New Physics Mining with Deep Learning at the LHC

Maurizio Pierini

CERN

European Research Council

O. Cerri et al., arXiv:1812.XXXXX
O. Cerri et al., arXiv:19YY.ZZZZZ

About this seminar

- An idea we are working on since a while (there is always something useful to make it better)
- Work is still in progress. Results are preliminary, but we are refining them to get soon a paper out
- Work done in collaboration with Caltech CMS group

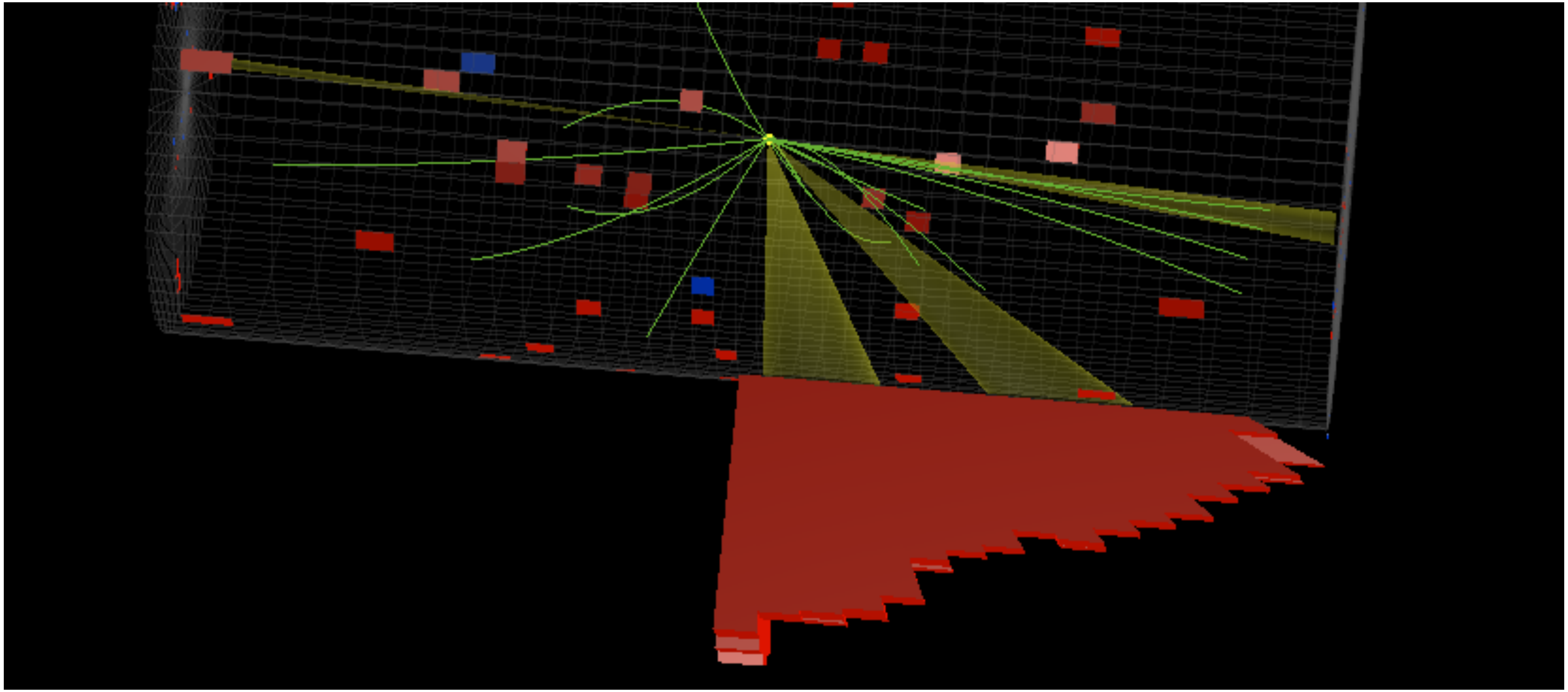
OLMO CERRI
(PHD STUDENT)

THONG NGUYEN
(PHD STUDENT)

MARIA SPIROPULU

JEAN-ROCH VLIMANT

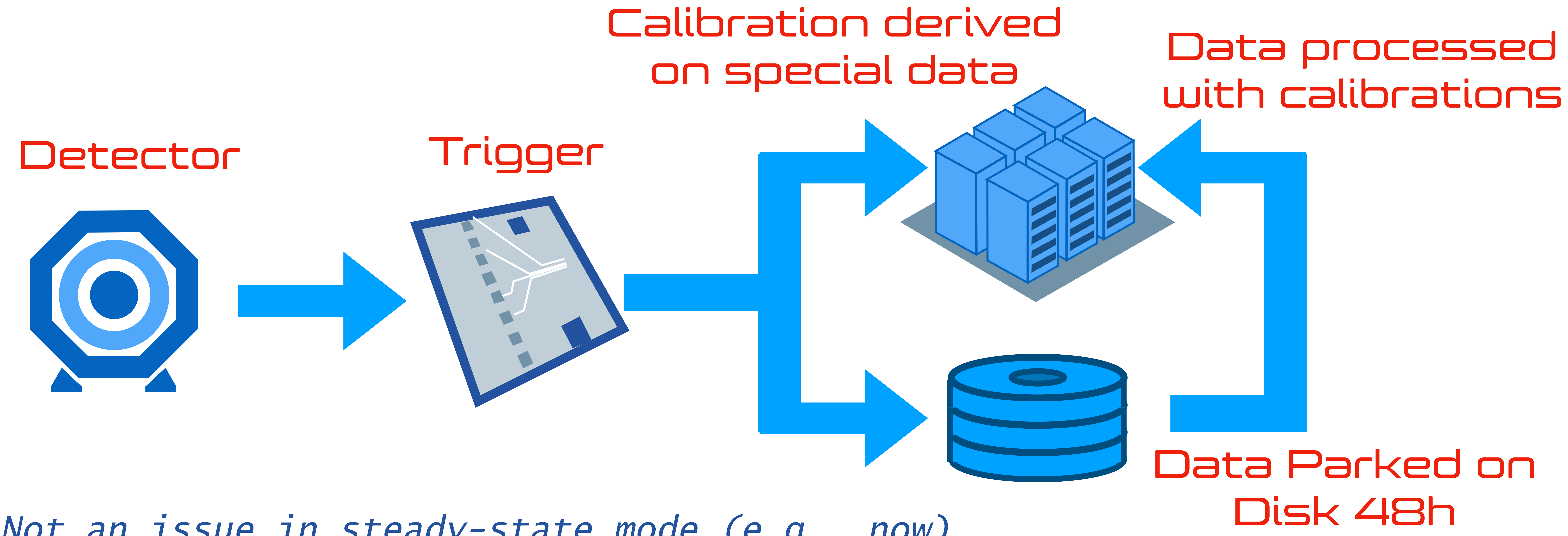
European Research Council



BSM as Anomalous events at LHC Run I

The CMS data-flow

- Traditionally, data take 2-7 days before becoming available for analysis



  - Not an issue in steady-state mode (e.g., now)
  - A substantial delay at startup, particularly if you hope for an early discovery
- This is why we implemented in 2009 an alarm system for special physics events

The Exotica Hotline

- Back in 2009, we implemented a set of triggers to catch rare (and possibly interesting events)
  - high-Pt jets, muons, electrons, photons or taus
  - large lepton multiplicity
  - large di-object invariant mass
- Stored O(10) events/day, processed in real time
- Studied by experts (visual inspection of event displays)

- Jets
  - at least 1 jet with p_T > 200 GeV [updated from 150
  - at least 5 jets with pT > 40 GeV
- MET/HT
  - MET > 250 GeV
  - HT > 300 GeV , calculated summing jets with pT > 30 GeV
- Electrons
  - at least 1 electron with pT > 100 GeV
  - at least 2 electrons with pT > 15 GeV
- Photons
  - at least 1 photon with pT > 100 GeV
  - at least 3 photons with pT > 20 GeV
- Muons
  - at least 1 muon with pT > 100 GeV
  - at least 2 muons with pT > 15 GeV
- Tracks
  - at least 600 tracks
- dE/dX
  - at least 1 track with pT > 50 GeV, dE/dx > 5.6 MeV/cm

European Research Council

The ringing bell…

- *The system was deployed in 2010, with a real-time alert system and a team of expert scanners*
- *It got some attention back then*
- *It was actually very effective in discovering something (which unfortunately was not new physics)*

CERN Bulletin

News Articles | Official News | Training | General Information | Staff Association

english | français

search

Issue No. 46-47/2010 - Monday 15 November 2010

Printable version - Subscribe:

"See you soon" – a retrospective on the first LHC proton run

A milestone in safety

The Latest from the LHC: Successful switch to ions

ATLAS: Exceeding all expectations

CMS: Beyond all possible expectations

ALICE: The best is yet to come

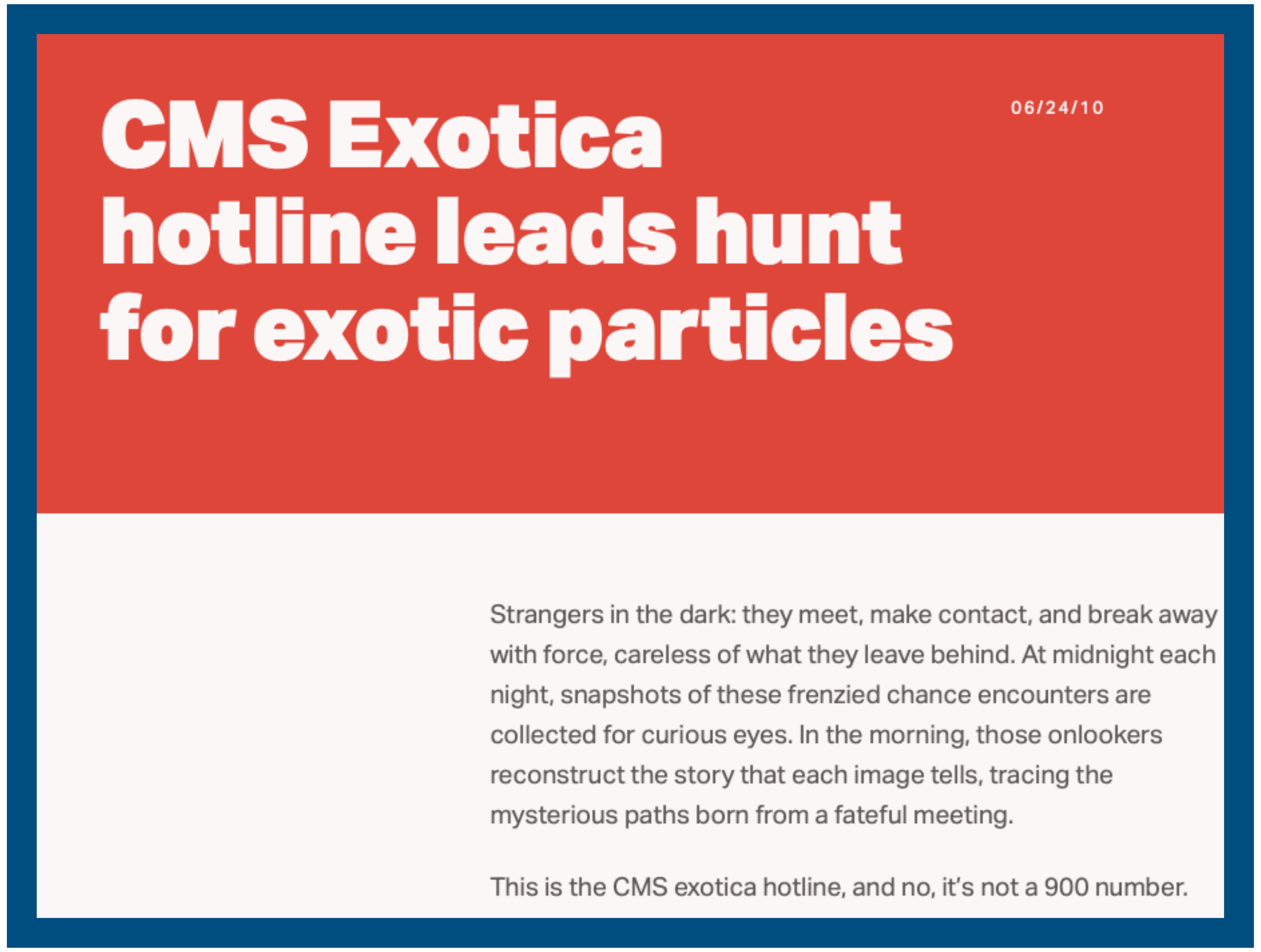
LHCb: Not just a precision experiment but also a detector ready for discoveries

TOTEM: Thousands of interesting events

IS THE BELL RINGING?

During the Nobel prize-winning UA1 experiment, scientists in the control room used to ring a bell if a particularly interesting event had occurred. Today, the "*CMS Exotica hotline*" routine produces a daily report that lists the exotic events that were recorded the day before.

Take just a very small fraction of the available data (max. 5%); define the events that you want to keep and set the parameters accordingly; run the *Exotica* routine and only look at the very few images



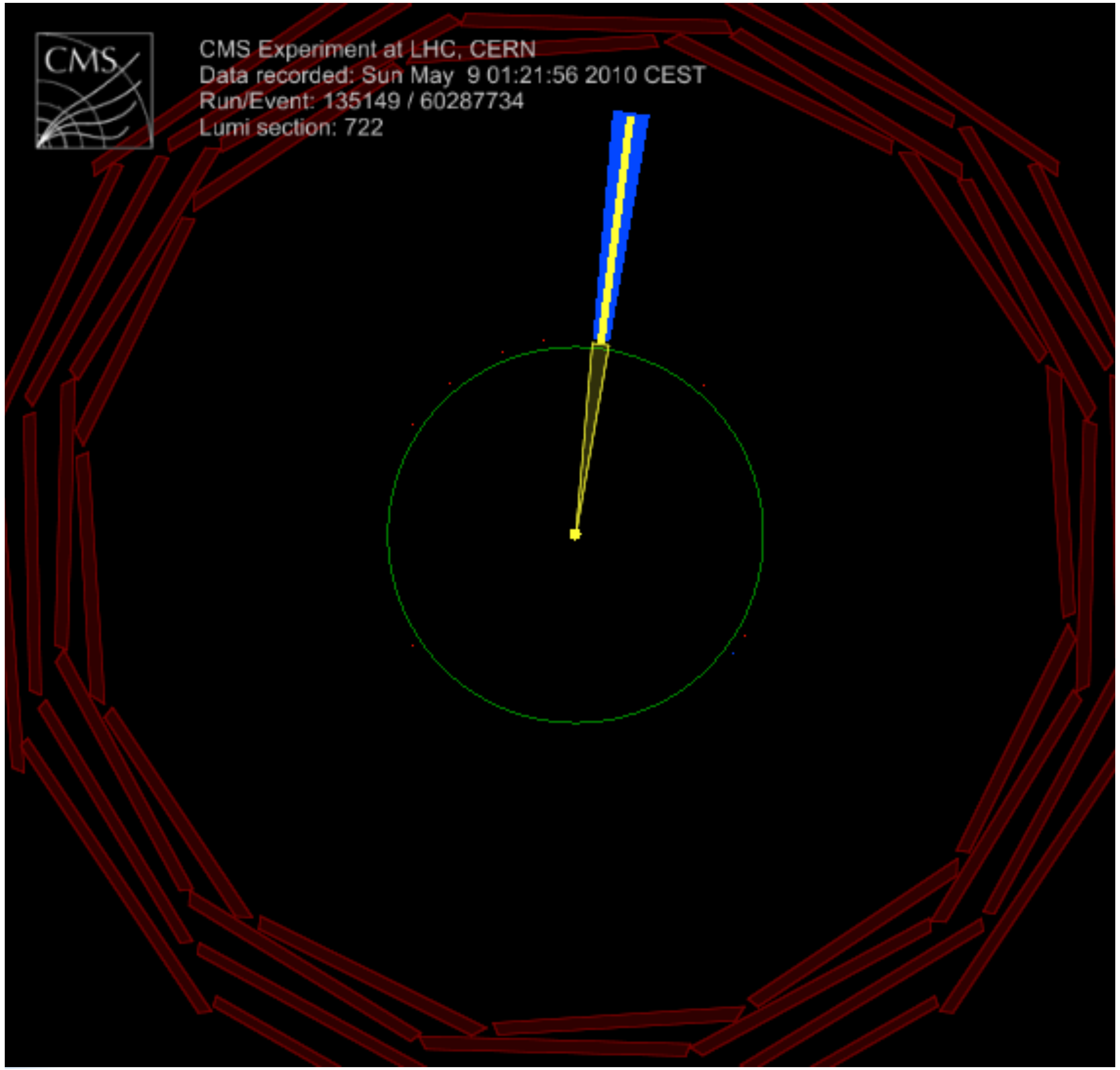
06/24/10

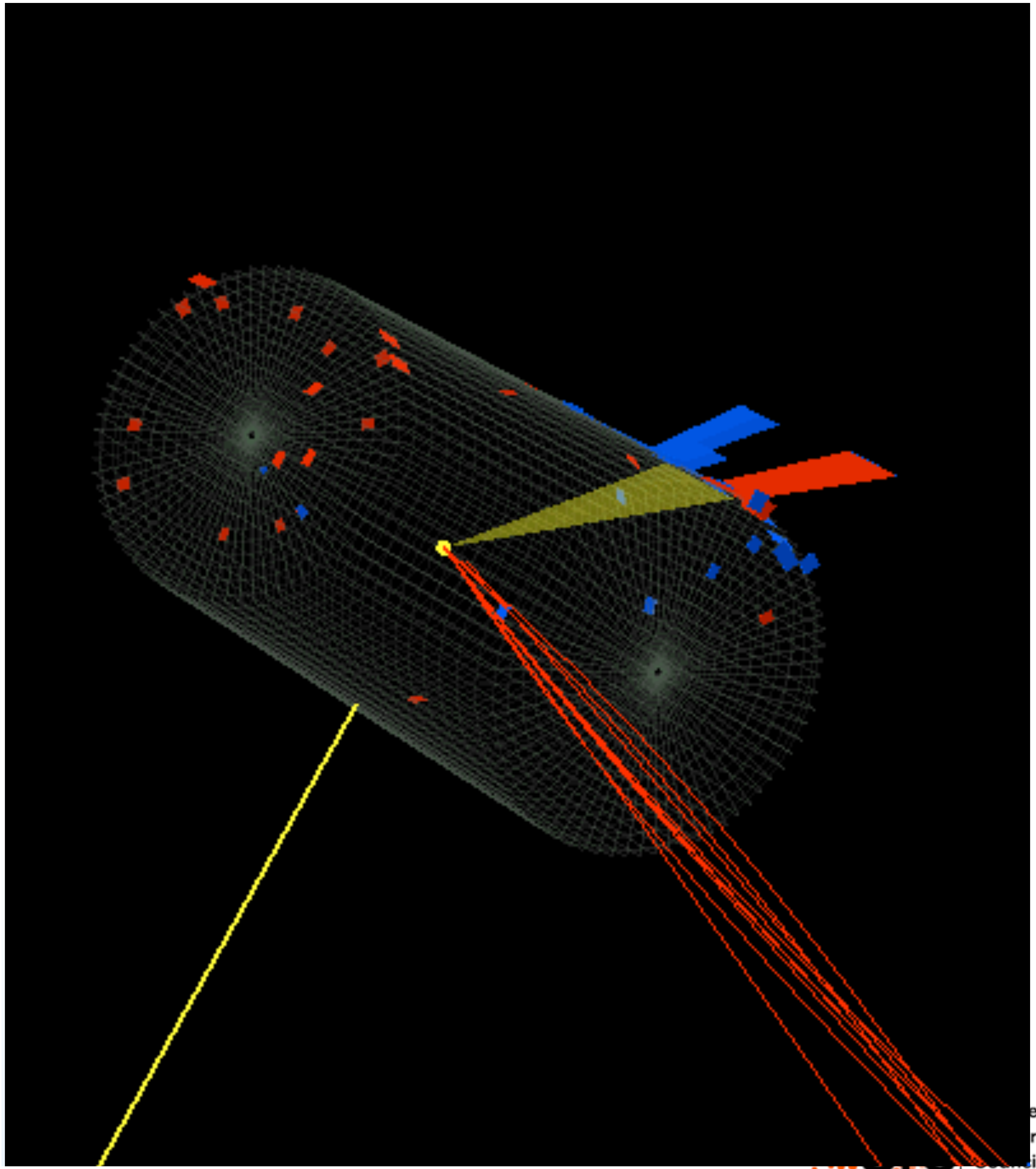
CMS Exotica hotline leads hunt for exotic particles

Strangers in the dark: they meet, make contact, and break away with force, careless of what they leave behind. At midnight each night, snapshots of these frenzied chance encounters are collected for curious eyes. In the morning, those onlookers reconstruct the story that each image tells, tracing the mysterious paths born from a fateful meeting.

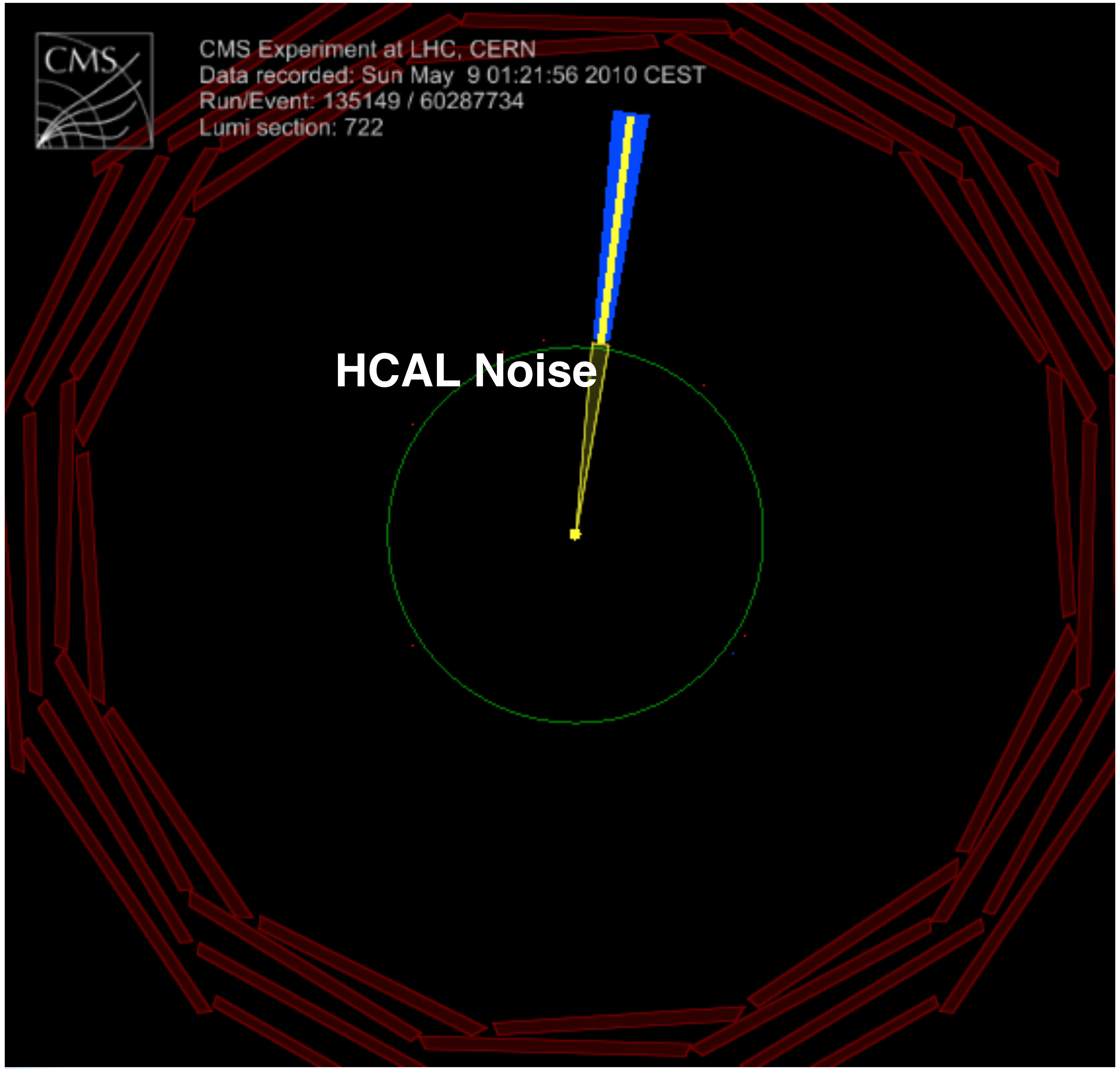
This is the CMS exotica hotline, and no, it's not a 900 number.

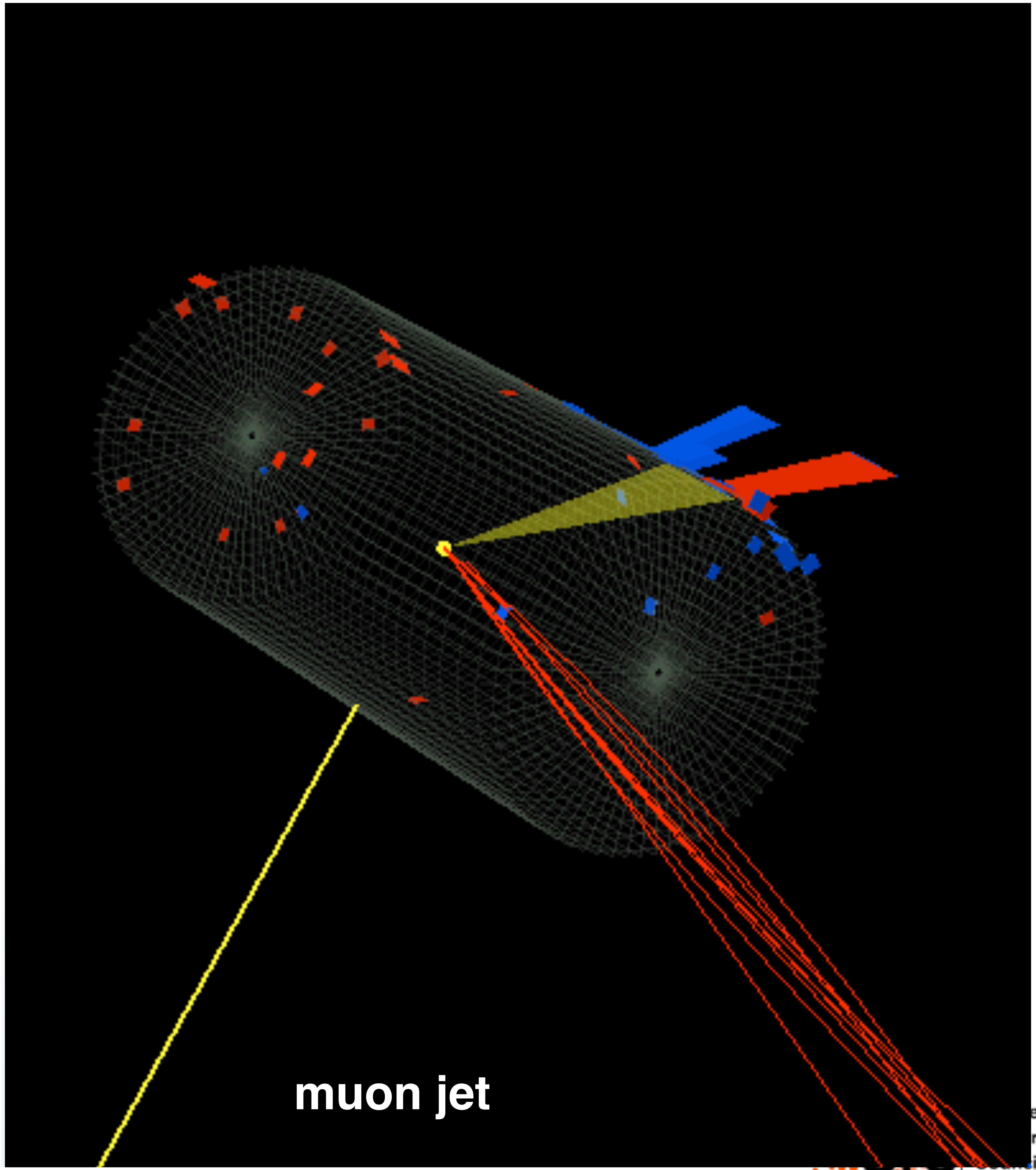
What was "found"



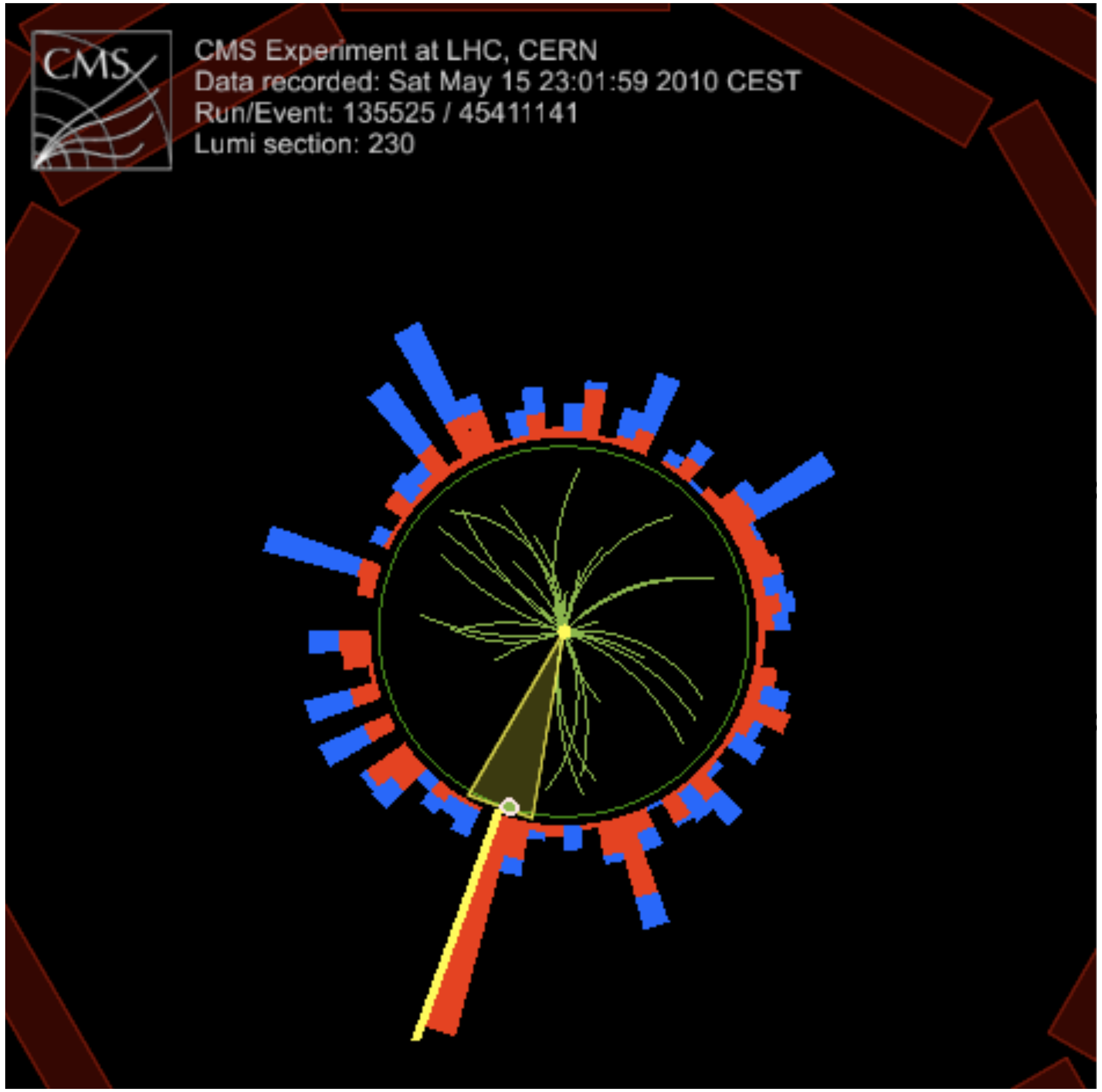


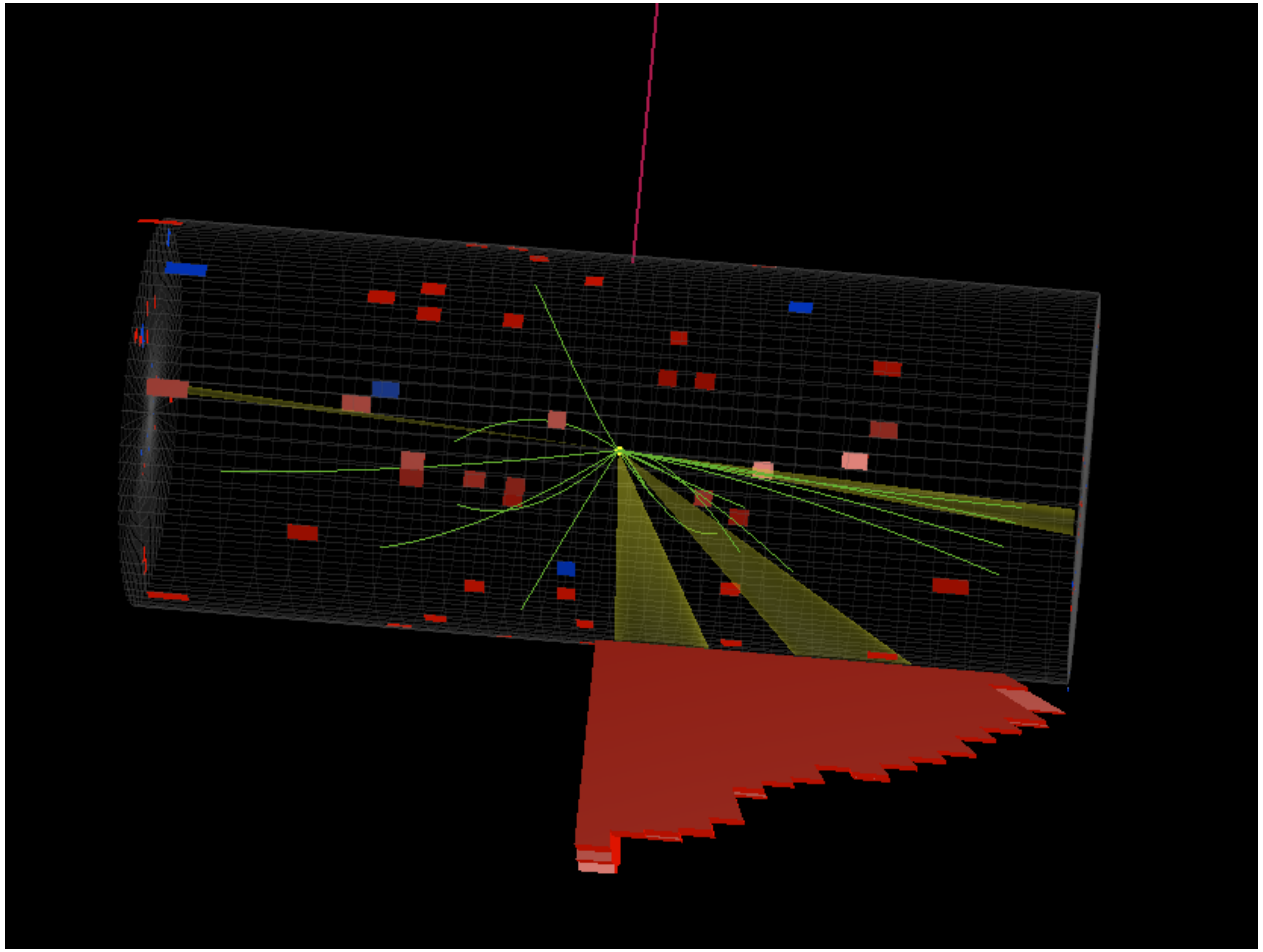
What was "found"





