UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

ML4Jets 2024, Paris
07.11.2024

How to Unfold Top Decays

Luigi Favaro[1], Roman Kogler[2], Alexander Paasch[3],
Sofia Palacios Schweitzer[1], Tilman Plehn[1], Dennis Schwarz[4]

1 - Institut für theoretische Physik, Universität Heidelberg
2 - Deutsches Elektronen-Synchrotron DESY, Hamburg
3 - Institut für Experimentalphysik, Universität Hamburg
4 - Institut für Hochenergiephysik, Österreichische Akademie der Wissenschaft, Wien

SPONSORED BY THE
Federal Ministry of Education and Research

with Anja Butter

What I am not going to talk about

1. Coolness of ML-based unfolding

You find yourself in the unfolding session at ML4Jets

2. Methodology of generative unfolding

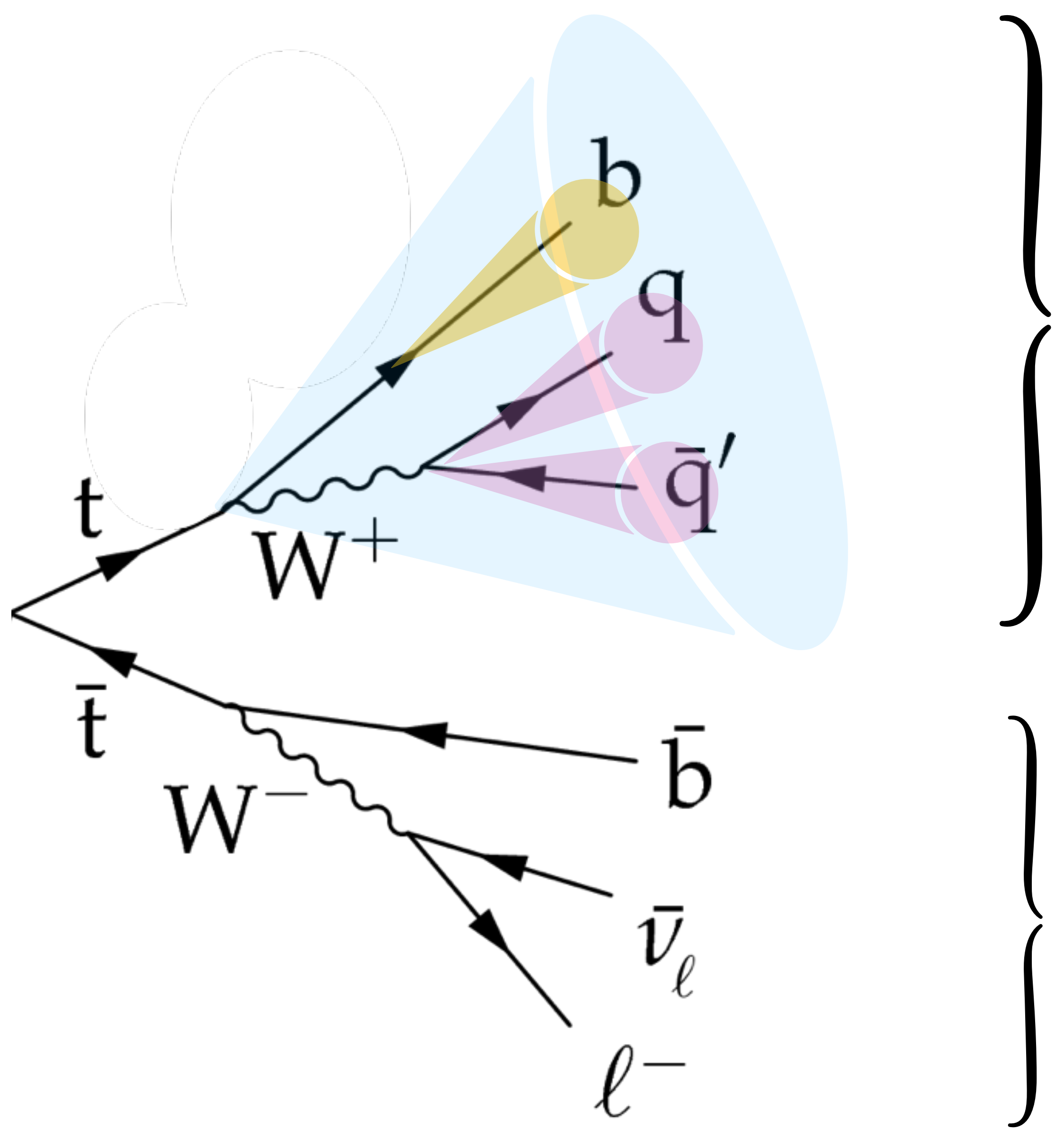
Nathan's overview talk

3. Generative Models (in particular CFMs)

Timo's talk, Jonas' talk, Luigi's talk, Dmitrii's talk …

Boosted top decays

$p_{T,J} > 400$ GeV

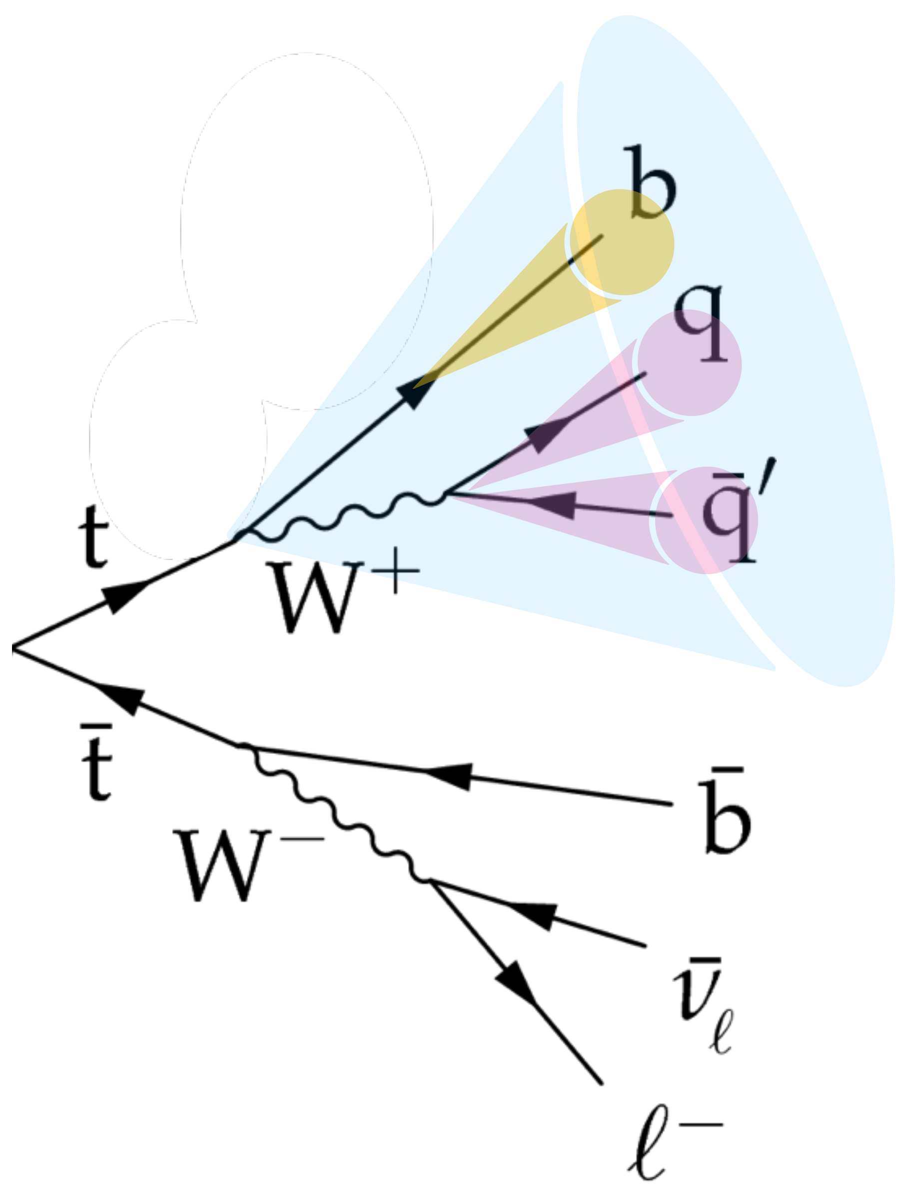


Reconstruct triple jet mass M_{jjj} to measure m_t

Tag side

Boosted top decays

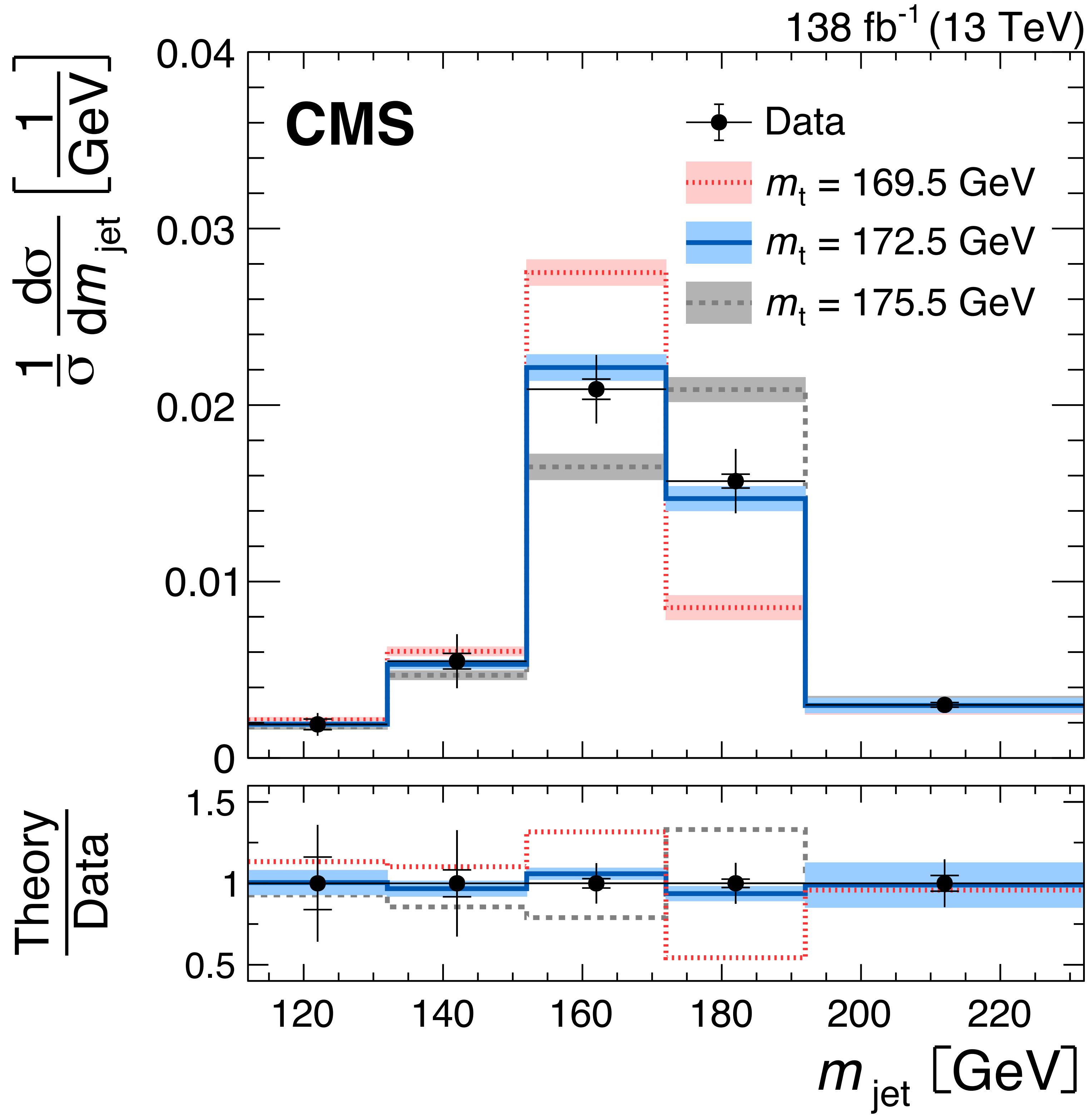
$p_{T,J} > 400$ GeV



Reconstruct triple jet mass M_{jjj} to measure m_t

Tag side

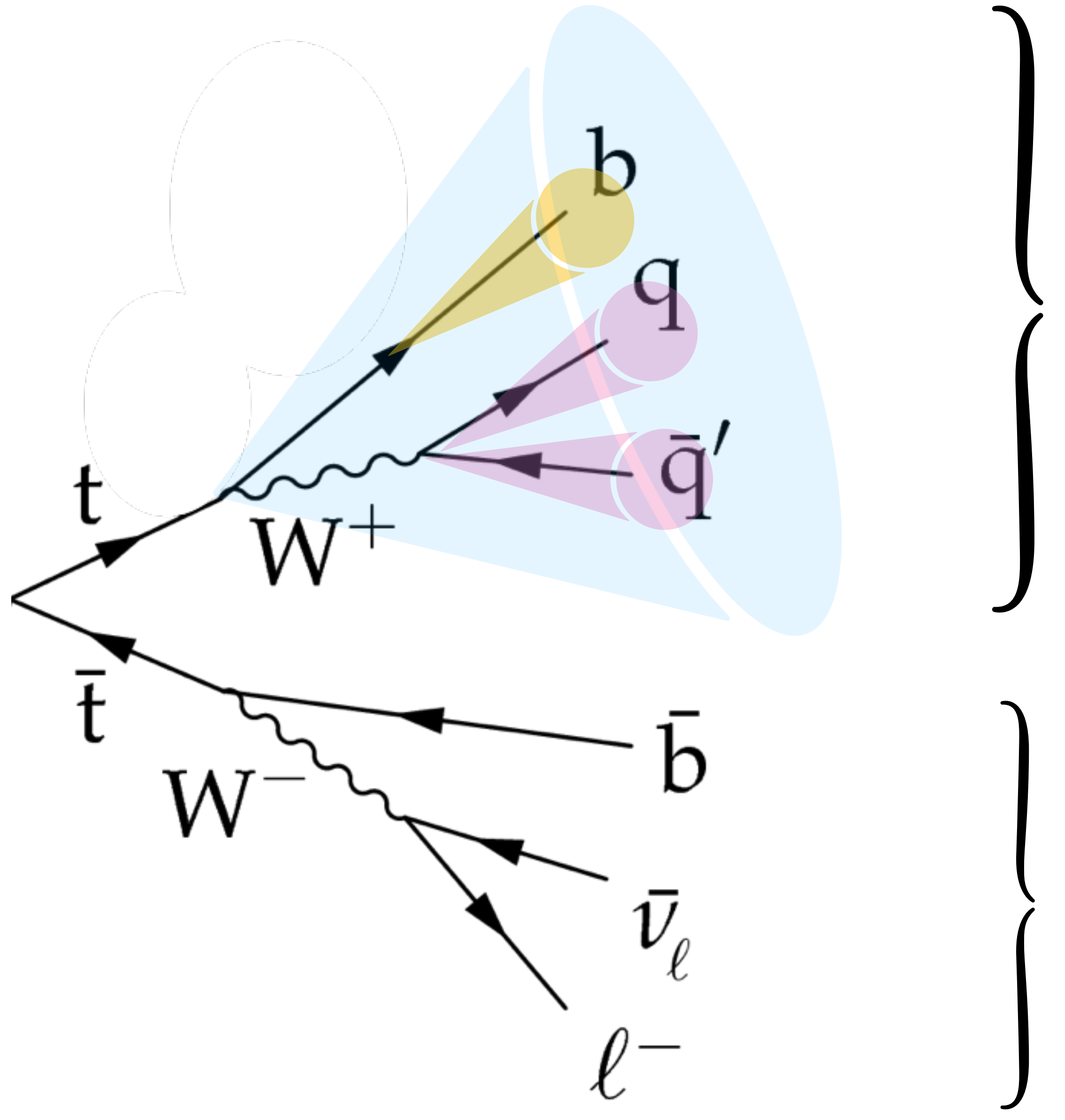
Previously done in CMS with TUnfold (classical binned unfolding algorithm)



CMS **2211.01456**

Boosted top decays

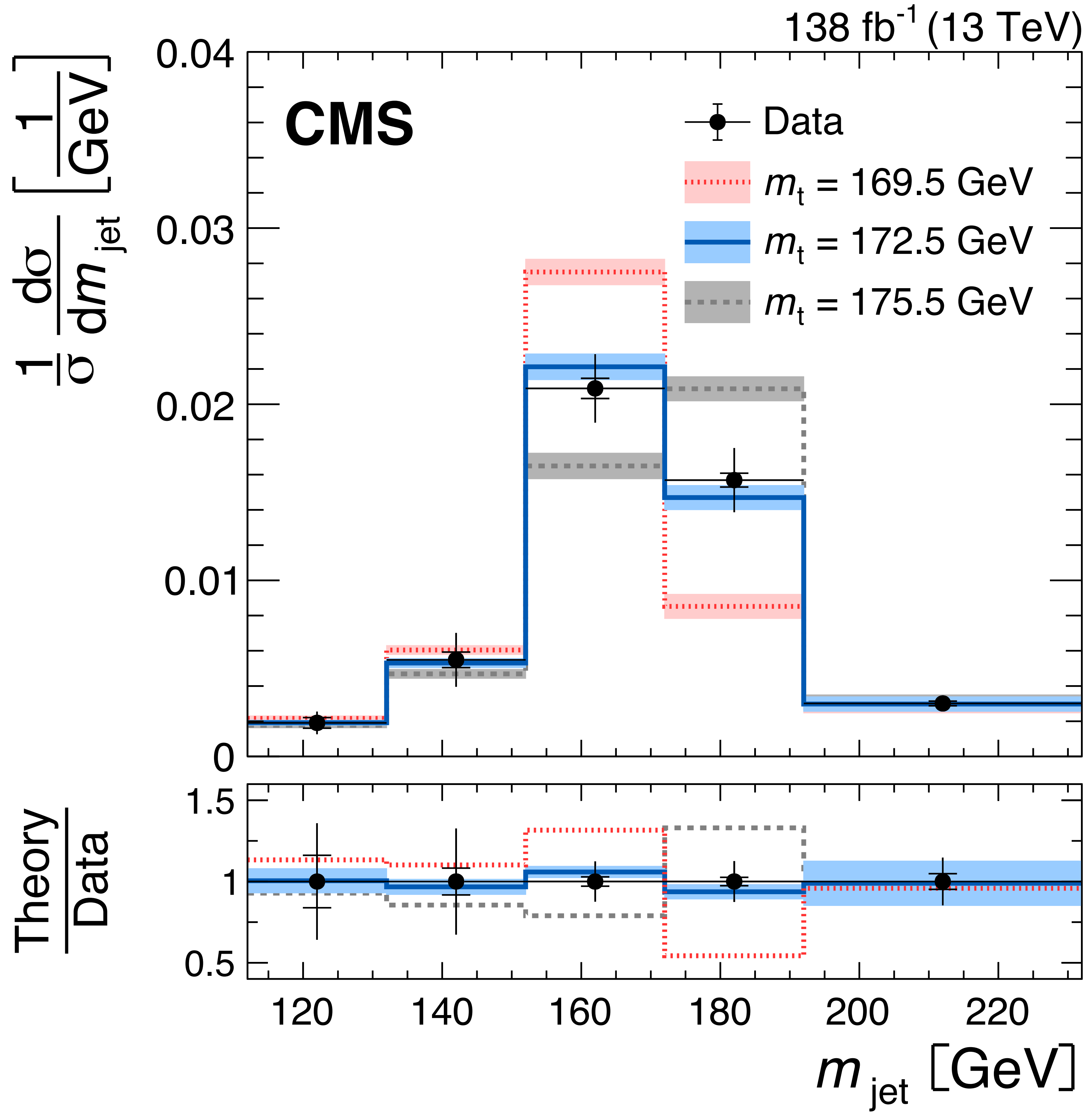
$p_{T,J} > 400$ GeV



Reconstruct triple jet mass M_{jjj} to measure m_t

Tag side

Previously done in CMS with TUnfold (classical binned unfolding algorithm)



