Boosted ggH → cc @CMS

YSF: Search for boosted Higgs boson decays to charm quark pairs

European Research Council
Established by the European Commission

Andrzej Novak on behalf of the CMS Collaboration

arXiv:2207.00043

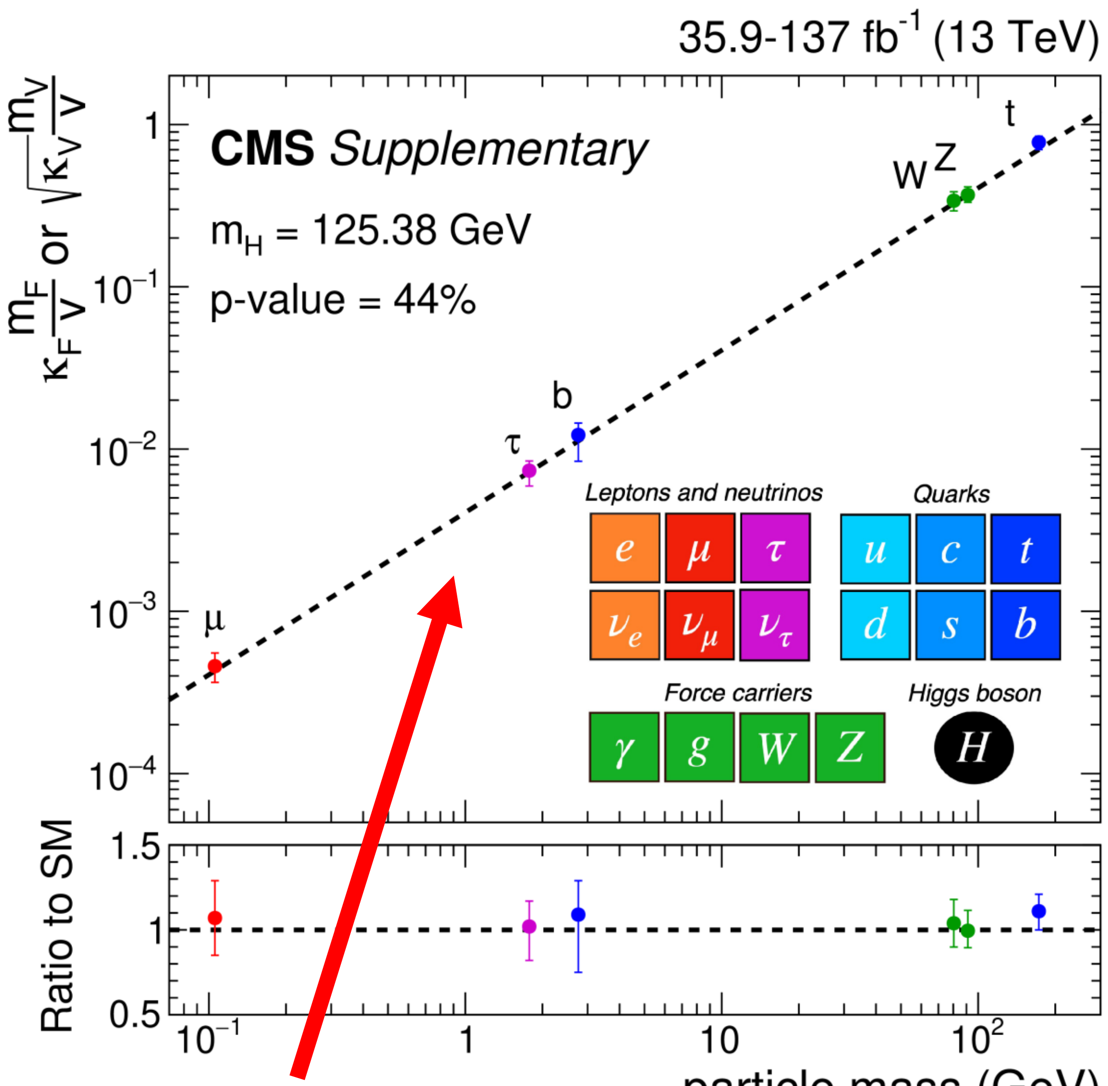
Motivation

Complete description of the Higgs boson

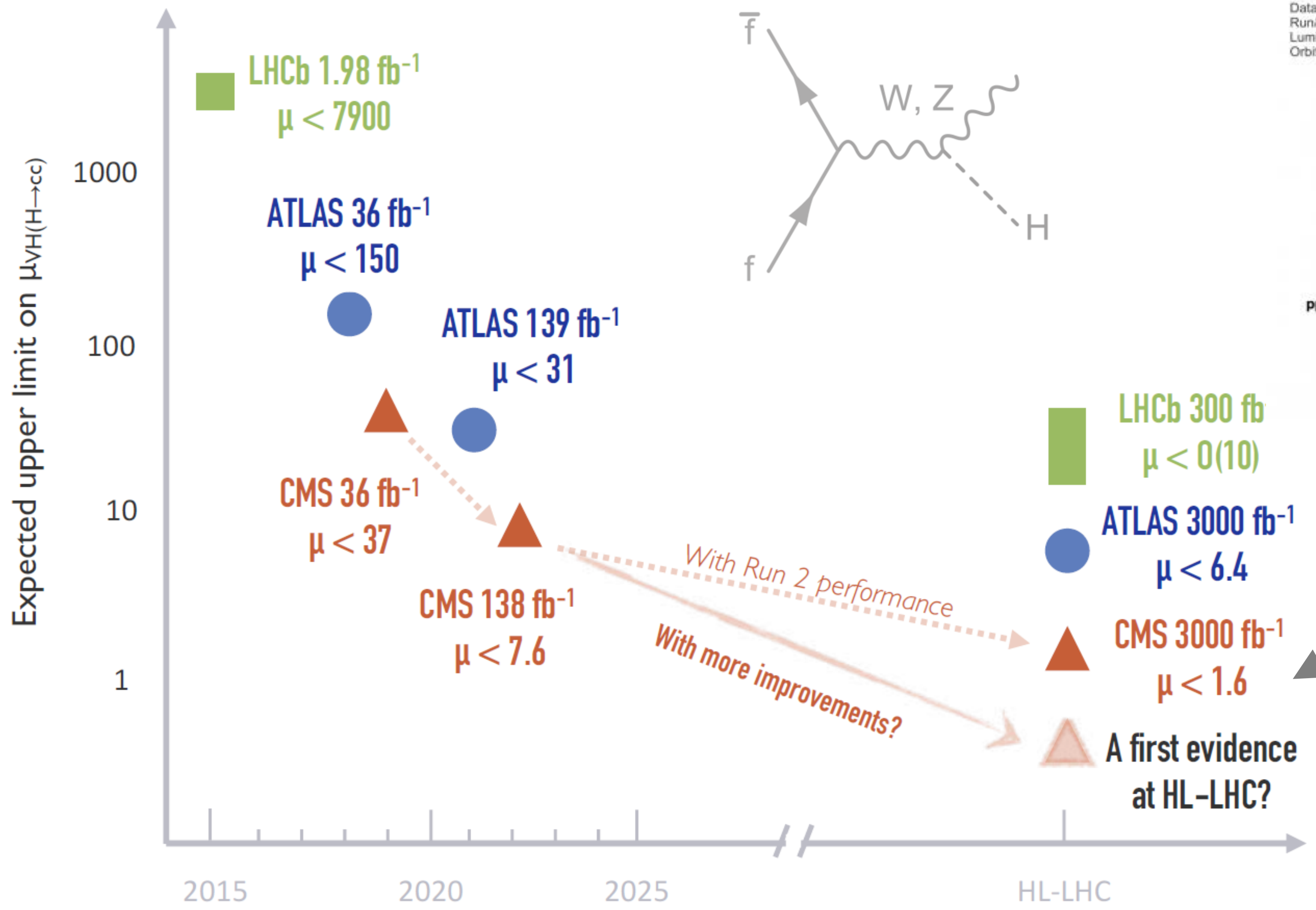
- Observed couplings (~90% of BR)
  - Vector bosons (WW, ZZ)
  - 3rd generation fermions (ttH, bb, ττ)
- Evidence for 2nd generation couplings
  - H → μμ

