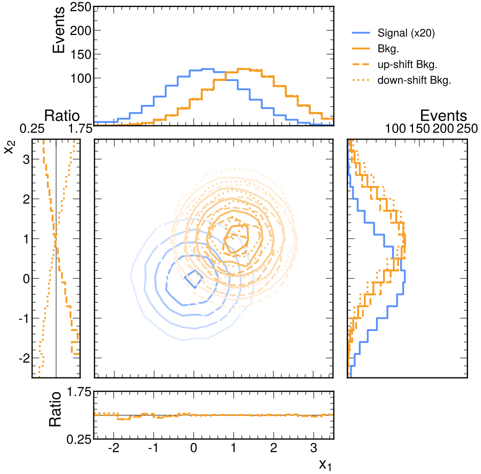
Development of systematic-aware neural network trainings for binned-likelihood-analyses at the LHC

Based on CMS-PAS-MLG-23-005

11. 09. 2024

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

Starting point: a synthetic example



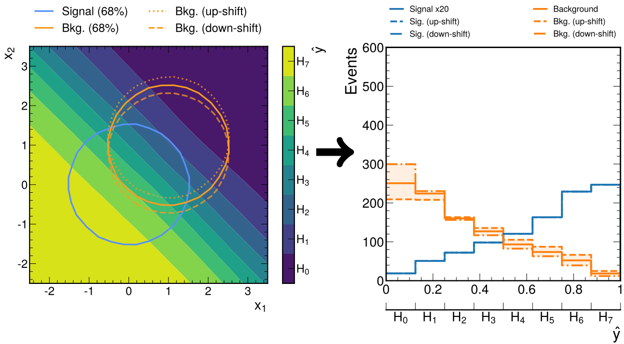
Objective: signal strength $r_s \pm \Delta r_s$ in a given model

Common problems for classification tasks (i.e. left):

- High imbalance between process yields
- Processes overlap in some phase space regions
- Processes are (usually) affected by uncertainties

$\rightarrow$ **Utilizing NN models for process classification**

NN models for conventional process separation



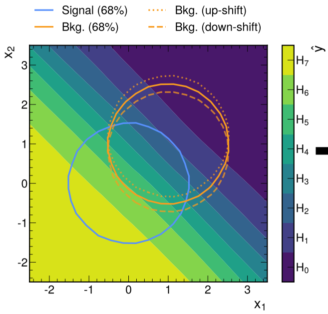
Cross entropy NN training (CENNT)

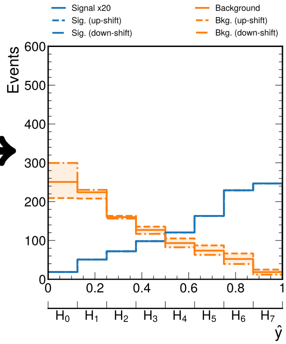
- Training objective: process separation
- Utilization of nominal signal (S) and background (B) dataset

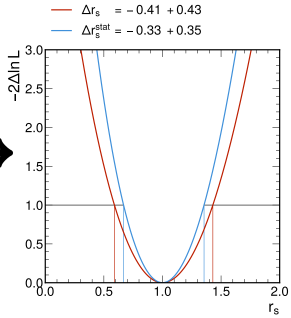
Resulting NN output

- Nominal S and B datasets
- Systematic variations ΔS_j, ΔB_j for each uncertainty source j

$r_s \pm \Delta r_s$



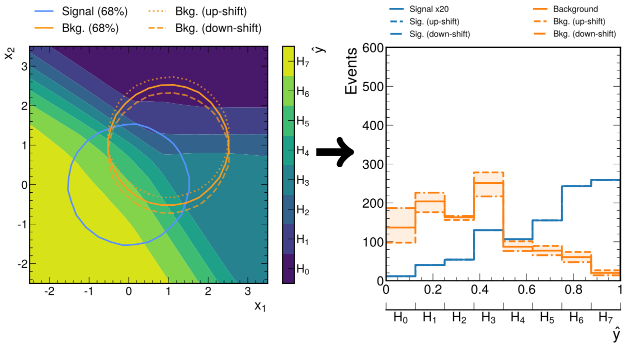




Final results obtained from statistical inference based on binned likelihood

Additional information about systematic variations

NN model for Δr_s minimization



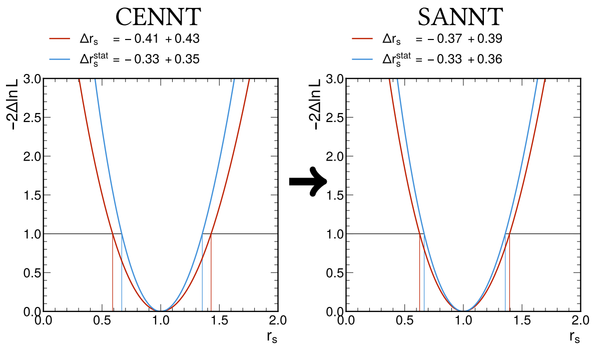
Systematic aware NN training (SANNT)

