Machine Learning to Reduce PDF Uncertainties

Jason Gombas-Salazar

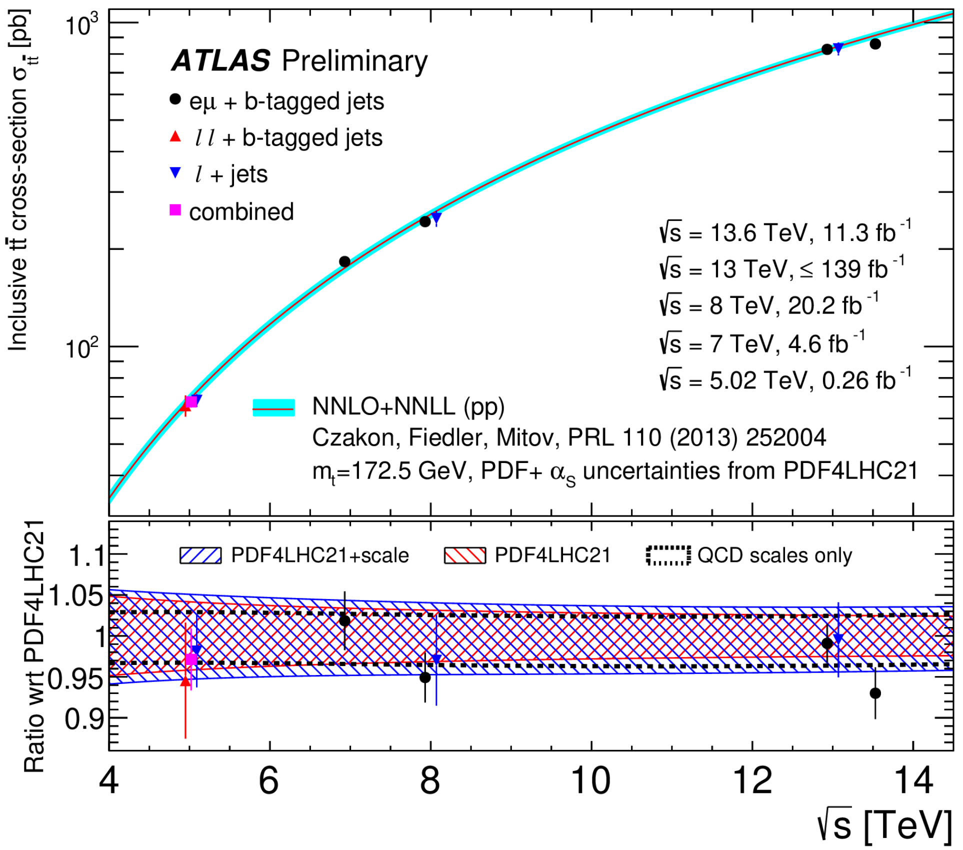
Reinhard Schwienhorst
Binbin Dong
Jarrett Fein

