Measurement of quark and gluon jet fractions at the CMS: methods, results and outlook for Run-3

S. Shulha

JINR (RU)
F. Skorina GSU (BY)

Conference on High Energy Physics
11-14 September, 2023
Erevan, Armenia

- This work is part of the CMS analyses, which deals with recognition and tagging of q- and g-jets
- Recognition of q/g-jets is based on the discriminator - each jet is assigned a discriminator value V
- Examples of V are simple Macro Parameters (MP's): particle multiplicity inside jet, jet radius in (η, φ)-space, or combinations of simple MP's (like QGL – "Quark-Gluon Likelihood",...)
- Discriminator is "trained" on MC jets: "training" means obtaining a MC normalized distributions over V for q/g-jets → $H^g_{\mathrm{MC}}(V)$ and $H^q_{\mathrm{MC}}(V)$ –

  $H^g_{\mathrm{MC}}(V)$ and $H^q_{\mathrm{MC}}(V)$ are also called "q/g-templates"
- "q/g-templates" are key objects in q/g-tagging: "q/g-templates" allow one to say whether a given jet is a q- or g-jet with a given probability
- True "q/g-templates" $H^f_{\mathrm{DAT}}(V)$ in data differ from model ones: $H^f_{\mathrm{DAT}}(V) \neq H^f_{\mathrm{MC}}(V)$
- Calculation of $H^f_{\mathrm{DAT}}(V)$ using data is referred to as obtaining "data-driven Scale Factor" (SF) for MC q/g-templates: $S^f(V) \equiv H^f_{\mathrm{DAT}}/H^f_{\mathrm{MC}}$. SF is a key issue in q/g-tagging task

- To obtain **TWO** corrected q/g-templates H^q_{DAT} and H^g_{DAT} (or SF's) we need **TWO Eqs** → need **TWO** jet samples with known g-fractions
- To date (Sept 2023), the official CMS recommendation for RUN-1 and RUN-2 is to use MC fractions for two channels (dijets and Z+jets) - $\alpha^g_{1\,\mathrm{MC}}$ and $\alpha^g_{2\,\mathrm{MC}}$:

$$H_{1,\mathrm{DAT}} = \alpha^g_{1,\mathrm{MC}} \cdot H^g_{\mathrm{DAT}} + (1 - \alpha^g_{1,\mathrm{MC}}) \cdot H^q_{\mathrm{DAT}} \quad (1)$$

$$H_{2,\mathrm{DAT}} = \alpha^g_{2,\mathrm{MC}} \cdot H^g_{\mathrm{DAT}} + (1 - \alpha^g_{2,\mathrm{MC}}) \cdot H^q_{\mathrm{DAT}}$$

- Solution of this system of Eqs. gives us data-driven corrected q/g-templates:

$$H^q_{\mathrm{DAT}} = \frac{\alpha^g_{2,\mathrm{MC}} H_{1,\mathrm{DAT}} - \alpha^g_{1,\mathrm{MC}} H_{2,\mathrm{DAT}}}{\alpha^g_{2,\mathrm{MC}} - \alpha^g_{1,\mathrm{MC}}} \quad (2)$$

$$H^g_{\mathrm{DAT}} = (g \to q, 1 \leftrightarrow 2)$$

- We showed the first measurements of g-fractions in 2018.
- Recommendation for us was to apply SF in measurement of g-fraction
- But, in current official form, Eqs.(2) were written w/o normalization and with hidden MC g-fractions. It is not difficult to guess that measured g-fraction with corrected q/g-templates in the data will give **exactly** the MC g-fractions!

Tip for the careful listener: measured $\alpha^g_{1,\mathrm{DAT}}$ is a solution of Eq.:

$$H_{1,\mathrm{DAT}} = \alpha^g_{1,\mathrm{DAT}} \cdot H^g_{\mathrm{DAT}} + (1 - \alpha^g_{1,\mathrm{DAT}}) \cdot H^q_{\mathrm{DAT}} \quad (1')$$

$$\alpha^g_{1,\mathrm{DAT}} = \alpha^g_{1,\mathrm{MC}}$$

- We proposed (2020) to use in CMS the modified SF for q/g-templates:

$$H^q_{\mathrm{DAT}} = \frac{\alpha^g_{2,\mathrm{DAT}} H_{1,\mathrm{DAT}} - \alpha^g_{1,\mathrm{DAT}} H_{2,\mathrm{DAT}}}{\alpha^g_{2,\mathrm{DAT}} - \alpha^g_{1,\mathrm{DAT}}}$$

$$H^g_{\mathrm{DAT}} = (q \leftrightarrow g, 1 \leftrightarrow 2) \quad (3)$$

- Before obtaining SF and $H^{q/g}_{\mathrm{DAT}}(V)$ we need to measure g-jet fractions. So, measurement of g-jet fraction becomes a key task for q/g-tagging**!**

We have found another important correction to SF Eqs.(3):

- Eqs.(3) give universal q/g-templates for any channel and any jet kinematics and environment. But, MC q/g-templates depend on kinematics! We proposed (2021, PEPAN Lett) method to introduce in Eqs.(3) corrections for kinematical non-universality

Very important remark:

- **Proposition**: g-fractions in data **with corrected** q/g-templates Eqs.(3) $\alpha^{f'}_{1,\mathrm{DAT}}$ are the same: $\alpha^{g'}_{1,\mathrm{DAT}} \equiv \alpha^g_{1,\mathrm{DAT}}$
- So, 1st measurement of g-fractions with MC q/g-templates cannot be improved by SF – iteration process is impossible!

Tip for the careful listener: to prove this, we start two equations:
1st iteration $\alpha^g_{1,\mathrm{DAT}}$ is a solution of Eq.: $H_{1,\mathrm{DAT}} = \alpha^g_{1,\mathrm{DAT}} \cdot H^g_{\mathrm{MC}} + (1 - \alpha^g_{1,\mathrm{DAT}}) \cdot H^q_{\mathrm{MC}}$
2nd iteration $\alpha^{g'}_{1,\mathrm{DAT}}$ is a solution of Eq.: $H_{1,\mathrm{DAT}} = \alpha^{g'}_{1,\mathrm{DAT}} \cdot H^g_{\mathrm{DAT}} + (1 - \alpha^{g'}_{1,\mathrm{DAT}}) \cdot H^q_{\mathrm{DAT}}$

Tip for the careful listener (cont.):

Proof:

$$H_{1,\mathrm{DAT}} = \alpha^{g\prime}_{1,\mathrm{DAT}} \cdot H^{g}_{\mathrm{DAT}} + (1 - \alpha^{g\prime}_{1,\mathrm{DAT}}) \cdot H^{q}_{\mathrm{DAT}}$$

$$H^{q}_{\mathrm{DAT}} = \frac{\alpha^{g}_{2,\mathrm{DAT}} H_{1,\mathrm{DAT}} - \alpha^{g}_{1,\mathrm{DAT}} H_{2,\mathrm{DAT}}}{\alpha^{g}_{2,\mathrm{DAT}} - \alpha^{g}_{1,\mathrm{DAT}}}$$

$$H^{g}_{\mathrm{DAT}} = \frac{(1 - \alpha^{g}_{1,\mathrm{DAT}}) H_{2,\mathrm{DAT}} - (1 - \alpha^{g}_{2,\mathrm{DAT}}) H_{1,\mathrm{DAT}}}{\alpha^{g}_{2,\mathrm{DAT}} - \alpha^{g}_{1,\mathrm{DAT}}}$$

$$\alpha^{g\prime}_{1,\mathrm{DAT}} = \frac{H_{1,\mathrm{DAT}} - H^{q}_{\mathrm{DAT}}}{H^{g}_{\mathrm{DAT}} - H^{q}_{\mathrm{DAT}}}$$

$$H_{1,\mathrm{DAT}} - H^{q}_{\mathrm{DAT}} = \frac{\alpha^{g}_{1,\mathrm{DAT}} (H_{2,\mathrm{DAT}} - H_{1,\mathrm{DAT}})}{\alpha^{g}_{2,\mathrm{DAT}} - \alpha^{g}_{1,\mathrm{DAT}}}$$

$$H^{g}_{\mathrm{DAT}} - H^{q}_{\mathrm{DAT}} = \frac{H_{2,\mathrm{DAT}} - H_{1,\mathrm{DAT}}}{\alpha^{g}_{2,\mathrm{DAT}} - \alpha^{g}_{1,\mathrm{DAT}}}$$

$$\alpha^{g}_{1,\mathrm{DAT}} = \frac{H_{1,\mathrm{DAT}} - H^{q}_{\mathrm{MC}}}{H^{g}_{\mathrm{MC}} - H^{q}_{\mathrm{MC}}} \qquad (4)$$

