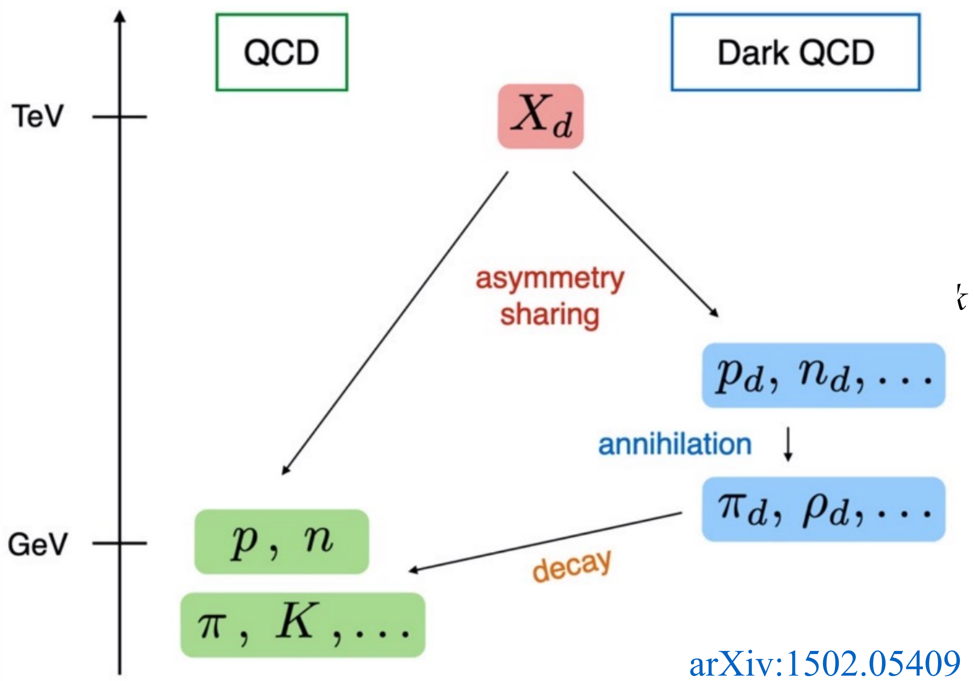
Search for emerging jets with graph neural networks using CMS Run 2 data

Claire Savard on behalf of the CMS collaboration
University of Colorado, Boulder
19 February 2024

Emerging jets (EMJ) theory

- Dark matter model = QCD-like hidden sector
- Dark hadrons with ~ Λ_{dark} (GeV), dark pions unstable $m_{\pi_{dark}} < \Lambda_{dark}$
- Heavy mediator particle ~ TeV couples to dark and visible sectors

- Energy scales reachable at LHC



arXiv:1502.05409

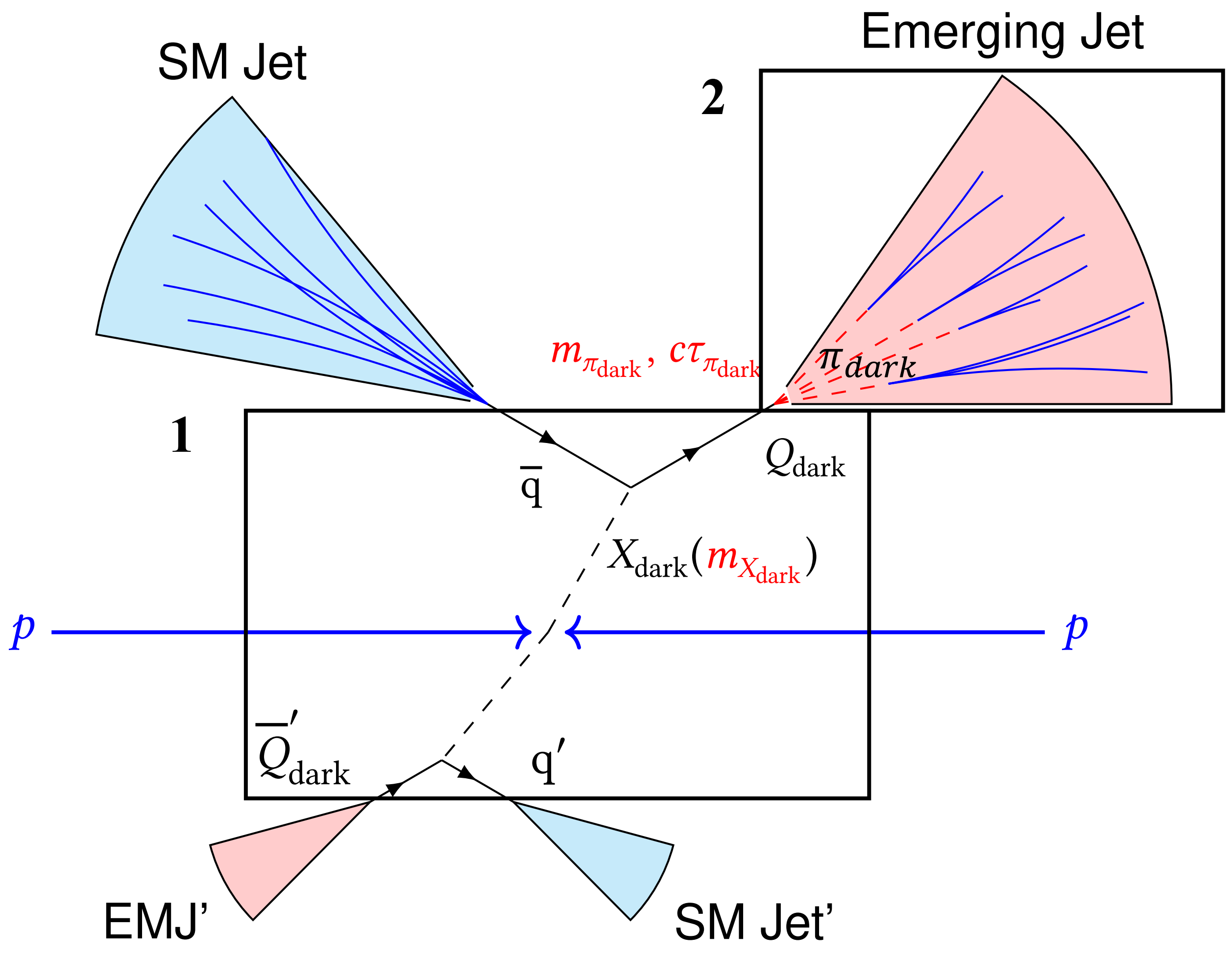
EMJ production at the LHC

1. $pp \rightarrow 2X_{dark} \rightarrow 2(qQ_{dark})$
2. $Q_{dark} \xrightarrow{\text{hadronizes}} N\,\pi_{dark}$ &

