



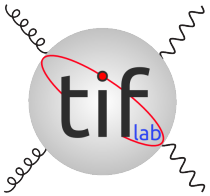
# REGRESSION NETWORKS: PRECISION AND UNCERTAINTY ESTIMATION

STEFANO FORTE
UNIVERSITÀ DI MILANO & INFN

WITH TUTORIALS BY

TOMMASO GIANI
NIKHEF

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI FISICA





MACHINE LEARNING IN PARTICLE THEORY

MITP, JULY 4, 2023

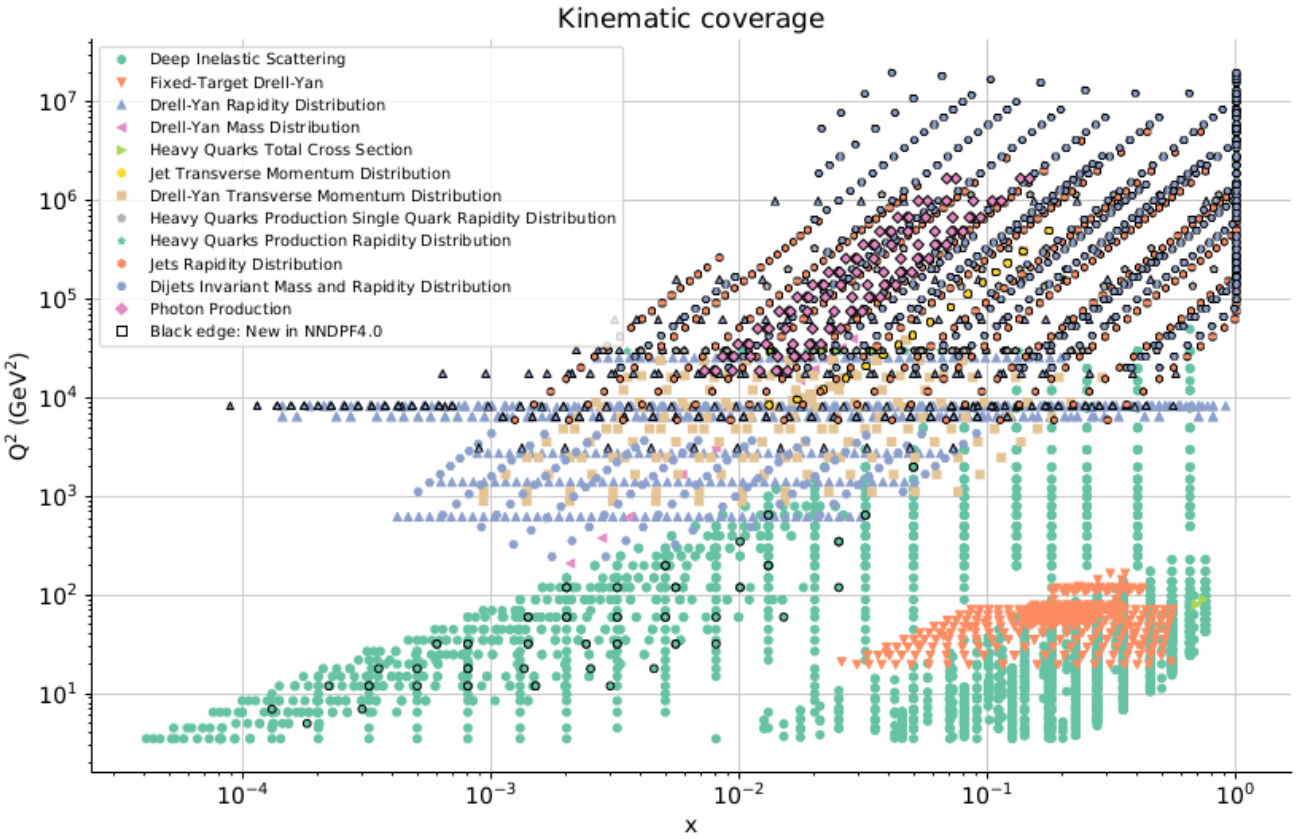
# II: UNCERTAINTIES AND REGRESSION

- INVERSE PROBLEMS
  - BAYESIAN INFERENCE
  - MAP ESTIMATION
- MONTE CARLO UNCERTAINTIES
  - MONTE CARLO REGRESSION
  - IMPORTANCE SAMPLING AND REWEIGHTING
- NEURAL NETWORK MODELS
  - ARCHITECTURE AND ACTIVATION FUNCTIONS
  - PREPROCESSING AND FEATURE SCALING
- NEURAL NETWORK TRAINING
  - STOCHASTIC VS. DETERMINISTIC MINIMZATION
  - REQUIREMENTS
- THE THEORY PIPELINE
  - REQUIREMENTS
  - THE PIPELINE

# PDF DETERMINATION

DATA $\rightarrow$ PARTON DISTRIBUTIONS

## Experimental data in NNPDF4.0




More than 4000 datapoints!

