Machine Learning the Top Mass

Katherine Fraser
Department of Physics & Miller Institute
University of California, Berkeley
and Lawrence Berkeley National Lab
kfraser@berkeley.edu

ML4Jets 2024

LPNHE, Paris, France

Based on:
2412.XXXXX with A. Bhattacharya, M. Schwartz

Outline

1. Measuring the Top Mass with Energy Correlators
2. Machine Learning Energy Correlator Space
3. Shape Optimization

Katherine Fraser, kfraser@berkeley.edu

Outline

1. Measuring the Top Mass with Energy Correlators
2. Machine Learning Energy Correlator Space
3. Shape Optimization

Katherine Fraser, kfraser@berkeley.edu

Why Energy Correlators?

Studies of jet substructure have been productive, but often rely on (groomed) observables which cannot be precisely computed. Energy correlators offer a precise alternative.

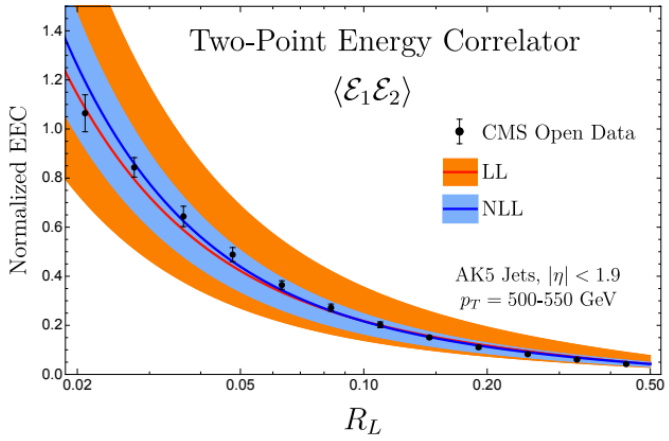
Energy flow operators describe the flux of energy along a particular direction, integrated over time. They provide a theoretical definition of calorimeter cells.

Energy Momentum Tensor

$$\varepsilon(\hat{n}) = \int_0^\infty dt \lim_{r\to\infty} r^2 n^i T_{0i}(t, r\vec{n})$$

Direction

Energy correlators are small angle correlation functions of these operators. They are not defined event wise.



Two point correlation functions obey a definite scaling, while other scales imprint in higher point correlators.

[0803.1467: Hoffman, Maldecena]
[Pietarinen, Stirling, PLB 84; 2004.11381: Chen et al]
[Image - 2205.03414: Lee et al]

Katherine Fraser, kfraser@berkeley.edu

Measuring the Top-quark Mass

Since the Top-quark has a three pronged decay, its mass should be embedded in the E3C. The E3C is

$$E3C(\zeta_{12}, \zeta_{23}, \zeta_{31}) = \sum_{i<j<k} \int d\sigma \frac{E_i E_j E_k}{Q^3} \delta_{\text{shape}}(\zeta_{ij}, \zeta_{jk}, \zeta_{ki})$$

Sum over triplets

in terms of energies (or transverse momenta, for hadron colliders) and pairwise distances ζ_{ij}. The pairwise distances are

$$\hat{\zeta}_{ij} = (1 - \cos(\theta_{ij}))/2 \quad (e^+e^-)$$
$$\hat{\zeta}_{ij} = \Delta\eta_{ij}^2 + \Delta\phi_{ij}^2 \quad \text{(hadron)}$$

[Holguin et al: 2201.08393]

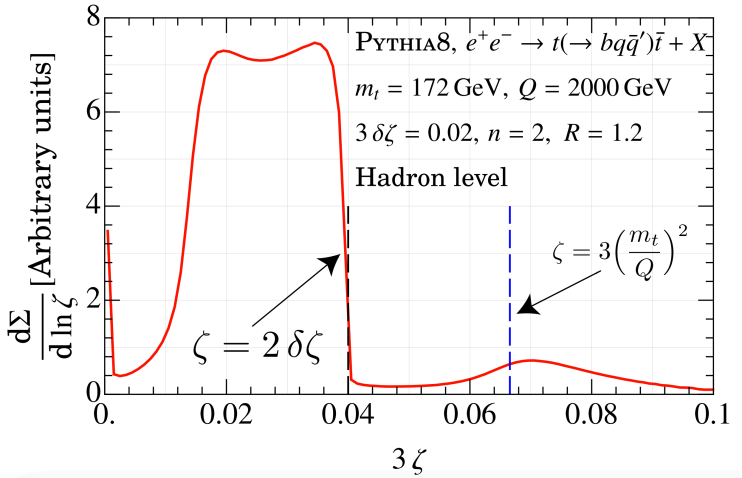
Katherine Fraser, kfraser@berkeley.edu

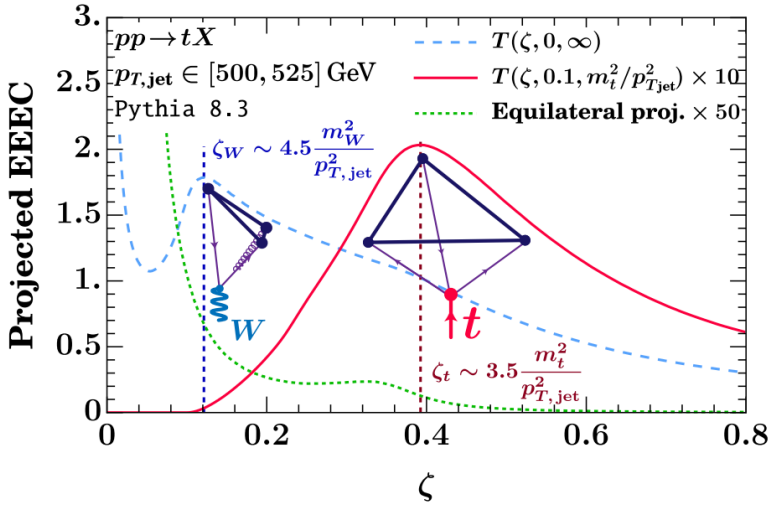
Shape Dependence

In order to compare to precision theory, these are projected onto 1d observables using δ_{shape}. The simplest choice is equilateral

