An Overview of the b-Tagging Algorithms in the CMS Offline Software

Christophe Saout

CERN, Karlsruhe Institute of Technology (KIT)

for the CMS experiment
on behalf of the b-Tag and Vertexing
Physics Objects Group

- Introduction
- CMS tracking system
- Input Objets
- Algorithms
- MVA Framework
- Conclusions

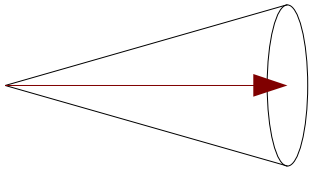
Introduction

Why b-tagging? Among list are discoveries involving Top. Higgs, SUSY...

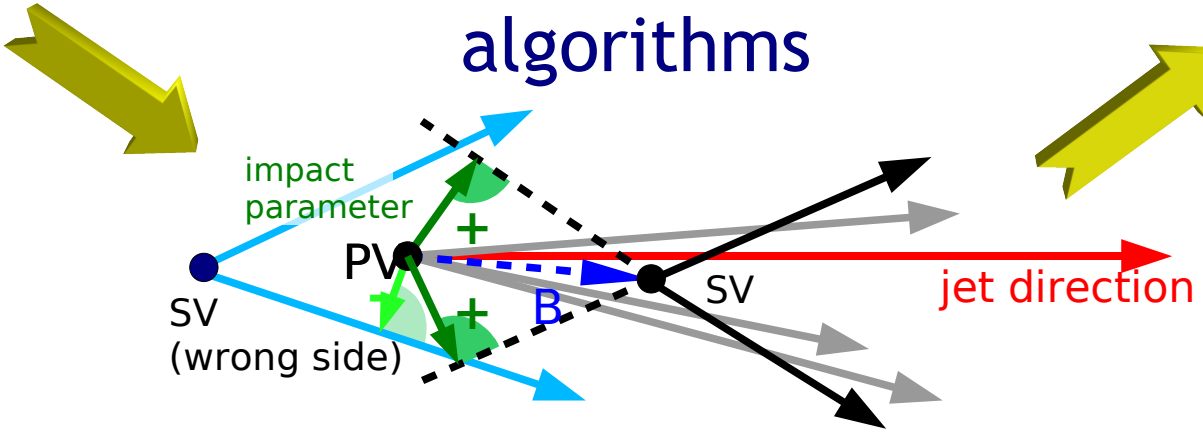
b-quarks significantly differ from light flavour quarks by:

- mass: m = 4.2 GeV
- lifetime: τ ≈ 1.5 ps → ~1.8mm *(at 20 GeV)* before decay
- decay: weak, mostly into c-quarks *(→ 3rd decay)* → 20% into leptons
- tracks: high decay multiplicity, significant displacement
- Secondary vertices *(SV)*: tracks intersecting at a common vertex

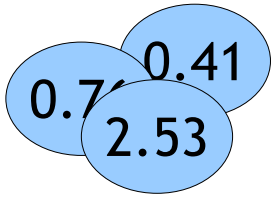
jets



algorithms

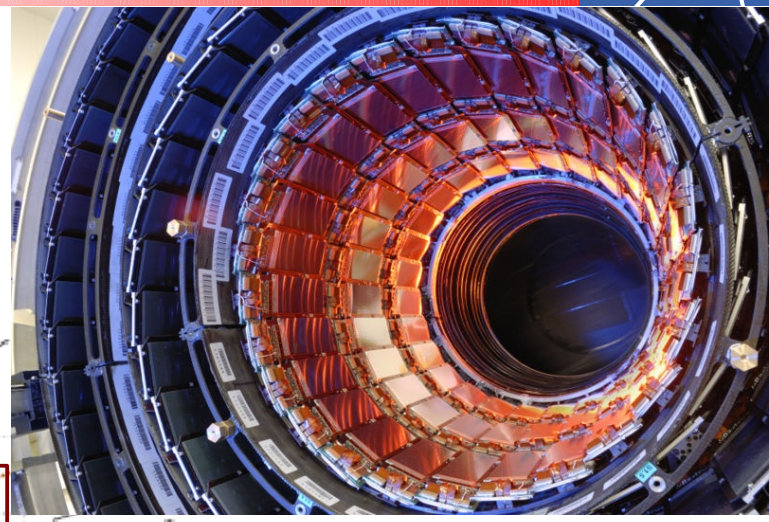


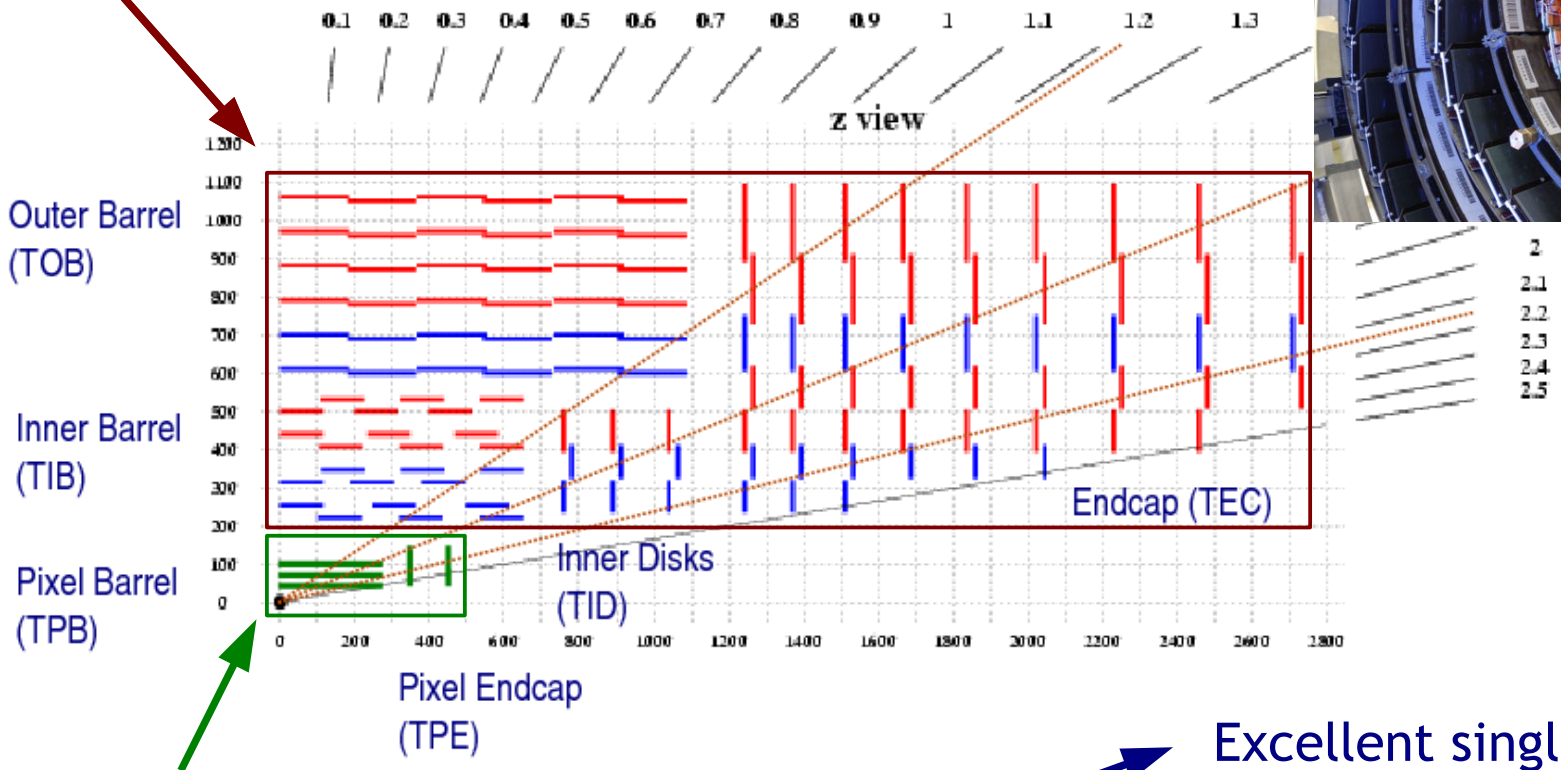
discriminators



The CMS Tracking System

10(*) layers of silicon strip detectors
- r-φ strip pitch of 80μm-180μm
- stereo layers: angle of 5.7°





