Discovering CP-Sensitive Observables with Symbolic Regression

Akanksha Bhardwaj

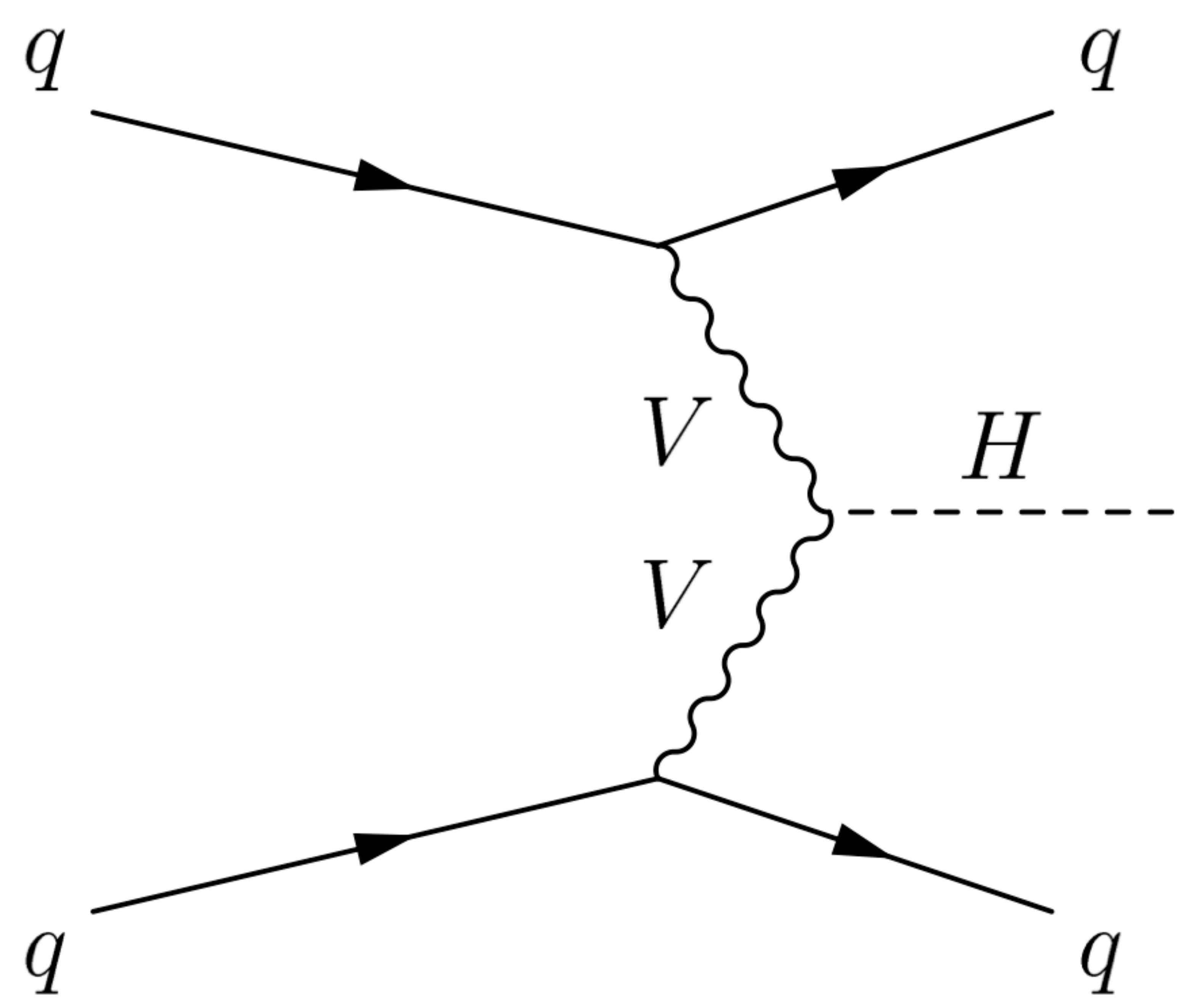
IIT PALAKKAD

ML4HEP-V4 School
Workshop

Interference, Angles, and Symbolic Regression

In Higgs plus two-jet production, CP-odd EFT effects appear mainly through interference between the Standard Model amplitude and a dimension-six CP-odd amplitude

- Goal: use symbolic regression to discover a compact analytic observable that predicts the sign of the interference
- We seek an observable that is not only accurate, but also interpretable and physically meaningful.
- Next Steps during workshop: VBF variables, CP properties, asymmetries, NN comparison, symmetry-constrained search.