Proposition : $\alpha^{g\prime}_{1,\mathrm{DAT}} \equiv \alpha^{g}_{1,\mathrm{DAT}}$ $\otimes$

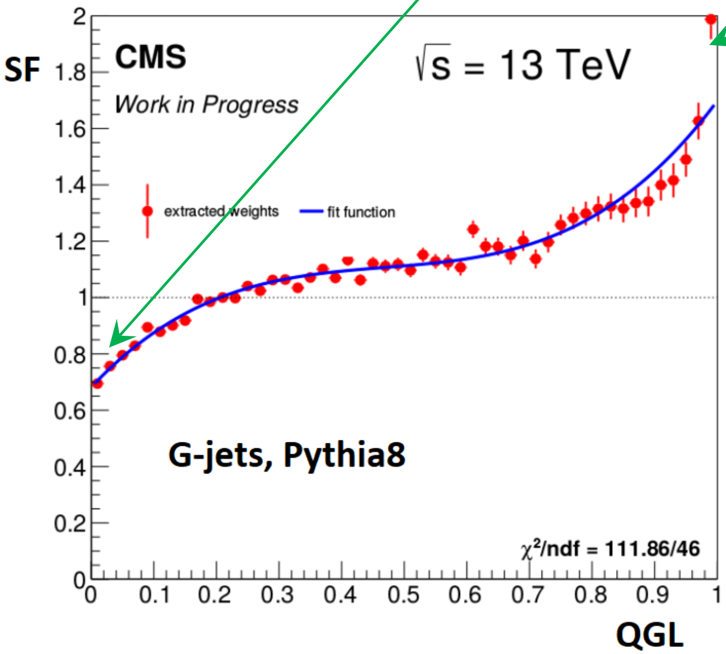
- 2nd iteration for g-fraction measurement is impossible!
- Model determines g-fraction in data unambiguously and does not allow it to be corrected within current model
- However, there is a way to define quantitatively discrepancy **between model and data** in measured g-fractions within one model. It is Model Uncertainty (M.U.).
- M.U. is low edge of Theoretical Uncertainty in g-fraction measurements

Remarks

- If $\alpha^g_{\mathrm{DAT}} \approx \alpha^g_{\mathrm{MC}}$ then official SF Eq.(2) ≈ new SF Eq.(3)
- Spoiler: we found strong g-jet suppression in region $P_T^{jet} < 200$ GeV:

$$\alpha^g_{\mathrm{DAT}} \approx (0.5 \div 0.7) \cdot \alpha^g_{\mathrm{MC}} \quad \Rightarrow \quad \text{official SF} \gg \text{new SF}$$

- Thus, official CMS SF's Eq.(2) developed for Run-1 and Run-2 are **wrong**: they significantly change true g-factions $\alpha^g_{\mathrm{DAT}} \to \alpha^g_{\mathrm{MC}}$ and MC q/g-templates are changed significantly also: -35% at small QGL≈ 0 and up to +100% QGL≈ 1



- If we use new SF Eq.(3) with measured g-fractions then q/g-templates are changed to a maximum of 4% w/o changing the used g-fractions

- It should be taken into account in CMS Run-3 q/g-tagging: measuring g-fractions should be the first task to obtain correct q/g-tagging

Now we are moving to g-fraction measurements…

- Careful listener may suggest already a method for measuring – the main formula has been written on page 5:

$$\alpha^g_{\mathrm{DAT}} = \frac{H_{\mathrm{DAT}}(V) - H^q_{\mathrm{MC}}(V)}{H^g_{\mathrm{MC}}(V) - H^q_{\mathrm{MC}}(V)} \qquad (4)$$

where $H_{\mathrm{DAT}}(V)$ – measured distribution, $H^f_{\mathrm{MC}}(V)$ - MC q/g-templates

- But right part depends on V-bin?

- Well! Each V-bin can be considered as independent experiment and we'll define measured α^g_{DAT} as averaged value…

Method of "bin averaging"

S.S. PEPAN Lett. **2023/2024** (in preparation)

- For any MP (jet macro parameter) $V \equiv V_{1,2,3,4,\dots}$:

$$H(V) = \alpha^g H^g(V) + (1-\alpha^g)\, H^q(V) \qquad (5)$$

- In case of **MC**, Eq.(5) has the same solution α^g for all V-bins:

$$\alpha^g = \frac{H^{\mathbf{MC}}(V) - H^q(V)}{H^g(V) - H^q(V)} = const(V)$$

- In case of **DATA**, solution of Eq. (5) is not a V-constant:

$$\alpha^g_V = \frac{H^{\mathbf{DAT}}(V) - H^q(V)}{H^g(V) - H^q(V)} \qquad (6)$$

Each bin is a separate independent experiment to measure α^g

- Definition: measured g-fraction is averaged value:

N_V - number of V-bins

$$\alpha^g \equiv \langle \alpha^g_V \rangle = \frac{\sum_{V=1}^{N_V} \alpha^g_V}{N_V} \qquad (7)$$

with uncertainty $\Delta\alpha^g \equiv \frac{\sqrt{\langle {\alpha^g_V}^2 \rangle - \langle \alpha^g_V \rangle^2}}{\sqrt{N_V}}$

- In 2023 we implemented this method and showed results in CMS (June 2023, SMP-HAD)
- Deprecated method: So far, we have used a more complex method with $V =$ QGL and with fit by WLS or LS methods (ROOT/MINUIT): $H_{\mathrm{DAT}} \sim \alpha^g_{\mathrm{DAT}} \cdot H^g_{\mathrm{MC}} + (1-\alpha^g_{1,\mathrm{DAT}}) \cdot H^q_{\mathrm{MC}}$

- At previous page we used one MP and obtained V–bin averaged g-fraction

- We can use "full set of independent MP's"[1] $V_{1,2,3,4,\ldots}$ to obtain several averaged g-fractions $\alpha_1^g, \alpha_2^g, \alpha_3^g, \ldots$

- In case of MC, calculation with any MP $V_{1,2,3,\ldots}$ gives the same $\alpha_1^g = \alpha_2^g = \alpha_3^g = \ldots = \alpha^g$ because q/g-templates are true for MC

$$\alpha^{\boldsymbol{g}} = \frac{H(V_k) - H^{\boldsymbol{q}}(V_k)}{H^{\boldsymbol{g}}(V_k) - H^{\boldsymbol{q}}(V_k)} = \boldsymbol{const}(\boldsymbol{k})$$

- In case of DATA $\alpha_1^g \neq \alpha_2^g \neq \alpha_3^g \neq \ldots$ because MC q/g-templates are not true for DATA

- Maximum of differences $|\alpha_1^g - \alpha_2^g|, |\alpha_1^g - \alpha_3^g|, |\alpha_2^g - \alpha_3^g|, \ldots$ describes the deviation of MC q/g-templates from true ones = Model Uncertainty (**M.U.**)

$$\textbf{M.U.} = \frac{1}{2} \cdot \max\{|\alpha_1^g - \alpha_2^g|, |\alpha_1^g - \alpha_3^g|, |\alpha_2^g - \alpha_3^g|, \ldots\}$$

