MACHINE LEARNING IN HIGH-ENERGY PHYSICS

STEFANO FORTE
UNIVERSITÀ DI MILANO & INFN

VBS TRAINING SCHOOL

MILANO BICOCCA, SEPT. 3, 2021

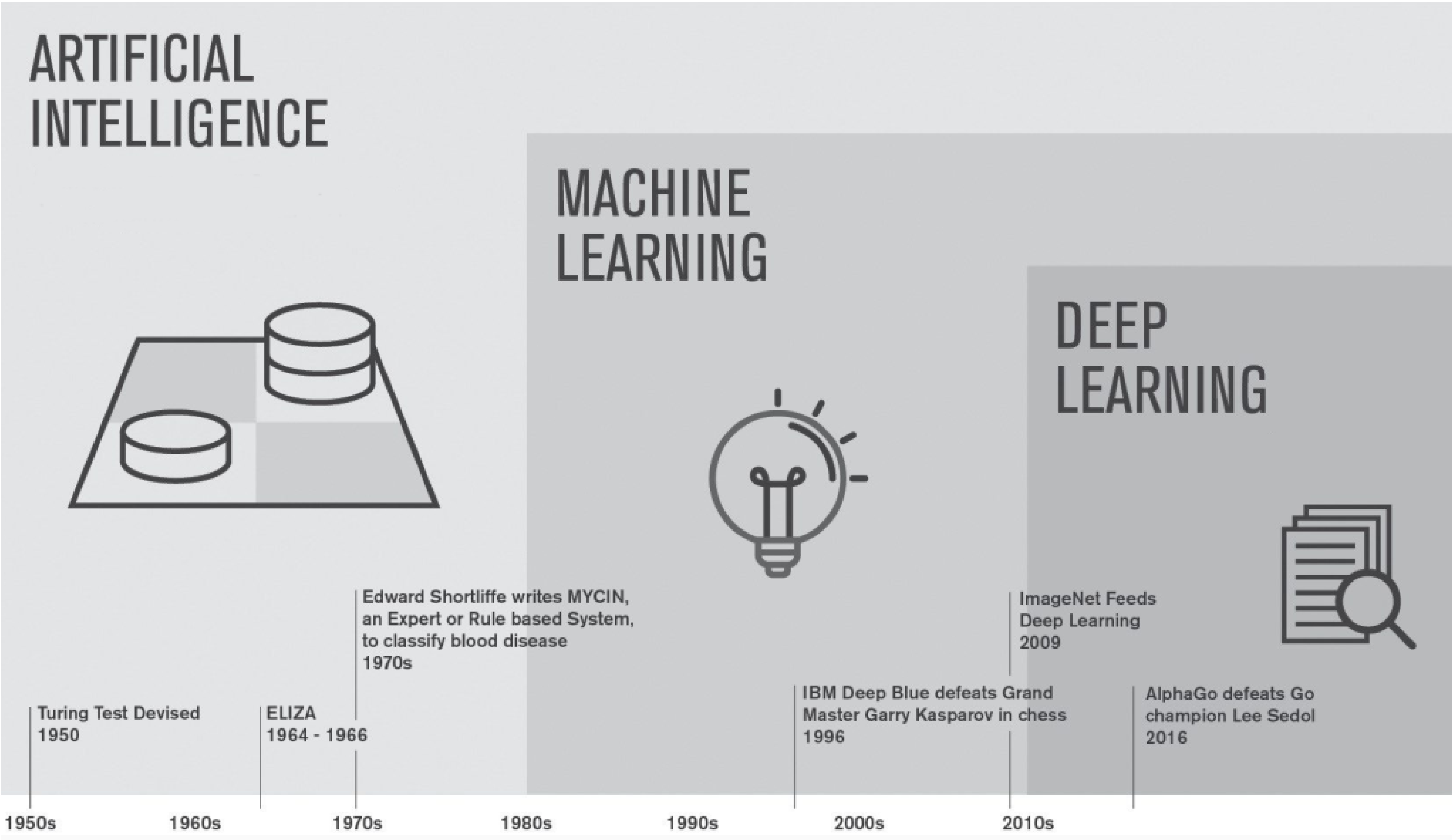
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 740006

SUMMARY

- INTRODUCTION: AI VS. ML
- ML IN HEP: SOME EXAMPLES
  - GAN EVENT UNWEIGHTING
  - ML CLASSIFIERS FOR OPTIMAL EFT SENSITIVITY
  - MAPPING ML ONTO HUMAN LEARNING
- A CASE STUDY: PDFS AS A ML PROBLEM
  - PDFS AND NNPDFS
  - NEURAL NETWORKS
  - MINIMIZATION: STOCHASTIC AND DETERMINISTIC
  - UNDER- AND OVER-LEARNING
  - CROSS-VALIDATION
  - HYPEROPTIMIZATION
  - K-FOLDING
  - GAN COMPRESSION

AI vs. ML

FROM AI TO ML



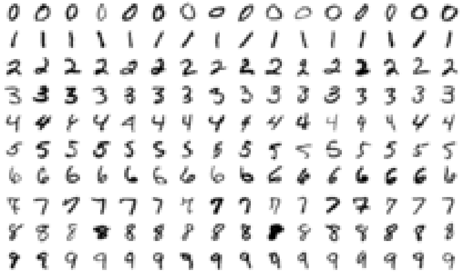
ARTIFICIAL INTELLIGENCE
MACHINE LEARNING
DEEP LEARNING
Edward Shortliffe writes MYCIN, an Expert or Rule based System, to classify blood disease 1970s
ImageNet Feeds Deep Learning 2009
Turing Test Devised 1950
ELIZA 1964 - 1966
IBM Deep Blue defeats Grand Master Garry Kasparov in chess 1996
AlphaGo defeats Go champion Lee Sedol 2016
1950s
1960s
1970s
1980s
1990s
2000s
2010s

SHIFTING OF PARADIGMS

"KNOWLEDGE BASED" AI

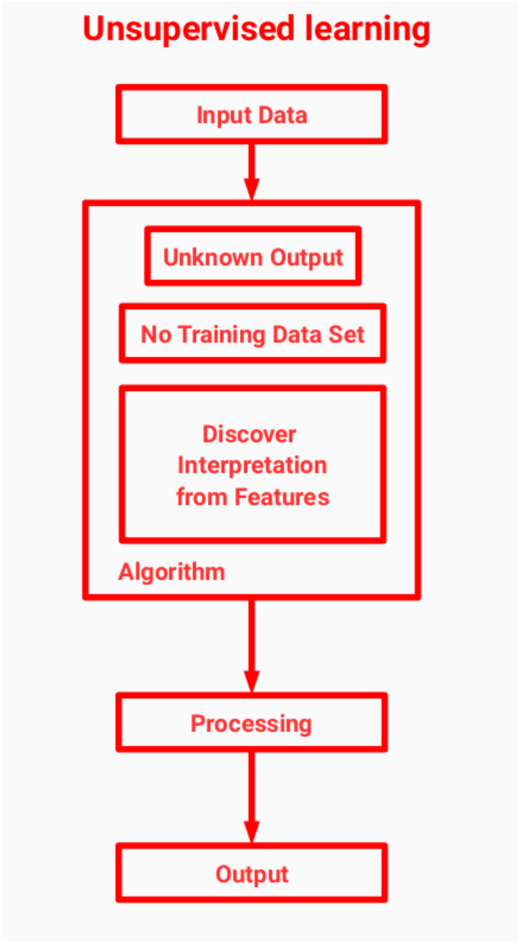
- LEARN AND IMPLEMENT A SET OF RULES
- GOOD FOR CHESS, BAD FOR REAL LIFE

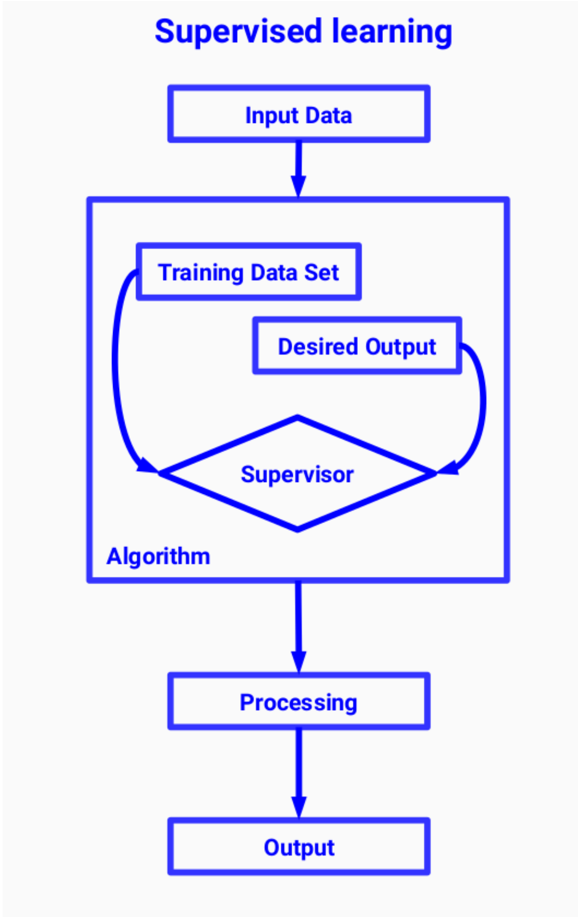
MACHINE LEARNING



- "INTUITIVE" REPRESENTATION
- THE AI AGENT BUILID UP ITS OWN KNOWLEDGE

MACHINE LEARNING ALGORITHMS





EXTRACT AND OPTIMIZE

DATA FEATURES

OPTIMIZE A PROPERTY

LEARNING FROM DATA

LEARN FROM DATA

THE LEARNING STRATEGY

ML in HEP

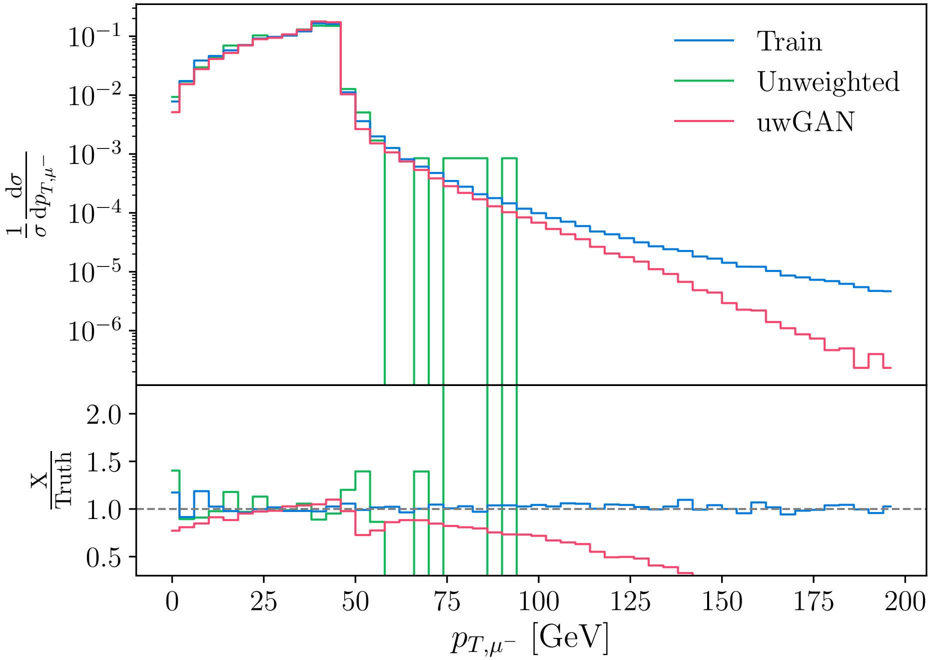
RECENT EXAMPLES

GANs for event unweighting

(Backes, Butter, Plehn, Winterhalder, 2021)

- A classic problem: determine weights for integration: $\sigma = \int dx w(x) = \int dy \tilde{w}(y)$, $\tilde{w}(y) \approx$ const.
- Standard solution: importance sampling $\Rightarrow$ rescale based on sampling (VEGAS)
- GAN: use events to train GAN
- Produce unweighted events with GAN

Muon p_T distribution in W^- production



500K training, 1k standard unweighted, 30M uwGAN events

- Faster event generation
- Reiliability?

