A Matrix-Based Approach for Jet-Parton Assignment Leveraging Mass and Momentum Using CMS Open Data

Eric Reinhardt[1] and Sergei Gleyzer[1] on behalf of the CMS collaboration

[1]The University of Alabama

Jet-parton assignment involves trying to properly pair jets originating from the same parent parton

- Matrix element method
- Chi-squared minimization
- Machine learning

Typical Mass-Based Approach

$$\chi^2_{t\bar{t}} = \frac{(m_{b_1q_1q_1} - m_t)^2}{\sigma_t^2} + \frac{(m_{b_2q_2q_2} - m_t)^2}{\sigma_t^2} + \frac{(m_{q_1q_1} - m_W)^2}{\sigma_W^2} + \frac{(m_{q_2q_2} - m_W)^2}{\sigma_W^2}$$

- Mass Formula: $M = \sqrt{(\Sigma_i E_i)^2 - (\Sigma_i p_i)^2}$
- Separate b-tagged and non-b-tagged jets
- Permute all combinations of quark jets and compare masses to what you expect to observe

Downsides to mass-based approach

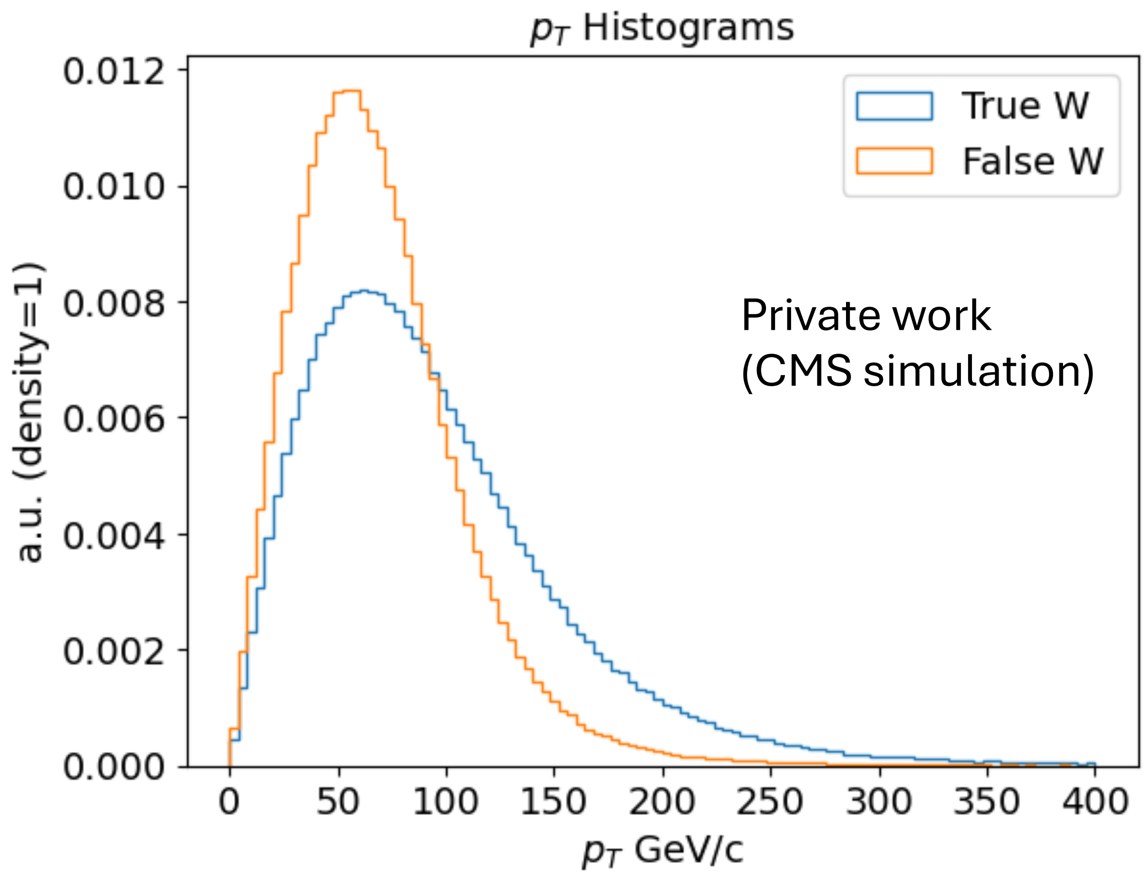
- Not effective for high jet multiplicity
- Fails to reconstruct outliers in mass distribution
- Fail rate increases with complex event signatures like 4 top etc.

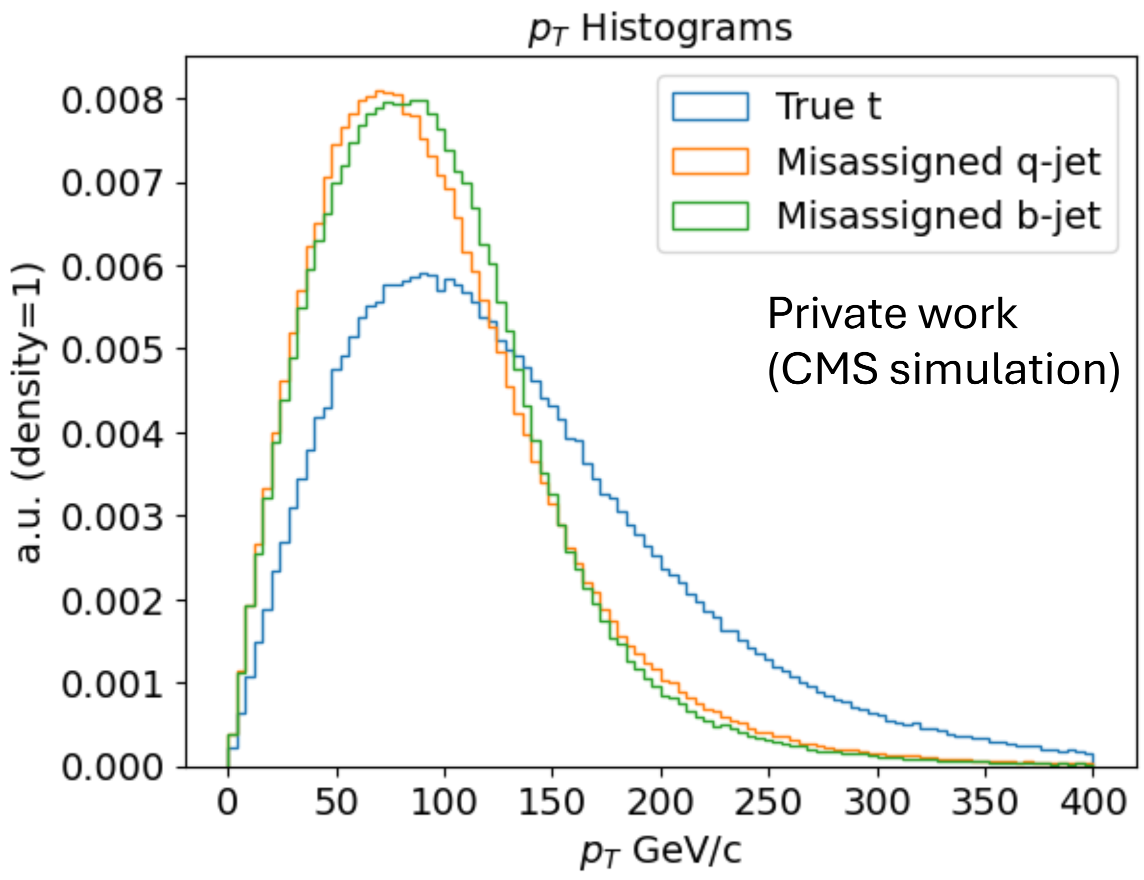
Initial cuts

- Pythia8 + Powheg simulated dataset [1]
- Exactly two jets with medium b-tag of >0.7
- Between 6 and 8 jets ***Note: typical approach often cuts to only 6 jets**
- "Assignable" events must have all 6 unique reco-to-gen truth-matched assignments below a threshold of $\Delta\mathrm{R} = 0.4$ and with a pT ratio of above 0.7
- All just must have a valid b-tag score
- All jets must have $|\eta| < 2.4$
- All just must have pT > 20GeV ***Note: typical approach often cuts to 55GeV**
- Total Events after cuts: N=1,013,419
- Ground-Truth Assignable Events: N=103,481

[1] CMS Collaboration (2021). Simulated dataset TT_TuneCUETP8M1_mtop1735_13TeV-powheg-pythia8 in MINIAODSIM format for 2015 collision data. CERN Open Data Portal. DOI:10.7483/OPENDATA.CMS.Q22T.BNJT

We can use pT as a discriminator for W boson and top quark assignments





Matrix-Computation Approach:

Create pairing weight matrices where row and column numbers correspond to specific non-b-tagged jets in the event and matrix number corresponds to specific b-tagged quarks

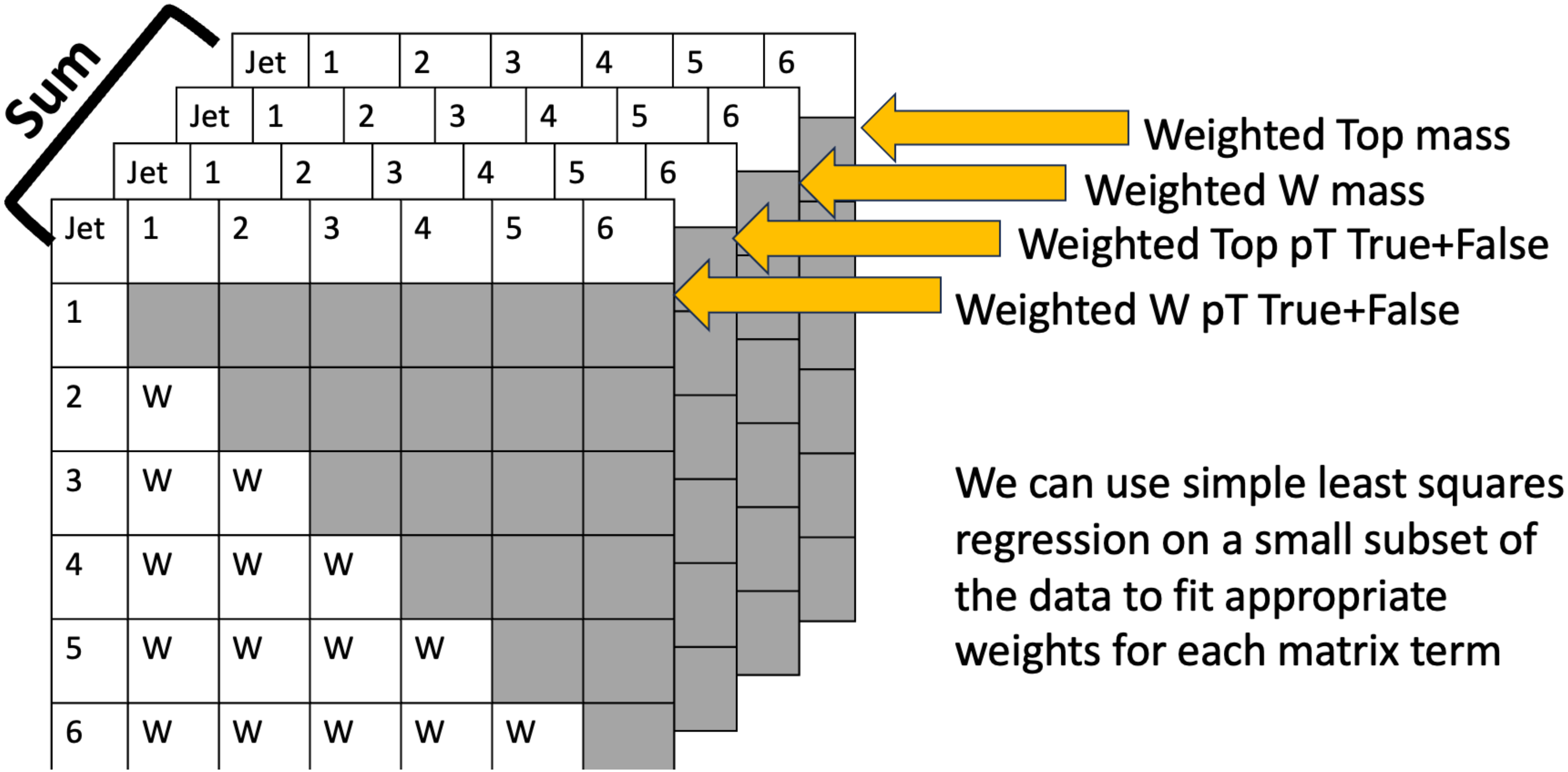
Bottom quark 1

Jet	1	2	3	4	5	6
1						
2	W					
3	W	W				
4	W	W	W			
5	W	W	W	W		
6	W	W	W	W	W	

Bottom quark 2

Jet	1	2	3	4	5	6
1						
2	W					
3	W	W				
4	W	W	W			
5	W	W	W	W		
6	W	W	W	W	W	

Our final pairing likelihood matrices are sums of mass and pT matrices



We can use simple least squares regression on a small subset of the data to fit appropriate weights for each matrix term

Shortcut approach to save computation time

Select most probable pairing then mask all rows and columns from the other matrix that share the same jet numbers then choose the second assignment from remaining options

Bottom quark 1

Jet	1	2	3	4	5	6
1						
2	W					
3	W	W				
4	W	W	W			
5	W	Best	W	W		
6	W	W	W	W	W	

Bottom quark 2

Jet	1	2	3	4	5	6
1						
2	W					
3	W	W				
4	W	W	W			
5	W	W	W	W		
6	W	W	W	W	W	

Low-cost improvement: Top-2 Selection

Conflicts present an issue when the most probable index could fit either top quark. Here the conflict is on jet 2 being assigned to both top quarks

Bottom quark 1

Jet	1	2	3	4	5	6
1						
2	W					
3	W	W				
4	W	W	W			
5	W	Best	W	W		
6	W	W	W	W	W	

Bottom quark 2

Jet	1	2	3	4	5	6
1						
2	W					
3	W	Best				
4	W	W	W			
5	W	W	W	W		
6	W	W	W	W	W	

Low-cost improvement: Top-2 Selection

We first consider one top quark matrix as having "priority" in selection and select the best entry for that matrix

Bottom quark 1

Jet	1	2	3	4	5	6
1						
2	W					
3	W	W				
4	W	W	W			
5	W	Best	W	W		
6	W	W	W	W	W	

Bottom quark 2

Jet	1	2	3	4	5	6
1						
2	W					
3	W	W				
4	W	W	W			
5	W	W	W	W		
6	W	W	W	W	W	

Low-cost improvement: Top-2 Selection

We then consider the other top quark matrix as having "priority" in selection and select the best entry for that matrix then compare the best combination of soft-minimum weights across both matrices

Bottom quark 1

Jet	1	2	3	4	5	6
1						
2	W					
3	W	W				
4	W	W	W			
5	W	W	W	W		
6	W	W	W	W	W	

Bottom quark 2

Jet	1	2	3	4	5	6
1						
2	W					
3	W	Best				
4	W	W	W			
5	W	W	W	W		
6	W	W	W	W	W	

Baseline efficiency comparison

- Same cuts as before + all jets must have pT > 55GeV
  - This cut is very standard as classical computing, mass-only approach to improve accuracy of jet-parton assignment
- Total Events after cuts: N=40,889
- Ground-Truth Assignable Events: N=6,072

Baseline efficiency comparison with cut of each jet pT > 55GeV with only events which are ground-truth assignable

Private work (CMS simulation)	***Mass-Only***		***Mass + Momentum***	
Num Jets	***Events***	***Efficiency***	***Events***	***Efficiency***
6	1493	0.997	1493	**0.998**
7	2236	0.859	2236	**0.912**
8	2343	0.694	2343	**0.791**

Baseline efficiency comparison with cut of each jet pT > 55GeV with only events which are ground-truth assignable with χ^2 cutoffs*

Private work (CMS simulation)	***Mass-Only***		***Mass + Momentum***	
Num Jets	***Events***	***Efficiency***	***Events***	***Efficiency***
6	1217	0.998	**1473**	**0.999**
7	1612	0.888	**2040**	**0.931**
8	1586	0.734	**1770**	**0.862**

***Note: The decision threshold used for mass-only is $\chi^2 < 20$ which is a common choice in recent literature**

Baseline efficiency comparison with cut of each jet pT > 55GeV for all events including ones which aren't ground-truth assignable with χ^2 cutoffs*

Private work (CMS simulation)	***Mass-Only***		***Mass + Momentum***	
Num Jets	***Events***	***Efficiency***	***Events***	***Efficiency***
6	1679	**0.724**	**1712**	0.689
7	**3017**	0.474	2892	**0.504**
8	**4014**	0.290	2939	**0.377**

***Note: The decision threshold used for mass-only is $\chi^2 < 20$ which is a common choice in recent literature**

Back to new approach without 55 GeV pT requirement

- Total Events after cuts: N=304,441
- Ground-Truth Assignable Events: N=31,193

New approach without 55GeV pT cut with only events which are ground-truth assignable

Private work (CMS simulation)	***Mass-Only***		***Mass + Momentum***	
Num Jets	***Events***	***Efficiency***	***Events***	***Efficiency***
6	55071	0.860	55071	**0.867**
7	33002	0.649	33002	**0.705**
8	15408	0.503	15408	**0.590**

***Note: The decision threshold used for mass-only is $\chi^2 < 20$ which is a common choice in recent literature**

New approach without 55GeV pT cut for all events including ones which aren't ground-truth assignable with χ^2 cutoffs*

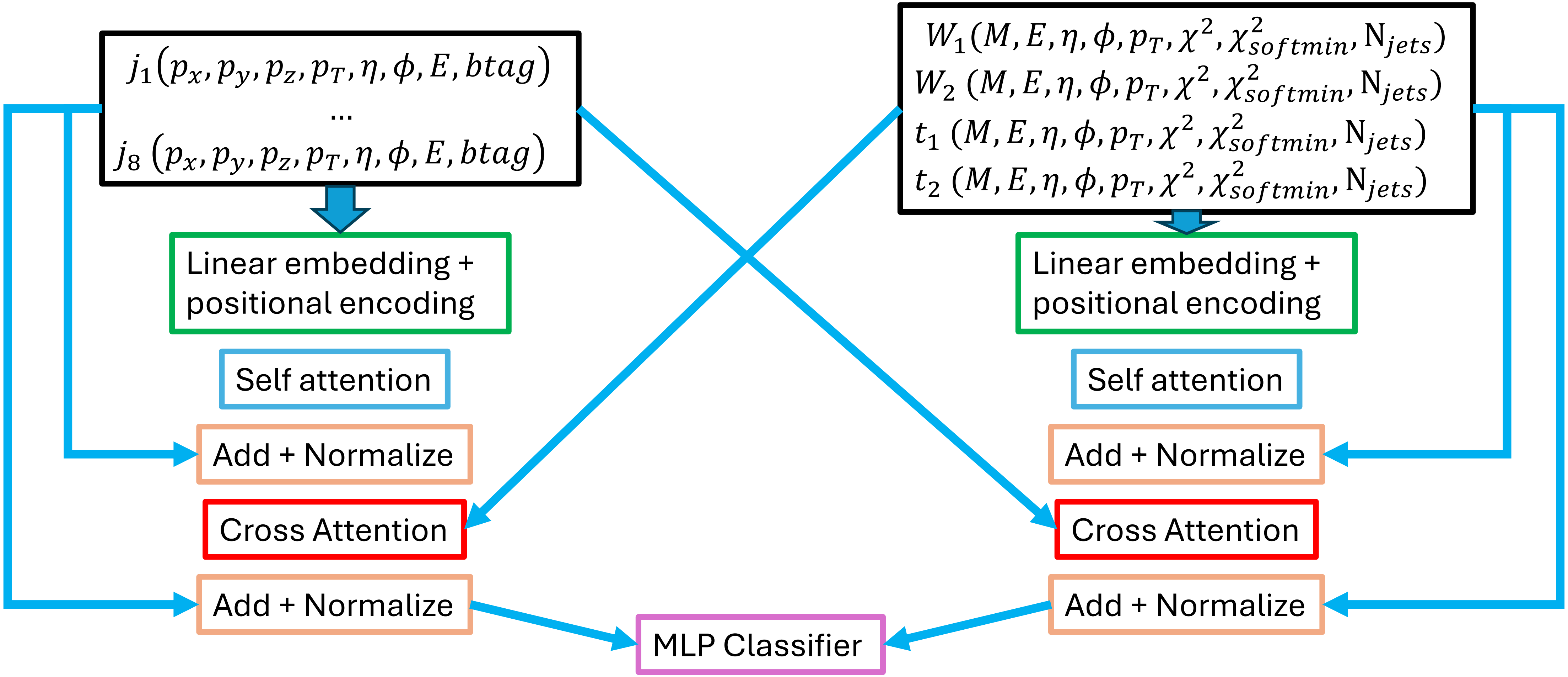
Private work (CMS simulation)	***Mass-Only***		***Mass + Momentum***	
Num Jets	***Events***	***Efficiency***	***Events***	***Efficiency***
6	1814	**0.687**	**2186**	0.560
7	**1769**	0.415	1541	**0.482**
8	**1267**	0.222	779	**0.421**

***Note: The decision threshold used for mass-only is $\chi^2 < 1$**

Approach comparison

- The mass-only approach is superior at discriminating between events which can be properly assigned and can't be properly assigned
- Momentum-only approach is superior when only looking at ground-truth assignable events
- 55 GeV pT removes many statistics but overall improves efficiency when exactly 6 jets are selected and mass-only approach is used

Machine learning can be used to remove unassignable events without requiring aggressive cuts on χ^2 cutoff values



New approach without 55GeV pT cut for all events including ones which aren't ground-truth assignable using machine learning to exclude unassignable events with a decision threshold

Private work (CMS simulation)	***Mass-Only + Machine Learning***		***Mass + Momentum + Machine Learning***	
Num Jets	***Events***	***Efficiency***	***Events***	***Efficiency***
6	2990	**0.881**	**3177**	0.871
7	**3050**	0.721	2964	**0.749**
8	3733	0.504	**4045**	**0.547**

