Divorcing Soft Substructure from Hard Kinematics

BOOST 2013 Workshop
Parallel Talk

August 14 2013

David Curtin
Yang Institute for Theoretical Physics
Stony Brook

Based on
WORK IN PROGRESS (DC, Rouven Essig, Brian Shuve)

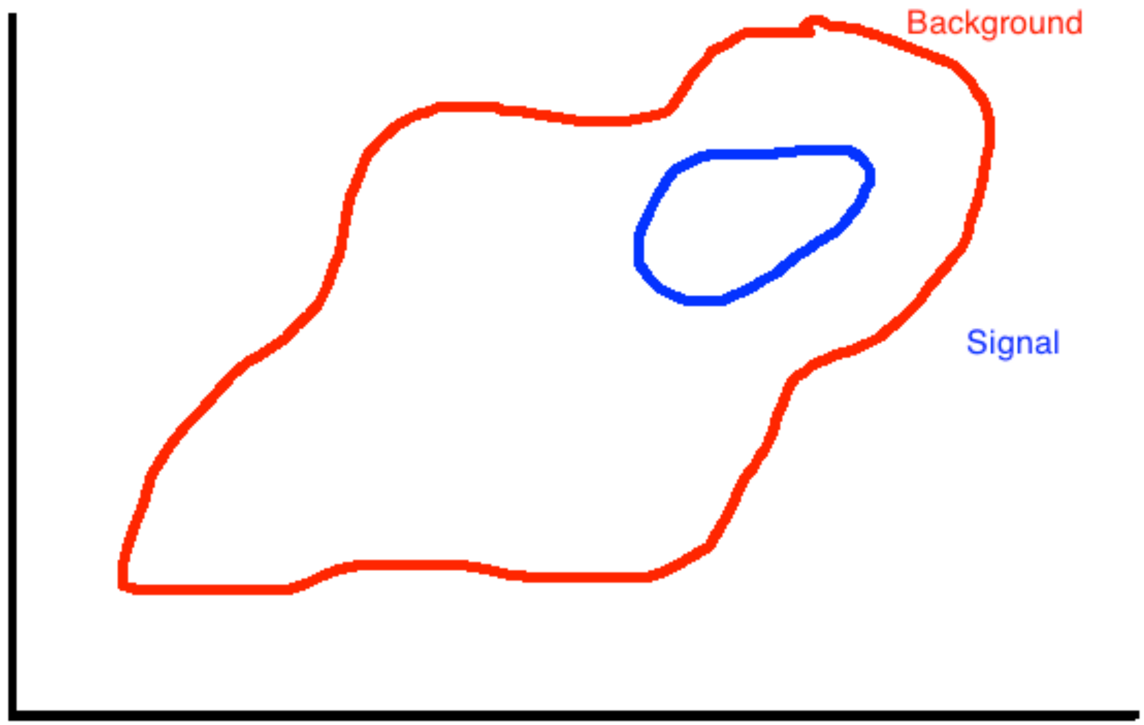
Motivation

- Huge number of Jet Substructure techniques/variables/search strategies
- While new techniques can always be useful, there is a need to go back to basics (common theme of this workshop):
  - Theoretical Calculations:
    Understand why techniques that were algorithmically devised work the way they do. (c.f. talks by Thaler, Salam, Marzani, ...). Improves robustness & reliability.
  - Different approach to MC Simulations:
    Rather than theorists doing a straightforward Signal vs Background MC study which includes their favorite substructure technique, maybe we should make things 'a bit harder' for ourselves (beyond checking correlation plots).

    We should quantify how much `new information' is made available by adopting new techniques (which carry a high 'deployment cost' for experimentalists) which is **not available by other means**.

Experimental Point of View

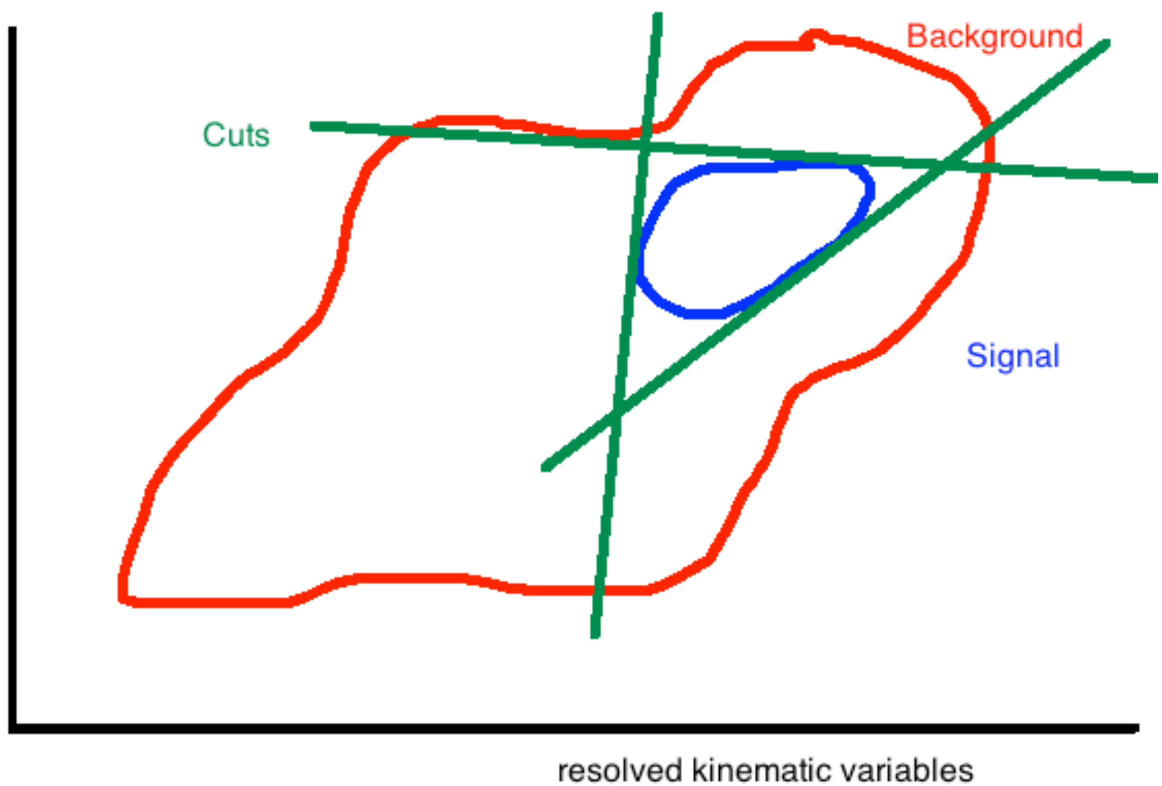
In a given search, signal and background have some weight distribution in the space of 'well understood' kinematic variables.



Experimental Point of View

Perform cuts in these variables to define a series of search regions that are optimized for Signal.

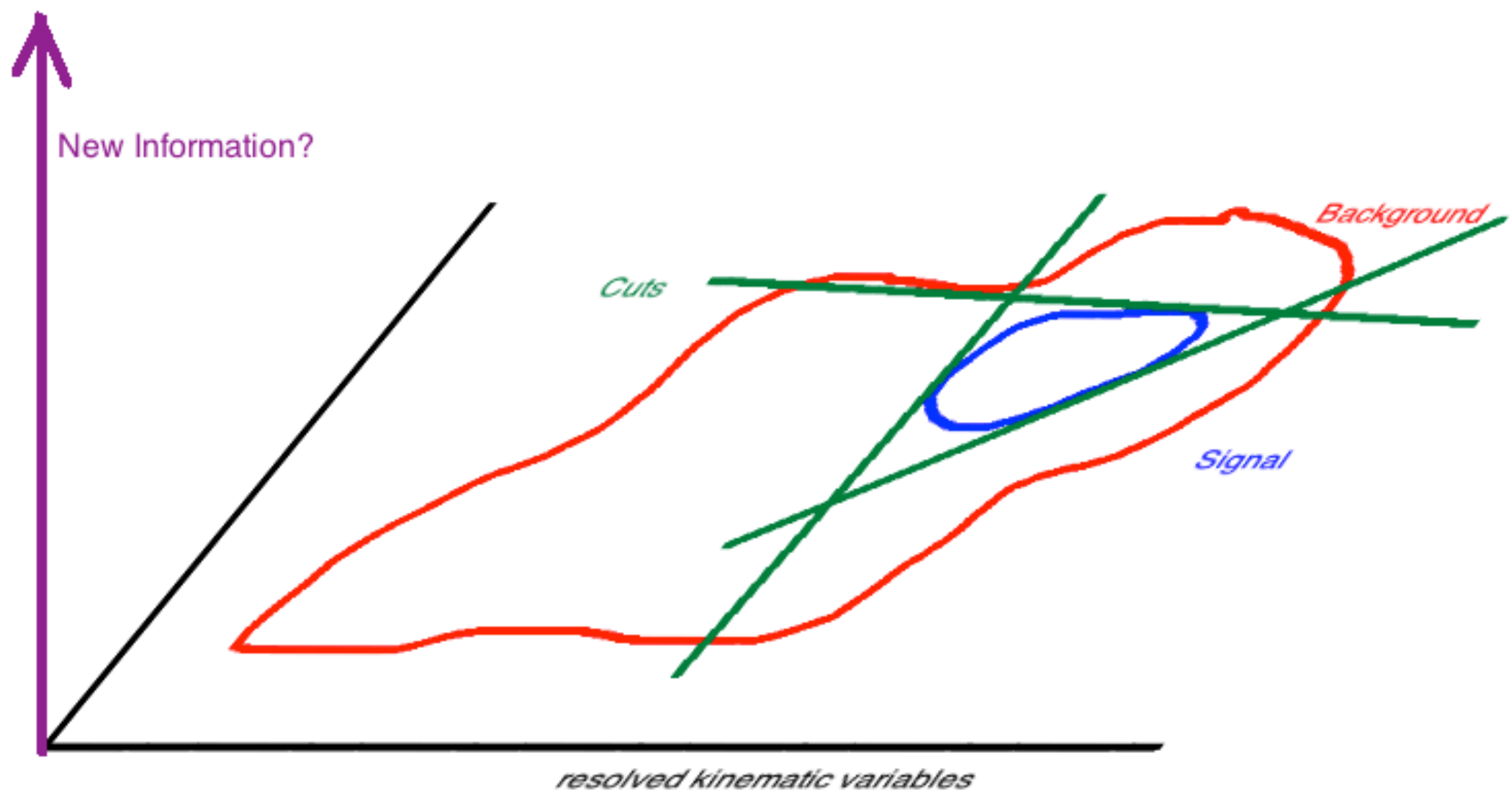
If this is done 'perfectly', the S and B distributions should be 'identical' in each search region. (Otherwise you can exploit kinematic differences to cut better, at least ideally.)



Experimental Point of View

Any new substructure techniques we as theorists devise should improve a search by making completely new information accessible.

How can we make sure this is the case?



