Machine Learning approaches for composite model analyses at LHC

F.A. Conventi[1,2] & E. Rossi[1,3]

[1] INFN-Napoli

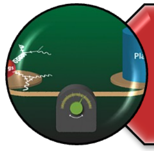
[2] Università degli Studi di Napoli “Parthenope”

[3] Università degli Studi di Napoli “Federico II”

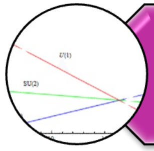
Overview

- Introduction and brief status of the LHC searches on the composite models
- Use of an analysis example based on the Higgs self-coupling measurements to explore Machine Learning tools and where and how they can be used
- Overview of some Machine Learning (ML) tools:
  - Deep Neural Network (DNN)
  - Convolutional Neural Network (CNN)
  - Recurrent Neural Networks (RNN)
  - Graph Neural Network (GNN)
- Extra slides (ML applied to b-tagging)

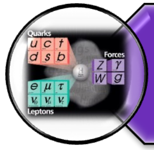
Introduction: SM incompleteness



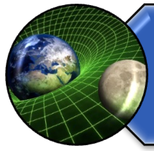
Hierarchy Problem: Why is $M_{Pl}/M_{EW} \sim 10^{17}$



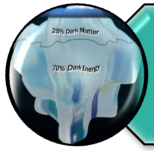
Unification of Gauge couplings: Why are gauge couplings so different, are they unified at a higher scale? Are there more forces in nature?



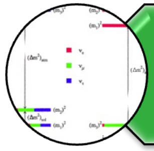
Origin of generations: Why do quarks and leptons come in three generations? Are they elementary particles?



Gravity: SM describes three of the four fundamental interactions at the quantum level (microscopically) but gravity is only treated classically.



Dark matter: What is 25% of the Universe made off, and how does it interact with ordinary matter?



Neutrino masses: What is the origin and nature of neutrino masses?

CP Violation: What is the origin?

Composite models at LHC

- ✦ *Compositeness mechanisms (i.e. confinement) can naturally generate mass scales well below the Planck scale.*
- ✦ *The quarks and leptons in the Standard Model (SM) could be composed of more fundamental particles.*
- ✦ *Constituents are normally SM-charged, but hadrons can be SM-neutral, leading to suppressed interactions.*

Today, there are two mass scales that observations say clearly exist but lack a natural explanation:

- ✓ *Dark matter exists and it should be more massive than neutrinos (cold dark matter, not axions).*
- ✓ *Higgs boson has a mass of 125 GeV but the Standard Model mechanism is severely fine-tuned.*

Higgs discovery

- 2012 higgs boson discovery

Higgs properties

- *After the discovery of the Higgs boson, focus has been given in precisely measuring its properties*

Compositness in Higgs signatures

- *Obvious starting point to explore the compositeness hypothesis is the analysis of Higgs signatures, and of their possible deviations from the predictions of the SM.*

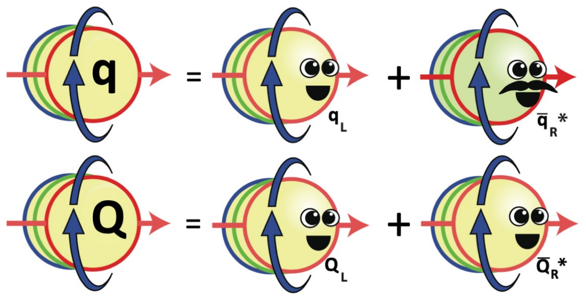
Higgs mass, spin, width and couplings can indeed set constraints on the kind of new physics possibly connected to the Higgs, and in particular, on composite interpretations

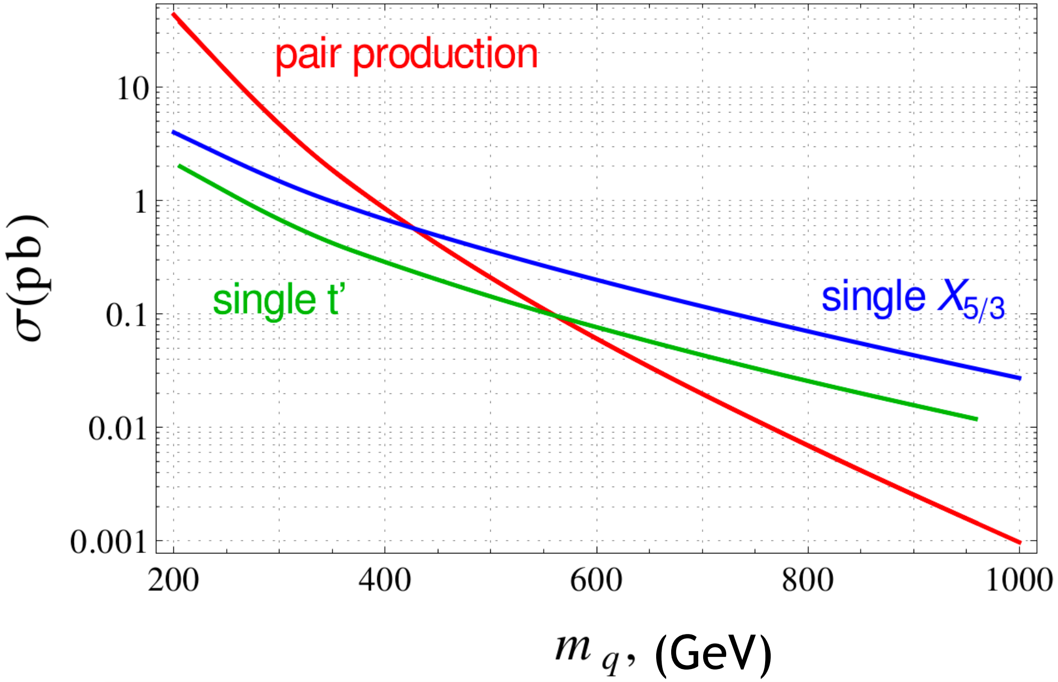
LHC is suitable for the exploration of signatures:

- Heavy Neutral Leptons (HNLs)
- Vector-like quarks (VLQs) and Vector-like leptons (VLLs)
- New heavy mass resonances (spin-0 objects) signatures can be very effective for constraining composite scenarios

Composite-Higgs models and vector-like fermions

- Spin 1/2 particles with color charge
- Left and right chiralities behave the same ➔ Vector-like interaction with weak force
- Mass not from Higgs boson and decay to SM boson and quark
- explains the lack of CP-violation in the strong interaction
- Simplest extensions with VLQ ($T^{2/3}$, $B^{-1/3}$ and $X^{5/3}$) singlets, doublets, and triplets.
- Vector-like quarks have the same mass heirarchy as SM quarks
- *The Higgs boson is a composite pseudo-Nambu-Goldstone boson (pNGB) from spontaneous breaking of a global symmetry in a new strongly coupled sector ➔ This protects the Higgs mass.*





QCD pair-production (via strong interactions): Mass-independent, dominant at low mass
Single-production (via EW interactions): Scales with coupling, model dependent, significant at high mass

Composite Models predicting new vector-like fermions:

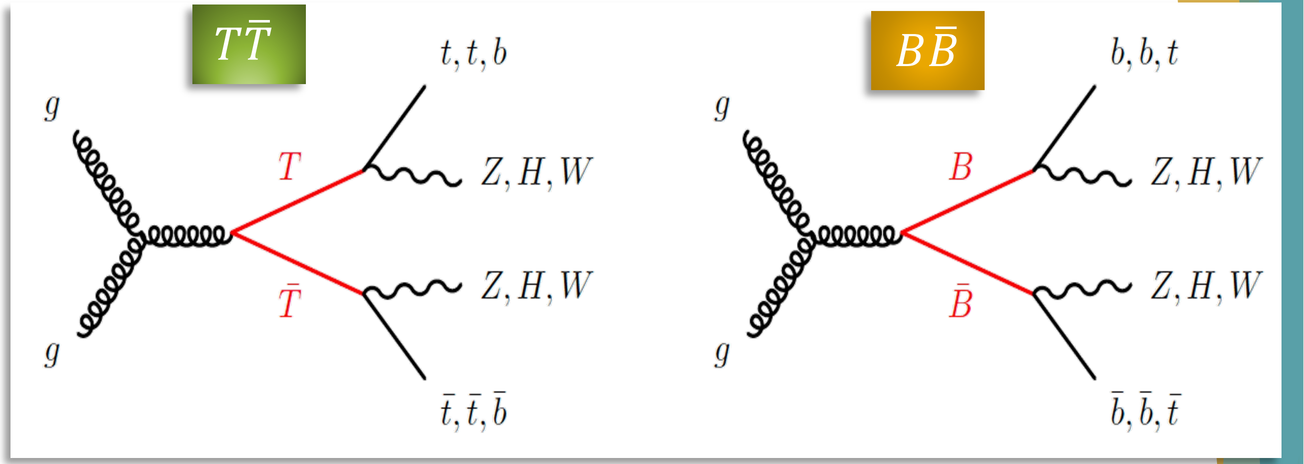
- ***Warped or universal extra-dimensions**: KK excitations of bulk fields*
- ***Composite Higgs models**: VLQ appear as excited resonances of the bounded states which form SM particles*
- ***Little Higgs models**: partners of SM fermions in larger group representations which ensure the cancellation of divergent loops*
- ***Gauged flavour group** with low scale gauge flavour bosons required to cancel anomalies in the gauged flavour symmetry*
- ***Non-minimal SUSY extensions**: VLQs increase corrections to Higgs mass without affecting EWPT*
- Predicted in other models such as the **Left Right Mirror Model Model**

Pair-produced vector-like top and bottom partners in events with large E_T^{miss}

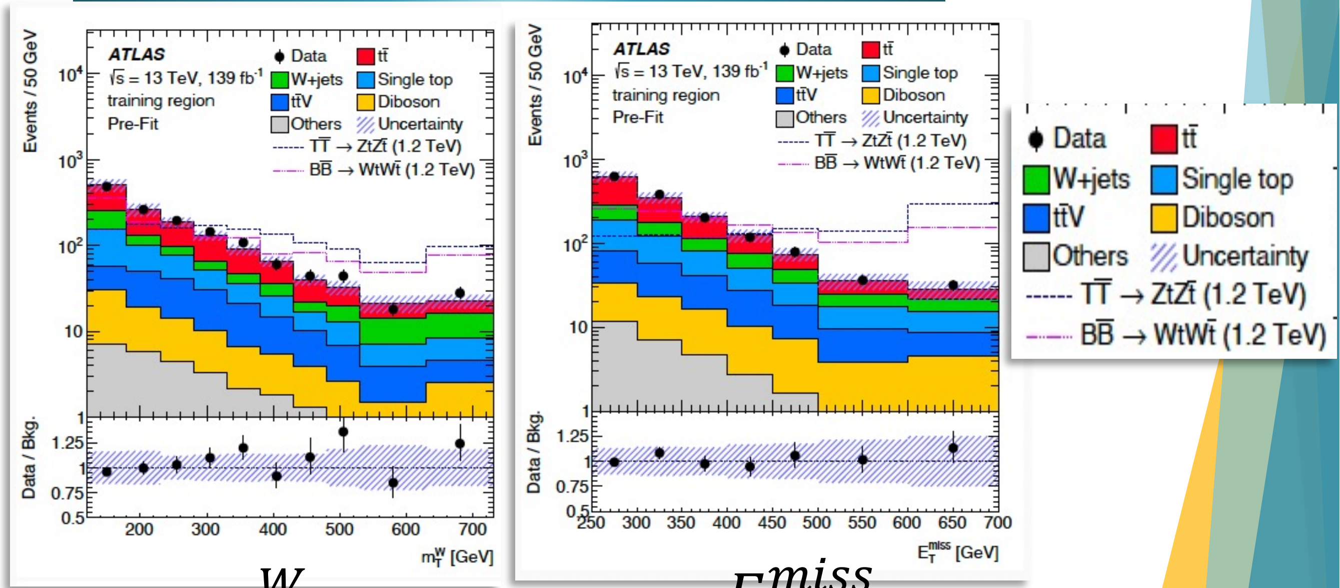
$T\bar{T}$ & $B\bar{B}$ production

Eur. Phys. J. C 83 (2023) 719

- The search uses 139 fb^{-1} data collected with the ATLAS detector
- **Masses of the VLQs >800GeV**
- **Vector-like $T^{2/3}$, $B^{-1/3}$ and $X^{5/3}$ considered**
- **Branching ratios: T:** $\mathcal{B}(Zt; Ht; Wb) \approx (0.25; 0.25; 0.5)$ and **B:** $\mathcal{B}(Zb; Hb; Wt) \approx (0.25; 0.25; 0.5)$
- **Events characterized by low lepton-multiplicity, high jet-multiplicity, and large missing transverse energy (E_T^{miss}):**
- High missing transverse momentum $E_T^{miss} > 250\ GeV$
- Only one lepton $\ell(e\ or\ \mu)$ ➔ veto for a second lepton
- At least 4 jets including a b-tagged jet
- *At least one top quark from the signal expected to have a high p_T ➔ requirement on large-R jets*
- **Dominant backgrounds: $t\bar{t}$ and W+jets** ➔ reduced using cuts on transverse mass; Other BKGs: $t\bar{t}H$, tWZ and $Z + jets$
- **Systematic Uncertainties resolution and scale of:** $t\bar{t}$ background, Jet mass, Effciency of lepton identification, isolation, reconstruction and energy



Examples of discriminating variables



m_T^W E_T^{miss}

Pair-produced vector-like top and bottom partners in events with large Emiss

$T\bar{T}$ & $B\bar{B}$ production

Eur. Phys. J. C 83 (2023) 719

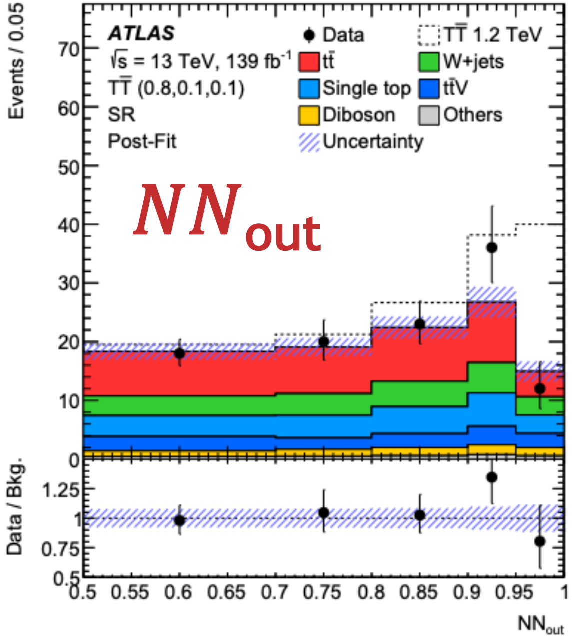
Neural networks (NN) used to discriminate between signal and background

- Input variables such as high m_{eff} for VLQ mass, properties of large-R jets, b-jet multiplicity, transverse mass etc. used
- The NNs are implemented as a three-layer feedforward NN with one input node for each variable, 15 nodes in the hidden layers and one output node which gives a continuous NN output score (NN_{out})
- Training on different Branching Ratios:
  - For TT 4NN: $(0.8; 0.1; 0.1)$; $(0.2; 0.4; 0.4)$; $(0.4; 0.1; 0.5)$; $(0.4; 0.5; 0.1)$
  - For BB 3NN: $(0.1; 0.1; 0.8)$; $(0.4; 0.1; 0.5)$ and $(0.1; 0.4; 0.5)$;
  - All the main backgrounds are used in the training

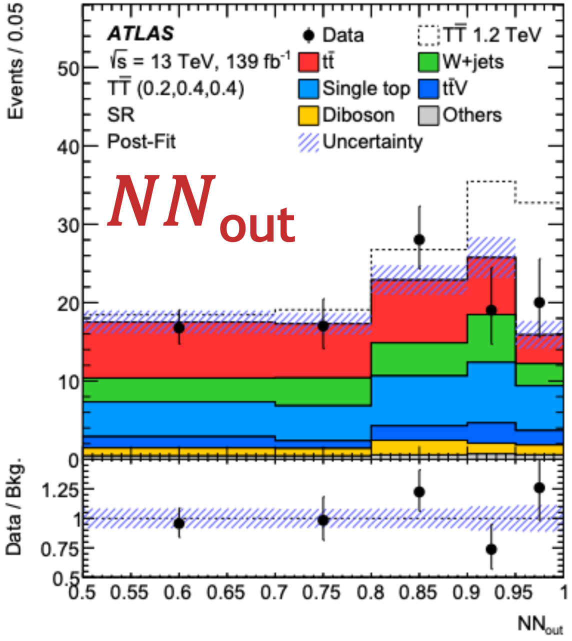
NN input variables

Variable	Description
m_{eff}	scalar sum of the transverse momenta of leptons, jets, and $E_{\text{T}}^{\text{miss}}$
$N_{b\text{-jets}}$	b-jet multiplicity
m_{T}^{W}	transverse mass of lepton and $E_{\text{T}}^{\text{miss}}$
am_{T2}	asymmetric transverse mass
$p_{\text{T}}(\text{large-}R\ \text{jet}_2)$	transverse momentum of second-highest-p_{T} large-R jet
$\lvert\Delta\phi(\text{jet}_1, E_{\text{T}}^{\text{miss}})\rvert$	azimuthal angle between $E_{\text{T}}^{\text{miss}}$ and highest-p_{T} jet
$E_{\text{T}}^{\text{miss}}$	missing transverse momentum
$\eta(\text{jet}_1)$	pseudorapidity of highest-p_{T} jet
$m(\text{large-}R\ \text{jet}_1)$	mass of highest-p_{T} large-R jet
$N_{\text{const}}(\text{large-}R\ \text{jet}_1)$	number of small-R jets reclustered to the highest-p_{T} large-R jet
$p_{\text{T}}(\ell)$	transverse momentum of lepton
$p_{\text{T}}(\text{jet}_3)$	transverse momentum of third-highest-p_{T} jet
$p_{\text{T}}(\text{jet}_2)$	transverse momentum of second-highest-p_{T} jet

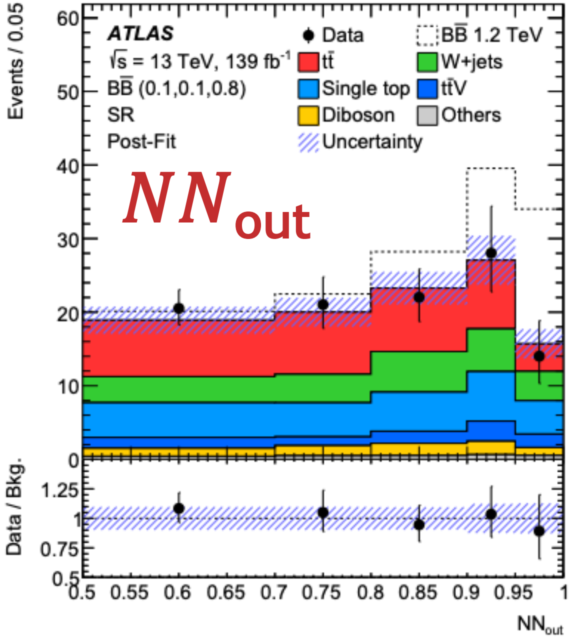
$T\bar{T}$ signal of mass 1.2 TeV
B (Zt, Ht, Wb) = (0.8, 0.1, 0.1)



$T\bar{T}$ signal of mass 1.2 TeV
B (Zt, Ht, Wb) = (0.2, 0.4, 0.4)

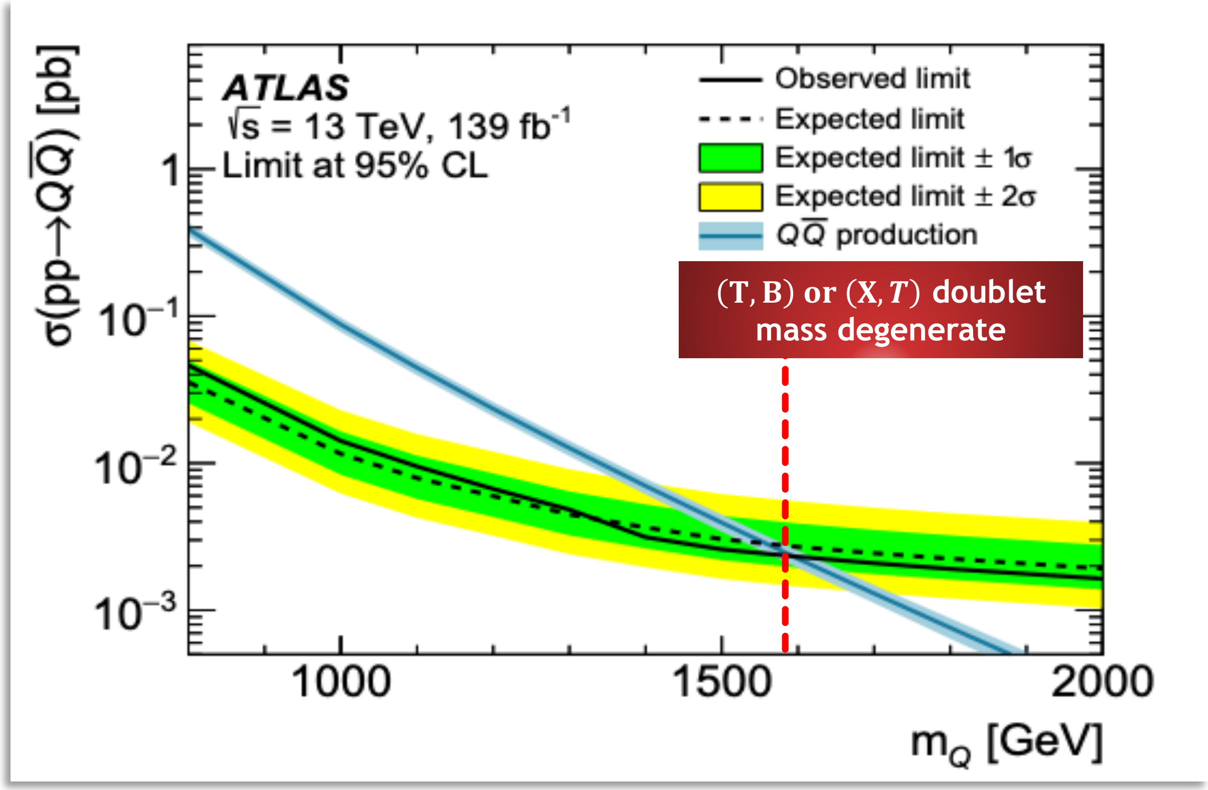


$B\bar{B}$ signal of mass 1.2 TeV
B $(Z$b, Hb, Wt) = (0.1, 0.1, 0.8)



Eur. Phys. J. C 83 (2023) 719

Results



Strongest limits corresponding to the weak-isospin doublet model

➔ (T,B) and (X,T) when $m_X = m_T = m_B$ are at 1.59 TeV

For masses of the VLQs >800GeV:

- ✓ No significant excesses
- ✓ Expected and observed mass limits as a function of the T and B branching ratios $\mathcal{B}$
- ✓ Analysis most sensitive to the T → Zt and B→Wt decay modes
  - ✓ $\mathcal{B}(\mathrm{T'} \to \mathrm{Zt}) = 100\%$
  - ✓ $\mathcal{B}(\mathrm{B'} \to \mathrm{Wt}) = 100\%\%$
- ✓ limit at 1.47 TeV for exclusive T → Zt decays
- ✓ limit at 1.46 TeV for exclusive B/X → Wt decays
- ✓ Lower limits on the T and B quark masses are derived for all possible branching ratios
- ✓ The obtained mass limits are 300 to 400 GeV higher than in the earlier ATLAS analysis in the same final state using a subset of the Run 2 data.

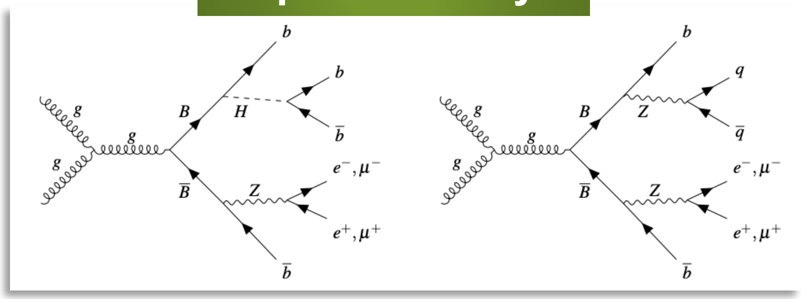
A search for bottom-type, vector-like quark pair production in leptonic and fully hadronic final states

B2G-20-014 Submitted to Phys. Rev. D

February 2024

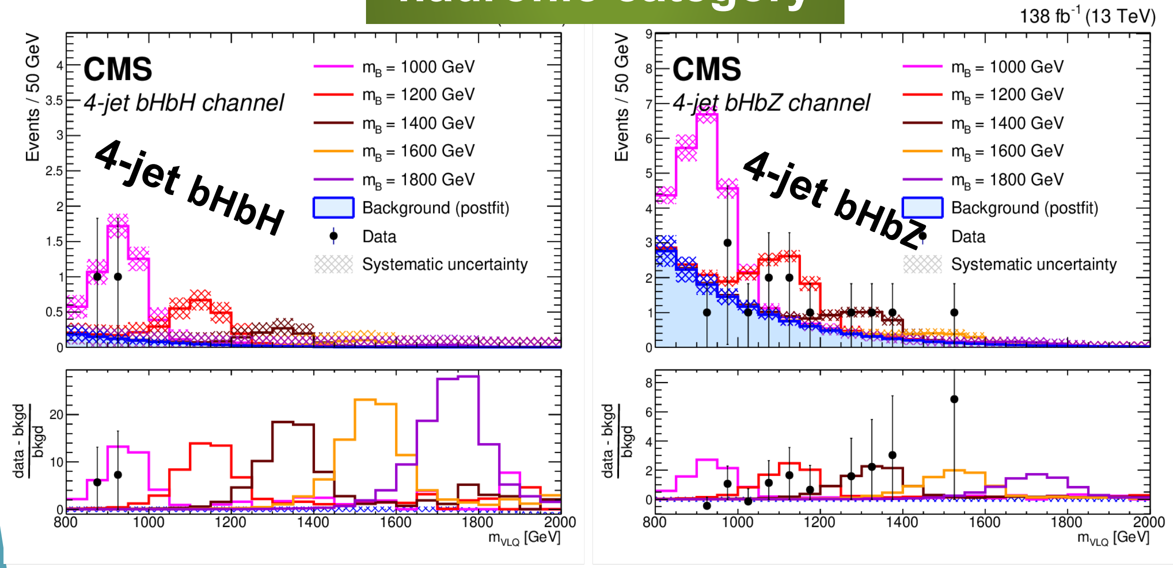
- **Masses of the VLQs from 1000 to 1800 GeV**
- **The search uses 138 fb^{-1} data collected with the CMSS detector**
- **First analysis to combine fully hadronic and leptonic categories**
- Branching Ratios $\mathcal{B}$: **Leptonic: $\mathcal{B}(Zb, Hb, Wt)$** & **Hadronic: $\mathcal{B}(Zb, Hb)$**

Leptonic decays

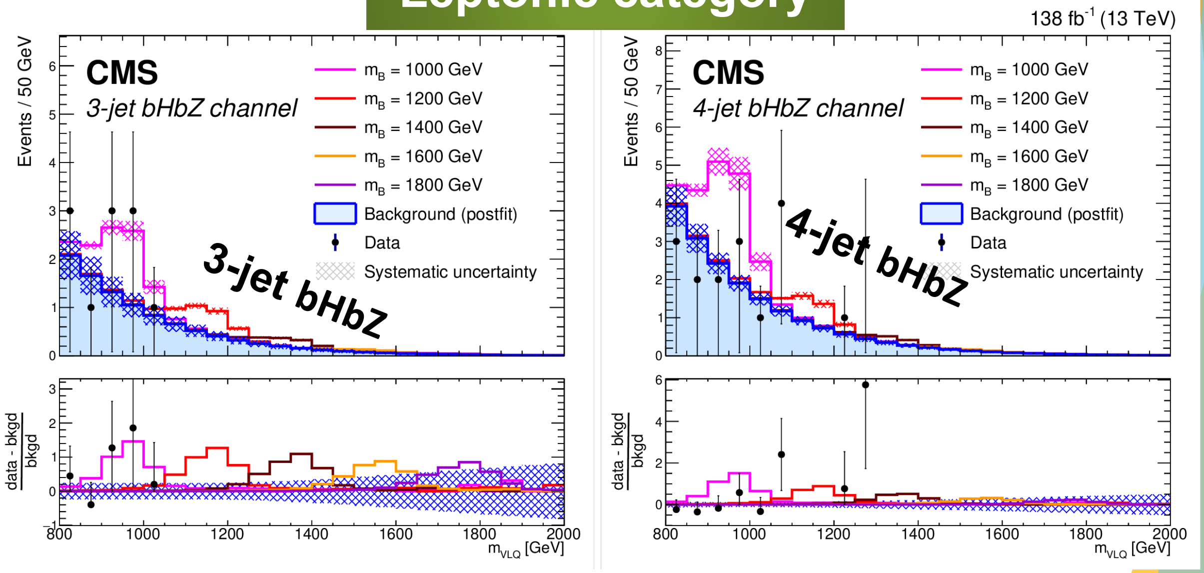


Reconstructed VLQ mass

hadronic category



Leptonic category



expected postfit background (blue histogram)
signal plus background (colored lines)
observed data (black points)

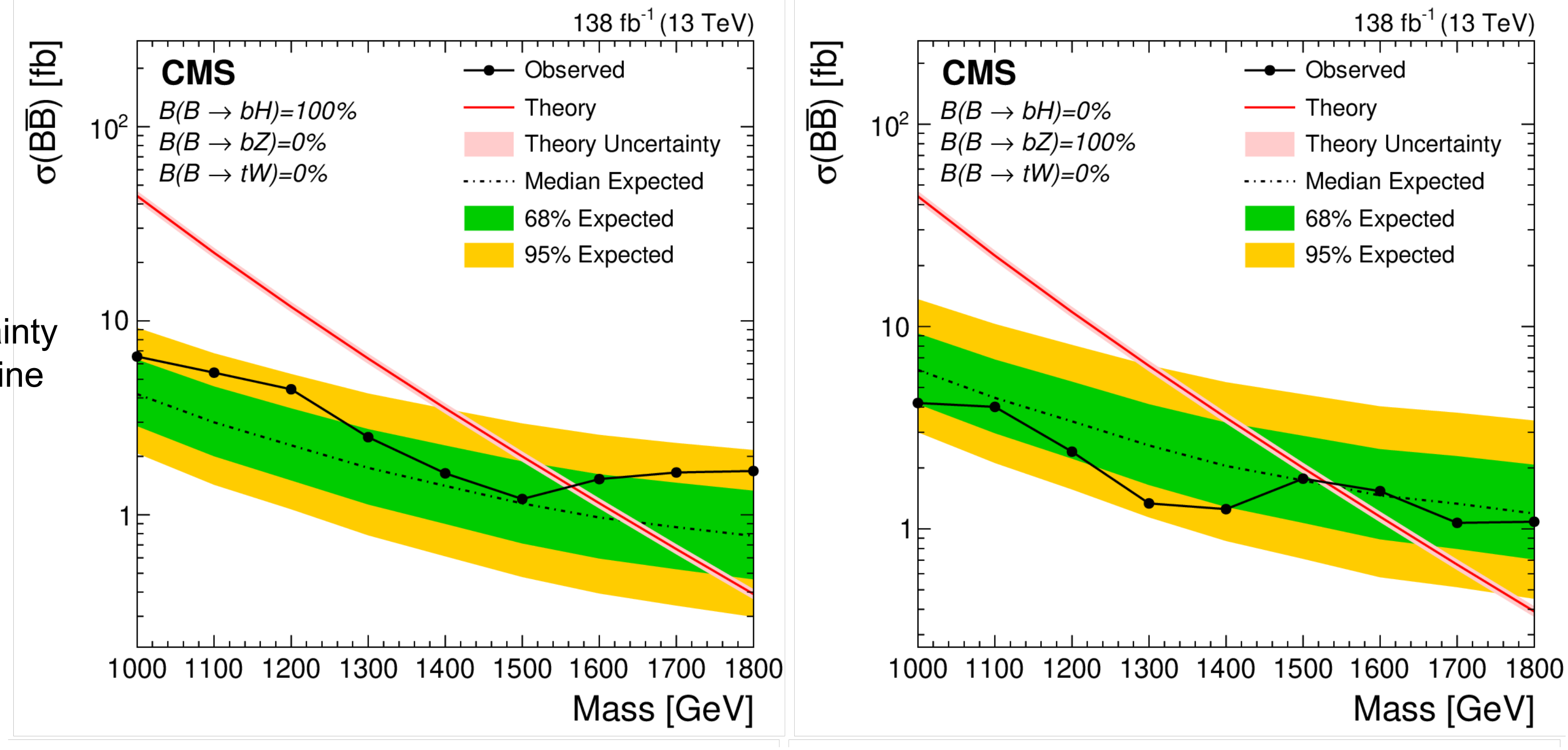
A search for bottom-type, vector-like quark pair production in leptonic and fully hadronic final states

B2G-20-014 Submitted to Phys. Rev. D

February 2024

Expected and observed limits on the cross section at 95% CL

The theoretical cross section and its uncertainty are shown by the red line and light-red band.



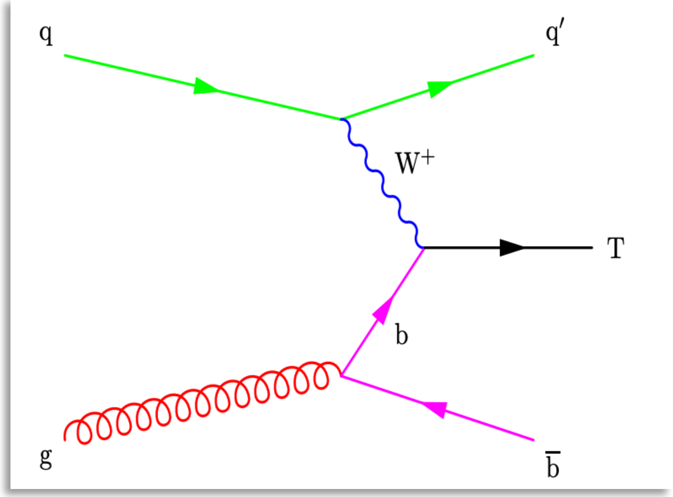
- ✦ No excess over the expected background is observed.
- ✦ Lower limits are set on the B VLQ mass at 95% confidence level and they depend on the B VLQ branching fractions:
  - ✦ for 100% B→bH ➔ 1.57 TeV
  - ✦ for 100% B→bZ ➔ 1.54 TeV
- ✦ **In most cases, the mass limits obtained exceed previous limits by at least 100 GeV.**

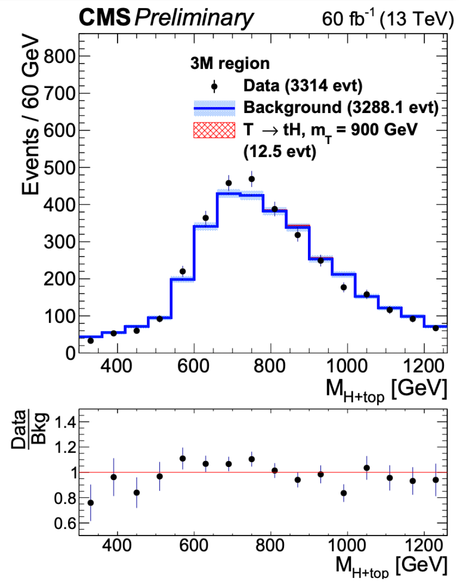
Search for electroweak production of a T in association with a b: T→ Ht / Zt →bb t

- Masses of the VLQs from 600 - 1200 GeV ➔ past searches signatures with a large Lorentz boost
- Search for electroweak production of a T in association with a b, which requires a nonzero TWb coupling for charged-current production
- The T with a narrow width can have charge 2/3 and can decay to a top quark (t) and a Higgs or Z boson ➔ Branching Ratios: B $(Zt, Ht, Wb) \approx (0.25, 0.25, 0.5)$
- Invariant mass reconstructed from 5 jets is used as the main discriminating variable
- Event kinematics and the presence of jets containing b hadrons are used to reconstruct the hadronic decays of the t and Higgs or Z boson.
- No discrepancy from the standard model prediction is observed in the data. The limits are stronger than those in the previous search by at least a factor three

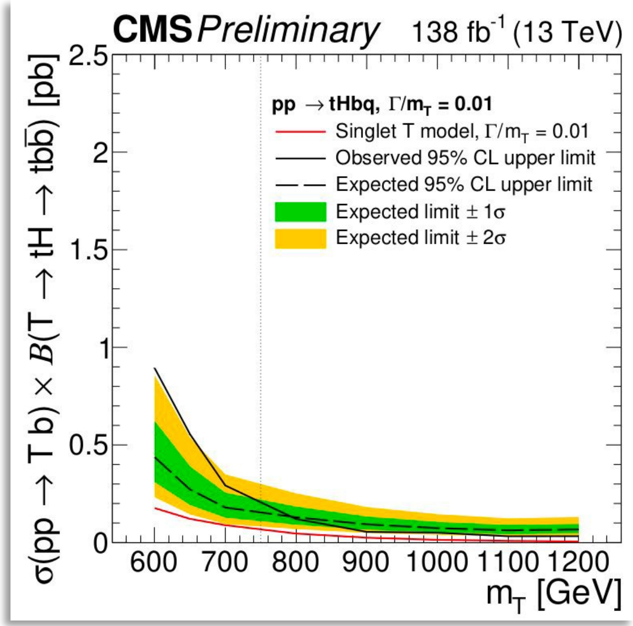
T→ Zt in ATLAS in backup slides

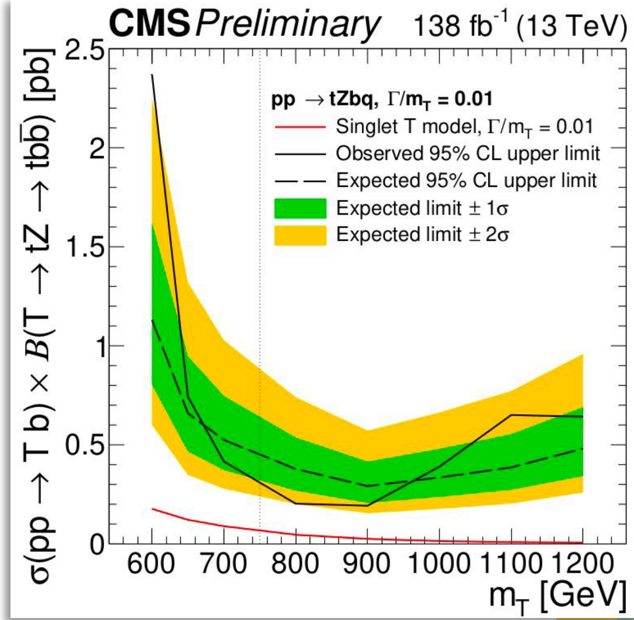
Associated production of a T and a b





95% CL upper limits on the cross-section for associated production with a b for final states tHbq and tZbq, for T masses from 600 - 1200 GeV



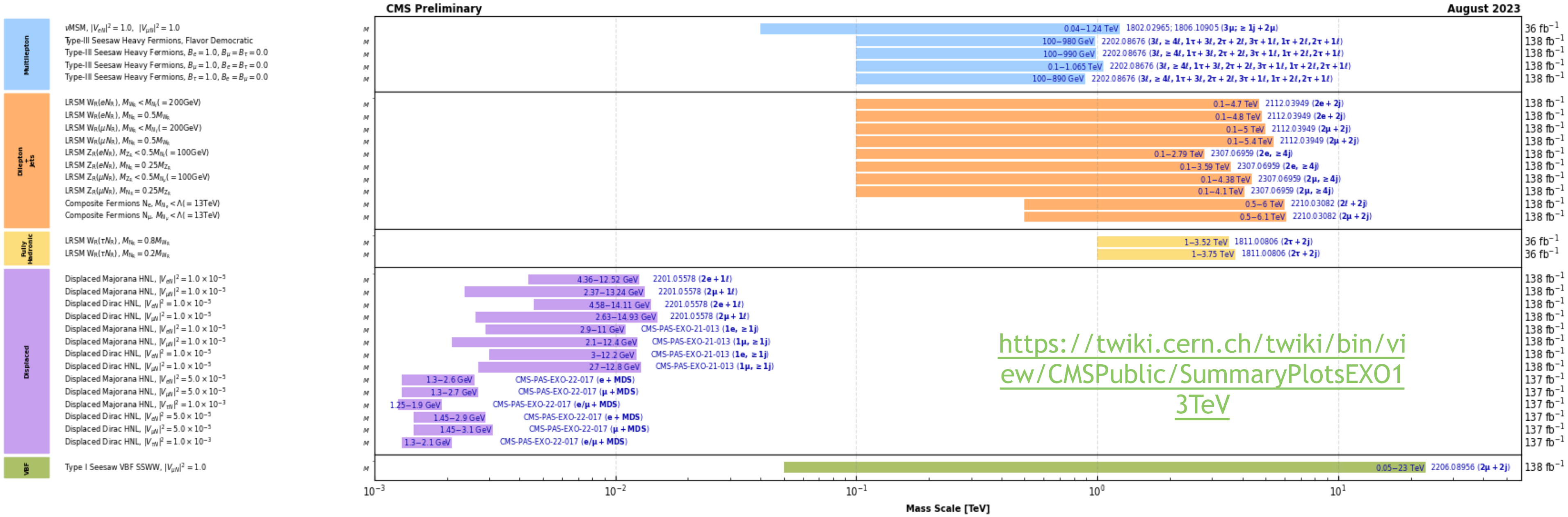


Heavy Neutral Leptons [HNLs]

Eur. Phys. J. C (2022) 82:1030
https://doi.org/10.1140/epjc/s10052-022-11011-7

Channel	Lepton flavour	Experiment	$\sqrt{s}$ (TeV)	$\mathcal{L}$ (fb^{-1})	M_N (GeV)
Prompt SS dilepton $pp \to \ell_\alpha^\pm N \to \ell_\alpha^\pm \ell_\beta^\pm + nj$	$ee/\mu\mu$	CMS'12 [87]	7	4.98	(50, 210)
	$\mu\mu$	CMS'15 [88]	8	19.7	(40, 500)
	$ee/e\mu$	CMS'16 [89]	8	19.7	(40, 500)
	$ee/\mu\mu$	ATLAS'15 [90]	8	20.3	(100, 500)
	$ee/e\mu/\mu\mu$	CMS'18 [69]	13	35.9	(20, 1600)
	$\mu\mu$	LHCb'20 [83]	7–8	3.0	(5, 50)
Prompt OS dilepton $pp \to \ell_\alpha^\pm N \to \ell_\alpha^\pm \ell_\beta^\mp + nj$	$\mu\mu$	LHCb'20 [83]	7–8	3.0	(5, 50)
Prompt trilepton $pp \to \ell_\alpha^\pm N \to \ell_\alpha^\pm \ell_\beta^\pm \ell_\gamma^\mp \nu$	$eee + ee\mu/\mu\mu\mu + \mu\mu e$	CMS'18 [91]	13	35.9	(1, 1200)
	$ee\mu/\mu\mu e$	ATLAS'19 [84]	13	36.1	(5, 50)
Displaced trilepton $pp \to \ell_\alpha N,\ N \to \ell_\beta \ell_\gamma \nu$	$\mu - e\mu/\mu - \mu\mu$	ATLAS'19 [84]	13	32.9	(4.5, 10)
	6 combinations of e, μ	ATLAS'22 [85]	13	139	(3, 15)
	6 combinations of e, μ	CMS'22 [86]	13	138	(1, 20)

Overview of CMS HNL results

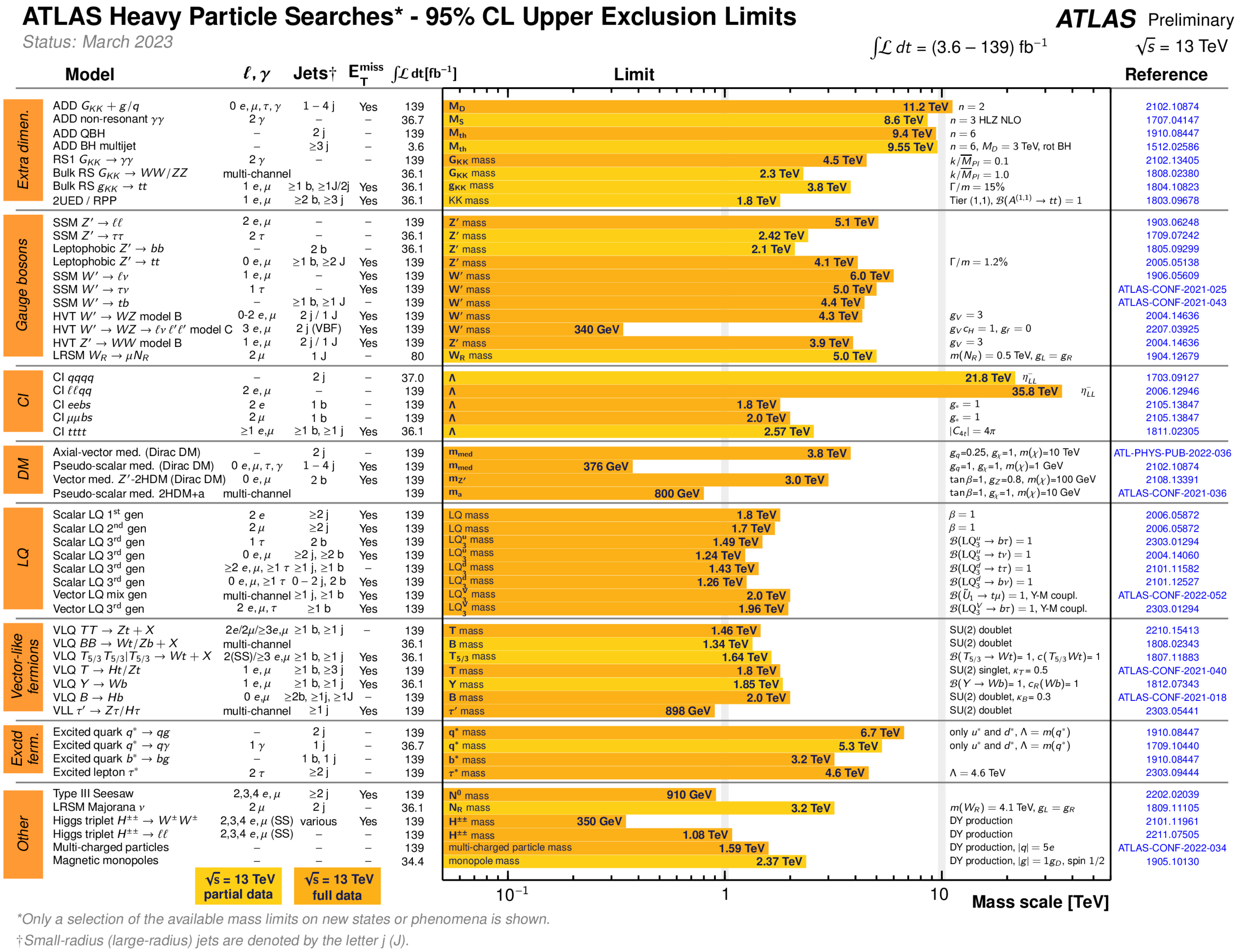


https://twiki.cern.ch/twiki/bin/view/CMSPublic/SummaryPlotsEXO13TeV

Selection of observed exclusion limits at 95% C.L. (theory uncertainties are not included).

Many searches, in different final states, and with both prompt and displaced signatures from ATLAS, CMS & LHCb

Ranges of new particle masses or energy scales excluded at the 95% confidence level

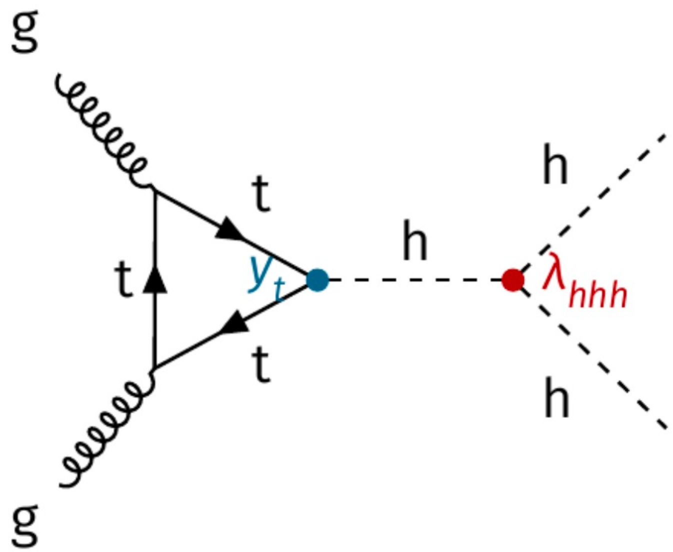


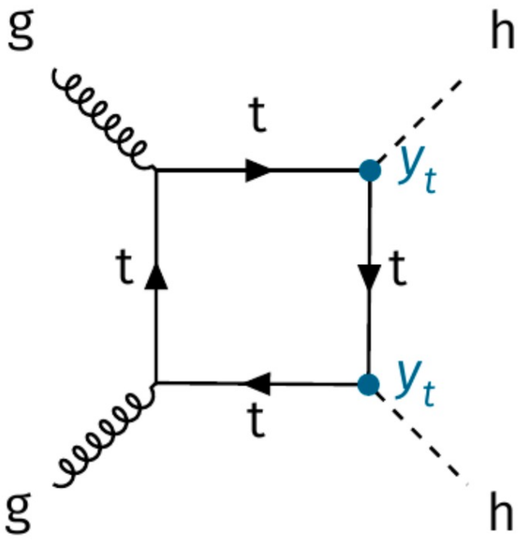
https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2023-008/

similar plot for CMS in slide

Analysis example: Higgs self-coupling measurements using ML

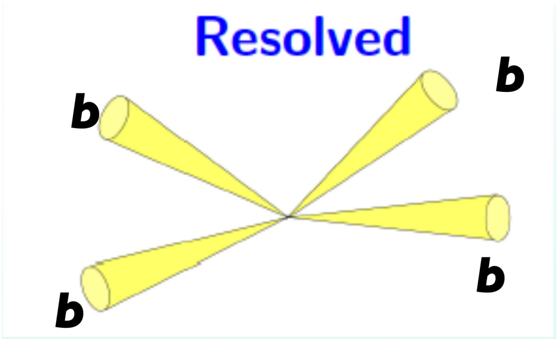
arXiv:2004.04240v3

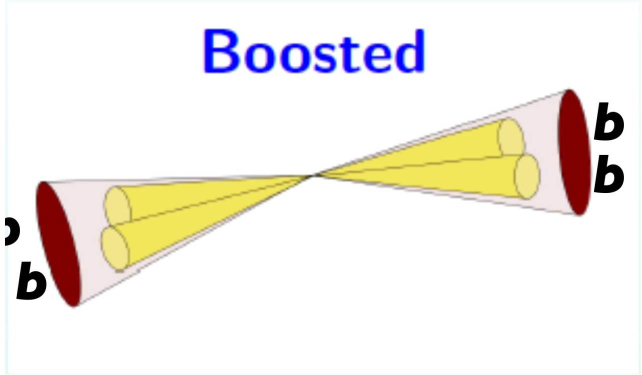




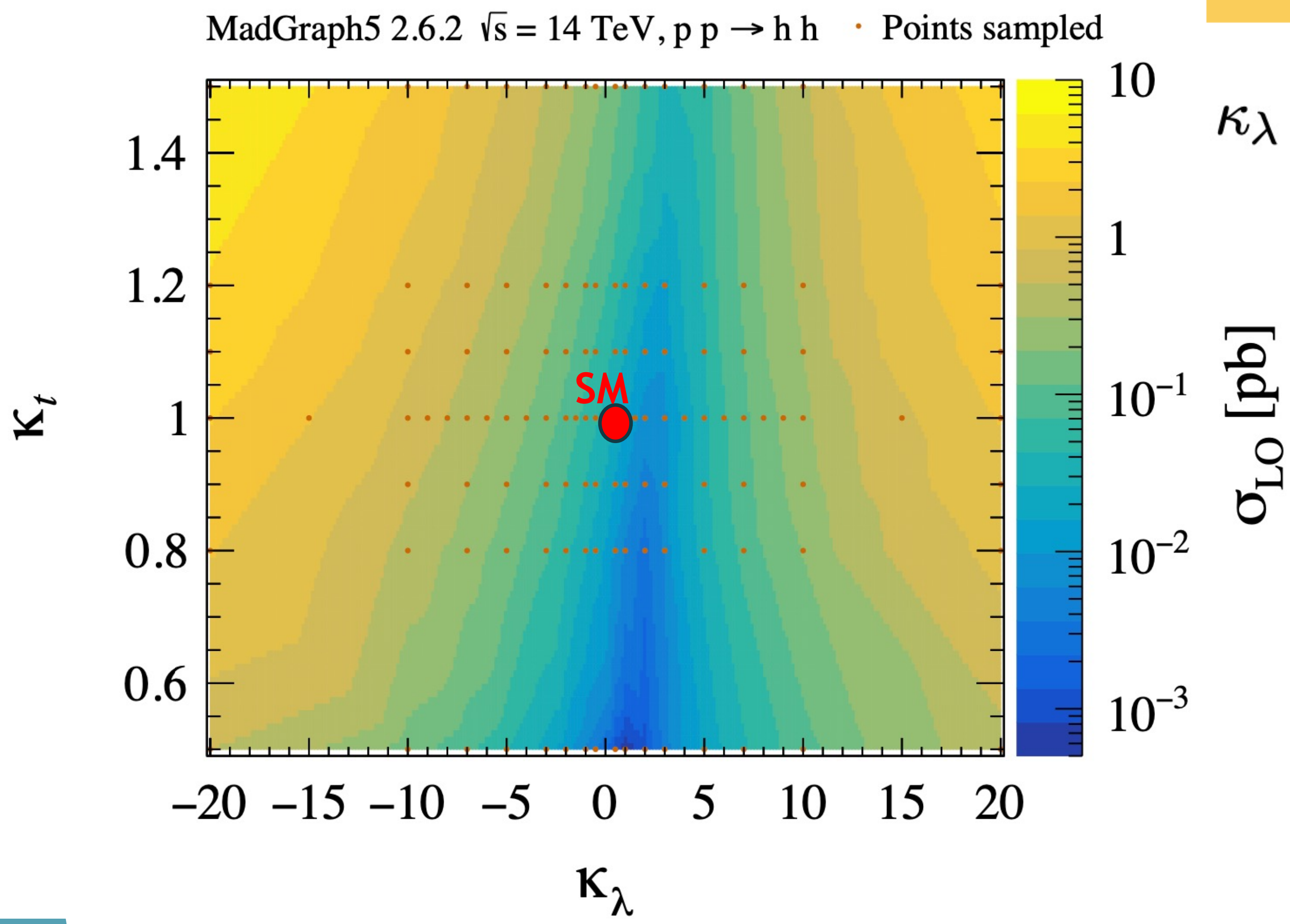
$$V(h) = m_h^2 h^2 + \lambda_{hhh} v h^3 + \lambda_{hhhh} h^4.$$

Exploit **4b final state**: highest signal yields, but **overwhelming QCD background** (by orders of magnitude!)





