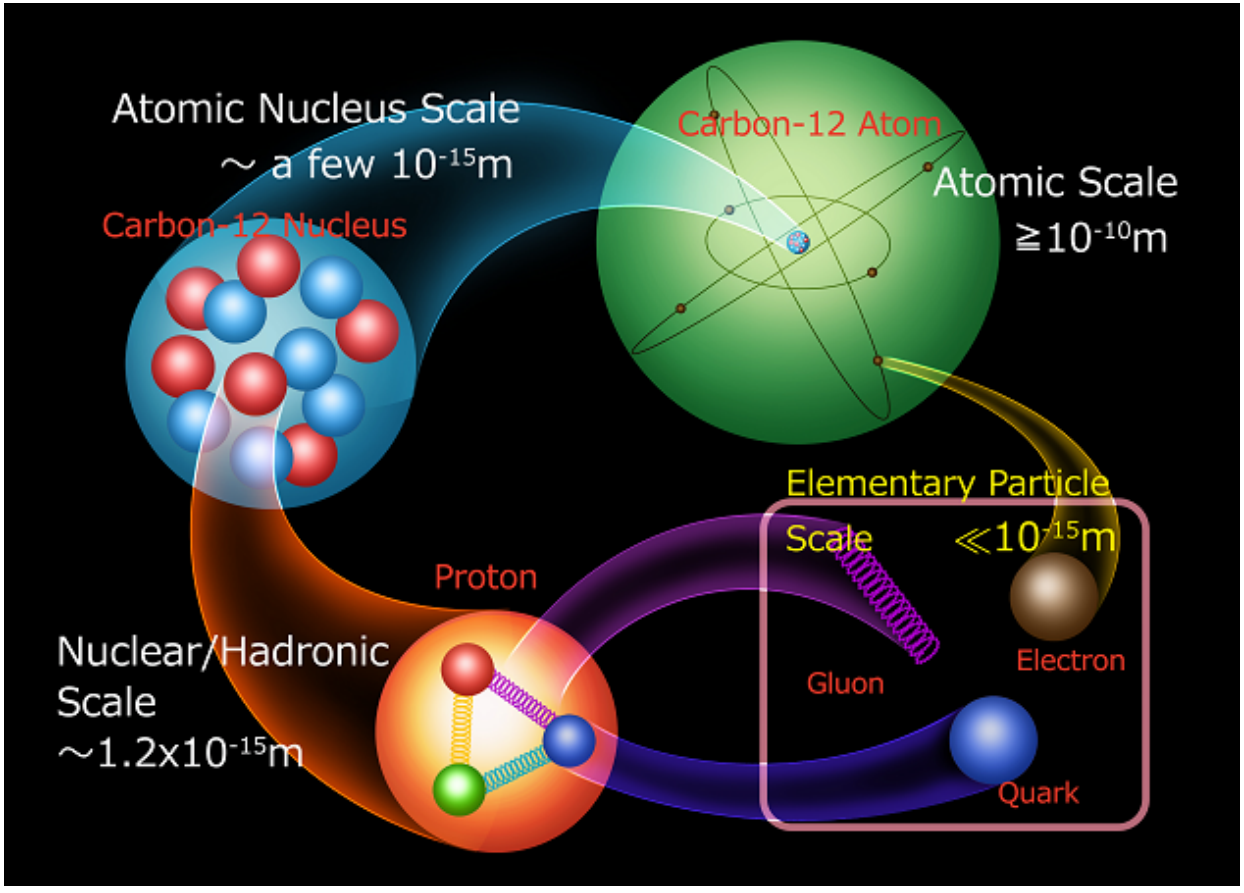
ML-based Model Agnostic Searches in Particle Physics

Oz Amram
CMU STAMPS
May 14th, 2026

Fermilab

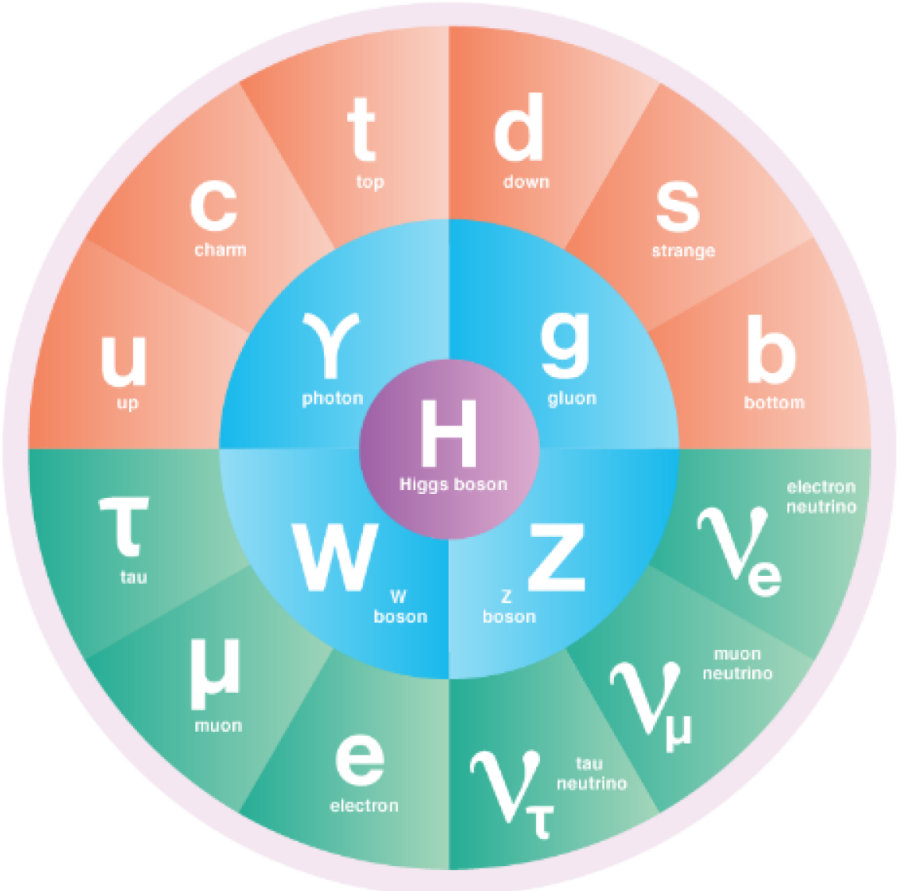
Inspired by ‘Model-Agnostic Signal Discovery with ML: Bridging the Gap Between Theory and Practice’
(in prep)
with Mikael Kuusela & Marco Letizia

Particle Physics



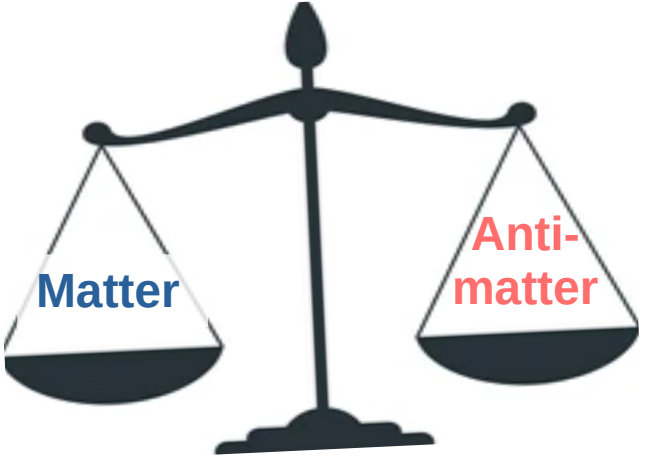
What are the fundamental particles and forces that shape our universe?

What We Know So Far:
The Standard Model

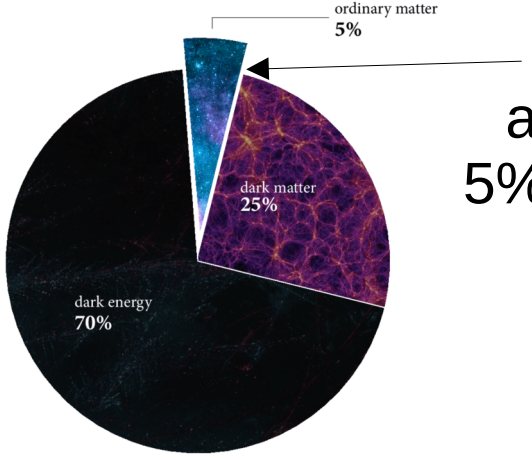


Beyond Standard Model Questions

Dark Matter/Energy?



Why is the universe dominated by matter and not anti-matter?



SM only accounts for 5% of universe!

Neutrino Mass?

Additional Higgs bosons?

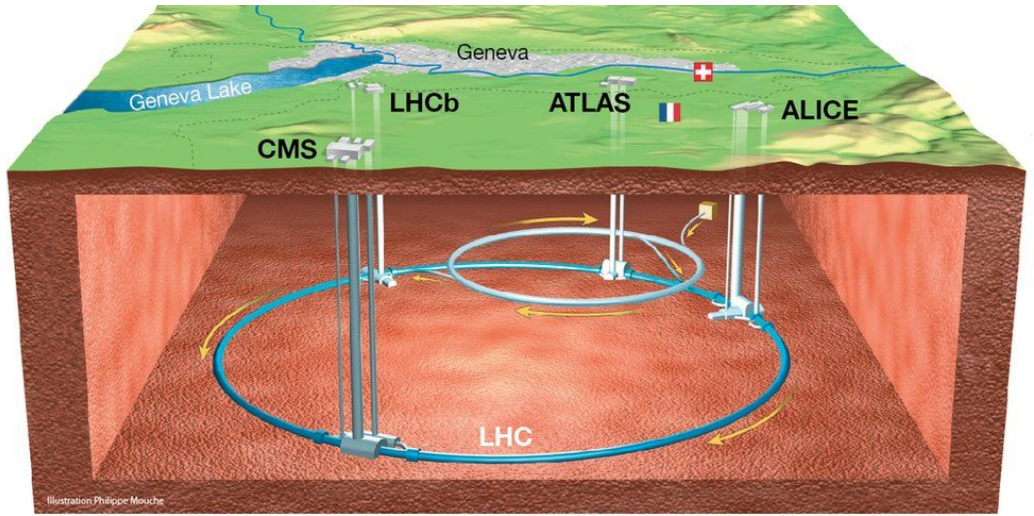
Grand Unification ?

And many more…

Strong CP problem?

Origin of SM parameters?

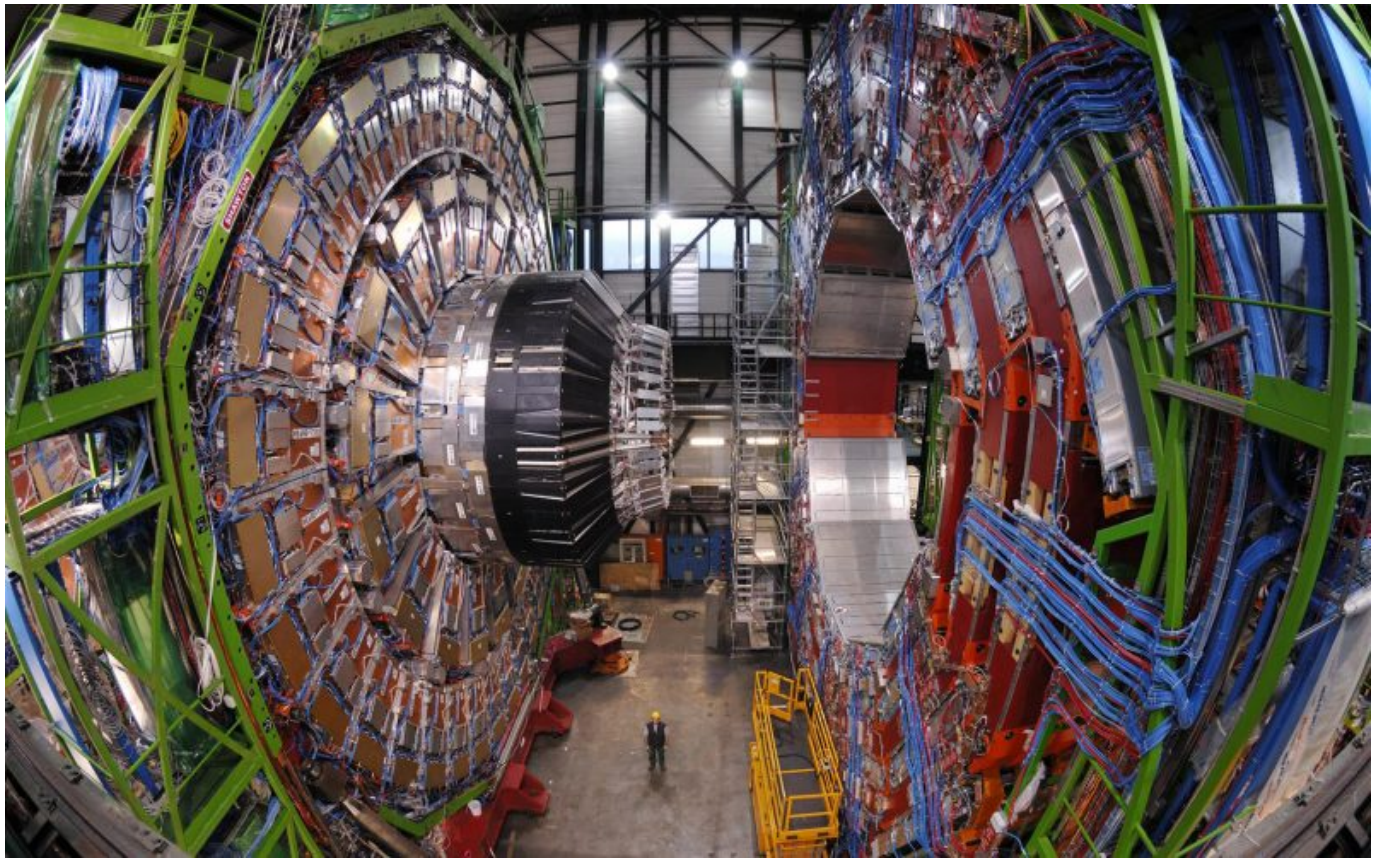
The LHC



LHC accelerates protons to highest energies **ever achieved** (6.8 TeV)

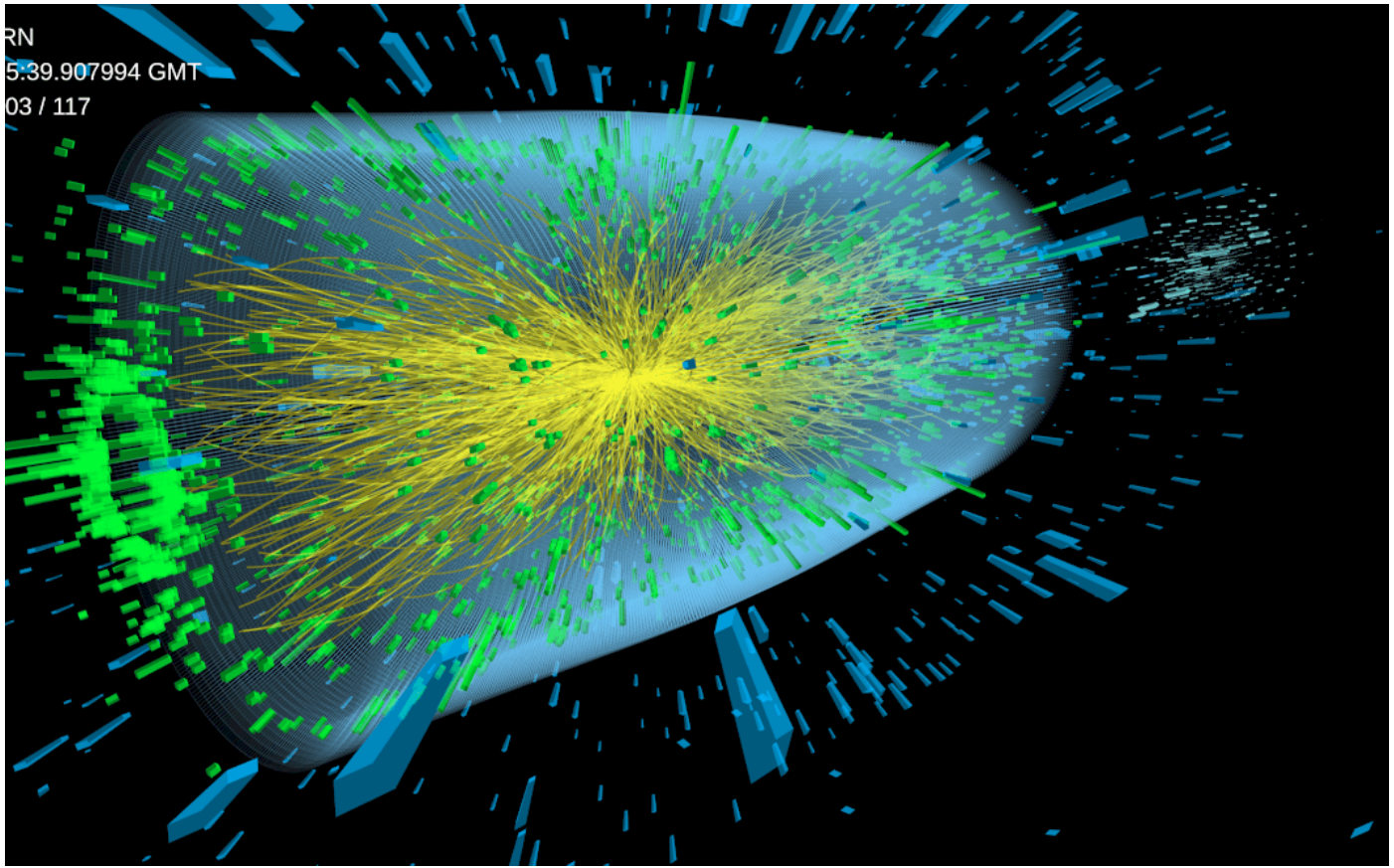
Collides bunches of protons **40 million times a second**

Detectors



Massive multi-layer detectors to record these collisions

Detectors



Massive multi-layer detectors to record these collisions

Needles in a Haystack

New particles in
< 1 in a billion collisions

Need to cut away ‘background’ collisions
And find those which may contain a signal

Hypothesis testing for discovery of new particles

Assume data is a mixture model of **signal** (new particle) and **background** (standard model)

- Known physics: *p_b(z)*
- New signal: *p_s(z)*
- Nature: *p_data(z)* = *(1−**α**)* *p_b(z)* + ***α*** *p_s(z)*

Want to test H_0: $\alpha = 0$ vs. H_1: $\alpha > 0$

Usually done with a (profile) likelihood ratio test statistic

$$t = \frac{p_{data}(z|\alpha)}{p_{data}(z|\alpha = 0)} = \prod_i \frac{(1-\alpha)p_b(z_i) + \alpha\, p_s(z_i)}{p_b(z_i)}$$

Reject H_0 at high significance (5σ) → claim discovery

Hypothesis testing for discovery of new particles

Assume a mixture model of **signal** (new particle) and **background** (standard model)

- Known physics: *p_b(z)*
- New signal: *p_s(z)*
- Nature: *p_data(z) = (1−**α**) p_b(z) + **α** p_s(z)*

