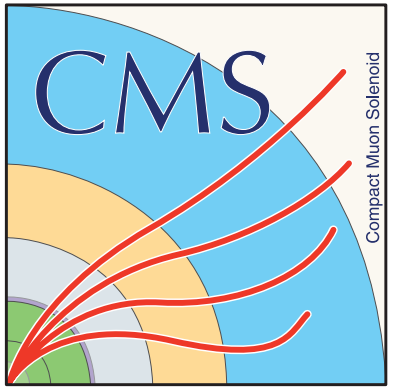
Using a Neural Network to Approximate the Negative Log Likelihood Function

Nathan Jamieson, Kevin Lannon, Shenghua Liu, Kelci Mohrman, Sirak Negash, Yuyi Wan, Brent Yates

for the CMS Collaboration

COLLEGE OF SCIENCE

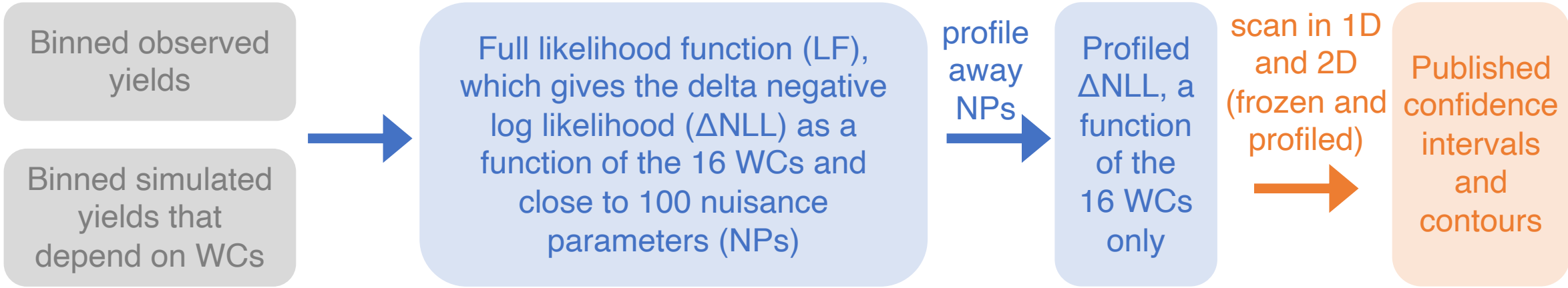
UNIVERSITY OF NOTRE DAME



Introduction

- This study builds on a previous CMS analysis [1], in which 16 Wilson Coefficients (WCs) in the framework of standard model effective field theory (SMEFT) were used to parameterize different types of new physics that could affect top quark production.

Data → Event Selection → Histograms → Statistical Analysis → Extracting Results



Limitation: Published 1D and 2D scans lose information about the WCs, and the other popular way to communicate LFs—the covariance matrix at a measured point—assumes Gaussianity.

[1]: https://doi.org/10.1007/JHEP03(2021)095

Introduction

- Goal of this study:
  - Following the proposal of efficiently distributing LFs as deep neural networks (DNNs) [2],
  - Train a DNN to learn the profiled ΔNLL
  - To produce a fast, differentiable, and portable approximation of the profiled ΔNLL with the NPs already profiled away—16 inputs (WCs), one output (ΔNLL).

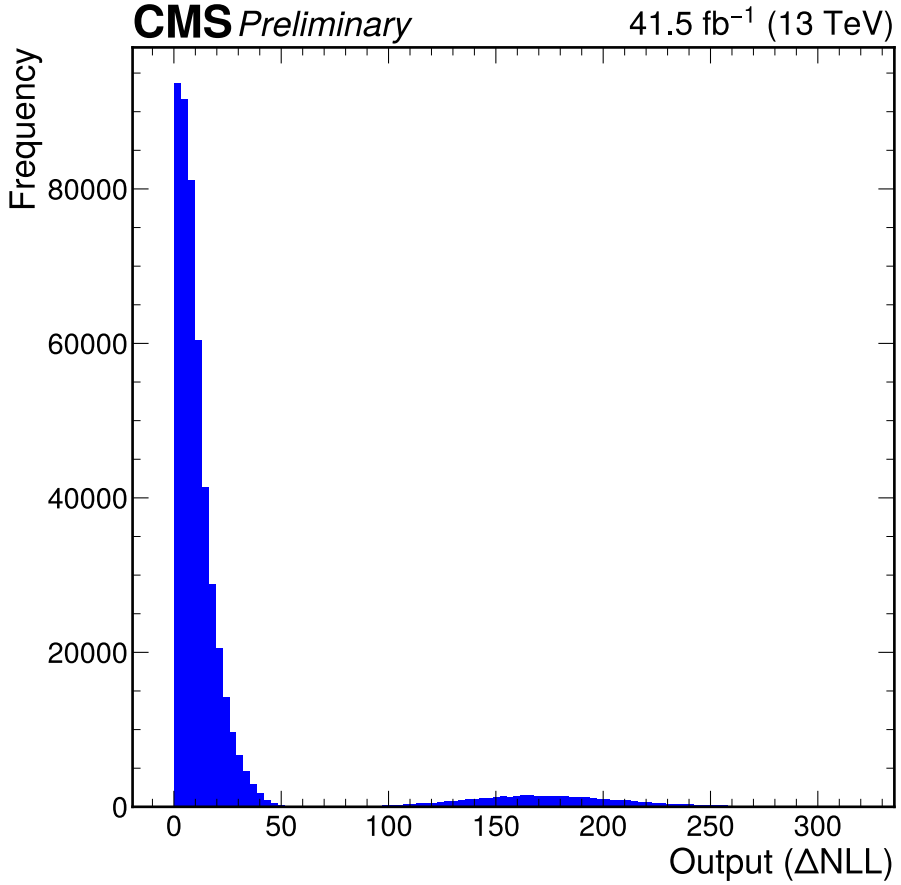
[2]: https://doi.org/10.1140/epjc/s10052-020-8230-1

DNN Architecture

- Fully-connected, feed-forward with these sequential layers:
  - A non-trainable standardization layer.
  - A non-trainable quadratic layer that appends all the square and cross terms between the 16 inputs ($c_i c_j$ for $1 \leq i \leq j \leq 16$) as additional inputs to the next layer.
  - Two hidden layers with 700 nodes each feeding into one output (the NLL).
  - A non-trainable layer that “de-standardizes” the output.