---

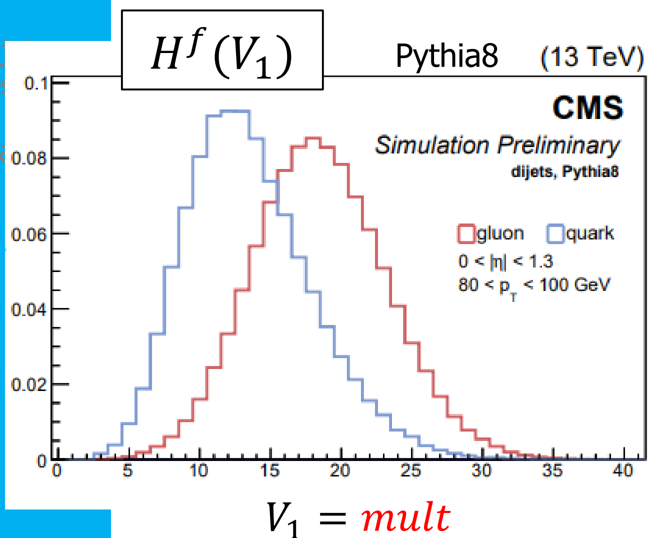
[1] How to define "full set of independent MP's" – it is interesting question. Whoever answers this question will make an important contribution to the "theory of quantum measurements"

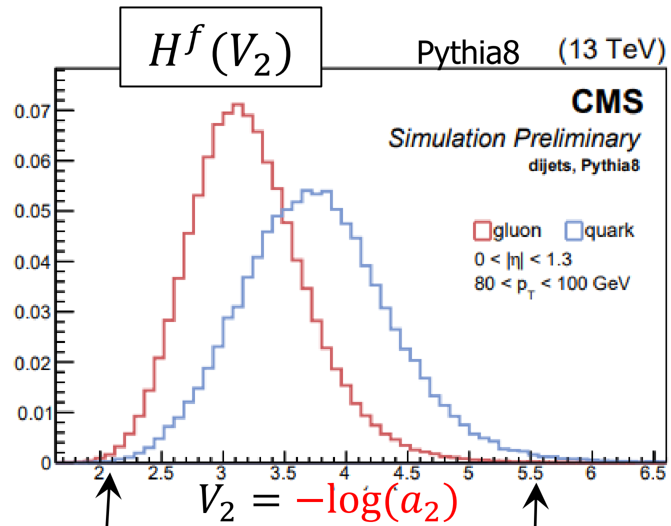
- Choose MP's which are the most sensitive to Jet Flavour[1]
  - Total multiplicity inside jet ($mult$)
  - Minor axis of jet ellipse in (η, φ)-space a_2
  - "Fragmentation function" $p_T D = \frac{\sqrt{\sum_i p_{T\,i}^2}}{\sum_i p_{T\,i}} \in [0,1]$

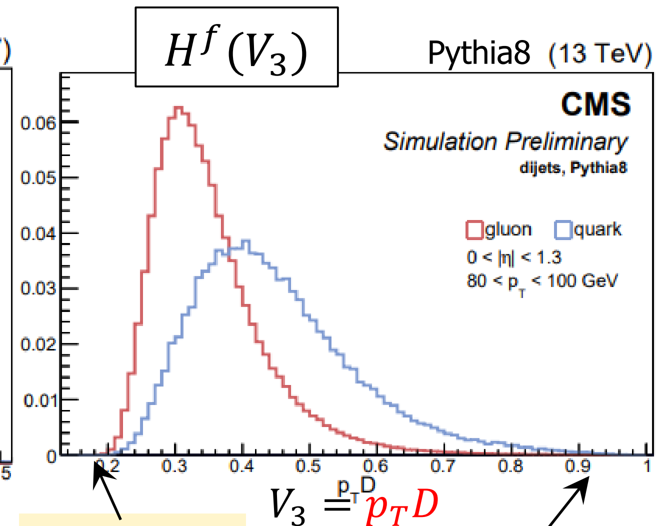
[1] CMS PAS JME-13-002
CMS PAS JME-16-003

$V_{1,2,3} = (mult,\ a_2,\ p_T D) \equiv \vec{V}$

Fig. 1: q/g-templates $H^f(V_1),\ H^f(V_2),\ H^f(V_3)$







$\frac{n^g}{n^q} \approx 1.5$

Wider jets

Narrower jets

P_T^{jet} is uniformly distributed among constituents

P_T^{jet} is concentrated in a limited number of constituents

- These three jet MP's are used to measure g-fractions

CMS PAS JME-13-002
CMS PAS JME-16-003

- QGL is a jet MP that is a combination of simple MP's :

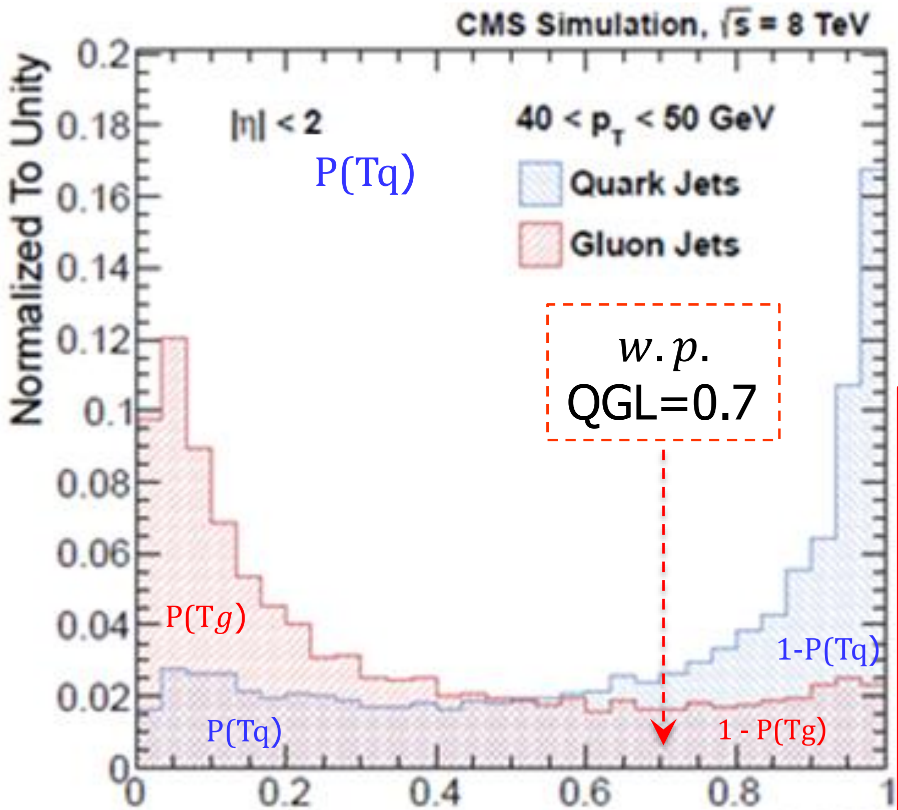
$$V_4 \equiv QGL = \frac{Q(\vec{V})}{Q(\vec{V}) + G(\vec{V})}$$

QGL – discriminator "Quark-Gluon Likelihood"

$$Q(\vec{V}) = \prod_{i=1}^{3} H^q(V_i), \qquad G(\vec{V}) = \prod_{i=1}^{3} H^g(V_i)$$

$V_{1,2,3} = (mult,\ a_2,\ p_T D) \equiv \vec{V}$

- Sensitivity of QGL to jet flavour is much stronger than that of original $mult, a_2, p_T D$.



