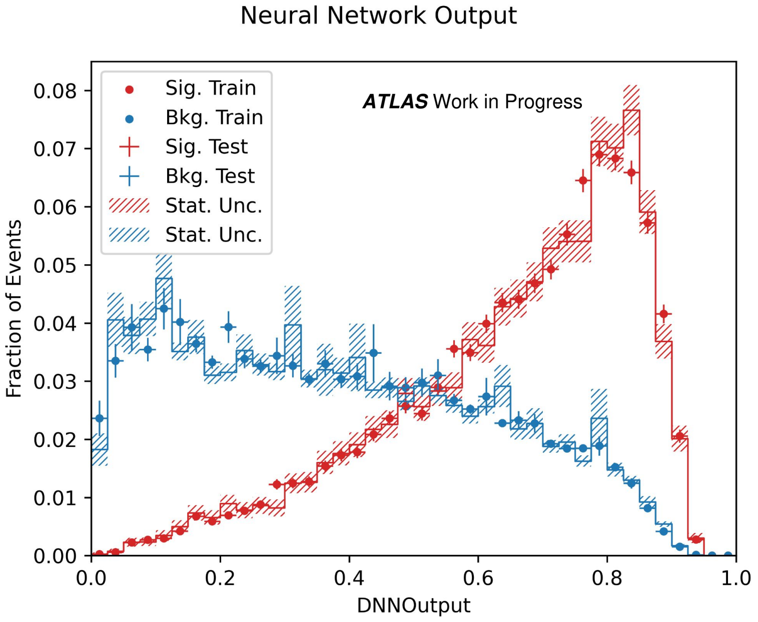
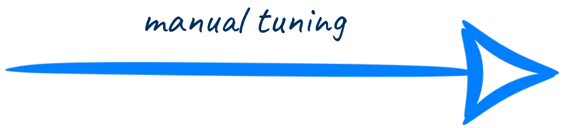
Automated Hyperparameter Optimization of Neural Networks for ATLAS analyses

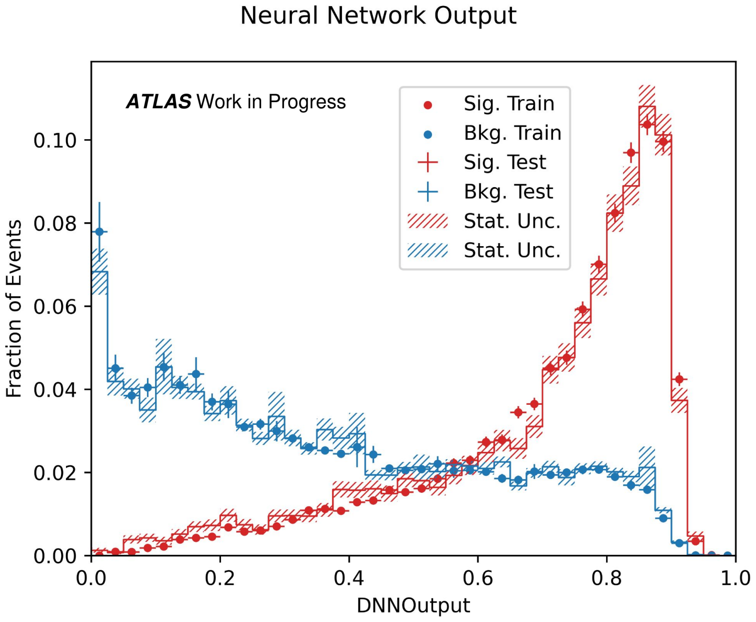
Erik Bachmann – DPG SMuK Dresden, March 22, 2023

Motivation

- Artificial neural networks are central part in many analyses
- Performance of the ANN determined by:
  - Training dataset:
    - size (fixed)
    - input variables (optimizable)
  - ANN's Hyperparameters (optimizable)







Motivation

- Artificial neural networks are central part in many analyses
- Performance of the ANN determined by:
  - Training dataset:
    - size (fixed)
    - input variables (optimizable)
  - ANN's Hyperparameters (optimizable)

Motivation

Goals:

- Improve results by using algorithms for
  - hyperparameter optimization
  - input variable selection
- Reduce manual work by automating the workflow
- Reduce turnaround time by using HPC resources

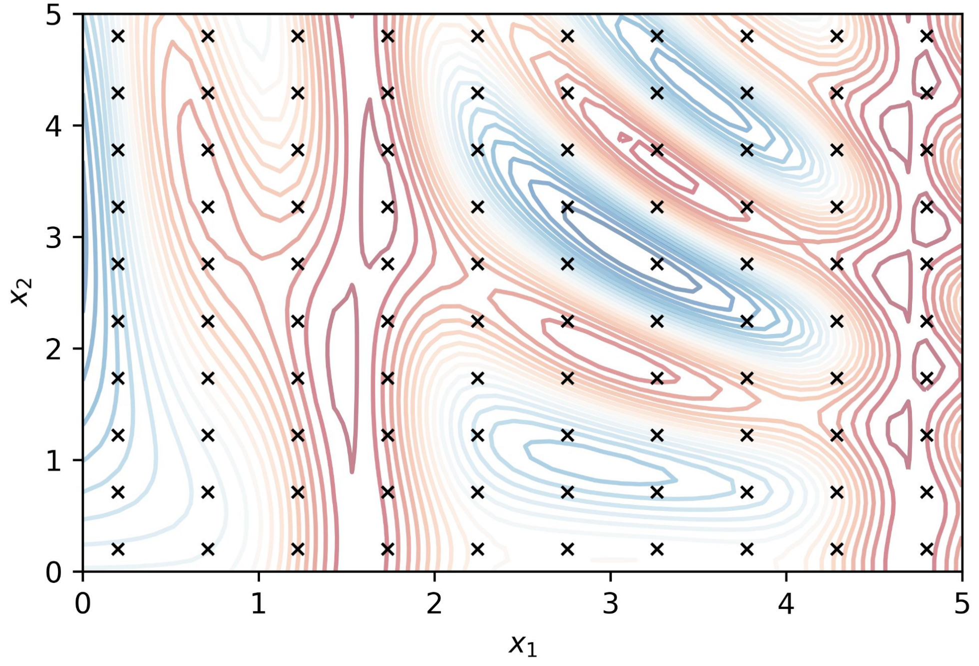
Hyperparameter optimization Framework

Hyperparameter Optimization Algorithms

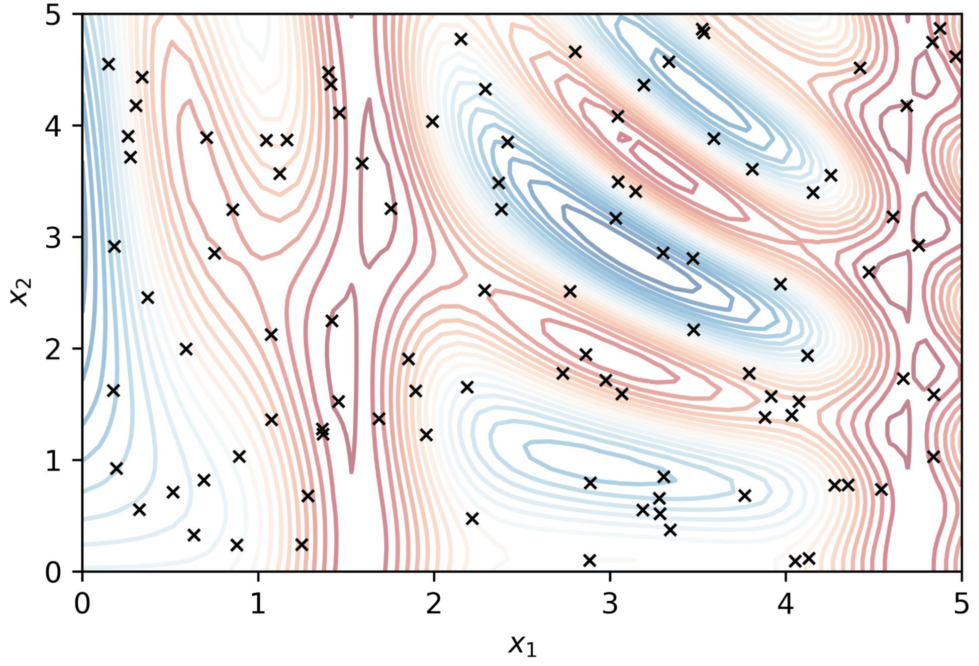
Easiest approach: grid or random search

- easy to implement, highly parallelizable
- but: no usage of collected knowledge

