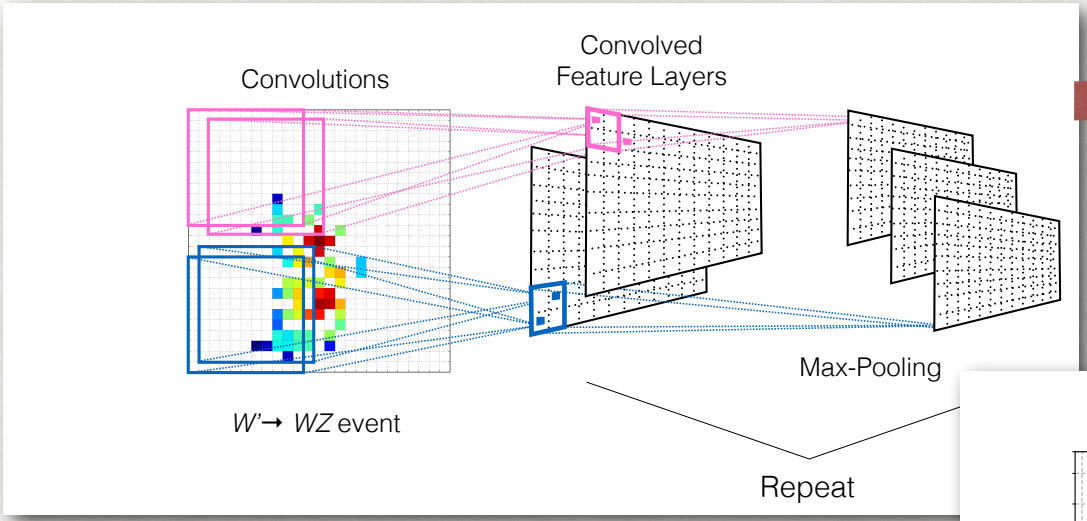
ジェットの深層学習の決定過程の解析と物理応用
(ミンコフキー汎関数を中心にして)

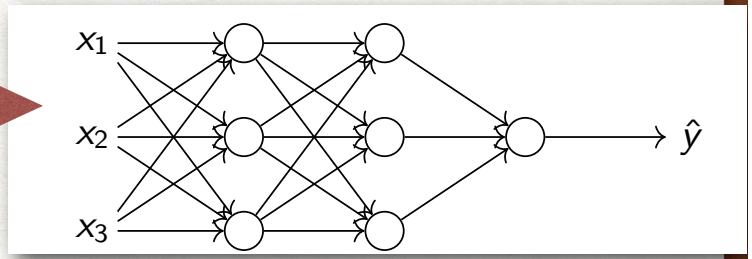
KEK 野尻美保子

ジェットの深層学習

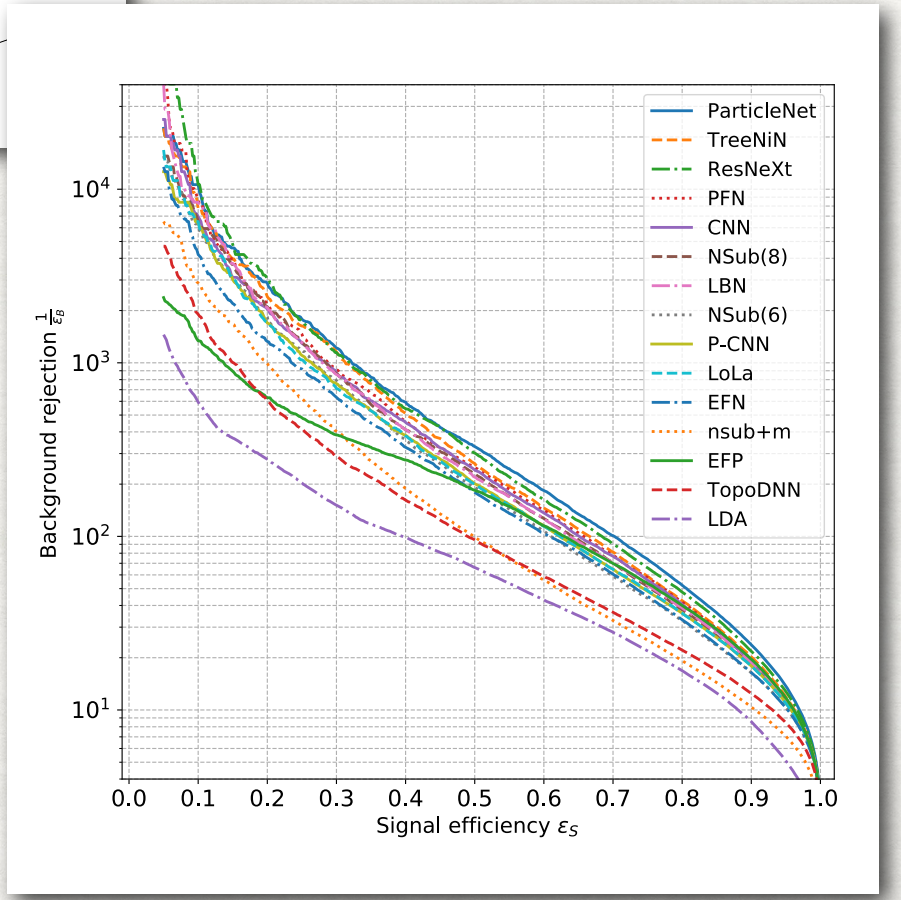
- Jet のsubstructure からHeavy 粒子の選別　今後の高輝度化、高エネルギー化→ high pT のHiggs 粒子　top 粒子の選別, mass drop tagger など摂動的にも良くわかった量から出発
- BDT (機械学習)から深層学習
- 様々なアルゴリズムの提案　高次の量の利用→低次の量(Jet image など)　　MC レベルではよりよい結果を示す。例えば top Higgs vs QCD ジェットなど

CNN とTOP TAGGING





- CNN, ResNeXT, Particle Net… は大体同じパフォーマンス　(右の図は同等の精度のinput で比べていないので注意)