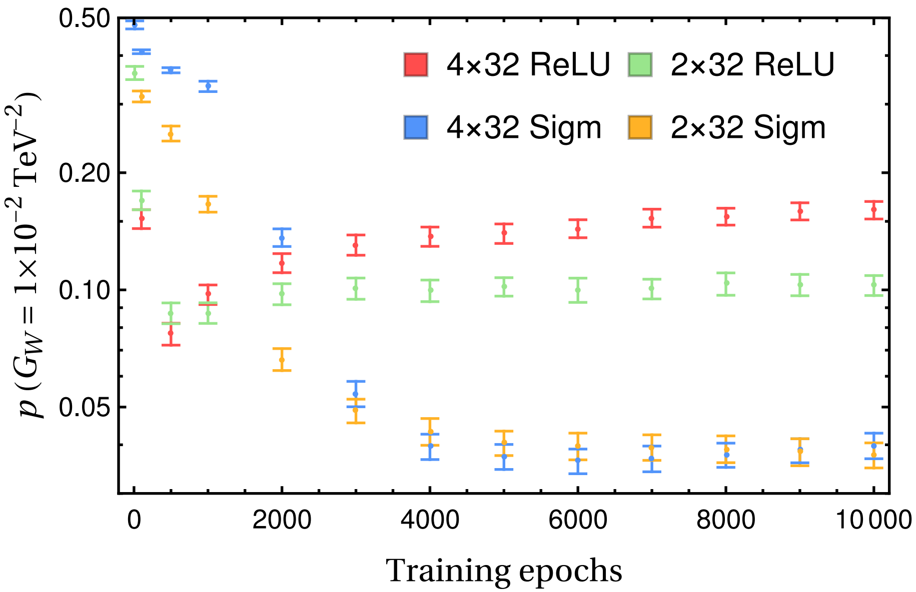
NEURAL NETWORK CLASSSIFIER FOR EFT BOUNDS

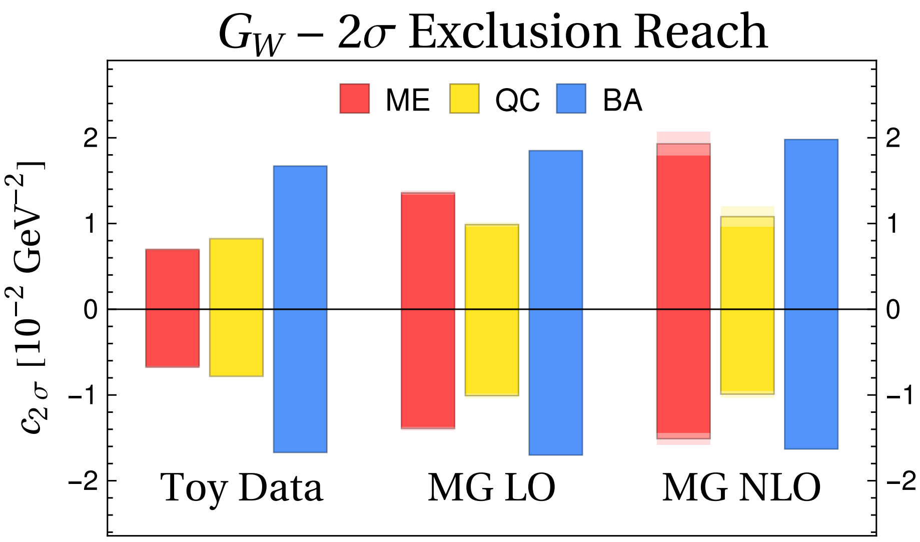
(Chen, Glioti, Panico, Wulzer, 2020)

- EFT CROSS SECTION $d\sigma_0(x;c) = d\sigma_1(x)[(1+c\alpha(x))^2 + (c\beta(x))^2]$:
  x kin. variables; SM $\Rightarrow c=0$; α, β coefficient functions for single operator
- TRAIN NEURAL NETWORKS TO REPRODUCE $\alpha(x)$ $\beta(x)$
  $\Leftrightarrow$ GENERATE MC SAMPLES WITH SEVERAL VALUES OF c & $c=0$
- OBTAIN RATIO $d\sigma_0(x;c)/d\sigma_1(x)$ FOR ALL c, x
- HYPEROPTIMIZE NEURAL NETWORK PARAMETERS

FULLY LEPTONIC ZW

HYPEROPT

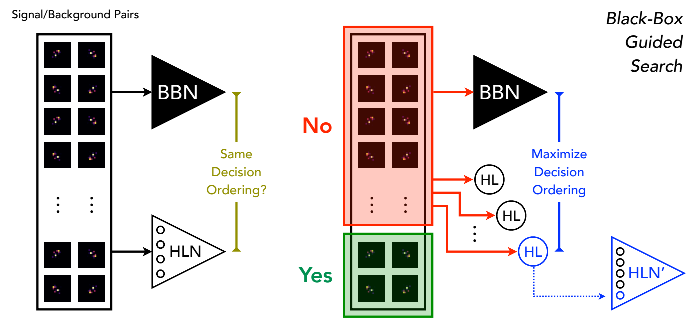




Matrix
Element vs NN Quadratic Classfifer
& Binned Analysis

- STUDY WITH TOTAL INTEGRATED HL-LHC LUMI
- COMPARISON TO MATRIX ELEMENT METHOD BASED ON ANALYTIC APPROX
  & BINNED ANALYSIS IN Pp_T^Z BASED ON THE SAME MC SIMULATIONS
- NO DETERIORATION AT NLO

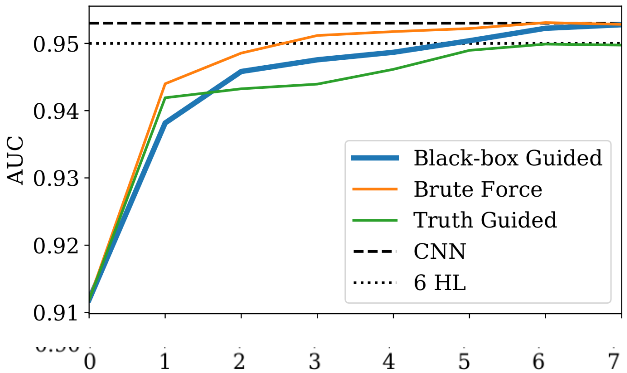
ML INSIGHTS ON HUMAN CLASSIFICATION (Faucett, Thaler, Witeson, 2021)



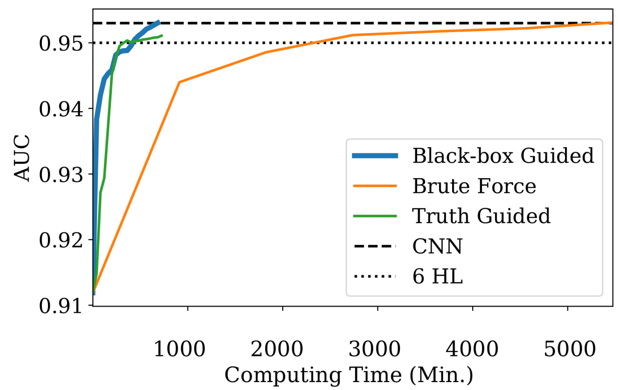
- CLASSIFICATION PROBLEM: IS EVENT SIGNAL OR BACKGROUND
  EXAMPLE: $W \to q\bar{q}$ SIGNAL: QUARK JETS
- START WITH SET OF HL OBSERVABLES & COMPARE TO BLACK-BOX NN CLASSIFIER
  EXAMPLE OF HL: JET MASS, ENERGY CORRELATION FUNCTIONS...
- SELECT HL_1 OBSERVABLE WITH HIGHEST AGREEMENT,
  LOOK AT EVENTS WITH HIGHEST DISAGREEMENT
- SELECT HL_2 OBSERVABLE WITH HIGHEST AGREEMENT & TRAIN NN ON HL_1 AND HL_2
- ITERATE UNTIL OPTIMAL SET OF HL_i DETERMINED

CLASSIFICATION PERFORMANCE

VS. NUMBER OF ADDED HL_i



VS. COMPUTING TIME



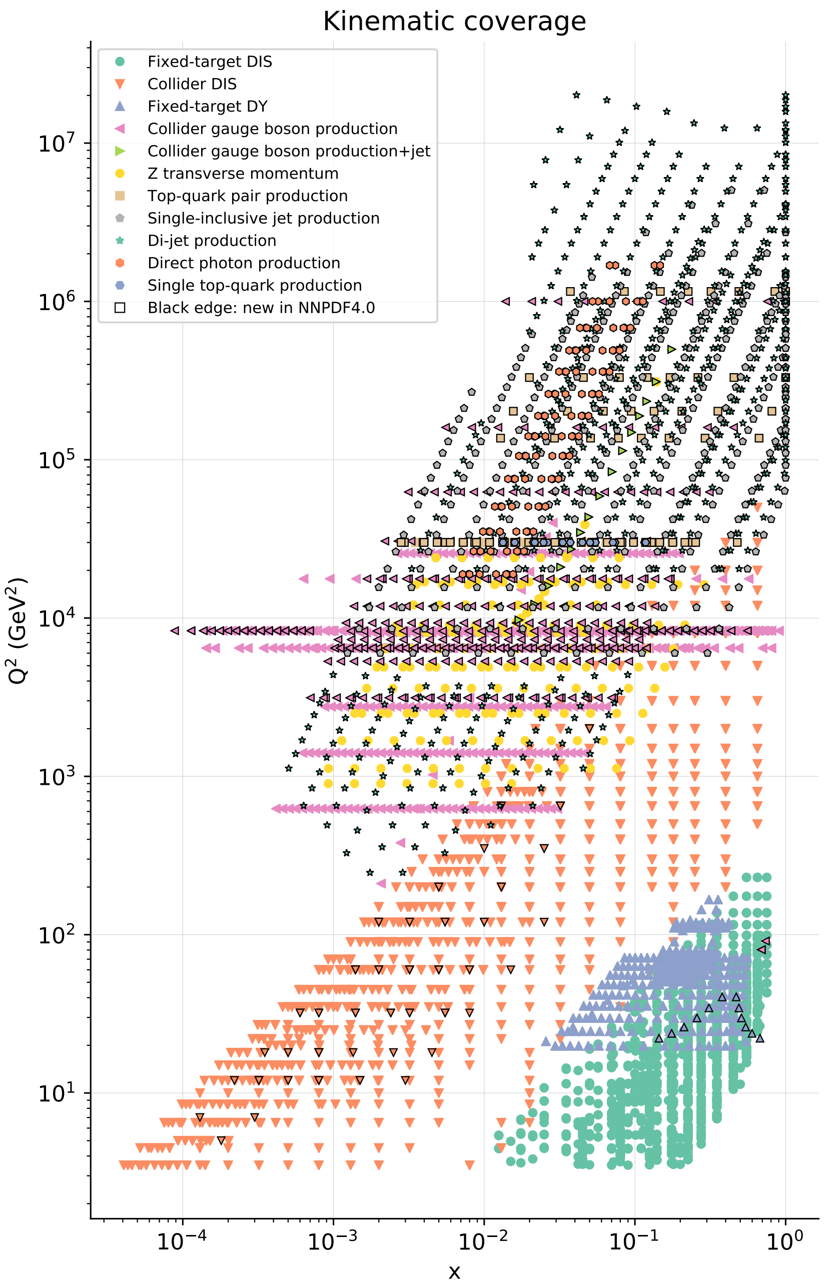
- MORE PERFORMANT THAN TRUTH-GUIDED, SLIGHTLY LESS THAN BRUTE-FORCE
- COMPUTATIONALLY AS EFFICIENT AS TRUTH-GUIDED, MUCH MORE THAN BRUTE FORCE
- PROVIDES INSIGHT ON HL OBSERVABLES

A CASE STUDY:
PDFs AS A ML PROBLEM

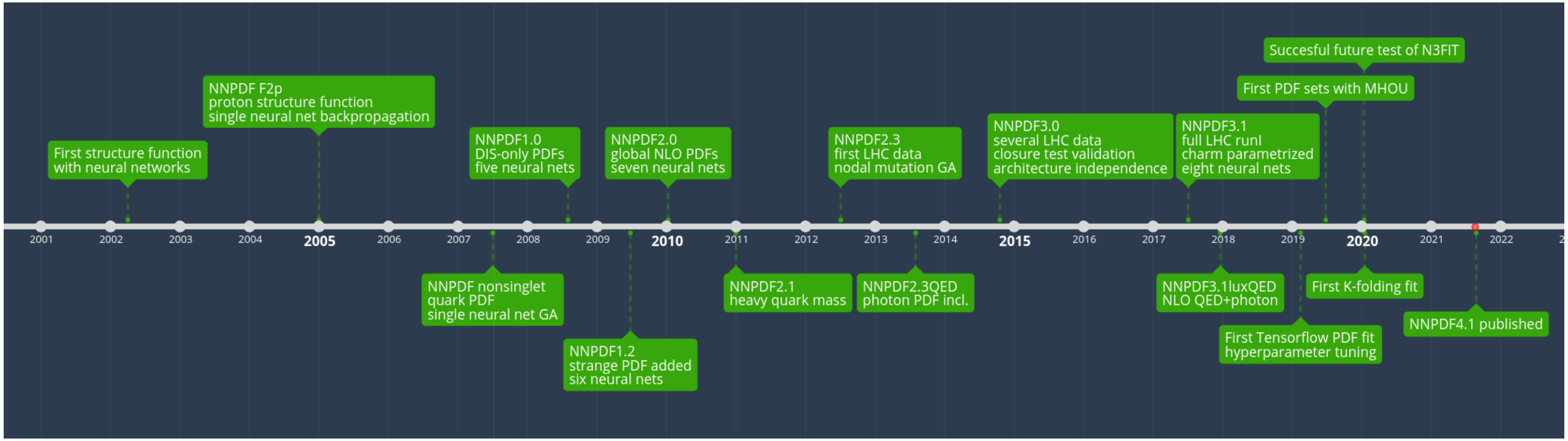
PDF DETERMINATION

- LHC CROSS SECTION:
  - $\sigma = \sum_{ij} \hat{\sigma}_{ij} \otimes f_i^{(1)} f_j^{(2)}$
  - $\hat{\sigma}_{ij}$ PARTONIC CROSS SECTION FOR WITH INCOMING PARTONS i, j
  - $f_i^{(j)}(x, Q^2)$ PDF FOR PARTON OF SPECIES i IN j-TH INCOMING PROTON
  - $\otimes$ CONVOLUTION OVER x
  - PDF DEPENDS ON Q^2 AND x, OTHER KINEMATIC VARIABLES IN $\hat{\sigma}$
- PARTONIC CROSS SECTION COMPUTED PERTURBATIVELY
- PDFs DETERMINED COMPARING σ TO DATA

