Karlsruhe Institute of Technology

Institute of Experimental Particle Physics

Implementation of an Optimal Statistical Inference to Reduce Systematic Uncertainties in the $\mathrm{H} \rightarrow \tau\tau$ Analysis at the CMS Experiment

Uncertainty aware training

12. 12. 2022

Markus Klute, **Artur Monsch**, Günter Quast, Lars Sowa, Roger Wolf

Statistical inference and systematic uncertainties in HEP

- Systematic variations provided in form of event weights and
- Incorporated in a statistical model as constraints

$$\mathcal{L}\left(N, \mu, \{\theta_j\}\right) = \prod_{i=1}^{N_{\text{bins}}} \mathcal{P}\Big(n_i \Big| \mu \underbrace{s_i\left(\{\theta_j\}\right)} + \underbrace{b_i\left(\{\theta_j\}\right)}\Big) \overbrace{\prod_{j=1}^{M} \mathcal{C}\left(\theta_j \middle| \mu_{\theta_j}, \sigma_{\theta_j}\right)}^{M \text{ systematic uncertainties as nuisance parameters}} ,$$

Impact of θ_j on signal and background models

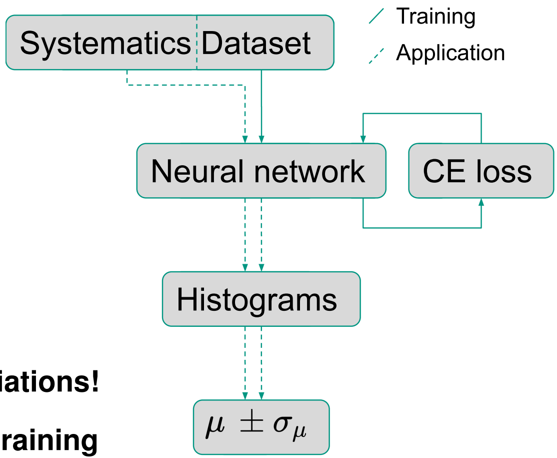
where

- $\mathcal{P}$ is the Poisson distribution
- μ is the signal strength
- s_i, b_i the expected number of signal and background events and
- n_i number of measured events in bin i

Target: Extraction of $\mu \pm \sigma_\mu$

Structure of NN-based analyses workflows (in HEP)

- Training task
  Separation of signal and background
- Application
  Propagation of nominal and shifted samples through trained NN
- Inference
  Extraction of $\mu \pm \sigma_\mu$ from nominal and shifted NN outputs



CE training is not aware of systematic variations!

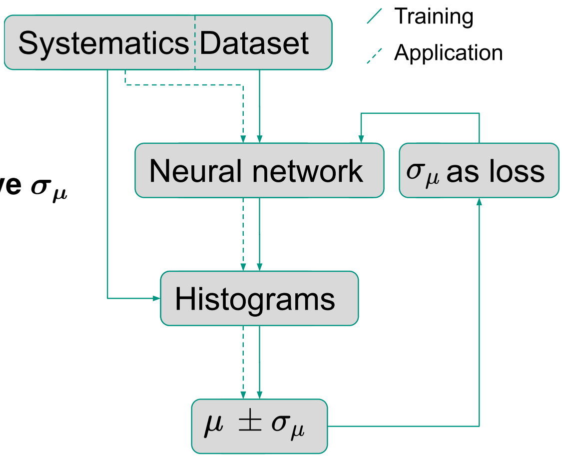
$\rightarrow$ **suboptimal training**

Uncertainty aware training

- Keep the application and inference step
- Change the training step by replacing

Training objective CE $\rightarrow$ Analysis objective σ_μ

- Incorporate systematic variations in calculation of σ_μ via
  - Event weights
  - Shifted data sets, propagated through NN



New Loss: σ_μ

- Starting with a binned Likelihood $\mathcal{L}\left(N, \mu, \{\theta_j\}\right)$ where systematic uncertainties for every bin i can be incorporated as

$$s_i = s_{i_0} + \sum_j^M \theta_j s_{i_{\text{shift}}}, \quad b_i = b_{i_0} + \sum_j^M \theta_j b_{i_{\text{shift}}}$$

  assuming an ideal case by
  - Obtain nominal value from Asimov data set: $n_i = \mu s_i + b_i$, $\mu = 1$ and
  - Applying no pull on the nuisance parameters $\theta_j = 0 \; \forall \; j$

- The estimation of σ_μ is obtained from the Fisher information:

$$\mathcal{F}_{ij} = \mathbb{E}\left[\left(\frac{\partial^2}{\partial x_i \partial x_j}\left(-\ln\mathcal{L}\right)\right)_{x_i, x_j = \mu, \{\theta_j\}}\right] \overset{\text{Asimov}}{=} \left(\text{Hess}\left(-\ln\mathcal{L}\right)\right)_{ij} \qquad \Rightarrow \qquad \left(\mathcal{F}_{ij}\right)^{-1} = V_{ij}$$

