

# XCC MC production plan for BDT training

Event-selection outputs across initial-states

G. Otero y Garzón

XCC Meeting - 28 August 2026

# Goal

- Extend the previous  $\gamma\gamma$  background study to all available initial-states: gg, eg, ge, ep, pe, ee, gp, pg, pp.
- Identify samples that survive the current event selection (designed for  $HH \rightarrow bbbb$ ).
- Group physically similar samples into approximately 10–20 background types for BDT training.
- Estimate additional MC needed per sample, with special treatment for brute-force requests above a few million events.

## Column definitions for the following tables

**generated** Events already produced for a state/sample. Here this corresponds to the expected pre-selection count for the assumed luminosity.

**selected** Events passing the current event selection.

**eff** Selection efficiency,  $N_{\text{selected}}/N_{\text{generated}}$ .

**group\_target\_selected** Desired selected MC statistics for the entire background type.

**target\_selected\_MC** Share assigned to an individual sample, proportional to its current selected contribution in the group.

**estimated\_total\_generated** Brute-force inclusive generation needed if current efficiency is unchanged.

**additional\_to\_generate** Brute-force additional events. For special samples this is a diagnostic, not a production request.

## Global status

| state | samples | non-zero | selected | overall eff |
|-------|---------|----------|----------|-------------|
| ee    | 60      | 4        | 1,144    | 1.44e-05    |
| eg    | 13      | 1        | 315      | 6.56e-06    |
| ep    | 105     | 9        | 1,112    | 1.75e-05    |
| ge    | 12      | 1        | 329      | 6.88e-06    |
| gg    | 72      | 28       | 40,984   | 4.89e-04    |
| gp    | 18      | 1        | 148      | 3.79e-06    |
| pe    | 84      | 8        | 1,003    | 1.71e-05    |
| pg    | 24      | 4        | 123      | 2.80e-06    |
| pp    | 62      | 2        | 5        | 1.92e-07    |

### Key point

Only 58 of 451 state/sample combinations have non-zero selected events. The production problem is therefore concentrated in a small subset of samples.

## Shape-similarity checks for grouping

| comparison                       | N1/N2      | median JS | JS to signal | recommendation |
|----------------------------------|------------|-----------|--------------|----------------|
| eg vs ge, AA→bb                  | 315/329    | 0.134     | 0.503        | YES            |
| gp vs pg, AA→bb                  | 148/119    | 0.222     | 0.558        | YES            |
| e-gamma vs positron-gamma, AA→bb | 644/267    | 0.155     | 0.520        | YES            |
| gg AA→bb vs gamma-lepton AA→bb   | 24,087/911 | 0.090     | 0.480        | YES            |
| ep vs pe, ffH                    | 423/373    | 0.137     | 0.260        | YES            |
| ep vs pe, eebb                   | 199/184    | 0.180     | 0.543        | YES            |
| ep vs pe, mumu bb                | 167/130    | 0.204     | 0.508        | YES            |
| ep vs pe, tautau bb              | 208/203    | 0.174     | 0.549        | YES            |
| ep/pe: eebb vs mumu bb           | 383/297    | 0.149     | 0.520        | YES            |
| ep/pe: eebb vs tautau bb         | 383/411    | 0.155     | 0.530        | YES            |
| ep/pe: mumu bb vs tautau bb      | 297/411    | 0.197     | 0.520        | YES            |
| ee: eA vs Ae assoc bb            | 443/409    | 0.131     | 0.570        | YES            |
| ee: eA vs ee_eebB                | 443/278    | 0.176     | 0.530        | YES            |
| ee: Ae vs ee_eebB                | 409/278    | 0.169     | 0.530        | YES            |
| gg WZ vs ep/pe WZ                | 112/106    | 0.362     | 0.400        | NO             |

- Low JS between backgrounds, compared with JS to signal, supports grouping.
- Do not merge gg WZ and ep/pe WZ for now: their mutual distance is comparable to their distance to signal.

