Searches with exotic jet substructure techniques

Simon Rothman (MIT)
For the CMS Collaboration

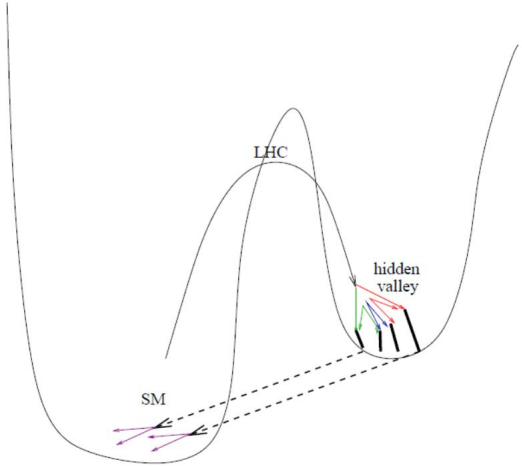
srothman@mit.edu

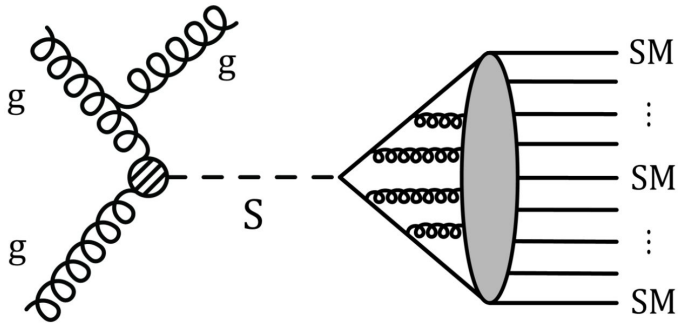
Overview

- LHC has produced amazing quantity of hadronic decays
- **Any naive physics analysis (search, measurement, …) will be completely overwhelmed by SM QCD background**
- Novel techniques in triggering, reconstruction, and selection are necessary to keep pushing our understanding and setting strong limits on new physics
- Today focus on on four recent CMS analyses:
  - CMS-EXO-23-002: Search for **Soft Unclustered Energy Patterns** (SUEPs) - a completely new signature in the far tails of the SM distribution
  - CMS-EXO-23-013: Search for **hadronically-decaying LLPs**, exploiting new Run 3 LLP triggers and deep learning for jet tagging
  - CMS-EXO-22-022: Search for **boosted diphoton resonances**, with completely custom reconstruction strategy based on CNNs
  - CMS EXO-24-007: Study of **boosted hadronic resonances**, with state-of-the art two-prong jet tagging and selections

Soft Unclustered Energy Patterns (CMS-EXO-23-002)

- Motivated by hidden valley dark sector models
  - Dark QCD separated from SM by potential barrier
  - LHC tunnels across barrier to probe DM
- Particularly interesting case is large dark t'Hooft coupling, leading to showers in the hidden valley before decaying back to SM
- Unique signature characterized by
  - Relatively soft tracks
  - High particle multiplicities
  - Isotropic topology
  - Similar signature to QCD instanton
- Very difficult to distinguish from pileup/ordinary high-multiplicity (typically gluon-initiated) jets
  - Requires novel analysis techniques
  - **First ever search for this signature at LHC**

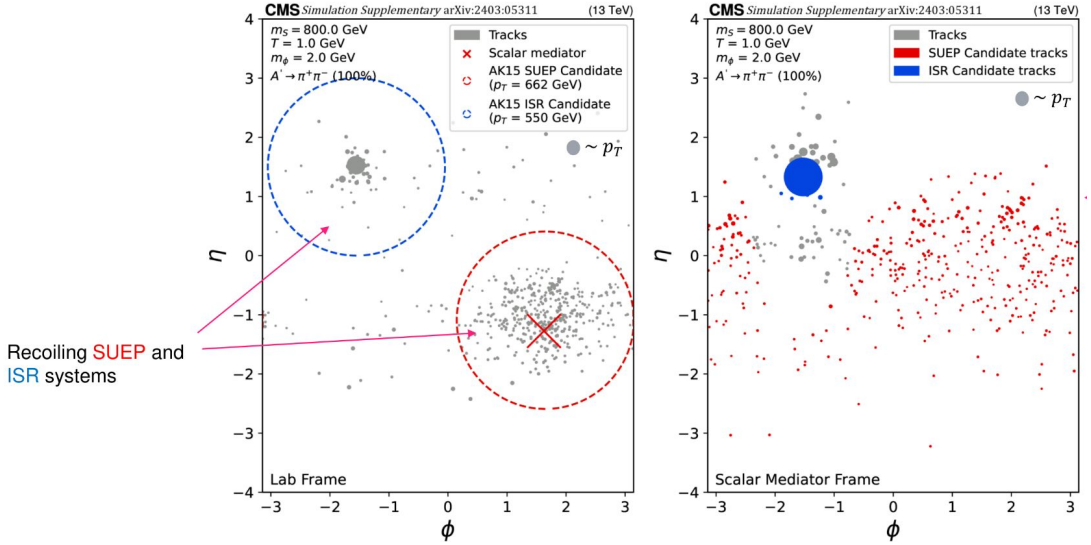




Soft Unclustered Energy Patterns: strategy

Strategy: **search for SUEPs boosted against large ISR**

- Trigger on hadronic activity
- Simplest possible selection (relatively model-independent) - boost into jet reference frame and tag based on sphericity, multiplicity



Soft Unclustered Energy Patterns: strategy

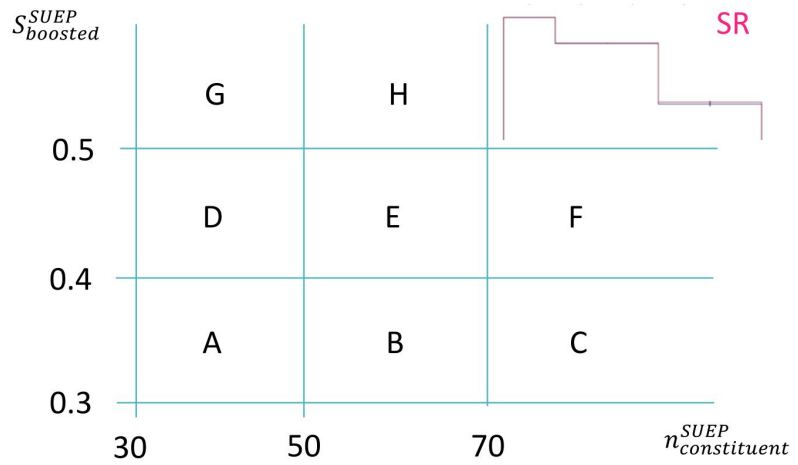
Strategy: **search for SUEPs boosted against large ISR**

- Trigger on hadronic activity
- Boost into jet rest frame and tag based on sphericity, multiplicity

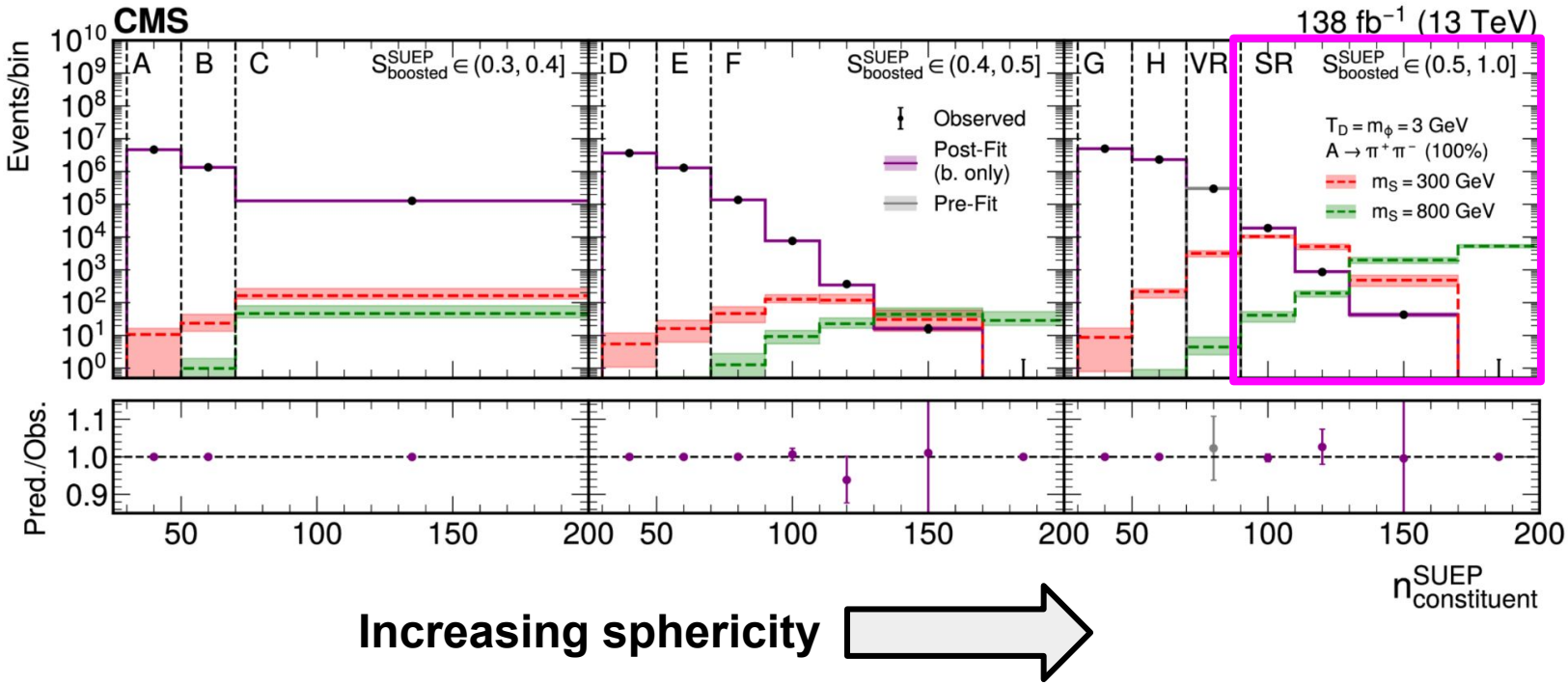
Background modeling is **data driven**

- MC modeling is not very good
- Use extended ABCD to characterize background in-situ

$$SR^{Bin\ i} \approx F^{Bin\ i} \frac{H^2 F\ D^2 B^2}{G\ C\ A\ E^4} + O(\Delta^4)$$