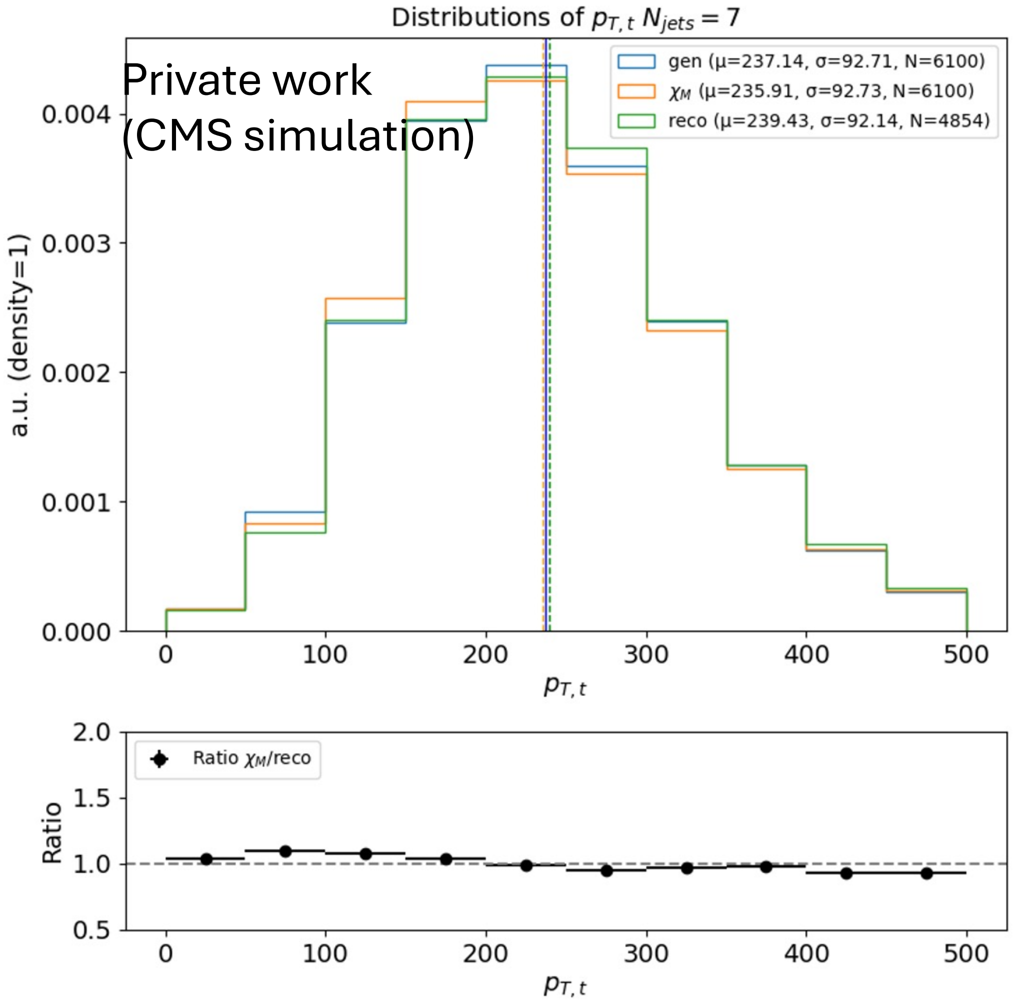
Total efficiency improvement using machine learning to exclude unassignable events compared to using jet pT cut of 55 GeV and χ^2 cutoff*

Private work (CMS simulation)	***Mass-Only Baseline***		***Mass + Momentum + Machine Learning***	
Num Jets	***Events***	***Efficiency***	***Events***	***Efficiency***
6	1679	0.724	**3177**	**0.871**
7	**3017**	0.474	2964	**0.749**
8	4014	0.290	**4045**	**0.547**

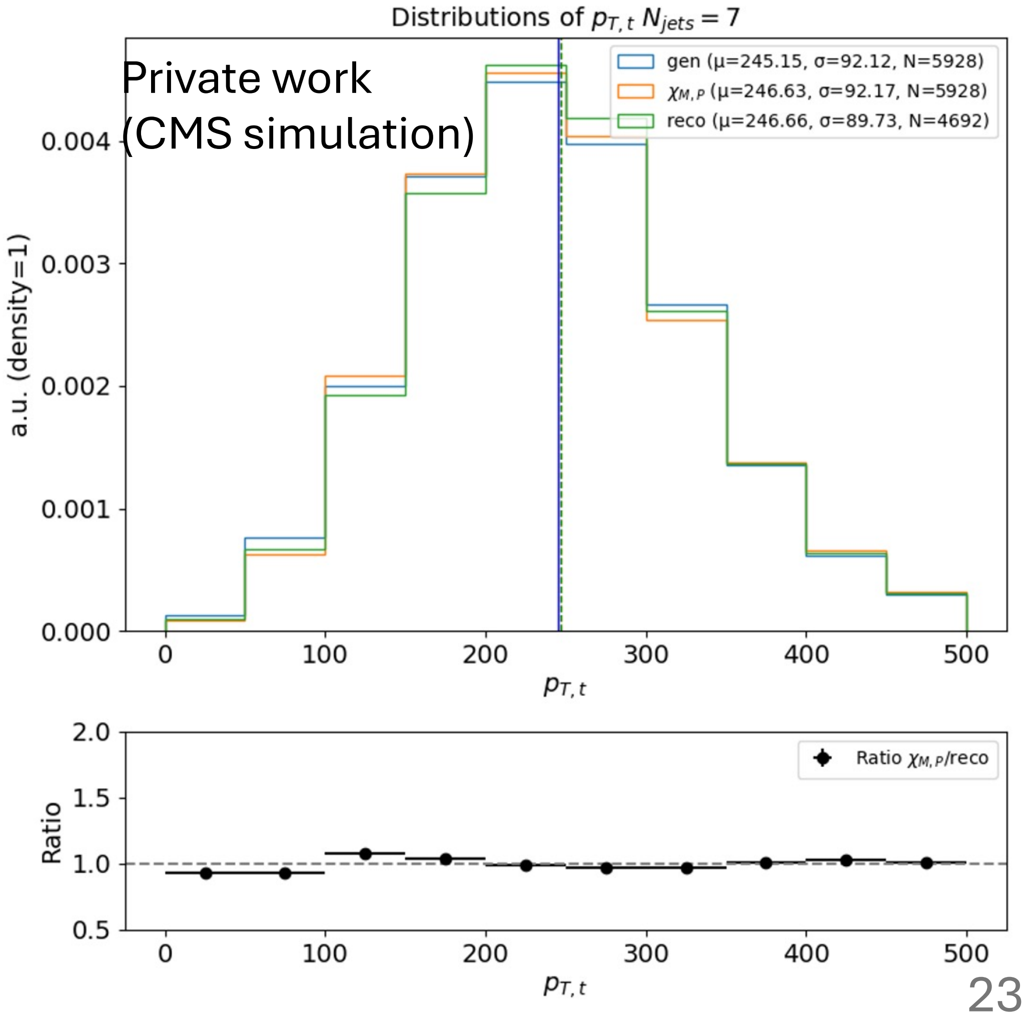
***Note: The decision threshold used for mass-only is $\chi^2 < 20$ which is a common choice in recent literature**

Reconstructed Kinematics: t pT, 7 jets

Mass Only



Momentum + Mass



Questions?

Backup Slides

“False” pairing matrix

Subtract away a weighted sum of the matrix elements from the opposite top-quark matrix which conflict with each pairing

Positive Matrix Element

Jet	1	2	3	4	5	6
1						
2	W					
3	W	W				
4	W	W	W			
5	W	W	W	W		
6	W	W	W	W	W	

Conflict Matrix Element

Jet	1	2	3	4	5	6
1						
2	W					
3	W	W				
4	W	W	W			
5	W	W	W	W		
6	W	W	W	W	W	

Additional machine learning details

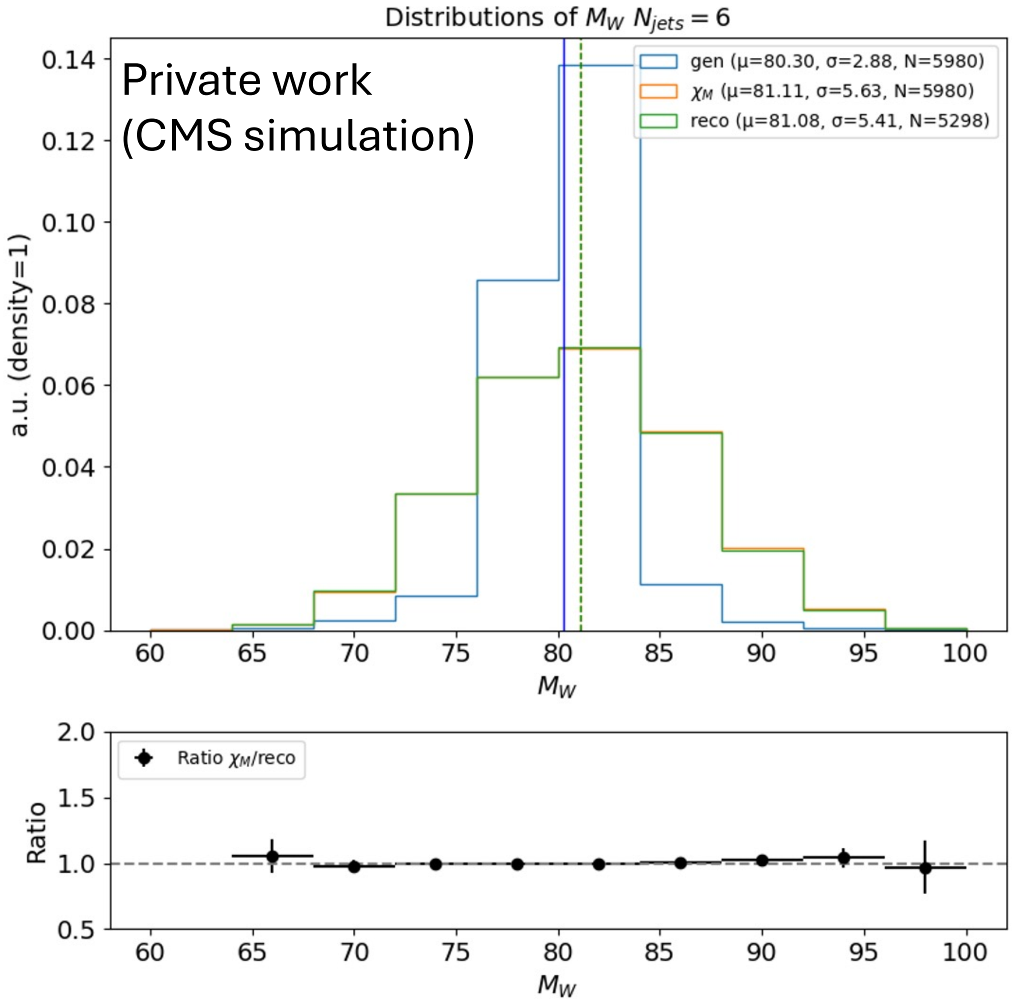
- Outliers determined using a 4.5*IQR Tukey fence are removed from both assignable and unassignable events with the same quartiles used for both
- Features are each normalized with the same mean and std used for assignable and unassignable events
- Weighted focal loss function is used to address class imbalance
- Nine different dropout layers are used to prevent overfitting
- The model is only considered valid if the performance on the test set exceeds the performance on the train and validation sets when dropout is not applied
- A minimum target of true positives (TP) is set and then beyond that number a decision threshold is selected which has best ratio of TP/FP on the test data

Note about accounting for additional radiated particles

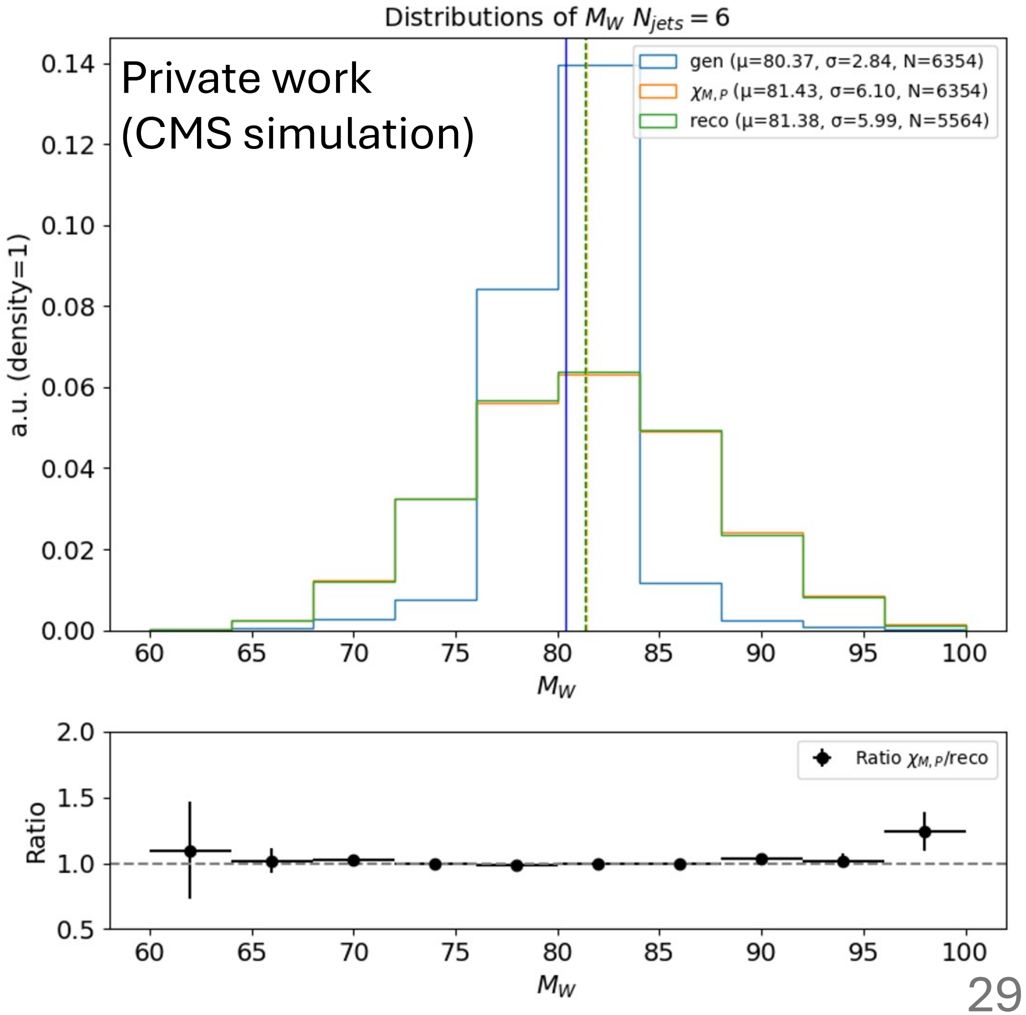
- Particles produced by interactions throughout the process here could be clustered separately from their associated originating particle
- Further work will aim to explore approaches combining overlapping clusters into joint objects in the initial truth-matching and matrix elements