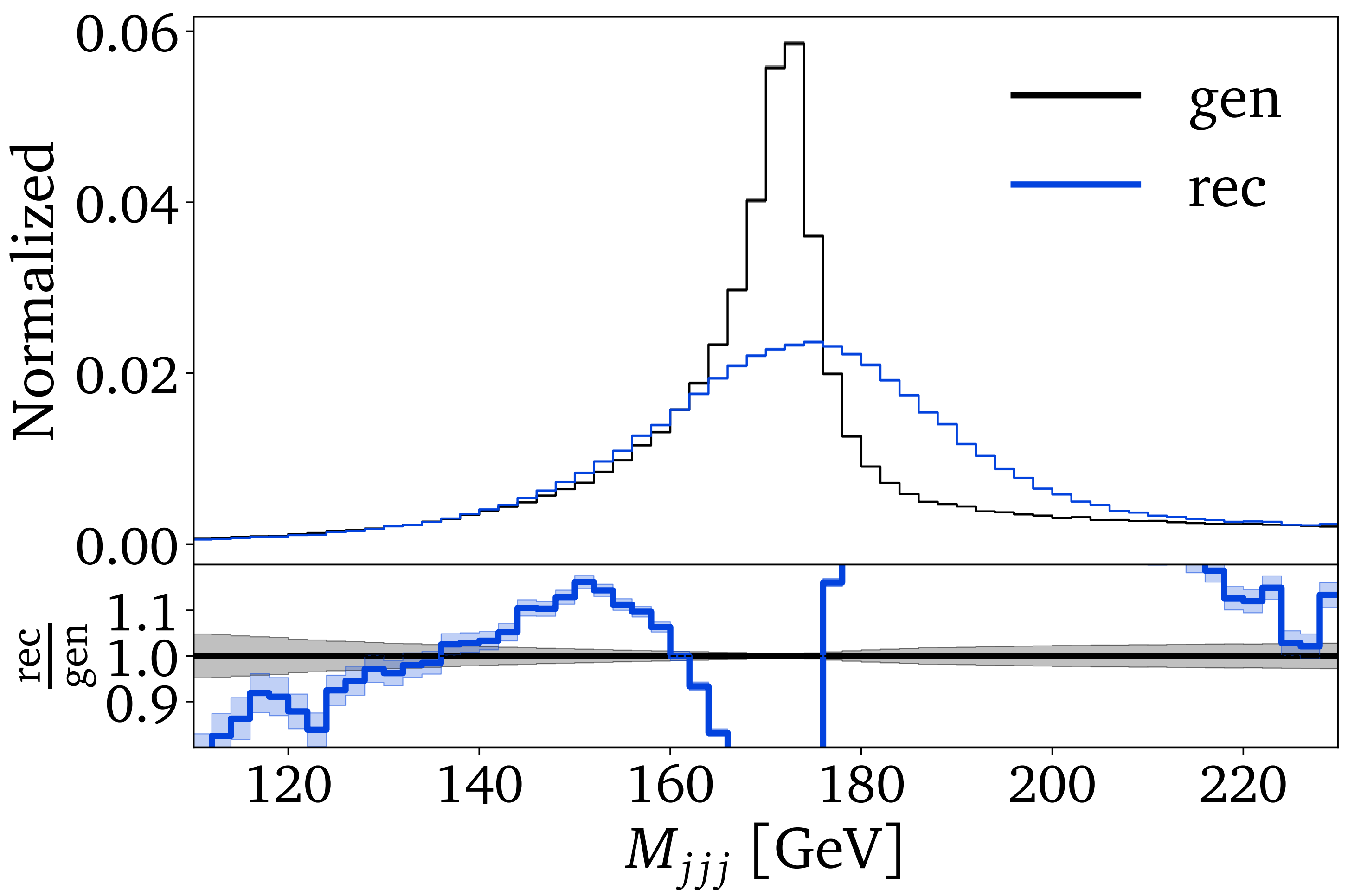
CMS **2211.01456**

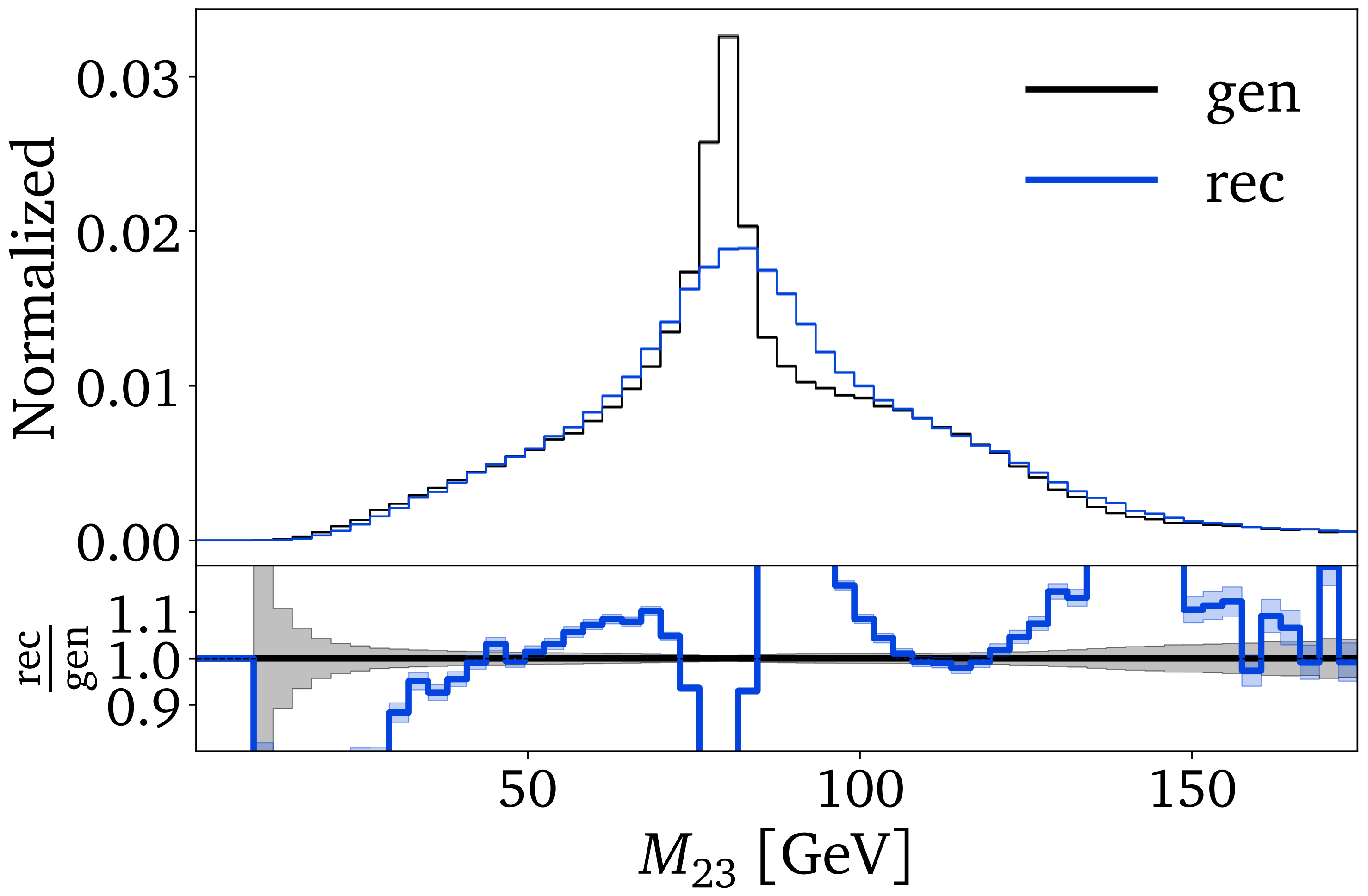
BUT leading uncertainty: choice of m_t in simulation + no access to full phase space