$$\delta_{\text{equilateral}} = \delta(3\zeta - \Sigma\zeta_{pair}) \prod \theta(\delta\zeta - |\zeta_{ij} - \zeta_{jk}|)$$

where $\zeta = (\zeta_{12} + \zeta_{23} + \zeta_{31})/3$





Equilateral triangles are not optimal, because a large fraction of triplets are discarded. Nearly isosceles triangles work better.

Picking an optimal shape is hard!

[Holguin et al: 2311.02157, 2311.14389, 2407.12900]

[Holguin et al: 2201.08393, Xiao et al: 2405.20001]

Katherine Fraser, kfraser@berkeley.edu

Outline

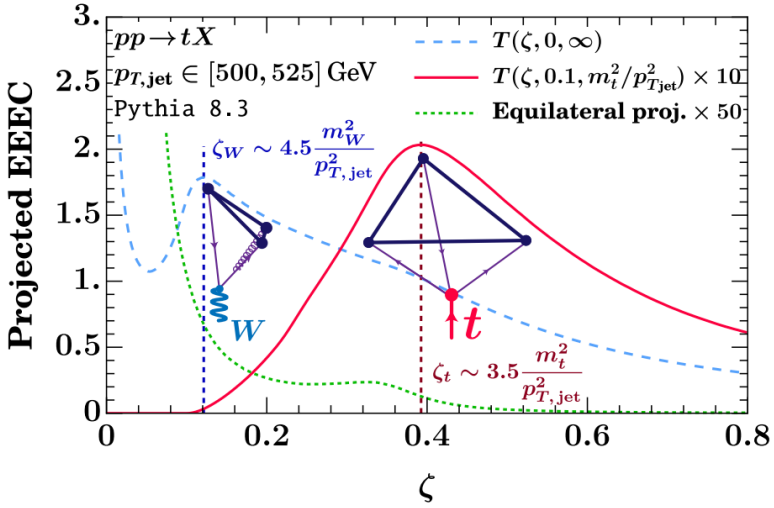
1. Measuring the Top Mass with Energy Correlators
2. Machine Learning Energy Correlator Space
3. Shape Optimization

Katherine Fraser, kfraser@berkeley.edu

Why Add ML?

Measuring m_t by comparing to precision theory calculations requires picking a 1d triangle "shape" to project.

But this choice is arbitrary, and affects the results! Different shapes perform differently. Can we choose the best shape, or avoid picking one all together?



[Image - Holguin et al: 2311.02157]

Yes! With ML, we can:

- Learn the full distribution & avoid shape choice
- Choose an optimal shapes

Katherine Fraser, kfraser@berkeley.edu

A two step process

Our ML approach follows a two step process:

1. Learn the full EC distribution from simulation using an unsupervised approach

Dense Network

- Outputs Energy Correlator $\overline{E}\ p(\zeta_i, m_t \,|\, \text{sim})$ directly
- Trained with energy weighted MLE + explicit integral constraint

Normalizing Flows (NFs)

- Using NFs for density estimation of joint prob $p(\zeta_i, E_i, m_t \,|\, \text{sim})$
- Integrate to get energy correlators

2. Extract the Top-quark mass from the distribution.

Fit to Data Histograms

- Goal: Get optimal performance with as much shape info as possible
- Minimize over m_t^{fit} :

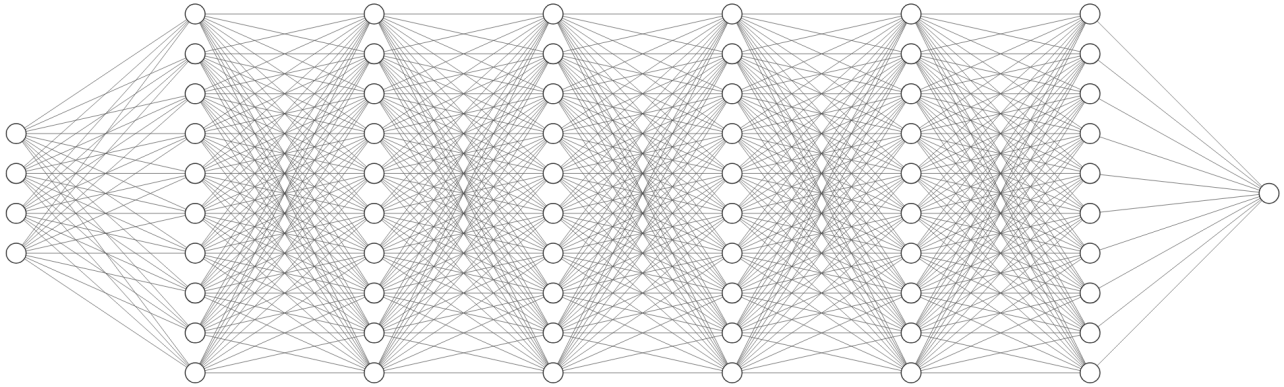
$$\sum_{bins} | EEC_{net}(\zeta_i, m_t^{fit}) - EEC_{data}(\zeta_i, m_t) |$$

