June 17th, 2025

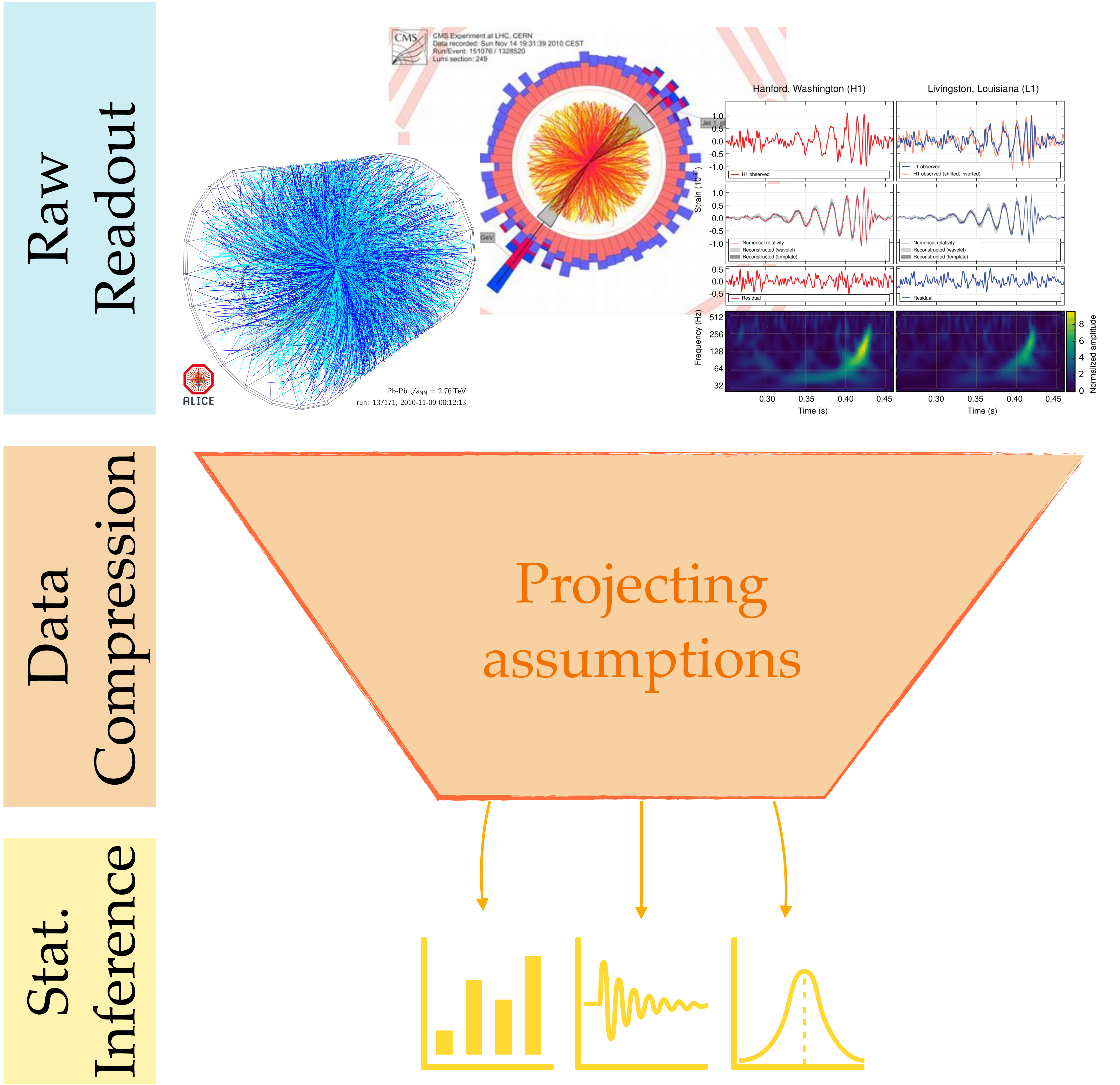
From High Dimensions to Statistical Discovery: a Contrastive Learning Approach to Anomaly Detection

Gaia Grosso[1,2], Kyle Metzger[3], Sam Bright-Thonney[1,2], Christina Reissel[2], Thea Aarrestad[3], Phil Harris[1,2], Ekaterina Govorkova[2], Andrzej Novak[2], Nathaniel Woodward[2], Eric Moreno[2], Lana Xu[2]

[1]Institute for Artificial Intelligence and Fundamental Interactions (IAIFI), [2]Massachusetts Inst, of Technology (MIT), [3]ETH Zurich

Inclusiveness vs assumptions: the never ending struggle

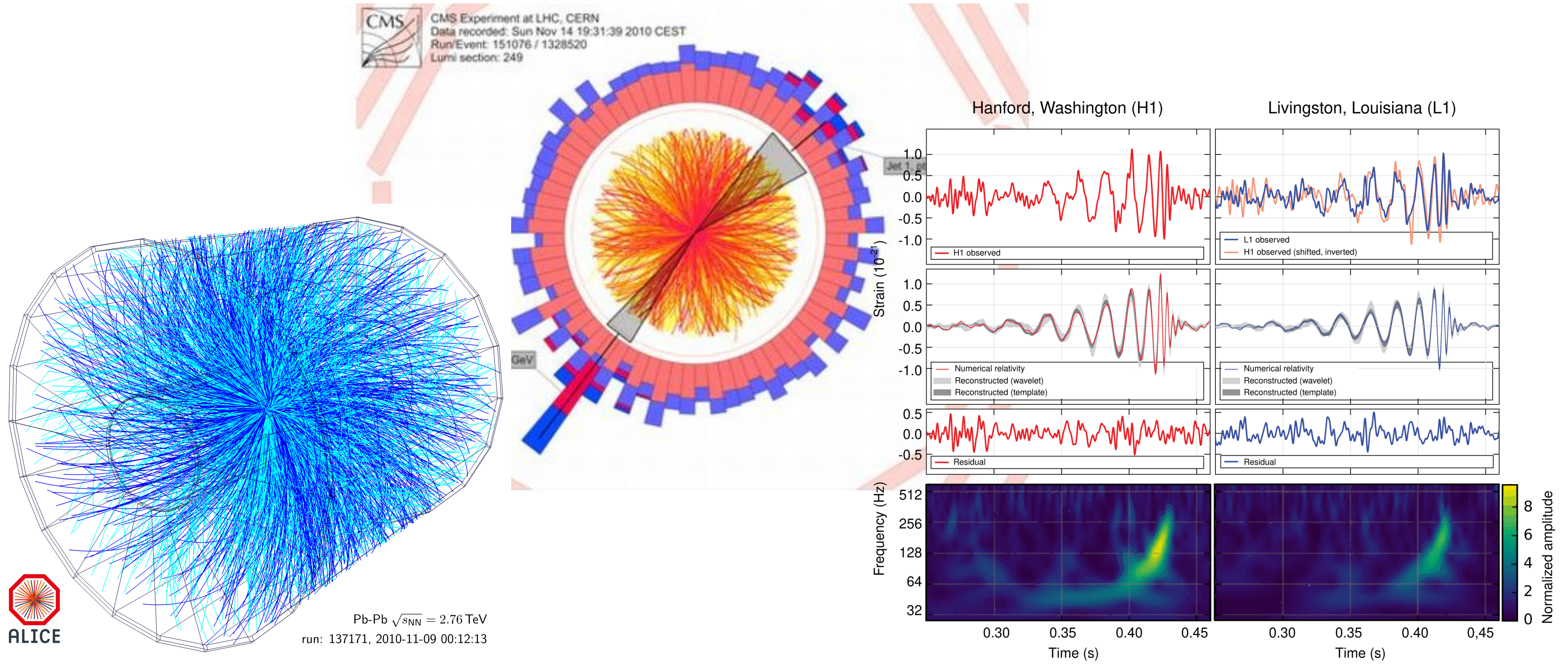
- Raw experimental HEP data are high dimensional and highly structured
- Statistical analysis are only possible after a data compression step which reduces complexity by projecting data in a lower dimensional representation.
- One has to make assumptions about which information is relevant to the downstream task (domain knowledge, current understanding of HEP).



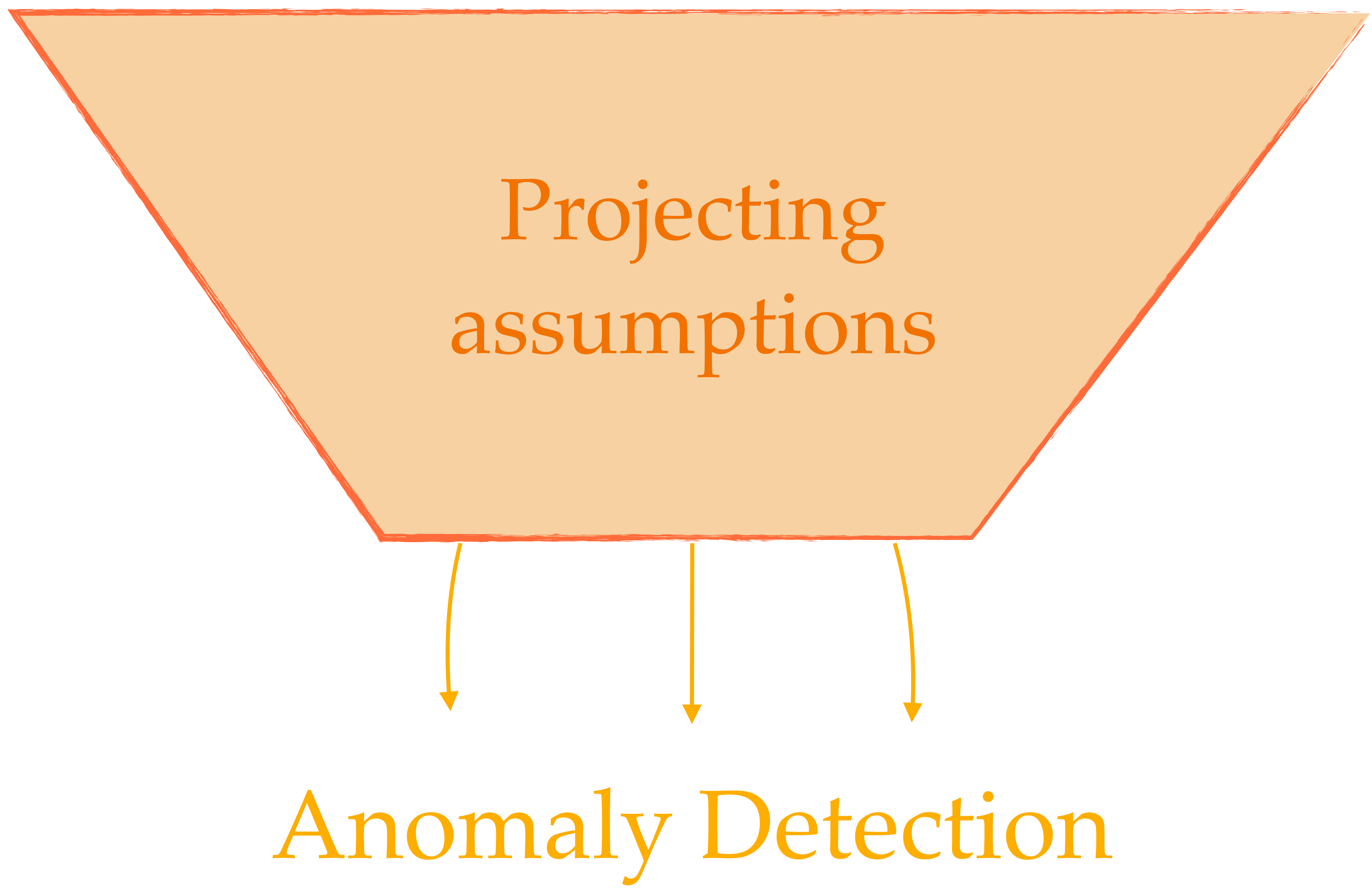
Inclusiveness vs assumptions: the never ending struggle

- In absence of a signal model (unsupervised anomaly detection) the best assumptions are not known a priori: we unavoidably induce a sensitivity loss to some signal by imposing lossy projections.
- How can we mitigate this phenomenon?