Standard Model Effective Field Theory

- Lack of experimental evidence of new physics indicate a mass gap between SM and BSM scales
- SMEFT Lagrangian $\mathscr{L} = \mathscr{L}_{\rm SM} + \sum_i \frac{c_i}{\Lambda^2}\tilde{\mathcal{O}}_i\,,$

$$d\sigma = d\sigma_{\rm SM} + \underbrace{\frac{C_i}{\Lambda^2} d\sigma_i}_{\text{Interference}} + \underbrace{\frac{C_i C_j}{\Lambda^4} d\sigma_{ij}}_{\text{Cross terms}}$$

- Any differential cross section follows:
- Operators Introduce new source of CP violation $\mathcal{O}_{\Phi\widetilde{W}} = \frac{c_{\Phi\widetilde{W}}}{\Lambda^2}\Phi^\dagger\Phi\, W^{i\,\mu\nu}\widetilde{W}^i_{\mu\nu}$
- New physics effect can be seen as deformation of these distribution

Observable effects at LHC

$$|\mathcal{M}|^2 = |\mathcal{M}_{\rm SM}|^2 + \frac{c_i}{\Lambda^2}\, 2\, \boxed{\Re\left[\mathcal{M}_{\rm SM}\mathcal{M}^*_{{\rm d6},i}\right]} + \frac{c_i\, c_j}{\Lambda^4}\mathcal{M}_{{\rm d6},i}\mathcal{M}^*_{{\rm d6},j}\,,$$

Leading order deviation from SM comes from interference terms

$$\text{sign\_w} \sim \text{sign}\left(d\sigma_{\rm int}\right).$$

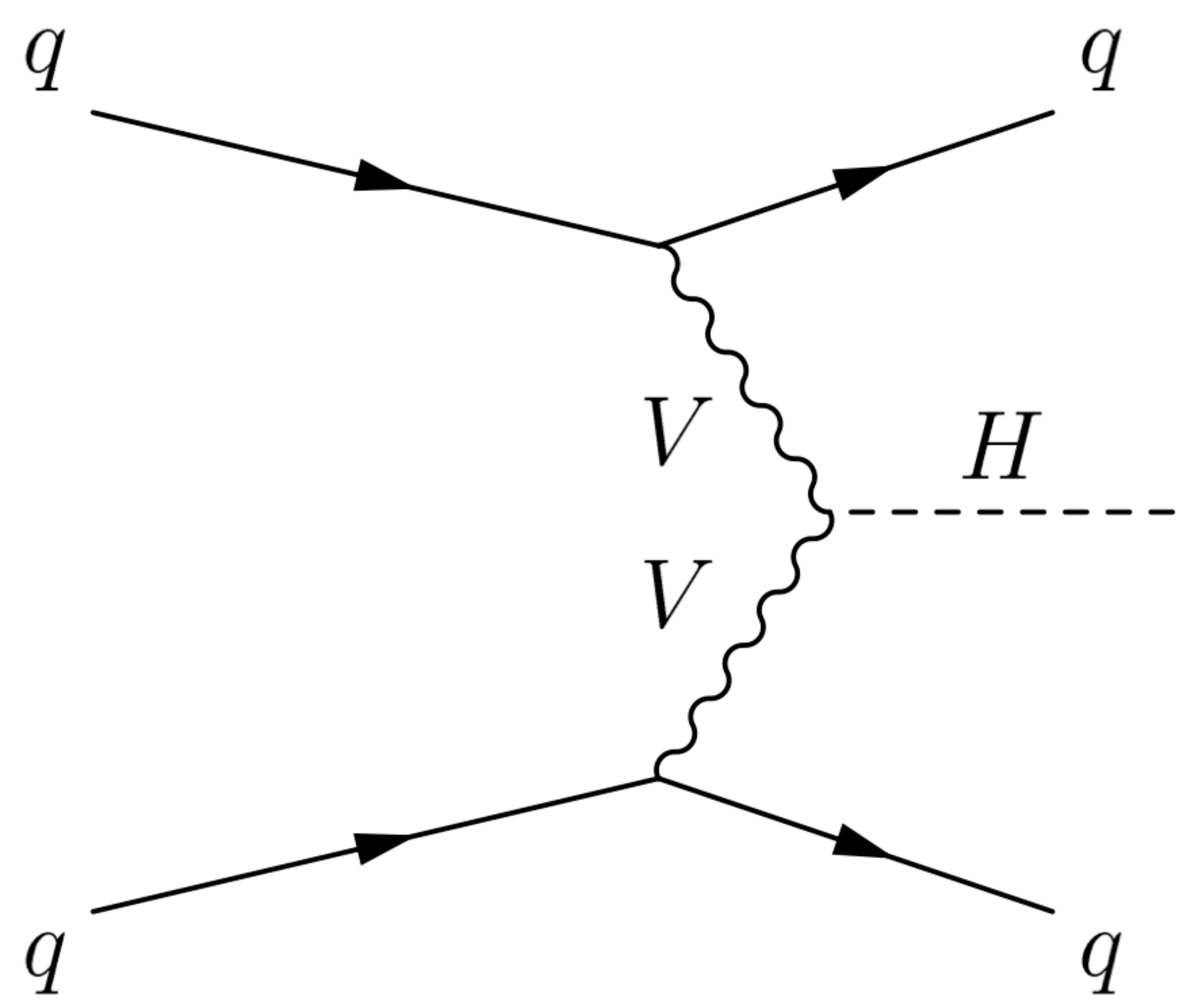
This interference changes sign across phase space under the CP transformation.

