ML4Jets 2023

Quark/gluon tagging in CMS Open Data with CWoLa and TopicFlow

Ayodele Ore

In collaboration with Matthew J. Dolan and John Gargalionis

Weakly-supervised Q/G tagging

- Fully-supervised learning not for Q/G ideal since:
  - ❗ Discrimination is sensitive to non-perturbative effects with large uncertainties in MC
  - ❗ Parton labels not well defined at detector level

- Instead, train on *mixed* samples — obtainable from exp.

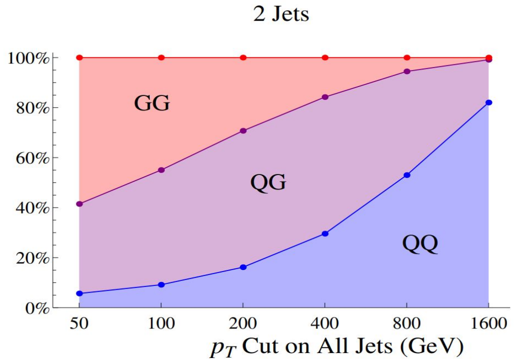
$$p_{M_1}(x) = f_1\, p_Q(x) + (1 - f_1) p_G(x)$$

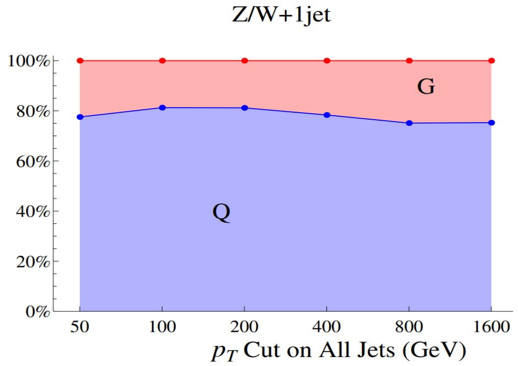
$$p_{M_2}(x) = f_2\, p_Q(x) + (1 - f_2) p_G(x)$$

- Same optimal classifier for M_1 vs M_2 as Q vs G (CWoLa)

[JHEP10(2017)174]

- CMS Open Data is a great testing ground





[JHEP10(2011)103]

Jet Topics: Disentangled distributions

[Phys.Rev.Lett.120,241602]

- If the mixture fractions are known, the pure distributions can be recovered:

$$p_{M_1}(x) = f_1\, p_Q(x) + (1 - f_1) p_G(x)$$
$$p_{M_2}(x) = f_2\, p_Q(x) + (1 - f_2) p_G(x)$$

Invert $\longrightarrow$

$$p_Q(x) = \frac{(1 - f_2) p_{M_1}(x) - (1 - f_1)\, p_{M_2}(x)}{f_1 - f_2}$$
$$p_G(x) = \frac{f_1 p_{M_2}(x) - f_2\, p_{M_1}(x)}{f_1 - f_2}$$

$\longrightarrow$ ROC curve from ε_Q ε_G

Jet Topics: Disentangled distributions

[Phys.Rev.Lett.120,241602]

- If the mixture fractions are known, the pure distributions can be recovered:

$$p_{M_1}(x) = f_1\, p_Q(x) + (1-f_1) p_G(x)$$
$$p_{M_2}(x) = f_2\, p_Q(x) + (1-f_2) p_G(x)$$

Invert

$$p_Q(x) = \frac{(1-f_2) p_{M_1}(x) - (1-f_1)\, p_{M_2}(x)}{f_1 - f_2}$$
$$p_G(x) = \frac{f_1 p_{M_2}(x) - f_2\, p_{M_1}(x)}{f_1 - f_2}$$

ROC curve from ε_Q ε_G

- Define "reducibility factors" : $\kappa_{ij} \equiv \min_x \frac{p_{M_i}(x)}{p_{M_j}(x)}$

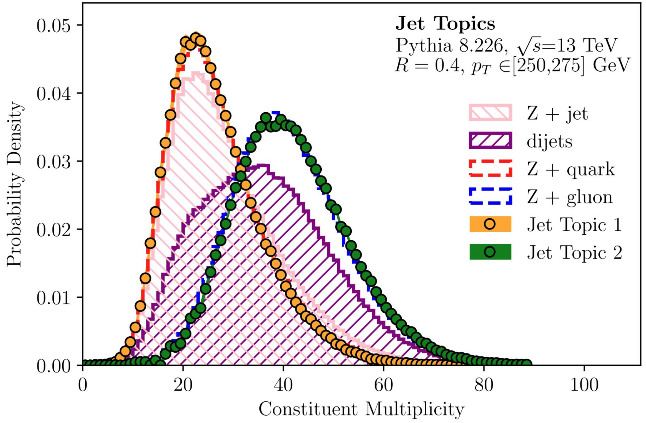
- If $\kappa_{QG} = \kappa_{GQ} = 0$ (mutual irreducibility) then:

$$f_1 = \frac{1-\kappa_{12}}{1-\kappa_{12}\kappa_{21}} \qquad f_2 = \kappa_{21} f_1$$

Measure directly from M_1, M_2!

… Otherwise you need to know κ_{QG} or κ_{GQ}

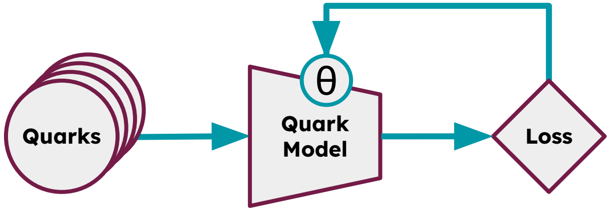
Estimate from Theory/MC



TopicFlow

*[**AO** & Matthew Dolan, PhysRevD.107.114003]*

- Generative models can be trained to learn topic distributions, given fractions.
  - ✅ Can apply in many dimensions
  - ✅ Can smooth statistics with oversampling
  - ✅ Can access quark/gluon likelihoods (with normalizing flow)

