ttH(cc) analysis overview

Zoja Rokavec, Jakob Novak, Lenart Jerala, Miha Muškinja, Andraž Tomšič, Tadej Novak, Borut Paul Kerševan

28.9.2026, Krvavec

UNIVERZA V LJUBLJANI
FMF Fakulteta za matematiko in fiziko
Jožef Stefan Institute

ttH(cc) analysis overview

- So far, measured Higgs couplings agree with the SM, while the **charm Yukawa coupling** remains poorly constrained - currently $|\kappa_c| < 3.0$ at 95% CL

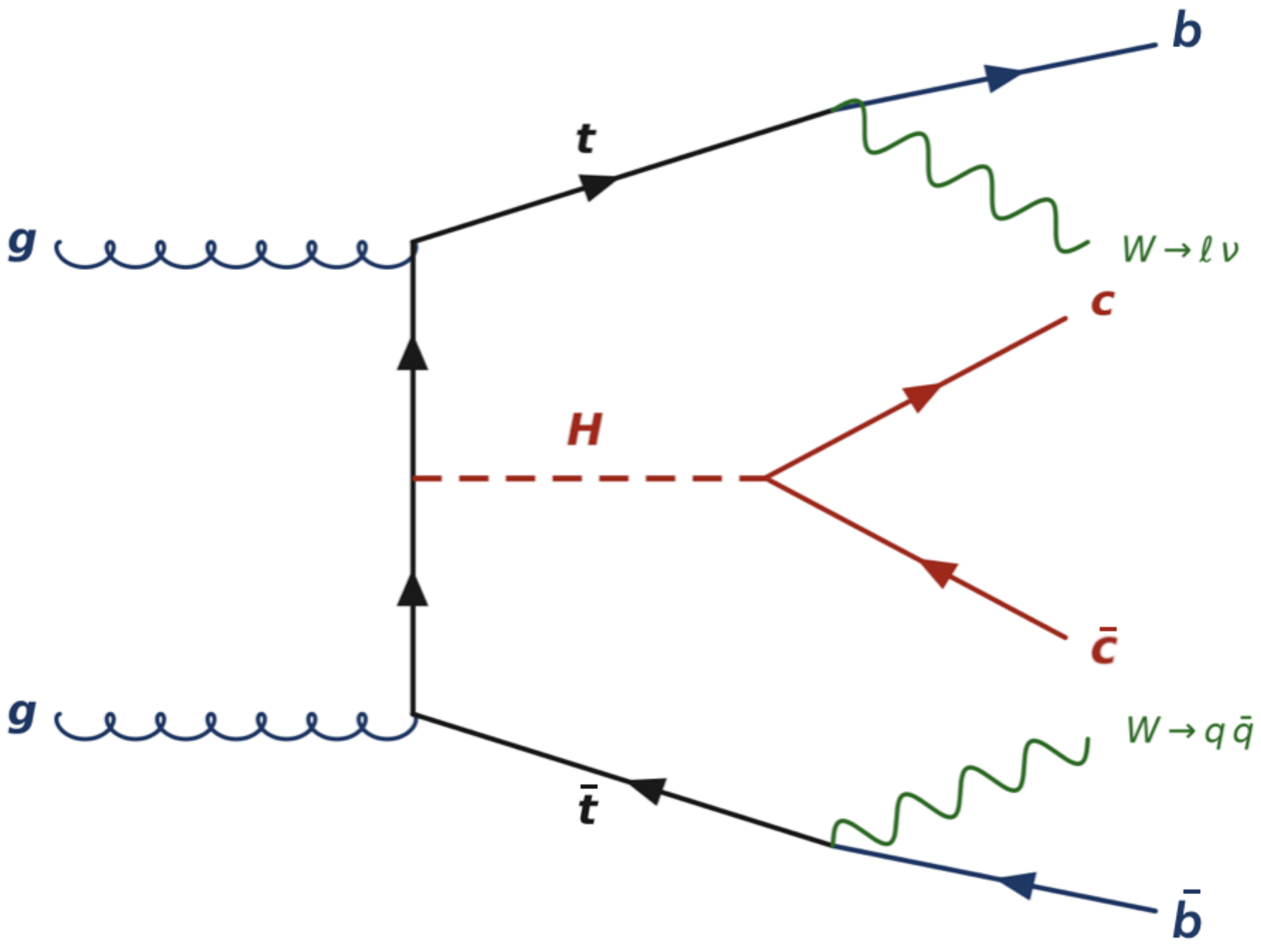
- Our goal is to measure/constrain $H \to c\bar{c}$ in the **ttH** production mode, complementary to existing VH(bb)/(cc) analysis

