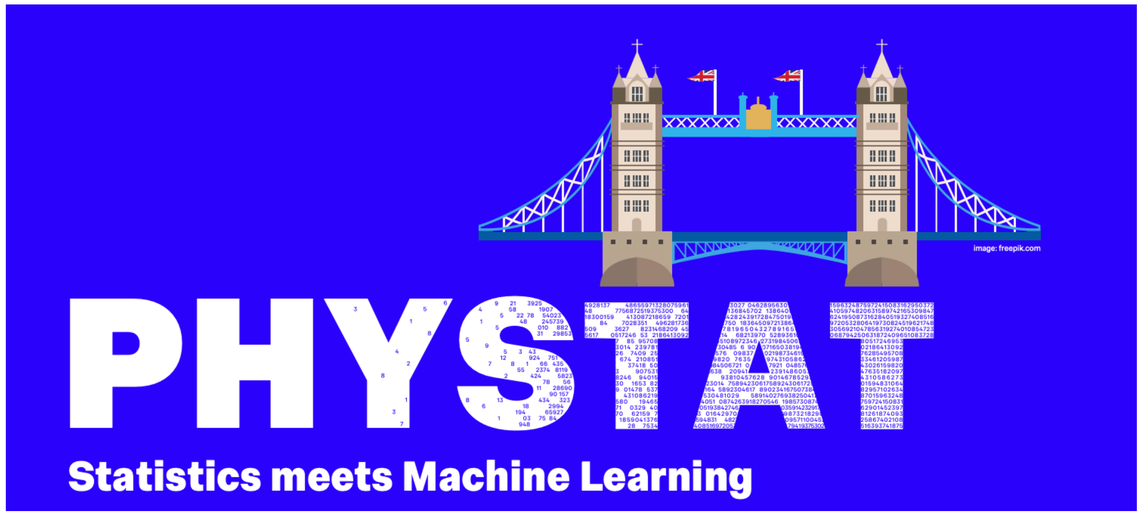
September 10th, 2024



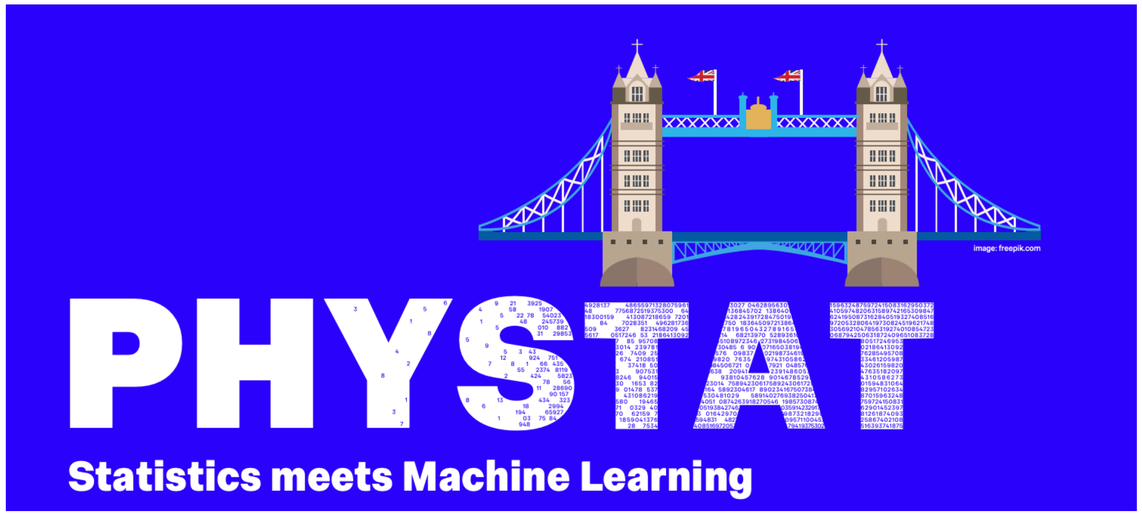
Statistical tests for anomaly detection in particle physics

Gaia Grosso
IAIFI fellow, MIT, Harvard

gaia.grosso@cern.ch

Massachusetts Institute of Technology

HARVARD UNIVERSITY



This talk

ML for statistical tests for anomaly detection at the LHC

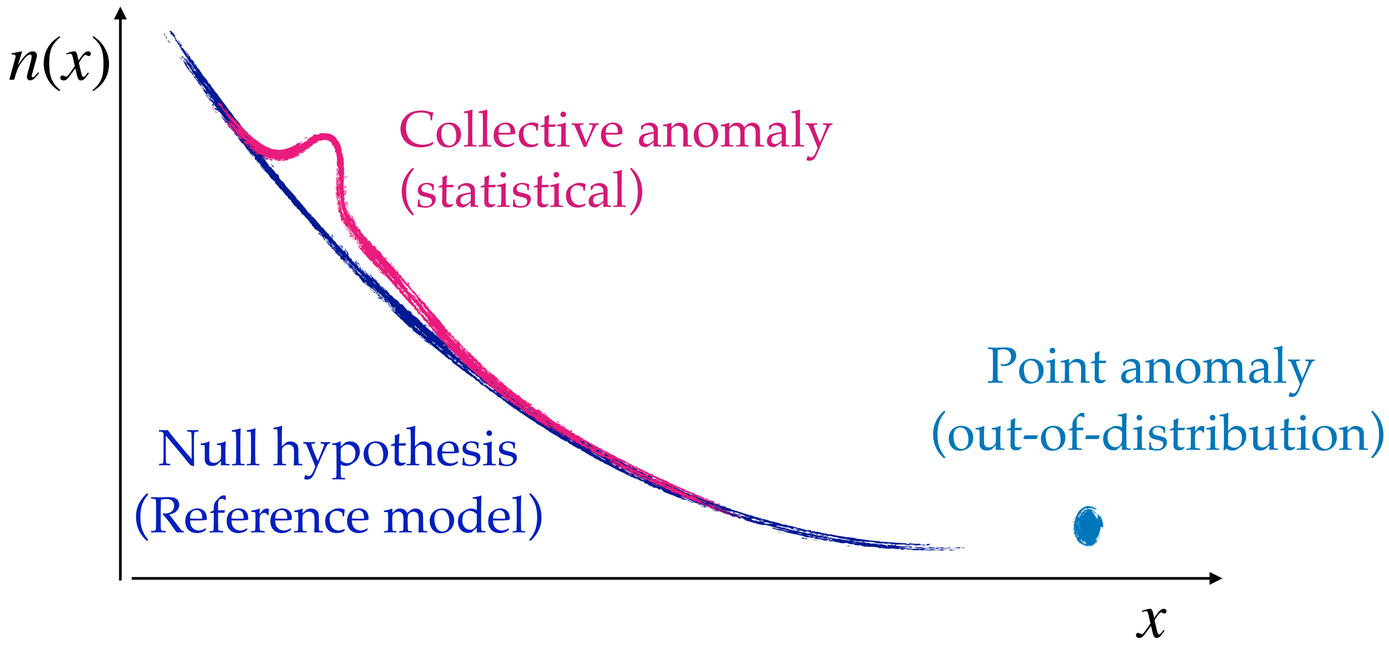
- Overview
- Challenges
- Possible solutions/ongoing work

[Disclaimer: biased selection]

Anomaly detection

finding *patterns* in data that do not conform
to a well defined notion of *normal* behavior

"Anomaly detection: A survey" Chandola et al. 2010

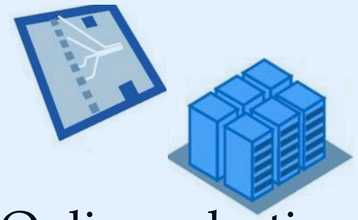


1. Detect anomalies
2. Assess their statistical significance

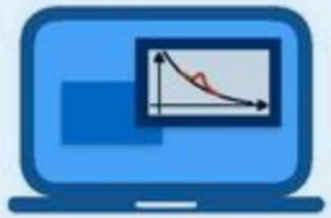
How good is your model?

LHC Experiment

Detector readout



Online selection



Offline analysis

Detector malfunctioning ?

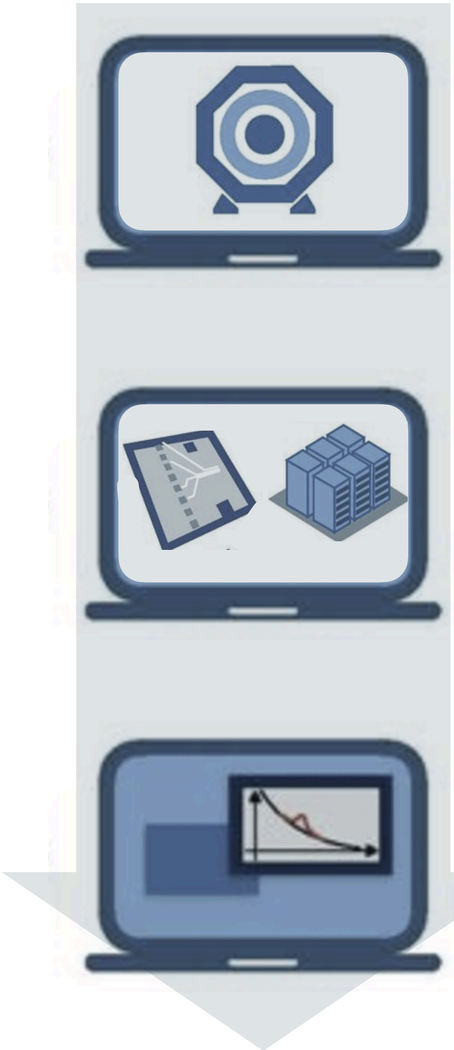
Physics Beyond the Standard Model ?

Typical question in experimental physics

How well do we understand experimental data?

How good is your model?

Simulation



Can we trust our simulations ?

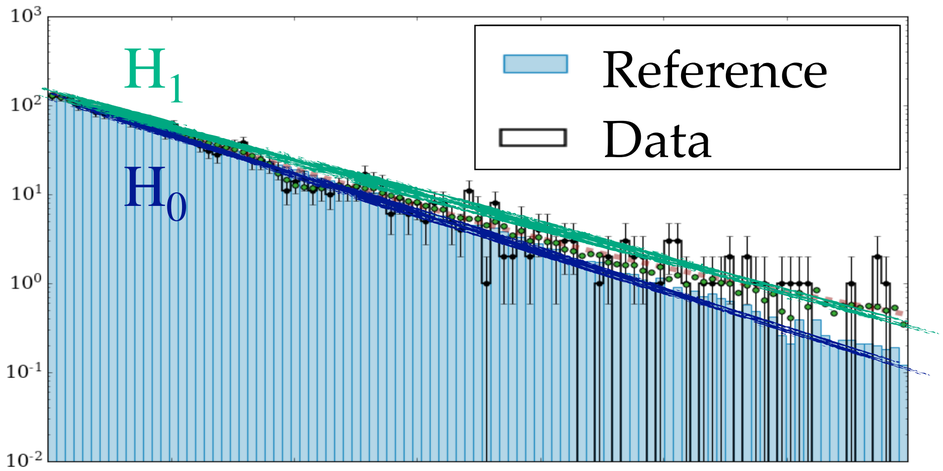
How to compare generative models ?

Can we build faithful surrogate models ?

… increasing interest

in the era of Generative AI

AI meets Physics [1/3]



Supervised:

Hypothesis testing

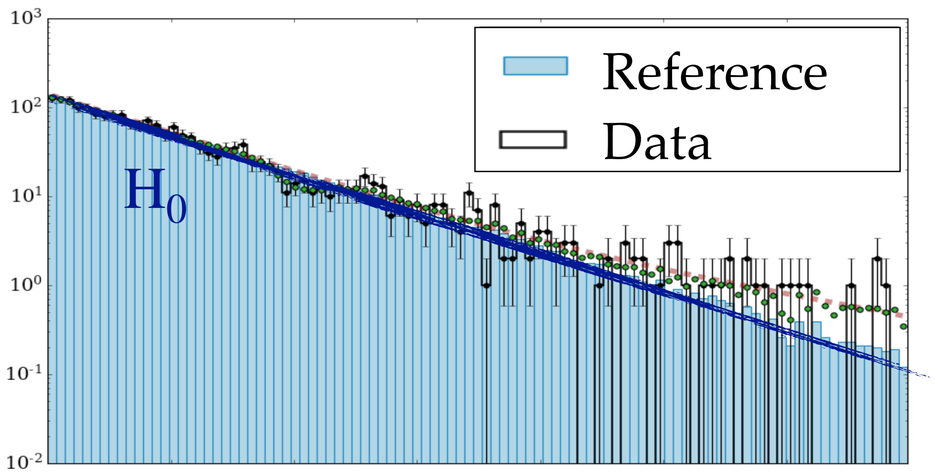
Likelihood-ratio test: traditionally employed in searches for physics Beyond the Standard Model (BSM) at the LHC.

Optimal test (Neyman-Pearson lemma)

Optimal detection can be designed
(we know how to do this)

$$t(\mathcal{D}) = 2 \log \frac{\mathcal{L}(\mathcal{D}|\mathrm{H_1})}{\mathcal{L}(\mathcal{D}|\mathrm{H_0})}$$

AI meets Physics [2/3]



What is optimal when measuring the *unexpected*?

Semi-supervised:

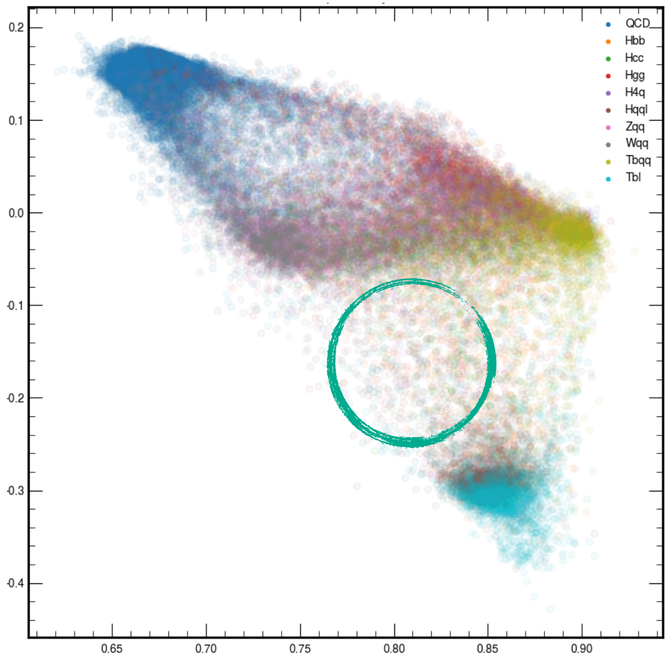
goodness-of-fit / two-sample test

- **Data**: experimental measurements of the natural process $D = \{x_i\}_{i=1}^{N_D}$
- **Reference model**: expected nominal behavior of the data (Standard Model, normal operating condition of a detector…).