→ **Could generative unfolding help?**

Challenging aspects of top - unfolding

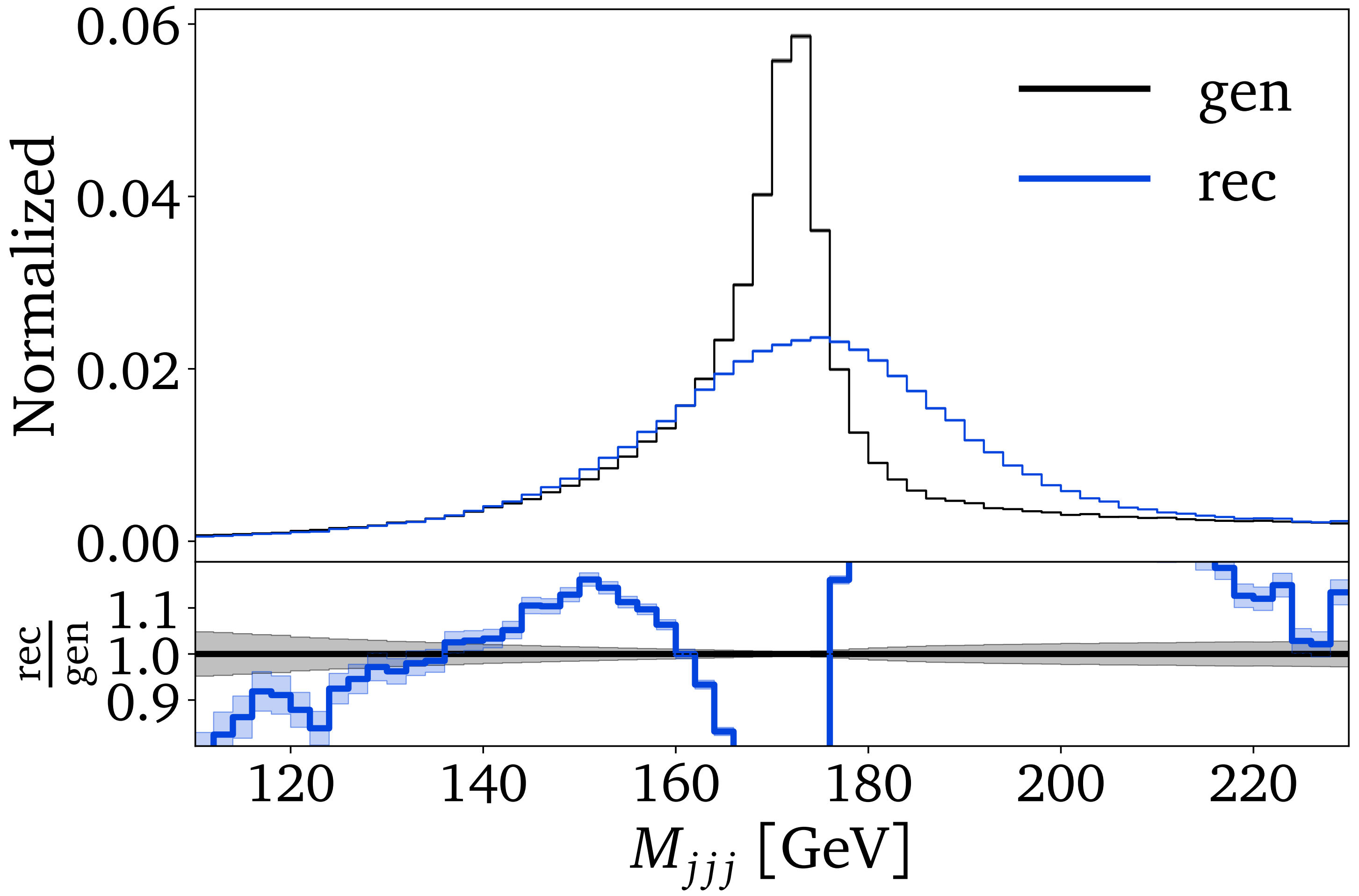
1. Mulitresonant phase space

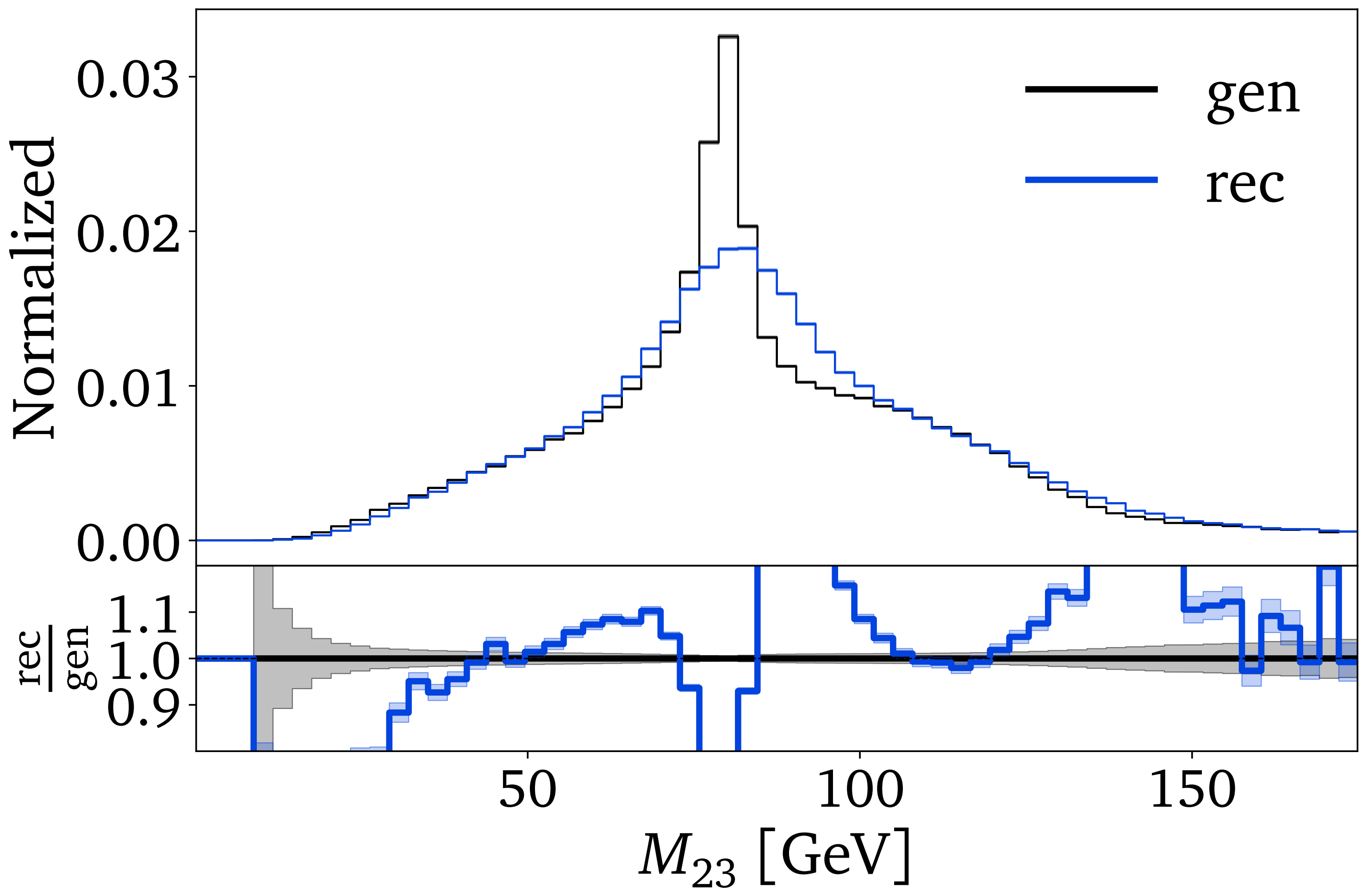




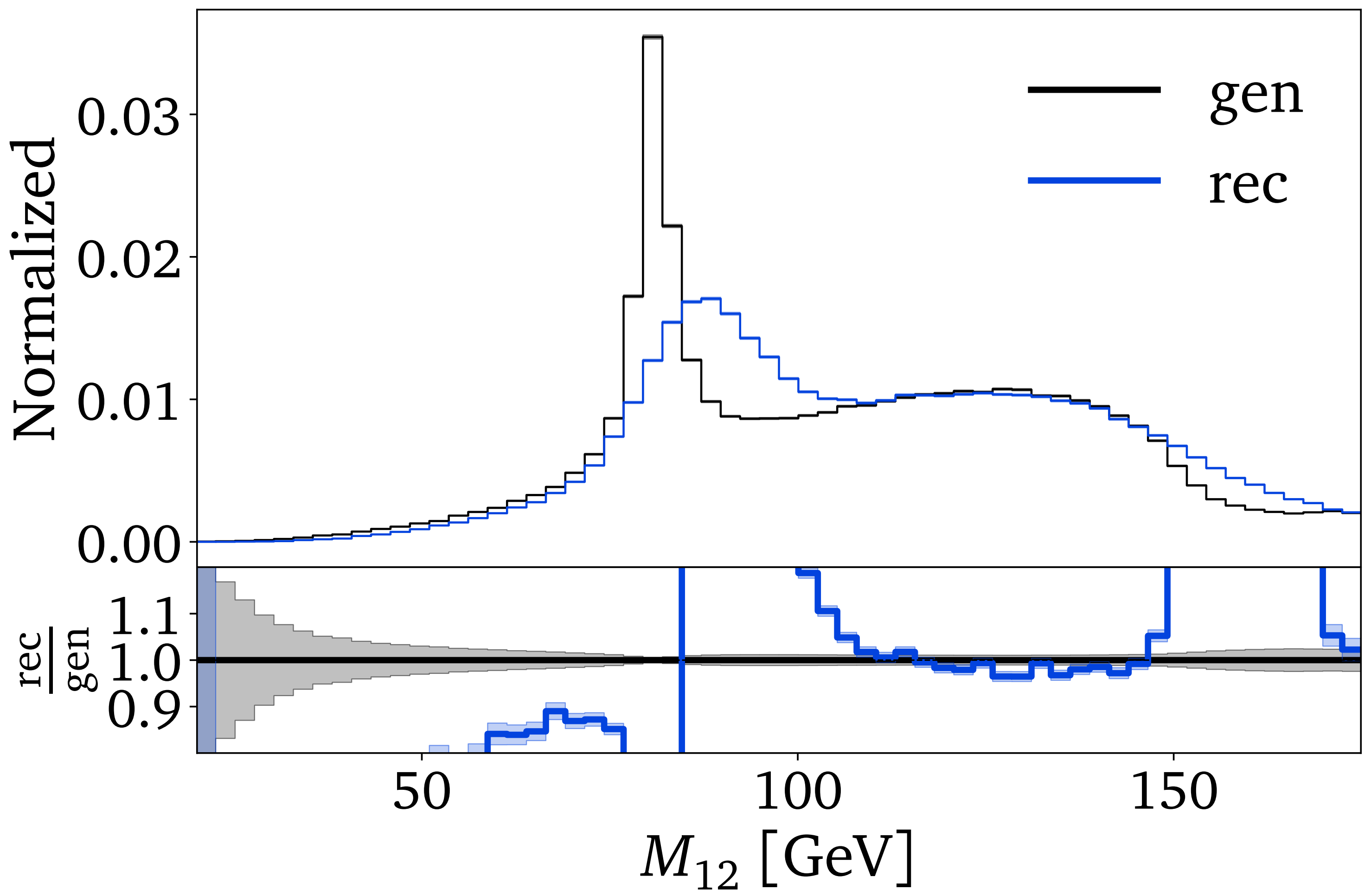
Challenging aspects of top - unfolding

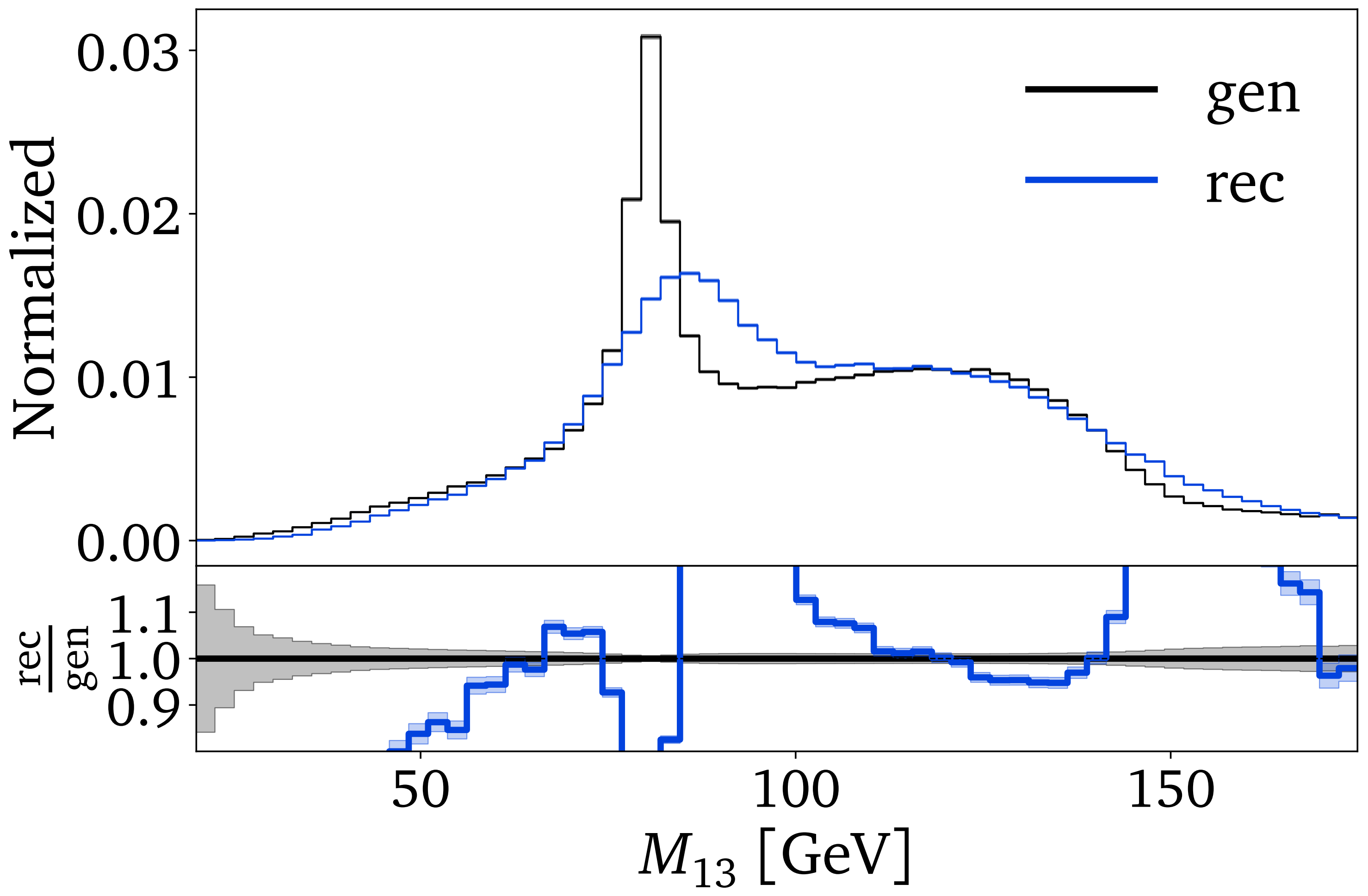
1. Mulitresonant phase space





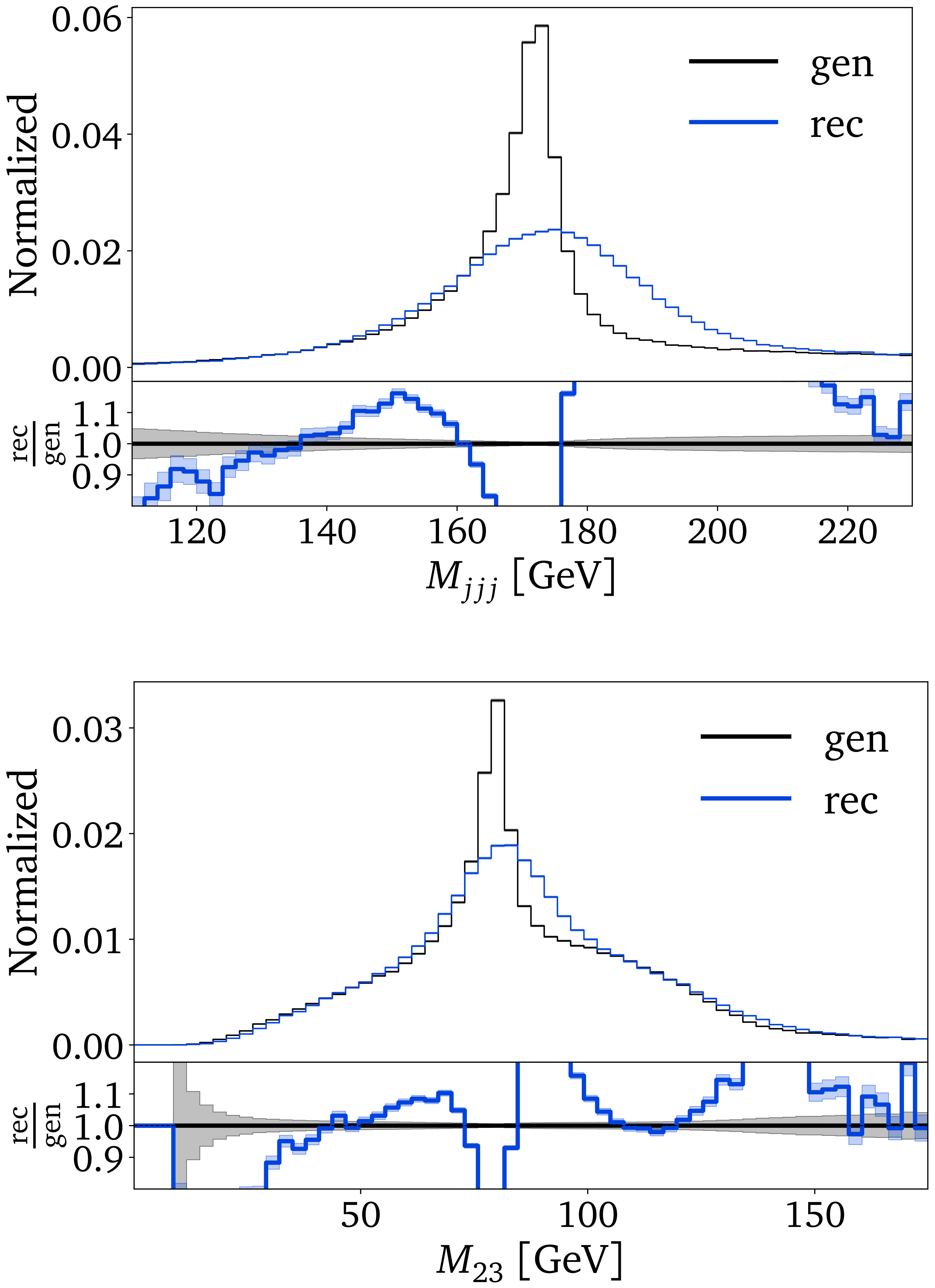
2. Combinatorics



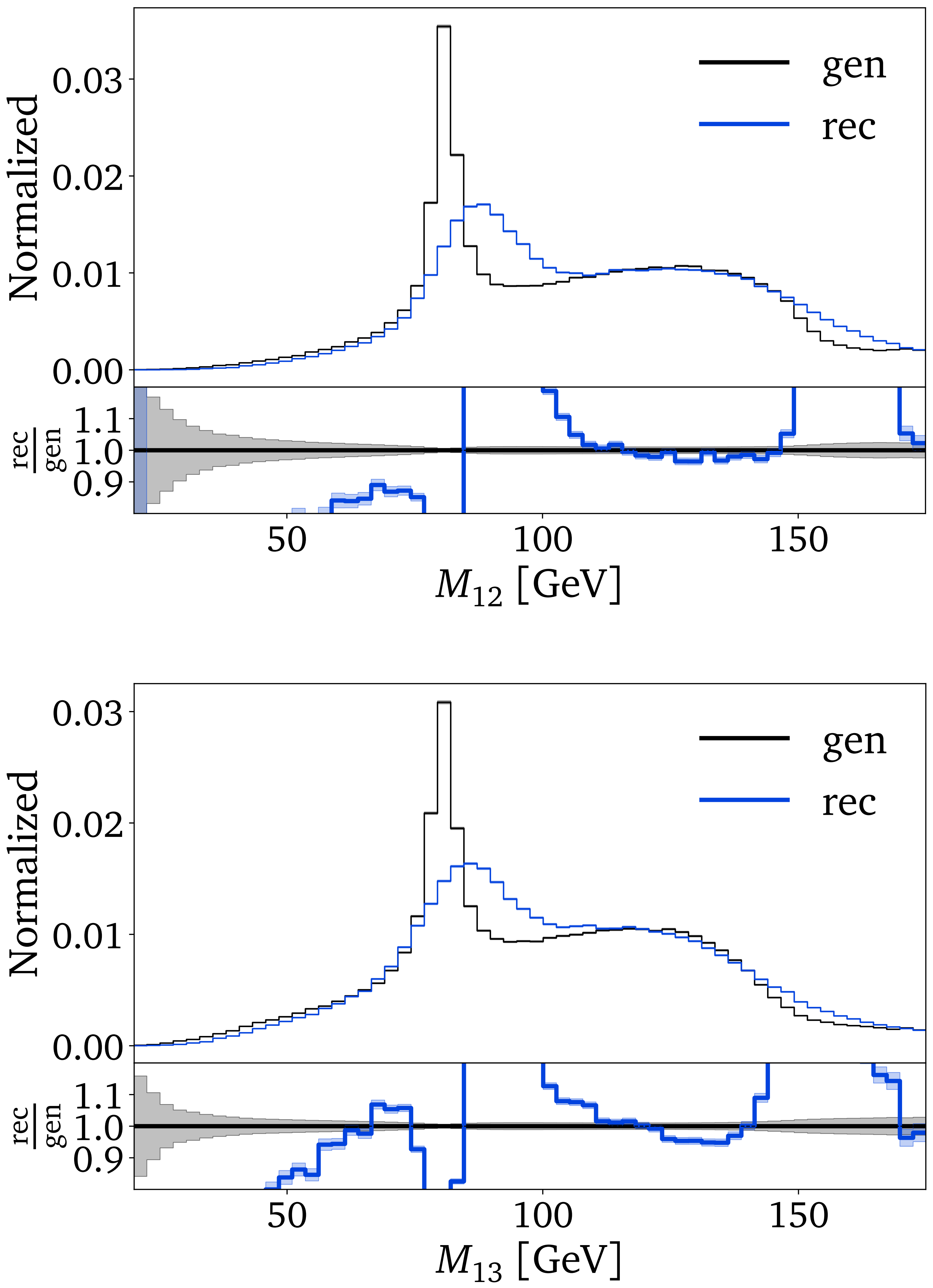


Challenging aspects of top - unfolding

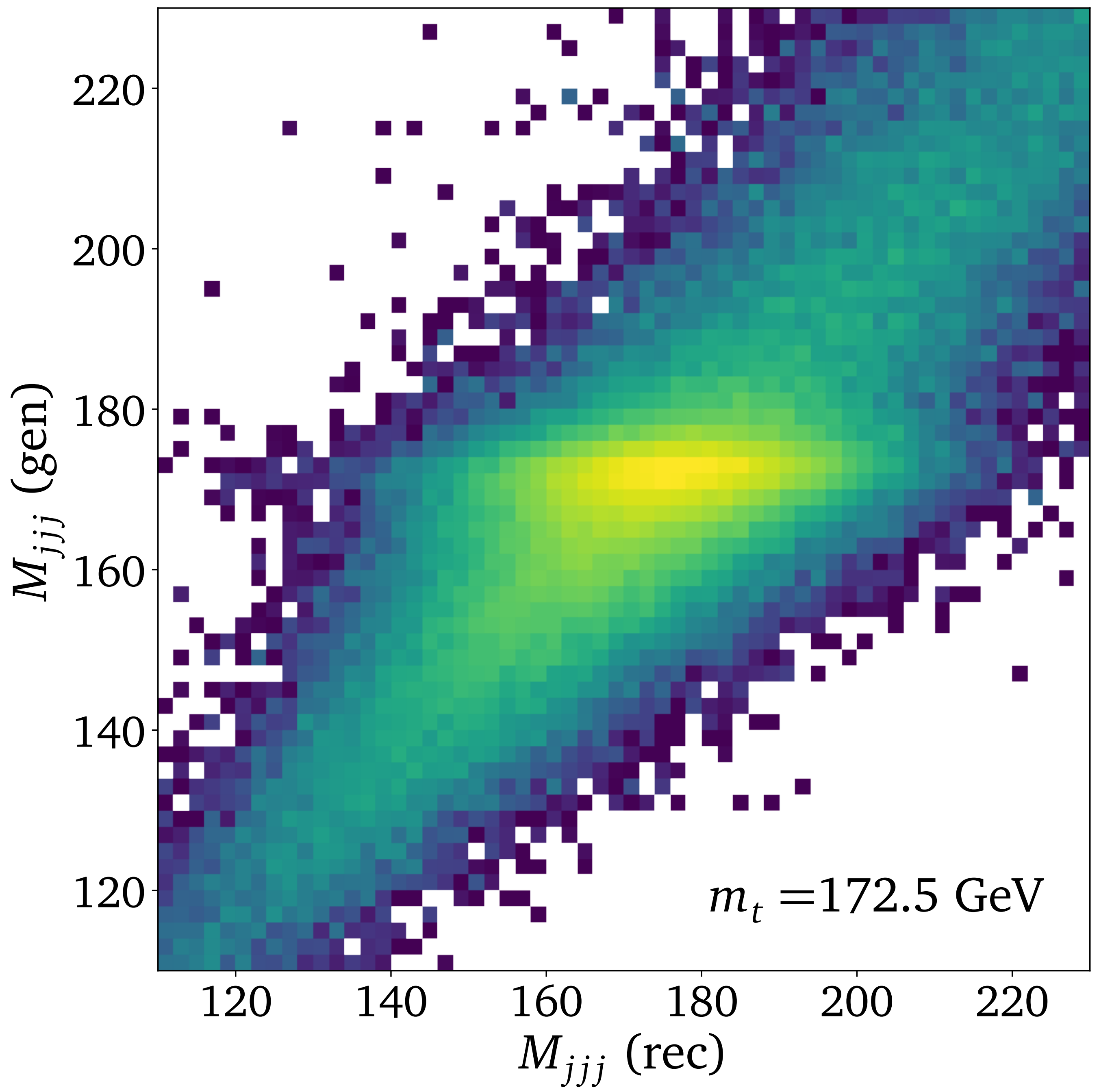
1. Mulitresonant phase space



2. Combinatorics



3. Detector Smearing



Choosing the right parametrization

1. The naive

$$p_1 = (E_1, \vec{p}_1)$$

$$p_2 = (E_2, \vec{p}_2)$$

$$p_3 = (E_3, \vec{p}_3)$$

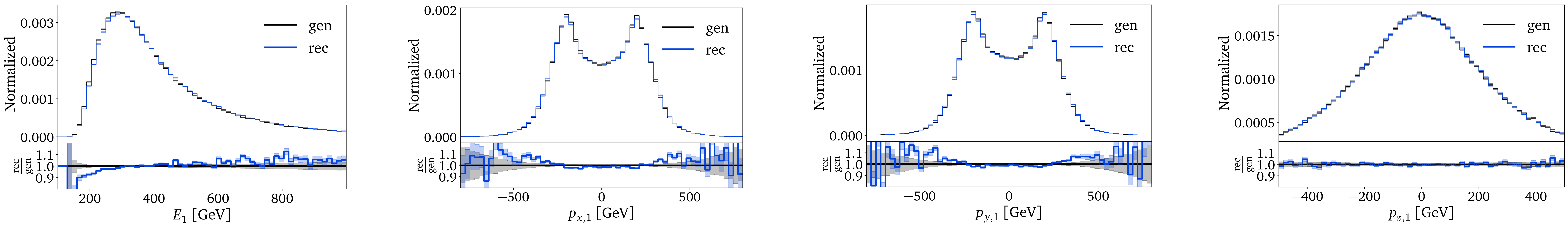
$$M_{jjj}(p_1, p_2, p_3)$$

$$M_{ij}(p_i, p_j)$$

12 dimensional correlation

8 dimensional correlation + combinatorics difficult

Reco and gen level difference not significantly visible, only in correlations



Choosing the right parametrization

2. The less naive

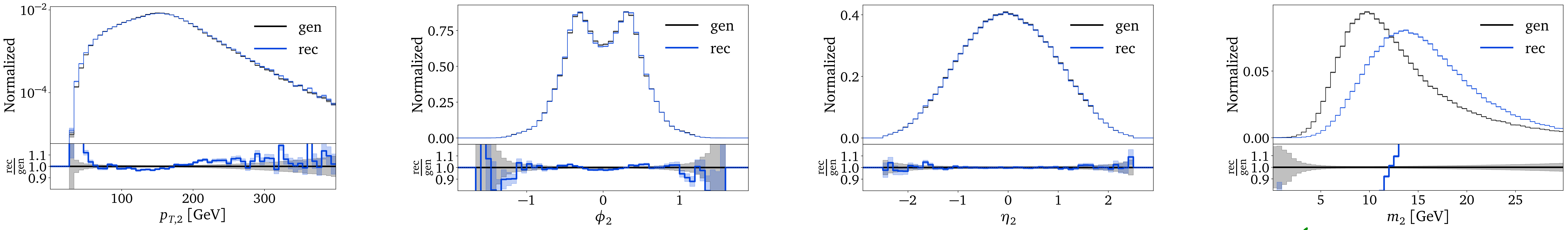
$$p_1 = (p_{T,1}, \phi_1, \eta_1, m_1)$$
$$p_2 = (p_{T,2}, \phi_2, \eta_2, m_2)$$
$$p_3 = (p_{T,3}, \phi_3, \eta_3, m_3)$$

