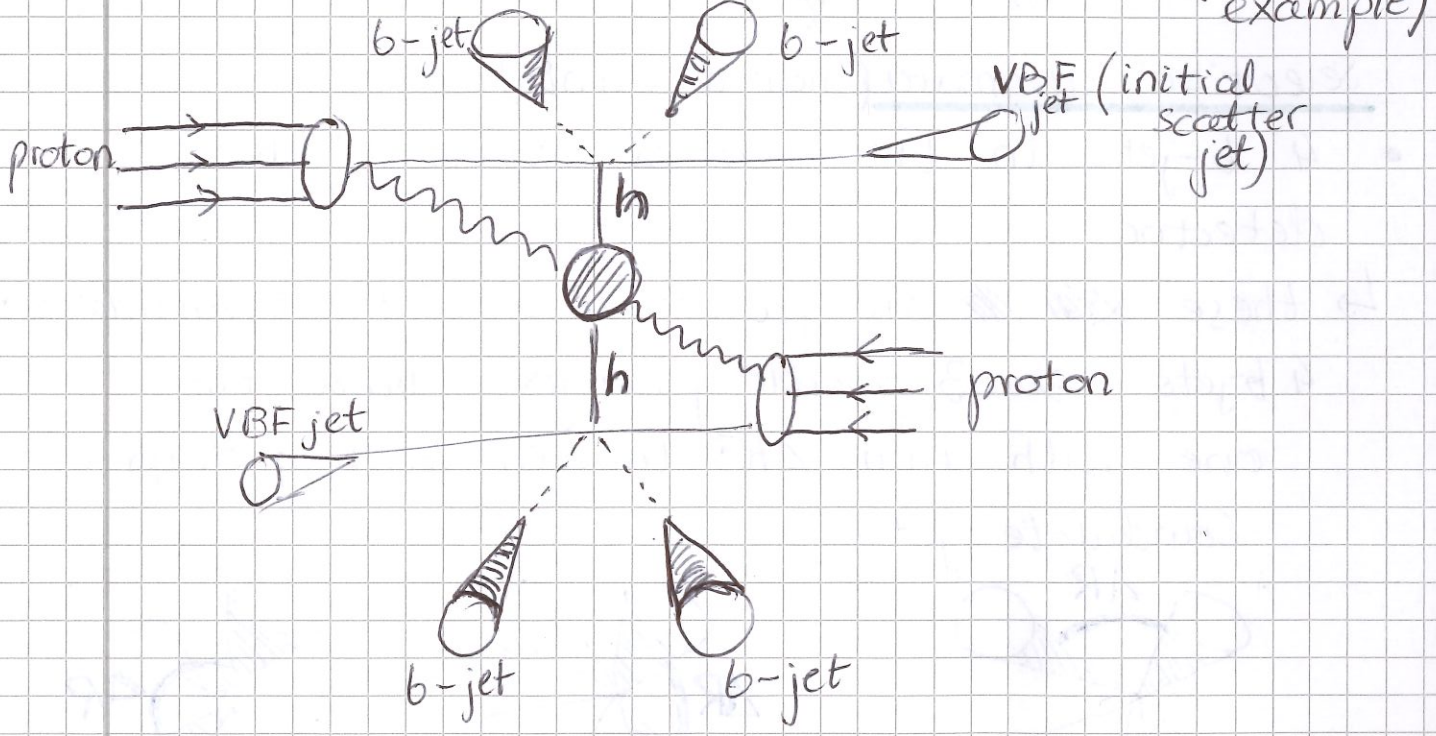
Lecture 3 : HH → 4b analysis and background estimation challenge

Last lecture : general overview of Higgs and di-Higgs status at LHC (ATLAS in particular)

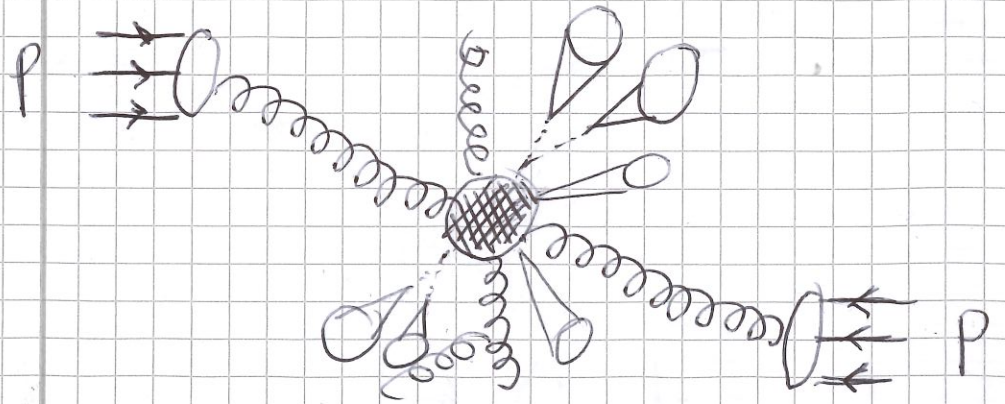
Today : details of 4b analysis, especially neural network reweighting use case for background estimation

