Novel Jet Observables from Machine Learning

Kaustuv Datta[a, b], Andrew Larkoski[b] and Benjamin Nachman[c]

[a] ETH Zürich

[b] Reed College

[c] Lawrence Berkeley National Laboratory

Outline

Can we use ML to identify how much information in a jet is useful for discrimination?

How do we extract this information?

Can we use this information to build powerful new observables ‘by hand’?

Can we automate the building of new observables for jet discrimination, using ML?

Can we generalize this procedure for higher dimensional inputs that resolve more differential phase space?

Part 1: Can we characterize the dimensionality of information inside a jet using machine learning?

How Much Information is in a Jet?

To answer this we require an organizing principle:

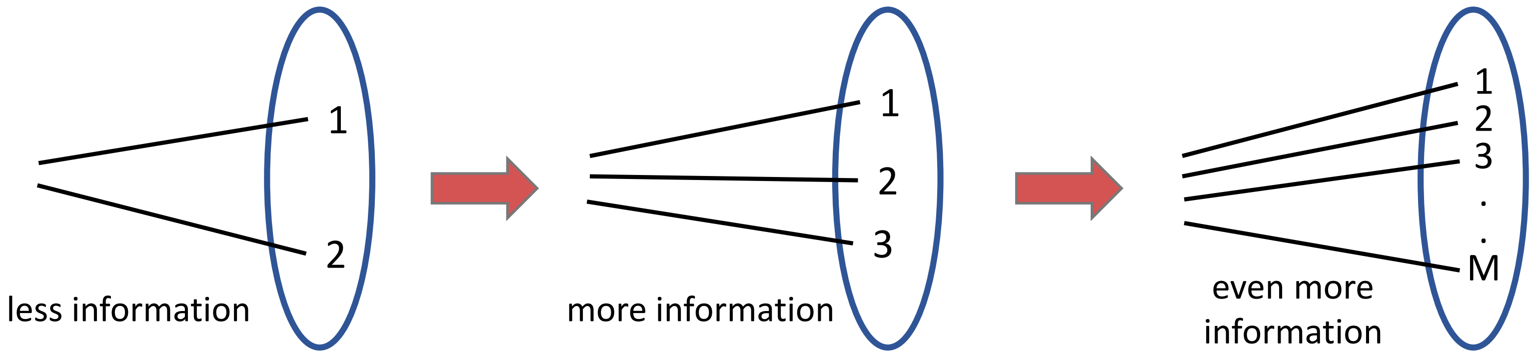
- Directly calculable:

$$\tau_N^{(\beta)} = \frac{1}{p_{T_J}} \sum_{i\in \text{Jet}} p_{T_i} \min\left\{\Delta R_{1i}{}^{\beta}, \Delta R_{2i}{}^{\beta}, \dots, \Delta R_{Ni}{}^{\beta}\right\}$$

Directly identifies *N* subjet directions in a jet

Measure a collection of 3*M* – 4 N-subjettiness observables

- Systematically improvable



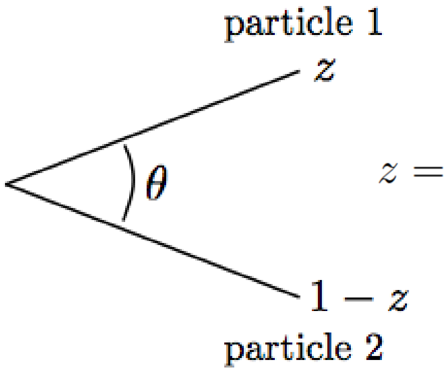
Minimal and complete bases of observables that resolve *M*-body phase space of particles in a jet

ie., when $M = n_{\text{particles}}$, one is sensitive to all of the information in the jet kinematics

From 2- to *M*-body phase space

2-body phase space:

(3x2 - 4) = 2 dimensional



$$z = \frac{p_{T1}}{p_{TJ}} \qquad 1 - z = \frac{p_{T2}}{p_{TJ}}$$

<u>Phase space defined by *z* and θ</u>

Measure two 1-subjettiness:

$$\tau_1^{(1)} = 2z(1-z)\theta, \;\; \tau_1^{(2)} = z(1-z)\theta^2$$

Can be inverted to find *z* and θ:

$$z(1-z) = \frac{\left(\tau_1^{(1)}\right)^2}{4\tau_1^{(2)}}, \;\; \theta = \frac{2\tau_1^{(2)}}{\tau_1^{(1)}}$$

M-body phase space is

(3M - 4) dimensional *

Measure more N-subjettiness to resolve *M*-body phase space

$$\left\{\tau_1^{(0.5)}, \tau_1^{(1)}, \tau_1^{(2)}, \tau_2^{(0.5)}, \tau_2^{(1)}, \tau_2^{(2)}, \ldots, \tau_{M-2}^{(0.5)}, \tau_{M-2}^{(1)}, \tau_{M-2}^{(2)}, \tau_{M-1}^{(1)}, \tau_{M-1}^{(2)}\right\}$$

(3*M* – 4) observables

Require:

M - 1 transverse momentum fractions

2*M* - 3 pairwise angles between subjets

<u>This is a set of IRC safe observables with minimal and complete information for the phase space of jet substructure for an *M* particle jet</u>

* Proof of this is an application of the Euler Characteristic formula: V – E + F = 2 [arXiv:1704.08249]

Saturation of Discrimination Power

<u>*M*-body phase space machine learning:</u>

Inputs: jet mass + 2-, 3-, 4-,5- or 6-body sets of observables

Calculate ROC curves for neural networks trained on each

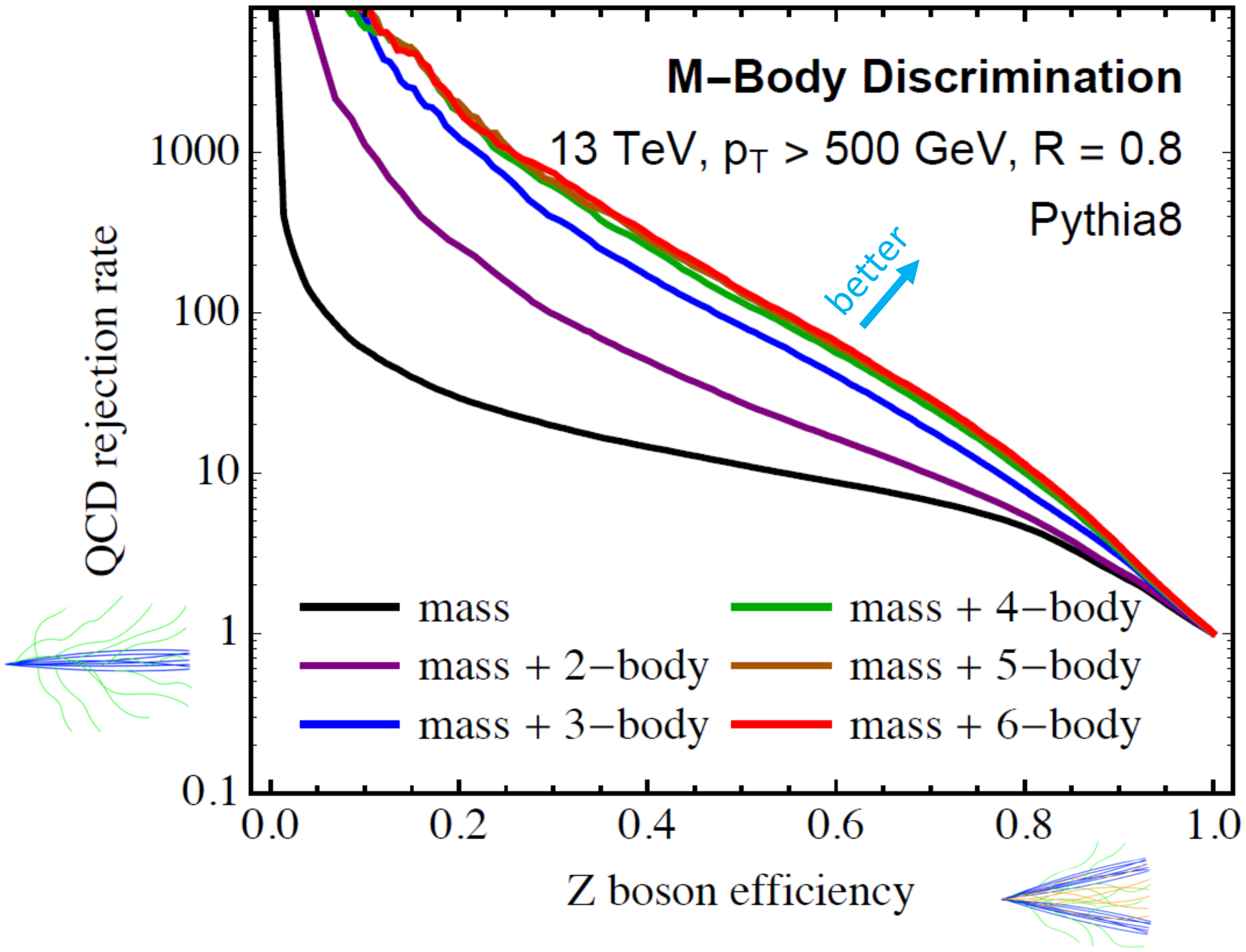
<u>For *Z* boson vs QCD discrimination</u>

Saturation is observed at 4-body
(3*4 - 4 = 8 dimensional) phase space

Only a small amount of information is useful for discrimination

Exploit this fact to construct powerful new observables

arXiv:1704.08249



Part 2: Can we use ML to make new, powerful discrimination observables with a compact, analytic form?

Making new observables

Determine the *M*-body phase space at which discrimination power saturates

$$\left\{\tau_1^{(0.5)}, \tau_1^{(1)}, \tau_1^{(2)}, \tau_2^{(1)}, \tau_2^{(2)}\right\}$$

Construct a function of these 3*M* - 4 variables with tunable parameters

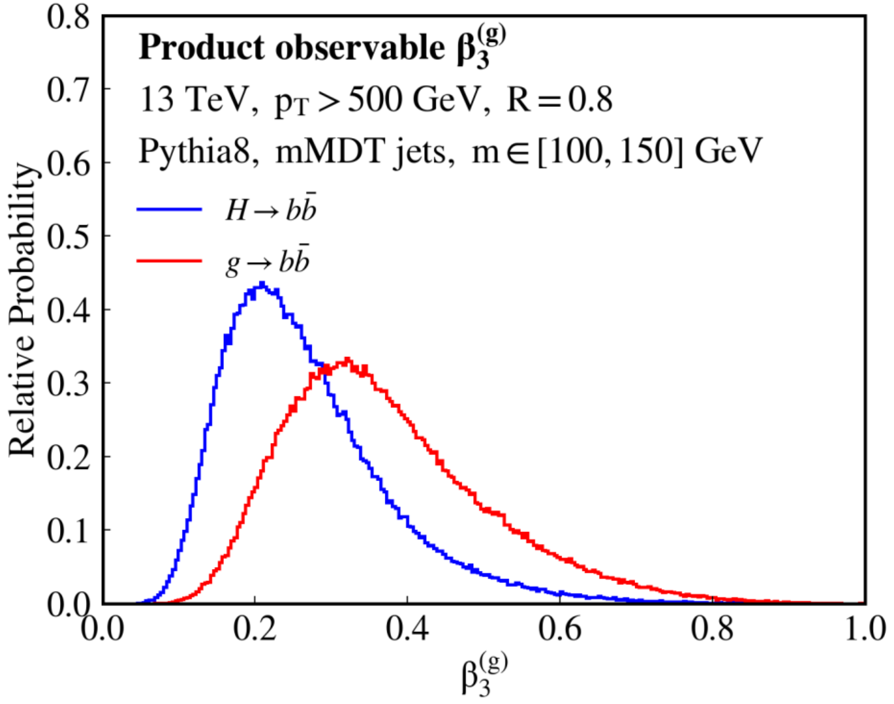
$$\beta_3 = \left(\tau_1^{(0.5)}\right)^a \left(\tau_1^{(1)}\right)^b \left(\tau_1^{(2)}\right)^c \left(\tau_2^{(1)}\right)^d \left(\tau_2^{(2)}\right)^e$$

