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)

XIXth Workshop on High Energy Spin Physics
dedicated to 90th anniversary of A.V. Efremov birth
4-8 September, 2023
JINR

- 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 (gluon jets have 1.5 times greater multiplicity), jet radius in (η, φ)-space (gluon jet is wider) or combination of simple MP's (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 $H^f_{\mathrm{DATA}}(V)$ using data is referred to as obtaining "data-driven Scale Factor" (SF) for 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 SF (or q/g-templates) we 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 $\alpha^g_{2\ \mathrm{MC}}$:

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

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

$$H^q_{\mathrm{DAT}} = \frac{\alpha^g_{2,\mathrm{MC}} H_{1,\mathrm{DAT}} - \alpha^g_{1,\mathrm{MC}} H_{2,\mathrm{DAT}}}{\alpha^g_{2,\mathrm{MC}} - \alpha^g_{1,\mathrm{MC}}} \qquad (2)$$
$$H^g_{\mathrm{DAT}} = (g \to q, 1 \leftrightarrow 2)$$

- 1st 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 from Eqs.(1) and (2) 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. like 1st Eqs (1):

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

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

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

$$H^q_{\rm DAT} = \frac{\alpha^g_{2,\rm DAT} H_{1,\rm DAT} - \alpha^g_{1,\rm DAT} H_{2,\rm DAT}}{\alpha^g_{2,\rm DAT} - \alpha^g_{1,\rm DAT}}$$
$$H^g_{\rm DAT} = (q \leftrightarrow g, 1 \leftrightarrow 2) \qquad (3)$$

- Before obtaining SF and $H^{q/g}_{\rm DAT}(V)$ 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 (3):

- Eqs.(3) give universal q/g-templates for any channel and any jet kinematics. But, MC q/g-templates depend on kinematics! We proposed method to introduce in Eqs.(3) kinematical non-universal terms (SS, D.Budkouski, PEPAN Lett 2021-2022)

Very important remark:

- g-fraction measurement with corrected q/g-templates Eqs.(3) gives the same $\alpha^g_{1,\rm DAT}$ and $\alpha^g_{2,\rm DAT}$. So, 1st measurement of g-fractions with MC q/g-templates cannot be improved by SF – iteration process is impossible!

Proposition: $\alpha^{g\prime}_{1,\rm DAT} \equiv \alpha^g_{1,\rm DAT}$

Tip for the careful listener: to prove this, we need to write two equations

1st iteration $\alpha^g_{1,\rm DAT}$ is a solution of Eq.: $H_{1,\rm DAT} = \alpha^g_{1,\rm DAT} \cdot H^g_{\rm MC} + (1 - \alpha^g_{1,\rm DAT}) \cdot H^q_{\rm MC}$

2nd iteration $\alpha^{g\prime}_{1,\rm DAT}$ is a solution of Eq.: $H_{1,\rm DAT} = \alpha^{g\prime}_{1,\rm DAT} \cdot H^g_{\rm DAT} + (1 - \alpha^{g\prime}_{1,\rm DAT}) \cdot H^q_{\rm DAT}$

Tip for the careful listener (cont.):

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

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

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

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

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

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

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

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

- 2nd iteration for g-fraction measurement is impossible!
- The model determines g-fraction in the data unambiguously and does not allow it to be corrected if the original model is not changed
- However, there is a way to define quantitatively discrepancy between model q/g-templates and data ones in measured g-fractions – it is Model Uncertainty (M.U.)
- To find M.U., we need to use several independent jet macro parameters…

- If $\alpha^g_{\text{DAT}} \approx \alpha^g_{\text{MC}}$ then official SF ≈ new SF
- Spoiler: we found strong g-jet suppression in region $P_T^{jet} < 200$ GeV:

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

- Thus, official CMS SF's developed for Run-1 and Run-2 are wrong: they correct the g-factions $\alpha^g_{\text{DAT}} \to \alpha^g_{\text{MC}}$

It is our negative contribution to CMS "q/g-tagging".
It should be taken into account in CMS Run-3 analyses

Now we are moving to g-fraction measurements…

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

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

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 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) $V \equiv V_{1,2,3,4,\dots}$:

$$H^{\text{MC,DAT}}(V) = \alpha^{g} H^{g}(V) + (1-\alpha^{g})\, H^{q}(V) \qquad (5)$$

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

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

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

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

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 June 2023 we implemented this method and showed results in CMS
- Deprecated method: So far, we have used a more complex method with QGL and with fit:

WLS or LS methods by ROOT/MINUIT: $H_{\text{DAT}} \sim \alpha^g_{\text{DAT}} \cdot H^g_{\text{MC}} + (1 - \alpha^g_{1,\text{DAT}}) \cdot H^q_{\text{MC}}$

- Method of “bin averaging” allows to find stable result for any/all jet MP’s $V_{1,2,3,4,\dots}$ with small statistics

- In case of MC, calculation with any jet MP $V_{1,2,3,\dots}$ gives the same $\alpha_1^g = \alpha_2^g = \alpha_3^g = \dots = \alpha^g$ 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, calculation with any MP $V_{1,2,\dots}$ gives different $\alpha_1^g \neq \alpha_2^g \neq \alpha_3^g \neq \dots$ because MC q/g-templates are not true for DATA

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

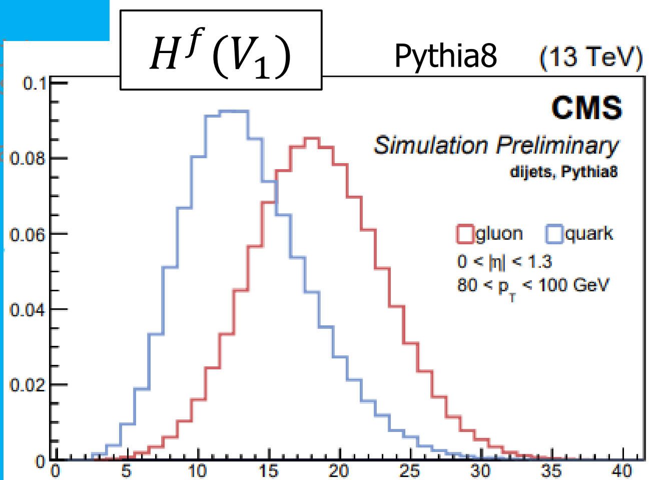
$$\textbf{M.U.} = \frac{1}{2} \cdot \max\{|\alpha_1^g - \alpha_2^g|, |\alpha_1^g - \alpha_3^g|, |\alpha_2^g - \alpha_3^g|, \dots\}$$