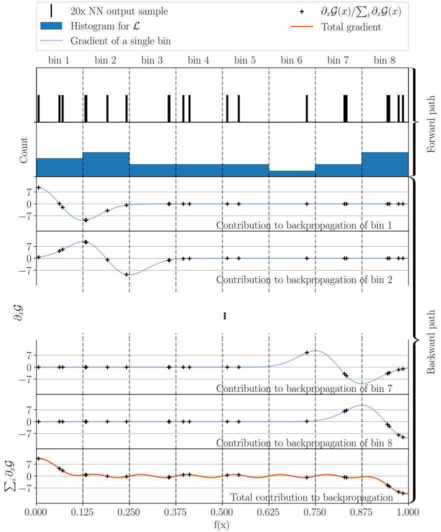
- Where $V_{11} = \sigma_\mu^2$ contains statistical and systematic uncertainties of μ

(Un)differentiable histograms

- $\mathcal{L}\left(N, \mu, \{\theta_j\}\right)$ is calculated on binned data
- NN backpropagation requires each step in the calculation of of the loss function to be differentiable
- Histogram gradient, described by delta functions at bin edges lacks continuity.

Replacement of histogram gradient necessary

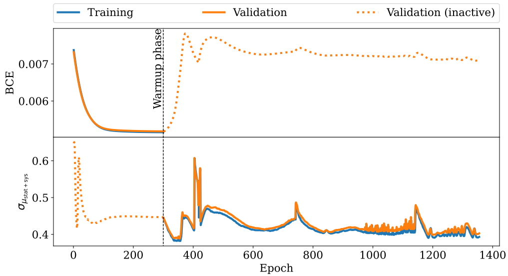
- Previously proposed gradient replacement [1]: Gaussian derivative for each bin
- Additional introduction of warm-up phase based on BCE
  - Statistical part of σ_μ can be reduced by a spatial separation of signal and background

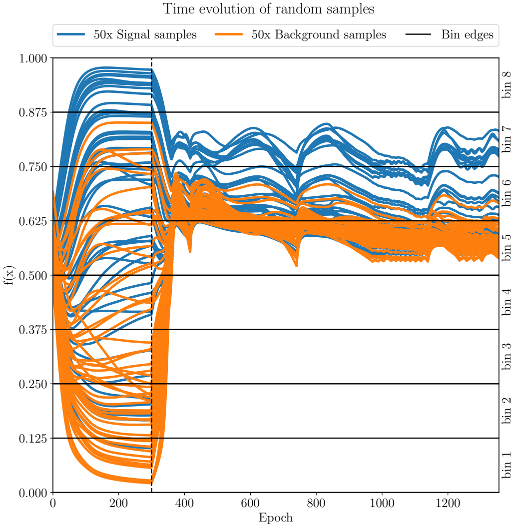


Evolution of NN output with proposed custom gradient

- Collapse of NN output function into 1-3 bins
  - Independent from concrete warm-up
  - More pronounced feature for more complicated tasks.
- Collapse is not improving the loss

→ **Unstable training, convergence is not ensured**





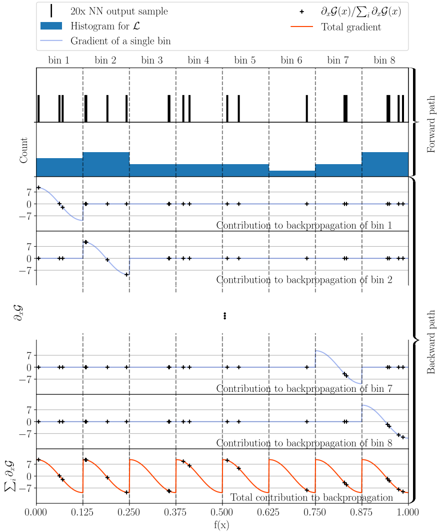
Improved custom histogram gradient

- Restrict Gaussian derivative to the respective bin

  $\rightarrow$ Removes long range effects across bins, which lead to low gradient amplitudes everywhere except the outer most bins

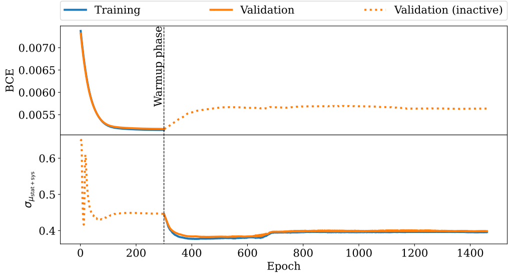
  $\rightarrow$ "Movement directive within a bin" rather than a Gaussian "smearing of a bin"

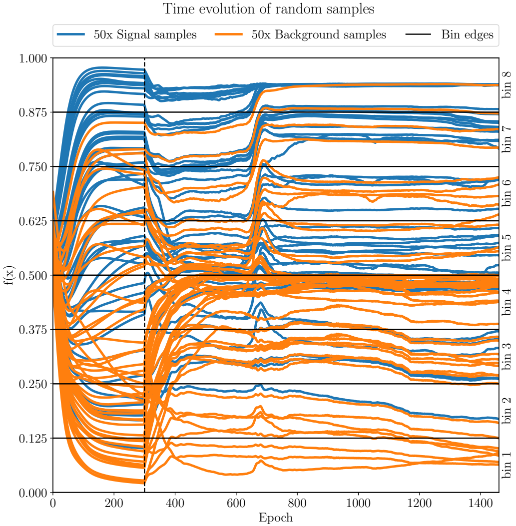
- Further adjustments to the training procedure:
  - Increase of learning rate
  - Change optimizer from Adam to NAdam