Determining New Information Content

- Substructure variables can be divided into those which are designed to have strong dependence on 'hard matrix element kinematics' ....
    - various h/W/top taggers distinguish hard EW decays from QCD splittings
    - N-subjettiness (though it also retains sensitivity to subjet shapes)
    - jet pruning/filtering algorithms are essentially 'selectively destructive mass filters' which are designed to leave jet mass from hard EW decays untouched while severely reducing mass due to QCD splittings.
- .. and those that do not, and are *primarily* designed to probe the parton shower and parton content of a jet (for lack of a better term 'soft substructure')
    - color flow
    - quark/gluon tagging
    - various jet moments and shape variables

Determining New Information Content

- Substructure variables can be divided into those which are designed to have strong dependence on 'hard matrix element kinematics' ....
  - various h/W/top taggers distinguish hard EW decays from QCD spl
  - N- apes)
  - jet tructive ma EW decays untouched while severely reducing mass due to QCD splittings.

Reliable & Well Established

- .. and those that do not, and are *primarily* designed to probe the parton shower and parton content
  - color flow
  - quark/gluon tagging
  - various jet moments and shape variables

⋆ **Many** different variables proposed
⋆ Varying degrees of correlation with kinematic variables
⋆ Systematics? MC dependence?
⋆ PU & grooming dependence?

Determining New Information Content

- These 'soft substructure' or 'radiation'-variables are challenging experimentally, and their usefulness must be clearly understood to motivate their practical deployment.

- q/g tagging is already being developed by experimental collaborations, but event topology is likely to be a crucial input into the discrimination.

  This likely applies to other radiation variables as well. e.g. CMS 1202.1416

➡ Which other variables (if any) should experimentalists focus on?

Determining New Information Content

Color Flow Example

DC, Rouven Essig,
Brian Shuve 1210.5523

~200 GeV R-hadron **pairs** decaying
into **3 jets each**,

color flow can improve discovery significance by

$\varepsilon_{sig}/\sqrt{\varepsilon_{background}} \sim (12\%)\sqrt{(0.2\%)} \sim 2.7$

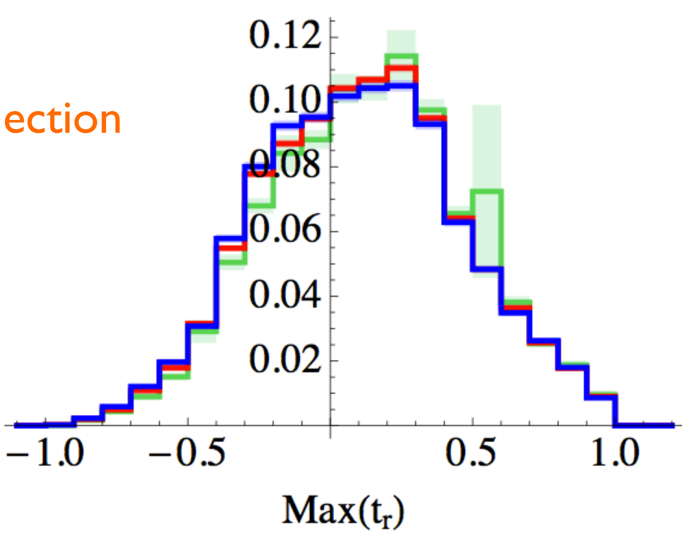
Aside: MC Comparison for RPV Gluino Study

Compared QCD background distributions in Sherpa to POWHeg + Pythia6.4, POWHeg + Pythia8.

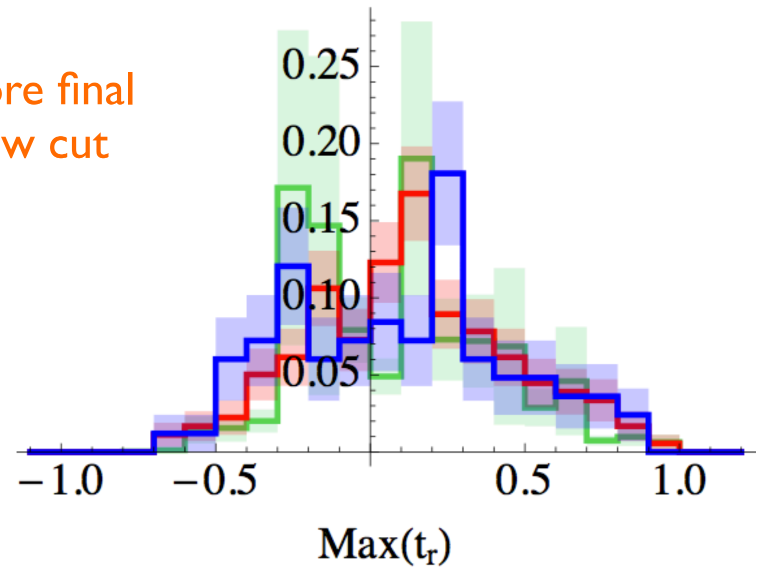
Shape agreement is generally good, but some important deviations in cut efficiencies & tails of distributions for color flow.

However, the *usefulness* of the color flow cut seems robust!

After preselection



Just before final color-flow cut



DC, Rouven Essig, Brian Shuve 1210.5523

Determining New Information Content

Color Flow Example

DC, Rouven Essig, Brian Shuve 1210.5523

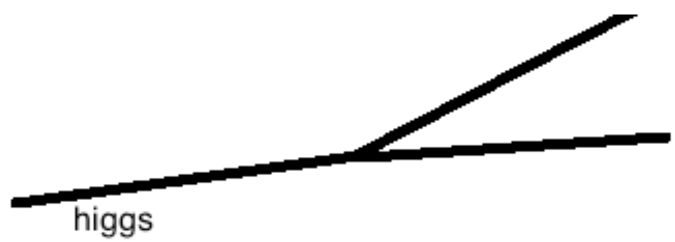
~200 GeV R-hadron **pairs** decaying into **3 jets each**,

color flow can improve discovery significance by

$$\boldsymbol{\varepsilon_{sig}/\sqrt{\varepsilon_{background}} \sim (12\%)/\sqrt{(0.2\%)} \sim 2.7}$$

Does not work for every topology

single 125 GeV higgs decaying into 2 gluons



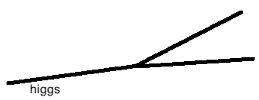
color flow doesn't help to improve significance at all, *after you work hard to optimize conventional kinematic cuts.*

What unique information is in radiative variables?

- It makes sense to organize the study of soft substructure variables by event topology: **how many singlets in event, and how many jets each do they decay to?**

more general — specialized

not sure if soft substructure helps



1 singlet → 2j

soft substructure likely very helpful

2 singlets → 2 x 3j

- Would like to narrow down the list of useful variables from the large number of possibilities. [Makes experimental adoption more likely.]

- **To make the most general statements, we have to avoid any accidental correlations with hard kinematic observables.**

Start with the most common one: 1 color singlet → 2 jets

Partially been studied before, e.g. Gallicchio, Huth, Kagan, Schwartz, Black, Tweedie '10
but not in boosted regime, and no studies have focussed on a complete lack of kinematic correlation.
(also lots of new variables to weed through now....)

Divorcing Kinematics from Soft Substructure

- Fairly standard sample selection for LHC14 study

  $W(h \to qq)$ & $W(h \to gg)$ vs Wqq, Wqg, Wgg

  with **leptonic W** to trigger on and moderately boosted higgs with $p_T > 200$ GeV. *(High boost sample later.)*

- Generate samples in Madgraph with very stringent generator-level cuts:
  $p_T^W > 200$ GeV $\Delta R_{jj} \lesssim 1.8$ $m_{jj} = m_H \pm 2$ GeV $p_{Tj} > 40$ GeV
  and apply further cuts after showering with Pythia8 and clustering with antikT(0.4):
  W-selection cuts, $\Delta R_{jj} < 1.2$, $p_{Tjj} > 200$ GeV, $p_{Tj} > 50$ GeV, $m_{jj} \sim m_H$

  ... but maybe we missed some clever kinematic correlation?

  How to ensure our results will be completely kinematics-independent?

Divorcing Kinematics from Soft Substructure

Modify parton-level events to have identical kinematics!

This amounts to 'flattening' the differences between Signal and Background kinematic distributions to emulate the 'perfectly optimized' experimental study that cannot be improved upon with any other techniques.

Completely artificial, but ensures that any new information is 'genuine' and that the relevant techniques can be 'pasted' on top of any experimental analysis.

Any discovered gains in $S/\sqrt{B}$ should be strictly cumulative.

Also allows us to study correlations purely between soft substructure variables, to separate correlation WITH kinematics from correlations due to DIFFERENCES in kinematics.

Current results are all Pythia8 shower without Pile-Up.

Variables that 'pass this test' would be checked with PU and with different showering schemes.

Methodology

- Kinematics were adjusted in LHE files by consistently adjusting 4-vectors, scale of event, QCD coupling but leaving parton content & color connections intact.

- Examined various implementations of

jet mass	Pull Angles	R-cores
girth	Radial Pull	N-subjettiness
N_charged_track	Axis Contraction	
	Dipolarity	
	Color Connectivity	
		(more being added)

- Automated cut optimization for any combination of one or two different variables.
  Could do MVA, but wanted to systematically work our way through various correlations.

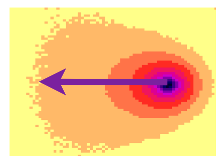
- Examine variable volatility and kinematic correlation by comparing full sample with/without modified kinematics and repeated showerings of a single parton-level event.

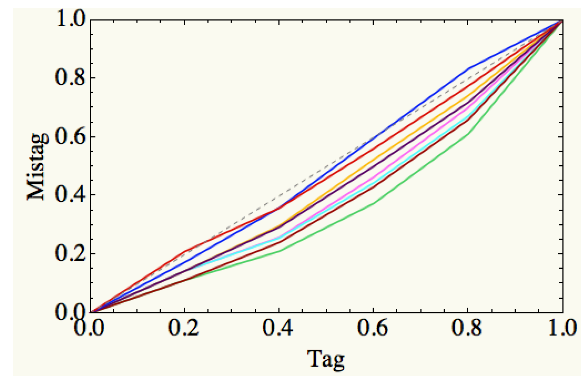
Summary of Main Results

- Color flow variables by themselves do not help to improve S/$\sqrt{}$B much since the final state has just two jets.
  But we confirm that **some of them** are largely uncorrelated with q/g tagging observables, so they do contain additional information.

➡ Simpler color flow variables do better than maximally 'clever' ones.