- 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

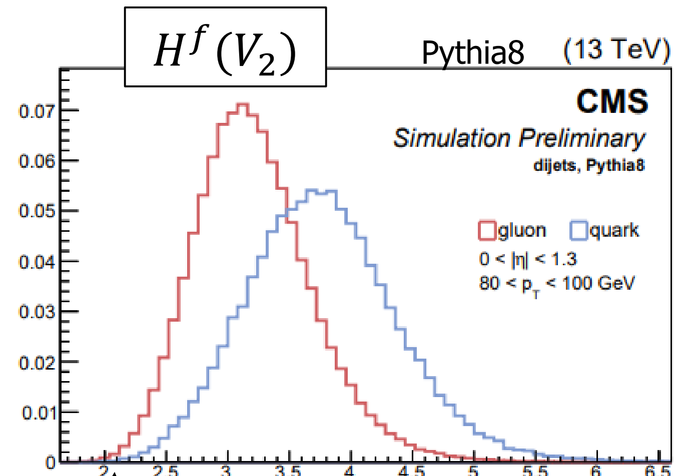
$V_{1,2,3} = (mult,\ a_2,\ p_T D) \equiv \vec{V}$

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



$V_1 = mult$

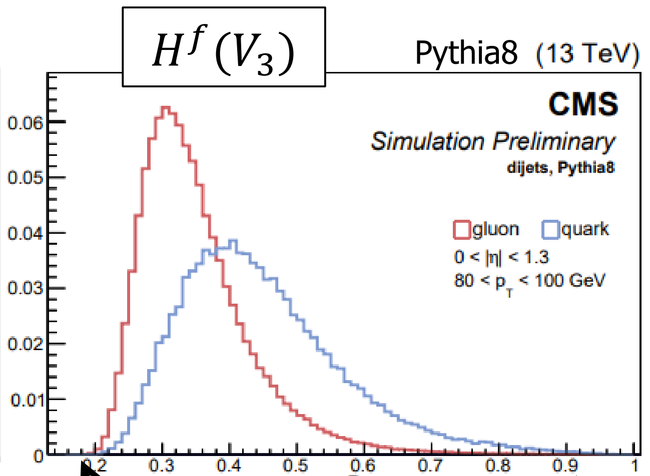
$\frac{n^g}{n^q} \approx 1.5$



$V_2 = -\log(a_2)$

Wider jets

Narrower jets



$V_3 = p_T D$

P_T^{jet} is uniformly distributed among constituents

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

These three q/g-templates 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 :

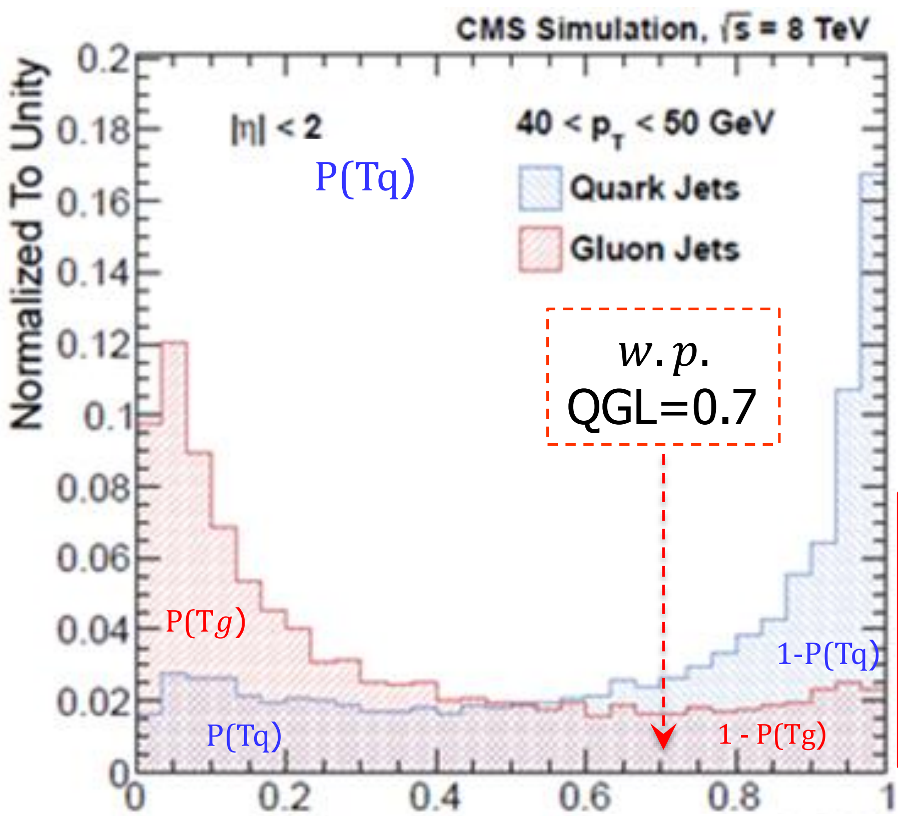
$$V_4 \equiv QGL = \frac{Q(\vec{V})}{Q(\vec{V}) + G(\vec{V})}$$

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

With respect to w.p.:

- Left/right red area= g-jet efficiency/rejection
- Left/right blue area= q-jet rejection/efficiency

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

- We show (July 2023) that QGL written in all CMS Run-2 datasets are wrong
- We prepared new QGL for CMS Run-2 and test them using g-fraction measurements

It is our contribution to CMS "q/g-tagging"
It should be implemented in
Run-2(2016-2018) analyses

