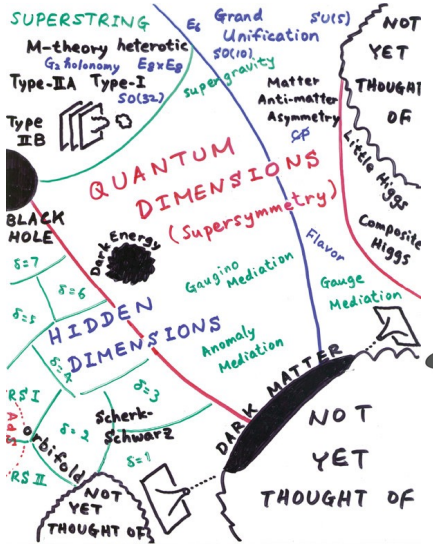
Where's the new Physics?

-or-

How to find something when you don't know what you're looking for

Patrick McCormack

May 3, 2023



H. Murayama

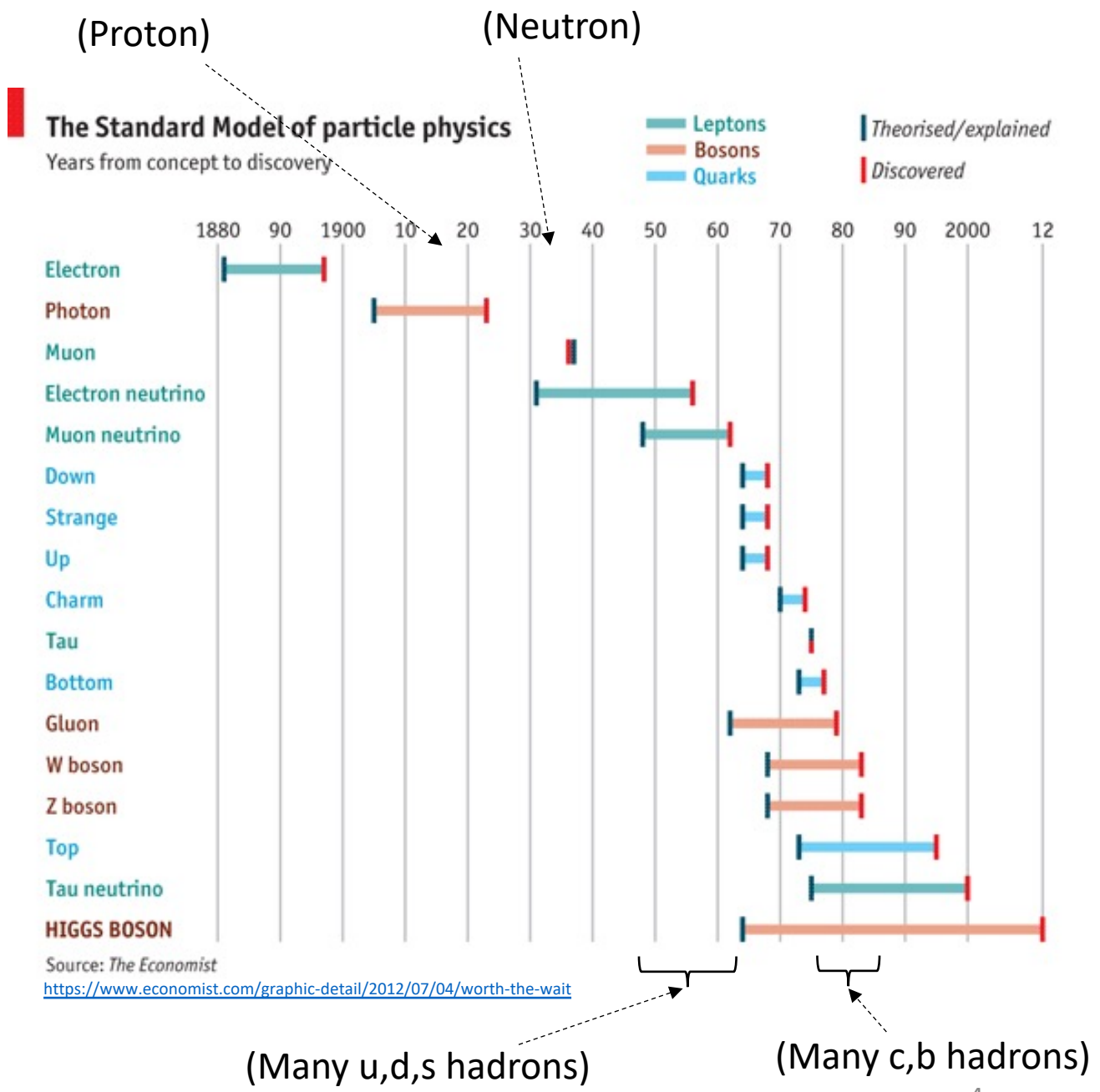
Contents

Contents

Context

- Main goal of particle physics: understand the particle content of the universe

- Step 1: catalogue all particles
  - You'll note on the right that most (fundamental) particles were predicted before being experimentally discovered*

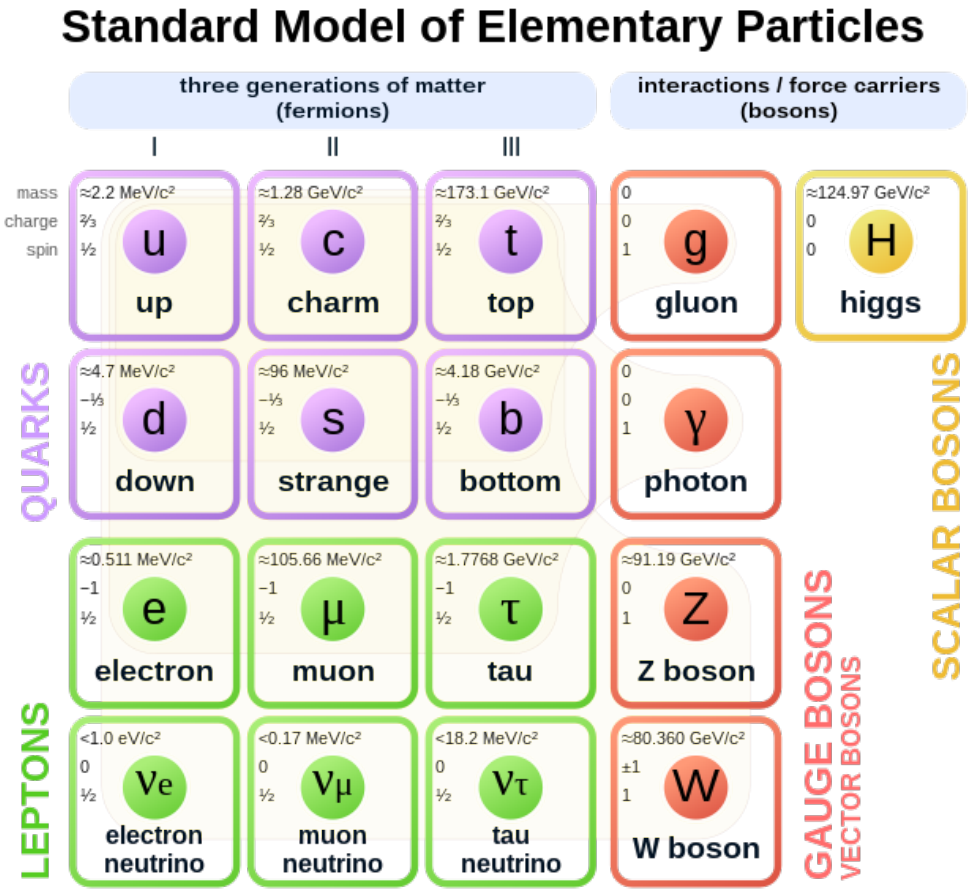


*There's a lot of subtlety left out of this chart – there have many unexpected results in particle physics, like the discovery of the positron and the hadron "particle zoo", which helped develop theory, which then led to more particle predictions
Check out Cahn and Goldhaber's "The Experimental Foundations of Particle Physics"

Context – the Standard Model

- The Standard Model (SM) is a quantum field theory that encapsulates all known particles and their interactions
- The dynamics of free particles are governed by relativistic wave equations, so their propagators* have denominators as below, where m is the mass of the particle in question:

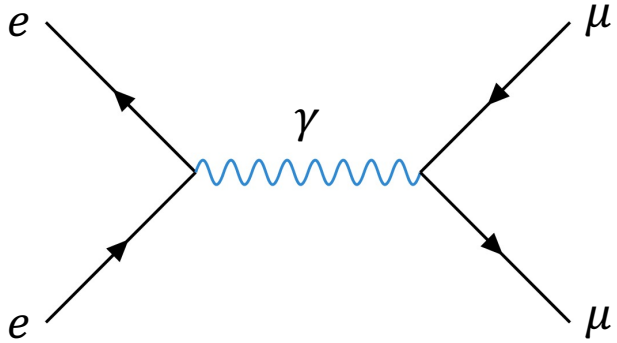
$$\tilde{G}_F(p) = \frac{1}{p^2 - m^2 + i\varepsilon}$$



*A propagator encapsulates the the probability amplitude for a particle to travel from one state to another (in time and space)

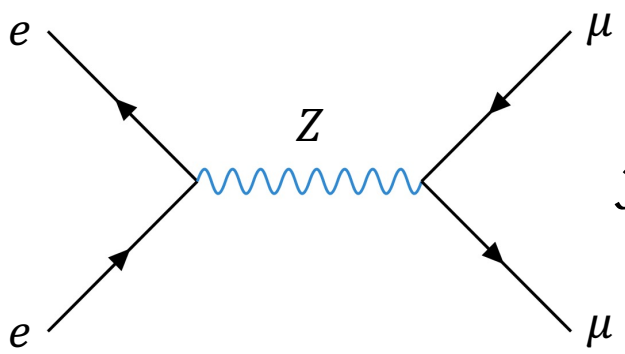
Context – Resonances

- This propagator is relevant when calculating the likelihood of 2->2 processes, such as $e^+e^- \rightarrow \mu^+\mu^-$
  - 2->2 processes are very important experimentally. They're why collider experiments are useful!
- In this scenario, one massless and one massive channel comprise the leading contributions



$$\mathcal{M}_\gamma \propto \frac{e^2}{q^2}$$

Massless propagator (photon)



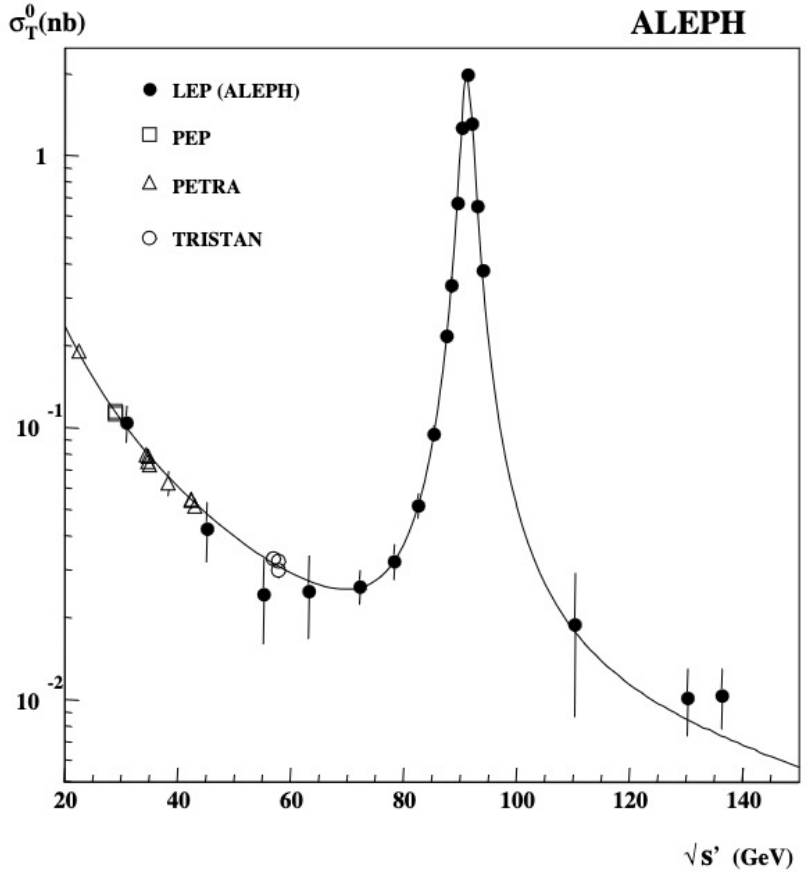
$$\mathcal{M}_Z \propto \frac{g_Z^2}{q^2 - m_Z^2 - i m_Z \Gamma_Z}$$

Massive propagator (Z boson)

Here, $\mathcal{M}$ denotes a "matrix element". You square the matrix element to get the cross section for a process (which is basically the likelihood that the process will occur). Also, q is the transferred momentum.

Context – Resonances

- When a process involves a massless propagator, the cross section goes as $\sigma \propto \frac{1}{s}$
  - $\sqrt{s}$ is the center of mass energy for the two initial state particles
- When a process involves a massive propagator, like the Z boson, the cross section goes as $\sigma \propto \frac{1}{(s-m_Z^2)^2+m_Z^2\Gamma_Z^2}$
  - This is a **resonance** around the propagator's mass
- **Experimental discoveries of new particles often revolve around finding a resonance**



https://cds.cern.ch/record/850586/files/0008cartagena1.pdf

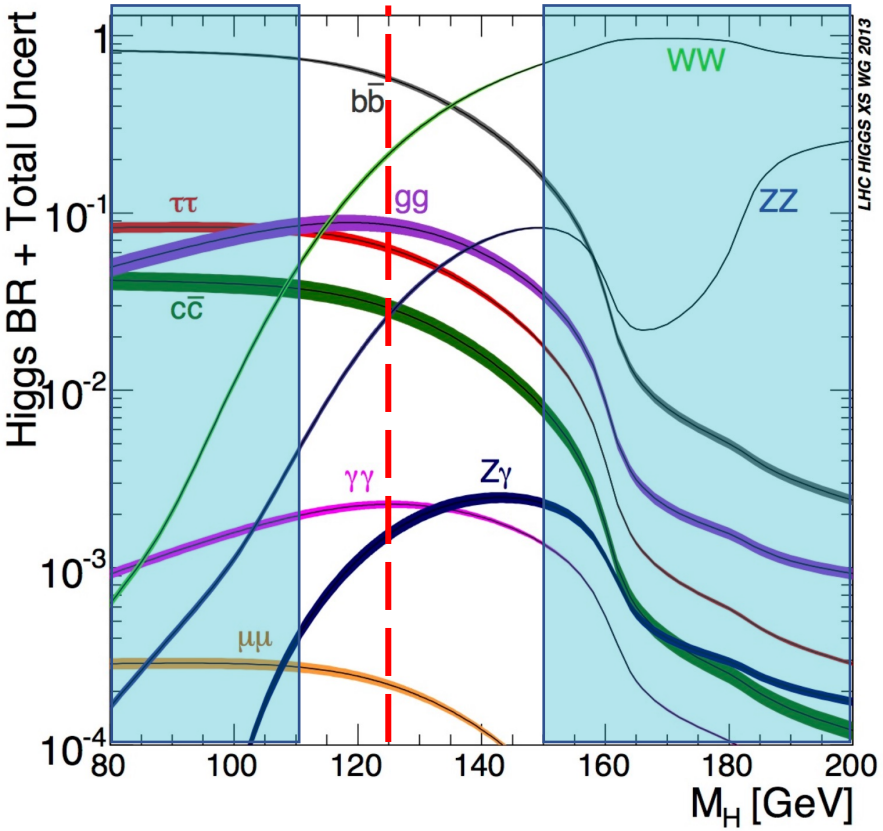
$e^+e^- \rightarrow \mu^+\mu^-$ data from the LEP collider at CERN. The falling spectrum comes from the photon mediator, and the resonance comes from the Z boson

Contents

Looking for something when you know what it is

Example: Higgs discovery

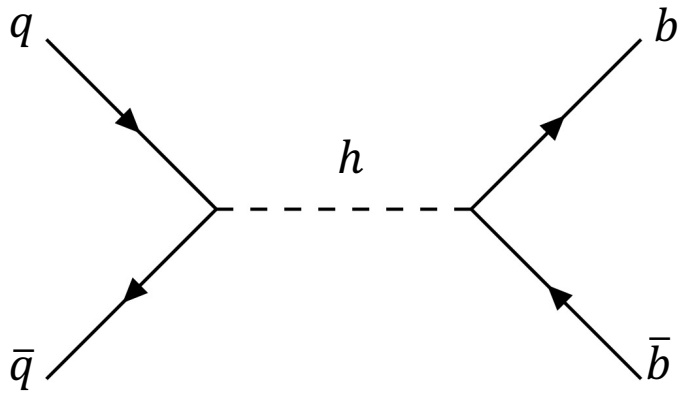
- The Higgs boson was discovered by ATLAS and CMS in 2012
  - Final particle of the SM to be experimentally confirmed
- The SM Higgs and its interactions were well understood before its discovery, other than its mass (and existence)
  - Note: it wasn't necessarily *guaranteed* that the Higgs boson existed
  - (Also, the "simple" discovery of the particle isn't the end of the story – you have to check in detail that it really behaves as expected)