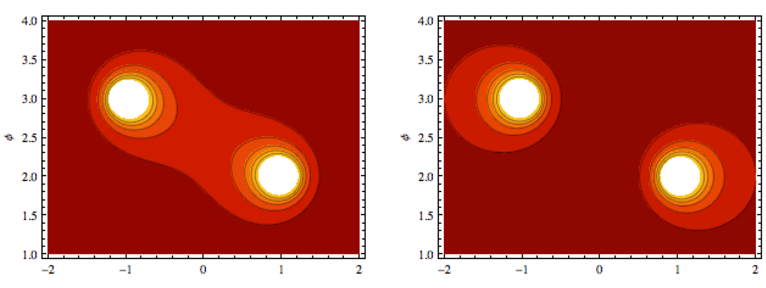
For example, radial pull = projection of pull along line connecting two jets [basically jet skew]

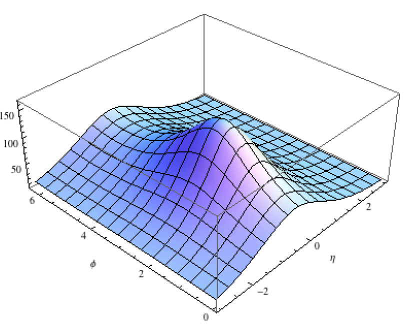


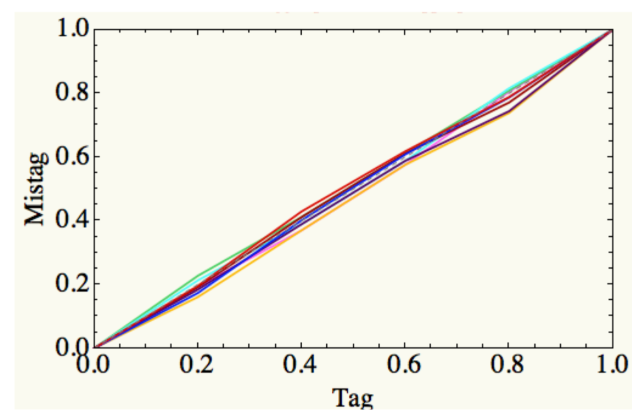


vs

color connectivity [developed with Jason Gallicchio and Yang-Ting Chien]
which matches 2-subjet radiation pattern to eikonal template





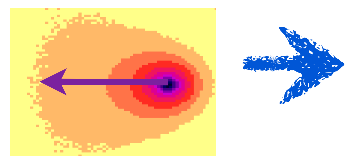


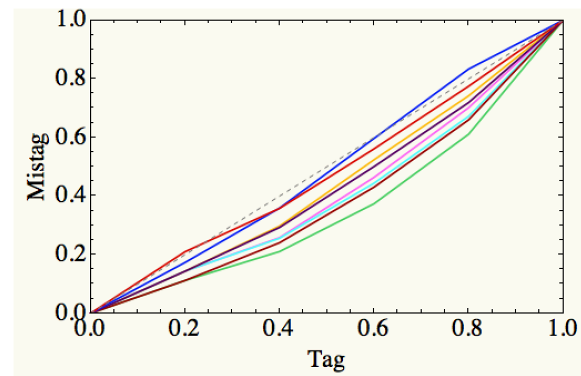
Summary of Main Results

- Color flow variables by themselves do not help to improve $S/\sqrt{B}$ much since the final state has just two jets.
  But we confirm that **some of them** are largely uncorrelated with q/g tagging observables, so they do contain additional information.

➡ Simpler color flow variables do better than maximally 'clever' ones.

For example, radial pull = projection of pull along line connecting two jets [basically jet skew]

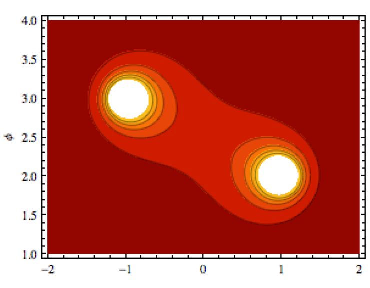


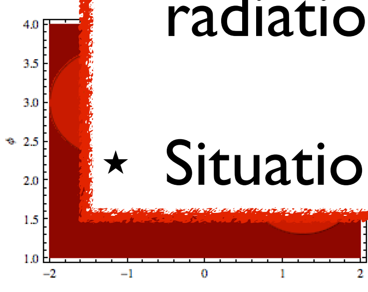


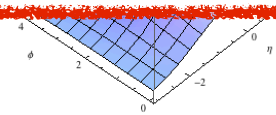
vs

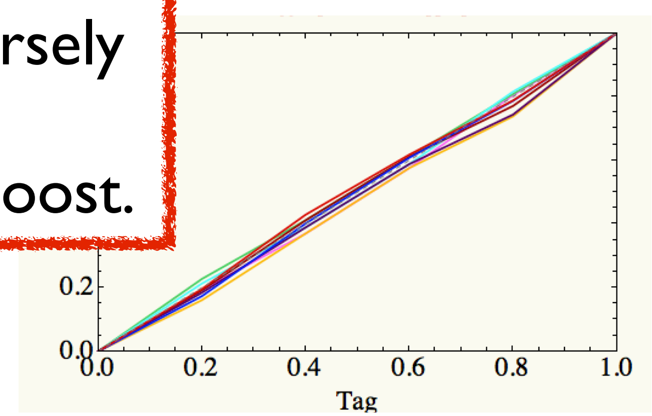
color conn[...]

which ma[...] late









★ Relies on tell-tale color singlet radiation that is emitted too sparsely

★ Situation may change at higher boost.

Summary of Main Results

- For all sample comparisons, quark-gluon distinguishing variables (girth, Ncharged) do as well as all other methods (though they may not be mutually exclusive, still under investigation)
    - ➡ These additive variables impose ‘hard’ limits on the $S/\sqrt{B}$ improvement of ~ 2. Often much lower.

- For (singlet) → 2 quarks or 2 gluons, dominant background is often quark-gluon QCD dijets. This is a difficult background to distinguish.

Summary of Main Results

- The most useful variables fall into two separate groups:
  - N_charged
    girth
    thin jet mass
    dipolarity (color flow variable that weighs radiation along line connecting two jets)
    are correlated amongst each other with correlation ~ 0.5
  - radial pull (projection of pull angle along line connecting jets)
    axis contraction (change of τ_{21} axes with changing minimization measure)
    are correlated amongst each other with correlation ~ 0.7
  - correlation between these two groups is < 0.1

Summary of Main Results

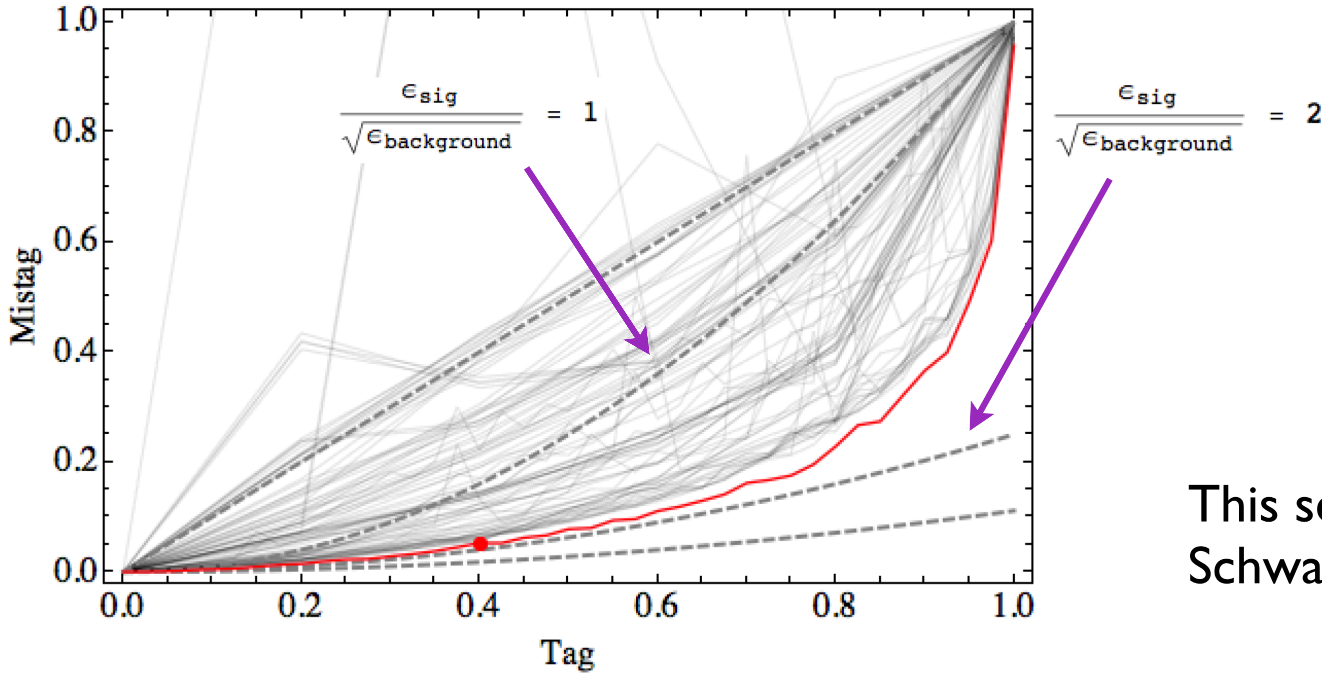
- The most useful variables fall into two separate groups:
  - N_charged
    girth
    thin jet mass
    dipolarity (color flow variable that weighs radiation along line connecting two jets)
    are correlated amongst each other with correlation ~ 0.5
  - radial pull (projection of pull angle along line connecting jets)
    axis contraction (change of τ_{21} axes with changing minimization measure)
    are correlated amongst each other with correlation ~ 0.7
  - correlation between these two groups is < 0.1

parton identity

bit of both?

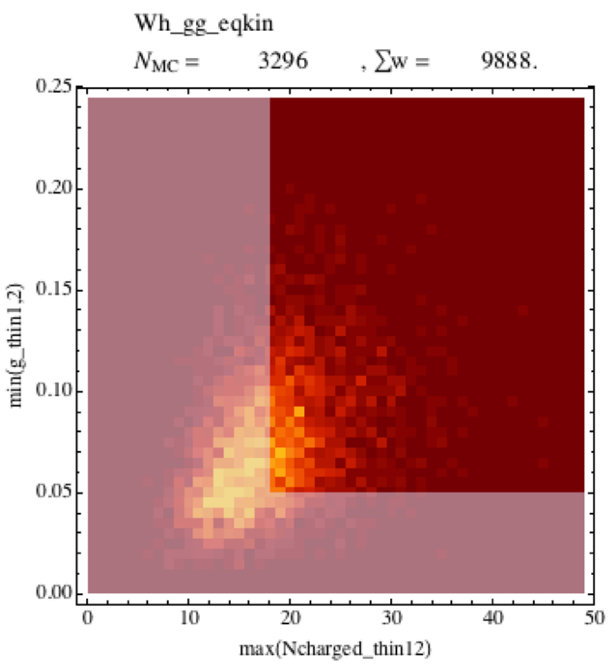
color connections

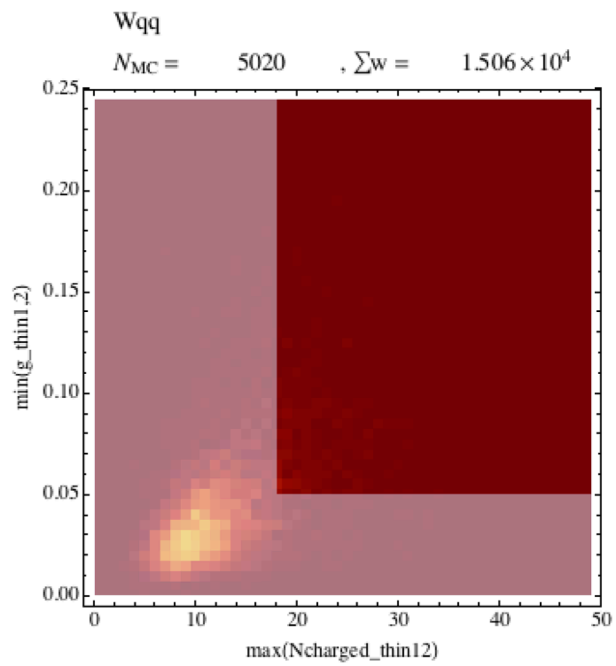
Sample Comparison: W(h→gg) vs Wqq



color-singlet (gg)
vs
non-singlet (qq)