Raw Readout



Data Compression



Stat. Inference

Inclusiveness vs assumptions: the never ending struggle

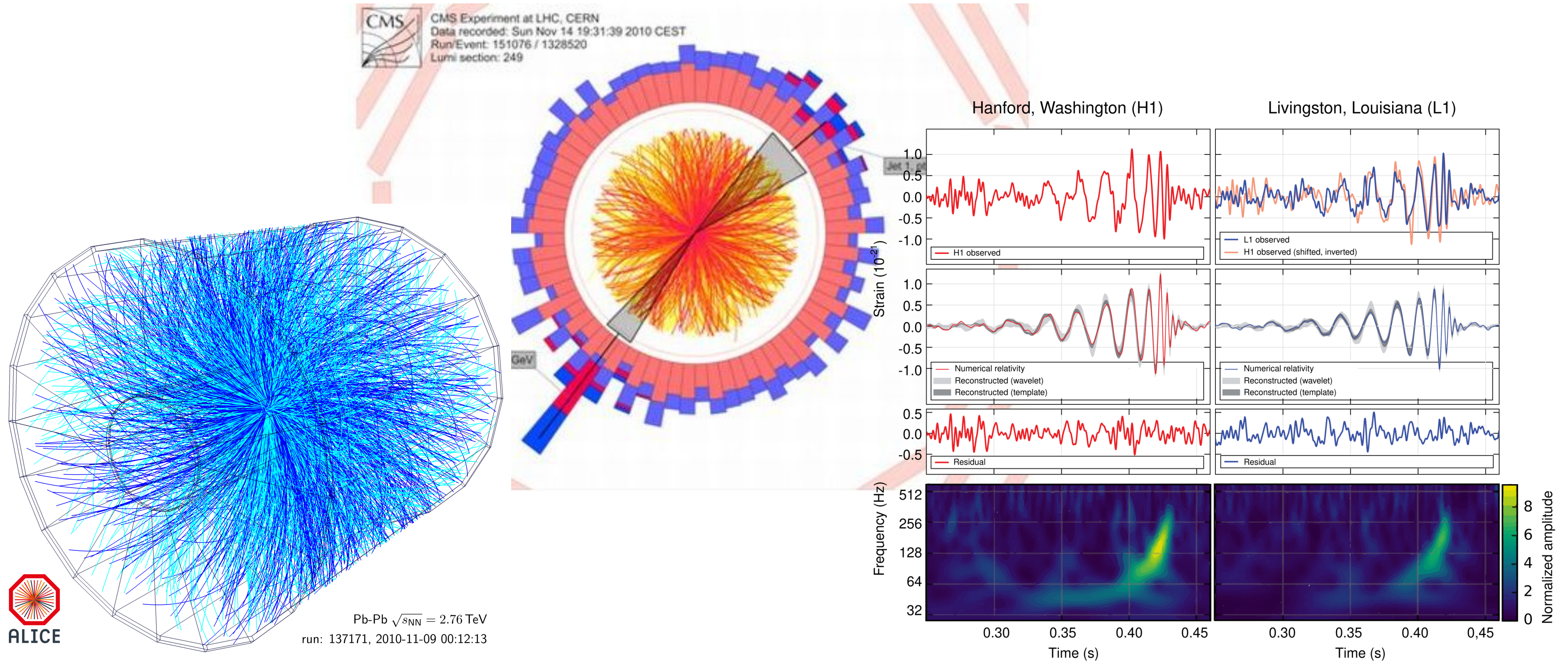
Two main research directions:

1. loose the compression and improve anomaly detection tools to work well in high dimensions

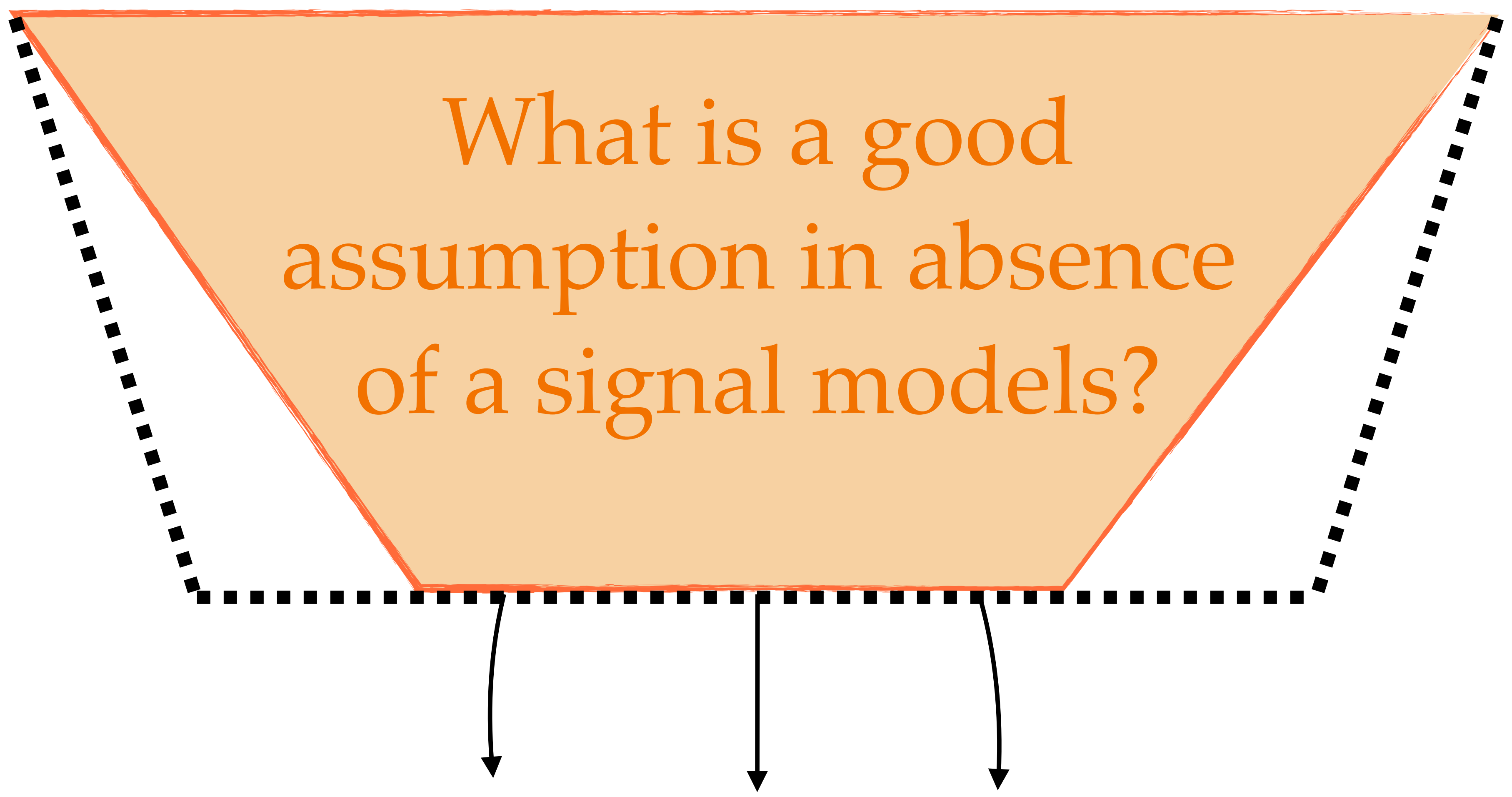
See for instance:

- Back to the roots [2309.13111]
- AD in presence of irrelevant features [JHEP02(2024)220]