- Training objective: Δr_s minimization, now aligns with analysis objective
- Not a process separation anymore: High S/B bins creation with low syst. variations

Resulting NN output

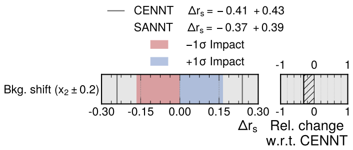
- Nominal S and B datasets
- Systematic variations ΔS_j, ΔB_j for each uncertainty source j

Consequences for Δr_s



- Statistical uncertainty comparable
- Redistribution of events reduces impact of uncertainty on r_s

  $\rightarrow$ Reduction of overall uncertainty



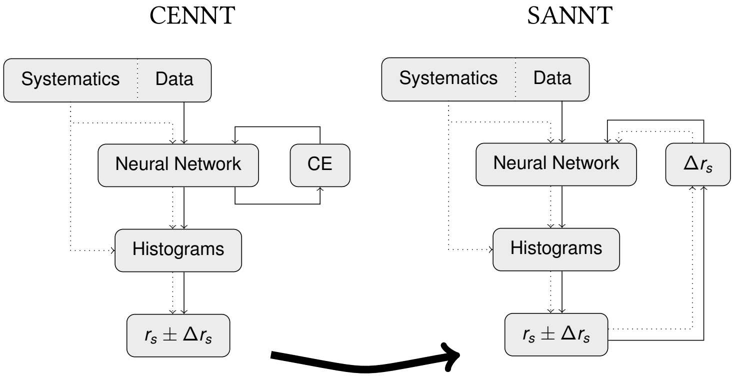
Changes to the training procedure summarized

Use CE pretraining achieving:

- Process separation
- Good starting point for SANNT

Main difference: SANNT vs. CENNT :

- Changed training objective
- Added information about systematic variations to the training



Δr_s as training objective

1. Modification to existing likelihood $\mathcal{L}\left(\{k_i\}, \{r_s\}, \{\theta_j\}\right)$ (backup):
   - Extend λ_i in $\mathcal{P}\left(k_i, \,|\, \lambda_i,\right)$ to $\lambda_i' = \sum_s r_s S_{si} + \sum_b B_{bi} + \tilde{\Delta}_i$ with systematic shifts $\tilde{\Delta}_i$, set $\theta_j = 0 \,\forall\, j$ and use Asimov dataset, replacing $k_i \to k_i^{\mathrm{A}}$

$$\tilde{\Delta}_i = \sum_{p \in \{s,b\}} \sum_j \max\left(0, \theta_{jp}\right) \left(\Delta_{jp}^{\text{up}}\right)_i + \min\left(\theta_{jp}, 0\right) \left(\Delta_{jp}^{\text{down}}\right)_i \,,$$

$\to$ Build computational graph of effects of syst. variations effects for each process $p \in \{s, b\}$
$\to$ Including asymmetries at $\theta_k^{\pm} \to 0$ for each deviation from nominal

Δr_s as training objective

1. Modification to existing likelihood $\mathcal{L}\left(\{k_i\}, \{r_s\}, \{\theta_j\}\right)$ (backup):
   - Extend λ_i in $\mathcal{P}\left(k_i, \middle| \lambda_i,\right)$ to $\lambda_i' = \sum_s r_s S_{si} + \sum_b B_{bi} + \tilde{\Delta}_i$ with systematic shifts $\tilde{\Delta}_i$, set $\theta_j = 0 \,\forall\, j$ and use Asimov dataset, replacing $k_i \to k_i^{\mathrm{A}}$

2. Asymptotically estimate Δr_s from $\sqrt{F^{-1}_{r_s, r_s}}$ using full, weighted training/validation dataset

$$F_{x_i x_j} = \mathbb{E}\left[\frac{\partial^2}{\partial x_i \partial x_j}\Big(-\log \mathcal{L}\Big)\right]_{x_i\, x_j \in \{\{r_s\}, \{\theta_j\}\}} = \Big(\mathrm{Hess}\Big(-\log \mathcal{L}\Big)\Big)_{x_i\, x_j \in \{\{r_s\}, \{\theta_j\}\}}$$

For n signals: sum over n diagonal elements of $\sqrt{F^{-1}_{r_s, r_s}}$

Caviat: Backpropagation

Training should have a **non breaking backpropagation**

- Every computation step of Δr_s must be differentiable
- Gradient for histograms is usually not provided (i.e. PyTorch, TensorFlow)

Approach: introduce bin-wise **custom functions** $\mathcal{B}_i$ replacing histogram gradient (adapted from [1])