the nnpdf4.0 dataset

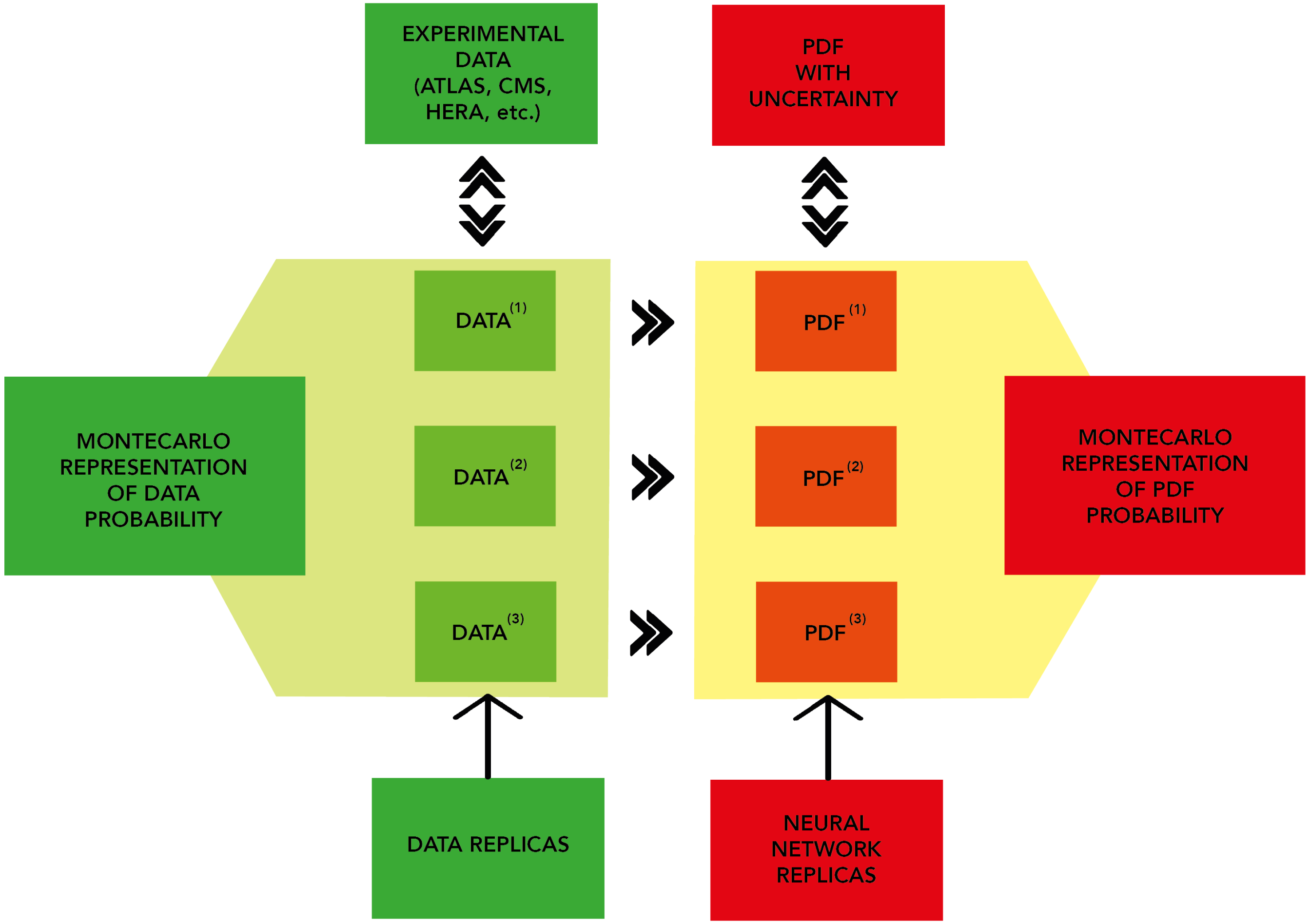


PROTON STRUCTURE AS AN AI PROBLEM: NNPDF



AI FOR PDFS: THE NNPDF APPROACH
THE FUNCTIONAL MONTE CARLO

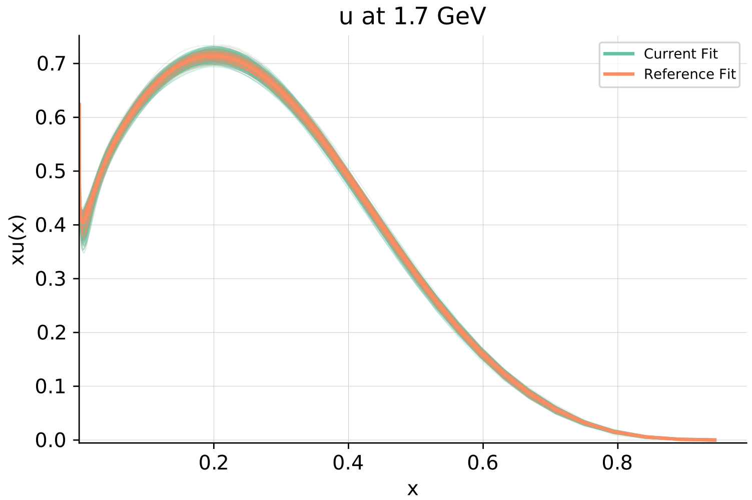
REPLICA SAMPLE OF FUNCTIONS $\Leftrightarrow$ PROBABILITY DENSITY IN FUNCTION SPACE
KNOWLEDGE OF LIKELIHHOD SHAPE (FUNCTIONAL FORM) NOT NECESSARY



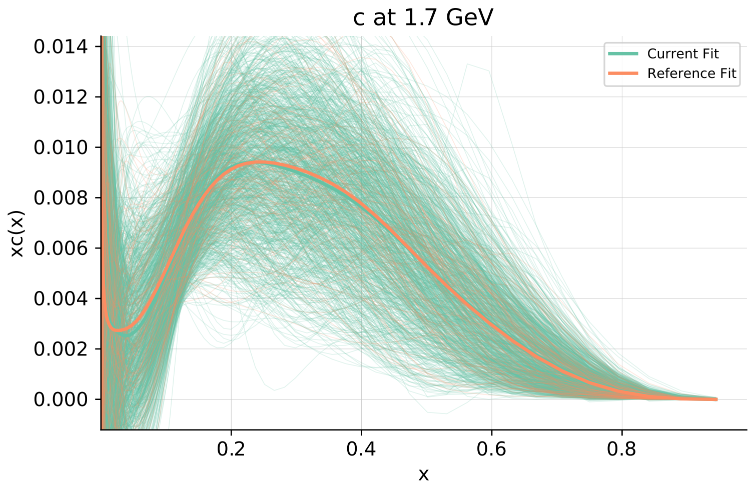
FINAL PDF SET: $f_i^{(a)}(x,\mu)$;
i =up, antiup, down, antidown, strange, antistrange, charm, gluon; $j = 1, 2, \ldots N_{\rm rep}$

THE PDFs

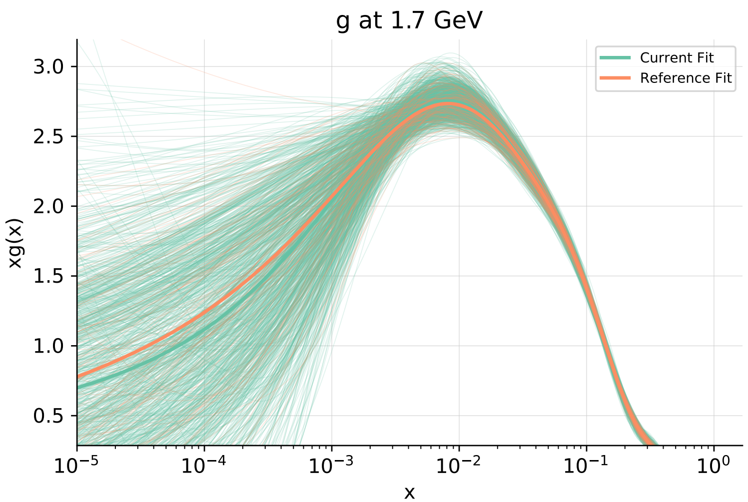
UP



CHARM



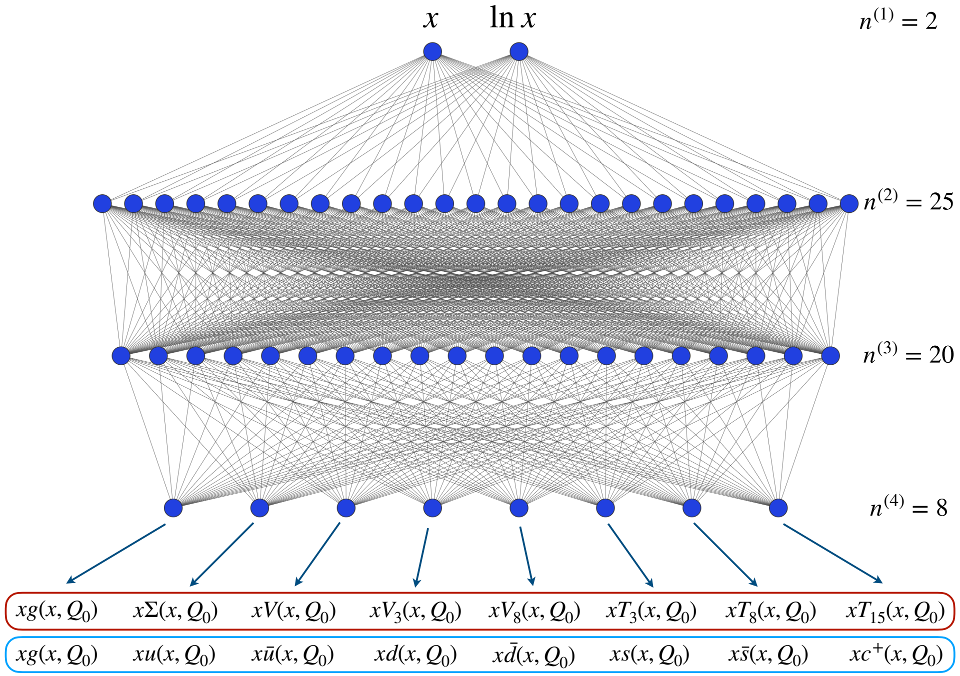
GLUON



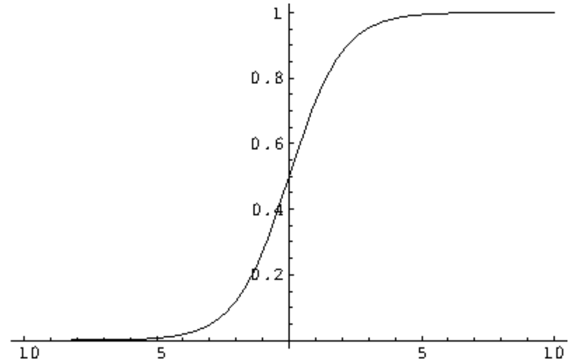
MC REPLICAS $\Leftrightarrow$ PROBABILITY DISTRIBUTION

NEURAL NETWORKS

ARCHITECTURE



ACTIVATION FUNCTIOM



$$F_{\text{out}}^{(i)}(\vec{x}_{\text{in}}) = F\left(\sum_j \omega_{ij} x_{\text{in}}^j - \theta_i\right)$$

- UNIVERSAL INTERPOLANT
- CAN REPRODUCSE ANY FUNCTIONAL FORM
- COMPLEXITY GROWS DURAING TRAINING

PARAMETERS

- WEIGHTS ω_{ij}
- THRESHOLDS θ_i

TRAINING: MINIMIZE LOSS FUNCION (E.G. χ^2)

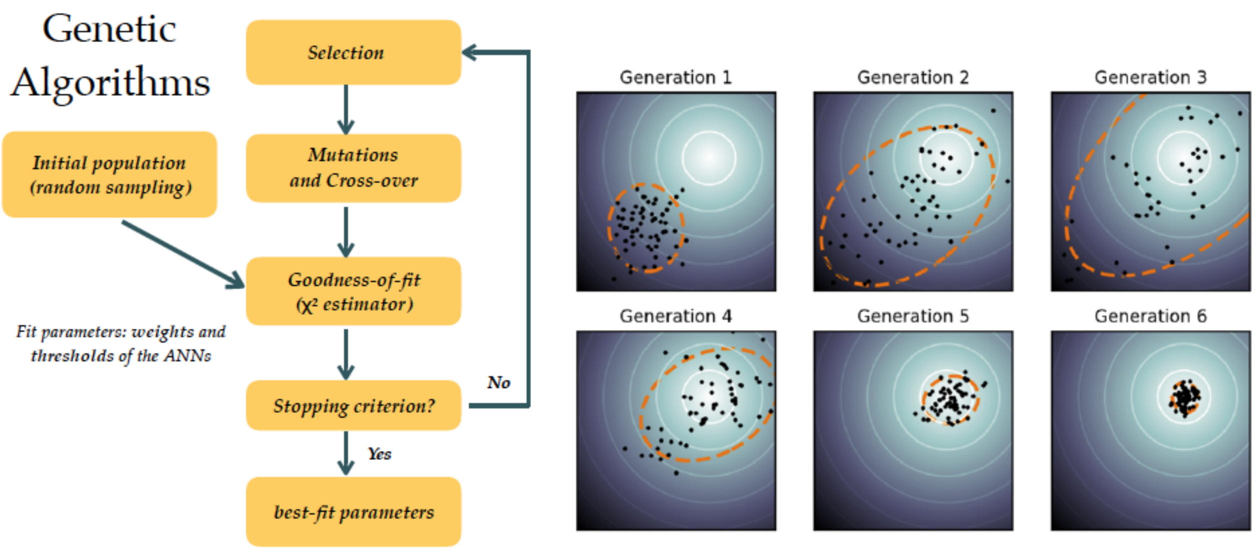
GENETIC ALGORITHMS

BASIC IDEA

- RANDOM MUTATION OF THE NN PARAMETER
- SELECTION OF THE FITTEST

FEATURES

- SLOW, COMPUTATIONALLY EXPENSIVE
- AVOIDS LOCAL MINIMA



CHOICES

- NUMBER OF MUTANTS
- MUTATION RATES
- NODAL VS LOCAL MUTATION
- ...