   $\pi_{dark} \xrightarrow{\text{travel } c\tau}$ SM particles

Free parameters:

- m_X: [1, 2.5] TeV
- $m_{\pi_{dark}}$: [6, 20] GeV
- $c\tau_{\pi_{dark}}$: [1, 1000] mm

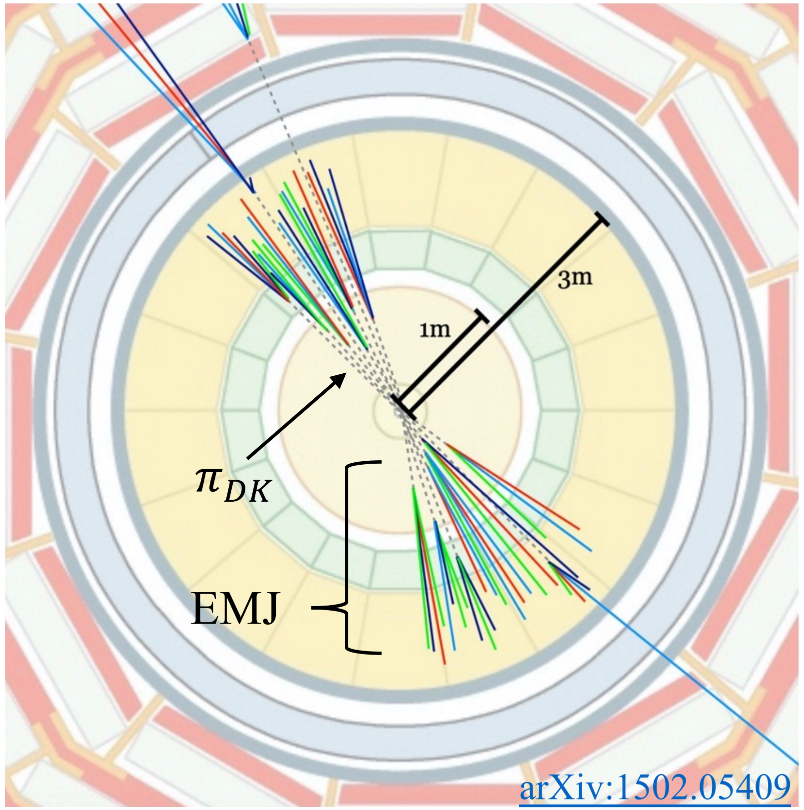


EMJs in CMS detector

- General-purpose particle detector
- Silicon tracker:
    - Charged particle reconstruction
    - Within 3.8 T solenoid for momentum resolution
    - Vertex z_0 resolution of ~15 μm
    - Extends from collision point to 1 m

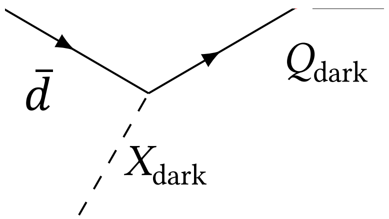
- $c\tau_{\pi_{dark}}$ 1 – 1000 mm, contained in tracker

CMS cross section

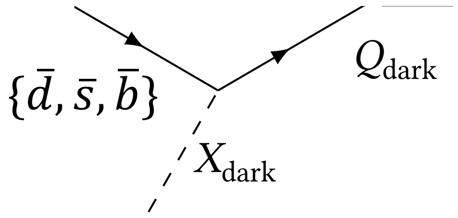


EMJ coupling scenarios

1. Unflavored down scenario
   - Dark quarks couple to down quarks ONLY
   - All π_{dark} have same $c\tau$
   - Previous CMS search (arXiv:1810.10069)



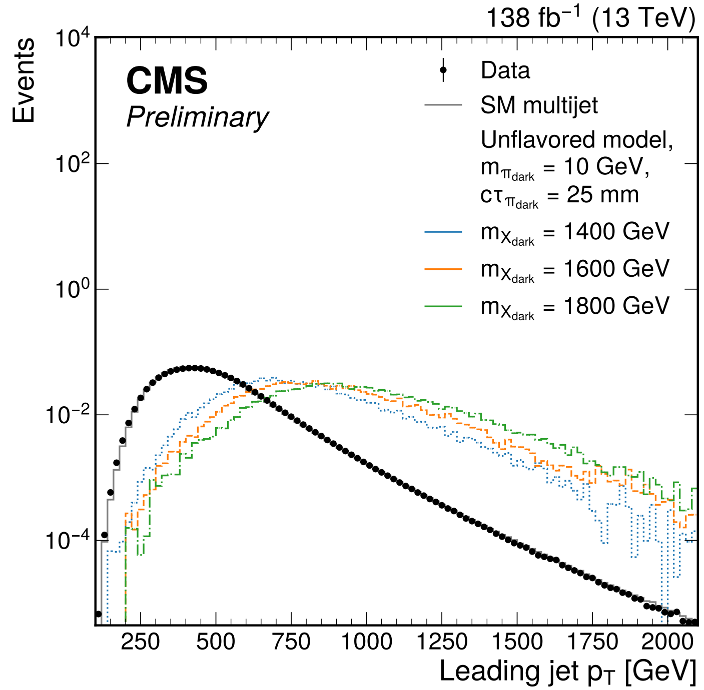
2. Flavor-aligned down scenario
   - Dark quarks couple to down-type quarks ONLY (d, s, b)
   - π_{dark} lifetime depends on dark pion composition

