Exploiting anomaly detection for new physics identification at the LHC

Maurizio Pierini
CERN

mPP

hls4ml

CMS

CERN

erc European Research Council

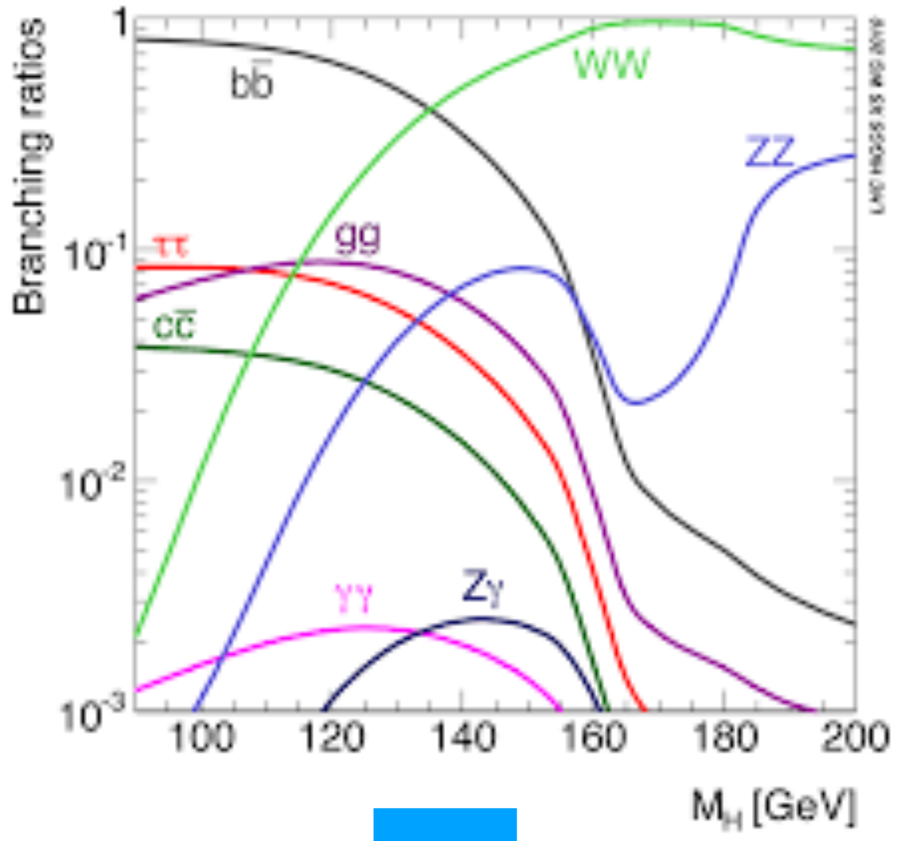
This talk in a nutshell

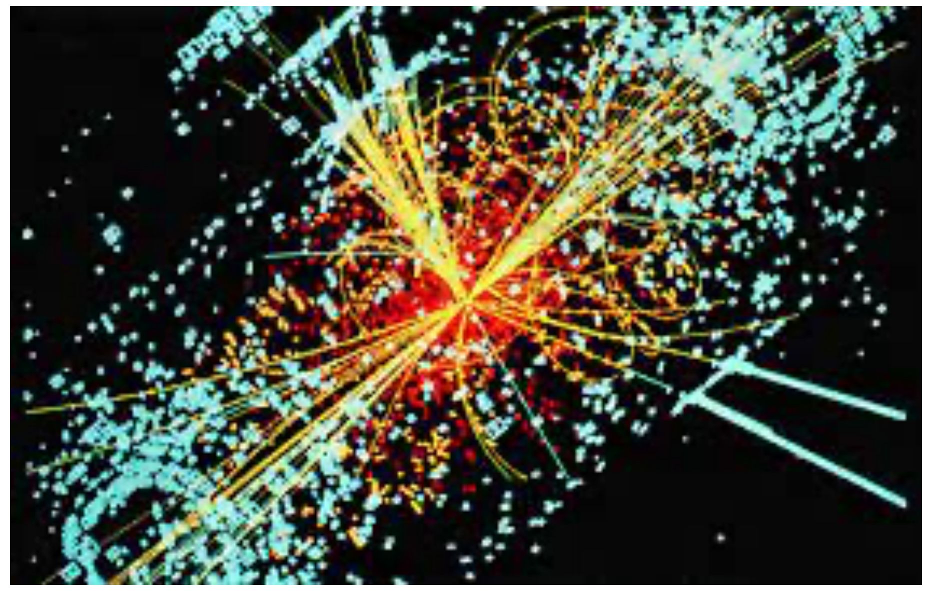
- *The LHC is a great discovery machine when you know what to search for*
- *Otherwise, you have to confront the limitations of the LHC big-data problem*
- *Since the SM was established, we followed an established discovery path. We had an easier life, but we have lost the capability of being surprised by data*
- *What we do is great, but we should (re)learn to look at data in a different way: observational particle physics, like astrophysics do*
- *Deep learning will be a crucial ingredient to this. And Run 3 is the right time.*

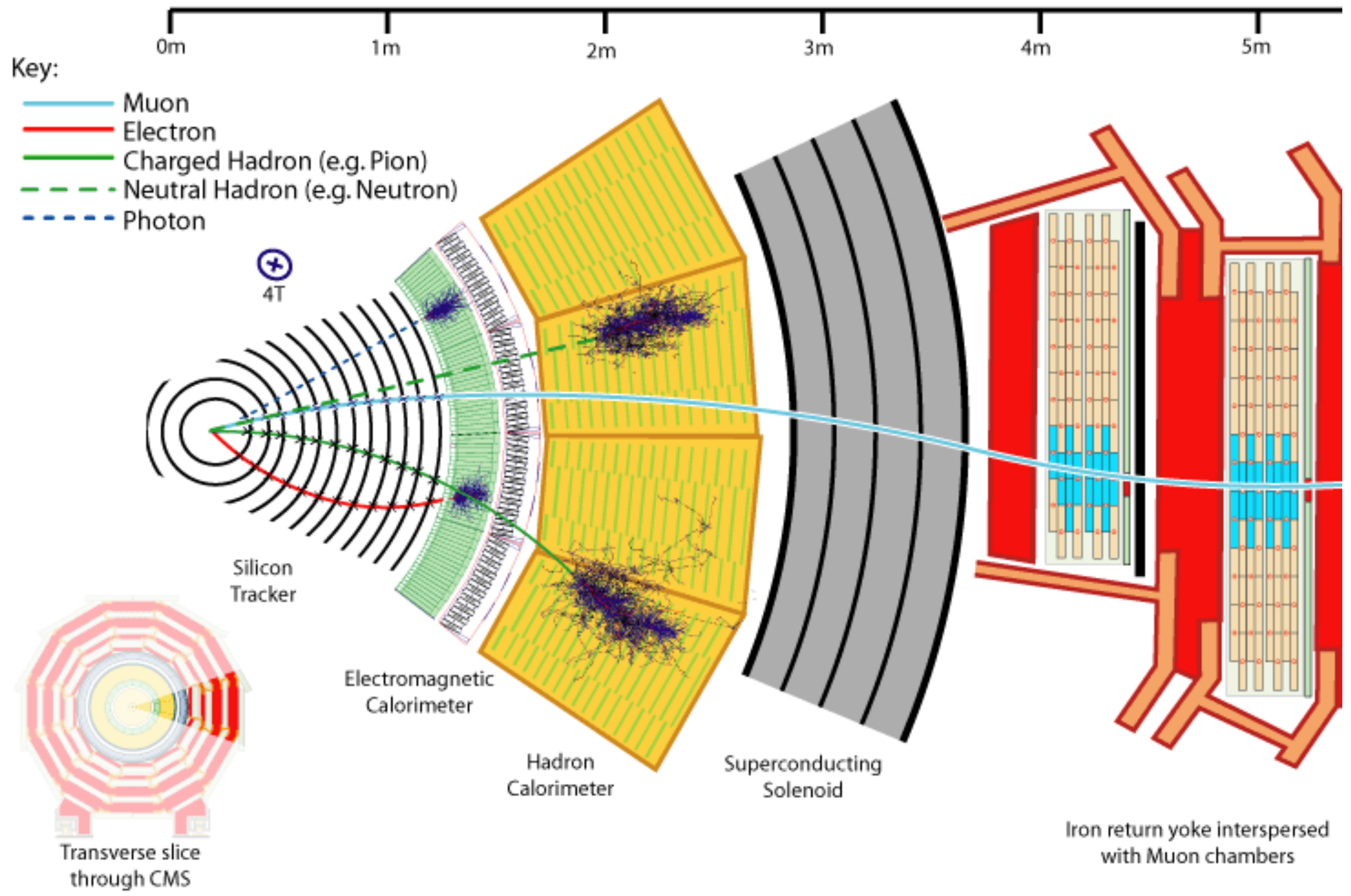
The ultimate discovery machine?

LHC as a discovery machine

- The LHC was mainly built to discover the Higgs boson
  - ATLAS & CMS were designed to cover the meaningful mass range for a particle that was fully characterized

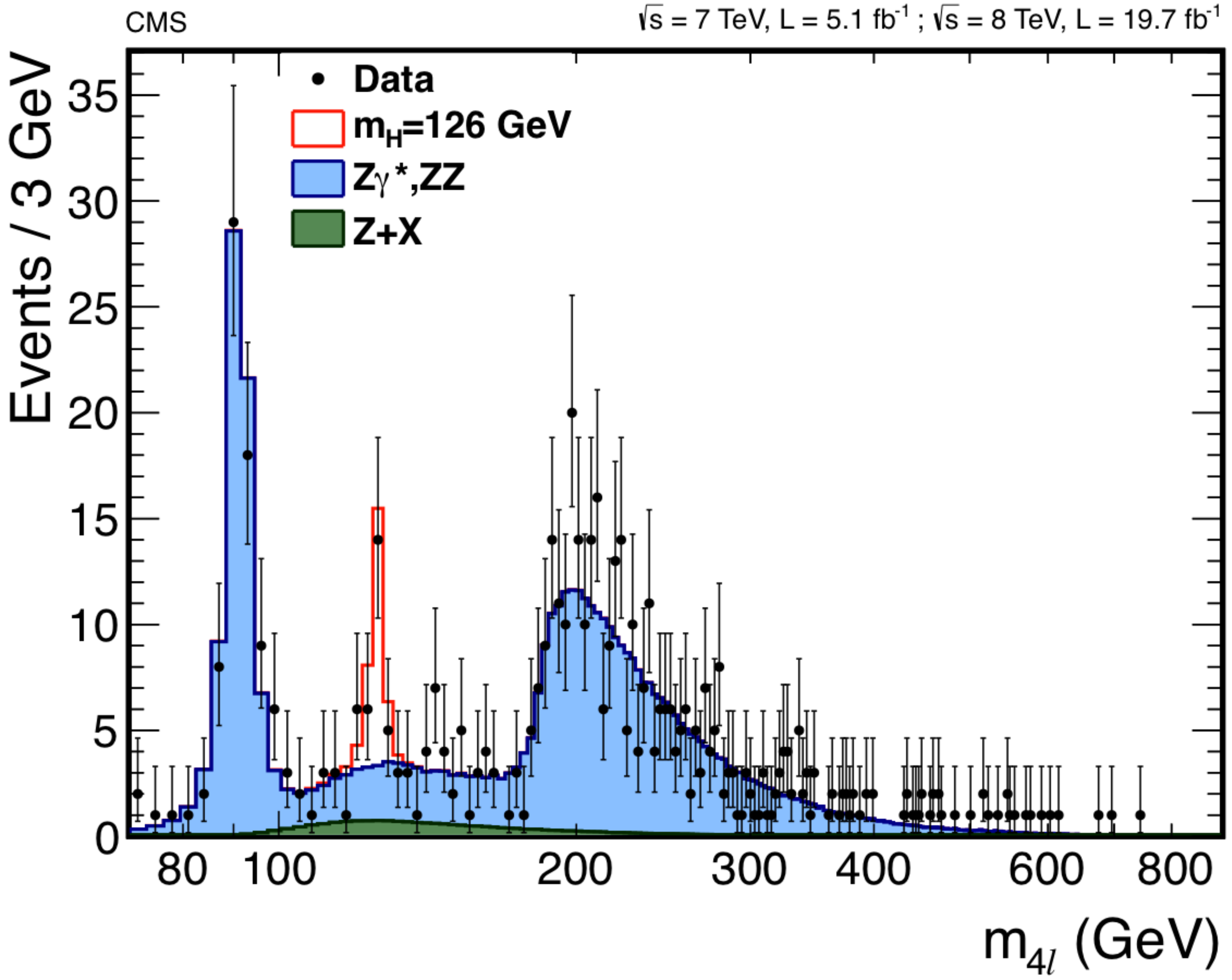


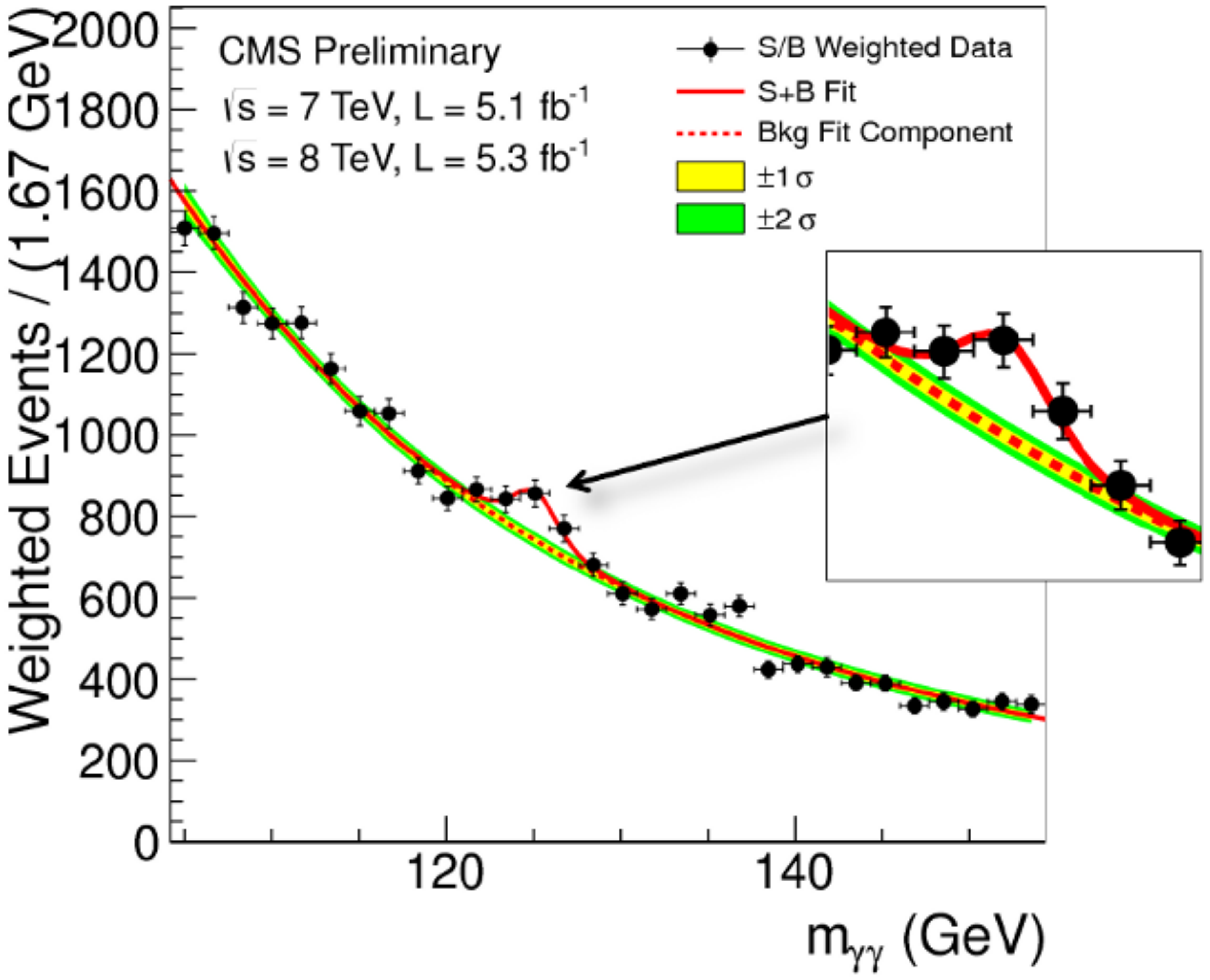




identify and measure muons, photons and electrons with high precision. The energy resolution for the above particles will be better than 1% at 100 GeV. At the core of the CMS detector sits a large superconducting solenoid generating a uniform magnetic field of 4 T. The choice of a strong magnetic field leads to a compact design for the muon spectrometer without compromising the momentum resolution up to rapidities of 2.5. The inner tracking system will measure all high p_t charged tracks with a momentum precision of $\Delta p/p \approx 0.1\, p_t$ (p_t in TeV) in the range $|\eta| < 2.5$. A high resolution crystal electromagnetic calorimeter, designed to detect the two photon decay of an intermediate mass Higgs, is located inside the coil. Hermetic hadronic calorimeters surround the intersection region up to $|\eta| = 4.7$ allowing tagging of forward jets and measurement of missing transverse energy.

And clearly it worked





Searches for something...

- At the LHC, you need a signal hypothesis
  - To design a trigger
  - To optimize your cuts
  - To compute the test statistics
  - To interpret the results
- so far so good...

CMS AN AN-11-065

CMS Draft Analysis Note

The content of this note is intended for CMS internal use and distribution only

2011/11/08
Head Id: 83705
Archive Id: 83789
Archive Date: 2011/11/07
Archive Tag: trunk

Trigger strategies for Higgs searches

The Higgs PAG

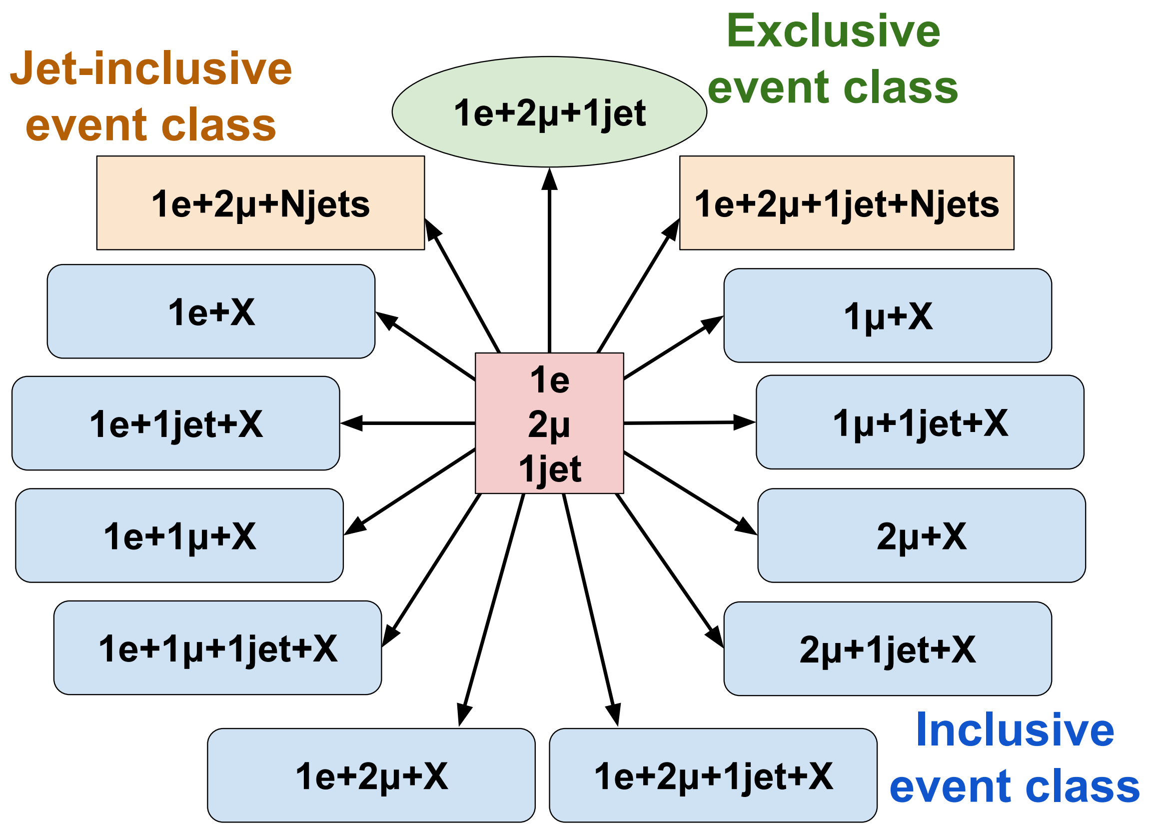
Abstract

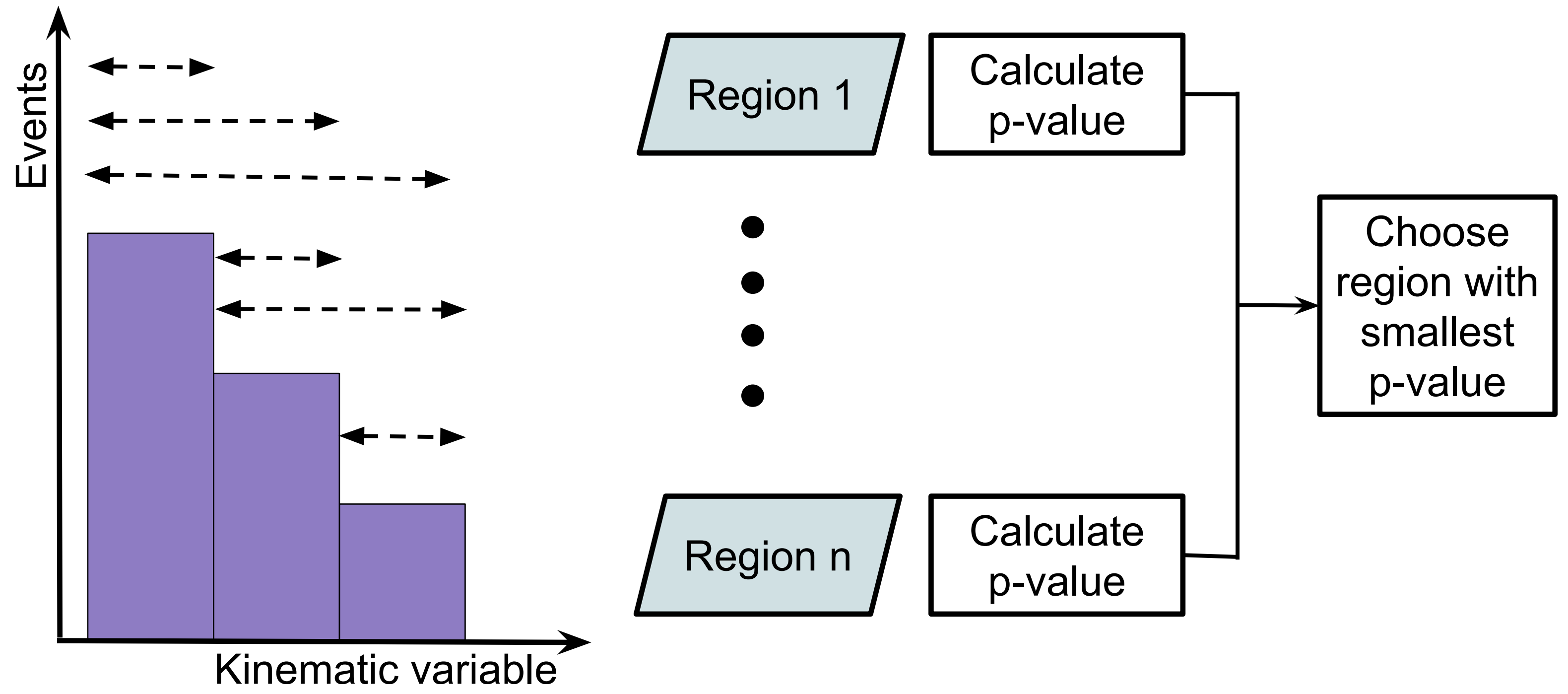
This document describes the triggers used in the Higgs analyses.

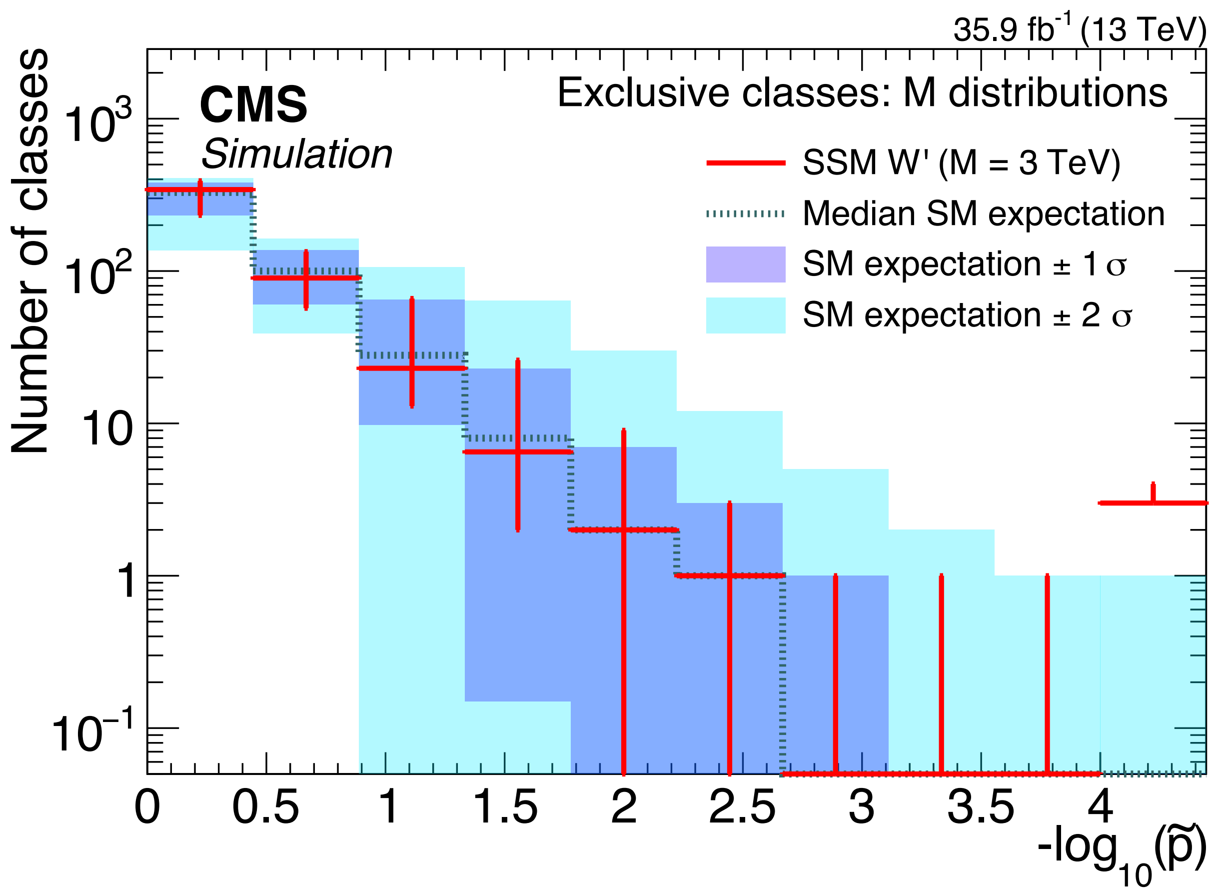
European Research Council

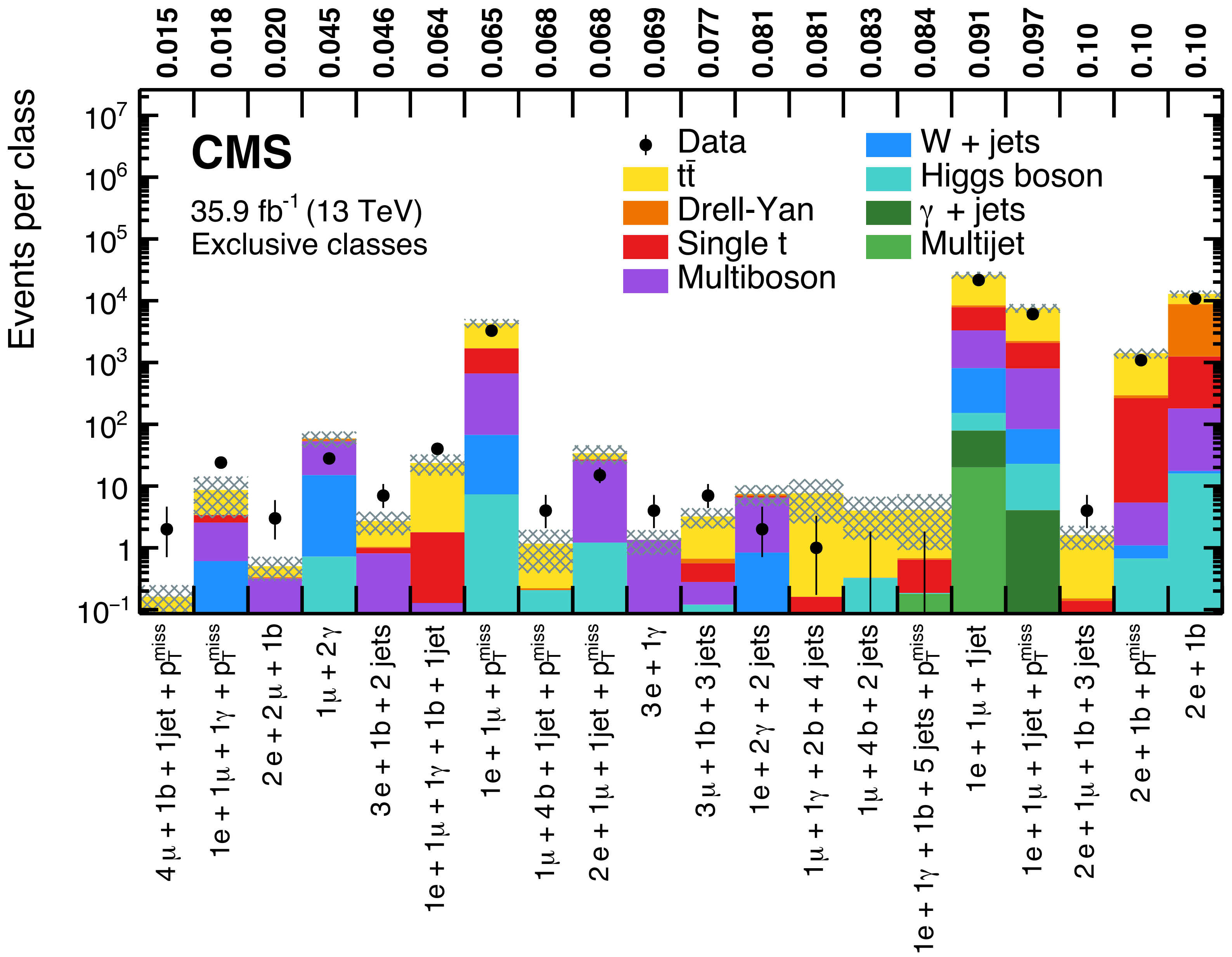
Searches for anything…

- What do you do when you don't know what to search for?
  - Any cut could be a signal killer
  - You need to look at as many signatures as possible
  - You can only look for some deviation from an expected distribution
- ***How do you know that the "right events" are there to start with?***





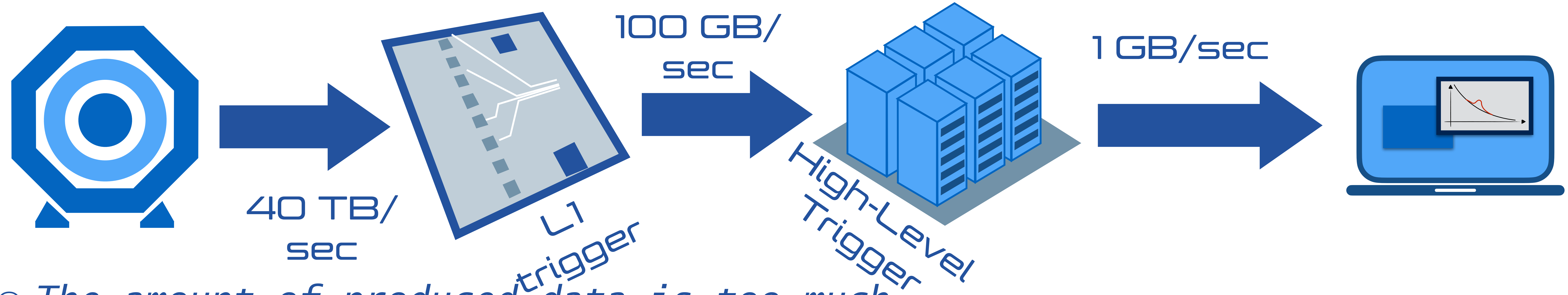




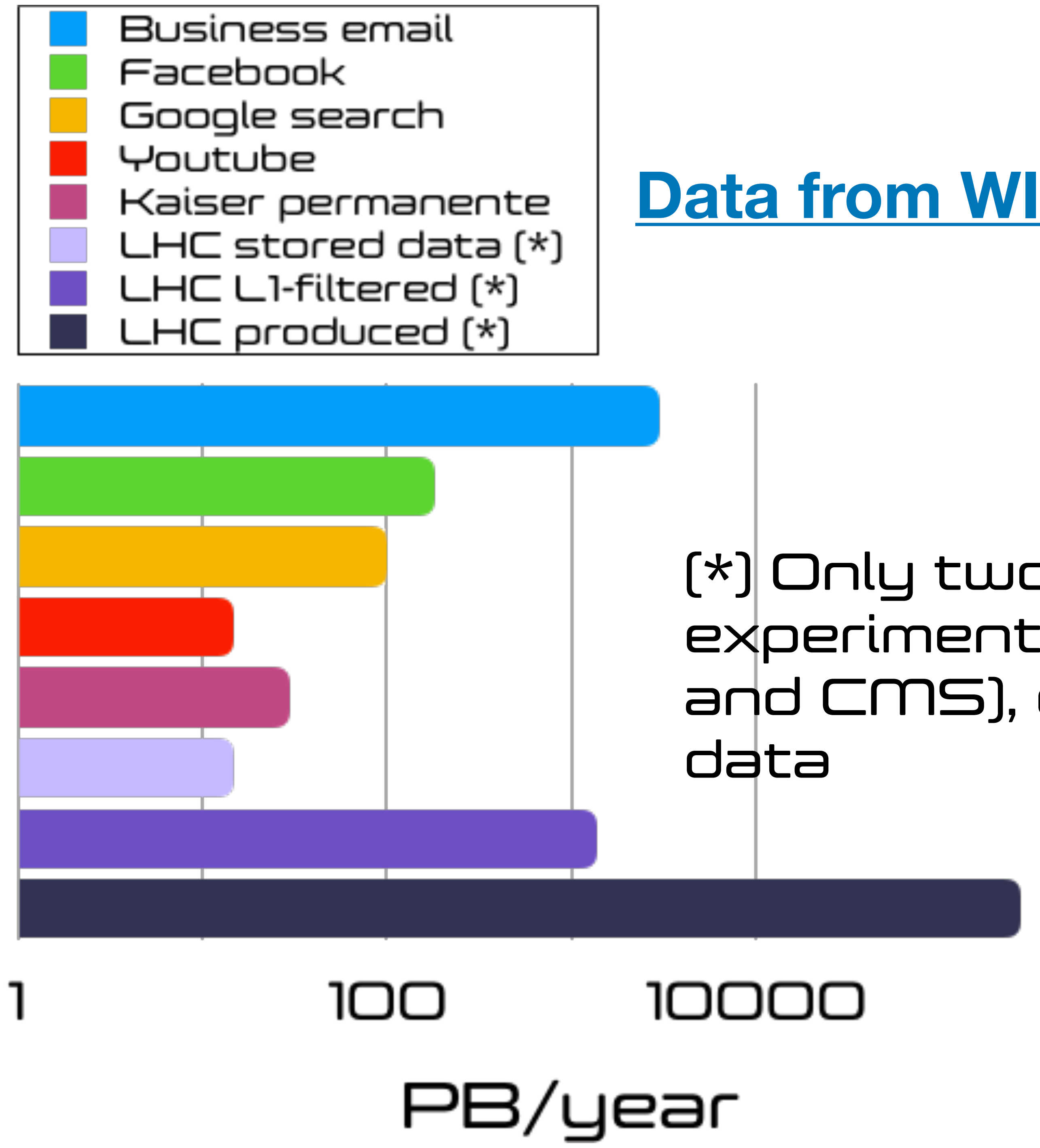
https://arxiv.org/pdf/2010.02984.pdf

European Research Council

Big Data @LHC



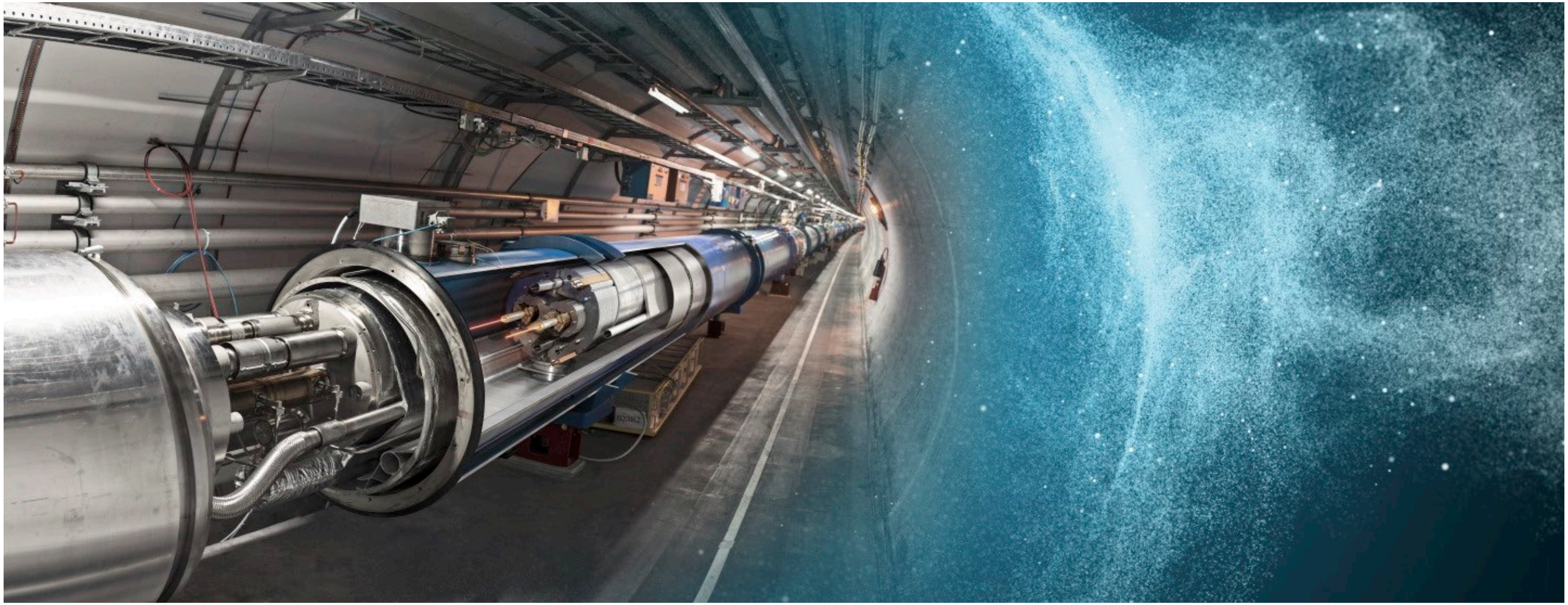
- *The amount of produced data is too much to be stored*
  - *1,000 times the data generated by google searches+youtube+facebook back in 2013*
  - *Reduced to 5x(google searches+youtube+facebook) after first filtering*
  - *Can only store 5% of those*



[Data from WIRED 2013]

