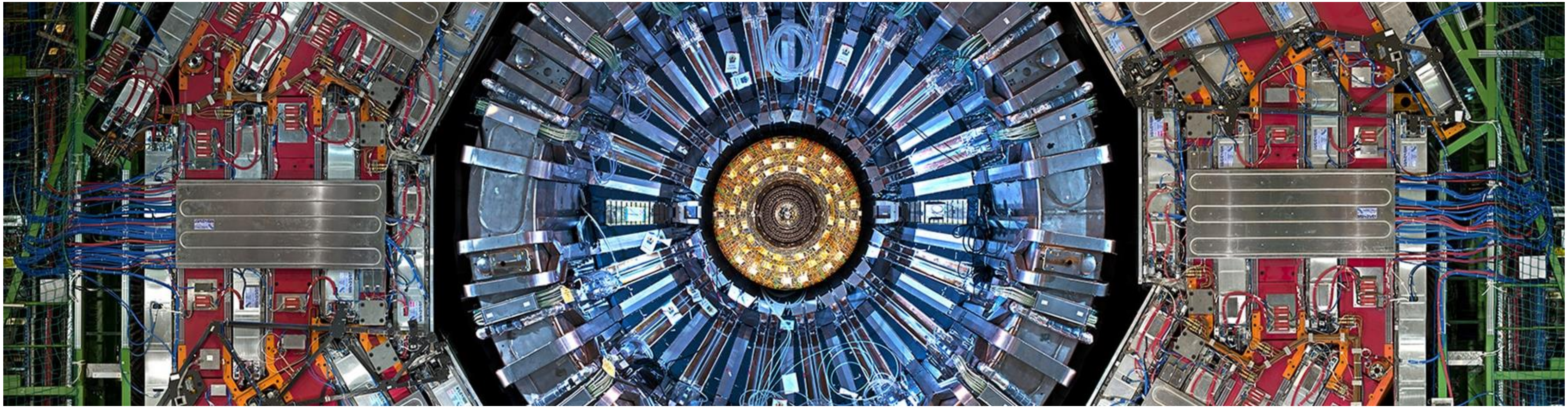
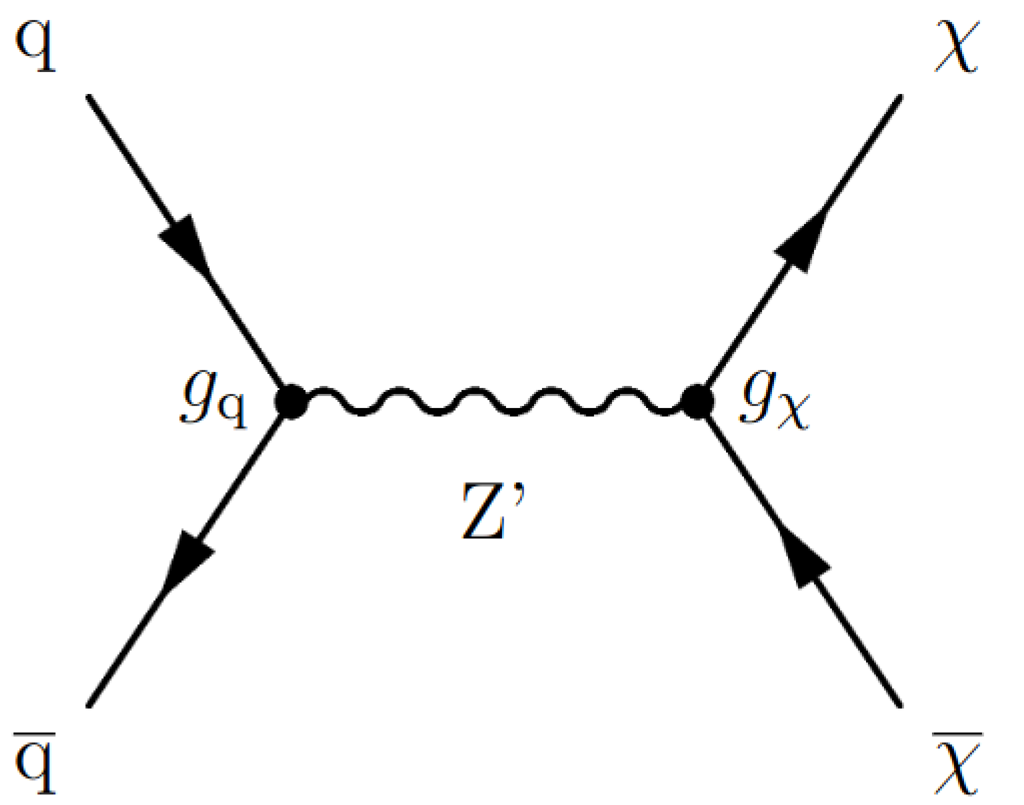
Search for Semivisible Jets at the CMS Experiment using Run 2 Scouting Data

Marcel Gaisdörfer, Markus Klute, Benedikt Maier, Brendan Regnery



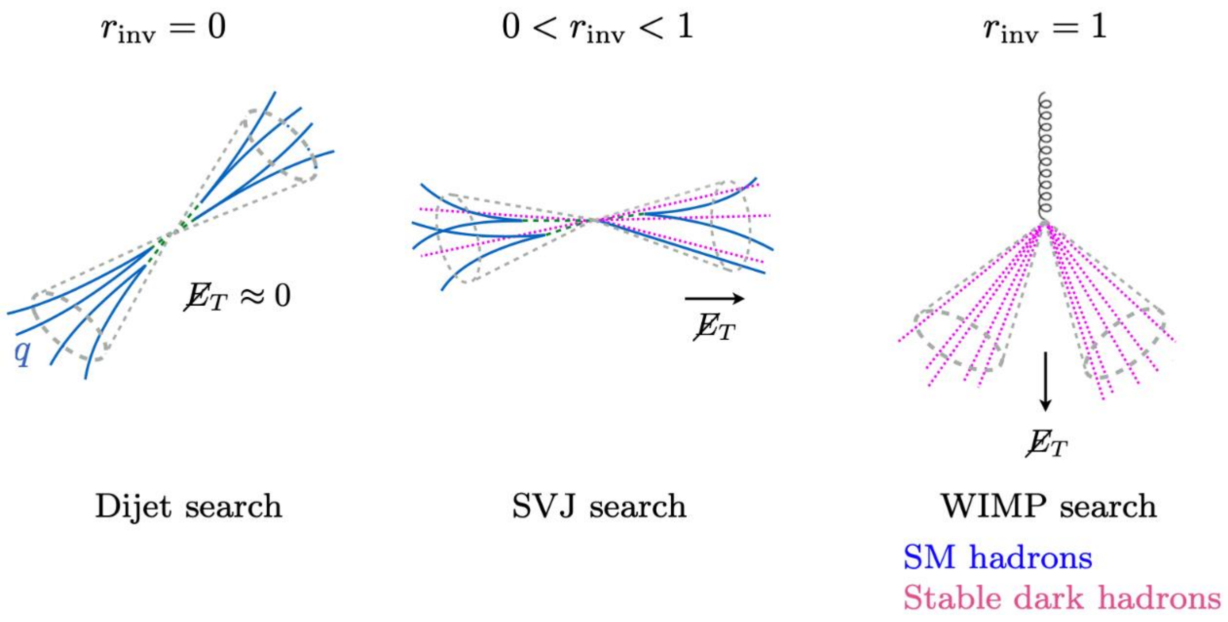
Hidden Valley Model

- Additional broken U(1) symmetry leads to massive Z‘ boson
- Z‘ acts as mediator to a QCD-like dark sector
- Coupling to the dark sector large compared to coupling to SM quarks (no coupling to leptons)



