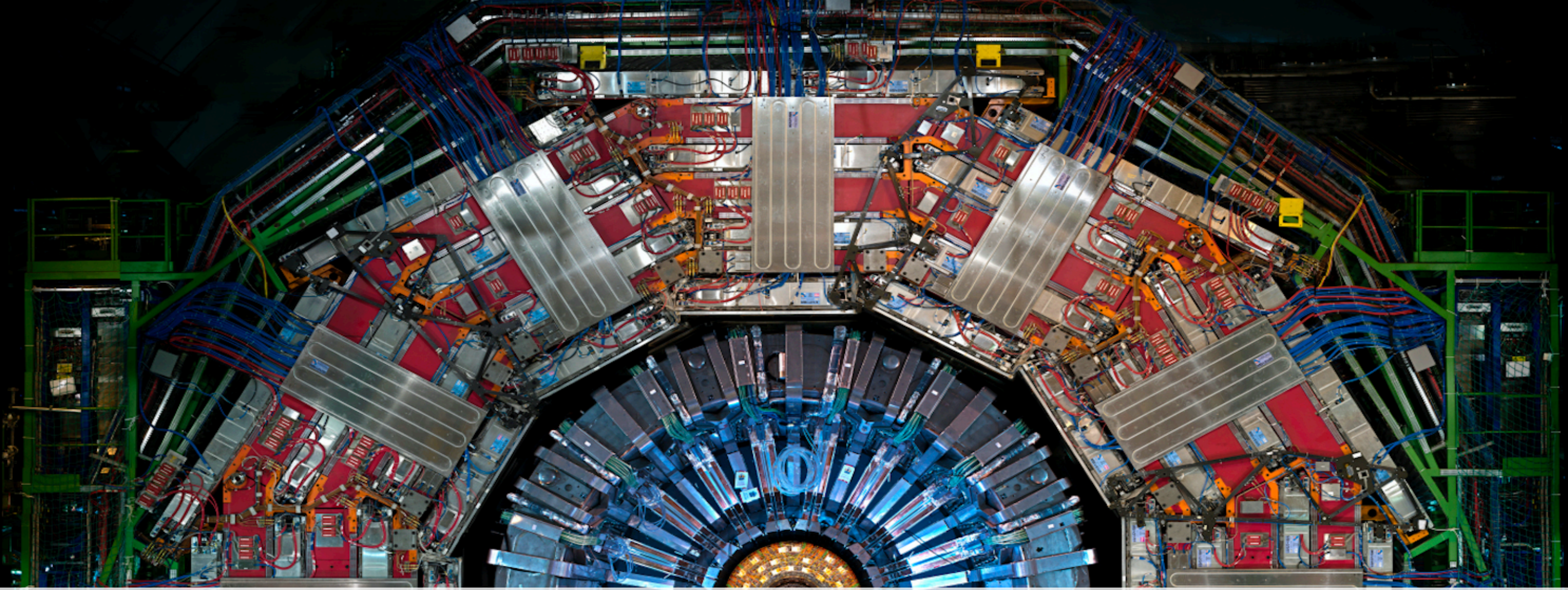
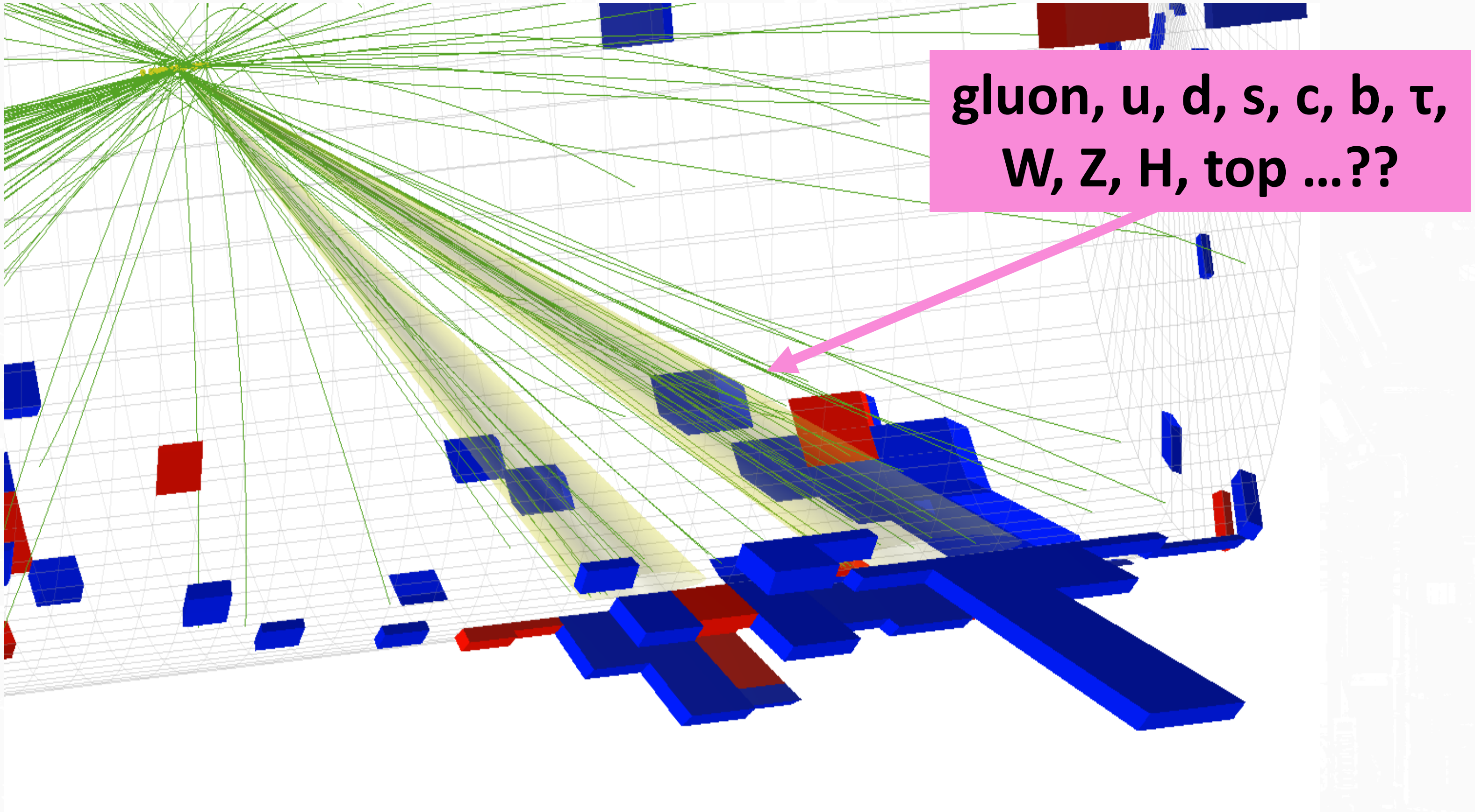


Jet flavor identification using Graph Neural Networks in CMS

Loukas Gouskos (CERN)
on behalf of the CMS Collaboration

CERN-Data Science seminar
Oct 5th, 2022

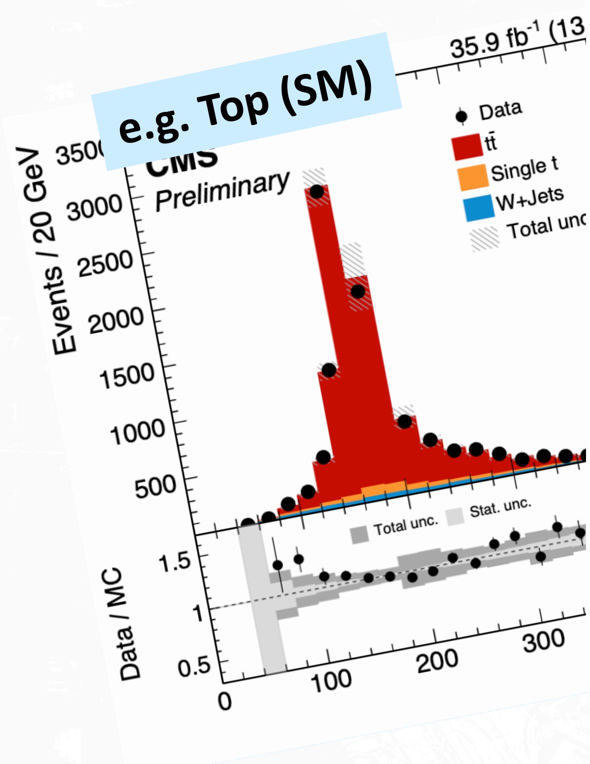
Jet identification ("tagging")

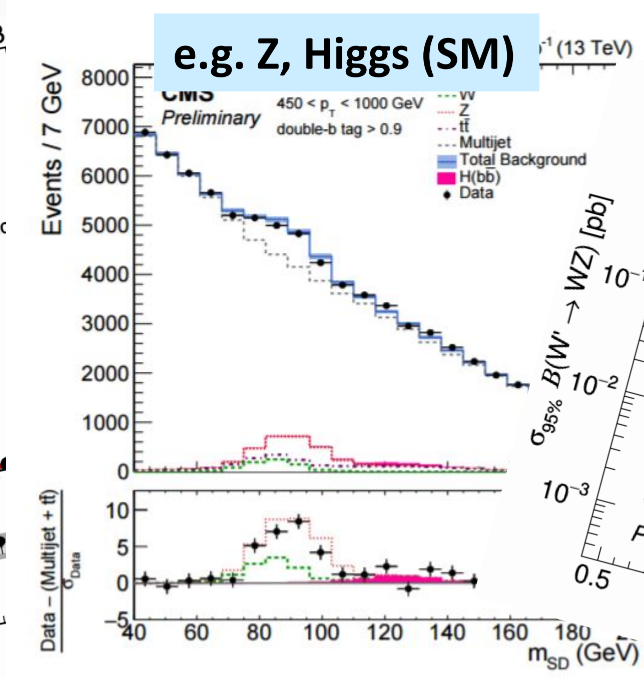


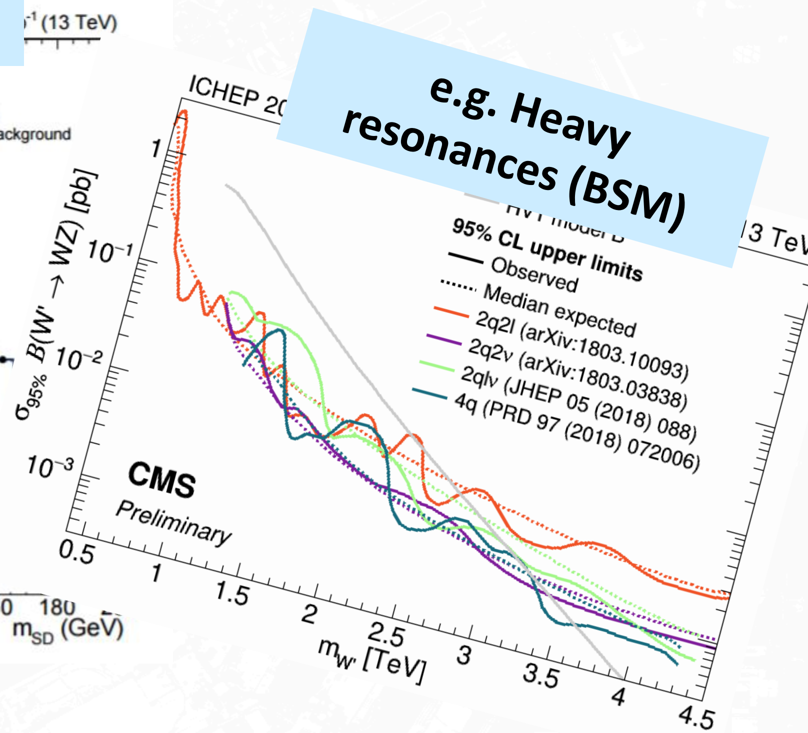
- Jet tagging: A topic of high interest in both TH and EXP communities
  - More than 30 years at colliders
    - b-jets at LEP and Tevatron; W, Z, H,.. at the LHC
- Recently: much more powerful algorithms w/ multi-object tagging capabilities
  - opened-up unchartered territory

Setting the scene

- Jets play a crucial role in the LHC physics program
  - both Standard Model (SM) and beyond (BSM)







- **Key for success:** Well calibrated jets & constantly improving our "JetToolbox"
  - particularly important now that LHC integrated luminosity increases only ~linearly with time & do not expect big jumps in collision energy

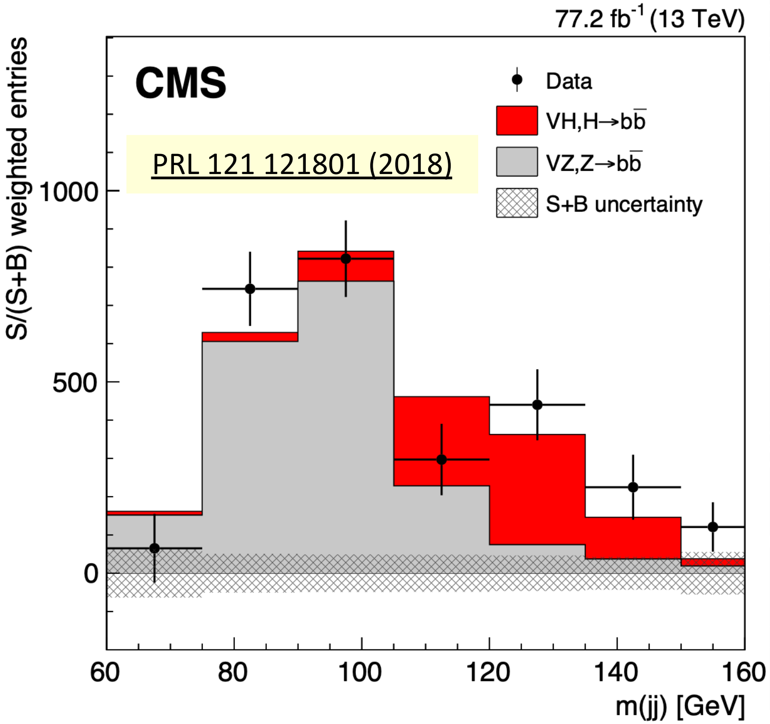
Setting the scene (II)

- Often exceeding expectations:

In conclusion, the extraction of a signal from $H \to b\bar{b}$ decays in the WH channel will be very difficult at the LHC, even under the most optimistic assumptions for the b-tagging performance and calibration of the shape and magnitude of the various background sources from the data itself.