Raw Readout



Data Compression



Stat. Inference

Anomaly Detection in high-D

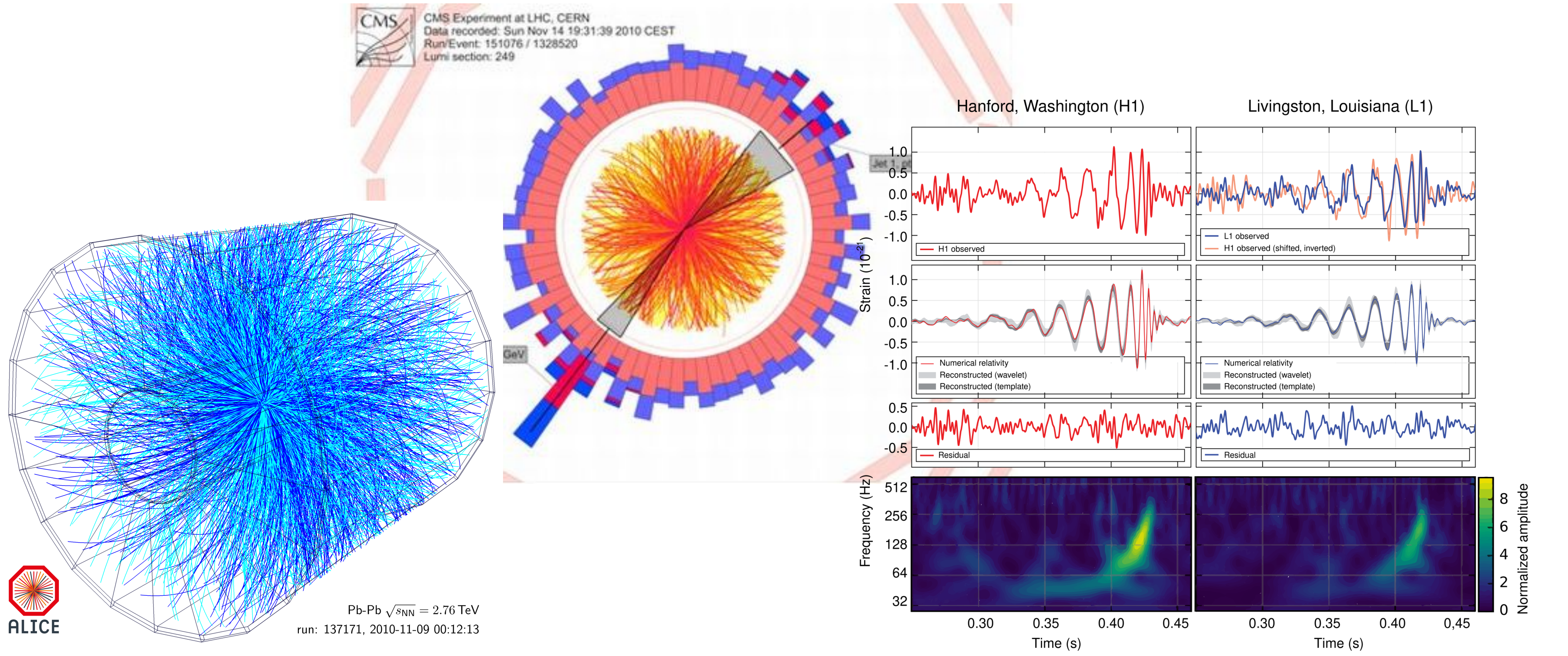
Inclusiveness vs assumptions: the never ending struggle

Two main research directions:

2. improve data compression approaches to reduce information loss

This talk

Raw Readout



Data Compression

What is a good assumption in absence of a signal models?

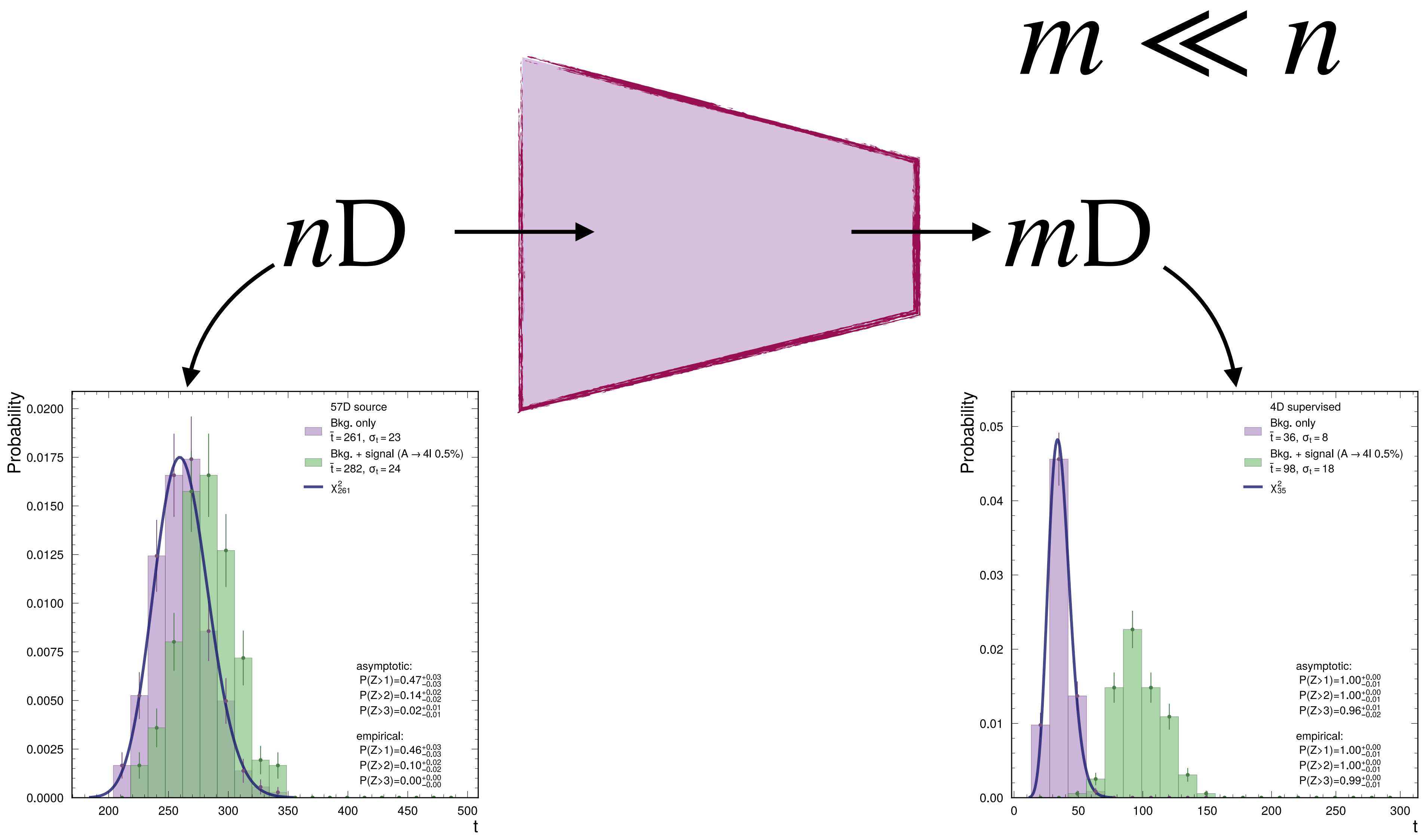
Stat. Inference

Anomaly Detection

Our approach to data compression for Anomaly Detection

Two steps:

1. Dimensionality reduction via neural embedding based on **supervised** contrastive learning
2. Statistical anomaly detection in the embedded space with NPLM (ML-based likelihood ratio test)

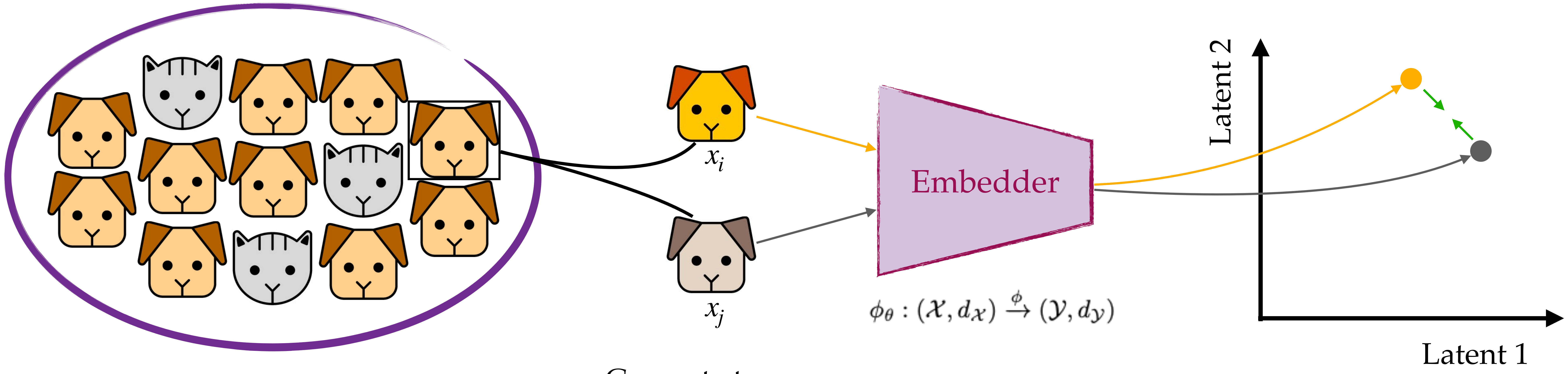


Based on:

"Anomaly preserving contrastive neural embeddings for end-to-end model-independent searches at the LHC" [2502.15926](https://arxiv.org/abs/2502.15926) (K. Metzger et al.)

"AutoSciDACT: Automated Scientific Discovery through Contrastive Embedding and Hypothesis Testing" (submitted to NeurIPS)

Contrastive learning [self-supervised]



unlabelled data points

Generate two distorted *views* of the same class

Constraint the organization of the views in the embedding space:

- Views of the same object are pushed together
- Views of different objects are pulled apart

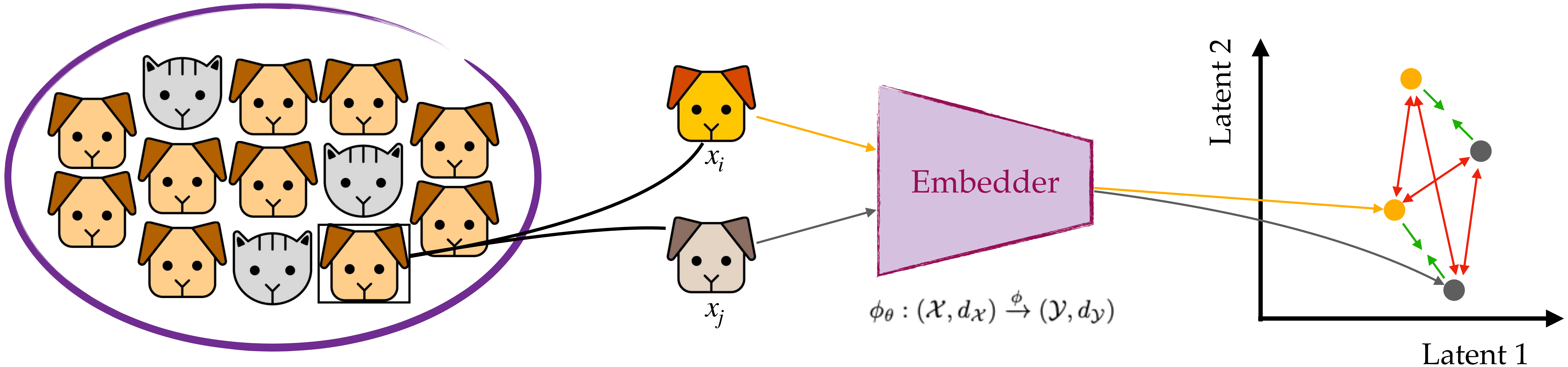
SimCLR loss [2002.05709]:

$$l_{\mathrm{SimCLR}}(x_i, x_j) = -\log \frac{e^{\mathrm{sim}(\phi_\theta(x_i), \phi_\theta(x_j))/\tau}}{\sum_{k=1}^{N} e^{\mathrm{sim}(\phi_\theta(x_i), \phi_\theta(x_k))/\tau}}$$

