Studying Hadronization with ML and the Road to "differentiable" Pythia

Stephen Mrenna
Fermilab (CSAID)

Based on work with MLHAD

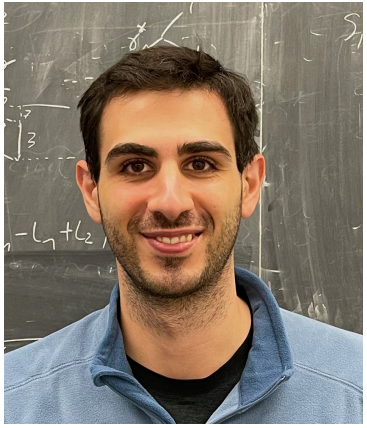
MLHAD


Menzo


Szewc


Wilkinson


Youssef


Assi


Ilten


Zupan


Bierlich

Post-hoc reweighting of hadron production in the Lund string model:
[arXiv:2505.00142]

Characterizing the hadronization of parton showers using the HOMER method:
[arXiv:2503.05667]

Describing hadronization via histories and observables for Monte-Carlo event reweighting:
doi:10.21468/SciPostPhys.18.2.054 [arXiv:2410.06342]

Towards a data-driven model of hadronization using normalizing flows: (MAGIC)
doi:10.21468/SciPostPhys.17.2.045 [arXiv:2311.09296]

Reweighting Monte Carlo predictions and automated fragmentation variations in Pythia 8:
doi:10.21468/SciPostPhys.16.5.134 [arXiv:2308.13459]

What IS Hadronization?

~ what Pythia does

Theme

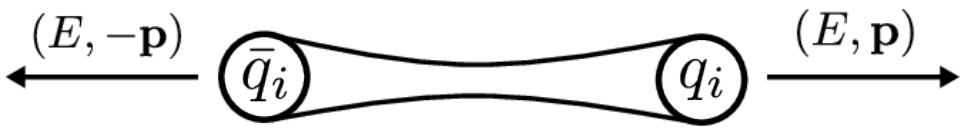
Use ML to correct a good model

Weighting one model to another

Differentiability

Weighting as a surrogate for *physics* parameter differentiability

The algorithm ($q\bar{q}$)



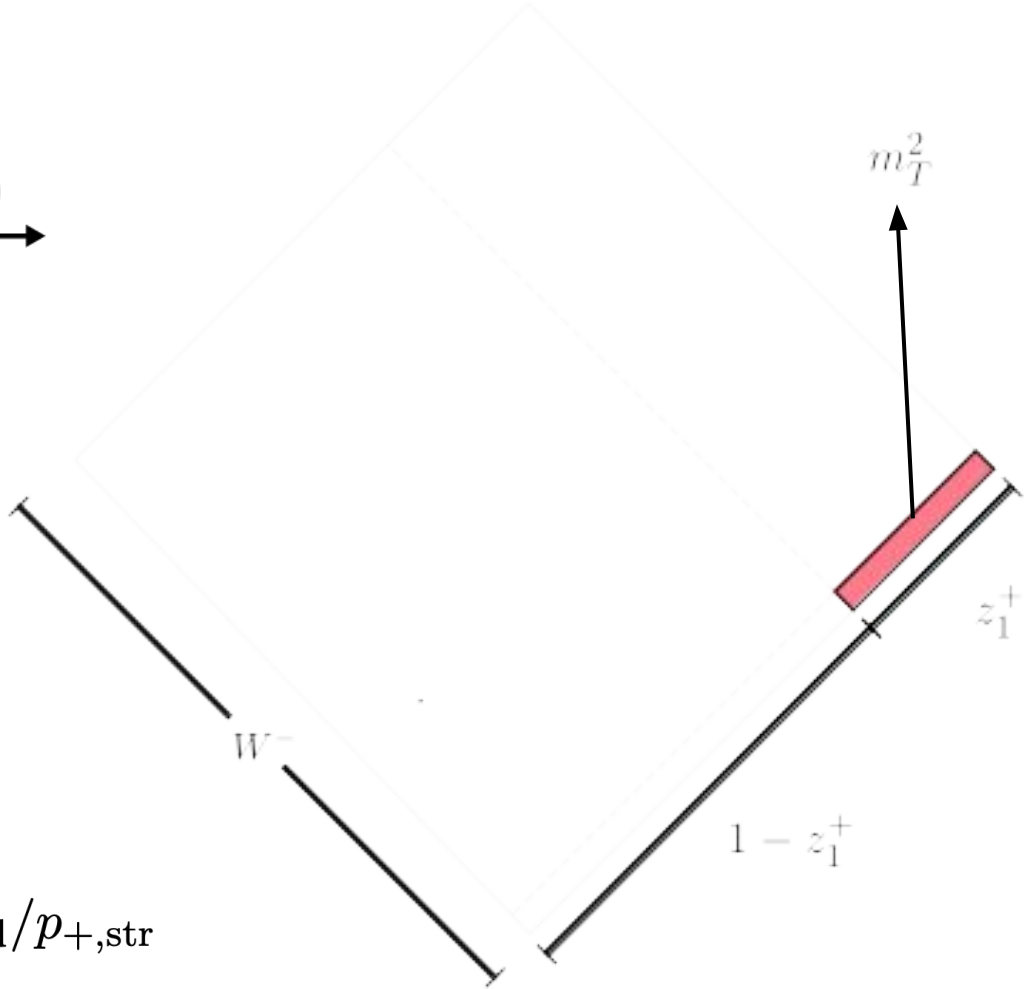
1. Randomly select one of the string ends
2. Sample new quark flavor
3. Sample transverse momentum of new quarks

$$\mathcal{P}(p_x, p_y; \sigma_{p_\perp}) = \frac{1}{2\pi\sigma_{p_\perp}^2} \exp\left(-\frac{p_x^2+p_y^2}{2\sigma_{p_\perp}^2}\right)$$

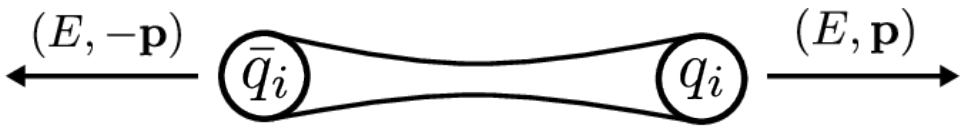
4. Sample longitudinal momentum fraction of new hadron

$$f(z) \propto \frac{(1-z)^a}{z} \exp\left(-\frac{bm_T^2}{z}\right), z \equiv p_{+,\mathrm{had}}/p_{+,\mathrm{str}}$$

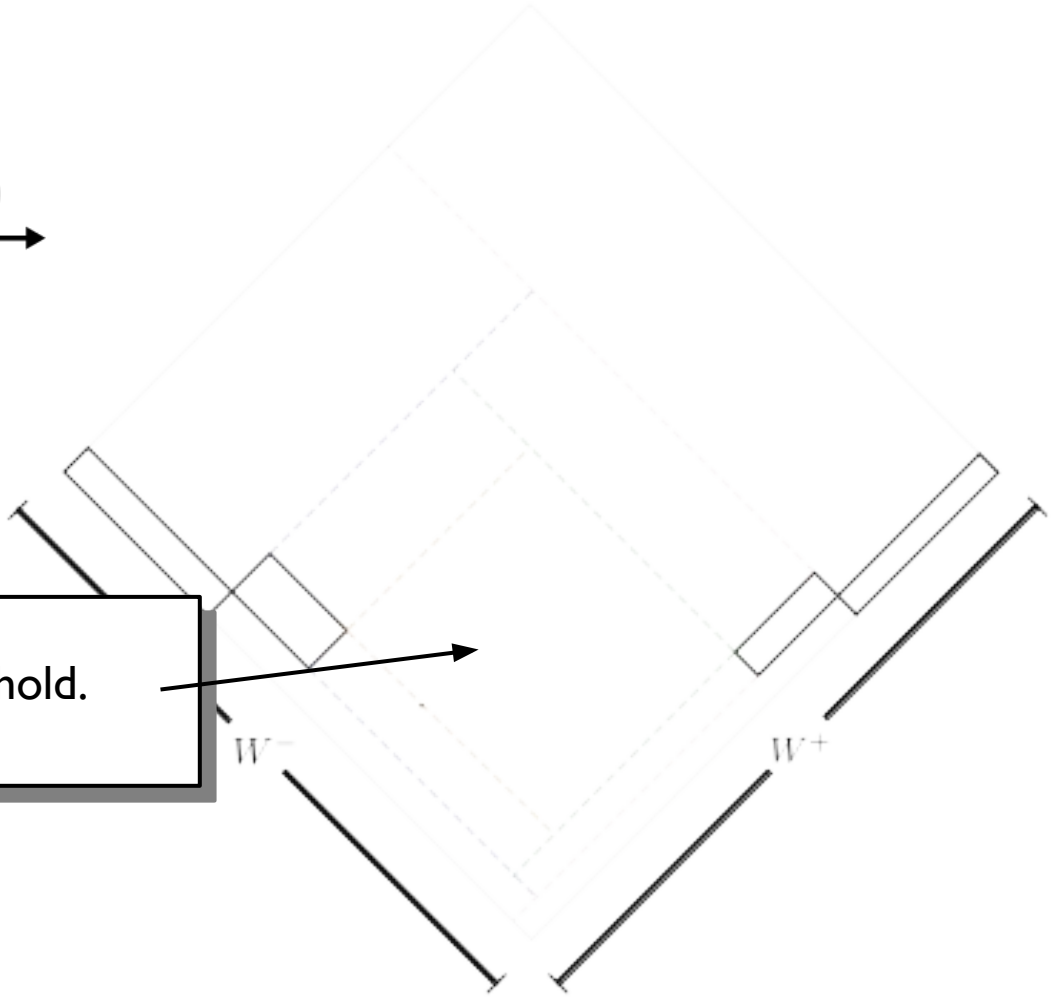
5. Repeat 1.-4.