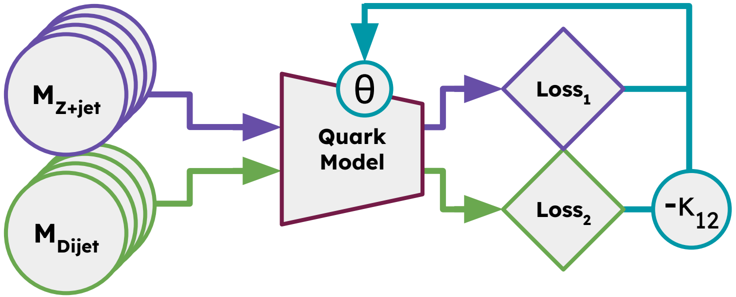


Typically requires pure sample!

$$\mathrm{Loss}_{\mathrm{Quark}} = \langle L_\theta(x) \rangle_{x\sim p_Q(x)}$$

$$= \langle L_\theta(x) \rangle_{x\sim p_{M_1}(x)} - \kappa_{12} \langle L_\theta(x) \rangle_{x\sim p_{M_2}(x)}$$

$\downarrow M_1$ $\uparrow M_2$



CMS Open Data

- 2011 data at @ 7 TeV

$50 < p_T^{j_1} < 80\text{GeV}$

$|m_{\mu\mu} - m_Z| < 20\text{GeV}$

$\Delta\phi(j_1, \mu\mu) > 2$

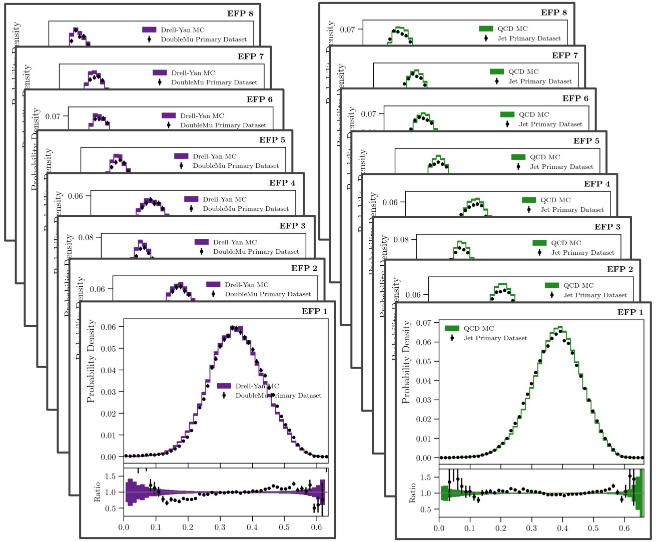
$p_T^{j_2} > 30\text{GeV}$

	Z+jet (mostly quark)	Dijet (mostly gluon)
Data	**Data[Zj]**	**Data[jj]**
MC	**MC[Zj]**	**MC[jj]**

MC[jj] → **Quarks**, **Gluons**

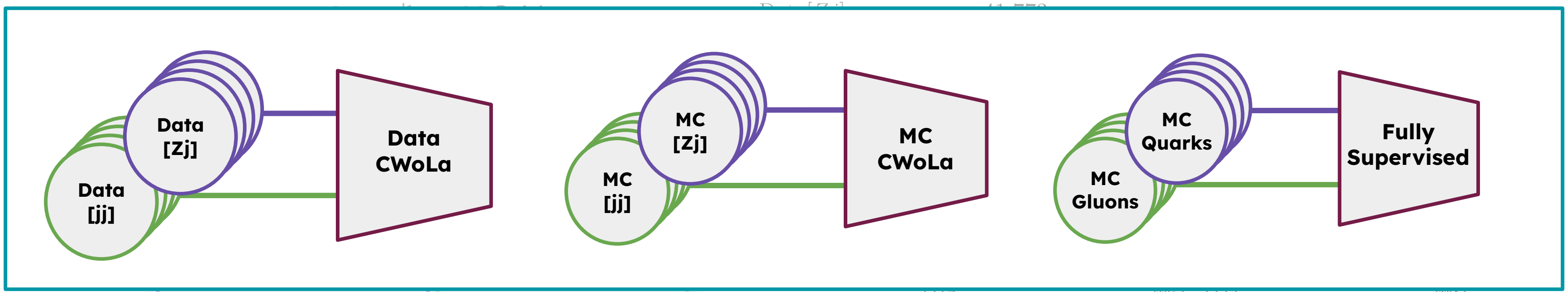
Including CMS detector simulation!

Dataset	Total events	Quarks	Gluons	
Data[Zj]	41,773	-	-	
Data[jj]	82,162	-	-	
MC[Zj]	95,324	70,568	24,756	~74% quarks
MC[jj]	3,064,713	868,556	2,196,157	~30% quarks