Soft Unclustered Energy Patterns: results

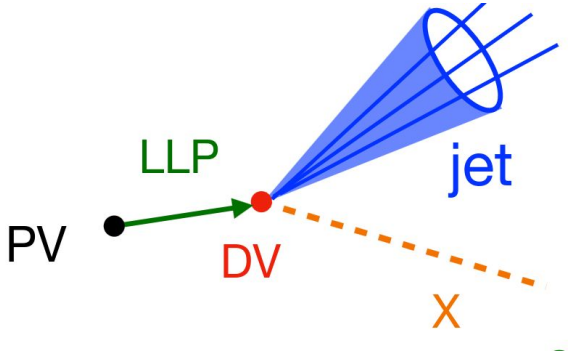


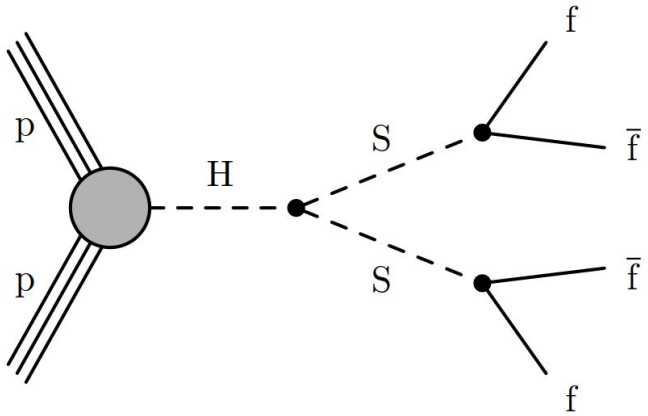
Signal region

Probing far tails of QCD! Excellent agreement with background model

Hadronically-decaying LLPs (CMS-EXO-23-013)

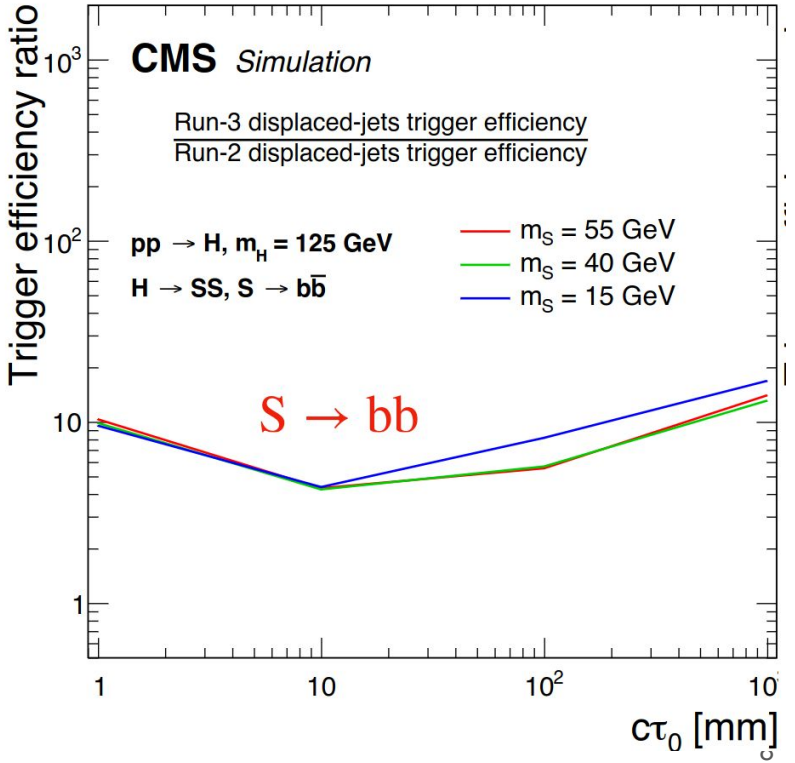
- Long-lived particles ubiquitous in beyond-standard-model theories (SUSY, hidden sectors, HNLs, …)
- Hadronically-decaying LLPs have unique signature of jets originating far from the interaction point
- Benchmark model is H -> SS -> 4 displaced jets
- **Challenge: standard reconstruction and trigger are optimized for prompt decays**





Hadronically-decaying LLPs: trigger and reconstruction

- **Novel techniques in triggering, reconstruction, and event selection**
- **New displaced jet triggers in Run 3**
  - Explicitly trigger on jets with ≤ 1 prompt track
  - **Improved signal efficiency by factor > 10** w.r.t. 2017/18 displaced jet triggers
- **Dedicated reconstruction algorithms for displaced vertices inside jets**



Hadronically-decaying LLPs: tagging

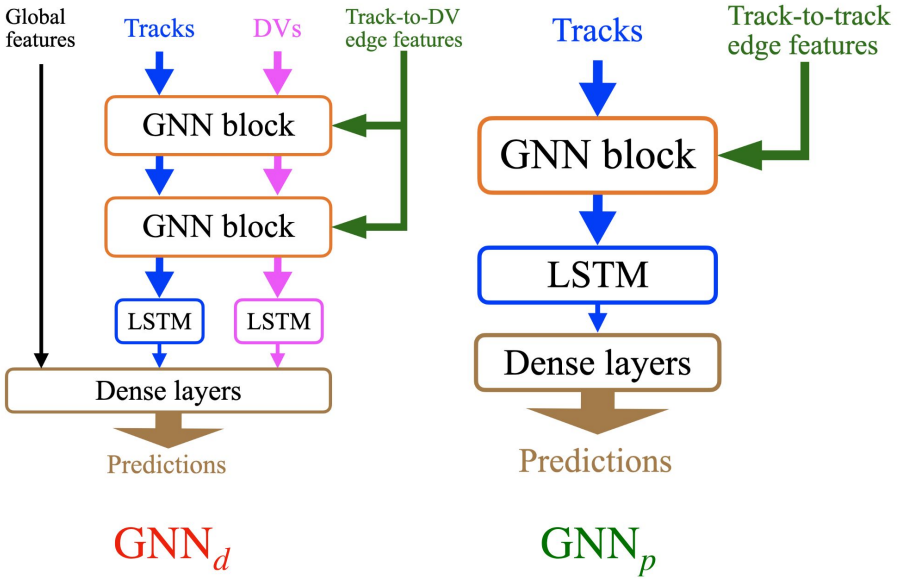
- **Developed GNN taggers to separate LLPs from QCD background**
  - Train two GNNs to obtain decorrelated prompt and displaced scores
- Decorrelated GNN taggers allow **ABCD method for background estimation**

Input to GNN_d:

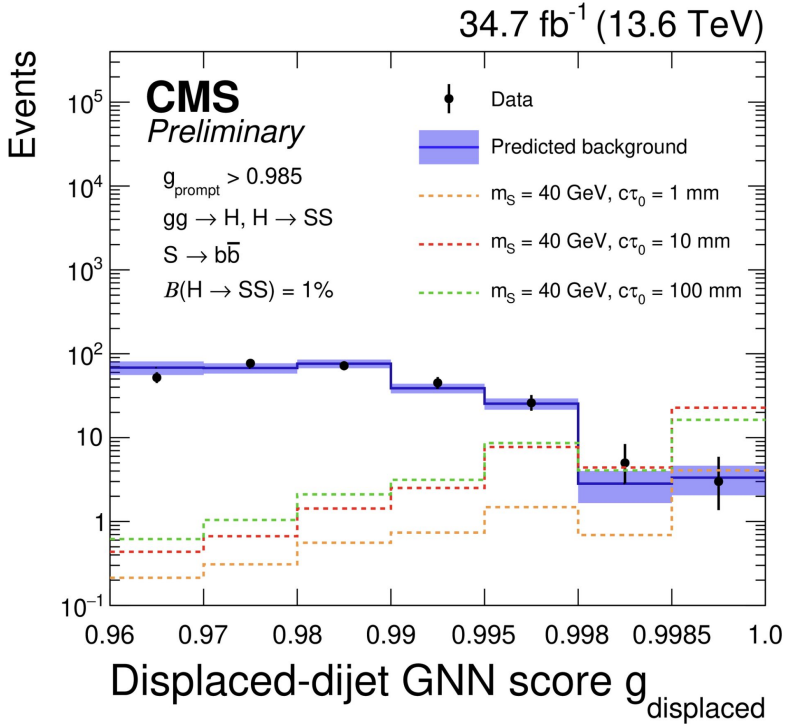
- Displaced tracks
- Displaced vertices
- DT-to-DV association

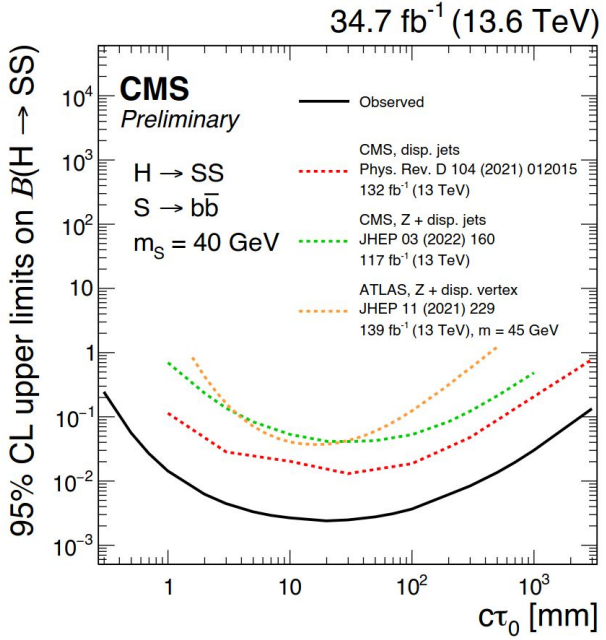
Input to GNN_p:

- Prompt tracks
- Whether tracks are associated to the same PV



Hadronically-decaying LLPs: results

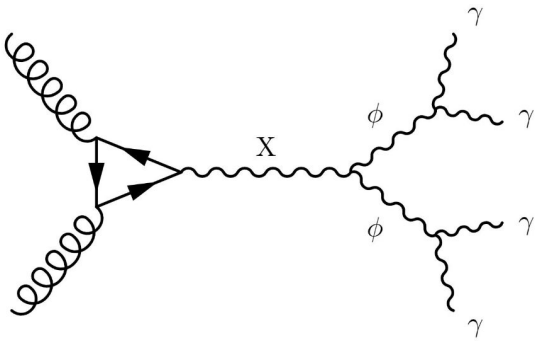


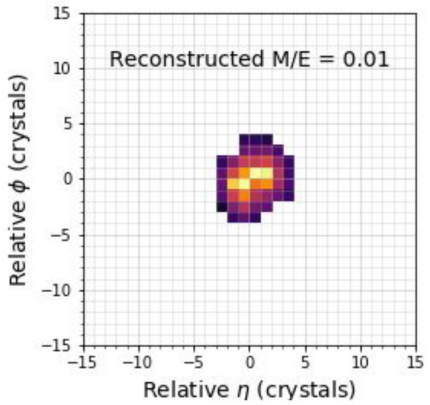


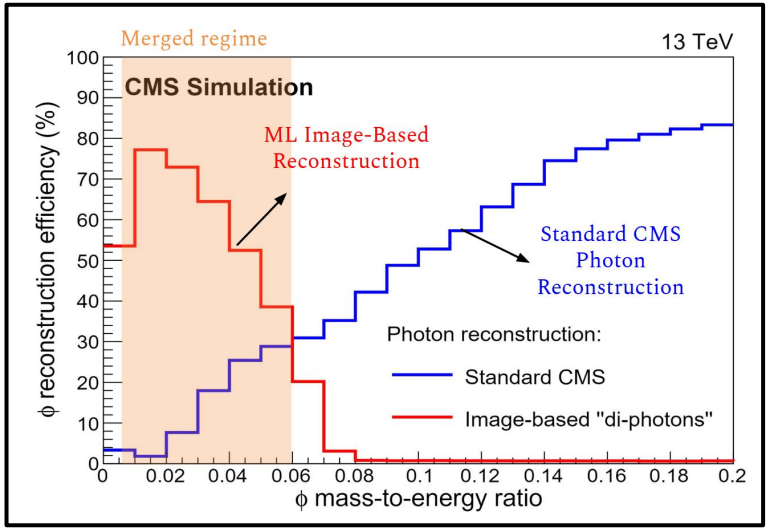
Improvement by order of magnitude w.r.t. Run 2 analysis despite factor 5 less luminosity