Fix parameters by maximizing a discrimination metric

such as the area under the ROC curve

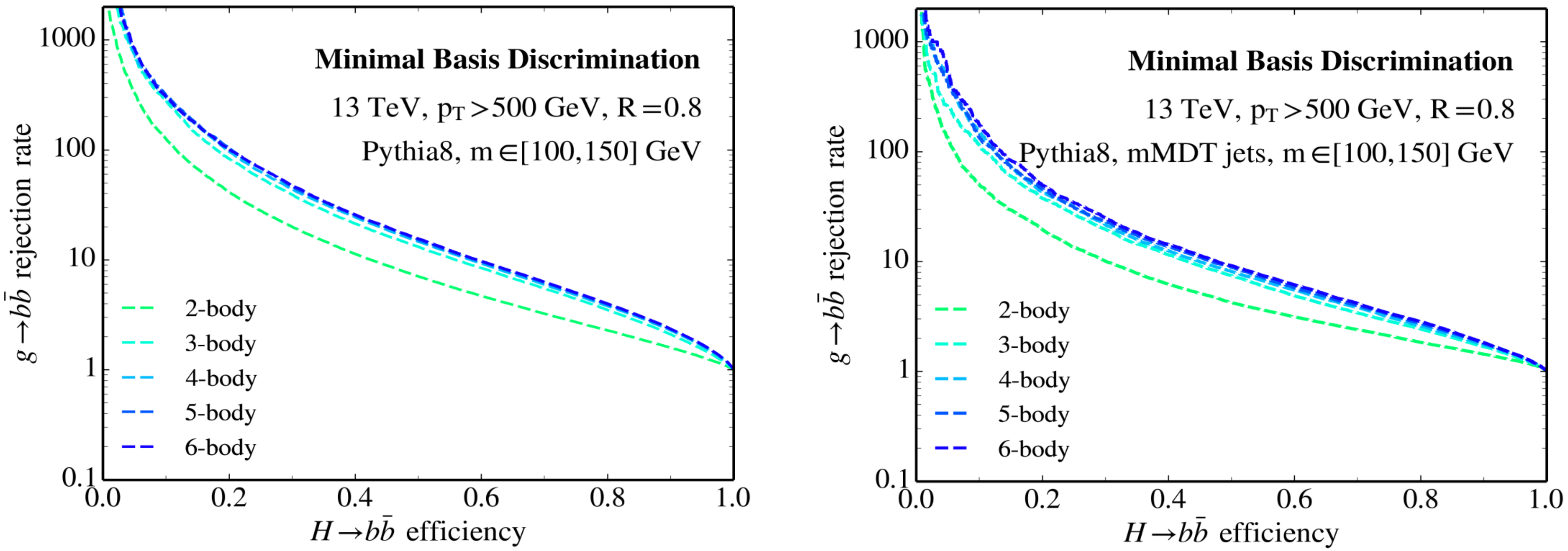
Choice of product form for the observable means function is sufficiently simple

to be studied analytically, and hopefully flexible enough to saturate performance



Example: $H \to b\bar{b}$ vs. $g \to b\bar{b}$ discrimination

Observe saturation of discrimination power at 3-body (5 dim.) phase space



Construct the product observable from the 5 phase space coordinates

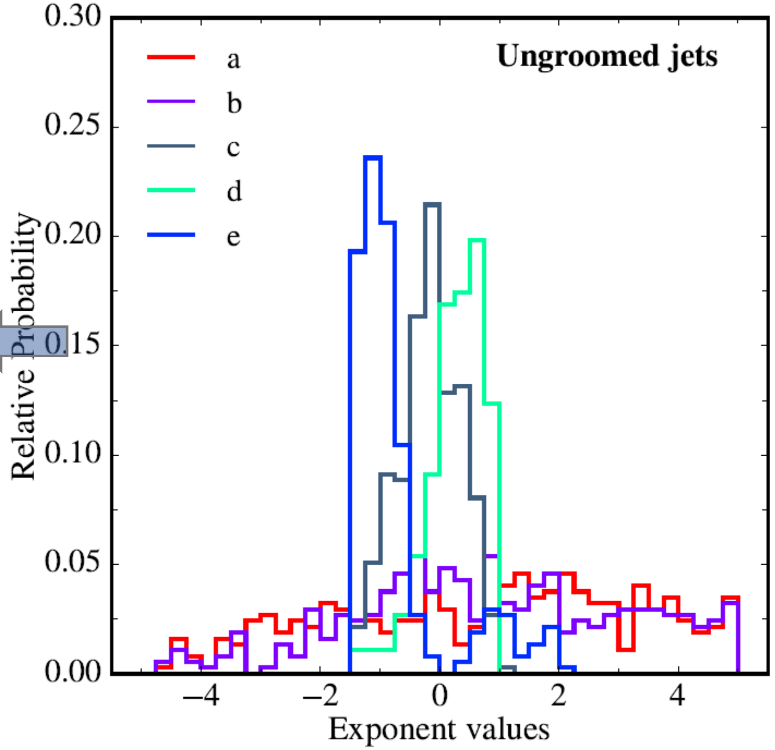
$$\beta_3 = \left(\tau_1^{(0.5)}\right)^a \left(\tau_1^{(1)}\right)^b \left(\tau_1^{(2)}\right)^c \left(\tau_2^{(1)}\right)^d \left(\tau_2^{(2)}\right)^e$$

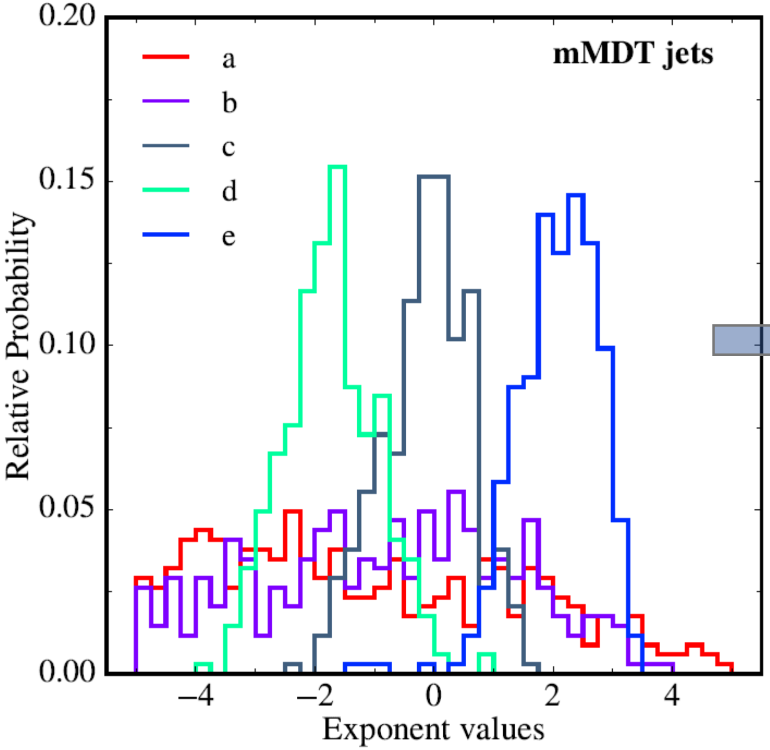
Example: $H \to b\bar{b}$ vs. $g \to b\bar{b}$ discrimination

To identify the optimal parameters for the problem:

Perform a random uniform scan over a fixed range (here, [-5,5]) for each parameter

$$\beta_3 = \frac{\left(\tau_1^{(0.5)}\right)^2 \left(\tau_2^{(1)}\right)^{0.5}}{\tau_2^{(2)}}$$





$$\beta_3^{(g)} = \left(\frac{\tau_2^{(2)}}{\tau_1^{(0.5)}\tau_2^{(1)}}\right)^2$$

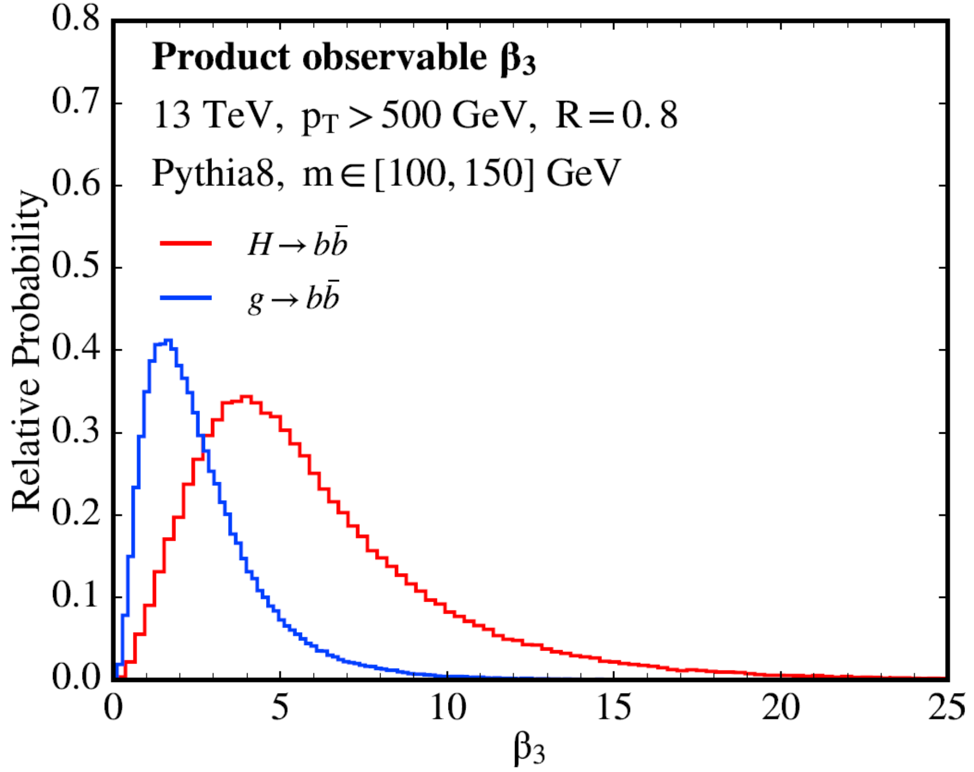
Observe the point in the input space that maximizes the metric from histograms of values of the exponents above a certain threshold cut on the metric space

Example: $H \to b\bar{b}$ vs. $g \to b\bar{b}$ discrimination

Final functional form for new observables:

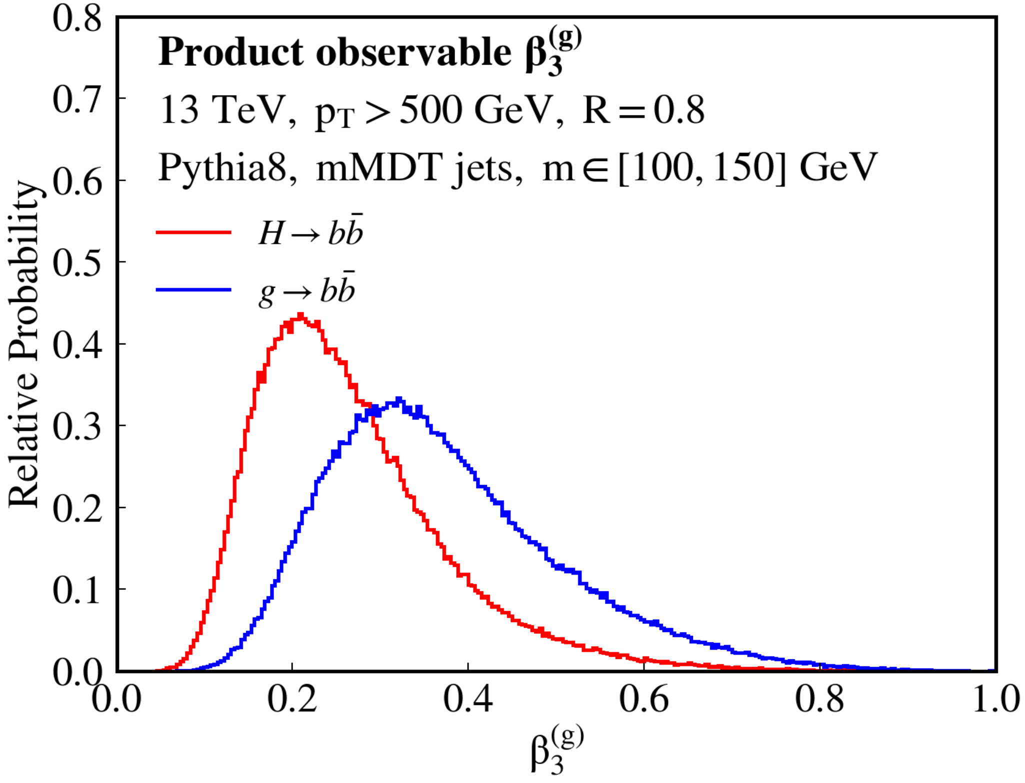
Ungroomed

$$\beta_3 = \frac{\left(\tau_1^{(0.5)}\right)^2 \left(\tau_2^{(1)}\right)^{0.5}}{\tau_2^{(2)}}$$