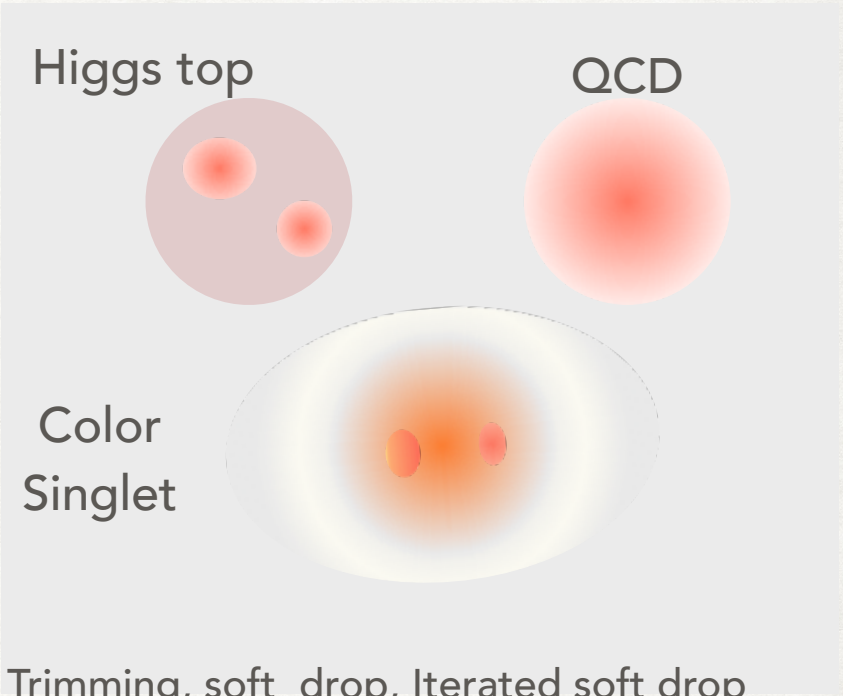
安心できる量と不安な量

IRC SAFETY

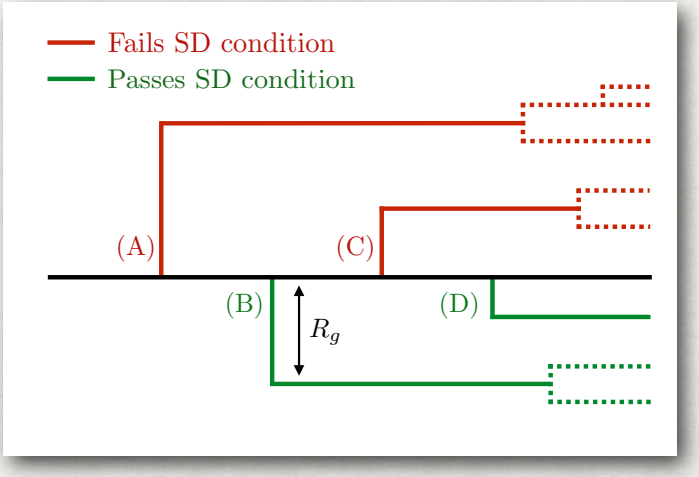
- IRC safe object: soft or collinear emission に対して安定な量, subjet
- Soft collinear に対して不安定な量　number of tracks, particles
  MC modeling に大きな差 (いつまで立っても収束しないPythia と Hewig 実験データなど)
- →Jet Image　の利用:　両者の間の区別が不明確　系統誤差の評価に問題はないか。実データで補正はやや楽観的ではないか。
- 理論研究：ジェットの構造の中でresummation でうまく記述できる部分を切りだす。

深層学習によるジェット分類の中で、
IRC SENSTIVE な部分がどのように働いているか。

- 典型的なジェットイメージで分類に
- Higgs ハードな部分が複数ある(mass drop) vs QCD: あまりない
- W, H, Z 粒子束が独立する傾向 vs QCD 撒き散らし。カラーコヒーレンスは以前からみたい量。
- クオーク (どちらかと言うと）芯がある vs グルーオンより広い
- 注意　underlying event とかmultiple scattering とかあるのでLHC の環境では難しい量. Soft の部分を切り捨てて、摂動計算と比較できる物理をやりたいという流れもある。
- **今のトーク　ソフトな粒子の分布に注目したい。**

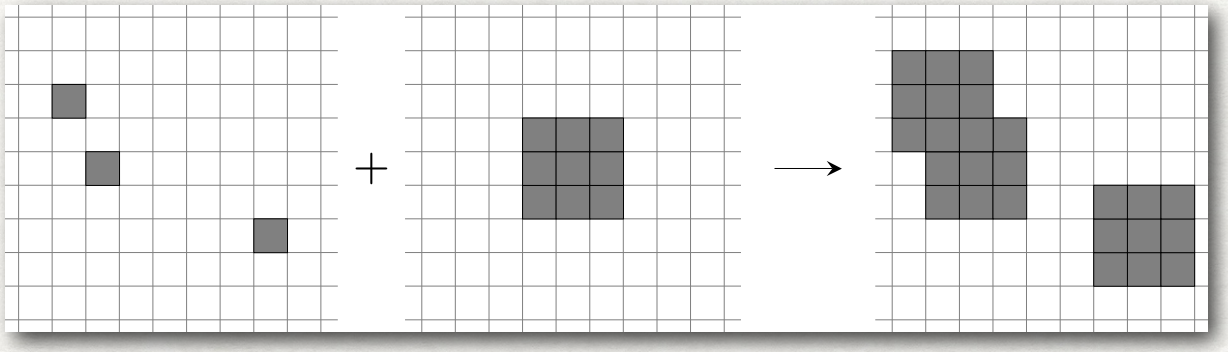


Trimming, soft drop, Iterated soft drop



粒子分布を定量化する

- ソフトなトラックの数だけではなくて、ソフトな粒子の相互の近接度を「定量的」に評価したい　　→キャリブレーションやMC turning に便利なように

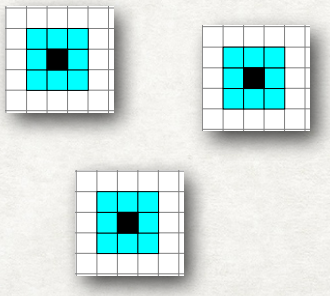


ジェットイメージ　　3x3 のマスク　　N(1)　　… N(n)

$N_0 = N_{pixel} = 3$

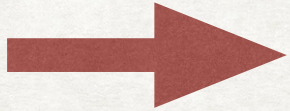
$N_1 = 25$

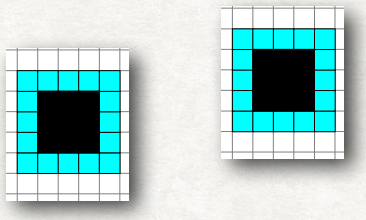
- N_1/N_0 is sensitive to how pixel clustered

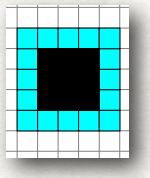


more clustered

$N_1/N_0 = 9$





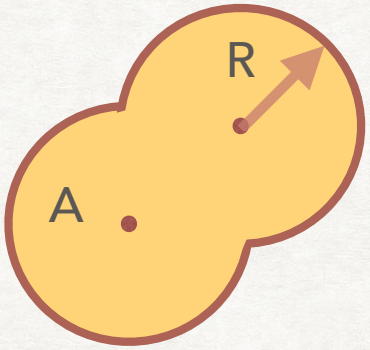


if all pixel appear in 3x3

$N_1/N_0 = 16/9 = 1.78$

背景にある数学 INTEGRAL GEOMETRY

Object の相対的な位置を定量化する理論



凸体の測度（大きさ）の満たすべき性質

並進　$M(gB) = M(B)$

加法　$M(B_1 \bigcup B_2) = M(B_1) + M(B_2) - M(B_1 \bigcap B_2)$

連続性　$M(K_i) \to M(K)$ as $K_1 \to K$ for $K, K_i \in \mathcal{R}$

- 凸体の測度（大きさに関する定理）n次元では、この性質を満たす関数は n+1 しかない (ハドビガーの定理)　これをミンコフスキー汎関数(MF) という。2次元であればこれは　1. Surface Area (A), 2. Length of the boundary(L) 3. Euler characteristic (χ)