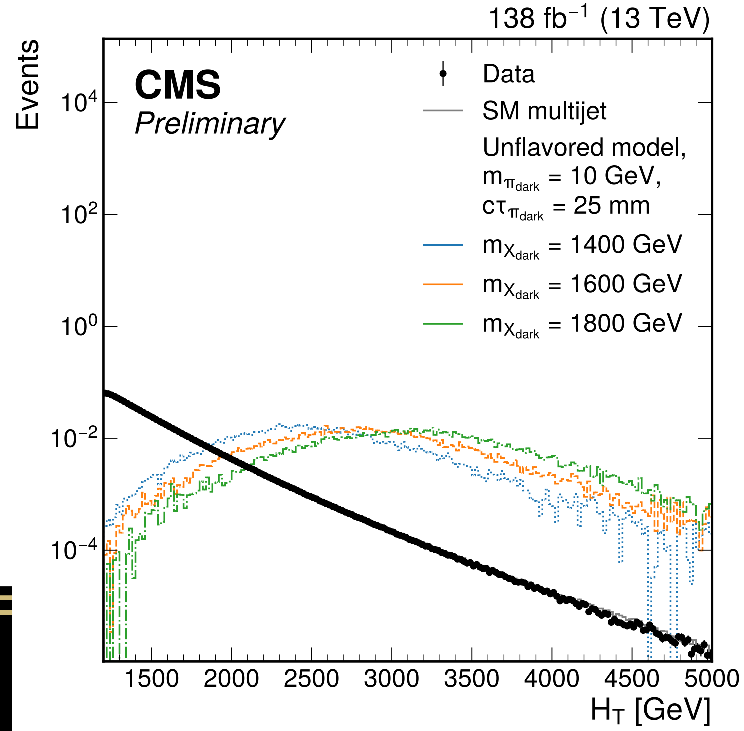


- Scenarios phenomenologically different, search methods tuned separately

Event selection

- ≥ 4 high p_T jets
- High event H_T ($= \sum_{jets} p_T$)
- ≥ 2 EMJ-tagged jets
  - Use graph neural network (GNN)

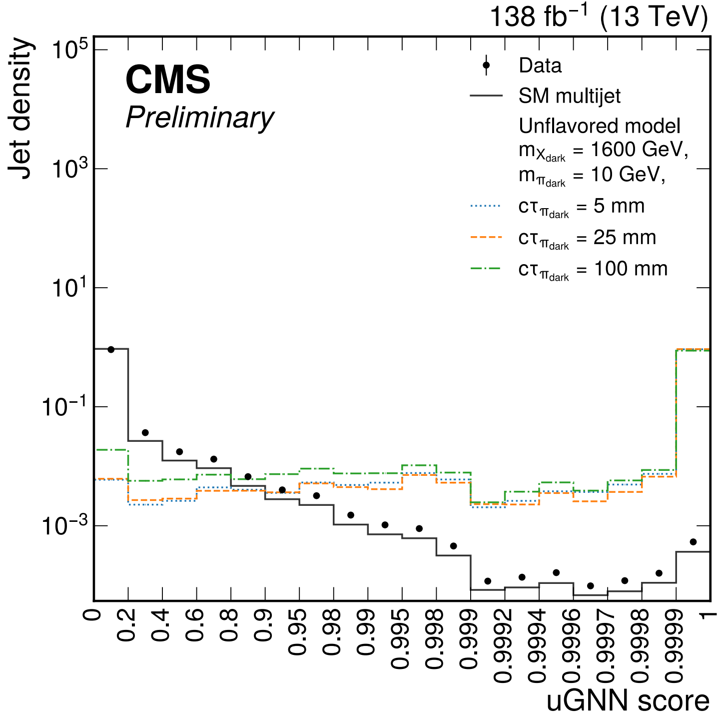


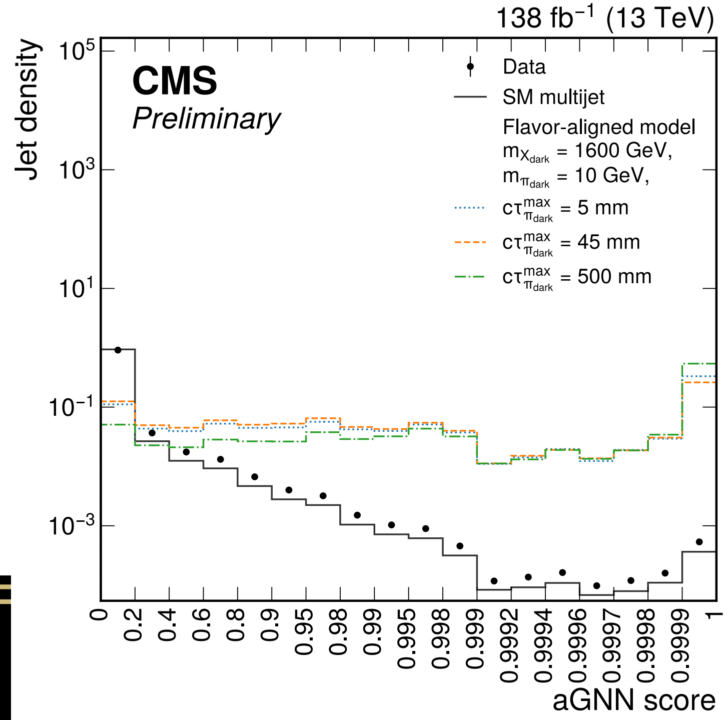


GNN EMJ tagger

- Associate tracks to jets within 0.8 angular separation from jet axis
- Track coordinates with respect to jet axis ($\Delta\phi$, $\Delta\eta$)
- Track features within jet:
  - Angular separation ΔR
  - p_T and p_T fraction
  - Impact parameters d_{xy} and d_z