Merged diphotons (CMS-EXO-22-022)

- Diphoton decay of light scalar will be highly boosted, yielding merged object in ECAL
  - Analogous to boosted two-prong jets in hadronic case
- Novel reconstruction techniques required to resolved merged diphoton invariant mass
- Benchmark model is decay of new heavy resonance into two light scalars



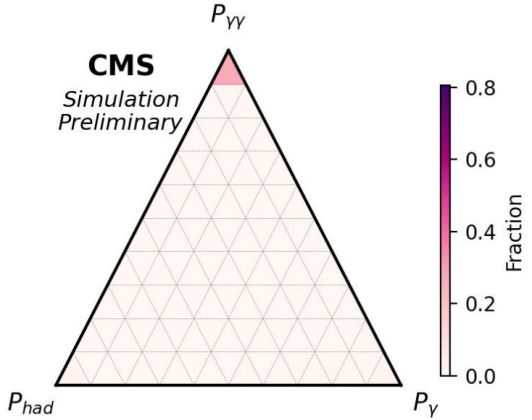




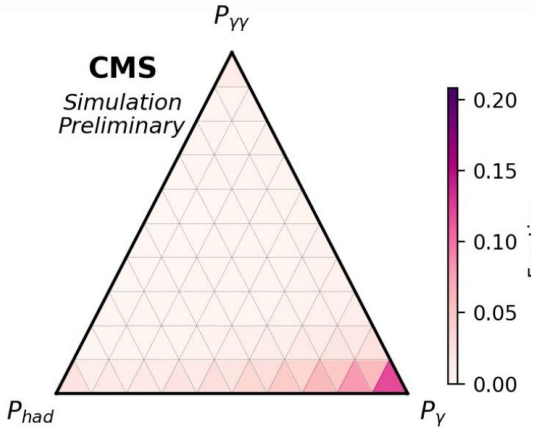
Merged diphotons: strategy

Treat merged diphoton object in crystal ECAL as 2D image and apply state-of-the-art ML techniques

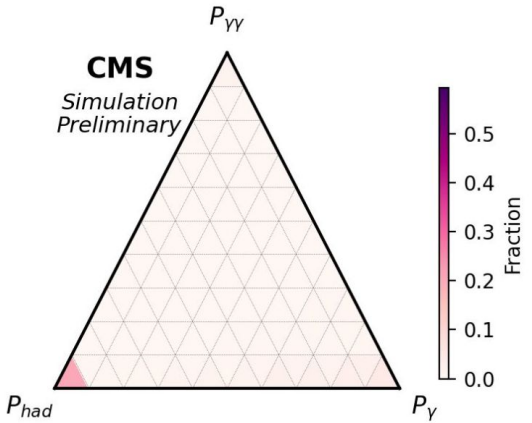
Classification CNN separates
single-photons, diphotons, hadrons



True diphoton



True monophoton

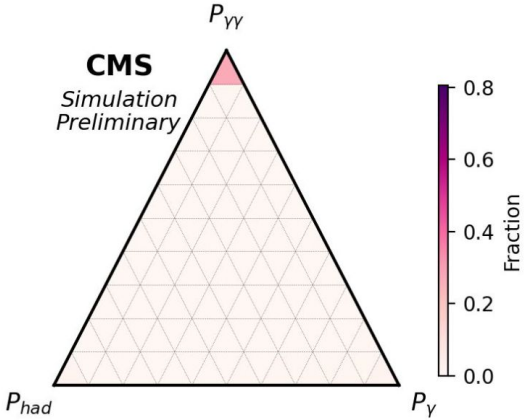


True hadron

Merged diphotons: strategy

Treat merged diphoton object in crystal ECAL as 2D image and apply state-of-the-art ML techniques

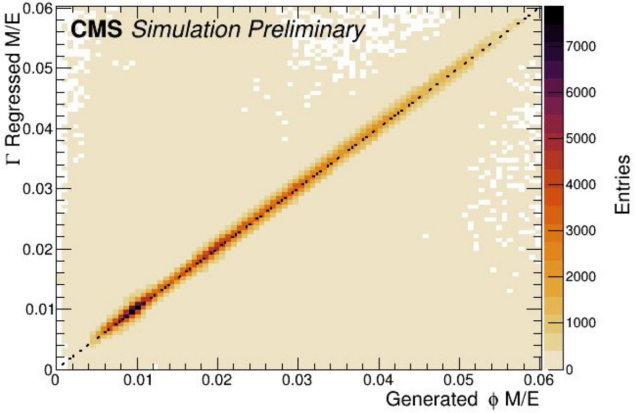
Classification CNN separates single-photons, diphotons, hadrons



True diphoton

Trained on flat m/E to avoid biases

Regression CNN reconstructs merged diphoton invariant mass



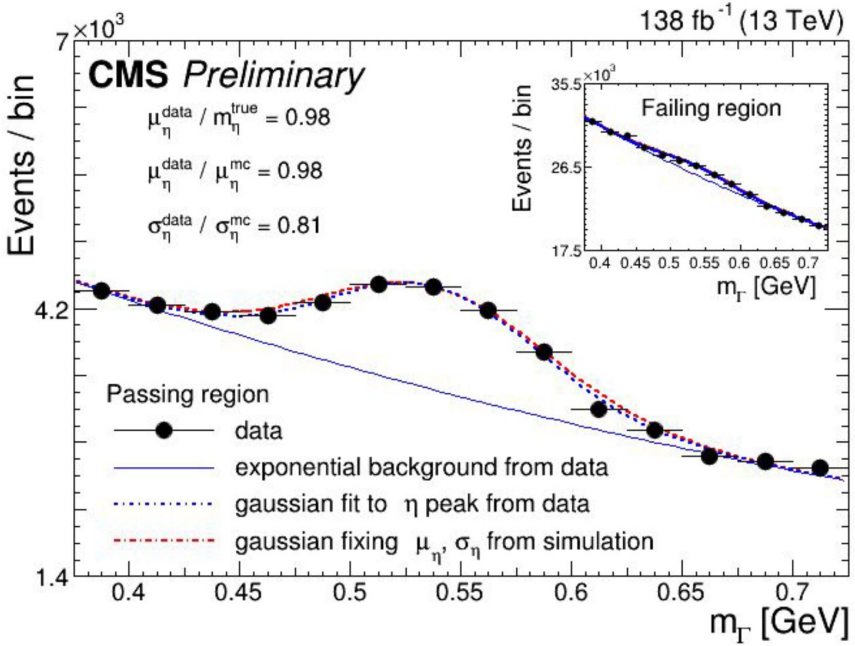
Predicted vs. True M/E in Signal MC

Merged diphotons: results

Validate with η→γγ decays inside jets

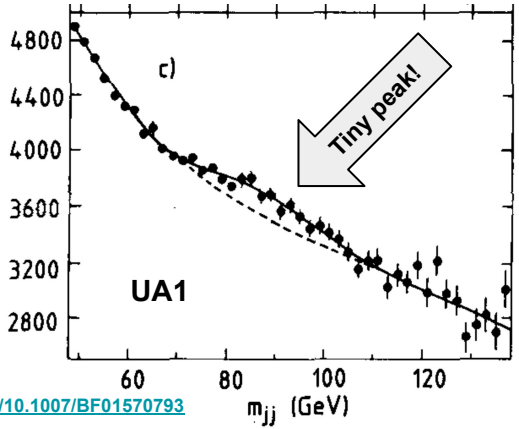
- Simultaneous fit in pass/fail regions to characterize tagger and regression

Perform bump hunt in di-diphoton mass, binned in ratio of diphoton mass to di-diphoton mass; no excess observed



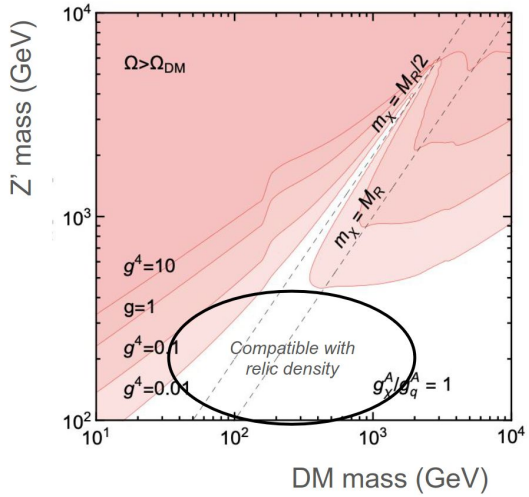
Boosted hadronic resonances (CMS PAS EXO-24-007)

- Low-mass (≲300 GeV) mediators arise naturally in many scenarios
  - Dark matter portals through dark photon
  - Light extra Higgs
  - …
- SM V(qq) very interesting and challenging
  - Small bump at SPS
  - Nothing at tevatron



https://doi.org/10.1007/BF01570793

Chala et al (1503.05916)

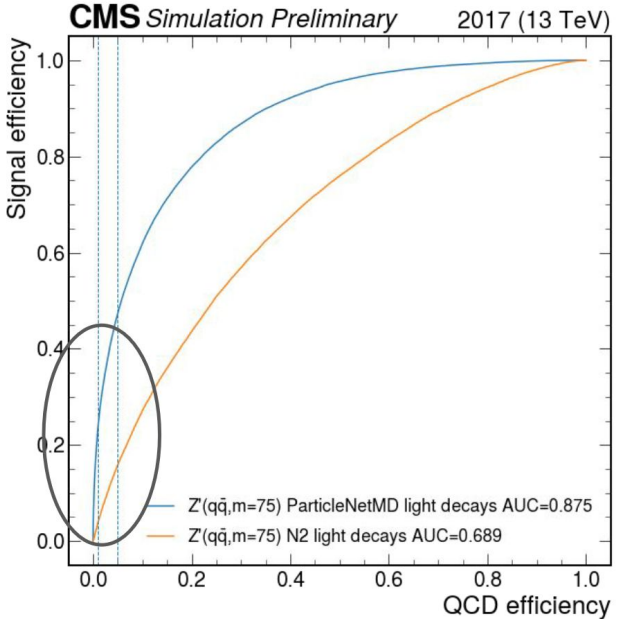


Boosted hadronic resonances: identifying signal events

- **Problem: huge amount of QCD background**
- Solution: new GNN tagger to identify boosted two-prong decays (CMS-DP-2020/002)
  - Significant gains in signal efficiency w.r.t. cutting on 2-subjettiness substructure variable

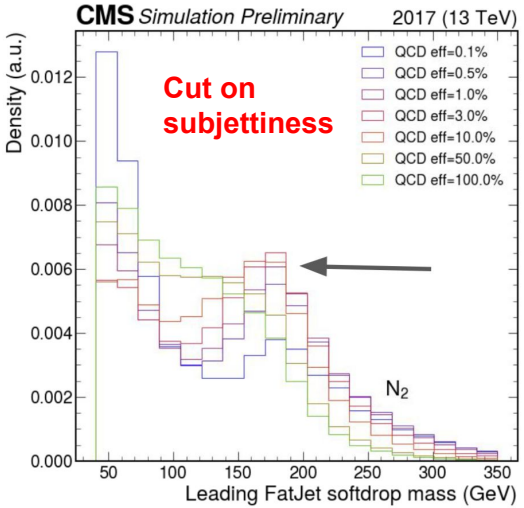
2017: τ21 substructure
-> 2020: N2 substructure
-> 2024: ParticleNet

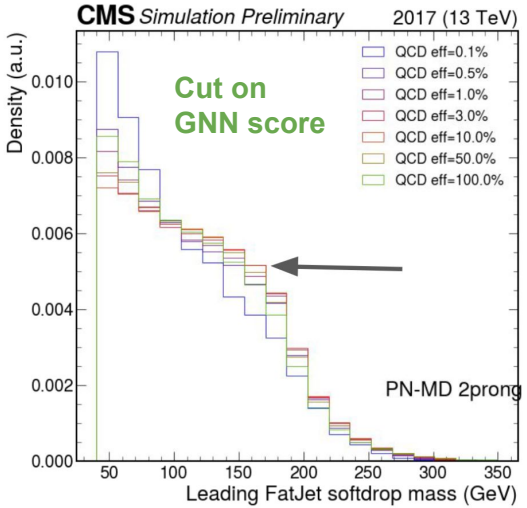
Same signal efficiency as N2 at ⅛ the QCD!