Cosine similarity:

$\mathrm{sim}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^\top \boldsymbol{v} / \|\boldsymbol{u}\| \|\boldsymbol{v}\|$

Contrastive learning [self-supervised]



unlabelled data points

Generate two distorted *views* of the same data point

Constraint the organization of the views in the embedding space:

- Views of the same object are pushed together
- Views of different objects are pulled apart

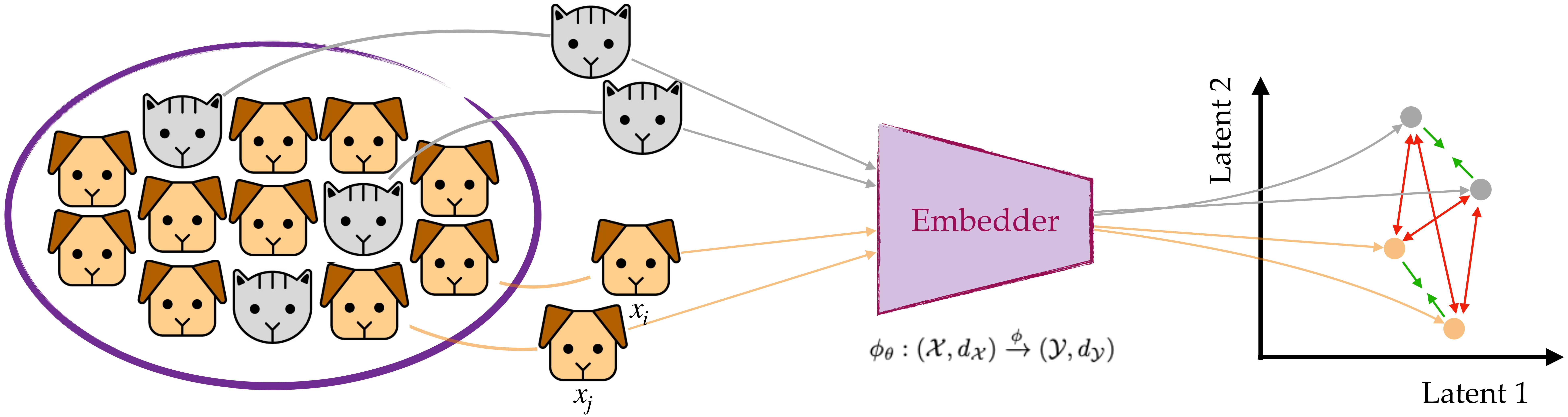
SimCLR loss [2002.05709]:

$$l_{\mathrm{SimCLR}}(x_i, x_j) = -\log \frac{e^{\mathrm{sim}(\phi_\theta(x_i), \phi_\theta(x_j))/\tau}}{\sum_{k=1}^{N} e^{\mathrm{sim}(\phi_\theta(x_i), \phi_\theta(x_k))/\tau}}$$

Cosine similarity:

$\mathrm{sim}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^\top \boldsymbol{v} / \|\boldsymbol{u}\| \|\boldsymbol{v}\|$

Contrastive learning [supervised]



labelled data points

Generate two *views* of the same class

Constraint the organization of the views in the embedding space:

- Views of the same object are pushed together
- Views of different objects are pulled apart

SimCLR loss [2002.05709]:

$$l_{\mathrm{SimCLR}}(x_i, x_j) = -\log \frac{e^{\mathrm{sim}(\phi_\theta(x_i), \phi_\theta(x_j))/\tau}}{\sum_{k=1}^{N} e^{\mathrm{sim}(\phi_\theta(x_i), \phi_\theta(x_k))/\tau}}$$

Cosine similarity:

$\mathrm{sim}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^\top \boldsymbol{v} / \|\boldsymbol{u}\| \|\boldsymbol{v}\|$

Two events generated according to the same hard core physics process (same label) are two realizations of the same physics *concept*.

Supervised contrastive learning. (Khosla, P. et al., NeurIPS 2020)

Numerical Experiments with ADC2021

Dataset:

MC simulations for CMS single-lepton L1 trigger:

- **4 background processes:** W production (59.2%), Z production (6.7%), $t\bar{t}$ production (0.3%), QCD multijet production (33,8%)
- **4 signal benchmarks:** LQ $\rightarrow b\tau$, $A \rightarrow 4l$, $h^0 \rightarrow \tau\tau$, $h^\pm \rightarrow \tau^\pm\nu$

Labels for supervised CL

Only seen at test time

- 19 objects: MET, first 4 most energetic electrons, first 4 most energetic muons, first 10 most energetic jets
- 3 observables: p_T, η, ϕ

19x3=57 D problem at the source

2502.15926 (K. Metzger et al.)

Numerical Experiments with ADC2021

Experiments:

- Supervised CL (transformer/MLP): **4D, …, 64D**
  - Pairing is based on physics labels

(Ours)

- **Self-supervised CL** (transformer/MLP): **4D, …, 64D**
  - Tried both augmentations inspired by AI vision (e.g. random masking), and physics-motivated ones (Dillon et al. 2301.04660)

Baselines

- Flatten input representation: **57D**
- Physics motivated selection: **6D**
  - MET, 2 highest pT electrons, 2 highest pT muons, 1 highest pT jets
- Latent space of a VAE: **6D** [Nat Mach Intell 4, 154–161 (2022)]

2502.15926 (K. Metzger et al.)

Numerical Experiments with ADC2021

Statistical test for anomaly: NPLM

ML-based Neyman-Pearson likelihood-ratio test:

$$\bar{t}(\mathcal{D}) = 2\max_{\mathbf{w}} \log\left[\frac{\mathcal{L}(\mathcal{D}\,|\,\mathrm{H}_{\mathbf{w}})}{\mathcal{L}(\mathcal{D}\,|\,\mathrm{R}_{\mathbf{0}})}\right] = 2\max_{\mathbf{w}}\left\{\log\left[\frac{e^{-N(\mathbf{w})}}{e^{-N(\mathrm{R})}}\prod_{i=1}^{\mathrm{N_D}}\frac{n(x_i|\mathbf{w})}{n(x_i|\mathrm{R})}\right]\right\}$$

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

$$f(x;\,\widehat{\mathbf{w}}) = \log\left[\frac{n(x\,|\,\mathrm{H}_{\widehat{\mathbf{w}}})}{n(x\,|\,\mathrm{R})}\right]$$ Kernel methods

Loss function

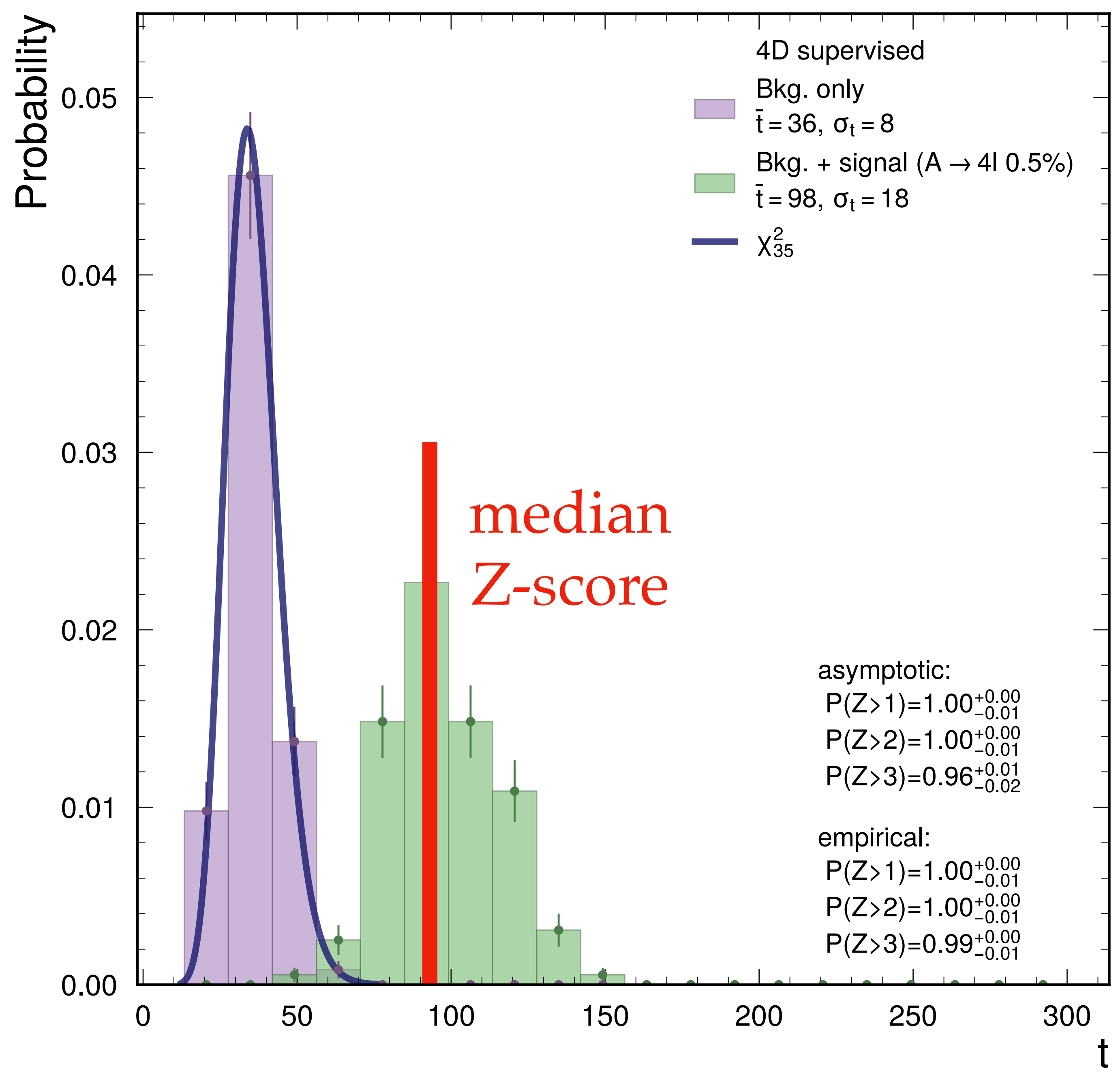
$$\bar{L}\left[f(x;\,\mathbf{w})\right] = -\sum_{x\in\mathcal{D}}\log\left[1+e^{-f_{\mathbf{w}}(x)}\right] + \sum_{x\in\mathcal{R}}\frac{\mathrm{N(R)}}{\mathrm{N}_{\mathcal{R}}}\log\left[1+e^{f_{\mathbf{w}}(x)}\right]$$ (+ L2 reg.)