MICHIGAN STATE UNIVERSITY

Introduction

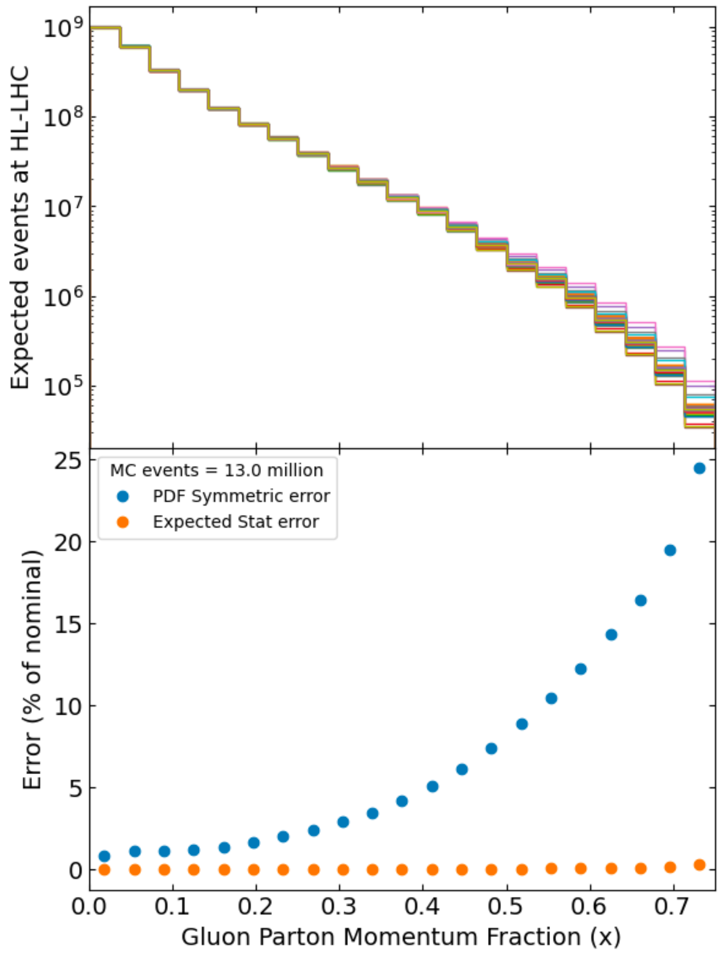
- PDF uncertainty will soon dominate many theory calculations and measurements including
  - $t\bar{t}$ production cross-sections
  - EFT fits
  - Higgs measurements
- Furthermore, other colliders that will further constrain the PDFs are far in the future (i.e. EIC)
- Global PDF fits utilize final-state variables, which do not provide the full information available for a given process
- Our focus is on the gluon PDF at high proton momentum fraction because of its impact on top quark measurements
  - See our 2021 Snowmass proceedings: https://arxiv.org/abs/2203.08064

Current areas where improvement is needed



PDFs will become more important in the next precision era which is revealed in the current $t\bar{t}$ production cross section as a function of COM energy

ATLAS (2023)