Analysis example: Higgs self-coupling measurements using ML



$$\sigma_{\text{triangle}} \sim \lambda_{hhh}^2 y_t^2, \quad \sigma_{\text{box}} \sim y_t^4, \quad \sigma_{\text{interference}} \sim -\lambda_{hhh} y_t^3.$$

$$\kappa_\lambda = \lambda_{hhh}/\lambda_{hhh}^{\text{SM}} \text{ and } \kappa_t = y_t/y_t^{\text{SM}}$$

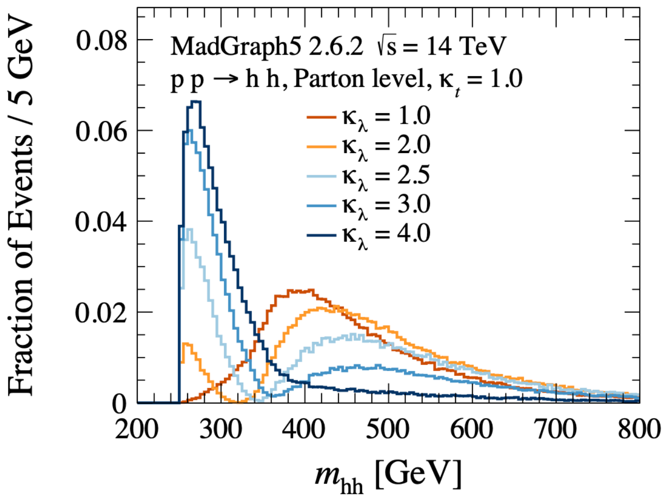
Current limits on λ_{hhh} from ATLAS:

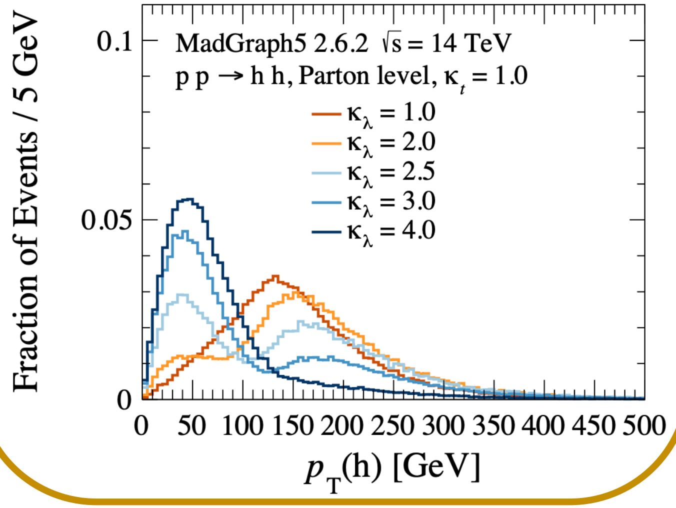
$$-2.3 < k_\lambda < 10.3$$

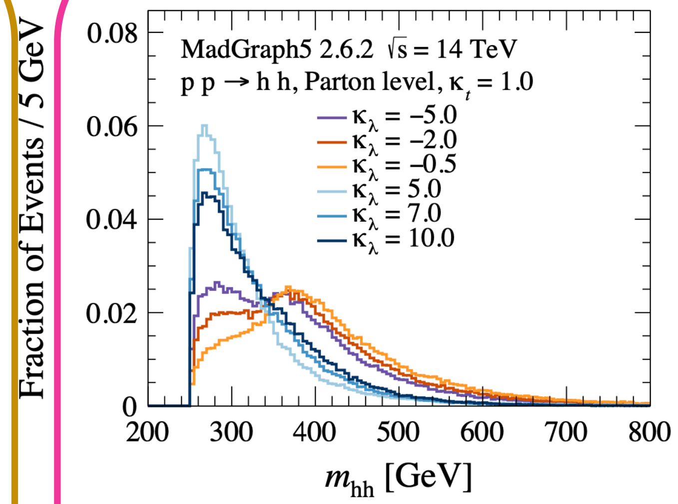
Ideally, analyses should have a high signal-to- background ratio S/B across different kinematic regimes of the Higgs boson but…

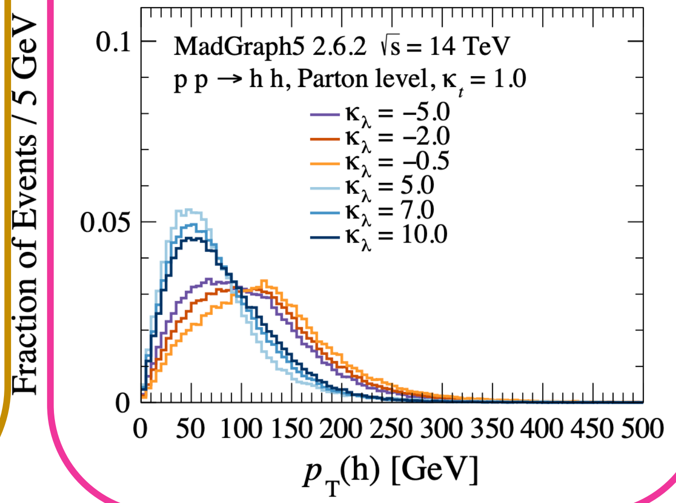
small k_λ variations cause dramatic changes to kinematical variables distributions as for m_{hh} and p_T(h)

Analysis example: Higgs self-coupling measurements using ML





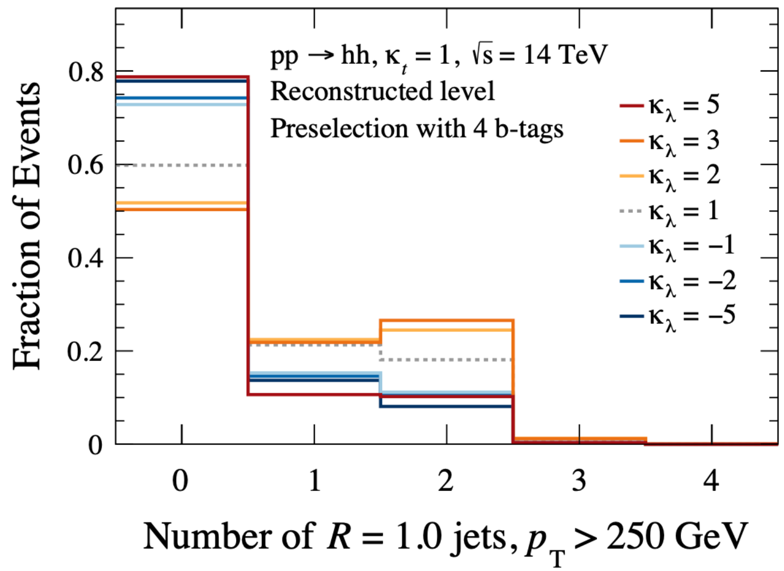


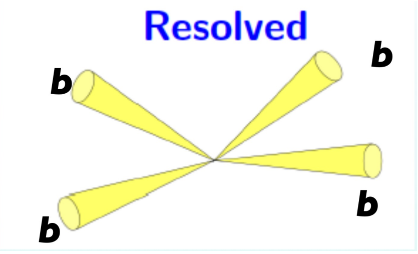


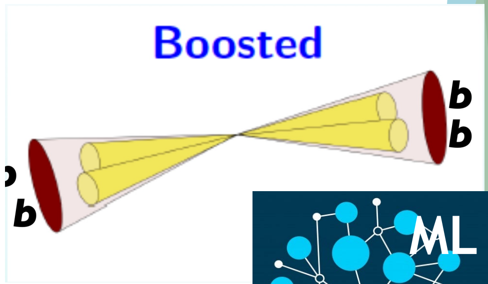
k_λ ***impact on*** m_{hh} and $p_T(h)$

k_λ ***sign impact on*** m_{hh} and $p_T(h)$

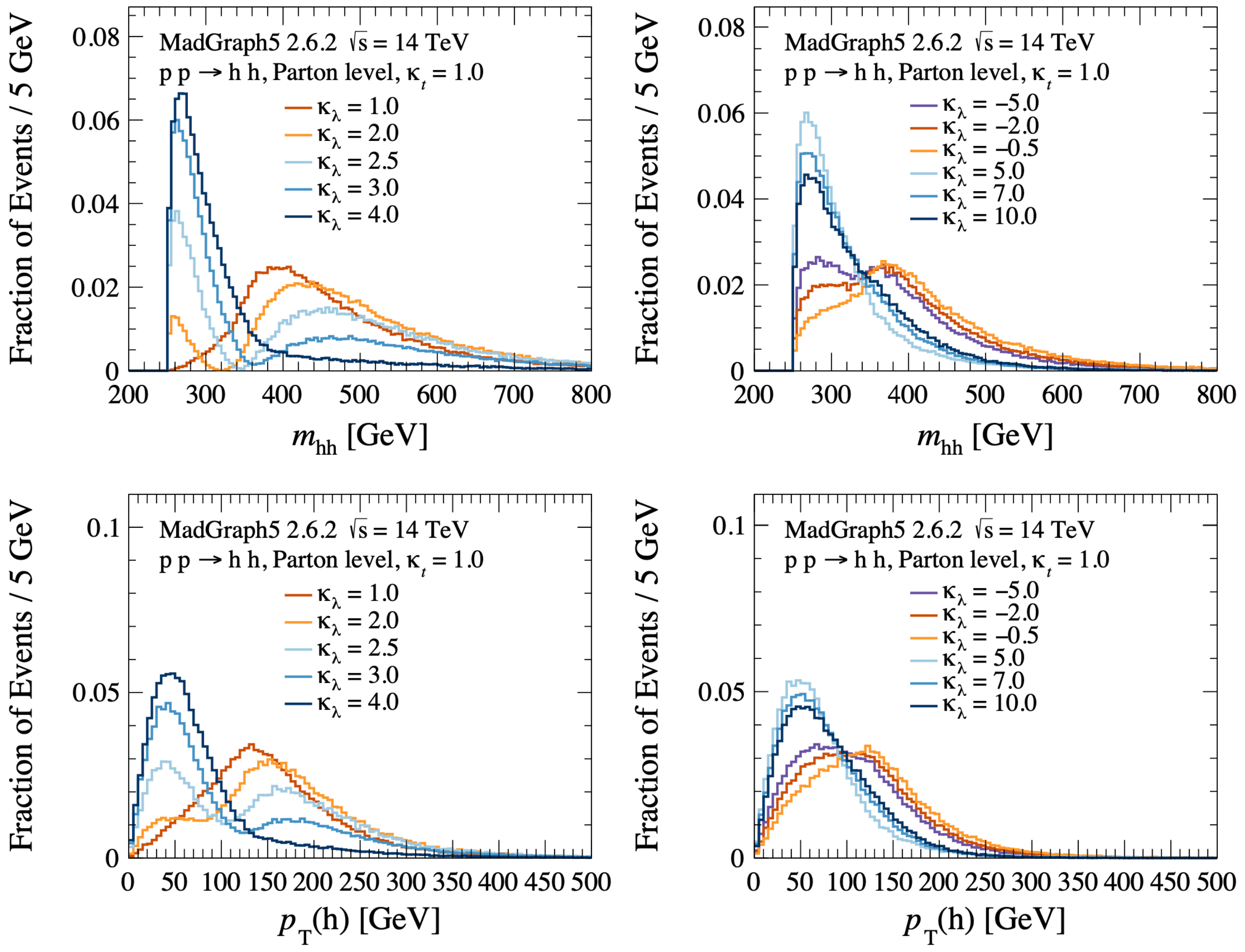
1) Resolved and/or boosted analysis?



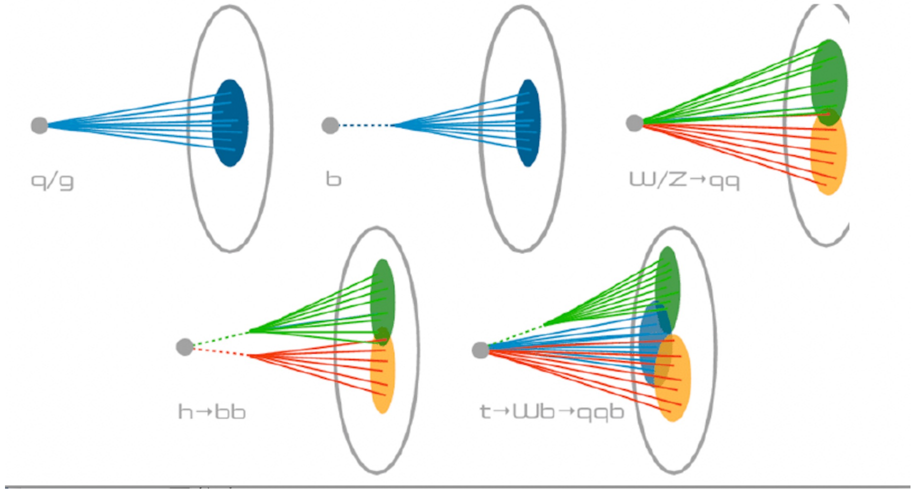


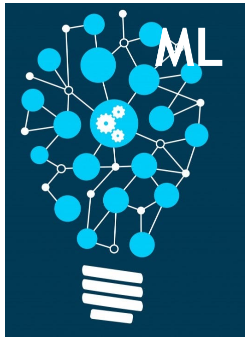


Analysis example: Higgs self-coupling measurements using ML

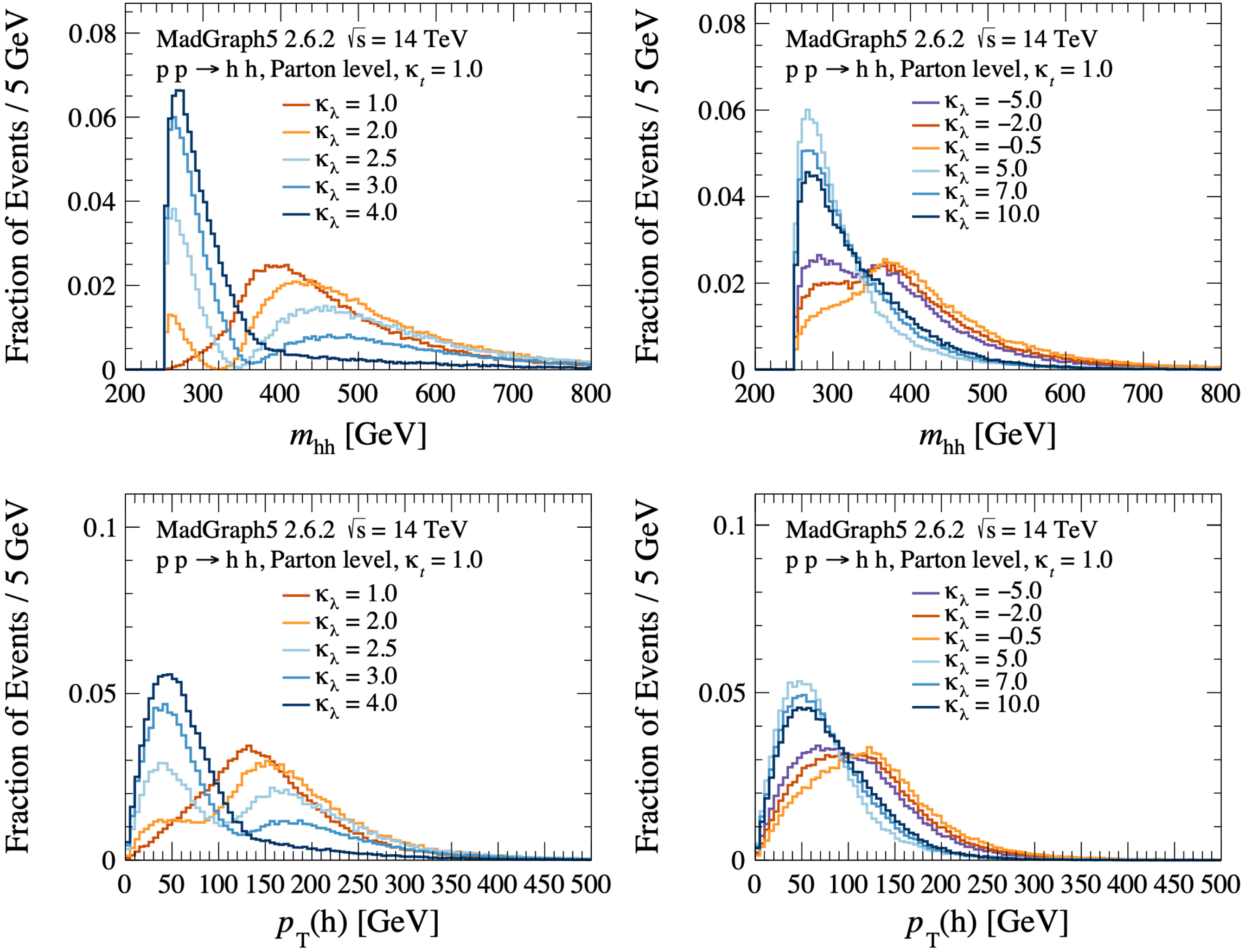


2) boosted analysis require Boson tagging



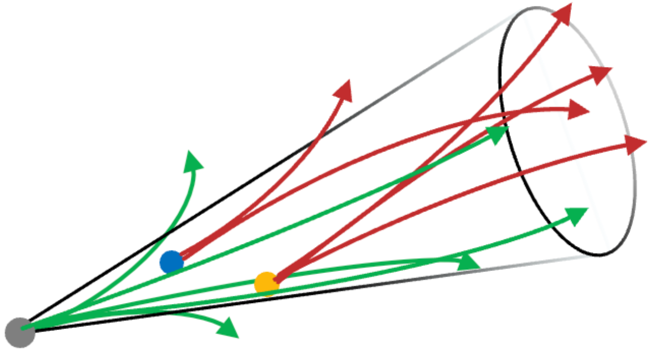


Analysis example: Higgs self-coupling measurements using ML



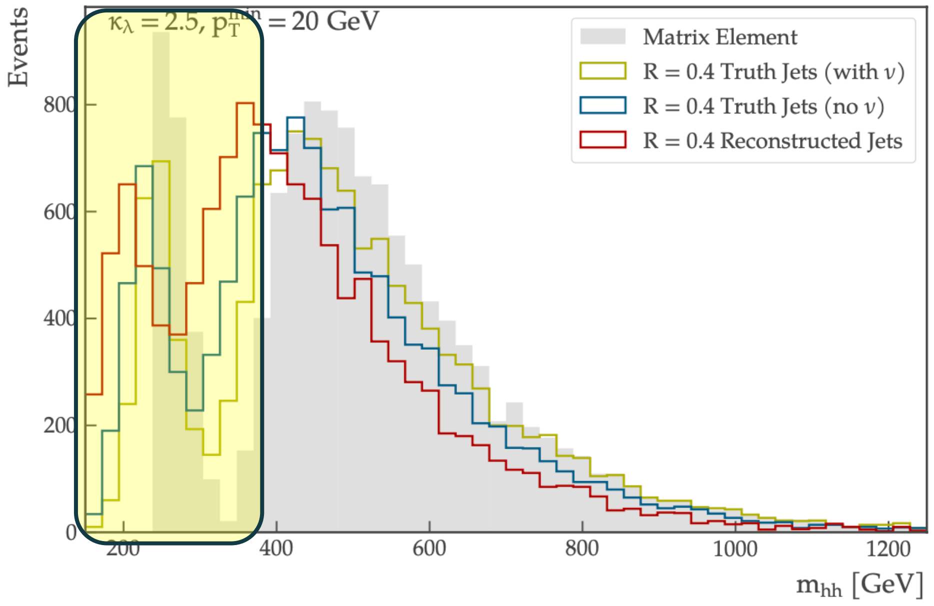
3) Resolved analysis require excellent b-tagging performance

- Due to the small signal-to-background ratios sensitivity is limited by systematics
- Reducing the multijet background mitigate the impact of systematics whose suppression relies on modern b-tagging algorithms

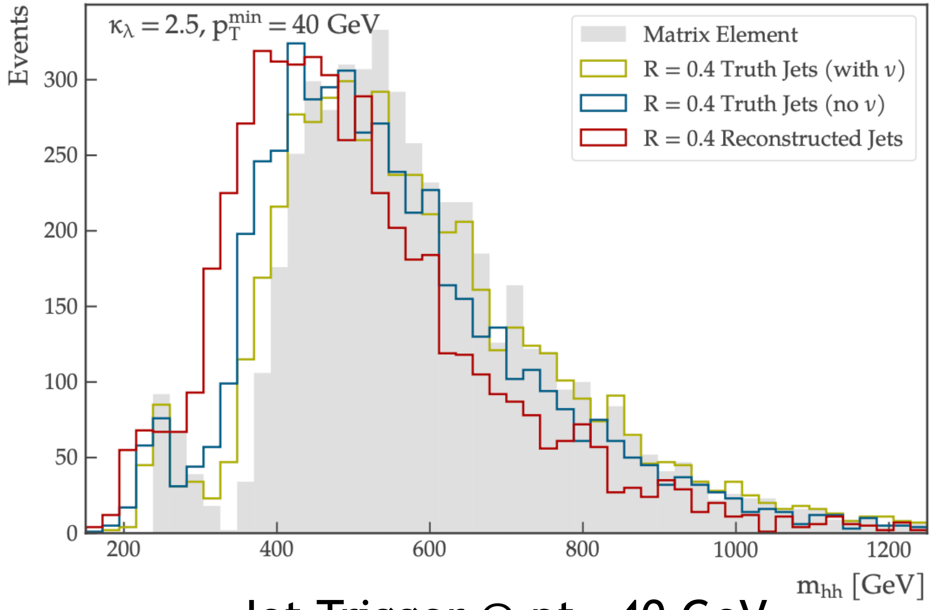


Analysis example: Higgs self-coupling measurements using ML

4) Online (trigger) selection:



Jet Trigger @ pt =20 GeV



Jet Trigger @ pt =40 GeV

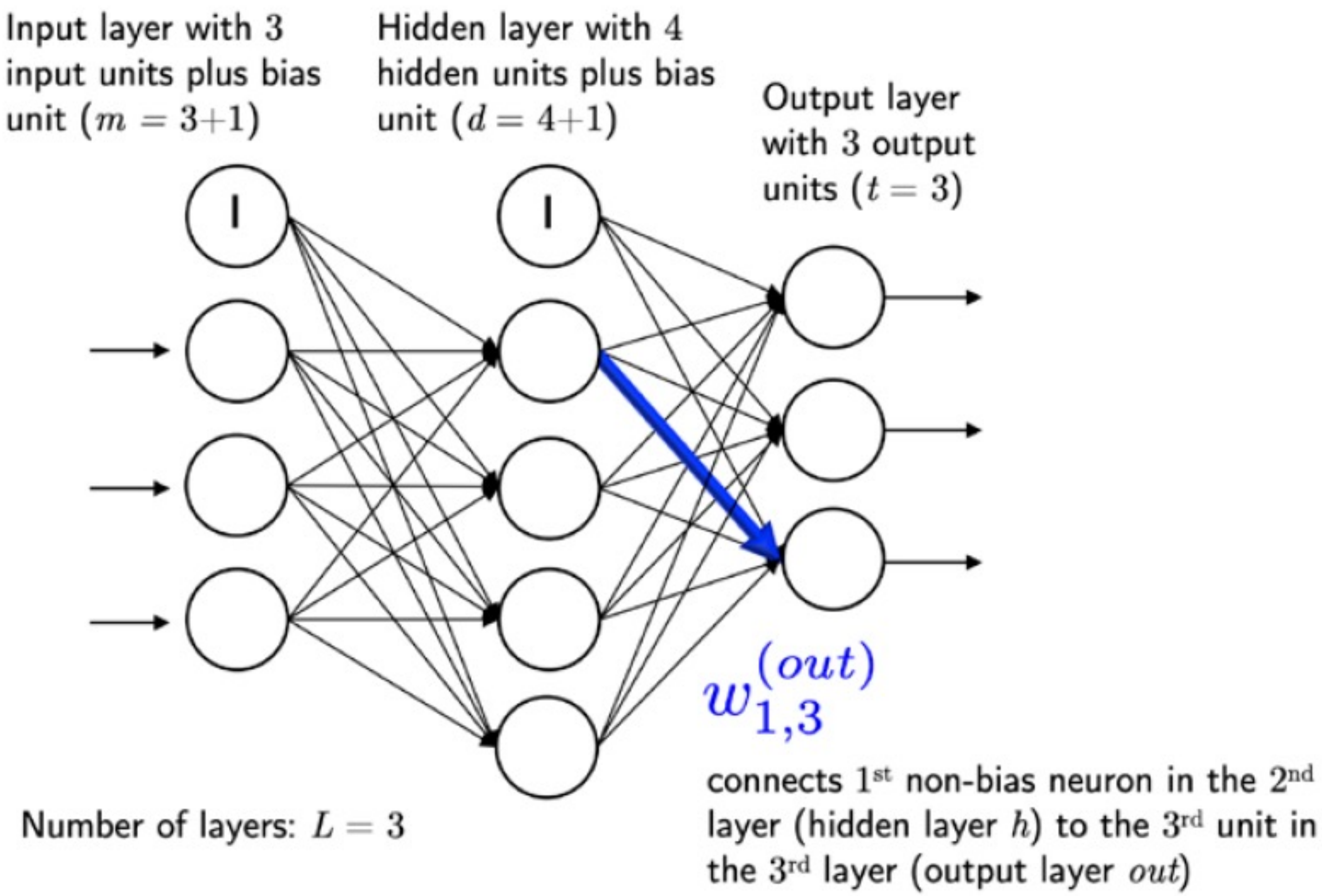
Maintaining sufficiently low trigger thresholds for the HL-LHC upgrades is of key importance for both discovery of the di-Higgs process as well as constraining $\boldsymbol{\lambda}_{hhh}$.

ML Classification in the hardware trigger

a.k.a. selection before storing data!

Multilayer perceptron

- The first example of a NN is the Multilayer perceptron, this is a net of fully connected perceptron
- In the schematical view on the right every circle is a perceptron with a fixed number of inputs and outputs
- In the example we have an **input layer,** only one **hidden layer** and an **output layer**



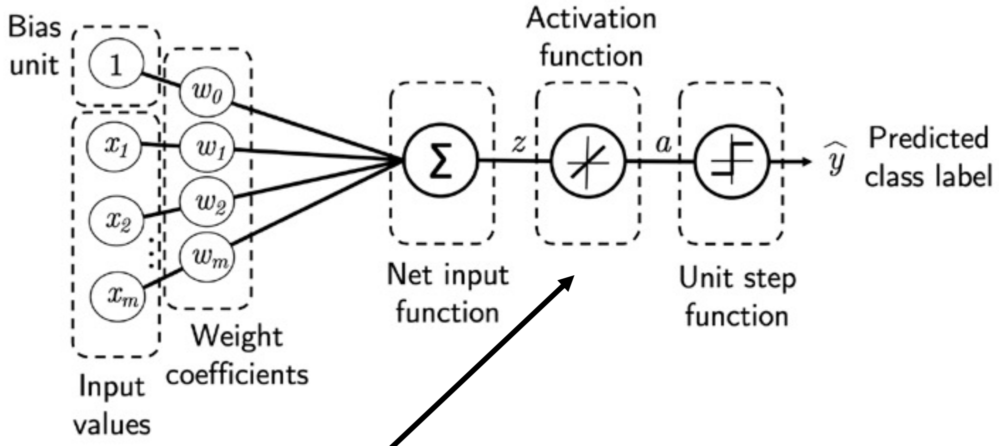
Activation Functions

The activation function provides to the activation or not of a node.

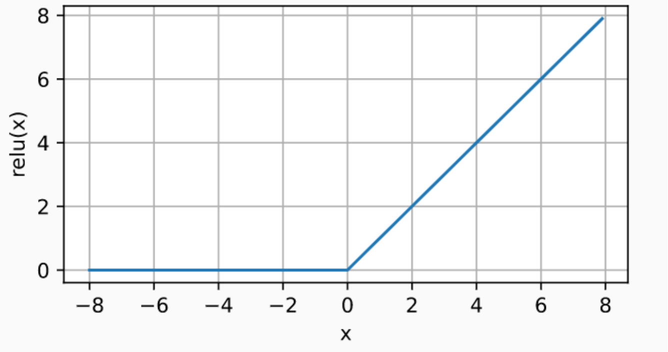
The functions are in general differentiable operators to transform the inputs to outputs

Most of them provides to add non-linearity to the model

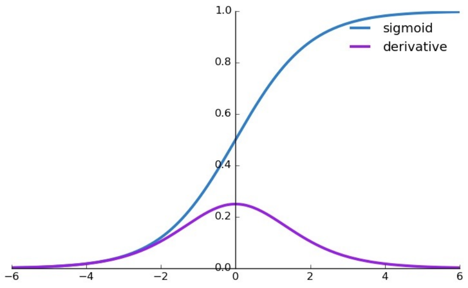
The activation function **σ** has as input the weighted sum of the input variables x, added with the bias b



Rectified Linear Unit (ReLU)



Sigmoid Function



$$\sigma = \sigma\left(\vec{w} \cdot \vec{x} + b\right)$$

$$ReLU(x) = max(0, x)$$

$$sigmoid(x) = \frac{1}{1 + e^{-x}}$$

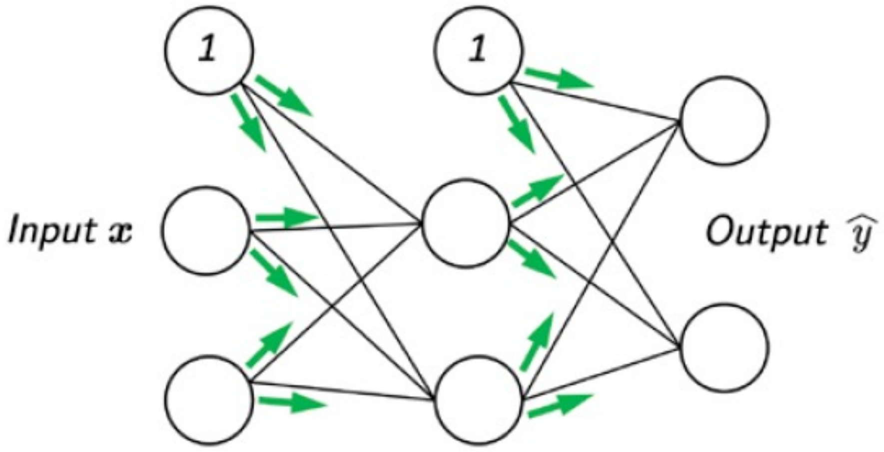
Classification

- How can we interpret the two values?
- In classification problem the goal is to understand how the input **x** is related to the belonging to a certain class.
- The output **o** could be seen as the vector of probabilities of belonging to each class.

However this is not straightforward:

$$\sum_{i=1}^{\#classes} o_i \neq 1$$

$$o_i \ngeq 0 \;\; \forall\, i \in [1, \#classes]$$



Softmax activation in multiclass regression

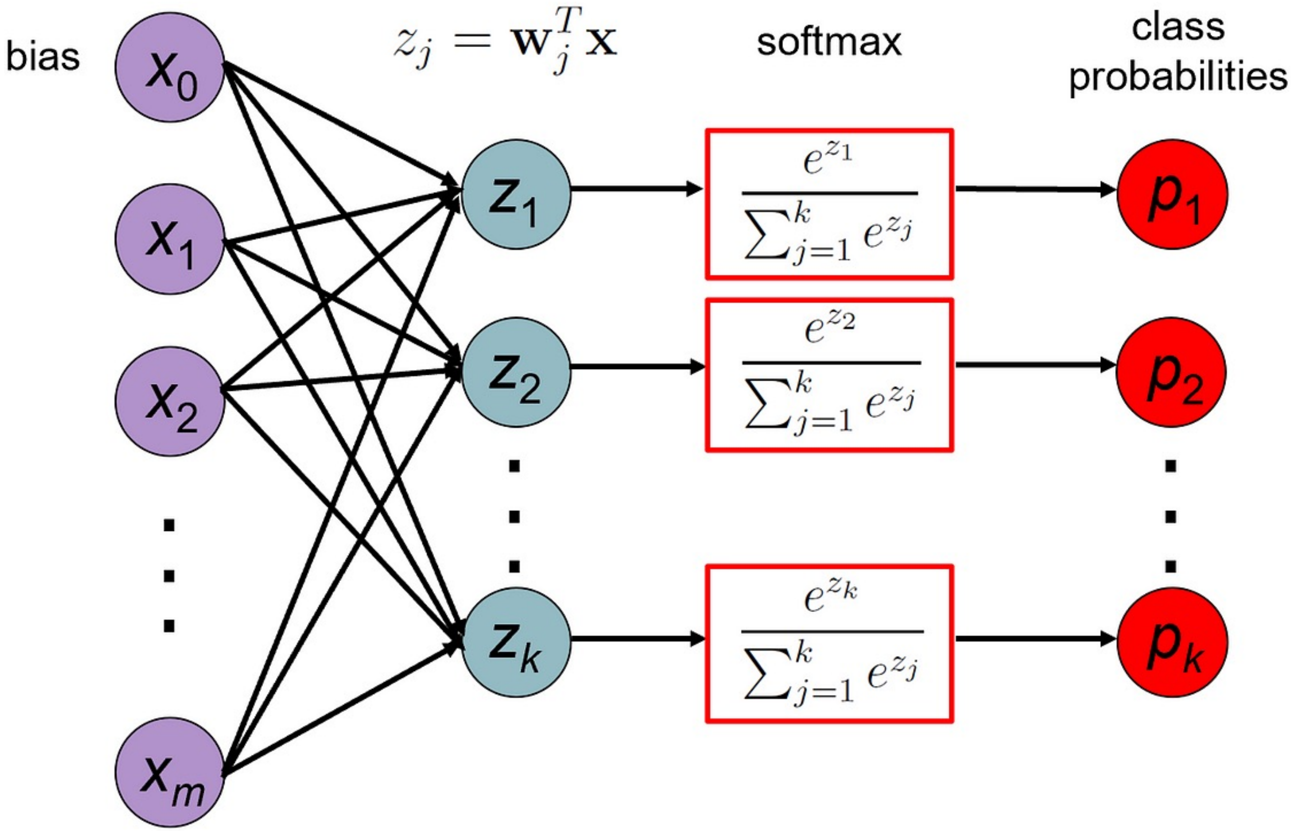
- However this is not straightforward:

$$\sum_{i=1}^{\#classes} o_i \neq 1$$

$$o_i \ngeq 0 \;\; \forall\, i \in [1, \#classes]$$

- **Softmax activation function:**

$$\boxed{\begin{aligned} \mathbf{y} &= softmax(\mathbf{o}) \\ y_i &= \frac{e^{o_i}}{\sum_k e^{o_k}} \\ y_{pred} &= \underset{j}{argmax}\; y_j \end{aligned}}$$



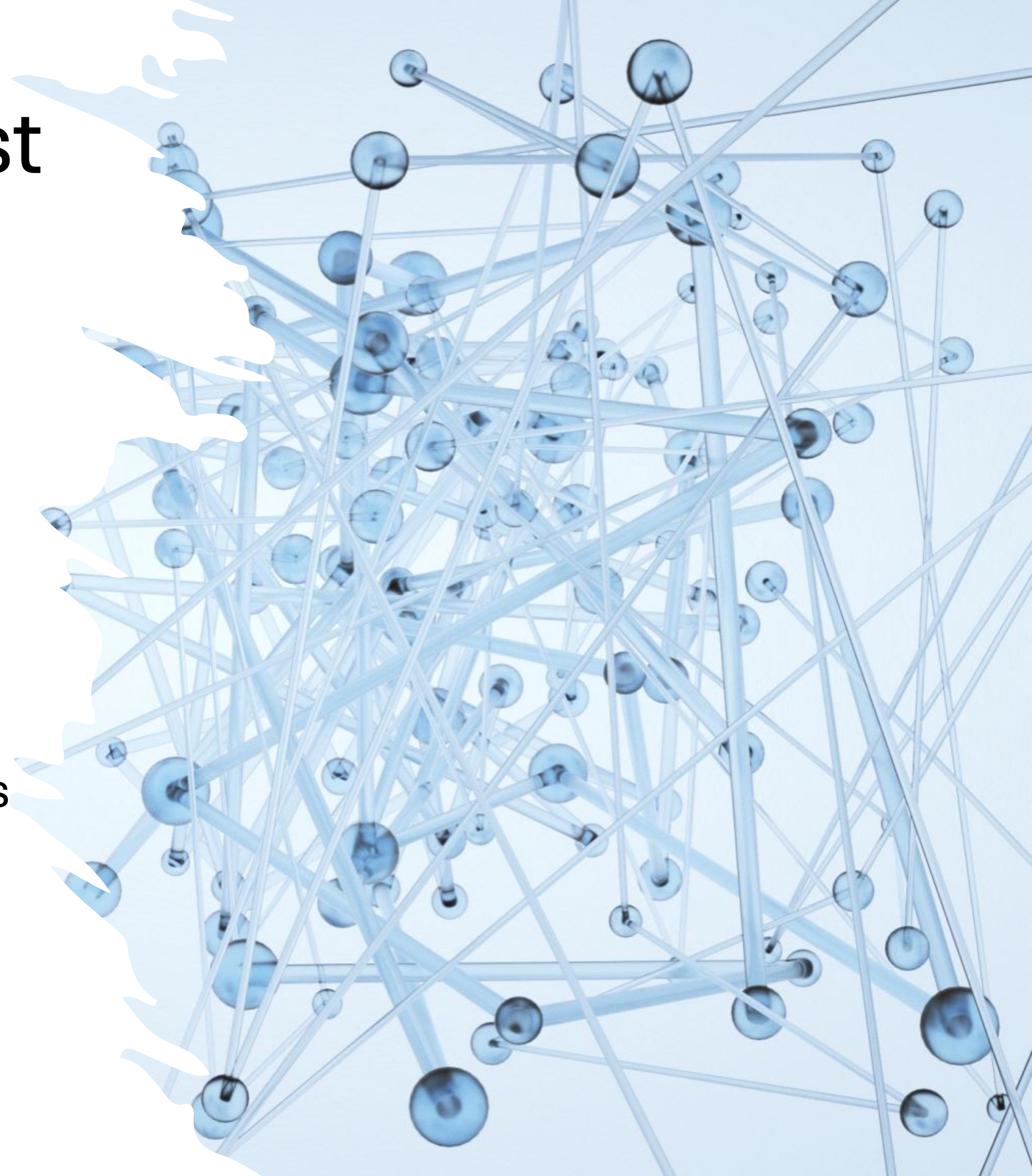
DNN Parametrization Cost

- Fully-connected layers are fundamental in the neural network building up process.

- Adding neurons to a network layer or adding a layer makes our model more complex and capable of facing a wide range of problems.

- The complexity of the model, however, faces directly with computational time, which could be extremely high. Suppose to have a hidden layer with d input and q outputs:

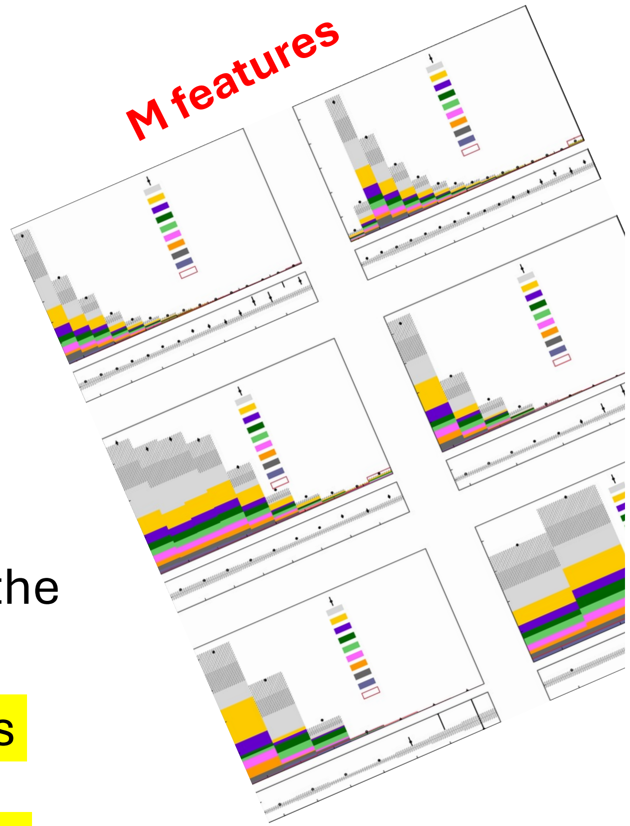
The parametrization cost is $\mathcal{O}(d \cdot q)$

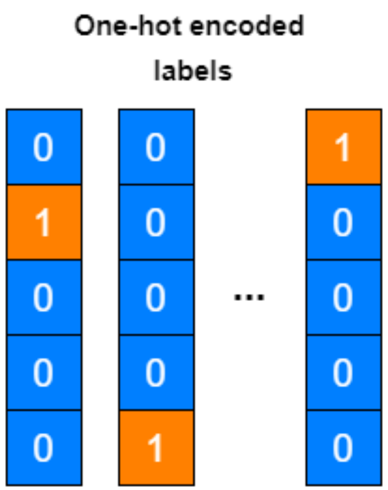


Loss function

- To measure the quality of our predicted probabilities we need a **loss function**.

- Let's suppose that the entire dataset has N examples {X,Y}. The i-th {X,Y} data entry is made by the **M features** vector **x-ith** and the one-hot label vector **y-ith**.
- It is possible to compare the predicted class with the real class by checking how probable the actual classes are according to our model.
- According to the maximum likelihood estimation, we want to maximize P(Y|X), or minimize the negative log-likelihood.





$$\mathcal{P}(\mathbf{Y}|\mathbf{X}) = \prod_{i=1}^{n} \mathcal{P}(\mathbf{y}^{(i)}|\mathbf{x}^{(i)})$$

$$-log\mathcal{P}(\mathbf{Y}|\mathbf{X}) = \sum_{i=1}^{n} -log\mathcal{P}(\mathbf{y}^{(i)}|\mathbf{x}^{(i)})$$

Which loss function?

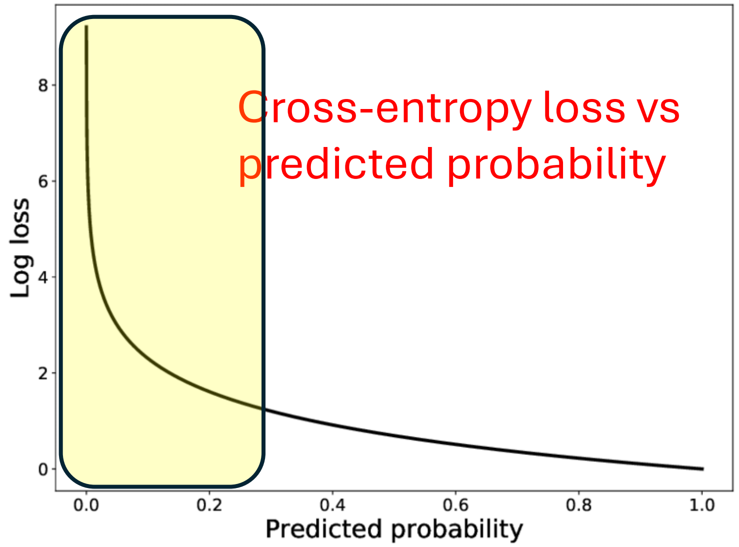
- The negative log-likelihood is equal to:

$$-log\mathcal{P}(\mathbf{Y}|\mathbf{X}) = \sum_{i=1}^{n} l(\mathbf{y}^{(i)}, \mathbf{y}_{pred}^{(i)})$$

- Usually loss can be the **cross-entropy**, defined as:

$$l(\mathbf{y}^{(i)}, \mathbf{y}_{pred}^{(i)}) = -\sum_{j=1}^{\#Classes} y_j^{(i)} \; log \, y_j^{(i)\,pred}$$

$$NB : y_j^{(i)\,pred} \leq 1 \;\rightarrow log \, y_j^{(i)\,pred} \leq 0$$



$$l(y, y^{pred}) = -\sum_{j=1}^{\#Classes} y_j \, log \frac{e^{o_j}}{\sum_{k=1}^{\#Classes} e^{o_k}}$$

$$= \sum_{j=1}^{\#Classes} y_j \log \sum_{k=1}^{\#Classes} e^{o_k} - \sum_{j=1}^{\#Classes} y_j \, o_j$$

$$= \log \sum_{k=1}^{\#Classes} e^{o_k} - \sum_{j=1}^{\#Classes} y_j \, o_j$$

Mean Squared Error (MSE)/ Quadratic Loss/ L2:

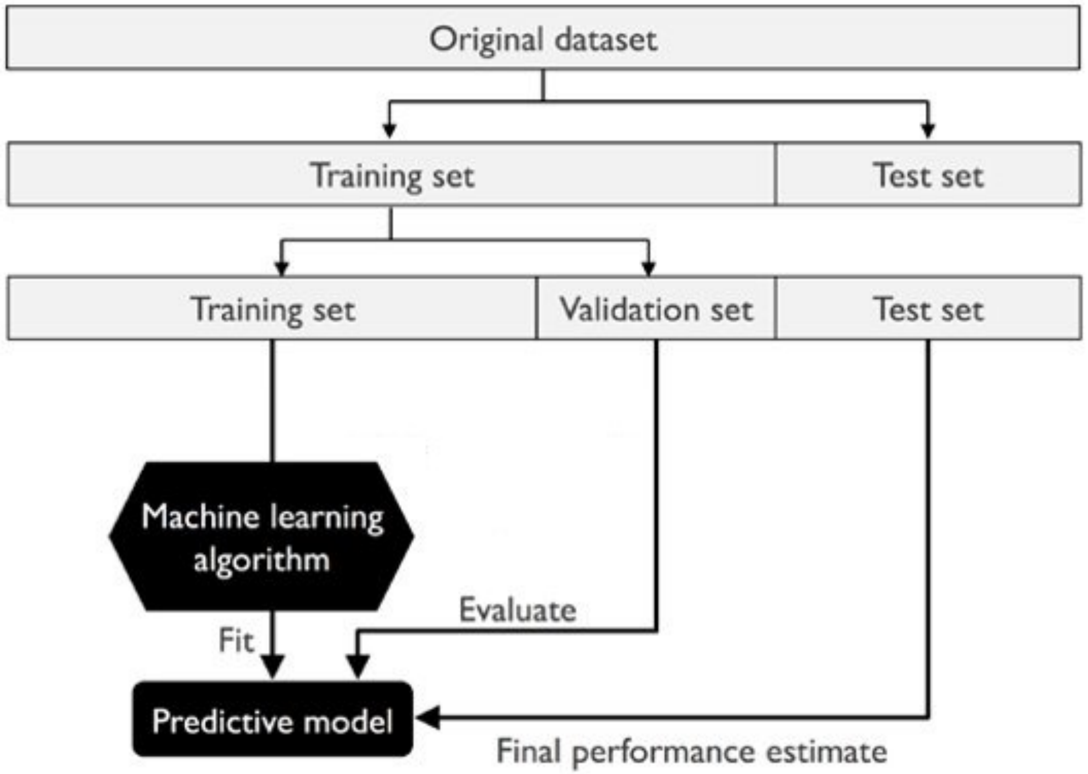
$$MSE(y^{(i)}, y_{pred}^{(i)}) = \frac{\left(y^{(i)} - y_{pred}^{(i)}\right)^2}{n}$$

Mean Absolute Error (MAE)/ L1 Loss:

$$MAE(y^{(i)}, y_{pred}^{(i)}) = \frac{\left|y^{(i)} - y_{pred}^{(i)}\right|}{n}$$

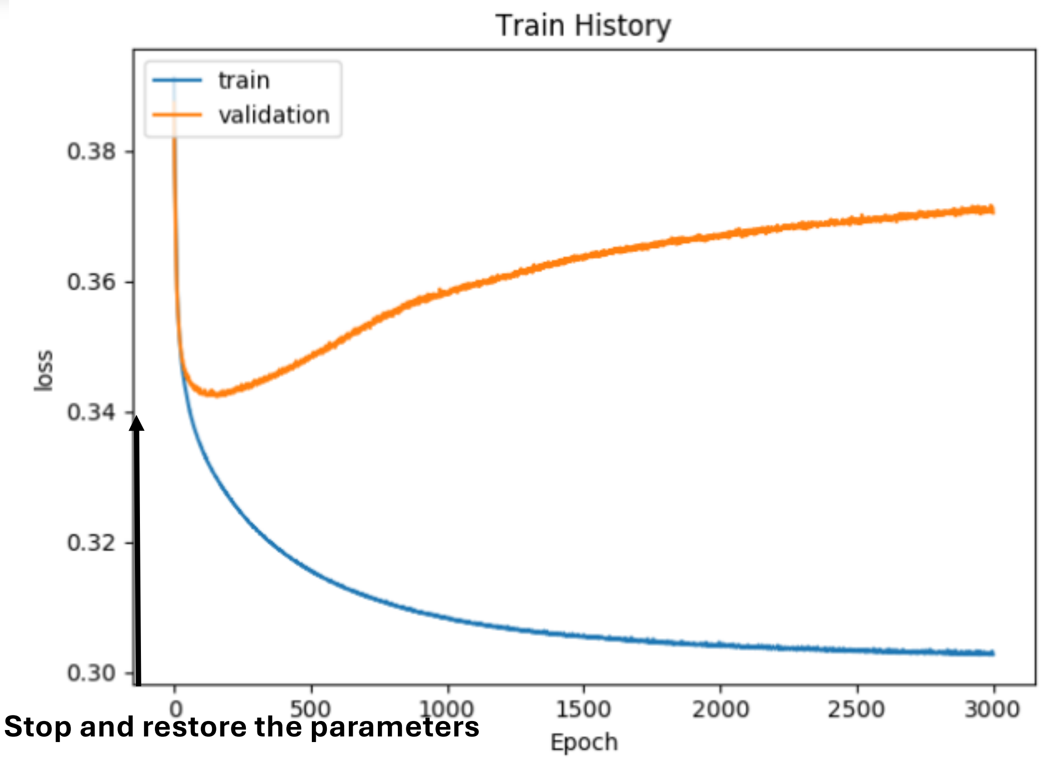
Validation procedure

- The common practice to address this problem is to split our data three ways, incorporating a ***validation set*** in addition to the **training** and **test** datasets.
- **Number of epochs:** The main idea of the training is to iterate over the network model different times.
- In each epoch is selected from the training dataset **k** stochastic **minibatches** of n (batchsize) entries.
- Model update (i.e. weights update) is done on the average loss over single minibatch. Then after k iterations the epoch ends.