Negative weighted event is a phase-space point where the **interference contribution** to the differential cross section is negative, so it must cancel positive regions when integrated.

Interference, Angles, and Symbolic Regression

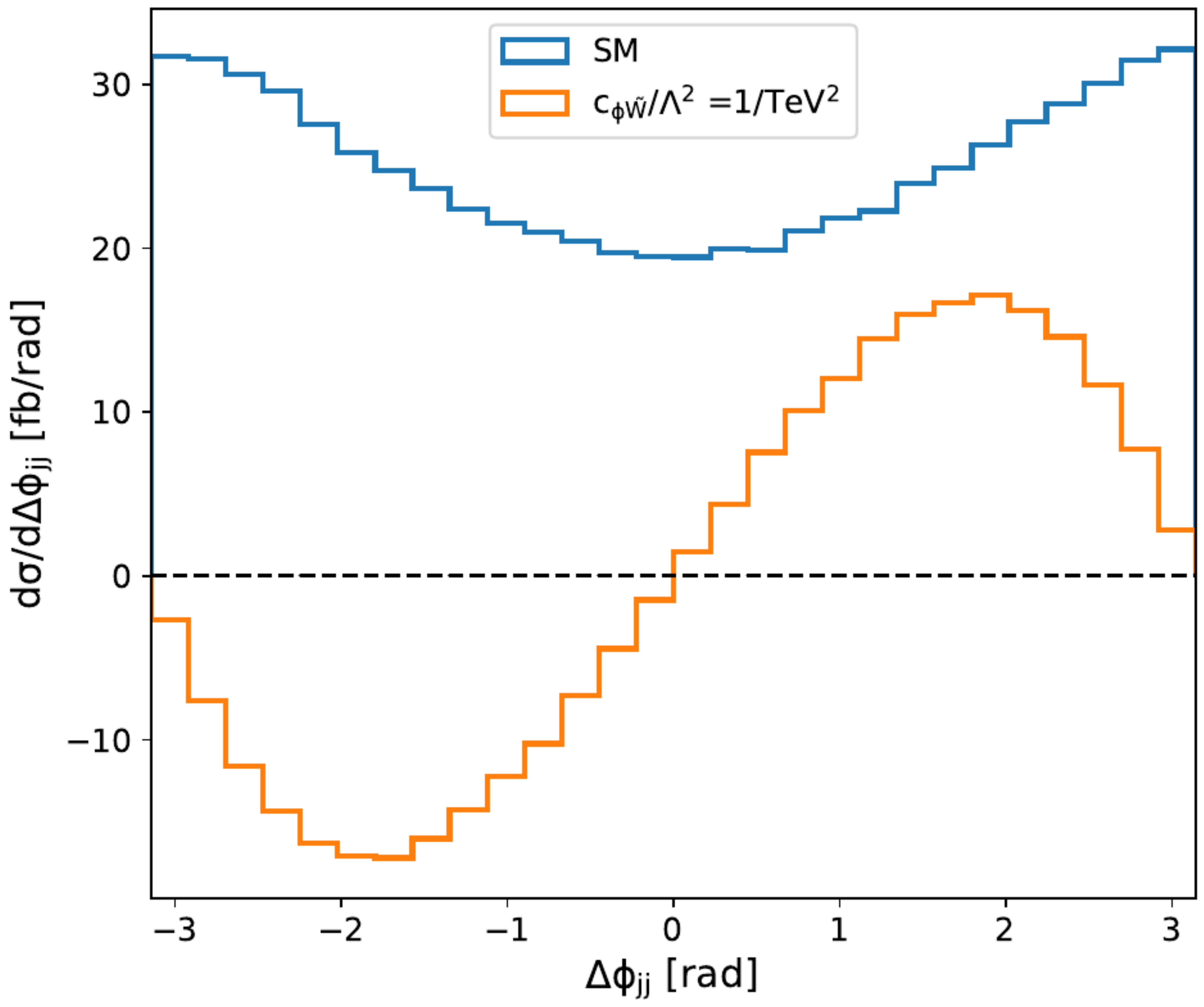
What is there in the data file ?