点-> 点を中心に半径R の円を追加→ A(R), L(R),X(R) が点分布を記述

MF の他の物理応用

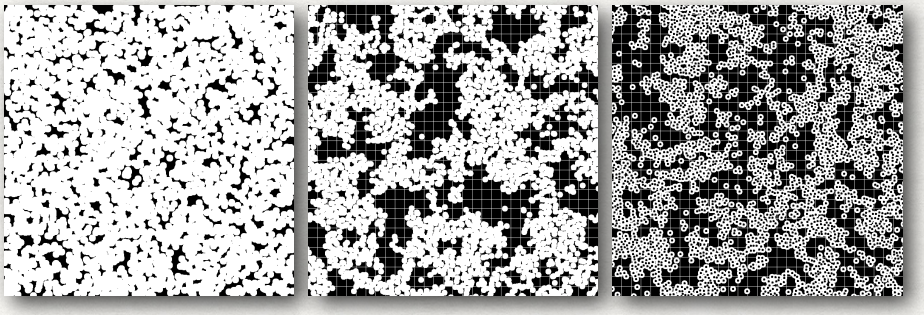
統計物理

左 多孔質体

真ん中: 微乳濁液

左 コロイド

体積の占有状況V, 表面の大きさ(S) 等に依存して物性が変わる　図は　Mecke and Stoyan (2000)



Kratochvil　1109.6334　Proving Cosmology with Weak Lensing Minkowski Functinal s

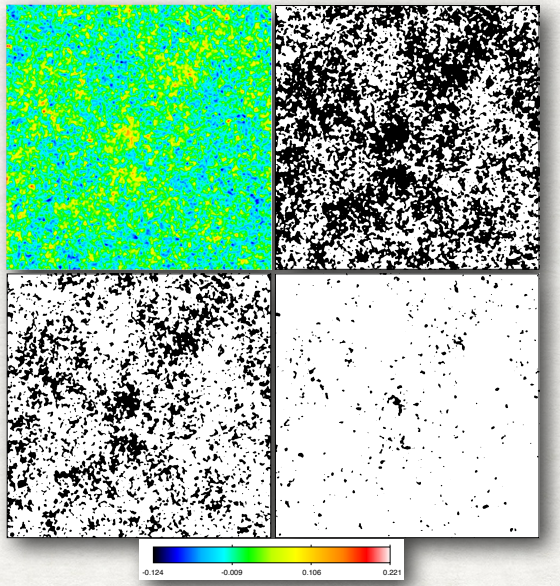


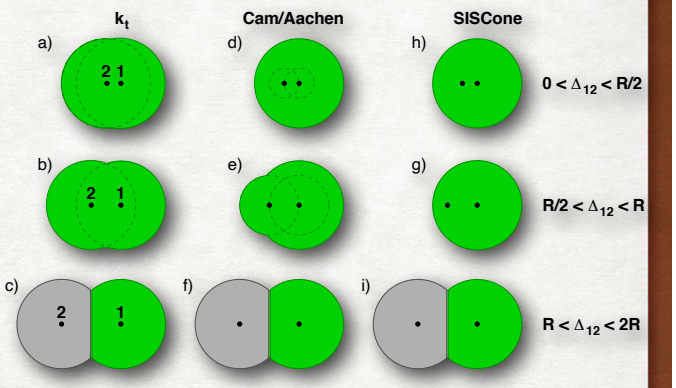
FIG. 1: *Top left panel: example of a simulated 12-square-degree convergence map in the fiducial cosmology, with intrinsic ellipticity noise from source galaxies and $\theta_G = 1$ arcmin Gaussian smoothing. A source galaxy density of $n_{\rm gal} = 15/{\rm arcmin}^2$ at redshift $z_s = 2$ was assumed. Other three panels: the excursion sets above three different convergence thresholds κ, i.e. all pixels with values above (below) the threshold are black (white). The threshold values are $\kappa = 0.0$ (top right), $\kappa = 0.02$ (bottom left), and $\kappa = 0.07$ (bottom right). The Minkowski Functionals V_0, V_1, and V_2 measure the area, boundary length, and Euler characteristic (or genus), respectively, of the black regions as a function of threshold.*

- 天文: 星の分布の定量化、 銀河分布、シミュレーション結果の定量化、non-Gaussinaity of CMB, weak lensing..

点の集まりの意味を定量的に表現する時に使う

CNN はMF(ミンコフスキー汎関数） をみているか

- MF の一意性: 原理的には、ドットイメージとMF はある意味で等価。（情報は落ちてない）
- MF はCNNとは相性は良さそう。
- 基準点がないのでジェットアルゴリズムと相性が良い。そもそも、 Jet Algorithm の Voronoi 領域と”The Catchment Area of Jets “(Cacciari & Salam 2008)も 同じ系列の数学
- 深層学習の分類問題との関係： QCD と t, Z, W のジェットイメージのMF が十分に違っていれば、CNN はMF を学習しているだろう。



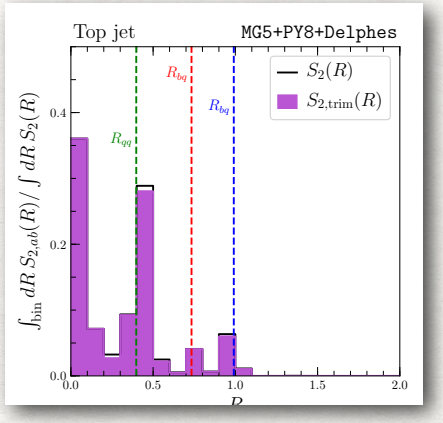
解析:TOP TAGGING の中で

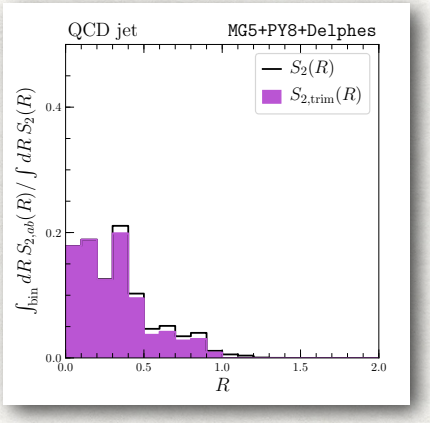
Chakraborty, Lim, Nojiri, Takeuchi 2003.11787