Reconstructed Kinematics: W Mass, 6 jets

Mass Only

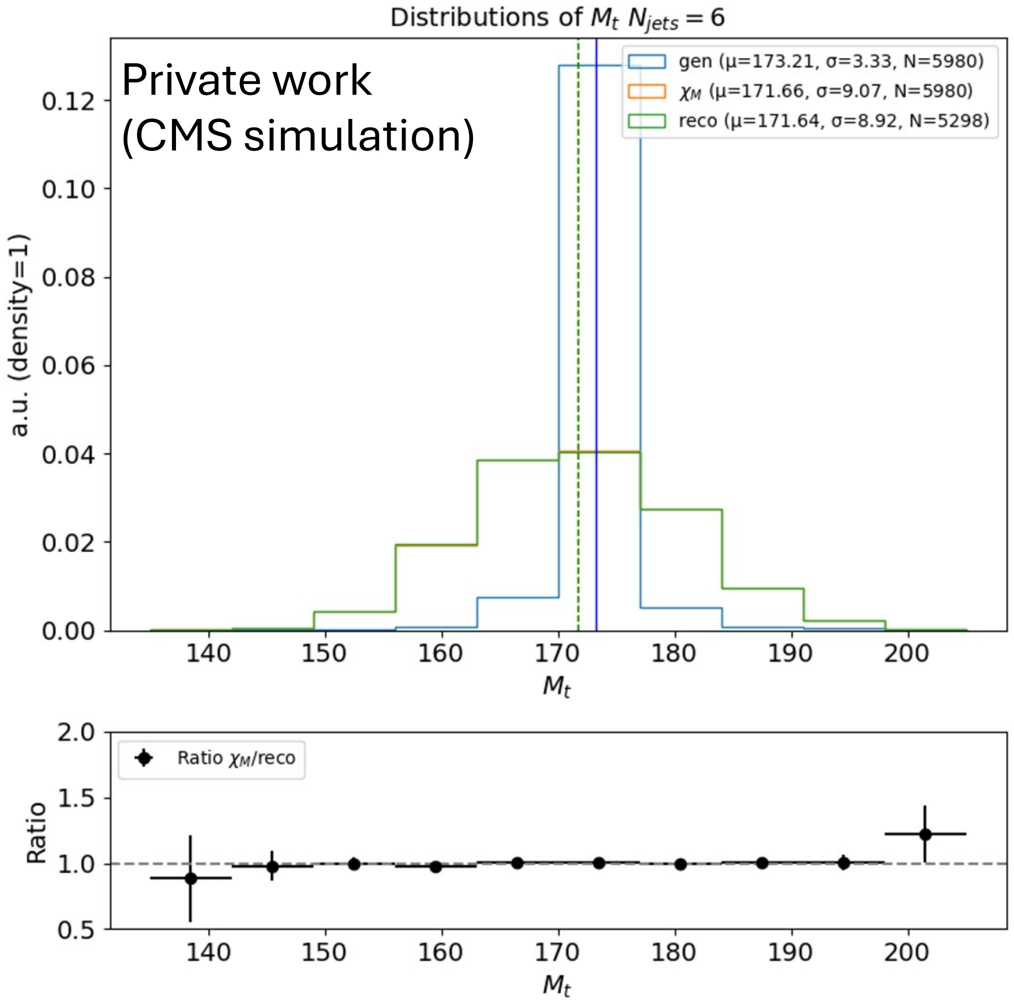


Momentum + Mass

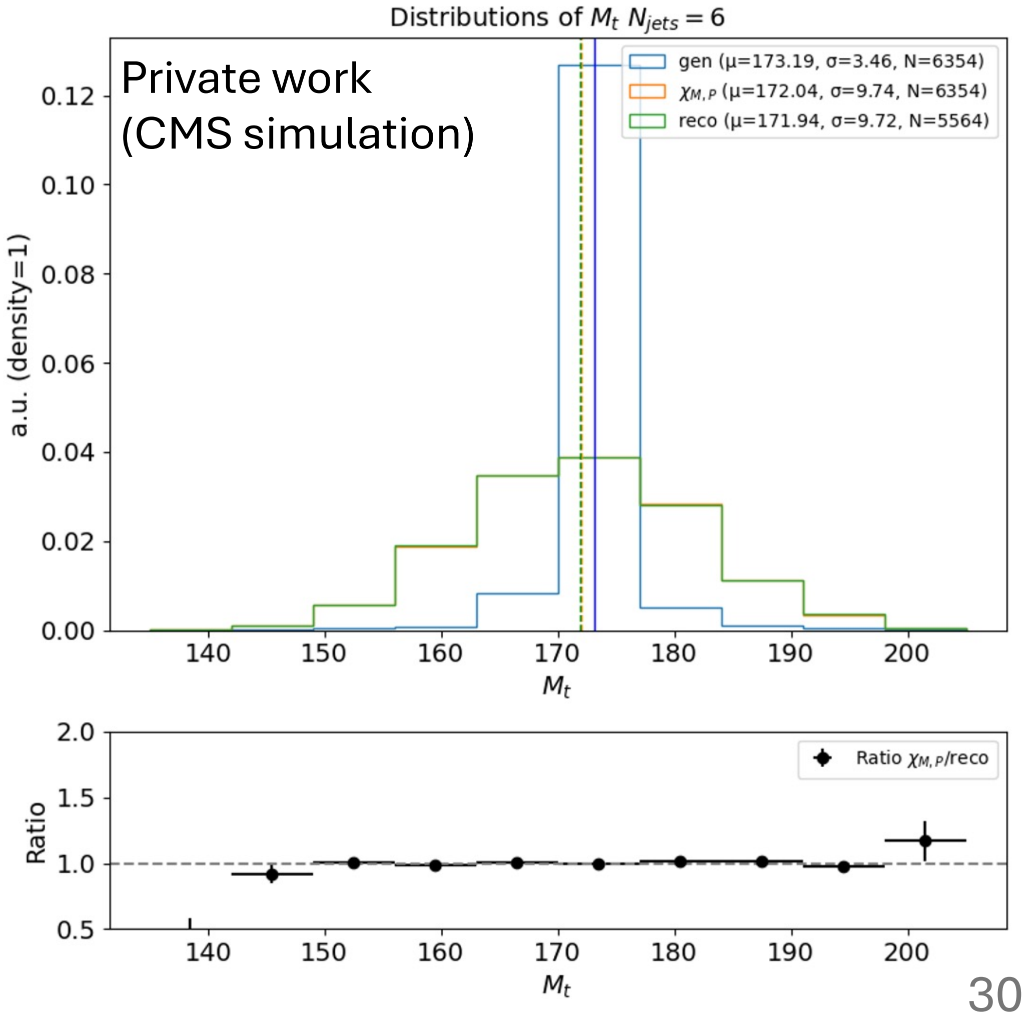


Reconstructed Kinematics: t Mass, 6 jets

Mass Only

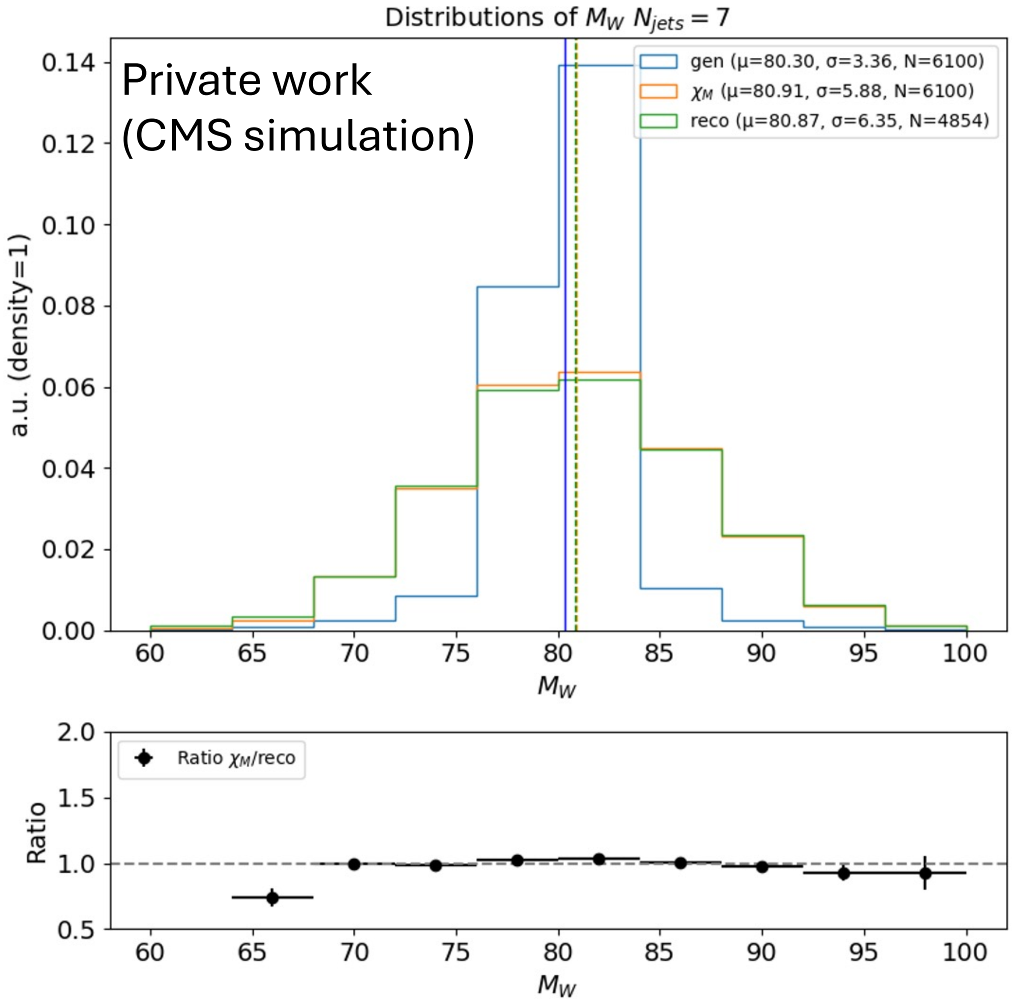


Momentum + Mass

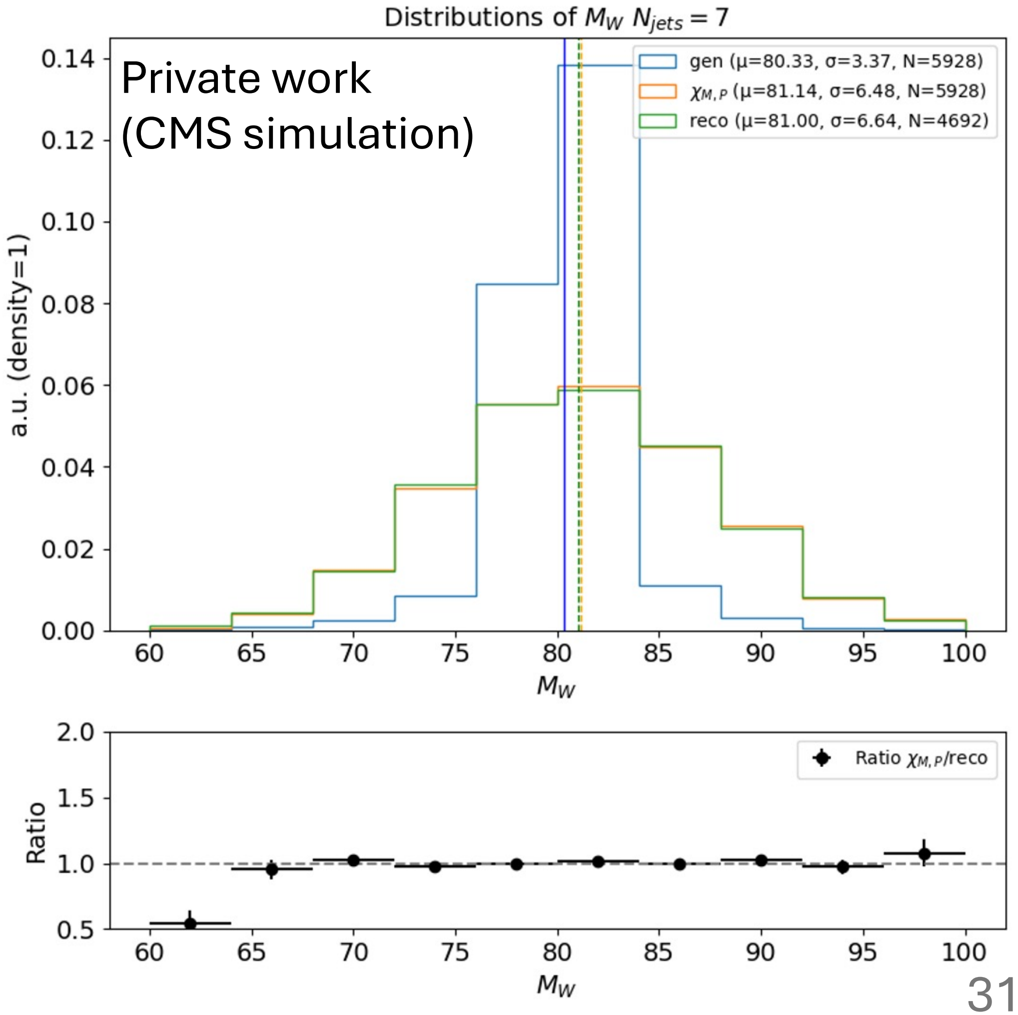


Reconstructed Kinematics: W Mass, 7 jets

Mass Only

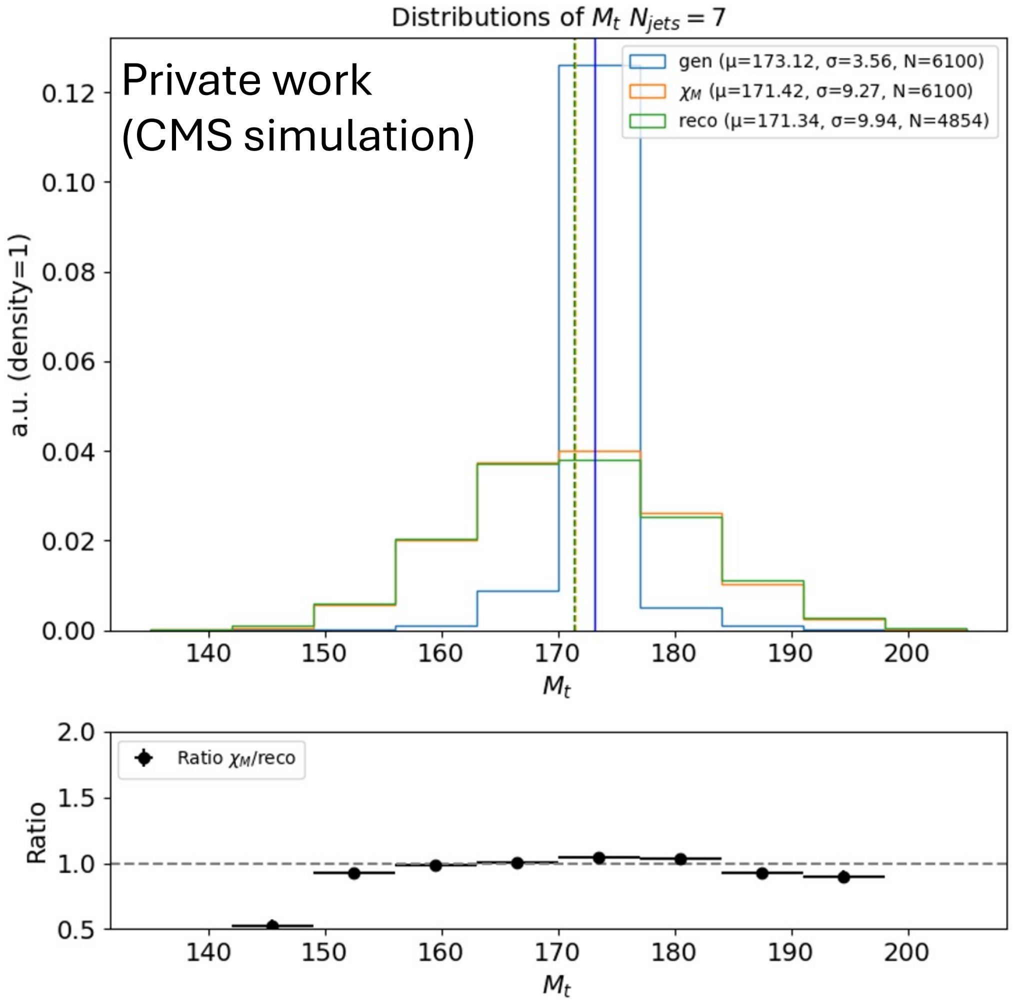


Momentum + Mass

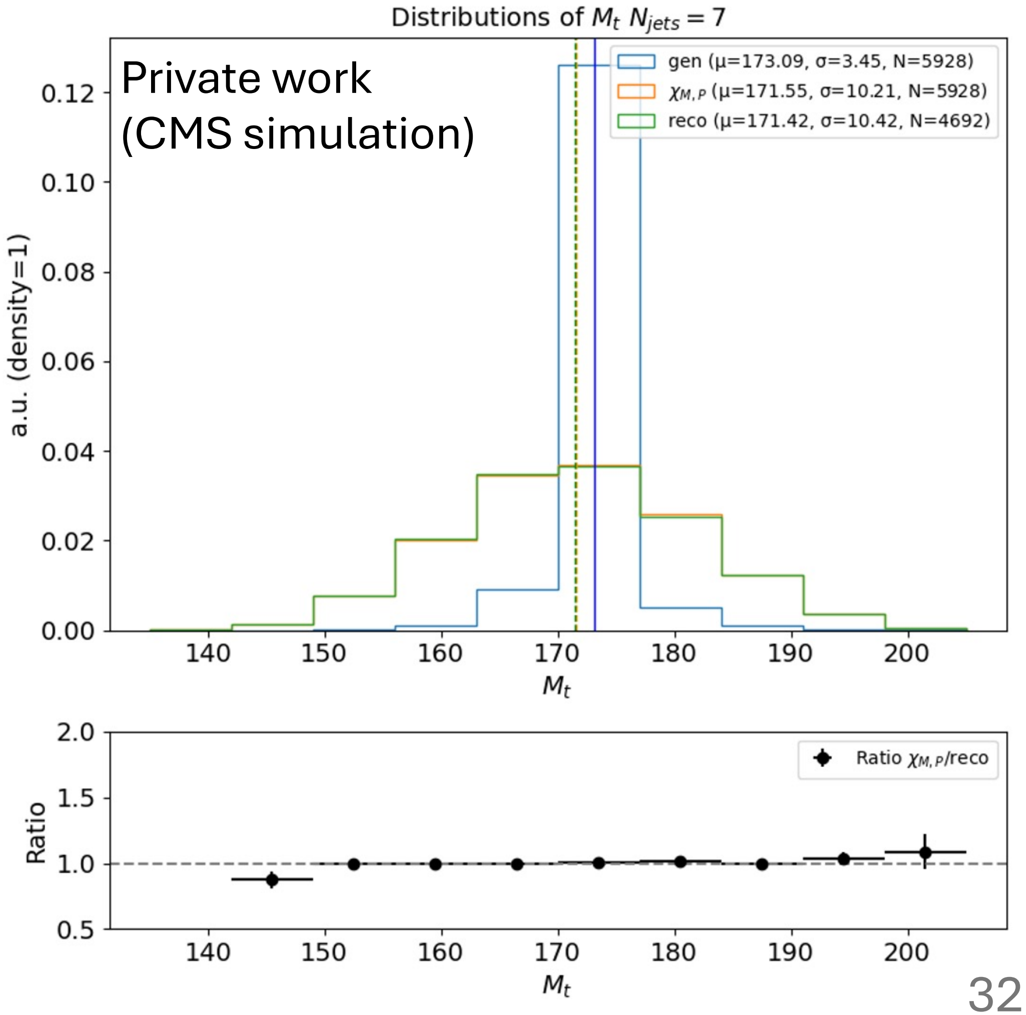


Reconstructed Kinematics: t Mass, 7 jets

Mass Only

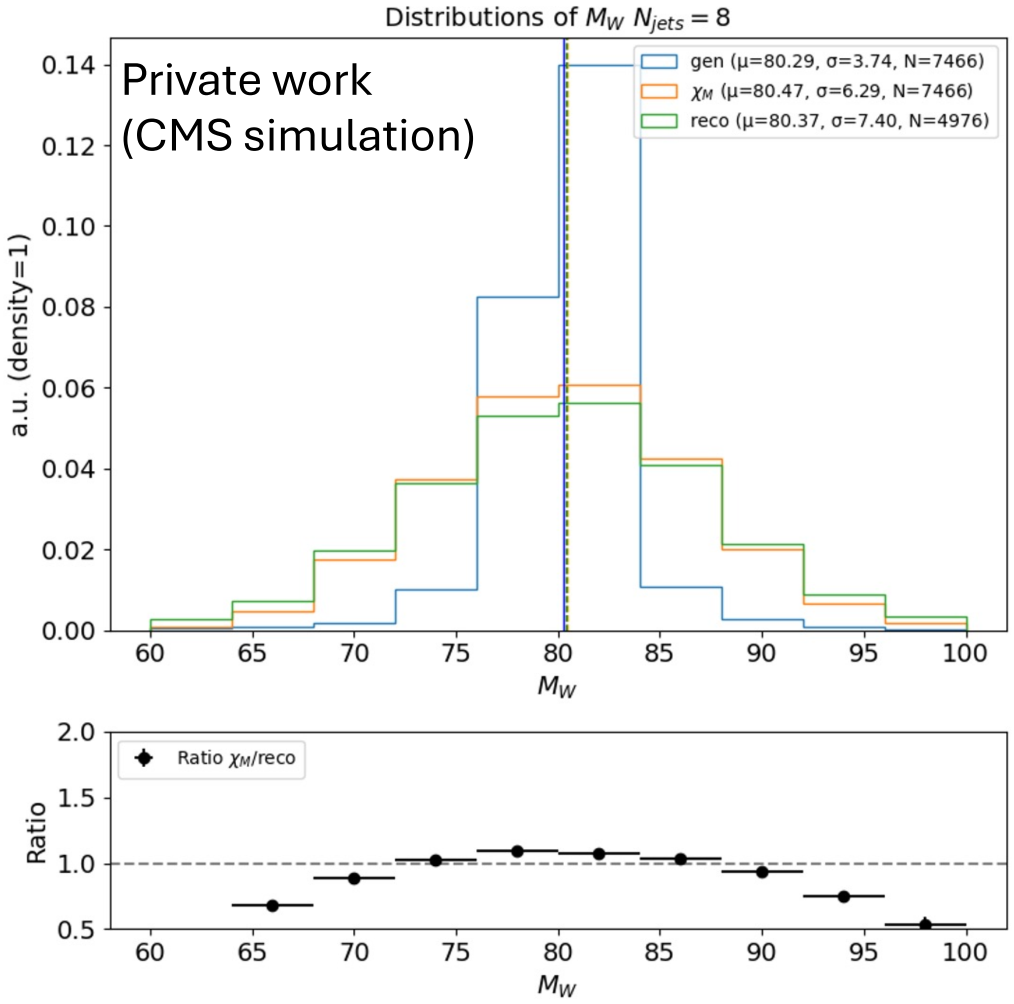


Momentum + Mass