Challenge!
p(z)'s not known well in high-dimensions

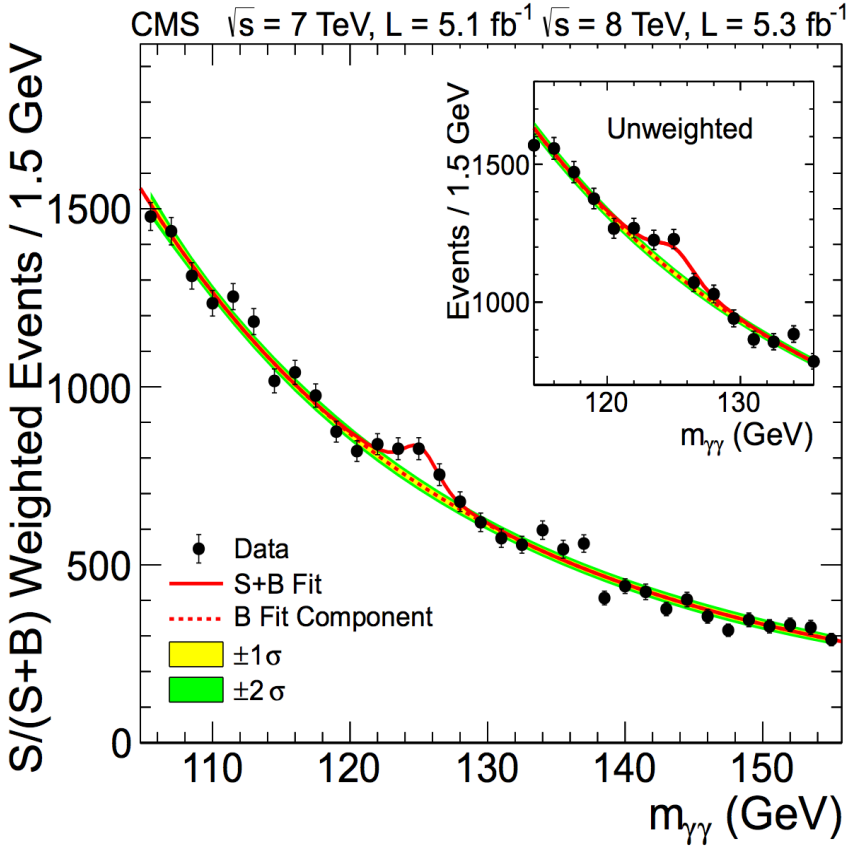
Want to test H_0: $\alpha = 0$ vs. H_1: $\alpha > 0$

Usually done with a (profile) likelihood ratio test statistic

$$t = \frac{p_{data}(z|\alpha)}{p_{data}(z|\alpha = 0)} = \prod_i \frac{(1-\alpha)p_b(z_i) + \alpha\, p_s(z_i)}{p_b(z_i)}$$

Reject H_0 at high significance (5σ) → claim discovery

Classic Solution: Collapse down to 1D observable with known analytic form

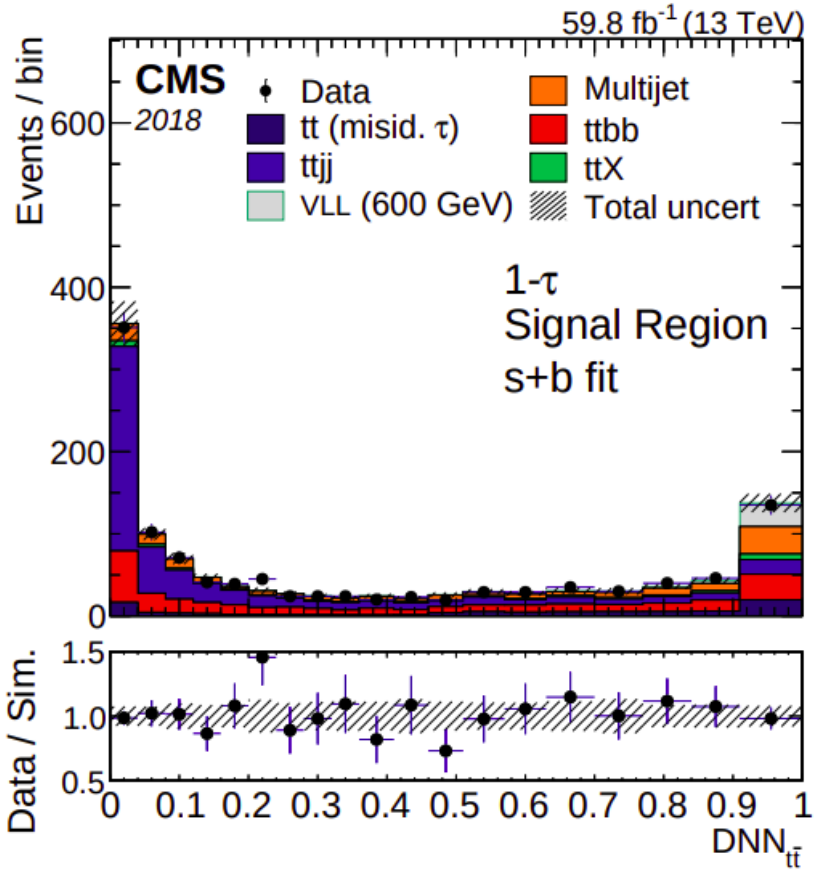


Classic Method : 'Bump Hunt'

- Apply selection criteria on high-dimensional data to restrict to part of mixture signal is assumed to live in
- Observable = Mass of some combination of objects (here two photons)
- Signal and background distributions via known parametric functional forms

→ Likelihood ratio hypothesis test

Now : Commonly Using Supervised ML

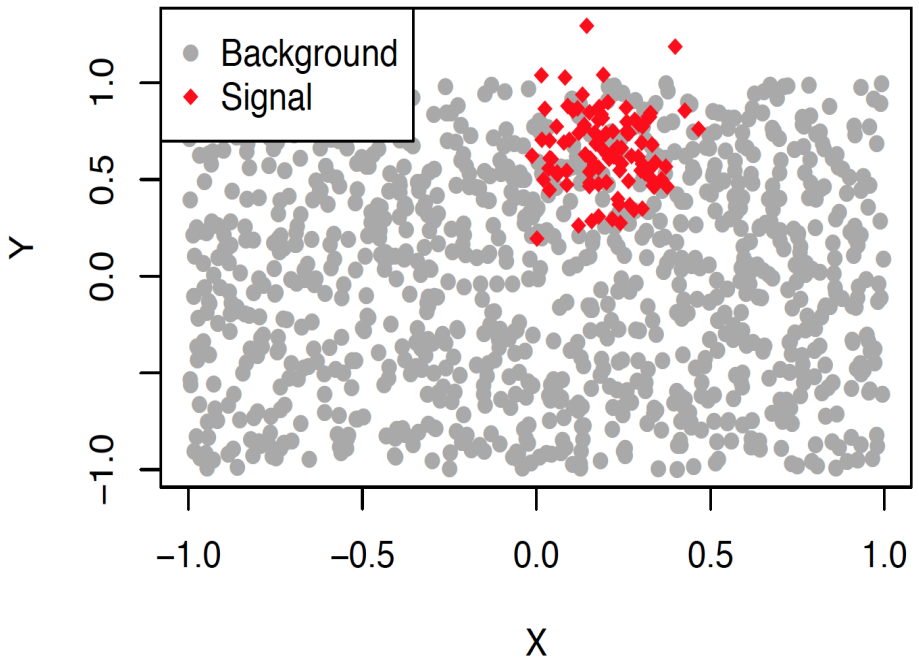


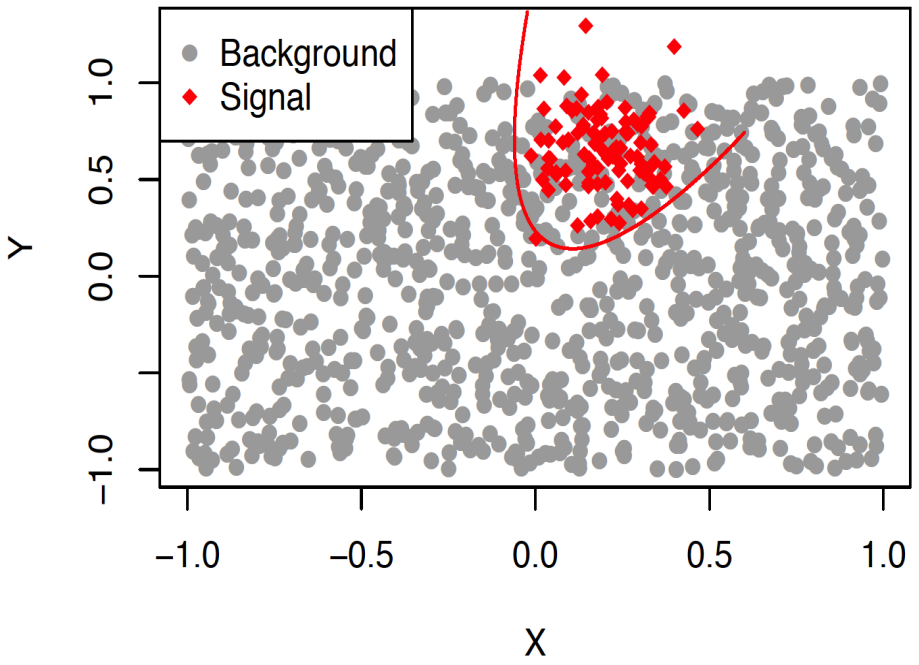
Classic Method : 'Bump Hunt'

- Train ML classifier on simulated signal sample
- Observable = ML classifier score (likelihood ratio)
- Binned histogram fit
- Signal and background distributions from simulation + bespoke methods

→ Likelihood ratio hypothesis test

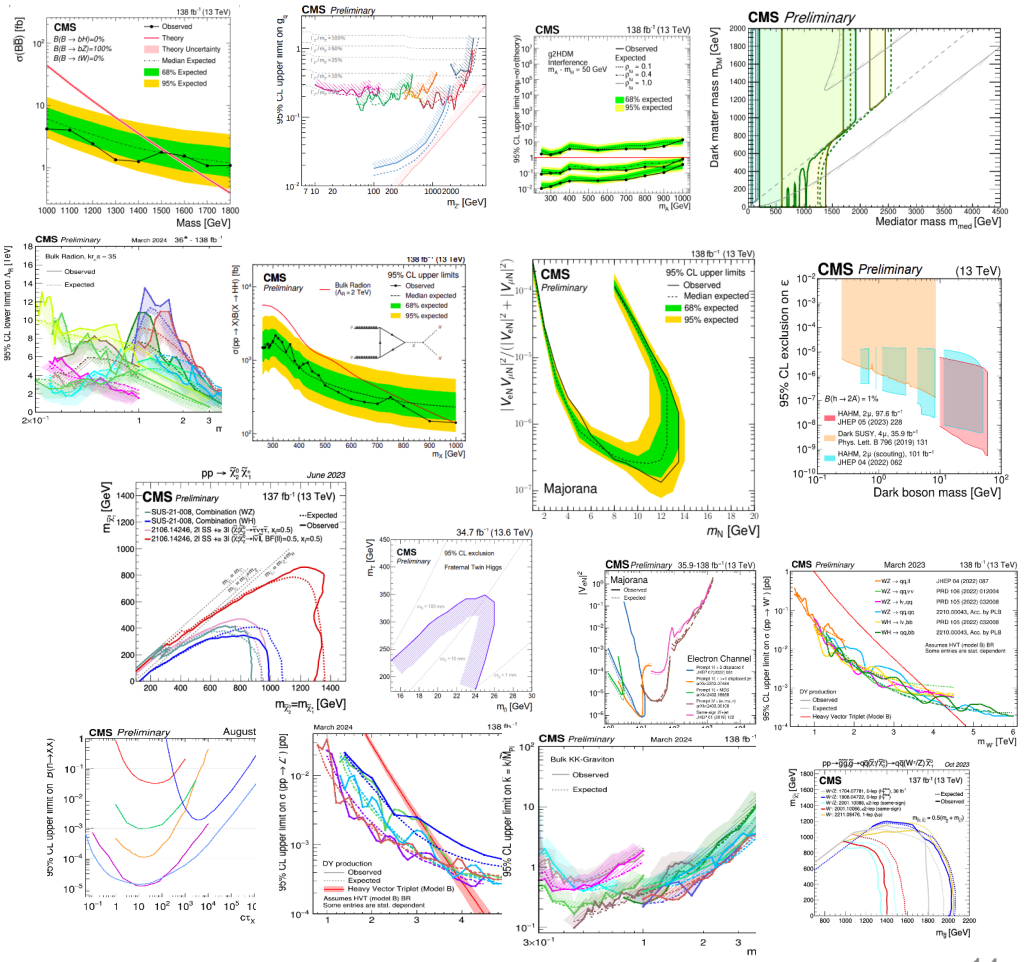
Classifier training





LHC Post-Higgs Discoveries?

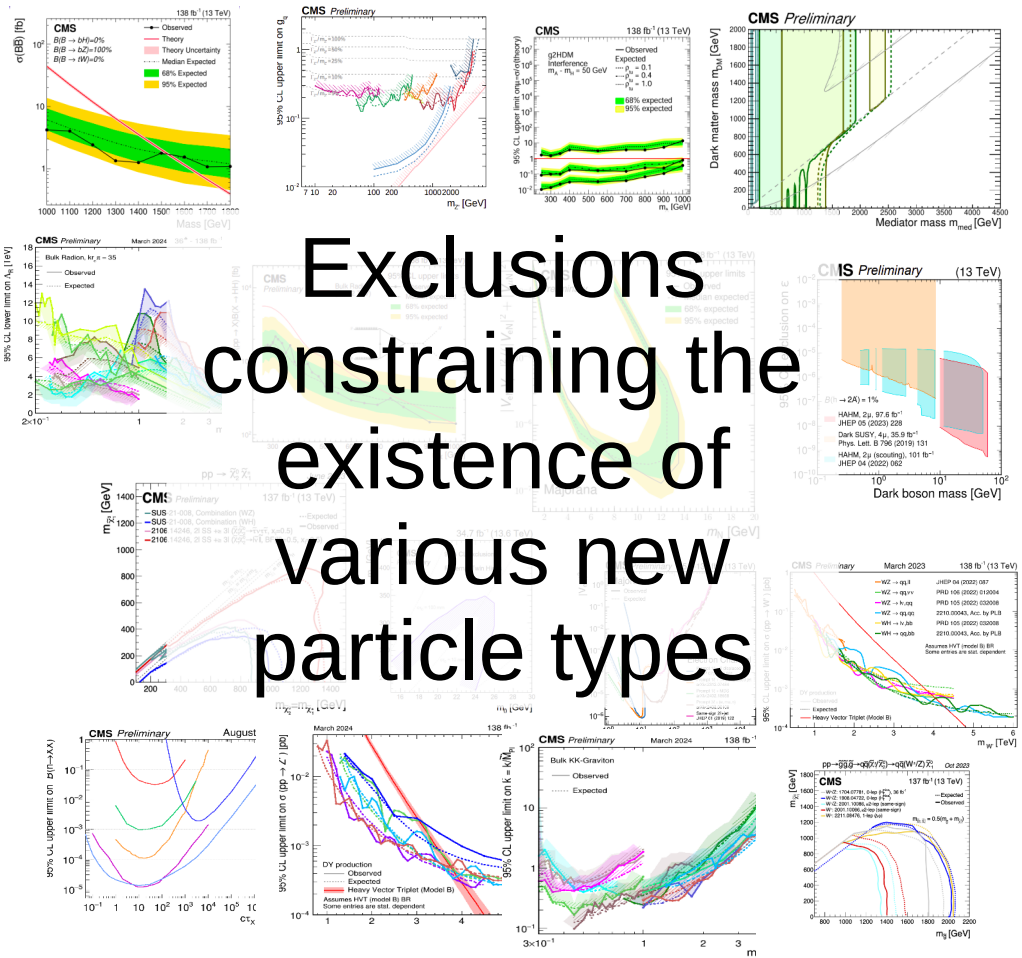
None so far...



DM graphic from T. Tait

LHC Post-Higgs Discoveries?

None so far...

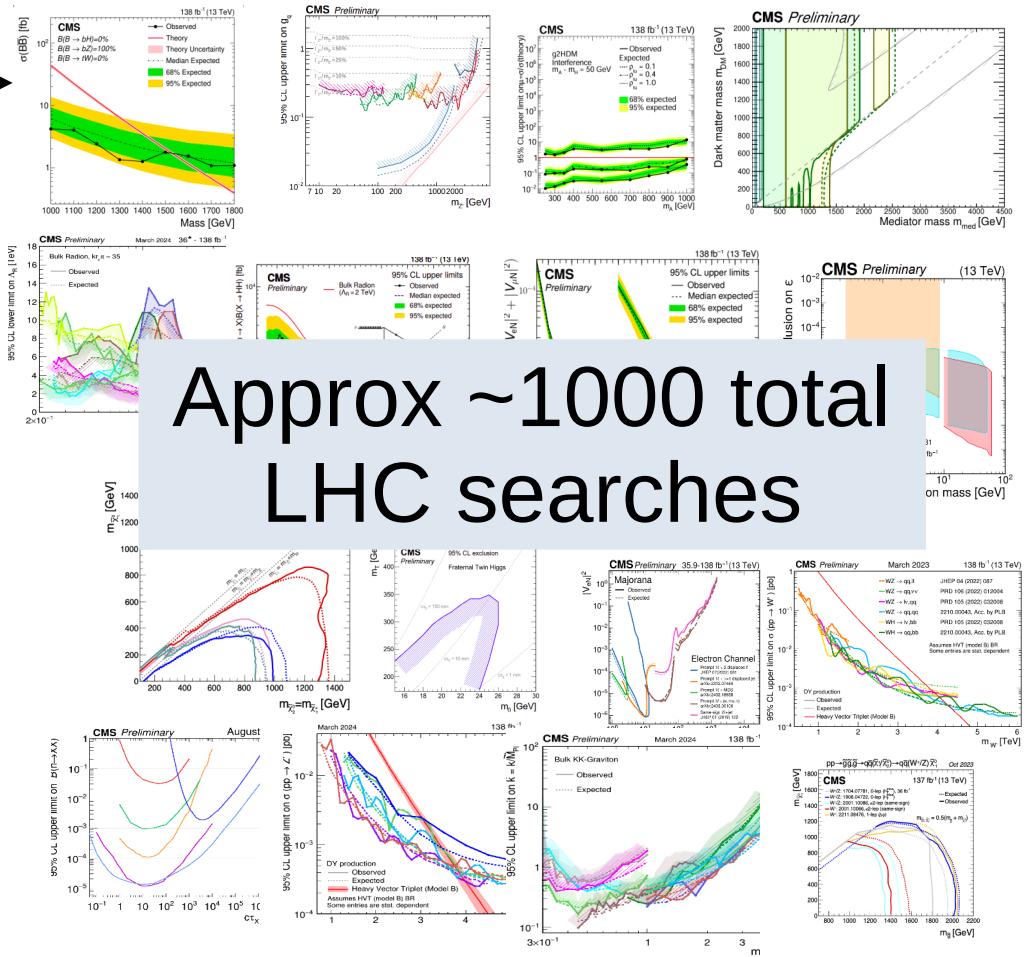


DM graphic from T. Tait

LHC Post-Higgs Discoveries?