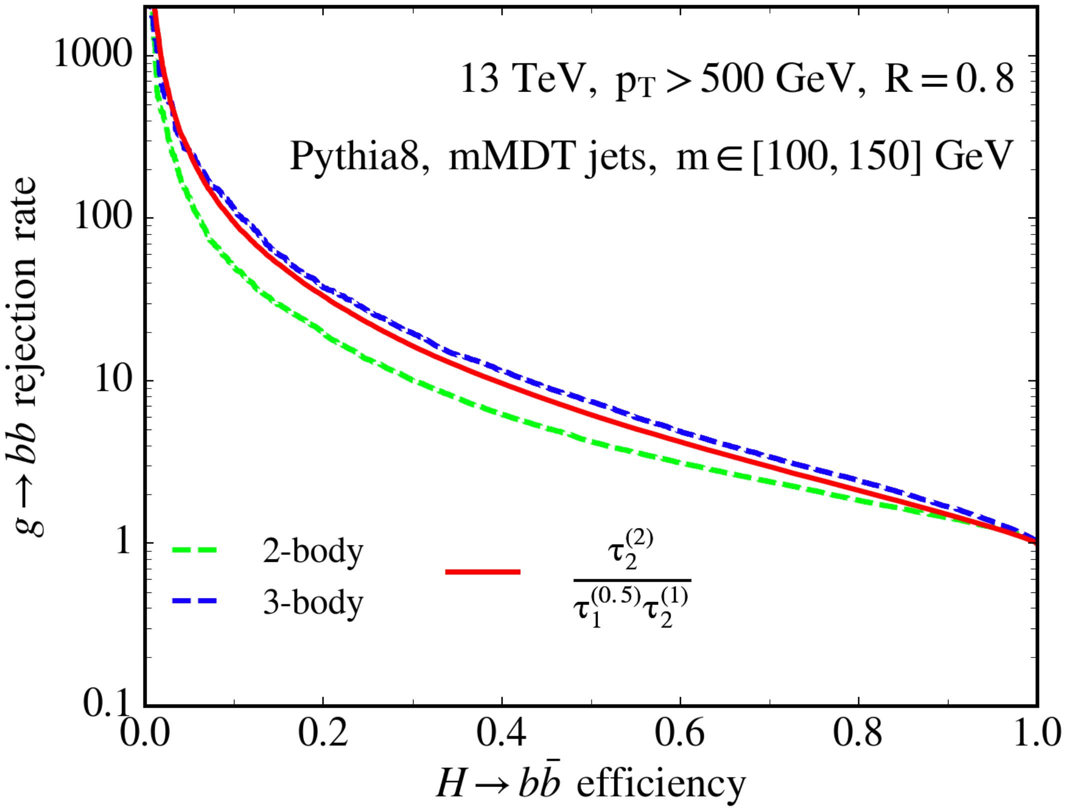
mMDT groomed

$$\beta_3^{(g)} = \frac{\tau_2^{(2)}}{\tau_1^{(0.5)}\tau_2^{(1)}}$$



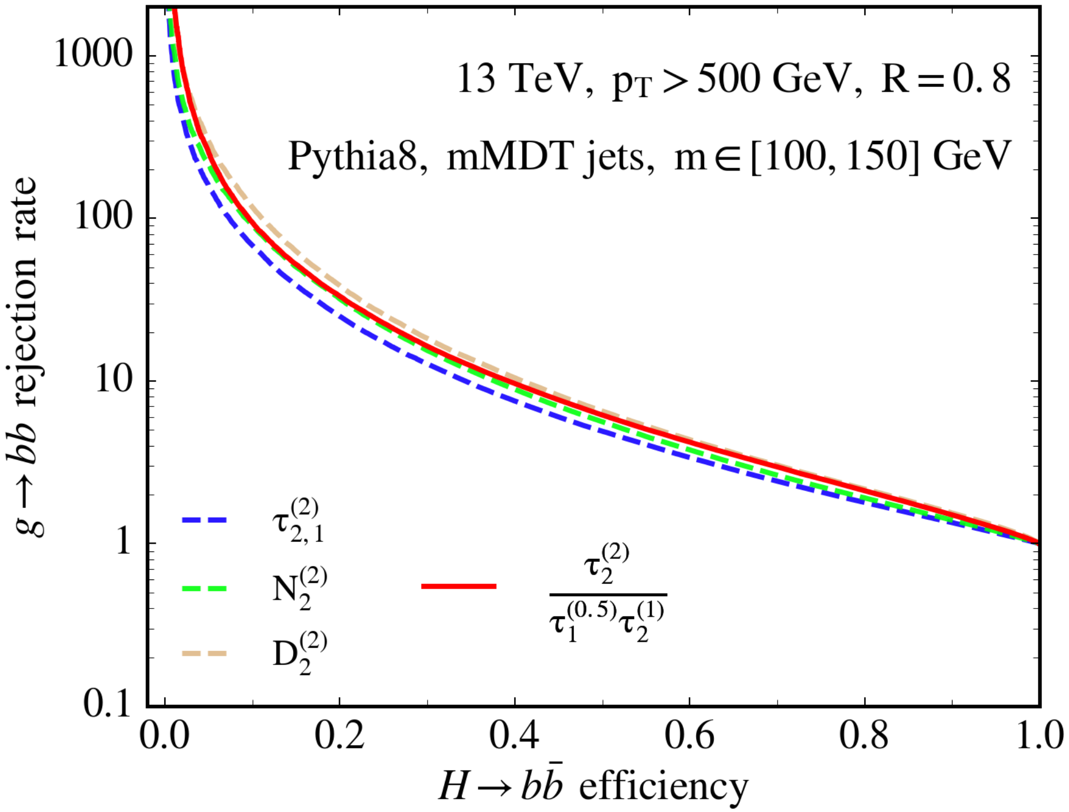
Example: Groomed $H \to b\bar{b}$ vs. $g \to b\bar{b}$ discrimination

Comparison to neural network



Recovers most of the 3-body phase space information identified by the neural network

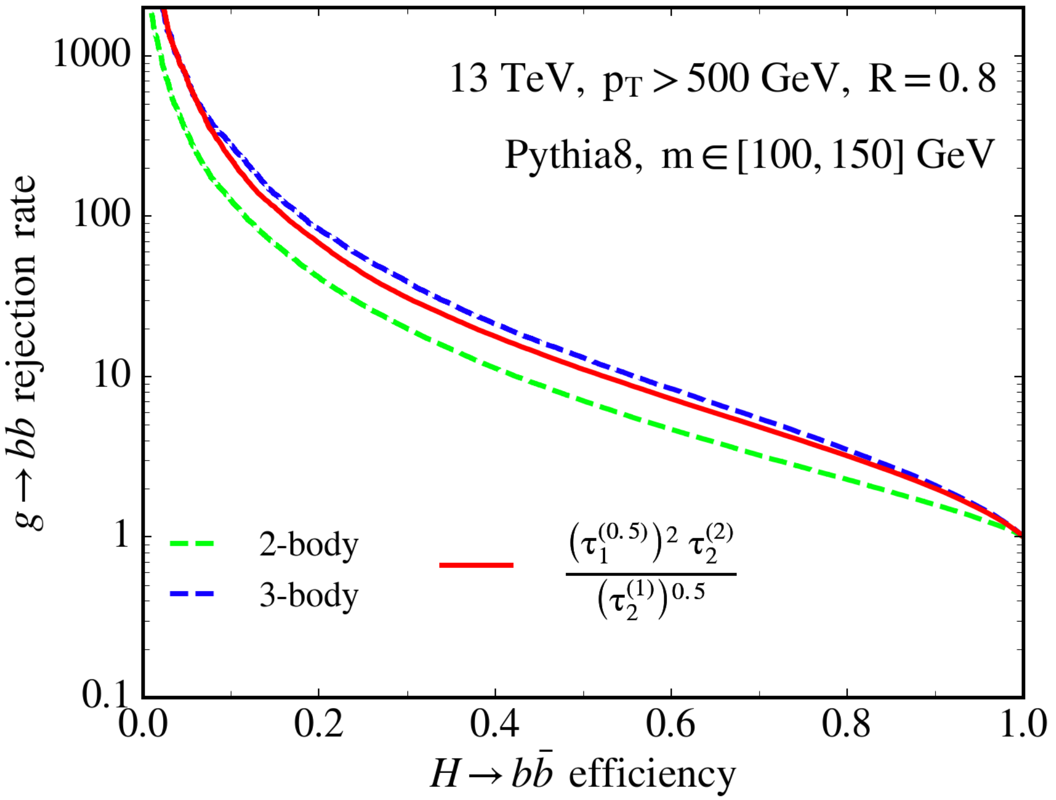
Comparison to standard observables



Performance at par with standard observables

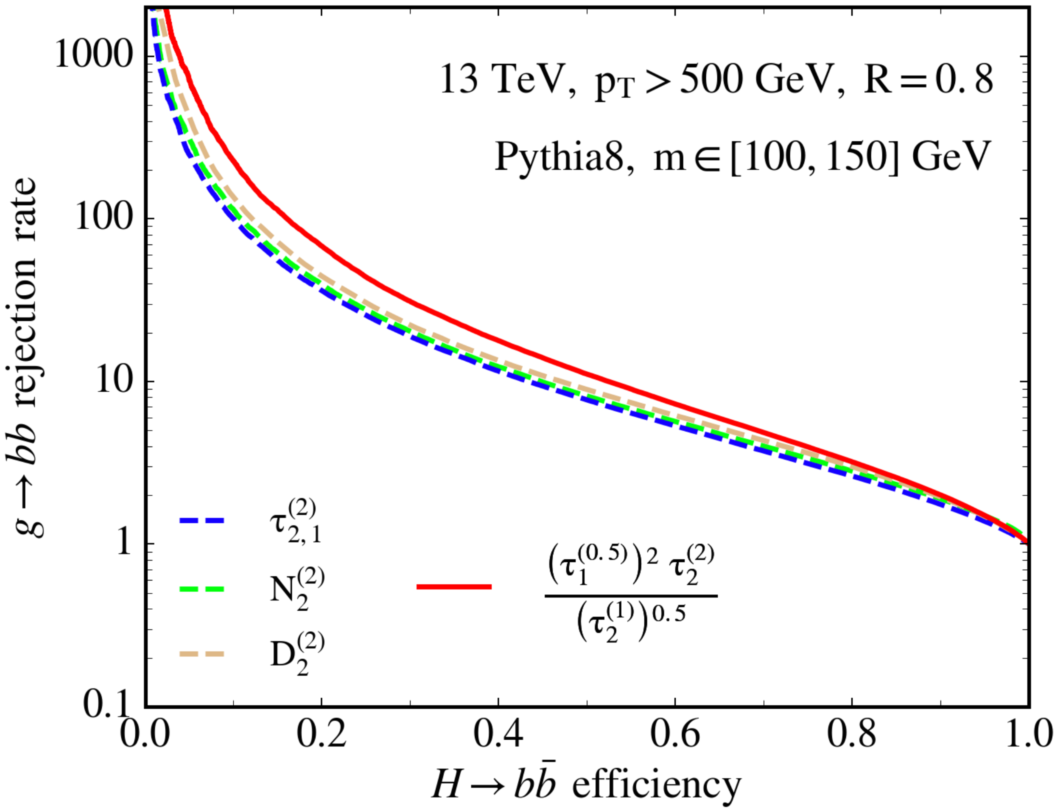
Example: Ungroomed $H \to b\bar{b}$ vs. $g \to b\bar{b}$ discrimination

Comparison to neural network



Recovers most of the 3-body phase space information identified by the neural network

Comparison to standard observables



<u>Outperforms standard observables</u>

BOOST 2018; Datta, Larkoski and Nachman

Conclusions and Caveats

Powerful new observables, with a compact, analytic form can be developed from information identified by machine learning classifiers

Procedure can, in principle, be automated

but

No guarantee the product form of the observable includes all discrimination power of resolved 3-body phase space
(but comparison to the NN demonstrates it captures most of this information)

Running a random scan over arbitrarily large M-body phase space is inefficient for finding the optimal product observable

Part 3a: Can we automate the development of new observables using machine learning?