[Ref: ATLAS TDR 1999]
feeling not different in CMS…

H→bb discovery in VH production



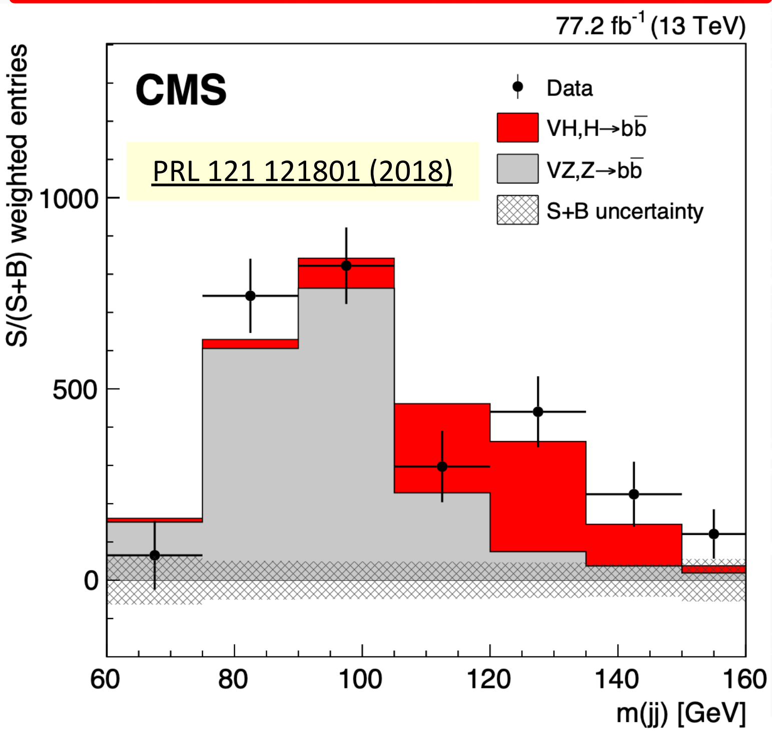
Setting the scene (II)

- Often exceeding expectations:

In conclusion, the extraction of a signal from $H \to b\bar{b}$ decays in the WH channel will be very difficult at the LHC, even under the most optimistic assumptions for the b-tagging performance and calibration of the shape and magnitude of the various background sources from the data itself.

[Ref: ATLAS TDR 1999]
feeling not different in CMS…

H→bb discovery in VH production



- Today's talk:
  - Review a decade of jet tagging [in CMS]
    - focus on "boosted" jet tagging
  - Jet tagging using GraphNNs
  - GNNs in jet physics beyond tagging
  - Highlights from CMS analyses

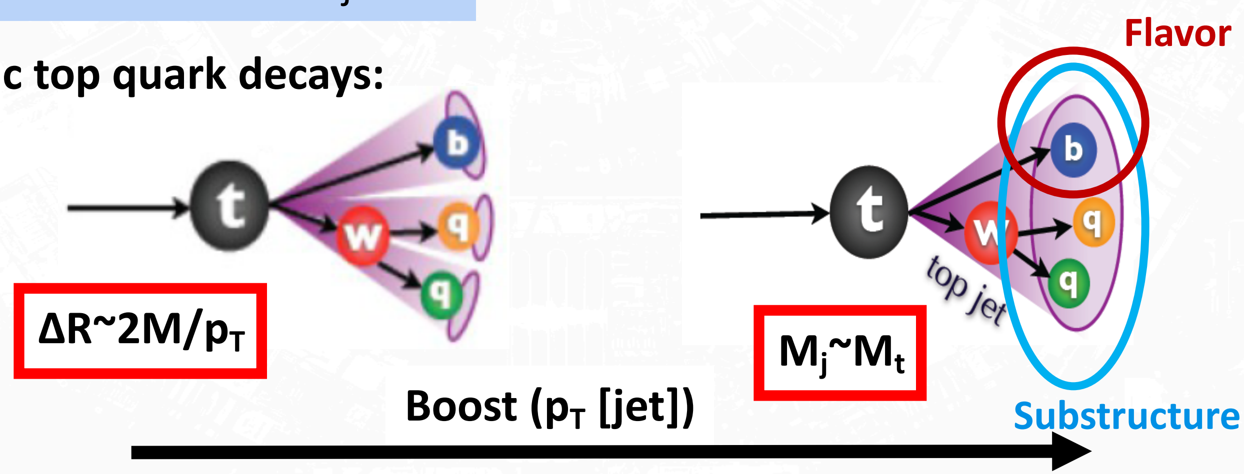
Jet identification: Basics and brief review

Today: focus on tools developed for Large-*R* jet

Low boost:
unmerged decay products
→ decay products resolved
as distinct small-*R* jets

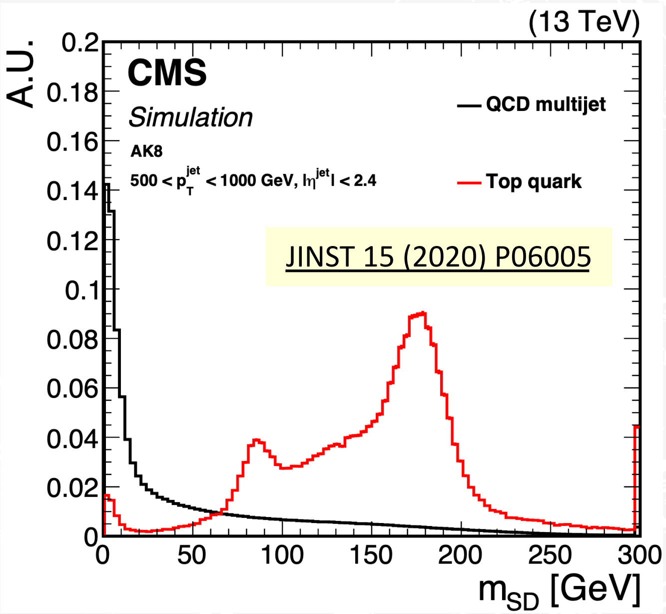
Large boost:
Decay products "merged"
in to a single large-*R* jet

e.g.,: hadronic top quark decays:

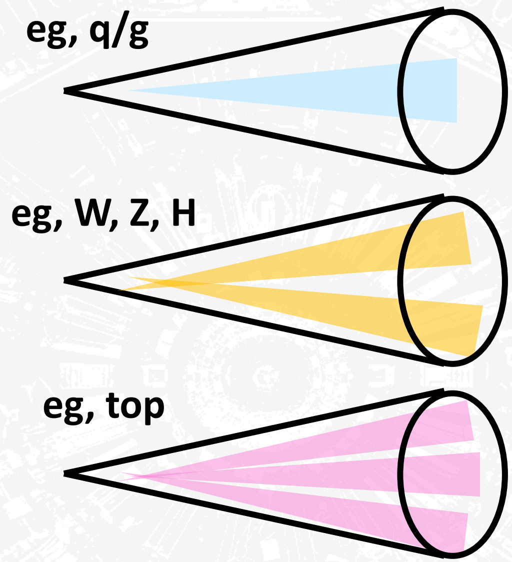


Basics for jet identification

Mass

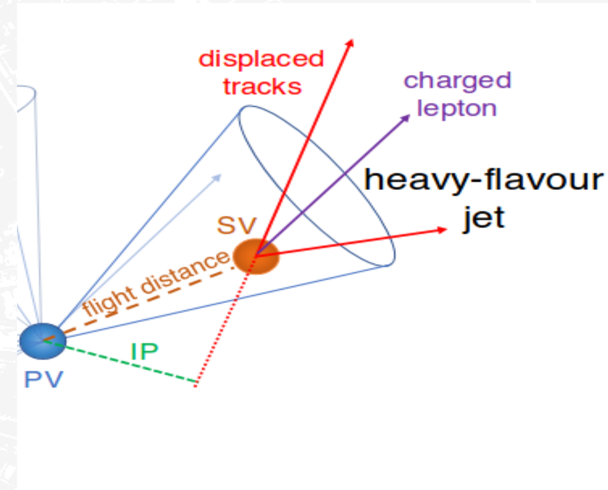


Substructure



- Explore the energy distribution inside jet

Flavor



- Exploit large lifetime of b/c hadrons
- Presence of displaced tracks & sec. vertices

In the beginning unclear what correlations existed among these

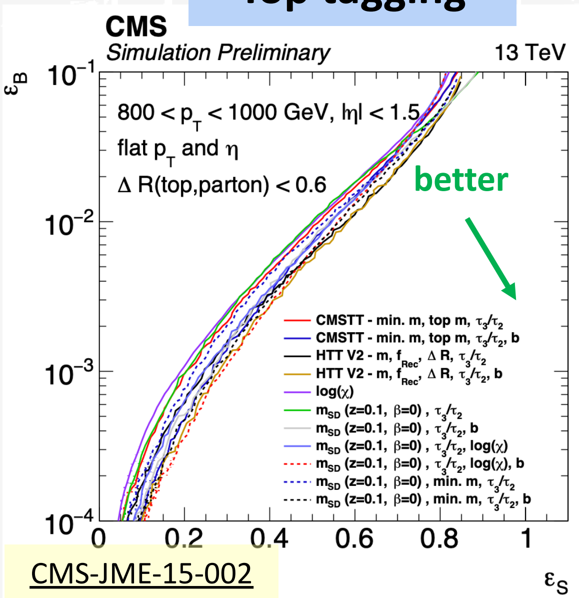
Review a decade [in CMS]

- Enormous progress since the LHC start

The early days:
Human +inspired
high-level variables
+
Cut-based or simple ML

Run 1 **Run 2** **Run 3**

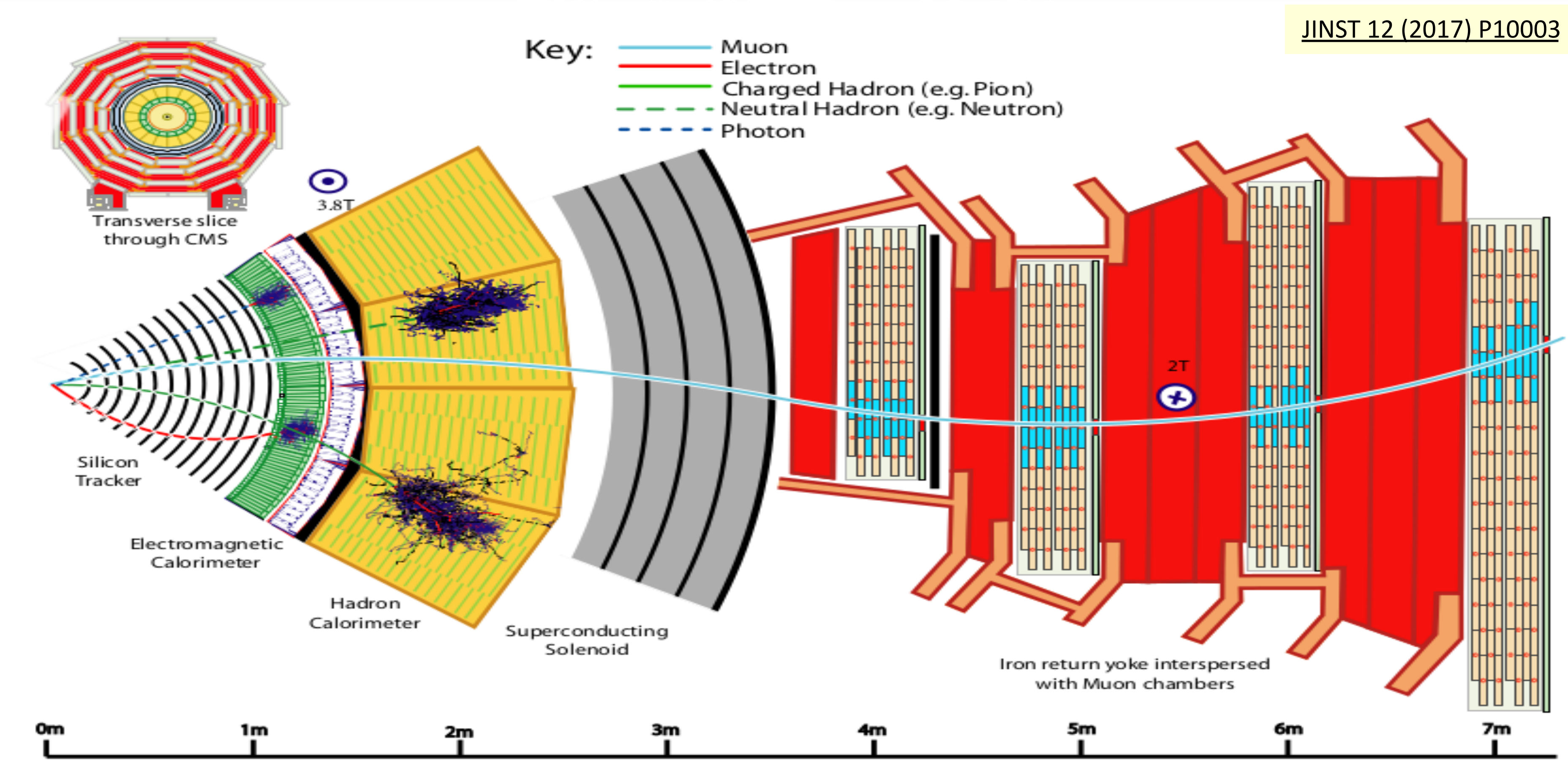
Top tagging



CMS-JME-15-002

Intermezzo: CMS event reconstruction

JINST 12 (2017) P10003

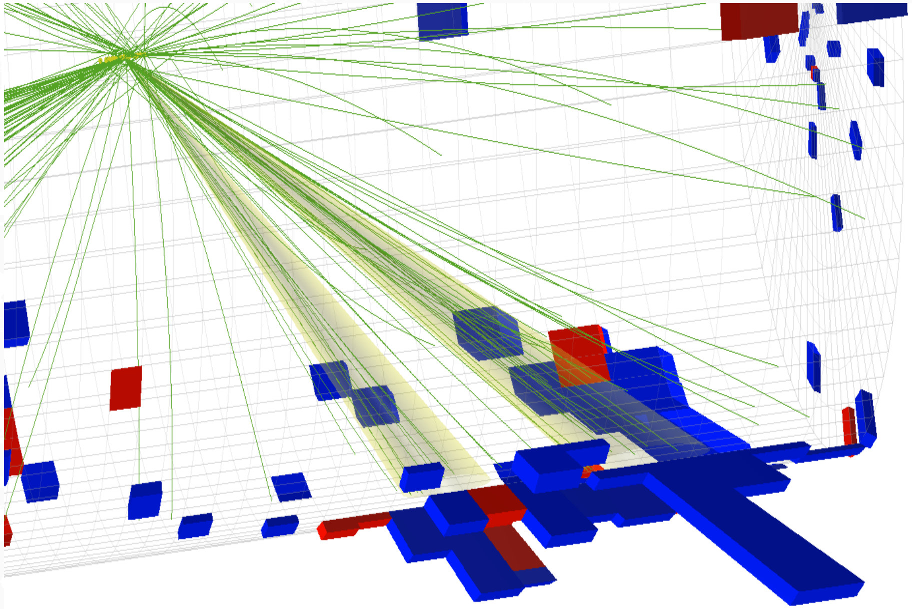


- CMS event reconstruction uses the **Particle Flow (PF)** algorithm
  - Combines information from all subdetectors
  - Output: mutually exclusive list of particles
    - Then: build higher-level objects (jets, ME_T,..)