Validation and Overfitting problem

Here a clear example of overfittig, the train loss keeps going down while the validation loss get worse!!



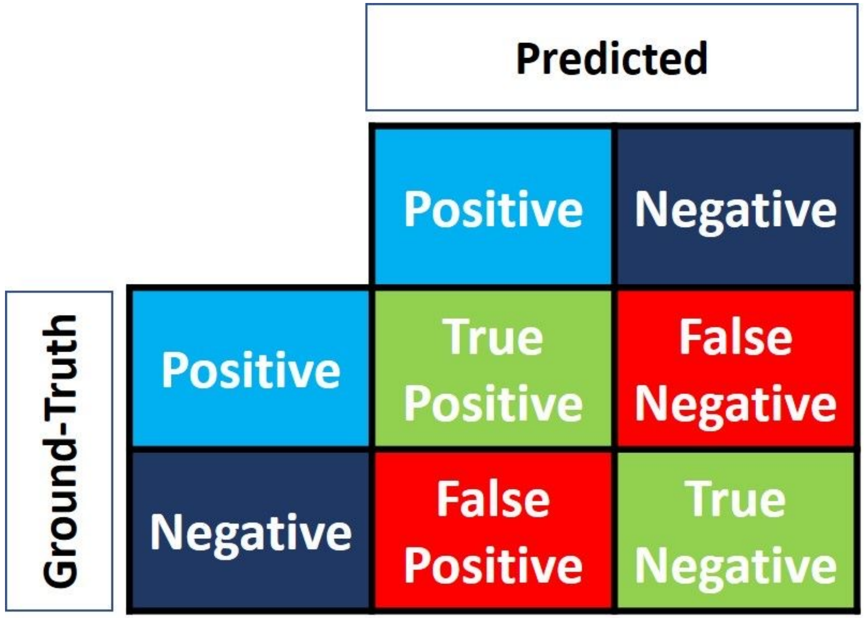
The more complex the model is, the higher is the risk of overfitting.

In order to avoid overfitting and make the training stable we have different approach:

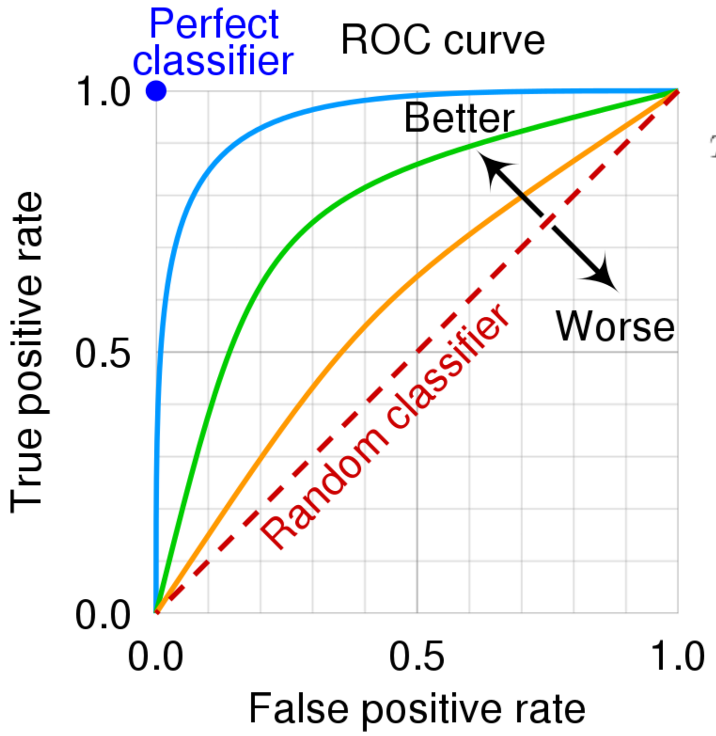
1. Introduce a callback function that stops the training if the validation loss get worse and restore the best parameters (**Early Stop function**). Reduce overtraining and time needed for the training.
2. "**Dropout**": injecting noise while computing each internal layer during forward propagation.

Performance evaluation

- The confusion matrix helps us visualize whether the model is "confused" in discriminating between two or more classes.



- Metrics:
  - **Accuracy** = TP+TN/ALL
  - **Precision** = TP/TP+FP
    - (TP+FP = all predicted positive)
  - **Recall** = TP/TP+FN
    - (TP+FN = all true positive)

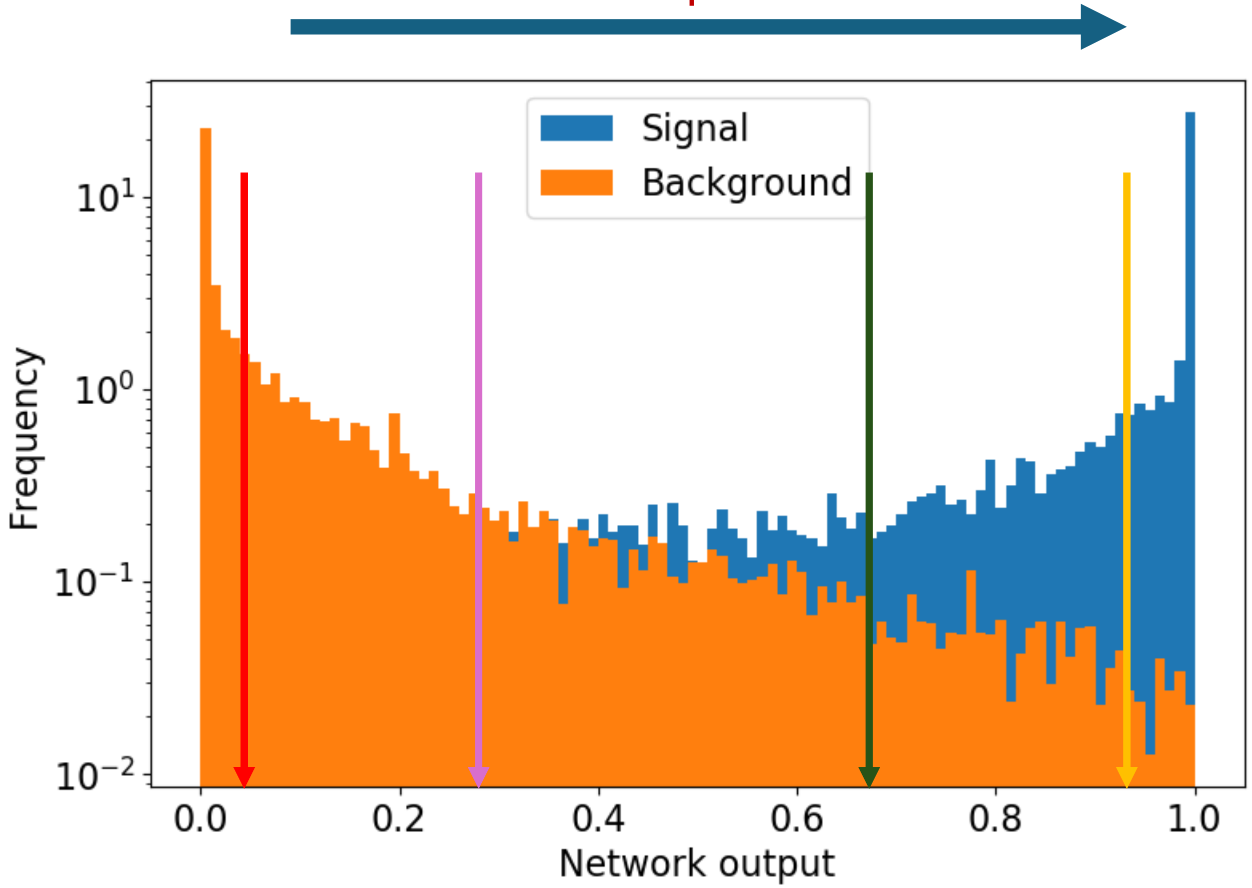


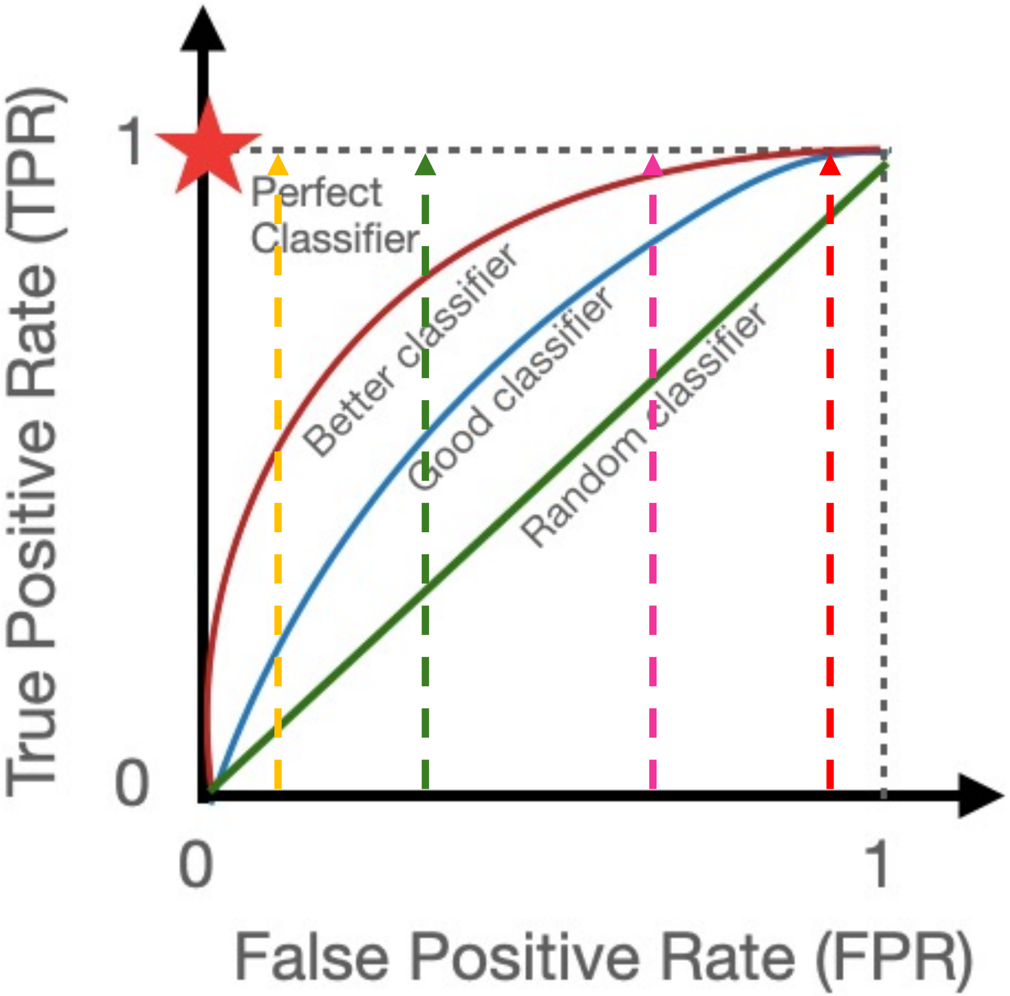
$$TPR = \frac{True_{positive}}{True_{positive} + False_{negative}}$$

$$FPR = \frac{False_{positive}}{False_{positive} + True_{negative}}$$

Roc curve

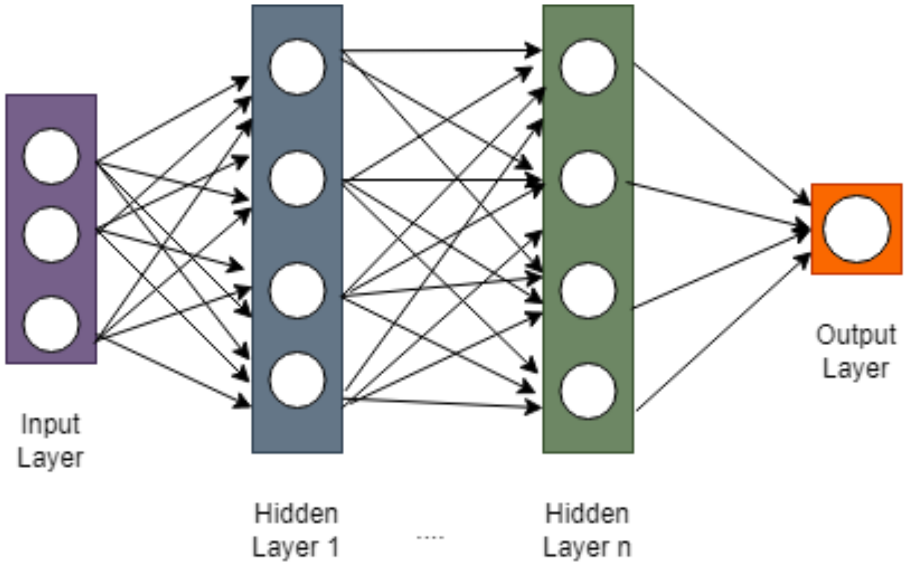
Events selected if: NN output value > arrow value





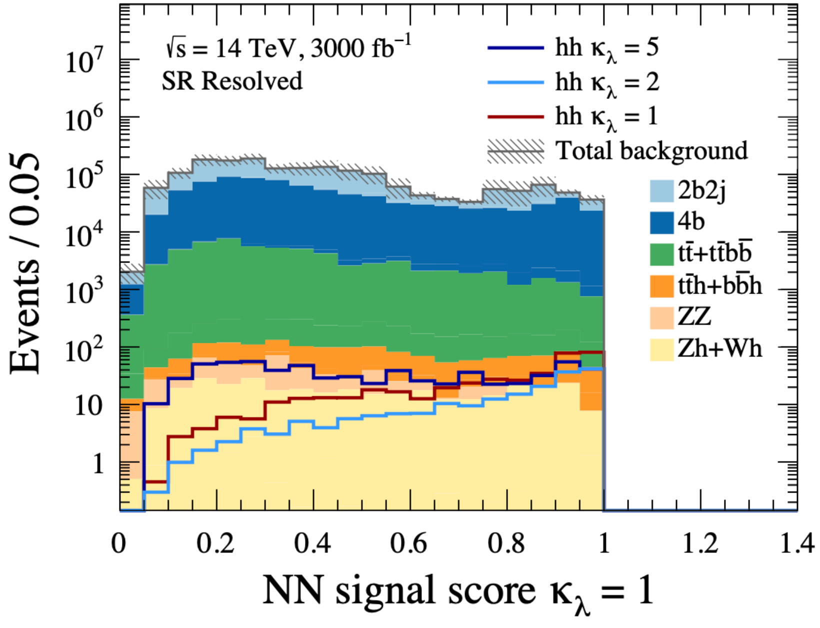
Analysis example: Higgs self-coupling measurements using ML: DNN results

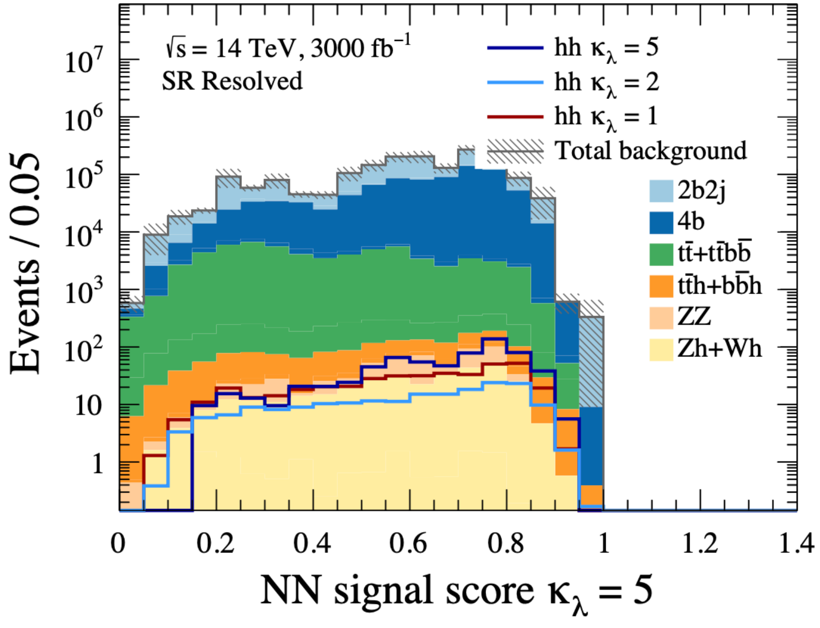
arXiv:2004.04240v3



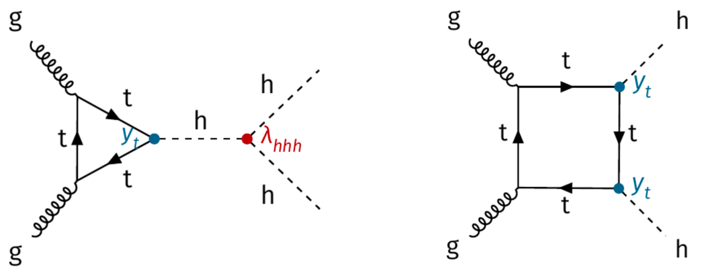
- **20 high level features as:**
candidates H candidates four-momenta, ΔR distance between the two subjets associated to Higgs candidate, b-tagging, the missing transverse momentum, the number of reconstructed electrons and muons (veto), and the mass and transverse momentum of the di-Higgs system
- DNN with 2 hidden layer
- 200 nodes for each layer
- Activation function is ReLu
- Loss is cross-entropy

- The signal vs background discrimination is improved across the categories.
- However, this depends on the value of κ_λ. DNN trained on $\kappa_\lambda = 1$ is optimal for $\kappa_\lambda = 1$ signal, but not for a $\kappa_\lambda = 5$ signal.

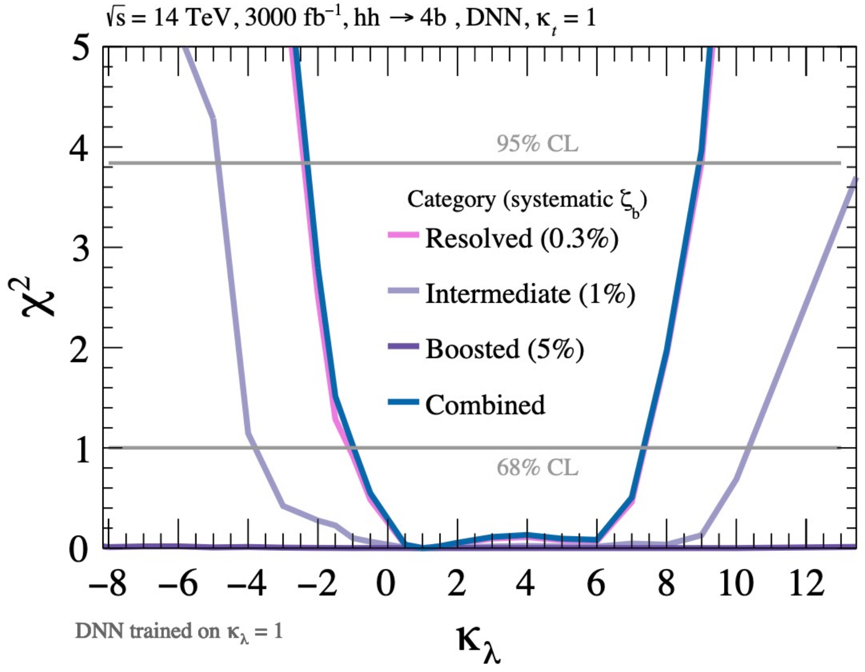




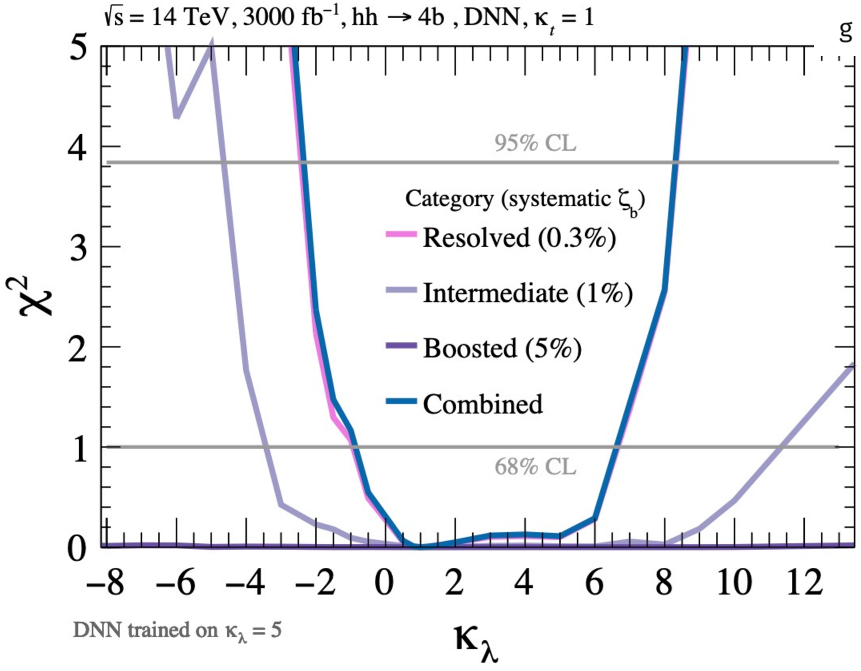
pDNN: Higgs self coupling λ_{HHH}



Parameterised neural network can be used to construct an observable that is a well- behaved function of **κλ**



(c) DNN trained on $\kappa_\lambda = 1$, nominal syst.

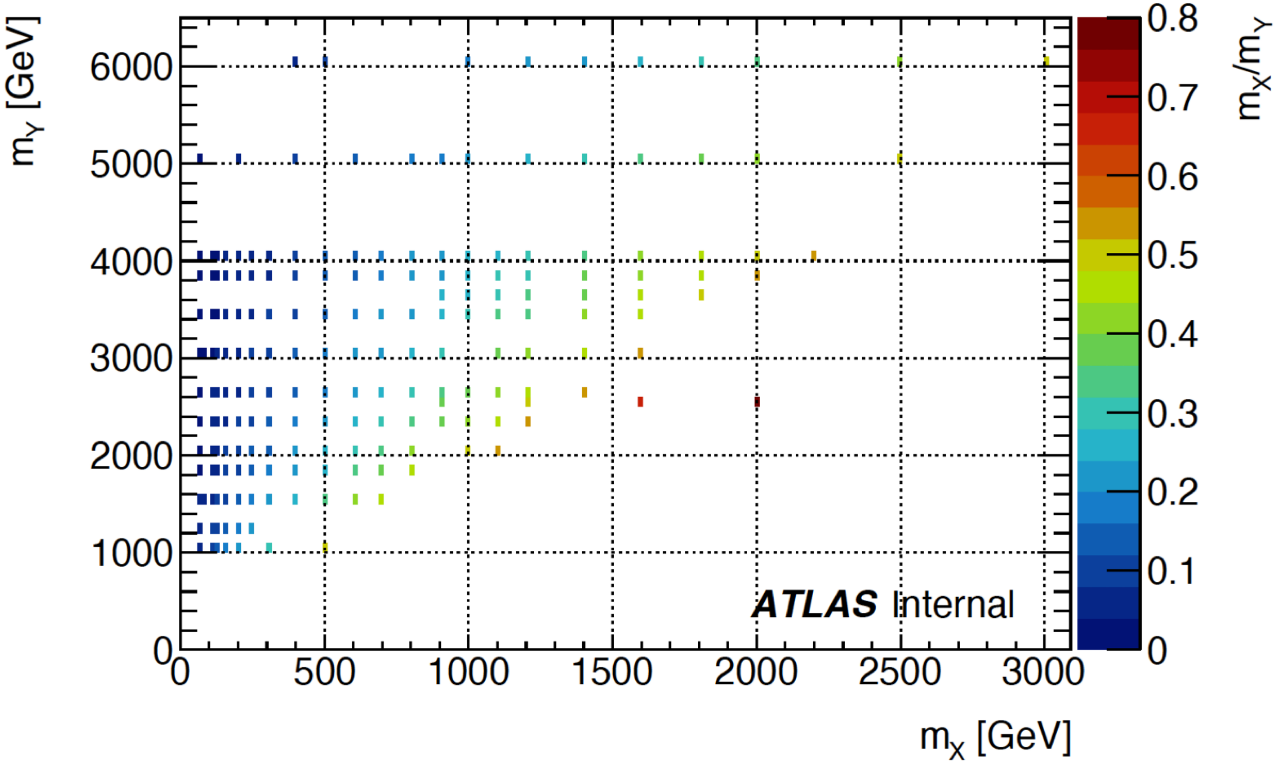


(d) DNN trained on $\kappa_\lambda = 5$, nominal syst.

- DNN training at **κλ** = 5 surpassing those achieved by the **κλ** = 1 DNN training (substantial triangle diagram contribution and therefore has a significant fraction of events at low m_{hh}).
- This underscores the importance of optimisation away from the signals with SM couplings for **κλ** constraints in the hh → 4b final state.

.

Parametrized DNN in HEP (pDNN)



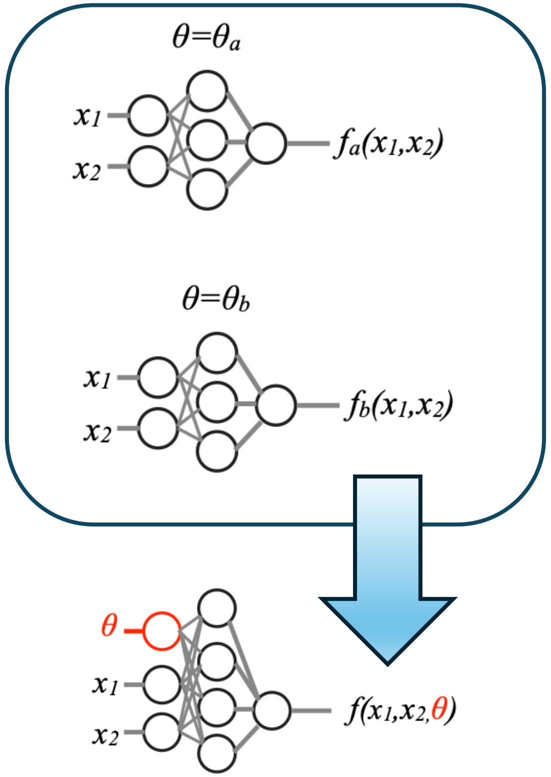
A single **parameterized network** can replace a set of individual networks trained for specific cases, as well as smoothly interpolate to cases where it has not been trained.

In a typical application of neural networks, one might consider various options:

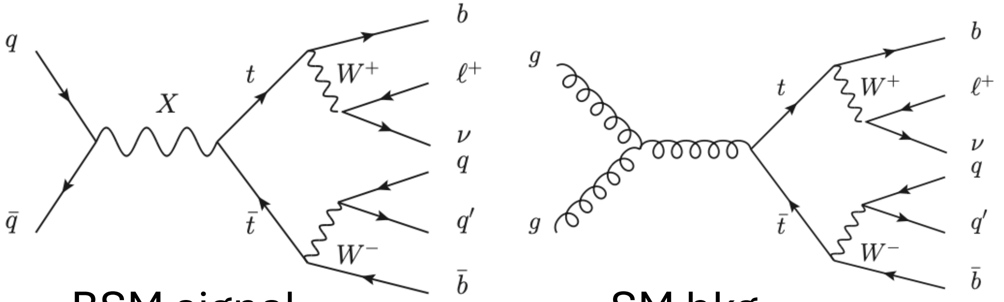
- Train a single neural network at one intermediate value of the mass and use it for all other mass values
  - quite usual but suboptimal (as in the previous example)
- Train a single neural network using a mixture of signal samples and use it for all other mass values
  - It's better but degrade performance almost everywhere
- Train a set of neural networks for a complete set of mass.
  - This approach gives the best signal-background but is time/computional expensive and doesn't interpolate

Parametrized DNN in HEP (pDNN)

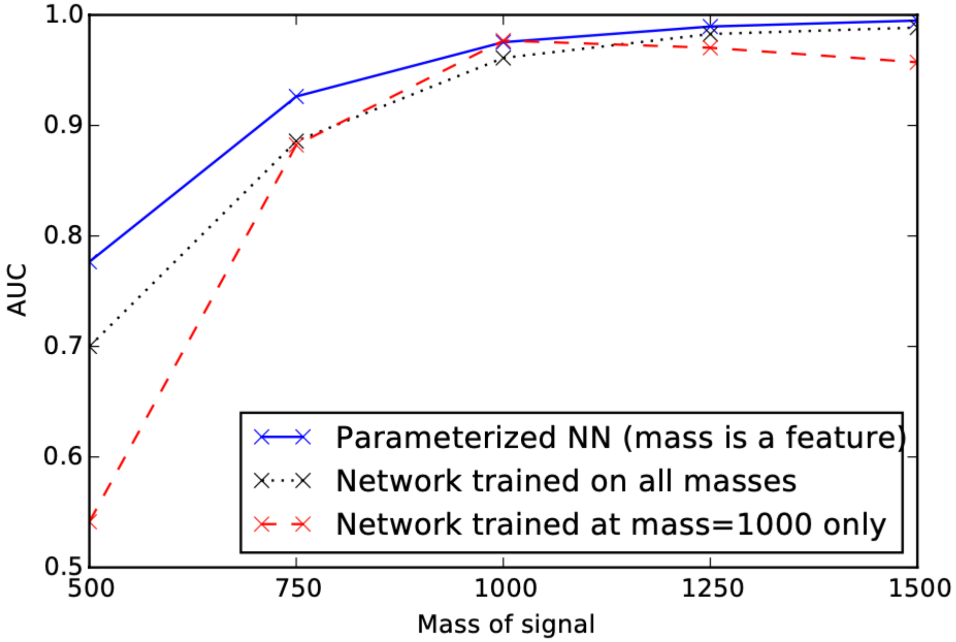
A parameterized network, however, provides a result that is parameterized in terms of **θ** : f(x₀,**θ**), yielding different output values for different choices of the parameters **θ** ;

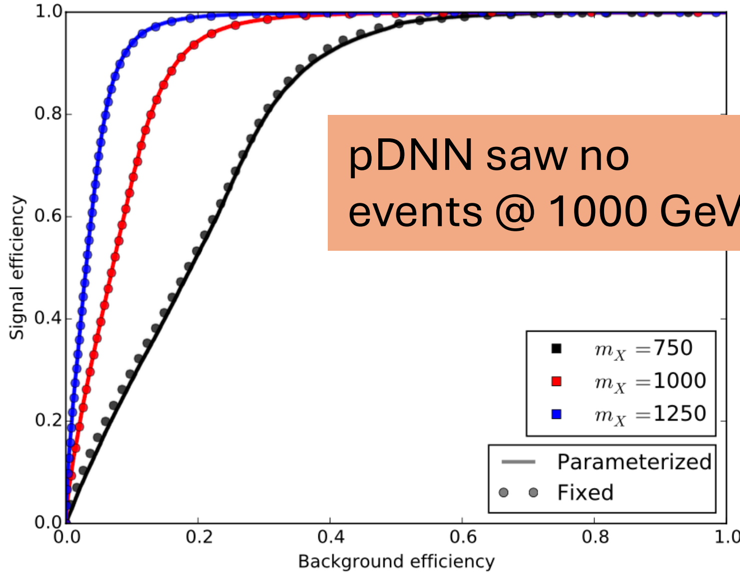


$X \rightarrow t\bar{t}$ *BSM resonance*



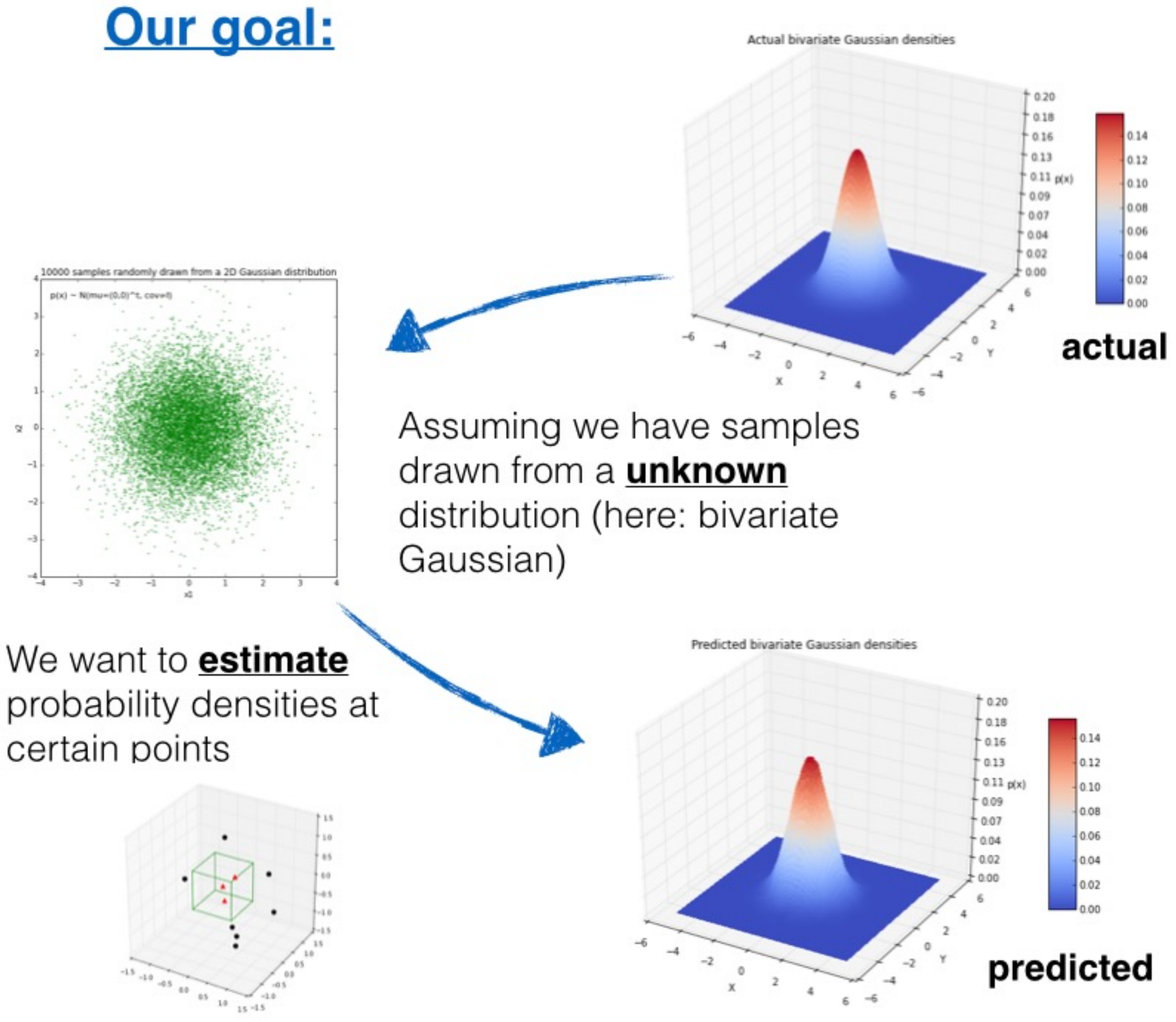
BSM signal SM bkg





pDNN saw no events @ 1000 GeV

DNN-based Reweighting procedure



The exact statistical solution to the problem of calculating the reweighting function would be to know the multi-dimensional pdfs of data.

Indeed, the weights between a region 1 and a region 2 are exactly the pdfs ratio between the two regions:

$$p_1(X) = w(X) \cdot p_0(X) \rightarrow w(X) = \frac{p_1(X)}{p_0(X)}.$$

The re-weighting function has the form of a probability density ratio (*direct importance estimation* field)
It can be directly estimated from data via a DNN algorithm, by minimizing a specific loss function

How to Direct Importance Estimation with DNN? (probability density estimation)

- Two regions 0 and 1 and #observables_X = B, N_0 measurements in region #0 and N_1 measurements in region #1
- The multidimensional pdf $p_0(\vec{X})$ and $p_1(\vec{X})$ are supposed unknown
- Our goal is the analytical ratio $w(\vec{X}) = \frac{p_0(\vec{X})}{p_1(\vec{X})}$ and our estimation: $\overline{w(\vec{X})} = \sum \alpha_f \phi_f(\vec{X})$

Loss function

$$J_0(\underline{\alpha}) = E_0[(\hat{w}(\underline{X}) - w(\underline{X}))^2] = \int p_0(\underline{X})d\underline{X}\,(\hat{w}(\underline{X}) - w(\underline{X}))^2 = \int p_0(\underline{X})d\underline{X}\,\hat{w}(\underline{X})^2 + \int p_0(\underline{X})d\underline{X}\,w(\underline{X})^2 - 2\int p_0(\underline{X})d\underline{X}\,\hat{w}(\underline{X})\cdot w(\underline{X})$$

This term does not depend on α, thus it can be ignored in the minimization

$$J_0(\underline{\alpha}) = E_0[\hat{w}(\underline{X})^2 - 2\cdot w(\underline{X})\cdot\hat{w}(\underline{X})] = E_0[\hat{w}(\underline{X})^2] - 2\cdot E_1[\hat{w}(\underline{X})]$$

If $\overline{w(\vec{X})}$ is intended as the output of the network (it is effectively a function of data) and the constants α as the weights of the network the loss function J_0 can be effectively minimized by a Neural Network algorithm!

How to Direct Importance Estimation with DNN?

- Two regions 0 and 1 and #observables_X = B, N_0 measurements in region #0 and N_1 measurements in region #1
- The multidimensional pdf $p_0(\vec{X})$ and $p_1(\vec{X})$ are supposed unknown
- Our goal is the

★ Intrinsic multi-dimensionality

★ Event-level weights, instead of histogram-level as in the previous an analysis

Loss function

$J_0(\underline{\alpha}) = E_0[(\hat{w}(\underline{X})$

This term does not depend on α, thus it can be ignored in the minimization

tended as the output
rk (it is effectively a
ata) and the
constants α
as the weights of the network the loss function J_0 can be effectively minimized by a Neural Network algorithm!

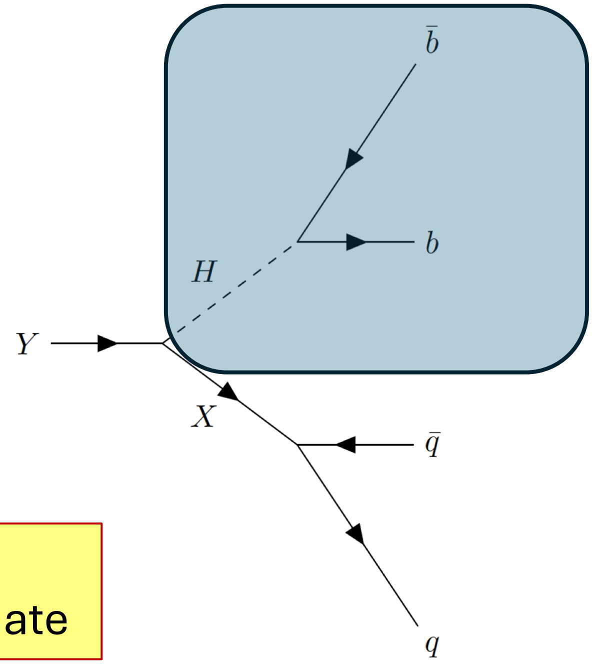
$$J_0(\underline{\alpha}) = E_0[\hat{w}(\underline{X})^2 - 2 \cdot w(\underline{X}) \cdot \hat{w}(\underline{X})] = E_0[\hat{w}(\underline{X})^2] - 2 \cdot E_1[\hat{w}(\underline{X})]$$

Real Life example of DNN reweighting

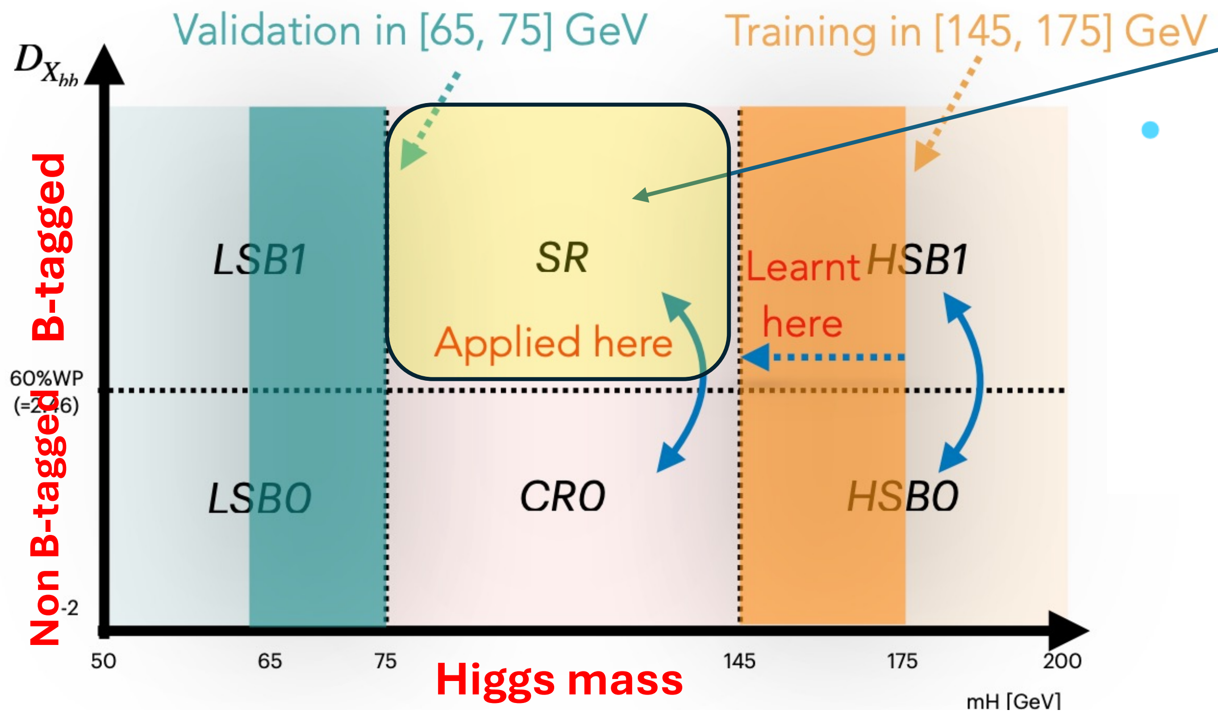
Background is dominated by QCD di-jet processes

Data-driven background estimation required to not incorporate MC mismodelling into background estimation

Strategy: a reweighting function is learnt by a Deep Neural Network (DNN) in HSB, validated in LSB and extrapolated in Higgs mass window to reweight control region to SR



Here you select a b-tagged Higgs candidate

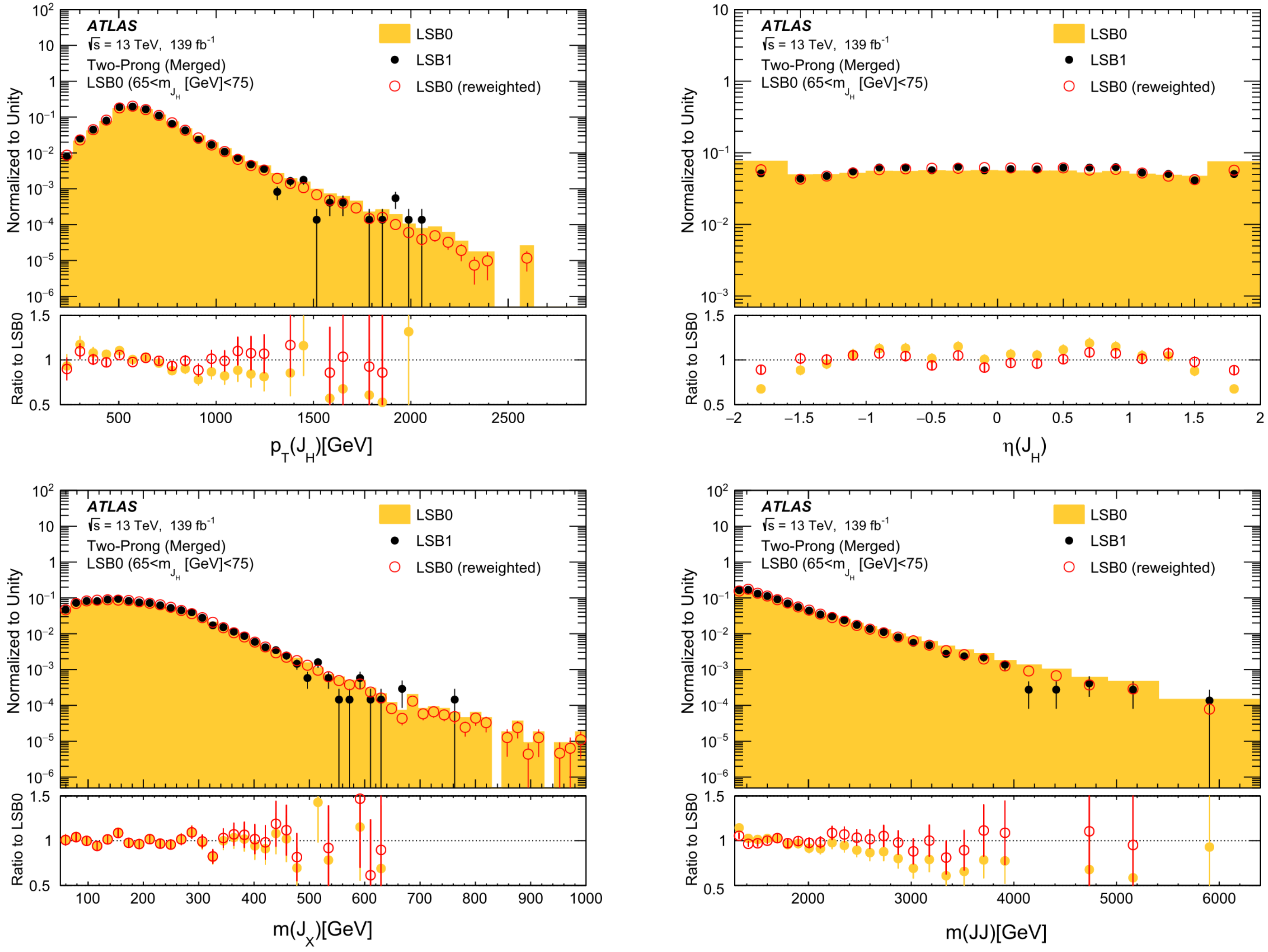


- Inputs:
  - H four-momentum
  - Ntracks associated to h
  - p_T leading and sub-leading track jets associated to H
    - p_T, η, ϕ, m

No feature from X candidates has been used!!