Evolution of NN output with modified custom gradient

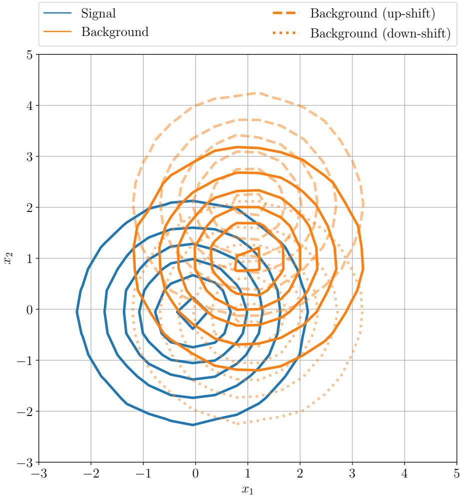
- Reduced event aggregation into fewer bins
- Improved training convergence





Demonstration on a binary toy model

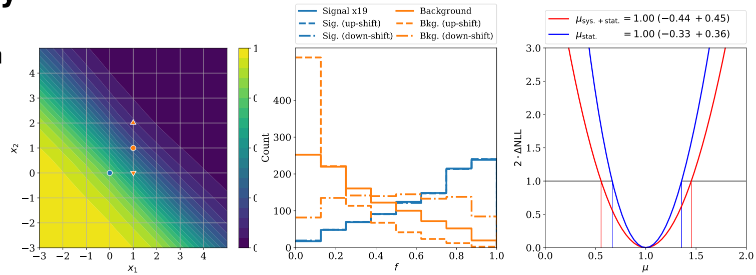
- Signal (background) modeled as 2D - Gaussians with 10^5 samples each, reweighted to 50 (1000) events for the final inference
- Systematic variation: $x_2 \pm 1$ (dashed lines in Figure)
- Test (training, validation) data sets: $2 \cdot 10^5$ (10^5, 10^5) independent events
- Fully connected feed forward NN
  - Input: x_1, x_2
  - One hidden layer with 100 nodes and ReLU activation
  - One output node with sigmoid activation
- Full-batch training, 1000 epochs patience on validation loss



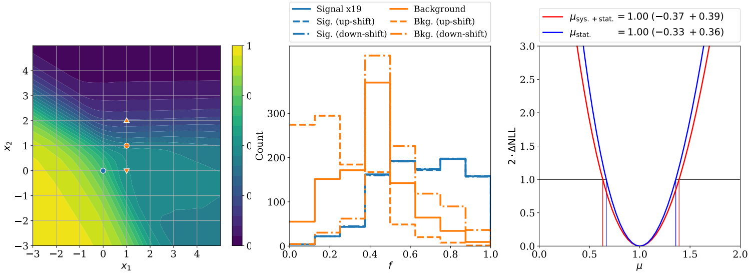
Karlsruhe Institute of Technology

Results of binary toy study

- Training on BCE: Spatial separation of processes in value space of NN output leads to reduction of statistical part of σ_μ
- Effect of systematic variation is clearly visible (as expected)



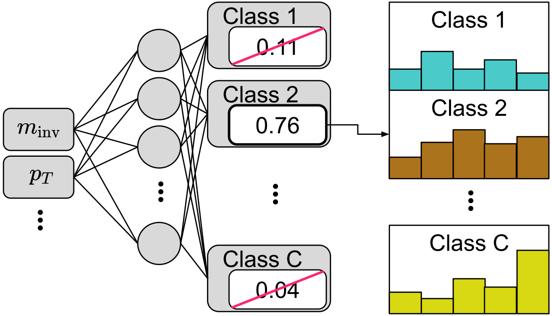
- Training on σ_μ reduces the combined uncertainty by $\approx 20\,\%$ relative to BCE based training
- BCE warm-up improves the process separation in cases of dominating statistical uncertainty



Idea of multi-class classification

- Assignment of signal/background processes to (different) classes
- Realization use multiple output nodes
- Activation function in output layer: Softmax

  $\rightarrow$ Probabilistic interpretation of the likelihood to find a corresponding event in a given class when using CE loss



- Modifications to final inference

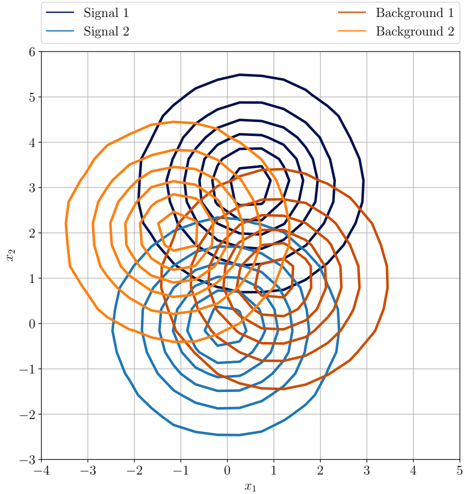
$$\mathcal{L}\left(N,\mu,\{\theta\}\right) = \prod_{\text{class } c=1}^{C} \prod_{\text{bin } i=1}^{N} \mathcal{P}\left(n_i \middle| \sum_{k=1}^{P} \mu_{k} s_{i,k}\left(\{\theta_j\}\right) + \sum_{k'=1}^{P} b_{i,k'}\left(\{\theta_j\}\right)\right) \prod_{j=1}^{M} \mathcal{C}\left(\theta_j \middle| \mu_{\theta_j}, \sigma_{\theta_j}\right)$$