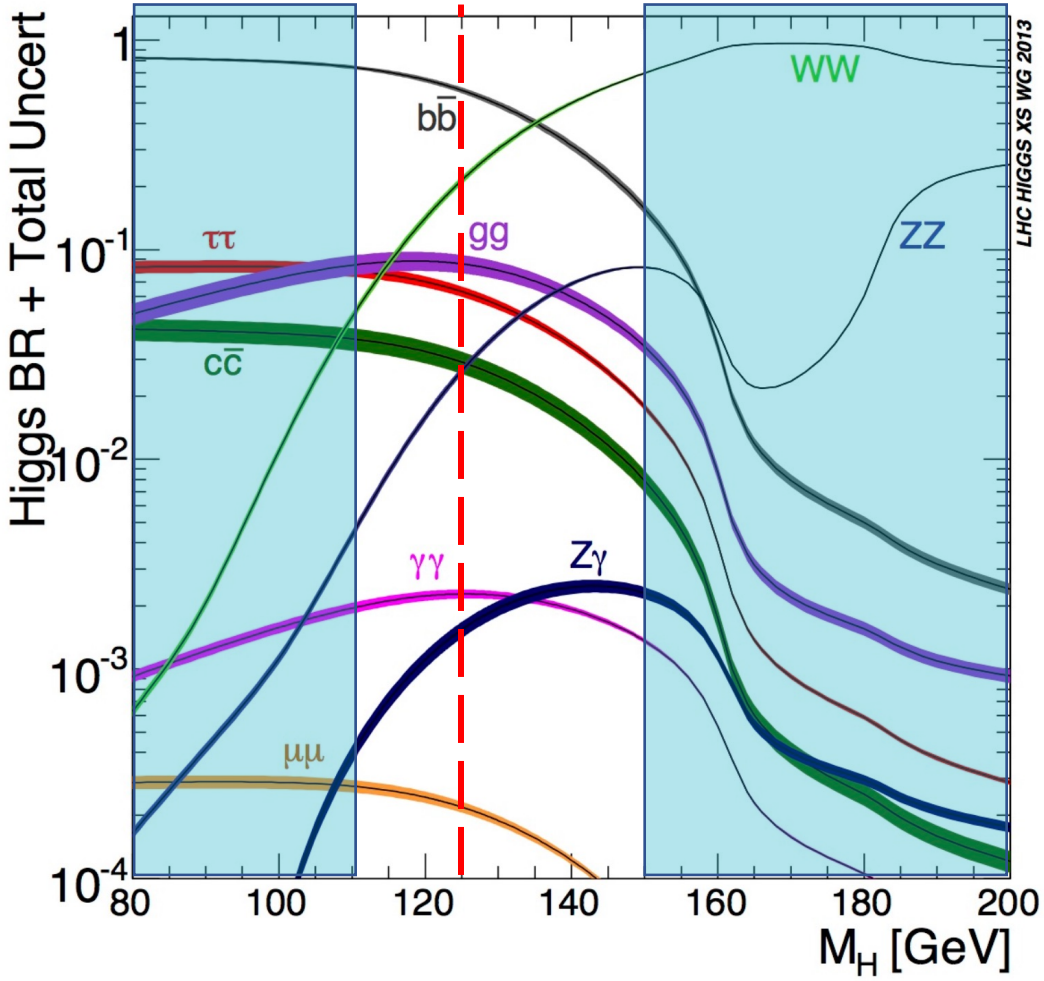


Higgs branching ratios for different Higgs masses. Shaded boxes represent pre-LHC exclusions (mostly from LEP and Tevatron)

Higgs: Designing a search

- Naively, if you looked at the plot on the right, you might think that the easiest way to look for the Higgs is to look for events with two bottom quarks*

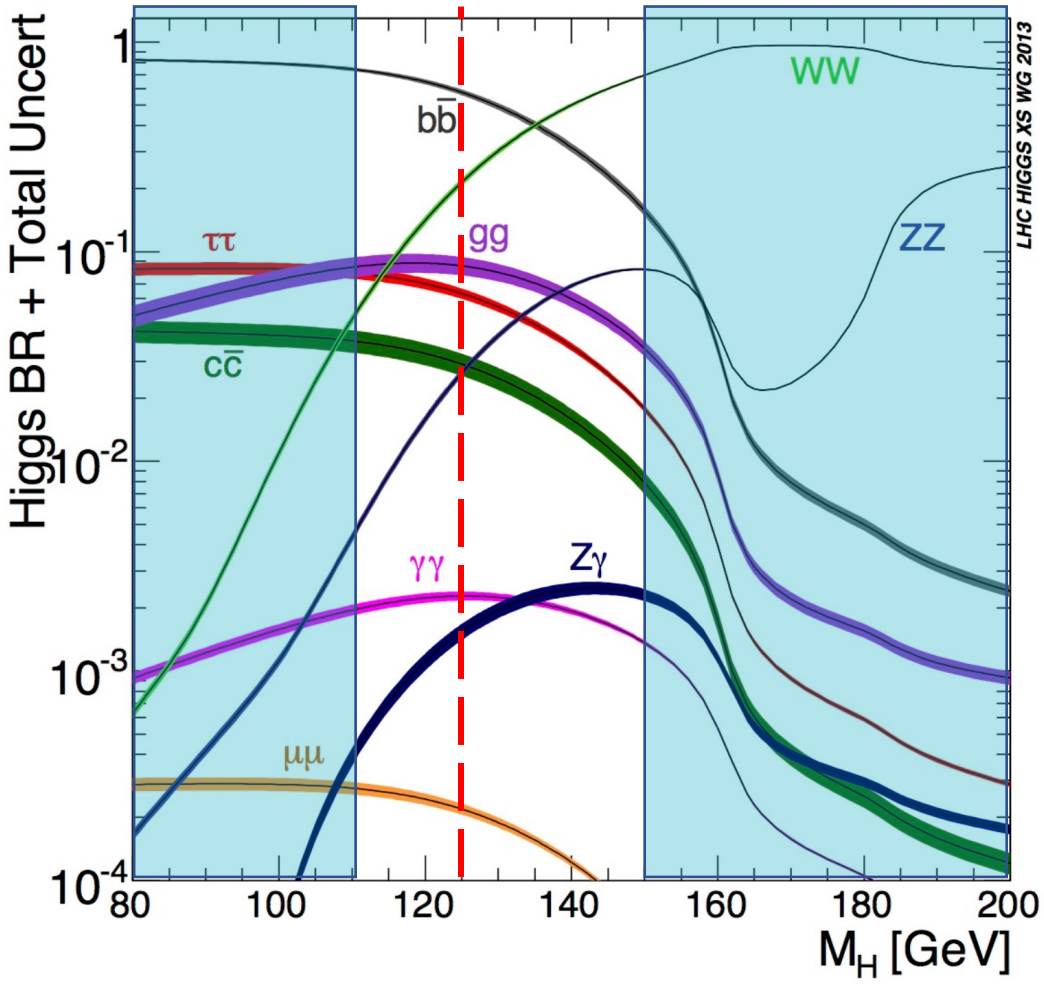




*Experimentally, these final state quarks manifest as “jets”. Final state quarks and gluons “hadronize” into columnated showers of baryons and meson, which are reconstructable experimentally and clustered together to create jets. It is possible to tag jets as coming from a b quark by identifying specific long-lived mesons within the jets, though not all b jets are correctly identified, and some non-b jets are incorrectly labelled as b-jets

Higgs: Designing a search

- Though millions of Higgs bosons *are* created at the LHC, and most of them decay into jets of some sort, the vast majority of LHC events involve the creation of jets
- **Critical point**: there would be **overwhelming backgrounds** if you looked in the naïvely preferred channels
  - The higgs's resonance would be smaller than the statistical error on the bins in your histogram of dijet mass
- Point #2: dijet mass resolution is worse than di-lepton or di-photon mass resolution
  - Poor resolution leads to a smeared-out peak, which is even harder to find

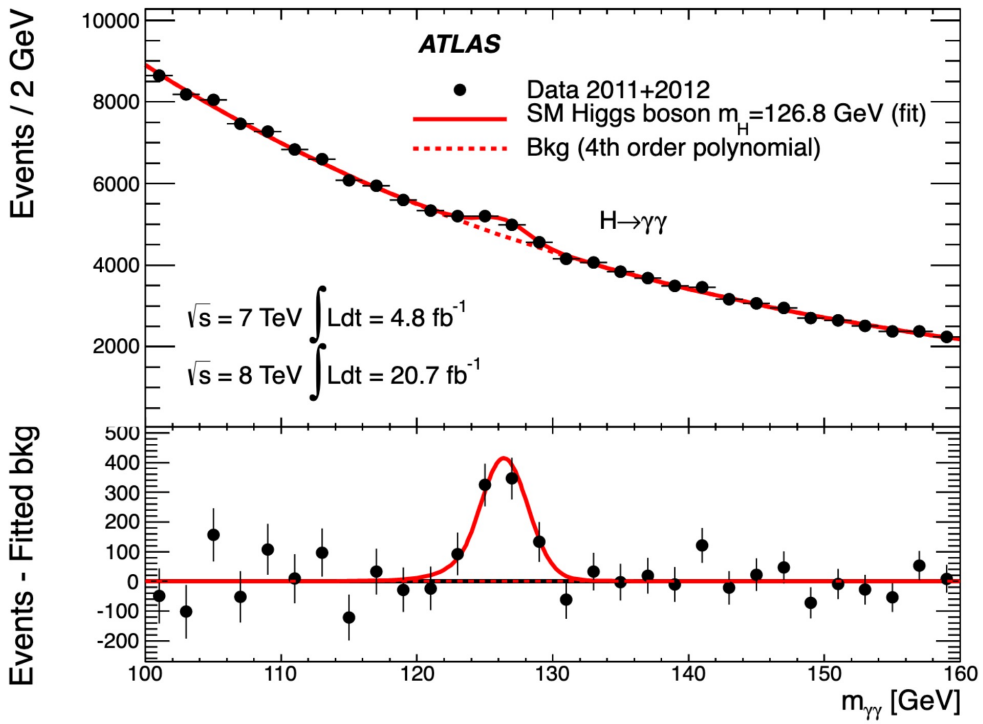


Higgs: Designing a search

*Technically, the search involved a multi-category fit. Different categories targeted different Higgs production mechanisms or different photon kinematics, such that some categories were more signal-pure than others. This approach increases sensitivity relative to an inclusive search.

- To actually find the Higgs, one critical channel was Higgs decays to two photons
- The **basic search idea** here is to **fit** the di-photon invariant **mass spectrum** with parametric functions for signal and background. From this fit, you can **extract the signal rate and significance***
- Events included in the "inclusive" spectrum on the right had to pass several pre-selections, such as:
  - Event passes di-photon trigger
  - Reconstructed photons must pass thresholds on transverse energy
  - Photons must be found in region of detector covered by trigger (and not in region with calorimeter gap)
  - Photons must pass "tight" identification criteria
  - Photons must be "isolated"
  - Etc. (please refer to papers to see full selection)

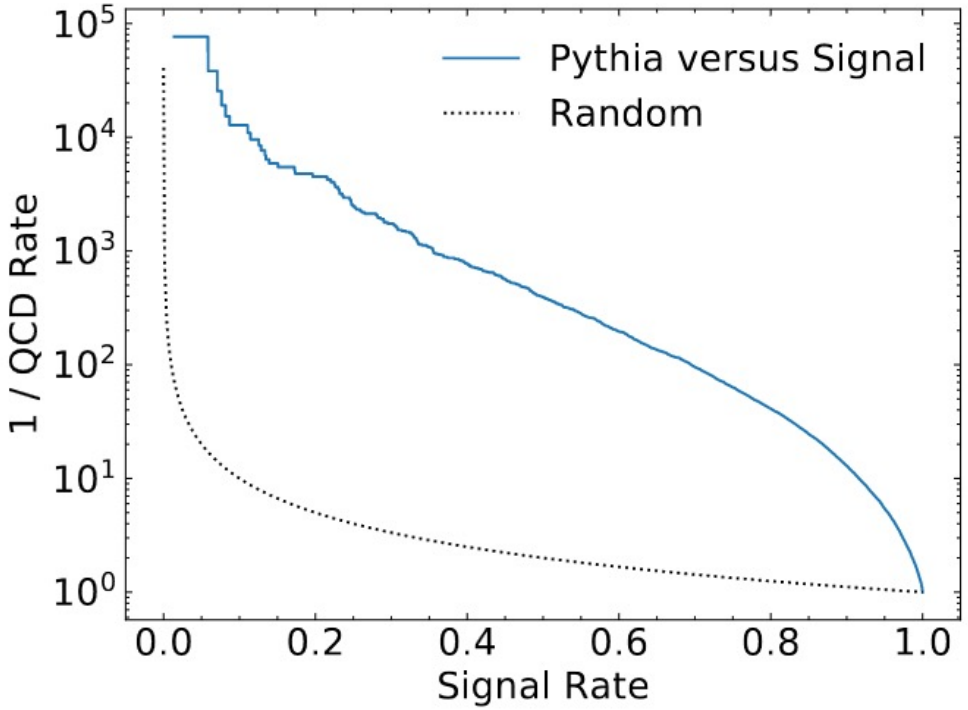
https://doi.org/10.1016/j.physletb.2013.08.010



Inclusive di-photon mass spectrum in 7 and 8 TeV LHC data from ATLAS. Signal and background fits are superimposed.

Interlude: Significance improvement

- On the previous slides, I introduced several cuts that were used in the search for the Higgs (using di-photons instead of jets, placing photon identification and isolation cuts, placing photon kinematic cuts, etc.)
- **Critical point**: each cut reduces the number of accepted true signal (Higgs) events, *but* the number of background events accepted is reduced even more
- There is generally a **tradeoff between signal acceptance and background rejection** when you place cuts

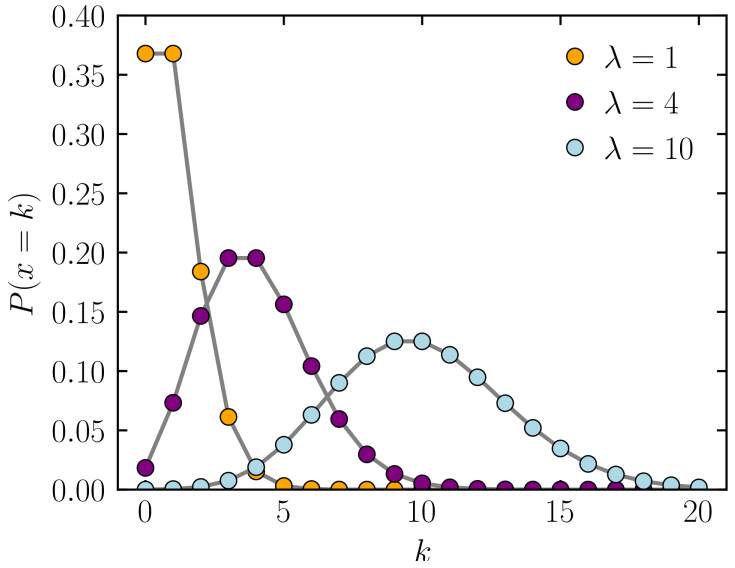


Receiver Operating Characteristic (ROC) curves are a popular way to represent the signal efficiency vs background rejection tradeoff. Here, up and to the right represents "better" performance (higher signal acceptance and lower background rates)

Interlude: Significance improvement

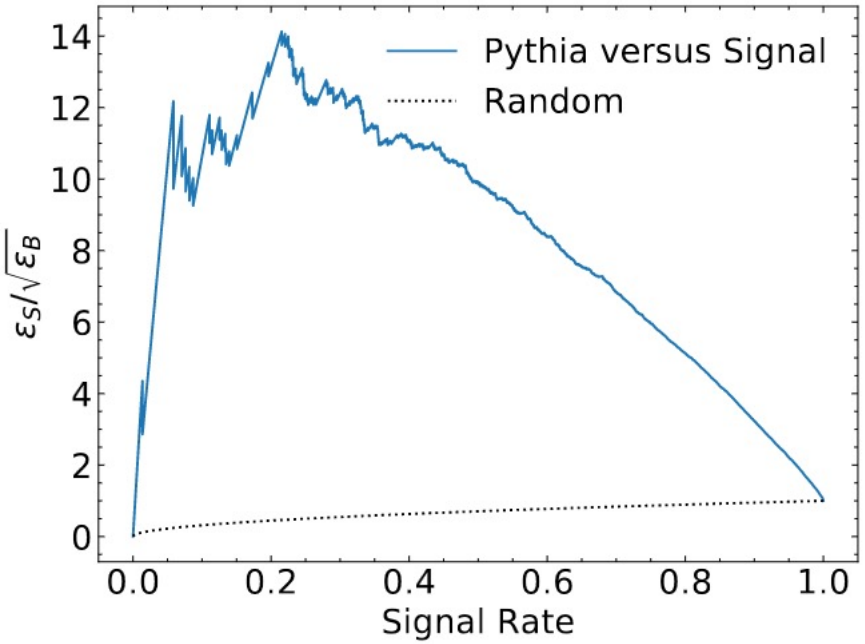
- Generally, in particle physics, we consider event rates to follow Poisson statistics
- The expected value of X (the observed number of events) is λ, with a standard deviation of $\sqrt{\lambda}$
- As λ increases, the Poisson distribution looks increasingly Gaussian
- A common approximation is used:
  - If we have an expected background rate, N_B, for a process (or histogram bin), then the standard deviation is $\sigma = \sqrt{N_B}$
  - E.g. if we expected 100 background events:
    - An observation of 110 would be a 1 sigma deviation
    - An observation of 120 would be a 2 sigma deviation
    - An observation of 130 would be a 3 sigma deviation
    - Etc.