$$M_{jjj}(p_1, p_2, p_3)$$

$$M_{ij}(p_i, p_j)$$

12 dimensional correlation

8 dimensional correlation + combinatorics difficult



Reco and gen level difference visible

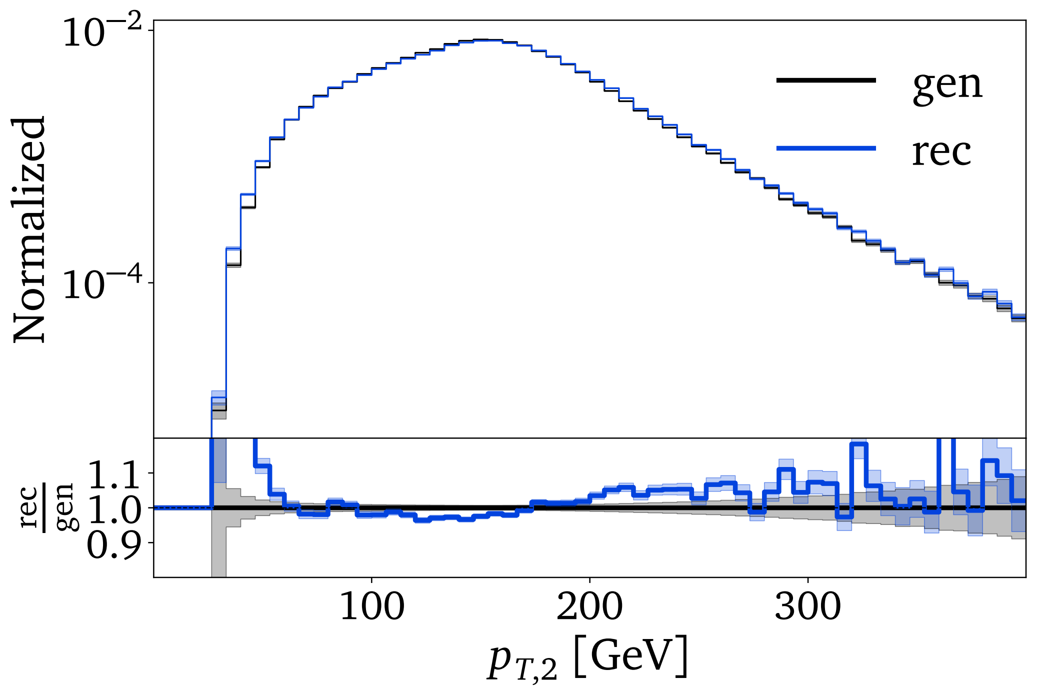
Choosing the right parametrization

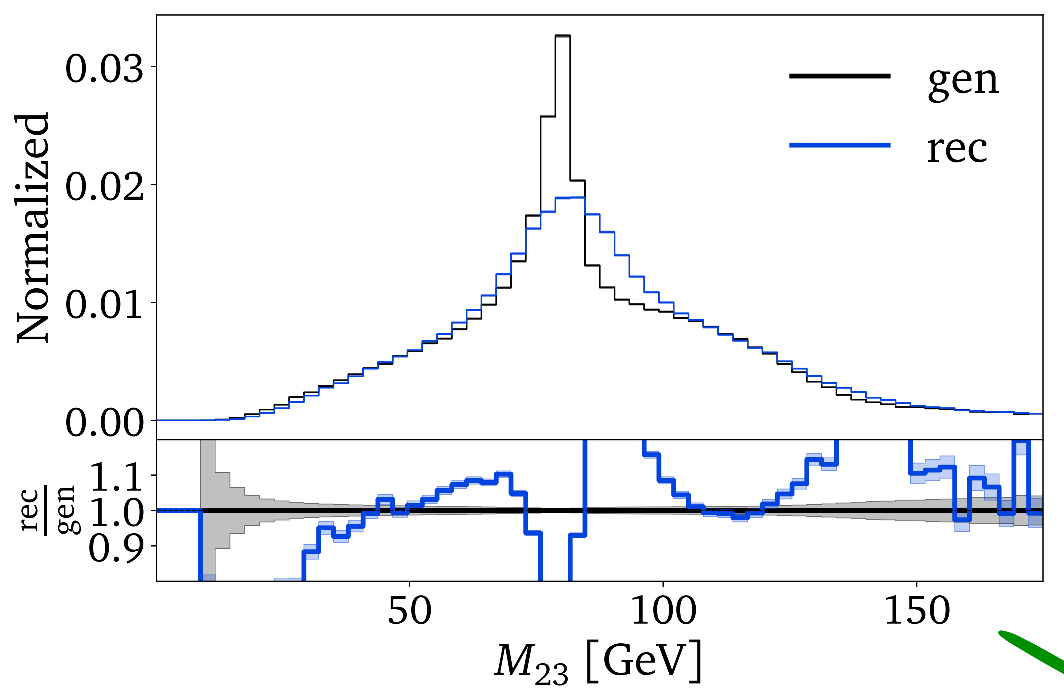
3. The least naive

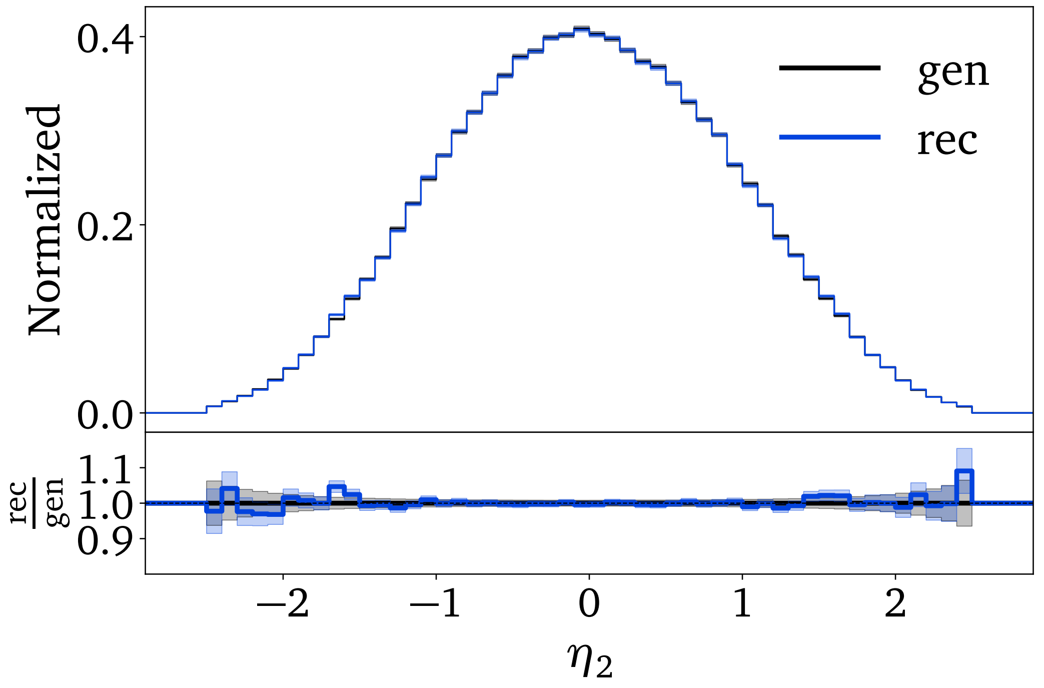
$$\left.\begin{array}{l} p_1 = (p_{T,1}, M_{12}, \eta_1, m_1) \\ p_2 = (p_{T,2}, M_{23}, \eta_2, m_2) \\ p_3 = (p_{T,3}, M_{13}, \eta_3, m_3) \end{array}\right\} \qquad M_{jjj}^2 = \sum_{ij,i>j} M_{ij}^2 - \sum_i m_i^2$$

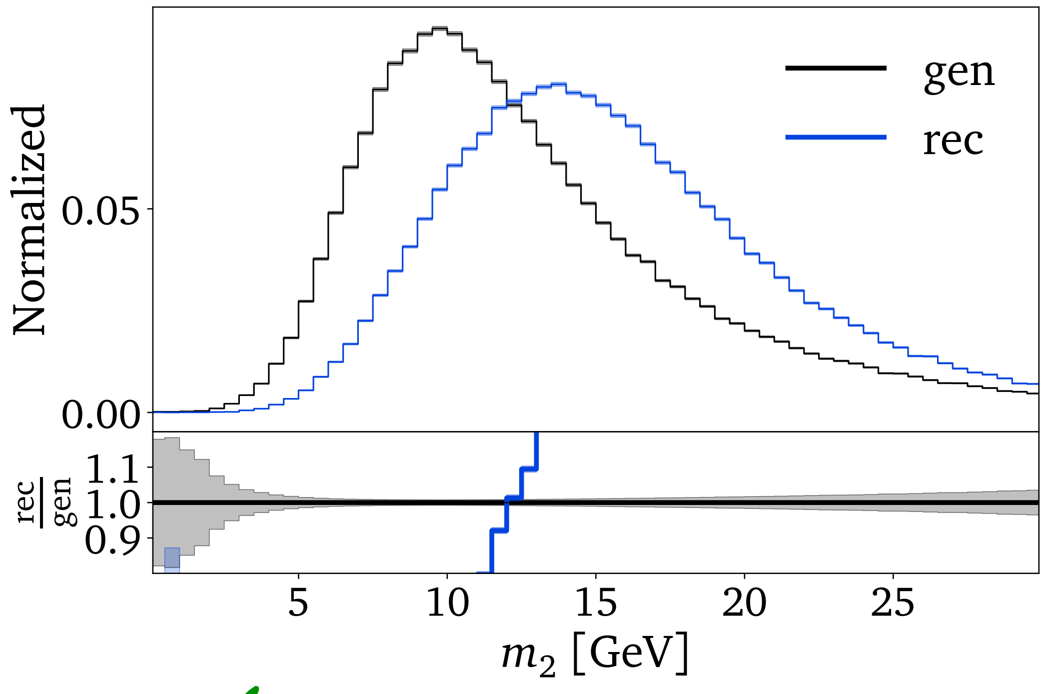
6 dimensional correlation

Direct input + combinatorics simple





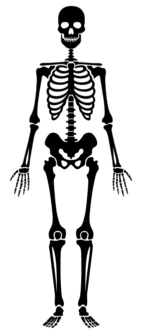




Reco and gen level difference visible

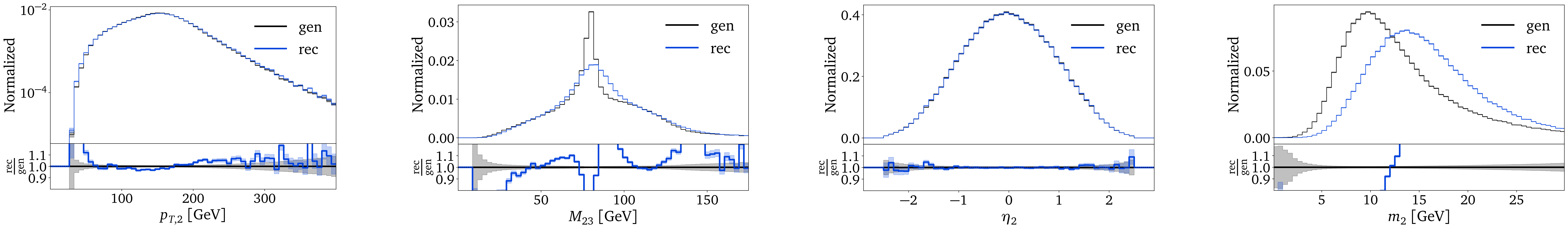
Choosing the right parametrization

3. The least naive



$|\Delta\phi_{ij}|$

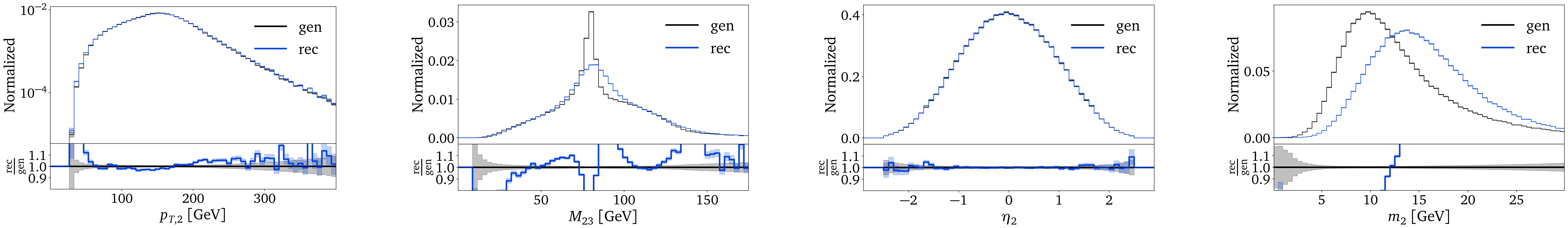
$$\left.\begin{aligned} p_1 &= (p_{T,1}, M_{12}, \eta_1, m_1) \\ p_2 &= (p_{T,2}, M_{23}, \eta_2, m_2) \\ p_3 &= (p_{T,3}, M_{13}, \eta_3, m_3) \end{aligned}\right\} \quad M_{jjj}^2 = \sum_{ij,i>j} M_{ij}^2 - \sum_i m_i^2$$



Choosing the right parametrization

3. The least naive

$$\left.\begin{aligned} p_1 &= (p_{T,1}, M_{12}, \eta_1, m_1) \\ p_2 &= (p_{T,2}, M_{23}, \eta_2, m_2) \\ p_3 &= (p_{T,3}, M_{13}, \eta_3, m_3) \end{aligned}\right\} \qquad M_{jjj}^2 = \sum_{ij,i>j} M_{ij}^2 - \sum_i m_i^2$$



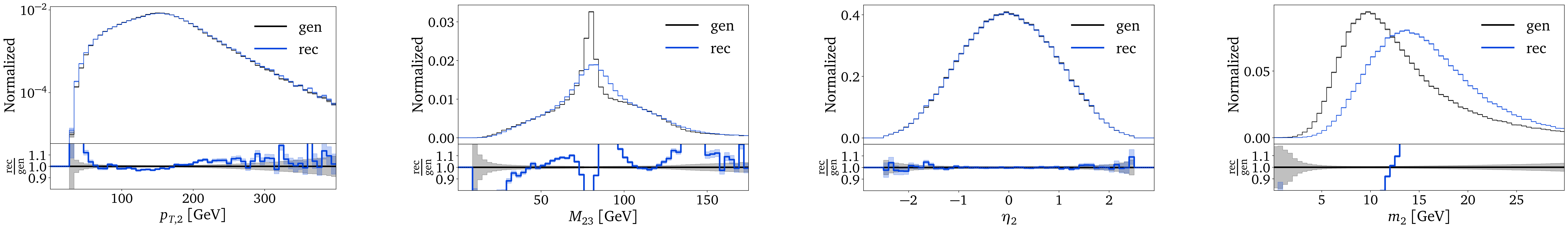
Choosing the right parametrization

3. The least naive

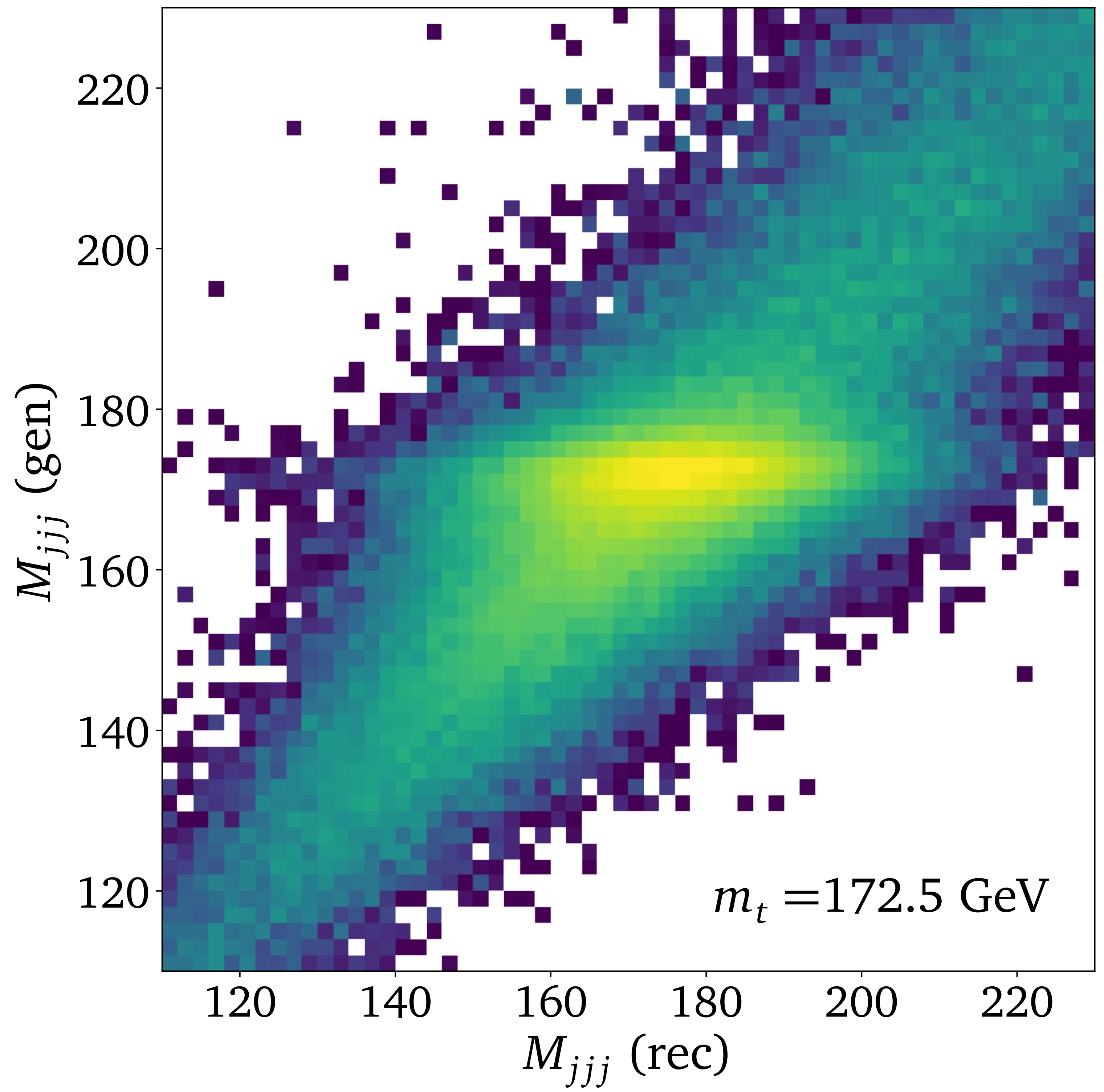
$$p_1 = (p_{T,1}, M_{12}, \eta_1, m_1)$$
$$p_2 = (p_{T,2}, M_{23}, \eta_2, m_2)$$
$$p_3 = (p_{T,3}, M_{13}, \eta_3, m_3)$$

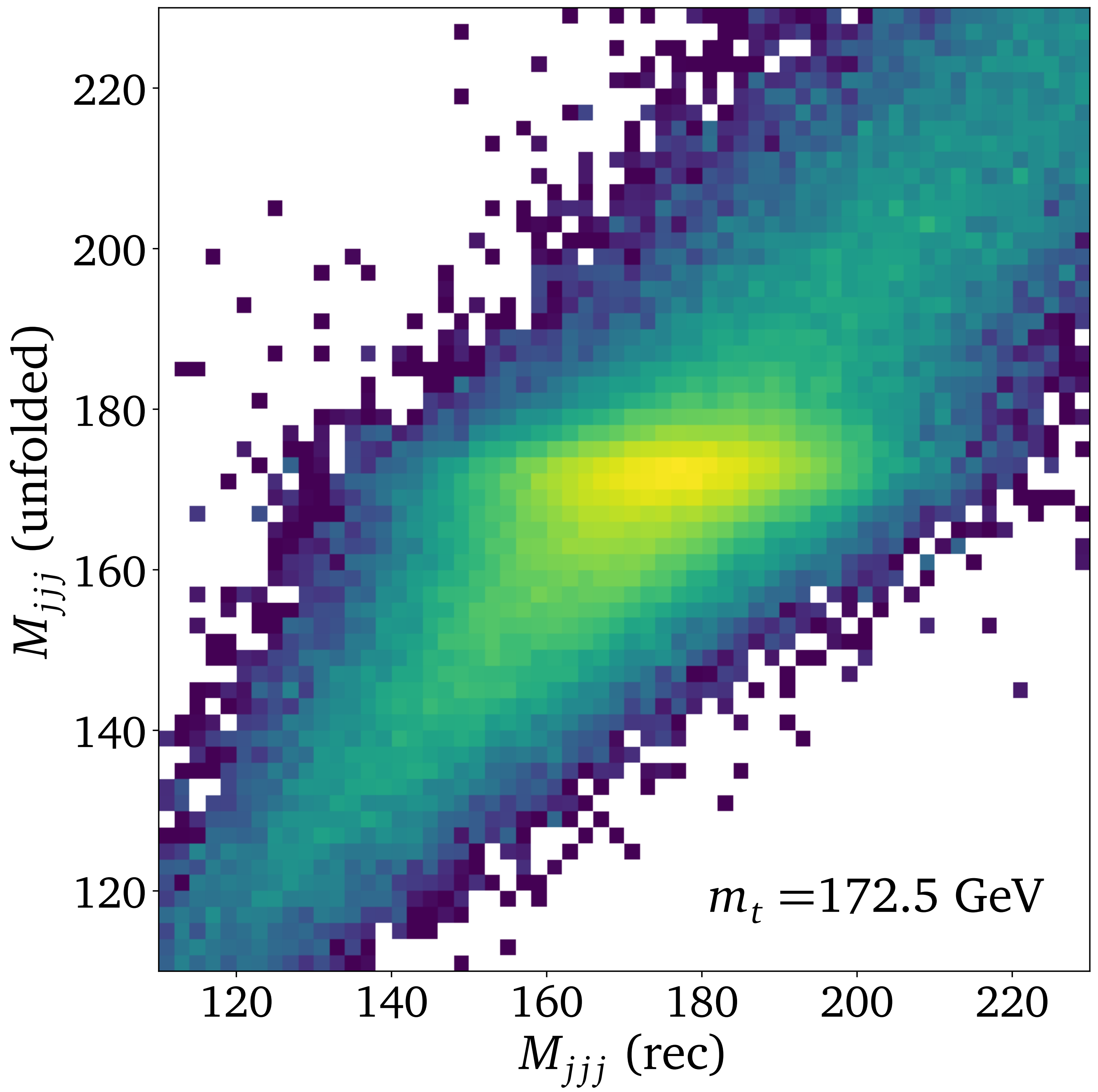
$$M_{jjj}^2 = \sum_{ij,i>j} M_{ij}^2 - \sum_i m_i^2$$

For mass measurement, we only use 6 dimensional subset of phase space to increase network performance