- Histogram remains unchanged in the forward pass
- Use $\mathcal{B}_i$ only during backward pass
- Choice of $\mathcal{B}_i$ has an effect on training procedure

$\rightarrow$ Improvement of $\mathcal{B}_i$ can enable application to more complex tasks

Effects of systematic aware training

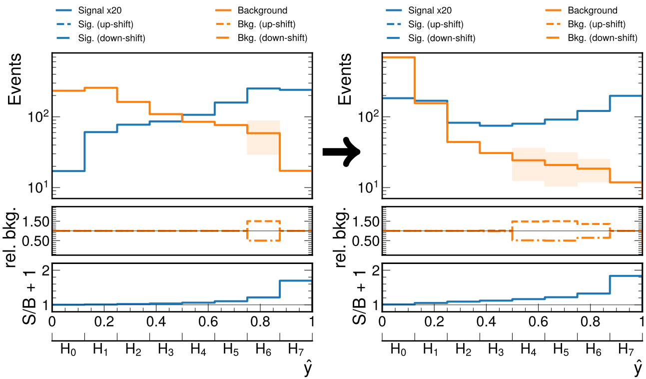
Visualizing effects of systematic variations

Injecting uncertainty for B in bin 6 after pretraining (left)

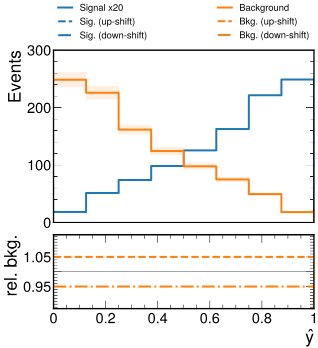
After subsequent SANNT training (right):

- S events
  - Approx. maintaned in S-enriched regions
- B events
  - Further removed from S-enriched region mostly to the left in the histogram
- Unct. affected B events moved to the left

$\rightarrow$ Effect of unct. is reduced



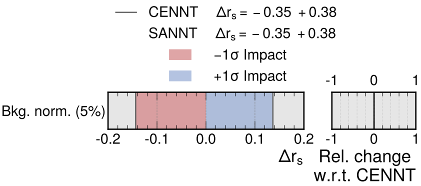
Systematic variations that can not be addressed



- Best process separation achieved after pretraining
  $\rightarrow$ lowest possible Δr_s^{stat} achieved

In case of pure normalization uncertainties:

- Events are present in all bins and are equally affected
  $\rightarrow$ no phase space that could lead to improvement of Δr_s w.r.t. Δr_s^{stat}



Systematic variations that should not be considered

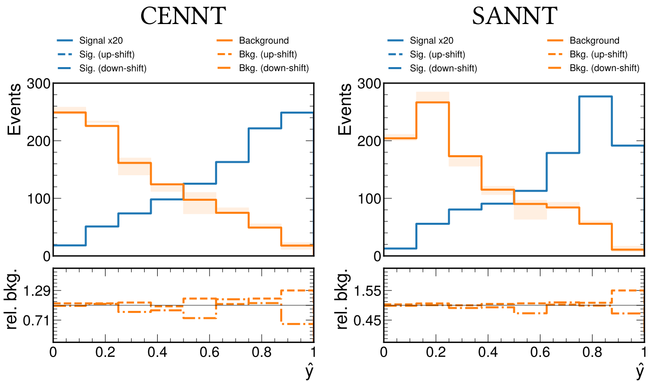
Low statistics uncertainty sample

- Downsampled B events from 5 % normalization shift (previous slide)

Fluctuating results are retrieved in both cases

SANNT will pick up the fluctuations and try to minimize their effect on Δr_s!

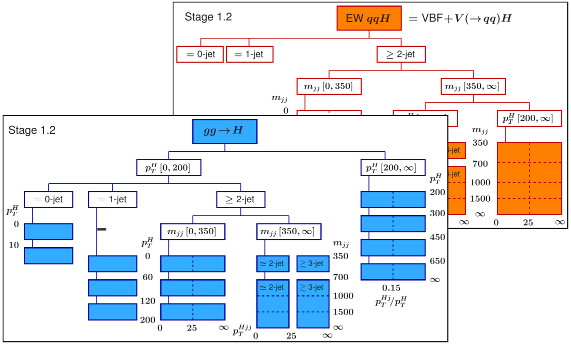
$\longrightarrow$ **Good uncertainty model description remains essential here**



Application on reduced $H(125) \rightarrow \tau\tau$ analysis

ML-based $\mathrm{H}(125) \rightarrow \tau\tau$ analysis (HIG-19-010, [2])

- **Differential cross-section** measurement of $\mathrm{H}(125)$ production based on STXS scheme
- Utilized multiclass classification with **5 background and up to 15 STXS signal classes**