- CNN （ベースモデル）を　高次量をインプットとするMLP と比較
- インプットの分割
  - S2 : C correlator(Energy correlator) f(θ）=Ei Ej δ(θ−θij)　２点関数

    Tkachov (hep-ph 960138) Lim, Nojiri 1807.03312, Chakrabory, Lim Nojiri 1904.02092　~ e_2^{β}の任意のβの情報を担っている。
  - Top なので、3点がメイン；：Leading subjetの粒子との２点関数

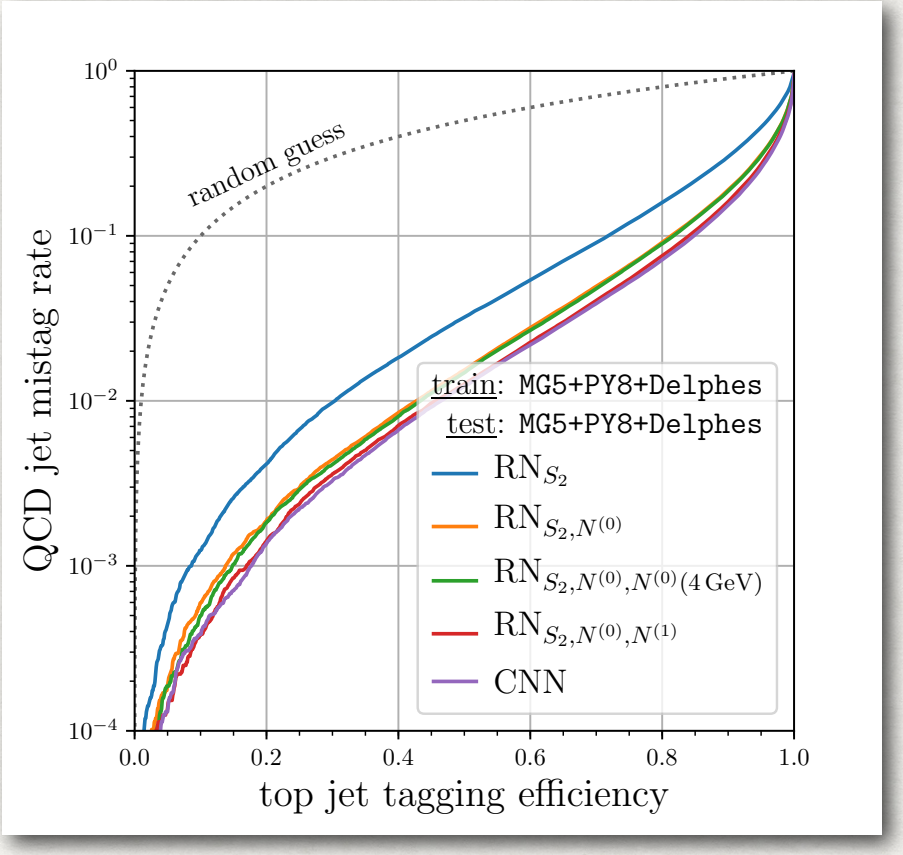
    +Leading subject の粒子を除いた2点関数
  - 上記を hard なsubjet 内の粒子に制限したもの　S2trim (groom でも良い。
  - New :IRC sensitive なインプット　(N0, N1, N0(pt> 4GeV), N1(pt> 4GeV)
- 最終的なMLP は RN (relation network) + global な量(jet pt, mass, trimmed jet 情報と　MF で作る





ROCはCNN と等価

- IRC insensitive な量だけ使った青い線とCNN は非常に差が大きい。
- Top jet vs QCD問題では、hit のあるカロリメータの数(N0) 、その周りのエリア数を加えるとCNN と完全に同じになる
- Top ジェットの場合、カラーを持つ粒子のせいか、機械学習は　ΔR= 0.1より遠距離のソフトな相関を学んでいない。
- Top がカラーを持っている。QCD 生成など



過去の同様なアプローチ

N-subjettiness

$$\tau_N^{(\beta)} = \frac{1}{p_{TJ}} \sum_{i \in \text{Jet}} p_{Ti} \min \left\{ R_{1i}^{\beta}, R_{2i}^{\beta}, \ldots, R_{Ni}^{\beta} \right\}.$$

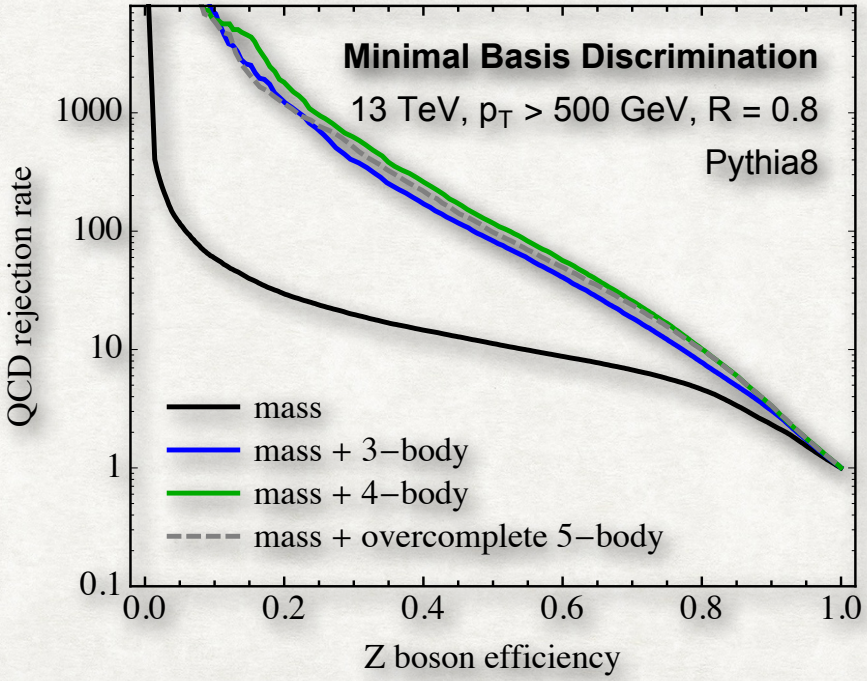
粒子を n 個の軸にうまく
取り込めるかという問題
N 個のスポットがあると τN で突然小さくなる