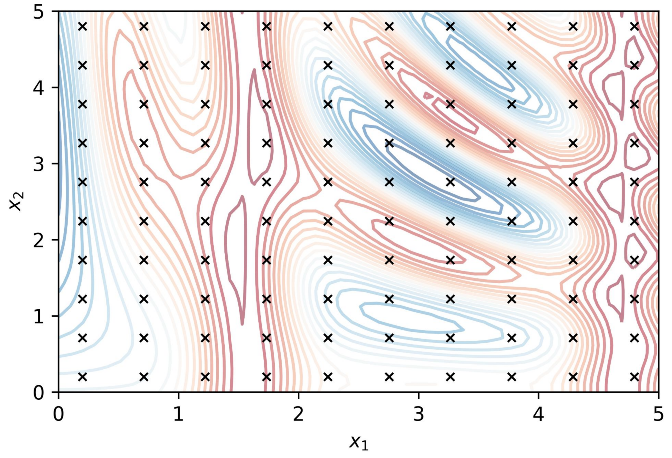
Grid search

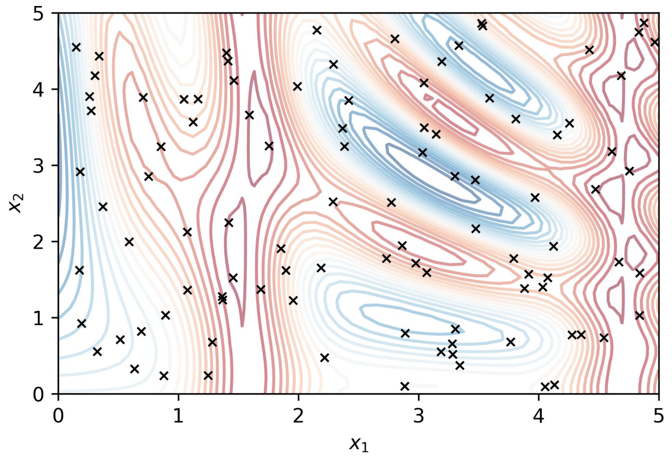


Random search

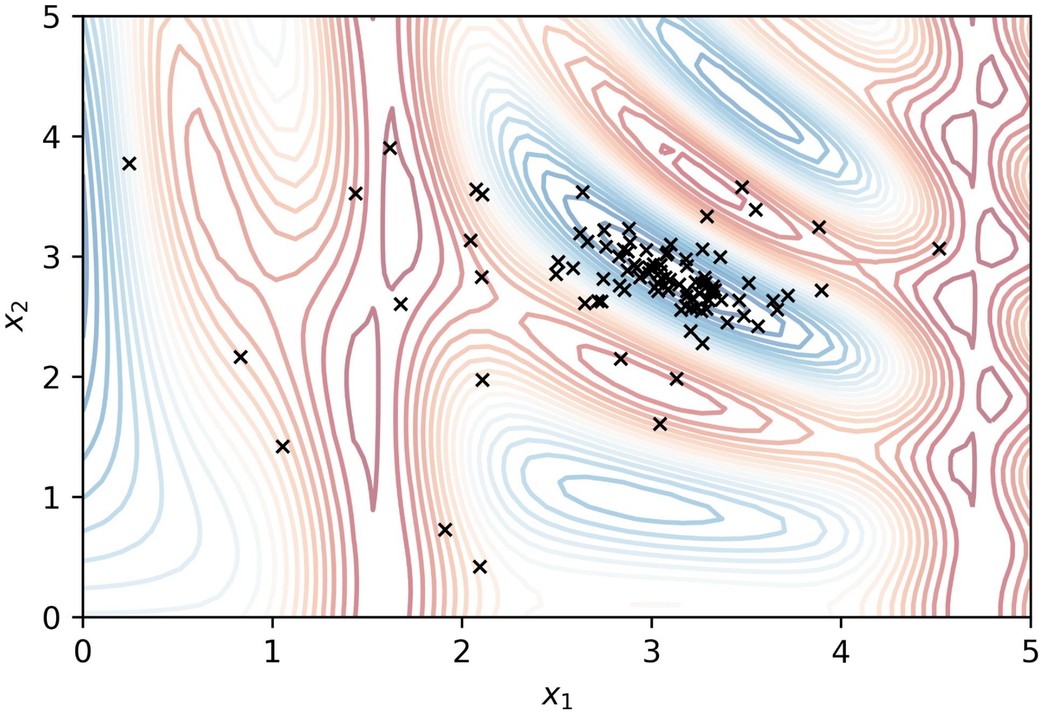


Hyperparameter Optimization Algorithms



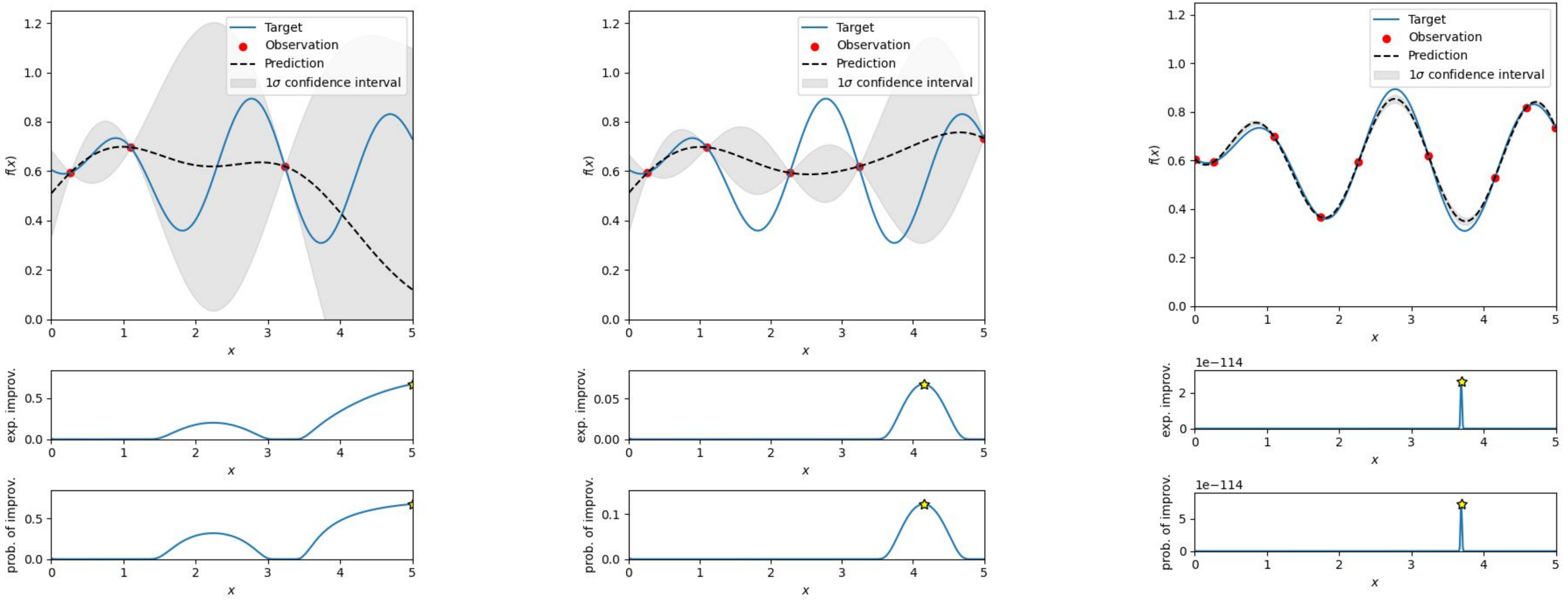


More efficient sampling techniques?



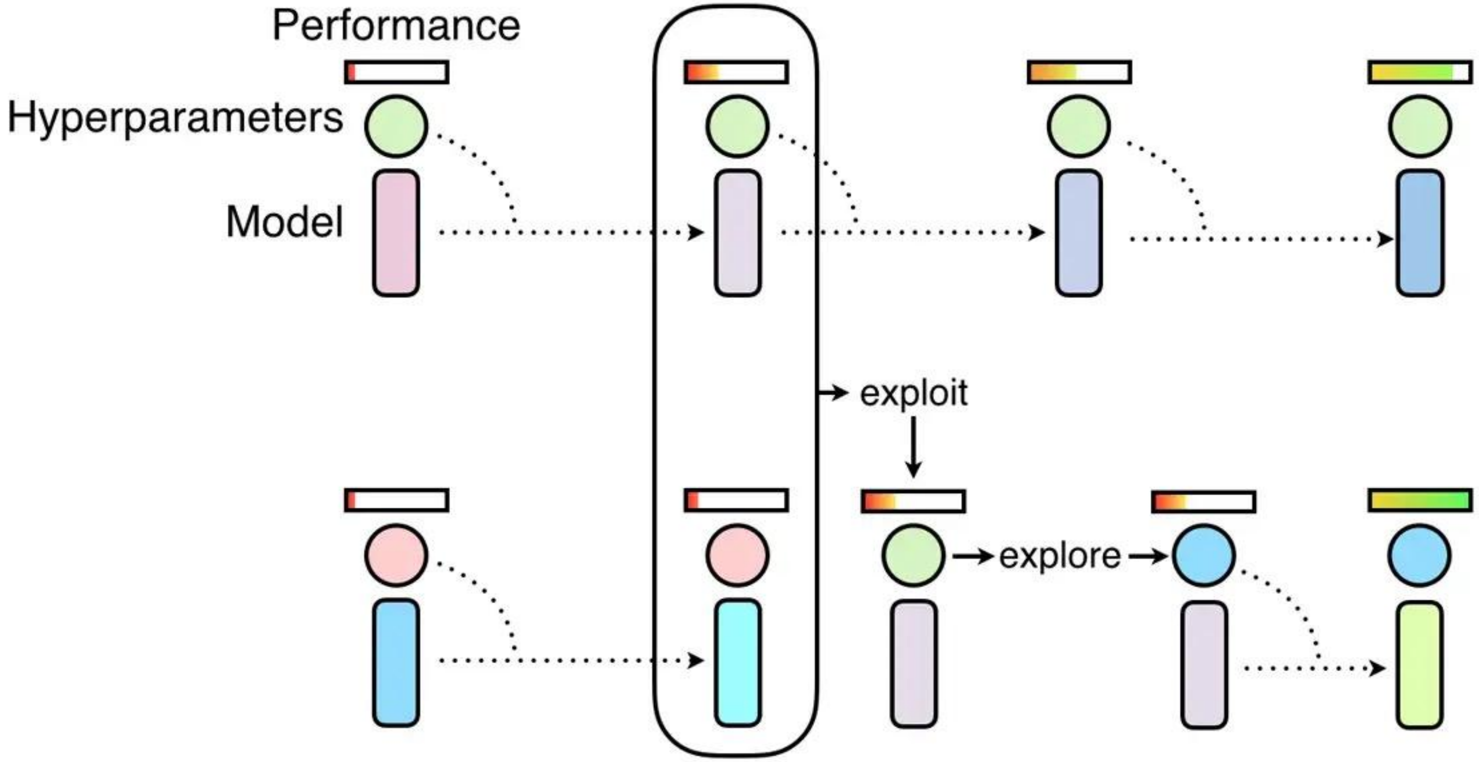
Hyperparameter Optimization Algorithms

Bayesian Optimization



Hyperparameter Optimization Algorithms

Population Based Training



Jaderberg, M. et al. (2017). Population Based Training of Neural Networks. arXiv. https://doi.org/10.48550/ARXIV.1711.09846