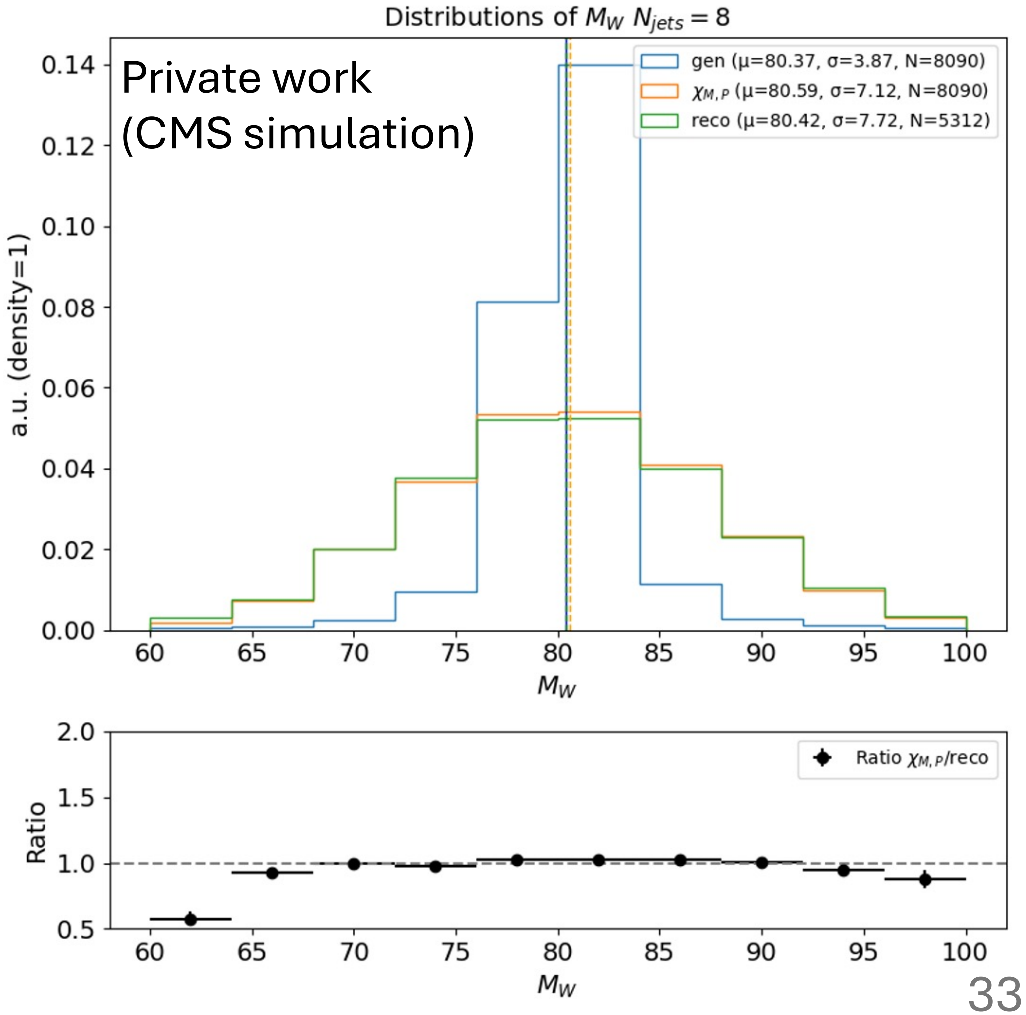


Reconstructed Kinematics: W Mass, 8 jets

Mass Only

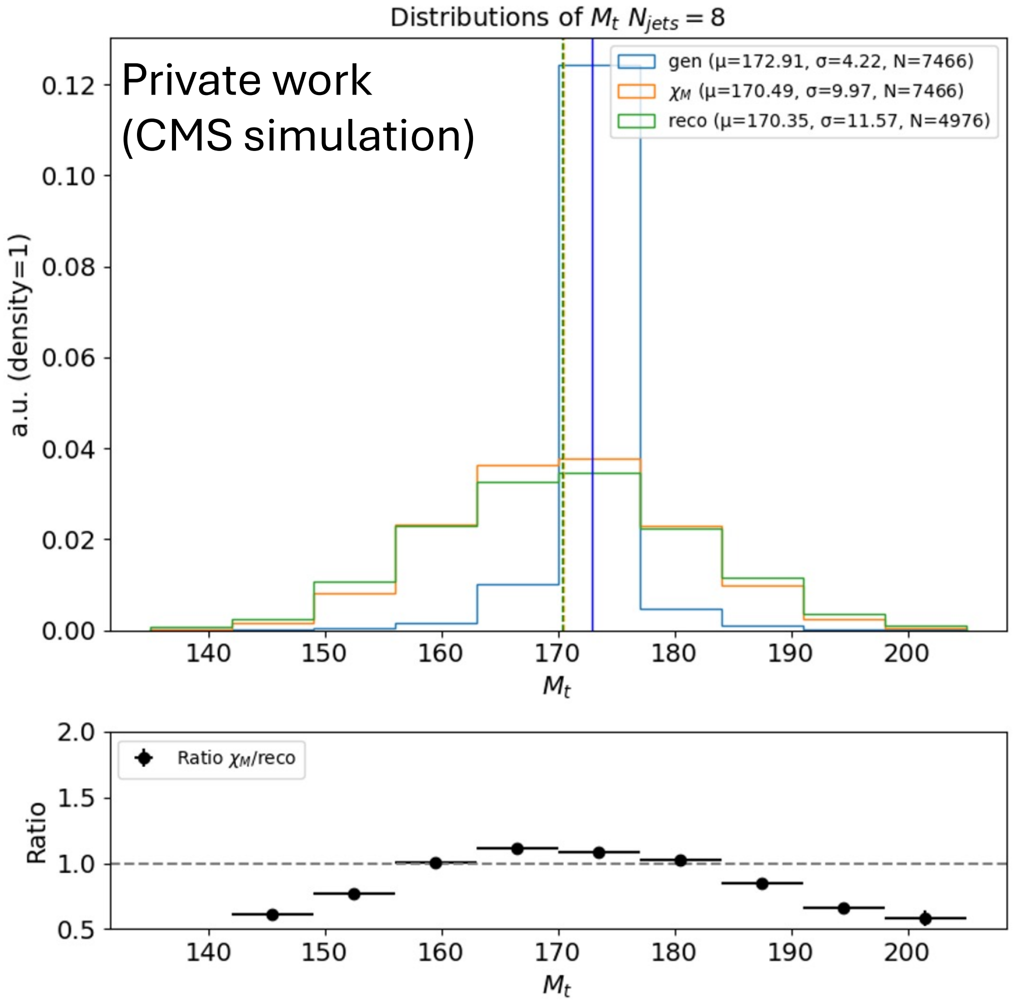


Momentum + Mass

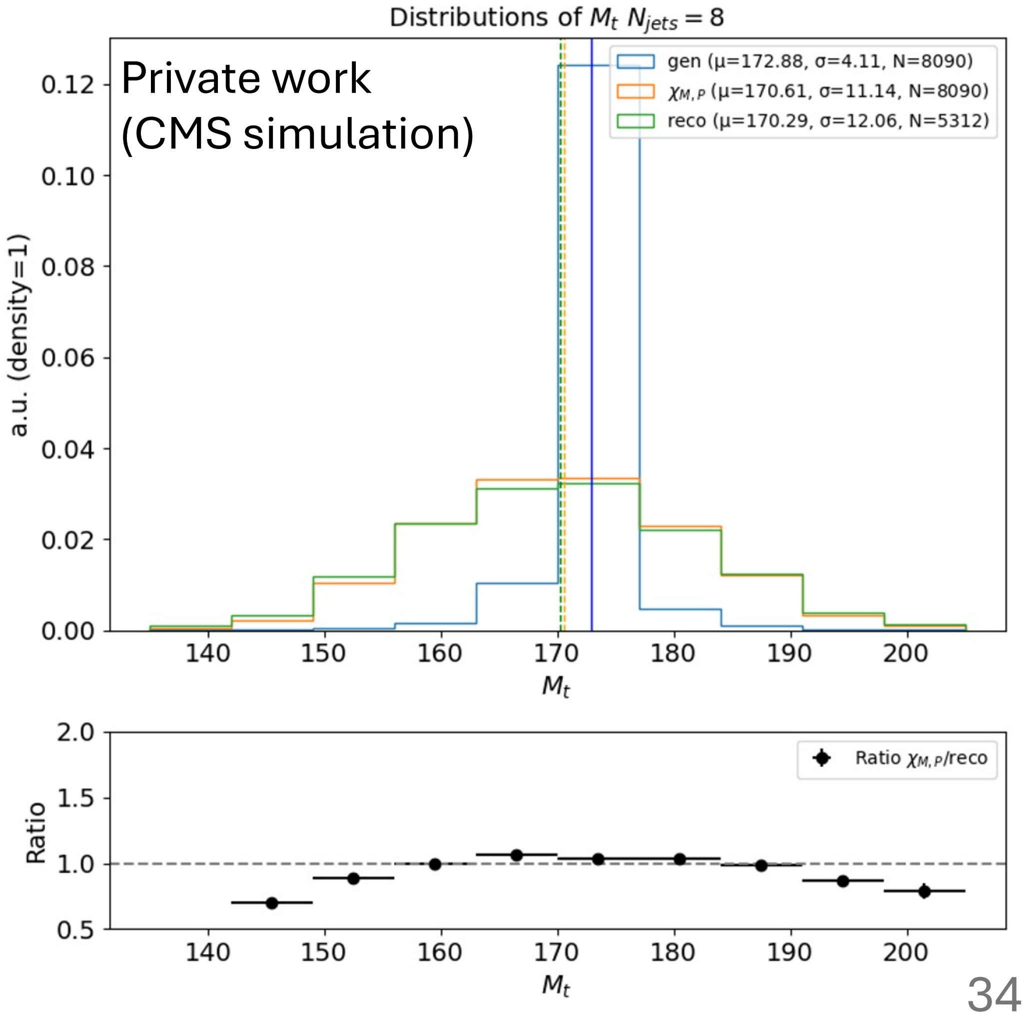


Reconstructed Kinematics: t Mass, 8 jets

Mass Only

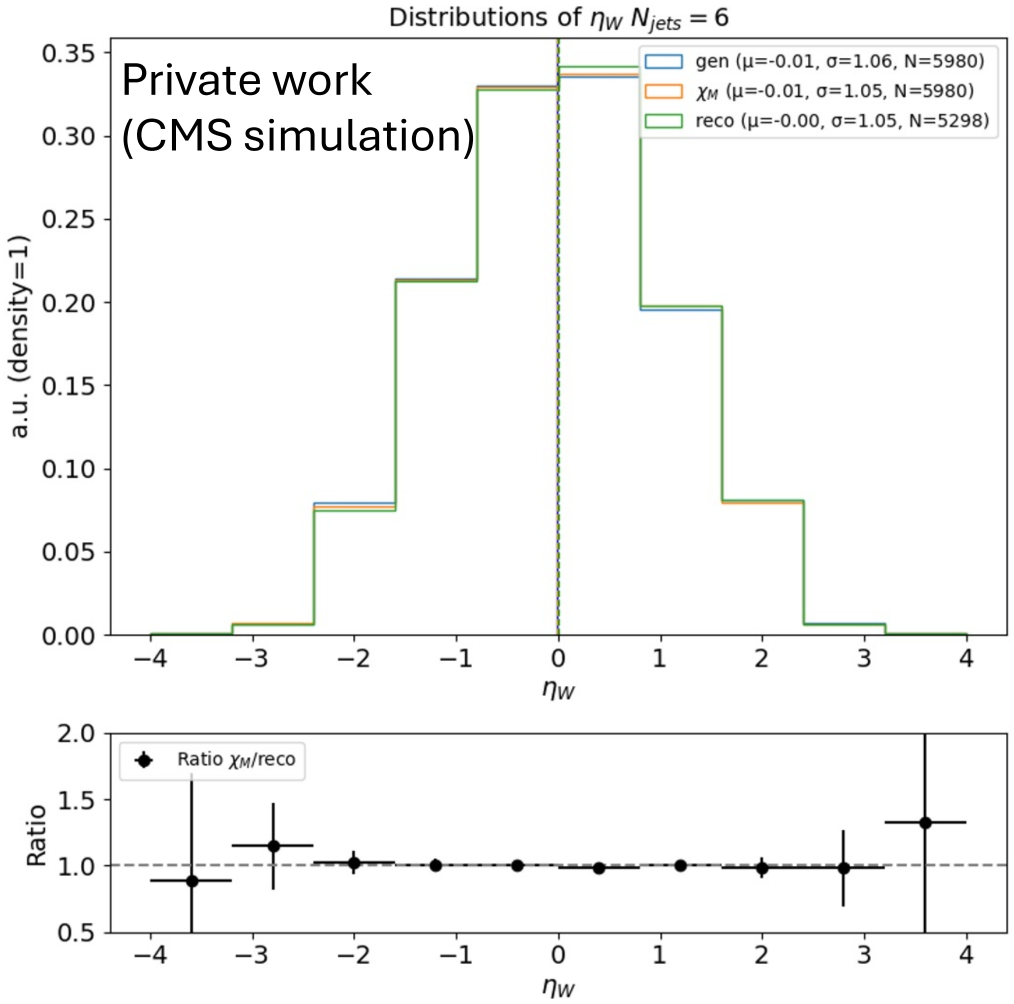


Momentum + Mass

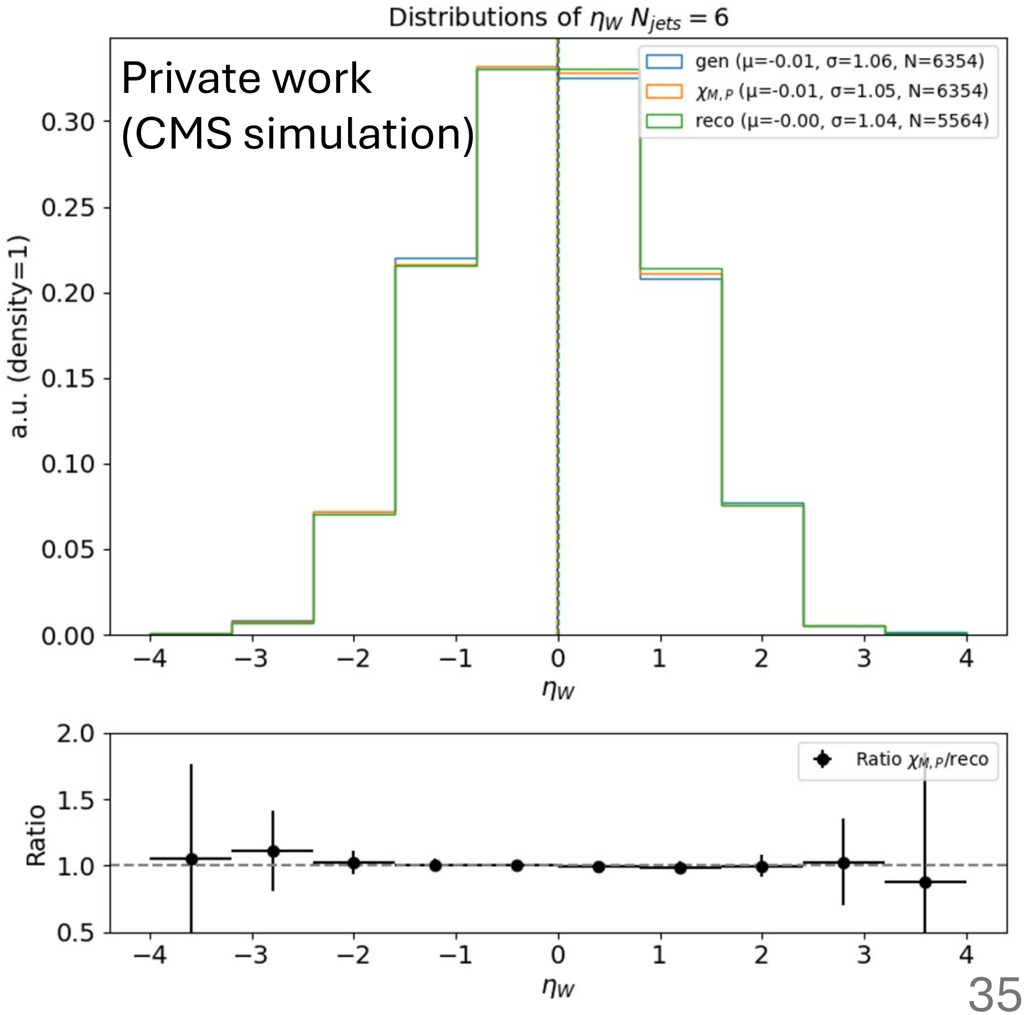


Reconstructed Kinematics: W eta, 6 jets

Mass Only

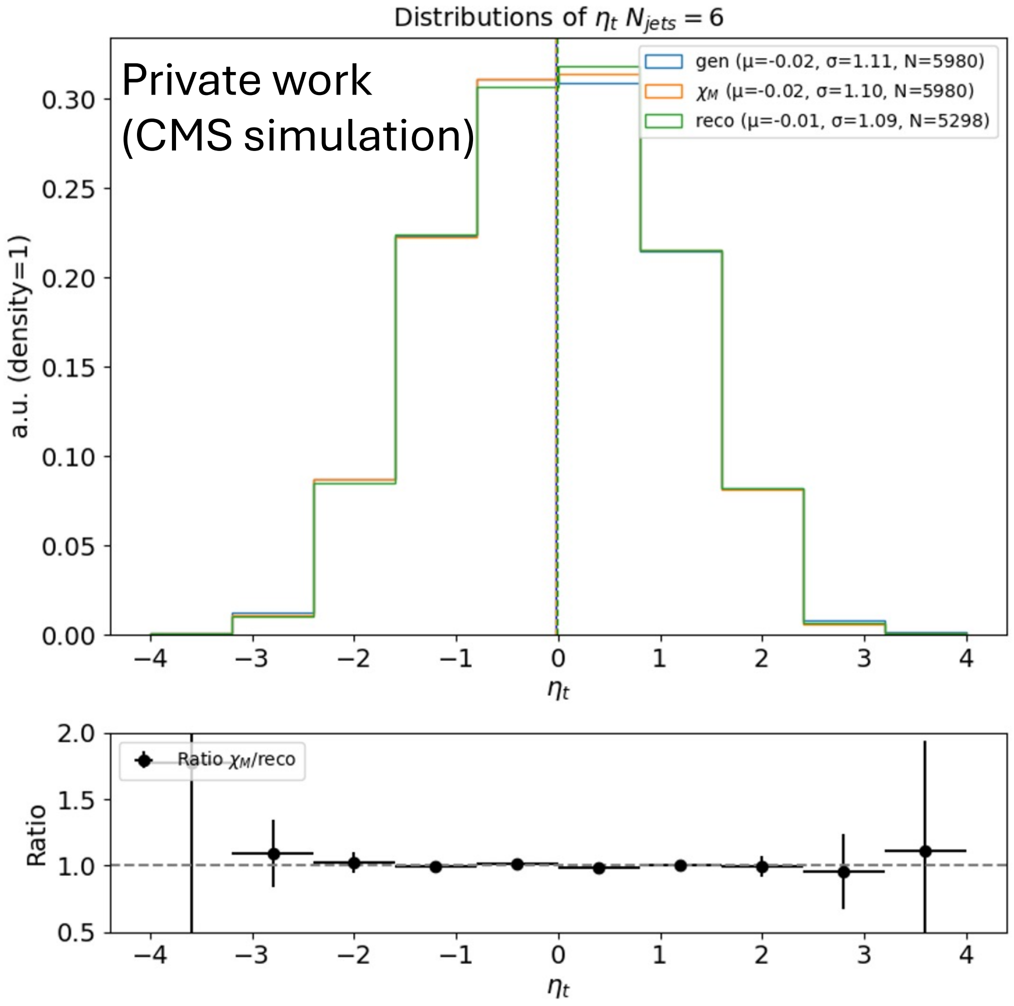


Momentum + Mass

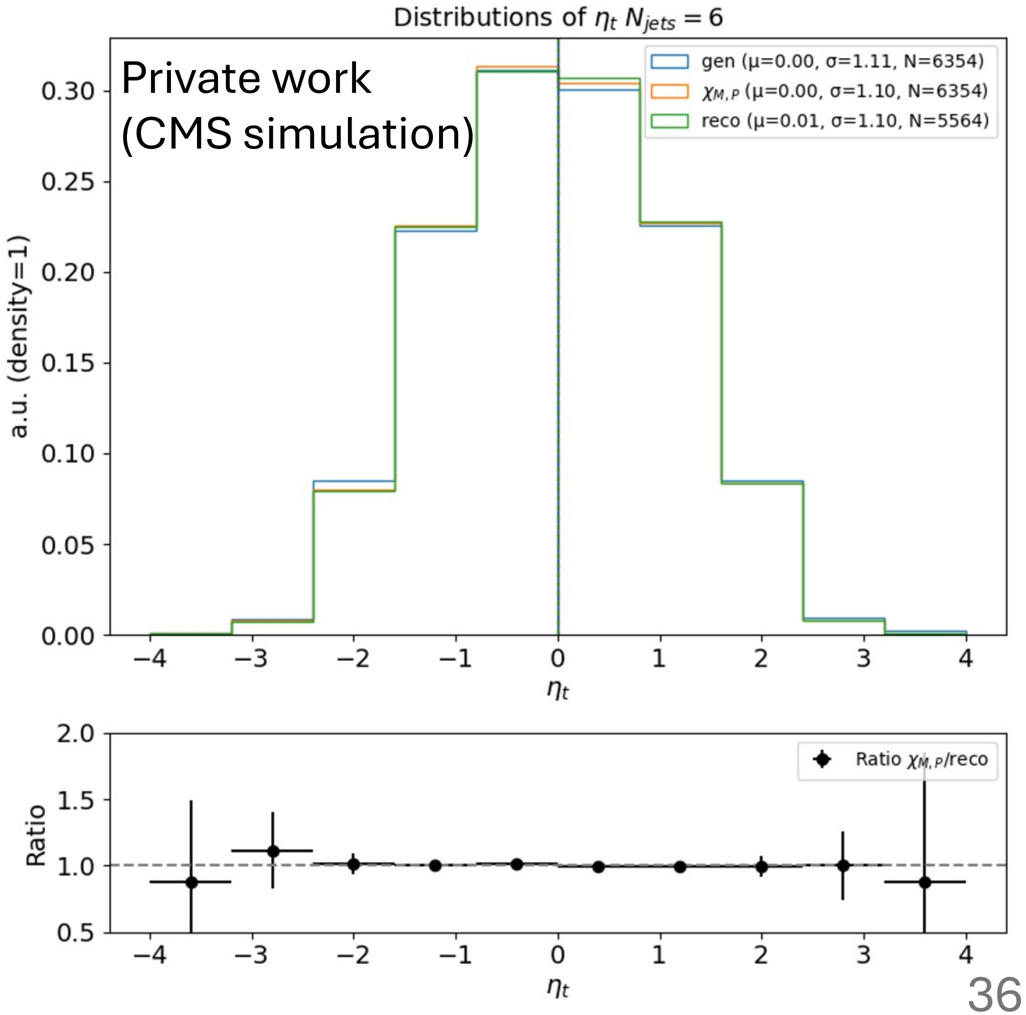


Reconstructed Kinematics: t eta, 6 jets

Mass Only

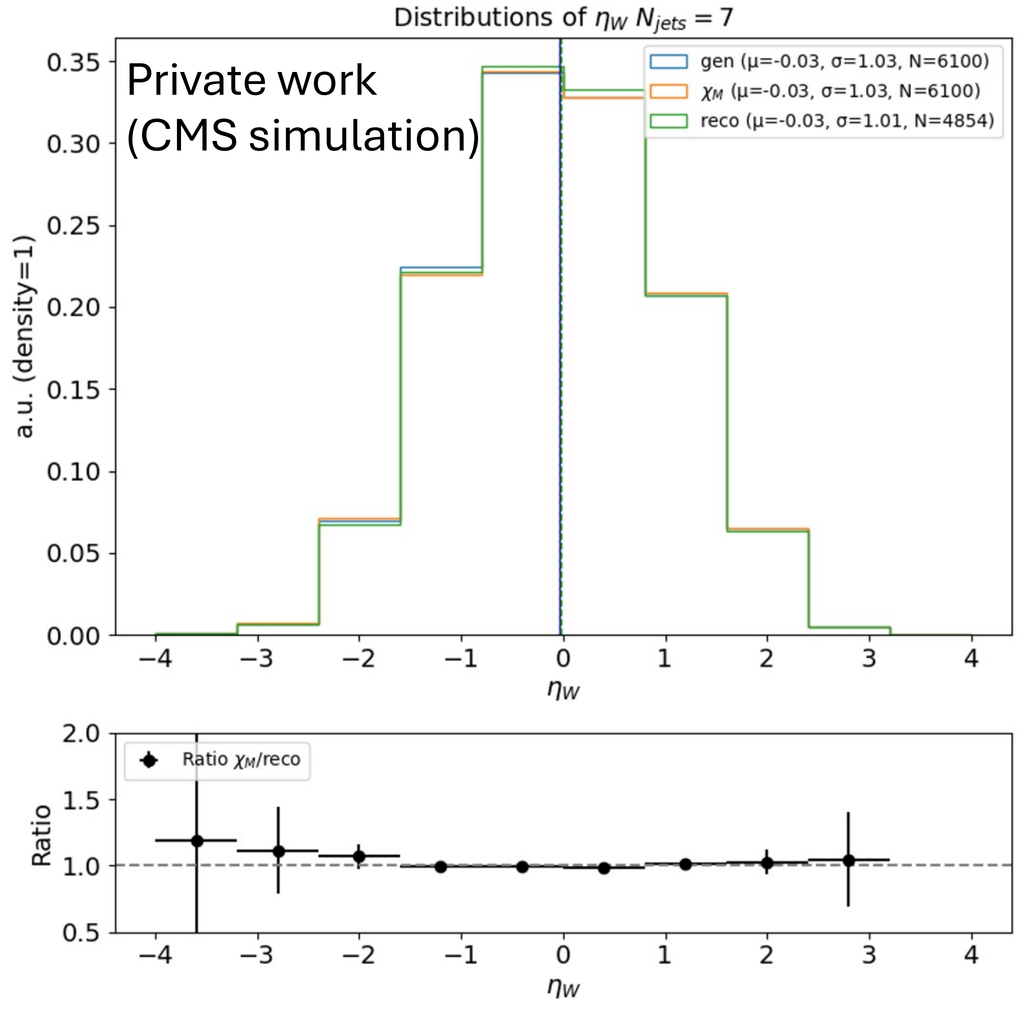


Momentum + Mass