摂動計算の試みあり

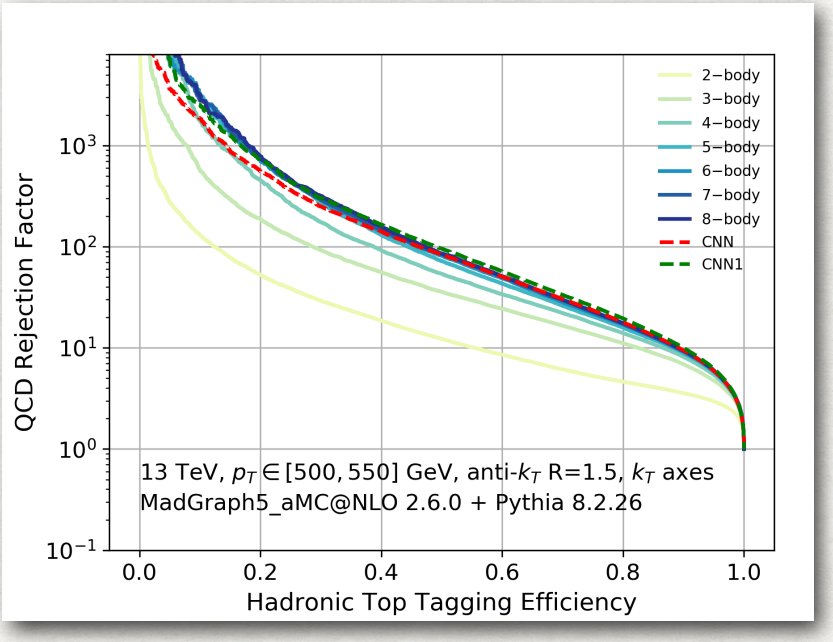
$$\left\{ \tau_1^{(0.5)}, \tau_1^{(1)}, \tau_1^{(2)}, \tau_2^{(0.5)}, \tau_2^{(1)}, \tau_2^{(2)}, \tau_3^{(0.5)}, \tau_3^{(1)}, \tau_3^{(2)}, \tau_4^{(0.5)}, \tau_4^{(1)}, \tau_4^{(2)}, \tau_5^{(1)}, \tau_5^{(2)} \right\}$$

高次のQCD まで取り込まない
と計算できないやば目の量
を入れないと再現しない。

arXiv 1704.08249 Datta Larkoski



CNN との比較 Liam Moore et al 1807.04769



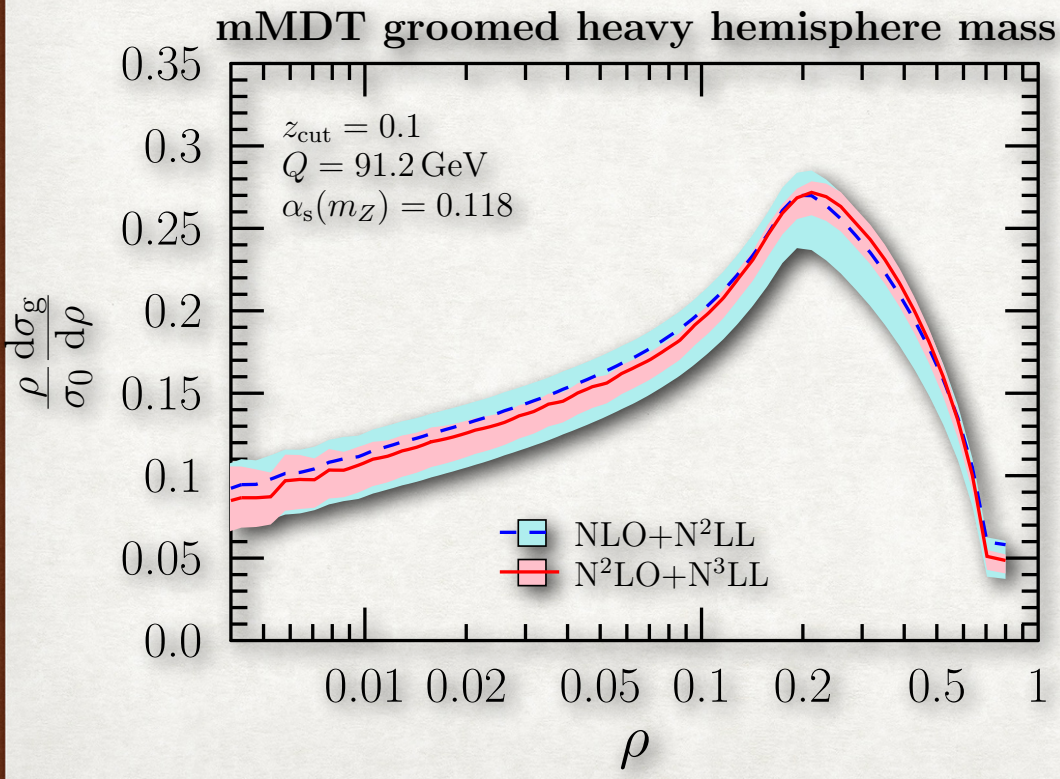
PREDICTABILITY

gloomed jet mass with NNLO+ N3LL resum
e+e -> hadrons (factorization を保証)