New processes:

- direct photon
- single top
- dijets
- W+jet
- DIS jet

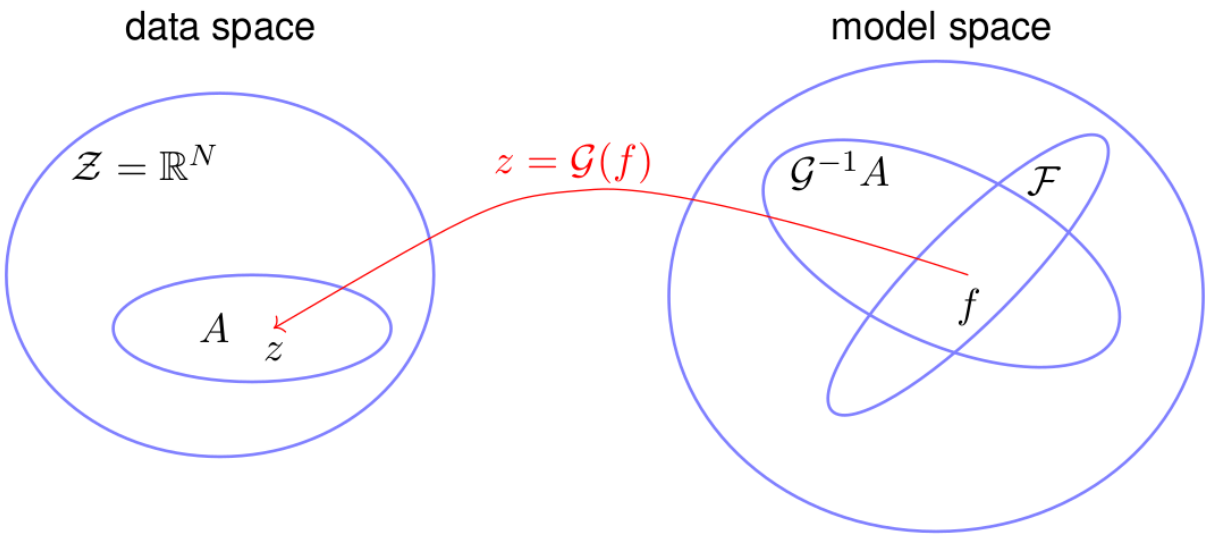
## ISSUES:

- PROBABILITY IN THE SPACE OF FUNCTIONS: HOW TO REPRESENT IT?
- UNCERTAINTY ON FUNCTIONS: HOW TO AVOID POLYNOMIAL REGRESSION?

# INVERSE PROBLEMS

# BAYESIAN INFERENCE

inverse problems




**Bayes theorem**

$$p(f|A) \propto p(A|f)p(f)$$

- OBSERVE DATA $z$ (E.G. $Z$ CROSS-SECTION), WITH UNCERTAINTIES $\Rightarrow$ COVMAT $C_z$
- DATA DEPEND ON MODEL (E.G. GLUON PDF) $f$ THROUGH FUNCTION $\mathcal{G}(f)$, WITH PRIOR PROBABILITY IN SPACE OF MODELS $\Rightarrow$ COVMAT $C_f$
- THEORY PREDICTIONS $t = \mathcal{G}(f)$, POSSIBLY SHIFTED BY THEORY ERROR $\Rightarrow$ COVMAT $C_t$
- LINEARIZATION AND SAMPLING: $f = f_i$; $t_i = \mathcal{G}_{ij} f_j$

# GAUSSIAN REGRESSION

## PRIOR PROBABILITIES

- (PRIOR) PROBABILITY OF DATA TRUE VALUES $z$ VS. OBSERVED VALUES $z_0$
  $\pi^0(z|z_0, C_z) \propto \exp\left(-\frac{1}{2}\left(|z - z_0|^2_{C_z}\right)\right.$
- (PRIOR) PROBABILITY OF TRUTH MODEL $f$ VS CENTRAL ANSATZ $f_0$
  $\pi^0(f|f_0, C_f) \propto \exp\left(-\frac{1}{2}\left(|f - f_0|^2_{C_f}\right)\right.$
- (PRIOR) PROBABILITY OF TRUTH PREDICTION $t$ VS PREDICTION FROM TRUTH MODEL $\mathcal{G}(f)$:
  $\pi^0(t|\mathcal{G}(f), C_t) \propto \exp\left(-\frac{1}{2}\left(|t - \mathcal{G}(f)|^2_{C_t}\right)\right.$

## MARGINALIZING THEORY UNCERTAINTIES

- TRUE DATA = TRUE THEORY PREDICTION: $\delta(t - z)$
- PROBABILITY OF OBSERVED DATA $z$ VS AVAILABLE THEORY PREDICTIONS $\mathcal{G}(f)$
  $\pi(z_0|\mathcal{G}(f), C_{\rm tot}) = \int d^N t d^N z \delta(t - z)\pi^0_z(z|z_0, C_z)\pi^0_{\mathcal{G}}(t|\mathcal{G}(f), C_t) \propto$
  $\exp\left(-\frac{1}{2}\,|z_0 - \mathcal{G}(f)|^2_{C_z+C_t}\right)$
- THEORY AND EXP COVARIANCE MATRICES ADD: $C_{\rm tot} = C_z + C_t$
  NATURE IS IGNORANT OF PHYSICISTS' ERRORS