Extension of the toy data set to multi-class classification

- Introduction of a second
  - background at $(-1, 2)$
  - signal at $(0.5, 3)$

- signal (background) processes are reweighted to 100 (1000) events each

- Used uncertainties
  - Background 1: $x_2 \pm 1$ (as before)
  - Background 2: $x_1 \pm 1$

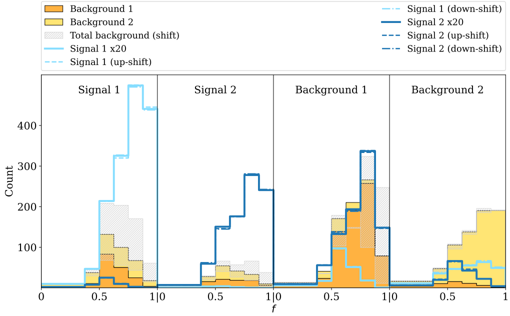


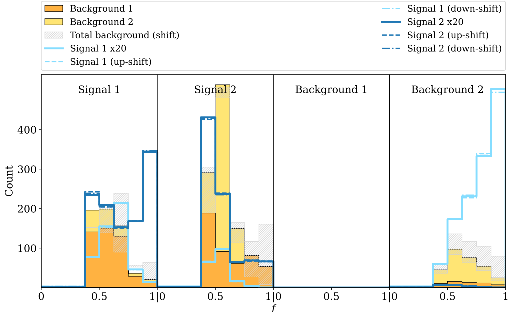
Multi-class classification: Problem

Expectation: Minor adjustments after CE warm-up to address systematic uncertainties, but:

- Classes act only as additional bins
- No penalty for misclassification in the training on σ_μ and final inference
- Predefined classes are not used during training as intended

$\rightarrow$ With increasing problem complexity: Empty classes and misclassified events within a class occurs





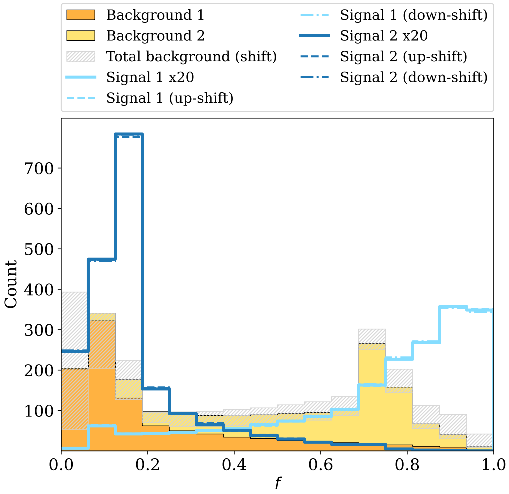
Applicable multi-class classification: Ansatz 1

- Not all NN classification information is used for loss calculation

Ansatz 1:

- Use only one ”class”
  - Change (back) to one output node with Sigmoid activation
  - Increase number of bins
- Separation of signal processes as due to $\sum_i \sigma_{\mu_i}$ minimization

$\rightarrow$ **Creation of regions with accumulated signal processes**



Applicable multi-class classification: Ansatz 2

Ansatz 2:

Preservation of class assignments by performing weight optimization in constrained phase space

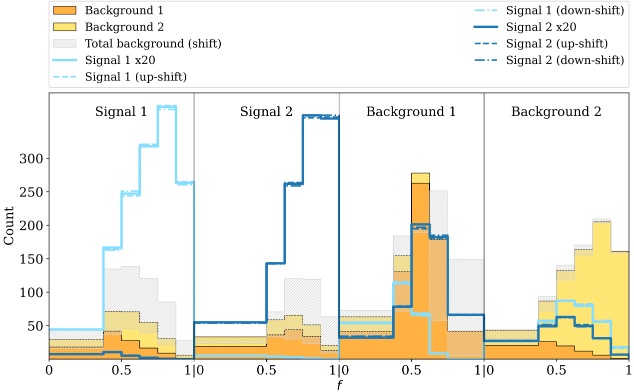
$\rightarrow$ Output nodes activation function: Sigmoid

$\rightarrow$ Construction of a loss with a penalty term

$$\text{Loss} = \sigma_\mu + \lambda \left(\mathrm{L}_{\mathrm{BCE}} - \mathrm{L}'_{\mathrm{BCE}} \right)$$

where

- $\mathrm{L}'_{\mathrm{BCE}}$ is the BCE loss as the end of the warm-up phase,
- $\mathrm{L}_{\mathrm{BCE}}$ the BCE loss of the current epoch and
- λ a learnable parameter in case of $\mathrm{L}_{\mathrm{BCE}} - \mathrm{L}'_{\mathrm{BCE}} > 0$ and 0 otherwise



$\rightarrow$ **Constraint preserves the initial classification during the σ_μ minimization**