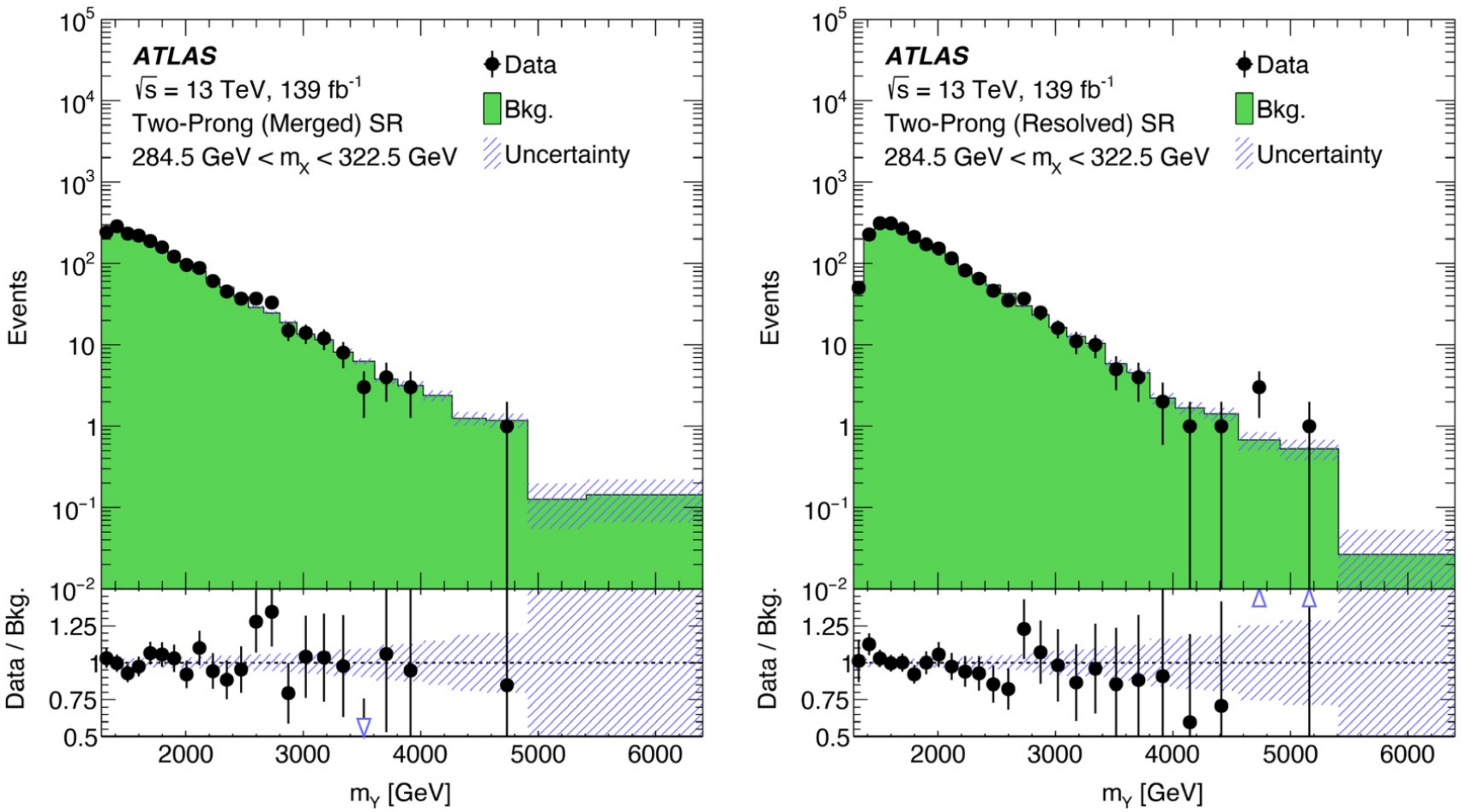
Phys. Rev. D 108, 052009 – Published 18 September 2023

DNN reweighting results in Validation Regoin



DNN reweighting results in Signal region

G. AAD *et al.* PHYS. REV. D **108,** 052009 (2023)



Background Uncertainties

Phys. Rev. D 108, 052009 – Published 18 September 2023

Background Uncertainties

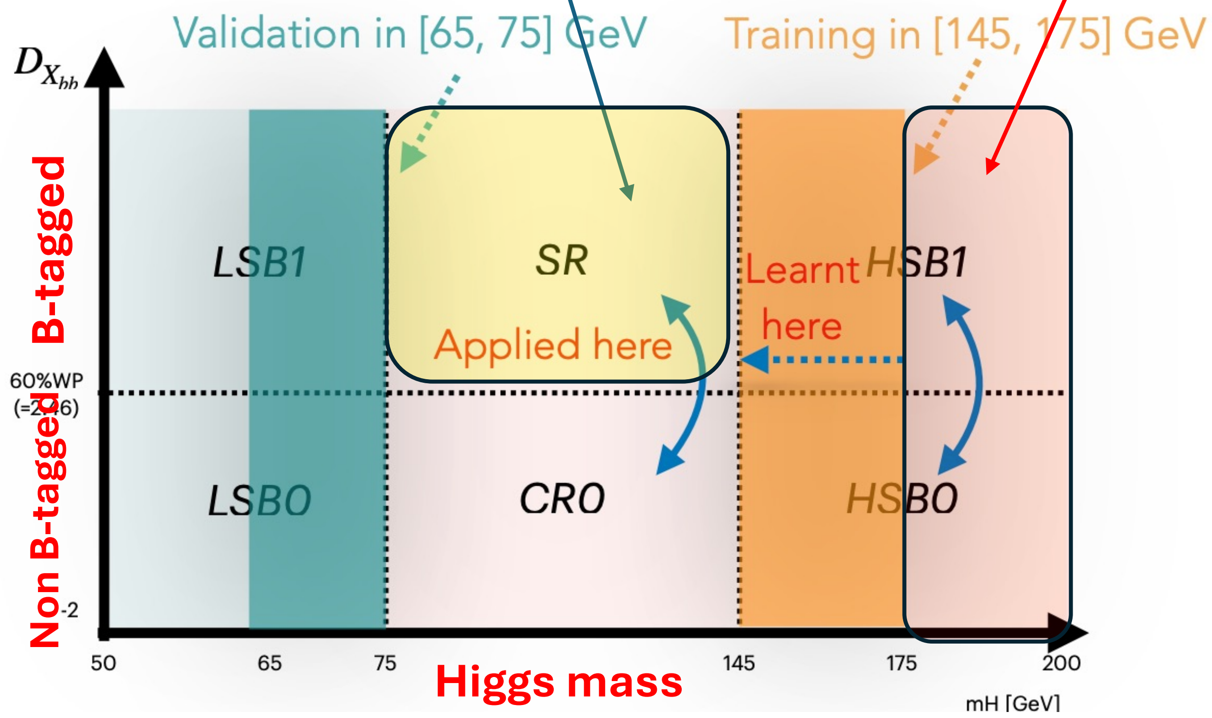
- **No standard Combined Performances from Atlas Recostruction or MC-based systematics on background, since it is fully data-driven.**

- **Three kinds of uncertainties considered:**
  1) Systematic, on the **choice of the training region** (~5-10%)
  2) Statistical, intrinsically related to the **training procedure** (<10%). Summed in square   with the Poissonian error in each bin
  3) Systematic, on the **extrapolation** of predictions across bins (~10%)

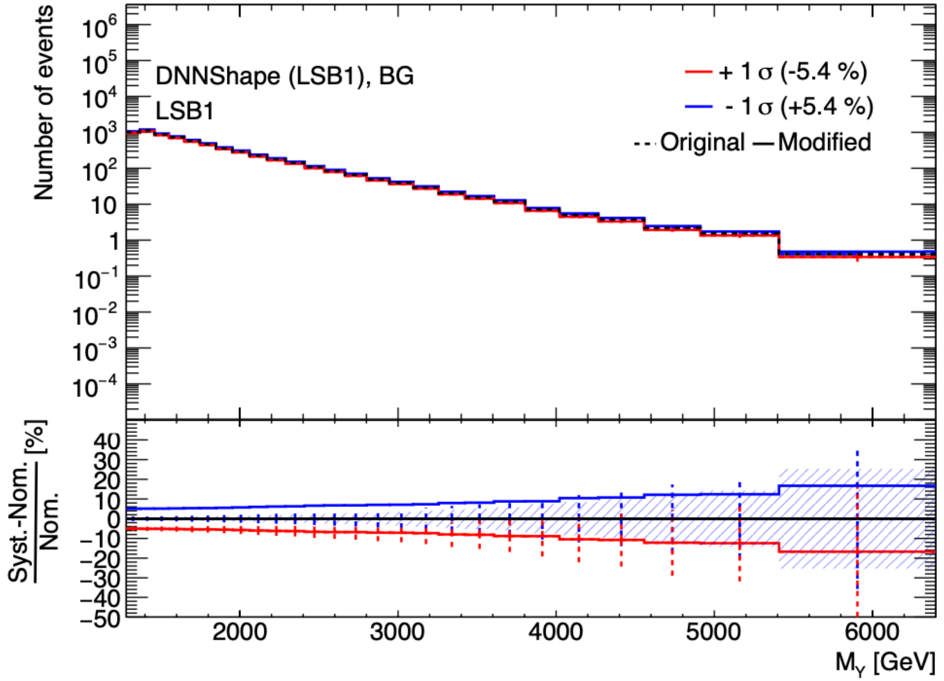
1) Shape Uncertainties - Training Region

Predictions in the SR may be different if the region used for training changes
To quantify this mismodelling, an additional kinematic region (in [165, 200] GeV) is used to train an alternate model (totally identical to the nominal one, only changing the training region)
The ratio of the alternate shape to the nominal shape is determined as the NN modeling shape uncertainty

Here you select a b-tagged Higgs candidate

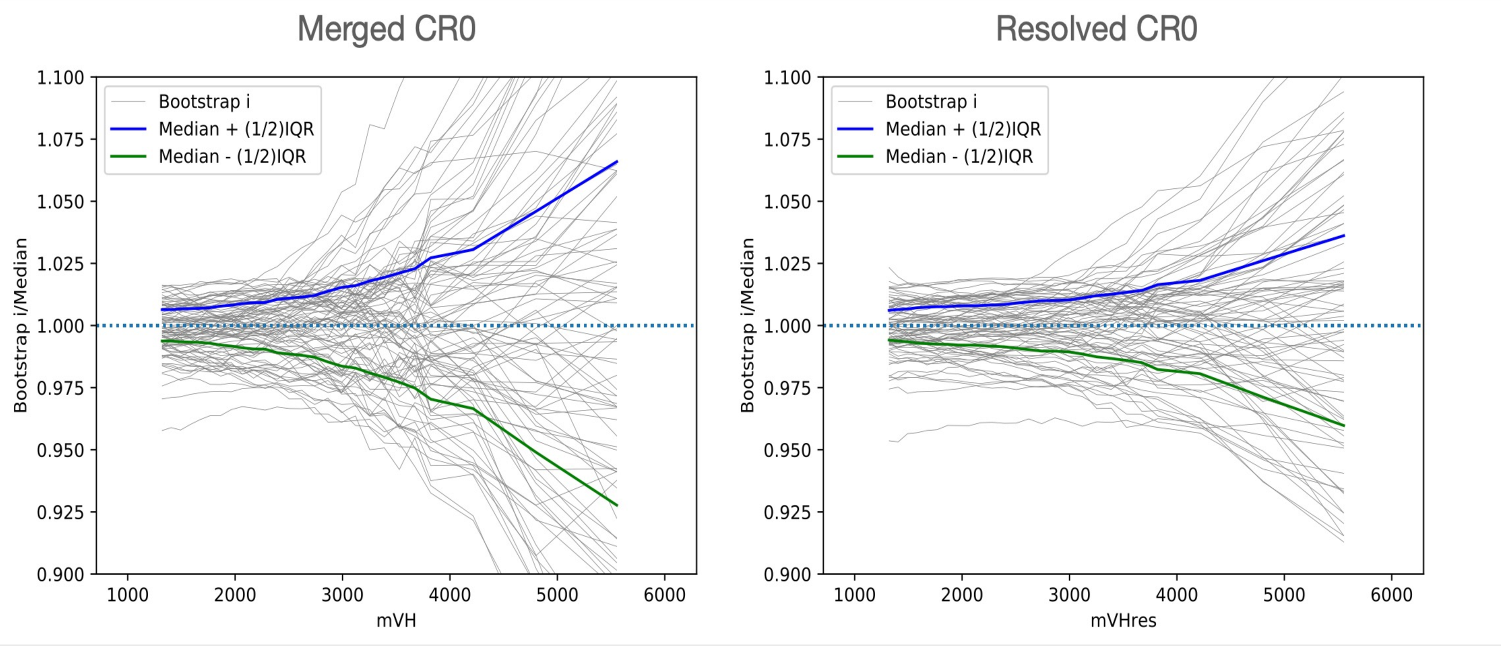


The ratio of the alternate shape to the nominal shape is determined as the DNN modeling shape uncertainty



Bootstrap for Statistical Error

- The DNN is trained on a sample with a finite number of events and the weights of the network are randomly initialized at the beginning of the training
- Uncertainty estimated repeating the training N=100 times, randomly sampling the training dataset
  - Nominal histogram reweighed with the median of weights distribution for each event and normalized with the median of the normalization factors
  - Up/down variations obtained taking the median +- half the interquartile range (IQR) of weights distribution for each event and normalized with the median +-IQR/2 of the normalization factors



*more details in backup

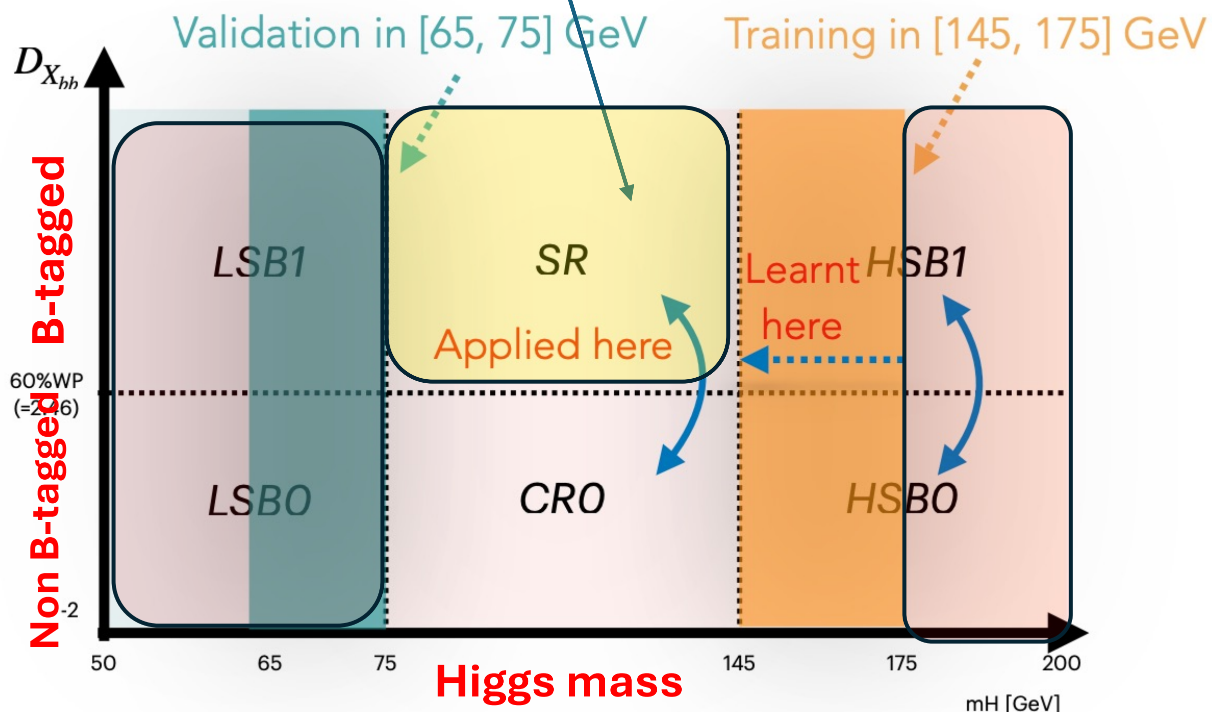
2) Shape Uncertainties – Extrapolation (Non Closure test)

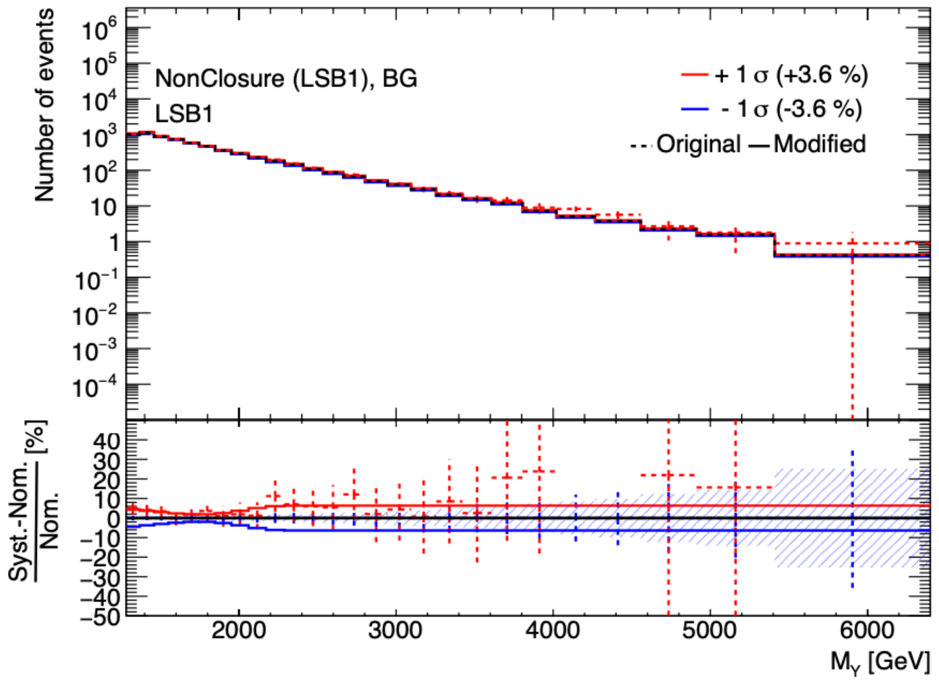
Weights extrapolation process from the training region to the SR may be an additional source of mismodelling.

Since it is not possible to directly estimate the discrepancy between reweighed data and the target distribution in **SR**, it is determined by looking at the ratio of data to estimated background in **LSB** (LSB1 over reweighed LSB0)

It is defined by the symmetrized shape difference between the data and predicted background in the LSB. The non-closure is negligible for low mJJ and rises to O(10)% in the mJJ tails.

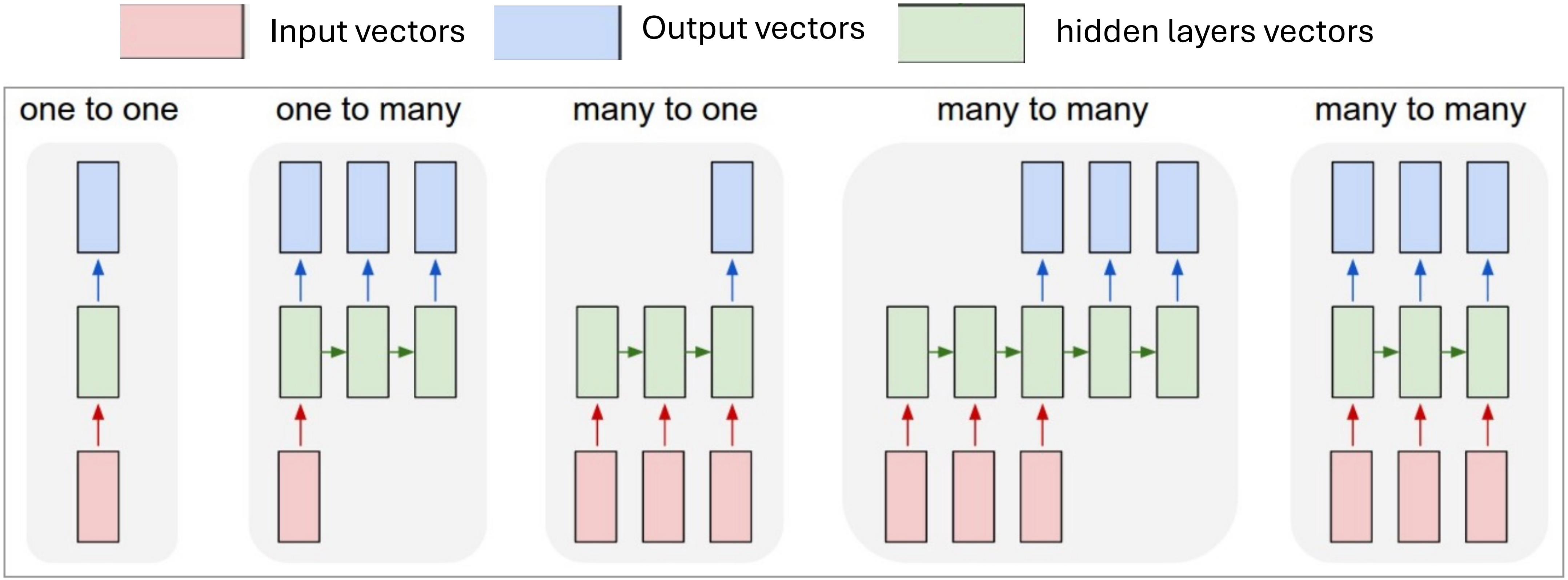
Here you select a b-tagged Higgs candidate





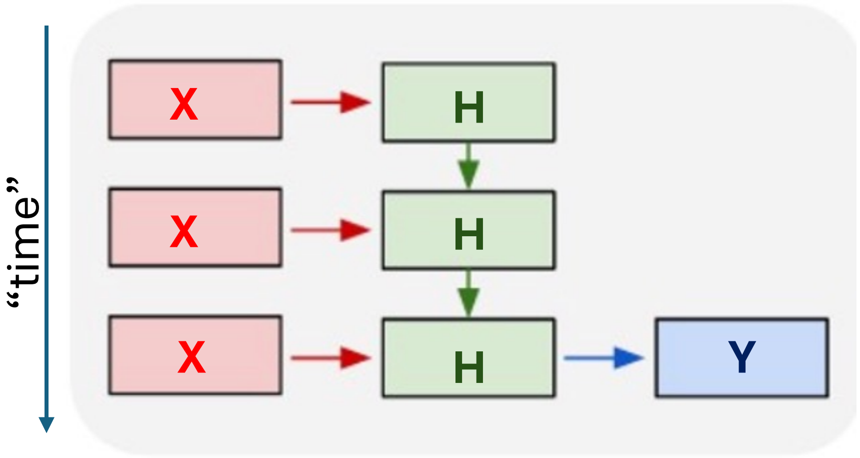
RNN: Recurrent Neural Networks

- Two main limitations from standard NN:
- Standard NN (and also Convolutional Networks) are constrained to accept fixed-sized vector as input (e.g. an image) and produce a fixed-sized vector as output (e.g. probabilities of different classes)
- These models perform this mapping using a fixed amount of computational steps (e.g. the number of layers in the model)



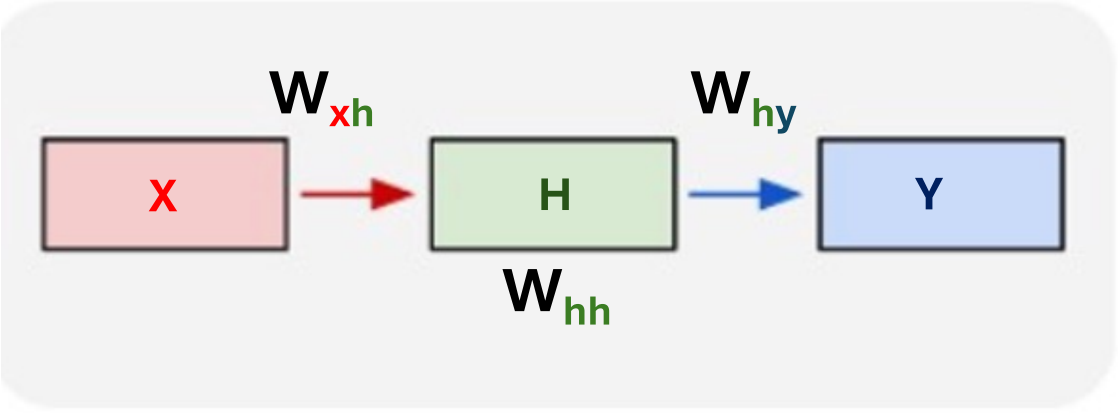
Standard NN | **Image to sequence** | **Sequence to single out** | **Sequence to sequence output** | **Synced sequence input/output**

What's "new" about RNN algorithm?



The output vector's **Y** contents are influenced not only by the input **X** you just fed in, but also on the entire history of inputs you've fed in in the past.

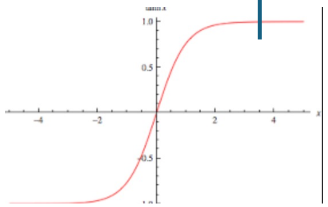
- Let's see how this mechanism works in the simplest case with a **single** *hidden* vector h:



When you feed the RNN with your step t input **X(t)** the hidden state **H(t)** is updated this way:

$$H(t) = \tanh(\boxed{W_{hh}H_{t-1}} + \boxed{W_{xh}H_t})$$

based on the past history **based on the current step**

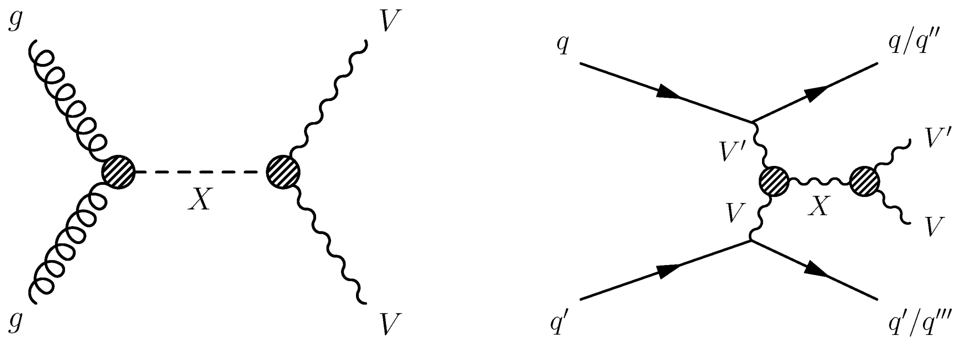


Activation function

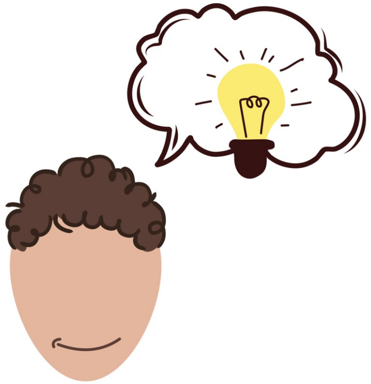
In practice most of us use a slightly different formulation called a ***Long Short-Term Memory* (LSTM)** network. The LSTM is a particular type of RNN that works slightly better with more powerful update equation and some appealing backpropagation dynamics.

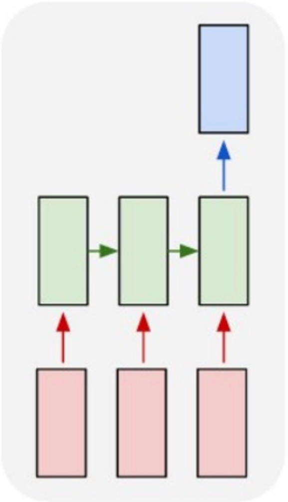
RNN real life application

Eur. Phys. J. C **2020**, 80, 1165
https://www.mdpi.com/2076-3417/13/5/3282

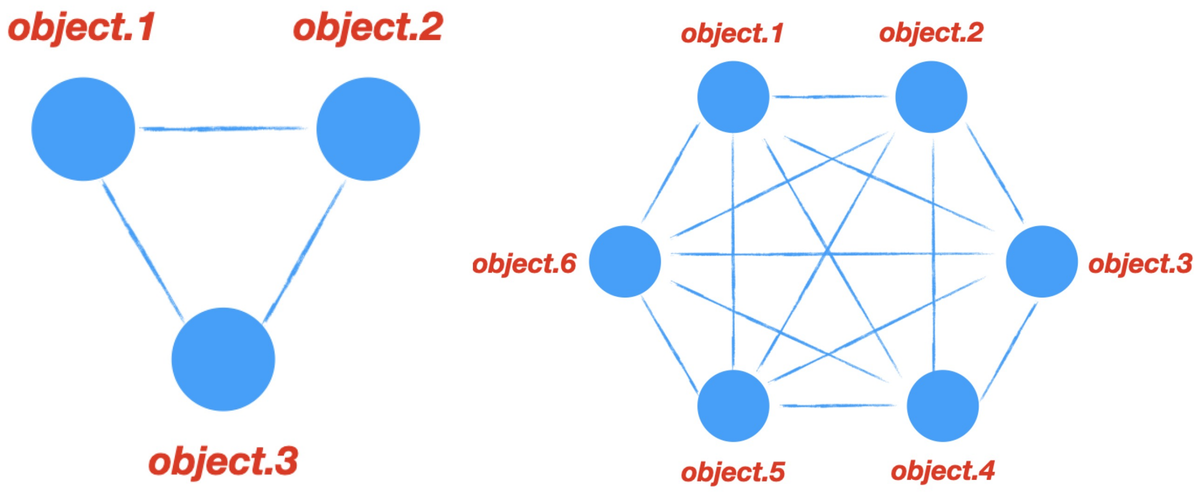


Given the complexity of the LHC and the detector effects each event has a **different jet multiplicity**, that may vary up to about tens of jets.



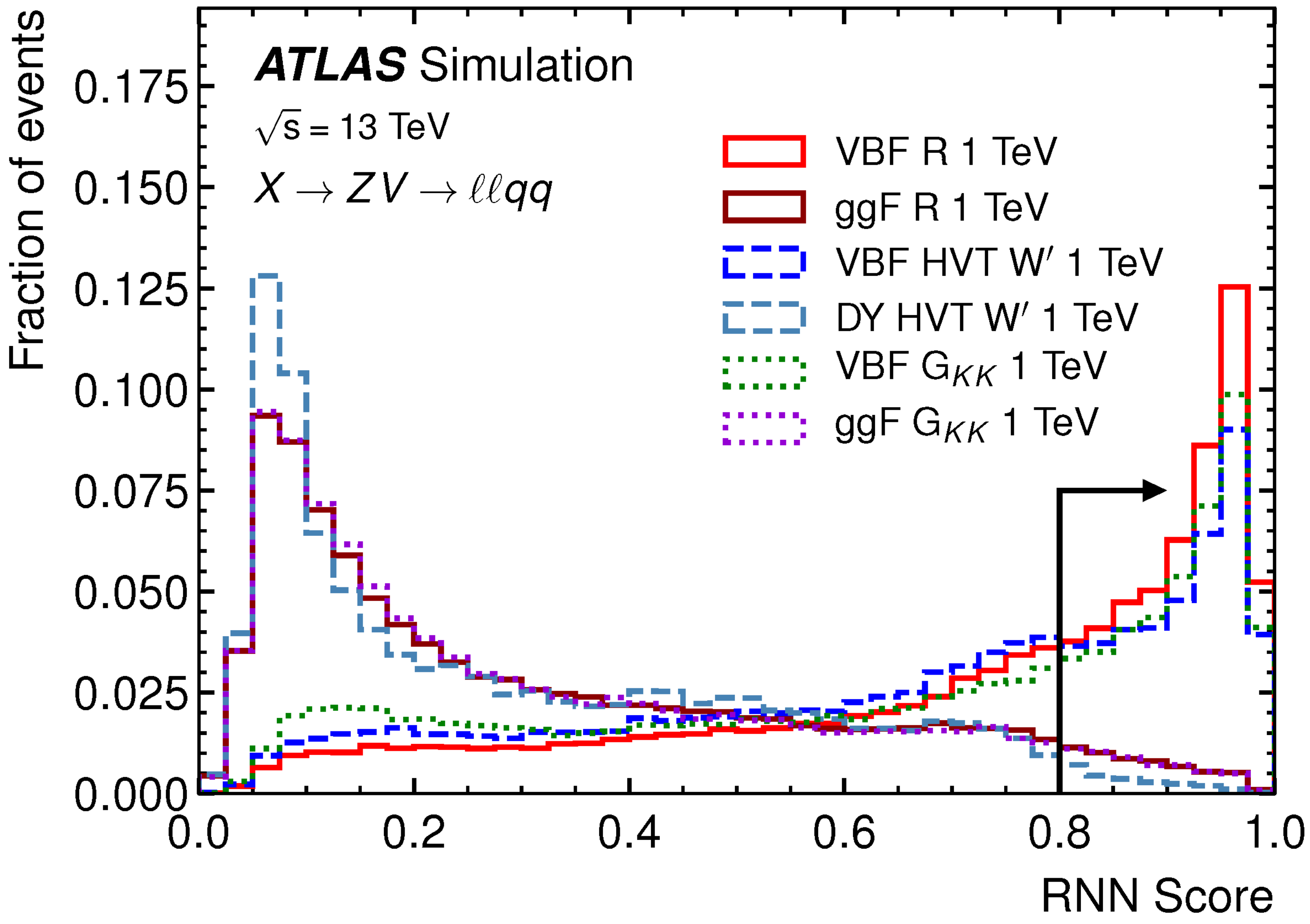


- Classification VBF events was strongly driven by the identification of **2** jets in the events that are directly related to the topology of the process.
- This represents a combinatorial problem in which all the possible pairs of objects may diverge quickly as the total amount of objects (N) available increases.
- For instance, with N=3 only 3 pairs are possible, but with N=6 the number of possible pairs increases to 15

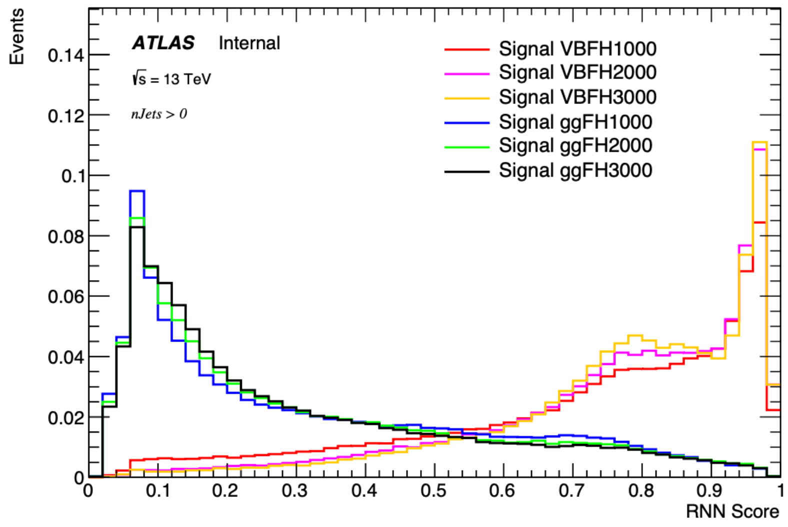


(a) (b)

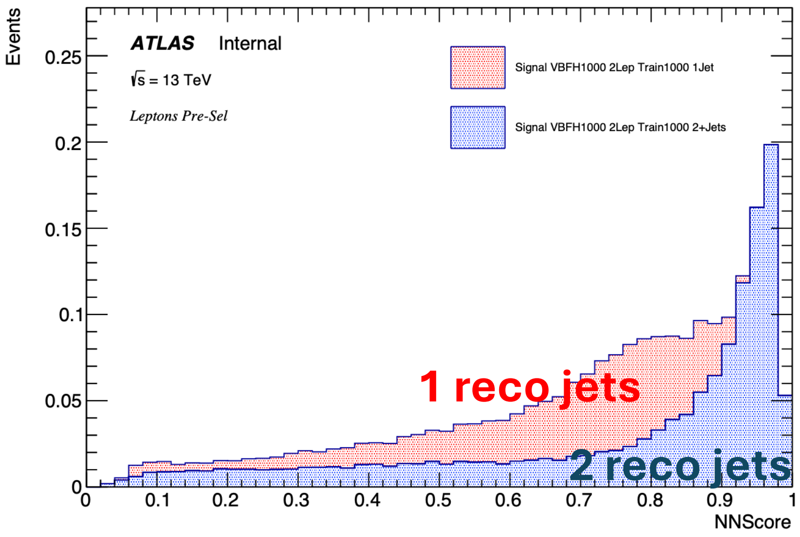
RNN real life application



Weak dependency on BSM resonance mass



1 or 2 VBF reco jets??

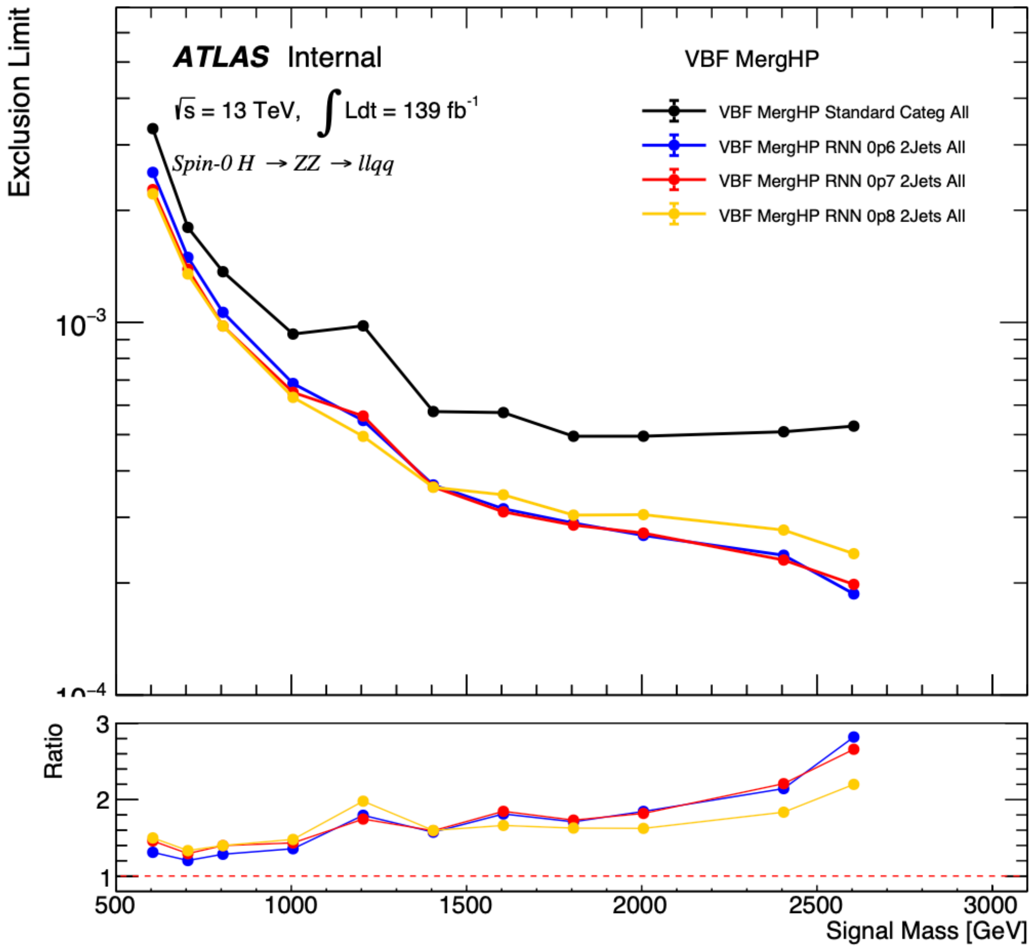


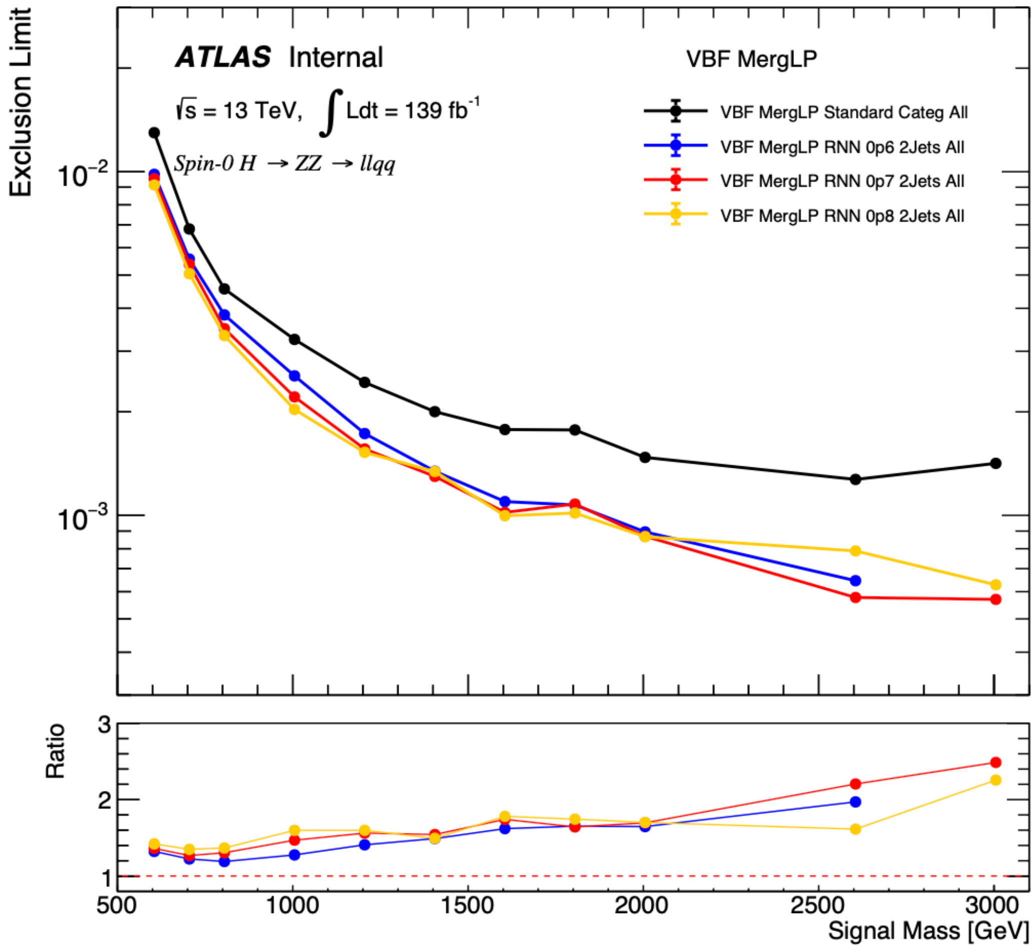
(b)

RNN real life application:

A. Giannini, A. Machine Learning Methods for Diboson Searches in Semi-Leptonic Final States with the ATLAS Experiment at LHC. Ph.D. Thesis

Expected exclusion limits to VBF signals



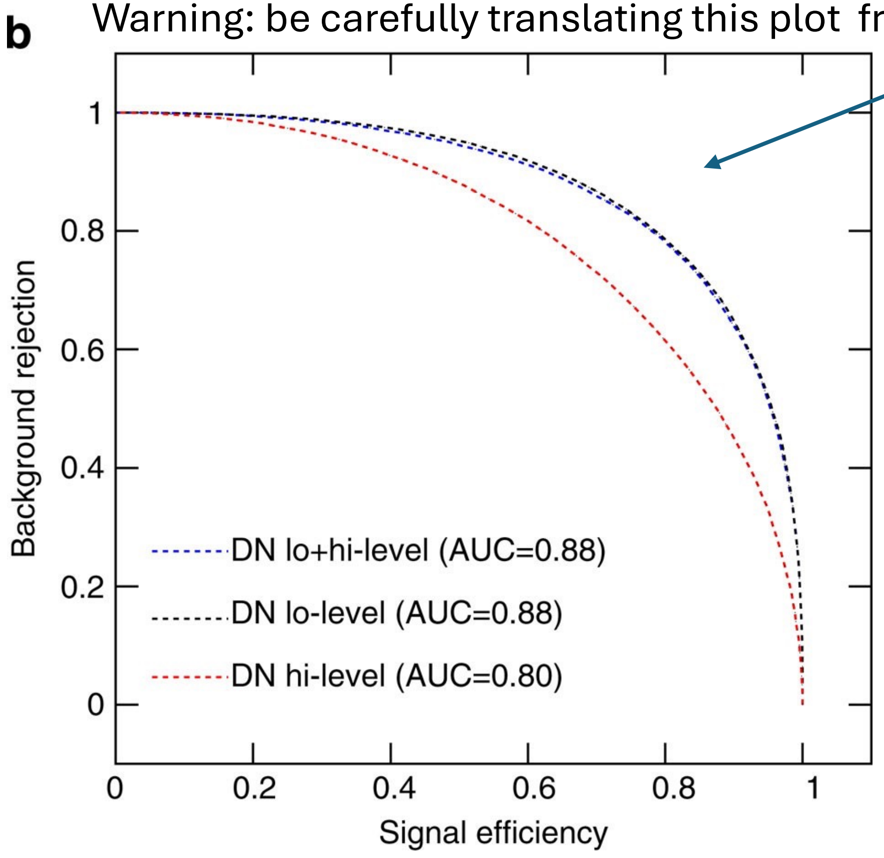


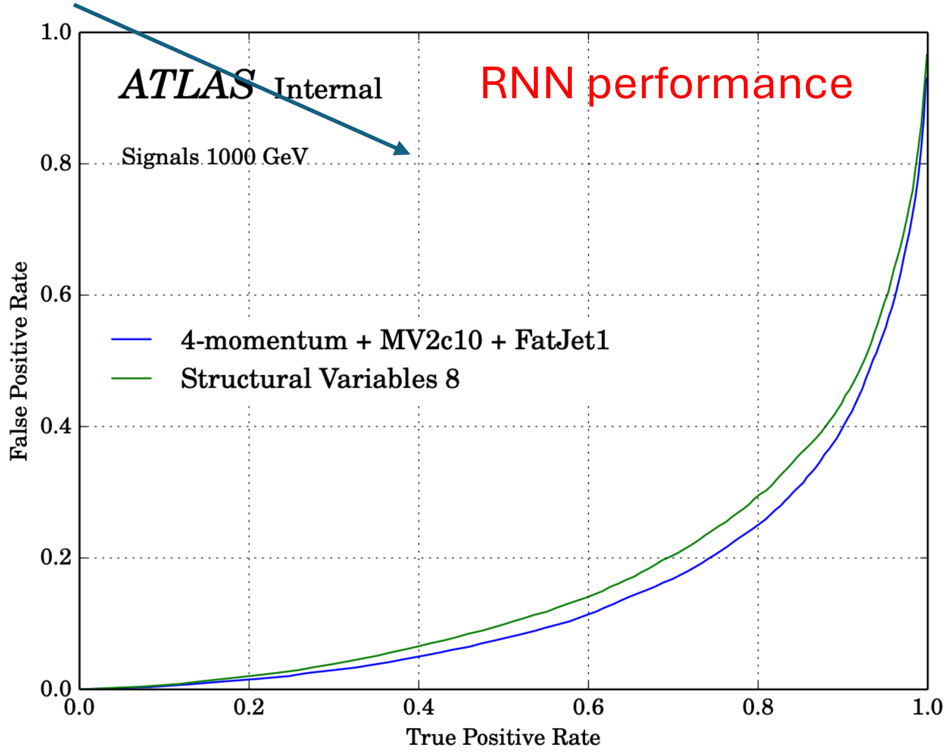
Low-level vs High-level features

From 2014 outbreaking result of Baldi, Sadowski and Whiteson as a result of the deep learning:

"A set of features with basic information (low-level) such as information coming directly form the detectors implies better performances wrt features built combining basic information (high-level)"

Warning: be carefully translating this plot from here to there





Theoretical and **experimental** questions motivate a deep exploration **of the fundamental structure of nature**

Dark matter

Hierarchy problem

Strong CP

Flavor puzzles

Dark energy

Baryogengesis

We have performed thousands of hypothesis tests & have no significant evidence for physics beyond the Standard Model

Three possibilities

(1) There is nothing new at accessible energies

(2) Patience! (new physics is rare)

(3) We are not looking in the right place

There are two complementary paths forward:

(1) Identify new, specific, well-motivated places to look

This is still an incredibly important direction and has resulted in new directions like long-lived particle searches

(2) Look in many places all at once

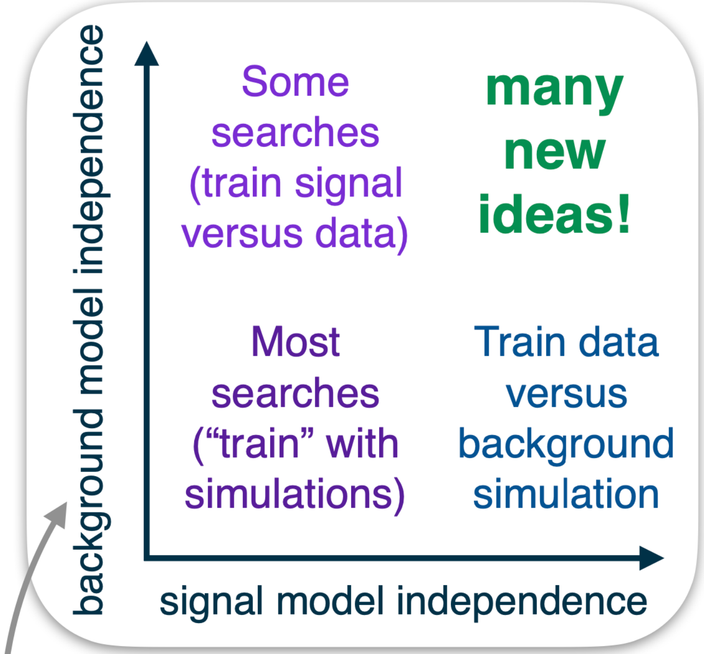
(1) There is nothing new at accessible energies

(2) Patience! (new physics is rare)

(3) We are not looking in the right place

Anomaly detection and model independent searches

[Nature Reviews Physics (2022), 2112.03769]



Signal sensitivity

Standard Model

Unsupervised = no labels
Weakly-supervised = noisy labels
Semi-supervised = partial labels
Supervised = full label information

This is most searches. You simulate the signal (label = 1), simulate the background (label = 0) and "train" a classifier to distinguish the 1's from the 0's.

There are many new ideas that make use of modern machine learning

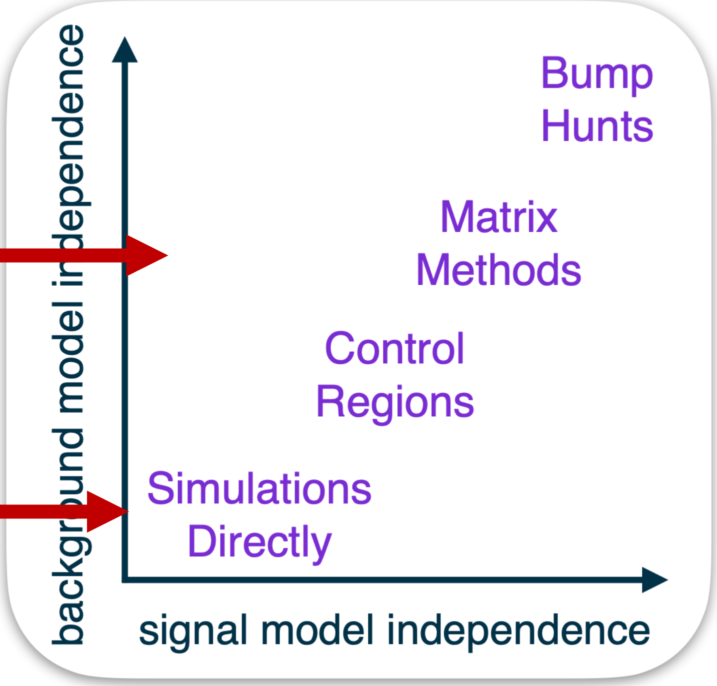
The goal is to learn **directly from data**, injecting as little bias as possible

Core idea:

create a reference sample and see if our target and reference are the same; if yes, limits; if no, discovery!

Here ML can help!

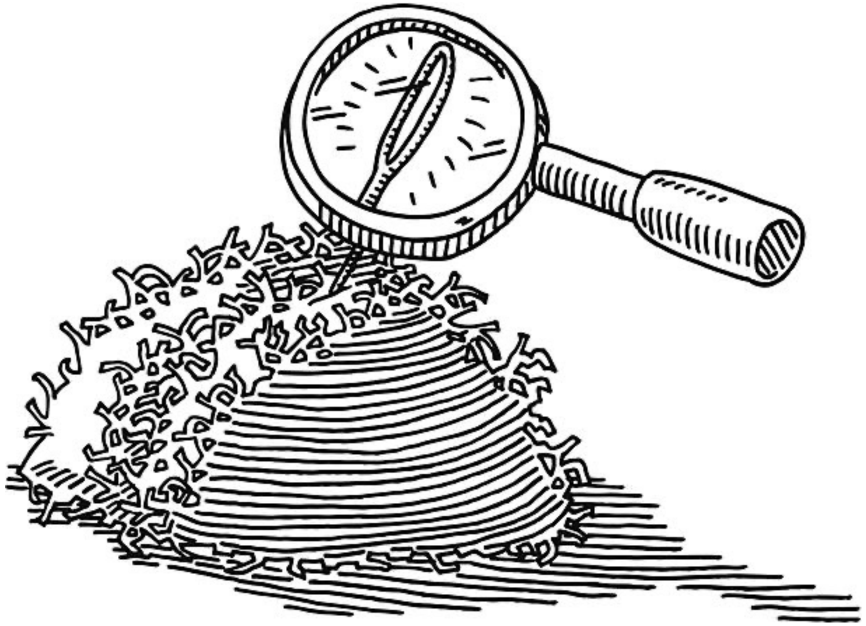
Almost no searches at the LHC with a few exceptions for very well-known processes like 4-lep

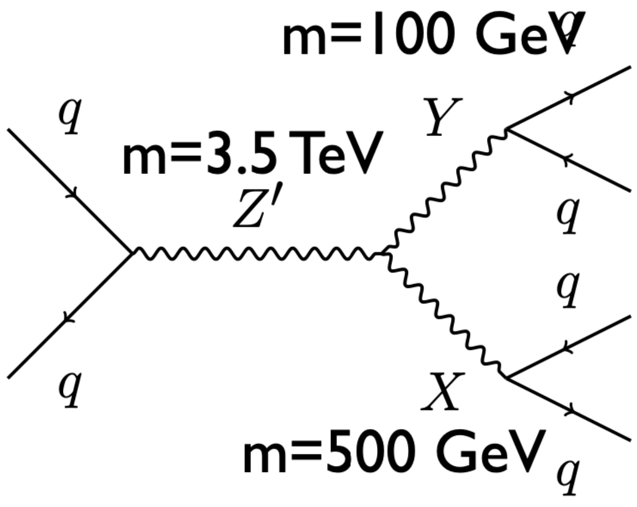


Background specificity

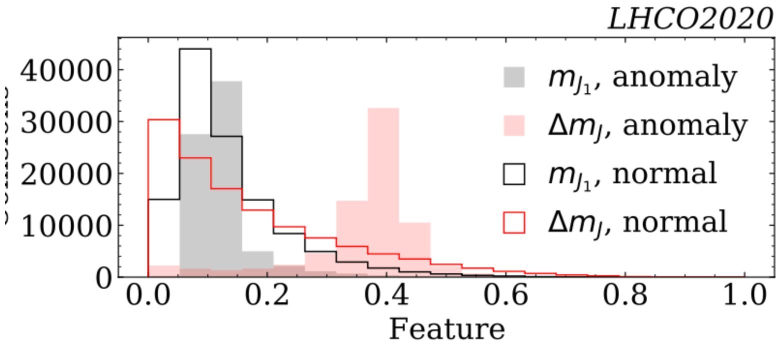
[Anomaly detection for LHC]

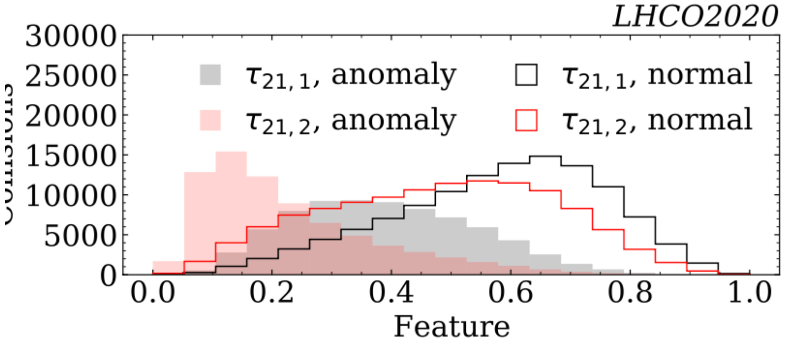
Relatively simple signal and well know features for s/b separation…but are we able to catch signal with an unsupervised approach? Or we lose it?