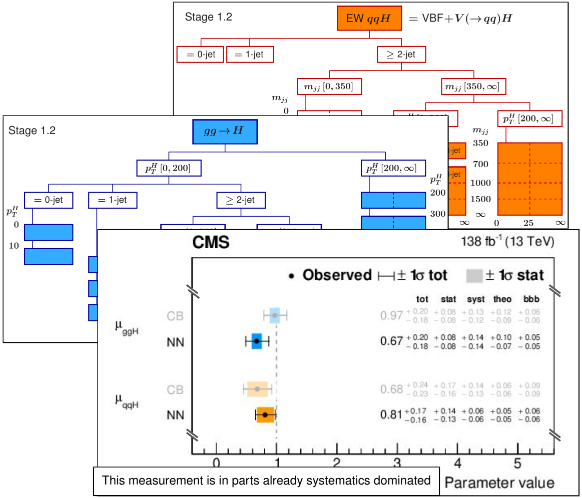
ML-based $\mathrm{H}(125) \rightarrow \tau\tau$ analysis (HIG-19-010, [2])

Karlsruhe Institute of Technology

- **Differential cross-section** measurement of $\mathrm{H}(125)$ production based on STXS scheme
- Utilized multiclass classification with **5 background and up to 15 STXS signal classes**

Future prospects for this and other analyses:

- Statistical uncertainties will decrease (Run3 and HL-LHC)
- Importance of systematic uncertainties will increase
  $\rightarrow$ Addressing them will become more important



This measurement is in parts already systematics dominated

Setup overview of application on reduced HIG-19-010 analysis ([3])

Karlsruhe Institute of Technology

- Using a subset of the dataset used by [2]
  - Final state: $\mathrm{e}\mu$, $\mu\tau_\mathrm{h}$, $\mathrm{e}\tau_\mathrm{h}$, $\tau_\mathrm{h}\tau_\mathrm{h}$
  - Era: 2016, 2017, 2018
- Selecting 86 theoretical and experimental uncertainties

- NN input: 15 variables used as in [2]
- Full-batch training with evaluation based on two-fold cross validation scheme

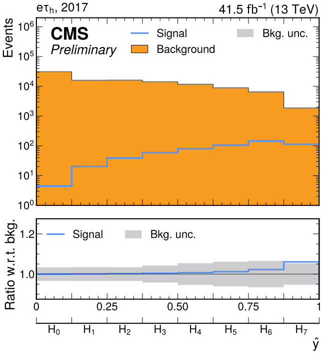
- Binary classification: all S_s (B_b) processes grouped
- Multiclass classification: 5 B and 2 S processes

Application on reduced analysis of HIG-19-010
Binary classification

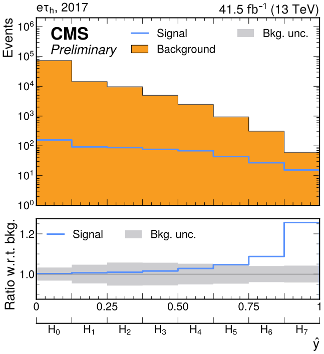
CENNT **vs.** SANNT: **Binary classification**

- Conceptual differences of SANNT:
  - S/B separation not the primary target
  - No relational bin information for SANNT: bin-wise ordering due to CE pretraining
- Improvements in process separation due to B reduction in S-enriched bins
- Largest total B unct. contributions move away from S-enriched bins

CENNT

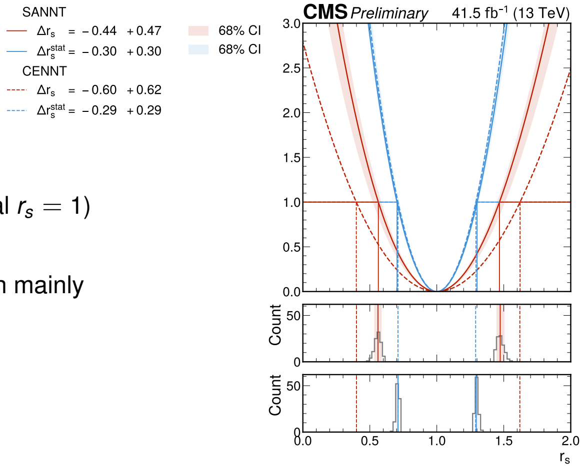


SANNT



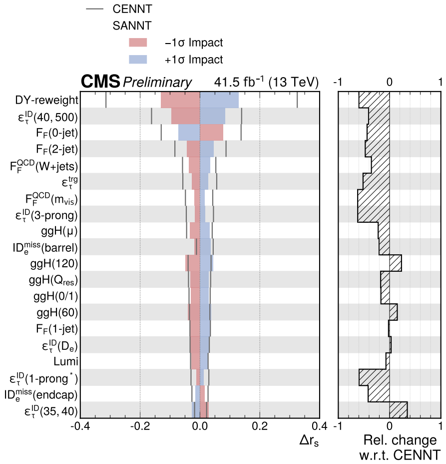
Results of an ensemble test

- Ensemble configuration:
  - Sample size: 100 repetitions
  - Changing NN weight initialization
  - Confidence interval derived from Δr_s
- Fit to Asimov data ($S + B$ model with SM-like signal $r_s = 1$)
- Median expectation of Δr_s w.r.t. CENNT reduction mainly by reducing systematic component of Δr_s