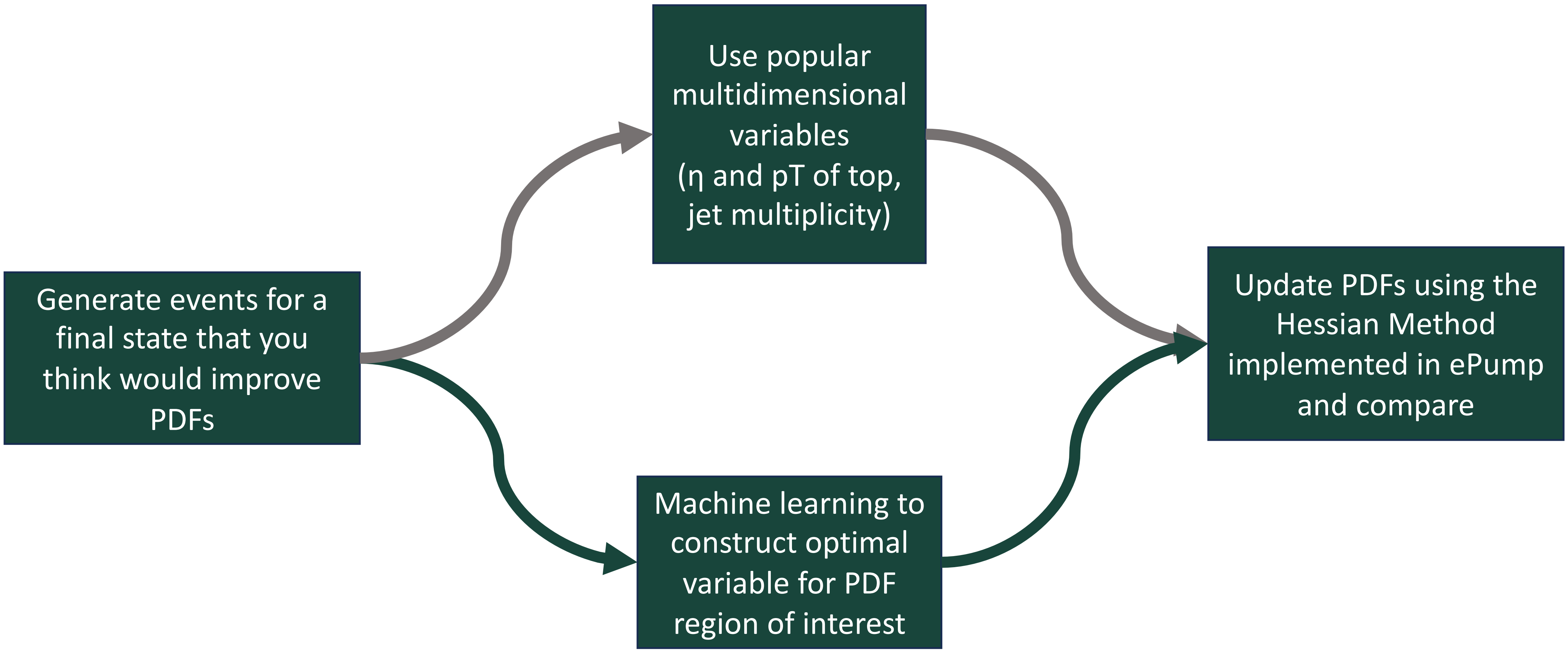


Expected number of events at HL-LHC of $t\bar{t}j$ at NLO and COM energy of 14 TeV as a function of the initial gluon parton momentum fraction (CT18NLO)

Strategies for PDF global fit

- Typical variables, like η or p_T of top quarks, have been chosen to be included in the global PDF fit because of motivated physics intuition
- Rapidity is good because it is straightforward to calculate and provides good information on the general boost of the hard scattered particles, but is it optimal?
- The goal here is to concentrate PDF sensitivity into a single non-linear variable using machine learning to include in future global PDF fits
- I will show that future fits with HL-LHC data might benefit from non-standard variables