This is for R&D and training step

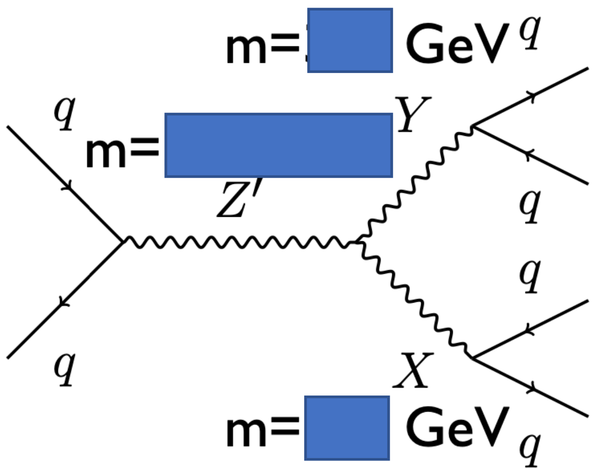


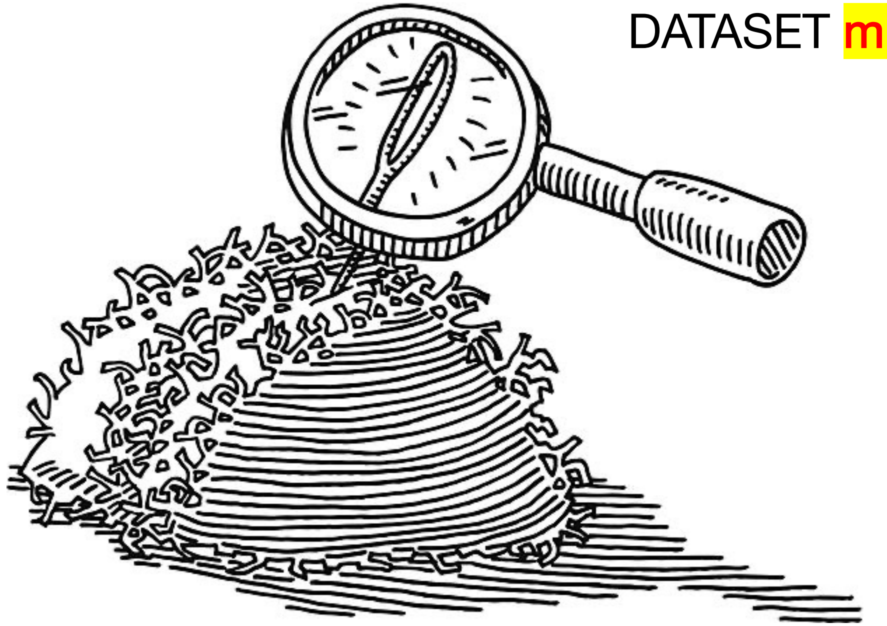


[LHC Olympics paper]

Anomaly detection for LHC

Relatively simple signal and well know features for s/b separation…but are we able to catch signal with an unsupervised approach? Or we lose it?

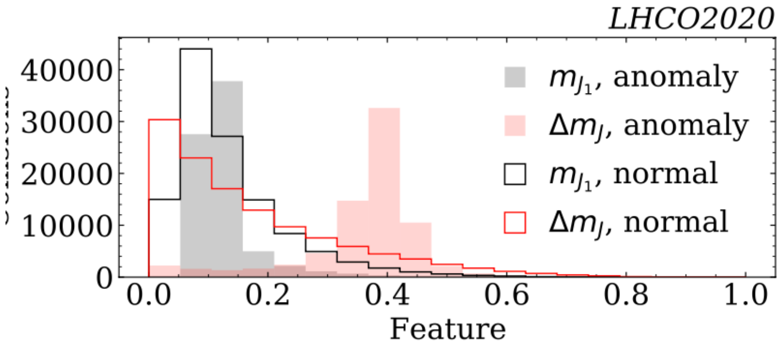


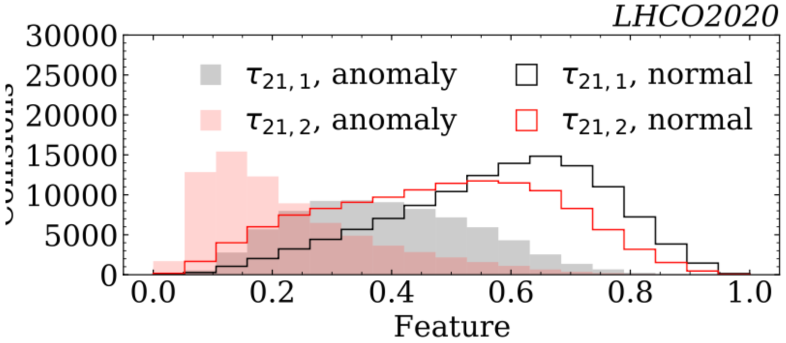


DATASET might contain signal

Short Name	Method Type
VRNN	Unsupervised
ANODE	Unsupervised
BuHuLaSpa	Unsupervised
GAN-AE	Unsupervised
GIS	Unsupervised
LDA	Unsupervised
PGA	Unsupervised
Reg. Likelihoods	Unsupervised
UCluster	Unsupervised
CWoLa	Weakly Supervised
CWoLa AE Compare	Weakly/Unsupervised
Tag N' Train	Weakly Supervised
SALAD	Weakly Supervised
SA-CWoLa	Weakly Supervised
Deep Ensemble	Semisupervised
Factorized Topics	Semisupervised
QUAK	Semisupervised
LSTM	Semisupervised

This is for TEST!!

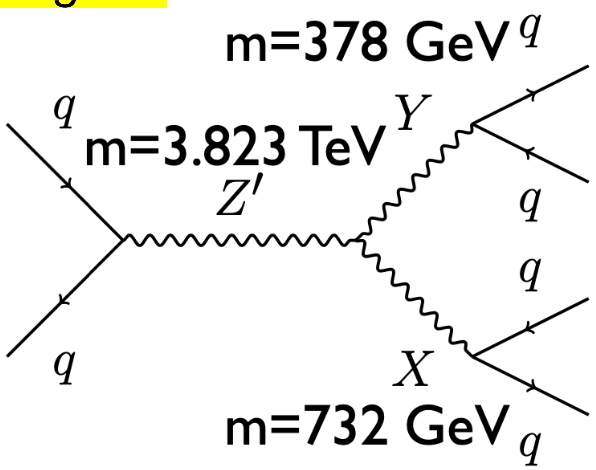




LHC Olympics paper

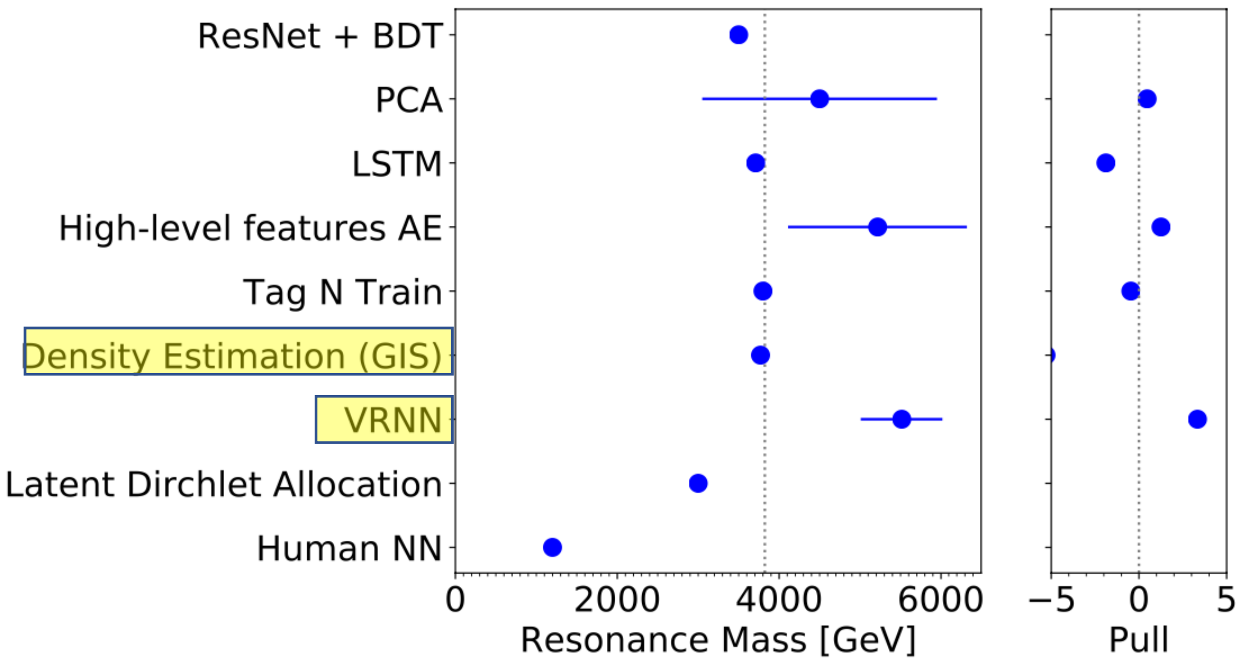
Anomaly detection for LHC

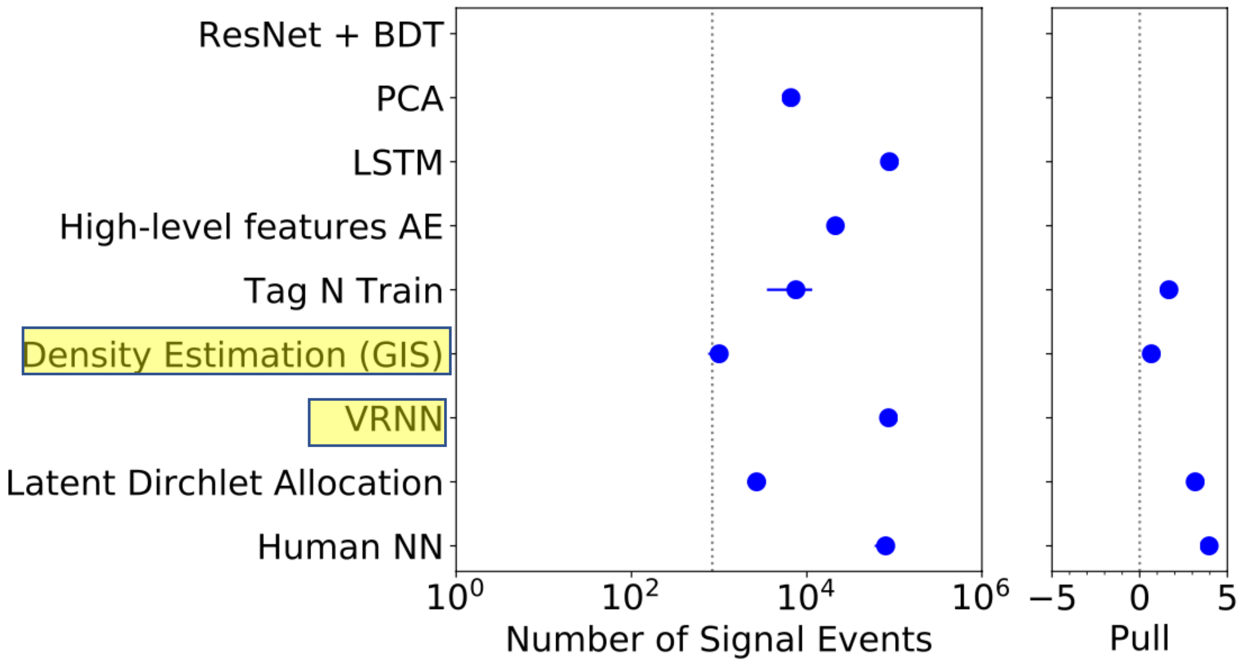
DATASET contain signal!!



Relatively simple signal and well know features for s/b separation…but are we able to catch signal with an unsupervised approach? Or we lose it?

This is for TEST!!





Anomaly detection for LHC

Relatively simple signal and well know features for s/b separation…but are we able to catch signal with an unsupervised approach? Or we lose it?

Different observations claimed, none identified the correct value of the mY signal….

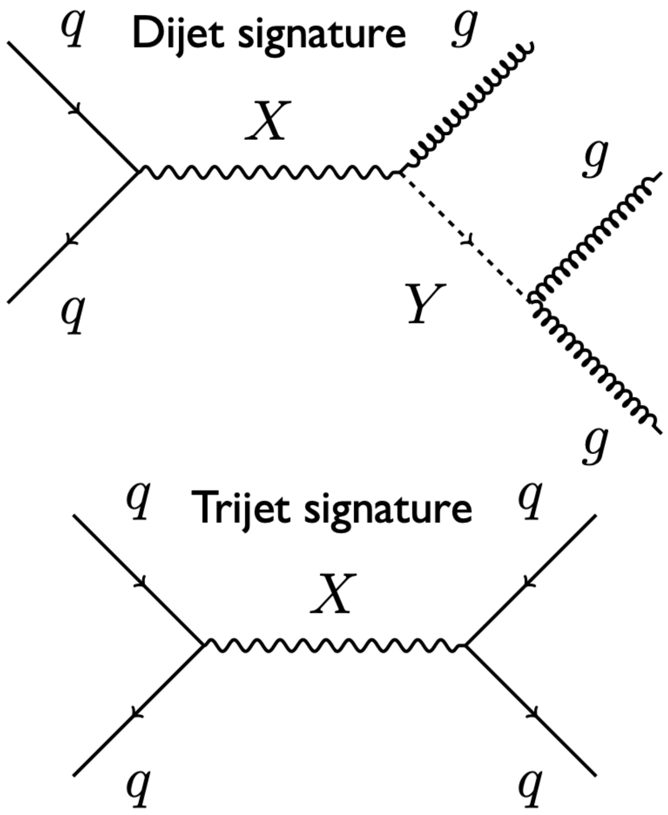
Lesson learned

- Anomaly detection is difficult
  - Even for “anomalies” close to already considered signals
  - Even more so for “exotic” signals
  - Value in blind studies

LHC Olympics paper

DATASET contain signal!!

mX = 4.2 TeV and two decay modes:
1200 signal events in di-jet signature
2000 signal events in tri-jet signature
(finding individual excess should not yield significance)



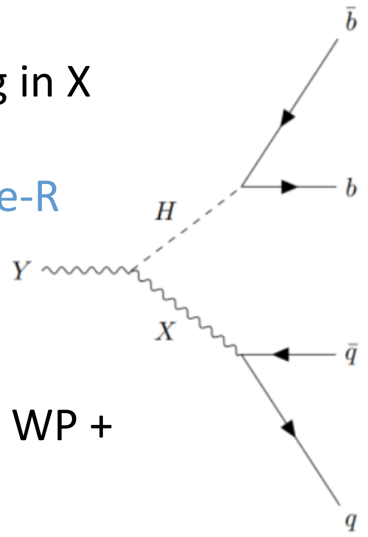
First application of Anomaly detection in ATLAS

[Phys. Rev. D 108, 052009 – Published 18 September 2023]

- Model-independent discovery region introduced with novel data-driven anomaly score (AS)
- AS determined from fully unsupervised variational recurrent neural network (VRNN) trained over jets modeled as sequence of constituent four-vectors.

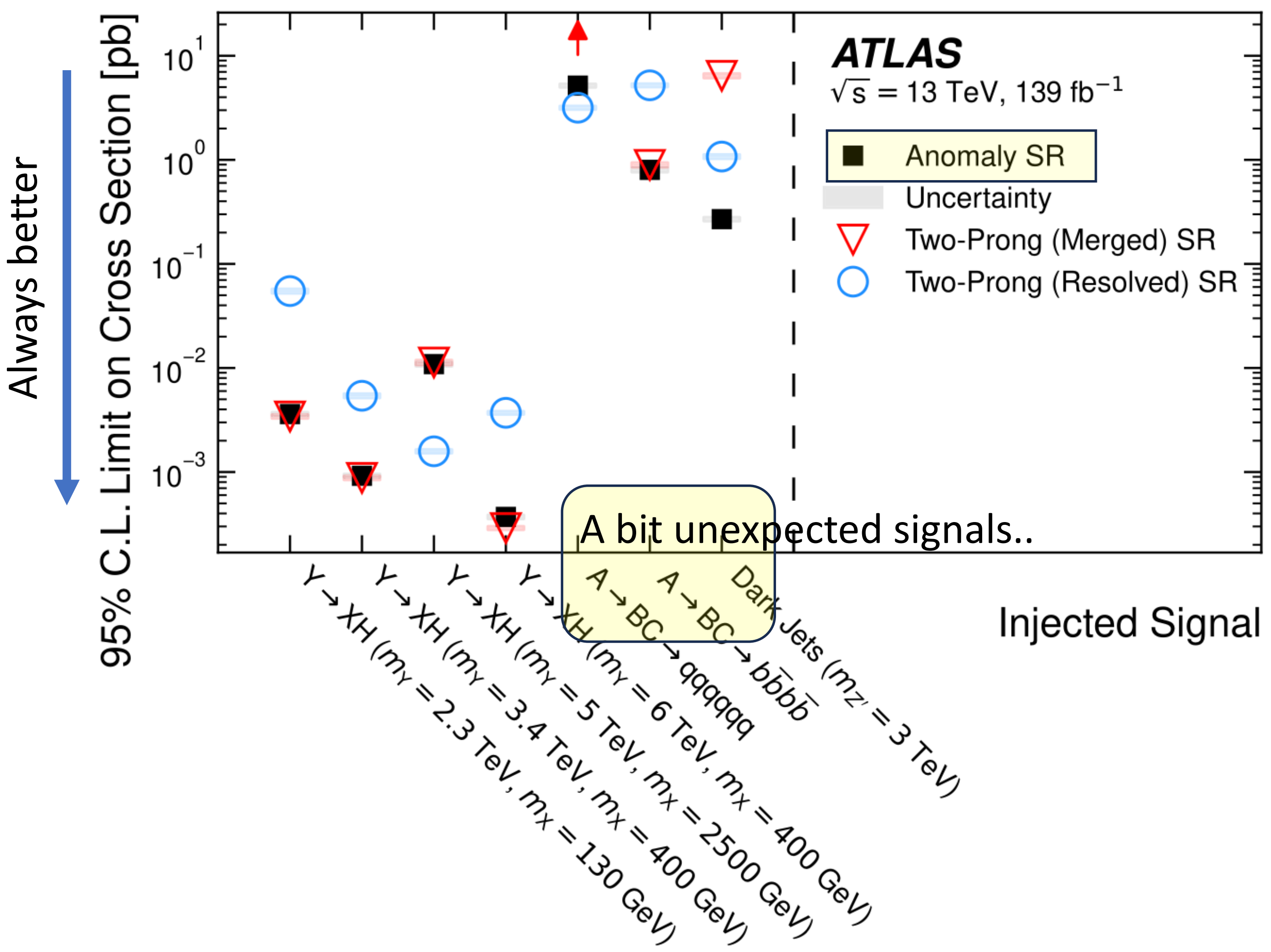
High (~1-6TeV) Y mass resulting in X and H boosted

- Y Reconstructed with two large-R jets



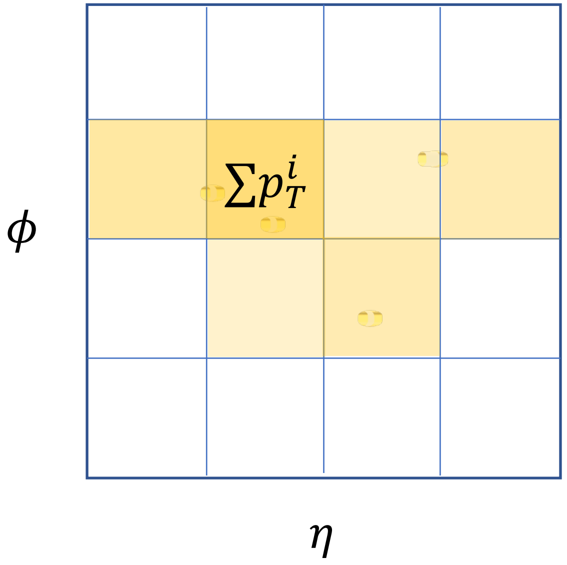
<u>Object Reconstruction:</u>

H Candidate: XbbTagger @60% WP + mass window (75GeV < mH < 145GeV)



What is a CNN?

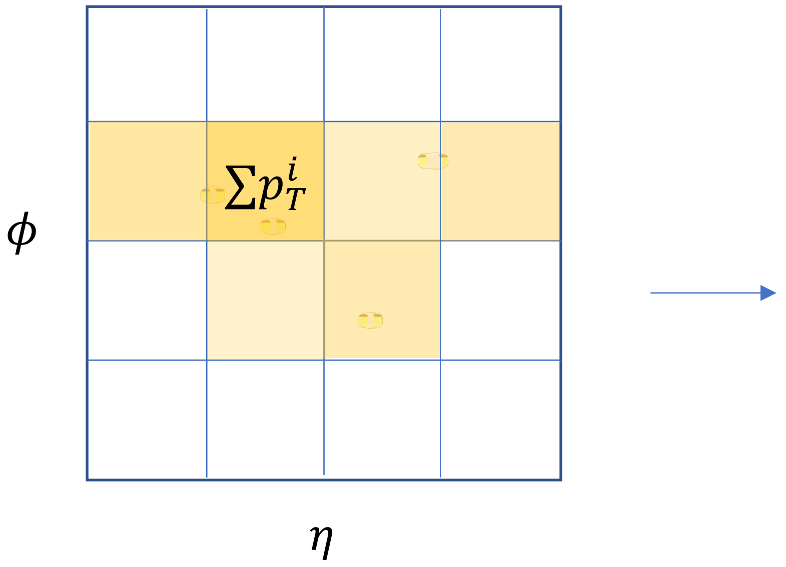
- Let's build intuition using jet tagging as example
- Begin with a jet of constituents (as defined by some conical clustering)



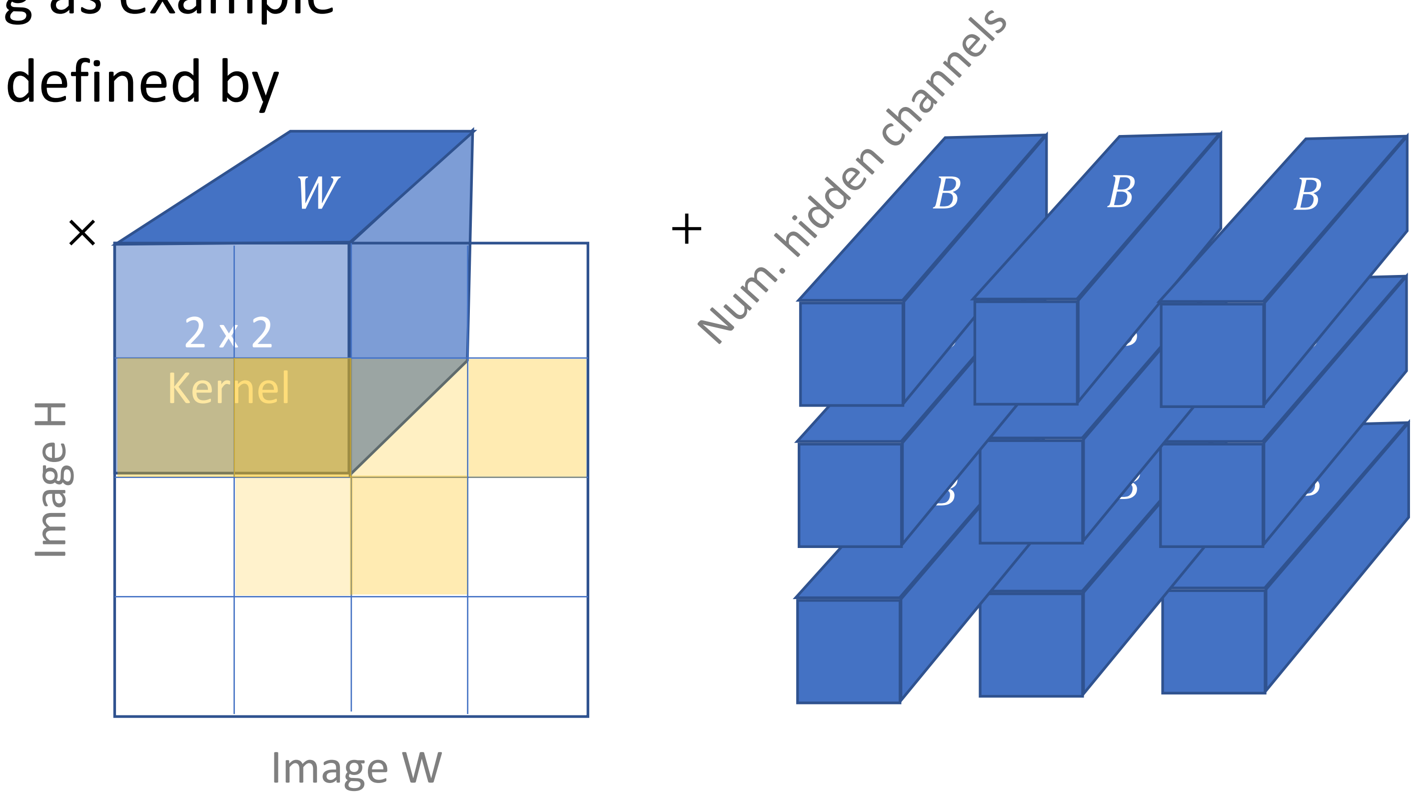
1. Make 2D histogram, weighted by p_T

What is a CNN?

- Let's build intuition using jet tagging as example
- Begin with a jet of constituents (as defined by some conical clustering)



1. Make 2D histogram, weighted by p_T

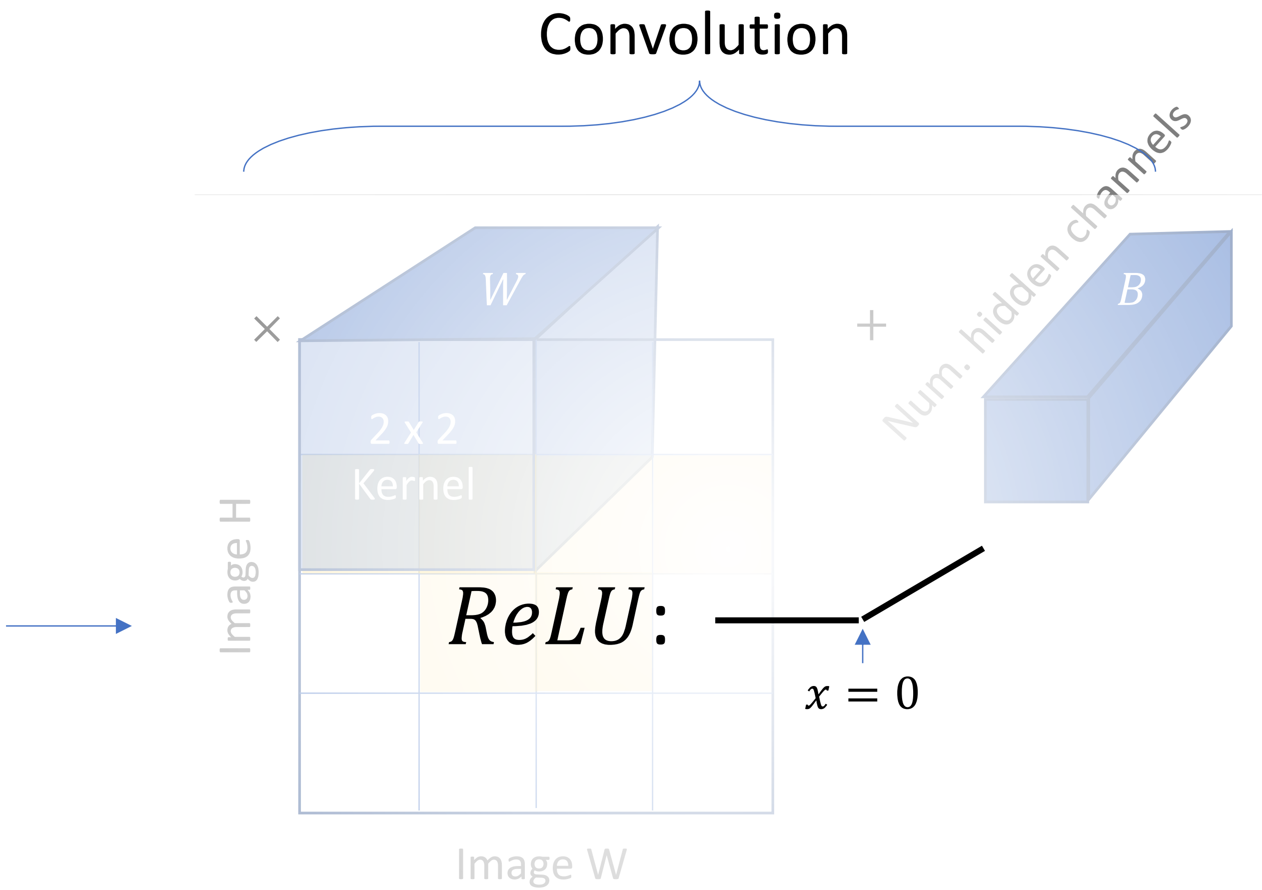


2. Apply a "2D MLP" = a convolution to each kernel block

What is a CNN?

What is a CNN?

- Let's build intuition using jet tagging as example
- Begin with a jet of constituents (as defined by some conical clustering)
- Grid avoids arbitrary ordering by applying a physical interpretation of vectors into $\eta - \phi$ space
- Has translational invariance (i.e. could translate all momenta by $\Delta\eta, \Delta\phi$ without affecting prediction)
- Stack CNN convolutions to a final vector

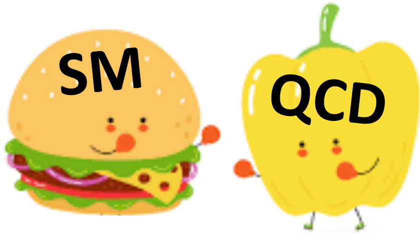


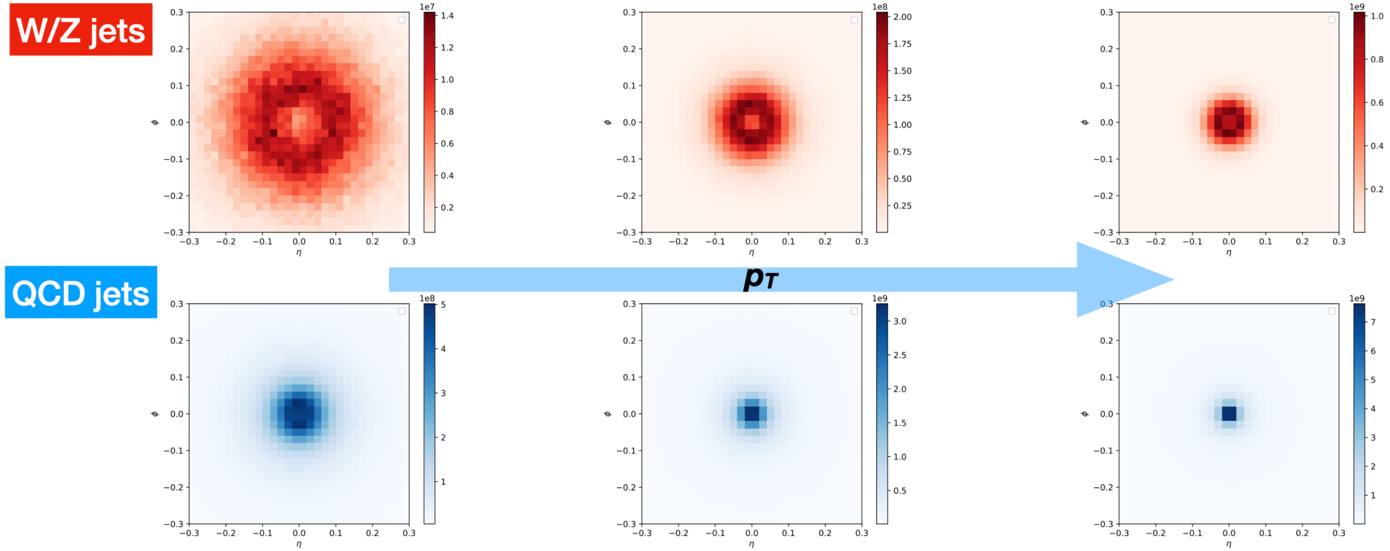
2. Apply a "2D MLP" = a convolution to each kernel block

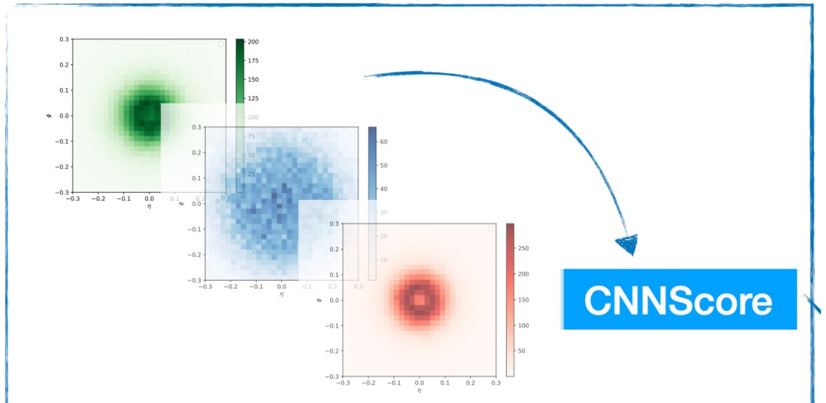
A.Giannini (USTC), FC

Using Jet Images technique

[Costituents based W/Z tagger in VVJJ analysis]

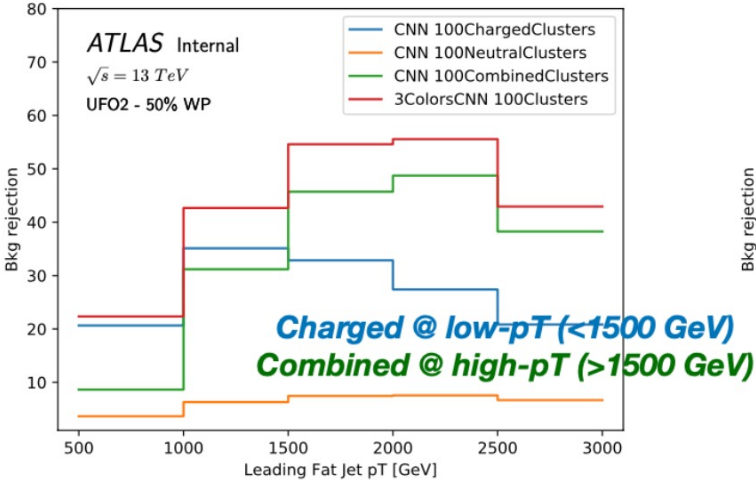




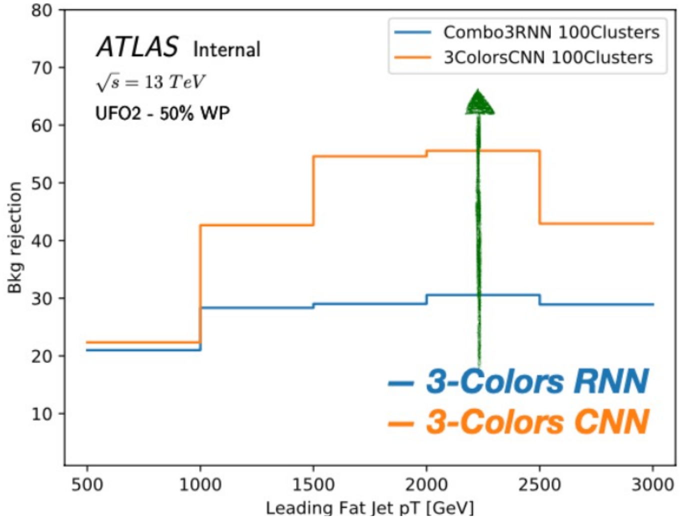


Low-level, jet constituents (taste-aware network)

Each single taste can say something and add information for the net

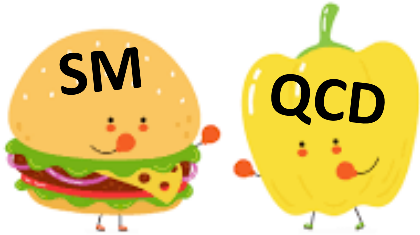


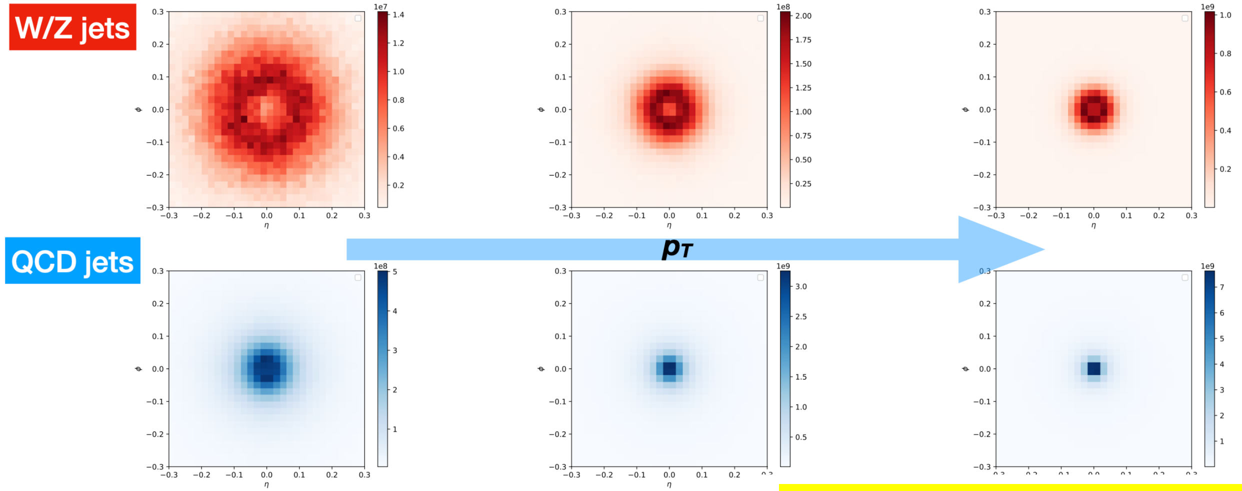
CNN better than RNN, image-based even more natural for the 2-prong sub-structure



Using Jet Images technique

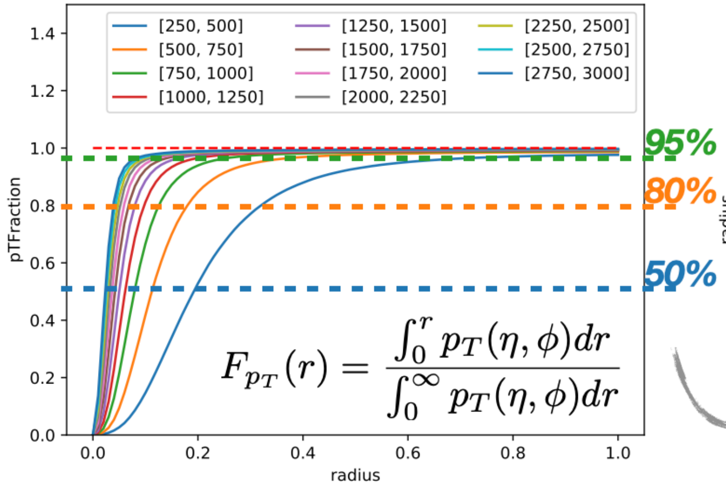
Costituents based W/Z tagger in VVJJ analysis



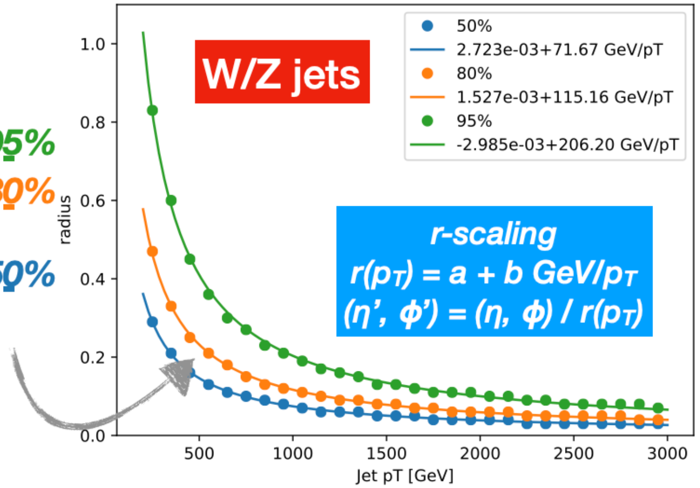


How to build the jet images?

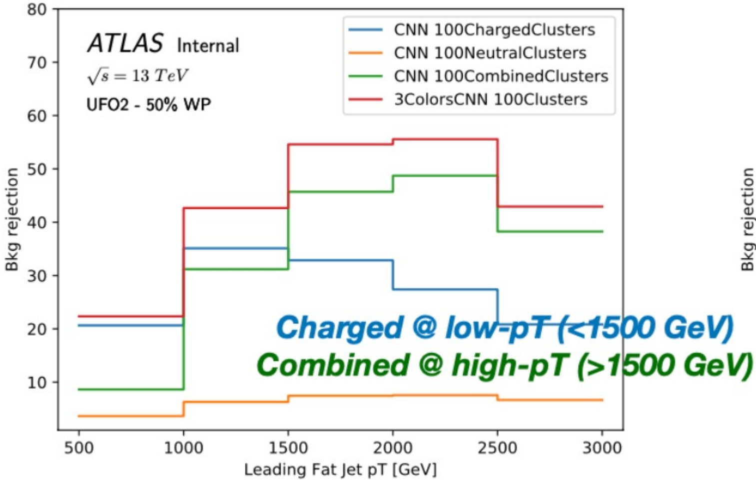
p_T fraction VS radius



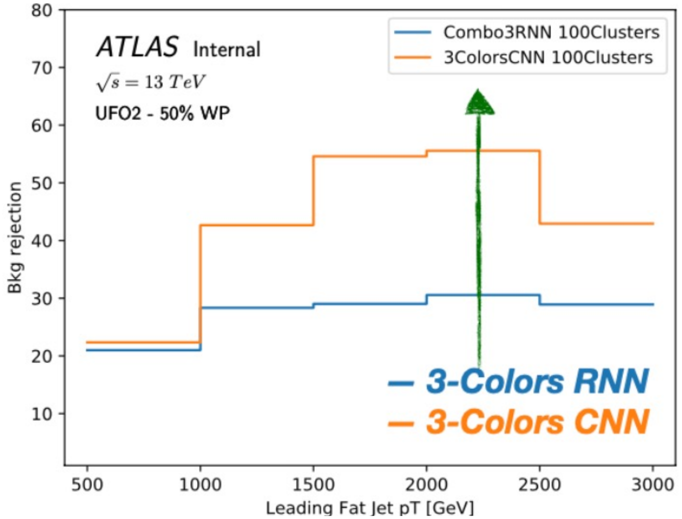
radius VS jet p_T



Each single taste can say something and add information for the net

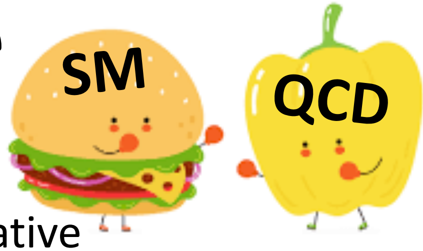


CNN better than RNN, image-based even more natural for the 2-prong sub-structure



A.Giannini (USTC)

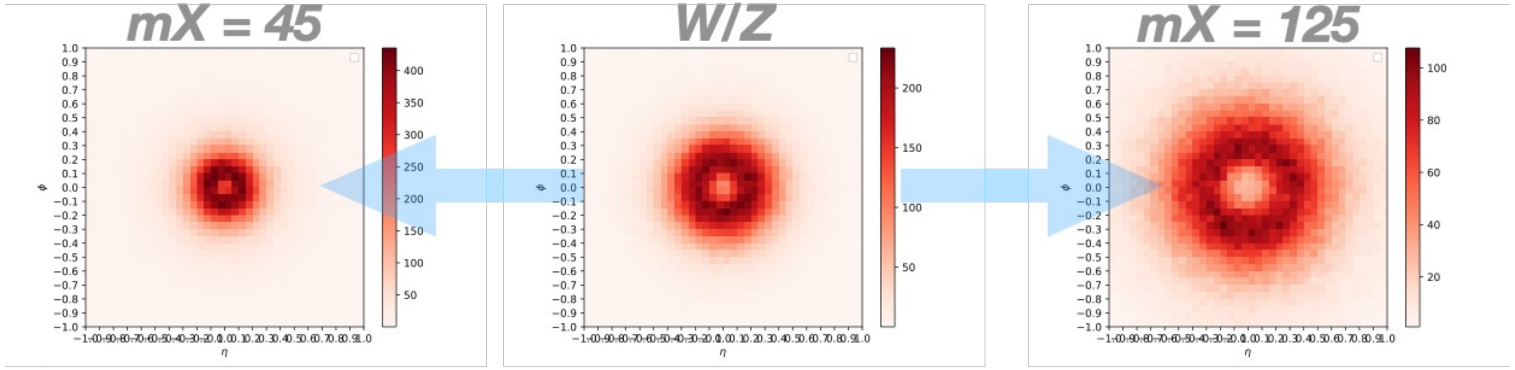
Using Jet Images technique



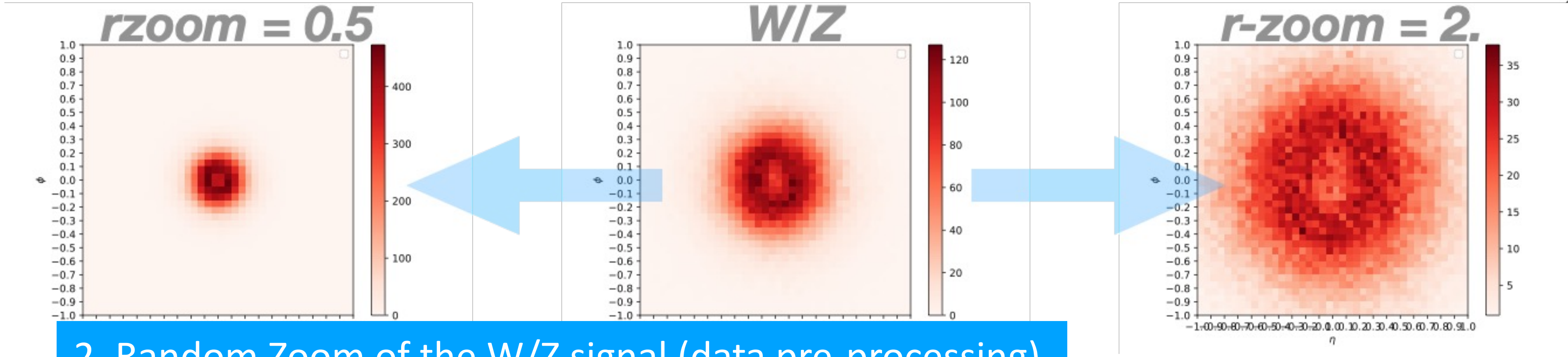
Costituents based W/Z tagger in VVJJ analysis

It is possible to mitigate the mass-peak correlation using alternative boson mass samples

This is nicely explainable in the jet images context

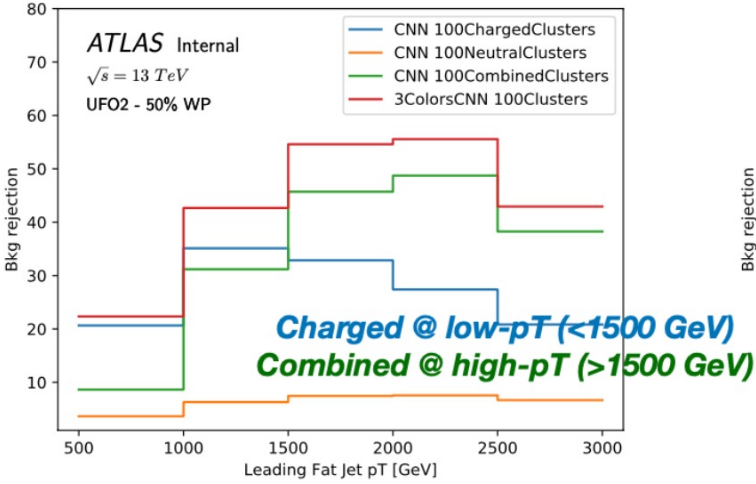


1. Variable mX samples

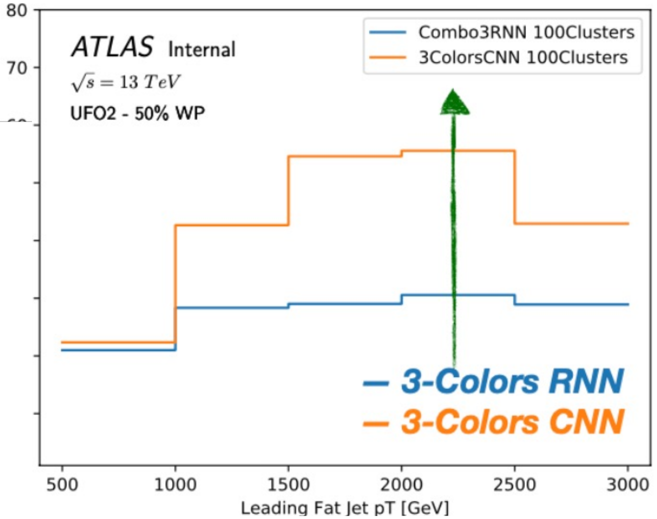


2. Random Zoom of the W/Z signal (data pre-processing)

Each single taste can say something and add information for the net



CNN better than RNN, image-based even more natural for the 2-prong sub-structure

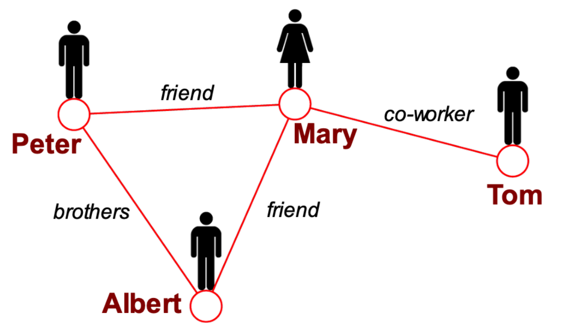


What are graphs?