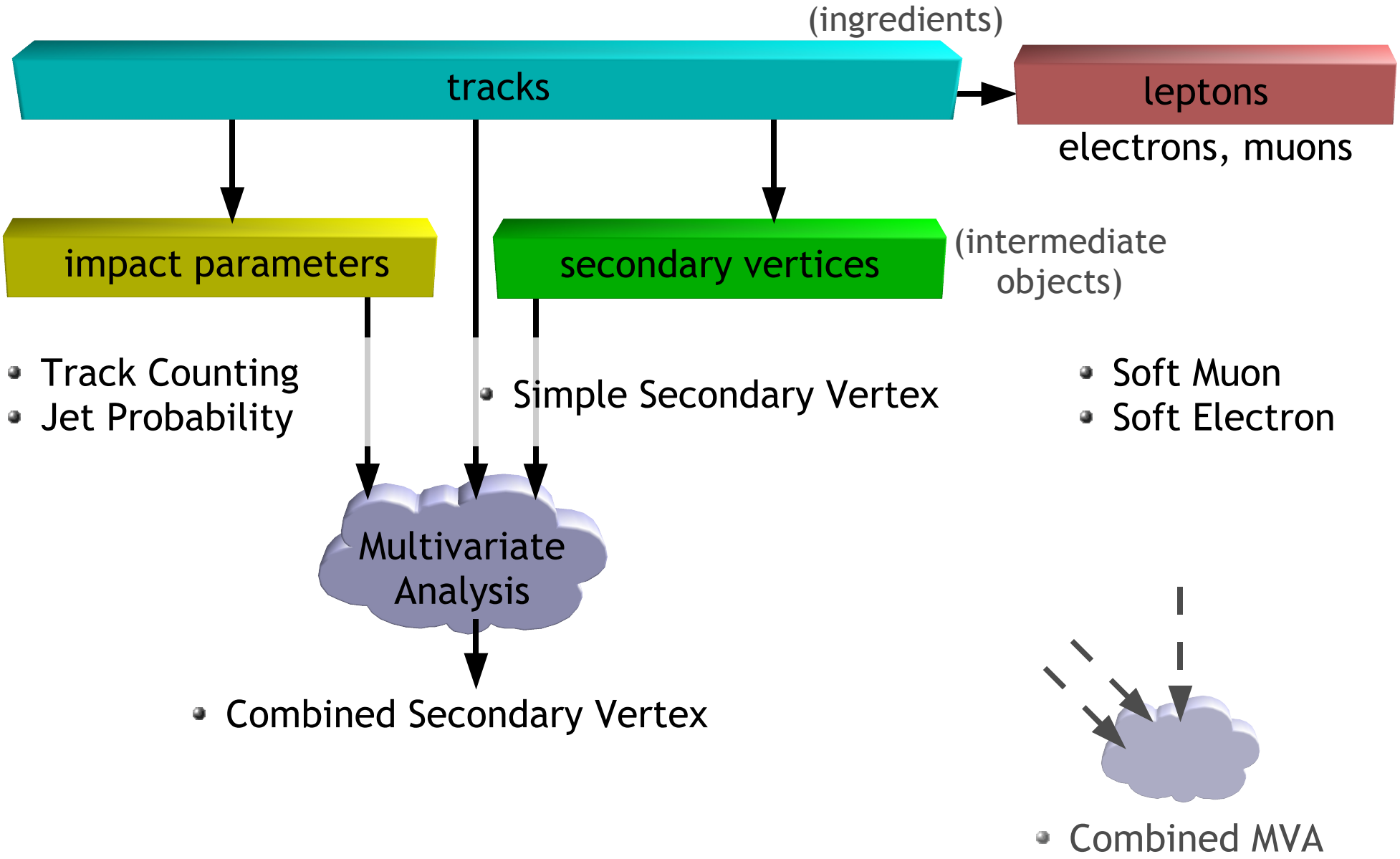
Three(*) layers of pixel detectors:
- 768 modules
- Inner ring at r = 4.4cm
- 100 μm × 150 μm pixel size

Excellent single-point resolution:
- 10μm in r-φ, 20μm in z

→ *good for b-tagging*

(*) in the central detector

Algorithm Structure



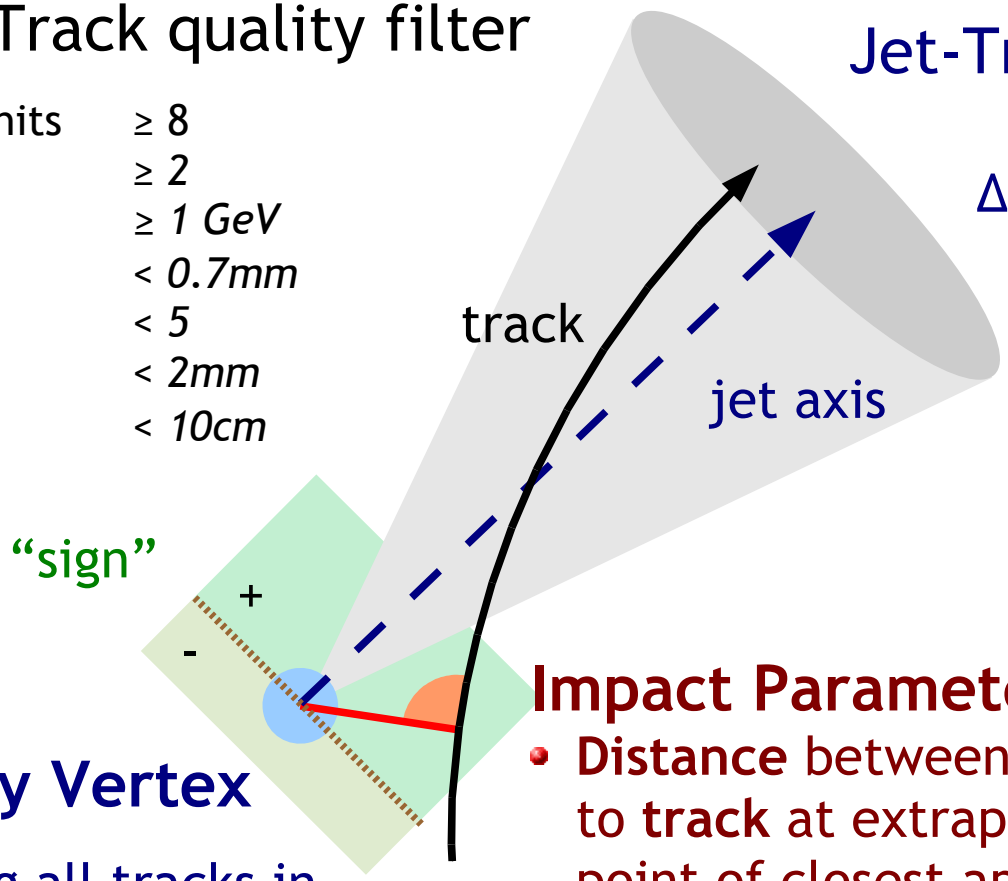
Impact Parameters

Track quality filter

total tracker hits	≥ 8
pixel hits	≥ 2
P_T	≥ *1 GeV*
jet axis dist.	< *0.7mm*
$x^2/ndof$	< *5*
IP_{xy}	< *2mm*
decay length	< *10cm*

Jet-Track Association

ΔR_{max} to jet axis:
0.5 or 0.3



Primary Vertex

reconstructed using all tracks in event using the

"Adaptive Vertex Fitter":

An iterative down-weighting Kalman vertex fit *(simulated annealing)*

Impact Parameter

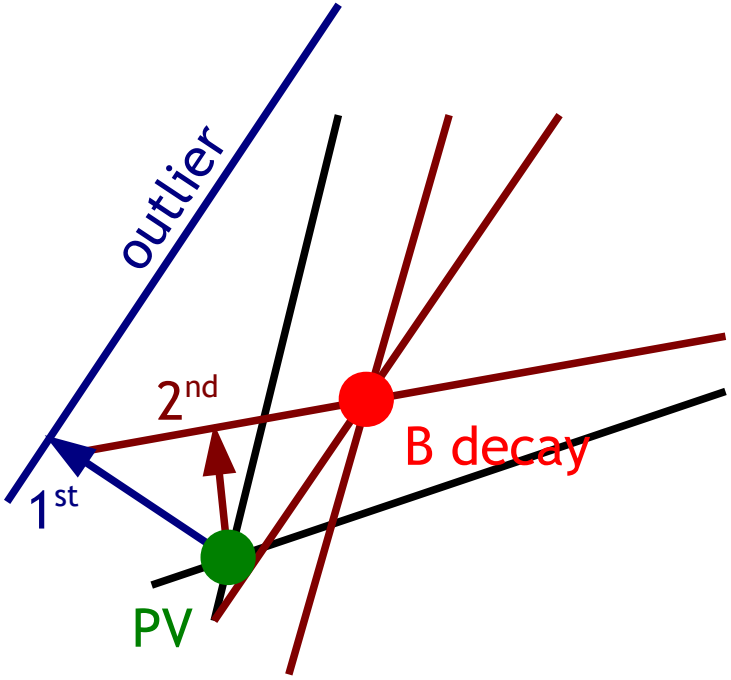
- **Distance** between **Primary Vertex** to **track** at extrapolated point of closest approach
- **Signed**
- Transverse **r-φ** or full **3D** value
- **Significance:** distance / error *using full PV fit and track extrapolation covariance matrices*

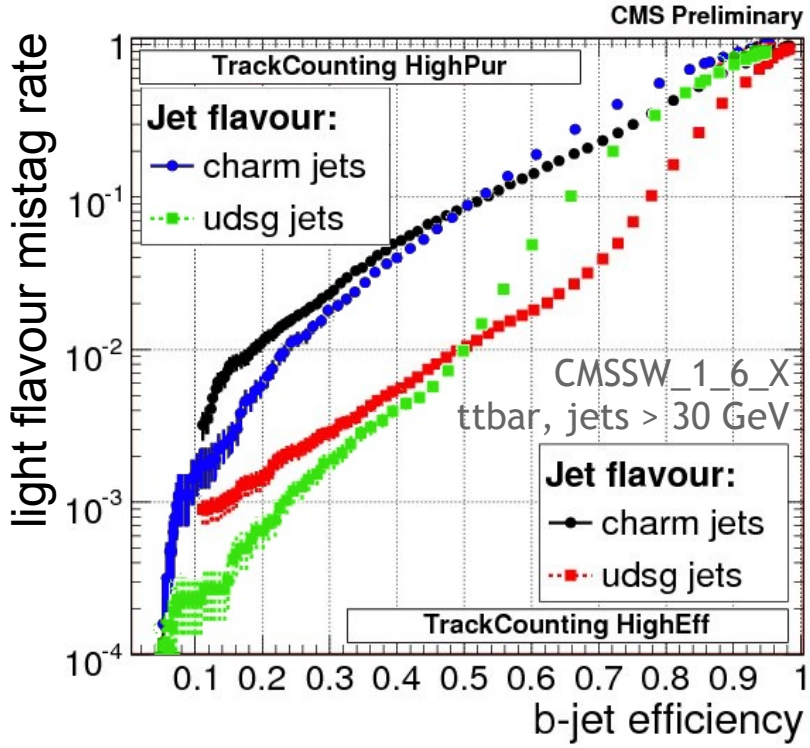
"Track Counting" algorithm

- Compute Impact Parameters for all tracks in jet
- Sort tracks by descending **Signed IP Significances** *(3D)*
- Select **nth track**
  - 2nd track → "high efficiency" tag
  - 3rd track → "high purity" tag
- Use IP significance as **discriminator**
- Simple, fast → suitable for HLT

simple & suitable for early data

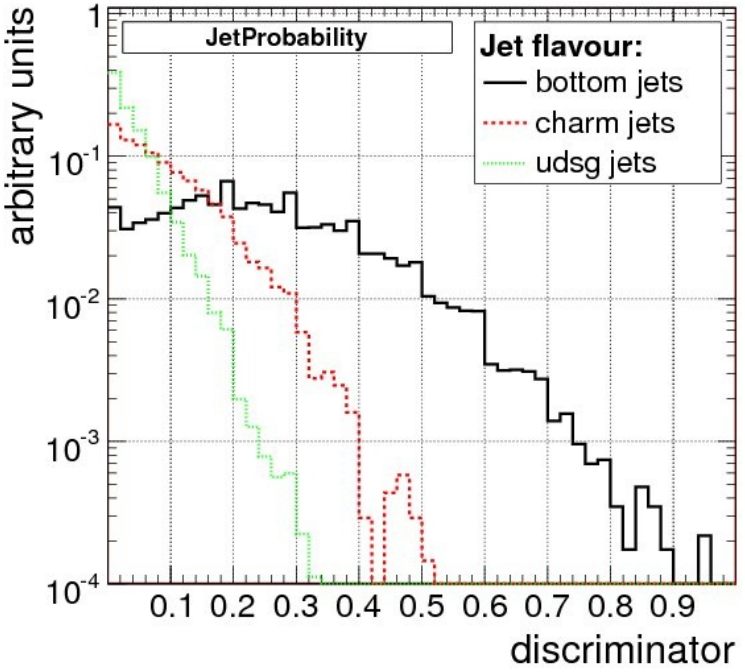
→ Eliminate non-b decay outliers
- Fake tracks
- V0 decays
- …





“Jet Probability” algorithm

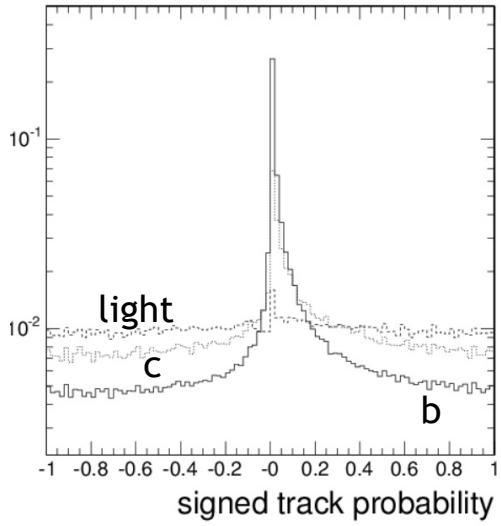
- Used at LEP, originally from ALEPH
- Compute “track probabilities” for each track
  - Probability for the track to originate from PV
  - PDFs for Impact Parameter Significance
  - divided in track quality categories
    - #hits total
    - #hits in pixel detector
    - Valid hit in first pixel layer
    - track pseudo-rapidity
    - track momentum
    - Track fit χ^2



- Compute total “jet probability” that all tracks originate from PV

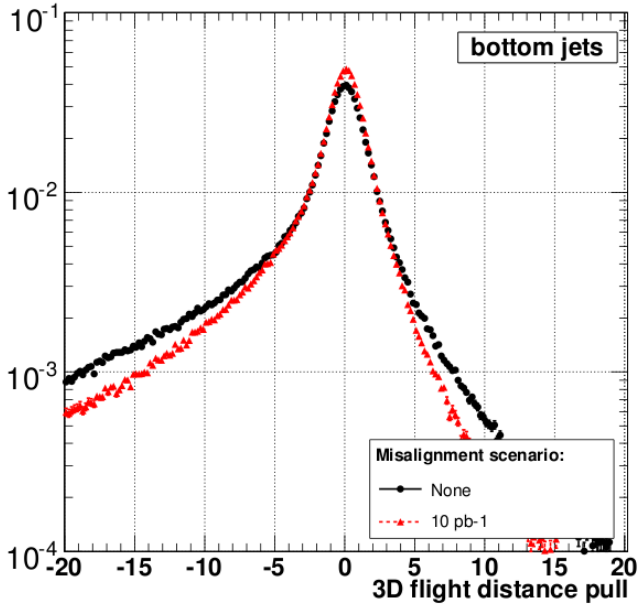
$$P_{jet} = \Pi \cdot \sum_{j=0}^{N-1} \frac{-\ln \Pi^{J}}{j!} \quad \text{with} \quad \Pi = \prod_{i=0}^{N} \tilde{P}_{tr}(i) \quad \text{and} \quad discr = -\log(\Pi)$$

- By default use only positive signed IP
- Can be calibrated from data using negative-side IP
- Variant giving more weight to 4 most b-like tracks: “Jet B Probability”



Secondary Vertices

- Inclusive vertex reconstruction in a jet
- Using the "Adaptive Vertex Reconstructor":
  - Iterative approach starting from all tracks:
  - Attempt to fit a vertex using the "Adaptive Vertex Finder"
    → will head for "best" vertex and downweight incompatible tracks
  - Repeat with tracks excluded from fit until track exhausted
- Check vertex compatibility with Primary Vertex
  - Cut on PV-SV distance and significance *(0.1mm < dxy < 2.5cm, dxy/σ > 3)*
  - Not more than 65% tracks shared with Primary Vertex
  - Maximum vertex mass of 6.5 GeV
  - Invariant mass window around K_S rejected
  - Vertex in jet direction *(ΔR < 0.5)*
- Vertex finding rate (*):
  - b-jets: 63% *(latest software ~70%)*
  - c-jets: 22%
  - Light: 2.7%



(*) CMSSW_1_6_X ttbar, jets > 30 GeV

"Simple Secondary Vertex"

simple & suitable for early data

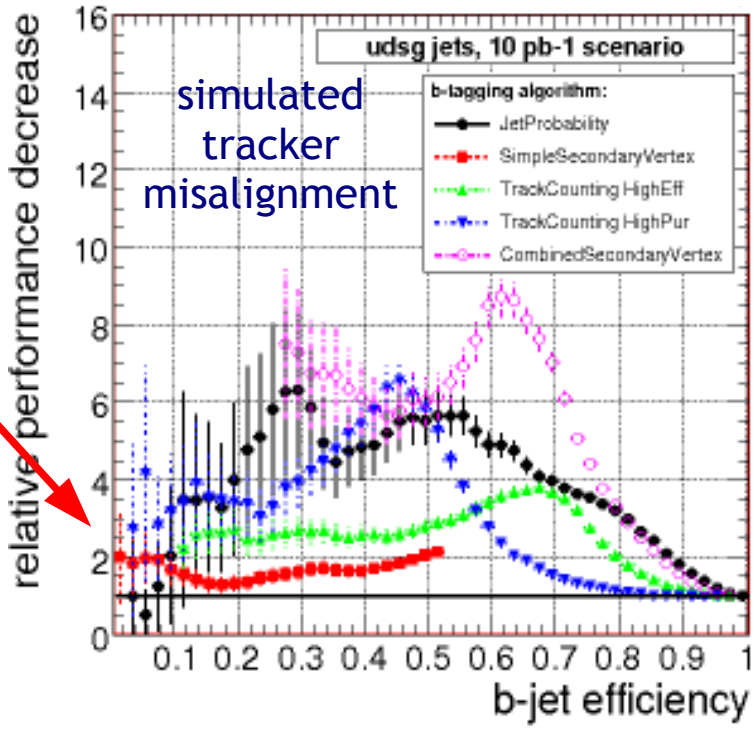
- Uses presence of a reconstructed Secondary Vertex as b-tag
- Use flight distance measurement as discriminator
  - In transverse plane or 3D
  - Distance PV-SV or its significance *(value/error)*

(defaults underlined)

- Will give no discriminator without reconstructed SV
  → *b-tagging efficiency limited to vertex finding efficiency*
  → *can be used as a yes/no tag*
- Most "robust" algorithm,
  least sensitive to detector alignment

  (CDF is still actively using the similar "SVX" tag)

- Performance comparable to the "track counting" algorithms
- Allows to define a "negative vertex tag" for purposes of mistag measurement



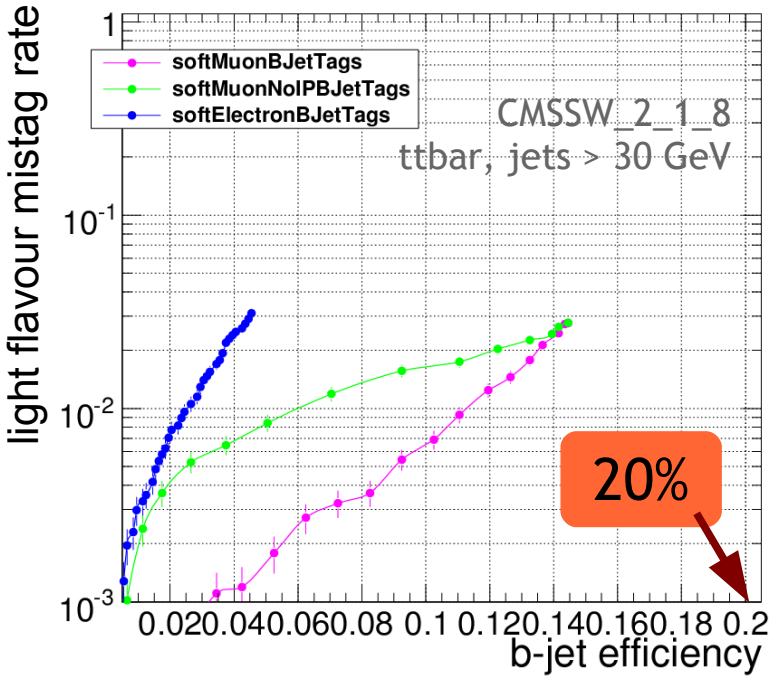
“Soft Lepton” algorithms

- In ~20% of the b-jets one gets a lepton from the weak decay
- Needs leptons in jets, not isolated ones!

robust & suitable for early data

- For muons:
  - Muon reco and ID *unproblematic* with the CMS standalone muon system
- For electrons:
  - Cannot use default electron reconstruction *(because of isolation)*
  - Using a dedicated in-jet electron ID (which is being worked on)
- Default algorithms use a simple feed-forward MLP *(neural network)* to compute the discriminator:
  - $p_{T\,rel}$ wrt. jet axis
  - ΔR wrt. jet axis
  - relative lepton momentum
  - signed IP significance
  - lepton quality
- Simple and robust variants for early data

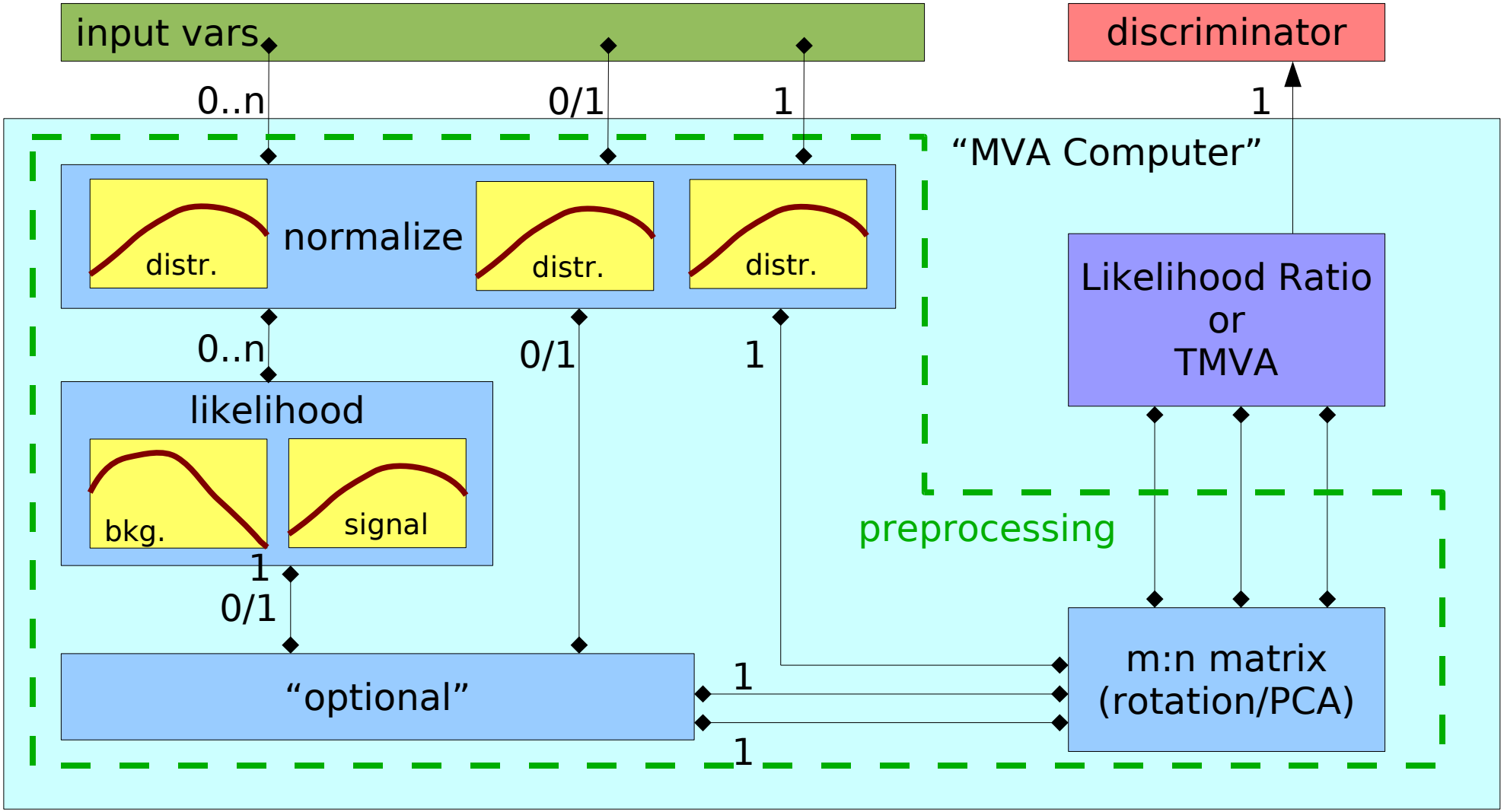
  e.g. muon $p_{T\,rel}$ tagger



(CMSSW) MVA Framework

- Modularized interface to Multivariate Analysis Techniques within the CMS software framework
- Especially designed with reconstruction software needs in mind
  - Native storage of training data in the CMS Conditions Database
    (allows live access to central run-dependent conditions over the Internet)
  - Fully compatible with the CMS "Event Data Model"
  - ***Small footprint:*** Evaluating networks is very resource-friendly
- Can deal with varying number of variables!
  e.g. per track-variables in b-tagging or missing secondary vertex variables
- Unlimited user-definable stacking of modules
- Many out-of-the box modules for common reco tasks
  - Variable preprocessors *(normalization, linear decorrelation)*
  - Classic Likelihood ratio, Fisher's Discriminant
  - User-definable categorized PDF histogramming
  - Variable counting, splitting, sorting, …
- Interface to powerful third-party MVA packages, e.g. ROOT ***TMVA***

MVA Layout Example



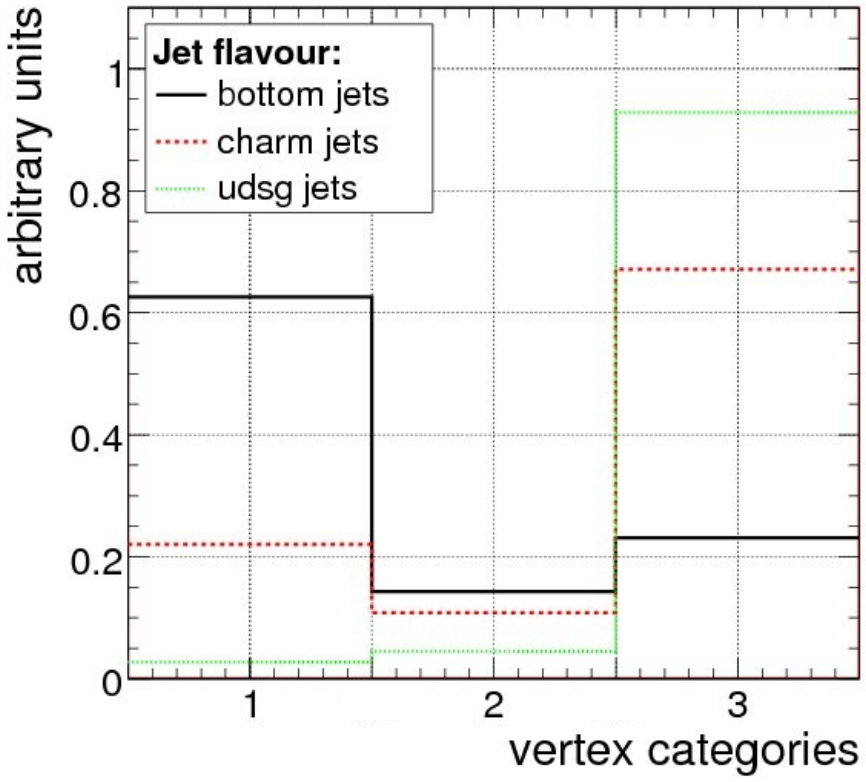
more complex example for “CombinedSV” b-tagger with a more advanced MVA

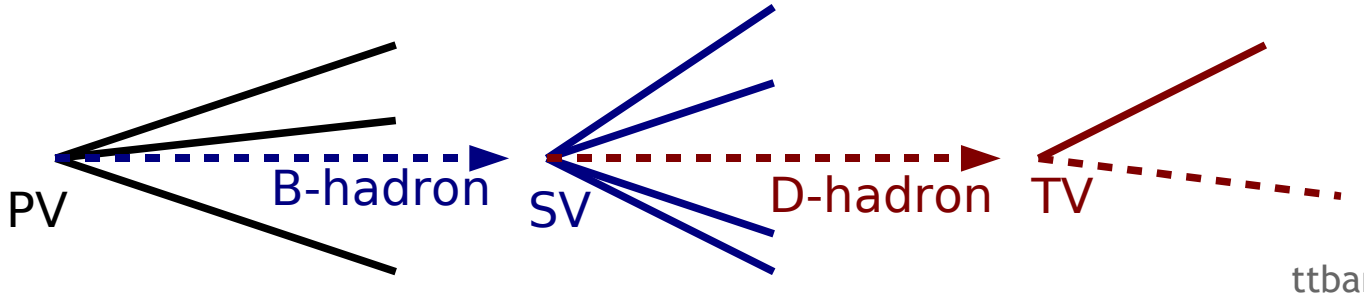
User-definable using an MVA layout description defined in XML

“Combined Secondary Vertex”

- Combines all information that can be gotten out of tracks
  → *impact parameters and vertices*

- Defines three vertex categories:
  1. “RecoVertex”:
     at least one good Secondary Vertex
  2. “PseudoVertex”:
     at least 2 track with IP/σ > 2
     (attempts to catch cases where b and c decay yield one track each)
  3. “NoVertex”:
     remaining cases



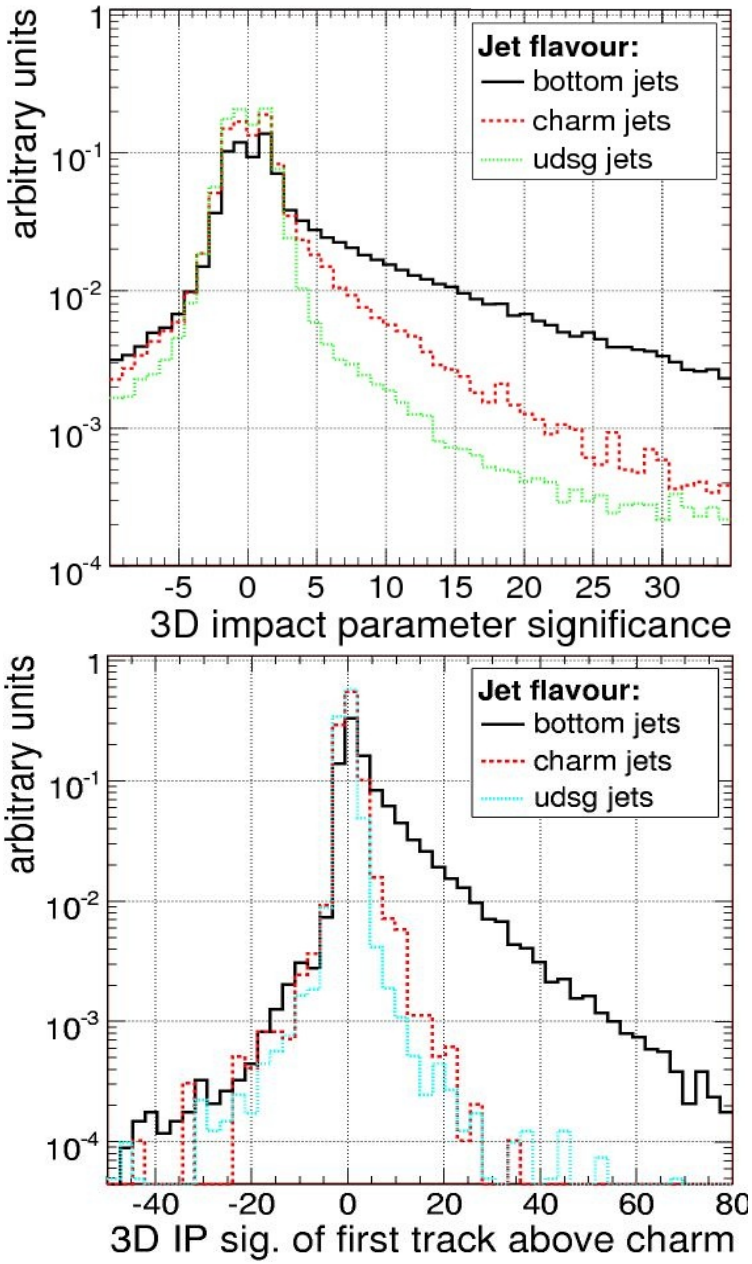


CMSSW_2_1_8
ttbar, jets > 30 GeV

"Combined Secondary Vertex"

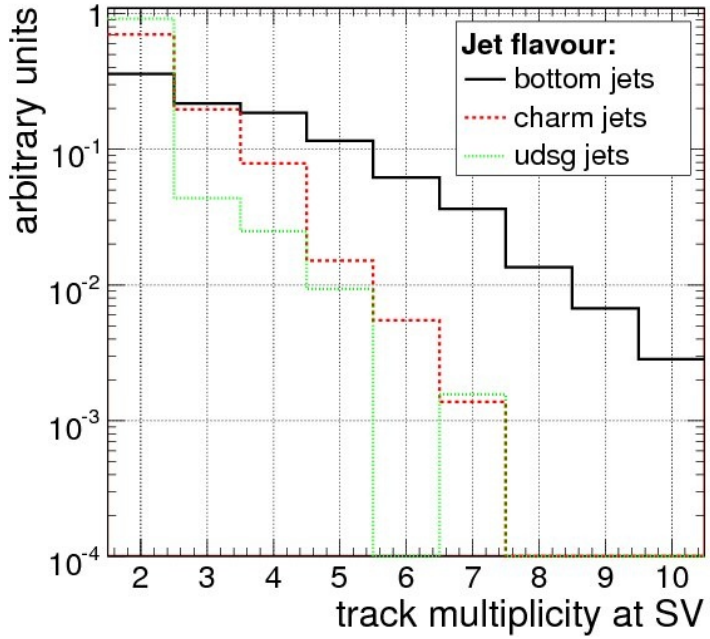
- Track Variables:
  - 3D signed IP significances
    (corresponds to variables used by "track counting" and "jet probability")
  - 3D signed IP significance of first track lifting the invariant mass above 1.5 GeV
    (iteratively adding tracks with highest IP/σ)
    → good b/c discrimination
  - With a secondary or pseudo vertex:
    Rapdities of SV tracks along jet axis

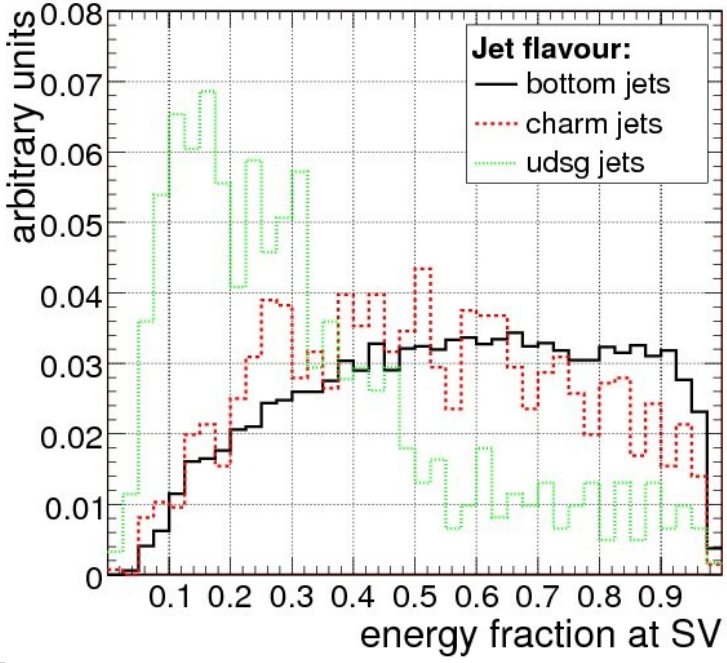
$$y = \frac{1}{2} \cdot \ln \frac{E + p_{par}}{E - p_{par}}$$

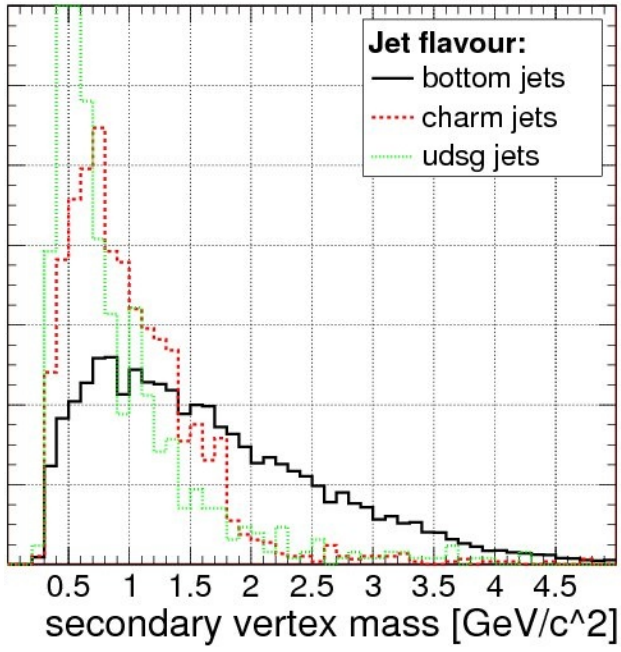


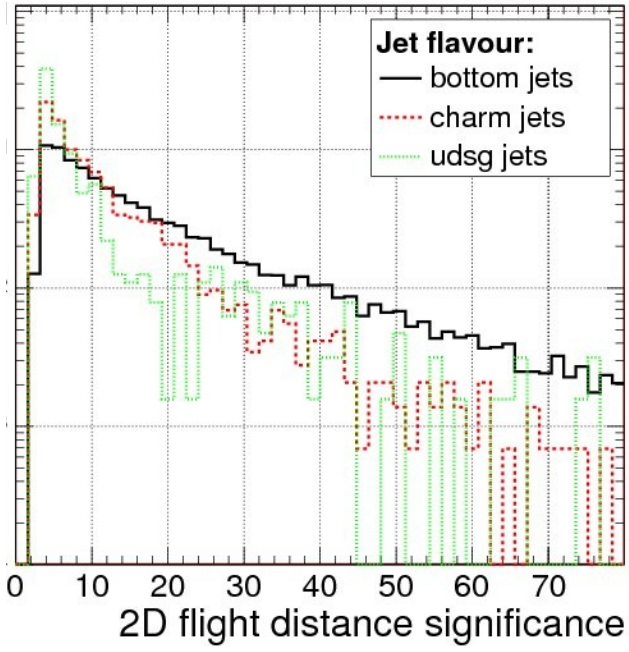
“Combined Secondary Vertex”

- Secondary/Pseudo Vertex Variables
  - 2D Flight Distance Significance
  - Invariant SV Mass
  - Fractional charged energy at SV
  - Track Multiplicity at SV
  - ΔR between SV direction and jet axis









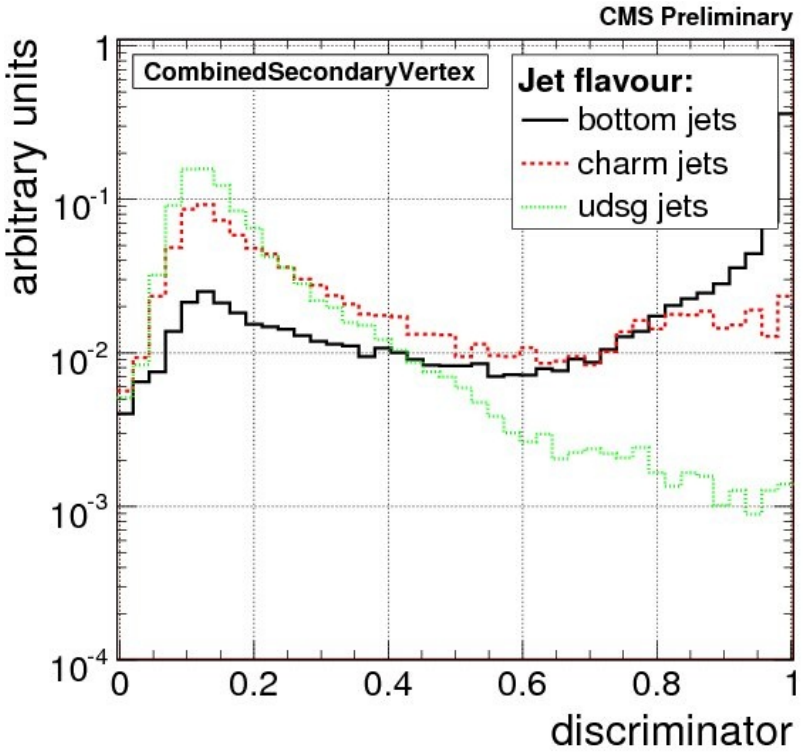
“Combined Secondary Vertex”

- Final discriminator is built as a likelihood ratio from all input variables

$$L^{b,c,q} = f^{b,c,q}(\alpha) \times \prod_i f_\alpha^{b,c,q}(x_i)$$

$$y = f_{BG}(c) \times \frac{L^b}{L^b + L^c} + f_{BG}(q) \times \frac{L^b}{L^b + L^q}$$

b ↔ c b ↔ udsg



- $f_{BG}(c)$: prior for charm content in non-b jets (default chosen from ttbar → 0.25)
  $f^{b,c,q}(\alpha)$: probability for flavour q to be in category α
  $f_\alpha^{b,c,q}(x_i)$: PDF of variable x_i for category α and flavour q (parametrized in bins of jet p_T and η)
- Full discriminator computation directly implemented using MVA framework directly on input variables
- Variant employing a neural network for the “RecoVertex” case instead of the likelihood ratio → small gain in b-efficiency

“Combined MVA” algorithm

- The Combined Secondary Vertex is the best-performaning b-tagger so far
- For leptonic b-decays the reconstruction only sees a displaced track
- By adding soft lepton information *(i.e. the lepton ID)* in addition one should be able to additionally gain some b-tagging efficiency

- Two possibilities:
  - Write a tagger using all input variables *(tracks, vertices, leptons)*
  - Combine already well-optimized algorithm outputs

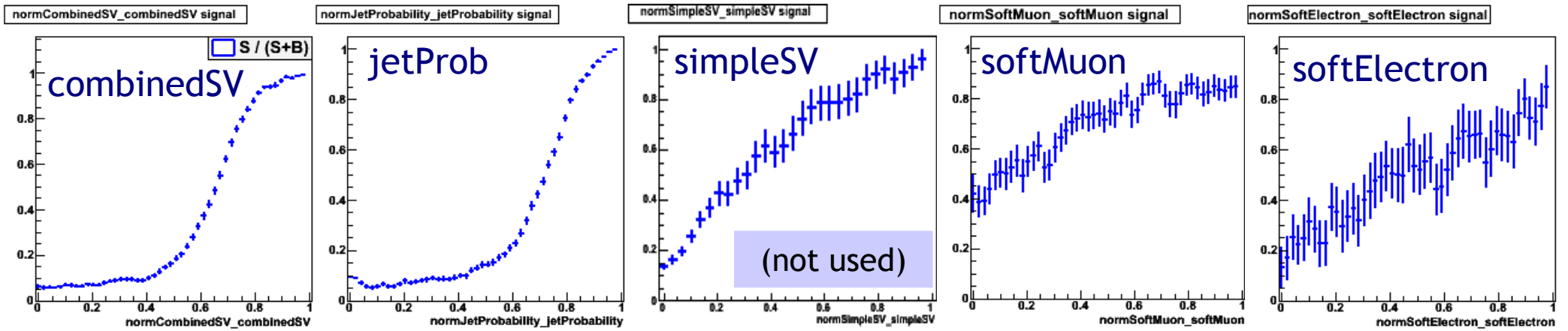
  → *currently implemented for demonstration purposes*

  - Combines discriminator outputs
  - In order to train only needs knowledge about
    - Discriminator distributions for b-jets and background
    - Correlations between algorithm output

      → needs only well-understood tagger output

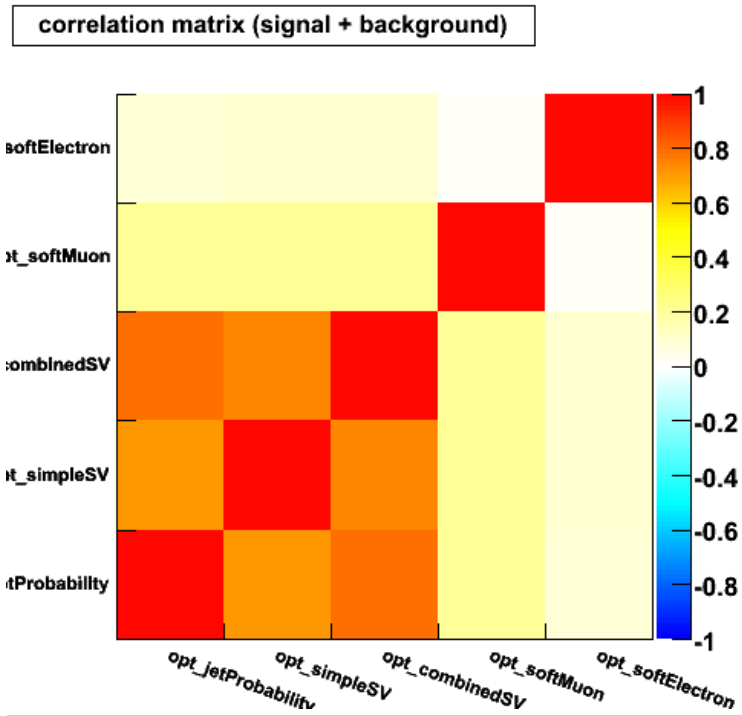
      (no need to understand all individual input variables)

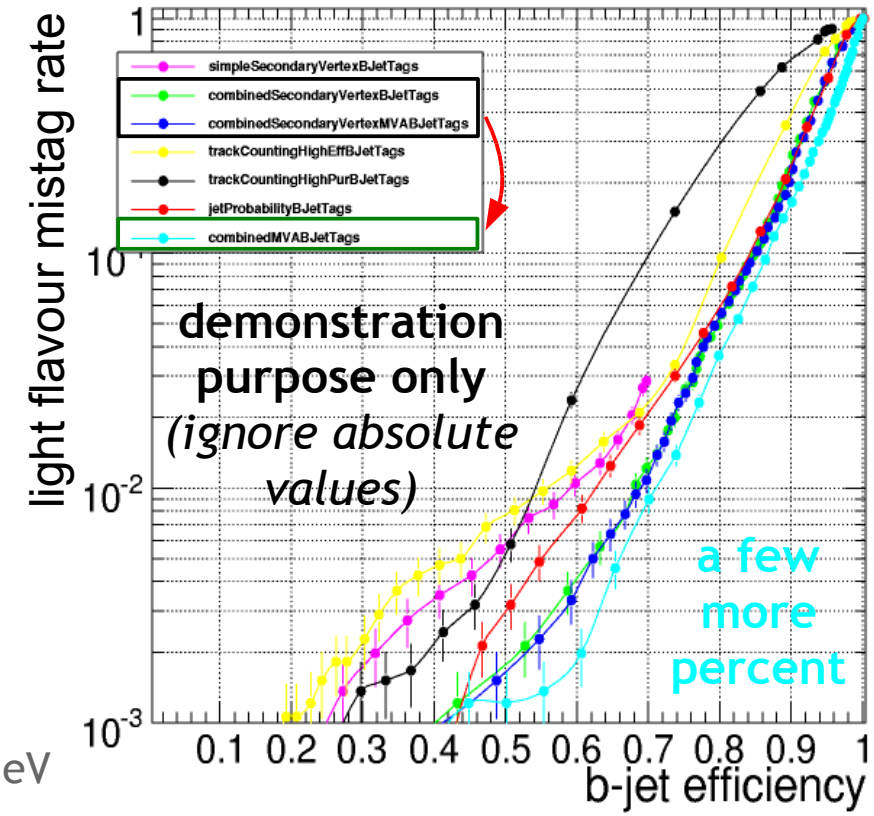
“Combined MVA” algorithm



Correlations of normalized input variables to target

S / (S+B) → if discriminator was ≥ 0





CMSSW_2_0_X
ttbar, jets > 30 GeV

Conclusions

- The CMS offline software has a wide variety of algorithms
  - Simple and fast ones → suitable for HLT
  - Simple and robust ones → suitable for early data
  - Algorithms suitable for efficiency and mistag measurements from data
  - Orthogonal algorithms *(lifetime / leptons)*
  - Algorithms trainable from data
  - High-performing algorithms for later
  - Multivariate analysis techniques for highest-possible performance

    → *everything in good shape for data-taking*

See poster from Victor E. Bazterra

- Will hopefully be able to commission first b-tagging algorithms early
  - b-Tagging depends on many subsystems *(especially tracker alignment)*
  - Data-driven techniques for efficiency/mistag measurements in place
  - And then we hope we will …

… make discoveries with b-jet final states!

References

- ”The CMS Physics Technical Design Report, Volume 1,” 2006
  CMS Collaboration
  Chapter 6: inner tracking system, Chapter 12.2: b-tagging
- CMS NOTE-2007/008: “Adaptive Vertex Fitting”,
  R.Fruehwirth, W.Waltenberger, P.Vanlaer
- CMS NOTE-2006/019: “Track impact parameter based b-tagging with CMS”,
  A.Rizzi, F.Palla, G.Segneri
- CMS NOTE-2006/014:
  “A Combined Secondary Vertex Based B-Tagging Algorithm”, C.Weiser
- CMS NOTE-2006/043: “Tagging b jets with electrons and muons at CMS”,
  A.Bocci, P.Demin, R.Ranieri, S.de Visscher
- CMS PAS BTV-07-003: “Effect of misalignment on b-tagging”, 2007
  CMS Collaboration
- “TMVA – Toolkit for Multivariate Data Analysis”, 2007
  A.Hoecker, P.Speckmayer, J.Stelzer, F.Tegenfeldt, H.Voss, A.Christov, S.Henrot-Versille, M.Jachowski, A.Krasznahorkay Jr., Y.Mahalalel, R.Ospanov, X.Prudent, M.Wolter, A.Zemla