  Often not known in close form:

  Reference sample $R = \{y_i\}_{i=1}^{N_R} \rightarrow$ ***Two-sample test***

AI meets Physics [3/3]



Unsupervised:
Clustering / self-organization

Can be exploited to <u>detect</u>

- out-of-distribution points
- Modes / patterns in data

Must be followed by a (semi)supervised method to asses the *statistical* significance

(what's the likelihood of the anomalous point under the null?)

Two-sample tests for anomaly detection at the LHC

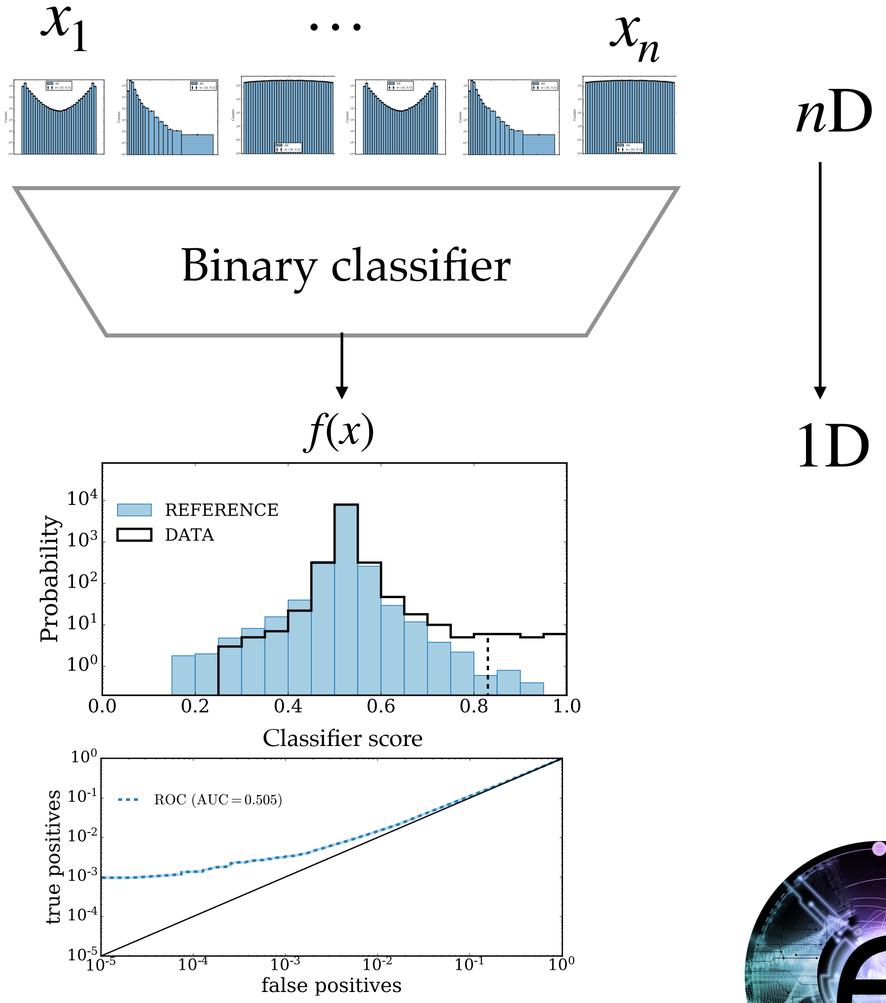
How to design a statistical test for anomalies with ML

Classifier-based two-sample tests

- Train a classifier to tell apart two samples, D and R.
- Output: density ratio function

$$f(x, w) = \log \left[\frac{n(x \mid D)}{n(x \mid R)} \right]$$

- Design a test $t_f(D, R)$:
  - Standard 1D GOF on the classifier score: KS, AD, χ^2, etc.[1]
  - Classification metrics: ACC, AUC, MCE [2]



[1] Friedman (2003)
[2] Charkavarti et al. (2021), Lopez et al. (2017)
[3] Baker and Cousins (1984), d'Agnolo, Wulzer (2018), d'Agnolo, Grosso et al. (2021)
[4] Metodiev, Nachman, Thaler (2017) […]

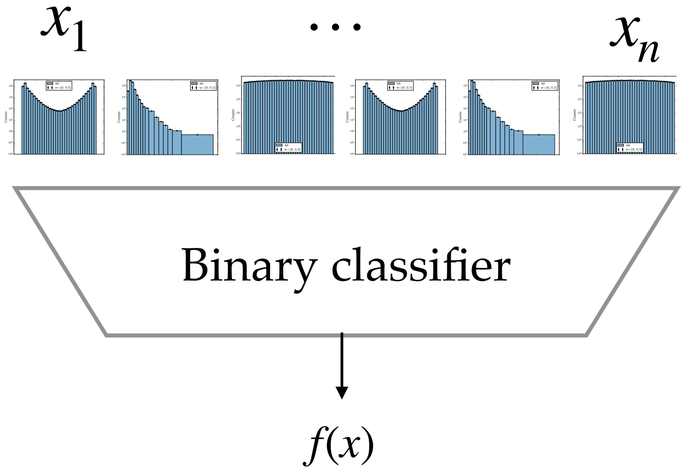
How to design a statistical test for anomalies with ML

Classifier-based two-sample tests

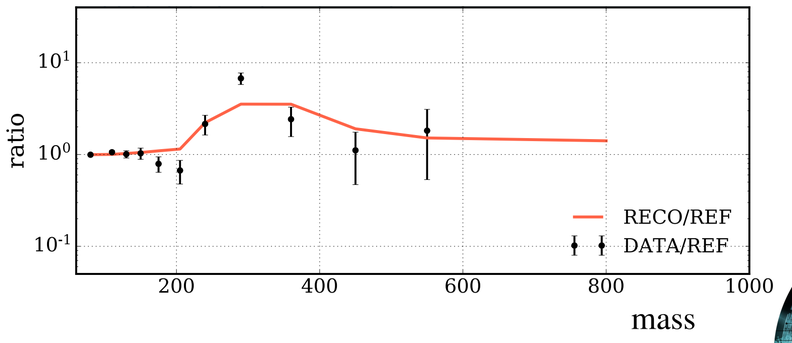
- Train a classifier to tell apart two samples, D and R.
- Output: density ratio function

$$f(x, w) = \log\left[\frac{n(x\,|\,D)}{n(x\,|\,R)}\right]$$

- Design a test $t_f(D, R)$:
  - Standard 1D GOF on the classifier score[1]
  - Classification metrics[2]
  - Neyman-Pearson test from log-density-ratio[3]



$$t_f(D) = 2\sum_{x\in\mathcal{D}} f_{\mathbf{w}}(x) - 2\sum_{x\in\mathcal{R}} \frac{\mathrm{N(R)}}{\mathrm{N}_{\mathcal{R}}}\left[e^{f(x;\mathbf{w})} - 1\right]$$



[1] Friedman (2003)
[2] Charkavarti et al. (2021), Lopez et al. (2017)
[3] Baker and Cousins (1984), d'Agnolo, Wulzer (2018), d'Agnolo, Grosso et al. (2021)
[4] Metodiev, Nachman, Thaler (2017) […]

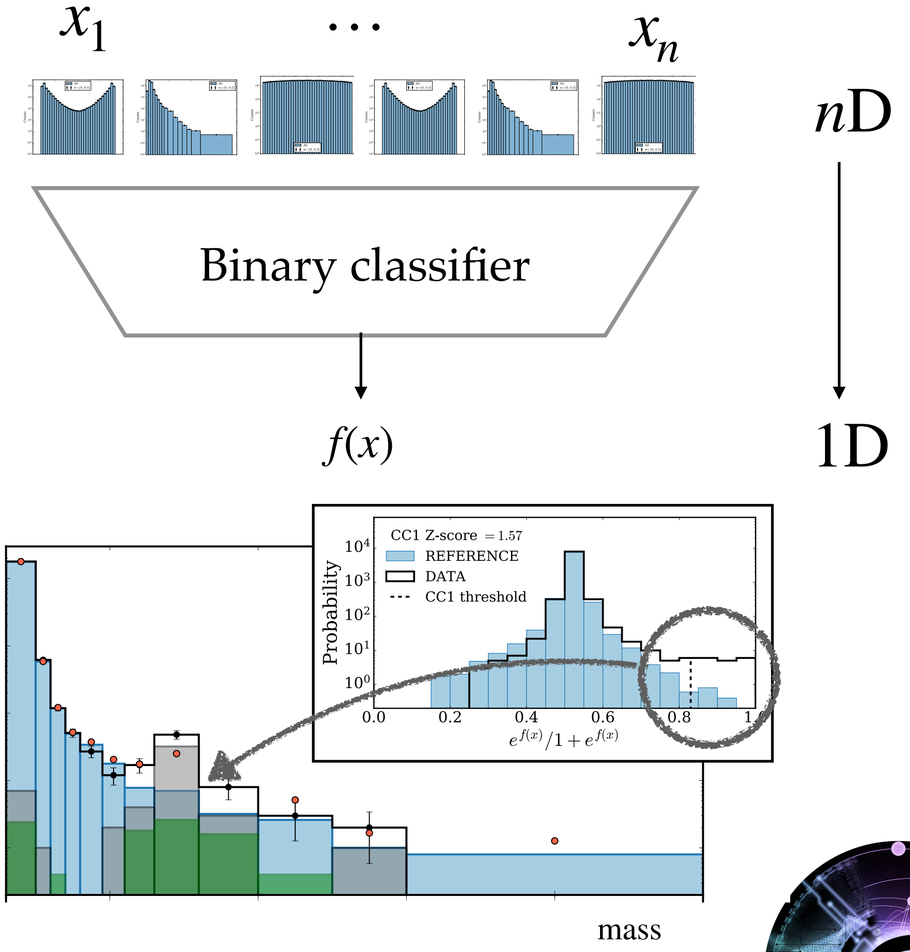
How to design a statistical test for anomalies with ML

Classifier-based two-sample tests

- Train a classifier to tell apart two samples, D and R.
- Output: density ratio function

$$f(x, w) = \log \left[\frac{n(x \mid D)}{n(x \mid R)} \right]$$

- Design a test $t_f(D, R)$:
  - Standard 1D GOF on the classifier score[1]
  - Classification metrics[2]
  - Neyman-Pearson test from log-density-ratio[3]
  - Select a signal region based on the classifier score, to enhance a signal hypothesis (e.g Bump-Hunt)[4]



[1] Friedman (2003)
[2] Charkavarti et al. (2021), Lopez et al. (2017)
[3] Baker and Cousins (1984), d'Agnolo, Wulzer (2018), d'Agnolo, Grosso et al. (2021)
[4] Metodiev, Nachman, Thaler (2017) [...]

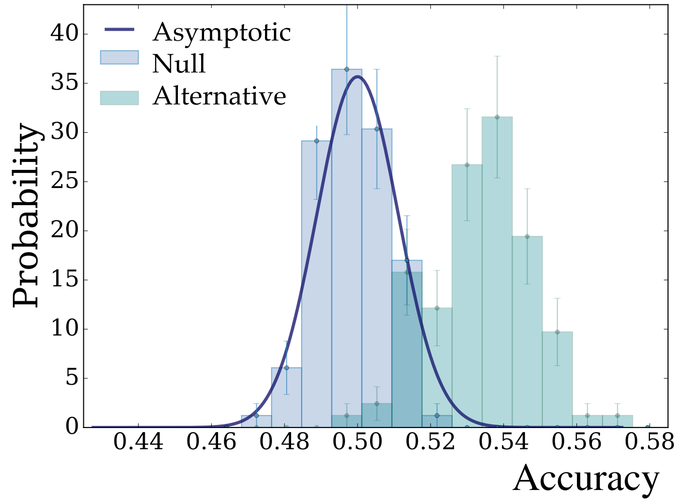
How to design a statistical test for anomalies with ML

Classifier-based two-sample tests

- Train a classifier to tell apart two samples, D and R.

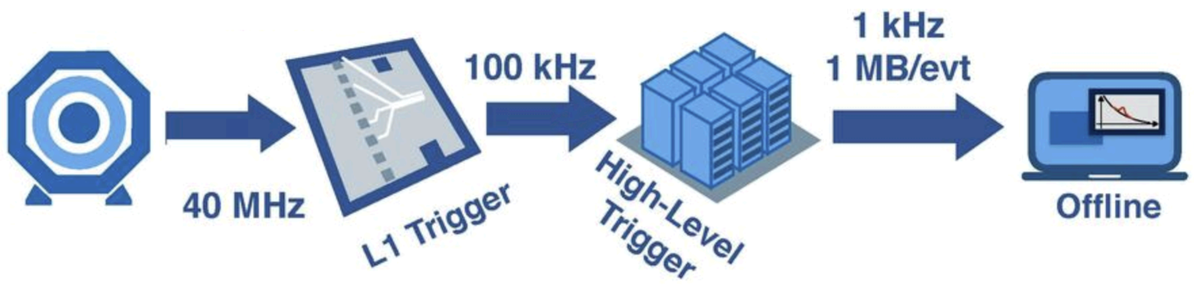
- Calibration and p-value
  - The single value of test statistic is not sufficient without a fair comparison to the test statistic distribution under the null
  - Build the null: asymptotic, bootstrap, permutation, slow permutation[2]

Example: Accuracy test calibration

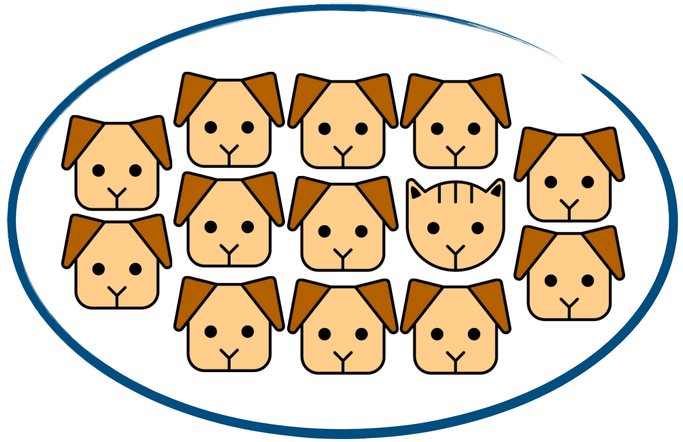


[1] Friedman (2003)
[2] Charkavarti et al. (2021), Lopez et al. (2017)
[3] Baker and Cousins (1984), d'Agnolo, Wulzer (2018), d'Agnolo, Grosso et al. (2021)
[4] Metodiev, Nachman, Thaler (2017) […]

What makes the problem so hard at the LHC?