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

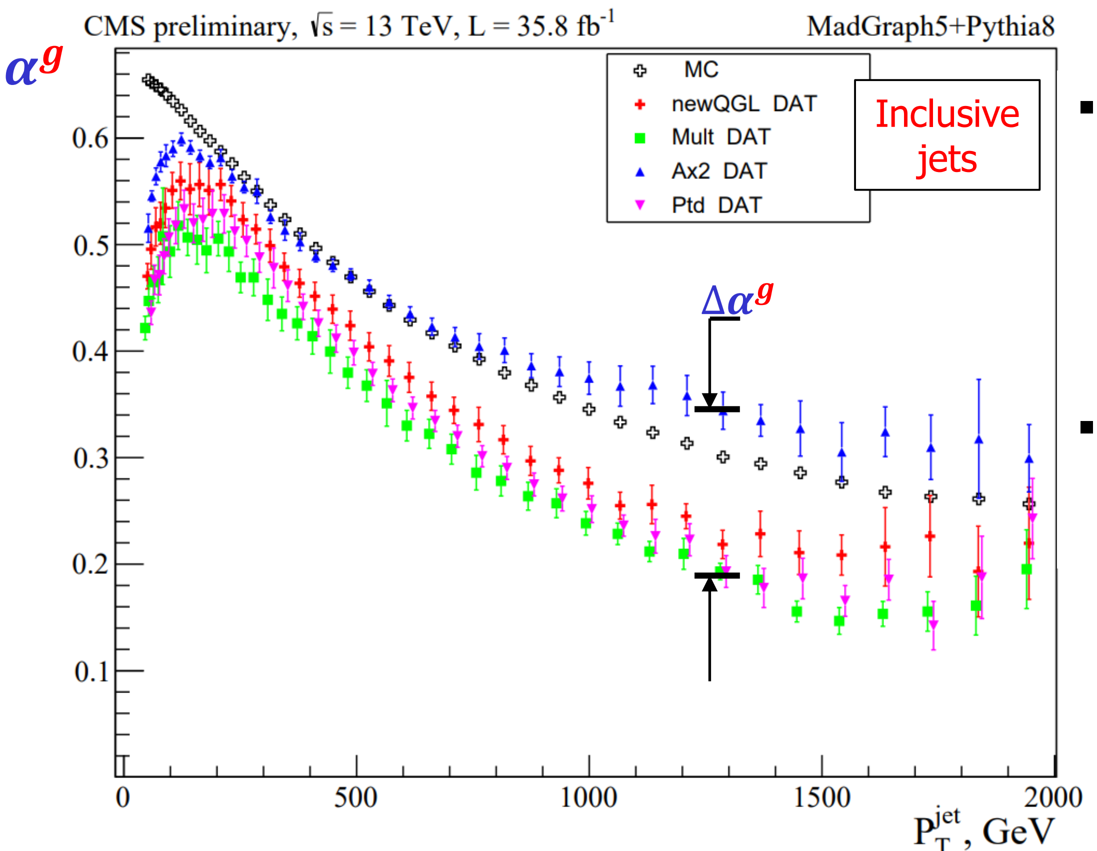


Fig. 3: Demonstration of M.U.

- This preliminary results were obtained in CMS group "Gluon-jet/Quark-jet analyses" [1]:

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

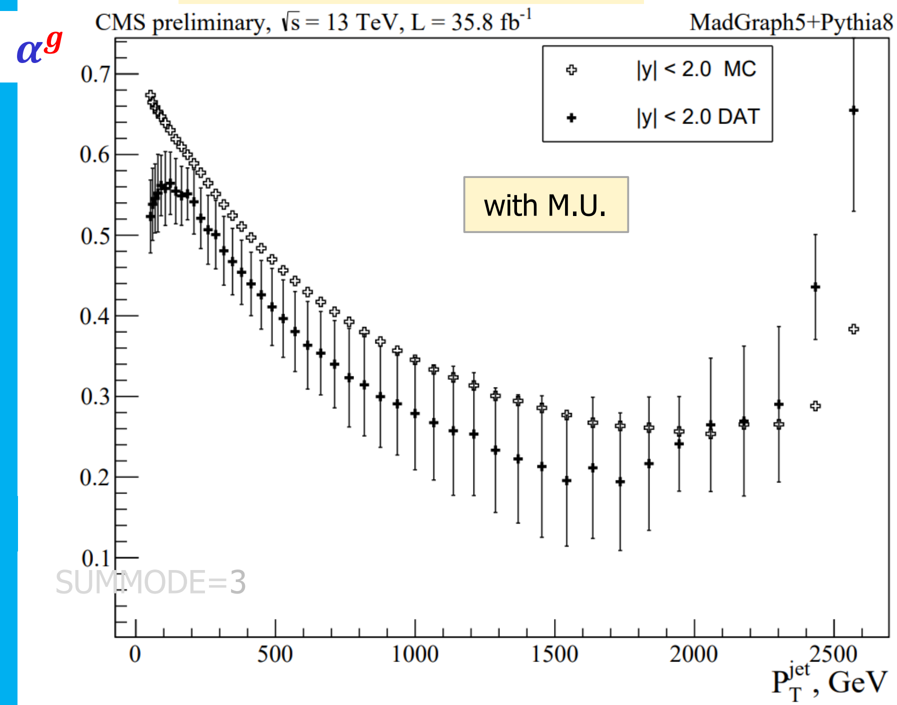
- This group was created in April 2021 purposefully to measure g-fractions in inclusive jet channel with Run-II data

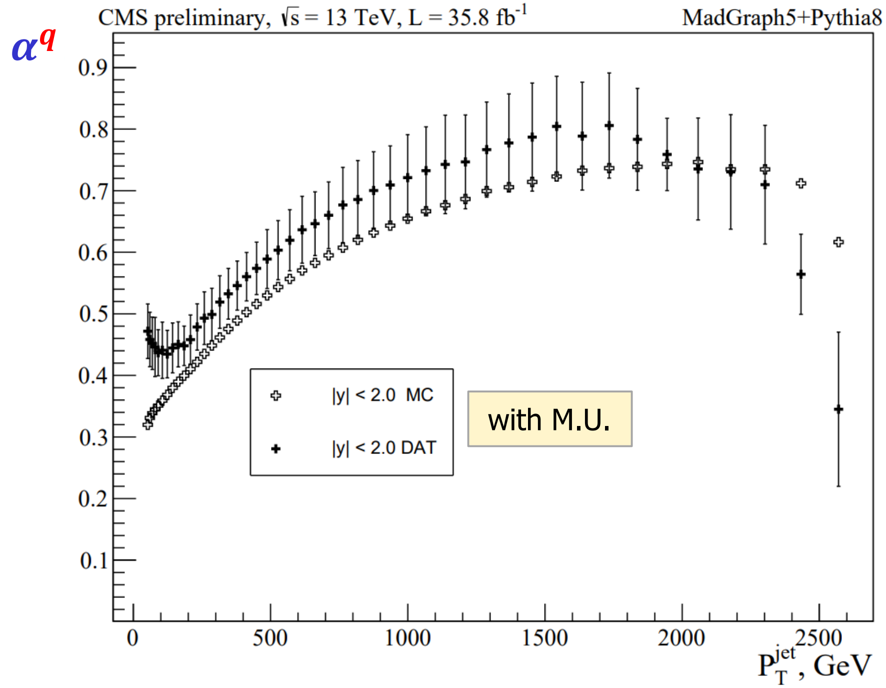
[1] https://indico.cern.ch/category/12755/