Shape Distillation

- Goal: Choose optimal shape which can be precisely computed
- Use neural ratio estimation

Katherine Fraser, kfraser@berkeley.edu

The Dense Net Approach



$$\text{Loss} = \sum_i E_i \ln NN(\zeta_i, m_t) + \lambda \times \log \int d\zeta_i dm_t NN(\zeta_i, m_t)$$

This is a modified version of maximum likelihood estimation:

$$\begin{aligned}\text{Loss} &= \sum_{ij} E_{ij} \ln NN(\zeta_i, m_t) \\ &= \ln \Pi_{ij} NN(\zeta_i, m_t)^{E_{ij}} \\ &= \ln \Pi_i NN(\zeta_i, m_t)^{\sum_j E_{ij}}\end{aligned}$$

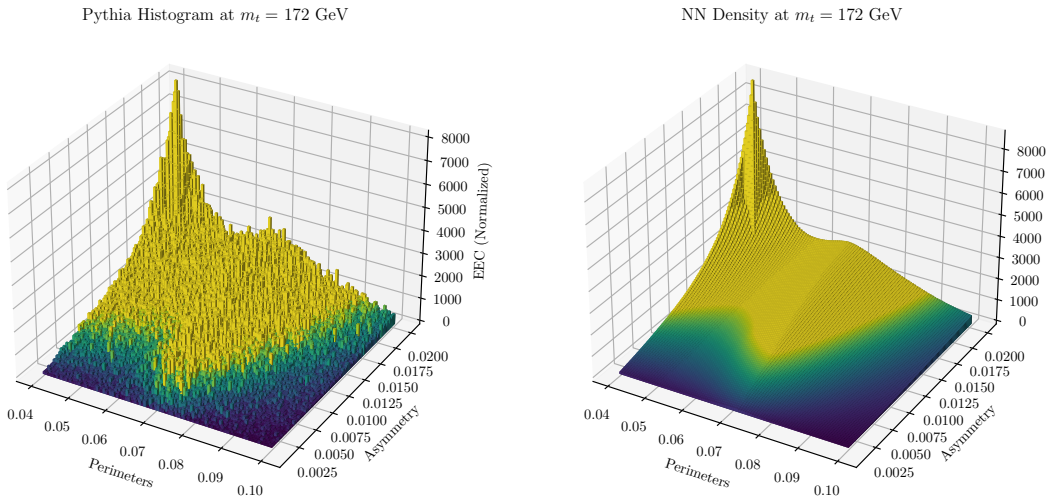
Details:

- Very sensitive to hyperparameters
- Trained with AdamW with weight decay and gradient clipping
- Need to preprocess inputs to be $\mathcal{O}(1)$
- Integral computed with gpu Torchquad
- Use stochastic weight averaging after half the epochs

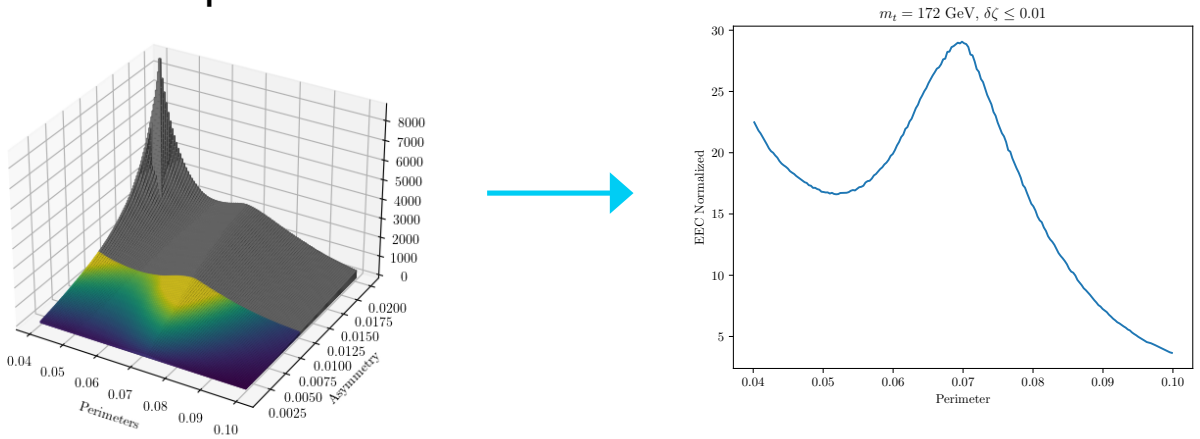
Katherine Fraser, kfraser@berkeley.edu

Dense Net Results

The dense net learns the full distribution differential in ζ_1, ζ_2, ζ_3, and m_t.



Getting the 1D projected ECs simply requires integrating the full distribution over a wedge determined by δ_{shape}.