## POSTERIOR MODEL PROBABILITY

- PROBABILITY OF MODEL $f$ VS OBSERVED DATA $z_0$
  $\pi(f) \propto \exp\left(-\frac{1}{2}\,|z_0 - \mathcal{G}(f)|^2_{C_{\rm tot}} - \frac{1}{2}\,|f - f_0|^2_{C_f}\right)$

# MAP ESTIMATION

- MAXIMUM A POSTERIORI PROBABILITY

$$f_* = \text{argmin}\left(-\frac{1}{2}\,|z_0 - \mathcal{G}(f)|^2_{C_{\text{tot}}} - \frac{1}{2}\,|f - f_0|^2_{C_f}\right)$$

- FOR A LINEAR MAP $z_i = \mathcal{G}_{ij} f_j \Rightarrow f_*$ GAUSSIANLY DISTRIBUTED
  - WITH COVMAT $\bar{C}^{-1} = \left(\mathcal{G}^T C_z^{-1} \mathcal{G} + C_f^{-1}\right) \Leftrightarrow$ UNCERTAINTY
  - CENTERED AT $\bar{f} = \bar{C}\left(\mathcal{G}^T C_z^{-1} z_0 + C_f^{-1} f_0\right) \Leftrightarrow$ BEST-FIT MODEL

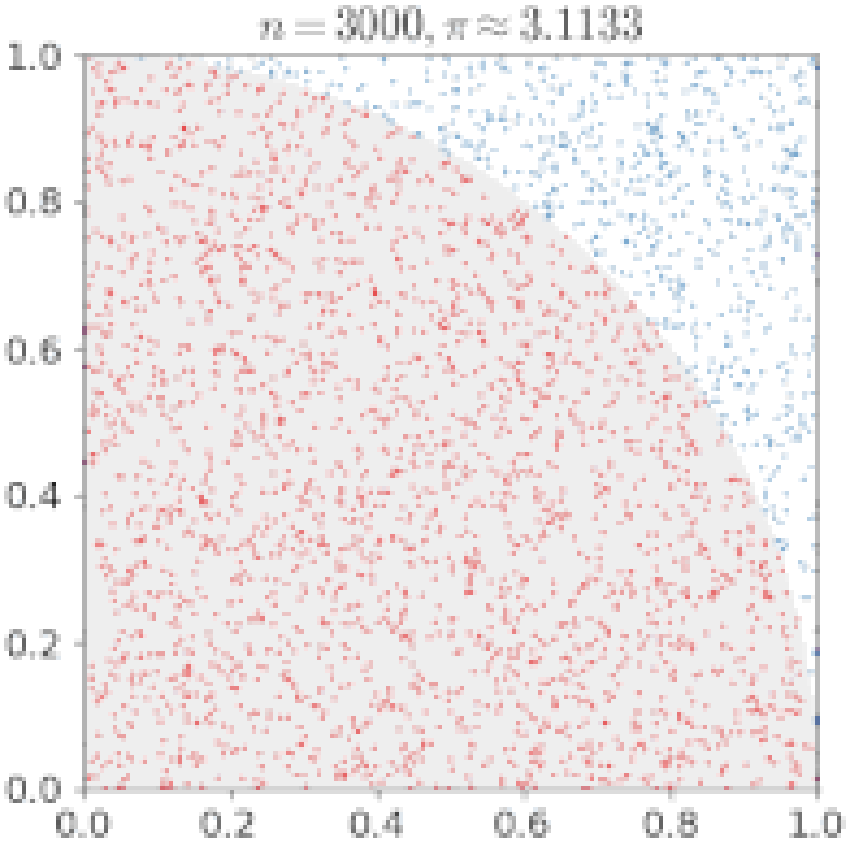
PROBLEMS:

- MINIMIZATION? $\Rightarrow$ NUMERICAL/ALGORITHMIC
- UNCERTAINTIES??