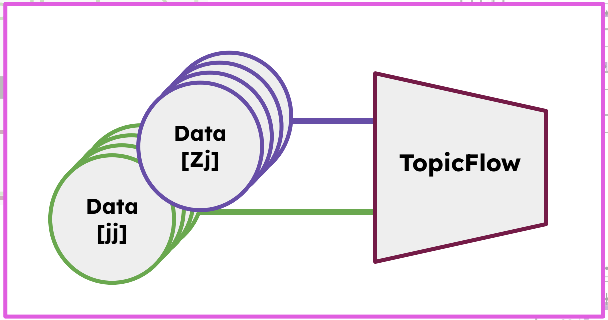
CMS Open Data

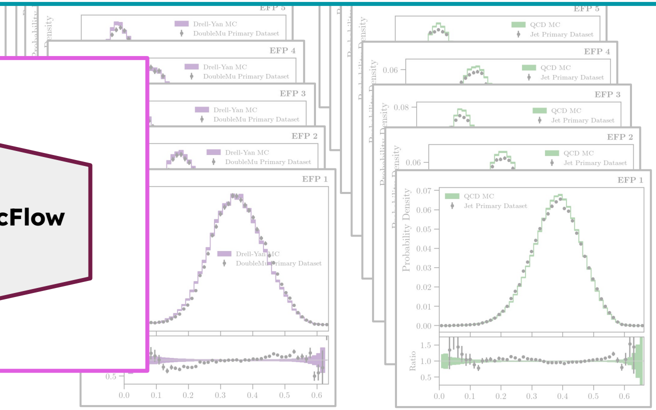
Dataset	Total events	Quarks	Gluons



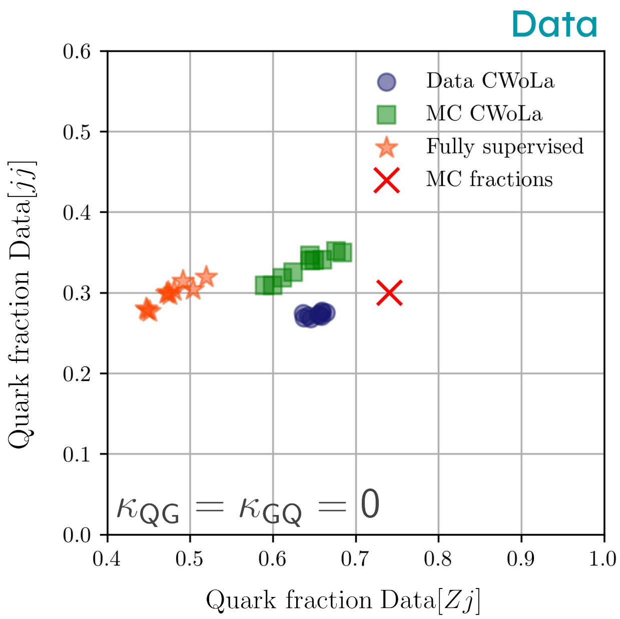
	Z+jet (mostly quark)	Dijet
Data	Data[Zj]	
MC	MC[Zj]	

Including CMS detector simulation!