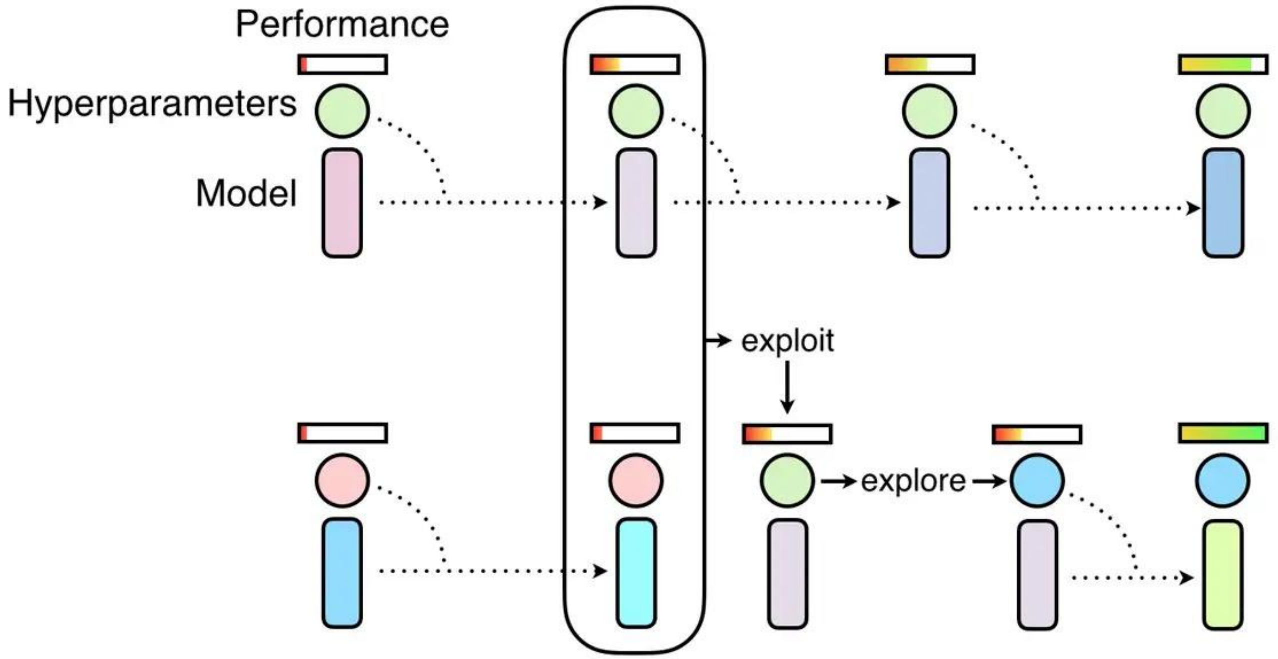
Hyperparameter Optimization Algorithms

Population Based Training

advantage: hyperparameter schedule instead of fixed values

disadvantage: not all hyper-parameters supported

→ **optimize with Bayesian Optimization first, then finetune with PBT**



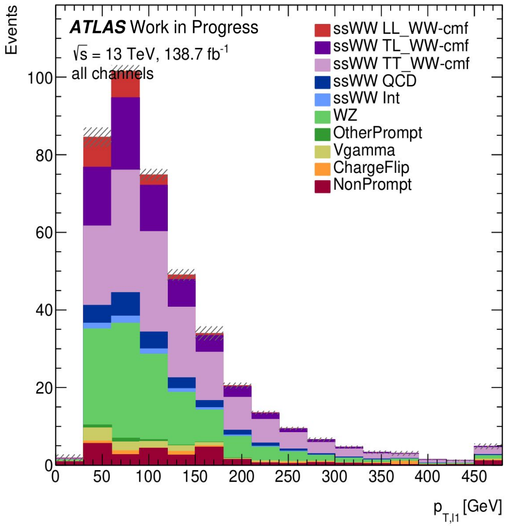
Jaderberg, M. et al. (2017). Population Based Training of Neural Networks. arXiv. https://doi.org/10.48550/ARXIV.1711.09846

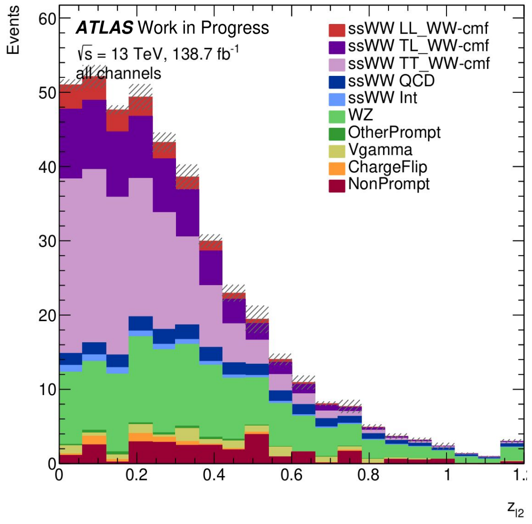
Input Variable Selection

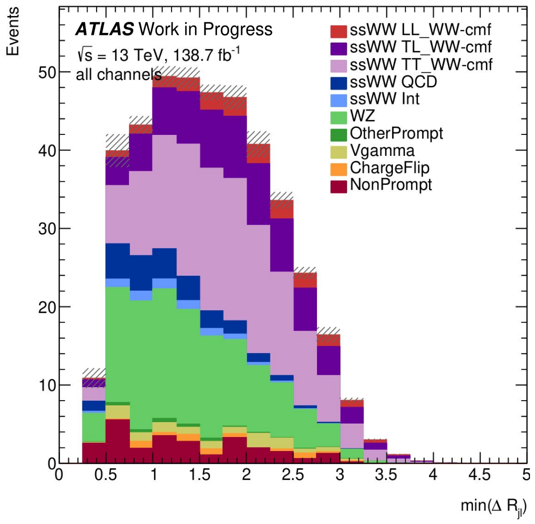
<u>Motivation:</u>

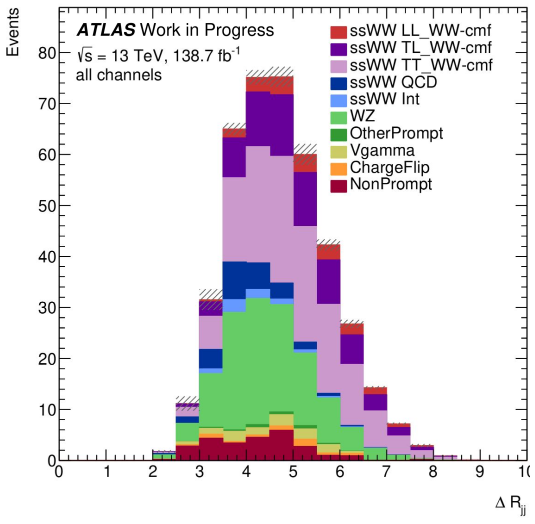
- reduce modeling uncertainties
- reduce overfitting (MLPs)

Importance of each variable depends on correlations with others
→ hard to judge beforehand





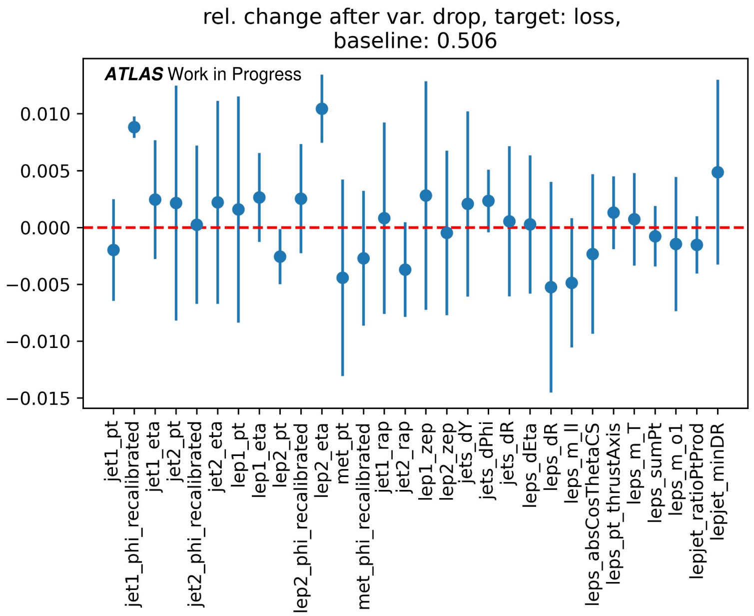


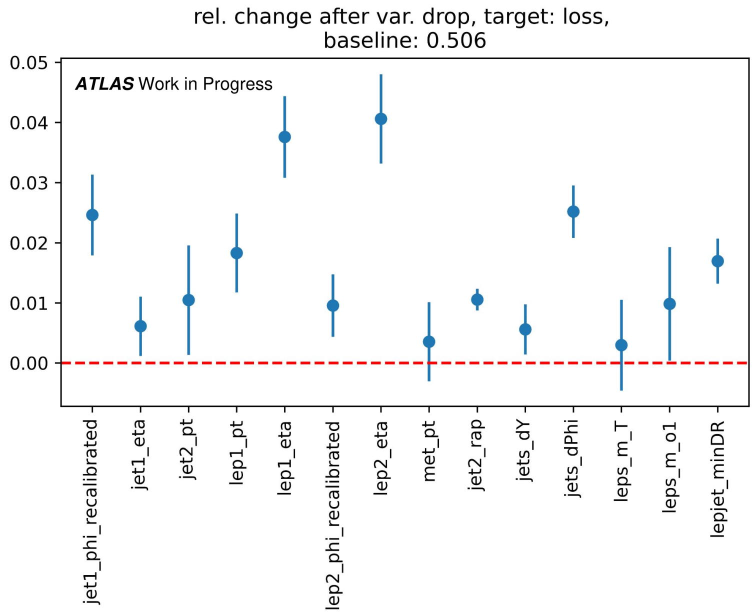


Input Variable Selection

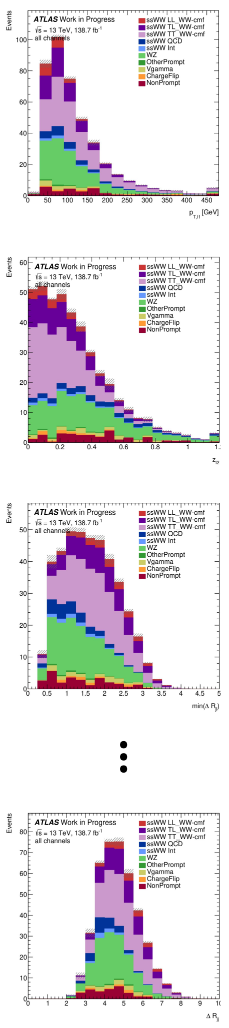
Idea: Backwards Elimination

- Baseline ANN with all input variables
- For input each variable, retrain ANN with variable removed (with k-fold cross-validation)
- Drop variable with highest relative improvement
- Use “early stopping”-like condition for termination