Workflow scheme

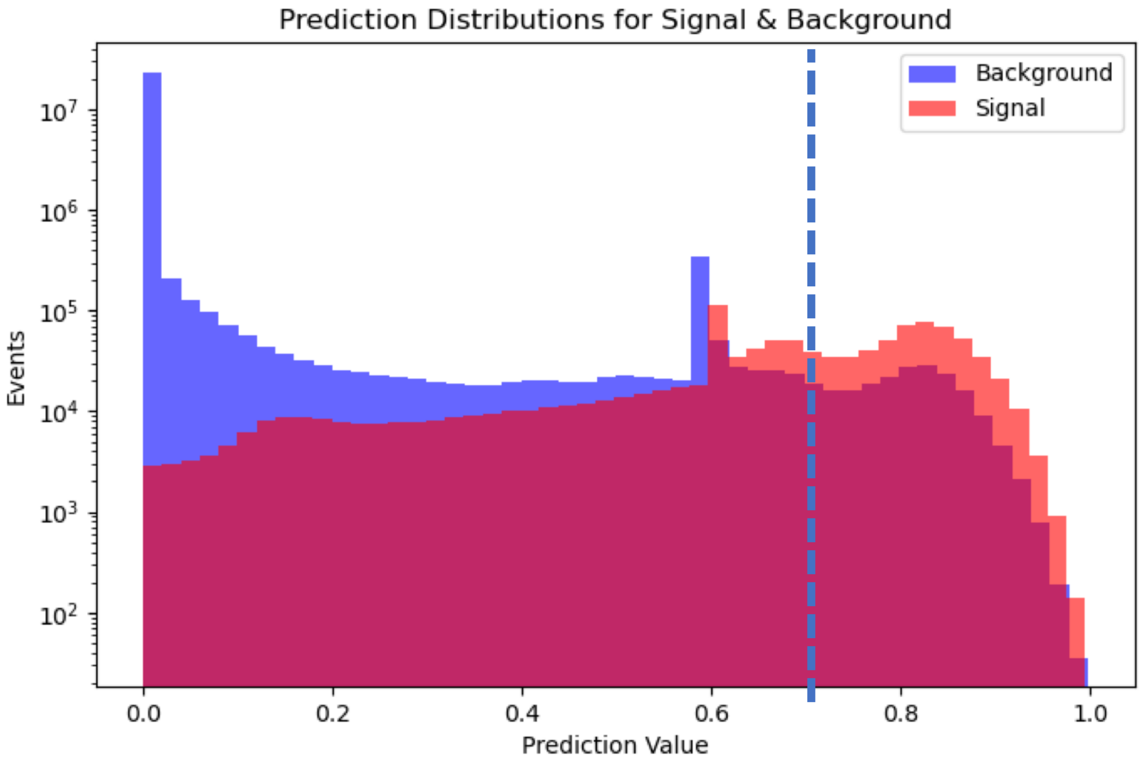


Details on the study and the sample

- 7.5 million events were generated using Madgraph aMC@NLO
  - The final state is $t\bar{t}$j at NLO
  - 14 TeV center of mass energy
  - Only truth level where the top quarks are not decayed
  - Normalize to 3,000 fb^{-1} (ie HL-LHC)
- Train and use a MLP to identify events with gluon $p_Z > 2$ TeV
- Form the rapidity distribution for **events that pass this filter** and for **all events**
- ePump is then run for both cases to compare a “traditional” variable with an “improved” variable
  - Who is familiar with ePump?