- CMS has reported the first ttH(cc) result - this analysis provides the corresponding ATLAS search

Signal signature

- 2 c-jets from the Higgs
- 2 b-jets from the two tops
- Number of leptons: 1, 2 or 0

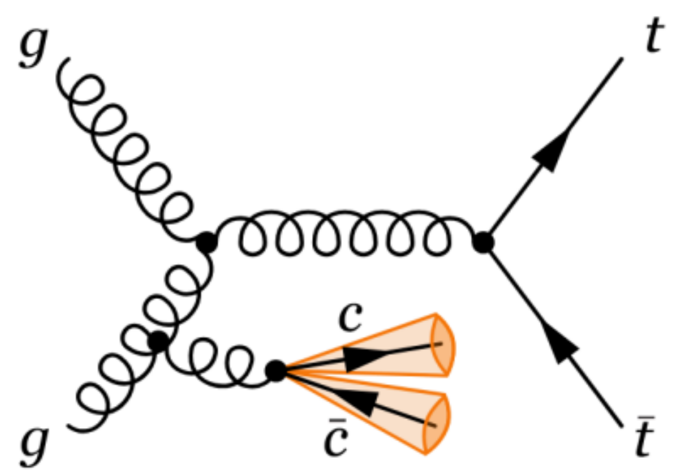
  (for 0L see next talk by Andraž)



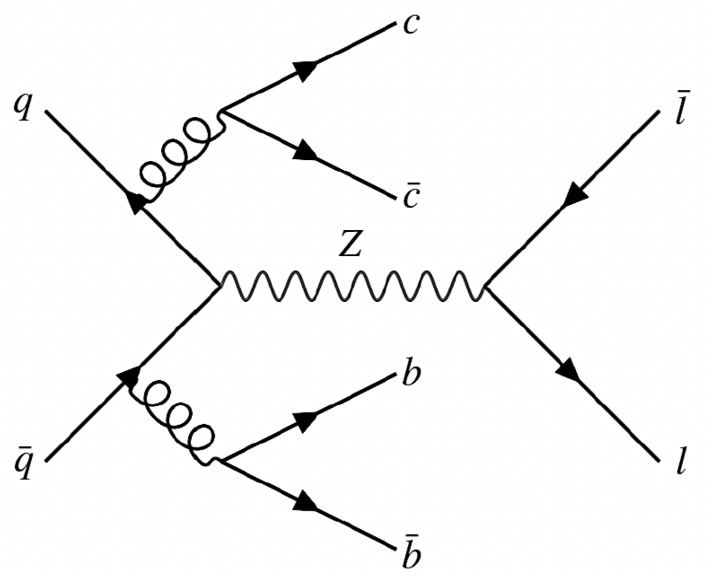
ttH(cc) - 1L channel

Challenges

- Rare signal: H(cc) branching ratio is small (~2.9%), about 20x below H(bb)
- The dominant tt+ heavy-flavour backgrounds contain the same jet flavours and very similar event topologies as the signal while they are orders of magnitude larger
- Charm tagging is less powerful than b-tagging, with larger mistag rates
- High jet multiplicity + leptons gives a complex event topology
- As a result, no simple cut-based selection isolates the signal