Boosted hadronic resonances: identifying signal events

- **Problem: huge amount of QCD background**
- Solution: new GNN tagger to identify boosted two-prong decays (CMS-DP-2020/002)
  - Significant gains in signal efficiency w.r.t. cutting on 2-subjettiness substructure variable
  - Special samples with flat mass spectrum generated for training, allowing **decorrelation of tagger selection with jet mass**





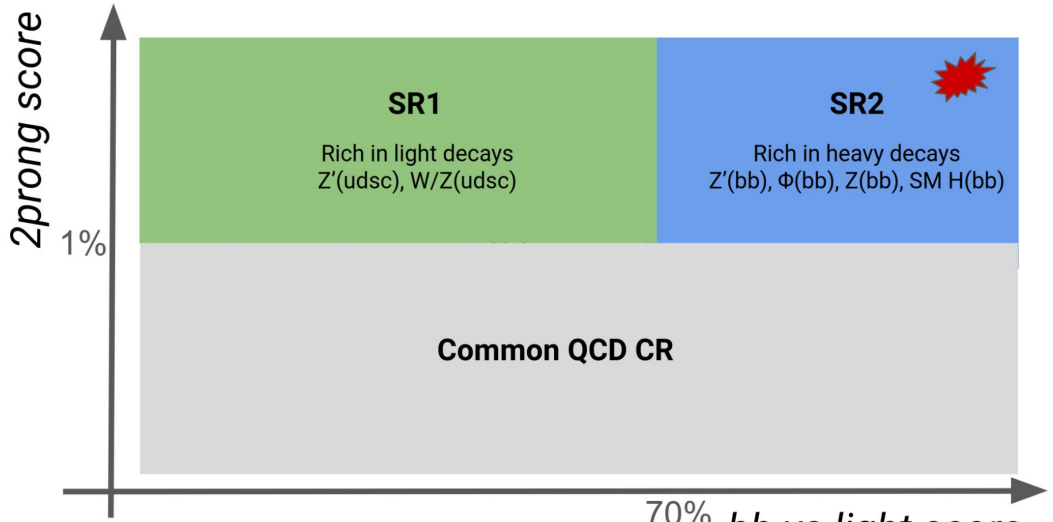
2020: use DDT to manually decorrelate w.r.t mass

2024: mass-decorrelation built into training and fitting procedure

Boosted hadronic resonances: heavy flavor

- GNN tagger discriminates between qq, cc, bb signals and QCD background
- Yukawa scalar decays dominated by bb decays -> **use two signal regions**

Simultaneous test of all possible resonances for first time at LHC



$$\mathrm{PN}^{\mathrm{MD}}_{\mathrm{2prong}} = \mathrm{PN}^{\mathrm{MD}}_{\mathrm{Xbb}} + \mathrm{PN}^{\mathrm{MD}}_{\mathrm{Xcc}} + \mathrm{PN}^{\mathrm{MD}}_{\mathrm{Xqq}}$$

$$\mathrm{PN}^{\mathrm{MD}}_{\mathrm{bvl}} = \frac{\mathrm{PN}^{\mathrm{MD}}_{\mathrm{Xbb}}}{(\mathrm{PN}^{\mathrm{MD}}_{\mathrm{Xbb}} + \mathrm{PN}^{\mathrm{MD}}_{\mathrm{cc}} + \mathrm{PN}^{\mathrm{MD}}_{\mathrm{Xqq}})}$$

Boosted hadronic resonances: background modeling

QCD background estimated in data with rhalphabet method (essentially differential ABCD)

- Simultaneous fit of pass/fail regions in data and MC
- **Constrains background in situ**
- **Only works because tagger is mostly mass-decorrelated**

$$N_{SR}^{QCD}(\rho, p_T) \sim TF^{MC} \times TF^{Data} \times N_{CR}^{QCD}(\rho, p_T)$$

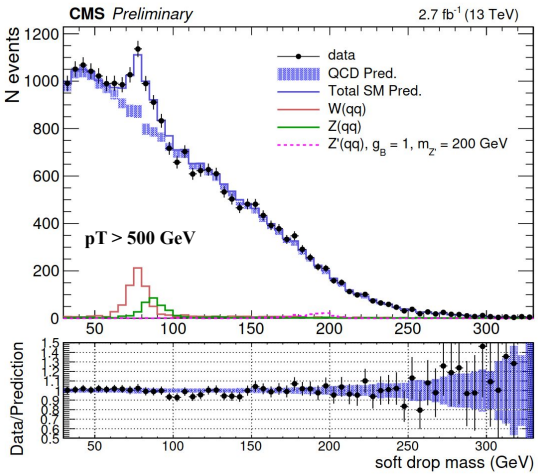
Polynomial transfer function to account for mass-sculpting

Polynomial transfer function to account for residual data/MC differences

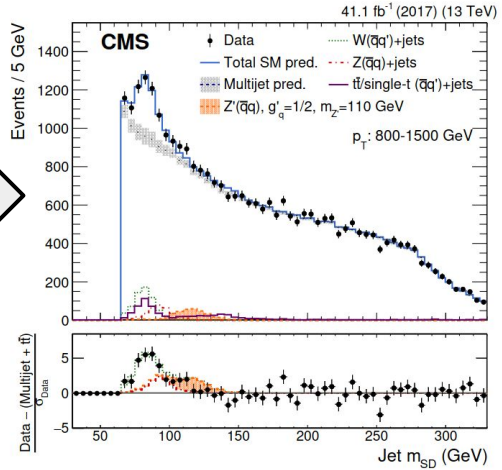
- **Residual mass sculpting and data/MC uncertainty is dominant systematic**