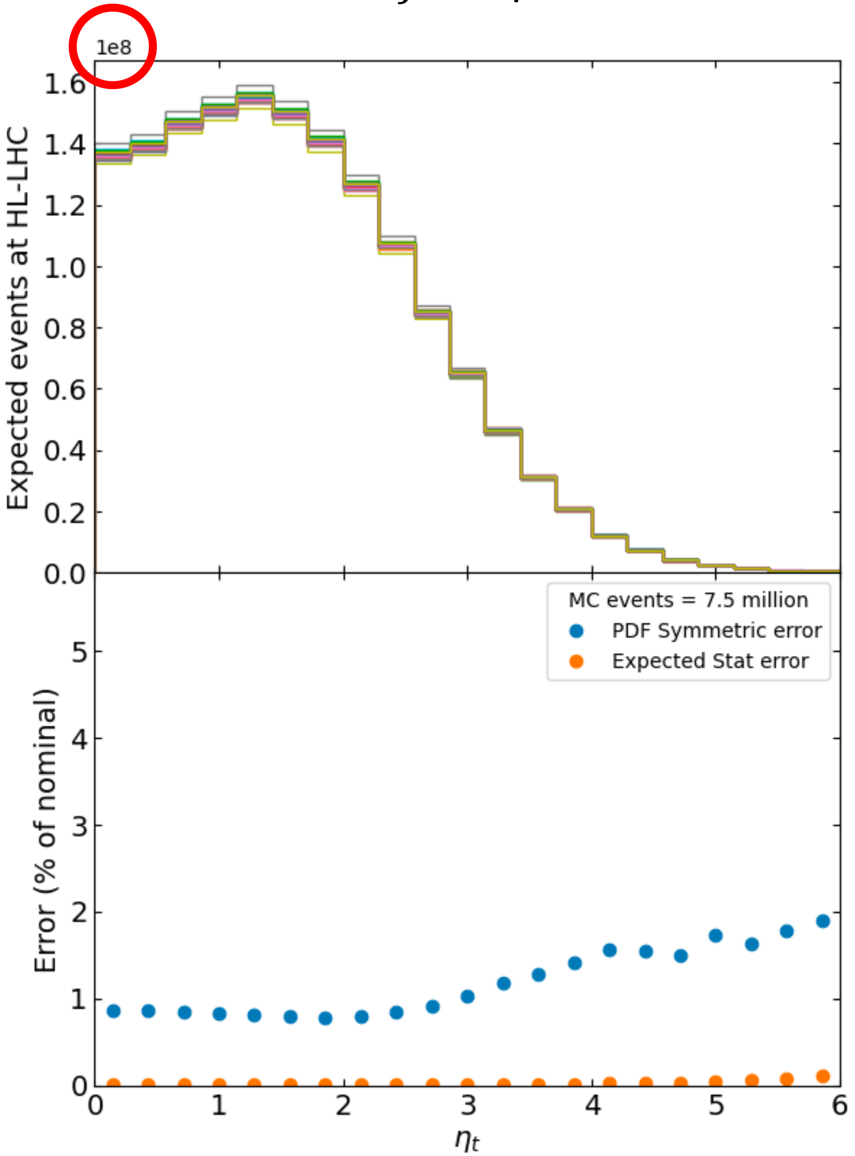
A NN high x gluon filter

- DNN input are the 4-vectors of each particle in the final state ($t\bar{t}j$ – 12 input floats)
- DNN is used with 3 hidden layers [128, 64, 32]
- Signal is classified as events with a gluon parton with $p_Z > 2$ TeV
- DNN output is classification score [0,1]
- Cut of 0.7 was chosen
- Background to signal ratio ~ 25:1