Model-Dependence





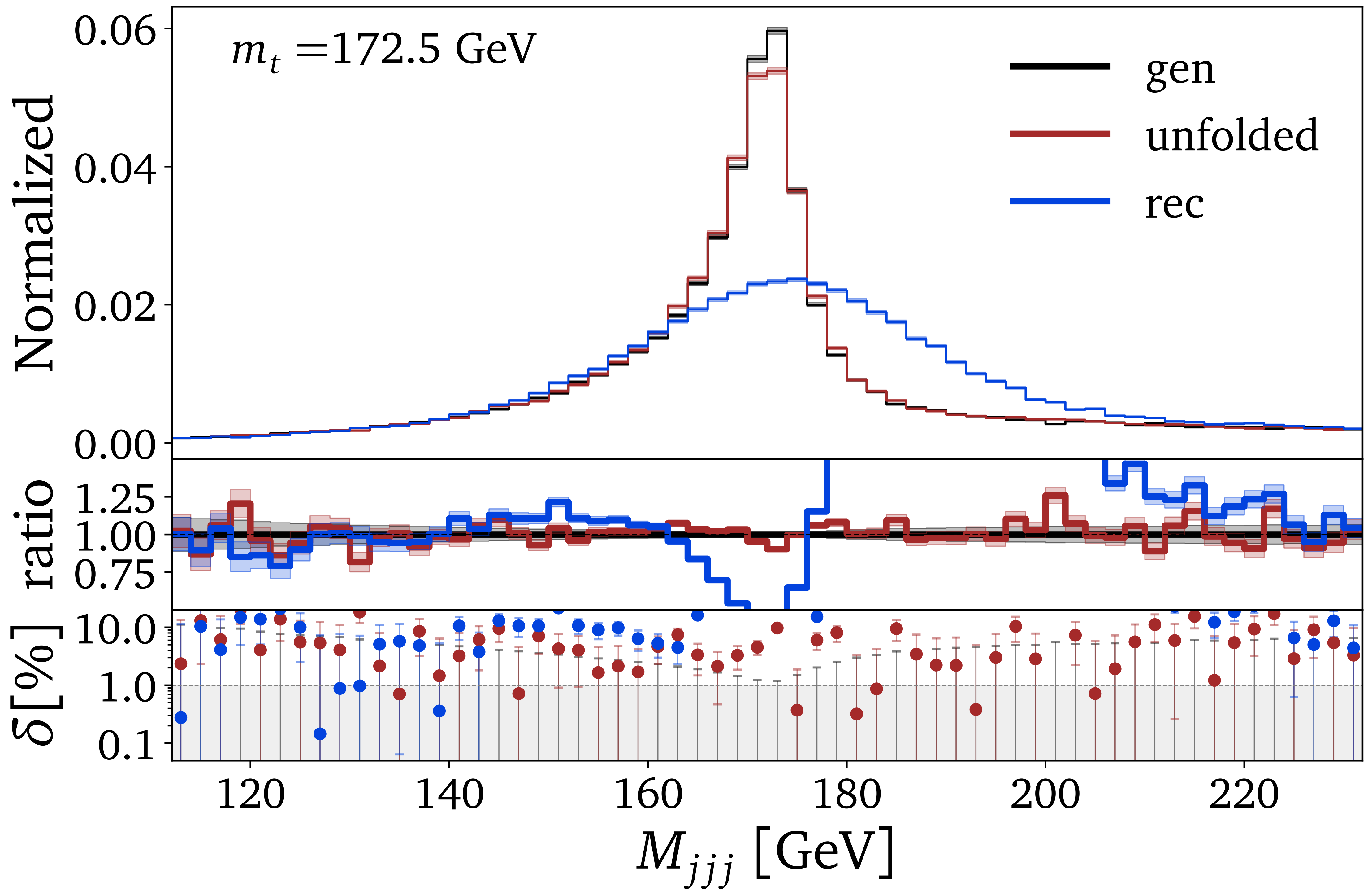
Correct migration learned?

Model-Dependence?

Train with full CMS simulation with
$m_t = 172.5$ GeV

Unfolded distribution of triple jet mass within
$\mathcal{O}(1\%)$ of truth gen level

BUT: Test data also simulation with
$m_t = 172.5$ GeV



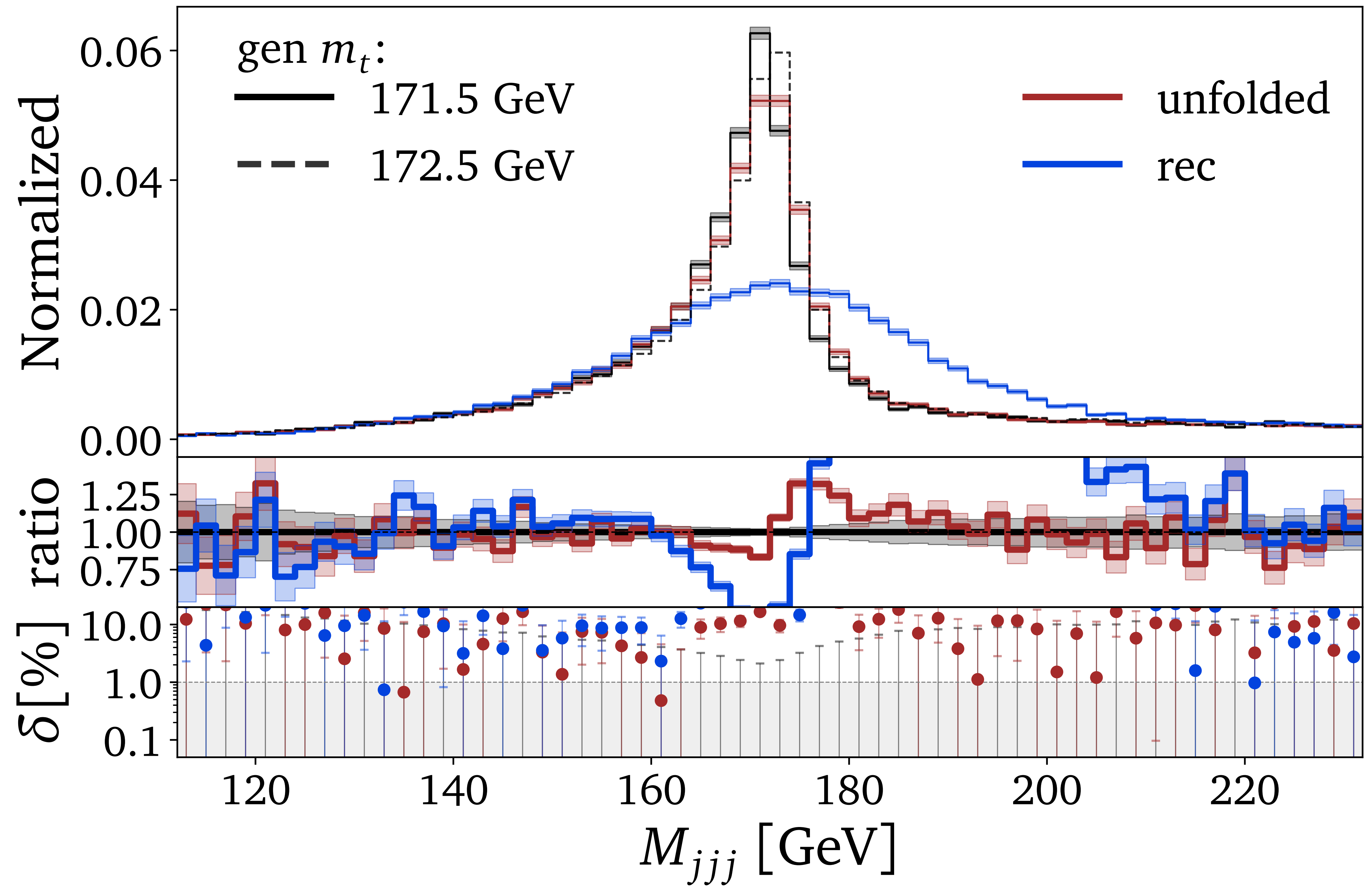
Model-Dependence!

Train with full CMS simulation with
$m_t = 172.5$ GeV

Unfolded distribution of triple jet mass within
$\mathcal{O}(1\%)$ of truth gen level

BUT: Test data also simulation with
$m_t = 172.5$ GeV

For pseudo-data with different top masses :
Algorithm falls back to prior ($m_t = 172.5$
GeV)



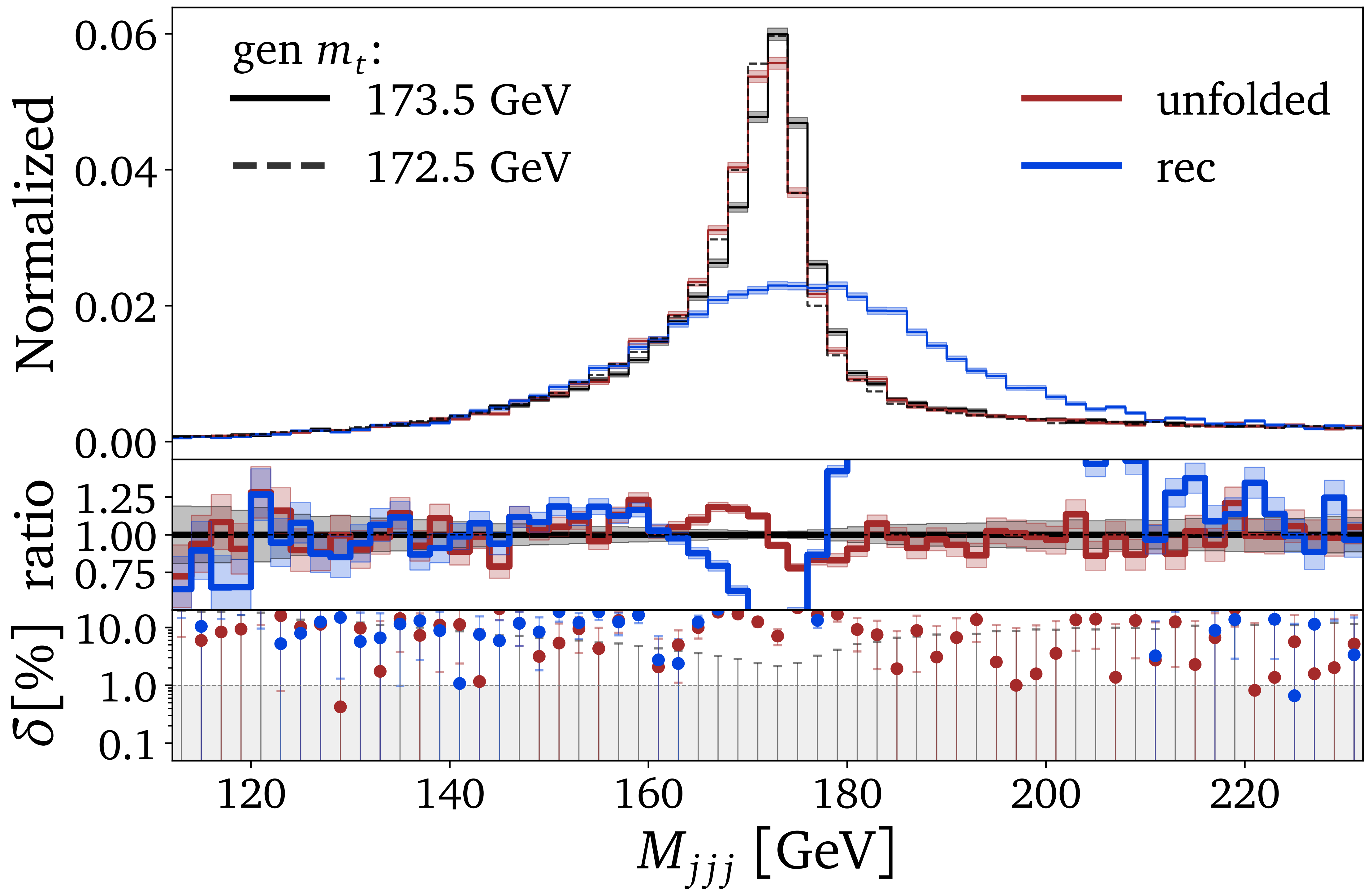
Model-Dependence!

Train with full CMS simulation with
$m_t = 172.5$ GeV

Unfolded distribution of triple jet mass within
$\mathcal{O}(1\%)$ of truth gen level

BUT: Test data also simulation with
$m_t = 172.5$ GeV

For pseudo-data with different top masses :
Algorithm falls back to prior ($m_t = 172.5$
GeV)



Removing Model-Dependence

$p_{sim}(x_{gen} | m_s)$

$p(x_{reco} | x_{gen})$ ↓

$p_{sim}(x_{reco} | m_s)$ ⟷ correspondance ⟷ $p_{data}(x_{reco} | m_d)$

$p_{model}(x_{gen} | x_{reco}, m_s)$ ↑

$p_{unfold}(x_{gen} | m_s, m_d)$

Removing Model-Dependence

$p_{sim}(x_{gen} | m_s)$

$p(x_{reco} | x_{gen})$

$p_{sim}(x_{reco} | m_s)$

correspondance

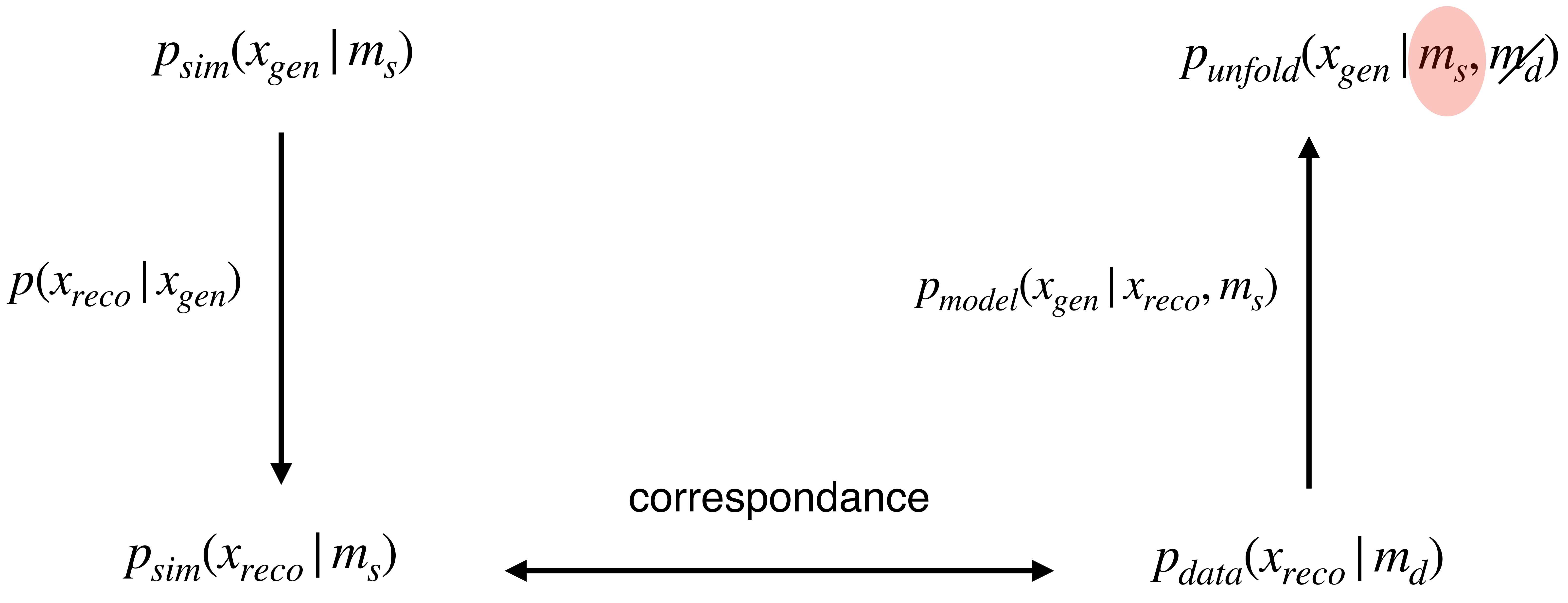
$p_{data}(x_{reco} | m_d)$

$p_{model}(x_{gen} | x_{reco}, m_s)$

$p_{unfold}(x_{gen} | m_s, \cancel{m_d})$

→ **Solution: Strengthen m_d dependence, but how?**

Removing Model-Dependence



→ **Solution: Strengthen m_d dependence, but how?**

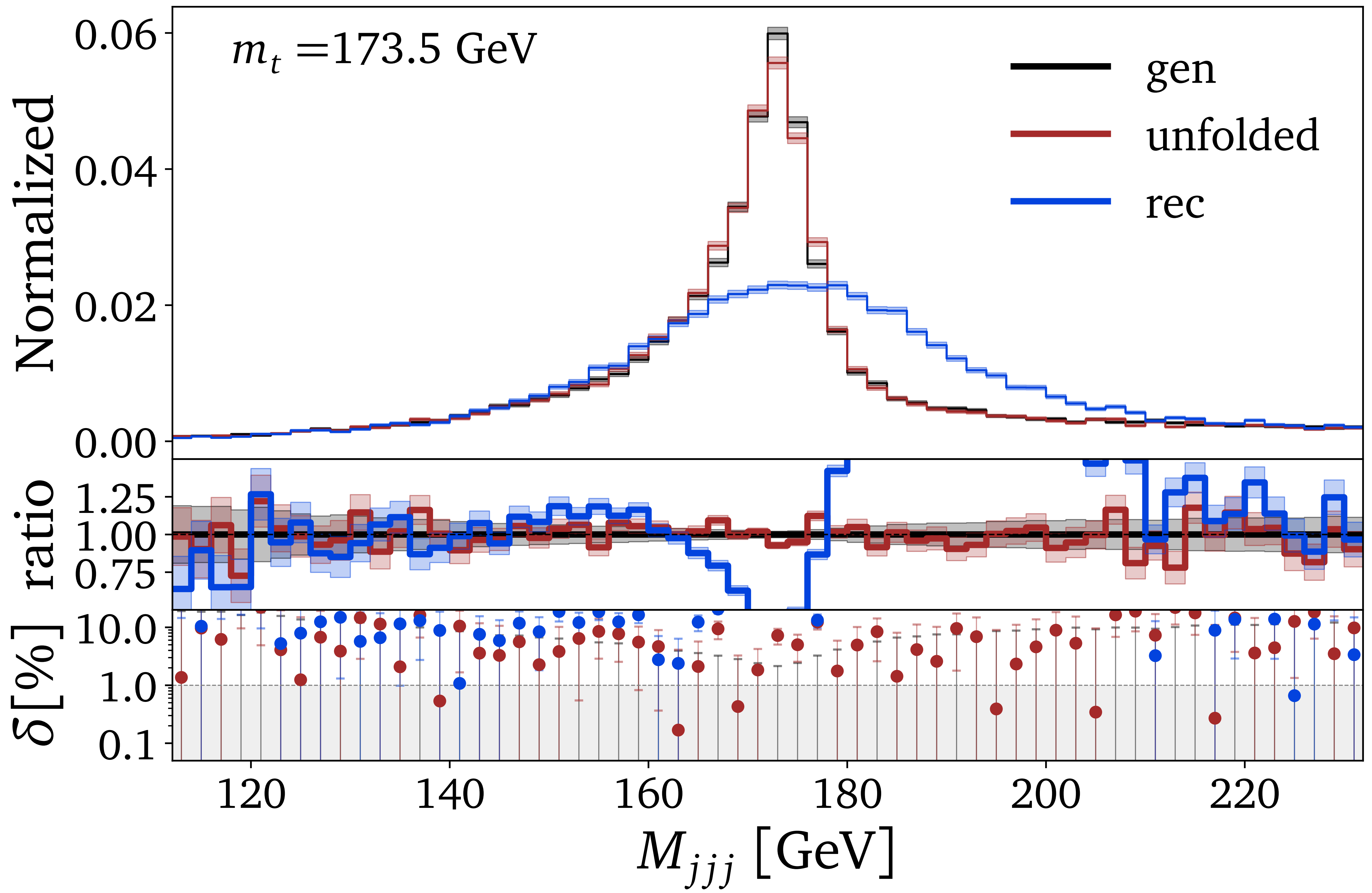
1. Augment training data with simulation from different top masses
2. Estimate batch-wise $m_d \approx$ weighted-median(M_{jjj}^{batch}) on reco level

Removing Model-Dependence!