Learning new physics efficiently with nonparametric methods. [Eur. Phys. J. C, 82(10)]

w: trainable parameters on the *f* model
D: data sample
R: reference sample (built according to the $\mathrm{R_0}$ hypothesis); could be weighted (*w*)

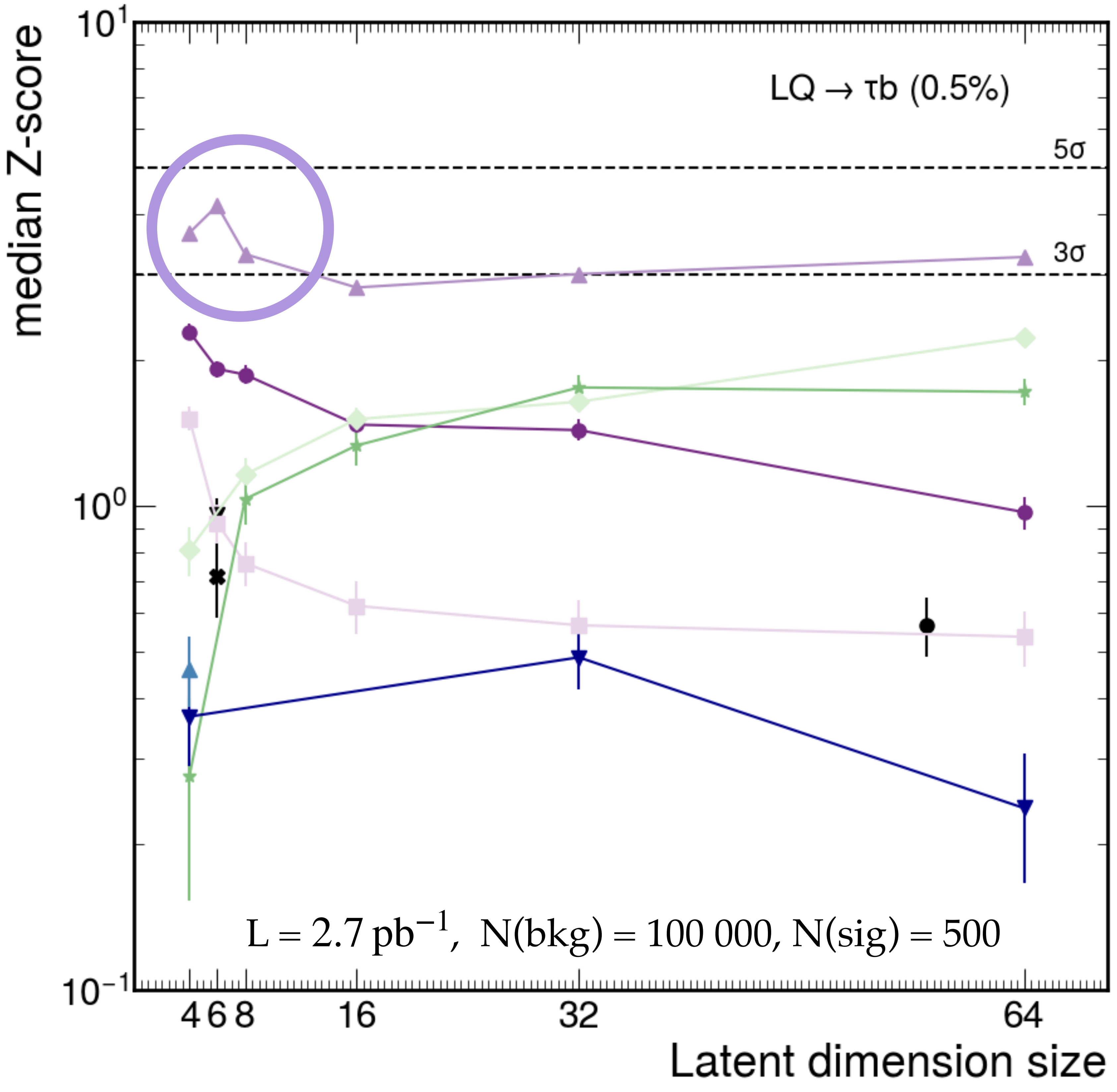
Assumptions:

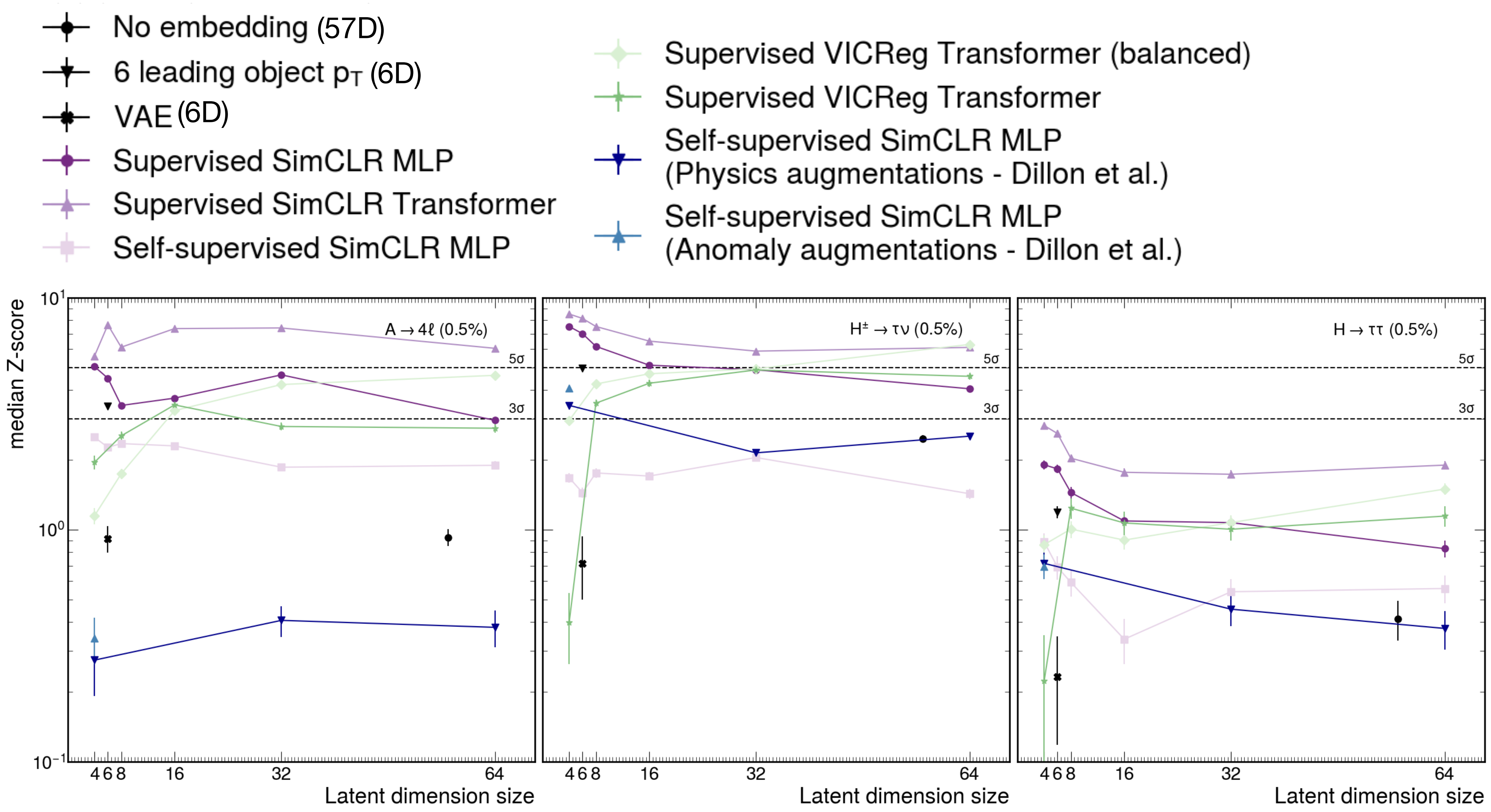
- $\mathrm{N_R} \gg \mathrm{N_D}$ the statistical fluctuations of the reference sample are negligible.
- the weights of the reference sample (*w*) are such that the reference sample is normalised to match the data sample luminosity



Numerical Experiments with ADC2021

Results:

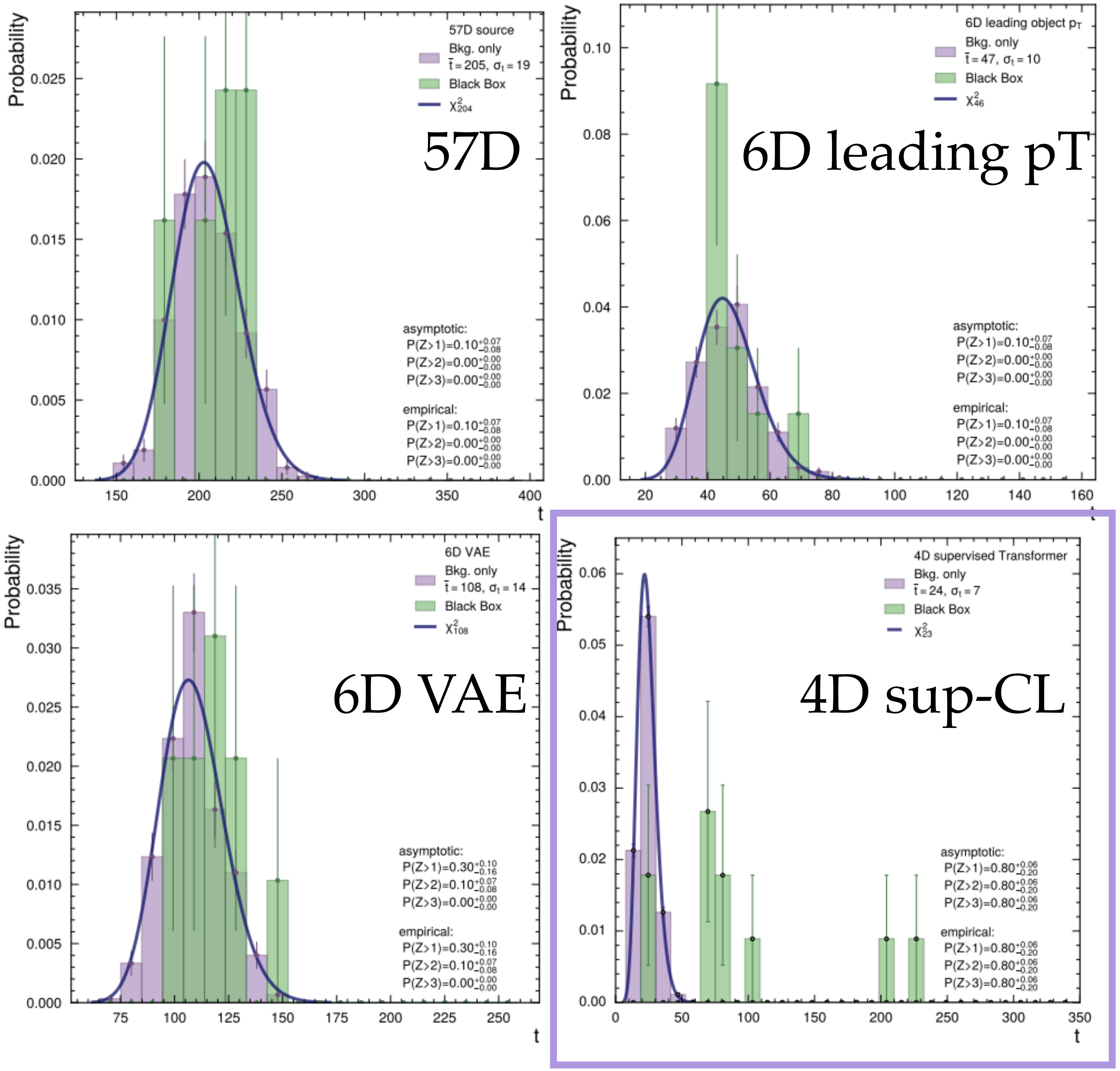




Supervised contrastive learning with SimCLR loss produces **low dimensional** latent spaces that enhance anomaly detection

2502.15926 (K. Metzger et al.)