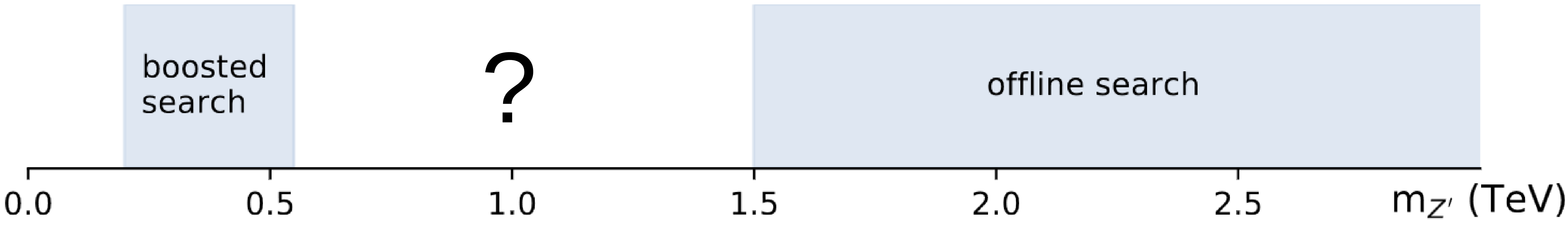
What is a Semivisible Jet?

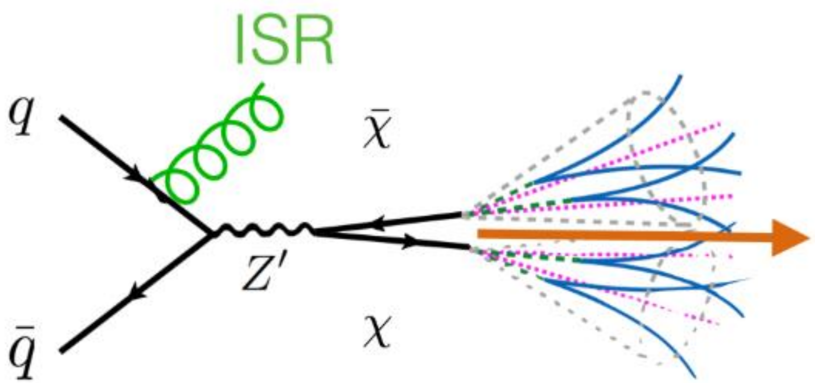
- Dark sector consists of 2 dark quarks that hadronize
- A fraction of dark quarks decay back to SM quarks
  → large jets with invisible particles, called semivisible jets
- Shower dynamics depend on fraction of invisible particles r_{inv}

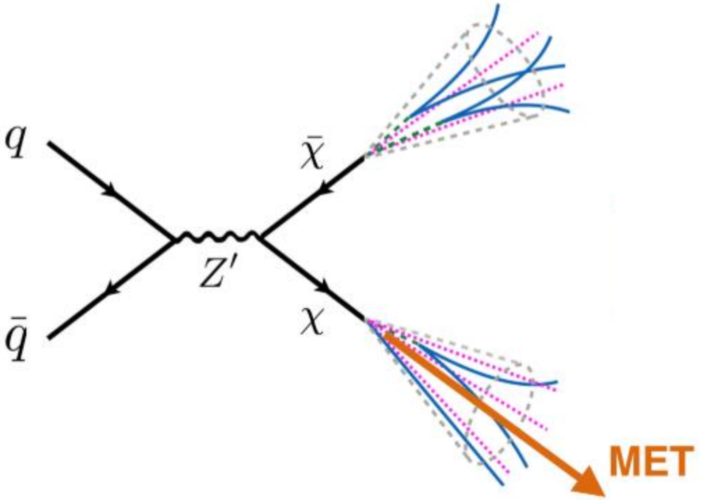


https://indico.cern.ch/event/1319442/contributions/5666522/attachments/2762759/4812368/SVJ_EXO_Workshop.pdf

SVJ Phase Space

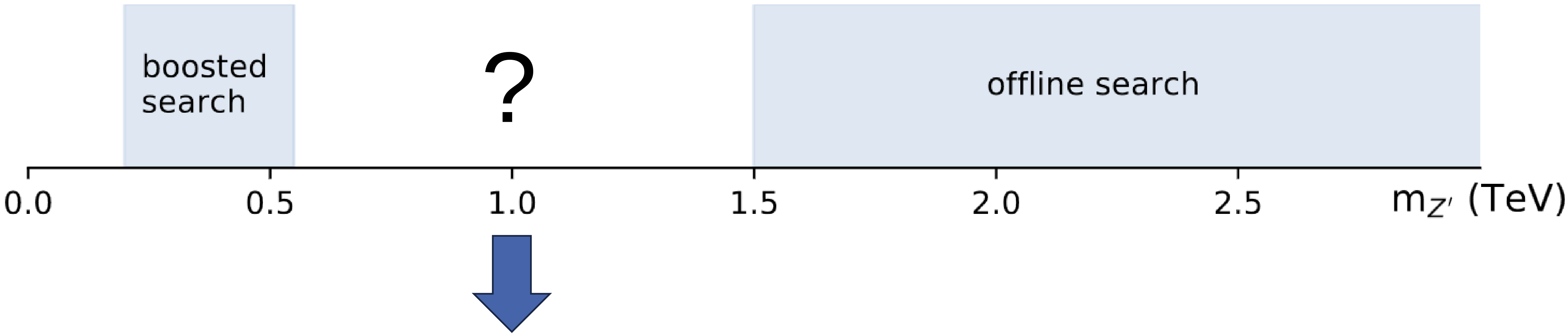






https://indico.cern.ch/event/1319442/contributions/5666522/attachments/2762759/4812368/SVJ_EXO_Workshop.pdf

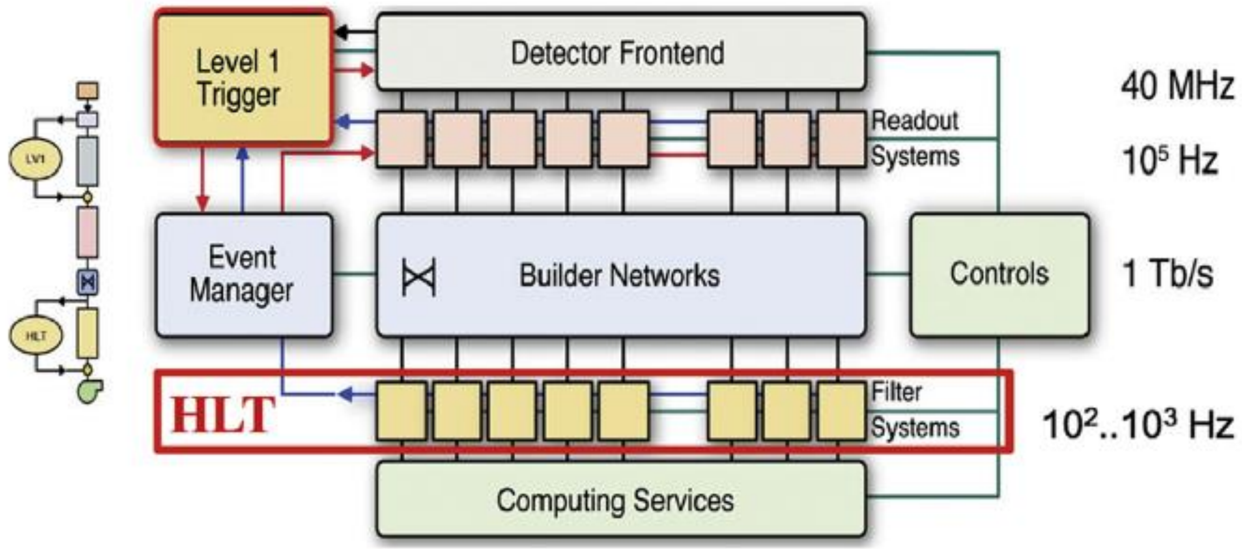
SVJ Phase Space



Accessible through scouting data!