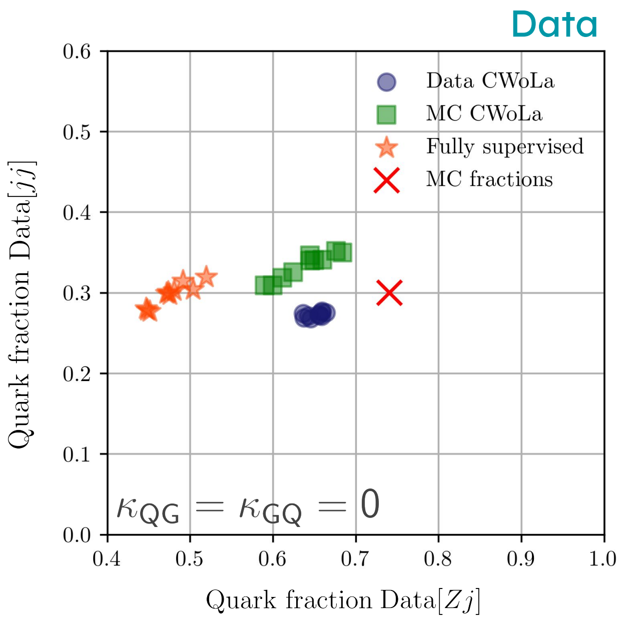


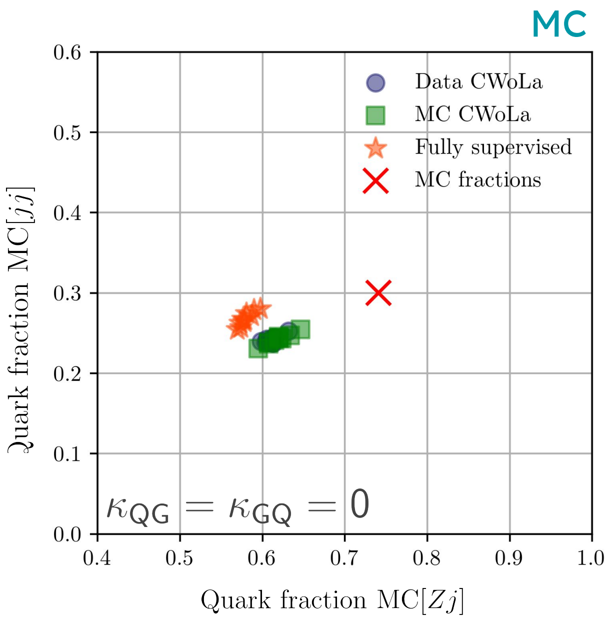


Quark fractions

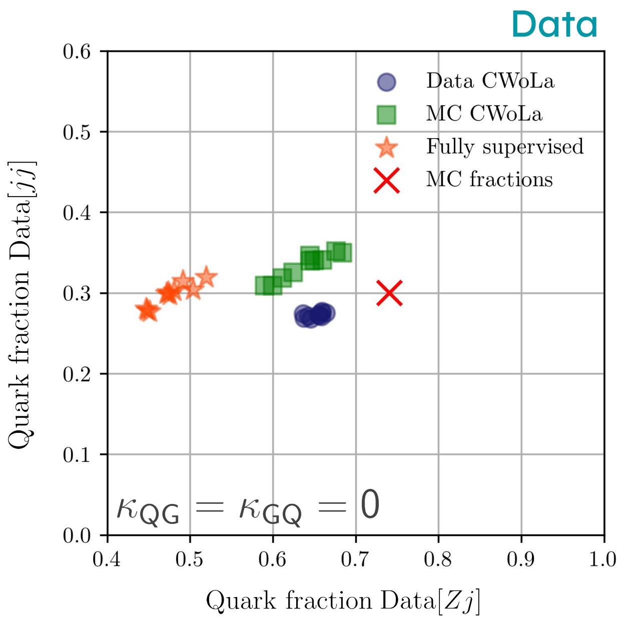


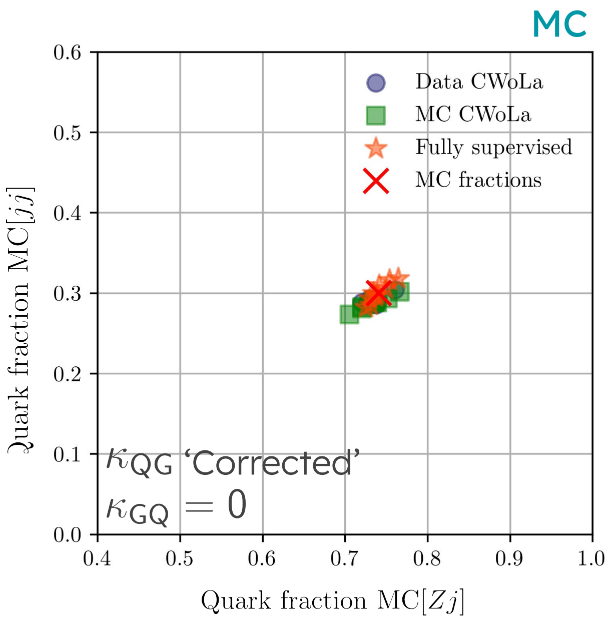
Quark fractions



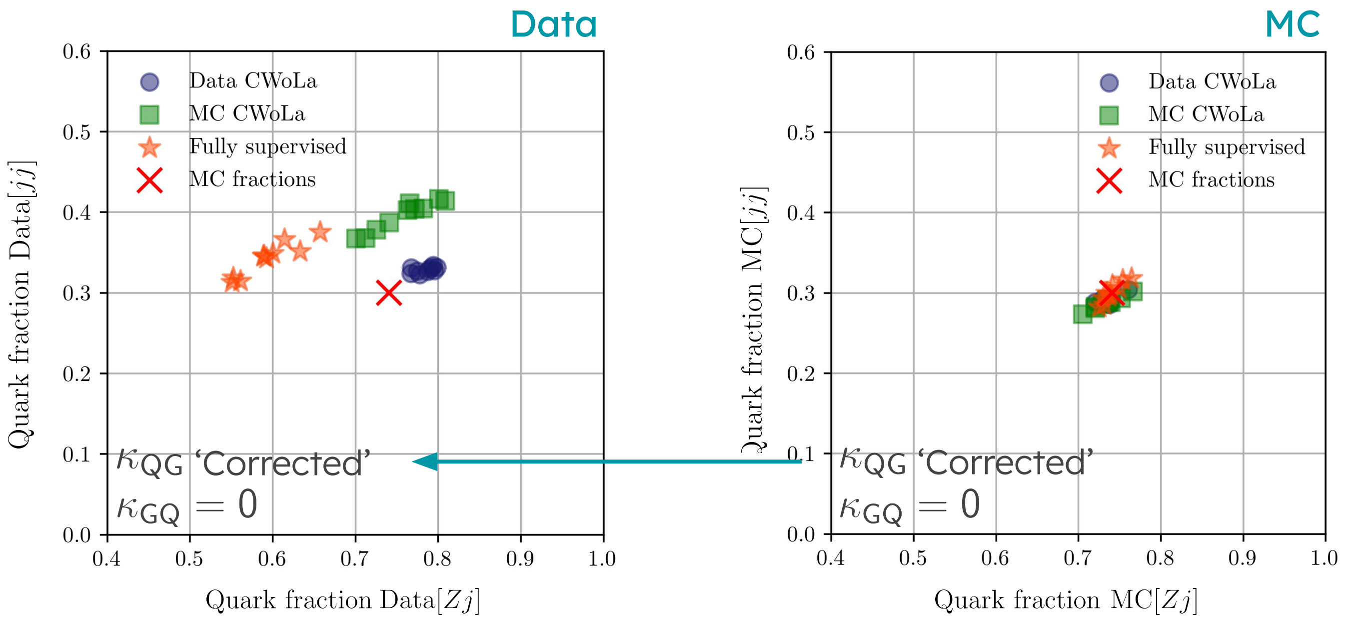


Quark fractions





Quark fractions



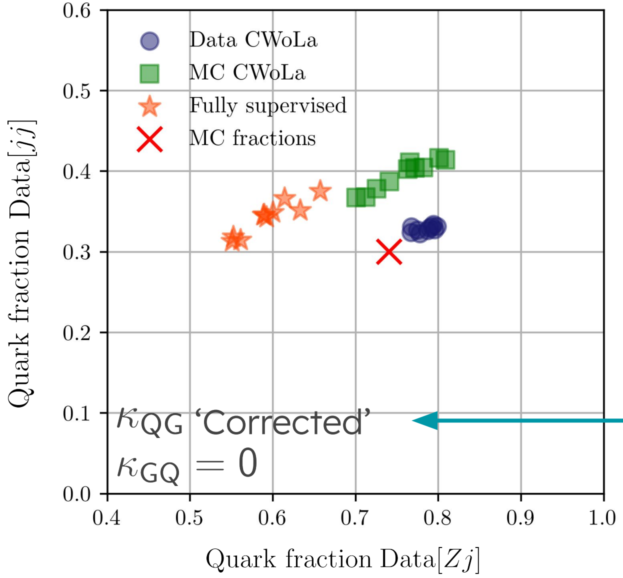
Quark fractions

From Data CWoLa:

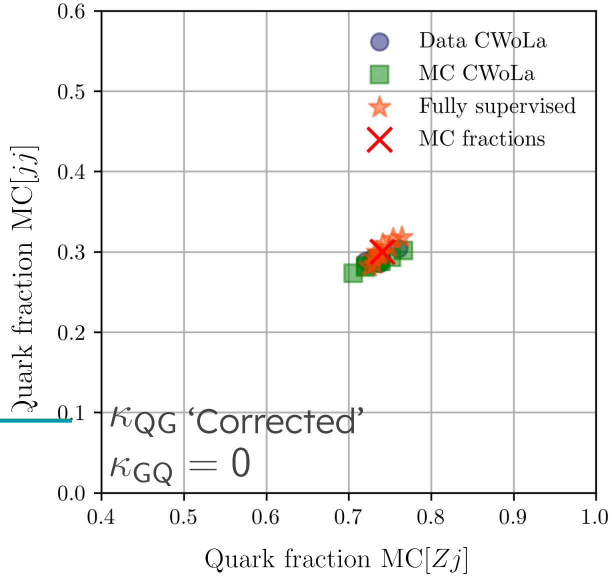
$\kappa_{QG} = 0$: f_1= 0.651 , f_2= 0.273