↳ (Yes ☺)

Path to Automation

Measure parametrized product observables on samples of signal and background jets

Previous procedure is inefficient for problems that require random scans over more than a few parameters

Replace the Monte-Carlo based analysis segment with a neural network regression analysis and minimization routine

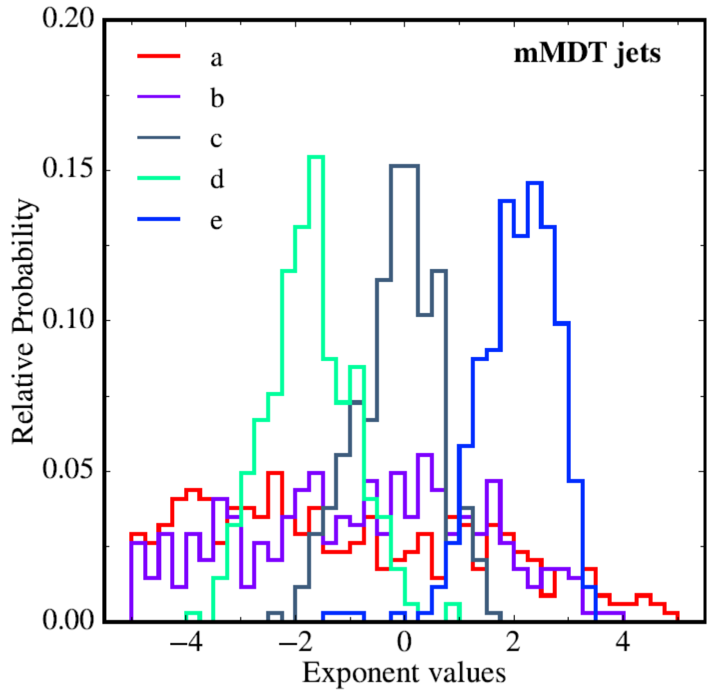
Neural network learns the mapping from the *M*-body (3M - 4 dim.) input space to the PDF of the product observable for signal and/or background

Network output is a measurement of the observables on 25k signal or background ‘events’

Groomed $H \to b\bar{b}$ vs. $g \to b\bar{b}$ discrimination

Carry out new ML procedure on a problem with known solution:

Expect a minima around [-2, 0, 0, -2, 2] for mMDT $H \to b\bar{b}$ case



Neural network architecture for regression (only fully-connected layers):

{Parameters, 1/0} -> 250 nodes -> 100 nodes -> 25,000 nodes
{a,b,c,d,e, sig./bkg. switch} LeakyReLU LeakyReLU Linear

Optimization over the input space

Calculate AUC from likelihood distributions of network output for signal and background

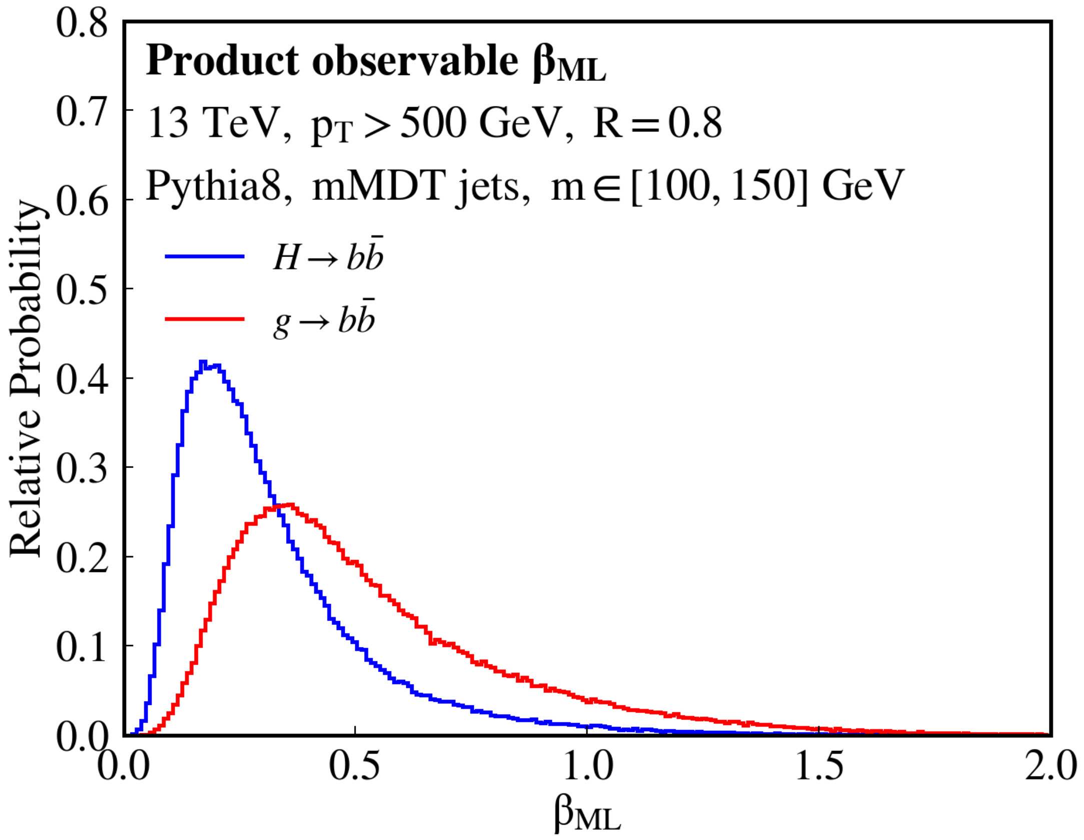
Minimize (1 – AUC) metric using scipy's basin hopping routine, with the COBYLA optimizer

Constrained Optimization by Linear Approximation

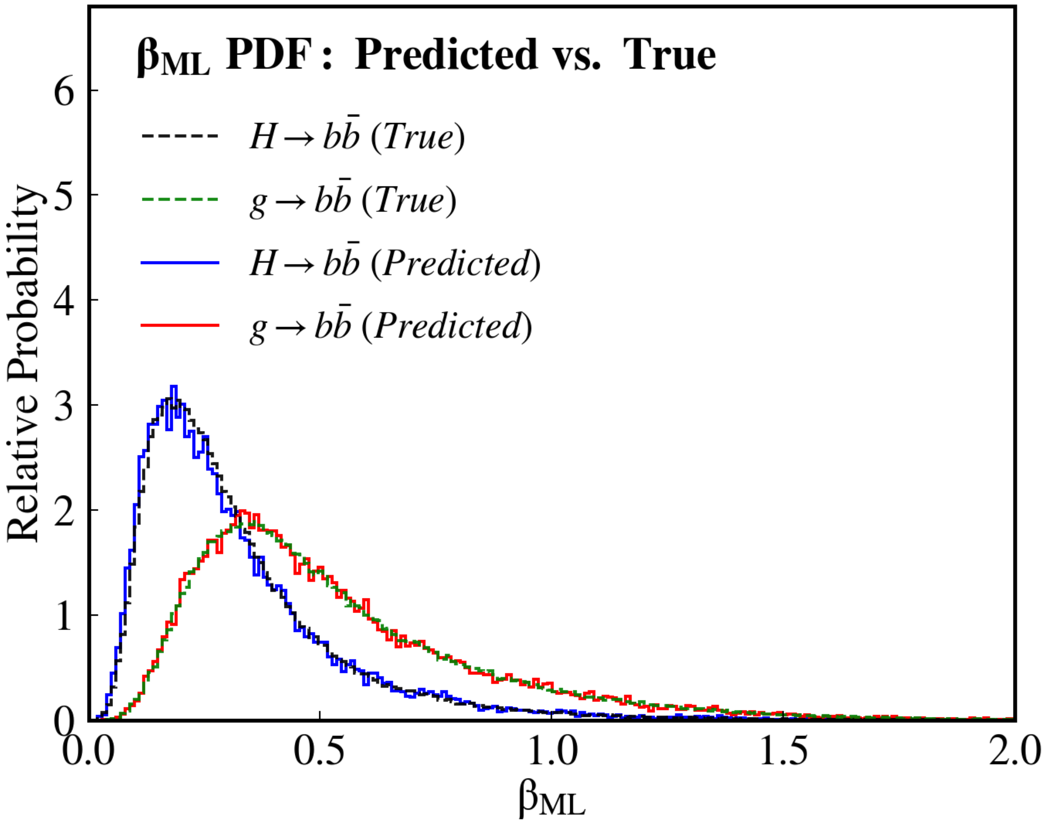
Results: Groomed $H \to b\bar{b}$ vs. $g \to b\bar{b}$ discrimination

Minima found at [-1.67, 0.2, -0.34, -1.84, 1.72] (close to the expected solution: [-2, 0, 0, -2, 2])

$$\beta_{ML} = \left(\tau_1^{(0.5)}\right)^{-1.67}\left(\tau_1^{(1)}\right)^{0.2}\left(\tau_1^{(2)}\right)^{-0.34}\left(\tau_2^{(1)}\right)^{-1.84}\left(\tau_2^{(2)}\right)^{1.72}$$



New observable measured on 250k signal and background jets

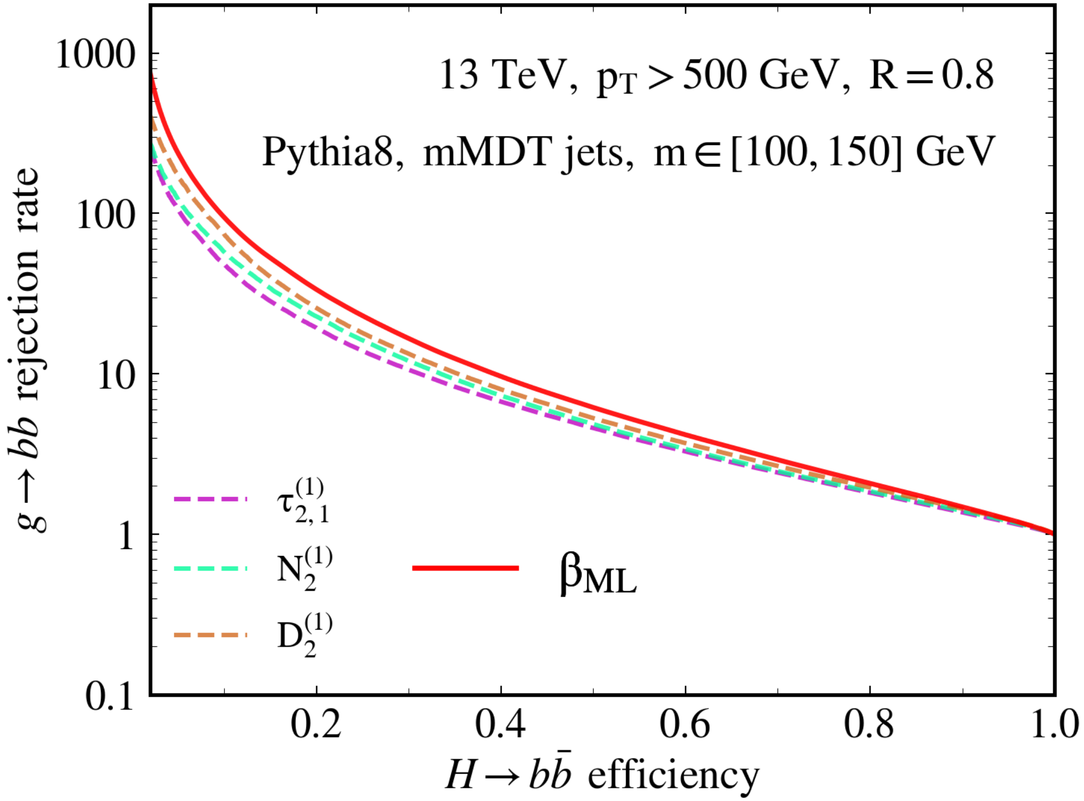


Comparison of NN output of 25k 'events' to measurement on larger statistics (250k) (normalized)