Each is ~3 person-years of work

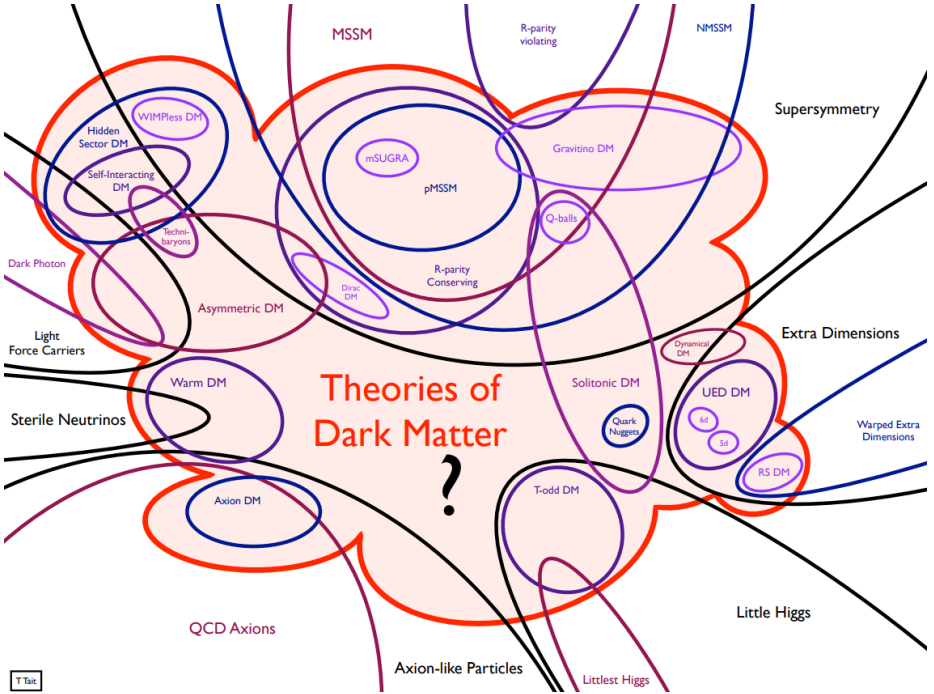
Huge amount of effort!



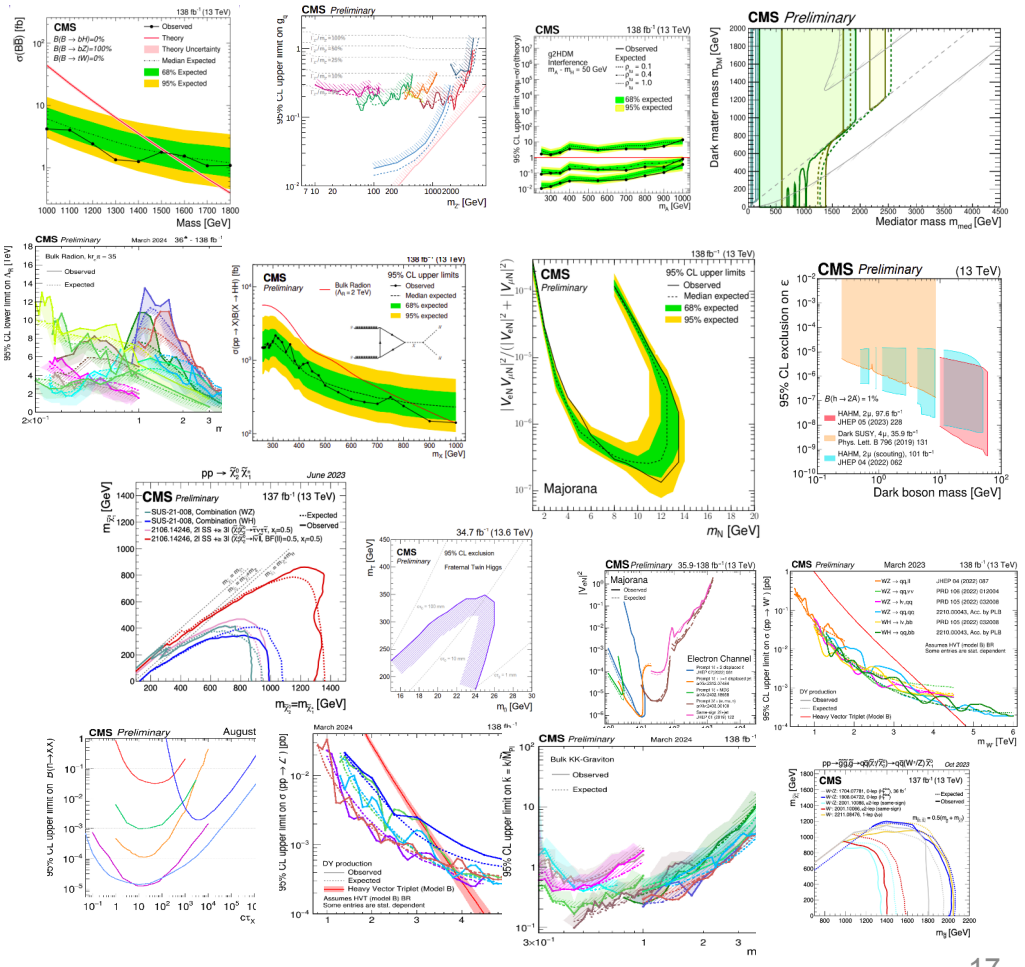
DM graphic from T. Tait

LHC Post-Higgs Discoveries?

But...



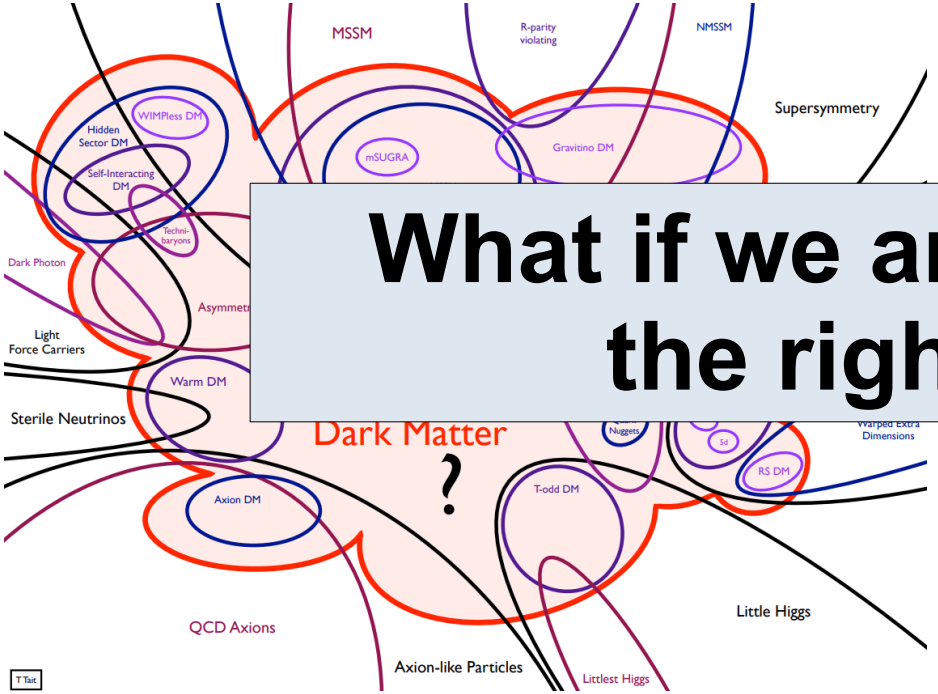
>



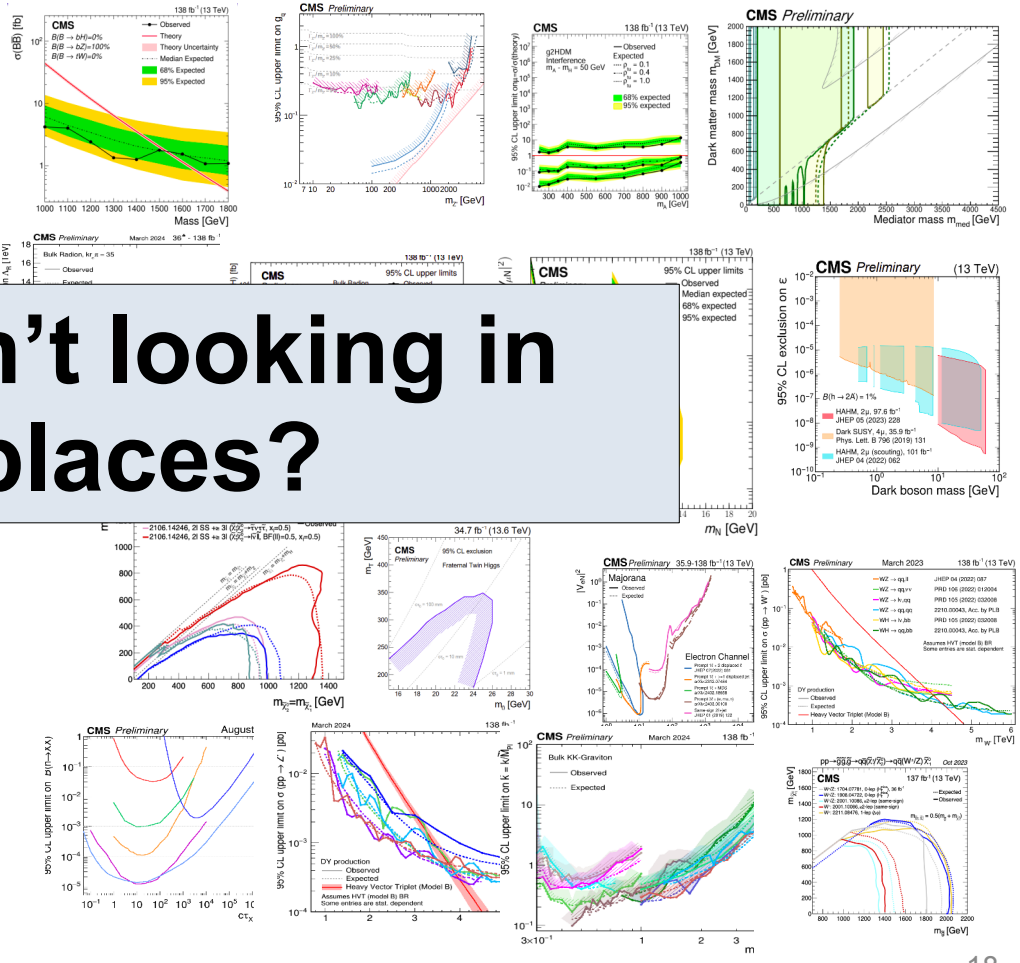
DM graphic from T. Tait

LHC Post-Higgs Discoveries?

But...

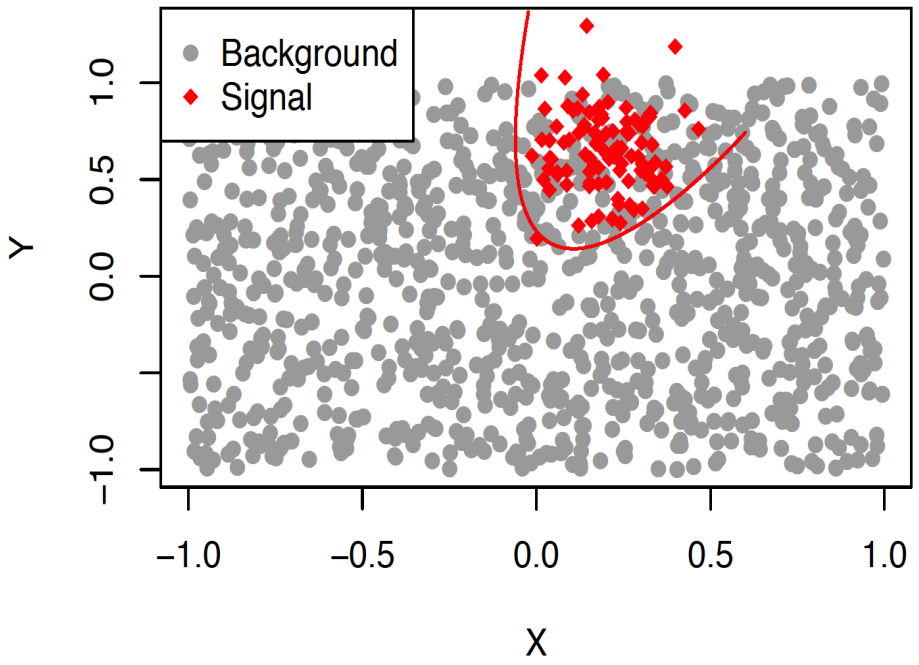


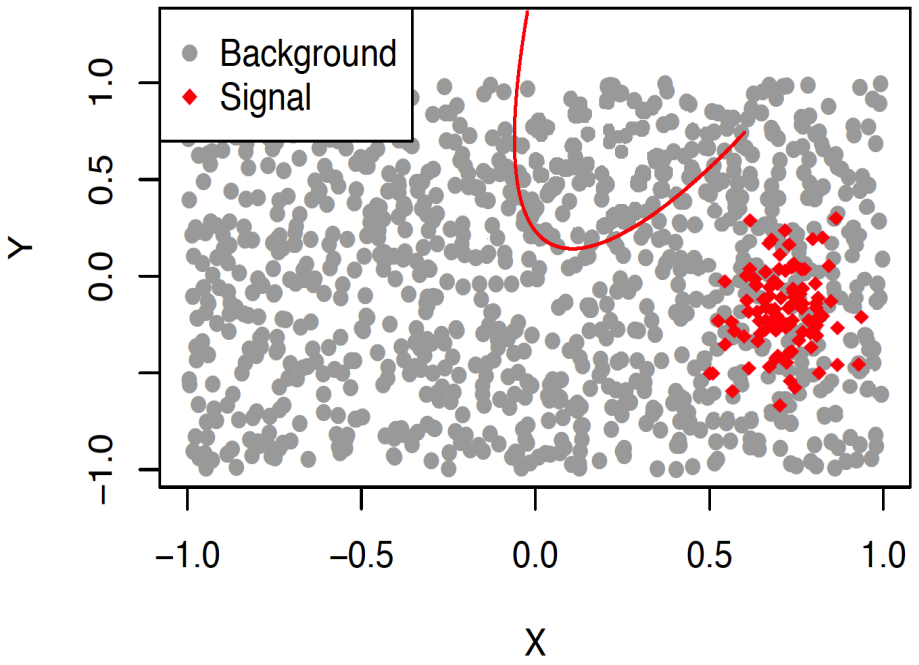
What if we aren't looking in the right places?



DM graphic from T. Tait

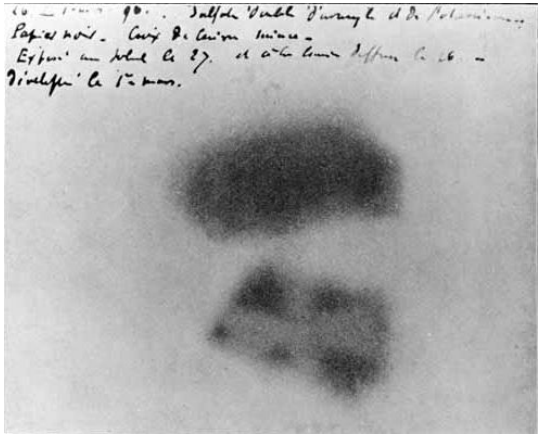
Systematically misspecified signal





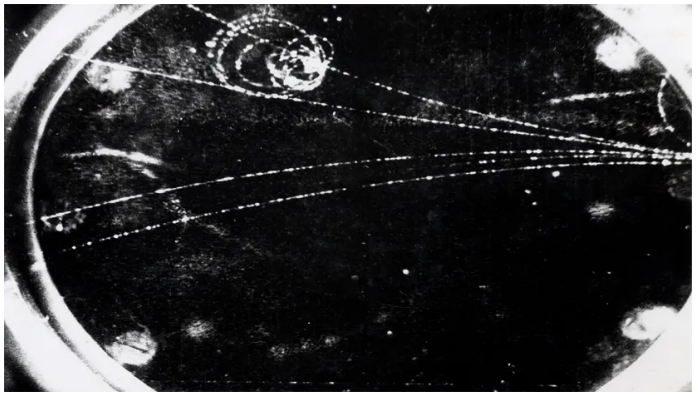
Surprise Discoveries in Fundamental Physics

Radioactivity (1896)



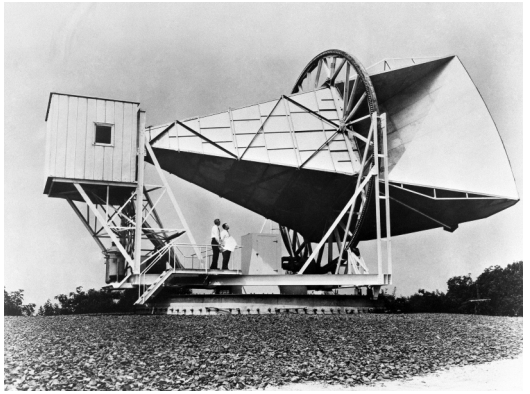
Source accidentally left alone with film in a closet

Muon (1936)



"Who ordered that?" – I.I. Rabi

Cosmic Microwave Background (1964)



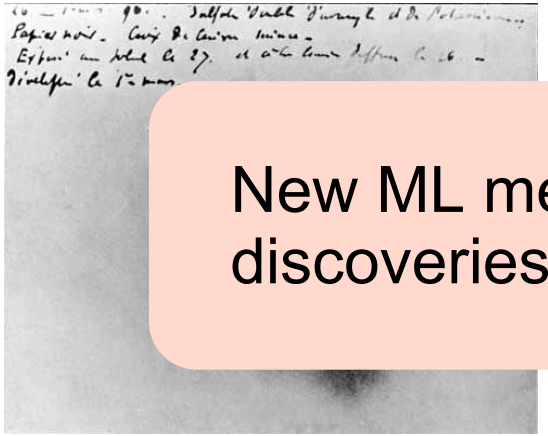
First thought to be bird poop on the antenna!

Arguably many others as well!

Kaon, Parity Violation, CP violation, neutrinos, neutrino oscillation, dark energy, ...

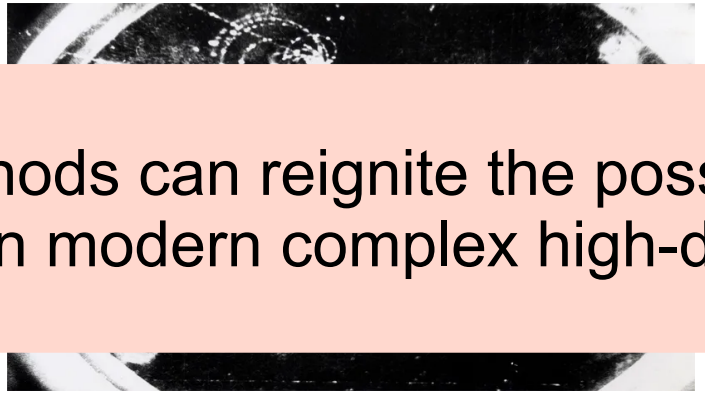
Surprise Discoveries in Fundamental Physics

Radioactivity (1896)



Source accidentally left alone with film in a closet

Muon (1936)



"Who ordered that?" – I.I. Rabi

Cosmic Microwave Background (1964)

First thought to be bird poop on the antenna!

New ML methods can reignite the possibility of surprise discoveries in modern complex high-data experiments!

Arguably many others as well!

Kaon, Parity Violation, CP violation, neutrinos, neutrino oscillation, dark energy, ...