κ_{QG} 'Corrected' : f_1= 0.784 , f_2= 0.329

Data

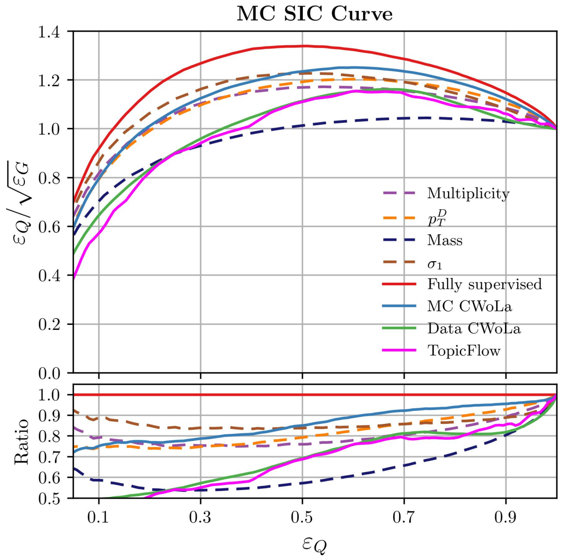


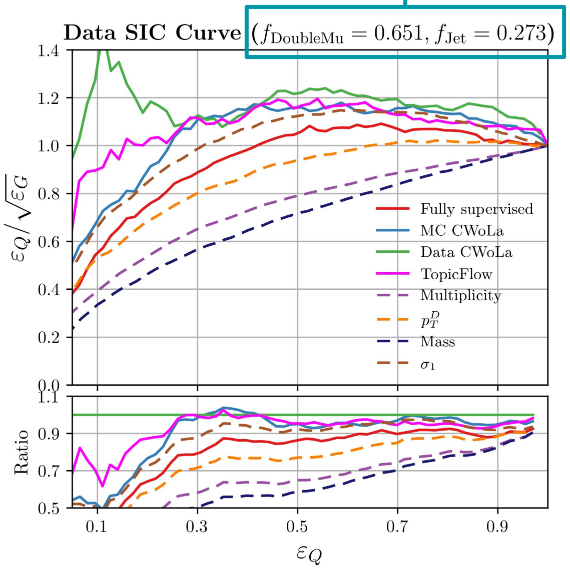
MC

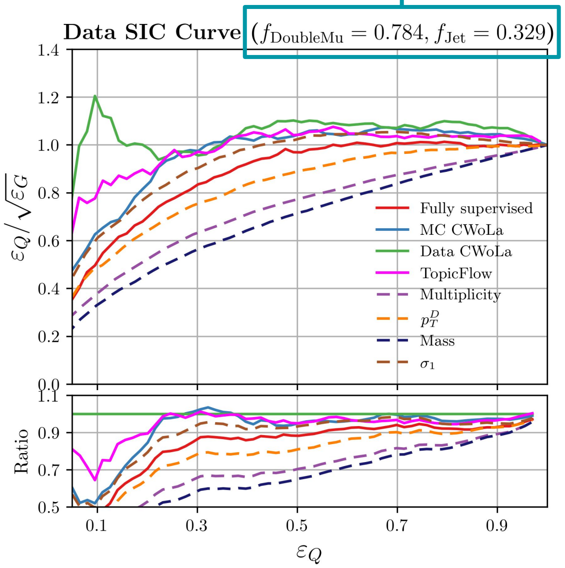


Tagging performance

- Recall estimate for efficiencies on data: $\varepsilon_G(t) = \dfrac{f_1 \varepsilon_{M_2}(t) - f_2\, \varepsilon_{M_1}(t)}{f_1 - f_2}$ $\quad \varepsilon_Q(t) = \dfrac{(1-f_2)\varepsilon_{M_1}(t) - (1-f_1)\, \varepsilon_{M_2}(t)}{f_1 - f_2}$





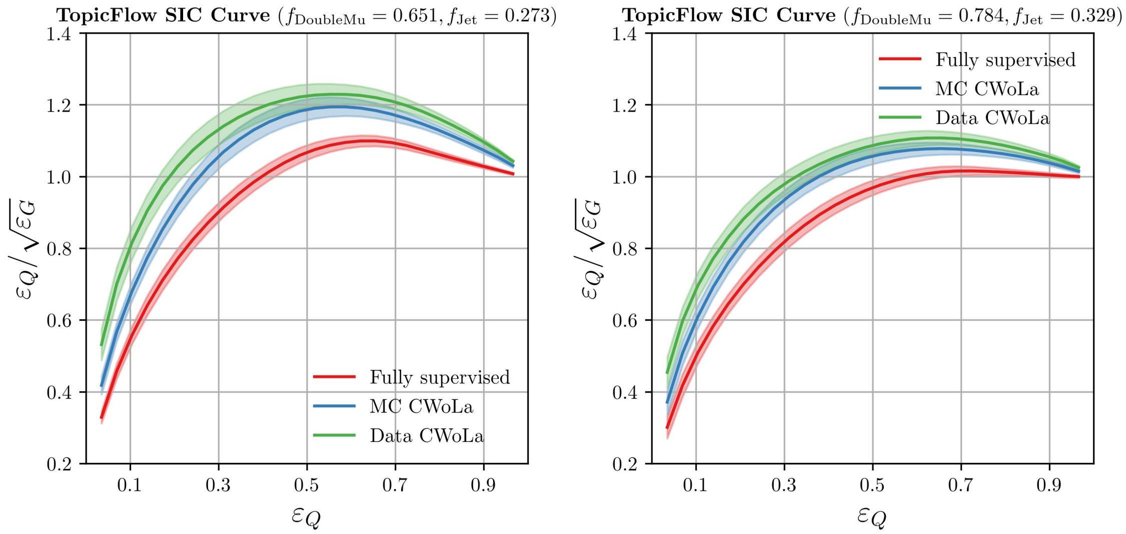


Smoothed ROCs with TopicFlow

$$\varepsilon_Q(t) = \langle \Theta(h(x) - t) \rangle_{x \sim p_Q}$$

$$\varepsilon_G(t) = \langle \Theta(h(x) - t) \rangle_{x \sim p_G}$$

- TopicFlow lets us avoid subtraction and evaluate ROC on samples:
- Error bars from ensemble: (Different distributions compatible with the test dataset)
- Bands smaller than impact of assumed fractions



Summary

- Fully-supervised learning causes train-test domain shift for jet tagging.
- Weakly-supervised learning facilitates training on data:

Classification

CWoLa outperforms MC-supervised models in CMS data.