Next target – search for **H → cc**

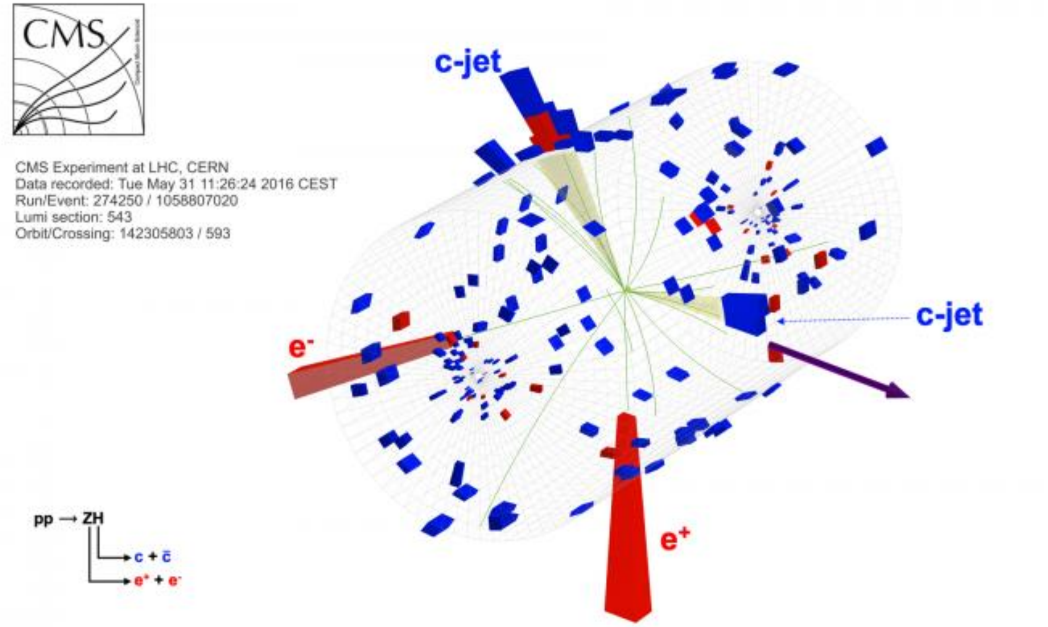
- Largest missing fraction of BR
- Establishing couplings to 2nd gen. quarks
- Measuring any enhancements to coupling would hint at new physics



VH → cc Searches



H. Qu & L. Mastrolorenzo, LHC Seminar 2022

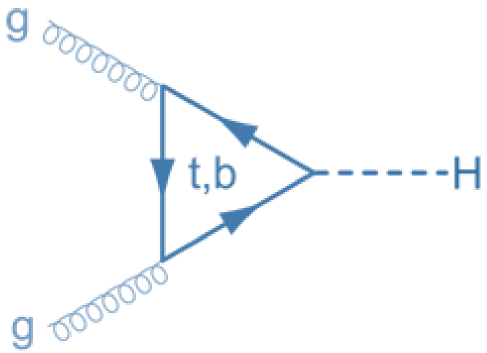


Not quite good enough

Need to improve sensitivity and/or combine with other channels

Direct Search Channels

Gluon Fusion



x 20

Premise:

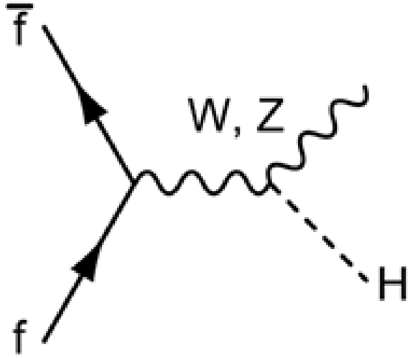
Largest Higgs production cross-section

- σ(pp → H)@13 TeV = **48.6 pb**

Very large QCD contamination: g → qq

- **σ**(pp → qq)@13 TeV ~ **360 000 pb** (HT > 300 GeV)

Higgs-Strahlung



Premise:

Lower cross-section

- σ(pp → VH)@13TeV = **2.25 pb**

Clean signature: trigger on leptons from Z,W

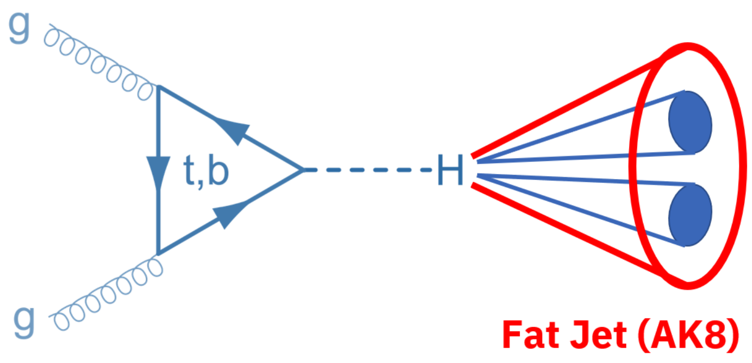
Mostly V+jets background

- σ(pp → V+qq)@13TeV ~ **600 pb** (HT > 400 GeV)