---

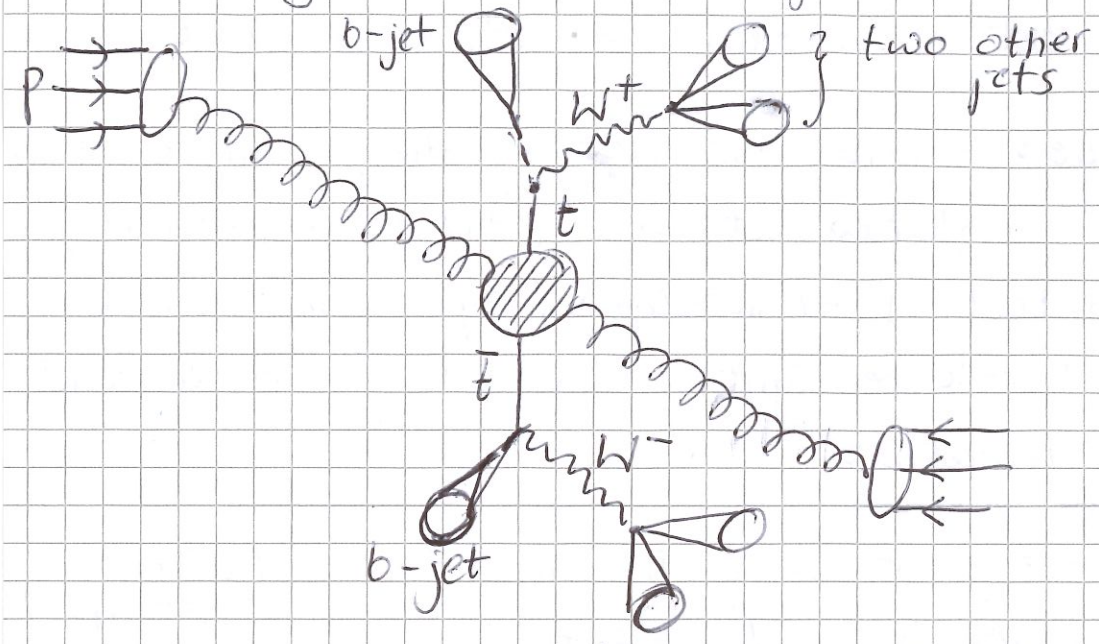
4b event we want to see in the detector (VBF example)



However, what we also see → a lot of background, mainly (≈95%): QCD multijet background (example process)

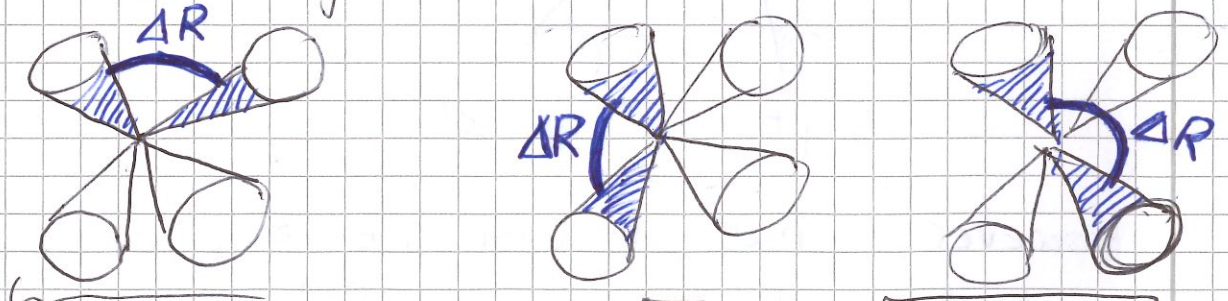


Remaining 5% : $t\bar{t}$ background



Selection summary (signal events):