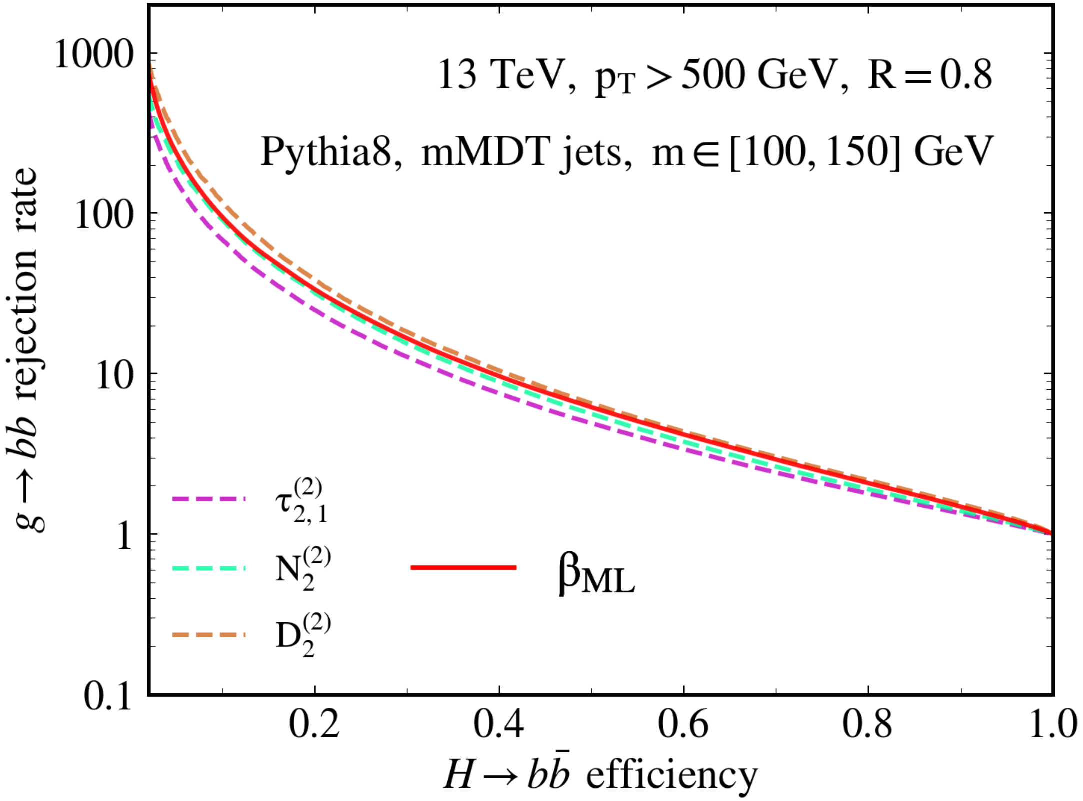
Results: Comparison to standard observables

Performs basically at par with standard observables

Results are as expected from the random scan procedure



Standard observables with angular exponent = 1



Standard observables with angular exponent = 2

Part 3b: Can this procedure be generalized to higher *M*-body phase space?

Z boson vs. QCD discrimination

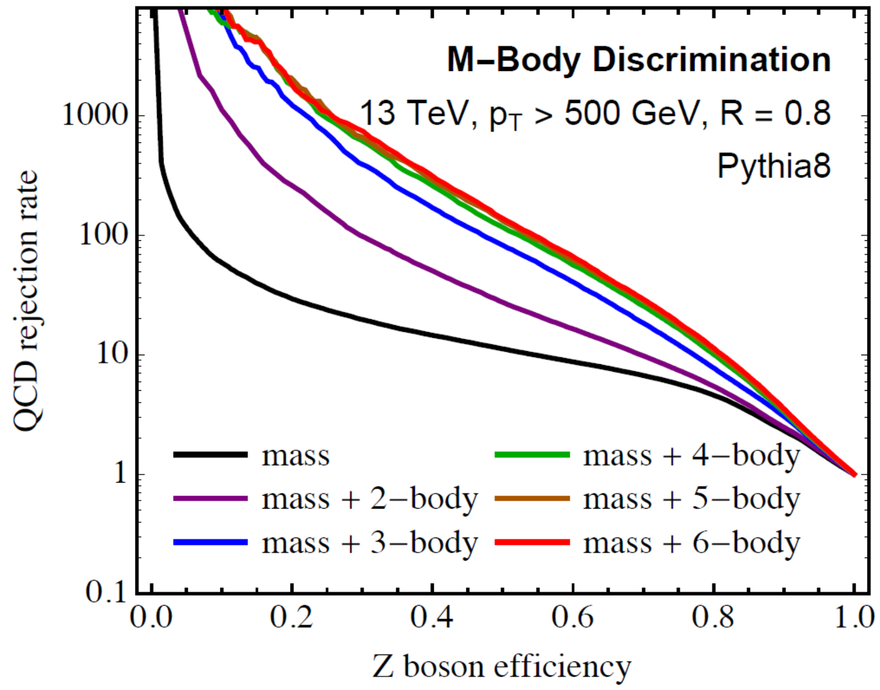
Apply same general ML and minimization procedure to Z vs QCD discrimination

Harder problem, since saturation observed at 4-body (8 dimensional) phase space

The 4-body product observable:

$$\beta_4 = \left(\tau_1^{(0.5)}\right)^a \left(\tau_1^{(1)}\right)^b \left(\tau_1^{(2)}\right)^c \left(\tau_2^{(0.5)}\right)^d \left(\tau_2^{(1)}\right)^e \left(\tau_2^{(2)}\right)^f \left(\tau_3^{(1)}\right)^g \left(\tau_3^{(2)}\right)^h$$

Solution not known a priori (we have to trust the machine!)



Z boson vs. QCD discrimination

Two separate neural networks trained for signal and background jets:

{4-body parameters} -> 500 nodes -> 250 nodes -> 100 nodes -> 25,000 nodes

{a, b, c, d, e, f, g, h}	LeakyReLU	LeakyReLU	LeakyReLU	Linear

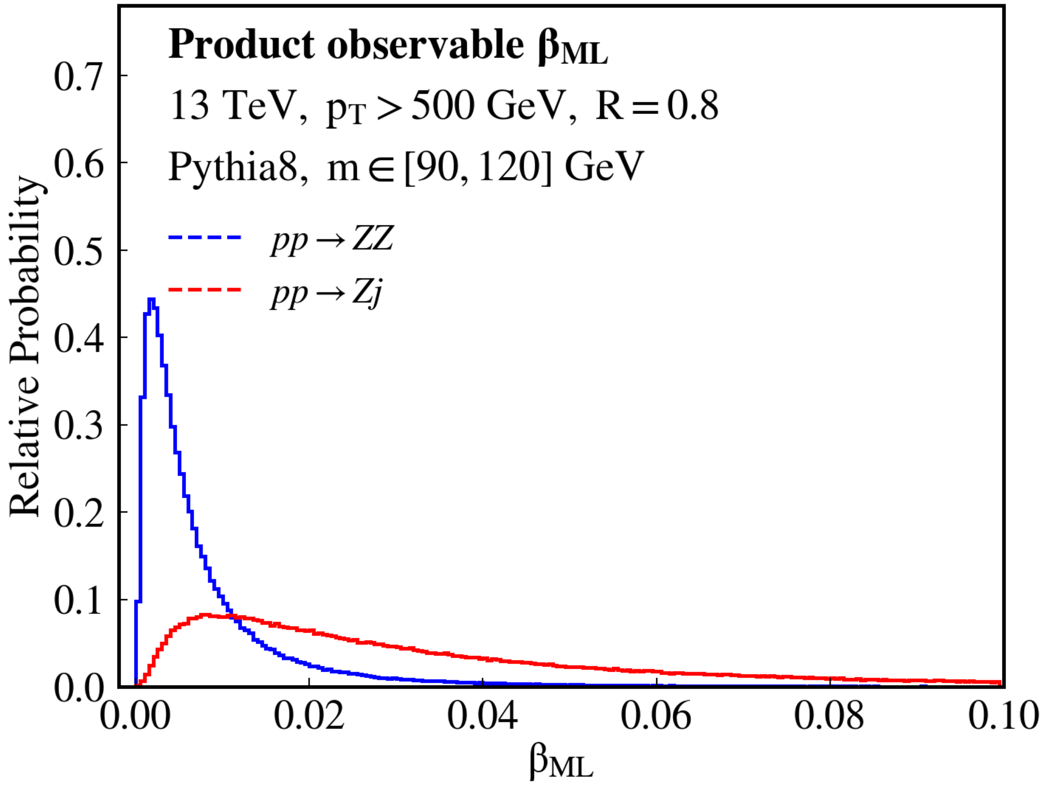
We find a minima at the following point in R^8

[a, b, c, d, e, f, g, h] = [1.02, -2.83, 0.36, -0.64, -0.57, 0.95, 1.87, 0.]

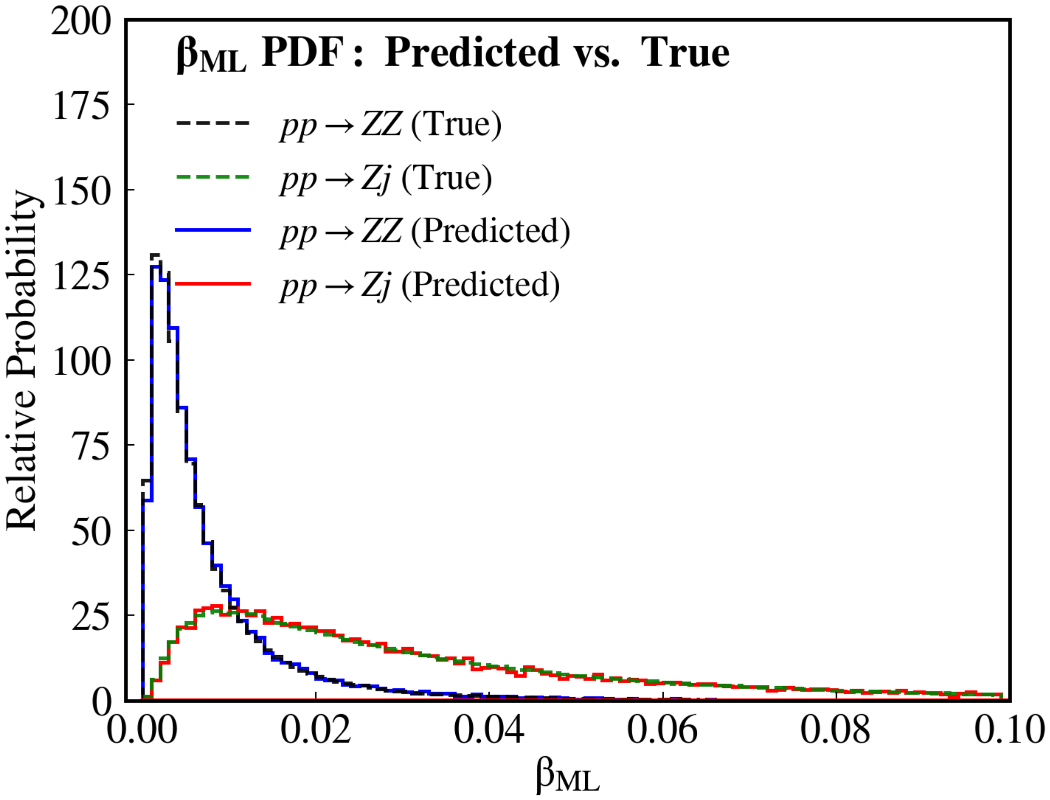
Results: *Z* boson vs. QCD discrimination

