Data-driven hadronization with **HOMER**

Vienna, ML4Jets 2026

Susie Kim, Ayodele Ore, Sofia Palacios Schweitzer, Tilman Plehn,
Manuel Szewc and MLHAD Collaboration

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Why hadronization with ML?

Motivation:

- Hadronization is not understood from first principles
- Reliance on phenomenological models with parametric forms
- Bottleneck in modern LHC simulations

Why hadronization with ML?

Motivation:

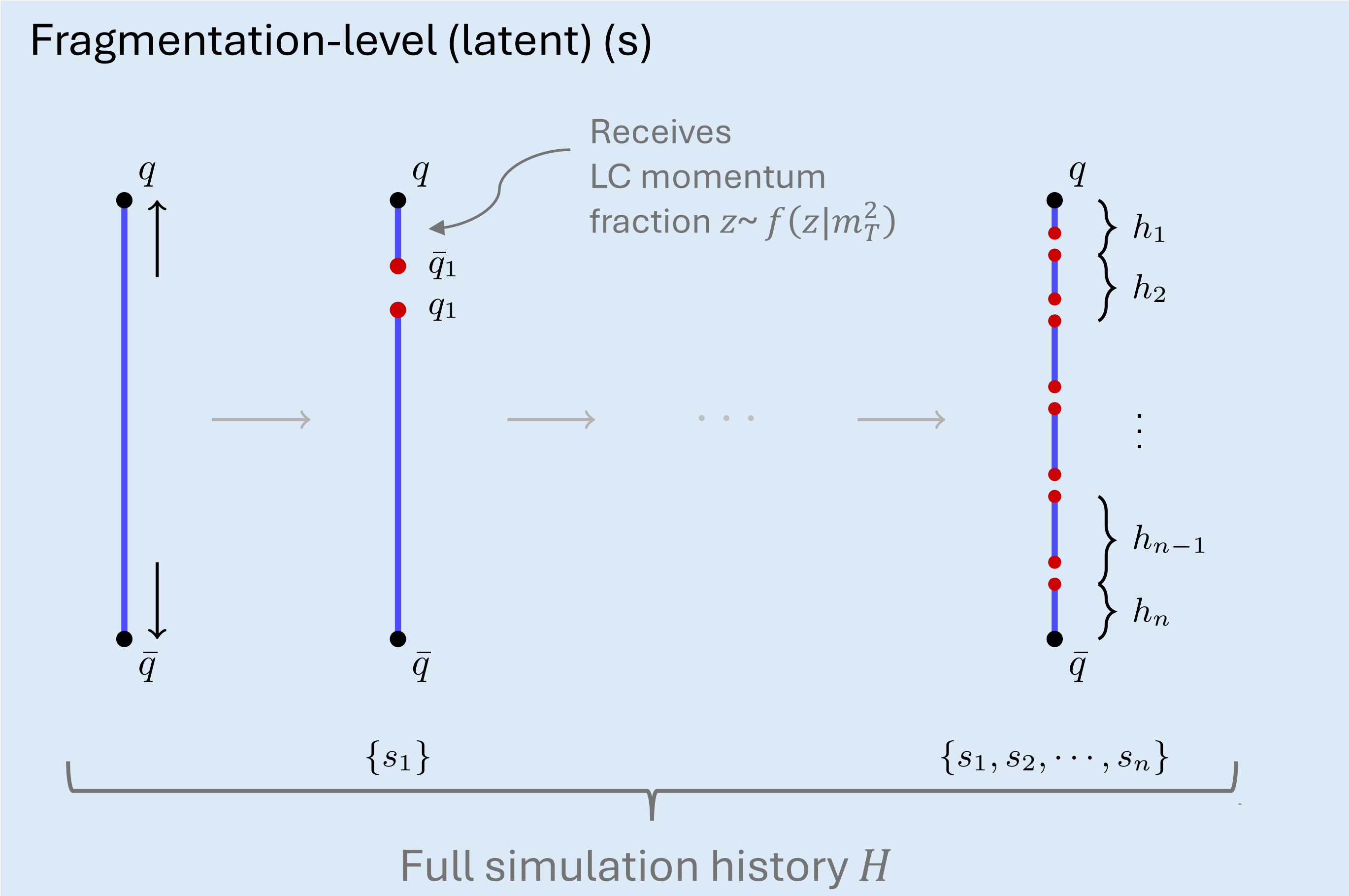
- Hadronization is not understood from first principles
- Reliance on phenomenological models with parametric forms
- Bottleneck in modern LHC simulations

⇩

GOAL:
Learn hadronization **directly from data** through neural network

- No parametric form assumed
- Flexibility beyond the traditional models

String fragmentation

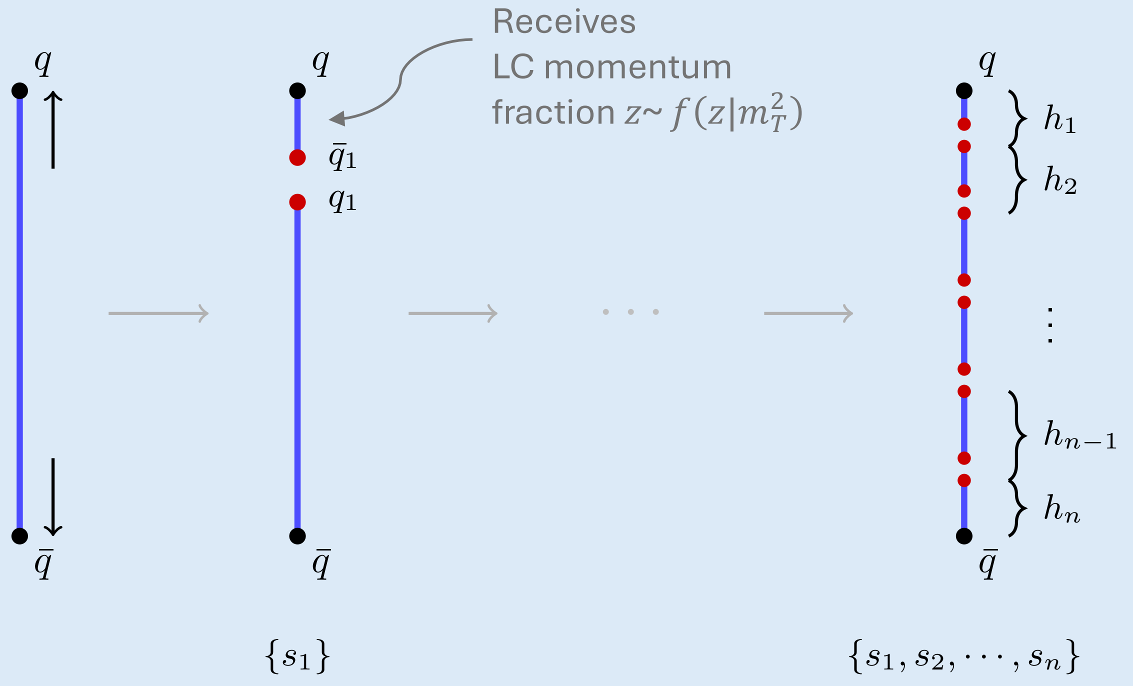


Lund fragmentation function:

$$f(z|m_T^2) \propto \frac{(1-z)^a}{z} \exp\left(-\frac{bm_T^2}{z}\right)$$

String fragmentation

Fragmentation-level (latent) (s)