$$f(k;\lambda) = \Pr(X{=}k) = \frac{\lambda^k e^{-\lambda}}{k!}$$



Interlude: Significance improvement

- Given the preceding approximation, the statistical significance of an observation of N_S signal events is roughly $\frac{N_S}{\sqrt{N_B}}$
- If we place a cut on the events that accepts signal events with an efficiency of ε_S and background events with an efficiency of ε_B, the significance becomes $\frac{\varepsilon_S N_S}{\sqrt{\varepsilon_B N_B}}$, and if we want to compare the relative significances achieved from two different working points, then we need to consider the ratio $\frac{\varepsilon_{S,1}}{\sqrt{\varepsilon_{B,1}}} / \frac{\varepsilon_{S,2}}{\sqrt{\varepsilon_{B,2}}}$ (factors of N_S and N_B cancel)
- Alongside ROC curves, it can be also be popular to show Significance Improvement Curves (SIC), showing $\frac{\varepsilon_S}{\sqrt{\varepsilon_B}}$ as a function of the cut



Example SIC curve for a cuts based on a classifier shown on slide 14. **By sacrificing some signal efficiency, a large increase in overall sensitivity is achieved**

Main takeaway

- The best search for a new particle **maximizes sensitivity**
- This is often done by placing cuts that sacrifice some signal efficiency for a compensatory decrease in background contamination

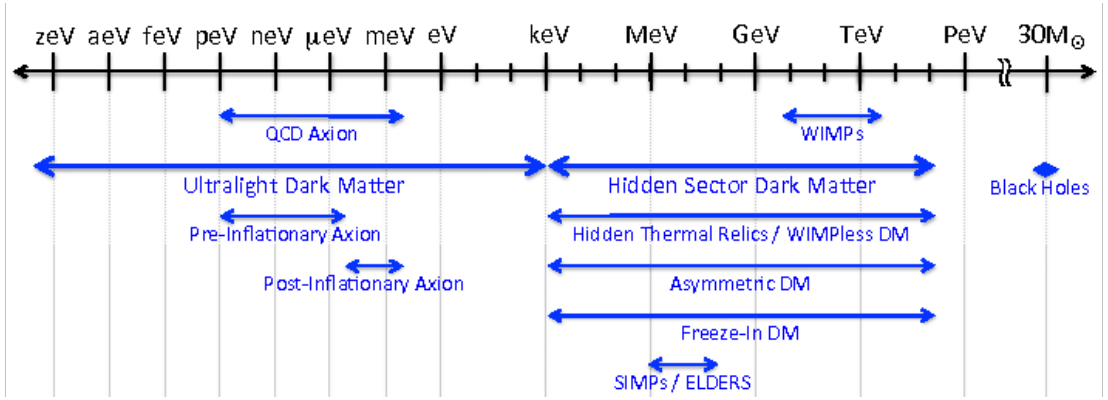
Contents

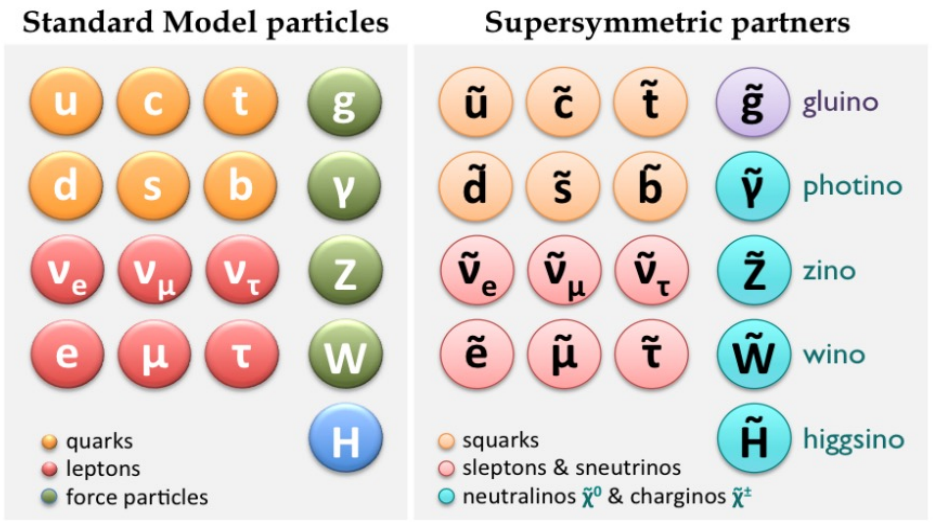
End of the road?

*There are many more arguments than the two I highlighted

- The Higgs boson completed the SM. Do we think there are new particles left to discover?
- Many (most?) people think there should be *something* new*
  - "Experimental" argument: Dark matter doesn't seem to be comprised of normal matter
  - "Theoretical" argument: SM quantum corrections should make the Higgs mass very large. Bizarre fine tuning required for observed mass
- There are *lots* of postulated Beyond the Standard Model (BSM) scenarios

2017 Report





Looking for something new

- In the case of the Higgs, what we were looking for was very specific: a scalar particle with well understood couplings and a narrow mass range it could hide in

- For BSM physics, there's a huge range of parameter space that new particles could be hiding in

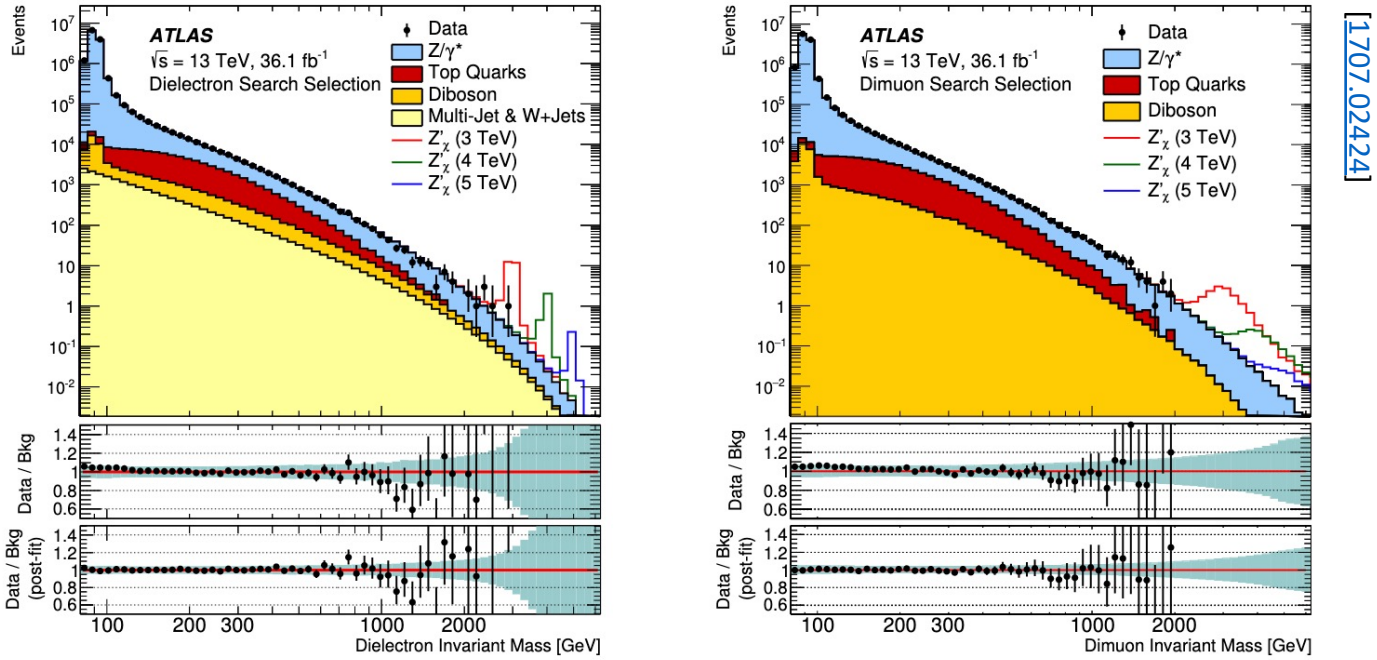
Looking for something new

- In general, BSM searches will fall somewhere on a spectrum of specificity

- On one hand, you can do something non-specific, like an inclusive search for a resonance
  - Sensitive to many BSM signatures, but generally weaker sensitivity
- On the other, you can search for some specific signature
  - Little breadth, but “optimized” sensitivity

Inclusive example

- One of the first analyses performed at a new collider is to look for resonances in di-object spectra
  - E.g. di-muon, di-electron, di-photon, or di-jet mass spectra
- Not super sensitive to rare BSM signatures, as there are high backgrounds

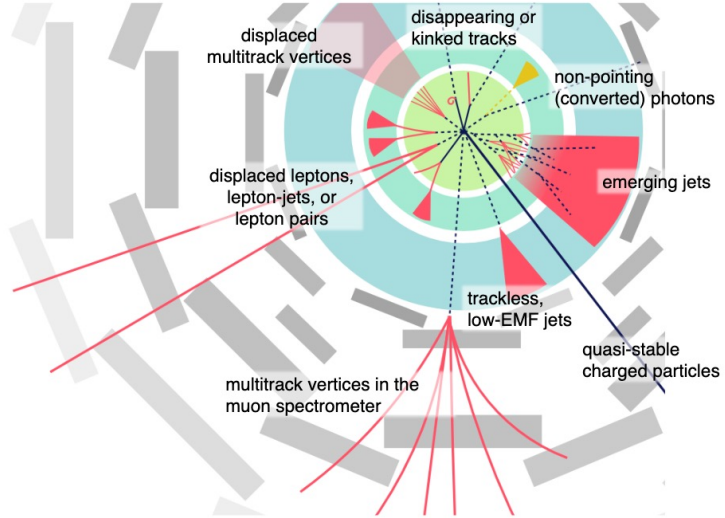


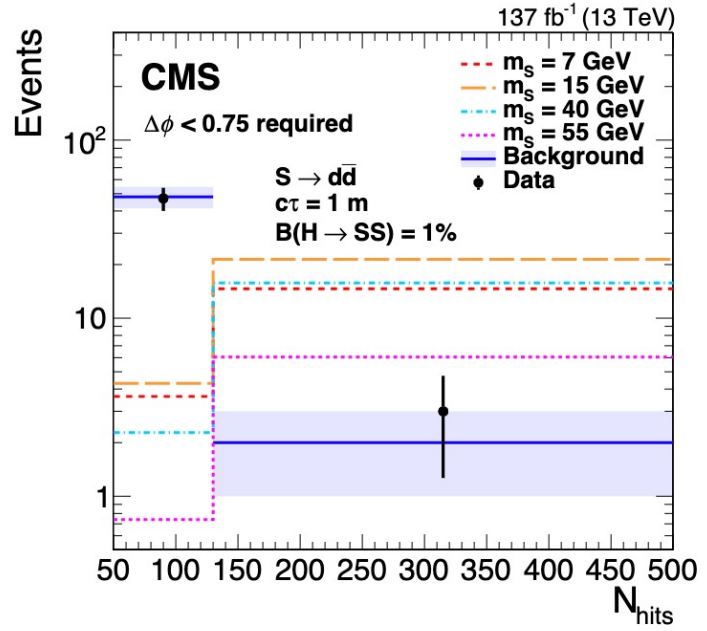
[1707.02424]

ATLAS search for di-electron and di-muon resonances. Typically interpreted in the context of a new, high-mass BSM boson

Specific example

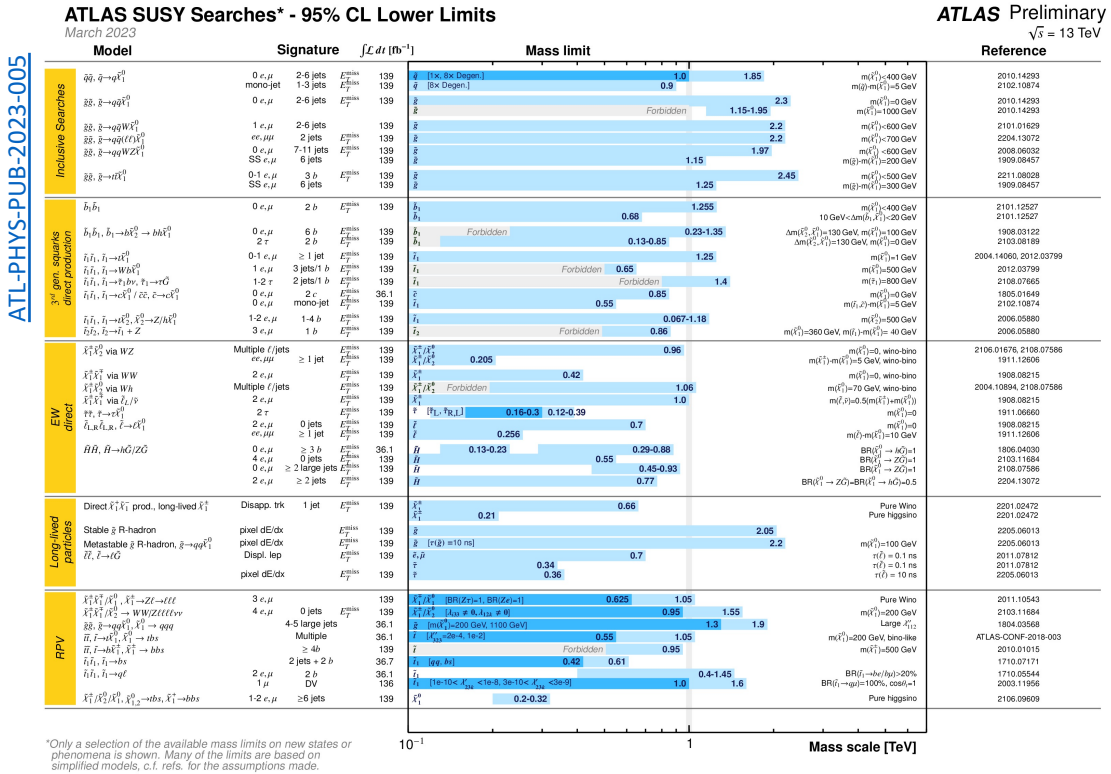
- One can consider a specific model or class of models that will result in some unique final state
  - While there may be low backgrounds, the more “unique” a signature is, the fewer BSM signatures could cause it
- E.g. [2107.04838]:
  - Search for Higgs decays into pair of neutral long-lived scalars that then decay to bottoms, taus, or downs in the endcap muon detectors
  - For given parameters (e.g. scalar mass and lifetime), one can fully simulate what these events would look like in the detector, and search for them accordingly

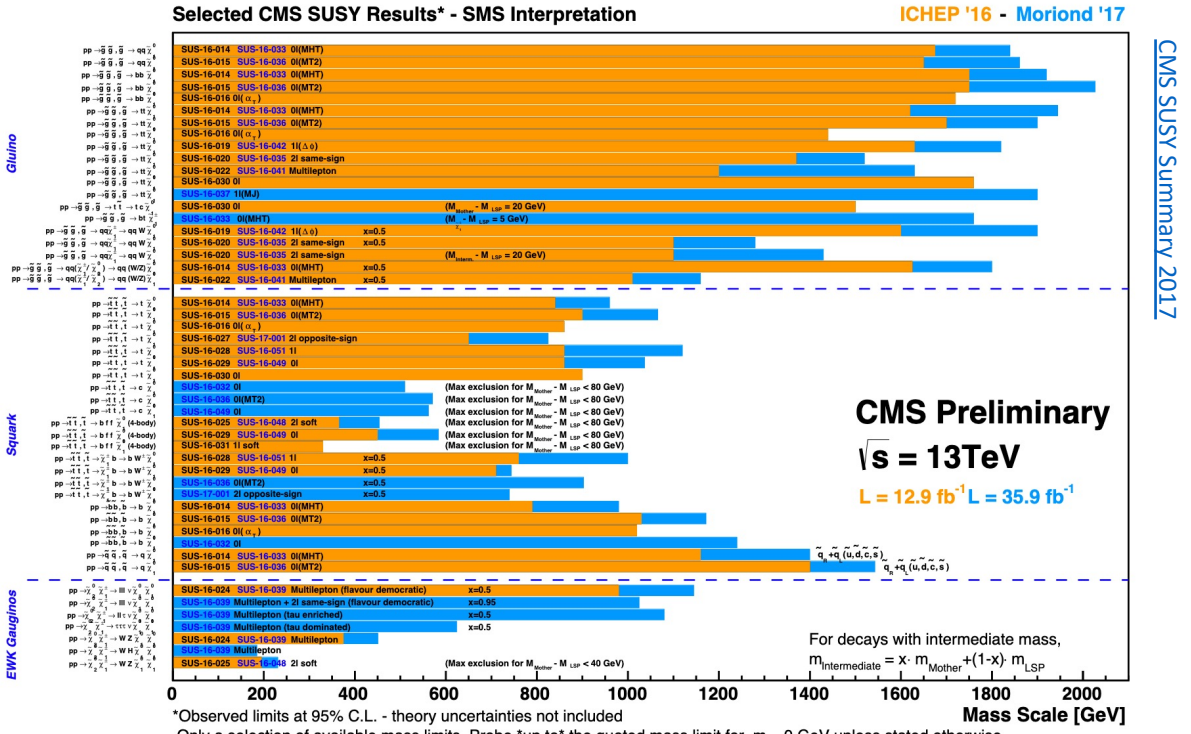




Exclusions limits

- It would be impossible to exclude every potential BSM scenario with a specific search, but searches for many popular possibilities have been performed
  - So far, no new particles discovered – we get exclusion limits though