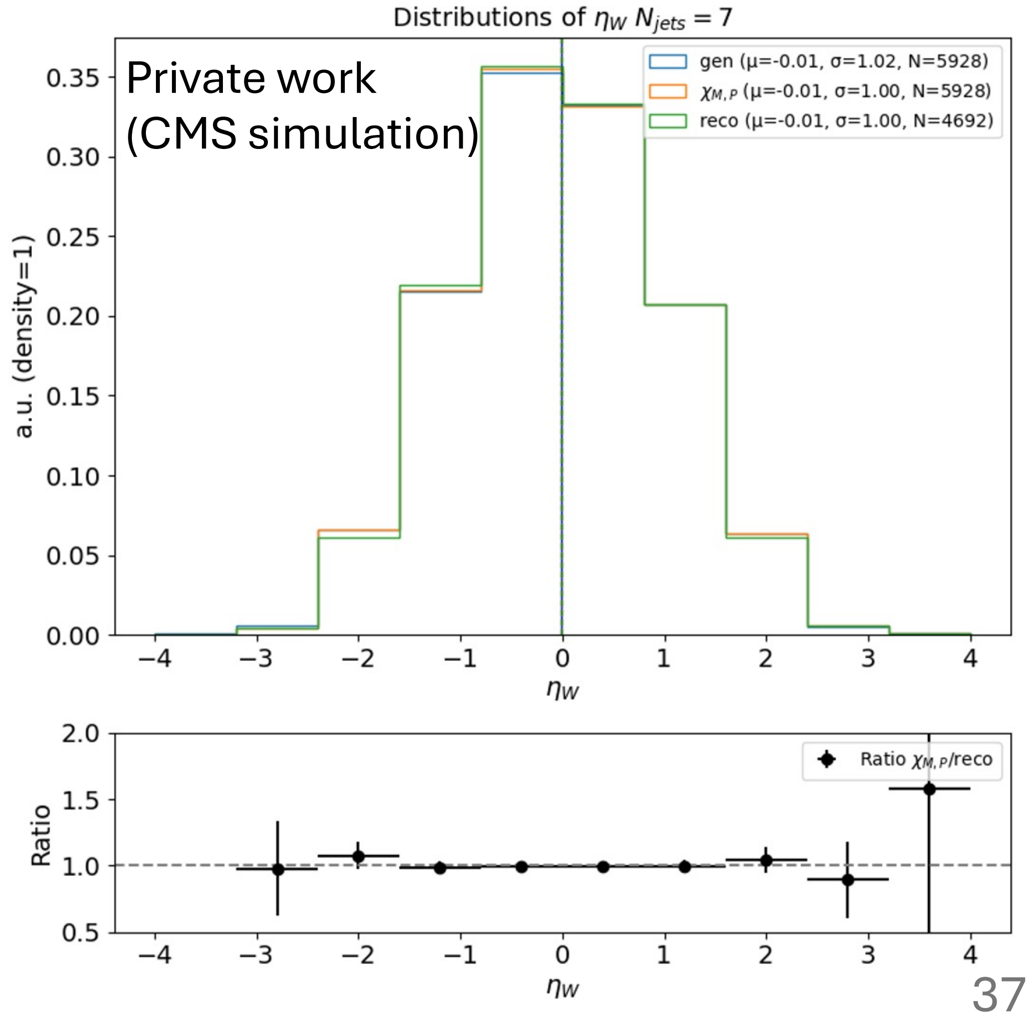


Reconstructed Kinematics: W eta, 7 jets

Mass Only

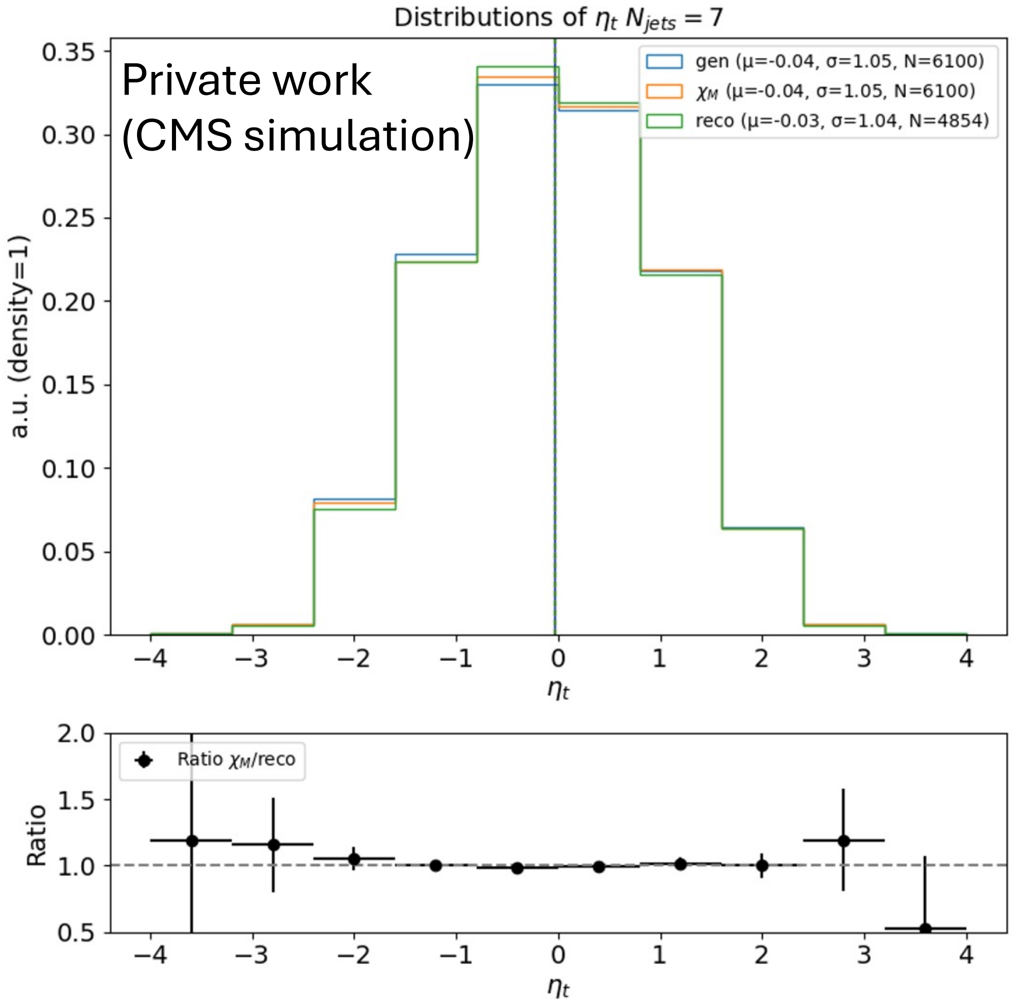


Momentum + Mass

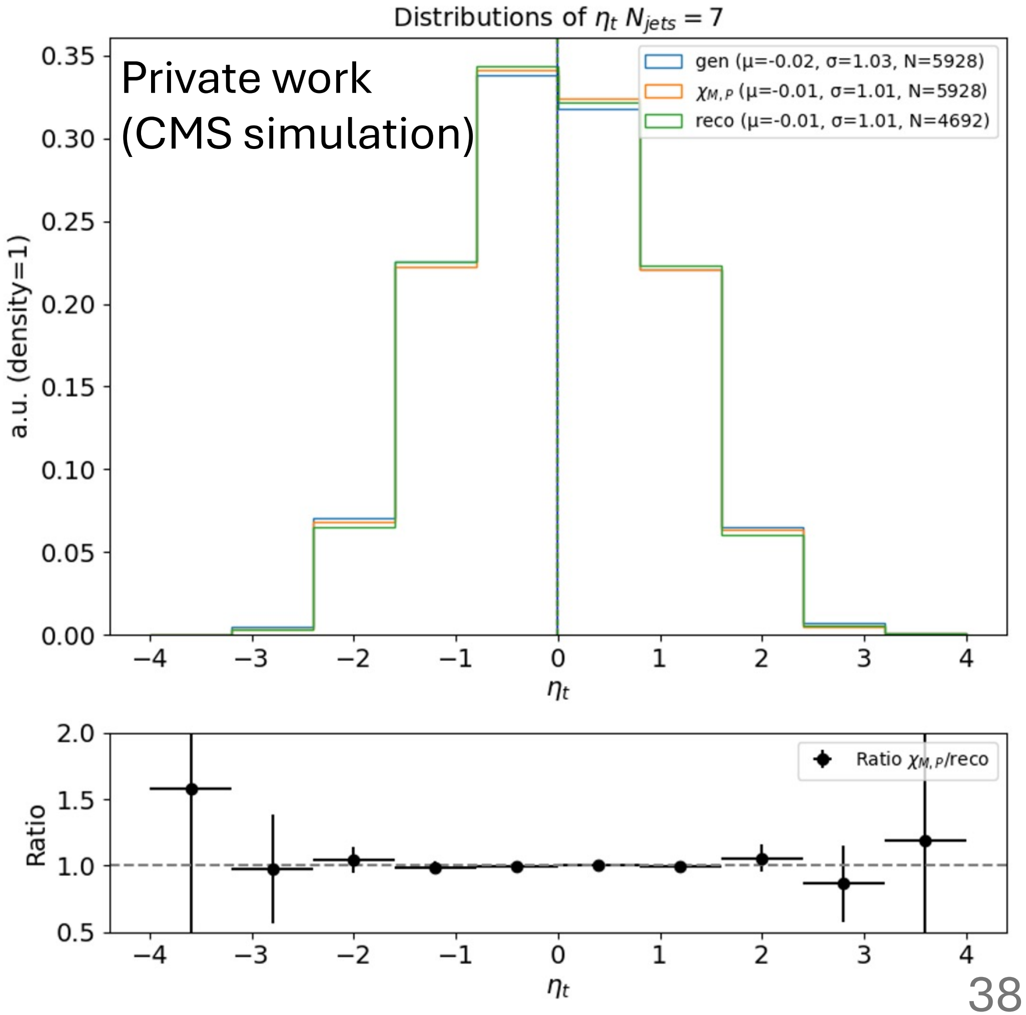


Reconstructed Kinematics: t eta, 7 jets

Mass Only

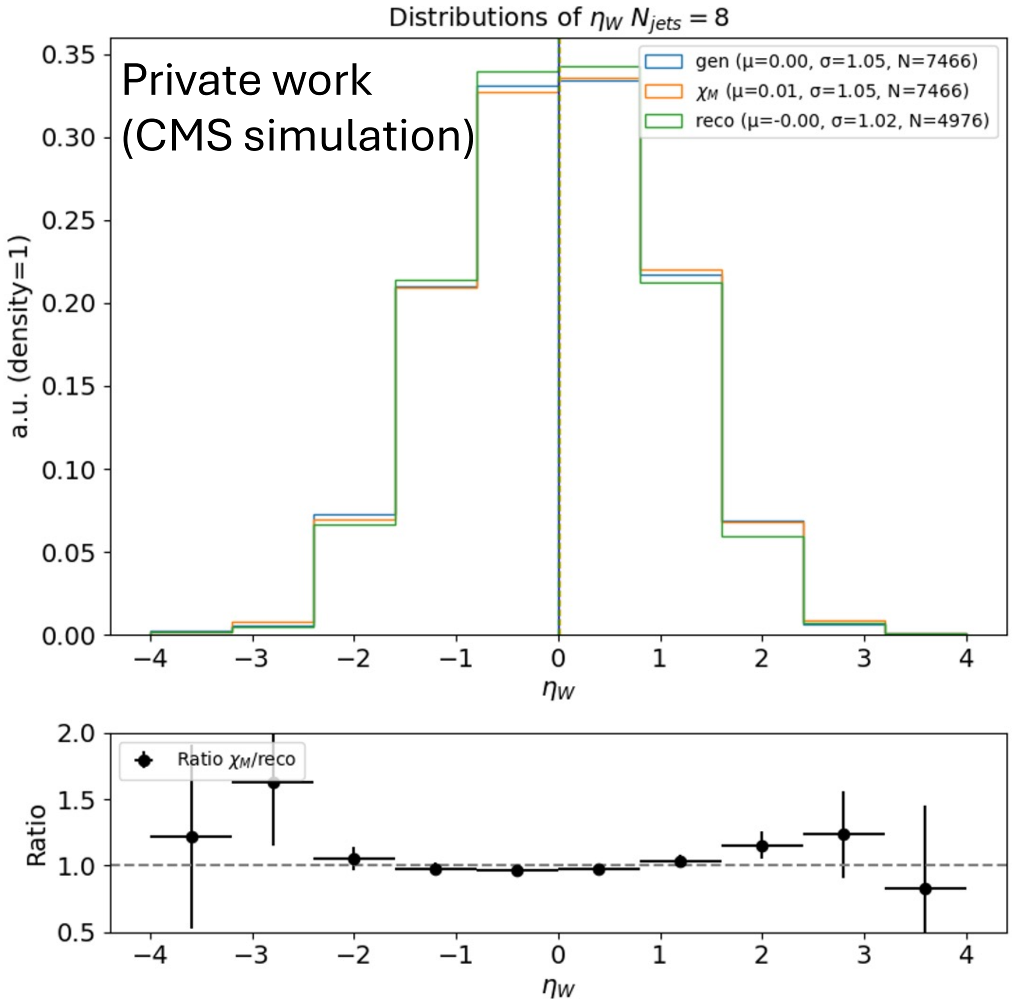


Momentum + Mass

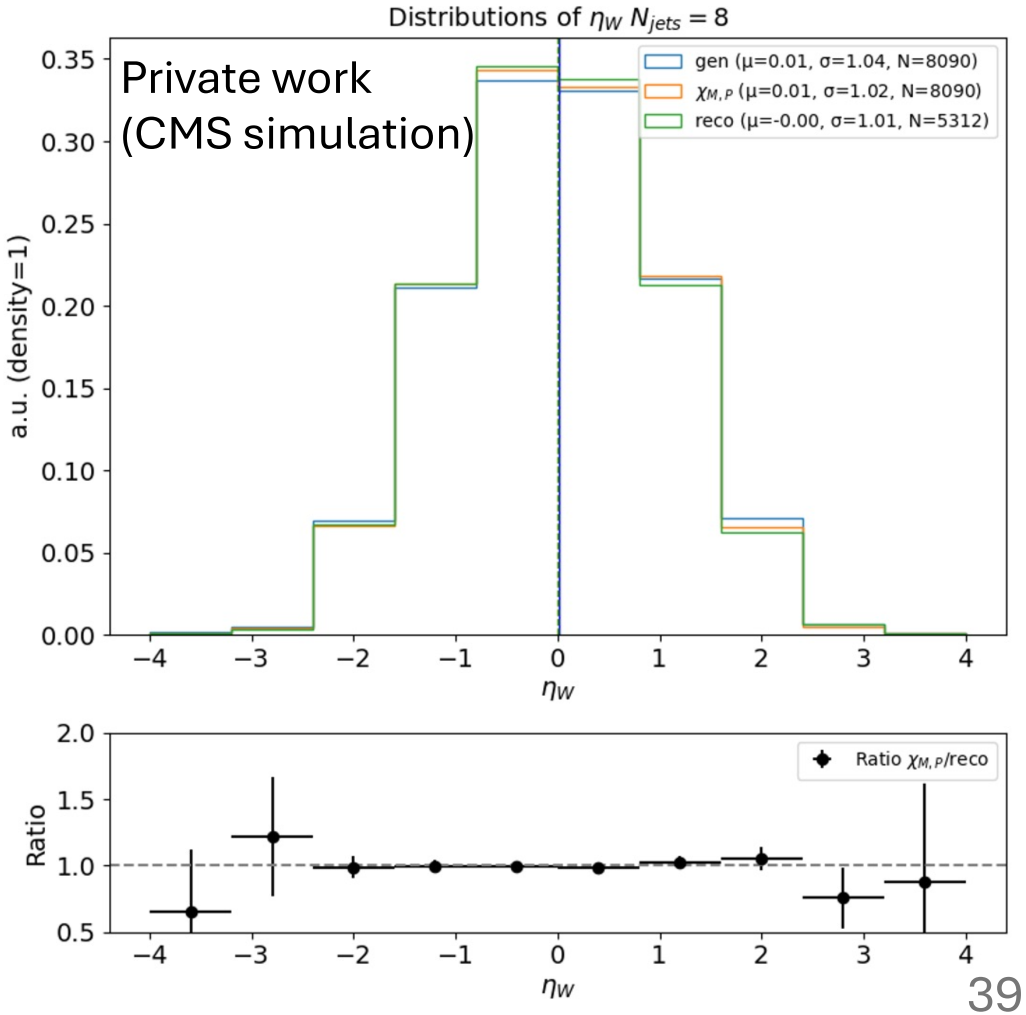


Reconstructed Kinematics: W eta, 8 jets

Mass Only

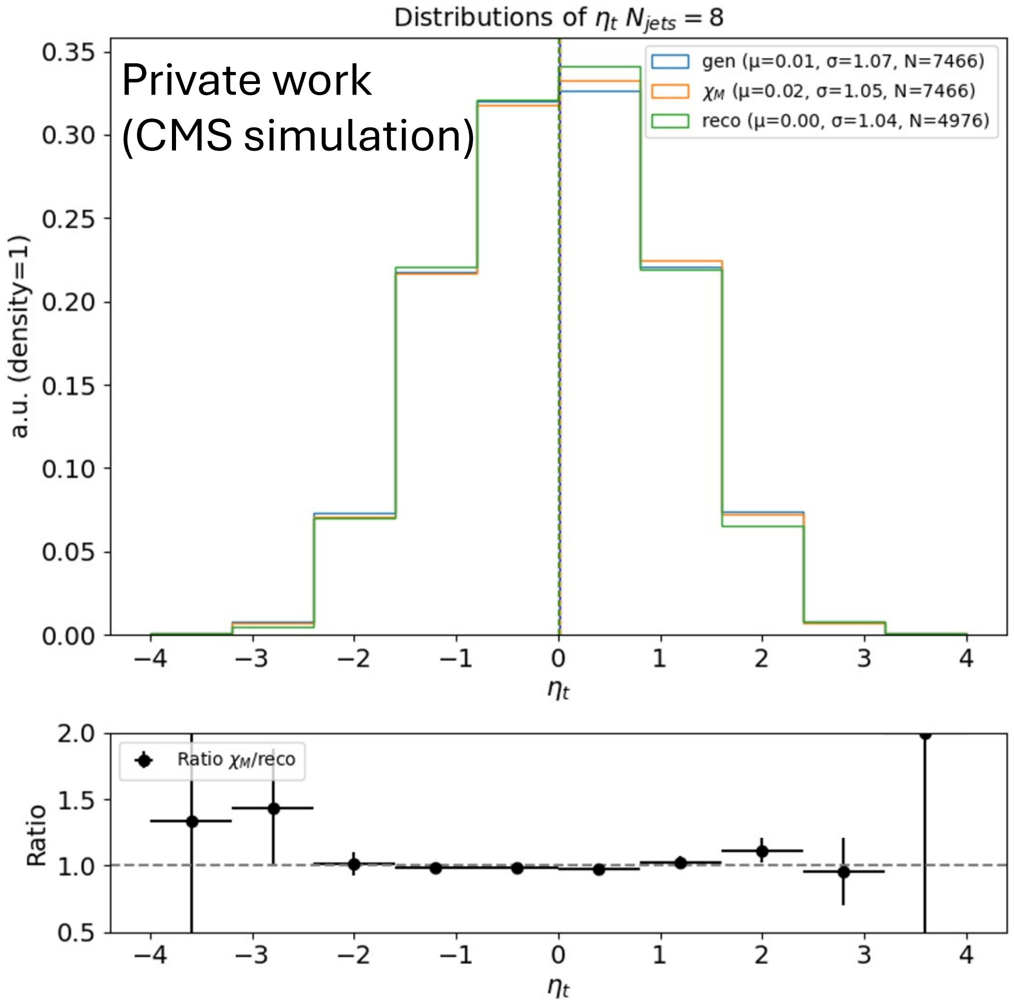


Momentum + Mass

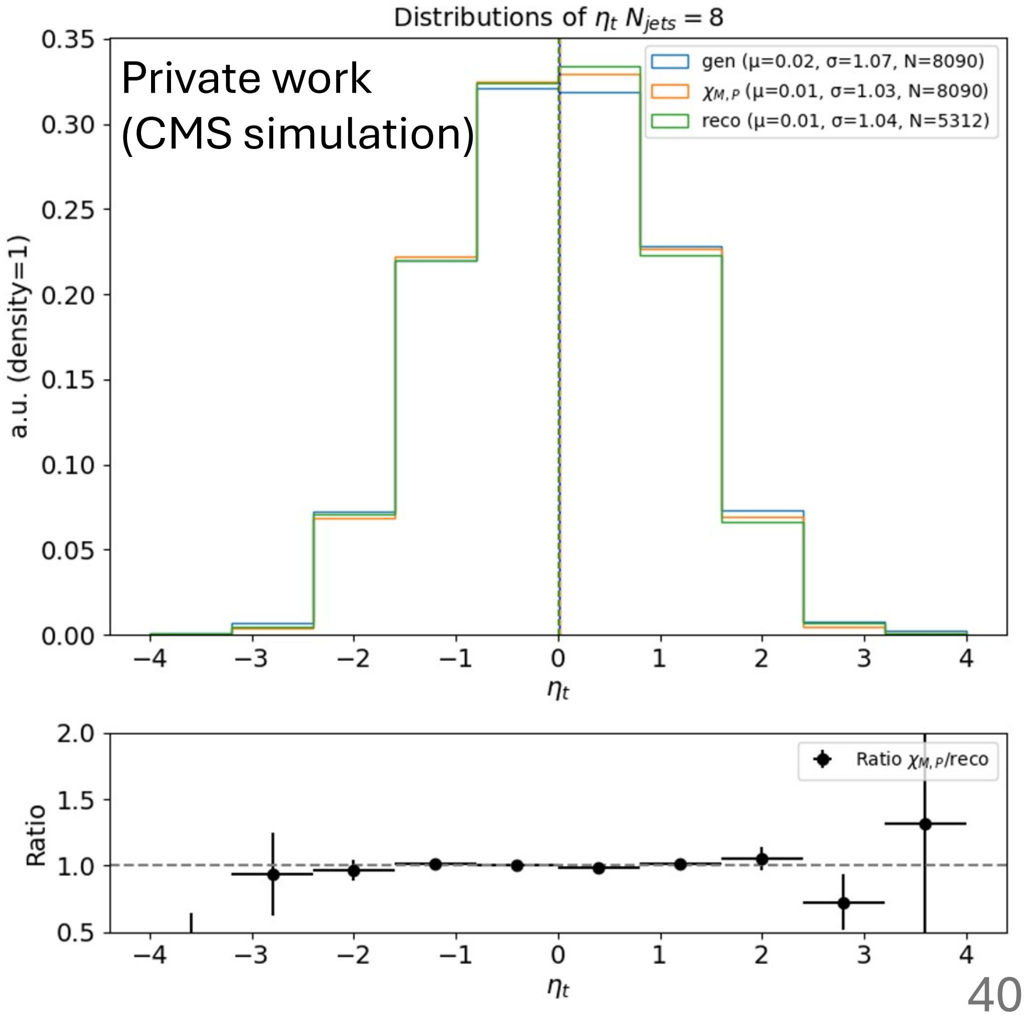


Reconstructed Kinematics: t eta, 8 jets

Mass Only

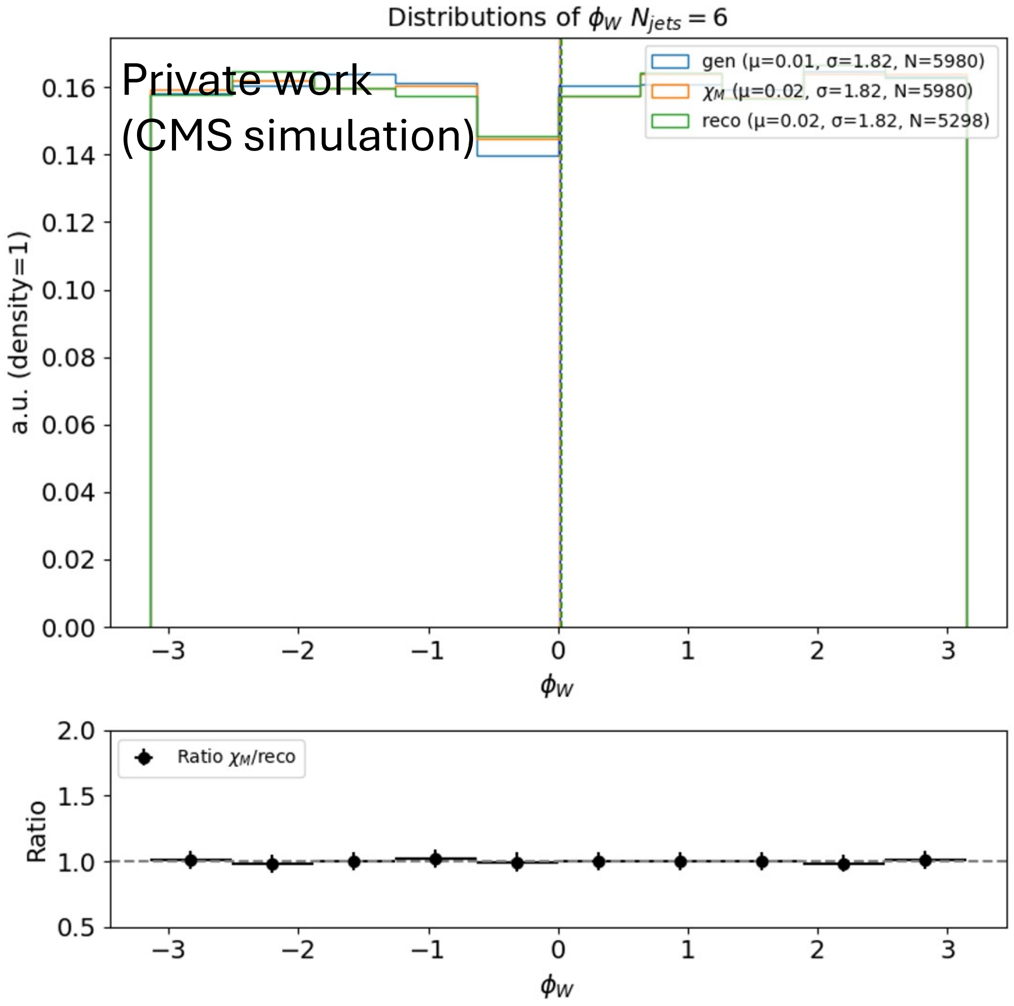