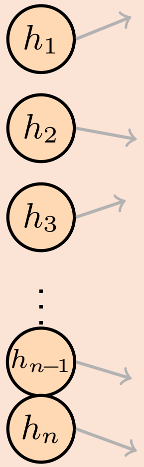


➢ This information is not accessible in experiment!

Fundamental information loss

Many $f(z|m_T^2)$ can lead to same event (x)

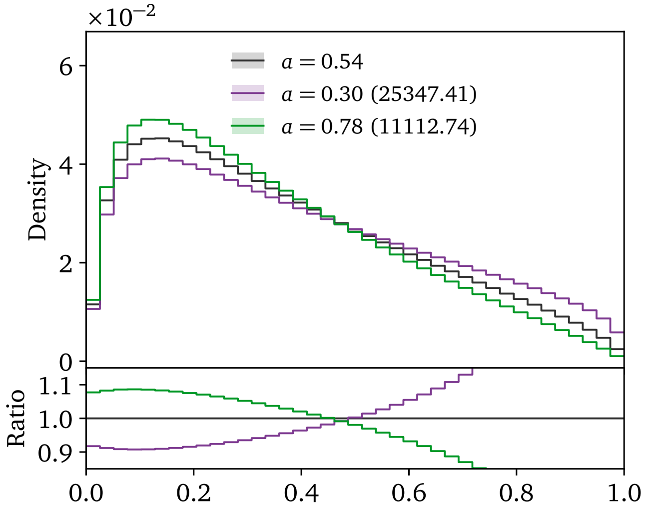
Event-level (x)



- 4-momenta
- charges

Degeneracy in fragmentation level

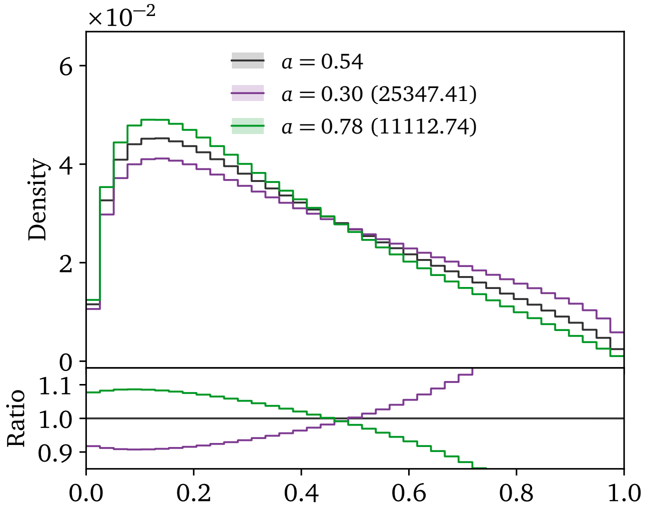
Fragmentation (H)



Lightcone momentum fraction z

Degeneracy in fragmentation level

Fragmentation (H)

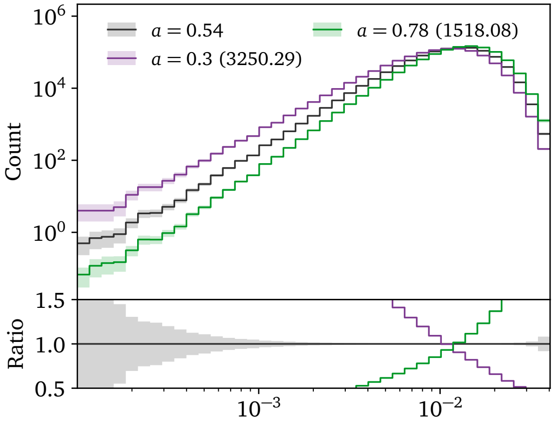


Lightcone momentum fraction z

Without shower

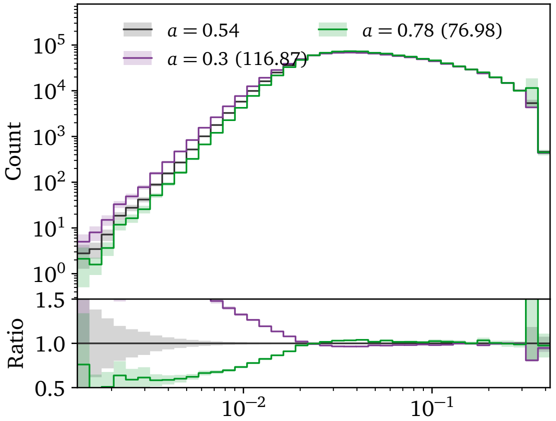
With shower

Event observable (x)



Change caused by shift in a is still visible

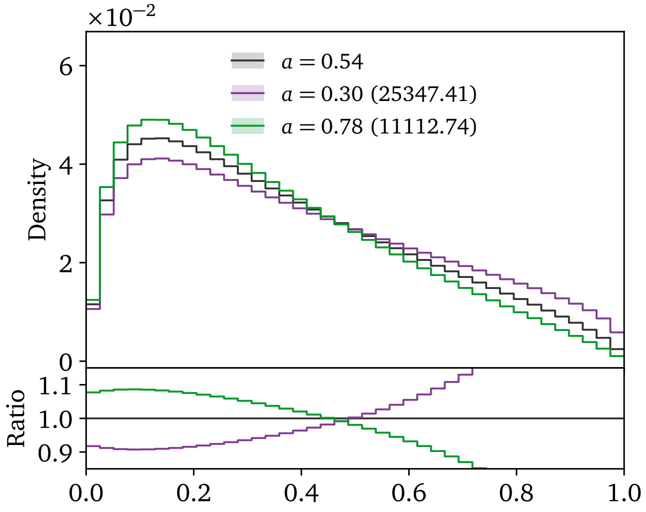
1- Thrust



Sensitivity to a is strongly reduced in bulk

Degeneracy in fragmentation level

Fragmentation (H)