Reminder: The CMS Trigger System

- Impossible to record all events produced at the LHC
- L1 Trigger: hardware-based
- HLT: software-based, uses a fast online reconstruction for decision making
- Selected events are subjected to prompt offline reconstruction

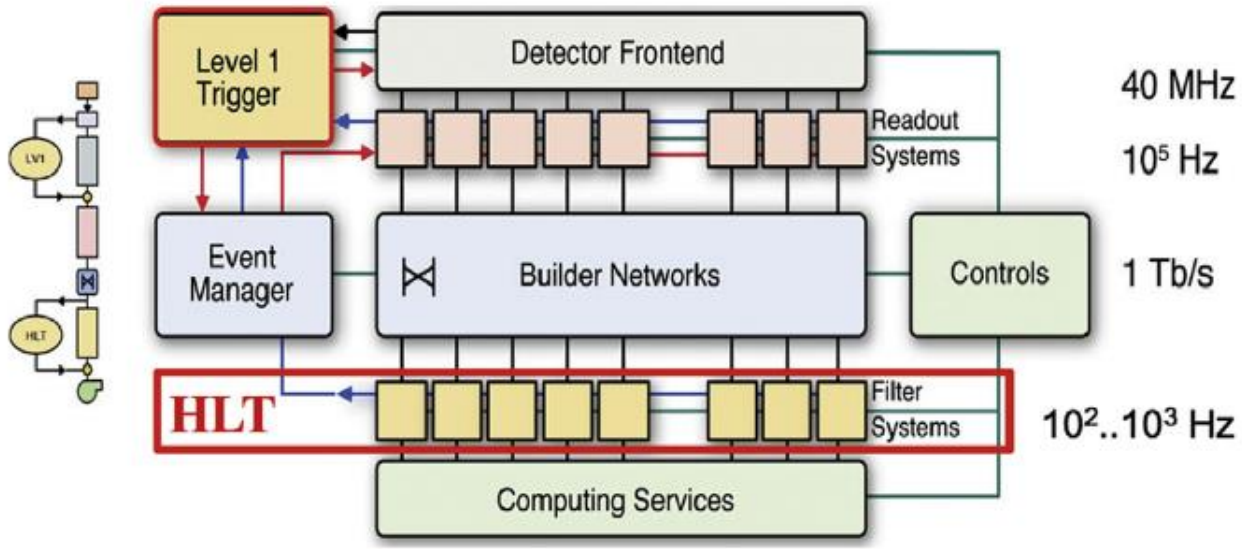


https://www.sciencedirect.com/science/article/pii/S0168900212009941?ref=cra_js_challenge&fr=RR-1

Reminder: The CMS Trigger System

Problems:

- BSM physics could hide at low energies, below trigger thresholds
- Trigger thresholds rise with increasing luminosity (HL-LHC!)

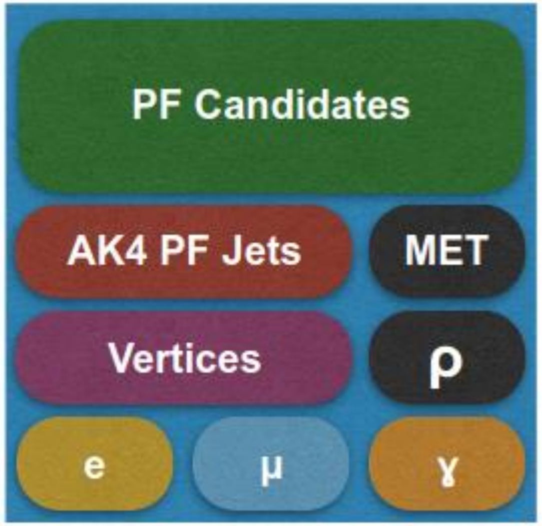


https://www.sciencedirect.com/science/article/pii/S0168900212009941?ref=cra_js_challenge&fr=RR-1

Data Scouting and Data Parking at CMS

- Use online reconstruction capabilities of the HLT to only save reconstructed physics objects

- Pros:
  - Low file size (10-15kB/event vs 1MB/event)
    → can record more events at lower trigger thresholds
  - Low disk space needed
  - Almost no additional strain on DAQ
  - HLT reconstruction not much worse than offline reconstruction



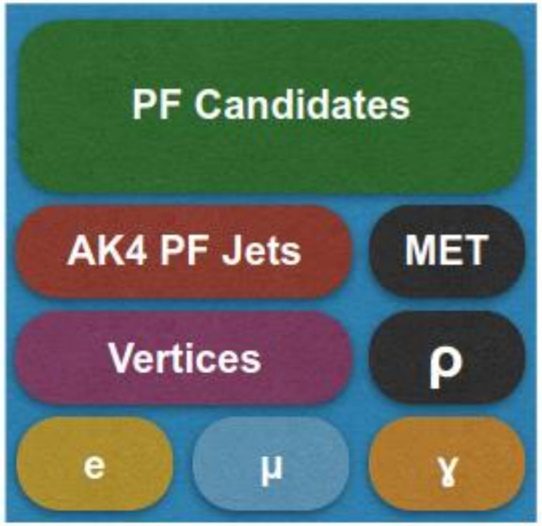
https://arxiv.org/abs/1808.00902

Data Scouting and Data Parking at CMS

- Use online reconstruction capabilities of the HLT to only save reconstructed physics objects

- Cons:
  - Loss of some accuracy (HLT uses slightly simplified PF)
  - Loss of flexibility: custom reconstruction instead of PF impossible → data parking

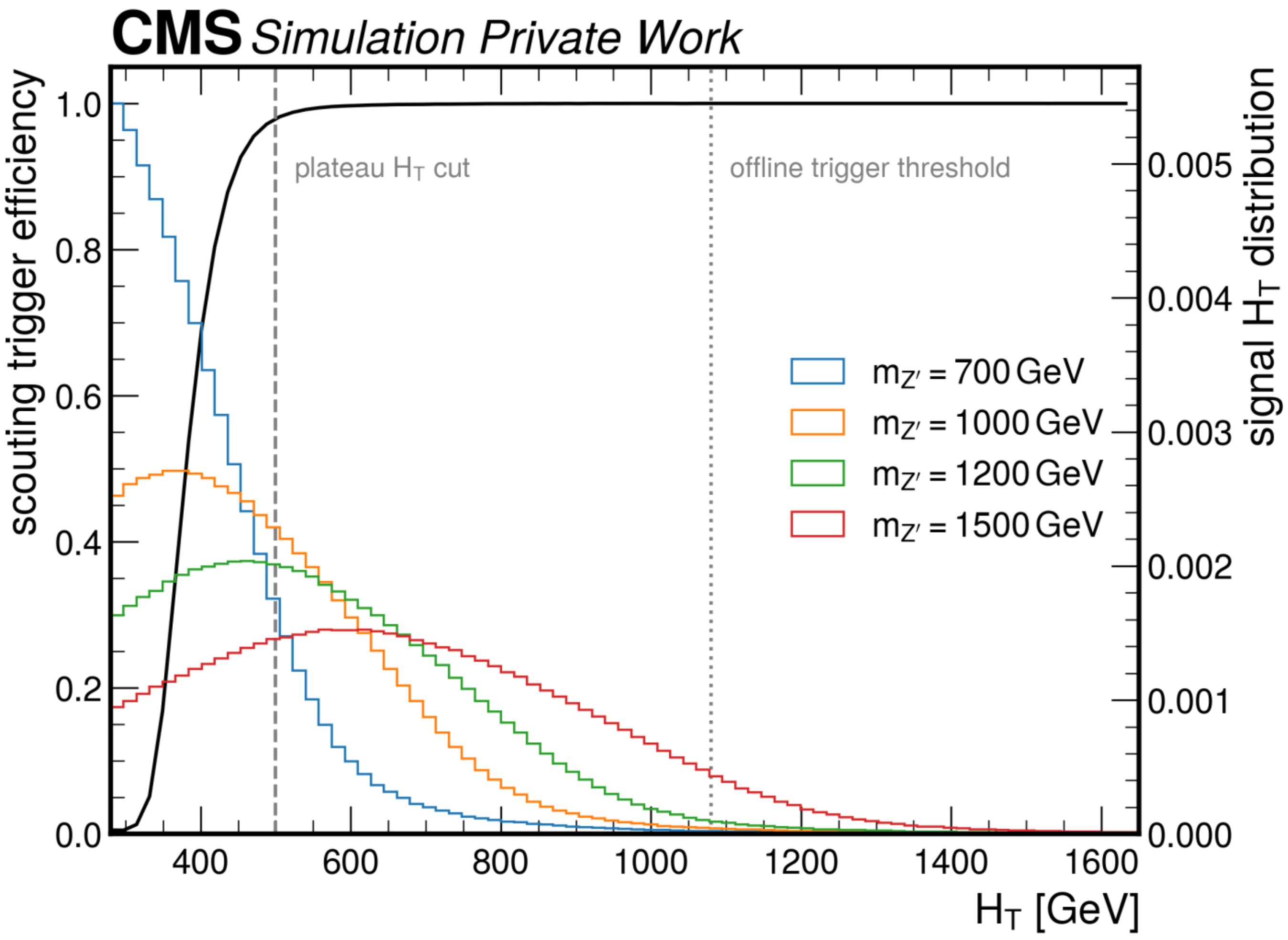
Data parking: save full event information on tape without running a reconstruction (can be analyzed later if needed)