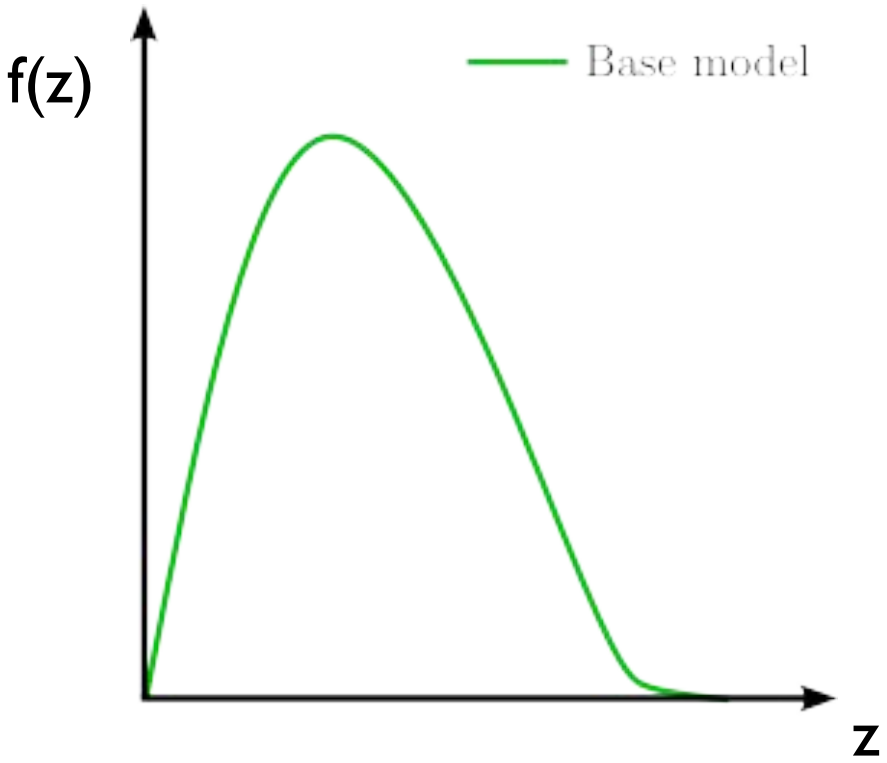
The algorithm ($q\bar{q}$)

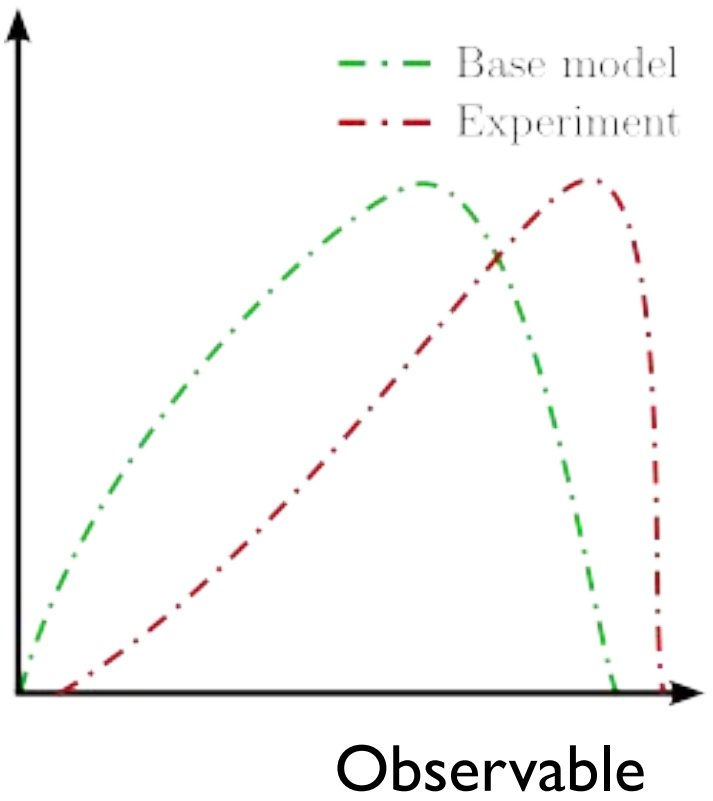


Special treatment at low mass threshold.