Anomaly Detection Landscape

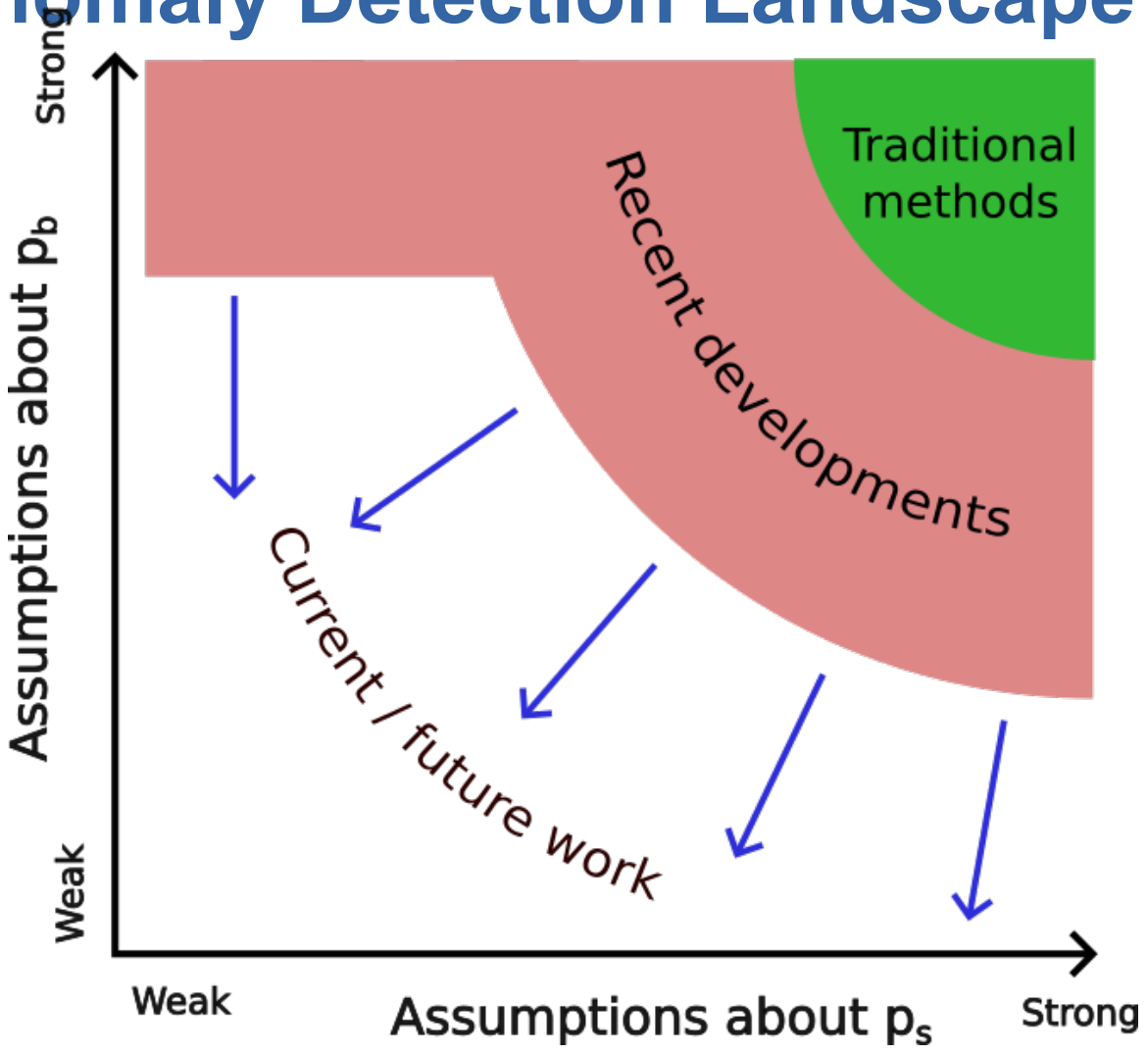


Figure credit: Mikael Kuusela

Types of anomalies

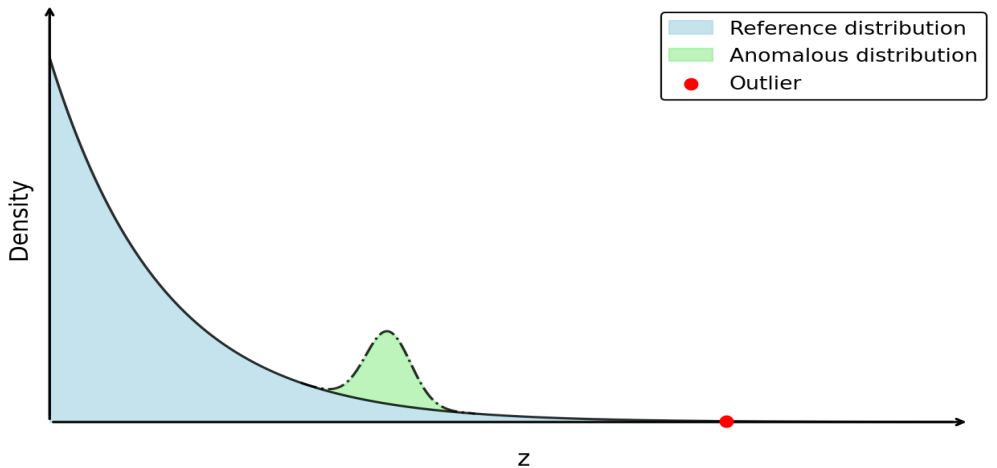
Many anomaly detection applications in HEP can be understood as **collective anomaly detection** (e.g., Chandola 2009)

- Is there a *collection* of data points which taken together deviate from the anticipated data?

Notice that:

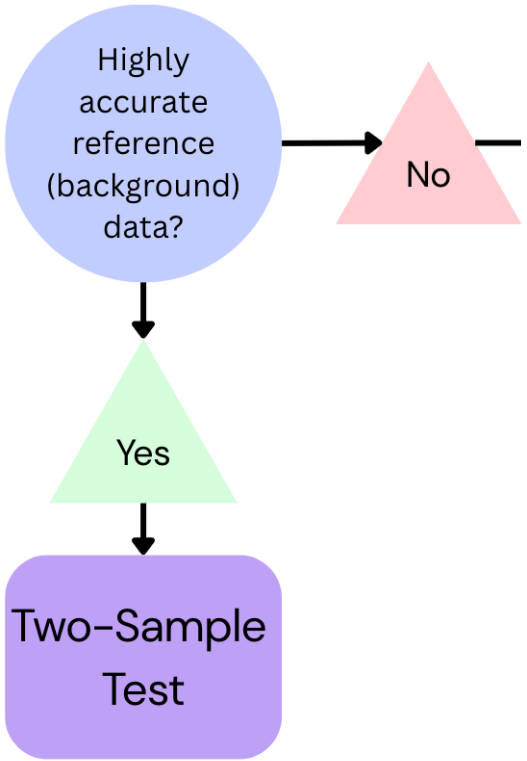
collective anomaly detection ≠ outlier detection

In collective anomaly detection, each signal event is typically indistinguishable from the background on its own; it is the *collection* of many signal events that defines the excess



Flowchart of Anomaly Detection Methods

(Personal opinion)



Two-sample testing

Two-sample testing refers to the following hypothesis testing problem:

Let $X_1,\ldots,X_n \sim p_1$ i.i.d. and $Y_1,\ldots,Y_m \sim p_2$ i.i.d.
Test $H_0: p_1 = p_2$ vs. $H_1: p_1 \neq p_2$

Lots of classical tests in the univariate case (KS, AD, Cramér–von Mises, …)

New in recent years: use **ML-based classifiers** to perform the test in high-dimensional spaces

- Basic idea: train a classifier to separate $X_1,\ldots,X_n$ from $Y_1,\ldots,Y_m$
  - If classifier can distinguish between the two samples → evidence $p_1 \neq p_2$

By taking $p_1 \equiv$ ***p_b*** and $p_2 \equiv$ ***p_data***, the model-agnostic search problem can be framed as a two-sample testing problem

(D'Agnolo & Wulzer 2018; D'Agnolo et al. 2019; Letizia et al. 2022; Chakravarti et al. 2023)

- **Assumes** that a reliable sample from ***p_b*** is available

Example: Two-sample tests with NPLM

The **New Physics Learning Machine** is a ML-based GoF / two-sample test designed as a likelihood-ratio test with data-driven hypotheses

D'Agnolo & Wulzer (2018); D'Agnolo et al. (2021); Letizia et al. (2022); D'Agnolo et al. (2022); Grosso et al. (2024)

Two main steps:

1) learning the density ratio with a classifier trained on measurements $\mathscr{X}$ and a background sample $\mathscr{Y}$:

$$f_{\hat{w}}(z) \approx \log \frac{n_{\text{data}}(z)}{n_b(z)}$$

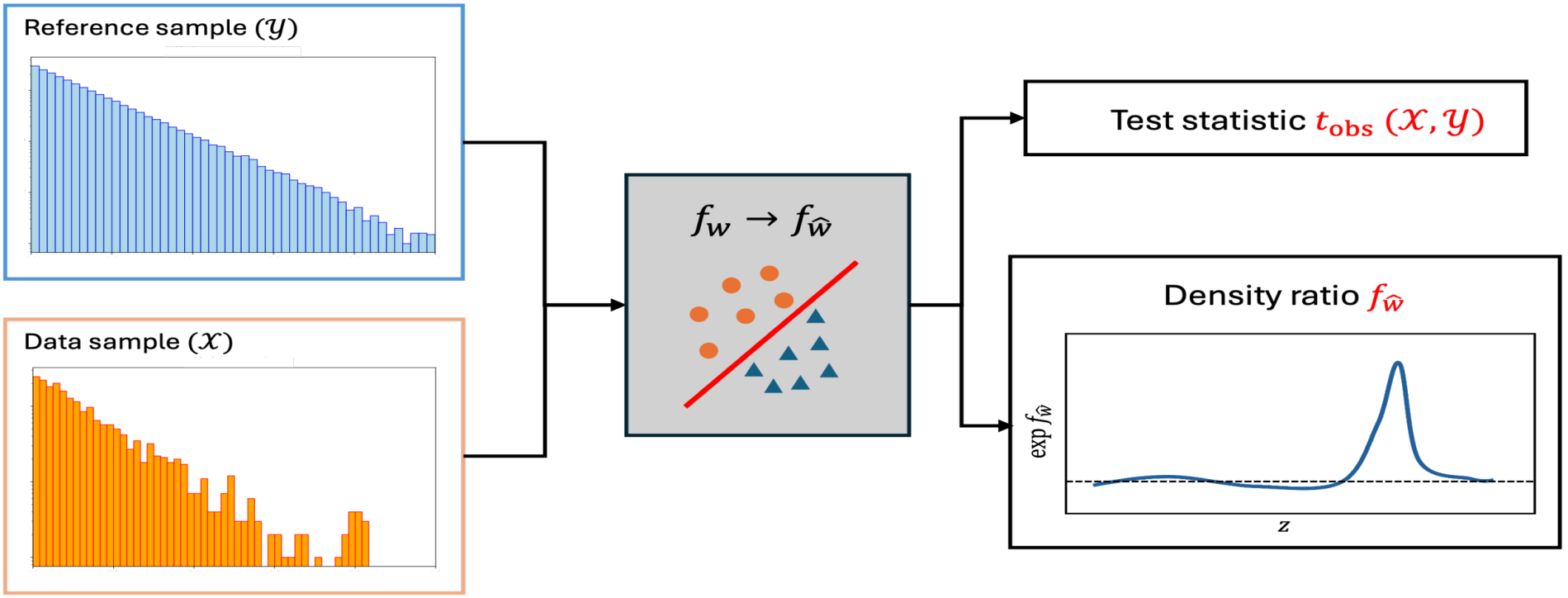
2) evaluating a Monte Carlo formulation of the extended LRT on the data:

$$t_{\text{obs}}(\mathcal{X}, \mathcal{Y}) = -2\left[\frac{N(b)}{m}\sum_{z\in\mathcal{Y}}(e^{f_{\hat{w}}(z)}-1) - \sum_{z\in\mathcal{X}} f_{\hat{w}}(z)\right] \approx 2\log \max_{w} \frac{\mathcal{L}(H_w \mid \mathcal{X})}{\mathcal{L}(b \mid \mathcal{X})}$$

Note that the likelihood ratio btwn the s+b mixture vs bkg is monotonic wrt likelihood ratio s vs b

$$L_{s+b,b} = \frac{p_{s+b}(z)}{p_b(z)} = \frac{(1-\alpha)p_b(z)+\alpha p_s(z)}{p_b(z)} = (1-\alpha)+\alpha\frac{p_s(z)}{p_b(z)} = (1-\alpha)+\alpha L_{s,b}.$$

Two-sample test with NPLM



An illustration of the NPLM method (input data are unbinned).

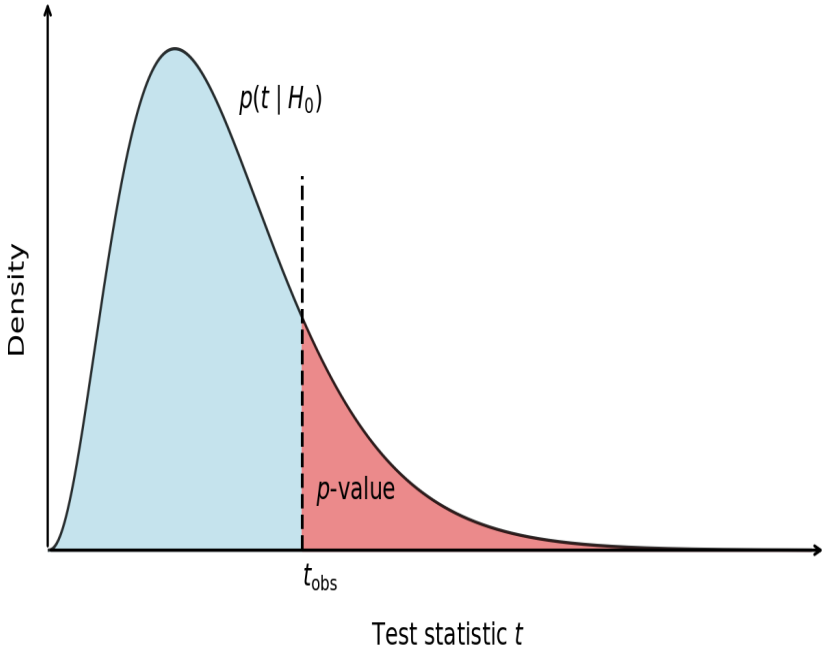
Generating Null Distribution

To compute the p-value associated with *t_obs*, we must estimate $p(t \mid H_0)$, where the null hypothesis is *p_b* = *p_data*

Strategy:

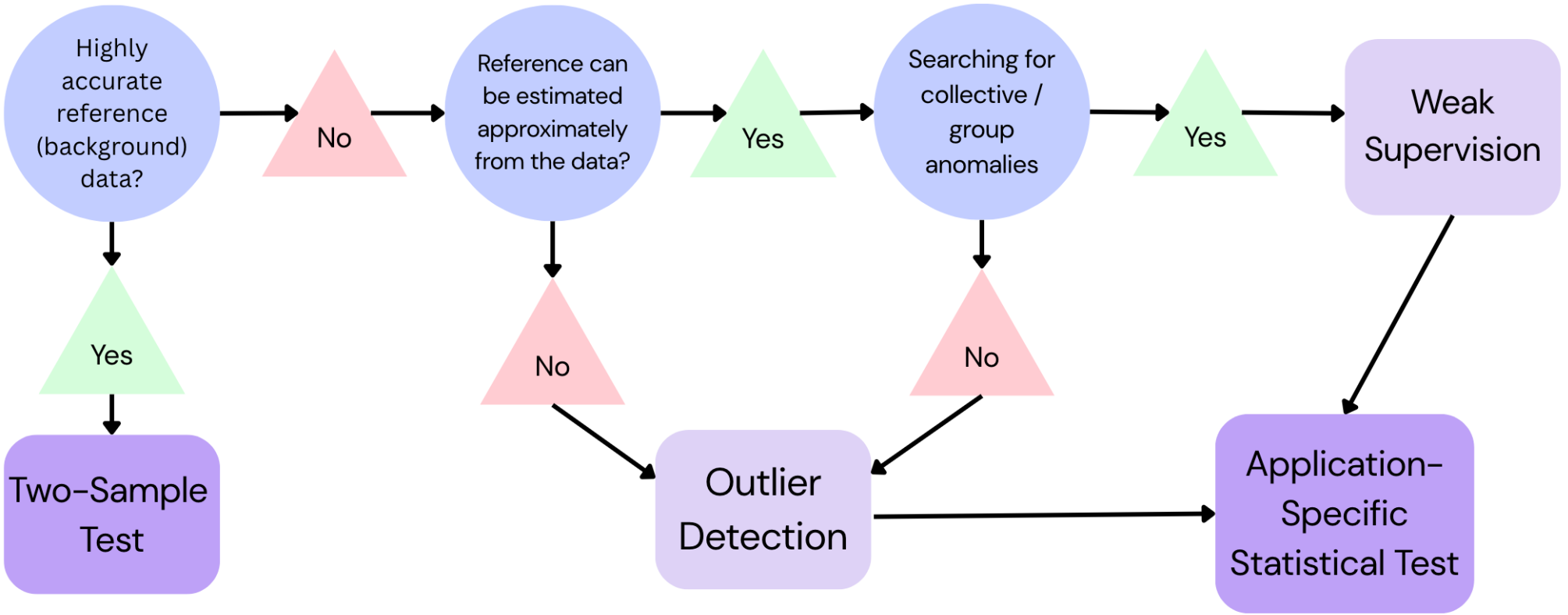
- Generate toy datasets (bootstrap samples) from *p_b*
- Alternatively, permute class labels between *p_b* and *p_data* (permutation test)
- Repeat full training + evaluation
- Build empirical distribution of *t*

$$p\text{-value} = P(t \geq t_{\text{obs}} \mid H_0)$$



Flowchart of Anomaly Detection Methods

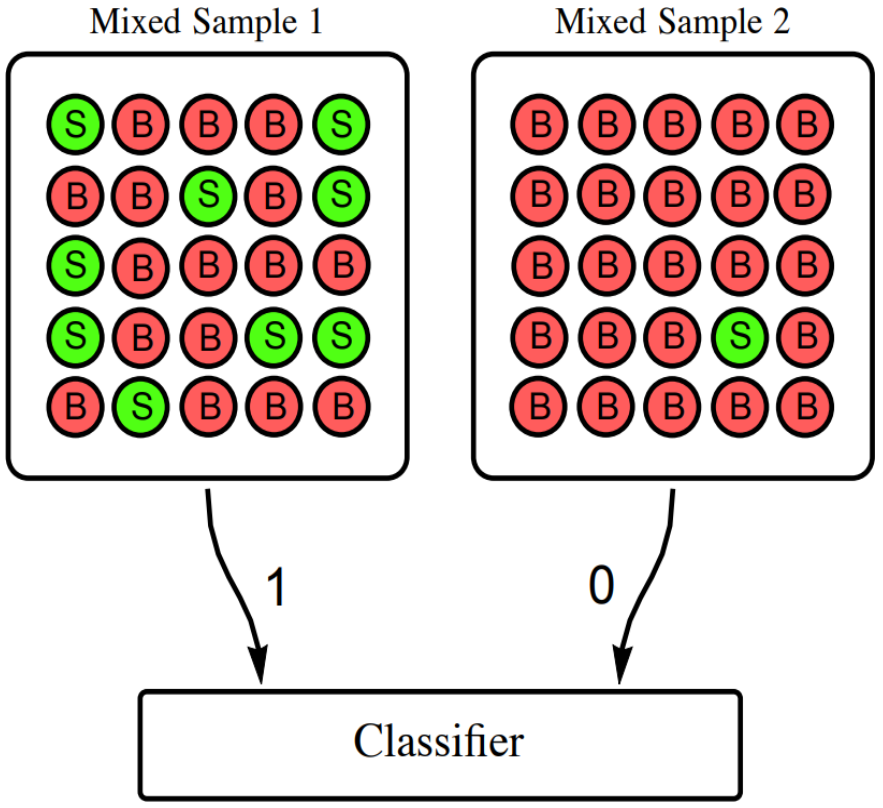
(Personal opinion)