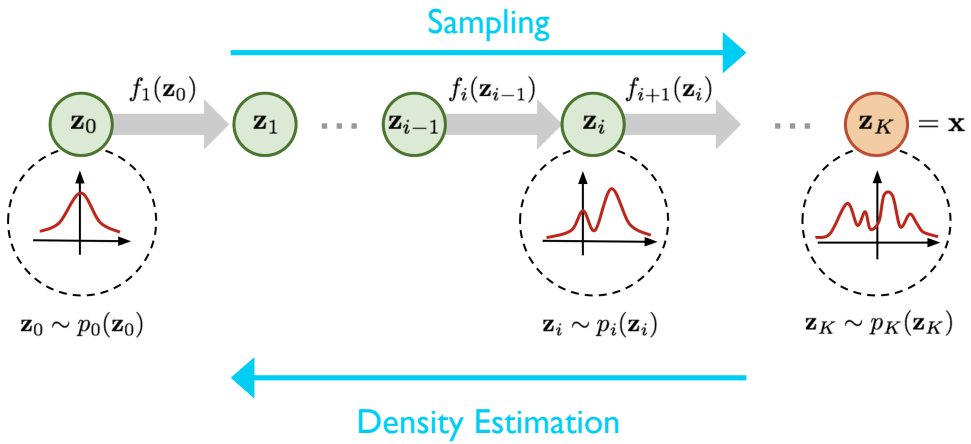


Katherine Fraser, kfraser@berkeley.edu

Normalizing Flows Background

Normalizing flows are an invertible map between a Gaussian distribution and a more complicated one. One direction estimates the density, while the other samples.

The loss is again the negative log likelihood, but normalization is now automatic.



Details:

- Spline coupling flow with 3 layers, each followed by a batch norm.
- Trained with Adam for 20 epochs
- Inputs with $\zeta < 0.005$ masked
- All inputs rescaled using logit

[Papamakarios et al: 1906.04032]
[Image - LiLian Wang 2018]

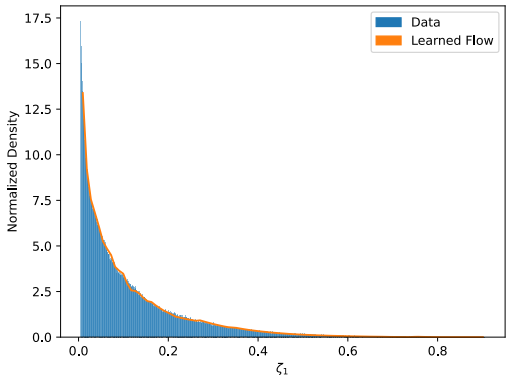
Katherine Fraser, kfraser@berkeley.edu

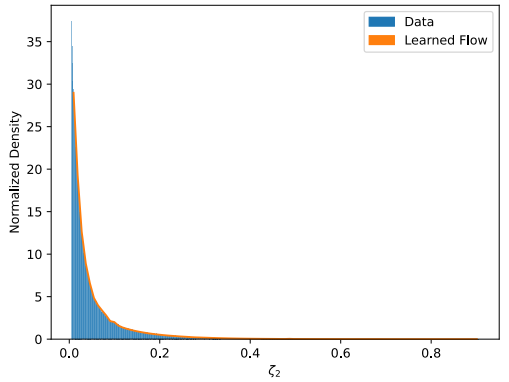
Normalizing Flows

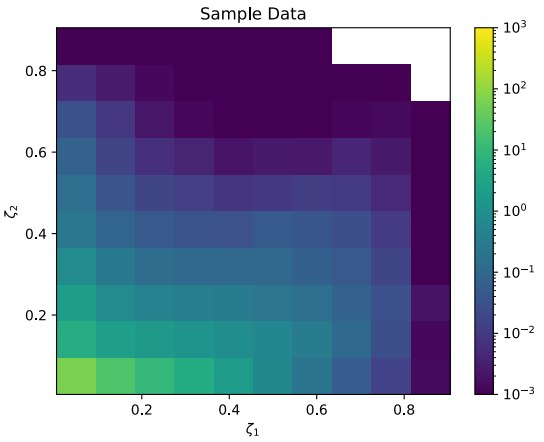
We learn the 6d joint probability $p(\zeta_i, E_i)$ with NFs. We are working on the 7d probability $p(\zeta_i, E_i, m_t)$.

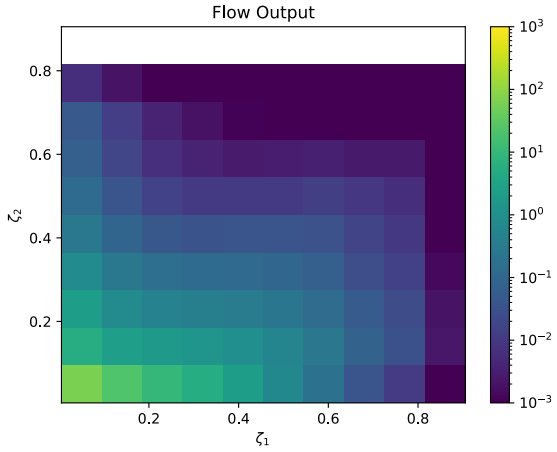
We multiply by and integrate over E_i to compute the 3d ECs. 1d correlators are computed by integrating slices over ζ_i.

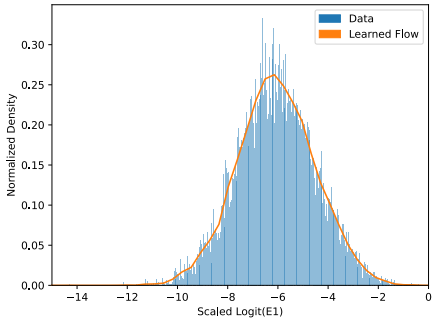
This avoids integration during training, but requires a very precise distribution since energy weighting isn't naturally learned.