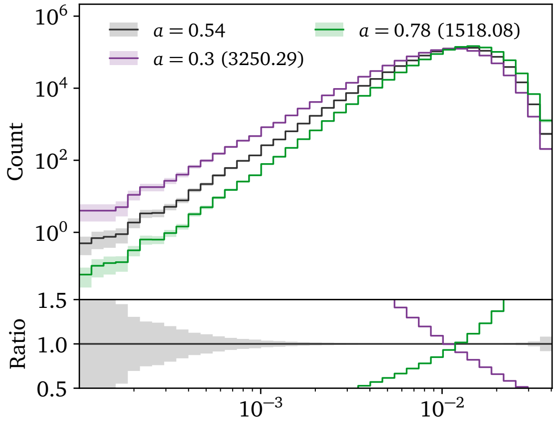
Lightcone momentum fraction z

Without shower

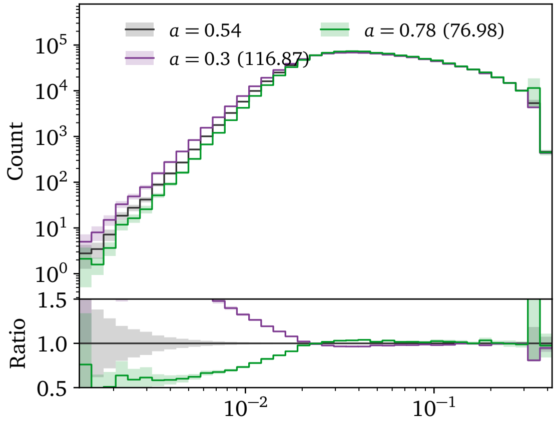
With shower

Non-invertible mapping $p(x|H)$

Event observable (x)



1- Thrust



Change caused by shift in a is still visible

Sensitivity to a is strongly reduced in bulk

Event x can be reproduced by many different $f(z|m_T^2)$

HOMER method

[Bierlich et al, 2410.06342, Assi et al, 2503.05667, Butter et al, 2509.03592]

Goal: $f_{\mathrm{HOMER}}(z|m_T^2) = w(s) f_{\mathrm{sim}}(z|m_T^2)$

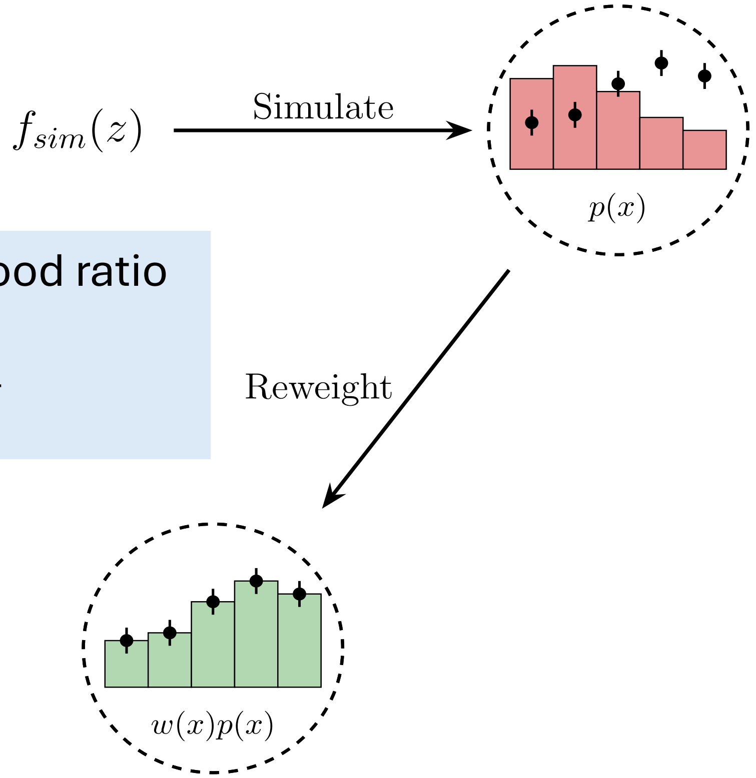
Reproduces target event distribution

Individual break-level weight

Goal: $f_{\text{HOMER}}(z|m_T^2) = w(s) f_{\text{sim}}(z|m_T^2)$

HOMER method

[Bierlich et al, 2410.06342, Assi et al, 2503.05667, Butter et al, 2509.03592]