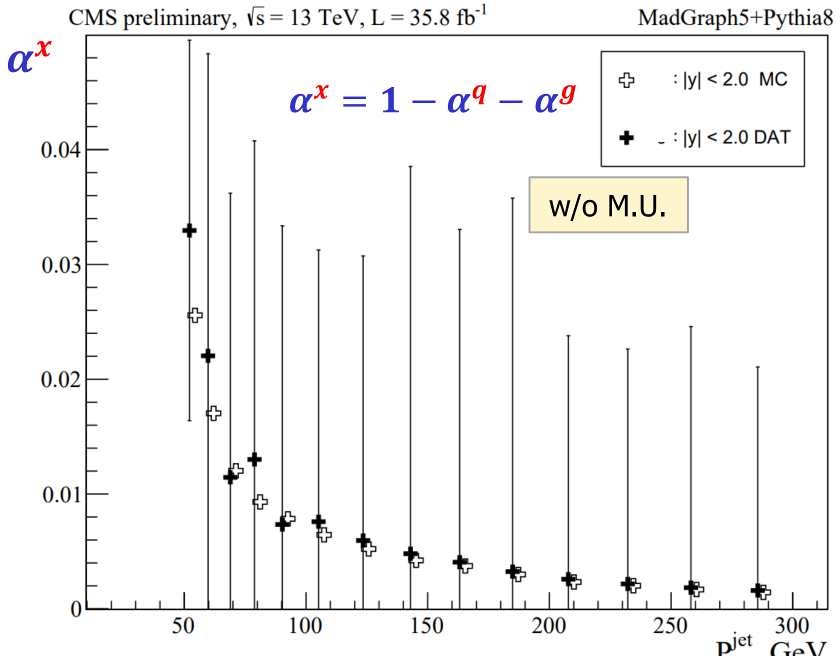
- Measurement of g-fraction demonstrates indirectly large deviation of true unknown DATA q/g-templates from Pythia8 ones

q/g/x-jet fractions

Inclusive jets







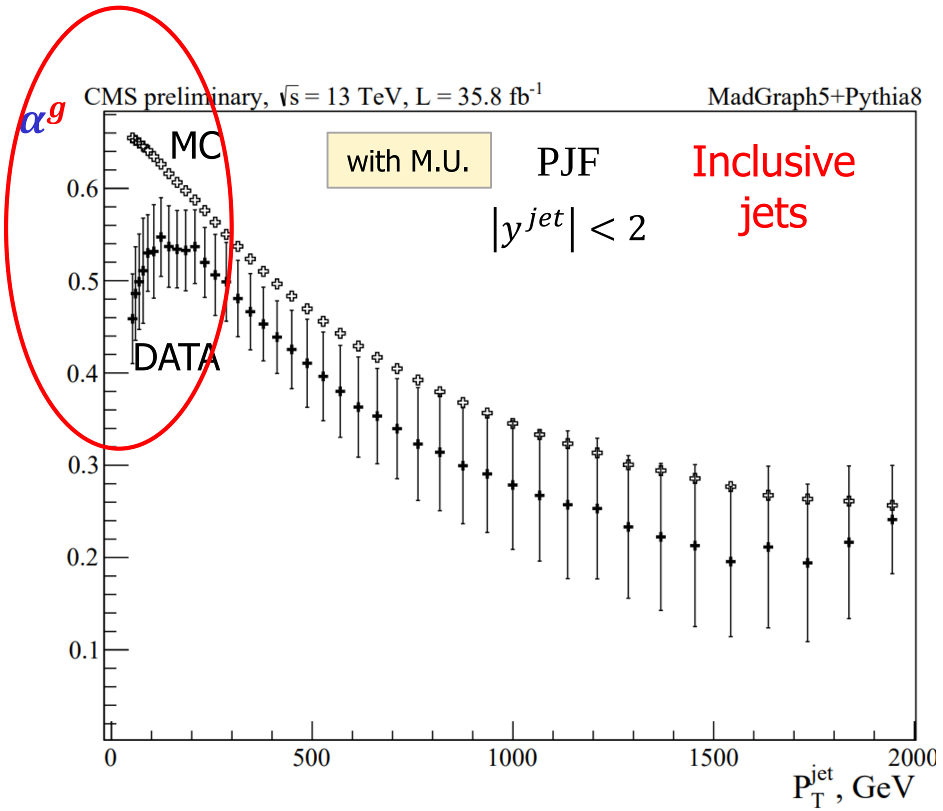
- This measurement results were obtained in CMS group "Gluon-jet/Quark-jet analyses"[1] :

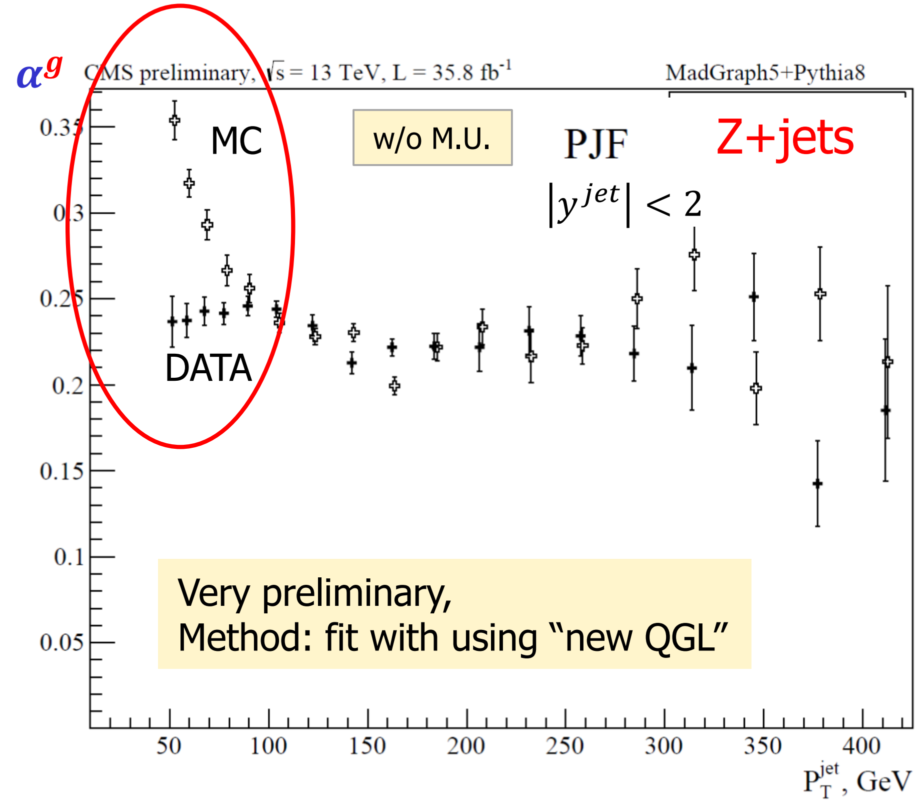
  S.S., D.Budkouski(JINR), J.Strologas (GR), O.Atakisi(TR)

