Based on CMS-NOTE-2023-013

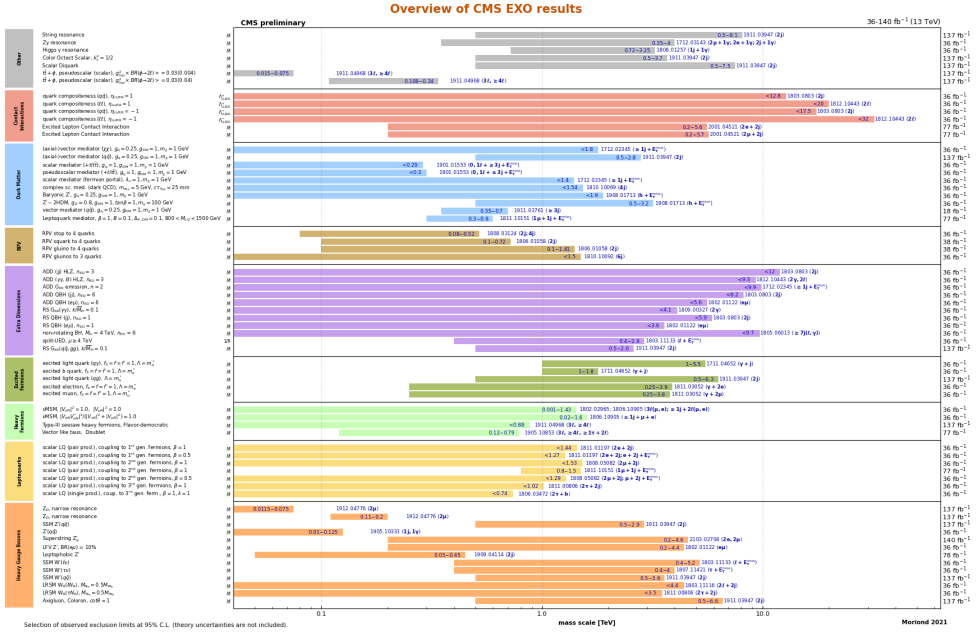
Techniques for ML-based Model Agnostic Searches in CMS

Oz Amram
Dec. 15th, 2023
US LUA Meeting

Fermilab

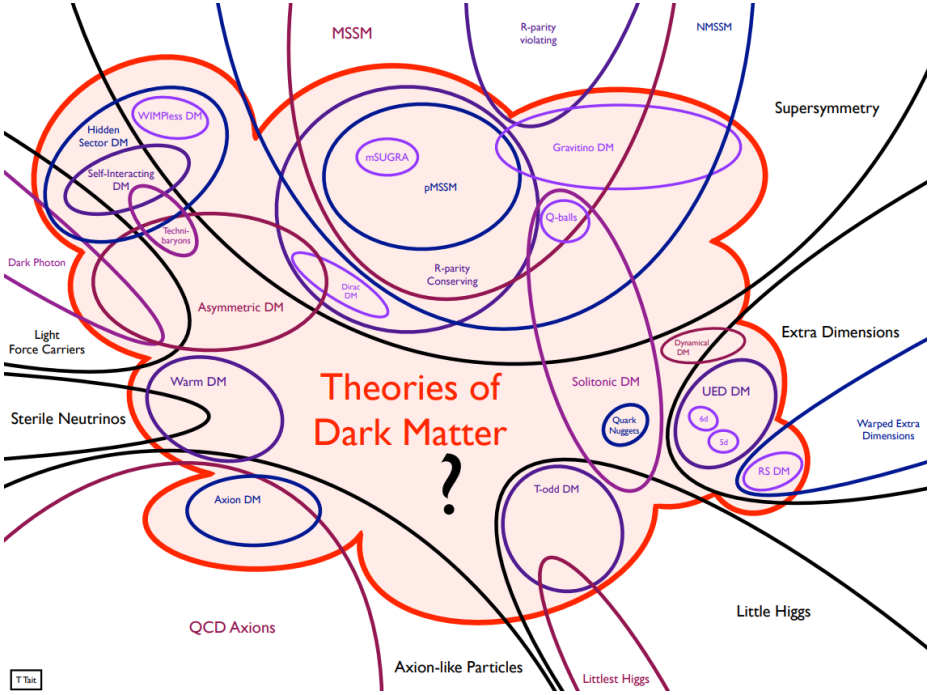
Why Model Agnostic Searches?

Already have a vibrant search program at the LHC

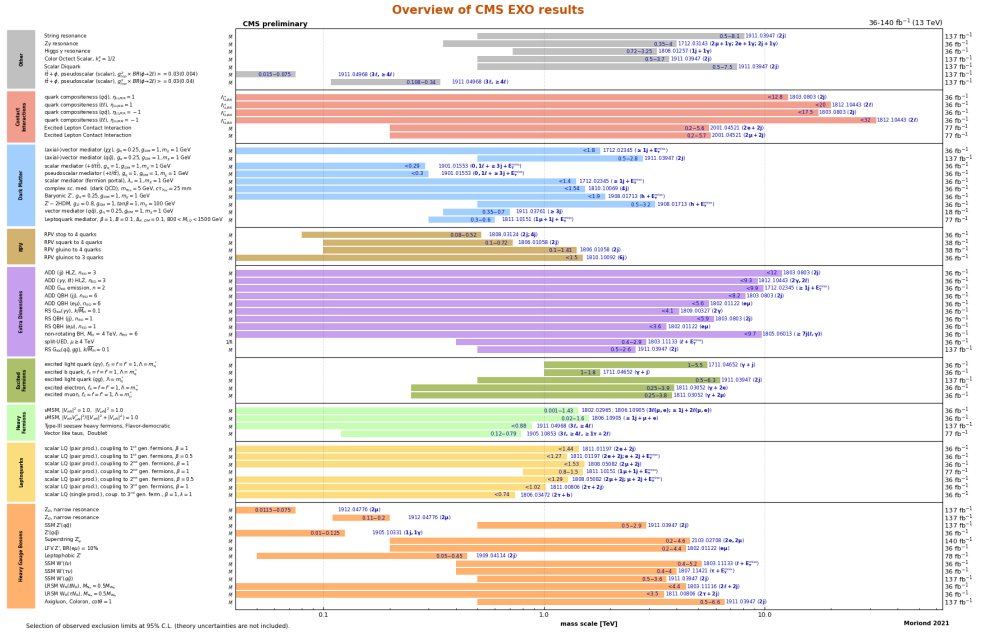


But no conclusive signs of new physics yet…

Why Model Agnostic Searches?



