ZHH production at 3 TeV CLIC

Matthias Weber (CERN)

HHZ at 3 TeV

At 3 TeV **WW fusion dominating production mode** of double H, ee→HHνν

Information from 3 TeV HHZ production helpful in behaviour of EFT fits on double H events

HHZ production at 3 TeV studied at CLIC in fast simulation samples so far, first study of this process using **full simulation** and **new detector model CLICdet**

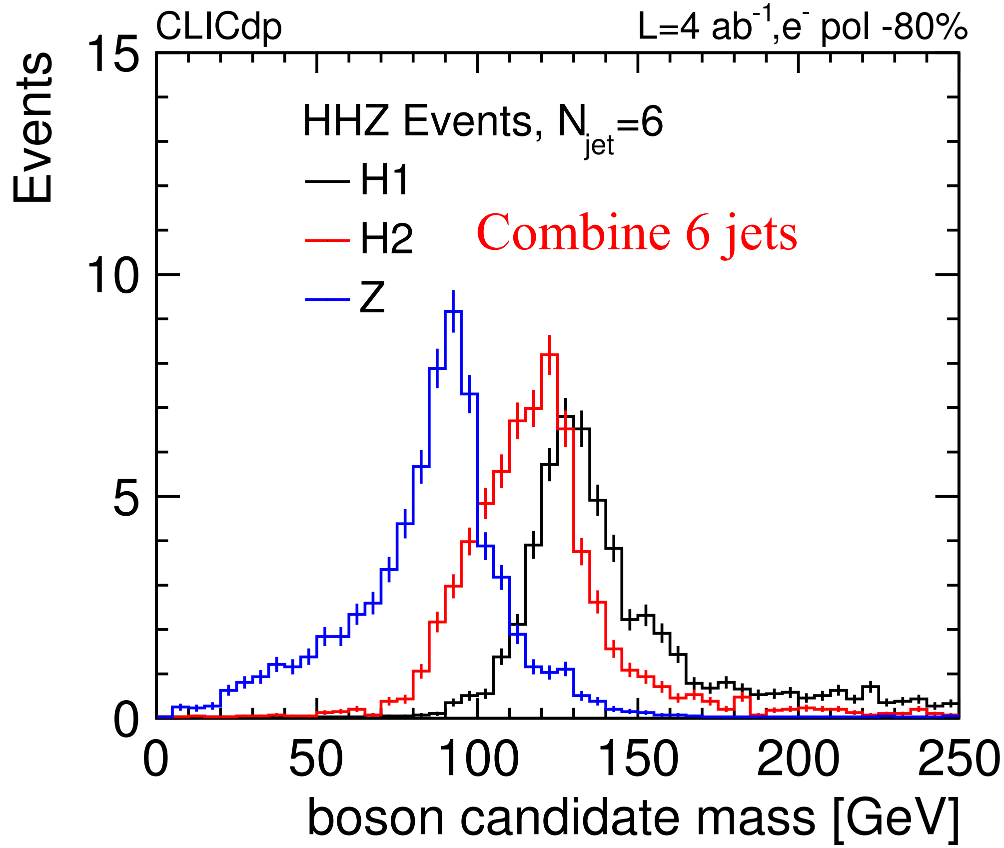
Concentrate on HH Z →bb bb qq:

Signal sample: HH qq, concentrate on four b-state **HH qq→bb bb qq**
For -80 % polarisation: cross section: 4.18e-2 fb^{-1}
→ 167 events in total
→ 68 events in desired phase space of bb bb qq
For +80 % polarisation: cross section: 2.30e-2 fb^{-1}
→ 29 events in total
→ 12 events in bb bb qq final state

Baseline: combine 6 jets

Use jets with VLC algorithm R=1.10, $\beta=\gamma=1.0$, run in exclusive mode with 6 jets

Combine 6 jets into 3, minimizing sum= $(rj(x)-m_H)^2 + (rj(y)-m_H)^2 + (rj(z)-m_Z)^2$



Using 6 jets leads to 3 clear peaks with mass peaks close to Z and H masses. For jet with largest mass clear shift to higher values.