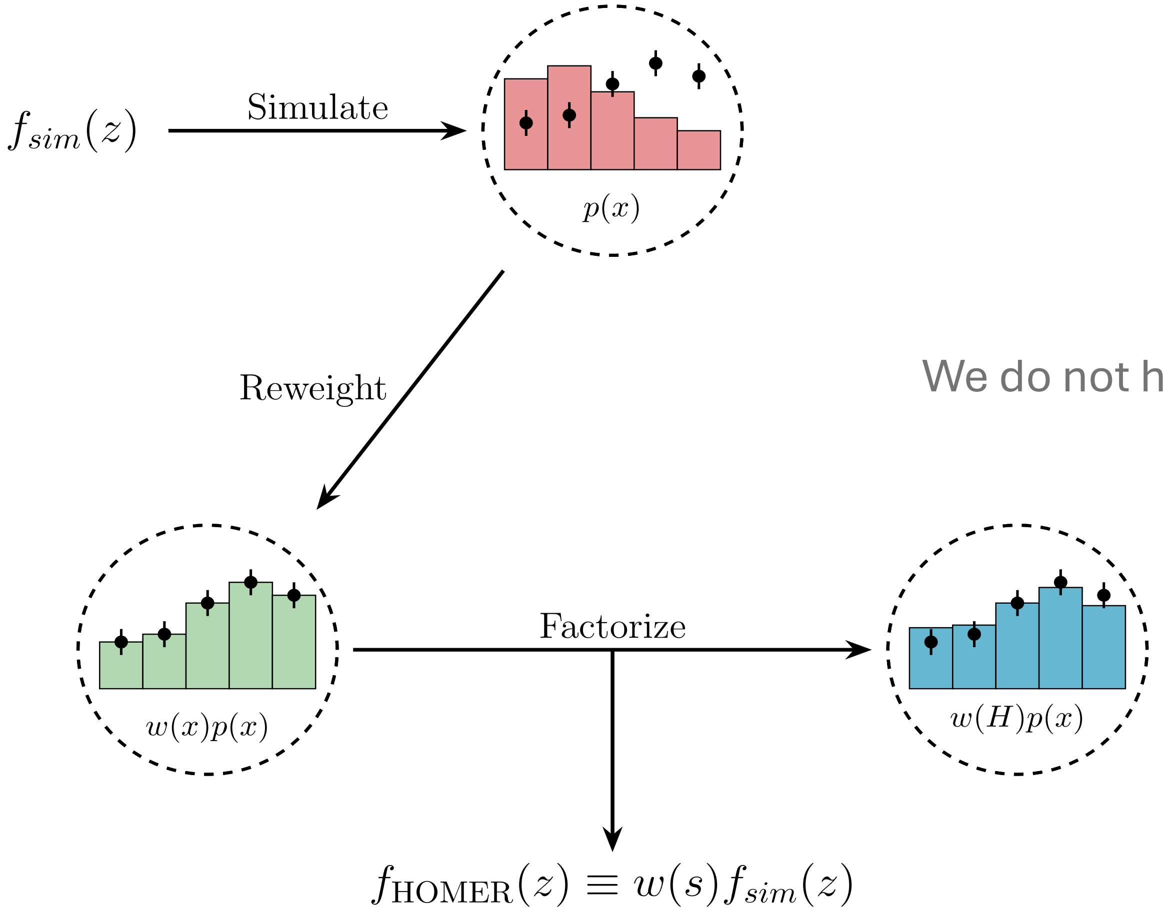
Step 1: event-level likelihood ratio

$$w(x) = \frac{p_{\text{data}}(x)}{p_{\text{sim}}(x)}$$

Goal: $f_{\mathrm{HOMER}}(z|m_T^2) = w(s) f_{\mathrm{sim}}(z|m_T^2)$

HOMER method

[Bierlich et al, 2410.06342, Assi et al, 2503.05667, Butter et al, 2509.03592]



We do not have $p_{\mathrm{data}}(s)$!

Step 2: pull $w(x)$ to break-level

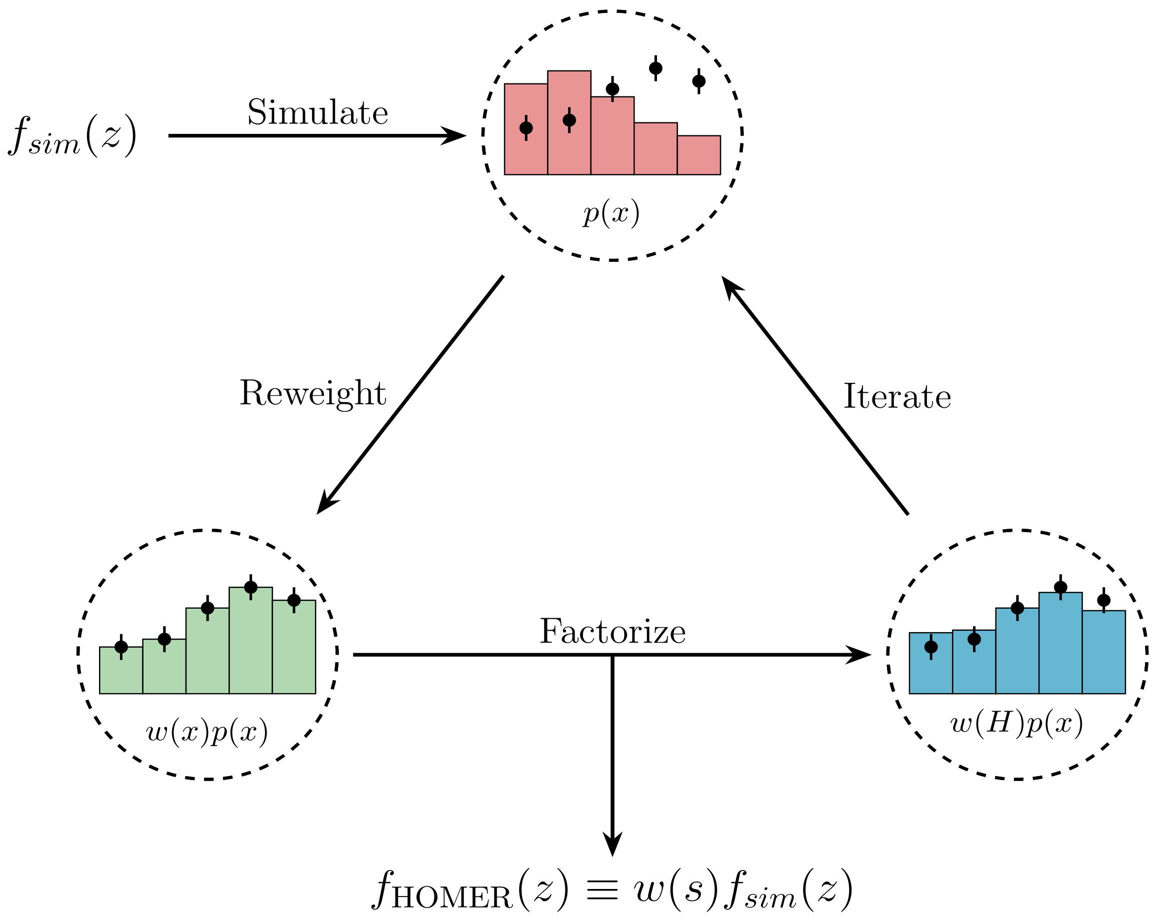
$$\prod_{s \in H} w_\phi(s) \Leftrightarrow w(x)$$

Assumes $p(x|H) \approx \delta(x - H)$

HOMER method

Goal: $f_{\mathrm{HOMER}}(z|m_T^2) = w(s) f_{\mathrm{sim}}(z|m_T^2)$

[Bierlich et al, 2410.06342, Assi et al, 2503.05667, Butter et al, 2509.03592]



Iteration:
$f_{\mathrm{sim}}(z) \to w(s) f_{\mathrm{sim}}(z)$

+

$$\int dH\, p_{\mathrm{sim}}(x|H) \prod_{s \in H} w_\phi(s)$$

- Gaussian smearing
- Neural importance sampling
- AUSSIE [Ore et al, 2602.24282]

Results 1 – Lund closure test

Q: Can HOMER find the "ground truth" $f(z|m_T^2)$?