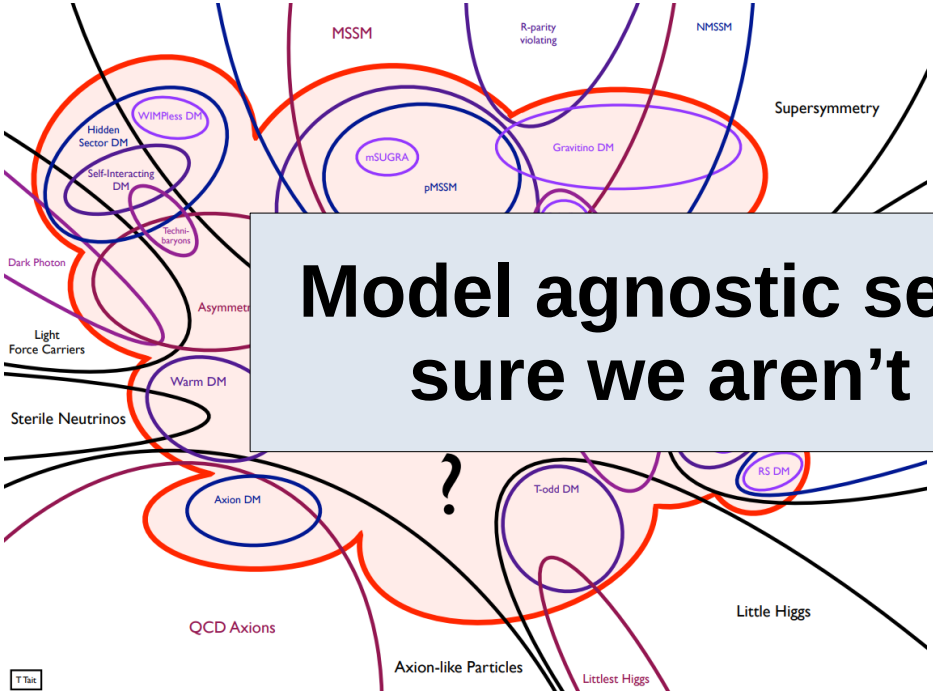
>

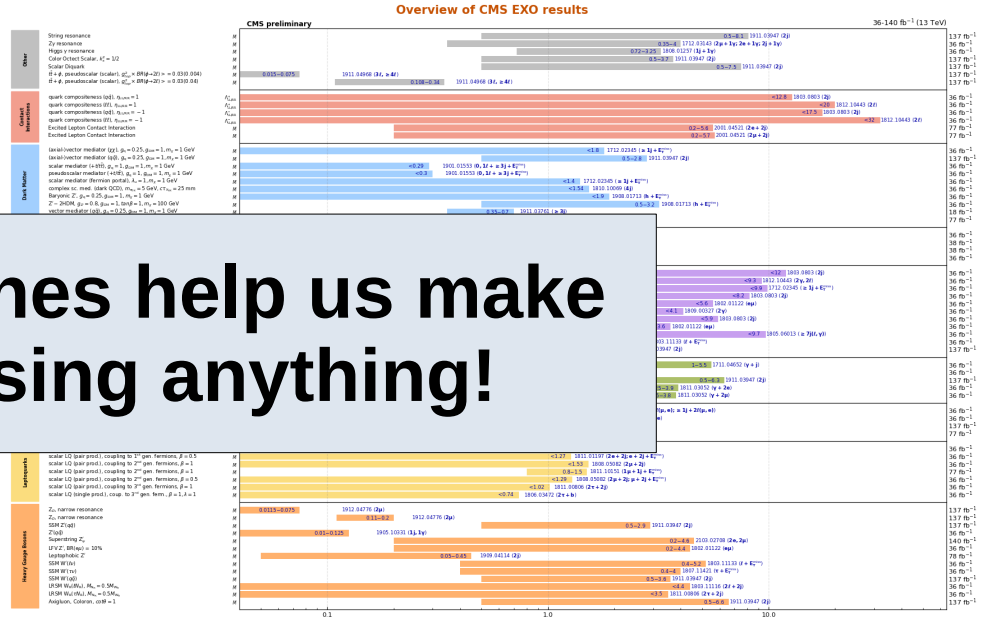


But theory space will always be larger our dedicated search coverage!

DM graphic from Tim Tait

Why Model Agnostic Searches?

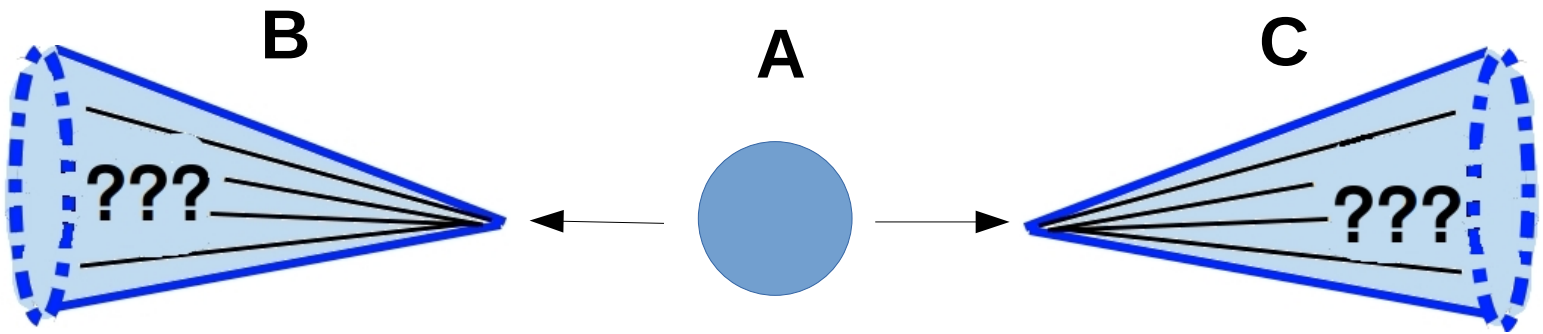




Model agnostic searches help us make sure we aren't missing anything!

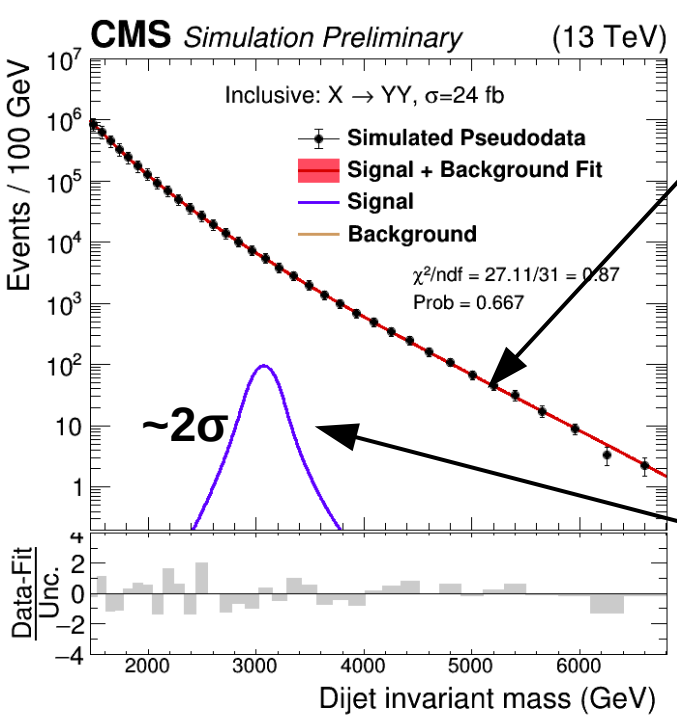
DM graphic stolen from Tim Tait

Dijet Anomaly Search



- Targeting **A→BC** topology
  - Heavy A → both B and C contained in large-radius jets
- Huge background from QCD → apply anomaly detection
  - Employing **data-driven machine learning** methods to reject QCD and select 'anomalous' jets
- Goal is **broad sensitivity** to many different kinds of B and C with different kinds of substructure