tt+jets



Z+jets - mimics ttH(cc) signal

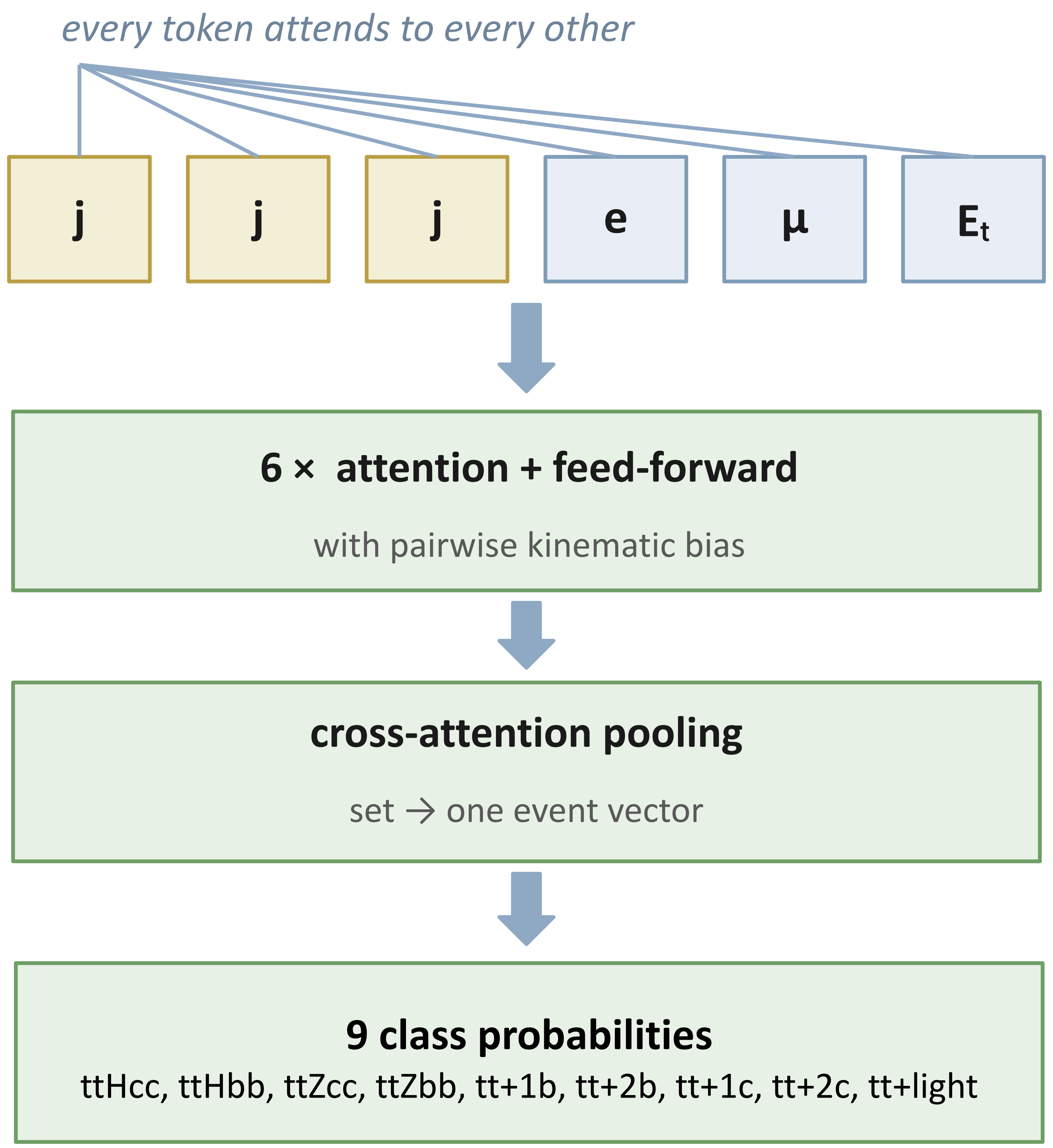
Our strategy

- Use multiclass ML to separate the different physics processes
- Build a discriminant from the ML outputs
- Use it to define background-like and signal-like regions for the fit

ML model: Particle Transformer

Based on work by Lenart Jerala and Jan Gavranovič

- **Input**: variable-length set of objects: up to 15 jets, leptons and MET – each with its own kinematic features
- Each objects features are encoded and fused into one token
- Each token attends to every other, with a **physics-informed pairwise bias** from jet-pair kinematics (ΔR, invariant mass, …) added to the attention
- Run 2 + Run 3 are trained together, with an era/source embedding
- We train **9 physics classes** (+QCD class in 0L) - the classifier outputs one probability $q(k|x)$ for each class