Hyperparameter Optimization Framework

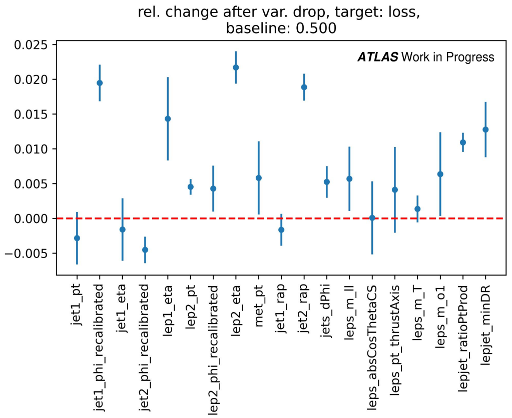


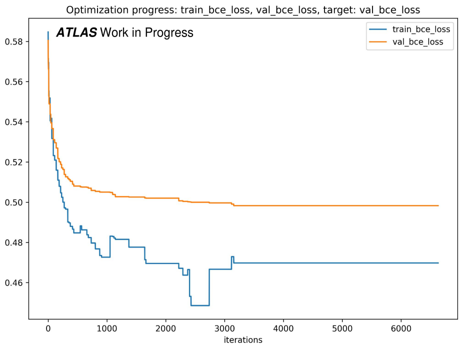
Input Variable Selection → Bayesian Optimization → PBT

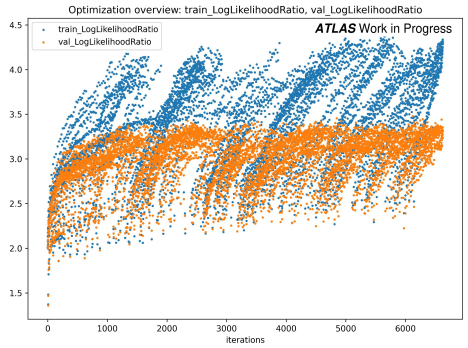
- pre-optimization (BO) with all input variables
- optimize input variables with hyper-parameters from pre-optimization

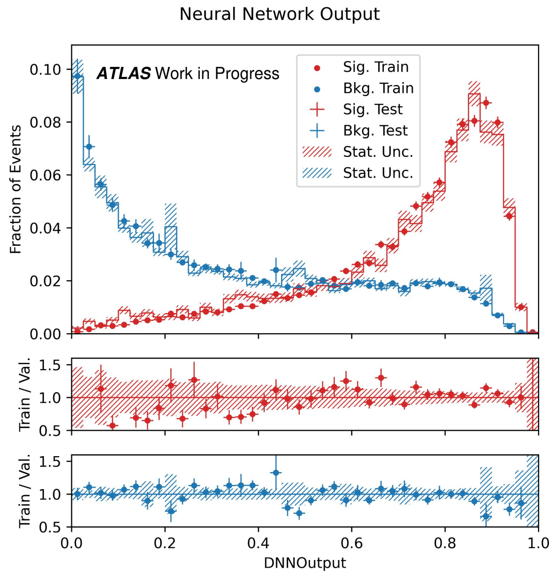
- find best hyper-parameters with optimized variables

- does HP-schedule improve result?







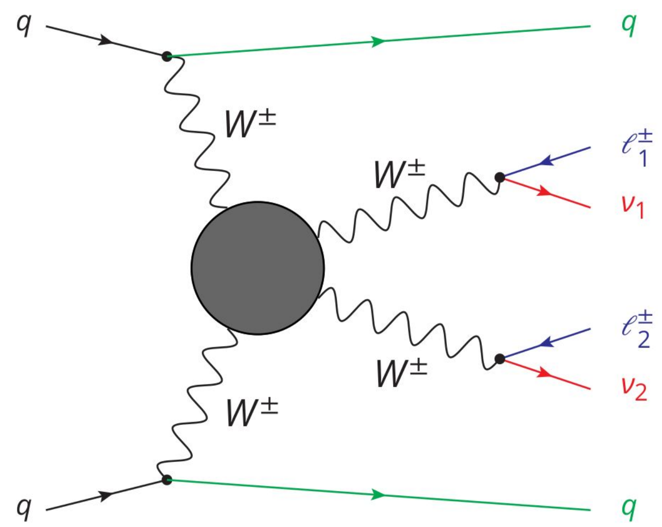


Hyperparameter Optimization Framework

Goals:

- *Generality:*
  - applicable to all Keras models for supervised learning
  - all metrics computable from ANN output and target labels are supported
  - extensively configurable via config-file
- *Scalability:* Based on Ray and Tune for distributed execution

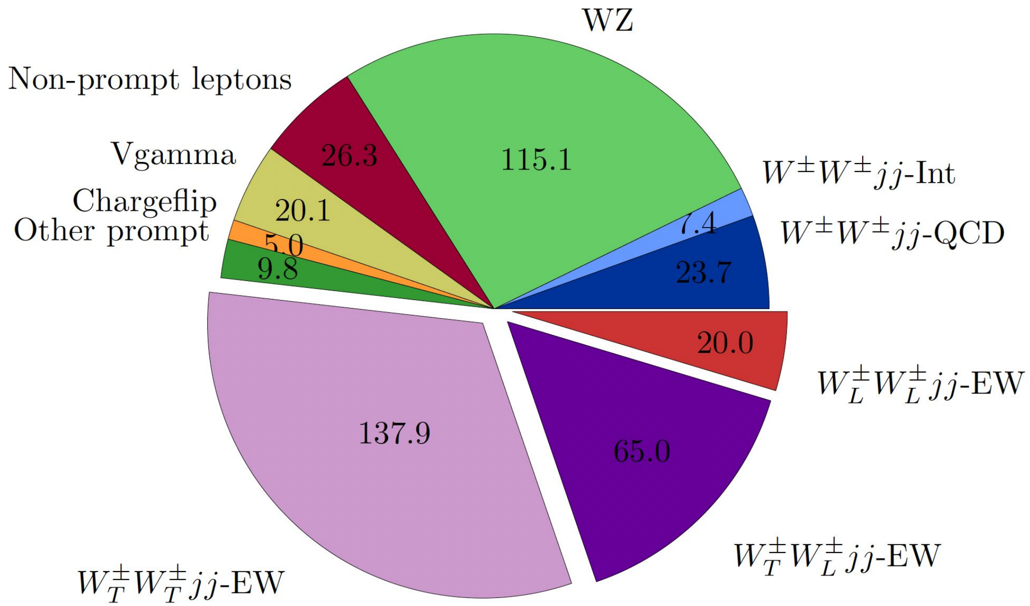
Application: Polarization in $W^{\pm}W^{\pm}jj$



- Use the full Run-2 data (139 fb^{-1})
- Multilayer perceptron to classify $W_L^{\pm}W_L^{\pm}$ vs. Bkg.
- Extract discovery significance for LL from fit on MLP output
- more details: talk T 6.5 by Max Stange

ssWW signal region

Exactly two Tight leptons with identical electrical charge
Electrons pass the charge flip rejection (ECIDS)
ee-channel: $\lvert\eta\rvert < 1.37$, $\lvert m_{ee} - m_Z\rvert > 15\,GeV$
$m_{ll'} > 20\,GeV$
$E_T^{miss} > 30\,GeV$
At least two jets Leading jet: $p_T > 65\,GeV$ Subleading jet: $p_T > 35\,GeV$
b-jet veto
$\lvert\Delta y_{jj}\rvert > 2$
$m_{jj} > 500\,GeV$



Stange, M. (2023). Machine Learning Application for Single Boson Polarization Measurement in Same-Charged WW Scattering Within the Atlas Experiment [Conference presentation]. DPG SMuK 2023, Dresden, Germany.

Application: Polarization in $W^\pm W^\pm jj$

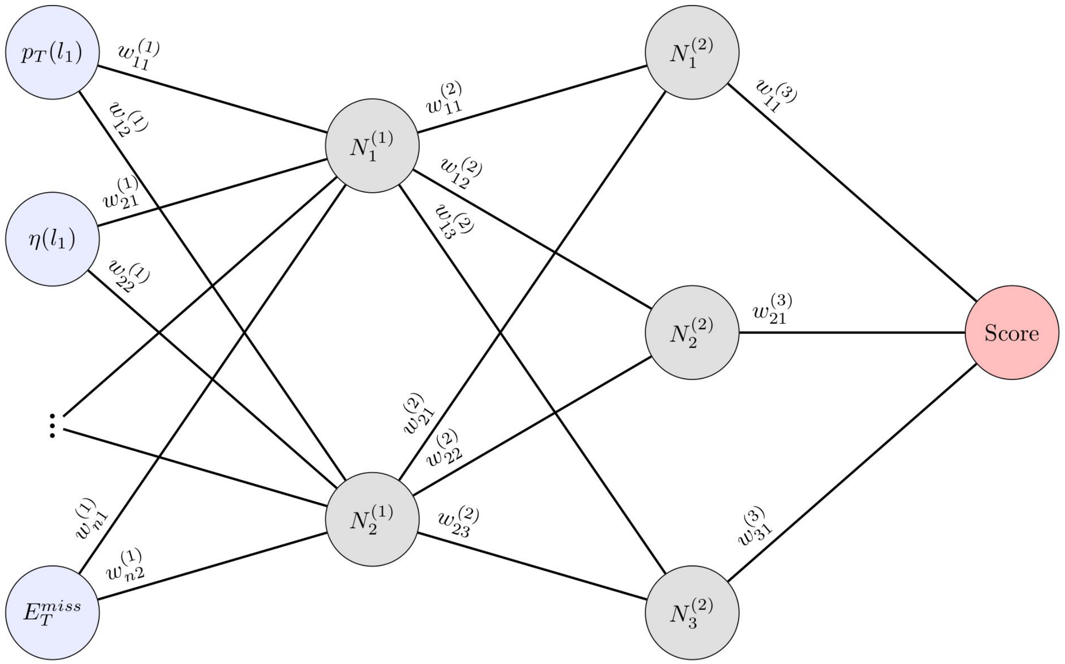
Dataset:

- $\approx$ 250k MC events, 80% used for training
- 30 high- and low-level input variables

Hyperparameter search space:

- number of hidden layers
- number of neurons per hidden layer
- dropout rate
- strength of L1 and L2 regularization
- batch size
- parameters of ADAM optimizer
- weight of signal events

$\longrightarrow$ *11 free parameters*



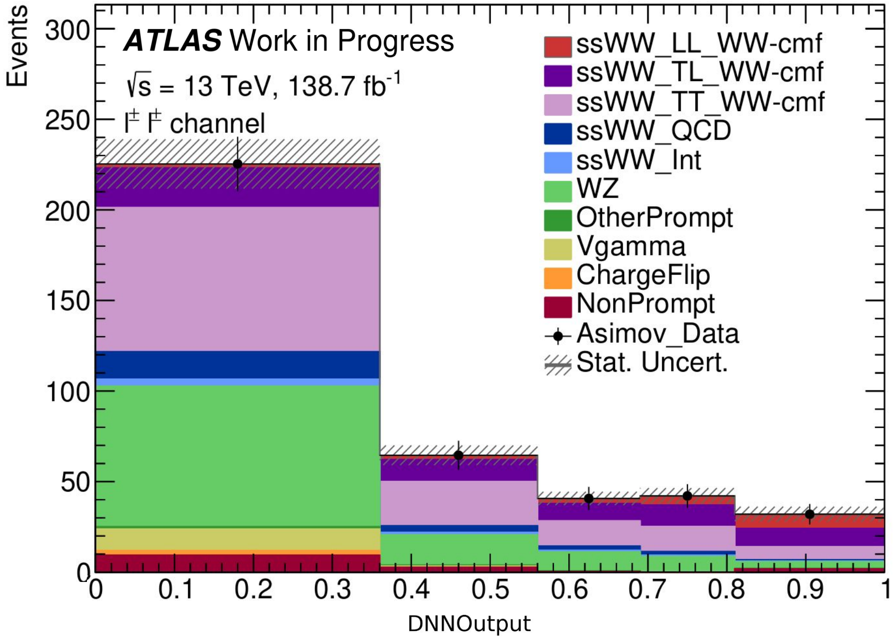
Application: Polarization in $W^{\pm}W^{\pm}jj$

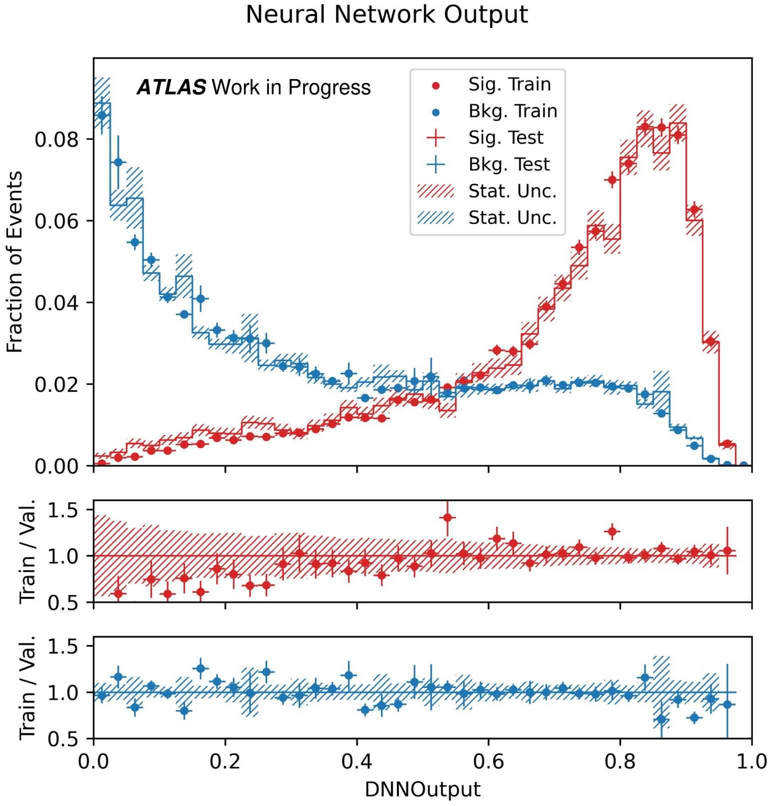
Optimization target:

binary cross-entropy loss $L(y, \hat{y}) = -\left[y \ln \hat{y} + (1-y)\ln(1-\hat{y})\right]$

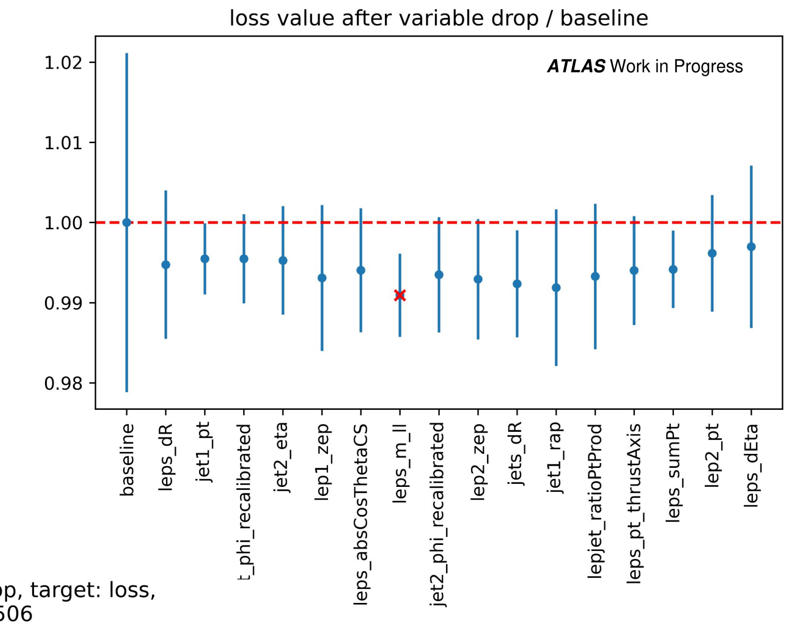
Pre-optimization (all 30 input variables):

Stat.-only significance: $Z_{stat} = 1.33$



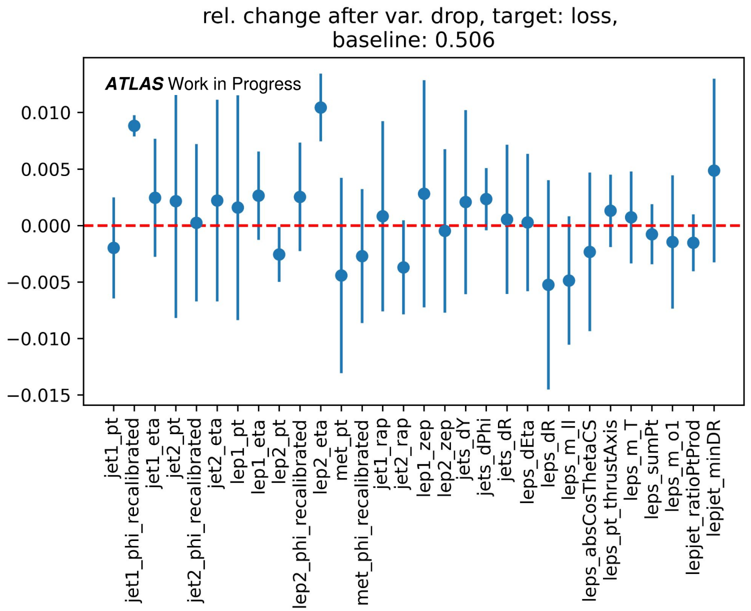


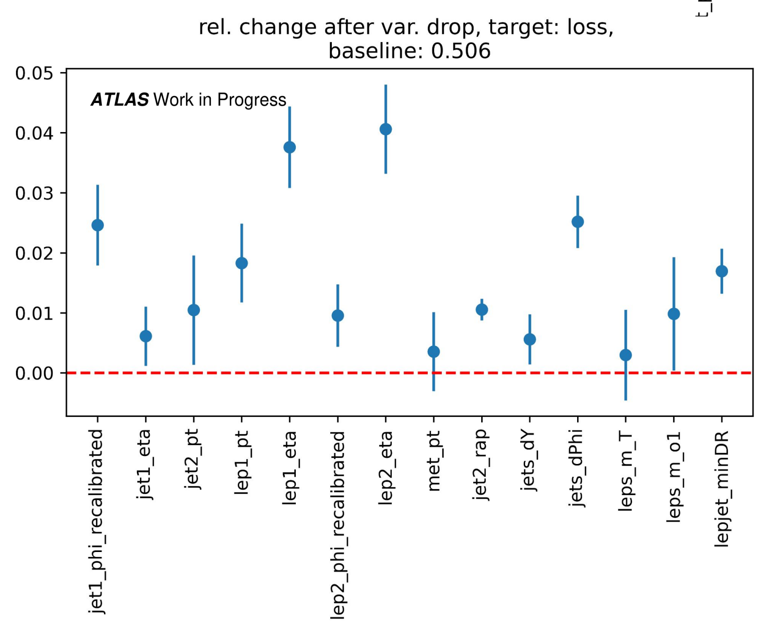
Application: Polarization in $W^{\pm}W^{\pm}jj$



Input variable selection:

- stopped when loss after drop was worse than baseline
- revert to best iteration
  $\rightarrow$ 8 variables dropped



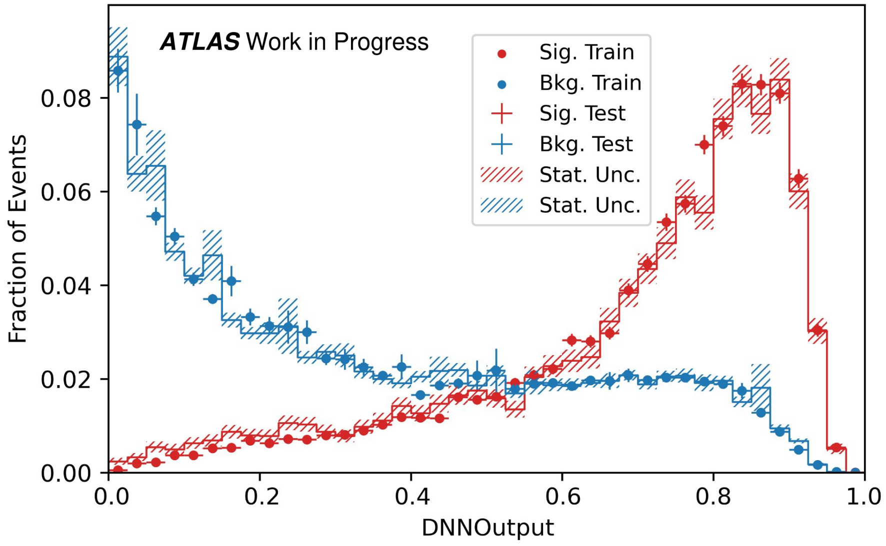


Application: Polarization in $W^{\pm}W^{\pm}jj$

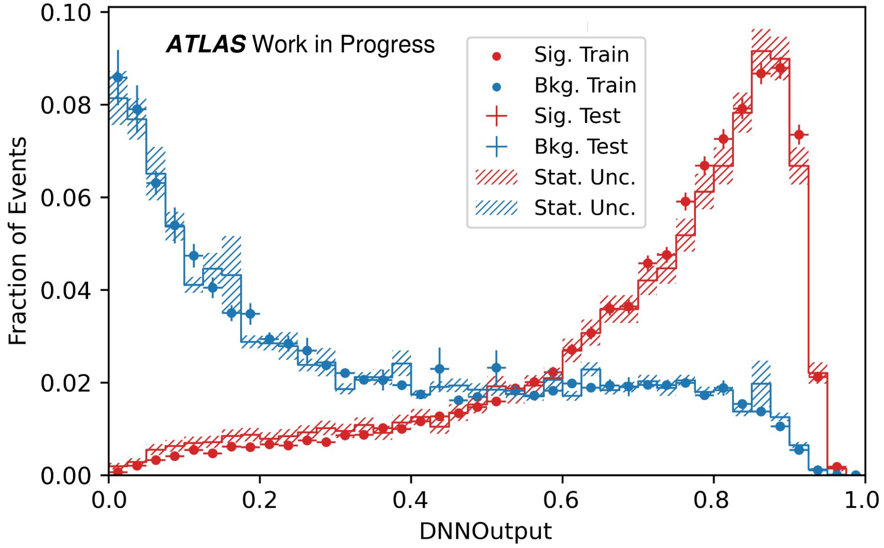
Input variable selection:

- stopped when loss after drop was worse than baseline
- revert to best iteration
  $\rightarrow$ 8 variables dropped
- smoother output distribution after main optimization

30 input variables



22 input variables

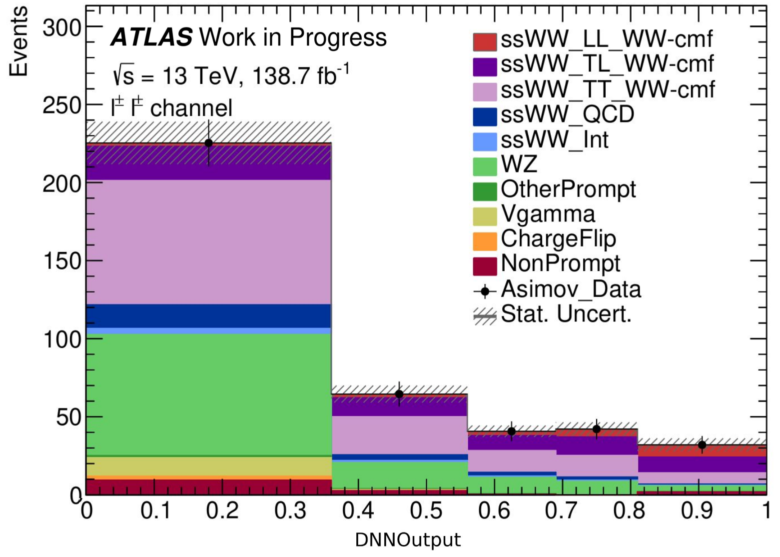


Application: Polarization in $W^{\pm}W^{\pm}jj$

Input variable selection:

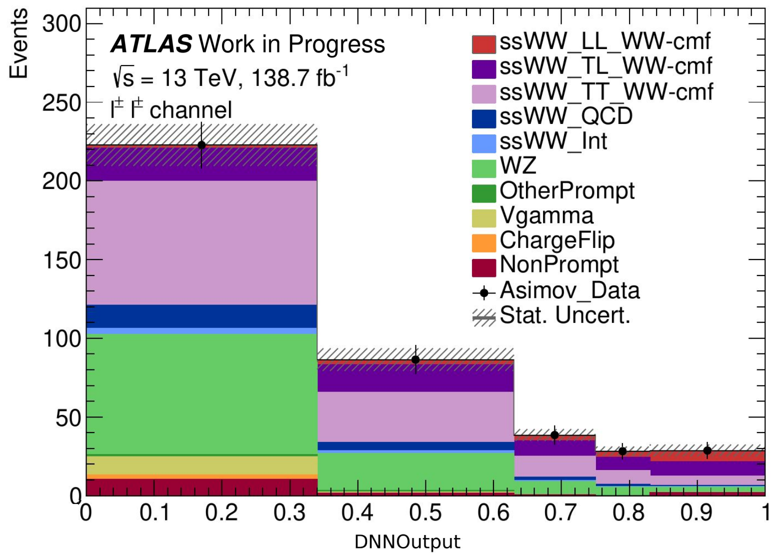
- stopped when loss after drop was worse than baseline
- revert to best iteration
  $\rightarrow$ 8 variables dropped
- smoother output distribution after main optimization

30 input variables



Stat.-only fit: $Z_{stat} = 1.33$

22 input variables

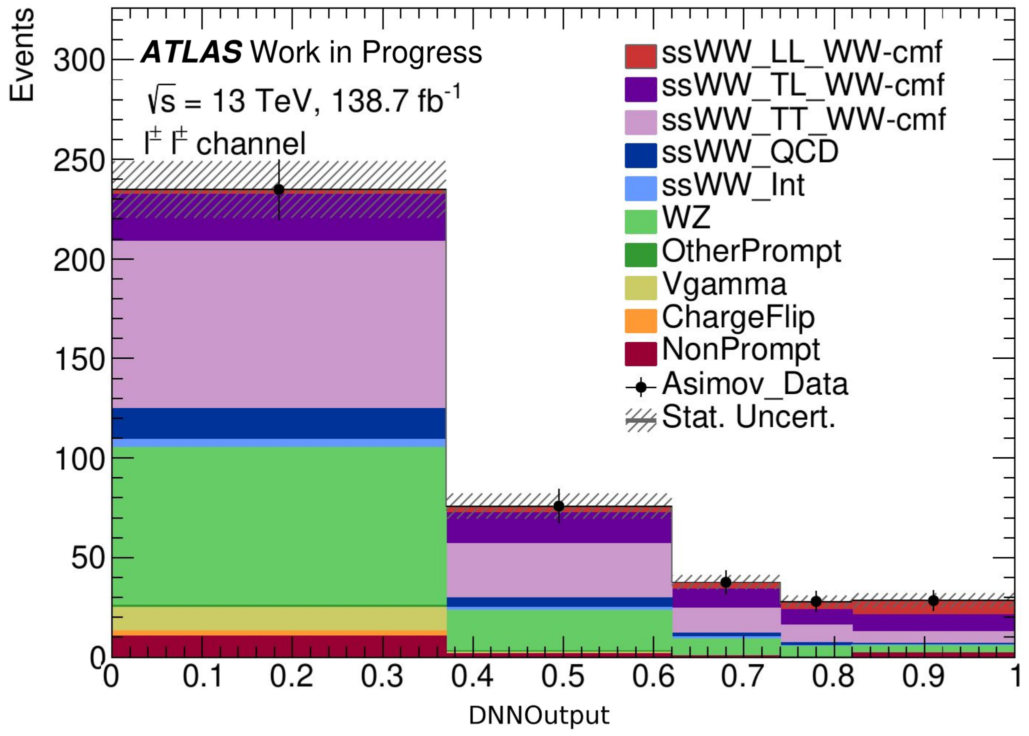


Stat.-only fit: $Z_{stat} = 1.37$

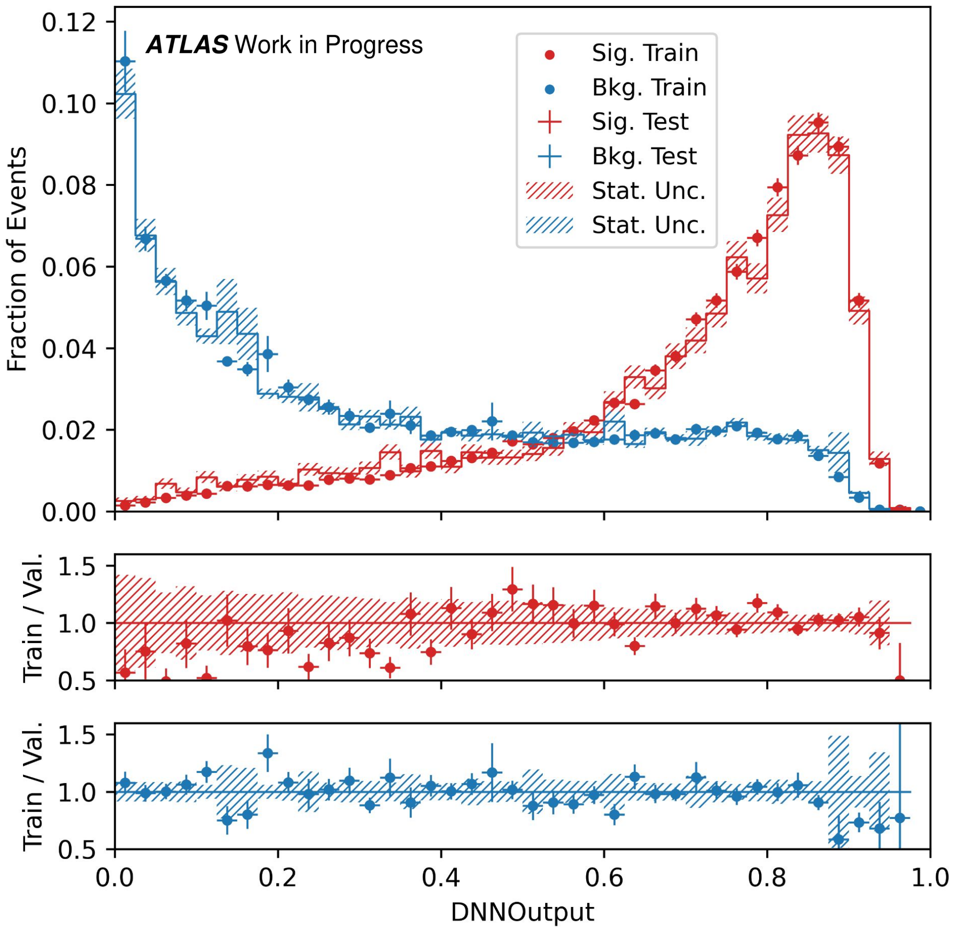
Application: Polarization in $W^{\pm}W^{\pm}jj$

Population Based Training:

Stat.-only significance: $Z_{stat} = 1.39$



Neural Network Output



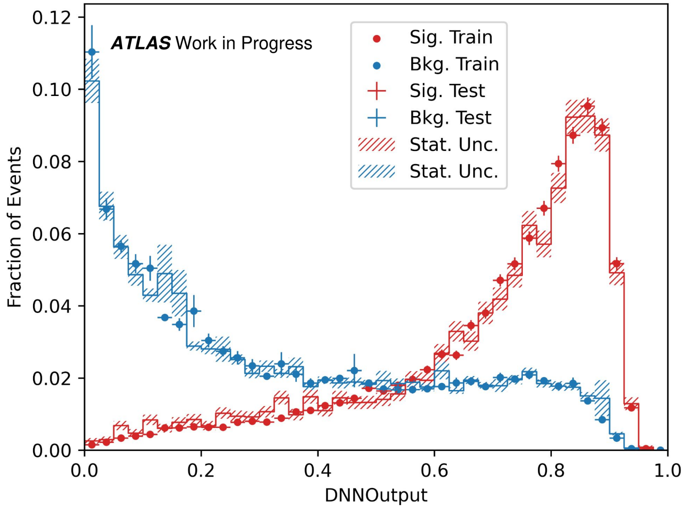
Application: Polarization in $W^{\pm}W^{\pm}jj$

Alternative optimization target: *Stat.-only Asimov Log-Likelihood ratio*

$$\log \frac{\mathcal{L}_B}{\mathcal{L}_{S+B}} = -2 \sum_{i=1}^{N} \left[(S_i + B_i) \log \frac{B_i}{S_i + B_i} + S_i \right]$$

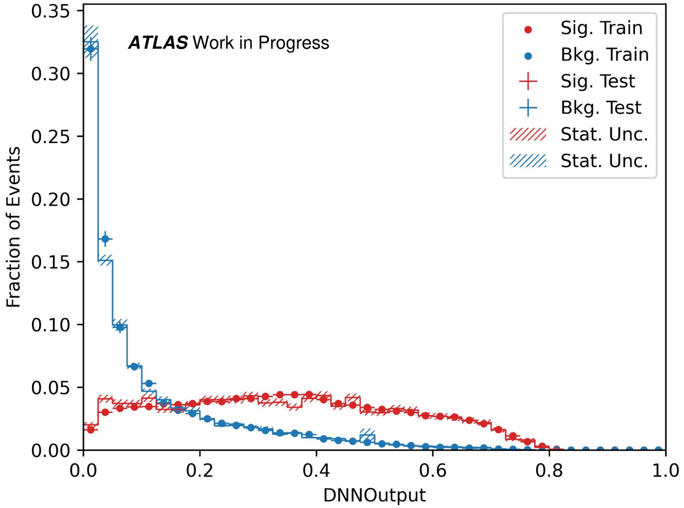
Significantly different output shape, but comparable performance

Binary Cross-entropy Loss



Stat.-only fit: $Z_{stat} = 1.39$

Log-Likelihood Ratio



Stat.-only fit: $Z_{stat} = 1.40$

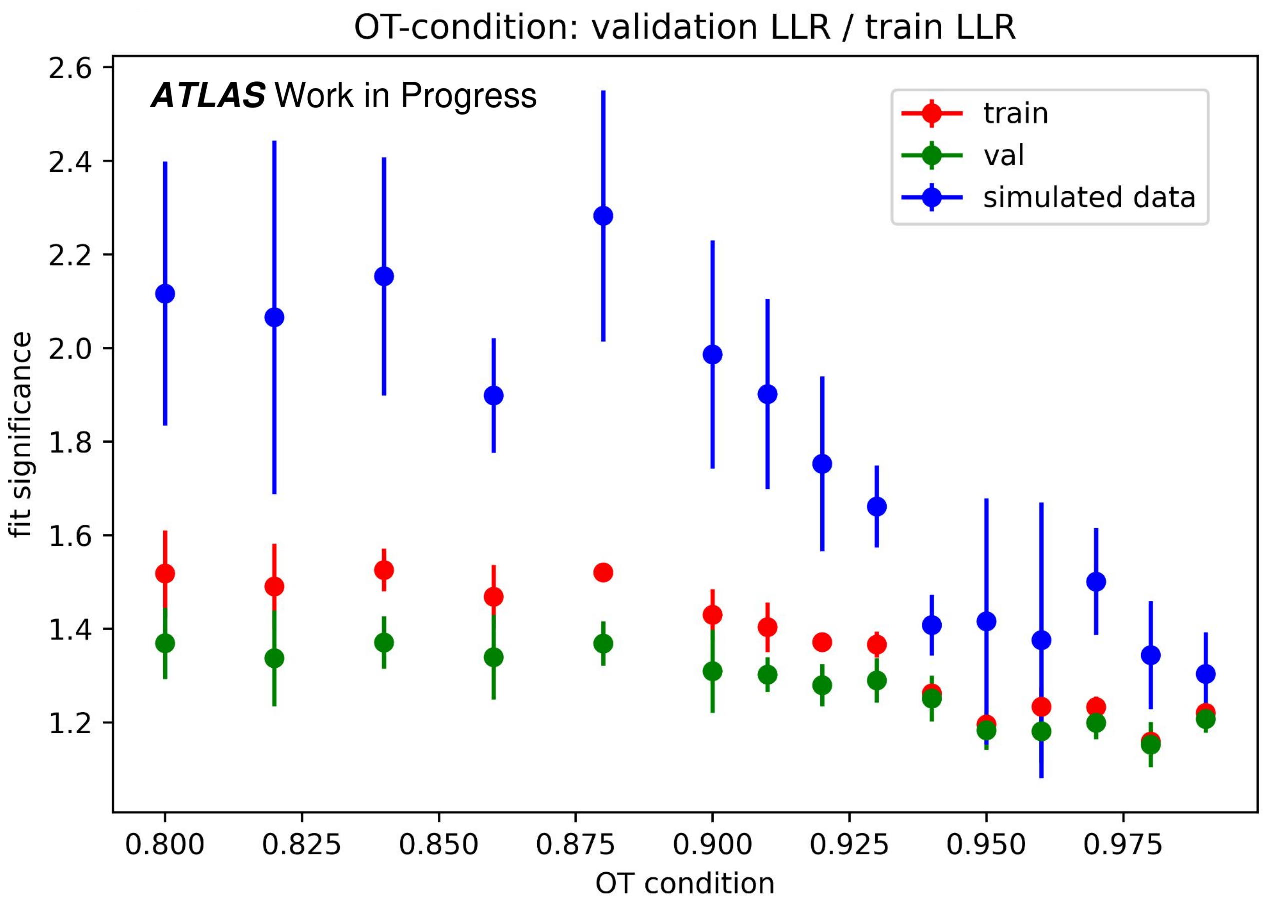
Summary

- Algorithms for hyperparameter optimization and input variable selection were introduced
- Framework for automated and distributed optimization of ANNs was presented
- First application in the ATLAS ssWW polarization analysis was shown

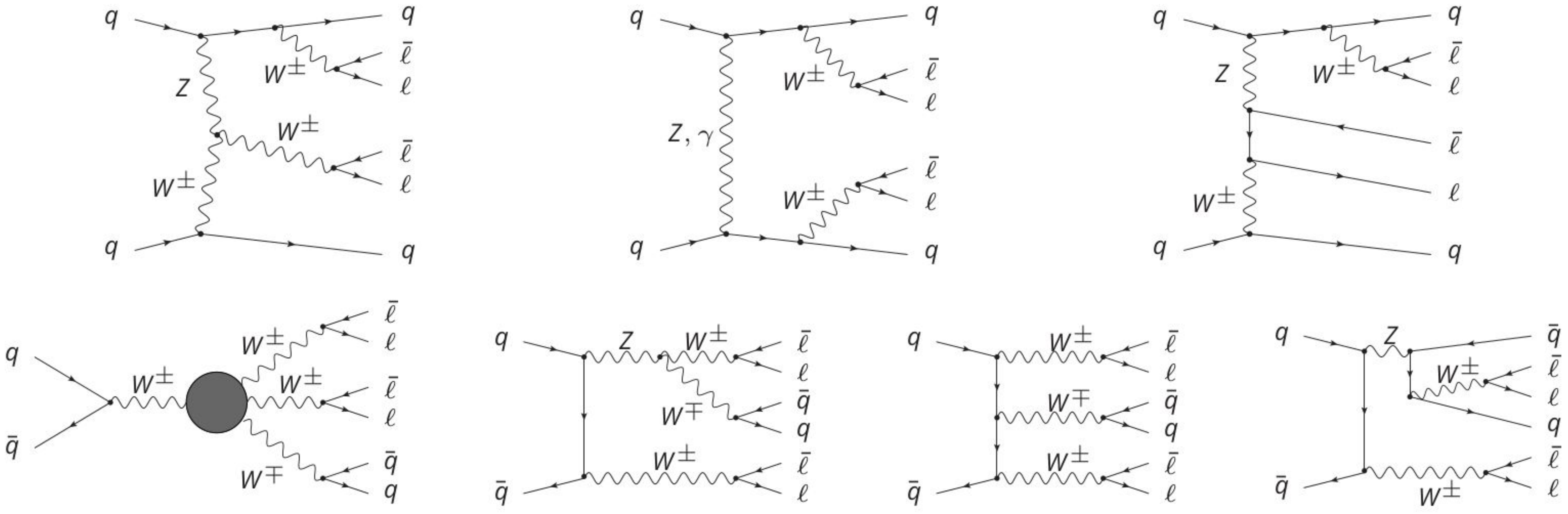
APPENDIX

Overtraining conditions

Comparing metrics on training and validation datasets allows the definition of overtraining conditions

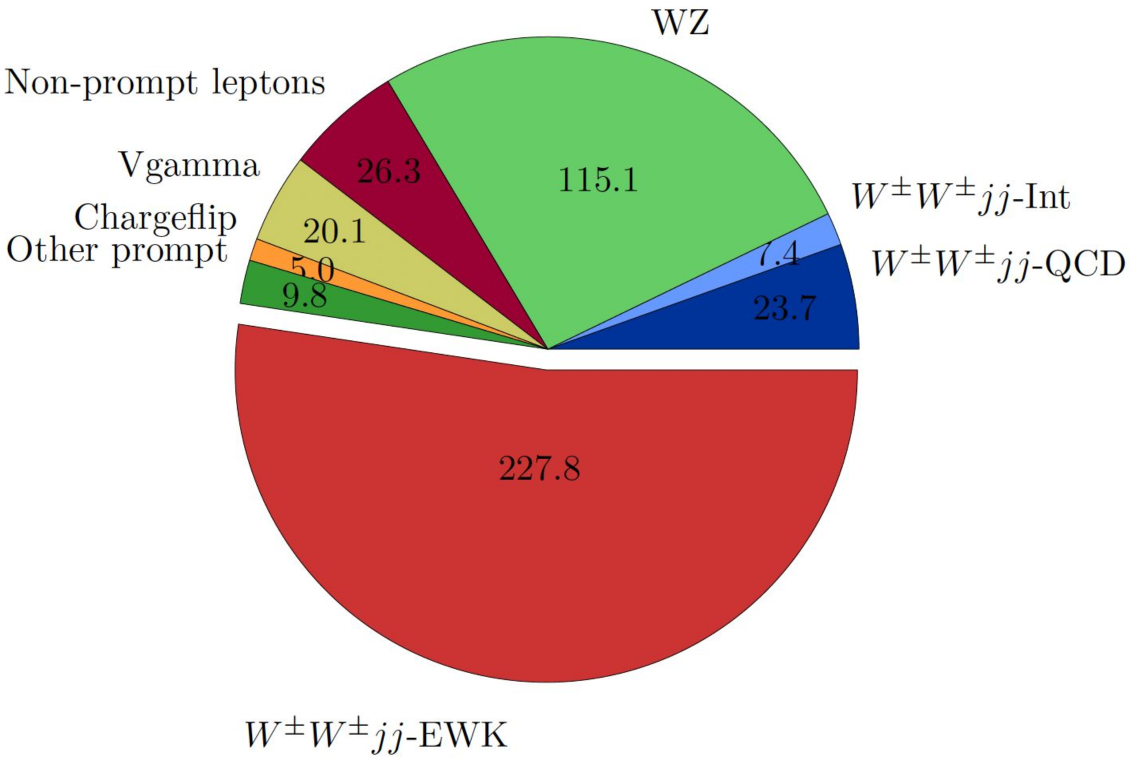


Same-sign WW-boson scattering at the LHC

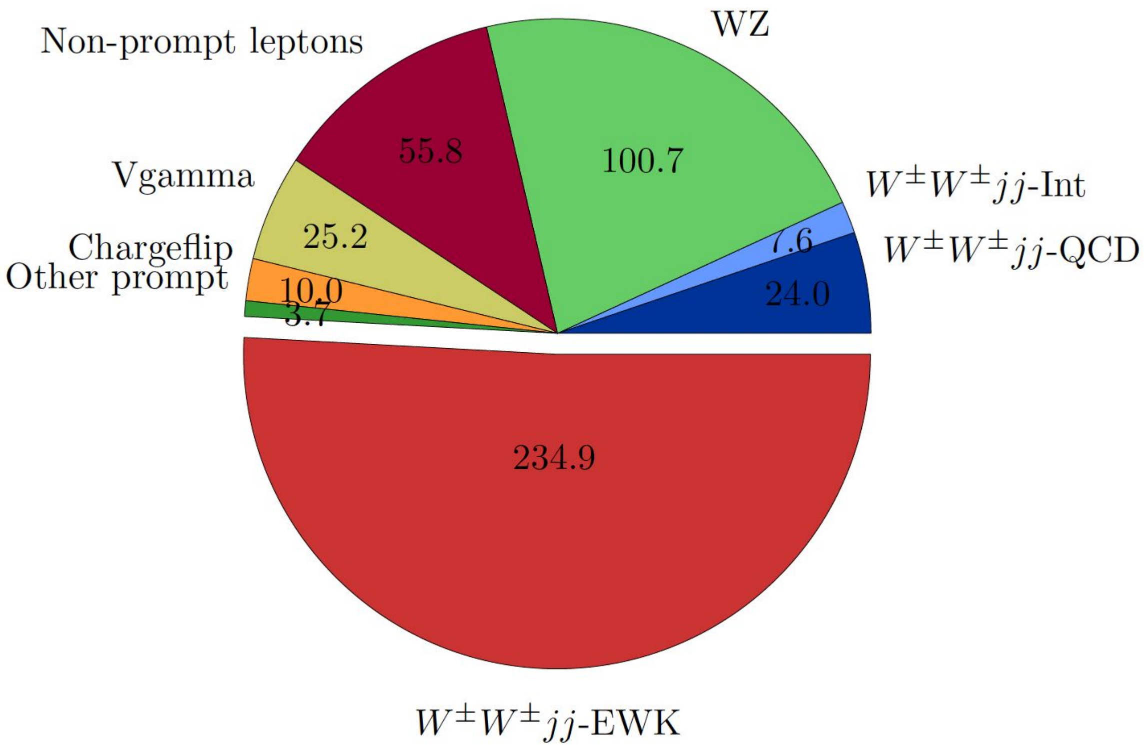


Polarization vs. Nominal Analysis

Polarization Analysis



Nominal Analysis



Polarization vs. Nominal Analysis

	$\ell^-\ell^-$	$\ell^+\ell^+$	$\ell^\pm\ell^\pm$
Other prompt	1.171 ± 0.093 ± 0.000 ± 0.324	2.519 ± 0.144 ± 0.000 ± 0.730	3.690 ± 0.172 ± 0.000 ± 1.052
WZ	39.040 ± 0.601 ± 0.769 ± 2.367	61.706 ± 1.012 ± 1.619 ± 3.924	100.746 ± 1.177 ± 2.382 ± 6.262
Chargeflip	5.331 ± 0.237 ± 0.000 ± 1.897	4.762 ± 0.200 ± 0.000 ± 1.778	10.093 ± 0.310 ± 0.000 ± 3.675
Vgamma	12.576 ± 2.609 ± 0.000 ± 4.586	12.624 ± 1.935 ± 0.000 ± 5.913	25.200 ± 3.248 ± 0.000 ± 10.117
Non-prompt	25.310 ± 3.439 ± 0.000 ± 4.306	30.528 ± 4.154 ± 0.000 ± 5.838	55.838 ± 5.408 ± 0.000 ± 9.900
$W^\pm W^\pm$jj QCD	7.096 ± 0.086 ± 2.511 ± 0.183	16.946 ± 0.132 ± 5.901 ± 0.453	24.042 ± 0.157 ± 8.410 ± 0.635
$W^\pm W^\pm$jj Int	2.161 ± 0.041 ± 0.141 ± 0.048	5.433 ± 0.065 ± 0.309 ± 0.074	7.594 ± 0.077 ± 0.444 ± 0.116
asimovData	156.382 ± 12.505 ± 0.000 ± 0.000	305.721 ± 17.485 ± 0.000 ± 0.000	462.103 ± 21.497 ± 0.000 ± 0.000
$W^\pm W^\pm$jj EW	63.697 ± 0.000 ± 2.291 ± 1.175	171.203 ± 0.000 ± 5.011 ± 2.665	234.900 ± 0.000 ± 6.638 ± 3.831
Total Expected	156.382 ± 4.367 ± 3.488 ± 7.893	305.721 ± 4.702 ± 7.915 ± 11.594	462.103 ± 6.430 ± 10.985 ± 19.046

Table 22: Pre-fit yields in the signal region. Shown are central values with statistical, theory, and experimental systematic uncertainties.

ssWW Polarization		[Prompt+ChargeFlip+Fake+Unknown]	
ssWW LL	506204	MGPy8_ssWWjj_lep_LL_EW6_WWcmf_LO	p4252
ssWW LT	506206	MGPy8_ssWWjj_lep_TL_EW6_WWcmf_LO	p4252
ssWW TT	506201	MGPy8_ssWWjj_lep_TT_EW6_WWcmf_LO	p4252
ssWW QCD		[Prompt]	
QCDSigMGH7EG_LO	500990	MGH7EG_LO_EW4_ssWWjj	p4252
ssWW Int		[Prompt]	
INTSigMGH7EG_LO	500991	MGH7EG_LO_INT_ssWWjj	p4252
WZ		[Prompt]	
WZQCD_Sherpa222	364253	Sherpa_222_NNPDF30NNLO_lllv	p3975
WZEW_MGPy8EG	364739 - 364742	MGPy8EG_NNPDF30NLO_A14NNPDF23LO_lvlljjEW6_[...]	p3975
OtherPrompt		[Prompt]	
ttV_aMcAtNloPythia8	410155	aMcAtNloPythia8EvtGen_MEN30NLO_A14N23LO_ttW	p3975
ttV_aMcAtNloPythia8	410081	MadGraphPythia8EvtGen_A14NNPDF23_ttbarWW	p4252
ttZ_aMcAtNloPythia8	410218 - 410220	aMcAtNloPythia8EvtGen_MEN30NLO_A14N23LO_tt[...]	p3975
singletop_tV	410560	MadGraphPythia8EvtGen_A14_tZ_4fl_tchan_noAllHad	p3975
singletop_tV	410644, 410645	PowhegPythia8EvtGen_A14_singletop_schan_lept_[...]	p4434
singletop_tV	410648, 410649	PowhegPythia8EvtGen_A14_Wt_DR_dilepton_[...]	p4434
singletop_tV	410658, 410659	PhPy8EG_A14_tchan_BW50_lept_[...]	p4434
Vgamma		[Prompt+Unknown]	
ZgammaSherpa222	364500 - 364509	Sherpa_222_NNPDF30NNLO_[...]	p4252
WgammaSherpa222	364522 - 364535	Sherpa_222_NNPDF30NNLO_[...]	p3975
ChargeFlip		[Prompt+Chargeflip+Unknown]	
Zee_MGPy8EG	363147 - 363170	MGPy8EG_N30NLO_Zee_Ht[...]	p3975
NonPrompt		[ChargeFlip+Fake+Unknown]	
WjetsMadgraph	363600 - 363671	MGPy8EG_N30NLO_W[...]	p3975
ttbar_nonallhad	410470	PhPy8EG_A14_ttbar_hdamp258p75_nonallhad	p3975
ttbar_allhad	410471	PhPy8EG_A14_ttbar_hdamp258p75_allhad	p3975

	pol. study	support note
ssWW LL	20.0 ± 0.1	
ssWW TL	65.0 ± 0.4	
ssWW TT	137.9 ± 0.7	
ssWW LX	85.0 ± 0.5	
ssWW pol	222.9 ± 0.8	
ssWW Madgraph	227.8 ± 0.5	234.900
ssWW QCD	23.7 ± 0.1	24.042
ssWW Int	7.4 ± 0.1	7.594
WZ	115.1 ± 1.2	100.746
OtherPrompt	9.8 ± 0.4	3.690
Vgamma	20.1 ± 7.9	25.200
ChargeFlip	5.0 ± 1.3	10.093
NonPrompt	26.3 ± 3.6	55.838
Background	207.5 ± 8.9	227.203