Background composition

Sample Pol (e^-) -80 %	Events	Cross-section [fb]	Produced Events
HHqq	167	4.18e-2	9600
HHqq→bbbbqq	68	1.70e-2	3948
Hqq	15320	3.83	115174
ee→qq	5.07 M	1269	1.56 M
ee→qqqq	3.61 M	902	1.9 M
ee→qqqqqq			2.41 M

qqqqqq is mixture of several samples, containing tri-bosons as well as ttbar, but not ZZH and WWH

WWH: 4.116 fb →16464 evts
ZZH: 1.394e-1 fb →558 evts

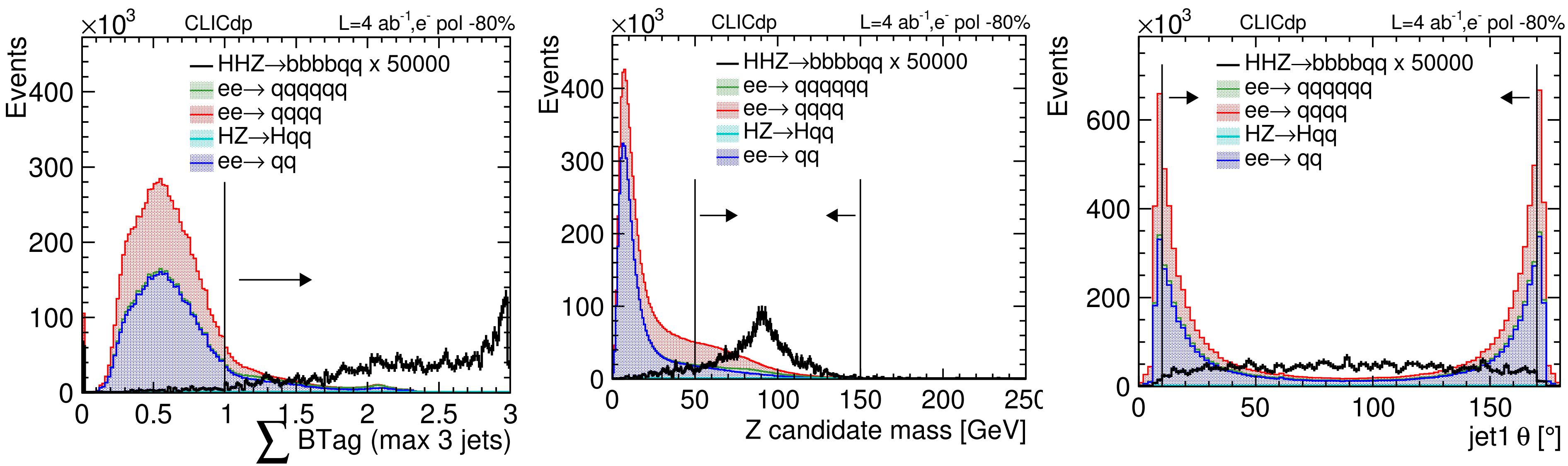
Sample Pol (e^-) +80 %	Events	Cross-section [fb]	Produced Events
HHqq	29	2.898e-2	9552
Hqq	2670	2.67	29034
ee→qq	786 000	786	388 190
ee→qqqq	120 000	120	478 995
ee→qqqqqq			589440

WWH: 4.176e-1 fb →418 evts
ZZH: 7.1645e-2 fb →72 evts

Without Preselection BDT tuning in TMVA fails

Running BDT on full statistics using TMVA reaches a significance of below 1 σ
→try retuning after pre-selection

final state	Events -80%	Evts after cut -80%	Efficiency -80%, in [%]	Events +80%	Evts after cut +80%	Efficiency +80% [in%]
$HHq\bar{q}$, both $H \rightarrow b\bar{b}$	69	54	78	12	9.4	78
$HHq\bar{q}$, all H decays	167	110	66	29	19	66
$Hq\bar{q}$, all H	15300	3370	22	2670	590	22
$q\bar{q}$	5 070 000	61 100	1.2	787 000	5820	0.74
$q\bar{q}q\bar{q}$	3 610 000	35 900	0.99	120 000	2180	1.8
$q\bar{q}q\bar{q}q\bar{q}$	311 000	124 000	40	23 900	12 500	52