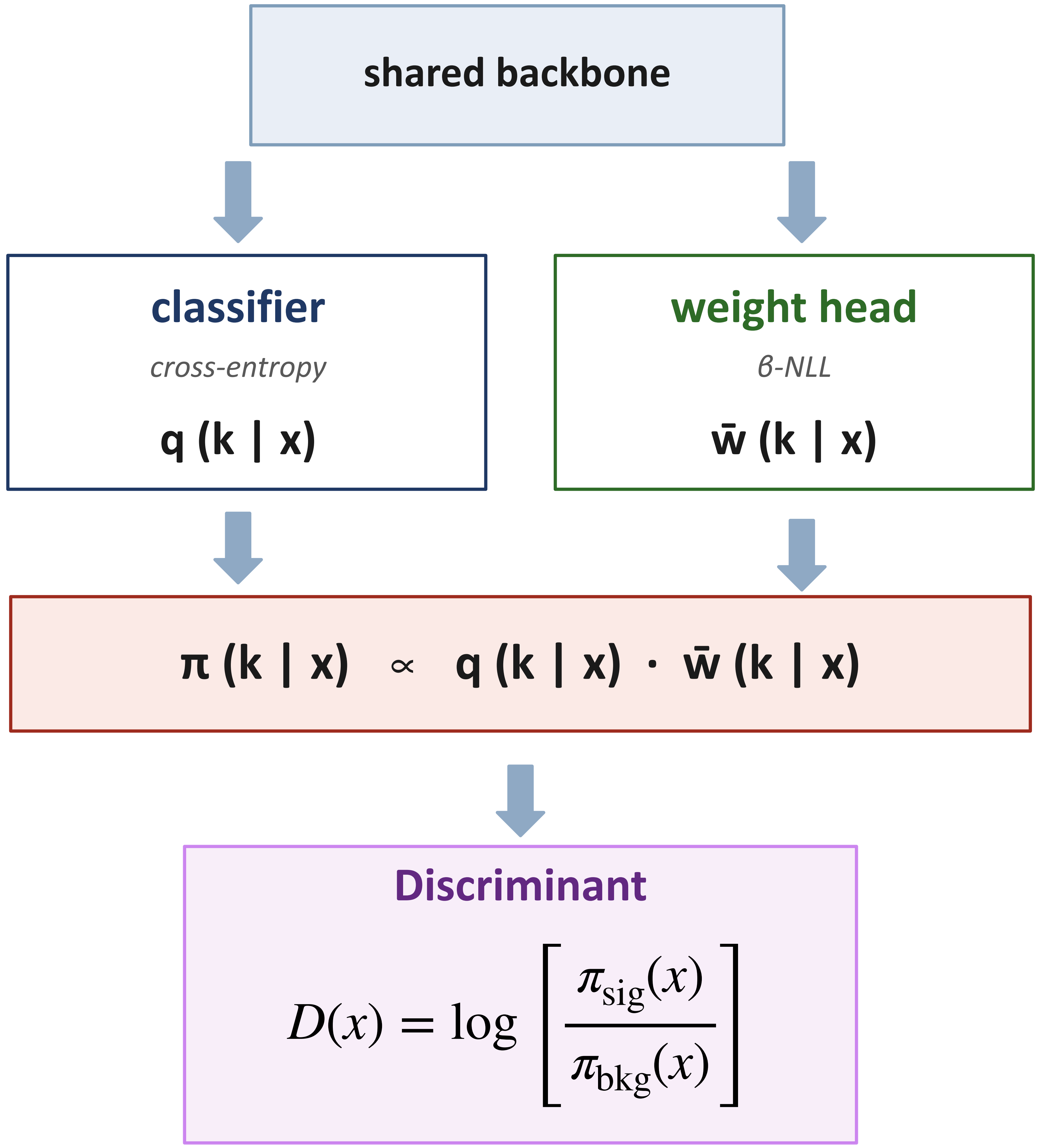


More info: SeeSawML Review

CMOK: training with MC weights

Based on work by Lenart Jerala

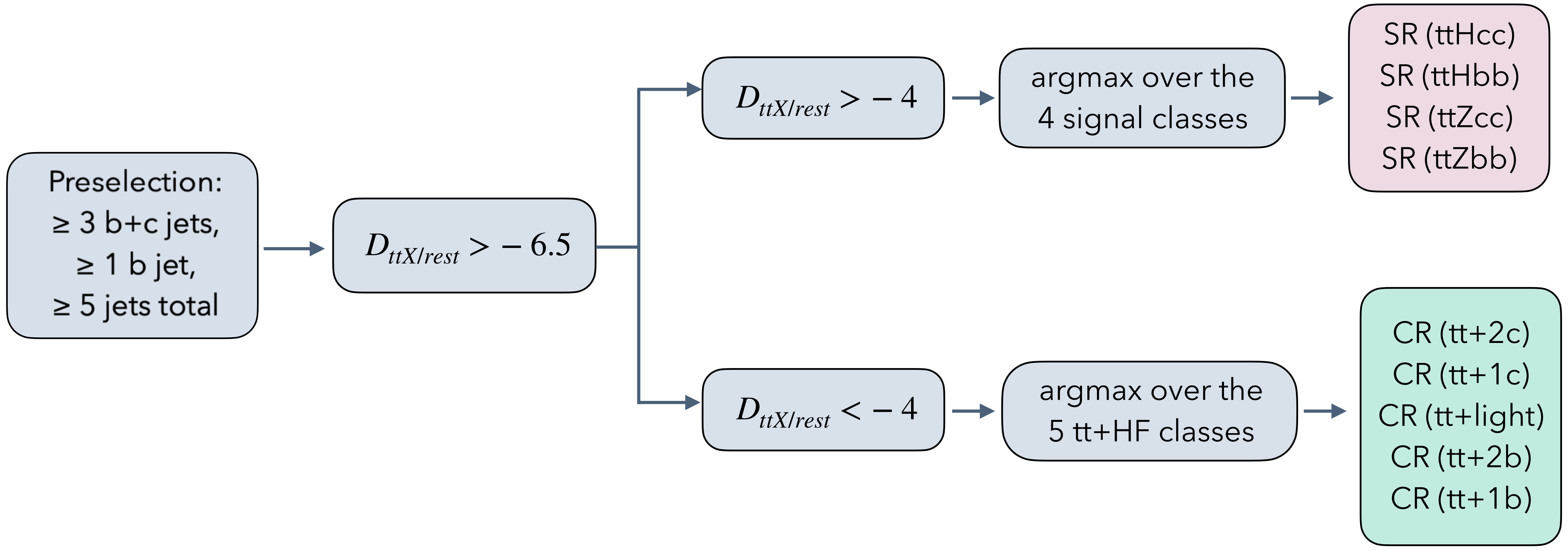
- Simulated events carry **weights,** so the physical signal and background densities are not just given by the raw number of generated events
- Our classifier just learns *q(k|x),* the posterior for the generated sample
- **CMOK**: predicts one conditional mean weight per class
- For a given event, it asks: if this event belonged to class *k*, what mean MC weight would events like this carry?
- The classifier and weight head are combined at inference (π) - this gives the **physical posterior** used to build the ttH(cc) discriminant
- **Neyman–Pearson**: the **optimal test statistic** for separating signal from background is the likelihood ratio $t(x) = p_S(x)/p_B(x)$