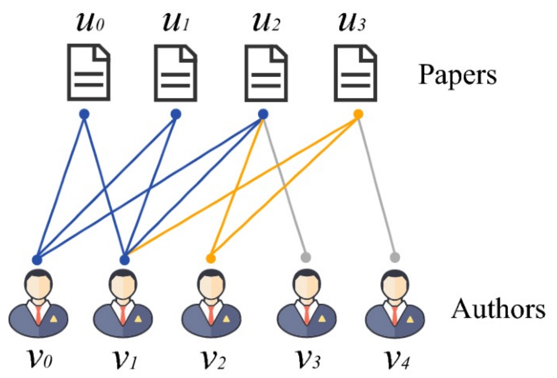
Some data must be arranged in array-like objects in order to be processed by machine learning algorithms, but sometimes it just doesn't feel intuitive (protein chains, social networks between people, ecc.)

Graph representation!

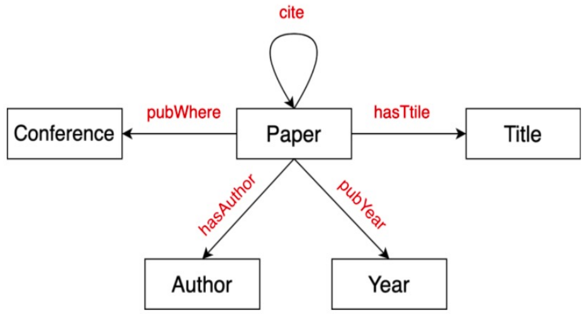
- Structured objects composed of entities used to describe and analyze relations and interactions (edges) between such entities (nodes).
  - Nodes and edges typically contain features specific to each element and each pair.
  - Many types of graphs based on the relations: directed, heterogeneous, bipartite, weighted ecc.



Undirected graph



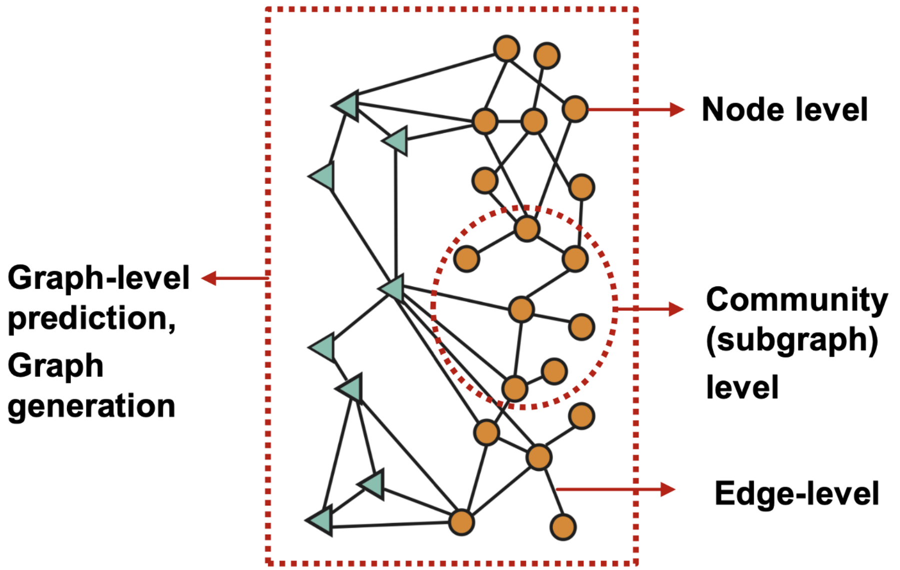
Bipartite graph

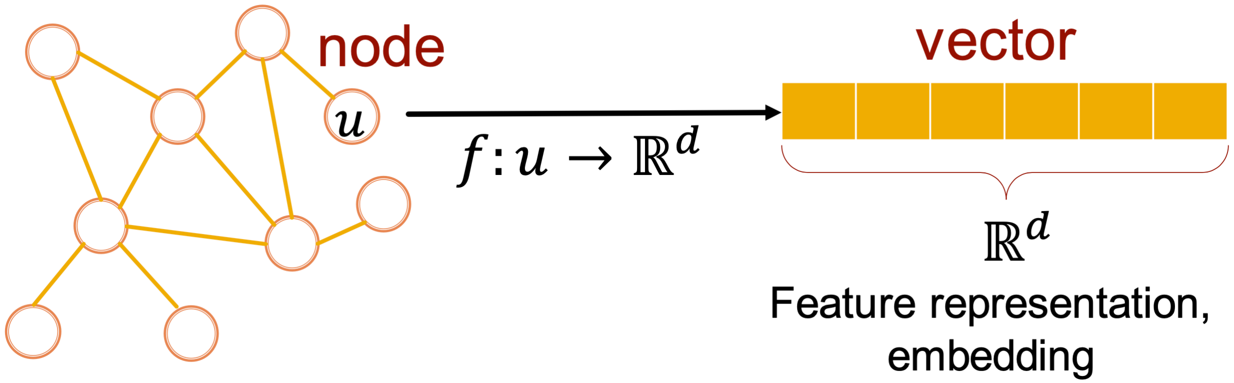


Directed graph

How? Graph Neural Networks!

- Graph Neural Networks (GNNs) are ML architectures built specifically to make predictions on graphs, exploiting their relational nature.
    - The network during training learns the vector representation (embedding $\mathrm{h_v}$) of each node of the input graphs.
    - **Embedding** of a target node depends in some way on what the embeddings of the other nodes are and from its structure.





Several task levels, carried out by processing the final node embeddings in certain ways.

1. **Node prediction** by prediction $\hat{y}(\mathrm{h_v})$, e.g. for identification of malevolous users among social network.
2. **Edge prediction** by predction $\hat{y}(\mathrm{h_v}, \mathrm{h_u})$, e.g. estimate the probability of affiliation between two people of interest of a person to an object.
3. **Graph identification** by prediction $\hat{y}(\mathrm{Pool}(\mathrm{h_v}))$, e.g. classification of protein structure.

Message passing

- The embeddings are updated at each layer by aggregating the information passed between the **target node** and the nodes from its closest neighbourhood → <u>message passing</u>

$$\begin{cases} h_v^i = \text{initial features of node } \nu \\ \boxed{h_v^{(l+1)}} = \sigma(\boxed{W_l}\boxed{\sum_{u\in N(v)}}\boxed{\frac{\boxed{h_u^{(l)}}}{|N(v)|}} + \boxed{B_l} h_v^{(l)}), \boxed{\forall l \in \{0, \dots, L-1\}} \end{cases}$$

Bias term

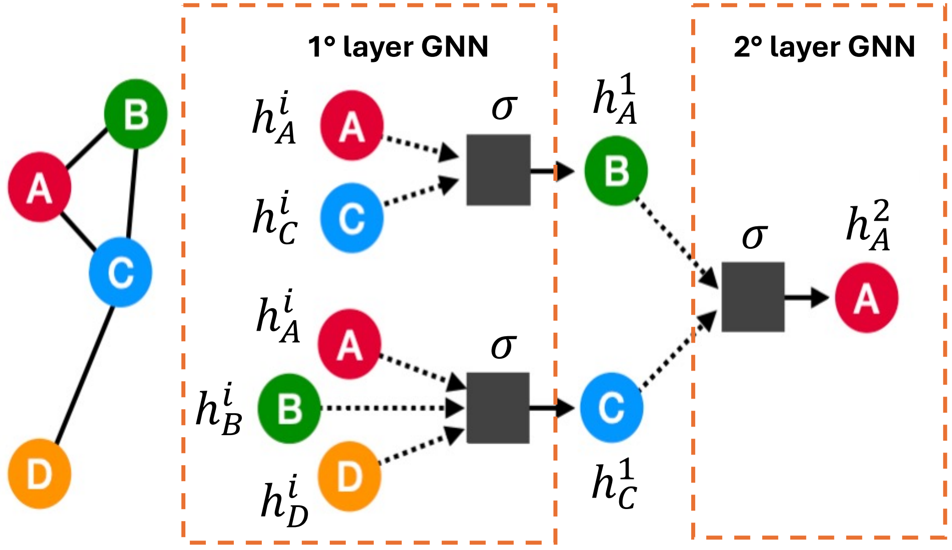
weights matrix

aggregation over N(ν) neighbours u

message passed from u nodes to node ν

set of L layers

embedding representation of node ν at layer (l+1)



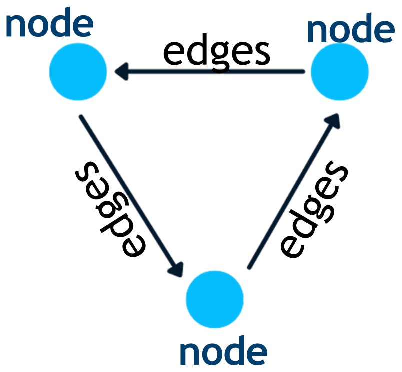
Each layer of GNN extends the neighbour range

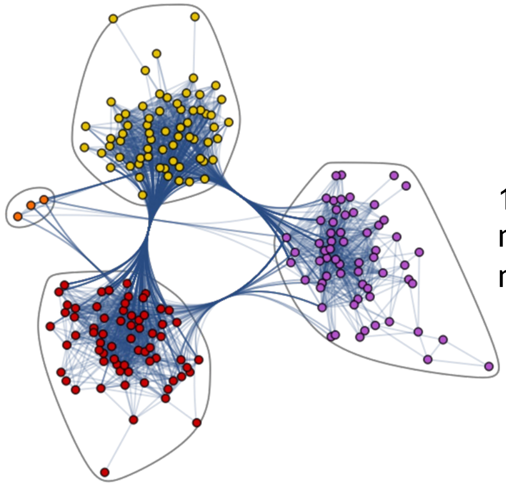
- Graph G embedding is obtained by pooling the nodes embedding at the final layer into one global representation
  - Global sum pooling: $h_G = Sum(\{h_v^L \in \mathbb{R}^d, \forall \nu \in G\})$
  - Global mean pooling: $h_G = \text{Mean}(\{h_v^L \in \mathbb{R}^d, \forall \nu \in G\})$
  - Global max pooling: $h_G = \text{Max}(\{h_v^L \in \mathbb{R}^d, \forall \nu \in G\})$

Graph Neural Network

Graph Neural Network (GNN) is a deep learning model that handles a graph as input data.

A **Graph** is the type of data structure that contains nodes and edges. A node can be a person, place, or thing, and the edges define the relationship between nodes. The edges can be directed and undirected based on directional dependencies.

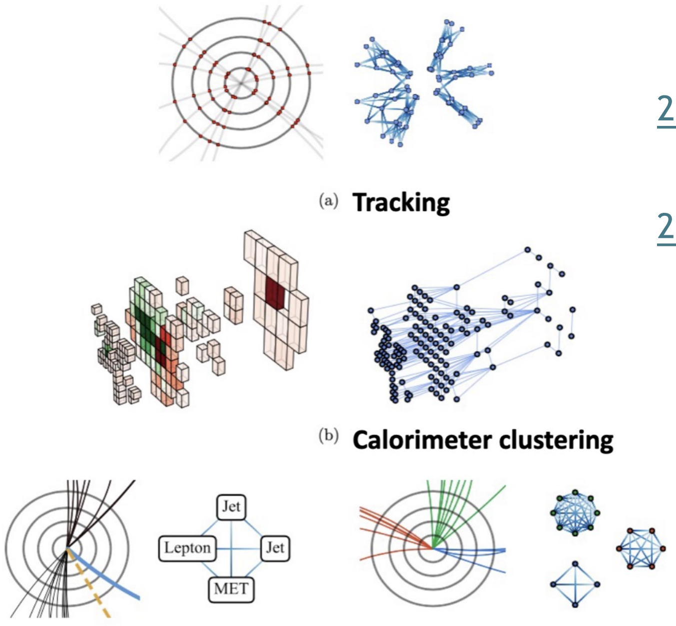




Community Graph Plot by dataset
[Jazz Musicians Network]
198 nodes and 2742 edges: different colors of nodes represent various communities of Jazz musicians and the edges connecting them

Graphs are excellent in dealing with complex problems with relationships and interactions.
They are used in pattern recognition, social networks analysis, recommendation systems, and semantic analysis. Creating graph-based solutions is a whole new field that offers rich insights into complex and interlinked datasets.

Data structure in HEP



(a) Tracking

(b) Calorimeter clustering

(c) Event classification

(d) Jet classification

2007.13681

2203.12852

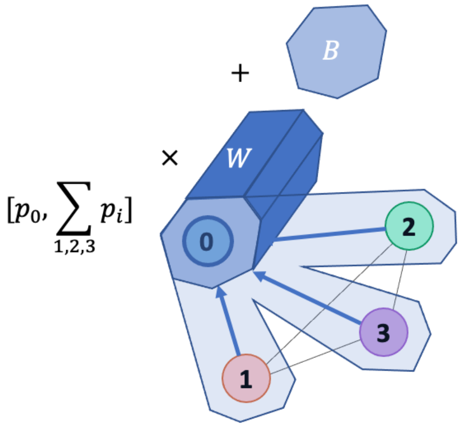
Task Definition: the first step is to decide what function one wants to learn with the GNN. In some applications this is trivial - for example jet, event or particle classification. In those cases a GNN is used to learn some representation of the node or the entire graph/set and a standard classifier is trained on that representation. For tasks such as segmentation or clustering this definition is less trivial.

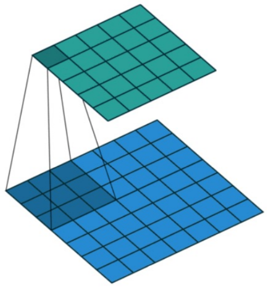
Jure Leskovec et al., CS224W: Machine Learning with Graphs Stanford Course, Fall 2021

CNN vs GNN

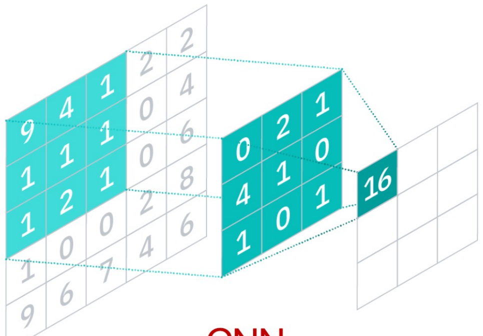
- CNNs are special GNNs with fixed neighbour size and nodes ordering of the input graphs.
  - Heterogeneous objects can be treated as nodes
  - Graphs typically have arbitrary number of connections between nodes, as opposed to images.
  - Possibility to assign any kind of information to nodes and edges (structural and features).

- GNN message passing updating: $h_v^{(l+1)} = \sigma(\mathbf{W}_l \sum_{u \in N(v)} \frac{h_u^{(l)}}{|N(v)|} + B_l h_v^{(l)}), \forall l \in \{0, \dots, L-1\}$
- CNN convolution updating: $h_v^{(l+1)} = \sigma(\sum_{u \in N(v) \cup \{v\}} W_l^u h_u^{(l)}), \forall l \in \{0, \dots, L-1\}$
  - Rewritten as: $h_v^{(l+1)} = \sigma(\sum_{u \in N(v)} \mathbf{W}_l^u h_u^{(l)} + B_l h_v^{(l)}), \forall l \in \{0, \dots, L-1\}$





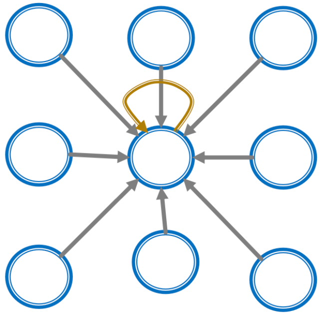
Image



Image

CNN weights

Output

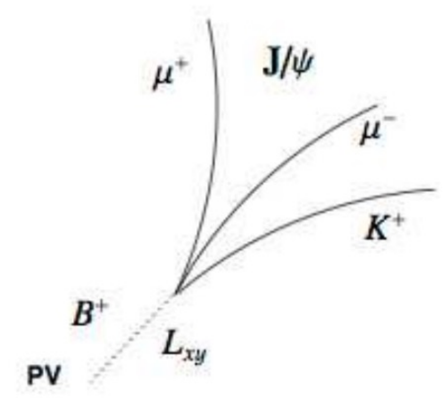


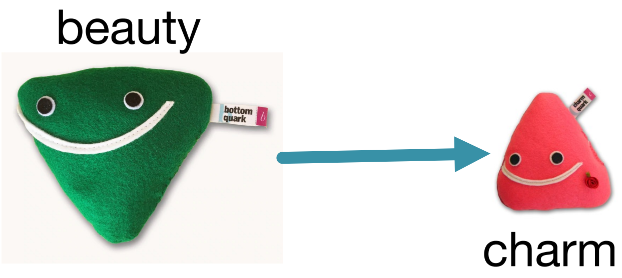
Graph

B and W: weight parameters
$N(v)$: set of neighbours of node v
σ: non-linear activation function
$h_u^{(l)}$: embedding at layer l of node u

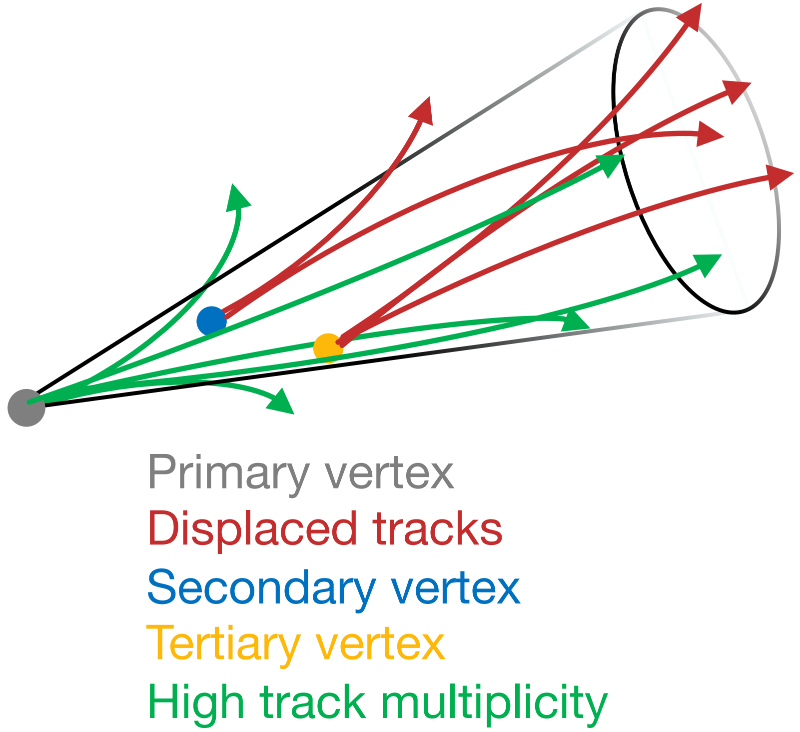
b-jet tagging

The **tracks from b-hadron decay** products tend to have **large impact parameters** which can be distinguished from tracks stemming from the primary vertex





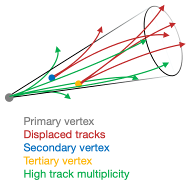
b quark:
Lifetime: ~1.5 ps ($c\tau \approx 450$ μm)
<l> = $\beta\gamma c\tau$ ~3mm in the transverse direction
Mass: ~5 GeV
Decays to charm



b-jet tagging algorithm:

- *Lifetime-based tagging algorithms*
- *Vertex-based algorithms*
- *Combined tagging algorithms* ➔ The vertex-based algorithms ➔ lower mistag rates but their efficiency for actual b jets is limited by the secondary vertex finding efficiency ➔ Both vertex- and lifetime-based approaches are therefore combined to define versatile and powerful tagging algorithms.

b-jet tagging: evolution of the algorithms



JINST 11 P04008 (2016) Eur.Phys.J.C 83 (2023) 7, 681 ATL-PHYS-PUB-2022-027

ATLAS Two-stage approach:

- ✦ **Low-level flavour-taggers:** reconstruct the characteristic features of the heavy-flavour jets via **two complementary approaches**: one that uses the properties of individual charged-particle tracks associated with a hadronic jet (**transverse and longitudinal parameters of impact**), and a second which combines the tracks to explicitly reconstruct **displaced vertices**
- ✦ **High-level flavour-taggers:** high-level algorithms consisting of multivariate classifiers using the results of low-level algorithms

Run 1

- the low-level algorithms first introduced during Run 1
- algorithms based on boosted decision trees or likelihood discriminants

Run 2

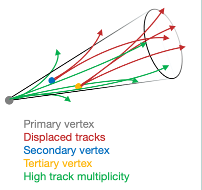
- improvements and retuning of the low-level algorithms
- introduction of new low-level algorithms based on recurrent and deep neural networks
- introduction of new high-level algorithms based on deep neural networks

Run 3

- a new machine learning algorithm based on graph neural networks is introduced.
- It uses information from a variable number of charged particle tracks within a jet, to predict the jet flavour without the need for intermediate low-level algorithms
- the model predicts which physics processes produced the different tracks in the jet, and groups tracks in the jet into vertices.

JINST 11 P04008 (2016)

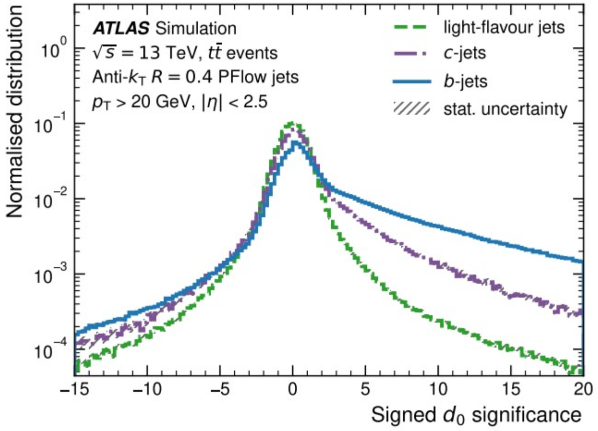
Low Level b-jet tagging algorithms

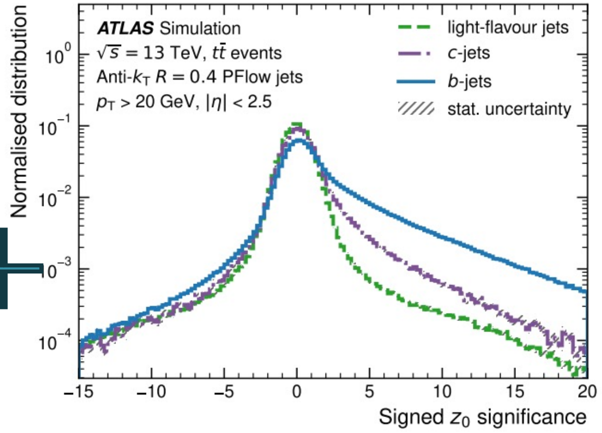


1st approach: Lifetime-based tagging algorithm

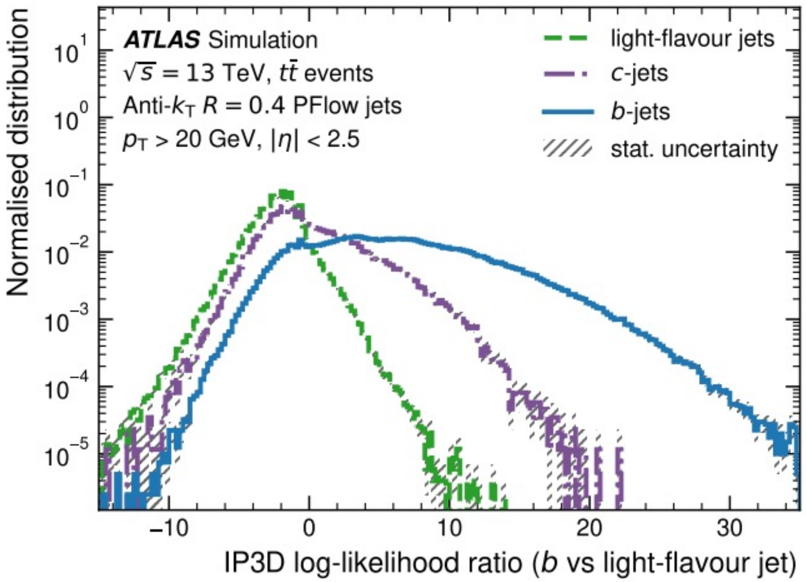
- **IP2D and IP3D:** IP2D makes use of the signed transverse impact parameter significance of tracks to construct a discriminating variable; IP3D uses both the signed transverse and signed longitudinal impact parameter significances in a two-dimensional template to account for their correlation ➔ **Log-likelihood ratio (LLR) discriminants** are then defined as the sum of the logarithms of the per-track probability ratios for each jet-flavour hypothesis
- **RNNIP algorithm:** developed during Run-2 exploits a **recurrent neural network** to learn track impact-parameter correlations in order to further improve the jet flavour discrimination.

transverse and longitudinal impact parameter significances





IP3D Log-likelihood ratio discriminants



JINST 11 P04008 (2016)

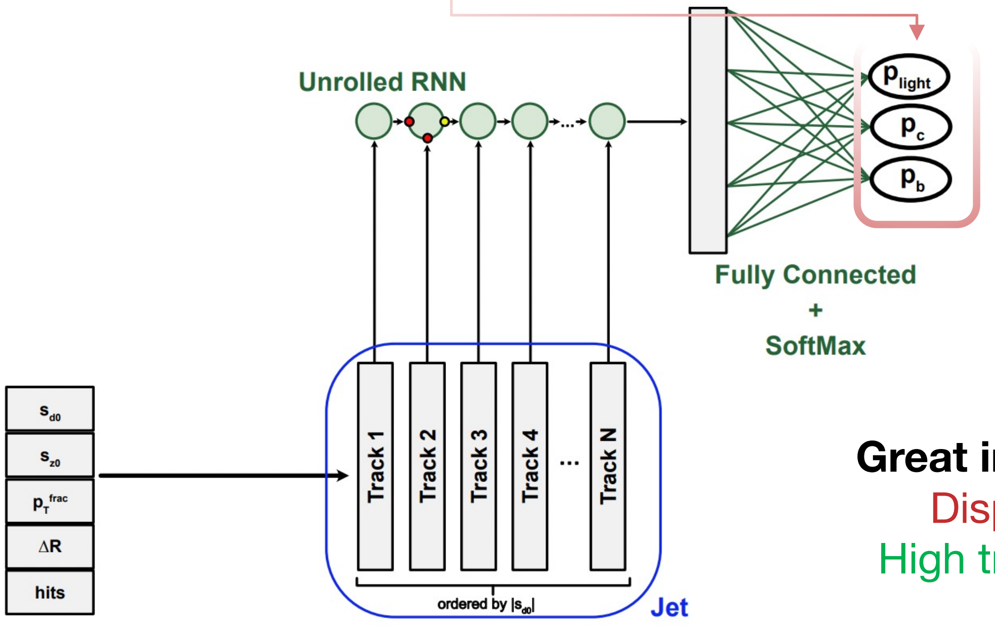
Low Level b-jet tagging algorithms

1st approach: Lifetime-based tagging algorithm

- **IP2D and IP3D:** IP2D makes use of the signed transverse impact parameter significance of tracks to construct a discriminating variable; IP3D uses both the signed transverse and signed longitudinal impact parameter significances in a two-dimensional template to account for their correlation ➔ **Log-likelihood ratio (LLR) discriminants** are then defined as the sum of the logarithms of the per-track probability ratios for each jet-flavour hypothesis
- **RNNIP algorithm:** developed during Run-2 exploits a **recurrent neural network** to learn track impact-parameter correlations in order to further improve the jet flavour discrimination.

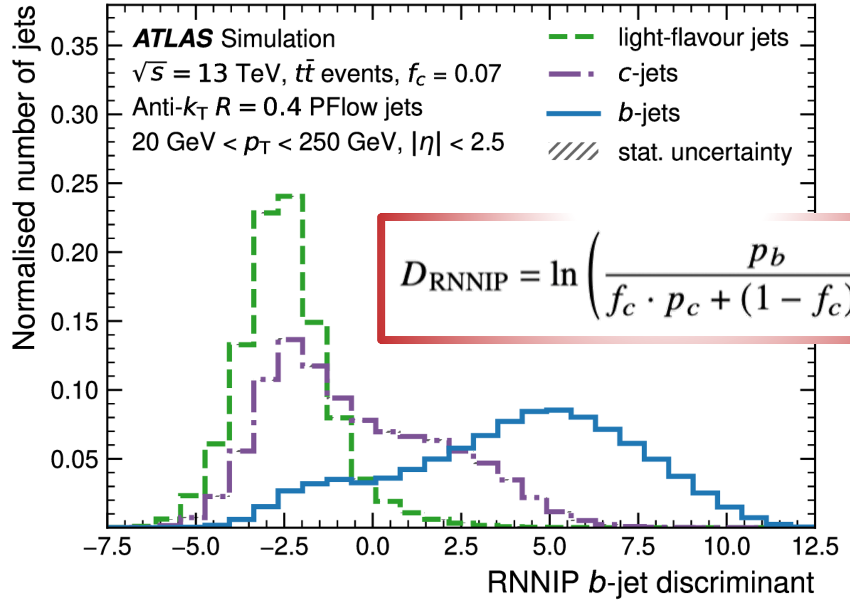
RNNIP scheme

The outputs provided by the network correspond to the b-jet, c-jet, and light-flavour jet probabilities



Great improvement on:
Displaced tracks
High track multiplicity

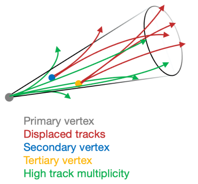
The outputs of the RNN are combined into the b-tagging discriminant function (fc denotes the c-jet fraction)



$$D_{\mathrm{RNNIP}} = \ln\left(\frac{p_b}{f_c \cdot p_c + (1 - f_c) p_{\mathrm{light}}}\right)$$

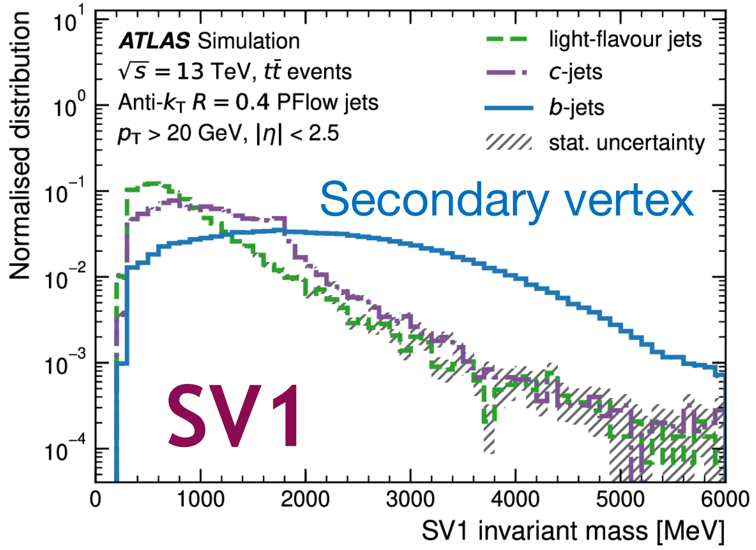
JINST 11 P04008 (2016)

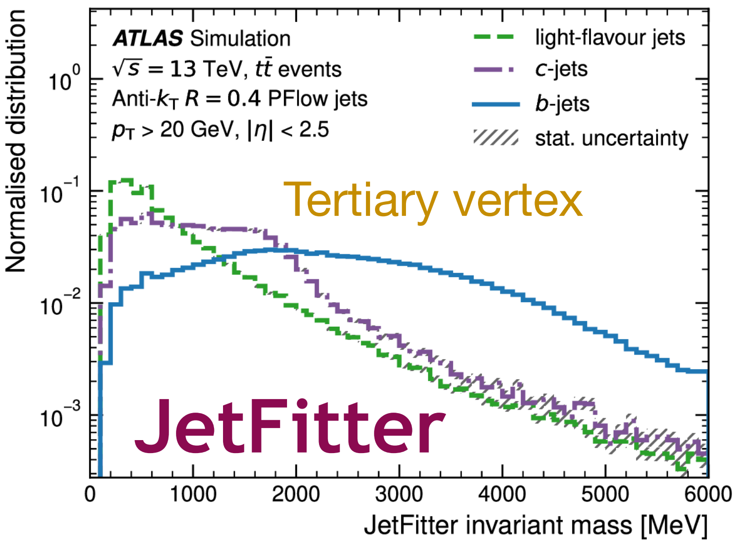
Low Level b-jet tagging algorithms

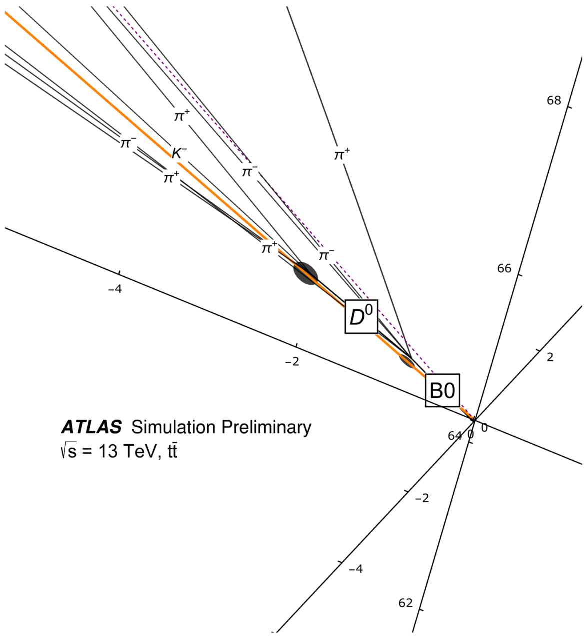


2nd approach: explicit reconstruction of displaced vertices

- **SV1 algorithm:** secondary-vertex-tagging algorithm attempts to reconstruct an inclusive secondary vertex; iterative procedure that starts with two-track vertices built with all tracks associated with the jet. In each iteration, the track-to-vertex matching is evaluated using a χ^2 test
  - developed for Run-1 but improved in Run-2 (increased pile-up rejection and an overall enhancement of the performance at high jet p_T)
- **JetFitter algorithm:** JetFitter algorithm aims to reconstruct the full b- to c-hadron decay chain. A modified Kalman filter is used to find a common line on which the primary, b- and c-hadron decay vertices lie, approximating the b-hadron flight path as well as the vertex positions ➔ possible to resolve the b- or c-hadron decay vertex
- Both were developed for Run-1 but improved for Run-2 (increased pile-up rejection and an overall enhancement of their performance)

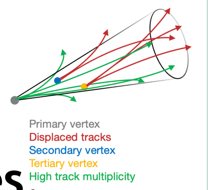






JINST 11 P04008

High Level b-jet tagging algorithms

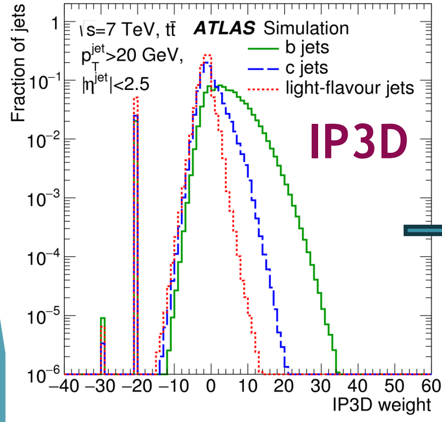


MV1 and MV2: MV1 based on perceptron with two hidden layers consisting of three and two nodes, respectively, and an output layer with a single node which holds the final discriminant variable MV1 based on **boosted decision trees (BDTs).**

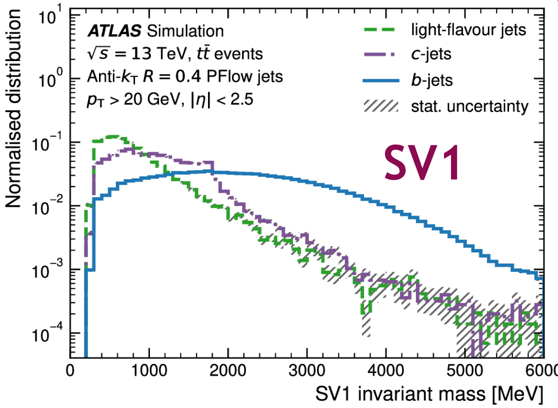
➔ **input variables low-level tagger: IP3D, SV1 and JetFitter discriminants** ➔ IP3D (Lifetime-based tagging algorithm); SV1(Vertex-based algorithm) and JetFitter (Vertex-based algorithm)

➔ ***discrimination between b-jets and light-flavour jets***

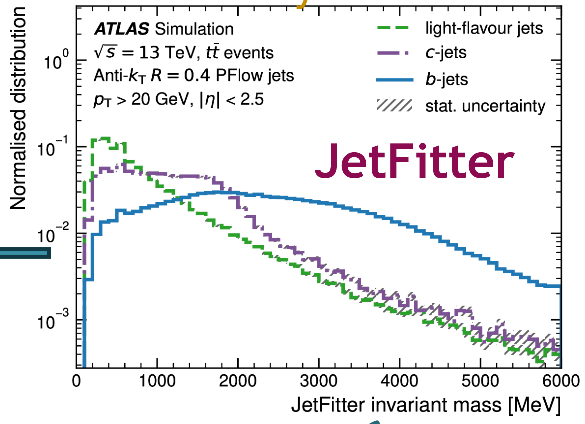
Displaced tracks



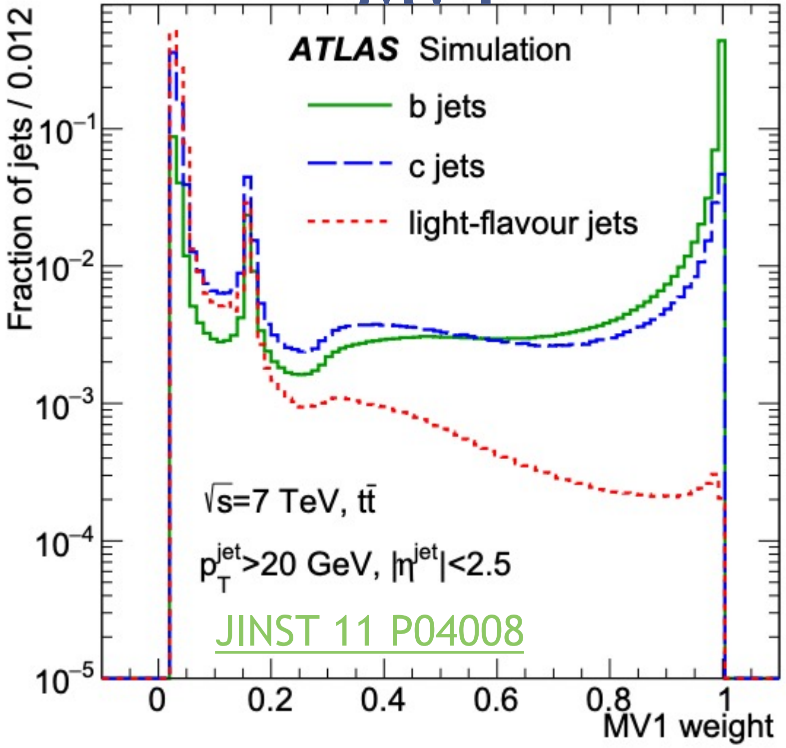
Secondary vertex



Tertiary vertex

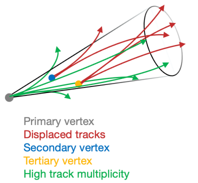


MV1



Mostly used in Run-1

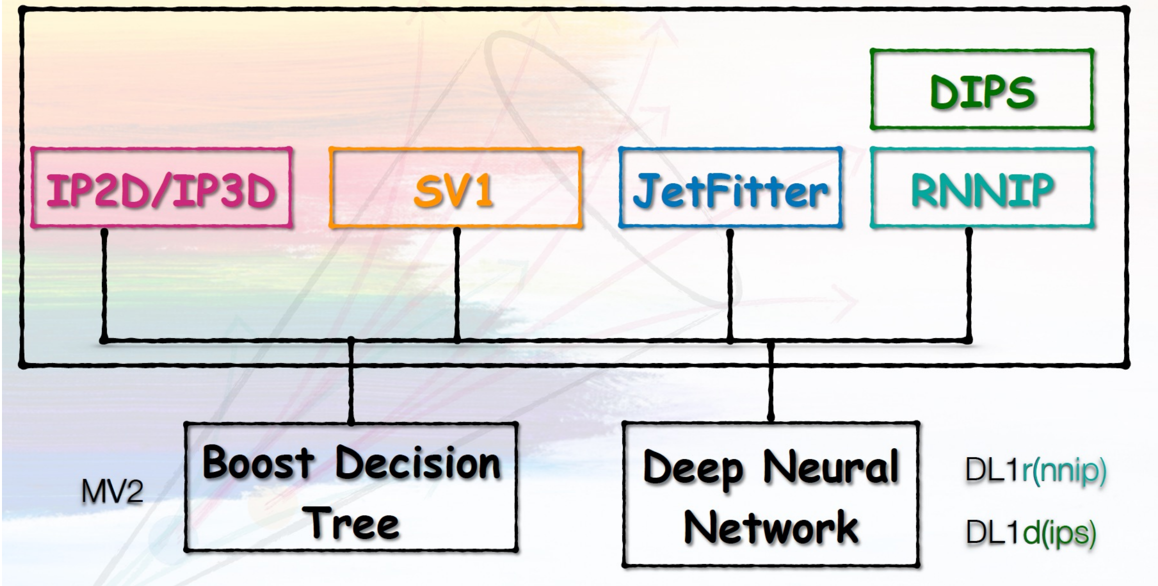
High Level b-jet tagging algorithms



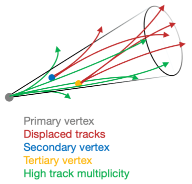
DL1 algorithm series: output quantities of the low-level algorithms are combined using **deep-learning classifiers**, based on **fully connected multi-layer feed-forward neural networks (NN)**

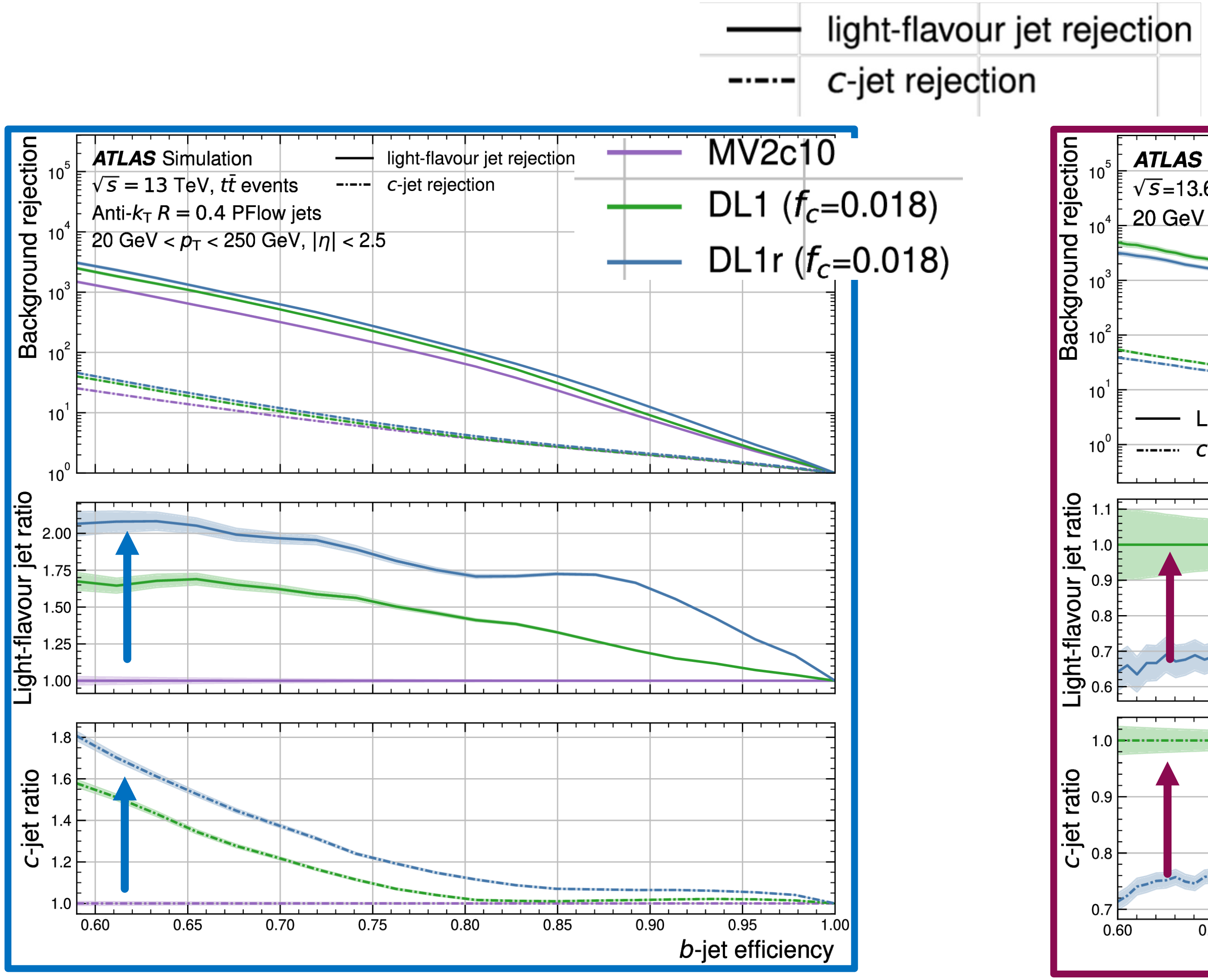
- Trained with a hybrid training sample: 70% of the jets are from tt events 30% of the jets are from $Z' \rightarrow qq$ events
- TensorFlow with the Keras front-end and the Adam optimiser
- **DL1 algorithm exploits as input the IP2D, IP3D, SV1 and JetFitter algorithm outputs**
- **DL1r algorithm also includes the jet RNNIP output probabilities** ➔ multidimensional output corresponding to the probabilities for a jet to be a b-jet, a c-jet or a light-flavour jet
- **DL1d algorithm: Flavour Tagging based on Deep Sets (DIPS)**, which models the jet as a set of tracks, in order to identify the experimental signatures of jets containing heavy flavour hadrons using the impact parameters and kinematics of the tracks. This approach is an evolution with respect to the RNN

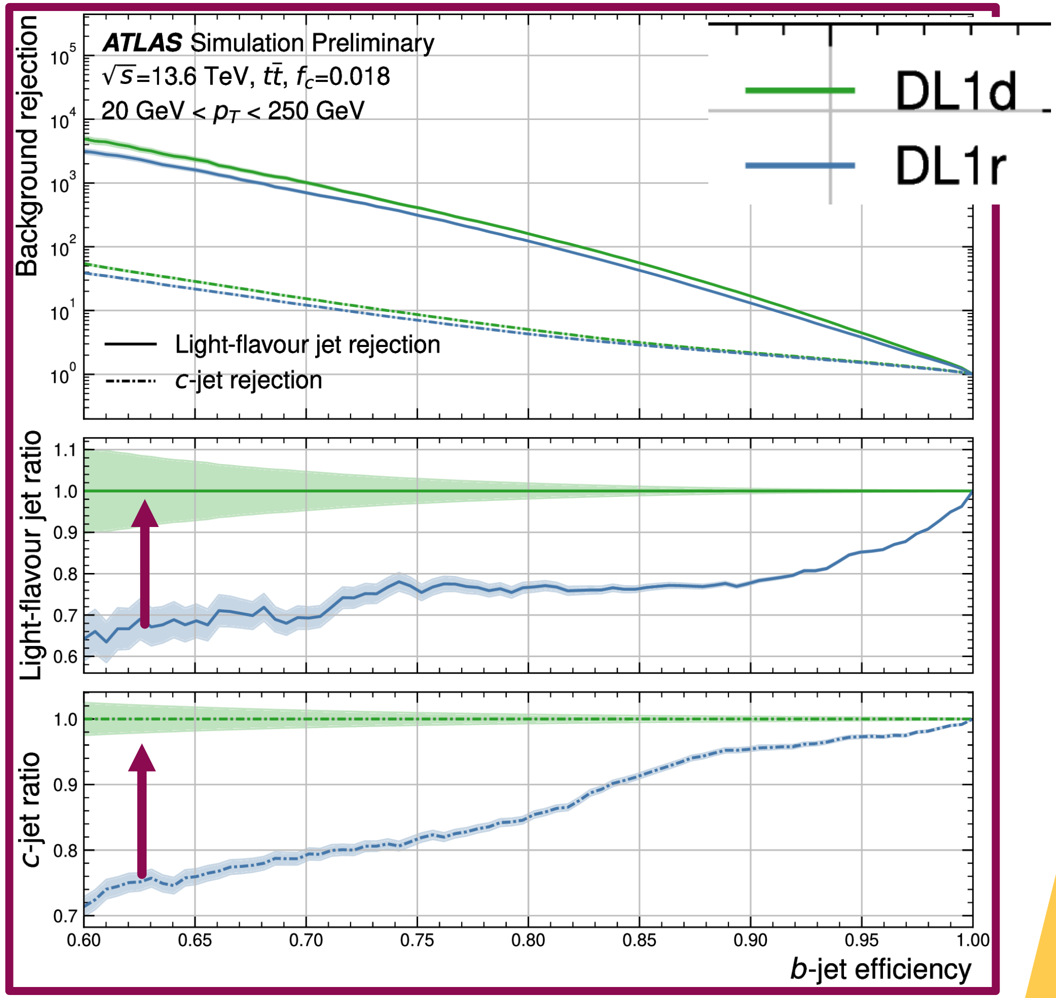
Run-2 & early Run-3



High Level b-jet tagging algorithms



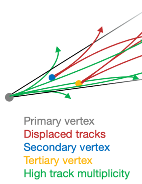




Early Run 2 → Run 2 → Early Run 3

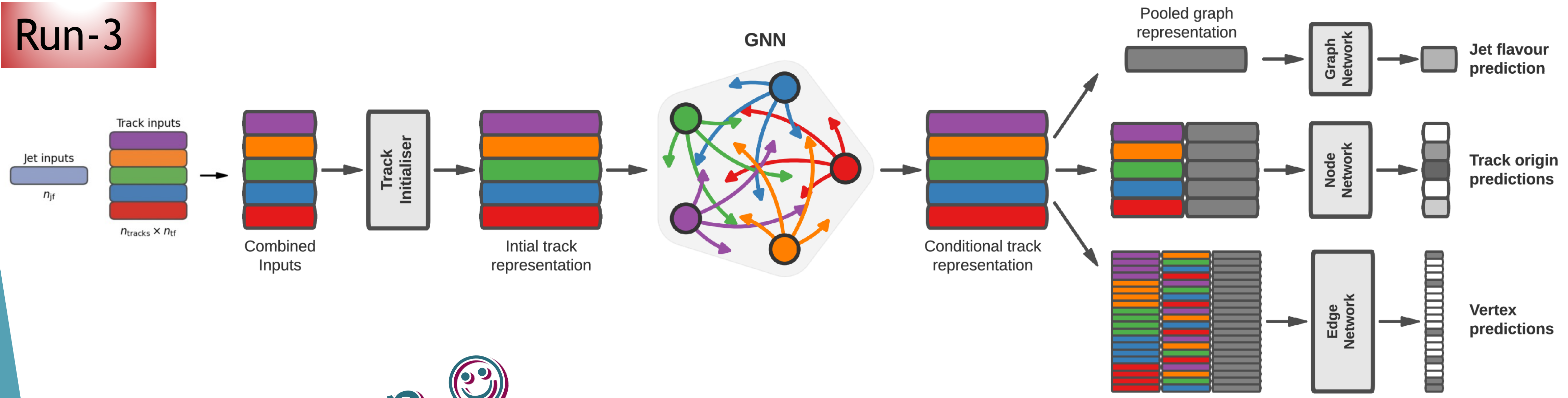
ATL-PHYS-PUB-2022-027

High Level b-jet tagging algorithms



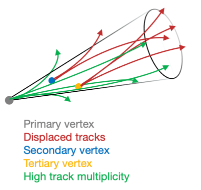
GN1 algorithm: uses Graph Neural Networks (GNNs) is a new machine learning algorithm based on graph neural networks. GN1 uses information from a variable number of charged particle tracks within a jet, to predict the jet flavour *without the need for intermediate low-level algorithms*.
Alongside the jet flavour prediction, the model predicts which physics processes produced the different tracks in the jet, and groups tracks in the jet into vertices ➔ useful additional information on the contents of the jet and improve performance
GN1lep algorithm: the GN1 Lep variant includes an additional track-level input, lepton ID, which indicates if the track was used in the reconstruction of an electron, a muon or neither.