# MONTE CARLO UNCERTAINTIES

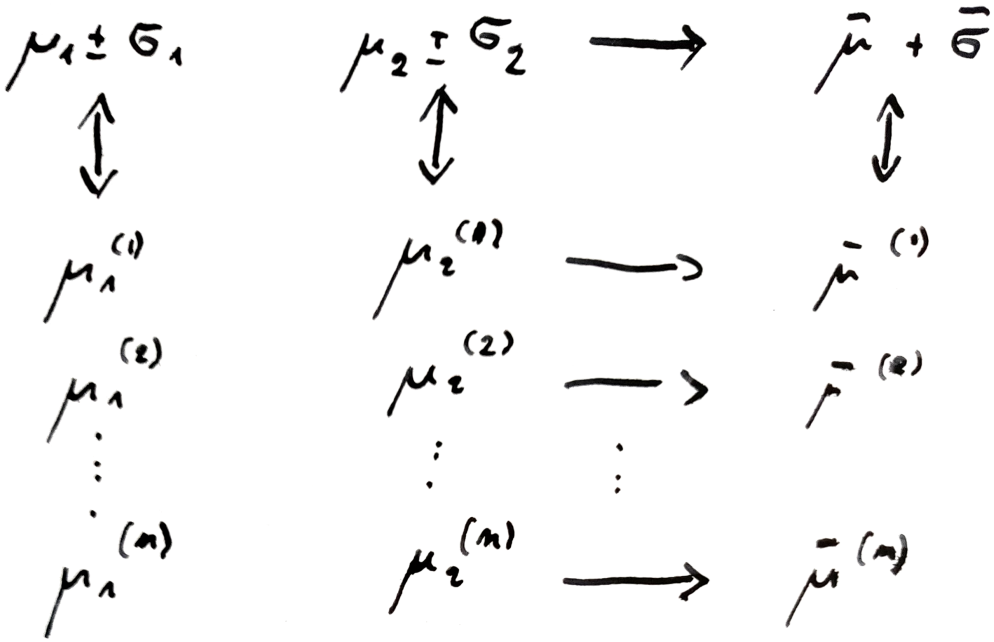
# THE MONTECARLO METHOD TO THE RESCUE

MONTECARLO COMPUTATION OF $\pi$



# MONTE CARLO COMBINATION

- TWO DATA WITH UNCERTAINTY $z_i = \mu_i \pm \sigma_i$
- SAMPLE OF DATA REPLICAS $\mu_i^{(k)} \to \mu_i = \langle \mu_i^{(k)} \rangle$; $\sigma_i^2 = \langle \left( \mu_i^{(k)} - \mu_i \right)^2 \rangle$.
- MAP COMBINATION $\mu_1^{(k)}, \mu_2^{(k)} \to \bar{\mu}^{(k)}$
- $\mu^{(k)}$ REPLICA SAMPLE $\Rightarrow$ REPRESENTATION OF MAP PROBABILITY $\bar{\mu} \pm \bar{\sigma}$
  $\bar{\mu} = \langle \bar{\mu}^{(k)} \rangle$; $\bar{\sigma}^2 = \langle \left( \bar{\mu}^{(k)} - \bar{\mu} \right)^2 \rangle$.



# MONTE CARLO REGRESSION

## PROOF FOR GAUSSIAN CASE

### MC REPLICAS

- GENERATE REPLICAS $\bar{z}^{(k)} = z_0 + \eta^{(k)}$, $\eta$ DRAWN FROM GAUSSIAN CENTERED AT ZERO WITH COVMAT $C_z$
  SO PROBABILITY OF DATA REPLICAS $z$ VS OBSERVED DATA $z_0$
  $\pi^0(\bar{z}|z_0, C_z) \propto \exp\left(-\frac{1}{2}\,|\bar{z} - z_0|^2_{C_z}\right)$

### MODEL REPLICAS

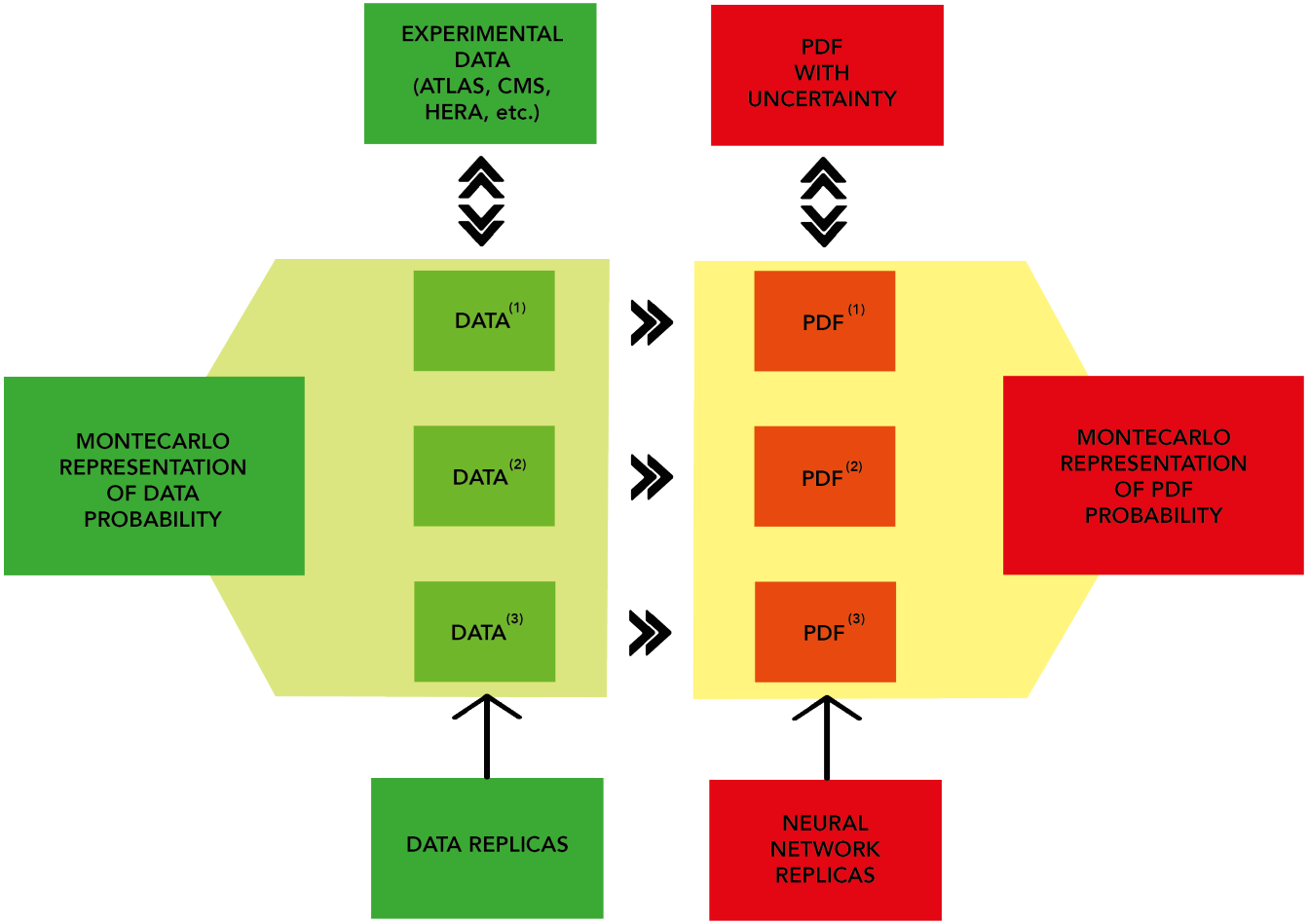
- DETERMINE MAP SOLUTION ON A REPLICA-BY-REPLICA BASIS
  $f_*{}^{(k)} = \left(\mathcal{G}^T C_z^{-1} \mathcal{G}\right)^{-1} \left(\mathcal{G}^T C_z^{-1} z_0 + \mathcal{G}^T C_z^{-1} \eta^{(k)}\right)$