$V_4 \equiv$ QGL$(\vec{V})$

Fig.2: QGL-templates

- QGL-templates are used to tag q/g-jets. It is very important tool to select channels
- We measured g-fractions with QGL-templates also to check QGL written in datasets

- We show (June 2023, SMP-HAD) that QGL written in all CMS Run-2 datasets are wrong
- It is necessary to inform everyone who uses q/g-tagging in Run-2 analyses
- We prepared new QGL for CMS Run-2 and test them using g-fraction measurements

- α^g was found by V = $mult, a_2, p_T D$ and "new QGL"

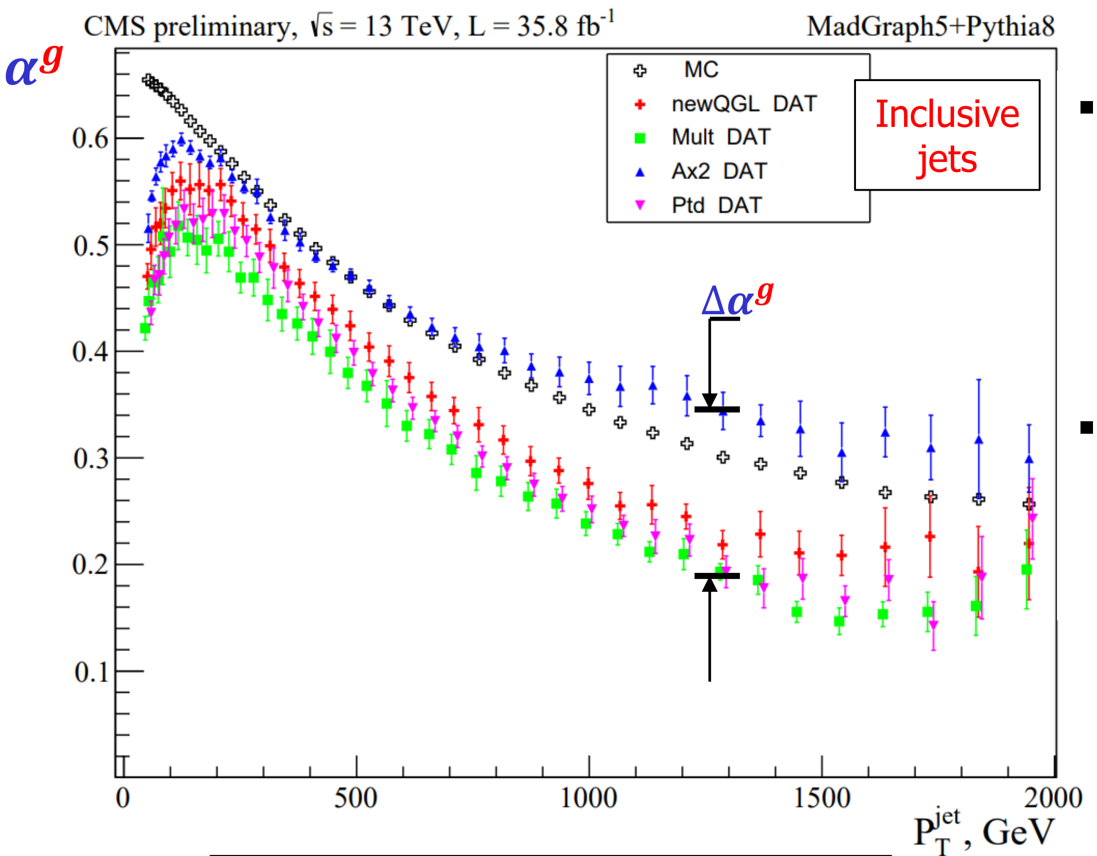


Fig. 3: Demonstration of M.U. (S.S., SMP-HAD, June 2023)

- This preliminary results were obtained in CMS group "Gluon-jet/Quark-jet analyses":

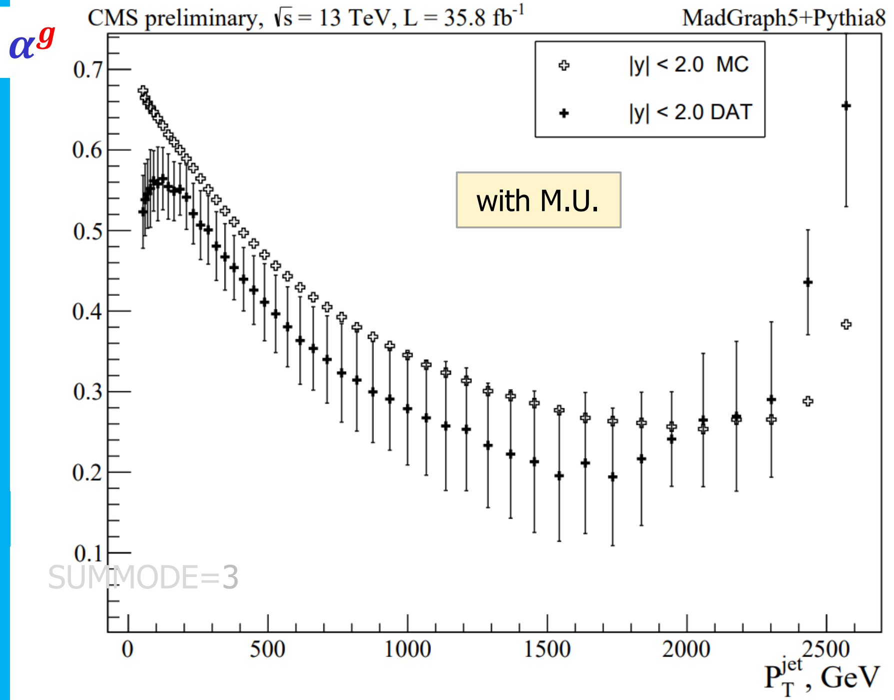
  S.S., D.Budkouski(JINR), J.Strologas (GR), O.Atakisi(TR)

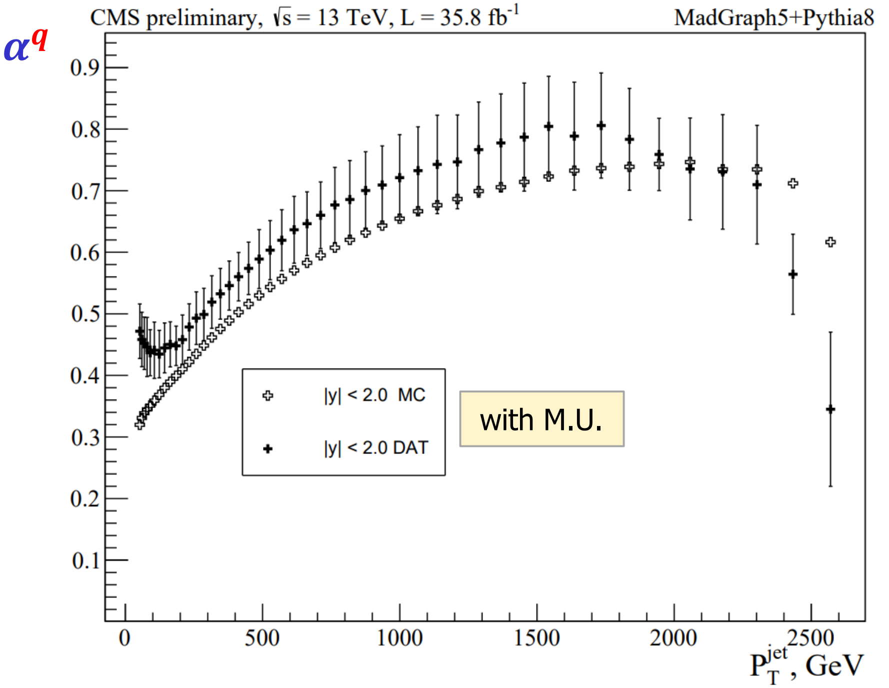
- This group was created within CMS SMP-HAD group in April 2021 purposefully to measure g-fractions in inclusive jet channel with Run-II data

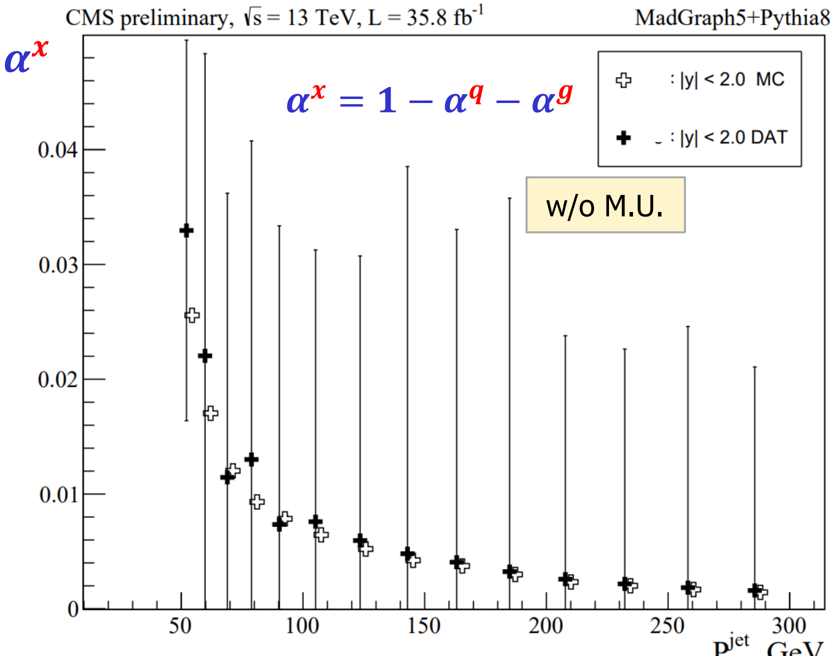
- Measurement of g-fraction demonstrates indirectly large deviation of true unknown DATA q/g-templates from Pythia8 ones

q/g/x-jet fractions

Inclusive jets







- This preliminary results were obtained in CMS group "Gluon-jet/Quark-jet analyses":