(*) Only two big experiments (ATLAS and CMS), only RAW data

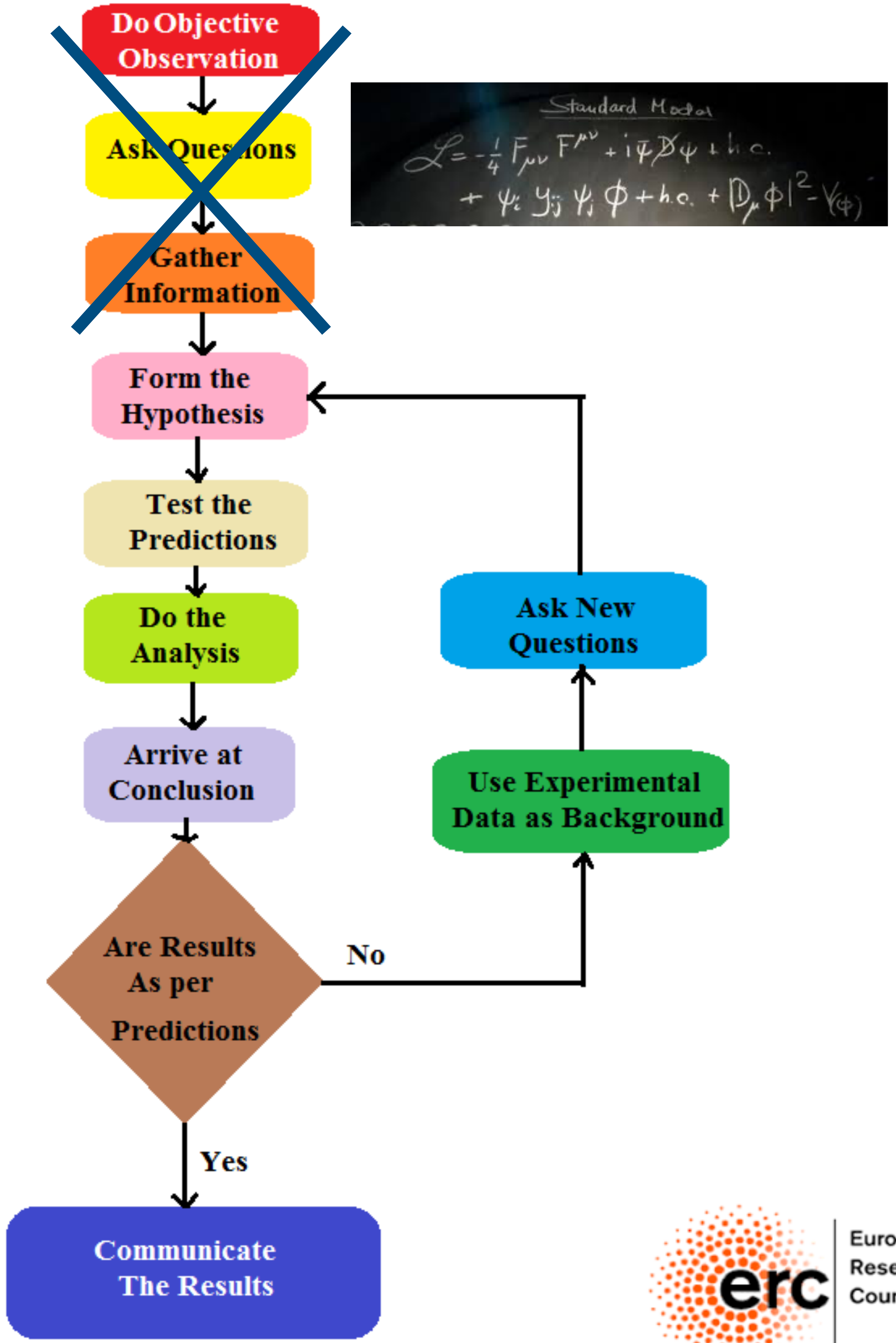
European Research Council



Unsupervised searches & Observational Particle Physics

HEP searches in LHC era

- Research under the scientific method starts gathering information about nature
- Instead, our baseline is the SM, which was formed once these informations were gathered
- We are victim of our success:
  - Since 1970s, we start always from the same point
  - We have lost the value of learning from data
  - Not by chance, we totally endorsed blind analysis as the ONLY way to search



Learning from Data

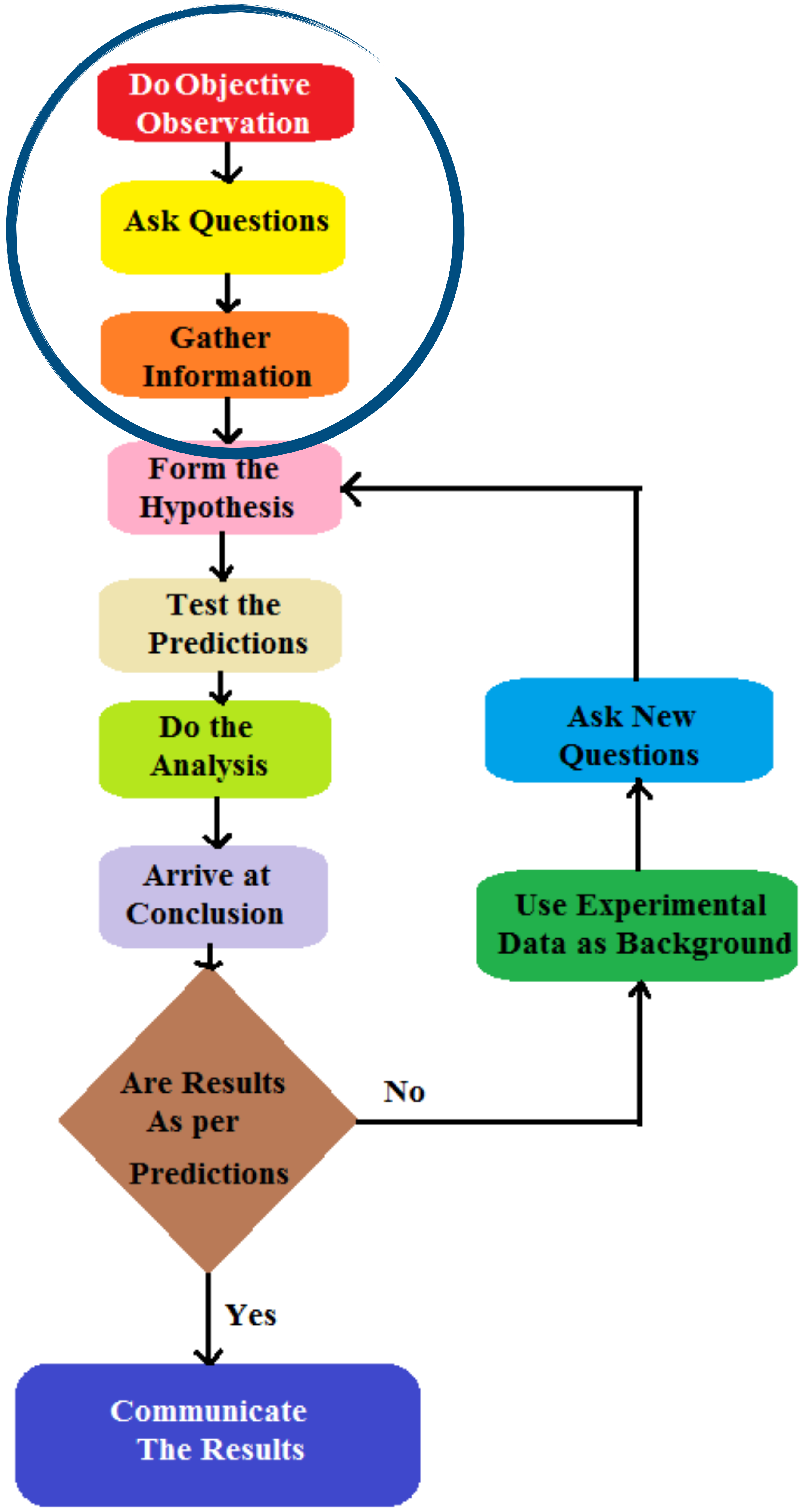
- *Rather than specifying a signal hypothesis upfront, we could start looking at our data*

- *Based on what we see (e.g., clustering alike objects) we could formulate a signal hypothesis*

- *EXAMPLE: star classification was based on observed characteristics*

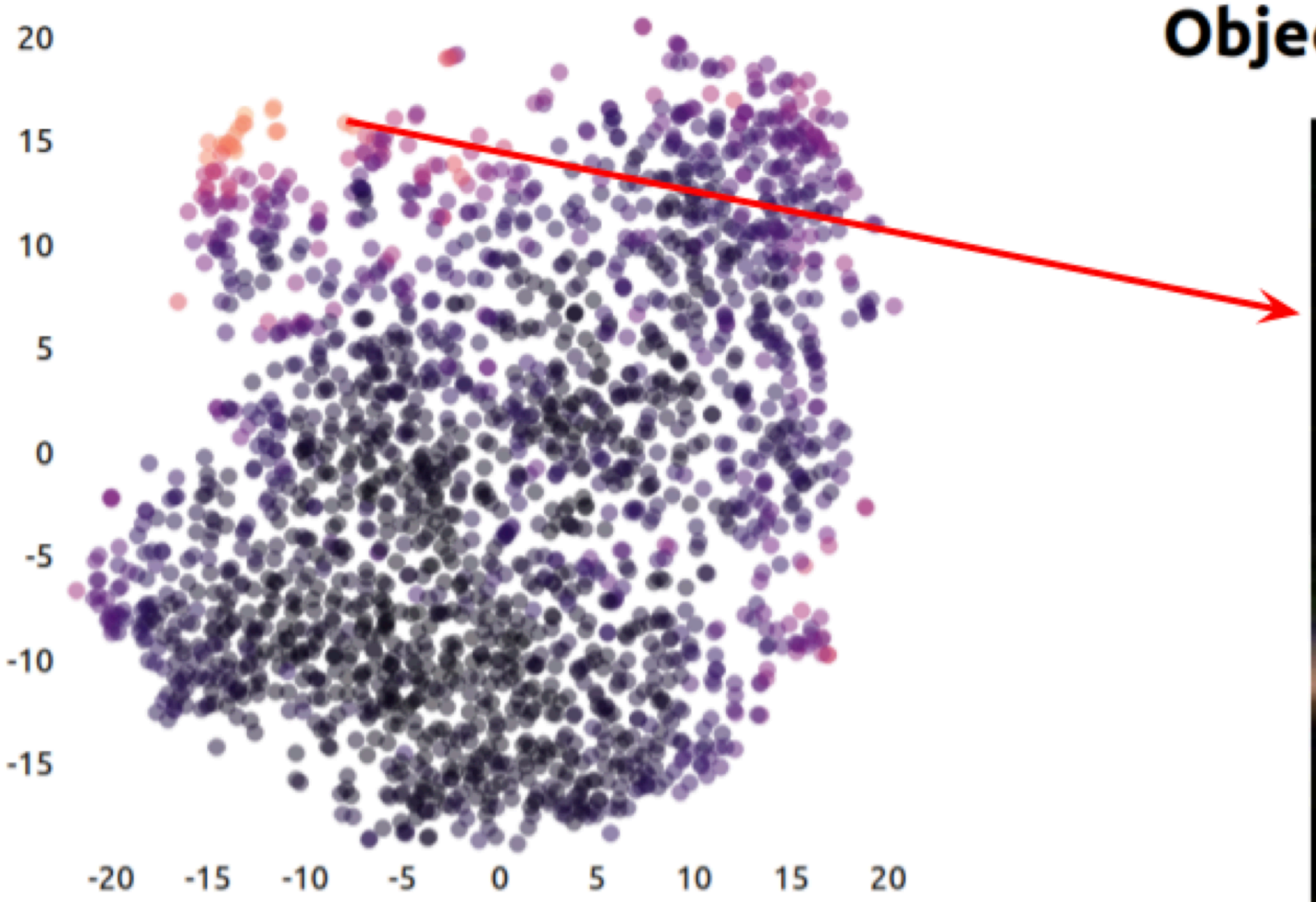
Class	Effective temperature[1][2]	Vega-relative chromaticity[3][4][a]	Chromaticity (D65)[5][6][3][b]	Main-sequence mass[1][7] (solar masses)	Main-sequence radius[1][7] (solar radii)	Main-sequence luminosity[1][7] (bolometric)	Hydrogen lines	Fraction of all main-sequence stars[8]
O	≥ 30,000 K	blue	blue	≥ 16 $M_\odot$	≥ 6.6 $R_\odot$	≥ 30,000 $L_\odot$	Weak	~0.00003%
B	10,000–30,000 K	blue white	deep blue white	2.1–16 $M_\odot$	1.8–6.6 $R_\odot$	25–30,000 $L_\odot$	Medium	0.13%
A	7,500–10,000 K	white	blue white	1.4–2.1 $M_\odot$	1.4–1.8 $R_\odot$	5–25 $L_\odot$	Strong	0.6%
F	6,000–7,500 K	yellow white	white	1.04–1.4 $M_\odot$	1.15–1.4 $R_\odot$	1.5–5 $L_\odot$	Medium	3%
G	5,200–6,000 K	yellow	yellowish white	0.8–1.04 $M_\odot$	0.96–1.15 $R_\odot$	0.6–1.5 $L_\odot$	Weak	7.6%
K	3,700–5,200 K	light orange	pale yellow orange	0.45–0.8 $M_\odot$	0.7–0.96 $R_\odot$	0.08–0.6 $L_\odot$	Very weak	12.1%
M	2,400–3,700 K	orange red	light orange red	0.08–0.45 $M_\odot$	≤ 0.7 $R_\odot$	≤ 0.08 $L_\odot$	Very weak	76.45%

- *Afterwords, it was realised that different classes correspond to different temperatures*

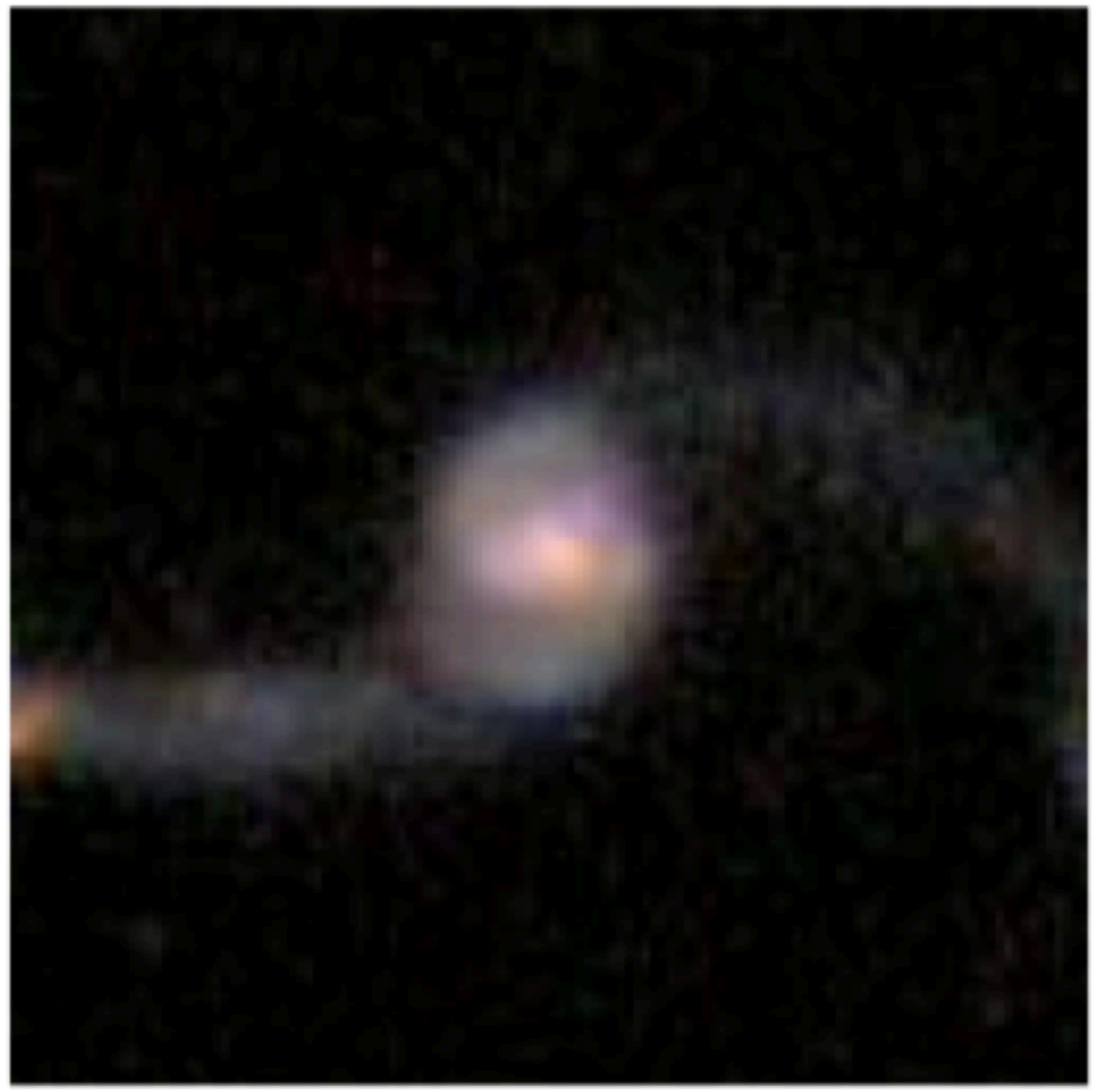


Learning from Anomalies

- Anomaly detection is one kind of data mining technique
  - One defines a metric of "typicality" to rank data samples
  - Based on this ranking, one can identify less typical events, tagging them as anomalies
  - By studying anomalies, one can make hypotheses on new physics mechanisms



Object ID: 960415



Anomaly Score: 4.470837

Back to 1984

- In the 1984 the UA1 experiment reported an excess of events with large missing transverse energy
- Before than, events with this signatures were extensively discussed with theorists (see “” for a first hand account of this)
- The community was looking for explanations (which eventually was provided by a combination of calorimeter cracks and tau decays)

EXPERIMENTAL OBSERVATION OF EVENTS WITH LARGE MISSING TRANSVERSE ENERGY ACCOMPANIED BY A JET OR A PHOTON (S) IN $p\bar{p}$ COLLISIONS AT $\sqrt{s}$ = 540 GeV

UA1 Collaboration, CERN, Geneva, Switzerland

G. ARNISON [m], O.C. ALLKOFER [g], A. ASTBURY [m,1], B. AUBERT [b], C. BACCI [ℓ], G. BAUER [p], A. BÉZAGUET [d], R.K. BOCK [d], T.J.V. BOWCOCK [h], M. CALVETTI [d], P. CATZ [b], P. CENNINI [d], S. CENTRO [2], F. CERADINI [ℓ], S. CITTOLIN [d], D. CLINE [p], C. COCHET [n], J. COLAS [b], M. CORDEN [c], D. DALLMAN [d,o], D. DAU [d,g], M. DeBEER [n], M. DELLA NEGRA [b,d], M. DEMOULIN [d], D. DENEGRI [n], D. DiBITONTO [d], A. DiCIACCIO [ℓ], L. DOBRZYNSKI [j], J. DOWELL [c], K. EGGERT [a], E. EISENHANDLER [h], N. ELLIS [d], P. ERHARD [a], H. FAISSNER [a], M. FINCKE [g,1], P. FLYNN [m], G. FONTAINE [j], R. FREY [k], R. FRÜHWIRTH [o], J. GARVEY [c], S. GEER [e], C. GHESQUIÈRE [j], P. GHEZ [b], W.R. GIBSON [h], Y. GIRAUD-HÉRAUD [j], A. GIVERNAUD [n], A. GONIDEC [b], G. GRAYER [m], T. HANSL-KOZANECKA [a], W.J. HAYNES [m], L.O. HERTZBERGER [i], D. HOFFMANN [a], H. HOFFMANN [d], D.J. HOLTHUIZEN [i], R.J. HOMER [c], A. HONMA [h], W. JANK [d], G. JORAT [d], P.I.P. KALMUS [h], V. KARIMÄKI [f], R. KEELER [h,1], I. KENYON [c], A. KERNAN [k], R. KINNUNEN [f], W. KOZANECKI [k], D. KRYN [d,j], P. KYBERD [h], F. LACAVA [ℓ], J.-P. LAUGIER [n], J.-P. LEES [b], H. LEHMANN [a], R. LEUCHS [g], A. LÉVÊQUE [d], D. LINGLIN [b], E. LOCCI [n], M. LORET [n], T. MARKIEWICZ [p], G. MAURIN [d], T. McMAHON [c], J.-P. MENDIBURU [j], M.-N. MINARD [b], M. MOHAMMADI [p], M. MORICCA [ℓ], K. MORGAN [k], F. MULLER [d], A.K. NANDI [m], L. NAUMANN [d], A. NORTON [d], A. ORKIN-LECOURTOIS [j], L. PAOLUZI [ℓ], F. PAUSS [d], G. PIANO MORTARI [ℓ], E. PIETARINEN [f], M. PIMIÄ [f], D. PITMAN [k], A. PLACCI [d], J.-P. PORTE [d], E. RADERMACHER [a], J. RANSDELL [k], H. REITHLER [a], J.-P. REVOL [d], J. RICH [n], M. RIJSSENBEEK [d], C. ROBERTS [m], J. ROHLF [e], P. ROSSI [d], C. RUBBIA [d], B. SADOULET [d], G. SAJOT [j], G. SALVINI [ℓ], J. SASS [n], A. SAVOY-NAVARRO [n], D. SCHINZEL [d], W. SCOTT [m], T.P. SHAH [m], I. SHEER [k], D. SMITH [k], J. STRAUSS [o], J. STREETS [c], K. SUMOROK [d], F. SZONCSO [o], C. TAO [j], G. THOMPSON [h], J. TIMMER [d], E. TSCHESLOG [a], J. TUOMINIEMI [f], B. Van EIJK [i], J.-P. VIALLE [b], J. VRANA [j], V. VUILLEMIN [d], H.D. WAHL [o], P. WATKINS [c], J. WILSON [c], C.-E. WULZ [o] and M. YVERT [b]

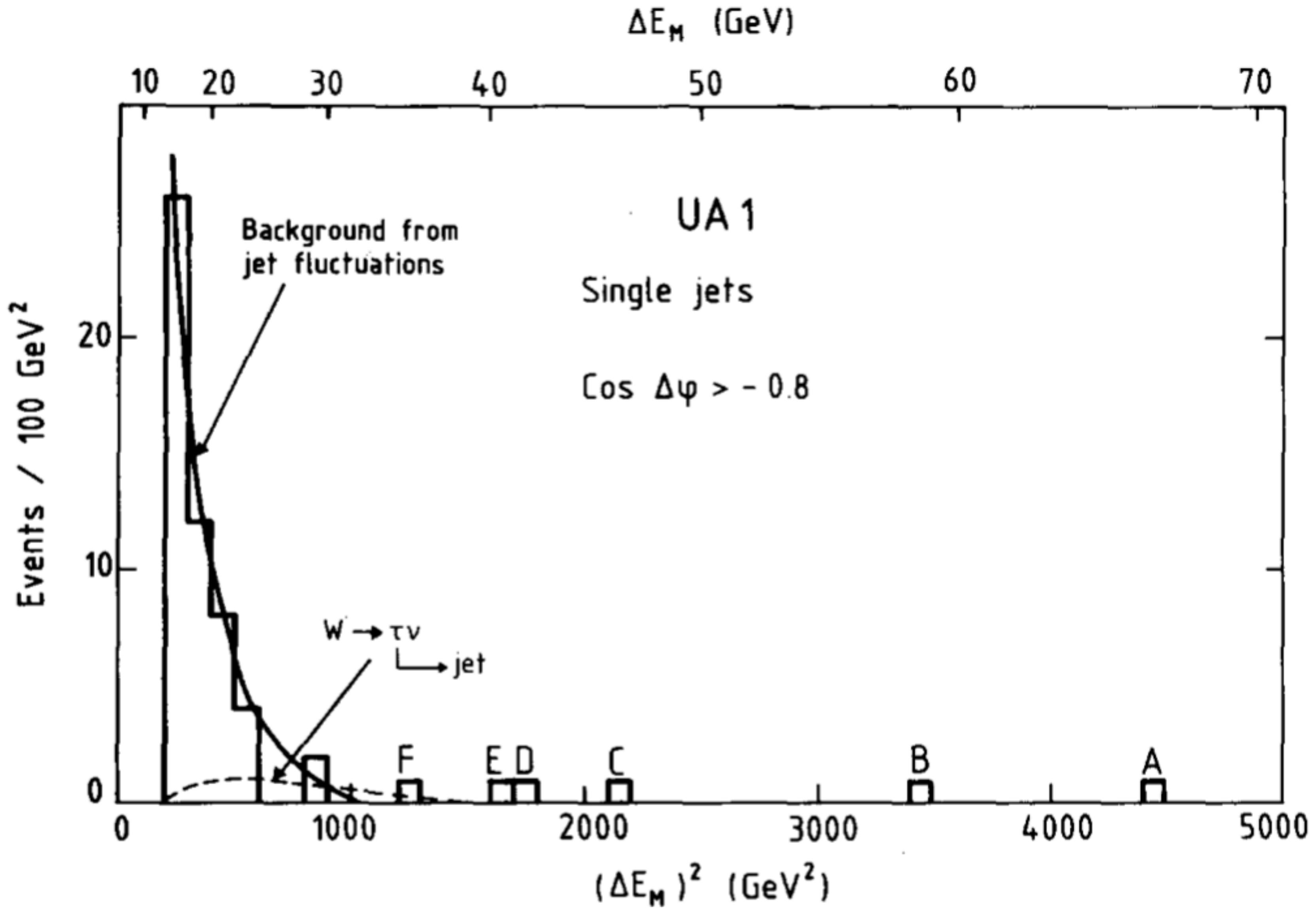
Aachen [a]–Annecy (LAPP) [b]–Birmingham [c]–CERN [d]–Harvard [e]–Helsinki [f]–Kiel [g]–Queen Mary College, London [h]–NIKHEF, Amsterdam [i]–Paris (Coll. de France) [j]–Riverside [k]–Roma [ℓ]–Rutherford Appleton Lab. [m]–Saclay (CEN) [n] Vienna [o]–Wisconsin [p] Collaboration

Received 30 March 1984

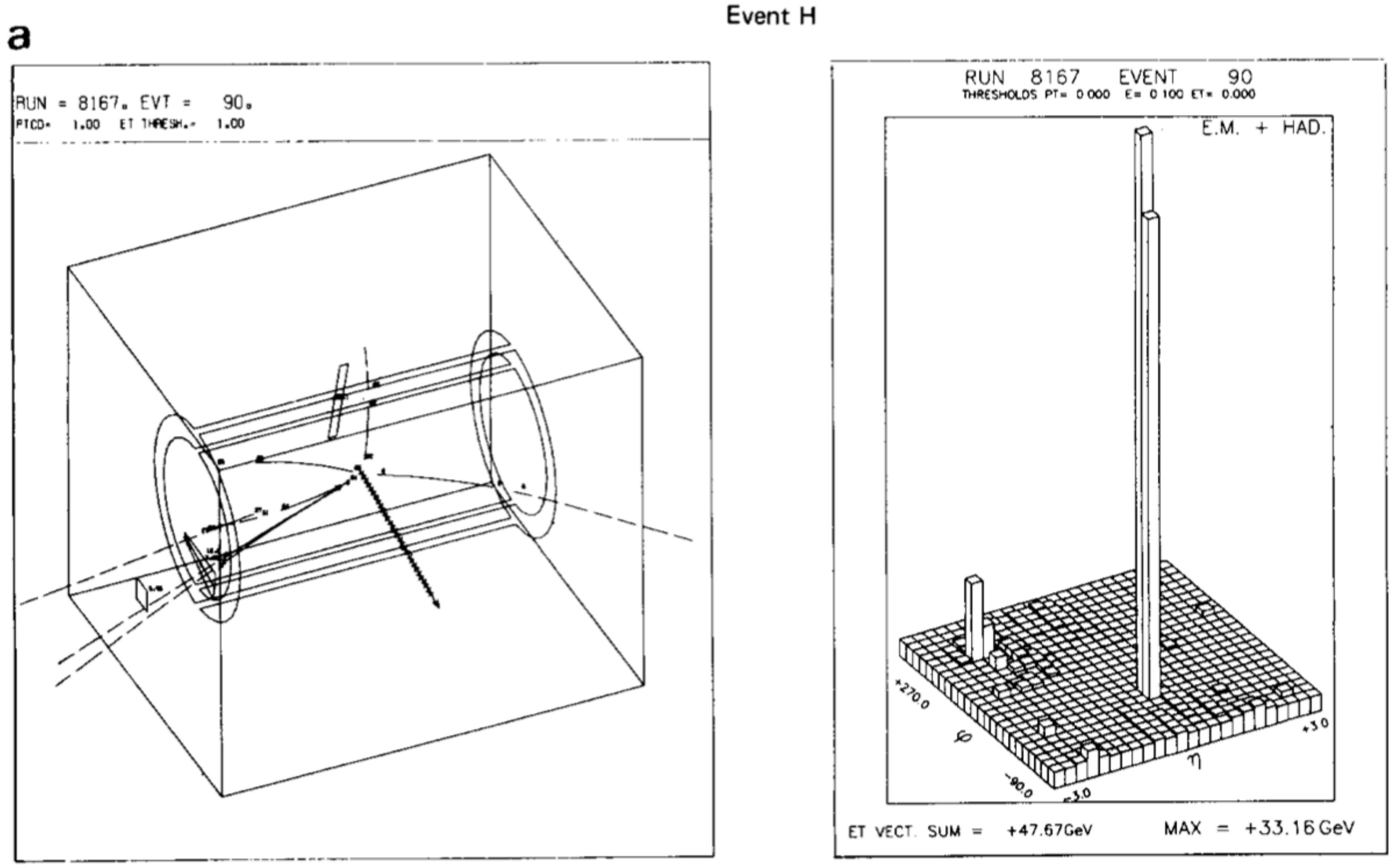
European Research Council

Back to 1984

- *In the article, one sees the seeds of modern large-scale data analysis techniques*
- *But the paper is more about single events, event displays, etc. and not just significance, limits, p-value and interpretation*
- *Data, and not their statistical interpretation, was central*
- *Certainly, we moved away from that for good reason (blind analysis, etc.)*
- *On the other hand, aren't we missing something?*



Event H



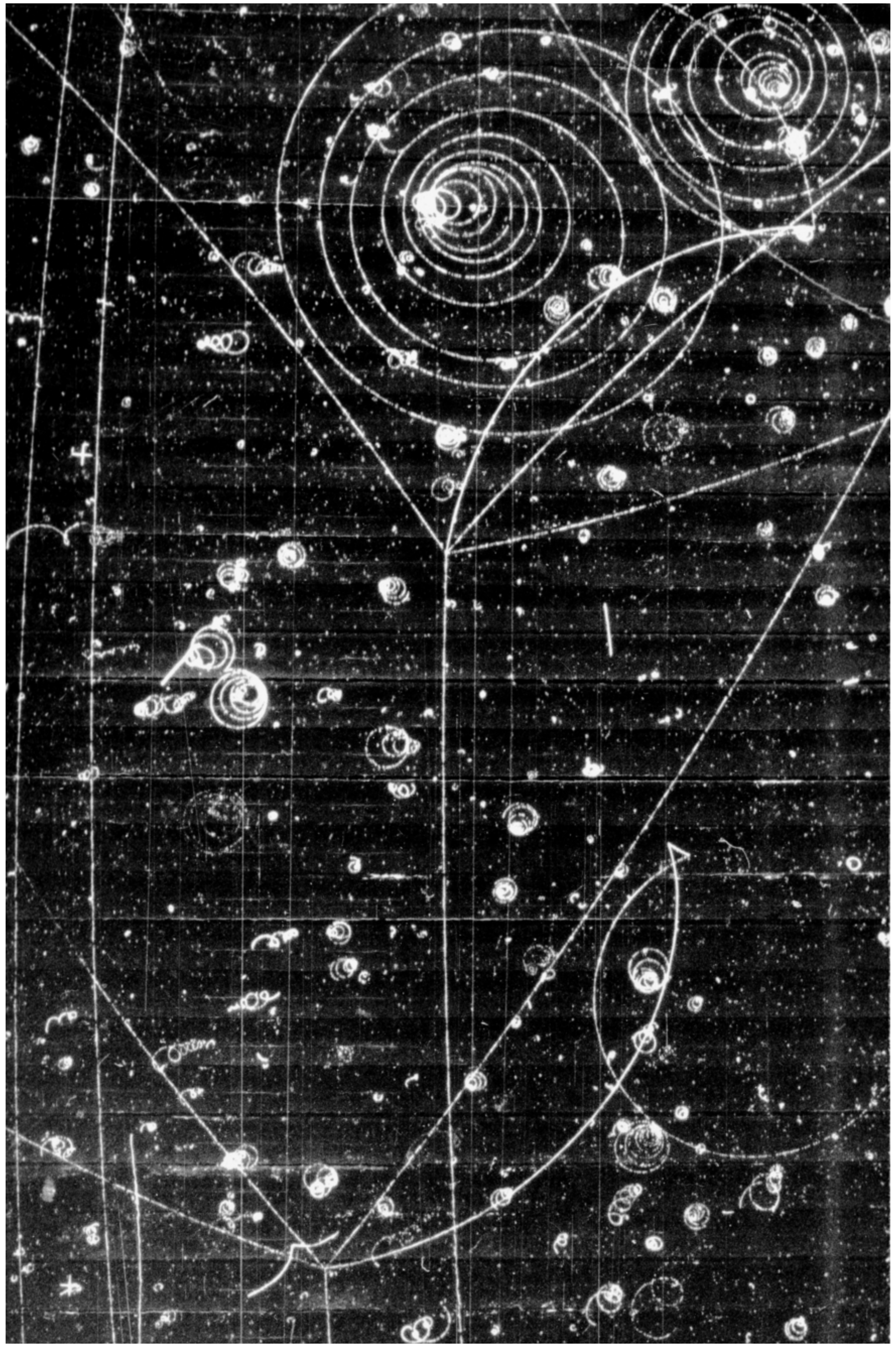
Looking at data used to be OK

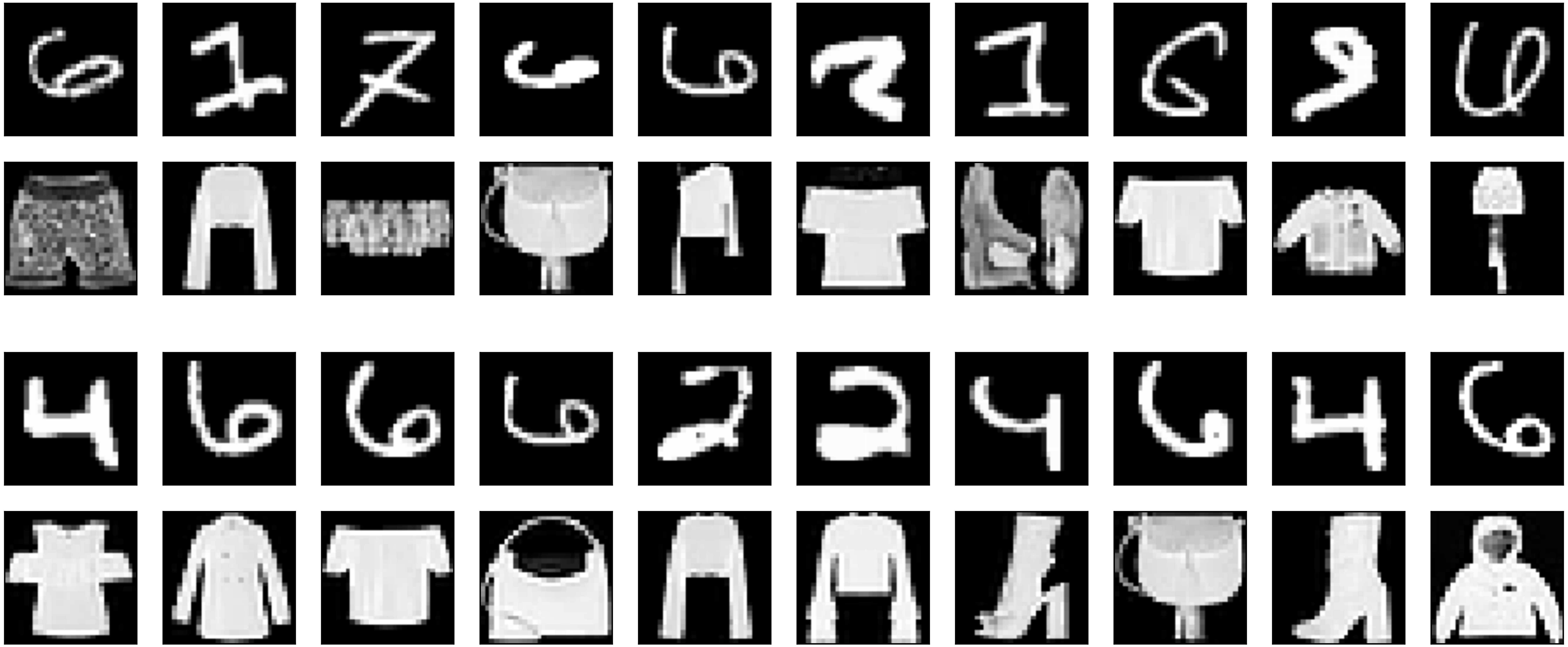
- *Our community looked at data for decades. It was the standard before the new standard (large-scale blind statistical analyses) became a thing*

- *I am not saying we should go back (Discoveries have to be based on reasonable statistical procedures)*

- *I am saying that we should have a pre-analysis step in which we look at data to identify reasonable signatures.*

- *Model independent searches are a way to do this. But there are other ways, in which data are made more central*





Autoencoders for anomaly detection

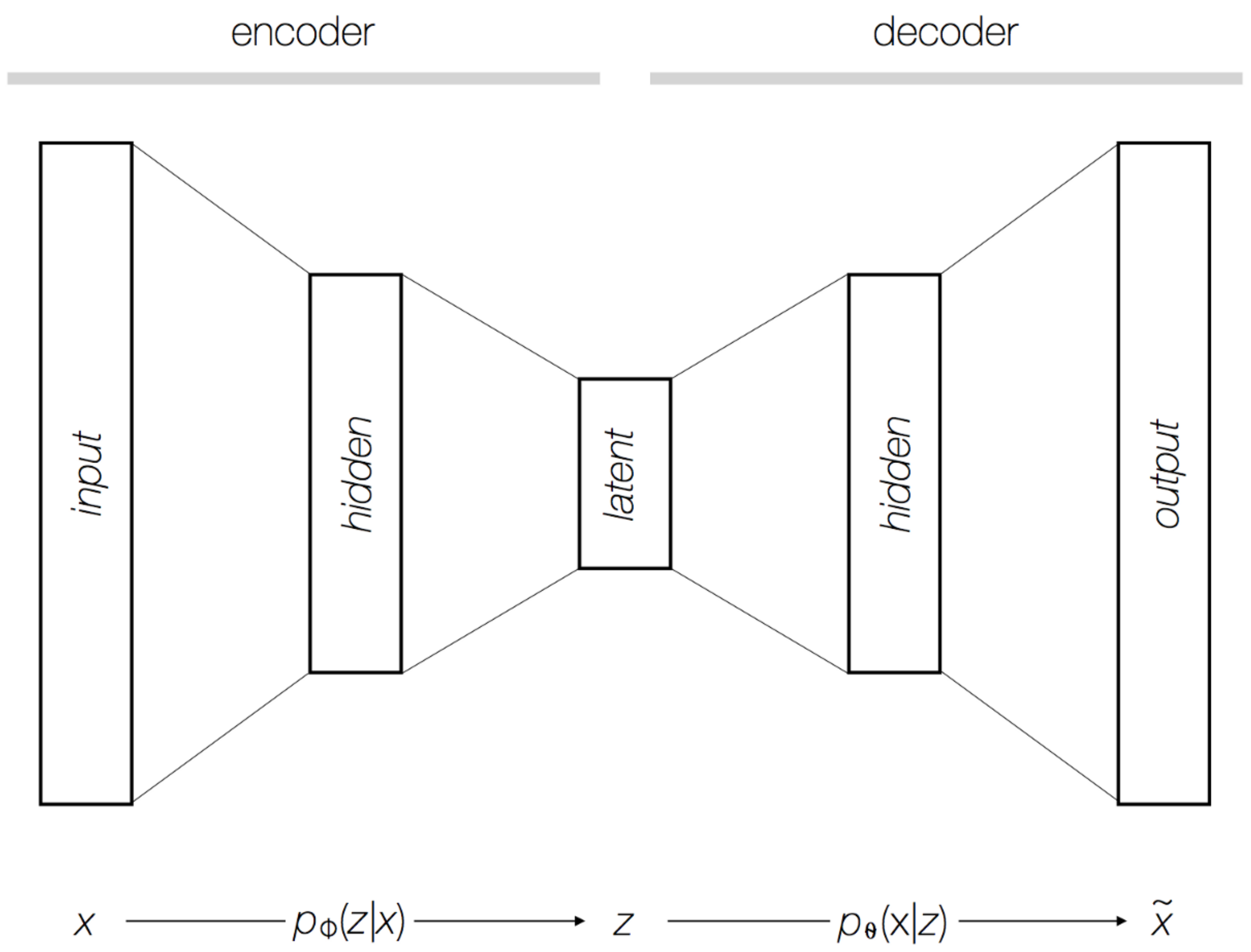
Autoencoders in a nutshell

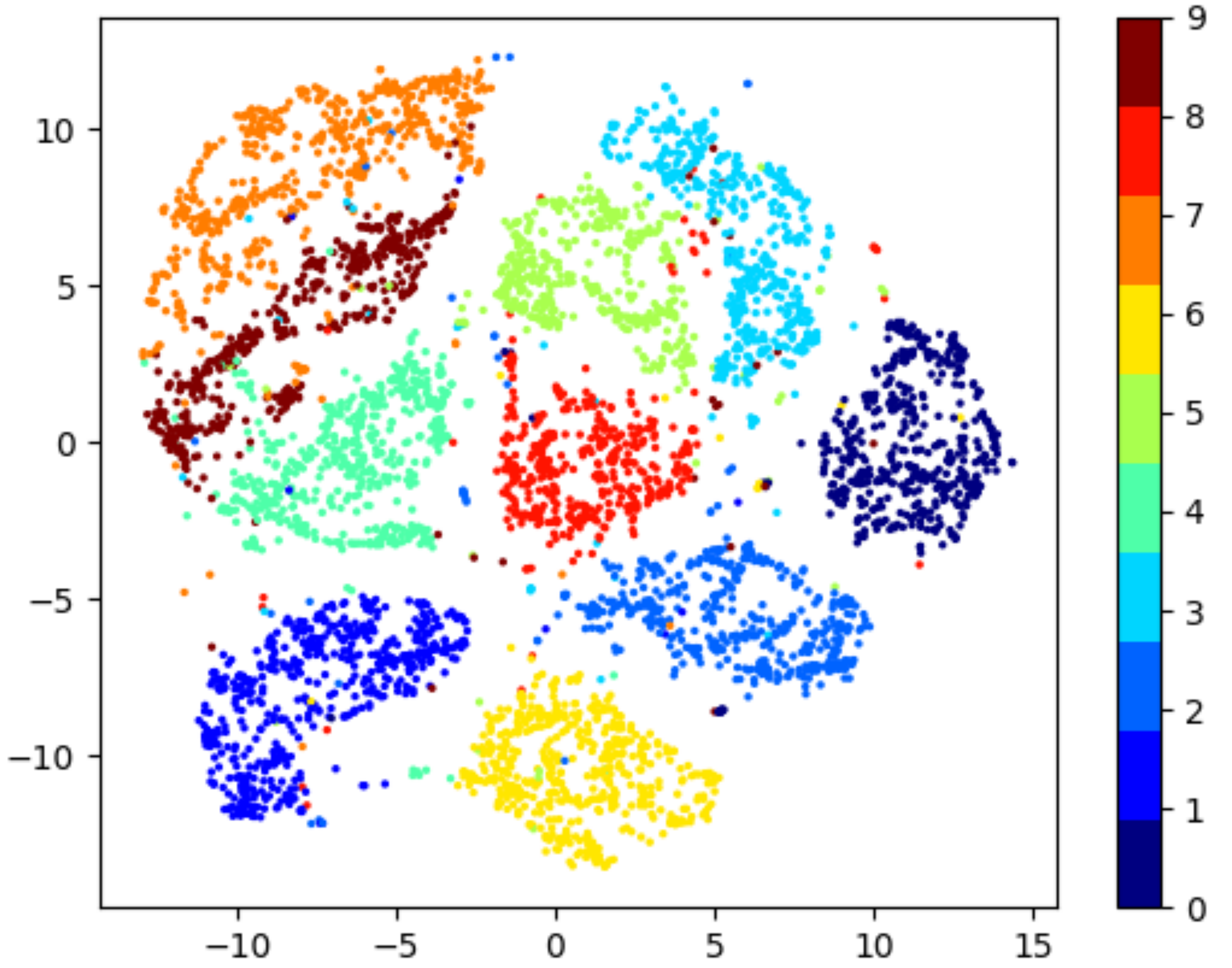
- Autoencoders are compression-decompression algorithms that learn to describe a given dataset in terms of points in a lower-dimension latent space

- UNSUPERVISED algorithm, used for data compression, generation, clustering (replacing PCA), etc.