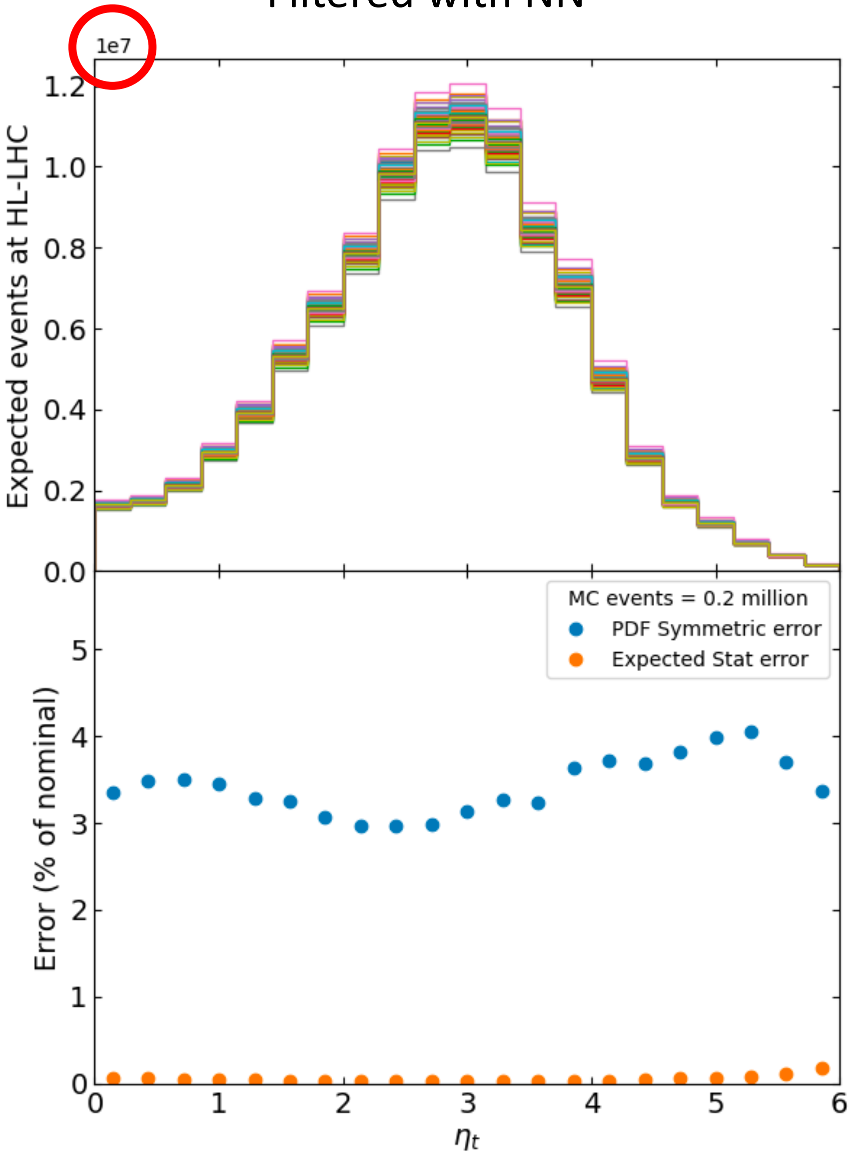


Pseudo-data for our PDF updates – Rapidity of top quark

Full $t\bar{t}j$ sample

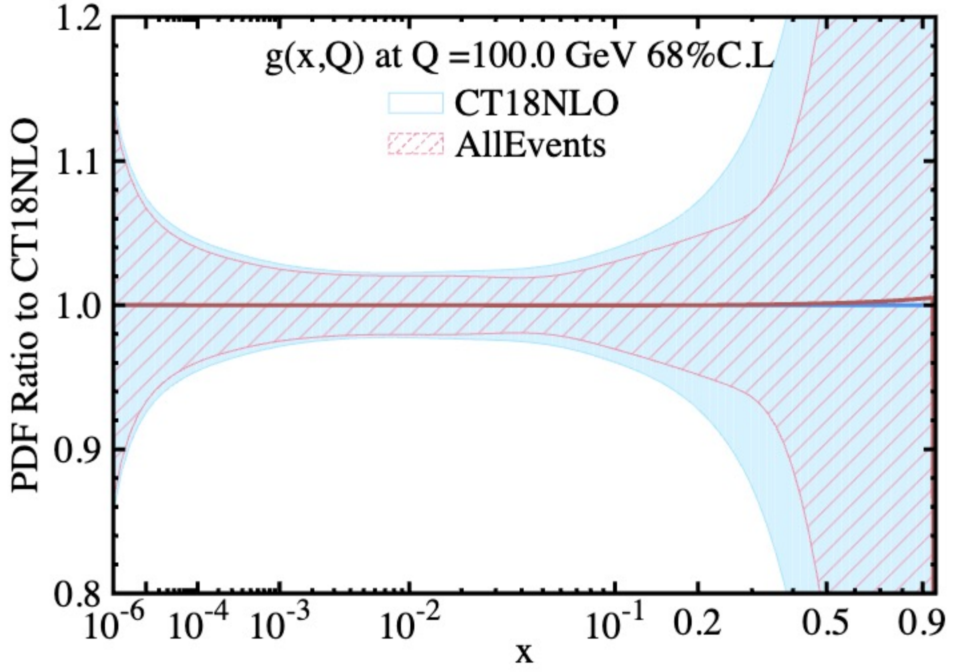


Filtered with NN

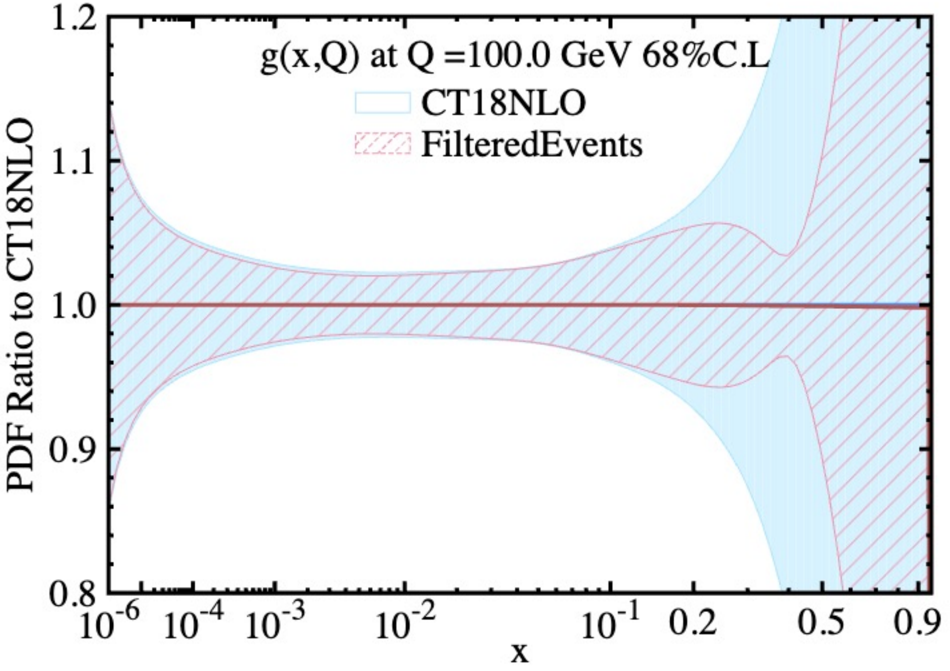


Updating PDFs with the pseudo-data

Full $t\bar{t}j$ sample



Filtered with NN

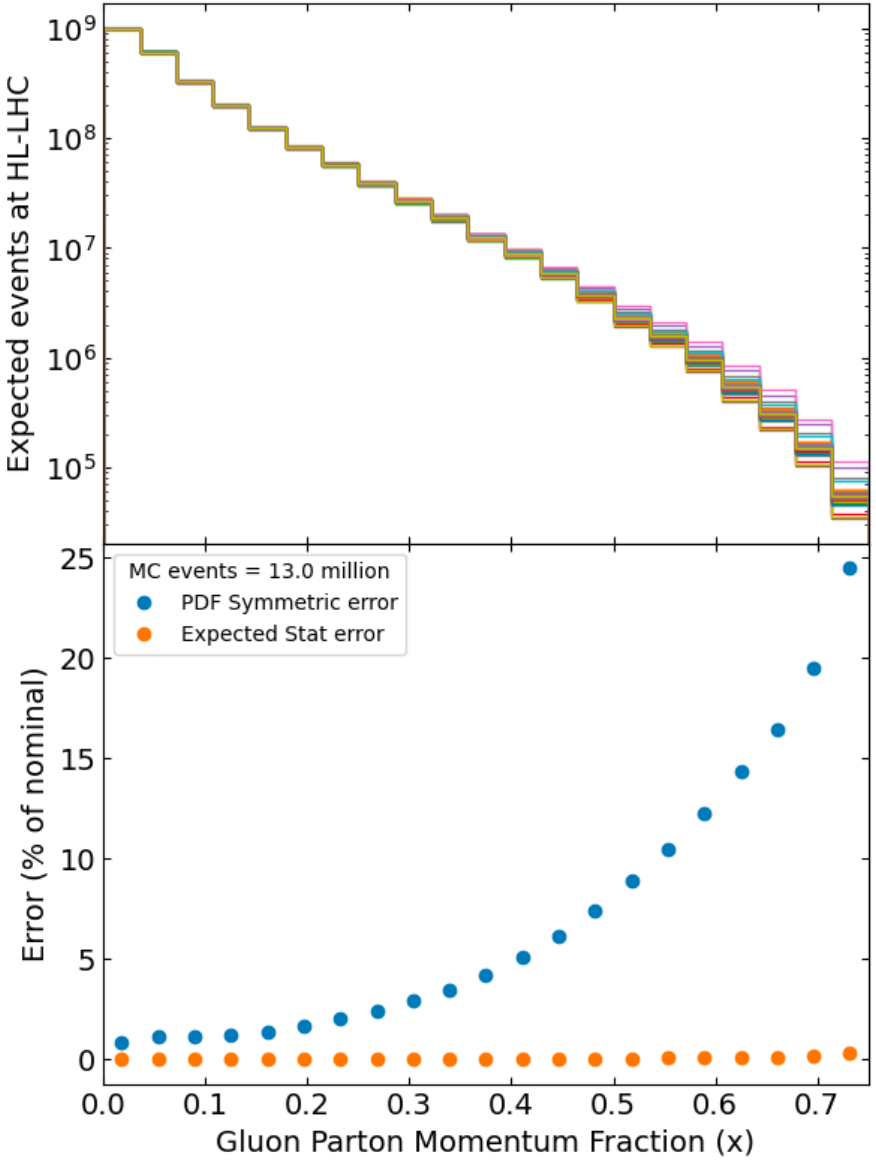


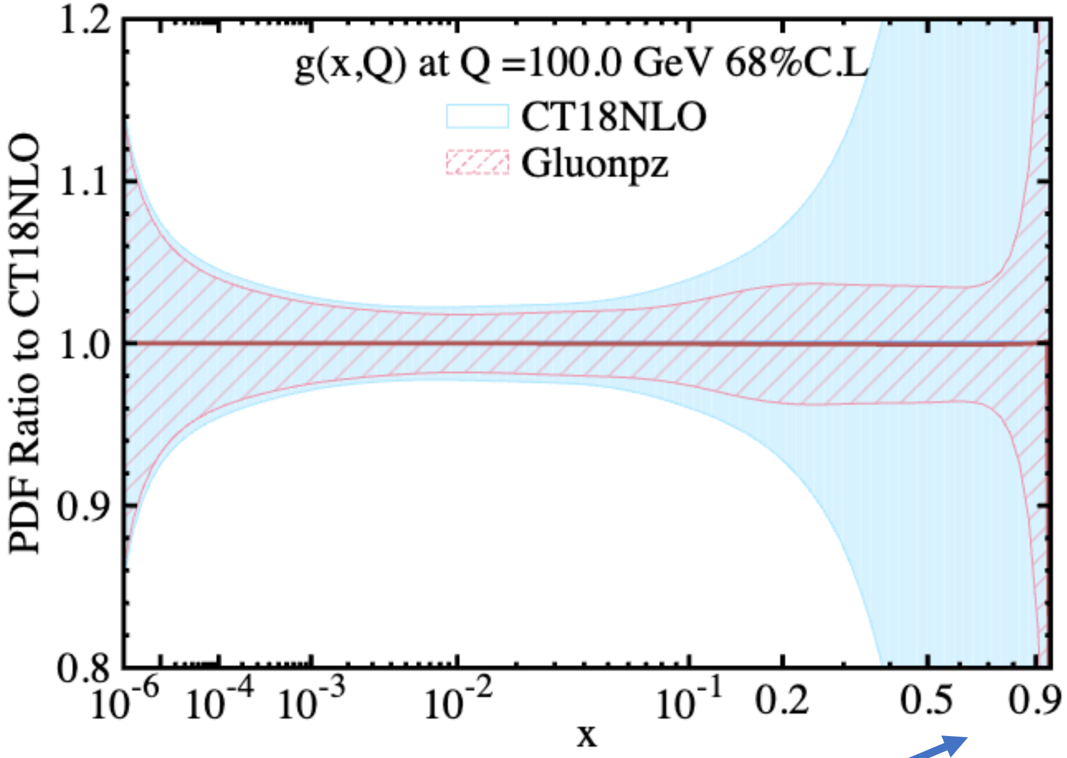
Fits use nominal distribution as data with a difficult but achievable systematic uncertainty of 1%
This comparison shows that the high x gluon PDF region gets targeted by design

Just the beginning…

- These are nice, but are obviously too clean
- Our future perspective is to do much more including:
  - Decaying the tops
  - Reconstructing from detector simulation
  - Find a better variable than just rapidity with a filter
- Collaboration opportunities:
  - Experimental side with reconstruction of the partons from messy experimental data
  - Theory side with accurate modeling of the partons and introduction of such a constructed variable into the global PDF fit

Where this could go… a best-case scenario





* Because of approximately infinite statistics from the HL-LHC

Uncertainty reduction all the way to x = 0.9!

Outlook

- Reducing the PDF uncertainty will be vital with the next generation of precision measurements
- New techniques using machine learning can improve variables to be included in future global PDF fits
- Next steps:
  - Complete our simulation with top decays and detector simulation
  - Neural network regression to approximate the gluon momentum fraction
  - Neural network architecture that concentrates and captures uncertain regions in the gluon PDFs in a variable
  - Test with ePump and compare with each other

Thank you!
and
Questions?