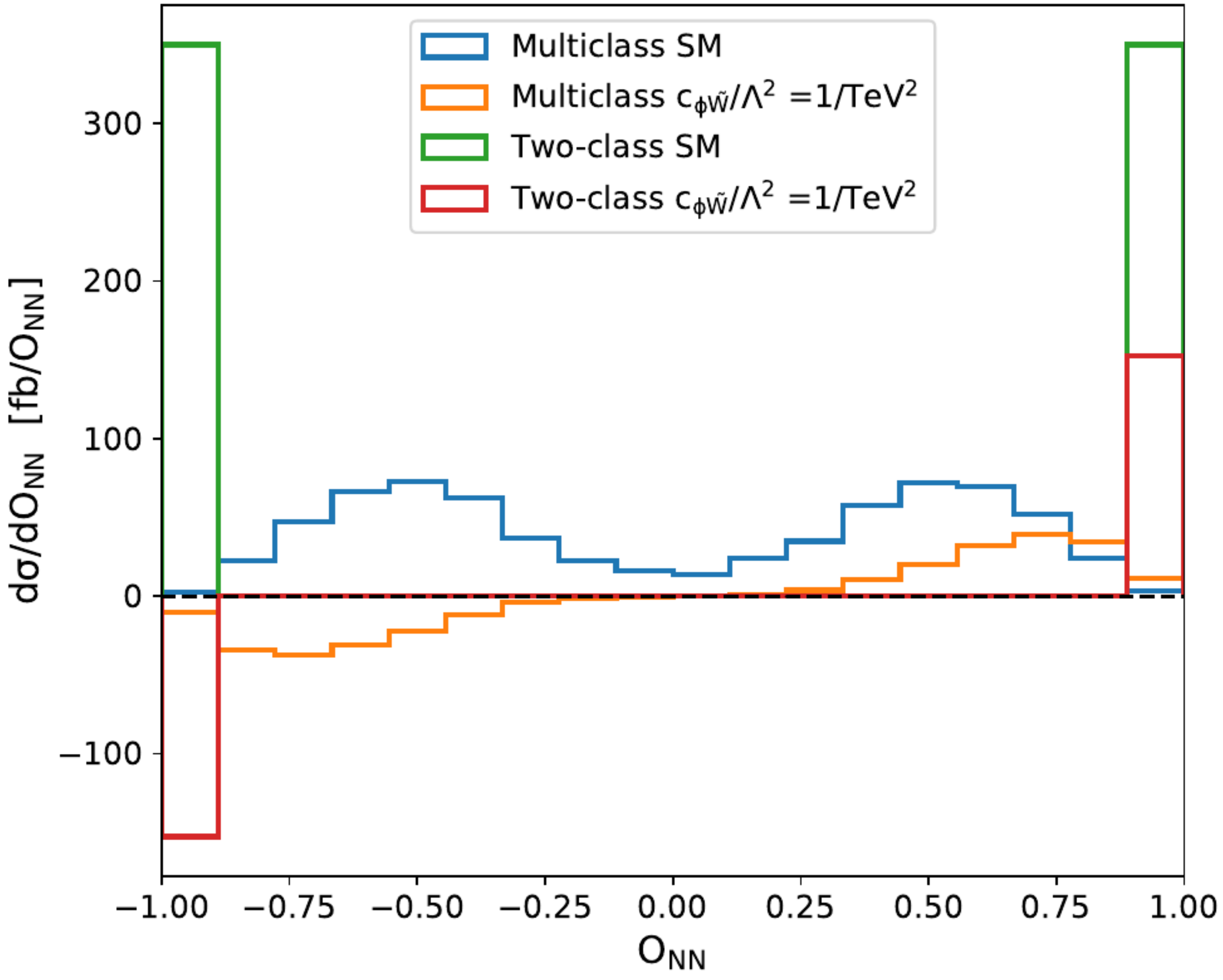
- Each event has kinematic information for two jets and the Higgs boson: j_1, j_2, h
- Input variables: p_{T,j_1}, η_{j_1}, ϕ_{j_1}, p_{T,j_2}, η_{j_2}, ϕ_{j_2}, $p_{T,h}$, η_h, ϕ_h
- The target lable is sign_w $= +1$ or -1.
- The symbolic regression task is to find $f(x)$ such that $\text{sign}\big(f(x)\big) \approx$ sign_w .



