Anomaly Detection in HEP

A3D3 Kick-off Meeting, Nov 7, 2021

Sang Eon Park, Phil Harris
Patrick McCormack, Mikaeel Yunus (MIT)

Motivation

Strong motivation that there must be Beyond Standard Model (BSM) physics

Dark matter/energy, origin of neutrino mass, and so on

No BSM physics found at the LHC with searches targeting specific models

After the discovery of the Higgs (2012), effort focused on model specific searches

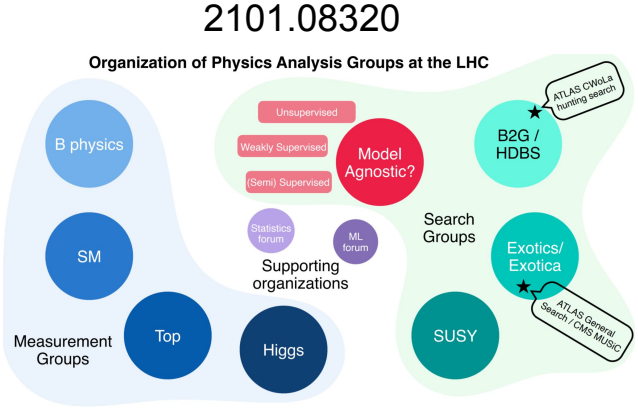
Supersymmetry, extra dimensions, extended Higgs and many more, but no convincing evidence yet

We need to start thinking about **model agnostic searches** : “anomaly detection”

Anomaly detection at colliders = searches not targeting specific models

Use ideas from “anomaly detection” in applied ML, but have to solve problems unique to HEP

2101.08320



Challenges of Anomaly Detection

At the moment there is no way to generalize performance of each search method

Unlike model specific searches, usually **ROC is not enough**

No one evaluation metric can summarize the performance, since the target is truly unknown

Each method performs best in different scenarios

There's no fixed "test dataset"

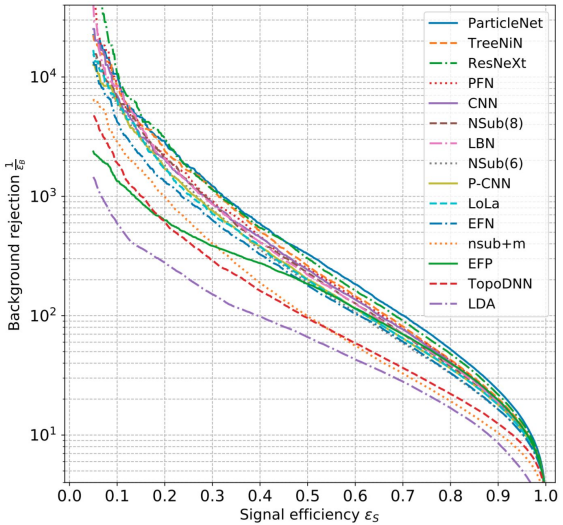
Performance will depend on which signal model we test on, signal properties, the region of phase space the true signal lies in, S/B ratio, etc

Also depends on which figure of merit we choose (ROC/PR/Significance)

No one approach is strictly better than every other approach

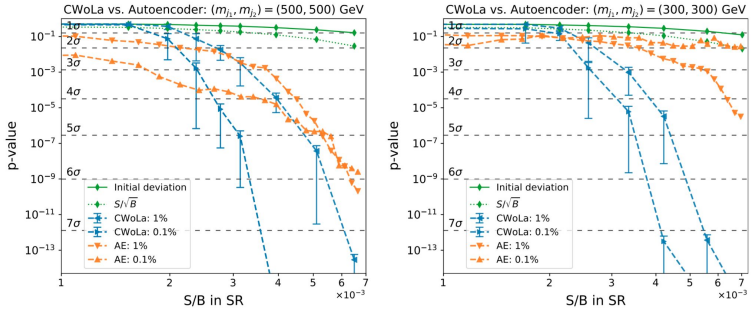
Methods are complementary to each other

No clear winner in anomaly detection, need to try a variety of methods!



1902.09914

2104.02092



Community Wide Effort

We need to have multiple strategies

Different methods will have different sensitivities to different **regions of search space** and **S/B ratio**

Big **community wide efforts** to come up with a wide variety of search strategies

LHC Olympics (2020) = challenge with 3 “black box” datasets with hidden embedded signal

- 18 algorithms submitted (2101.08320)

DarkMachines Challenge (2021) = challenge to detect a wide ensemble of signals

- 16 algorithms submitted (2105.14027)

34 algorithms + a lot more, all with unique approaches!