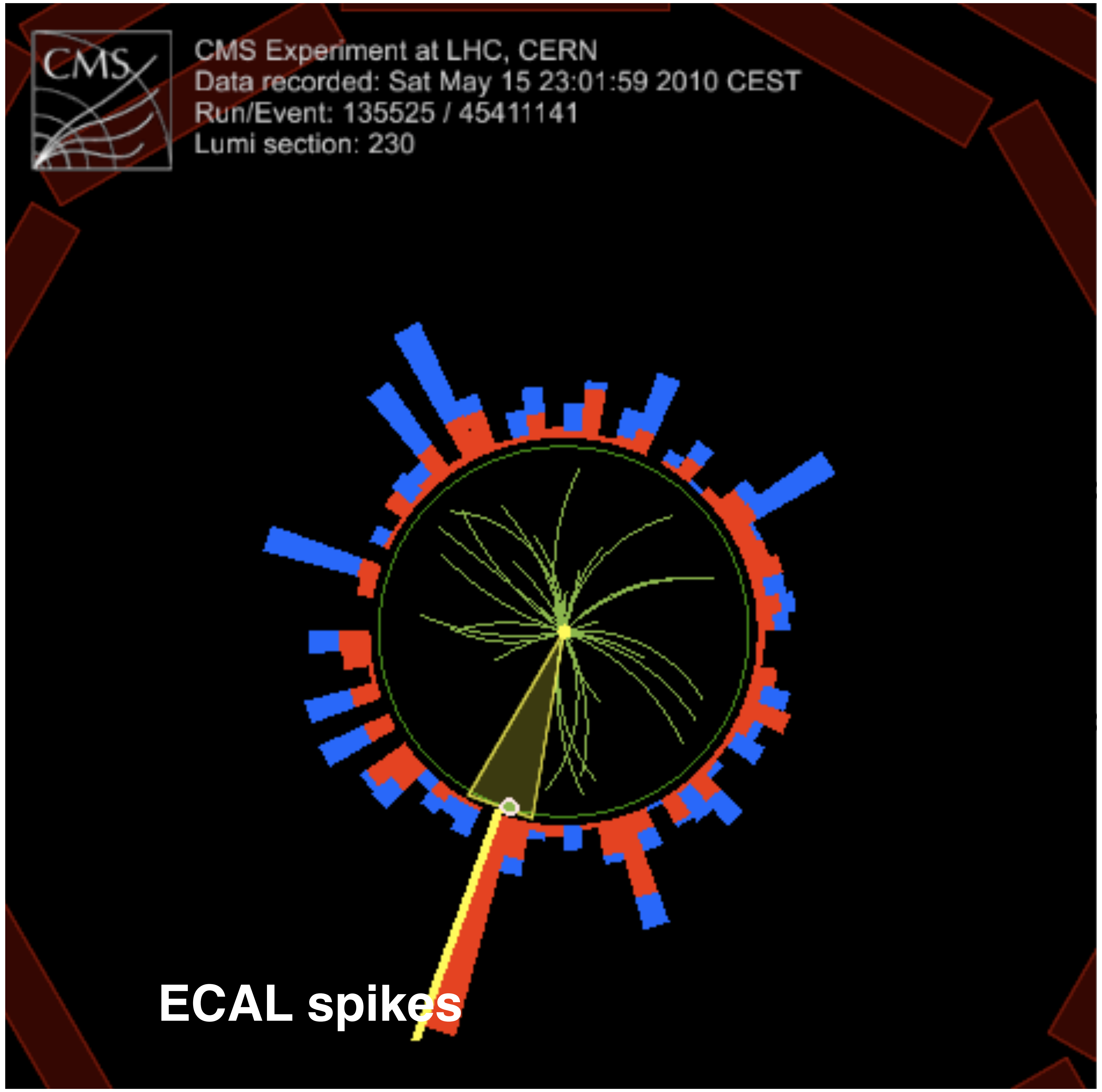
What was "found"

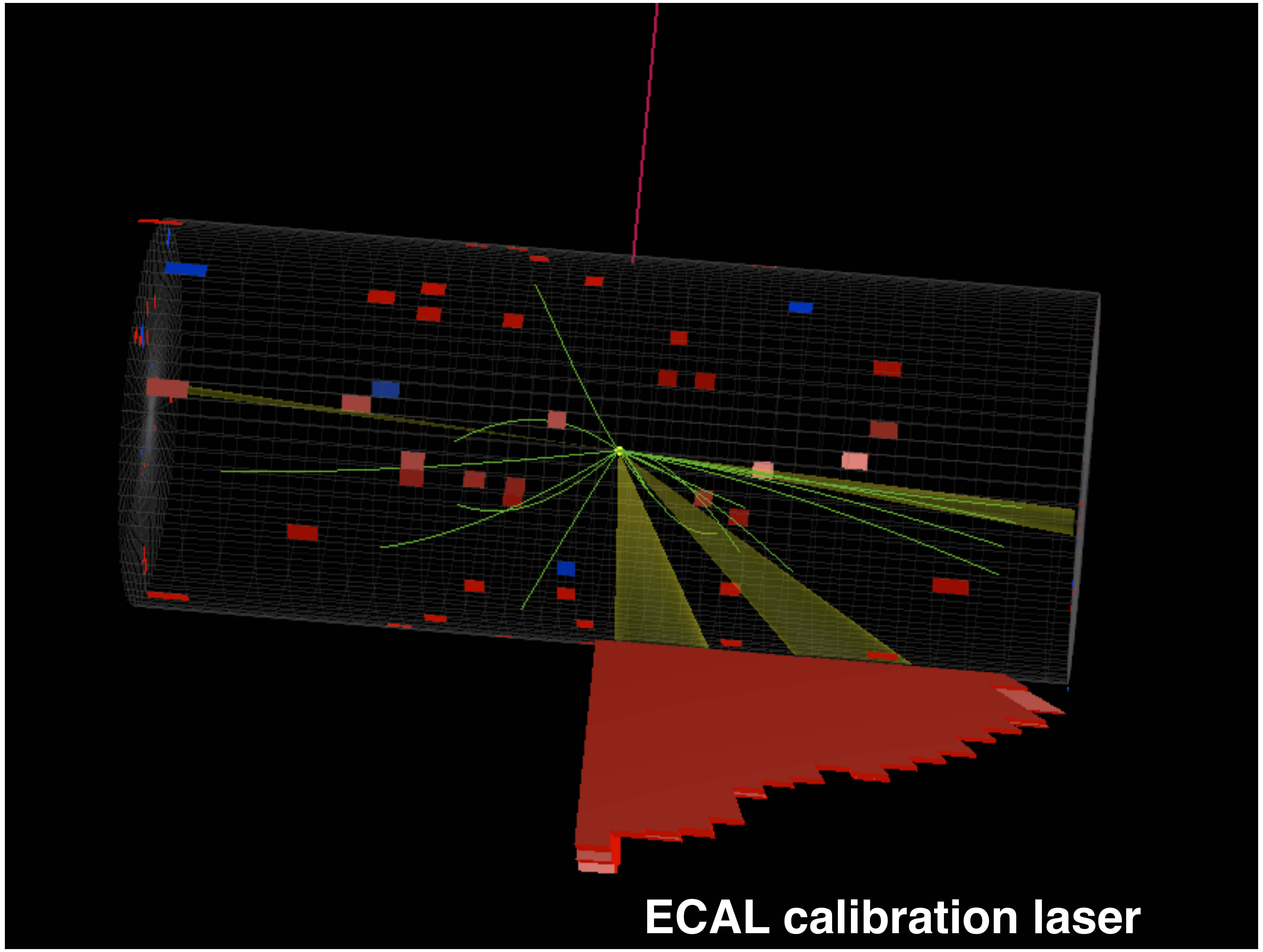




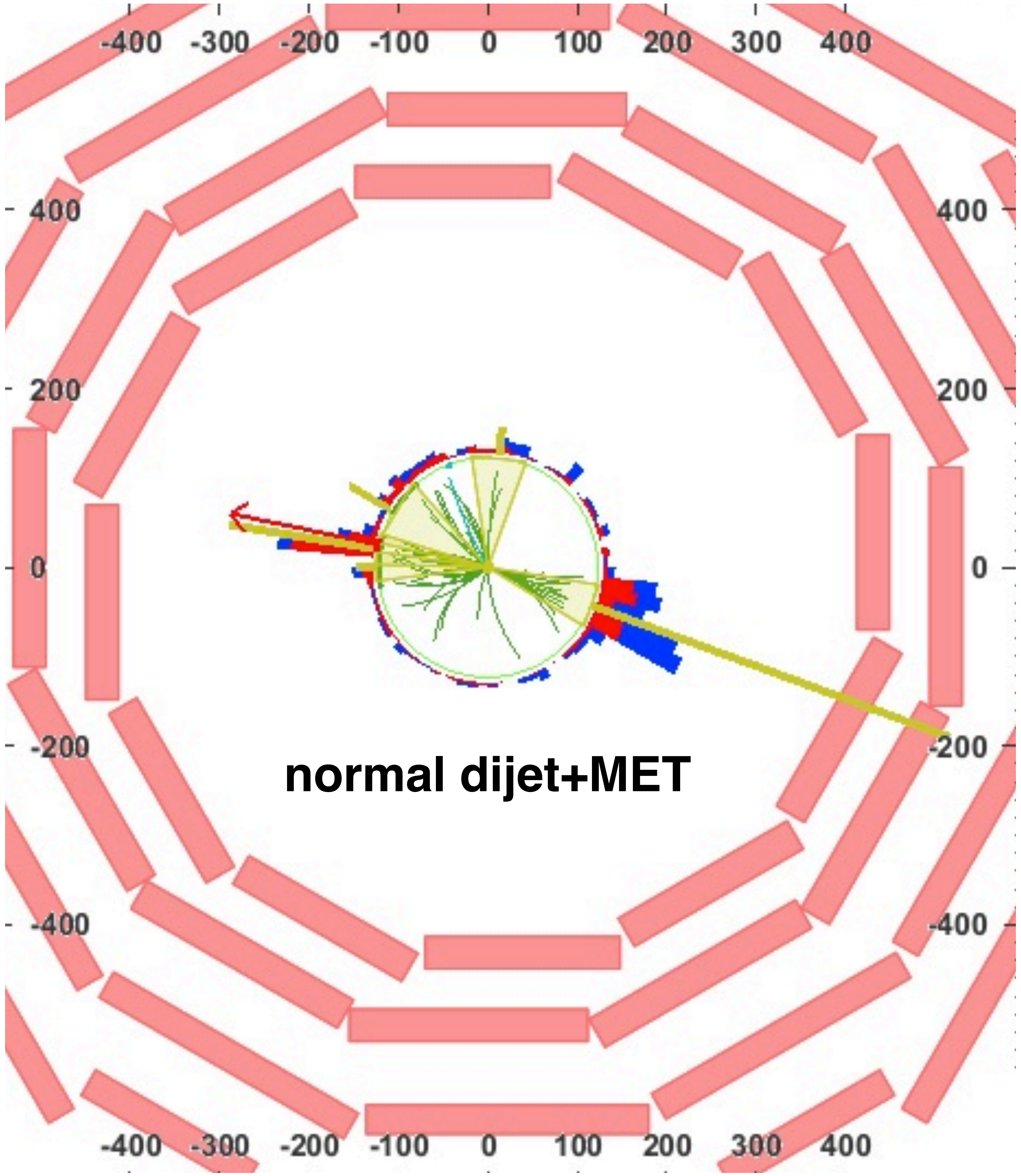
European Research Council

What was "found"



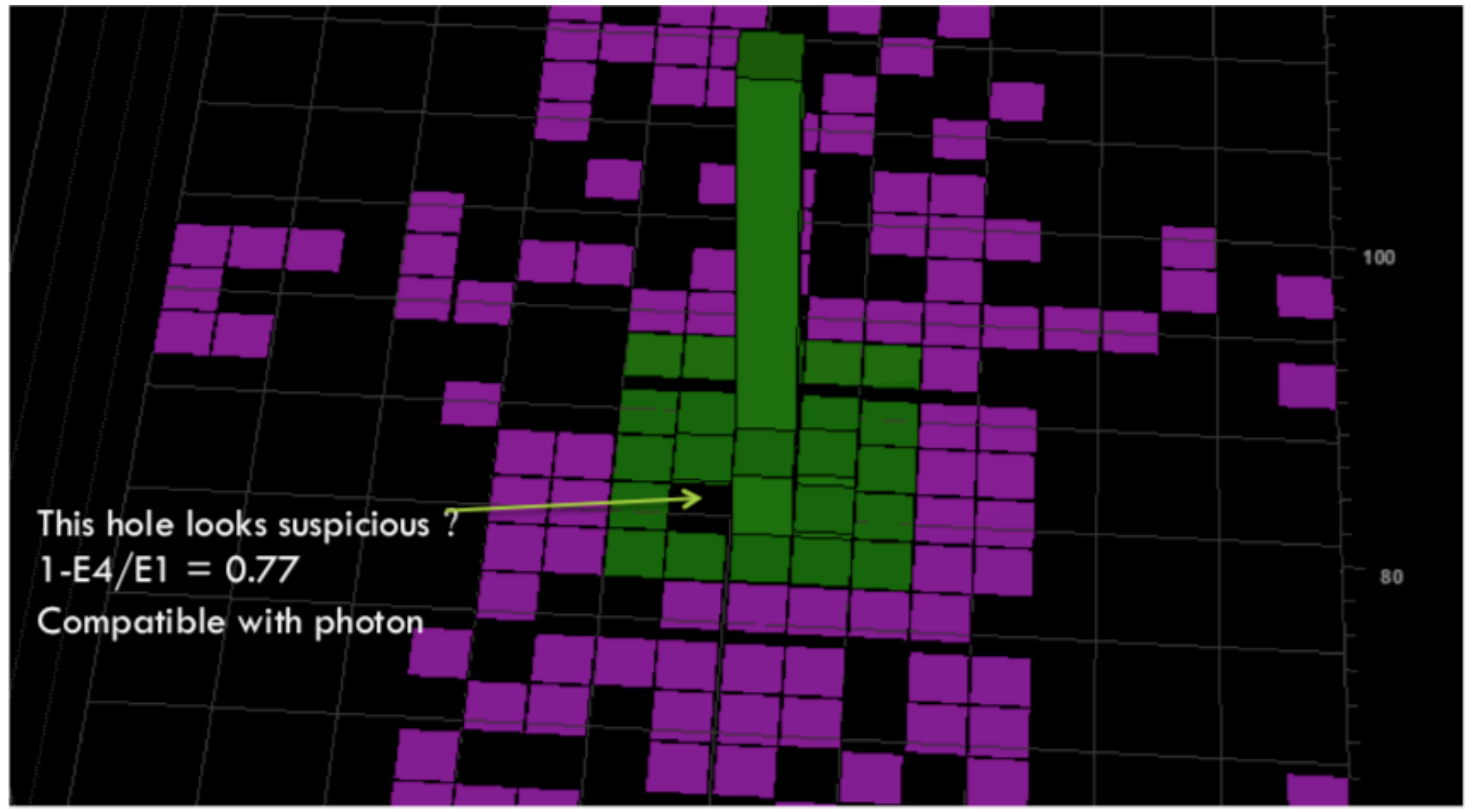


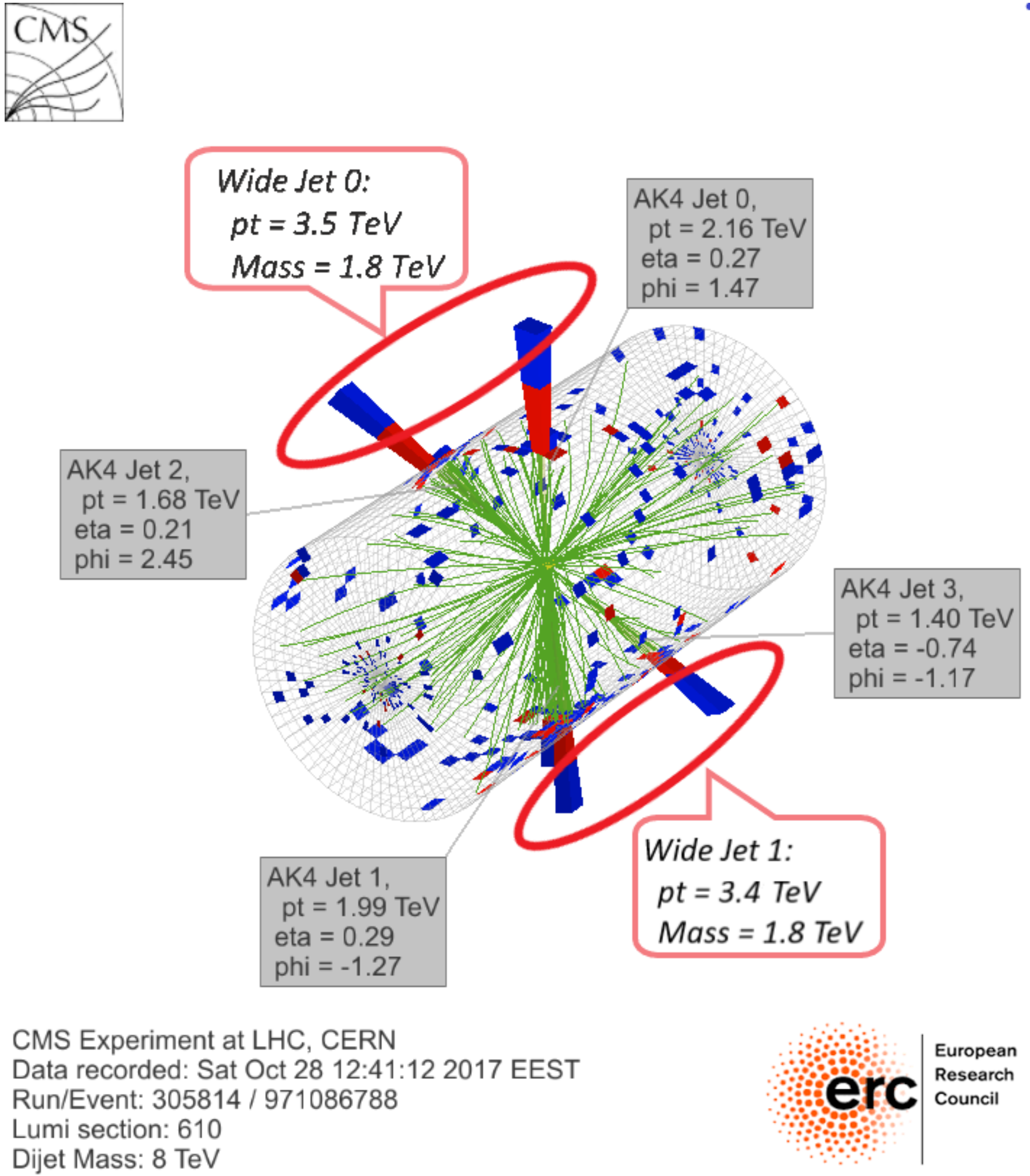
What was "found"



Exotica hotline today

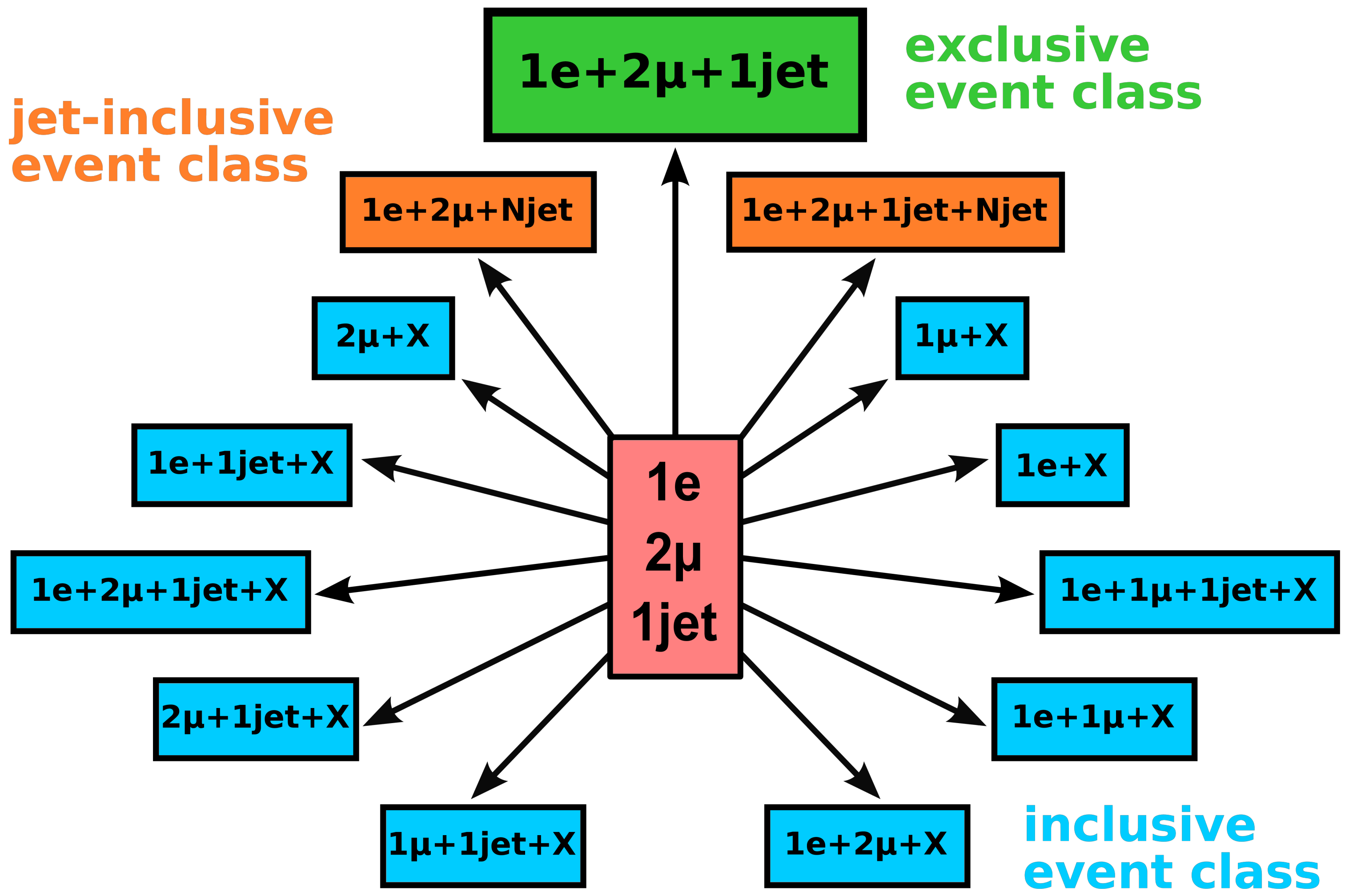
- Given the integrated luminosity and the typical turnaround time for an analysis, physics can wait 48h
- Anomalous events are now looked for using standard data stream
- Instead, exotica hotline is still useful at startup, to early catch problems with reconstruction (e.g., with MET)
- The early-alert system was retired in 2015, just after Run II started

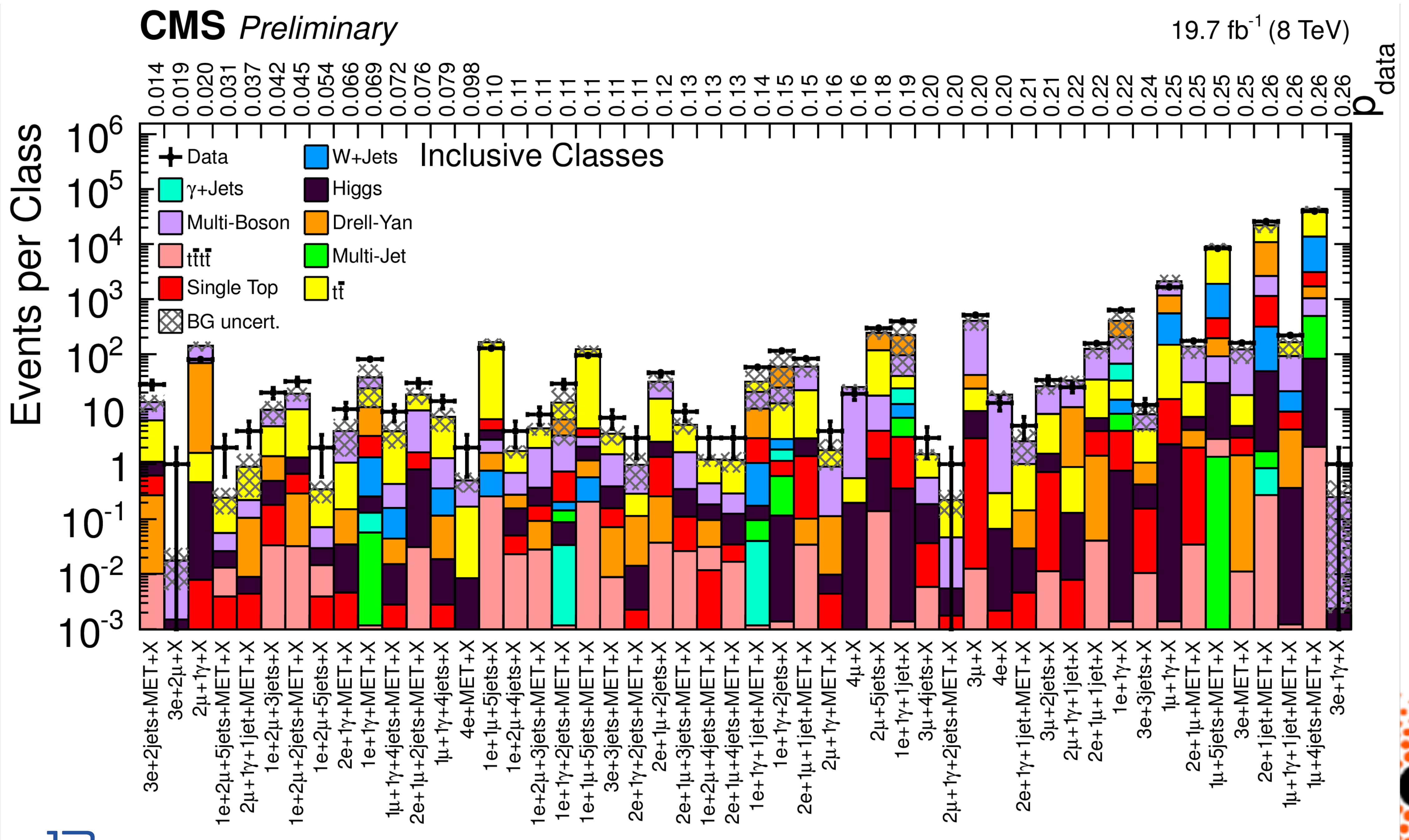




What we do today

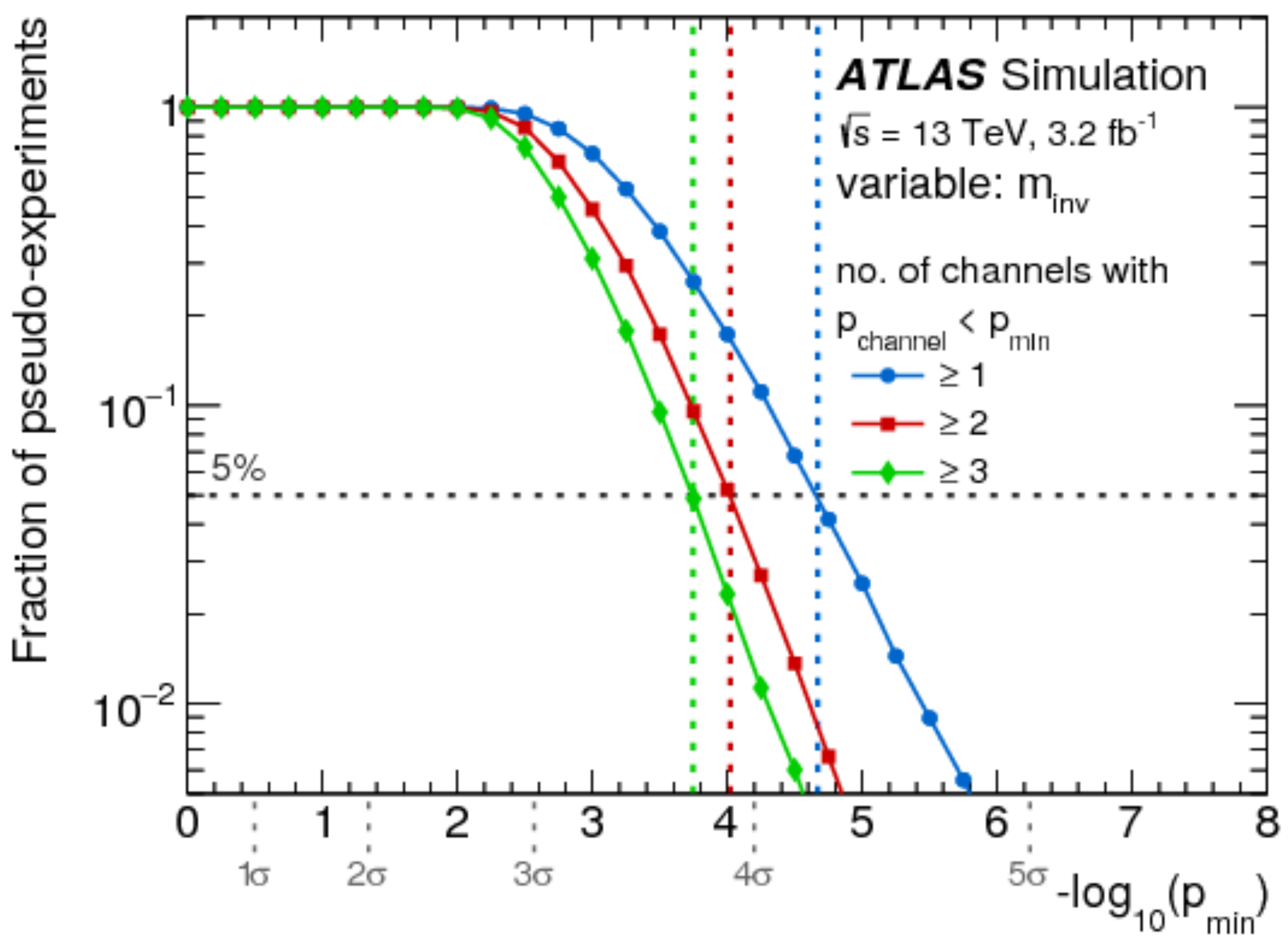
- Model independent analyses are performed at colliders since Tevatron
- Plot a lot of histograms for data and compare them to what you expect on Monte Carlo
- Look for a discrepancy
- If you find it, try to exclude any instruments-driven explanation

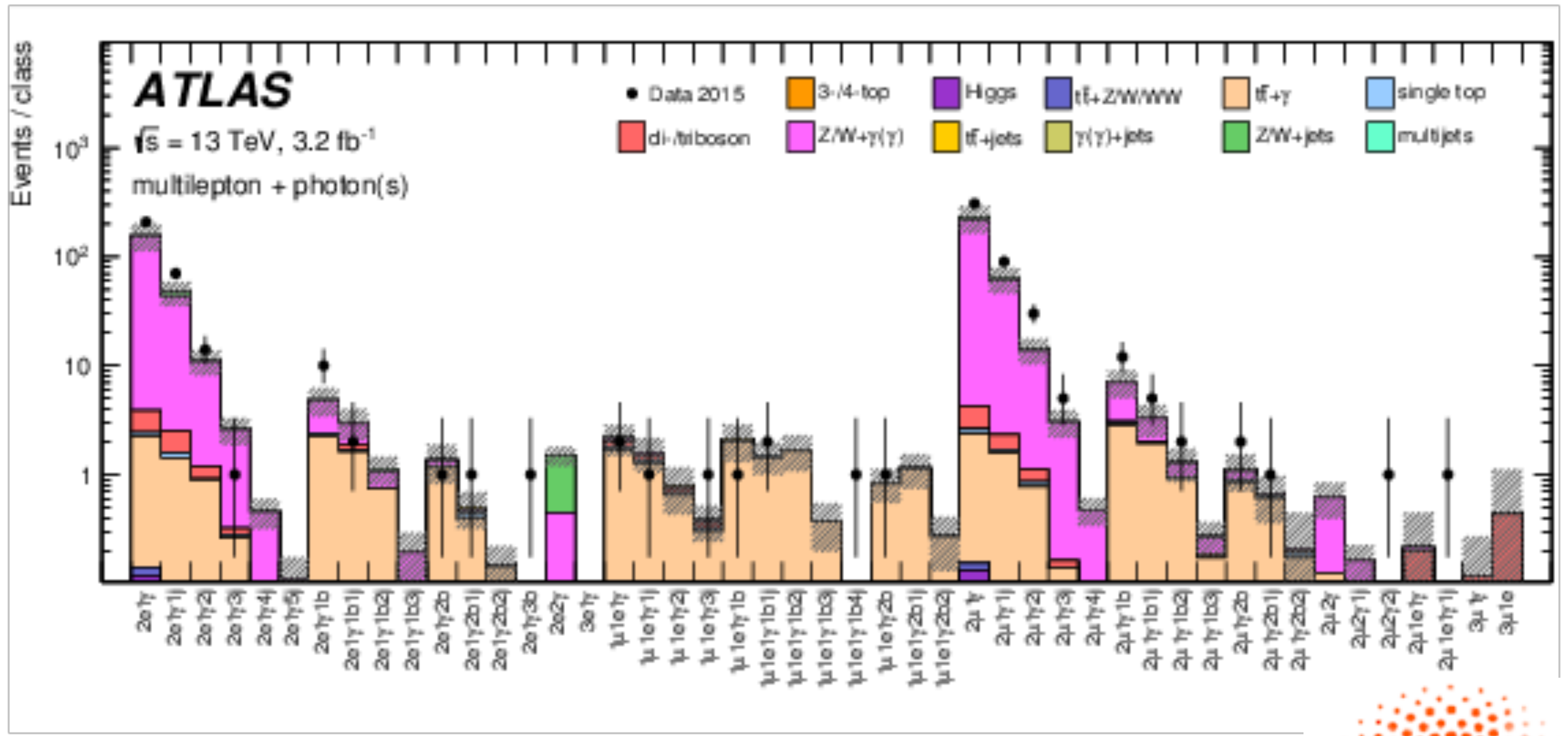




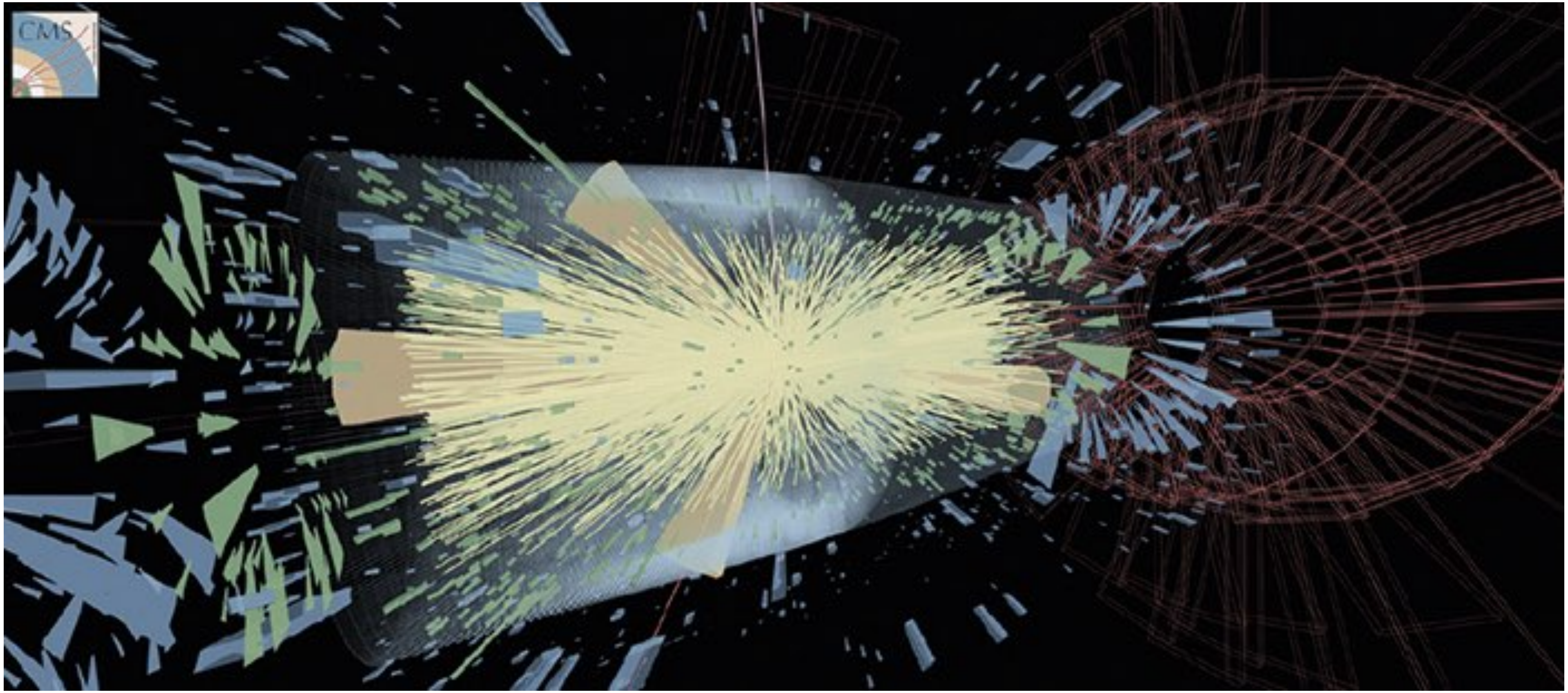
Trial factor

- *The issue with this approach is the trial factor (look-elsewhere effects)*
    - *one sets a p-value threshold α (e.g., 5%) to define the alarm*
    - *a fraction ~ α of the bins will be off even in absence of an anomaly*
    - *for large number of bins, this dilutes discovery power*
- *ATLAS came out with a proposal: use the analysis to identify an excess, but establish the significance with a traditional method on an independent dataset*
- *This is the same spirit we have in mind for what follows*