Classifier rankings unaffected by estimated fraction.

Generative Modelling

TopicFlow can learn pure quark/gluon distributions.

Generative classification competitive with CWoLa.

Oversampling smooths ROC curves.

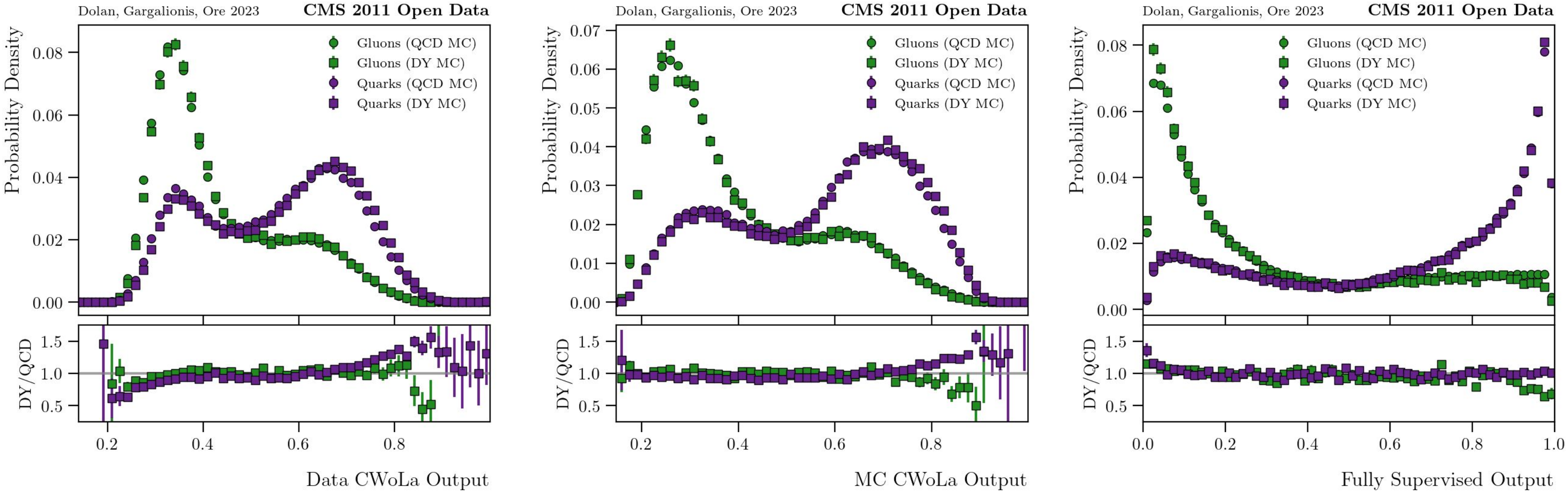
- Future directions:

★ Model/quantify sample dependence

★ Add f_1, f_2 to TopicFlow optimization

Backup Slides

Backup: Sample independence



Backup: All Jet Topics formulas

$$p_{M_1}(x) = f_1\, p_Q(x) + (1 - f_1) p_G(x)$$

$$p_{M_2}(x) = f_2\, p_Q(x) + (1 - f_2) p_G(x)$$

$$p_G(x) = \frac{f_1 p_{M_2}(x) - f_2\, p_{M_1}(x)}{f_1 - f_2}$$

$$p_Q(x) = \frac{(1 - f_2) p_{M_1}(x) - (1 - f_1)\, p_{M_2}(x)}{f_1 - f_2}$$

Define "reducibility factors" $\kappa_{ij} \equiv \min_x \frac{p_{M_i}(x)}{p_{M_j}(x)}$

Measure directly from M_1, M_2

If $\kappa_{QG} = \kappa_{GQ} = 0$ ("mutual irreducibility") then: $f_1 = \frac{1 - \kappa_{12}}{1 - \kappa_{12}\kappa_{21}}, \quad f_2 = \kappa_{21} f_1$

For non-zero κ_{QG} or κ_{GQ} :

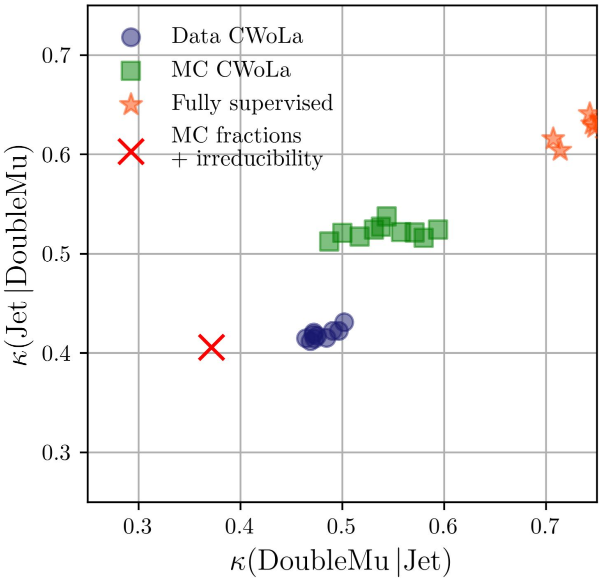
Calibrate with MC/Theory

$$f_1 = \frac{1}{1 - \kappa_{12}\kappa_{21}} \left(\frac{1 - \kappa_{12}}{1 - \kappa_{QG}} - \frac{\kappa_{12}\kappa_{GQ}(1 - \kappa_{21})}{1 - \kappa_{GQ}} \right)$$