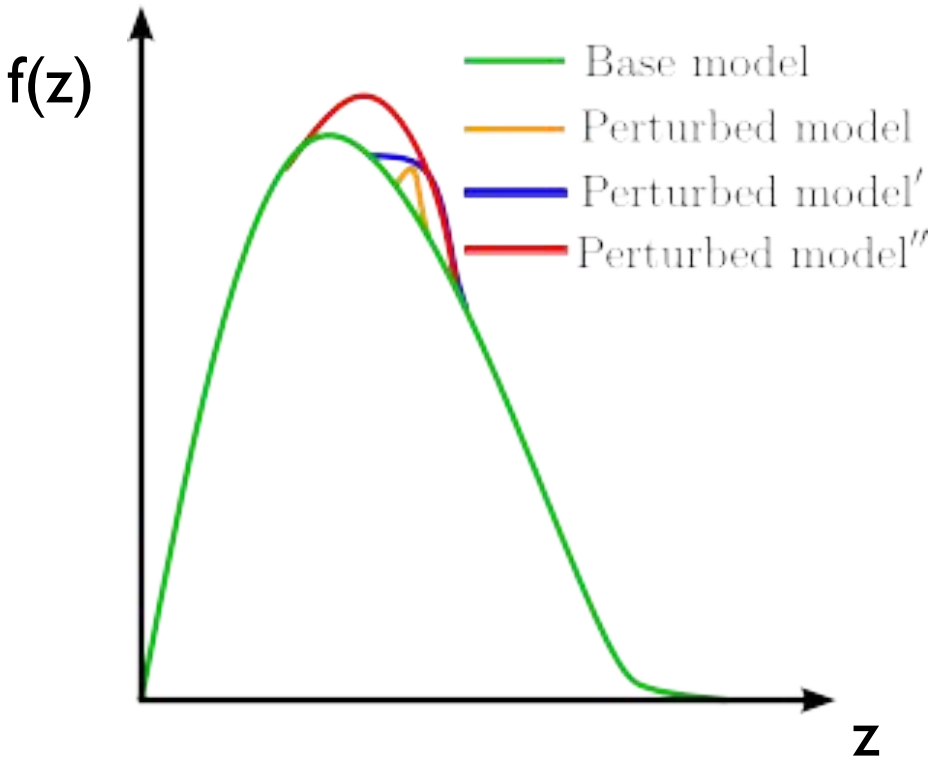


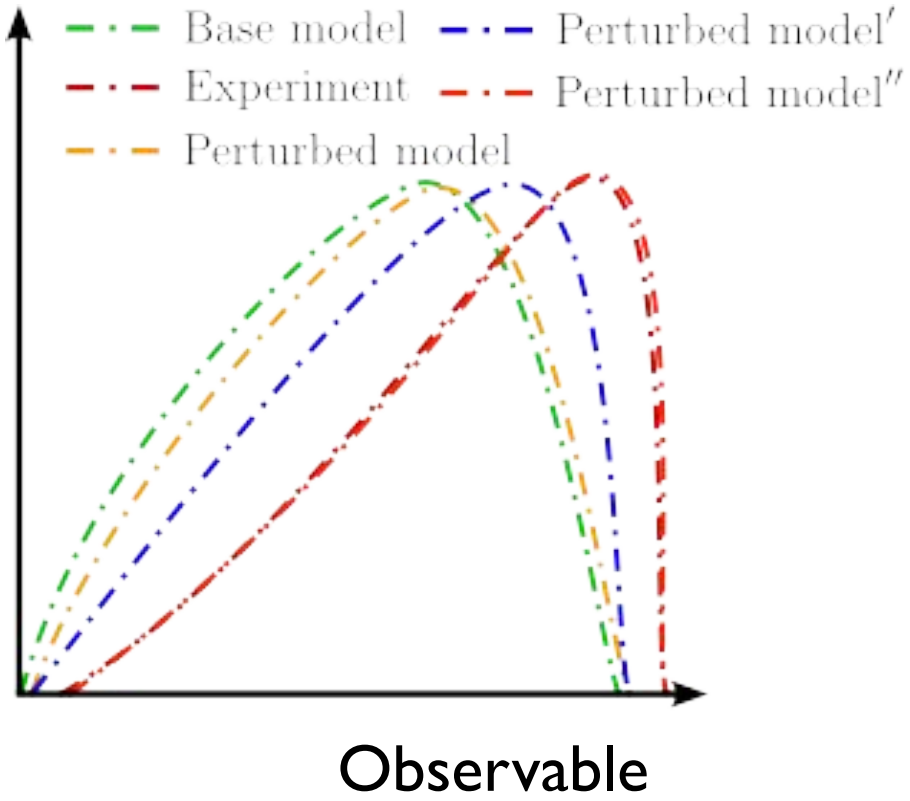
MLHAD





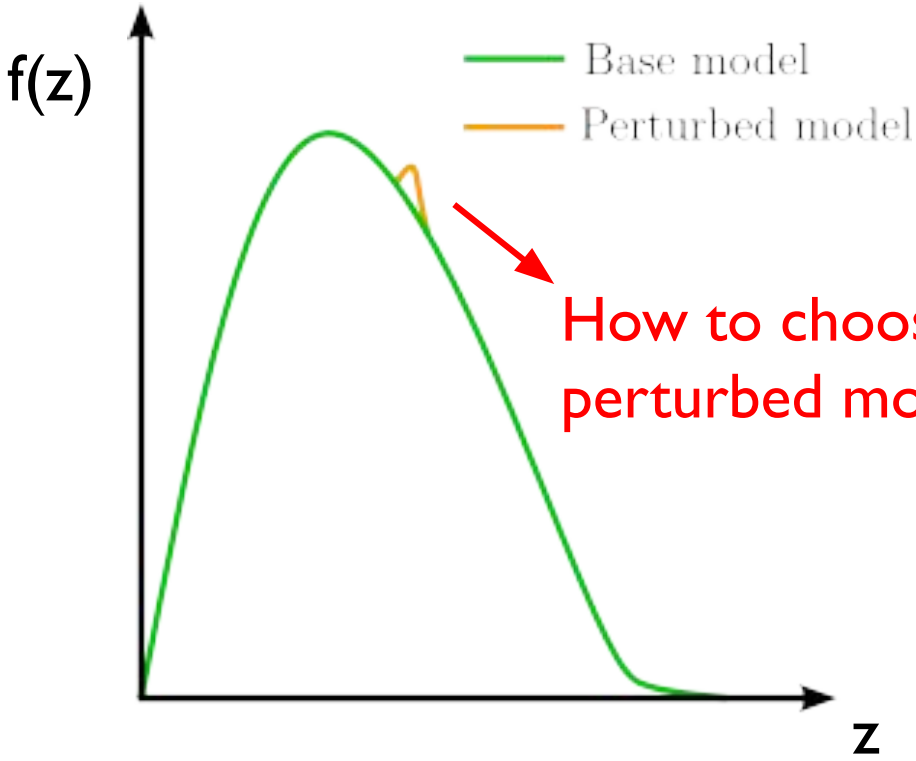
MLHAD


f(z)
Base model
Perturbed model
Perturbed model′
Perturbed model″
z


Base model
Perturbed model′
Experiment
Perturbed model″
Perturbed model
Observable

MLHAD

~~ML~~ vs ML



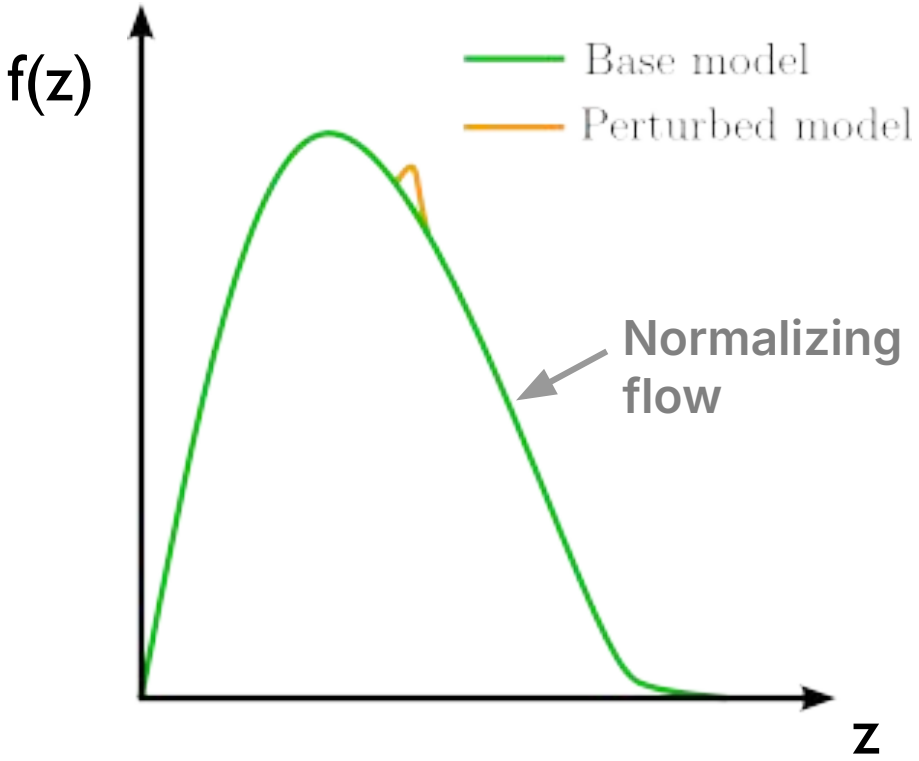
How to choose perturbed model?