[1]https://indico.cern.ch/category/12755/

Run-II(2016)

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

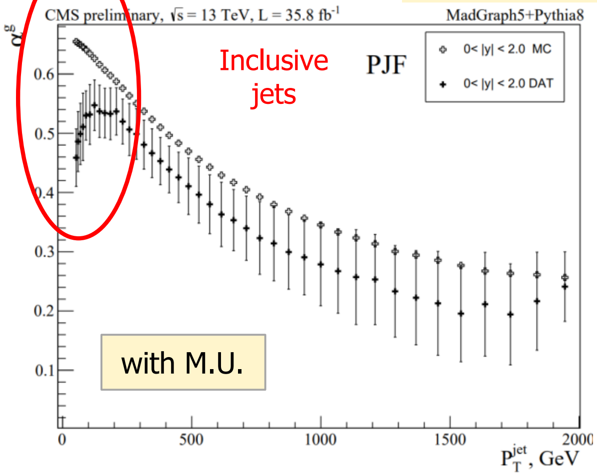


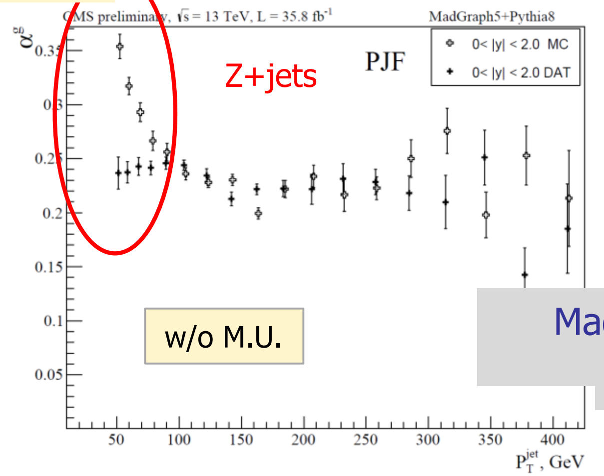


MadGraph5+Pythia8

ak4-jets: R = 0.4

Run-II(2016)



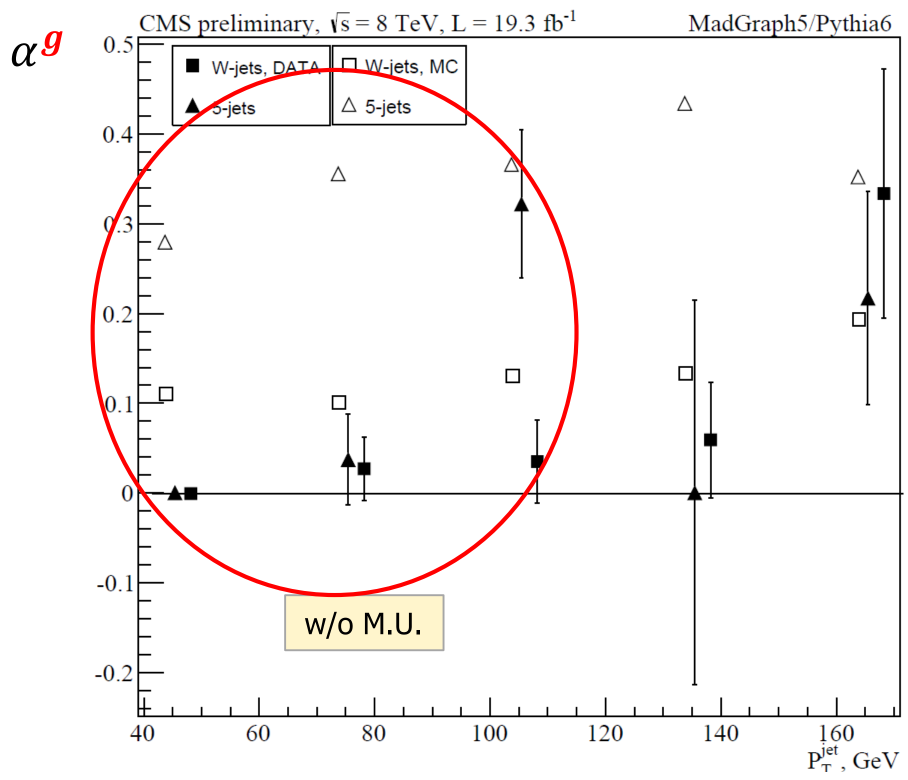


MadGraph5+**Pythia8**

ak4-jets: **R = 0.4**

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

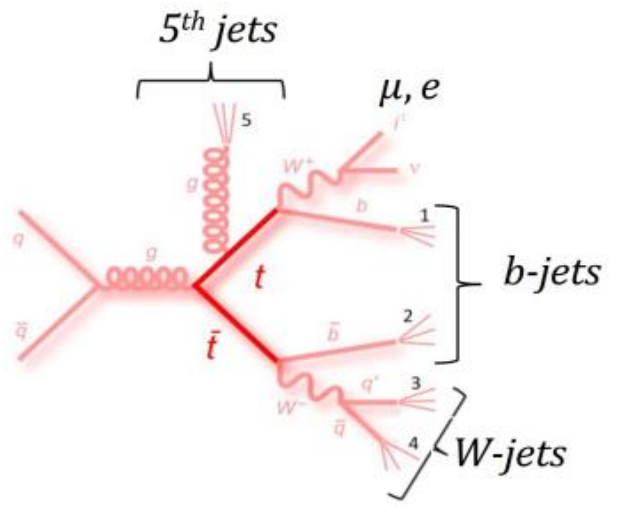
S.S., S.Shmatov, A.Zarubin: CMS AN-2018-131, **2018**
S.S. D.Budkouski, CMS AN-2020-143, **2020**
S.S. D.Budkouski, CMS AN-2021-024, **2021**
S.S. SMP-HAD Workshop, 11 Feb **2020**, https://indico.cern.ch/event/861896/
S.S. SMP-HAD Meeting, 1 June **2018**, https://indico.cern.ch/event/732652/