Standard Bump Hunt

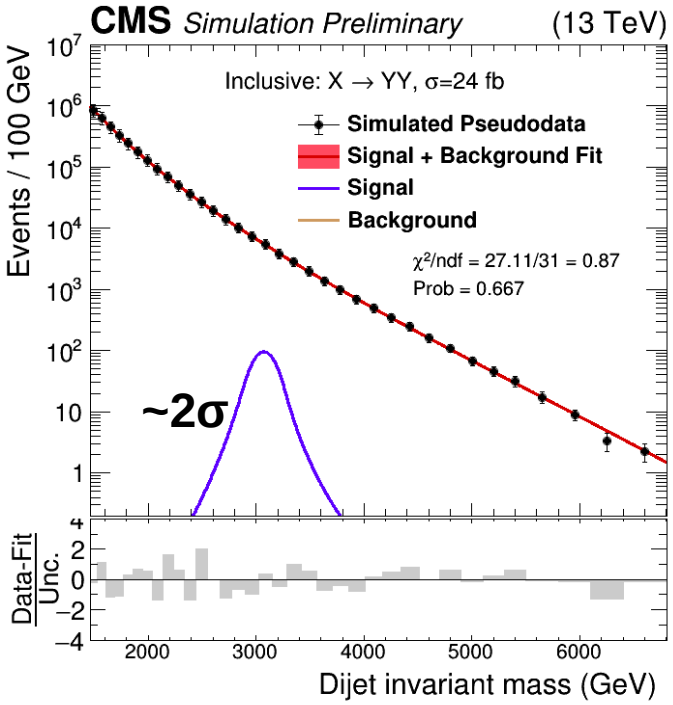


Background fit with standard analytic functions

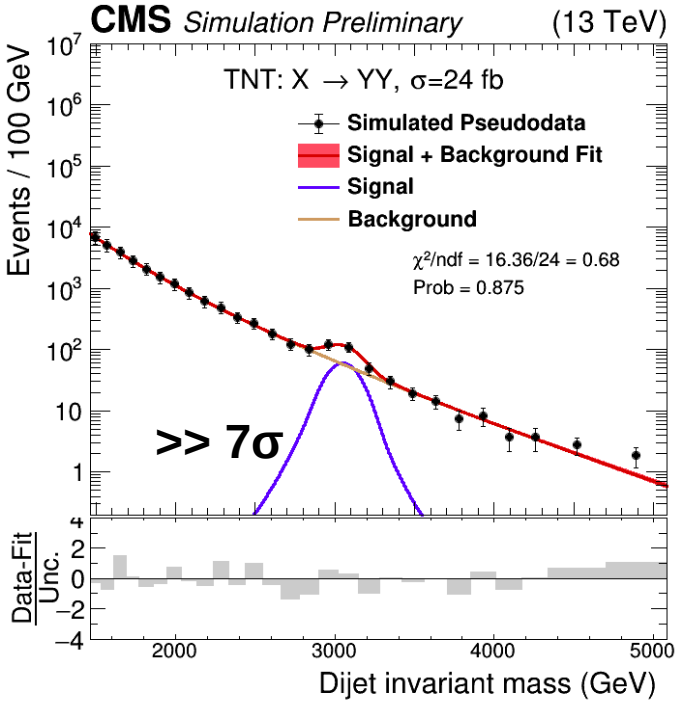
Double Crystal Ball signal shape

Without any substructure cuts → Signal swamped by QCD background…

+Anomaly Detection