Black box



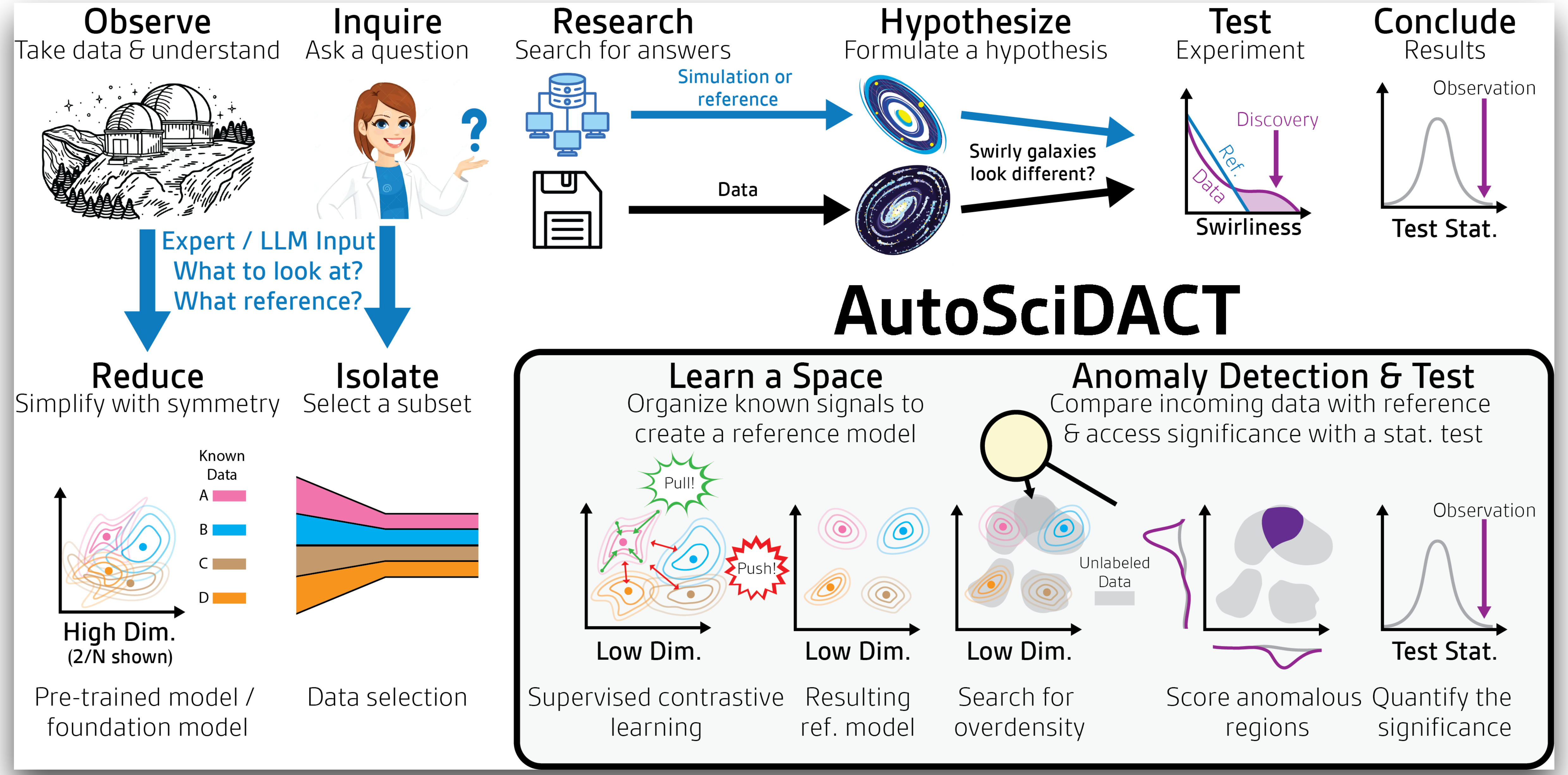
Supervised contrastive learning allows to discover the anomaly in 8/10 data batches.

N(sig) = 1000
N(bkg) = 1 000 000
Split in 10 batches

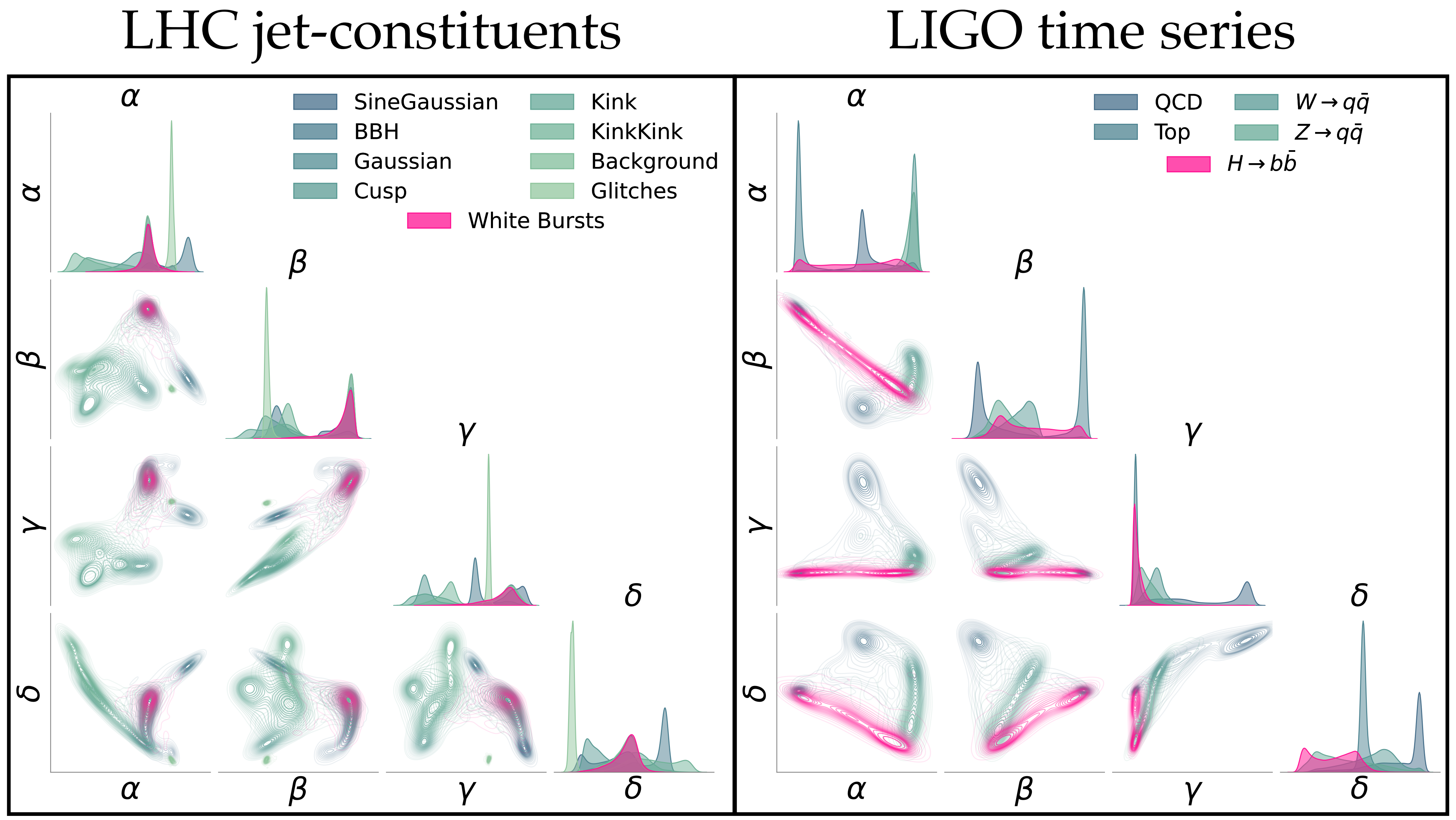
2502.15926 (K. Metzger et al.)

A *generalized* pipeline for all scientific domains

- Can be applied to any use case for which domain knowledge about meaningful augmentations can be incorporated in labels
- Highly automatized end-to-end pipeline



"AutoSciDACT" (Submitted to NeurIPS)



Examples of other HEP applications with **different data representations**

Sensitivity to %, ‰ signal injections

Plots show the contour plots of the learnt latent representations (anomalous signal in purple)

Future directions

- Integrating uncertainties in the pipeline
  - NPLM provides a framework to treat systematic uncertainties.
    (Eur. Phys. J. C (2021) d'Agnolo, Grosso, Pierini, Wulzer, Zanetti)
  - We have to provide: background models and their uncertainties in the latent space → Ongoing work (See EuCaif poster by Giada Badaracco)
  - Can we mitigate the impact of domain shifts due to systematic uncertainties in the latent space? → Ongoing work
- Integrate LLMs agents for an automatic pipeline?
- Understanding scaling laws: we need labeled data, but how many?