This seems consistent with Gallicchio, Schwartz results (e.g. 1106.3076)

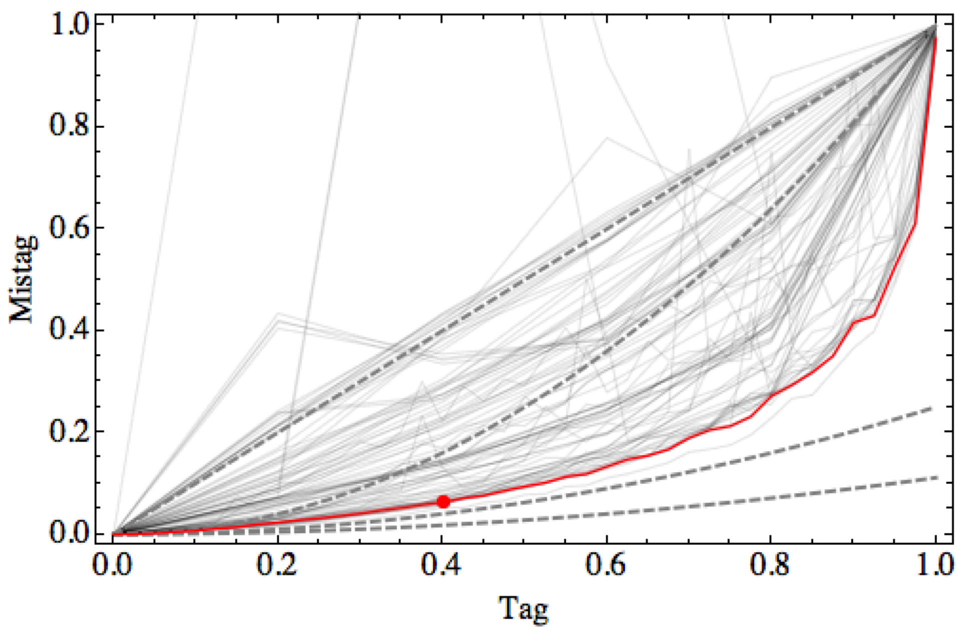


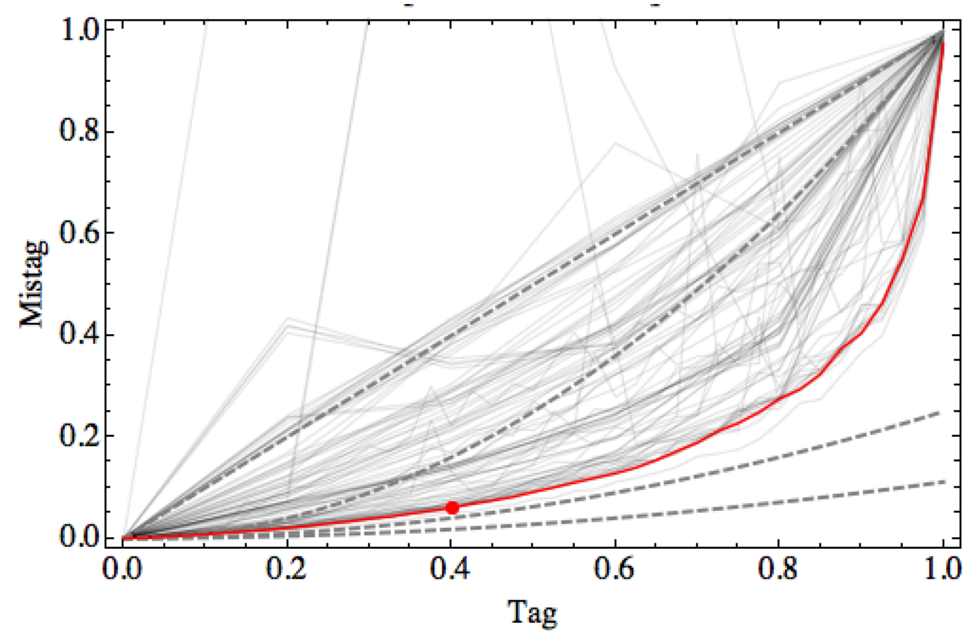


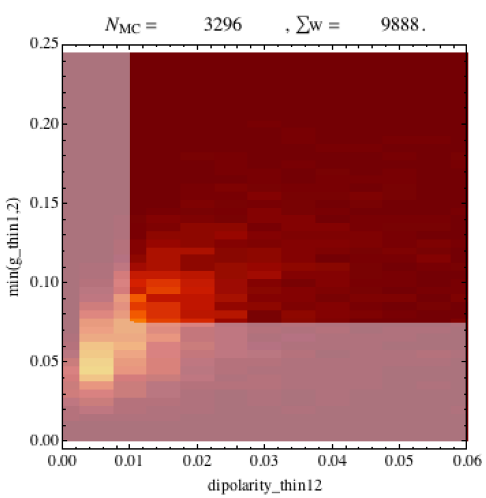
Sample Comparison: W(h→gg) vs Wqq

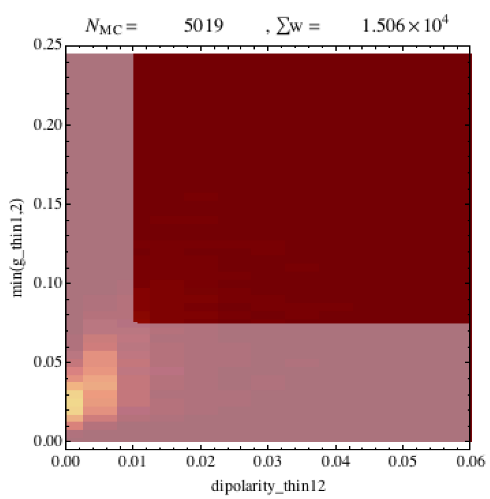
Replacing Ncharged (or girth) by **dipolarity** is just as good.

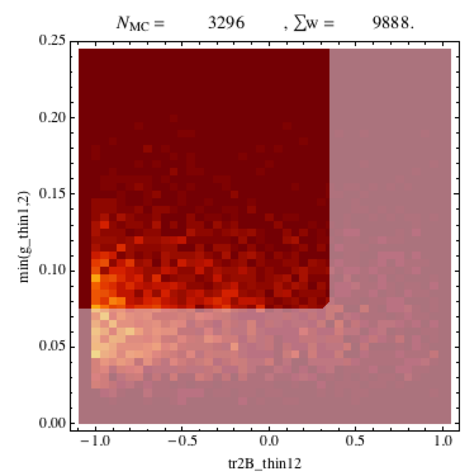
Replacing Ncharged (or girth) by **radial pull or axis contraction** is just as good.