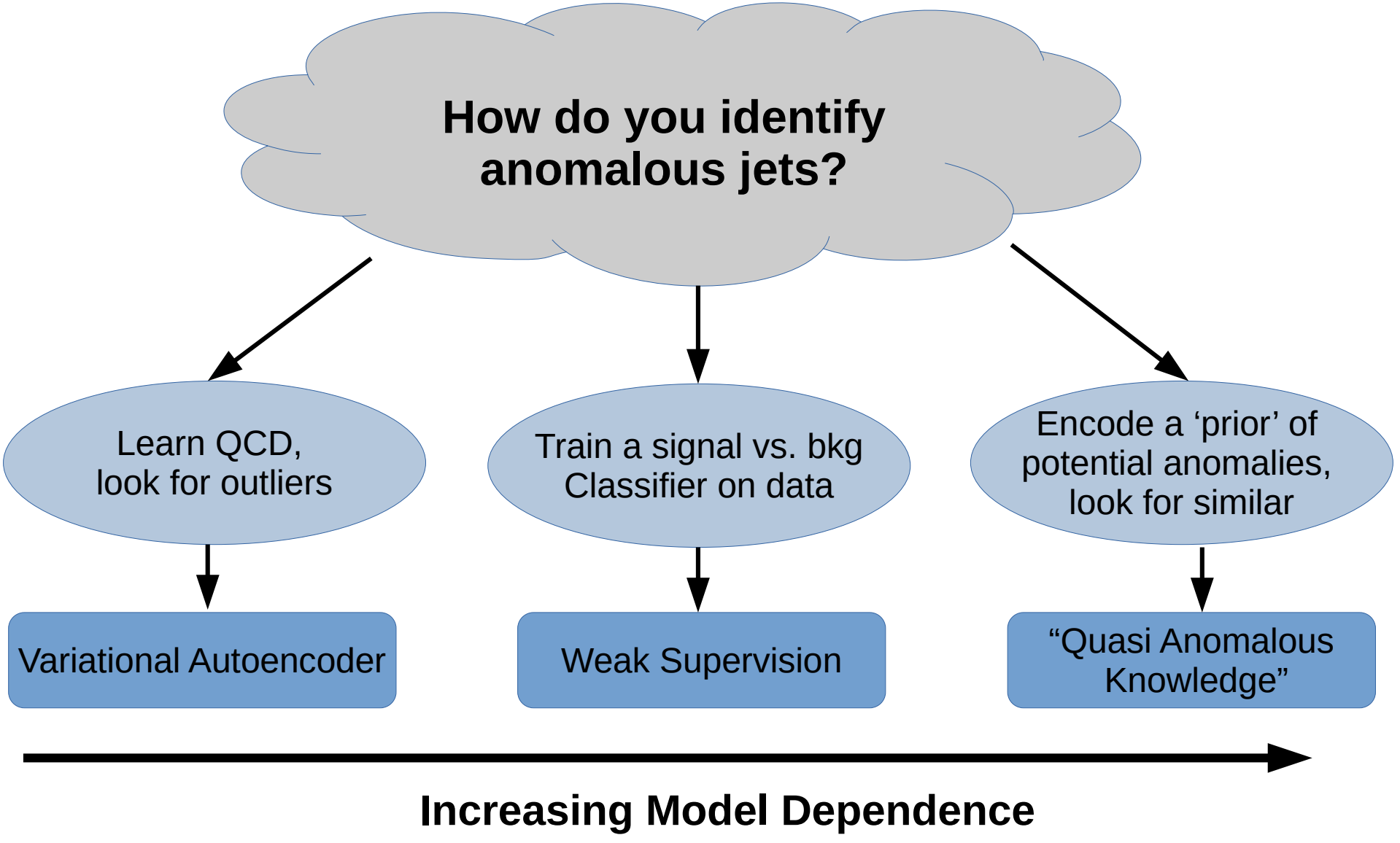


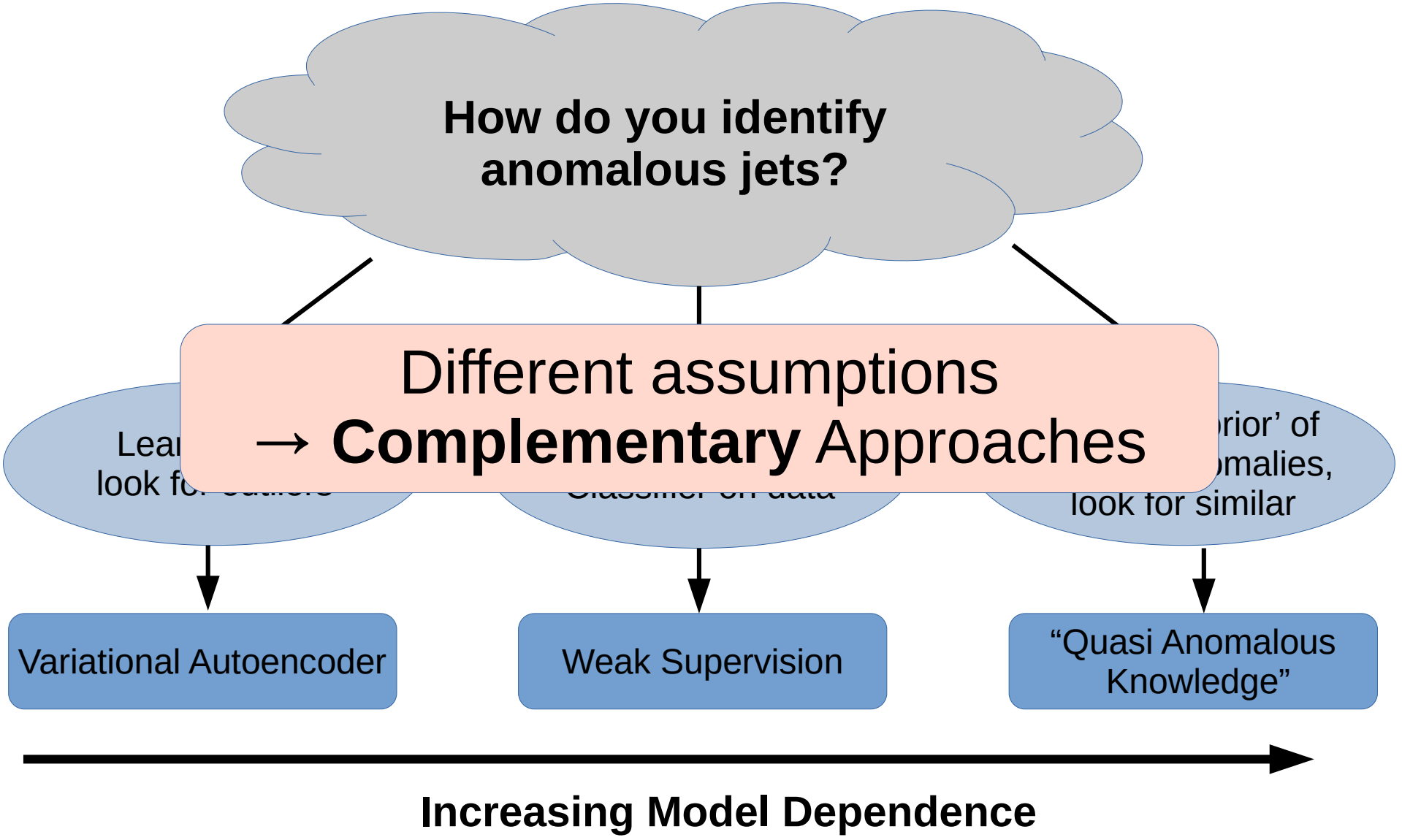
Cut on anomaly score



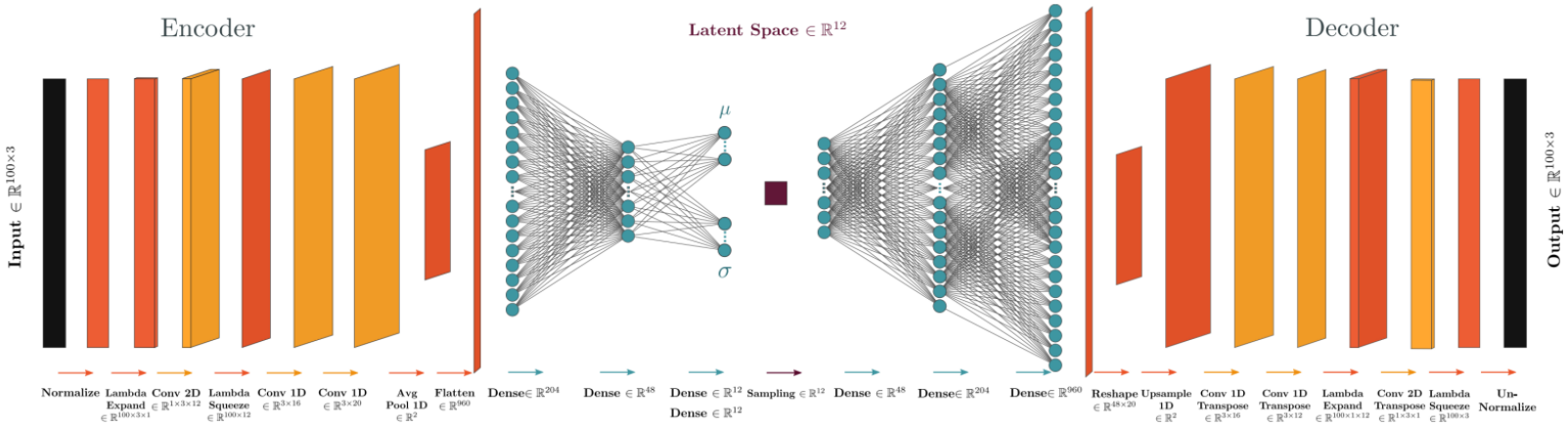
Without any substructure cuts →
Signal swamped by QCD background…

Anomaly detection finds hidden resonance!