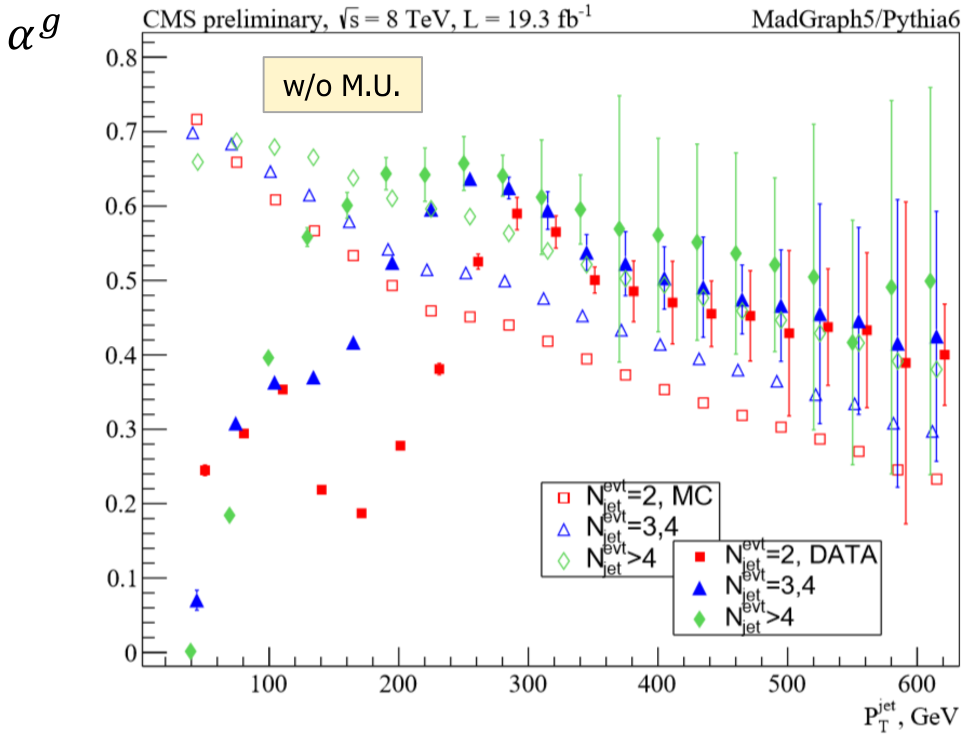
MadGraph5+**Pythia6**

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	5^{th}-jets	30÷90	0÷3 (±5)	28÷34



MadGraph5+Pythia6

ak5-jets: **R = 0.5**

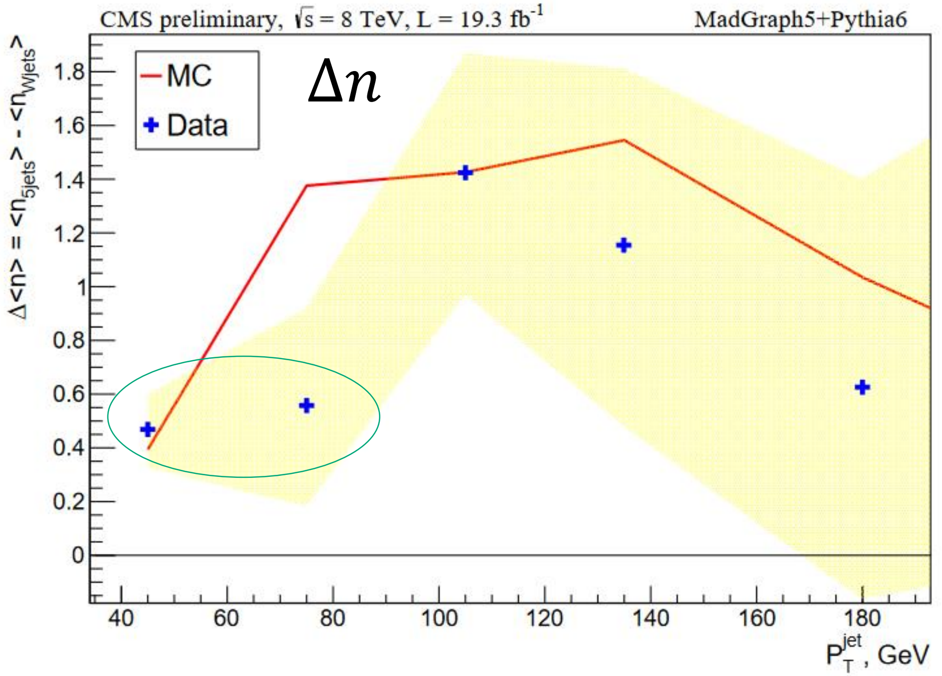
- Dijet, Run-I(2012)
- **M.U.** is not taken into account
- HLT prescaling is not taken into account

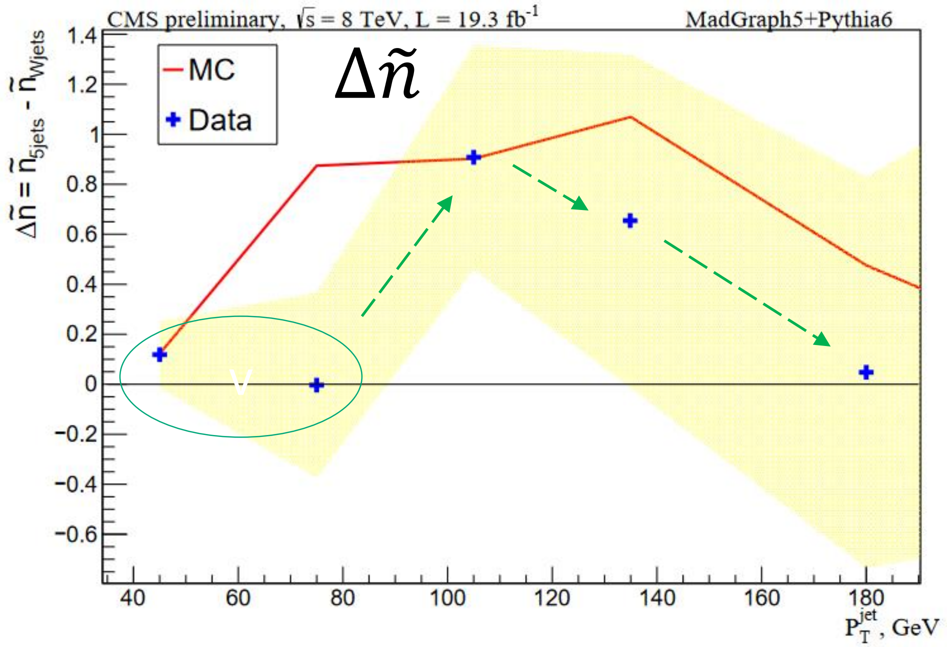
N^{evt}_{jets}	P^{jet}_T, GeV	$\alpha^{g,DAT}_k$, %	$\alpha^{g,MC}_k$, %	Name
2	30÷210	16÷35	72÷50	"dijet-1" (red)
3,4	30÷180	6÷40	70÷60	"dijet-2" (blue)
≥5	30÷120	0÷40	65÷69	"dijet-3" (green)
4	30÷150	0÷5 (±5)	10÷11	W-jets
≥ 5	30÷90	0÷3 (±5)	28÷34	5th-jets