https://arxiv.org/abs/1808.00902

Data Scouting for SVJ

Karlsruhe Institute of Technology

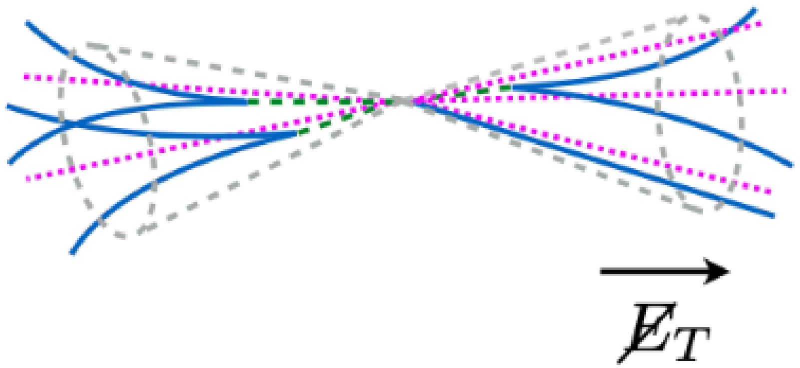


- Offline trigger threshold: 1080 GeV
- Scouting trigger threshold: 410 GeV (500 GeV to be at the plateau)

→ Scouting allows for BSM searches at lower energies (meaning here: lower Z' mediator masses)

Signal-Background Discrimination

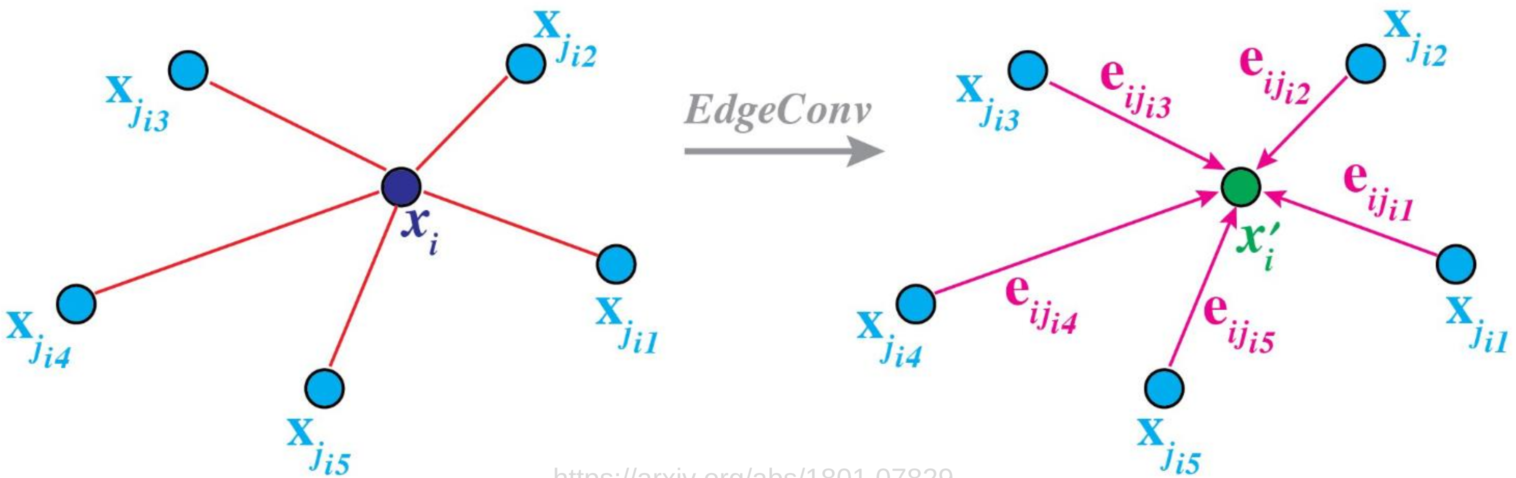
- Event signature: 2 large Jets, moderate amount of missing transverse energy aligned with one of the jets
- Main SM backgrounds: QCD multijet events, ttbar
- Two approaches: model-independent and model-dependent
  - Model-independent: cut-based event selection
  - Model-dependent: cuts + machine learning-based tagger



https://indico.cern.ch/event/1319442/contributions/5666522/attachments/2762759/4812368/SVJ_EXO_Workshop.pdf