$$\frac{\min[E_i, E_j]}{E_i + E_j} > z_{\rm cut}$$

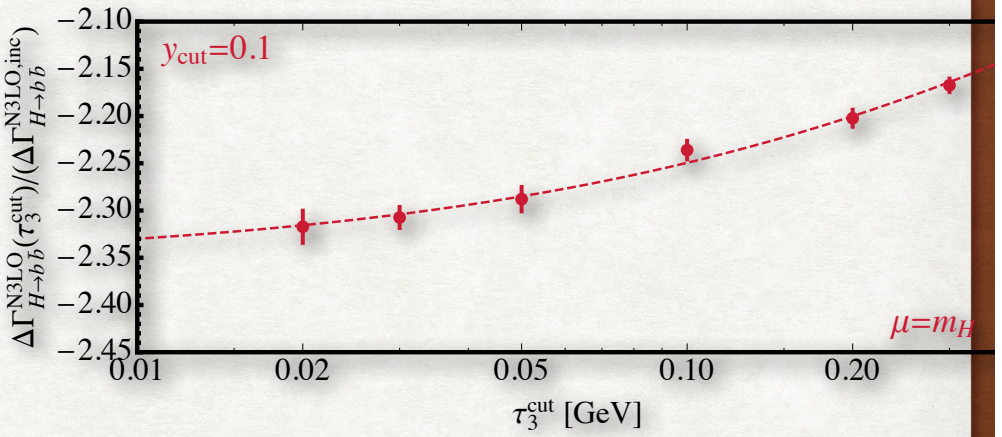
うまく高次効果を計算できる
フェーズスペースに特化した量を
計算する。

Kardos et al 2002.00942



Higgs の τ3 カットへの応答を見る

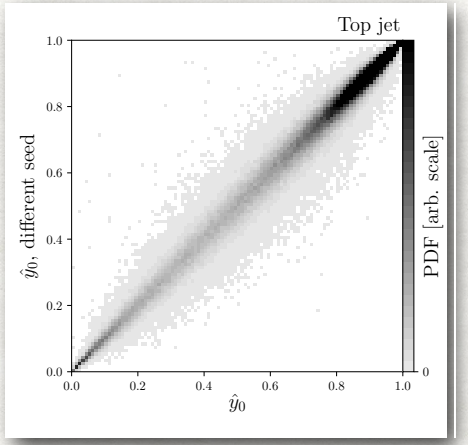
N3LO 計算 (Mondini et al 1904.08960

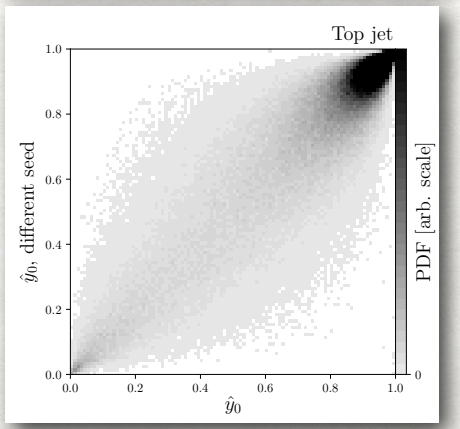


QCD との関係がつくのはいずれにしても
3 POINT くらいまでで、後は実データ
ベースで議論することになる。

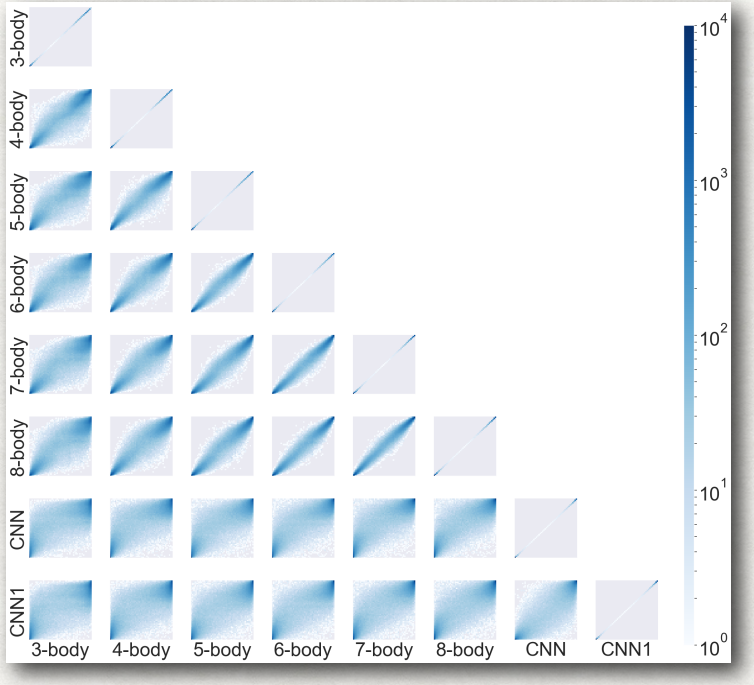
LOCAL MINIMUM 問題の改善

- CNN のloss function の最小化で、「真のminimum にたどり着くことはあまりない。ROC は安定しているが、個々のイベントに対して、違うseed で使ったclassifier は違う結果を出す。RN + MF は input が少ないのでevent ごとの結果も遙かに安定　(900-> 85)