  S.S., D.Budkouski(JINR), J.Strologas (GR), O.Atakisi(TR)

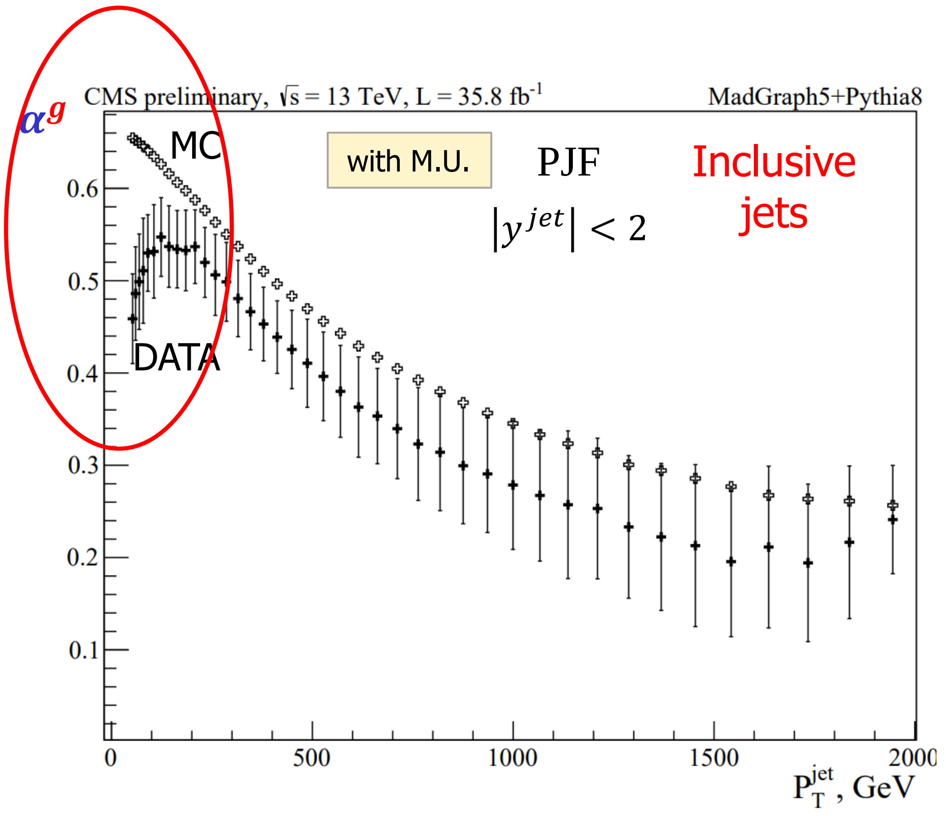
Figs from my SMP-HAD talk (June 2023)

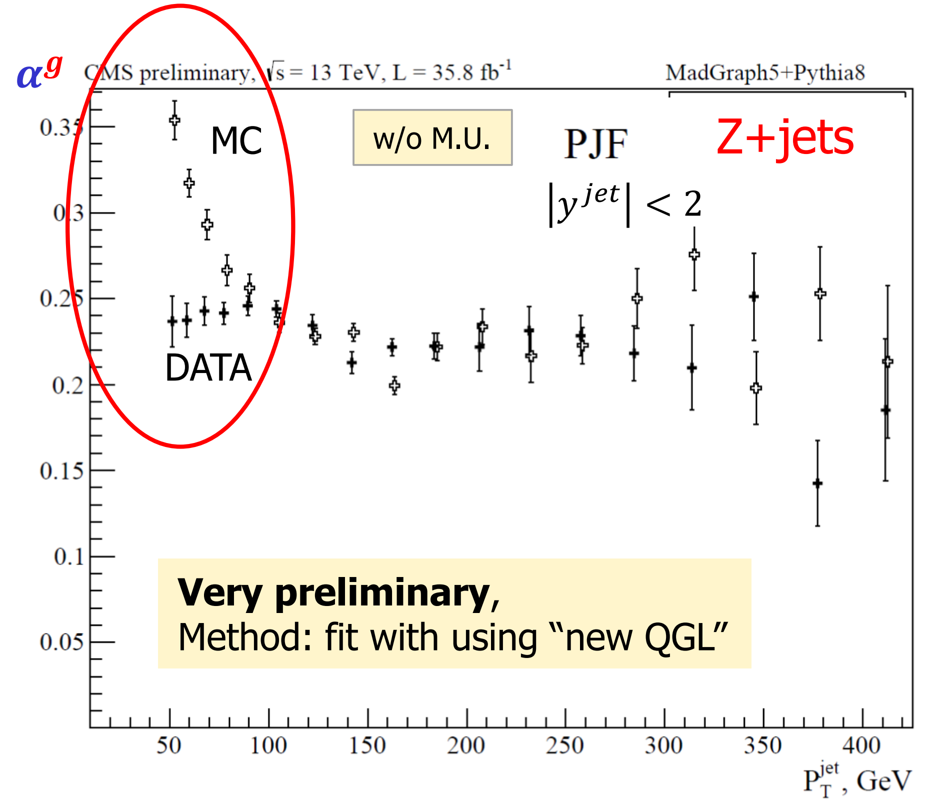
Run-II(2016)

MadGraph5+Pythia8

ak4-jets: R = 0.4

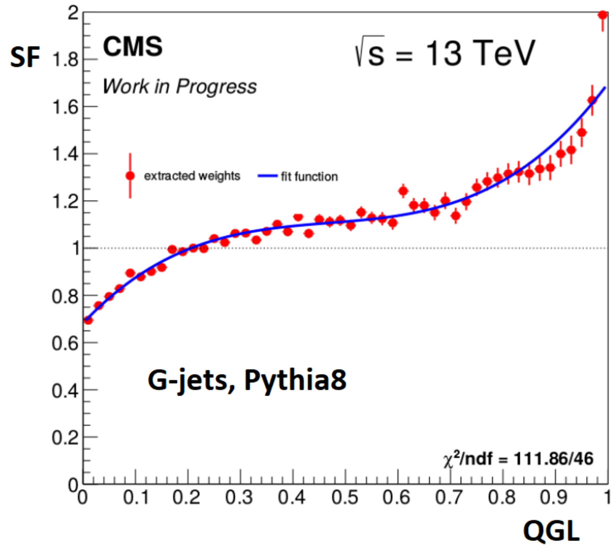
- **g-jet suppression** is visible at low P_T^{jet} in "**Inclusive jets**" and in "**Z+jets**"