How do you identify anomalous jets?
Learn QCD, look for outliers
Train a signal vs. bkg Classifier on data
Encode a ‘prior’ of potential anomalies, look for similar
Variational Autoencoder
Weak Supervision
“Quasi Anomalous Knowledge”
Increasing Model Dependence


How do you identify anomalous jets?
Different assumptions
→ Complementary Approaches
look for similar
Variational Autoencoder
Weak Supervision
“Quasi Anomalous Knowledge”
Increasing Model Dependence

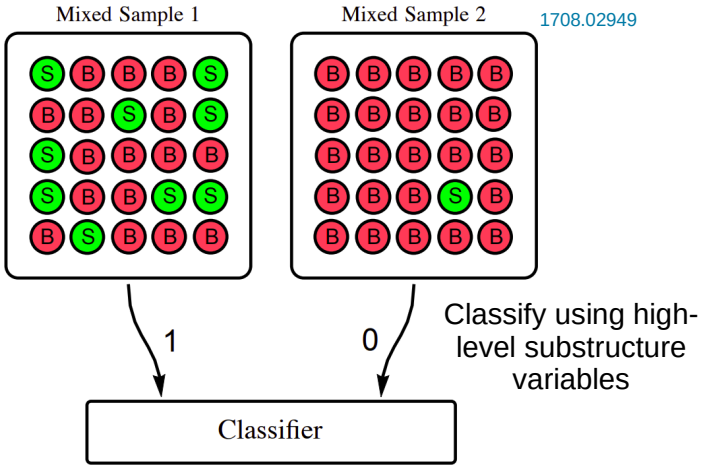
Variational Autoencoder (VAE)

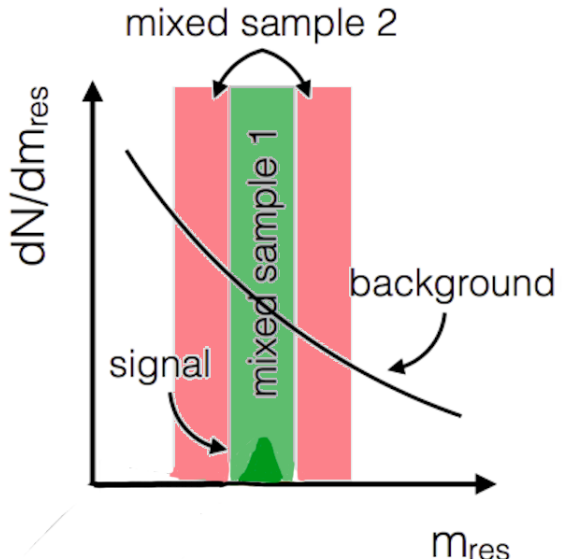


- Take in up to 100 constituents of the jet
- Learn to compress & decompress using sample of background events in data
  - Network won't learn how to do this for 'anomalous events'
- Use difference between original & reconstructed as an anomaly score
- Quantile Regression (QR) used to ensure no sculpting of Mjj shape

Weak Supervision

- Train a classifier between **signal-rich** and **background-rich** mixed samples
  - → Learns to tag signal vs. bkg
- Three methods to construct mixed samples, all assume a narrow resonance
  - **CWoLa** : take bkg. samples directly from sideband events
  - **CATHODE**: bkg. interpolated from sidebands
  - **Tag N' Train**: enrich purity of anomalies before training by using autoencoder

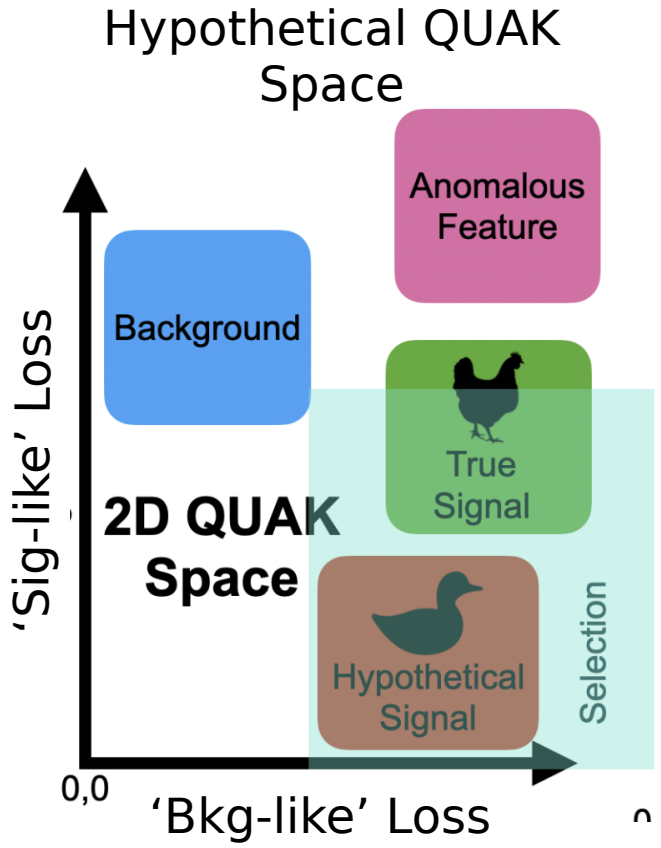




CWoLa: 1902.02634 / **CATHODE**: 2109.00546 / **TNT**: 2002.12376

Quasi Anomalous Knowledge (QUAK)

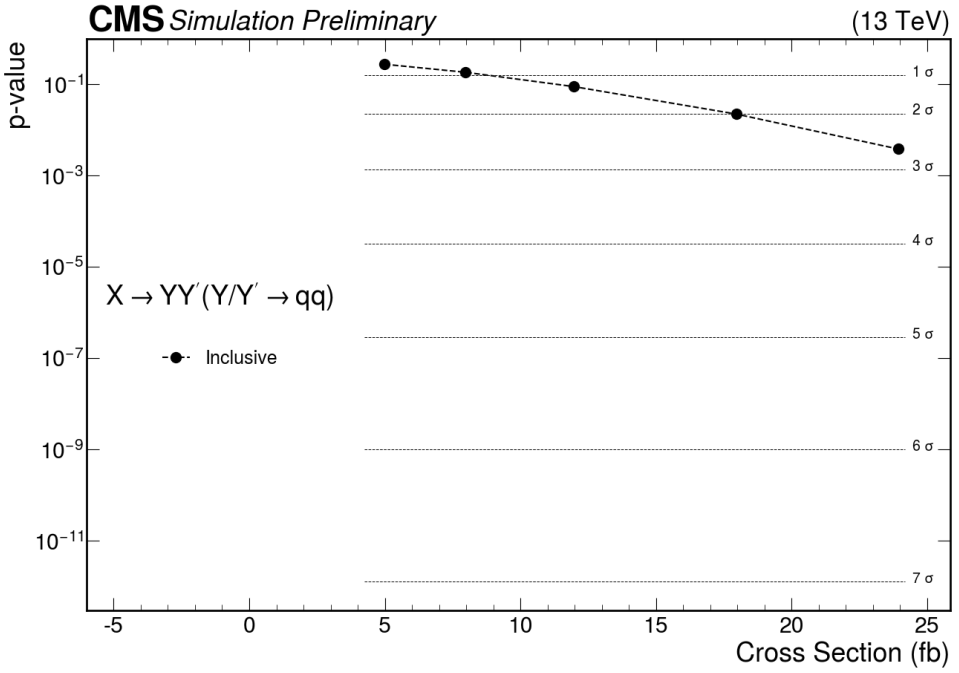
- **Hybrid approach** between fully model-indep. and standard search
- Idea: **encode a prior** on what a potential signal may look like
  - AE trained on a mixture of signal MC's
- Construct 'QUAK space':
  - Loss of signal AE vs bkg AE
- Select events with low sig loss and high bkg loss