Contents

Searching for an anomaly

- In the past ~4-5 years, there's been increased focus on model agnostic searches powered by ML
  - The idea is to create a search that's sensitive to many BSM scenarios while being significantly more sensitive than traditional inclusive bump hunts

- What you need:
  - A final state that could be created in many BSM models
  - Some sort of classifier that can distinguish events created by SM processes from those created by BSM processes (then a cut can be made with $\frac{\varepsilon_S}{\sqrt{\varepsilon_B}} > 1$)

Designing an anomaly search

- What is a final state that could be created in many BSM models?
  - At the LHC, the most generic type of event would be di-jet production. If a new particle can be created in the collisions of two protons (think s-channel production), then it can almost certainly decay to two jets

- How do you create a classifier to distinguish SM from non-SM?
  - **We've finally made it to the main point of this lecture**

Framing device: LHC Olympics

- In 2020, the HEP anomaly detection community put together the LHC Olympics, which was a chance for researcher teams to try out their algorithms on pre-assembled black box simulated datasets
  - Black boxes could have either some simulated signal injected or no signal
- [2101.08320] – Good overview of the methods used and their motivations

Anomaly detection techniques

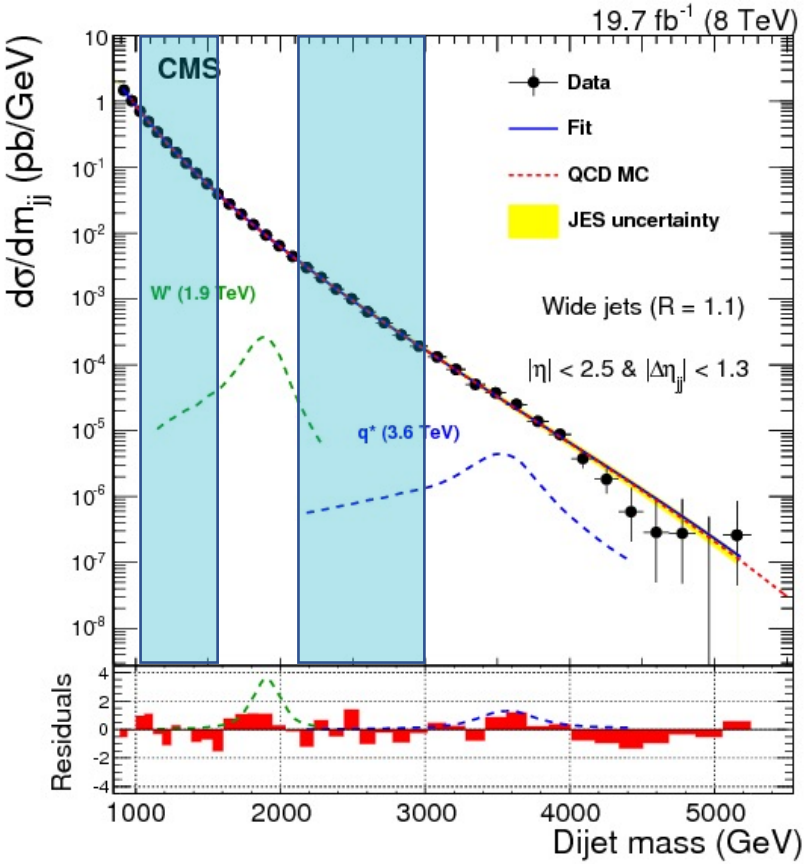
- Broadly speaking, we can group many anomaly detection schemes into three categories based on the classifier architecture used and how training is performed

Anomaly detection techniques

- **Unsupervised Methods**
  - Classifier does not use any labels in training
  - Training performed on data (which is mostly background)
  - Goal: determine the probability that an event is just a background event
  - Events with low probability are more “anomalous” and more likely to be BSM

Anomaly detection techniques

- **Unsupervised Methods**
  - Goal: determine the probability that an event is just a background event
- **Weakly Supervised Methods**
  - Train with noisy labels
  - Typically assume that signal is localized in at least one event-level variable (like dijet mass)
  - Subsets of data can be trained against each other
  - Goal: if one subset of data is signal enriched, classifier will learn to distinguish signal from background



If this W' resonance actually existed, the middle region would be a signal-enriched subset of the data. Sidebands would be a similar dataset, but with less signal

Anomaly detection techniques

- **Unsupervised Methods**
  - Goal: determine the probability that an event is just a background event
- **Weakly Supervised Methods**
  - Goal: if one subset of data is signal enriched, classifier will learn to distinguish signal from background
- **Semi Supervised Methods**
  - These techniques typically attempt to enhance one of the above with some information about expected signals

E.g. maybe you expect that your signal will bark instead of hiss

Contents

Unsupervised searches

- For unsupervised learning, we need to use ML algorithms that *identify patterns within a dataset* without external guidance
- I'll highlight two popular techniques:
  - **Variational Autoencoders (VAE)**: a probabilistic generative algorithm that attempts to learn a low-dimensional representation of the data that still encapsulates all information
  - **Normalizing Flows (NF):** also a generative algorithm, but one that produces tractable distributions for high-dimensional probability density estimation

Variational Autoencoders for AD

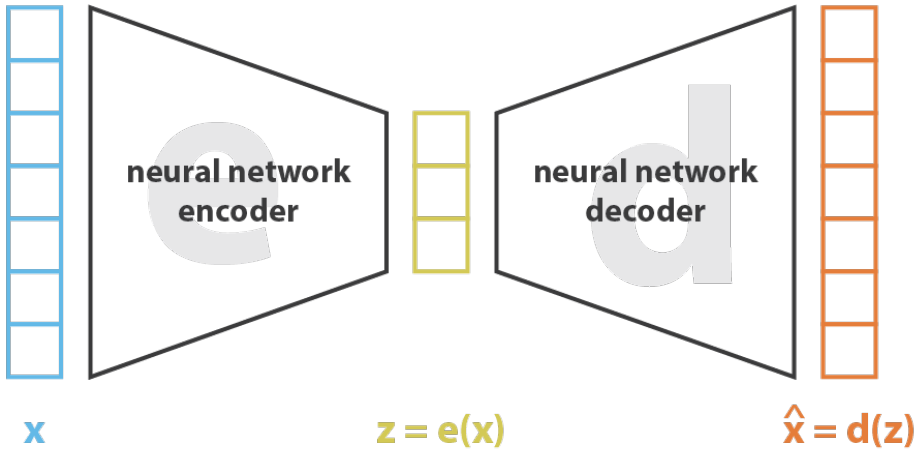
[Lectures on VAE]

VAE paper [1906.02691]

[Accessible intro article]

- Regular autoencoder

Component 1:
Encoder network performs "dimensionality reduction"*. High dimensional data is sent to a "latent space" of reduced dimension. The latent space is formed from (potentially) nonlinear combinations of input values.



*Another (linear) example of dimensionality reduction would be Primary Component Analysis (PCA), if you've encountered that before

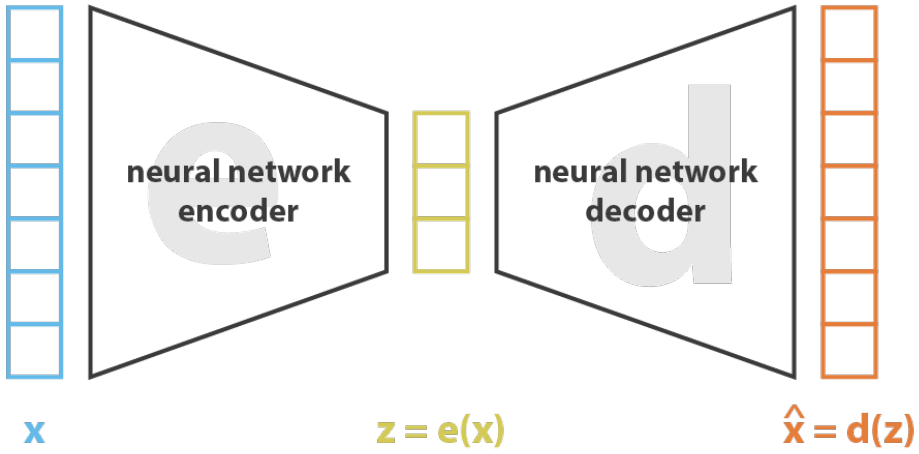
Variational Autoencoders for AD

Lectures on VAE

VAE paper [1906.02691]

Accessible intro article

- Regular autoencoder



Component 2:
Decoder network performs attempts to reconstruct the original inputs based on the latent space information alone

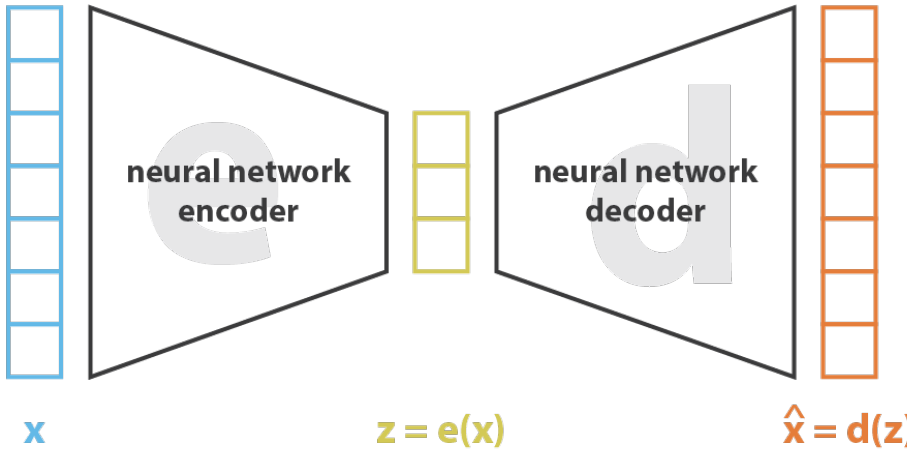
Variational Autoencoders for AD

Lectures on VAE

VAE paper [1906.02691]

Accessible intro article

- Regular autoencoder



$$\text{loss} = \| x - \hat{x} \|^2 = \| x - d(z) \|^2 = \| x - d(e(x)) \|^2$$

Training aims to minimize "reconstruction" loss, which is the difference between the initial and reconstructed data

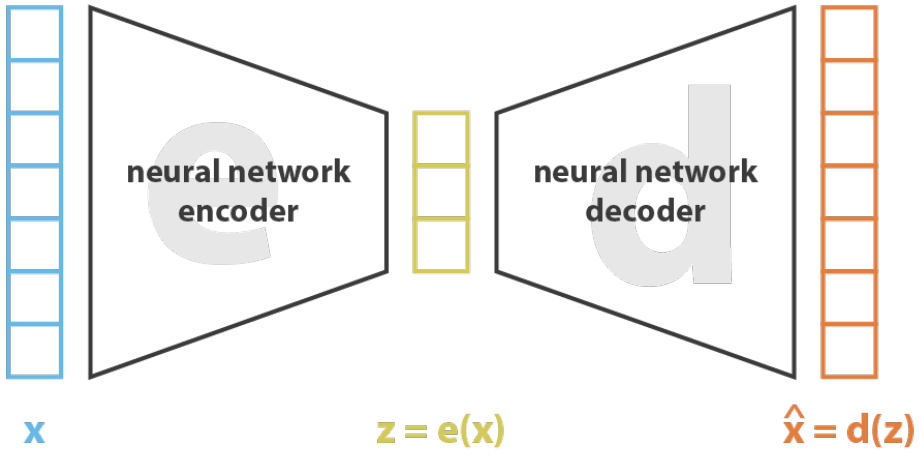
Variational Autoencoders for AD

Lectures on VAE

VAE paper [1906.02691]

Accessible intro article

- Regular autoencoder



"Subtle" problem with regular AEs: there is no guarantee that latent space will have any regularity or structure so to speak. There can be overfitting in the latent space and events out of the training distribution can have nonsensical decoding

$$\text{loss} = \| x - \hat{x} \|^2 = \| x - d(z) \|^2 = \| x - d(e(x)) \|^2$$

Variational Autoencoders for AD

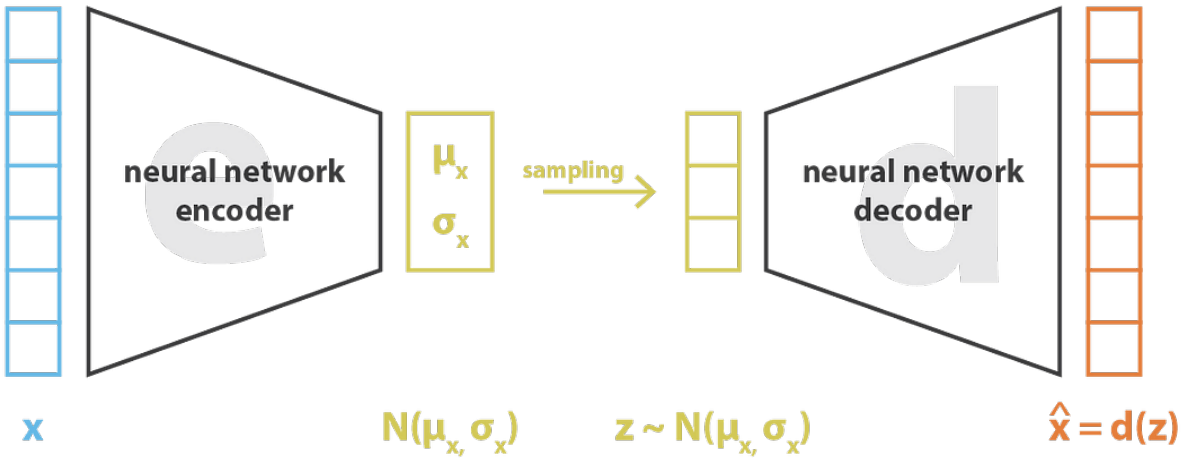
[Lectures on VAE]

VAE paper [1906.02691]

[Accessible intro article]

- Variational autoencoder (VAE)

A VAE is a restructured autoencoder that aims to have a regularized latent space



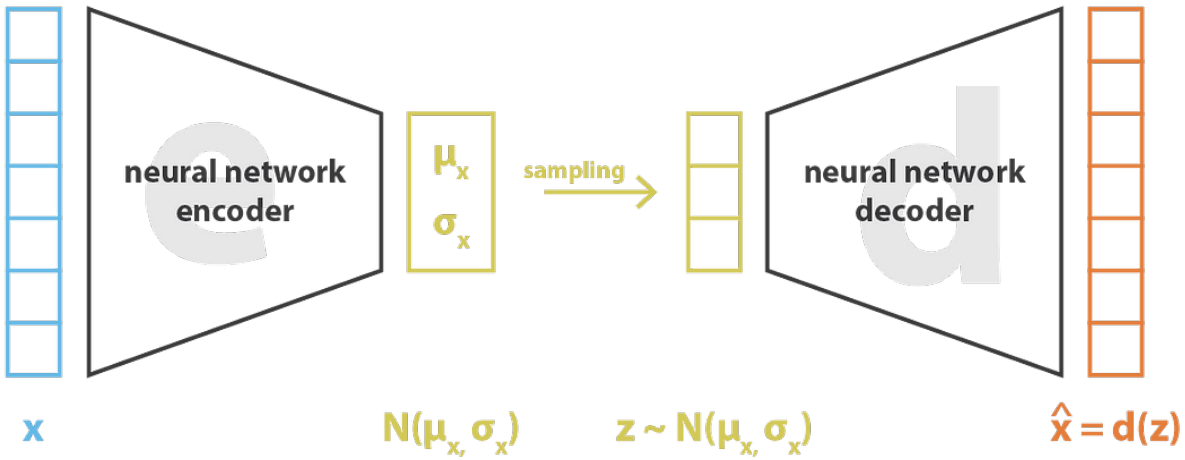
Variational Autoencoders for AD

[Lectures on VAE]

VAE paper [1906.02691]

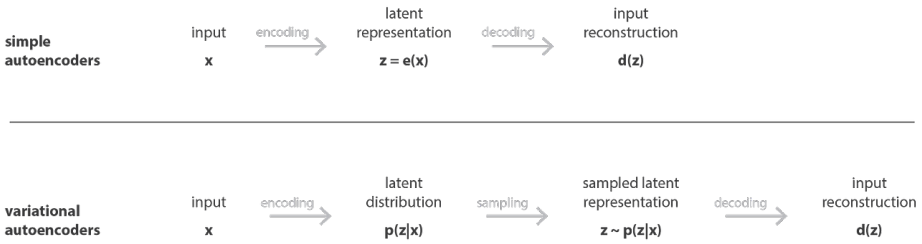
[Accessible intro article]