Results 1 – Lund closure test

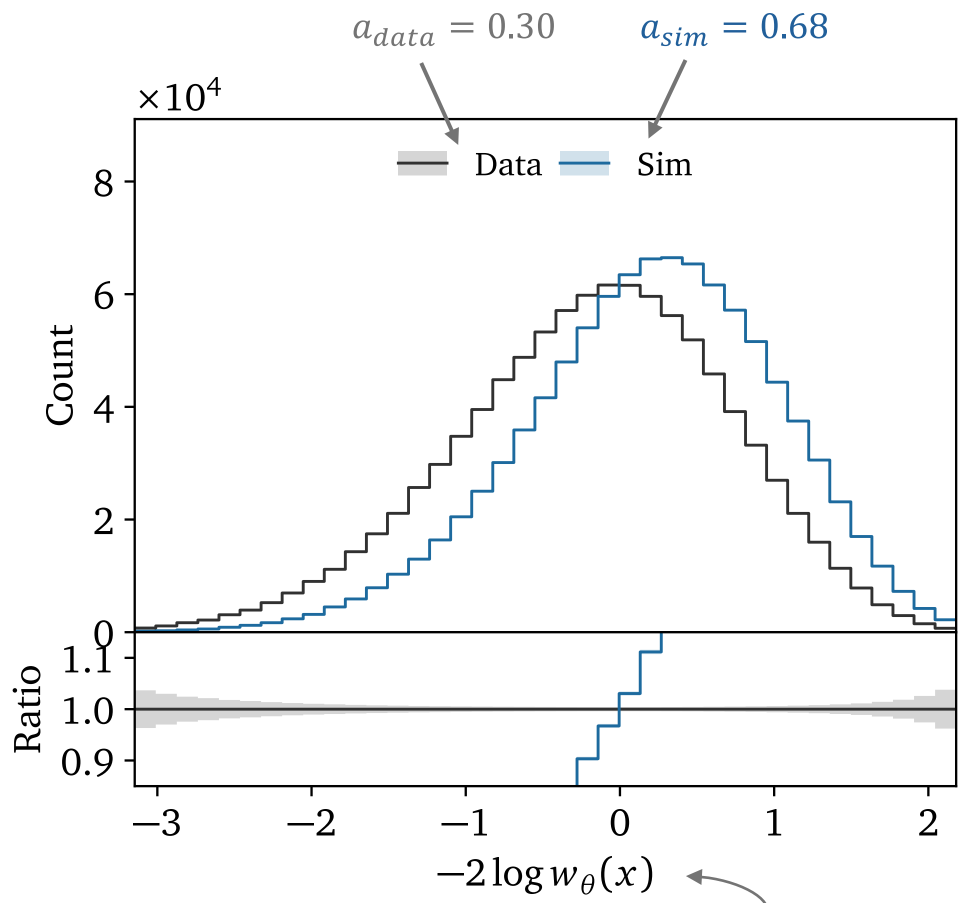
Q: Can HOMER find the "ground truth" $f(z|m_T^2)$?

Setup:

⇒ Data and sim only differ by Lund parameter **a**

$$f(z|m_T^2) \propto \frac{(1-z)^{a}}{z} \exp\left(-\frac{bm_T^2}{z}\right)$$

⇒ With parton shower ⟶ high degeneracy!



Derived from the Step 1 classifier output

Results 1 – Lund closure test

Q: Can HOMER find the "ground truth" $f(z|m_T^2)$?

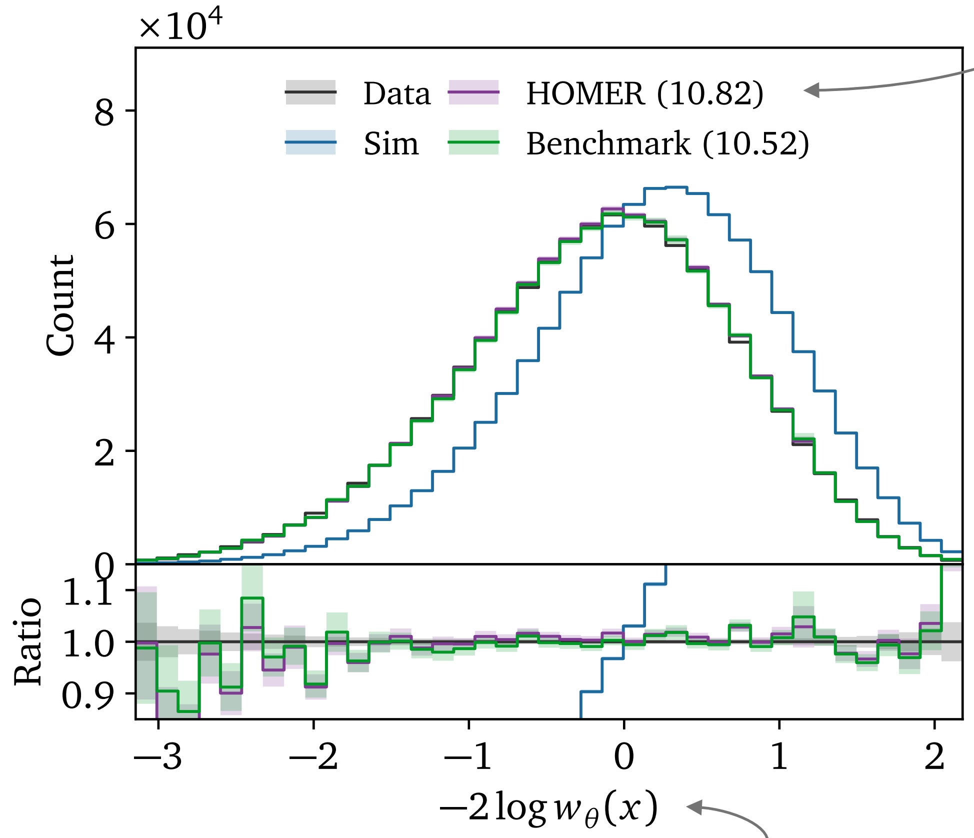
Benchmark:

Network trained directly on $P_{sim}(s)$ and $P_{data}(s)$

Result:

✓ HOMER agrees with Benchmark in event optimal observable

χ^2/N_{bins}

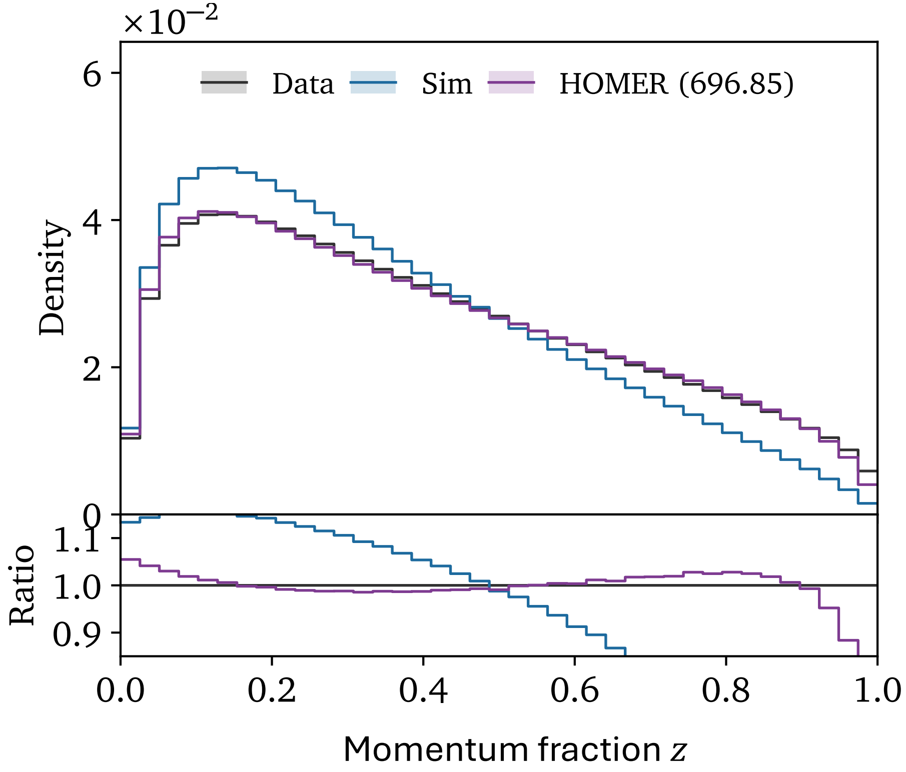


Derived from the Step 1 classifier output

Results 1 – Lund closure test

Q: Can HOMER find the "ground truth" $f(z|m_T^2)$?

Benchmark:
Network trained directly on $P_{sim}(s)$ and $P_{data}(s)$



Result:

- ✓ HOMER agrees with Benchmark in event optimal observable
- ✓ z distribution recovered in the bulk

- ➢ HOMER can reweight one Lund fragmentation function into the target Lund fragmentation function

Result 2 – Beyond Lund fit

Q. Can HOMER fit $f(z|m_T^2)$ outside the Lund family?

Result 2 – Beyond Lund fit

Q. Can HOMER fit $f(z|m_T^2)$ outside the Lund family?

Setup:

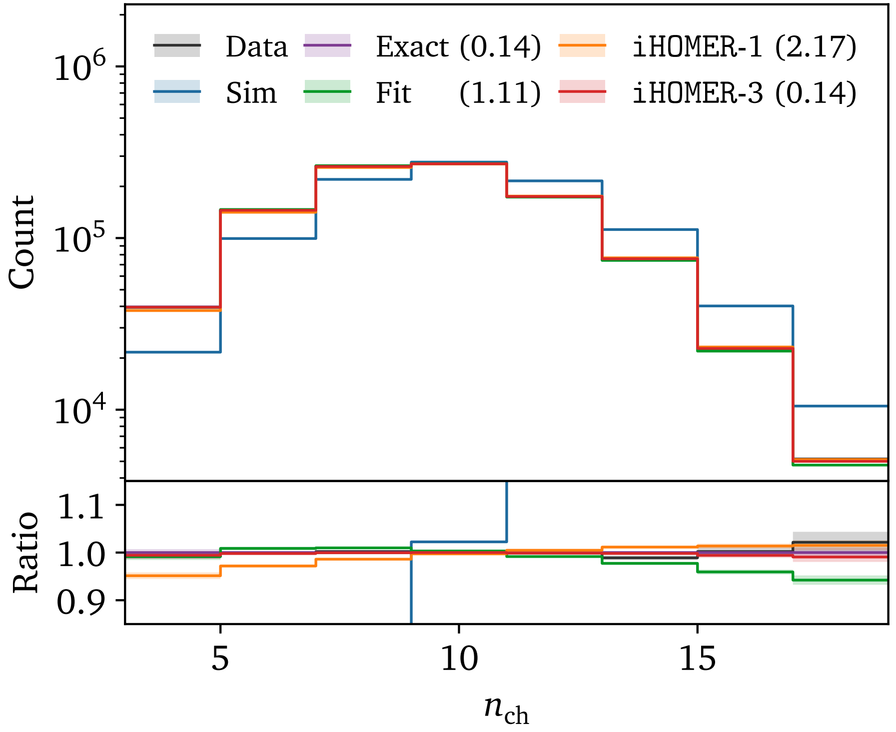
⇒ Data and sim only differ by fragmentation

Sim: standard $f(z|{m_T}^2)$ $a_{sim} = 0.68$

Data: $\frac{1}{2} f_1(z|{m_T}^2) + \frac{1}{2} f_2(z|{m_T}^2)$ ← Doesn't belong to Lund family

$a_1 = 0.30$ and $a_2 = 0.68$

⇒ Without parton shower ($q\bar{q}$)