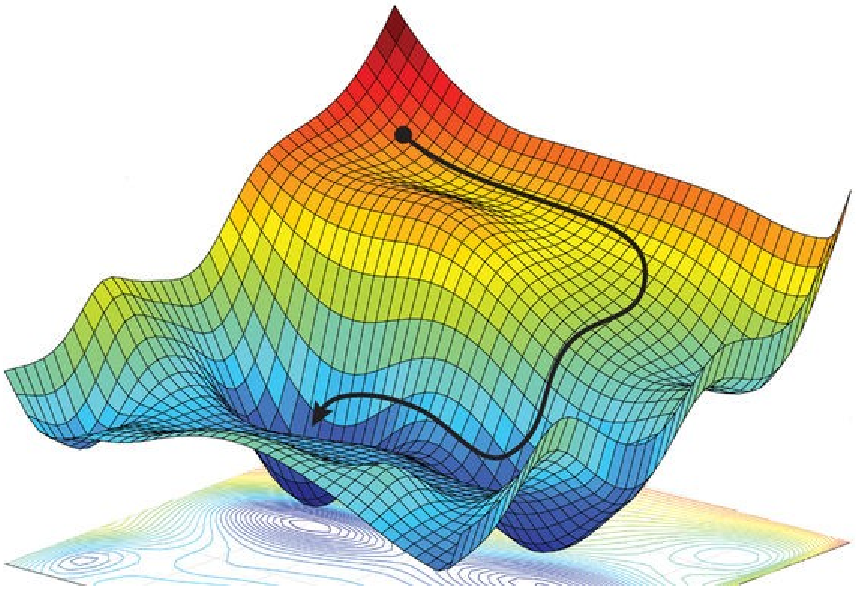
GRADIENT DESCENT

BASIC IDEA

- COMPUTE GRADIENT OF LOSS WR TO PARAMETERS
- STEEPEST DESCENT PATH

FEATURES

- LARGE MEMORY FOOTPRINT
- FAST



CHOICES

- GRADIENT SAMPLING AND BATCHES
- MOMENTUM (MEMORY OF PREVIOUS GRADIENT)
- ADAPTIVE PER-PARAMETER RATE
- ...

NNPDF4.0 PDF LEARNING: AN ANIMATION

NEURAL NETWORK TRAINING

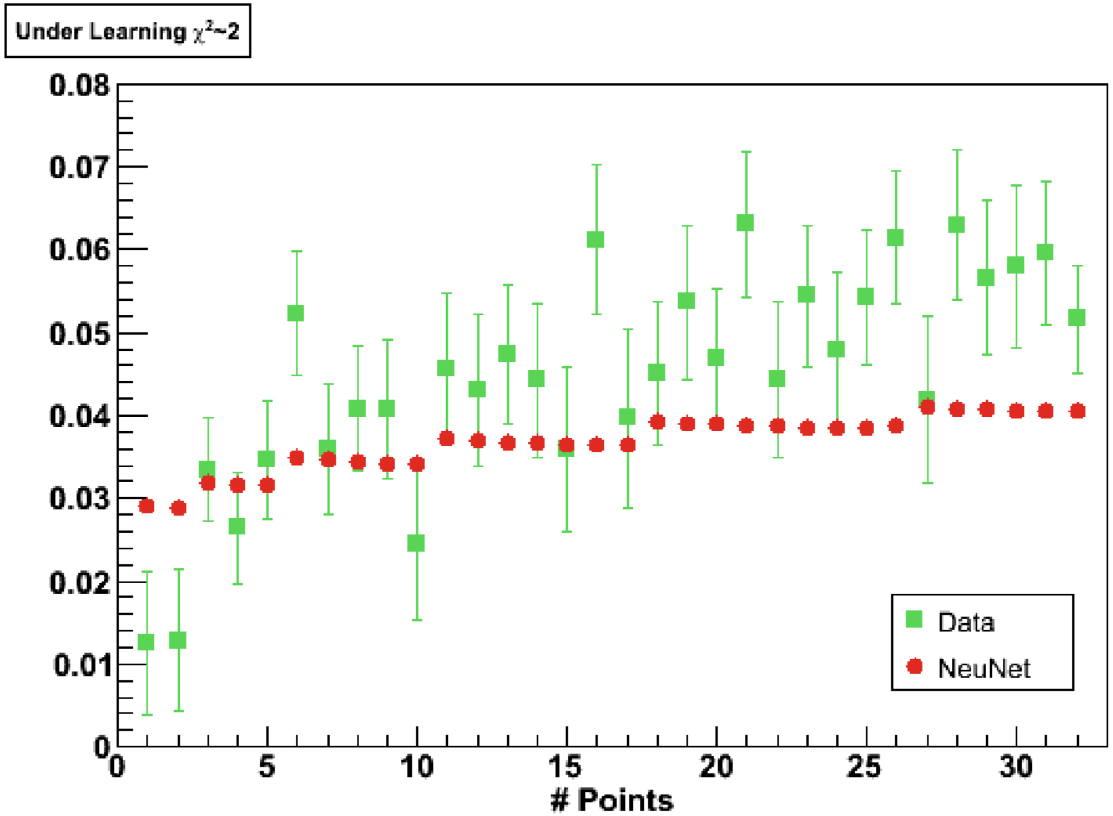
SOME FEATURES: GRADIENT DESCENT OPTIMIZATION SHOWN (NADAM)

- STRUCTURE BUILDS UP
- OUTLIERS BROUGHT UNDER CONTROL
- FEWER RANDOM FLUCTUATIONS
- UNCERTAINTIES SHRINK

NEURAL LEARNING

- COMPLEXITY INCREASES AS THE FITTING PROCEEDS
- UNTIL LEARNING NOISE
- WHEN SHOULD ONE STOP?

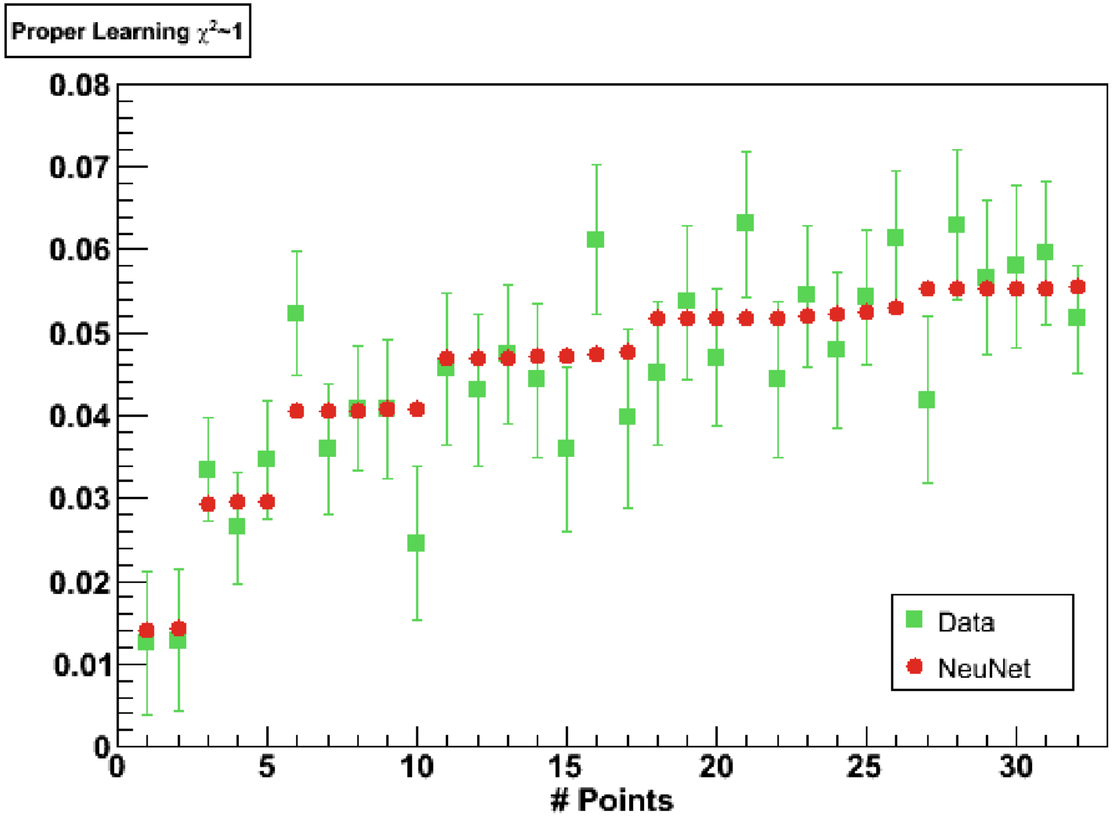
UNDERLEARNING



NEURAL LEARNING

- COMPLEXITY INCREASES AS THE FITTING PROCEEDS
- UNTIL LEARNING NOISE
- WHEN SHOULD ONE STOP?

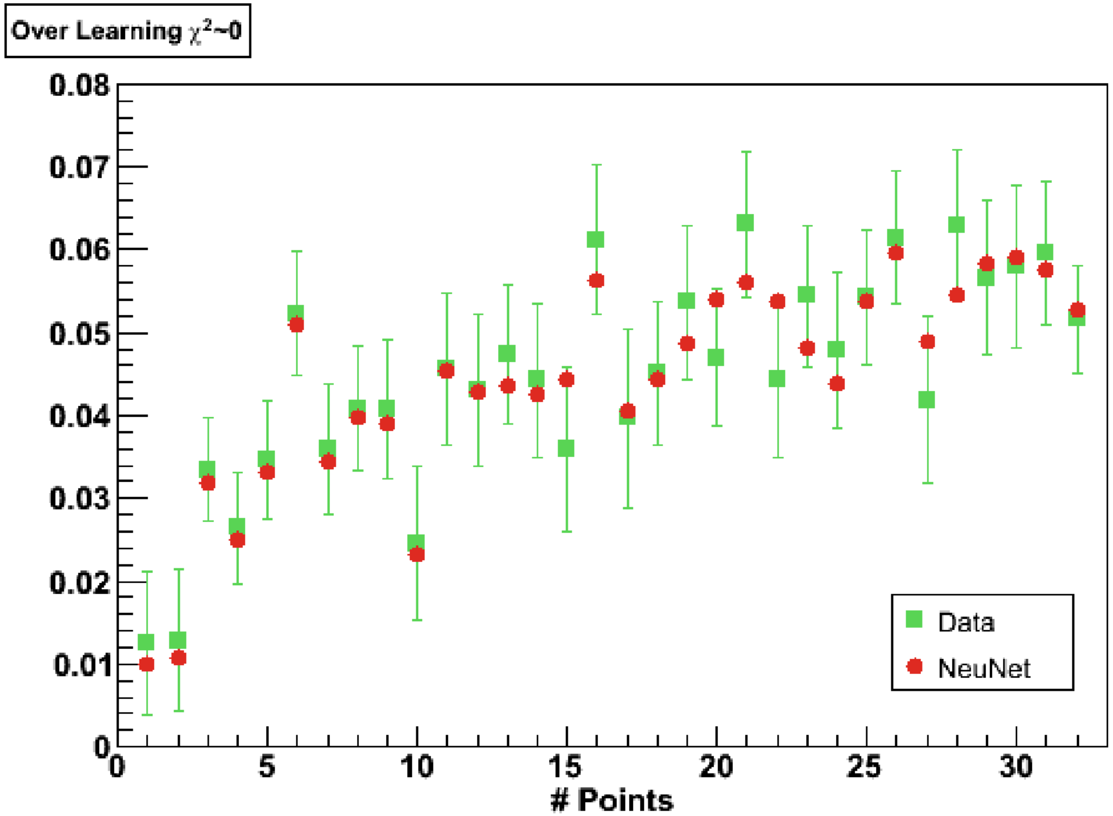
PROPER LEARNING



NEURAL LEARNING

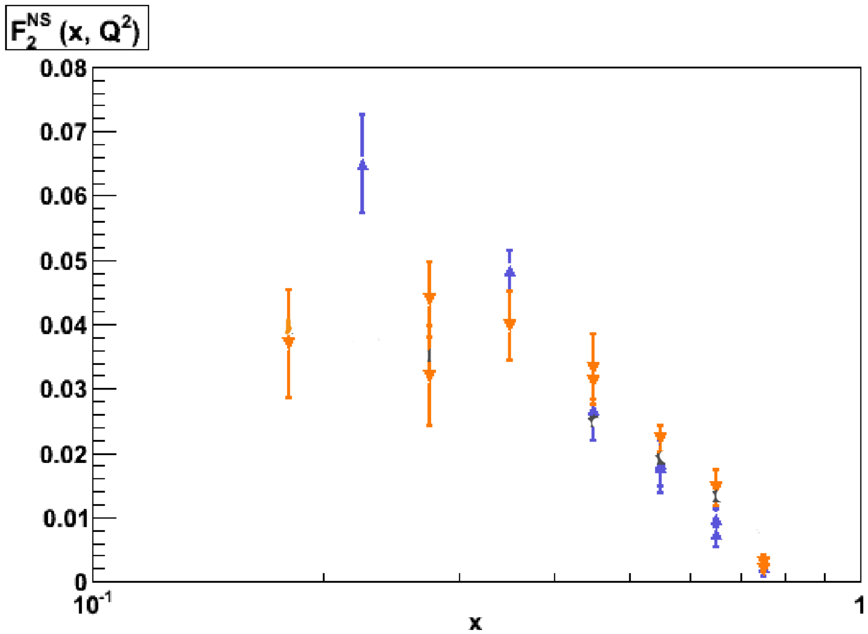
- COMPLEXITY INCREASES AS THE FITTING PROCEEDS
- UNTIL LEARNING NOISE
- WHEN SHOULD ONE STOP?