Katherine Fraser, kfraser@berkeley.edu

Outline

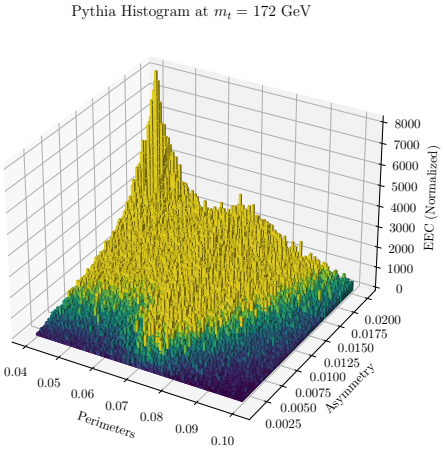
1. Measuring the Top Mass with Energy Correlators
2. Machine Learning Energy Correlator Space
3. Shape Optimization

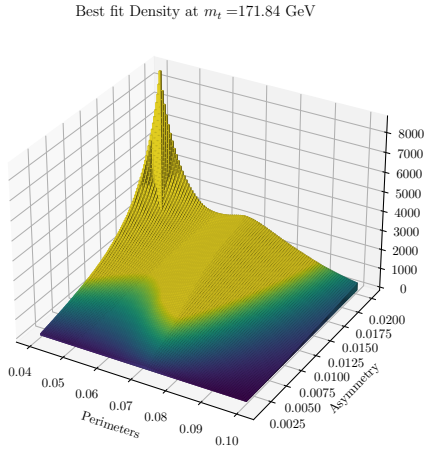
Katherine Fraser, kfraser@berkeley.edu

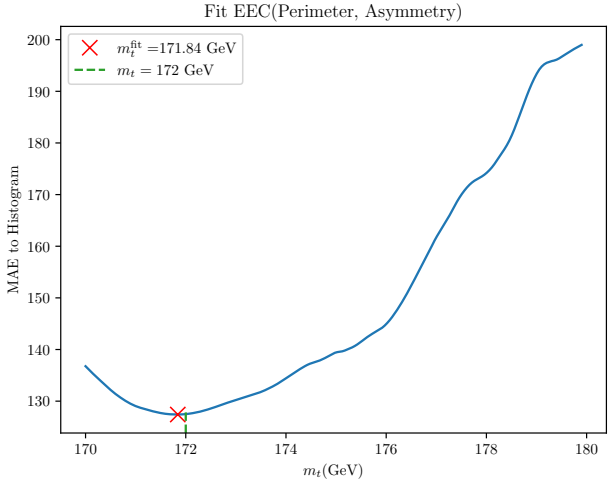
Fit to Data Histograms

- Goal: to find the best performing energy correlator (projection), even if it can't be compared directly to data, including both the full distribution and lower dimensional marginalizations.
- Compare EC value the network learns for an arbitrary m_t^{fit} value to a histogram with fixed m_t value
- True value of m_t^{fit} should minimize

$$\sum_{bins} |EEC_{net}(\zeta_i, m_t^{fit}) - EEC_{data}(\zeta_i, m_t)|$$

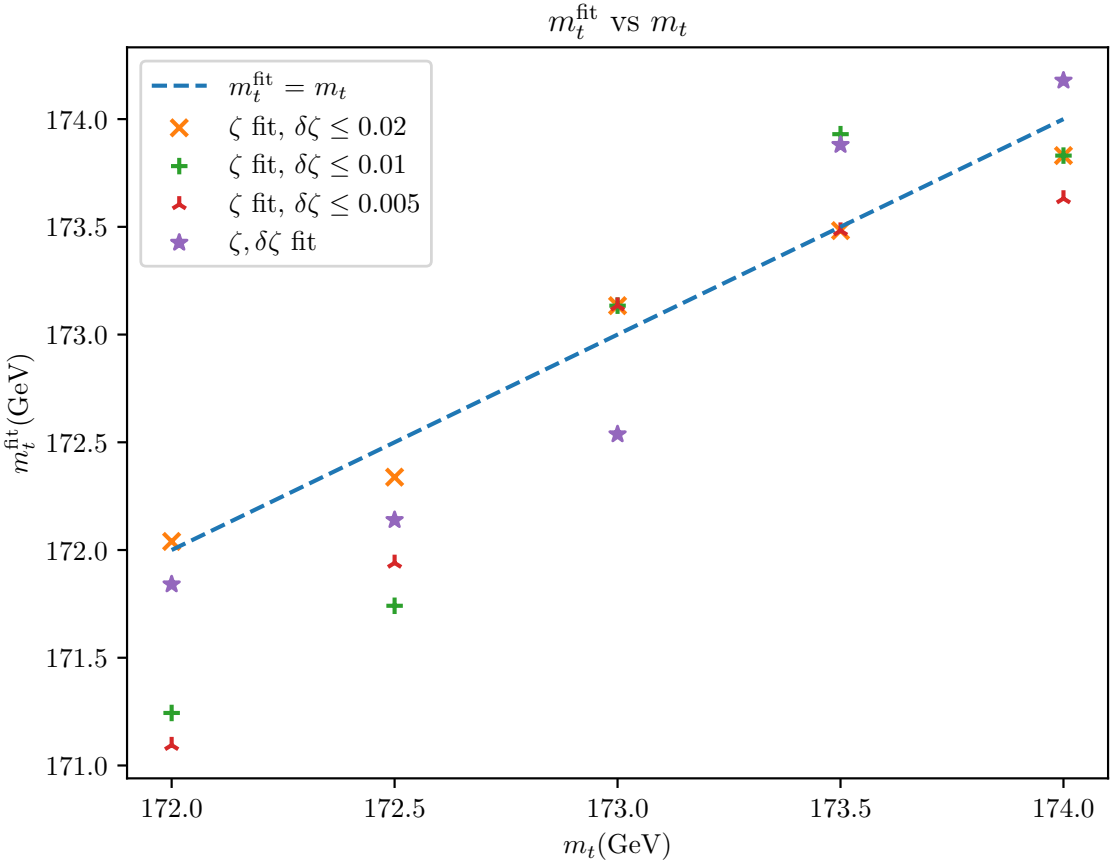






Katherine Fraser, kfraser@berkeley.edu

Top Mass Results I



Top Mass Errors:

ζ fit, $\delta\zeta \leq 0.02$: 0.104 GeV

ζ fit, $\delta\zeta \leq 0.01$: 0.45 GeV

ζ fit, $\delta\zeta \leq 0.005$: 0.396 GeV

ζ fit, 2D: 0.308 GeV

Katherine Fraser, kfraser@berkeley.edu

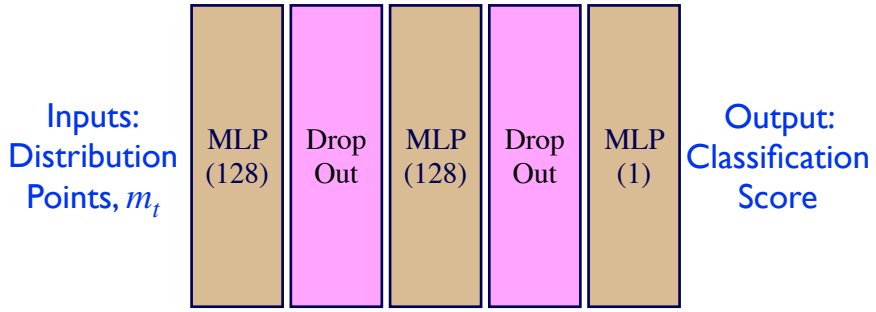
Neural Ratio Estimation

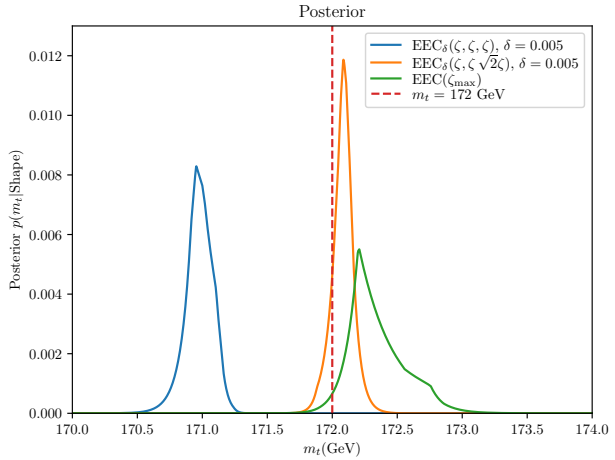
NRE learns the likelihood to evidence ratio

$$\frac{p(\theta, data)}{p(\theta)p(data)}$$

via a classifier which distinguishes between samples with true m_t and samples with fake m_t drawn from a uniform distribution.

Since we know $p(\theta)$, this can be used to extract $p(\theta \,|\, data)$ for a new data sample once the classifier is trained. The true m_t maximizes the posterior.