A wide variety of ideas, guiding principles, choice of input features and training set

Different ways to categorize these approaches

What does the main algorithm do?

Dimensionality Reduction / Density Estimation / Overdensity / Clustering ...

PCA, AE, VAE, flows, deep sets, noisy labels, isolation forest, k-means, BDT

How much signal information is used?

Unsupervised / Weakly Supervised / Semi-Supervised

What kind of input is used?

Images (CNN) / Particles (RNN, graphs) / Processed Jet Variables (MLPs)

Is it trained on data / simulation?

Density estimation based searches

Density estimation based methods are **most common** - 26/34 methods

Many ways to do it: deep AEs, VAEs, normalizing flows, kernel methods

Here you want the model to estimate the probability distribution of high dimensional data

Typically: train (on background events) a model to learn the distribution of the background, then in testing select events with low p(background)

Can we use information from what we already know?

In dijet searches, with density estimation based approaches you train a model to learn the distribution of background (QCD) and choose events with low p(background)

Unsupervised searches use only information of background (SM physics)

Can we incorporate certain aspects of known physics into the search?

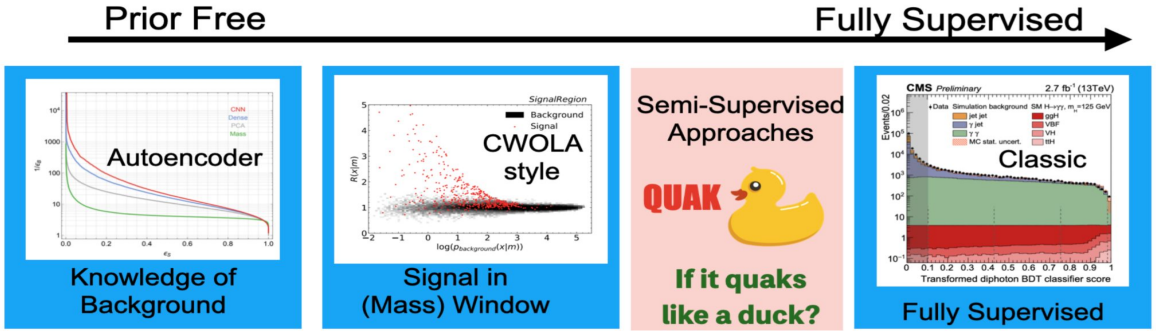
Incorporating signal models into the search can improve the sensitivity

Semi-supervised methods try to do this (Many ways to do this!)

Choice of loss metric, training on ensemble, and many more

Key is to **preserve model independence** while incorporating some aspect of signal data

2011.03550

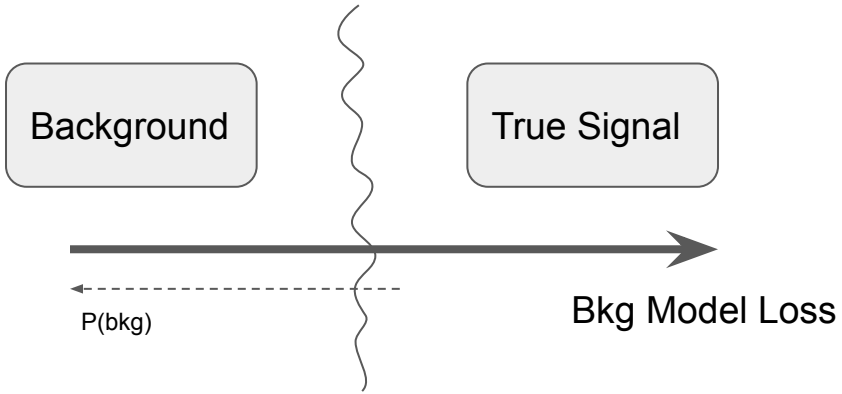


One way to do this - QUAK

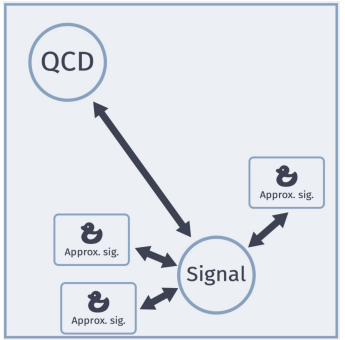
Quasi Anomalous Knowledge: Searching for new physics with embedded knowledge 2011.03550

Conventional density estimation based search

learning latent representation of background (train on background)