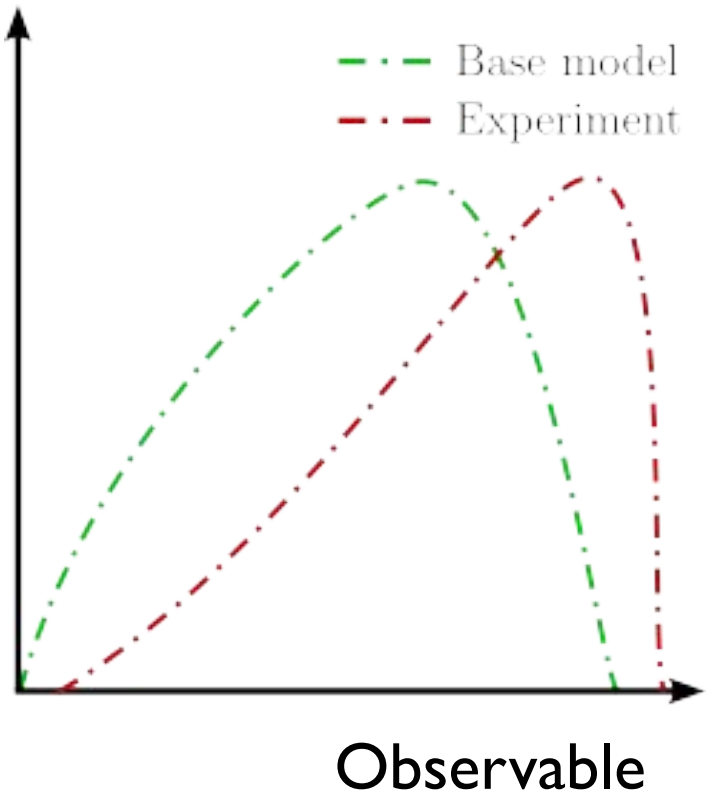
Vary parameters
Brute force

ML-based (data-driven) fragmentation function
- MAGIC, HOMER

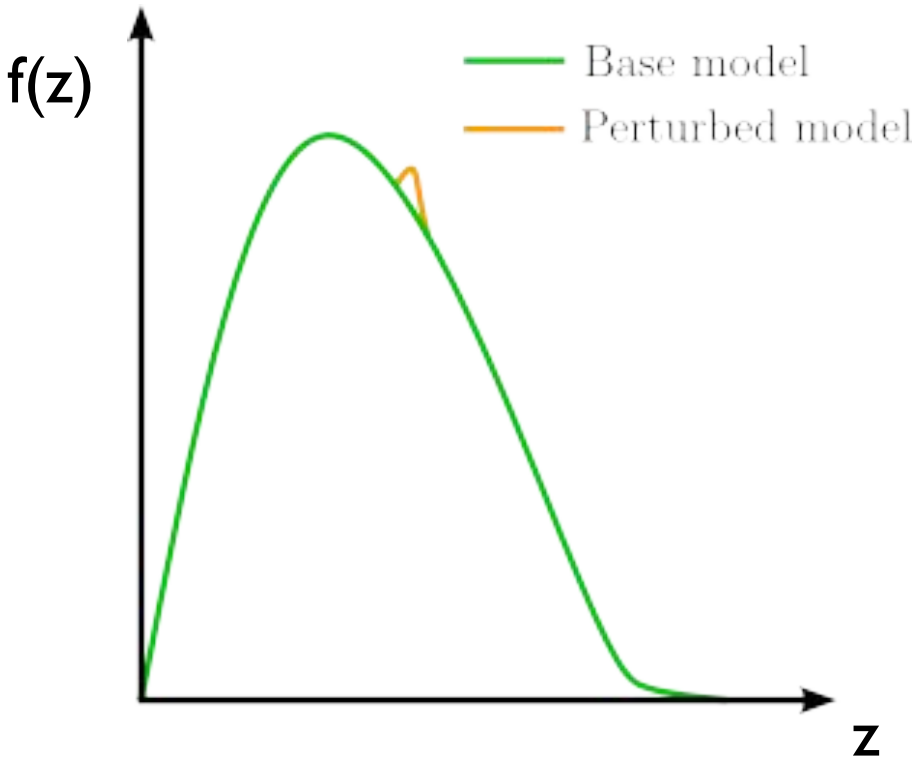
Hybrid – keep Lund fragmentation function, promote Lund parameters to differentiable objects
- Rejection sampling with autodifferentiation (RSA)