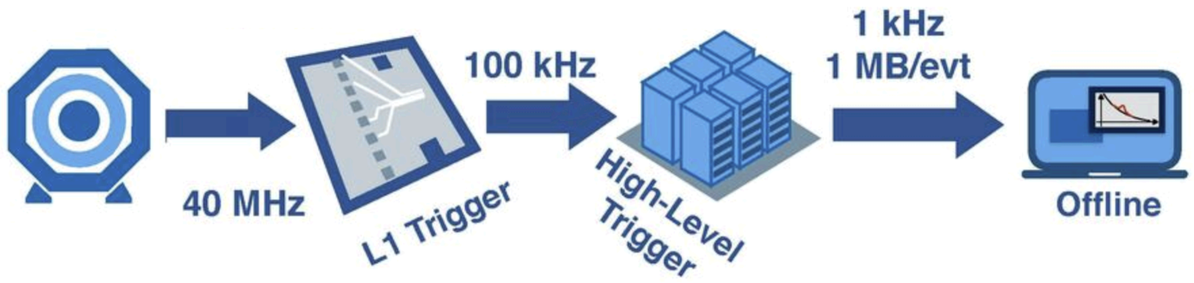
vs.



Challenges of *Two-sample test* in particle physics

- Big data challenge: high dimensional problems, large size samples
- Anomalies are rare! (highly diluted anomalous points)
- Reference models are affected by uncertainties
- Beyond human intervention: we need a statistical evaluation to claim anomaly detection

Challenges of statistical anomaly detection (at the LHC)



$$t = t(D, R_\nu)$$

Are the input data *informative* about existing anomalies?

- DATA REPRESENTATION
- DATA COMPRESSION
- SCALABILITY

How can AI help?

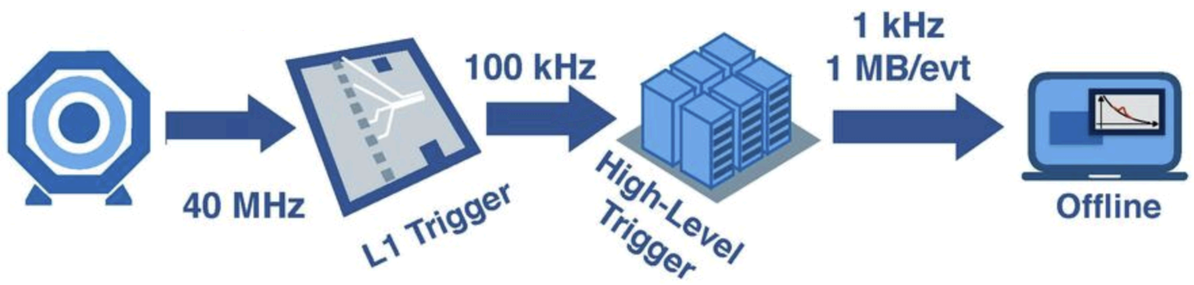
How to design a good test?

- MODEL SELECTION
- INDUCTIVE BIAS
- INTERPRETABILITY

How to propagate *systematic uncertainties* in the Reference model?

- ROBUSTNESS

Challenges of statistical anomaly detection (at the LHC)



$$t = t(D, R_\nu)$$

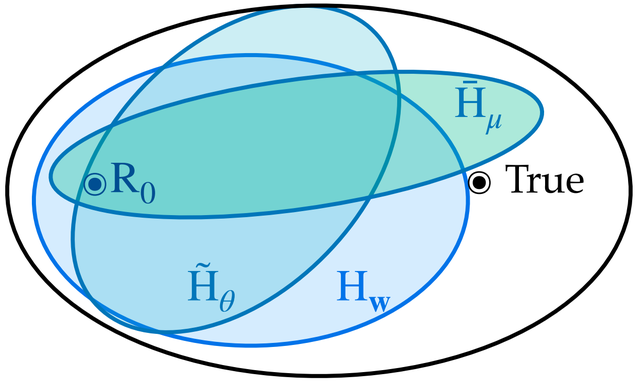
How to design a good test?

- MODEL SELECTION
- INDUCTIVE BIAS
- INTERPRETABILITY

The problem of *model selection*

How to define the family of universal approximants?

$$f(x, w) = \log \left[\frac{n(x \mid D)}{n(x \mid R)} \right]$$

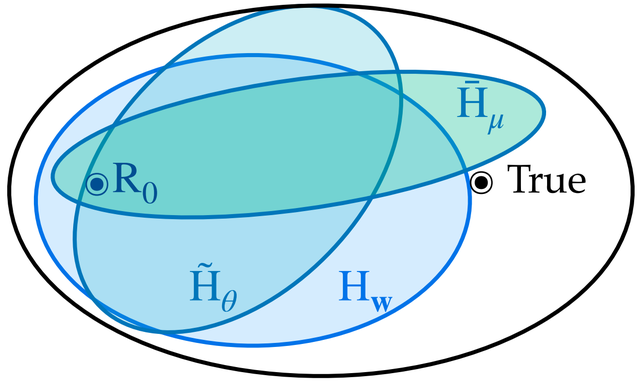


- Trade-off between *expressivity* and *specificity* is required
  - *Model selection*: hyper-parameters choice poses hard constraints.
  - *Regularization* is a powerful form of *inductive bias* (e.g. smoothness) affecting the **learning dynamics**

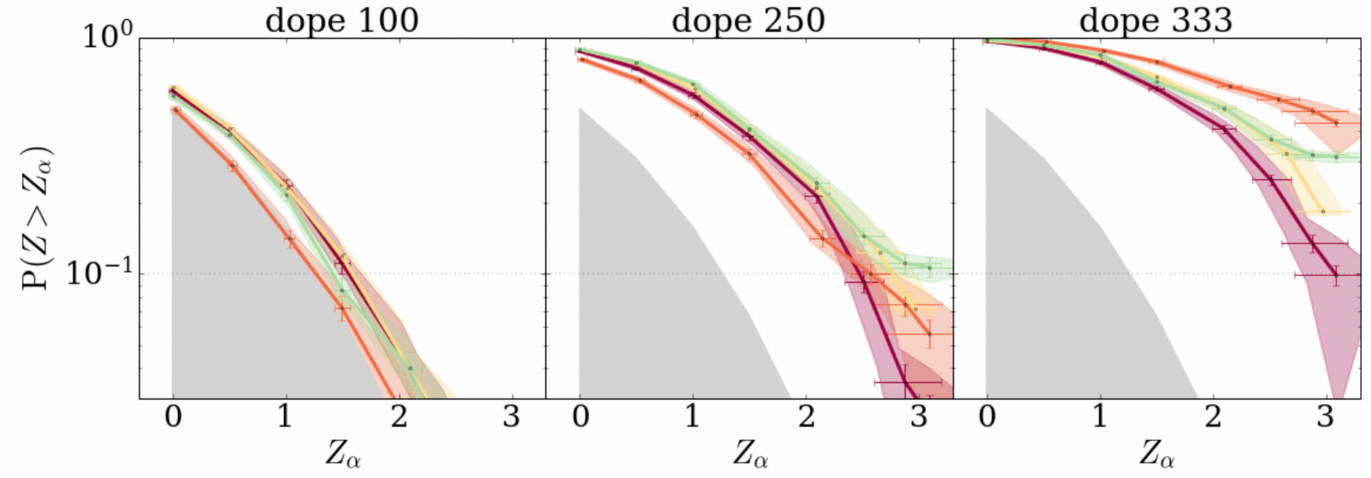
The problem of *model selection*

How to define the family of universal approximants?

$$f(x, w) = \log \left[\frac{n(x \mid D)}{n(x \mid R)} \right]$$



A unique *optimal* solution doesn't exists!



LHCO dataset

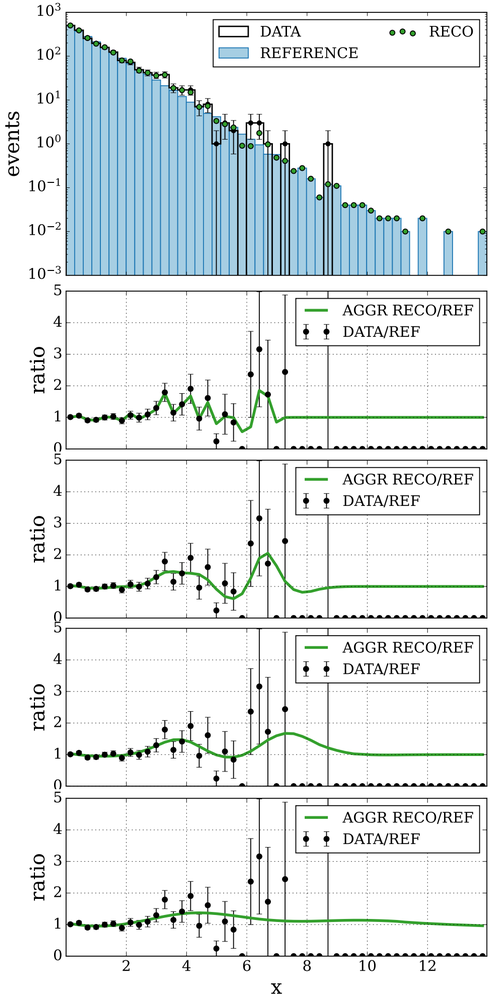
BDT binary classifier for signal region selection

- leaf=100, $\lambda = 10^{-2}$, thr=0.8
- leaf=100, $\lambda = 10^{-1}$, thr=0.99 → best at dope 333
- leaf=100, $\lambda = 10^{-1}$, thr=0.8 → best at dope 100
- leaf=31, $\lambda = 10^{-1}$, thr=0.8 → best at dope 250

(Work in preparation, credits to D. Sengupta)

The problem of *model selection*

Multiple testing for robust Neyman-Pearson two-sample test



Kernel width

$\sigma = 0.1$

$\sigma = 0.7$

$\sigma = 1.4$

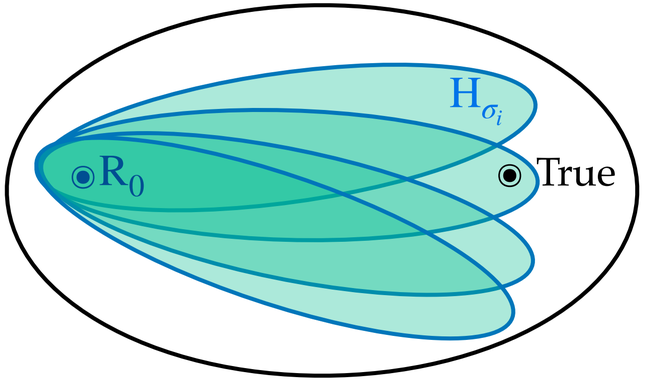
$\sigma = 3.0$

Kernel-based model

$$f_w(x) = \sum_{i=1}^{M} w_i k_\sigma(x, \tilde{x}_i)$$

$$k_\sigma(x, x') = e^{-\|x-x'\|^2/2\sigma^2}$$

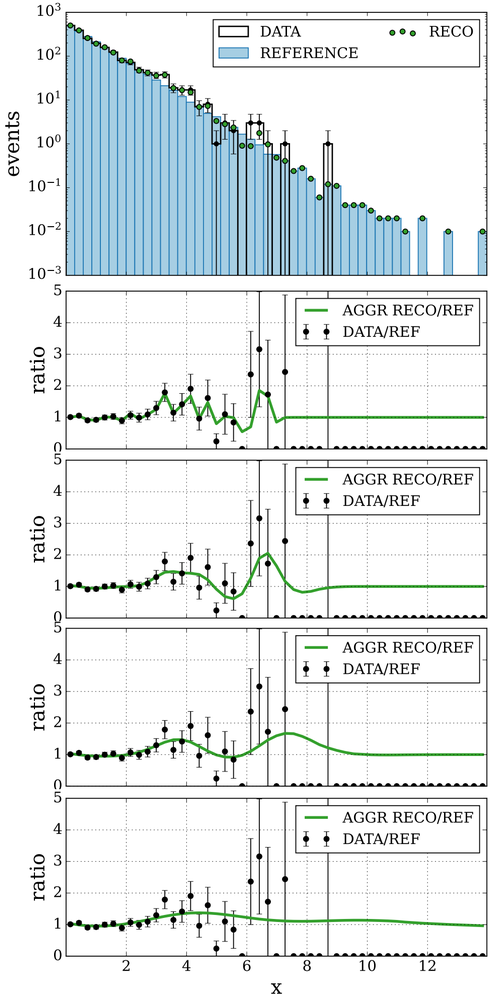
Multiple testing over model hyper-parameters



"Multiple testing for signal-agnostic searches of new physics with machine learning" 2408.12296 (Grosso, Letizia)

The problem of *model selection*

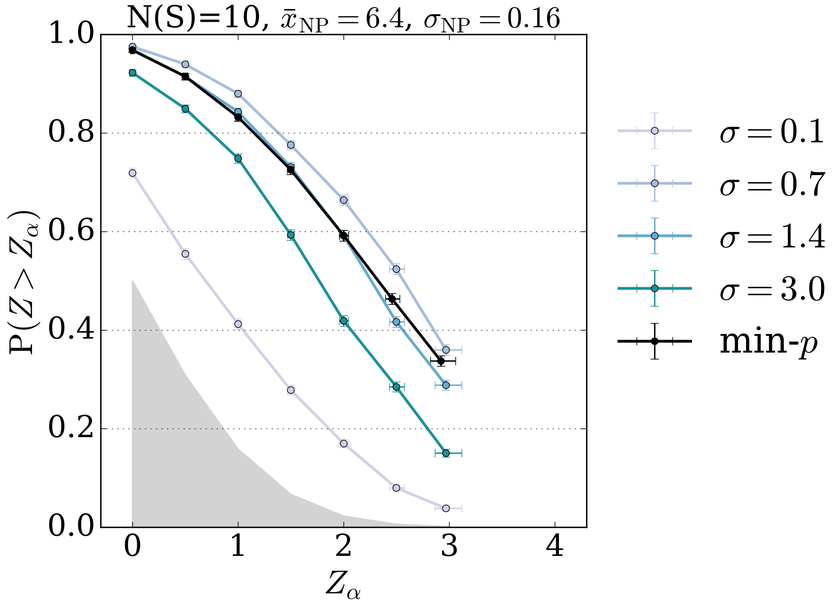
Multiple testing for robust Neyman-Pearson two-sample test



Kernel width	Neyman Pearson test	p-value
$\sigma = 0.1$	$\rightarrow t_\sigma(D,R)$	$\rightarrow p_\sigma(D,R)$
$\sigma = 0.7$	$\rightarrow t_\sigma(D,R)$	$\rightarrow p_\sigma(D,R)$
$\sigma = 1.4$	$\rightarrow t_\sigma(D,R)$	$\rightarrow p_\sigma(D,R)$
$\sigma = 3.0$	$\rightarrow t_\sigma(D,R)$	$\rightarrow p_\sigma(D,R)$

p-aggregation

$$p_{\min}(D,R) = \min_\sigma p_\sigma(D,R)$$

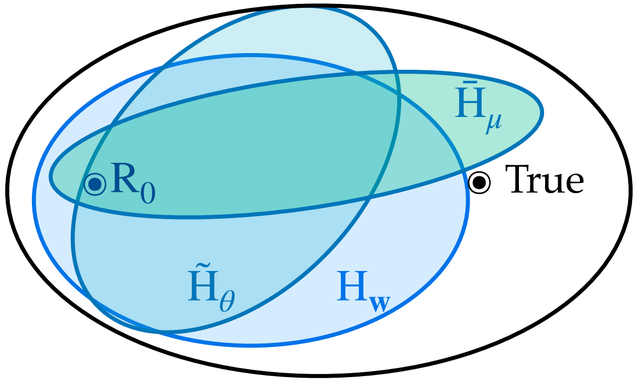


"Multiple testing for signal-agnostic searches of new physics with machine learning" 2408.12296 (Grosso, Letizia)

The problem of *model selection*

How to define the family of universal approximants?

$$f(x, w) = \log \left[\frac{n(x \mid D)}{n(x \mid R)} \right]$$



- Trade-off between *expressivity* and *specificity* is required
  - *Model selection*: hyper-parameters choice poses hard constraints.
  - *Regularization* is a powerful form of *inductive bias* (e.g. smoothness) affecting the **learning dynamics**

Interpretability vs. sensitivity

How to design $f(x)$ to capture *rare* and *unexpected* subtle perturbations on top of the known physics?