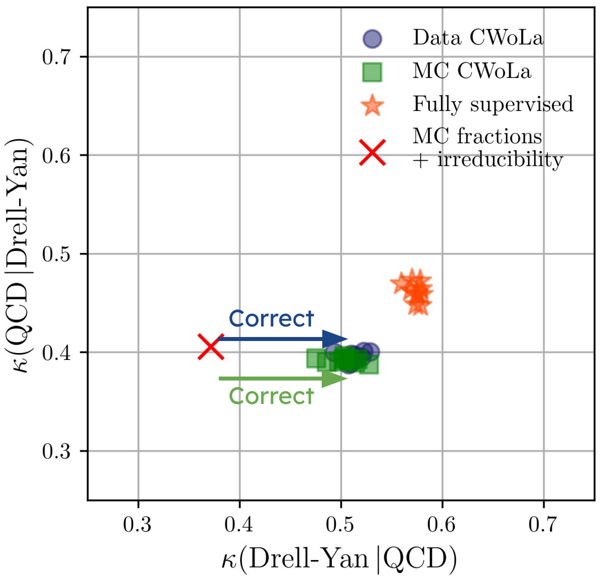
$$f_2 = \frac{1}{1 - \kappa_{12}\kappa_{21}} \left(\frac{\kappa_{21}(1 - \kappa_{12})}{1 - \kappa_{QG}} - \frac{\kappa_{GQ}(1 - \kappa_{21})}{1 - \kappa_{GQ}} \right)$$

Backup: Reducibility correction

Data



MC



$$\kappa_{\mathsf{QG}} = 1 - \frac{1 - \kappa_{12}}{\mathsf{f}_1 - \kappa_{12}\mathsf{f}_2}$$

Backup: TopicFlow (v2)

$$p_Q(x) = \frac{p_{M_1}(x) - \kappa_{12}\, p_{M_2}(x)}{1 - \kappa_{12}}$$

$$p_G(x) = \frac{p_{M_2}(x) - \kappa_{21}\, p_{M_1}(x)}{1 - \kappa_{21}}$$

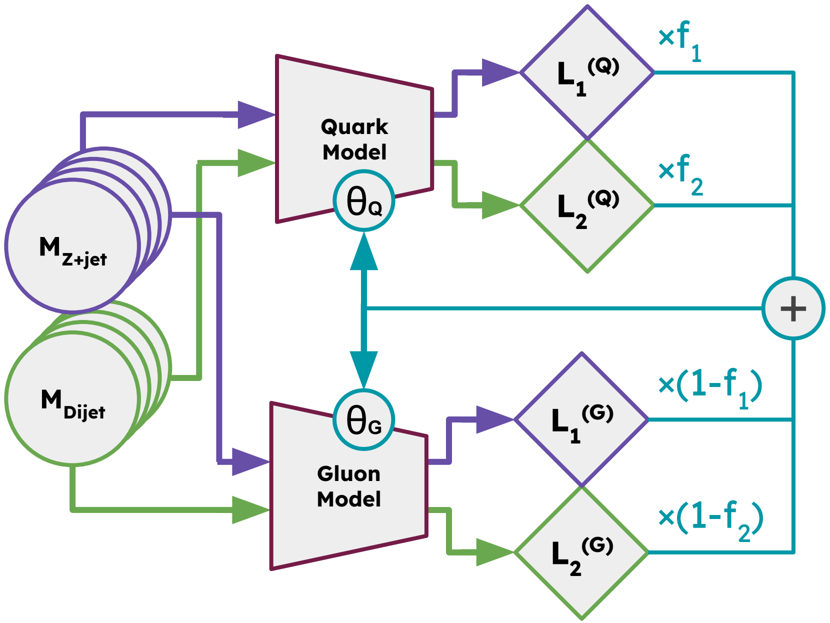
v1

v2

$$p_{M_1}(x) = f_1\, p_Q(x) + (1 - f_1) p_G(x)$$

$$p_{M_2}(x) = f_2\, p_Q(x) + (1 - f_2) p_G(x)$$

- Joint training of the quark and gluon models gives convex loss:

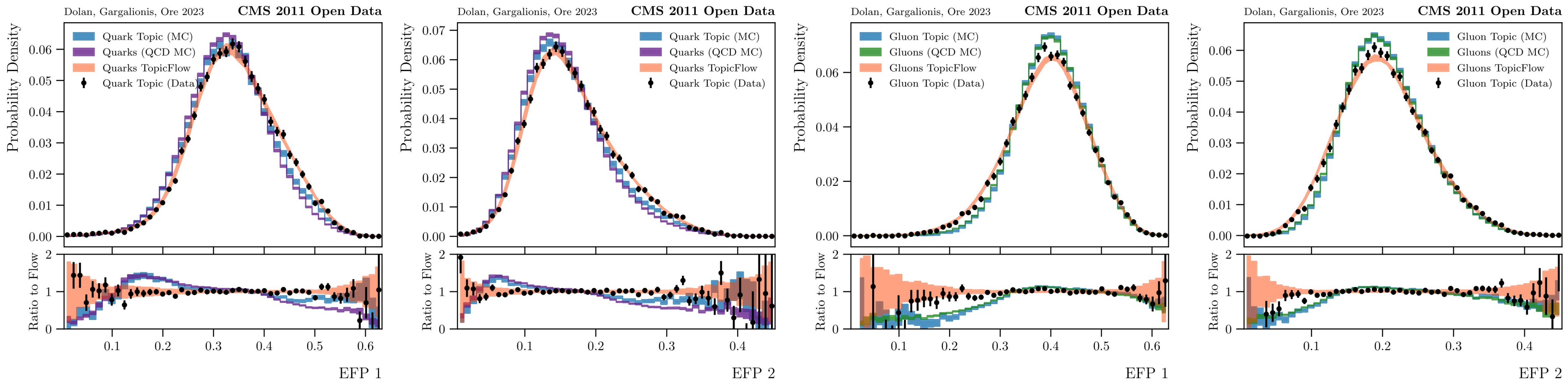


$$\text{Loss} = \left\langle L_{\theta_Q,\theta_G}(x) \right\rangle_{x\sim p_{Z+jet}(x)} + \left\langle L_{\theta_Q,\theta_G}(x) \right\rangle_{x\sim p_{dijet}(x)}$$