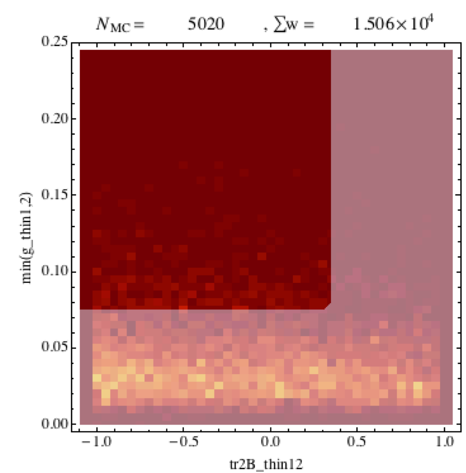








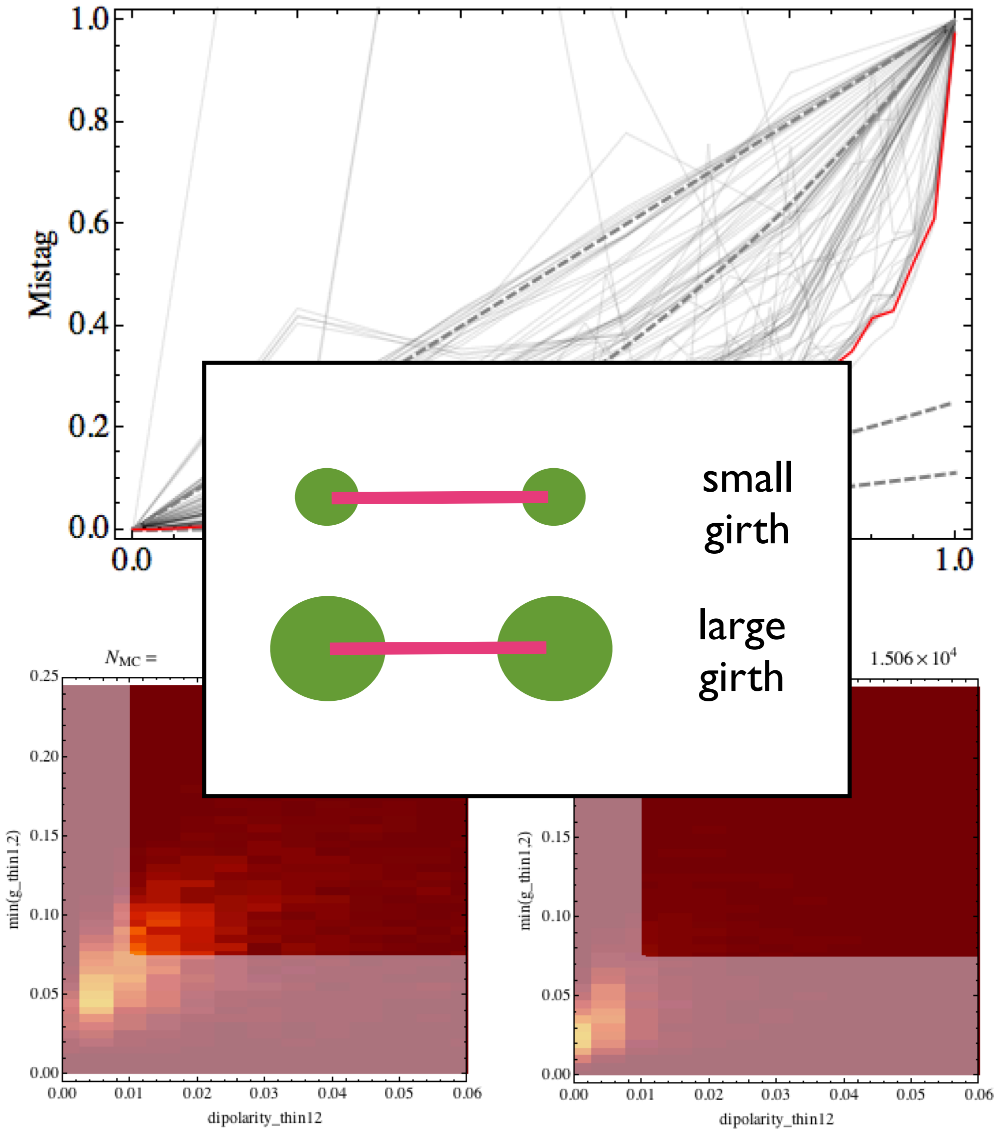


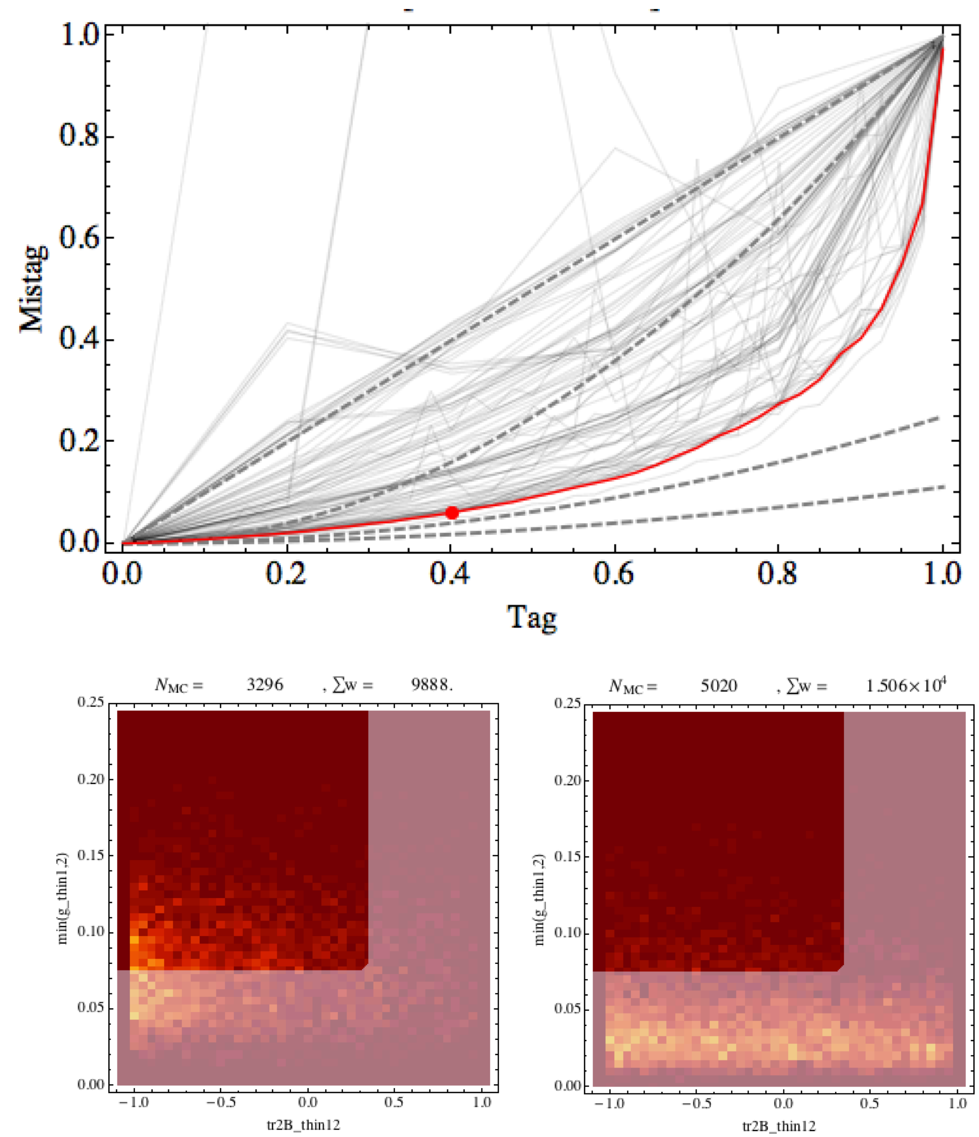


Sample Comparison: W(h→gg) vs Wqq

Replacing Ncharged (or girth) by **dipolarity** is just as good.

Replacing Ncharged (or girth) by **radial pull or axis contraction** is just as good.

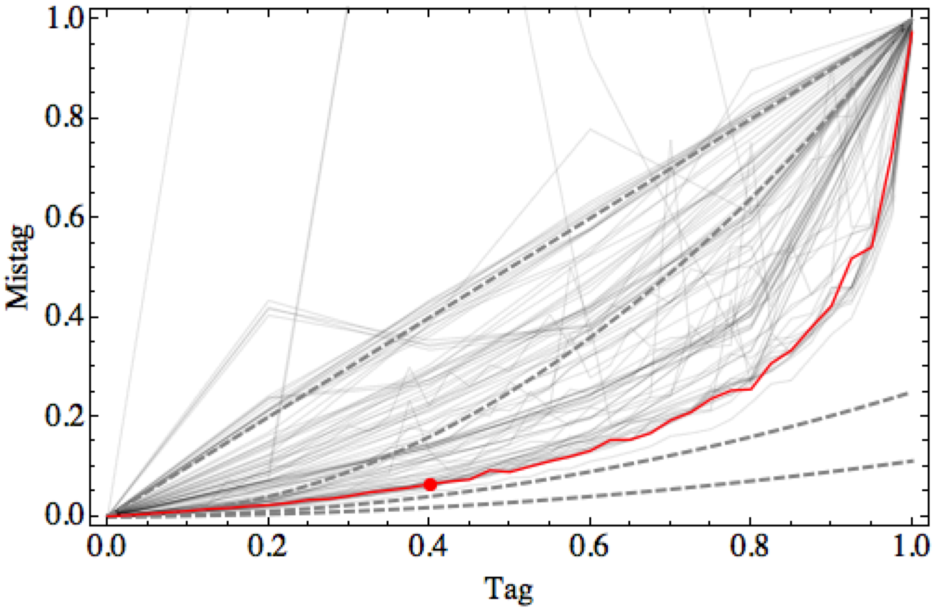


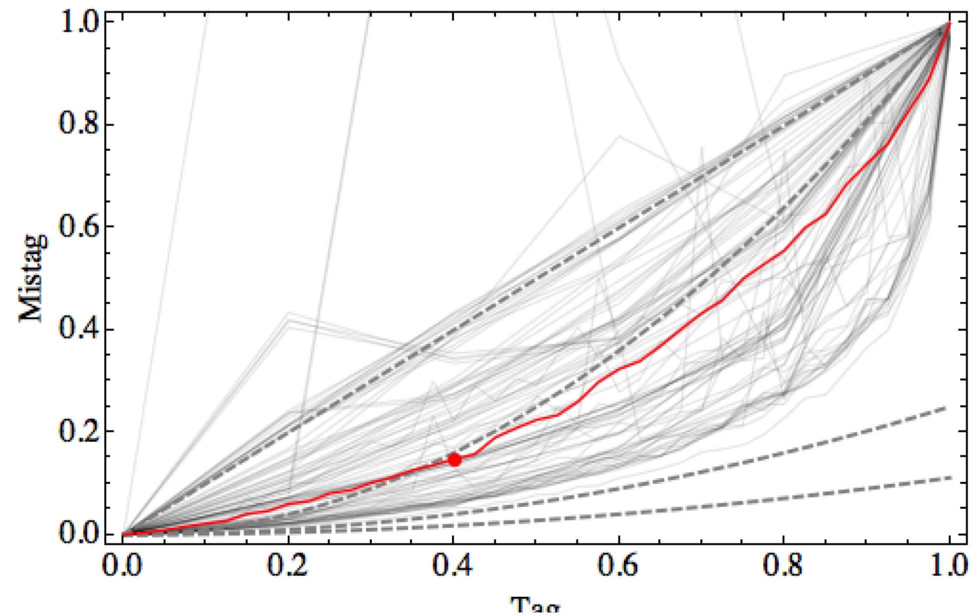


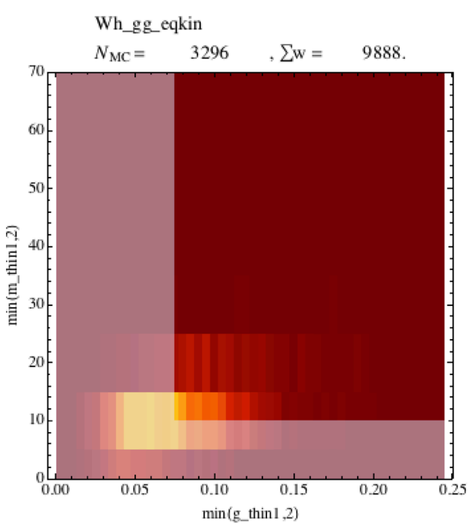
Sample Comparison: W(h→gg) vs Wqq

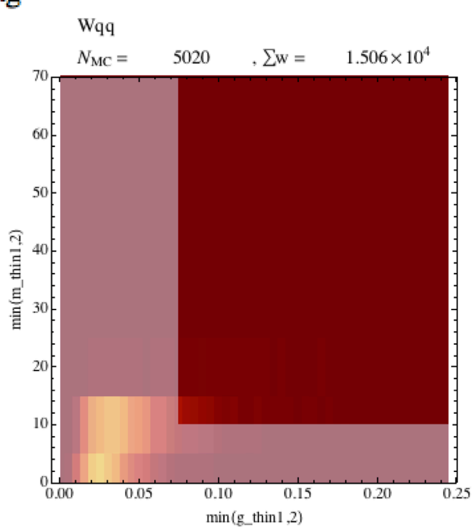
Replacing Ncharged (or girth) by **thin jet mass** is just as good.

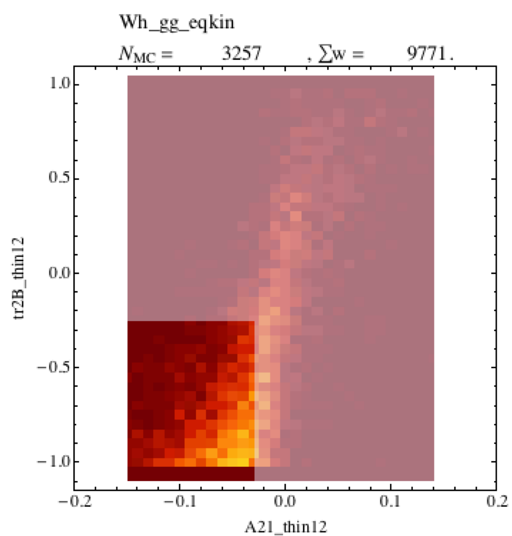
‘Pure’ color flow 2-variable cuts (axis contraction vs radial pull) just manage not to reduce significance.