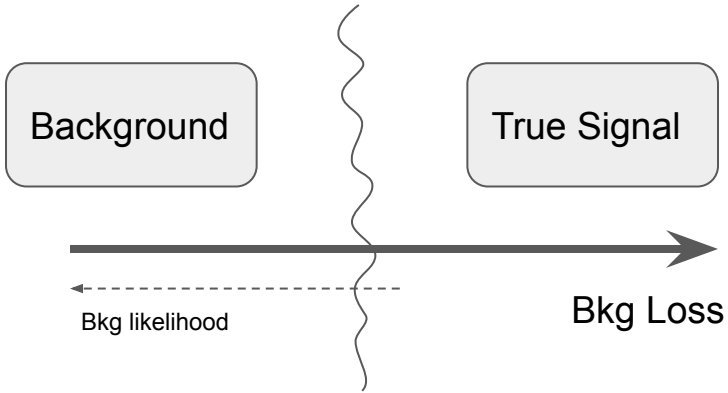


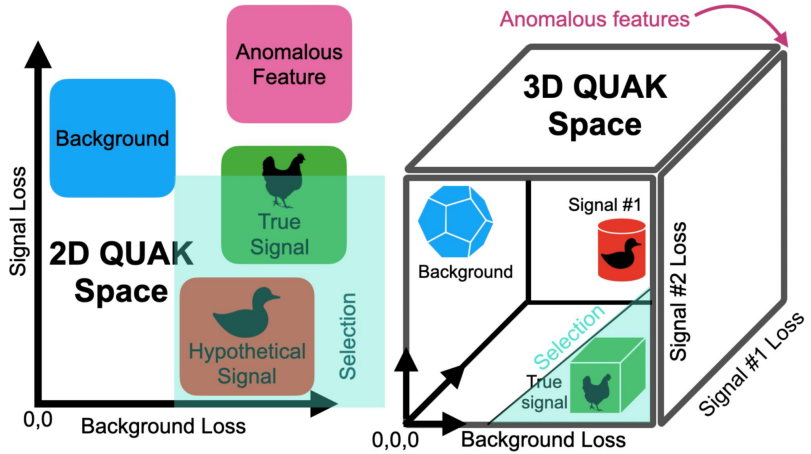
Core Idea of QUAK



QUAK adds another axis to this search by training multiple models on approximate hypothetical signal priors

This gives us control we didn't have in 1D,can separate out more categories





2011.03550

We also checked model independence and performed comparison with supervised methods

How would it be done in practice?

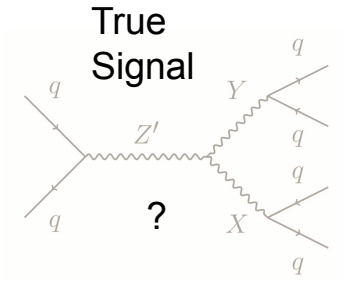
LHC Olympics dijet anomaly search

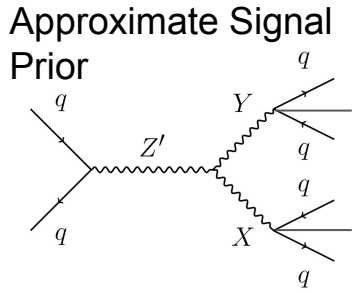
Build 2 dimensional loss space

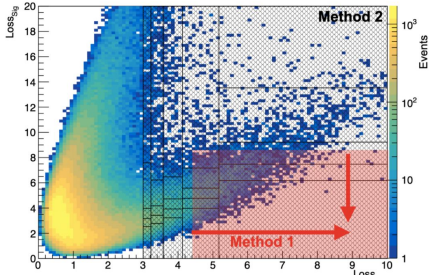
- One with background + One choice of signal prior
- Do the full pvalue scan to avoid look elsewhere effect and to get maximum significance

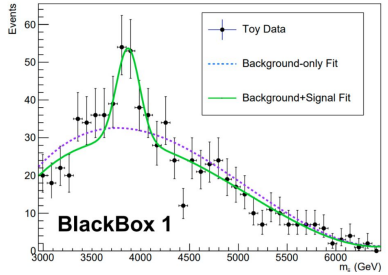
More general dijet anomaly search

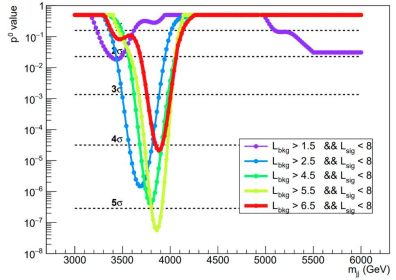
- Can be more ambitious, add more signal priors
- (masses, pronginess etc.) and build higher dimensional QUAK space

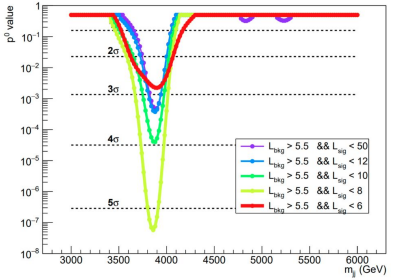












Can we run these algorithms online?