- DISTRIBUTION OF MC REPLICAS GAUSSIAN WITH THE SAME MEAN AND COVARIANCE AS PREVIOUS MAP SOLUTION
  - COVMAT $\bar{C}^{-1} = \left(\mathcal{G}^T C_z^{-1} \mathcal{G}\right) \Leftrightarrow$ UNCERTAINTY
  - CENTERED AT $\bar{f} = \bar{C}\left(\mathcal{G}^T C_z^{-1} z_0\right) \Leftrightarrow$ BEST-FIT MODEL

- COINCIDES WITH MAP ESTIMATE FOR FLAT PRIOR

# MONTE CARLO REGRESSION
# CHOOSE THE REGRESSOR (NEURAL NETWORKS)

## MONTE CARLO REPRESENTATION



# MONTE CARLO REPRESENTATION
# IMPORTANCE SAMPLING

- PROBABILITY DISTRIBUTION SAMPLED DIRECTLY
  $\Rightarrow$ ALL INSTANCES EQUALLY WEIGHTED $\langle f \rangle = \frac{1}{N} \sum_{I=1}^{N} f_i$
- CONTRAST TO A MODEL DEPENDING ON PARAMETERS $\theta_i$ WITH KNOWN PROBABILITY $p(\theta_i)$:
- $\langle f \rangle = \frac{1}{N} \sum_{i=1}^{N} f(\theta_i) p(\theta_i) = \frac{1}{N} \sum_{i=1}^{N} f(\theta_i^p)$;
  $\theta_i^p$ sampled with probability $p(\theta_i)$
  - IF $p(\theta_i)$ SMALL FOR SOME $\theta_i \Rightarrow$ INEFFICIENCY
  - REDEFINE $\langle f \rangle = \frac{1}{N} \sum_{I=1}^{N} f(\theta_i) \frac{p(\theta_i)}{q(\theta_i)} q(\theta_i) = \frac{1}{N} \sum_{I=1}^{N} f(\theta_i^q) \frac{p(\theta_i^q)}{q(\theta_i^q)}$;
    $\theta_i^q$ sampled with probability $q(\theta_i)$
  - OPTIMIZE CHOICE OF $q(\theta_i)$
- EQUAL WEIGHTING $\Rightarrow$ OPTIMAL CHOICE

# MONTE CARLO REPRESENTATION
# REWEIGHTING

- NEW INFORMATION FROM DATA $z_{\rm new}$ CAN BE ADDED USING BAYES' THEOREM:
  $P(f|z_{\rm new}) \propto P(z_{\rm new}|f)P(f)$

- MONTECARLO SAMPLING BOTH SIDES $\Rightarrow$ $f^i_{\rm new} = f^i_{\rm new} w_k$
  $w_k$ SAMPLING OF $P(z_{\rm new}|f)$

- WEIGHTS $w_k$ CAN BE COMPUTED FROM LIKELIHOOD OF DATA $z_{\rm new}$ GIVEN OLD MODEL
  $$w_k = \frac{(\chi^2_k)^{\frac{1}{2}(n-1)} e^{-\frac{1}{2}\chi^2_k}}{\frac{1}{N}\sum_{k=1}^{N} (\chi^2_k)^{\frac{1}{2}(n-1)} e^{-\frac{1}{2}\chi^2_k}}.$$