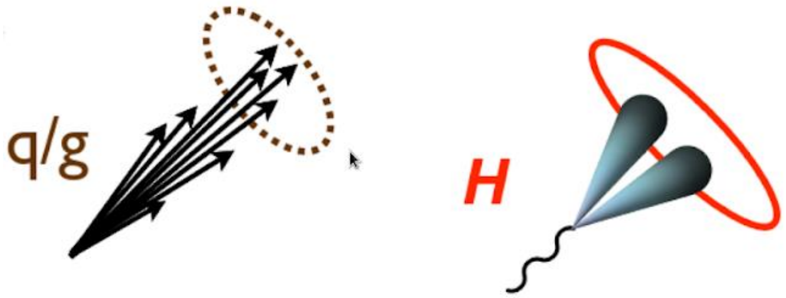
x 600

General Strategy

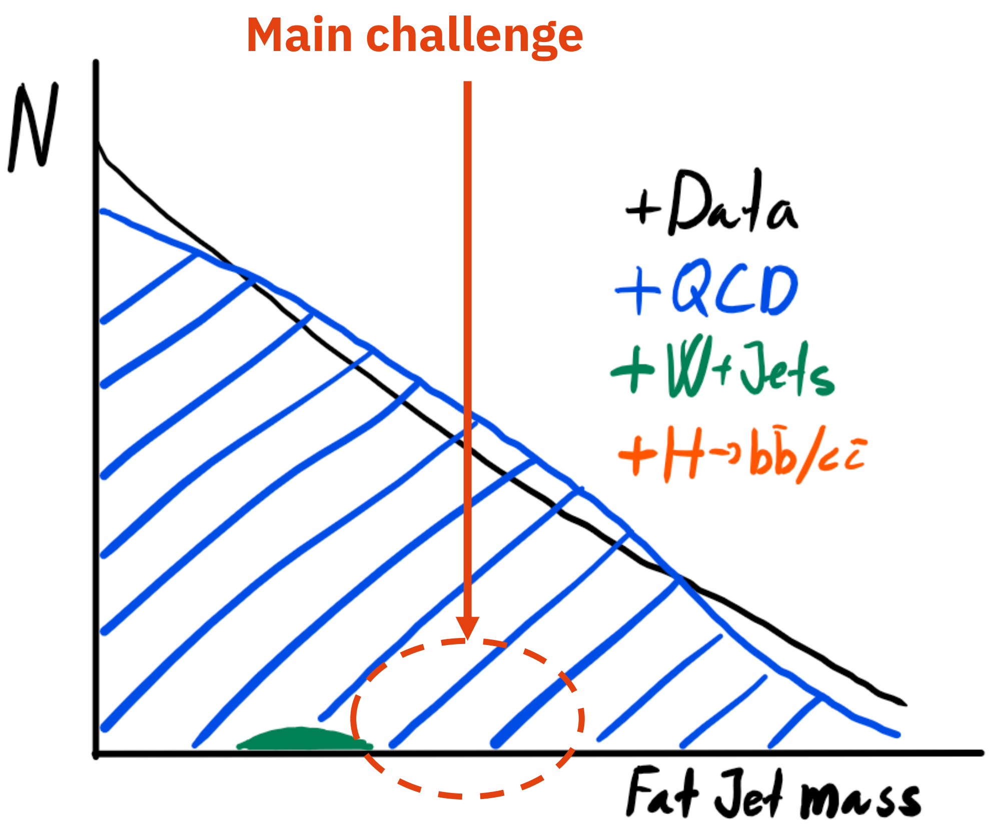
Gluon Fusion Production Mode



Jet Identification



Data-driven fit in jet mass



Analysis In a Nutshell

Event Selection

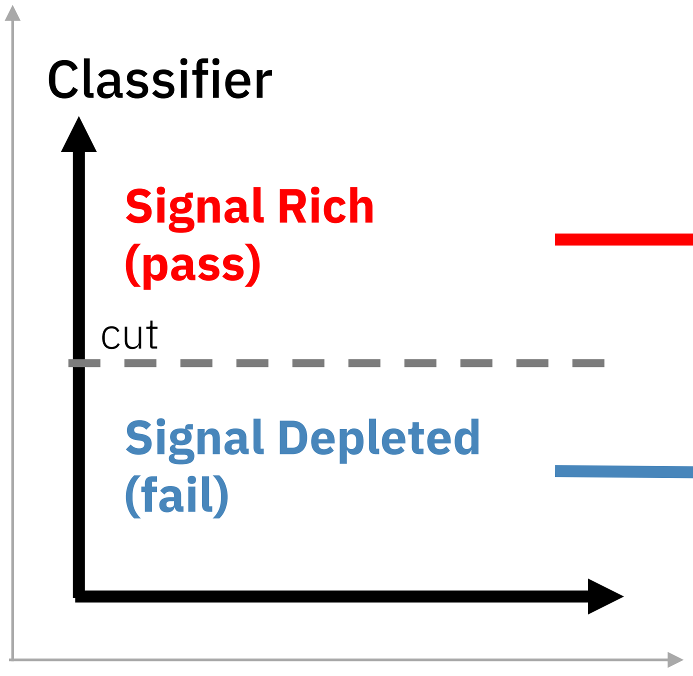
Lepton veto

Top veto
- $E_T^{miss} < 140$
- b-tag veto

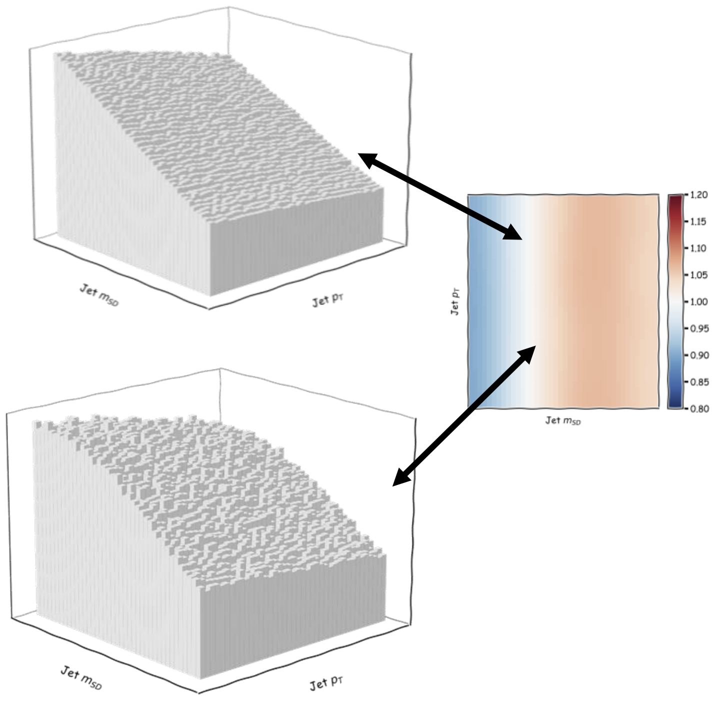
*Mix of Fat Jet Triggers

AK8 Jet (Higgs Cand.)
- m_{SD}, p_T cuts
- N_2 substructure cut
- Charm vs. bottom cut

Signal Region Definition



Simultaneous Fit



Fit jet soft-drop mass in bins of p_T

Constrain QCD **from data *in-situ***

Analysis In a Nutshell

Event Selection

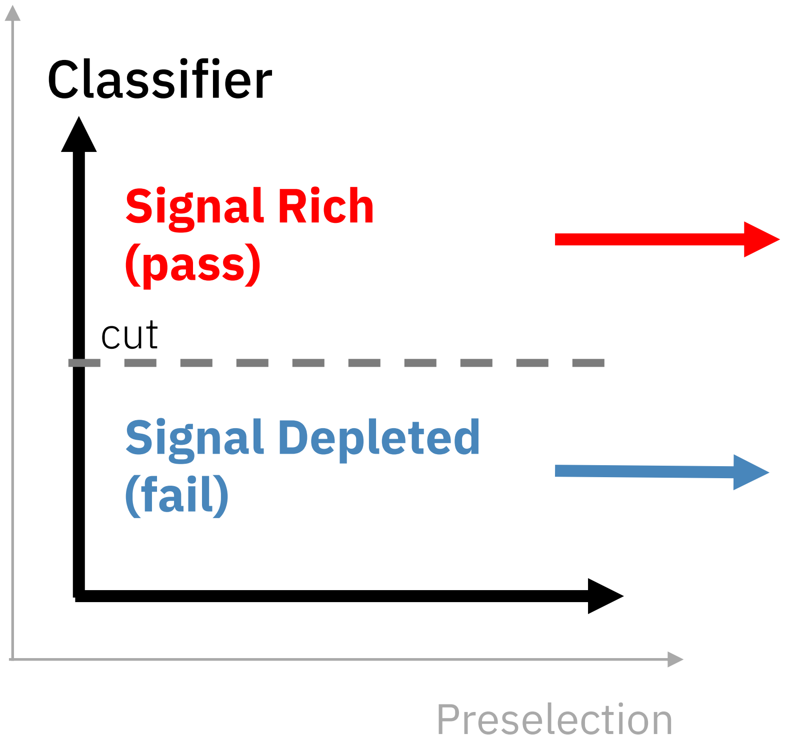
Lepton veto

Top veto
- $E_T^{miss} < 140$
- b-tag veto

*Mix of Fat Jet Triggers

AK8 Jet (Higgs Cand.)
- m_{SD}, p_T cuts
- N_2 substructure cut
- Charm vs. bottom cut

Signal Region Definition



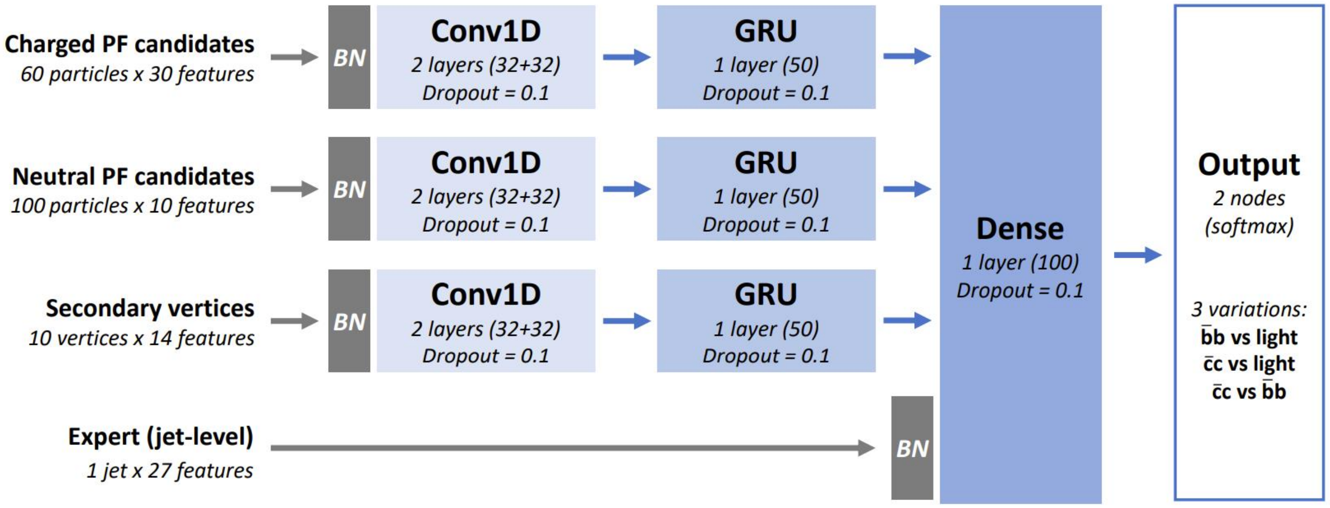
Simultaneous Fit

Fit jet soft-drop mass in bins of p_T

Constrain QCD **from data *in-situ***

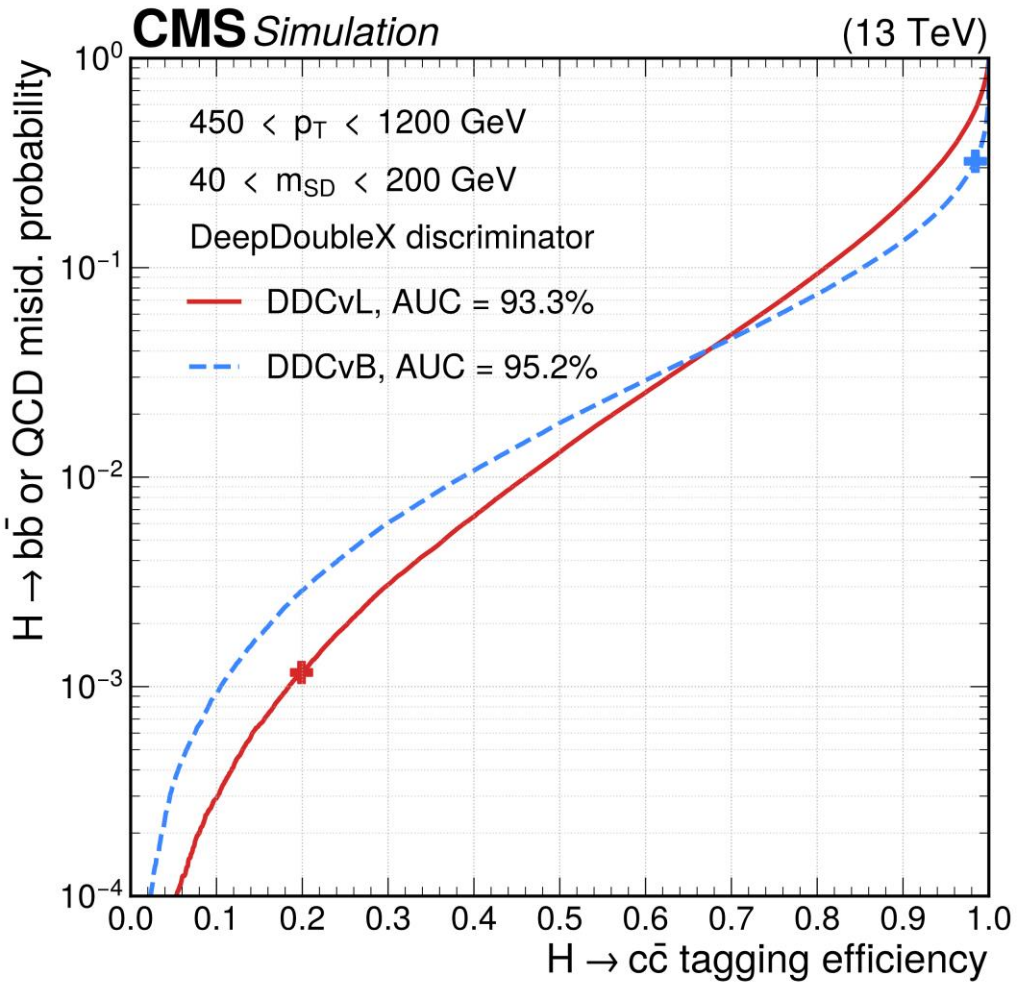
Charm Tagging – DeepDoubleX

DeepJet (CNN) inspired architecture with extended range of low-level inputs



Input relevance studies (IG, DeepTaylor)