Train with full CMS simulation with
$m_t = [172.5$ GeV, 169.5 GeV, 175.5 GeV$]$

Test by unfolding simulation with
$m_t = 171.5$ GeV & 173.5 GeV

Unfolded distribution of triple jet mass within
$\mathcal{O}(1\%)$ of truth gen level **without** bias

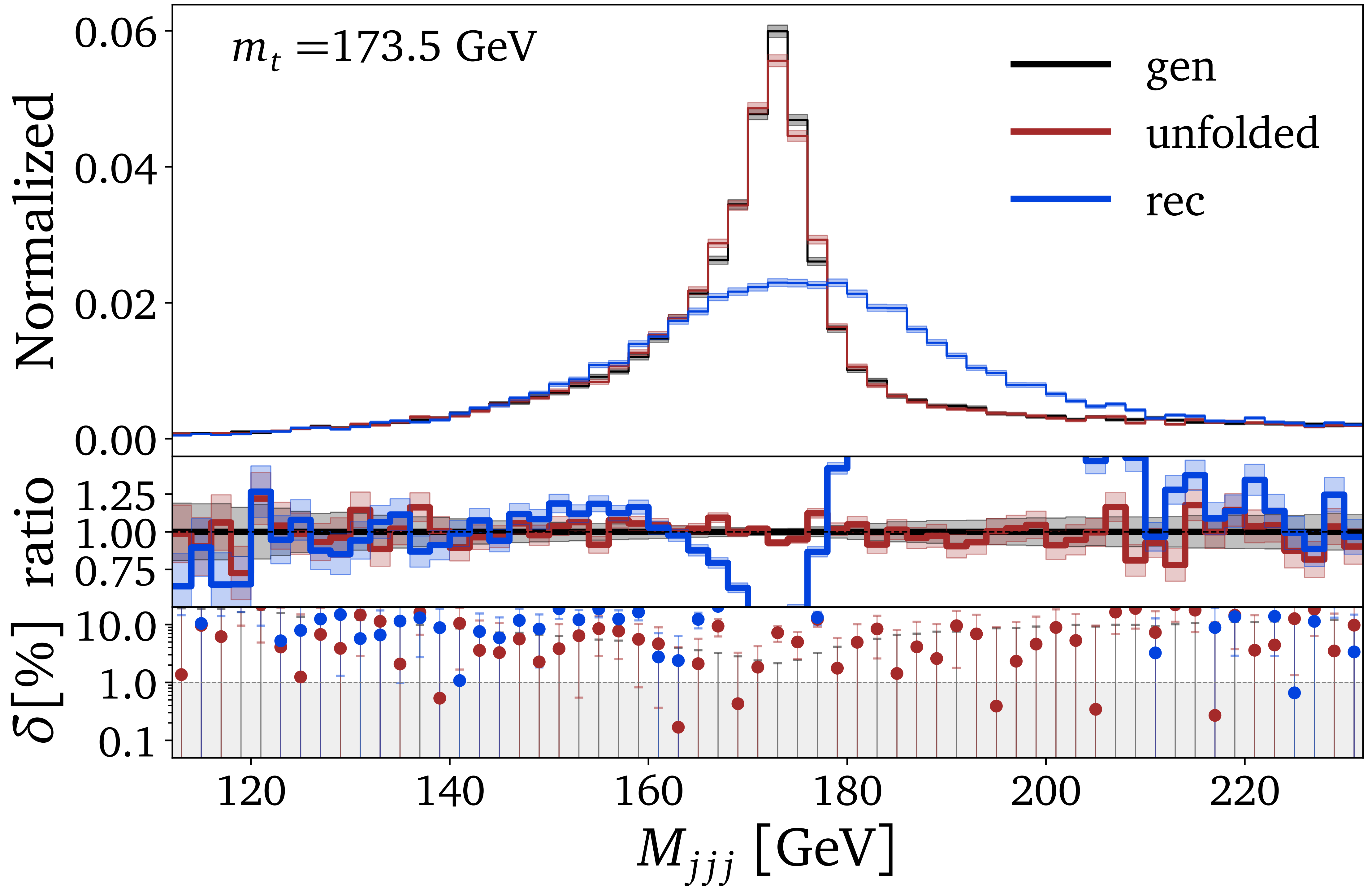


Removing Model-Dependence!

Train with full CMS simulation with
$m_t = [172.5$ GeV, 169.5 GeV, 175.5 GeV$]$

Test by unfolding simulation with
$m_t = 171.5$ GeV & 173.5 GeV

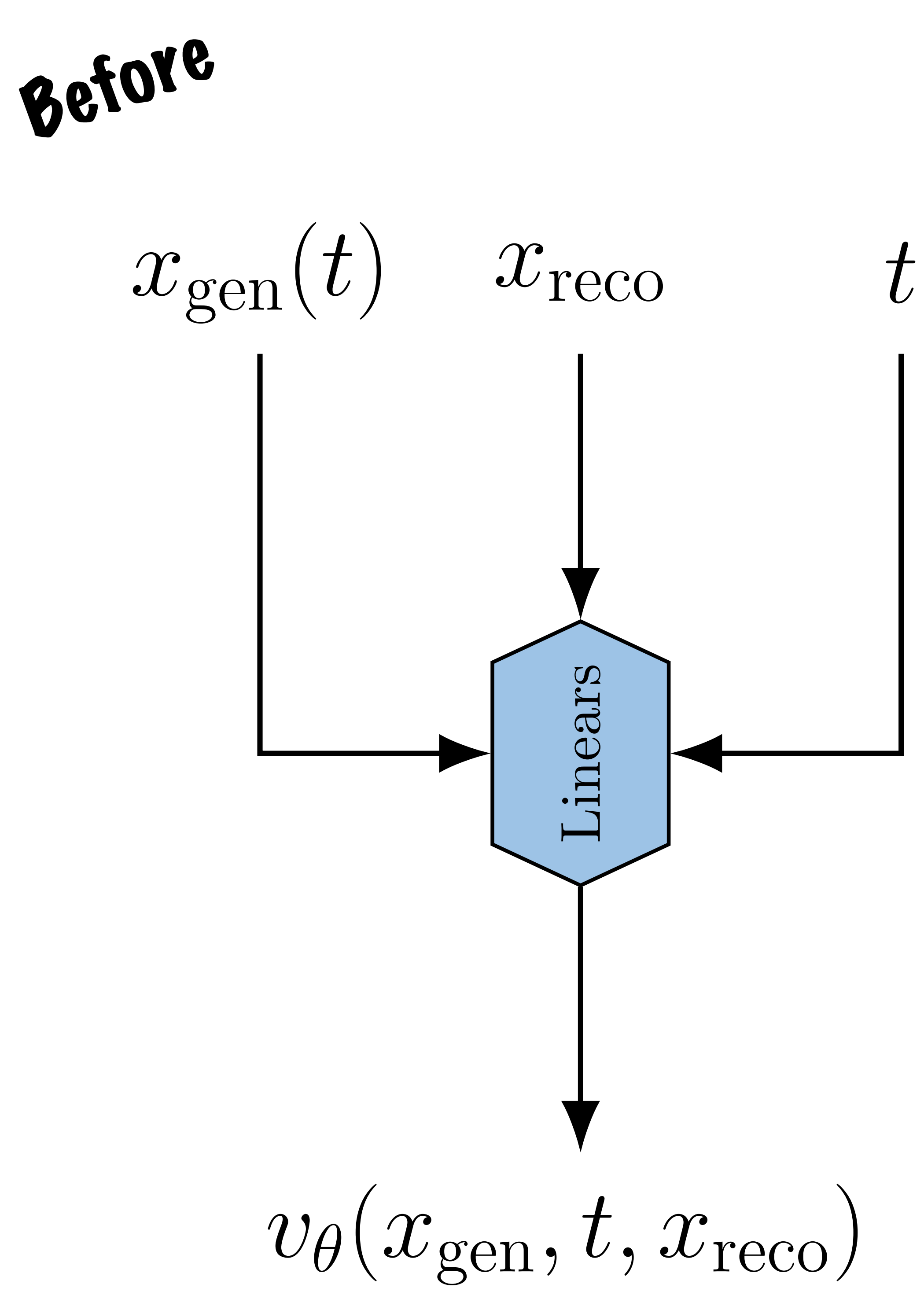
Unfolded distribution of triple jet mass within $\mathcal{O}(1\%)$ of truth gen level **without** bias



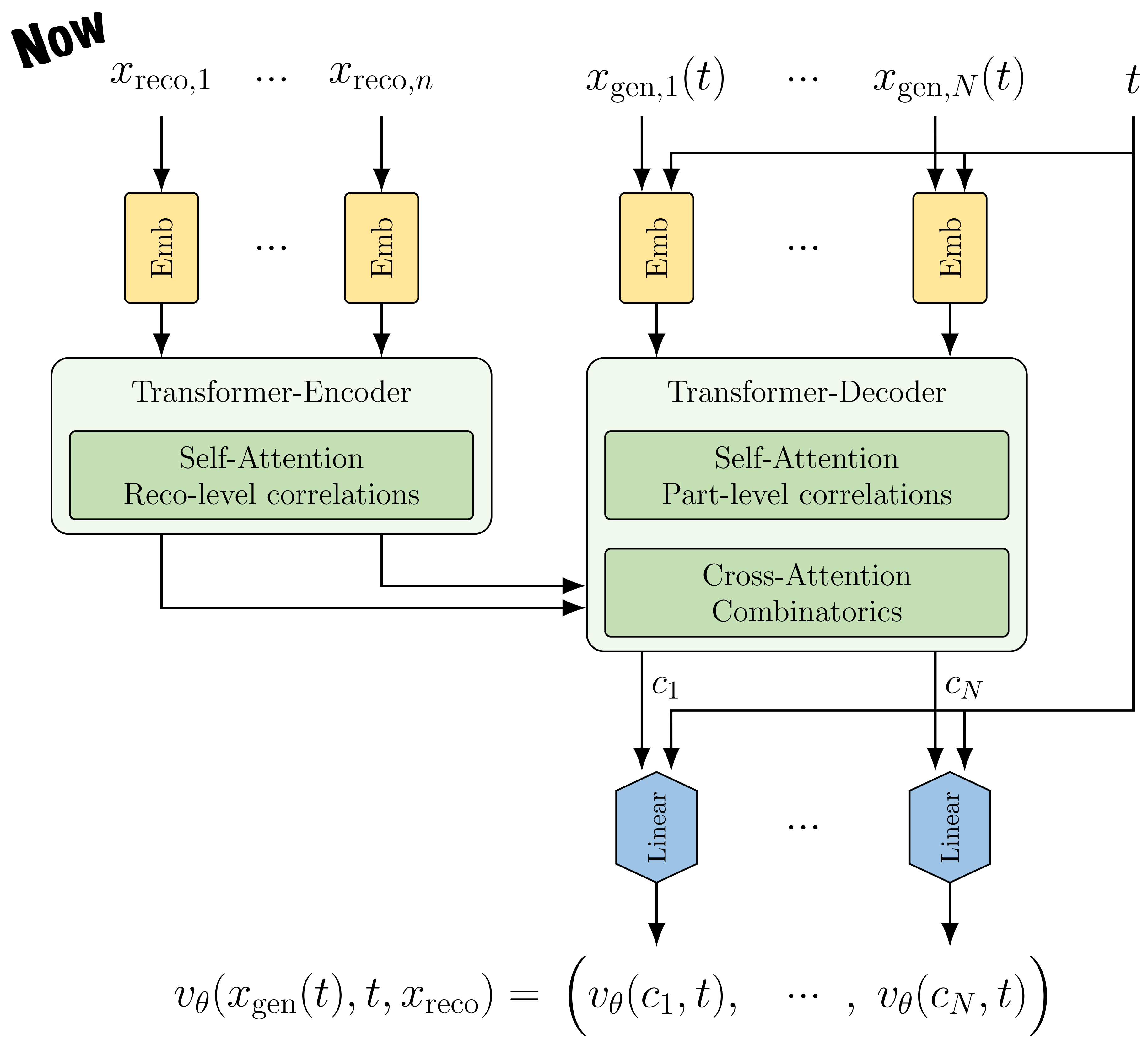
ML task becomes much harder

Removing Model-Dependence!

ML task becomes much harder



vs



Adapted from **arXiv:2310.07752**

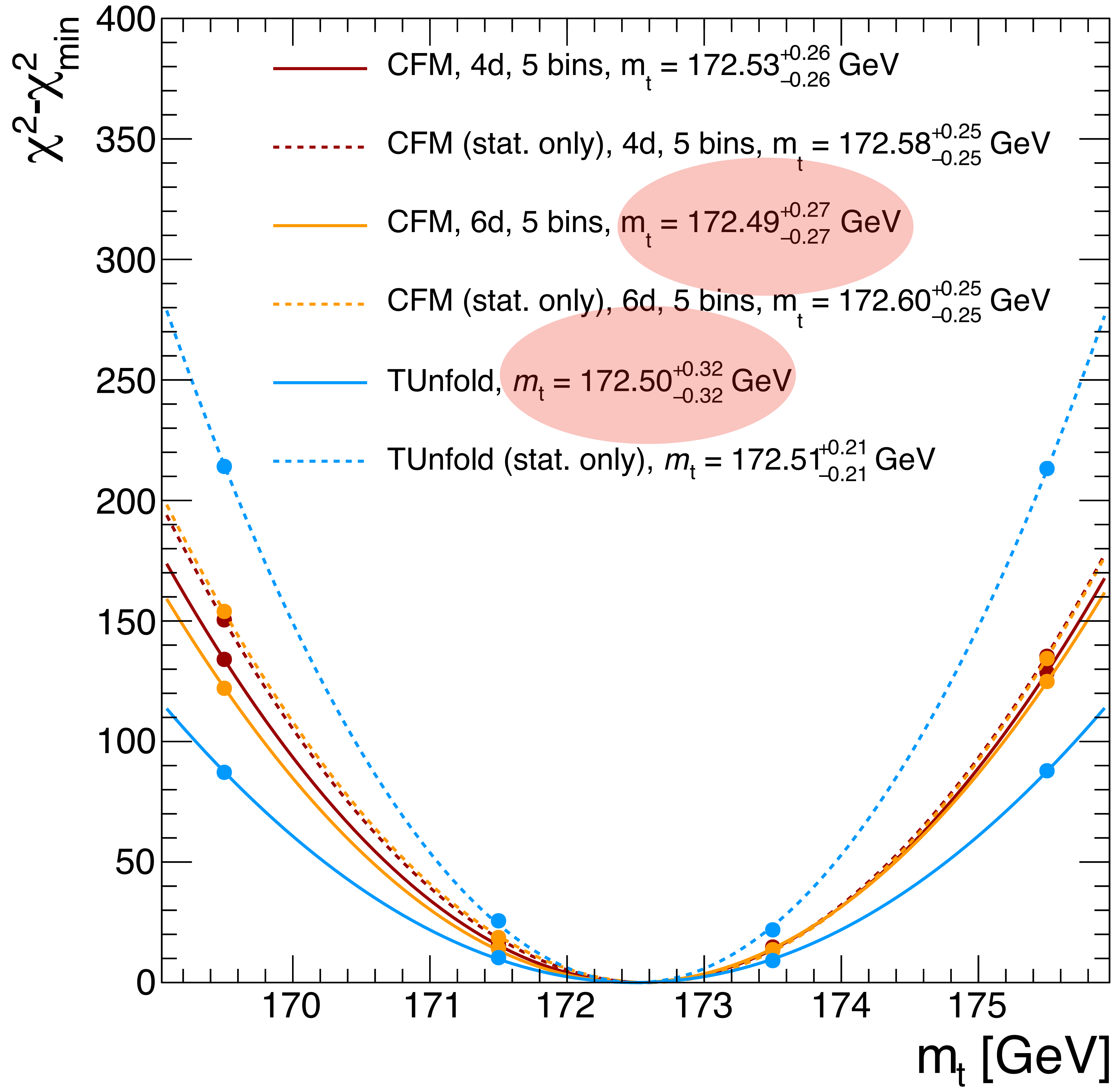
Mass Measurement

For a fixed top mass:

Choose subset of test data of 41000 reco level events

Unfolded 1000 bootstrapped replicas

Estimate covariance matrix and mean by 1000 different unfolded distributions



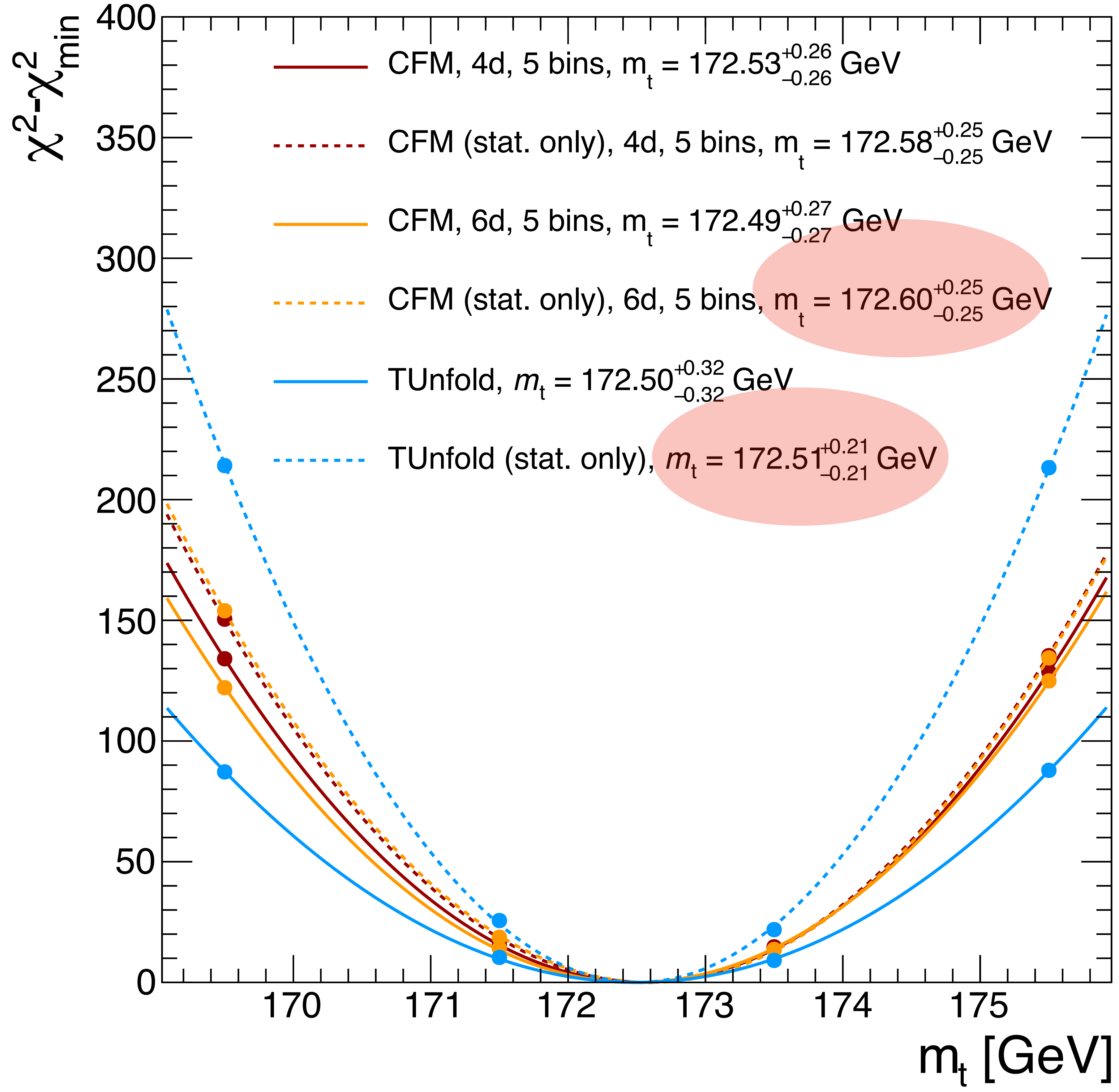
Mass Measurement

For a fixed top mass:

Choose subset of test data of 41000 reco level events

Unfolded 1000 bootstrapped replicas

Estimate covariance matrix and mean by 1000 different unfolded distributions



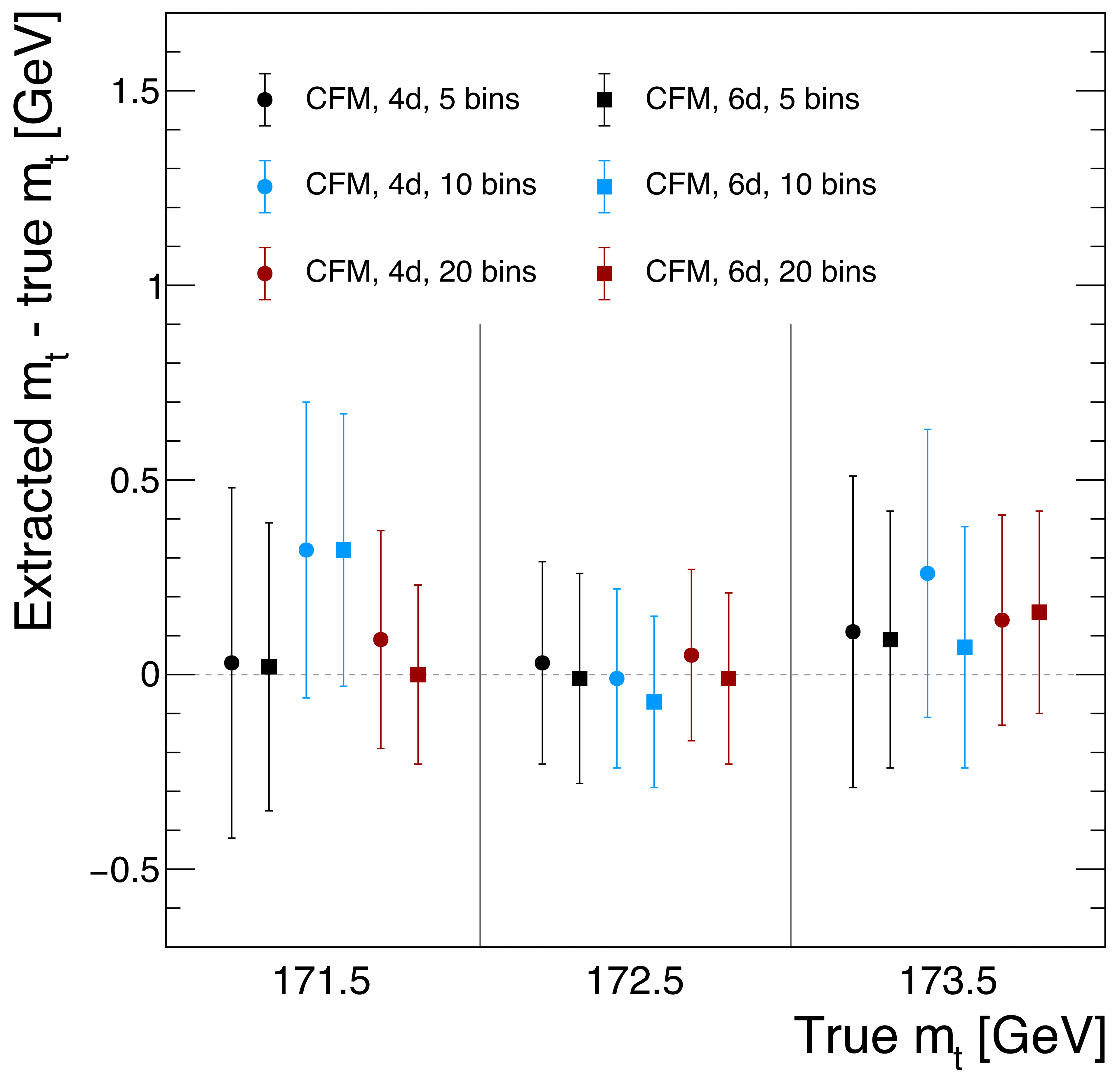
Mass Measurement

For a fixed top mass:

Choose subset of test data of 41000 reco level events

Unfolded 1000 bootstrapped replicas

Estimate covariance matrix and mean by 1000 different unfolded distributions



→ Reliably unfold triple jet mass without bias

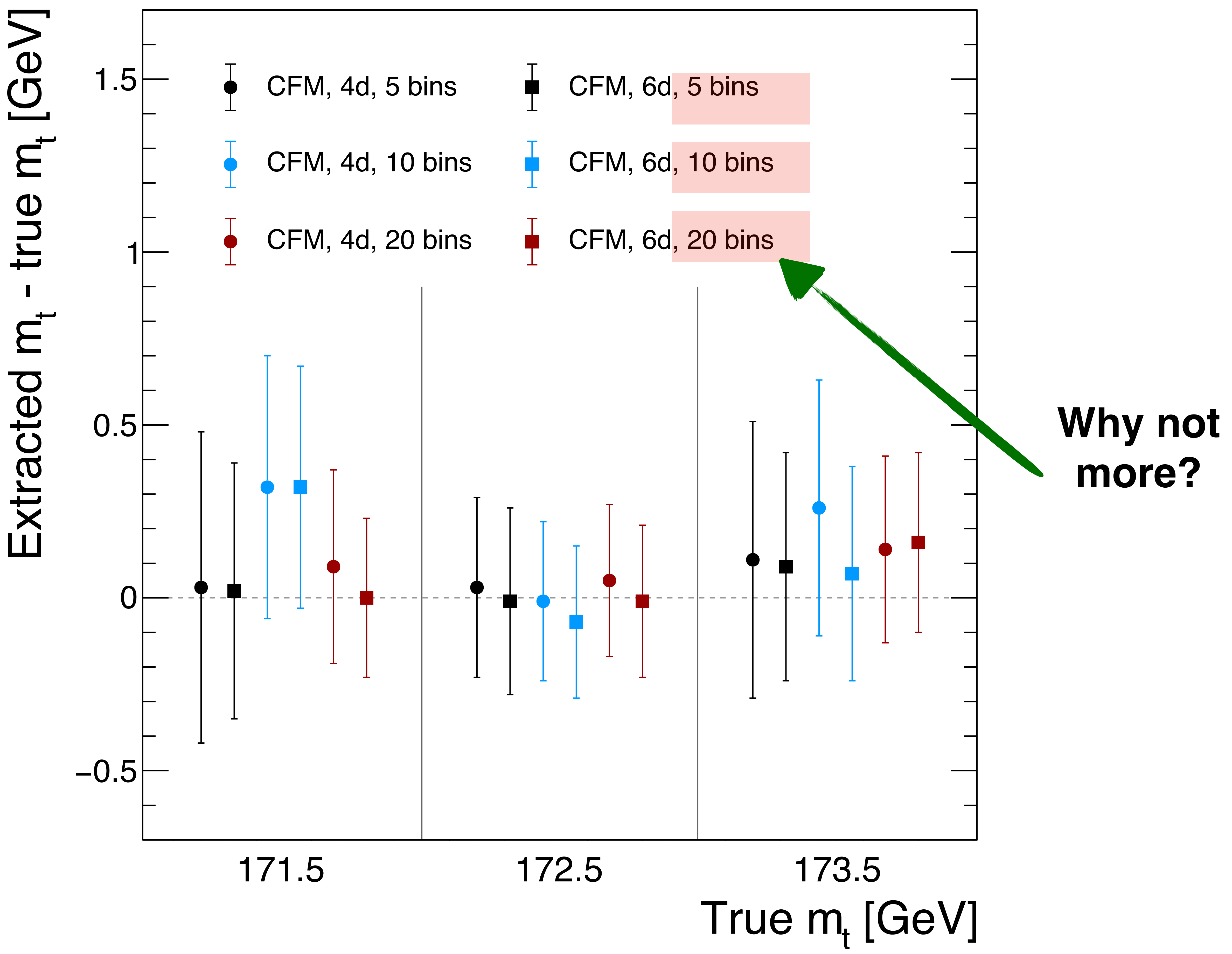
Mass Measurement

For a fixed top mass:

Choose subset of test data of 41000 reco level events

Unfolded 1000 bootstrapped replicas

Estimate covariance matrix and mean by 1000 different unfolded distributions



→ Reliably unfold triple jet mass without bias

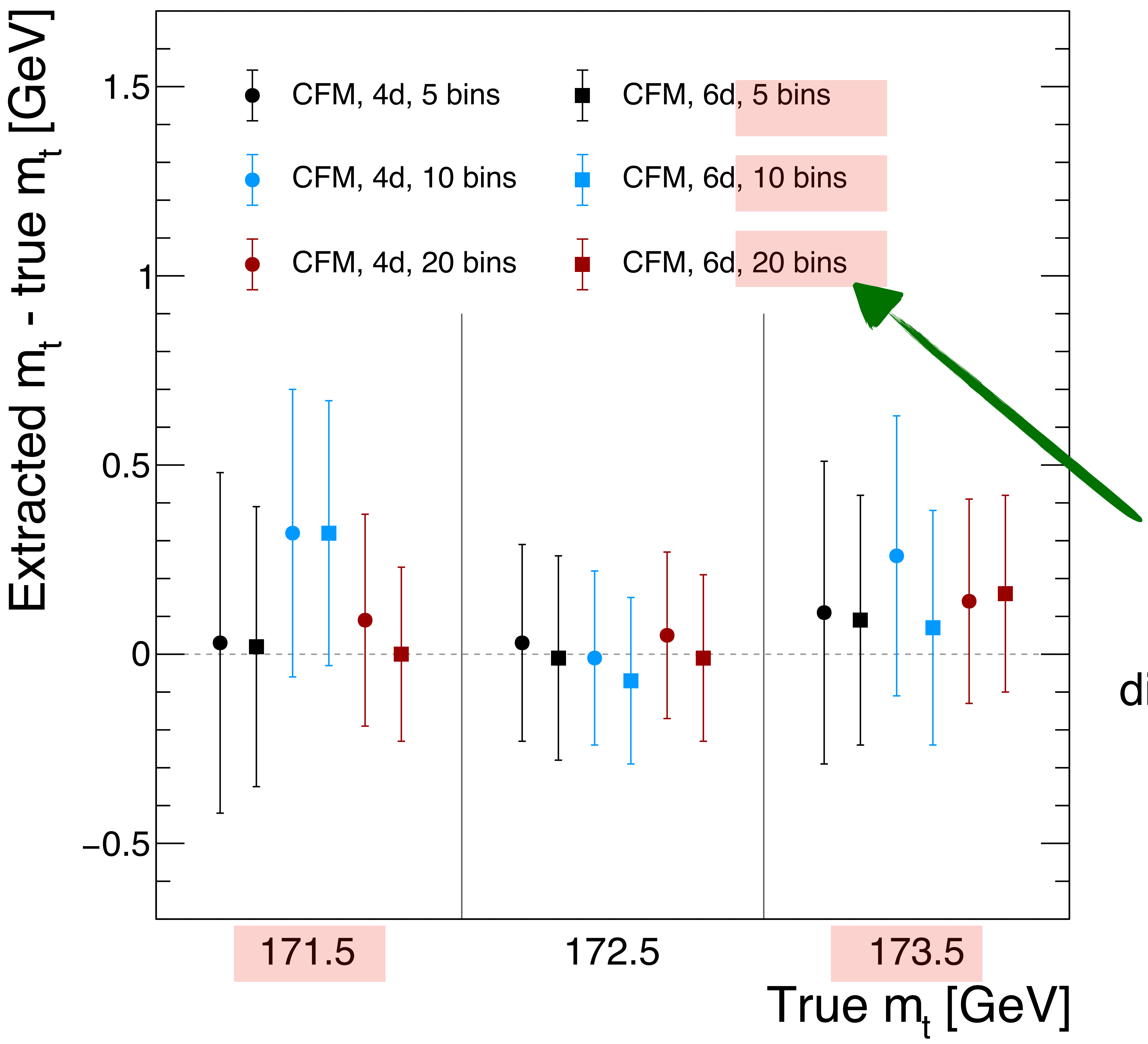
Mass Measurement

For a fixed top mass:

Choose subset of test data of 41000 reco level events

Unfolded 1000 bootstrapped replicas

Estimate covariance matrix and mean by 1000 different unfolded distributions



Why not more?

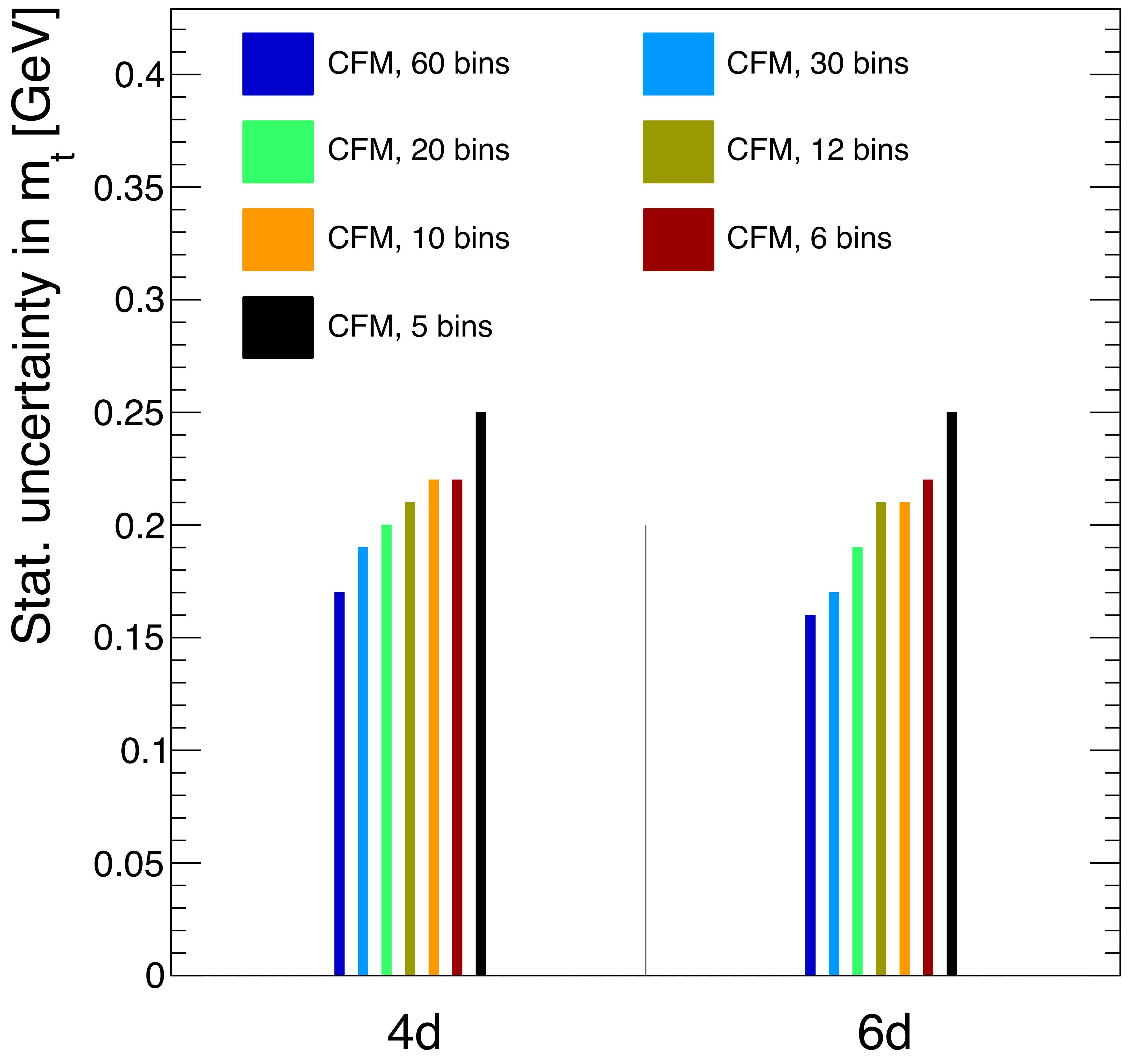
Limited by discrete grid of available m_t simulations

→ Reliably unfold triple jet mass without bias

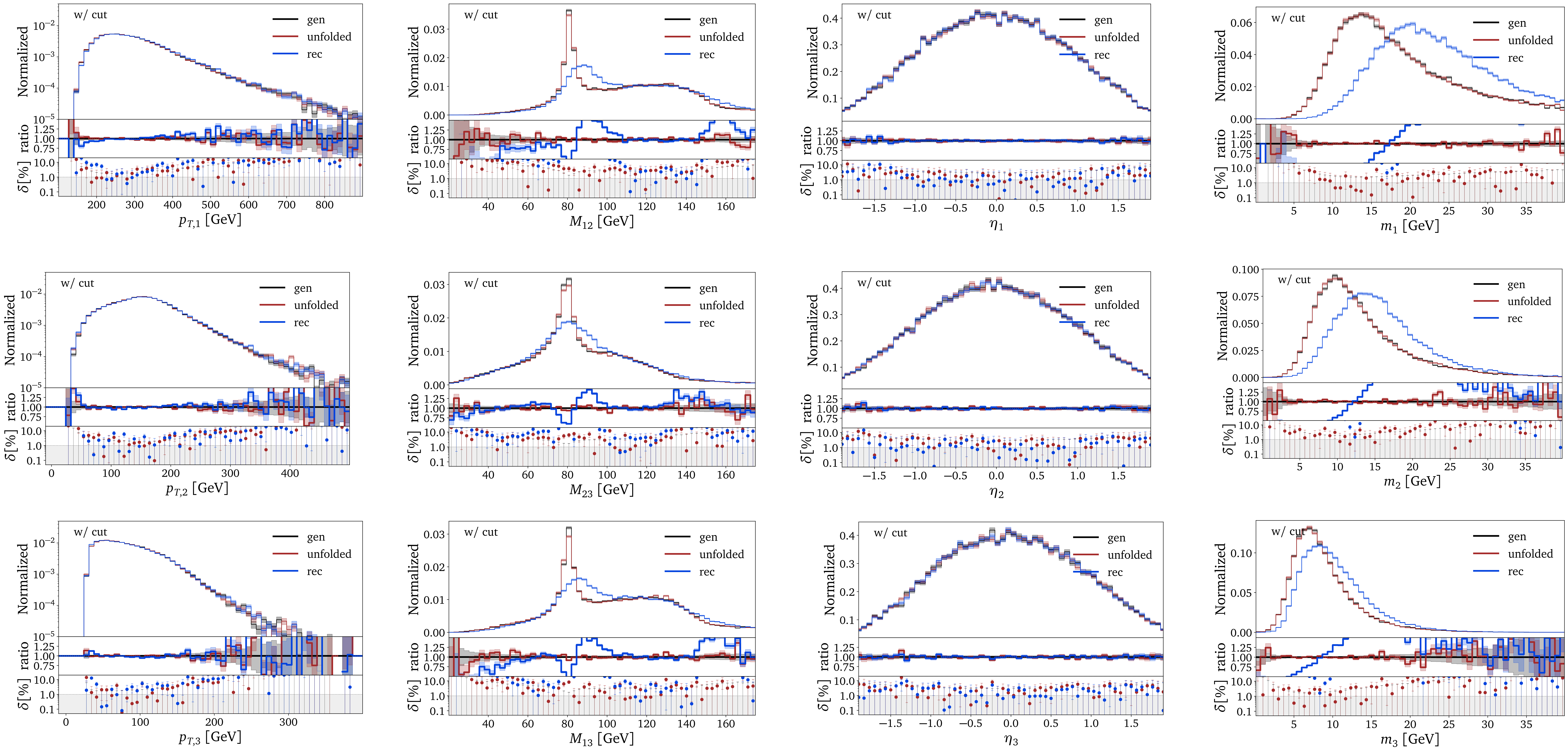
Mass Measurement

For $m_t = 172.5$ GeV, we have a close grid of available simulations (± 1 GeV)

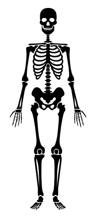
Statistical uncertainty for 60 bins decreases by 36%