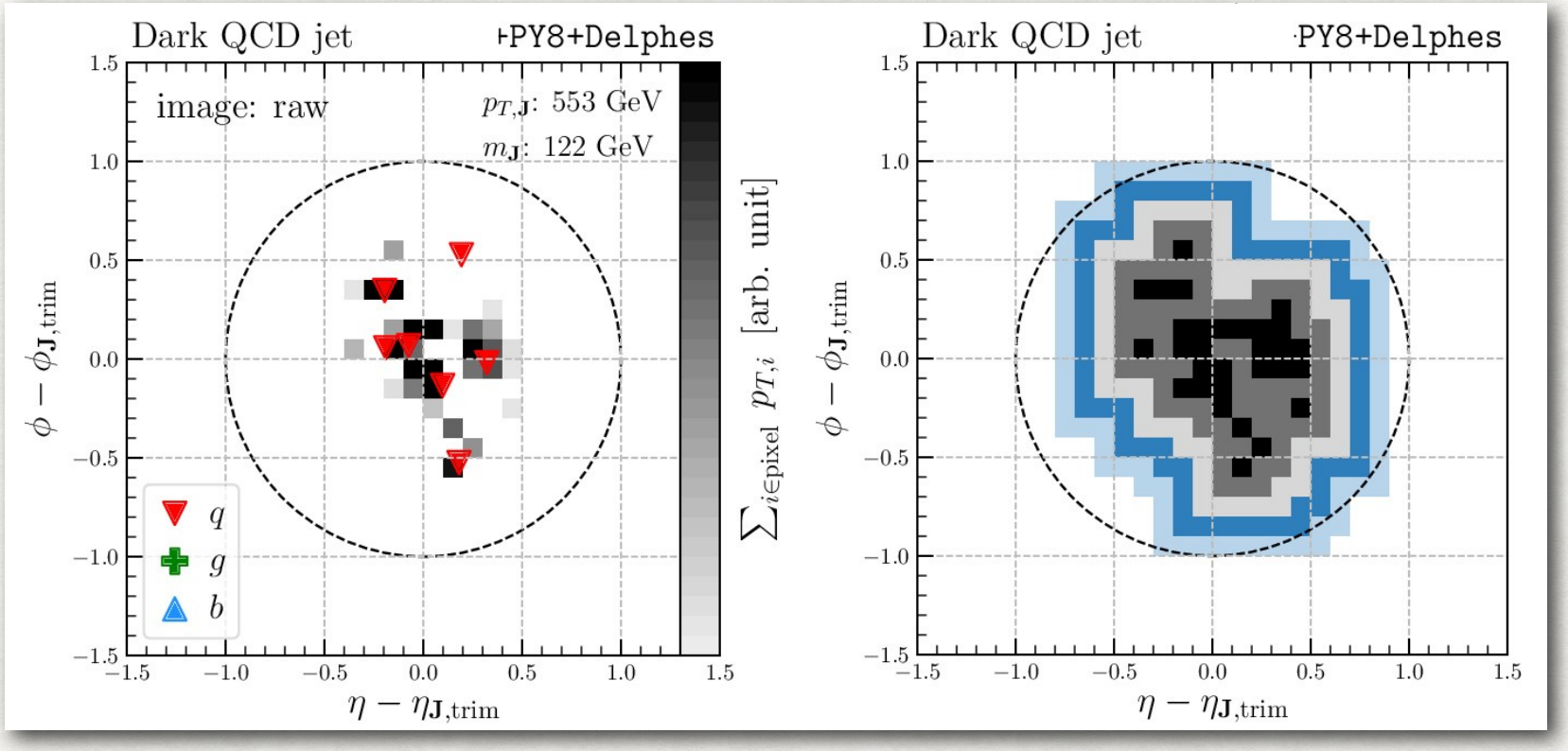
N-subjettiness の場合 1807.04769



DARK JET の場合

Lim, Nojiri　in preparation

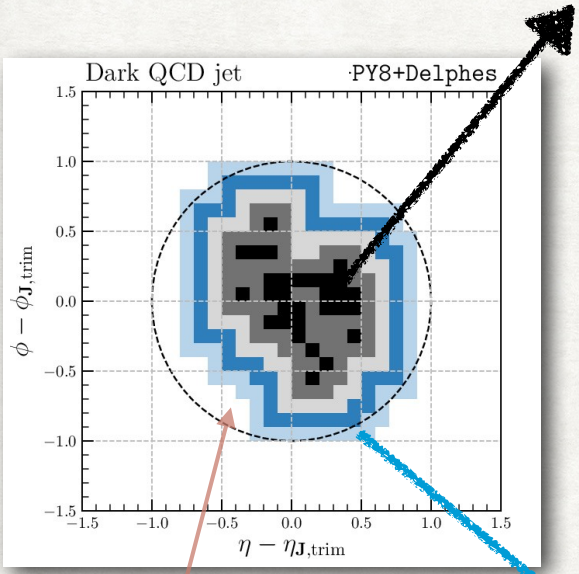
- Dark Jet　pp →Z′ → qD qD→ dark Parton shower → ρdiag→qq
- カラーシングレットなシャワー:粒子がたくさんあるが、いくつかのカラーシングレットなクラスターがある状態



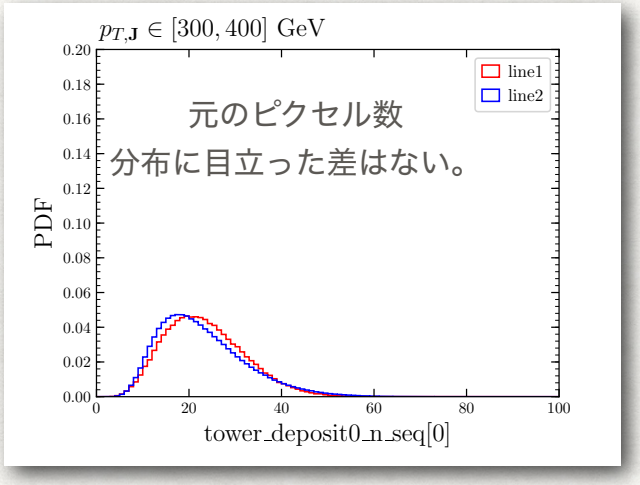
mρ=20GeV

CNN の学習結果

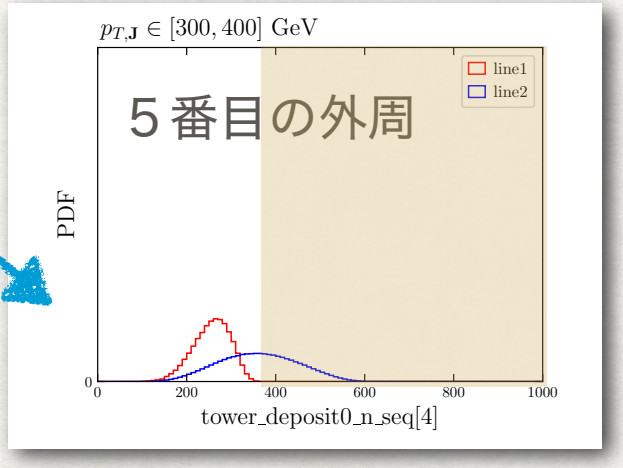
- mρ＝20GeV, 300GeV<pT<400GeV CNN のイベント選択は、MF(n>3) でカットをかけに行っていた。