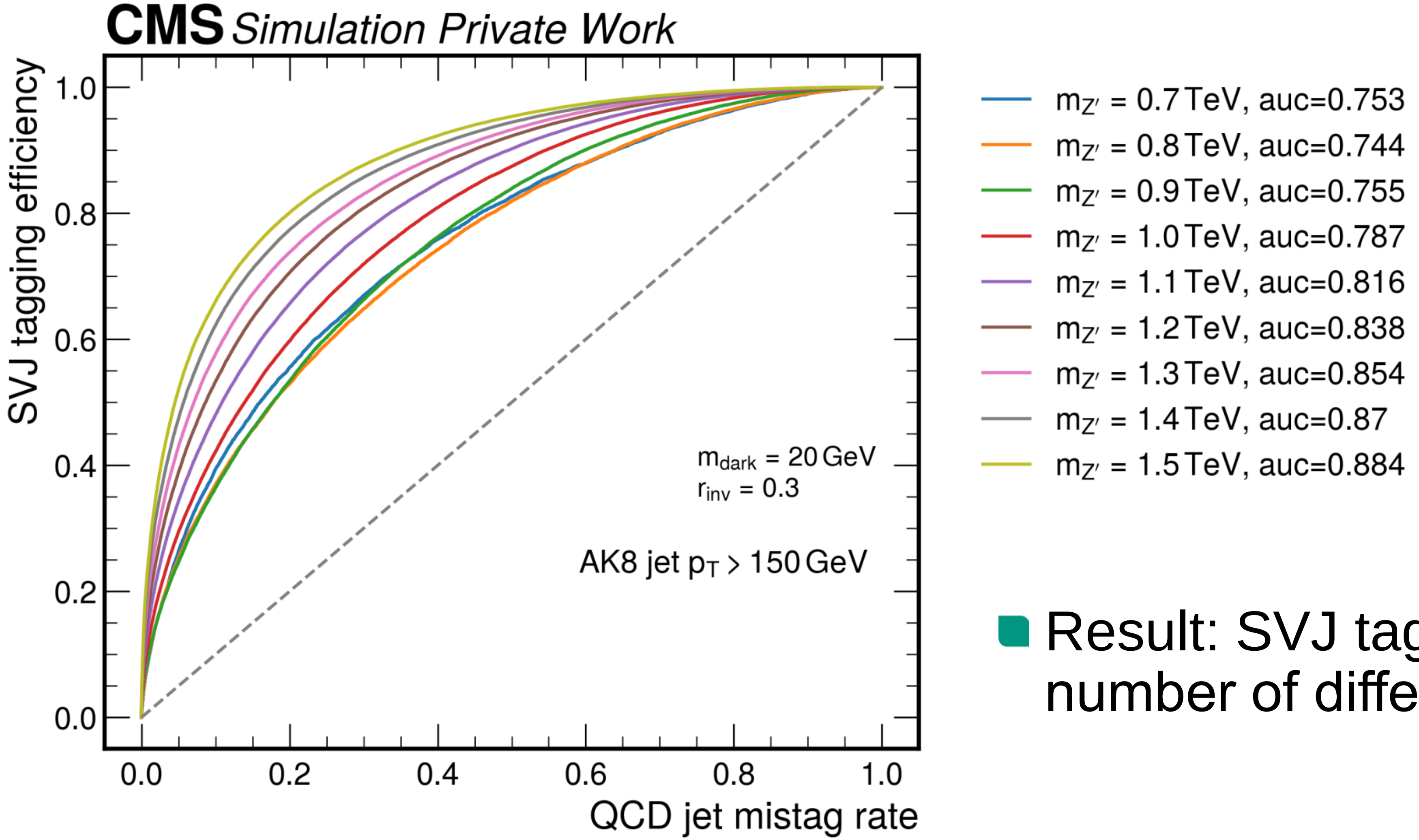
GNN-based SVJ tagger

- Input: features of the constituents of the 2 leading jets (p_T, η, ϕ, mass, charge, pdgID), represented as a particle cloud
- Dynamic Edge Convolution used to construct local graphs from k-nearest neighbors
- Stacking of Edge Convolution builds a deep network, dynamically updating the graphs and learning jet substructure



https://arxiv.org/abs/1801.07829

GNN-based SVJ tagger

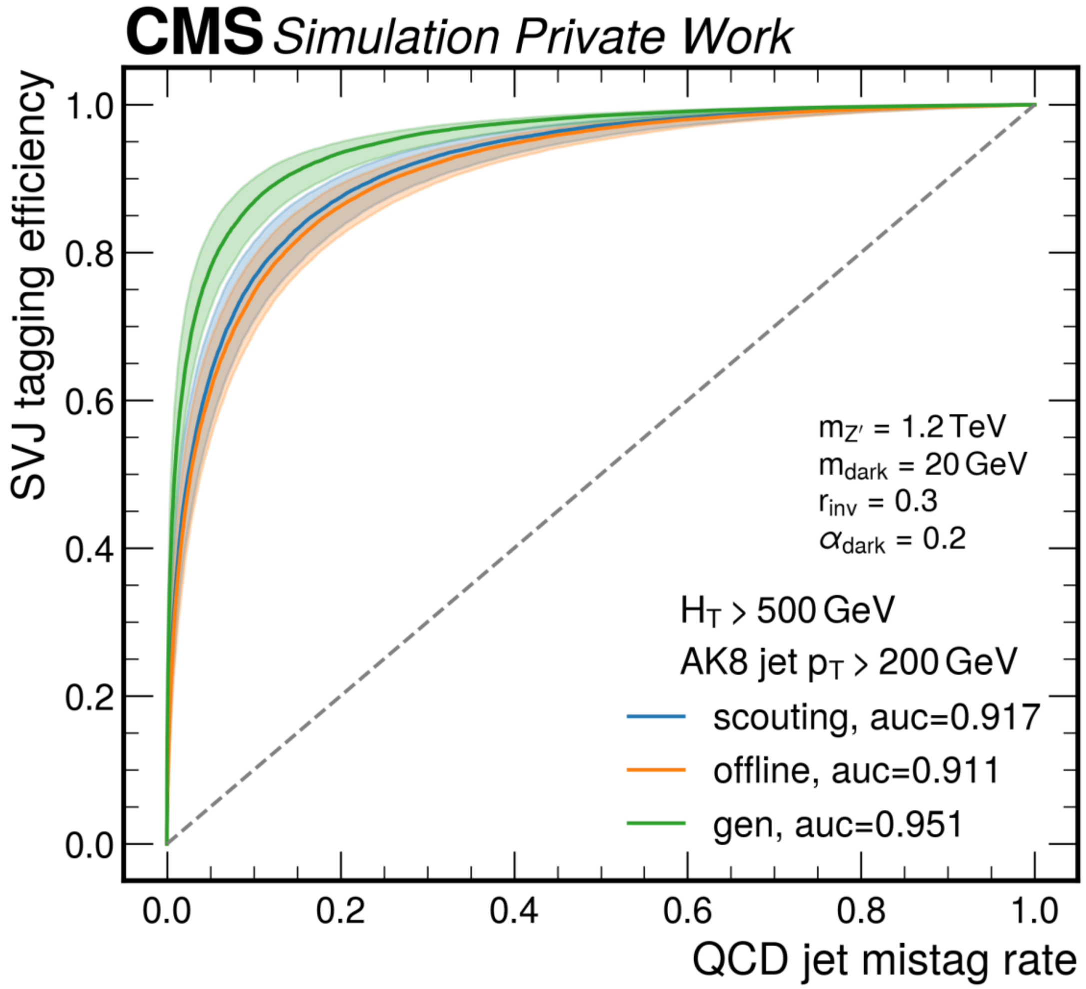


- Result: SVJ tagger trained on a number of different Z‘ masses

Scouting Study

- Performance of the tagger evaluated on
  - Scouting data
  - Offline reconstruction
  - Genparticles