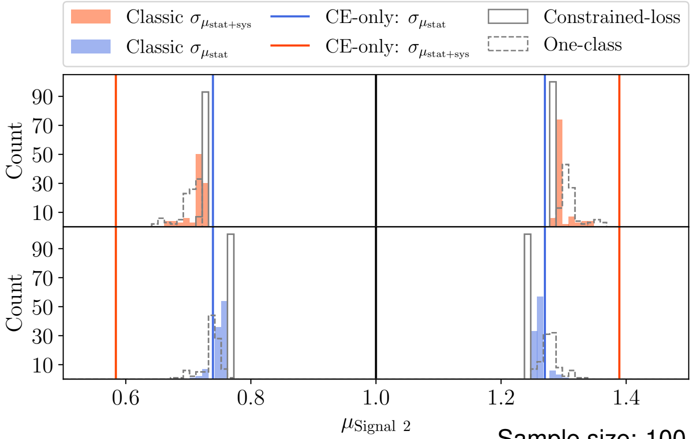
Ansatz comparison

One class:

- No additional hyper parameters
- Worse results compared to constrained approach
- With increasing number of signal processes separation worsens

Constrained loss approach:

- Provides stable and better results
- Separation still possible also with/despite of increasing number of processes
- Introduction of additional hyperparameters during training, which are not present in the final inference



Sample size: 100

Application to Standard Model $H \rightarrow \tau\tau$ analysis [3]

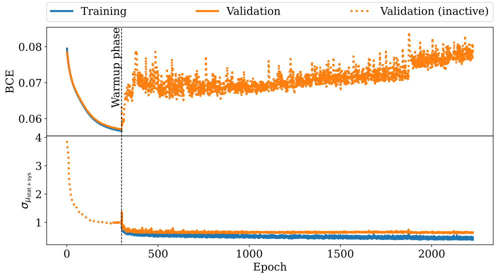
- Goal: Differential measurements of Higgs boson production
- $\mathrm{H} \rightarrow \tau\tau$: Highest branching ratio $(6.3\,\%)$ after $\bar{b}b$ but with less background contributions
- NN output is used for the final inference

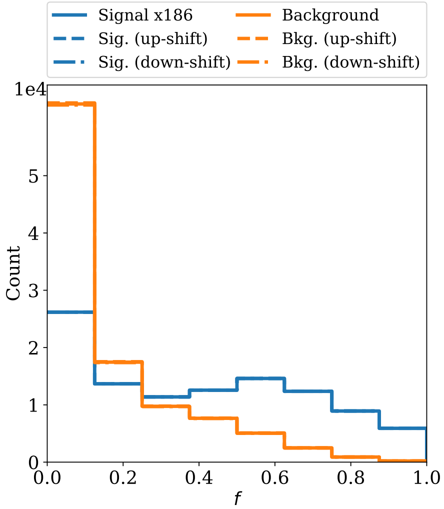
Restriction for this study:

- Final decay mode: $\tau_\mathrm{h}\tau_\mathrm{h}$, $\mu\tau_\mathrm{h}$, $\boldsymbol{\mathrm{e}\tau_\mathrm{h}}$, $\mathrm{e}\mu$
- Data set: 2016, $\mathbf{2017}$, 2018
- Using **86 systmatic variations in form of event weights** (no jer, jes...)

$\mathrm{H} \rightarrow \tau\tau$: Binary classification

- Combination of all background and signal processes correspondingly (inclusive measurement)
- All uncertainties as shape uncertainty
- Stable and converging training - comparable to toy study



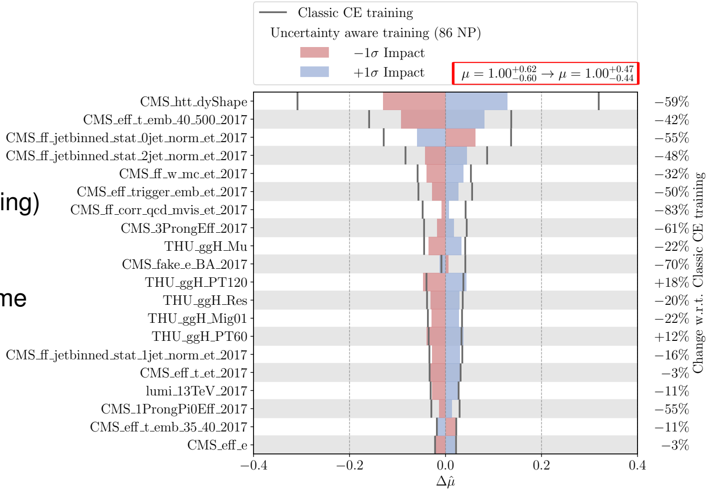


$H \to \tau\tau$: Results

uncertainty aware training reduces σ_μ by $\approx 25\,\%$

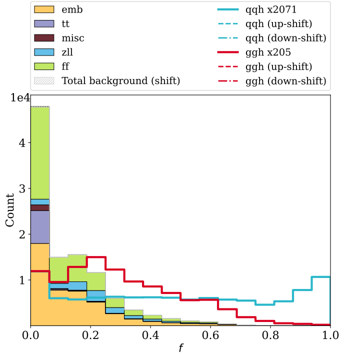
- Use of less uncertainties (i.e. 30 in training) already leads to similar reduction in σ_μ