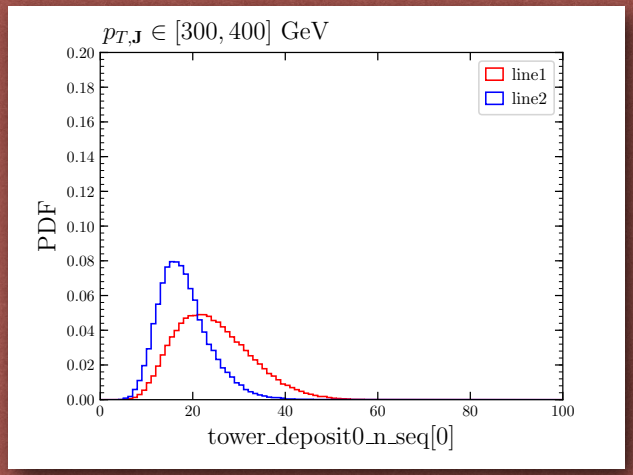
何もない空間に注目している。

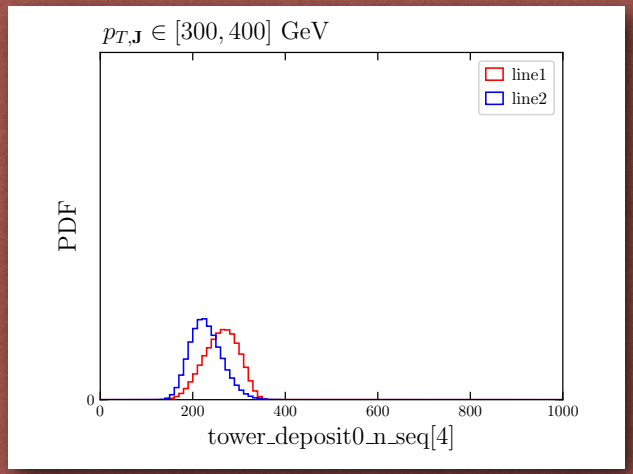


元の分布



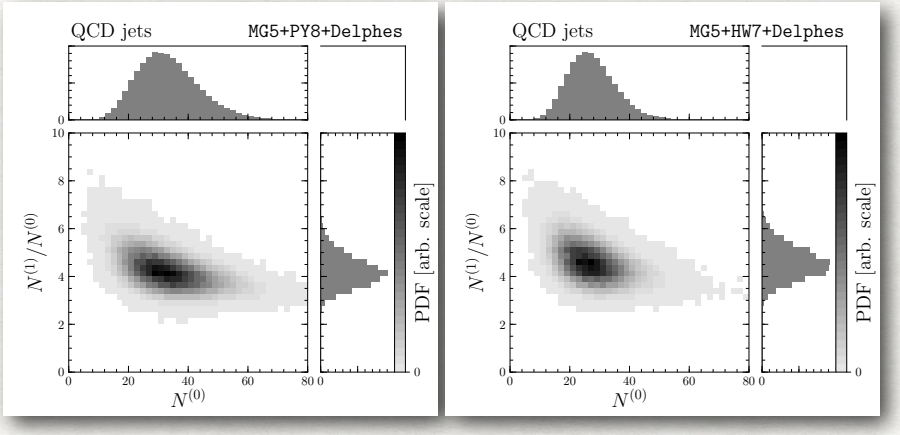
CNN でカット

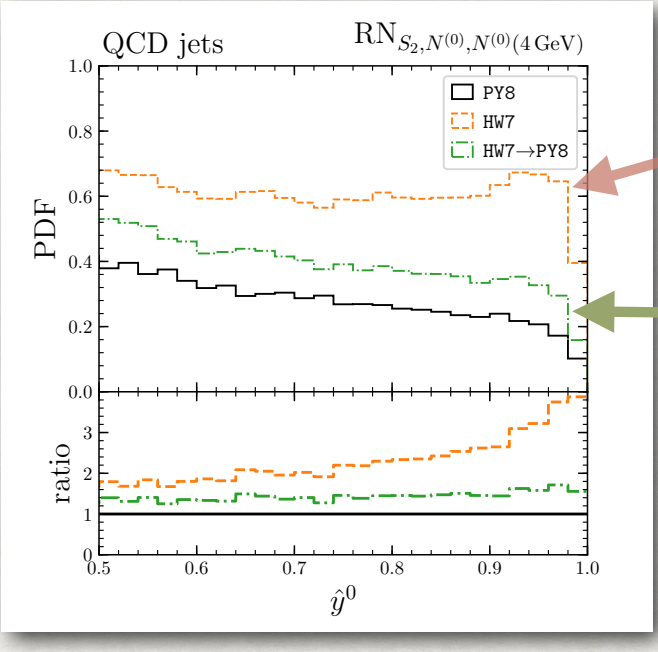




MC TURNING とかCALIBRATION とか

- Top jet vs QCD ジェットは event generator が違うと結構違う結果になる。
- 一番差にきいているのがQCDジェットの粒子数と広がり。結構違う、しかも実験データとも違う。
- [MC を実データで補正] MF の値が同じになるように、イベントにウエイトをつけると、一致がよくなる。





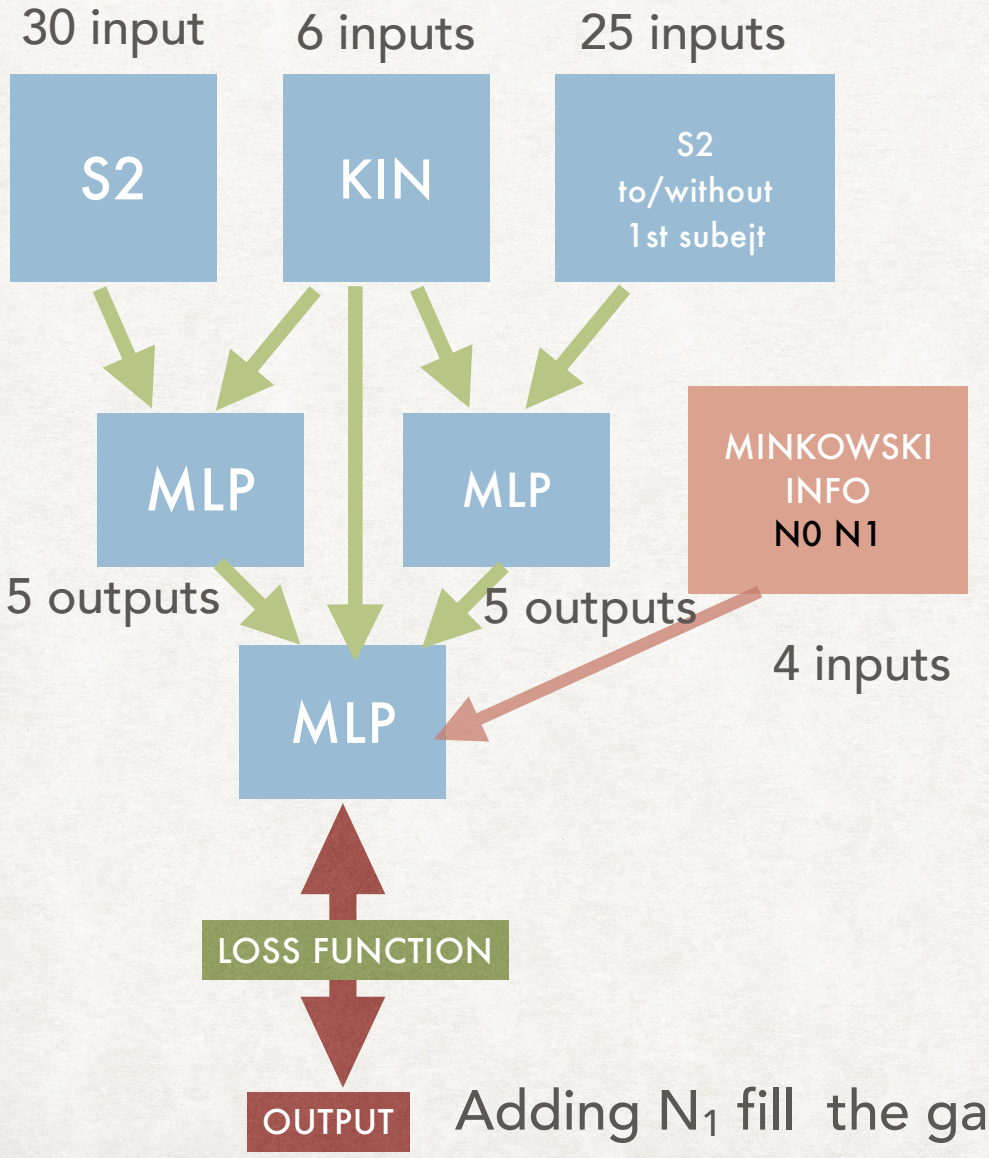
Phythia で　training した classifier で Herwig のQCD jet を分類

N0 分布を補正した後

教訓とやれそうなこと

- CNN などのジェットイメージを使った訓練はインフラの物理の違いを使って分類を強化している。
- pixel wise な情報じゃなくて、MF のような「まとめ指標」の方がキャリブレーションにも便利かもしれない。
- MF: カラー構造の違う粒子の性質を効率的に捉えているように見える。
- N-subjettiness などの従来の指標ともコンシステント

おまけ　NN のシステム



process
pp →tt vs pp→2j
500GeV<pT<600GeV
150GeV<mj<200GeV

case 1
modulation for two point correlation
two point correlation + Kin→5 outputs
correlation to/without leading jet
→ 5 outputs
→ROC

case2　+ N0(number of active pixel)
→ROC

case 3 + N0,N1→ROC

Adding N_1 fill the gap between CNN and our approach.