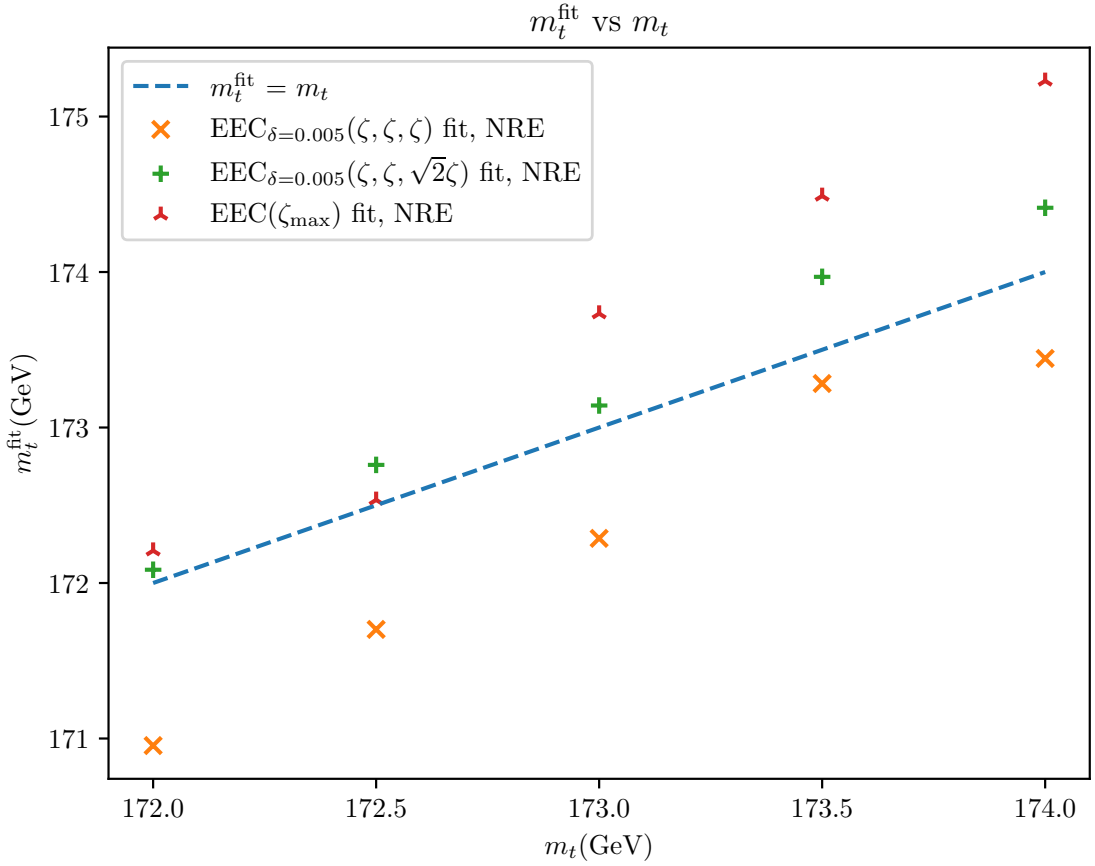




[Hermans et al: 1903.04057]

Katherine Fraser, kfraser@berkeley.edu

Top Mass Results II



Top Mass Errors:

NRE, $(\zeta,\zeta,\zeta)_{\delta=0.005}$: 0.668 GeV

NRE, $(\zeta,\zeta,\sqrt{2}\zeta)_{\delta=0.005}$: 0.274 GeV

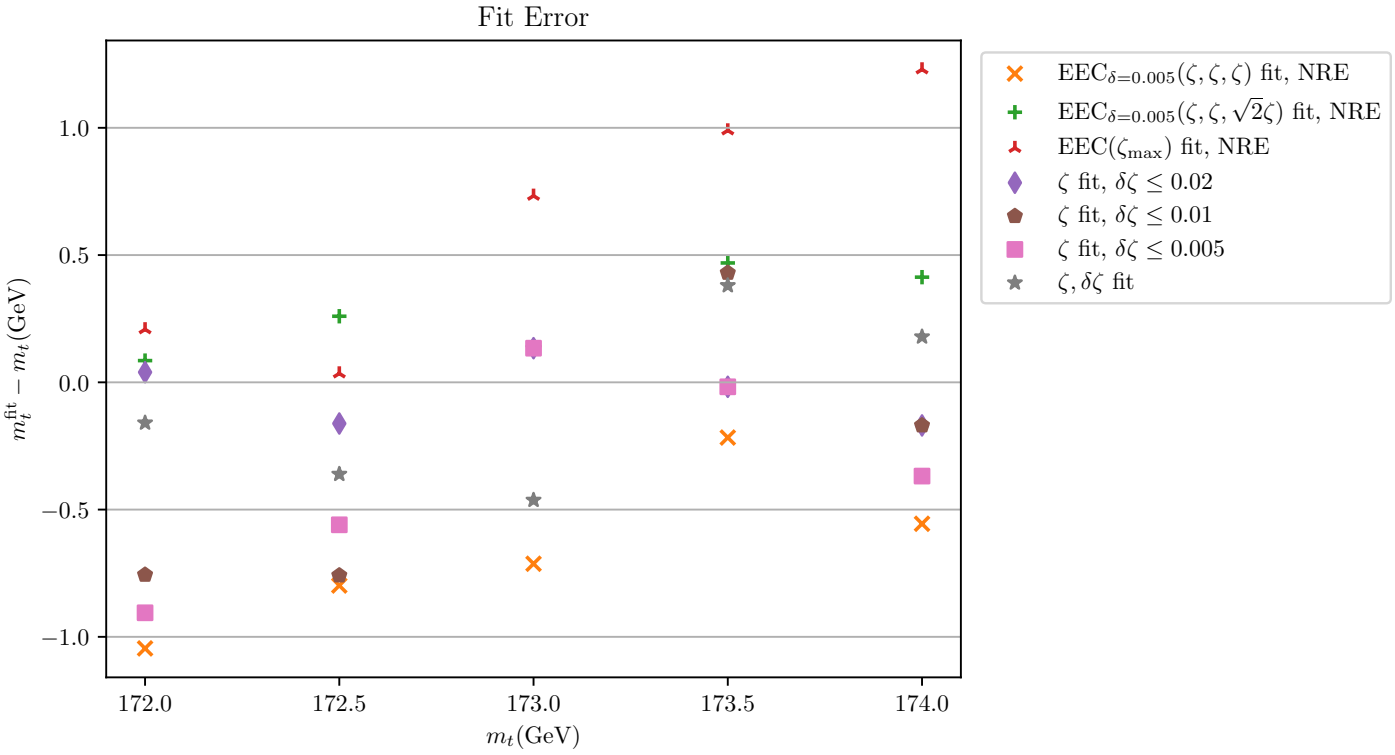
NRE, (ζ_{max}): 0.638 GeV

Katherine Fraser, kfraser@berkeley.edu

Conclusions

ML can be used to learn the full energy correlator distribution.