Evolution of boosted dijet tagging

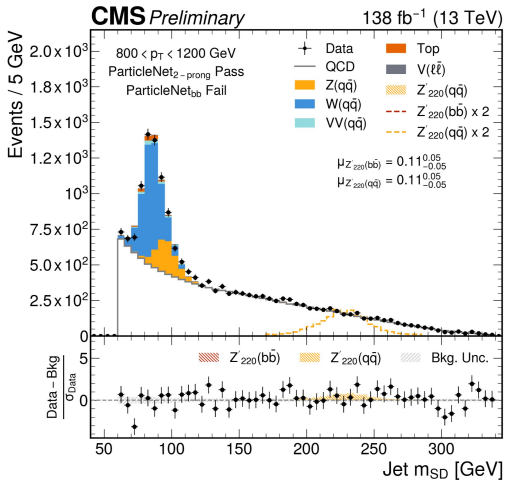
2017

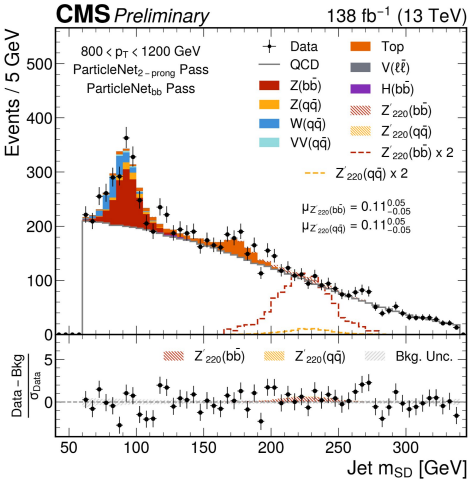


2020



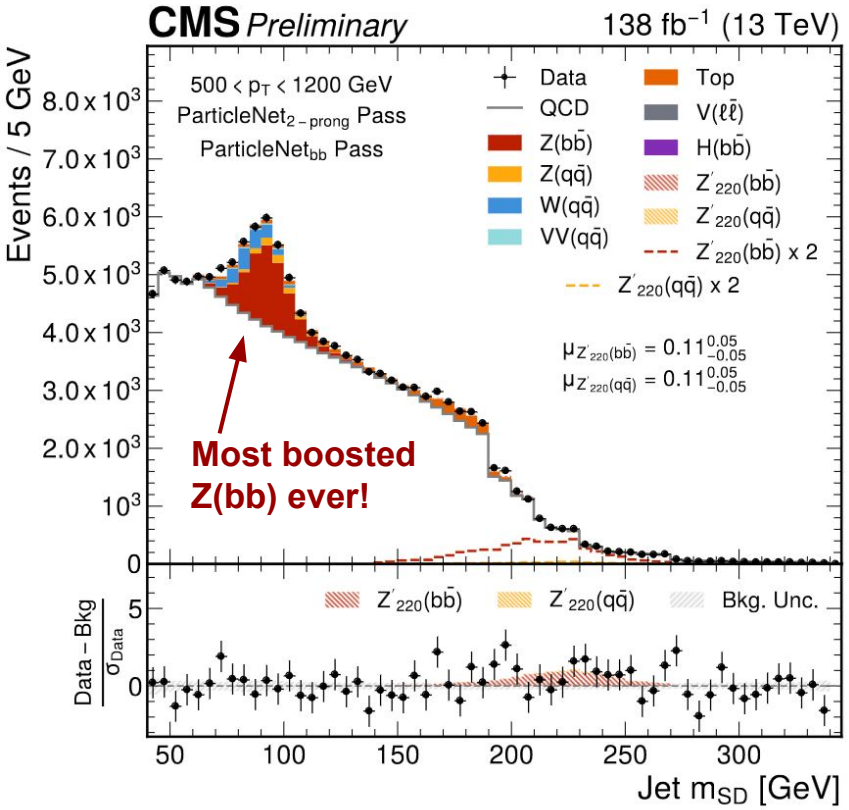
2024

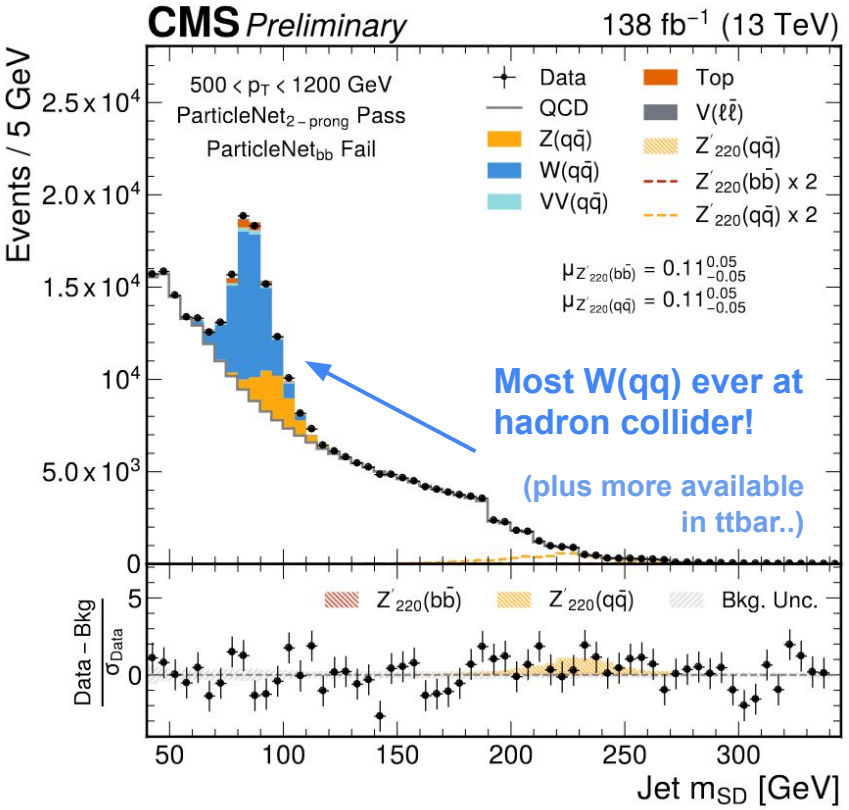




All pT bins in backup

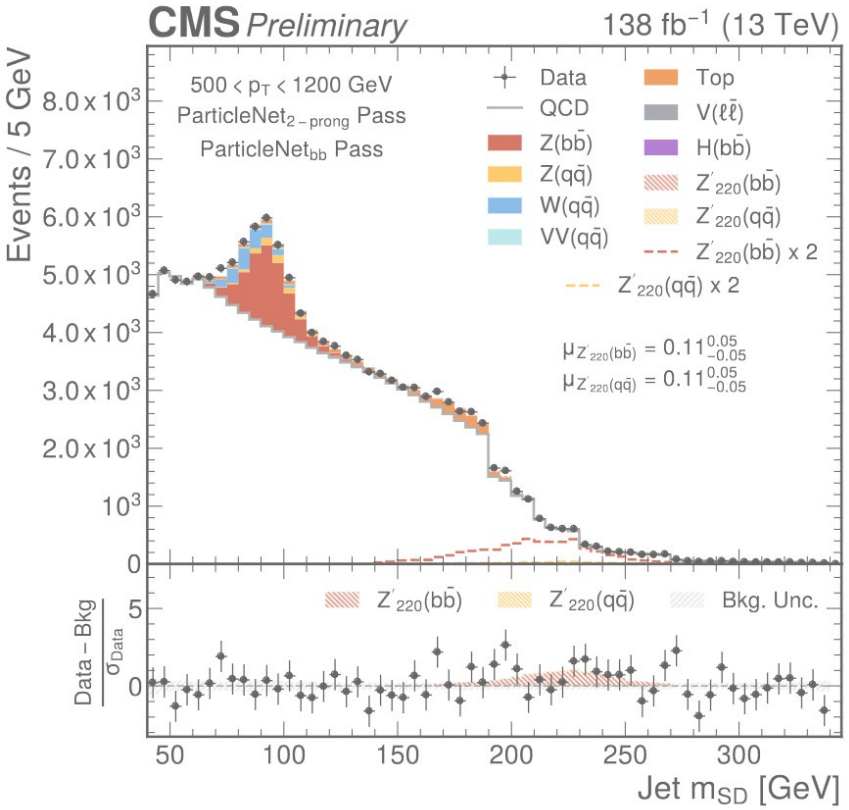
Boosted hadronic resonances: results

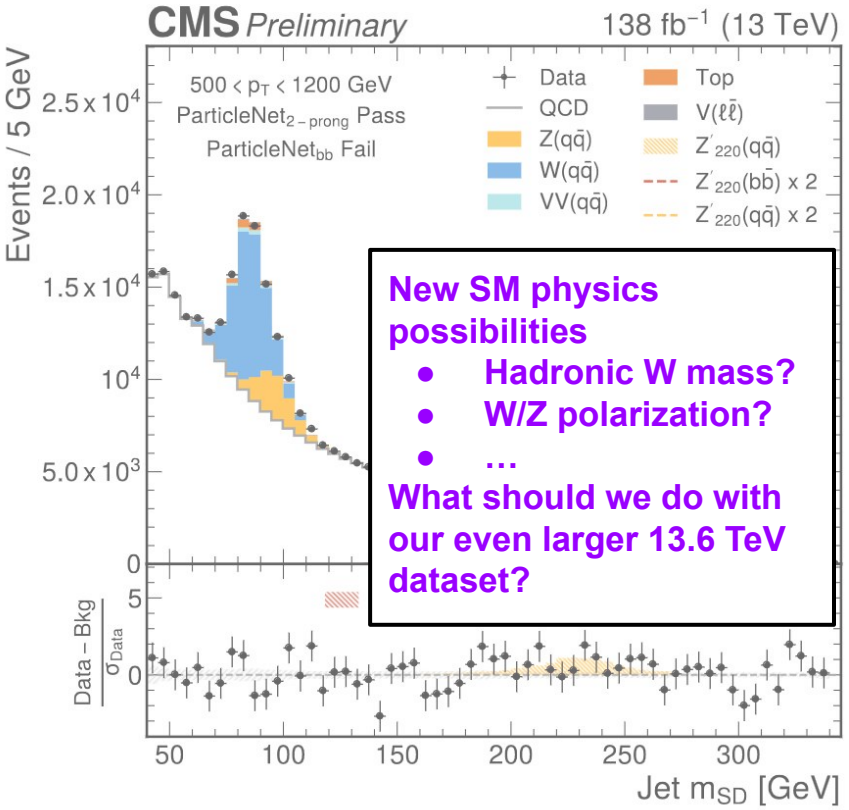




All pT bins in backup

Boosted hadronic resonances: results





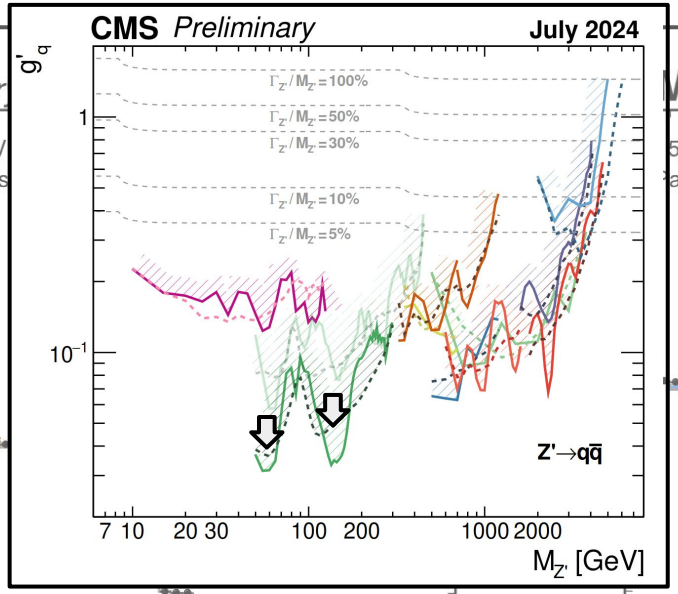
New SM physics possibilities

- **Hadronic W mass?**
- **W/Z polarization?**
- **…**

What should we do with our even larger 13.6 TeV dataset?

All pT bins in backup

Boosted hadronic resonances: results



World-leading limits on BSM

- Interpret in terms of vectors, scalars, pseudoscalars
- Limits on g_q scale like lumi^{-4}
- New techniques are driving improvements!

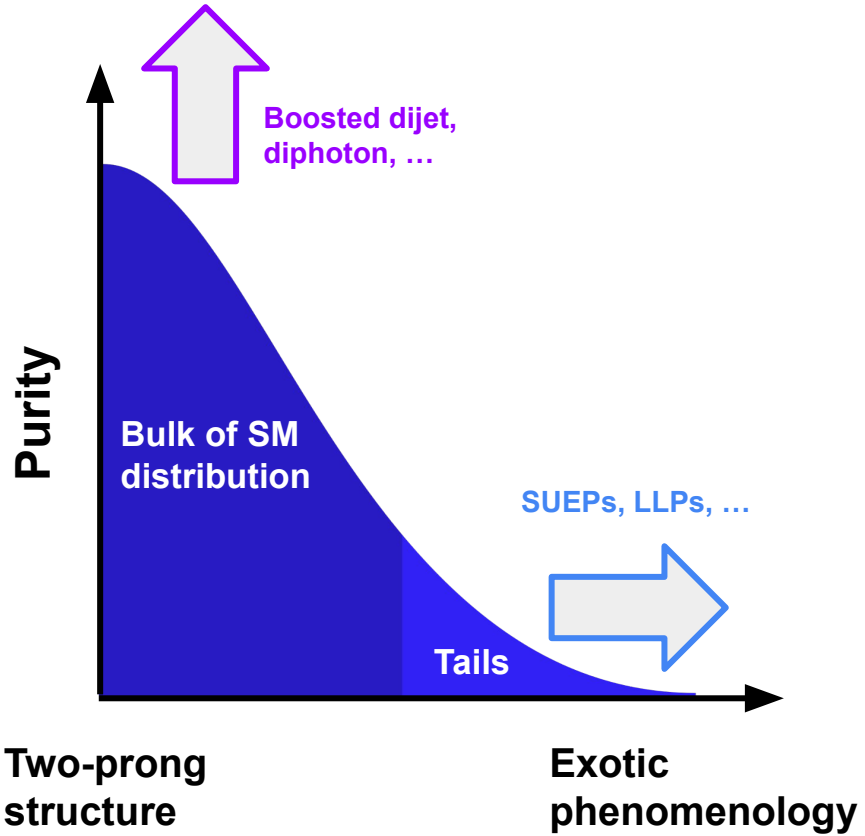
New SM physics possibilities

- Hadronic W mass?
- W/Z polarization?
- …

What should we do with our even larger 13.6 TeV dataset?

Conclusions

- Lots of advances in triggering, reconstruction, tagging, analysis techniques, …
- Upshot is dramatic improvements in analysis sensitivity
- In addition we have new sensitivity to exotic signatures that probe the tails of our understanding
- We are only continuing to develop and refine these techniques - expect even more exciting results for BOOST 2025!