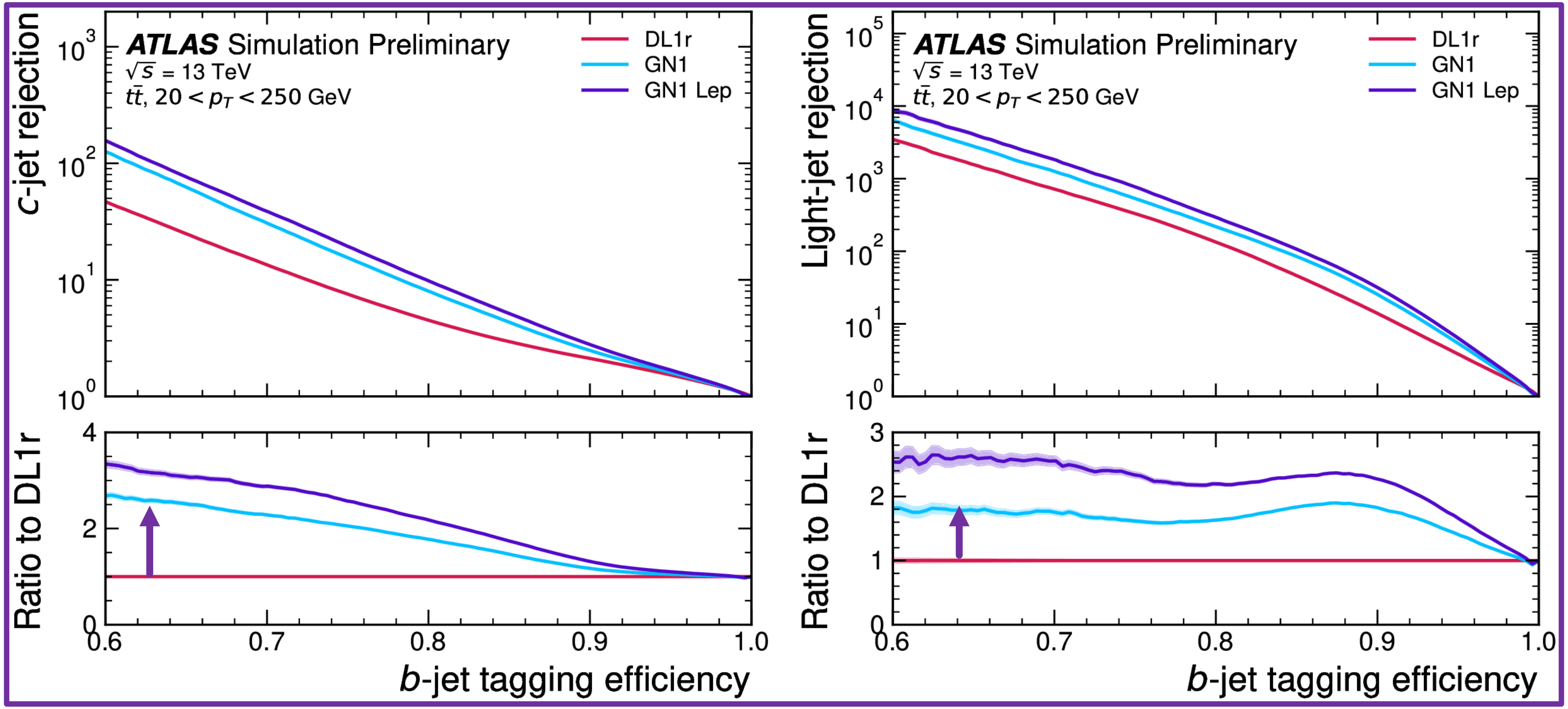
Run-3



Last born ☺

High Level b-jet tagging algorithms

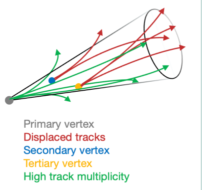




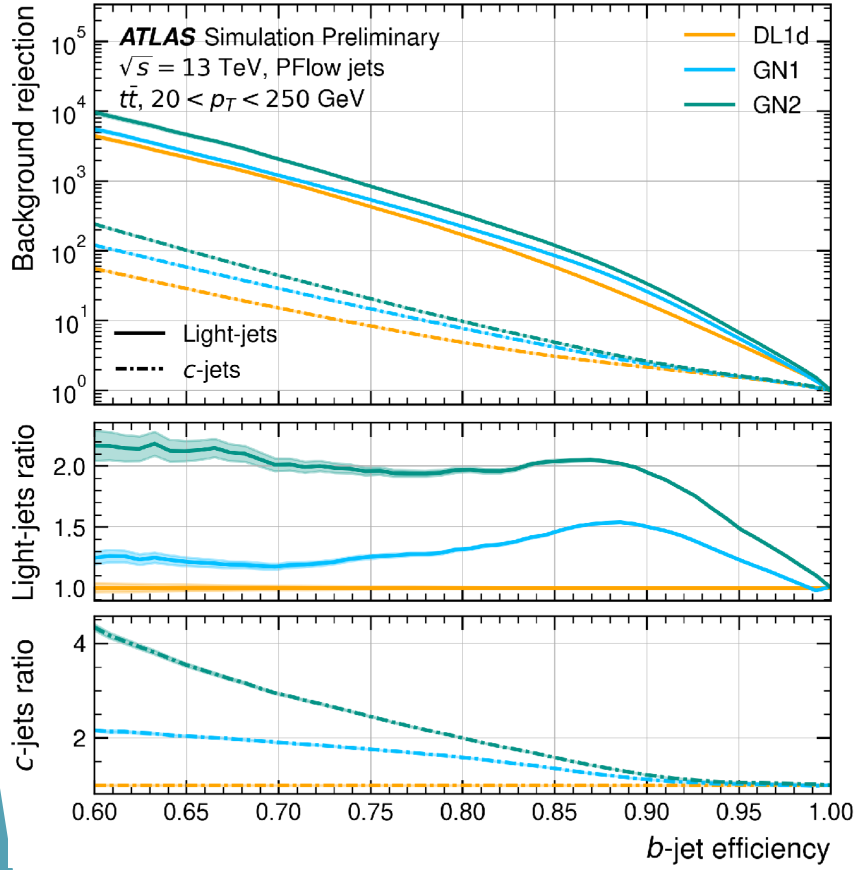
Early Run 2 → Run 2 → Early Run 3 → Run 3

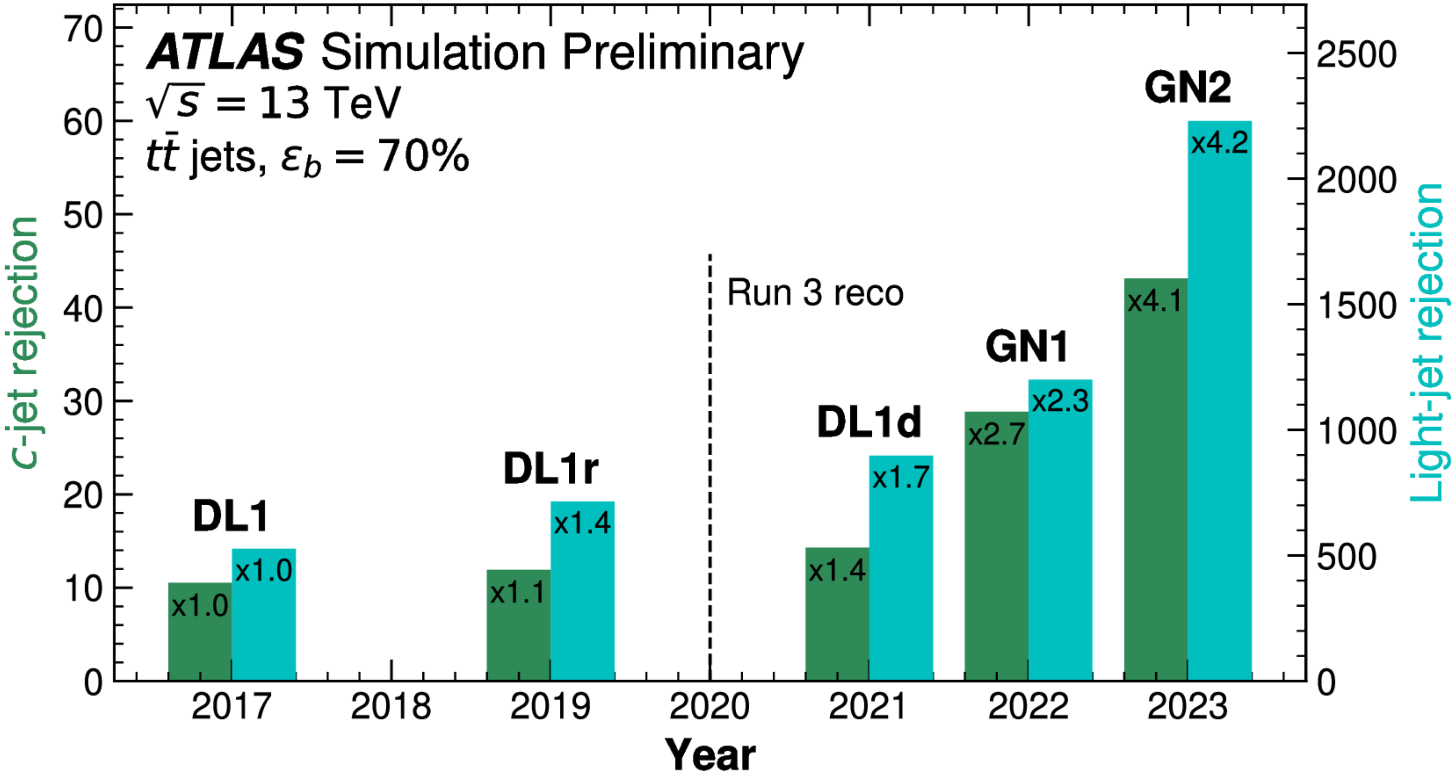
- ✗ For jets coming from tt decays with 20 < $p\mathrm{T}$ < 250 GeV: b-jet efficiency of 70% & the light-jet rejection is improved by a factor of ~1.8 and c-jet rejection of ~2.1 for jets coming from tt decays with 20 < $p\mathrm{T}$ < 250 GeV.
- ✗ For jets coming from Z' decays with transverse momentum 250 < $p\mathrm{T}$ < 5000 GeV: b-jet efficiency of 30% & the light-jet rejection improves by a factor ~6 and and c-jet rejection of ~2.8 for a comparative 30% b-jet efficiency.

High Level b-jet tagging algorithms



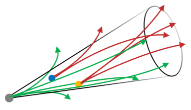
GN2 algorithm: updated version of GN1 with the **transformer architecture** adopted achieves the best performance!





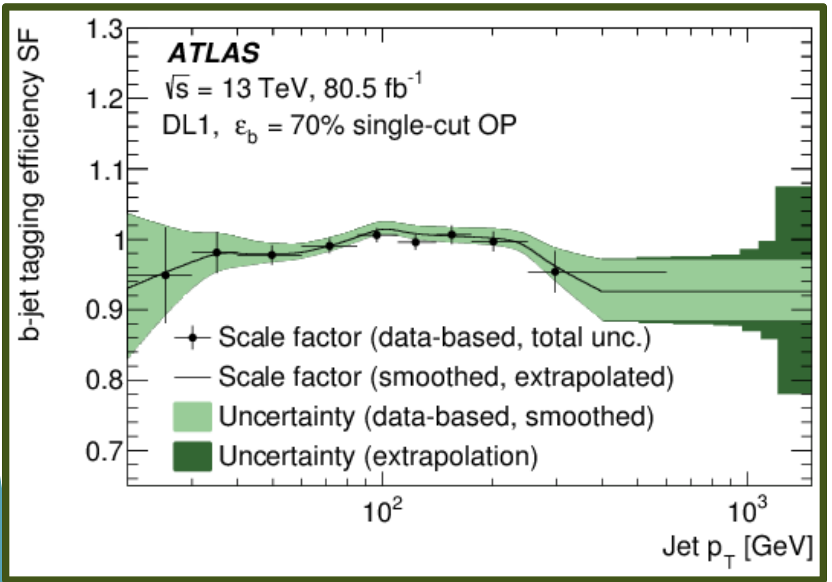
Last born and immediately improved ☺

b-jet tagging algorithms and data

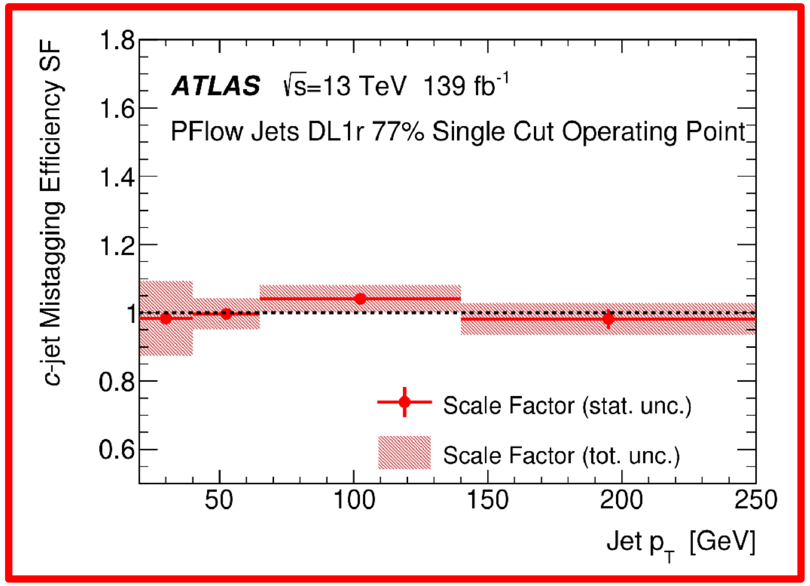


Taggers are developed using simulation
Do the work well in data?

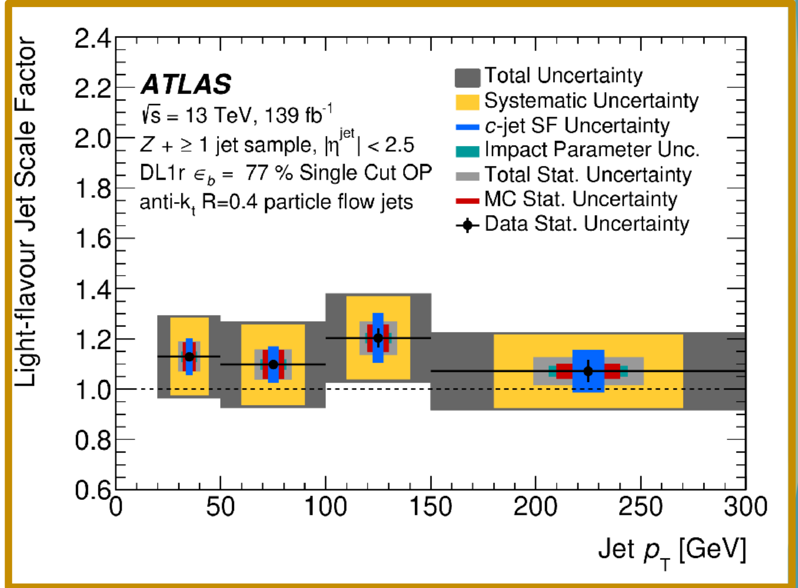
Eur. Phys. J. C 79 (2019) 970



Eur. Phys. J. C 82 (2022) 95



CERN-EP-2022-211



Efficiencies SF in data measured for b-,c- and light-jet

Backup

Vector Like Quarks (VLQ)

What are Vector-Like Quarks (VLQs)?

Vector-like fermions,ψ, have left- and right-handed chiralities that transform in the same way under the SM gauge group:

$$SU(3)_C \times SU(2)_L \times U(1)_Y$$

$$\mathcal{L}_W = \frac{g}{\sqrt{2}} \left(J^{\mu+} W^+_\mu + J^{\mu-} W^-_\mu \right)$$ Charged current Lagrangian

- SM chiral quarks: ONLY left-handed charged currents

$$J^{\mu+} = J_L^{\mu+} + J_R^{\mu+} \quad \text{with} \quad \begin{cases} J_L^{\mu+} = \bar{u}_L \gamma^\mu d_L = \bar{u} \gamma^\mu (1-\gamma^5) d = V - A \\ J_R^{\mu+} = 0 \end{cases}$$

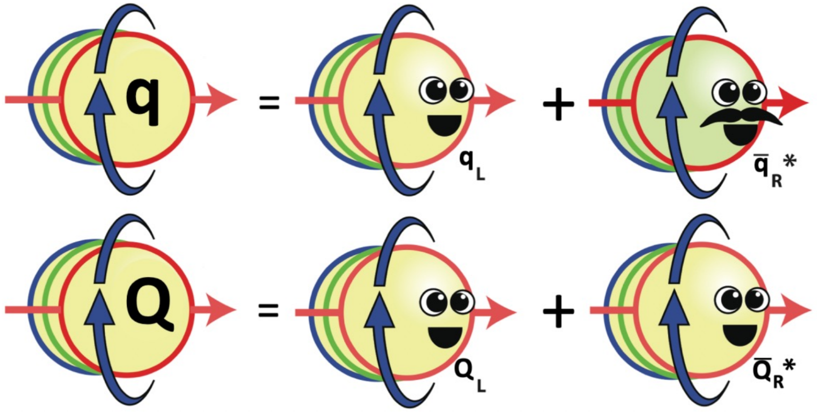
- vector-like quarks: BOTH left-handed and right-handed charged currents

$$J^{\mu+} = J_L^{\mu+} + J_R^{\mu+} = \bar{u}_L \gamma^\mu d_L + \bar{u}_R \gamma^\mu d_R = \bar{u} \gamma^\mu d = V$$

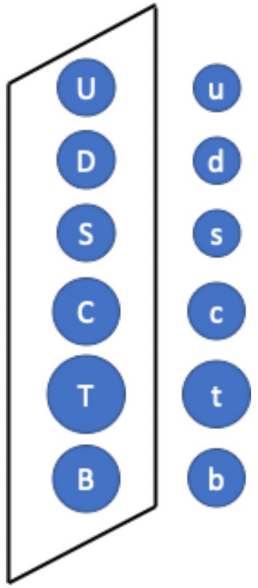
Additionally, gauge-invariant mass terms,$-M\bar{\psi}\psi$, allowed without the need of Higgs.

Vector-like quarks in many models of New Physics

- ✗ Spin 1/2 particles with color charge
- ✗ Left and right chiralities behave the same ➔ Vector-like interaction with weak force
- ✗ Mass not from Higgs boson
- ✗ Decay to SM boson and quark
- ✗ explains the lack of CP-violation in the strong interaction
- ✗ Vector-like quarks have the same mass heirarchy as SM quarks

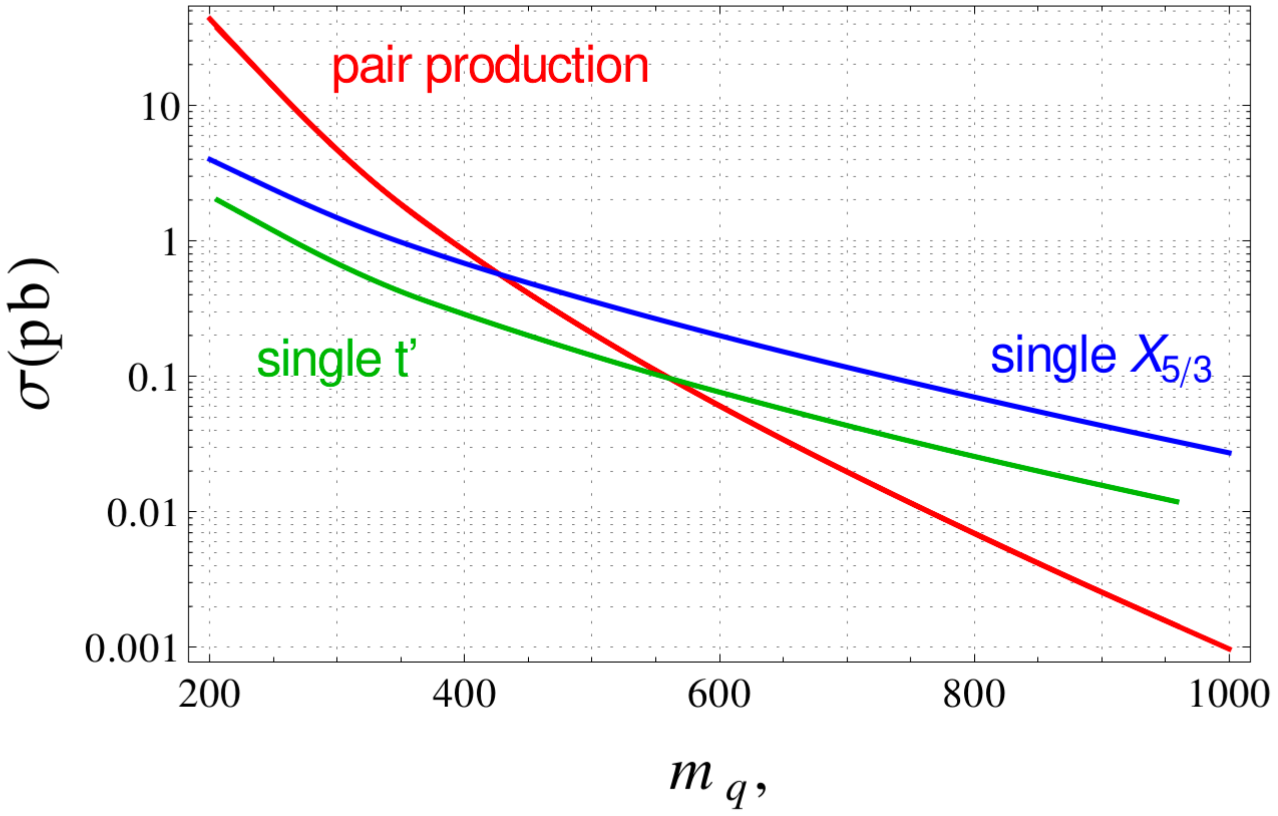


- ***Warped or universal extra-dimensions***: *KK excitations of bulk fields*
- ***Composite Higgs models***: *VLQ appear as excited resonances of the bounded states which form SM particles*
- ***Little Higgs models***: *partners of SM fermions in larger group representations which ensure the cancellation of divergent loops*
- ***Gauged flavour group*** *with low scale gauge flavour bosons required to cancel anomalies in the gauged flavour symmetry*
- ***Non-minimal SUSY extensions***: *VLQs increase corrections to Higgs mass without affecting EWPT*
- Predicted in other models such as the **Left Right Mirror Model Model**



Composite-Higgs models and vector-like quarks

- *The Higgs boson is a composite pseudo-Nambu-Goldstone boson (pNGB) from spontaneous breaking of a global symmetry in a new strongly coupled sector ➔ This protects the Higgs mass.*
- *Models with partial compositeness predict new vector-like fermions.*
- *Simplest extensions with VLQ ($T^{2/3}$, $B^{-1/3}$ and $X^{5/3}$) singlets, doublets, and triplets.*
- *VLQs assumed to decay via charged and neutral currents to 3rd generation quarks.*

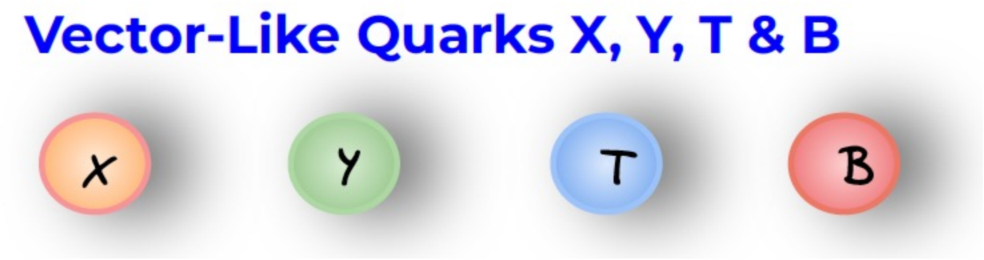


QCD pair-production:
Mass-independent, dominant at low mass

Single-production:
Scales with coupling, model dependent, significant at high mass.

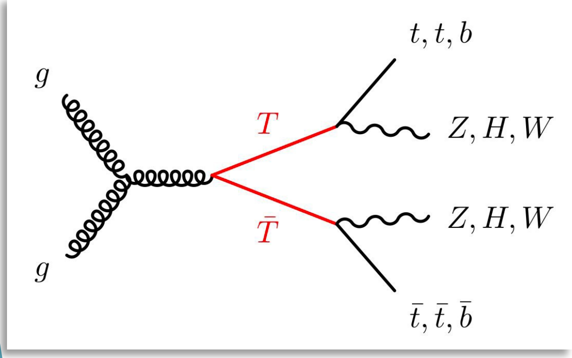
Vector-Like Quarks (VLQs)

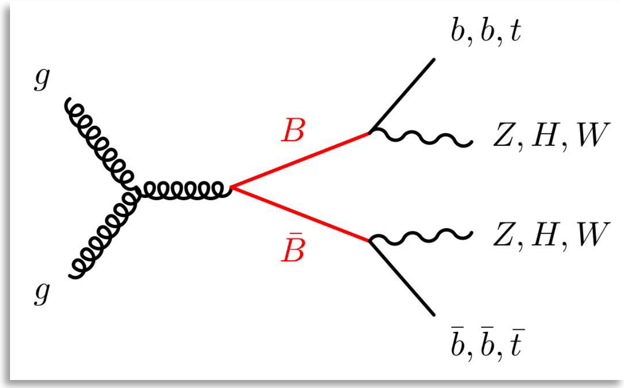
- *Spin 1/2.*
- *Left and right-handed chiralities transform in the same way under the SM gauge group.*
- *Decay to qZ, qW or qH where q = { t, b }*



QCD pair-production

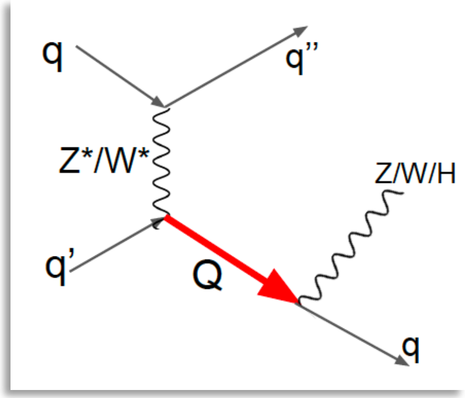
Model Independent production cross section

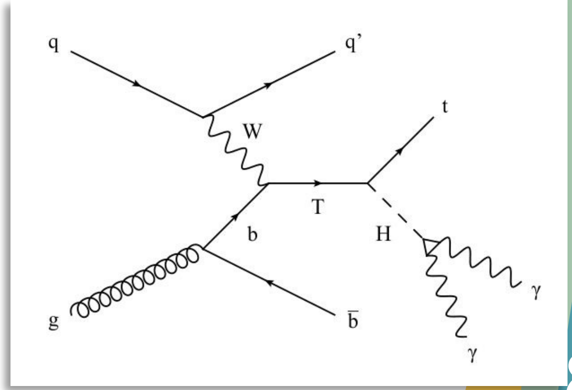




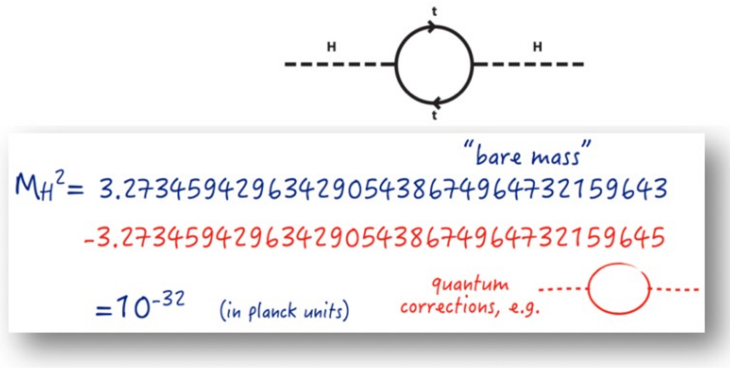
Single-production:

- *Dependent on qQ coupling (constraints from flavor physics and EW precision tests)*
- *Becomes dominant at high energies*



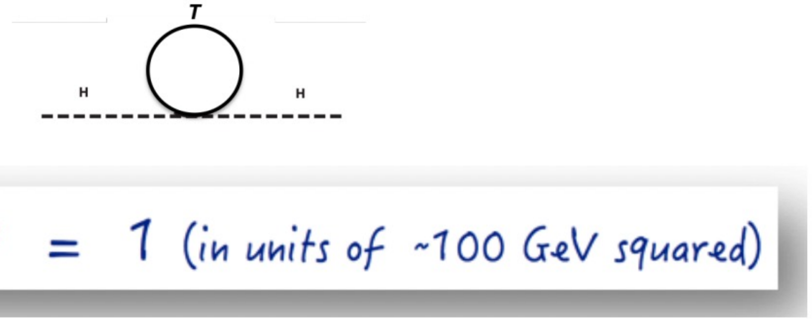


Why Vector-Like Quarks?



from
Roni Harnik

- Quantum corrections from top quark $\rightarrow$ quadratic divergence
  - Called the "Hierarchy Problem"



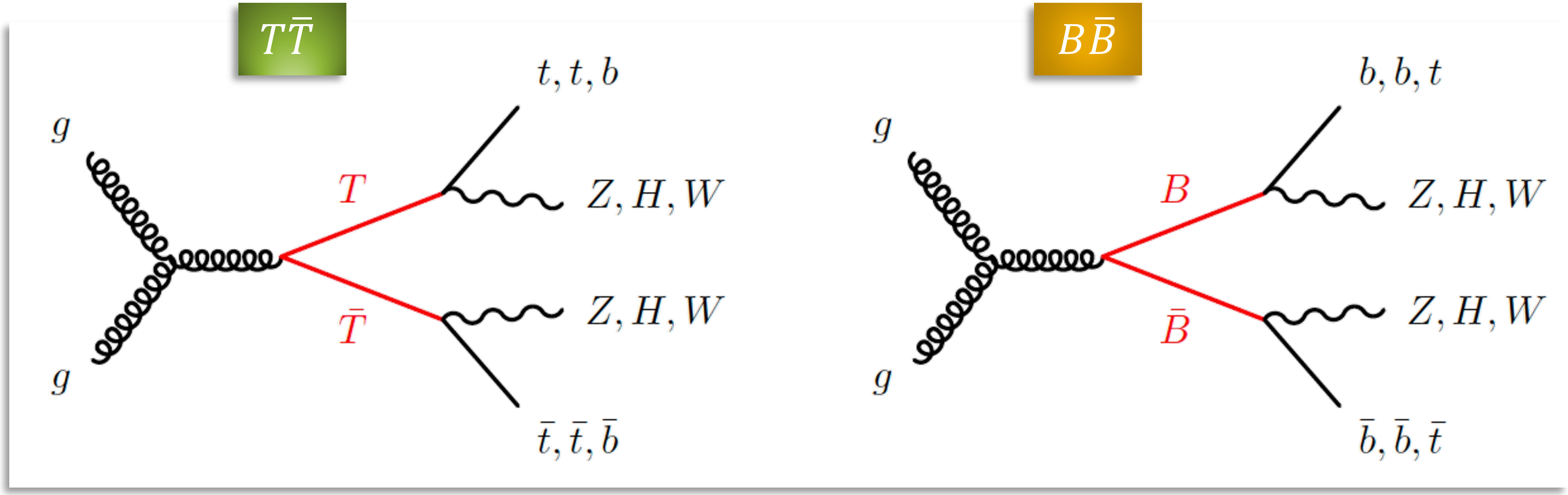
- VLT corrections $\rightarrow$ removes quadratic divergence
- VLQs are included in many models that solve the Hierarchy Problem

QCD pair-production

Pair-produced vector-like top and bottom partners in events with large Emiss

$T\bar{T}$ & $B\bar{B}$ production

Eur. Phys. J. C 83 (2023) 719

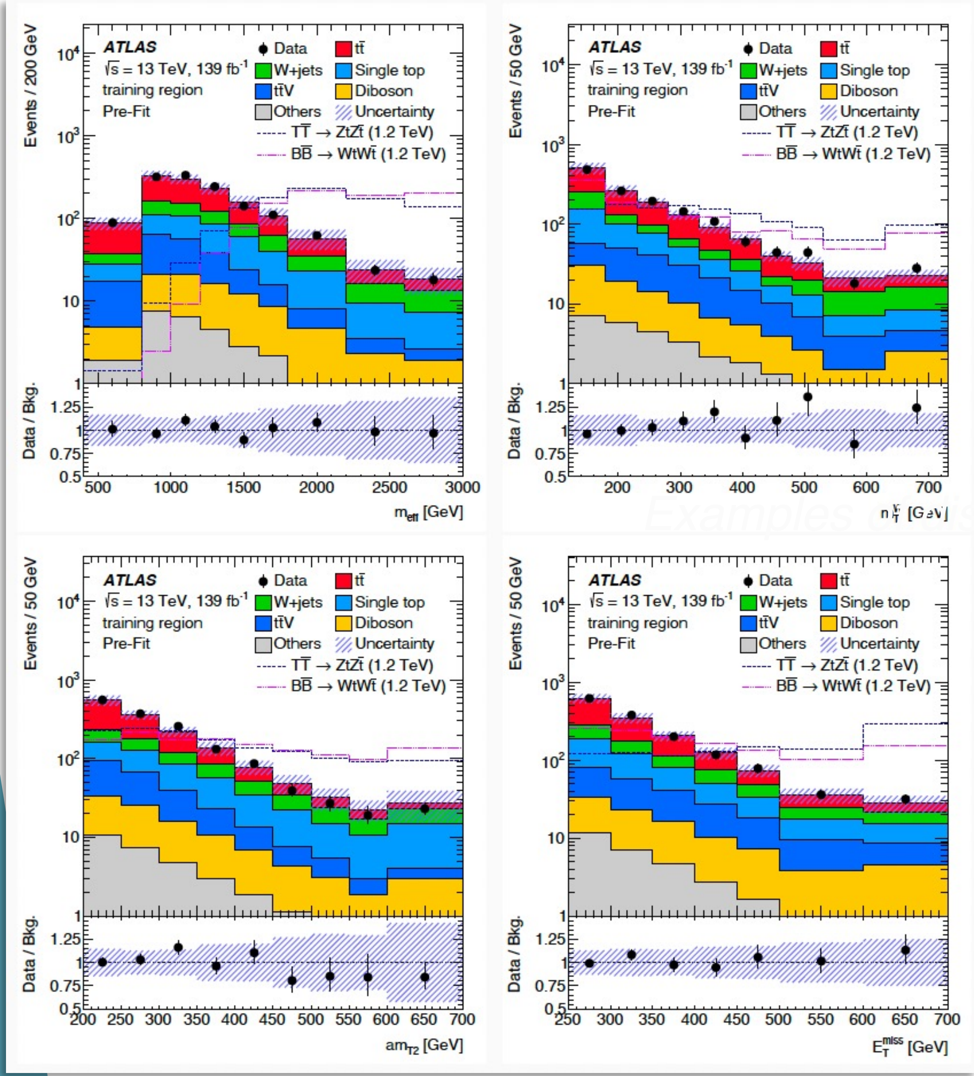


- ***Vector-like $T^{2/3}$, $B^{-1/3}$ and $X^{5/3}$ considered***
- *Events characterized by low lepton-multiplicity, high jet-multiplicity, and large missing transverse energy (E_T^{miss})*
- ***Dominant backgrounds: $t\bar{t}$ and W+jets*** ➔ *reduced using cuts on transverse mass*
- *At least one top quark from the signal expected to have a high p_T ➔requirement on large-R jets*
- ***Neural networks used to discriminate between signal and background*** ➔*Input variables such as high me for VLQ mass, properties of large-R jets, b-jet multiplicity, transverse mass etc. used*
- ***The search uses 139 fb^{-1} data collected with the ATLAS detector***

Pair-produced vector-like top and bottom partners in eventswith large Emiss

Eur. Phys. J. C 83 (2023) 719

Examples of discriminating variables



- For masses of the VLQs >800GeV
- Branching ratios $\mathcal{B}$:
  - T: $\mathcal{B}(Zt; Ht; Wb) \approx (0.25; 0.25; 0.5)$
  - B: $\mathcal{B}(Zb; Hb; Wt) \approx (0.25; 0.25; 0.5)$
- Final state signature:
  - High missing transverse momentum $E_T^{miss} > 250\ GeV$
  - Only one lepton $\ell(e\ or\ \mu)$ ➔ veto for a second lepton
  - At least 4 jets including a b-tagged jet
- Dominant background: $t\bar{t}$ and $W + jets$
  - Others: $t\bar{t}H, tWZ$ and $Z + jets$
- Neural Networks (NN) covering sections on the Branching ratios $\mathcal{B}$ plane:
  - For TT 4NN: $(0.8; 0.1; 0.1)$; $(0.2; 0.4; 0.4)$; $(0.4; 0.1; 0.5)$; $(0.4; 0.5; 0.1)$
  - For BB 3NN: $(0.1; 0.1; 0.8)$; $(0.4; 0.1; 0.5)$ and $(0.1; 0.4; 0.5)$;
- More sensitive to $T' \rightarrow Zt$, $B' \rightarrow Wt$
- Systematic Uncertainties resolution and scale of:
  - $t\bar{t}$ background
  - Jet mass
  - Effciency of lepton identification, isolation, reconstruction and energy

Pair-produced vector-like top and bottom partners in events with large Emiss

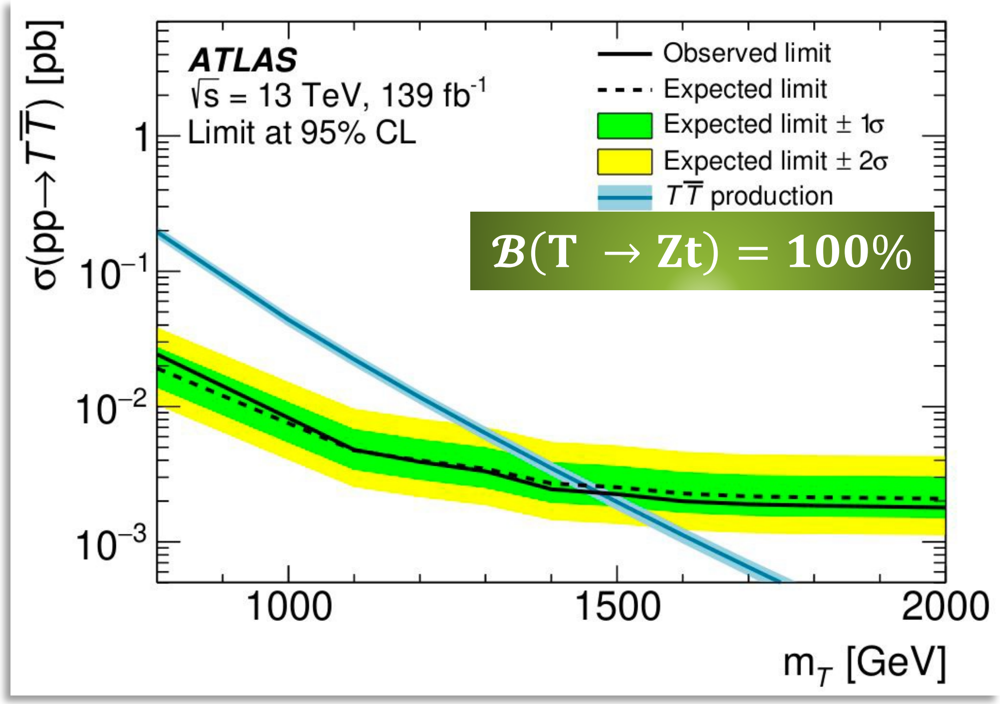
Eur. Phys. J. C 83 (2023) 719

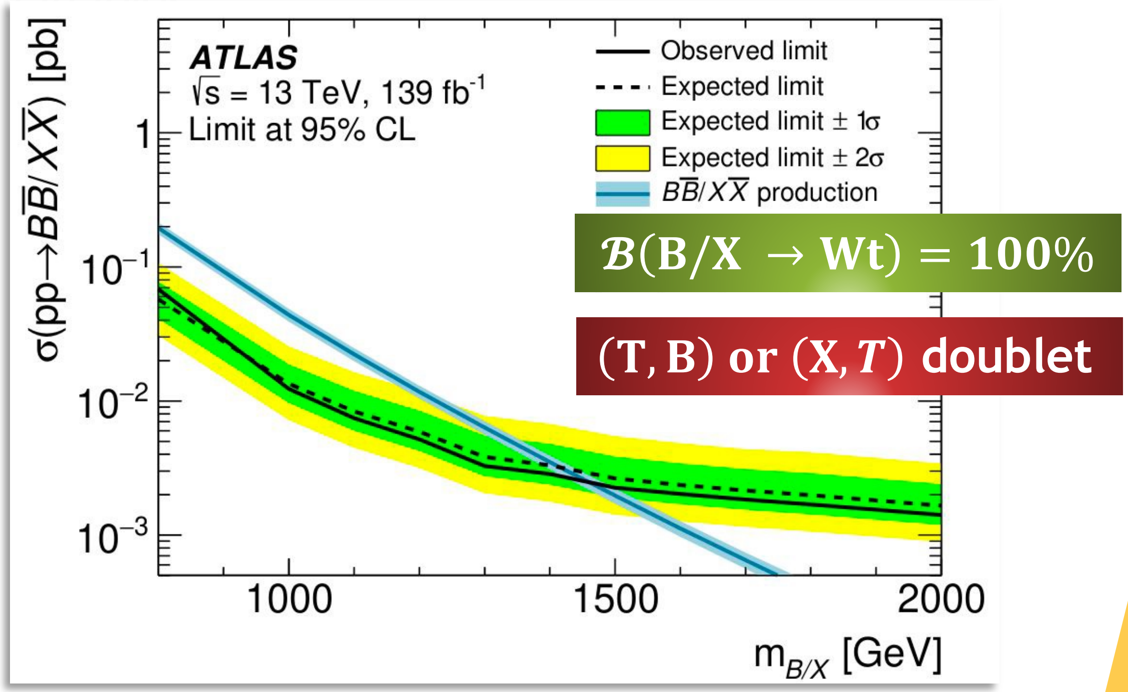
Results

VLQ	Scenario	Obs. limit (TeV)
$T(B)$	Singlet	1.26 (1.33)
T	(T, B) or (X, T) doublet	1.41
B/X	(T, B) or (X, T) doublet or $\mathcal{B}(B/X) \to Wt = 100$	1.46
$T/B/X$	(T, B) or (X, T) doublet mass-degenerate	1.59

- ✓ *No significant excesses*
- ✓ *Analysis most sensitive to the $T \to Zt$ and $B \to Wt$ decay modes*
- ✓ *Strongest limits for the (T,B) and (X,T) when $m_X = m_T = m_B$ are at 1.59 TeV*

Expected and observed upper limits on the signal cross-section

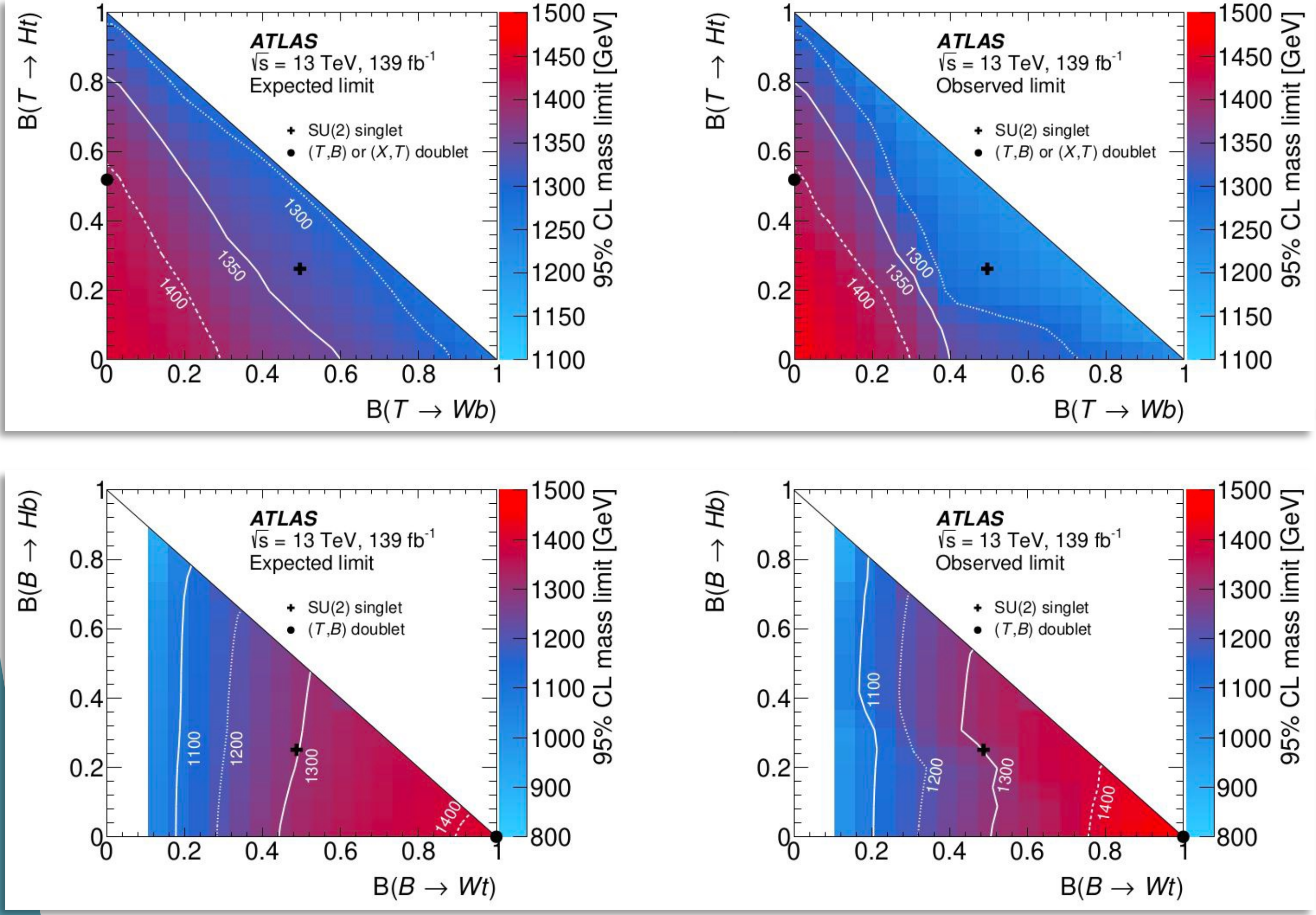




Pair-produced vector-like top and bottom partners in events with large Emiss

Eur. Phys. J. C 83 (2023) 719

Results



- ✓ *No significant excesses*
- ✓ *Expected and observed mass limits as a function of the T' and B' branching ratios $\mathcal{B}$*
- ✓ *Analysis most sensitive to the T → Zt and B→Wt decay modes*
  - ✓ $\boldsymbol{\mathcal{B}(\mathrm{T}' \rightarrow \mathrm{Zt}) = 100\%}$
  - ✓ $\boldsymbol{\mathcal{B}(\mathrm{B}' \rightarrow \mathrm{Wt}) = 100\%\%}$
- ✓ *Strongest limits corresponding to the weak-isospin doublet model ➔ (T,B) and (X,T) when $m_X = m_T = m_B$ are at 1.59 TeV*
- ✓ *1.47 TeV for exclusive T → Zt* decays
- ✓ *1.46 TeV for exclusive B/X → Wt* decays
- ✓ *Lower limits on the T and B quark masses are derived for all possible branching ratios*

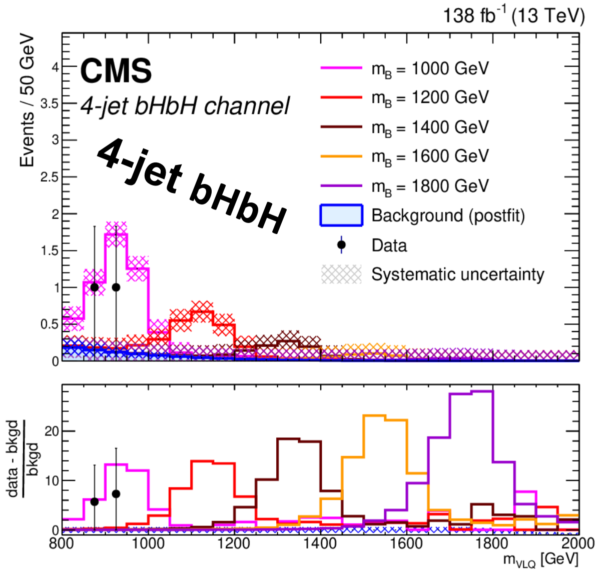
A search for bottom-type, vector-like quark pair production in leptonic and fully hadronic final states

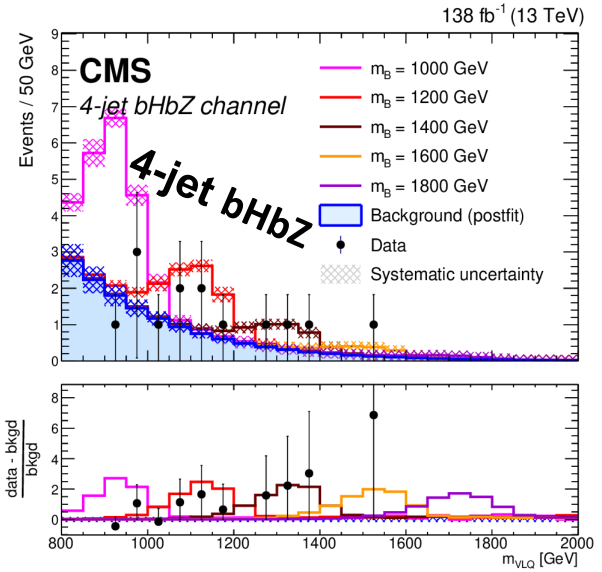
B2G-20-014 Submitted to Phys. Rev. D

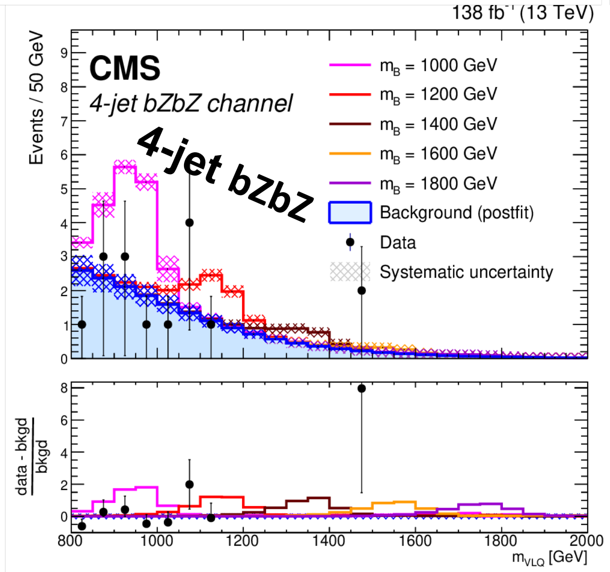
February 2024

- ***For masses of the VLQs from 1000 to 1800 GeV***
- *Branching Ratios $\mathcal{B}$:*
  - ***Leptonic: $\mathcal{B}(Zb, Hb, Wt)$***
  - ***Hadronic: $\mathcal{B}(Zb, Hb)$***
- ***Fully hadronic category:***
  - *At least 4 (<=6) AK4 jets $P_T > 50$ GeV $|\eta| < 2.4$, $H_T > 1350$ GeV*
  - No isolated e or μ $P_T > 50$ GeV
  - Bkg: SM jets produced through the strong interaction (QCD multijet events).
- **Leptonic category:**
  - At least 3 (<=5) AK4 jets $P_T > 50$ GeV and $|\eta| < 2.4$
  - At least one pair of leptons $80 < m_{ll} < 102$ GeV
  - Bkg: Drell-Yan dilepton production in association with jets
- **Systematic uncertainties:**
  - Integrated luminosity, trigger, dilepton Z boson efficiency, scale factors…