The problem of *model selection*

Sparse linear combination of Gaussian Kernels (SparKer)

$$f_{\boldsymbol{\mu},\boldsymbol{w}}(x) = \sum_{i=1}^{\mathrm{M}} w_i\, k(x;\, \mu_i, \sigma_i)$$

Local interpretability
Active kernels highlight anomalous regions

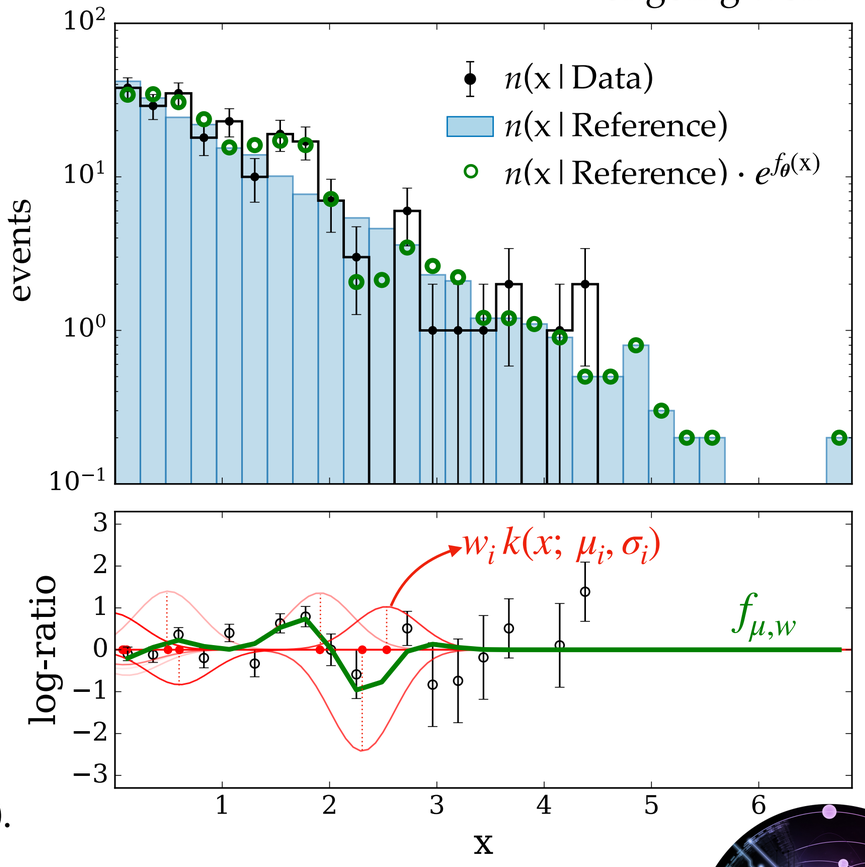
$$k(x;\, \mu_i, \sigma_i) = \mathrm{A} \exp\left[-\frac{||x-\mu_i||^2}{2\sigma_i^2}\right]$$

Sparse model (M $\ll$ N)
competition between data points to attract the kernels

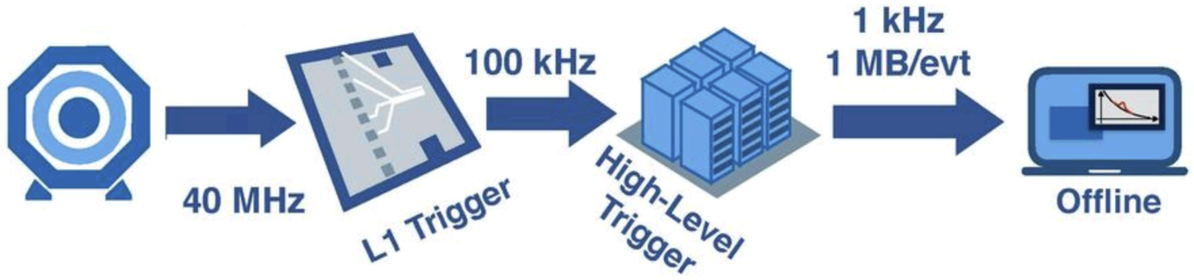
Adaptive model (learnable μ)
directing *attention* to anomalous features

Smooth model ($\sigma^2 = \sigma^2_{\mathrm{exp}} + \sigma^2_X$)
Physics constraints (e.g. experimental resolution).
What is the scale of New Physics?

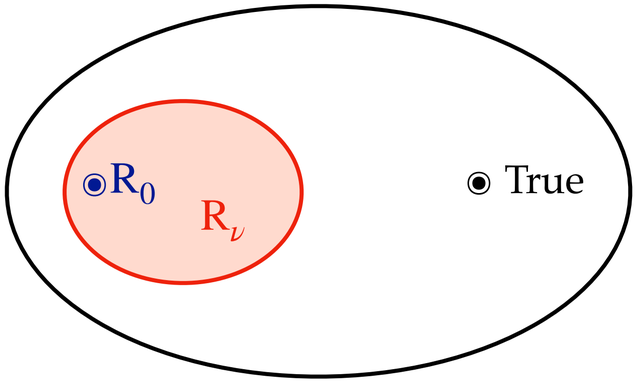
ongoing work

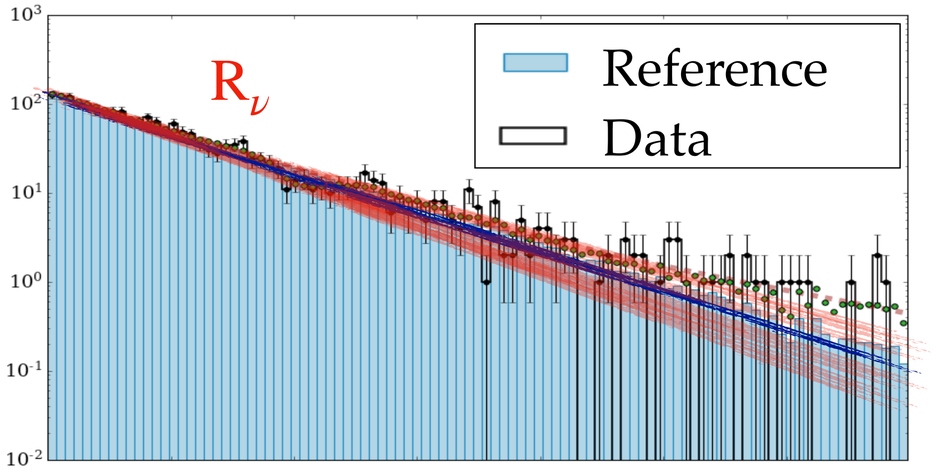


Dealing with *systematic uncertainties*



$$t = t(D, R_\nu)$$





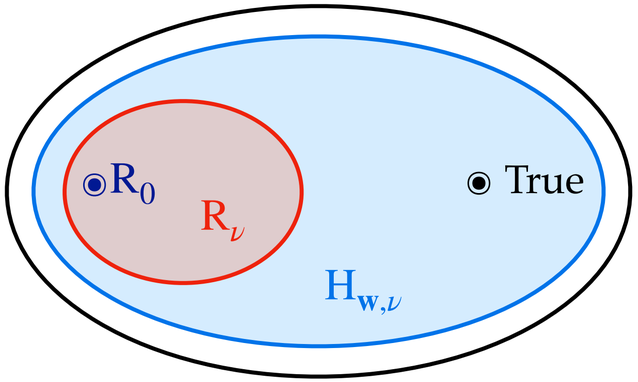
How to propagate *systematic uncertainties* in the Reference model?

- ROBUSTNESS

Dealing with *systematic uncertainties*

Frequentist approach (discovery at the LHC)

$$t(\mathcal{D},\,\mathcal{A}) = 2\log\left[\frac{\max\limits_{\mathbf{w},\,\boldsymbol{\nu}}\mathcal{L}(\mathrm{H}_{\mathbf{w},\,\boldsymbol{\nu}}|\mathcal{D},\,\mathcal{A})}{\max\limits_{\boldsymbol{\nu}}\mathcal{L}(\mathrm{R}_{\boldsymbol{\nu}}|\mathcal{D},\,\mathcal{A})}\right] = 2\log\left[\frac{\max\limits_{\mathbf{w},\,\nu}\mathcal{L}(\mathrm{H}_{\mathbf{w},\,\boldsymbol{\nu}}|\mathcal{D})\mathcal{L}(\boldsymbol{\nu}|\mathcal{A})}{\max\limits_{\boldsymbol{\nu}}\mathcal{L}(\mathrm{R}_{\boldsymbol{\nu}}|\mathcal{D})\mathcal{L}(\boldsymbol{\nu}|\mathcal{A})}\right]$$



$\mathrm{R}_{\boldsymbol{\nu}}$: reference hypothesis (null)

$\mathrm{H}_{\mathbf{w},\,\boldsymbol{\nu}}$: alternative hypothesis

$\mathbf{w}$: trainable parameters on the NN model

$\boldsymbol{\nu}$: set of nuisance parameters modeling the uncertainties effects

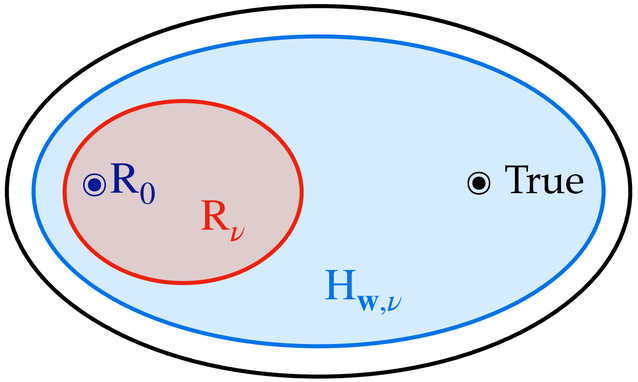
$\mathcal{D}$: data sample

$\mathcal{A}$: auxiliary sample (used to constrain $\boldsymbol{\nu}$)

Dealing with *systematic uncertainties*

Frequentist approach (discovery at the LHC)

$$t(\mathcal{D},\,\mathcal{A}) = 2\log\left[\frac{\max\limits_{\mathbf{w},\,\boldsymbol{\nu}} \mathcal{L}(\mathrm{H}_{\mathbf{w},\,\boldsymbol{\nu}}|\mathcal{D},\,\mathcal{A})}{\max\limits_{\boldsymbol{\nu}} \mathcal{L}(\mathrm{R}_{\boldsymbol{\nu}}|\mathcal{D},\,\mathcal{A})}\right] = 2\log\left[\frac{\max\limits_{\mathbf{w},\,\nu} \mathcal{L}(\mathrm{H}_{\mathbf{w},\,\boldsymbol{\nu}}|\mathcal{D})\mathcal{L}(\boldsymbol{\nu}|\mathcal{A})}{\max\limits_{\boldsymbol{\nu}} \mathcal{L}(\mathrm{R}_{\boldsymbol{\nu}}|\mathcal{D})\mathcal{L}(\boldsymbol{\nu}|\mathcal{A})}\right]$$



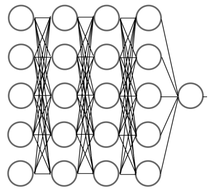
Inform the model about *known* data transformations

Uncertainties-aware parametrization of the alternative hypothesis:

$$n(x\,|\,\mathrm{H}_{\mathbf{w},\,\nu}) = n(x\,|\,\mathrm{R}_0)\;\frac{n(x\,|\,\mathrm{R}_\nu)}{n(x\,|\,\mathrm{R}_0)}\;e^{f(x;\,\mathbf{w})}$$

central value (ν=0)

NN model



Model of systematic uncertainties effects

$$\hat{r}(x;\nu) = \frac{n(x\,|\,\mathrm{R}_\nu)}{n(x\,|\,\mathrm{R}_0)} = \exp\left[\hat{\delta}_1(x)\,\nu + \hat{\delta}_2(x)\,\nu^2 + \ldots\right]$$

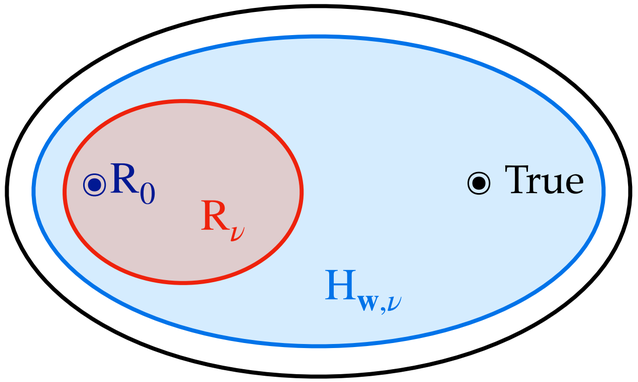
"Learning New Physics from an Imperfect Machine"
Eur. Phys. J. C (2021) (Grosso, d'Agnolo, Wulzer, Zanetti, Pierini)