OVERLEARNING



OPTIMAL FIT: CROSS-VALIDATION

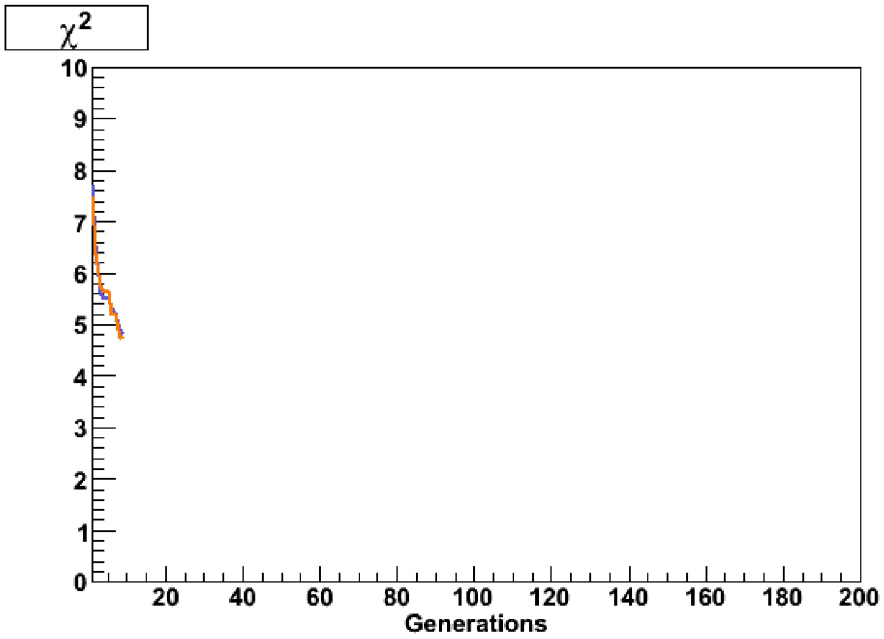
- DIVIDE THE DATA IN TWO SETS: TRAINING AND VALIDATION
- MINIMIZE THE χ^2 OF THE DATA IN THE TRAINING SET
- AT EACH ITERATION, COMPUTE THE χ^2 FOR THE DATA IN THE VALIDATION SET (NOT USED FOR FITTING)
- WHEN THE VALIDATION χ^2 STOPS DECREASING, STOP THE FIT

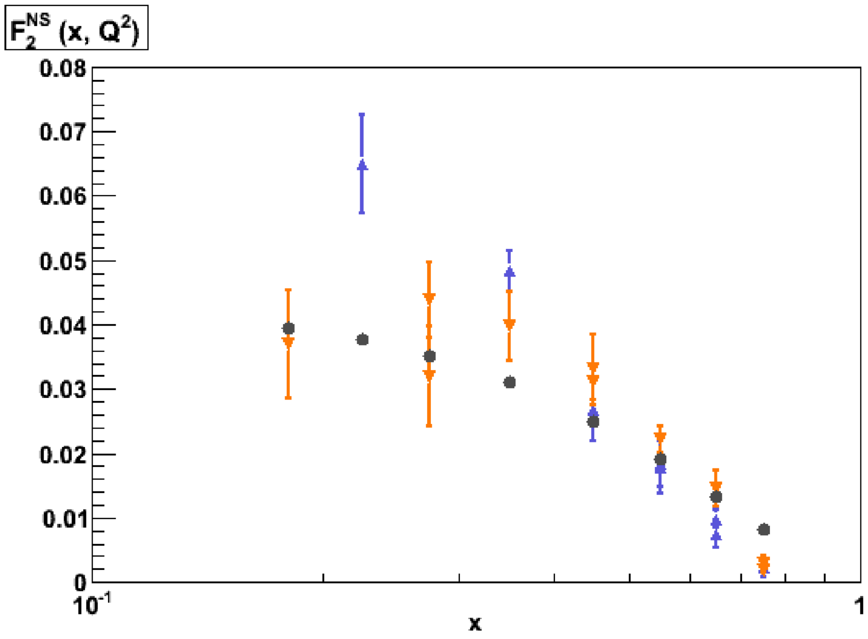


OPTIMAL FIT: CROSS-VALIDATION

- DIVIDE THE DATA IN TWO SETS: TRAINING AND VALIDATION
- MINIMIZE THE χ^2 OF THE DATA IN THE TRAINING SET
- AT EACH ITERATION, COMPUTE THE χ^2 FOR THE DATA IN THE VALIDATION SET (NOT USED FOR FITTING)
- WHEN THE VALIDATION χ^2 STOPS DECREASING, STOP THE FIT

GO!

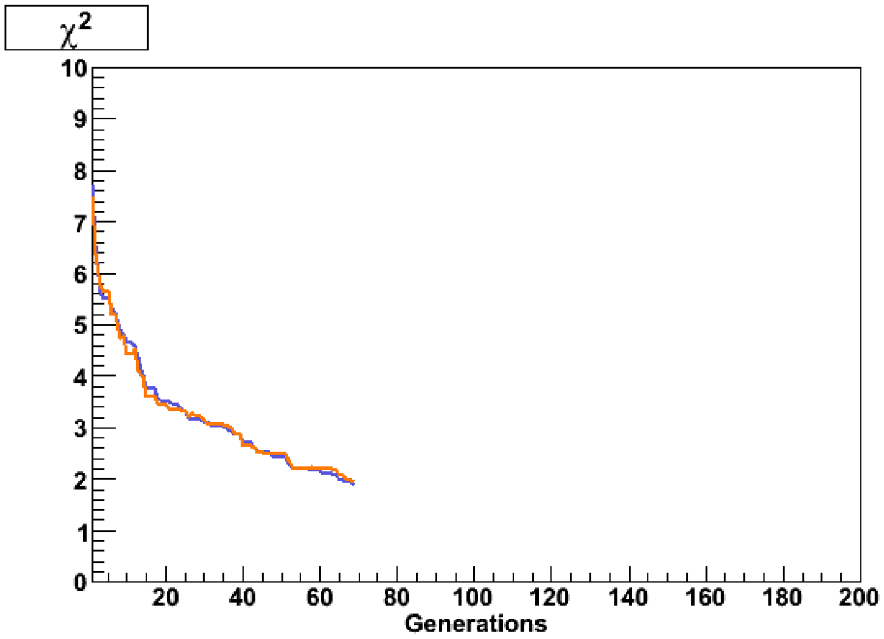


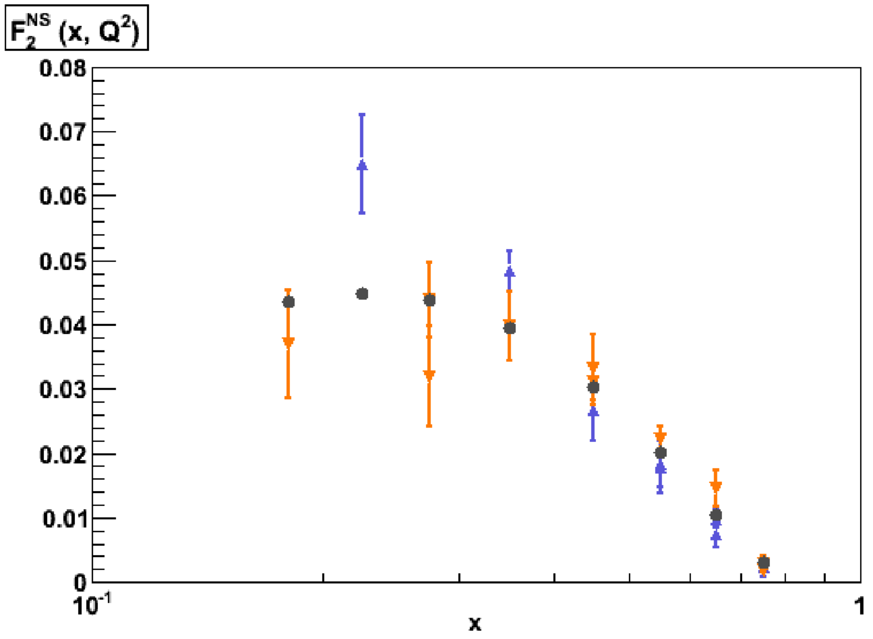


OPTIMAL FIT: CROSS-VALIDATION

- DIVIDE THE DATA IN TWO SETS: TRAINING AND VALIDATION
- MINIMIZE THE χ^2 OF THE DATA IN THE TRAINING SET
- AT EACH ITERATION, COMPUTE THE χ^2 FOR THE DATA IN THE VALIDATION SET (NOT USED FOR FITTING)
- WHEN THE VALIDATION χ^2 STOPS DECREASING, STOP THE FIT

STOP!

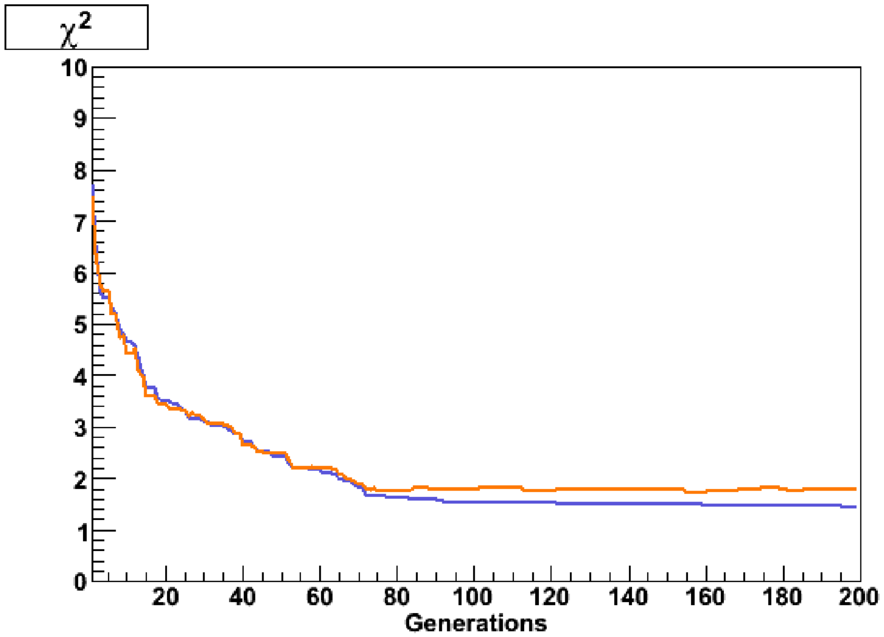


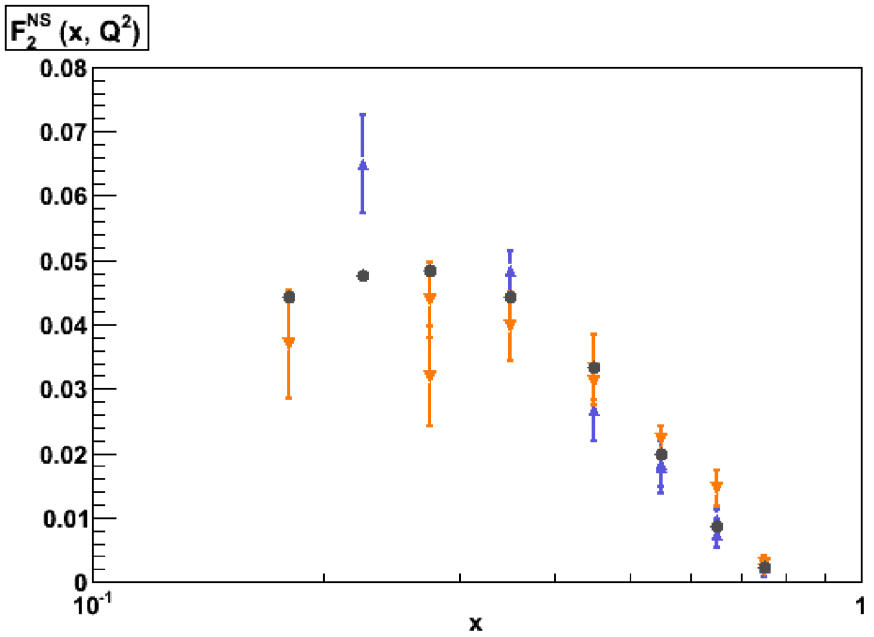


OPTIMAL FIT: CROSS-VALIDATION

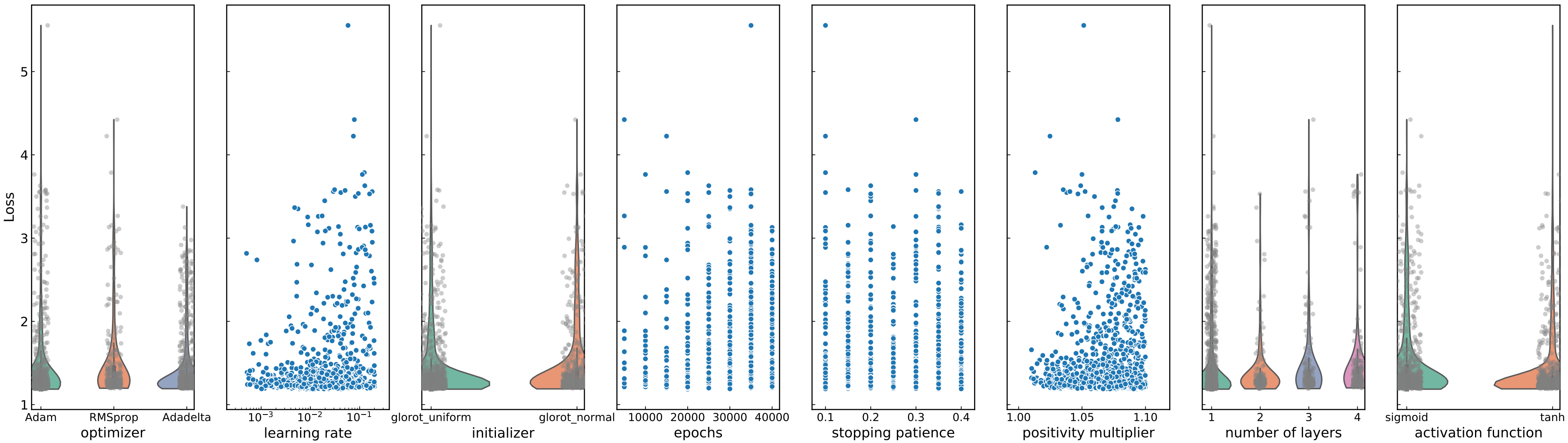
- DIVIDE THE DATA IN TWO SETS: TRAINING AND VALIDATION
- MINIMIZE THE χ^2 OF THE DATA IN THE TRAINING SET
- AT EACH ITERATION, COMPUTE THE χ^2 FOR THE DATA IN THE VALIDATION SET (NOT USED FOR FITTING)
- WHEN THE VALIDATION χ^2 STOPS DECREASING, STOP THE FIT

TOO LATE!