Significant improvement in object performance wrt traditional approaches

Exploiting more of CMS potential

- **A Jet in theory:**
  - A spray of particles produced by the hadronization of quarks and gluons
- **Experimentally:**
  - A cone of reconstructed particles in the detector



- Towards **particle-based jet tagging**
  - CMS PF algo: Rich set of info / particle
    - Energy/momentum
    - Position
    - Particle type
    - Displacement from the PV
    - Reconstruction quality
    - ....

Inputs to jet substructure

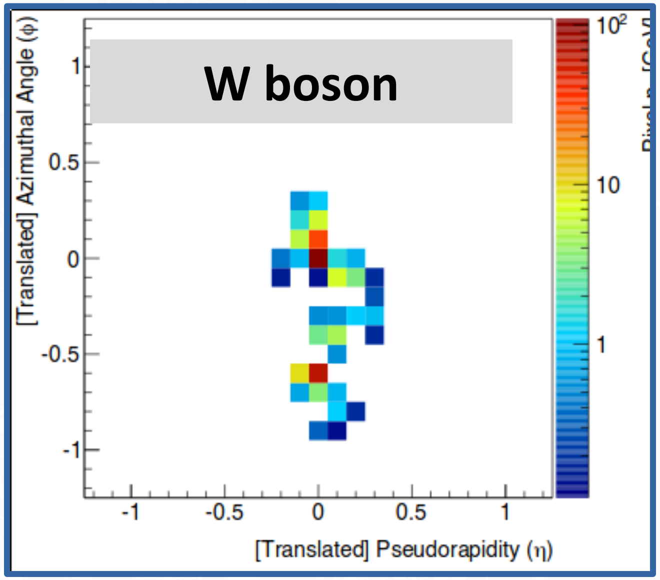
Inputs for flavour tagging

  - [O(50) properties/particle] x [~50-100 particles/jet] ~O(1000) inputs/jet
- Ideal case for Deep Learning (DL) based algorithms with low-level inputs

Jet representation

- Key ingredient for powerful DL-based algorithms

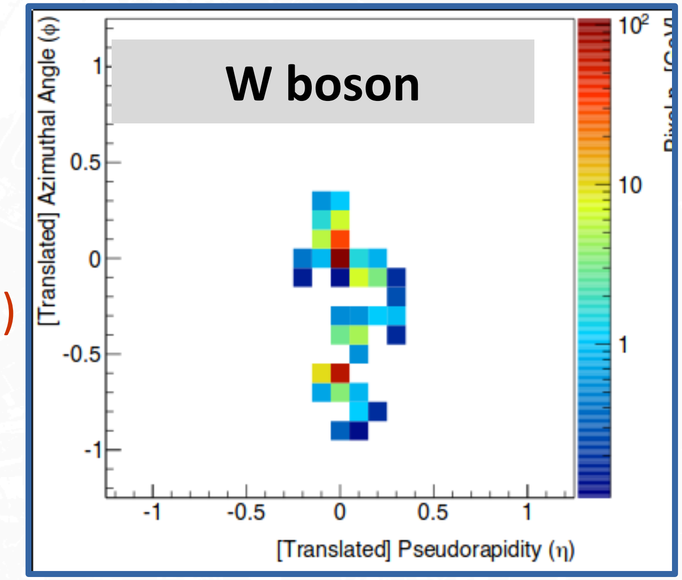
- Jet as an **image**:
  - Treat **detector** (i.e. calorimeters) as a **camera** & the **jet** as an **image**
  - Apply techniques from image recognition (CNN-2D)
  - **But:** jet images are very sparse
  - **Also:** CMS very heterogeneous/complex
    - difficult to include info from other subdetectors [eg., tracker]



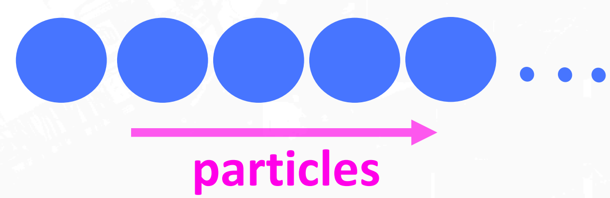
Jet representation

- Key ingredient for powerful DL-based algorithms

- Jet as an **image**:
  - Treat **detector** (i.e. calorimeters) as a **camera** & the **jet** as an **image**
  - Apply techniques from image recognition (CNN-2D)
  - **But:** jet images are very sparse
  - **Also:** CMS very heterogeneous/complex
    - difficult to include info from other subdetectors [eg., tracker]



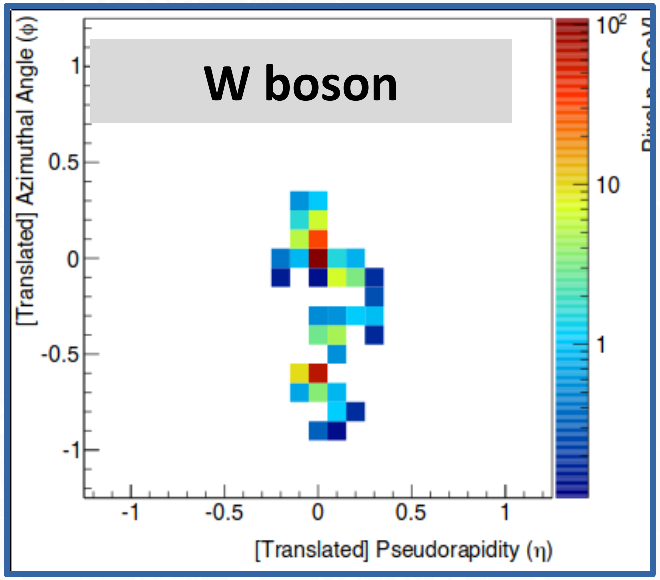
- Jet as **particle sequence**:
  - **Jet** as a **sequence of constituent particles**
  - Apply techniques from **natural language processing** [e.g. CNN-1D, RNNs ..]



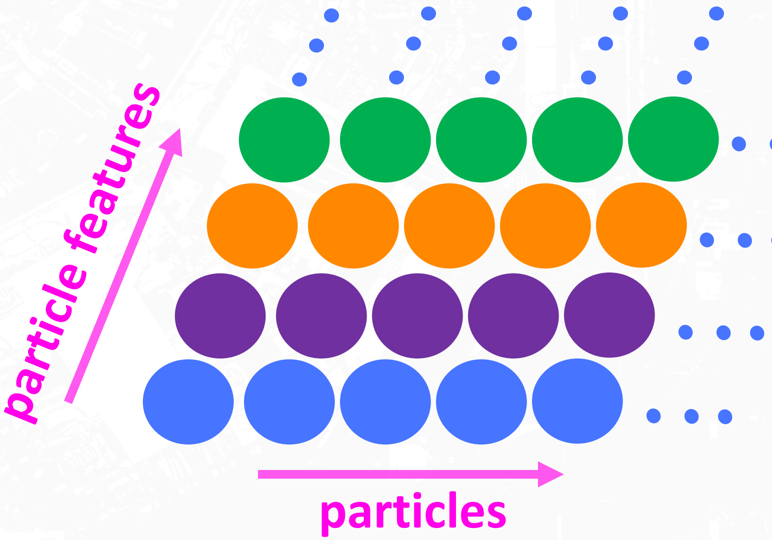
Jet representation

- Key ingredient for powerful DL-based algorithms

- Jet as an **image**:
  - Treat **detector** (i.e. calorimeters) as a **camera** & the **jet** as an **image**
  - Apply techniques from image recognition (CNN-2D)
  - **But:** jet images are very sparse
  - **Also:** CMS very heterogeneous/complex
    - difficult to include info from other subdetectors [eg., tracker]



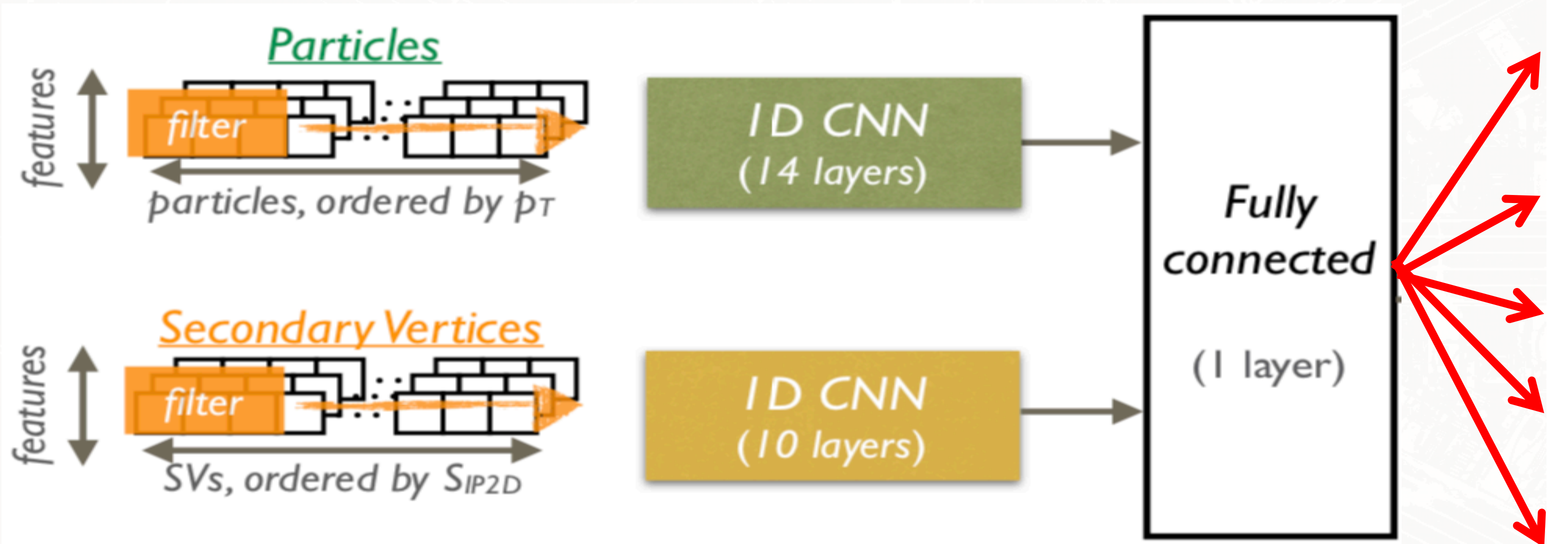
- Jet as **particle sequence**:
  - **Jet** as a **sequence of constituent particles**
  - Apply techniques from **natural language processing** [e.g. CNN-1D, RNNs ..]
  - Inclusion of more information straight forward
  - Explore **more** of the **CMS detector & CMS event reconstruction potential**



Particle-based jet tagging: *DeepAK8*

- Advanced DL-based boosted jet tagger [using particle sequences]
  - multi-class classifier for top, W, Z, Higgs, and QCD jets
  - inspired by ResNeXt50 [K. He et al.]
  - **Simultaneously** exploit *substructure* and *flavor* information

NeurIPS 2017
CMS-DP-2017-049
JINST 15 (2020) P06005



Category	Label
Top	bcq bqq bq cq
Higgs	bb cc VV*→qqqq
Z	bb cc qq
W	cq qq
QCD	g→bb g→cc b c others

Inputs:
→ Up to 100 particles (42 features/particle)
→ Up to 7 Secondary vertices (15 features/SV)

A very versatile boosted jet tagger
→ various decay modes with different flavor content

Review a decade [in CMS]

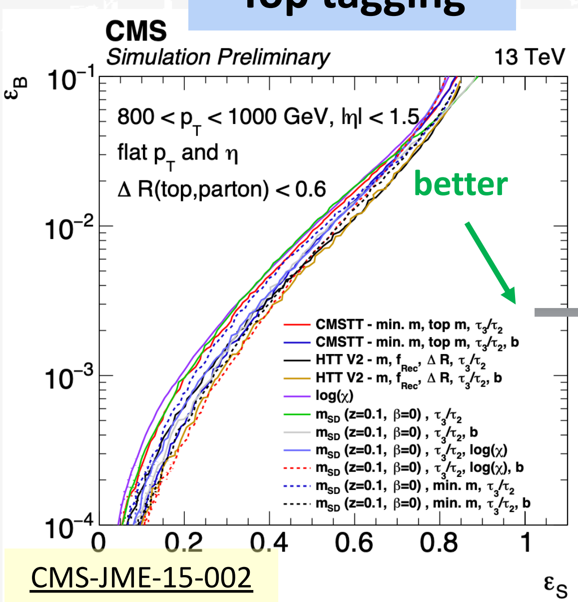
- Enormous progress since the LHC start

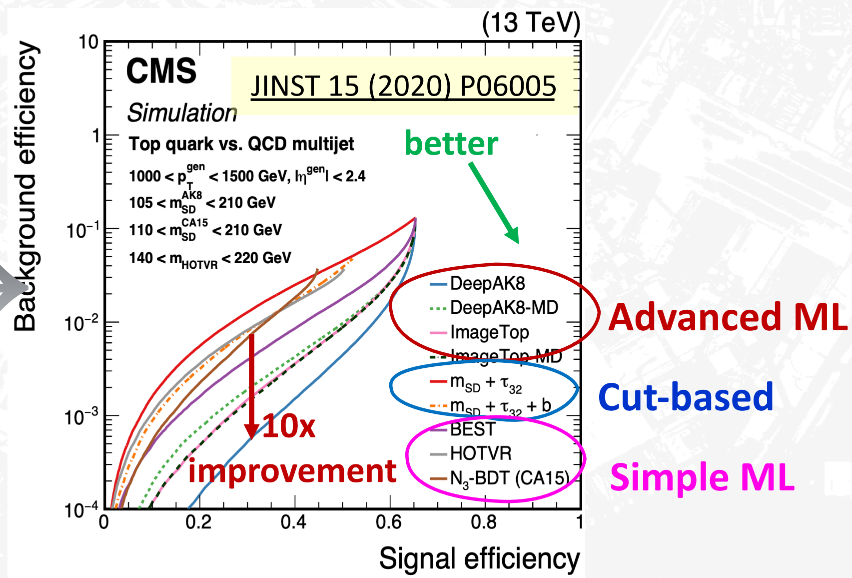
The early days:
Human +inspired,
high-level variables
+
Cut-based or simple ML

The "Game changer":
Inputs: low-level info
+
Advanced ML (CNN, RNN)

Run 1 **Run 2** **Run 3**

Top tagging





Review a decade [in CMS]

- Enormous progress since the LHC start

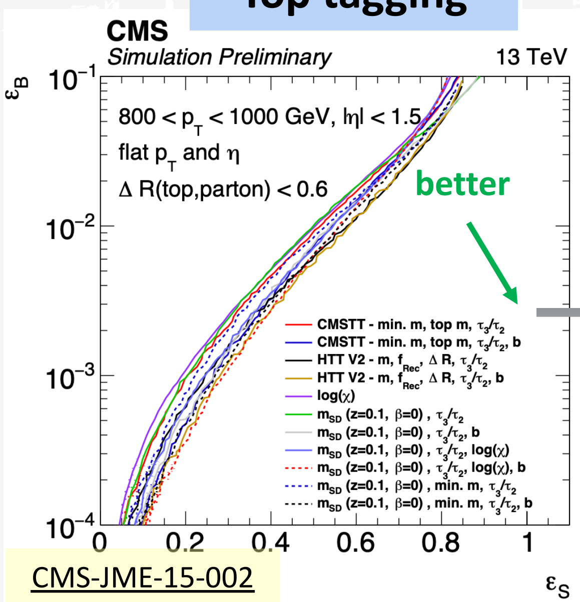
The early days:
Human +inspired, high-level variables
+
Cut-based or simple ML

The "Game changer":
Inputs: low-level info
+
Advanced ML (CNN, RNN)

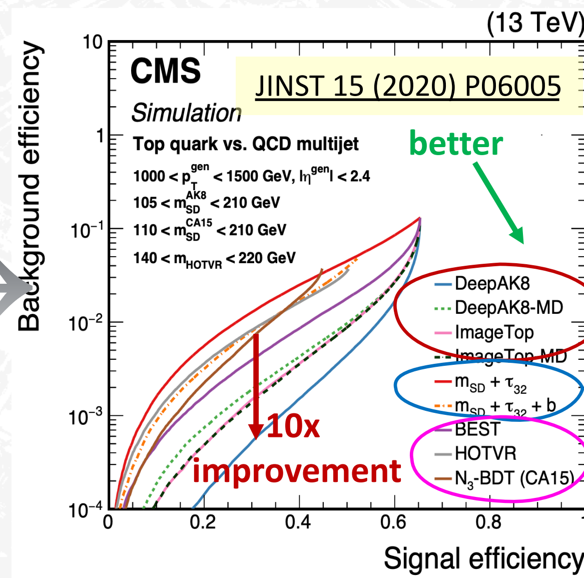
Pushing the limits further:
Inputs: low-level info
+
Graph Neural Networks

Run 1 **Run 2** **Run 3**

Top tagging



CMS-JME-15-002



JINST 15 (2020) P06005

Advanced ML

Cut-based

Simple ML

?