Momentum + Mass

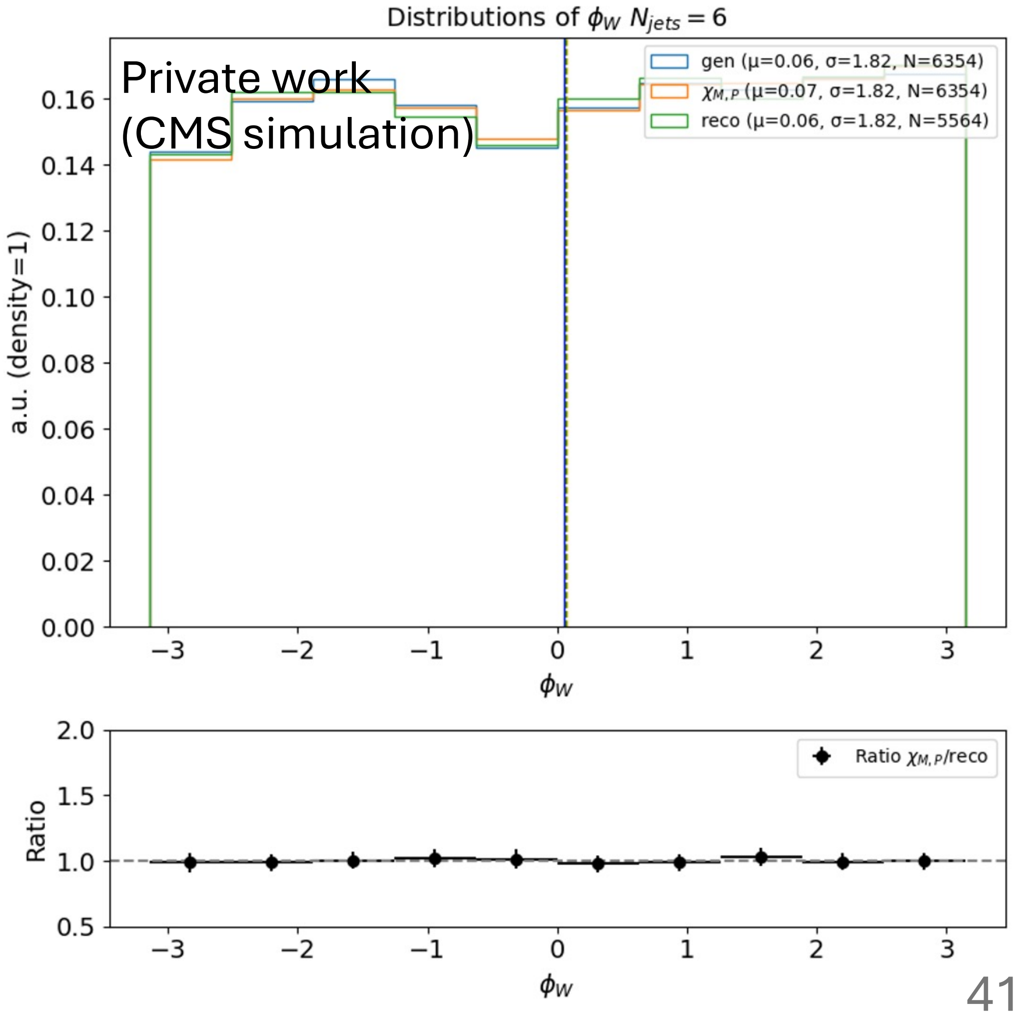


Reconstructed Kinematics: W phi, 6 jets

Mass Only

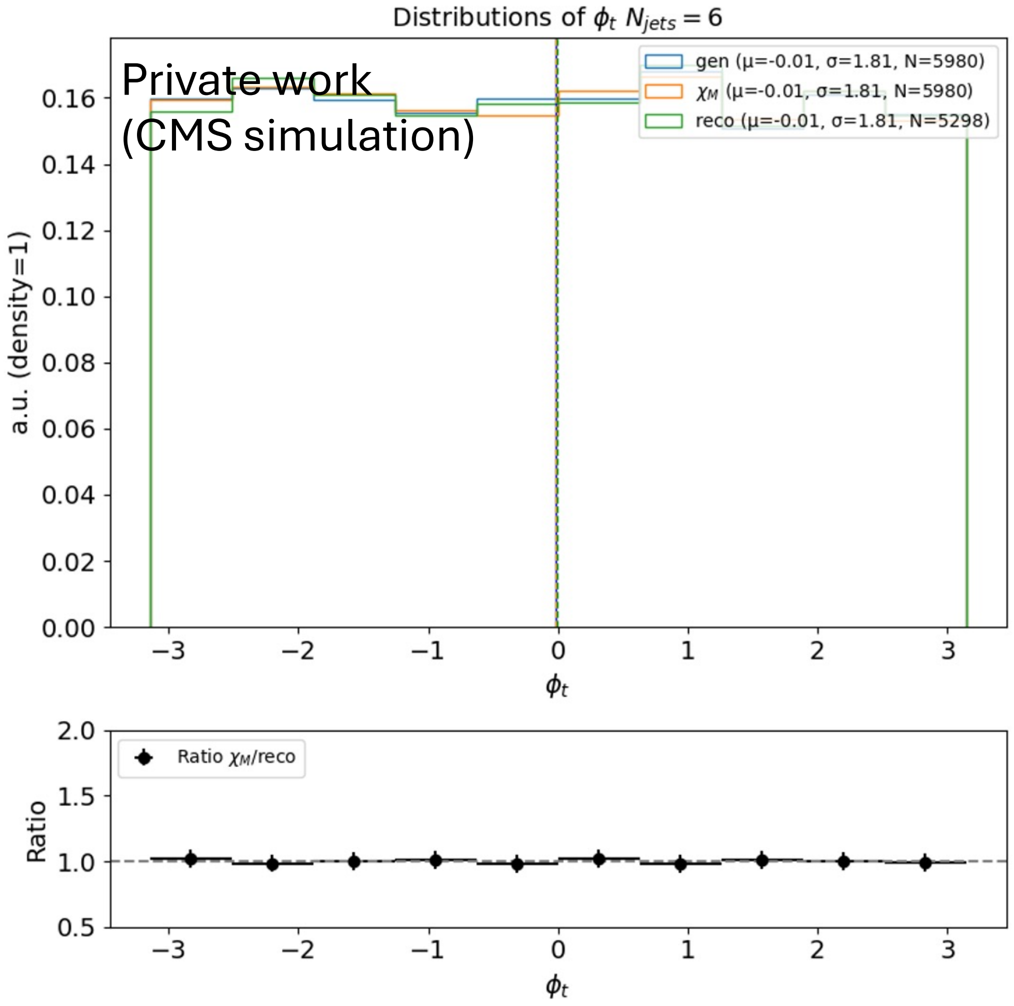


Momentum + Mass

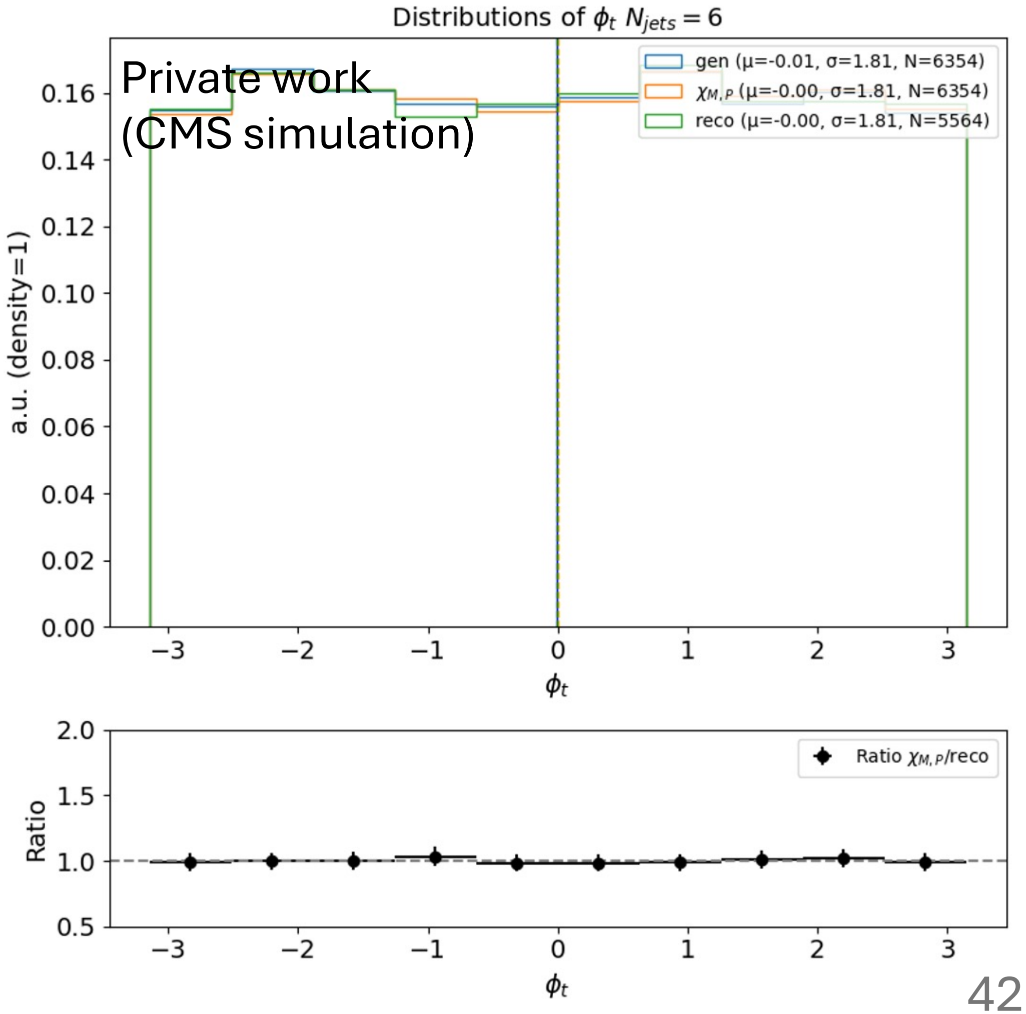


Reconstructed Kinematics: t phi, 6 jets

Mass Only

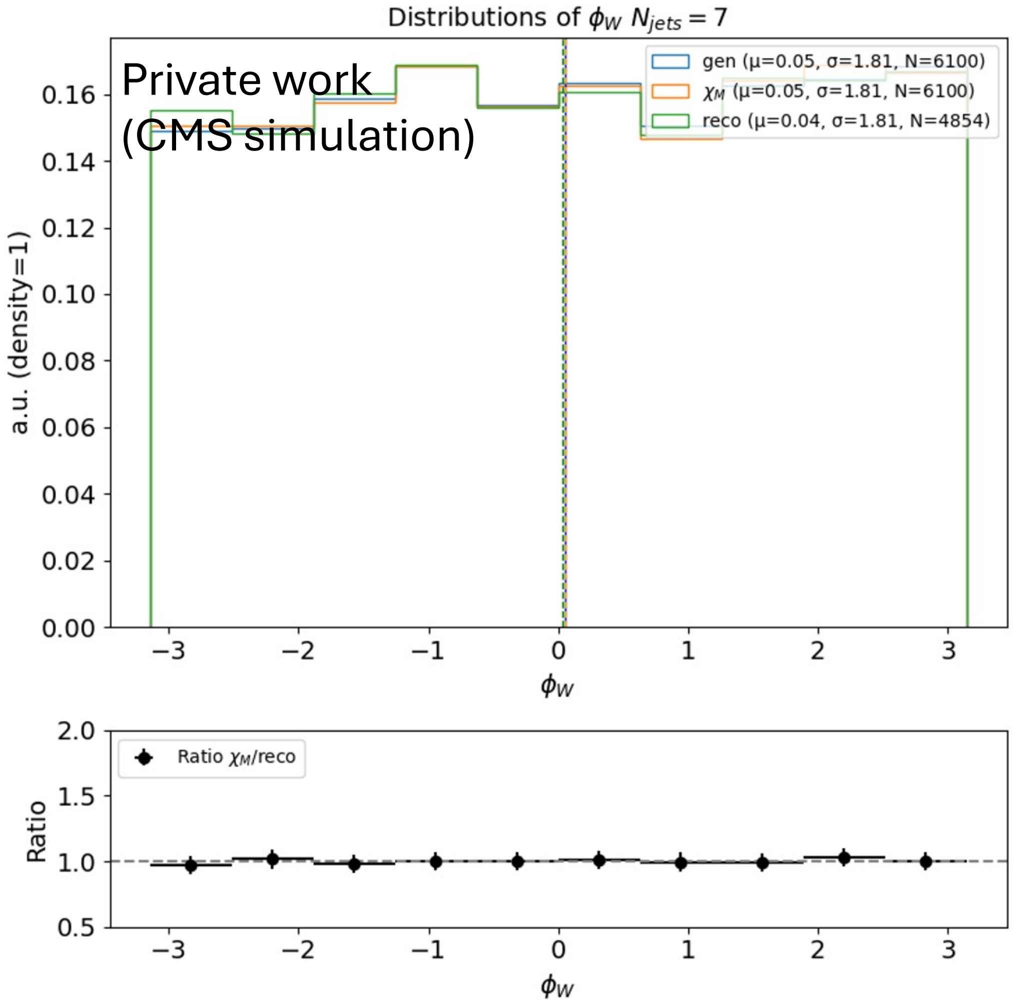


Momentum + Mass

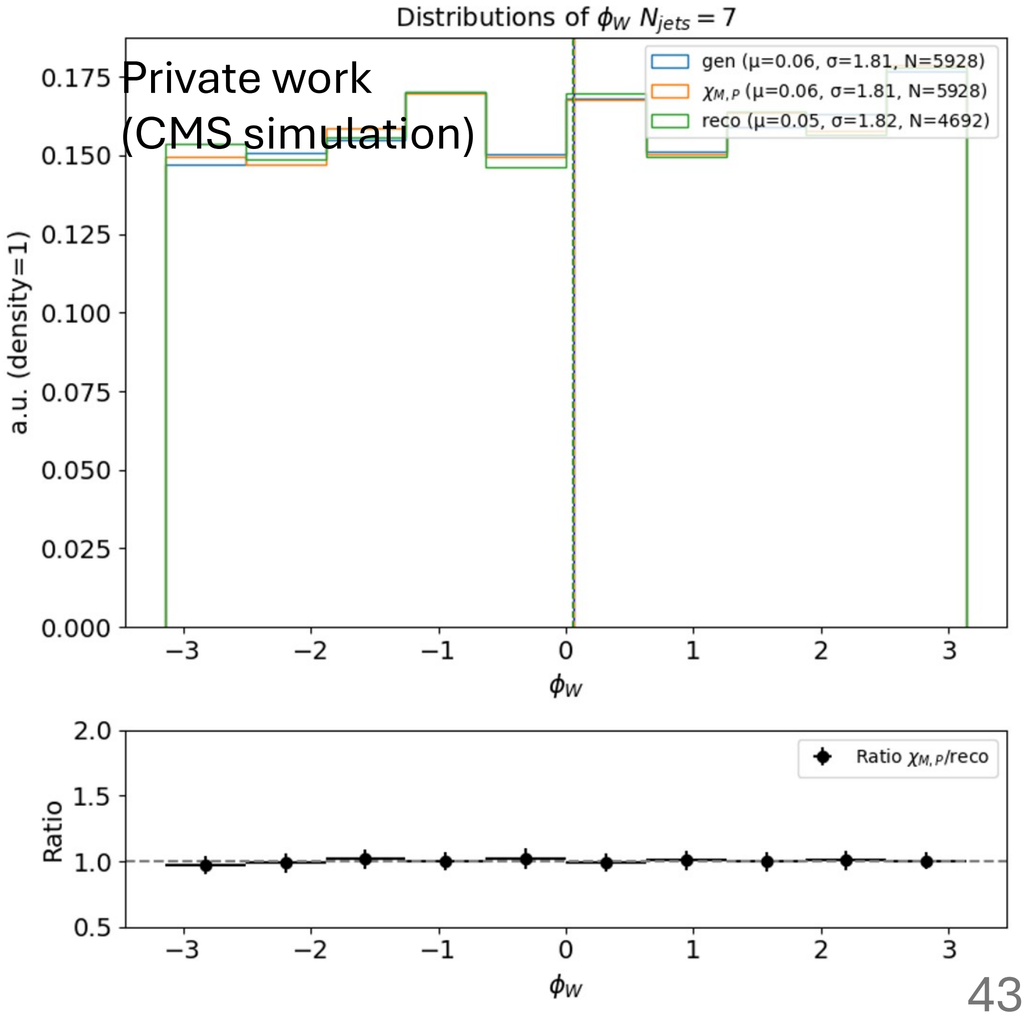


Reconstructed Kinematics: W phi, 7 jets

Mass Only

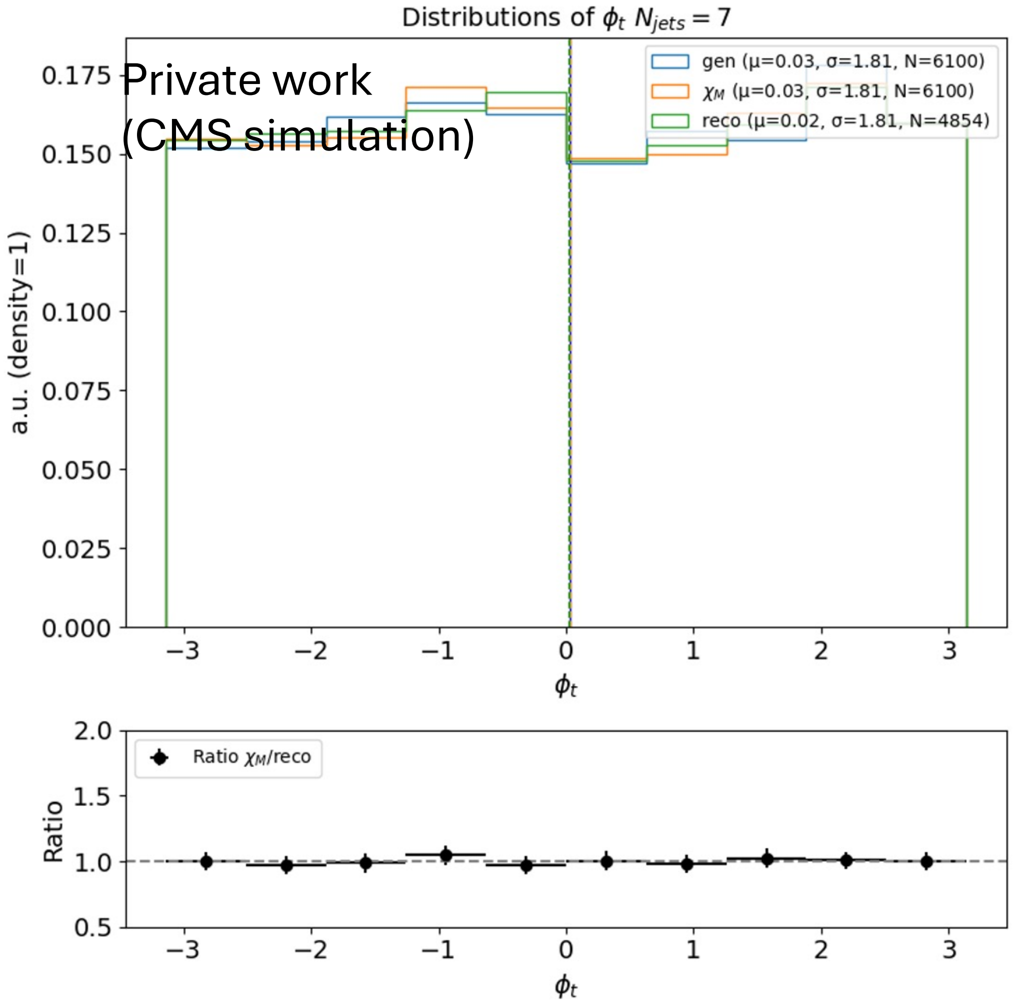


Momentum + Mass

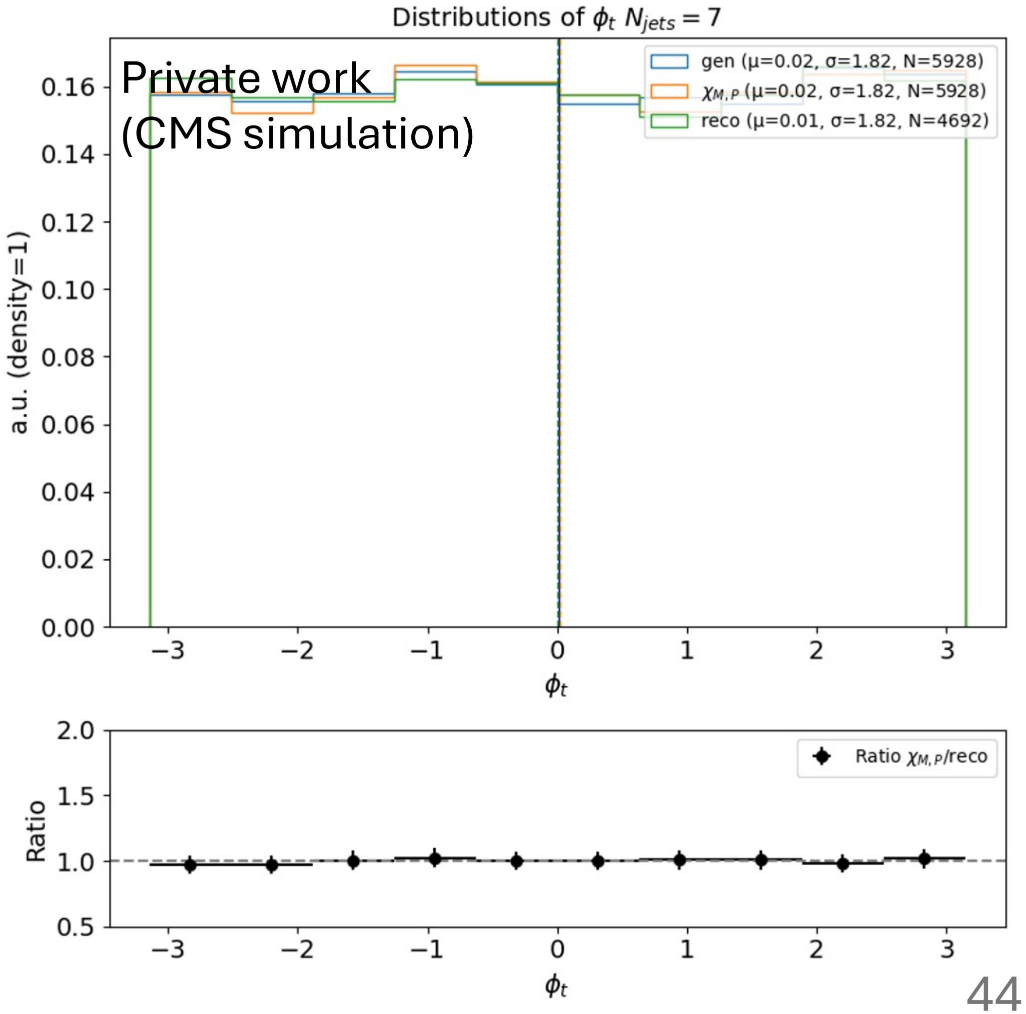


Reconstructed Kinematics: t phi, 7 jets

Mass Only