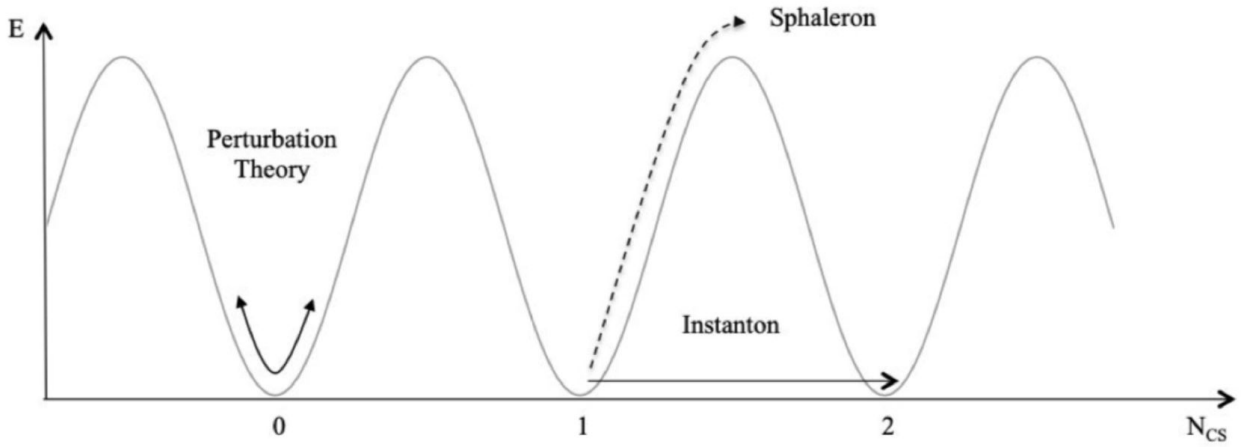


Backup

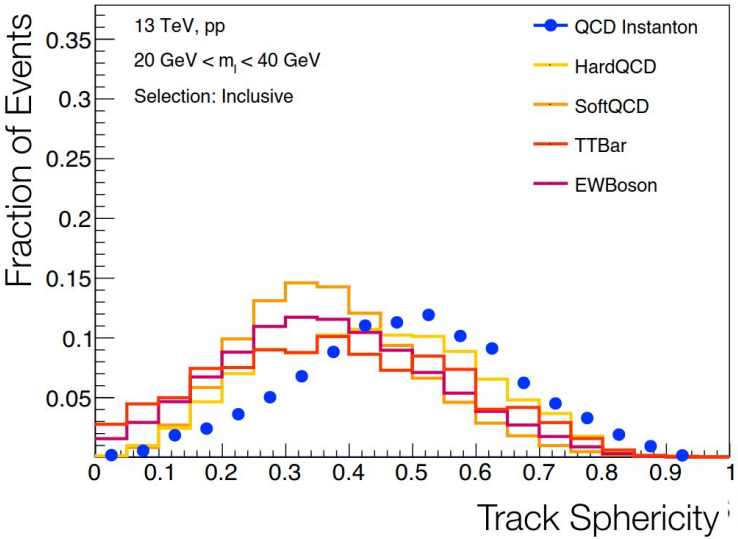
QCD instanton at LHC: theory

Yang-Mills theories have topologically non-trivial vacuum -> can have tunneling between minima

This is not expressible in perturbation theory



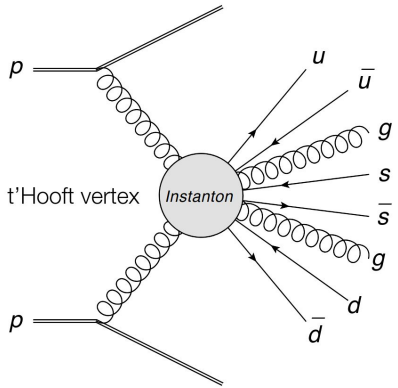
QCD instanton at LHC: practicalities



Signature is lots of quarks and gluons, similar to SUEPs

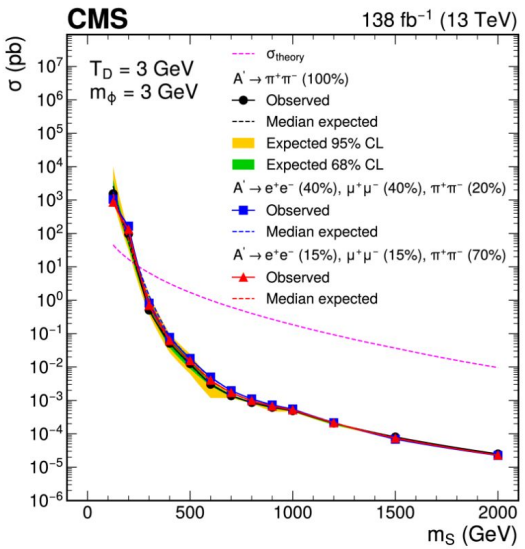
Problem/difference is can't boost against ISR

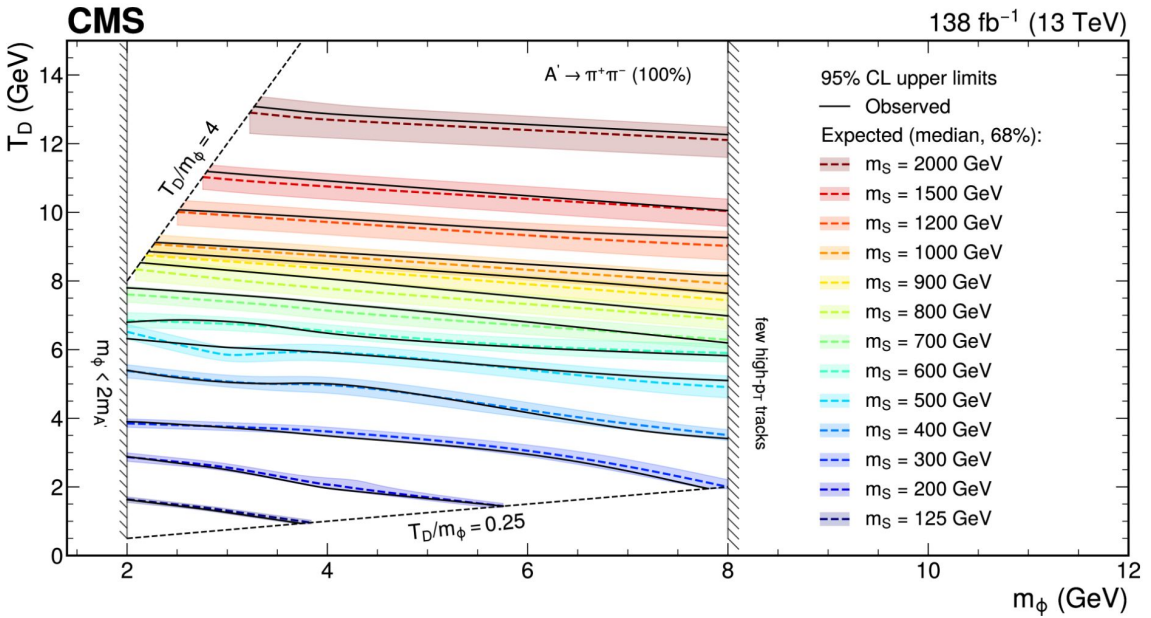
See talk by Simone Amoroso for more details



SUEP limits

Benchmark model has four parameters (shower temperature, scalar mass, dark photon mass, dark meson mass), so four-dimensional limits (difficult to visualize)



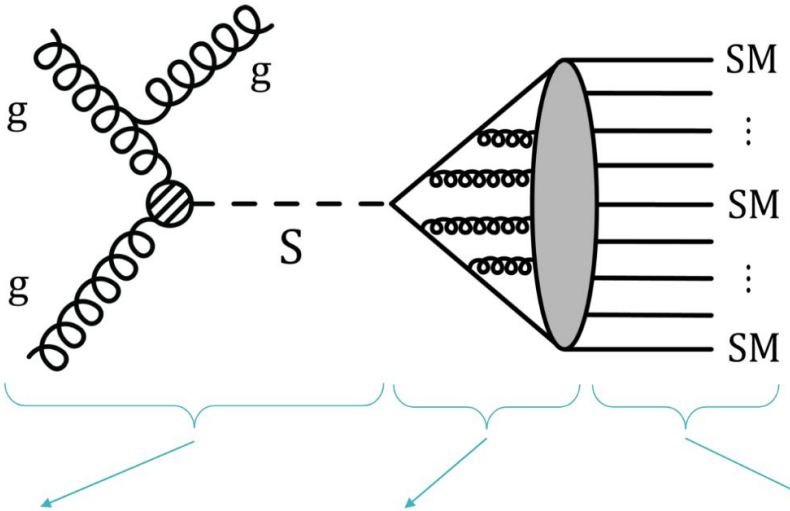


From talk at APS DPF
by Luca Lavezzo

SUEP model parameters

Model Parameters

* Thanks to Simon Knapen for putting together a suep_generator!



m_S
- Affects final state multiplicity

m_ϕ, T_D
- Affect momenta and number of particles produced

$m_{A'}$
- Demanding promptness, can fix A'-SM coupling: $m_{A'}$ free param.
- Varies what type of charged particle is produced

LLP jet triggers

Displaced-jets triggers have been significantly improved in Run 3, in order to greatly increase the efficiencies for low-mass LLPs:

- Removing/loosening the online displaced track requirements;
- Tuning the prompt-track-veto requirement to control the trigger rates;

New Run-3 displaced jets triggers

L1HTT-seeded path (main):

- HLT $H_{\mathrm{T}} > 430\mathrm{GeV}$
- At least two online CALO jets, each satisfying:
  - $p_{\mathrm{T}} > 40\mathrm{GeV}$, $|\eta| < 2.0$;
  - At most 1 track with $\mathrm{IP_{2D}} < 0.5\mathrm{mm}$, $\mathrm{Sig[IP_{2D}]} < 5.0$, $p_{\mathrm{T}} > 1\mathrm{GeV}$

Matched with offline $H_{\mathrm{T}} > 450\mathrm{GeV}$

L1Mu6HTT240-seeded path (auxiliary):