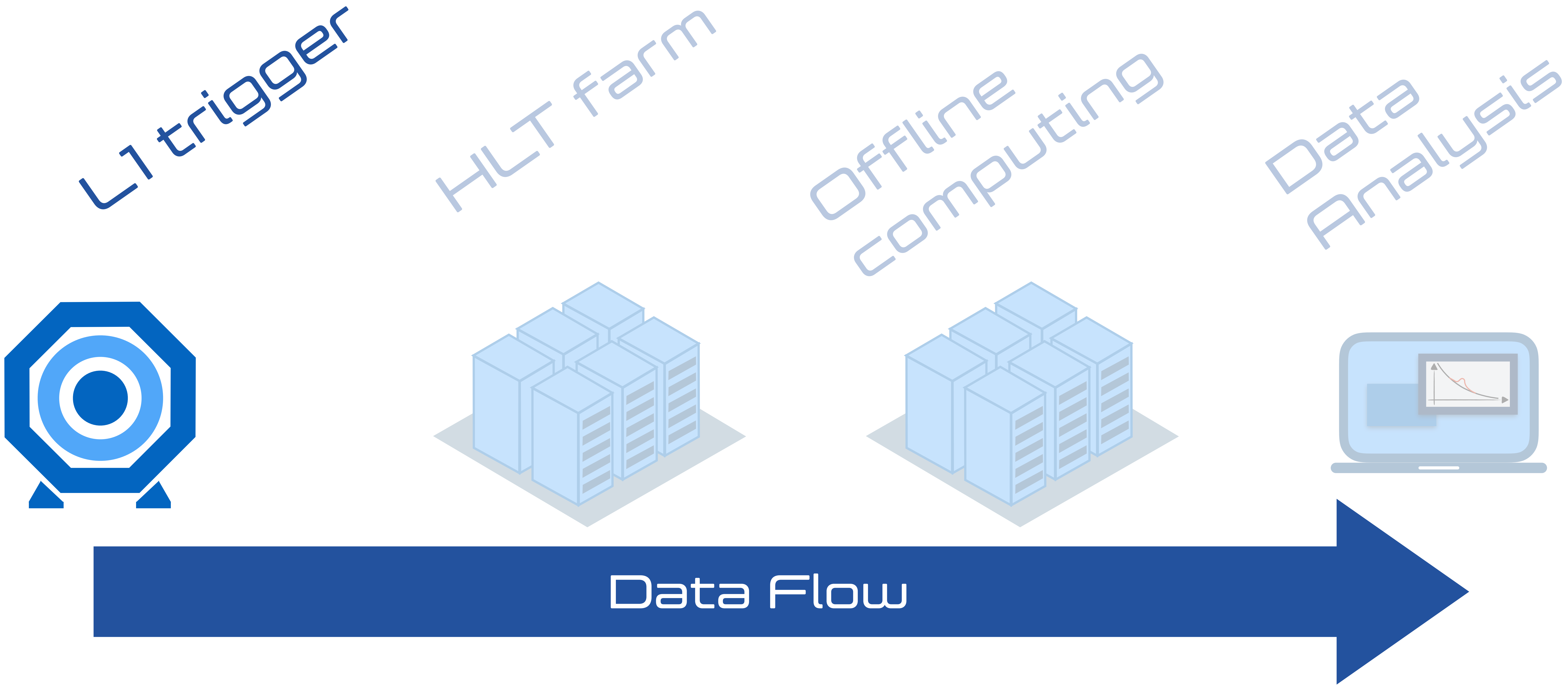


CERN-EP-2018-070



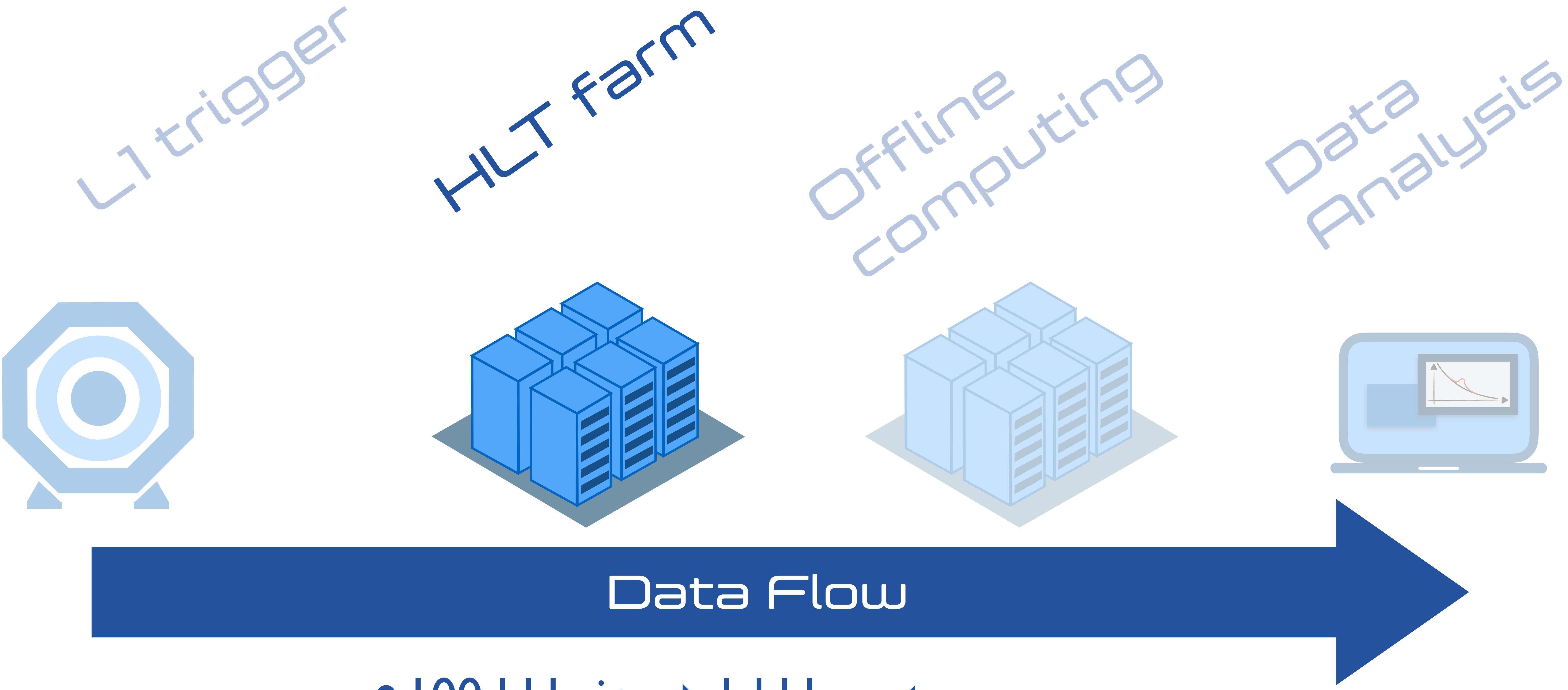
The LHC Big Data Problem

The LHC Big Data problem



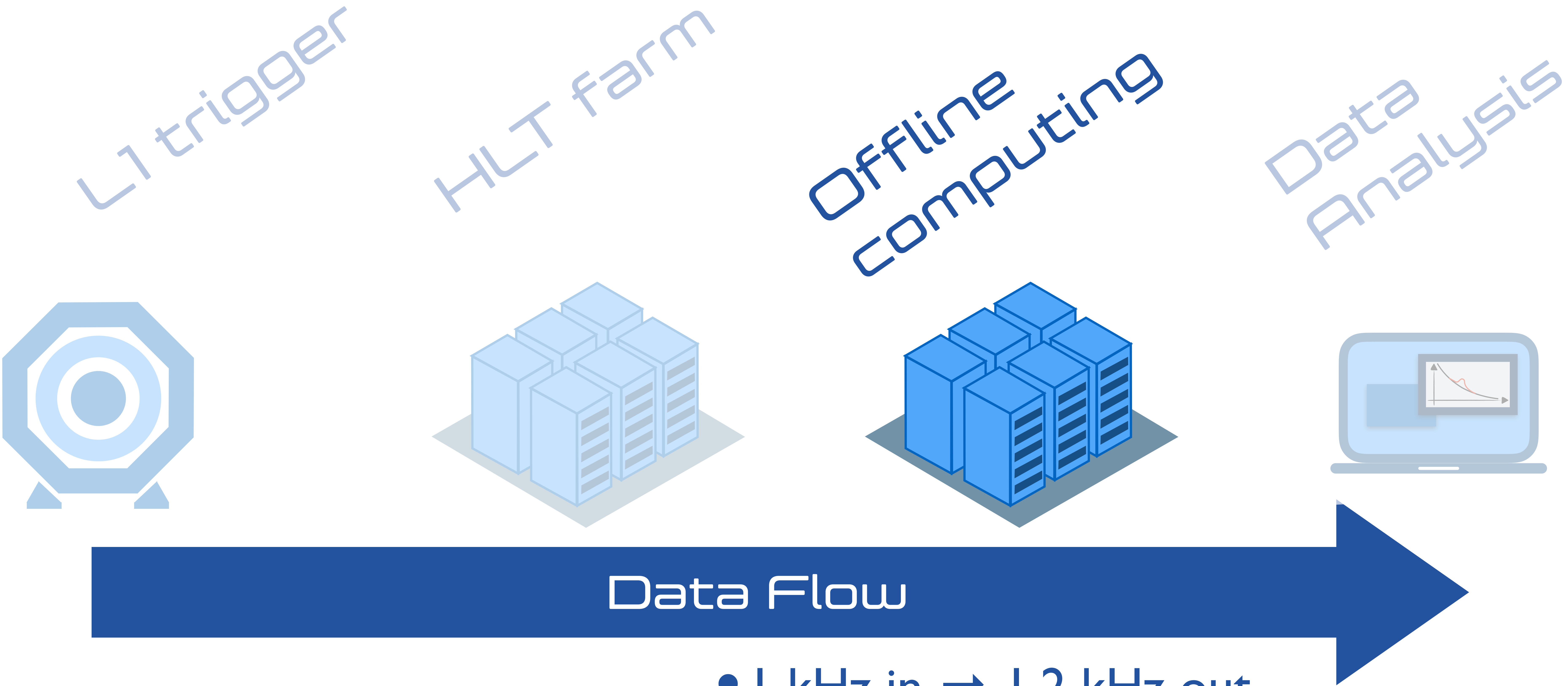
- 40 MHz in → 100 kHz out
- ~ 500 KB / event
- Processing time: ~10 μs
- Based on coarse local reconstructions
- FPGAs / Hardware implemented

The LHC Big Data problem



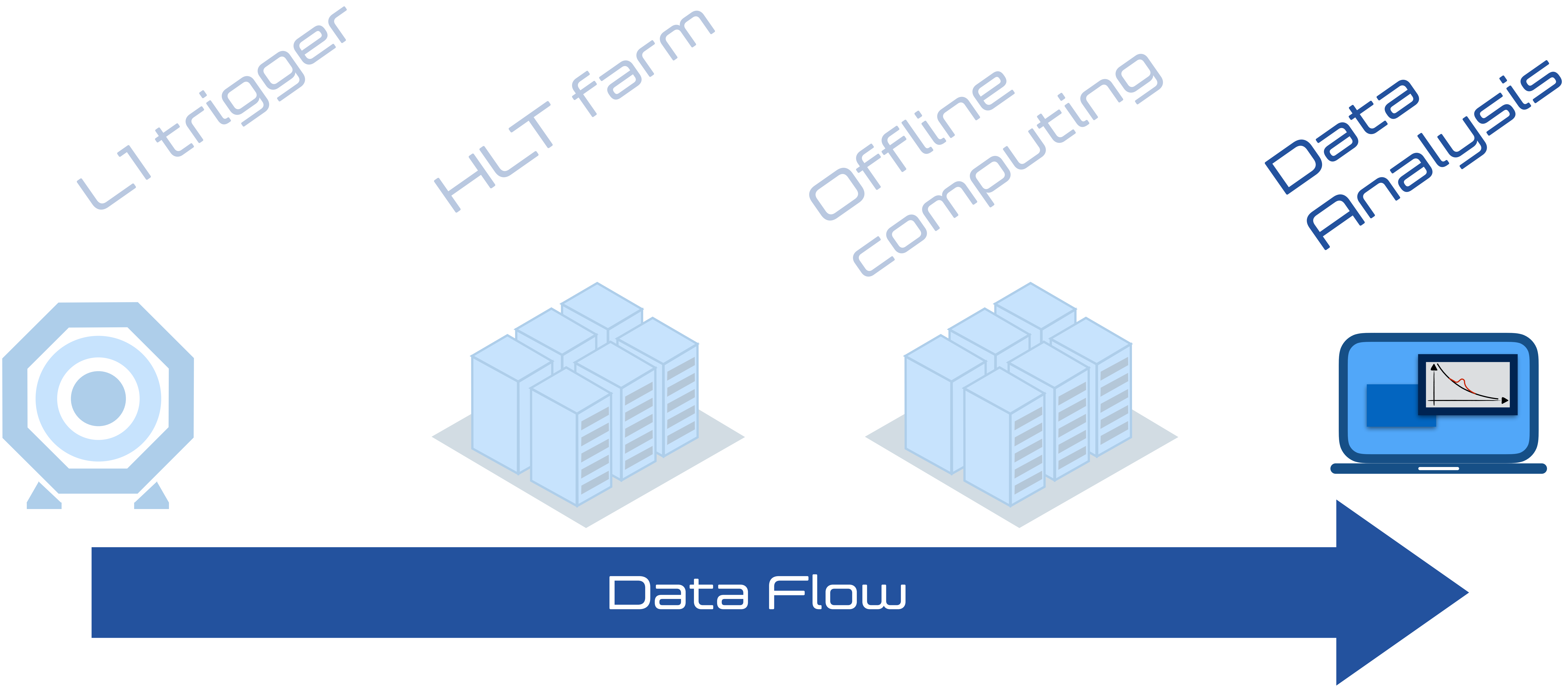
- 100 kHz in → 1 kHz out
- ~ 500 KB / event
- Processing time: ~30 ms
- Based on simplified global reconstructions
- Software implemented on CPUs

The LHC Big Data problem



- 1 kHz in → 1.2 kHz out
- ~ 1 MB / 200 kB / 30 kB per event
- Processing time: ~20 s
- Based on accurate global reconstructions
- Software implemented on CPUs

The LHC Big Data problem



- Up to ~ 500 Hz In → 100-1000 events out
- <30 KB per event
- Processing time irrelevant
- User-written code + centrally produced selection algorithms

New Physics Mining

- With such a tight selection to be made, the risk of discarding events is not negligible
  - Particularly because we found no new physics in the data we collected
  - The problem starts with the need to assume a specific model, to then make sure that we trigger on it. What if we never considered the right model?
- We would like to deploy in the trigger system an algorithm that selects anomalous events
  - Data-driven approach (data mining) that could guide the next generation of new-physics searches
  - We don't want to define what "anomaly" means based on BSM hypotheses (as it was done with the hotline)
  - We would like to do this using Deep Learning

Anomaly Detection

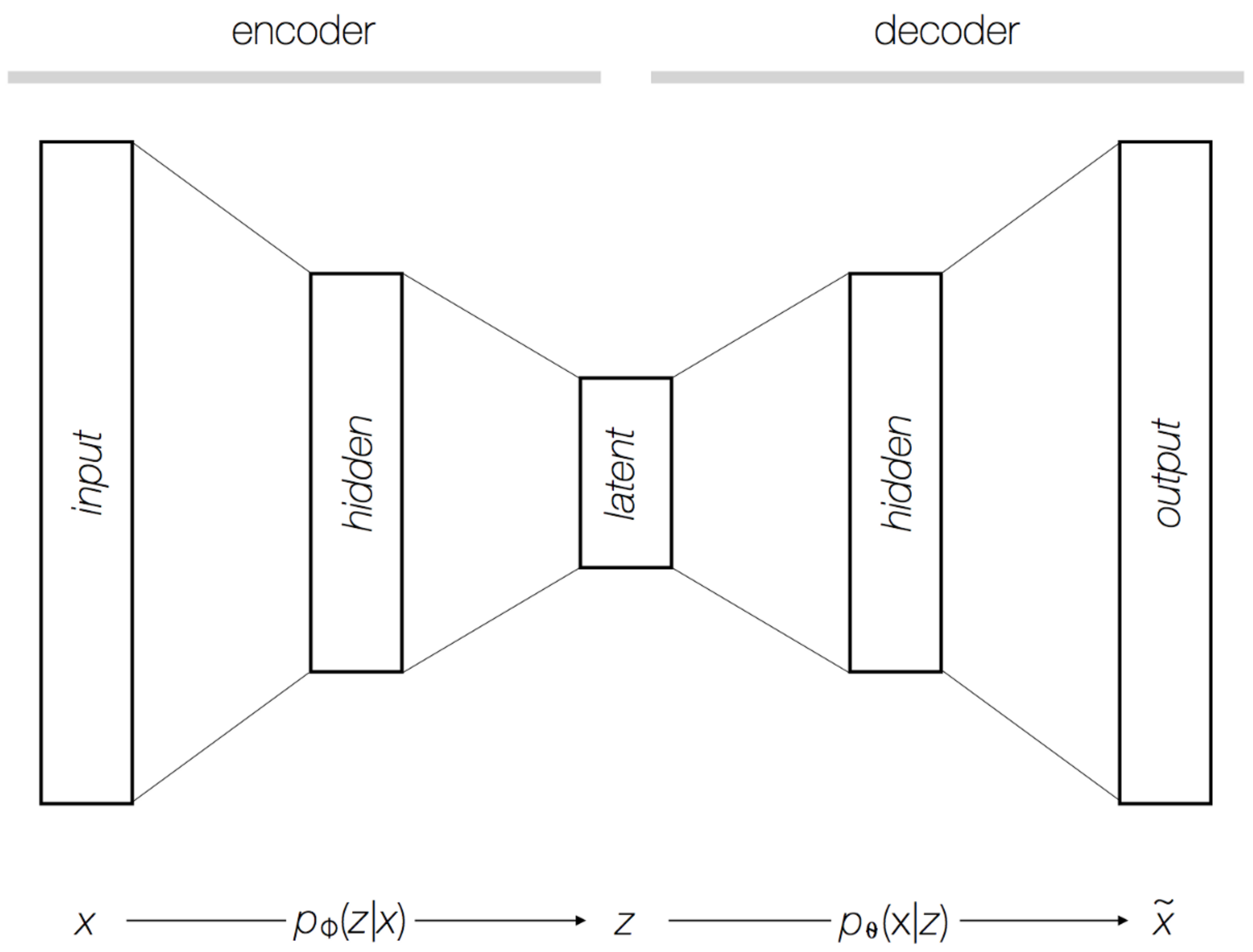
Autoencoders in a nutshell

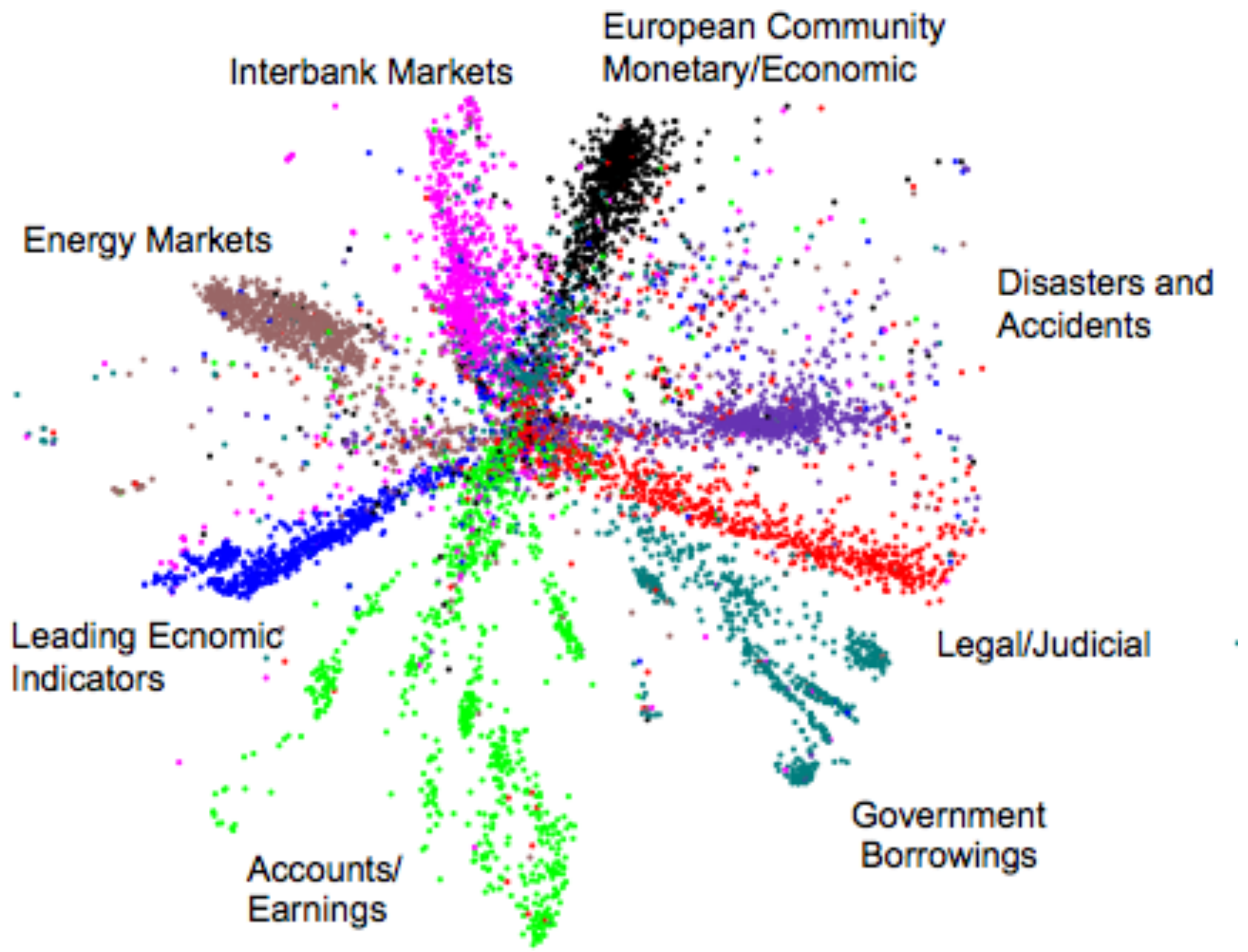
- Autoencoders are compression-decompression algorithms that learn to describe a given dataset in terms of points in a lower-dimension latent space

- 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

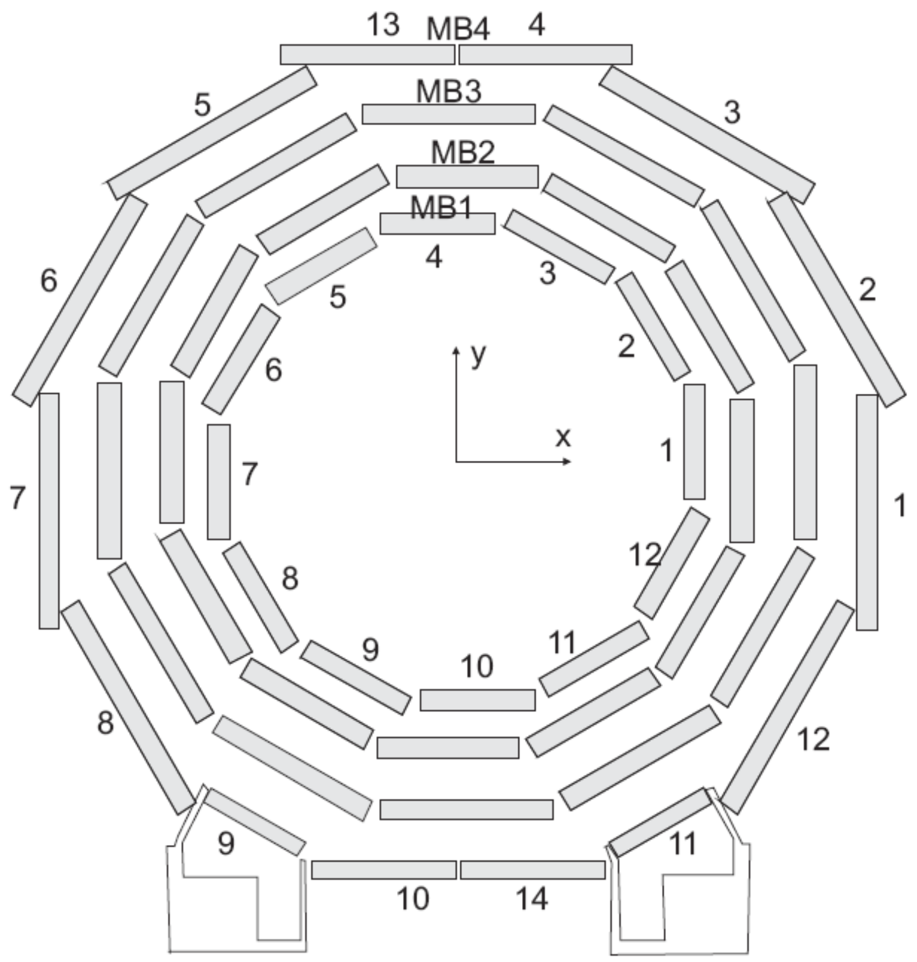
- One can define anomalous any event whose decompressed output is “far” from the input, in some metric (e.g., the metric of the auto-encoder loss)

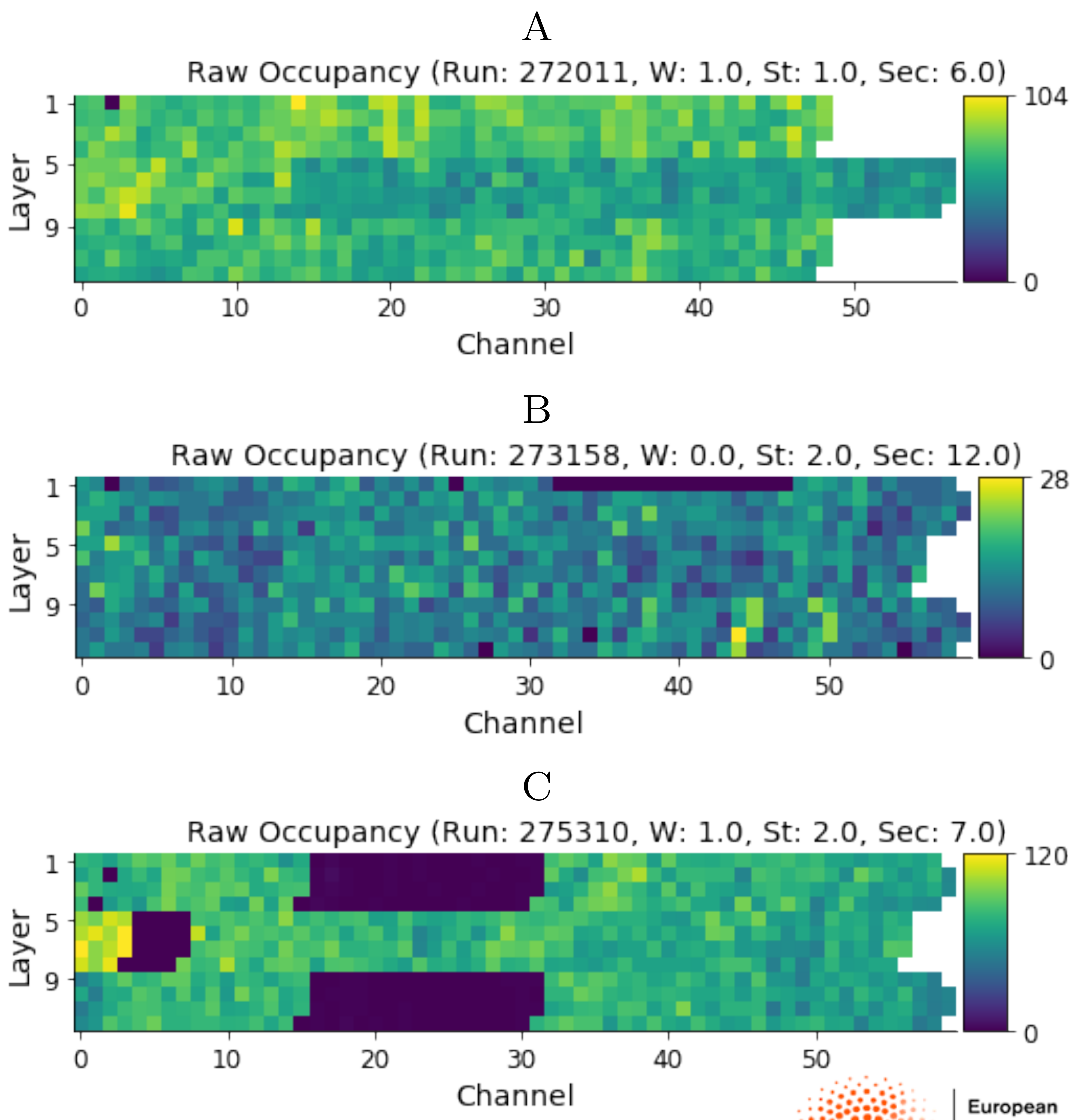




Example: Data Quality Monitoring

- *When taking data, >1 person watches for anomalies in the detector 24/7*
- *At this stage no global processing of the event*
- *Instead, local information from detector components available (e.g., detector occupancy in a certain time window)*





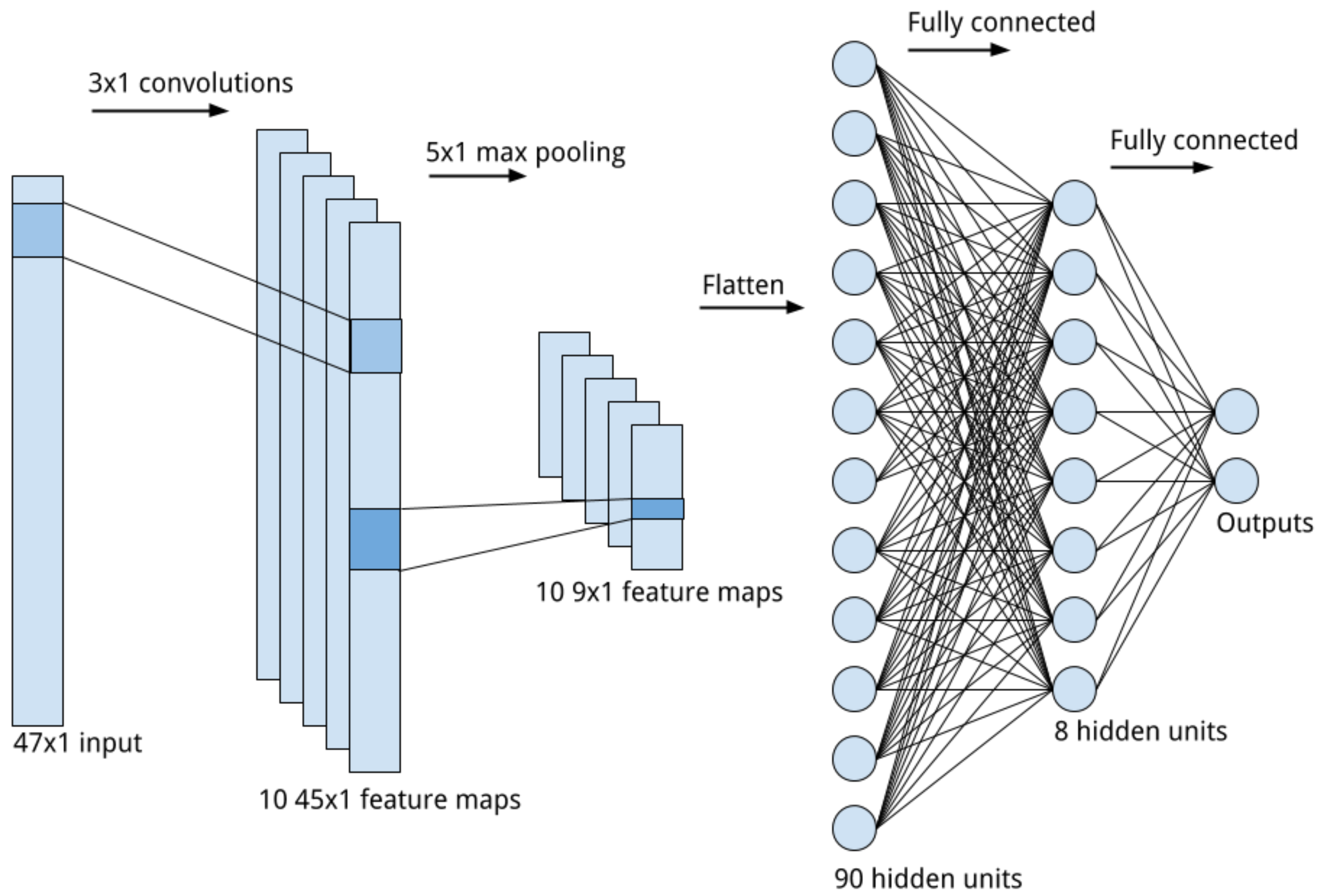
[Pol, G. Cerminara, C. Germain, MP and A. Seth arXiv:1808.00911](https://arxiv.org/abs/1808.00911)

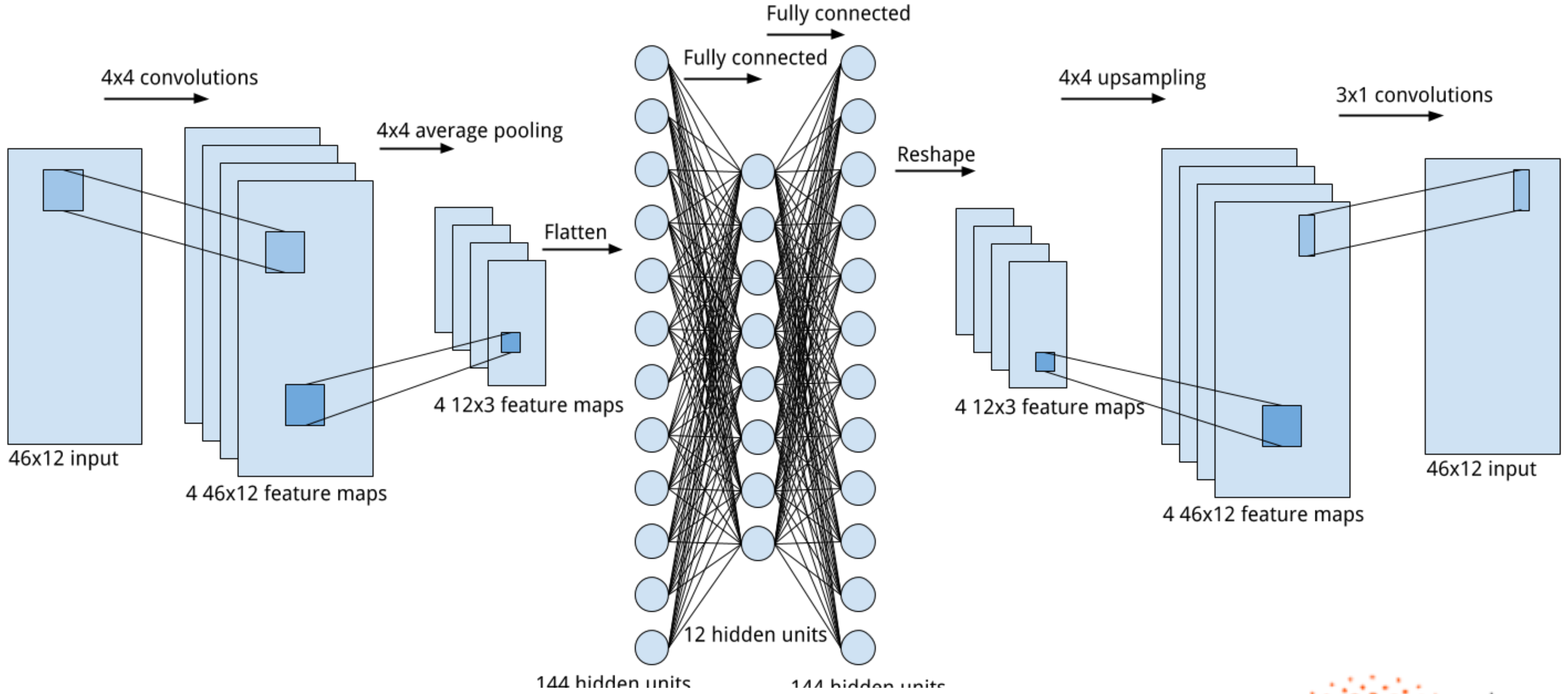
European Research Council

Example: Data Quality Monitoring

- Given the nature of these data, ConvNN are a natural analysis tool. Two approaches pursued
  - Classify good vs bad data. Works if failure mode is known
  - Use autoencoders to assess data “typicality”. Generalises to unknown failure modes