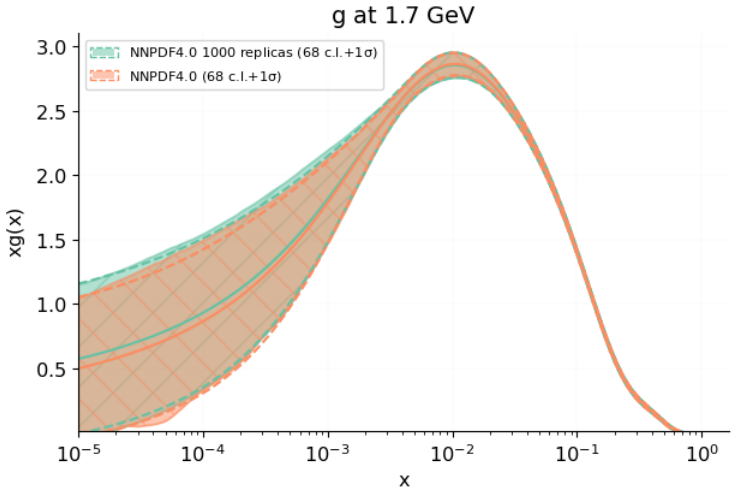
- EFFICIENCY LOSS: SHANNON ENTROPY $\Rightarrow$ EFFECTIVE # OF REPLICAS
  $N_{\rm eff} = \exp\left\{\frac{1}{N}\sum_{k=1}^{N} w_k \ln(N/w_k)\right\}$

# NEURAL NETWORKS

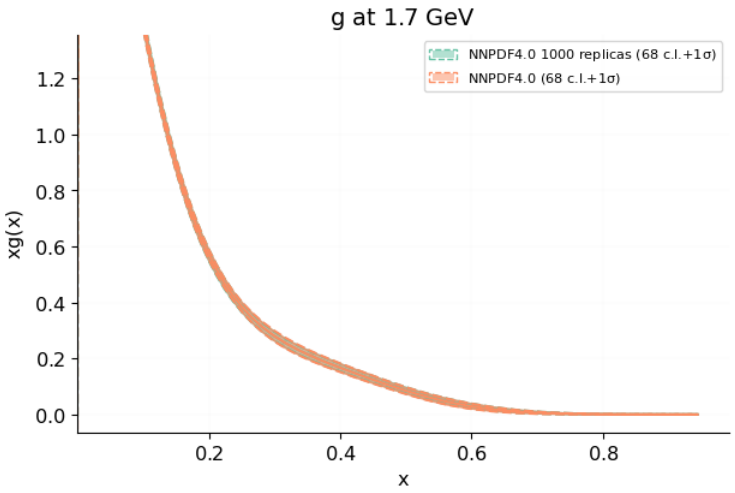
# NN MODELING

## THE PDFS

LOG SCALE




LINEAR SCALE

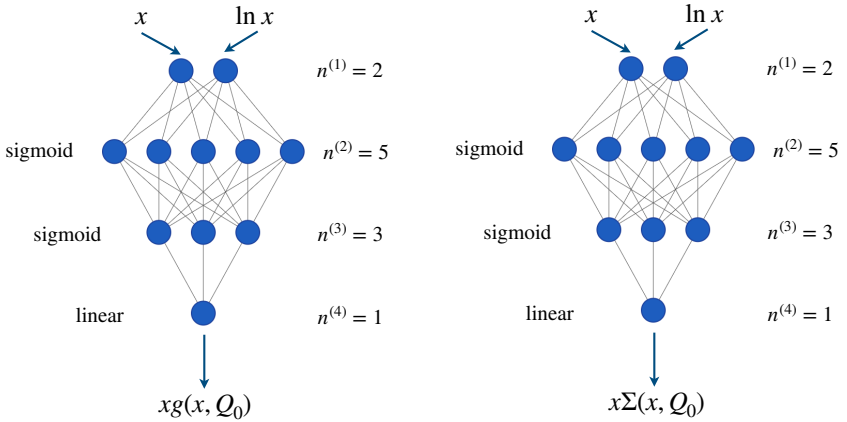



- PDF POWERLIKE IN $\ln x$; AS $x \to 1 \Rightarrow$ POWERLIKE IN $1 - x$
- ALL PDFS DETERMINED AT ONCE