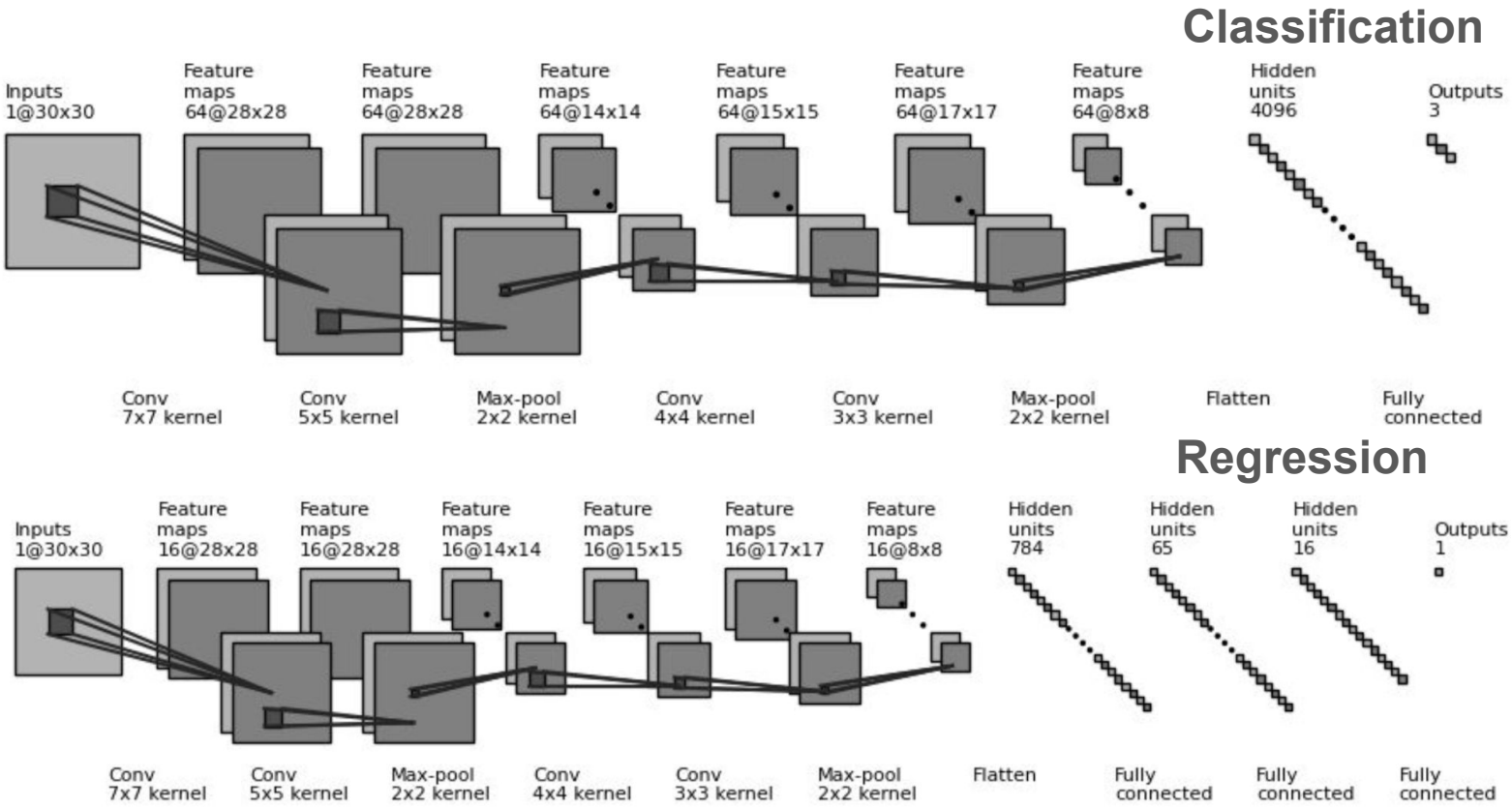
- At least two online CALO jets, each satisfying:
  - $p_{\mathrm{T}} > 40\mathrm{GeV}$, $|\eta| < 2.0$;
  - At most 1 track with $\mathrm{IP_{2D}} < 0.5\mathrm{mm}$, $\mathrm{Sig[IP_{2D}]} < 5.0$, $p_{\mathrm{T}} > 1\mathrm{GeV}$;
    - ✦ If there is exactly one associated prompt track, further require at least one track satisfying $\mathrm{IP_{2D}} > 0.3\mathrm{mm}$

Matched with offline $H_{\mathrm{T}} > 240\mathrm{GeV}$

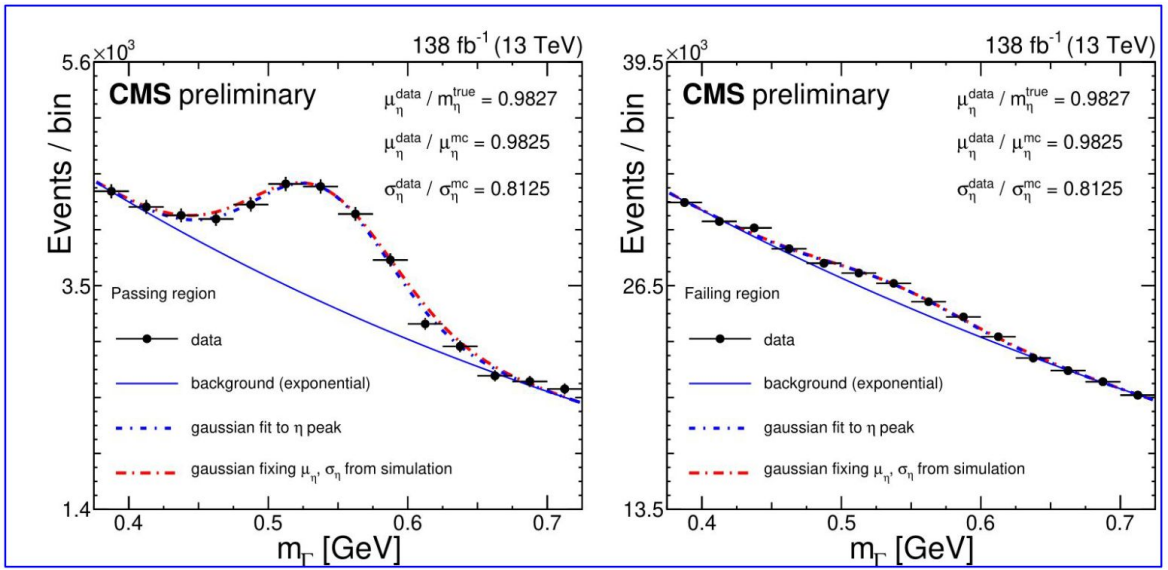
Hadronically-decaying LLPs: training the taggers

- **Signal sample:** S -> bb decays
  - m_S = 30, 40, 55 GeV
  - $c\tau$ = 1, 10, 100, 1000 GeV
- **Background sample:** QCD MC
- GNN_d working point is 10^4 background rejection and 50% signal efficiency
- GNN_p working point is 10^3 background rejection and 30% signal efficiency

Mergd diphoton: ML architectures



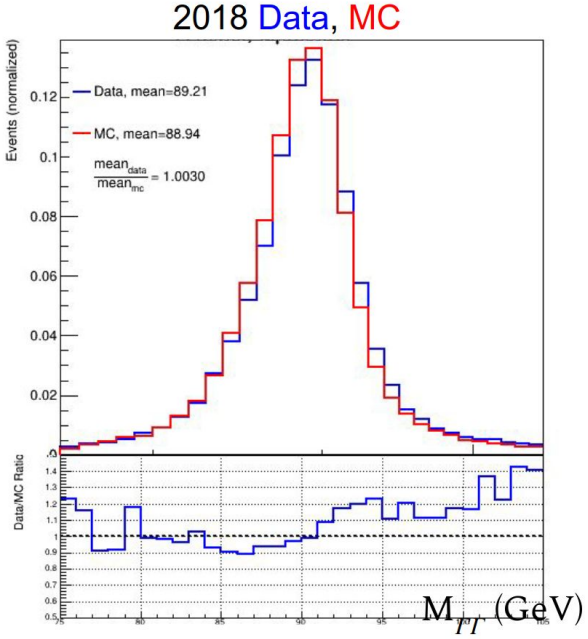
Merged diphoton: validation



Data/MC Fit Parameters:

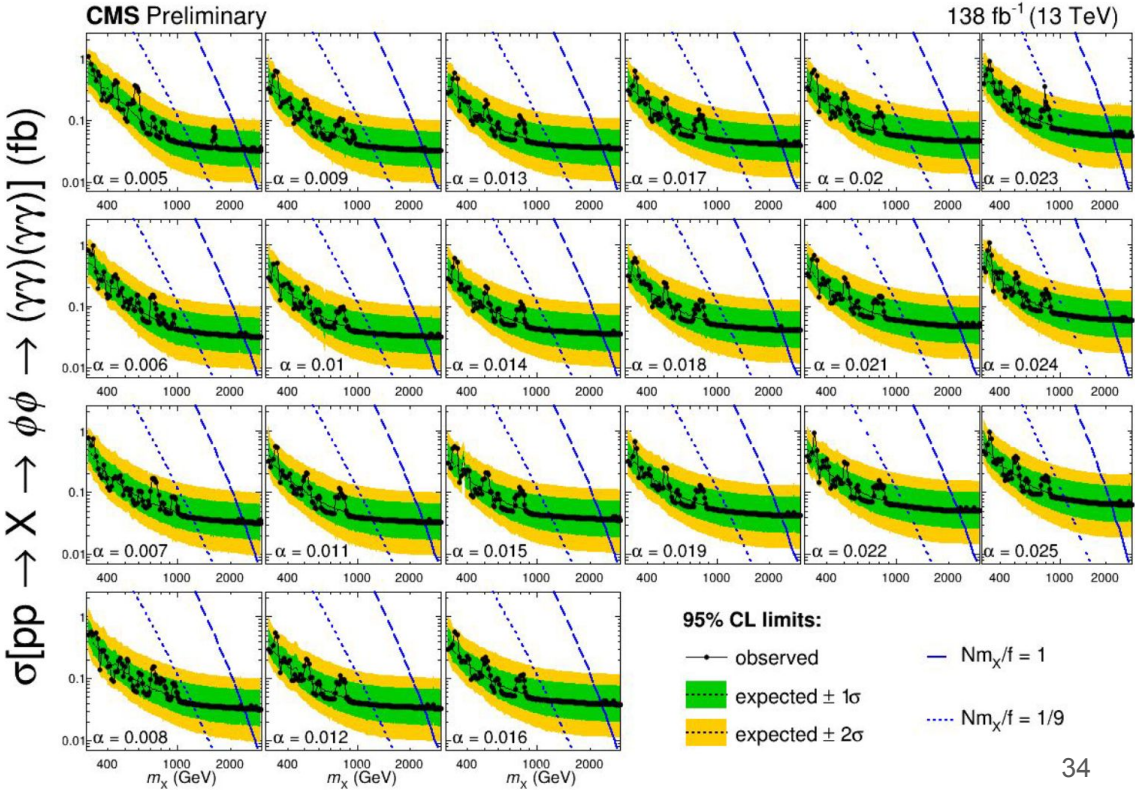
Parameter	Data	GJets MC
μ	0.5384 ± 0.0001	0.5480 ± 0.0127
σ	0.0403 ± 0.0001	0.0496 ± 0.0113