Core task

Find a symbolic observable f(x) whose sign predicts sign_w.
Compare f = Δϕjj, f = sin(Δϕjj), and PySR-discovered f.





[AB, Christoph Englert, Robert Hankache, Andrew D. Pilkington,. PLB`23]

Interference, Angles, and Symbolic Regression

- Level 0: Rediscover the known angular variable $f(x) \sim \Delta\phi_{jj}$ or $\mathrm{f(x)} \sim \sin(\Delta\phi_{\mathrm{jj}})$
- Level 1: Do not use $\Delta\phi_{jj}$ directly, Use only the raw variables (p_T, η, ϕ) for the two jets and the Higgs.
- **Question:** Can PySR reconstruct the relevant angular information from raw φ variables?

- Once we have a symbolic score f(x), we can build an asymmetry:

$$A_f = \frac{N(f > 0) - N(f < 0)}{N(f > 0) + N(f < 0)}$$

- Then Compare different observables: $A_{\Delta\phi}$, $A_{\sin\Delta\phi}$, A_{PySR}
- Which observable gives the cleanest separation between positive and negative interference regions?

Interference, Angles, and Symbolic Regression

- Does the CP-sensitive observable work equally well everywhere?
- Test VBF-like regions using cuts such as $m_{jj} > 400$ GeV, $|\Delta\eta_{jj}| > 2.5, \eta_{j_1}\eta_{j_2} < 0$.
- Then compare PySR before and after cuts.

Questions to study:

- Does the accuracy improve in the VBF-like region?
- Does the symbolic formula become simpler?
- Does PySR rely more strongly on angular variables after VBF cuts?
- Does the sign of the interference become cleaner in this region?

Things to do during workshop

- **VBF kinematics:** Compute mjj, $\Delta\eta jj$, ΔRjj, Higgs centrality. Study performance with VBF cuts.
- **CP structure:** Classify variables into CP-even and CP-odd. Check whether the PySR formula has correct CP behavior.
- **Optimal observable interpretation:** Relate sign_w to the SMEFT interference term and explain PySR as an approximation to the optimal observable.
- **Asymmetry construction:** Turn the symbolic score into an asymmetry and compare different observables.
- **Black-box versus symbolic:** Train a NN/BDT and then compare with PySR in terms of accuracy and interpretability.
- **Symmetry-constrained PySR:** Force the observable to be $\sin(\Delta\phi jj) \times$ CP-even function and let PySR learn the CP-even modulation.

Thank you