Preselection efficiencies

Sample Pol (e^-) -80 %	Efficiency Jet E-cuts In %	Efficiency Jet E & θ cuts In %	Efficiency Jet E,θ & Btag cuts In %
HHqq→bbbbqq	95.7	92.5	85.1
Hqq	74	70	56
ee→qq	16.4	14.9	6.9
ee→qqqq	38.2	25.6	1.1
ee→qqqqqq	92	78	60
WWH			29
ZZH			84

Previous Results, missing WWH and ZZH

final state	Events -80%	Events +80%	Events -80% and +80%
$HHq\overline{q}$, both $H \rightarrow b\overline{b}$	8.63 ± 0.19	2.27 ± 0.04	10.90 ± 0.20
$HHq\overline{q}$, all H decays	10.24 ± 0.21	2.81 ± 0.05	13.05 ± 0.22
$Hq\overline{q}$, all H	0.53 ± 0.27	0.83 ± 0.28	1.36 ± 0.39
$q\overline{q}$	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
$q\overline{q}q\overline{q}$	1.88 ± 1.88	1.00 ± 0.50	2.89 ± 1.95
$q\overline{q}q\overline{q}q\overline{q}$	15.29 ± 1.44	3.60 ± 0.42	18.9 ± 1.5

Previous (incomplete) result was 2.30 σ, about 13 HHZ events left (11 for bbbb qq)
→ Most relevant background are six quark events with about 19

New pre-selection, new BDT

Using the same BDT previously leads to a significantly worse result, retune BDT with new tightened preselection BTag-3 leading jets > 2.20 instead of 1.10

Sample Pol (e^-) -80 % and +80 %	BDT(-)>0.385 BDT(+)>0.325
HHqq→bbbbqq	11.4
HHqq all decays	13.6
Hqq	1.43
ee→qq	2.02
ee→qqqq	2.14
ee→qqqqqq	15.6
WWH→qqqqH	1.80
ZZH→qqqqH	5.83
Significance in σ: S/sqrt(S+B)	2.09

This pre-selection rejects substantially more six-quark events, works also on ZZH and WWH (think e.g. on ttbar, which has two b-jets)

→ Previous pre-selection leads to a tuning with a significane of about 1.7 σ

→ Previous (incomplete) result was 2.30 σ, so even after addition of relevant triboson backgrounds decent result can be achieved (slightly larger efficiency on signal)

Beyond TMVA: XGBoost – technical details

Industry tool for optimized distributed gradient boosting with C++ and python packages, offering more flexibility compared to TMVA boosting with more tunable parameters. So far I use it on my laptop, but it can also be run over distributed computing. More complex parameter sets run faster than in TMVA
https://xgboost.readthedocs.io/en/latest/index.html

Lives outside of the ROOT environment, so ntuples have to be prepared differently:

- Use uproot tool, developed within the HEP community to transform e.g. root trees into pandas dataframes, then work from pandas dataframe to convert input into the needs of xgboost
- For python a standalone version of XGBoost exist and a version integrated into scikit-learn:
  - I cross-checked that results are indeed identical
  - Default in scikit-learn xgboost gives only classification of events (1 for signal like events, 0 otherwise), our way of selecting on the BDT score is also possible, but not set as default.
  - scikit-learn is most popular industry python machine-learning library → offers prebuilt parameter tuning functionality, validation of results

XGBoost: Training

Classification of HHZ vs all backgrounds, train on optimizing for area below ROC curve, split training vs test dataset 0.8:0.2 (tunable parameter), can stop creating additional trees, in case test dataset shows decreasing performance vs number of trees