Historical overview [in CMS]

- A topic of high interest in both TH and EXP communities
  - Enormou … LHC start

Focus for today

The early days:
Human +inspired,
high-level variables
+
Cut-based or shallow ML

The "Game changer":
Inputs: low-level info
+
Advanced ML (CNN, RNN)

Pushing the limits further:
Inputs: low-level info
+
Graph Neural Networks

Run 1 **Run 2** **Run 3**

Top tagging

Advanced ML
Cut-based
Simple ML

?

Pushing the limits in jet tagging using Graph Neural Networks

[How] Can we do better?

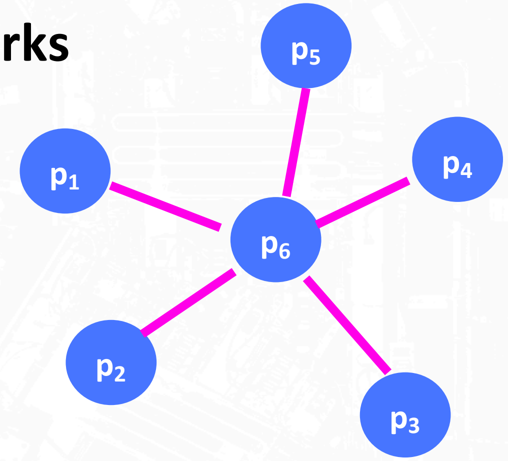
- **Jet as particle sequence:** striking improvement in jet tagging performance
  - **Important limitation:** Must impose a "human-chosen" ordering [in p_T, displacement, etc..]
  - **However**, a Jet is an intrinsically unordered set of particles with relationships between the particles
- Beyond sequences: **Point clouds**

- A very active research area in ML community
- A set of unordered data points in space (x,y,z) with no fixed structure
- Points "close" in space represent physical objects

ParticleNet for jet tagging

- Improve jet representation: *"**Particle Sequences**"* → *"**Particle Clouds**"*
  - Treat the jet as an unordered set of particles
  - Rich set of information per particle [same ones as for DeepAK8]
    - can be "viewed" as the coordinates of each particle in an abstract space

- Improved Network architecture: **Graph Neural Networks**
  - Particle cloud represented as a graph
    - Each particle: **vertex** of the graph
    - Connections between particles: the **edges**

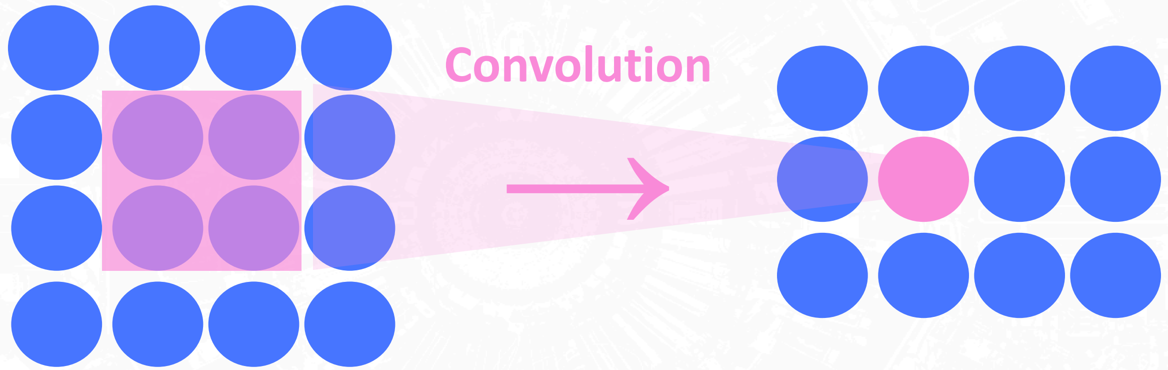


- **Build** the graph:
  - One approach: Fully connected Graph [but computationally very expensive]
  - Another possibility: apply some criteria
    - e.g., ParticleNet uses *k*-Nearest Neighbors (*k*NN)

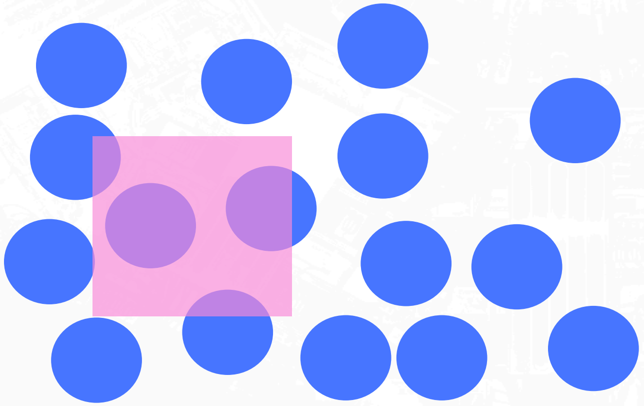
ParticleNet for jet tagging (II)

- Last step: **Learn** from the graphs
  - Follow a **hierarchical learning** approach:
    First learn <u>local</u> structures and **then more <u>global</u>** ones
- Convolution operations proven to be very powerful

Fixed grid:



point/particle cloud:

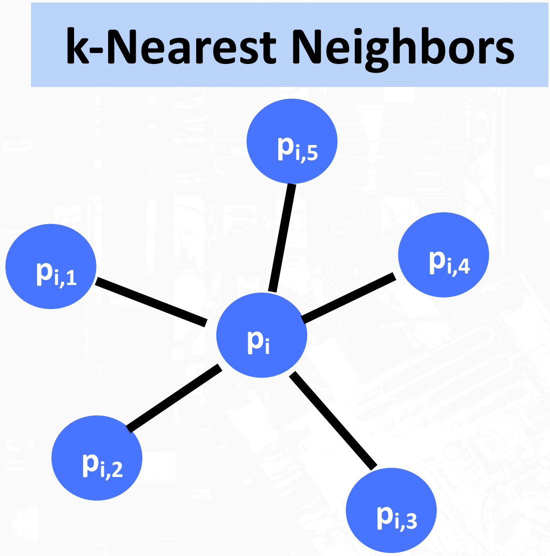


… but not straightforward on point/particle clouds
- Irregular and unordered sets
- Requires a permutation invariant convolutional operation

EdgeConv: Convolution on point clouds

Y. Wang et al.

- Find the ***k*-nearest neighbors** of each point

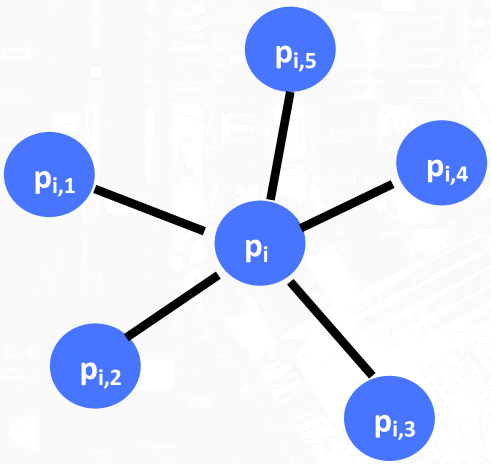


EdgeConv: Convolution on point clouds

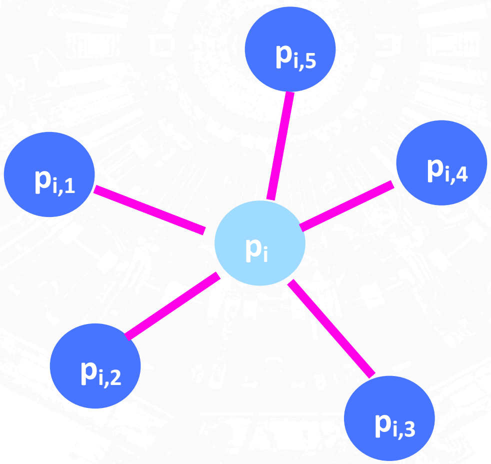
Y. Wang et al.

- Find the ***k*-nearest neighbors** of each point
- Design a permutation invariant **convolution operation**
  - Define an **edge feature** function → **aggregate** edge features w/ a symmetric func.

k-Nearest Neighbors



Convolution operation



- In a nutshell:

$$p_i' = \underset{j=1}{\overset{k}{\square}} h_\Theta(p_i, p_{ij} - p_i)$$

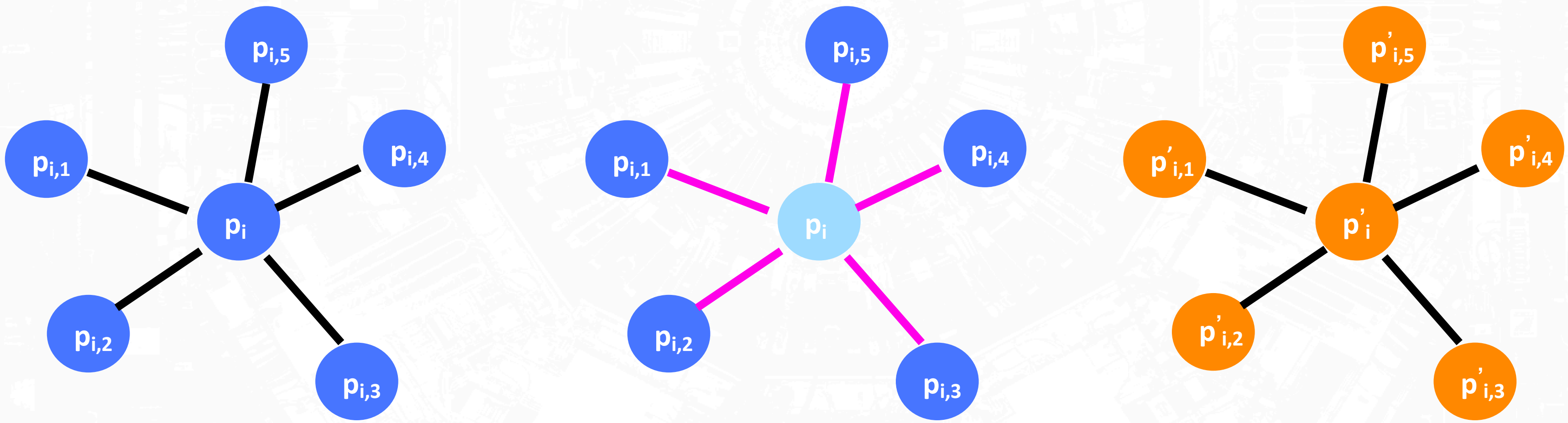
ParticleNet:
h_Θ: MLP [shared across edges]
$\square$ **: average over all *k*-NN**

EdgeConv: Convolution on point clouds

Y. Wang et al.

- Find the ***k*-nearest neighbors** of each point
- Design a permutation invariant **convolution operation**
  - Define an **edge feature** function → **aggregate** edge features w/ a symmetric func.
- **Update Graph (ie Dynamic Graph CNN, DGCNN):**
  Using *k*NN in the feature space produced after EdgeConv
  - Can be viewed as a mapping from one particle cloud to another

k-Nearest Neighbors | **Convolution operation** | **Update Graph**



- In a nutshell:

$$p'_i = \mathop{\square}_{j=1}^{k} h_\Theta(p_i, p_{ij} - p_i)$$

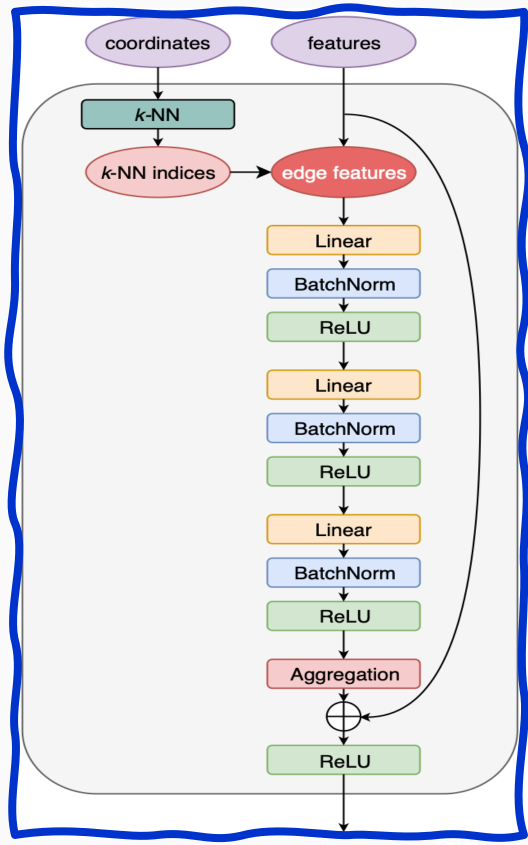
ParticleNet:
h_Θ: MLP [shared across edges]
$\square$ **: average over all *k*-NN**

ParticleNet architecture

H. Qu and LG
PRD 101 056019 (2020)

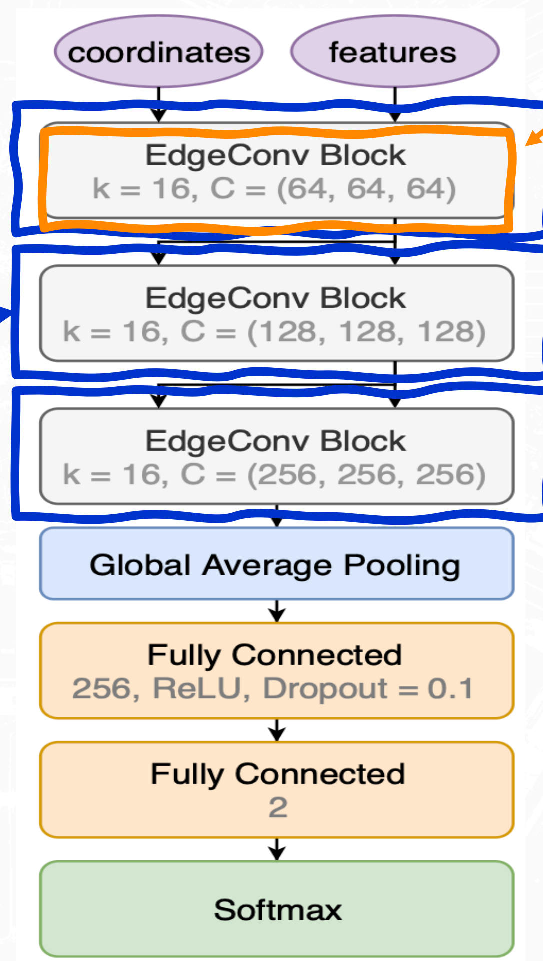
- Based on EdgeConv and DGCNN
  - but customized for the jet tagging task

EdgeConv block



Introduced:
- features beyond spatial coordinates
- residual connections
- MLP conf.