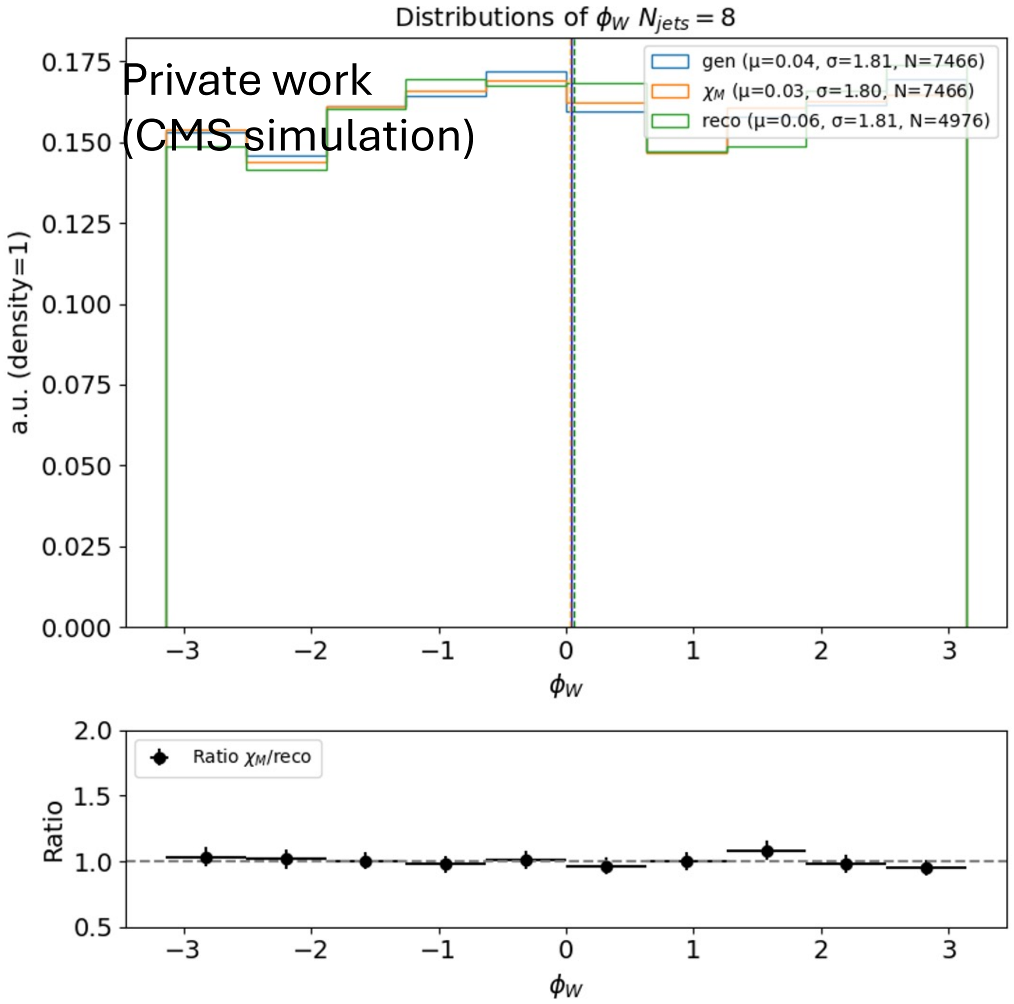


Momentum + Mass

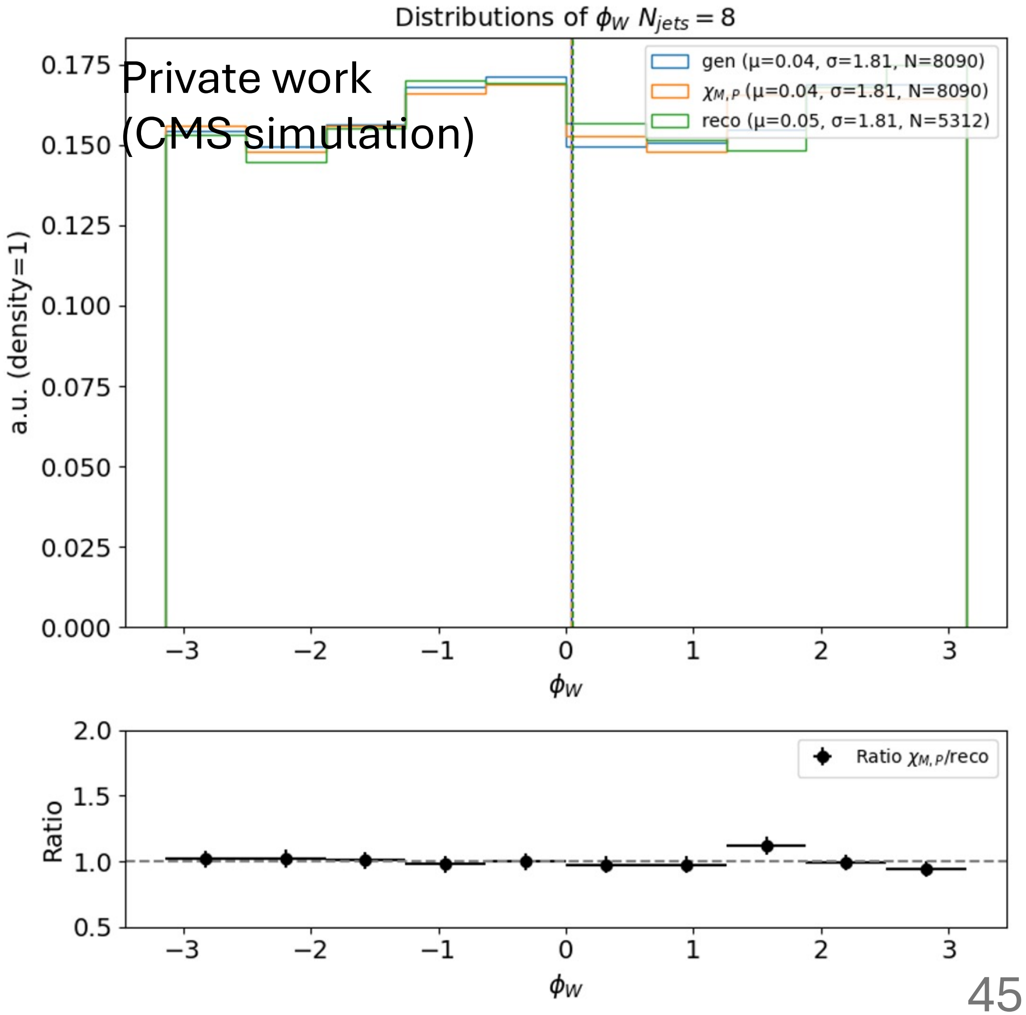


Reconstructed Kinematics: W phi, 8 jets

Mass Only

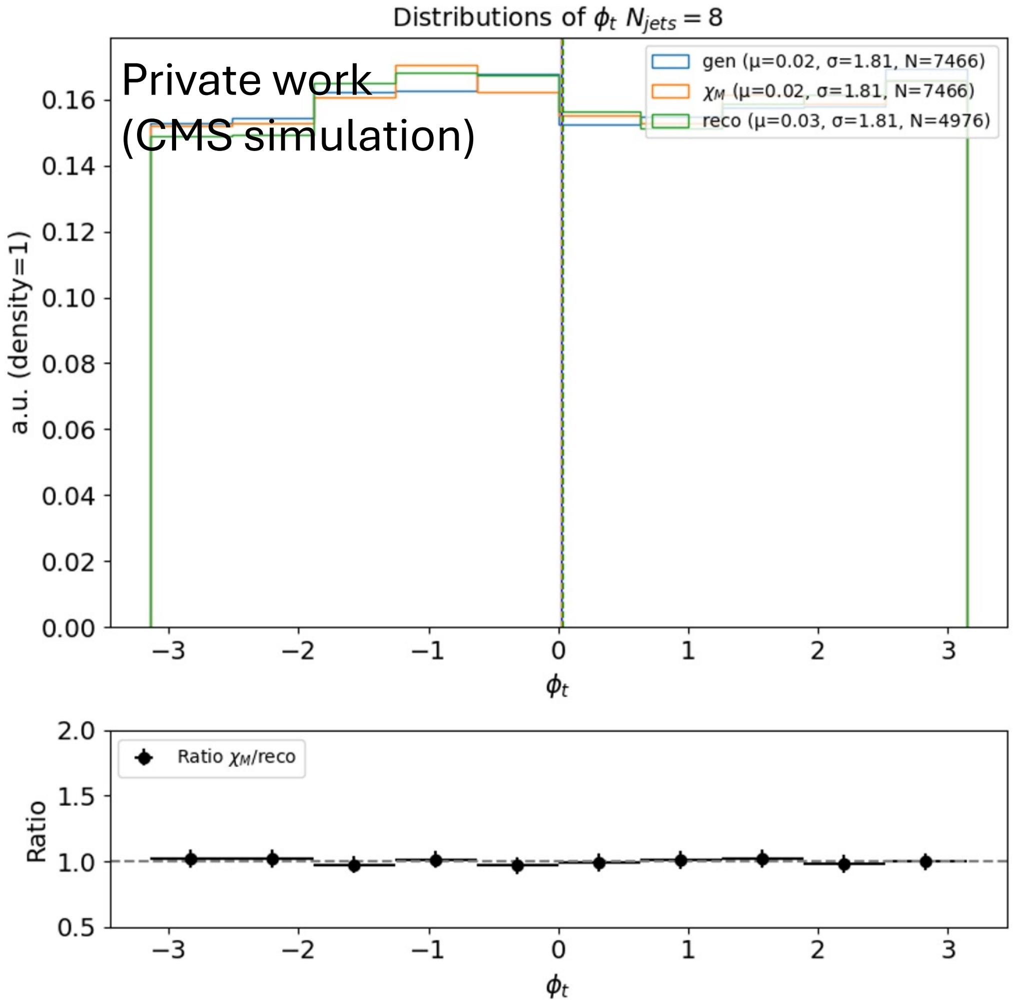


Momentum + Mass

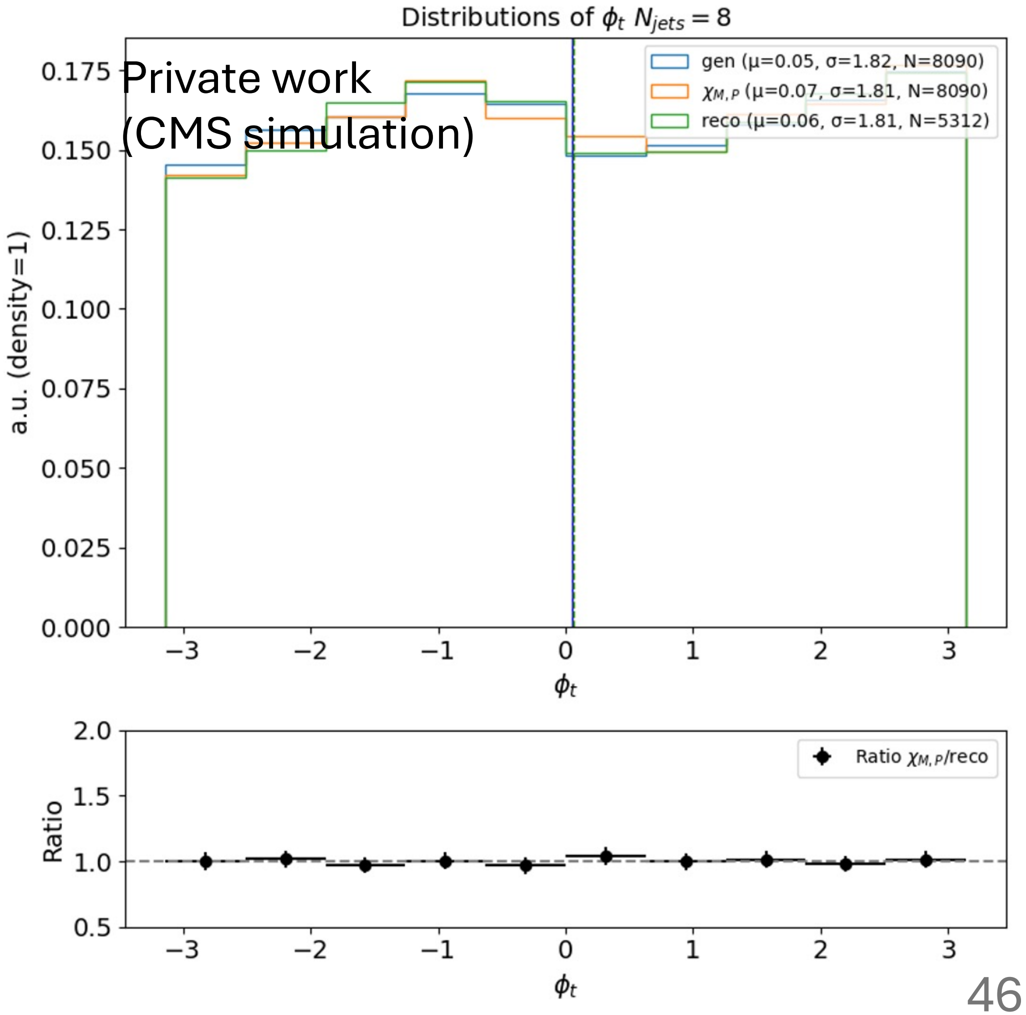


Reconstructed Kinematics: t phi, 8 jets

Mass Only

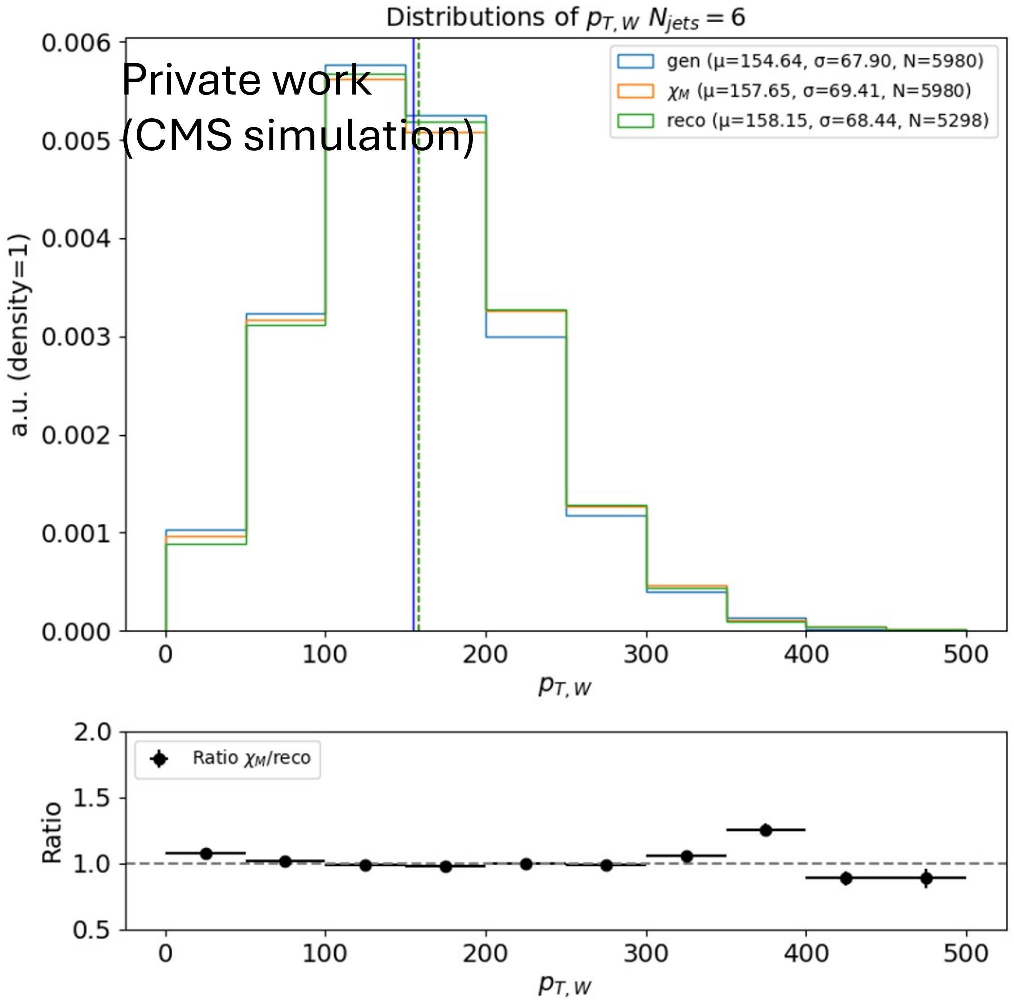


Momentum + Mass

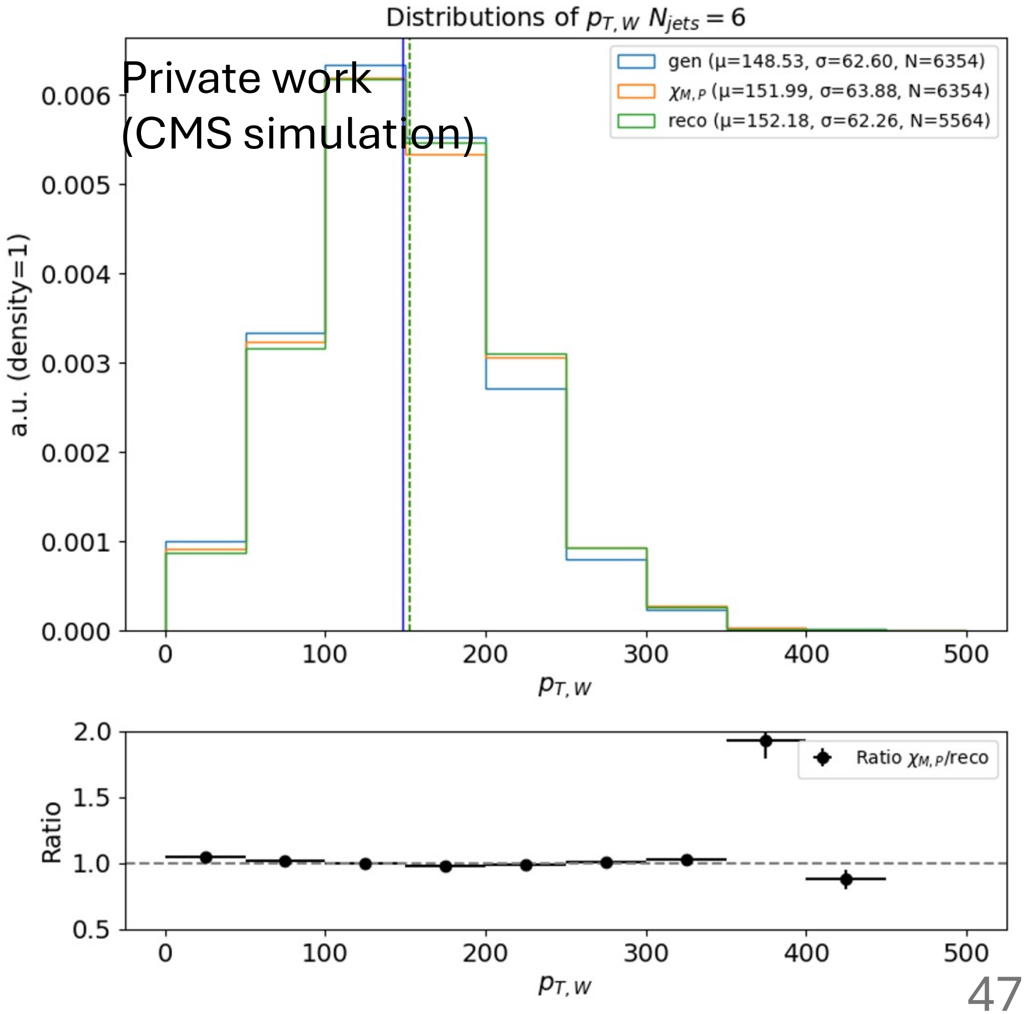


Reconstructed Kinematics: W pT, 6 jets

Mass Only

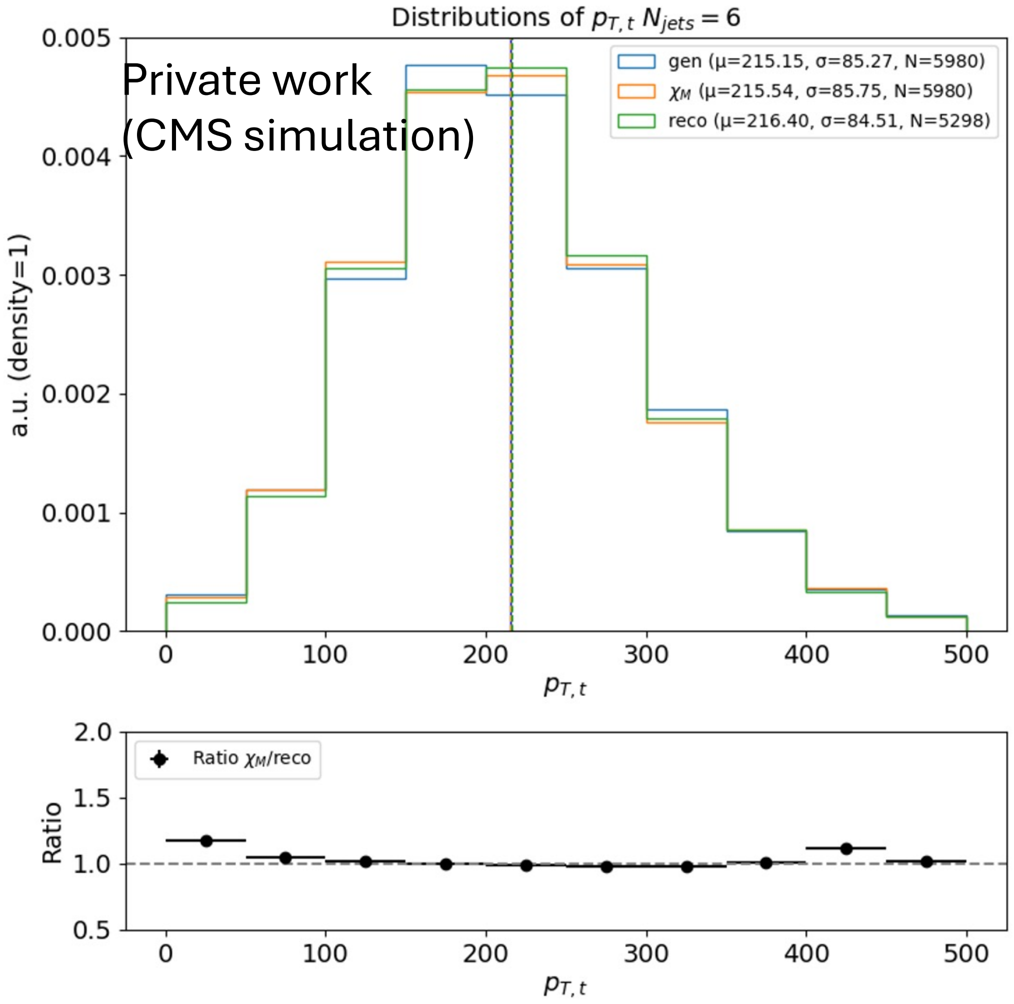


Momentum + Mass

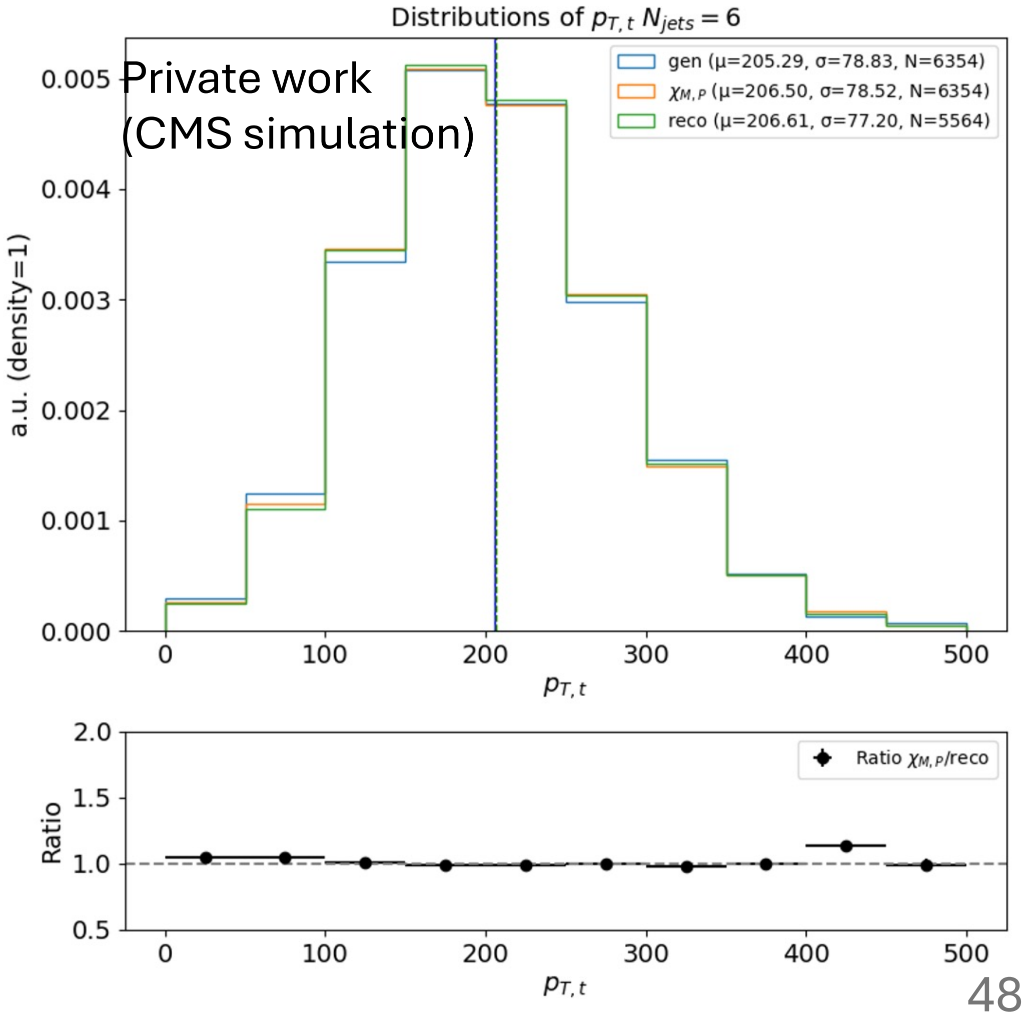


Reconstructed Kinematics: t pT, 6 jets