Methods for model-agnostic signal selection

- While two-sample testing does signal localization and hypothesis testing within the same procedure, some other methods only seek to select a subset of the data that is enriched in signal
- Some additional assumption is then used to perform a full statistical test (eg bump hunt)
- Or data events are merely saved / flagged for further human analysis (e.g., trigger, alert system, …)
- The two main classes of methods here are **outlier detection** and **weak supervision**

Weak supervision

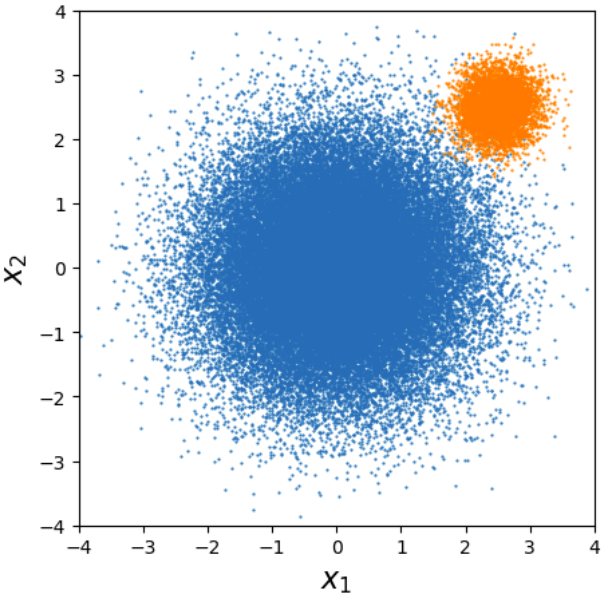


(Metodiev et al. 2017)

- Train a classifier between data (sig + bkg) and approximated reference (~ bkg) → Learns to identify signal
  - Same logic as NPLM (learn likelihood ratio)
- Requires collective anomalies and construction of an approximate reference sample
  - e.g., data-driven methods
- Various methods differ in how the two samples are constructed

Outlier detection

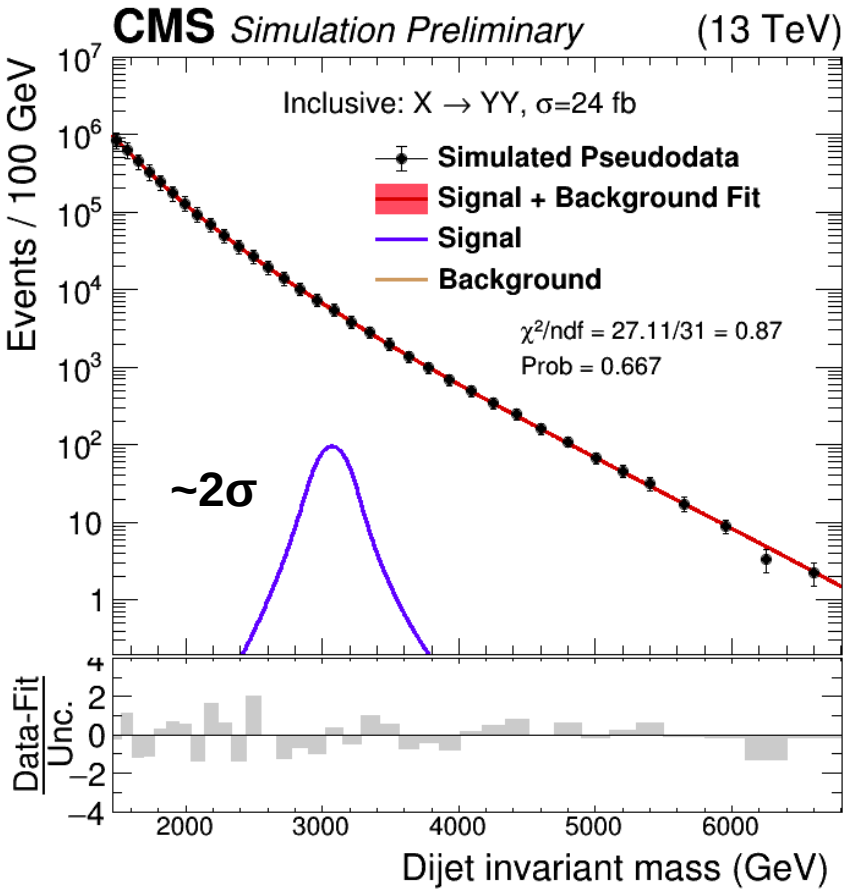
- Train a machine learning model to explicitly or implicitly learn the background distribution $p_b(z)$
  - E.g., autoencoder/VAE (implicit), norm flow (explicit)
- Determine an anomaly score based on how likely events are under the background distribution
  - Autoencoder/VAE: high reco error implies outside training distribution
  - Explicit model: low $p_b(z)$ indicates outlier
- **Challenge: Coordinate invariance**
  - Probability distributions are **not invariant under nonlinear coordinate transformations**
    - Altered by the Jacobian



Lesson: Data representation is a large inductive bias for outlier detection

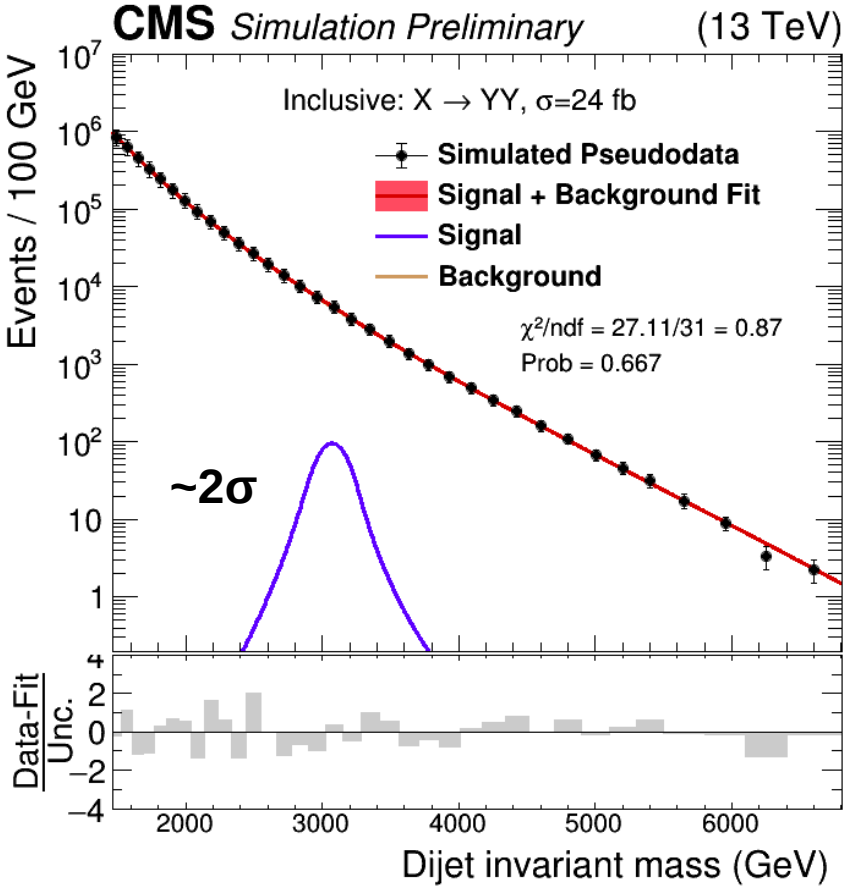
Case study: Resonance Searches

The Bump Hunt

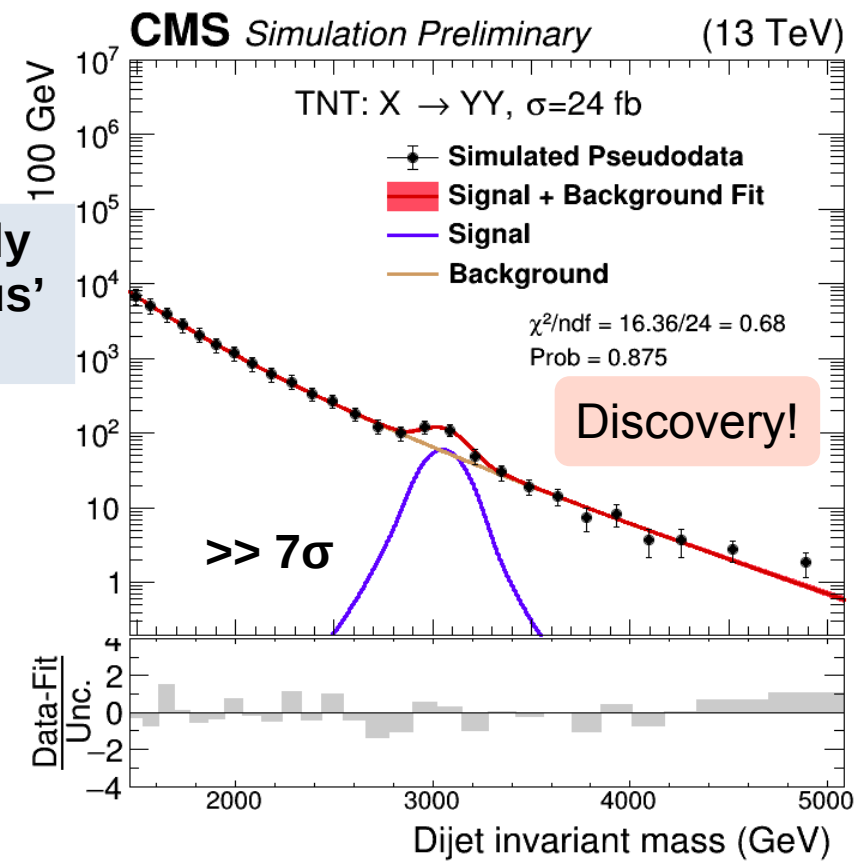


Jet background too high!
Miss discovery

The Bump Hunt



Select only 'anomalous' jets!

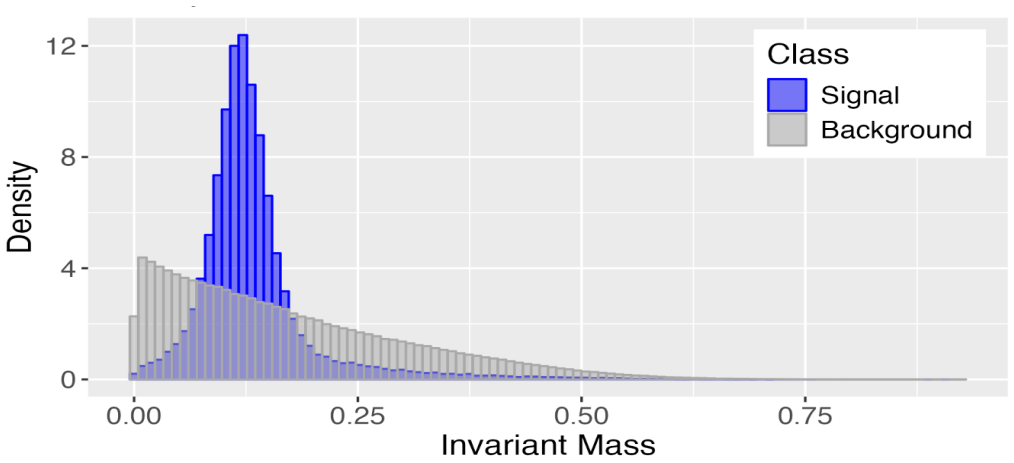


Pitfall: Mass sculpting

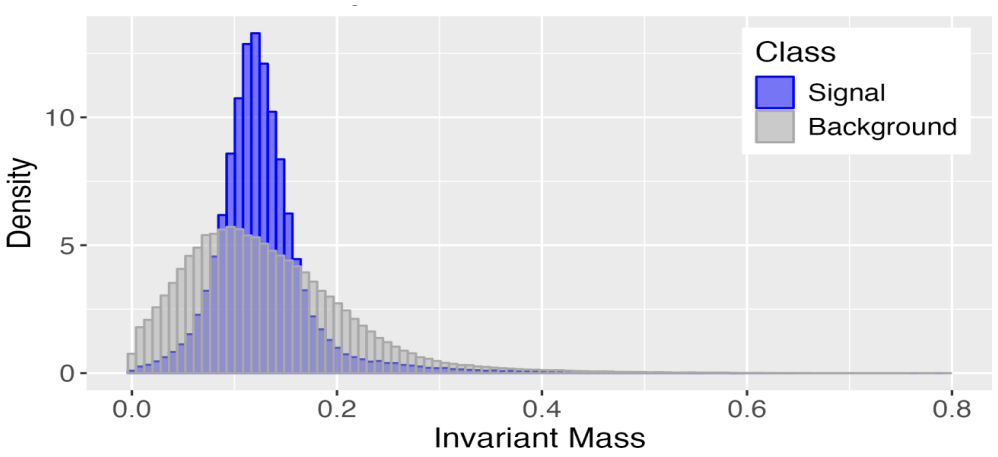
- Key pitfall: Selection of anomalous events must not bias the follow-up statistical test
  - Eg distort the background mass distribution (sculpting)
- Accomplished through various *decorrelation procedures*

  → Choose features that are uncorrelated with mass, custom training strategies, and/or post-training correction procedures

Distribution of Mass



Distribution of Mass after Cut



Mass sculpting distorts the shape of the backgrund distribution (Figure from Chakravarti et al. 2024)

Validation of Null

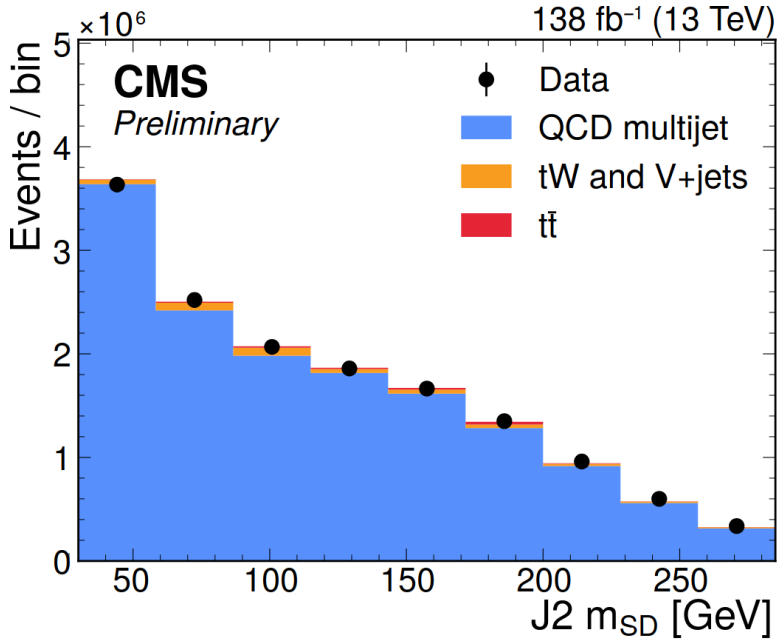
Multiple methods to validate that the test is unbiased
(eg no mass sculpting)

Method	Pros	Cons
Monte Carlo	Full control, interpretation, ease of use	Limited sample size, data-MC mismodeling
Data control region	Potentially larger statistics, no worry of MC mismodeling	Not always easy to identify, potential distribution shift
Synthetic dataset (from gen. model)	Can run many toys, theoretically small domain shift	Potential artifacts from generative model?

Validation of True Positives

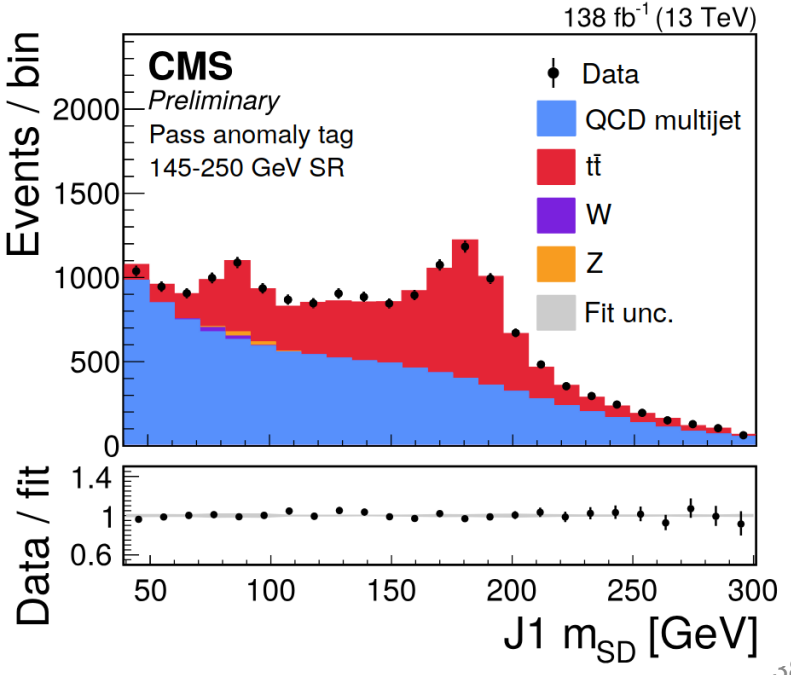
- Can be tested by injecting simulated benchmark signals into control datasets (simulation, data control region, synthetic, …)
- Sometimes, known rare processes can also be used as true anomalies to validate algorithms in data

‘Rediscovering’ high momenta **top quarks** in data with anomaly detection!



Anomaly Detection!