Comparison of uncertainty impacts

- Importance ordering as obtained after CENNT (Detailed description: backup)
- Reduction of uncertainties with largest impact on Δr_s
  $\rightarrow$ Comparable reduction of Δr_s can be achieved using only a subset of uncertainties with largest impact on Δr_s
- Note: Normalization uncertainties in binary classification can show a shape-changing effect if applied on a subset process of S, B



Application on reduced analysis of HIG-19-010
Multiclass classification

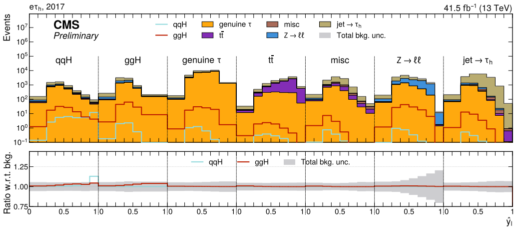
SANNT: **Multiclass classification**

- Considering uncertainty for two signal processes
  - Gluon fusion (ggH)
  - Vector boson fusion (qqH)
- No relational information to predefined classes nor bins
- Sustain concept of output classes

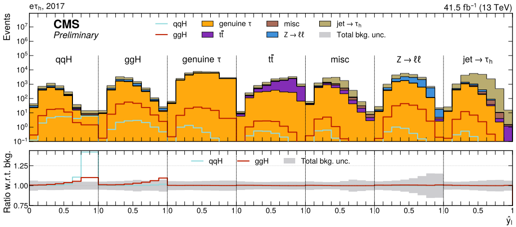
$$L_{\mathrm{SANNT}}^{\mathrm{mult.}} = \sum_{s} \Delta r_s + \omega_\lambda g\,(\,\cdot\,)\,, \text{ with}$$

  - $g\,(\,\cdot\,) = \max(\mathrm{CE}' - \mathrm{CE}'^{\,\mathrm{pretrain.}}_{\mathrm{min}}, 0)$
  - and learnable ω_λ [4]

CENNT



SANNT



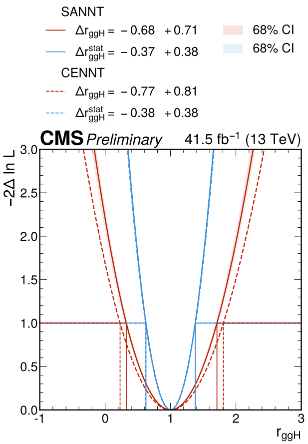
Uncertainty reduction for SANNT multiclass classification

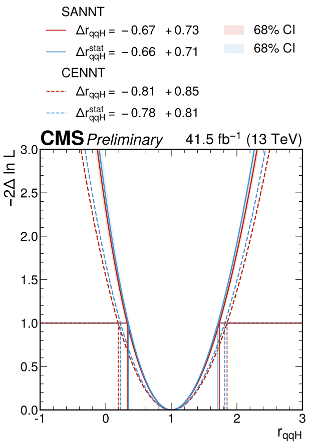
qqH has low process yield in comparison to ggH

- Δr_{qqH} still statistically dominated
- Reduction of Δr_{qqH} through minimization of $\Delta r^{\mathrm{stat}}_{\mathrm{qqH}}$
  $\rightarrow$ Improvement in process separation

ggH is more distributed across multiple classes

- Process enrichment in few bins more difficult
  $\rightarrow$ only minor improvement in $\Delta r^{\mathrm{stat}}_{\mathrm{ggH}}$
- Major reduction of Δr_{ggH} through reduction of systematic uncertainty contribution





r_{inc}: backup

Summary

Summary

- Fundamental SANNT studies:
  - General idea of functionality through synthetic examples
  - Effects of various uncertainty sources on training outcomes
  - Effects of poorly understood uncertainty models, regardless of applied method
- Application of SANNT:
  - Systematic NN training is applicable to real-life analysis tasks with considerably higher complexity
  - Extension to multiple classes while maintaining interpretability is possible
  - Significant sensitivity improvement over classical CE-based training using 86 uncertainties

References I

[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". In: Computing and Software for Big Science 5.1 (Jan. 2021), p. 4. ISSN: 2510-2044. DOI: 10.1007/s41781-020-00049-5. URL: https://doi.org/10.1007/s41781-020-00049-5.