- Variational autoencoder (VAE)



Normally, the distributions used are normal distributions, so for each input instance, the encoder generates a set of means and sigmas

Instead of mapping each input to a single point in the latent space, in a VAE, each input is mapped to a *distribution* over the latent space



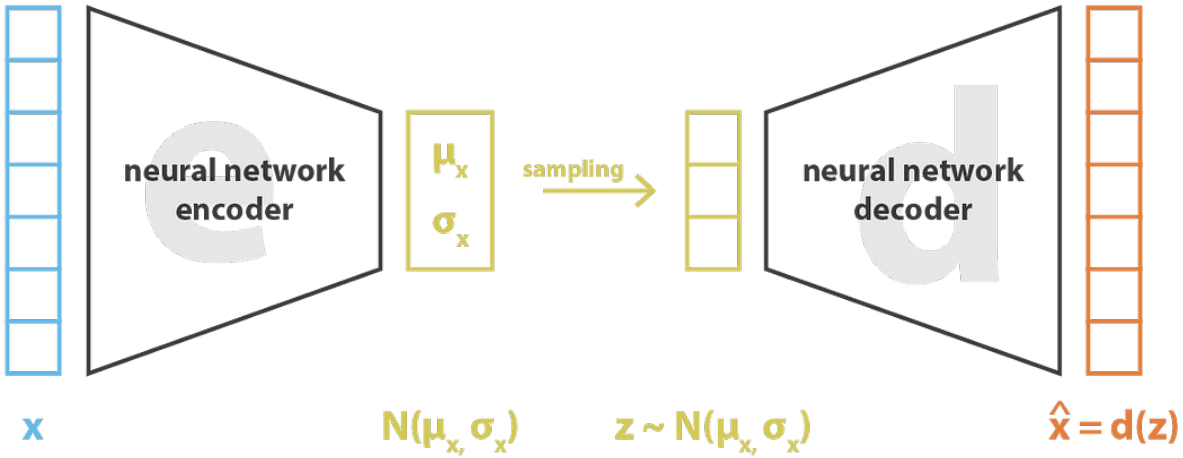
Variational Autoencoders for AD

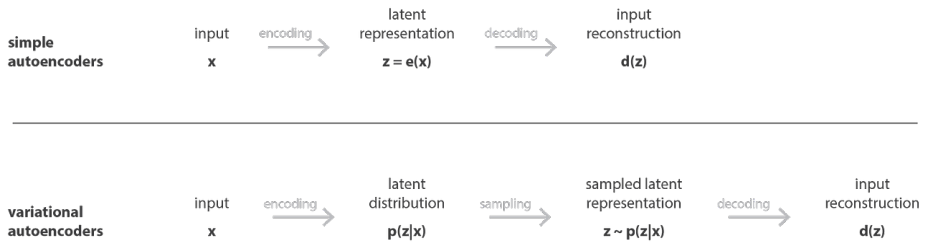
[Lectures on VAE]

VAE paper [1906.02691]

[Accessible intro article]

- Variational autoencoder (VAE)





For the decoding step, a point is sampled from the distribution associated with the input, and *that* is the point that is sent through the decoder network

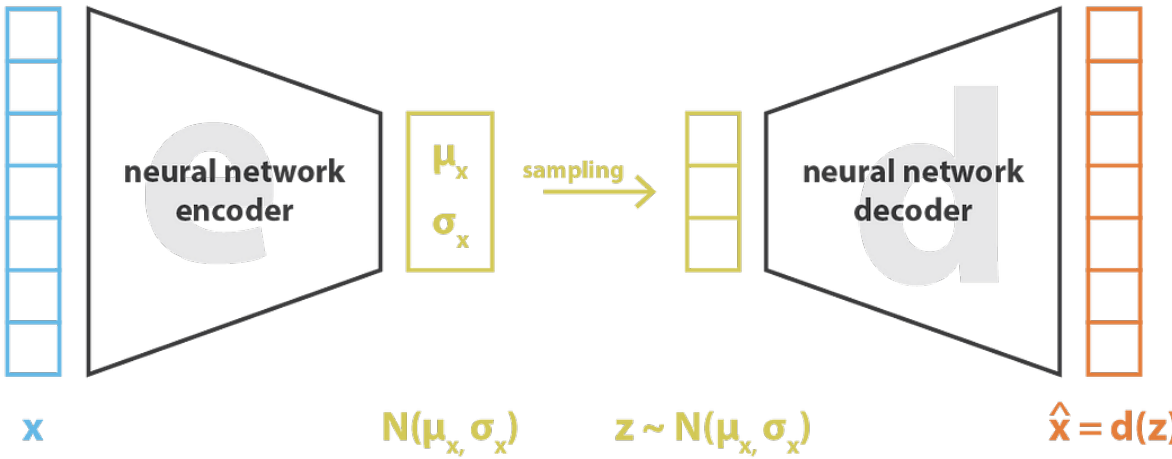
Variational Autoencoders for AD

Lectures on VAE

VAE paper [1906.02691]

Accessible intro article

- Variational autoencoder (VAE)



Regularization term is typically Kullback-Leibler divergence

$$\text{loss} = \|x - \hat{x}\|^2 + KL[N(\mu_x, \sigma_x), N(0, I)] = \|x - d(z)\|^2 + KL[N(\mu_x, \sigma_x), N(0, I)]$$

The loss used is a combination of the same reconstruction loss as before, but now there is a regularization term that aims to make the latent distributions close to a standard normal distribution

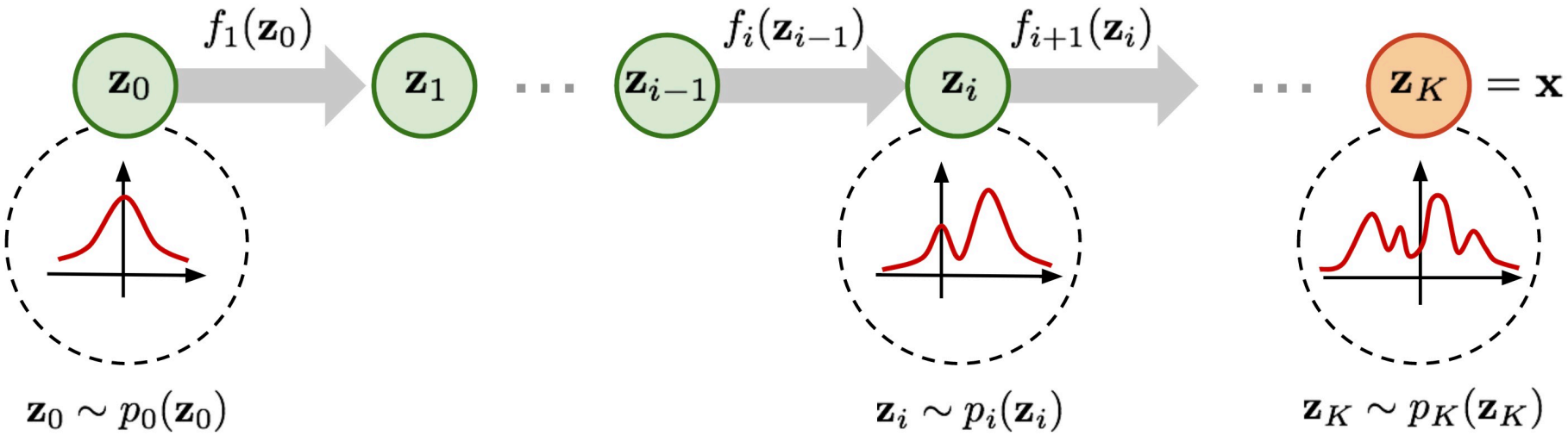
Variational Autoencoders for AD

- To use a VAE for AD in HEP:
  - First you would train a VAE on a dataset (if you can train on some “control region” that you expect to be signal-free, that might be best)
    - The VAE should be able to learn what typical events look like, and should be able to reconstruct them well
  - Then for each event in the region of interest, the “Anomaly Score” can be the reconstruction loss
    - Signal events should be “new” to the VAE, so it shouldn’t know how to reconstruct those events particularly well

Normalizing flows for AD

NF review paper [1908.09257]
Lecture notes

- A normalizing flow is a transformation of a simple probability distribution (often a normal distribution) to some other, complicated PDF
  - This is achieved via a sequence of *invertible* and *differentiable* mappings
- Density estimation: if you learn the PDFs for the input variables of a dataset, then for any input point, you can determine how rare or common it is (relative to the training dataset)



Normalizing flows for AD

NF review paper [1908.09257]
Lecture notes

- Mathematically speaking, you can imagine a flow as an N-dimensional mapping

$$X = f(Z) \text{ and } Z = f^{-1}(X)$$

- You can parameterize your flow* as below

$$X = f_\theta(Z) \text{ and } Z = f_\theta^{-1}(X)$$

- Requirements:
  - Input and output dimensions must be same
  - Transformation must be invertible
  - Computing the determinant of the Jacobian needs to be efficient and differentiable

$$p_X(\mathbb{x}) = p_Z(f^{-1}(\mathbb{x})) \left| \det \left(\frac{\partial f^{-1}(\mathbb{x})}{\partial \mathbb{x}} \right) \right|$$

Probability distribution transformation involves the Jacobian (accounts for distribution volume changes)

$$p_X(\mathbb{x}; \theta) = p_Z(f_\theta^{-1}(\mathbb{x})) \left| \det \left(\frac{\partial f_\theta^{-1}(\mathbb{x})}{\partial \mathbb{x}} \right) \right|$$

*There are many commonly used flow models that meet the basic requirements. Masked autoregressive rational quadratic splines have been used in HEP before. There are also simpler flows, like planar flows are radial flows (see the NF review paper linked above)

Normalizing flows for AD

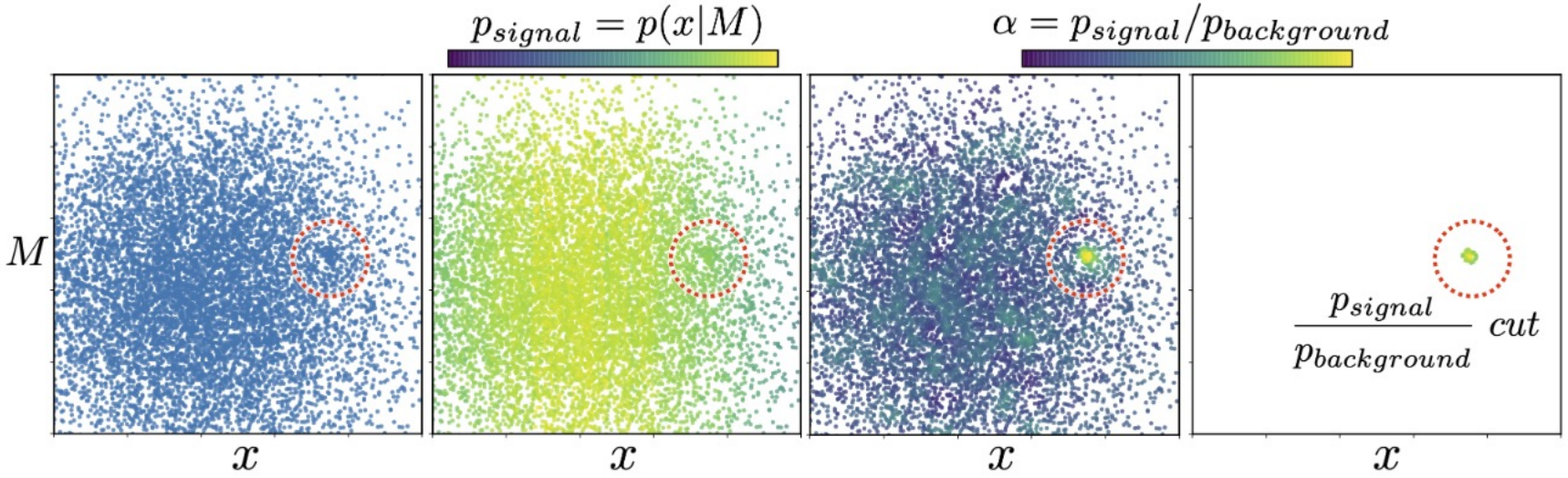
GIS [2012.11638]

- To use a NF for AD in HEP:
  - Simple idea: find the likelihood for different events, and use that as an anomaly score that can be cut on
  - Can be used in a semi-supervised approach
- One idea that worked well at the LHC Olympics was “Gaussianizing Iterative Slicing” (GIS)
  - This is an algorithm aimed at finding *in-distribution* anomalies via relative overdensities

Normalizing flows for AD

GIS [2012.11638]

- GIS algorithm:
  - Use flow to calculate density conditional on feature where you might expect an overdensity. New physics should result in an overdensity of events with energy near a new particle's mass (resonances!)
    - If we're talking about a di-jet search, we might denote this as $p(x|M_{JJ})$, where x are other event features
  - To check for an overdensity at a given M, you calculate the density at nearby masses, $M \pm \Delta$. Then you interpolate* to get the density estimation at M ($p_{background}$)
  - You then compare the actual density observed at M to the interpolated density. If $\frac{p_{signal}}{p_{background}} > 1$, that's a hint of an overdensity. Treat this ratio as the anomaly score



*GIS authors seem to leave interpolation technique and choice of Δ as an exercise for the reader

Normalizing flows for AD

GIS [2012.11638]

- GIS was one of the best performing algorithms at the LHC Olympics

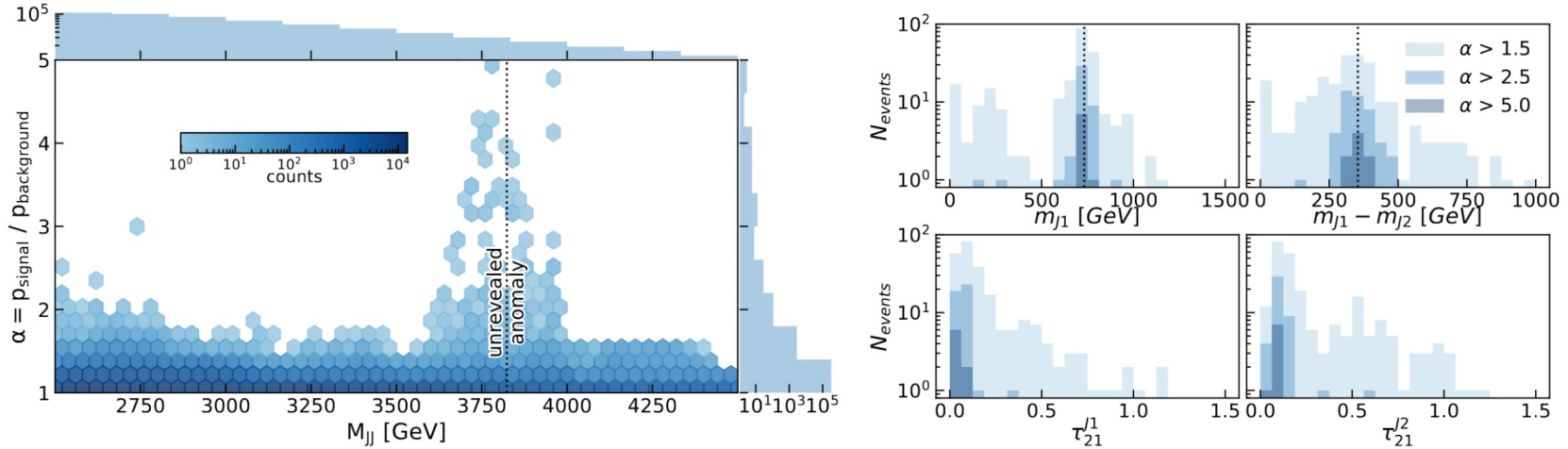


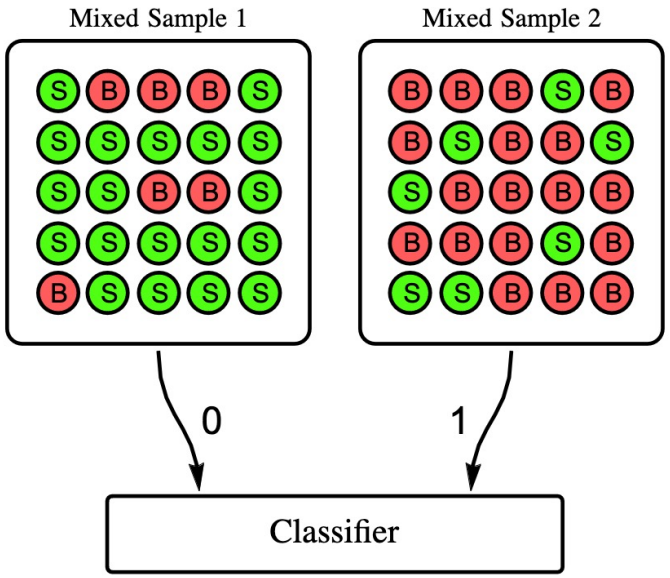
Figure 2: **Left**: The anomaly score for each event as a function of the invariant mass of the leading two jets. A number of anomalous events are clearly seen near $\mathrm{M_{JJ}} \approx 3750\mathrm{GeV}$. **Right:** parameter distributions of the events that remain after imposing cuts on the anomaly score α. Vertical dashed lines are the true anomalous events that were unveiled after the close of the competition.