HYPEROPTIMIZATION



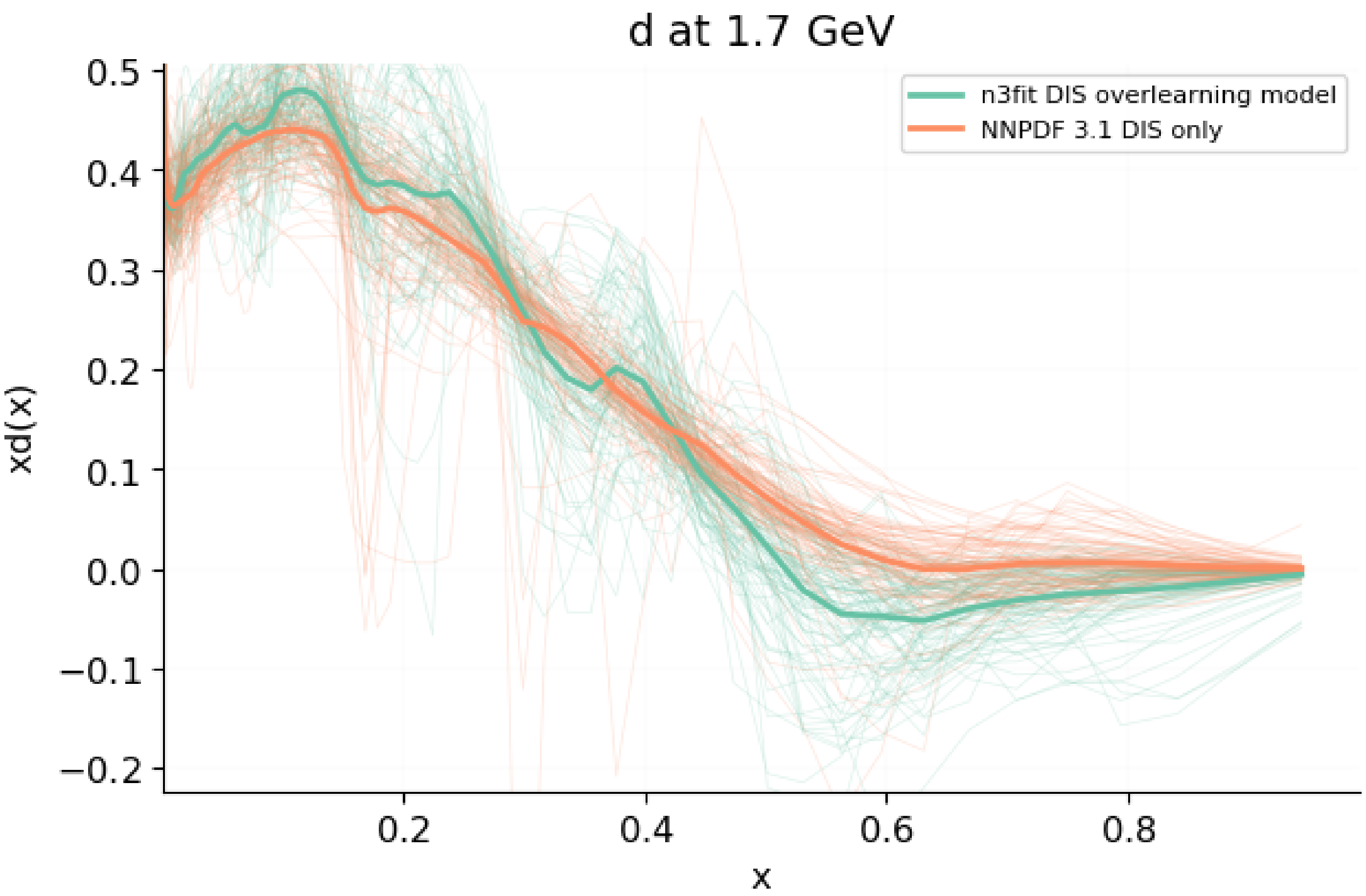
HYPEROPT PARAMETERS

Neural Network	Fit options
Number of layers (*)	Optimizer (*)
Size of each layer	Initial learning rate (*)
Dropout	Maximum number of epochs (*)
Activation functions (*)	Stopping Patience (*)
Initialization functions (*)	Positivity multiplier (*)

- SCAN PARAMETER SPACE
- OPTIMIZE FIGURE OF MERIT: VALIDATION χ^2
- BAYESIAN UPDATING

HYPEROPTIMIZATION: OVERFITTING

DOWN QUARK: HYPEROPTIMIZED VS. HAND-PICKED



- NOT HYPEROPTIMIZED: WIGGLES: FINITE SIZE $\Rightarrow$ WILL GO AWAY AS $N_{\rm rep}$ GROWS
- N3FIT: WIGGLY PDFS $\Leftrightarrow$ OVERFITTING $\Rightarrow$ WILL NOT GO AWAY ($\chi^2_{\rm train} \ll \chi^2_{\rm valid}$!!)

WHAT HAPPENED?

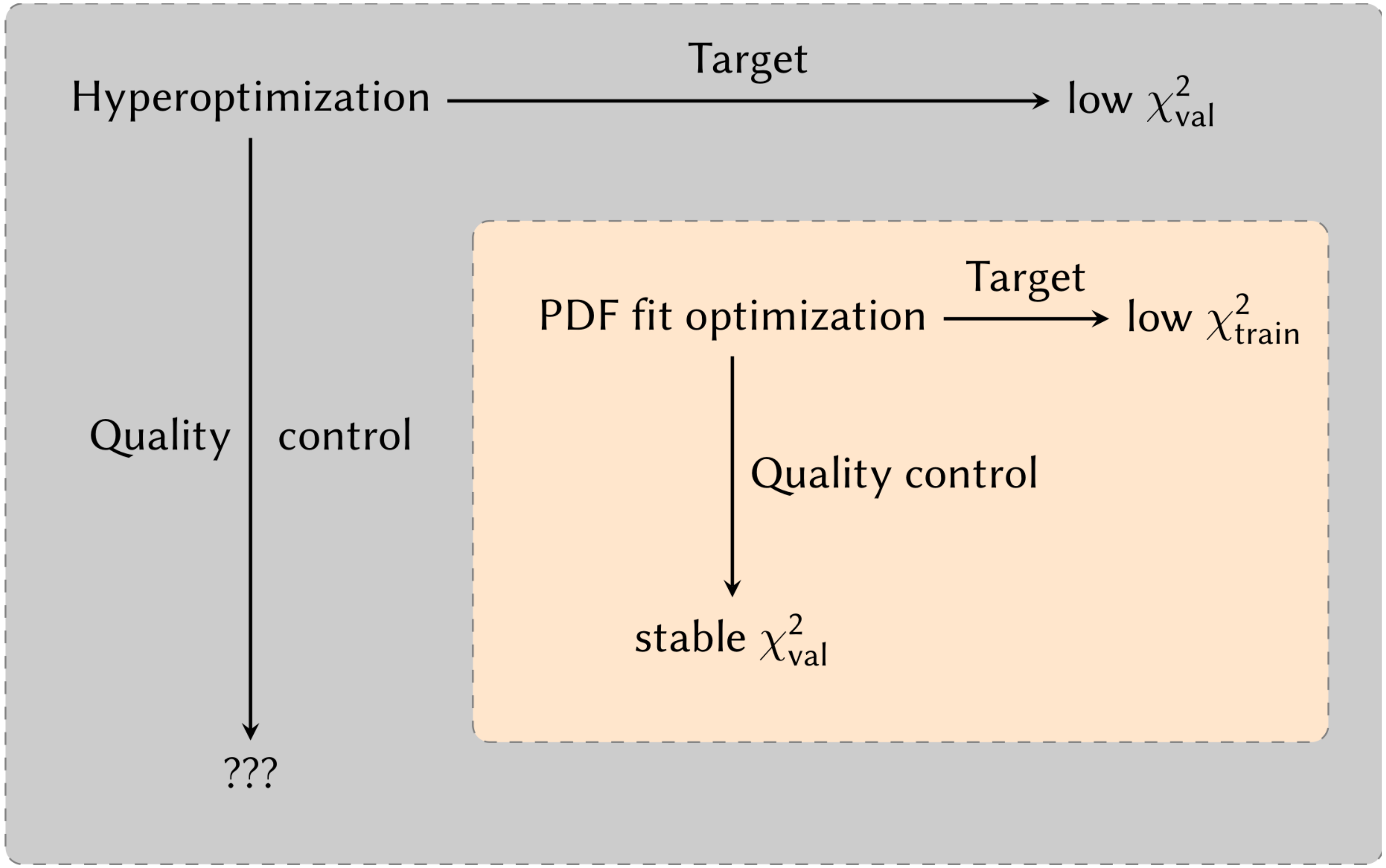
OPTIMIZATION

$$\begin{tikzcd}
\text{PDF fit optimization} \arrow[r, "\text{Target}"] \arrow[d, "\text{Quality control}"] & \text{low } \chi^2_{\text{train}} \\
\text{stable } \chi^2_{\text{val}} &
\end{tikzcd}$$

CROSS-VALIDATION SELECTS THE OPTIMAL MINIMUM

WHAT HAPPENED?