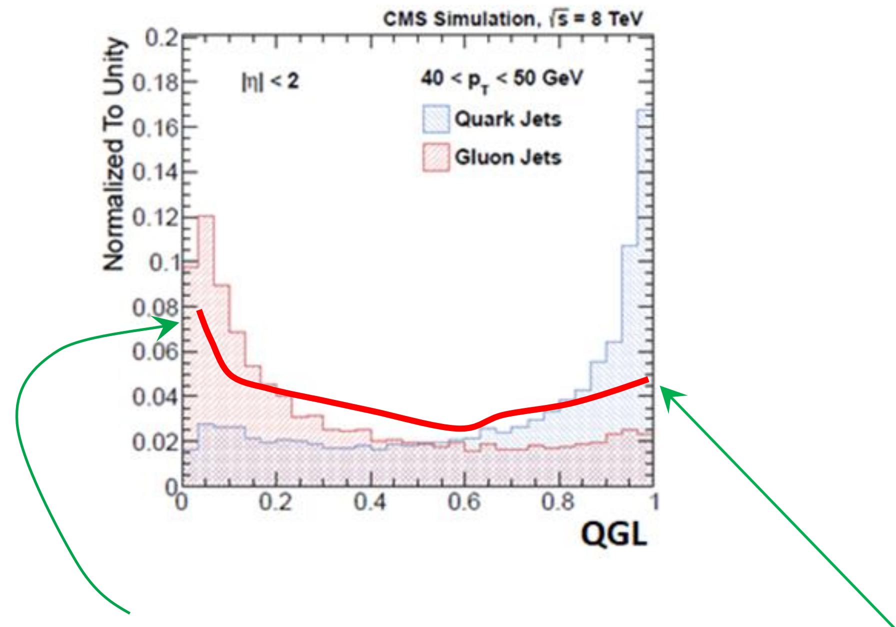




- First indirect observation of g-jet suppression was demonstrated in q/g-tagging group for Run-1(in PAS JME-13-002) and Run-2(2016) :

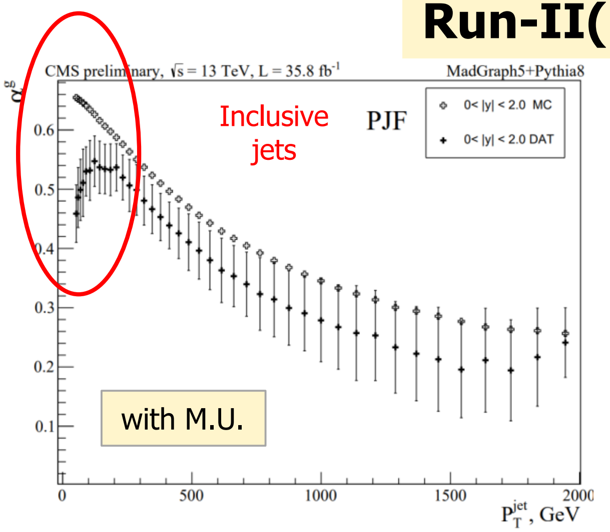
  This has been demonstrated a long time ago. But **only now we understand why gluon SF was so big** - the reason for this is wrong g-factions used in official SF.

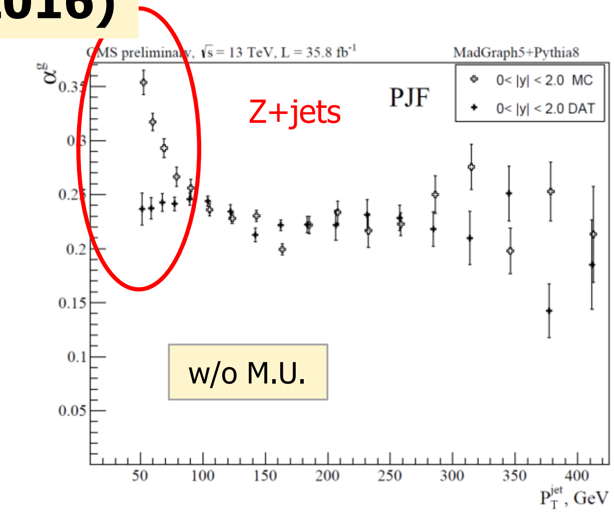




- SF modifies g-template: left gluon peak is 35% lower and right quark peak is 100% higher than original MC g-template

Run-II(2016)

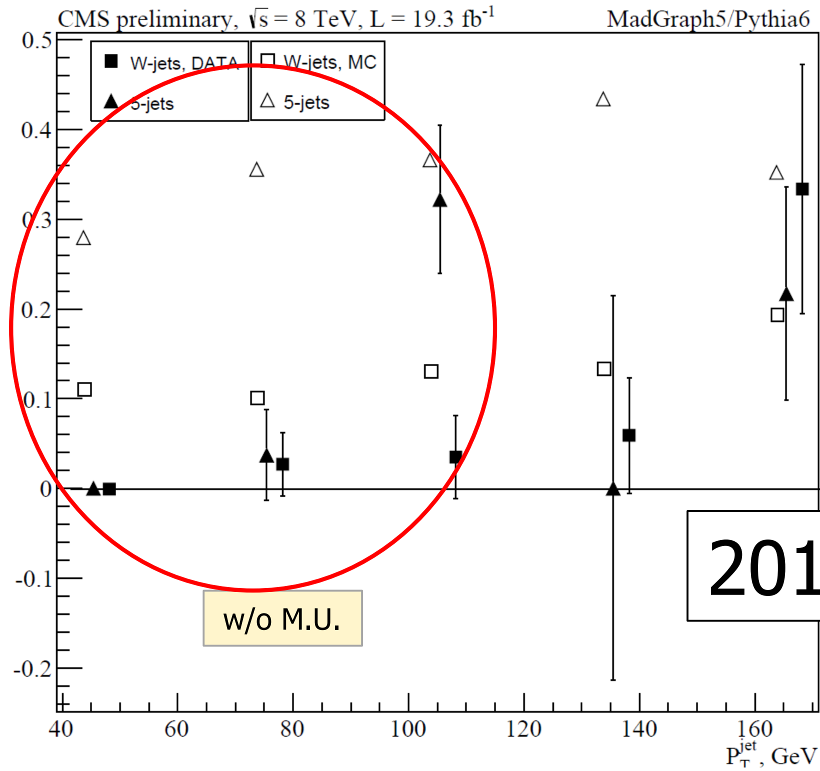




- **S**imilar results we obtained earlier for **Run-I (2012)**
- **R**un-I results are documented:

S.S., S.Shmatov, A.Zarubin: CMS AN-2018-131, **2018**
S.S. D.Budkouski, CMS AN-2020-143, **2020**
S.S. D.Budkouski, CMS AN-2021-024, **2021**
S.S. SMP-HAD Workshop, 11 Feb **2020**, https://indico.cern.ch/event/861896/
S.S. SMP-HAD Meeting, 1 June **2018**, https://indico.cern.ch/event/732652/