14



Merged diphoton: limits

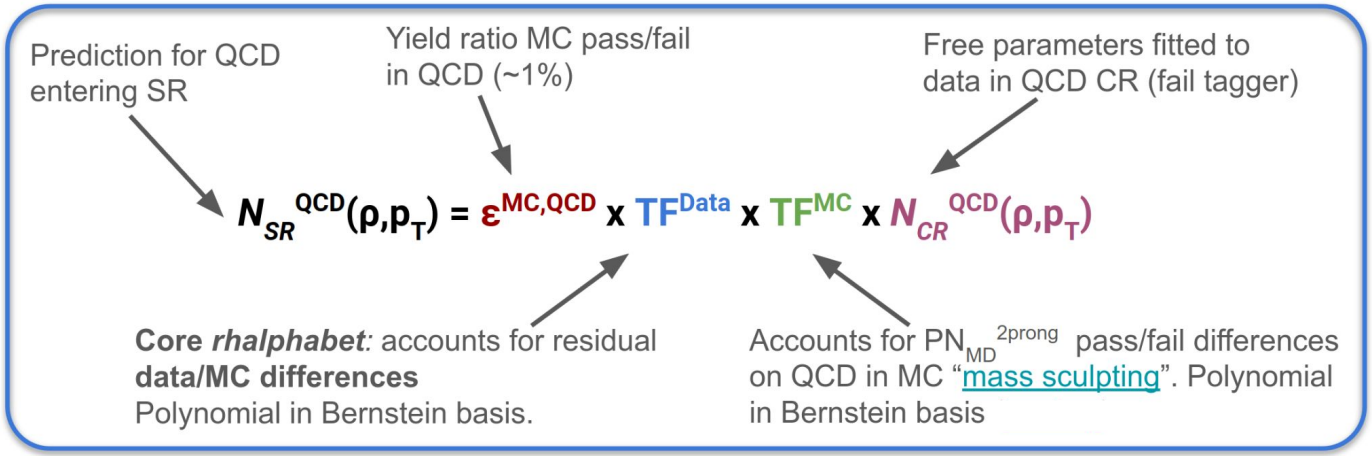
- Estimate background by fitting smooth falling QCD background
- Use resolution/scale parameters from eta and Z control regions
- No signal observed



Boosted hadronic resonances: background modeling (1)

QCD background estimated in data with rhalphabet method (essentially differential ABCD)

- Simultaneous fit of pass/fail regions in data and MC
- **Only works because tagger is mostly mass-decorrelated**

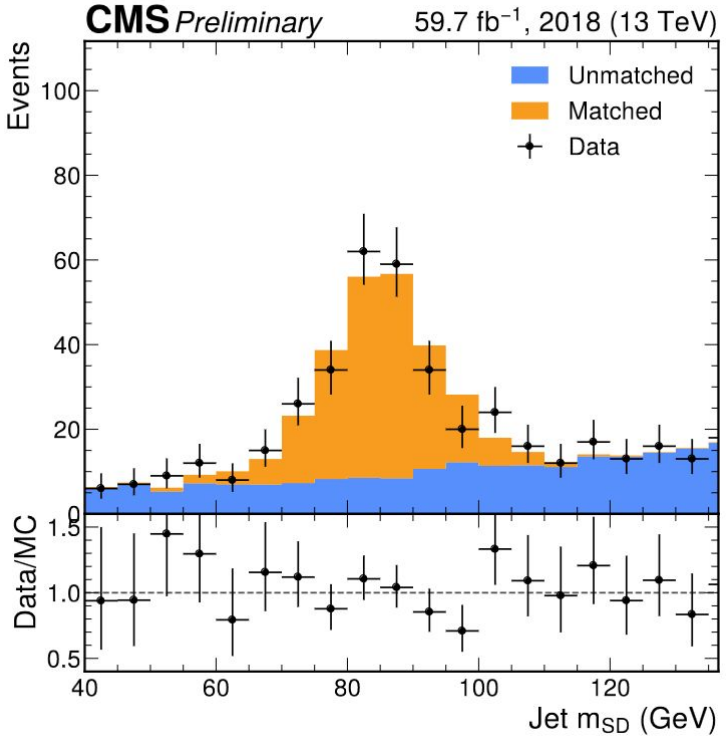


- **Residual mass sculpting and data/MC uncertainty is dominant systematic**

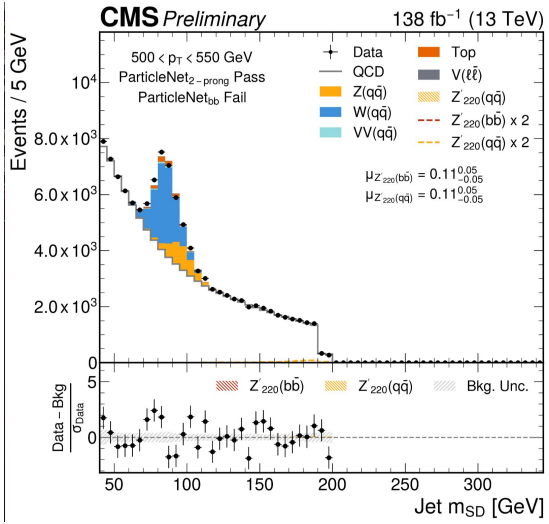
Boosted hadronic resonances: background modeling (2)

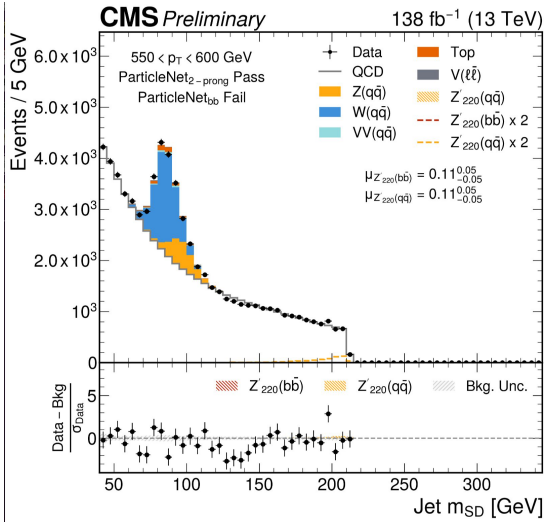
Resonant backgrounds estimated in MC with dedicated corrections

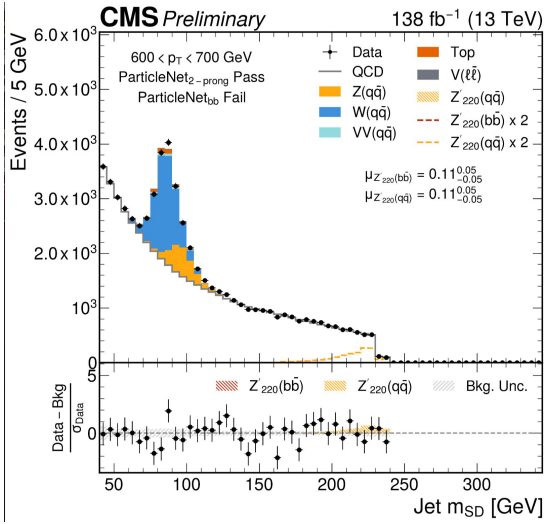
Dedicated W(qq) control region from ttbar semileptonic decays



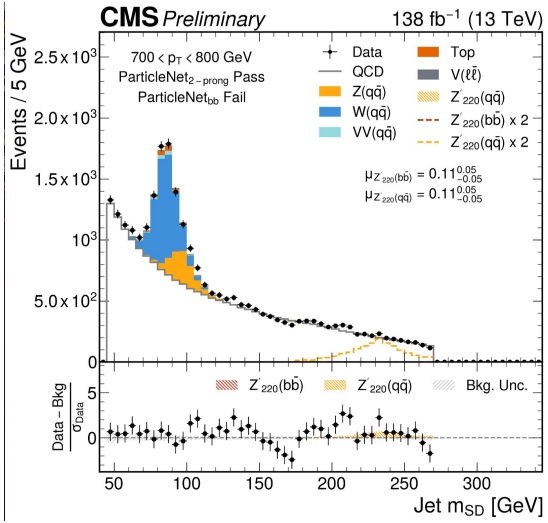
All pT bins: b tag fail (1)

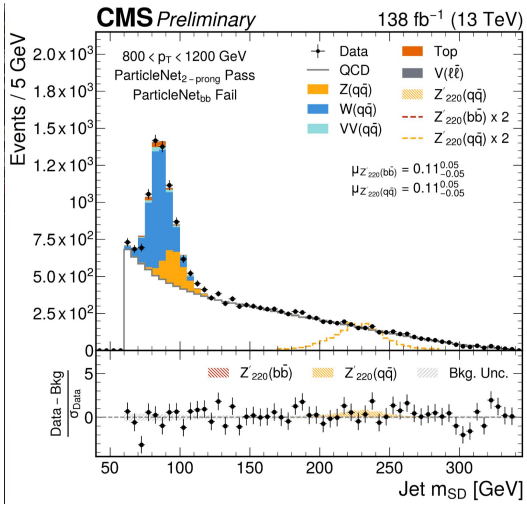




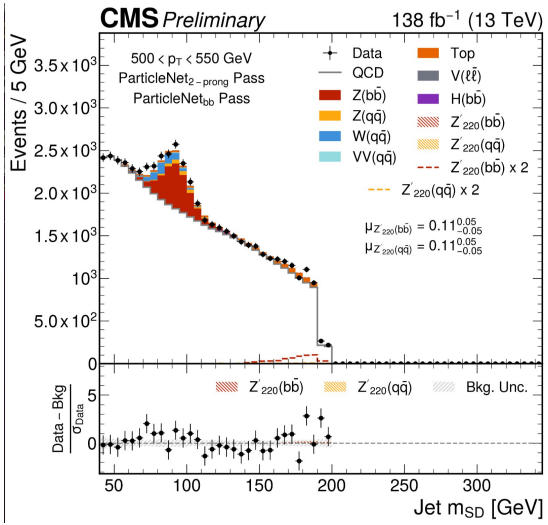


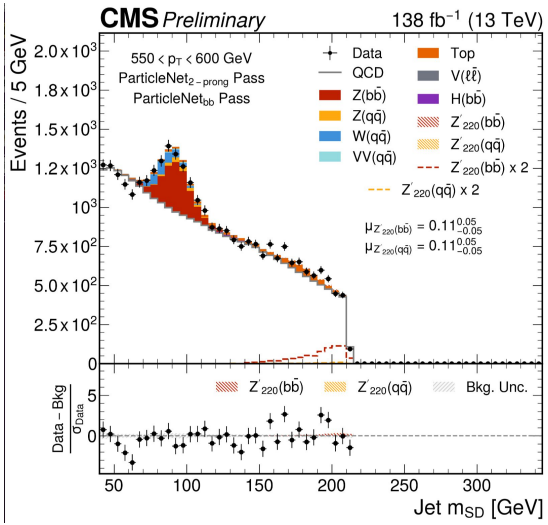
All pT bins: b tag fail (2)

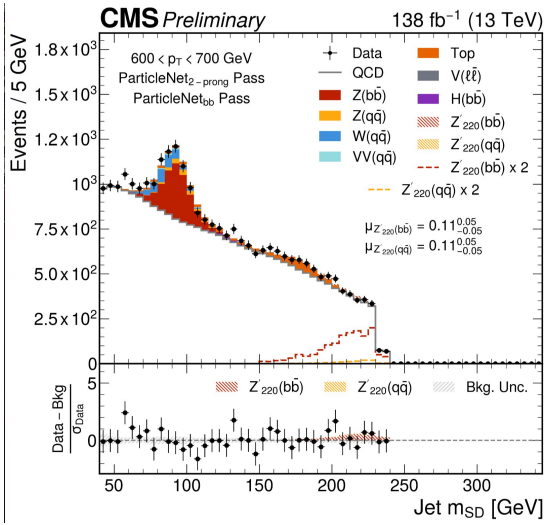




All pT bins: b tag pass (1)







All pT bins: b tag pass (2)