Mass Only

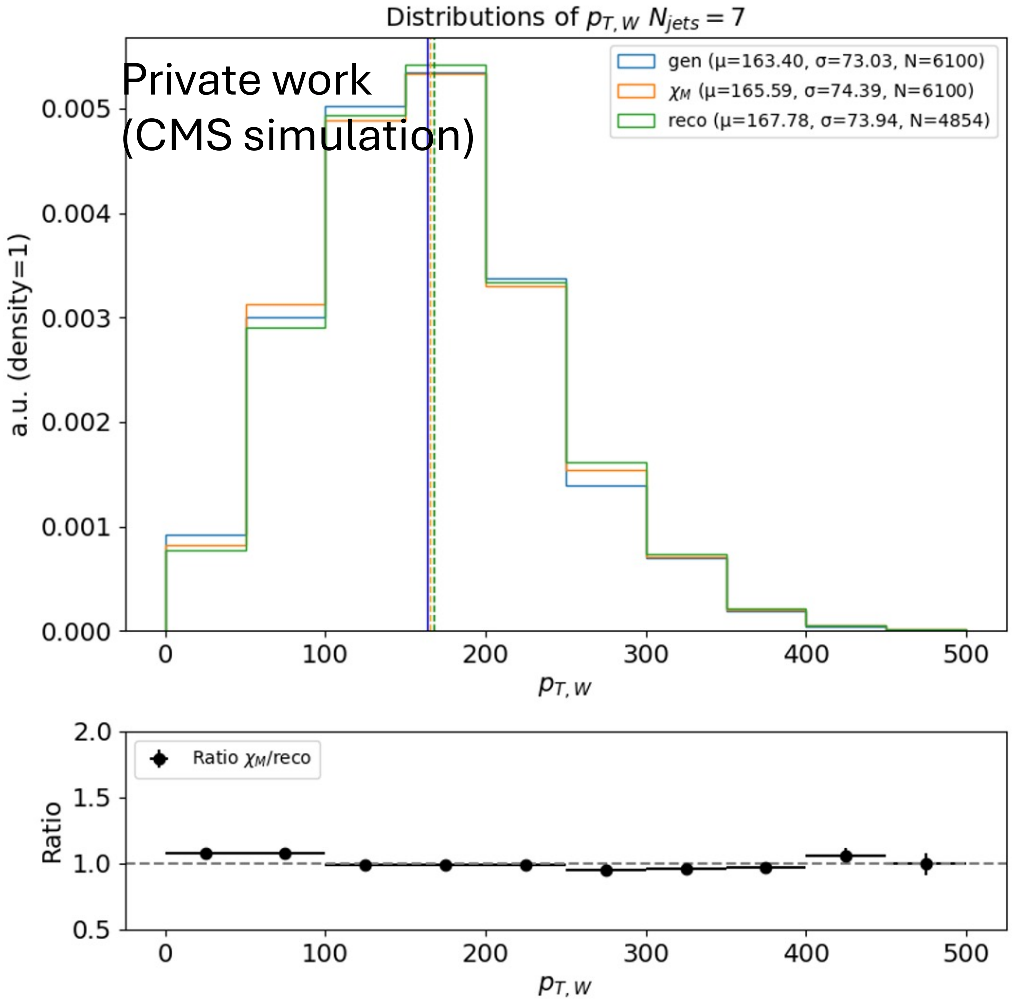


Momentum + Mass

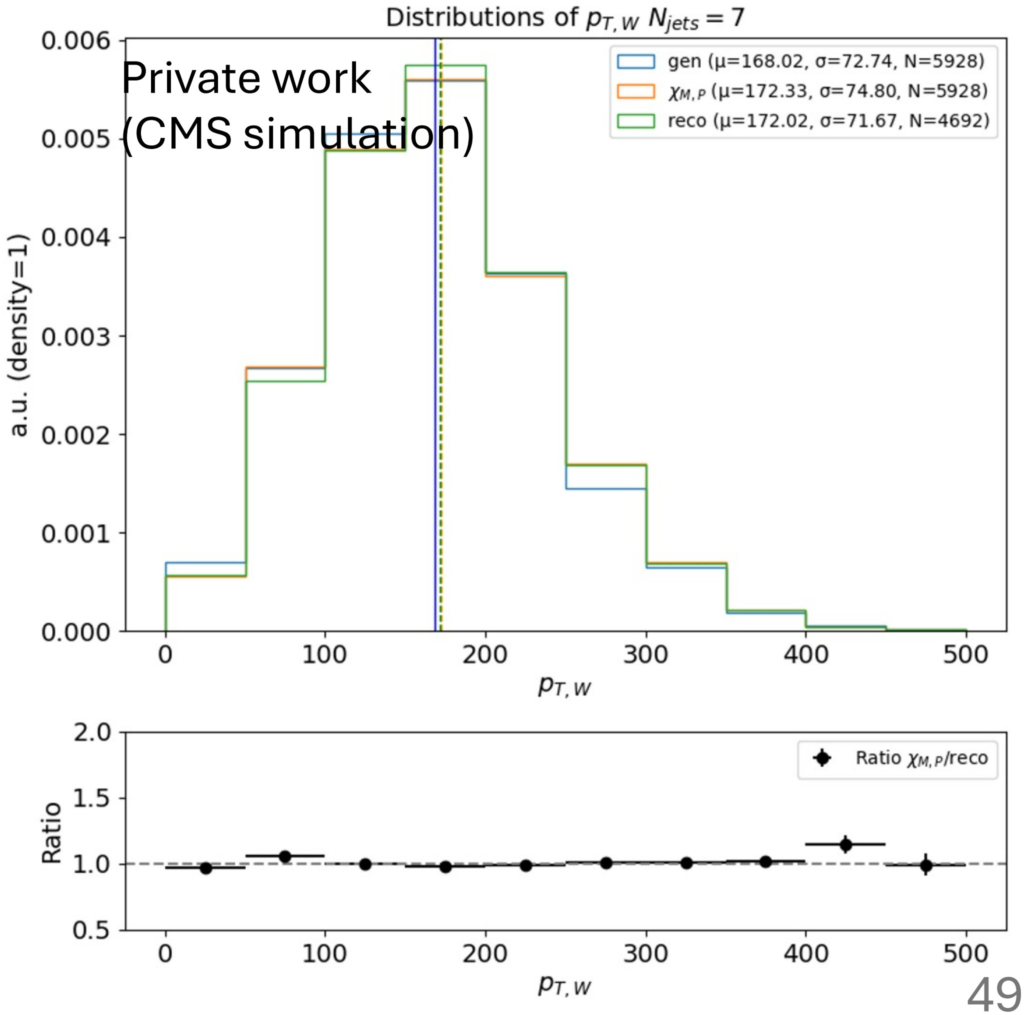


Reconstructed Kinematics: W pT, 7 jets

Mass Only

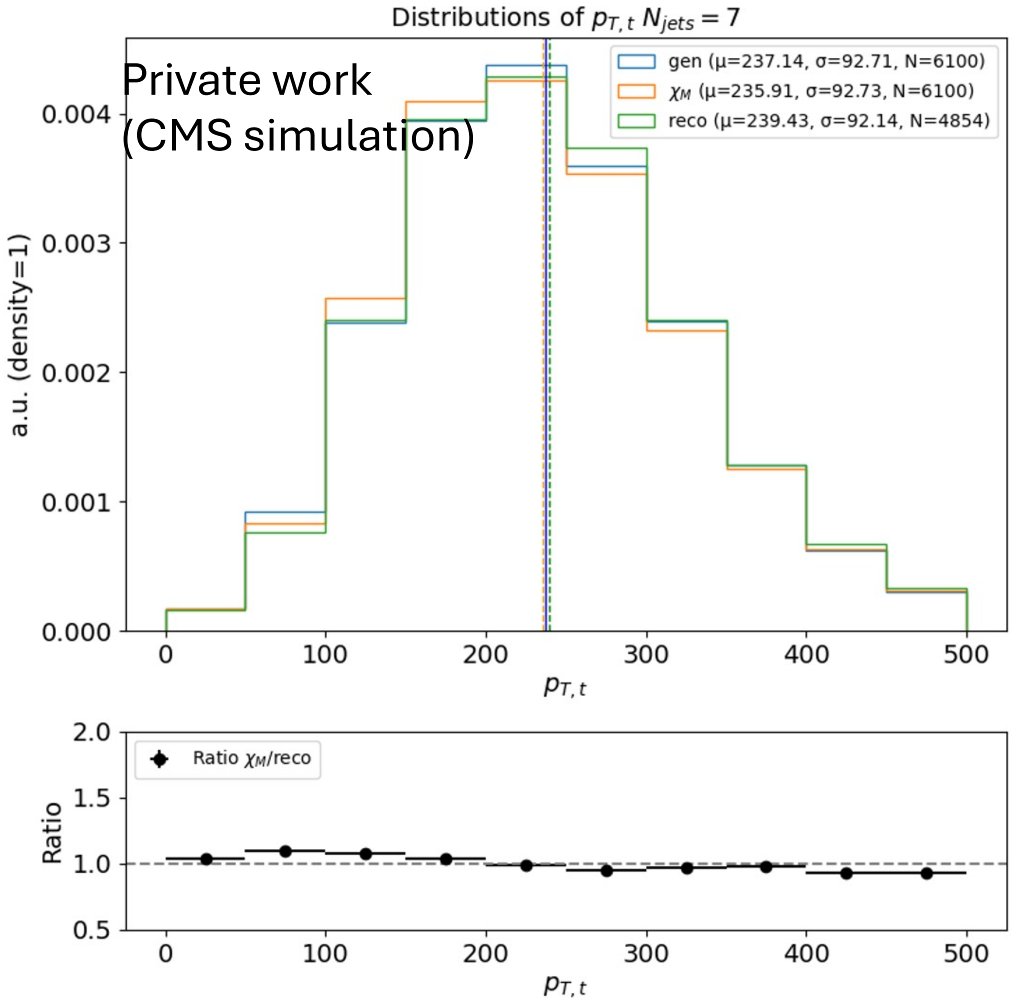


Momentum + Mass

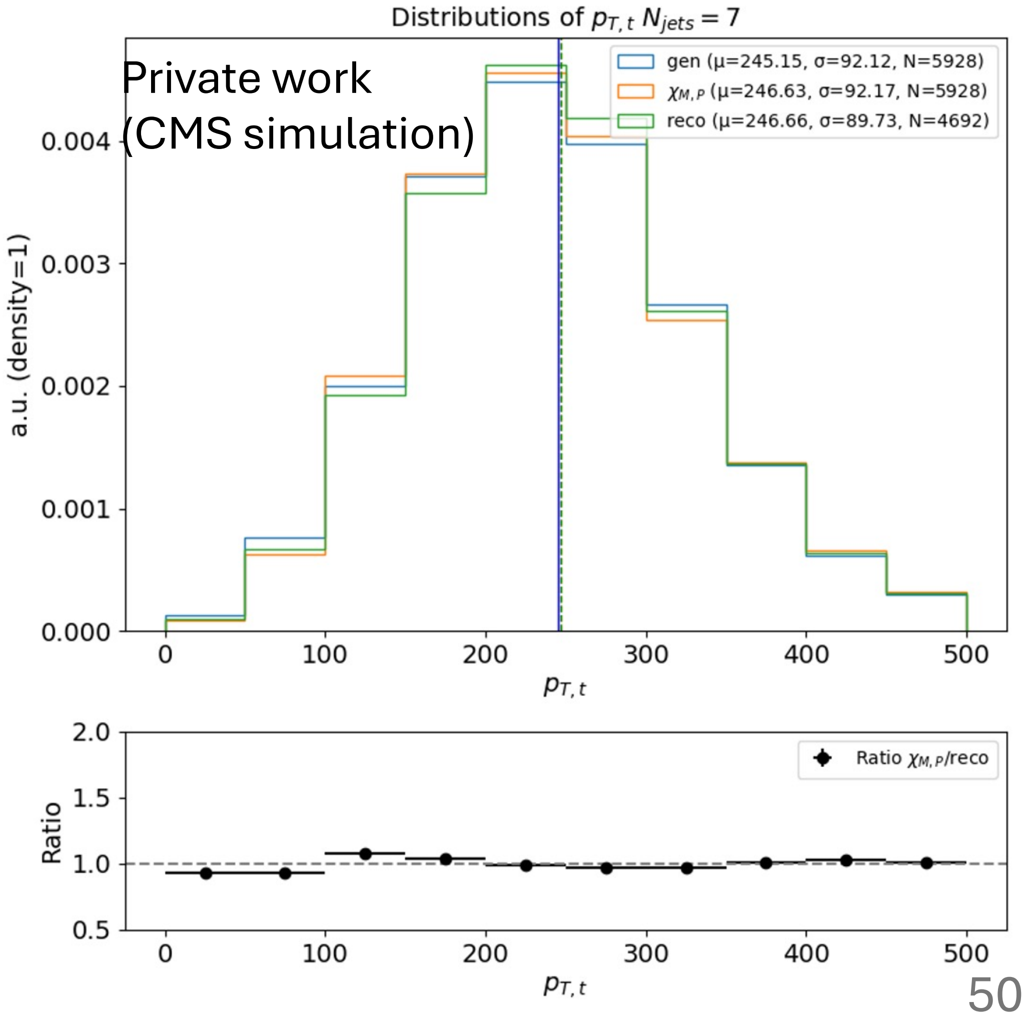


Reconstructed Kinematics: t pT, 7 jets

Mass Only

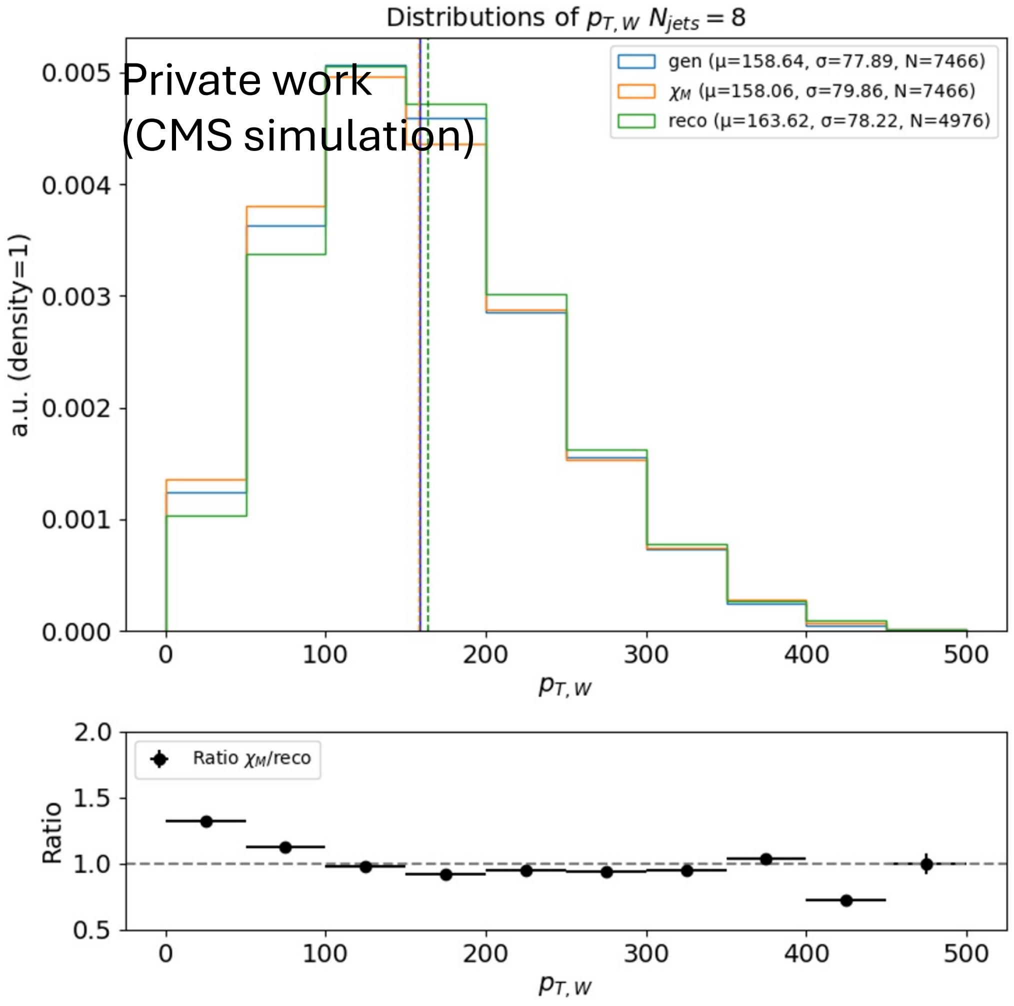


Momentum + Mass

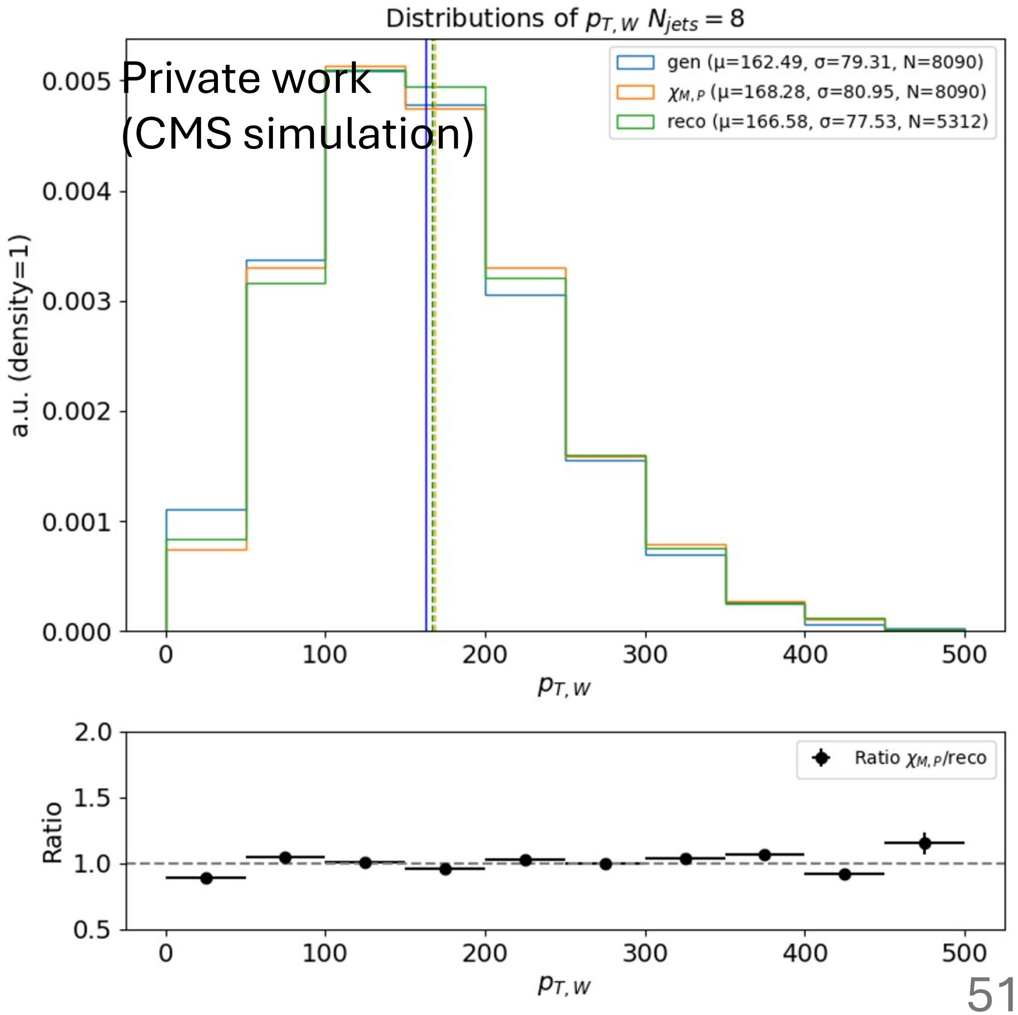


Reconstructed Kinematics: W pT, 8 jets

Mass Only

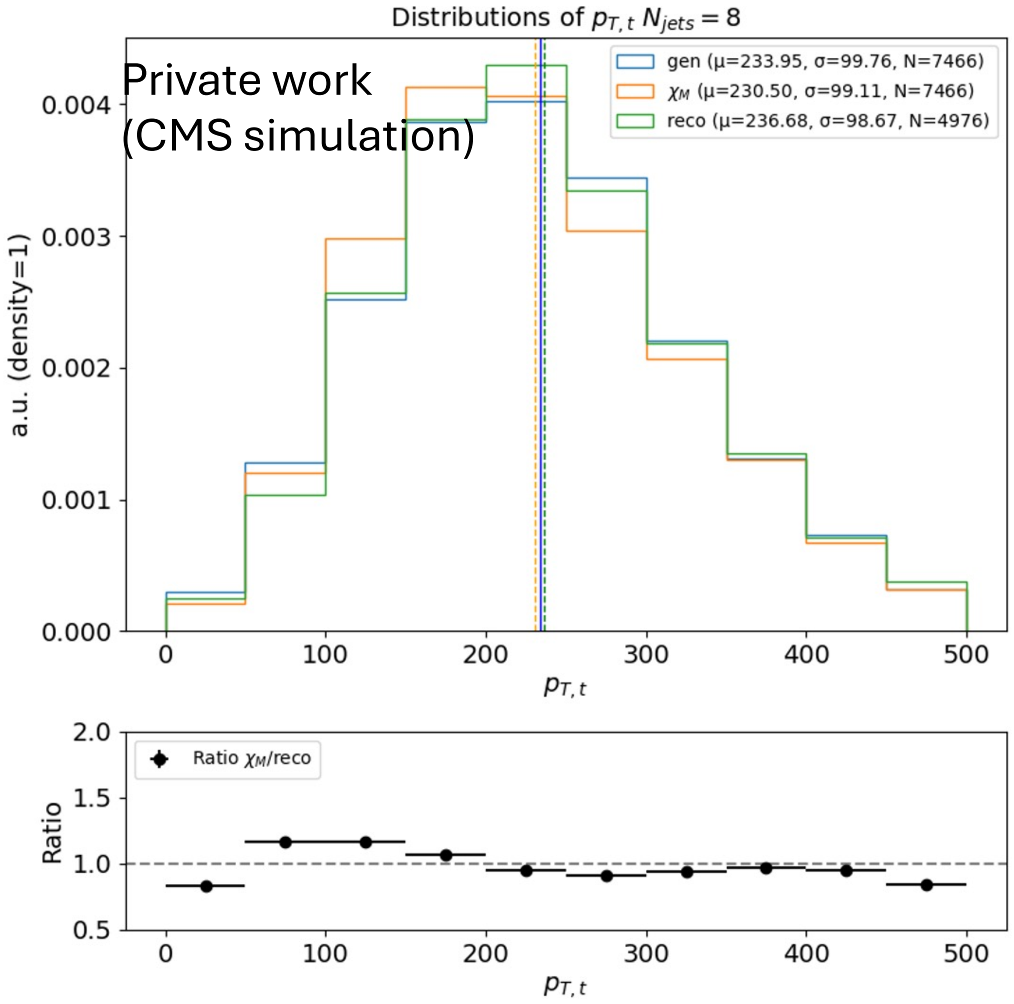


Momentum + Mass



Reconstructed Kinematics: t pT, 8 jets

Mass Only



Momentum + Mass