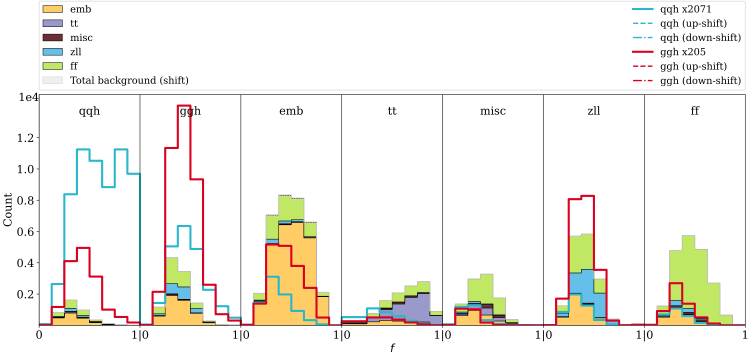
  $\to$ Potential reduction in computing time ($\approx 4\,\mathrm{h} \to \approx 1\,\mathrm{h}$) training time



$H \to \tau\tau$: Multi-class classification

- qqh and ggh Higgs boson production mechanisms as signal processes
- Separation of qqh events as a result of statistical uncertainty reduction (dominant uncertainty), larger confusion of ggh events with background processes



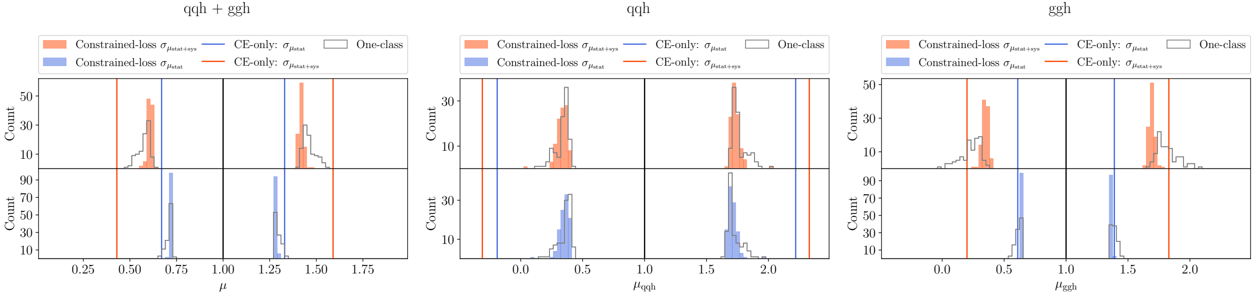


Ansatz comparison on $H \rightarrow \tau\tau$

Both approaches:

- Allow for differential measurements of the signal strength of selected Higgs boson production modes
- Are able so separate qqh better than classic CE-training based on the same NN architecture
- Avoid the problem of empty classes.

A constraint on σ_μ loss improves the result at the cost of introducing additional hyperparameters, which are not adressed or motivated in the final inference



Summary

- Improved stability during the training on σ_μ by modifying the custom histogram gradient
- Extension to uncertainty-aware multi-class classification
- Successfull application on a subset of the SM $\mathrm{H} \rightarrow \tau\tau$ analysis [3]

Outlook

- Apply to the full $\mathrm{H} \rightarrow \tau\tau$ analysis with more differentiable Higgs boson production process

- Incorporation of a method to avoid the use of histograms for uncertainty calculation
- Adaption of statistical inference to the multi-class classification based on uncertainty aware training

References

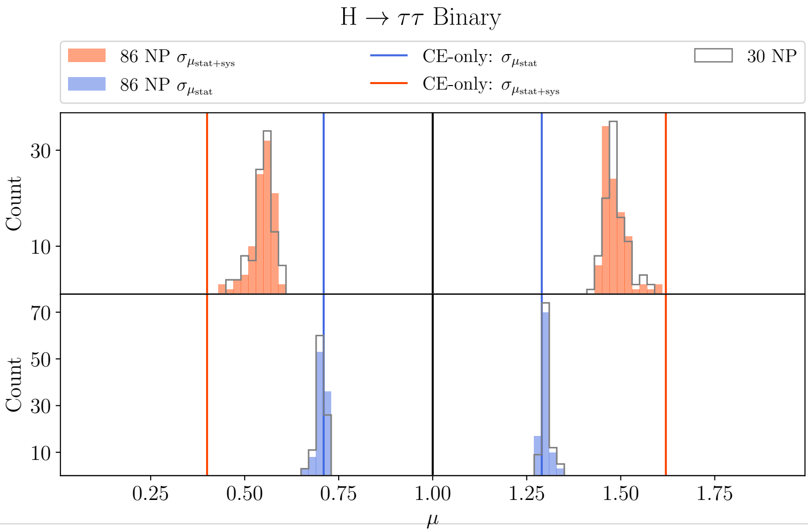
[1] Stefan Wunsch et al. “Optimal statistical inference in the presence of systematic uncertainties using neural network optimization based on binned Poisson likelihoods with nuisance parameters”. en. In: *Comput. Softw. Big Sci.* 5.1 (Dec. 2021).

[2] John Platt and Alan Barr. “Constrained Differential Optimization”. In: *Neural Information Processing Systems*. Ed. by D. Anderson. Vol. 0. American Institute of Physics, 1987. URL: https://proceedings.neurips.cc/paper/1987/file/a87ff679a2f3e71d9181a67b7542122c-Paper.pdf.

[3] CMS Collaboration. *Measurements of Higgs boson production in the decay channel with a pair of leptons in proton-proton collisions at $\sqrt{s}$ = 13 TeV*. 2022. DOI: 10.48550/ARXIV.2204.12957. URL: https://arxiv.org/abs/2204.12957.

Backup

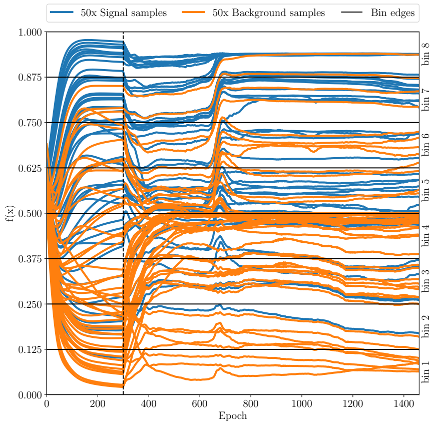
$\mathrm{H} \rightarrow \tau\tau$ Binary: Reduced number of NP



$\mathrm{H} \rightarrow \tau\tau$ Used training variables

- $p_T(\tau_1)$
- $p_T(\tau_2)$
- m_{vis}
- $p_{T_{\mathrm{vis}}}$
- $m_{\mathrm{sv}_{\mathrm{Puppi}}}$
- N_{Btag}
- $p_T(j_1)$
- N_{Jet}
- $\Delta\eta_{jj}$
- m_{jj}
- $\mathrm{MELA}_Q^2(V_1)$
- $p_T(jj)$
- $\mathrm{MELA}_Q^2(V_2)$
- $p_T(j_2)$
- $\Delta R_{\tau_1\tau_2}$

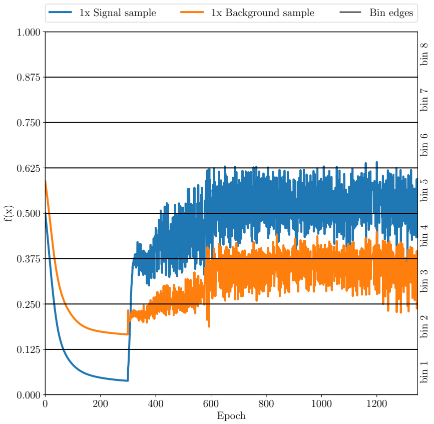
Impact of Further adjustments to the training procedure



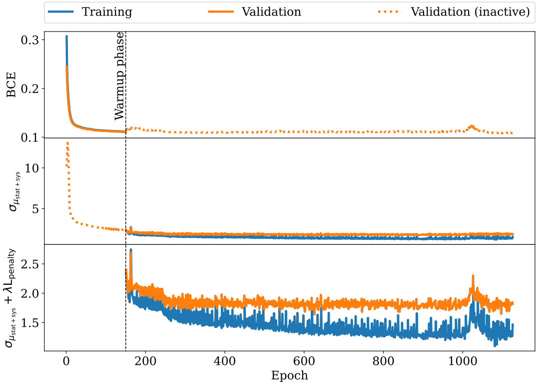
Applied Change:

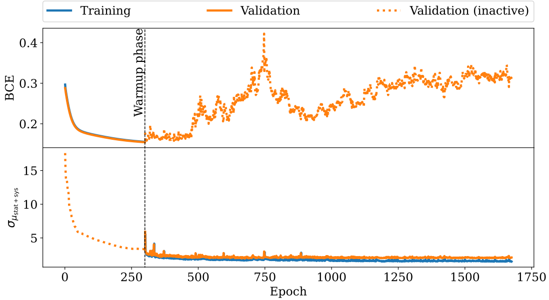
- increased learning rate
- Adam $\rightarrow$ NAdam

A change in σ_μ only occurs when minimum one event change the bin

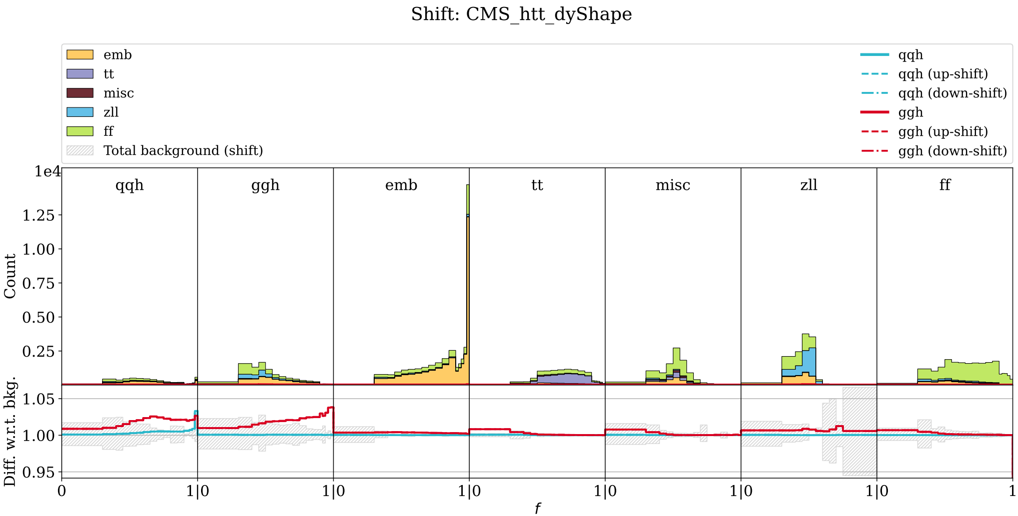


Loss evolution: One class and Constrained-loss

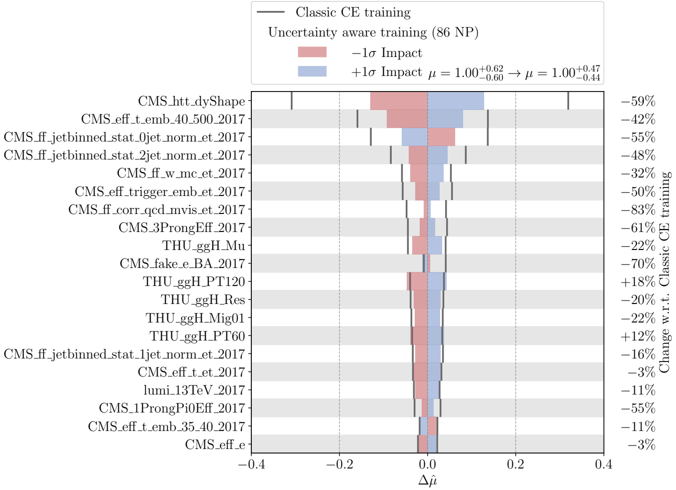




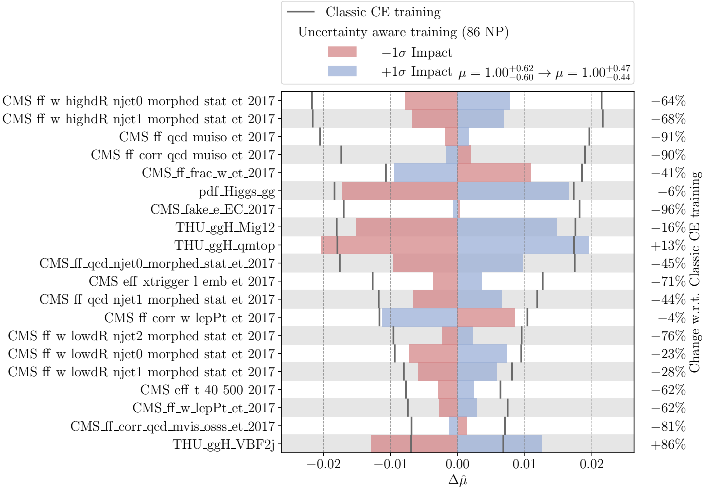
$\mathrm{H} \rightarrow \tau\tau$ Multi-class classification: CE benchmark, exemplary shift



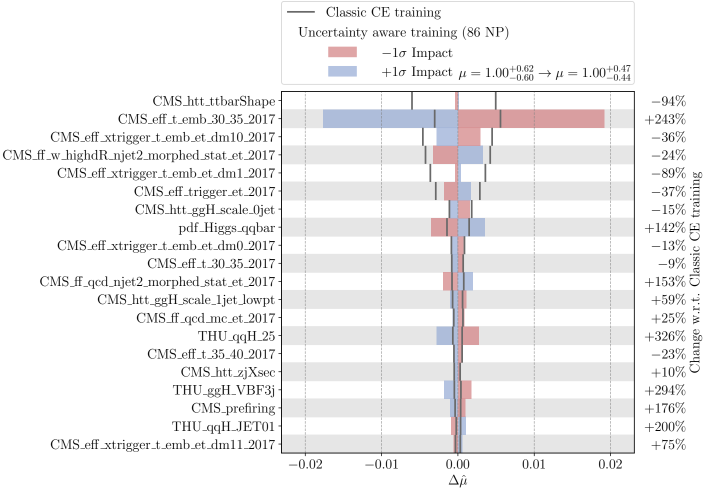
List of used systematic uncertainties 1/5



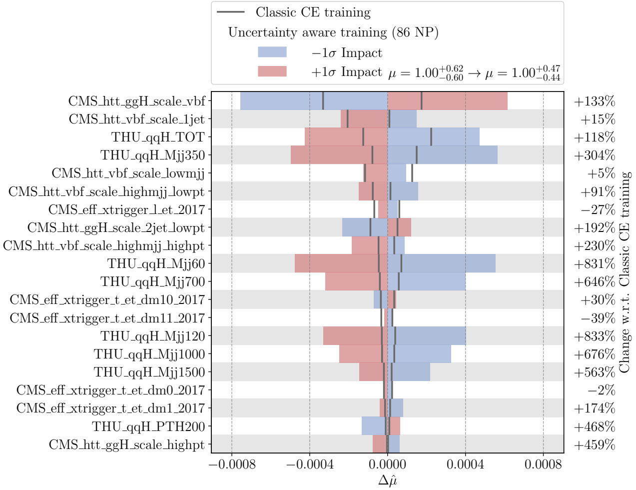
List of used systematic uncertainties 2/5



List of used systematic uncertainties 3/5



List of used systematic uncertainties 4/5



List of used systematic uncertainties 5/5