ParticleNet Architecture



particles distributed in η-ϕ

From **local** to more **global** structures

Performance comparison

- Comparison against various DL-based jet tagging algorithms
  - tested on a common top-tagging dataset

G. Kasieczka et al.
SciPost Phys. 7, 014 (2019)

	AUC	Acc	$1/\epsilon_B$ ($\epsilon_S = 0.3$) single	mean	median	#Param
CNN [16]	0.981	0.930	914±14	995±15	975±18	610k
ResNeXt [31]	0.984	0.936	1122±47	1270±28	1286±31	1.46M
TopoDNN [18]	0.972	0.916	295±5	382± 5	378 ± 8	59k
Multi-body N-subjettiness 6 [24]	0.979	0.922	792±18	798±12	808±13	57k
Multi-body N-subjettiness 8 [24]	0.981	0.929	867±15	918±20	926±18	58k
TreeNiN [43]	0.982	0.933	1025±11	1202±23	1188±24	34k
DeepAK8 P-CNN	0.980	0.930	732±24	845±13	834±14	348k
ParticleNet [47] (v1)	0.985	0.938	1298±46	1412±45	1393±41	498k
LBN [19]	0.981	0.931	836±17	859±67	966±20	705k
LoLa [22]	0.980	0.929	722±17	768±11	765±11	127k
LDA [54]	0.955	0.892	151±0.4	151.5±0.5	151.7±0.4	184k
Energy Flow Polynomials [21]	0.980	0.932	384			1k
Energy Flow Network [23]	0.979	0.927	633±31	729±13	726±11	82k
Particle Flow Network [23]	0.982	0.932	891±18	1063±21	1052±29	82k
Ensemble of all taggers: GoaT	0.985	0.939	1368±140		1549±208	35k
ParticleNet	**0.986**	**0.940**	**1615 ± 93**			**366k**

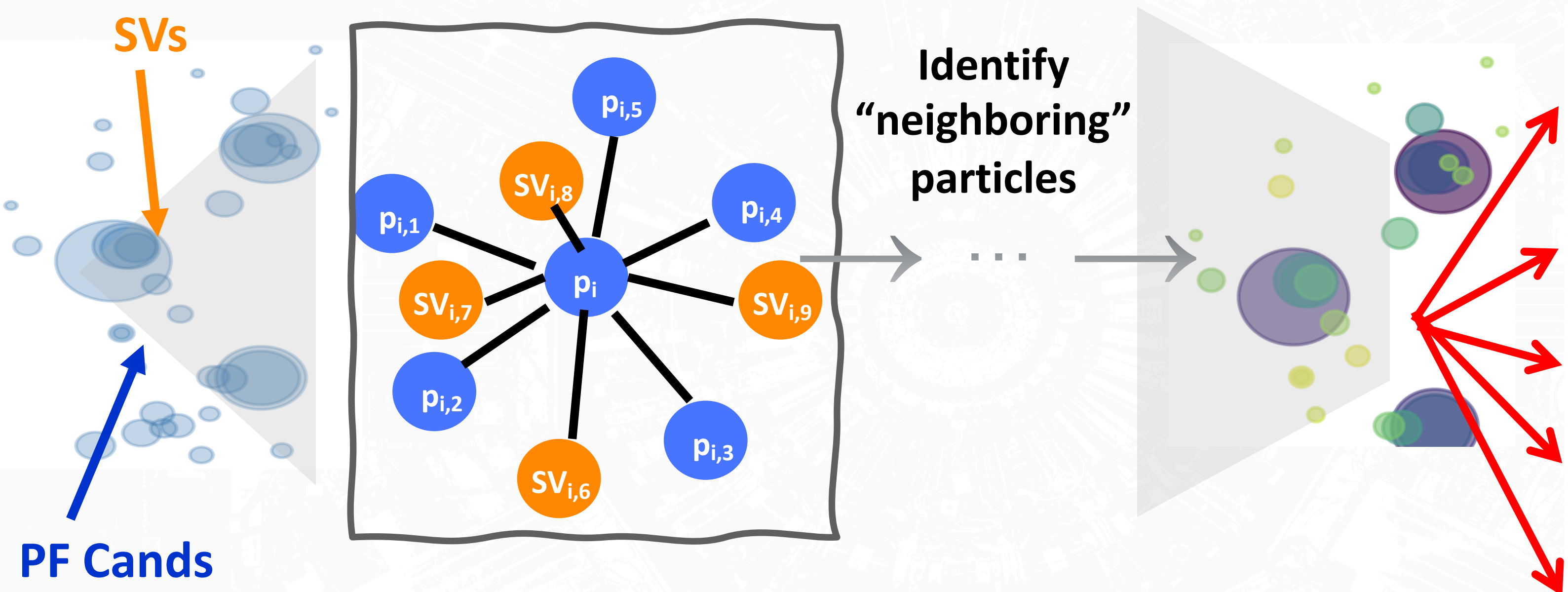
Strong improvement in performance

Jet tagging with ParticleNet in CMS analyses

ParticleNet @ CMS

CMS-DP-2020-002

- Similar ParticleNet architecture used for CMS
  - Same inputs as DeepAK8: **PF candidates** and **Secondary vertices (SVs)**
  - Same multiclass output as for DeepAK8

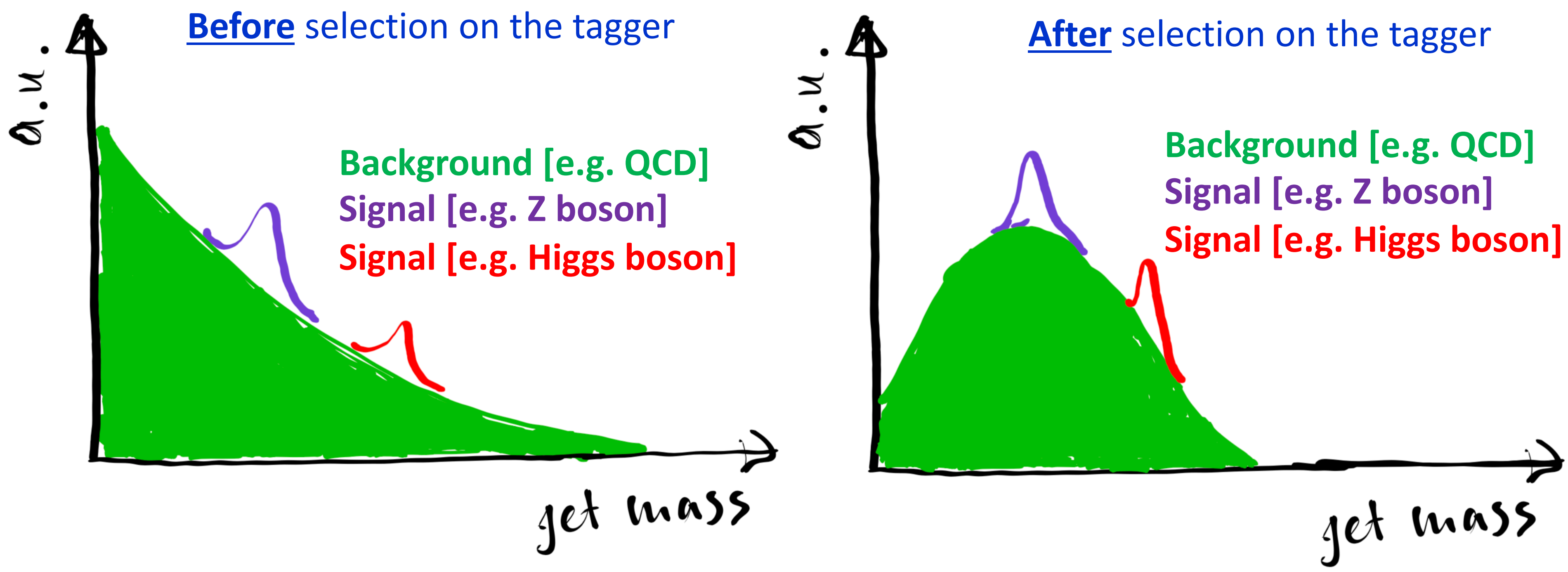


Category	Label
Top	bcq bqq bq cq
Higgs	bb cc VV*→qqqq
Z	bb cc qq
W	cq qq
QCD	g→bb g→cc b c others

- ... but first we need to tackle a few more things:
  - (de-) correlation with jet mass
  - Calibration using data

Jet mass (de-)correlation

- DL-based jet taggers correlated with jet mass



Jet Mass distribution in **Background** jets becomes similar to that from signal: **"Mass sculpting"**

- Depending on the analysis this may not be welcome

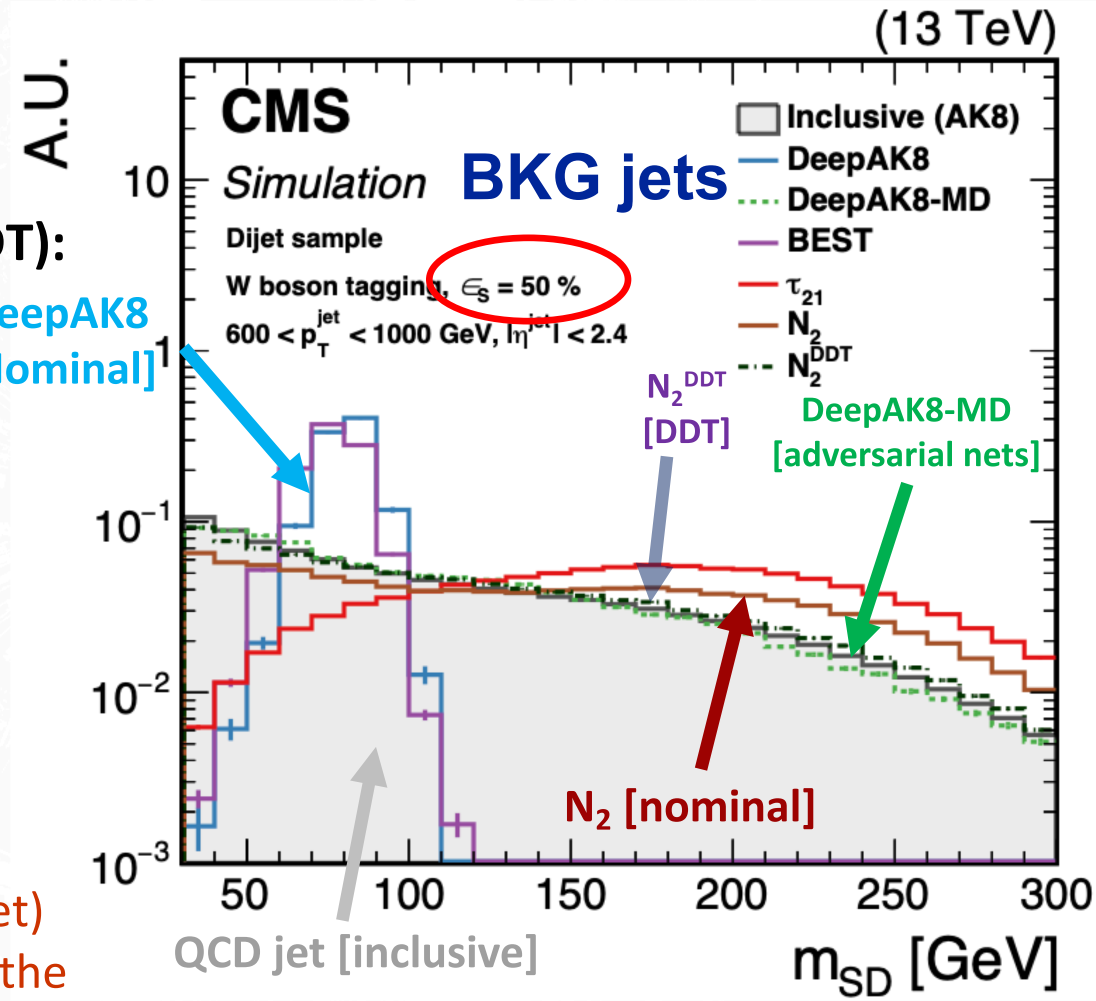
Jet mass (de-)correlation methods

JINST 15 (2020) P06005

- Several jet mass decorrelation techniques explored so far
- **Sample reweighting:**
  - reweight QCD m(j) to match the signal one
- Design **D**ecorrelated **T**agger **(DDT):**
  - Define a metric e.g.,

$$\boxed{\rho = \ln(m_{\mathrm{SD}}^2 / p_{\mathrm{T}}^2)}$$

    to capture correlation m(jet)
  - **Then:** transform response to preserve constant BKG rejection across m(j): **$\mathrm{Tagger}^{\mathrm{DDT}} = \mathrm{Tagger} - X_{(\%)}$**
- **Adversarial networks:**
  - Introduce an NN to predict m(jet) from the features extracted by the nominal network;
  - Acts as penalty term in classifier network