- Used in particular for anomaly detection: when applied on events of different kind, compression-decompression tuned on refer sample might fail

- One can define anomalous any event whose decompressed output is “far” from the input, in some metric (e.g., the metric of the auto-encoder loss)





Proof of concept: ℓ+X @HLT

- *Consider a stream of data coming from L1*
  - *Passed L1 because of 1 lepton (e,m) with pT>23 GeV*
  - *At HLT, very loose isolation applied*
  - *Sample mainly consists of W, Z, tt & QCD (for simplicity, we ignore the rest)*

Standard Model processes					
Process	Acceptance	Trigger efficiency	Cross section [nb]	Events fraction	Event /month
W	55.6%	68%	58	59.2%	110M
QCD	0.08%	9.6%	$1.6 \cdot 10^5$	33.8%	63M
Z	16%	77%	20	6.7%	12M
$t\bar{t}$	37%	49%	0.7	0.3%	0.6M

- *We consider 21 features, typically highlighting the difference between these SM processes (no specific BSM signal in mind)*

- The isolated-lepton transverse momentum p_T^ℓ.
- The three isolation quantities (CHPFISO, NEUPFISO, GAMMAPFISO) for the isolated lepton, computed with respect to charged particles, neutral hadrons and photons, respectively.
- The lepton charge.
- A boolean flag (ISELE) set to 1 when the trigger lepton is an electron, 0 otherwise.
- S_T, i.e. the scalar sum of the p_T of all the jets, leptons, and photons in the event with $p_T > 30$ GeV and $|\eta| < 2.6$. Jets are clustered from the reconstructed PF candidates, using the FASTJET [23] implementation of the anti-k_T jet algorithm [24], with jet-size parameter R=0.4.
- The number of jets entering the S_T sum (N_J).
- The invariant mass of the set of jets entering the S_T sum (M_J).
- The number of these jets being identified as originating from a b quark (N_b).
- The missing transverse momentum, decomposed into its parallel ($p_{T,\parallel}^{\text{miss}}$) and orthogonal ($p_{T,\perp}^{\text{miss}}$) components with respect to the isolated lepton direction. The missing transverse momentum is defined as the negative sum of the PF-candidate p_T vectors:

$$\vec{p}_T^{\,\text{miss}} = -\sum_q \vec{p}_T^{\,q} \,. \tag{2}$$

- The transverse mass, M_T, of the isolated lepton ℓ and the E_T^{miss} system, defined as:

$$M_T = \sqrt{2p_T^\ell E_T^{\text{miss}}(1-\cos\Delta\phi)} \,, \tag{3}$$

with $\Delta\phi$ the azimuth separation between the lepton and $\vec{p}_T^{\,\text{miss}}$ vector, and E_T^{miss} the absolute value of $\vec{p}_T^{\,\text{miss}}$.

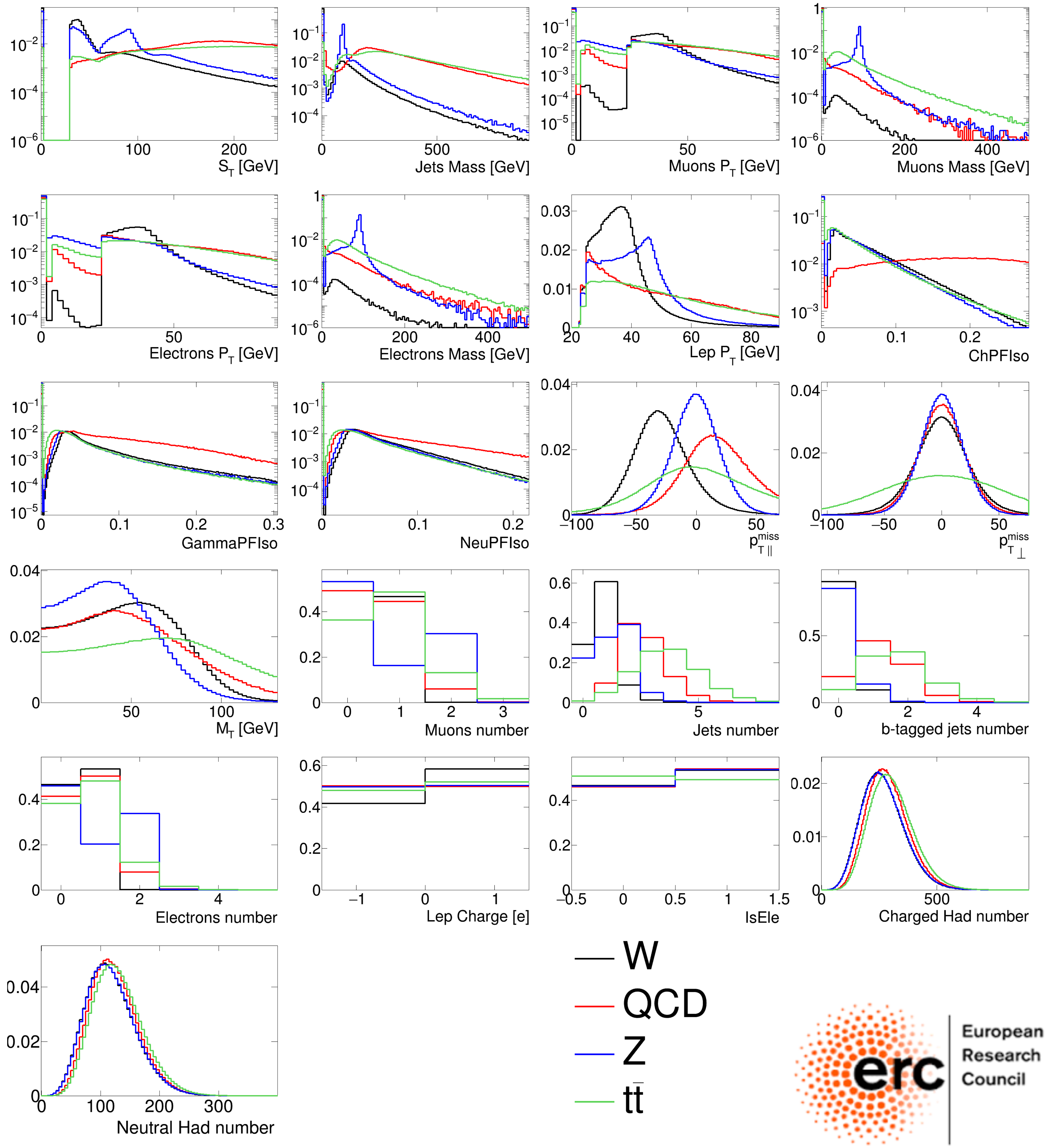
- The number of selected muons (N_μ).
- The invariant mass of this set of muons (M_μ).
- The total transverse momentum of these muons ($p_{T,TOT}^\mu$).
- The number of selected electrons (N_e).
- The invariant mass of this set of electrons (M_e).
- The total transverse momentum of these electrons ($p_{T,TOT}^e$).
- The number of reconstructed charged hadrons.
- The number of reconstructed neutral hadrons.

European Research Council

Proof of concept: ℓ+X @HLT

- Consider a stream of data coming from L1
  - Passed L1 because of 1 lepton (e,m) with pT>23 GeV
  - At HLT, very loose isolation applied
  - Sample mainly consists of W, Z, tt & QCD (for simplicity, we ignore the rest)

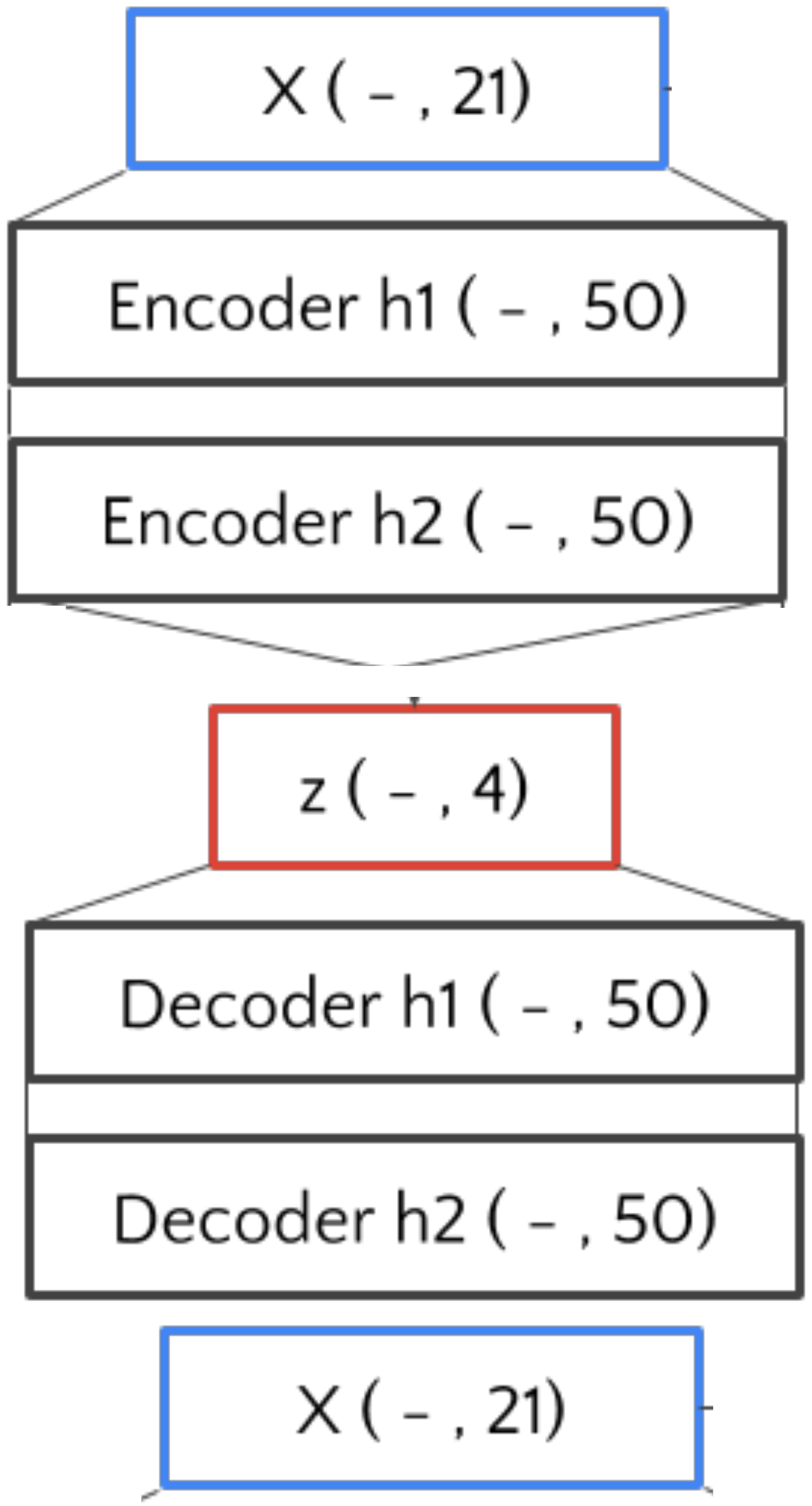
Standard Model processes					
Process	Acceptance	Trigger efficiency	Cross section [nb]	Events fraction	Event /month
W	55.6%	68%	58	59.2%	110M
QCD	0.08%	9.6%	$1.6 \cdot 10^5$	33.8%	63M
Z	16%	77%	20	6.7%	12M
$t\bar{t}$	37%	49%	0.7	0.3%	0.6M



- We consider 21 features, typically highlighting the difference between these SM processes (no specific BSM signal in mind)

Standard Model AE

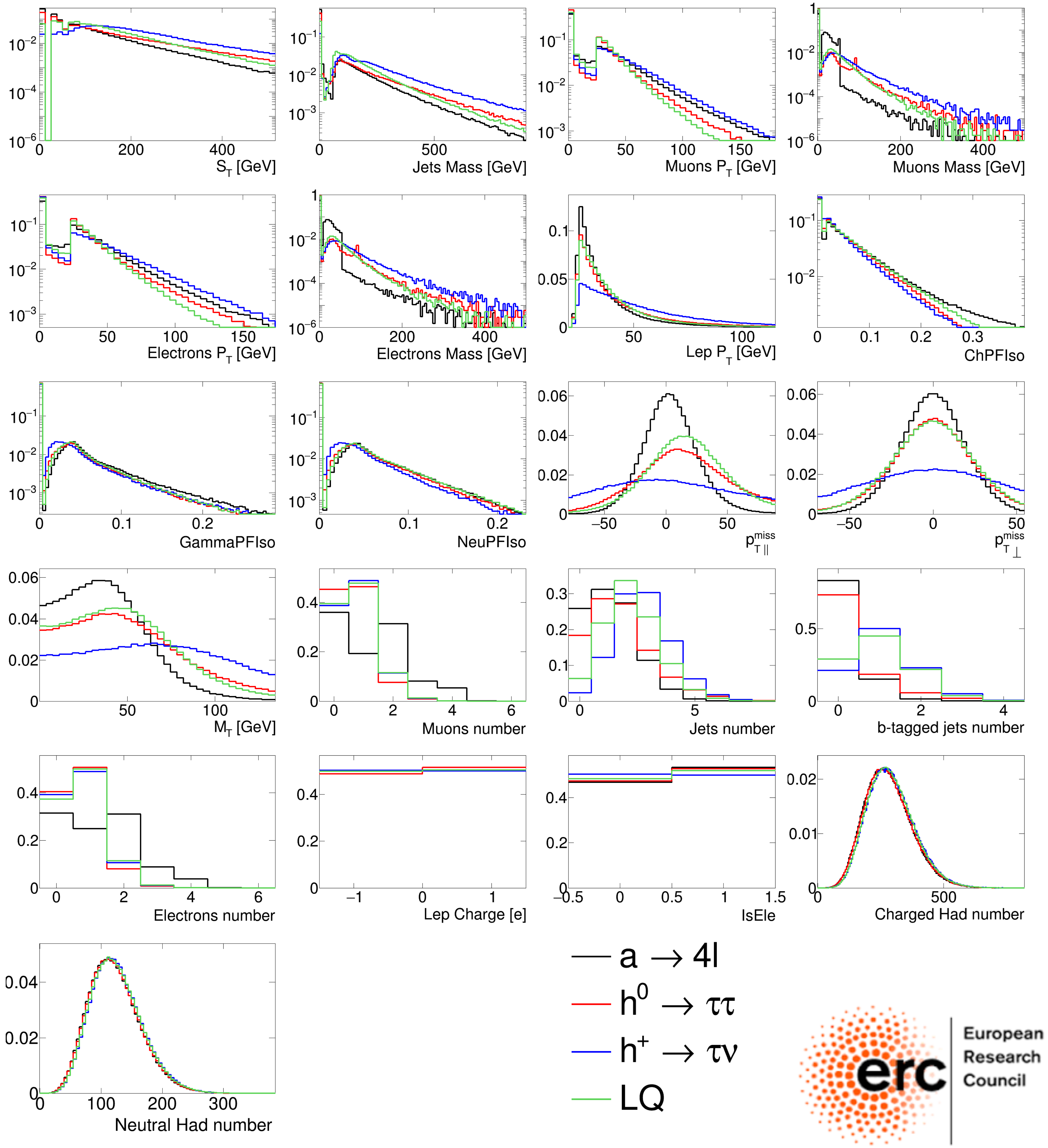
- *We train a VAE on a cocktail of SM events (weighted by xsec)*
- ***ENCODER:*** *21 inputs, 2 hidden layers → 4Dim latent space*
- ***DECODER:*** *from a random sample in the 4D space → 2 hidden layers → 21 outputs*



Some BSM benchmark

- We consider four BSM benchmark models, to give some sense of VAEs potential
  - leptoquark with mass 80 GeV, LQ→bτ
  - A scalar boson with mass 50 GeV, $a \to Z^*Z^* \to 4\ell$
  - A scalar scalar boson with mass 60 GeV, $h \to \tau\tau$
  - A charged scalar boson with mass 60 GeV, $h^{\pm} \to \tau\nu$

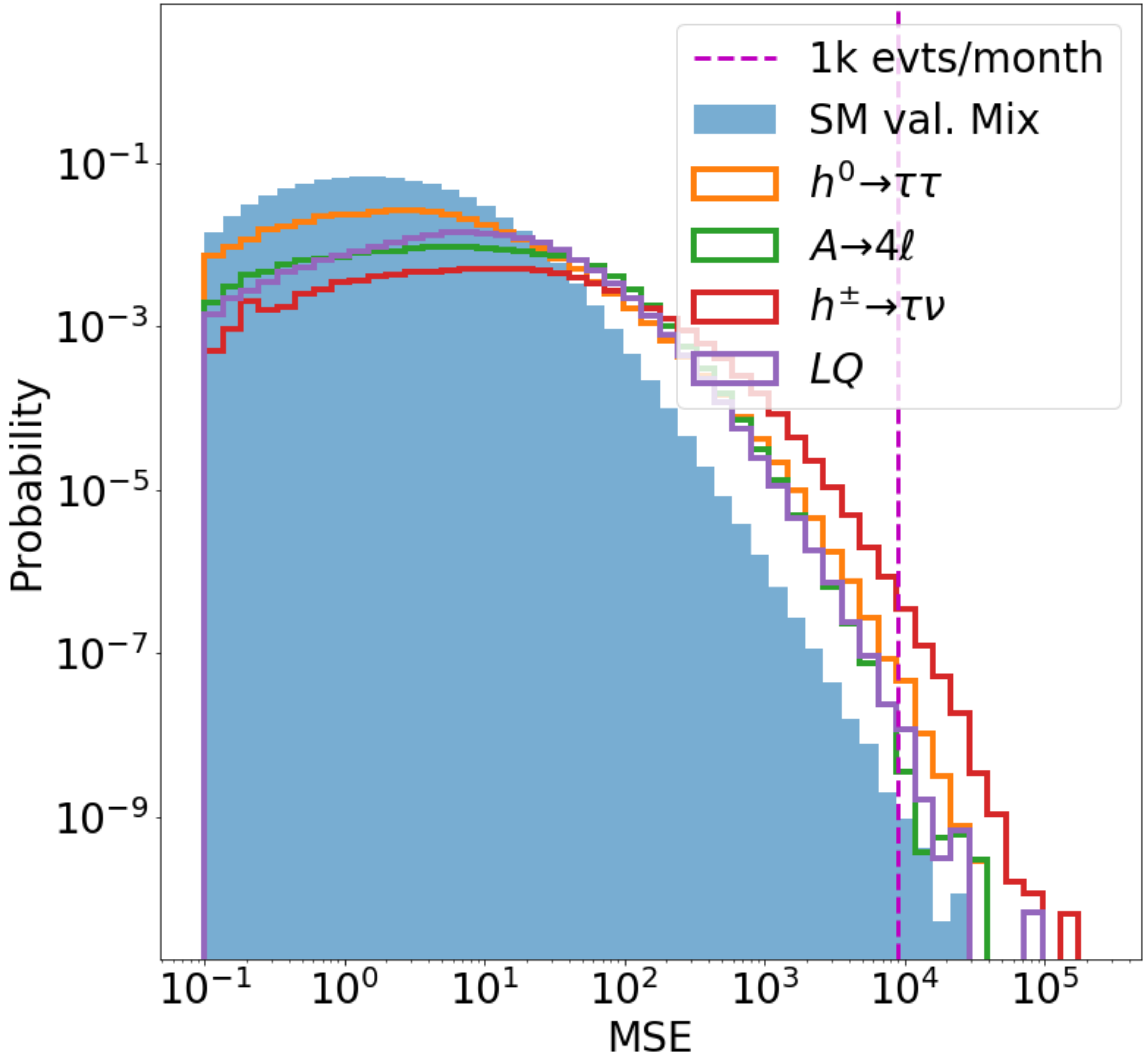
BSM benchmark processes				
Process	Acceptance	Trigger efficiency	Total efficiency	Cross-section 100 events/month
$h^0 \to \tau\tau$	9%	70%	6%	335 fb
$h^0 \to \tau\nu$	18%	69%	12%	163 fb
$LQ \to b\tau$	19%	62%	12%	166 fb
$a \to 4\ell$	5%	98%	5%	436 fb



European Research Council

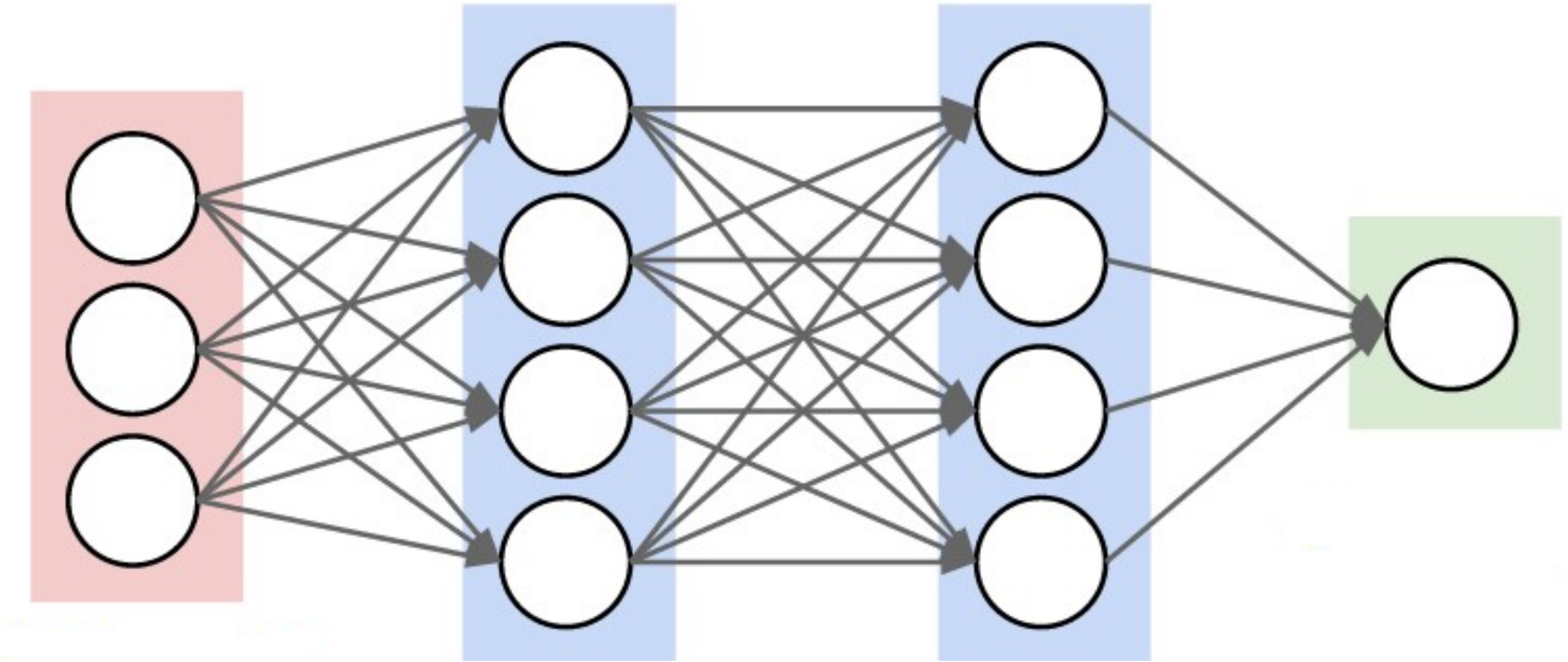
Defining anomaly

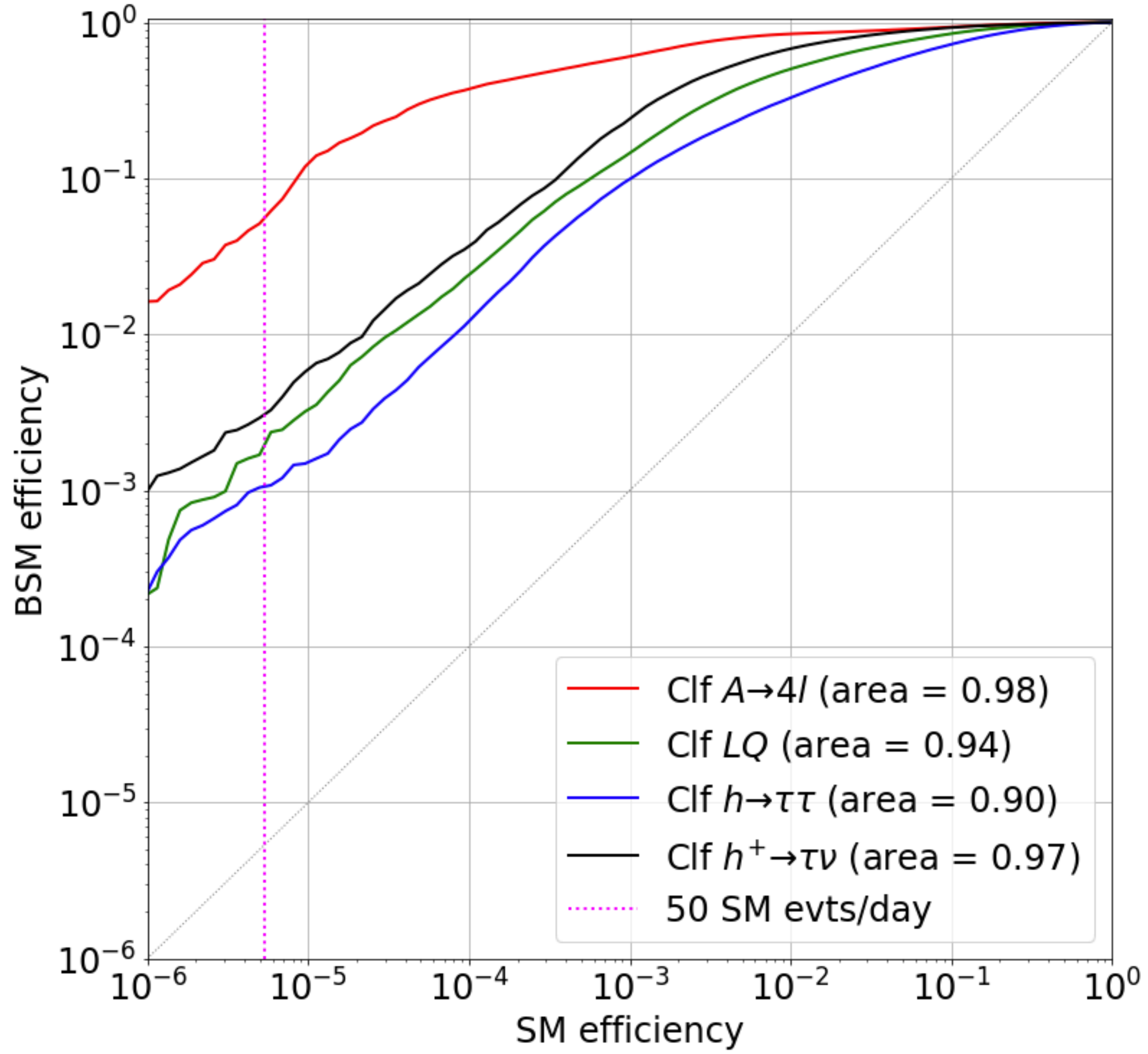
- *Anomaly defined as a p-value threshold on a given test statistics*
  - *Loss function an obvious choice*
  - *Some part of a loss could be more sensitive than others*
  - *We tested different options and found the total loss to behave better*



Benchmark comparison

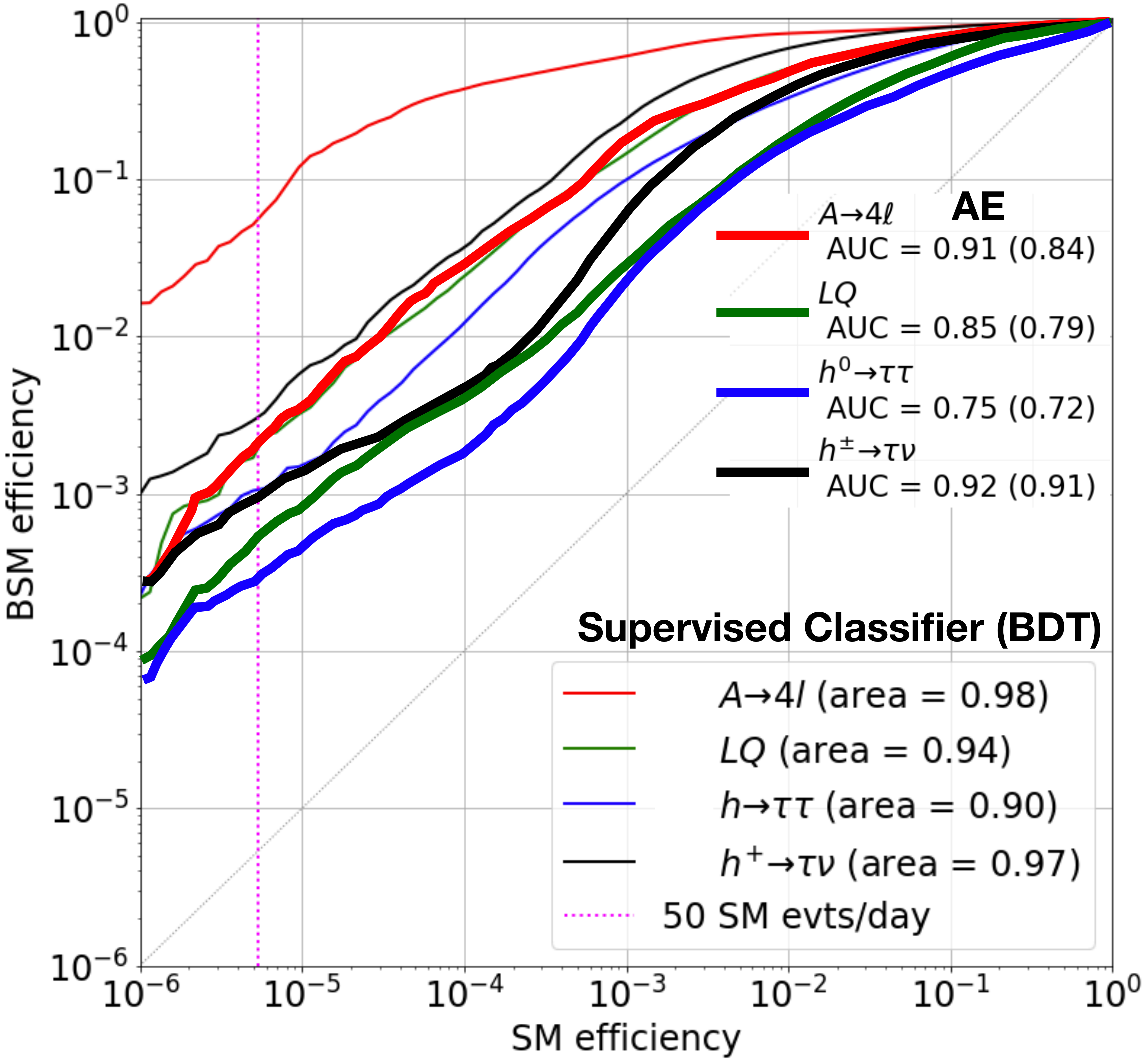
- VAE's performances benchmarked against supervised classifiers
- For each BSM model
  - take same inputs as VAE
  - train a fully-supervised classifier to separate signal from background
  - use supervised performances as a reference to aim to with the unsupervised approach
  - Done for our 4 BSM models using dense neural networks





Performances

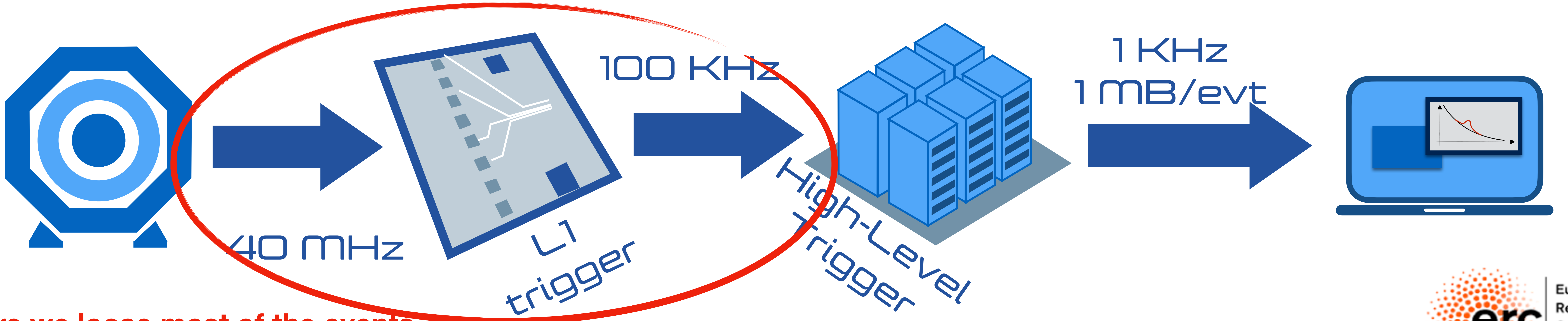
- *Evaluate general discrimination power by ROC curve and area under curve (AUC)*
  - *clearly worse than supervised*
  - *but not so far*
- *Fixing SM acceptance rate at 50 events/day*
  - *competitive results considering unsupervised nature of the algorithm*



Performances

- *Small efficiency but still much larger than for SM processes*
- *Allows to probe 10-100 pb cross sections for reasonable amount of collected signal events*

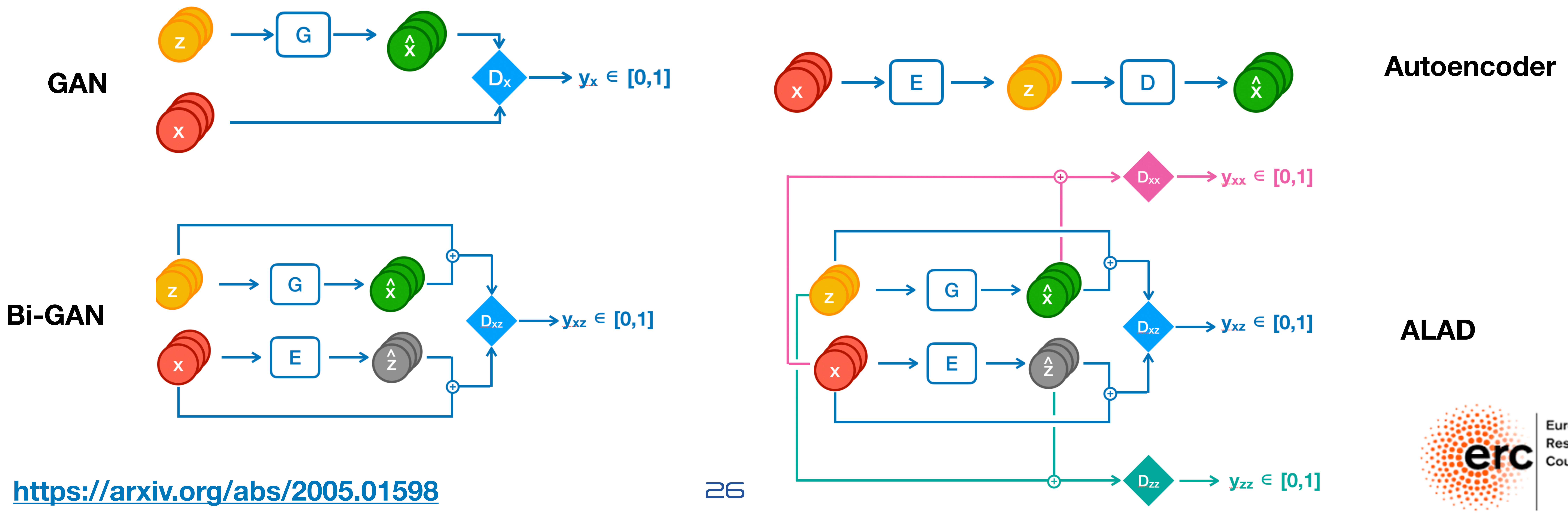
Process	Efficiency for ~30 evt/day	xsec for 100 evt/month [pb]	xsec for S/B~1/3 [pb]
$a \to 4\ell$	$2.8 \cdot 10^{-3}$	7.1	27
$LQ \to \tau b$	$6.5 \cdot 10^{-4}$	31	120
$h \to \tau\tau$	$3.6 \cdot 10^{-4}$	56	220
$h^{\pm} \to \tau\nu$	$1.2 \cdot 10^{-3}$	17	67