- ~50% reduction in input features, no loss in perf.
- Much faster inference



Charm Tagging – Mass Decorrelation

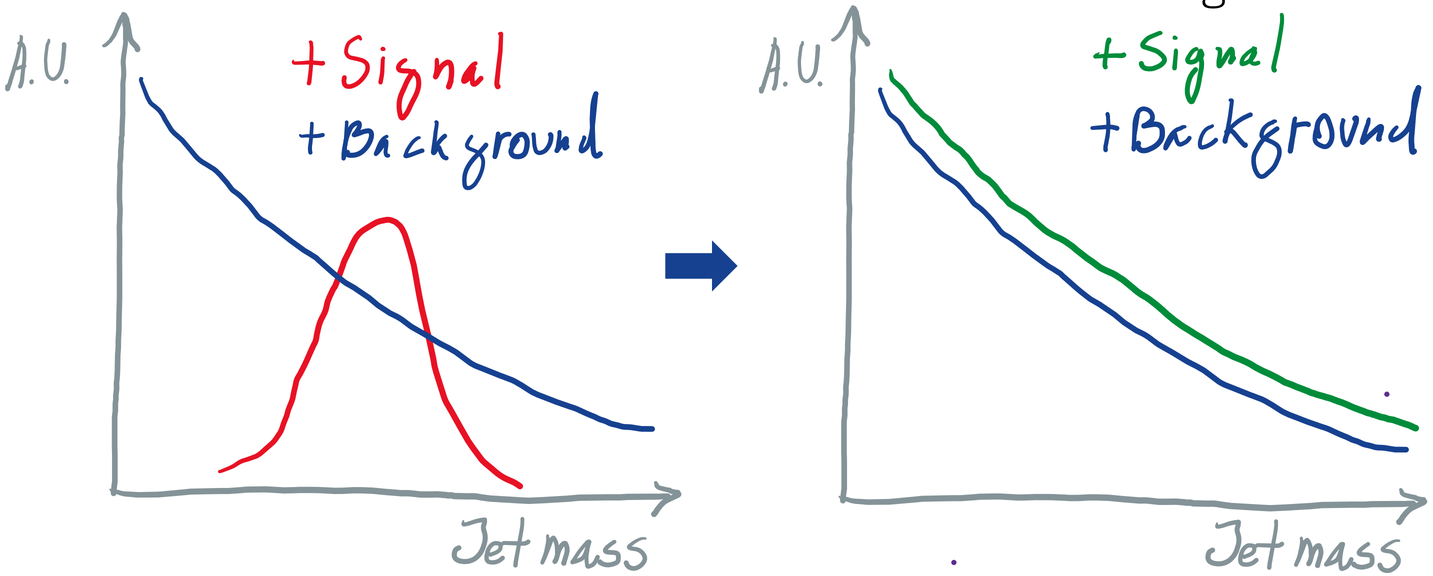
Independence of the fitted variable on the selection is crucial

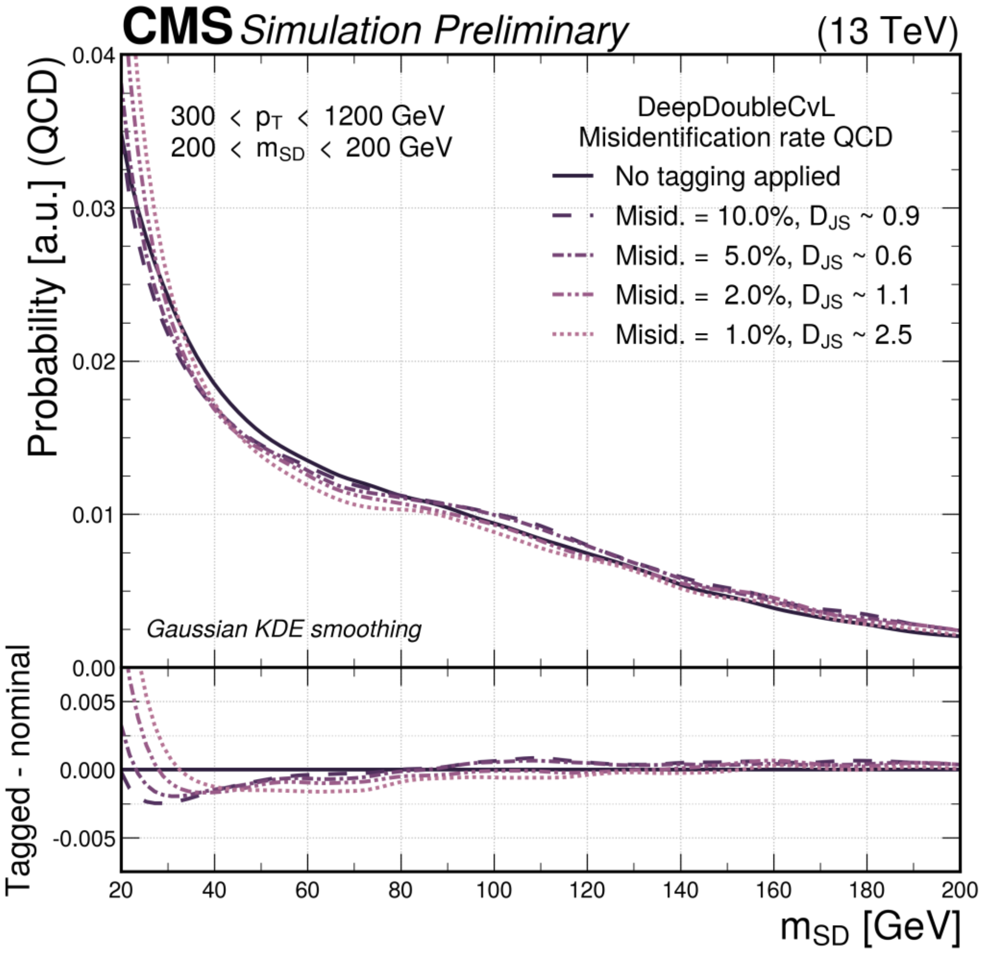
- In particular for data-driven background estimation

Train classifiers with flat mass spectrum signal

Plain training: no mass decorrelation

Mass-decorrelated training





Analysis In a Nutshell

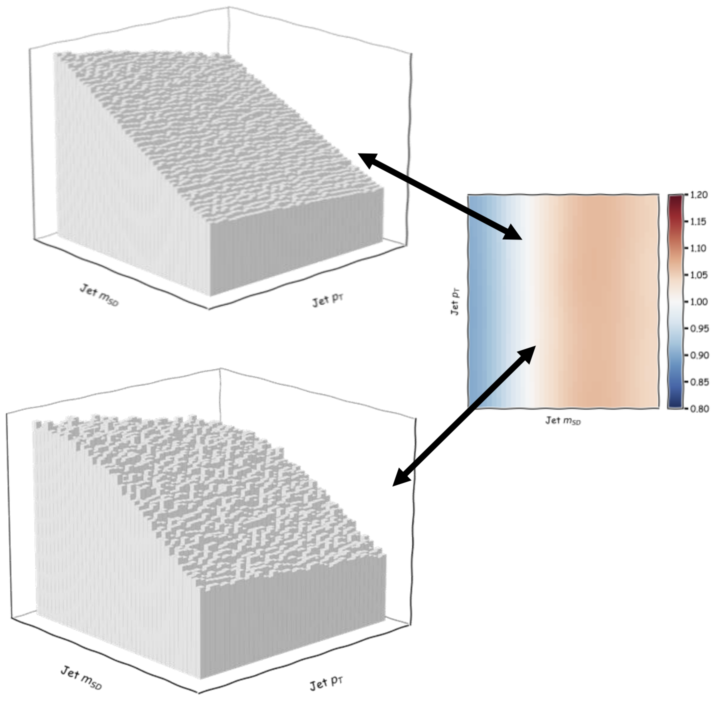
Event Selection

Lepton veto
Top veto
- $E_T^{miss} < 140$
- b-tag veto

AK8 Jet (Higgs Cand.)
- m_{SD}, p_T cuts
- N_2 substructure cut
- Charm vs. bottom cut

Signal Region Definition

Simultaneous Fit



Fit jet soft-drop mass in bins of p_T
Constrain QCD **from data *in-situ***

Transfer Factor Fit

Rhalphabet

- Differential **Alphabet** (ABCD)