Dealing with *systematic uncertainties*

Frequentist approach (discovery at the LHC)

$$t(\mathcal{D},\mathcal{A}) = 2\log\left[\frac{\max\limits_{\mathbf{w},\nu}\mathcal{L}(\mathrm{H}_{\mathbf{w},\boldsymbol{\nu}}|\mathcal{D})\mathcal{L}(\boldsymbol{\nu}|\mathcal{A})}{\max\limits_{\boldsymbol{\nu}}\mathcal{L}(\mathrm{R}_{\boldsymbol{\nu}}|\mathcal{D})\mathcal{L}(\boldsymbol{\nu}|\mathcal{A})}\right]\cdot\frac{\mathcal{L}(\mathrm{R}_{\mathbf{0}}|\mathcal{D})\mathcal{L}(\mathbf{0}|\mathcal{A})}{\mathcal{L}(\mathrm{R}_{\mathbf{0}}|\mathcal{D})\mathcal{L}(\mathbf{0}|\mathcal{A})}$$

$$= \tau(\mathcal{D},\mathcal{A}) - \Delta(\mathcal{D},\mathcal{A})$$



Tau term:

$$\tau(\mathcal{D},\mathcal{A}) = 2\max_{\mathbf{w},\boldsymbol{\nu}}\log\left[\frac{\mathcal{L}(\mathrm{H}_{\mathbf{w}},\boldsymbol{\nu}|\mathcal{D})\,\mathcal{L}(\boldsymbol{\nu}|\mathcal{A})}{\mathcal{L}(\mathrm{R}_{\mathbf{0}}|\mathcal{D})\,\mathcal{L}(\mathbf{0}|\mathcal{A})}\right] = -2\min_{\mathbf{w},\boldsymbol{\nu}} L\left[f(x,\mathbf{w}),\, r(x;\boldsymbol{\nu})\right]$$

NN model

Depends on the NN model, sensitive to unknown distortions

Delta term:

$$\Delta(\mathcal{D},\mathcal{A}) = 2\max_{\boldsymbol{\nu}}\log\left[\frac{\mathcal{L}(\mathrm{R}_{\boldsymbol{\nu}}|\mathcal{D})\,\mathcal{L}(\boldsymbol{\nu}|\mathcal{A})}{\mathcal{L}(\mathrm{R}_{\mathbf{0}}|\mathcal{D})\,\mathcal{L}(\mathbf{0}|\mathcal{A})}\right] = -2\min_{\boldsymbol{\nu}} L\left[r(x;\boldsymbol{\nu})\right]$$

Sensitive only to uncertainties related discrepancies

"Learning New Physics from an Imperfect Machine"
Eur. Phys. J. C (2021) (Grosso, d'Agnolo, Wulzer, Zanetti, Pierini)

Example: Signal-agnostic tool for New Physics searches

Controlling false positives (aka test calibration)

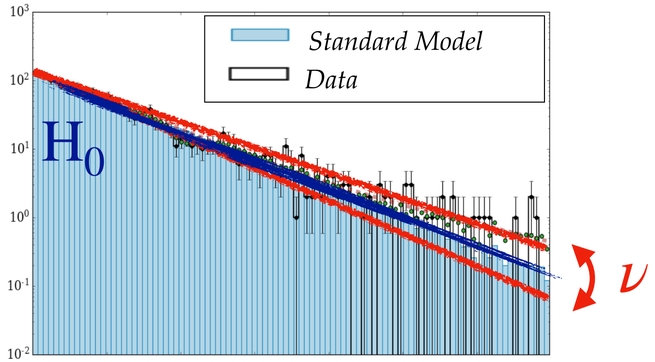
Validation of the algorithm

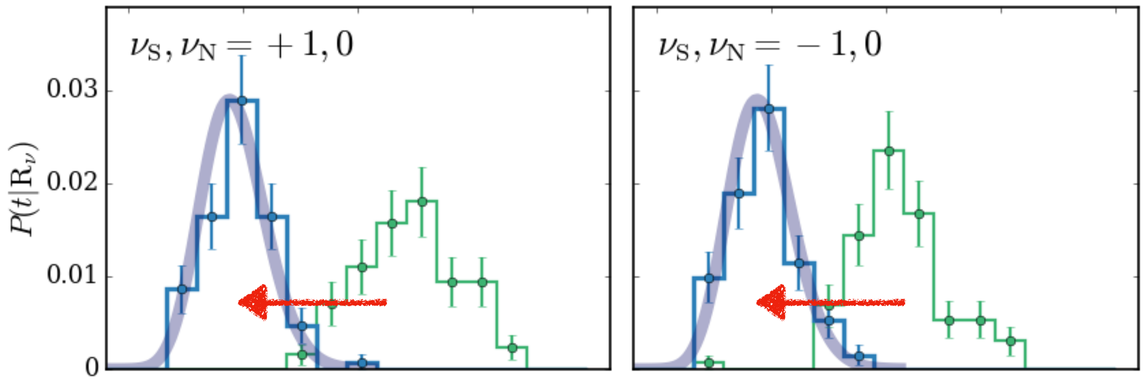
$\tau(D,\ A)$

$t(D,\ A) = \tau(D,\ A) - \Delta(D,\ A)$

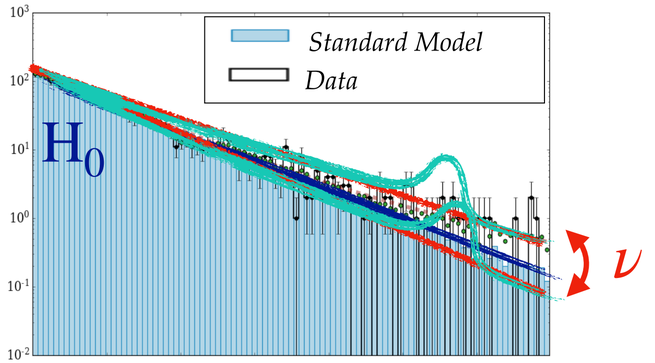
$t(D, A)$ in absence of distortions

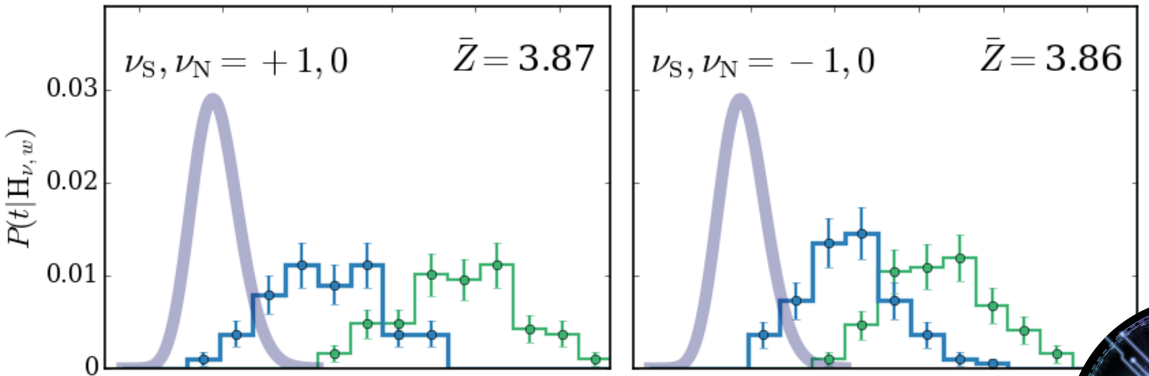
1. Data deformations within systematic uncertainties





2. Data deformations within systematic uncertainties + signal events





Dealing with *systematic uncertainties*

Data-driven reference samples

Estimating the background model from the control regions

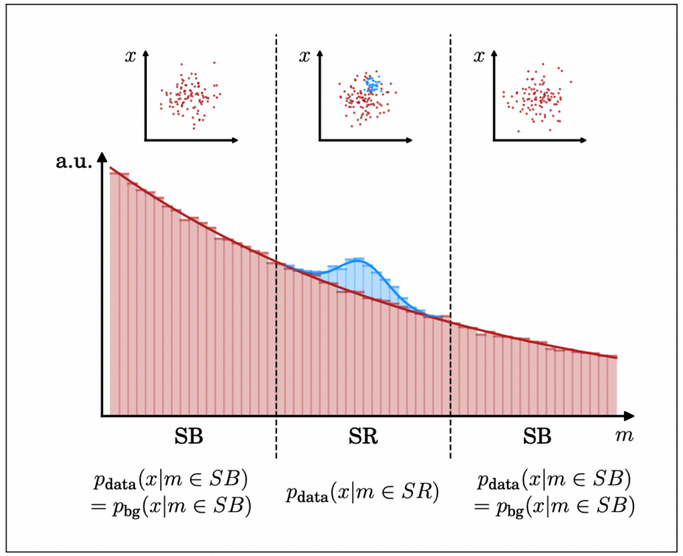
- Many systematic effects cancel out (good when possible!)
- Must make assumptions about the signal location (e. g. Bump hunt)
- Uncertainties on transfer functions

CURTAINs (2022) flow matching

CATHODE (2021) conditional normalizing flow

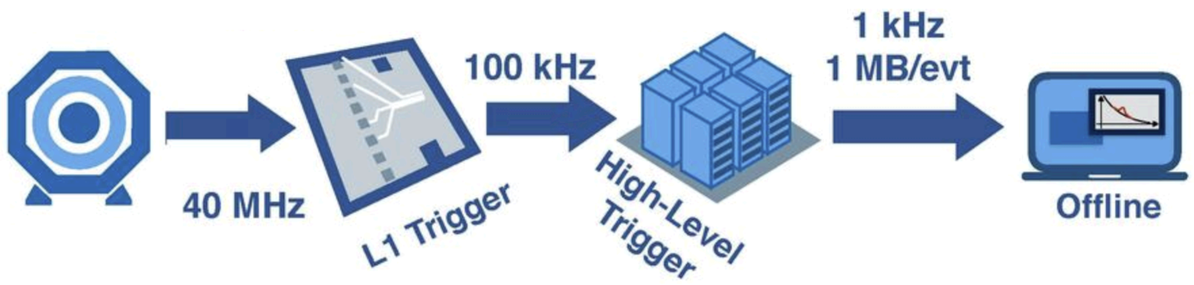
SALAD (2020) simulation assisted (reweighing + morphing)

[…]



Example from CATHODE (2021)

Challenges of statistical anomaly detection at the LHC

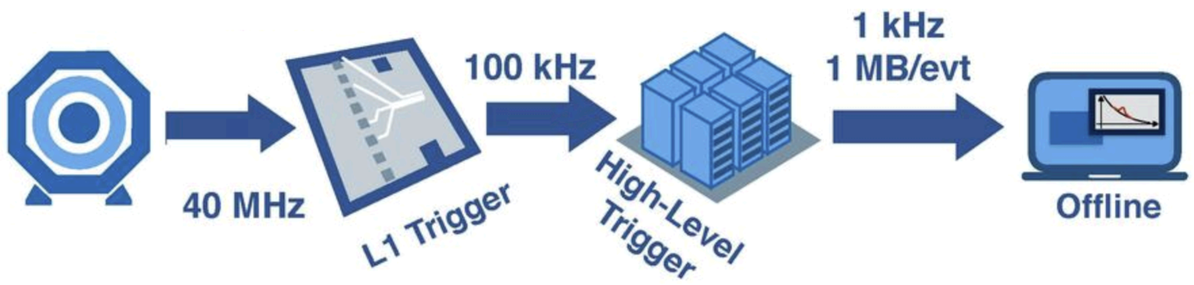


$$t = t(D, R_\nu)$$

Are the input data *informative* about existing anomalies?

- DATA REPRESENTATION
- DATA COMPRESSION
- SCALABILITY

Challenges of statistical anomaly detection at the LHC



$$t = t(D, R_\nu)$$

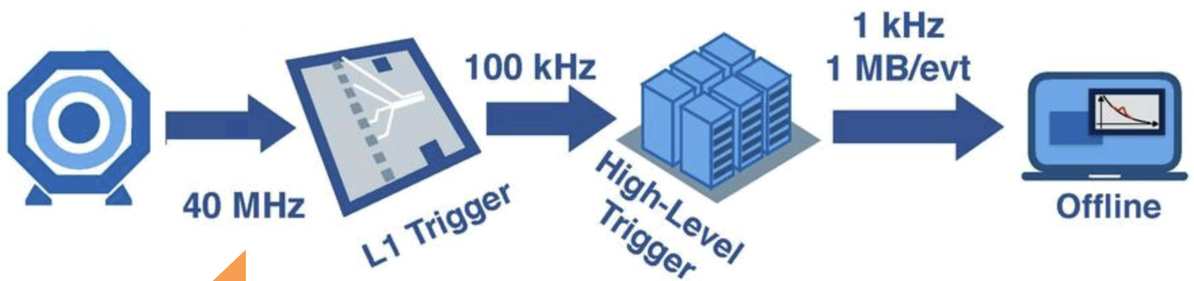
Are the input data *informative* about existing anomalies?

- DATA REPRESENTATION
- DATA COMPRESSION
- SCALABILITY

The nature of our experiments is intrinsically and unavoidably *biased*:

- ⊙ Detector design
- ⊙ Data selection (triggers)
- ⊙ Reconstruction
- ⊙ Data dimensionality reduction
- ⊙ Statistical strategy

Challenges of statistical anomaly detection at the LHC



$$t = t(D, R_\nu)$$

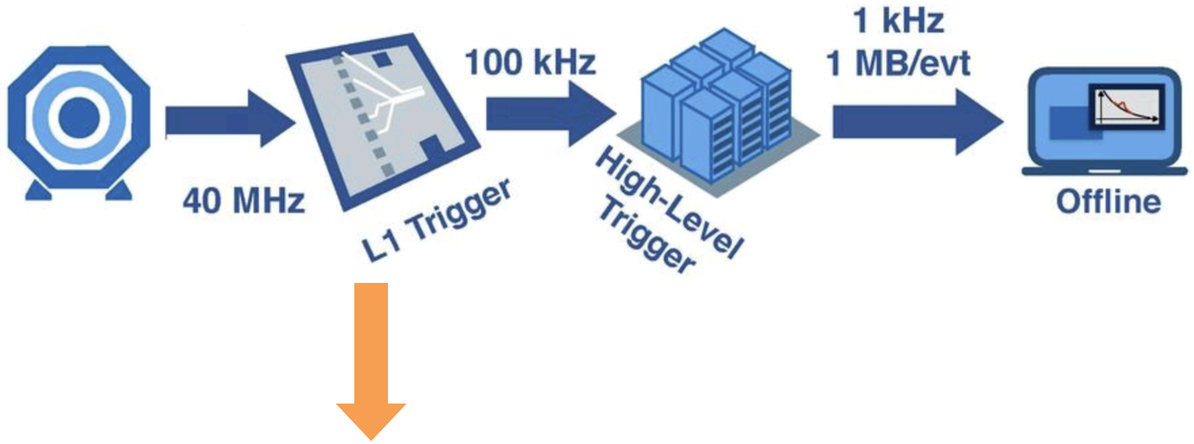
Are the input data *informative* about existing anomalies?