- GNN score output used to classify
- 2 GNNs: unflavored and aligned



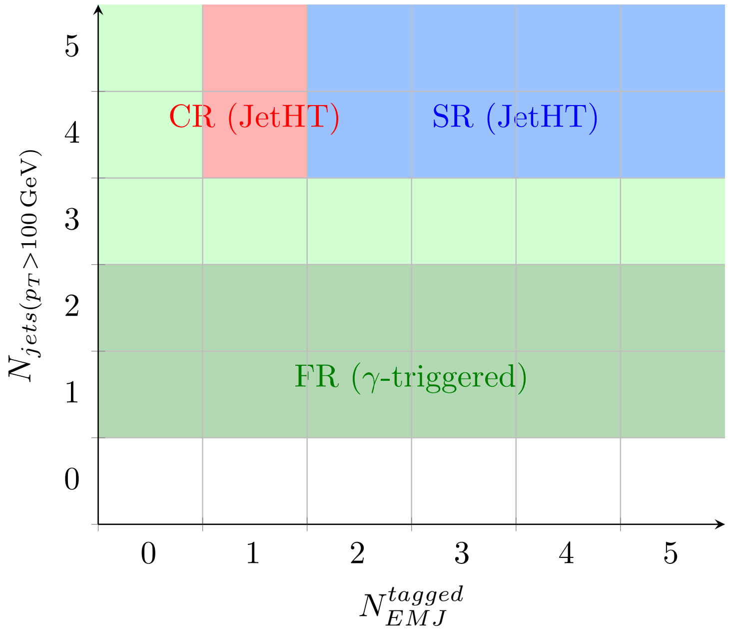


Event selection criteria

- Different ($m_{X_{dark}}$, $m_{\pi_{dark}}$, $c\tau$) regions have slightly different kinematics
- Each EMJ sample assigned 1 of 6 selection criteria

	Selection set	H_T [GeV] (>)	Jet p_T [GeV] (>)	GNN score (>)
Unflavored	uGNN set 1	1350	170, 120, 120, 100	0.9997
	uGNN set 2	1750	300, 260, 250, 250	0.9998
	uGNN set 3	1800	240, 180, 180, 100	0.9996
Flavor-aligned	aGNN set 1	1300	200, 140, 120, 100	0.9953
	aGNN set 2	1650	300, 250, 200, 200	0.9993
	aGNN set 3	1400	270, 220, 220, 120	0.9983

Background estimation



Estimate # of bkg. events pass into SR using CR events and mistag rates from FR

FR

$$N_{\mathrm{SR}} = \sum_{evt \in \mathrm{CR}} SF \sim \sum_{evt \in \mathrm{CR}} \frac{1}{2} \sum_{j \notin tagged} \epsilon(p_{T,j})$$

CR

- Fully data-driven estimation
- Mistag rate (ϵ) binned along jet p_T

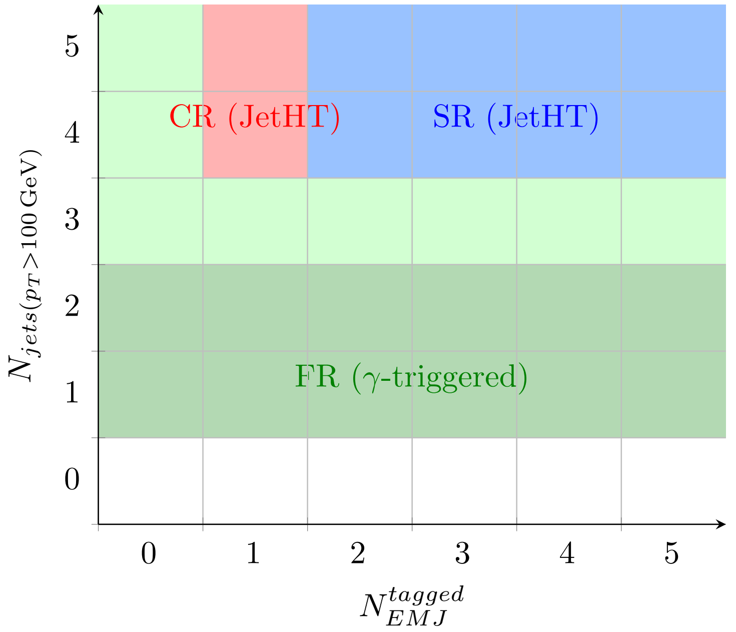
Background estimation

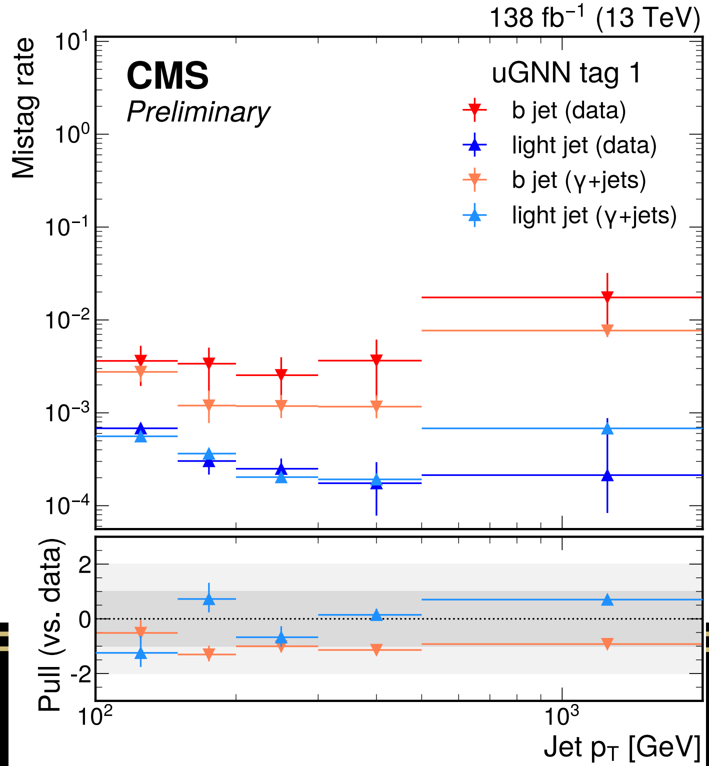
Estimate # of bkg. events pass into SR using CR events and mistag rates from FR

$$N_{\text{SR}} = \sum_{evt \in \text{CR}} SF \sim \sum_{evt \in \text{CR}} \frac{1}{2} \sum_{j \notin tagged} \overset{\text{FR}}{\epsilon}(p_{T,j})$$