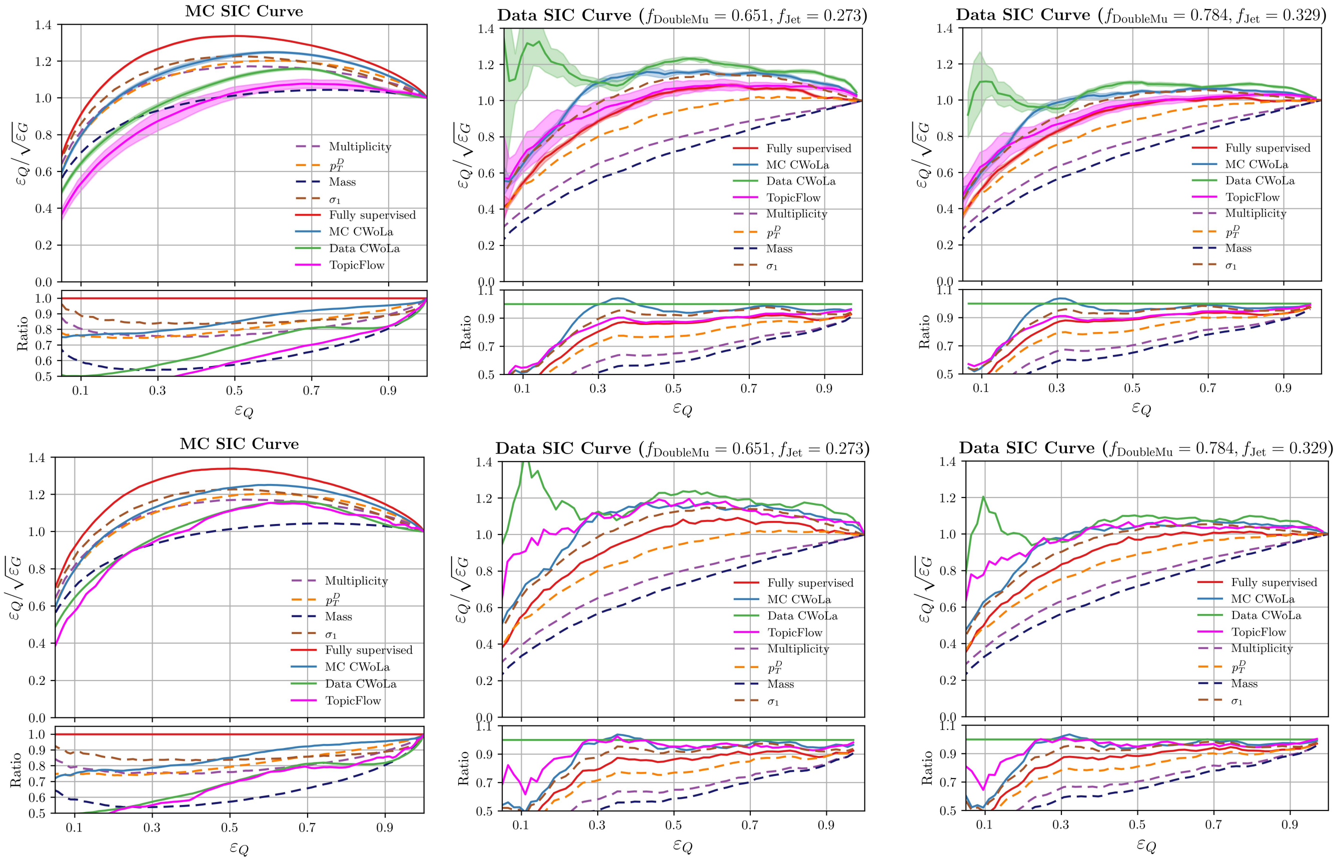
$$= f_1 \left\langle L_{\theta_Q}(x) \right\rangle_{x\sim p_{Z+jet}(x)} + (1 - f_1) \left\langle L_{\theta_G}(x) \right\rangle_{x\sim p_{Z+jet}(x)}$$

$$+ f_2 \left\langle L_{\theta_Q}(x) \right\rangle_{x\sim p_{dijet}(x)} + (1 - f_2) \left\langle L_{\theta_G}(x) \right\rangle_{x\sim p_{dijet}(x)}$$

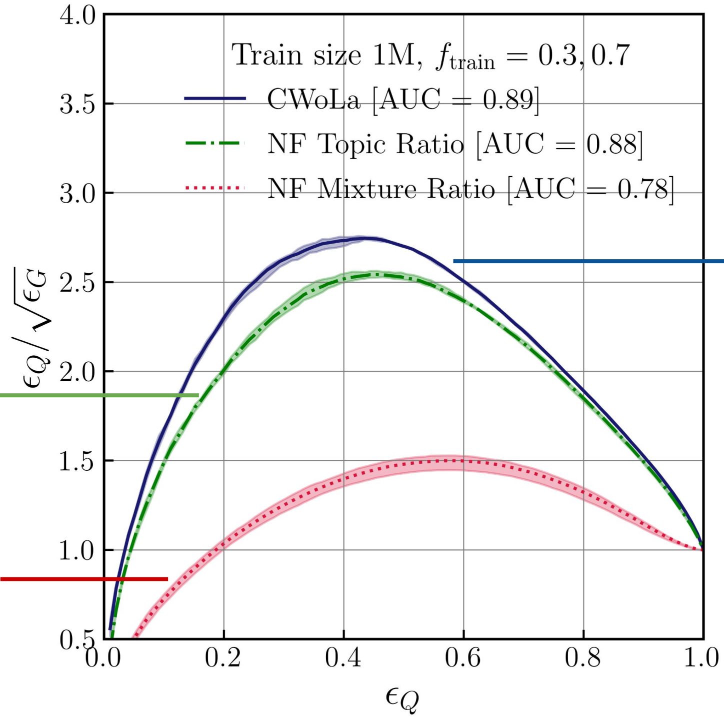
Backup: TopicFlow samples



Backup: Ensemble tagging



Backup: Generative Classification



Q vs. G Generative (TopicFlow)

M_1 vs. M_2 Generative

M_1 vs. M_2 Discriminative (CWoLa)