- 4 b-jets in the central region of the detector
  - ↳ these are paired into 2 Higgs Candidates:
    4 b-jets ⇒ 3 possible pairings : choose the one with min ΔR^* between leading Higgs Candidate jets



↳ min ΔR pairing

$^*\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2}$

- for VBF : additionally 2 jets in the forward region with large rapidity separation

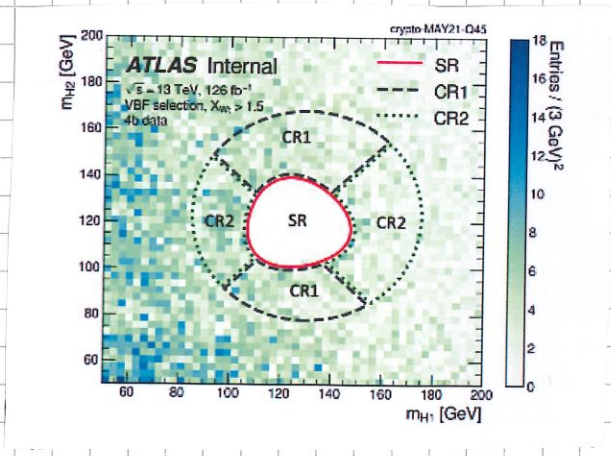
$$\Delta\eta_{jj} > 3 \; , \; m_{jj} > 1\,\text{TeV}$$

- additionally signal events must pass Signal Region requirement: this is an area centered around SM Higgs mass defined in $m_{H1} - m_{H2}$ massplane

- finally, reduce background, define a variable that vetos the events coming from the hadronic top decays
- the rest of the background has to be estimated, classic approach: Monte Carlo simulations, however, these are not precise enough for our QCD background case ⇒ use data-driven methods instead

Background estimation

↳ make use of the massplane: $m_{H_1} - m_{H_2}$
↳ define control regions (CR) around the signal region (SR)
↳ by definition: signal events are placed in SR, CR is rich in background



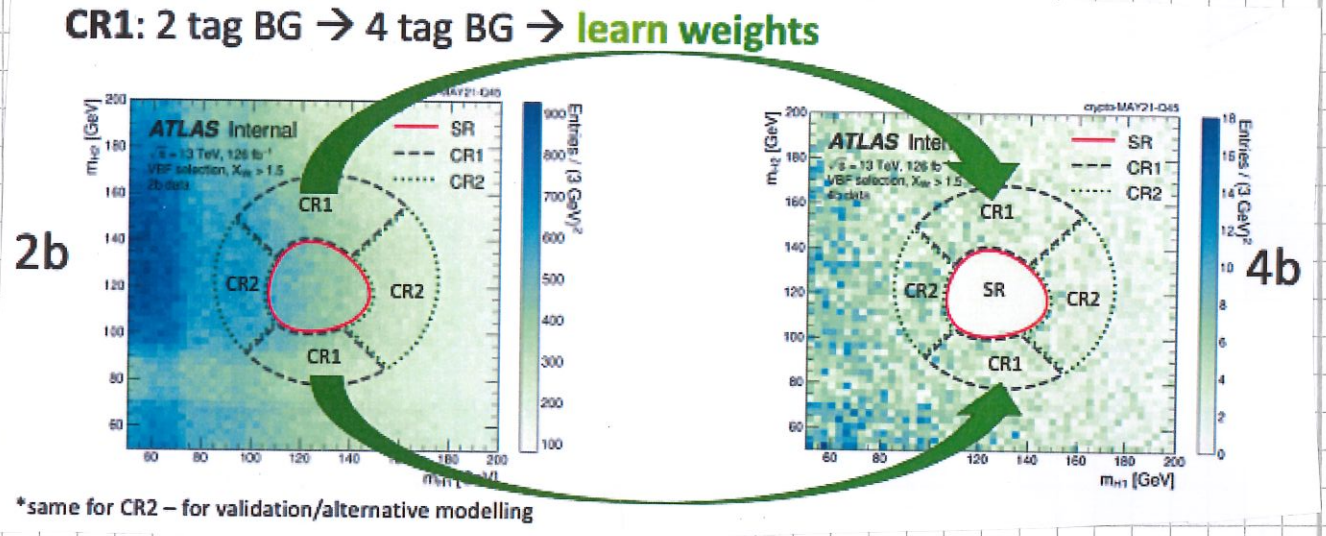
- in simplest case, one could interpolate the background from control Regions into the SR (ABCD method), but it assumes no correlations between variables (here $m_{H_1} - m_{H_2}$)
  ↳ not the case here