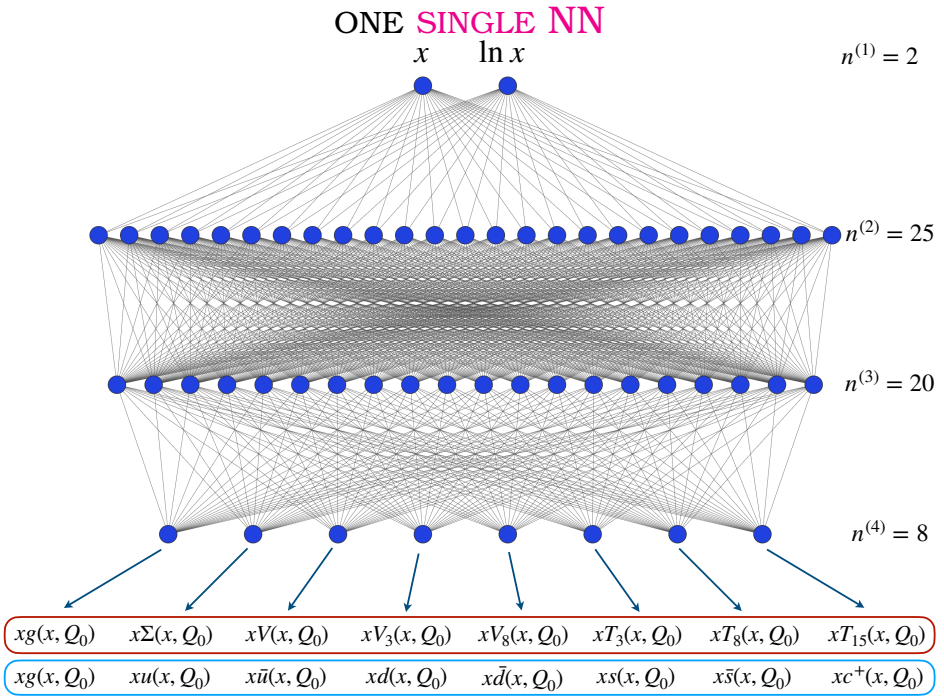
# NEURAL NETWORKS
## ARCHITECTURE

- HOW MANY INPUTS?
- HOW MANY INDEPENDENT NNs?

ONE NN PER PDF




ONE SINGLE NN



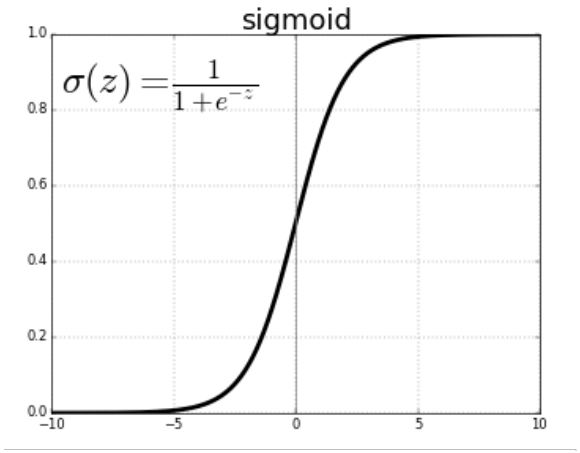
# NEURAL NETWORKS

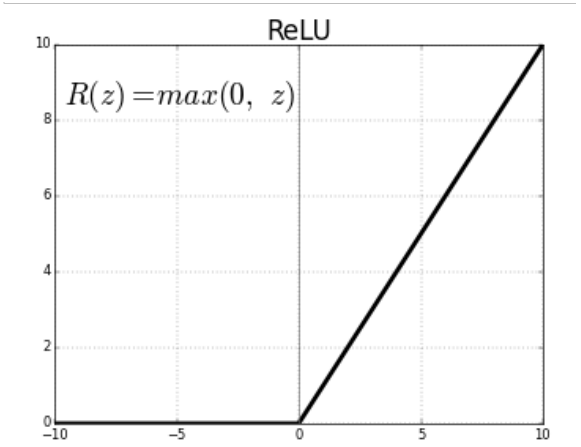
## ACTIVATION FUNCTION

- LINEAR ACTIVATION $\Rightarrow$ MULTILINEAR REGRESSION
- + NONLINEAR PROFILE $\Rightarrow$ UNIVERSAL INTERPOL.

$$F^{(i)}_{\rm out}(\vec{x}_{\rm in}) = F\left(\sum_j \omega_{ij} x^j_{\rm in} - \theta_i\right)$$

- sigmoid $F(x) = \frac{1}{1+e^{-x}}$
- arctan $F(x) = \frac{1}{2} + \frac{1}{\pi}\arctan x$
- RELU $F(x) \begin{cases} 0; & x < 0 \\ x; & x > 0 \end{cases}$