Transfer Factor – residual correction

- Accounting for different tagger response
- **Flat TF → Regular ABCD method**

TF defined (constrained) bin **barycenters**

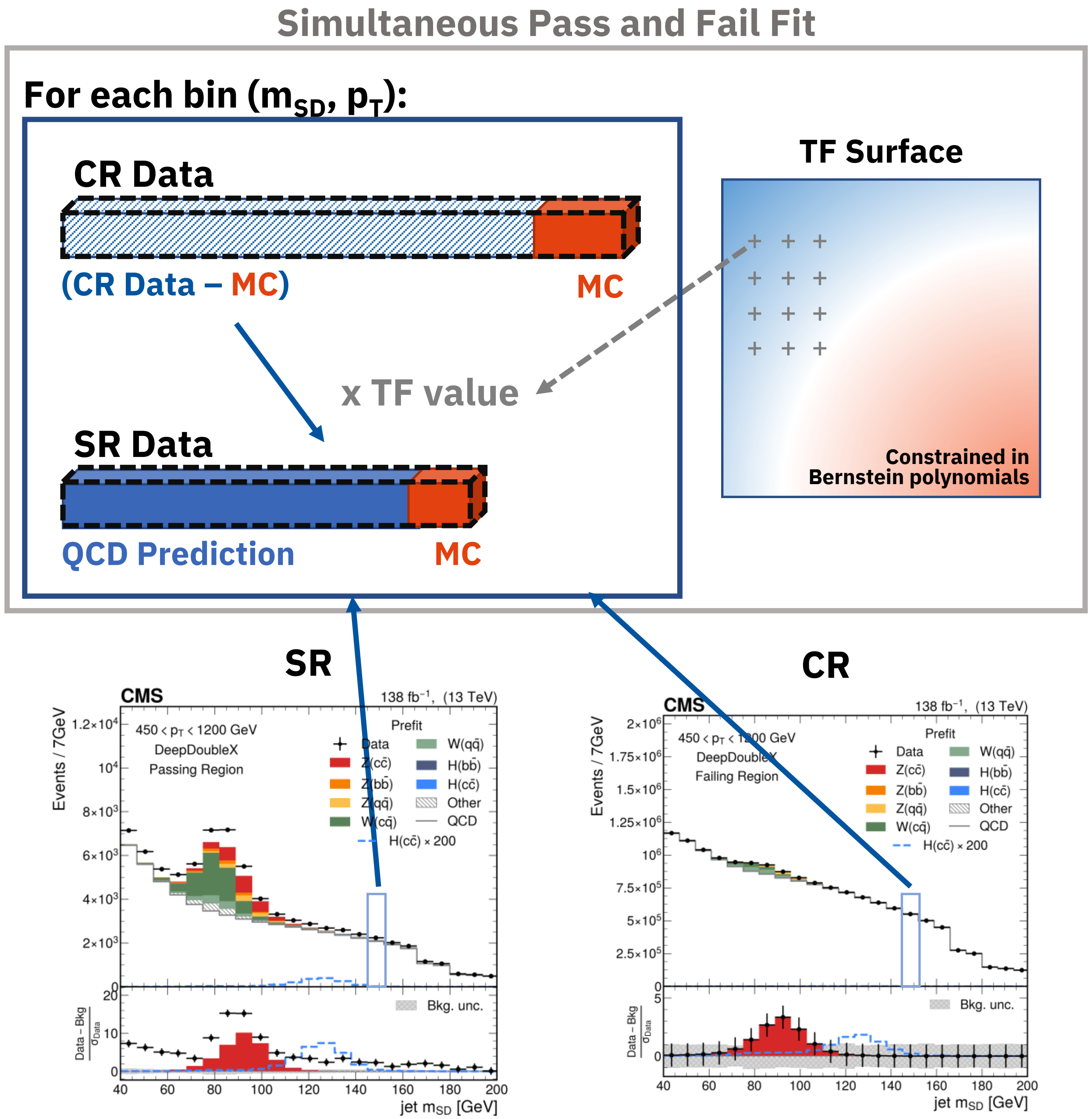
- 6 pT x 23 ρ (mass) bins
- Surface parametrized in **Bernstein basis**

$$R(\rho, p_T) = \sum_{k=0}^{n_\rho} \sum_{\ell=0}^{n_{p_T}} a_{k,\ell} b_{k,n_\rho}(\rho) b_{\ell,n_{p_T}}(p_T) \epsilon^{\mathrm{QCD}}(\rho, p_T)$$

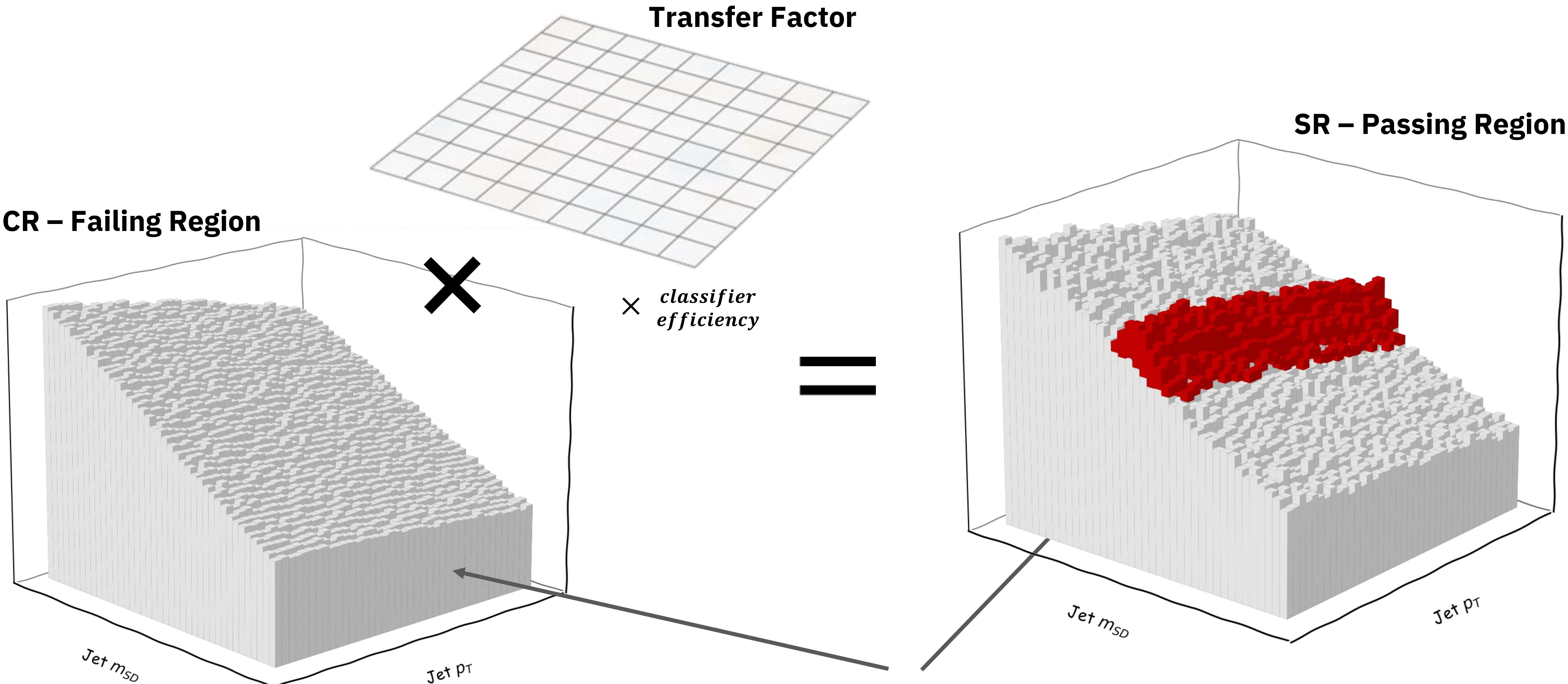
$$b_{\nu,n}(x) = \binom{n}{\nu} x^\nu (1-x)^{n-\nu} \qquad \rho = log \frac{p_T^2}{m_{SD}^2}$$

Order of polynomials is arbitrary

- Determine optimal configuration based from goodness of fit (F-test)