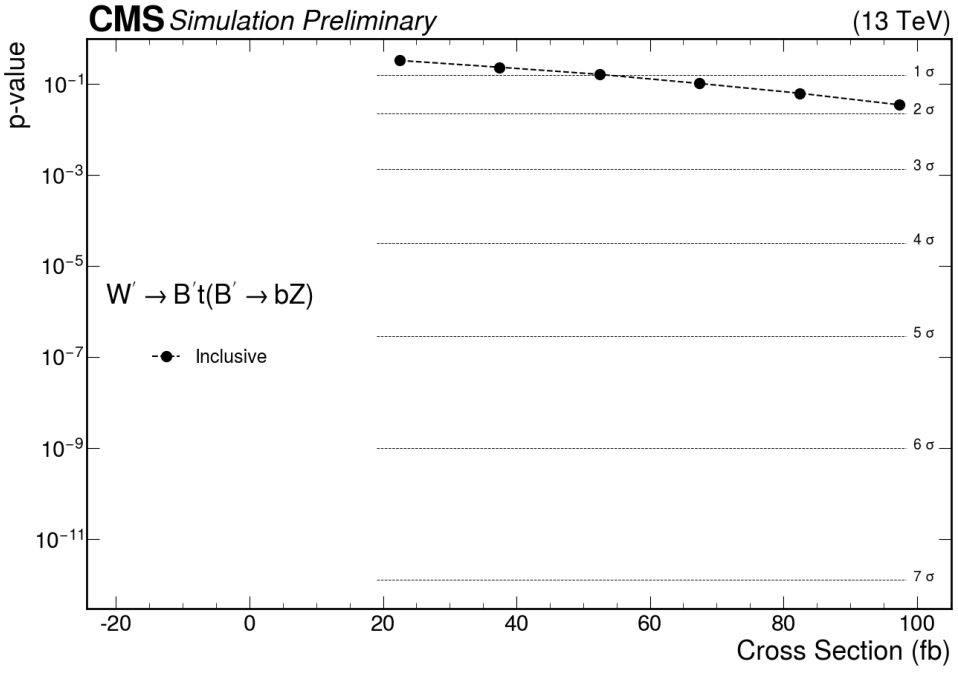
2011.03550

Sensitivity Study

2 Pronged Signal



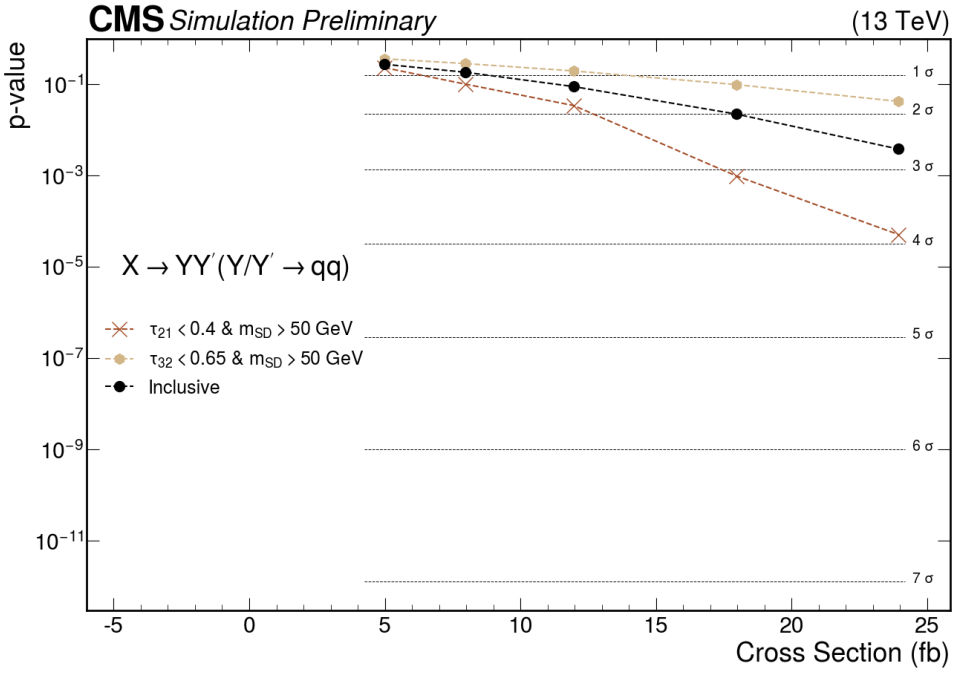
3 Pronged Signal



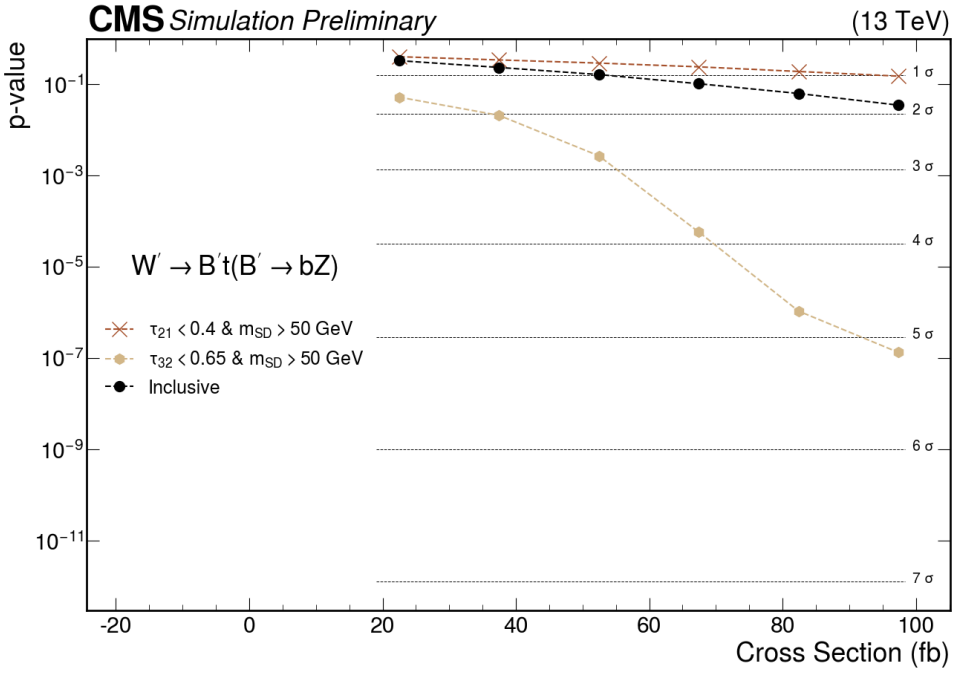
Inclusive analysis (no substructure cuts) sees only “hints”