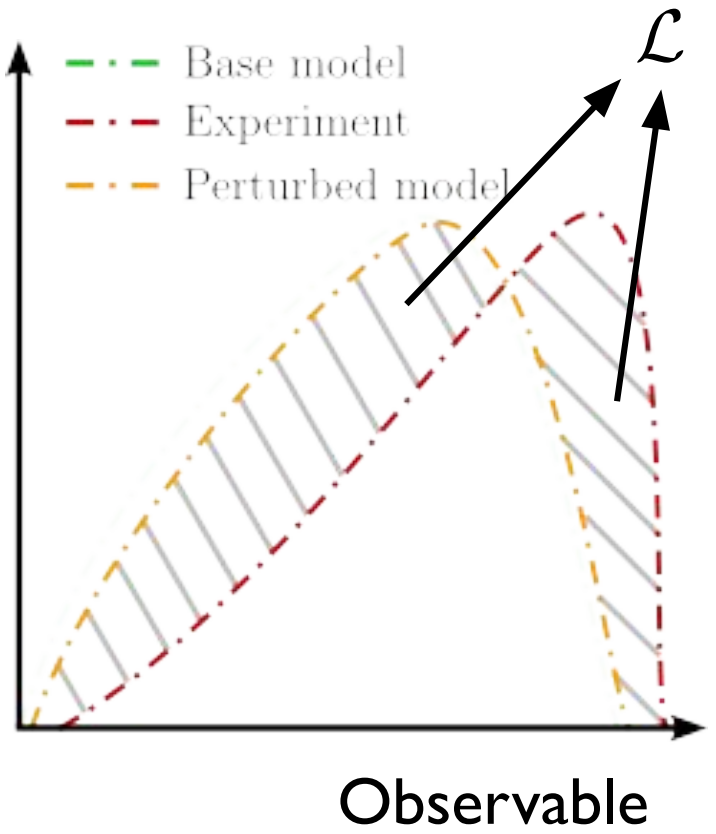
MLHAD **MAGIC**


f(z)
Base model
Perturbed model
Normalizing flow
z

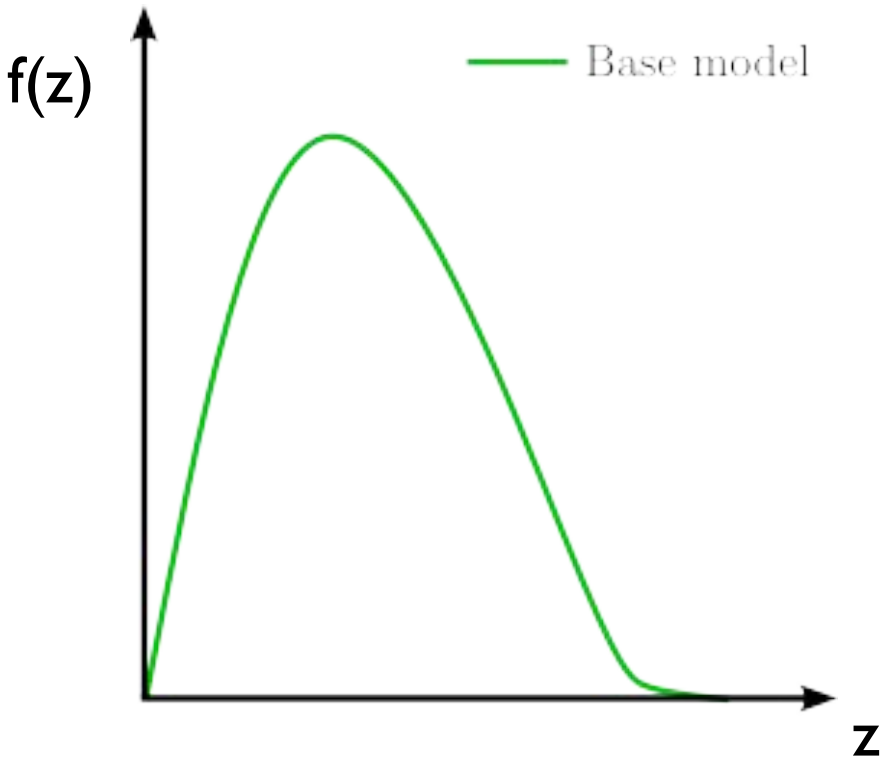

Base model
Experiment
Observable

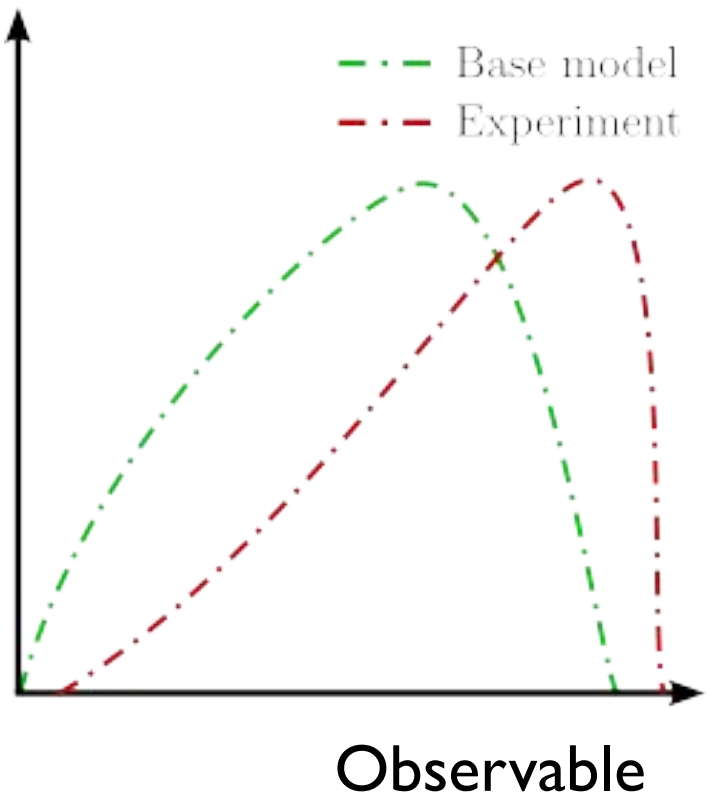
MLHAD **MAGIC**


f(z)
Base model
Perturbed model
z

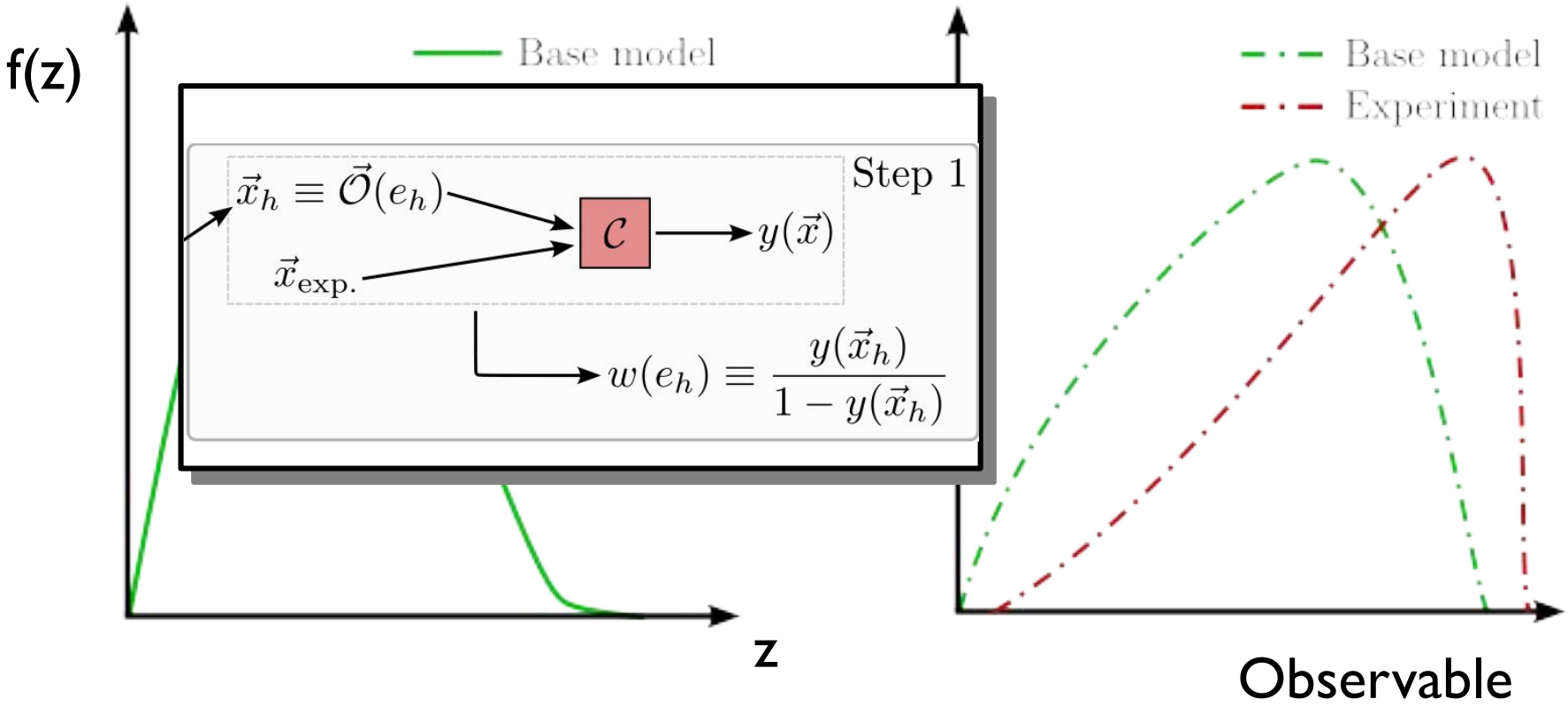

Base model
Experiment
Perturbed model
$\mathcal{L}$
Observable

MLHAD HOMER

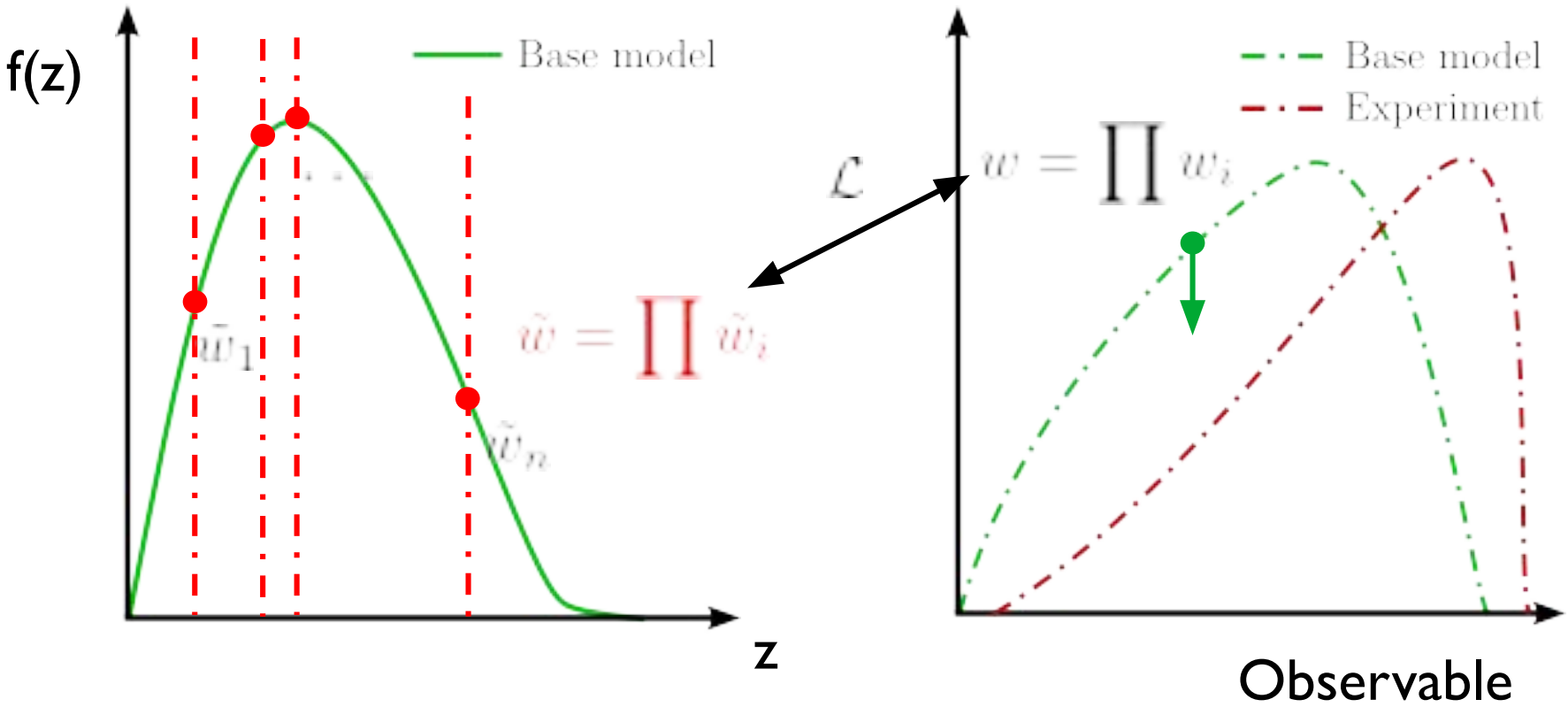




MLHAD HOMER


f(z)
Base model
Step 1
$\vec{x}_h \equiv \vec{\mathcal{O}}(e_h)$
$\mathcal{C}$
$y(\vec{x})$
$\vec{x}_{\text{exp.}}$
$w(e_h) \equiv \frac{y(\vec{x}_h)}{1 - y(\vec{x}_h)}$
z
Base model
Experiment
Observable