This is where we loose most of the events
-> This is where one would run this

Re-discovering the top quark

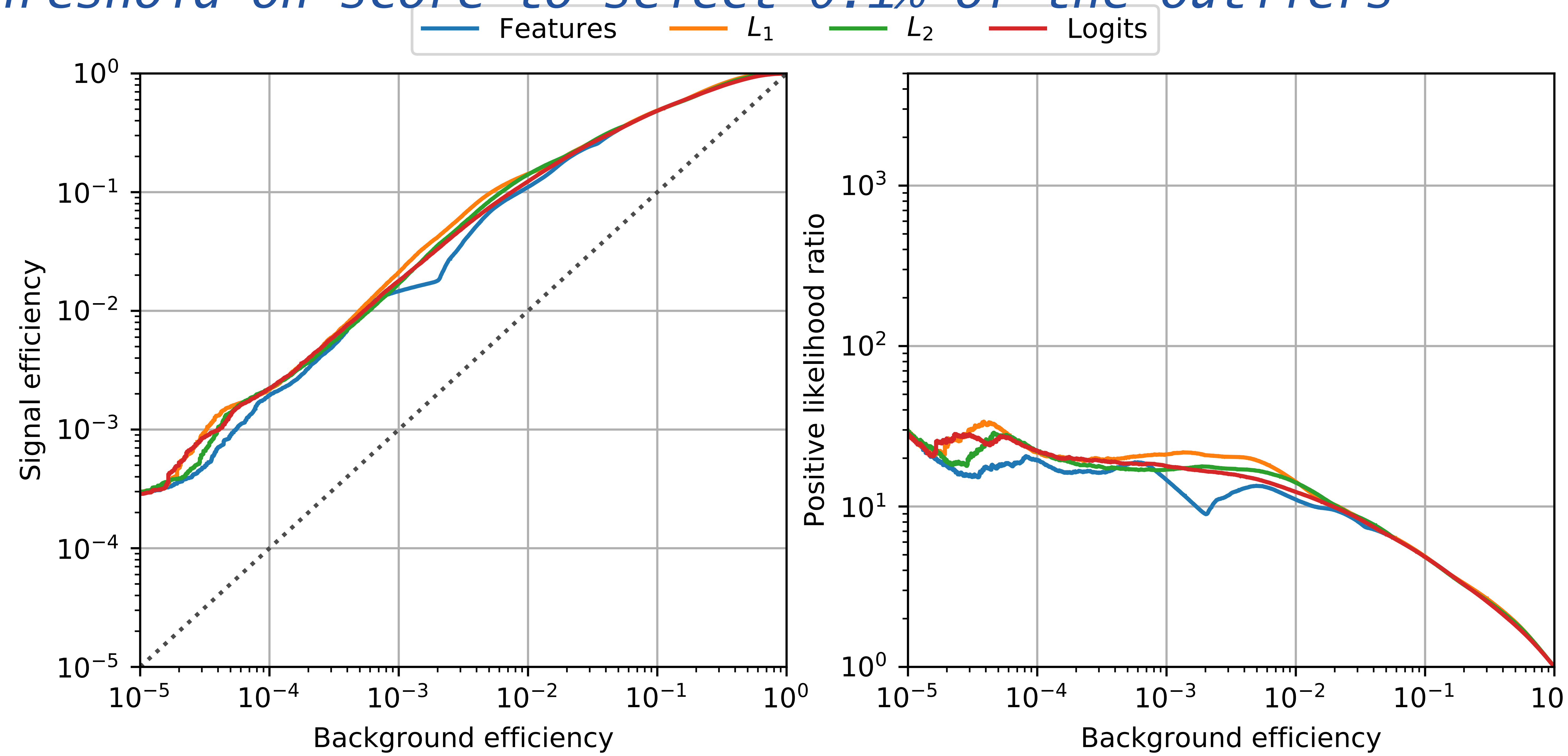
- *We use one kind of ADA on real CMS data to re-discover the top quark*
  - *5 fb^{-1} of 8 TeV CMS Open Data from 2012*
  - *SingleMu dataset*
- *We trained an Adversarially Learned Anomaly Detection algorithm*
  - *a GAN powered with an encoder*
  - *or an auto-encoder powered with adversarial training*
- *We apply threshold on score to select 0.1% of the outliers*



https://arxiv.org/abs/2005.01598

Re-discovering the top quark

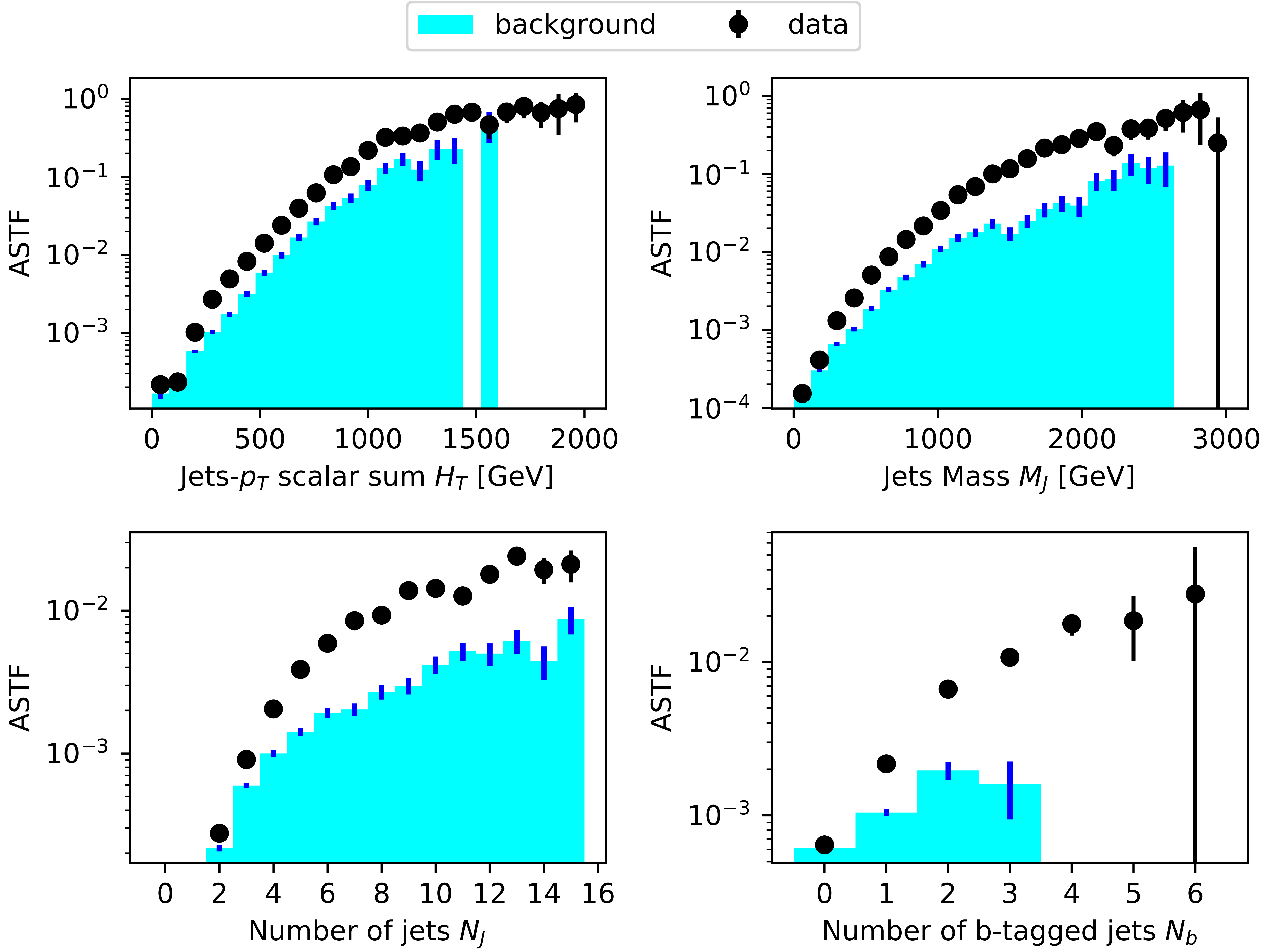
- *We applied this idea to real data*
  - *5 fb^{-1} of 8 TeV CMS Open Data from 2012*
  - *SingleMu dataset*
- *We trained an Adversarially Learned Anomaly Detection algorithm*
  - *a GAN powered with an encoder*
  - *or an auto-encoder powered with adversarial training*
- *We apply threshold on score to select 0.1% of the outliers*



https://arxiv.org/abs/2005.01598

Re-discovering the top quark

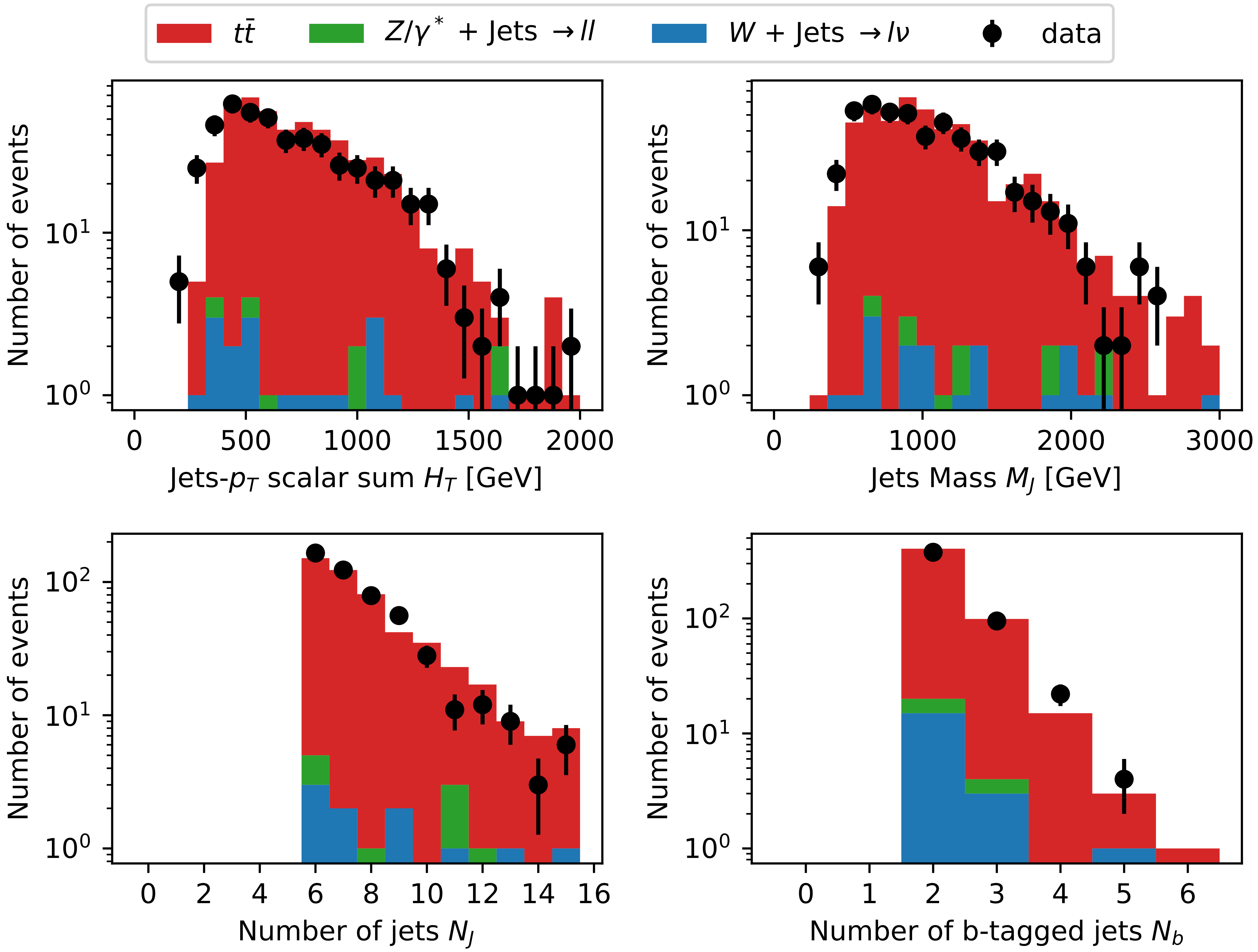
- We then look at differential accept/reject ratios (data vs MC) to get an idea of where the anomalies (if any) are clustering
- In this case, we have indication that anomalies come with many jets, some of which are b-jets
- We require >5j and >1 b-jet and expect ~ 0 standard events
- We see a lot of them: an almost pure sample of anomalies that we can further inspect (and that the MC is telling us are actually tt events)



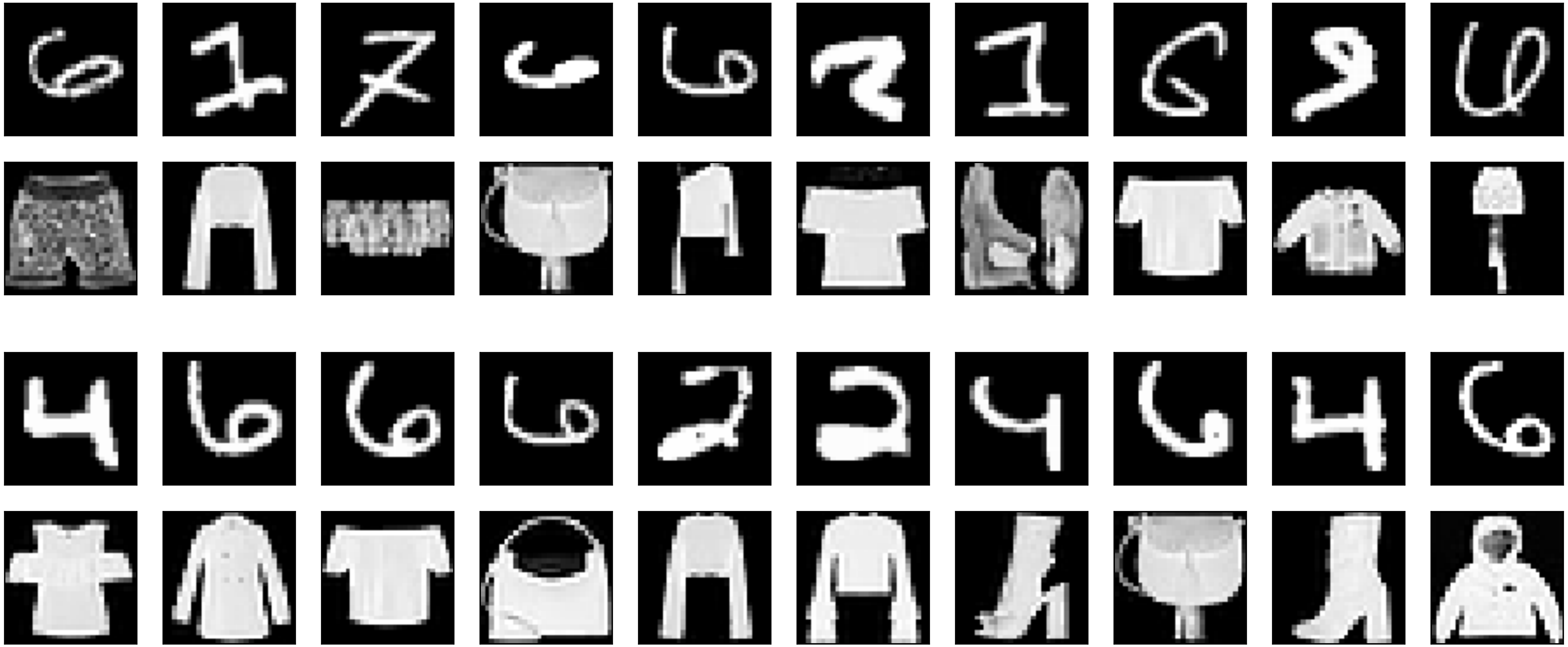
https://arxiv.org/abs/2005.01598

Re-discovering the top quark

- *We then look at differential accept/reject ratios (data vs MC) to get an idea of where the anomalies (if any) are clustering*
- *In this case, we have indication that anomalies come with many jets, some of which are b-jets*
- *We require >5j and >1 b-jet and expect ~ 0 standard events*
- *We see a lot of them: an almost pure sample of anomalies that we can further inspect (and that the MC is telling us are actually tt events)*



https://arxiv.org/abs/2005.01598

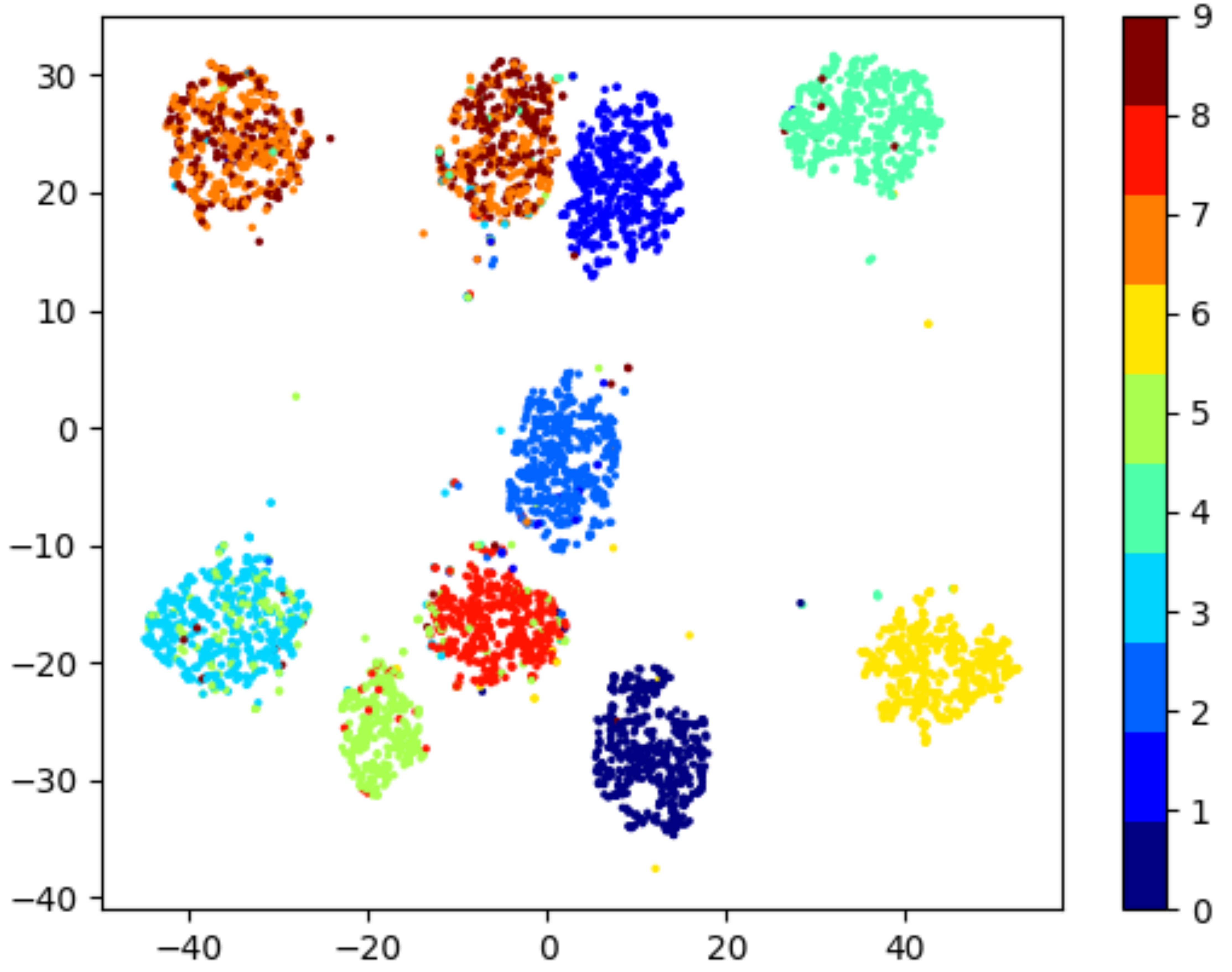


What to do with anomalies?

European Research Council

What to do with these data?

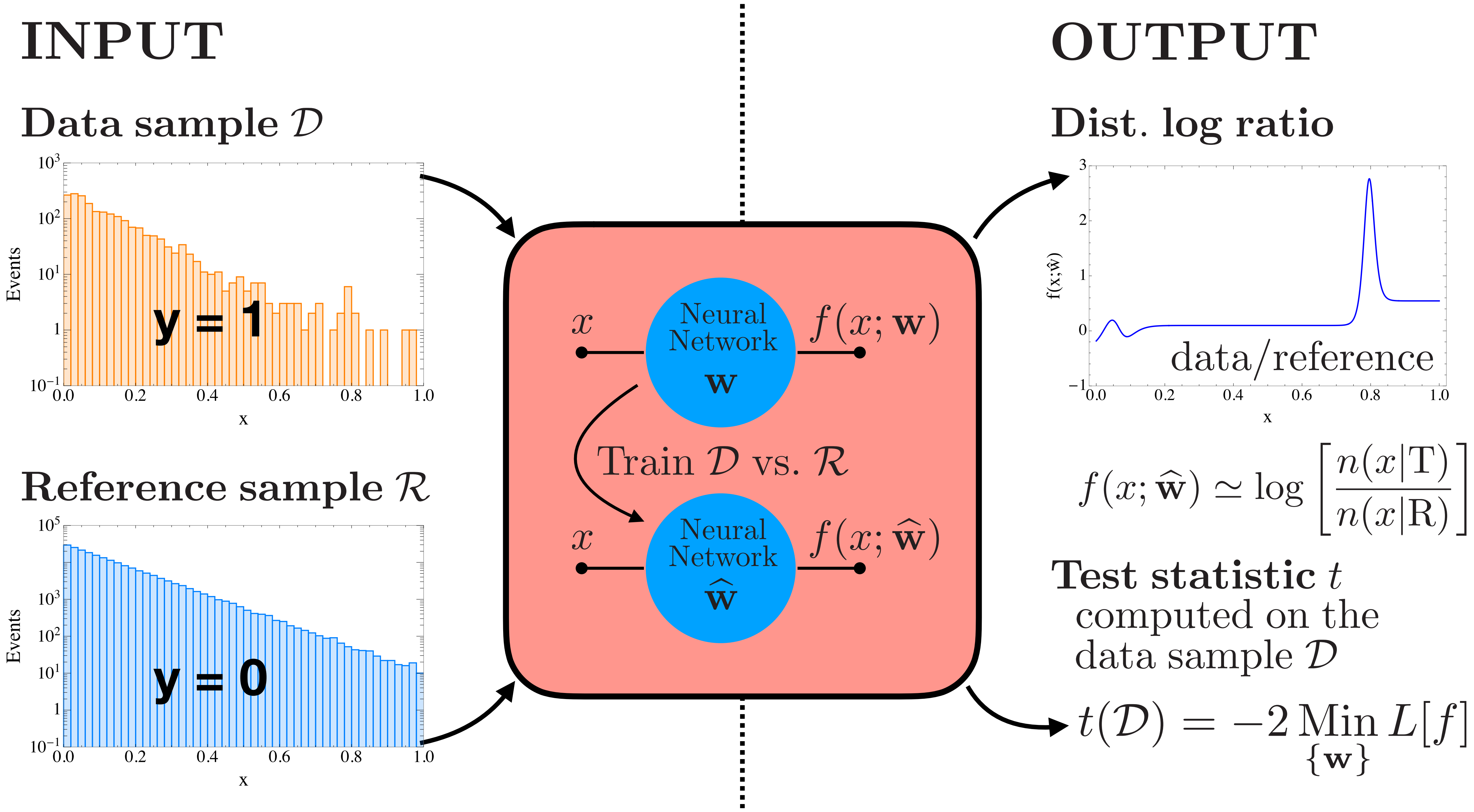
- *We could learn a lot running clustering algorithms (KNN, etc) on these data*
  - *In the latent space of the AE*
  - *In the natural space of the input*
  - *With any other similar technique*
- *In my mind, a descriptive paper on such an analysis would be a valuable publication, particularly before a long shutdown.*
- *Provided control on the background distribution (not for granted), we could run a statistical analysis on them and quote a significance (e.g., with https://arxiv.org/abs/1806.02350)*
- *Publishing the dataset as a catalog could incentive new ideas in view of HL-LHC*
- *While we sort out the technical details (e.g., with TSG and L1), we would like to request the EXO PAG to support the idea*



“Model-independent” hypothesis test

- Deep Learning could help relaxing the underlying hypotheses of a new-physics search
  - stay within the hypothesis test framework
  - replace the fully specified (model dependent) signal hypothesis with a neural network trained on data
  - exploit neural networks to express different model shapes at once
- Training setup to learn the likelihood ratio of a traditional search
- Formally, still a fully-supervised learning process

$$\boxed{L[f] = \sum_{(x,y)} \left[(1-y)\frac{N(\mathrm{R})}{\mathcal{N}_{\mathcal{R}}}\left(e^{f(x)}-1\right) - y\, f(x) \right]}$$



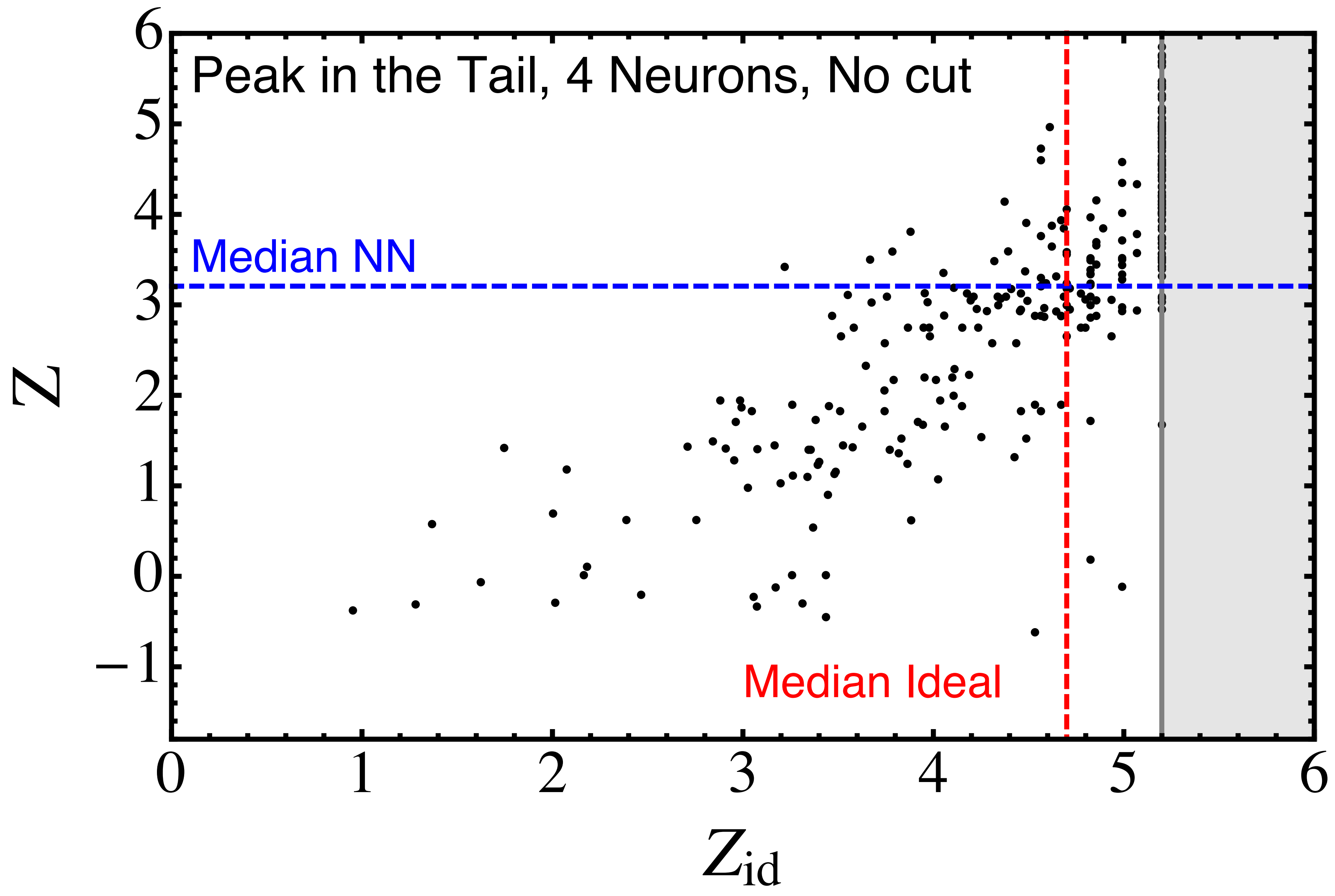
$$\underset{\{\mathbf{w}\}}{\mathrm{Min}}\, L = -\underset{\{\mathbf{w}\}}{\mathrm{Max}} \left\{ \log\left[\frac{e^{-N(\mathbf{w})}}{e^{-N(\mathrm{R})}} \prod_{x\in\mathcal{D}} \frac{n(x|\mathbf{w})}{n(x|\mathrm{R})} \right] \right\} = -\frac{t(\mathcal{D})}{2}\,.$$

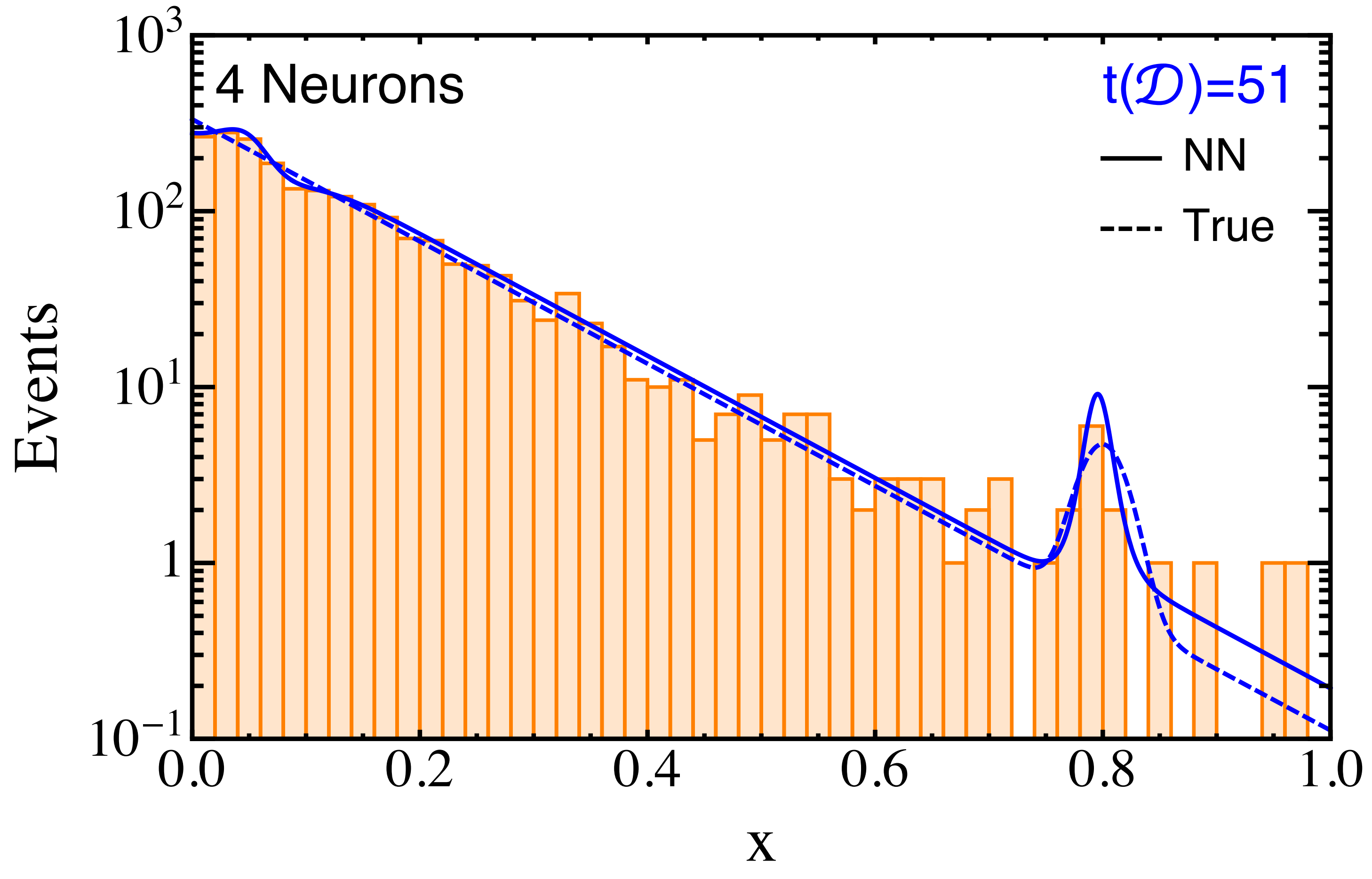
D’Agnolo et al., arXiv:1806.02350
D’Agnolo et al., arXiv:1912.12155

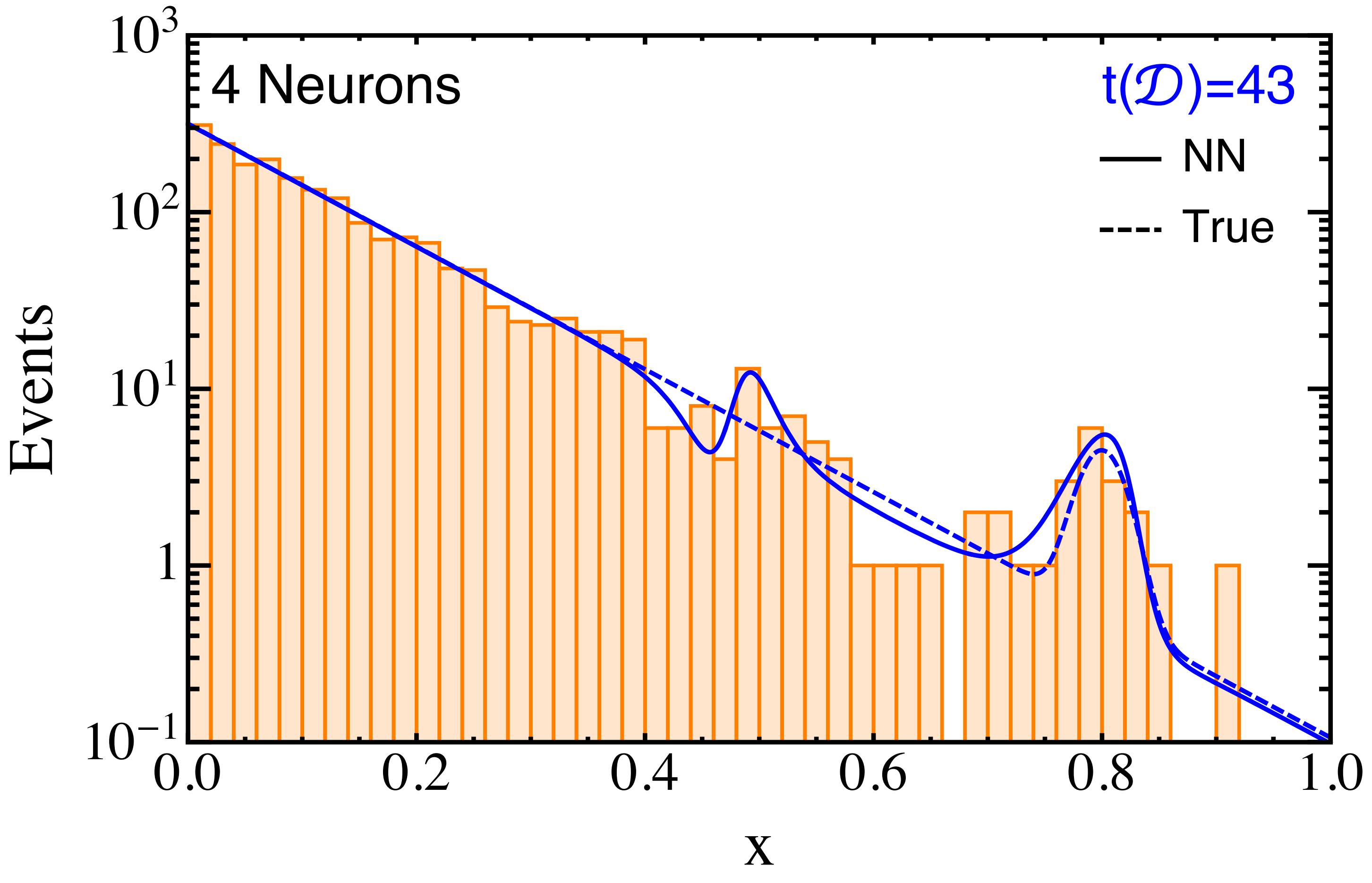
European Research Council

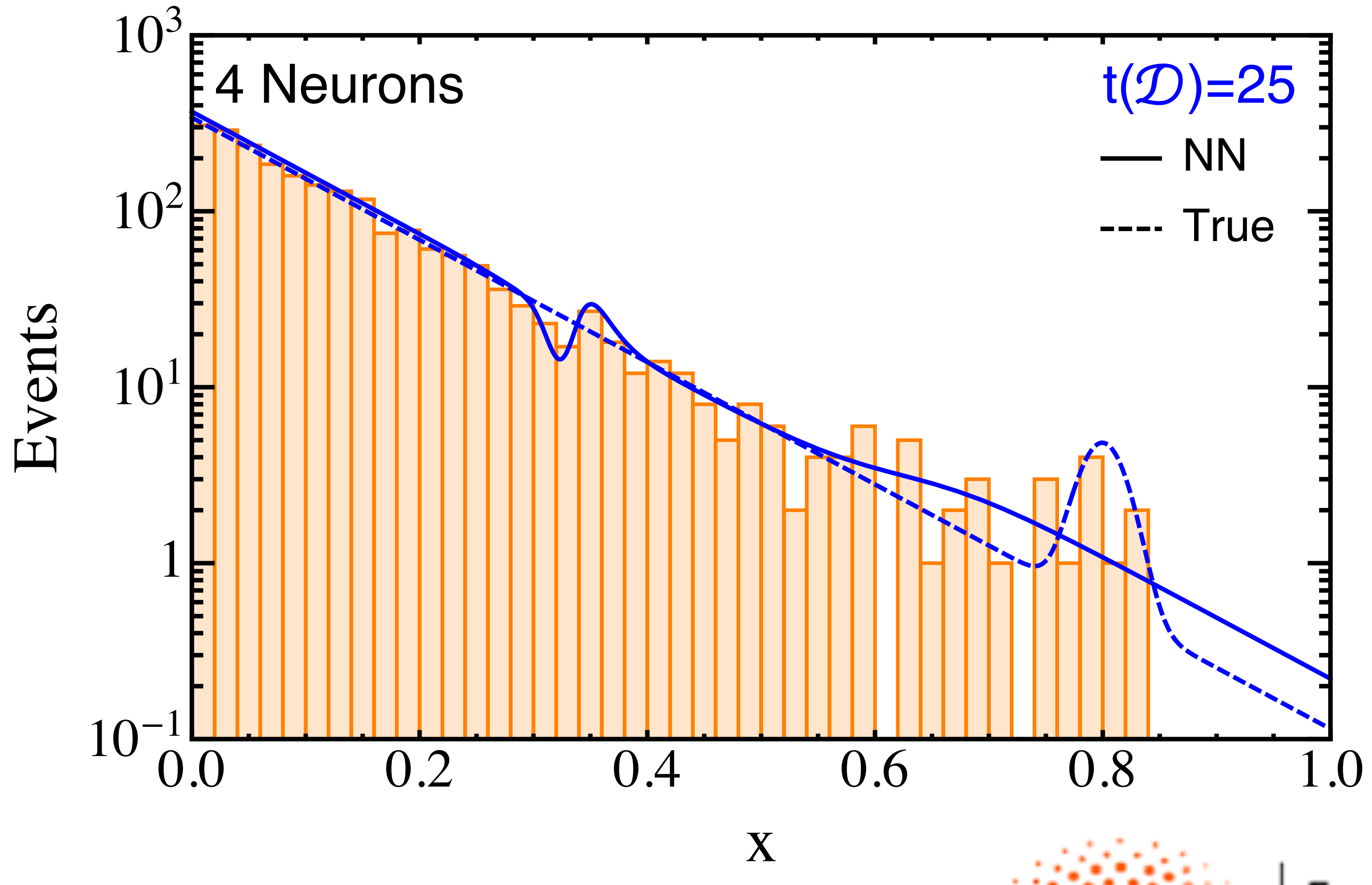
“Model-independent” hypothesis test

- *In 1D, this method can detect new physics presence in D (but not in R)*
  - *performance reduced wrt fully-specified hypothesis test*
  - *still, sensitivity retained*
  - *no explicit assumption on signal shape*





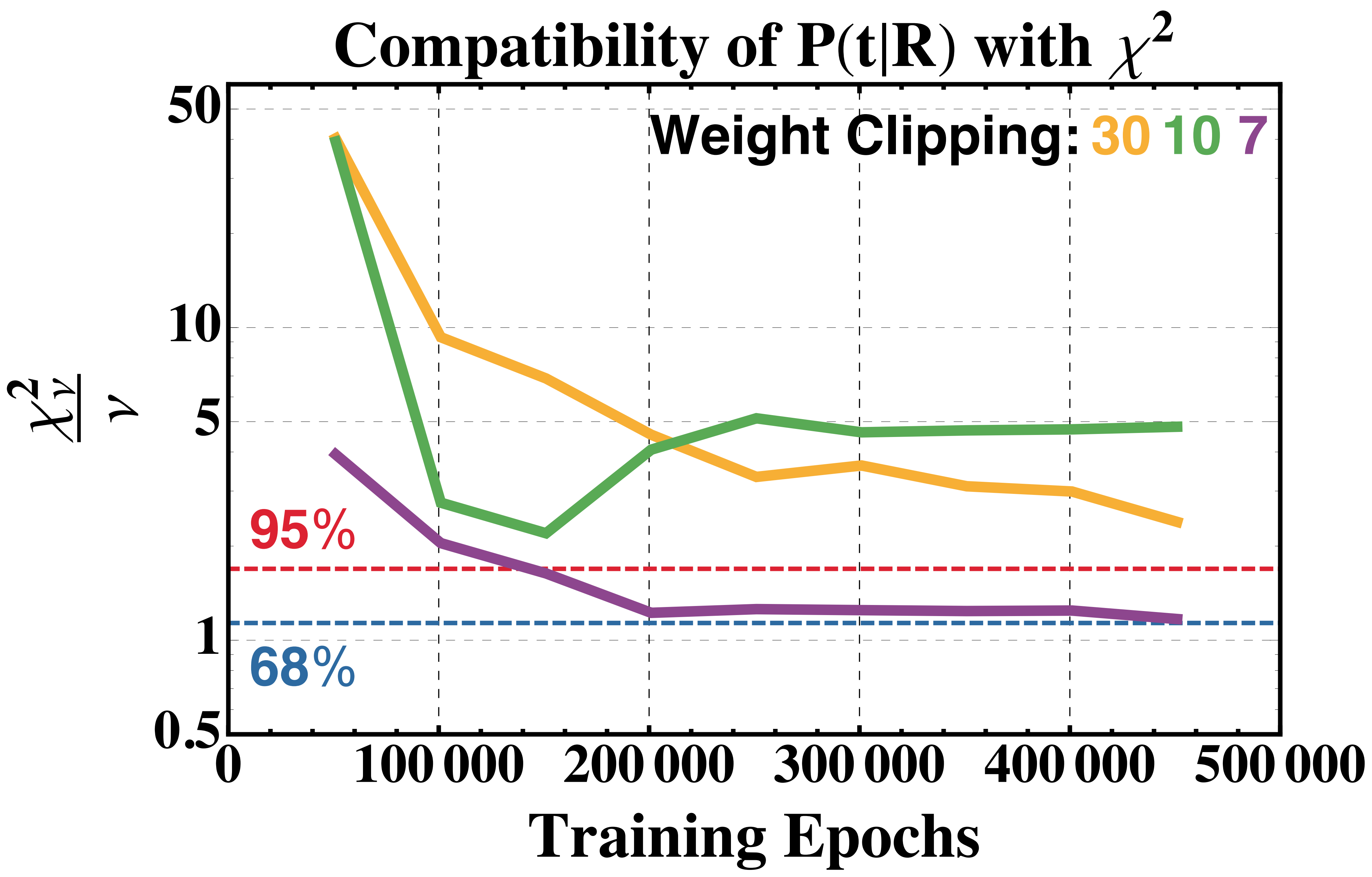


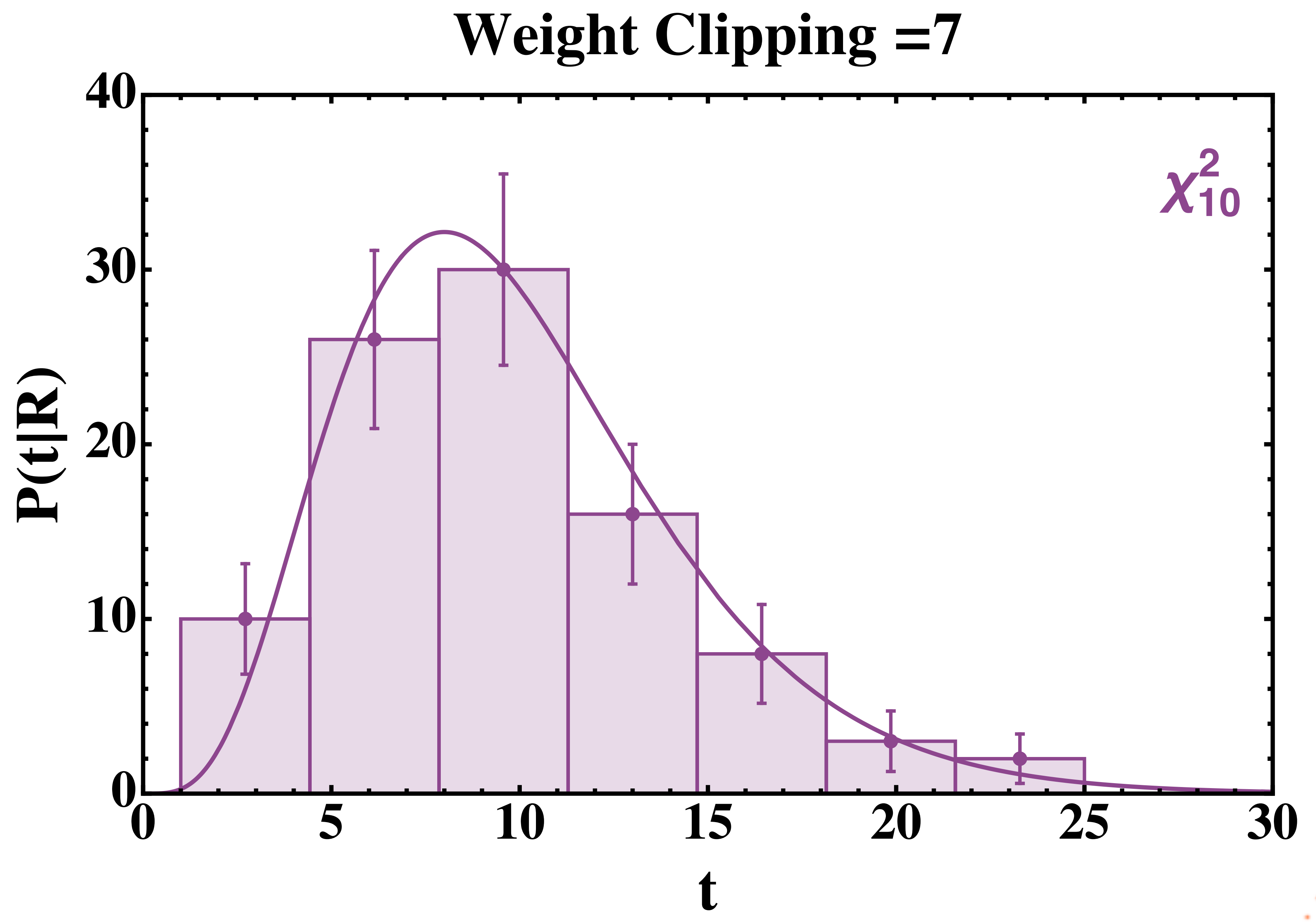


D’Agnolo et al., arXiv:1806.02350
D’Agnolo et al., arXiv:1912.12155

“Model-independent” hypothesis test

- *The N-Dim generalization requires regularisation mechanism*
  - *weight clipping enforced to prevent over-fitting*
  - *with converge, test statistics recovers χ^2 distribution for standard events, with Ndof fixed by number of network parameters*