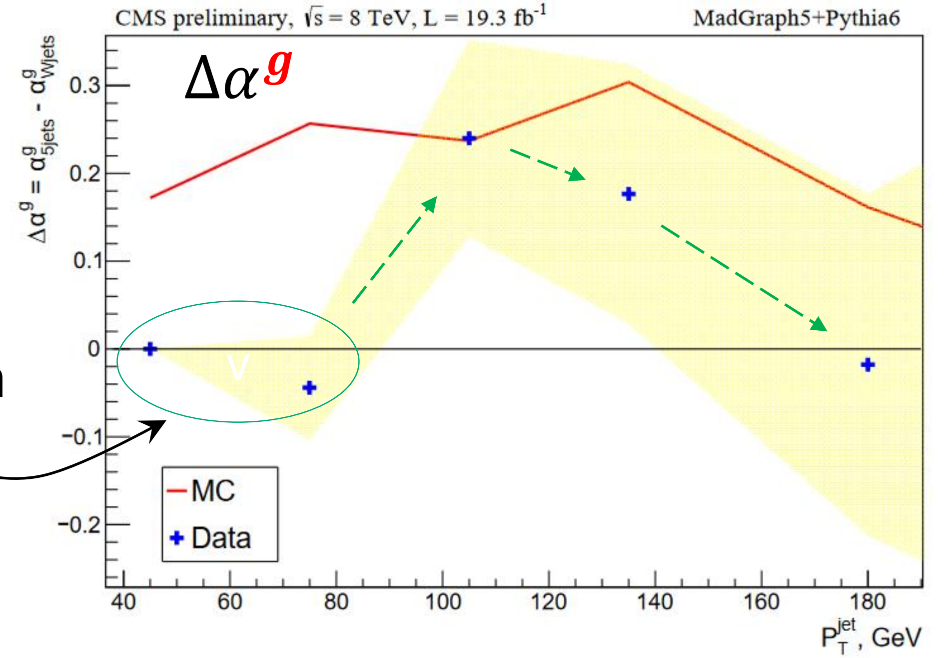
Semi-leptonic $t\bar{t}$

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

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







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

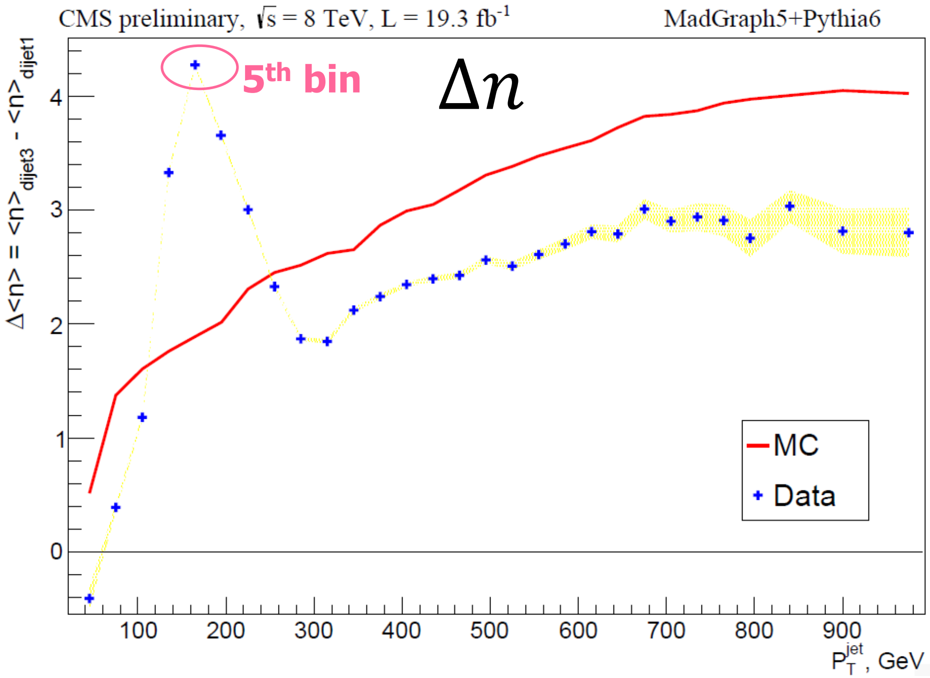
$\Delta\tilde{n} = A\,\Delta\alpha^g \approx 0$ in 1st and 2nd bins **!**

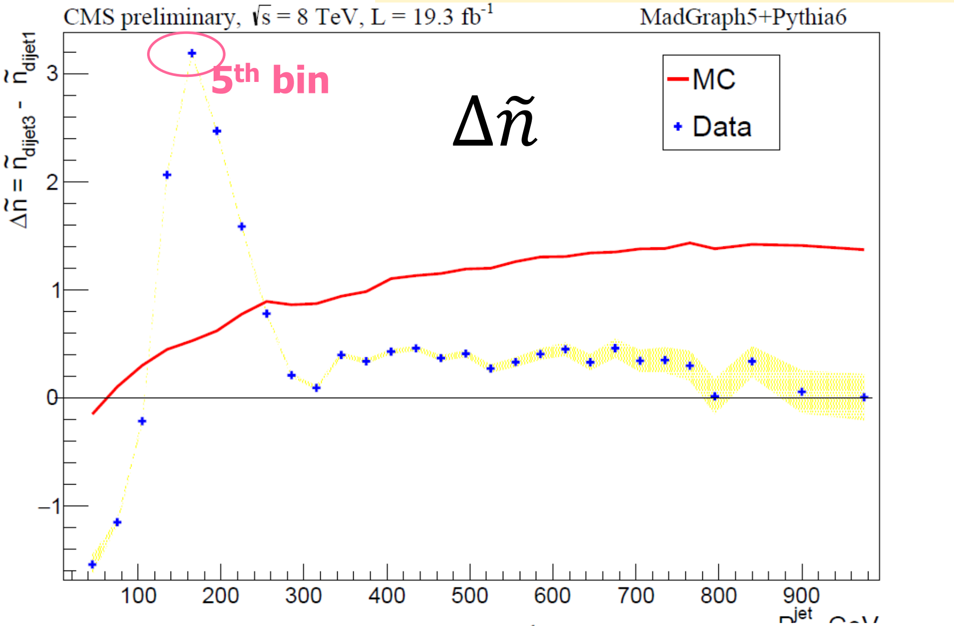
- Measurement of mean jet **C.P.M**'s indirectly confirms g-jet suppression