⇓

more complex method is needed

=3=

- introduce background-only 2 b-tagged events: those are events chosen in the same way as 4b events, except in the central region we have exactly 2 b-tagged jets and 2 other jets (highest p_T); pairing of the jets is done as for 4b events



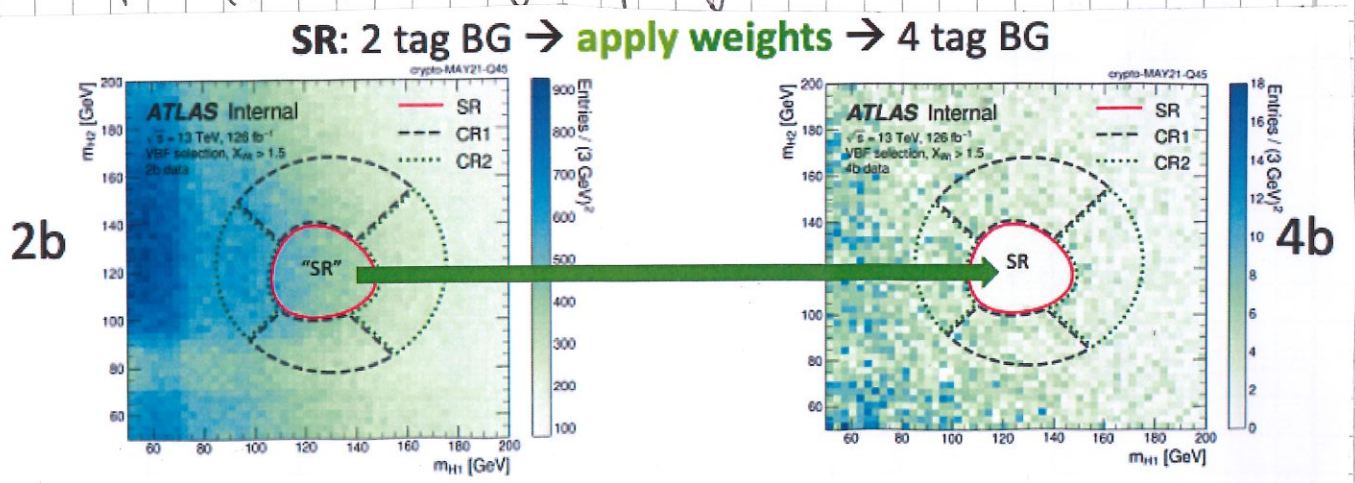
assume: $p_{2b}(x) \cdot w(x) = p_{4b}(x)$

$$\Rightarrow \quad w(x) = \frac{p_{4b}(x)}{p_{2b}(x)}$$

$w(x)$ → reweighting function (to be learned by the network)

$p_{4b}(x)$ → 4-tag probability density function

$p_{2b}(x)$ → 2-tag pdf



How is this concept implemented in practice?
↳ can only m_{HH} be used?
↳ biased estimate, may result in influence from signal-like fluctuations

↓

solution: a high dimensional model based on kinematic variables correlated to m_{HH} (e.g. p_T of Higgs candidates)

simplest idea: take n variables in form of 2b and 4b histograms
↳ perform a bin-by-bin reweighting from 2b to 4b
↳ multiply the results from each individual reweightings to get the final result/weight
↳ problems: no correlations between variables considered, curse of dimensionality

⇒ solution: Neural Network Reweighting

- input data: same/similar variables as for histograms method
- loss function (the way network learns): aim: minimize the loss function