Sensitivity Study

2 Pronged Signal



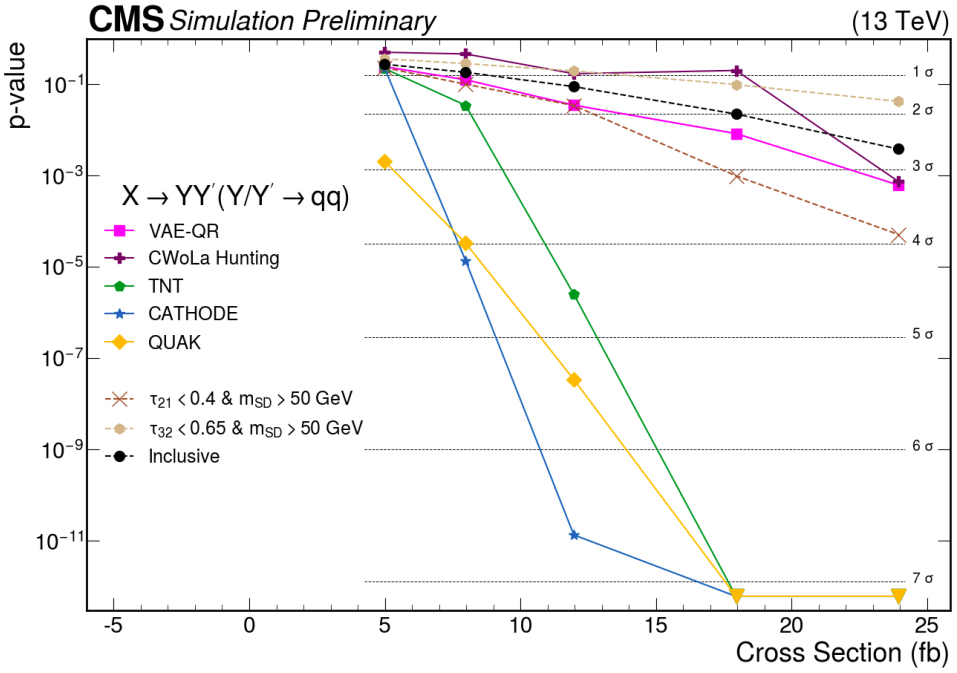
3 Pronged Signal



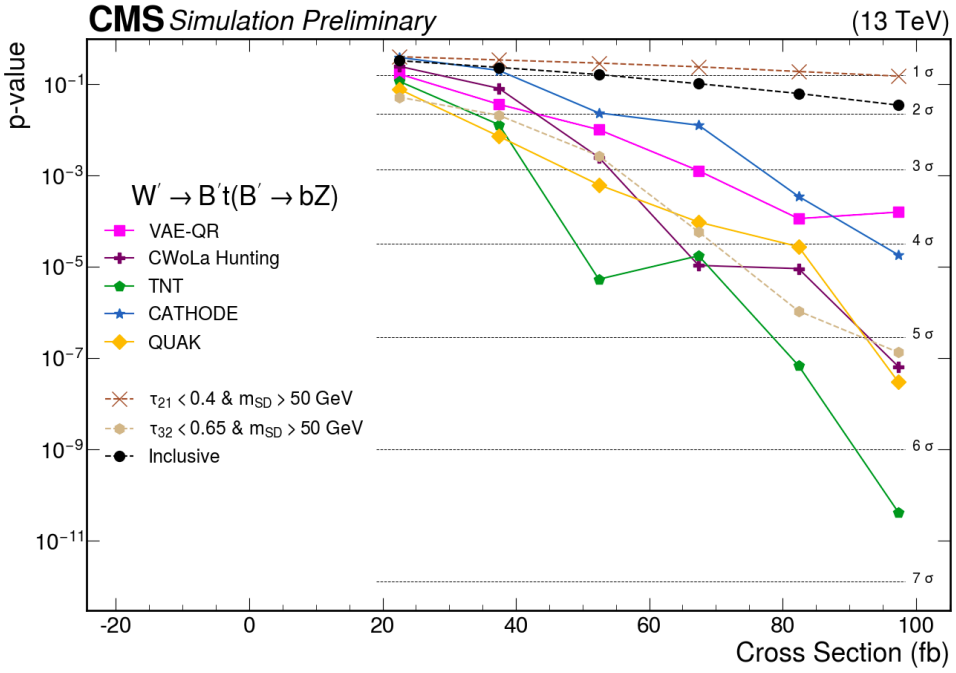
Traditional substructure cuts enhance sensitivity for a specific model, but not others

Sensitivity Study

2 Pronged Signal



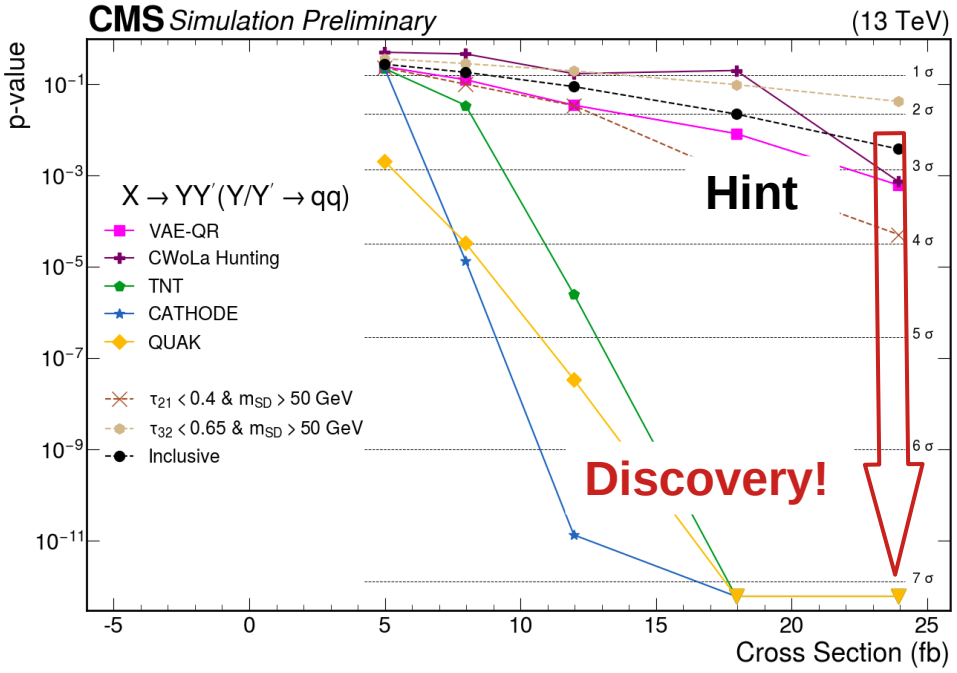
3 Pronged Signal



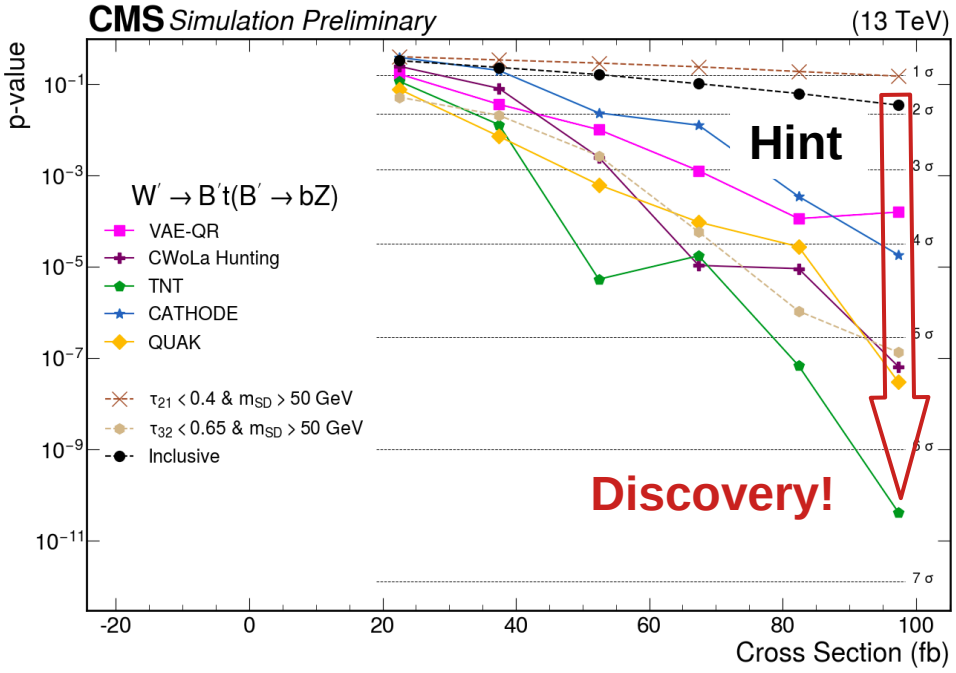
Anomaly detection enhances sensitivity for many models at once!