A. Pol et al., to appear soon



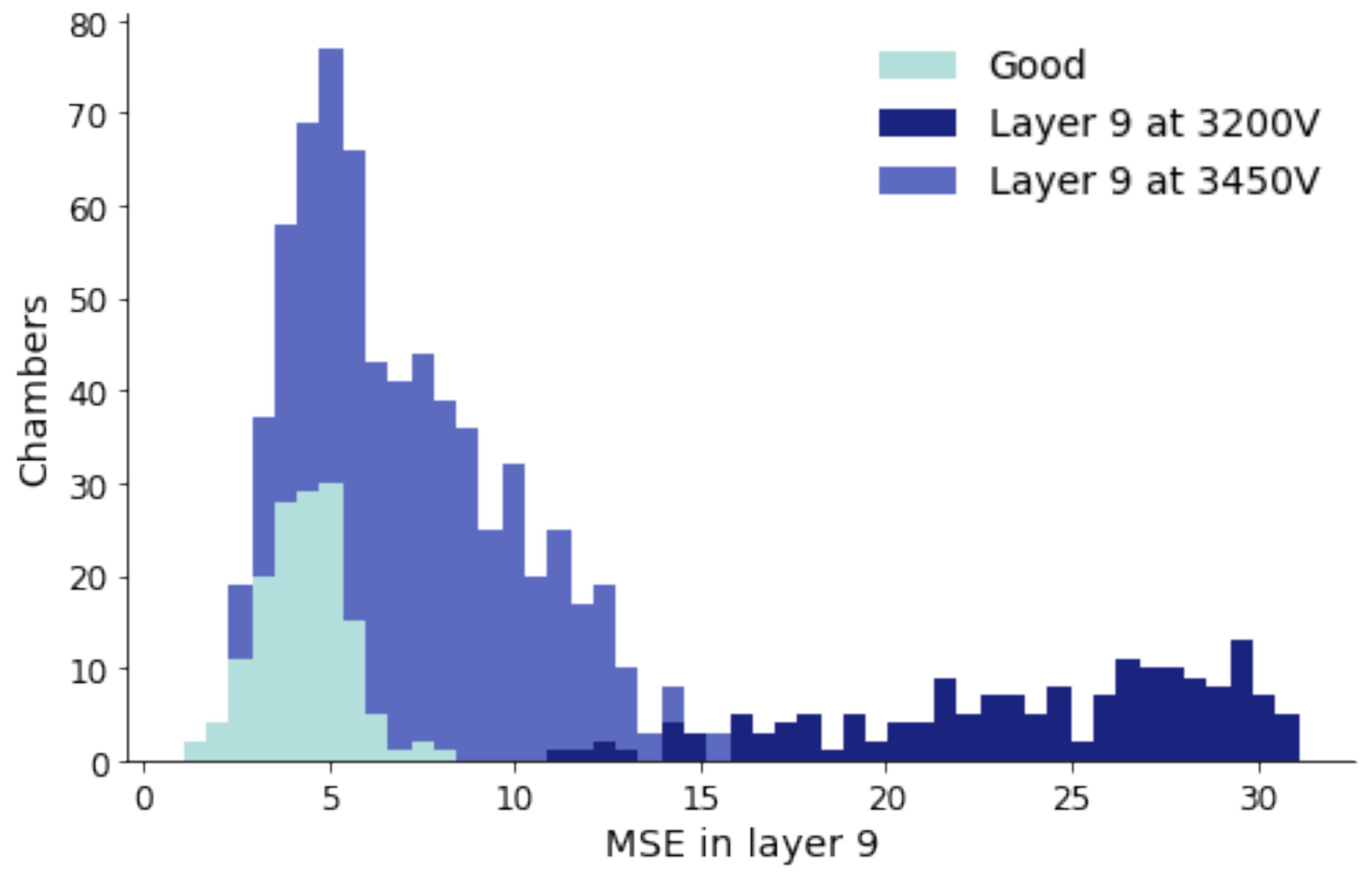


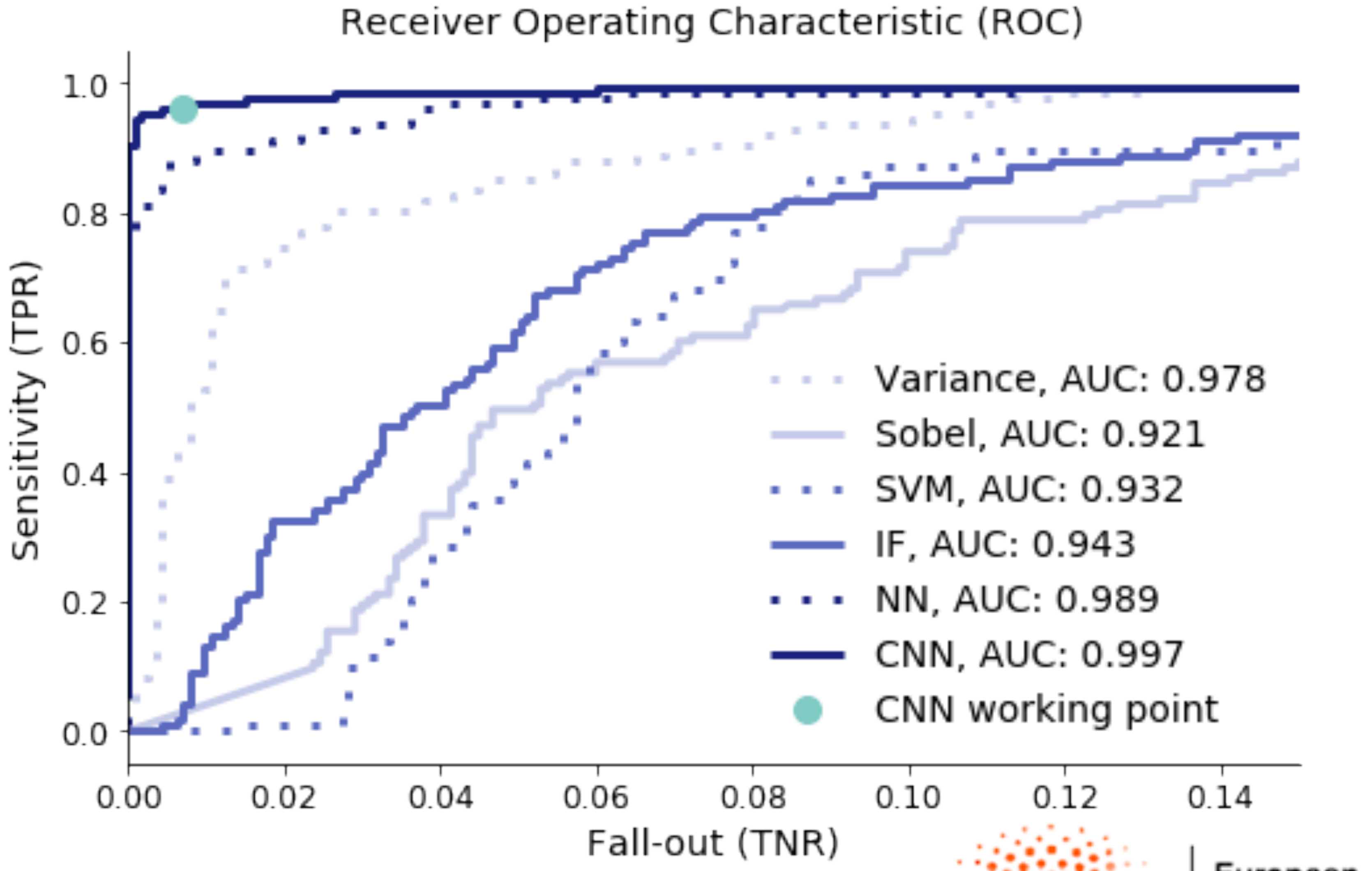
[Pol, G. Cerminara, C. Germain, MP and A. Seth arXiv:1808.00911]

European Research Council

Example: Data Quality Monitoring

- *Given the nature of these data, ConvNN are a natural analysis tool. Two approaches pursued*
  - *Classify good vs bad data. Works if failure mode is known*
  - *Use autoencoders to assess data "typicality". Generalises to unknown failure modes*





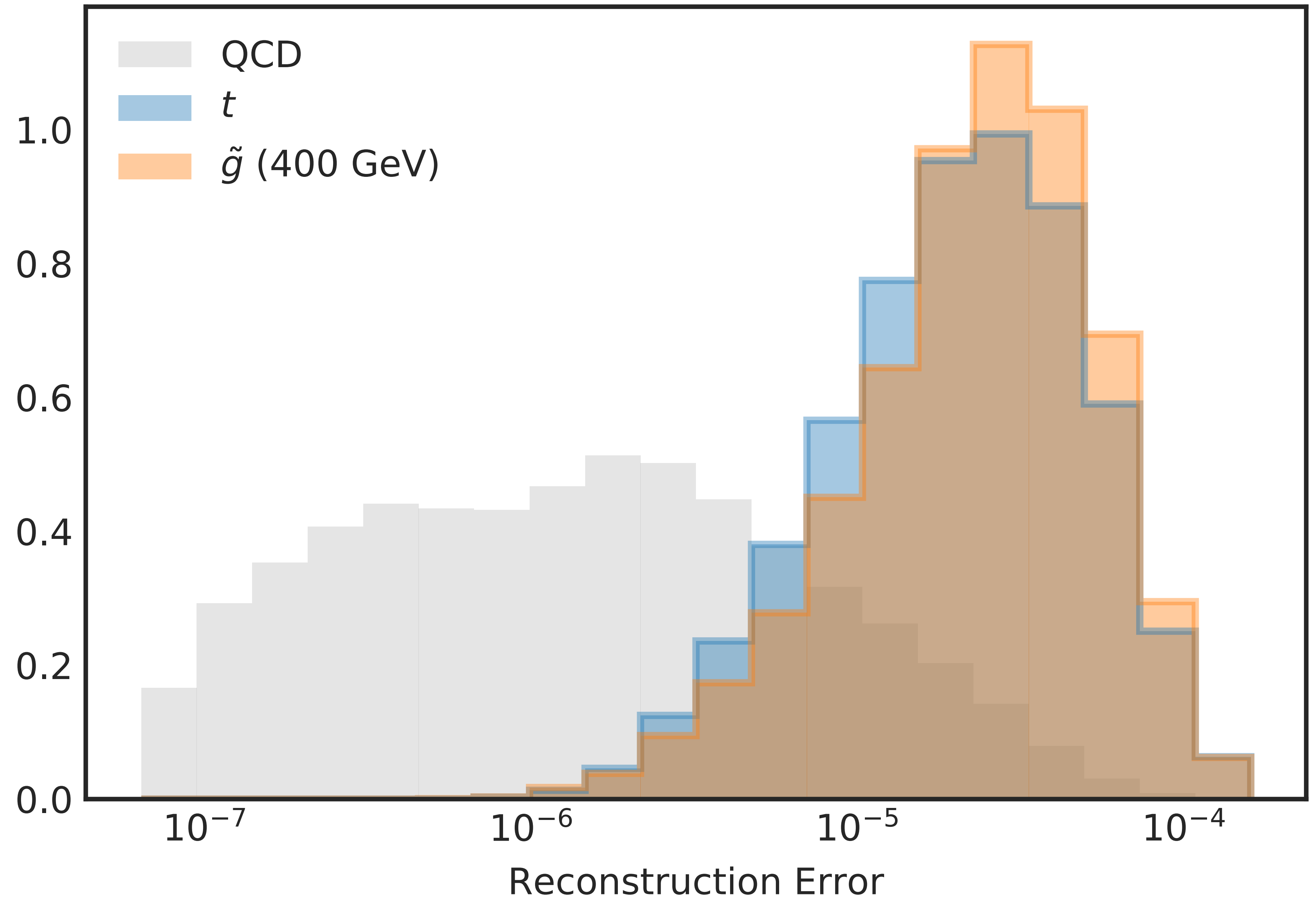
A. Pol et al., to appear soon

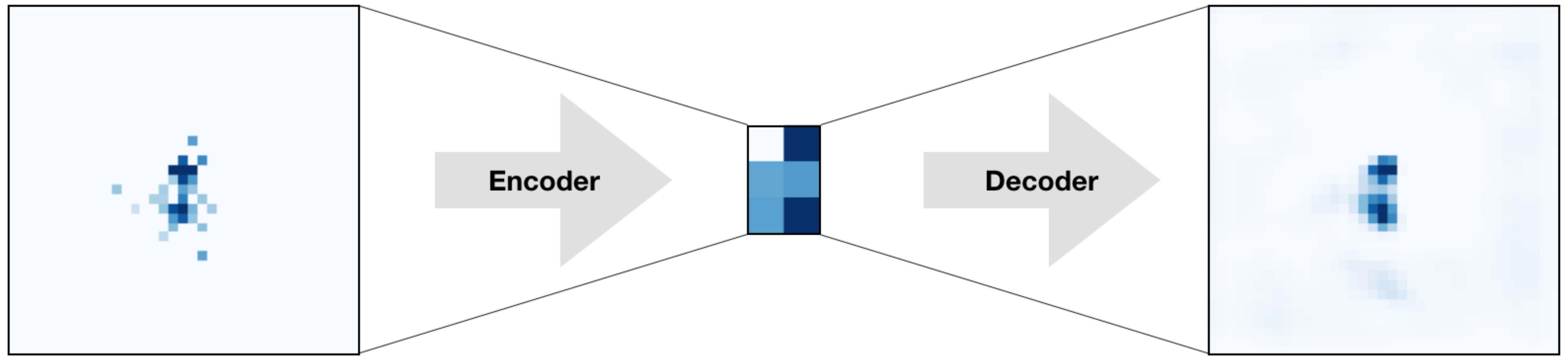
[Pol, G. Cerminara, C. Germain, MP and A. Seth arXiv:1808.00911]()

European Research Council

Example: Jet autoencoders

- *Idea applied to tagging jets, in order to define a QCD-jet veto*
- *Applied in a BSM search (e.g., dijet resonance) could highlight new physics signal*
- *Based on image and physics-inspired representations of jets*





$$\tilde{k}_j = \begin{pmatrix} \tilde{k}_{0,j} \\ \tilde{k}_{1,j} \\ \tilde{k}_{2,j} \\ \tilde{k}_{3,j} \end{pmatrix} \xrightarrow{\text{LoLa}} \begin{pmatrix} \tilde{k}_{0,j} \\ \tilde{k}_{1,j} \\ \tilde{k}_{2,j} \\ \tilde{k}_{3,j} \\ \sqrt{\tilde{k}_j^2} \end{pmatrix} .$$

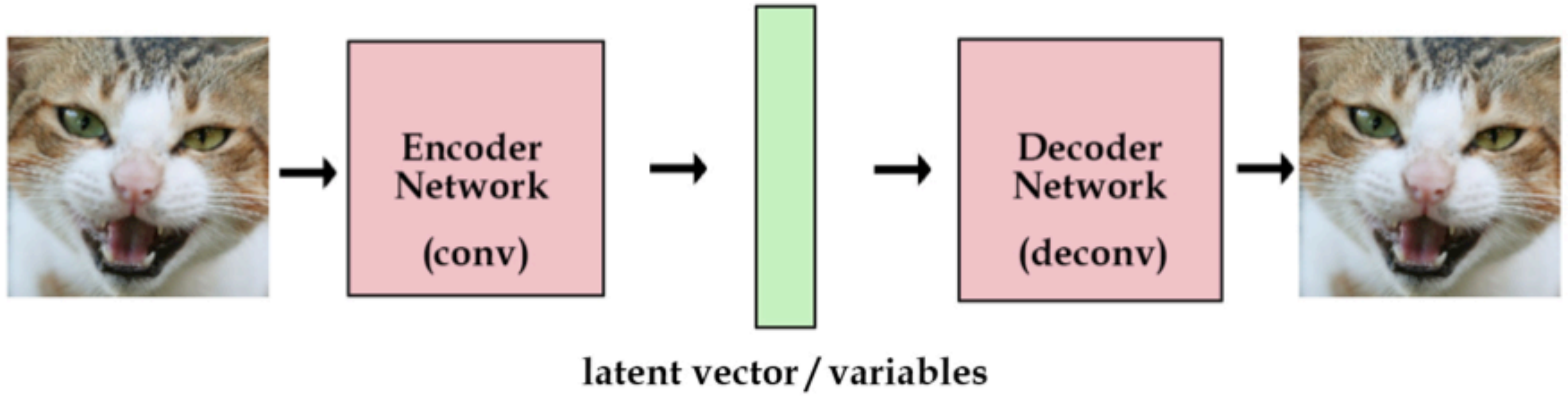
[Farina et al., arXiv:1808.08992]

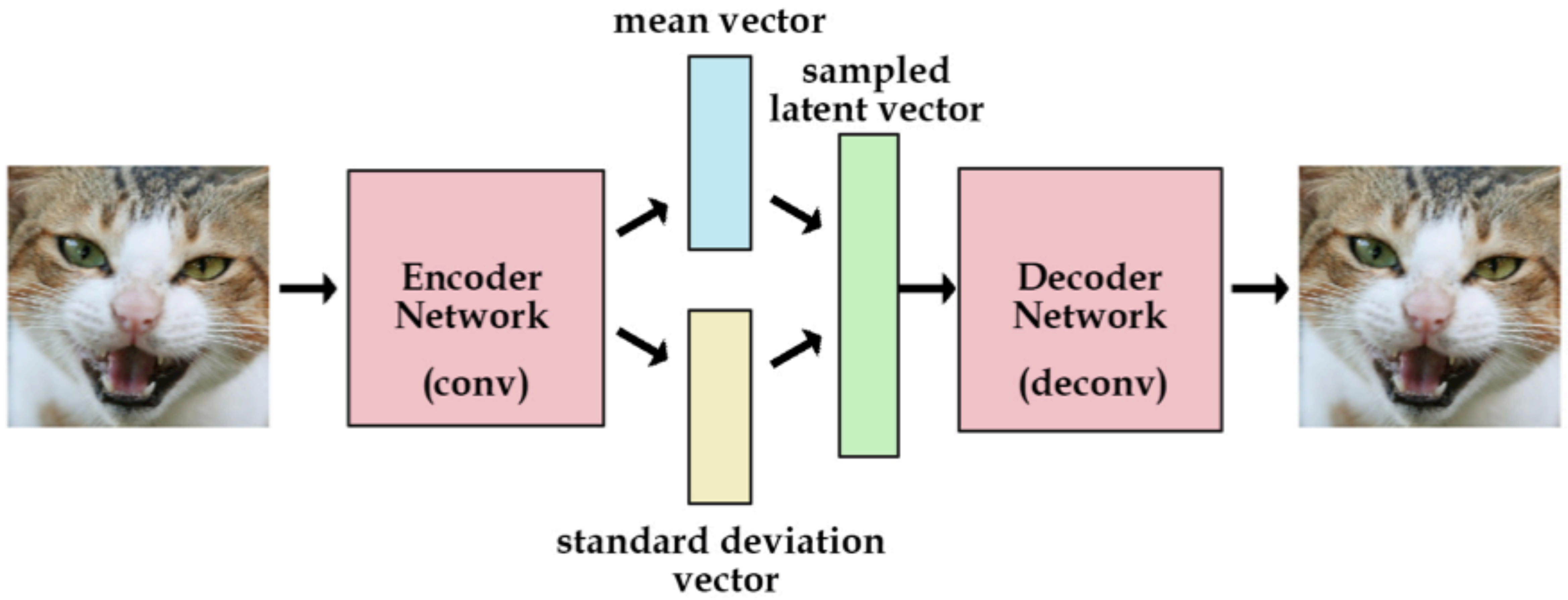
[Heimel et al., arXiv:1808.08979]

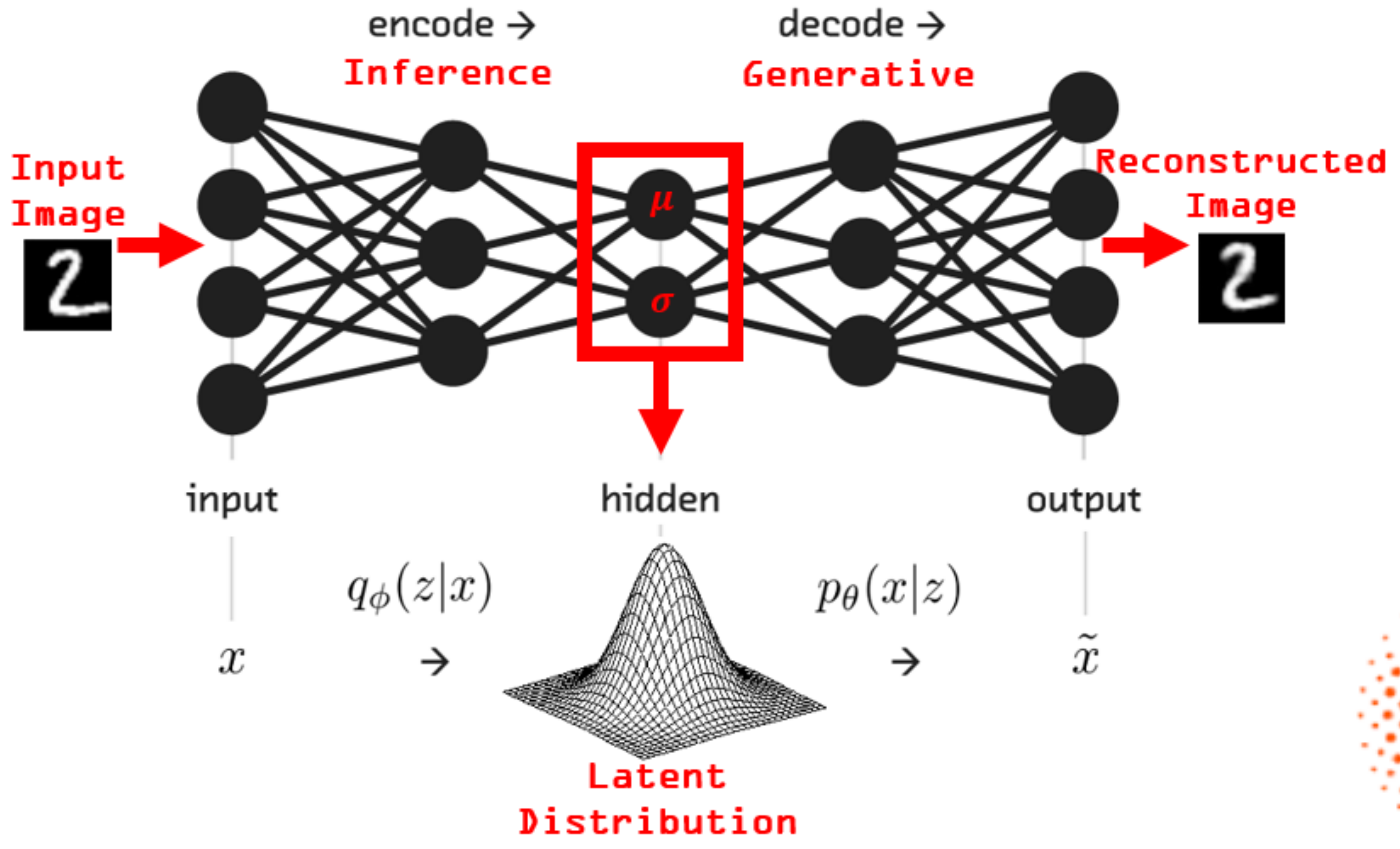
European Research Council

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







Our use case: 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 ($p_{T,\|}^{\text{miss}}$) 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:

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

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

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

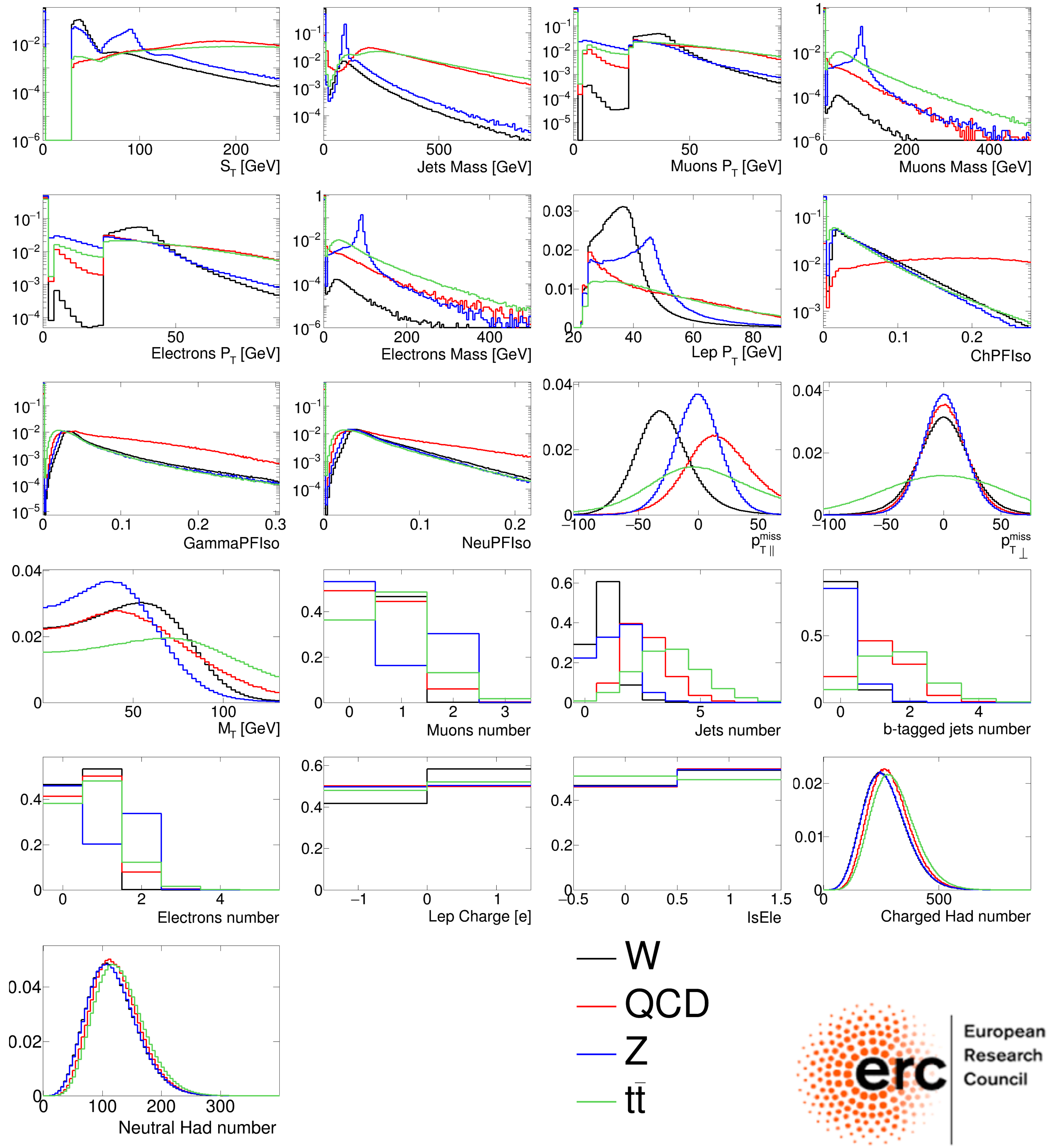
with $\Delta\phi$ the azimuth separation between the lepton and $\vec{p}_T^{\text{miss}}$ vector, and E_T^{miss} the absolute value of $\vec{p}_T^{\text{miss}}$.
- The number of selected muons (N_μ).
- The invariant mass of this set of muons (M_μ).
- The total transverse momentum of these muons ($p_{T,TOT}^\mu$).
- 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.

Our use case: 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)*

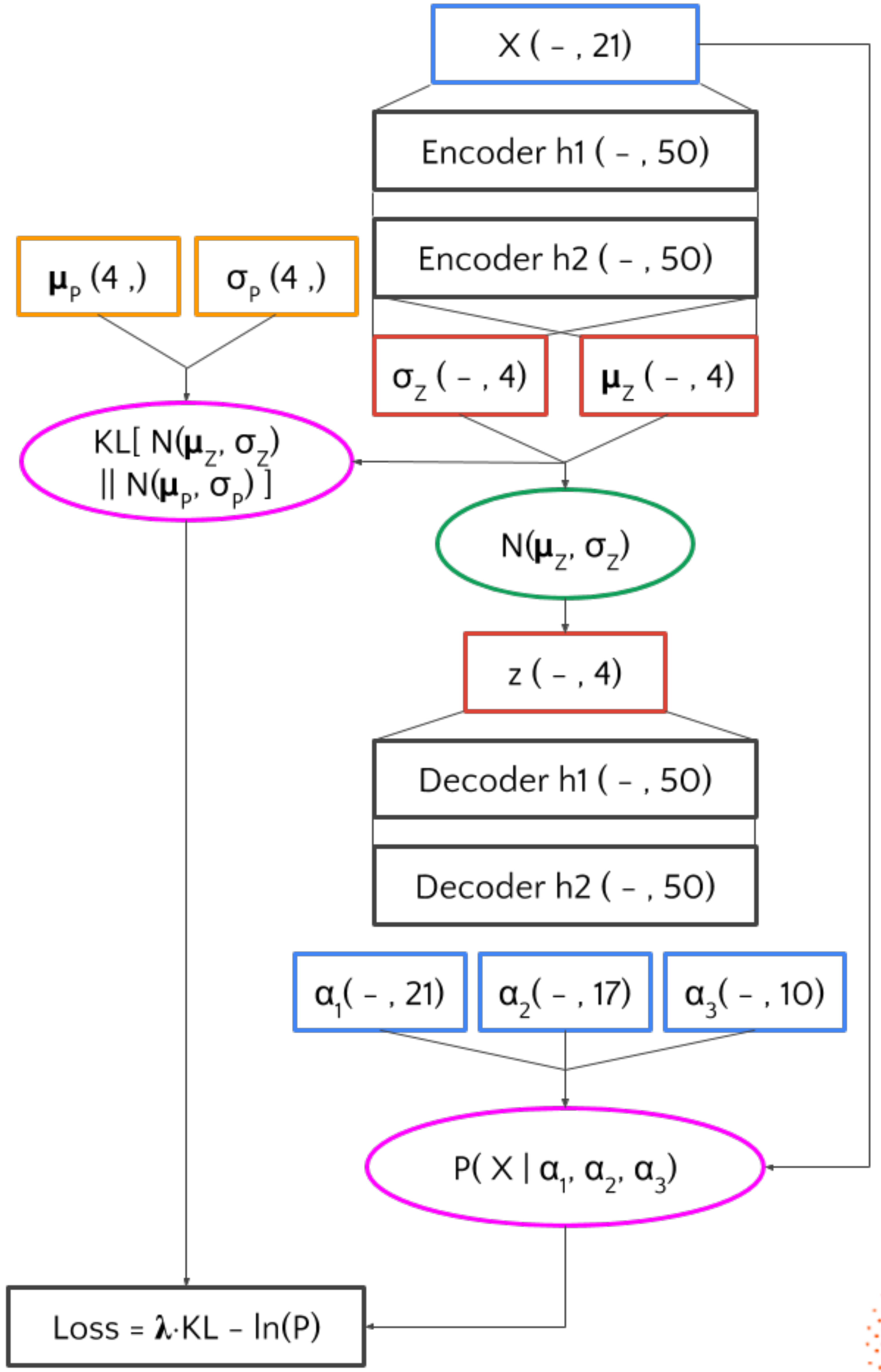


Standard Model VAE

- *We train a VAE on a cocktail of SM events (weighted by xsec)*

- ***ENCODER:*** *21 inputs, 2 hidden layers → 4Dim latent space*
  - *hidden nodes = μ and σ of the Gaussian pdfs describing the hidden variables*

- ***DECODER:*** *from a random sample in the 4D space → 2 hidden layers → parameters describing the shape of the 21Dim input space*



The Loss Function

$$\text{Loss}_{\text{Tot}} = \text{Loss}_{\text{reco}} + \lambda D_{\text{KL}}$$

- *Loss function described as the sum of two terms (scaled by a tuned λ parameter that makes the two contribution numerically similar)*
  - *Reconstruction loss: likelihood of the input 21Dim point, given the shape parameters reconstructed from it*
  - *KL loss:*

$$\text{Loss}_{\text{reco}} = -\frac{1}{k}\sum_i \ln\left(P(x \mid \alpha_1, \alpha_2, \alpha_3)\right)$$

$$= -\frac{1}{k}\sum_{i,j} \ln\left(f_j(x_{i,j} \mid \alpha_1^{i,j}, \alpha_2^{i,j}, \alpha_3^{i,j})\right)$$

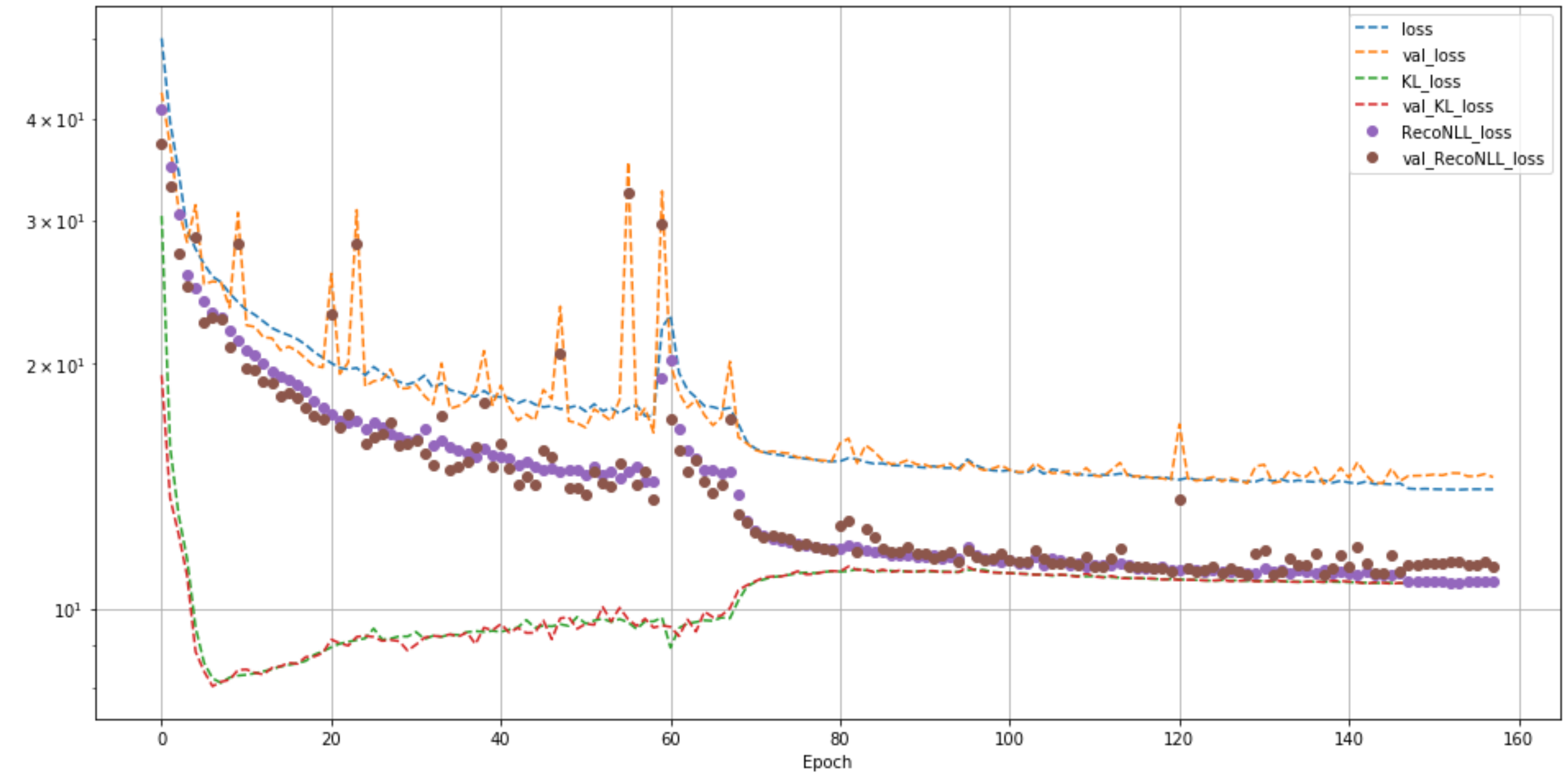
$$D_{\text{KL}} = \frac{1}{k}\sum_i D_{\text{KL}}\left(N(\mu_z^i, \sigma_z^i) \;||\; N(\mu_P, \sigma_P)\right)$$

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

European Research Council

Training

- *Thanks to choice of l, two terms simultaneously minimized by minimizing the sum*
- *Training converges after ~100 epochs (i.e., looping 100 times on the input dataset)*
- *Model implemented in Keras+TensorFlow. Trained on iBanks GPU cluster @Caltech*



Keras

TensorFlow

erc European Research Council

CERN

Standard Model encoding

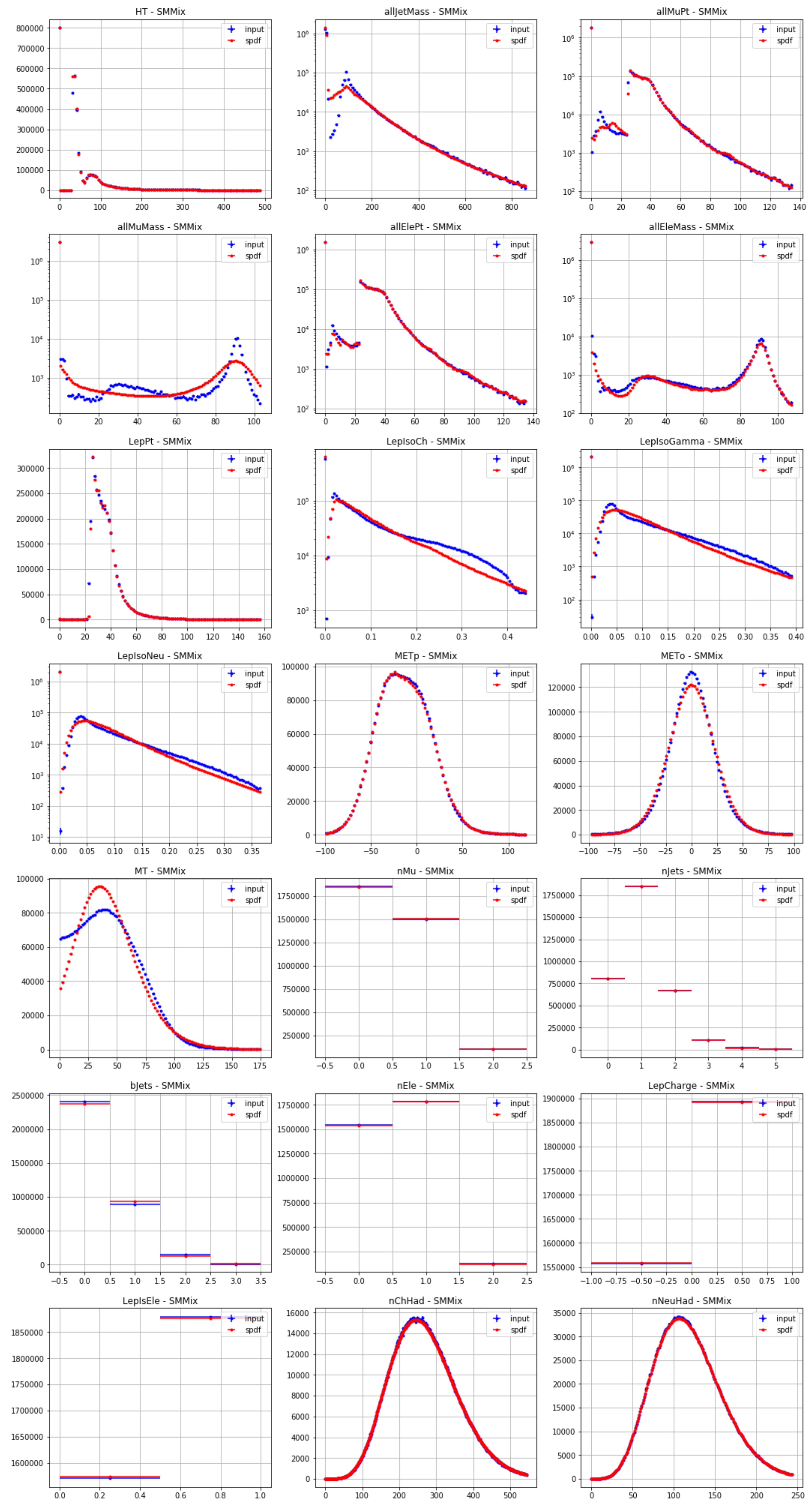
- First post-training check consists in verifying encoding-decoding capability, comparing input data to those generated sampling from decoder

- Reasonable agreement observed, with small discrepancy here and there

- NOTICE THAT: this would be a suboptimal event generator, but we want to use it for anomaly detection

    - no guarantee that the best autoencoder is the best anomaly detector (no anomaly detection rate in the loss function)

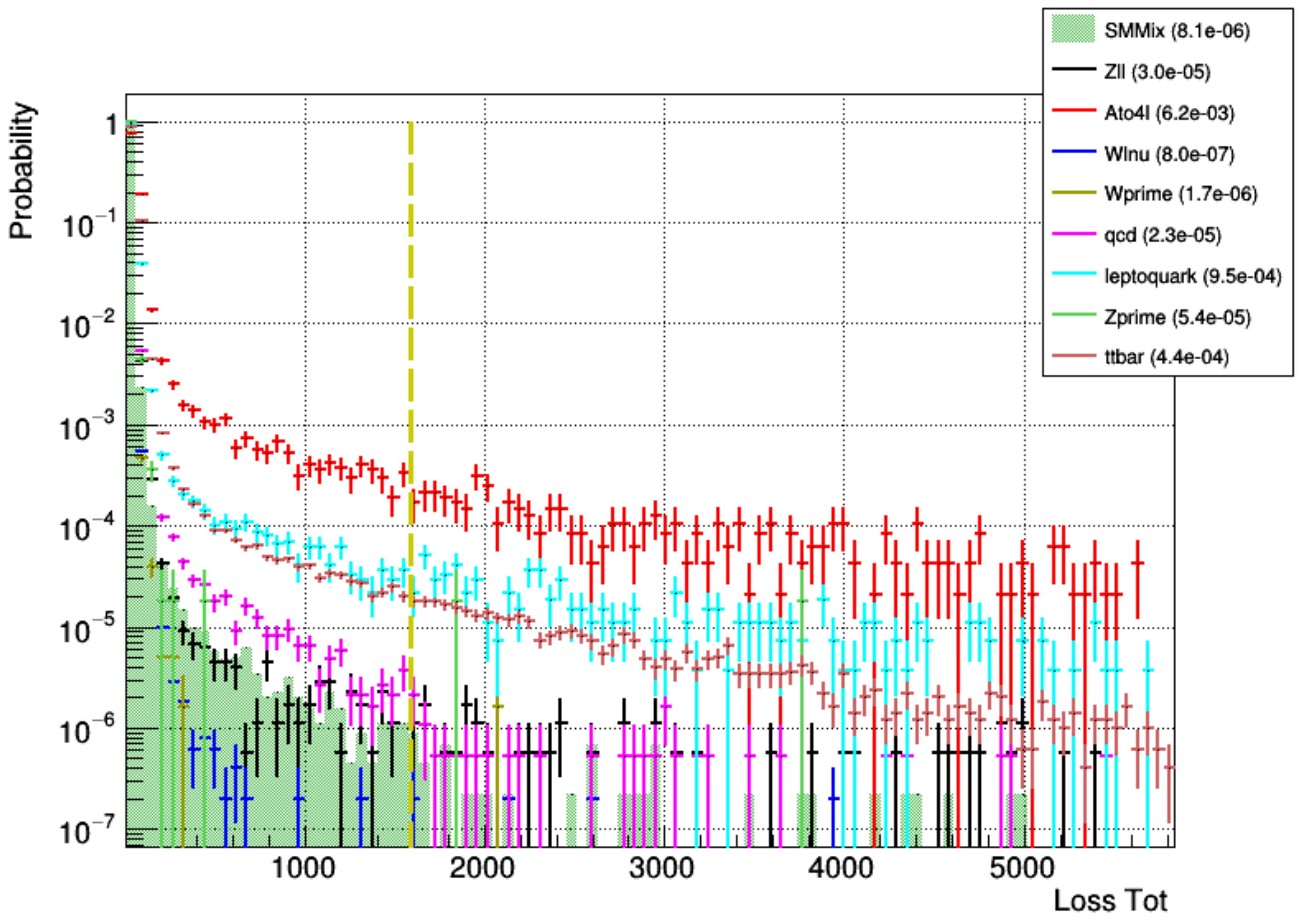
    - pros & cons of an unsupervised/ semisupervised approach

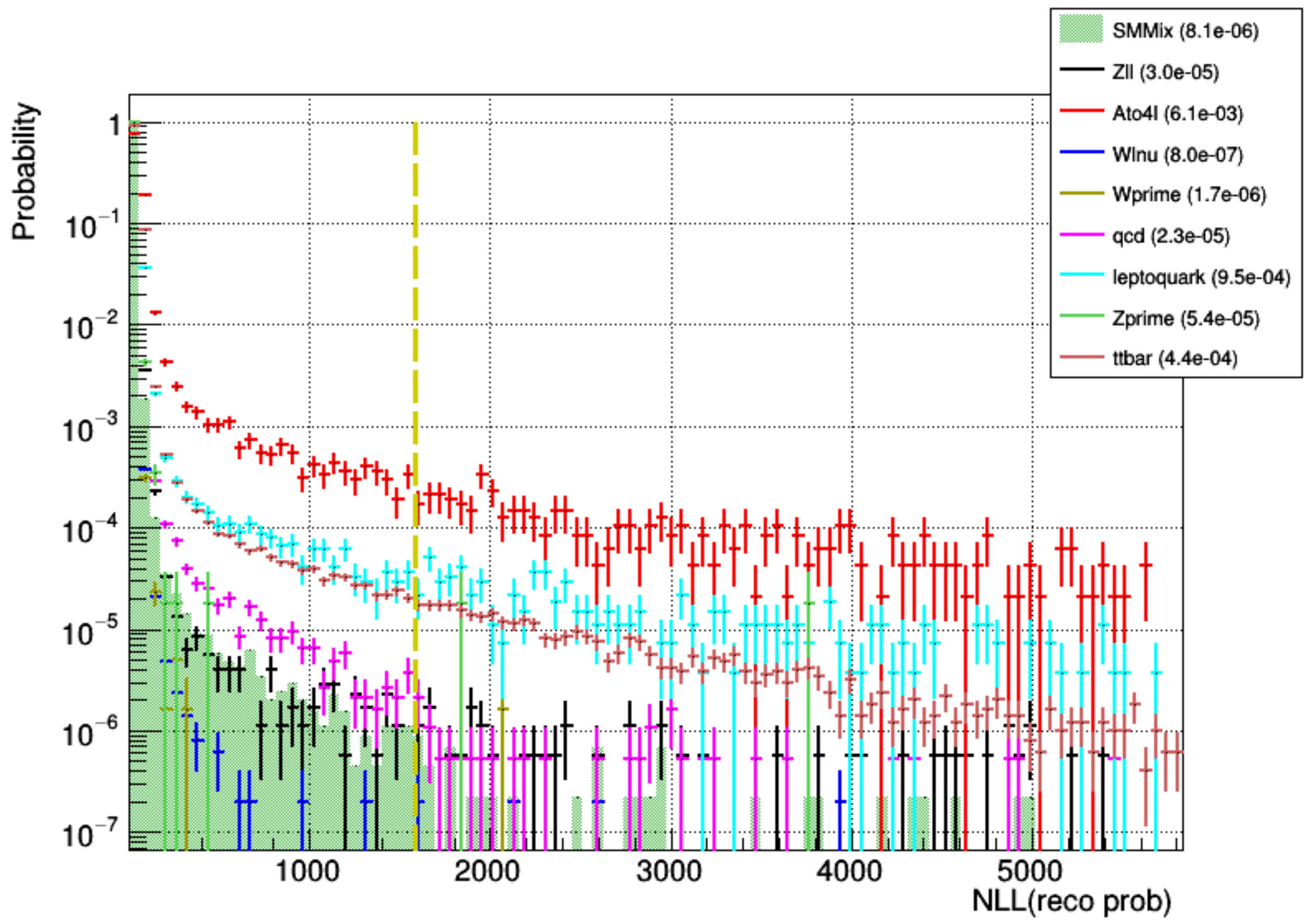


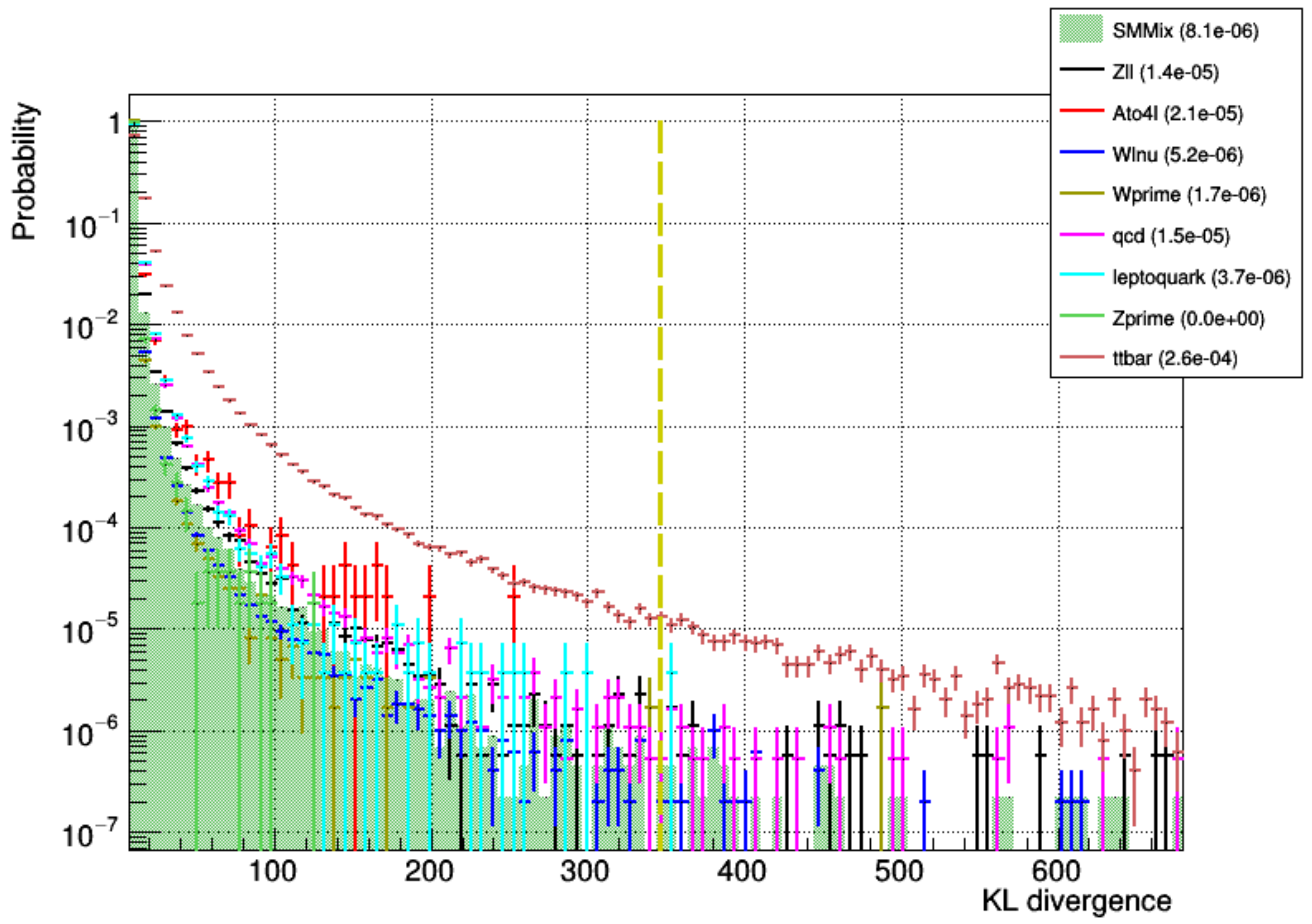
European Research Council

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





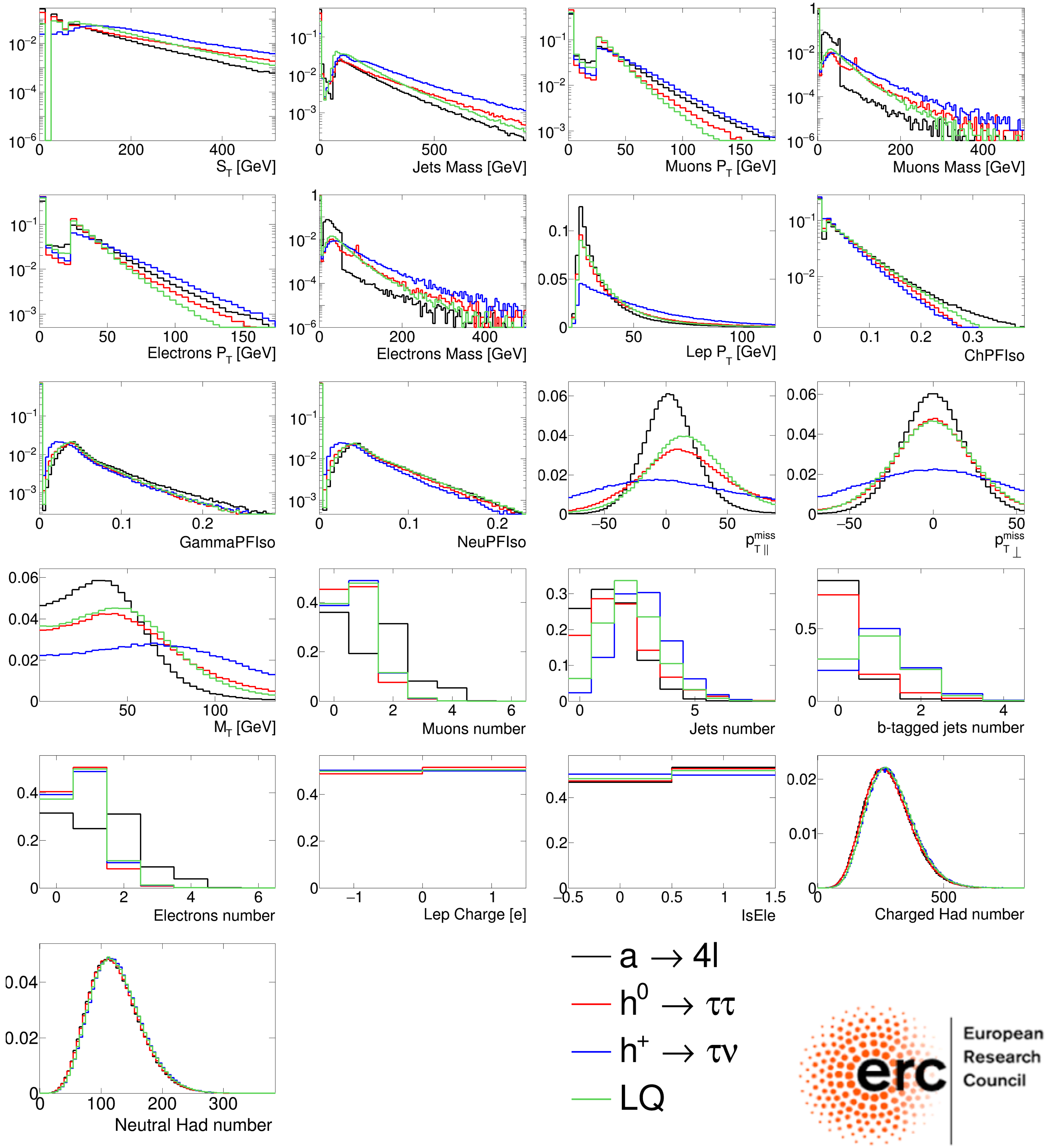


Some BSM benchmark

- *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→Z*Z*→4ℓ*
  - *A scalar scalar boson with mass 60 GeV, h→ττ*
  - *A charged scalar boson with mass 60 GeV, h±→τν*

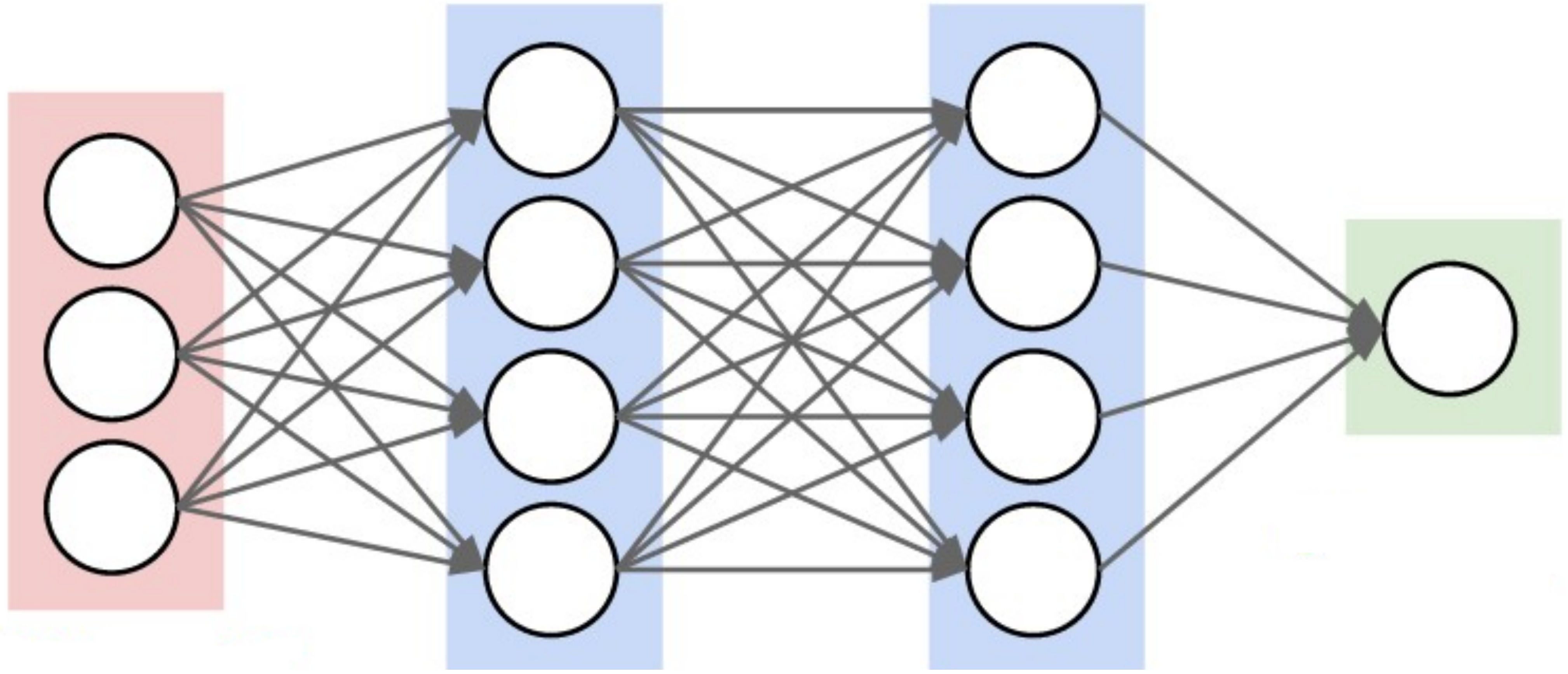
BSM benchmark processes

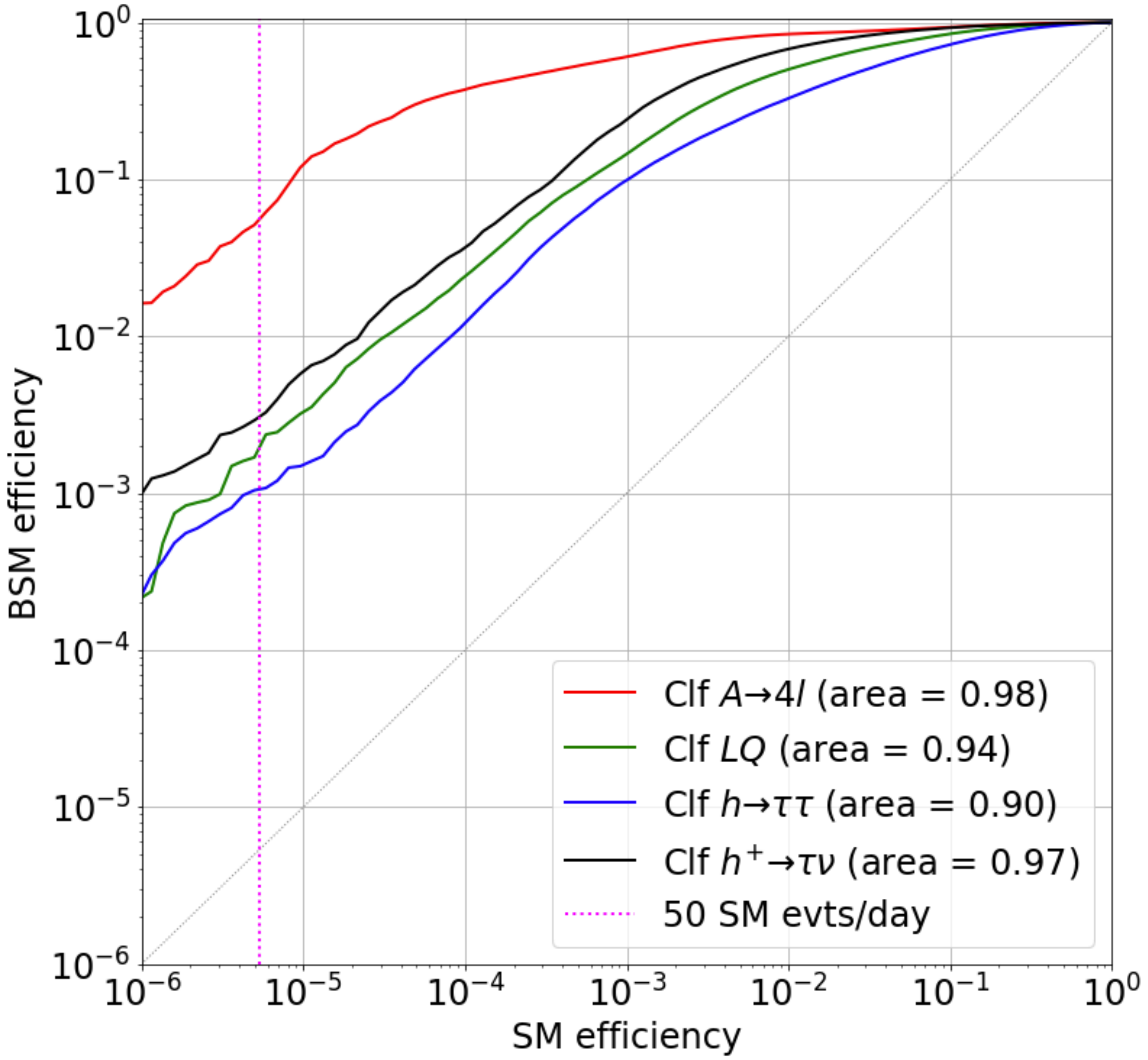
Process	Acceptance	Trigger efficiency	Total efficiency	Cross-section 100 events/month
$h^0 \to \tau\tau$	9%	70%	6%	335 fb
$h^0 \to \tau\nu$	18%	69%	12%	163 fb
$LQ \to b\tau$	19%	62%	12%	166 fb
$a \to 4\ell$	5%	98%	5%	436 fb



Benchmark comparison

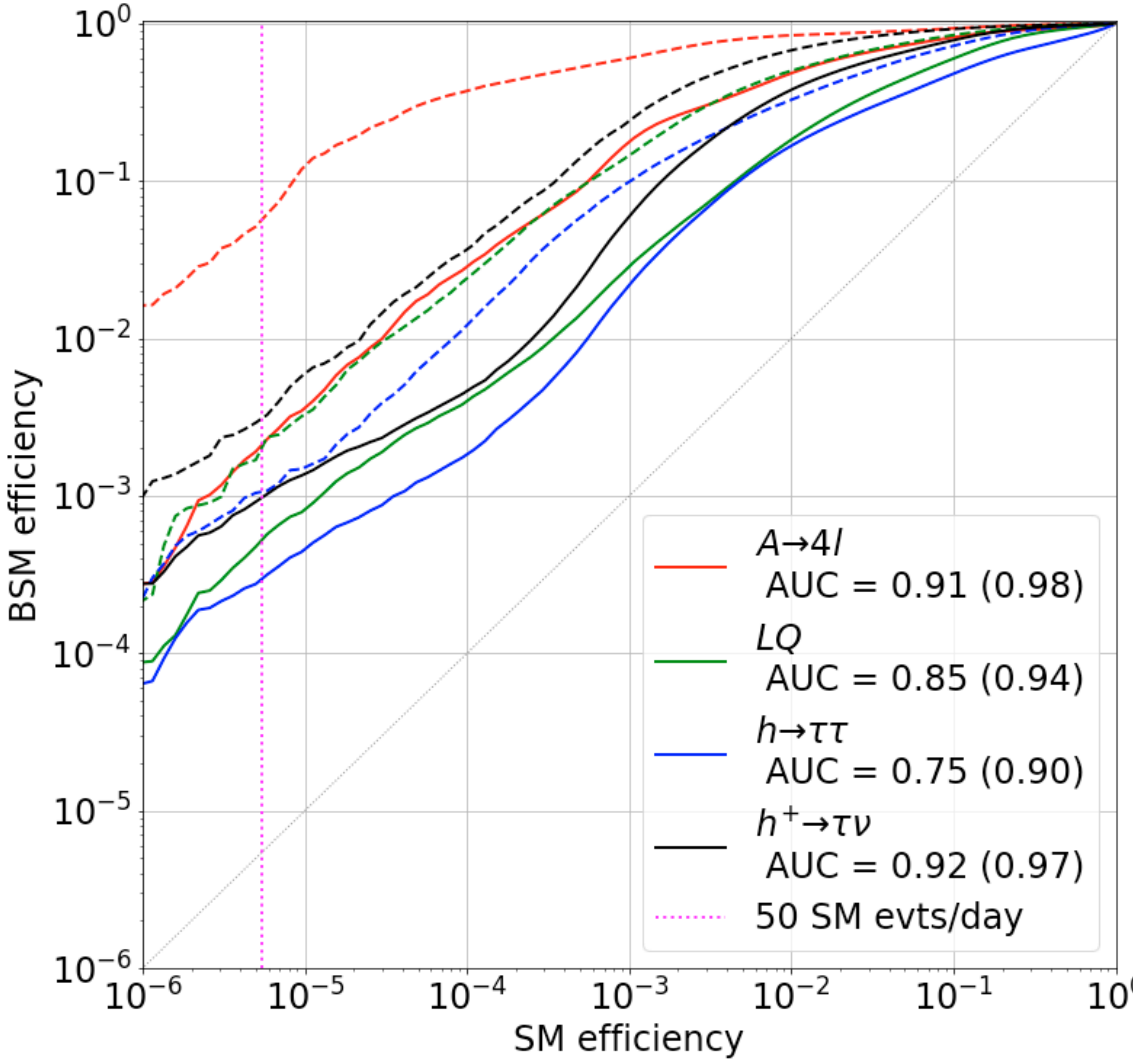
- *VAE's performances benchmarked against supervised classifiers*
- *For each BSM model*
  - *take same inputs as VAE*
  - *train a fully-supervised classifier to separate signal from background*
  - *use supervised performances as a reference to aim to with the unsupervised approach*
  - *Done for our 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 (assuming L=XXX*
  - *competitive results considering unsupervised nature of the algorithm*



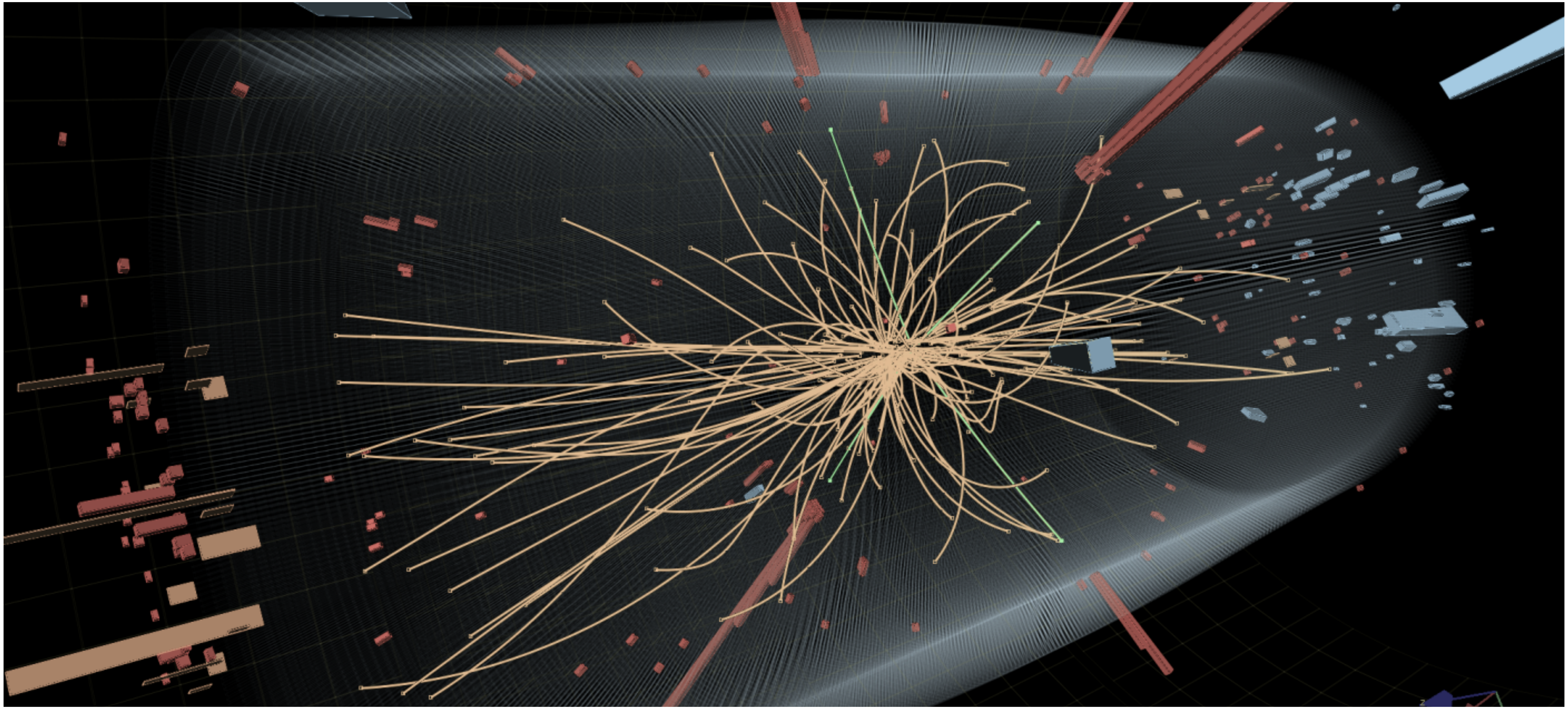
Performances

- *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

1/2 way to model independence

- *Procedure designed to be model independent*
  - *Training done only on SM*
  - *Algorithm that defines anomaly tuned only on number of selected SM events (false positive rate)*
- *Still, residual model dependence present*
  - *Based on physics-motivated observables*
  - *List not tailored on specific models and general enough to offer good performances in principle*
  - *But one cannot prove that performances on specific BSM models will generalise*
- *Can we go beyond this limitation and define something really BSM agnostic?*

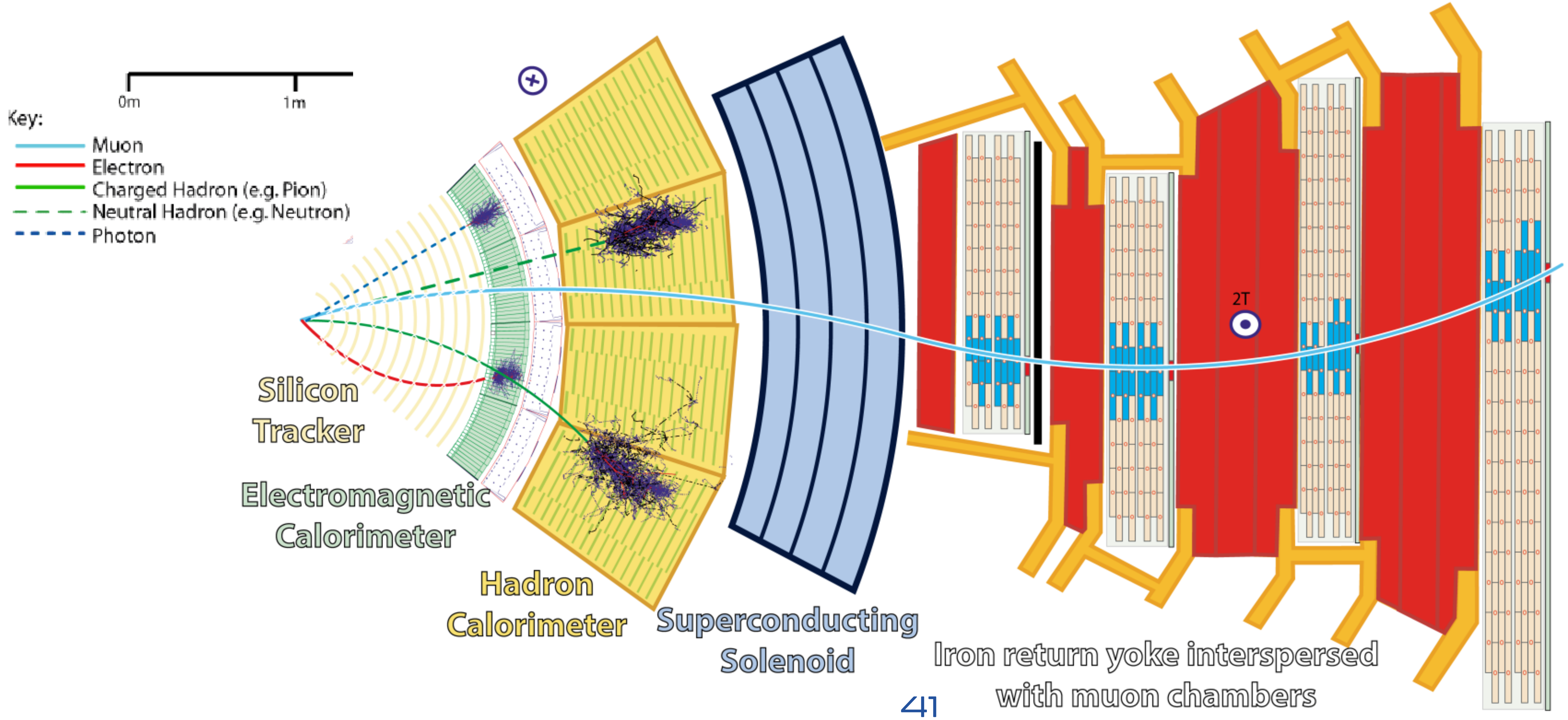


Particle Flow, Recurrent Networks & Model Independence

European Research Council

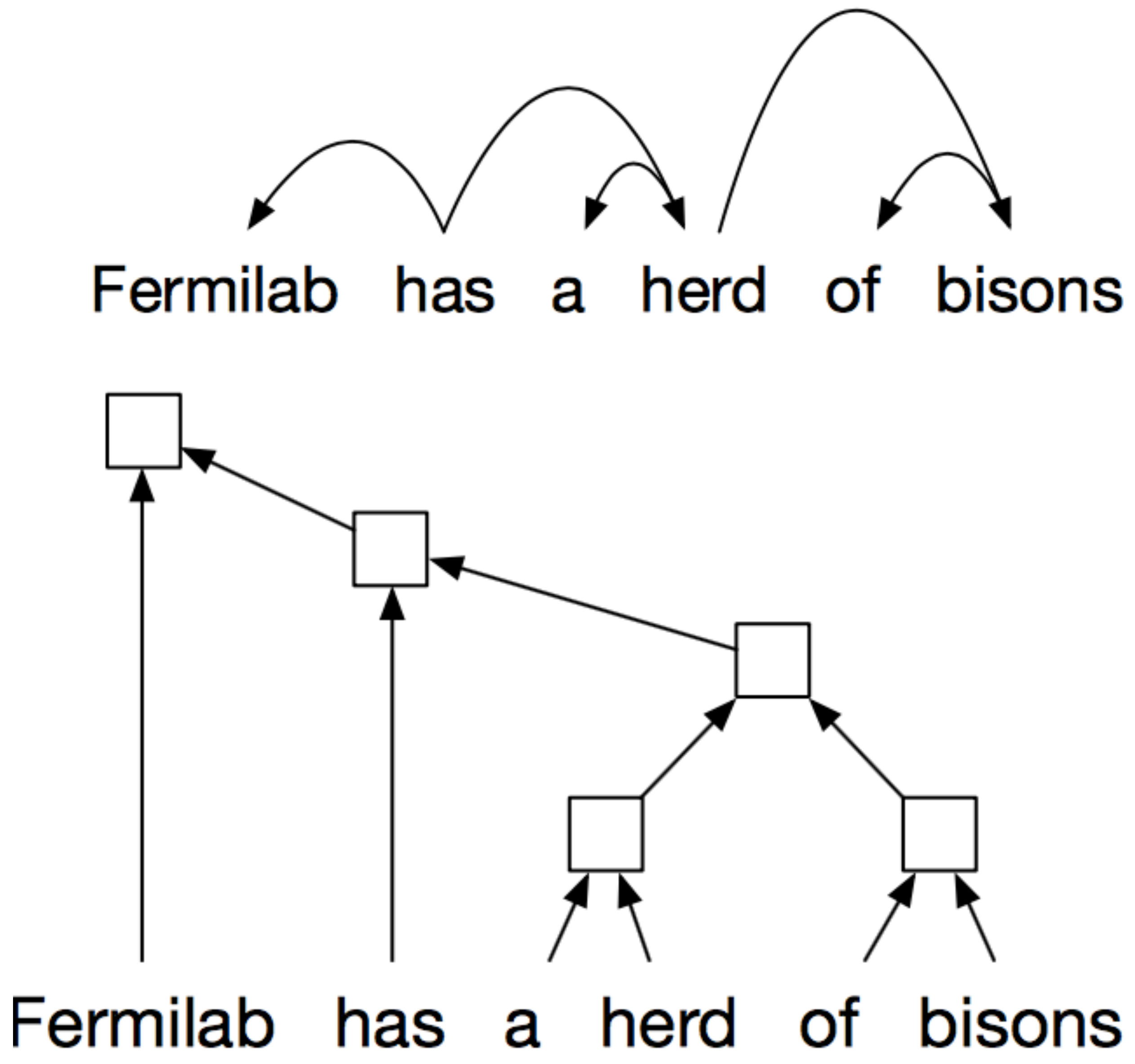
Particle Flow

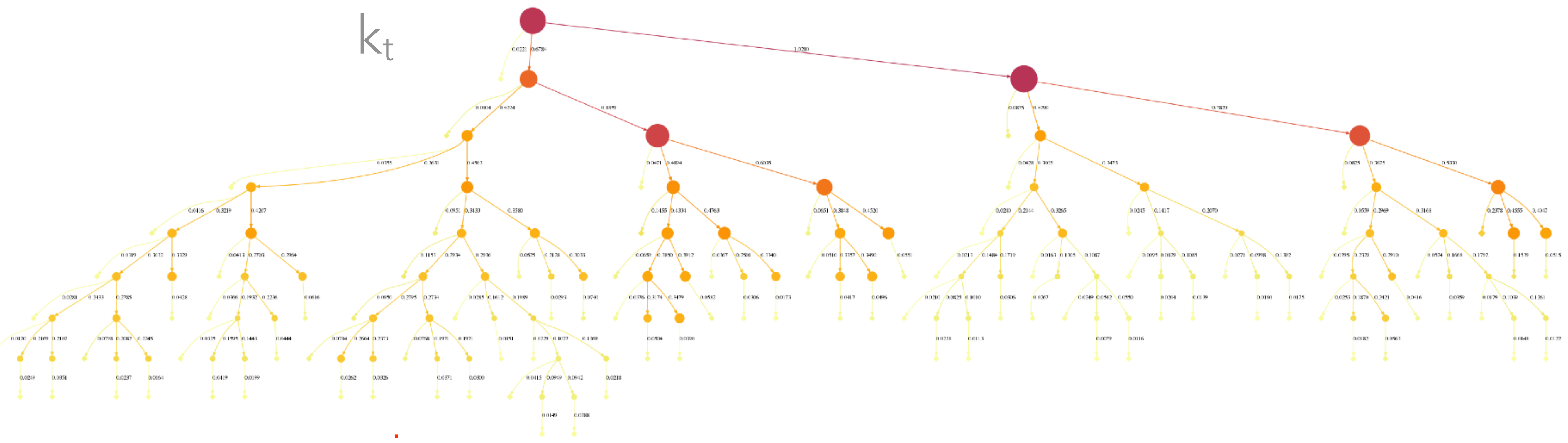
- CMS uses PF to combine sub-detector information and produce a list of reconstructed particles
- Anything (jets, MET, resonances, etc) is reconstructed from these particles
- One could generalise the VAE new-physics-detection algorithm and make it PF compliant
    - integrated in the reconstruction flow @HLT
    - can abstract from model dependence inherited by any physics-motivated HLF choice



LHC events & language processing

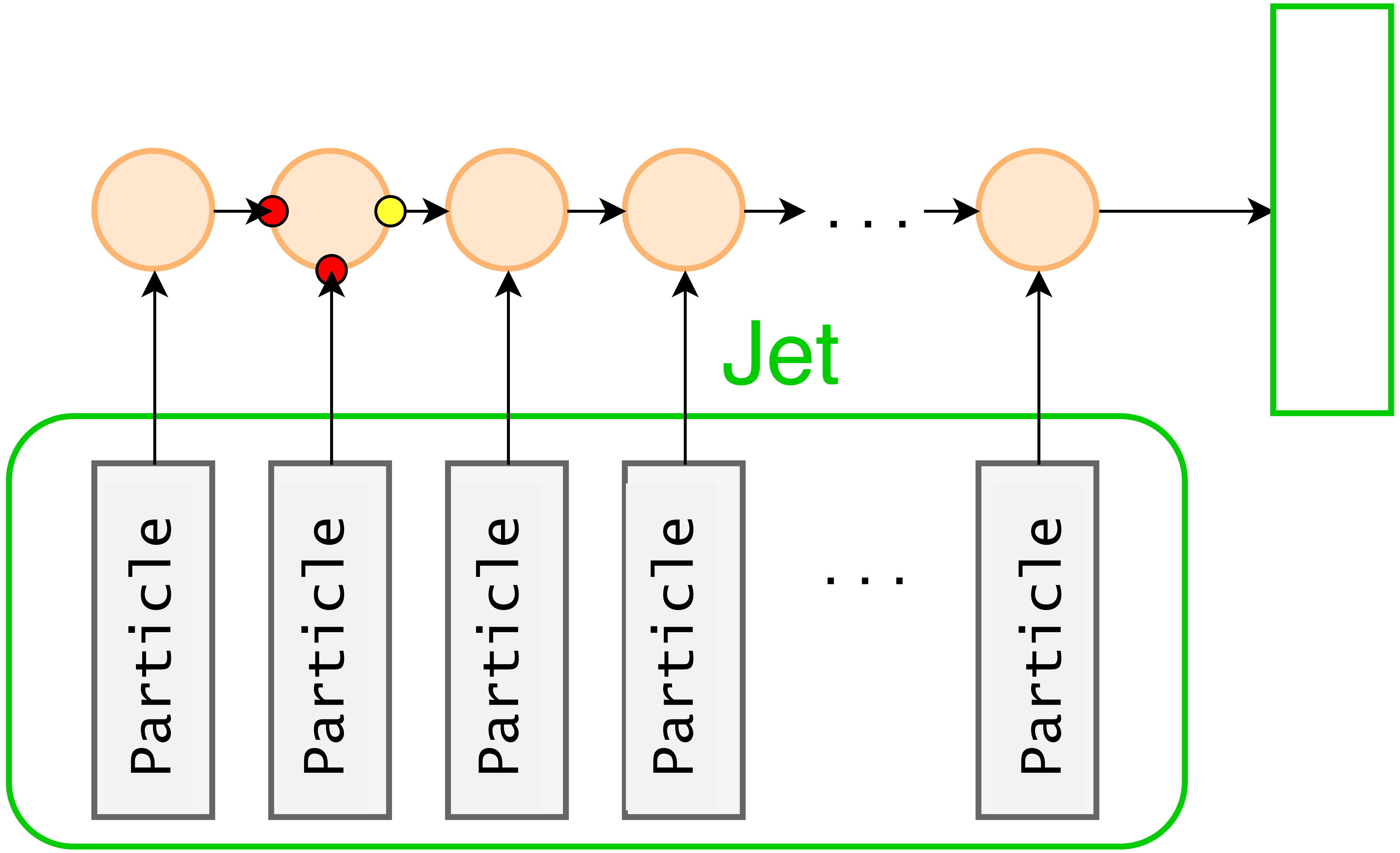
- *PF reco is not the best match for computing vision techniques (e.g., convolutional neural networks) don't work*
  - *one would have to convert the particles to a pixelated images, loosing resolution*
- *Instead, list of particles can be processed by Deep Learning architectures designed for natural language processing (RNN, LSTMs, GRUs, …)*
  - *particles as words in a sentence*
  - *QCD is the grammar*





Recurrent Neural Networks

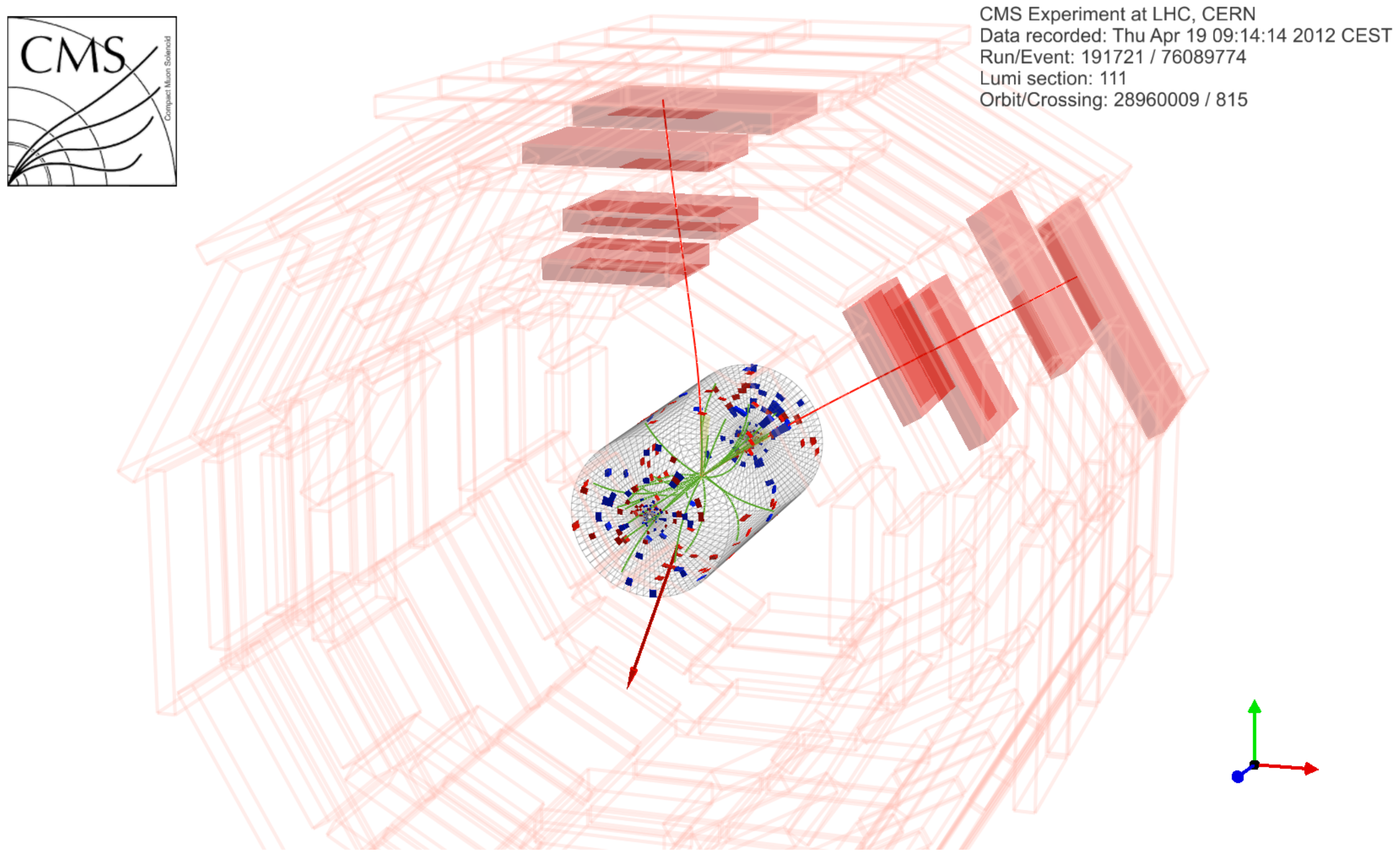
- *A network architecture suitable to process an ordered sequence of inputs*
  - *words in text processing*
  - *a time series*
  - *particles in a list*
- *Could be used for a single jet or the full event*
- *Next step: graph networks (active research direction)*

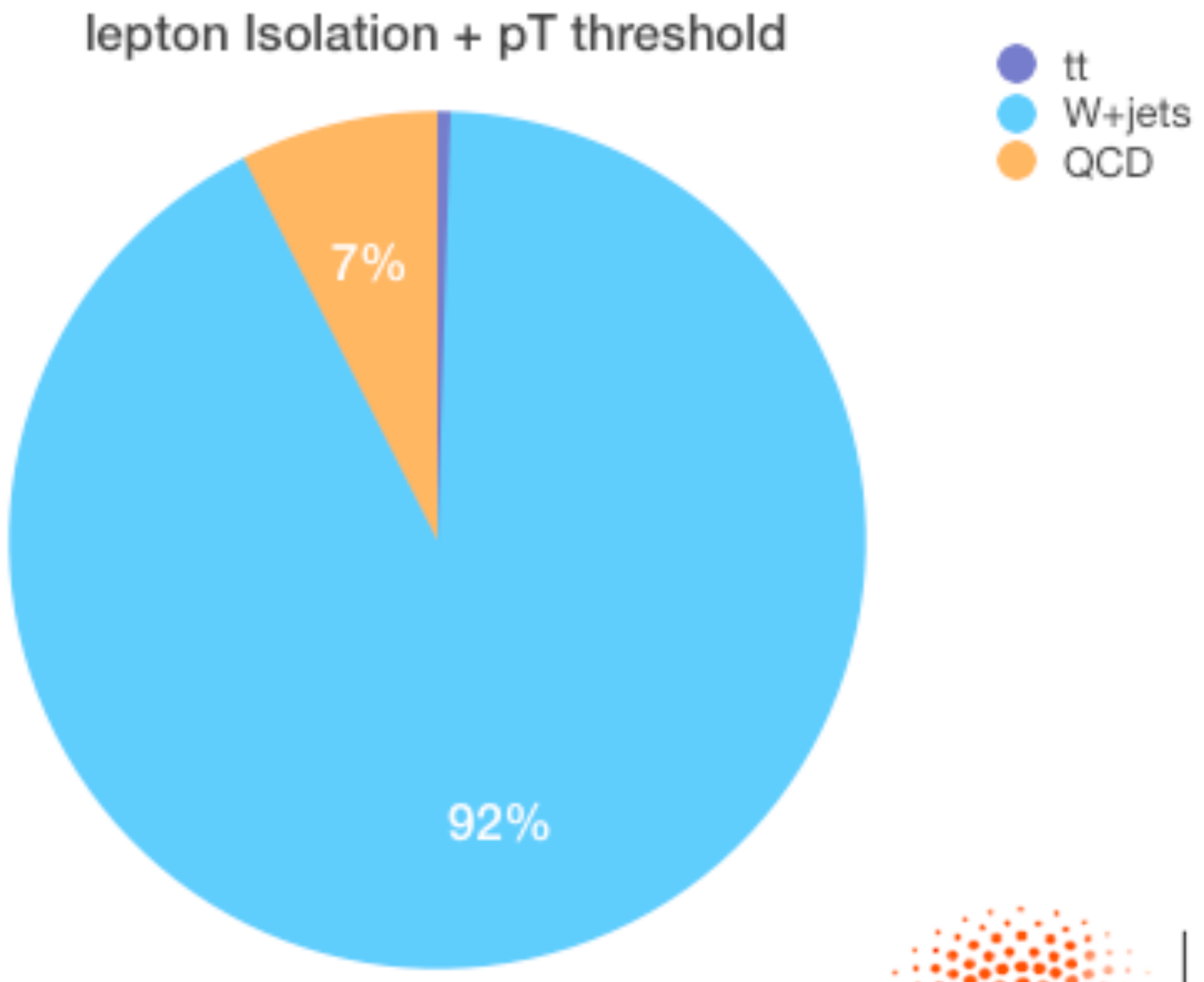


Example: A Topology Classifier

A typical example: leptonic triggers

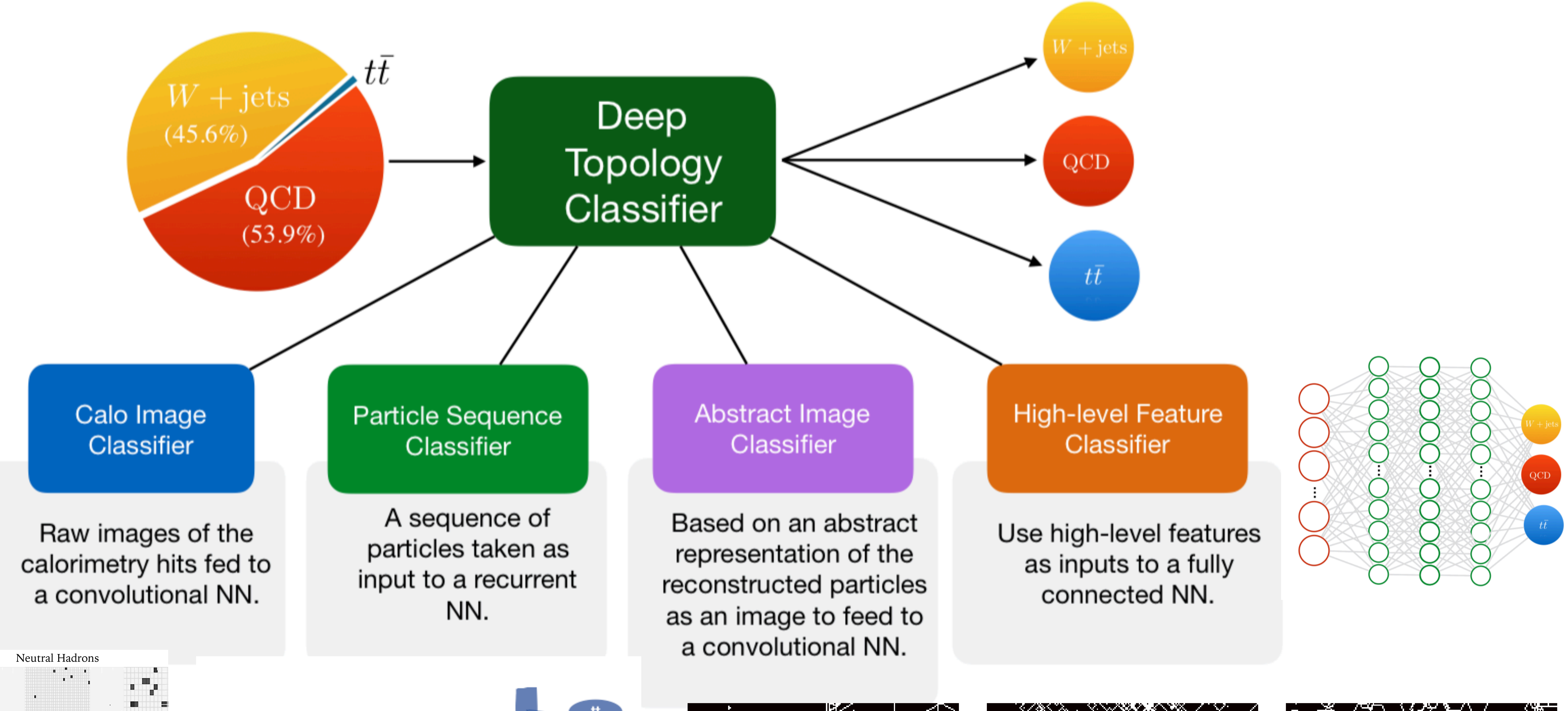
- at the LHC, producing an isolated electron or muon is very rare. Typical smoking gun that something interesting happened (Z,W,top,H production)-> **TAKE THEM!**
- Triggers like those are very central to ATLAS/CMS physics
- **The sample selected is enriched in interesting events, but still contaminated by non-interesting ones**
- Can we clean this up w/o biasing the physics? yes, with ML





See contribution to NIPS workshop

A Topology Classifier



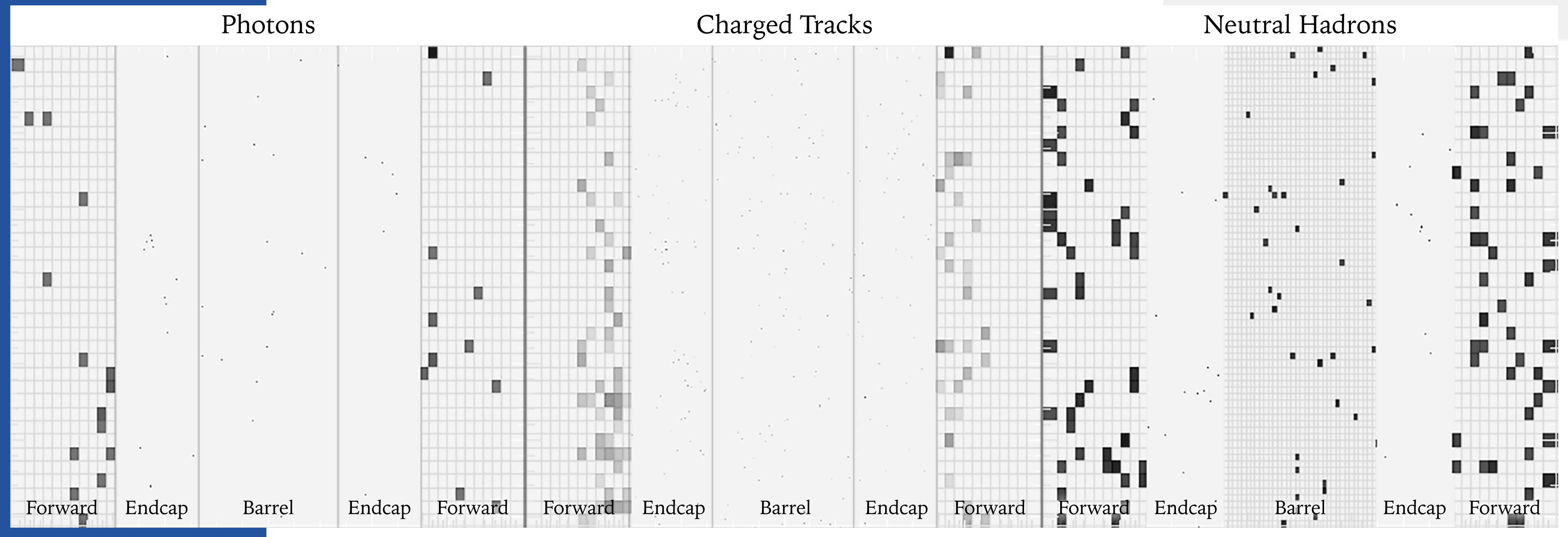
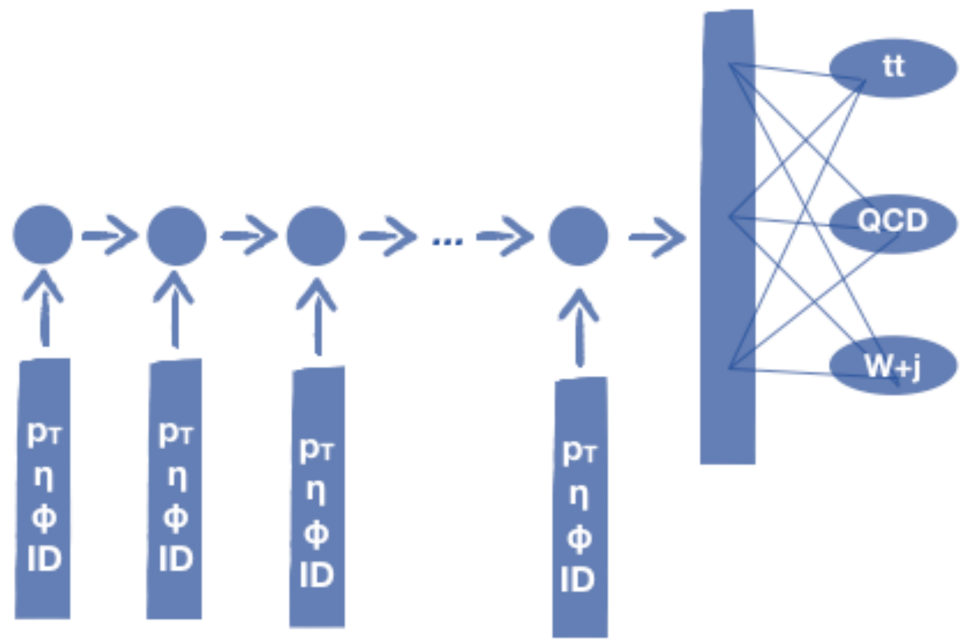
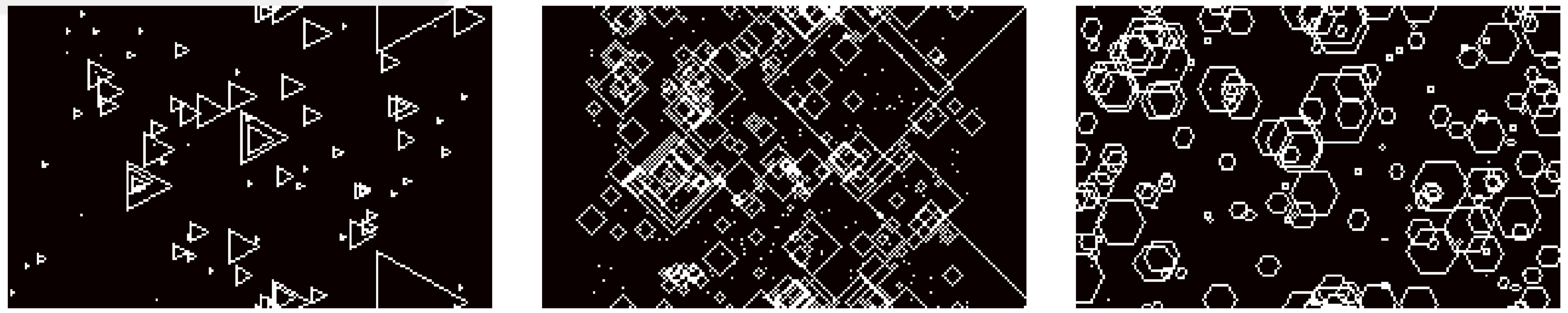


Figure 2: An example of a $t\bar{t}$ event as the input of the raw-image classifier.

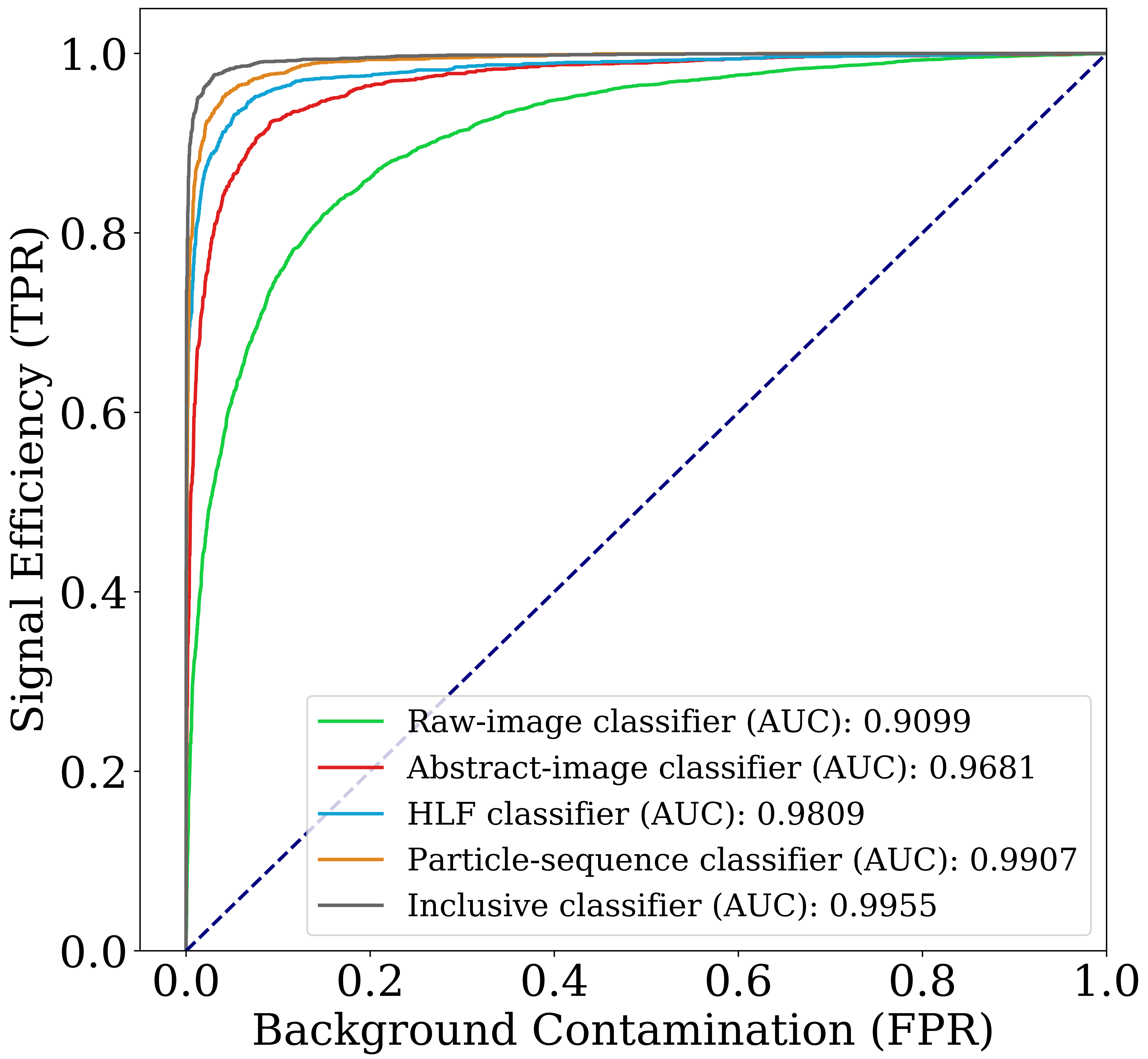




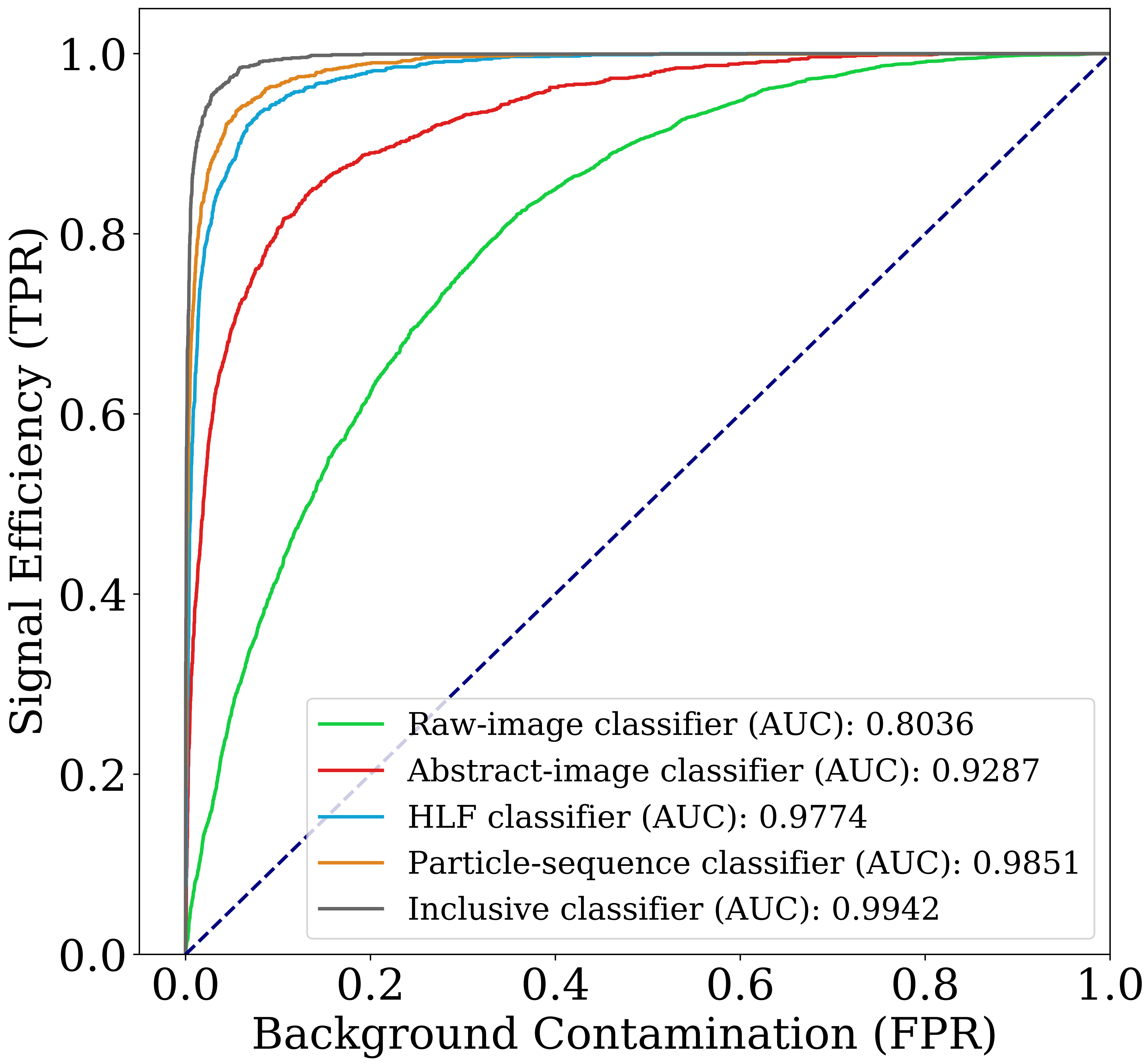
(a) Photons (b) Charged Particles (c) Neutral Hadrons

European Research Council

Selection performances



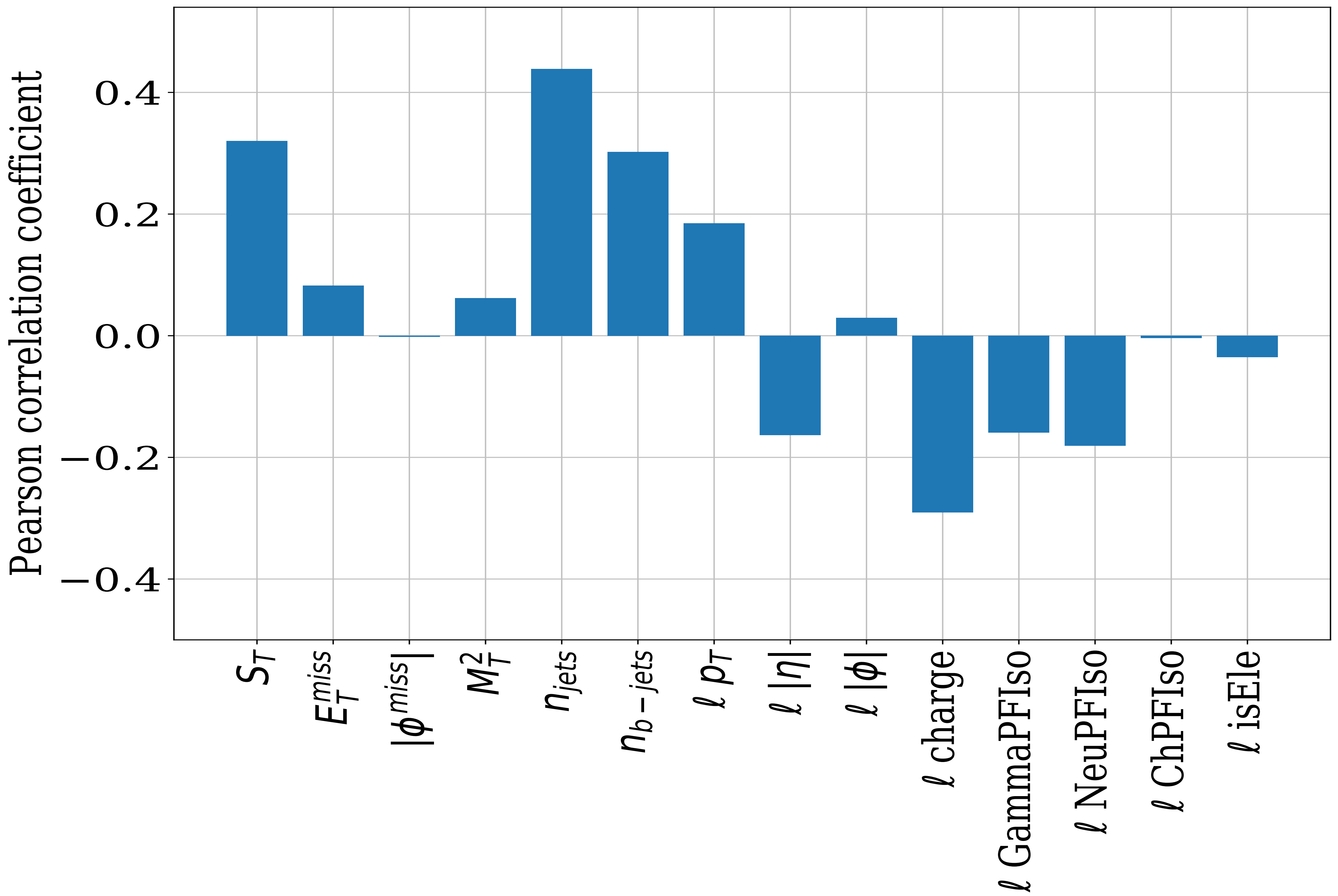
(a) $t\bar{t}$ selector

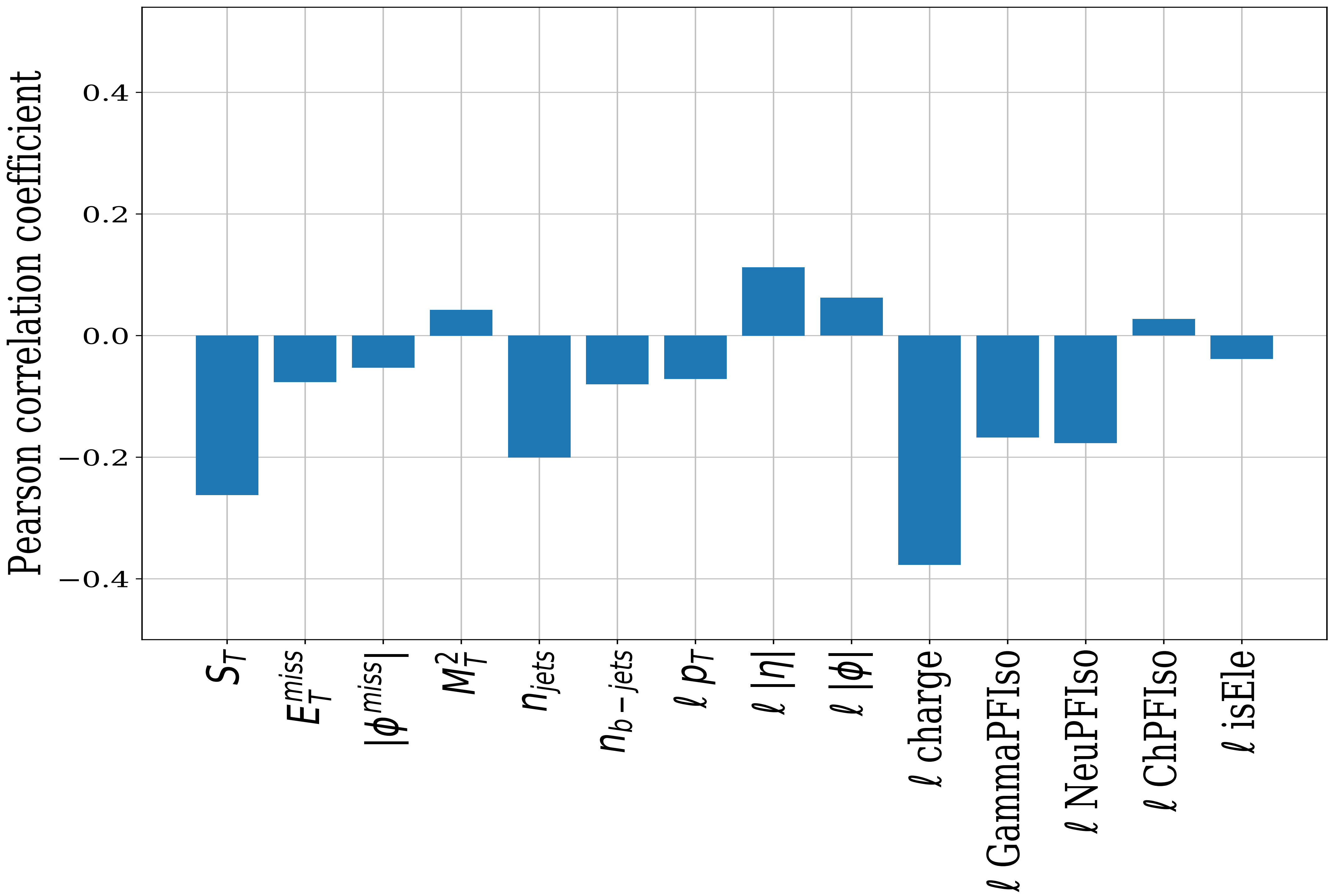


(b) W selector

Can select 99% of the top events and reduce the fraction of written events by a factor ~ 7

Selection performances



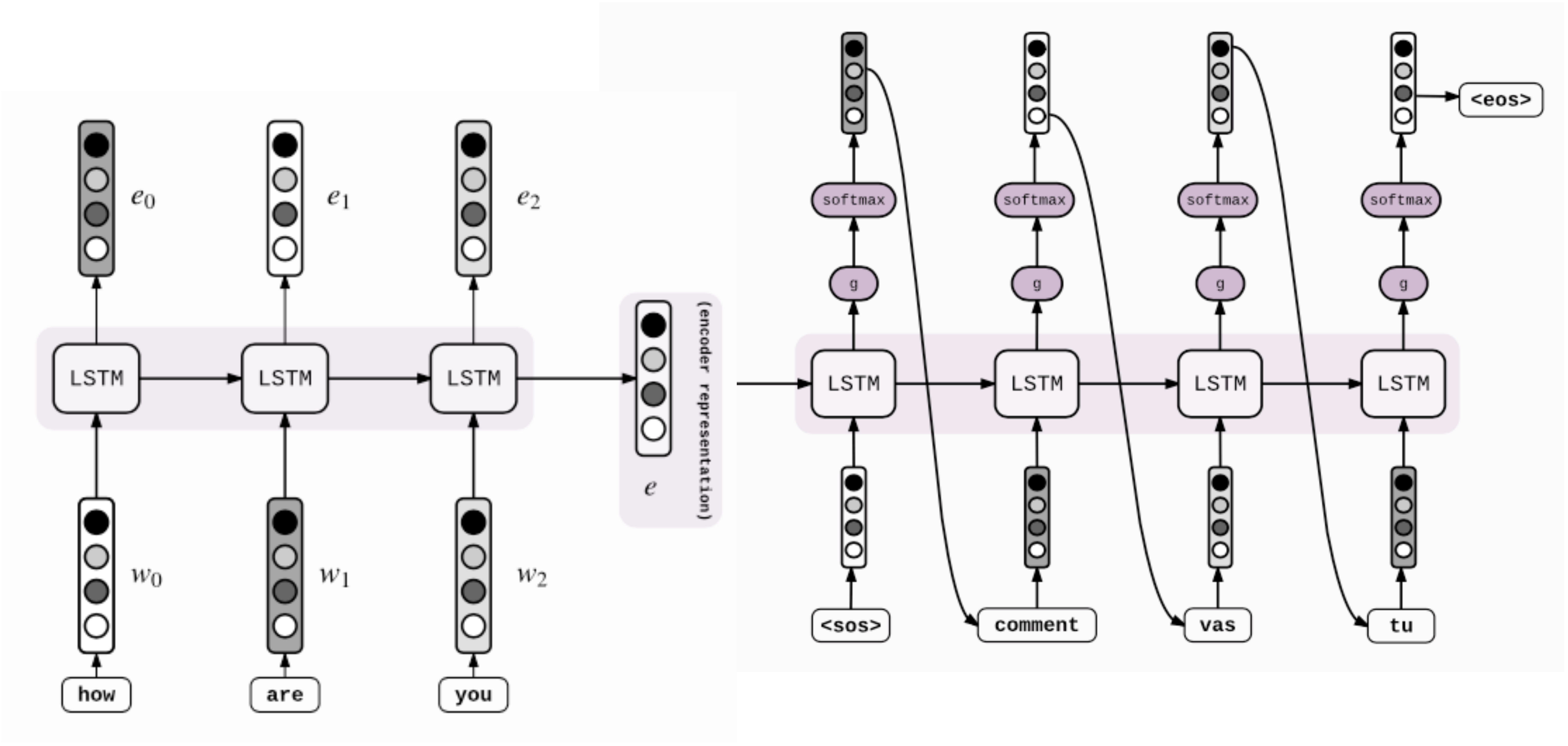


What is the network learning?

- tt events are more crowded that W events
- leptons in W and tt events are isolated from other particles

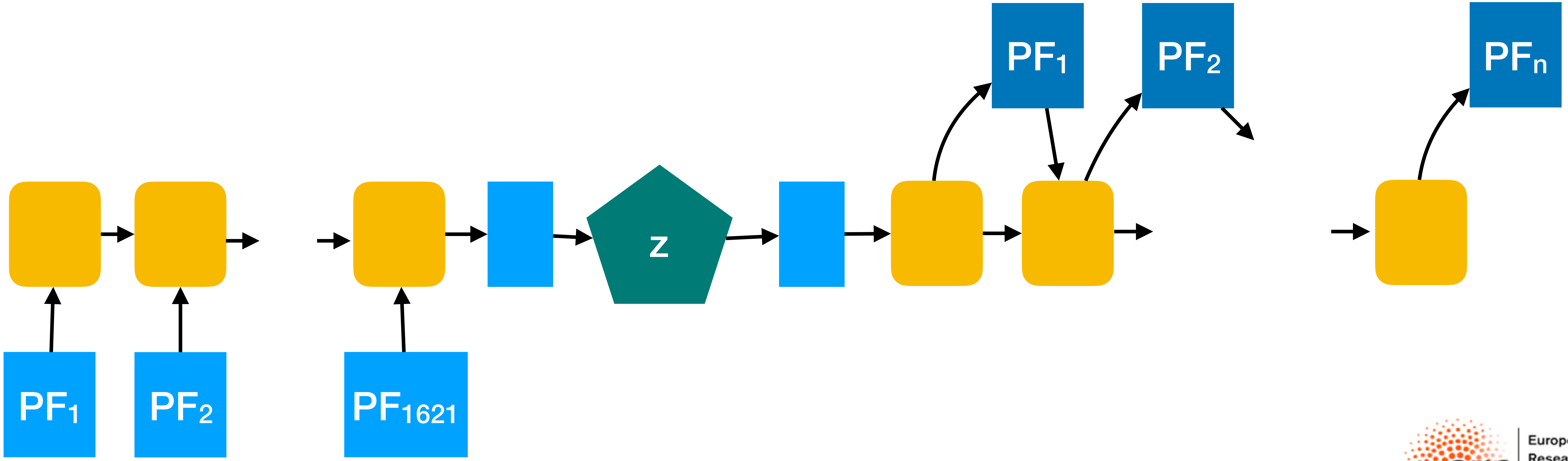
VAE with PF particles

- Issues:
  - variable number of particles/event as input
  - need to return particles as output
- Networks used for translation
  - start from a sentence in language
  - code its meaning in some latent space z
  - translate to some other language, generating words from z



VAE with PF particles

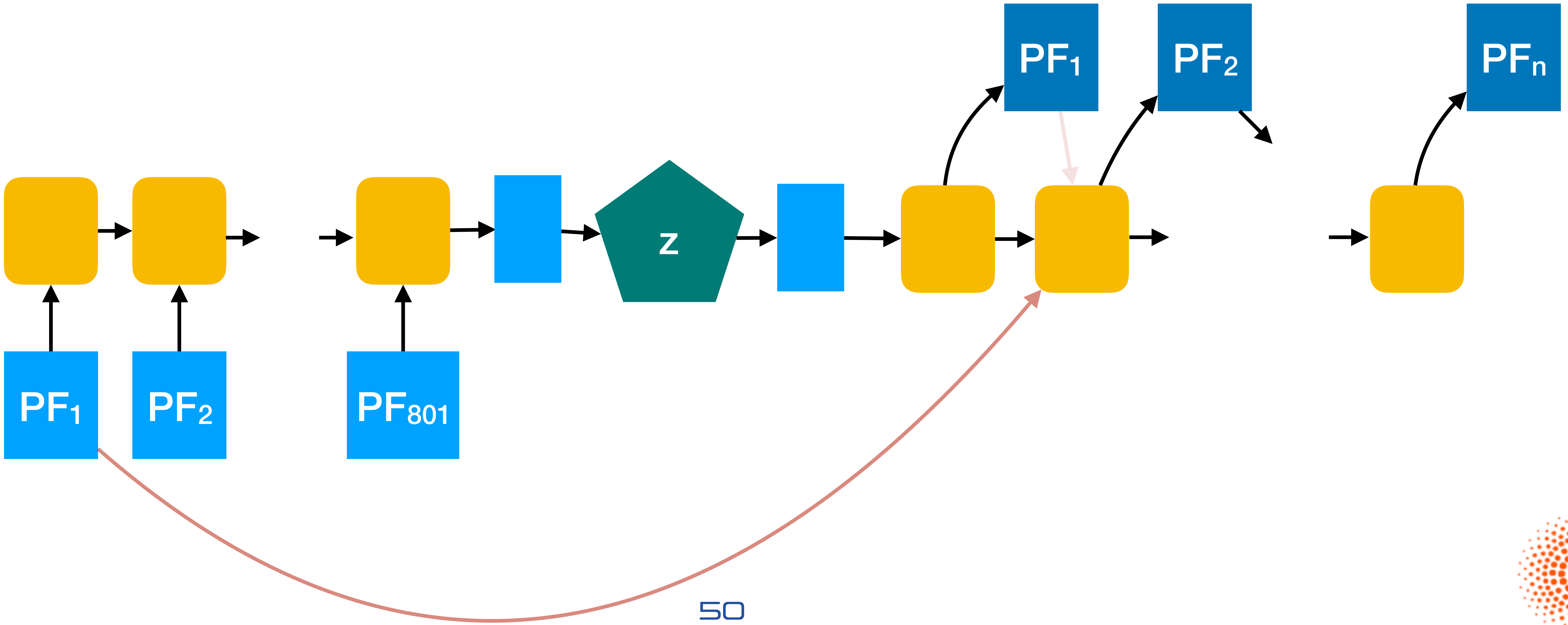
- Issues:
  - variable number of particles/event as input
  - need to return particles as output



Teacher forcing

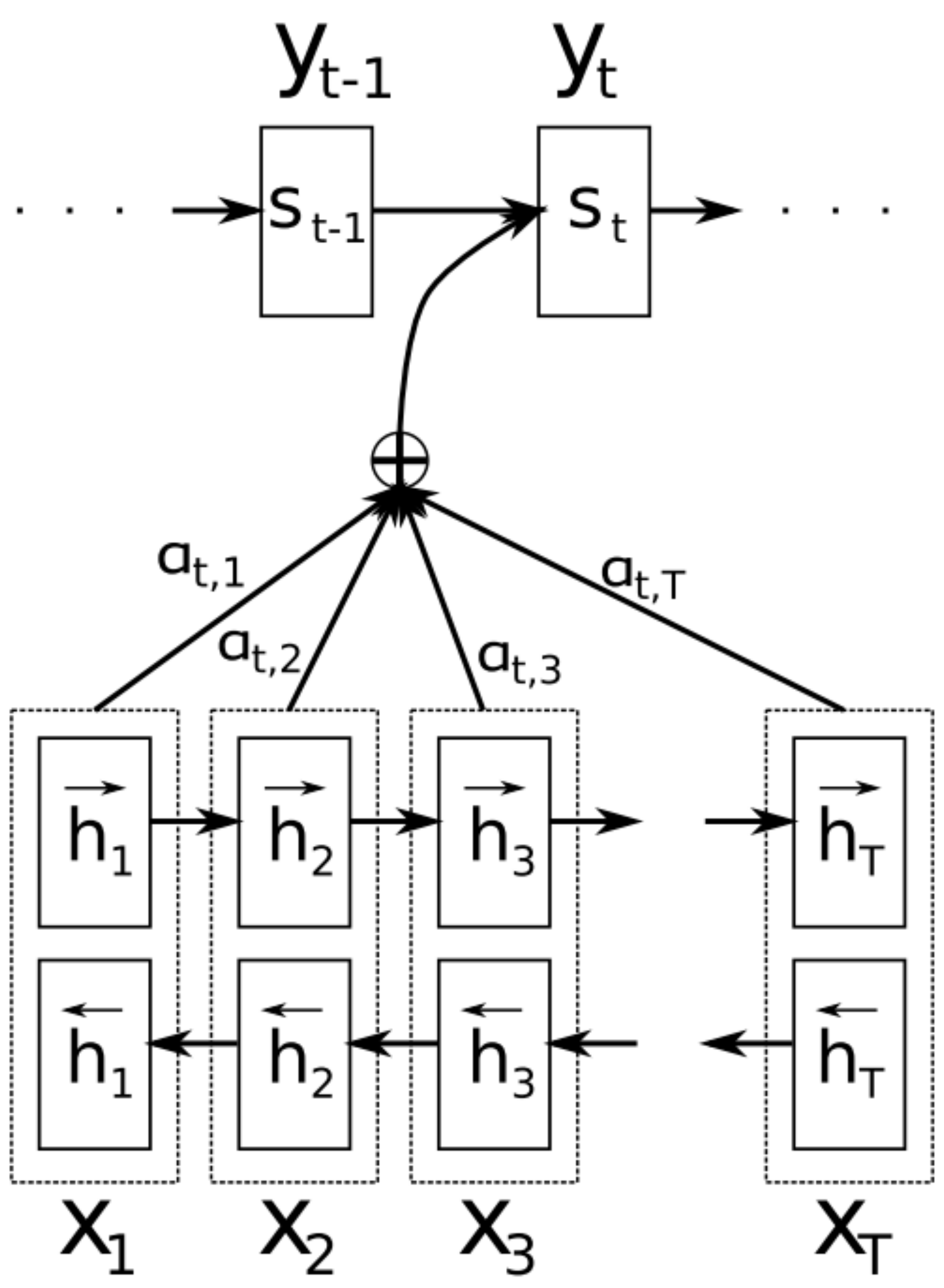
- *At early stage of training, the decoder can't reconstruct a reasonable first PF candidate; autoregressive mechanism propagates it into a wrong chain of particles.*

- ***Teacher-forcing:*** *under some probability k, feed the target as the next input instead of using the previous prediction. k decreases as the epoch number increases.*



Adding Attention

- *Attention allows the decoder to focus on which part of the inputs is relevant to the next prediction.*



[Bahdanau et al., arXiv:1409.0473]

the **Encoder** generates **h1,h2,h….hT** from the inputs **X1,X2,X3…XT**

a is the **Alignment model** which is a **feedforward neural network** that is trained with all the other components of the proposed system

$$e_{ij} = a(s_{i-1}, h_j)$$

The **Alignment model** scores (e) how well each encoded input (h) matches the current output of the decoder (s).

The alignment scores are normalized using a **softmax function.**

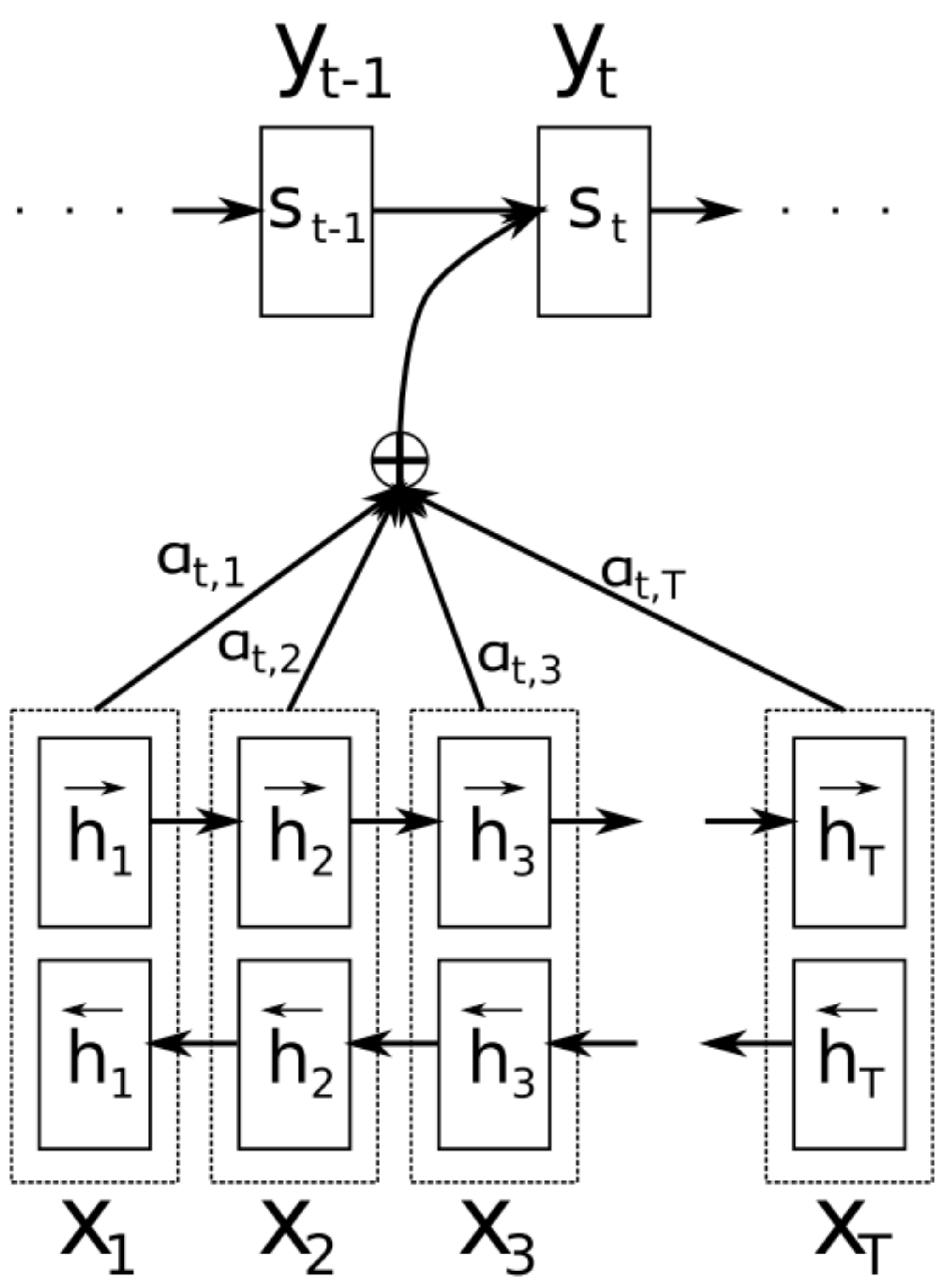
$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$

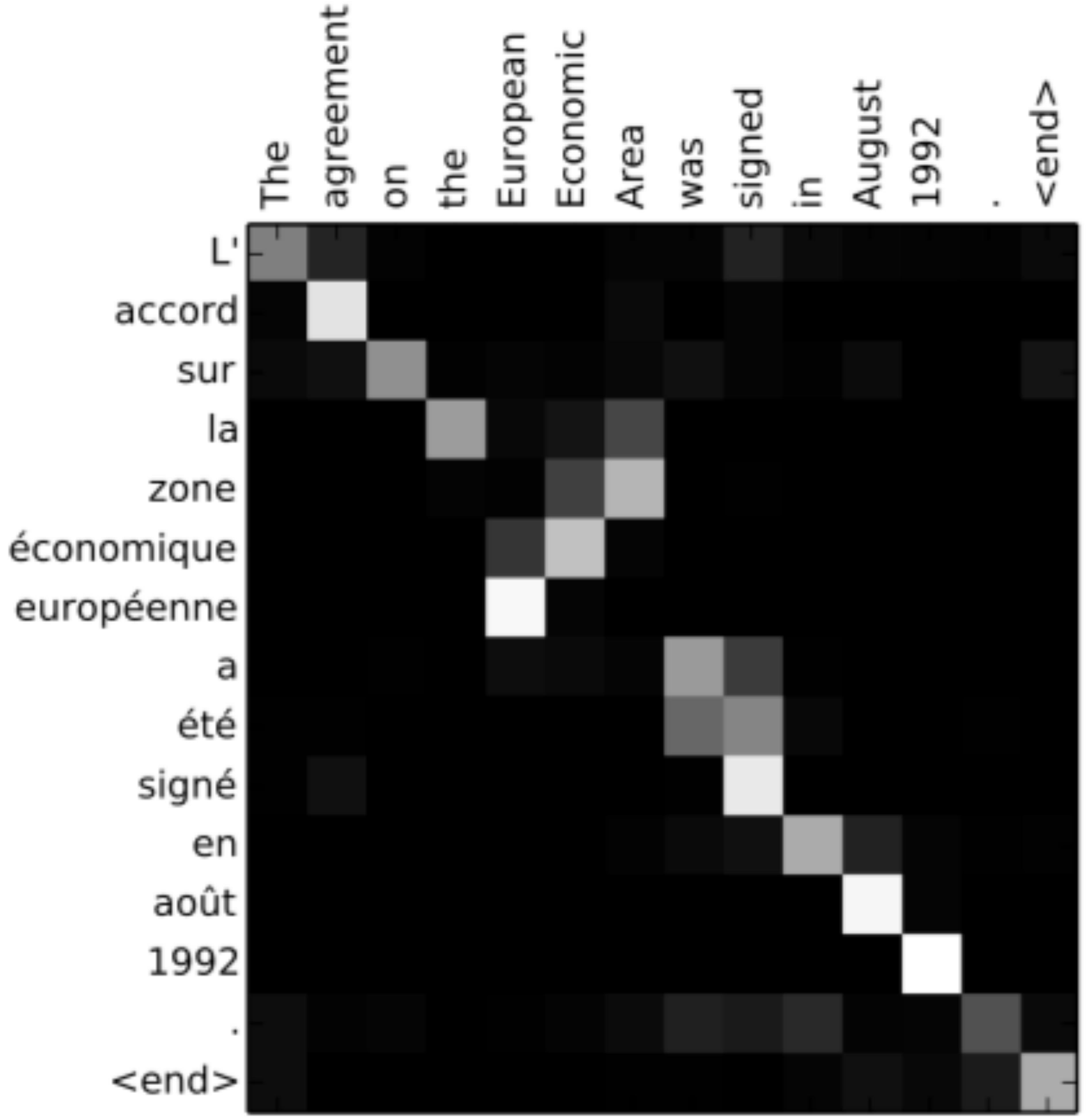
The context vector is a weighted sum of the **annotations** (hj) and **normalized alignment scores.**

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j.$$

Adding Attention

- *Attention allows the decoder to focus on which part of the inputs is relevant to the next prediction.*

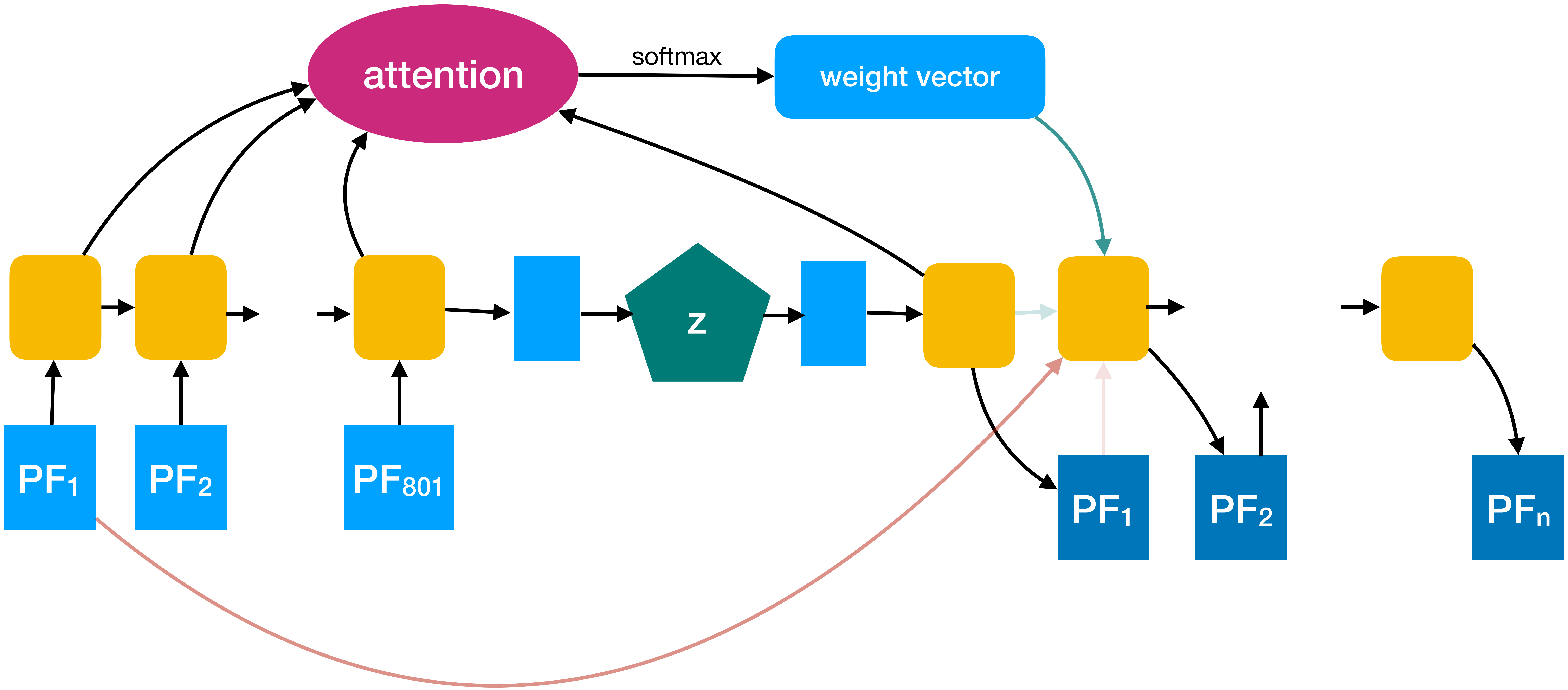




Bahdanau et al., arXiv:1409.0473

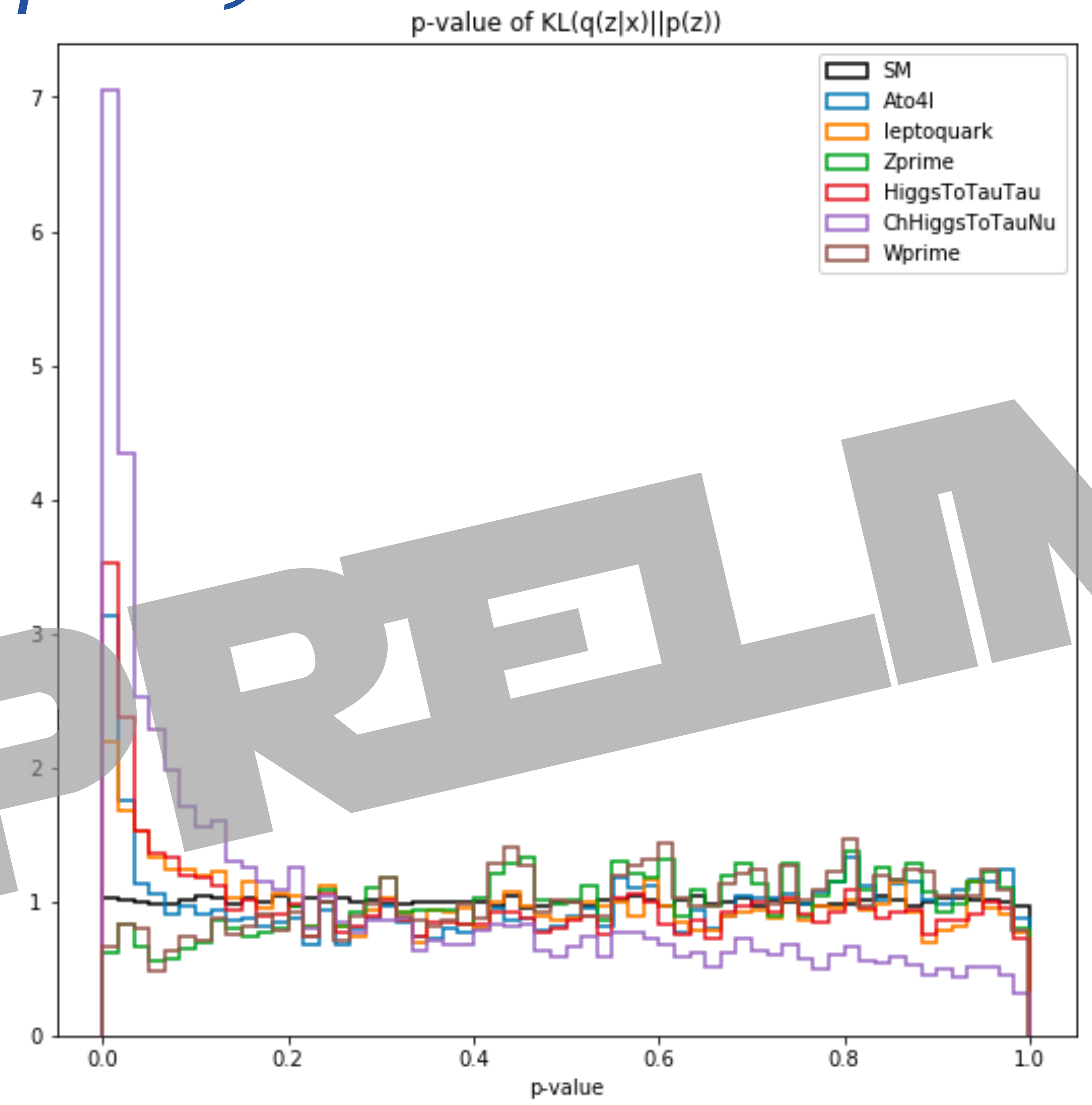
Adding Attention

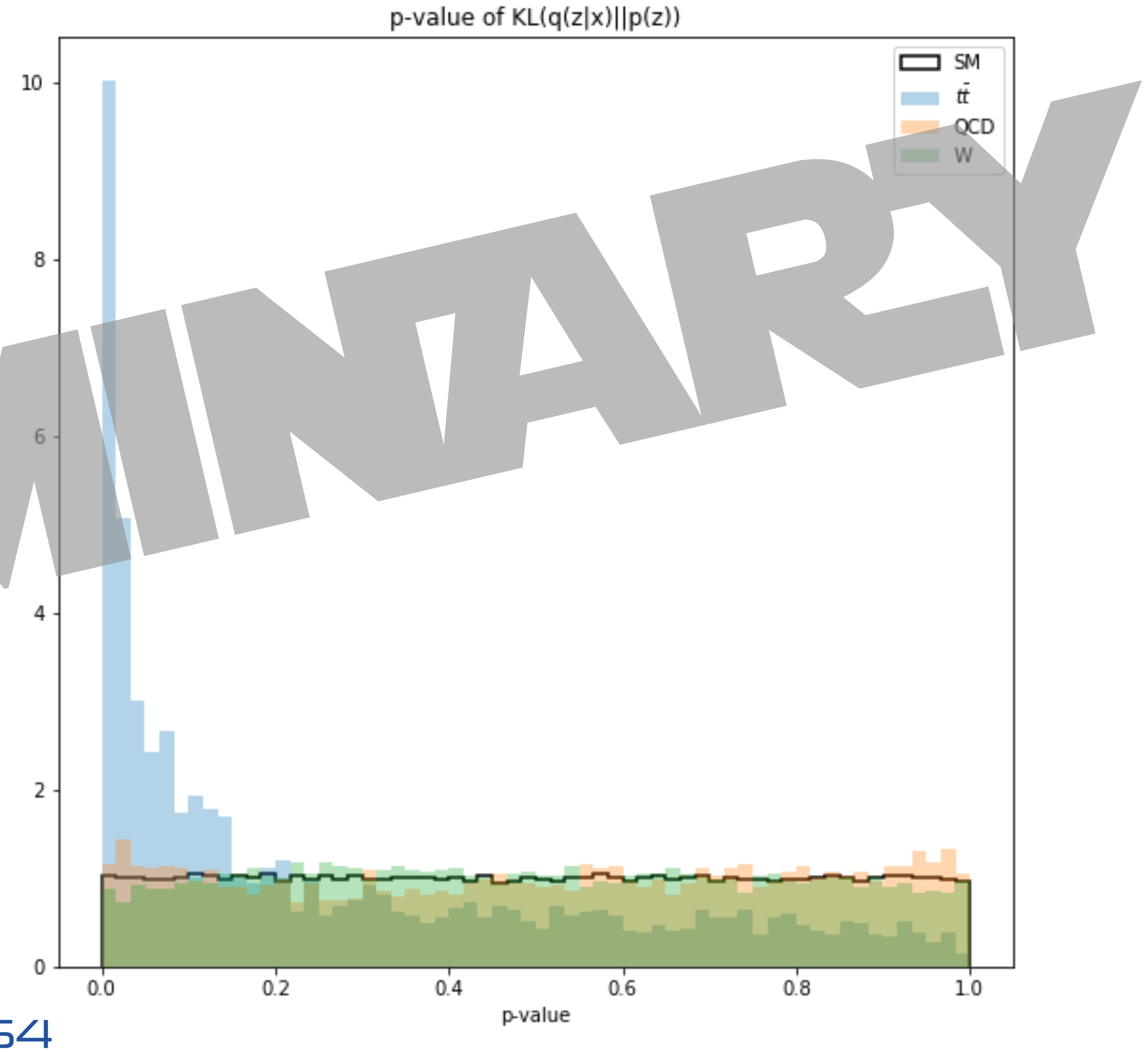
- *Attention allows the decoder to focus on which part of the inputs is relevant to the next prediction.*



Performances

- *(Preliminary) results trained on a small subset of the initial dataset (90K events)*
- *Due to architecture complexity, training is much slower (6h/epoch)*





Performances

- *(Preliminary) results trained on a small subset of the initial dataset (90K events)*

- *Due to architecture complexity, training is much slower (6h/epoch)*

Process	p-value = 0.05	p-value = 0.01	p-value = 0.001	p-value = 0.0001
$a \to 4\ell$	0.100	0.036	0.007	0.002
$LQ \to \tau b$	0.090	0.021	0.003	0.001
$h \to \tau\tau$	0.124	0.040	0.010	0.004
$h^{\pm} \to \tau\nu$	0.232	0.079	0.018	0.006

PRELIMINARY

Performances

- *(Preliminary) results trained on a small subset of the initial dataset (90K events)*
- *Due to architecture complexity, training is much slower (6h/epoch)*

Process	Efficiency for ~300 evt/day	xsec for 10 evt/ month [pb]	xsec for S/B~1/3 [pb]
$a \to 4\ell$	$3.3 \cdot 10^{-4}$	7.2	$1.5 \cdot 10^{3}$
$LQ \to \tau b$	$5.8 \cdot 10^{-4}$	4.1	850
$h \to \tau\tau$	$1.1 \cdot 10^{-3}$	2.2	450
$h^{\pm} \to \tau\nu$	$1.4 \cdot 10^{-3}$	1.7	340

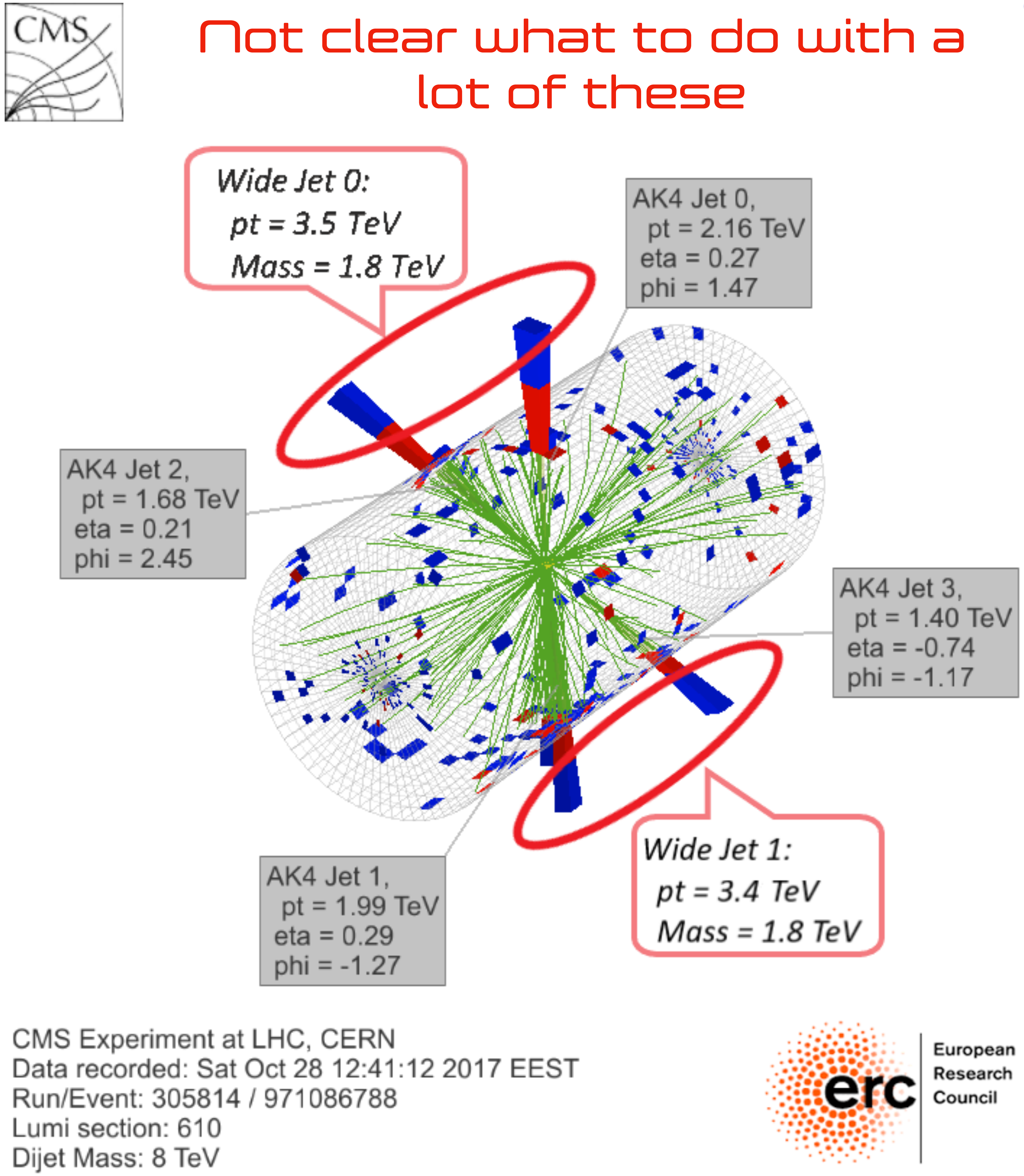
PRELIMINARY

How to use such an algorithm

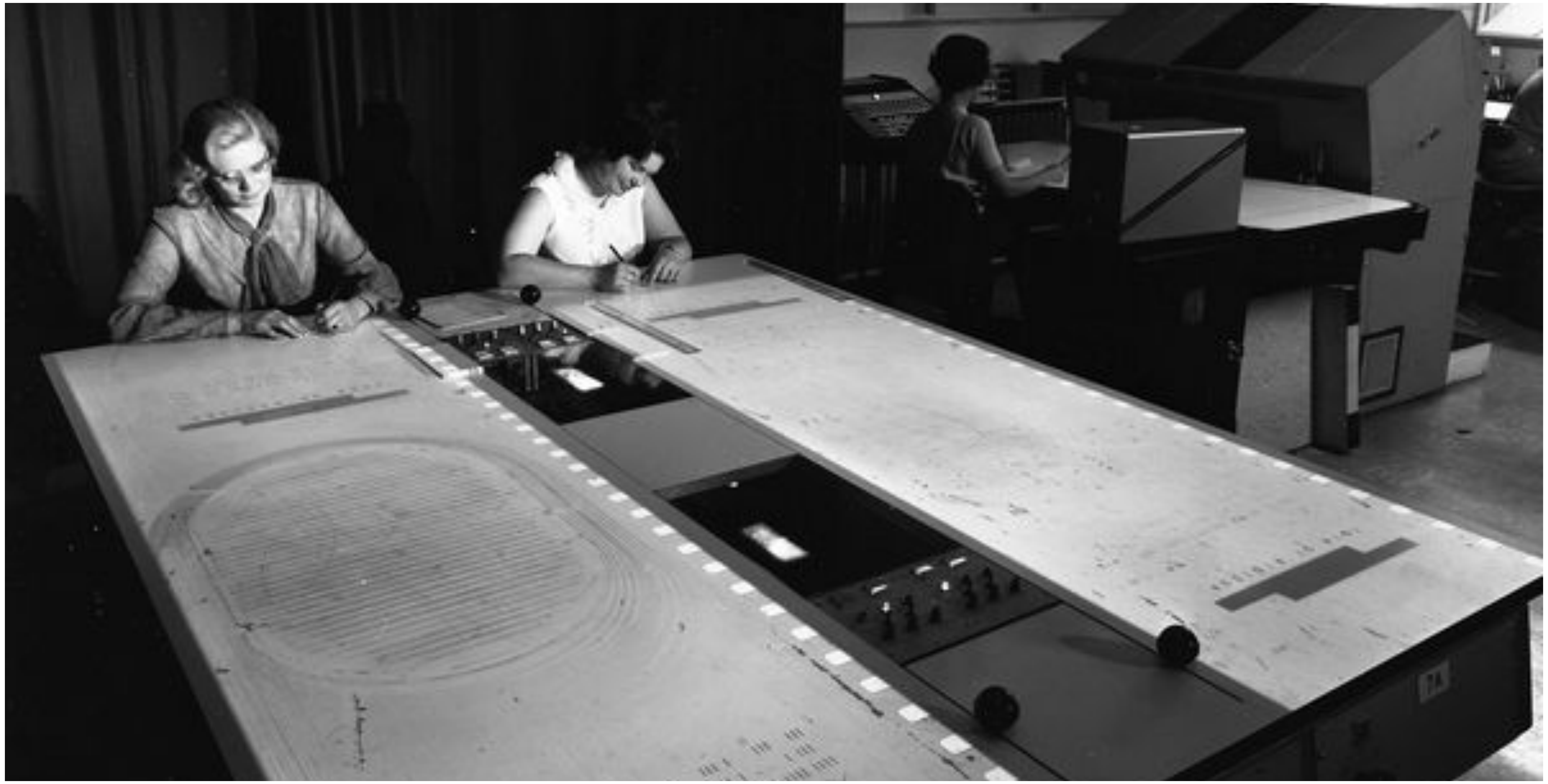
Not a discovery per-se

- As for model-independent searches, not a discovery tool per se
- One needs extra ingredients to translate what is found into a meaningful hypothesis test
  - Learn from data (data mining) and use the knowledge on new data
  - Add information (e.g., expected background model) and use “bsm-agnostic” hypothesis testing
  - Scan the events with some advanced tool
  - …

Not clear what to do with a lot of these



Visual inspection



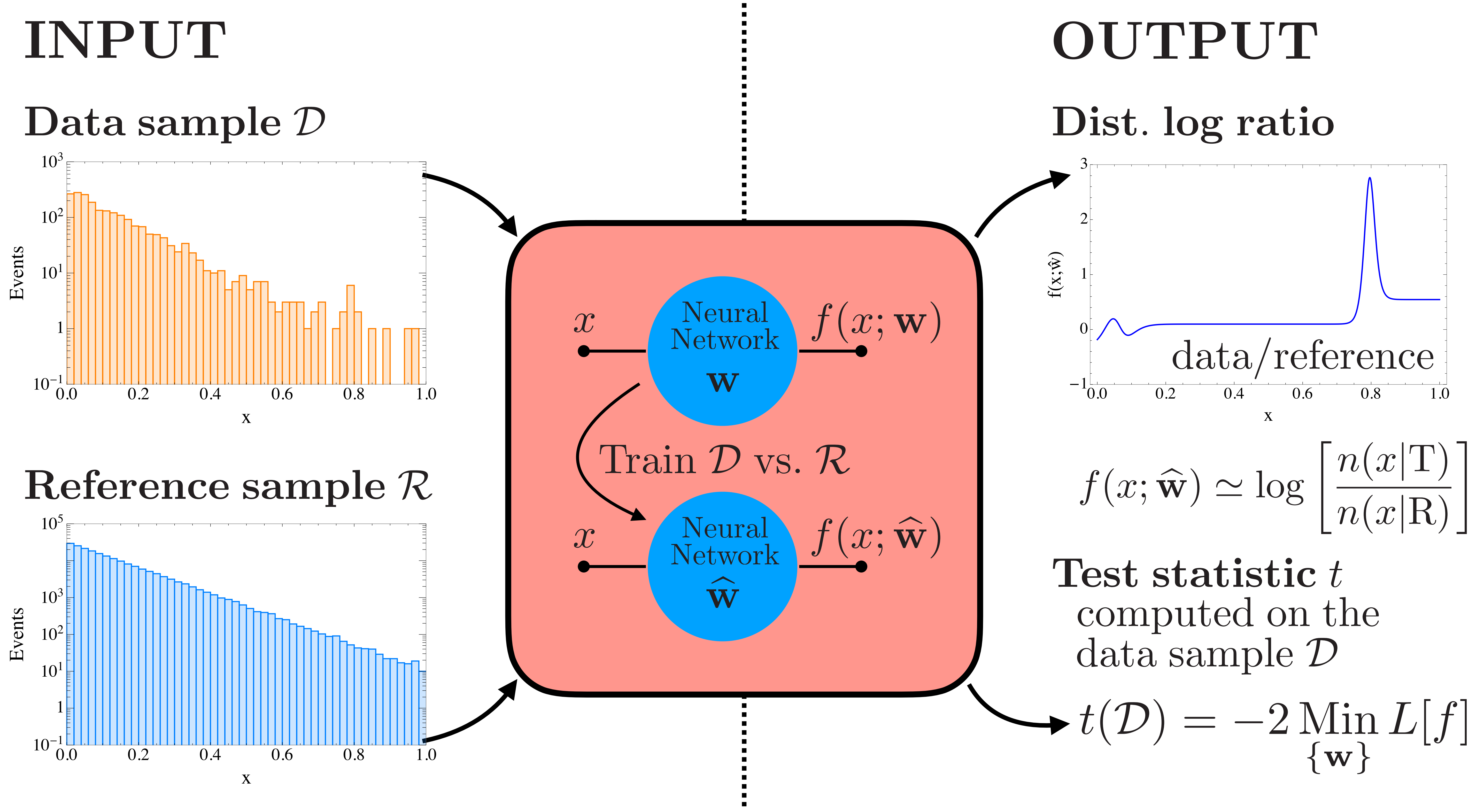
- *Nothing new (we used to do this in the past)*
- *In principle, one could release a catalog for you to play with it*

Learning NP from a Machine

- *With description of the SM samples that would be selected (e.g., MC if MC accurate) one could run hypothesis testing w/o specifying the signal model*

- *This would allow to "isolate" the anomalous events looking at the returned contribution to the likelihood ratio*

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



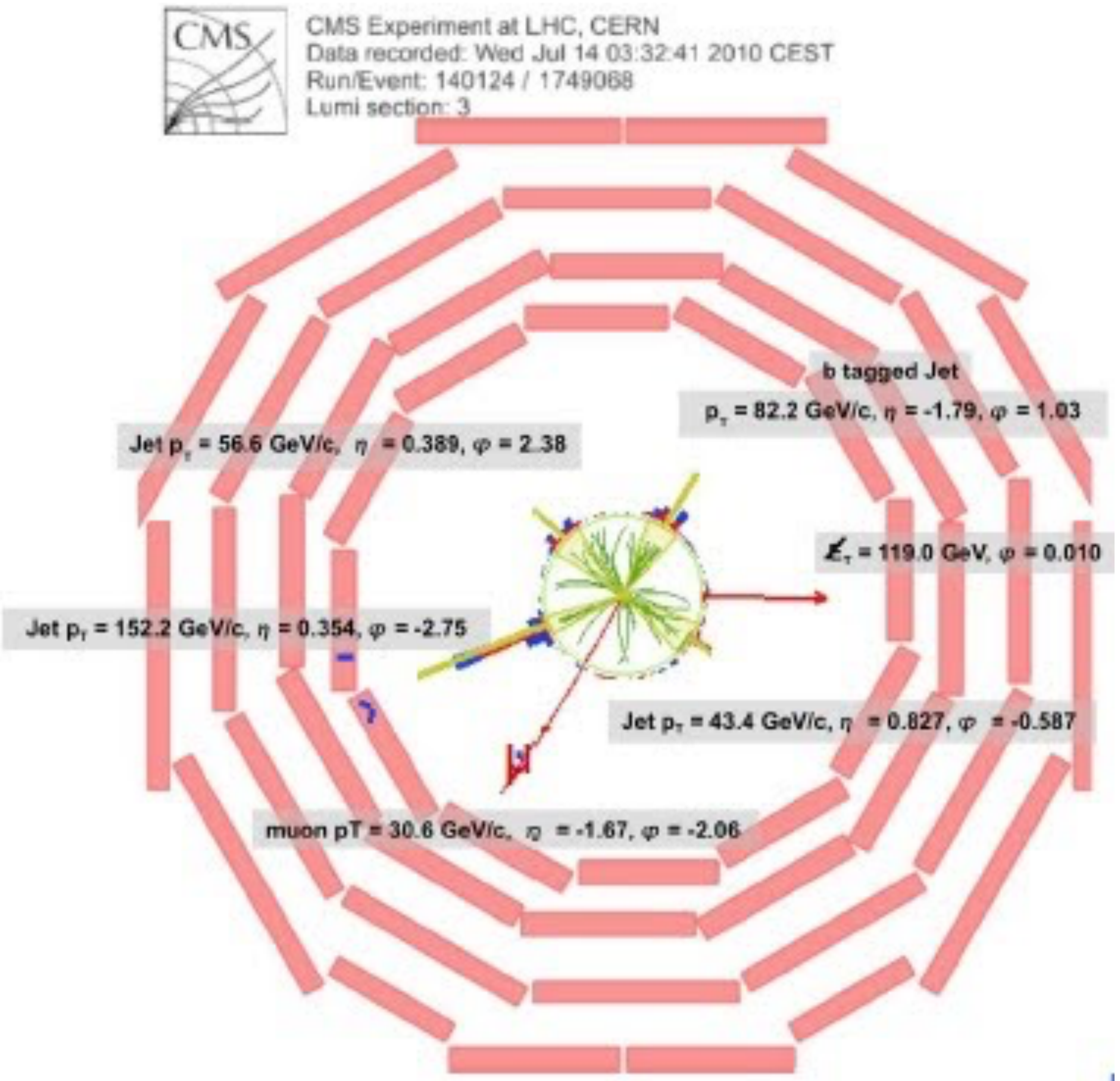
[Dagnolo and Wulzer, arXiv:1806.02350]

SSD

- BSM events might still be made of standard objects (jets, leptons, etc)

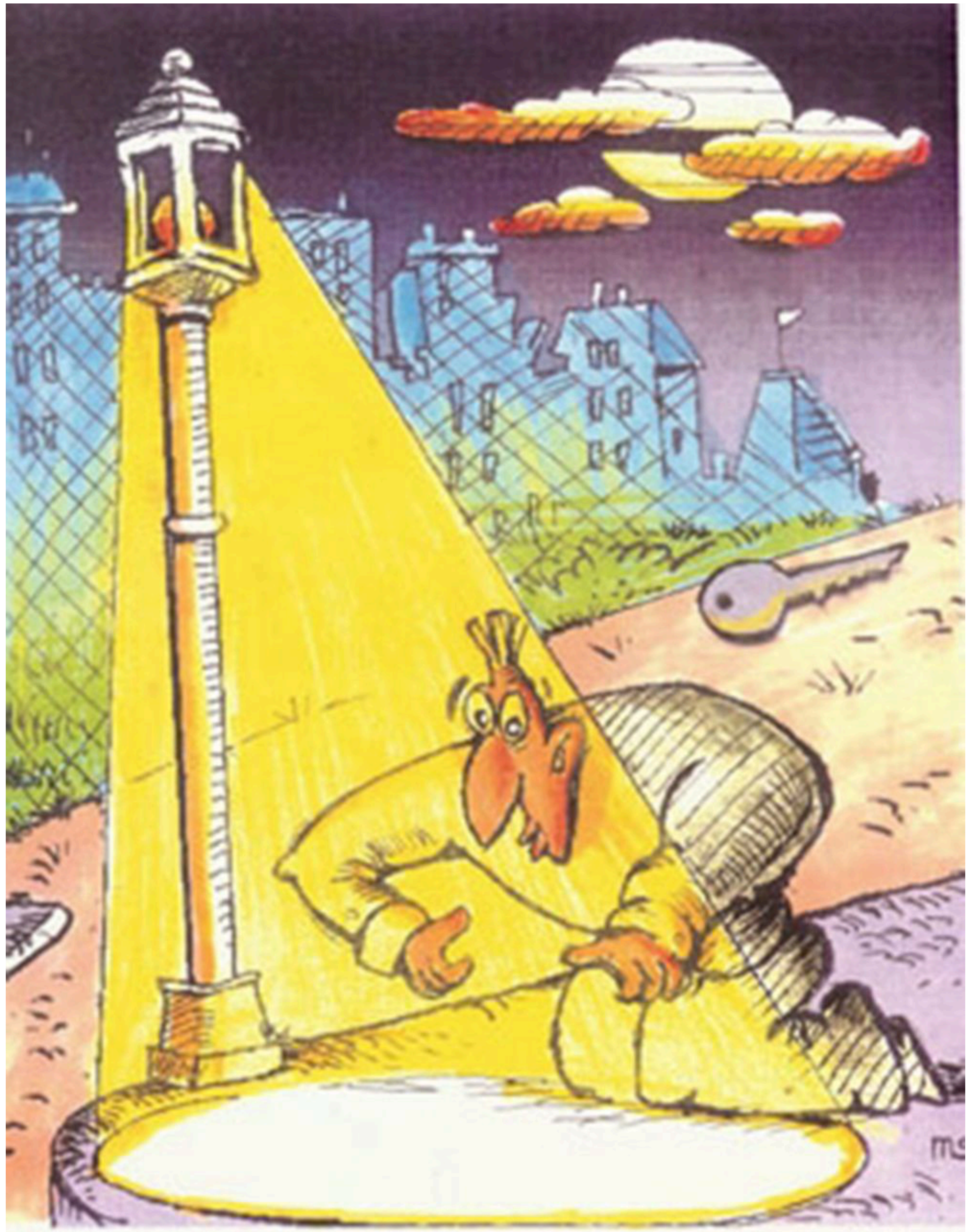
- Single Shot Multibox Detector could tell us what is in each event

- Multiplicity might be enough to highlight a pattern in the selected anomalies



Conclusions

- *The LHC Big Data problem might be fooling us: are we rejecting new physics events because we started with the wrong portfolio of BSM scenarios?*
- *We need an alternative strategy to act as an insurance, while we keep following the canonical strategy*
- *We propose to use autoencoders as anomaly detection tools running in the trigger, to let the data guide our search*
- *The ultimate goal is to select O(10) events/day and create a catalog of anomalous events, for further study within and outside the collaborations*
- *Hopefully, this might open our eyes towards new directions*



Backup

Pdf modeling

- **Clipped Log-normal + δ function**: used to describe S_T, M_J, p_T^μ, M_μ, p_T^e, M_e, isolated-lepton p_T, ChPFIso, NeuPFIso and GammaPFIso:

$$P(x \mid \alpha_1, \alpha_2, \alpha_3) = \begin{cases} \alpha_3 \delta(x) + \frac{1-\alpha_3}{x\alpha_2\sqrt{2\pi}} \exp\left(\frac{(\ln x - \alpha_1)^2}{2\alpha_2^2}\right) \ for\ x \geq 10^{-4} \\ 0\ for\ x < 10^{-4} \end{cases} . \quad (9)$$

- **Gaussian**: used for $p_{T,\parallel}^{\text{miss}}$ and $p_{T,\perp}^{\text{miss}}$:

$$P(x \mid \alpha_1, \alpha_2) = \frac{1}{\alpha_2\sqrt{2\pi}} \exp\left(\frac{(x-\alpha_1)^2}{2\alpha_2^2}\right) . \quad (10)$$

- **Truncated Gaussian**: a Gaussian truncated for negative values and normalized to unit area for $X > 0$. Used to model M_T:

$$P(x \mid \alpha_1, \alpha_2) = \Theta(x) \cdot \frac{1 + 0.5 \cdot (1 + \text{erf}\frac{-\alpha_1}{\alpha_2\sqrt{2}})}{\alpha_2\sqrt{2\pi}} \exp\left(\frac{(x-\alpha_1)^2}{2\alpha_2^2}\right) . \quad (11)$$

- **Discrete truncated Gaussian**: like the truncated Gaussian, but normalized to be evaluated on integers (i.e. $\sum_{n=0}^{\infty} P(n) = 1$). This function is used to describe N_μ, N_e, N_b and N_J. It is written as:

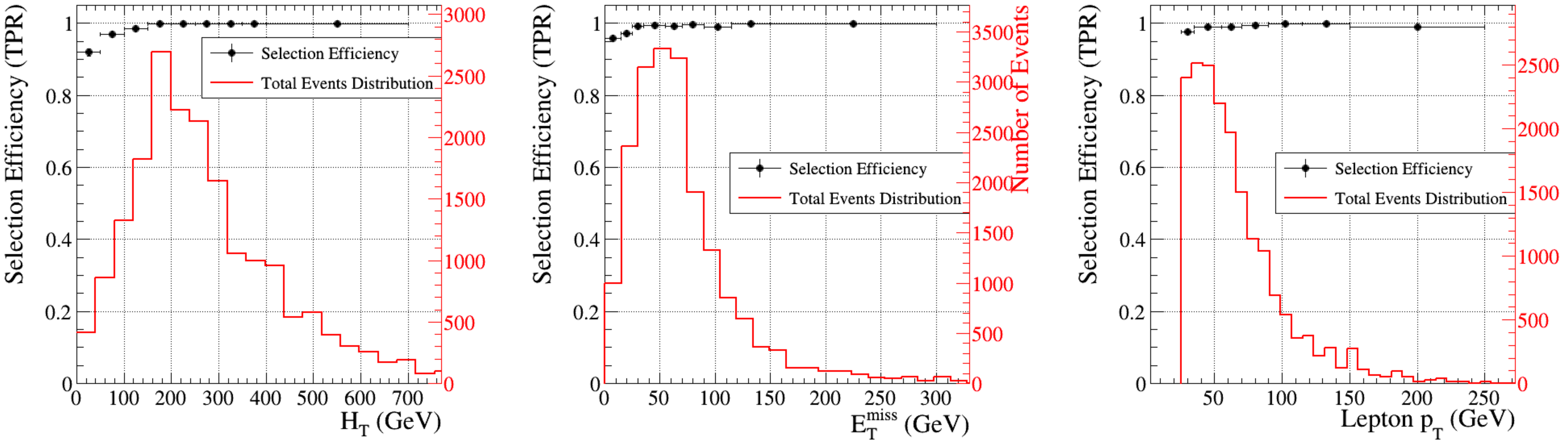
$$P(n \mid \alpha_1, \alpha_2) = \Theta(x) \left[\text{erf}\left(\frac{n + 0.5 - \alpha_1}{\alpha_2\sqrt{2}}\right) - \text{erf}\left(\frac{n - 0.5 - \alpha_1}{\alpha_2\sqrt{2}}\right)\right] \mathcal{N} , \quad (12)$$

where the normalization factor $\mathcal{N}$ is set to:

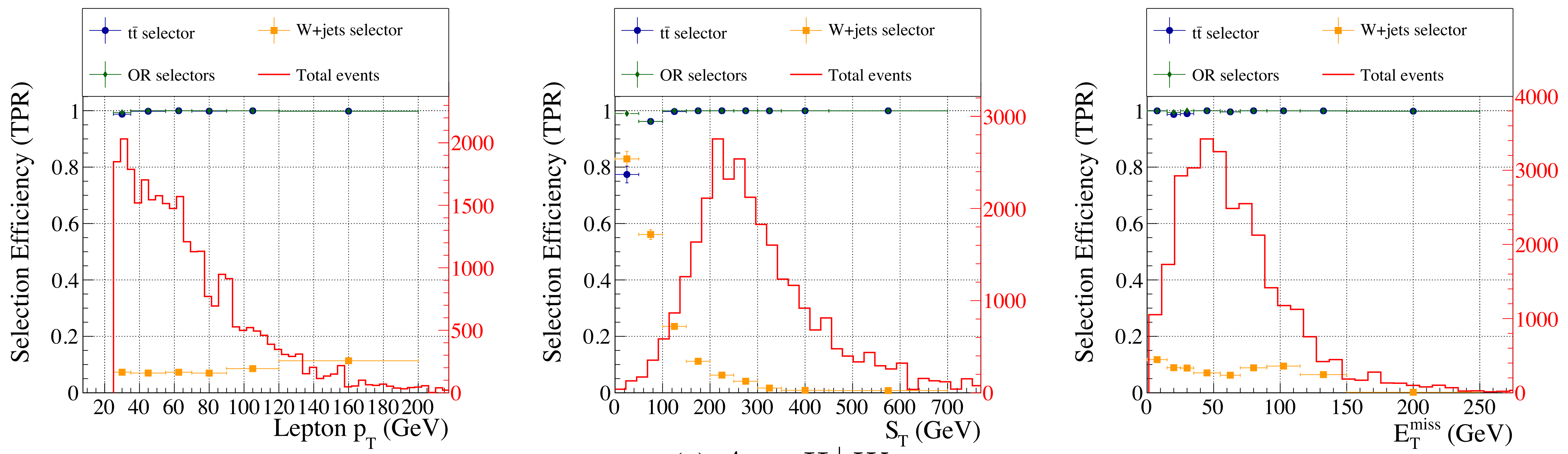
$$\mathcal{N} = 1 + \frac{1}{2}\left(1 + \text{erf}\left(\frac{-0.5 - \alpha_1}{\alpha_2\sqrt{2}}\right)\right) \quad (13)$$

Kinematic Bias?

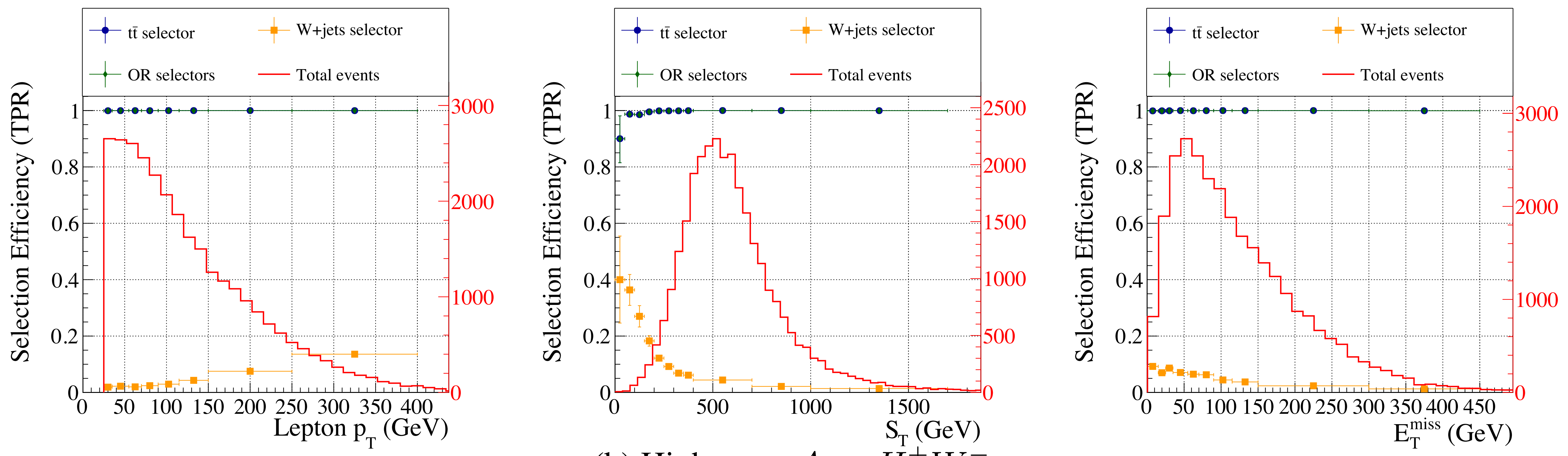
- *With 99% signal efficiency, bias on kinematic variables within the uncertainty of a trigger-efficiency measurement*



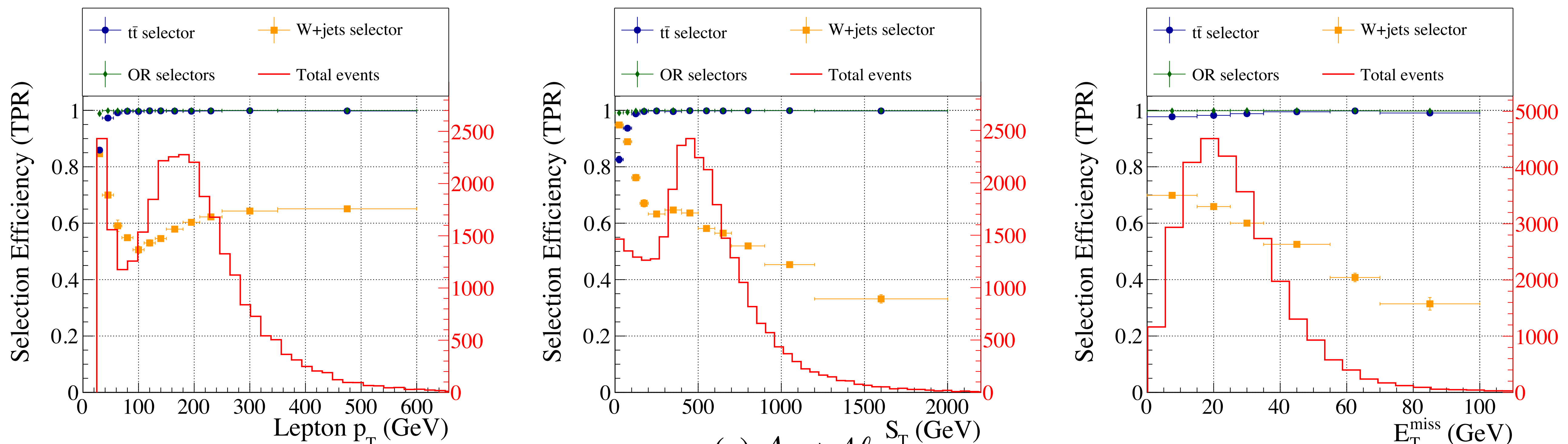
TOPCLASS: do we kill New Physics?



(a) $A \to H^+W^-$

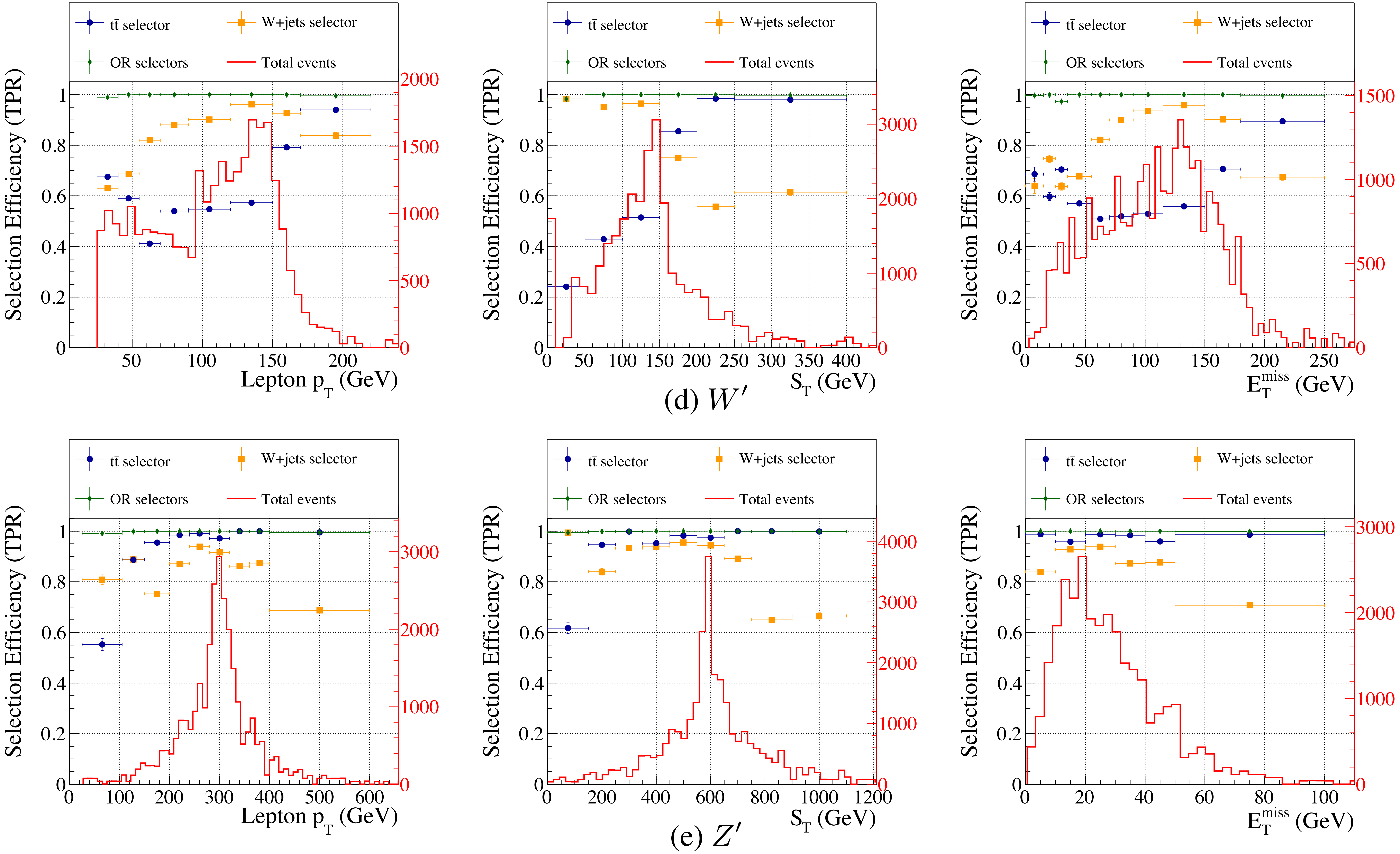


(b) High-mass $A \to H^+W^-$



(c) $A \to 4\ell$

TOPCLASS: do we kill New Physics?



(d) W'

(e) Z'