Backup slides

Test for anomaly: New physics learning machine (NPLM)

Likelihood ratio from Binary Cross Entropy

Test statistic

$$\bar{t}(\mathcal{D}) = 2\max_{\mathbf{w}}\log\left[\frac{\mathcal{L}(\mathcal{D}\,|\,\mathrm{H}_{\mathbf{w}})}{\mathcal{L}(\mathcal{D}\,|\,\mathrm{R}_{\mathbf{0}})}\right] = 2\max_{\mathbf{w}}\left\{\log\left[\frac{e^{-N(\mathbf{w})}}{e^{-N(\mathrm{R})}}\prod_{i=1}^{\mathrm{N}_{\mathrm{D}}}\frac{n(x_i|\mathbf{w})}{n(x_i|\mathrm{R})}\right]\right\}$$

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

$$f(x;\,\widehat{\mathbf{w}}) = \log\left[\frac{n(x\,|\,\mathrm{H}_{\widehat{\mathbf{w}}})}{n(x\,|\,\mathrm{R})}\right]$$ Kernel methods

w: trainable parameters on the NN model

D: data sample

R: reference sample (built according to the R_0 hypothesis); could be weighted (*w*)

Assumptions:

- $\mathrm{N_R} \gg \mathrm{N_D}$ the statistical fluctuations of the reference sample are negligible.
- the weights of the reference sample (*w*) are such that the reference sample is normalised to match the data sample luminosity

Loss function

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

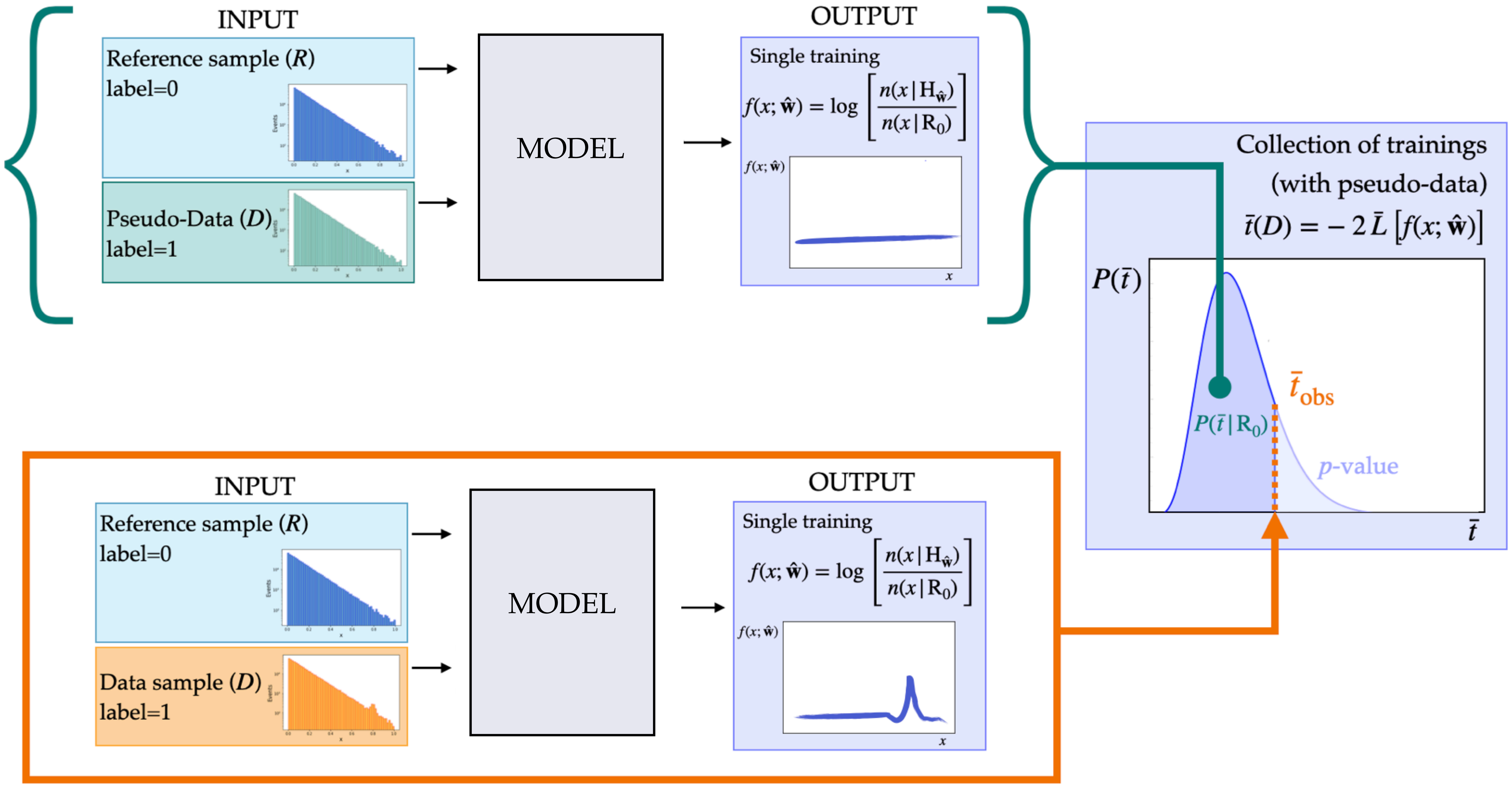
+ L2 regularization

Learning new physics efficiently with nonparametric methods. [Eur. Phys. J. C, 82(10)]

Figure of merit [1]: NPLM test statistic

Frequentist p-value (aka calibration):

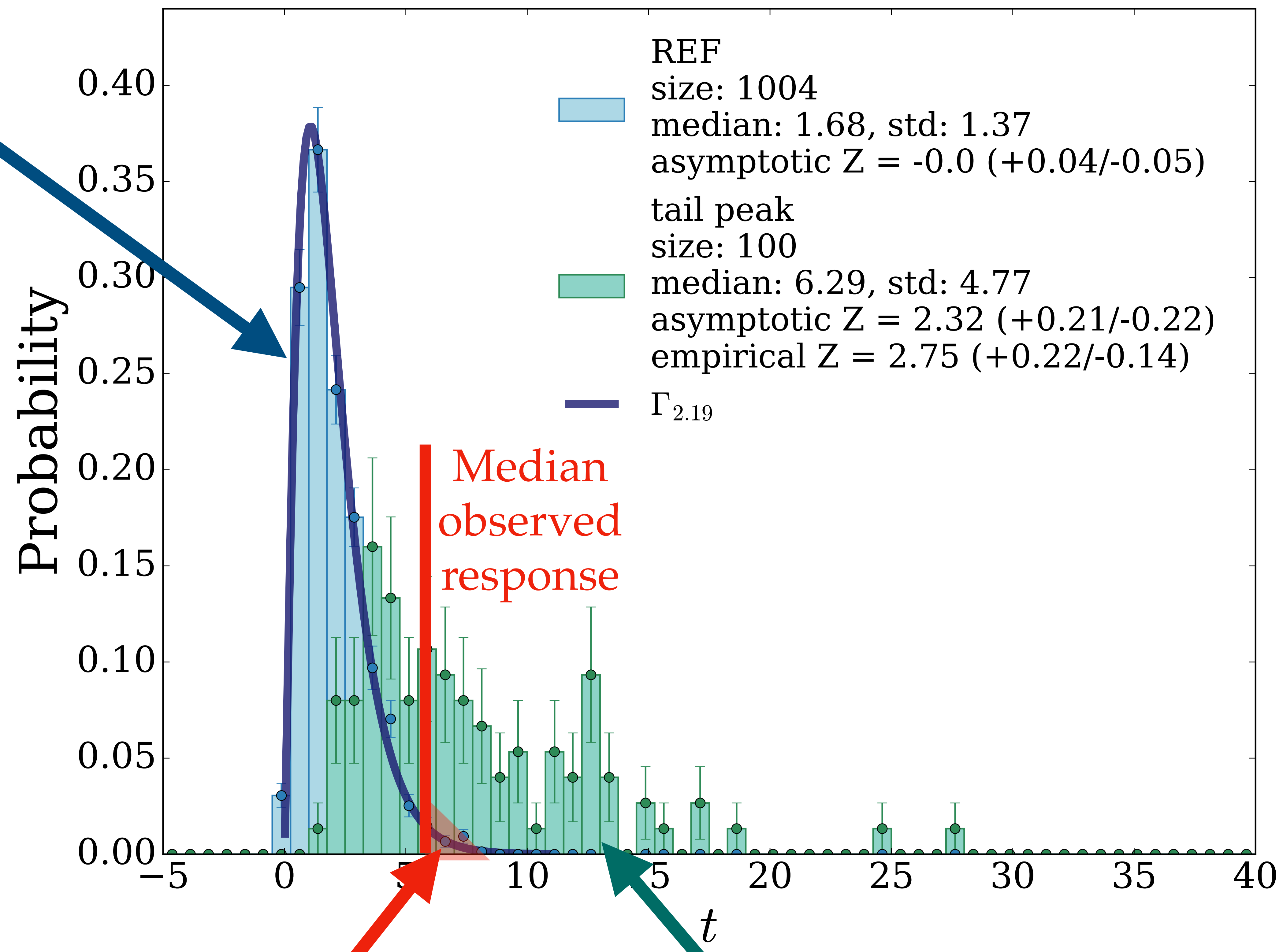
1. Run NPLM test on toy experiments to simulate the response under the null hypothesis
2. Run NPLM test on the data of interest and check where the test outcome falls to compute an exclusion p-value



Numerical Experiments with ADC2021: Figure of merit

NPLM test statistic

Test statistic distribution in absence of signal (baseline response of the test)



Z-score: $Z_p = \Phi^{-1}(1-p)$
(standardized parametrization of the p-value)