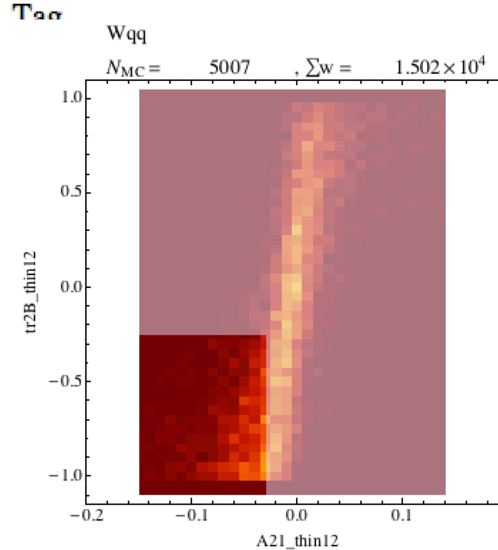




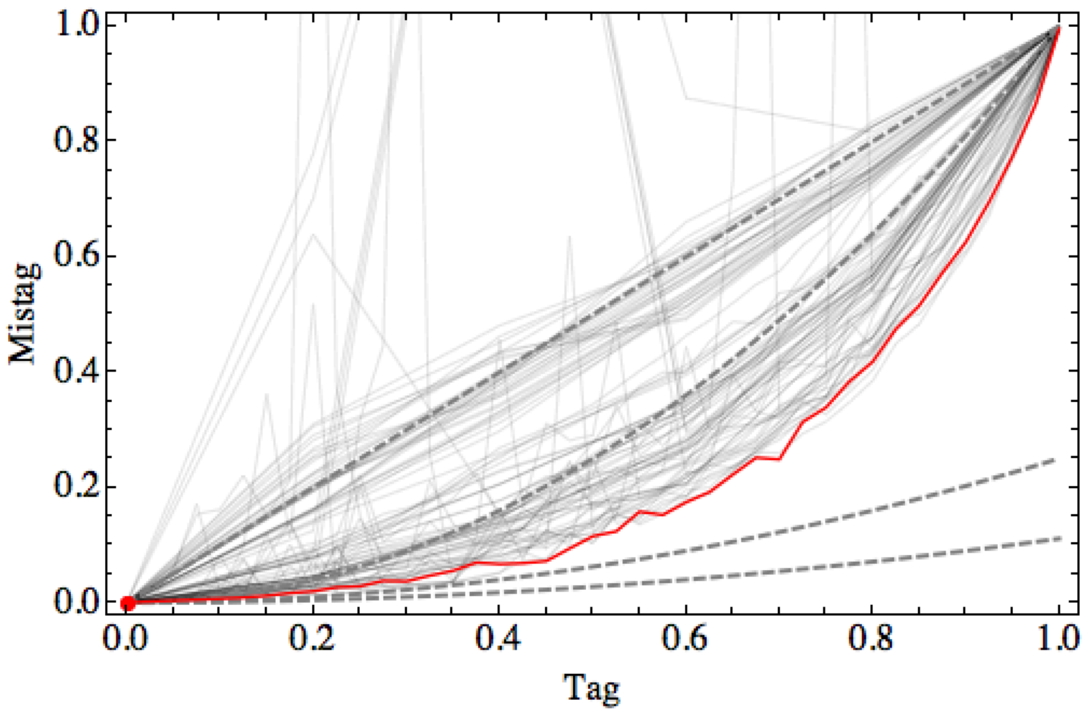








Sample Comparison: W(h→qq) vs Wgg



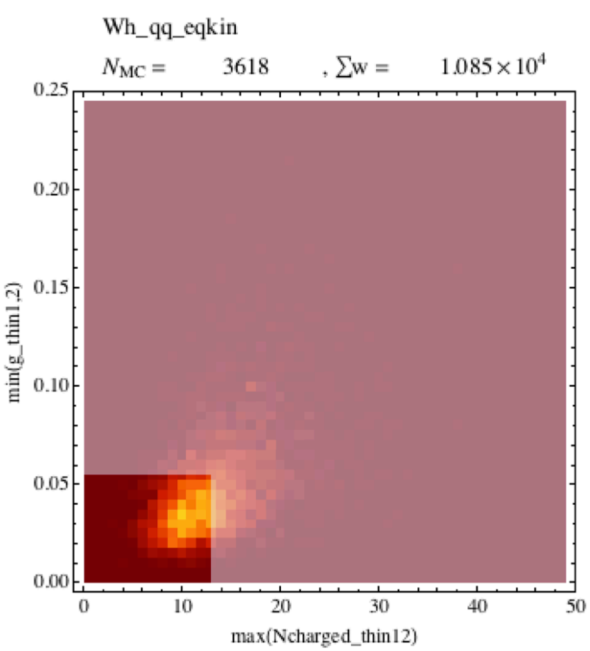
color-singlet (qq)

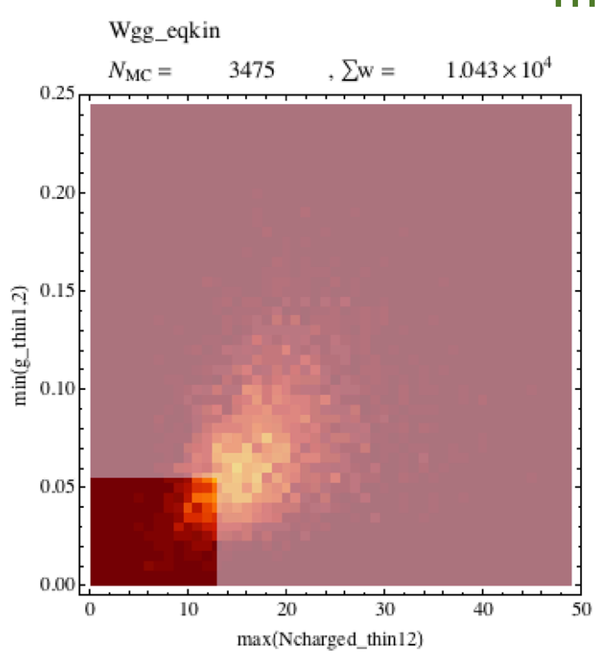
vs

non-singlet (gg)

Same story, no surprise there.
(A little worse than hgg vs qq since the gg singlet radiates more.)

Again, either Ncharged or girth can be replaced by color flow or thin jet mass

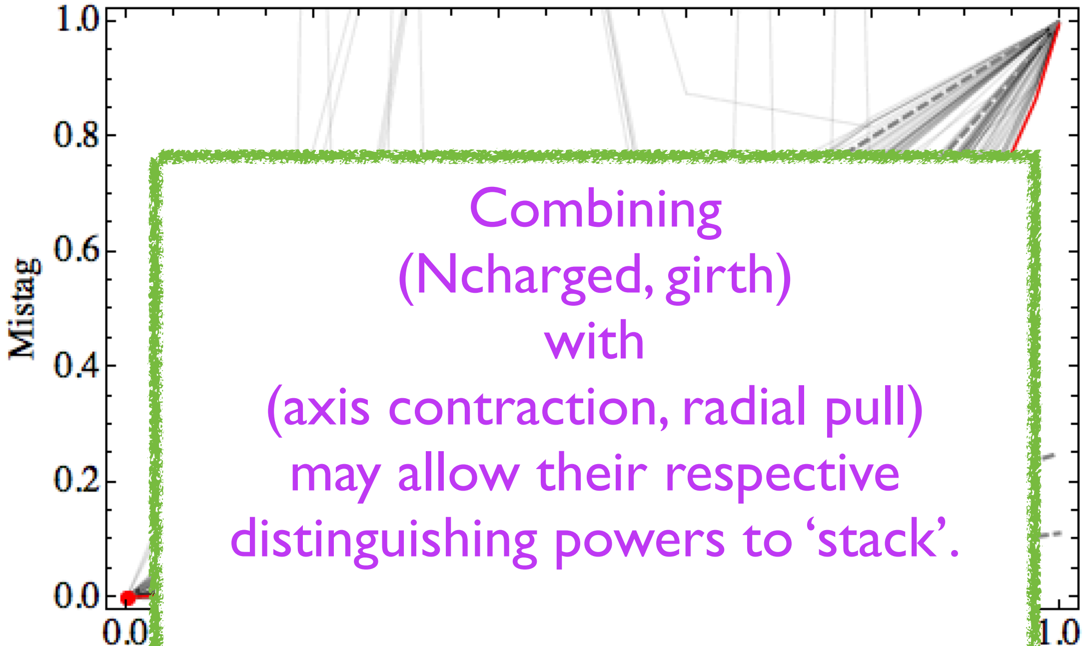




For
color singlets (XX)
vs
color non-singlet (YY)

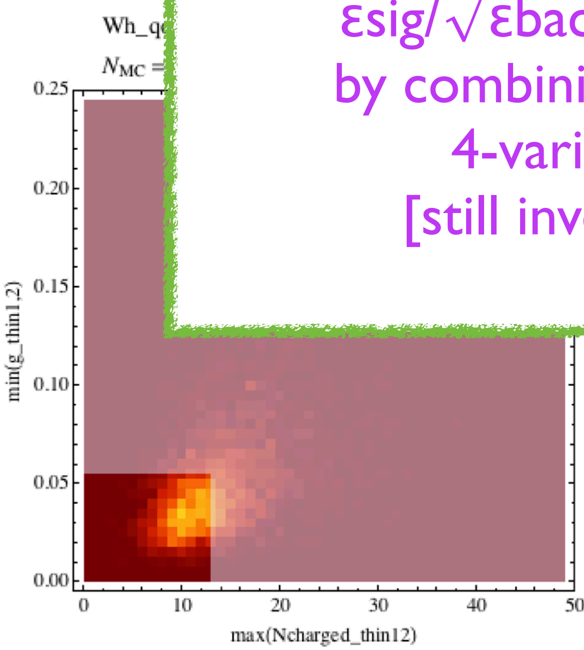
you can gain
$\varepsilon_{sig}/\sqrt{\varepsilon_{background}} = 2$

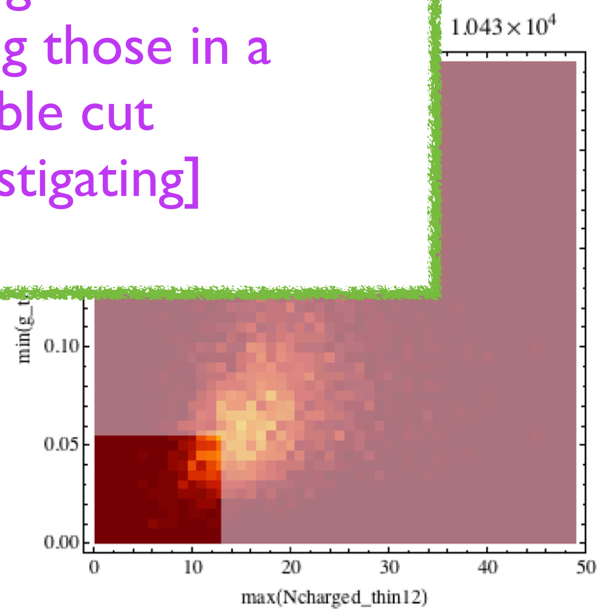
Sample Comparison: W(h→qq) vs Wgg



Combining
(Ncharged, girth)
with
(axis contraction, radial pull)
may allow their respective
distinguishing powers to ‘stack’.