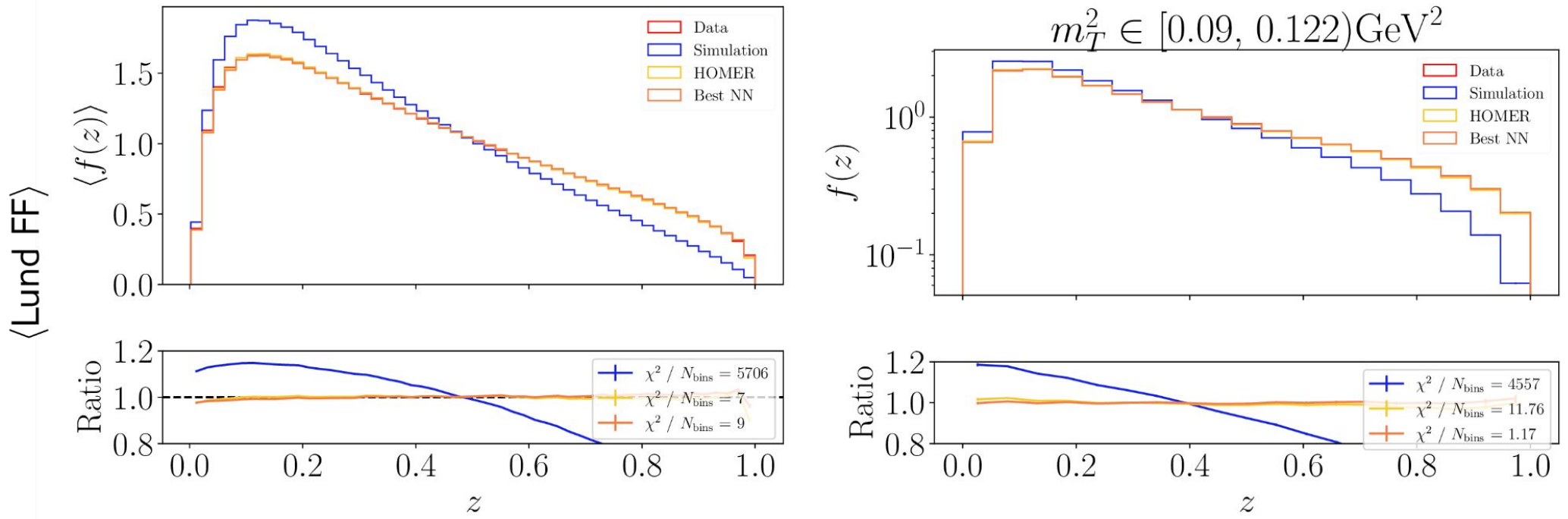
MLHAD HOMER


f(z)
Base model
$\tilde{w} = \prod \tilde{w}_i$
$\tilde{w}_1$
$\tilde{w}_n$
z
$\mathcal{L}$
Base model
Experiment
$w = \prod w_i$
Observable

Simulation reweighted to DATA using HOMER

HOMER uses Thrust, N_{charge}, $\cdots$ AND hadron p

Kinematic Slice of LUND FF



$z = \frac{p}{E_{\text{st}}}$

MLHAD collaboration

Satisfies closure on SIM:: NEXT STEP: REAL DATA

MAGIC vs HOMER

NF trained on simulated kinematics	NN trained on correlations between string fragmentation function and higher level observables (e.g. Thrust)
NF modified to match data	
Weighting speeds up ML	Weighting is learned output
Yields Likelihood	Yields Likelihood Ratios

Weighting in Pythia

MC methods in Pythia fully deterministic – can track every decision

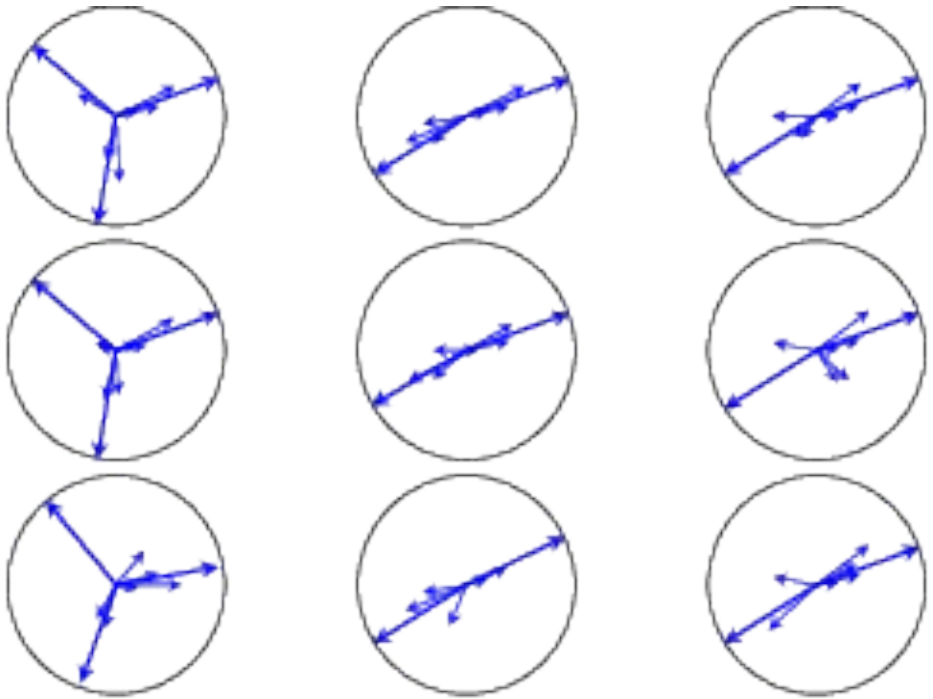
Careful accounting of accepts AND rejects: how would accepts and rejects change for *different* physics parameters

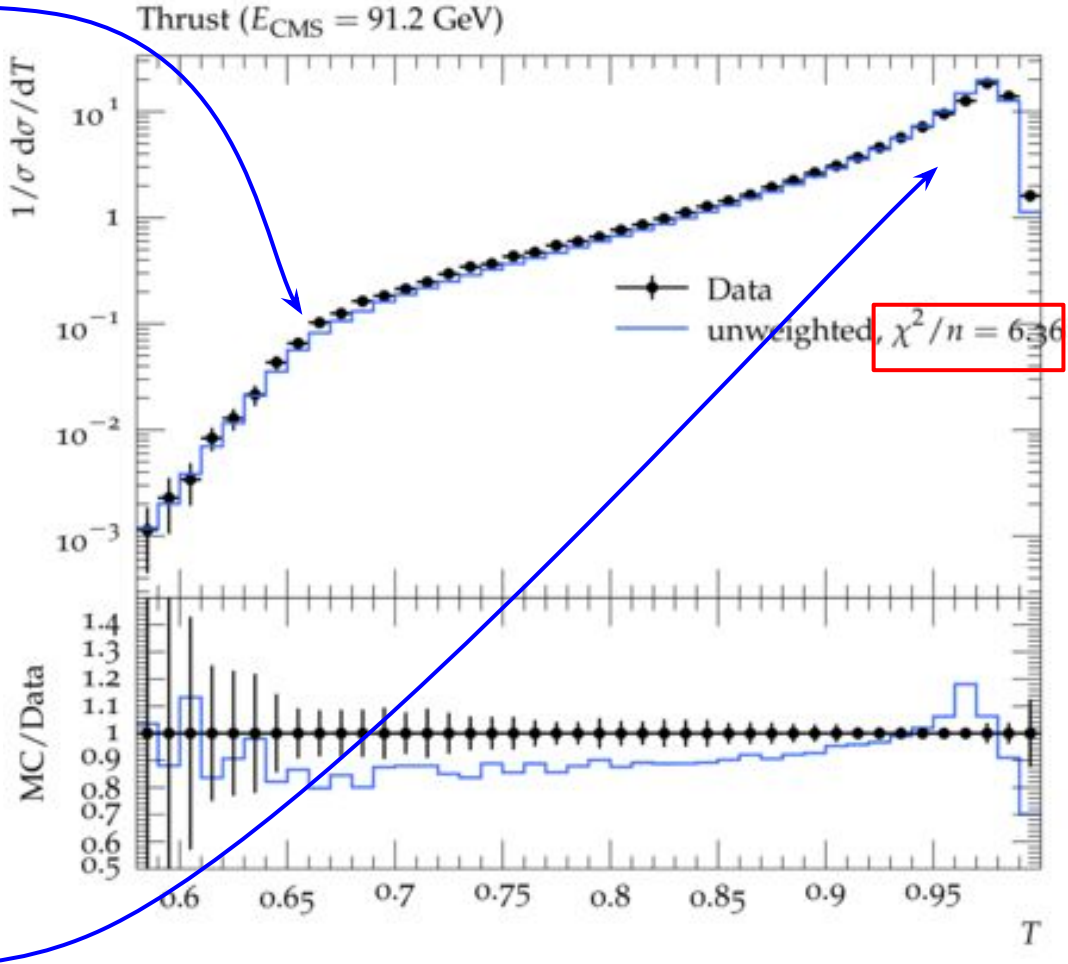
alphaS, aLund, bLund, probBaryon, etc.