## Proposed background groups

| group                     | states      | selected | target sel. MC | add. gen.     |
|---------------------------|-------------|----------|----------------|---------------|
| bb                        | gg          | 24,087   | 100,000        | 9,918,904     |
| llbb                      | gg          | 11,224   | 75,000         | 16,520,647    |
| AZ                        | gg          | 5,284    | 75,000         | 3,718,635     |
| same-sign lepton assoc bb | ee,pp       | 1,135    | 26,100         | 28,968,829    |
| e+e- llbb                 | ep,pe       | 1,091    | 50,400         | 11,041,771    |
| gamma-lepton bb           | eg,ge,gp,pg | 911      | 25,200         | 328,195,014   |
| e+e- ffH                  | ep,pe       | 796      | 25,100         | 375,717       |
| WZ                        | gg          | 112      | 19,000         | 5,917,082     |
| e+e- WZ                   | ep,pe       | 106      | 10,100         | 1,773,209     |
| VVbb                      | gg          | 105      | 25,000         | 6,675,938     |
| e+e- assoc bb             | ep,pe       | 83       | 10,100         | 20,625,348    |
| qq/mistag                 | gg          | 33       | 4,000          | 1,422,925,093 |
| e+e- VVbb                 | ep          | 29       | 10,000         | 1,648,209     |
| nu+jets                   | gg          | 22       | 7,000          | 7,643,997,778 |
| ee secondary AA→bb        | ee          | 14       | 2,000          | 284,120,142   |

## Non-zero selected samples outside gg: lepton/gamma states

| state | sample        | selected | eff      | group                     | add. gen.     |
|-------|---------------|----------|----------|---------------------------|---------------|
| ee    | eA_ebB_eqq    | 443      | 8.00e-04 | same-sign lepton assoc bb | 11,701,987    |
| ee    | Ae_bBe_qqe    | 409      | 7.39e-04 | same-sign lepton assoc bb | 11,768,201    |
| ee    | ee_eebB       | 278      | 1.95e-03 | same-sign lepton assoc bb | 3,035,505     |
| ee    | AA_bb_mb2.8   | 14       | 6.99e-06 | ee secondary AA→bb        | 284,120,142   |
| eg    | AA_bb_mb2.8   | 315      | 7.89e-05 | gamma-lepton bb           | 106,228,319   |
| ge    | AA_bb_mb2.8   | 329      | 8.56e-05 | gamma-lepton bb           | 102,433,726   |
| gp    | AA_bb_mb2.8   | 148      | 6.58e-05 | gamma-lepton bb           | 60,089,710    |
| pg    | AA_bb_mb2.8   | 119      | 5.35e-05 | gamma-lepton bb           | 59,443,259    |
| pg    | pA_vWZ        | 2        | 4.54e-03 | gamma-p rare EW           | 308,559       |
| pg    | AA_dd_md0.005 | 1        | 2.98e-07 | gamma-p qq mistag         | 3,351,430,215 |
| pg    | pA_pZZ        | 1        | 6.76e-03 | gamma-p rare EW           | 103,852       |
| pp    | Ap_bBp_qqp    | 4        | 1.03e-03 | same-sign lepton assoc bb | 484,102       |
| pp    | pA_pbB_pqq    | 1        | 2.52e-04 | same-sign lepton assoc bb | 1,979,034     |

## Non-zero selected samples outside gg: ep/pe

| state | sample      | selected | eff      | group                | add. gen.   |
|-------|-------------|----------|----------|----------------------|-------------|
| ep    | ep_ffH      | 423      | 6.44e-02 | e+e- ffH             | 200,434     |
| ep    | ep_taTabB   | 208      | 8.15e-03 | e+e- llbb            | 1,153,469   |
| ep    | ep_eEbB     | 199      | 2.59e-03 | e+e- llbb            | 3,471,257   |
| ep    | ep_muMubB   | 167      | 6.47e-03 | e+e- llbb            | 1,164,199   |
| ep    | ep_WZvE     | 49       | 5.07e-03 | e+e- WZ              | 918,326     |
| ep    | eA_ebB_eqq  | 35       | 4.07e-04 | e+e- assoc bb        | 10,484,964  |
| ep    | ep_vvbB     | 29       | 6.05e-03 | e+e- VVbb            | 1,648,209   |
| ep    | ep_vtacS    | 1        | 5.96e-06 | e+e- rare/mistag     | 167,535,297 |
| ep    | AA_bb_mb2.8 | 1        | 1.64e-06 | e+e- secondary AA→bb | 304,435,909 |
| pe    | pe_ffH      | 373      | 6.52e-02 | e+e- ffH             | 175,283     |
| pe    | pe_taTabB   | 203      | 8.97e-03 | e+e- llbb            | 1,025,374   |
| pe    | pe_eEbB     | 184      | 2.63e-03 | e+e- llbb            | 3,159,112   |
| pe    | pe_muMubB   | 130      | 5.50e-03 | e+e- llbb            | 1,068,360   |
| pe    | pe_WZvE     | 57       | 6.25e-03 | e+e- WZ              | 854,883     |
| pe    | Ae_bBe_qqe  | 48       | 5.67e-04 | e+e- assoc bb        | 10,140,384  |