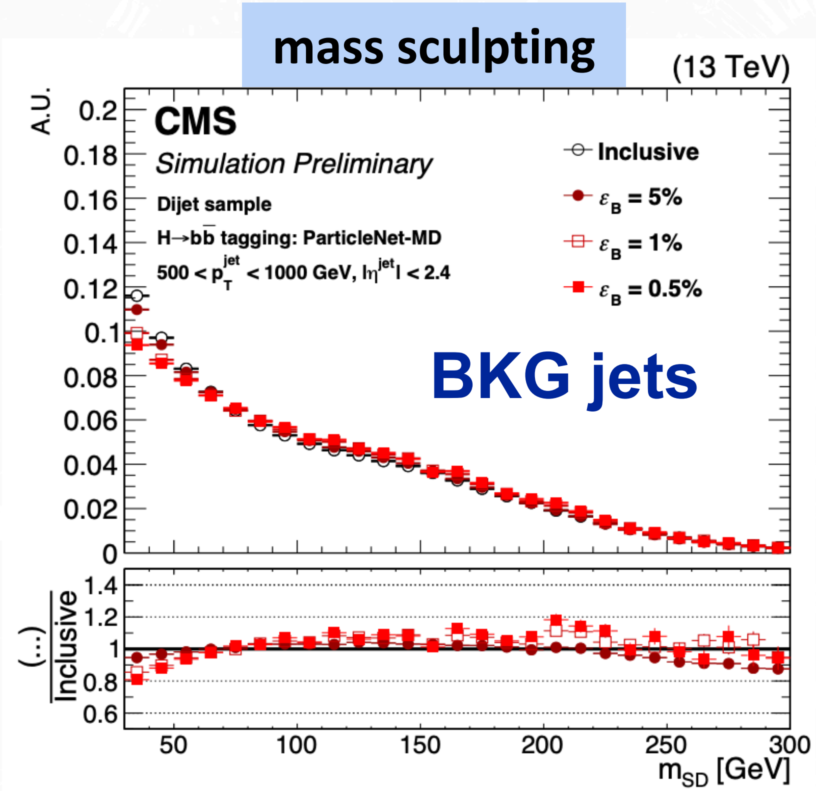


Mass sculpting greatly reduced but not perfect …

Jet mass (de-)correlation methods (II)

CMS-DP-2020-002

- A new jet mass de-correlation method for 2-prong tagging
  - Developed in the context of ParticleNet [but applicable to any DL-based tagger]
- Strategy:
  - Design a dedicated "signal" samples w/ flat m(X) [X: spin-0 particle]
    - hadronic decays of X → bb, cc, qq [on equal fractions]
  - Signal and Background jets re-weighted to a flat distribution in m(jet) and p_T(jet)



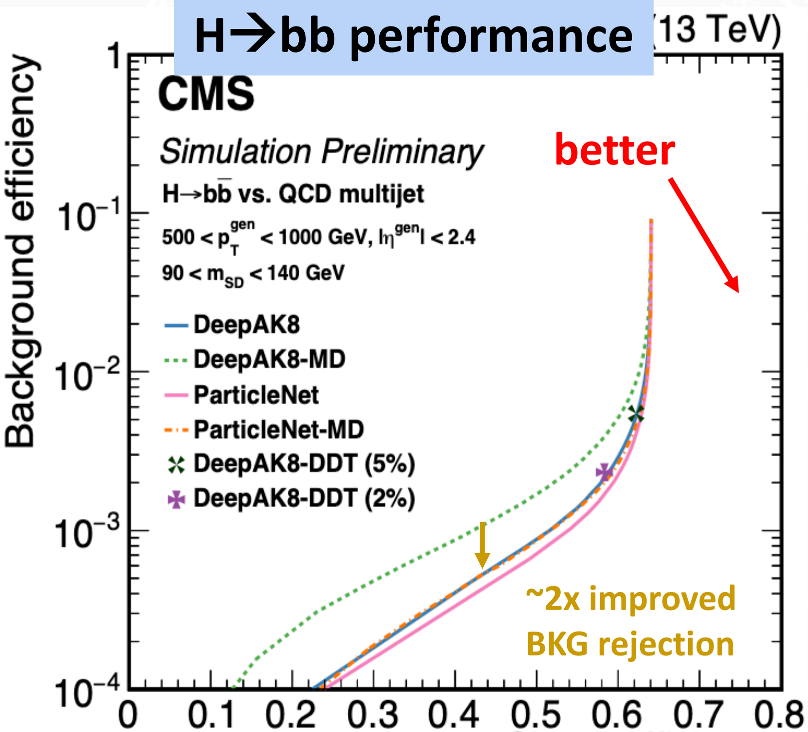
→ **Improved mass decorrelation**
No indication of mass sculpting
- even for very tight WPs

Jet mass (de-)correlation methods (II)

CMS-DP-2020-002

- A new jet mass de-correlation method for 2-prong tagging
  - Developed in the context of ParticleNet [but applicable to any DL-based tagger]
- Strategy:
  - Design a dedicated "signal" samples w/ flat m(X) [X: spin-0 particle]
    - hadronic decays of X → bb, cc, qq [on equal fractions]
  - Signal and Background jets re-weighted to a flat distribution in m(jet) and p_T(jet)

H→bb performance



MD := mass decorrelated

→ **Improved mass decorrelation**
No indication of mass sculpting
- even for very tight WPs

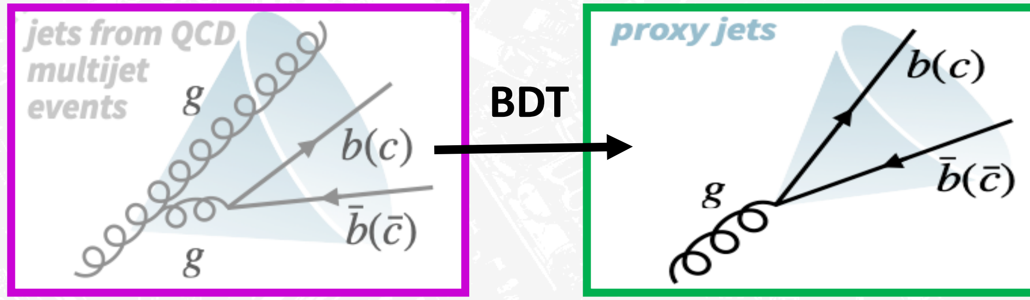
→ **Minimal performance loss compared to the nominal ParticleNet**
- ~2x improved performance compared to DeepAK8—MD (ie adversarial training)

→ **Training significantly easier**

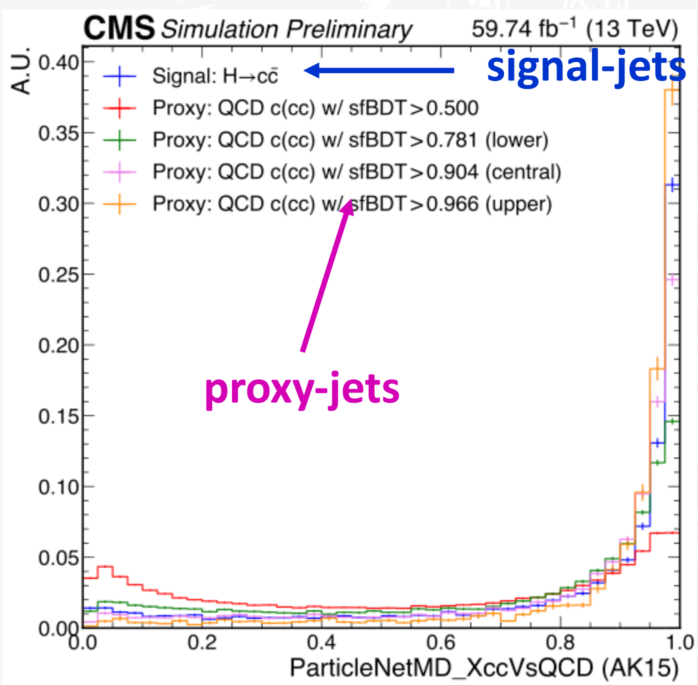
ParticleNet calibration

CMS-DP-2022-005

- **Challenge:** No pure H/Z→bb,cc sample; rely on proxy jets from g→bb,cc
  - Yet: difficult to select proxy jets w/ similar characteristics to signal jets
- **New method:**
  Develop a BDT to distinguish *hard 2-prong* splittings, from *soft-bb/cc* radiations

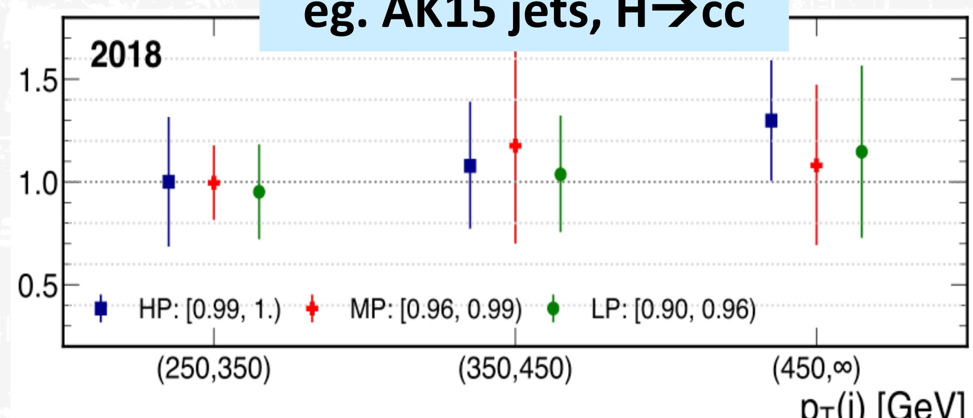


eg. AK15 jets, H→cc

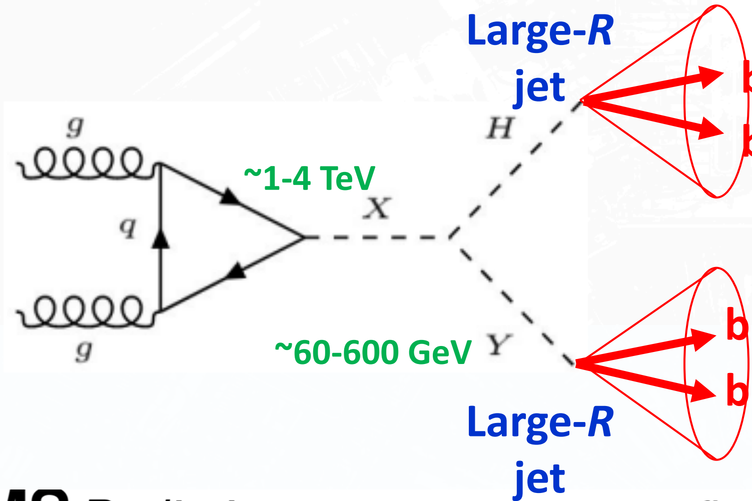


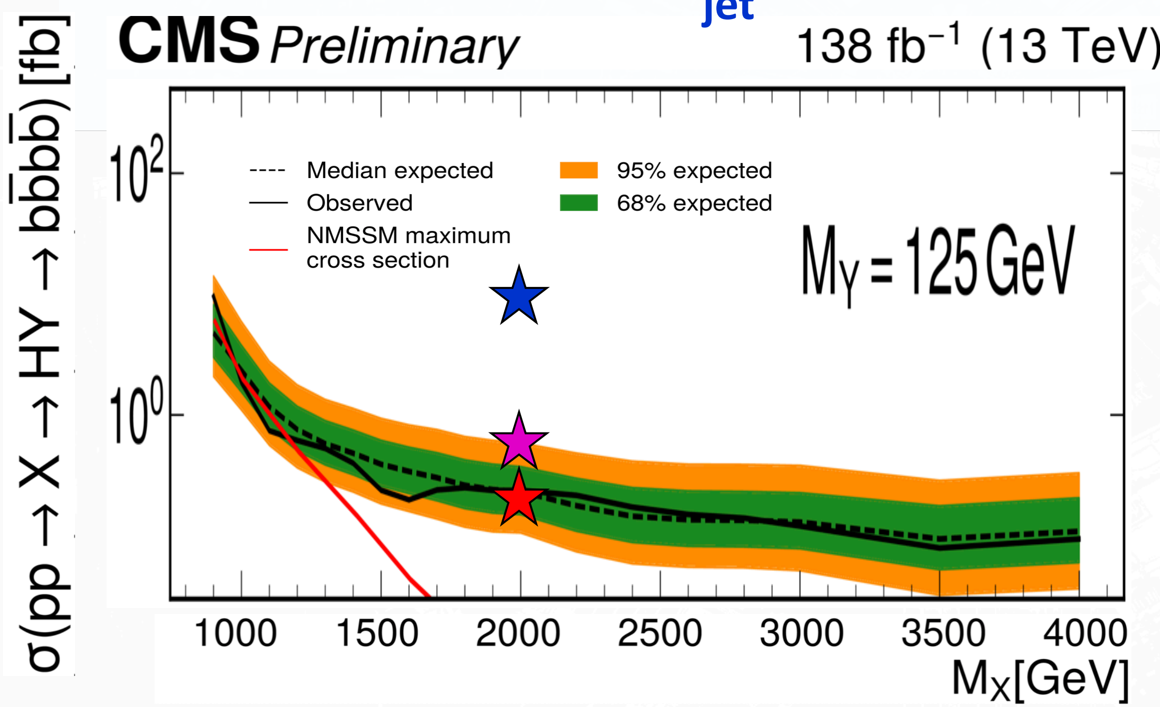
- Extract SFs from simultaneous fit of m(SV) in "pass" and "fail" categories
- For different values of the BDT

eg. AK15 jets, H→cc



Intermezzo: Impact on physics analyses





special case: mY = mH

2016 data [PLB 781 (2018) 244]
bb-tagging:
High-level inputs +BDT
[double-b]

Full Run 2 [CMS-B2G-20-004]
bb-tagging:
Low-level inputs + CNN 1D
[DeepAK8]

Full Run 2 [CMS-B2G-21-003]
bb-tagging:
Low-level inputs + Graphs
[ParticleNet]

Graph Neural Networks in jet physics beyond tagging

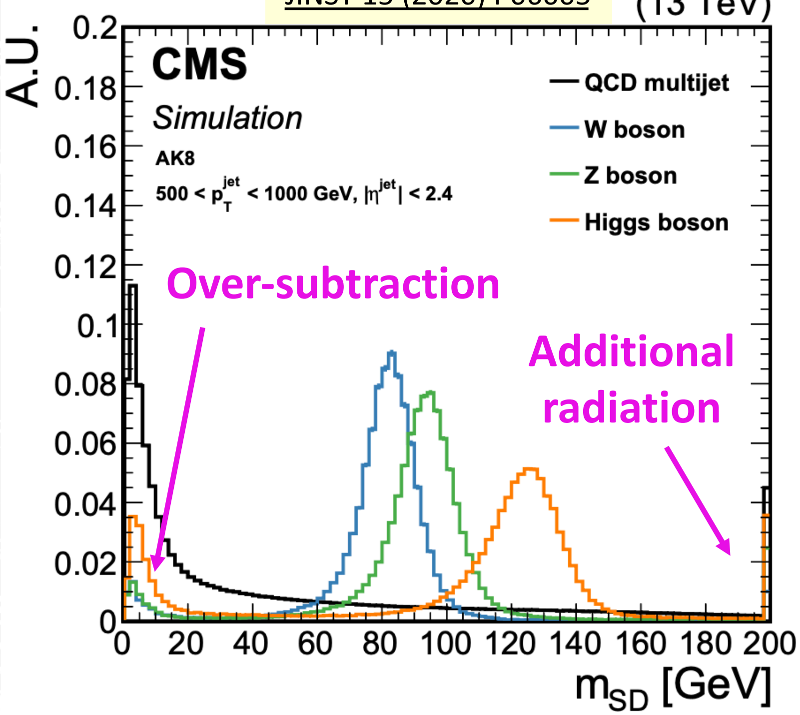
Jet mass regression

- Jet mass: powerful observable to discriminate signal (e.g. H→bb jets) from BKGs [e.g. QCD jets]
  - but very sensitive to soft radiation, pileup, …