Might get
εsig/√εbackground > 2
by combining those in a
4-variable cut
[still investigating]





color-singlet (qq)
vs
non-singlet (gg)

Same story, no surprise there.
(A little worse than hgg vs qq since the gg singlet radiates more.)

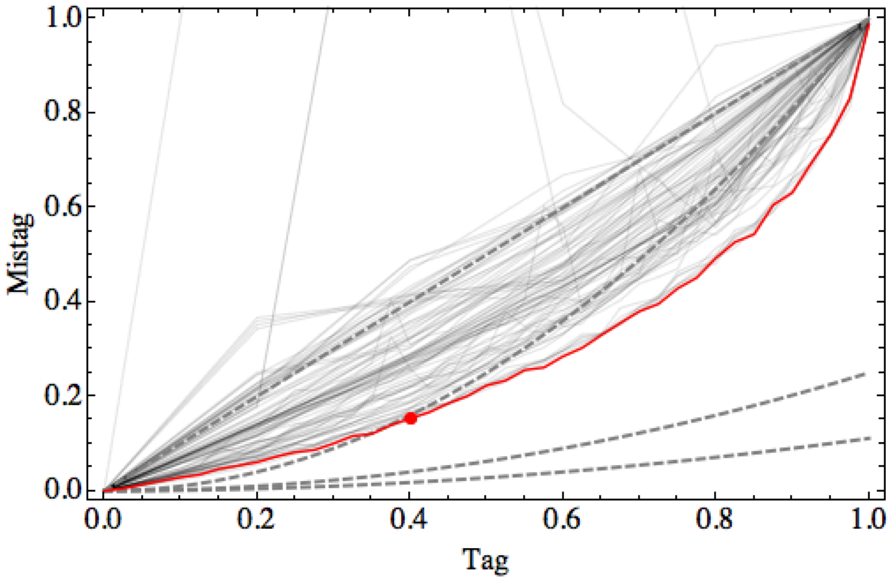
Again, either Ncharged or girth can be replaced by color flow or thin jet mass

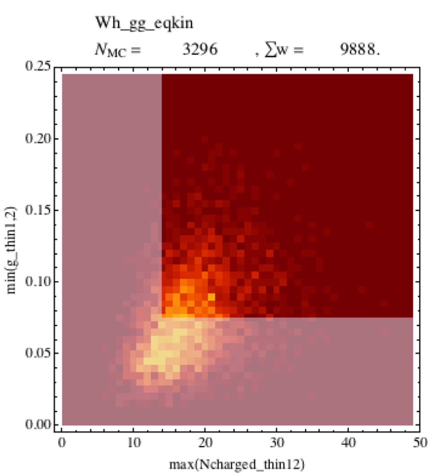
For
color singlets (XX)
vs
color non-singlet (YY)

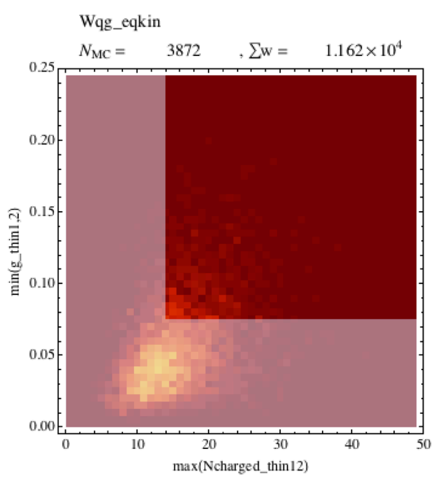
you can gain
εsig/√εbackground = 2

Sample Comparison: W(h→gg) vs Wqg

Again Ncharged x girth is best,
or replacing either by color flow or thin mass.



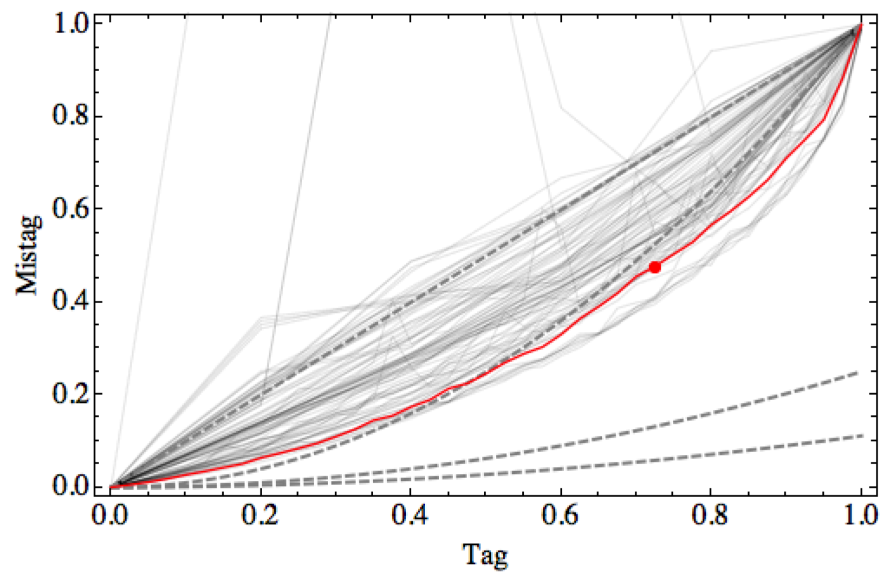




color-singlet (gg)
vs
non-singlet (qg)

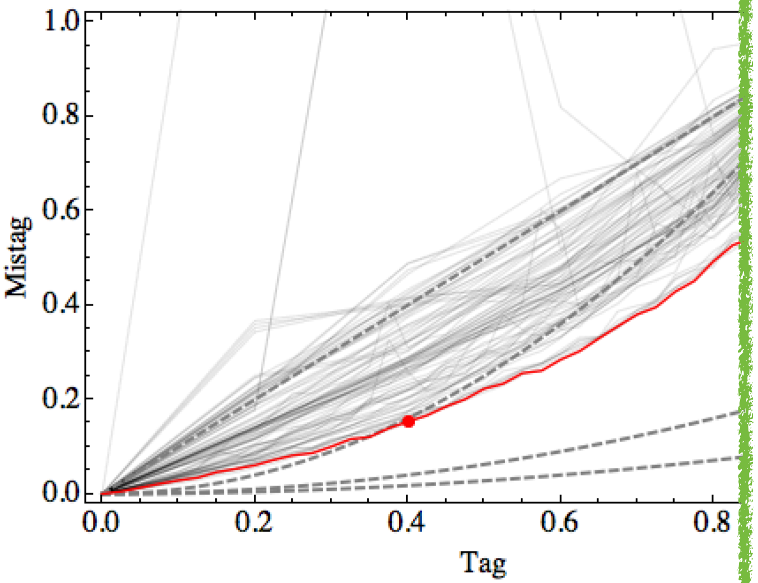
Often a dominant Background!

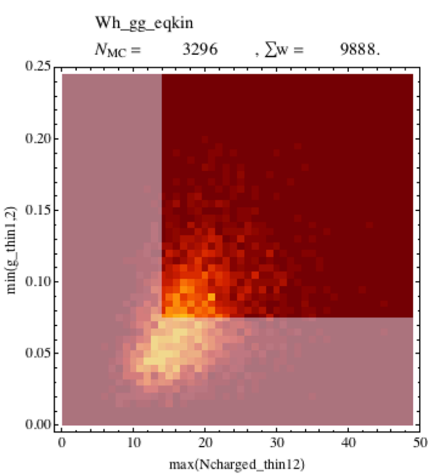
Again, pure color flow cuts a bit worse, but these sets of variables seem uncorrelated.

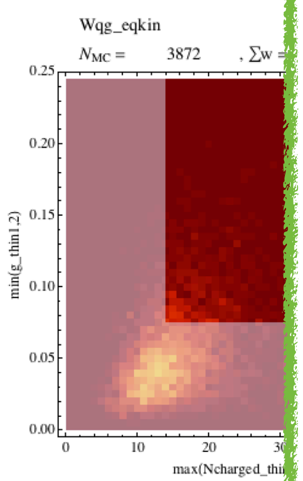


Sample Comp … qg

Again Ncharged x girth is
or replacing either by color flow



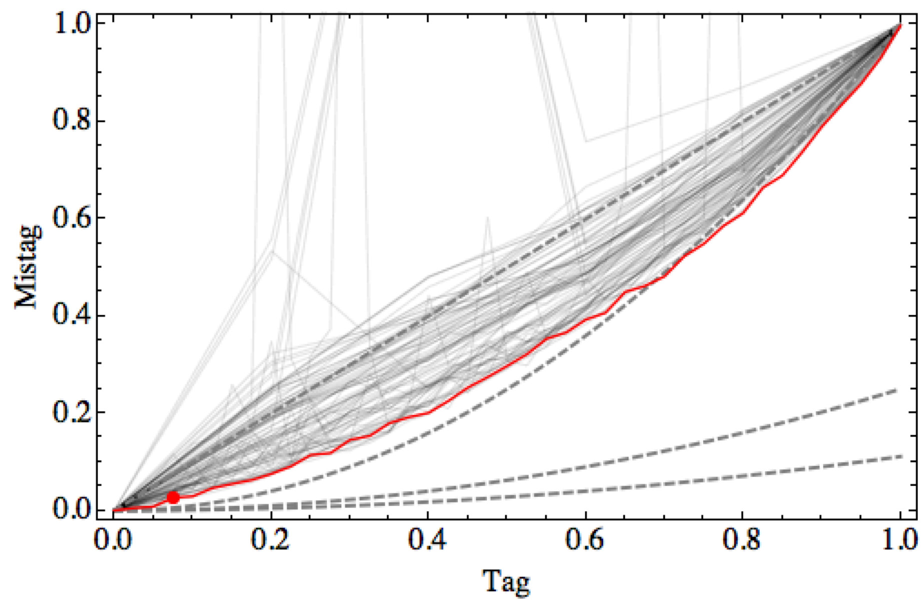




Similar story

for

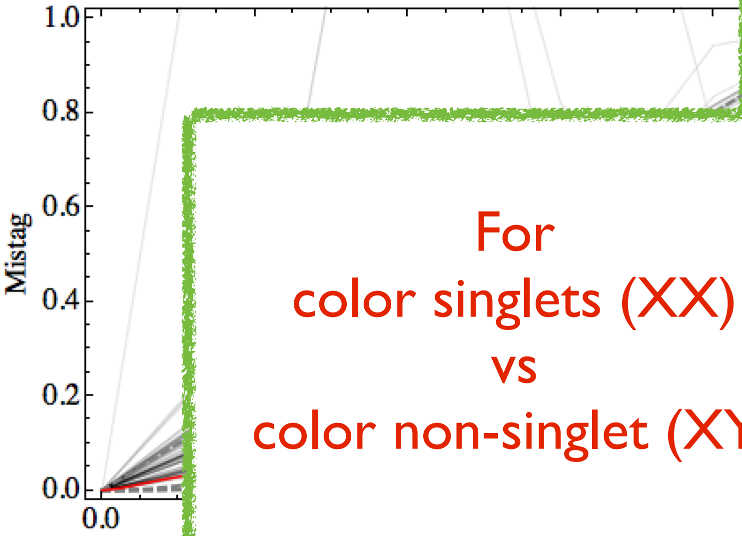
W(h→qq) vs Wqg



(a little worse, again because of less radiating singlet)

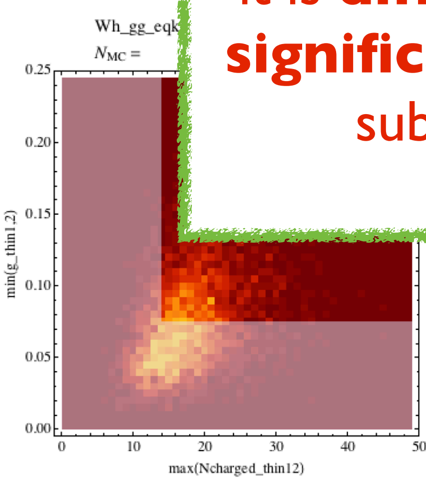
Sample Comp

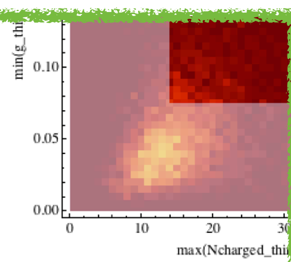
Again Ncharged x girth is
or replacing either by color flow



For
color singlets (XX)
vs
color non-singlet (XY)

it is **difficult to gain significance** using soft substructure.