<u>New observable from machine learning that resolves most of 4-body phase space</u>

$$\beta_{ML} = \left(\tau_1^{(0.5)}\right)^{1.02}\left(\tau_1^{(1)}\right)^{-2.83}\left(\tau_1^{(2)}\right)^{0.36}\left(\tau_2^{(0.5)}\right)^{-0.64}\left(\tau_2^{(1)}\right)^{-0.57}\left(\tau_2^{(2)}\right)^{0.95}\left(\tau_3^{(1)}\right)^{1.87}\left(\tau_3^{(2)}\right)^{0}$$



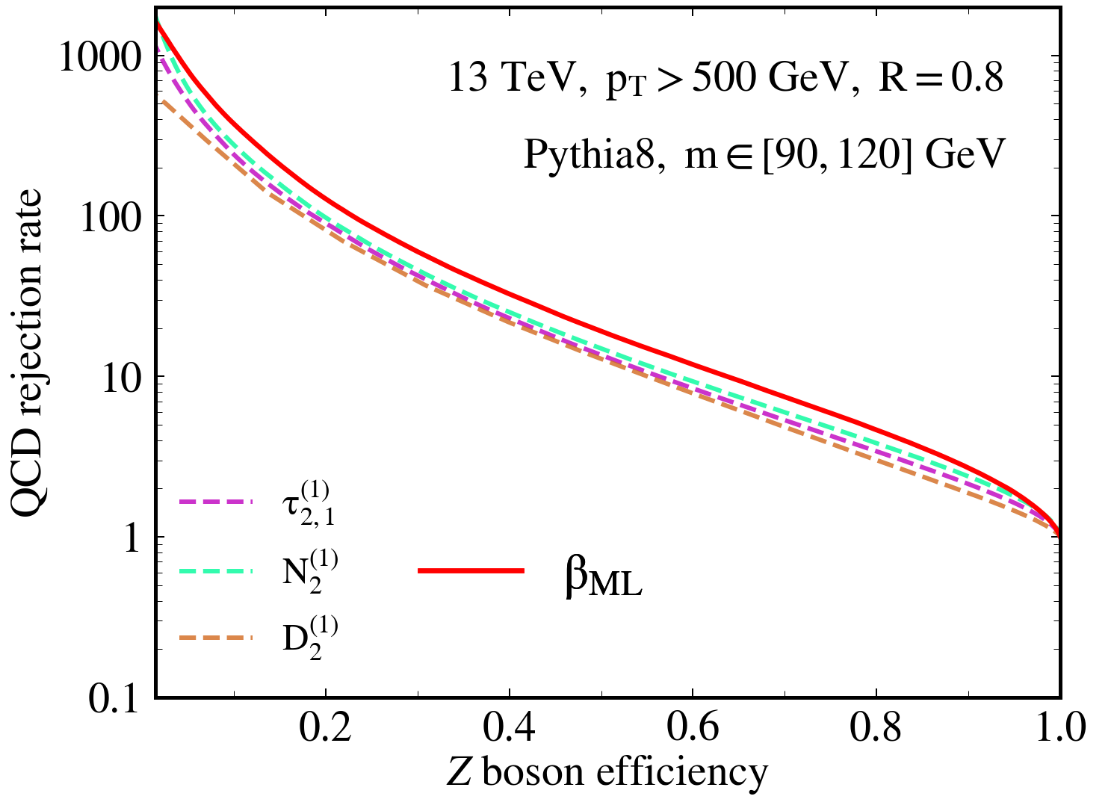
New observable measured on 500k signal and background jets



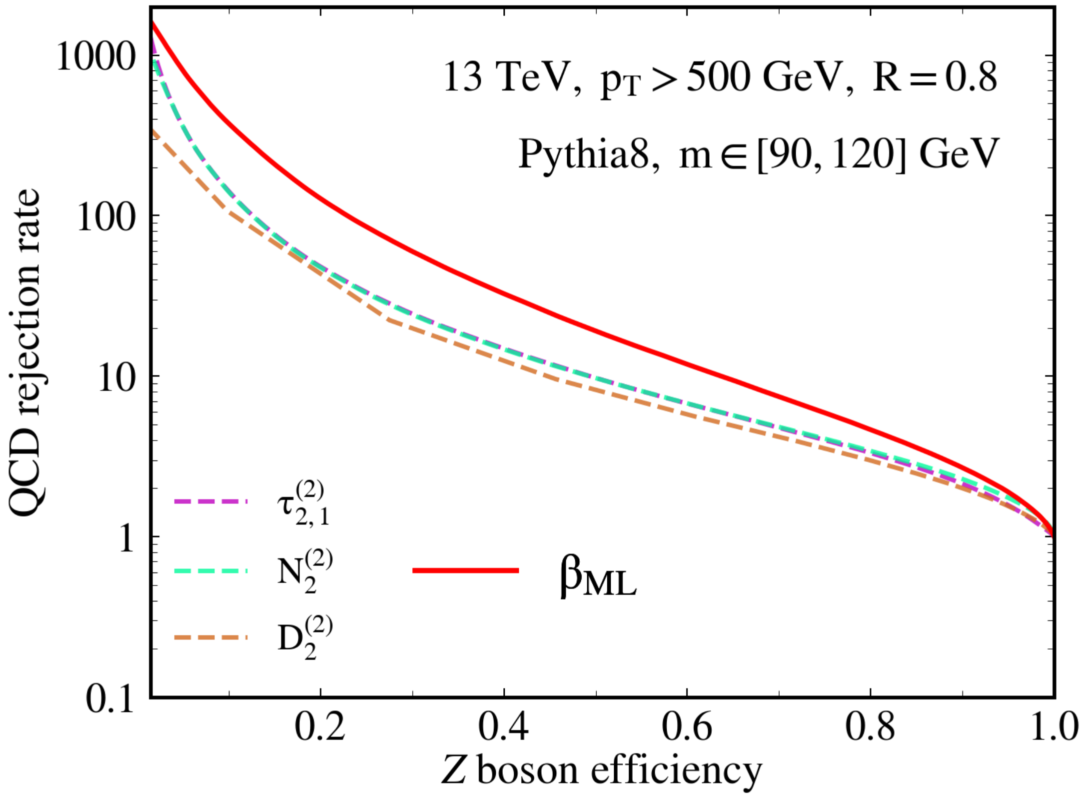
Comparison of NN output of 25k 'events' to measurement on larger statistics (500k)

Results: Comparison to standard observables

<u>Outperforms standard observables for *Z* vs. QCD discrimination</u>



Standard observables with angular exponent = 1



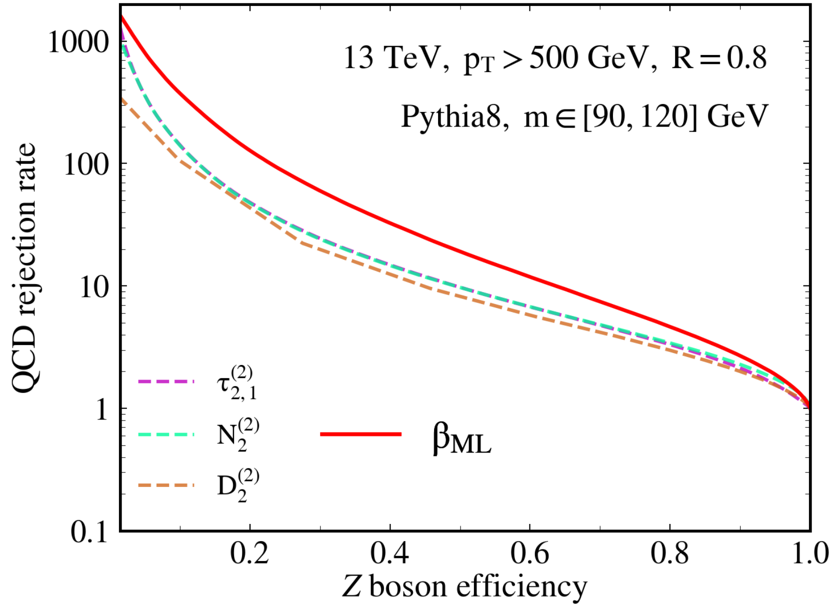
Standard observables with angular exponent = 2

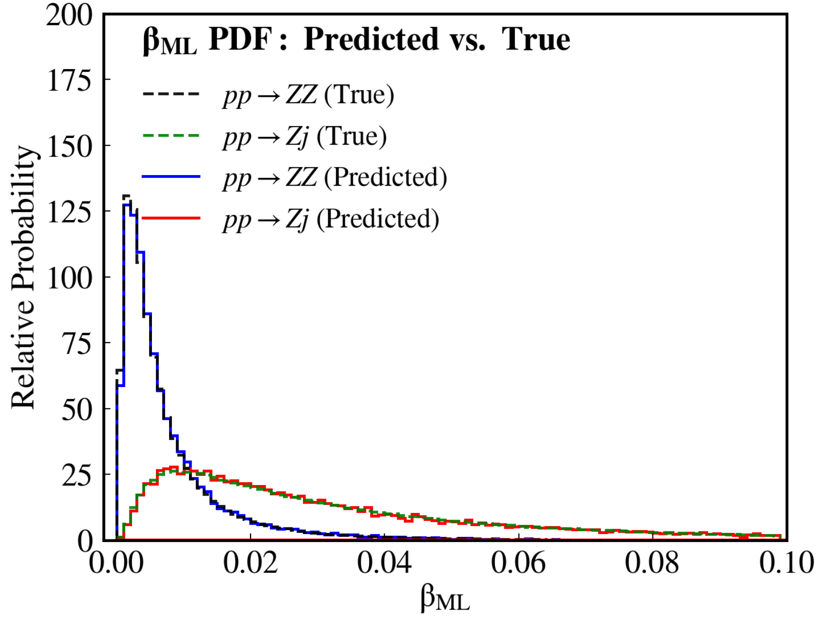
Conclusions

Powerful new observables for jet discrimination can be developed essentially automatically once point of saturation of discrimination power is determined

A neural network can reliably learn how the PDFs of products/ratios of N-subjettiness observables change with the exponents – we can exploit this to automate the optimality scan

This procedure can be efficiently extended to arbitrarily higher *M*-body phase space and can thus be implemented generically to develop optimal discriminants for any jet classification problem

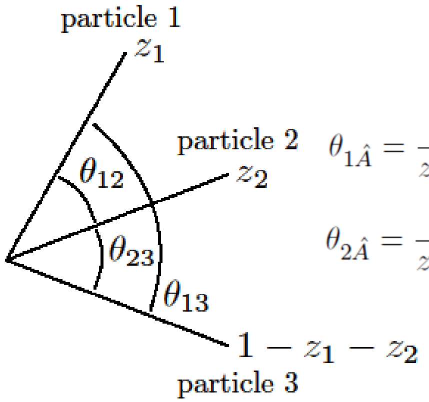




Thank you!