## Examples of direct additional production

| state | sample    | selected | eff      | group     | add. gen. |
|-------|-----------|----------|----------|-----------|-----------|
| pe    | pe_ffH    | 373      | 6.52e-02 | e+e- ffH  | 175,283   |
| ep    | ep_ffH    | 423      | 6.44e-02 | e+e- ffH  | 200,434   |
| pe    | pe_WZvE   | 57       | 6.25e-03 | e+e- WZ   | 854,883   |
| ep    | ep_WZvE   | 49       | 5.07e-03 | e+e- WZ   | 918,326   |
| pe    | pe_taTabB | 203      | 8.97e-03 | e+e- llbb | 1,025,374 |
| pe    | pe_muMubB | 130      | 5.50e-03 | e+e- llbb | 1,068,360 |
| ep    | ep_taTabB | 208      | 8.15e-03 | e+e- llbb | 1,153,469 |
| ep    | ep_muMubB | 167      | 6.47e-03 | e+e- llbb | 1,164,199 |

### Interpretation

These samples can be handled with standard oversampling and physical weights; the required additional generation is of order  $10^5$ – $10^6$  events.

## Samples requiring special treatment

| state | sample        | selected | eff      | group                     | brute-force add. gen. |
|-------|---------------|----------|----------|---------------------------|-----------------------|
| pg    | AA_dd_md0.005 | 1        | 2.98e-07 | gamma-p qq mistag         | 3,351,430,215         |
| ee    | AA_bb_mb2.8   | 14       | 6.99e-06 | ee secondary AA→bb        | 284,120,142           |
| eg    | AA_bb_mb2.8   | 315      | 7.89e-05 | gamma-lepton bb           | 106,228,319           |
| ge    | AA_bb_mb2.8   | 329      | 8.56e-05 | gamma-lepton bb           | 102,433,726           |
| gp    | AA_bb_mb2.8   | 148      | 6.58e-05 | gamma-lepton bb           | 60,089,710            |
| pg    | AA_bb_mb2.8   | 119      | 5.35e-05 | gamma-lepton bb           | 59,443,259            |
| ee    | Ae_bBe_qqe    | 409      | 7.39e-04 | same-sign lepton assoc bb | 11,768,201            |
| ee    | eA_ebB_eqq    | 443      | 8.00e-04 | same-sign lepton assoc bb | 11,701,987            |

### Do not generate these inclusively

Brute-force estimates above a few million events should trigger enriched/filtered generation or tag-importance sampling, not inclusive production.

## Recommended strategy for low-efficiency samples

### Kinematic inefficiency

Use generator-level filtering/enrichment looser than the reco selection. Keep the filter efficiency and physics weight.

### Mistag-driven inefficiency

Use tag importance sampling: enhance mistag probabilities in the training production and reweight tag patterns back to nominal.

### Tiny physical yield

Merge with a similar group or keep for evaluation only. The BDT does not need a separate class for every negligible contribution.

### Training vs evaluation

Balance samples for training; use physics weights for final yield/significance evaluation.

## Proposed next steps

- ❶ Freeze a first set of background types: around 12–14 groups, within the desired 10–20 range.
- ❷ Produce direct-oversampling samples for  $ffH$ ,  $llbb$ ,  $WZ$  and  $VVbb$ .
- ❸ Define filtered/enriched generation for secondary  $AA \rightarrow bb$  and gamma-lepton  $AA \rightarrow bb$  samples.
- ❹ Define tag-importance-sampling validation for mistag-dominated light/charm backgrounds.
- ❺ Train the first BDT using balanced training weights; evaluate with physical weights and the original state/sample metadata.

## Take-away

The additional MC plan is manageable if the target is set by background type, not by individual sample.

Ultra-low efficiencies should be handled with enriched generation or importance sampling, not brute-force inclusive production.