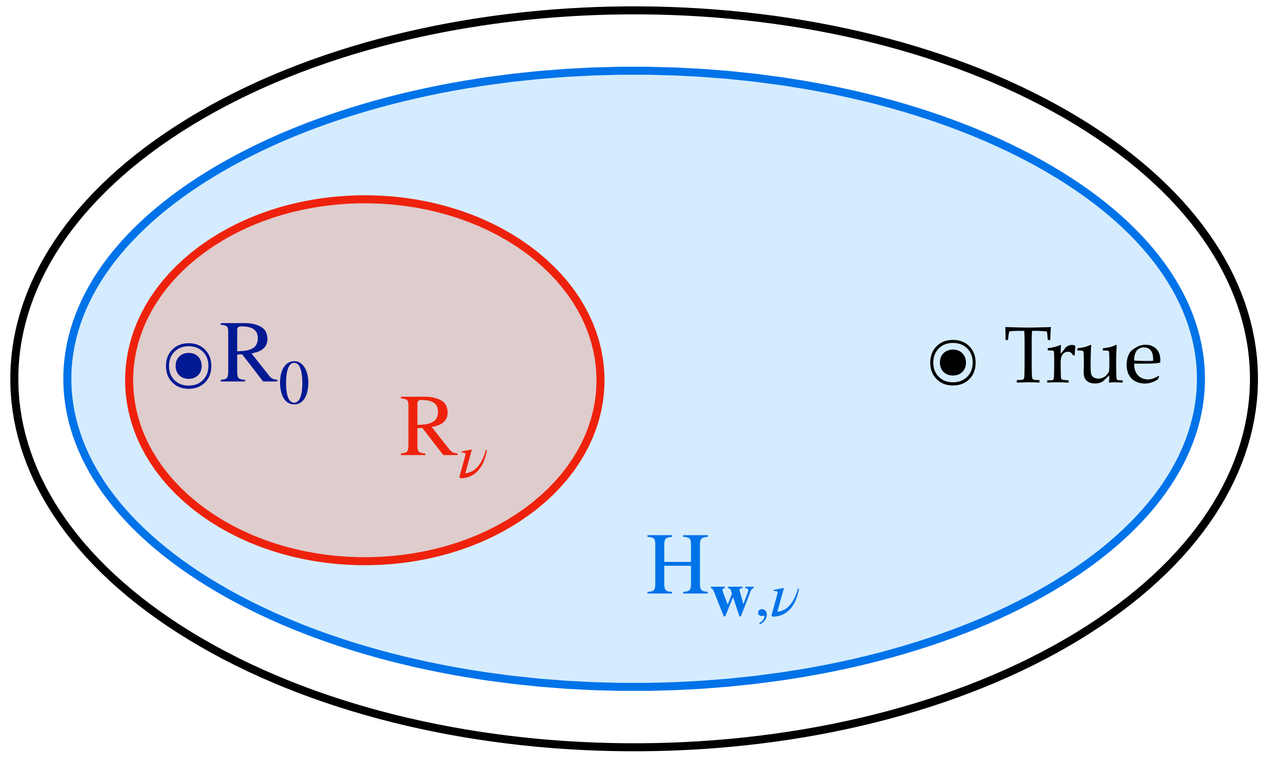
Median observed p-value (false positives rate)

Test statistic distribution in presence of signal

Dealing with *systematic uncertainties*

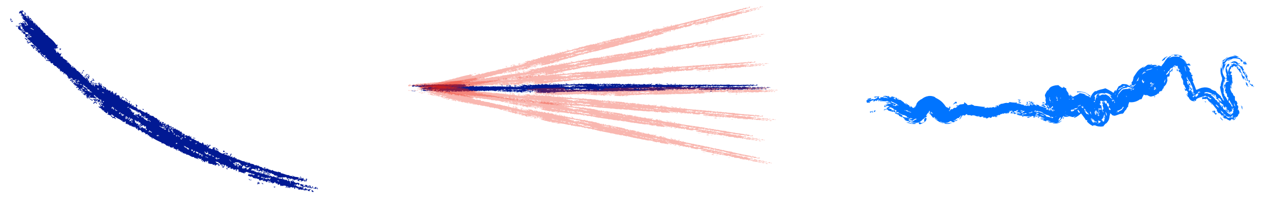
Frequentist approach (discovery at the LHC)

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



Inform the model about *known* data transformations

Uncertainties-aware parametrization of the alternative hypothesis:



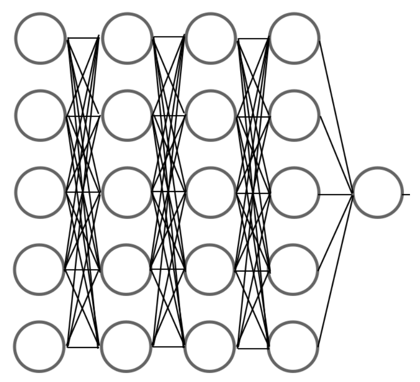
$$n(x\,|\,\mathrm{H}_{\mathbf{w},\,\nu}) = n(x\,|\,\mathrm{R}_0)\,\frac{n(x\,|\,\mathrm{R}_\nu)}{n(x\,|\,\mathrm{R}_0)}\,e^{\boxed{f(x;\,\mathbf{w})}}$$

central value
(ν=0)

Model of systematic uncertainties effects

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

NN model



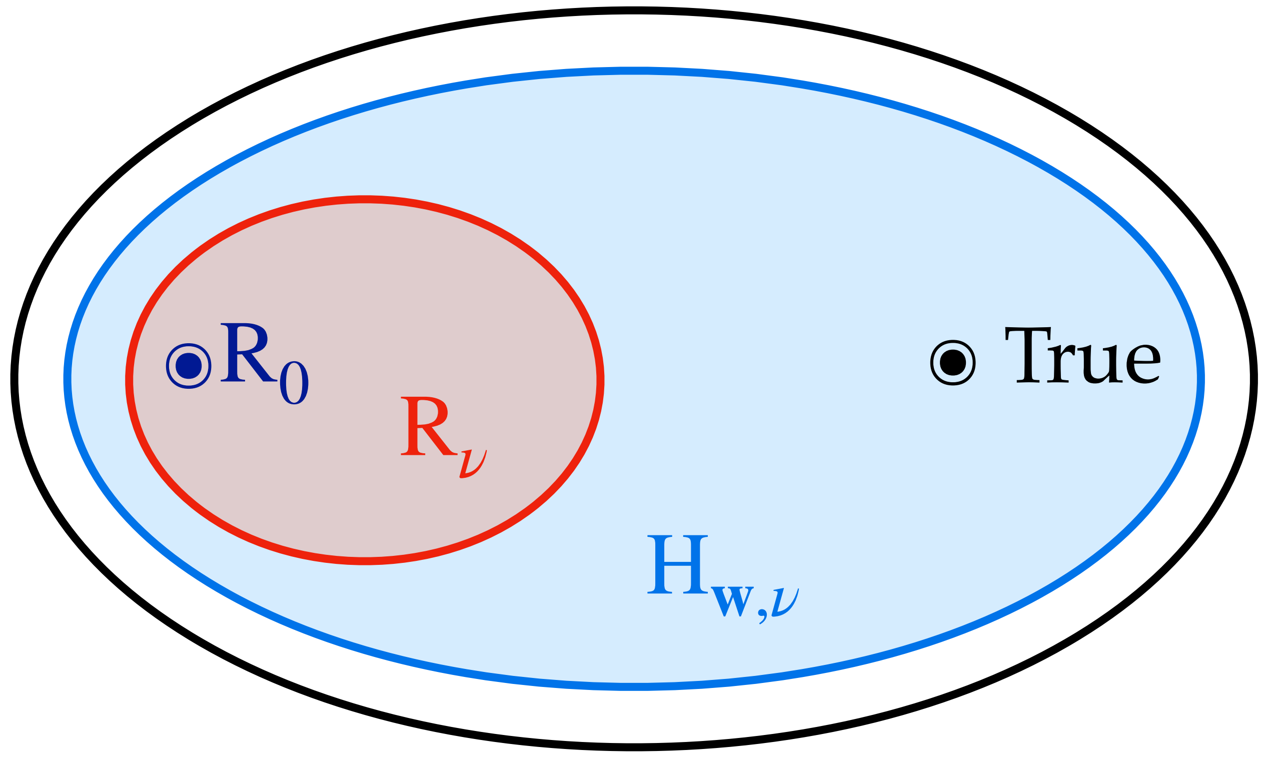
"Learning New Physics from an Imperfect Machine"
Eur. Phys. J. C (2021) (Grosso, d'Agnolo, Wulzer, Zanetti, Pierini)

Dealing with *systematic uncertainties*

Frequentist approach (discovery at the LHC)

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

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



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

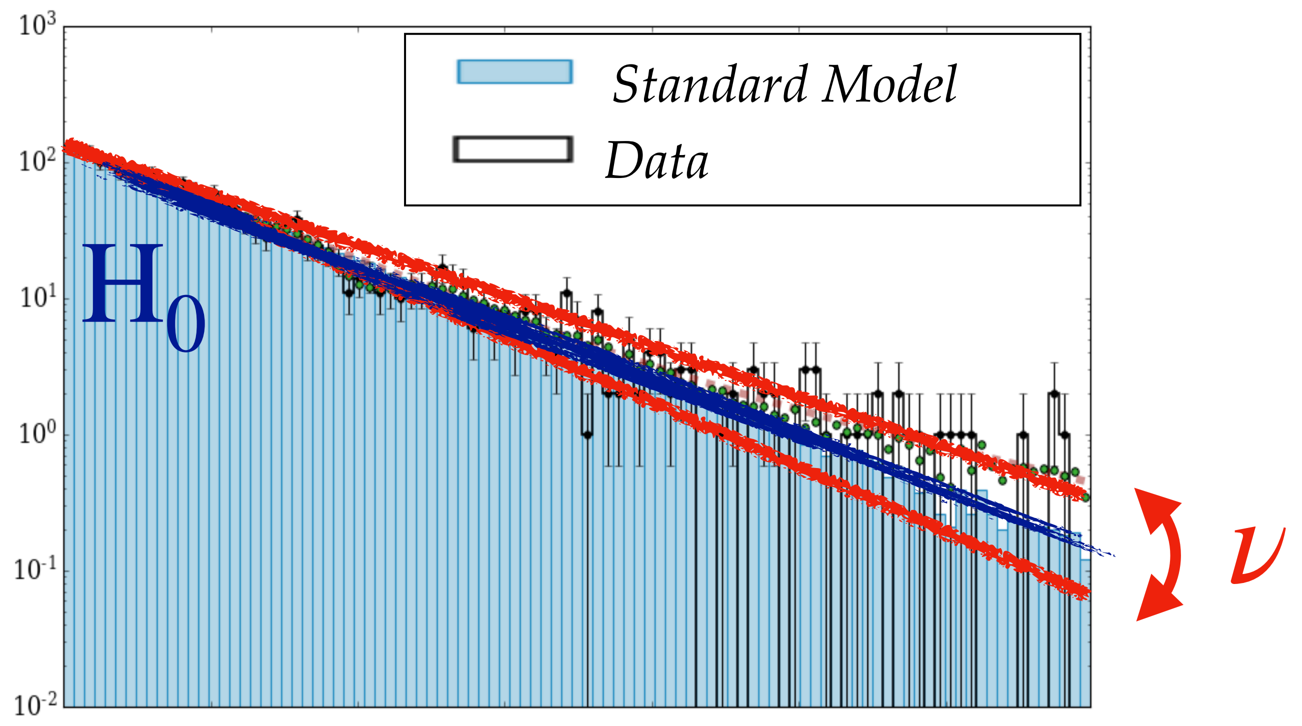
"Learning New Physics from an Imperfect Machine"
Eur. Phys. J. C (2021) (Grosso, d'Agnolo, Wulzer, Zanetti, Pierini)

Example: Signal-agnostic tool for New Physics searches

Controlling false positives (aka test calibration)

Validation of the algorithm

1. Data deformations within systematic uncertainties



2. Data deformations within systematic uncertainties + signal events