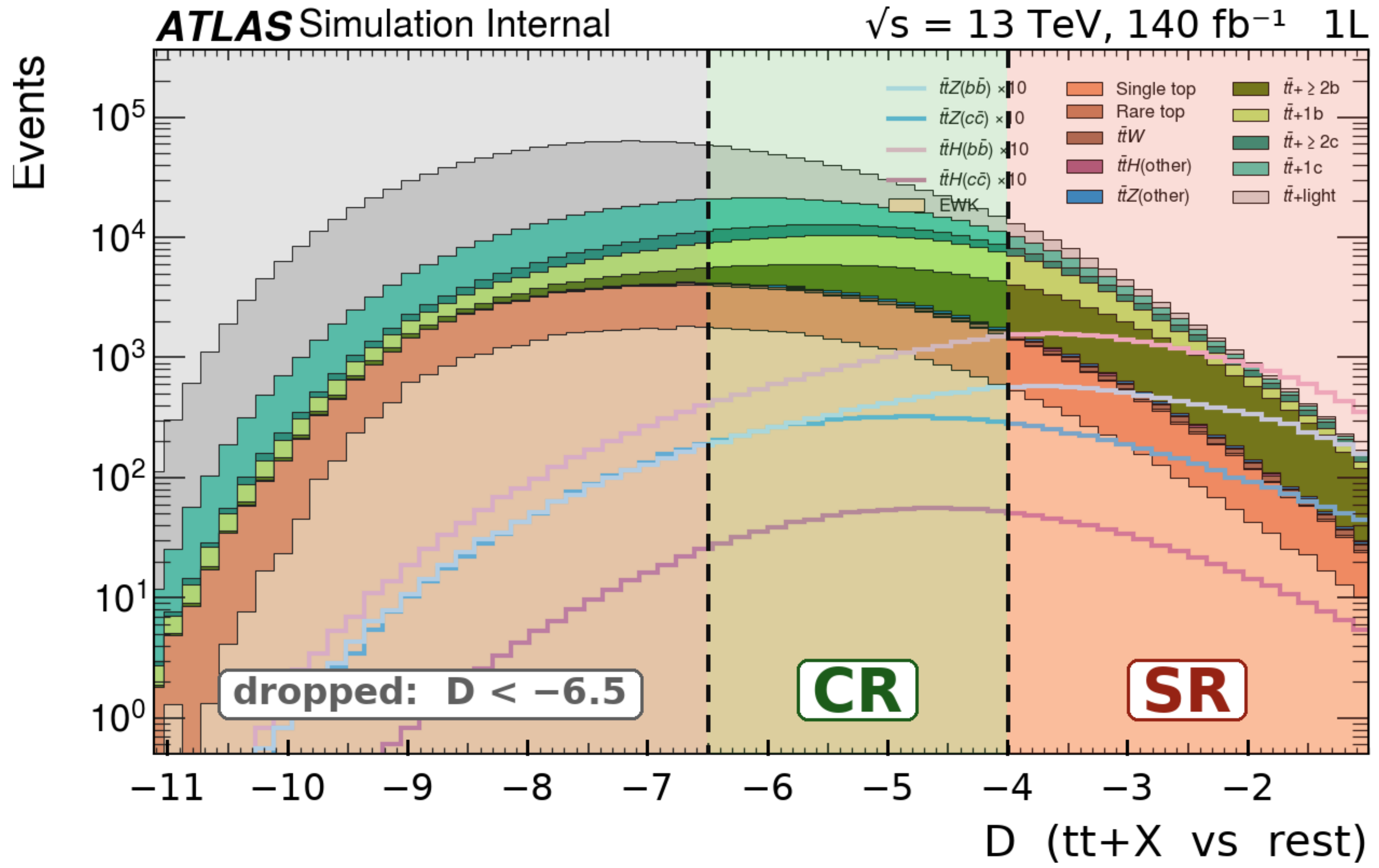
More info: CMOK

From ML outputs to fit regions - 1L channel

- The class posteriors are used to define the analysis regions
- We build a D_{ttX}, where the ttX posteriors are reweighted by the process cross sections - this prevents the region definition from being driven by the much larger ttH(bb) contribution
- D_{ttX} then separates ttX-like from tt+jets-like events
- Within each category, the largest class posterior assigns the event to its final ML region