TNT method



Performance evaluation

- Performance of AD needs to be put *in context,* requires comparison methods
- When possible, comparisons against simple / non-ML strategies provide useful context
  - E.g., compare against a bump-hunt on the same dataset without any model-agnostic signal enhancement
- **Model-specific supervised ML methods** provide an upper limit on possible performance
  - Often AD performance is well short of this

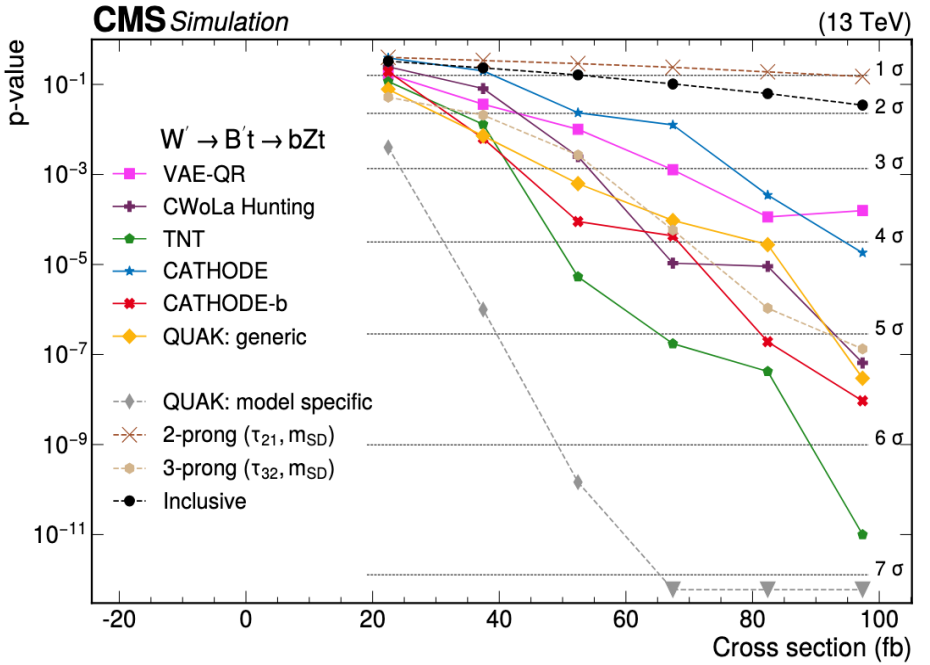
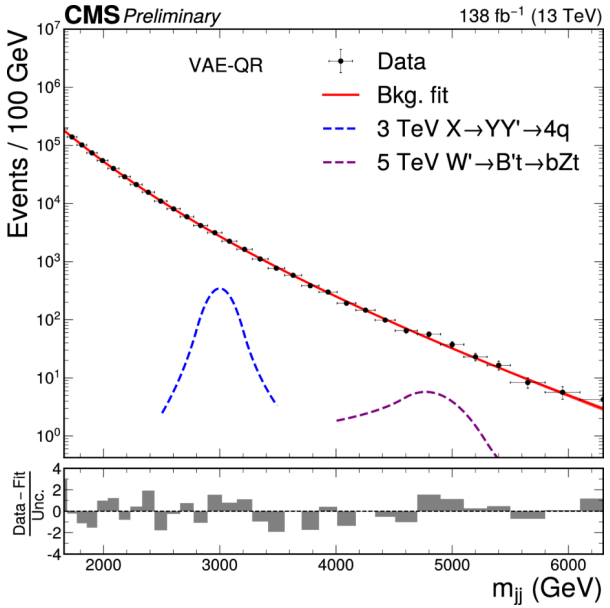
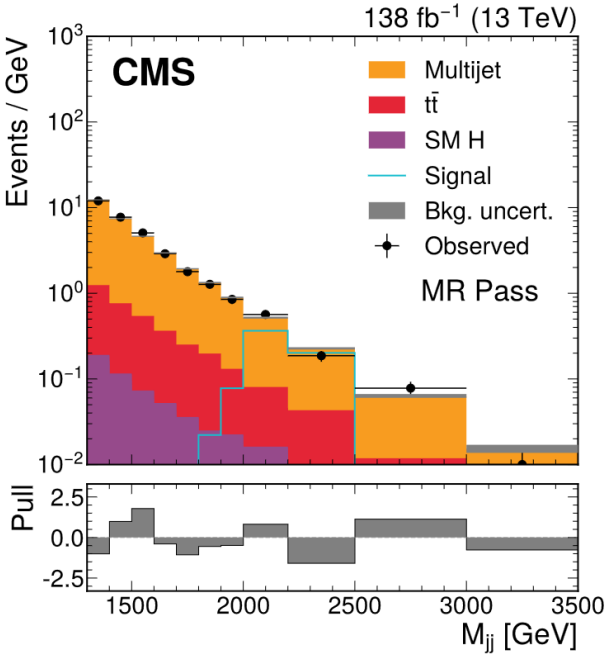


Figure from CMS Search

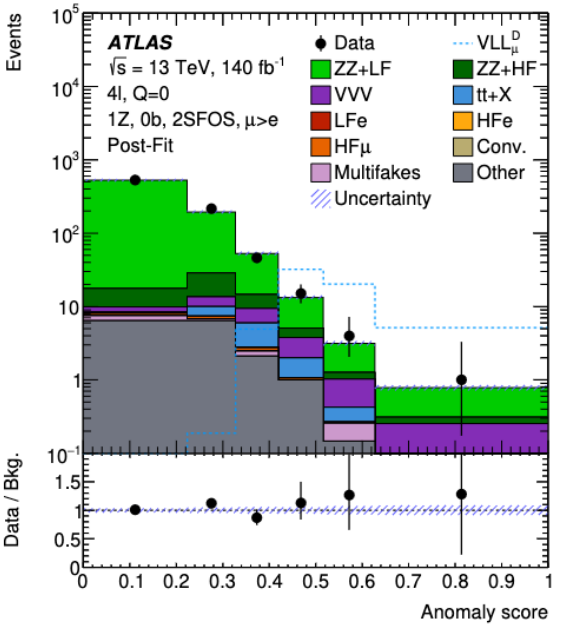
Results on Data!



CMS : Two Anomalous Jets
2412.03747,
2512.20395



CMS :
Resonance to Higgs + Anomaly
2509.13635

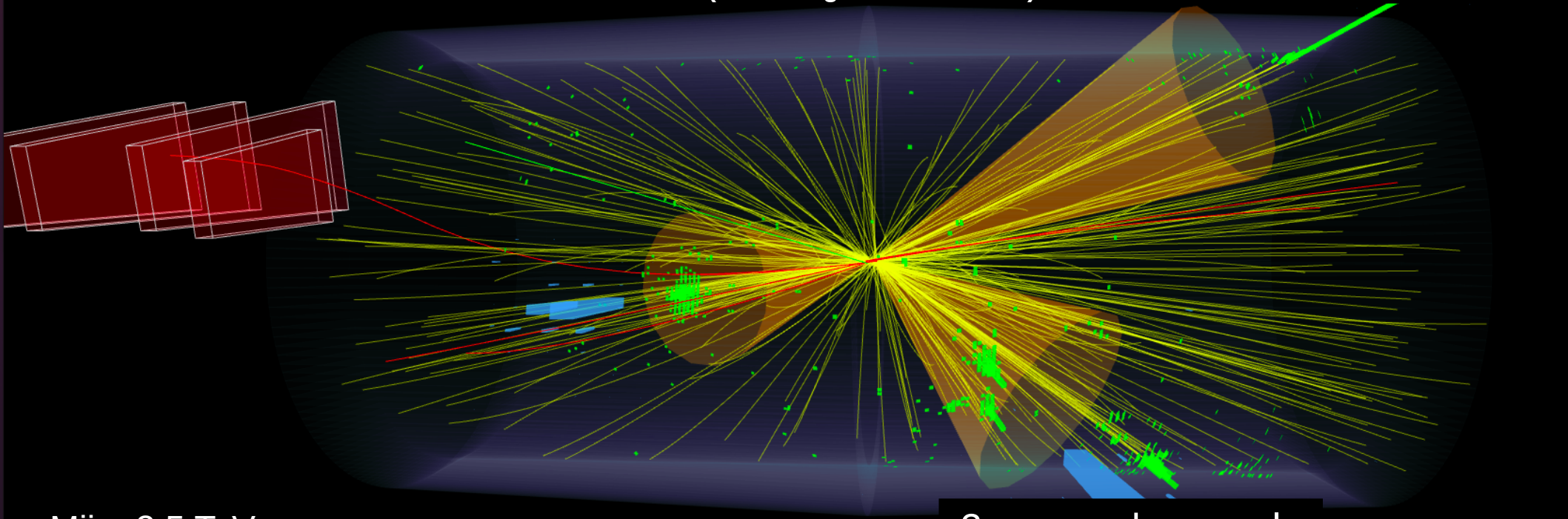


ATLAS : Non-resonant Multi-Lepton Anomaly
2508.19778

+ others from ATLAS ...

One of CMS's most anomalous events!

(according to VAE method)



High energy constituents anomaly

2-pronged anomaly

Mjj = 2.5 TeV
Evt: 851591650
Run: 322332
Era : 2018D

(Not statistical significant)

Search interpretation

Strategies to interpret excess events will be crucial if an anomaly is found

- General methods: Investigate features of highest-confidence anomaly events, interpret ML models, …
- More future work in this area is likely needed

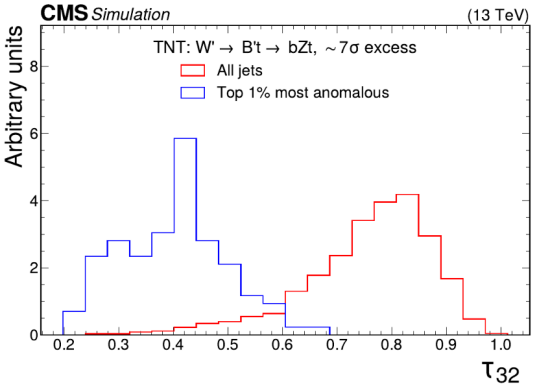
In case of no excess, search interpretation in terms of limits ruling out models of new particles is challenging

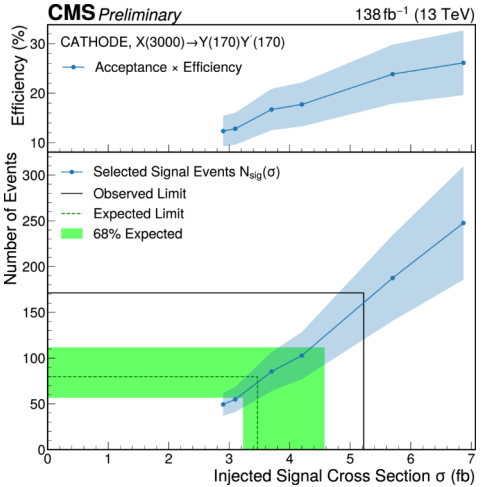
- Computationally costly, requires choice of benchmark signals
- But benefits the broader HEP community to report this

Janssen (2000) showed that it is **not possible to have power for all alternatives**

- Better methods to control & characterize where the high / low power are

Example excess interpretation strategy





Exclusion limits

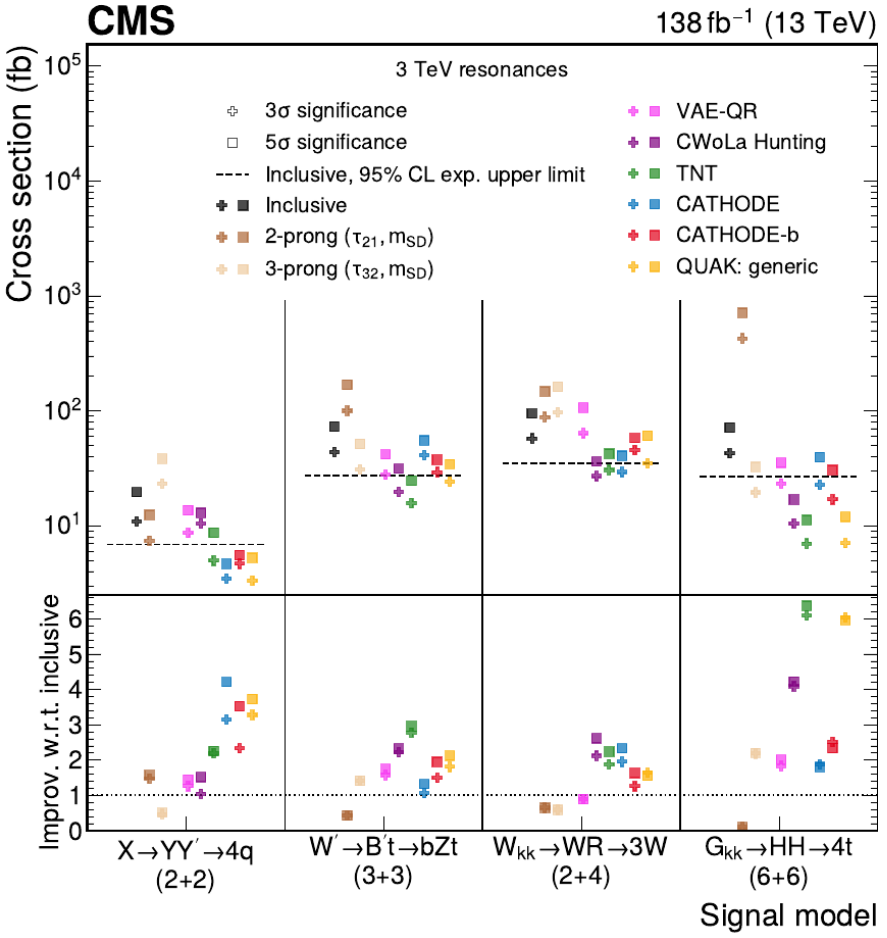
Conclusions

- Model-agnostic methods provide sensitivity for unexpected signals in **complex high dimensional data**
- Methods differ in terms of the strength of their assumptions about p_s and p_b
  - Two-sample testing
  - Outlier detection
  - Weak supervision
- Common challenges:
  - Validation under the null
  - Assessing sensitivity (power) to detect signals and dependence on model / method selection
  - Interpreting model-agnostic search results (both when reject or accept null)

Backup

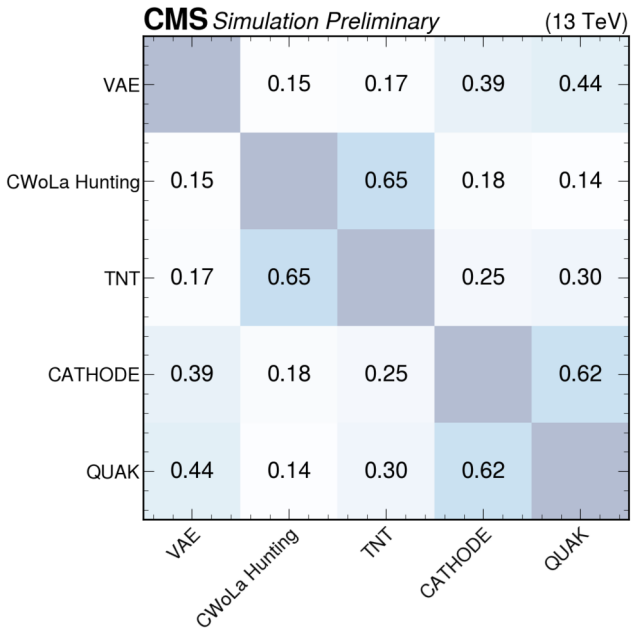
Improved Sensitivity

- "How strong of a signal do I need to get an expected 3σ/5σ excess?"
- **Anomaly detection** improves signal sensitivity by **3-7x**!
  - Ie would need **10-50x** less data for same discovery!

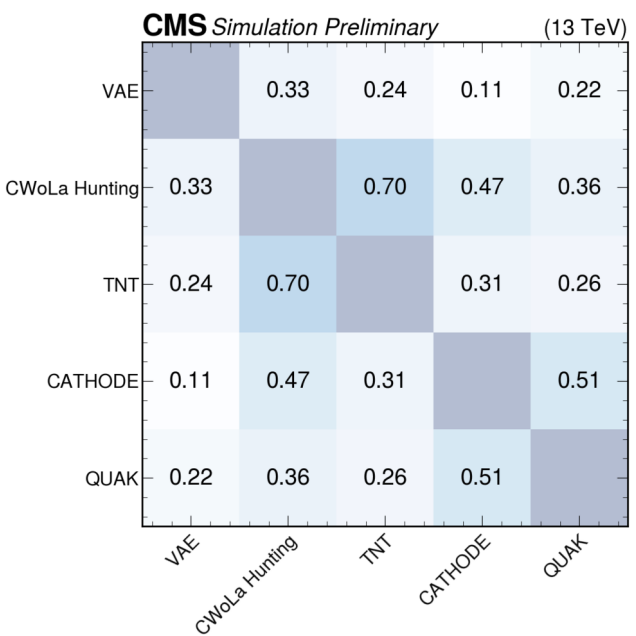


“Are these five methods just learning the same thing?”

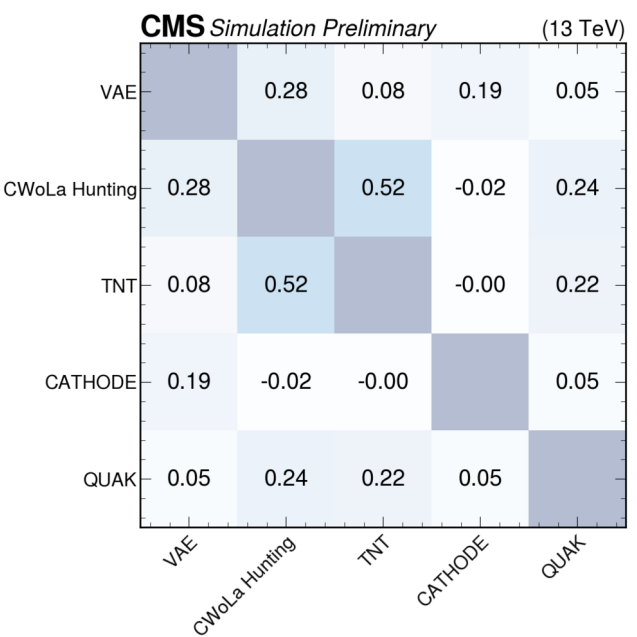
X → YY → qq qq



W’ → B’t → bqq bqq



QCD Bkg.



- Compute **correlation coefficients** between different anomaly scores
- Relatively **low** correlations → methods are

References

CMS Collaboration, Model-agnostic search for dijet resonances with anomalous jet substructure in proton-proton collisions at sqrt(s)=13 TeV, arXiv:2412.03747.
CMS Collaboration, Machine-learning techniques for model-independent searches in dijet final states. ArXiv:2512.20395
CMS Collaboration Search for resonances decaying to an anomalous jet and a Higgs boson in proton-proton collisions at \sqrt{s} = 13 TeV. ArXiv:2509.13635

ATLAS Collaboration, Dijet resonance search with weak supervision using sqrt(s)=13 TeV pp collisions in the ATLAS detector, Phys. Rev. Lett. 125 (2020) 131801 [arXiv:2005.02983].
ATLAS Collaboration, Anomaly detection search for new resonances decaying into a Higgs boson and a generic new particle X in hadronic final states, Phys. Rev. D 108 (2023) 052009 [arXiv:2306.03637].
ATLAS Collaboration, Search for new phenomena in two-body invariant mass distributions using unsupervised machine learning for anomaly detection, JHEP 04 (2024) 122 [arXiv:2307.01612].
ATLAS Collaboration, Search for resonant pair production of Higgs bosons in the $b\bar{b}b\bar{b}$ final state using pp collisions at sqrt(s)=13 TeV with the ATLAS detector, Phys. Rev. D 105 (2022) 092002 [arXiv:2202.07288].
ATLAS Collaboration, Search for non-resonant production of semi-visible jets using Run 2 data in the ATLAS detector, Phys. Lett. B 848 (2024) 138324 [arXiv:2305.18037].
ATLAS Collaboration Search for Beyond the Standard Model physics with anomaly detection in multilepton final states in pp collisions at 13 TeV with the ATLAS detector arXiv:2508.19778

V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: A survey, ACM Comput. Surv. 41 (2009) 1–58.
B. M. Chakravarti, M. Kuusela, J. Lee, Model-independent searches of new physics in DARWIN with a semi-supervised deep learning pipeline, PRD 108 (2023) 092013.
B. M. Chakravarti, A. de Oliveira Souza, M. Kuusela, Robust two-sample anomaly detection at the LHC, arXiv:2412.16776.
R. T. D'Agnolo, A. Wulzer, Learning new physics from a machine, Phys. Rev. D 99 (2019) 015014.
R. T. D'Agnolo, G. Grosso, M. Pierini, A. Wulzer, M. Zanetti, Learning multivariate new physics, Eur. Phys. J. C 81 (2021) 89.
R. T. D'Agnolo, G. Grosso, M. Pierini, A. Wulzer, M. Zanetti, Learning new physics from an imperfect machine, Eur. Phys. J. C 82 (2022) 275.
G. Grosso, M. Letizia, M. Pierini, A. Wulzer, Goodness of fit by Neyman-Pearson testing, SciPost Phys. 16 (2024) 123.
G. Grosso, M. Letizia, Multiple testing for signal-agnostic searches of new physics with machine learning, Eur. Phys. J. C 85 (2025) 19.
A. Hallin, J. Isaacson, G. Kasieczka, et al., CATHODE: Classifying anomalies through outer density estimation, Phys. Rev. D 106 (2022) 055006.
A. Janssen, Global power functions of goodness of fit tests, Ann. Statist. 28 (2000) 239–253.
A. Janssen, Testing nonparametric statistical functionals with applications to rank tests, J. Statist. Plann. Inference 81 (2008) 71–93.
I. Kim, A. B. Ramdas, A. Singh, L. Wasserman, Classification accuracy as a proxy for two-sample testing, Ann. Statist. 49 (2021) 411–434.
M. Kim, K. Cranmer, Global and local two-sample tests via regression, Electron. J. Statist. 13 (2019) 5253–5305.
M. Letizia, G. Losapio, M. Rando, et al., Learning new physics efficiently with nonparametric methods, Eur. Phys. J. C 82 (2022) 879.
E. M. Metodiev, B. Nachman, J. Thaler, Classification without labels: Learning from mixed samples in high-energy physics, JHEP 10 (2017) 174.