Most algorithms published so far are designed for **offline analysis**

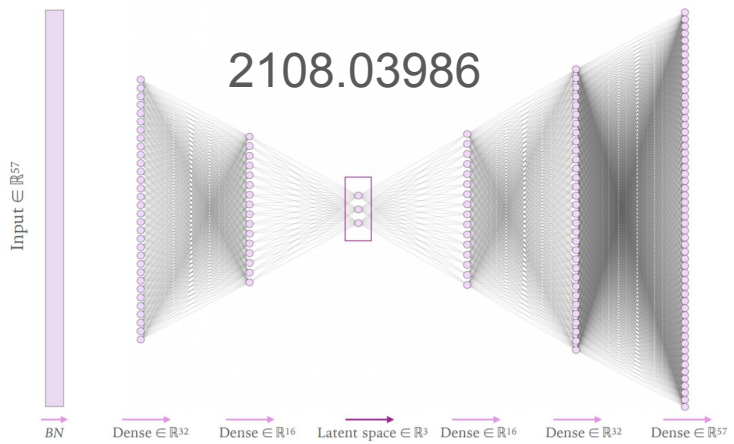
Algorithms should meet these two criteria to be run online:

1. Only look at data once
2. Be able to meet throughput and latency constraints

Autoencoder based methods are good candidates to be used in online setting

Hardware acceleration: deep autoencoder can be put on FPGAs, meeting L1 trigger restraints (2108.03986)

However there are many more possibilities!



Outlook

Anomaly detection

- **Need many ideas** : Wide community efforts to come up with diversity of ideas
- Big sensitivity improvement might come from incorporating physics knowledge into the search (semi-supervised approaches)

Lots of work needs to be done

- Development of **online strategies**
- Development of different evaluation metrics

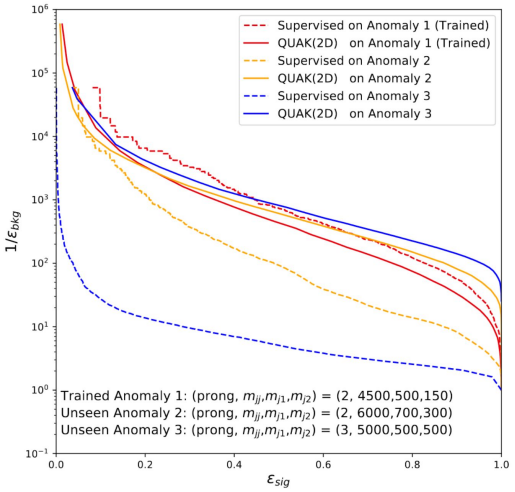
Backup

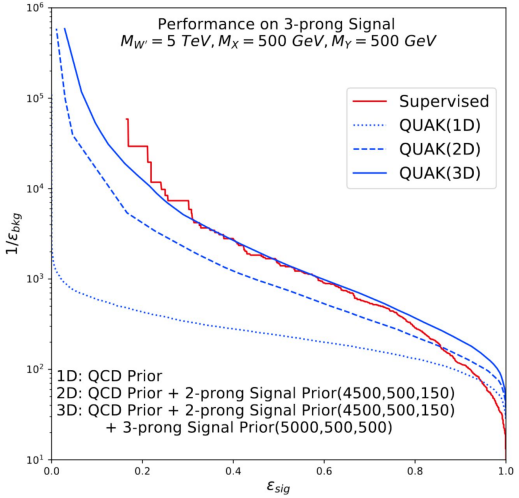
What we show in the paper

1. These approximate priors don't have to be accurate to help with the search

   model-independence

2. We can approach supervised classifier's performance if we have accurate prior





In the space of distance metric

What we do(in context of QUAK):

We train new generative model for each signal priors, Run the same training multiple times

Each event evaluated with multiple distance(in our case, likelihood)