Reconstructed VLQ mass







hadronic category

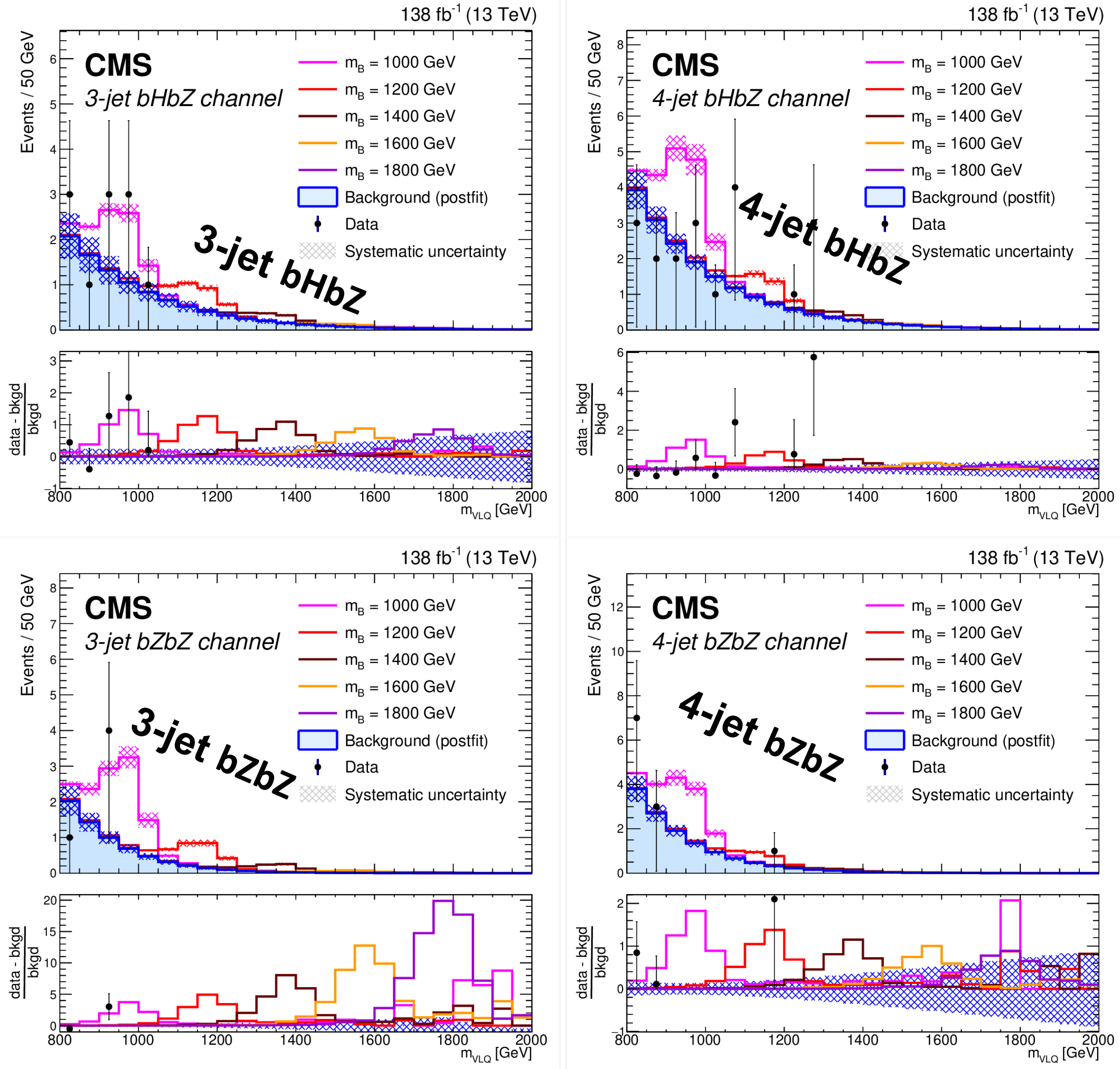
expected postfit background (blue histogram)
signal plus background (colored lines)
observed data (black points)

A search for bottom-type, vector-like quark pair production in leptonic and fully hadronic final states

B2G-20-014 Submitted to Phys. Rev. D

February 2024

Reconstructed VLQ mass



Leptonic category

expected postfit background (blue histogram)

signal plus background (colored lines)

observed data (black points)

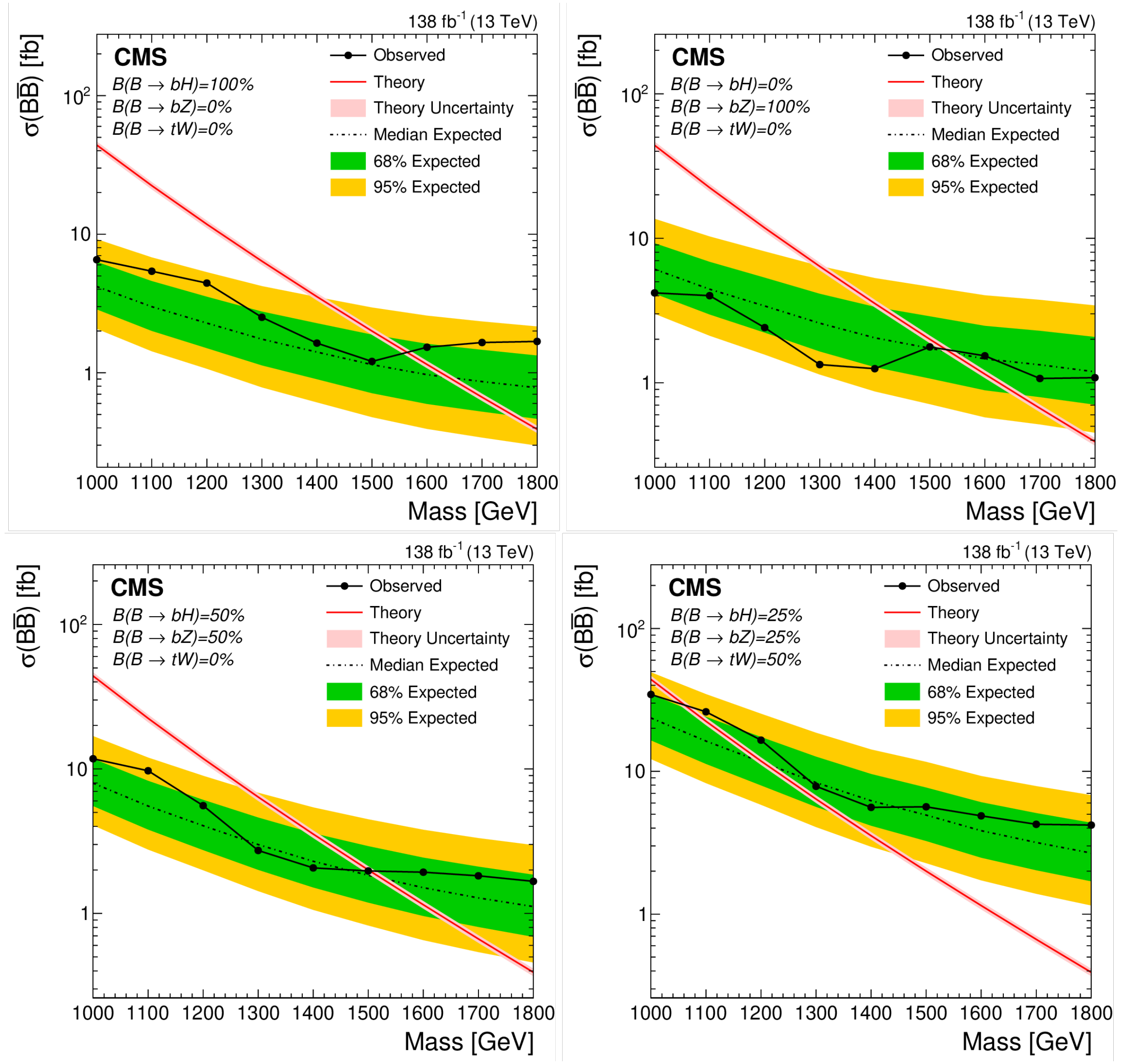
A search for bottom-type, vector-like quark pair production in leptonic and fully hadronic final states

B2G-20-014 Submitted to Phys. Rev. D

February 2024

Expected and observed limits on the cross section at 95% CL

- No excess over the expected background is observed.
- Lower limits are set on the B VLQ mass at 95% confidence level.
- These depend on the B VLQ branching fractions and are 1570 and 1540 GeV for 100% B→bH and 100% B→bZ, respectively.
- In most cases, the mass limits obtained exceed previous limits by at least 100 GeV.



The theoretical cross section and its uncertainty are shown by the red line and light-red band.

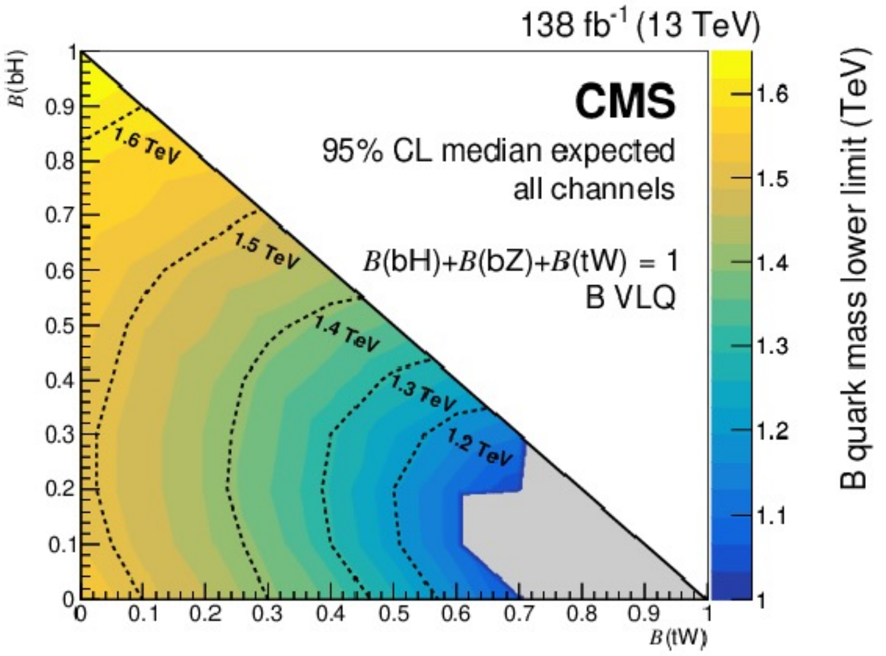
A search for bottom-type, vector-like quark pair production in leptonic and fully hadronic final states

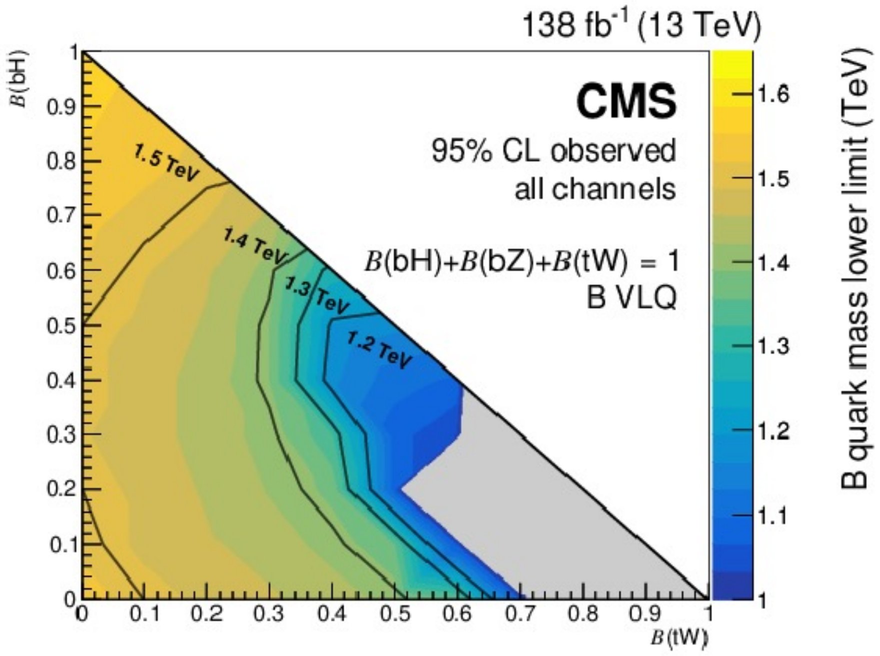
B2G-20-014 Submitted to Phys. Rev. D

February 2024

Expected exclusion limits on the VLQ mass at 95% CL as a function of the branching fractions

- No excess over the expected background is observed.
- Lower limits are set on the B VLQ mass at 95% confidence level.
- These depend on the B VLQ branching fractions and are 1570 and 1540 GeV for 100% B→bH and 100% B→bZ, respectively.
- In most cases, the mass limits obtained exceed previous limits by at least 100 GeV.



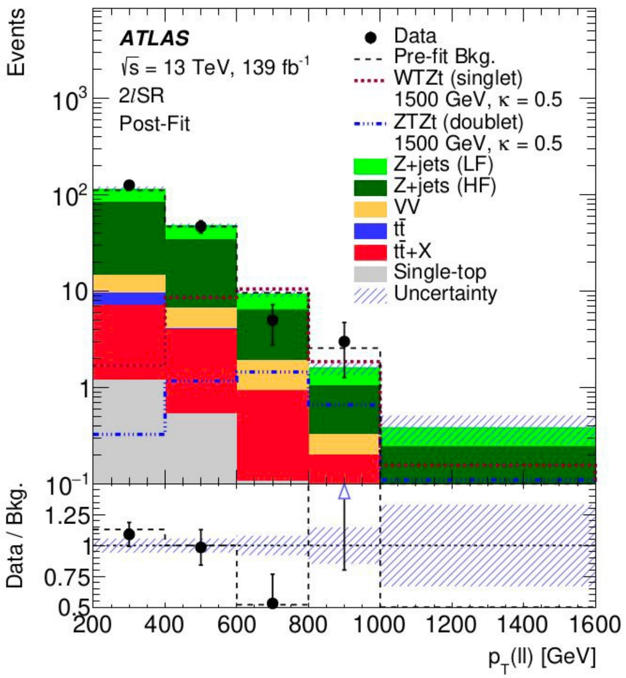


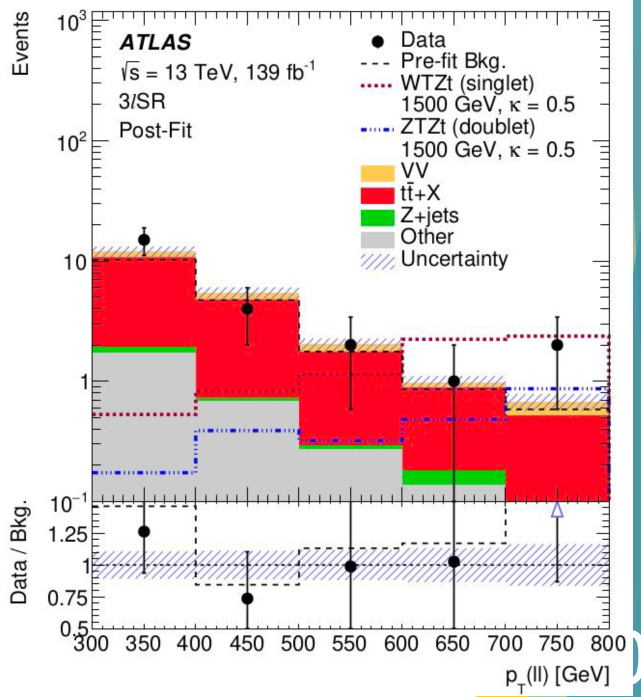
- ***Single-production***

$T \to Zt$ (multileptonic)

2307.07584

- For masses of the VLQs from 1 TeV to 2 TeV
- Relative strength coupling ξ
  - ξ: (ξW, ξZ, ξH) ≈ (0.5, 0.25, 0.25)
  - Signal samples normalized to NLO cross-section (NWA)
- Final state signature 2l:
  - Z boson: $|m_{\ell\ell} - m_Z| < 10$ GeV
  - $P_{T\ell\ell} > 200$ GeV
  - $H_T > 300$ GeV
  - $H_T + E_T^{miss} < m_{\ell\ell J}$
  - 1 FWD jet; at least 1 b-jet & 1 top-tagged jet
- Final state signature 3l:
  - 3 leptons passing the preselection
  - At least 1 FWD jet, 1 b-tagged jet
  - Z boson candidate PT(ℓℓ) > 300 GeV
  - Leading lepton PT(ℓ) > 200 GeV
  - $H_T \cdot n(\text{jets}) < 6$ TeV
- Main Backgrounds:
  - 2l: Z+jets, minor contributions from VV and tt-processes
  - 3l: Diboson processes and tt+Z and other small contributions from tt+W and tttt

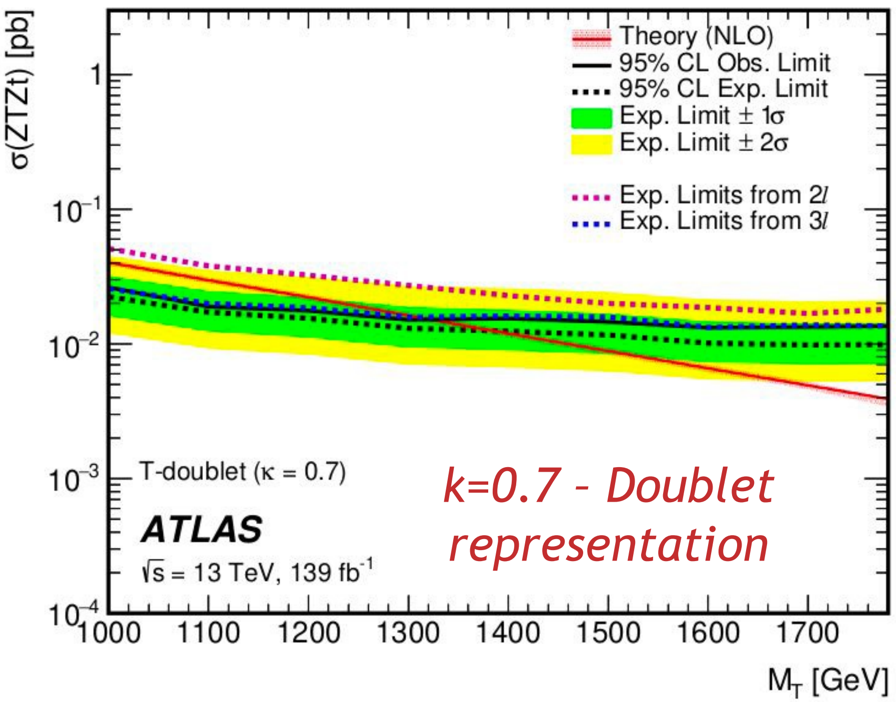


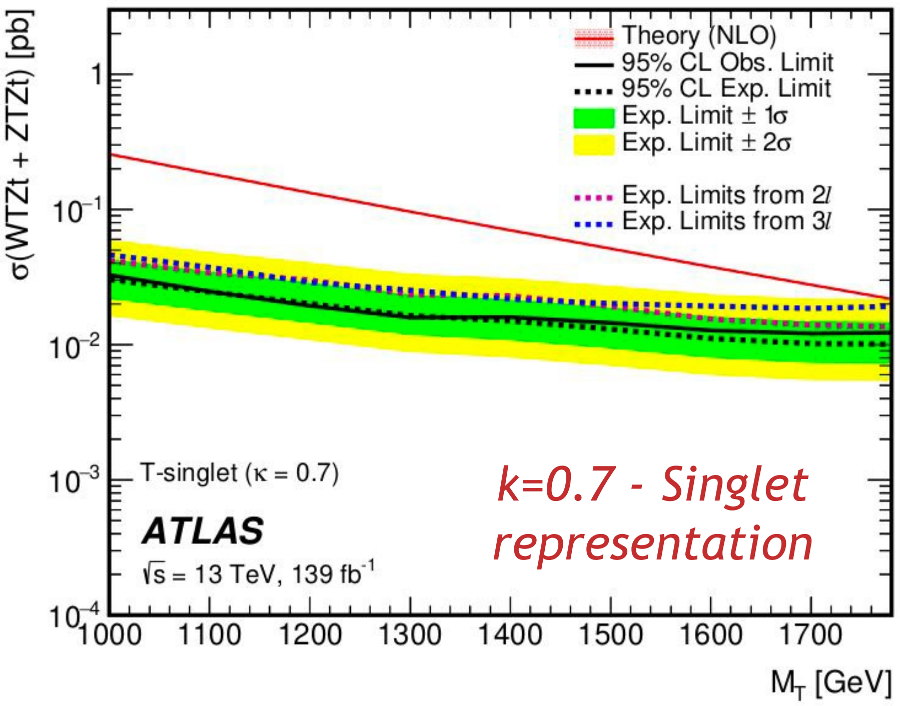


T’→ Zt (multileptonic)

2307.07584

- Systematic uncertainties
  - Experimental sources:
    - electron energy scale and resolution
    - muon momentum scale and resolution
    - flavor tagging, jet energy scale and resolution…
  - Theoretical sources:
    - cross section
    - other modeling uncertainties for all background samples
- Jets misidentified as leptons



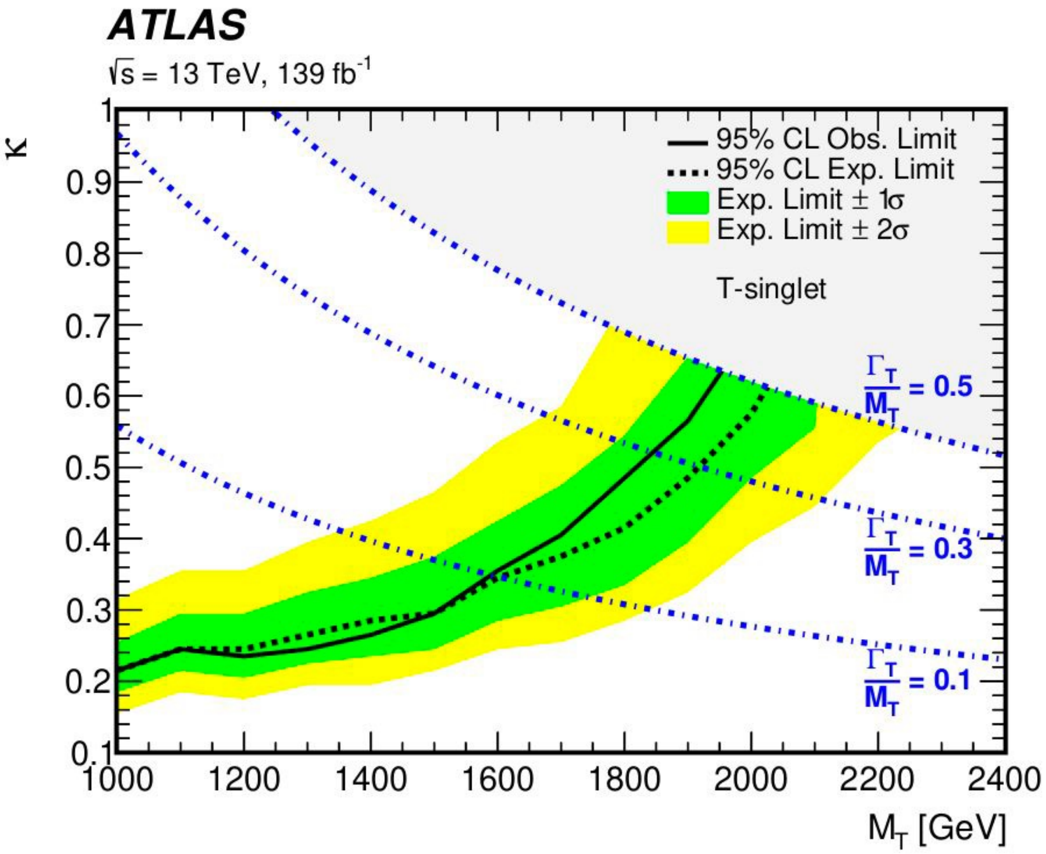


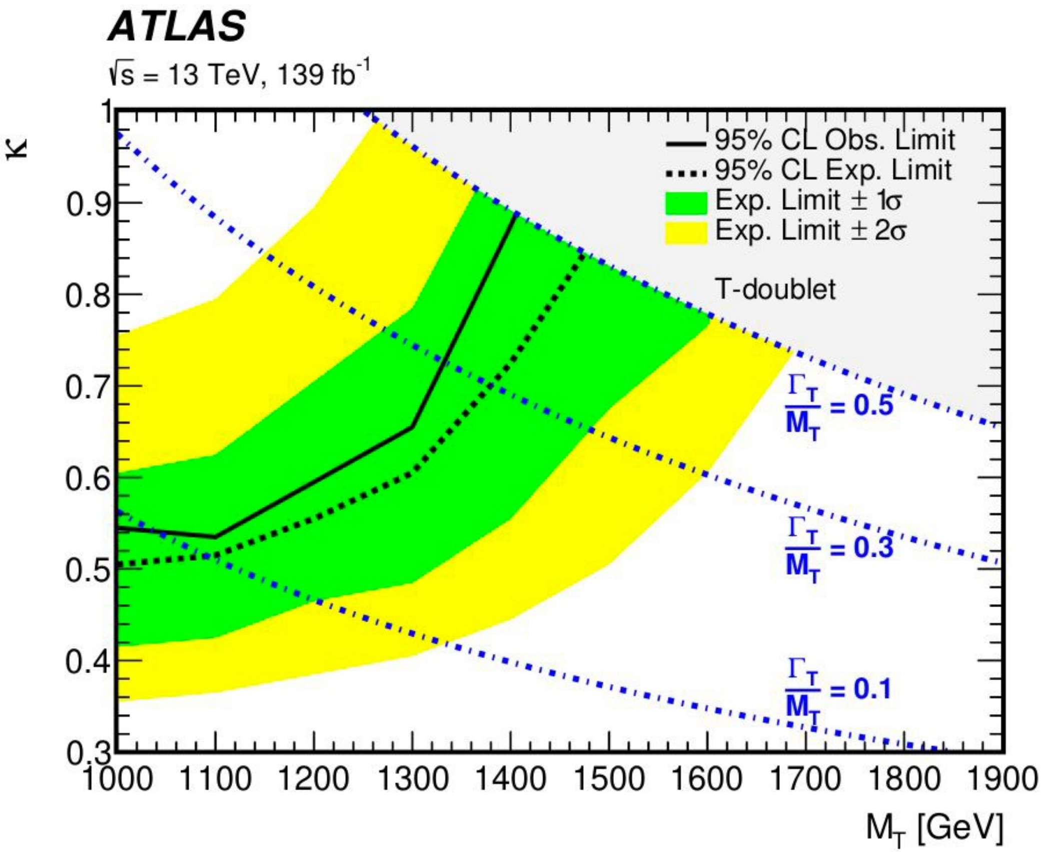
Observed and Expected limits at 95% on total cross-section
- k=0.7 - Singlet representation
- k=0.7 - Doublet representation

$T' \rightarrow Zt$ (multileptonic)

2307.07584

Observed and expected limits at 95% CL on the top partner coupling as a function of the T mass



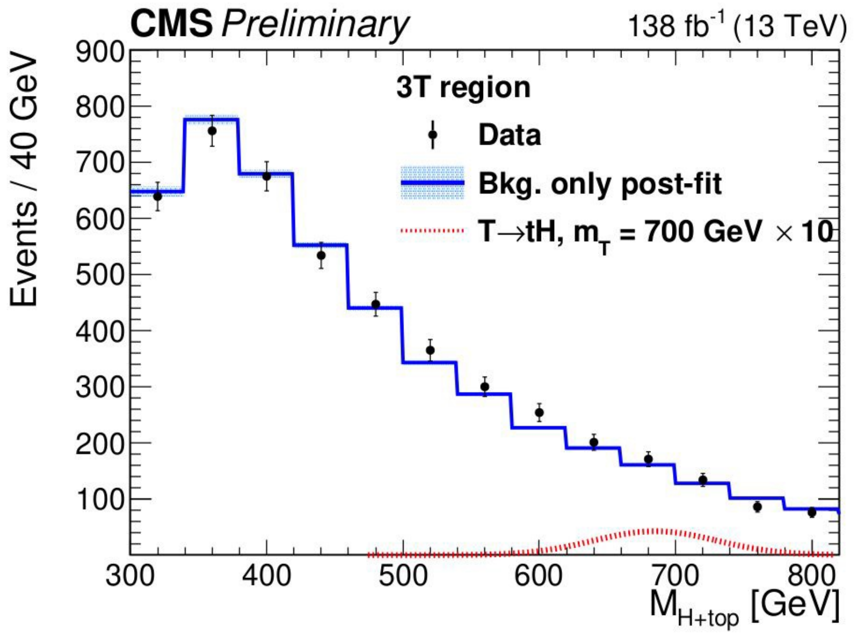


The strongest exclusion is observed for singlet representation with ξ_W approx 0.5 where masses up to 1975 GeV are excluded at relative decay width of Γ_T/M_T=0.5 for the top partner.

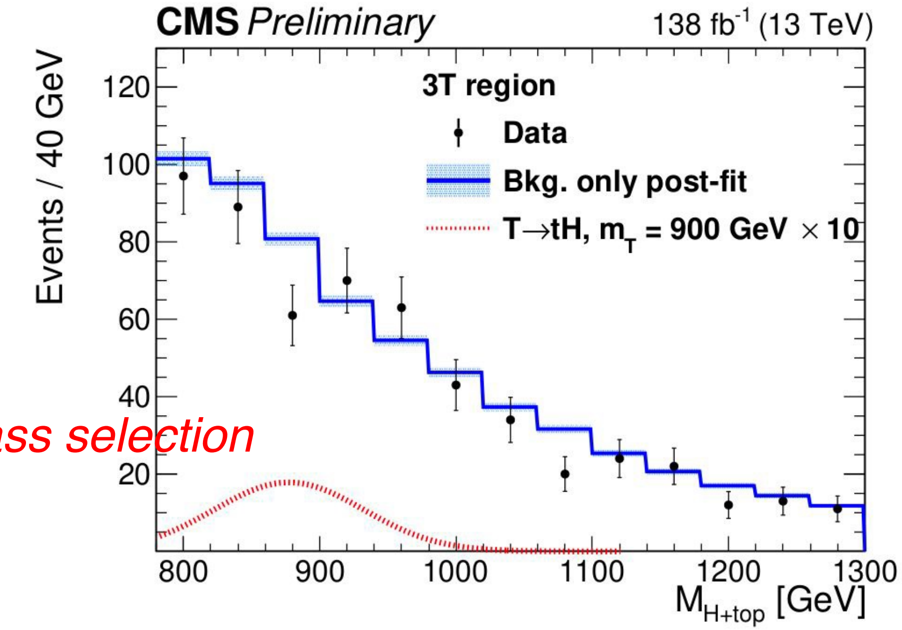
$T' \to Ht\ /\ Zt$

B2G-19-001

- For masses of the VLQs from 600 - 1200 GeV
- Branching Ratios B:
  - T': B $(Zt, Ht, Wb) \approx (0.25, 0.25, 0.5)$
- Final state signature:
  - 5 jets, single production 2 additional jets 3 of them b-jets
  - P_T > 400 GeV (2016)
  - P_T > 300 GeV (2017 & 2018)
  - m_T up to 700 GeV (low-mass selection)
  - m_T above 800 GeV (high-mass selection)
- Main Bkg process:
  - multijet
  - tt+ jets
- Systematic Uncertainties
  - Trigger efficiency
  - Jet energy and resolution uncertainties
  - b tagging efficiency scale factor for jets
- Invariant mass reconstructed from 5 jets is used as the main discriminating variable



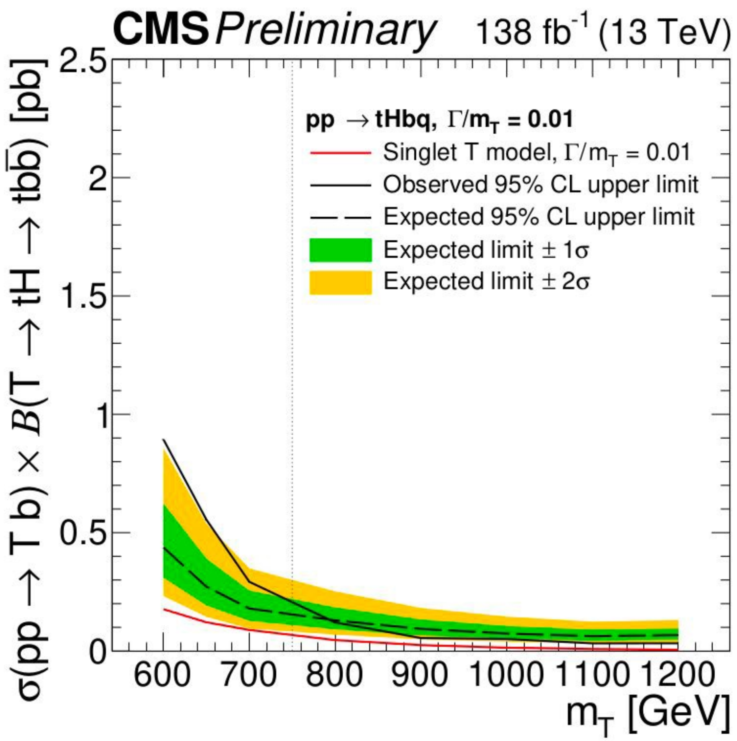
Low-Mass selection designed to avoid distorting the 5-jet invariant mass distribution and producing artificial peaks (see backup).

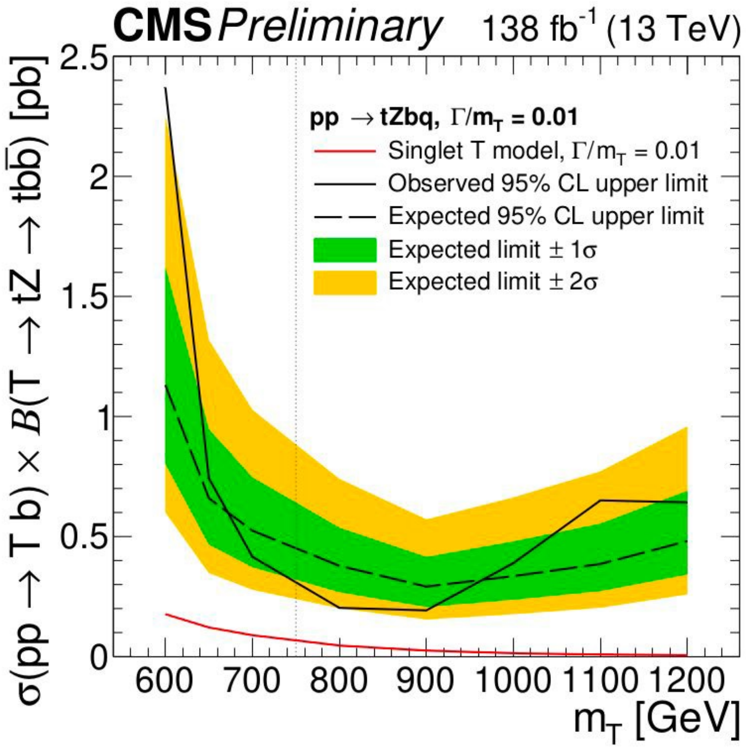


High-Mass selection

$T' \rightarrow Ht$ / Zt

Expected and Observed 95% CL upper limits on the cross-section for associated production with a b for final states tHbq and tZbq, for T masses from 600 - 1200 GeV



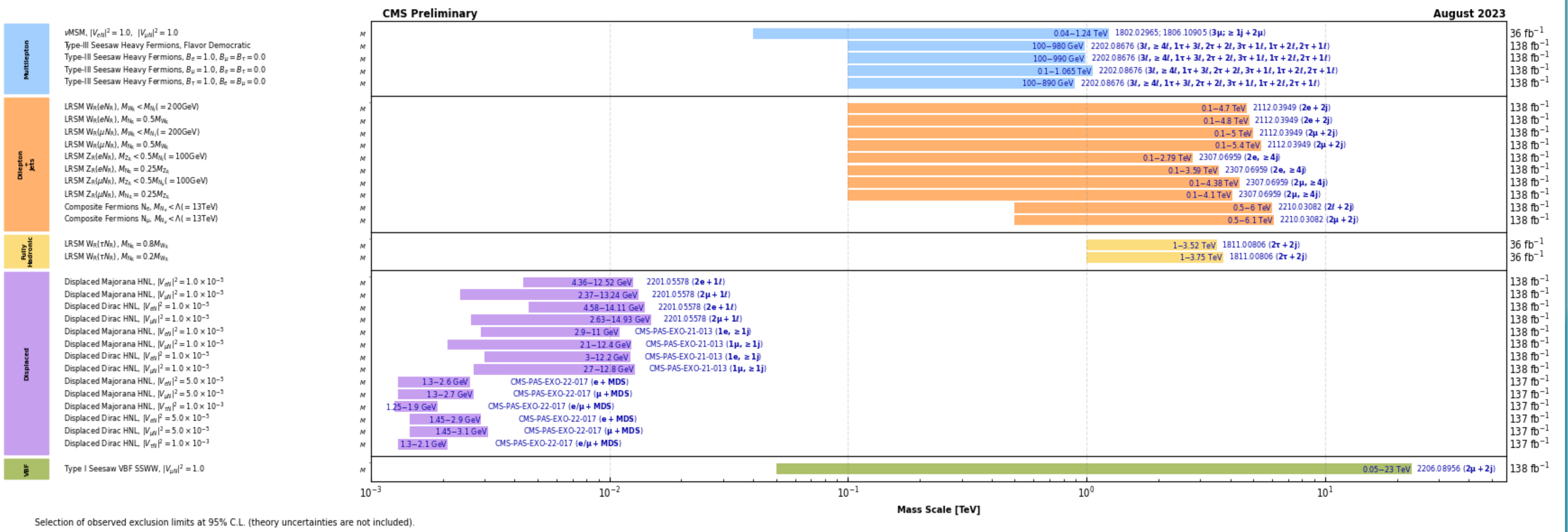


Excess in the tH final state found in [1909.04721], is not observed with a larger dataset.
The limits are stronger than those in the previous search by at least a factor of three

Exotics summary plots

Heavy Neutral Leptons [HNLs]

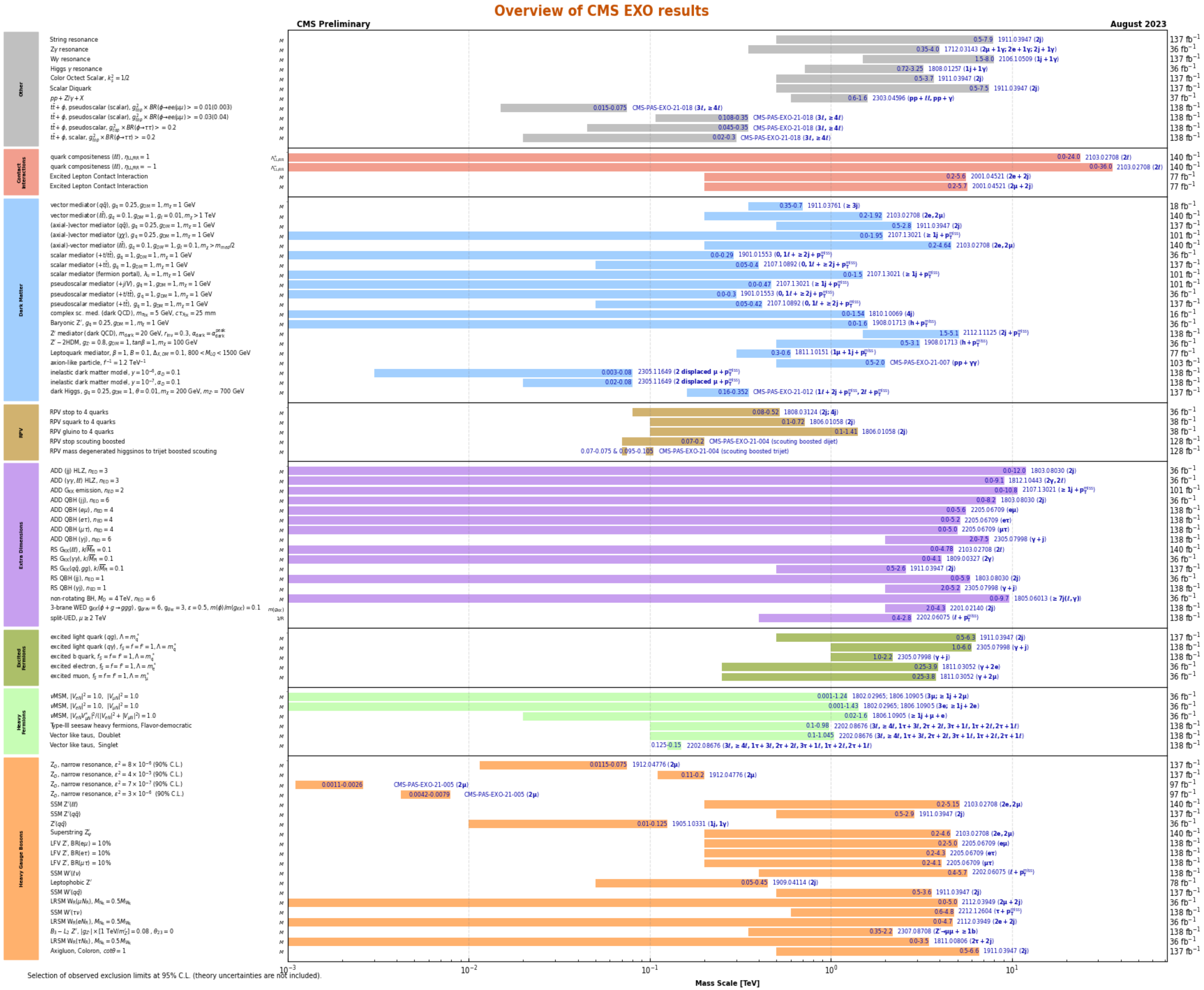
Overview of CMS HNL results



Many searches, in different final states, and with both prompt and displaced signatures from ATLAS, CMS & LHCb

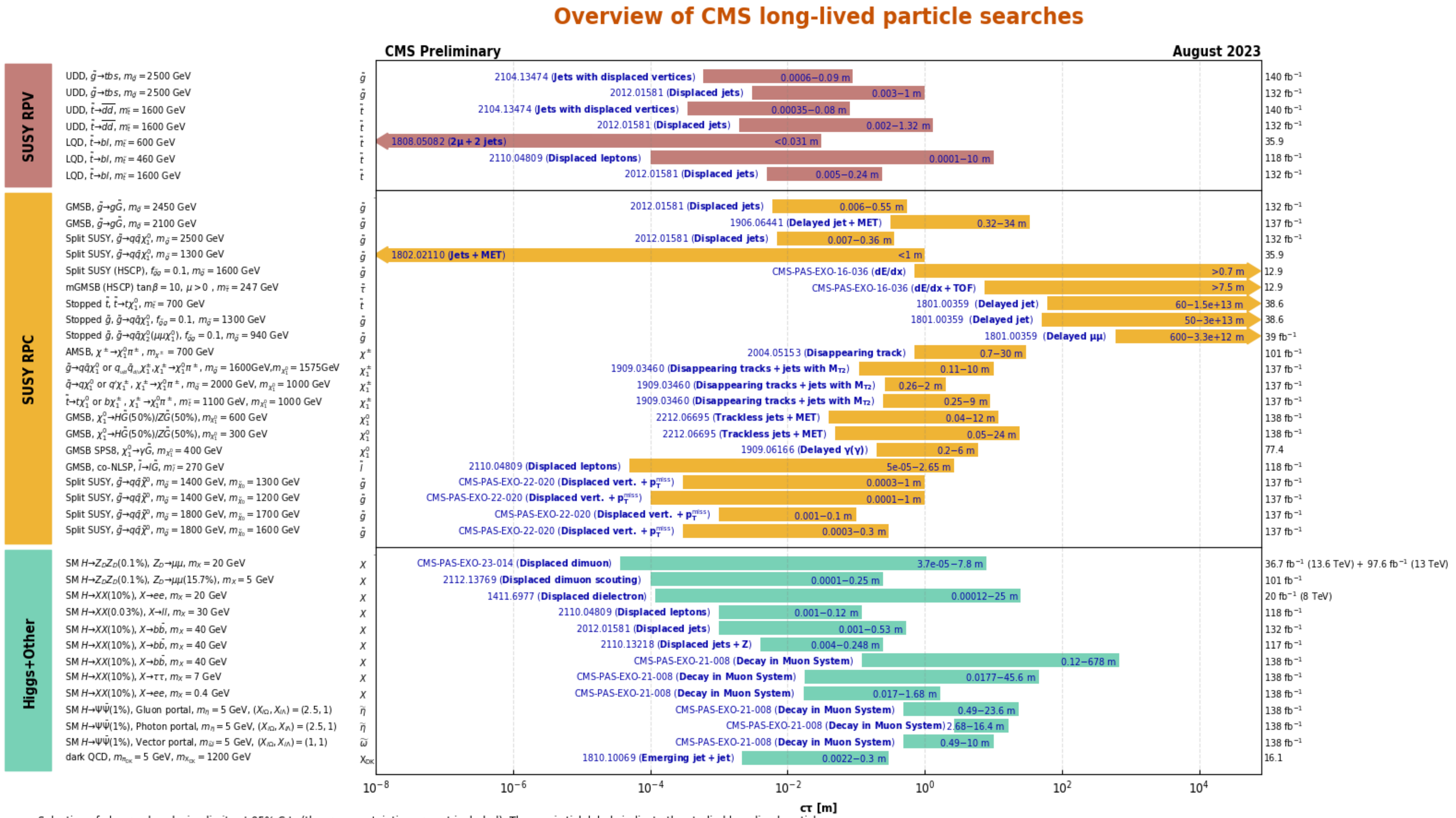
https://twiki.cern.ch/twiki/bin/view/CMSPublic/SummaryPlotsEXO13TeV

Ranges of new particle masses or energy scales excluded at the 95% confidence level



https://twiki.cern.ch/twiki/bin/view/CMSPublic/SummaryPlotsEXO13TeV

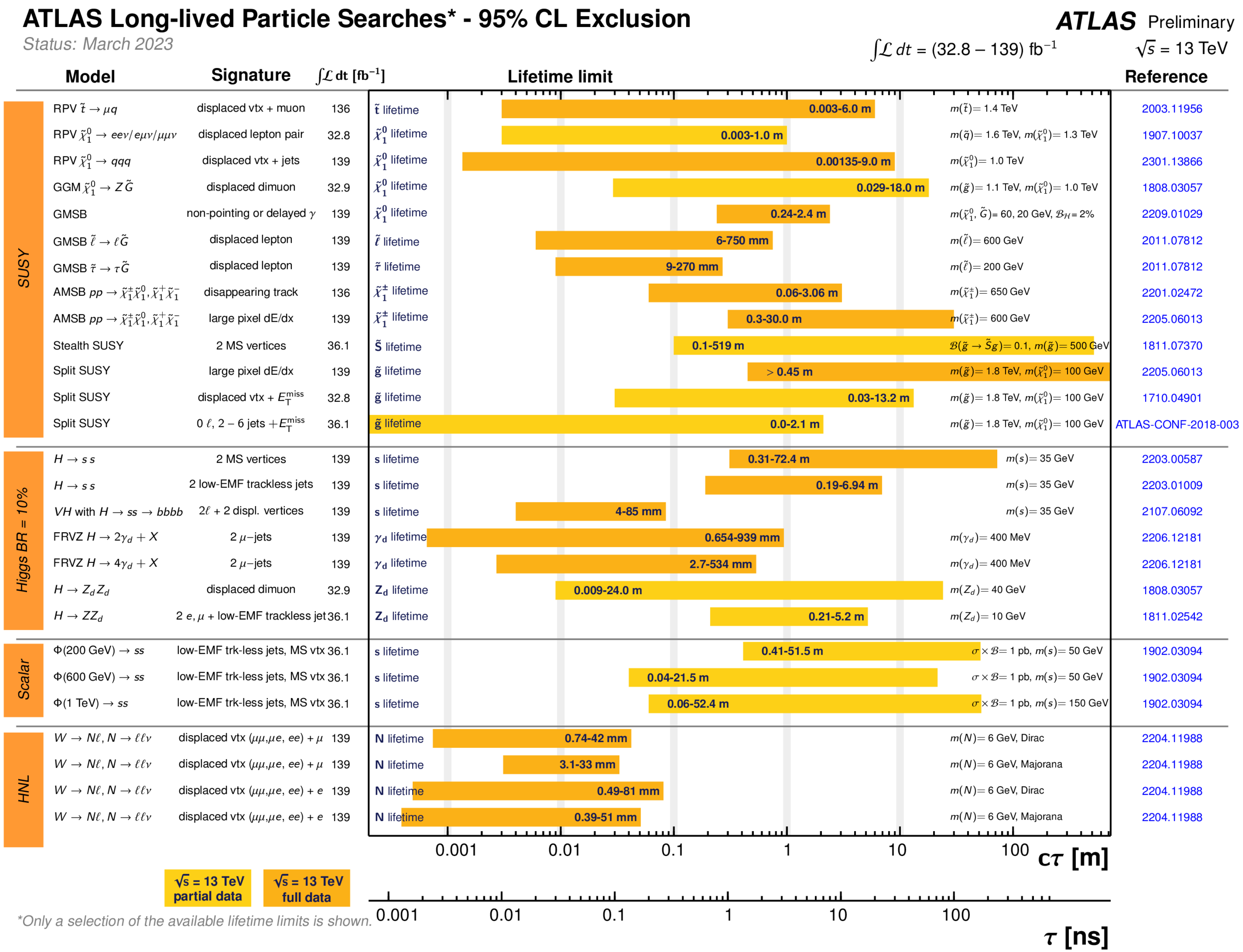
Ranges of new particle lifetimes excluded at the 95% confidence level



Selection of observed exclusion limits at 95% C.L. (theory uncertainties are not included). The y-axis tick labels indicate the studied long-lived particle.

https://twiki.cern.ch/twiki/bin/view/CMSPublic/SummaryPlotsEXO13TeV

Ranges of new particle lifetimes excluded at the 95% confidence level



https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2023-008/