CR

- Fully data-driven estimation
- Mistag rate (ϵ) binned along jet p_T





Background estimation

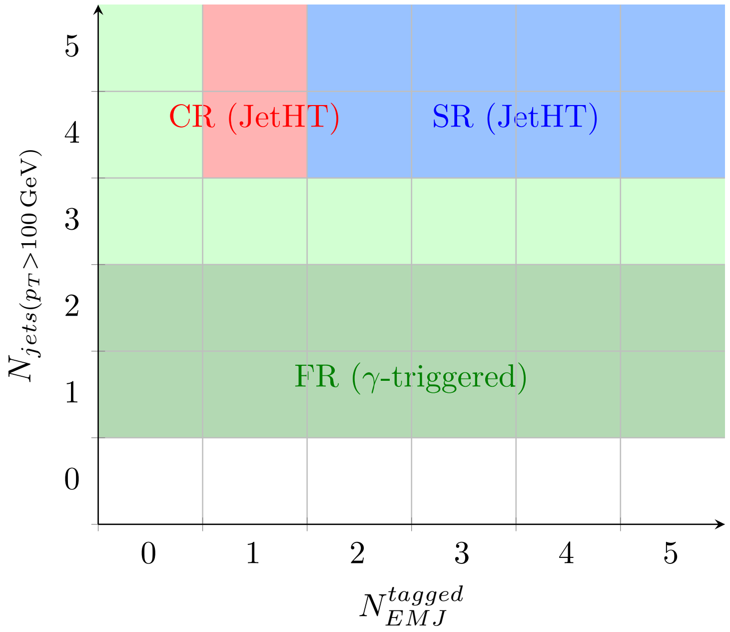
Estimate # of bkg. events pass into SR using CR events and mistag rates from FR

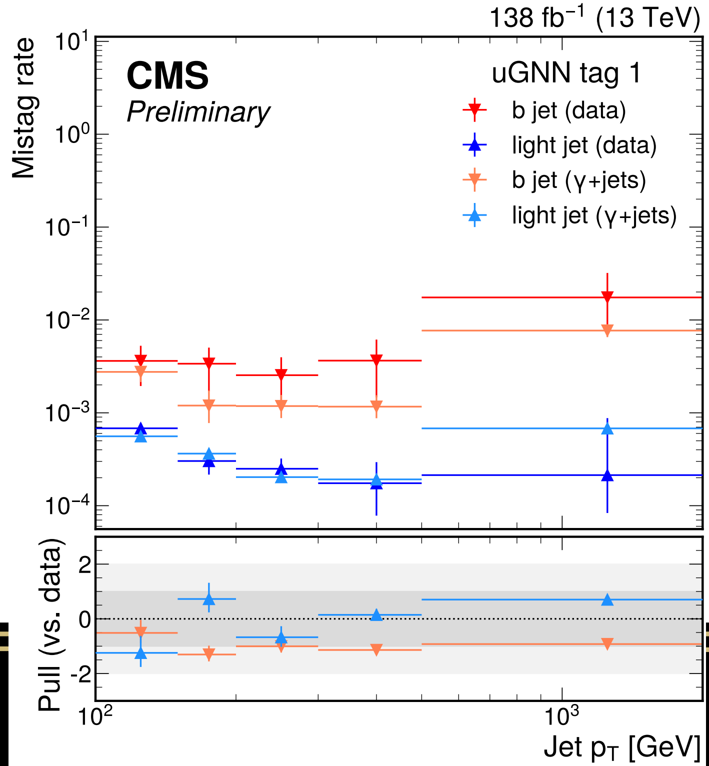
$$N_{\mathrm{SR}} \sim \sum_{evt \in \mathrm{CR}} \frac{1}{2} \sum_{j \notin \mathrm{tagged}} B^{\mathrm{CR}} \epsilon(b, p_{T,j}) + (1 - B^{\mathrm{CR}}) \epsilon(l, p_{T,j})$$

FR

CR

- Fully data-driven estimation
- Mistag rate (ϵ) binned along jet p_T
- B-jet discriminator to calculate mistag on bs separately and b-jet fraction (B^{CR})





Results

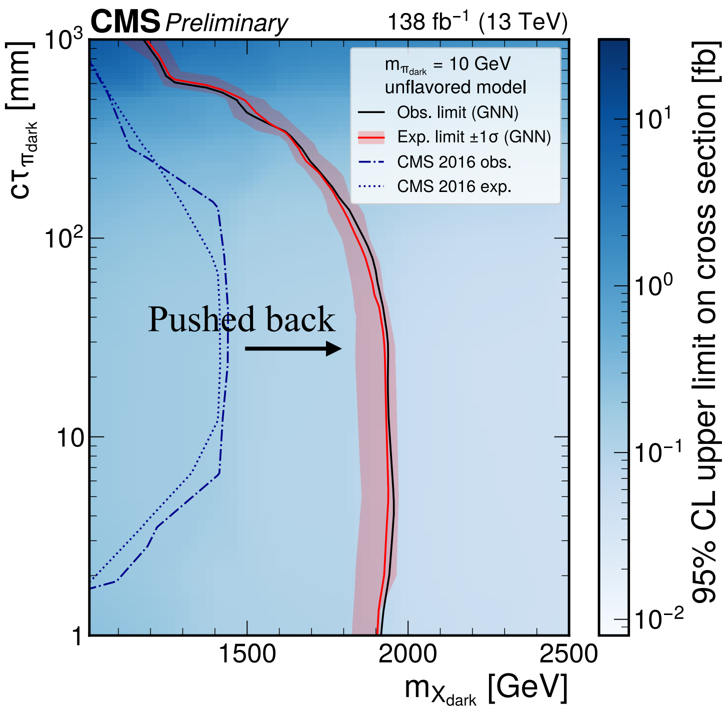
Jet flavor determination is largest systematic uncertainty