[2] Armen Tumasyan et al. "Measurements of Higgs boson production in the decay channel with a pair of τ leptons in proton–proton collisions at $\sqrt{s} = 13$ TeV". In: Eur. Phys. J. C 83.7 (2023). All the figures and tables, including additional supplementary figures and tables, can be found at http://cms-results.web.cern.ch/cms-results/public-results/publications/HIG-19-010 (CMS Public Pages), p. 562. DOI: 10.1140/epjc/s10052-023-11452-8. arXiv: 2204.12957. URL: https://cds.cern.ch/record/2807752.

[3] Development of systematic-aware neural network trainings for binned-likelihood-analyses at the LHC. Tech. rep. Geneva: CERN, 2024. URL: http://cds.cern.ch/record/2905411.

References II

[4] 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.

Backup

Likelihood formulation

go back

- All histogram bins (i) of all classes (c) enter as inputs to an extended binned likelihood function $\mathcal{L}$, including all nuisance parameters (θ_j)

$$\mathcal{L}\left(\{k_i\},\{r_s\},\{\theta_j\}\right)=\prod_c\left[\prod_i\mathcal{P}\left(k_i|\lambda_i\right)\prod_j\mathcal{C}_j\left(\tilde{\theta}_j|\theta_j\right)\right]_c$$

$$\lambda_i=\sum_s r_s S_{si}\left(\{\theta_j\}\right)+\sum_b B_{bi}\left(\{\theta_j\}\right)$$

with

- Poisson distribution $\mathcal{P}\left(\cdot\,|\,\cdot\right)$ constructed from k_i observed and λ_i expected events per bin i
- Use $S+B$ model with scaling parameter(s) r_s
- Uncertainties are taken into account in the form of nuisance parameters θ_j, following predifined pdf's $\mathcal{C}_j\left(\cdot\,|\,\cdot\right)$

- Estimation of the uncertainty Δr_s on r_s through Asimov dataset $D_H^{\mathcal{A}}$: $k_i\rightarrow k_i^{\mathrm{A}}$

Choice of $\mathcal{B}_i$

go back

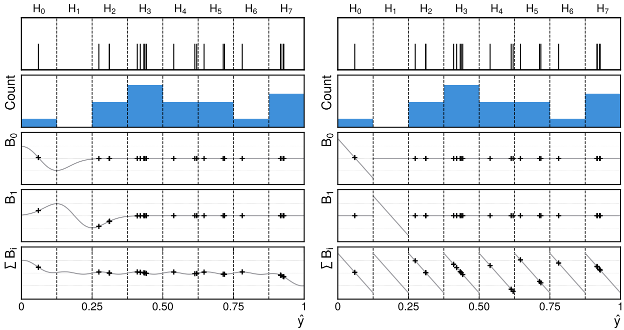
From [1] (left): Derivative of Gaussian PDF

- Low Gradient amplitude at center of bins
- Observing undesired concentration of events in very few bins in H($\hat{y}$)

  $\rightarrow$ Hard to apply to more complex tasks

Modification (right): linear function within bin boundaries

- Maintains injectivity
- Restricts range to corresponding bin H_i
- Keeps bins indistinguishable



Event evolution during training

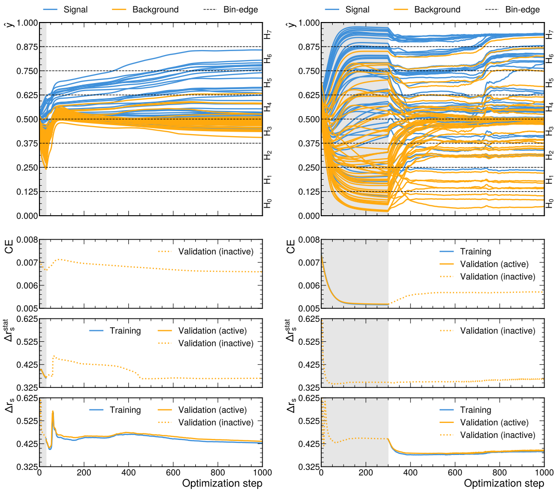
Using $\mathcal{B}_i$ from [1] with same setup ($\Delta x_2 = \pm 1$)

- Strong concentration in $\hat{y}$
- Independent from choice of pretraining

Pretraining change to CE:

- Equivalent to Δr_s^{stat} minimization
- Providing better starting point for minimization of systematic component of Δr_s

Identity operation (STE) for $\mathcal{B}_i$: similar behaviour to [1]



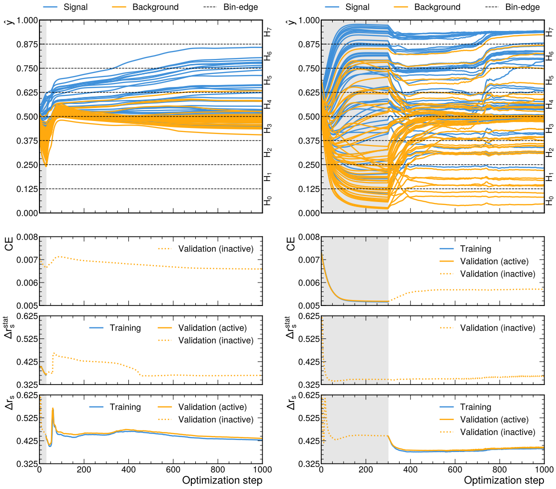
go back

Choice of $\mathcal{B}_i$: evolution during training

go back

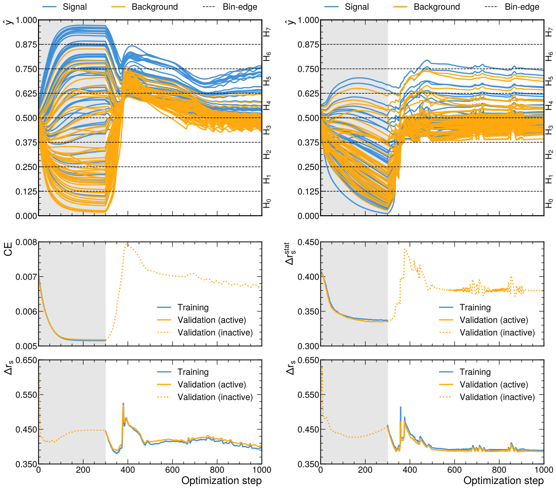
Difference between CE and $\Delta r_s^{(\text{stat})}$

Due to the product in $\mathcal{L} = \prod_i \mathcal{P}(k_i|\lambda_i)$ there is no sequential bin information when minimizing $\Delta r_s^{(\text{stat})}$



Choice of $\mathcal{B}_i$: Evolution during training - different pretraining

[go back](#)

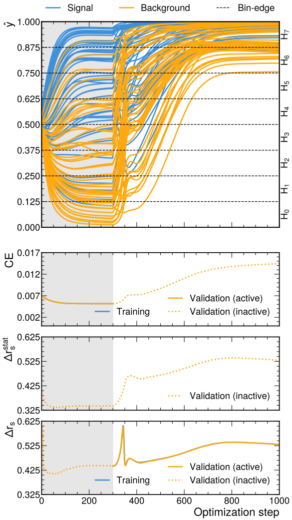


Derivative of Gaussian PDF as choice for $\mathcal{B}_i$

- Collapse of $\hat{y}$ into single bins independent of pretraining loss (CE or $\Delta r_s^{\mathrm{stat}}$) or duration (here 300 optimization steps)
- Collapse still prominent, $\hat{y}$ is less spread after pretraining/with shorter pretraining duration
- Best result achieved here (left) for 300 optimization steps of CE as pretraining: during collapse phase

Evolution using Idendity operation (STE) as $\mathcal{B}_i$

go back



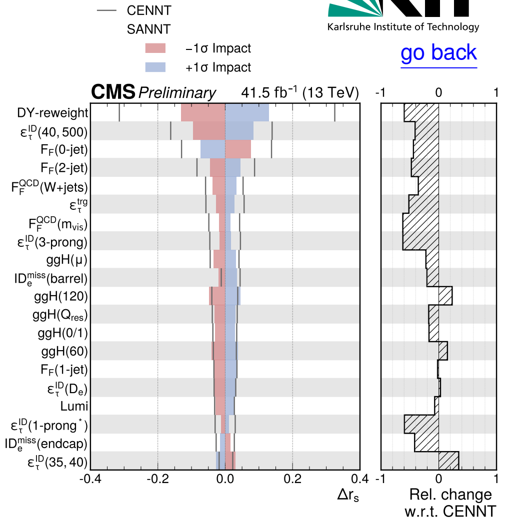
- Using identity for $\mathcal{B}_i$ leads to same problem as $\mathcal{B}_i$ as proposed in [1]
- Bins where the collapse into single bins occurs vary, depending on the seed used for weight initialization

Description of most impactful uncertainties (binary classification)

Karlsruhe Institute of Technology

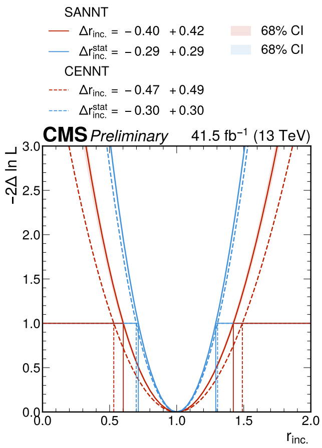
go back

Label	Type	Process	Rank	Norm	Shape	Comment
$\epsilon_{\tau}^{\mathrm{trig}}$	τ-Trigger	EMB	6	–	✓	–
$\epsilon_{\tau}^{\mathrm{ID}}(D_{e})$	τ-ID	MC, EMB	16	✓	–	Disrc. against e
$\epsilon_{\tau}^{\mathrm{ID}}(35, 40)$	τ-ID	EMB	20	–	✓	$35 < p_{\mathrm{T}}^{\tau_{\mathrm{h}}} < 40$ GeV
$\epsilon_{\tau}^{\mathrm{ID}}(40, 500)$	τ-ID	EMB	2	–	✓	$40 < p_{\mathrm{T}}^{\tau_{\mathrm{h}}} < 500$ GeV
$\epsilon_{\tau}^{\mathrm{ID}}(\text{1-prong}^{*})$	τ-ID	EMB	18	–	✓	One $\pi^{+}+\pi^{0}$'s
$\epsilon_{\tau}^{\mathrm{ID}}(\text{3-prong})$	τ-ID	EMB	8	–	✓	Three π^{+}'s
$F_{\mathrm{F}}(\text{0-jet})$	Norm.	F_{F}	3	–	✓	$N_{\mathrm{jet}} = 0$
$F_{\mathrm{F}}(\text{1-jet})$	Norm.	F_{F}	15	–	✓	$N_{\mathrm{jet}} = 1$
$F_{\mathrm{F}}(\text{2-jet})$	Norm.	F_{F}	4	–	✓	$N_{\mathrm{jet}} = 2$
$F_{\mathrm{F}}^{\mathrm{QCD}}(m_{\mathrm{vis}})$	Non-closure	F_{F}	7	–	✓	In m_{vis}
$F_{\mathrm{F}}^{\mathrm{QCD}}(\mathrm{W{+}jets})$	Subtr.	F_{F}	5	–	✓	Subtr. of MC
$\mathrm{ggH}(\mu)$	Theory	ggH	9	–	✓	μ_r and μ_f
$\mathrm{ggH}(Q_{\mathrm{res}})$	Theory	ggH	12	–	✓	Resummation
$\mathrm{ggH}(0/1)$	Theory	ggH	13	–	✓	$0 \rightarrow 1$ jet migr.
$\mathrm{ggH}(60)$	Theory	ggH	14	–	✓	$p_{\mathrm{T}}^{\mathrm{H}}$ migr.
$\mathrm{ggH}(120)$	Theory	ggH	11	–	✓	$p_{\mathrm{T}}^{\mathrm{H}}$ migr.
$\mathrm{ID}_{\mathrm{e}}^{\mathrm{miss}}(\mathrm{barrel})$	e-miss-ID	MC	10	–	✓	Barrel
$\mathrm{ID}_{\mathrm{e}}^{\mathrm{miss}}(\mathrm{endcap})$	e-miss-ID	MC	19	–	✓	Endcap
DY-reweight	Reweight	MC	1	–	✓	In $p_{\mathrm{T}}^{\mu\mu}$ and $m_{\mu\mu}$
Lumi	Luminosity	MC	17	✓	–	–



Uncertainty reduction for multiclass classification $r_{\mathrm{inc.}}$

go back



Differences to CENNT: In multiclass classification case

go back

Effects on NN output

- Conceptually:
  - No relational information about predefined classes (nor bins)
  - Given Δr_s or $\sum_s \Delta r_s$ no penalty for events that swap predefined classes
    $\rightarrow$ may lead to emergence of empty classes (less interpretable)

$\rightarrow$ Introduce a penalty term modifying the loss to: $L^{\mathrm{mult.}}_{\mathrm{SANNT}} = \sum_s \Delta r_s + \omega_\lambda g(\,\cdot\,)$ as introduced in [4]

  - Aim is to sustain the concept of output classes
  - $g(\,\cdot\,) = \mathrm{CE}' - \mathrm{CE}'_{\mathrm{min}}$ ensures a class assignment, that corresponds to $\mathrm{CE}'_{\mathrm{min}}$ as obtained from pretraining
  - ω_λ updated at each optimization step (using backpropagation) and set to 0 if $g(\,\cdot\,) < 0$ (improvement w.r.t. $\mathrm{CE}'_{\mathrm{min}}$)

- Technical differences:
  - Different activation function (Sigmoid vs. Softmax)
  - In case of SANNT: modified (binary) CE' for pretraining and as constraint

Related work

Estimation of statistical quantities from binned variables (in HEP) is often necessary due to

- Computational cost
- Availability of corrections/shifts

Transition to NN application requires differentiable histogram operation.

The two solutions to which this work is most comparable are (also mentioned in the paper):

`INFERNO`[1]: Histogram approximation trough tunable softmax operator

- Training objective: Diagonal element of inverse Fisher-Information matrix (Δr_s)
- Uses mini-batches during training and explicit sampling of F_{x_i,x_j}

`neos`[2]: Histogram approximation trough kernel density estimation

- Training objective: likelihood ratio
- Uses mini-batches during training

---

[1] DOI: 10.1016/j.cpc.2019.06.007
[2] DOI: 10.1088/1742-6596/2438/1/012105

Downsampled normalization shift

go back