- Performance found to be similar between scouting and offline reconstruction



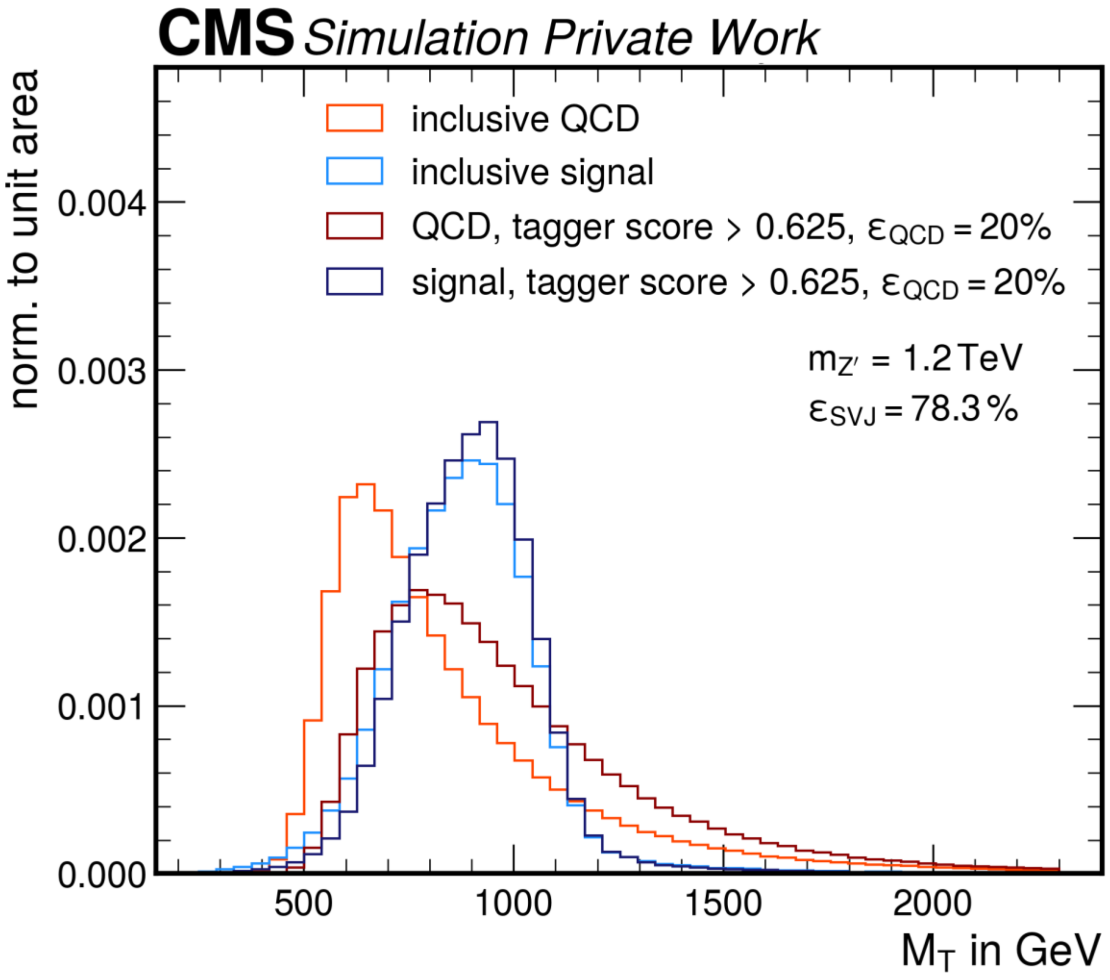
Sculpting

- Search strategy: typical bump hunt in M_T distribution

$$M_T^2 = \left[E_{T,JJ} + E_T^{miss}\right]^2 - \left[\vec{p}_{T,JJ} + \vec{p}_T^{miss}\right]^2$$
$$= M_{JJ}^2 + 2p_T^{miss}\left(\sqrt{M_{JJ}^2 + p_{T,JJ}^2} - p_{T,JJ}\cos(\phi_{JJ,miss})\right)$$

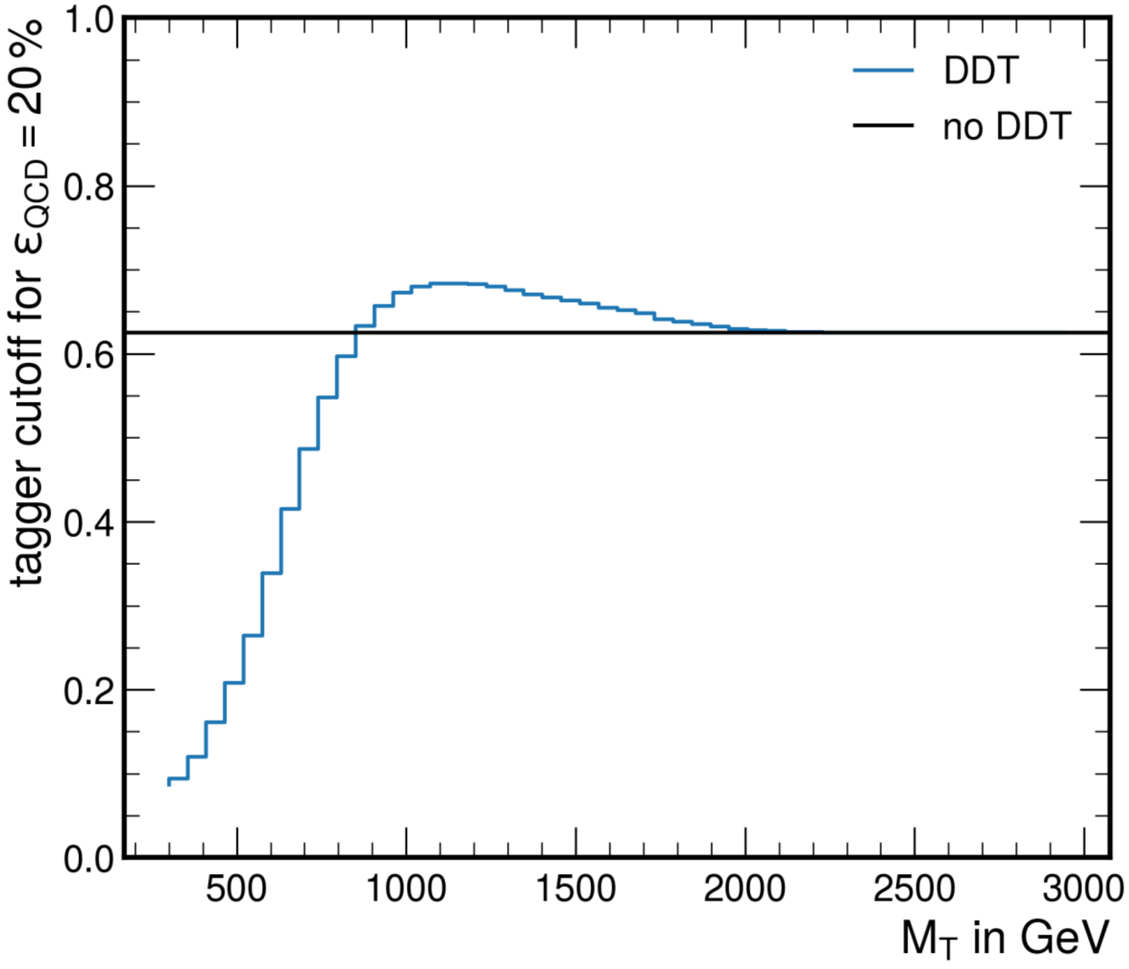
- **Problem:** applying the tagger “sculpts” the shape of the background distribution due to correlation with M_T