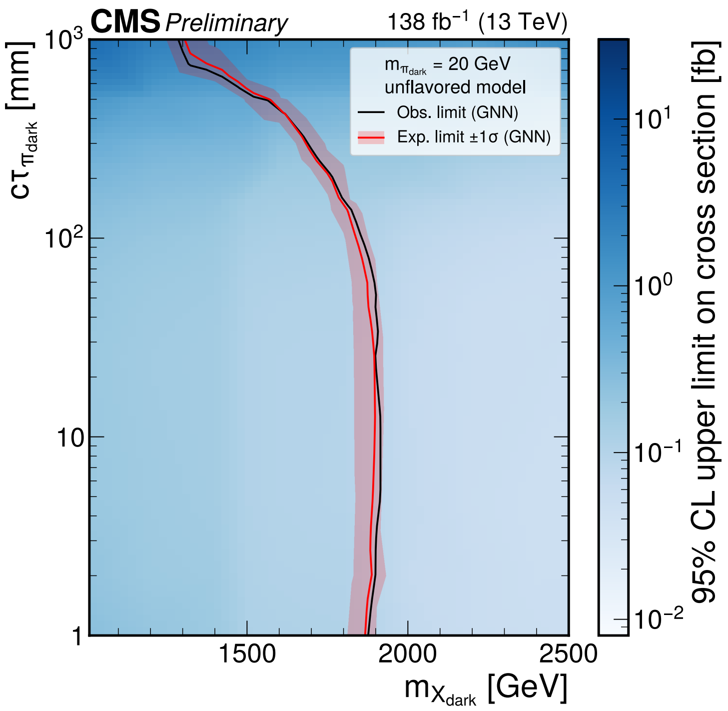
	Selection set	Estimation $^{+\text{stat. up}}_{-\text{stat. dn}}$ $\pm$ syst.	Observed yield
Unflavored	uGNN set 1	$15.6\ ^{+5.4}_{-1.9} \pm 3.8$	18
	uGNN set 2	$0.73\ ^{+0.44}_{-0.16} \pm 0.27$	0
	uGNN set 3	$7.6\ ^{+3.5}_{-1.3} \pm 2.3$	9
Flavor-aligned	aGNN set 1	$45\ ^{+18}_{-8} \pm 16$	59
	aGNN set 2	$0.30\ ^{+0.23}_{-0.07} \pm 0.18$	1
	aGNN set 3	$3.8\ ^{+2.2}_{-0.7} \pm 2.0$	5

No statistically significant excess between estimated and observed # of events

Exclusion limit results - unflavored

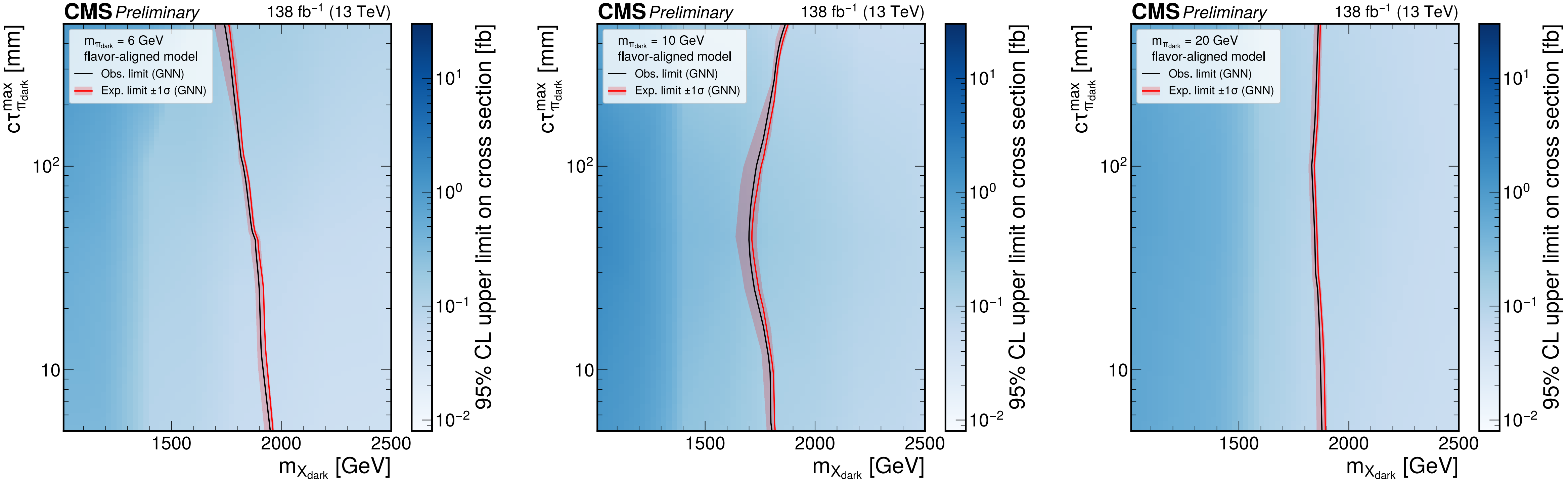
- Limits pushed back by ~ 400 GeV compared to previous publication
- Large gain in sensitivity using GNN at lower $c\tau$