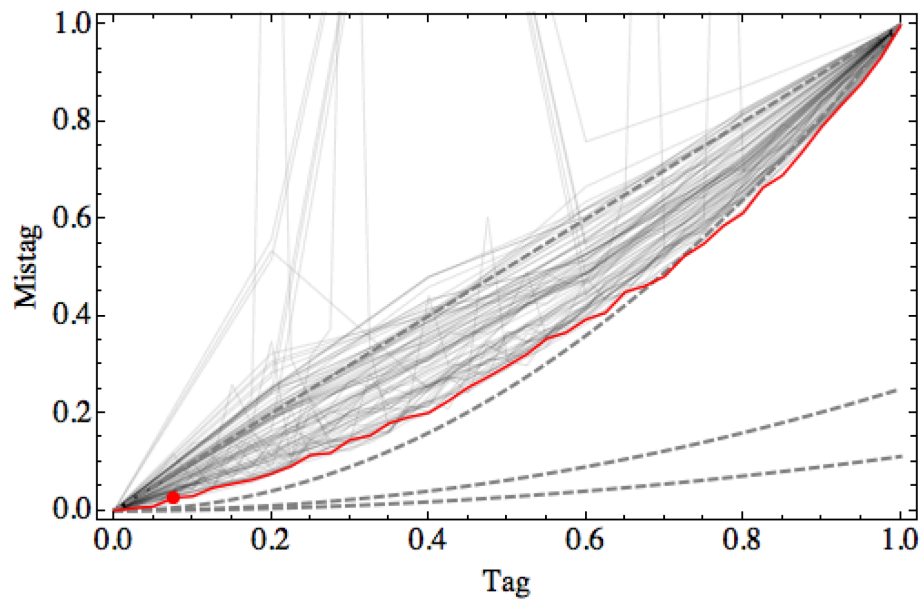




Similar story

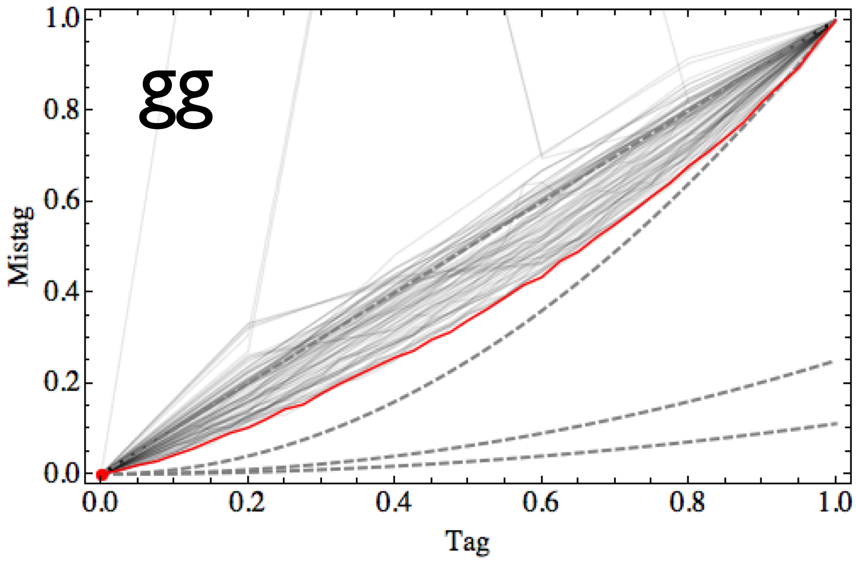
for

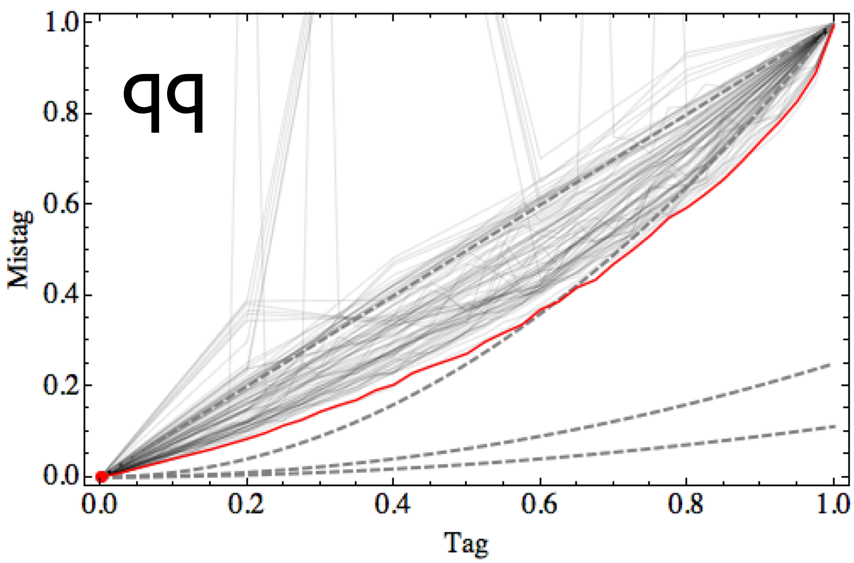
W(h→qq) vs Wqg



(a little worse, again because of less radiating singlet)

Sample Comparison: W(h→gg/qq) vs Wgg/qq





Color flow 1- or 2- variable cuts do comparably 'badly'.

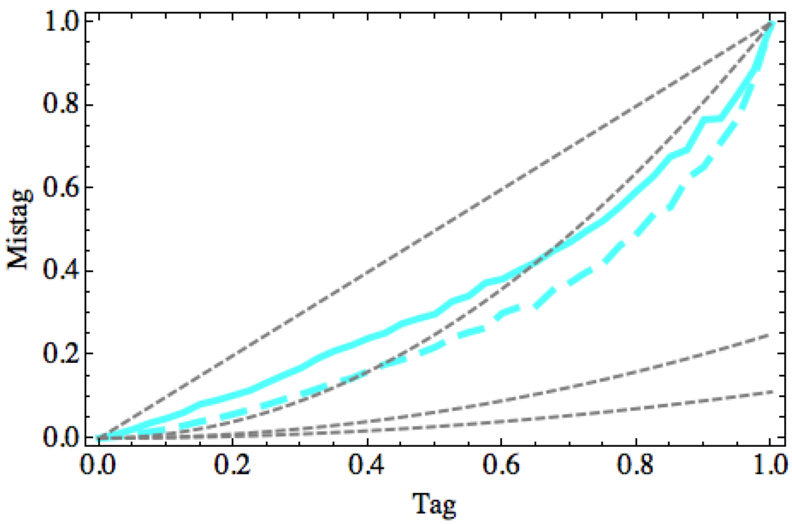
Even q/g-distinguishing variables do comparably.

For
color singlets (XX)
vs
color non-singlet (XX)

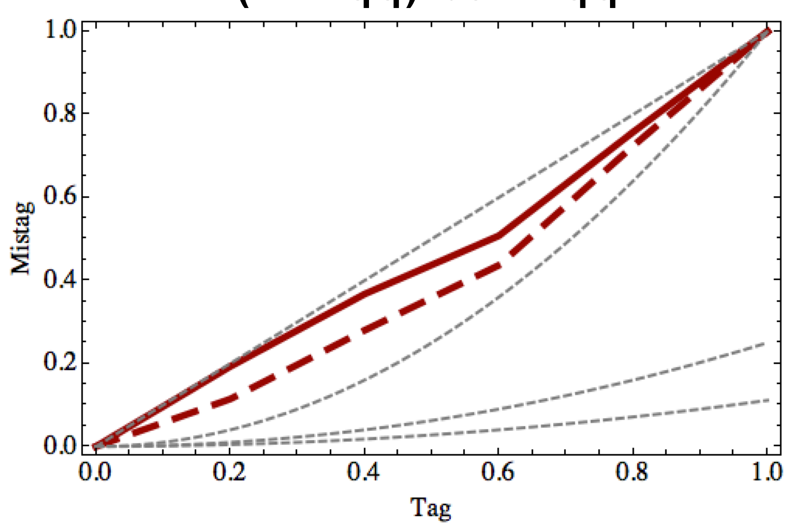
it is ***very* difficult to gain significance** using soft substructure.

Kinematic Dependencies

(Ncharged, m_thin) for
W(h→gg) vs Wqg



R-core for
W(h→qq) vs Wqq



equalized kinematics
original kinematics

Significant improvement in discrimination power for un-equalized kinematics.

Underlines importance of our equalization procedure for making general and 'factorizable' search strategy statements.

Still to do...

- q/g-tagging type variable seem relatively uncorrelated from color-flow type variables.
  - Need to establish how much the respective significance improvements factorize.
  - Hopefully identify a minimal variable combination that eliminates redundancy and gives clear guidance for experimental deployment.
  - Preliminary results suggest a good choice would be (girth, Ncharged) for q/g tagging [established] and (radial pull and/or axis contraction) for color flow
- Will need to compare MC generators and robustness under PU (e.g. using only charged tracks)

Conclusions

- Our equalization of kinematics allows us to make very general statements about 'factorizable' search strategies to guide experimental deployment.
  - Studies should be organized by topology (# singlets & # jets/singlet) and boost factor
  - Generalizing our approach to high boost might take some care (in progress)
  - *A priori* these results should be able to be 'pasted' on top of existing experimental studies, pending verification.

Important check for your favorite substructure technique

- For (1 singlet → 2 jets)
  - We have identified two uncorrelated groups of radiation variables: q/g tagging type variables and a subset of color flow variables.
  - In this topology, the utility of radiation variables is strongly limited, even in our optimistic preliminary study.
  - What # singlets & # jets/singlet do you need to make this worth the effort?