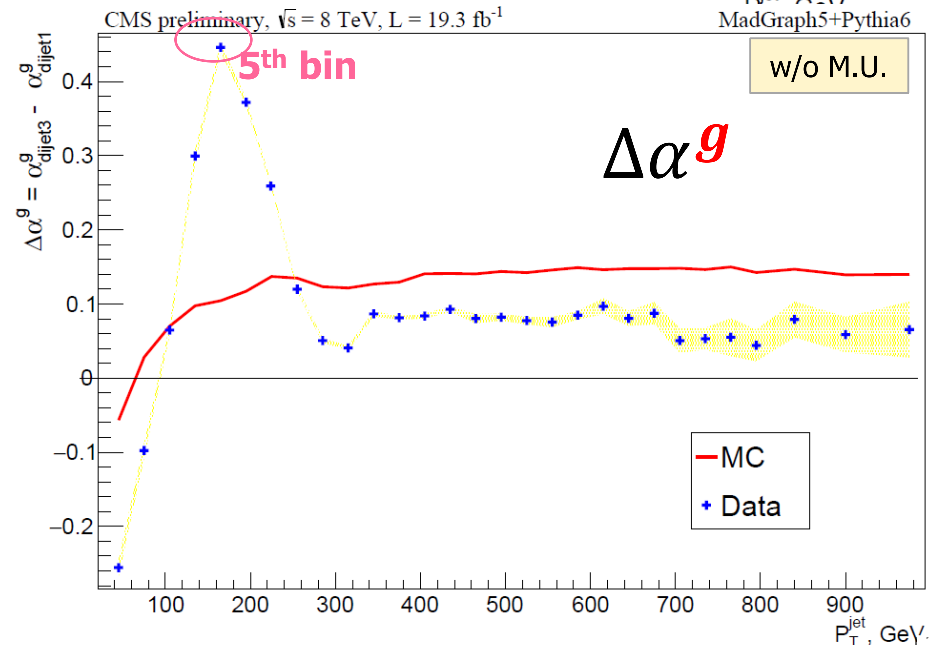
"Test"

Run-I(2012) Dijet

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







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

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

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

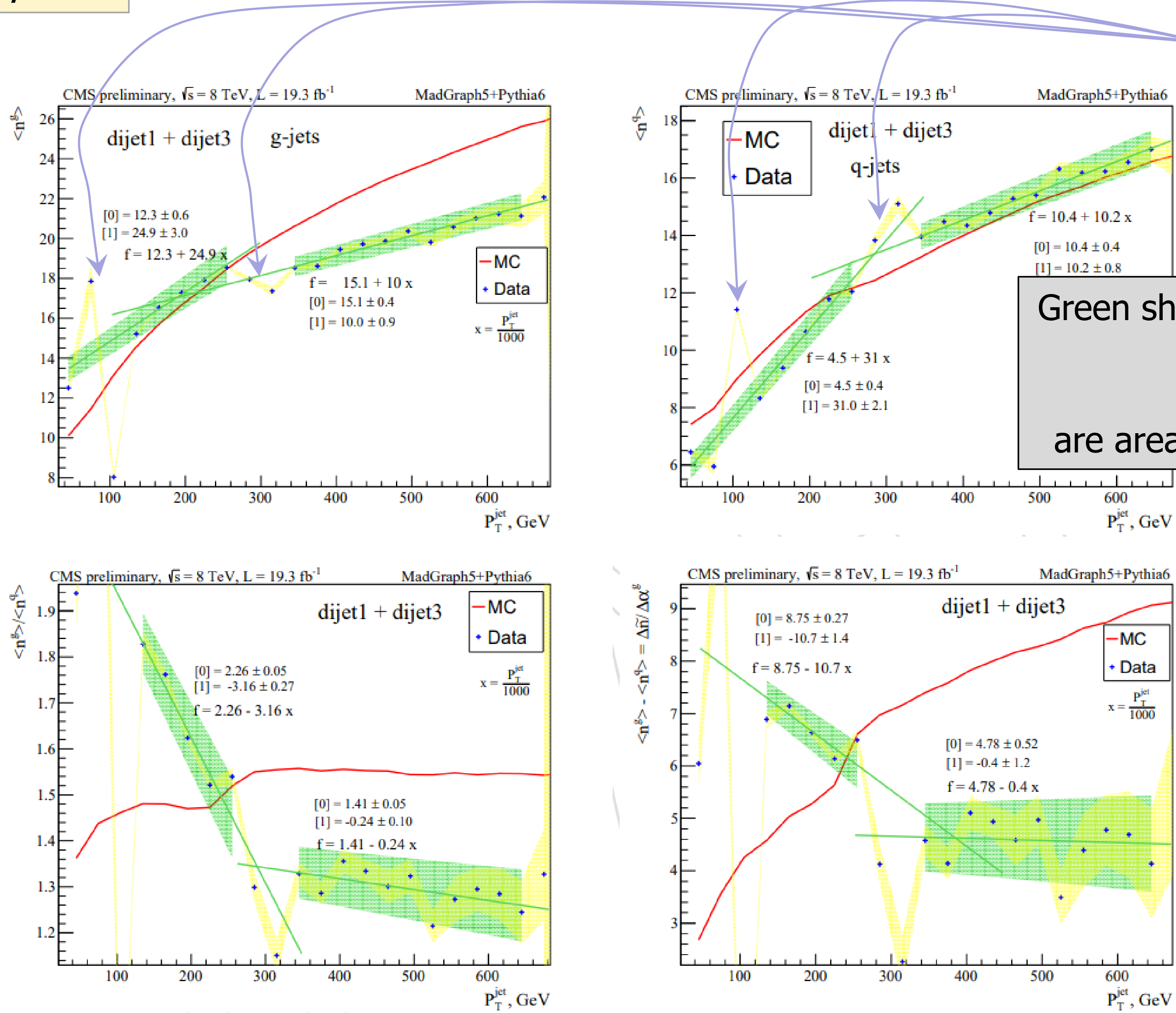
"Test"

- Measurement of g-fractions in many channels was proposed, developed and implemented in CMS (Run-1 and Run-2)
- It was shown that g-fraction measurement is a 1st stage in preparation of q/g-templates used in q/g-tagging
- Possible phenomenon of g-jet suppression in low P_T^{jet} region is observed by all studied channels, for CMS Run-1 and Run-2

Run-I(2012) Dijet

w/o M.U.

1st measurement of q/g-jet mean C.P.M's using measured g-fractions in two channels



Small denominator $\alpha_1^g - \alpha_2^g \sim 0$ in formulas makes solutions unstable

Green shaded areas with large denominator $\alpha_1^g - \alpha_2^g$ are areas of robust solutions

Two dijet samples used:

dijet1: $N_{jets}^{evt} = 2$

dijet3: $N_{jets}^{evt} \geq 5$