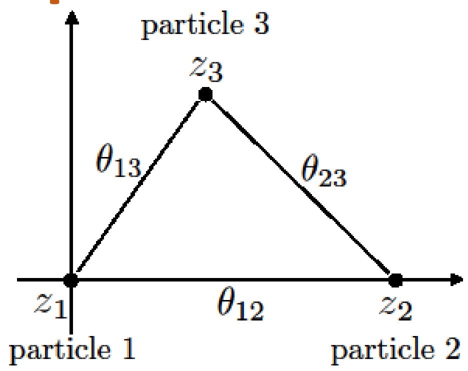
Backups

Example: 3-body phase space



$$\theta_{1\hat{A}} = \frac{z_2}{z_1+z_2}\theta_{12} \qquad \tau_2^{(1)} = z_1 \cdot \frac{z_2}{z_1+z_2}\theta_{12} + z_2 \cdot \frac{z_1}{z_1+z_2}\theta_{12} = \frac{2z_1z_2}{z_1+z_2}\theta_{12}\,,$$

$$\theta_{2\hat{A}} = \frac{z_1}{z_1+z_2}\theta_{12} \qquad \tau_2^{(2)} = z_1\left(\frac{z_2}{z_1+z_2}\theta_{12}\right)^2 + z_2 \cdot \left(\frac{z_1}{z_1+z_2}\theta_{12}\right)^2 = \frac{z_1z_2}{z_1+z_2}\theta_{12}^2$$



3-body phase space is 3*3 - 4 = 5 dimensional

We measure 4 N-subjettiness observables

$$\{\tau_1^{(0.5)}, \tau_1^{(1)}, \tau_1^{(2)}, \tau_2^{(1)}, \tau_2^{(2)}\}$$

One of 2-subjettiness axes along direction of a particle

⇓

ie – measuring 2-subjettiness sensitive to one relative energy fraction and one pairwise angle

Axis for 1-subjettiness observables displaced from direction of particles in jet (E-scheme recomb. instead of WTA)

E-scheme recomb. conserves momentum and so axis is degenerate to dir. of a particle if another particle has 0 energy or is exactly collinear to another

$$\text{particle 3:} \quad \left(\frac{\theta_{12}^2+\theta_{13}^2-\theta_{23}^3}{2\theta_{12}}, \frac{\sqrt{2\theta_{12}^2\theta_{23}^2+2\theta_{12}^2\theta_{13}^2+2\theta_{13}^2\theta_{23}^2-\theta_{12}^4-\theta_{23}^4-\theta_{13}^4}}{2\theta_{12}}\right)$$

$$\text{jet center:} \quad \left(z_2\theta_{12}+z_3\frac{\theta_{12}^2+\theta_{13}^2-\theta_{23}^3}{2\theta_{12}}, z_3\frac{\sqrt{2\theta_{12}^2\theta_{23}^2+2\theta_{12}^2\theta_{13}^2+2\theta_{13}^2\theta_{23}^2-\theta_{12}^4-\theta_{23}^4-\theta_{13}^4}}{2\theta_{12}}\right)$$

$$\begin{aligned}
\theta_{1\hat{A}}^2 &= z_2^2\theta_{12}^2 + z_3^2\theta_{13}^2 + z_2z_3(\theta_{12}^2+\theta_{13}^2-\theta_{23}^2)\,,\\
\theta_{2\hat{A}}^2 &= z_1^2\theta_{12}^2 + z_3^2\theta_{23}^2 + z_1z_3(\theta_{12}^2+\theta_{23}^2-\theta_{13}^2)\,,\\
\theta_{3\hat{A}}^2 &= z_1^2\theta_{13}^2 + z_2^2\theta_{23}^2 + z_1z_2(\theta_{13}^2+\theta_{23}^2-\theta_{12}^2)\,.
\end{aligned}$$

$$\begin{aligned}
\tau_1^{(0.5)} &= z_1\theta_{1\hat{A}}^{0.5} + z_2\theta_{2\hat{A}}^{0.5} + z_3\theta_{3\hat{A}}^{0.5}\,,\\
\tau_1^{(1)} &= z_1\theta_{1\hat{A}} + z_2\theta_{2\hat{A}} + z_3\theta_{3\hat{A}}\,,\\
\tau_1^{(2)} &= z_1\theta_{1\hat{A}}^2 + z_2\theta_{2\hat{A}}^2 + z_3\theta_{3\hat{A}}^2 = z_1z_2\theta_{12}^2 + z_1z_3\theta_{13}^2 + z_2z_3\theta_{23}^2
\end{aligned}$$

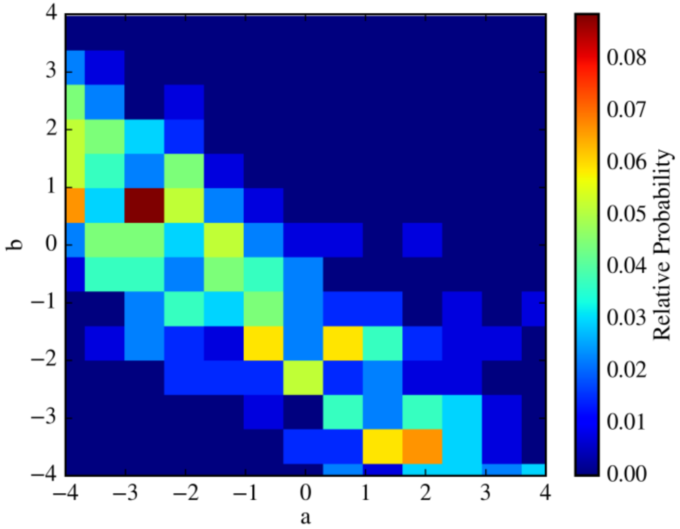
Proof of this (3M-4) is an application of the Euler Characteristic formula:

V − E + F = 2

Number of vertices V is number of particles in the jet, M. Number of faces F is number of triangles that tesselate the plane, with vertices located at the particles, ie F=M − 1, as we include the face outside the region where the points are located. It then follows that the number of edges E, that is, the number of pairwise angles necessary to uniquely specify their distribution, is E = 2M − 3.

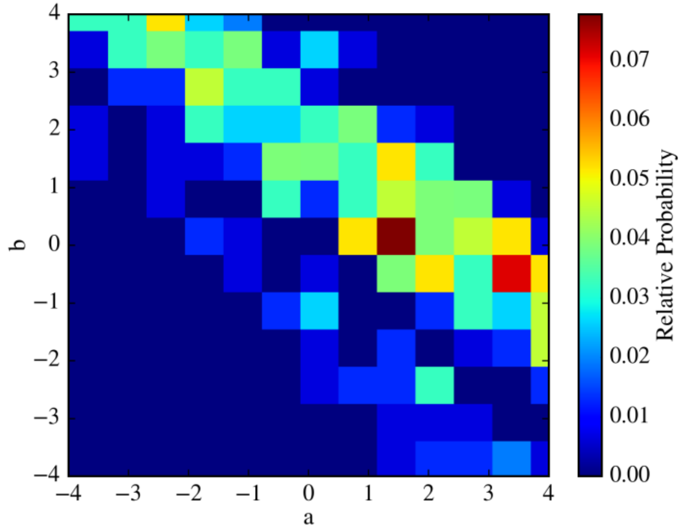
Correlation of exponents from H -> bb random scan

a+b = -2



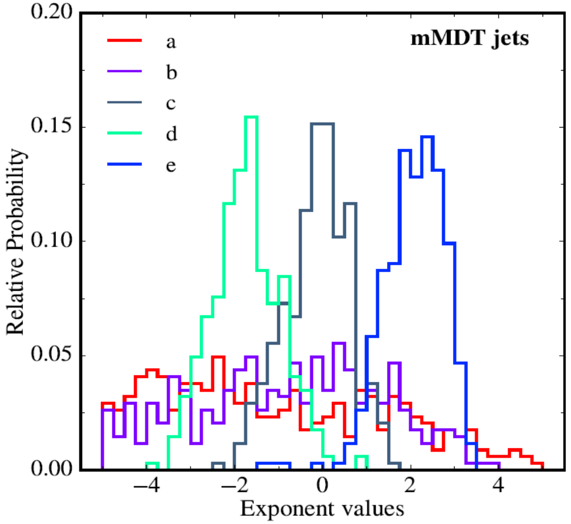
mMDT groomed

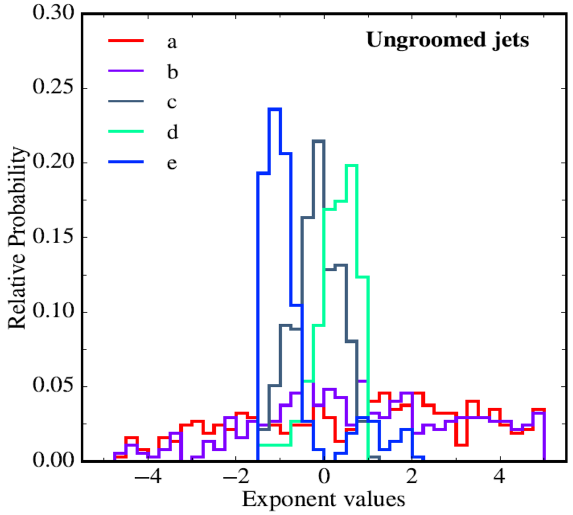
a+b = 2



Ungroomed

$$\beta_3^{(g)} = \left(\frac{\tau_2^{(2)}}{\tau_1^{(0.5)}\tau_2^{(1)}}\right)^2$$