→Problems with background fit

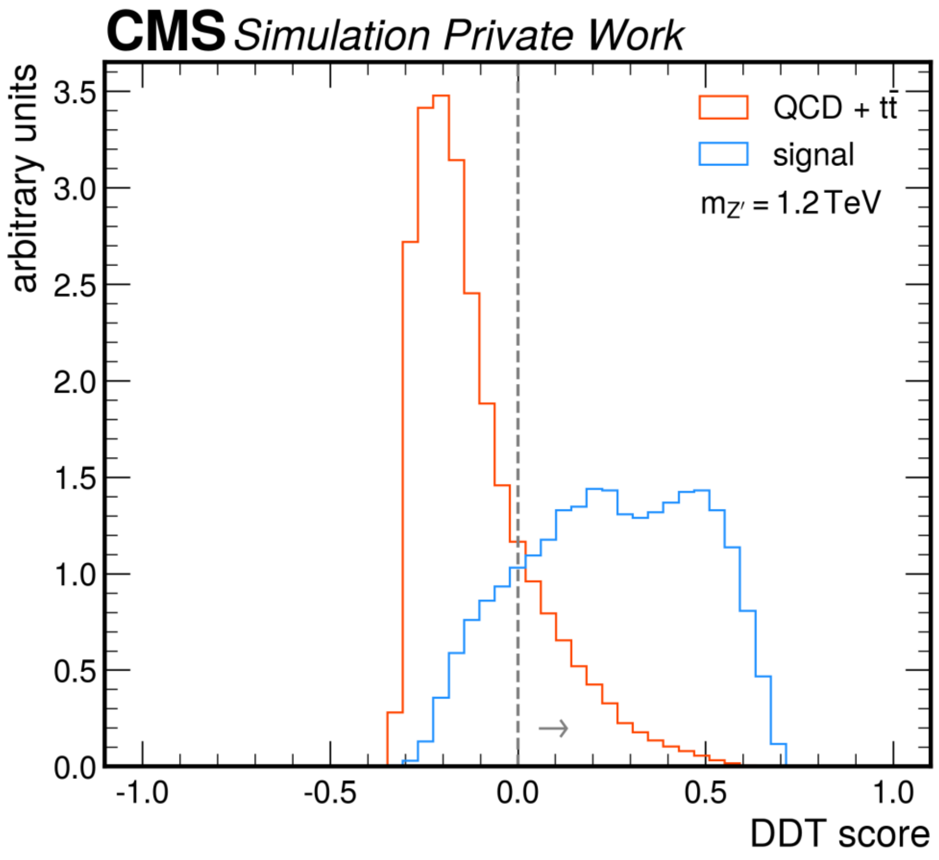


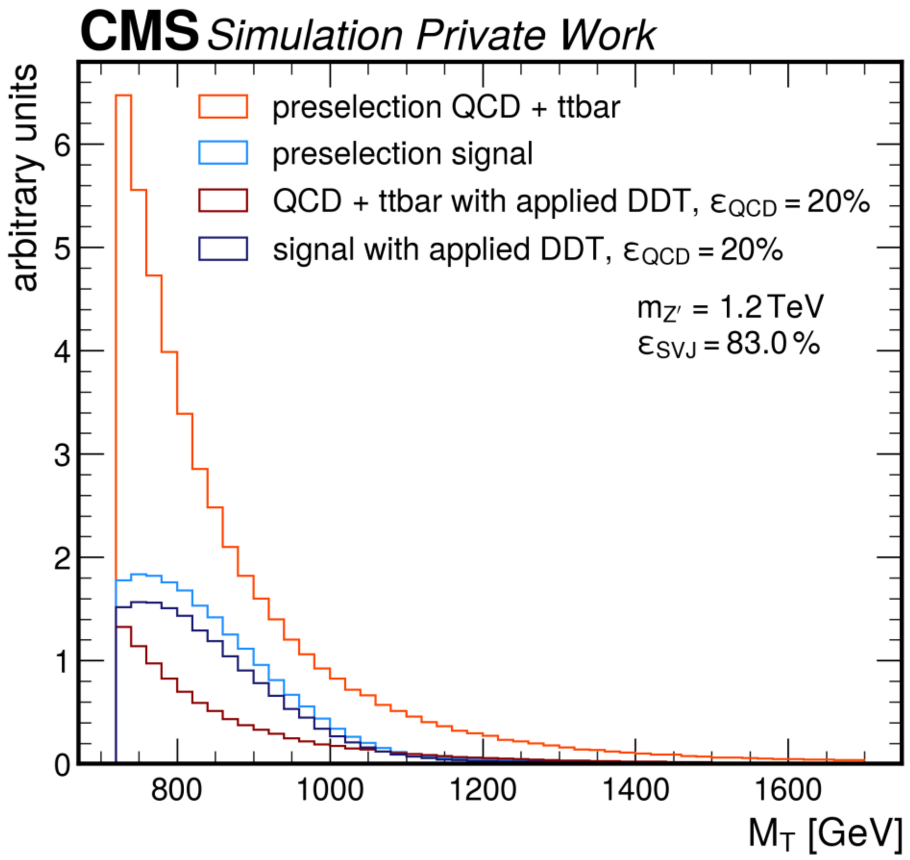
Decorrelation of the Tagger

- Separate events into M_T bins
- Calculate tagger cutoff per bin to cut off 80% of QCD events
- DDT = min(Score_{J1}, Score_{J2}) – cutoff
- This preserves the shape of the distribution while removing most background