DNN Training

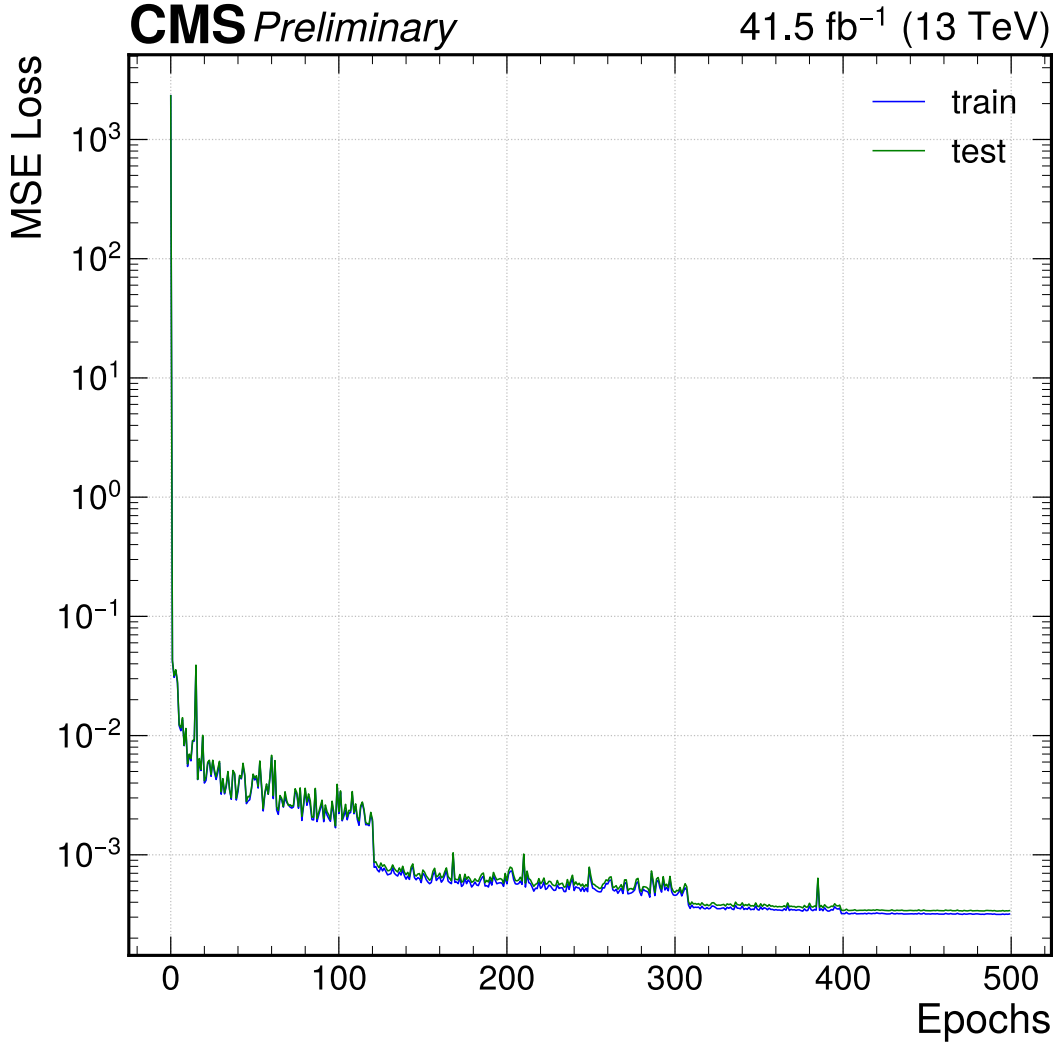
- The sample has around 50 million points across the 16D space with the NPs profiled away in the process.
- We use the MSE loss function.



- Training set output (ΔNLL) distribution. Regions in the WC space with low NLL (high likelihood) are sampled more heavily to improve DNN performance in these regions of interest.
- Experimental data underlying the LF are multilepton events selected from 2017 CMS data [1].

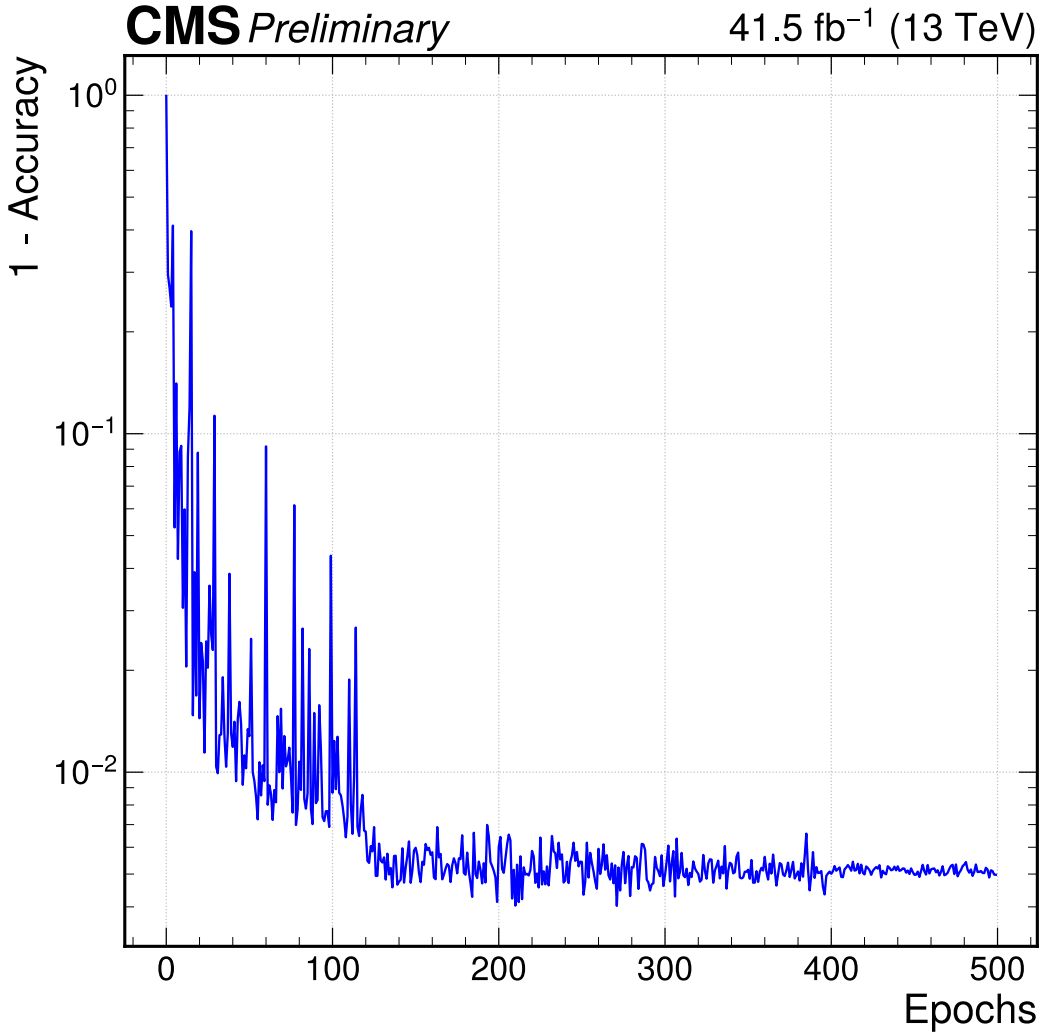
[1]: https://doi.org/10.1007/JHEP03(2021)095

DNN Training



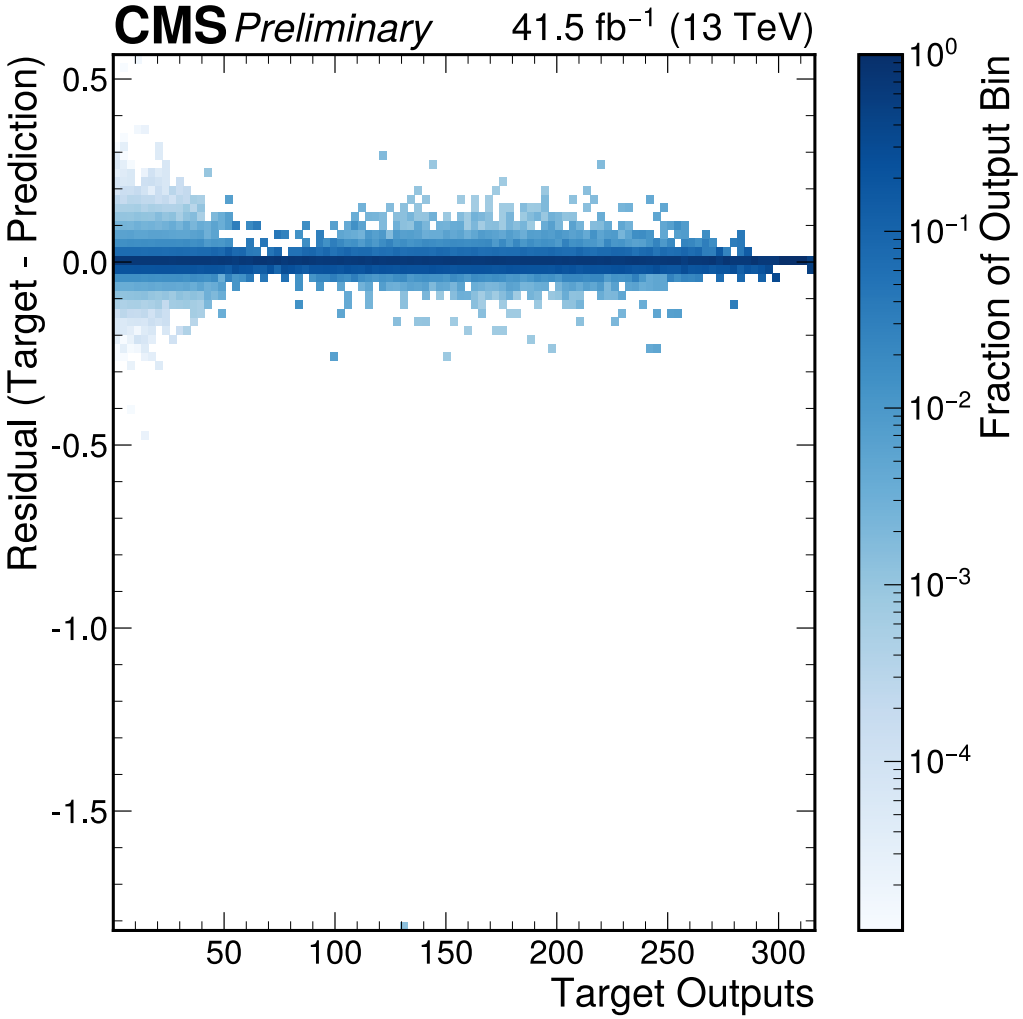
- Loss curve. Loss eventually approaches 10^{-4} without the training and testing curves separating significantly, indicating an excellent fit without overfitting.

DNN Training



- Accuracy curve. A point is considered accurate if the predicted output is within 0.05 absolutely of or 1% relative to the target output.
- Eventually, the DNN reaches over 99% accuracy, confirming an excellent fit.

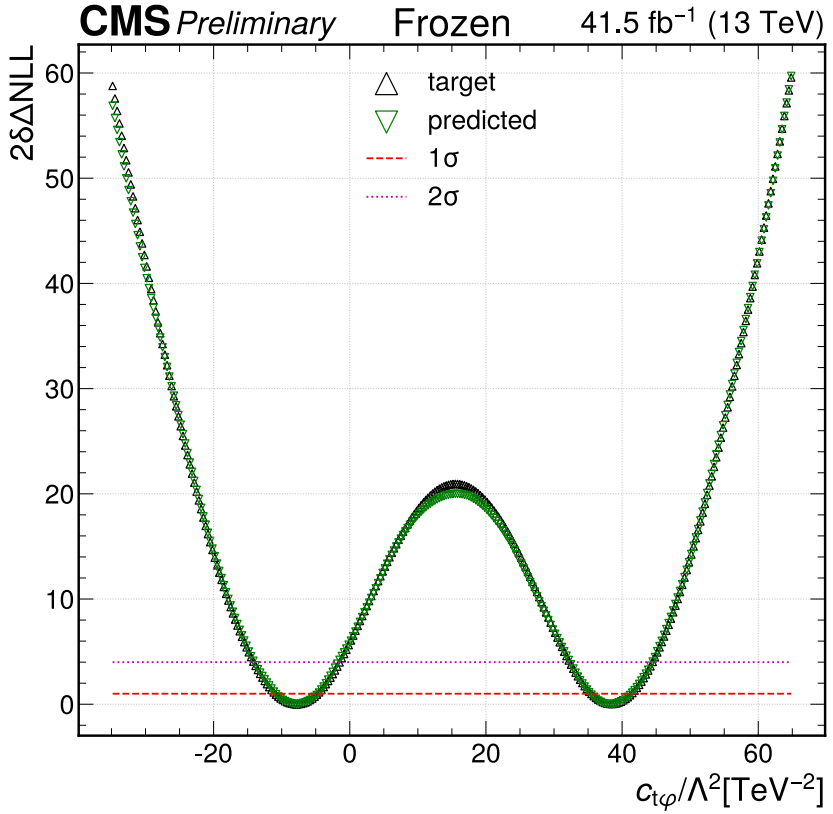
DNN Training

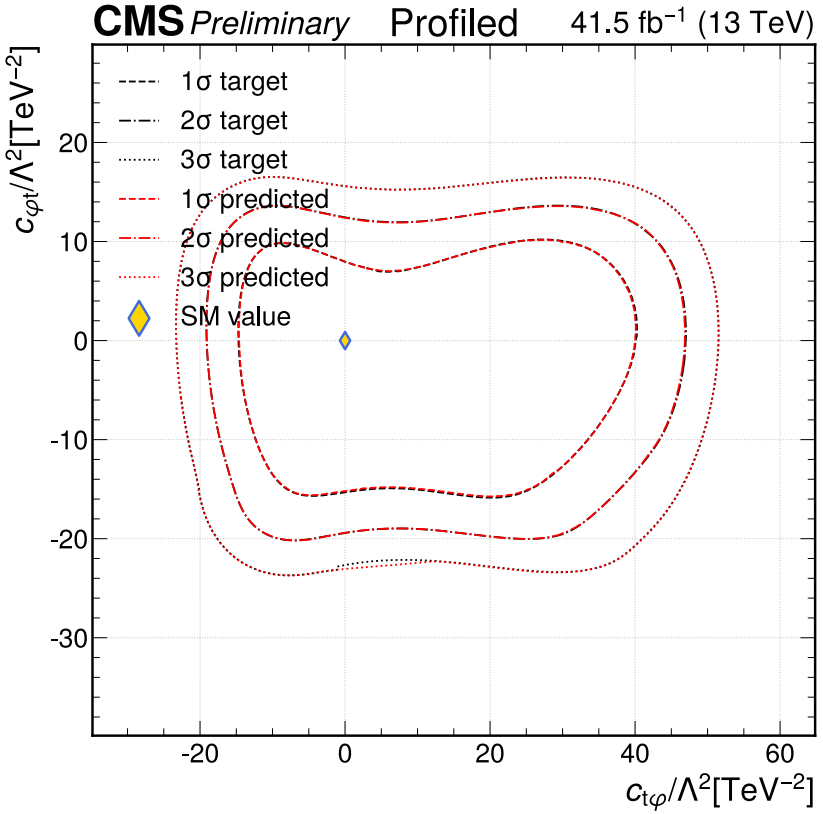


- Residual plot. Other than the very few outliers at the bottom of the plot near the output value of 130, the residuals show no obvious pattern and concentrate around 0, indicating a good fit across the range of outputs in the testing set.

DNN Validation

- To further visualize the accuracy of the DNN, we show select 1D and 2D scans published in Ref. [1] compared to DNN predictions.
- For statistical reasons, both the published and DNN-predicted curves (or surfaces in 2D scans) are shifted vertically so that their respective minimum is 0.

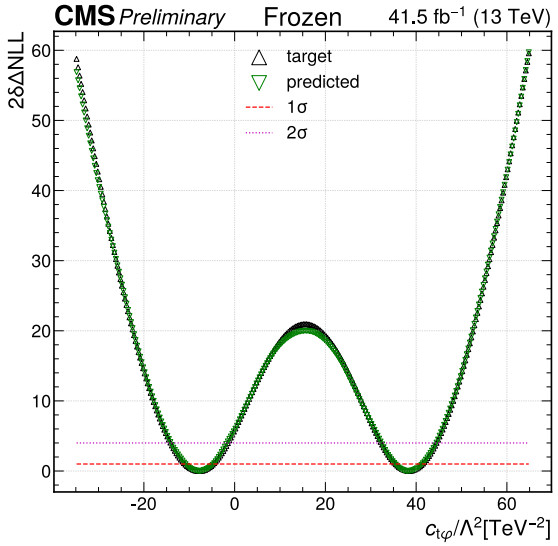


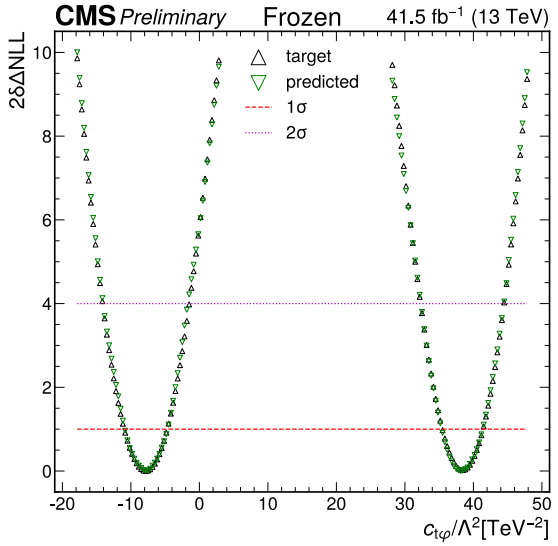


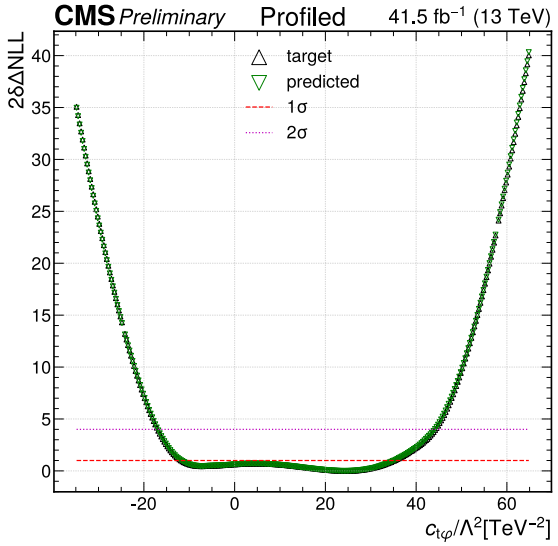
- Example scans.
- Reminder: All the NPs are already profiled away, so these are “slices” of the 16D WC space.

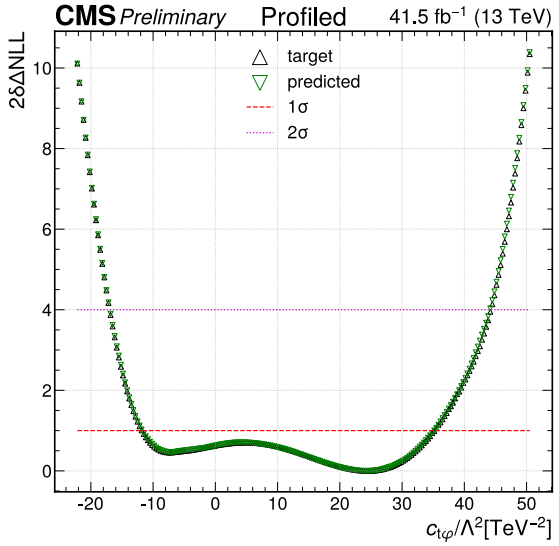
[1]: https://doi.org/10.1007/JHEP03(2021)095

DNN Validation



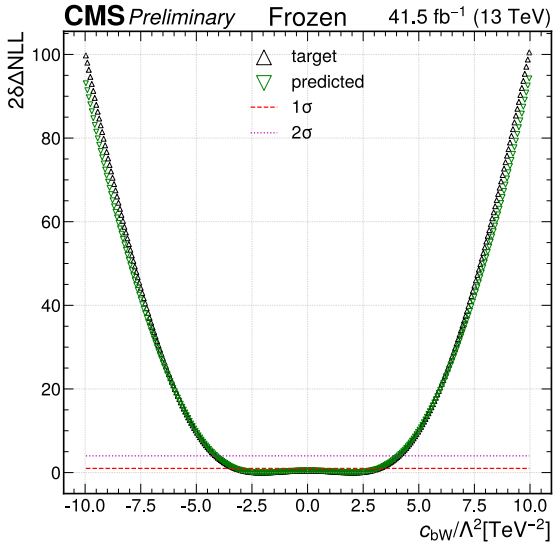


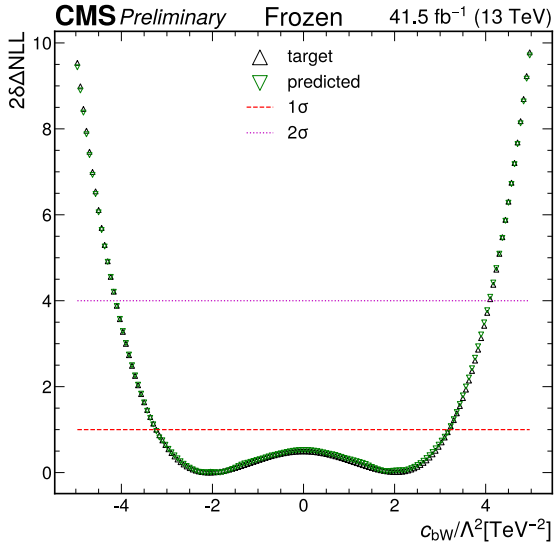


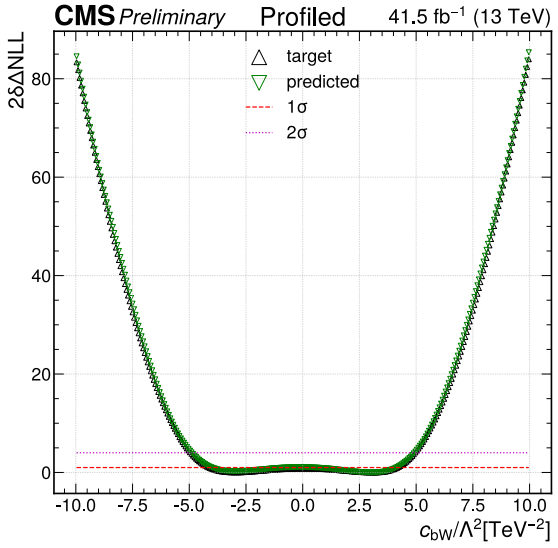


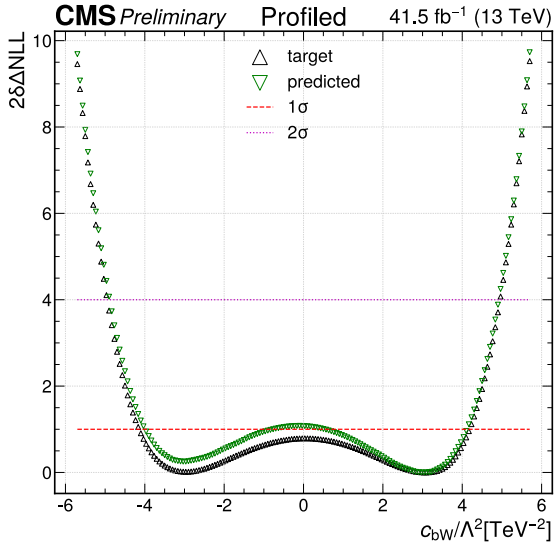
- 1D scans of $c_{t\varphi}$.
- Good agreement in all scans.

DNN Validation



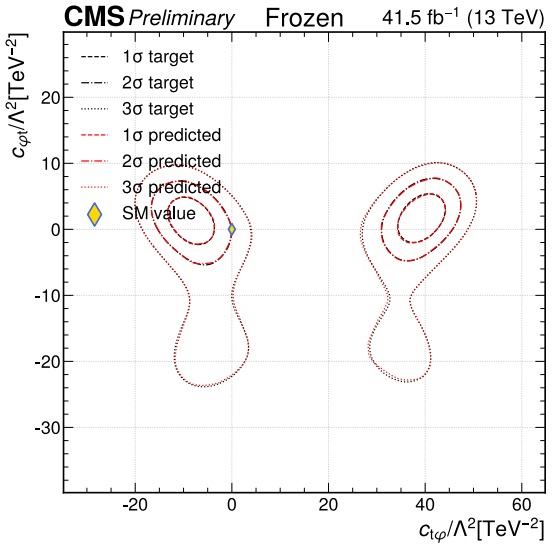


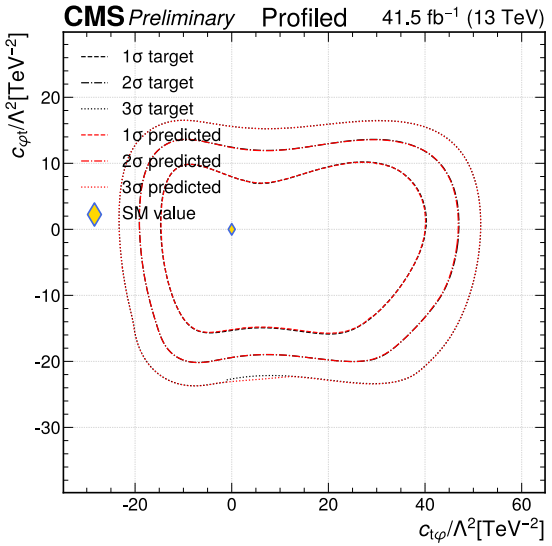




- 1D scans of c_{bW}.
- Slight deviation in the bottom right plot (profiled, zoomed-in), due to different minimization algorithms settling in slightly different minima.
- Good agreement for the rest.

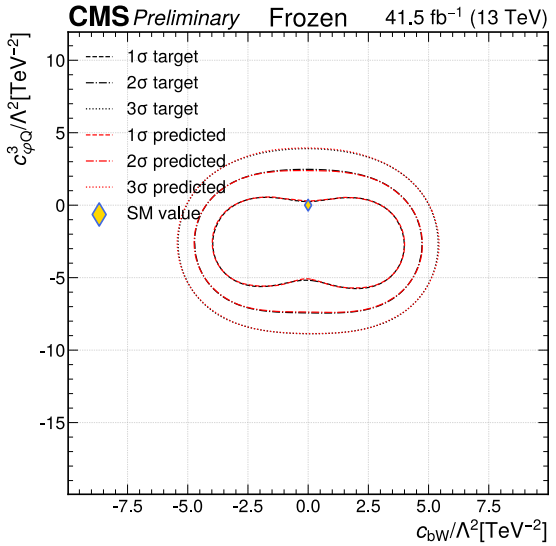
DNN Validation

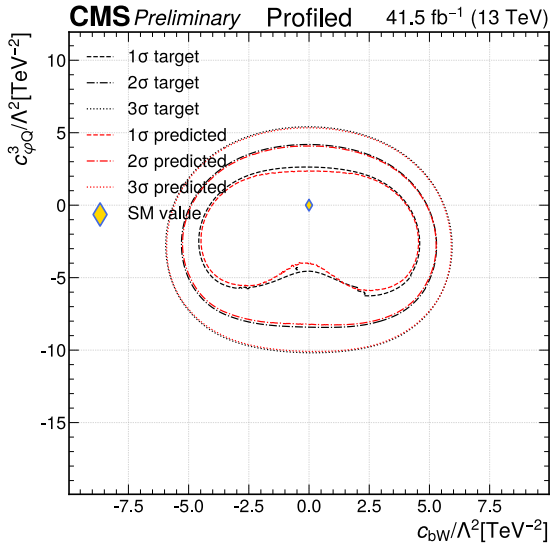




- Confidence contours of 2D scans of $c_{\varphi t}$ and $c_{t\varphi}$.
- Good agreement in both scans.

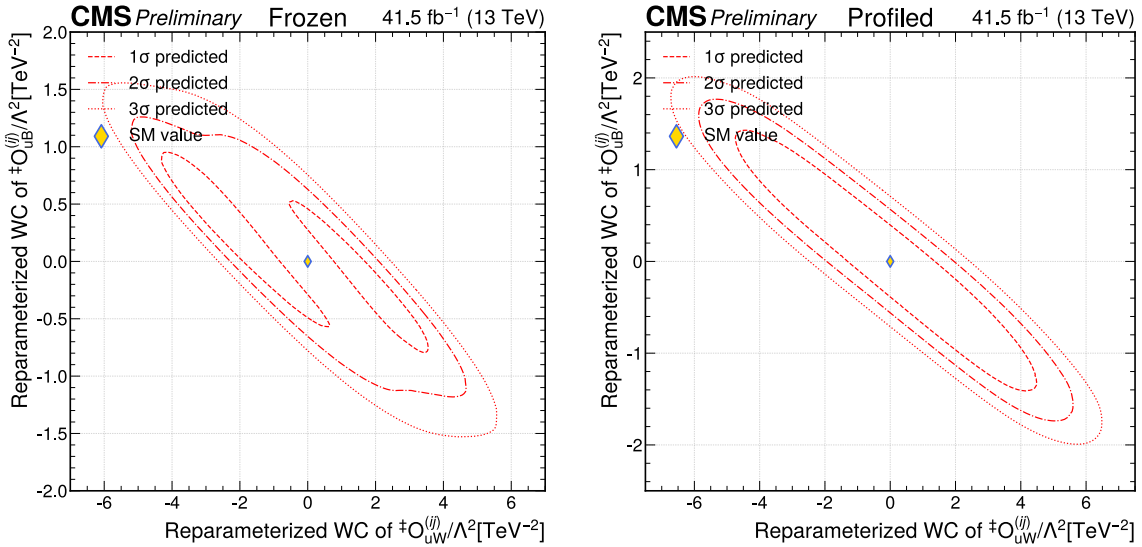
DNN Validation





- Confidence contours of 2D scans of $c^3_{\varphi \mathrm{Q}}$ and c_{bW}.
- Slight deviation in the profiled plot, due to different minimization algorithms settling in slightly different minima.
- Good agreement for the frozen plot.

Linear Combination of WCs Analysis



- Confidence contours of 2D scans of linear combinations of c_{tW} and c_{tZ} as defined on page 6 in Ref. [1].
- An example showing that the trained DNN is easily reusable for reparameterizations of the 16D WC space by adding a non-trainable linear layer, without retraining.

[1]: https://doi.org/10.1007/JHEP03(2021)095

DNN Advantages and Limitations

- Evaluation speed up
  - 5 orders of magnitude of speed up on the profiled NLL, which can be really useful depending on the use case.
- Portability
  - The DNN is small (megabytes) and portable across software environments.
- Limitations
  - The DNN is an approximation.
  - It does not retain information about the systematic uncertainties encoded by the NPs.
  - Needs an initial investment of ~50 M NLL evaluations with NPs profiled away and a few hours training on one GPU.

Conclusion

- We have a trained DNN that approximates the profiled LF with high accuracy.
- The trained DNN is easily reusable for reparameterizations of the WC space.

- For use cases that do not require
  - exact NLL values,
  - the systematic uncertainties encoded by the NPs,
- an initial investment of sampling the profiled ΔNLL and training the DNN yields
  - a differentiable, fast, and portable approximation of the profiled ΔNLL of our analysis
  - to share with the community without losing information about any of the WCs.

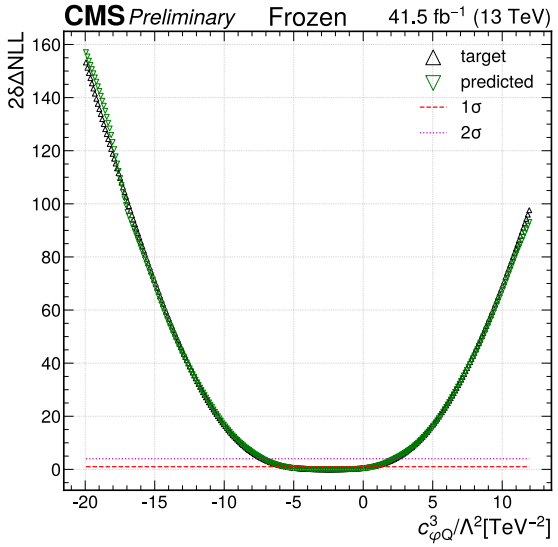
Backup Slides

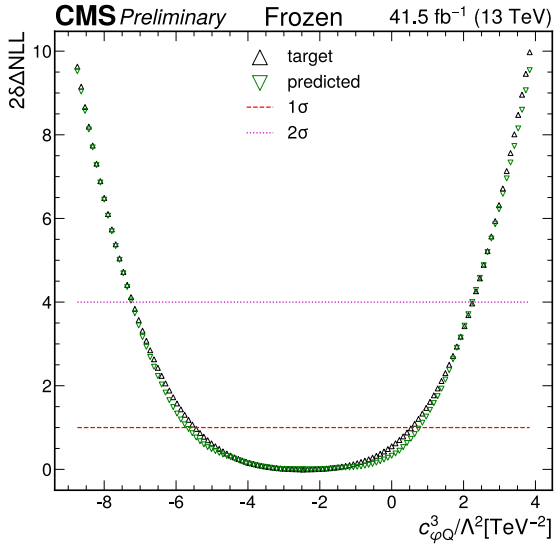
Hyperparameter Tuning

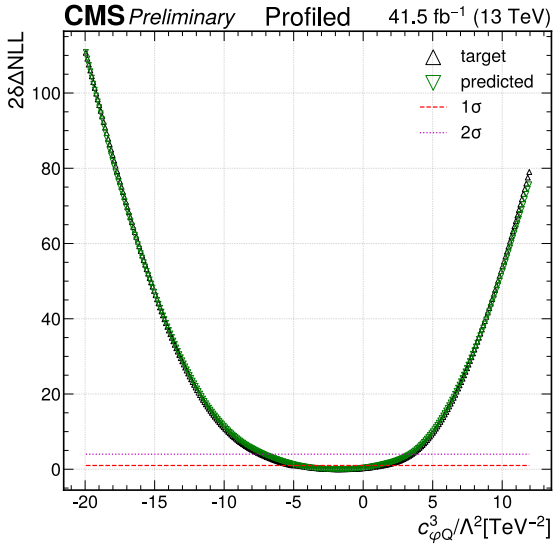
COLLEGE OF SCIENCE
UNIVERSITY OF NOTRE DAME

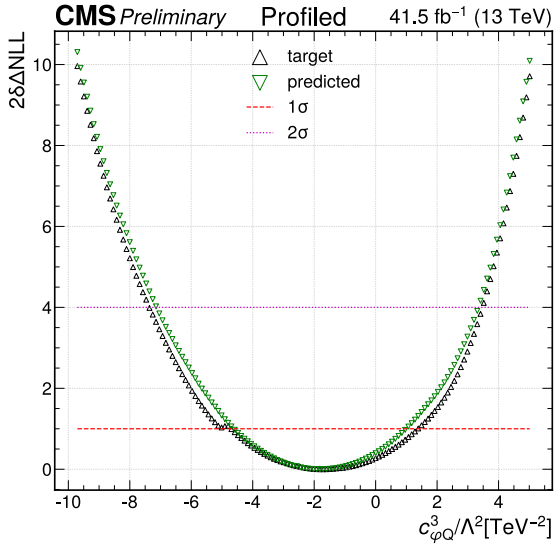
Hyperparameter	**Considered**	**Best trial**
Nodes	500-2000	700
Layers	2-4	2
Minibatch size	512, 1024	512
Epochs	500	500
Activation function	SELU, ReLU	ReLU
Loss function	Huber Loss, MSE	MSE
Optimizer	SGD, ADAM	ADAM
Initial learning rate	10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}	10^{-4}
Learning rate reduction factor	0.2	0.2
Learning rate reduction patience	5-25	20
Learning rate reduction threshold	10^{-6}	10^{-6}

DNN Validation



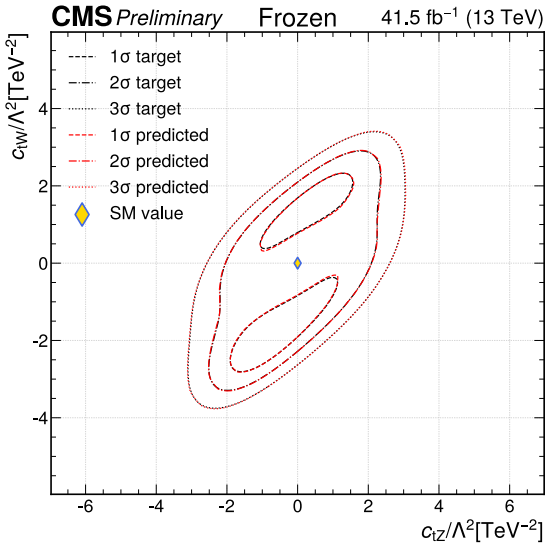


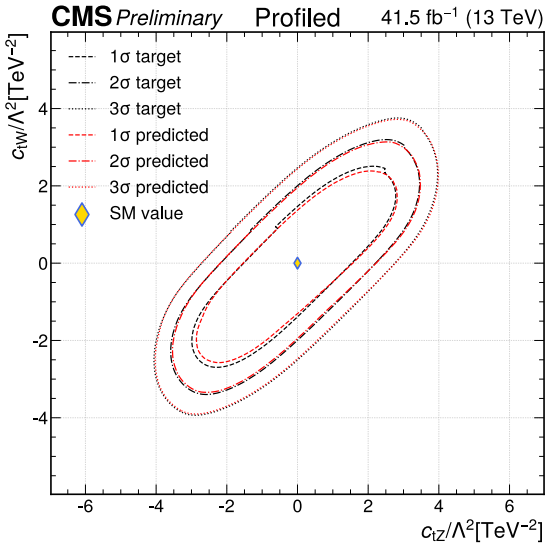




- 1D scans of $c^3_{\varphi \mathrm{Q}}$.
- Slight deviation in the bottom right plot (profiled, zoomed-in).
- Good agreement for the rest.

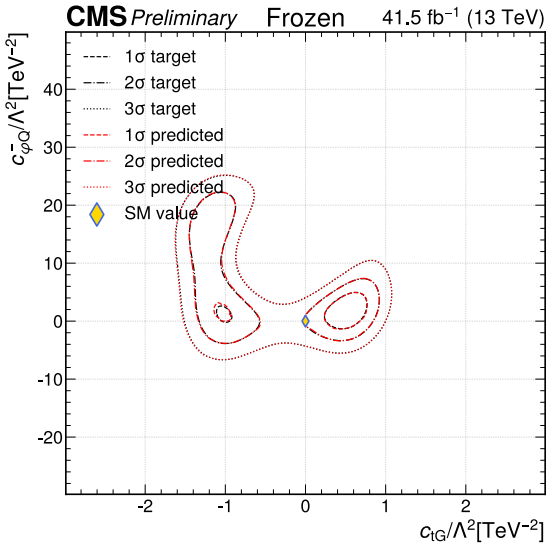
DNN Validation

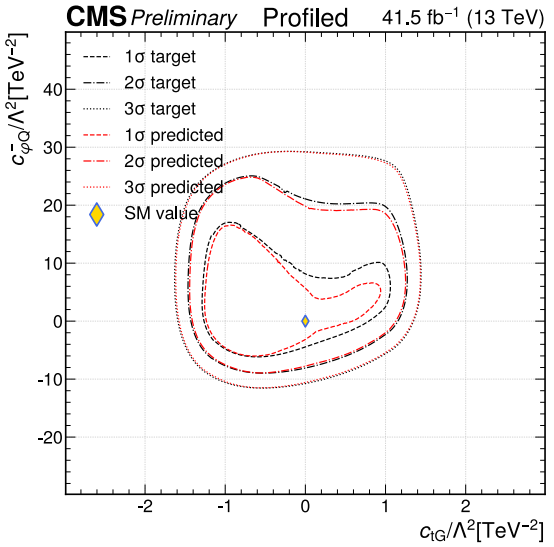




- Confidence contours of 2D scans of c_{tW} and c_{tZ}.
- Slight deviation in the profiled plot.
- Good agreement for the frozen plot.

DNN Validation





- Confidence contours of 2D scans of $c^{-}_{\varphi Q}$ and c_{tG}.
- Slight deviation in the profiled plot.
- Good agreement for the frozen plot.

Profiling in `PyTorch` [3]

- Use the same optimization tools as DNN training, but instead on the inputs (WCs) of the DNN to minimize the output (NLL).
  - Use `torch.autograd` on the trained NN.
- Minimizes sum(ln(ΔNLL + constant)) to
  - profile many points in the input space at once, taking advantage of GPU acceleration.
  - give more weight to smaller NLLs.

[3]: https://doi.org/10.48550/arXiv.1912.01703

References

[1]: The CMS collaboration. Search for new physics in top quark production with additional leptons in proton-proton collisions at √s = 13 TeV using effective field theory. J. High Energ. Phys. 2021, 95 (2021). https://doi.org/10.1007/JHEP03(2021)095

[2]: Coccaro, A., Pierini, M., Silvestrini, L. et al. The DNNLikelihood: enhancing likelihood distribution with Deep Learning. Eur. Phys. J. C 80, 664 (2020). https://doi.org/10.1140/epjc/s10052-020-8230-1

[3]: A. Paszke et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. e-Print: 1912.01703. https://doi.org/10.48550/arXiv.1912.01703