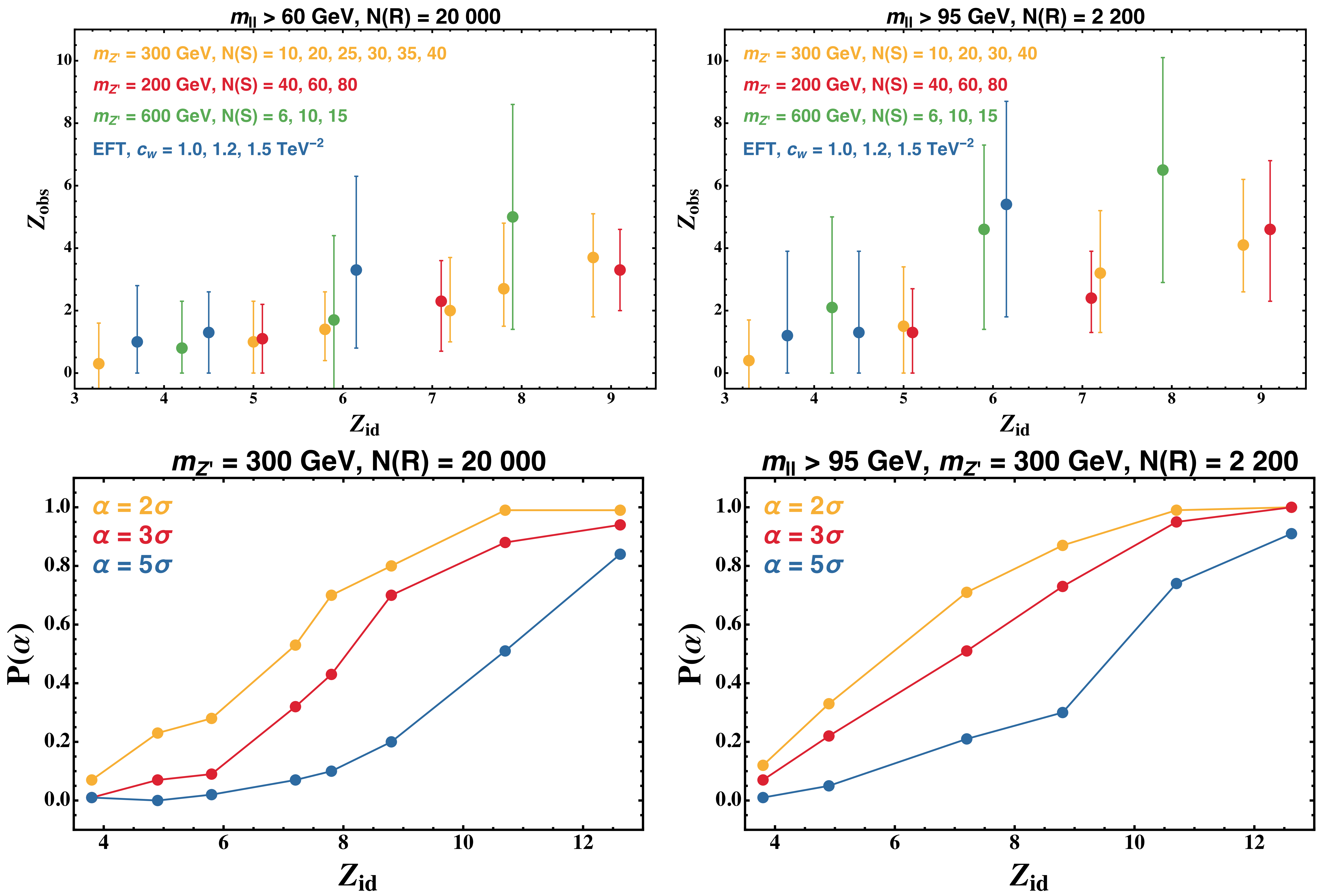




D’Agnolo et al., arXiv:1806.02350

D’Agnolo et al., arXiv:1912.12155

“Model-independent” hypothesis test

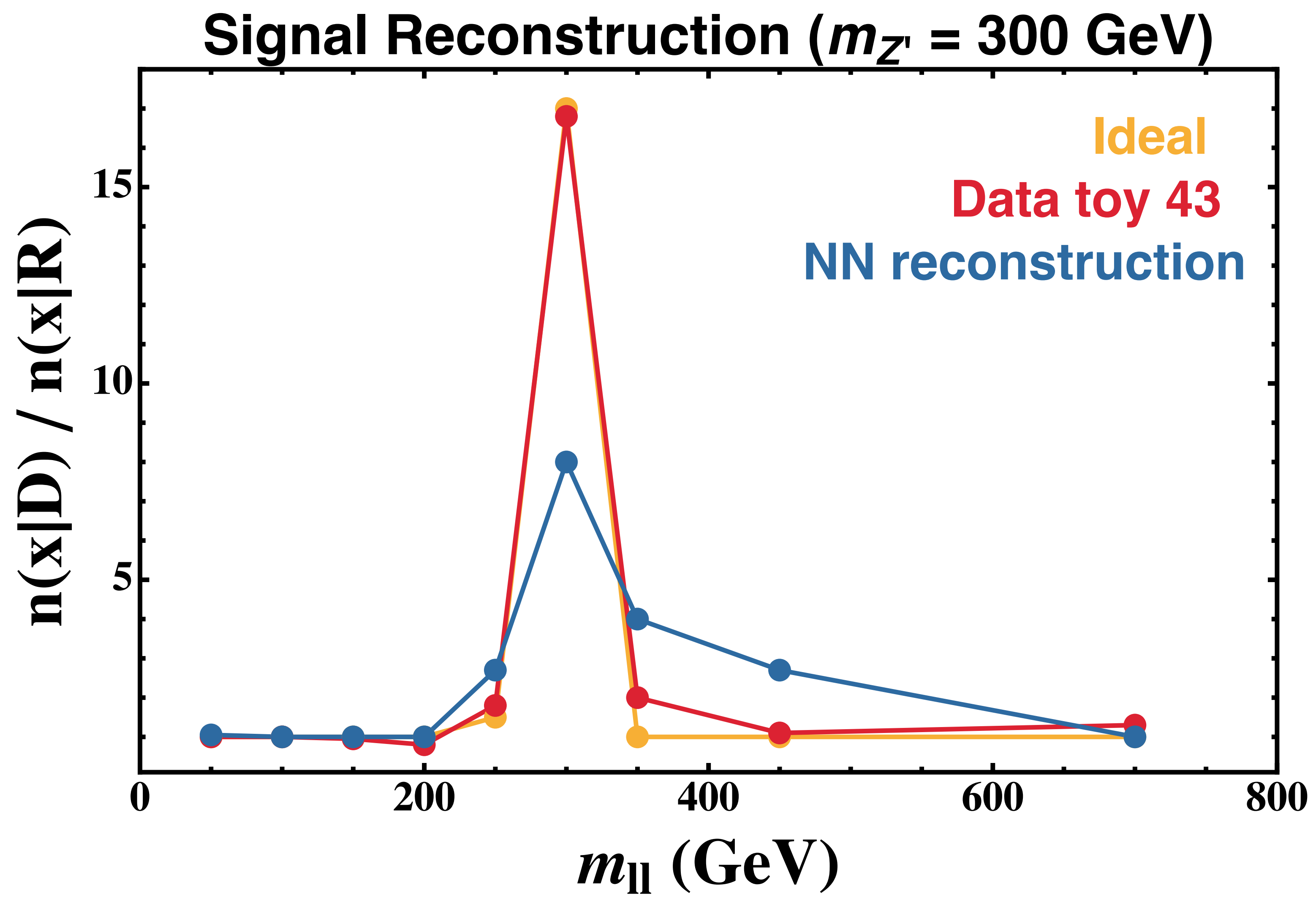


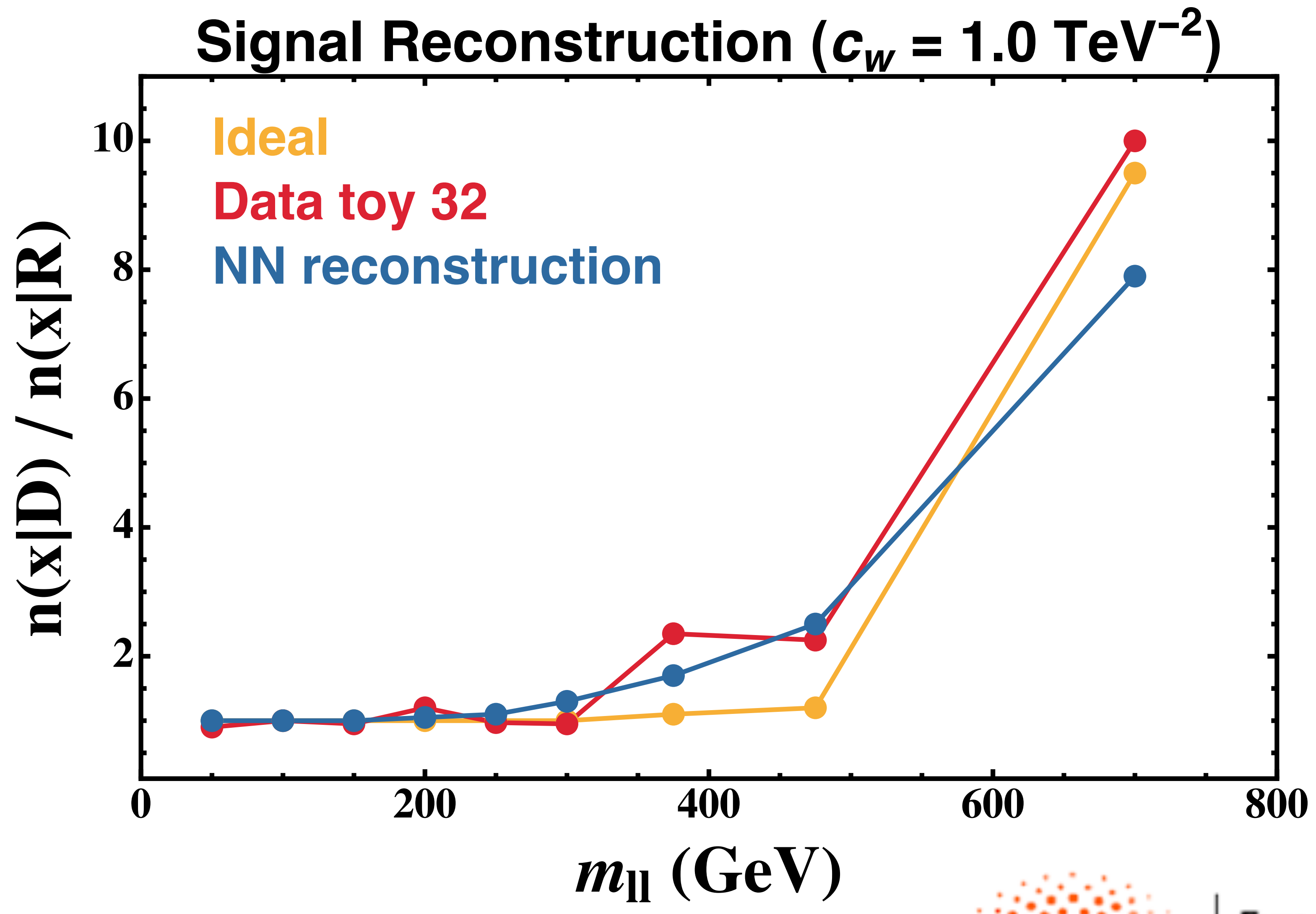
D’Agnolo et al., arXiv:1806.02350

D’Agnolo et al., arXiv:1912.12155

Characterizing the excess

- *A post-training analysis allows to characterize the nature of an excess that might have been found*
- *t(D) vs relevant quantities (not necessarily inputs to training) highlights clustering of signal events*
    - *Invariant mass peak for resonance signal*
    - *Tail excess for EFT signal*
- *The network is learning the nature of the underlying new physics and could guide its characterisation*





D'Agnolo et al., arXiv:1806.02350
D'Agnolo et al., arXiv:1912.12155

Conclusions

- *The LHC is a great discovery machine when you know what to search for*
- *Otherwise, you have to confront the limitations of the LHC big-data problem*
- *Since the SM was established, we followed an established discovery path. We had an easier life, but we have lost the capability of being surprised by data*
- *What we do is great, but we should (re)learn to look at data in a different way: observational particle physics, like astrophysics do*
- *Deep learning will be a crucial ingredient to this. And Run 3 is the right time.*

Backup

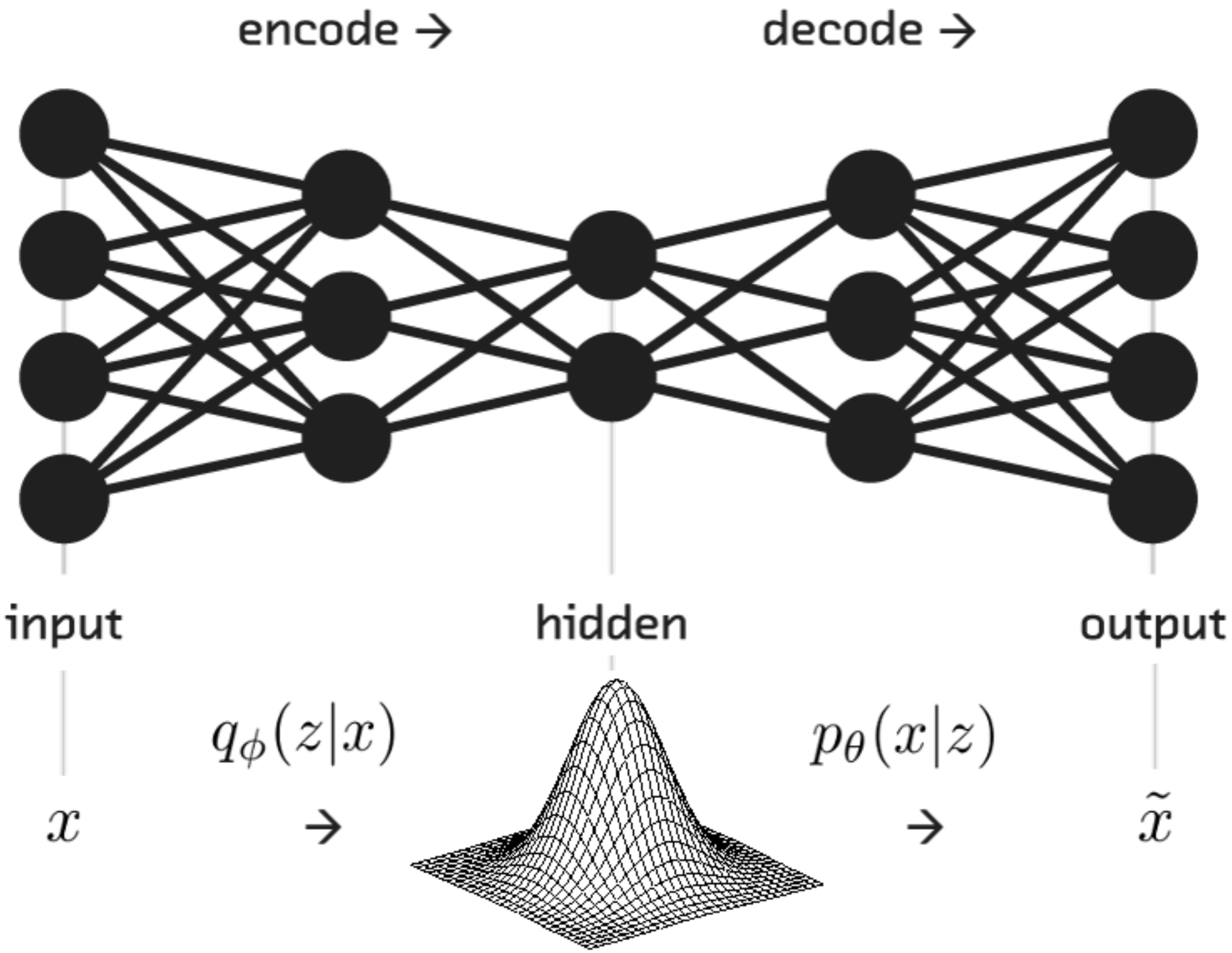
mPP

CERN

European Research Council

Variational Autoencoders

- We investigated variational autoencoders
- Unlike traditional AEs, VAEs try to associate a multi-Dim pdf to a given image
  - can be used to generate new examples
  - comes with a probabilistic description of the input
  - tends to work better than traditional AEs



The Loss Function

- *Loss function described as the sum of two terms (scaled by a tuned λ parameter that makes the two contribution numerically similar)*

  - *Reconstruction loss (e.g. MSE(output-input))*

  - *KL loss: distance between Gaussian pdfs (assumption on prior here)*

  - *Why Gaussian? KL loss can be written analytically*

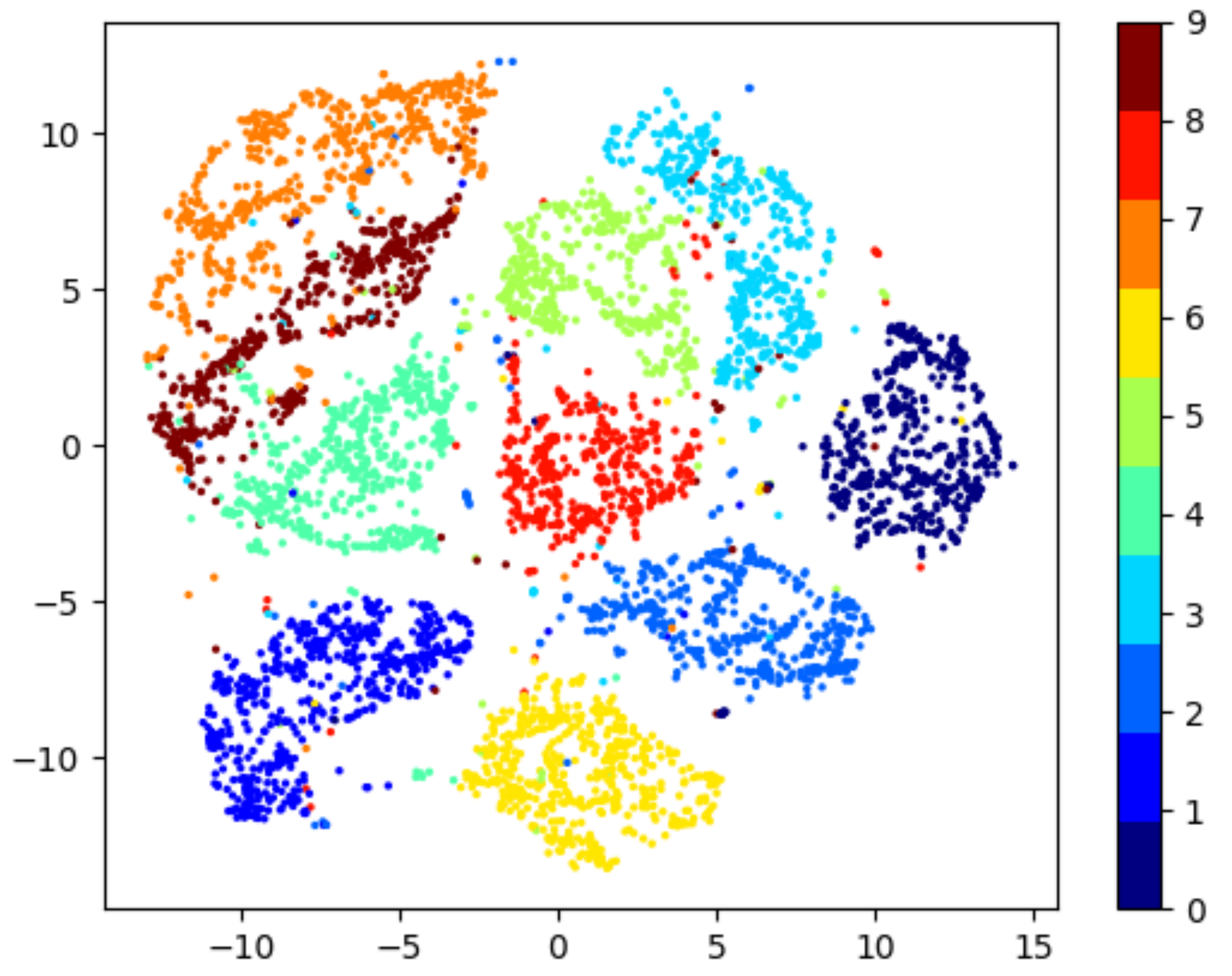
$$\mathrm{Loss}_{\mathrm{Tot}} = \mathrm{Loss}_{\mathrm{reco}} + \beta D_{\mathrm{KL}}$$

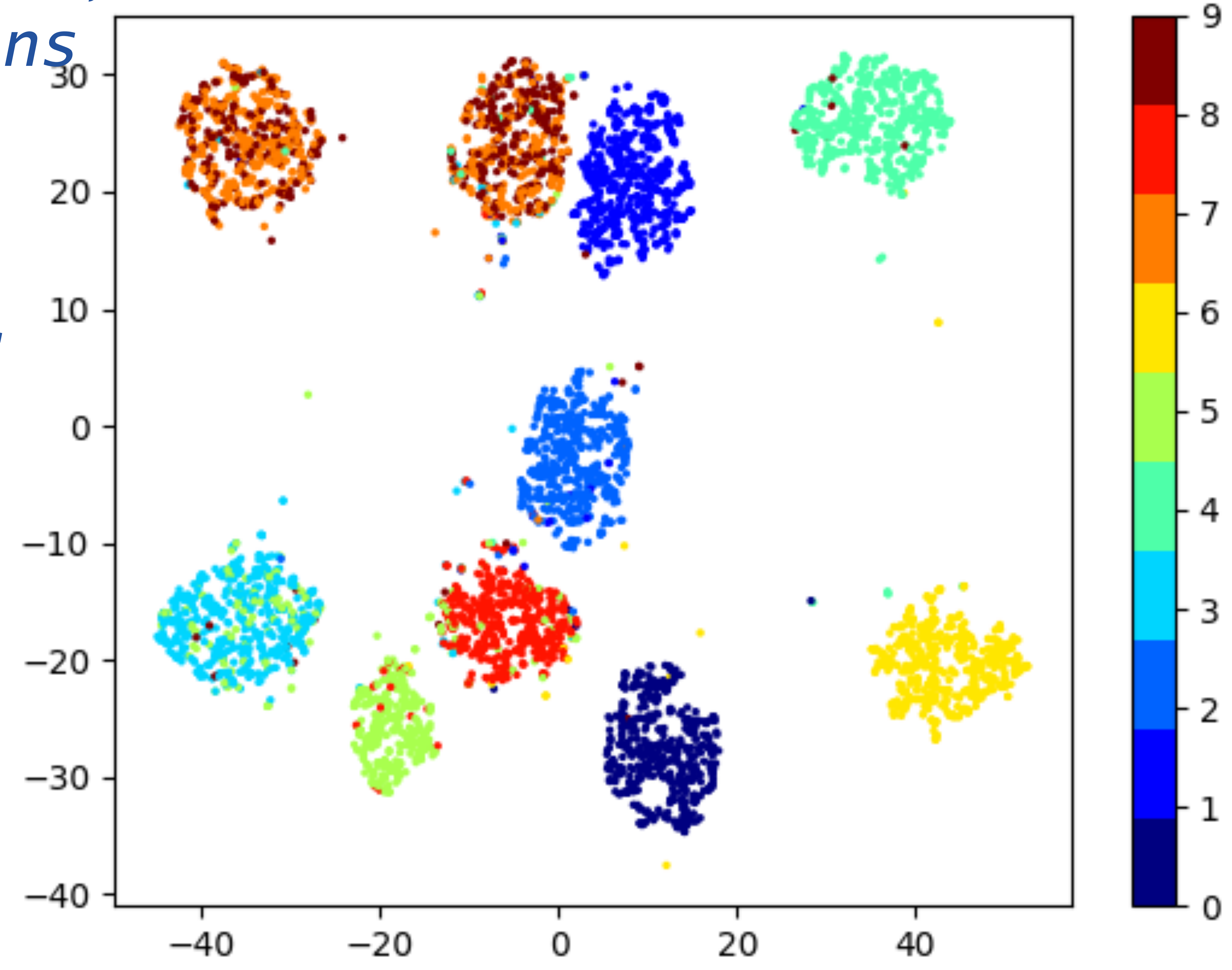
$$D_{\mathrm{KL}} = \frac{1}{k}\sum_i D_{\mathrm{KL}}\left(N(\mu_z^i, \sigma_z^i) \;||\; N(\mu_P, \sigma_P)\right)$$

$$= \frac{1}{2k}\sum_{i,j}\left(\sigma_P^j \sigma_z^{i,j}\right)^2 + \left(\frac{\mu_P^j - \mu_z^{i,j}}{\sigma_P^j}\right)^2 + \ln\frac{\sigma_P^j}{\sigma_z^{i,j}} - 1$$

Clustering with VAE

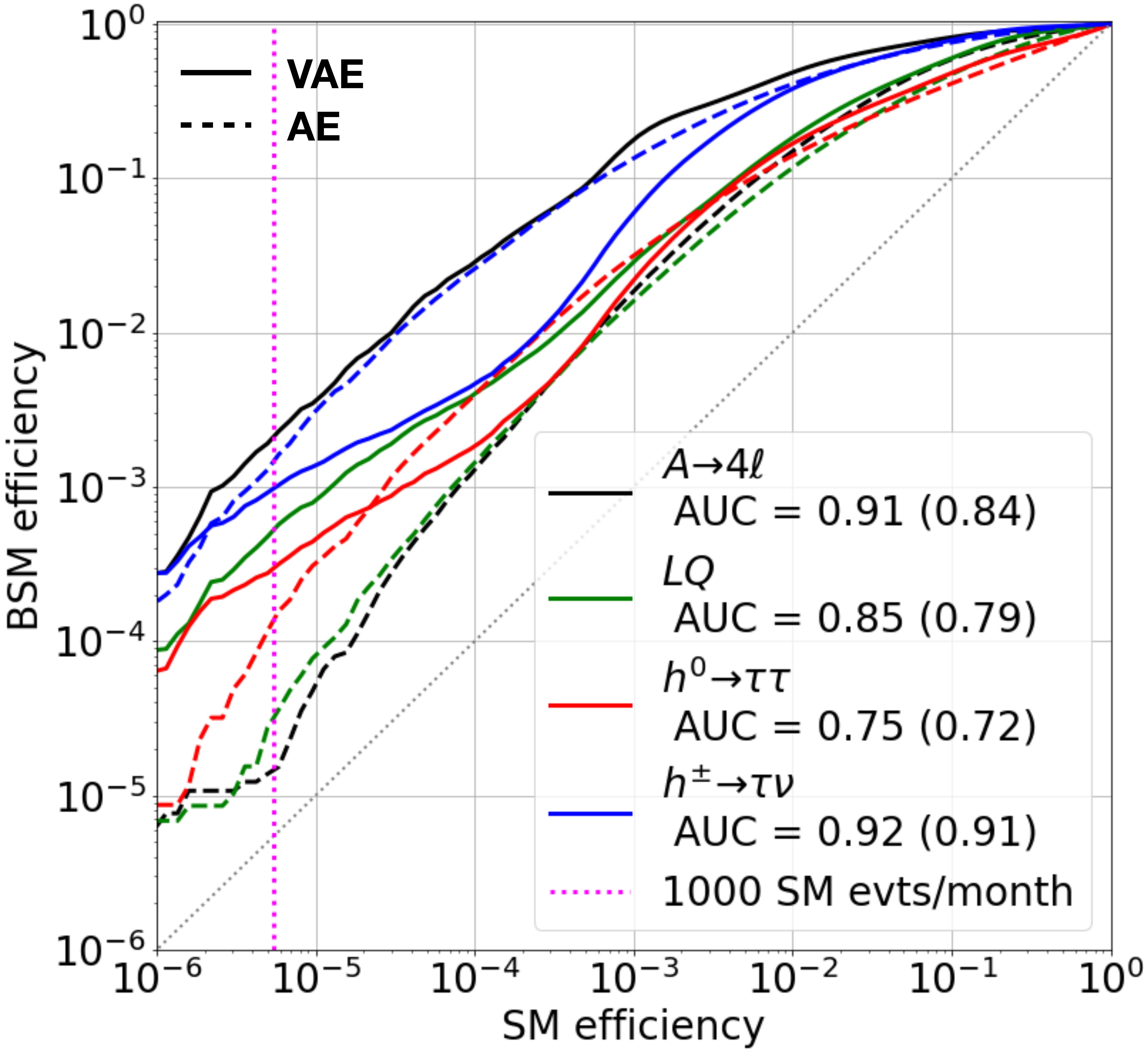
- *In the clustering example, the different populations are forced on sums of Gaussian distributions*
- *This gives more regular shape for the clusters*

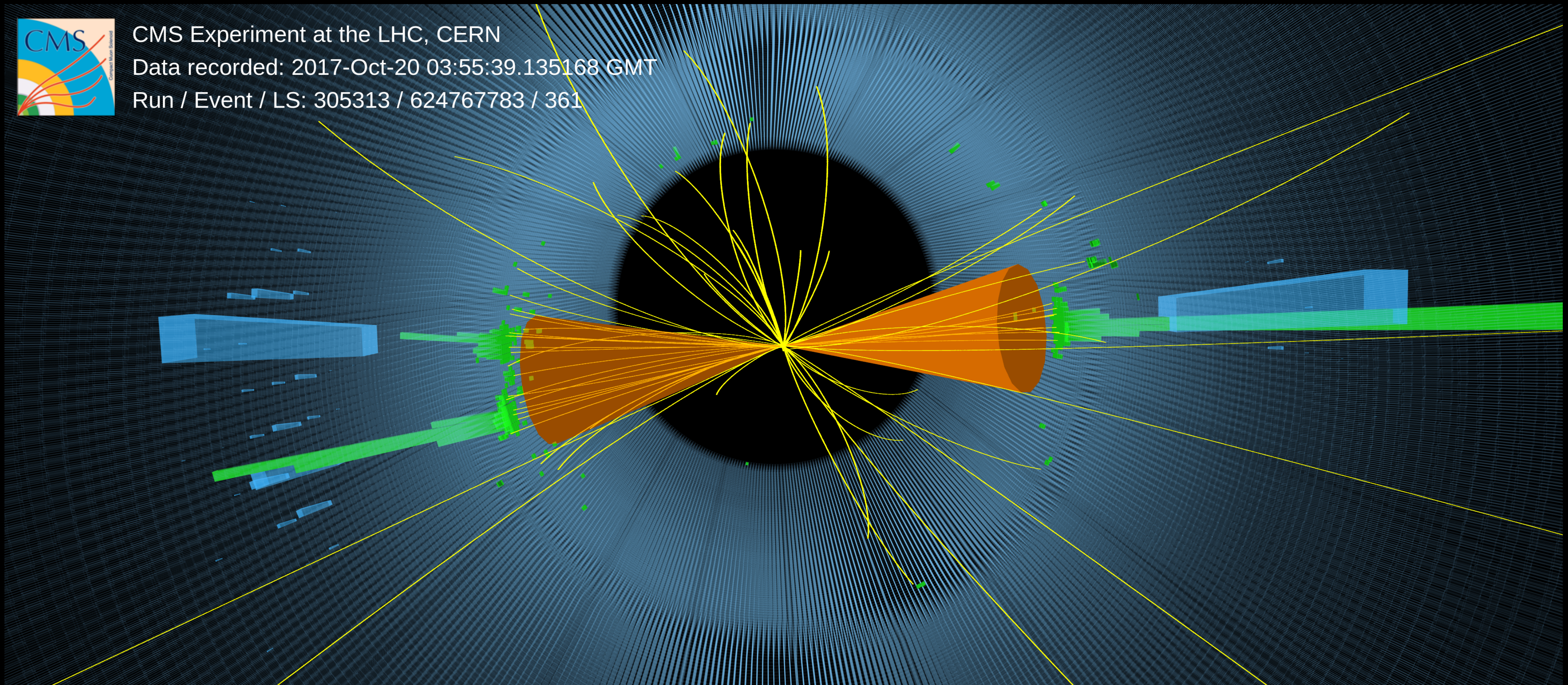




VAEs for anomaly detection

- *Evaluate general discrimination power by ROC curve and area under curve (AUC)*
  - *clearly worse than supervised*
  - *but not so far*
- *Fixing SM acceptance rate at 50 events/day*
  - *competitive results considering unsupervised nature of the algorithm*





PART 2:
NEW IDEAS FOR IMPLEMENTATION OF ANOMALY DETECTION ALGORITHMS AT THE LHC

From dijet resonance searches
to deployment online in the experiments trigger

Maurizio Pierini (CERN),
Jennifer Ngadiuba (FNAL/Caltech)

New Physics from Precision at High Energies
29 April 2021, KITP

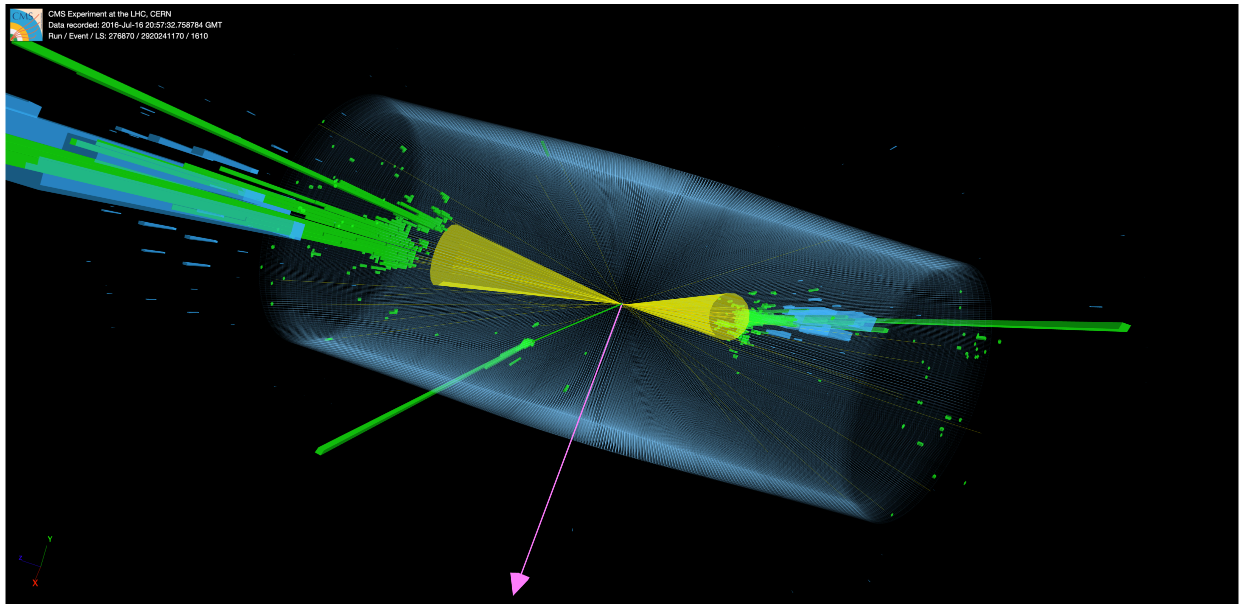
The physics case: dijet resonances

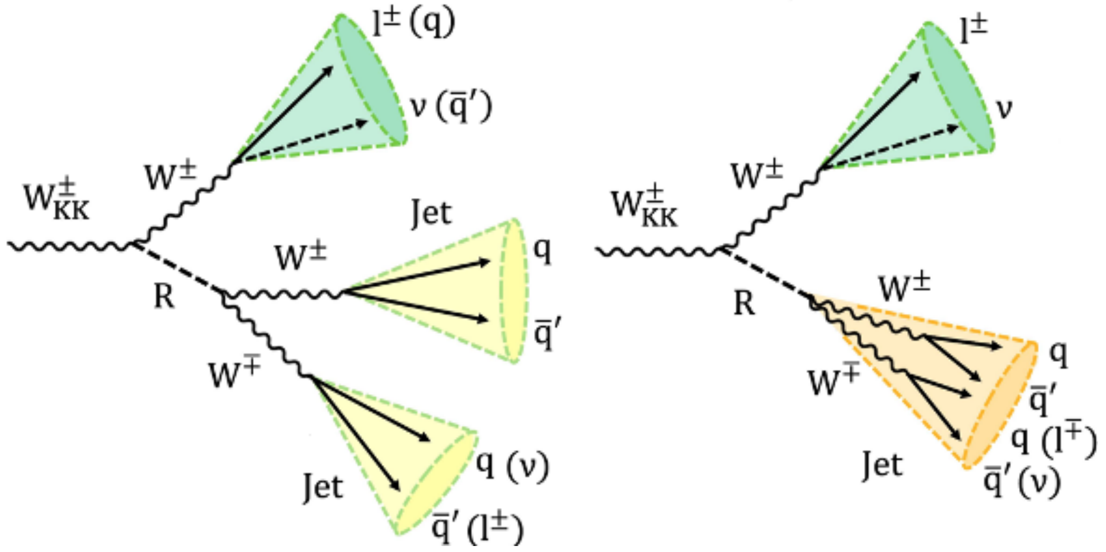
- Extensively studied at colliders
  - classic dijet w/ no jet tagging
  - $t\bar{t}$ w/ dedicated top tagging
  - diboson w/ dedicated SM boson jet tagging
  - **most recently: triboson!**
  - ...
- Many other possible BSM scenarios not covered by these searches
- Or there could be a BSM signal we never thought of

→ how to generalize?

Search for resonances decaying to triple W-boson final states in proton-proton collisions at $\sqrt{s}$ = 13 TeV

CMS-PAS-B2G-20-001

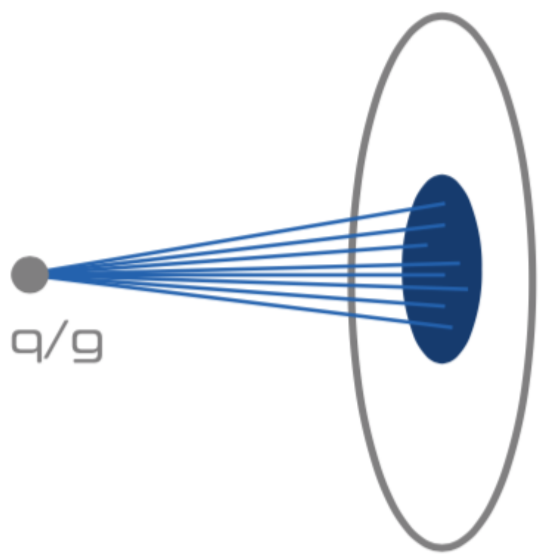


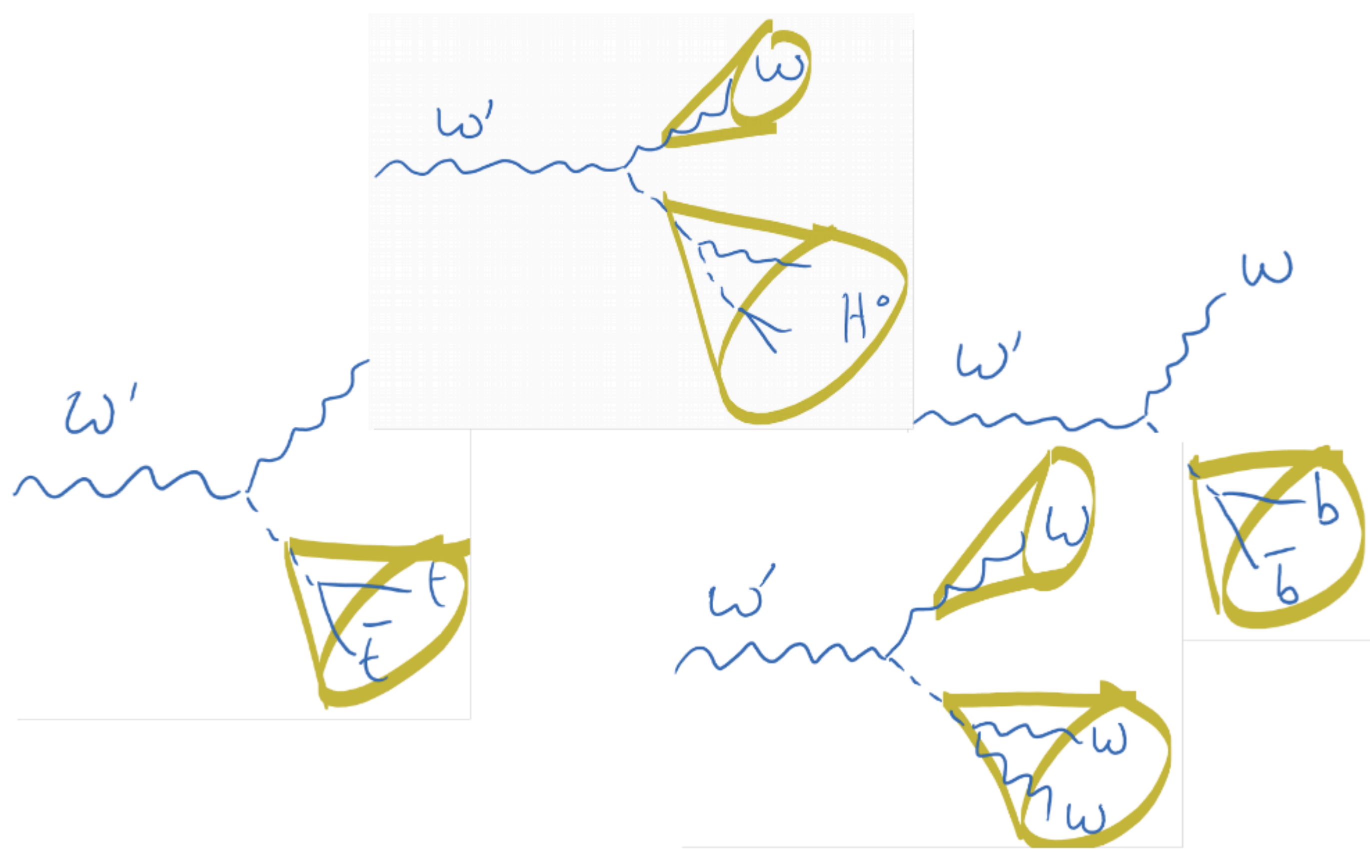


Building a QCD-jet veto

- Dijet searches overwhelmed by QCD multijet background
- How to be sensitive to an unknown and low-coupling BSM signal → **veto QCD jets**
- Novel signal-agnostic approaches uses **anomaly detection algorithms**

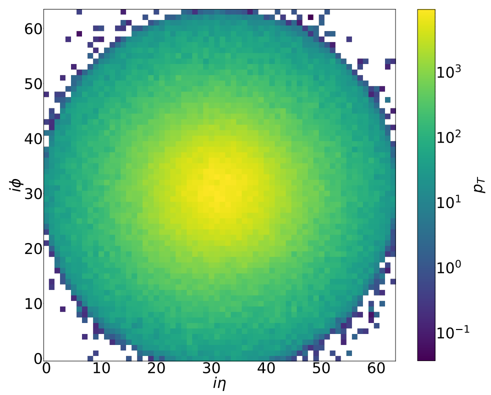
BACKGROUND QCD JET VS ANY OF THESE

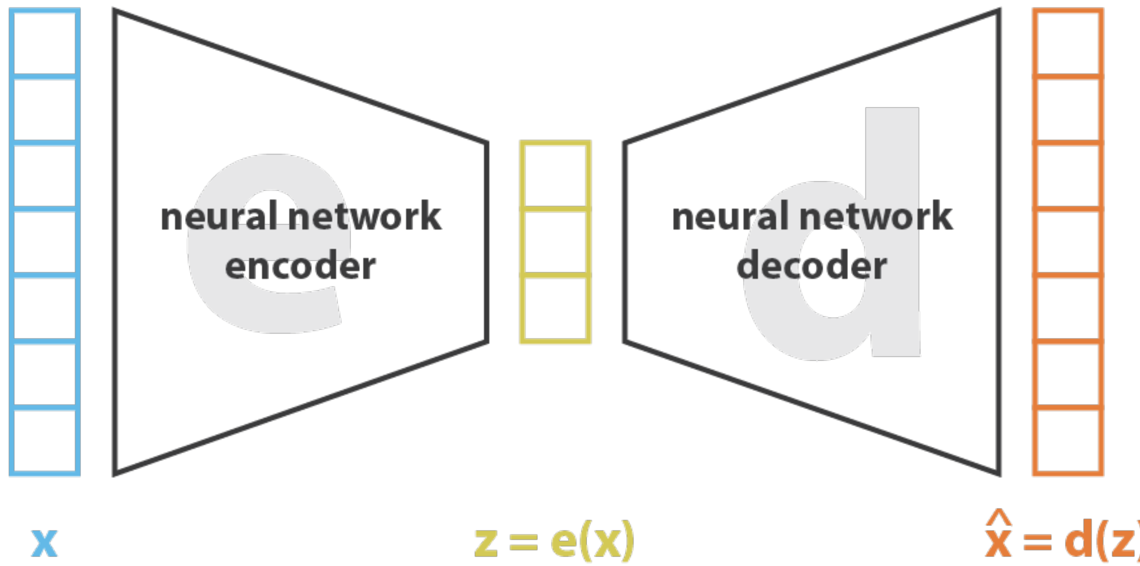




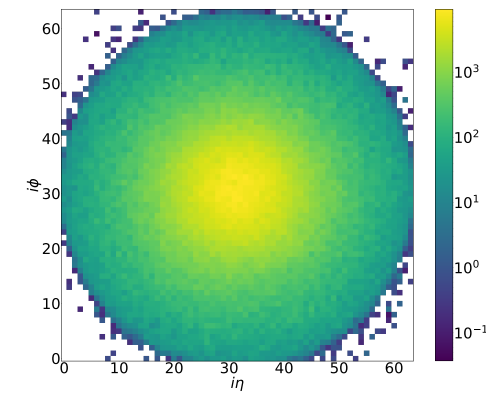
Autoencoders for jets

e.g, jet images



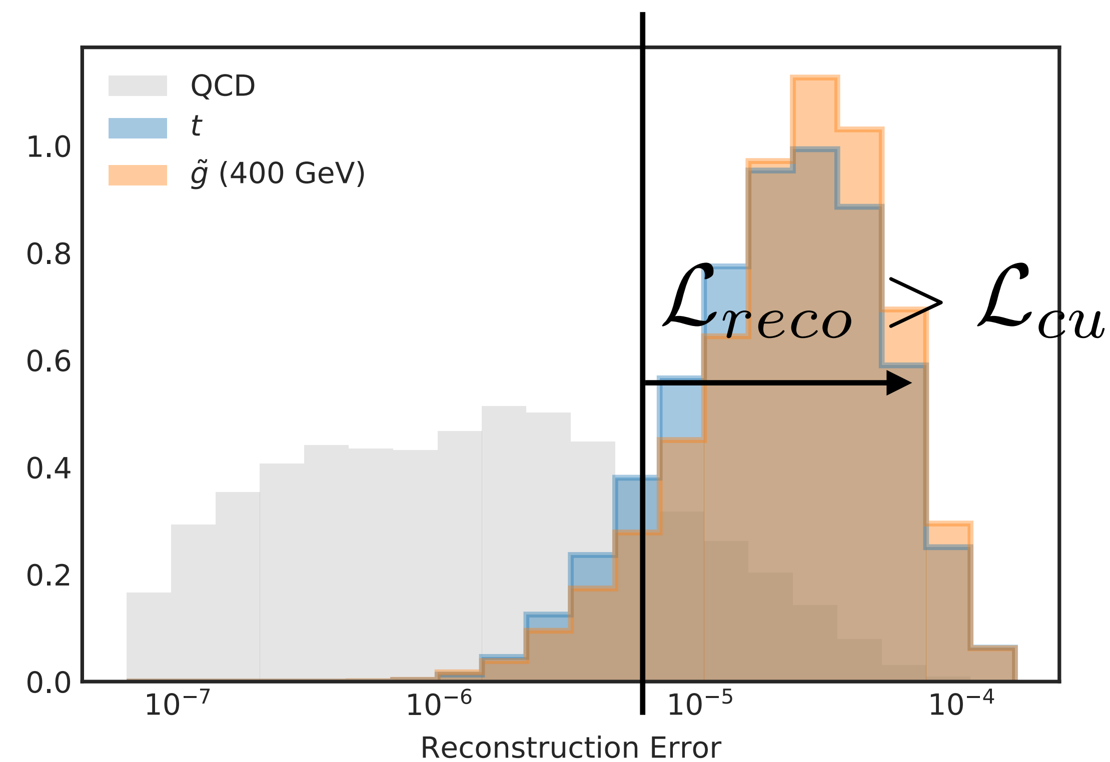


e.g, jet images



$$\mathcal{L}_{reco} = ||x - \hat{x}||^2 = MSE(input, output)$$

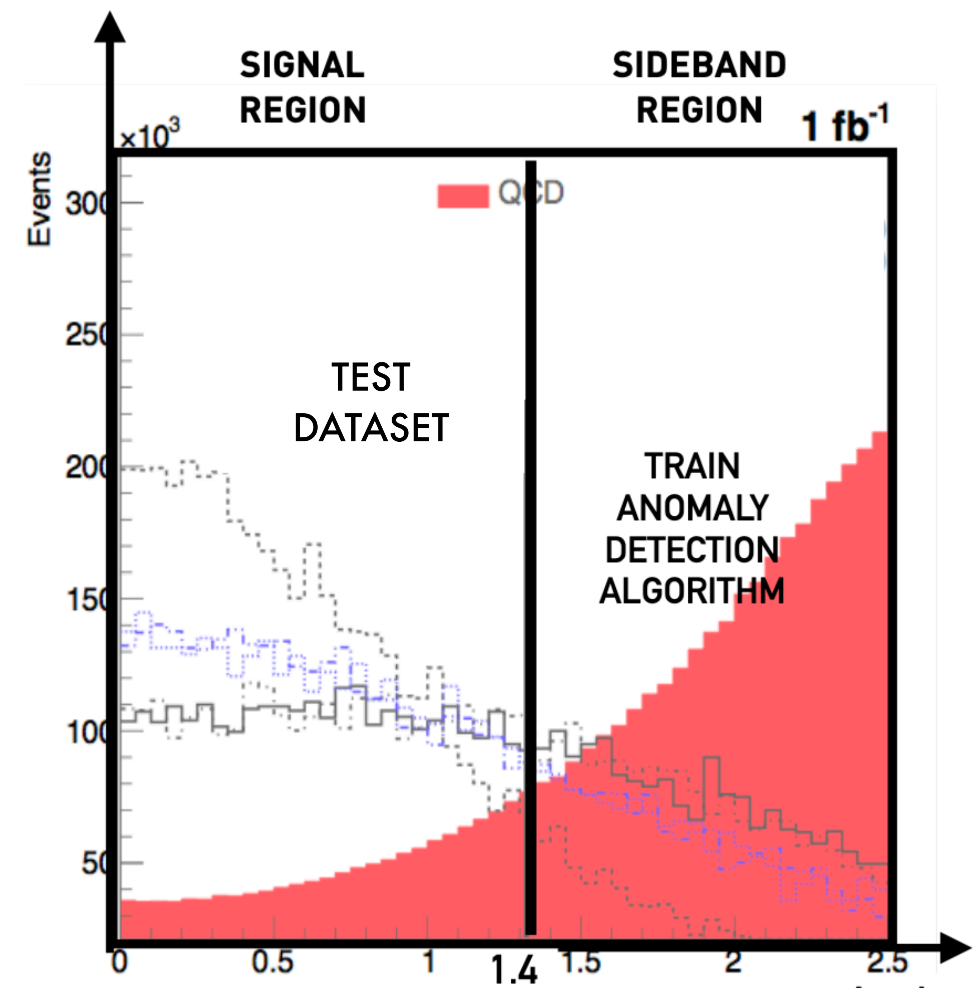
- Recent idea to use autoencoders for jet tagging, in order to define a QCD-jet veto [*]
- Based on jet images but other physics-inspired representations can be used
- Applied in a BSM search (e.g., dijet resonance) could highlight new physics signal



[*] Heimel et al.: SciPost Phys. 6, 030 (2019) , Farina et al.: Phys. Rev. D 101, 075021 (2020)

Apply it to the dijet search

- Train a jet autoencoder on each jet individually in observed dijet data
  - choose sample enriched in QCD multijet background: high $|\Delta\eta_{jj}|$ region
- Define an anomaly score:
  - loss function as obvious choice
  - evaluate on test dataset where a possible signal could live: low $|\Delta\eta_{jj}|$ region
- Go **from anomalous jets to anomalous dijet events** combining the two individual jet losses



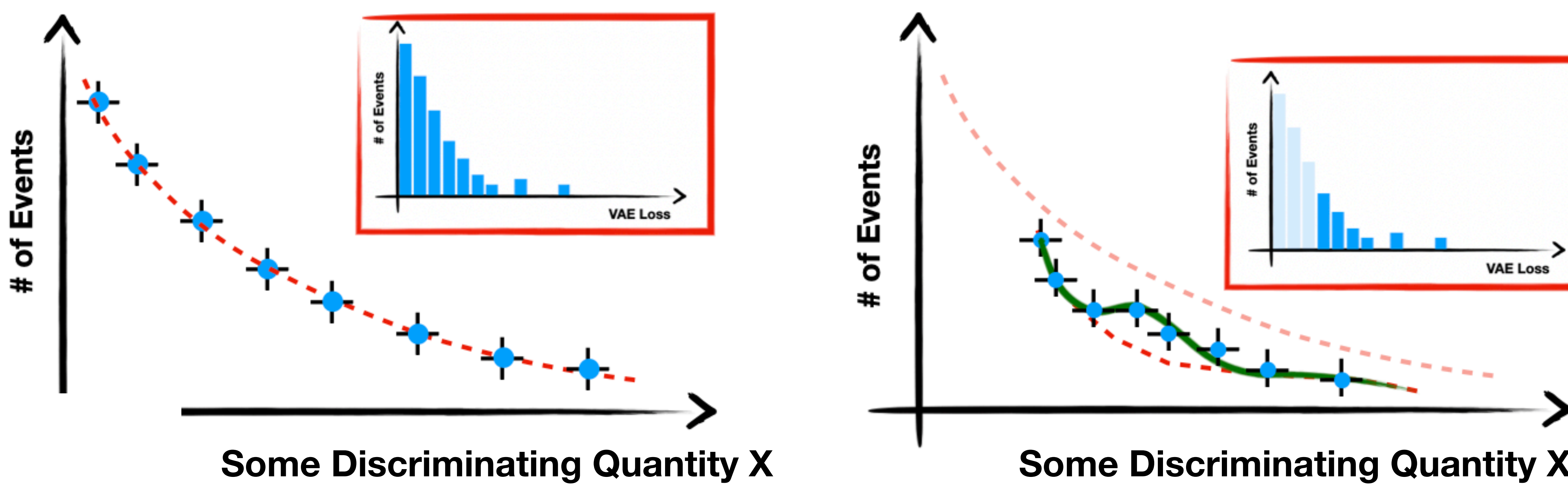
trained model

loss per jet

combined 2-jet loss

Apply it to the dijet search

- Doing so, one wants to avoid deformations in the background distribution that could fake a signal and/or disrupt the background estimation
  - bump hunt in $X=m_{jj}$ for dijet resonance search case

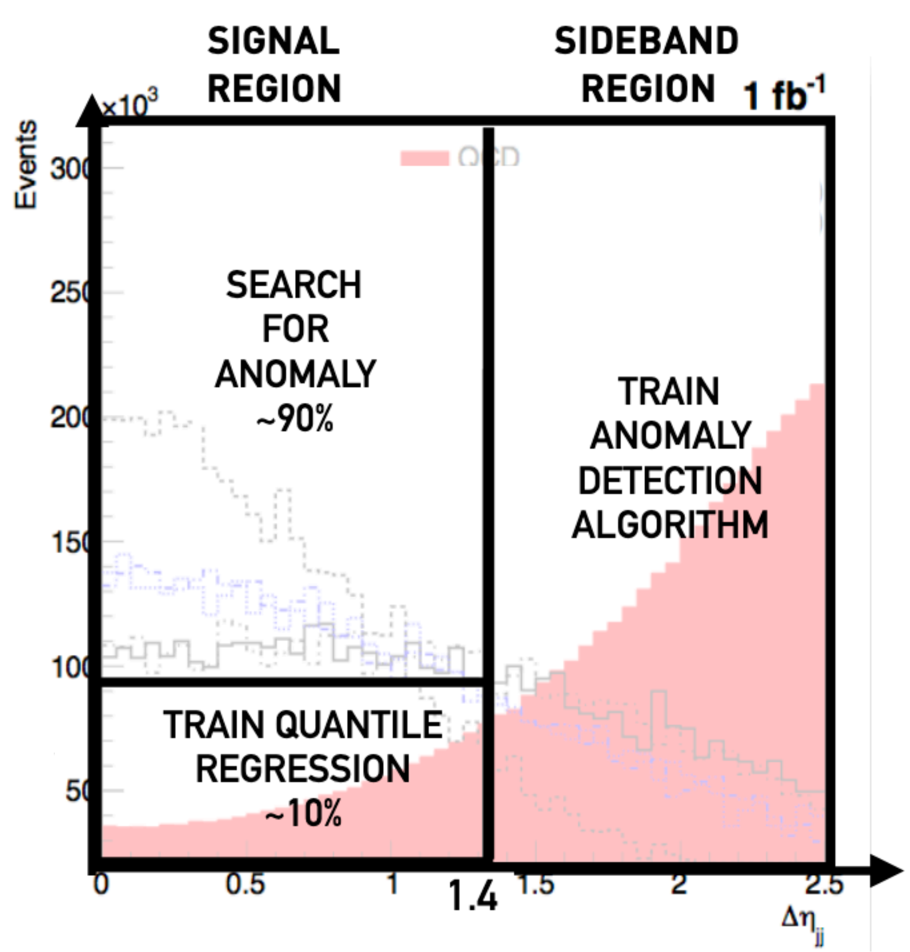


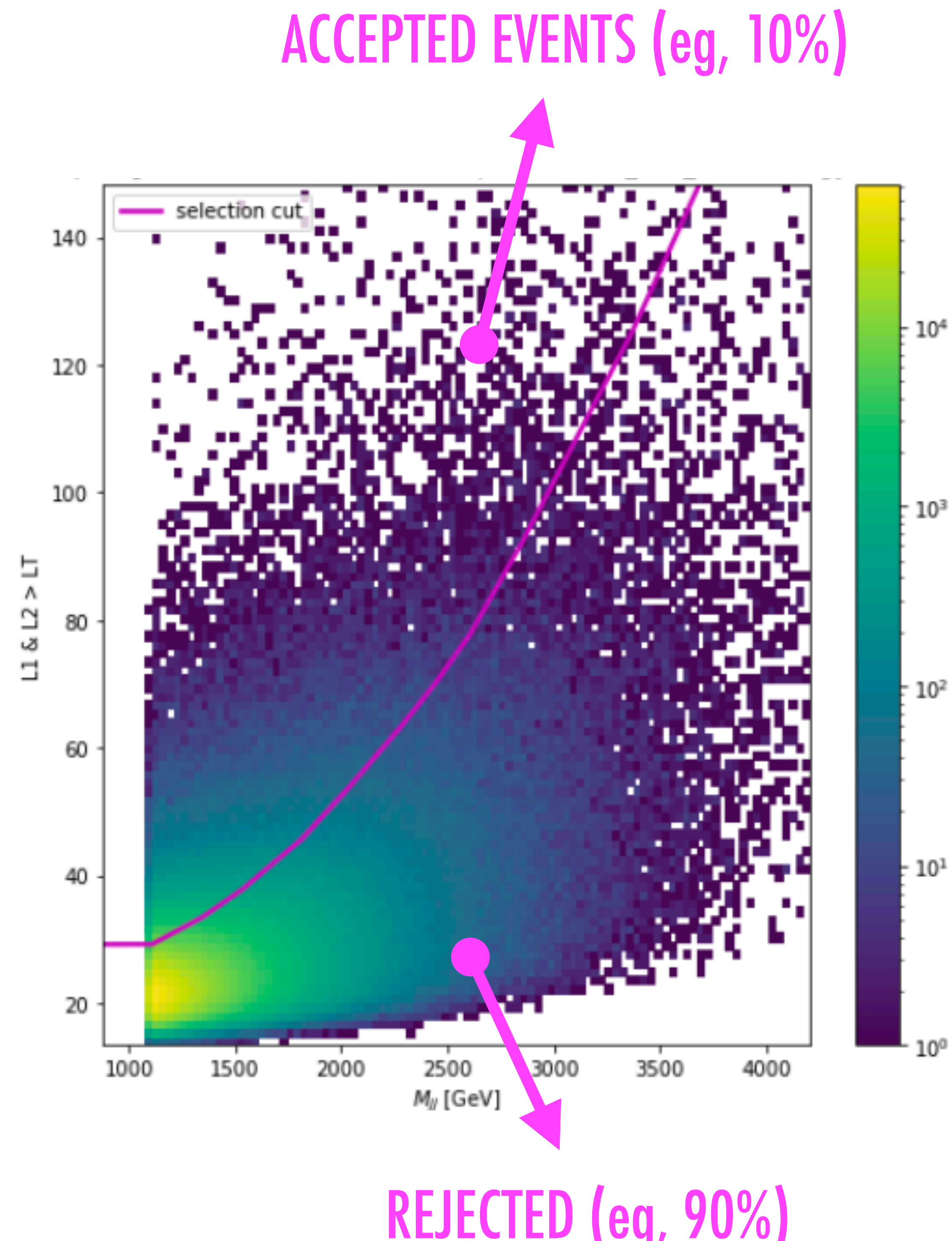
Apply it to the dijet search

- Use a quantile regression to obtain a X-dependent cut on the loss

$$\mathcal{L}_{reco}(X_i) > \mathcal{L}_{cut}(X_i)$$

  - chosen quantile value driven by the target background rejection rate
  - compute on a F fraction of the signal region data or use cross-training procedure





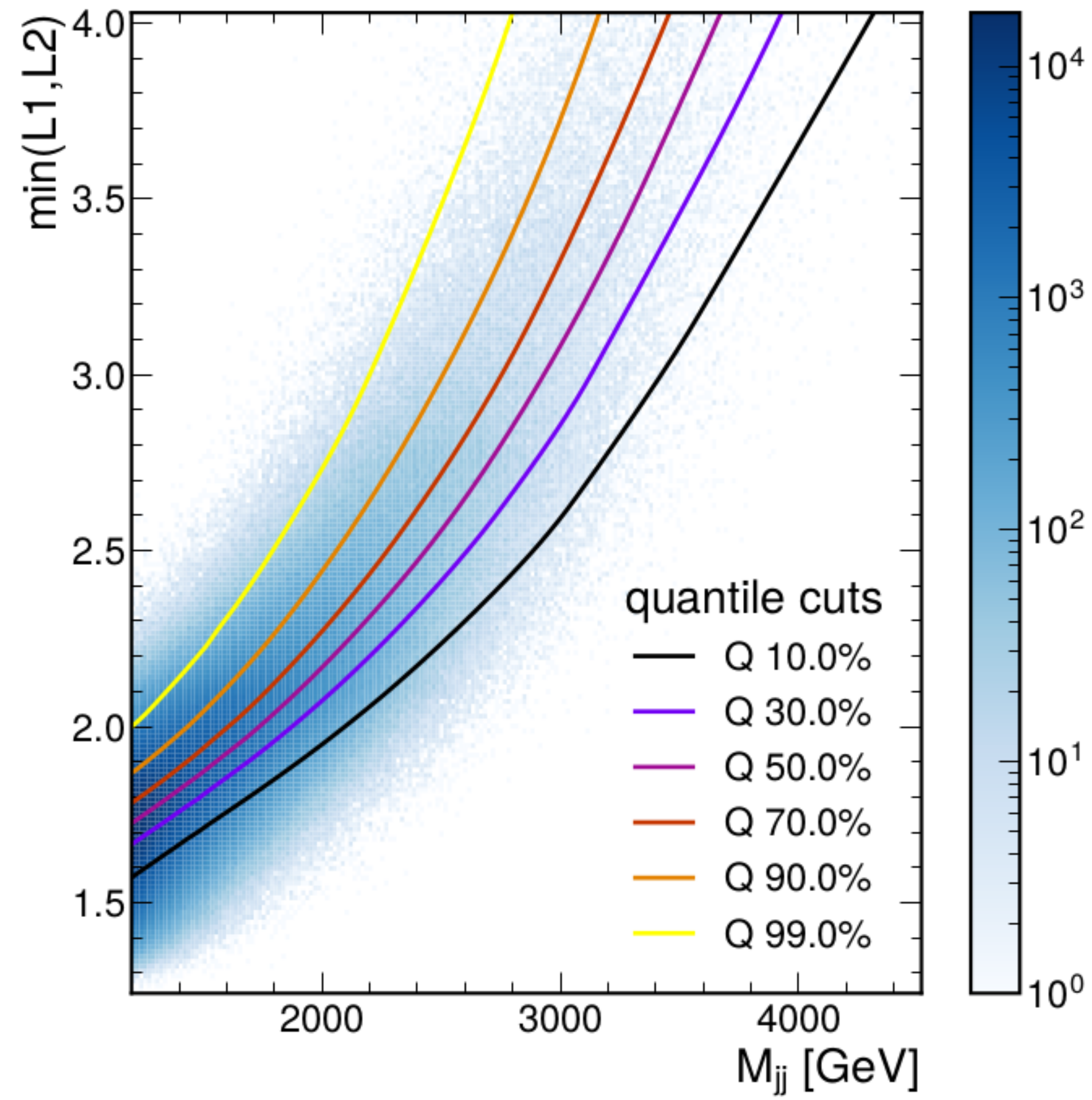
Apply it to the dijet search

- Use a quantile regression to obtain a X-dependent cut on the loss

$$\mathcal{L}_{reco}(X_i) > \mathcal{L}_{cut}(X_i)$$

  - chosen quantile value driven by the target background rejection rate
  - compute on a F fraction of the signal region data or use cross-training procedure

- Bin the sample in orthogonal quantile ranges
- Each bin with different signal vs background rates



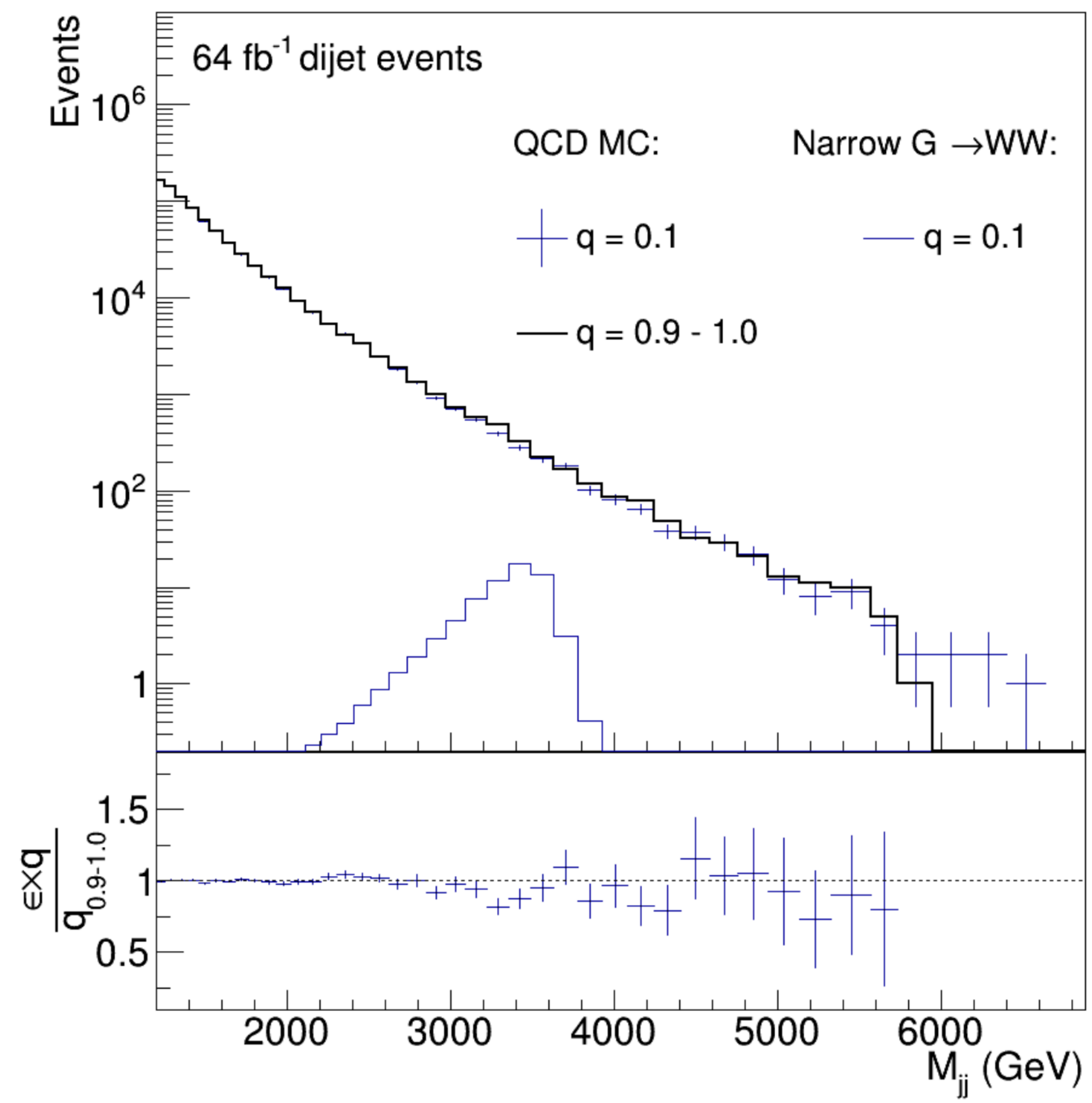
Apply it to the dijet search

- Use a quantile regression to obtain a X-dependent cut on the loss

$$\mathcal{L}_{reco}(X_i) > \mathcal{L}_{cut}(X_i)$$

  - chosen quantile value driven by the target background rejection rate
  - compute on a F fraction of the signal region data or use cross-training procedure

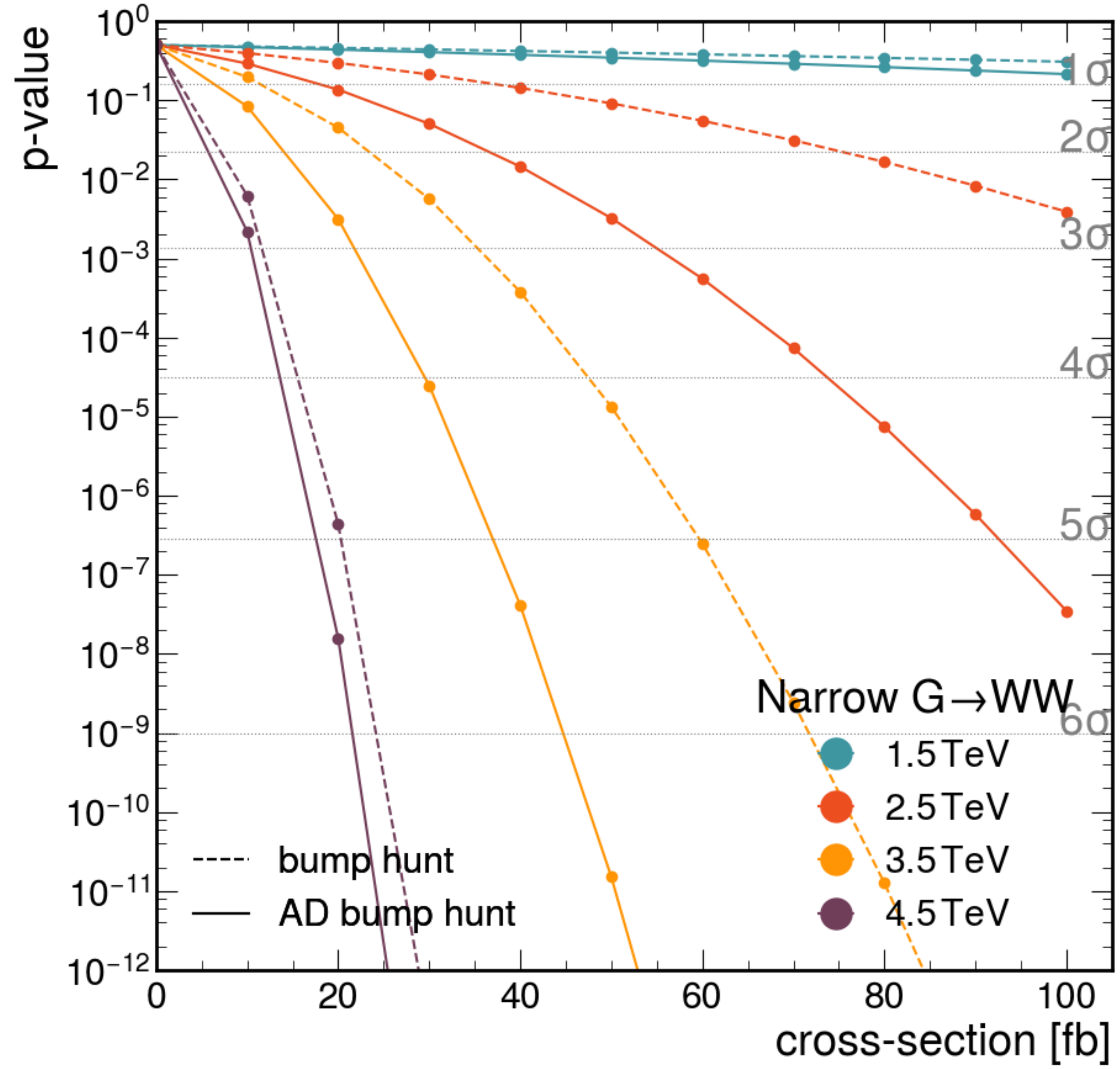
- Bin the sample in orthogonal quantile ranges
- Each bin with different signal vs background rates
- By construction and in absence of signal, background shape is the same in all quantile bins



Boosting sensitivity of dijet searches

- Method performance evaluated for a traditional signal
  - heavy resonance decaying to WW
  - narrow (1% width) and broad (35% width)
- Implement traditional bump hunt in dijet invariant mass spectrum
- Inject signal of increasing cross-section in QR training and observed dataset and compare p-values for:
  - fit to the inclusive dijet spectrum
  - simultaneous fit to all loss quantiles bins

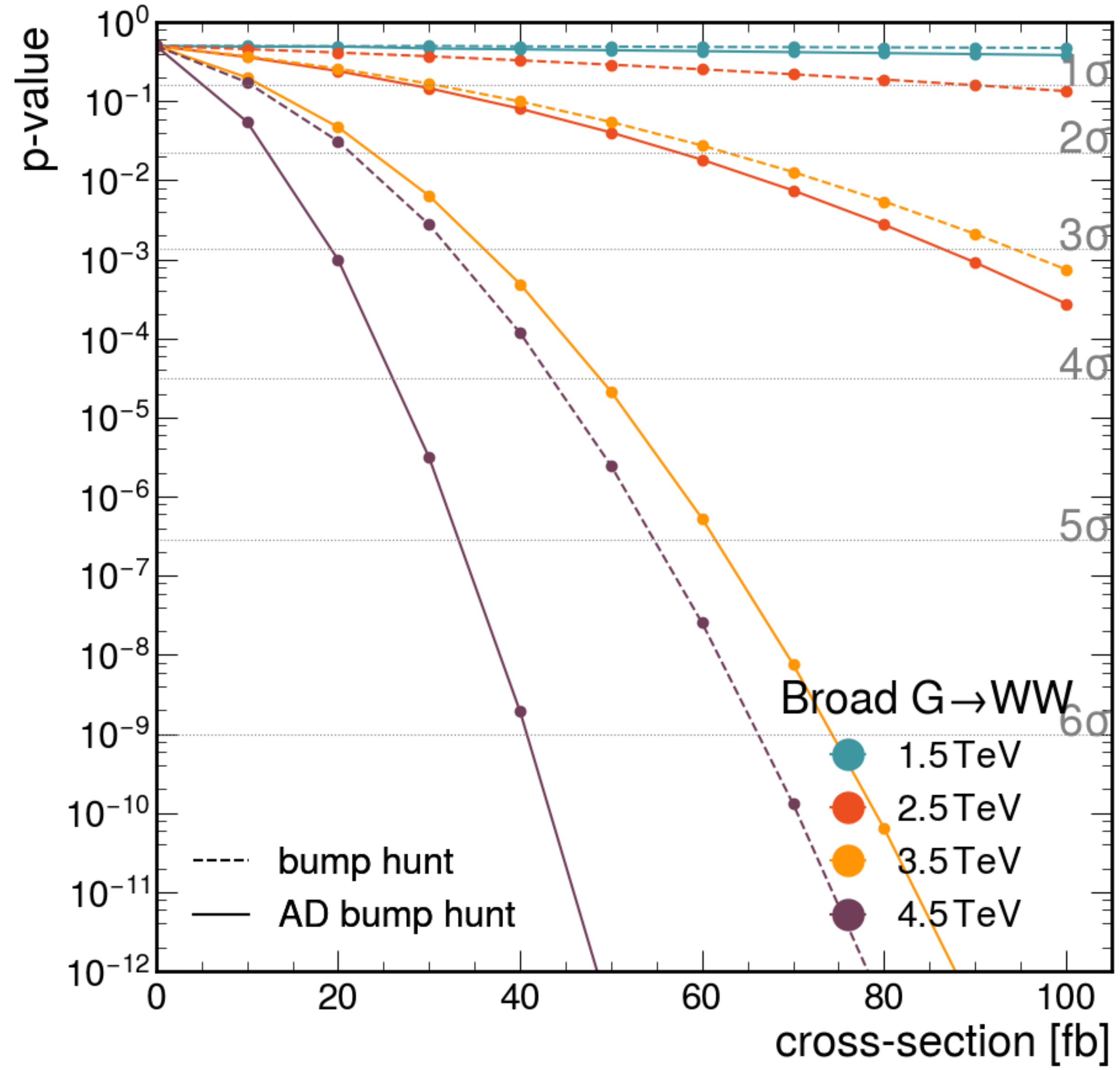
NARROW G→WW



Boosting sensitivity of dijet searches

- Method performance evaluated for a traditional signal
  - heavy resonance decaying to WW
  - narrow (1% width) and broad (35% width)
- Implement traditional bump hunt in dijet invariant mass spectrum
- Inject signal of increasing cross-section in QR training and observed dataset and compare p-values for:
  - fit to the inclusive dijet spectrum
  - simultaneous fit to all AE loss quantiles bins

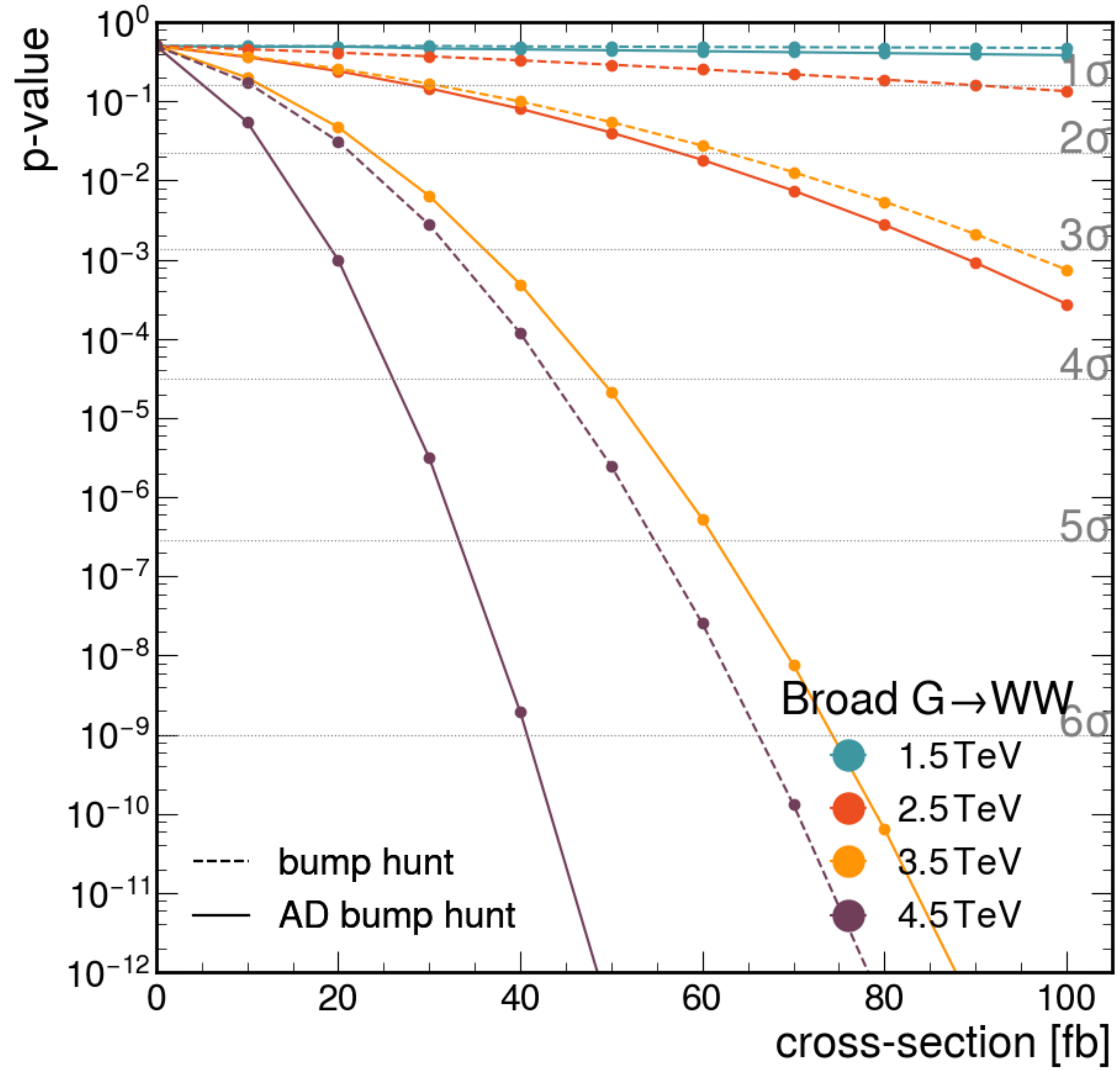
BROAD G→WW



Boosting sensitivity of dijet searches

- Method performance evaluated for a traditional signal
  - heavy resonance decaying to WW
  - narrow (1% width) and broad (35% width)
- Implement traditional bump hunt in dijet invariant mass spectrum
- Inject signal of increasing cross-section in QR training and observed dataset and compare p-values for:
  - fit to the inclusive dijet spectrum
  - simultaneous fit to all AE loss quantiles bins

BROAD G→WW



The same idea can be applied to any final states with N>=1 jets and for any discriminating variable X!

More boost: the 3D bump hunt

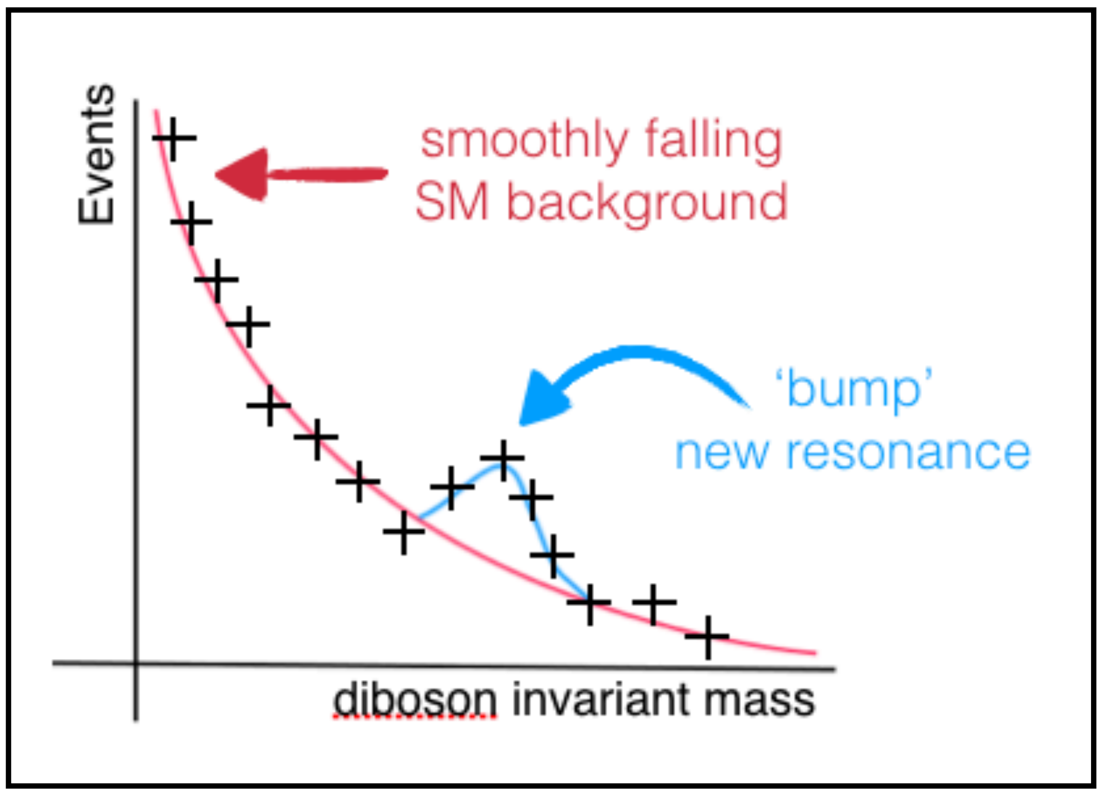
From **1D** bump hunt

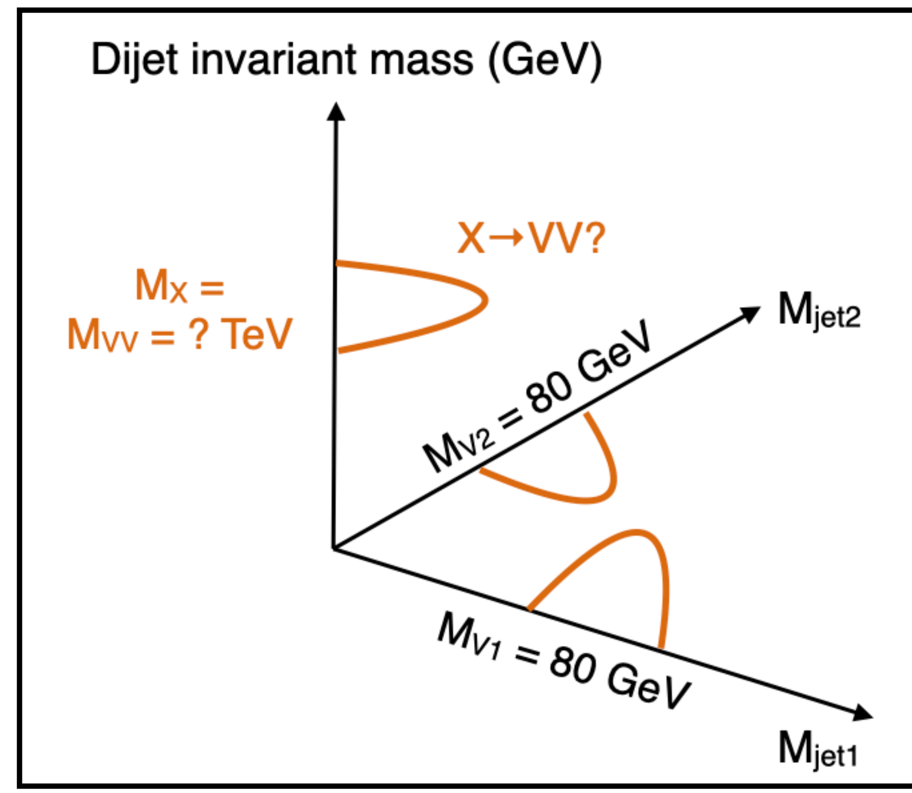
relax assumptions →

To **3D** bump hunt

fit to m_{jj} spectrum after cuts on jet mass and substructure

no cuts on the mass of the two jets

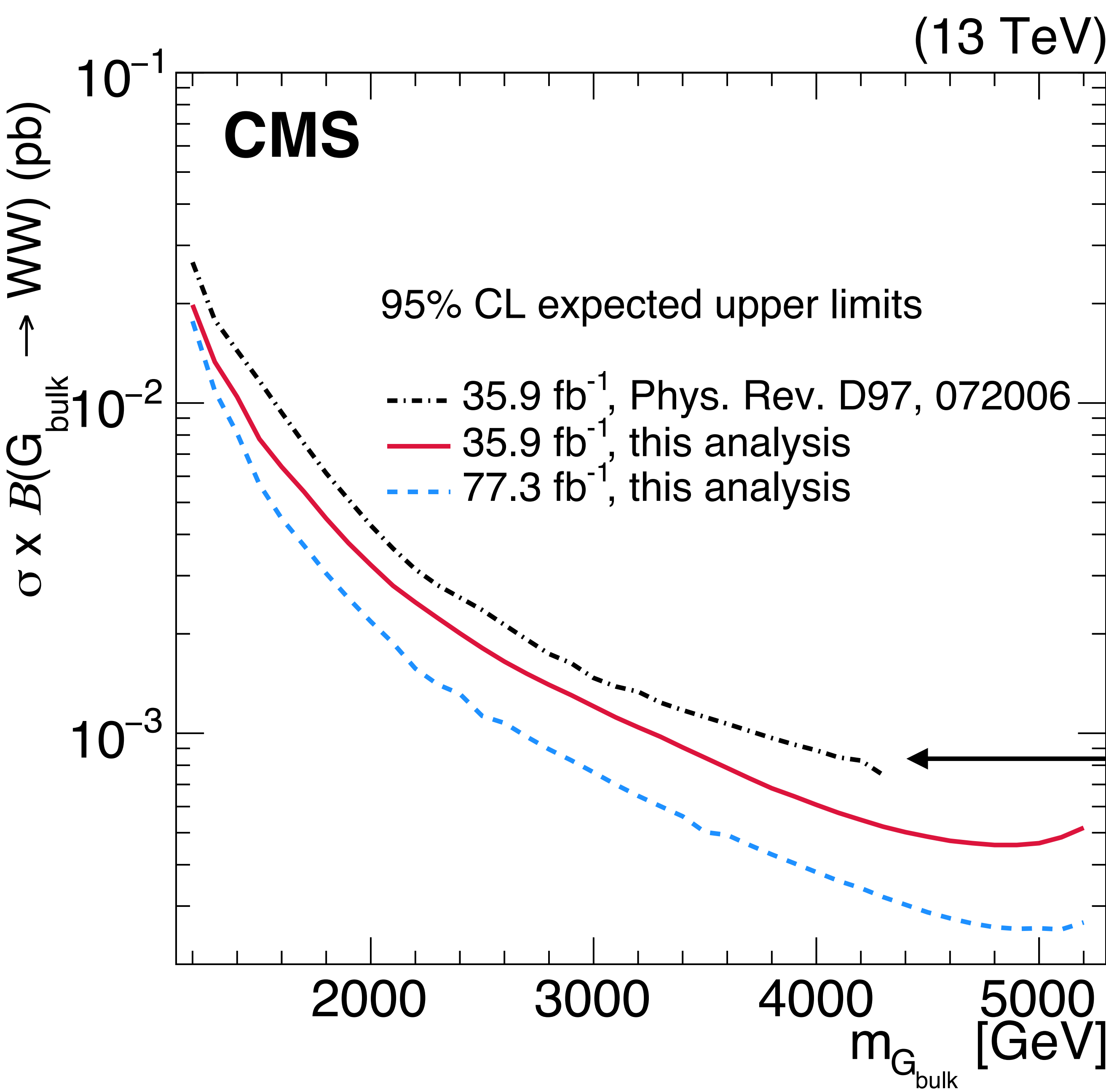


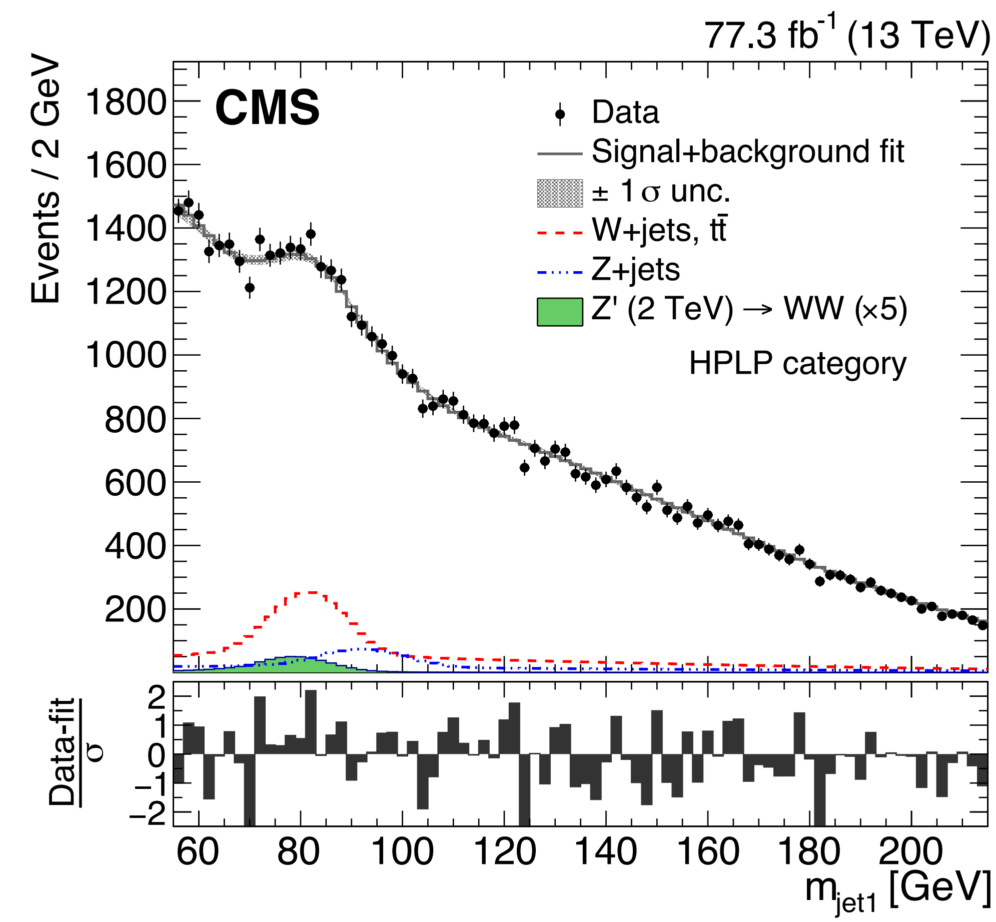


- Applied to Run 2 CMS data for heavy X→diboson→JJ search [EPJC 80 (2020) 237]
- Take advantage of signal peaking in both jet mass and dijet invariant mass and search for X→diboson in (M_{VV} - M_{jet1} - M_{jet2}) space

More boost: the 3D bump hunt

- Full modelling of correlation among m_{jj} and jet mass in QCD multijet background show improved sensitivity
  - more information inserted in the final fit





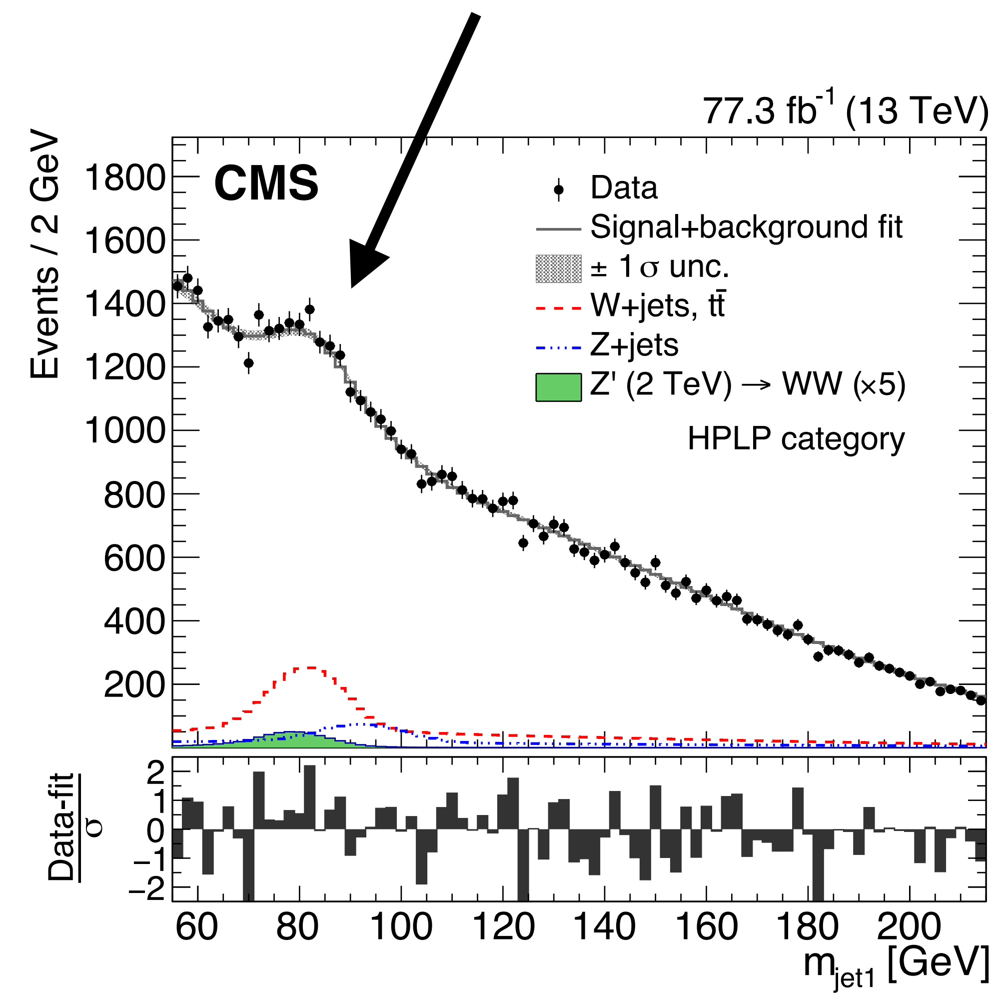
1D fit: 2016 run only

3D fit: 2016 run only

3D fit: 2016+2017 runs

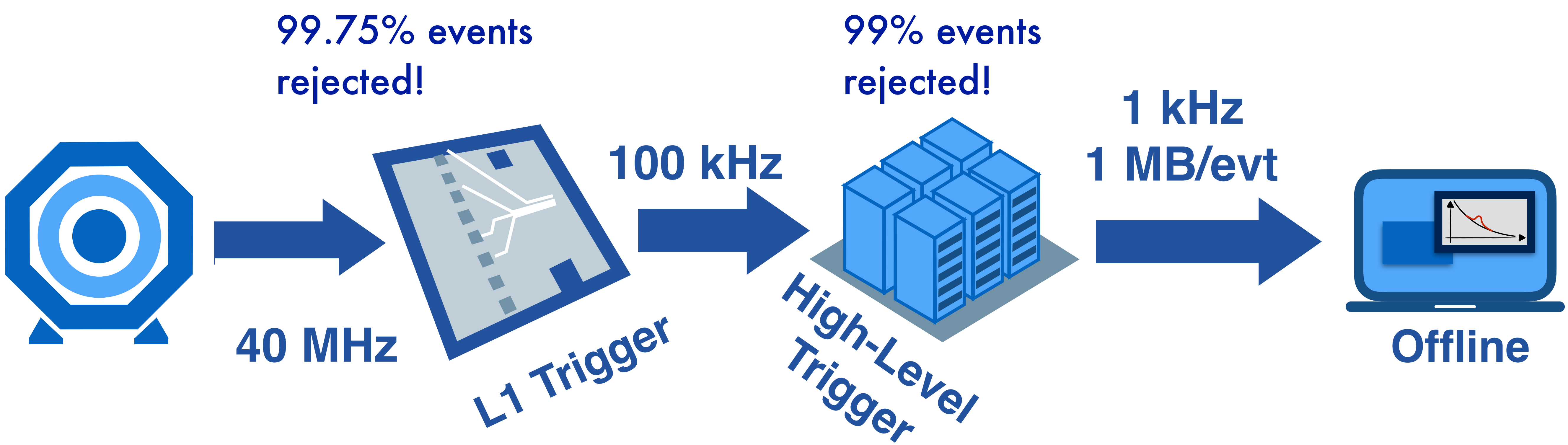
More boost: the 3D bump hunt

- Original CMS analysis used jet substructure targeting SM boson jet
- But ideal framework for anomalous dijet event tagging where mother and daughter particles are not known ($X \to YY'$, all three unknown)
- Could benefit from more controlled background model and jet mass calibrations
  - resonant backgrounds as V+jets or $t\bar{t}$ to be enhanced after cut on the anomaly score?
- A 3D quantile regression probably needed if this approach is applied



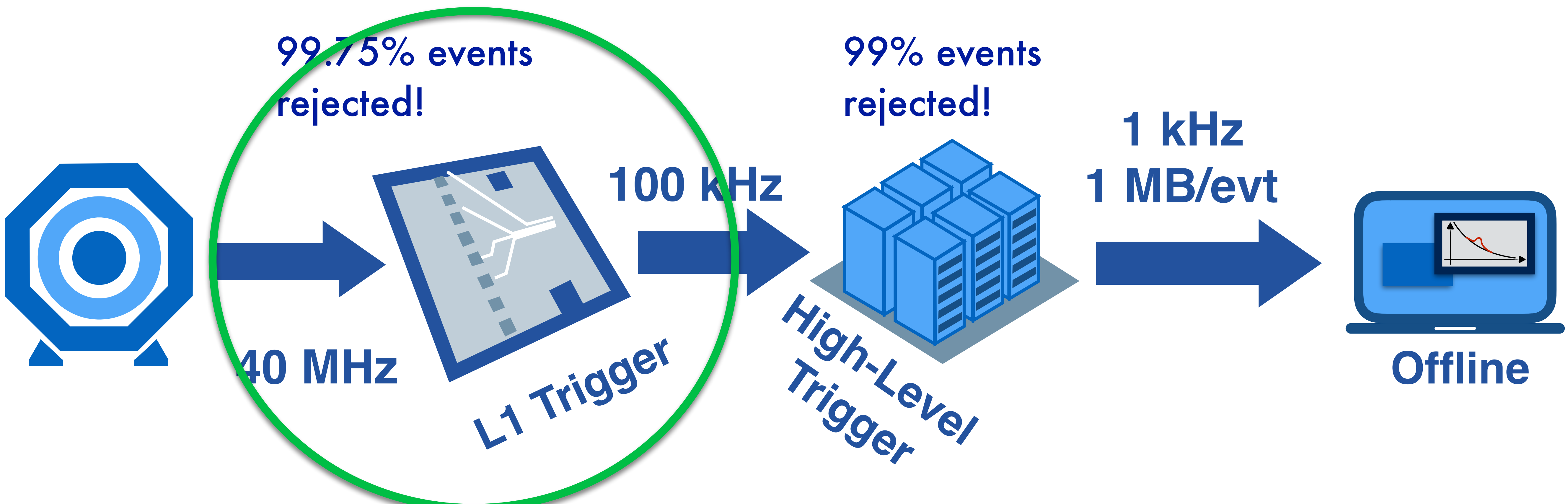
More boost: apply it to the trigger!

- With 40M collisions/seconds and 1000 stored, we might just being writing the wrong events
  - trigger algorithms quite model dependent
  - the anomaly that we look for offline could have easily be discarded



More boost: apply it to the trigger!

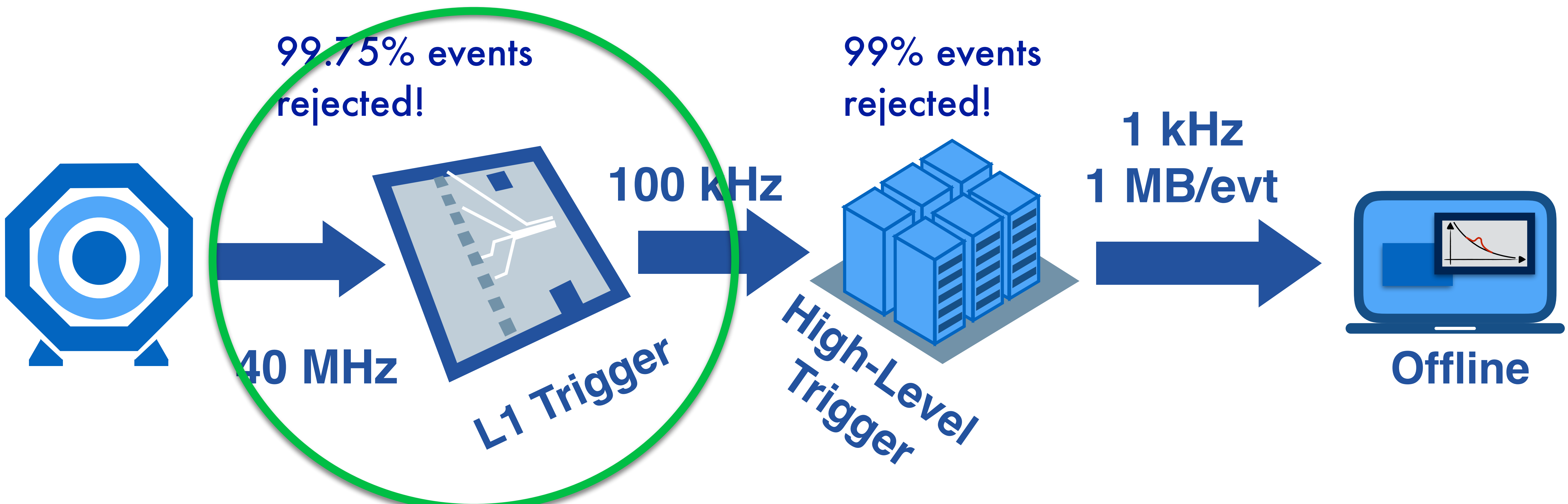
- With 40M collisions/seconds and 1000 stored, we might just being writing the wrong events
  - trigger algorithms quite model dependent
  - the anomaly that we look for offline could have easily be discarded



Correct the problem as early as possible in the data reduction flow!

More boost: apply it to the trigger!

- DL algorithms can become relatively large → memory and number of operations required for the inference can easily explode
- **Strict constraints at L1 trigger:**
  - latency of O(μs) → use FPGA hardware
  - scarse resources (mostly occupied to calibrate sensors, build physics objects, etc..)

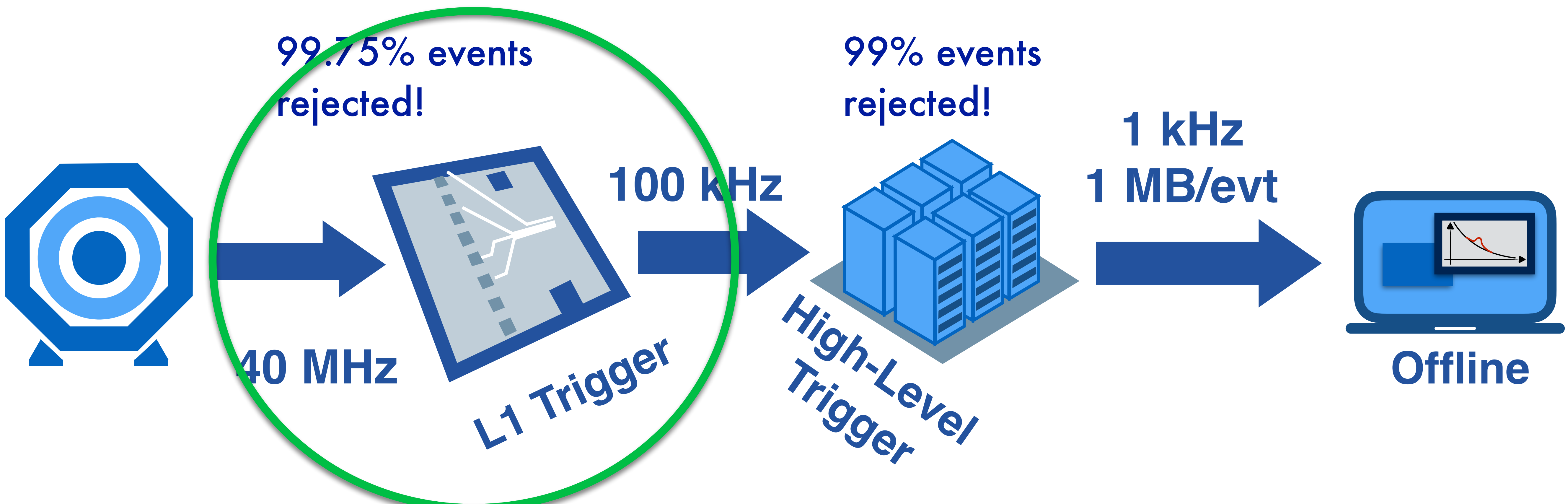


Correct the problem as early as possible in the data reduction flow!

More boost: apply it to the trigger!

- DL algorithms can become relatively large → memory and number of operations required for the inference can easily explode
- **Strict constraints at L1 trigger:**
  - latency of O(μs) → use FPGA hardware
  - scarse resources (mostly occupied to calibrate sensors, build physics objects, etc..)

How to fit a ML algo here?



Correct the problem as early as possible in the data reduction flow!

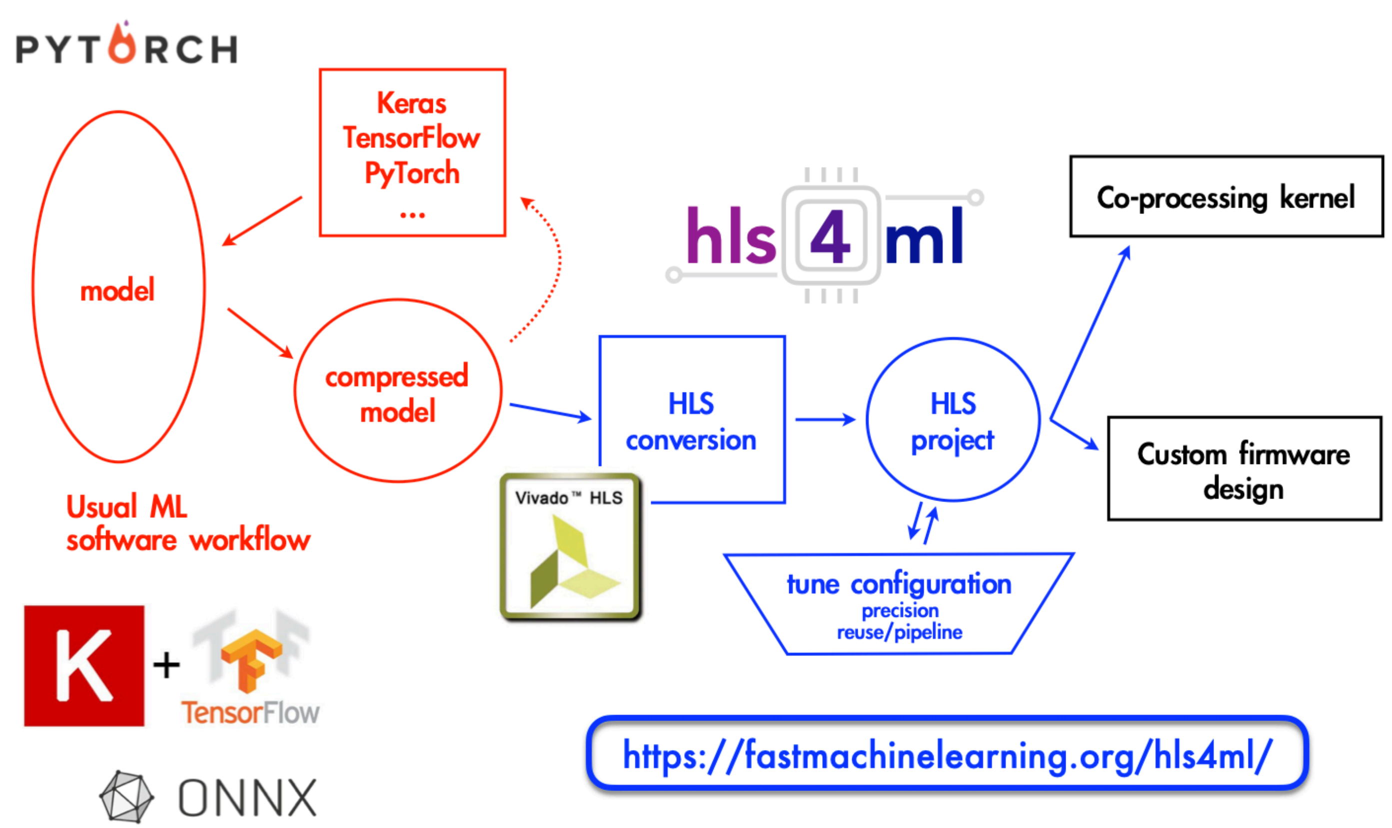
Bring DL to FPGA for L1 trigger with

high level synthesis for machine learning

- Automated tool to deploy DNN in FPGA with ultra low latency
- Easy to tune the inference performance for your specific application: *precision, resource vs latency/throughput tradeoff*

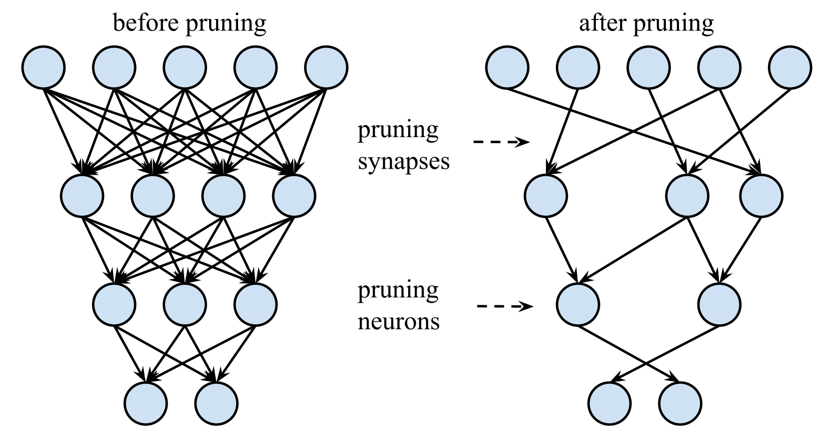
- Can be used as API
- Includes several debugging utilities

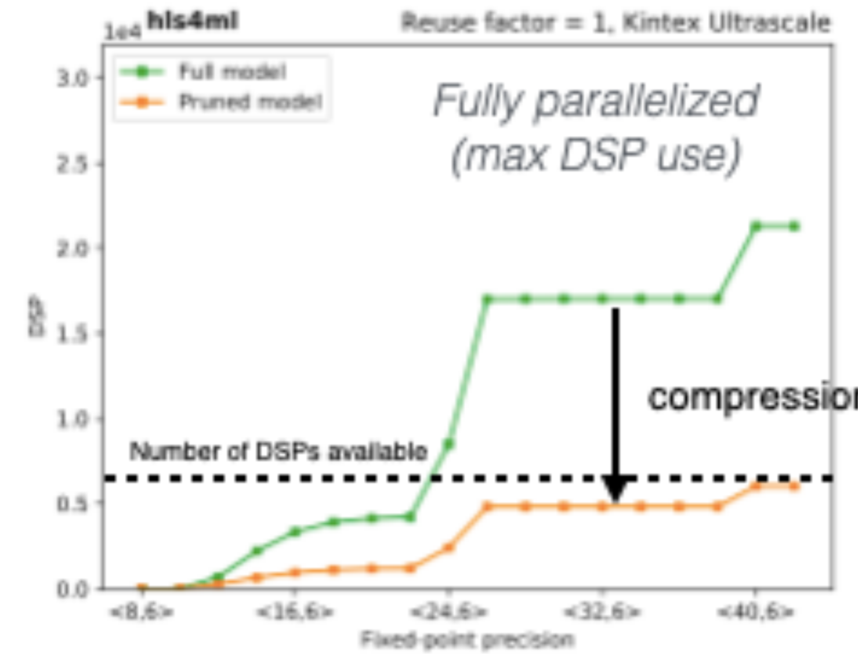
- Most common DL layers and activation functions supported



Make the model fit on one chip

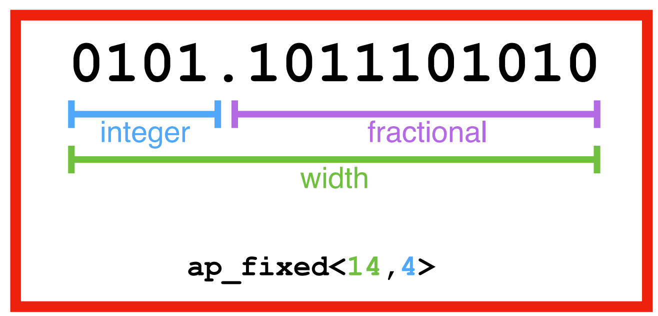
- Some tricks are needed here:
  - **Pruning:** remove the connections that play little role for final decision



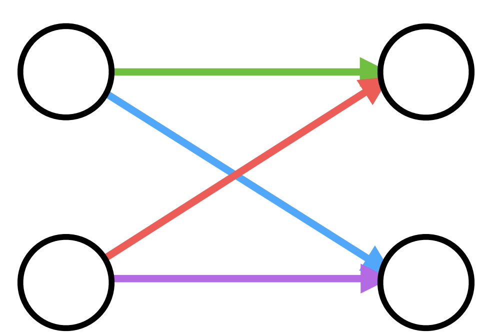


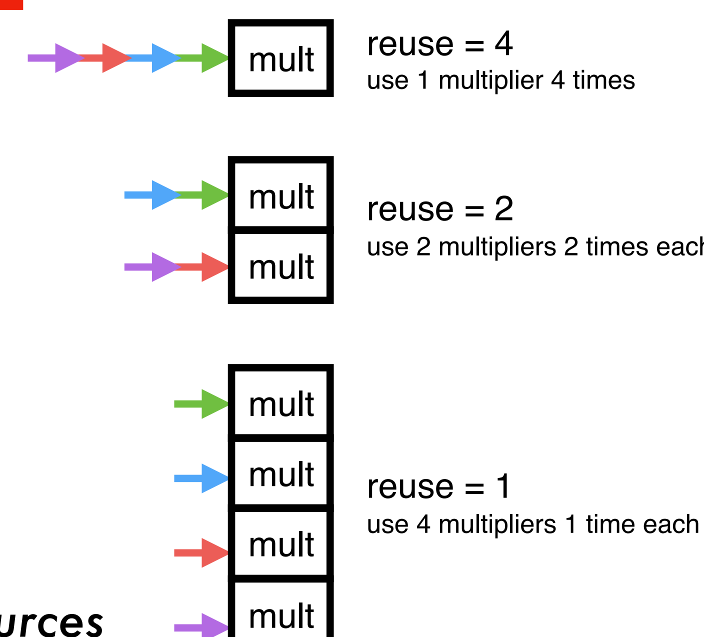
70% compression ~ 70% fewer DSPs

  - **Quantisation:** represents numbers with few bits reduce resources



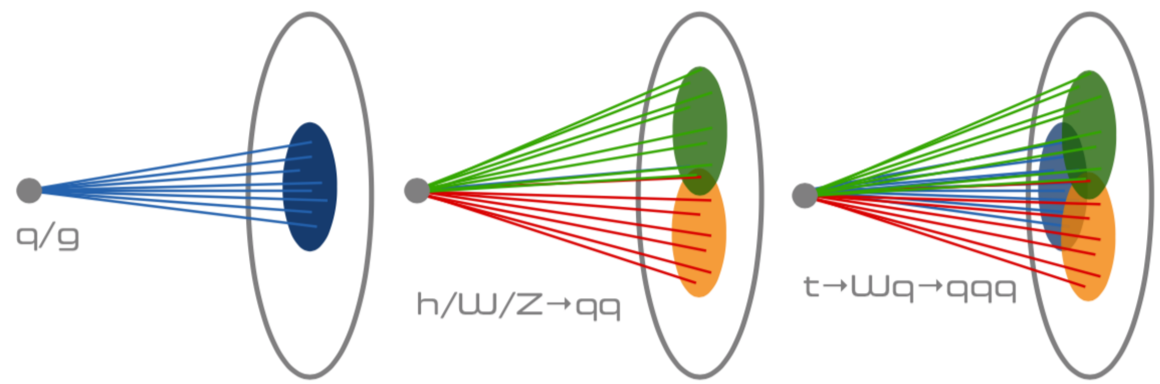
  - **Reuse:** allocate resources for each operation (run all network in one clock) vs spread calculation across several clock cycles