Benchmark signals

- Despite the model-agnostic nature of these methods, the use of benchmark signals is necessary for performance validation / demonstration
- The set of benchmarks should be chosen to cover a diverse range of signatures to demonstrate broad coverage
- Overly optimizing methods on benchmarks (choice of features, hyperparameters, etc.) risks an inaccurate assessment of performance
- **Recommendation:** A set of holdout signals, only tested once the search is unblinded, would demonstrate an unbiased assessment of performance (cf. LHC Olympics, Dark Machines)

Open Research: Model Selection

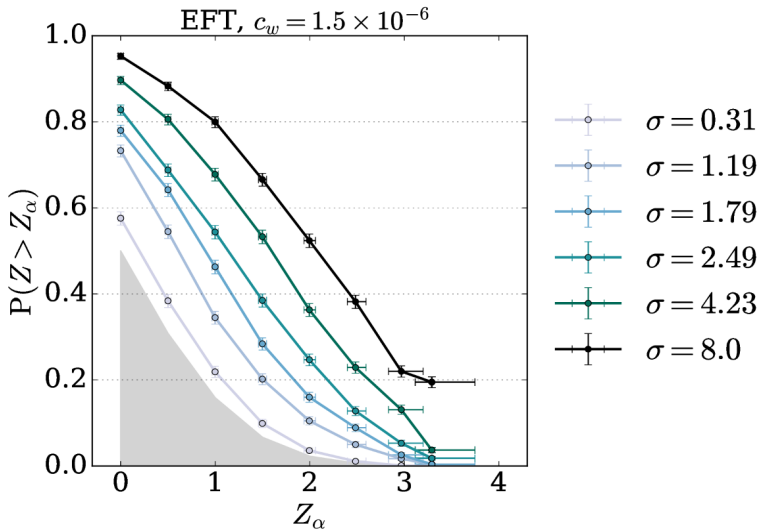
How do you parameterize the ML model?

Choice **introduces inductive biases and impacts sensitivity to signals**

→ Must be controlled for a signal-agnostic test

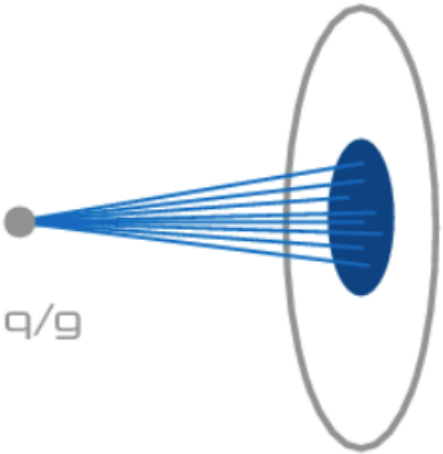
Example: in kernel-NPLM, density ratio is modeled as a sum of Gaussians

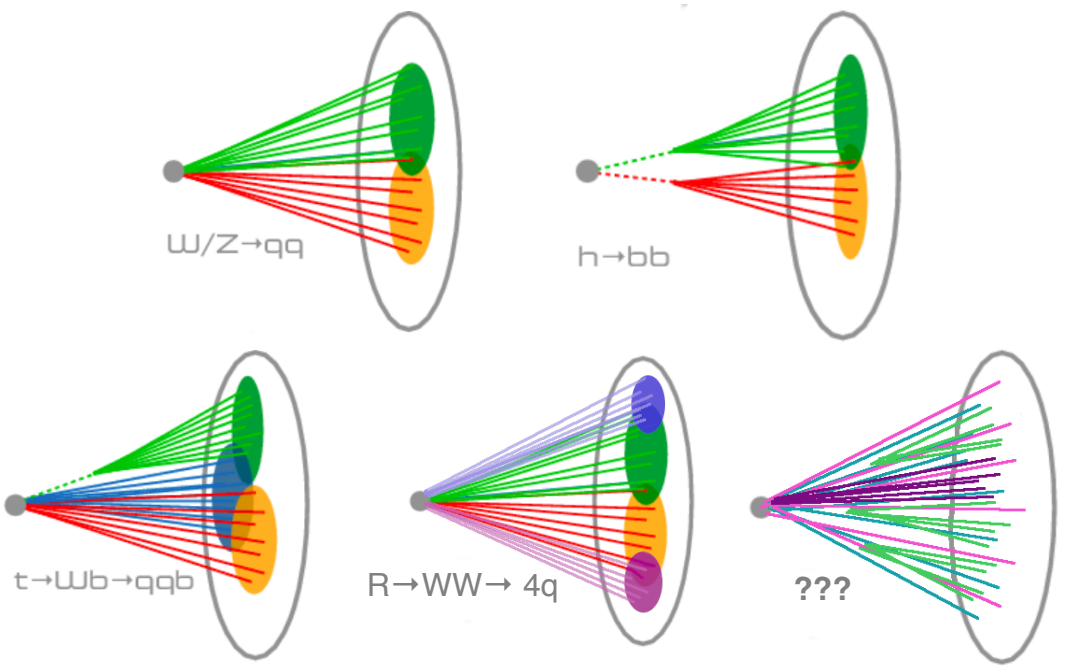
$$f_w(z) = \sum_i w_i \exp\left[-\frac{(z-z')^2}{2\sigma^2}\right]$$



→ *Multiple testing* is a possible solution: combine tests with different hyperparameters in ways that are robust against the LEE (Grosso & Letizia 2025)

Don’t Judge a Jet by its Cover





Typical jet

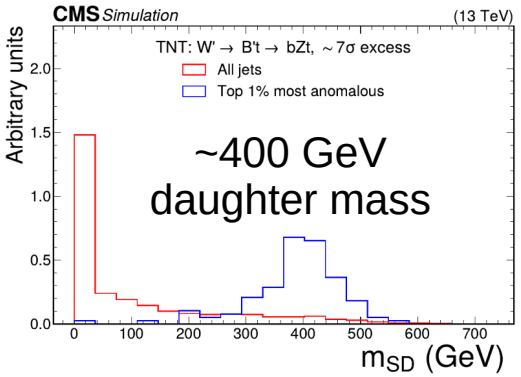
- One central axis (prong)
- From collision point (primary vertex)
- ...

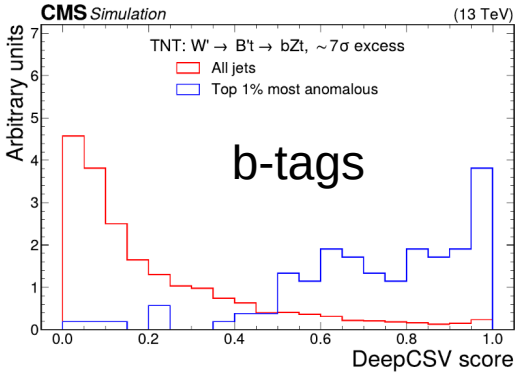
Anomalous jets

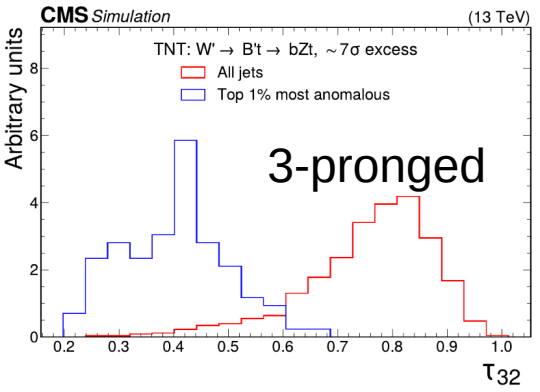
- Multiple prongs
- Displaced from collision point
- ???

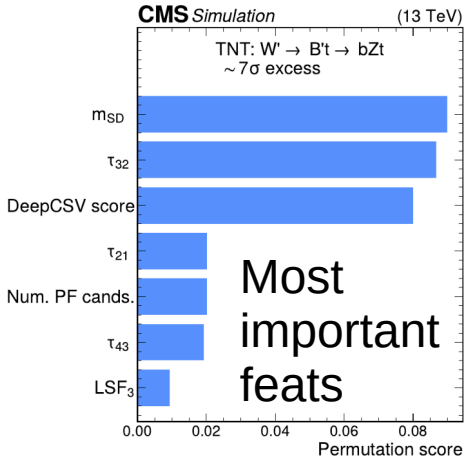
“What if you see an excess?”

Investigate features of most anomalous events!









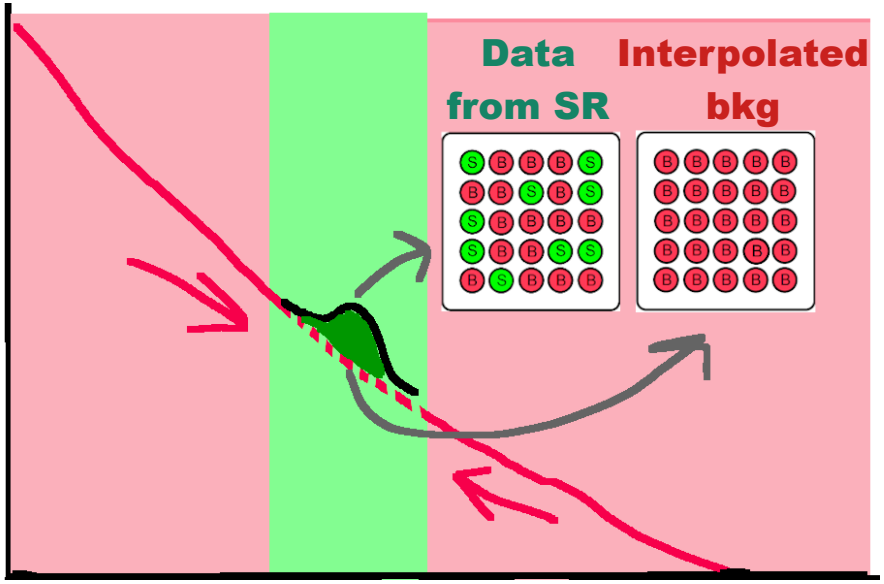
✔ Matches characteristics of injected signal

W’ → B’t, B’ → bZ

$M_{B'}$ = 400 GeV

CATHODE
Genertative AI to **interpolate** bkg into SR to construct sample

[Hallin et al 2109.00546]

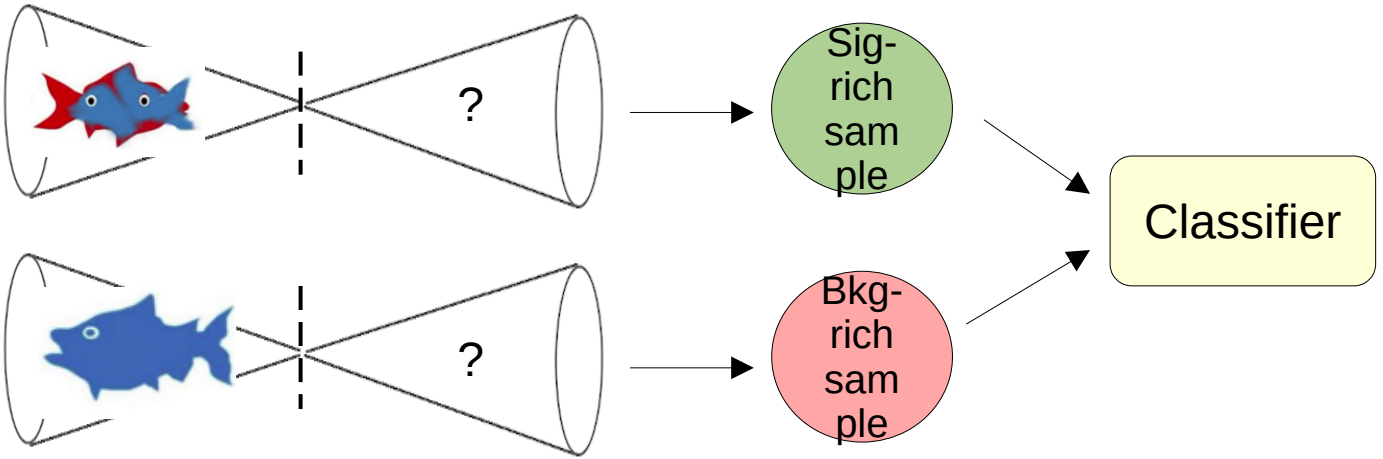


Can also be used for AI-based meas. of **Higgs self-coupling!**

OA & Szewc
JHEP 11 (2025) 129

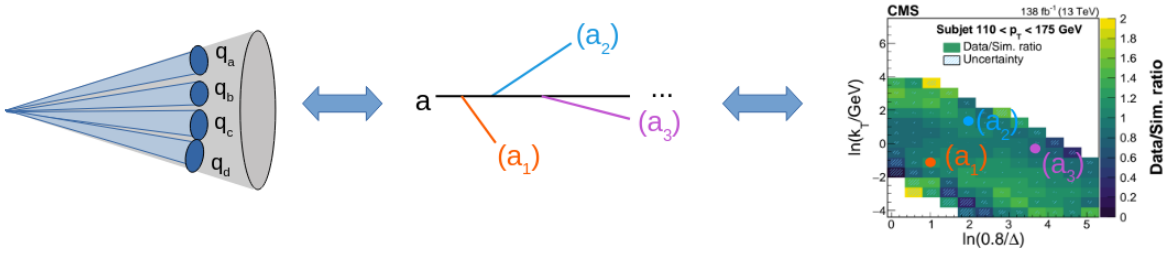
Tag N' Train
Looks for pairs of anomalies, purifies samples

OA & Suarez
JHEP 01 (2021) 153



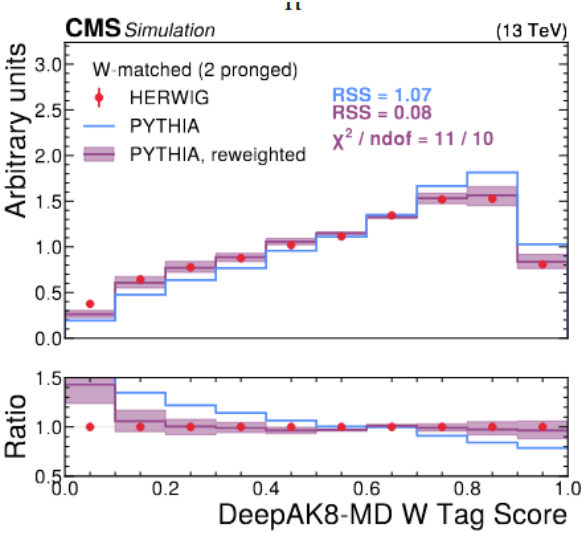
JHEP 11 (2025) 038

Uncertainties



$$W_{\text{jet}} = \prod_{\text{subjets}} W_{\text{subjet}} = \prod_{\text{splittings}} LPR(\text{splitting})$$

Use physics domain knowledge to factorize problem

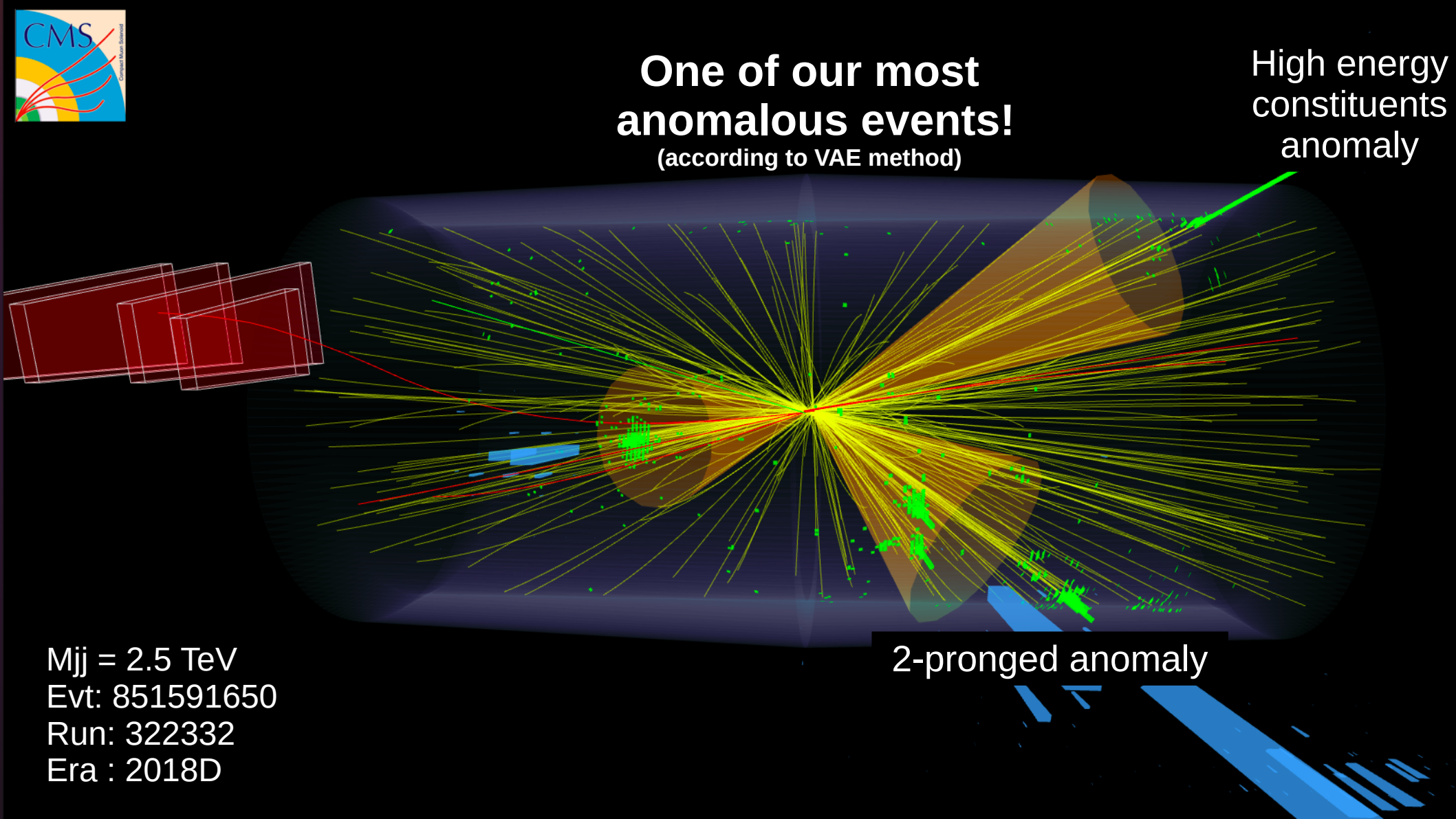


Account for data/sim domain shift

Method now standard within CMS, employed by multiple (5+) analyses!

One of our most anomalous events!