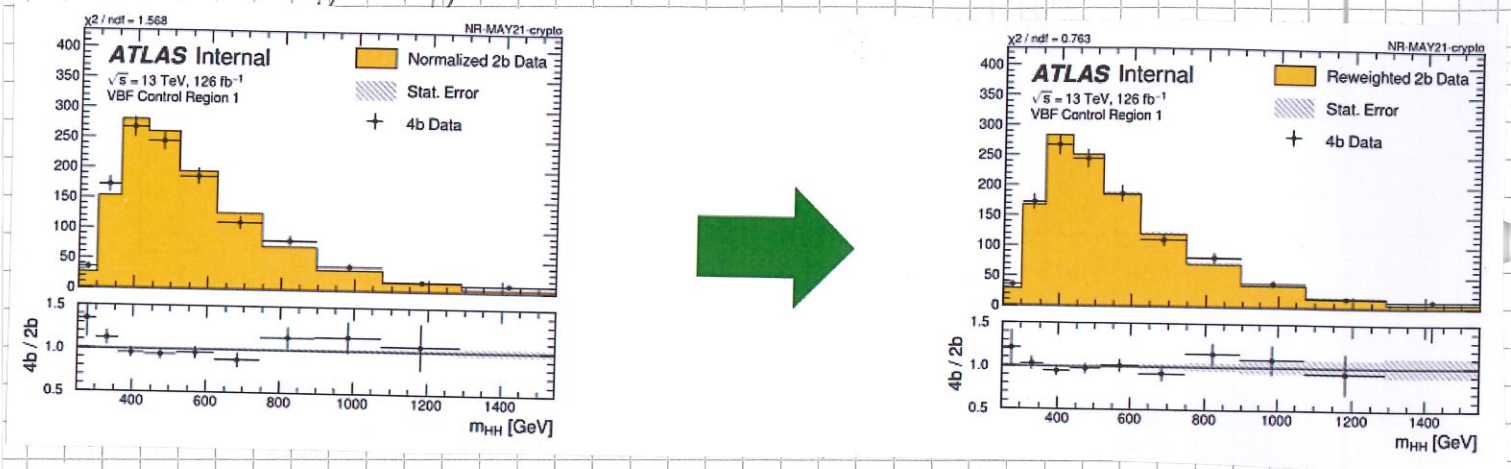
choose such a loss function that when minimized becomes: $w(x) = \frac{P_{4b}(x)}{P_{2b}(x)}$

↳ used here:

$$\mathcal{L}(R) = \mathbb{E}_{x\sim p_{2b}}\left[\frac{1}{R(x)}\right] + \mathbb{E}_{x\sim p_{4b}}\left[\frac{1}{R(x)}\right]$$

↳ find estimator $R(x)$ that minimizes the loss

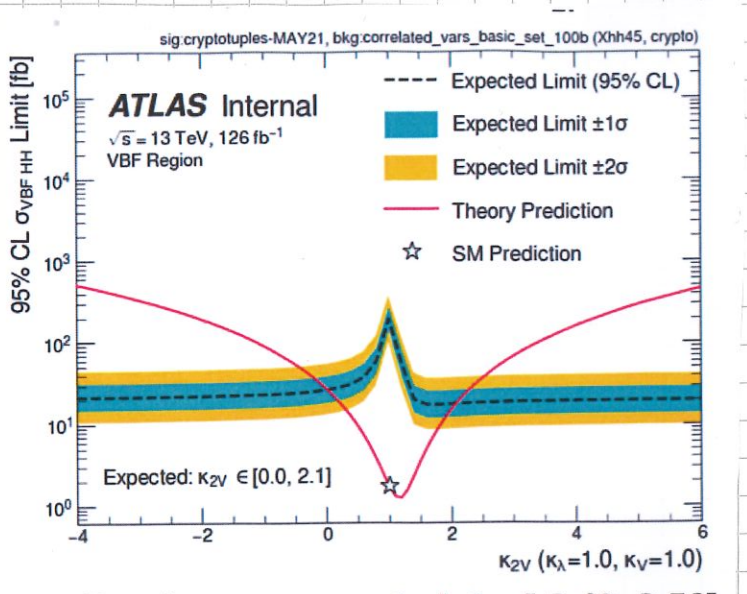
- perform the training
  ↳ train an ensemble of 100 individual networks (use median weight as a nominal event weight, derive an uncertainty from the spread of weights)
- evaluate the performance e.g. check pre and post reweighting distributions



- assign uncertainties: e.g. background shape uncertainty (systematic) derived from difference between CR1 and CR2 - trained networks
- finally: back to the statistical analysis profile likelihood ratio,

note: uncertainty on background estimation → most important Nuisance Parameter

- result (work in progress): expected κ_{2V} limits



Previous κ_{2V} constraints: [-0.43, 2.56]

Summary (take-home message)

- brief overview of the analysis steps
  - ↳ background estimation is complex, yet very important
- great example of how Machine Learning helps to improve results
- general note: di-Higgs is one of the key searches at the LHC, motivated by the need to explore the shape of Higgs potential ⇒ great motivation for HL-LHC and beyond