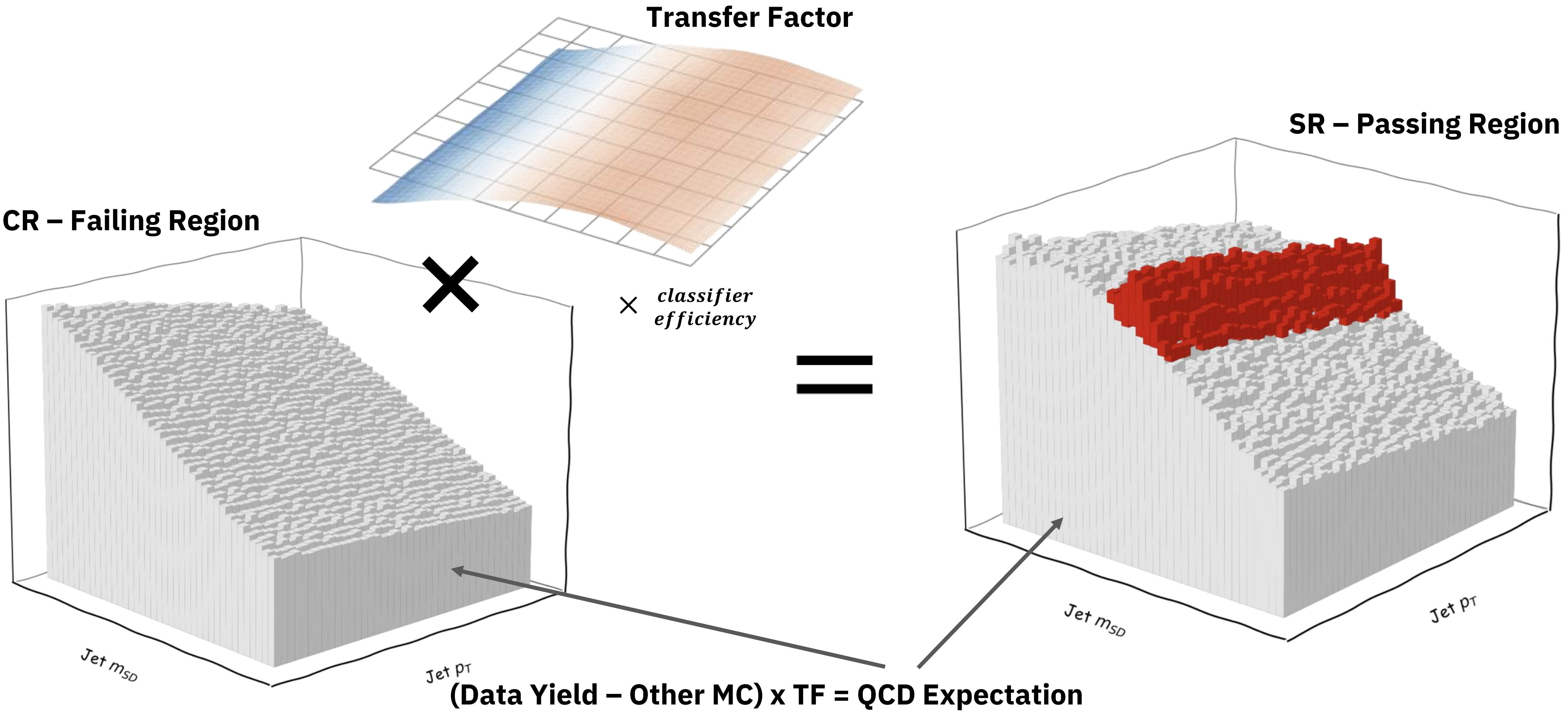


Transfer Factor Fit – Optimal Case (ABCD)



(Data Yield – Other MC) x TF = QCD Expectation

Transfer Factor Fit – Realistic Example



Transfer Factor Fit – In Practice

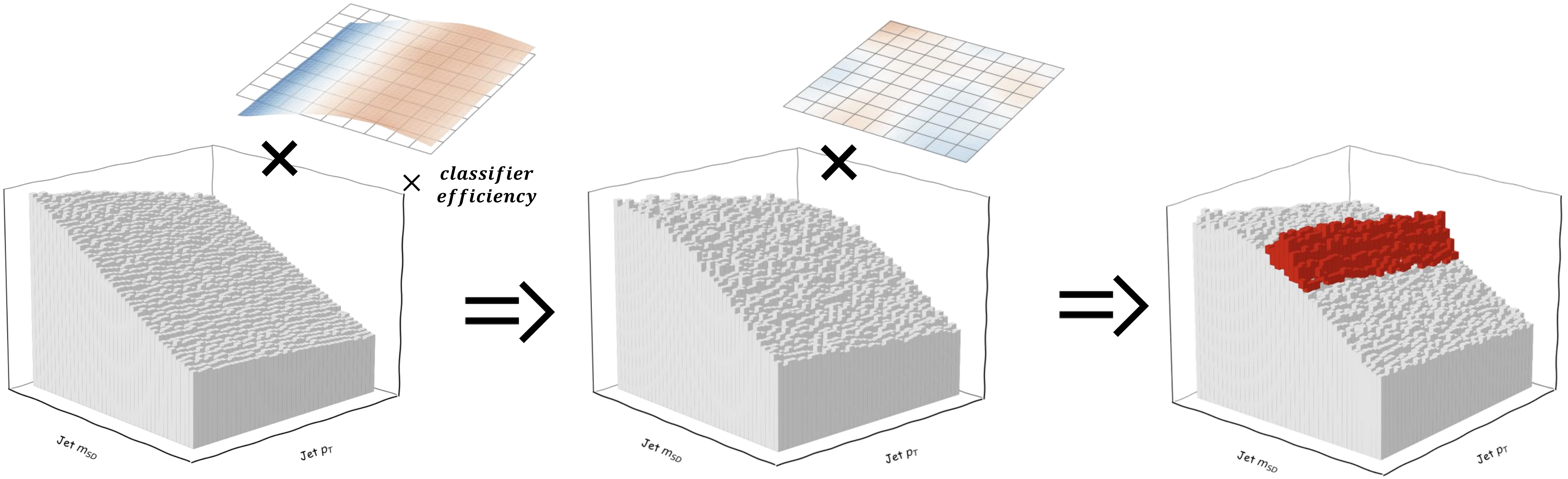
2 effects in to parametrize

- Classifier **Mass Sculpting** (MC only fit)
- **Discrepancies** due to **Data/MC** mismodelling

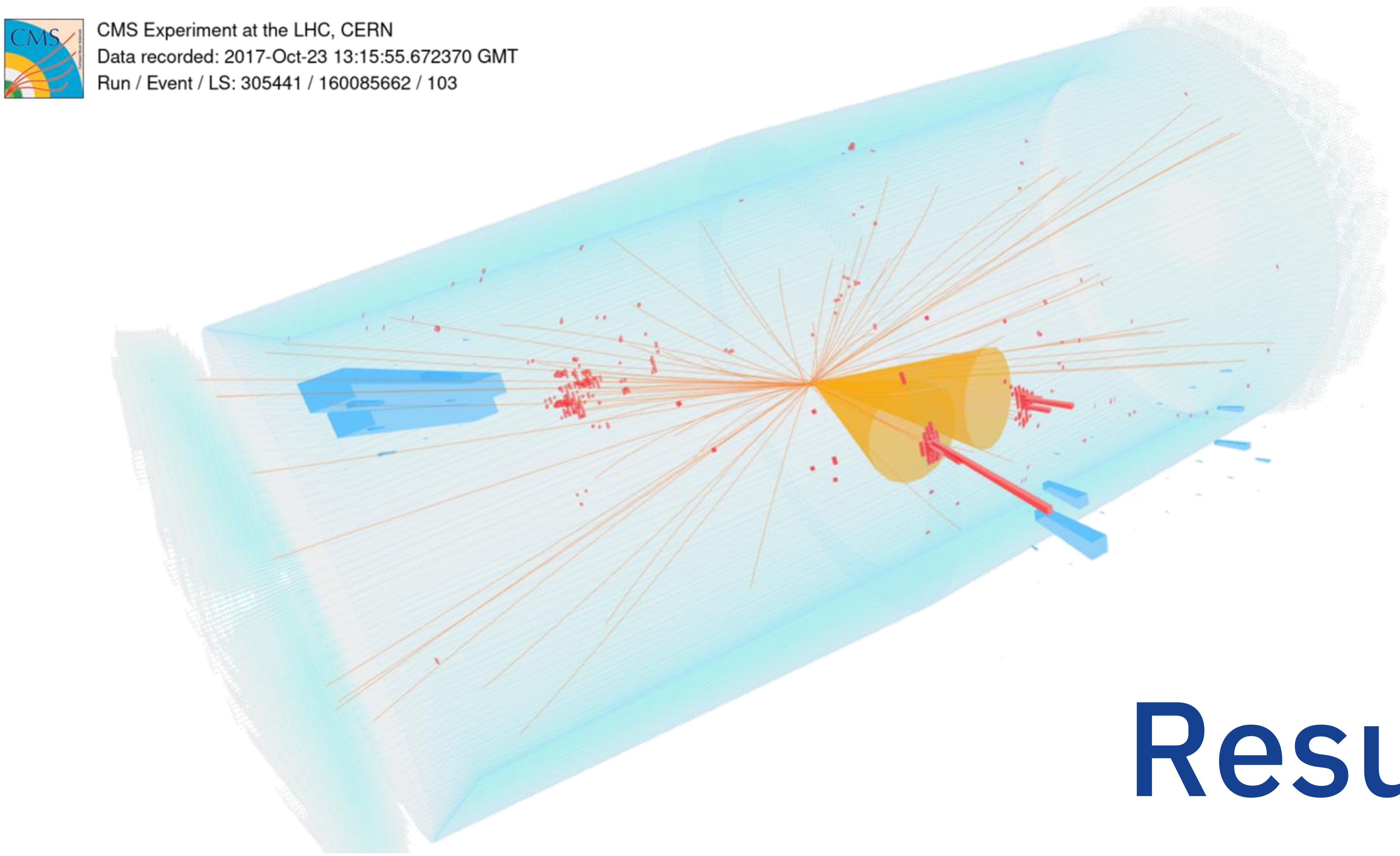
F-Tests to optimize #dof

Method tested against bias

- Spurious ‘’peaky” signal
- Different TF parametrizations

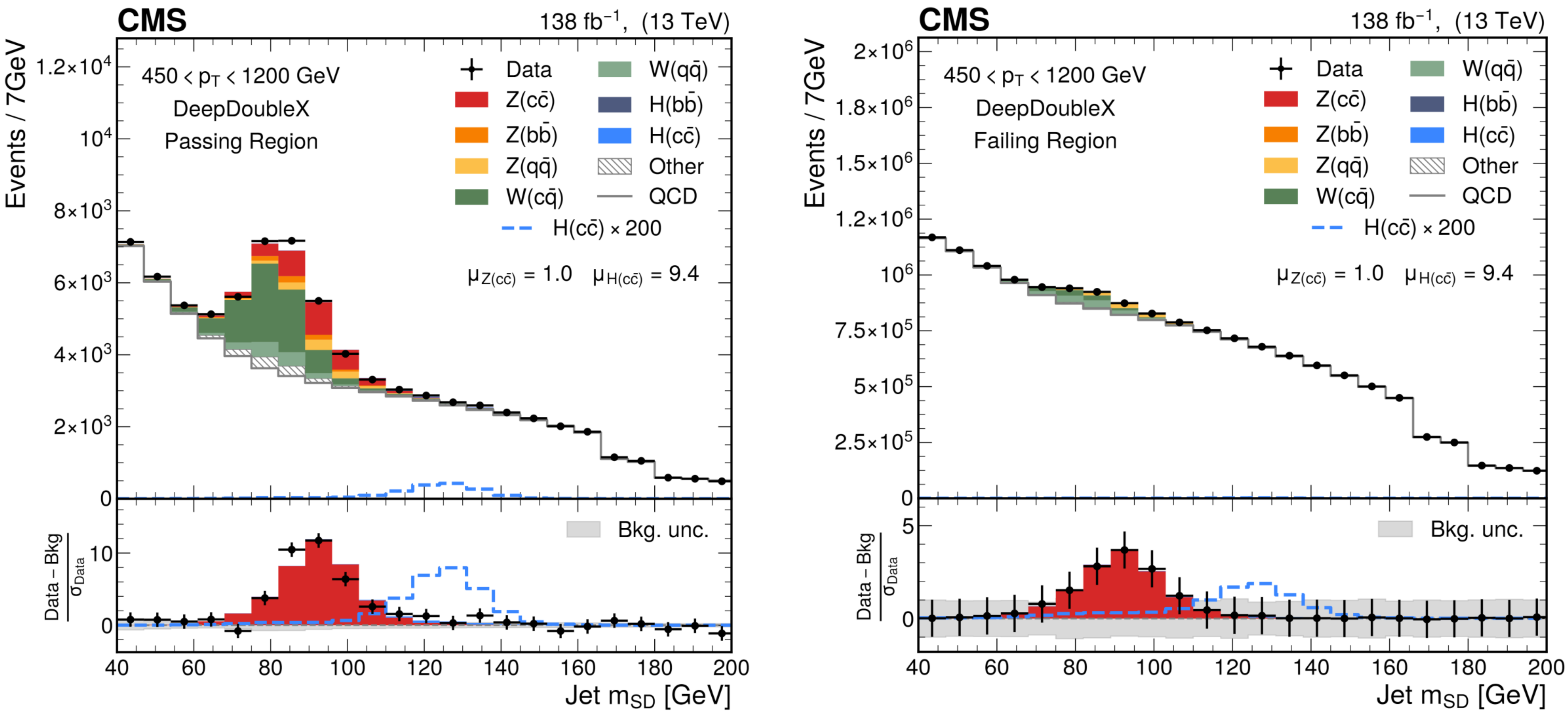


Live Event View



Results

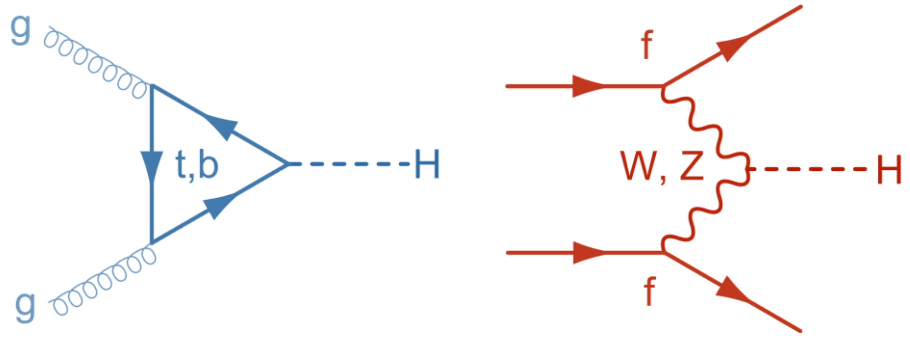
Results – Z → cc Observation



ggH → cc Limit

Inclusive in H production mode

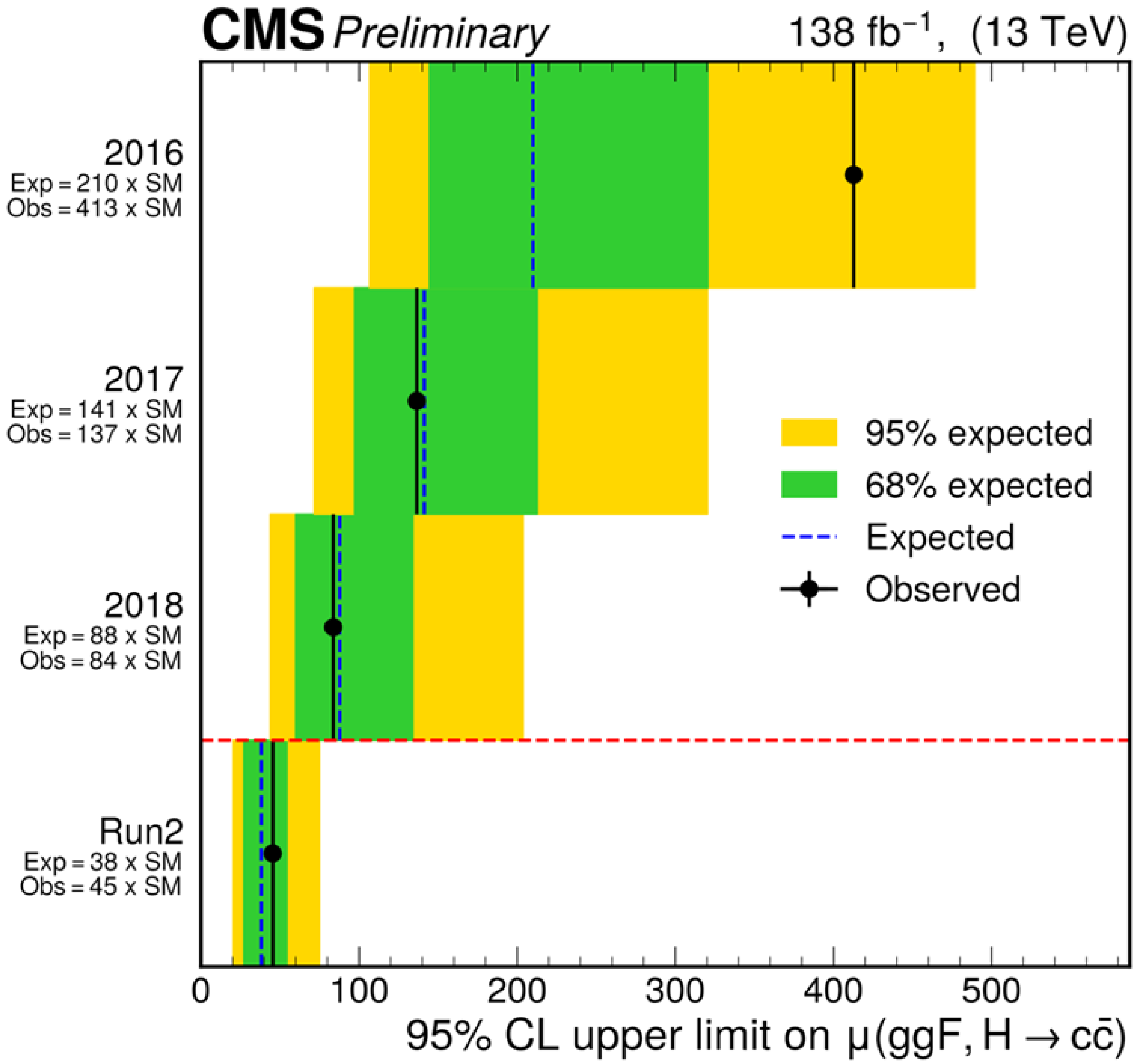
- ~ 50% ggF, ~30% VBF



$\mu_{H \to cc}$ < 47 (39) x SM@95%

- **Orthogonal to VH → cc**
- Higher p_T regime

Statistical uncertainties dominant



Summary

Full **Run2 ggH →cc Analysis** in a boosted regime, submitted to PRL

- Largely possible due to **DeepDoubleX** tagger
- First time this channel is explored at the LHC

Observation of **Z →cc** validating analysis method

- Significance >> 5σ
- Strongest constraint at the LHC yet
- First measurement in this production mode

Observed (expected) **H → cc Limit < 47 (39) x SM expectation**

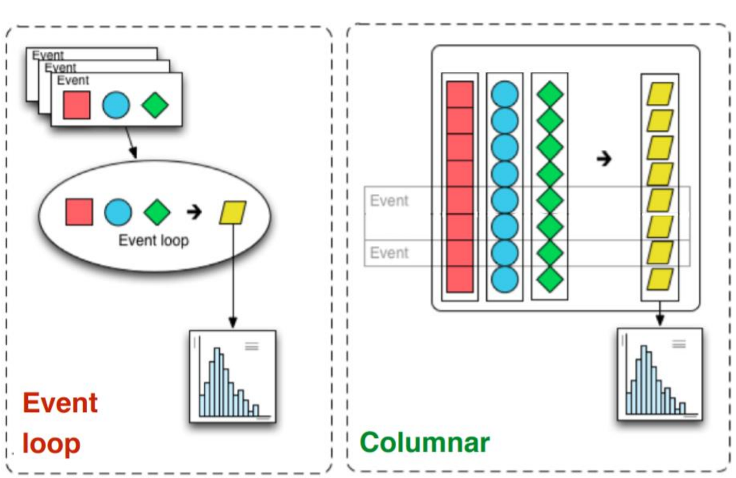
- Entirely orthogonal configuration to previous searches

European Research Council
Established by the European Commission

Thank You

This project has received funding from the European Research Council under the European Unions' Horizon 2020 research and innovation programme. [ID: 724704]

Technical Aspects



Implementation is *nearly* ROOT independent

Using Coffea framework with Scikit-HEP Software stack

- awkward – handling jagged data
- uproot – ROOT file reading
- (Fitting still dependent on RooFit/CMS Combine
- mplhep – plotting

Easy scale out Parsl DASK

Full Run2 analysis **~ 30TB data+sim, ~3TB branches accessed**

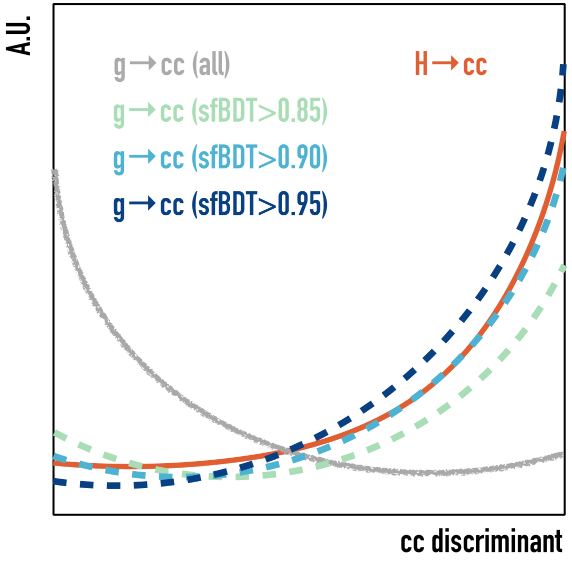
- Incl. systematic variations (JES/JER…)
- 10-20 nodes of 40-120 threads depending on availability

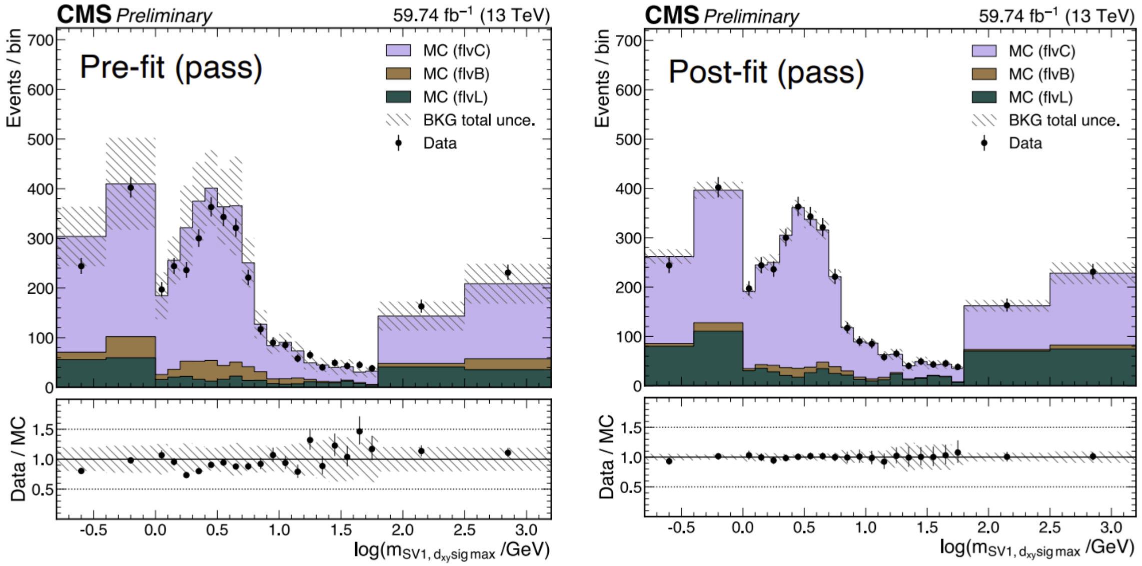
< 40 minutes total runtime (I/O limited) in optimal conditions

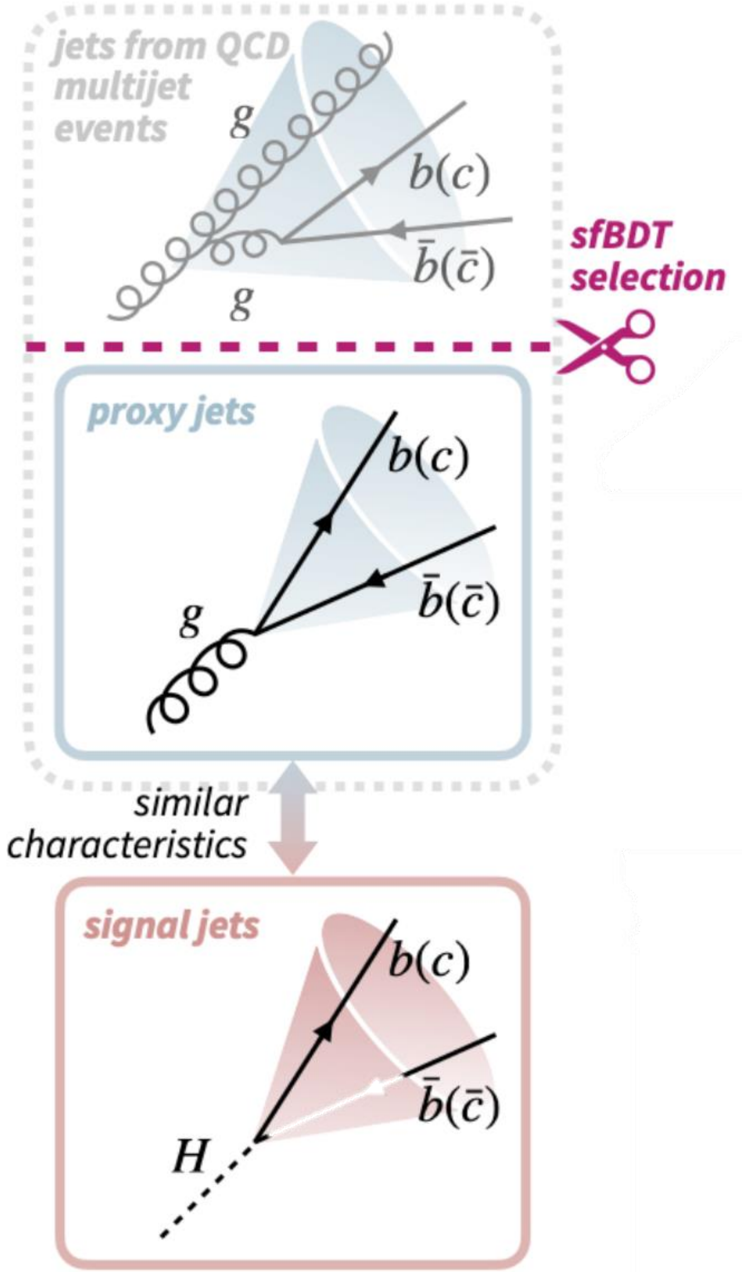
Tagger Calibration

- No existing pure H → cc or Z → cc region exists
- Use g → cc as a proxy (not entirely signal-like)
- Two ways of "forcing" a good proxy – compatible results
  1. Use soft-muon presence to select Higgs-like cc jets
  2. Train a BDT to select Higgs-like cc jets

Effects of the BDT







Background Parametrization Alternatives

- Bernstein – baseline
  - Previously used, easy to fit
  - Polynomial space

$$R_{\mathrm{p/f}}(\rho, p_{\mathrm{T}}) = \sum_{k=0}^{n_\rho} \sum_{\ell=0}^{n_{p_{\mathrm{T}}}} a_{k,\ell} \left[b_{k,n_\rho}(\rho) b_{\ell,n_{p_{\mathrm{T}}}}(p_{\mathrm{T}}) \right]$$

- Chebyshev
  - More sensitive to initial values
  - Polynomial space
  - Replace $b_{k,l}$ terms with $c_{k,l}$

- Exponential transform of Bernstein
  - Less stable
  - Independent parameter space

$$R_{\mathrm{p/f}}(\rho, p_{\mathrm{T}}) = \sum_{k=0}^{n_\rho} \sum_{\ell=0}^{n_{p_{\mathrm{T}}}} a_{k,\ell} \exp \left[b_{k,n_\rho}(\rho) b_{\ell,n_{p_{\mathrm{T}}}}(p_{\mathrm{T}}) \right]$$

1	-2	1
	2	-2
		1

Bernstein

1		
-1	2	
1	-8	8

Chebyshev

1		
	1	
		1

Polynomial