Contents

Weakly supervised searches

CWoLa paper

- In fully supervised case, training labels are correct
  - I.e. an event labelled “signal” really is signal, and if the label is “background”, that’s actually background
  - An alternative way of phrasing this is that you’re training a classifier to distinguish two populations: one that’s all signal from one that’s all background
- The idea for weak supervision comes from the idea of “learning from label proportions” or learning with noisy labels
  - In this case, you train a classifier to distinguish two populations, but the populations trained against each other are mixtures of signal and background



Weakly supervised searches

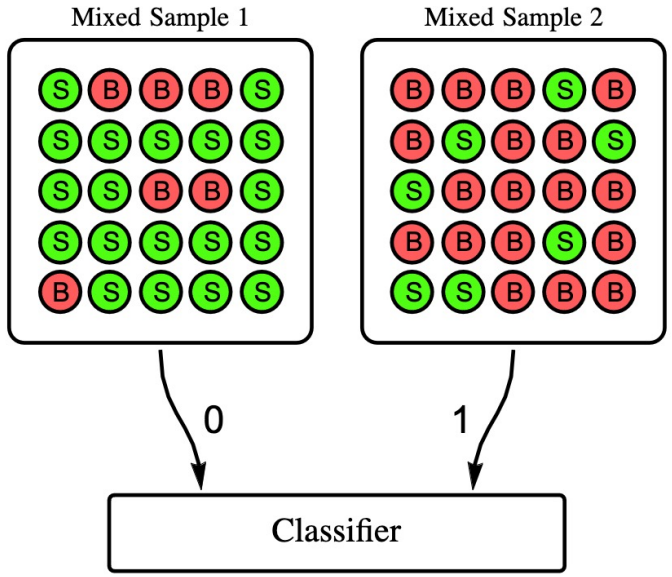
CWoLa paper

- Optimized classifiers:
  - Let's imagine that we have a classifier with signal efficiency ϵ_S and background efficiency ϵ_B for a given cut
  - The optimal classifier achieves the lowest ϵ_B for any given ϵ_S
  - According to the Neyman-Pearson Lemma, an optimal classifier is the likelihood ratio $p_S(x)/p_B(x)$ for given event, x
- ML-based classification:
  - Common ML models for classification include neural networks (NN) and boosted decision trees (BDT)
  - For a given architecture, you typically train to find a set of parameters that minimize a loss function like $\ell_{mse} = \frac{1}{N}\sum_{i=1}^{N}\left(c(x_i) - \mathbb{I}(u_i = S)\right)^2$
    - c is the classification function, $\mathbb{I}$ is the indicator function, u_i is the label for event x_i
  - For an appropriately expressive c, its performance should approach that of the optimal classifier

Weakly supervised searches

CWoLa paper

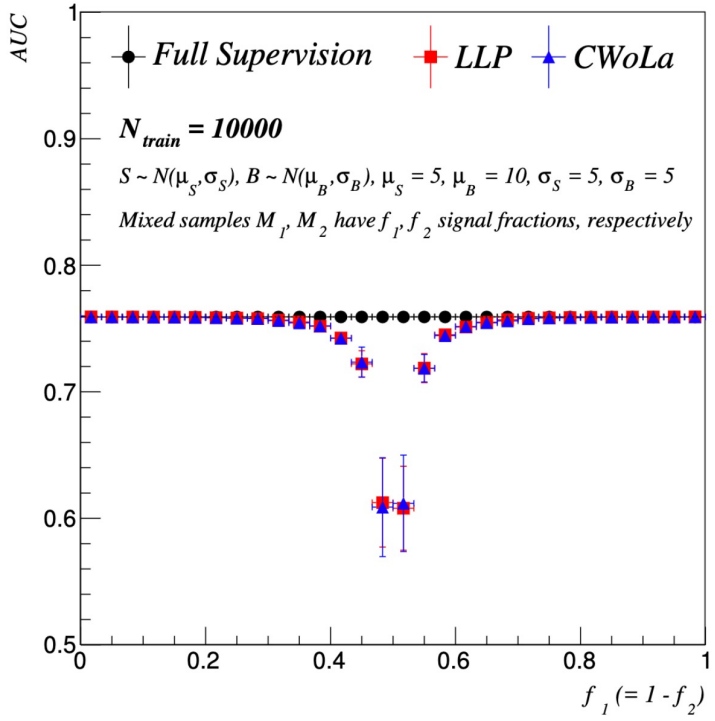
- Now, say that you have two mixed samples, M_1 and M_2
  - $p_{M_1}(x) = f_1 p_s(x) + (1 - f_1) p_B(x)$
  - $p_{M_2}(x) = f_2 p_s(x) + (1 - f_2) p_B(x)$
  - f_1 and f_2 are the respective signal fractions
- If you do some reshuffling, you can show that
  - $\frac{p_{M_1}}{p_{M_2}} = \frac{f_1 p_S + (1-f_1) p_B}{f_2 p_s + (1-f_2) p_B} = \frac{f_1 \frac{p_S}{p_B} + (1-f_1)}{f_2 \frac{p_S}{p_B} + (1-f_2)}$
  - That is to say that an **optimal classifier for M_1 vs M_2 is a monotonic rescaling of the optimal classifier for S vs B**



Classification Without Labels (CWoLa)

CWoLa paper

- Based on the previous slide, if we train a NN to distinguish M_1 from M_2, then the NN is also a classifier for signal vs background
- Since we don't need signal or background *labels* here, this approach is often called "Classification Without Labels" (CWoLa)
  - First CWoLa paper wasn't for AD, but was a quark vs gluon jet tagger
  - Relevant point for AD: **training can be performed exclusively with data**

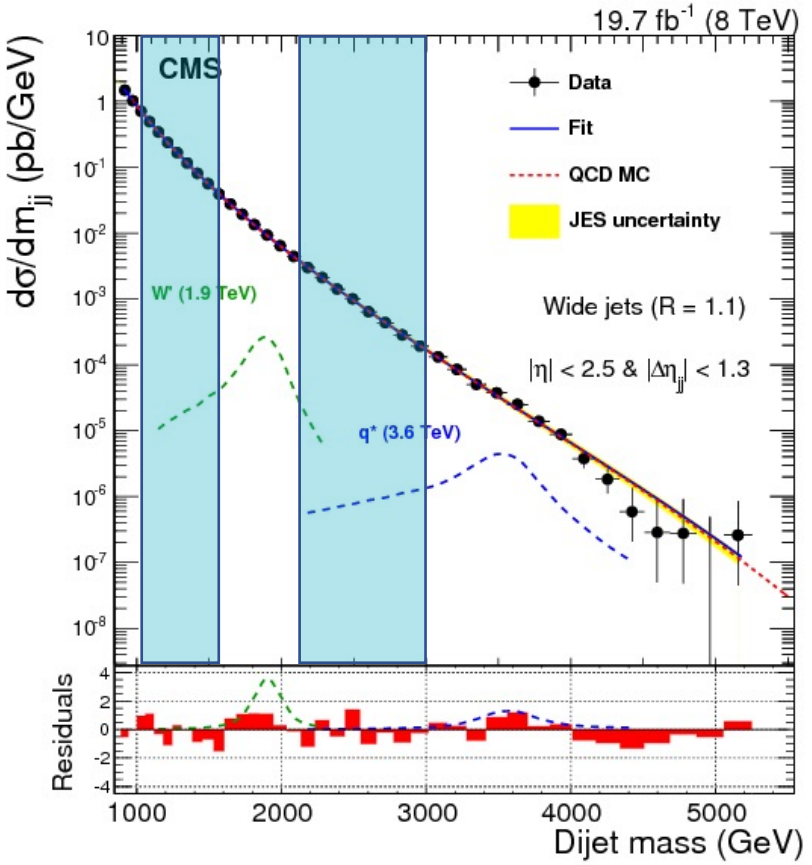


As long as signal fractions in the two samples are not too similar, a classifier based on CWoLa training approximates a fully trained classifier

CWoLa for AD

ATLAS application of CWoLa for AD: [2005.02983]

- Remember our goal: create a classifier that distinguishes BSM physics from SM physics
- How can CWoLa fit in?
  - We can use CWoLa to train on data. Ideally, if we train on BSM-enriched data, then our classifier will learn about the BSM physics that's there without us having to guess the new physics correctly
- For CWoLa to work, we would need a BSM-enriched subsample of the data
  - Hopefully I convinced you earlier that massive particles can lead to a resonance in a mass spectrum
  - The production rate of the massive particle increases when energy transfer ≈ particle mass

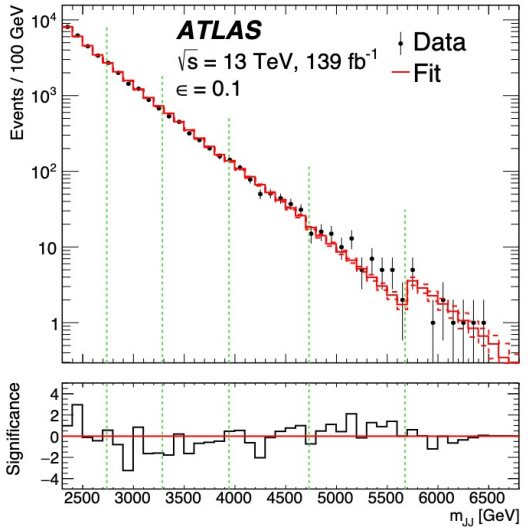


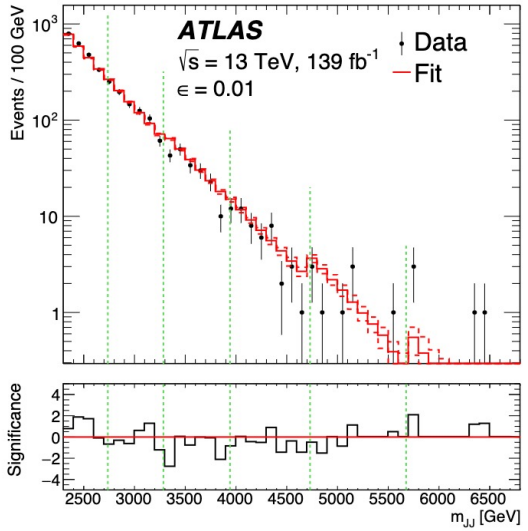
If a new particle is present in the data, then the subset of events with energy transfer near the particles constitute a signal-enriched subsample

CWoLa for AD

ATLAS application of CWoLa for AD: [2005.02983]

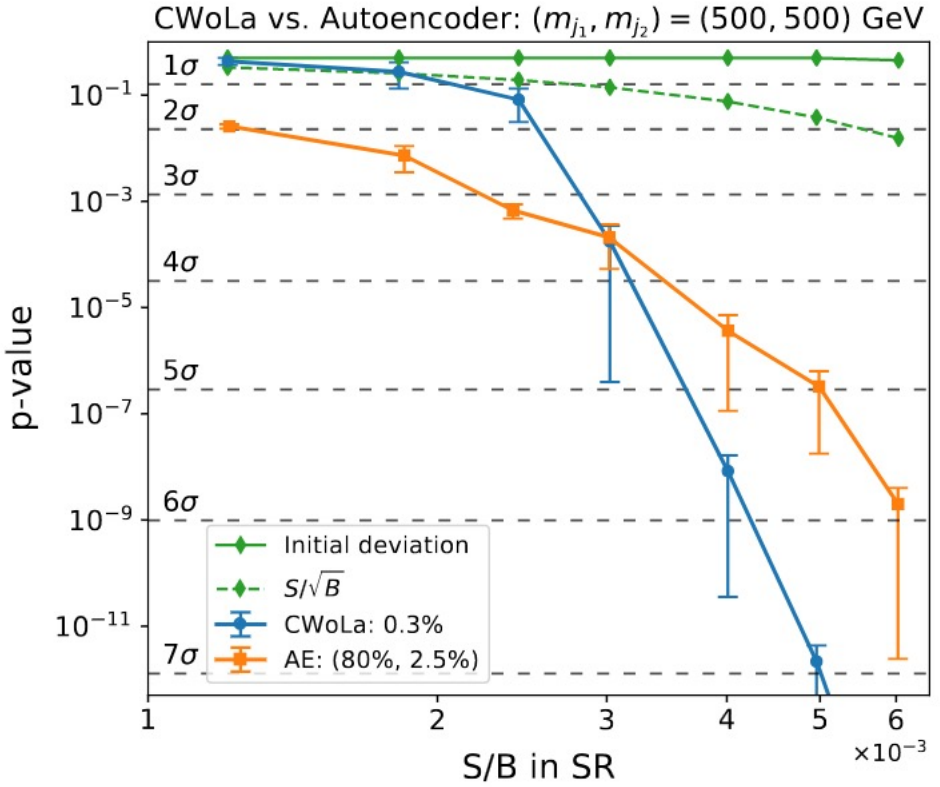
- CWoLa for AD has been demonstrated by ATLAS
  - Their strategy was to split up a di-jet mass spectrum into multiple mass “bins”
  - Then for a given bin, you train a classifier to distinguish that bin from its neighboring bins
  - Classifier inputs were just the masses of the two jets in each event
- One “weakness” of CWoLa is that the features used in the classifier should be uncorrelated with the resonant feature
  - Can be difficult to accomplish
  - ATLAS had to do some “interesting” rescaling of the individual jet masses to get around this





CWoLa vs VAE

- You might be wondering how CWoLa compares to the unsupervised methods discussed earlier
- A study at the LHC Olympics noted that performance for VAE was better for smaller signal rates, where CWoLa performed better for higher signal rates
  - Conclusion: there is some degree of complementarity between AD algorithms



Extending CWoLa

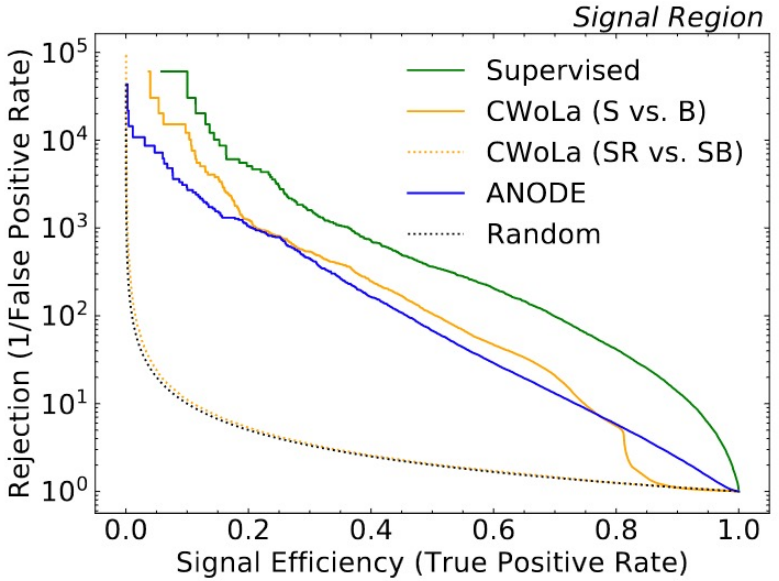
ANODE [2001.04990]
CATHODE [2109.00546]
TNT [2002.12376]

- Ways to improve CWoLa:
  - Reduce difficulties associated with feature-resonance correlations
  - Devise some means of creating even more signal-enriched subset of events than you get from bare mass windows
- Some variations resembling the CWoLa scheme have been developed
  - “Anomaly Detection with Density Estimation” (ANODE) and the related “Classifying Anomalies THrough Outer Density Estimation” (CATHODE)
  - “Tag n’ Train” (TNT)
- These methods combine CWoLa with unsupervised learning techniques

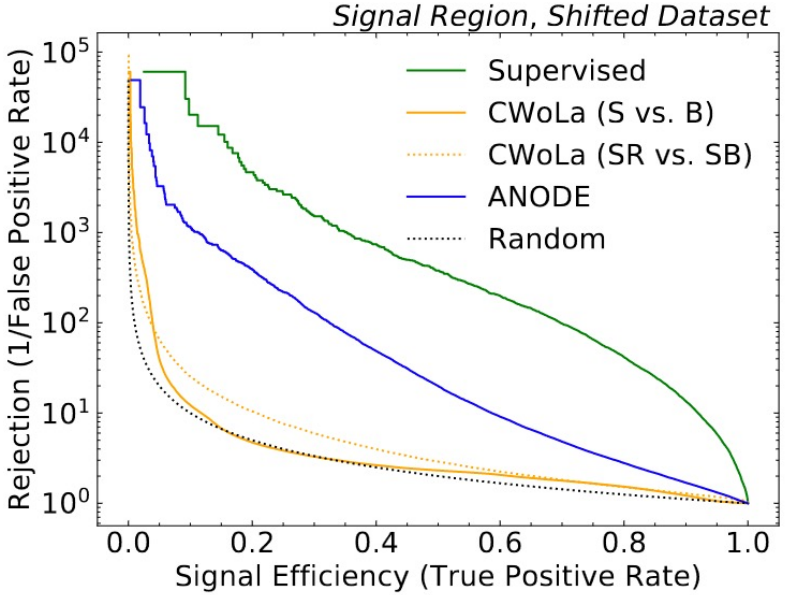
ANODE

ANODE [2001.04990]
CATHODE [2109.00546]