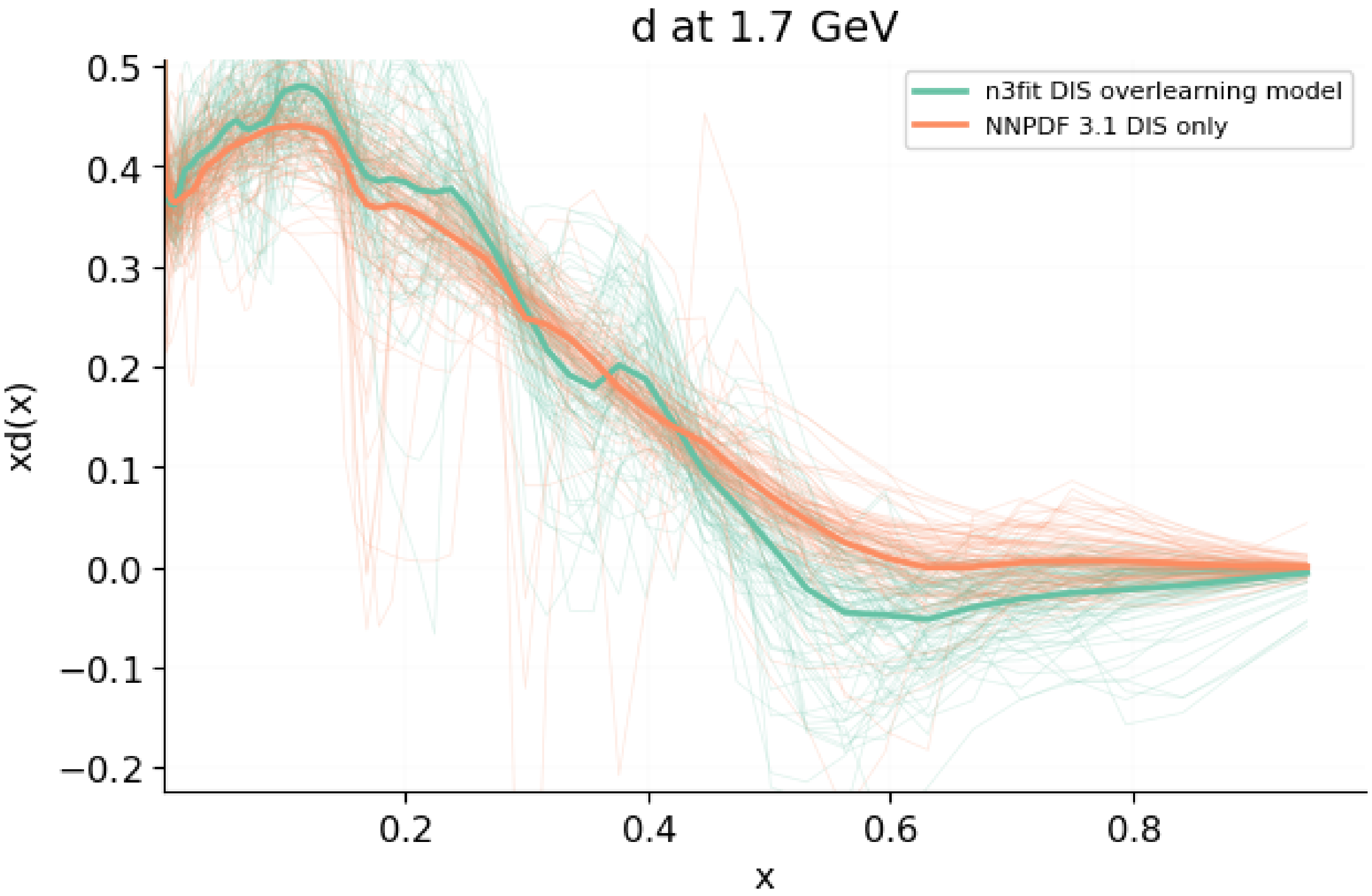
HYPEROPTIMIZATION



WE ARE MISSING A SELECTION CRITERION

HYPEROPTIMIZATION: OVERFITTING

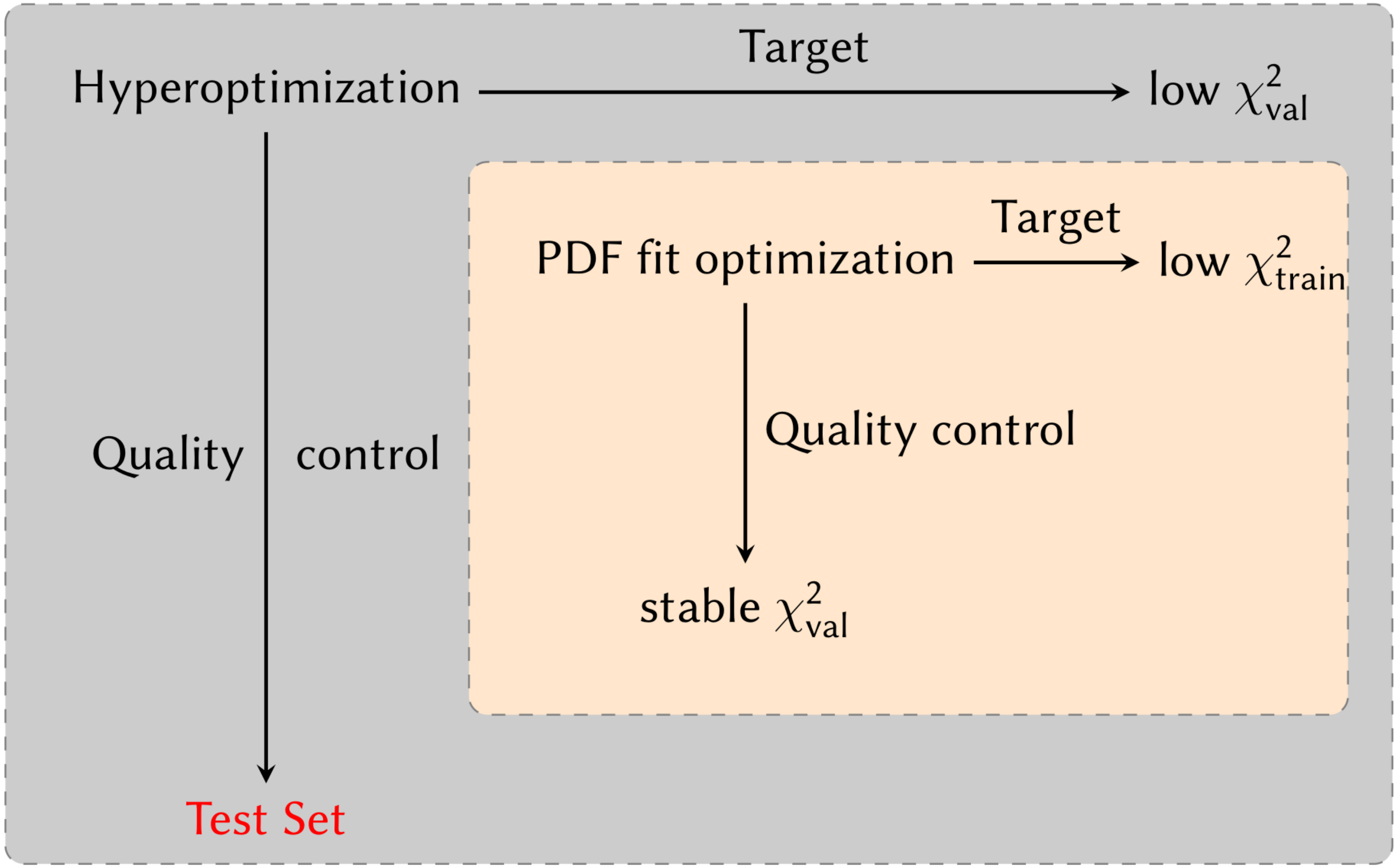
DOWN QUARK: HYPEROPTIMIZED VS. HANDPICKED



- HANDPICKED: WIGGLES: FINITE SIZE $\Rightarrow$ WILL GO AWAY AS $N_{\rm rep}$ GROWS
- N3FIT: WIGGLY PDFS $\Leftrightarrow$ OVERFITTING $\Rightarrow$ WILL NOT GO AWAY ($\chi^2_{\rm train} \ll \chi^2_{\rm valid}$!!)
- CORRELATIONS BETWEEN TRAINING AND VALIDATION DATA

THE SOLUTION

TUNED HYPEROPTIMIZATION



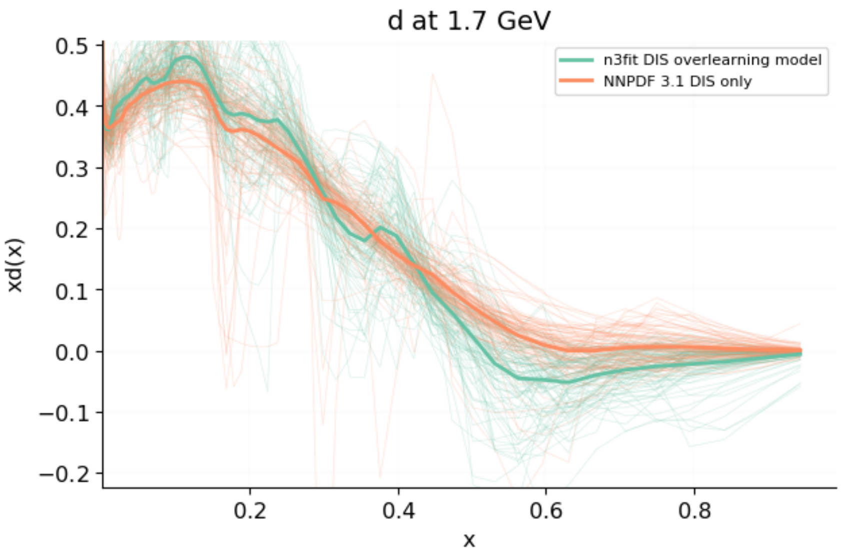
COMPARE TO A A TEST SET (NEW SET OF DATA PREVIOUSLY NOT USED AT AL)
TESTS GENERALIZATION POWER

THE TEST SET METHOD

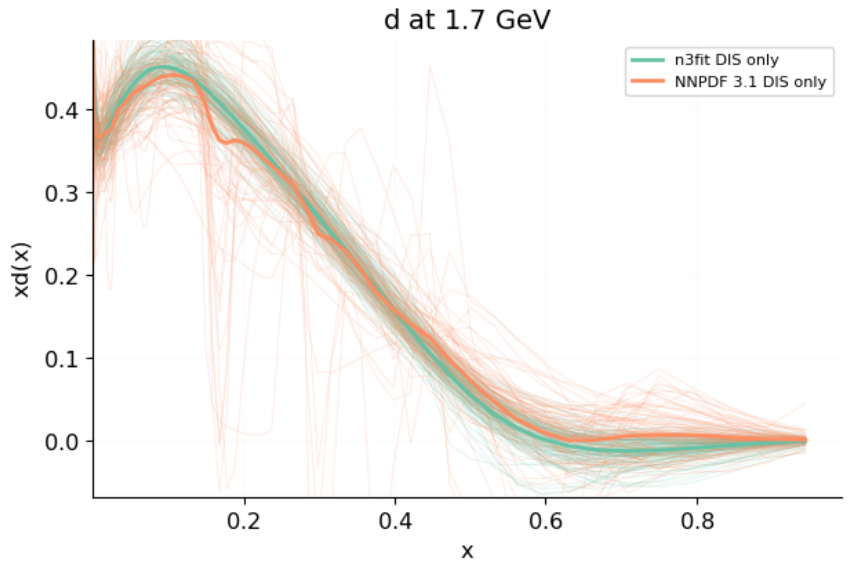
- COMPLETELY UNCORRELATED TEST SET
- OPTIMIZE ON WEIGHTED AVERAGE OF VALIDATION AND TEST
  $\Rightarrow$ NO OVERLEARNING

HYPEROPTIMIZED PDFs
DOWN QUARK

OVERFIT VS HANDPICKED



HYPEROPT VS HANDPICKED



- NO OVERFITTING
- COMPARED TO HANDPICKED
  - MUCH GREATER STABILITY $\Rightarrow$ FEWER REPLICAS FOR EQUAL ACCURACY
  - UNCERTAINTIES SOMEWHAT REDUCED

K-FOLDING

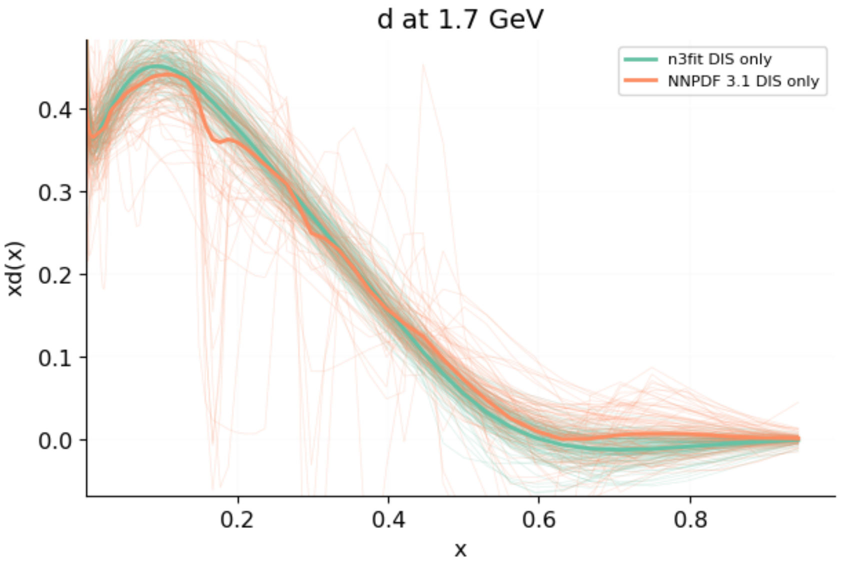
THE BASIC IDEA:

- DIVIDE THE DATA INTO n REPRESENTATIVE SUBSETS
  EACH CONTAINING PROCESS TYPES, KINEMATIC RANGE OF FULL SET
- FIT $n-1$ SETS AND USE n-TH SET AS TEST
  $\Rightarrow n$ VALUES OF $\chi^2_{\text{test, i}}$
- HYPEROPTIMIZE ON NON FITTED $\chi^2_{\text{test, i}}$
  $\rightarrow$ GOOD & STABLE GENERALIZATION

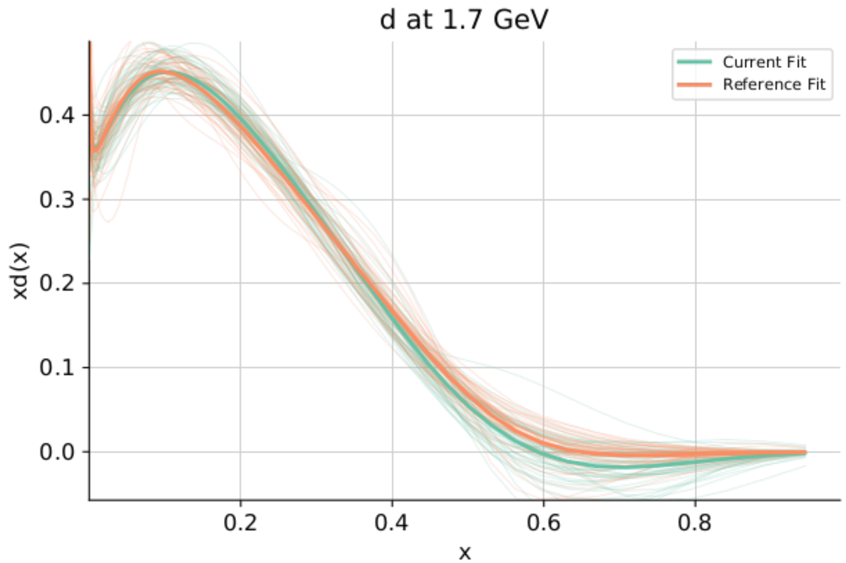
FOLDED PDFs

DOWN QUARK

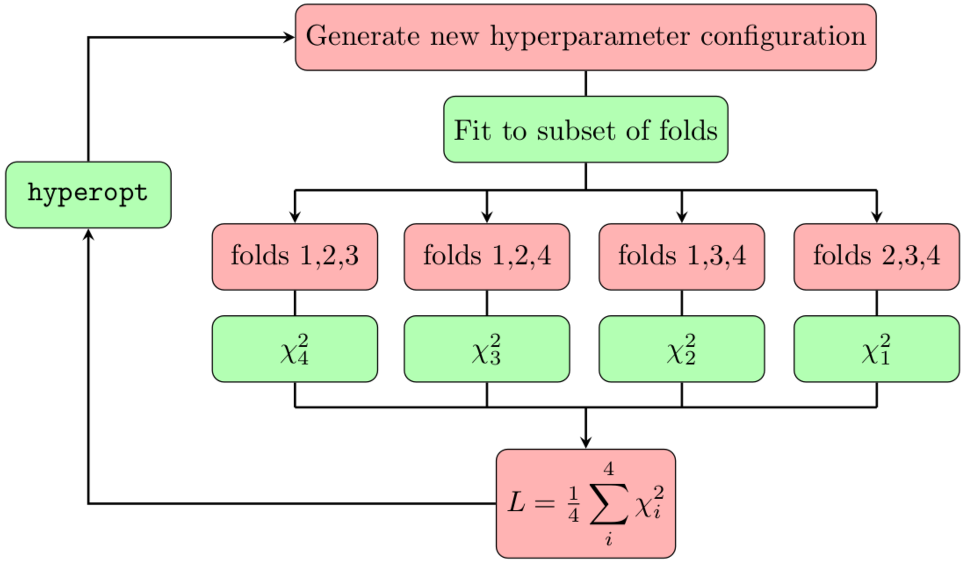
TEST-SET HYPER VS HANDPICKED



K-FOLD HYPER VS. TEST-SEY HYPER



K-FOLDING IMPLEMENTATION



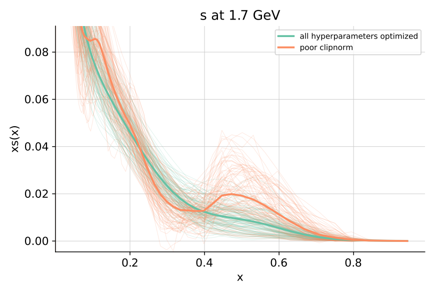
Fold 1		
CHORUS σ^{ν}_{CC}	HERA I+II inc NC e^+p 920 GeV	BCDMS p
LHCb Z 940 pb	ATLAS W, Z 7 TeV 2010	CMS Z p_T 8 TeV (p_T^{ll}, y_{ll})
DY E605 σ^{p}_{DY}	CMS Drell-Yan 2D 7 TeV 2011	CMS 3D dijets 8 TeV
ATLAS single-$\bar{t}$ y (normalised)	ATLAS single top R_t 7 TeV	CMS $t\bar{t}$ rapidity $y_{t\bar{t}}$
CMS single top R_t 8 TeV		

