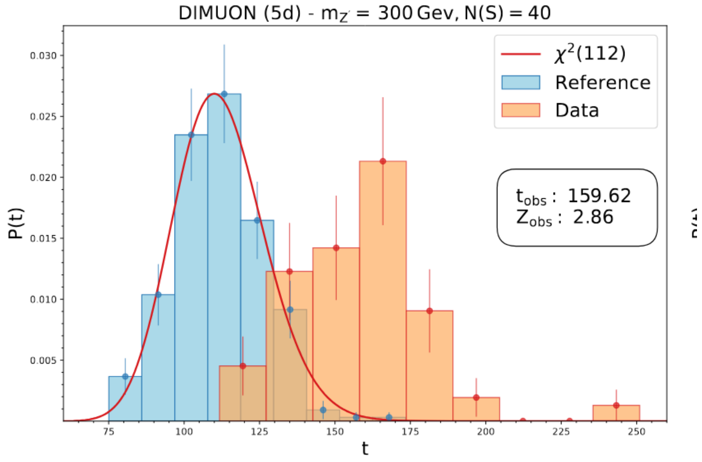
$$t_{\hat{w}}(\mathcal{D}, \mathcal{R}) = -2\left[\frac{N(SM)}{N_{SM}} \sum_{x \in \mathcal{R}} \left(e^{f_{\hat{w}}(x)} - 1\right) - \sum_{x \in \mathcal{D}} f_{\hat{w}}(x)\right] \approx 2 \log \frac{\mathcal{L}_{true}(\mathcal{D})}{\mathcal{L}_{SM}(\mathcal{D})}$$

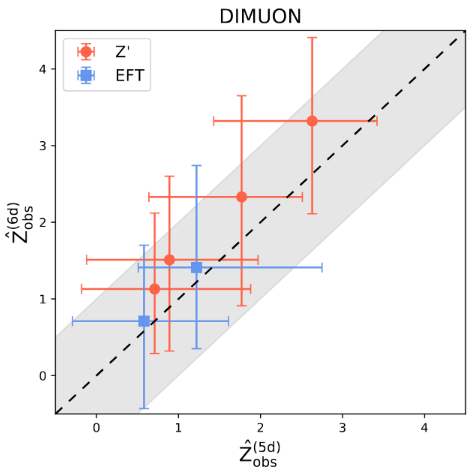
$$f_{\hat{w}}(x) \approx \log \frac{n_{\text{true}}(x)}{n(x|SM)}$$

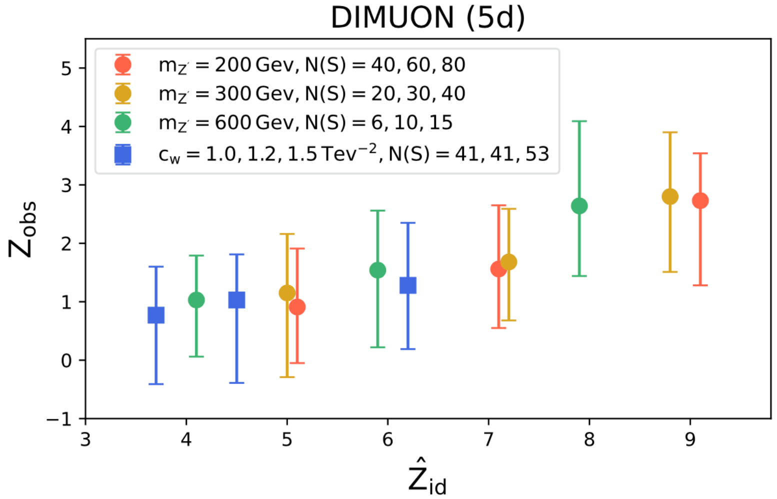
$$n(x) = N\, p(x)$$

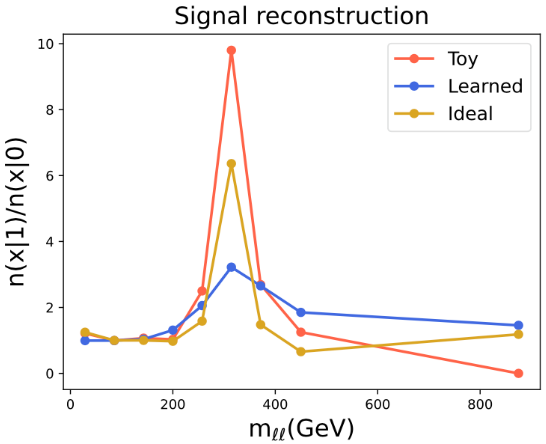
The New Physics Learning Machine

Performance studies [ML, Losapio, Rando, Grosso, Wulzer, Pierini, Zanetti, Rosasco (2022), 2204.02317]





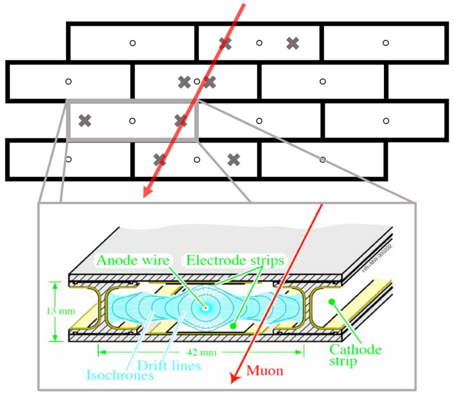


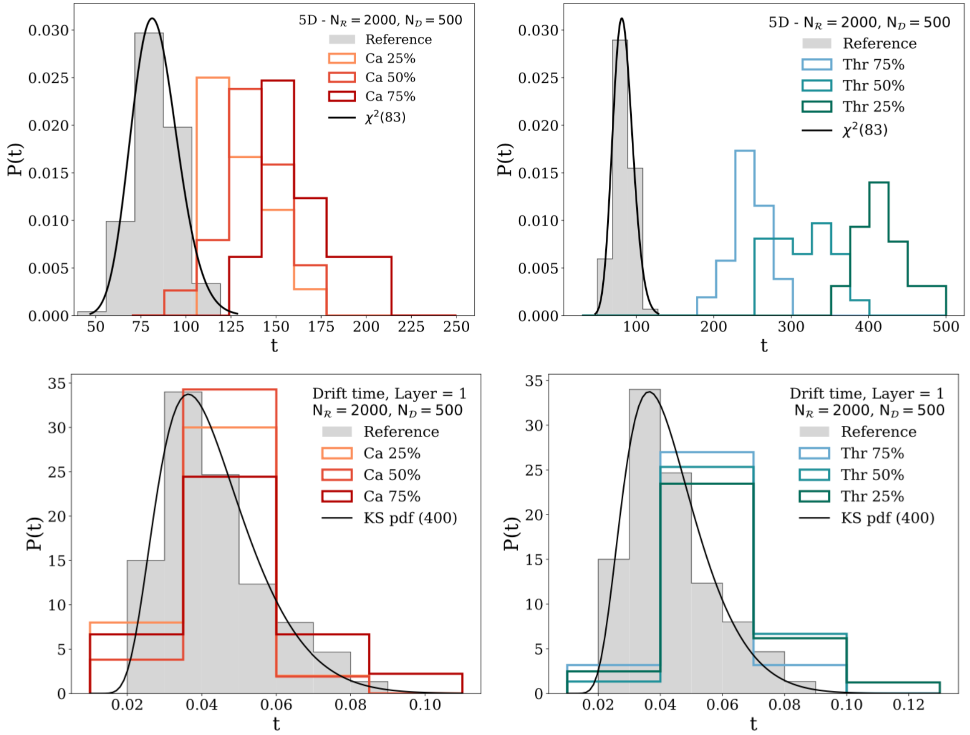


Data Quality Monitoring

[Grosso, Lai, ML, Pazzini, Rando, Rosasco, Wulzer, Zanetti, (2023), 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)

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

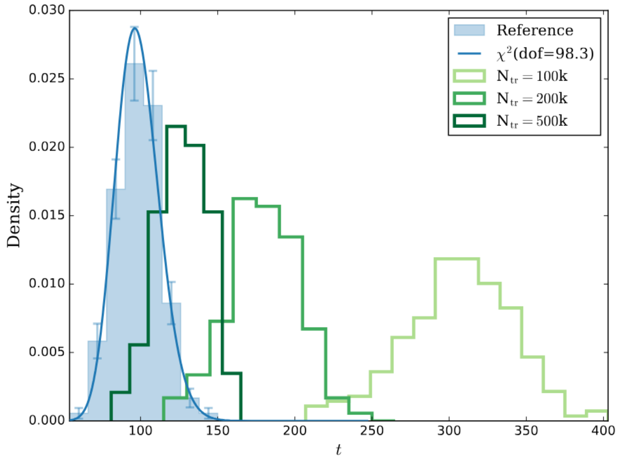
Data: https://zenodo.org/records/7128223

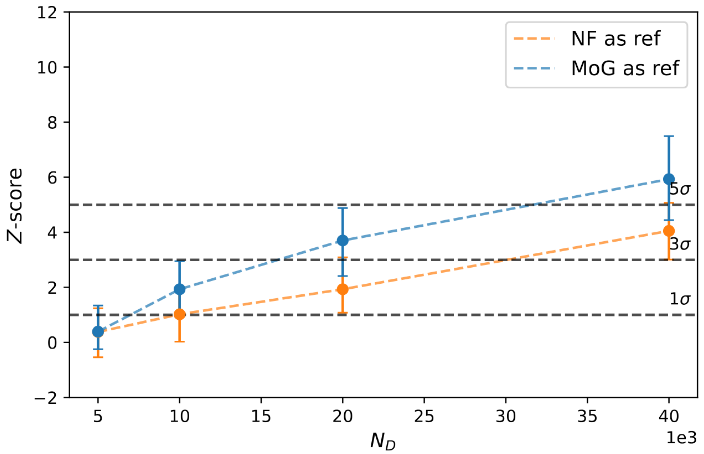
Evaluation of GenAI

Cappelli, Grosso, ML, Reyes-Gonzalez, Zanetti (out soon)

Compare data from a generative model against true data: relevant for fast simulations

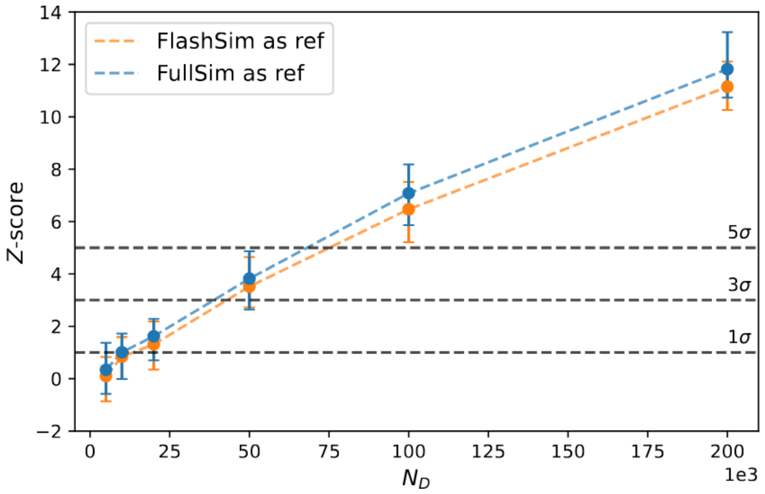
- Synthetic data: normalizing flows trained on mixture of Gaussians





- Realistic case: FlashSim

[Francesco Vaselli et al. *End-to-end simulation of particle physics events with Flow Matching and generator Oversampling*. 2024. 2402.13684]



Summary

- Two favors of model-agnostic searches:
  - Anomaly detection for the identification of interesting events/regions
  - Goodness-of-fit-based approaches: full agnostic statistical tests

- NPLM has yet to be applied to real analyses
  - Reliance on simulations → systematic uncertainties can be dealt with but tricky (2111.13633)
  - Current pipeline is slow → fast implementation is available but currently without systematics (2204.02317)
  - Scalability to high-dimensions: foundation models? Useful for low-level features (Metzger, Xu, Sodini, Arrestad, Govorkova, Grosso, Harris (2025), 2502.15926)
  - There is no notion of a "best GoF test" (with highest power against any new physics signal): certain tests can be better at detecting specific signals.
    NPLM is found to be more robust than the alternatives (less biased).

- Can be used for other tasks: Data Validation, DQM, validation of GenAI

Remarks

- How to evaluate signal-agnostic methods
  - Designed to detect the unexpected
  - Signal-benchmarks are reassuring (just for testing, no optimization) but not general
  - Comparison with ideal significance (fully aware: Neyman-Pearson) to esimate sensitivity loss

- Interpretability
  - How to characterize new physics if found?
  - Black box models do not help

- Re-intepretability
  - Reporting p-values is not enough
  - NPLM: trained model, background sample, $p(t|H_0)$

Thank you!