Large-R jet

- Grooming techniques [eg, SoftDrop] were developed to mitigate this effect:
  - Iteratively decluster the jet and remove constituents that are:
    - soft and/or wide angle
  - Pros: simple and well tested in data
  - Cons: some inefficiency
    - e.g., some two prong jet identified as 1-prong
- Decays to bb/cc:
  - additional energy loss via the (undetected) neutrinos from semileptonic decays

JINST 15 (2020) P06005



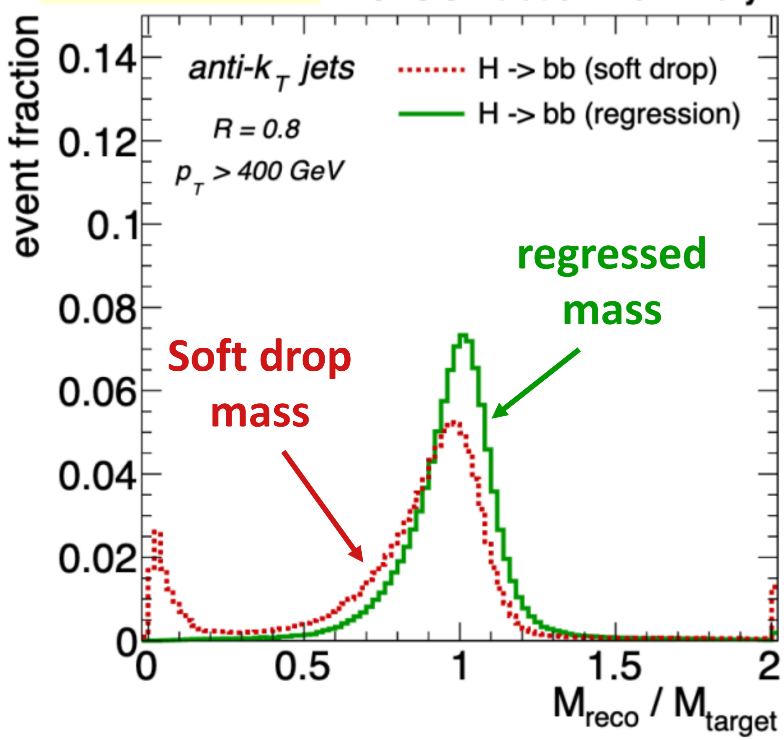
Jet mass regression (II)

- Develop algorithm to reconstruct jet mass with best possible scale & resolution
  - Meanwhile: avoid "sculpting" of the QCD jet mass distribution
- Exploit **ParticleNet** architecture to predict m(jet) directly from jet constituents
  - Same inputs (PF candidates + SV)
  - Same network architecture
  - Same training samples as for jet tagging
- Training details:
  - Target mass:
    - **Signal:** pole mass of spin-0 particle
    - **QCD:** Generated soft-drop mass
  - Loss function:

$$L(y, y^p) = \sum_{i=1}^{n} \log(\cosh(y_i^p - y_i))$$

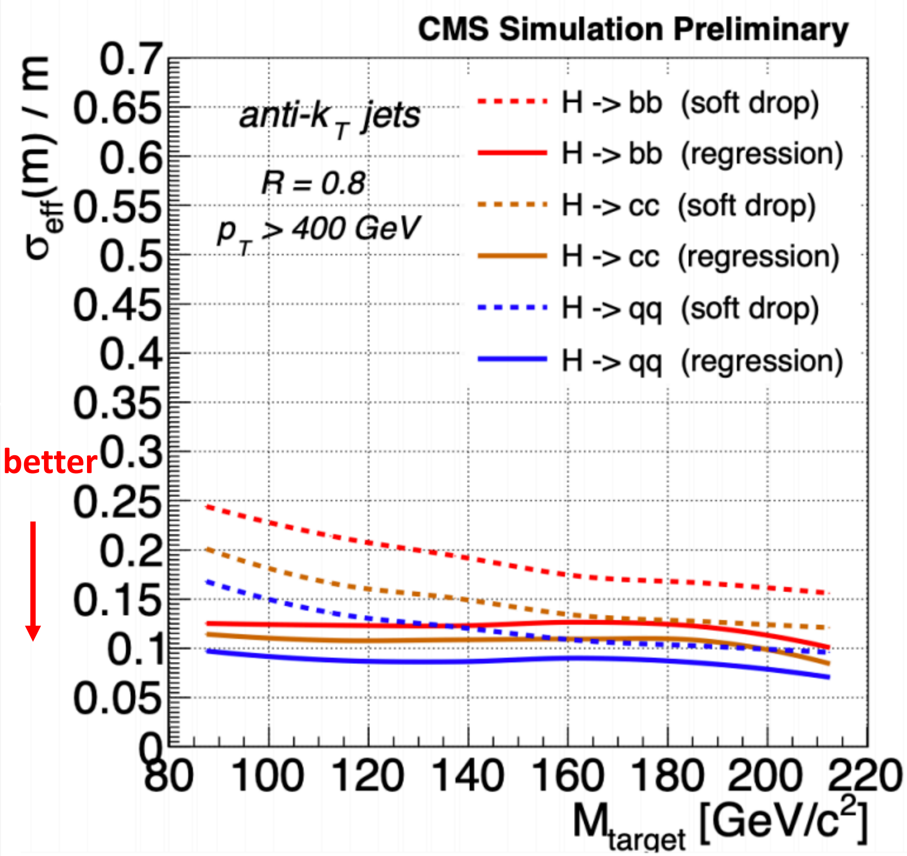
- Substantial improvement in both mass scale and mass resolution
- Tails in m(SD) significantly reduced

CMS-DP-2021-017

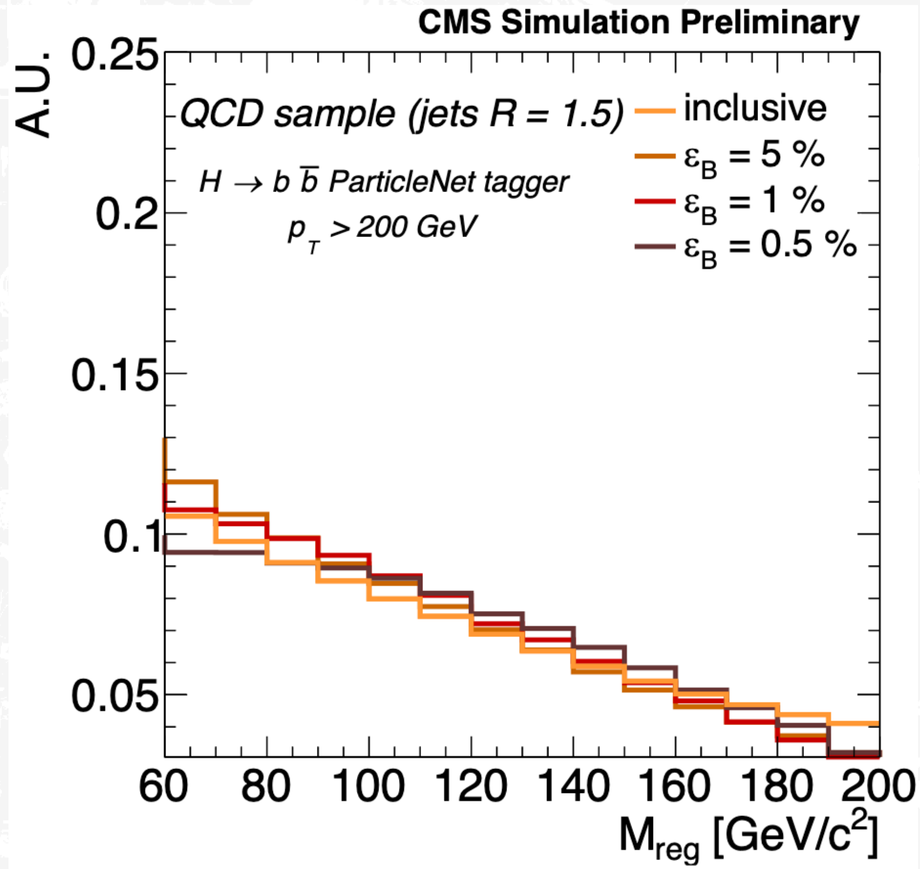


Jet mass regression: Performance

Mass resolution vs. m(X)



Regressed Mass vs. Tagger WP



- >2x improvement in mass resolution
- No indication of mass sculpting – even for very tight WPs
- Calibration using W jets: scale (resolution) correction < 1% (3%)

Putting pieces together: Highlights from CMS analyses

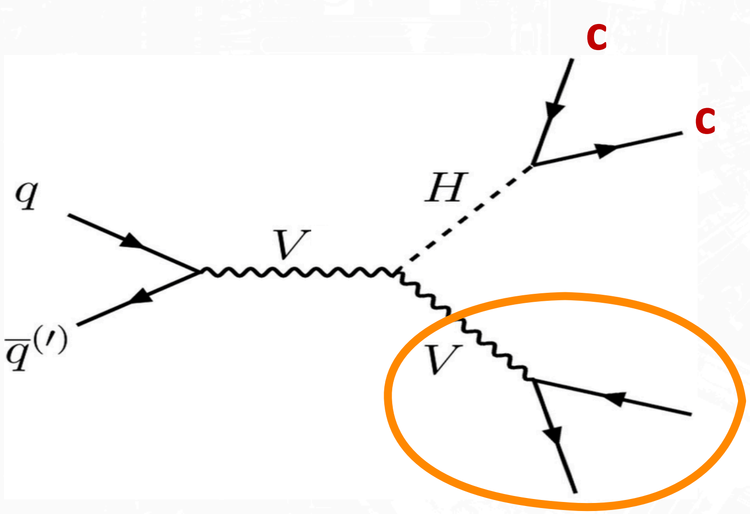
The search for H→cc in CMS

CMS-HIG-21-008

- Important priority for the Higgs program: measure couplings to 2nd-G fermions
  - but very [very..] challenging at the LHC
- Similar concept to H→bb but two huge challenges:
  - Much smaller signal
  - Charm –tagging more challenging than b-tagging

Need novel tools and techniques [& HL-LHC]

- The H→cc search in VH production



Target leptonic decays of V boson
- Suppress QCD background

The search for H→cc in CMS

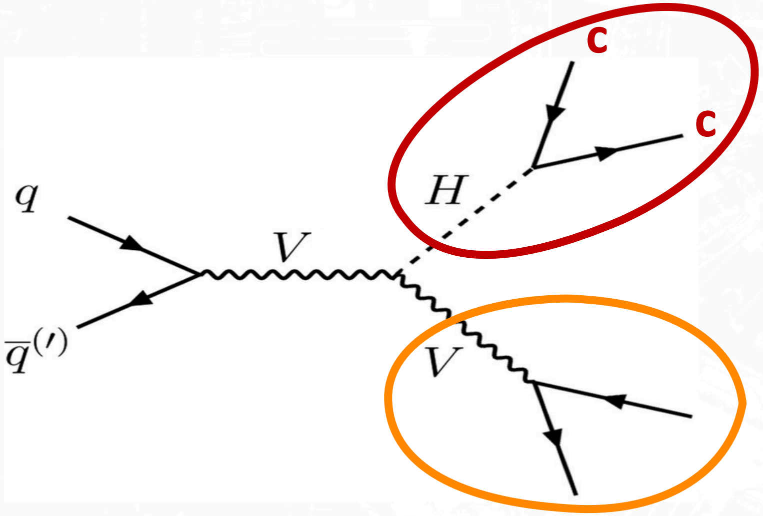
CMS-HIG-21-008

- Important priority for the Higgs program: measure couplings to 2nd-G fermions
  - but very [very..] challenging at the LHC
- Similar concept to H→bb but two huge challenges:
  - Much smaller signal
  - Charm –tagging more challenging than b-tagging

Need novel tools and techniques [& HL-LHC]

- The H→cc search in VH production

Fully explore the Higgs decay topology

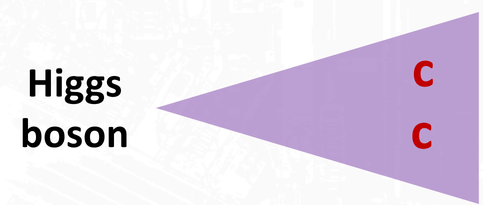


Target leptonic decays of V boson
- Suppress QCD background

Resolved-jet topology

→ Higgs candidate: two **small-*R*** jets
→ Larger fraction of σ(VH→cc) [w/ higher BKGs]

Merged-jet topology

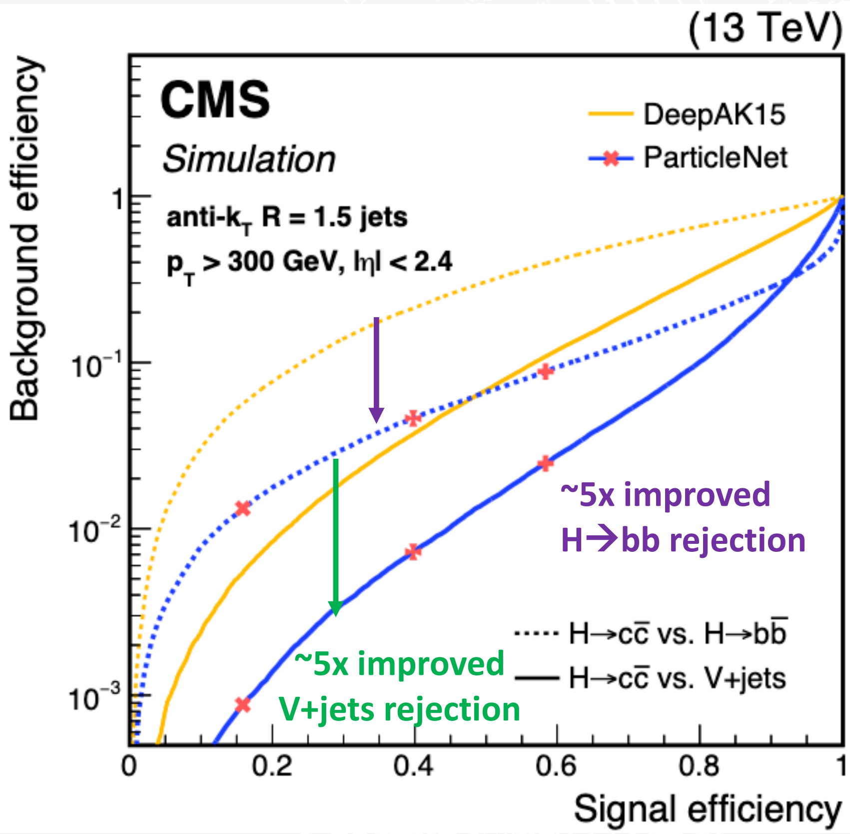


→ Single **large-*R*** (R=1.5) jet for H→cc
→ Better exploit correlations b/w two c-quarks
→ Use of state-of-the-art jet tools [ParticleNet]