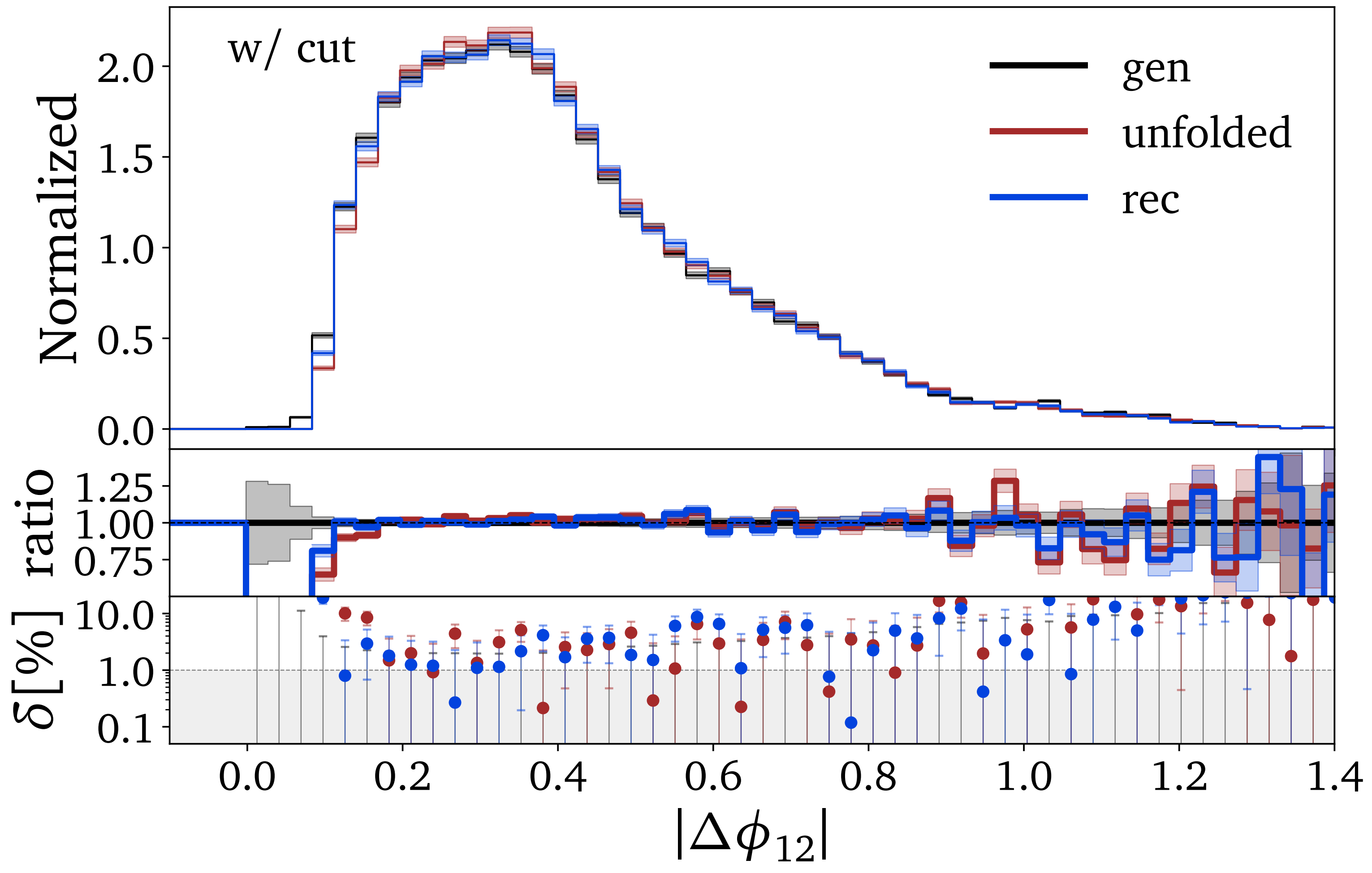
Full Phase Space Unfolding (12d)

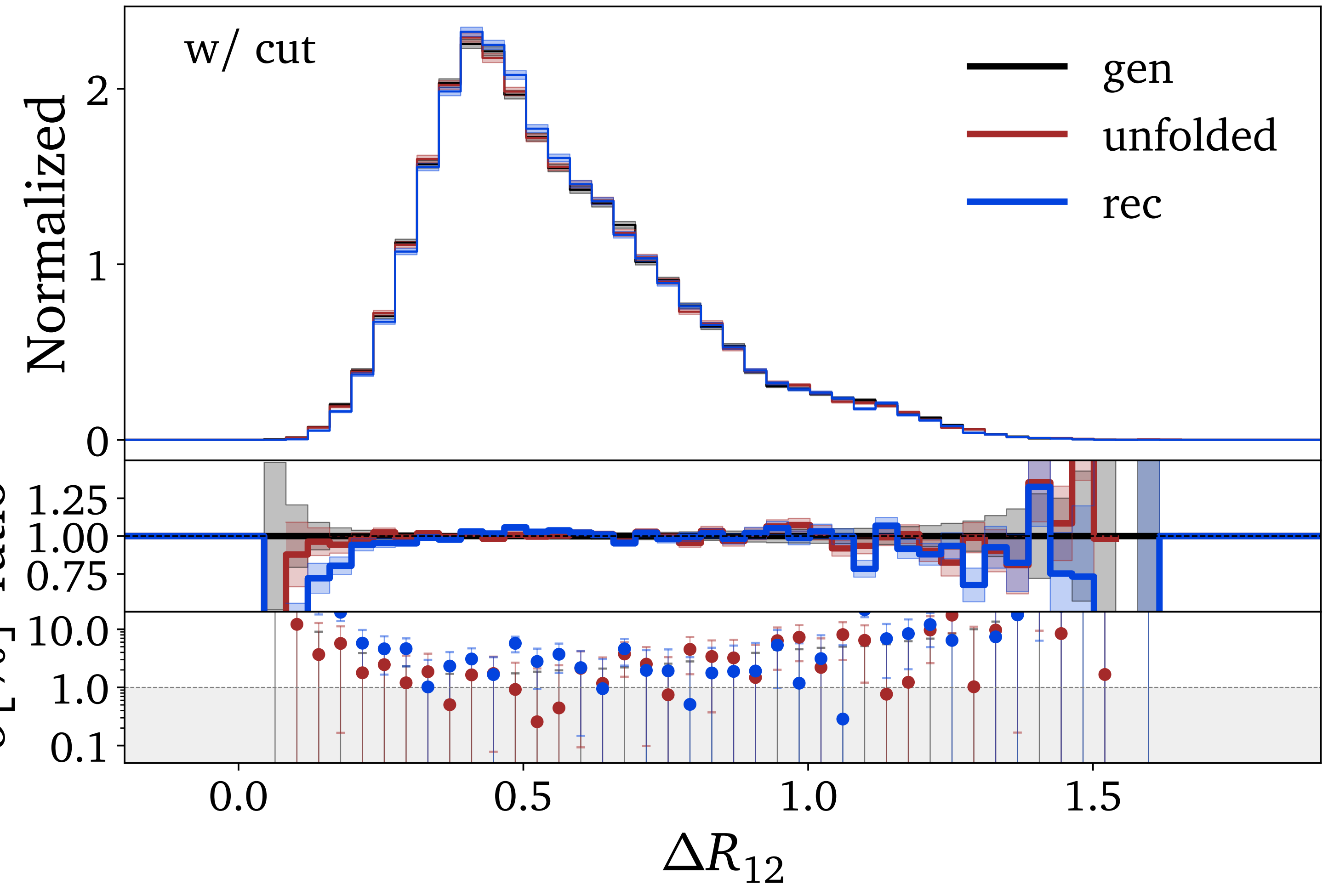


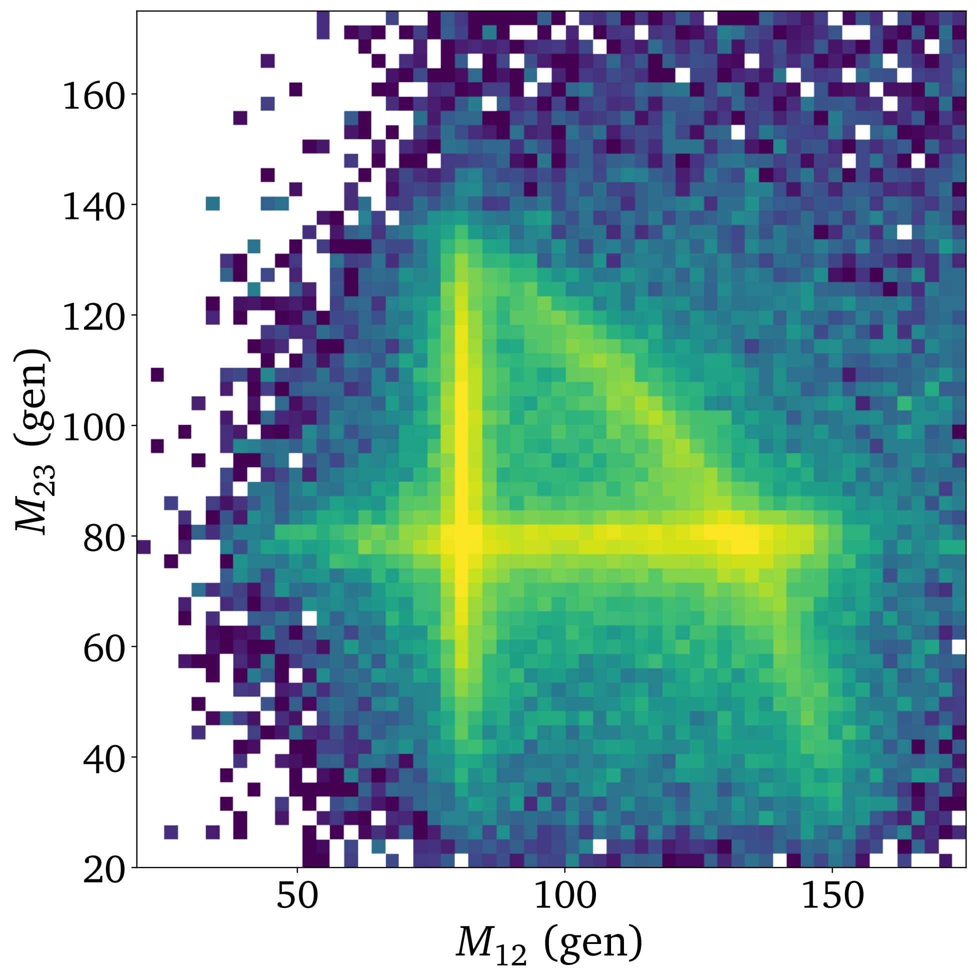
Full Phase Space Unfolding (12d) - Correlations

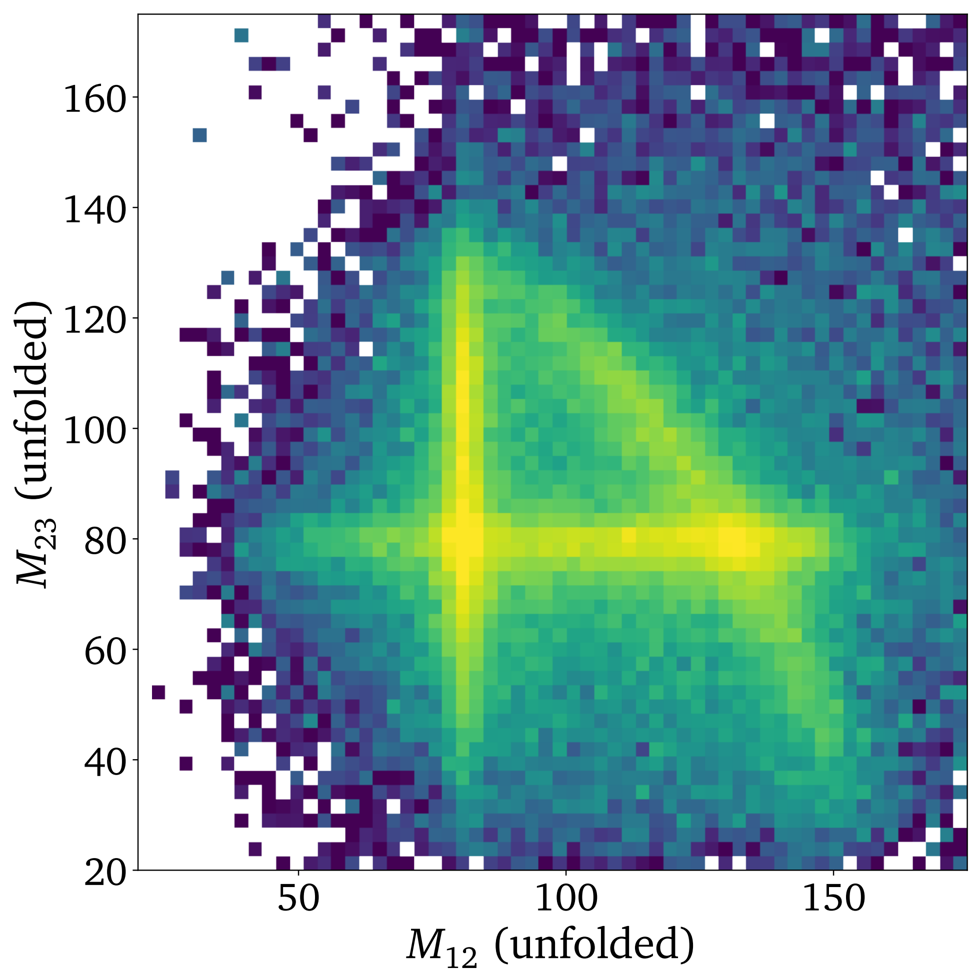


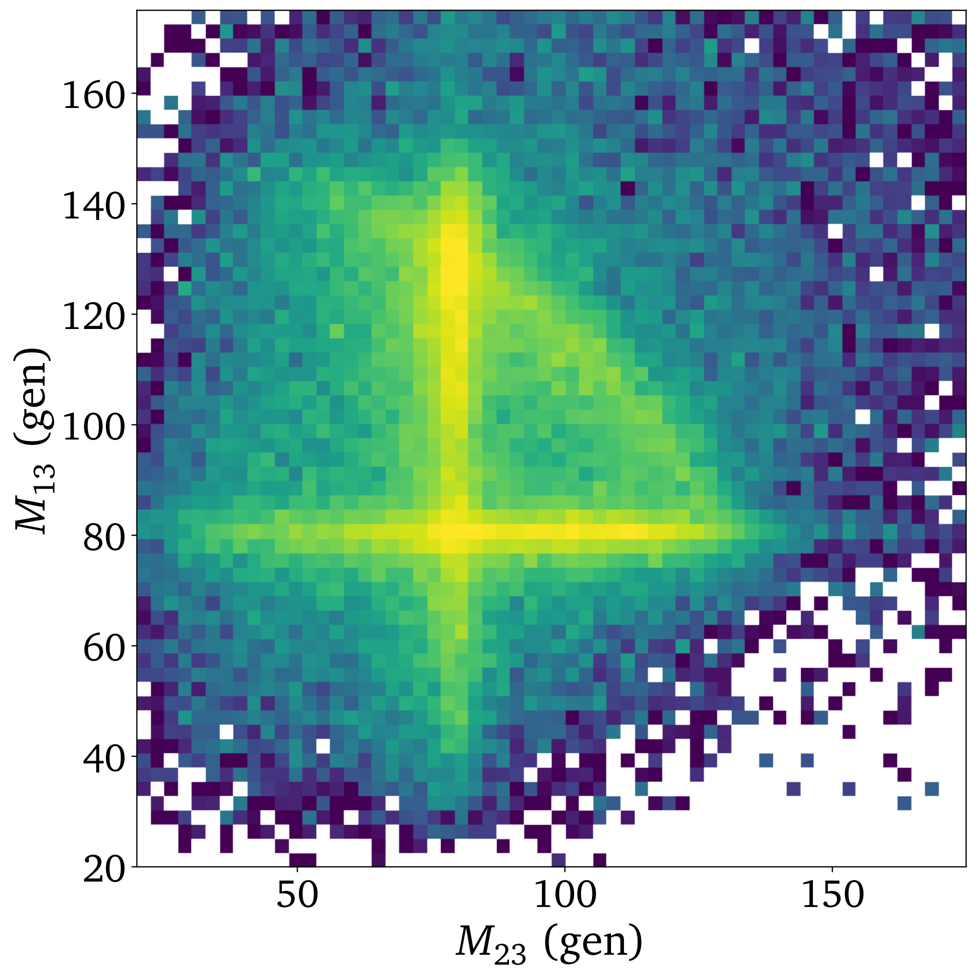
Cut at $|\Delta\phi_{ij}| > 0.1$

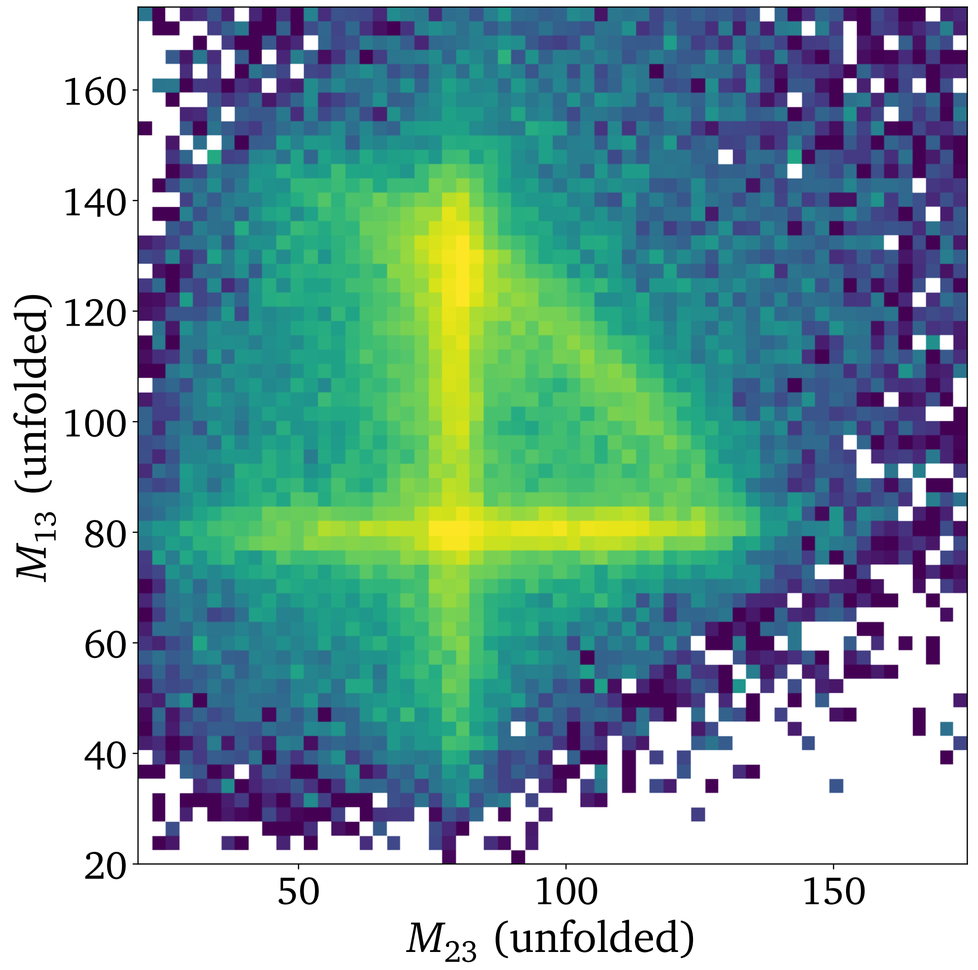












And now what?

Generative machine learning allows for unbinned, high dimensional unfolding

Unbiased networks can enhance precision in e.g. top mass measurement

Crucial step to build generative unfolding into existing LHC analysis

Proposal of analysis pipeline:

1. Event Selection
2. Jet calibration
3. Unfold subset
4. Measure top mass
5. Resimulate
6. Unfold full phasespace

And now what?

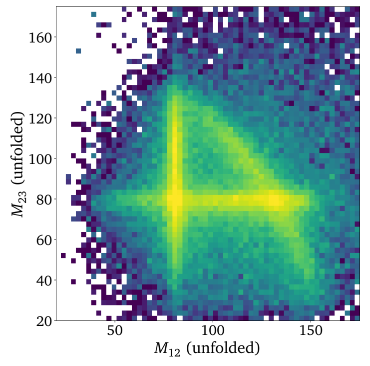
Generative machine learning allows for unbinned, high dimensional unfolding

Unbiased networks can enhance precision in e.g. top mass measurement

Crucial step to build generative unfolding into existing LHC analysis

Proposal of analysis pipeline:

1. Event Selection
2. Jet calibration
3. Unfold subset
4. Measure top mass
5. Resimulate
6. Unfold full phasespace



re there any questions?