- For CWoLa, we have the basic logical flow:
  - $\frac{p_{data}(x|SR)}{p_{data}(x|SB)} \rightarrow \frac{p_{data}(x|SR)}{p_{background}(x|SB)} \rightarrow \frac{p_{data}(x|SR)}{p_{background}(x|SR)}$
  - Implicit assumptions:
    - The PDFs in the sideband is the background PDF
    - The background PDF in the sideband is the same as the background PDF in the signal region
- Instead of a single SR vs SB classifier (like a NN), ANODE instead trains normalizing flows in the SR and SB
  - Signal region flow gives $p_{data}(x|SR)$, but sideband flow gives a smoother interpolation of background probability into signal region
  - This helps alleviate CWoLa's feature-resonance correlation problem
- The ratio of probabilities from the two flows is the anomaly score in this case



ANODE performs similarly to CWoLa in the case of uncorrelated features (though slightly worse)

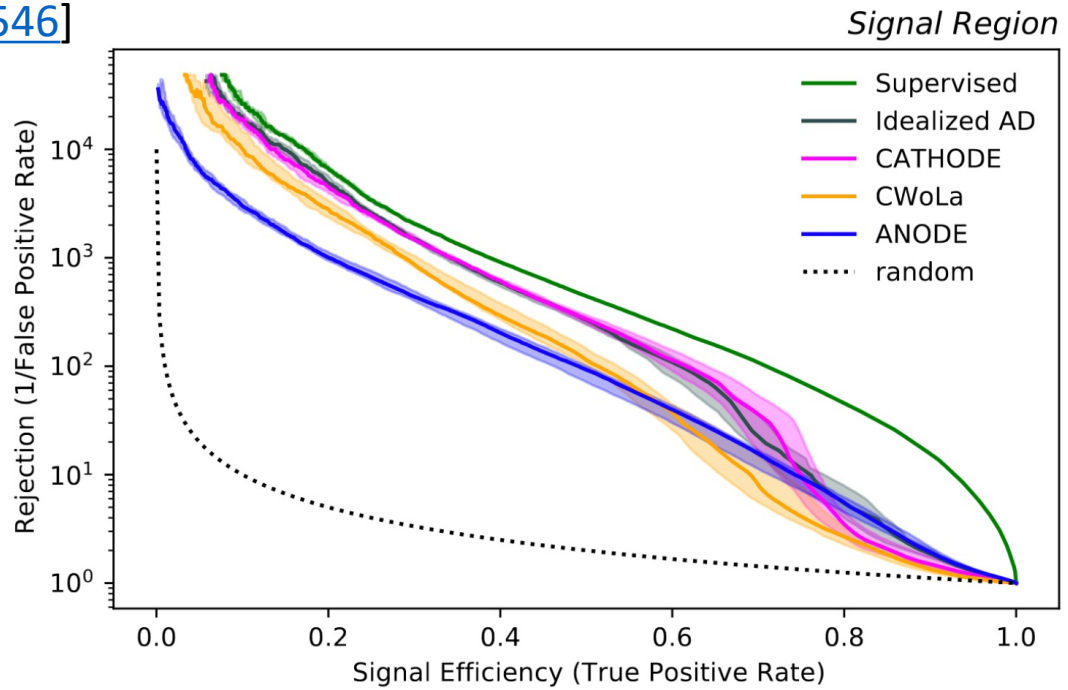


ANODE performs much better than CWoLa when there are feature-resonance correlations

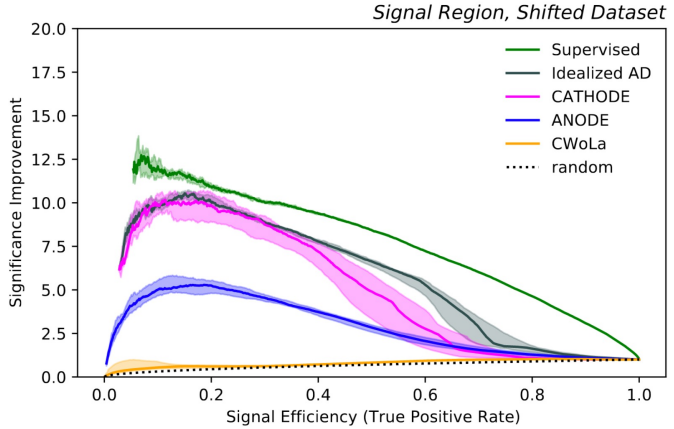
CATHODE

ANODE [2001.04990]
CATHODE [2109.00546]

- CATHODE extends ANODE by using the generative capabilities of NFs
- As in ANODE, a NF is trained on the sideband, but instead just of taking likelihood ratios from a SR NF and a SB NF, in CATHODE, events are *generated* from the SB NF to have masses in the SR
- Then a classifier (NN) is trained to distinguish data event in the SR from synthetic data in the SR



CATHODE performance approaches that of an "idealized" case – classifier specifically trained to distinguish "data" in the SR from the correct background in the SR (studies all performed in simulation)



CATHODE maintains improved performance in the correlated case as well

Tag N’ Train (TNT)

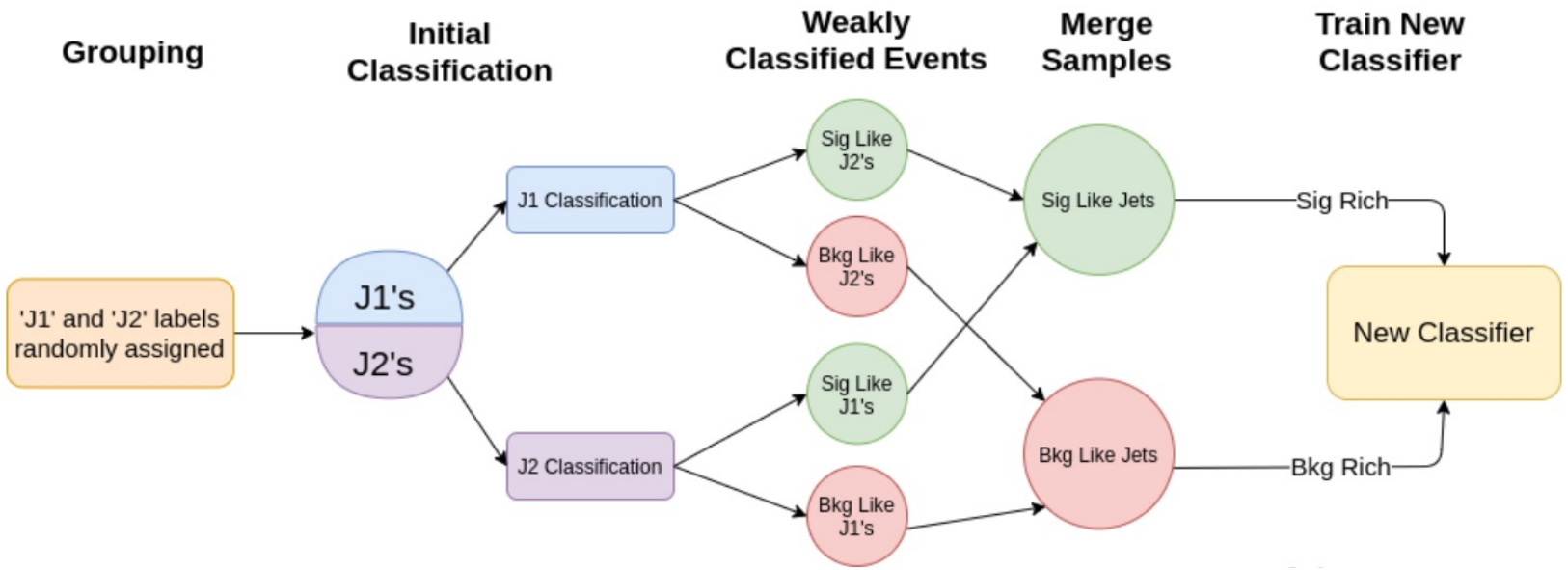
TNT [2002.12376]

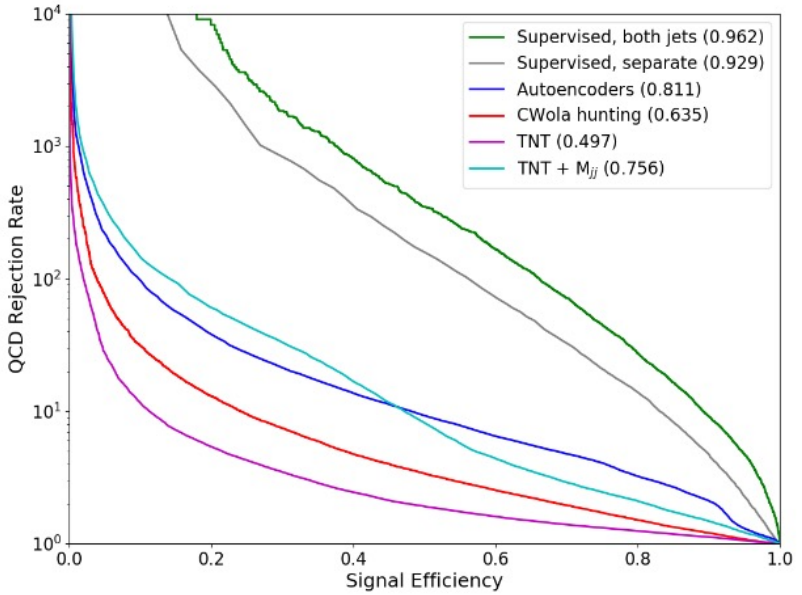
- TNT is a multistep approach to increasing signal fraction in a subsample for weak supervision training
  - It is also specifically for **events with two objects** (think of a dijet event), where **both** of the objects in the event **are anomalous** in some way (think of a process with A->BC, where A, B, and C are all not in the SM)
  - I’ll focus on the dijet case here

Tag N' Train (TNT)

TNT [2002.12376]

- Step 1: train an autoencoder *for the objects* on sideband events (ie a jet-level autoencoder, not an event-level autoencoder)
- Step 2: Split SR events into two subsets of events, one where you'll apply the AE to jet 1, and the other where you'll apply the AE to jet 2
- Step 3: After applying the AE to the selected jet in the event, you can get subsets of the subsets that are more signal like and more background like
- Step 4: combine the two signal-like subsets together and the two background like subsets together
- Step 5: train a classifier to distinguish jets in the signal-like combined set from jets in the background-like combined set
- Step 6: anomaly score can now be something like the product of the classifier score for the two jets in an event





In the case where signal event are rare, TNT can outperform CWoLa, and can outperform bare AEs at some working points (I'm focusing on TNT+Mjj here)

Contents

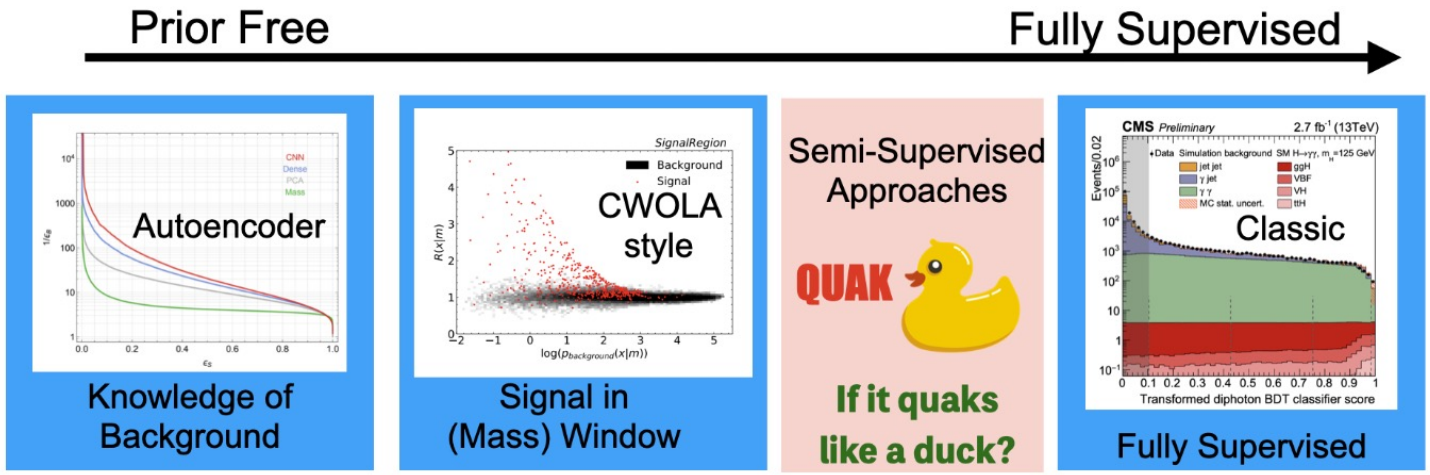
Semi-supervised searches

- Both unsupervised and weakly supervised methods are trained on data
  - The idea here is that classifier should learn what is *in* the data, rather than relying on (potentially incorrect) simulation or (potentially misguided) speculation about what signal should look like
  - However, both schemes become more sensitive when more signal is present (and are conversely *less* sensitive when *less* signal is present)
- Semi-supervised approach uses signal priors to maximize sensitivity while trying to maintain generality
  - Simulation may not be perfect, but it is generally pretty good

QUAK

QUAK [2011.03550]
MIT Masters's thesis from M. Yunus

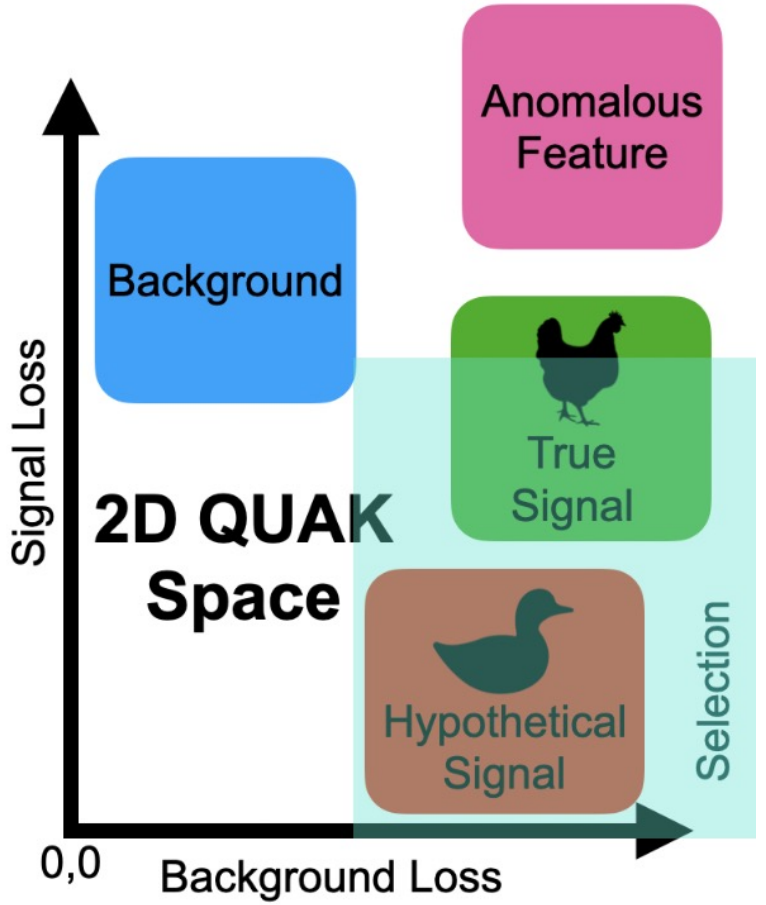
- "Quasi-Anomalous Knowledge" (QUAK) method has been developed for HEP
- QUAK functions almost as a branch of unsupervised techniques, but with the goal of avoiding selecting meaningless anomalous events (e.g. detector glitches)