H→cc: Impact of jet tagging developments

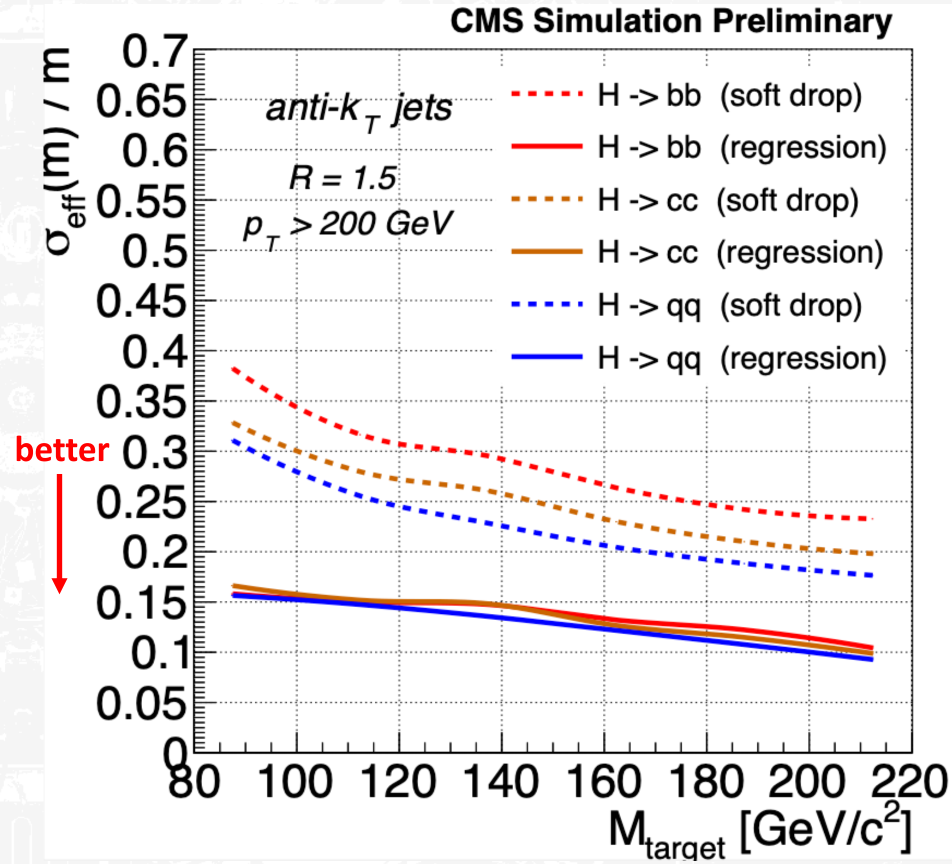
CMS-DP-2020-002
CMS-HIG-21-008

ParticleNet for H→cc tagging



>2x improvement in analysis sensitivity

ParticleNet for m(j) regression

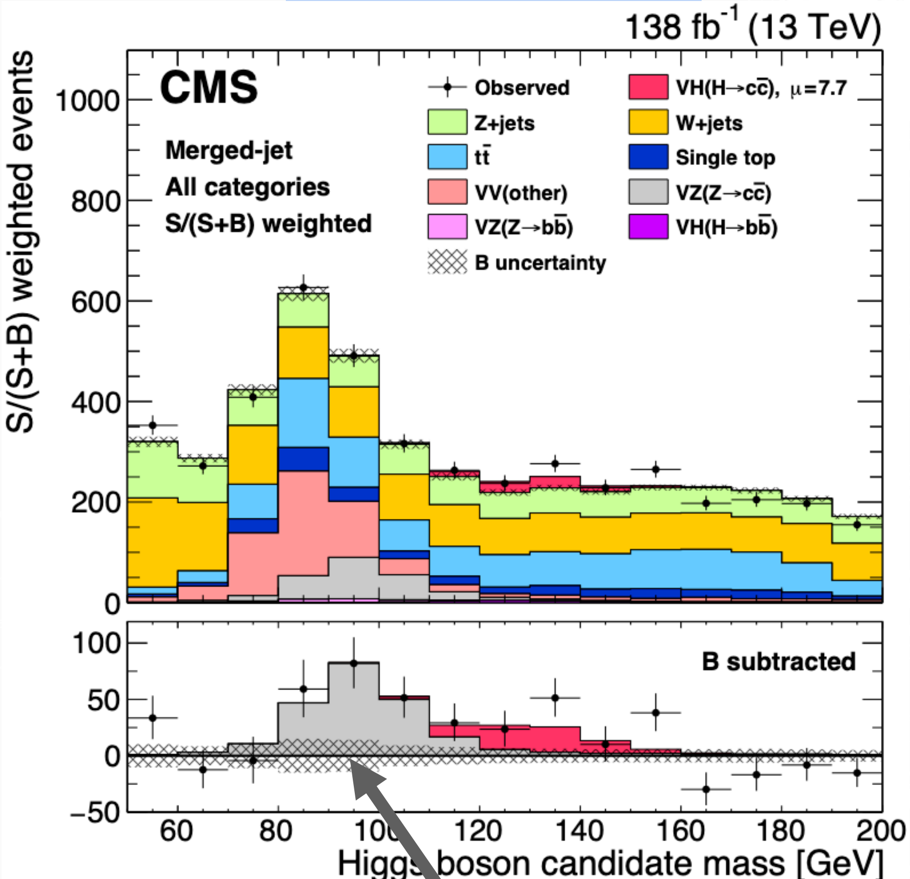


>20% improvement in analysis sensitivity

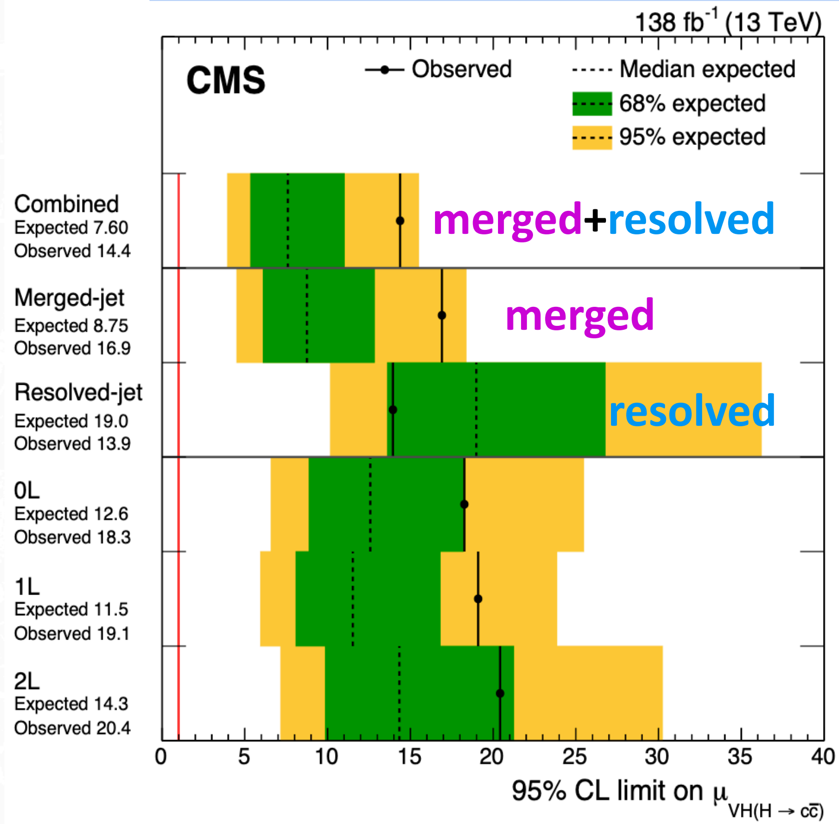
H→cc results

CMS-HIG-21-008

m_H [merged-jet]



95% UL on signal strength (μ)



<u>Technique proven on VZ(→cc)</u>

5.7σ (5.9σ) Obs (Exp)

$\mu = 1.01 \pm 0.22$

[1st observation at hadron colliders]

UL: 14.4 (7.6) Obs (Exp)

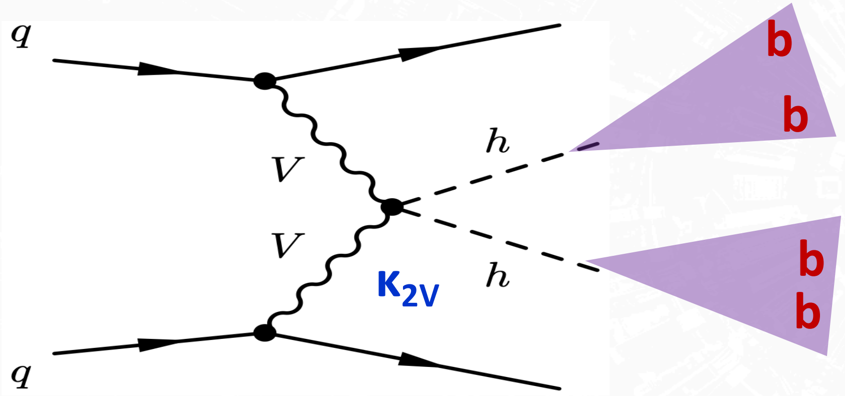
$1.1<\kappa_c<5.5$ ($|\kappa_c|<3.4$) Obs (Exp)

Most stringent constraints to date reaching sensitivity expected for HL-LHC

Nonresonant VBF HH→4b production

CMS-B2G-22-003

- Another big priority of the Higgs program: Search for Higgs-pair production
  - Usually searched for in the ggF channel (σ~31 fb)
  - VBF production: powerful probe of BSM physics [but very rare: σ~1.7 fb]
    - provides direct sensitivity to the VVHH (κ_{2V}) coupling

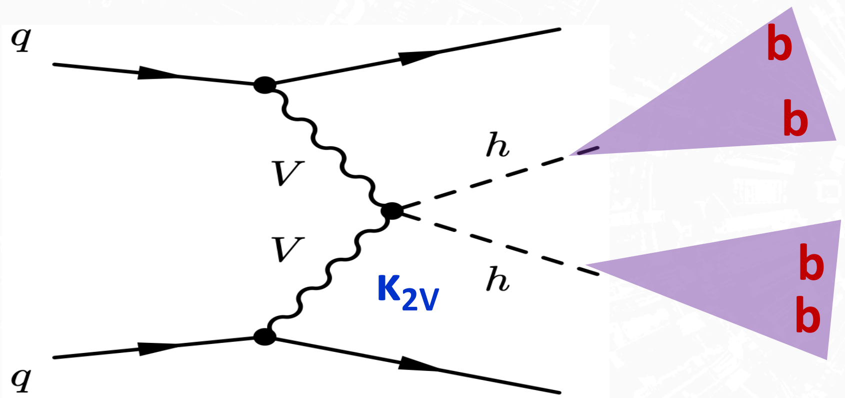


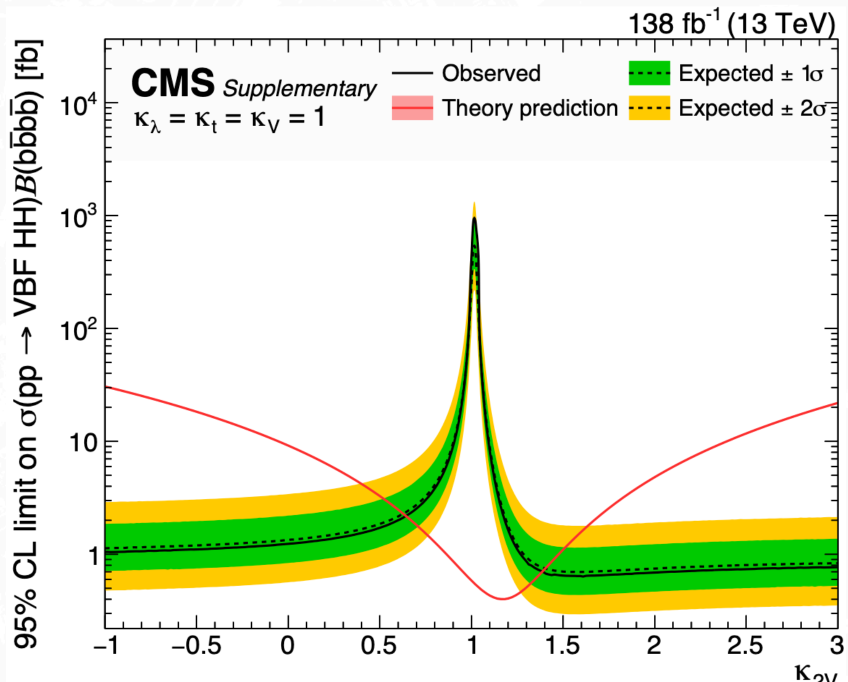
- Cornerstone of the search:
  - Reconstruct each H→bb as a single large-*R* jet
  - Exploit ParticleNet-based tools for jet tagging and mass regression

Nonresonant VBF HH→4b production

CMS-B2G-22-003

- Another big priority of the Higgs program: Search for Higgs-pair production
  - Usually searched for in the ggF channel (σ~31 fb)
  - VBF production: powerful probe of BSM physics [but very rare: σ~1.7 fb]
    - provides direct sensitivity to the VVHH (κ_{2V}) coupling





- Cornerstone of the search:
  - Reconstruct each H→bb as a single large-*R* jet
  - Exploit ParticleNet-based tools for jet tagging and mass regression

- Allowed values: **0.6 < κ_{2V} < 1.4**
- Strongest constraints on κ_{2V} to date:
  - ATLAS: $-0.6<\kappa_{2V}<3.1$ [JHEP07(2020)108]
  - CMS [resolved]: $-0.4<\kappa_{2V}<2.5$ [HIG-19-018]
- 1st time to exclude κ_{2V} = 0 hypothesis

Summary and outlook

Summary and outlook

- Physics with jets essential for the success of the LHC physics program
  - Large effort to improve/extend our jet tools

- **Enormous progress** in the development and performance of jet tools
  - Allows us to see much more of the true potential of the CMS apparatus
  - Key player in these developments: Advanced machine learning algorithms
    - particularly using Graph Neural Networks for inhomogeneous detectors
  - Still room for improvement:
    - on the performance side [more advanced techniques, fresh ideas,..]
    - robustness: calibration, systematics uncertainties [e.g., what the DNN learns?]

- **Effort pays off:** Large gain in performance translates to significant improvement in physics reach
  - even approaching sensitivity expected at HL-LHC [with a much larger dataset]

- Graph-based developments explored beyond Large-*R* jet tagging:
  - jet mass regression, flavour tagging to small-*R* jets, successfully implemented HLT@Run3, etc..
    - And of course beyond jet physics..