Decorrelated Distributions

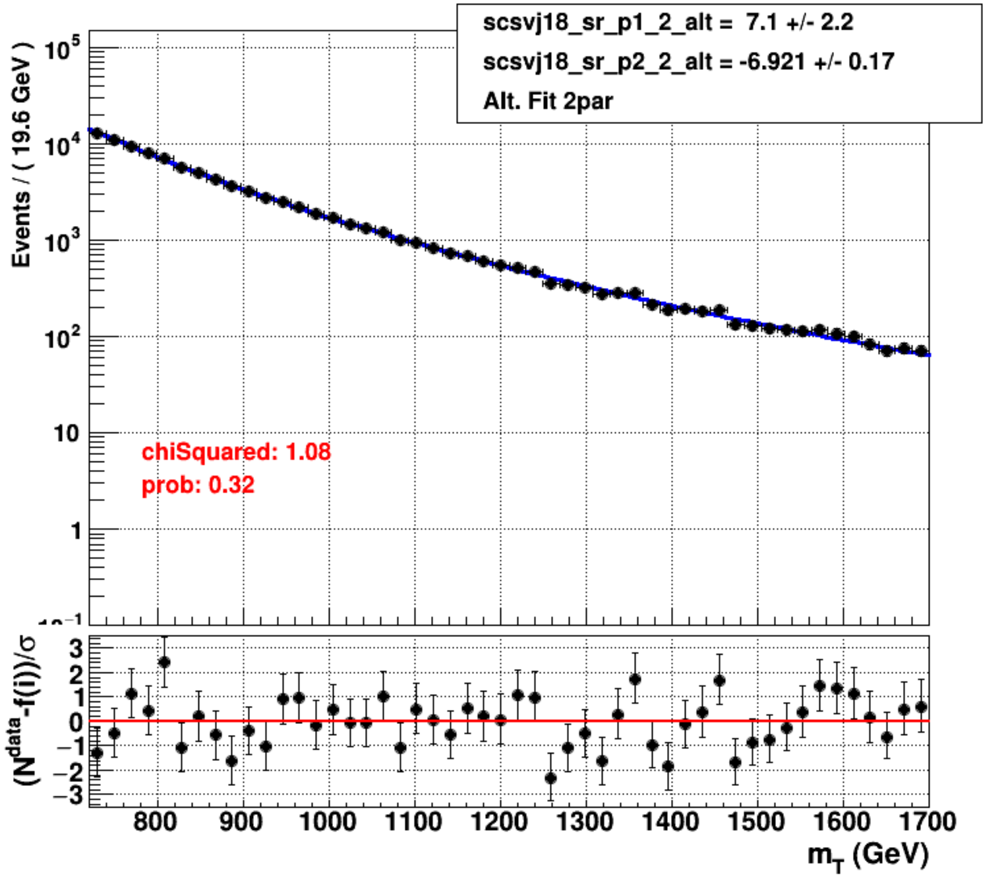




Background Estimation

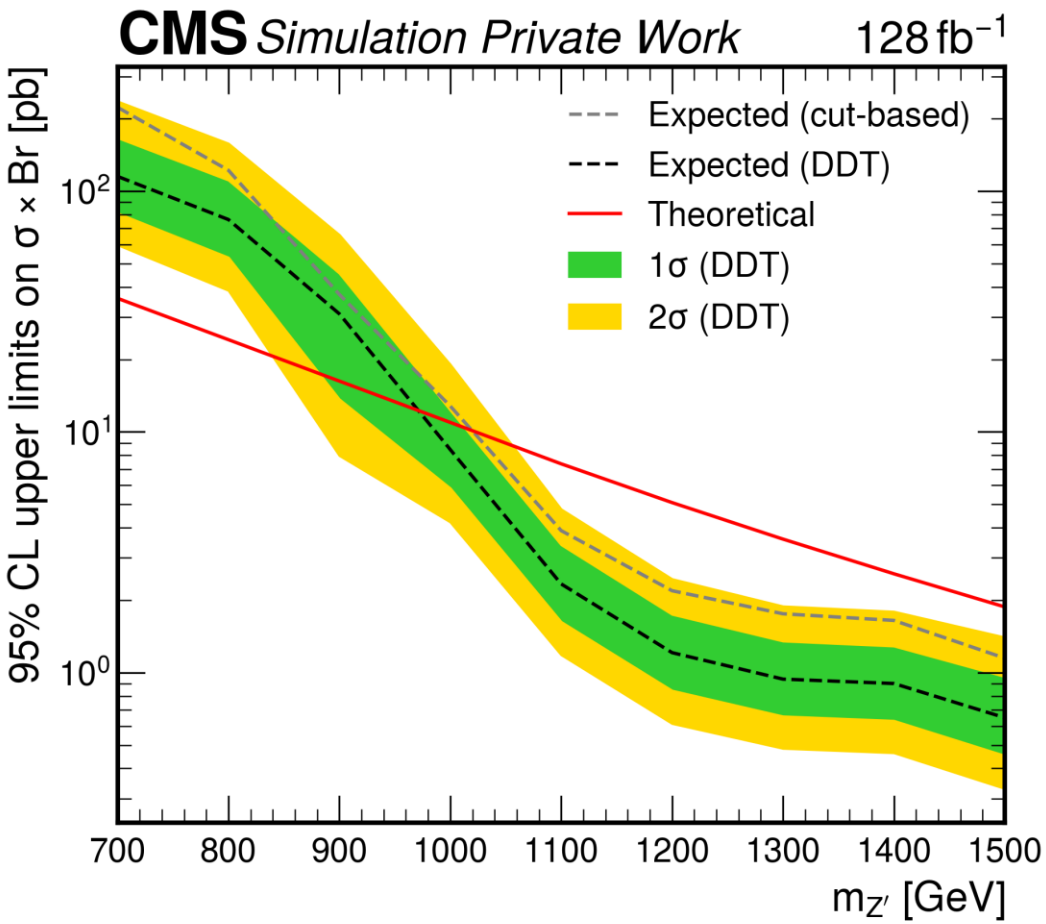
- Data-driven approach:
  fit analytic functions to the background spectrum

- Use Fisher test to determine optimal number of parameters



Expected Limits

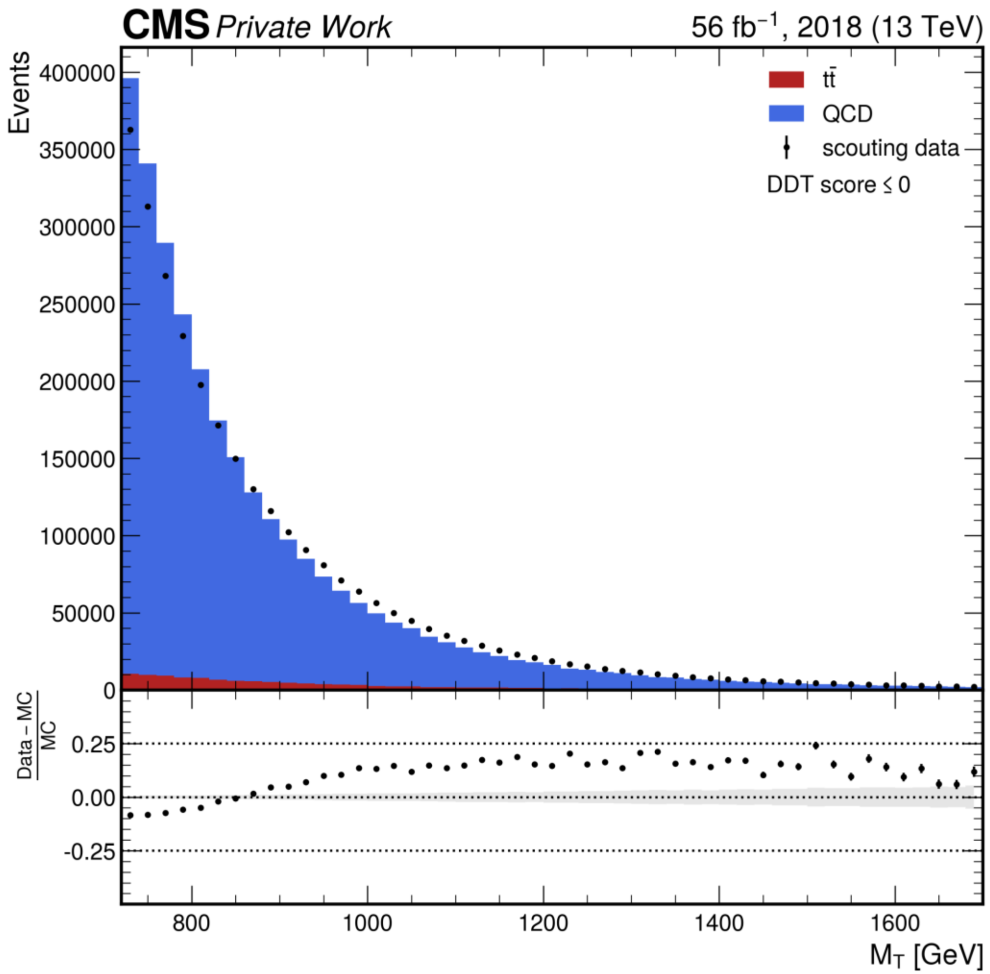
- Calculated expected limits from a small mock data set (CL_s method), scaled to the entire Run 2 scouting data set
- Expected possible exclusion above 1.0 TeV (1.1 TeV without tagger)
- Limits consistent with 2022 offline analysis at $m_{Z'}$ = 1.5 TeV



First Data Plots

- Tagger used to select background-like events and blind signal
- MC normalized to data for shape comparison

→ further corrections / additional backgrounds needed



Conclusion and next Steps

- Analysis framework has been set up
- Expected exclusion 1.0 (1.1) < $m_{Z'}$ < 1.5 TeV, limits consistent with offline analysis at shared mass point

- Next steps:
  - More expansive uncertainty estimation
  - Could add more SM background processes
  - Add further MC corrections
  - Could further optimize the GNN
  - …
  - Unblind the data to obtain final results

Backup: Selection and Cutflow

selection cuts	background efficiency	signal efficiency
2 AK8 jets with $p_{\mathrm{T}} \geq 150\,\mathrm{GeV}$ and $\eta \leq 2.4$	1.0000	1.0000
$H_{\mathrm{T}} > 500\,\mathrm{GeV}$	1.0000	1.0000
scouting trigger	0.9985	0.9917
$M_{\mathrm{T}} > 720\,\mathrm{GeV}$	0.7074	0.6509
$R_{\mathrm{T}} > 0.15$	0.0079	0.1778
$\Delta\eta_{JJ} < 1.5$	0.0064	0.1589
$\Delta\phi_{\mathrm{min}} < 0.8$	0.0061	0.1516

Backup: GNN Architecture

- 3-layer MLP
- 2 stacked DynamicEdgeConv operations (k=24)
- 5-layer MLP (2 10% dropout layers)
- GlobalSumPool aggregates outputs of all nodes into final prediction score

$$h_{\text{embed}} = \text{MLP}_{\text{embed}}(X)$$
$$h_{\text{DGC1}} = \textsc{DynamicEdgeConv}(h_{\text{embed}} | k = 24)$$
$$h_{\text{DGC2}} = \textsc{DynamicEdgeConv}(h_{\text{DGC1}} | k = 24)$$
$$h_{\text{enc}} = \text{MLP}_{\text{enc}}(h_{\text{DGC2}})$$
$$y_{\text{pred}} = \textsc{GlobalSumPool}(h_{\text{enc}})$$