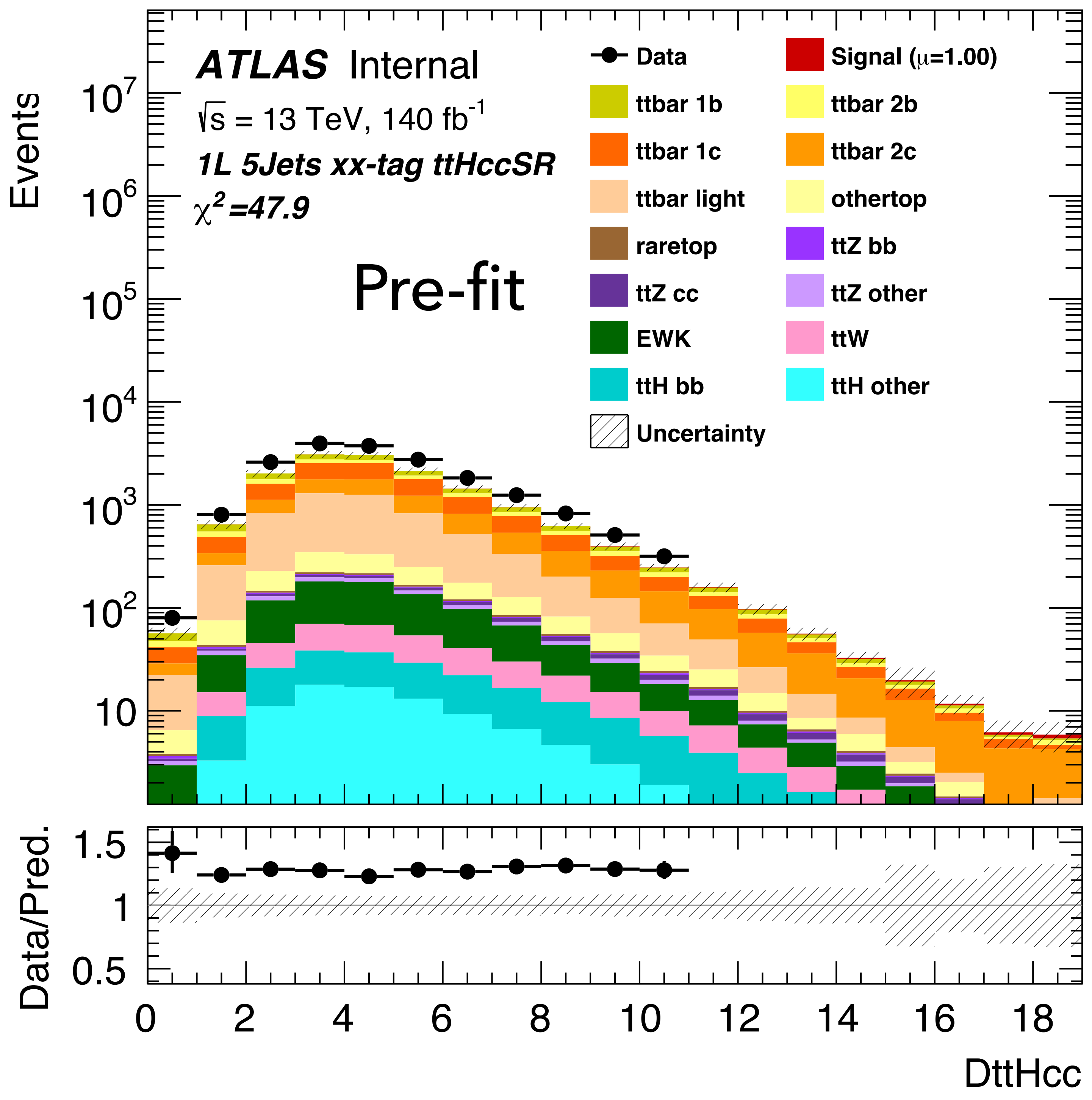
1L event selection scheme



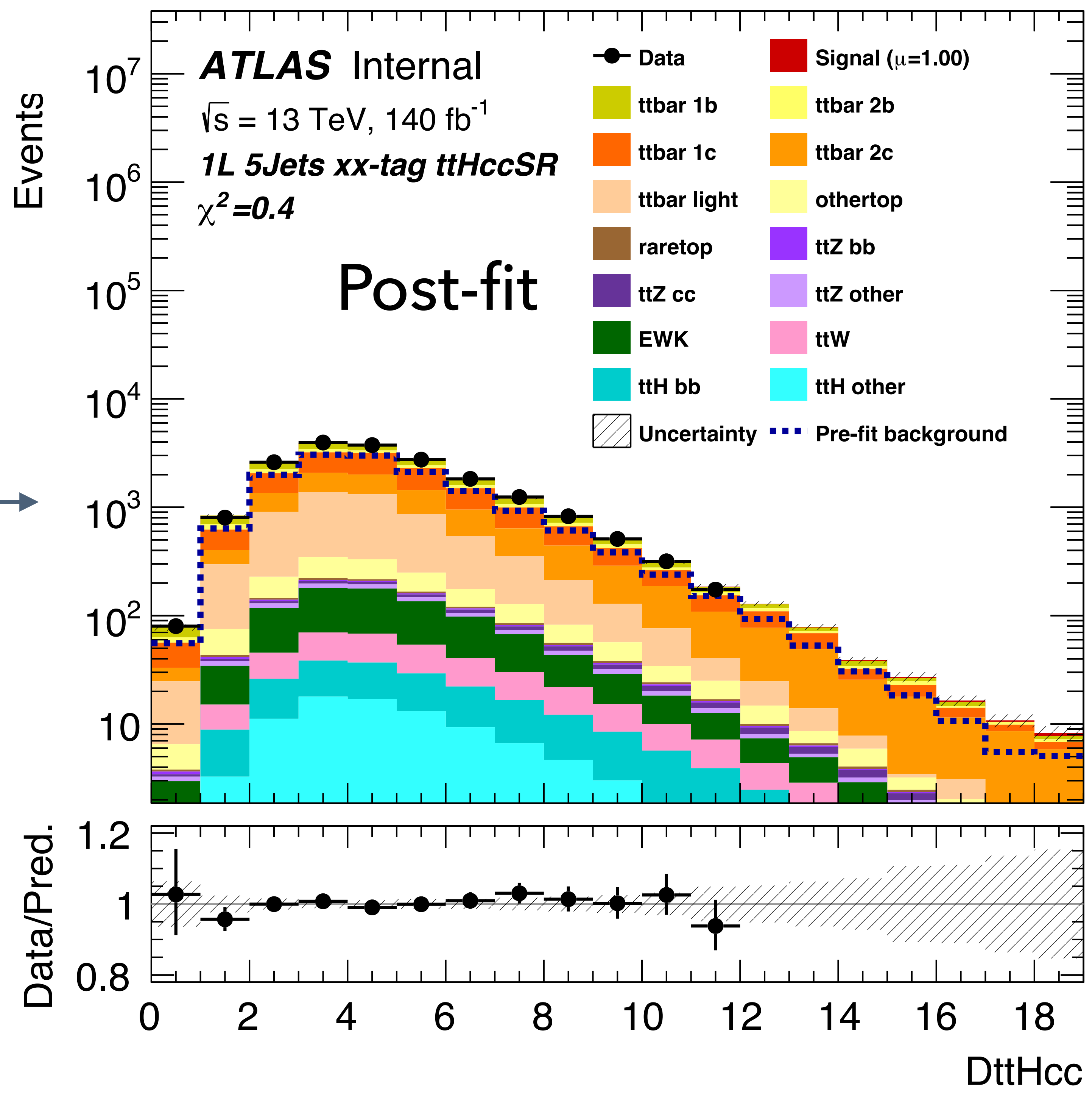
$$D_{ttX} = \log\left[\frac{\pi(ttHcc) + \pi(ttHbb) + \pi(ttZcc) + \pi(ttZbb)}{\pi(t\bar{t}+1c) + \pi(t\bar{t}+2c) + \pi(t\bar{t}+1b) + \pi(t\bar{t}+2b) + \pi(t\bar{t}+\text{light})}\right]$$

The likelihood fit - 1L channel

- The discriminant distributions from signal and control regions are fitted simultaneously
- Within each region, we build and bin its ML discriminant so the fit also uses the shape of the discriminant
- The dedicated control regions constrain the dominant tt+heavy-flavour components from data



→



Signal region ttH(cc), 1L, Run 2

Expected 95% CL upper limit 1L Run 2

Configuration	Expected limit on μ
Pre-fit	**4.46** +1.75/-1.25
Post-fit, no systematics	**5.13** +2.01/ -1.43
Post-fit with (theory) systematics	**7.28** +2.85 / −2.04

+ 60%

Current combined CMS limit: 7.8 x SM

Summary

- ttH(cc) gives direct sensitivity to $H \rightarrow c\bar{c}$, but the signal is rare and hidden under large tt+heavy-flavour backgrounds

- We use a multiclass ParticleTransformer to learn the full event topology and separate signal/background processes

- Current **1L Run 2 expected 95% CL limit is 7.28 x SM** including theory systematics

Next steps

- Finalise and validate the systematic uncertainties in the fit
- Optimise the ML regions/binning
- Move toward the final combined ttH(cc) result

Thank you!