Anomaly Detection

2 Pronged Signal



3 Pronged Signal



Anomaly detection enhances sensitivity for many models at once!

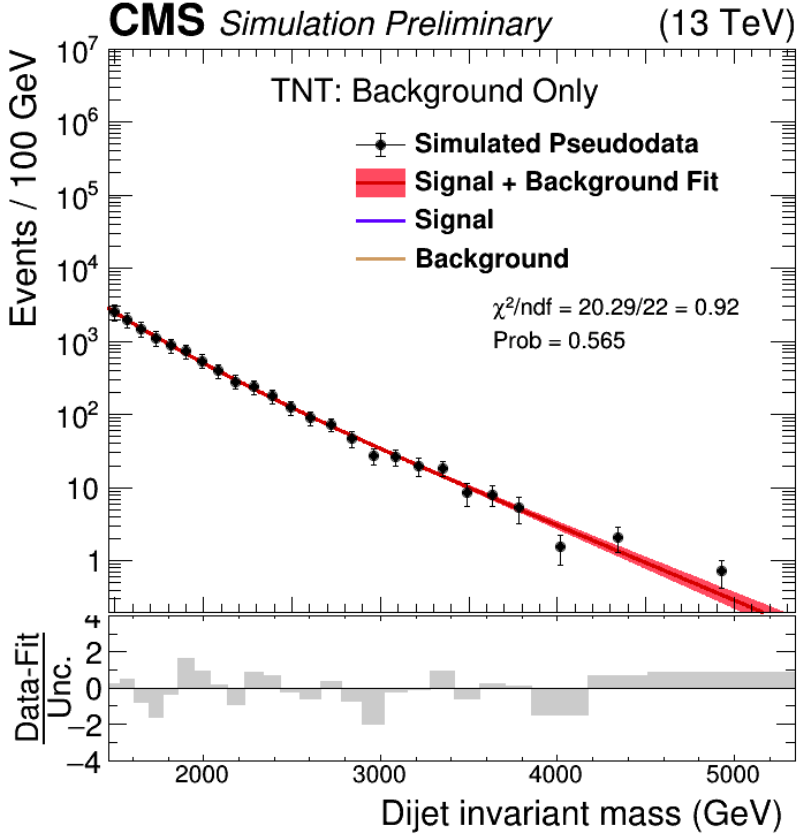
Conclusions

- CMS pursuing anomaly detection to enhance our search program
- Using multiple methods based on different philosophies → robust coverage
- Studies on simulation demonstrate enhanced discovery potential beyond traditional techniques
  - CDS Note (public): CMS-NOTE-2023-013
- Results on data being finalized, public in the **very** near future!
  - CADI Line (CMS internal): EXO-22-026

Backup

Mass Spectrum Bkg Only

- Decorrelation of anomaly score with Mjj
  - Crucial for weakly supervised methods
- Ensures no artificial bumps in the case of no signal

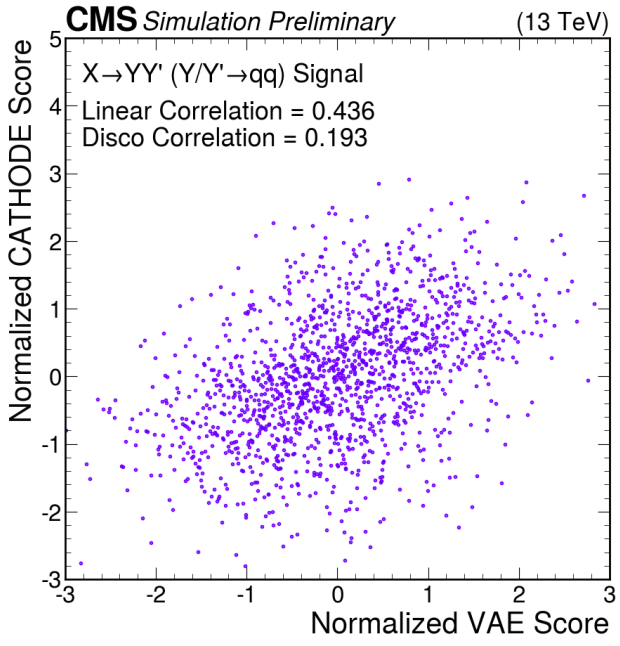


Anomaly Score Correlations

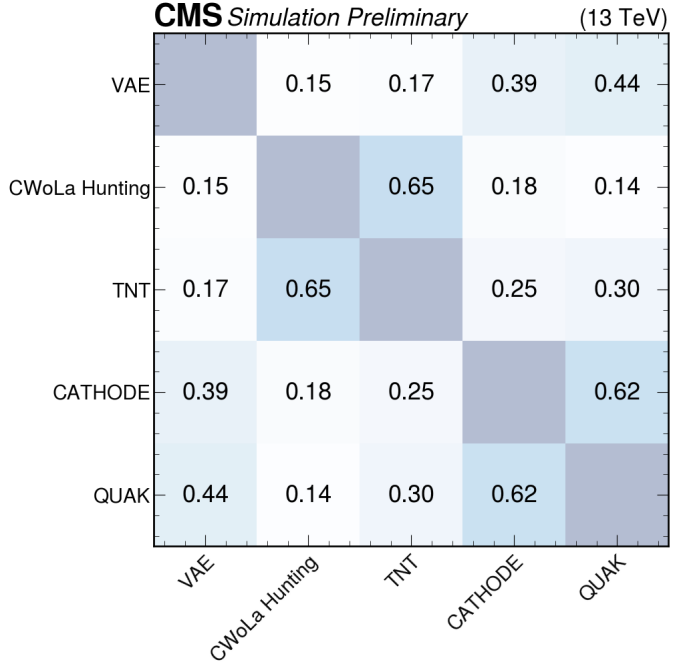
- Do all methods agree on anomalies?

  → Check correlation of anomaly scores

Single Comparison



Linear Correlation Summary



- Relatively low correlation of anomaly scores between methods (~0.5 or less)
- → Different methods are complementary

Input Features

- What features by each method are used defines what anomalies they could be sensitive to
- VAE: Uses p_T,η ,φ of all PF candidates inside the jet

  → quite ‘model agnostic’
- Weakly supervised & QUAK: uses typical jet substructure observables
  - soft-drop mass, n-subjettiness ratios, b-tagging info, lepton subjet fraction

Tagging Uncertainties

- For the analysis, need uncertainties on the signal tagging uncertainty to set limits
- Many of our signals don't have SM equivalents…
- → Developed new Lund Plane Reweighting method to correct MC QCD modeling & derive uncertainties
- Corrects density of splittings in MC per-prong
  - Validated to improve data/MC agreement on W and top events in data

CMS DP-2023/046