Fold 2		
HERA I+II inc CC e^-p	HERA I+II inc NC e^+p 460 GeV	HERA comb. $\sigma^{\mathrm{red}}_{b\bar{b}}$
NMC p	NuTeV $\sigma^{\bar{\nu}}_c$	LHCb $Z \to ee$ 2 fb
CMS W asymmetry 840 pb	ATLAS Z p_T 8 TeV (p_T^{ll}, M_{ll})	D0 $W \to \mu\nu$ asymmetry
DY E886 σ^{p}_{DY}	ATLAS direct photon 13 TeV	ATLAS dijets 7 TeV, R=0.6
ATLAS single antitop y (normalised)	CMS $\sigma^{\mathrm{tot}}_{tt}$	CMS single top $\sigma_t + \sigma_{\bar{t}}$ 7 TeV

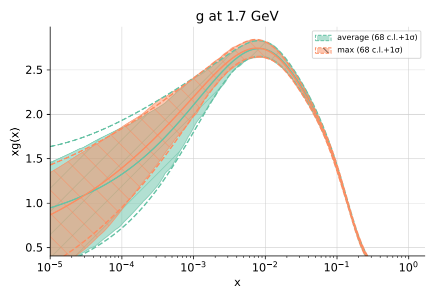
Fold 3		
HERA I+II inc CC e^+p	HERA I+II inc NC e^+p 575 GeV	NMC d/p
NuTeV σ^{ν}_c	LHCb $W, Z \to \mu$ 7 TeV	LHCb $Z \to ee$
ATLAS W, Z 7 TeV 2011 Central selection	ATLAS W^++jet 8 TeV	ATLAS HM DY 7 TeV
CMS W asymmetry 4.7 fb	DYE 866 $\sigma^{d}_{\mathrm{DY}}/\sigma^{p}_{\mathrm{DY}}$	CDF Z rapidity (new)
ATLAS $\sigma^{\mathrm{tot}}_{tt}$	ATLAS single top y_t (normalised)	CMS $\sigma^{\mathrm{tot}}_{tt}$ 5 TeV
CMS $t\bar{t}$ double diff. $(m_{t\bar{t}}, y_t)$		

Fold 4		
CHORUS $\sigma^{\bar{\nu}}_{CC}$	HERA I+II inc NC e^+p 820 GeV	LHCb $W, Z \to \mu$ 8 TeV
LHCb $Z \to \mu\mu$	ATLAS W, Z 7 TeV 2011 Fwd	ATLAS W^-+jet 8 TeV
ATLAS low-mass DY 2011	ATLAS Z p_T 8 TeV (p_T^{ll}, y_{ll})	CMS W rapidity 8 TeV
D0 Z rapidity	CMS dijets 7 TeV	ATLAS single top y_t (normalised)
ATLAS single top R_t 13 TeV	CMS single top R_t 13 TeV	

NO K-FOLDING



K-FOLDING VARIATION

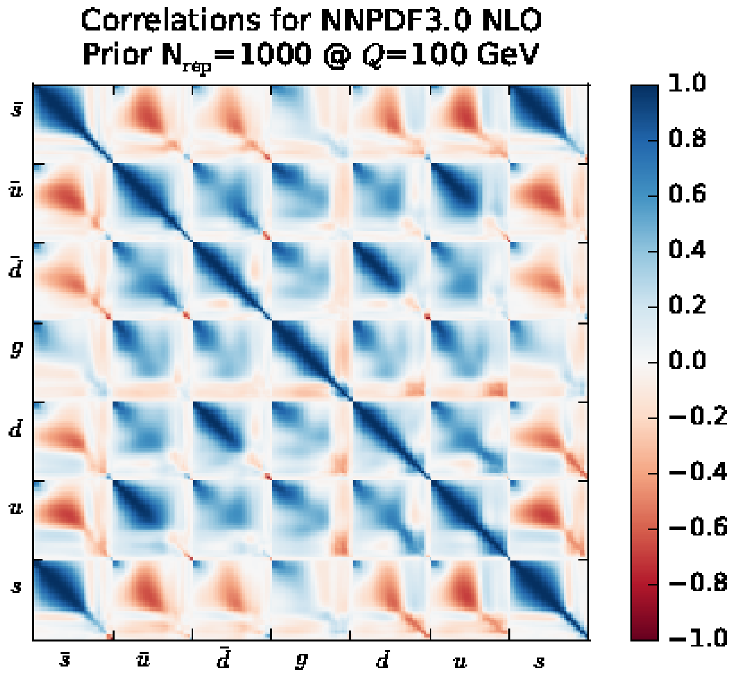


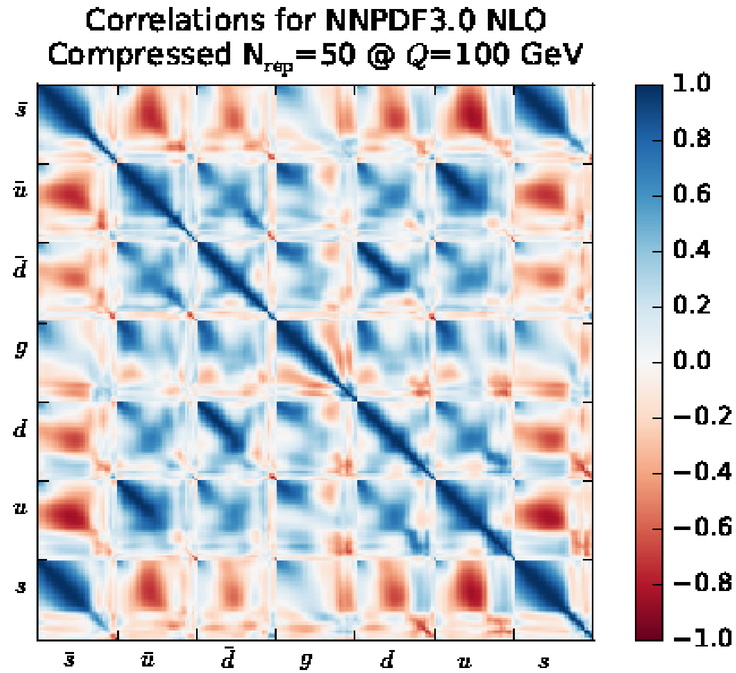
- EACH FOLD REPRODUCES FEATURES OF FULL DATASET
- DIFFERENT CHOICES POSSIBLE FOR LOSS (NON-FITTED)
  - BEST WORST
  - BEST AVERAGE
- RESULTS STABLE

MONTECARLO COMPRESSION

CAN WE REDUCE THE NUMBER OF REPLICAS?

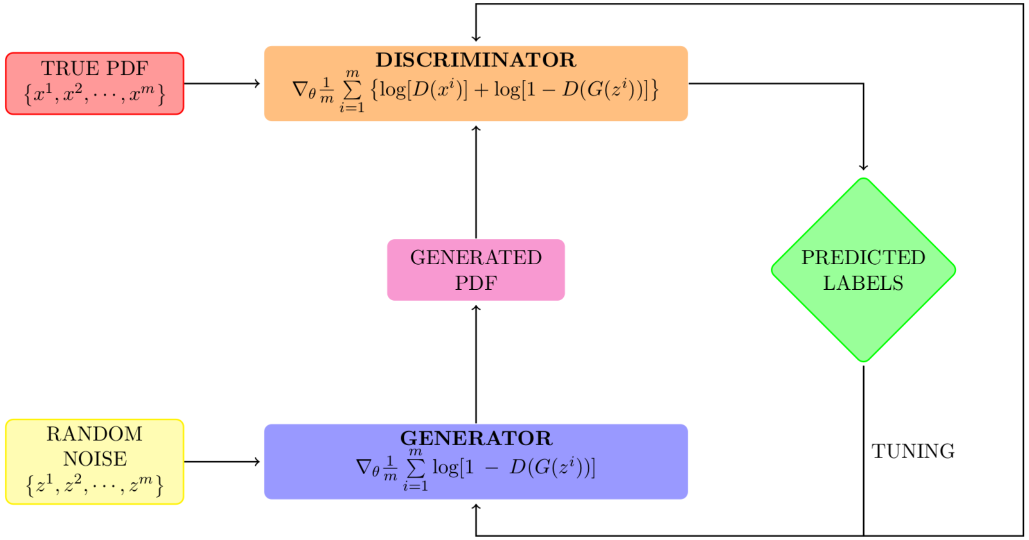
- START WITH LARGE REPLICA SAMPLE
- SELECT BY GENETIC ALGORITHM SUBSET OF REPLICAS $\Rightarrow$ STATISTICAL FEATURES OPTIMIZED TO PRIOR
- MINIMIZE LOSS: DIFFEREMCE OF MOMENTS, KL DIVERGENCE, ...
- 50 COMPRESSED REPLICA REPRODUCE 1000 REPLICA SET TO PRECENT ACCURACY





GAN ENHANCEMENT

CAN WE FURTHER REDUCE THE NUMBER OF COMPRESSED REPLICAS WITHOUT LOSS OF INFORMATION? GENERATIVE ADVERSARIAL NETWORKS

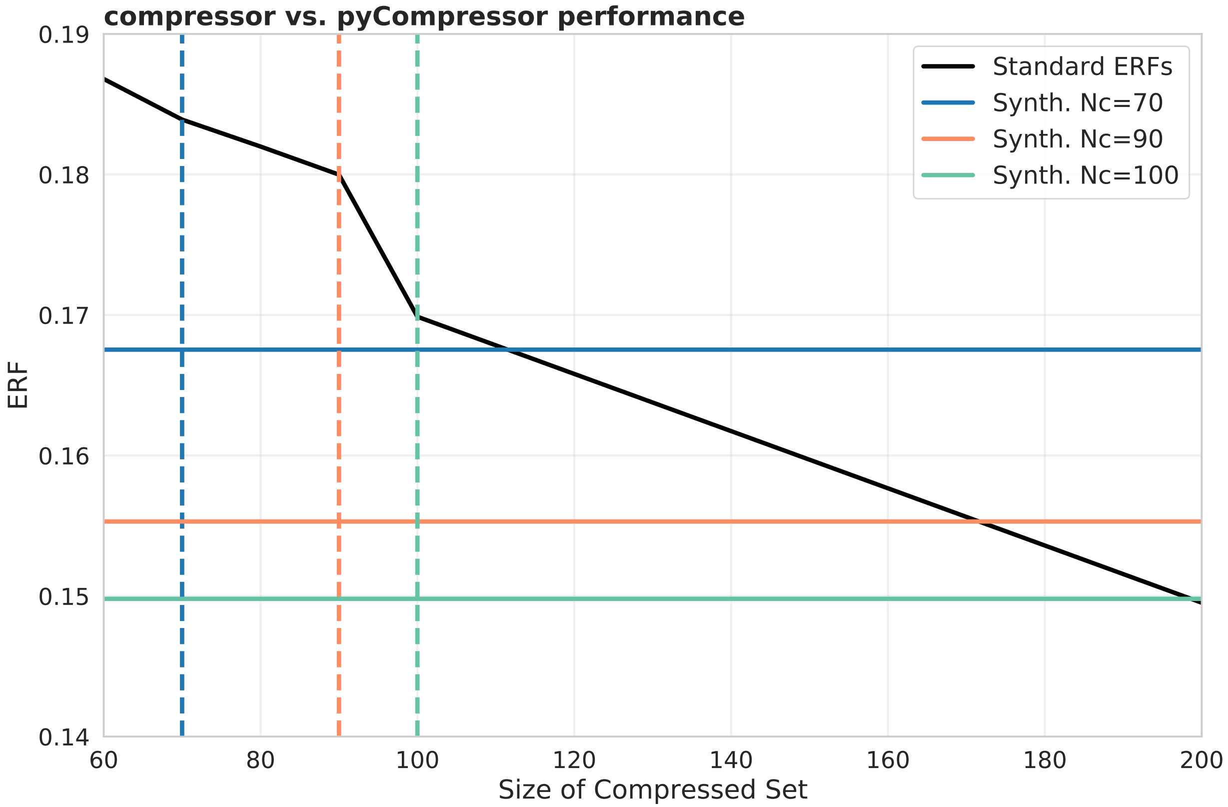


- TRAIN A NETWORK TO SIMULATE THE TRUE DISTRIBUTION (GENERATOR)
- TRAIN A NETWORK TO DISCRIMINATE TRUTH FROM SIMULATION (DISCRIMINATOR)
- TRAIN THE GENERATOR TO TRICK THE DISCRIMINATOR

GAN ENHANCEMENT

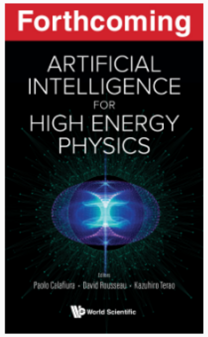
- ENHANCE THE STARTING PDF SET BY ADDING GAN-PDFS TO IT
- PERFORM COMPRESSION OF THE ENHANCED SET

PERFORMANCE



ENHANCED: NUMBER OF REPLICAS CUT IN HALF FOR SAME TARGET ACCURACY

IN LIEU OF A CONCLUSION



Artificial Intelligence for High Energy Physics

https://doi.org/10.1142/12200 | December 2021
Pages: 720

Edited By: Paolo Calafiura (*Lawrence Berkeley National Laboratory, USA*), David Rousseau (*Laboratoire de Physique des 2 Infinis Irène Joliot-Curie, France*) and Kazuhiro Terao (*SLAC National Accelerator Laboratory, USA*)

Tools   Share

Description

The Higgs boson discovery at the Large Hadron Collider in 2012 relied on boosted decision trees. Since then, high energy physics (HEP) has applied modern machine learning (ML) techniques to all stages of the data analysis pipeline, from raw data processing to statistical analysis. The unique requirements of HEP data analysis, the availability of high-quality simulators, the complexity of the data structures (which rarely are image-like), the control of uncertainties expected from scientific measurements, and the exabyte-scale datasets require the development of HEP-specific ML techniques. While these developments proceed at full speed along many paths, the nineteen reviews in this book offer a self-contained, pedagogical introduction to ML models' real-life applications in HEP, written by some of the foremost experts in their area.

Contents: