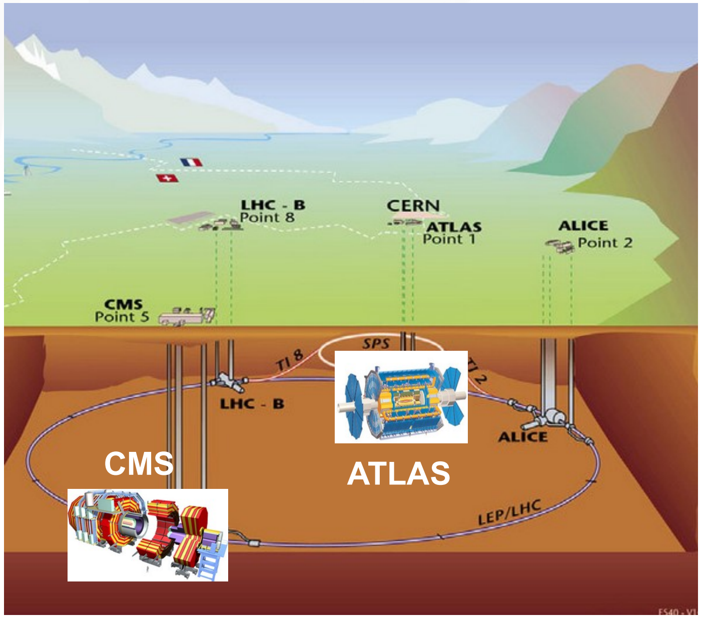
Search for New Physics at the LHC Using Unsupervised Machine Learning for Anomaly Detection

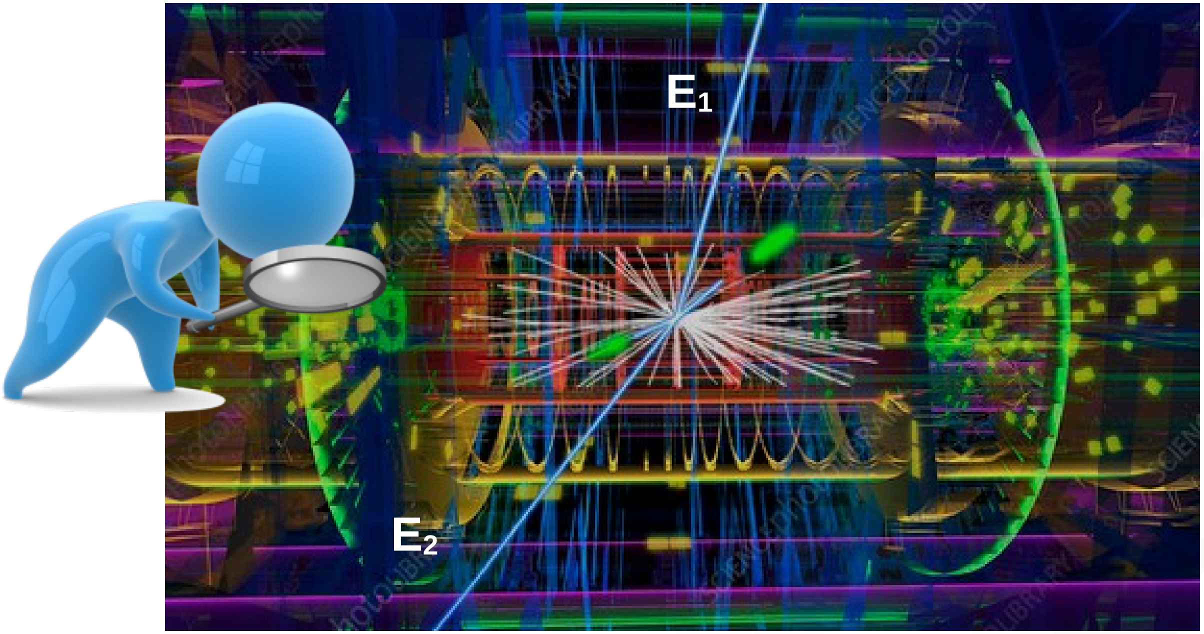
S.Chekanov (HEP/ANL)

Feb 26, 2026

In collaboration with colleagues from HEP/ANL and ATLAS collaboration

LHC experiments





Abstract

Title:

Search for New Physics at the LHC Using Unsupervised Machine Learning for Anomaly Detection

Abstract:

Traditionally, searches for new physics at the LHC rely on selections designed to target regions of phase space where sensitivity to beyond-the-Standard-Model (BSM) models is highest. This talk will summarize an alternative, model-agnostic approach, largely developed by the HEP/ANL group, using an autoencoder — a deep neural network that learns to reconstruct the primary kinematic features of collision events. Events that are poorly reconstructed are flagged as outliers (anomalous regions), providing a model-independent way to search for potential new physics.The talk will review the foundations of this method as presented in journal publications, along with related ATLAS results. Future prospects for the High-Luminosity LHC will also be discussed.

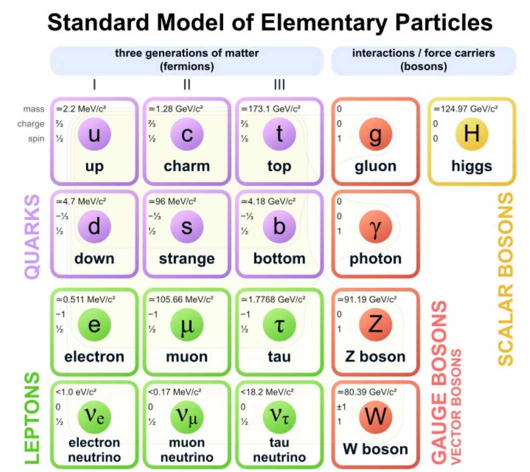
New particles at collider experiments

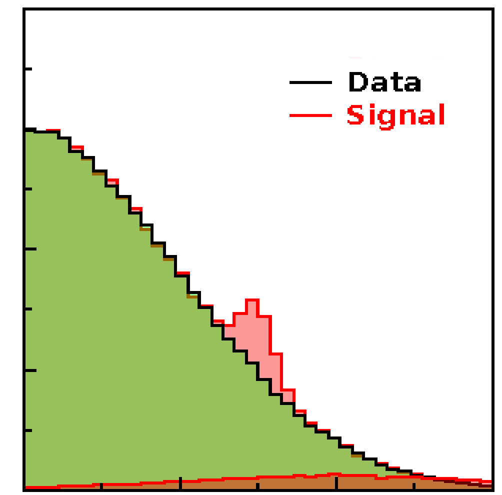
- Standard Model (SM) is successful for particle collisions
- Discrepancies may indicate new physics ≡ new particles/fields

- **<u>Direct observations</u> of new particles**
  - Combine known particles to create "invariant masses" & search for "resonance" enhancements above background
  - Or observe through unusual signatures in detector (anomalously high dE/dx tracks etc)

- **<u>Indirect observations</u> of new particles**
  - Compare SM predictions with data
  - Search for any discrepancy with SM background
  - Explain using theoretical frameworks beyond SM (BSM)

No evidence yet but no shortage of models predicting exotic heavy particles





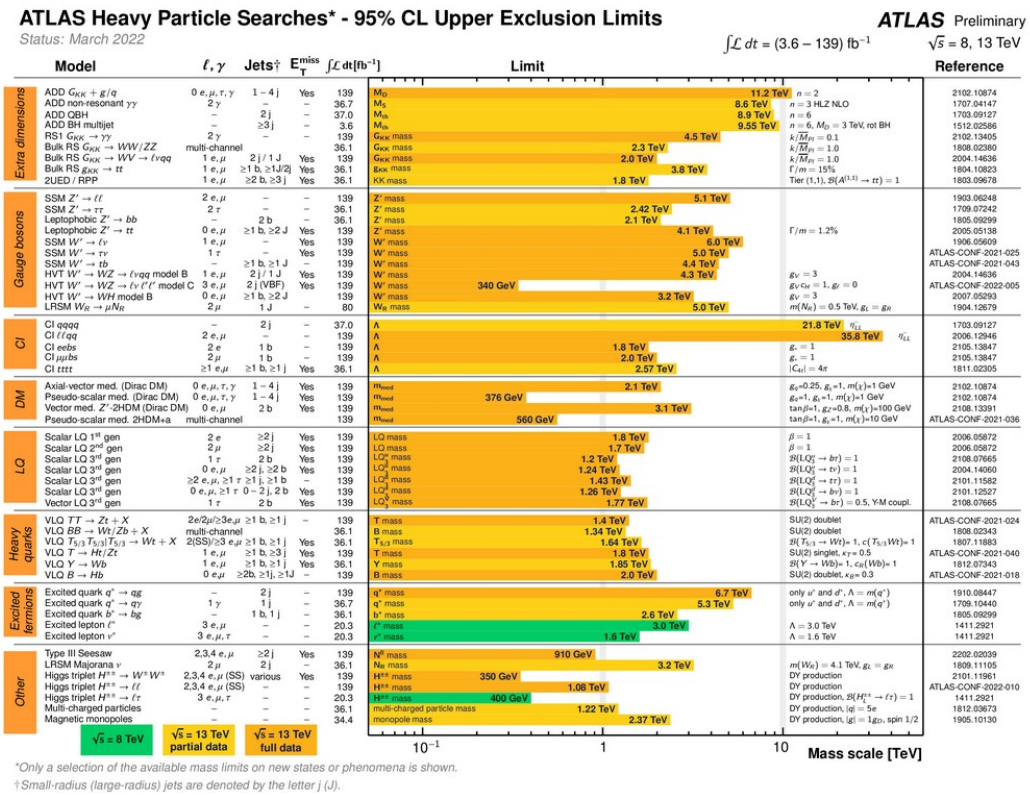
$$M^2 = (E_1 + E_2)^2 - \|\mathbf{p}_1 + \mathbf{p}_2\|^2$$

Invariant mass from known particles with energy E and **p**

LHC limits for direct and indirect BSM searches

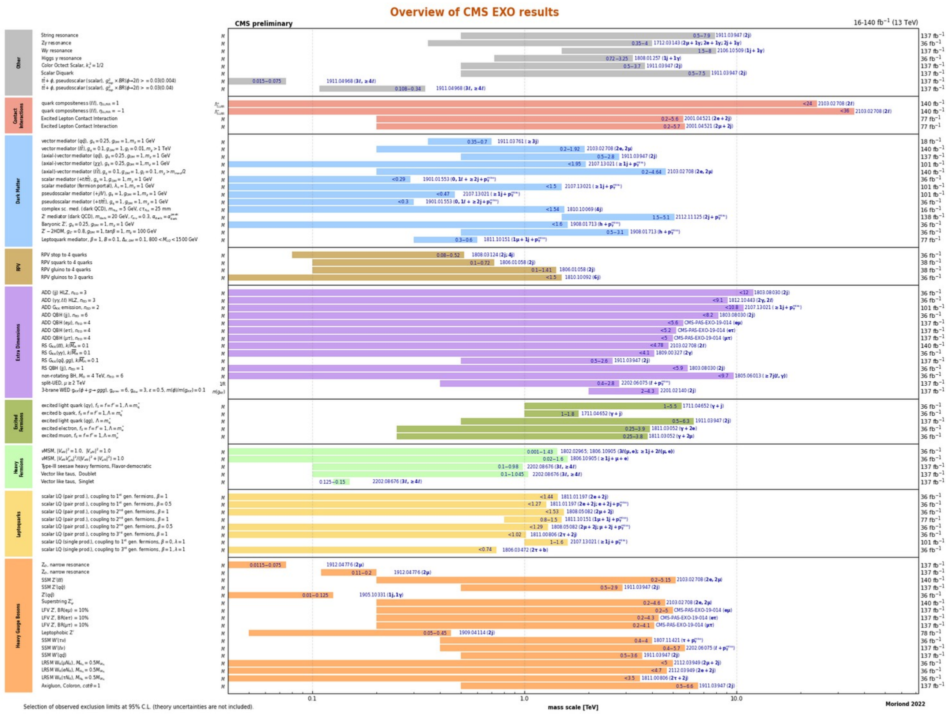
ATLAS

URL link to image



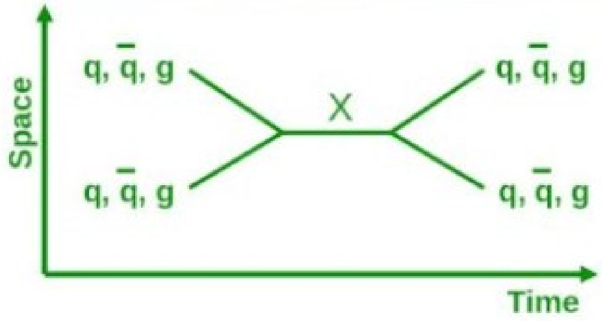
CMS

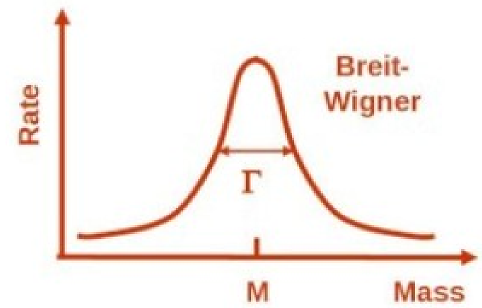
URL link to image



- ~50 decay channels studied for various BSM models *(extra dimensions, gauge bosons, contact interactions, dark matter, heavy quarks, excited fermions, leptoquarks etc)*
- Commonly excluded masses ~ **0.4 – 12 TeV**
- But plenty of models that predict too small cross section for exclusion!

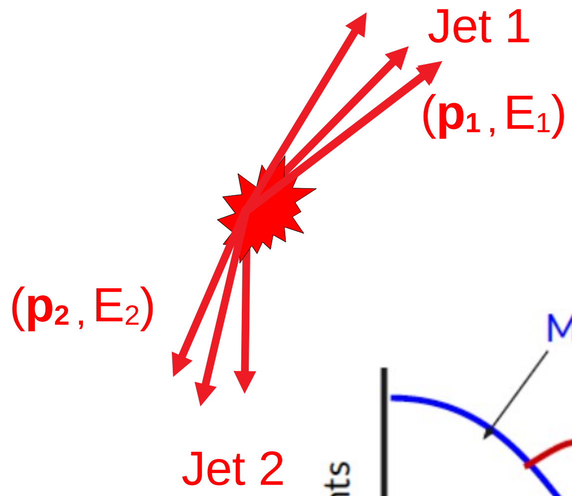
Search for high mass dijet resonances using model-independent approach

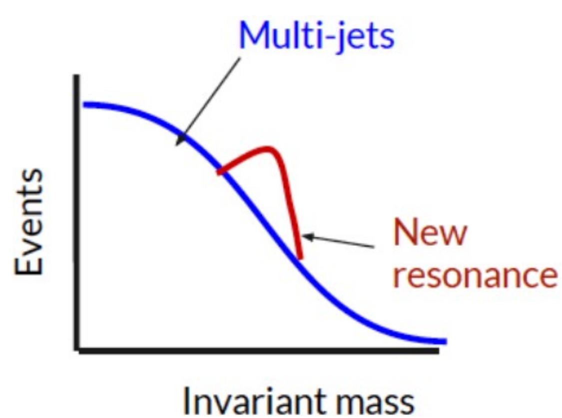




Resonance shapes depend on qq, qg and gg interactions

$$M^2 = (E_1 + E_2)^2 - \|\mathbf{p}_1 + \mathbf{p}_2\|^2$$





How to find a heavy state:

- Calculate invariant mass from 2 hadronic jets
- Fit with smooth analytic functions

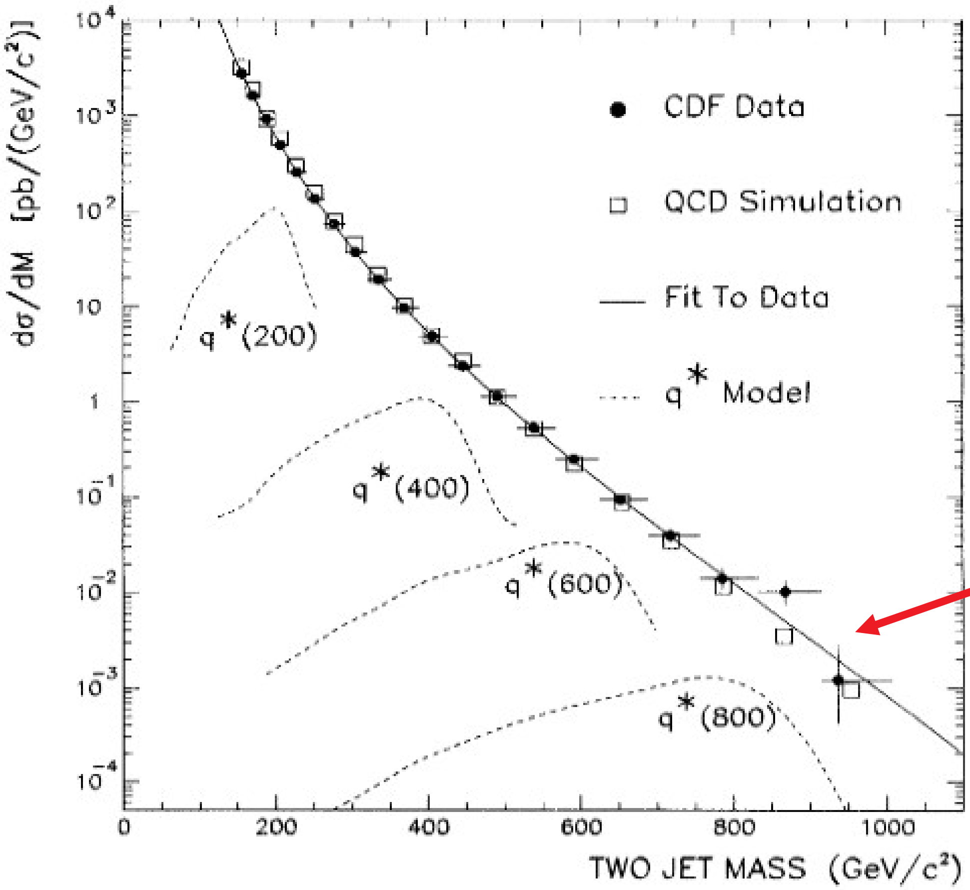
Signatures of Z'/W' bosons:
(review by P.Langacke Rev.Mod.Phys.81:1199 (2009))

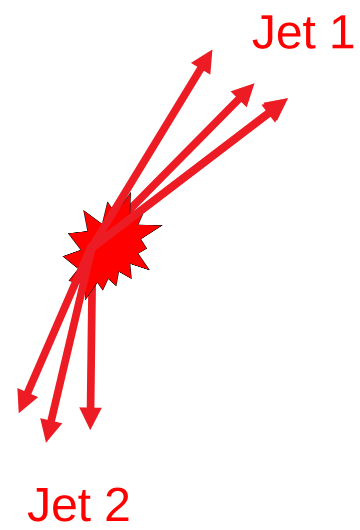
- *Similar to the SM W/Z bosons (but heavier)*
- *Extending SM to group SU(3) × SU(2) ×* ***U(1)***
- *Sequential Standard Model*
- *Grand unified theories, fine tuning problem*
- *Extra dimensions*
- *Dark matter mediator etc. etc.*

Early measurements – I (CDF/TEVATRON)

CDF Collaboration, Phys Rev Lett.74 (1995) 3538

L ~ 19 pb^{-1}





3 – parameter function:

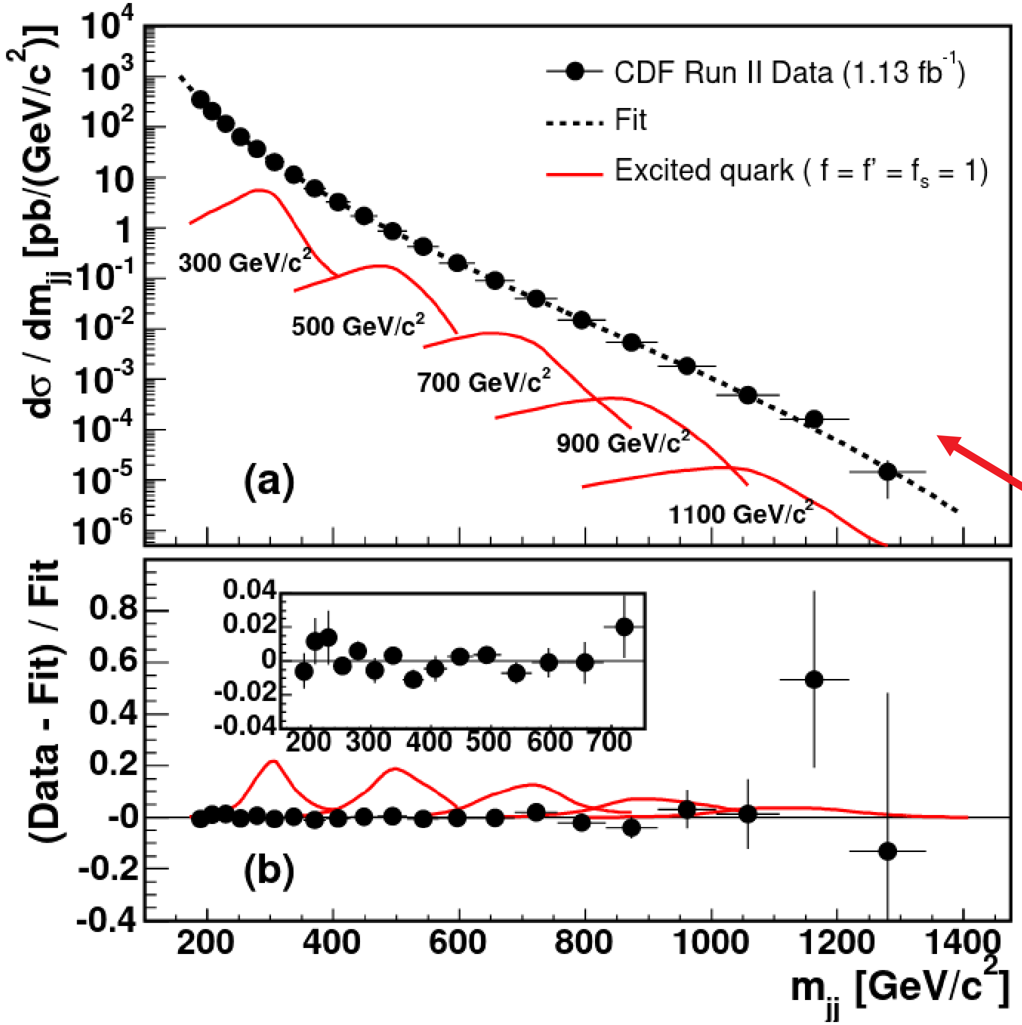
$$f(x) = p_1(1-x)^{p_2} x^{p_3}$$

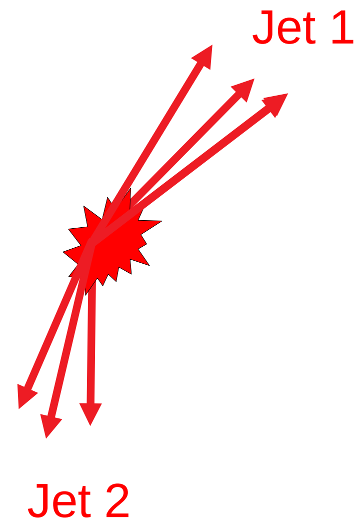
$$x \equiv m_{jj}/\sqrt{s}$$

Early measurements – II (CDF/TEVATRON)

CDF, Phys.Rev.D79 (2009) 112002

L ~ 1.1 fb^{-1}





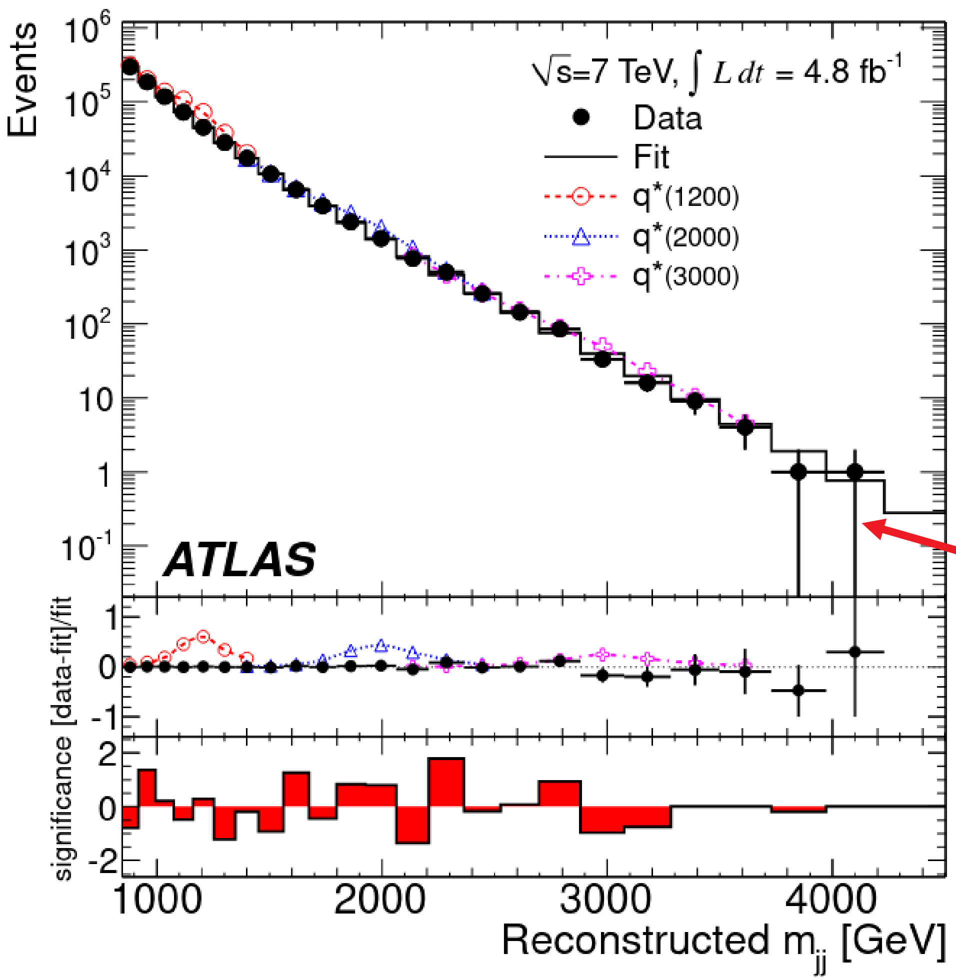
4 – parameter function:

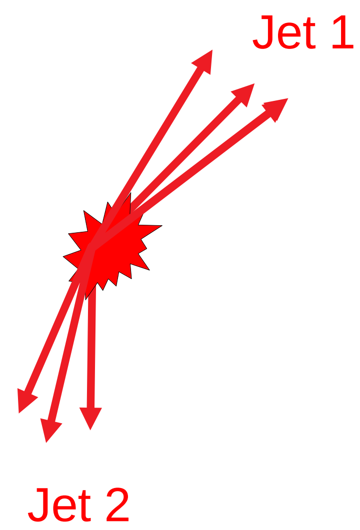
$$f(x) = p_1(1-x)^{p_2} x^{p_3 + \boxed{p_4 \ln x}}$$

$$x \equiv m_{jj}/\sqrt{s}$$

Early measurements – II (ATLAS/LHC)

ATLAS collaboration, JHEP 1301 (2013) 029





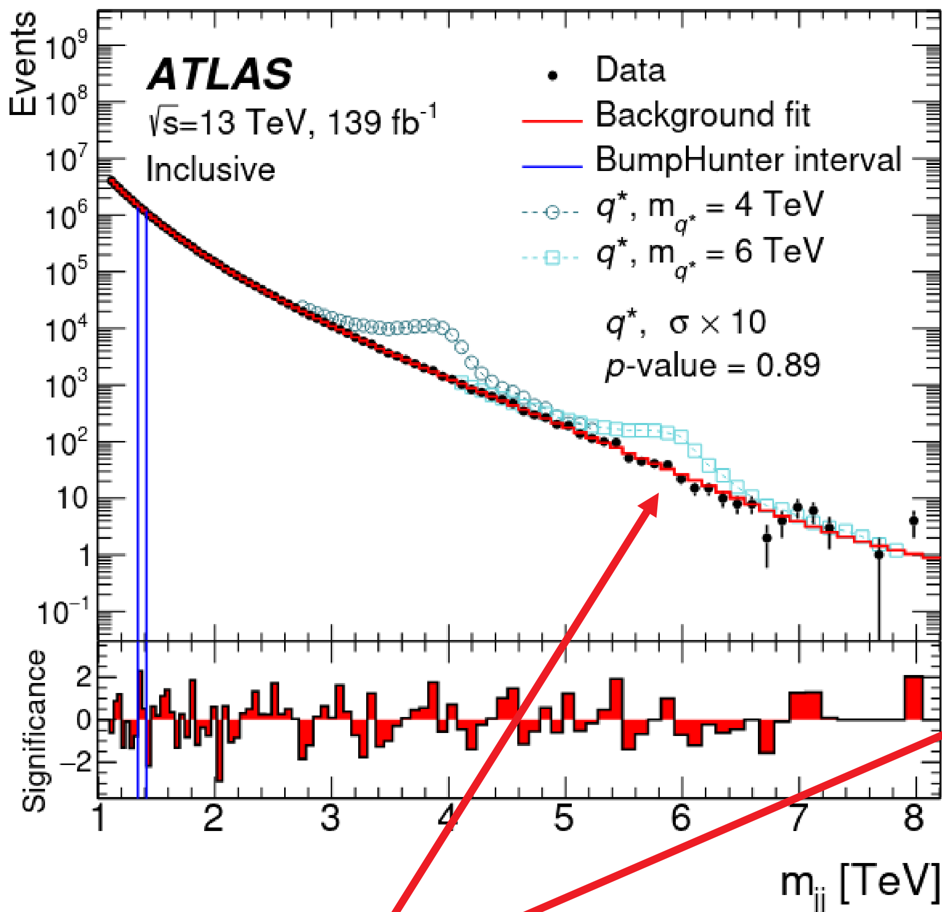
4 – parameter function:

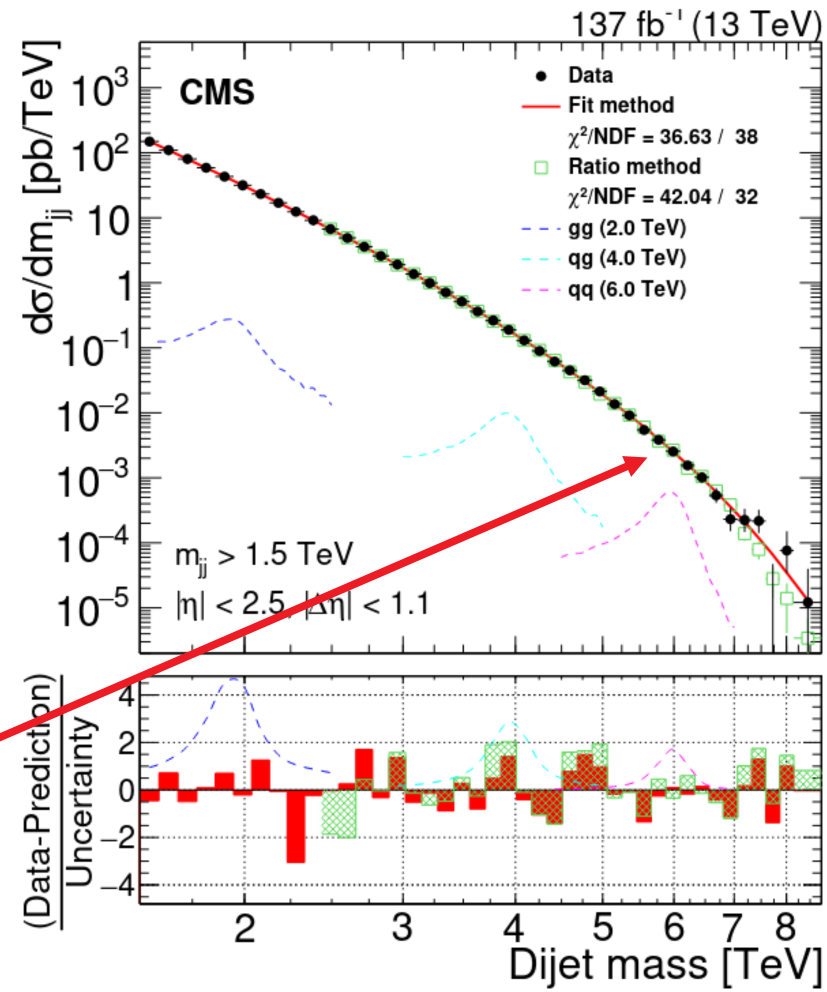
$$f(x) = p_1(1-x)^{p_2} x^{p_3 + \boxed{p_4 \ln x}}$$

$$x \equiv m_{jj}/\sqrt{s}$$

Recent high-statistics data

ATLAS [JHEP 03 (2020) 145](https://doi.org/)
CMS JHEP 05 (2020) 033





$$f(x) = p_1 (1-x)^{p_2} \, x^{p_3 + p_4 \ln x + \boxed{p_5 (\ln x)^2}} \qquad x \equiv m_{jj}/\sqrt{s}$$

* ATLAS uses "sliding" window fit

What do we know about the p5 fit function?

“p5” function:

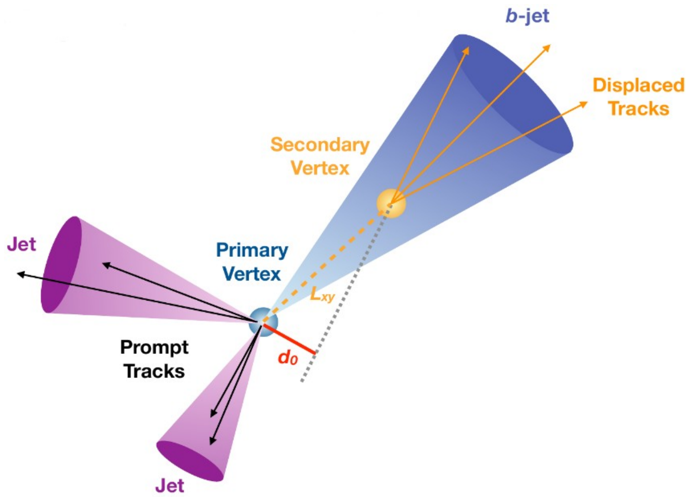
$$f(x) = p_1 (1-x)^{p_2} x^{p_3 + p_4 \ln x + p_5 (\ln x)^2}$$

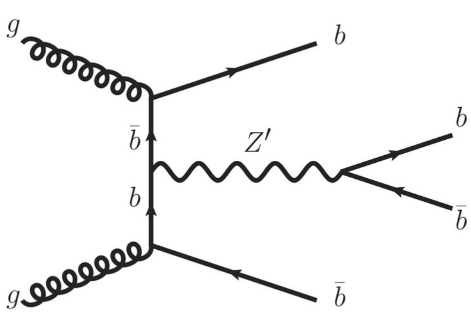
- First part - $p_1 (1-x)^{p2}$ - resembles the QCD splitting function
- Other 2 factors – pure phenomenology. But they work !
- Introduced / studied largely at ANL in 2018 by J.Bodwin – (ATLAS associate position / consultant).

Jets originating from b-quark

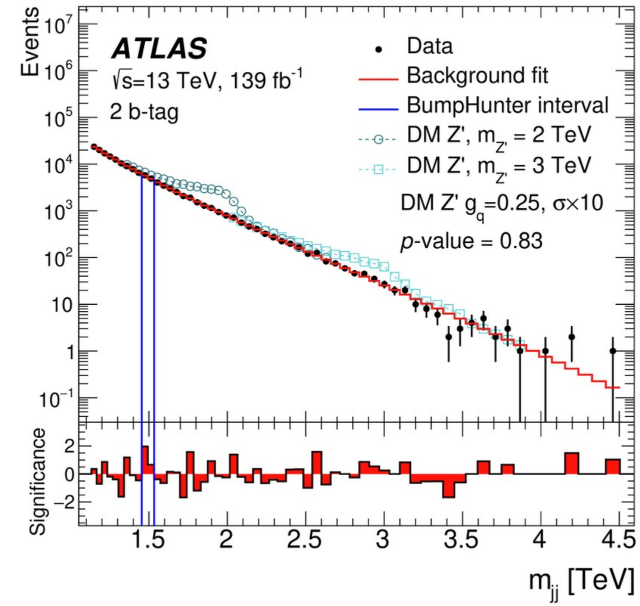
ATLAS JHEP 03 (2020) 145
ATLAS Phys. Rev. D 105 (2022) 012001
ANL/HEP contribution

- "Leptophobic" Z' (does not decay to leptons), can couple to third-generation quarks for some modes
- All-jet searches may fail. Instead, search for Z' → $b\bar{b}$ by combining jets from b-quarks

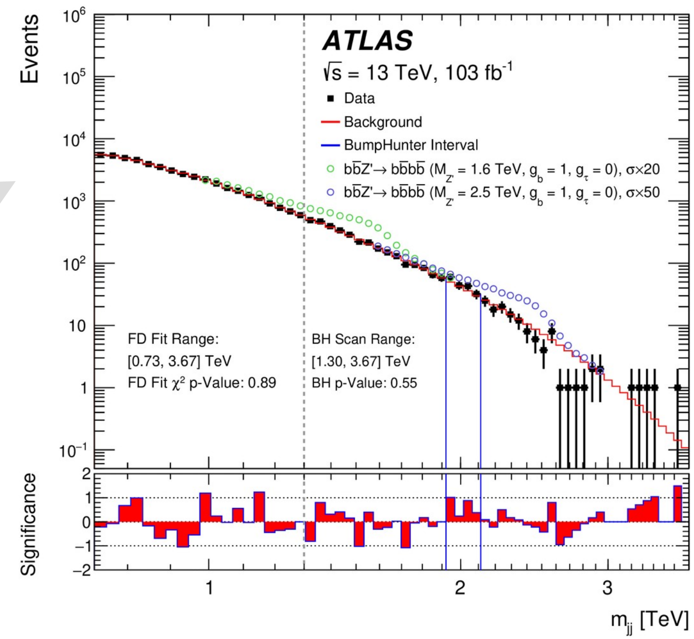




2-b jet mass in all events



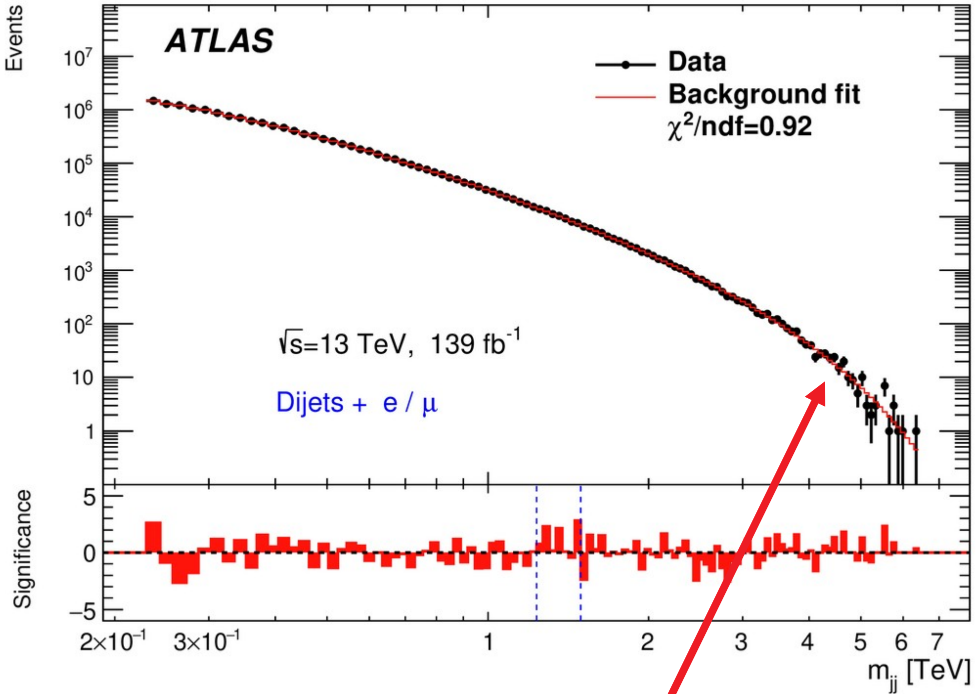
2-b jet mass in events with multiple b-quarks

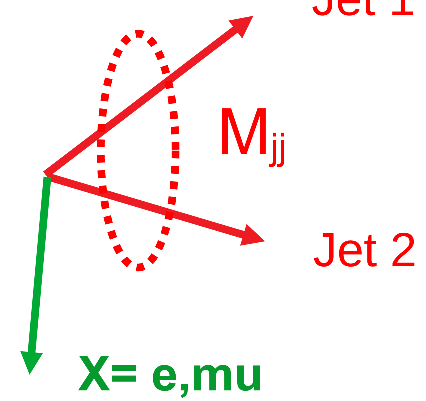


- No signals. Competitive limits for Z' → $b\bar{b}$ processes

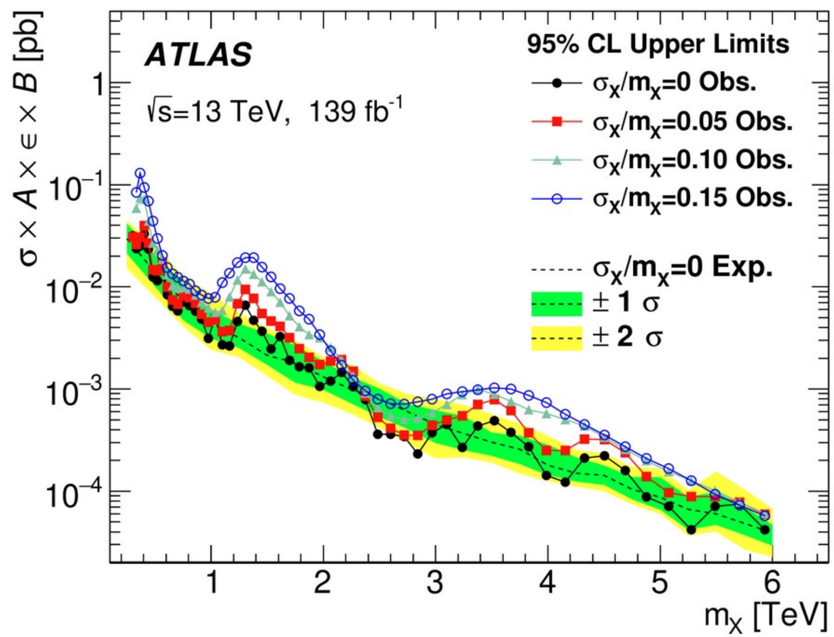
Searches using di-jets + X

- Additional object (lepton, jet, photon) helps triggering events with smaller masses (~ 300 GeV) and opens sensitivity to new BSM models





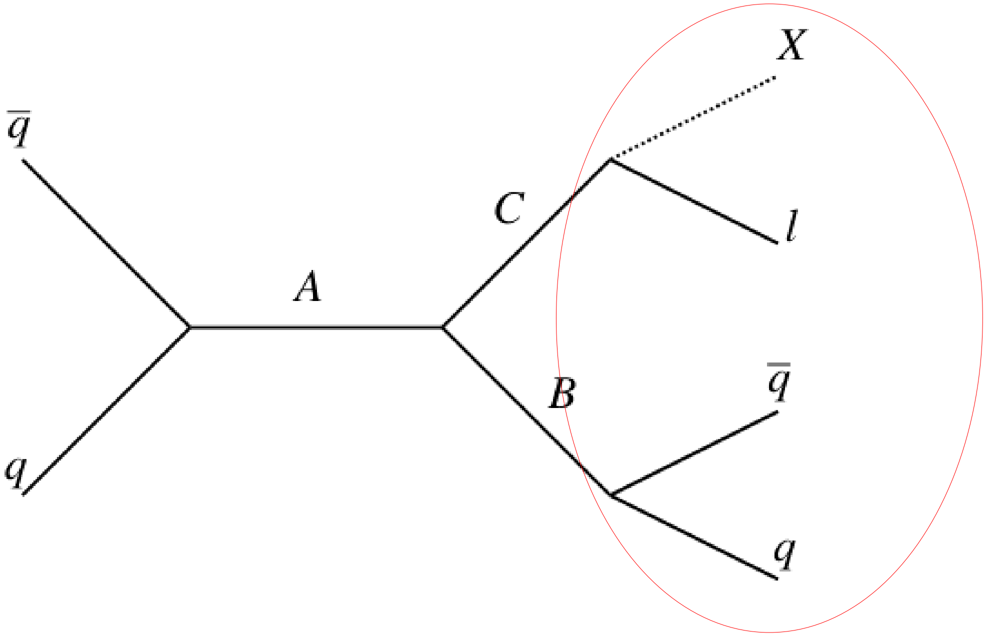
$$f(x) = p_1 (1-x)^{p_2} x^{p_3 + p_4 \ln x + p_5 (\ln x)^2}$$

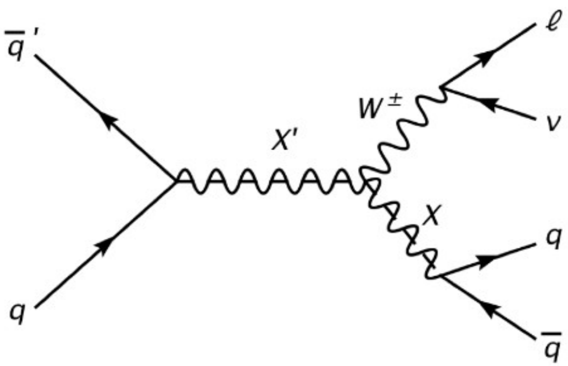


Still works down ~200 GeV !

Largest deviation at 1.2 TeV (~3 sigma)

Searches in multi-body invariant masses



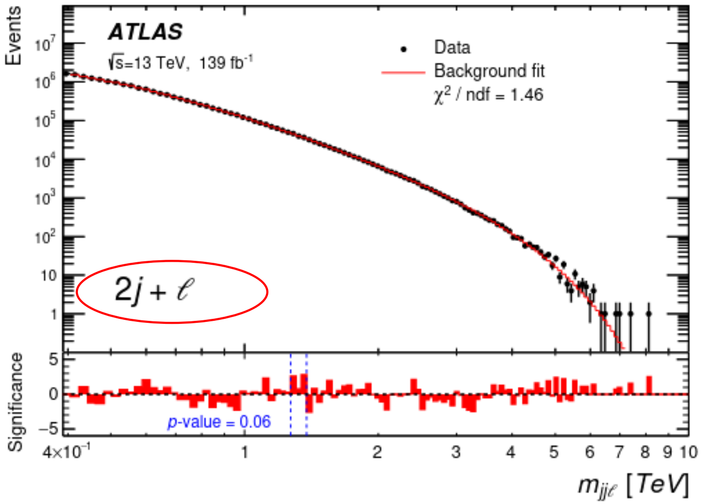


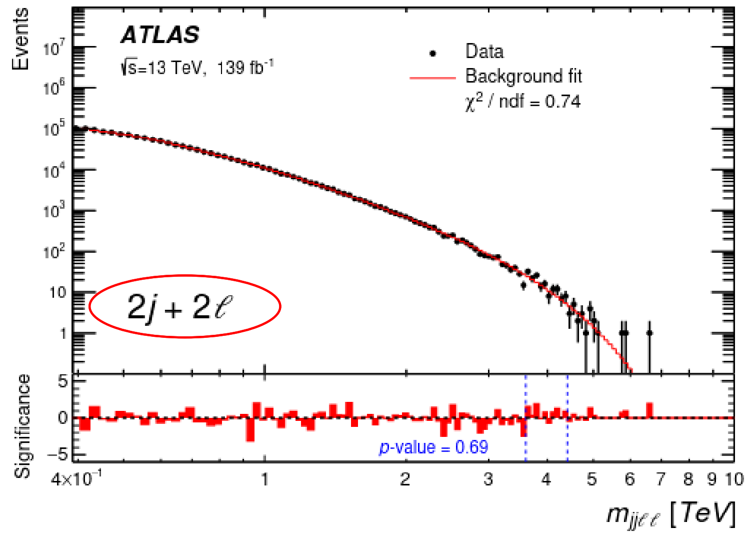
Sequential standard model example where mass splitting between X' and X are small

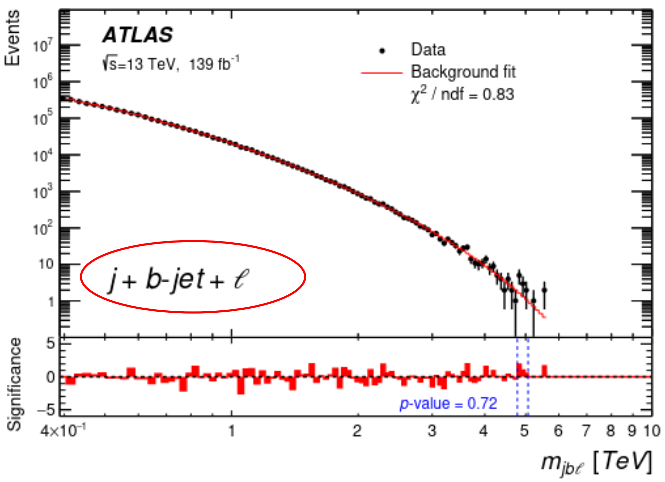
- Instead of focusing on 2-body decays, construct invariant masses of 3 and 4-objects (jets, b-jets, leptons)
- Covers scenarios where B$\rightarrow$ 2 jets are hard to detect (broad, large QCD background)
- Proposed in *S.C, S.Darmora, W.Islam, C. E. M. Wagner, J.Zhang (Universe 2021, 7, 333)*

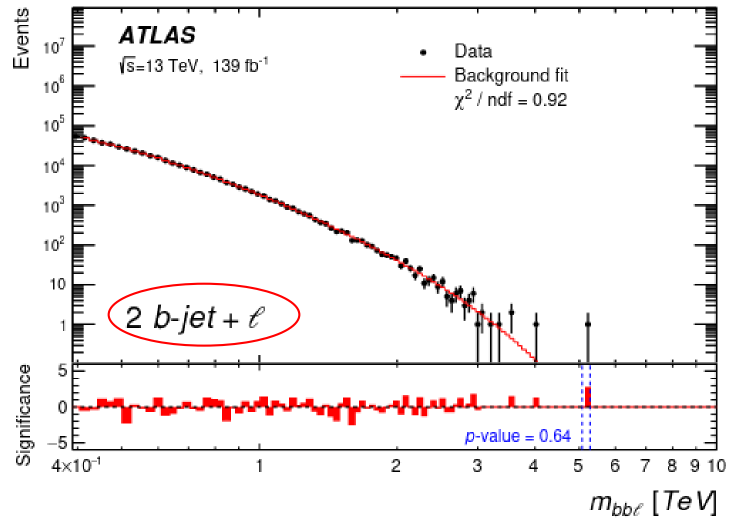
Searches in multi-body invariant masses

ANL contribution









- 4-types of multi-body masses studied: The p5 function works in all cases
- Excluded models: sequential SM, radion, composite lepton, dark matter models

What we have learned so far

- No sign of new physics using the most traditional method – invariant masses
- Background description using analytical functions becomes increasing complex as luminosity increases
- Requiring additional objects (leptons, b-jets) increase sensitivity to different BSM models and reduce the SM background by a factor of 10–100 (depending on the model).
- Given the limited set of available BSM models, can studies of this kind truly rule out the presence of new physics?

There is a growing recognition that the LHC data is far more complex than initially anticipated

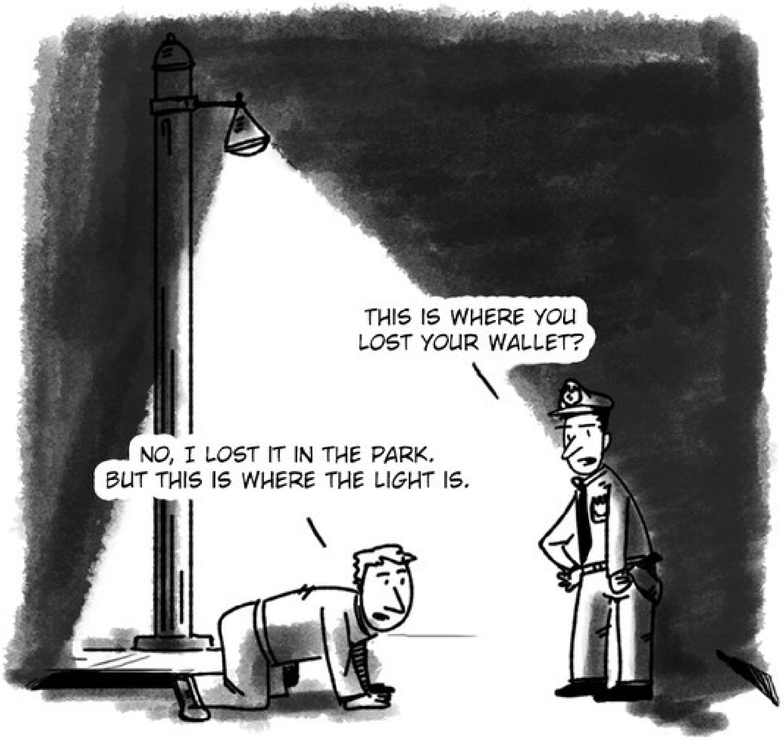
Why haven't particle physicists found any new physics ?

(extracted from the Reddit question in r/ParticlePhysics, 55 participants)

- New physics may be difficult to recognize, because the Standard Model, in many respects, is still a model, or a framework with numerous adjustable parameters — parameters that can only be determined empirically.

- We have explored a tiny fraction of data, and we cannot yet claim that we fully understand its aspects.

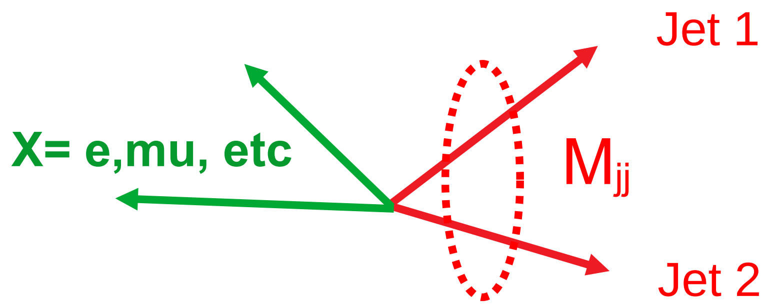
  SM measurements and targeted searches for specific BSM models do not provide a comprehensive framework to capture the full richness of the experimental data.



Is there any method to give a quantitative assessment of this statement?

From inclusive to exclusive

Lessons from the history of physics:



- After studies of inclusive events (such as dijet invariant mass), the next step is to focus on associated production with additional particles (such as dijets with W, Z,H,leptons).
- The importance of studies of exclusive event classes is well recognized for many BSM studies (such as SUSY, technicolor, sequential SM models, etc.), where the main focus are events with additional particles/jets.
- Background rates are drastically reduced compared to inclusive studies.
- **Two questions:**
  - **How many exclusive event classes should we expect at the LHC?**
  - **How many of such classes have been studied so far?**

Number of possible exclusive event types

S.C, Universe 2024, 10(11), 414

- Use semi-analytic combinatorial analysis with boundary conditions from either experiments or Standard Model simulations
- **Define a single event class as** $(Nm, Nj, Nb, Ne, N\mu, N\tau, N\gamma)$
  - Nm **- either 1 (missing energy>200 GeV) or 0 (no missing energy)**
  - Nj **- number of light-flavor jets,** Nb **- number of *b*-jets**
  - Ne, $N\mu$, $N\tau$, $N\gamma$ **- numbers of electrons, muons, taus, photons**
- **Now many unique event classes should we expect?**

$$\sum_{r=2}^{Nmax} \frac{(n+r-1)!}{(n-1)!r!}$$

For Nmax=20 (max number of objects) and n = 7 objects per event, the calculated number of exclusive event classes is 888,022 $\rightarrow$ Not realistic

- Apply **boundary conditions** to exclude events with unrealistic multiplicities (like 7 muons and electrons in the same event)
- Can be done numerically using SM Monte Carlo simulations

Boundary conditions

S.C, Universe 2024, 10(11), 414

From current LHC restrictions (from 36 fb^{-1} publications)

$$Nm < 2, \quad Nj < 7, \quad Nb < 5, \quad Nj + Nb < 10,$$
$$Ne < 4, \quad N\mu < 4, \quad N\tau < 4, \quad N\gamma < 4,$$
$$N\ell < 5, \quad N\ell + N\gamma < 5,$$
$$Nj + Nb + N\ell + N\gamma < 13.$$

$$N\ell = Ne + N\mu + N\tau$$

pT> 20 GeV for all objects

Nr of expected exclusive event classes:

6,676 – from numerical combinatorics
2,485 – predicted by PYTHIA8 (all SM)

From 154 fb^{-1} SM simulations (MC truth level)

$$Nm < 2, \quad Nj < 18, \quad Nb < 9, \quad Nj + Nb < 19,$$
$$Ne < 5, \quad N\mu < 5, \quad N\tau < 5, \quad N\gamma < 5,$$
$$Nj + N\ell < 19, \quad Nj + N\gamma < 19,$$
$$Nb + N\ell < 9, \quad Nb + N\gamma < 9,$$
$$N\ell < 6, \quad N\ell + N\gamma < 6,$$
$$Nj + Nb + N\ell + N\gamma < 21,$$

Nr of expected exclusive event classes:

19,497 – from numerical combinatorics
3,537 – predicted by PYTHIA8 (all SM)

PYTHIA (~ Standard Model) with all SM process failed to generate about 60% - 80% of event topologies

SM expectations vs "agnostic" reality

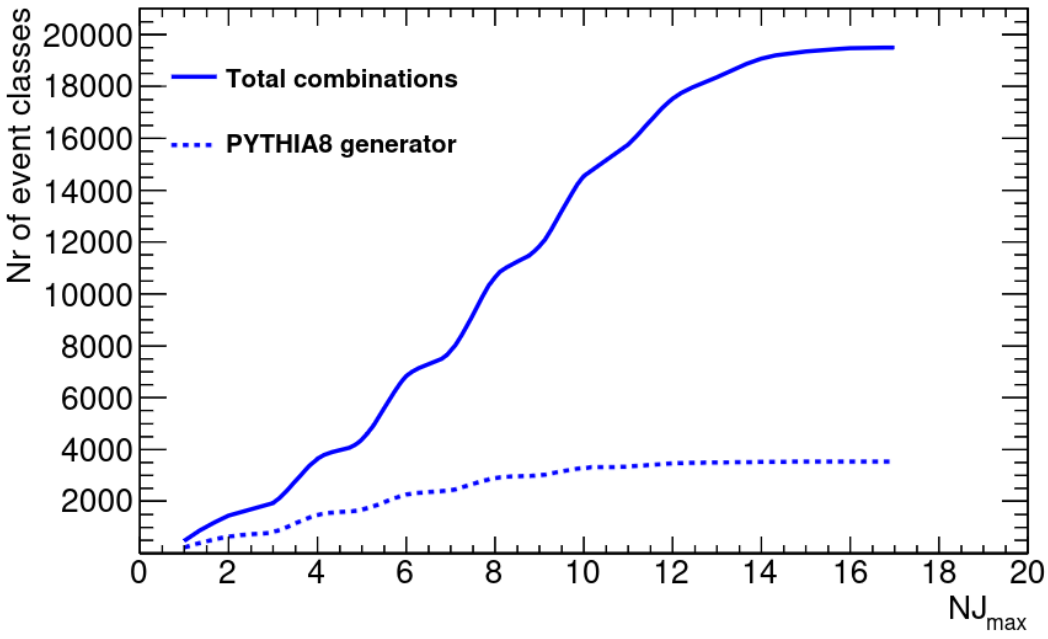
- PYTHIA8 (all SM events enabled) significantly under-predict the number of event classes (compare to what we potentially may expect)
- Simulations never contain event classes* such as:

$$(0m,\ 2j,\ 2b,\ 2e,\ 3\mu,\ 0\tau,\ 0\gamma),$$
$$(1m,\ 4j,\ 0b,\ 3e,\ 1\mu,\ 1\tau,\ 1\gamma),$$

* *pp collision, limited phase space,no lepton flavor conservation is expected*

The number of expected event classes from the combinatorial analysis and PYTHIA8 as a function of the maximum number of jets (NJ_{max}). It is assumed that the number of b−jet should be less than 50% of the total number of jets.

Could be NLO effects? Or some rare processes not in PYTHIA8?



How many 2-particle masses are expected?

S.C, Universe 2024, 10(11), 414

- Each event class leads to multiple invariant masses. Example:

$$(1j,\ 1b,\ 1e)$$

$$\downarrow$$

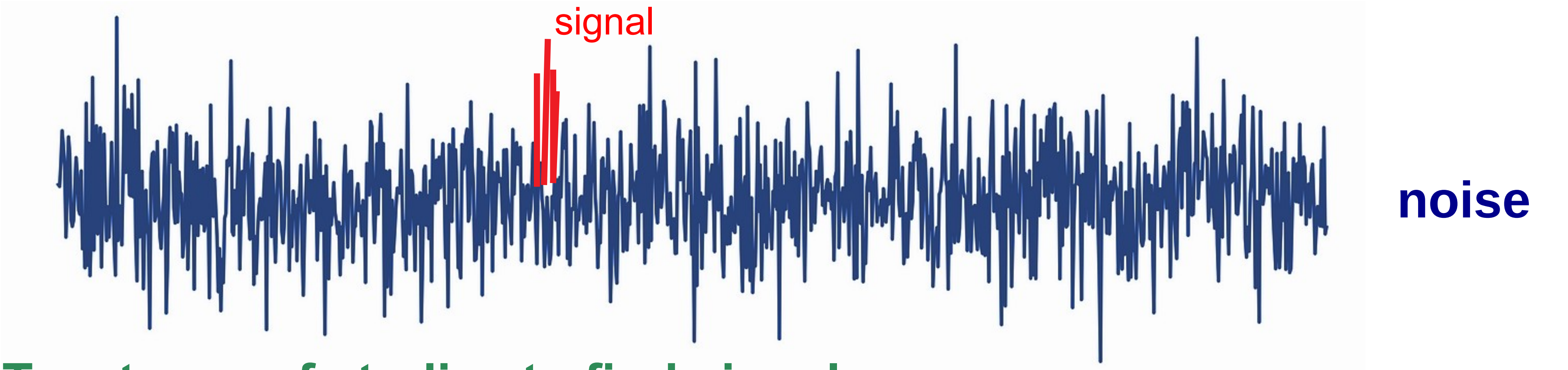
$$(0j,\ 1j1b,\ 1e),\quad (1j,\ 0b,\ 1b1e),\quad (1j1e,\ 1b,\ 0e)$$

- For the most conservative boundary condition with **6,676** event classes, the number of 2-body invariant masses is ~**53,000**

- ATLAS (Eur. Phys. J. C 79 (2019) 120) studied about 700 event classes using only 3.2 fb^{-1} of data — a tiny fraction of the available events. The total number of event classes investigated so far does not exceed ~**1000**.

The LHC currently cannot access this richness of final states using traditional methods. When long-lived signatures are included, the coverage of possible enhancements in invariant masses at the LHC becomes even smaller.

Standard model as the "noise" to filter out



Two types of studies to find signals:

➔(1) Use the knowledge of the "signal" (BSM models)

→ **Traditional LHC searches that use BSM Monte Carlo models**

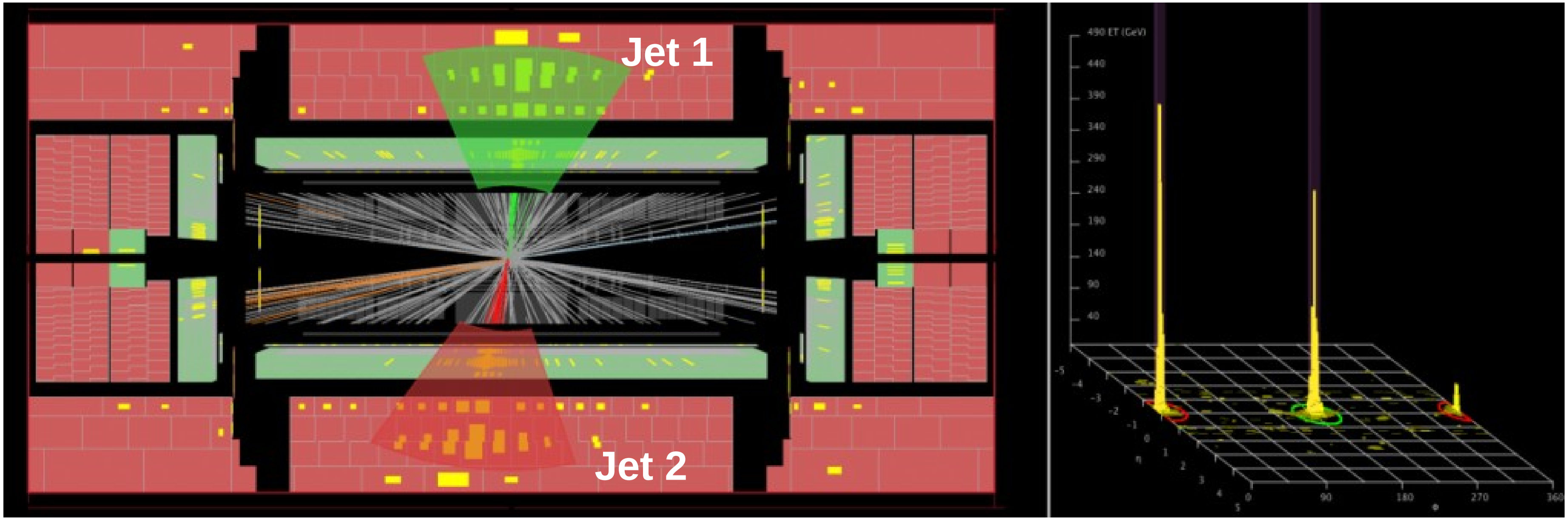
➔ (2) Use the knowledge of the "noise" (aka Standard Model)

→ **Larger coverage of unexpected signatures**

Focus on rejecting known Standard Model events while analyzing 'anomalous' events which may not be covered by specific BSM.

Create "Filter" to remove trivial SM kinetic signatures and look at the rest of events

Non-ML method for model-agnostic BSM searches

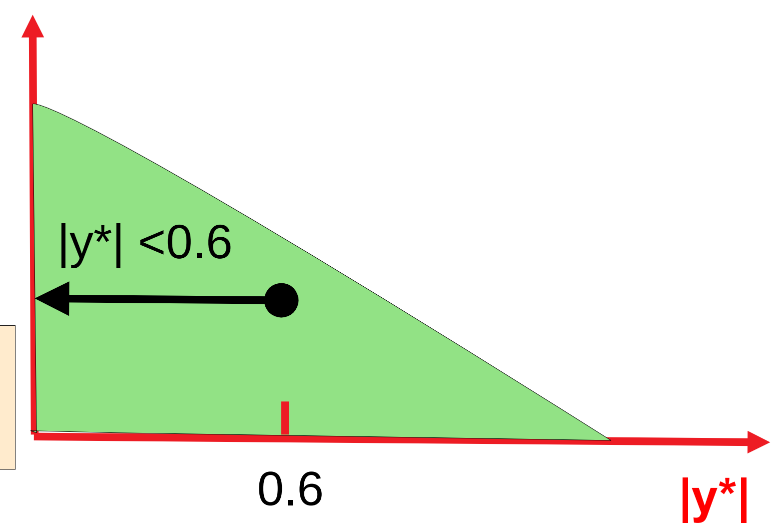


How to enhance sensitivity to heavy new particles using 2-jet production?

Select 2 jets in the central rapidity region:

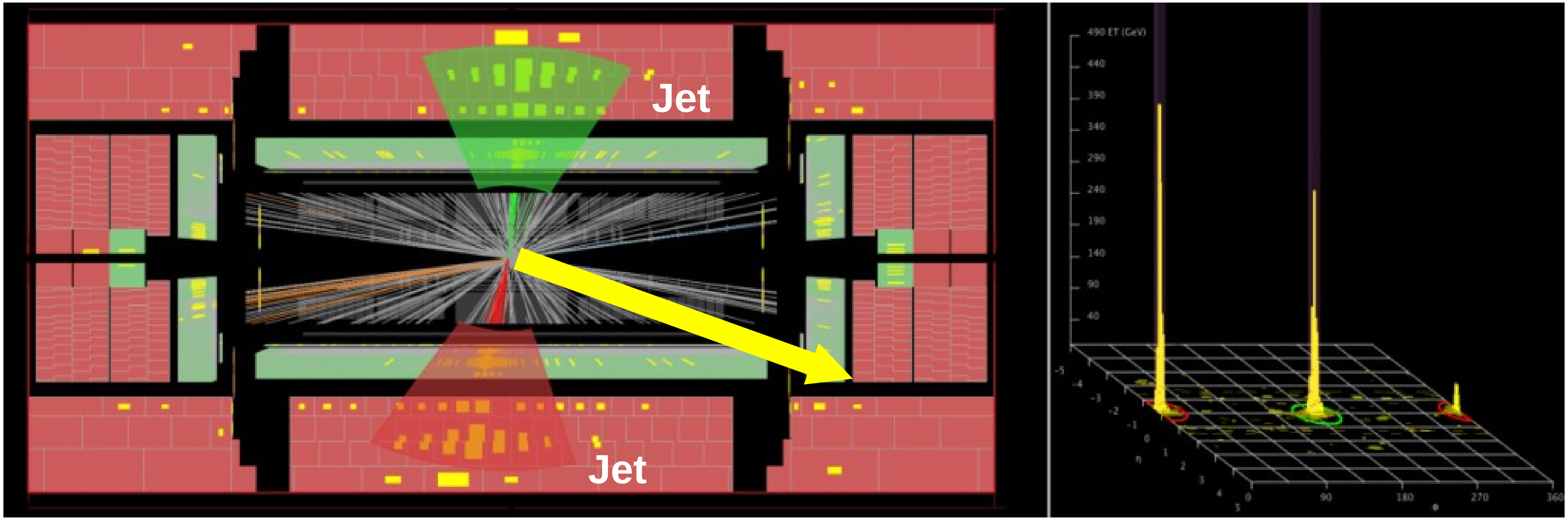
$|y^*| \equiv 0.5 * |\eta\ (\text{jet1}) - \eta\ (\text{jet2})|$ < 0.6

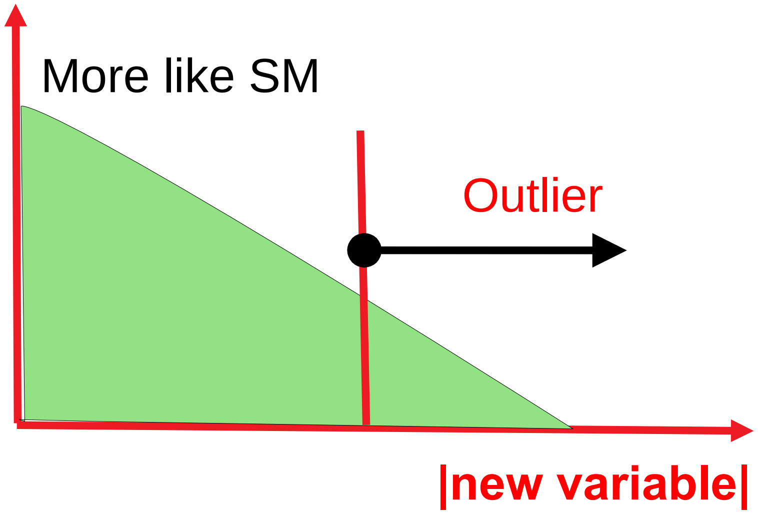
Increases sensitivity by ~40% for heavy BSM particles



This improvement disappears if events have extra objects (additional high-pT leptons, jets, etc)

What if there are additional high-pT objects?





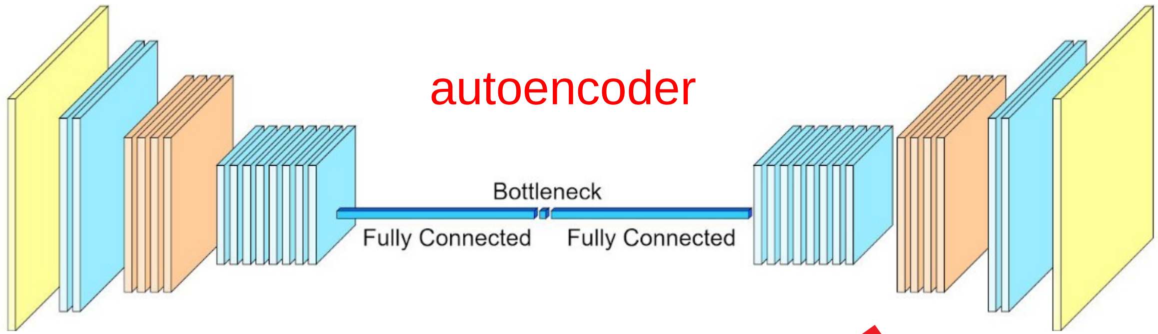
Design a new variable such that the presence of additional features (e.g., an extra particle, jet, or b-jet, or some correlation) that stand out from the "typical" or "bulk" Standard Model events leads to an increased value of this variable.

Larger values (the "outlier" or "anomaly" region) correspond to a higher degree of "strangeness" relative to typical Standard Model events.

Event-based anomaly detection

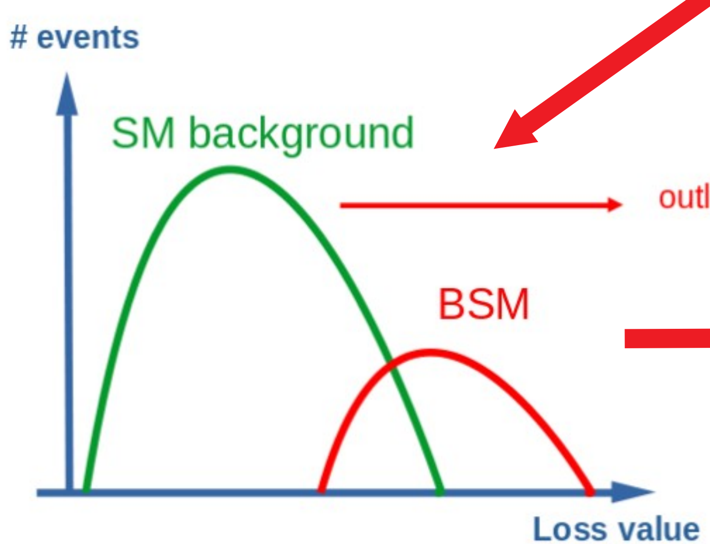
S.C. NIMA 931 (2019) 92-99
S.C, W. Hopkins, Universe 2022, 8(10), 494

Input LHC data



Replica of LHC data

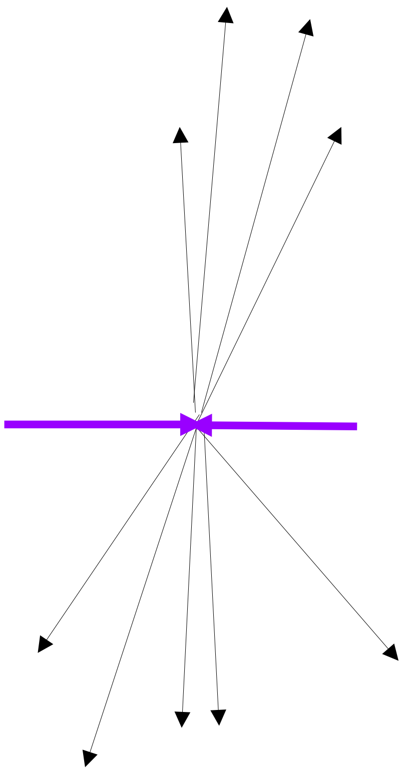
“Bad replica”



Reconstruct invariant masses in the “outlier” region

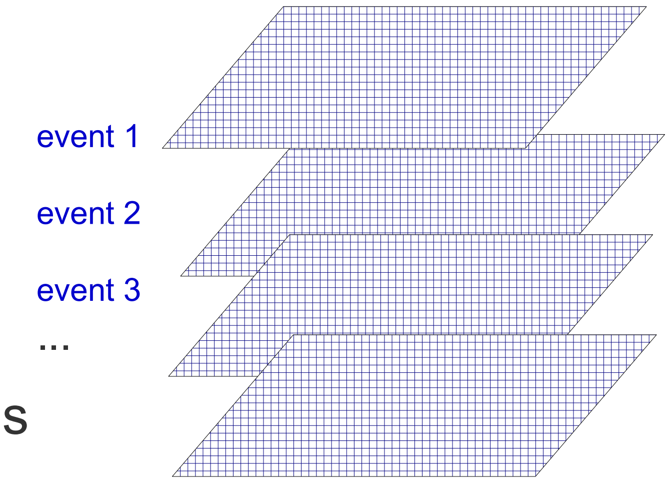
Inputs: “Imaging kinematics” of particle collision

(px,py,pz,E)



from variable-size list with particles → fixed-size matrices

- Fixed size (zero padding)
- Dimensionless
- Lorentz invariant
- Fixed range of values
- Single and 2-particle densities
- Small correlations between variables
- Similarity with images



Organizes variable-size list in compact fixed-size data structures.
Convenient input to ML & easy to visualize (similar to images)

Rapidity-mass matrix (RMM)

https://arxiv.org/abs/1805.11650
S.C. NIMA, A931 (2019) 92

Missing momentum and transverse masses

Invariant masses of pairs m(i,j)/E_{CM}

$$
\begin{pmatrix}
\mathbf{e_T^{miss}} & m_T(j_1) & m_T(j_2) & \ldots m_T(j_N) & m_T(\mu_1) & m_T(\mu_2) & \ldots m_T(\mu_N) \\
h_L(j_1) & \mathbf{e_T(j_1)} & m(j_1,j_2) & \ldots m(j_1,j_N) & m(j_1,\mu_1) & m(j_1,\mu_2) & \ldots m(j_1,\mu_N) \\
h_L(j_2) & h(j_1,j_2) & \delta\mathbf{e_T(j_2)} & \ldots m(j_2,j_N) & m(j_2,\mu_1) & m(j_2,\mu_2) & \ldots m(j_2,\mu_N) \\
\ldots & \ldots & \ldots & \ldots, & \ldots & \ldots & \ldots \\
h_L(j_N) & h(j_1,j_N) & \ldots & \ldots \delta\mathbf{e_T(j_N)} & m(j_N,\mu_1) & m(j_N,\mu_2) & \ldots m(j_N,\mu_N) \\
h_L(\mu_1) & h(\mu_1,j_1) & h(\mu_1,j_2) & \ldots h(\mu_1,j_N) & \mathbf{e_T(\mu_1)} & m(\mu_1,\mu_2) & m(\mu_1,\mu_N) \\
h_L(\mu_2) & h(\mu_2,j_1) & h(\mu_1,j_2) & \ldots h(\mu_2,j_N) & h(\mu_1,\mu_2) & \delta\mathbf{e_T(\mu_2)} & m(\mu_2,\mu_N) \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \\
h_L(\mu_N) & h(\mu_N,j_1) & h(\mu_N,j_2) & \ldots h(\mu_N,j_N) & h(\mu_N,\mu_1) & h(\mu_N,\mu_2) & \delta\mathbf{e_T(\mu_N)}
\end{pmatrix}
$$

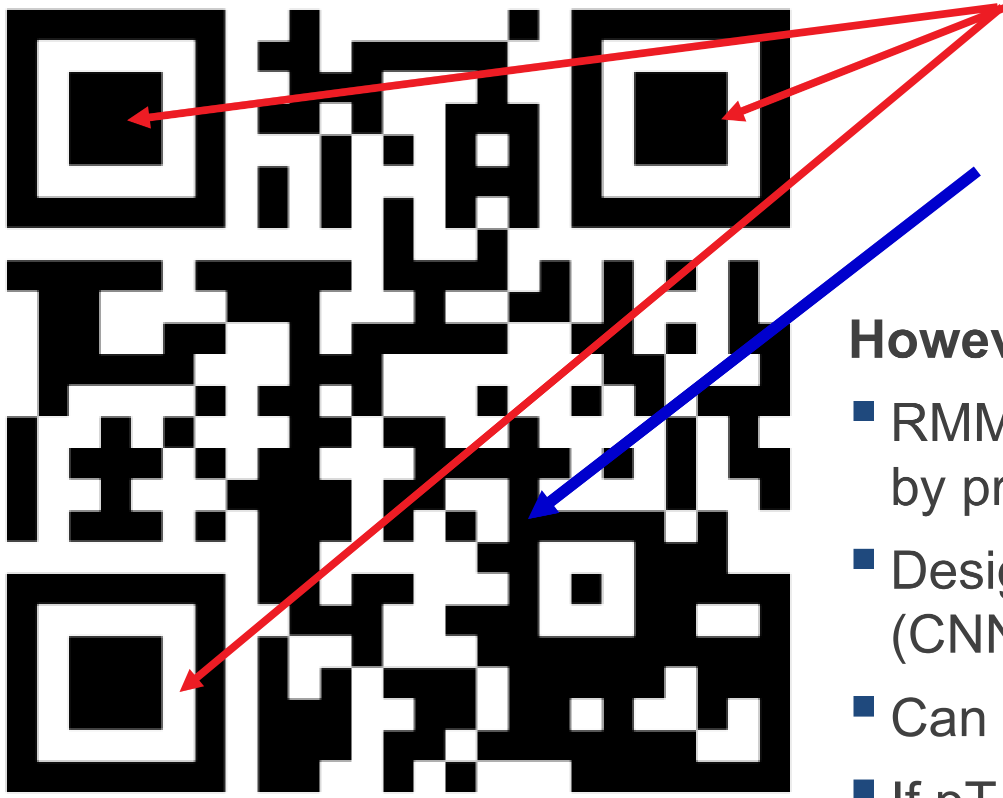
E_T / E_{CM} scaled transverse energies and energy imbalances

Lorentz factors

Rapidity difference: h(i,j) ~ cosh (y_j - y_i)

- Dimensionless, Lorentz invariant
- Single and two-particle densities for each identified particle or jet
- Cell values are ~ independent for SM processes → "almost" decorrelation by construction
- Re-scaling and normalization by construction
- Fixed sizes, well-defined mapping to input nodes
- Cells connected by proximity → good for visualization

Similarity with QR code



"Position markers:" MET and 1-particle densities (M_{Ti}, η_i)

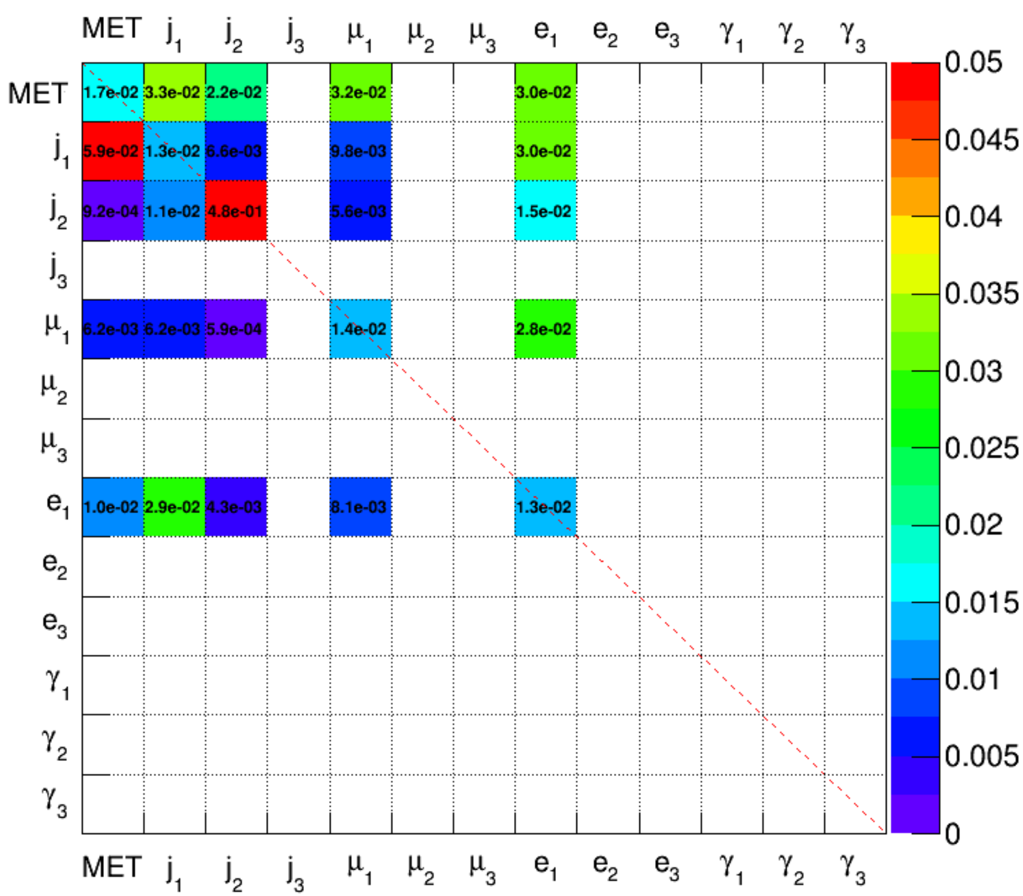
"Data module": 2-particle densities Boost Lorentz Invariant variables (M_{ij} and y_{ij})

However:

- RMM are structural data with cells connected by proximity ("images", not "noise")
- Designed for convolution neural networks (CNN) for image classification
- Can be studied/debugging by humans
- If pT of objects are large:
  - → fewer objects → less "image"
  - Can be treated as sparse matrices
  - Use standard matrix manipulation libraries

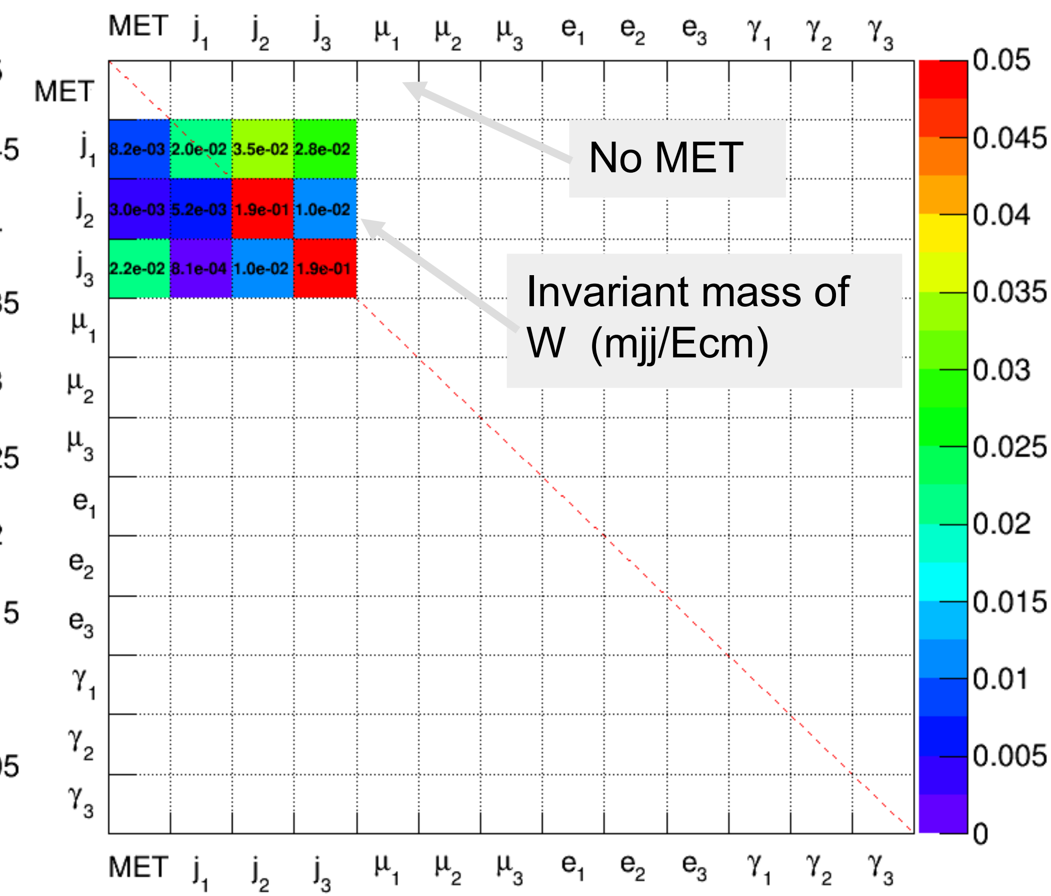
Example: Two PYTHIA8 events with $t\bar{t}$ as RMMs

t t → Wb Wb → e nu b μ nu b



Cell with MET, μ and e leptons activated

t t → Wb Wb → 6 jets



Many jets, no MET and leptons

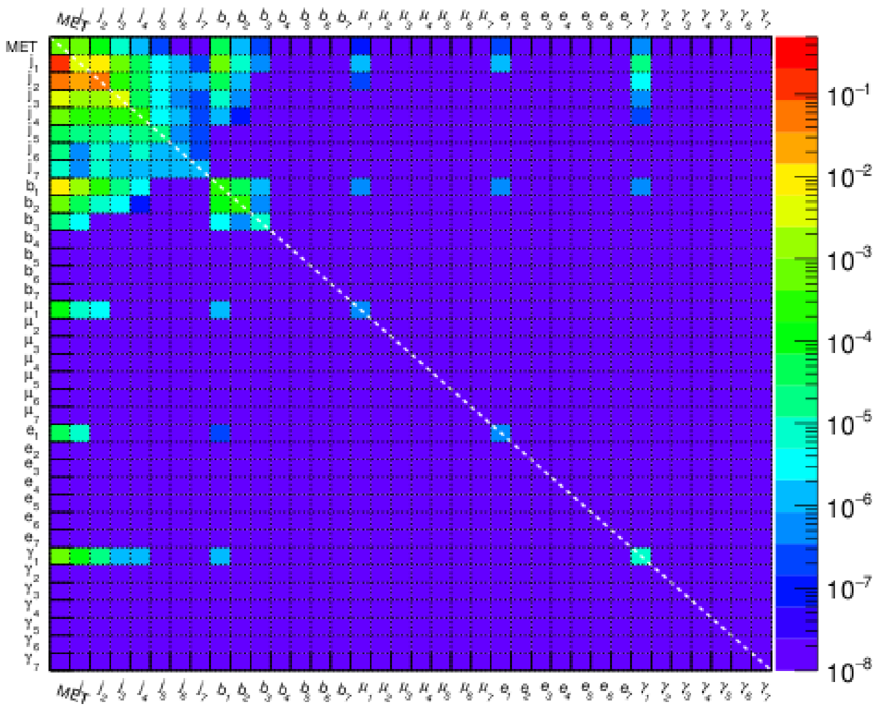
Each cell maps to an input neuron → "natural" language for ML

S.C, arXiv:1810.06669

- Premise of the RMM - generality. Includes single & 2-particle densities
- No need to hand-pick input variables for every event topology/decay
- Good choice for general event classifiers

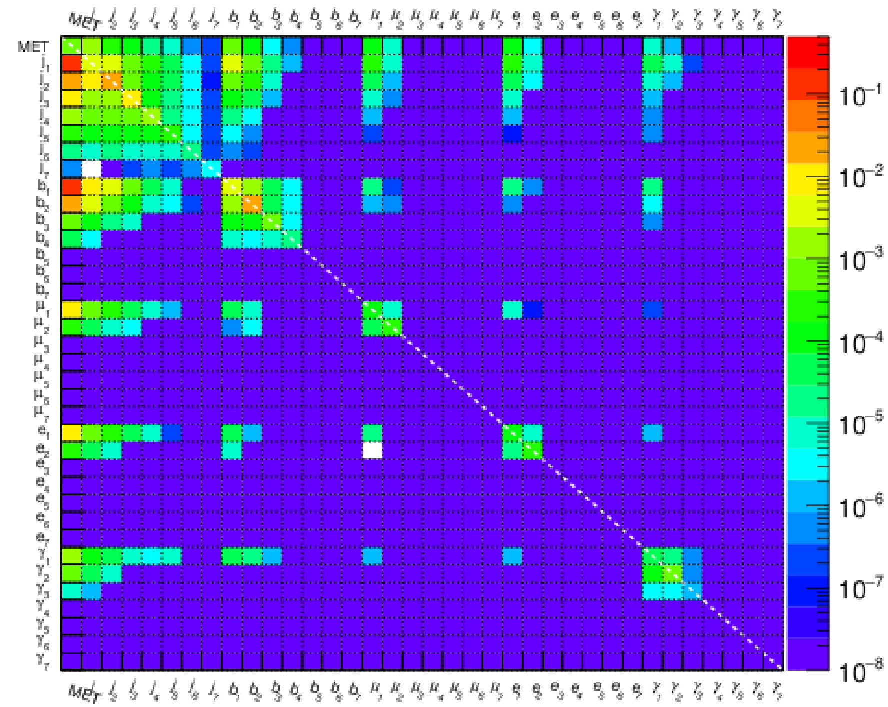
Example for 50k events:

Multi-jet QCD



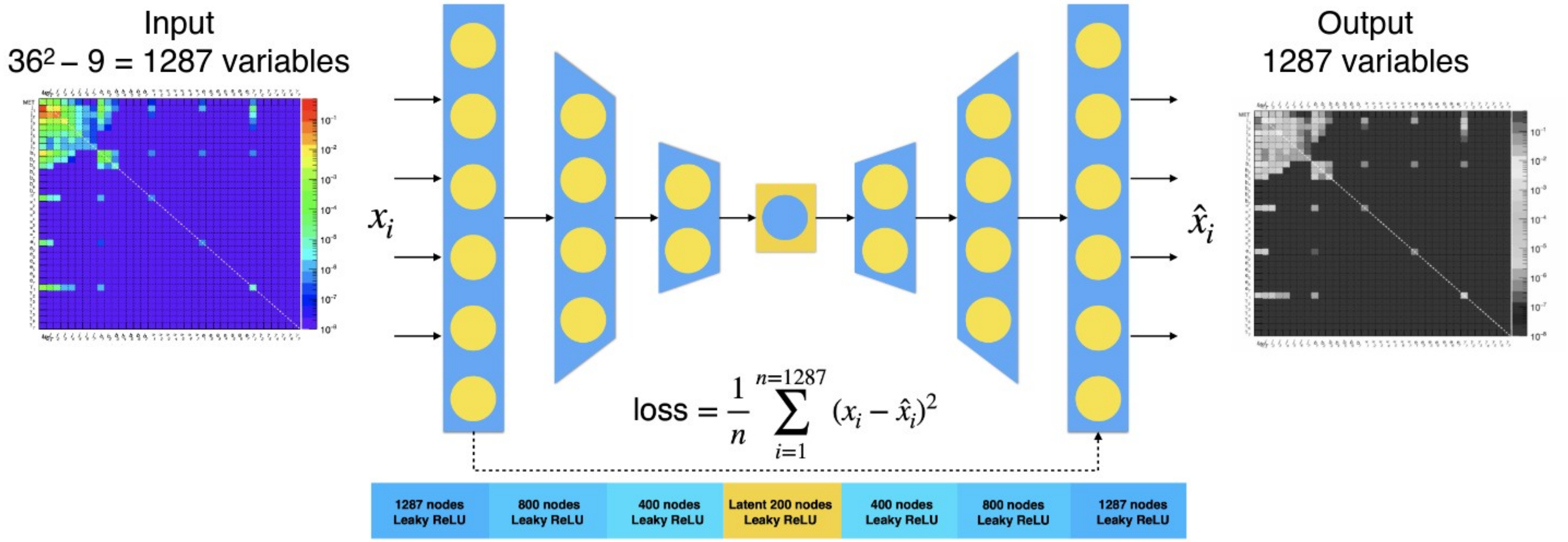
(a)Multi-jets QCD

Higgs productions (all decays)



(b)Higgs processes

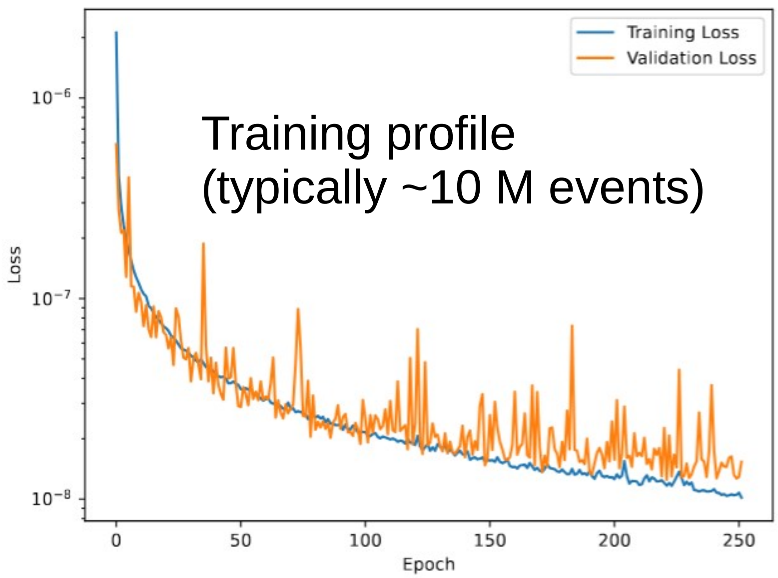
General searches using autoencoder

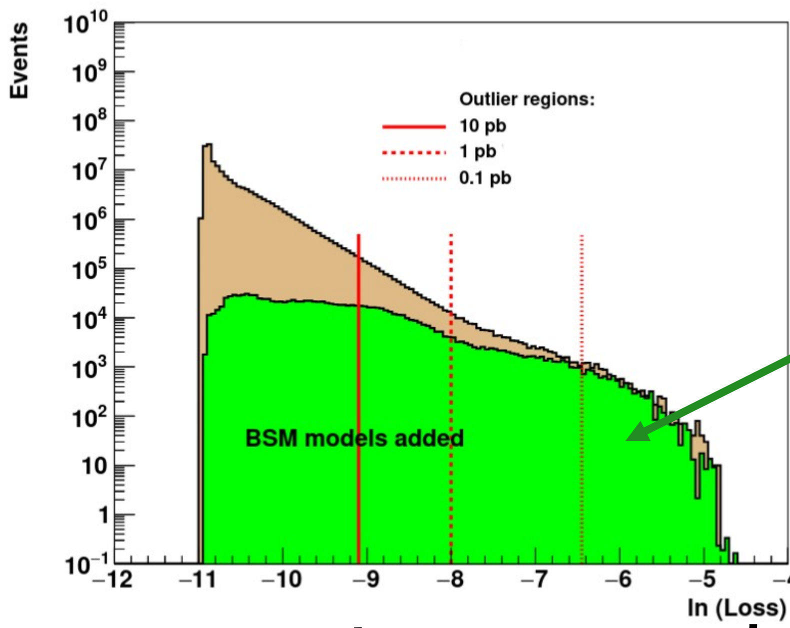


- Train autoencoder using 1% of real data
- Typical RMMs holds up to 10 jets, 10 b-jets and up to 5 leptons /photons (pT>30 GeV)
  - corresponds to 1287 variables (maximum capacity). On average, ~200 input values
- “Memorizes” LHC data using 2 million neuron connections
- Apply to real data. Large error (loss) will indicate “anomalous” events

Finding BSM signatures

ATLAS Collaboration
Phys. Rev. Lett. 132 (2024) 081801
ANL contribution

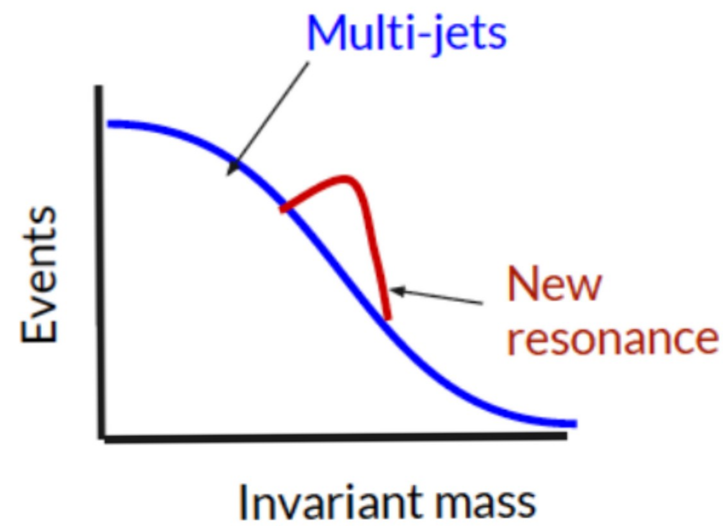




~ 10 different BSM models combined
(masses 0.4-4 TeV)

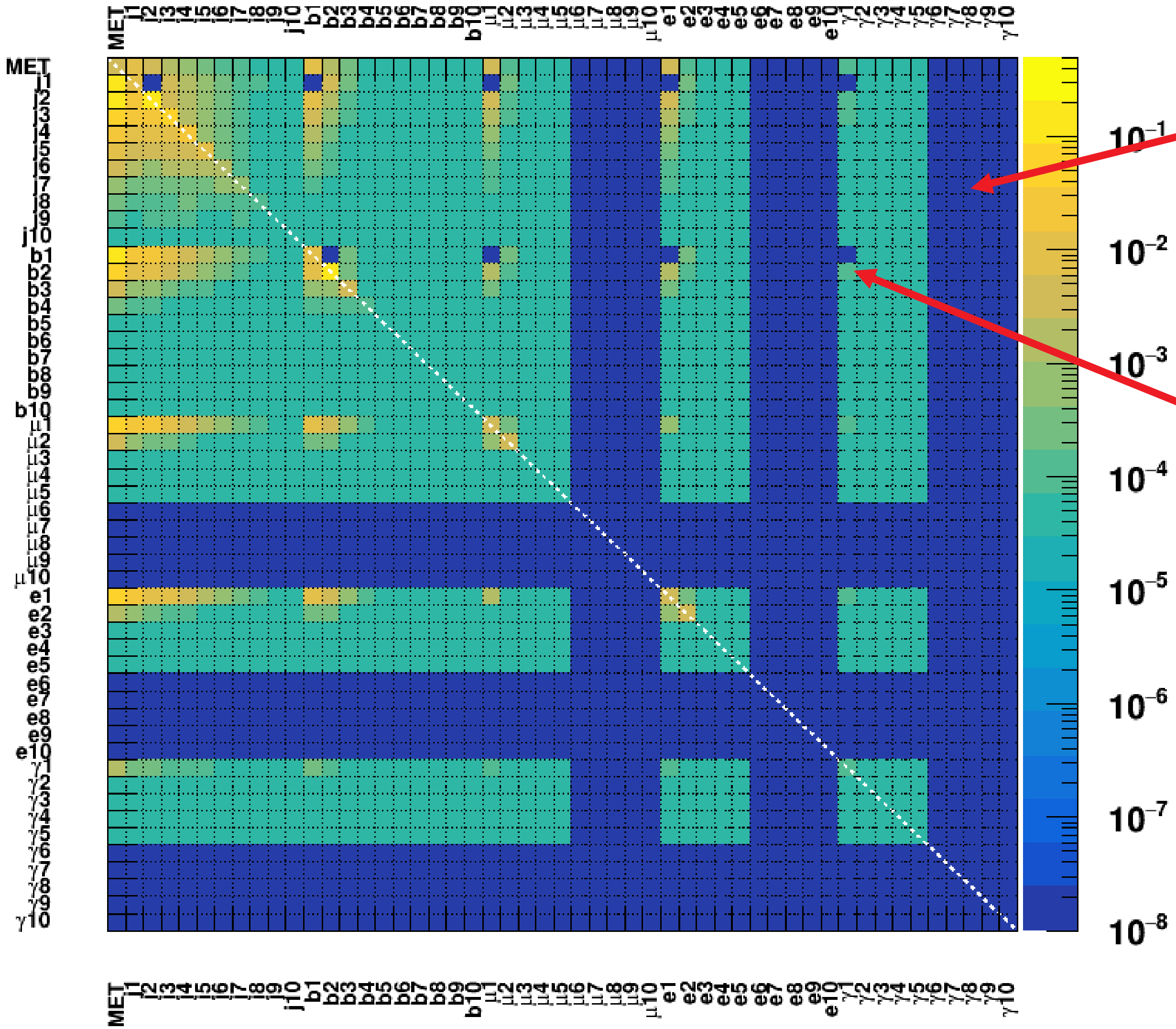
But note: an autoencoder can distort a smoothly falling SM background in an uncontrolled way.

The key idea here is that the RMM enables a factorization of the signal invariant-mass information from the remaining kinematic inputs!



General inputs (average over many events)

Example of RMM for typical SM Monte Carlo events (truth level, ttbar)
- 20k events, selected with at least one lepton above 60 GeV



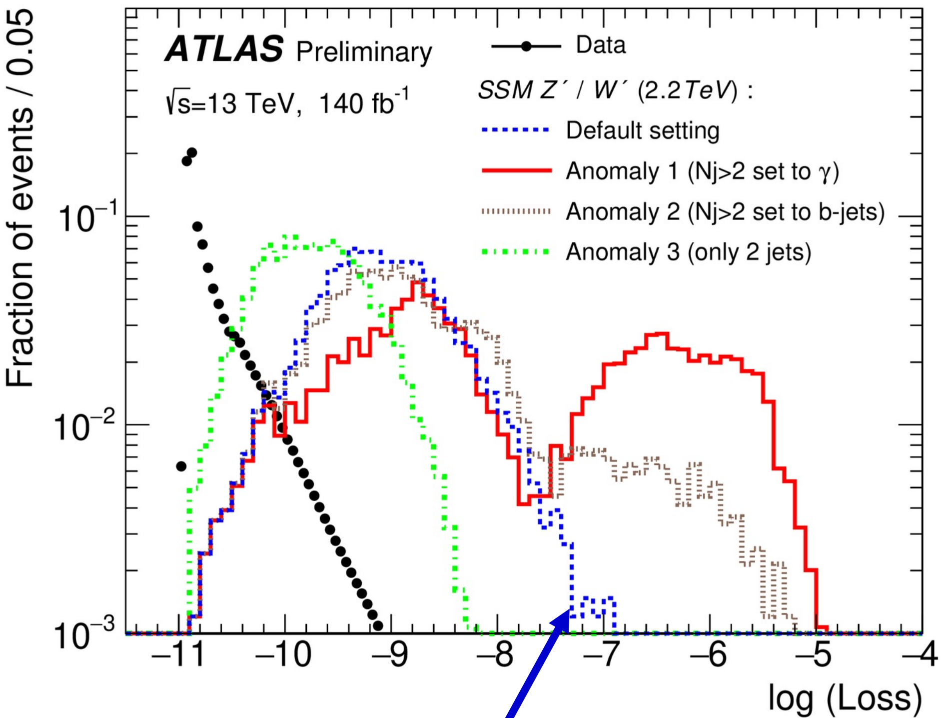
do not consider leptons/photons with N> 5 (simplify input)

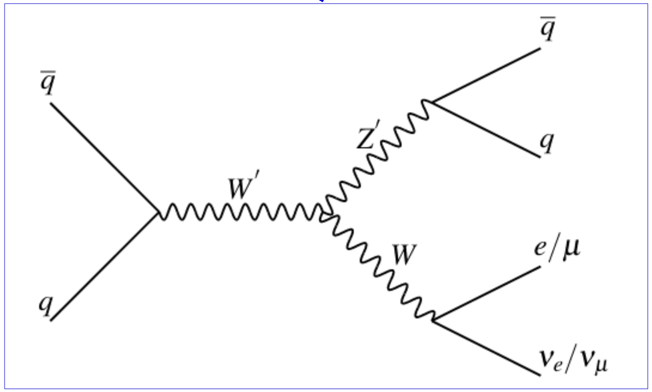
9 invariant masses jet + X – (X=j, b, e, mu, gamma) are signal region

- Blind these cells and do not use in training
- This avoids training biases on invariant masses

What do we expect from autoencoder?

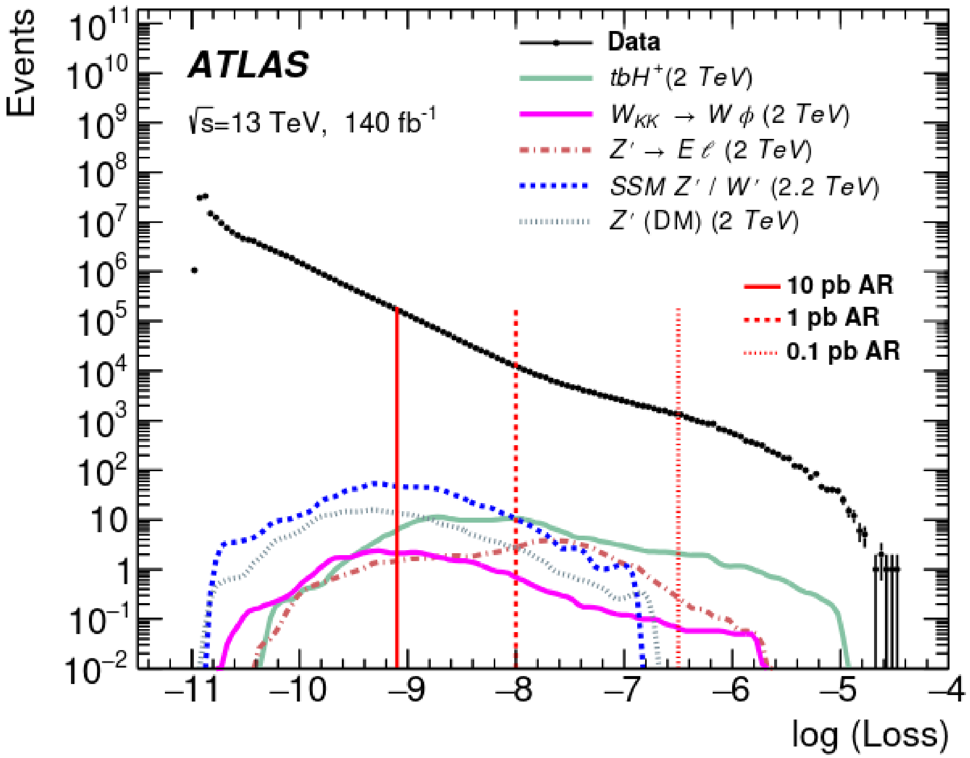
ATLAS Collaboration
Phys. Rev. Lett. 132 (2024) 081801
ANL contribution





- "Anomaly" expected to have:
  - ➔ unusual large multiplicity of all objects
  - ➔ unusual kinematics
  - ➔ sensitive to particle ID
- Tested about 20 architectures of autoencoder using modifications of sequential standard model
- Check separation of events for log (Loss)
- Autoender with "wide" architecture and LeakyRelu is a winner:
  - ➔ Input layer: 1287 neurons
  - ➔ 1st layer: 800 neurons
  - ➔ 2nd layer: 400 neurons
  - ➔ 3rd latent layer: 200 neurons
  - ➔ + repeat for the decoder
- ~ 2 million trained weights (~3 h training)

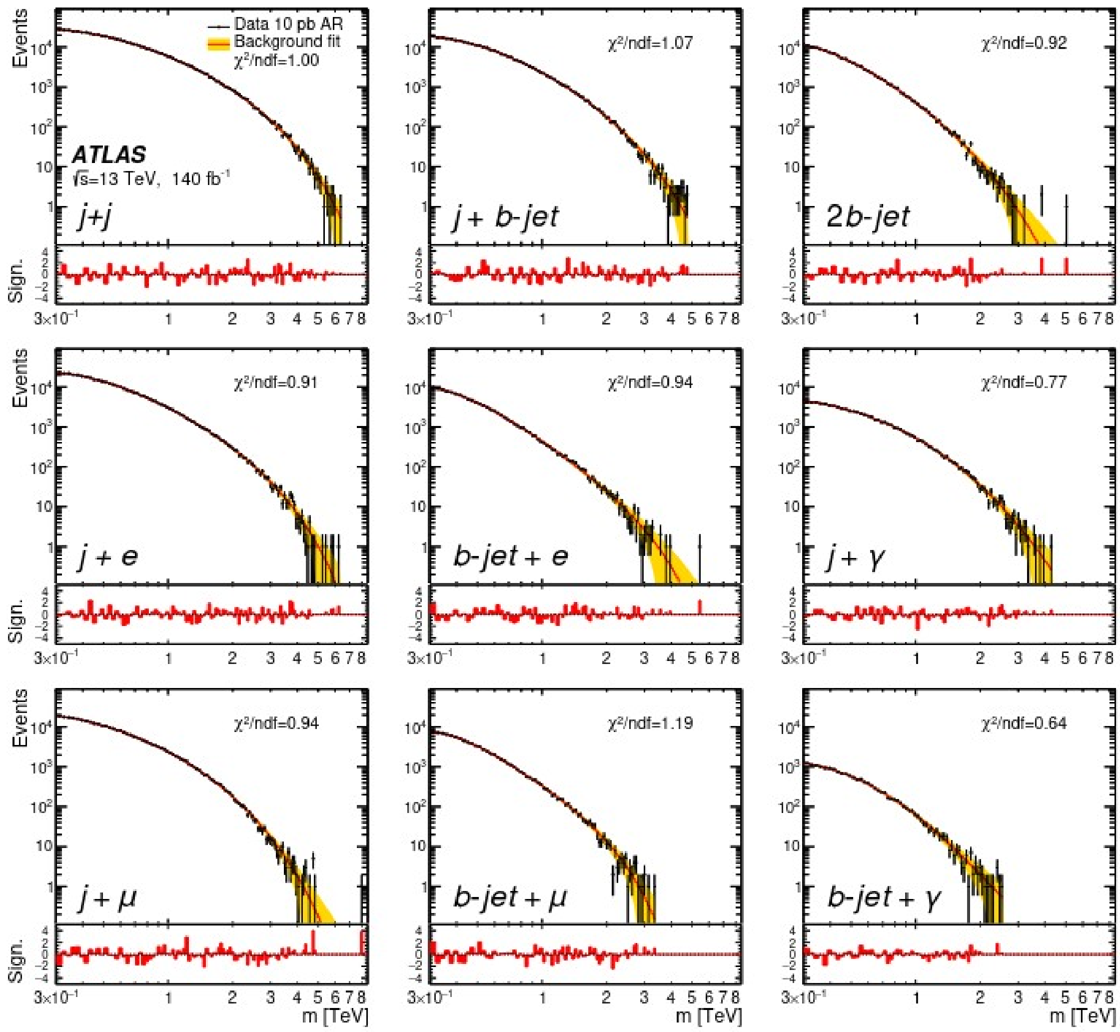
Results of the training



- AE was trained on 1% of data
- Anomaly region (AR) defined in terms of cross sections for BSM events:
  - 10 pb
  - 1 pb
  - 0.1 pb
- Compared with five BSM models with W'/Z'/H+ masses from 0.5 – 6 TeV
- Study invariant masses of jet+X in the anomaly regions using p5 function

$$f(x) = p_1 (1 - x)^{p_2} x^{p_3 + p_4 \ln x + p_5 (\ln x)^2}$$

Results for 10 pb Anomaly Region

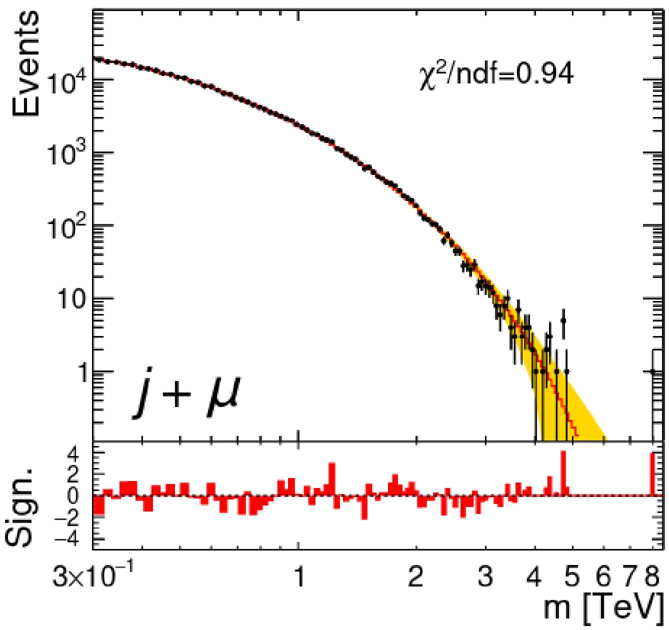


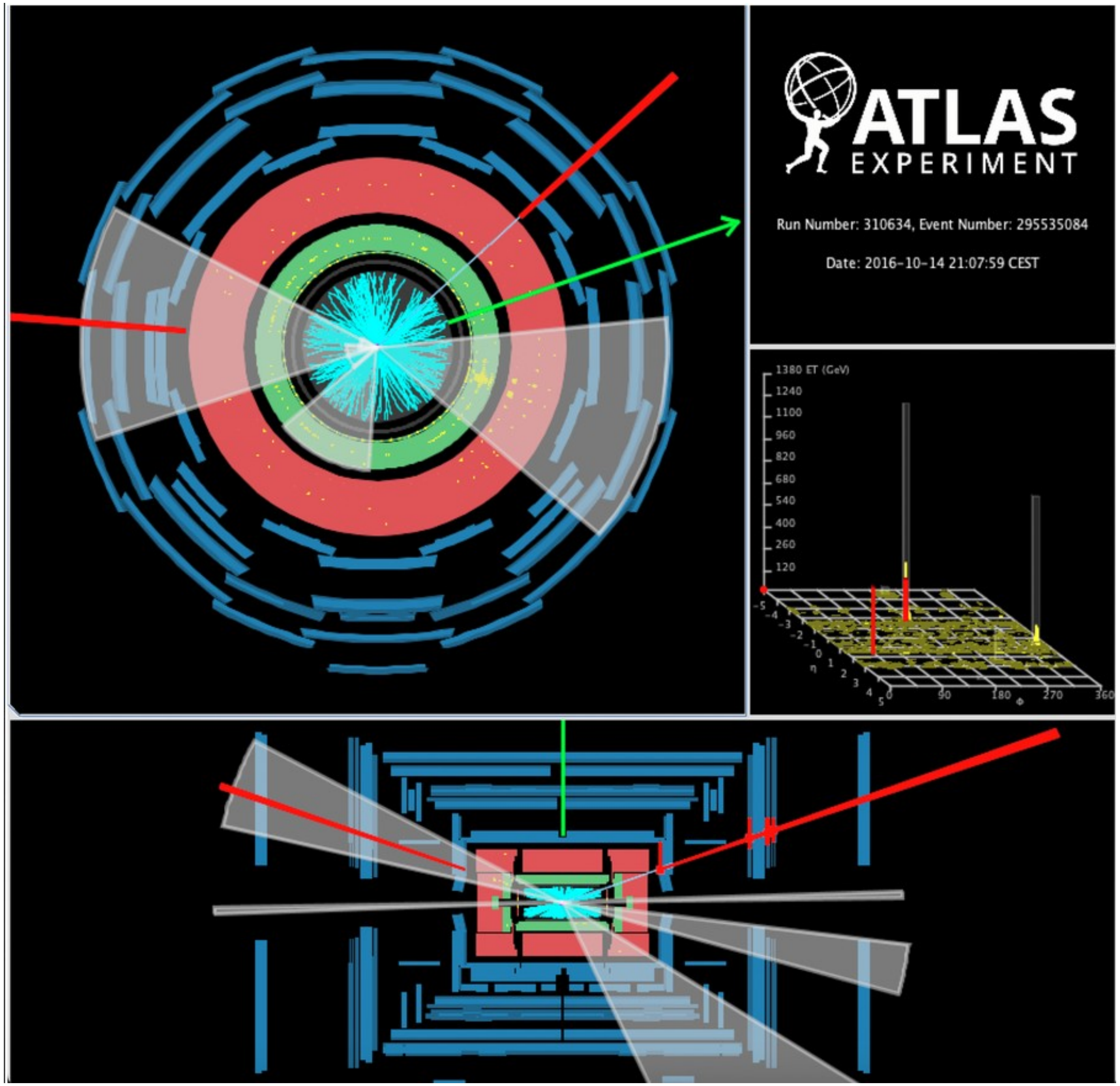
- Good agreement with p5

$$f(x) = p_1\,(1-x)^{p_2}\;x^{p_3+p_4\ln x+p_5(\ln x)^2}$$

- Largest deviation near 4.8 TeV for jet+μ
  - 2.9 sigma (local)

Typical event after AE for 4.8 TeV (largest deviation)





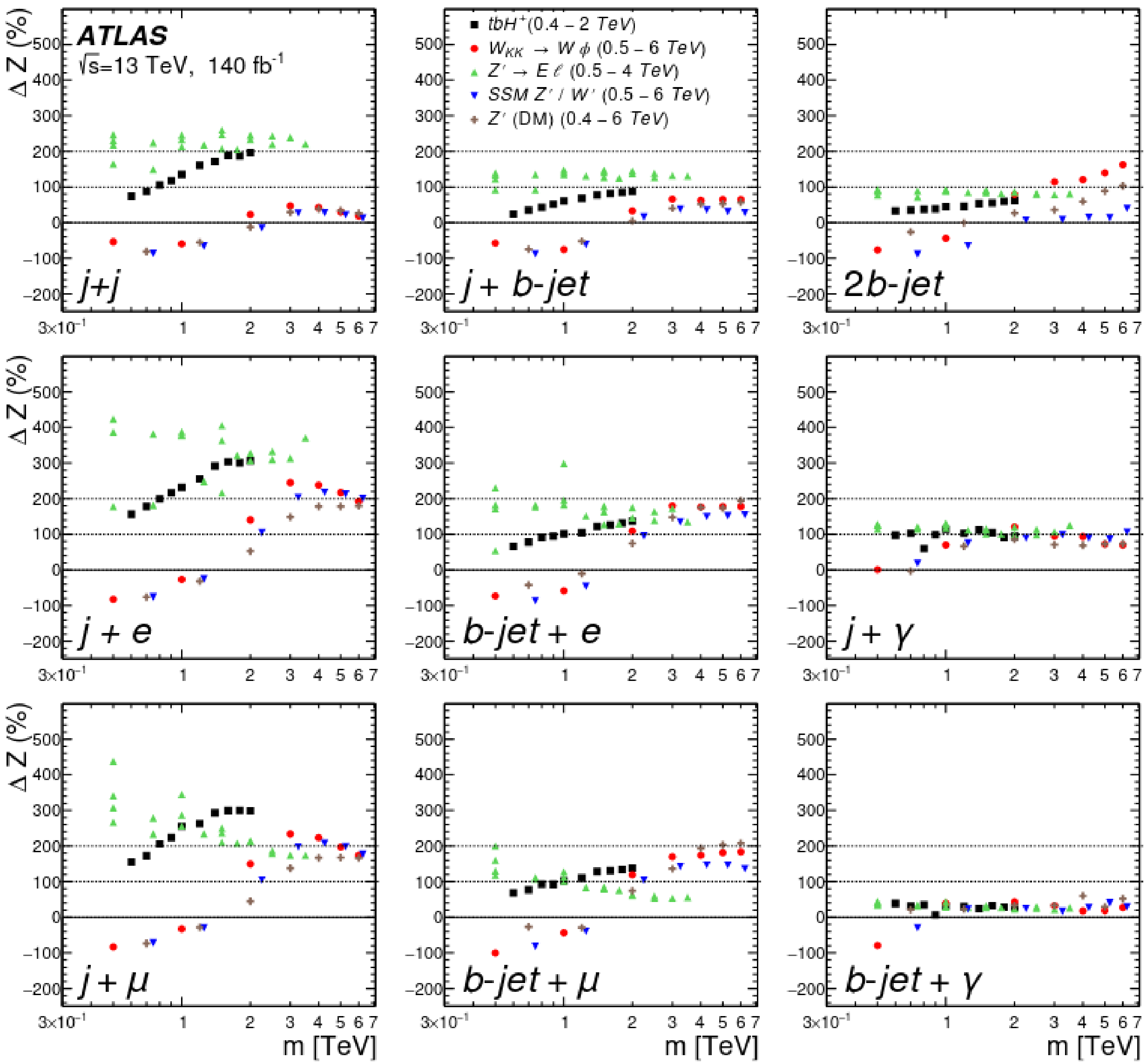
Local significance of this excess is 2.9 σ.

Discovery sensitivity for BSM models

ATLAS Collaboration

Phys. Rev. Lett. 132 (2024) 081801

ANL contribution



- Calculate improvement in discovery sensitivity:

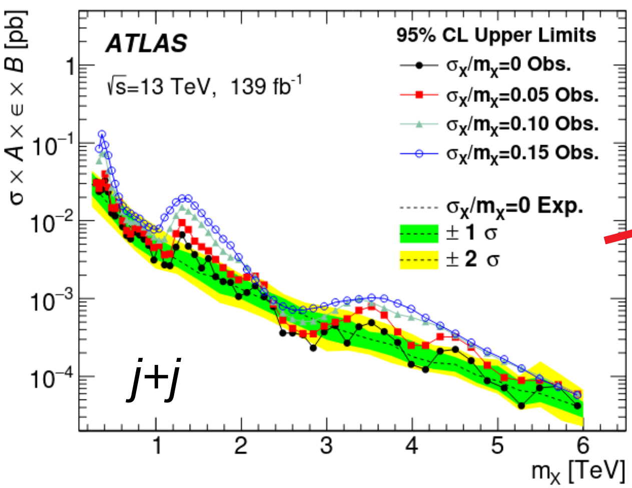
$$\Delta Z = ((Z_{AE}/Z) - 1) \times 100\%$$

where $Z \approx s/\sqrt{b}$

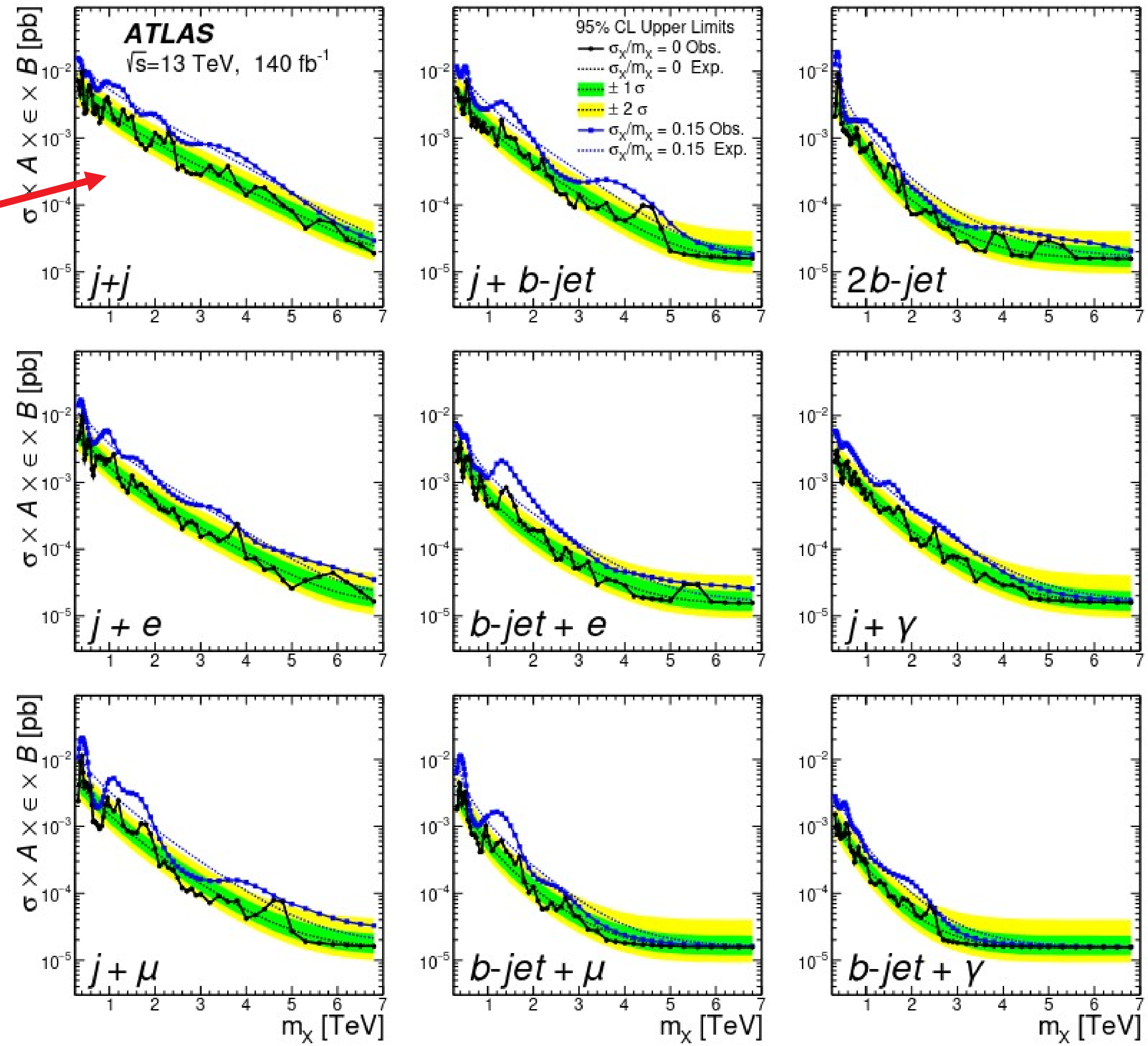
- Autoencoder improved the discovery sensitivity by 300-400% for most BSM
- **Reason:** autoencoder acceptance for BSM models ~80% while for data ~10% (at low mass)
- Directly translates into competitive limits

Gaussian limits in the Anomaly Region

Published: JHEP 06 (2020) 151



10 pb anomaly region after Autoencoder



- Calculate exclusion limits using Gaussian shapes using 0, 5, 10, 15% widths of exotic signal
- At masses < 1 TeV the limits are factor 2-3 better than for similar selection without autoencoder (JHEP 06 (2020) 151)

ATLAS Collaboration

Phys. Rev. Lett. 132 (2024) 081801

ANL contribution

Argonne NATIONAL LABORATORY 80 1946–2026

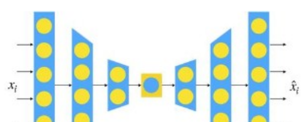
Machine learning could help reveal undiscovered particles within data from the Large Hadron Collider

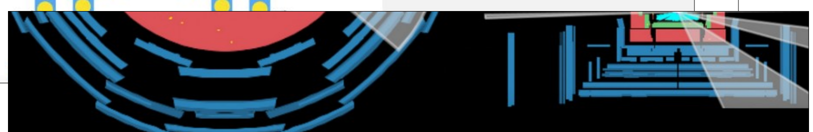
The study marks an innovative use of a neural network to analyze data from a collider experiment

BY SAVANNAH MITCHEM | APRIL 15, 2024

Scientists used a neural network, a type of brain-inspired machine learning algorithm, to sift through large volumes of particle collision data.

For over two decades, the ATLAS particle detector has recorded the highest energy particle collisions in the world within the Large Hadron Collider (LHC) located at CERN, the European Organization





f 264 · Tweet · Share · Email

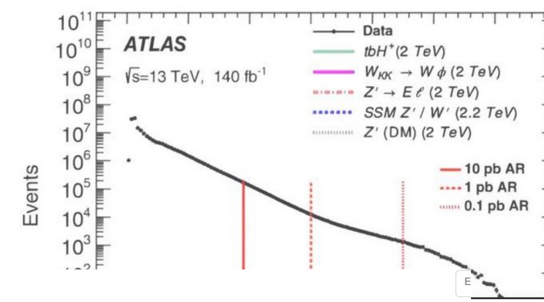
APRIL 15, 2024

Machine learning could help reveal undiscovered particles within data from the Large Hadron Collider

by Savannah Mitchem, Argonne National Laboratory

Editors' notes

Add as preferred sour



Physics Briefing

Tags: new physics, 2023 summer conferences, machine learning

ATLAS searches for new phenomena using unsupervised machine learning for anomaly detection

24 August 2023 | By ATLAS Collaboration

Since starting up in 2009, the Large Hadron Collider (LHC) has been at the forefront of scientific exploration – with researchers driven to uncover new particles and phenomena that go beyond the Standard Model. Over the years, thousands of scientists have channelled their expertise into refining analysis techniques and developing new ways to find these new physics phenomena.

Traditionally, searches for new physics use complex computer simulations to reproduce what Standard Model processes should look like in collisions recorded by the ATLAS Experiment. These are then compared to simulations of new physics models (e.g. dark matter, supersymmetry, etc.). Such models also help physicists determine the types of collisions where new physics processes would be very prominent or where the collisions cannot be described by Standard-Model simulations –

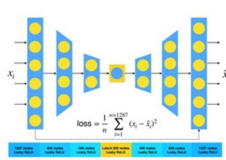
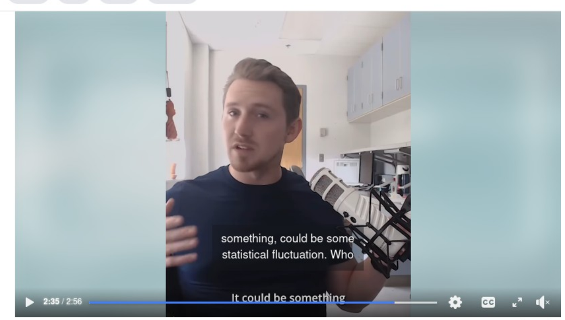


Figure 1: A schematic representation of the autoencoder architecture used for training and selection of the three anomaly regions. (Image: ATLAS Collaboration)

facebook

Video Home Live Reels Explore



Explained: Anomaly Detection

Like · Comment · Share · 87 · 7 comments · 1.2K views

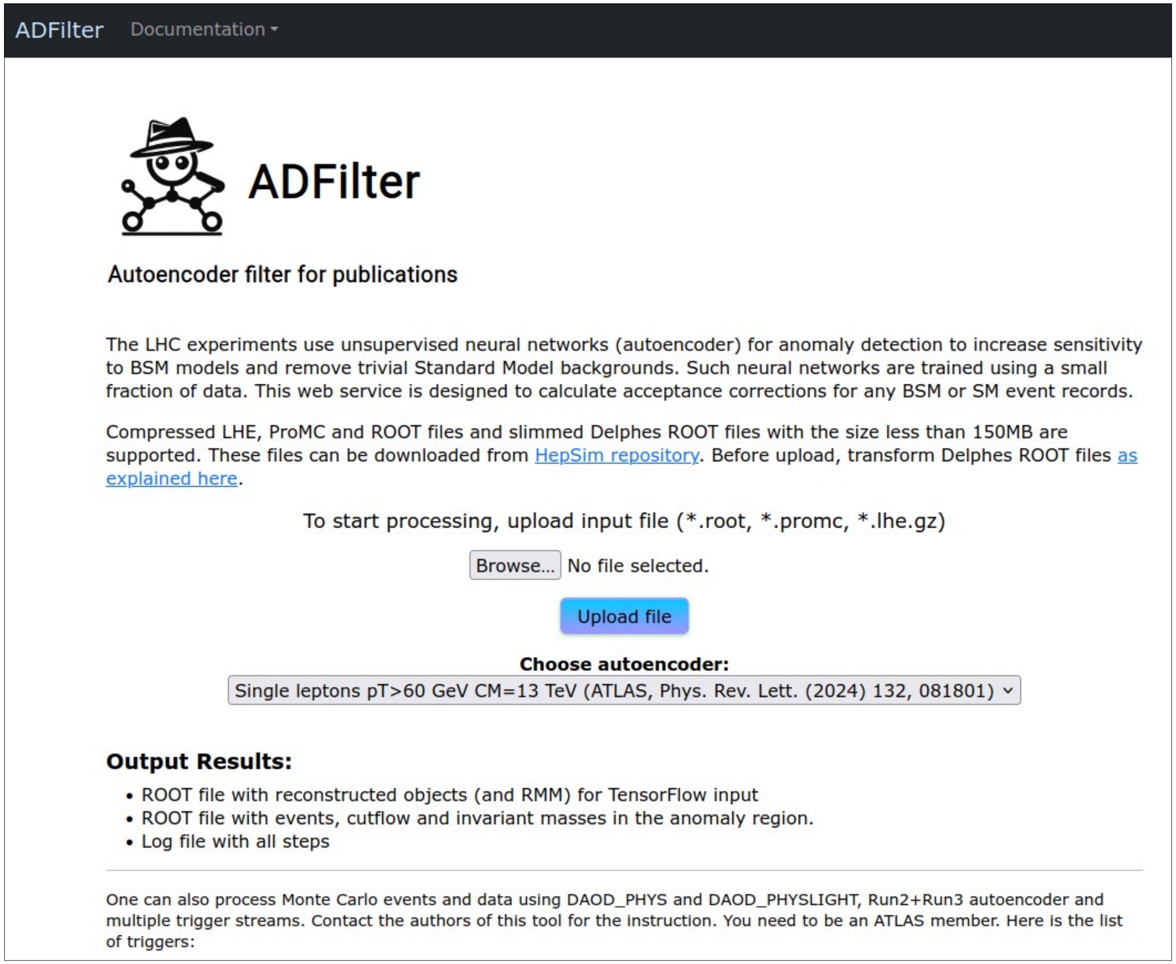
Limit interpretation by theorists

Information 2025, 16(4), 258;

ANL contribution

- To compare published limits, you need to process a BSM model using public autoencoder trained on ATLAS data
- **Theorists can exclude their favorite BSM model after estimating the autoencoder acceptance**
- This can be done with the ADFilter
- The description:
  - *SC, W.Islam, R.Zhang, N.Luongo*

    ADFilter —A Web Tool for New Physics Searches with Autoencoder-Based Anomaly Detection Using Deep Unsupervised Neural Networks
    arXiv:2409.03065



Test BSM models using 7-trigger streams with Run2+Run3 autoencoder after uploading LHE

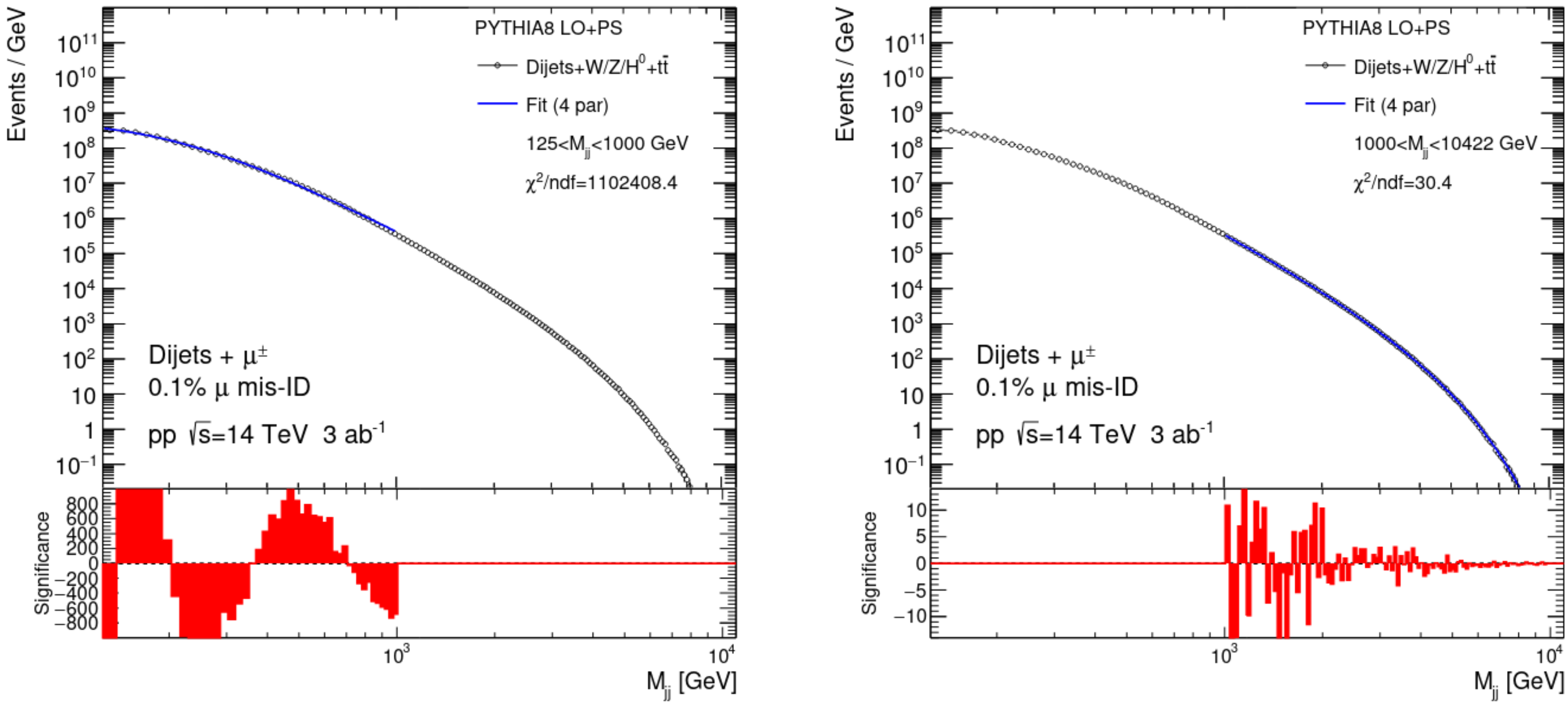
https://mc.hep.anl.gov/adfilter/

Future: HL-LHC studies

ANL contribution

HL-LHC data will be so precise that all analytic functions we tested fail to describe the Monte Carlo predictions, and no fully analytic Standard Model prediction exists.

How can BSM signals be extracted?



(a) 14 TeV, 3 ab^{-1}, 125 < M_{jj}^{fit} < 1 TeV

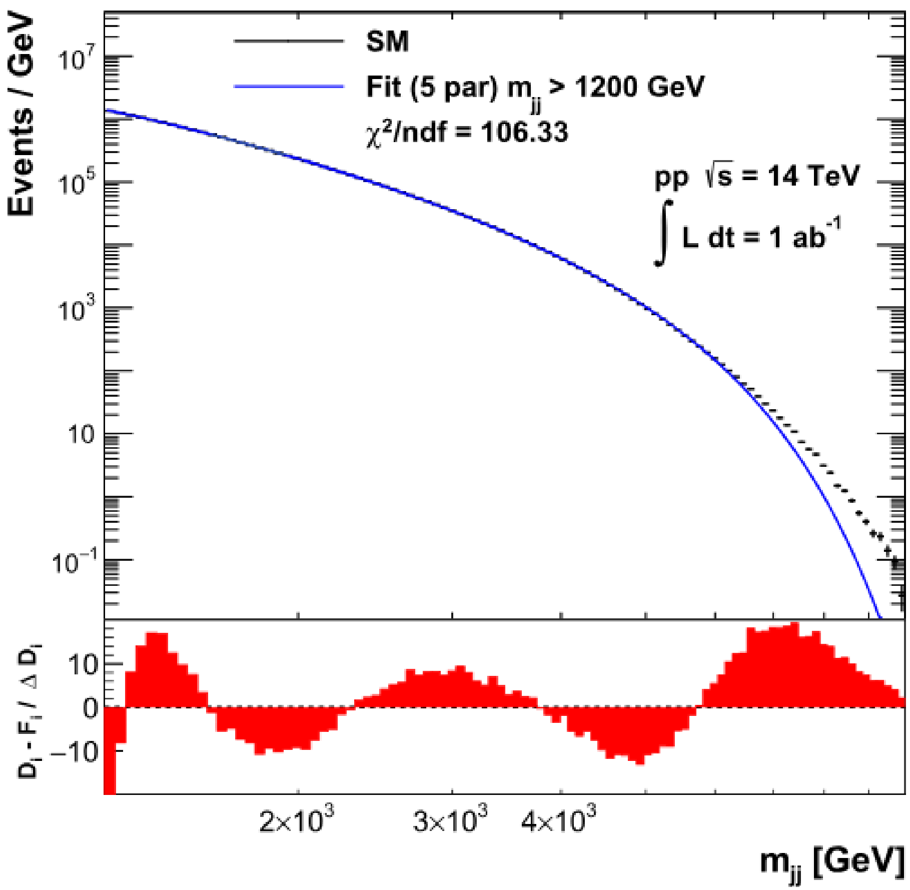
(b) 14 TeV, 3 ab^{-1}, M_{jj}^{fit} > 1 TeV

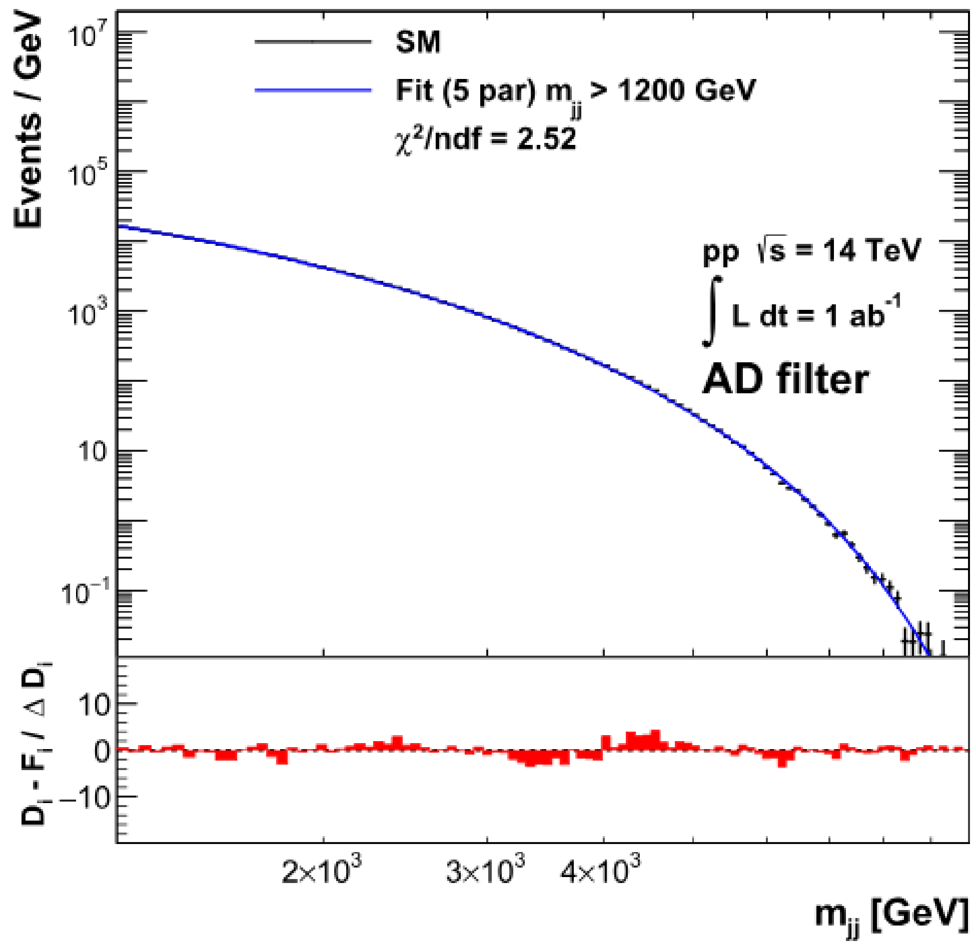
Remove SM and restore the fit!

S.C., R.Zhang,
Eur. Phys. J. Plus (2024) 139

ANL contribution

In the Anomaly Region





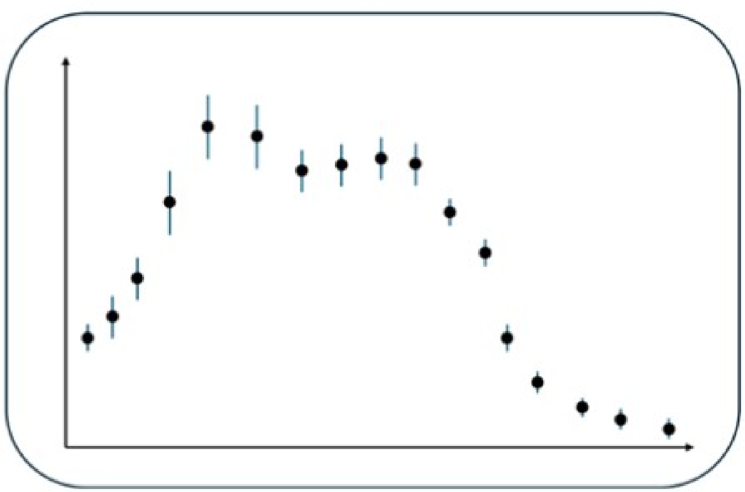
The validity of the description by p5 function has been restored after autoencoder!

If still does not walk? Use symbolic regression!

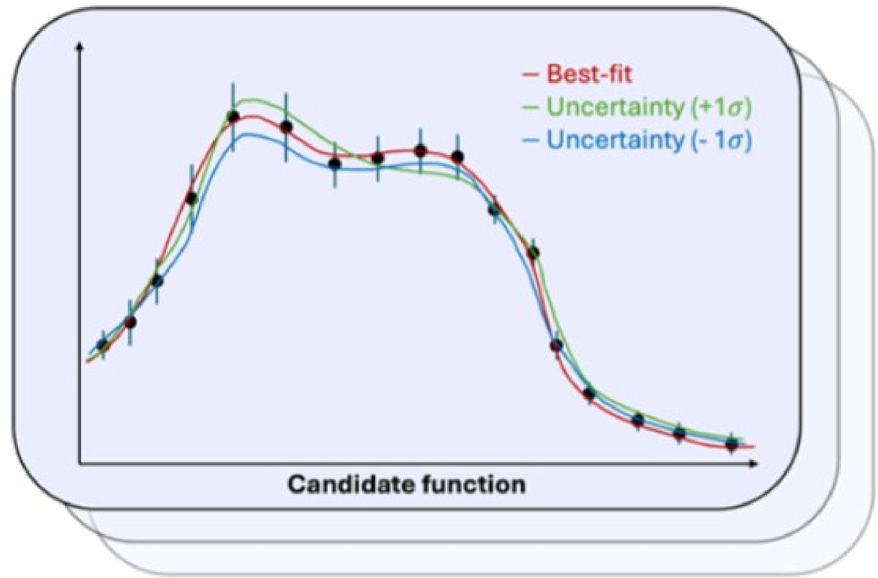
Use some AI method. Symbolic regression?

Symbolic regression is a machine learning technique used to discover the underlying mathematical expression (function) that best fits a given dataset, without pre-specifying the form of that function

Often implemented via “genetic programming”



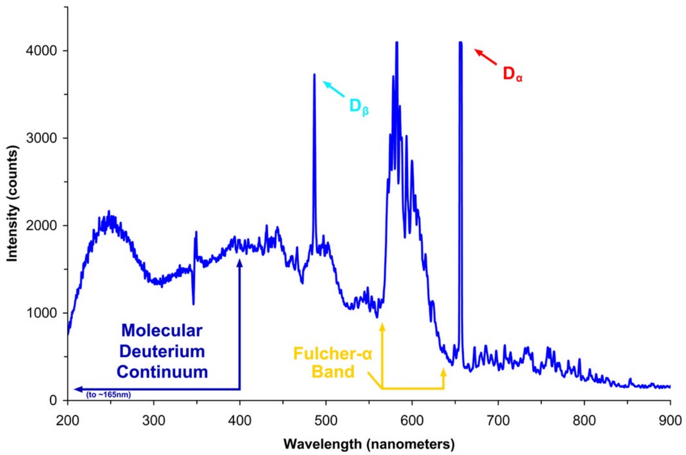
Input



See: Håkan Kjellerstrand etc. JGAP software, books on Constraint Solving and Planning with Picat

The Historical Context

1885, Johann Balmer derived an empirical formula that accurately described the wavelengths of the visible spectral lines of hydrogen



$$\lambda = B\left(\frac{m^2}{m^2 - n^2}\right) = B\left(\frac{m^2}{m^2 - 2^2}\right)$$

Where

- λ is the wavelength.
- B is a constant with the value of $3.645\,0682 \times 10^{-7}$ m or 364.506 82 nm.
- m is the initial state $(m \in \mathbb{N}, m > n)$.
- n is the final state $(n \in \mathbb{N}, n \geq 1)$ generally, and $(n = 2)$ for the Balmer series.

No any science, just human intuition and expectations that the nature should be described by simple mathematical patterns

1913 The function was later explained by quantum mechanics through Bohr's model

Now data are so complicated that making such pattern matching by people is impossible

We need some AI method to find simplest patterns, which may later indicate underplaying dynamics

There is something unexpected for AI studies!

What if data are "Heterogeneous", i.e. contain different scales, physics units?

Constant	Name	Value	$\pm\varepsilon$	$\pm\varepsilon^{rel}$ (%)
PI	π	3.14159	1e-05	0.0003
Fine-struct. (inv)	α^{-1}	137.036	0.001	0.0007
α_s at Z^0	α_S	0.1180	0.0009	0.7627
CKM constants		no units	$\pm\varepsilon$	$\pm\varepsilon^{rel}$ (%)
12-mix angle	θ_{12}	0.22501	0.00068	0.3022
23-mix angle	θ_{23}	0.04183	0.00079	1.8886
13-mix angle	θ_{13}	0.003732	9×10^{-5}	2.4116
CP-viol. phase	δ	1.147	0.026	2.2668
Mass		MeV	$\pm\varepsilon$	$\pm\varepsilon^{rel}$ (%)
electron mass	m_e	0.510998	10^{-6}	0.0002
muon mass	m_μ	105.658	0.001	0.0009
τ mass	m_τ	1776.93	0.09	0.0051
u-quark mass	m_u	2.16	0.07	3.2407
d-quark mass	m_d	4.70	0.07	1.4894
s-quark mass	m_s	93.5	0.8	0.8556
c-quark mass	m_c	1273.0	4.6	0.3614
b-quark mass	m_b	4183	7	0.1673
t-quark mass	m_t	172560	310	0.1796
Z-boson mass	m_Z	91188.0	2.0	0.0022
W-boson mass	m_W	80369.2	13.3	0.0165
H-boson mass	m_H	125200	110	0.0879

Create a library of analytic expressions that relate all fundamental parameters

~ 0.5 millions with "simplicity" rank up to 50

Filter out numeric noise:
1) numeric noise, i.e. expresions that do not pass dimensional analysis
2) expressions with masses which do not vanish when $m_H \to 0$

What remains is a set of linked equations which possible contain a structure reflecting true connections between SM

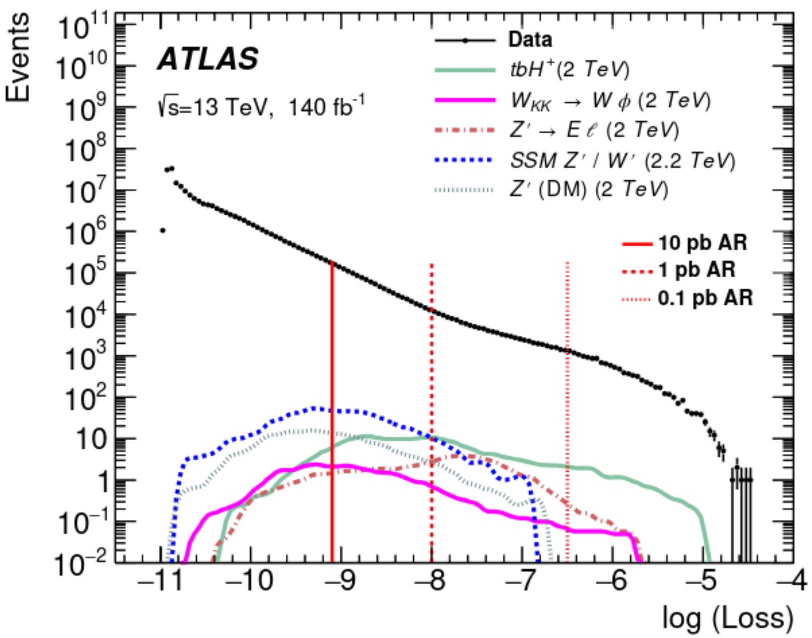
Still requires AI for further simplifications and identification of possible symmetries

Interested?
S. V. Chekanov and H. Kjellerstrand, "Discovering the Underlying Analytic Structure within Standard Mod
el Constants Using Artificial Intelligence", (2025) Particles , Vol8, 4, 95

The future use of Agentic AI

ATLAS Collaboration

Phys. Rev. Lett. 132 (2024) 081801



“anomaly scan”

- In ATLAS, we could do only 3 anomaly working points (0.1,1,10) pb
- For each point:
  1) Calculate 7 invariant masses
     (for Run2+Run3 analysis – 63 masses)
  2) Establish background hypotheses and check its robustness (p2,p3,.. p6), fit 63 masses
  3) Search for “bumps”, find a signal if any
- These (1)-(4) steps can be automated using agentic AI, smoothly scanning **log(loss)** from 0.001 .. to 100 pb
- Our division can do this using LHC open data

Conclusions

- Massive, pristine LHC dataset of 450 fb^{-1} are waiting to be analyzed.
- Traditional methods can only find what we think might be there, but model-agnostic approaches reveal a vast phase space of event categories that remains largely unexplored.
- We were among the first groups to bring event-based anomaly detection to the LHC (circa 2020).
- Still a tiny fraction from ~50,000 invariant masses in exclusive categories that can be explored!
- **This year:**

Search in 63 invariant masses in anomaly region using 7 independent triggers and Run2+Run3 (300 fb^{-1})!