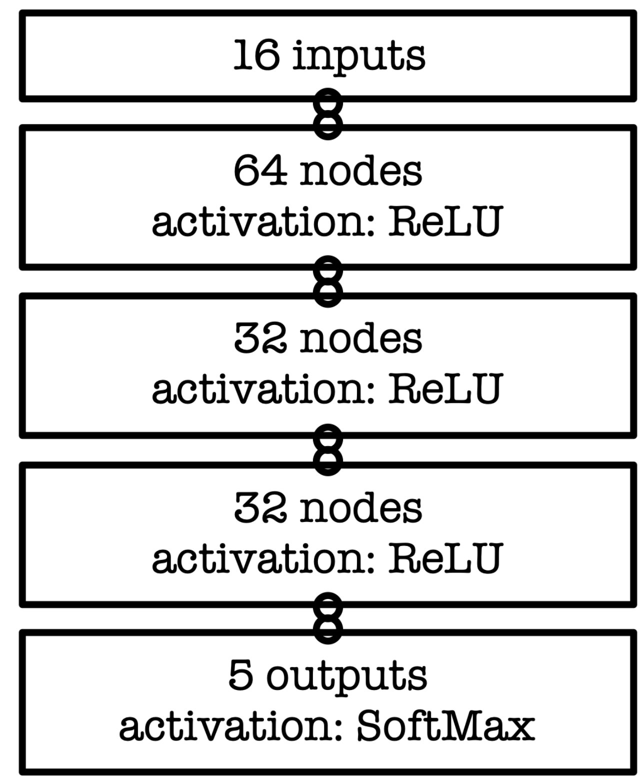




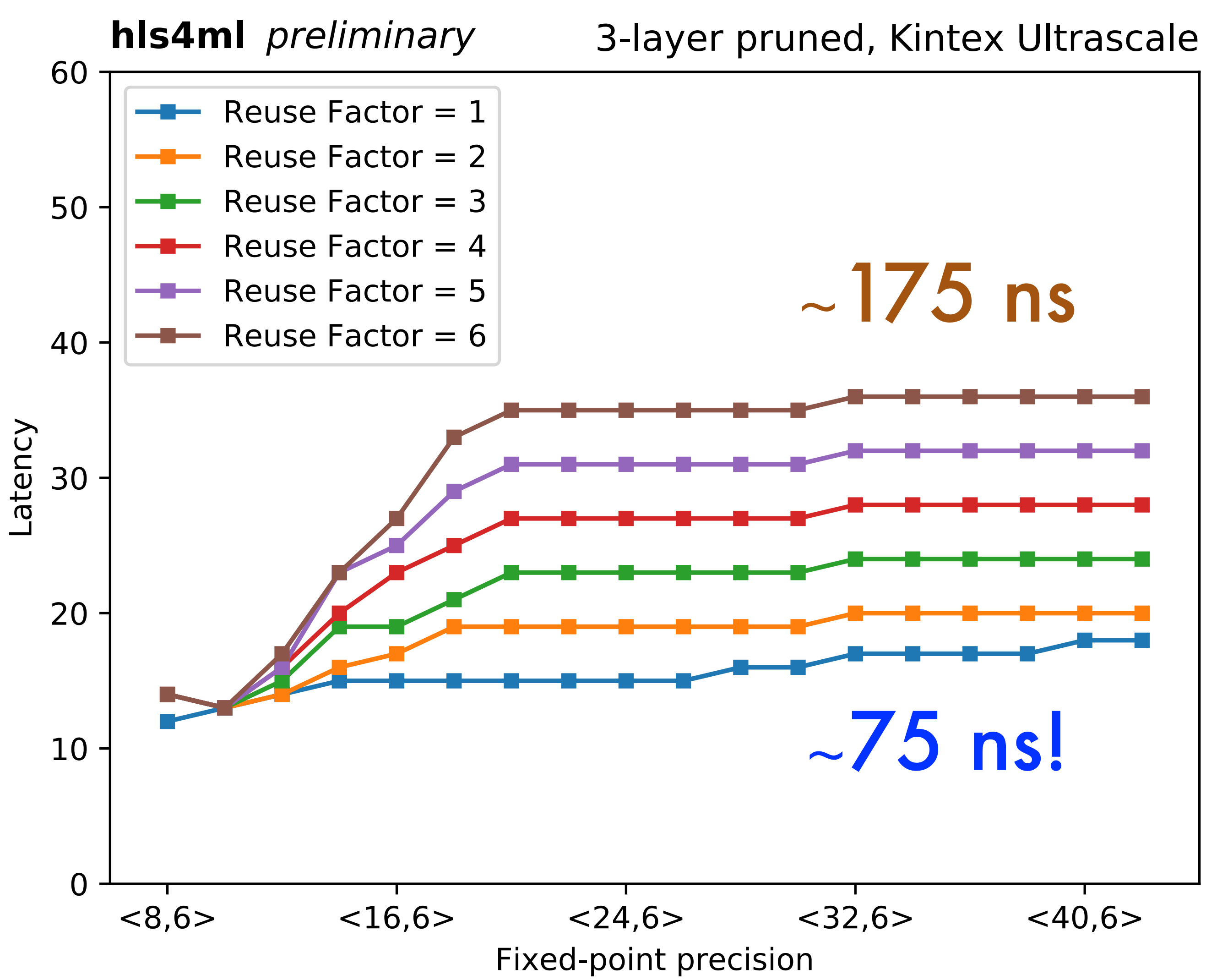
more parallelization → more resources

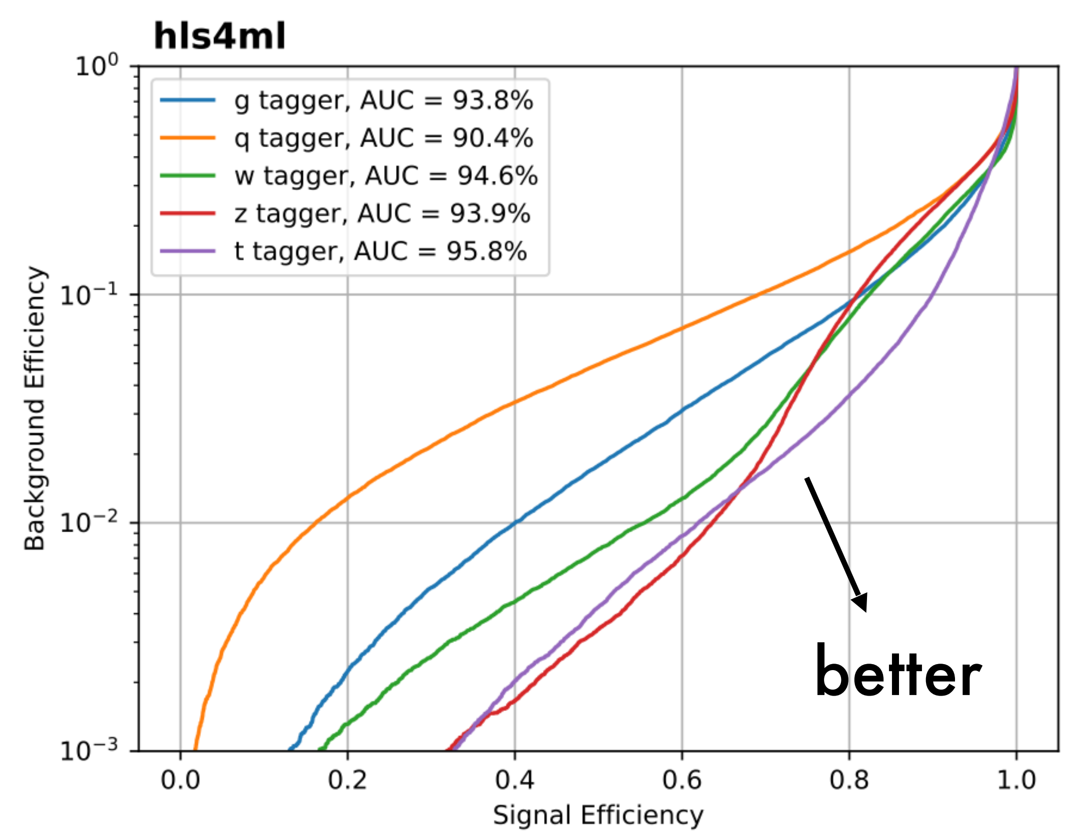
Ultra-low latency inference





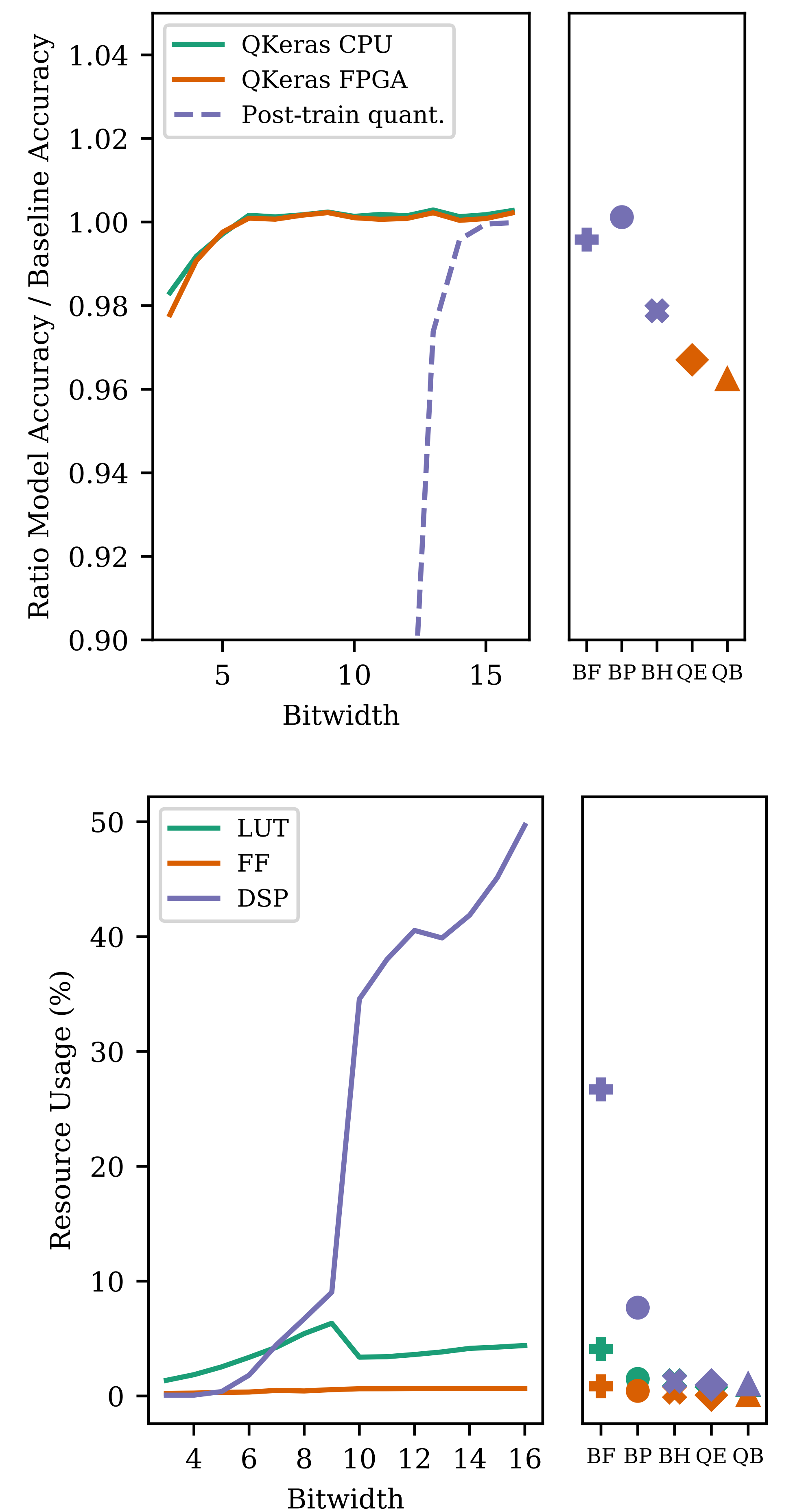
→ high-level features: jet mass, substructure, multiplicity, etc...





Quantization-aware training

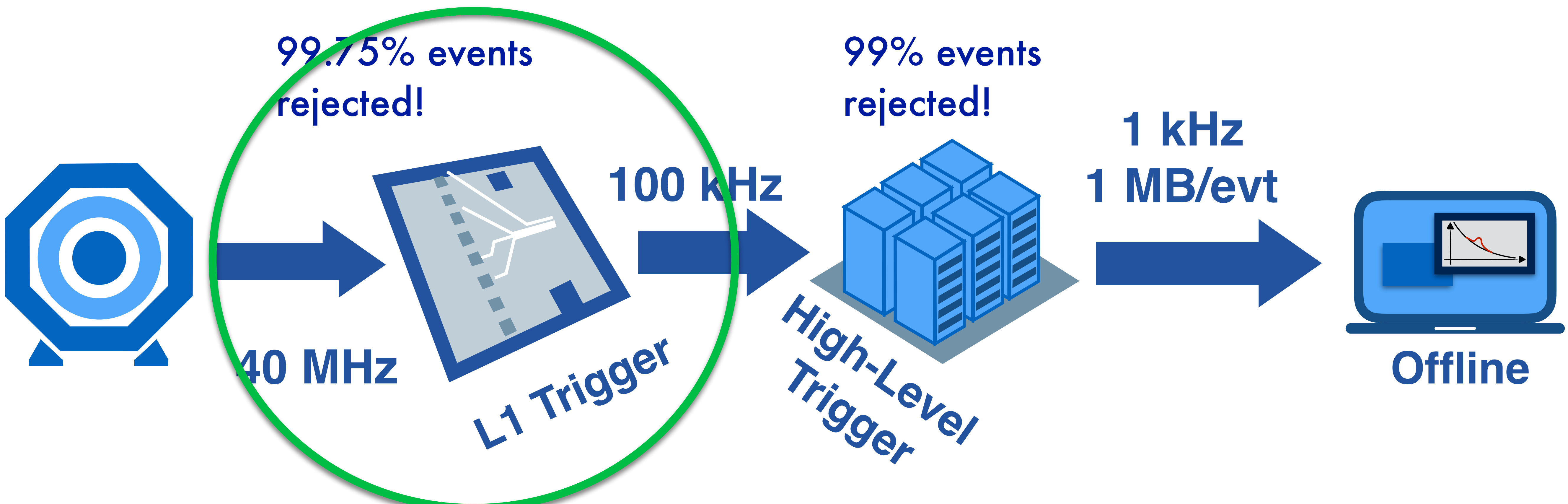
- Post-training quantization can affect accuracy
  - for a given bit allocation, the loss minimum at floating-point precision might not be the minimum anymore
- One could specify quantization while look for the minimum
  - maximize accuracy for minimal FPGA resources
- Workflow: quantization-aware training with Google QKeras and firmware design with hls4ml for best NN inference on FPGA performance



C. N. Coelho et al., arXiv:2006.10159 - Submitted to Nature Machine Intelligence

More boost: apply it to the trigger!

How to fit a ML algo here?



Correct the problem as early as possible in the data reduction flow!

Fast autoencoders @ L1

- We start from the single-lepton data stream discussed previously

- Move to momentum-based data representation

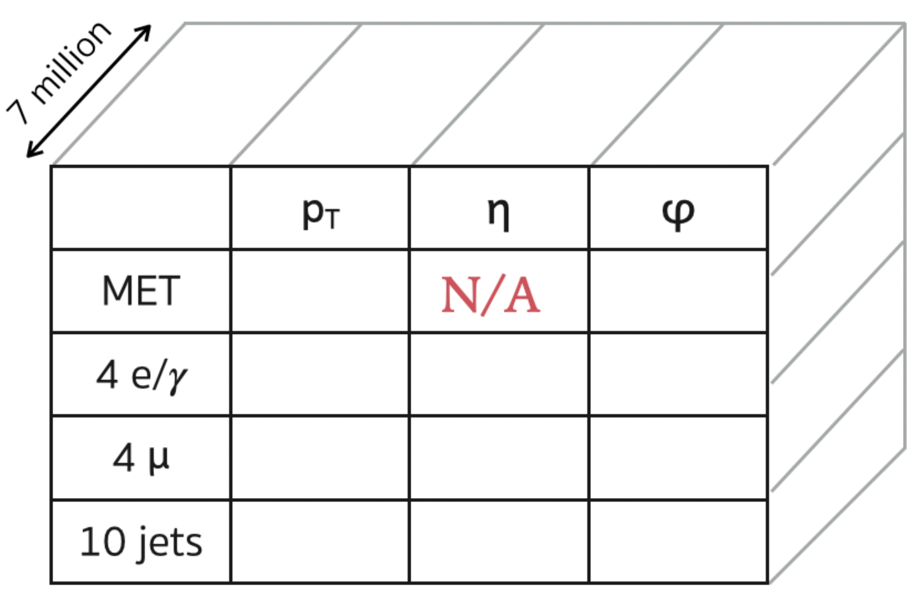
  - avoid need of computing high-level features at L1 which can be time or resource consuming

Standard Model processes					
Process	Acceptance	L1 trigger efficiency	Cross section [nb]	Event fraction	Events /month
W	55.6%	68%	58	59.2%	110M
QCD	0.08%	9.6%	$1.6 \cdot 10^5$	33.8%	63M
Z	16%	77%	20	6.7%	12M
$t\bar{t}$	37%	49%	0.7	0.3%	0.6M

BSM benchmark processes				
Process	Acceptance	L1 trigger efficiency	Total efficiency	Cross-section 100 BSM events/month
$A \to 4\ell$	5%	98%	5%	0.44 pb
$LQ \to b\tau$	19%	62%	12%	0.17 pb
$h^0 \to \tau\tau$	9%	70%	6%	0.34 pb
$h^\pm \to \tau\nu$	18%	69%	12%	0.16 pb

- We compare different architectures: CNN vs DNN and autoencoders (AE) versus variational AE (VAE)

- m_A = 50 GeV
- m_{LQ} = 80 GeV
- m_{h0} = 60 GeV
- $m_{h\pm}$ = 60 GeV



Number of objects chosen to emulate limited L1 bandwidth

Variational autoencoders

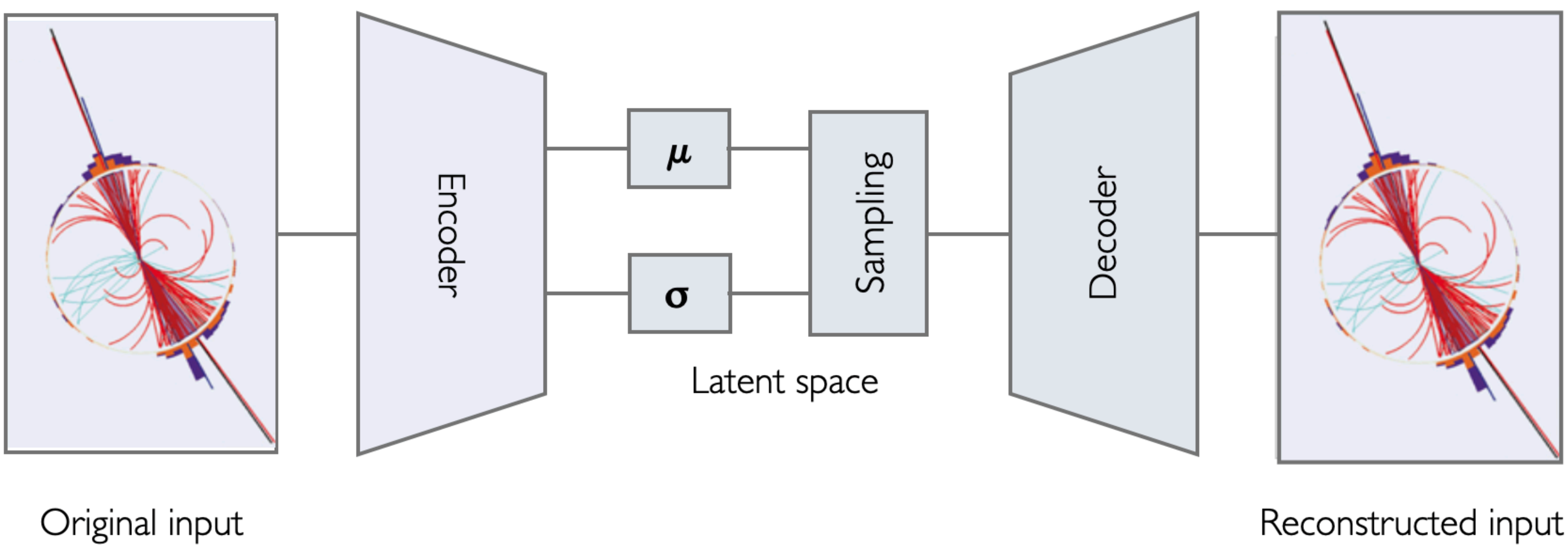
- Encode inputs as pdfs over latent space rather than single point
  → return $\vec{\mu}$ and $\vec{\sigma}$ of N-dim Gaussian

- Impose prior on latent space and add divergence to total loss

$$\mathcal{L}_{tot} = (1-\beta)\cdot L_{reco} + \beta \cdot D_{KL}(\overrightarrow{\mu}, \overrightarrow{\sigma})$$

MSE I/O
anomaly detection

Kullback-Leibler regularization term



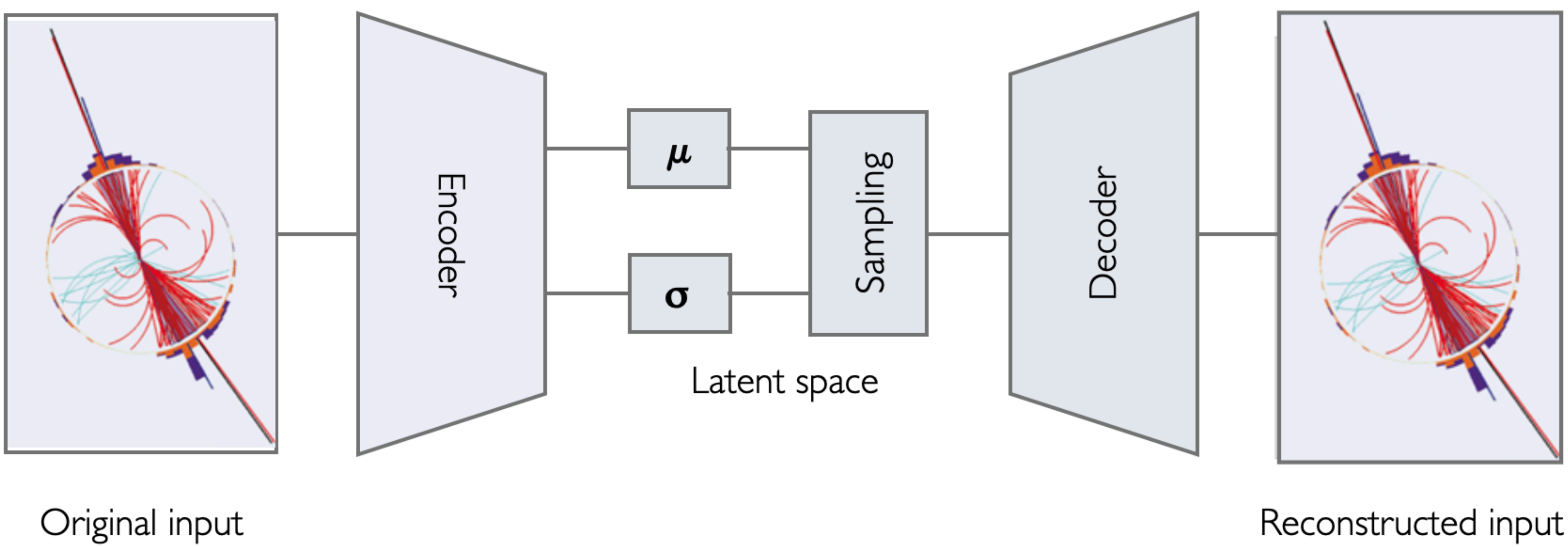
Variational autoencoders

$$\mathcal{L}_{tot} = (1-\beta)\cdot L_{reco} + \beta \cdot D_{KL}(\vec{\mu}, \vec{\sigma})$$

MSE I/O
anomaly detection

Kullback-Leibler regularization term

Baseline I/O AD sub-optimal @ L1:
- Random sampling not practical in L1 environment
- Trigger decision required to be deterministic

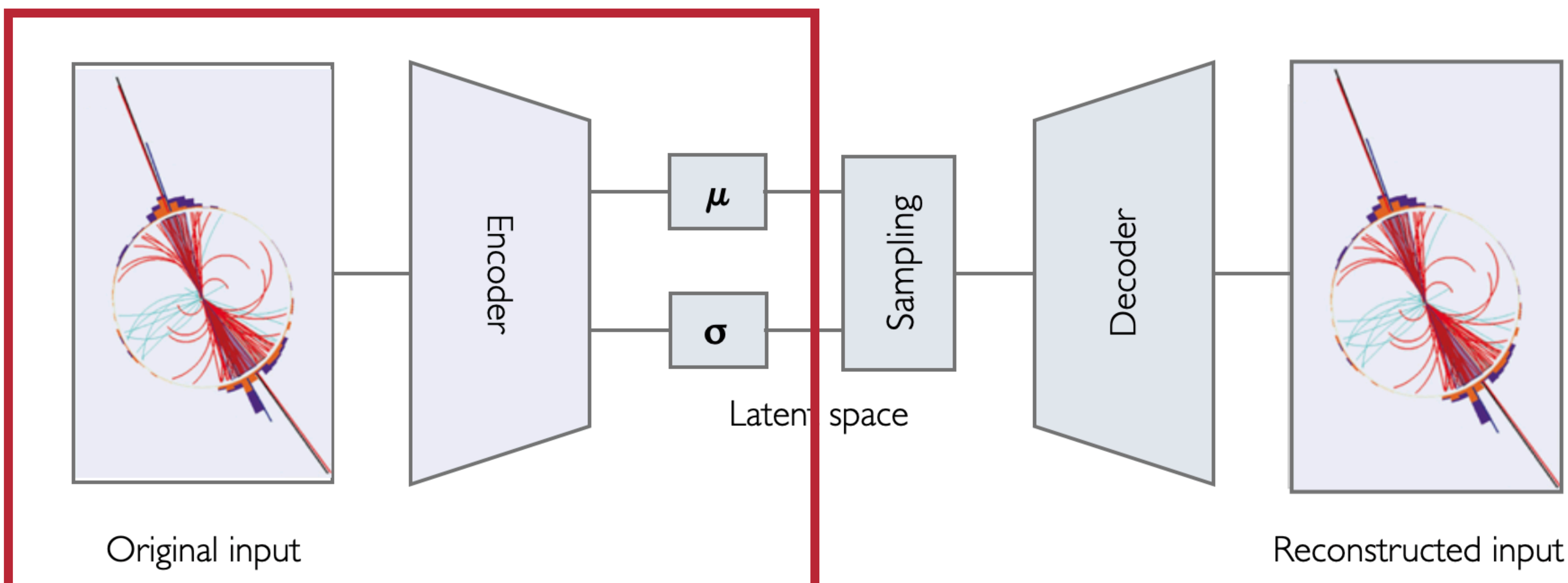


Fast autoencoders @ L1

ALTERNATIVE APPROACH:

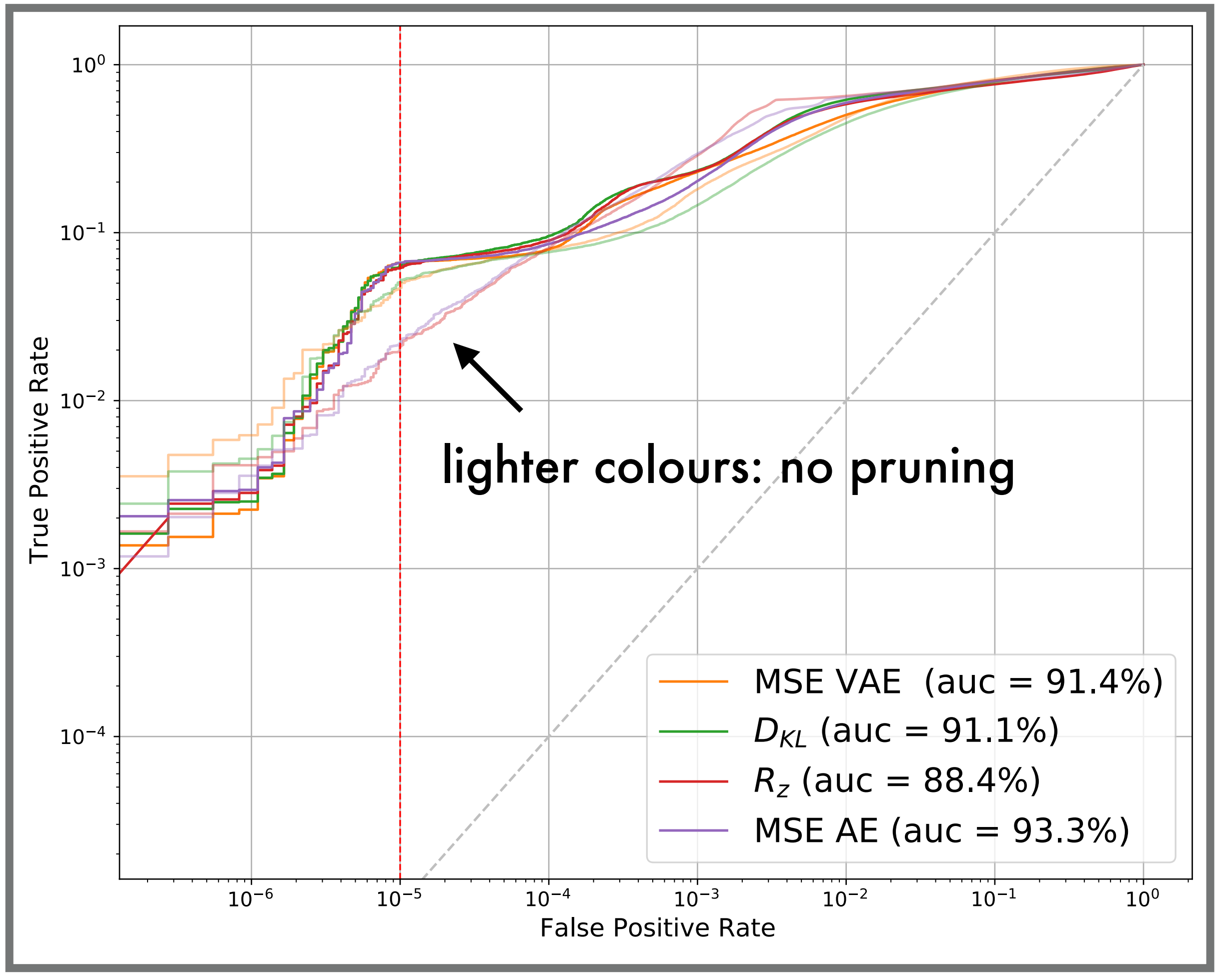
- Train encoder+decoder with $\boxed{\mathcal{L}_{tot} = (1-\beta)\cdot L_{reco} + \beta \cdot D_{KL}(\vec{\mu}, \vec{\sigma})}$
- Define an AD figure of merit in the latent space $\boxed{D_{KL}(\vec{\mu}, \vec{\sigma})}$ or $\boxed{R_z = \sum_i (\mu_i/\sigma_i)^2}$
- Advantages for L1 trigger application:
  - no sampling at inference
  - save resources and latency by not running decoder at inference

Pull of Gaussian from expectation (μ=0, σ=1) in the latent space



Fast autoencoders @ L1

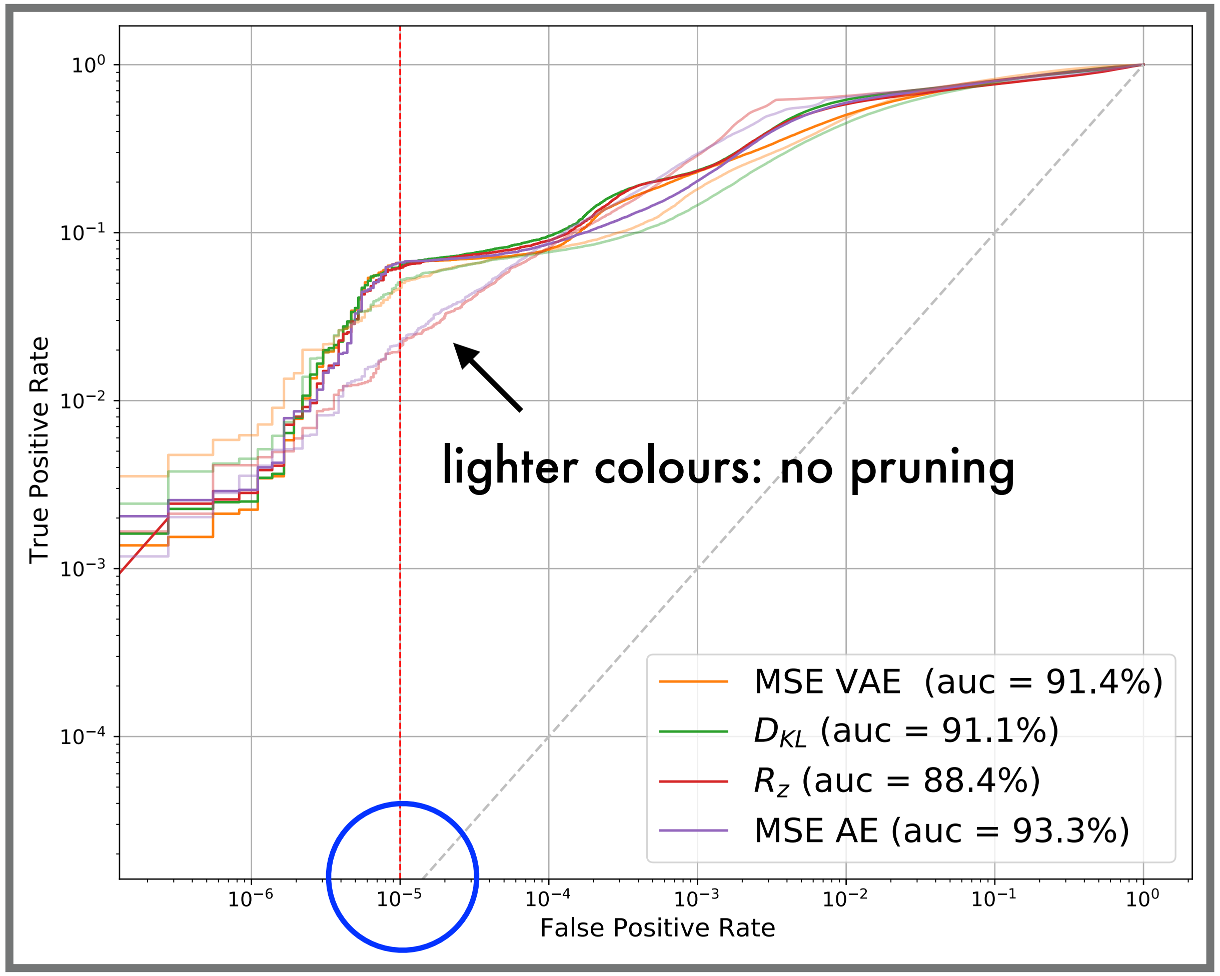
Dense NN
Signal: A → 4l



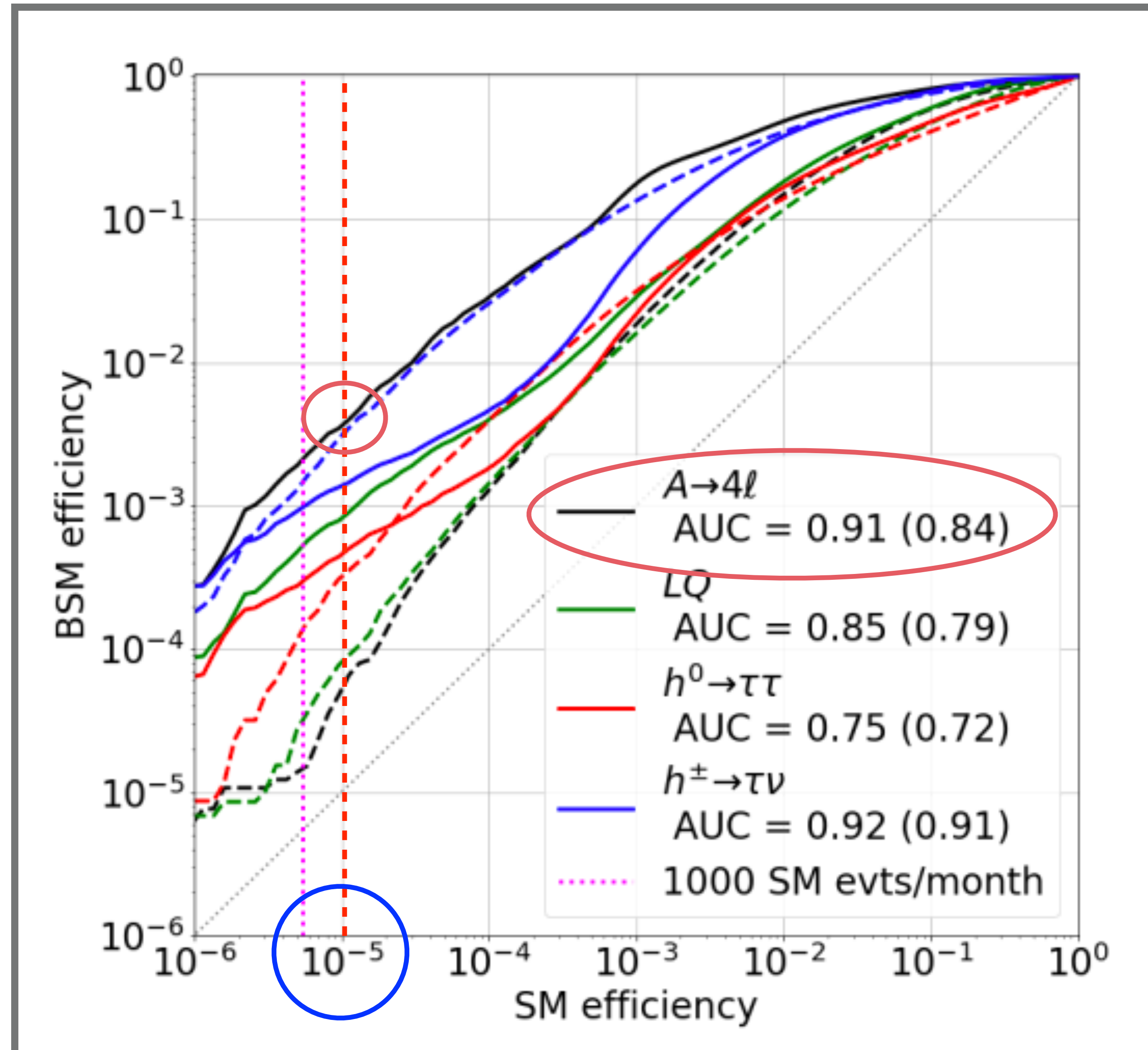
- $MSE_{VAE} \cong MSE_{AE} \cong D_{KL}$ → can run only encoder @ L1 without loss in performance
- Pruning preserves performance
- Can also be quantized during training with QKeras to reduce resources
- Similar conclusions for the CNN architecture
  - final choice mainly depends on resources and latency

Fast autoencoders @ L1

Dense NN
Signal: A → 4l



~x10 improvement wrt original study!



FPR = 10^{-5} → threshold for comparing figures of merit

Fast autoencoders @ L1

- $MSE_{VAE} \cong MSE_{AE} \cong D_{KL}$ → can run only encoder @ L1 without loss in performance
- Pruning preserves performance
- Can also be quantized during training with QKeras to reduce resources
- Similar conclusions for the CNN architecture
  - final choice mainly depends on resources and latency

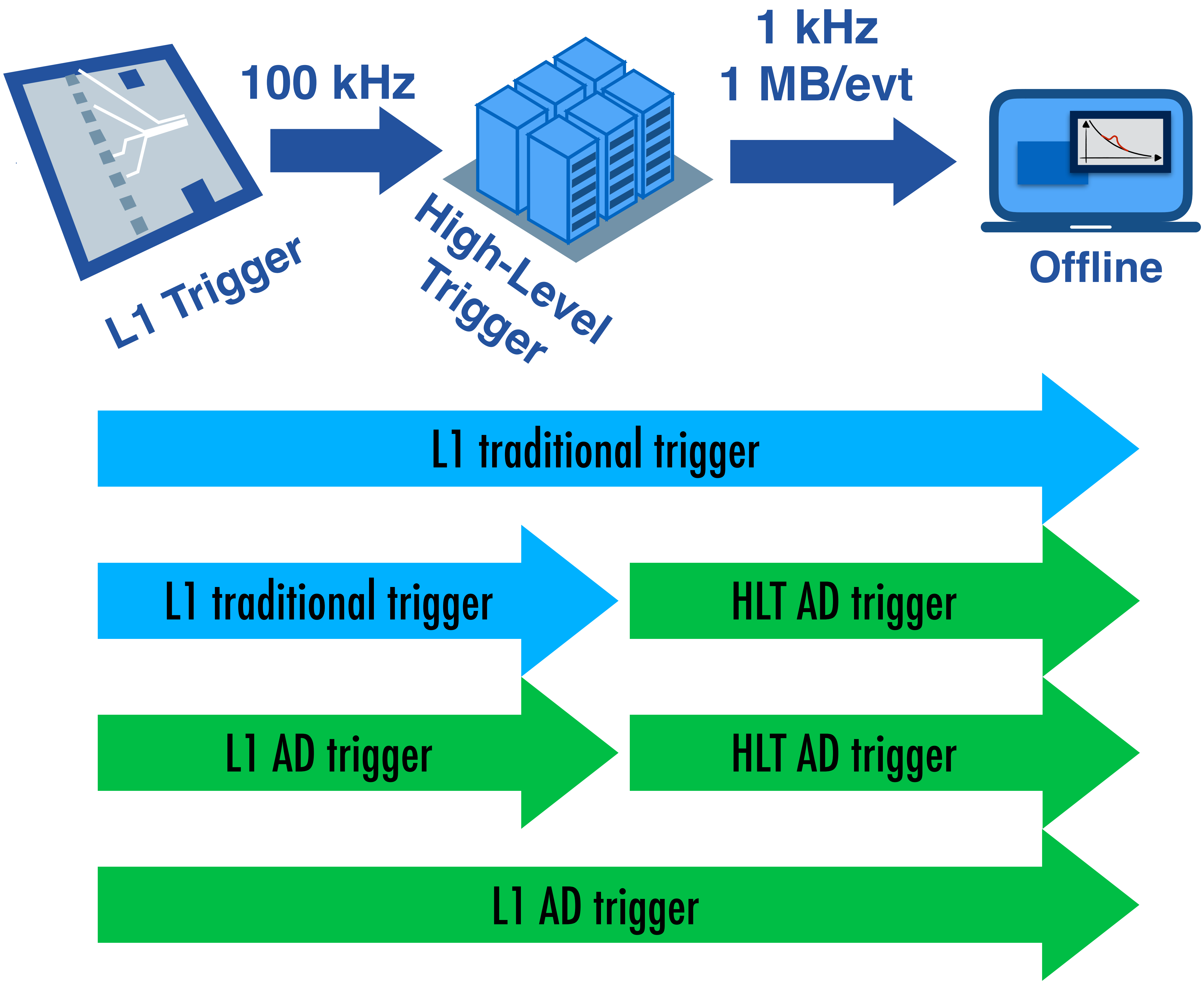
Q (8 bits)	Latency (ns)	DSPs (%)	LUTs (%)	FFs (%)
DNN AE	48	20	8	0.4
DNN VAE (encoder)	40	9	3	~0
CNN VAE (encoder)	275	21	18	3

→ Could already be implemented for Run 3

→ Target HL-LHC

nb, results for target device for Phase 2 CMS trigger system

Anomaly detection for Run 3



Anomaly detection for Run 3

- The obvious question is **what to do with these "anomalous" data?**

- The answer is an additional and new field of study
  - run clustering algorithms (eg, KNN) on these data in the latent space or natural space of the inputs
  - look at differential distributions then develop analysis/trigger tailored to a specific final state/signal
  - publish the data as a catalog to incentivate new ideas in view of HL-LHC
  - full statistical analysis also possible