Exclusion limit results - flavored

- Completely new limits
- Exclusion of up to ~1950 GeV in $m_{X_{dark}}$



Conclusions

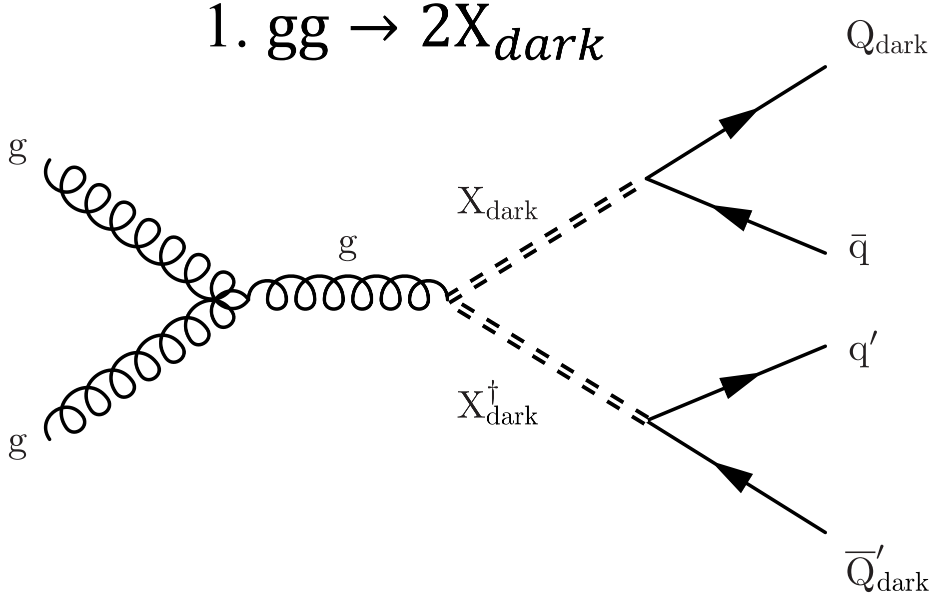
- Pushed back previous limits on unflavored $m_{\pi_{dark}} = 10$ GeV models
- Completely new limits for:
  - Unflavored $m_{\pi_{dark}} = 20$ GeV
  - All flavor-aligned models
- One of first analyses to use GNN tagger
  - Developed traditional cut-based algorithm in parallel and limits of GNN pushed back by 150 – 600 GeV in $m_{X_{dark}}$ everywhere

Backup

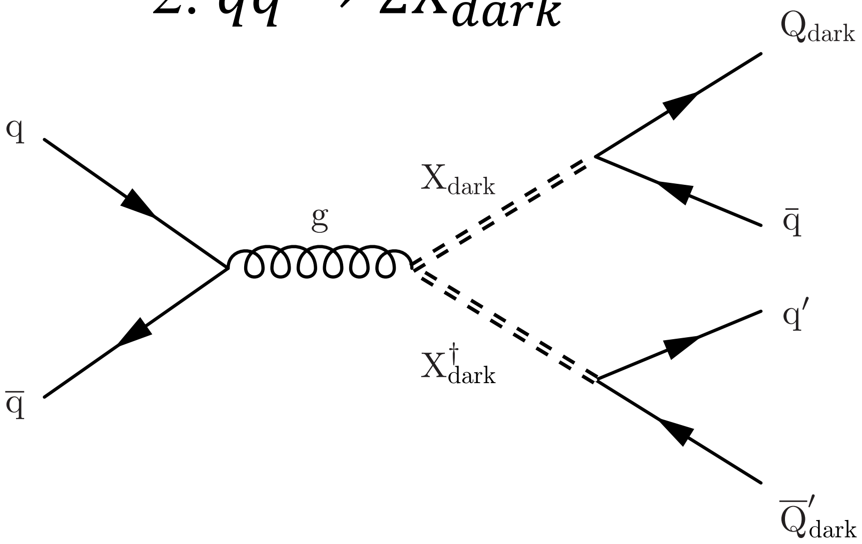
EMJ production at LHC

- 2 methods of X_{dark} production:

1. $\mathrm{gg} \rightarrow 2\mathrm{X}_{dark}$

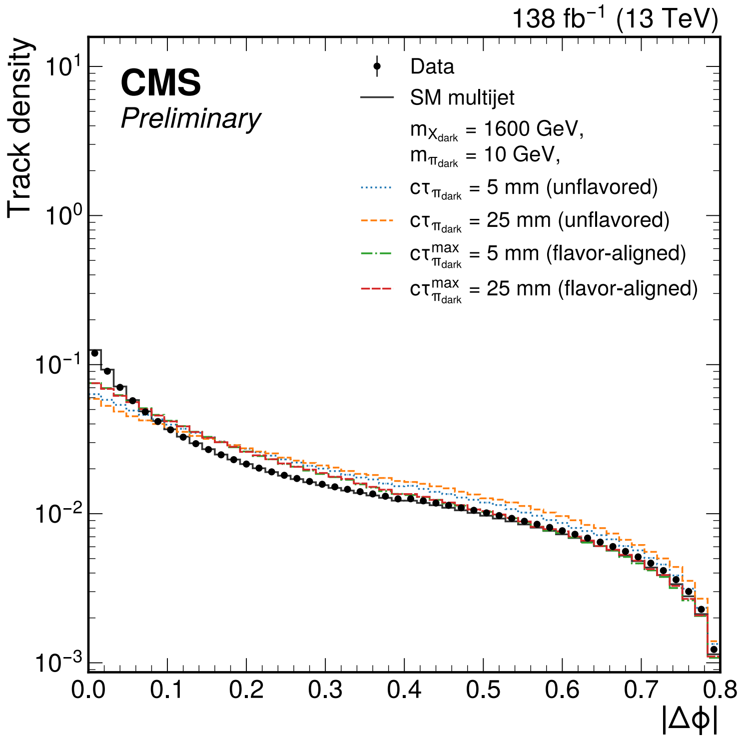


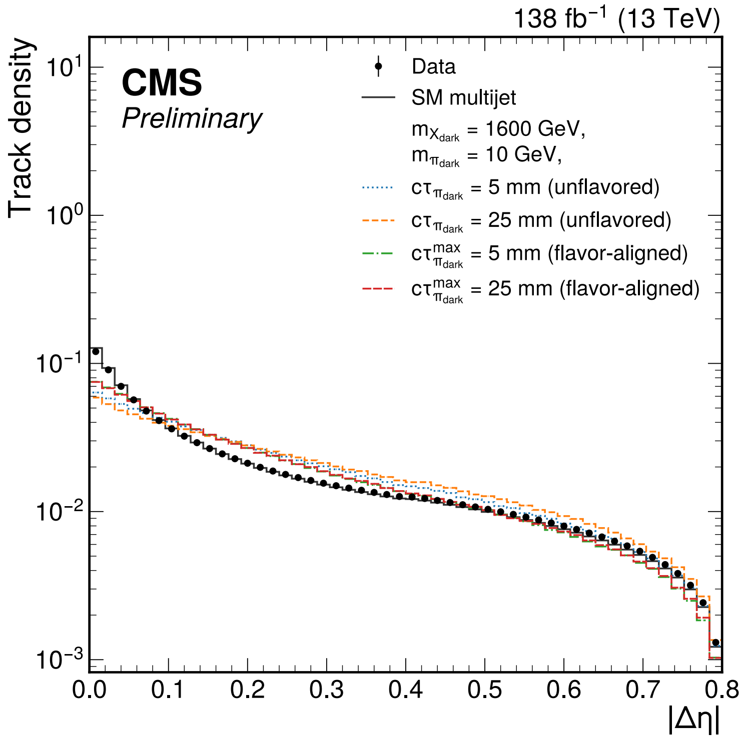
2. $qq \rightarrow 2\mathrm{X}_{dark}$



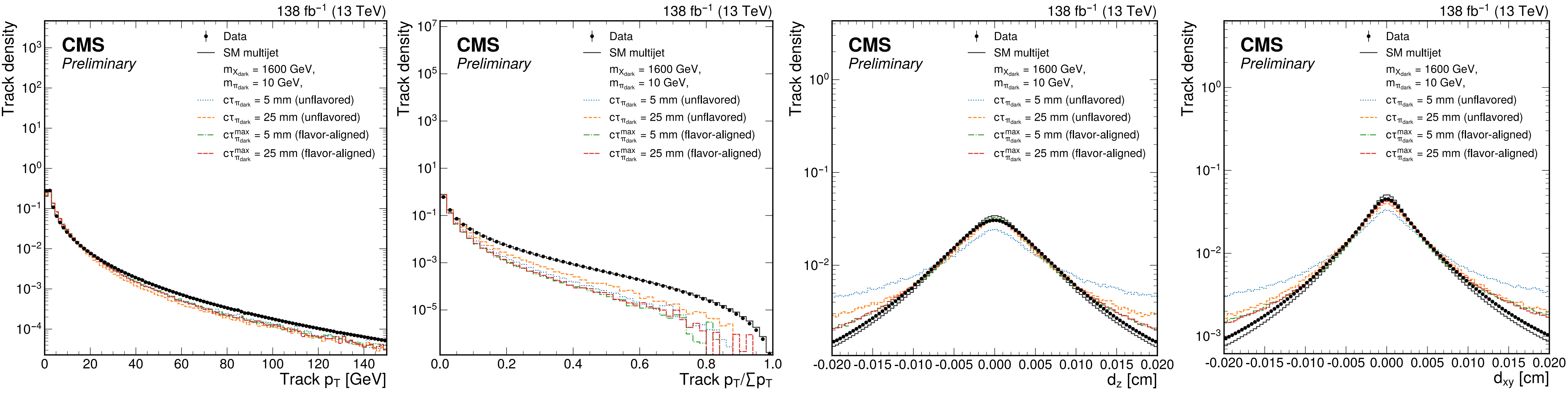
- Pairs of X_{dark} from gluon (g), decay to visible (q) and dark quarks ($\bar{Q}_{dark}$)

GNN tagger coordinates





GNN tagger features



Mistag rate calculations

2 equations, 2 unknowns: solve for $\epsilon(b/l, p_T)$
(inverse method)

Mistag rate equations:

$$FR_E: \quad \epsilon^E(p_T) = B^E(p_T)\boxed{\epsilon(b, p_T)} + (1 - B^E)\boxed{\epsilon(l, p_T)}$$
$$FR_S: \quad \epsilon^S(p_T) = B^S(p_T)\boxed{\epsilon(b, p_T)} + (1 - B^S)\boxed{\epsilon(l, p_T)}$$

b-enhanced and b-suppressed region defined using b-jet discriminator on a jet in FR, can measure mistag rate directly

Calculate b-jet fraction using DeepJet discriminator template

Mistag rate scale factor (SF)

Use flavor averaging* to get final background estimation:

$$N_{SR} \sim \sum_{evt \in CR} \frac{1}{2} \sum_{j \notin tagged} \epsilon\left(f_j, p_{T,j}\right)$$

$$N_{SR} \sim \sum_{evt \in CR} \frac{1}{2} \sum_{j \notin tagged} B^{CR} \epsilon\left(b, p_{T,j}\right) + (1 - B^{CR}) \epsilon(l, p_{T,j})$$

*Same method implemented in arXiv:1810.10069

Background uncertainties

Most accurate background estimation

$$Est^{JetHT}_{true}\left(\epsilon^{JetHT}_{true}(\vec{\theta}_{\infty})\right)$$

Cannot evaluate ϵ in infinitely fine jet kinematics bins

$$Est^{JetHT}_{true}\left(\epsilon^{JetHT}_{true}(p_T)\right)$$

Cannot evaluate ϵ in SR, potential signal contamination

$$Est^{JetHT}_{true}\left(\epsilon^{\gamma+jets}_{true}(p_T)\right)$$

Cannot determine flavor directly in data

$$Est^{JetHT}_{avg}\left(\epsilon^{\gamma+jets}_{inv}(p_T)\right)$$

What we CAN evaluate

Each change leads to uncertainty in final estimation

Signal uncertainties

Uncertainty source	Unflavored		Flavor-aligned	
	mean	std.	mean	std.
Integrated luminosity	1.8	0.6	1.8	0.6
Trigger efficiency	0.3	0.1	0.3	0.1
JES	1.3	0.9	0.7	0.4
JER	0.2	0.3	0.2	0.1
Pileup reweighting	0.9	0.8	1.0	0.9
Track modeling in sim.	0.3	0.8	0.5	0.6
PDF	<0.1	< 0.1	<0.1	<0.1
μ_F, μ_R	<0.1	<0.1	<0.1	<0.1

Evaluated per EMJ signal sample