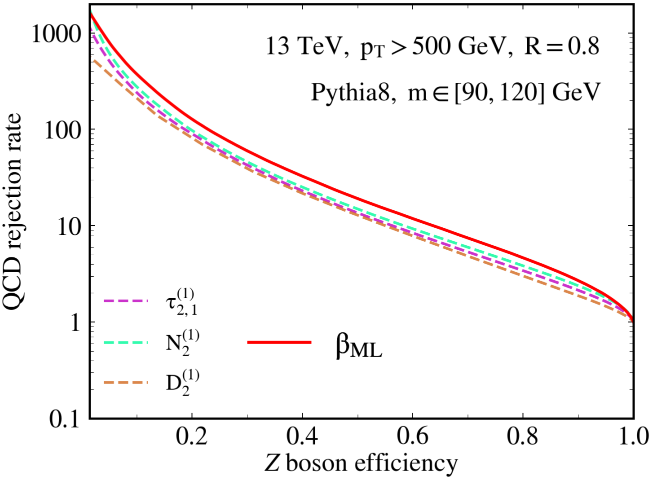


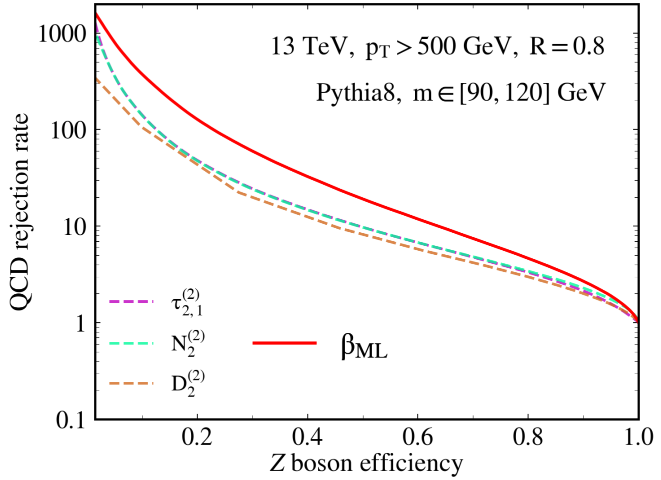


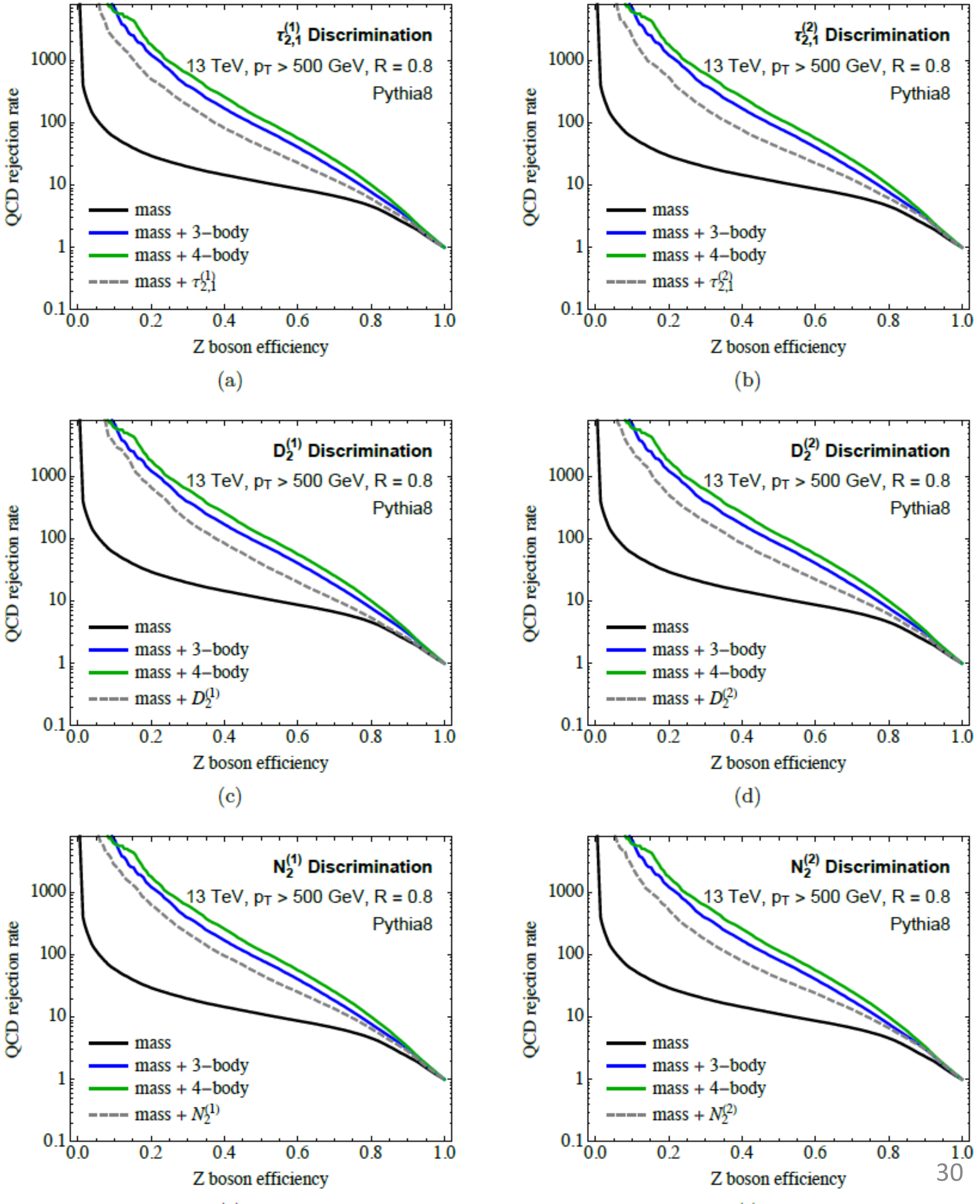
$$\beta_3 = \frac{\left(\tau_1^{(0.5)}\right)^2 \left(\tau_2^{(1)}\right)^{0.5}}{\tau_2^{(2)}}$$

More about the new observables: Z vs. QCD

New observable identifies information in 4-body phase space missed by the standard observables





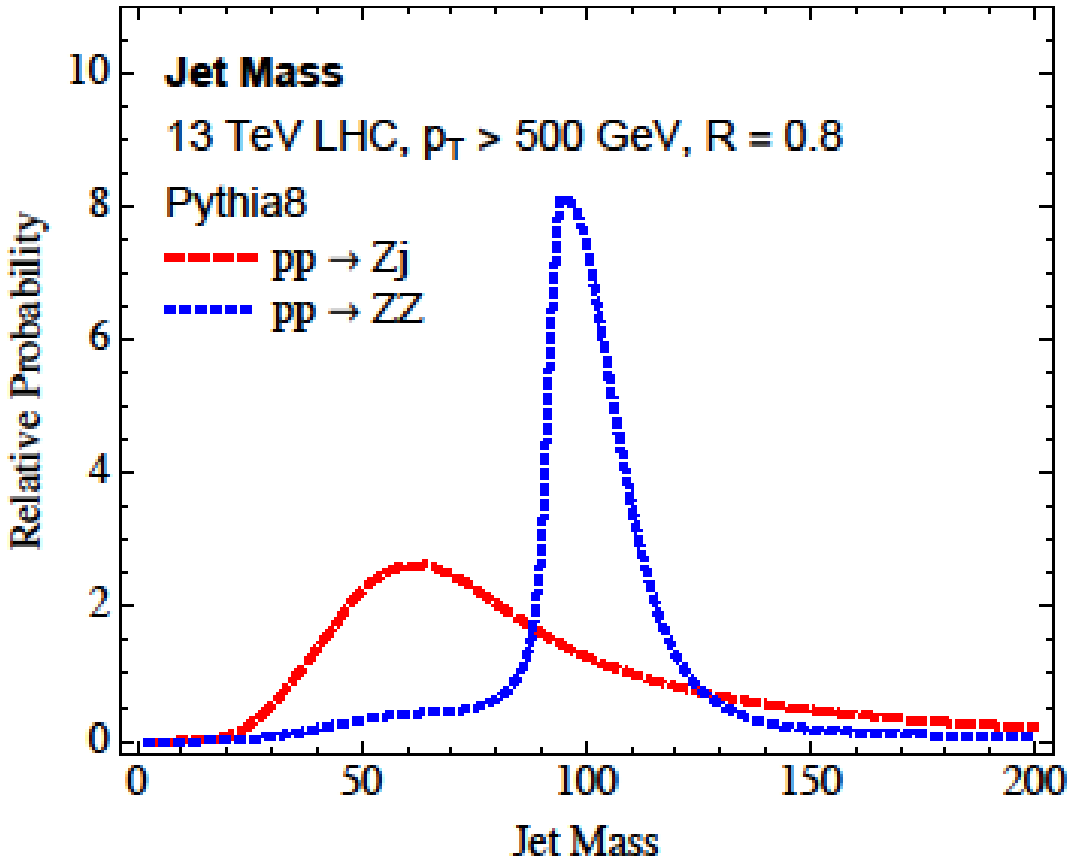


BOOST 2018; Datta, Larkoski and Nachman

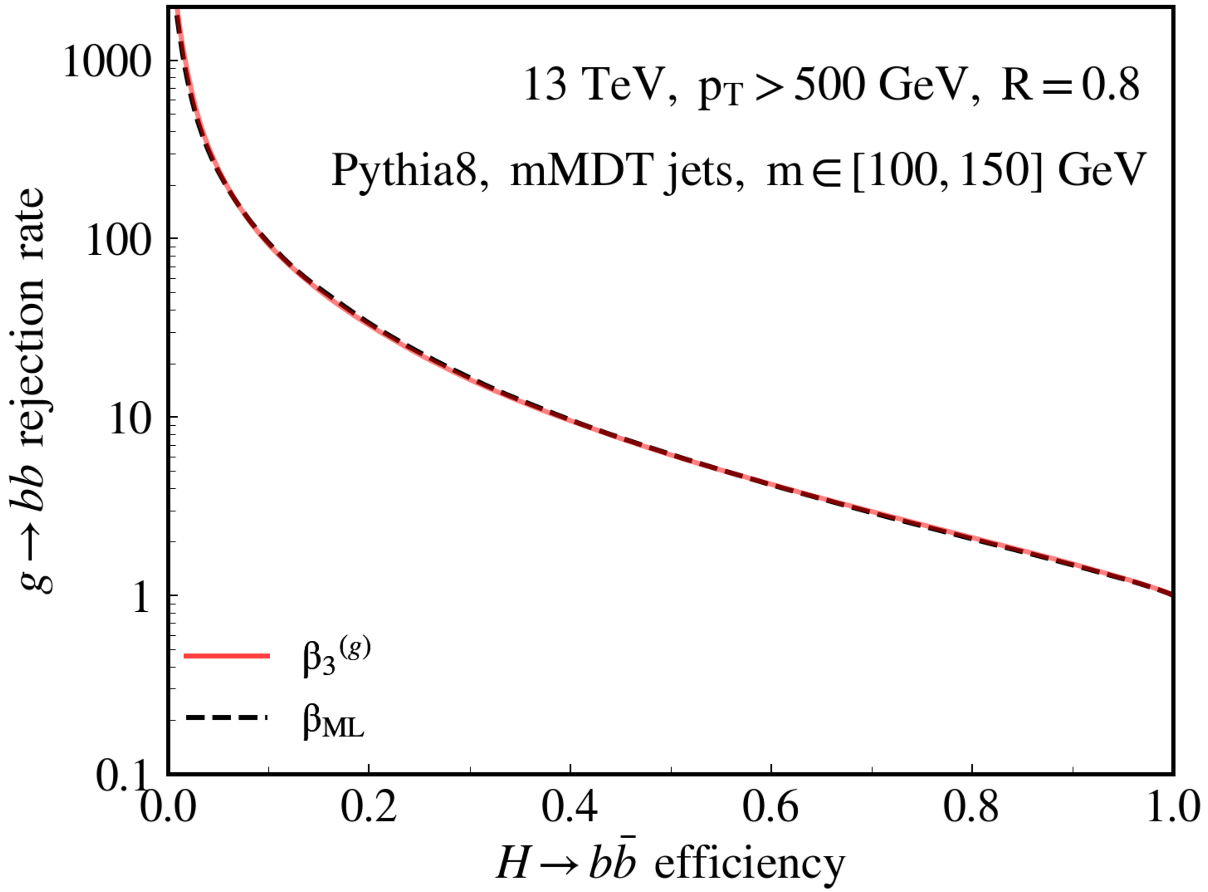
Jet Mass: *Z* vs. QCD

Why we use a [90,120] GeV mass cut for the *Z* vs. QCD case:

Signal peaks around 105 GeV



More about the new observables: $H \to b\bar{b}$ vs. $g \to b\bar{b}$



$$\beta_3^{(g)} = \frac{\tau_2^{(2)}}{\tau_1^{(0.5)}\tau_2^{(1)}}$$

$$\beta_{ML} = \left(\tau_1^{(0.5)}\right)^{-1.67}\left(\tau_1^{(1)}\right)^{0.2}\left(\tau_1^{(2)}\right)^{-0.34}\left(\tau_2^{(1)}\right)^{-1.84}\left(\tau_2^{(2)}\right)^{1.72}$$

Same performance of ML and random scan product observable as expected

Minimization with Basin Hopping and COBYLA

Basin hopping

Global minimum finder:
Two-phase method that combines a global stepping algorithm with local minimization at each step.

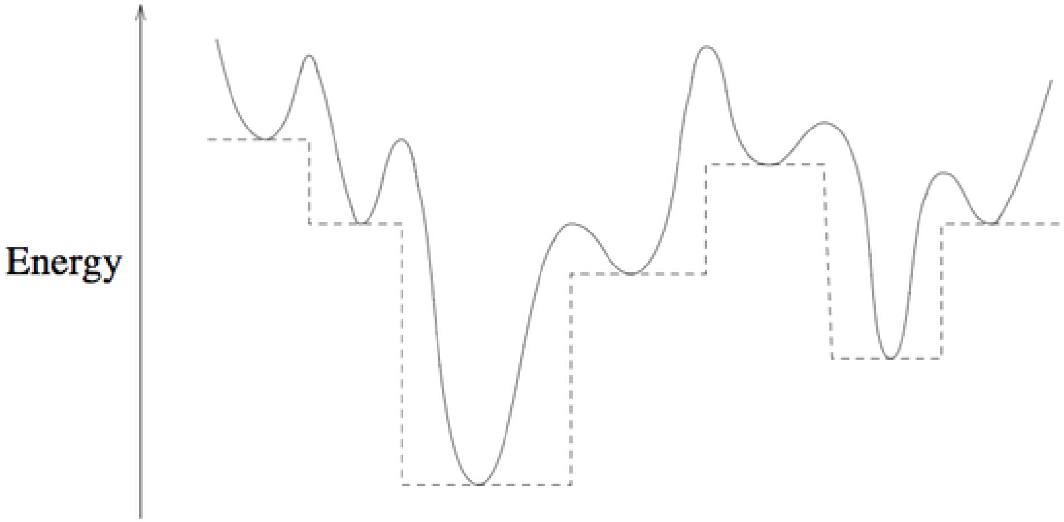


FIG. 2. A schematic diagram illustrating the effects of our energy transformation for a one-dimensional example. The solid line is the energy of the original surface and the dashed line is the transformed energy, $\tilde{E}$.

arXiv:cond-mat/9803344

COBYLA

Constrained optimization by linear approximation:
Numerical optimization method where derivative of function is not known

Powell, M.J.D., "A view of algorithms for optimization without derivatives"

Where can we improve?

Most improvement would be accomplished by construction of an optimal basis of functions with parameters that can be tuned to maximize discrimination power