Result from Butter et al, 2509.03592

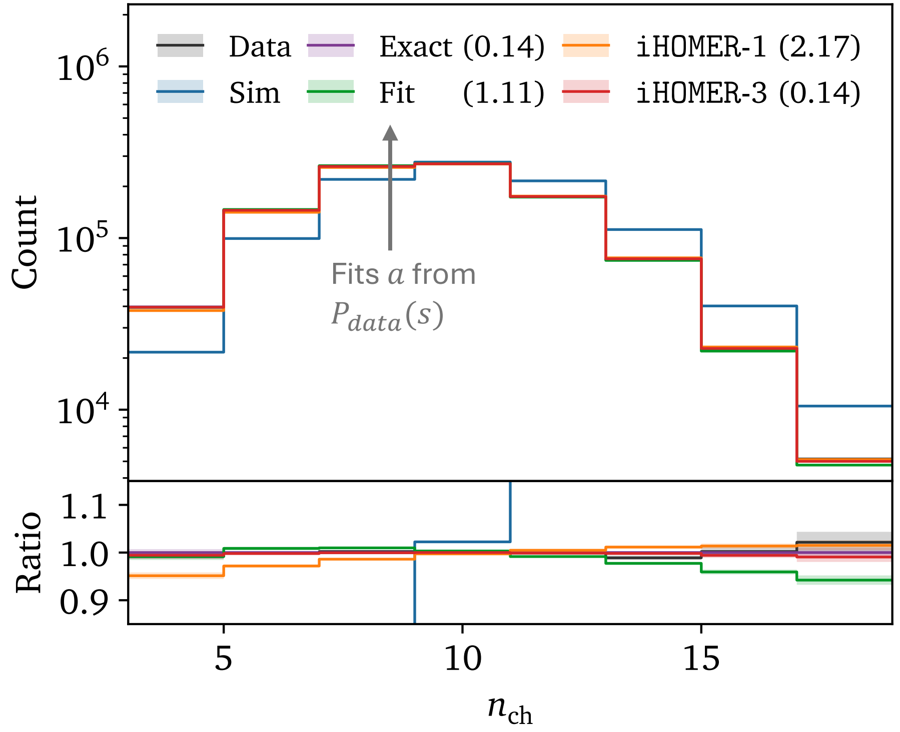
Result 2 – Beyond Lund fit

Q. Can HOMER fit $f(z|m_T^2)$ outside the Lund family?

Result:

✓ HOMER reweighting gets closer to the target than the Lund parameter fit

➢ HOMER provides flexibility to go beyond the Lund



Result from Butter et al, 2509.03592

Result 3 – Parton shower variation

Q. Does HOMER fit anything that is not fragmentation?

Result 3 – Parton shower variation

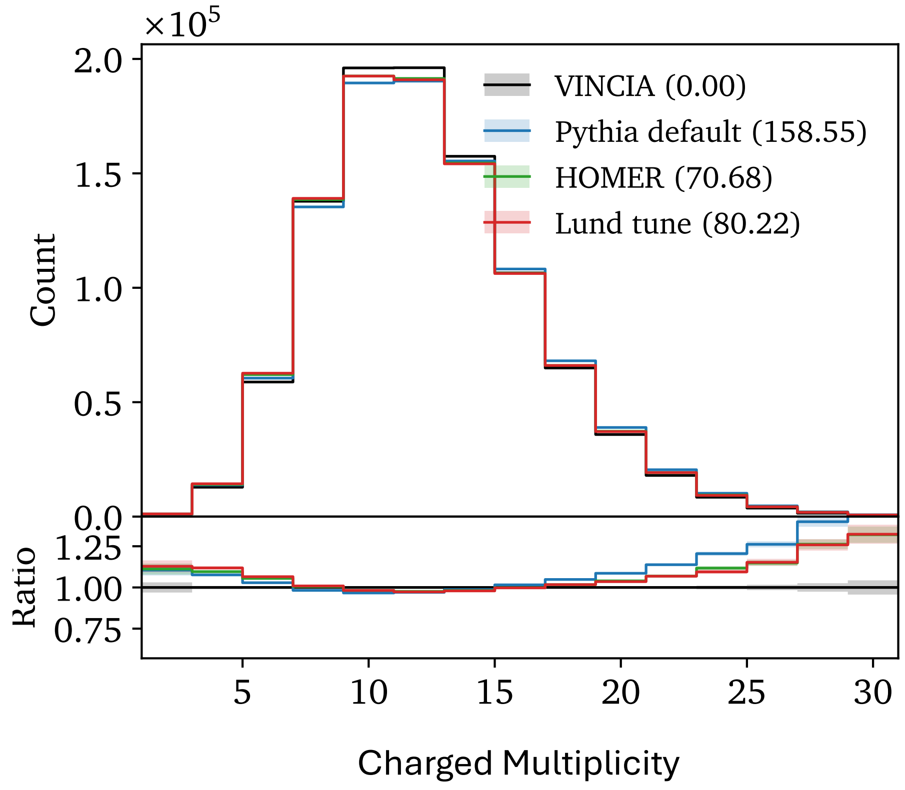
Q. Does HOMER fit anything that is not fragmentation?

Setup:

⇒ Data and sim only differ by **parton shower**

⇒ Fragmentation is the same,
both based on standard $f_{a,b}\,(z|m_T{}^2)$

Not what HOMER is designed to fix!



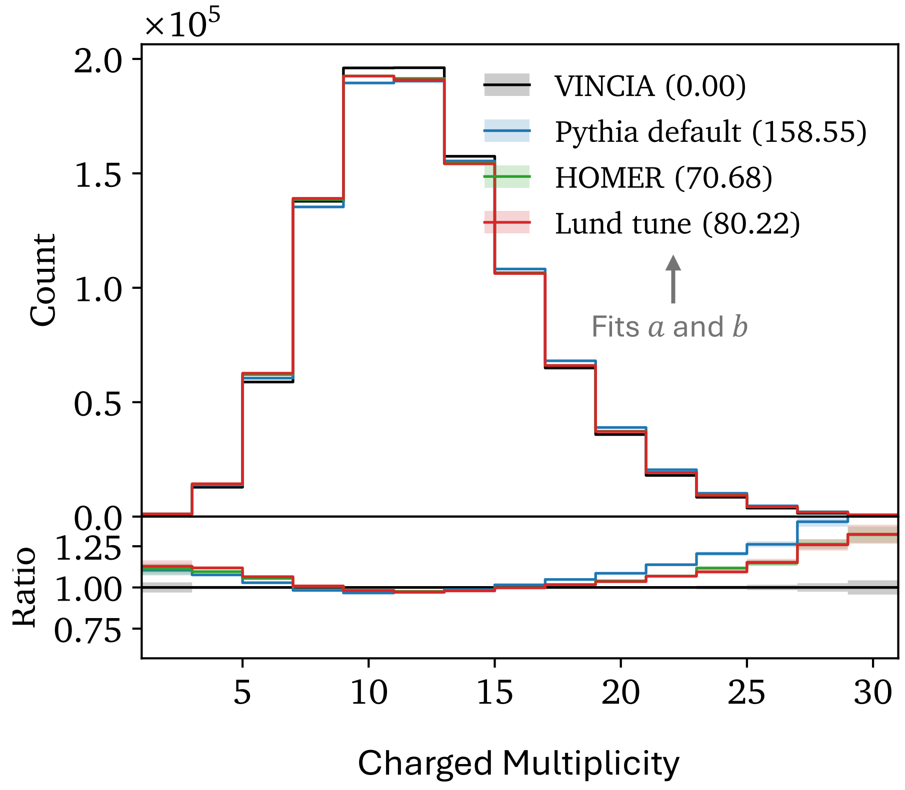
Result 3 – Parton shower variation

Q. Does HOMER fit anything that is not fragmentation?

Result:

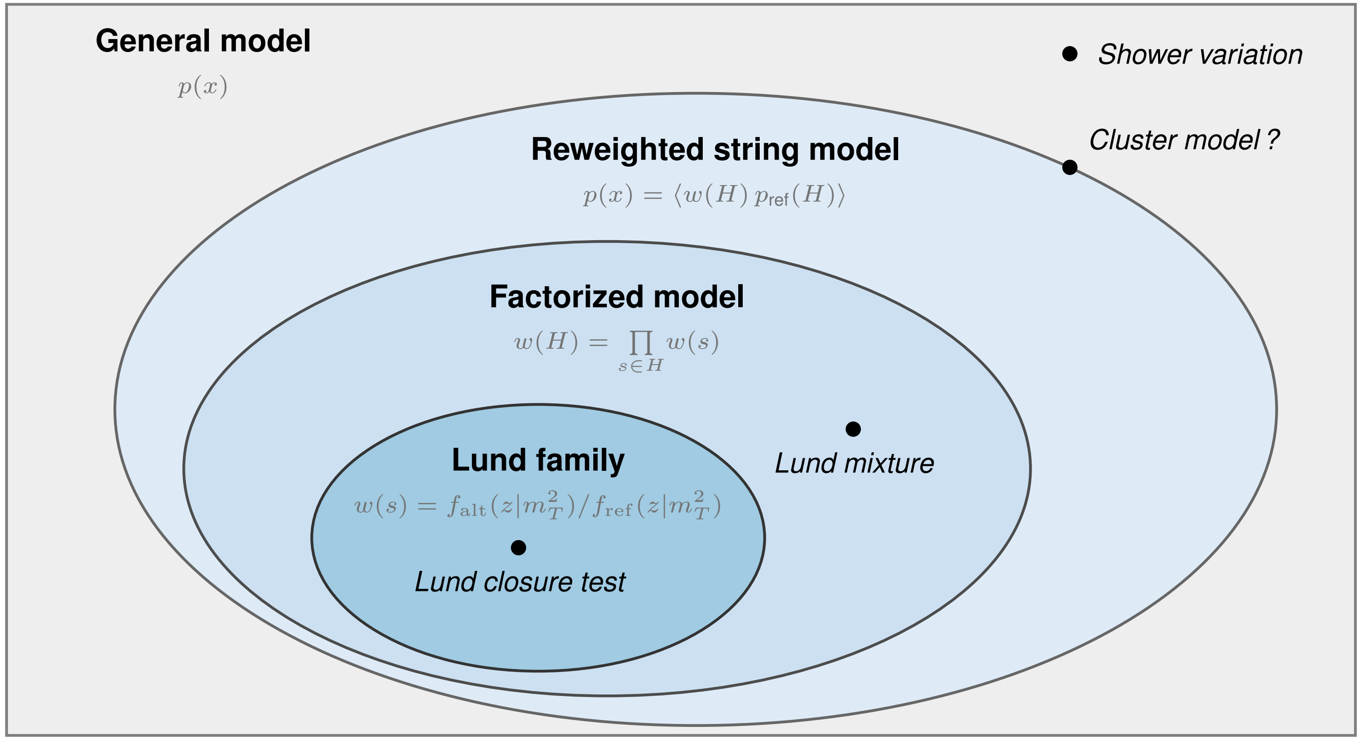
✓ HOMER and Lund fit stay close

➢ HOMER does not completely absorb the shower difference into $f_{\text{HOMER}}(z) = w(s) f_{\text{sim}}(z)$



Lund tune trained on $P_{sim}(s)$
Implementation based on [Heller et al, 2411.0219]

Summary



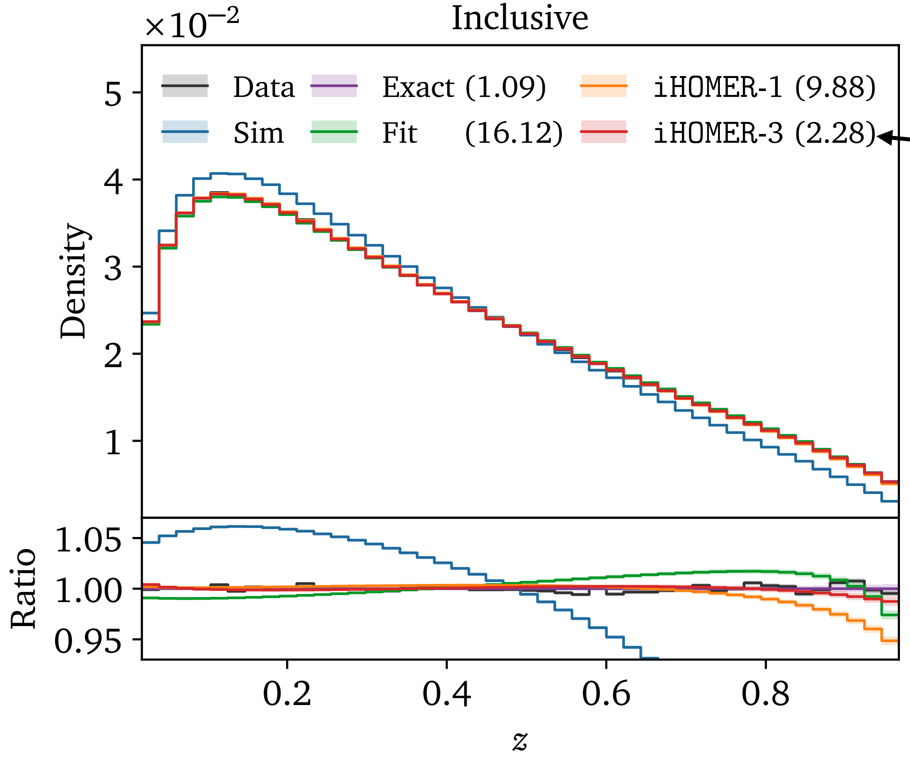
HOMER is
- ✓ Optimal
- ✓ Flexible
- ✓ Robust

… without requiring $p_{\text{data}}(s)$!

Next step:
cluster model fit
and eventually actual data

Additional Slides

(Lund mixture fragmentation level reweigting)



$f_{\text{HOMER}}(z|m_T^2) = w(s) f_{\text{sim}}(z|m_T^2)$
after 3 iteraitons of HOMER