- DATA REPRESENTATION
- DATA COMPRESSION
- SCALABILITY

Loose selections:

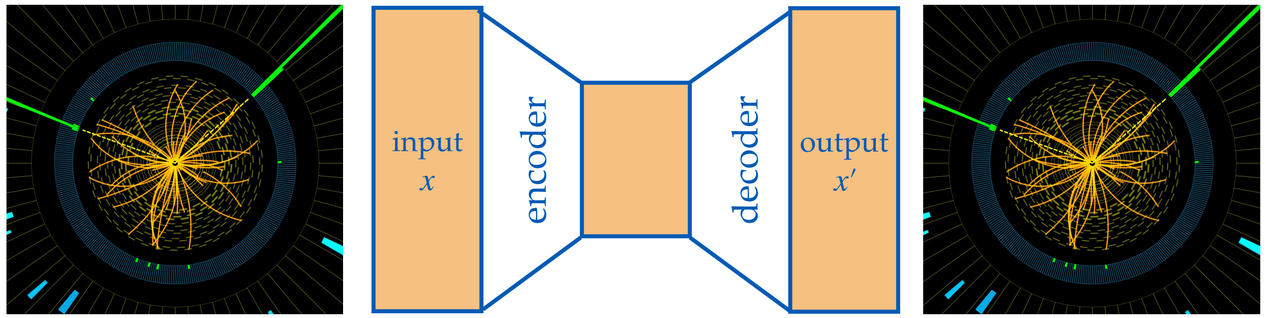
- Higher dimensional spaces
- Smarter triggers / Trigger-less readout
  - Increased signal acceptance (though highly diluted!)
  - Much larger datasets (computationally demanding!)

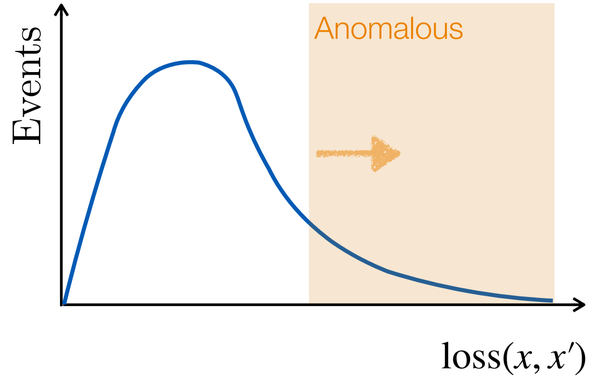
The problem of *unbiased data extraction* (at the LHC)



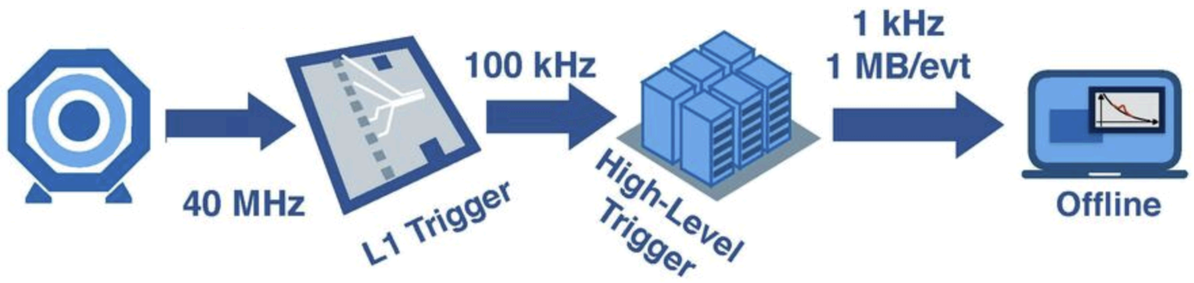
Save an alternative stream of data based on anomaly detection (See Thea's talk)

Anomaly detection at **trigger level**



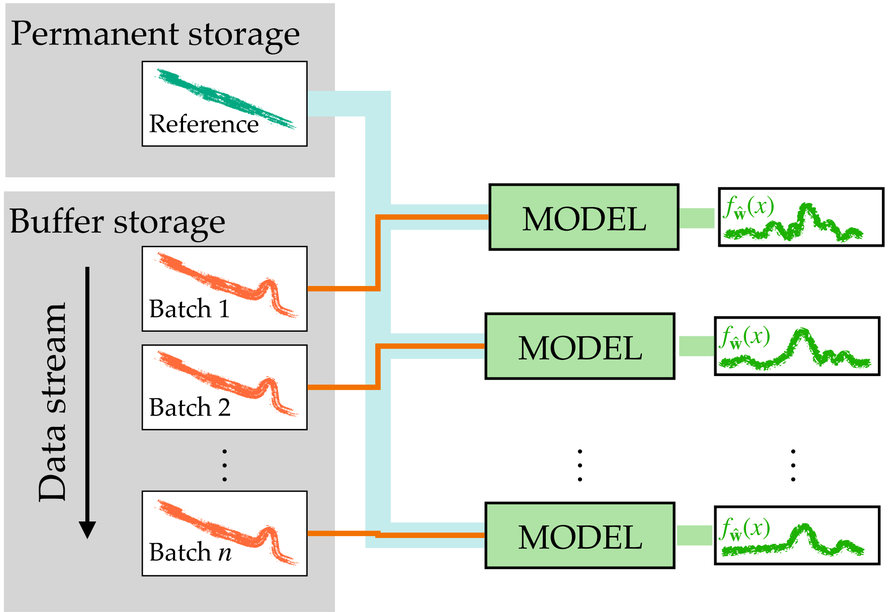


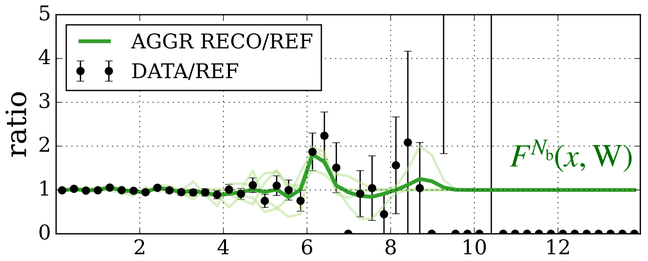
The problem of *unbiased data extraction* (at the LHC)



How to recover sensitivity of the full statistics?

Handling large datasets in batches





combination preserving the *local* structures

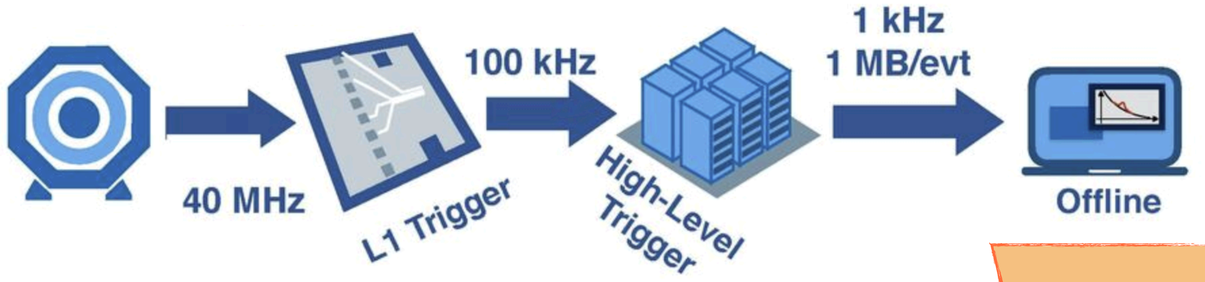
f-aggregation

$$F^{N_b}(x; \mathbf{W}) = \log \frac{n(x; \mathrm{H}_{\mathbf{w}}^{N_b})}{n(x; \mathrm{R})} = \log \left[\frac{1}{N_b} \sum_{i=1}^{N_b} e^{f_i(x; \mathbf{w})} \right]$$

"Anomaly-aware summary statistic from data batches"
arXiv:2407:01249 (Grosso)

The problem of *unbiased data extraction* (at the LHC)

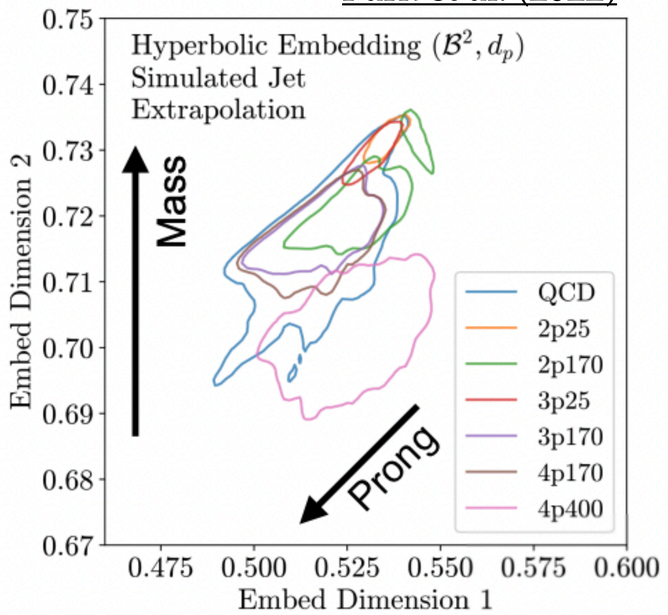
Park et al. (2022)
Dillon et al. (2024).
Li et al (2024)
Halling et al (2024)
Golling et al. (2024)



Anomaly preserving dimensionality reduction

$$t = t(D, R_\nu)$$

Park et al. (2022)



Data organization via self-supervision.
How to make self-supervision anomaly-preserving?
How to preserve interpretability?

Personal thought

AI has the potential to extend anomaly detection
beyond human abilities,
trading assumed knowledge for a richer basis of
acquired experimental evidence.

The design of rigorous statistical frameworks is essential to trust the tools and tackle scientific discoveries.

LHC challenges as a playground for STAT and CS:

- Interpretability
- Data compression
- Data representation
- Uncertainty quantification

Great opportunity for knowledge transfer!

Backup slides

New Physics Learning Machine (NPLM)
ML-based signal-agnostic likelihood-ratio test

"Learning New Physics from a Machine" Phys. Rev. D (d'Agnolo, Wulzer)
"Learning Multivariate New Physics" Eur. Phys. J. C 81, 89 (2021) (d'Agnolo, Grosso, Pierini, Wulzer, Zanetti)
"Learning New Physics from an imperfect machine" Eur. Phys. J. C 82, 275 (2022) (d'Agnolo, Grosso, Pierini, Wulzer, Zanetti)
"Goodness of fit by Neyman-Pearson testing" SciPostPhys.16.5.123 (Grosso, Letizia, Wulzer, Pierini)
"Learning New Physics Efficiently with non-parametric models" Eur. Phys. J. C 82, 879 (2022) (Losapio, Letizia, Grosso, et al.)
"Fast kernel methods for data quality monitoring as a goodness-of-fit test" Mach.Learn.Sci.Tech. 4 (2023) 3, 035029 (Lai, Grosso, Letizia, et al.)
"Anomaly-aware summary statistic from data batches" arXiv:2407:01249 (Grosso)
"Multiple testing for signal-agnostic searches of new physics with machine learning" 2408.12296 (Grosso, Letizia)

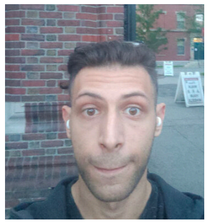
Work in collaboration with:

R.T. d'Agnolo M. Pieirini A. Wulzer M. Zanetti N. Lai M. Letizia L. Rosasco M. Rando





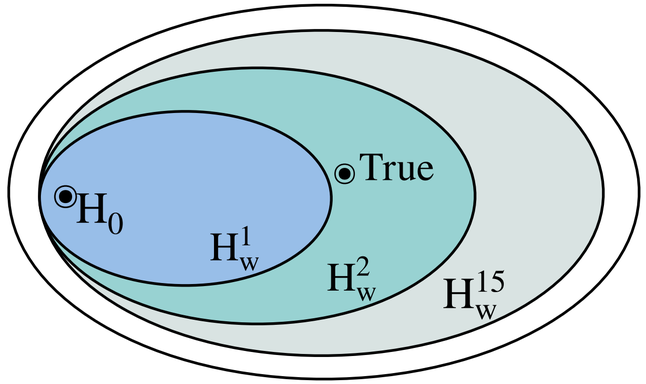


The problem of *model selection*

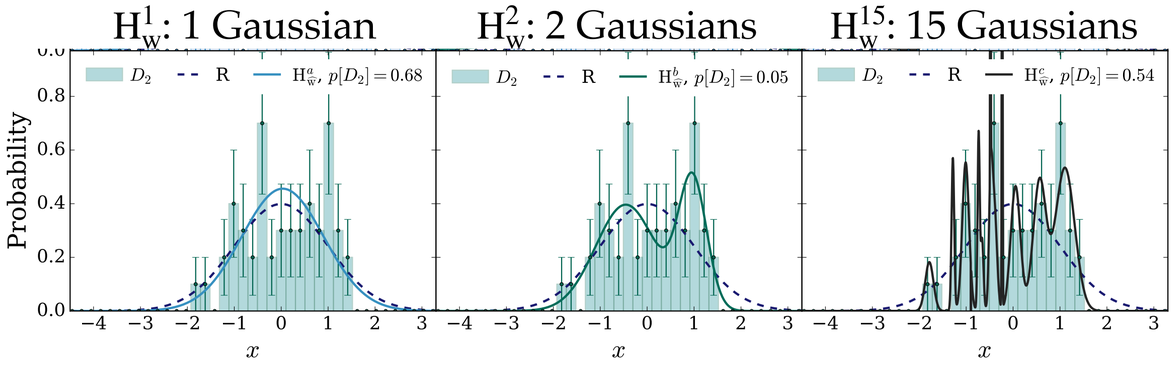
How to define the family of universal approximants?

$$f(x, w) = \log \left[\frac{n(x \mid D)}{n(x \mid R)} \right]$$



- Too simple models may not contain an element that approximate the True hypothesis well enough
- Too complex models could be over-sensitive to statistical fluctuations, independently on the nature of the data

Brute force scaling is not the solution

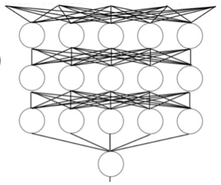


The problem of *model selection*

Regularization and statistical robustness

Neural Network model

$f(x, w) = \mathrm{DNN}(x, w)$



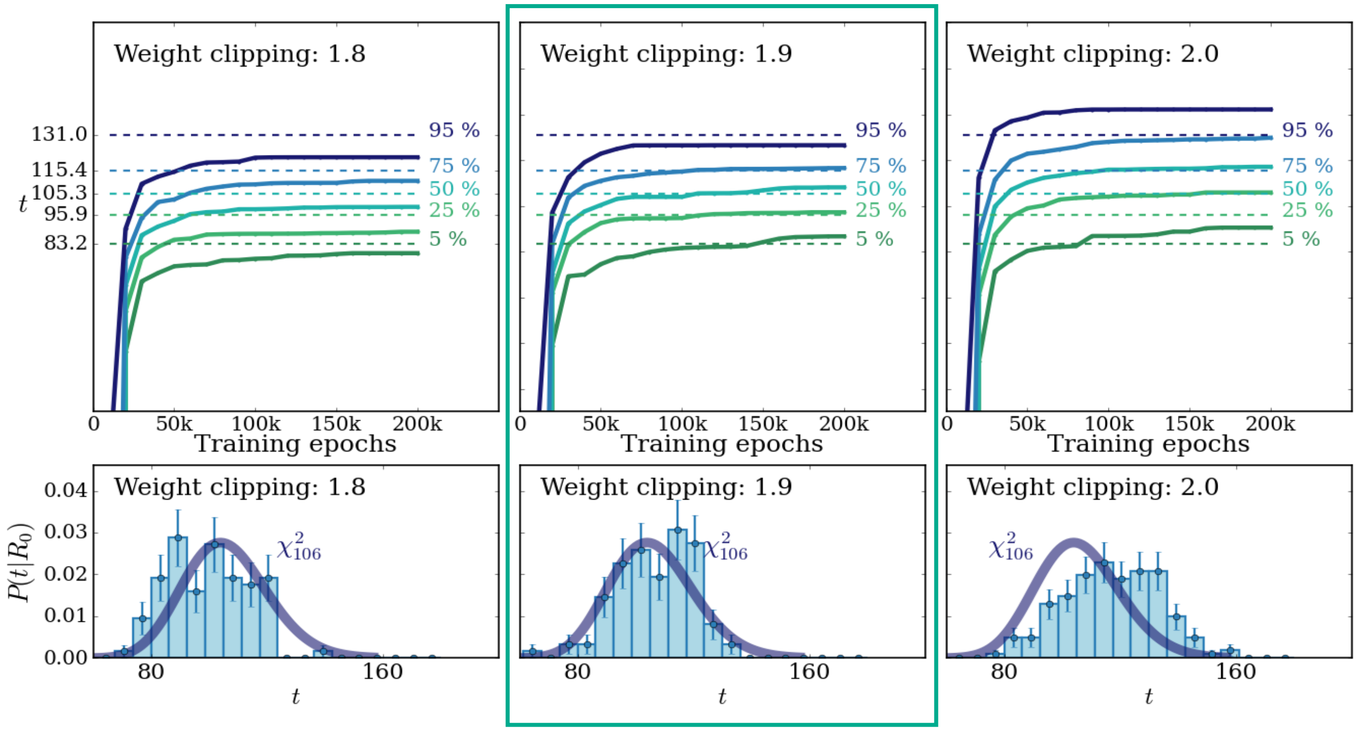
Weight clipping parameter:
Upper boundary to the magnitude that each trainable parameter can assume during the training.

$|w| = \max(|w|, w_{clip})$

For a chosen NN architecture, tuning the weight clipping allows to recover a good agreement of the empirical distribution of t under R_0 with a **target** χ^2_{df} distribution (df=# trainable parameters)

Legend:
- Percentiles of the empirical $\bar{t}$ distribution under R_0
- Percentiles of the target $\chi^2_{|\mathbf{w}|}$
- Empirical $\bar{t}$ distribution under R_0
- Target $\chi^2_{|\mathbf{w}|}$

Example:
NN model: 5-7-7-1, Number of parameters:106

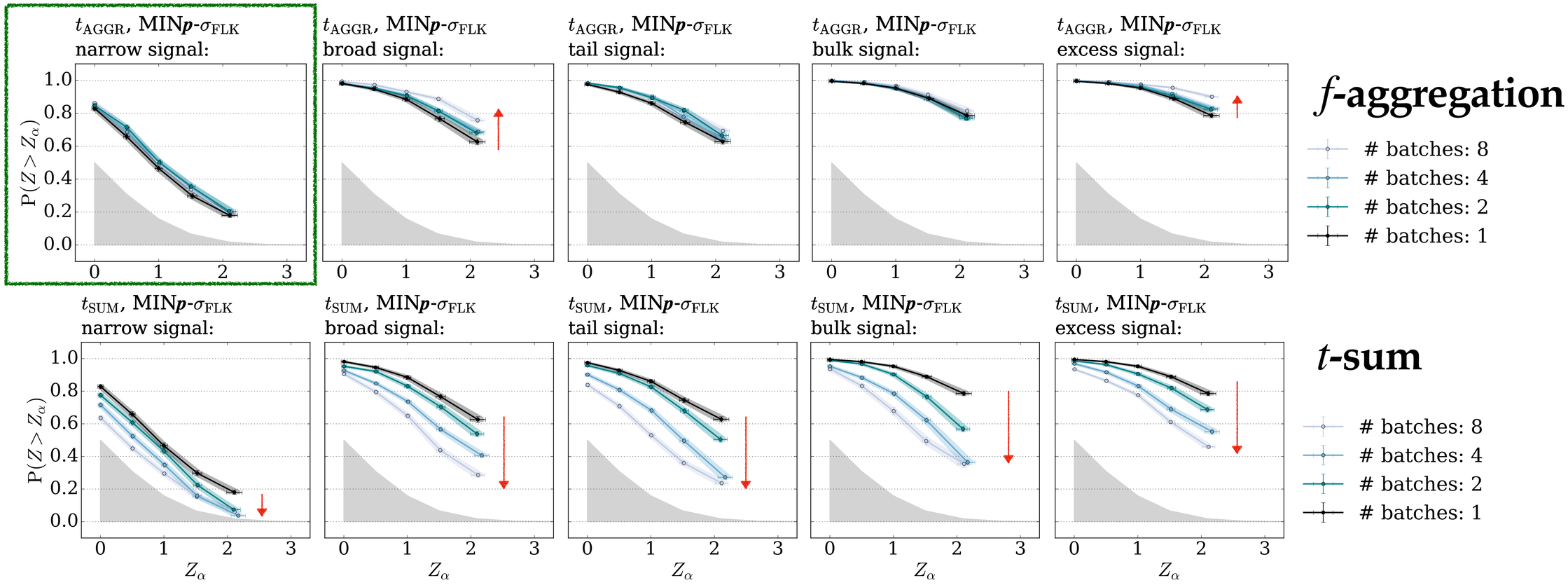


"Learning Multivariate New Physics" Eur. Phys. J. C (Grosso et al.)

A *batched* approach to likelihood-ratio with NPLM

1D experiments:

Previous slide



"Anomaly-aware summary statistic from data batches"
arXiv:2407:01249 (Grosso)

A *batched* approach to likelihood-ratio with NPLM

Neyman-Pearson test over the batches:

$$t_{id}(\mathcal{D} = \cup_i \mathcal{D}_i) = 2 \sum_{i=1}^{N_b} \log \frac{\mathcal{L}(\mathcal{D}_i|\mathrm{T})}{\mathcal{L}(\mathcal{D}_i|\mathrm{R})}$$

T: true hypothesis underlying the data (unknown)
R: reference hypothesis (null)

If $\mathrm{H}^1_\mathrm{w} \approx \mathrm{H}^2_\mathrm{w} \approx \ldots \approx \mathrm{H}^\mathrm{i}_\mathrm{w} \approx \mathrm{T}$

Then **summing** *t* over the batches is a good approximation of the Neyman-Pearson test:

$$t^{N_b}_{\mathrm{SUM}}(\mathcal{D}) = \sum_{i=1}^{N_b} t(\mathcal{D}_i) = 2 \sum_{i=1}^{N_b} \log \frac{\mathcal{L}(\mathcal{D}_i|\mathrm{H}^i_\mathbf{w})}{\mathcal{L}(\mathcal{D}_i|\mathrm{R})} = 2 \sum_{i=1}^{N_b} \left[\sum_{x \in \mathcal{R}} w_\mathcal{R}(x)(1 - e^{f_{i,\hat{\mathbf{w}}}(x)}) + \sum_{x \in \mathcal{D}_i} f_{i,\hat{\mathbf{w}}}(x) \right]$$

But learning in batches is **not** optimal due to the reduced size of the training set.

The batch size determines the sensitivity and we don't get any gain by summing over *t*!

"Anomaly-aware summary statistic from data batches"
arXiv:2407:01249 (Grosso)

A *batched* approach to likelihood-ratio with NPLM

Neyman-Pearson test:

$$t_{id}(\mathcal{D} = \cup_i \mathcal{D}_i) = 2 \sum_{i=1}^{N_b} \log \frac{\mathcal{L}(\mathcal{D}_i|\mathrm{T})}{\mathcal{L}(\mathcal{D}_i|\mathrm{R})}$$

T: true hypothesis underlying the data (unknown)
R: reference hypothesis (null)

Since $\mathrm{H}^1_\mathrm{w} \approx \mathrm{H}^2_\mathrm{w} \approx \ldots \approx \mathrm{H}^\mathrm{i}_\mathrm{w}$ are all *distorted views* of T

Then combining the information at the level of the learnt model allows to build a unique alternative *shared* among batches

$$n(x; \mathrm{H}^{N_b}_\mathbf{w}) = \frac{1}{N_b} \sum_{i=1}^{N_b} n(x; \mathrm{H}^i_\mathbf{w}) \longrightarrow F^{N_b}(x; \mathbf{W}) = \log \frac{n(x; \mathrm{H}^{N_b}_\mathbf{w})}{n(x; \mathrm{R})} = \log \left[\frac{1}{N_b} \sum_{i=1}^{N_b} e^{f_i(x;\mathbf{w})} \right]$$

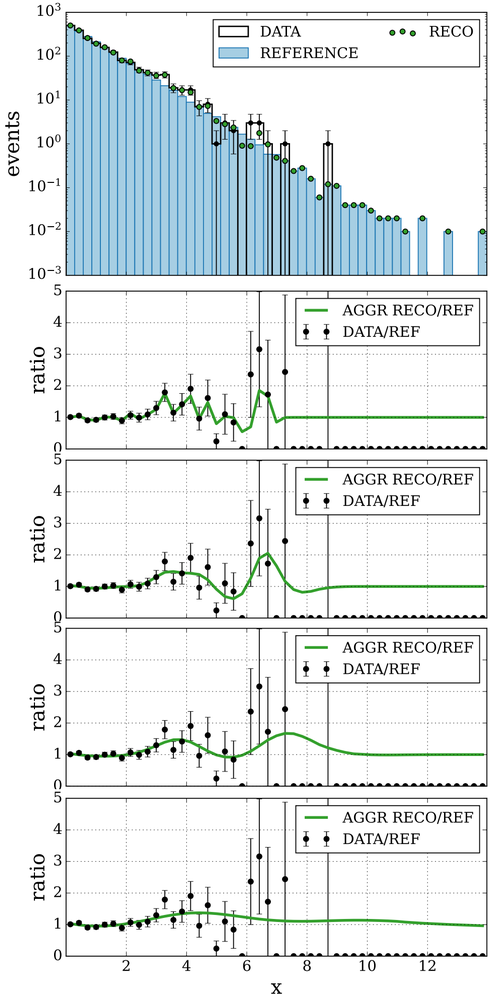
And use it a basis for a likelihood-ratio test

$$t^{N_b}_{\mathrm{AGGR}}(\mathcal{D}) = 2 \sum_{i=1}^{N_b} \log \frac{\mathcal{L}(\mathcal{D}_i|\mathrm{H}^{N_b}_\mathbf{w})}{\mathcal{L}(\mathcal{D}_i|\mathrm{R})} = 2 \sum_{i=1}^{N_b} \left[\sum_{x \in \mathcal{R}} w_\mathcal{R}(x)(1 - e^{F^{N_b}(x;\mathbf{W})}) + \sum_{x \in \mathcal{D}_i} F^{N_b}(x; \mathbf{W}) \right]$$

"Anomaly-aware summary statistic from data batches"
arXiv:2407:01249 (Grosso)

The problem of *model selection*

Multiple testing for robust statistical anomaly detection



Signal benchmarks: gaussian resonances with various width (σ_{NP}) and locations ($\bar{x}_{NP}$). [N(S) signal events injected over N(B) = 2000 events]

NP test for different kernel widths (σ)

Aggregation methods

N(S)	7	18	13	10	90
$\bar{x}_{NP}$	4	4	4	6.4	1.6
σ_{NP}	0.01	0.16	0.64	0.16	0.16
$\sigma = 0.01$	0.004 ± 0.002	0.0015 ± 0.0008	0.0005 ± 0.0004	0.001 ± 0.001	0.042 ± 0.005
$\sigma = 0.3$	$\mathbf{0.011 \pm 0.002}$	$\mathbf{0.094 \pm 0.007}$	0.005 ± 0.002	0.222 ± 0.009	0.62 ± 0.01
$\sigma = 0.7$	0.005 ± 0.002	$\mathbf{0.094 \pm 0.007}$	$\mathbf{0.008 \pm 0.002}$	$\mathbf{0.31 \pm 0.01}$	$\mathbf{0.65 \pm 0.01}$
$\sigma = 1.4$	0.004 ± 0.001	0.060 ± 0.005	0.007 ± 0.002	0.26 ± 0.01	0.49 ± 0.01
$\sigma = 4.5$	0.0015 ± 0.0008	0.013 ± 0.003	0.005 ± 0.002	0.062 ± 0.005	0.197 ± 0.009
$\sigma = 9.0$	0.0005 ± 0.0004	0.006 ± 0.002	0.004 ± 0.001	0.027 ± 0.004	0.117 ± 0.007
$\sigma = 2.3$	0.003 ± 0.001	0.044 ± 0.004	0.013 ± 0.002	0.193 ± 0.009	0.36 ± 0.01
min-p	$\mathbf{0.012 \pm 0.002}$	$\mathbf{0.106 \pm 0.007}$	$\mathbf{0.009 \pm 0.002}$	$\mathbf{0.32 \pm 0.01}$	$\mathbf{0.67 \pm 0.01}$
prod-p	0.006 ± 0.002	0.072 ± 0.006	$\mathbf{0.011 \pm 0.002}$	0.24 ± 0.01	0.63 ± 0.01
avg-p	0.004 ± 0.001	0.051 ± 0.005	$\mathbf{0.011 \pm 0.002}$	0.072 ± 0.006	0.50 ± 0.01
smax-t	0.002 ± 0.001	0.002 ± 0.001	0.002 ± 0.001	0.0015 ± 0.0008	0.001 ± 0.001

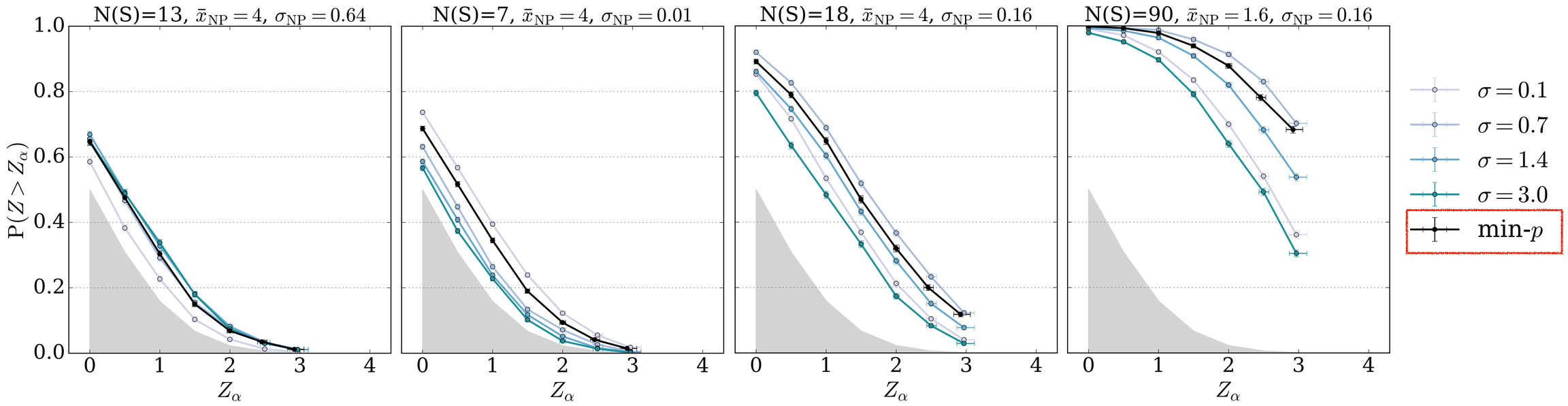
Table 1: EXPO 1D – probability of observing $Z \geq 3$.

Single tests have different power over the signal benchmarks.
The aggregation shows **uniform enhanced power** over the range of benchmarks.

"Multiple testing for signal-agnostic searches of new physics with machine learning" 2408.12296 (Grosso, Letizia)

The problem of *model selection*

Multiple testing for robust Neyman-Pearson two-sample test



Robust sensitivity among different signal patterns

"Multiple testing for signal-agnostic searches of new physics with machine learning" 2408.12296 (Grosso, Letizia)

A *batched* approach to likelihood-ratio

1D toy model

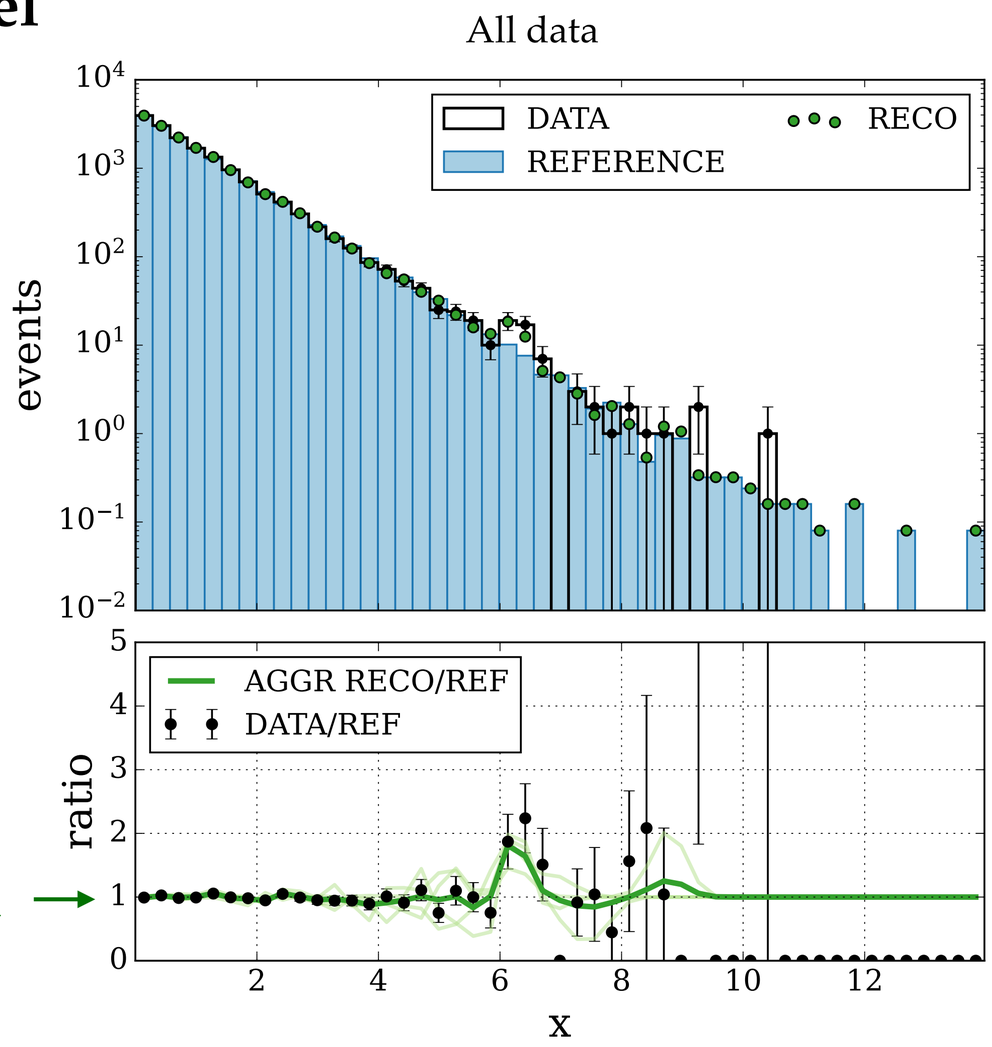
Single batch:

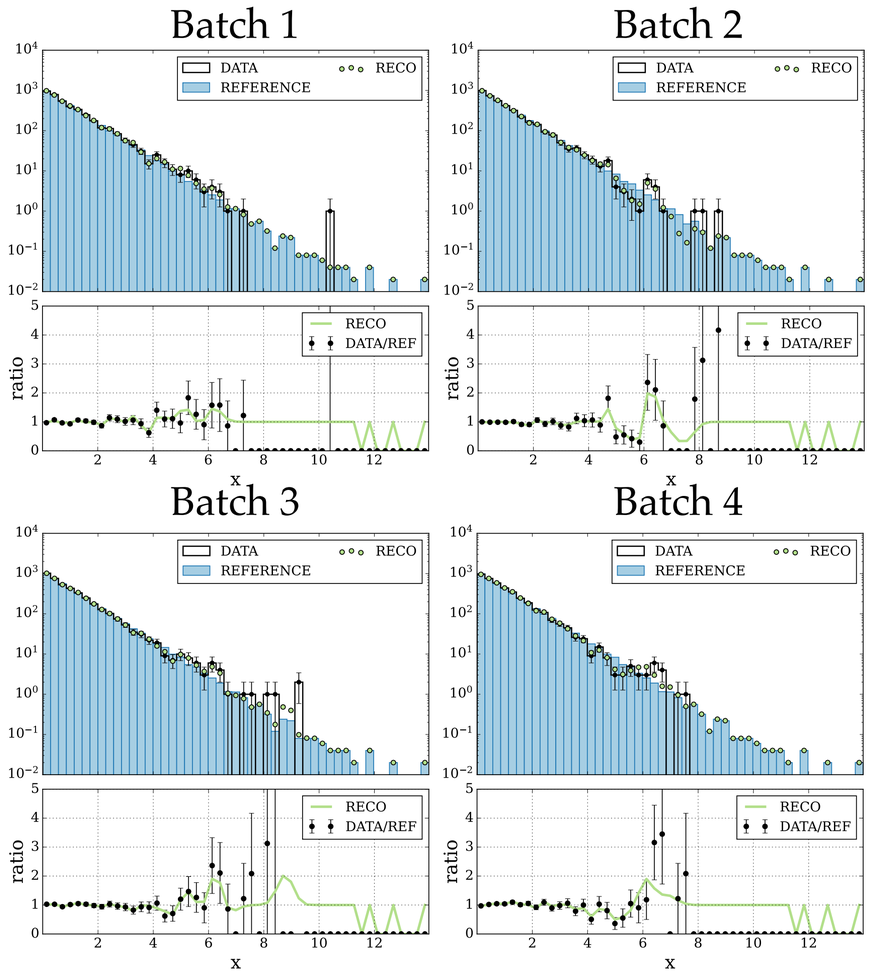
- N(B) = 4 000
- N(S) = 6
- $Z_{ideal} = 2.4$

All data:

- N(B) = 16 000
- N(S) = 24
- $Z_{ideal} = 4.8$

Averaged model →

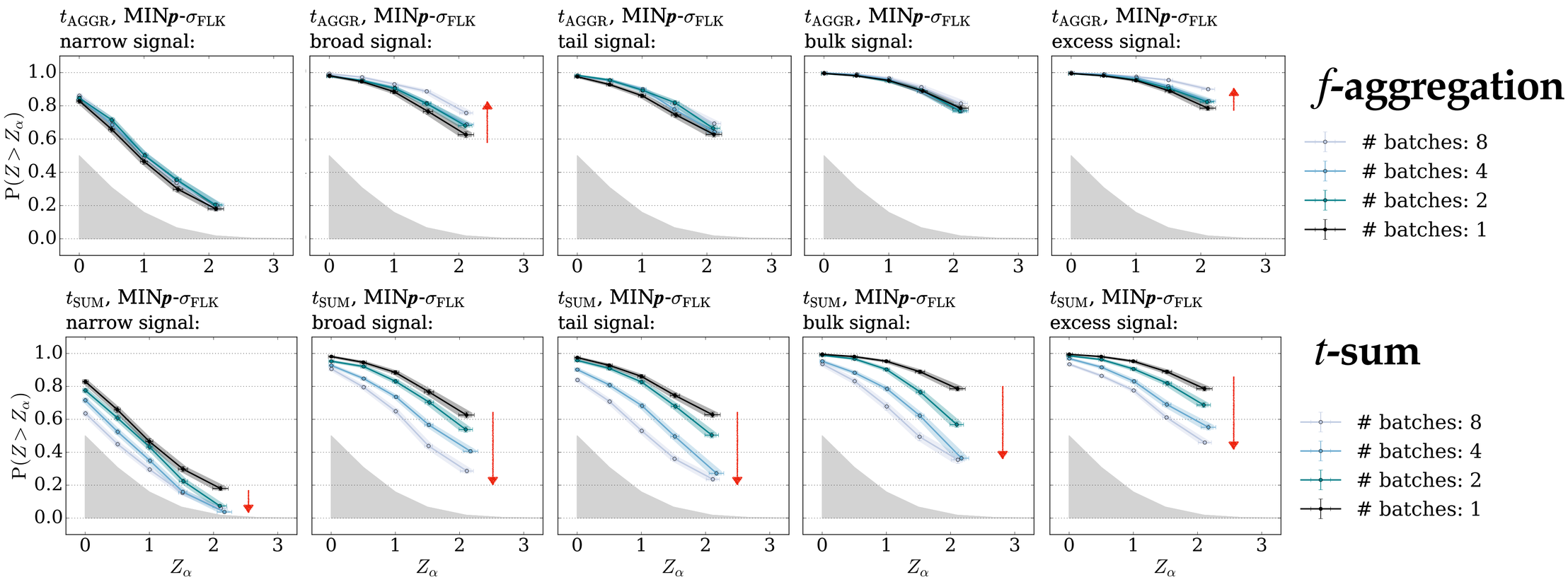




"Anomaly-aware summary statistic from data batches"
arXiv:2407:01249 (Grosso)

A *batched* approach to Neyman-Pearson test

1D experiments:



***f*-aggregation**

- # batches: 8
- # batches: 4
- # batches: 2
- # batches: 1

***t*-sum**

- # batches: 8
- # batches: 4
- # batches: 2
- # batches: 1

"Anomaly-aware summary statistic from data batches"
arXiv:2407:01249 (Grosso)

Less is more: *sparse* kernel methods with dictionary learning

Expressive, *regularized* and *interpretable* models for **statistical anomaly detection**

Gaia Grosso[1,2,3], Demba Ba[1,2], Phil Harris[1,3]

[1]NSF Institute for Artificial Intelligence and Fundamental Interaction (IAIFI)

[2]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA [3]MIT Laboratory for Nuclear Science, Cambridge, MA

EUROPEAN AI FOR FUNDAMENTAL PHYSICS CONFERENCE

EuCAIFCon 2024

SOLUTION

Sparse linear combination of Gaussian Kernels (SGK)

$$f_{\mu,w}(x) = \sum_{i=1}^{M} w_i \, k(x;\, \mu_i, \sigma_i)$$

Local interpretability

Active kernels highlight anomalous regions

$$k(x;\, \mu_i, \sigma_i) = \mathrm{A} \exp\left[-\frac{||x-\mu_i||^2}{2\sigma_i^2}\right]$$

Sparse model ($M \ll N$)

competition between data points to attract the kernels

Adaptive model (learnable μ)

directing *attention* to anomalous features

Smooth model ($\sigma^2 = \sigma^2_{\mathrm{exp}} + \sigma^2_X$)

Physics constraints (e.g. experimental resolution).
What is the scale of New Physics?