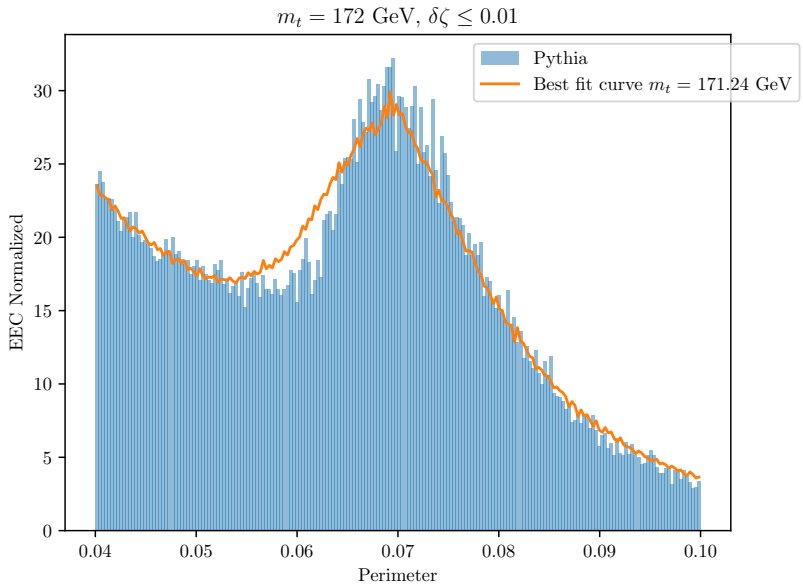
This distribution can be used to optimize the 1d Energy Correlator projection, and also to directly put a bound on how well the Top Mass can be extracted from the full correlator.

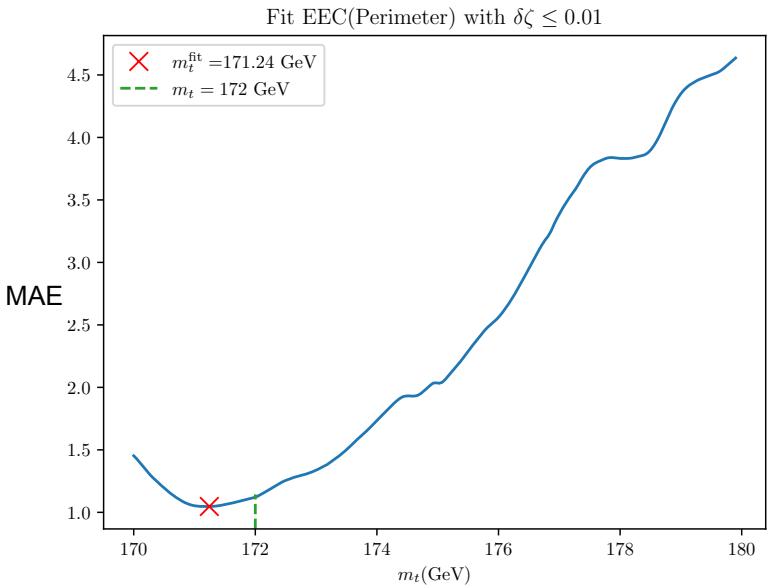


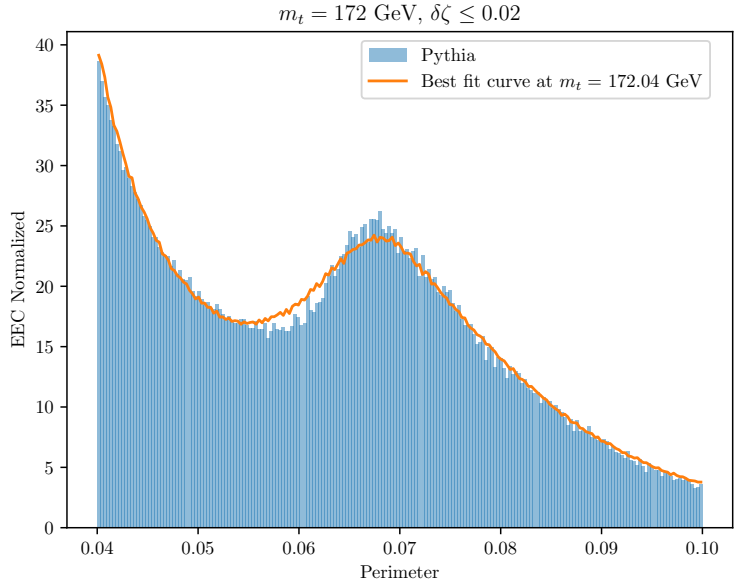
Katherine Fraser, kfraser@berkeley.edu

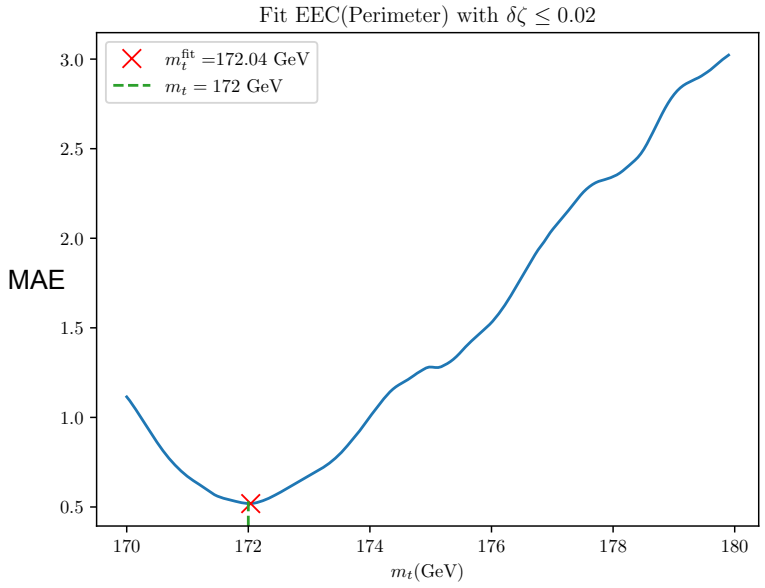
Back Up Slides

Katherine Fraser, kfraser@berkeley.edu









Katherine Fraser, kfraser@berkeley.edu