Available for parton shower and some hadronization kinematics and flavor parameters

ONE expensive detector simulation, many different physics models

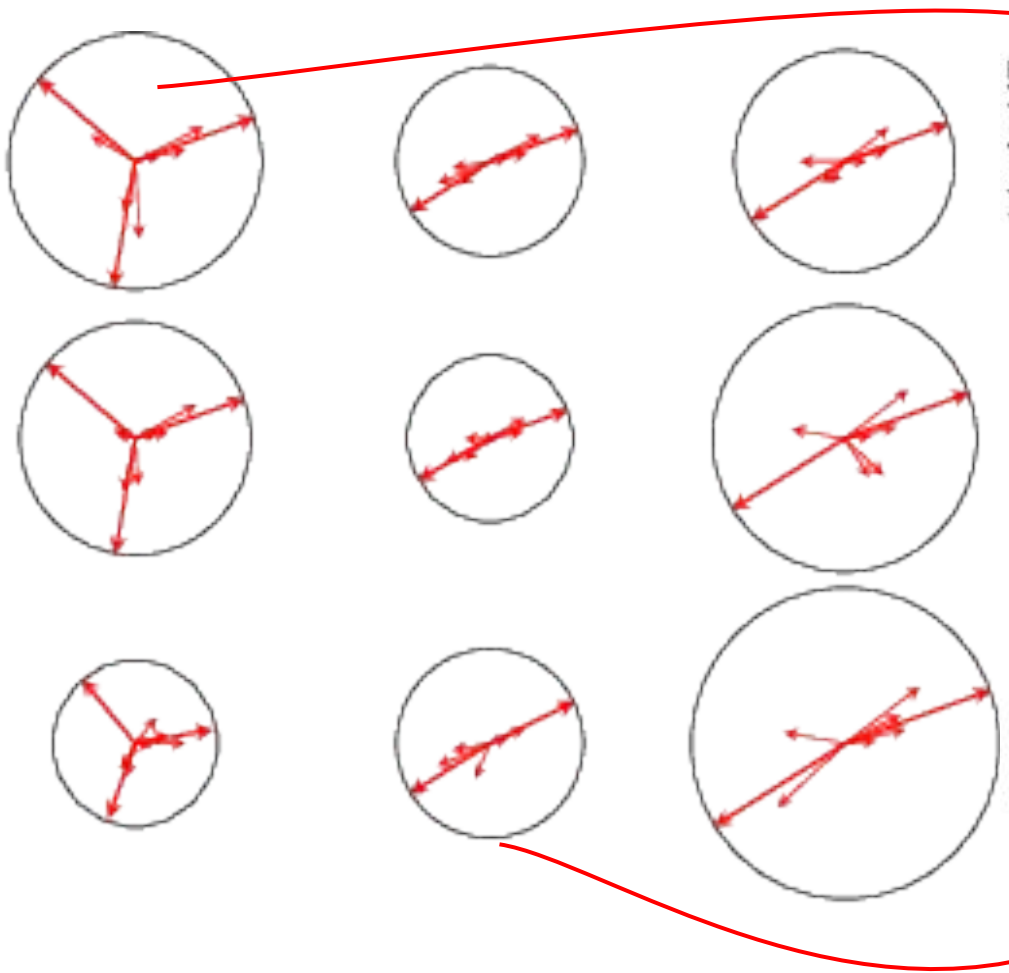
Snapshots of unweighted collider events

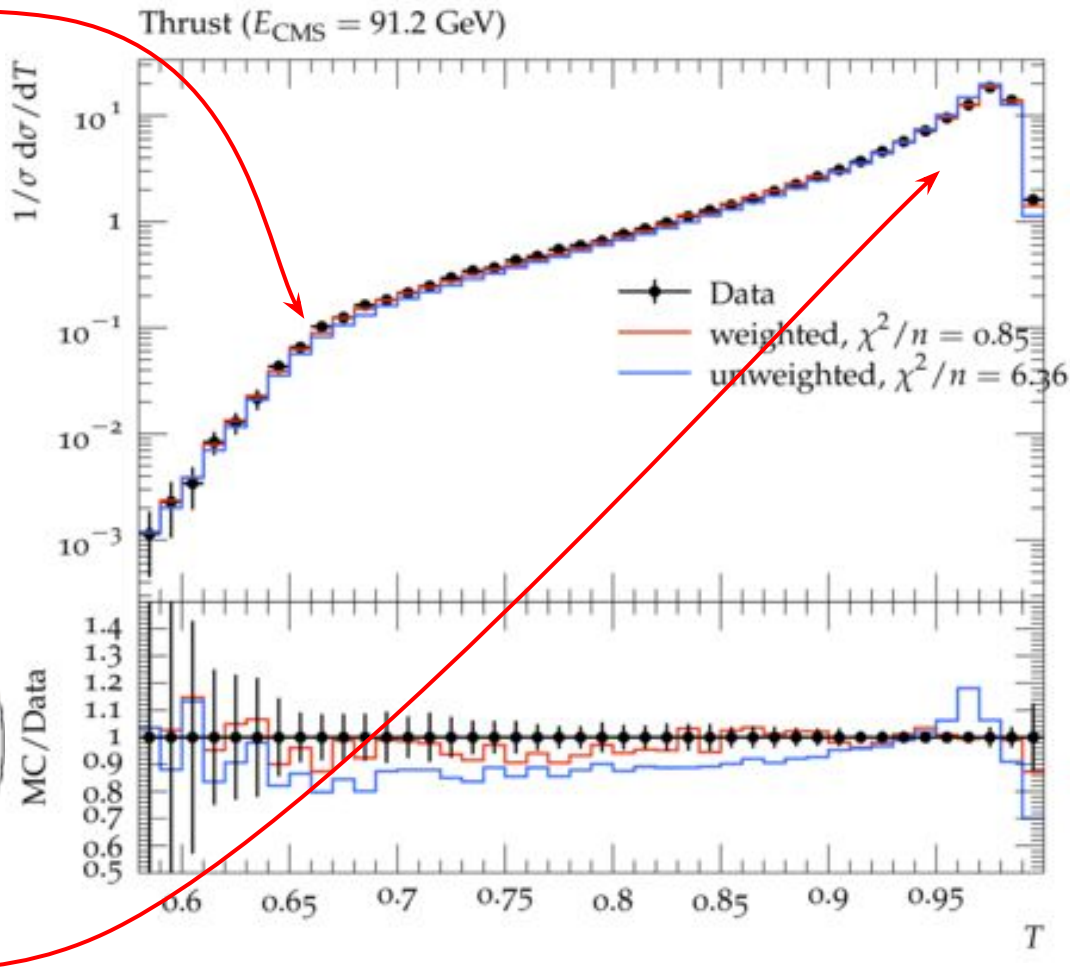




Size of circle represents “weight” of event, here = 1/9

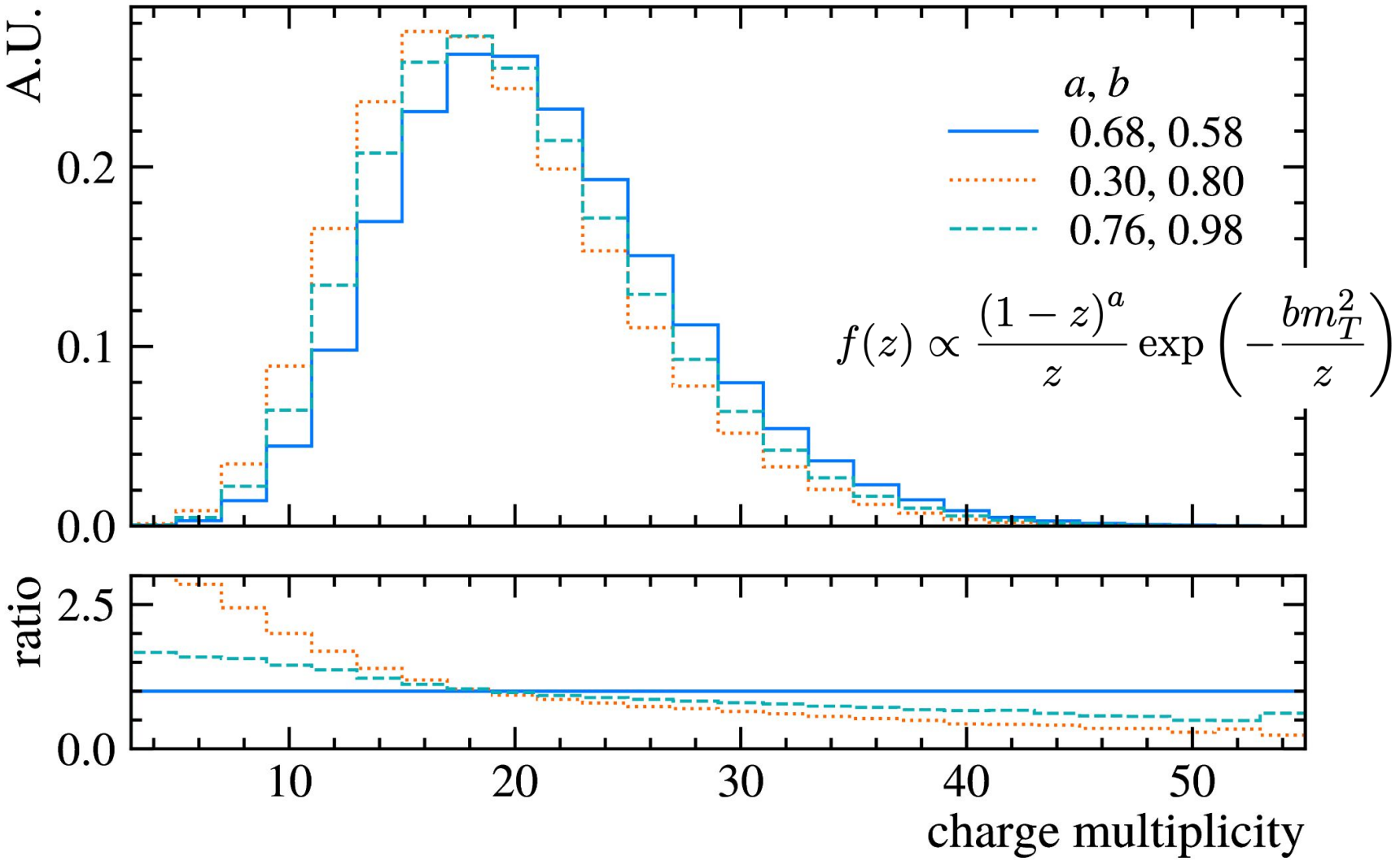
Snapshots of weighted collider events

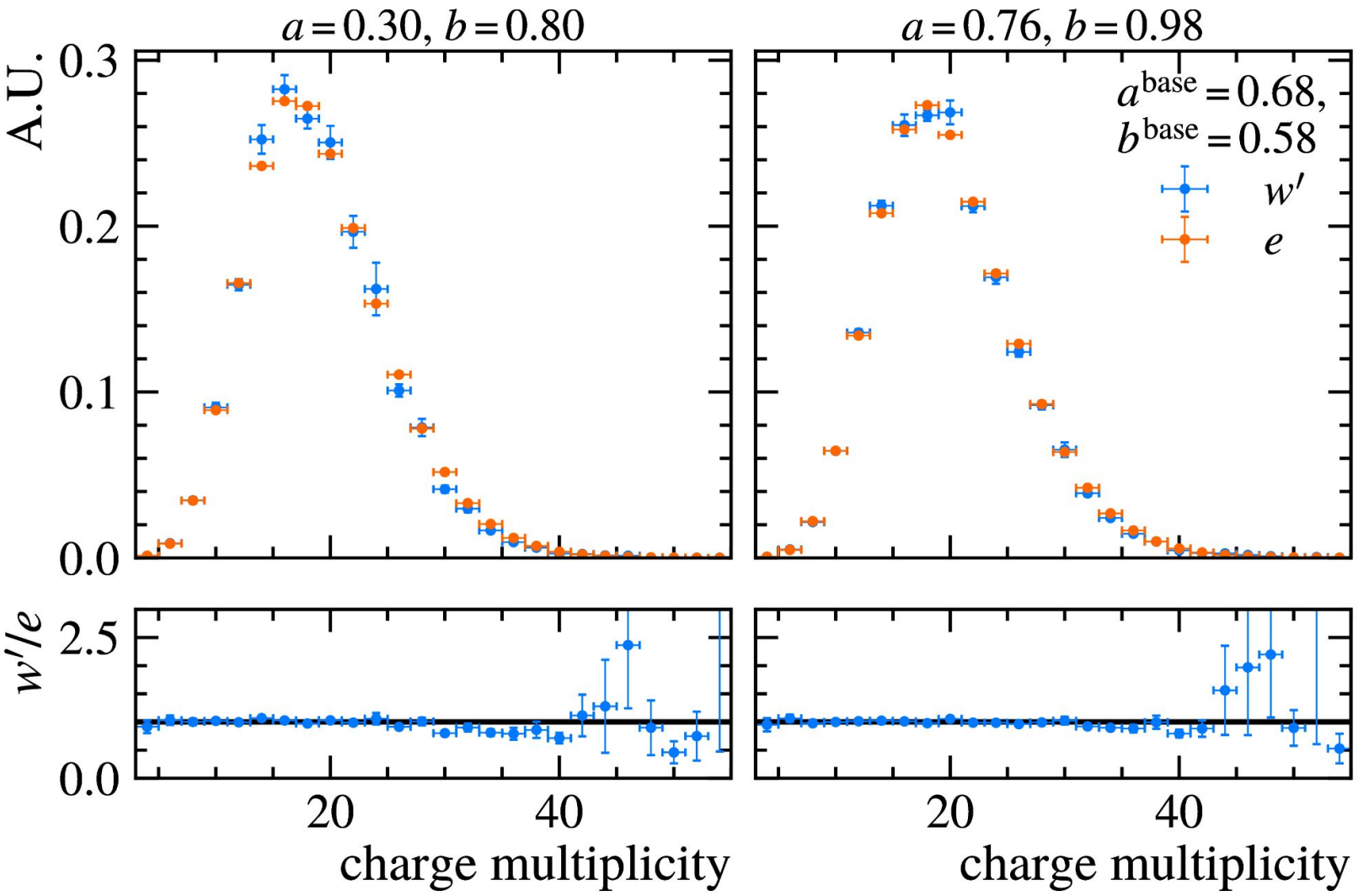




Size of circle represents “weight” of event, here adjusted to match data

LEP collisions




$a = 0.30,\ b = 0.80$
$a = 0.76,\ b = 0.98$
A.U.
0.3
0.2
0.1
0.0
$a^{\text{base}} = 0.68,$
$b^{\text{base}} = 0.58$
w'
e
w'/e
2.5
0.0
20
40
charge multiplicity
20
40
charge multiplicity

Conclusions

MLHAD collaboration driven by brilliant, young people.
Invite them to give more detailed talks. Give them jobs.

MAGIC and HOMER are methods for learning how to modify Lund string model to better match data and learn about hadronization.

Simulated data can be replaced by real data!

Weighting of predictions is a key component of this work.

A reweight-able Pythia makes event generators practically differentiable.