α^g

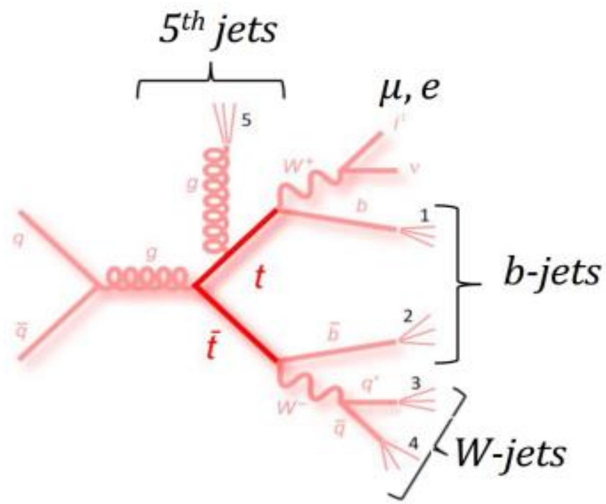


2018

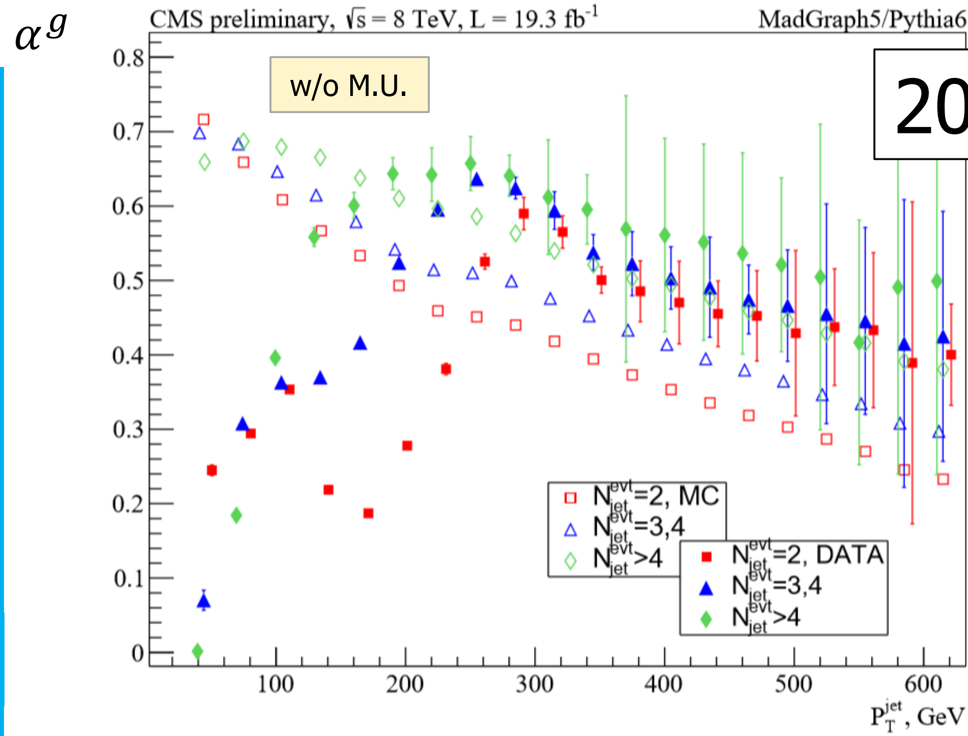
MadGraph5+Pythia6

ak5-jets: R = 0.5

- Semileptonic $t\bar{t}$ channel
- **M.U.** is not shown



N_{jets}^{evt}	Jet name	P_T^{jet}, GeV	$\alpha_k^{g,DAT}$, %	$\alpha_k^{g,MC}$, %
4	W-jets	30÷150	0÷5 (±5)	10÷11
≥ 5	5th-jets	30÷90	0÷3 (±5)	28÷34



2020

MadGraph5+Pythia6

ak5-jets: R = 0.5

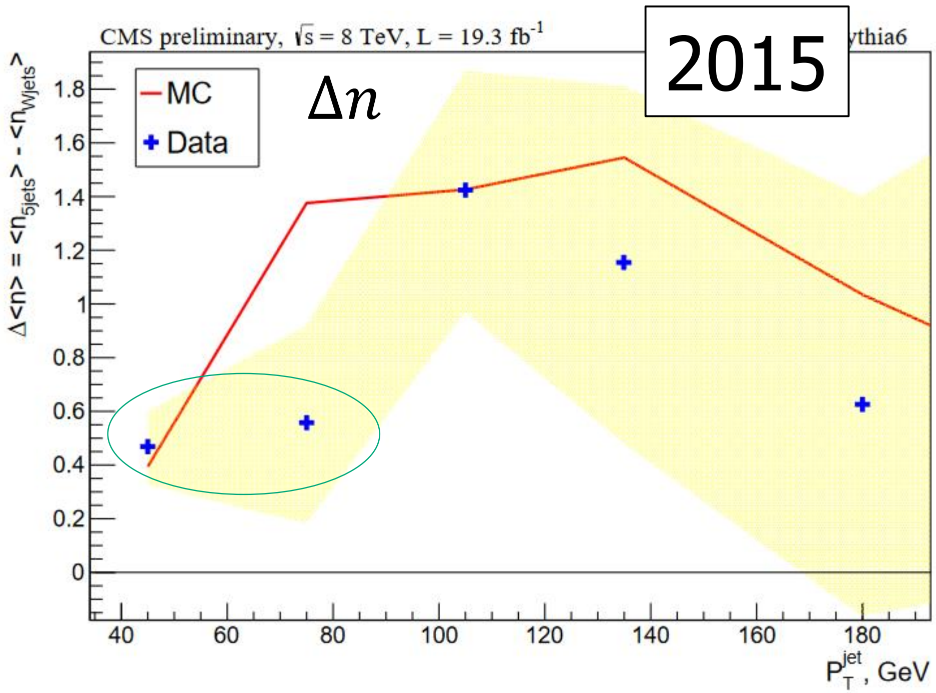
- Dijet, Run-I(2012)
- HLT prescaling is not taken into account

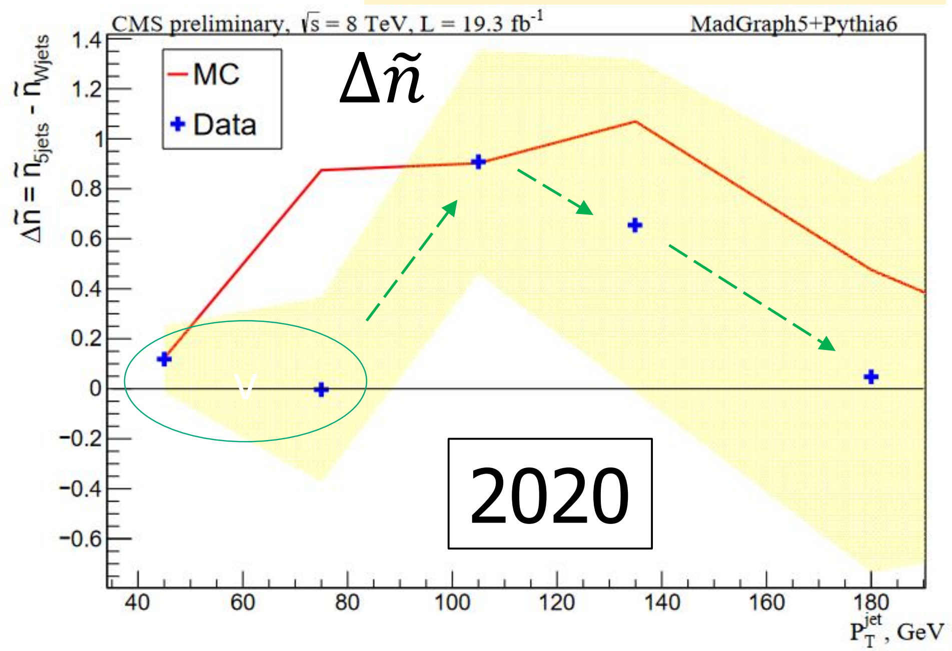
N_{jets}^{evt}	P_T^{jet}, GeV	$\alpha_k^{g,DAT}$, %	$\alpha_k^{g,MC}$, %	Name
2	30÷210	16÷35	72÷50	"dijet-1" (red)
3,4	30÷180	6÷40	70÷60	"dijet-2" (blue)
≥5	30÷120	0÷40	65÷69	"dijet-3" (green)
4	30÷150	0÷5 (±5)	10÷11	W-jets
≥ 5	30÷90	0÷3 (±5)	28÷34	5th-jets

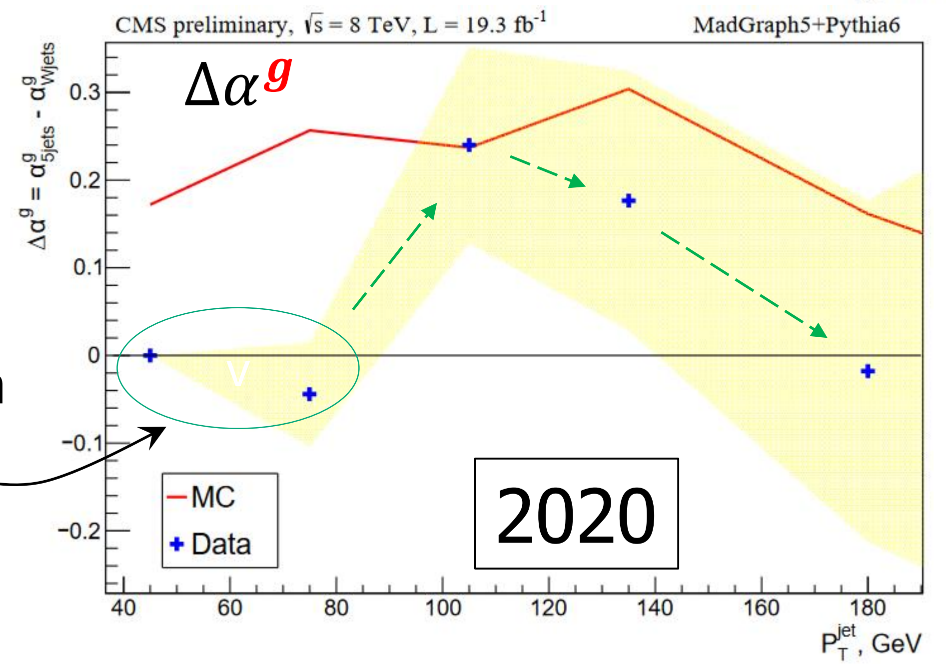
Semi-leptonic $t\bar{t}$

Run-I(2012) semileptonic $t\bar{t}$

$$A \cdot \Delta\tilde{n} = \Delta\alpha^g$$







- $\Delta\tilde{n}$ and $\Delta\alpha^g$ are similar:

$\Delta\tilde{n} = A\,\Delta\alpha^g \approx 0$ in 1st and 2nd bins **!**

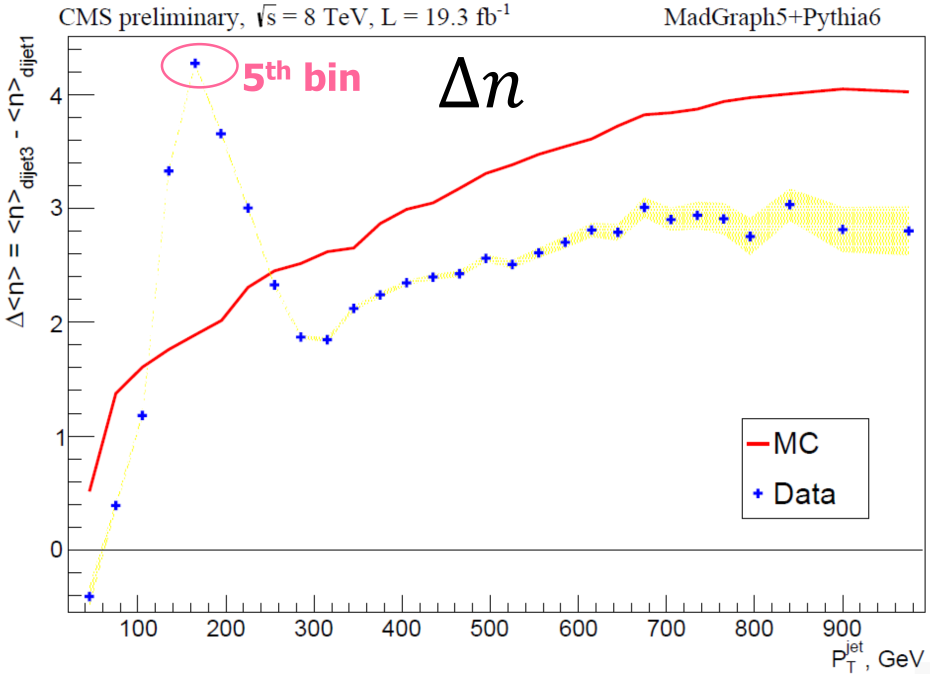
- Measurement of mean jet **C.P.M**'s indirectly confirms g-jet suppression

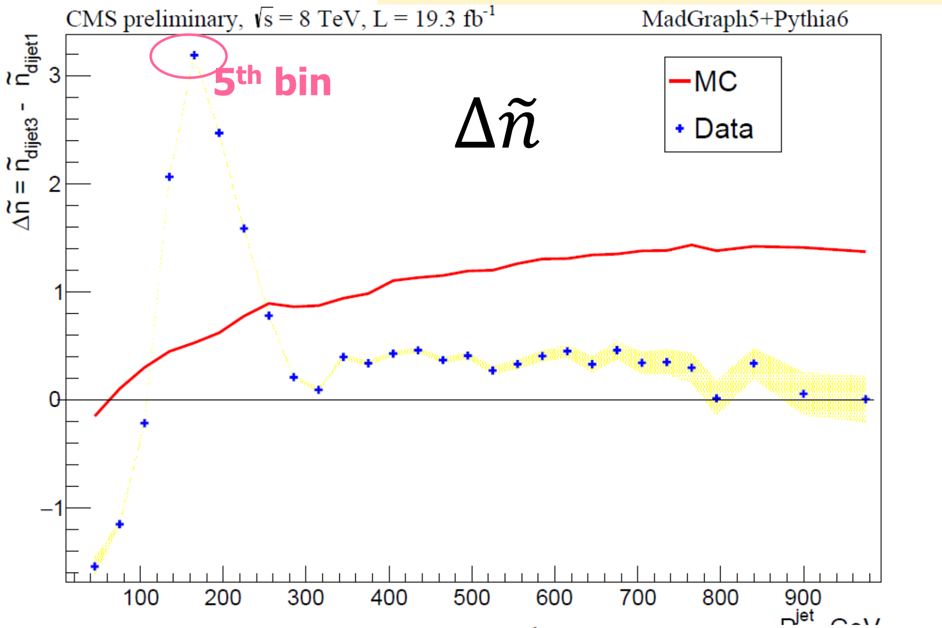
"Test"

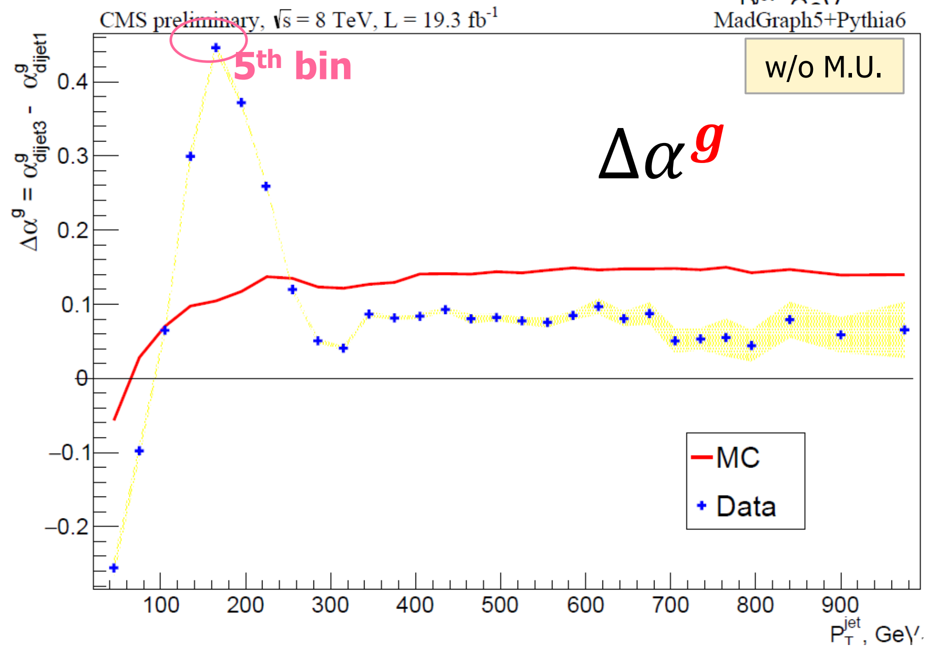
Run-I(2012) Dijet

2020

$A \cdot \Delta\tilde{n} = \Delta\alpha^g$







- $\Delta\tilde{n}$ and $\Delta\alpha^g$ are **similar** in all bins:

$$A \cdot \Delta\tilde{n} = \Delta\alpha^g \quad !$$

- Measurement of mean jet **C.P.M's** indirectly **confirms *g*-jet suppression at low P_T^{jet}**

"Test"

- Measurement of g-fractions was proposed, developed and implemented for many channels in CMS (Run-1 and Run-2)
- It was shown that g-fraction measurement should be a 1st stage in preparation of QGL-templates used in q/g-tagging
- Possible g-jet suppression in low P_T^{jet} region is observed by indirect model-independent measurement jet CPM, and by direct model-dependent g-fraction measurement, in several channels, for CMS Run-1 and Run-2 (**not approved** in CMS yet, but work is in final stage for inclusive jets channel)