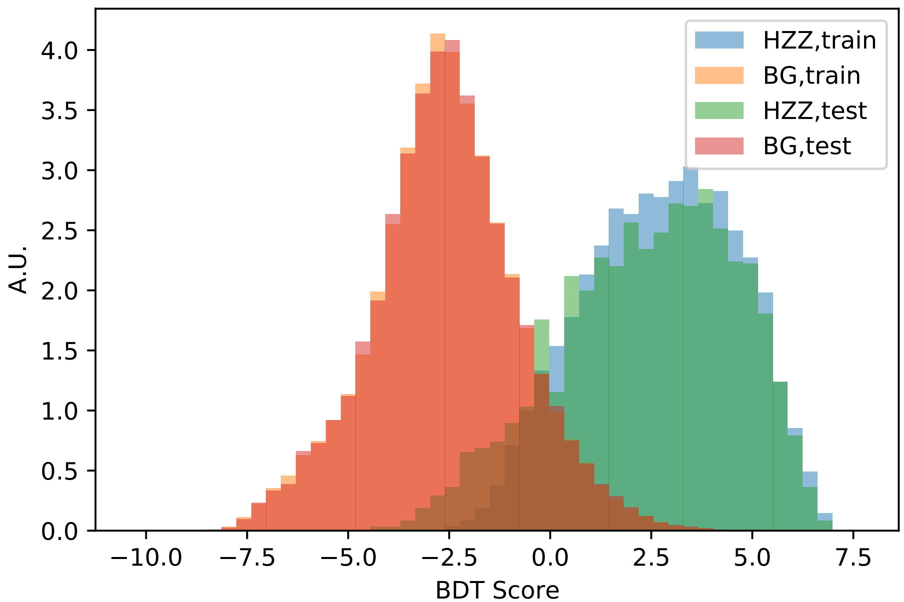
- Parameter tuning using grid parameter search is slow, considers optimising performance on test data → result is best possible result, but per default without a cross-check if test vs training BDT scores agree (aka if data become more overtrained): after 2 days of tuning overtraining checks did show an issue for signal, background distribution very stable

→ Used default setting for now, training parameters without comparing test vs training separately might need more conservative split of train vs test dataset

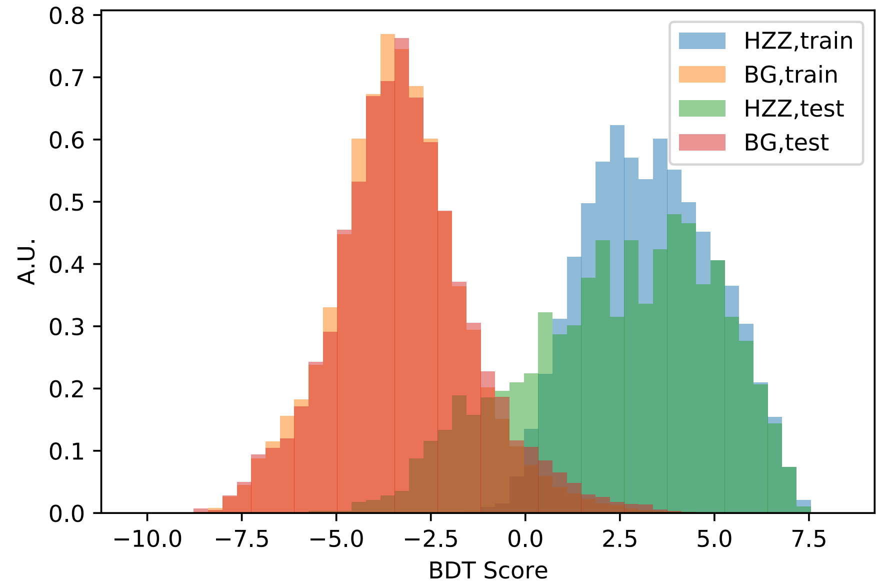
XGBoost: More overtrained wrt TMVA

Compare BDT scores for training and testing: good agreement for backgrounds, more overtrained for Signal, particularly for positive polarisation

Negative polarisation



Positive polarisation



Considering bbbb qq events only XGBoost is more aggressive

-80% Signal: 6.14 evts, BG 5.48

+80 % Signal: 2.21, BG 1.95

→ Just by summing these numbers we get 2.18 σ

Summary: HHZ

Not a boosted analysis, need to combine six individual jets

- Combining 6 jets into 3 jets, use refined VLC jets, R=1.1

Add relevant new backgrounds from WWH and ZZH → with new tighter preselection and new BDT achieve significance of 2.09 σ compared to 2.30 σ previously, signal count above 10

No sign of overtraining in BDT in TMVA

Investigate alternative BDT to check if it performs well without tight pre-selection cutoffs

→ For this preselection TMVA produced results of around 1 σ

→ Parameter set in use right now a bit overtrained for positive polarised data

→ Even for loose previous pre-selections results better