QUAK

QUAK [2011.03550]
MIT Masters's thesis from M. Yunus

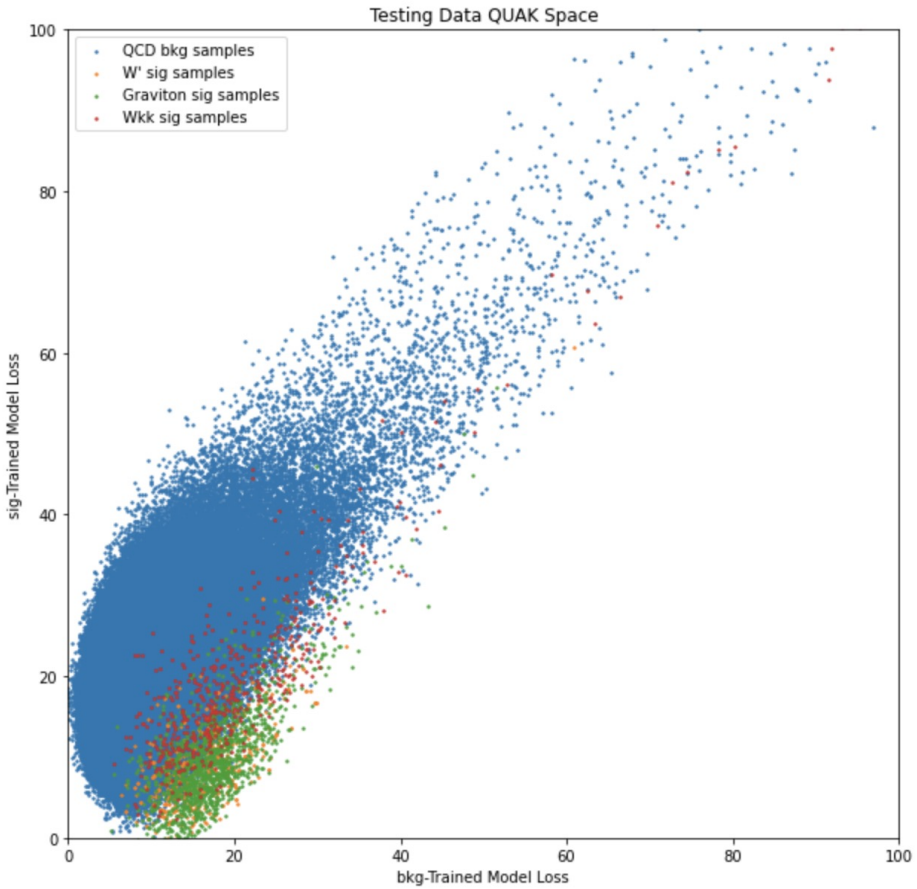
- QUAK works by constructing a "QUAK Space"
- For a typical implantation, you would train at least two normalizing flows (or VAEs)
  - One flow can be trained on background (either simulation or data). Anomalous events will have low likelihood (or high loss if using a VAE)
  - The other flow would be trained on a signal model simulation or combination of signals. Here, signal-like events will have higher likelihood (or low VAE loss)
  - You can do higher dimensional QUAK spaces as well, though it's unclear how useful that is
- Events populate the QUAK space with coordinates: (background loss, signal loss)
  - Background events should have low background loss and high signal loss
  - Signal events should have high background loss and low signal loss
  - "Weird events", like detector glitches, should have high background loss and high signal loss



QUAK

QUAK [2011.03550]
MIT Masters's thesis from M. Yunus

- Signal events should generally reside in the lower right part of QUAK space
- Two possible extremes:
  - If the signal axis is trained for a signal that is actually present in the data, then QUAK is about as sensitive as a dedicated search
  - If the signal axis is trained on a process that is completely unrelated to what's in the data, then QUAK roughly reduces to an unsupervised search (via the background axis)
- In general, QUAK is more flexible than a specific search
  - E.g. more robust to signal mismodelling than a directly trained NN

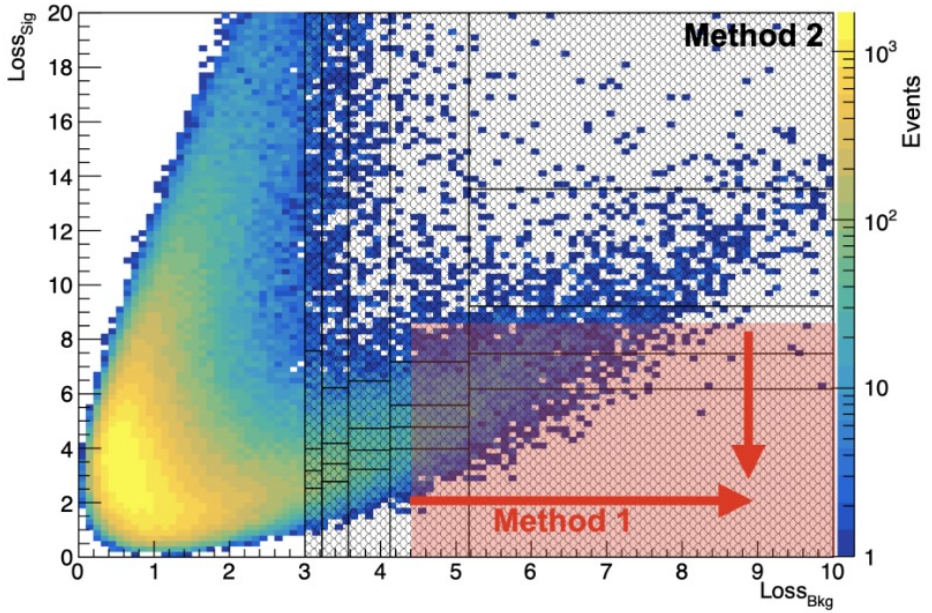


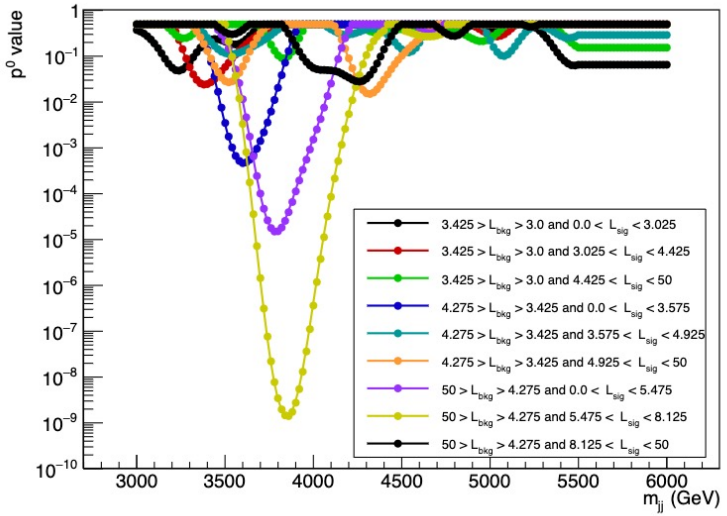
QUAK space showing distribution of different event classes. Signal axis flow was trained on a mixture of signal events from the 3 types presented

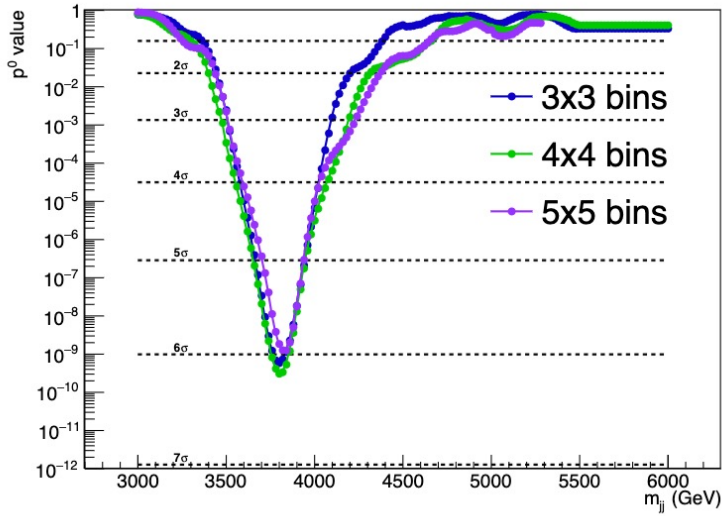
QUAK

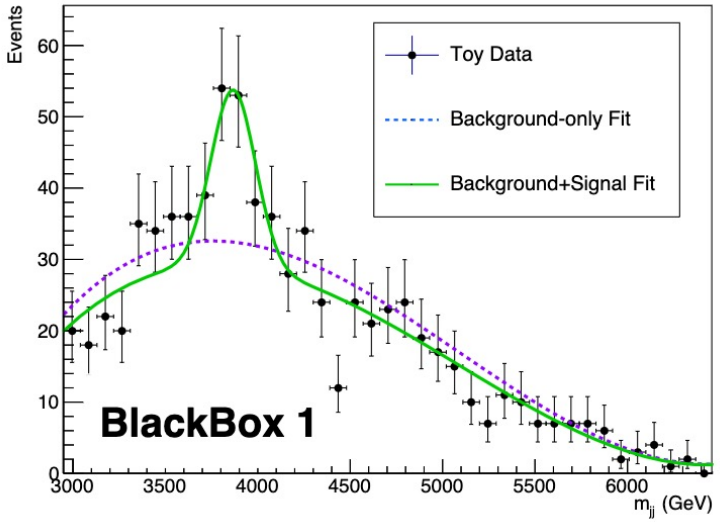
QUAK [2011.03550]
MIT Masters's thesis from M. Yunus

- A major trick to the algorithm is selecting events
  - QUAK has a "2D anomaly score", whereas most other algorithms' scores are in 1D
- For LHC Olympics challenge, 2 event selection methods were explored
  - Take events from a single rectangular bin and fit dijet mass spectrum with signal+background function
  - Use multiple non-overlapping bins to get independent spectra. Fit each, and combine p-values using Fisher's method
- Additional ideas: look for overdensities in QUAK space as a function of dijet mass, or cut along contours









Contents

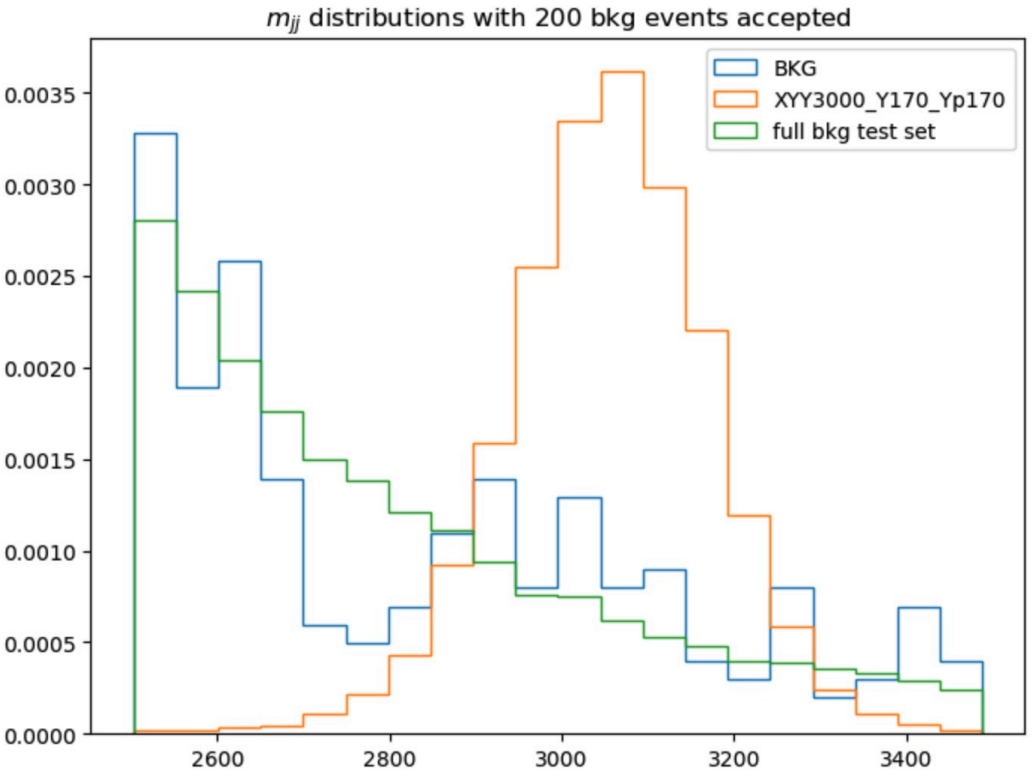
Discussion

- There isn't really a "best" algorithm for AD (at least right now)
  - Different algorithms have different strengths and weaknesses
  - Which algorithm you prefer can come down to preferences or philosophy
    - *How* model agnostic do you want to be?
    - E.g. from today, QUAK is *least* model agnostic, and there are some middle ground algorithms like TNT, that assume an A->BC topology with A, B, *and* C anomalous

Discussion

- A word of warning: it's important to be wary about finding signals that aren't really in the data
  - If a classifier "learns" the idea that something anomalous should be at a given mass, then *background* events near that mass can preferentially pass a classification cut, creating a false resonance
  - This is often referred to as "sculpting"
- Extensive validation is usually required to confirm robustness

Courtesy of Sam Bright-Thonney



E.g. if you train a NN to distinguish QCD events from a signal process with a resonance at 3 TeV, background events with mass near 3 TeV will preferentially pass a NN cut

Discussion

- Lastly, I want to highlight the fact that I mostly presented the algorithms here with a broad brush so that you get an idea of approaches to AD that people have come up with
  - There are many additional algorithms out there
- Also, for every algorithm, there are technical details that I glossed over
  - Selection and preprocessing of input features
  - Particulars of event preselection and cutting strategy
  - Exact architectures used
  - Algorithm validations
  - Etc
- Anomaly detection could probably fill up a whole semester
- If your research involves looking for rare processes with large backgrounds, hopefully you got a few ideas from this lecture!

Backup

LHC Olympics Results

Please refer to LHCO paper
I did not discuss every algorithm that made it to the LHCO

- LHC Olympics (LHCO) were styled as 3 black box datasets
  - Black Box 1 consisted of 834 W'->XY signal events with masses mW' = 3.823 TeV, mX = 732 GeV, and mY = 378 GeV
    - There had been an R&D dataset released with this same process (but with different masses)
  - Black Box 2 had *no* signal injected
  - For Black Box 3, there was a resonance at 4.2 TeV with two different decay modes

LHCO – BB 1

Please refer to LHCO paper

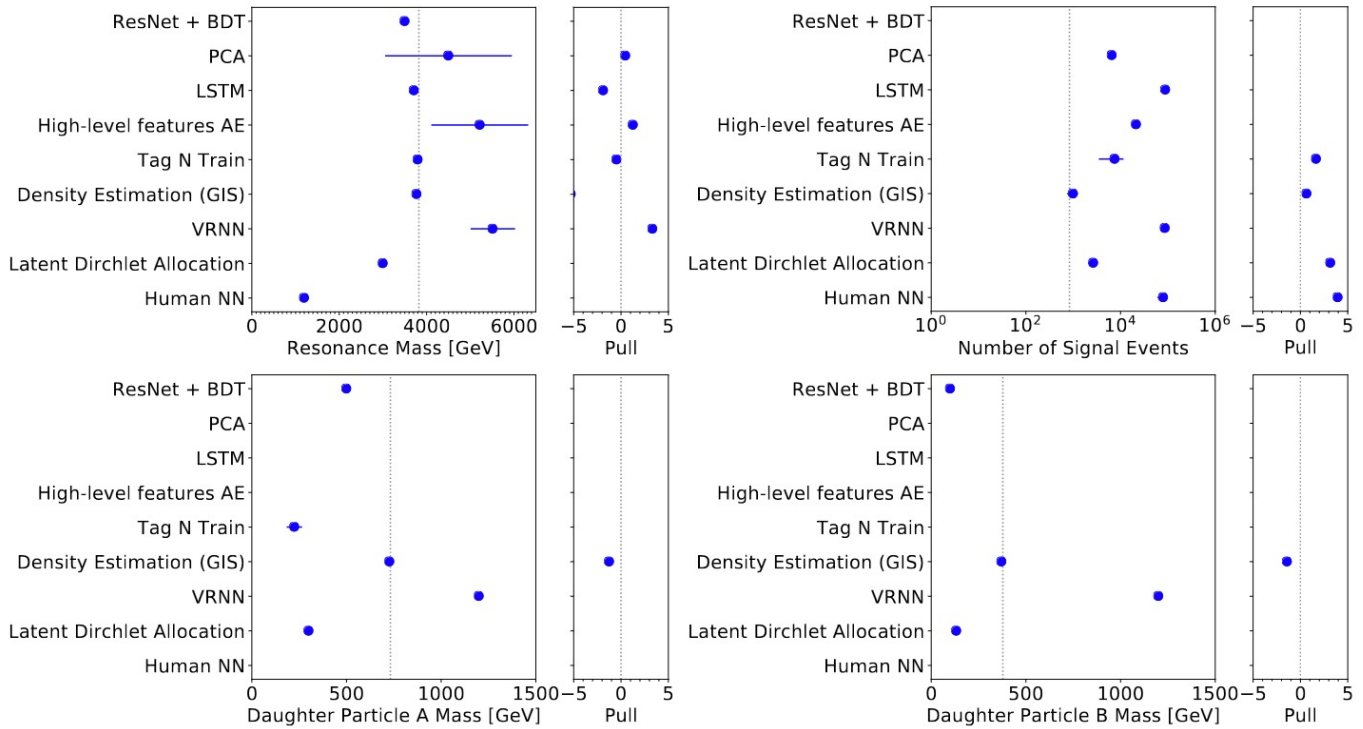


Figure 51. Results of unblinding the first black box. Shown are the predicted resonance mass (top left), the number of signal events (top right), the mass of the first daughter particle (bottom left), and the mass of the second daughter particle (bottom right). Horizontal bars indicate the uncertainty (only if provided by the submitting groups). In a smaller panel the pull (answer-true)/uncertainty is given. Descriptions of the tested models are provided in the text.

LHCO – BB 2

Please refer to LHCO paper

- Remember there was no signal
- Results
  - PCA: resonance at 4.8 TeV
  - VRNN: resonance at 4.2 TeV
  - Embedding clustering: resonance at 4.6 TeV
  - QUAK: resonance at 5 TeV
  - LDA: no resonance

LHCO – BB 3

Please refer to LHCO paper

- Remember resonance at 4.2 TeV with two different decay modes
- Results
  - PCA: “a resonance decaying to invisible particles”
  - VRNN: no resonance
  - Embedding clustering: resonance at 3.1 TeV
  - QUAK: resonance at 5 and 5.5 TeV
  - LDA: resonance with mass between 5.4 and 6.4 TeV

LHCO

- Several algorithms did well on the first black box
- The other black boxes two were much trickier
  - Having the R&D dataset probably helped for BB1 ;)