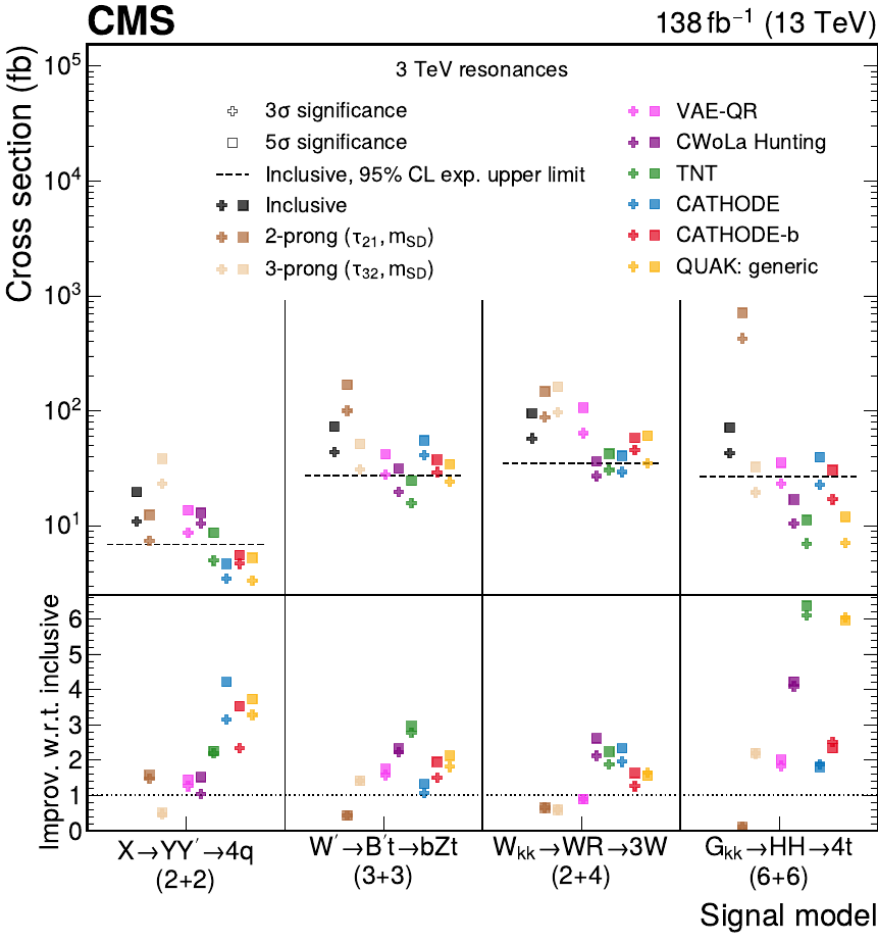
(according to VAE method)



Mjj = 2.5 TeV
Evt: 851591650
Run: 322332
Era : 2018D

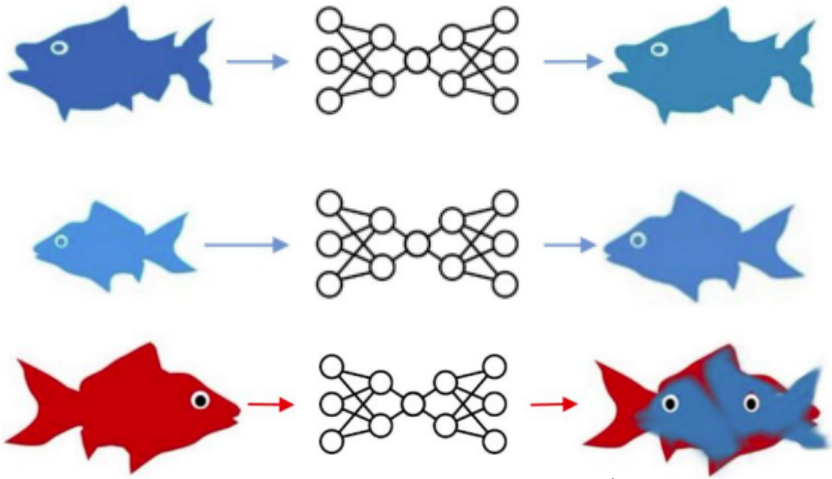
Improved Sensitivity

- "How strong of a signal do I need to get an expected 3σ/5σ excess?"
- **Anomaly detection** improves signal sensitivity by **3-7x**!
  - Ie would need **10-50x** less data for same discovery!



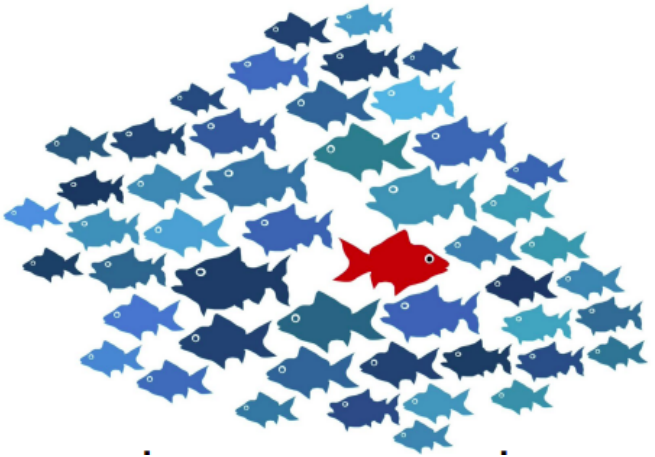
Looking for Outliers

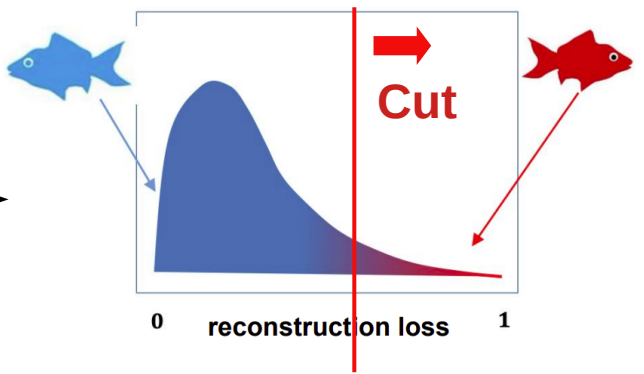
Apply Autoencoder



Take difference

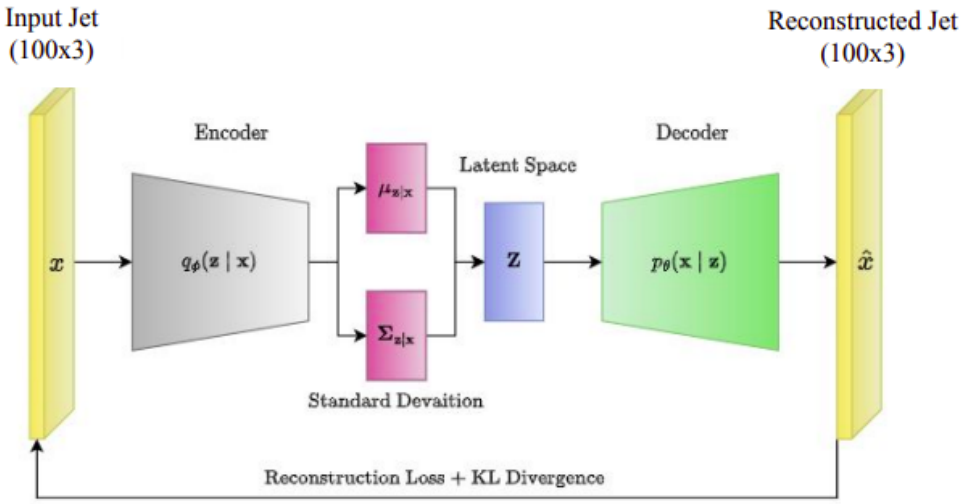
Data from signal region





Illustrations: J Gonski, A Kahn

Variational Autoencoder (VAE)

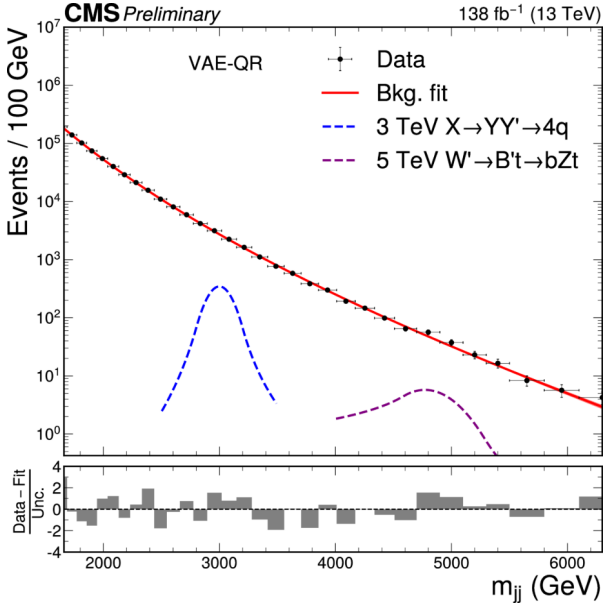


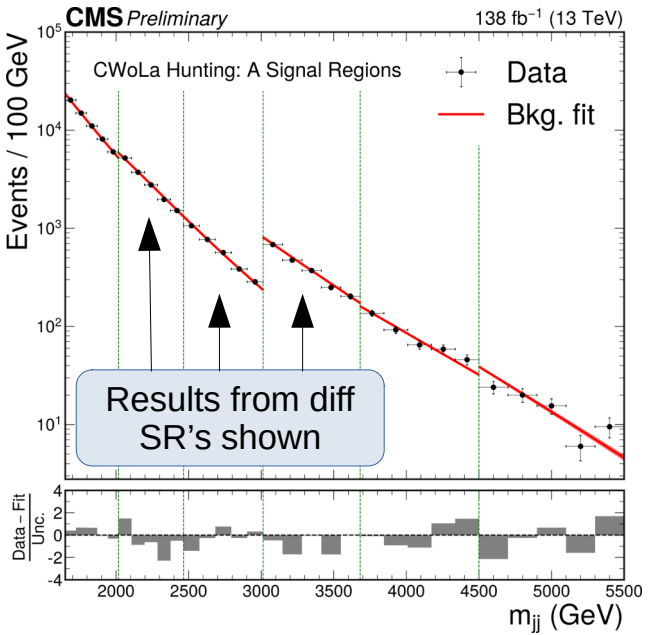
Latent space forced to be Gaussian thru additional term in loss

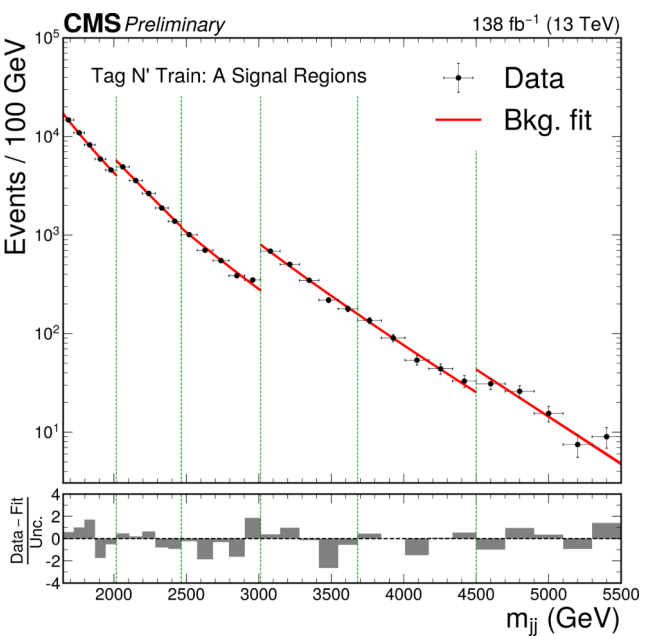
- Jet represented by up to 100 highest p_T constituents (p_x, p_y, p_z)
- 100x3 matrix compressed to latent space of size 12
- Trained on jets from |Δη| sideband
  - Sampled to match SR kin.

Search Results

No significant excesses from any method





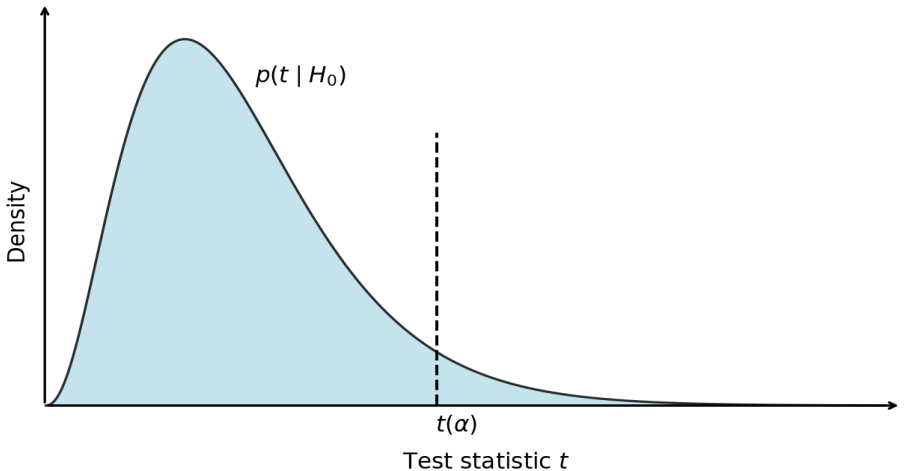


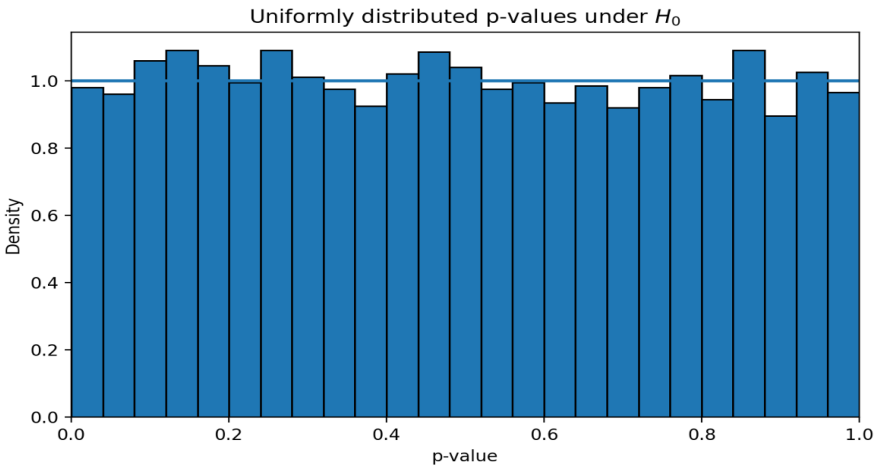
QUAK & CATHODE
results similar

Challenge: validation of null hypothesis

In two-sample testing, the null hypothesis H_0 asserts that the two data-generating distributions are identical → use background toy data to verify that the test behaves as intended when H_0 is true:

- type-I errors (false positives) are at the nominal level α
- p-values are uniformly distributed under H_0





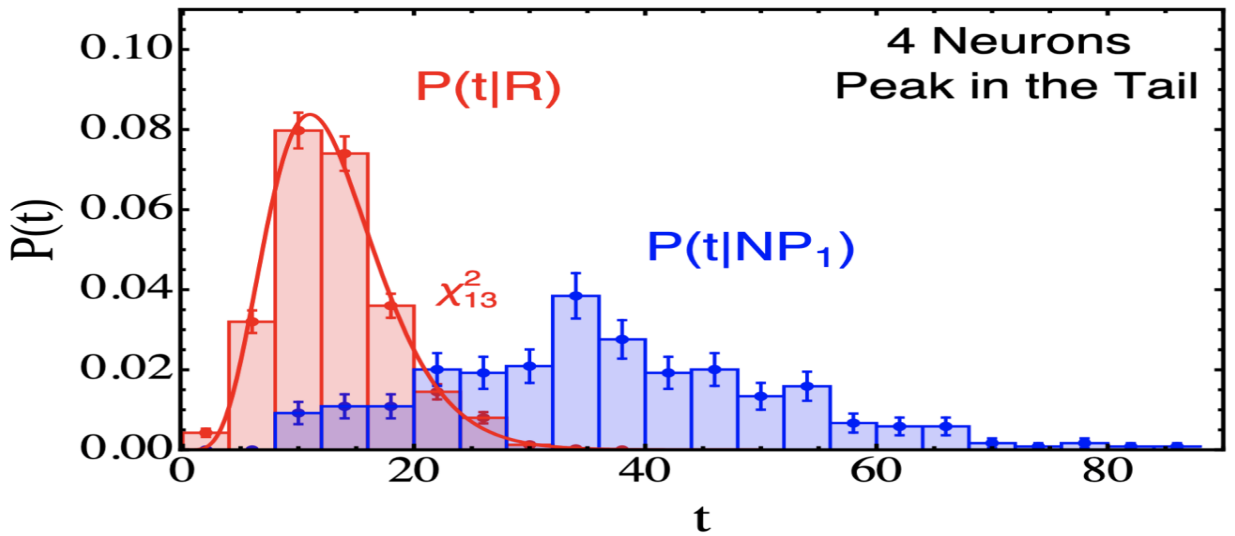
Guaranteed by construction in some cases but still good practice as it helps to identify implementation errors or bugs in the pipeline

Challenge: validation of null hypothesis

NPLM is inspired by the likelihood-ratio test: is the null a χ^2?

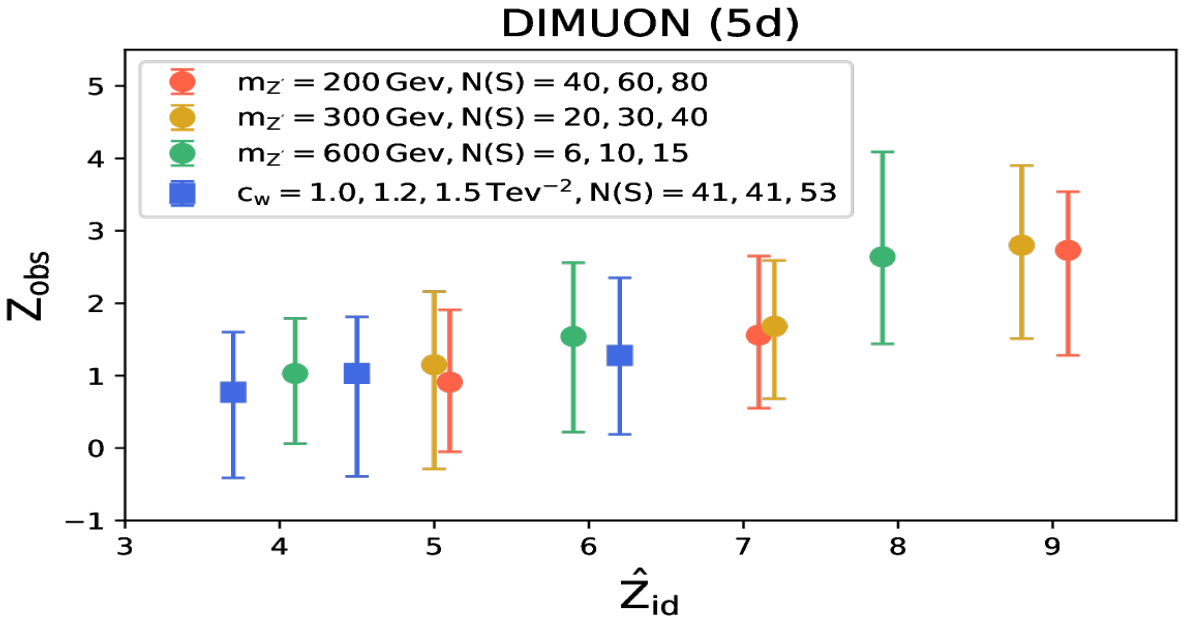
Only *empirical evidence* under careful model selection and regularization

→ Useful to control over/underfitting but p-values are computed empirically with toy data



Challenge: assessment of performance

Establish the sensitivity of the method in controlled benchmark scenarios: compare a signal-agnostic method against a model-specific search



Such comparisons cannot fully characterize performance across all possible new-physics scenarios

But they quantify the trade-off between sensitivity and signal agnosticity

Note: Statistical tests cannot have power to detect all alternative hypotheses (Janssen 2000)

→ Any model-independent search strategy is insensitive to some set of deviations