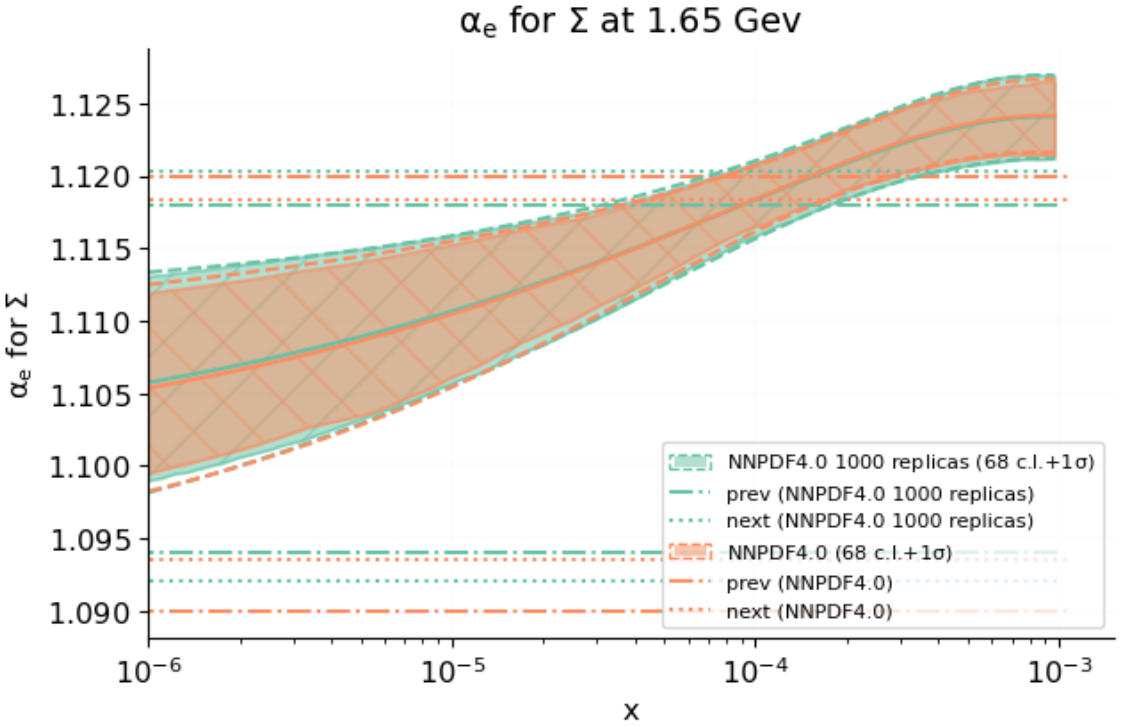
# PREPROCESSING

- ACTIVATION FUNCTION BOUNDED $\rightarrow$ NO DIVERGENCE
- LEADING BEHAVIOR DIVIDED OUT $\Rightarrow$ FASTER CONVERGENCE

$$f(x) = x^{\alpha}(1-x)^{\beta} NN(x)$$

- EXPONENTS VARIED RANDOMLY FOR EACH REPLICA,
- UNIFORM DISTRIBUTION IN FIXED RANGE
- RANGE DETERMINED SELF-CONSISTENTLY

QUARK SINGLET EFFECTIVE EXPONENT $df/d\ln(1/x)$

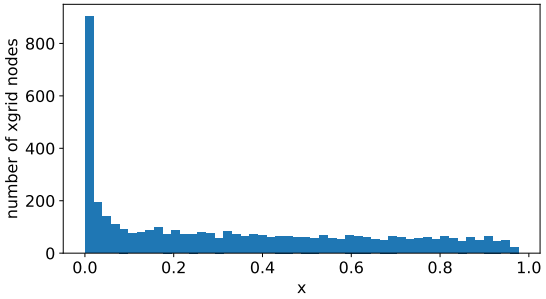



INEFFICIENT: NEED ITERATION; POTENTIAL BIAS

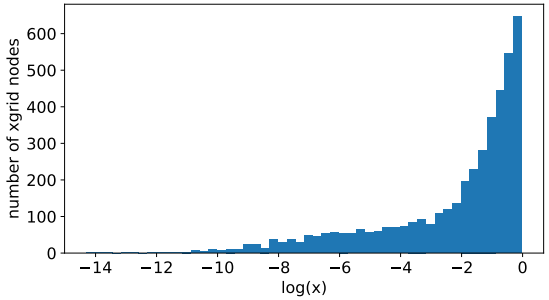
# FEATURE SCALING

INPUT NODE DISTRIBUTION

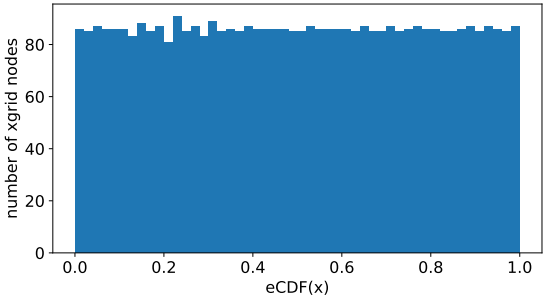
$x$ DISTN




$\ln x$ DISTN



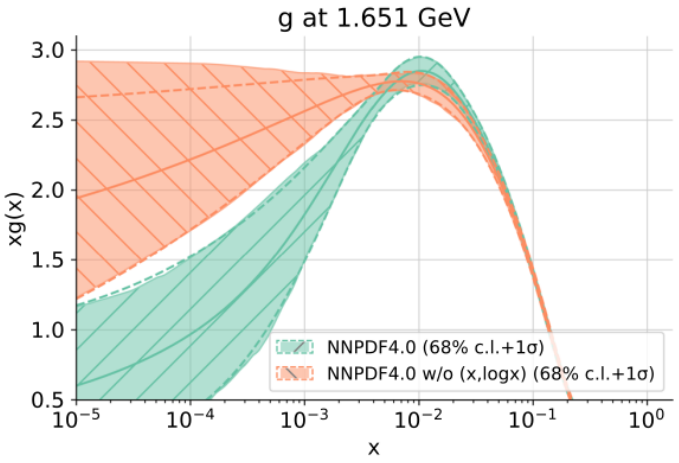

ECDF DISTN



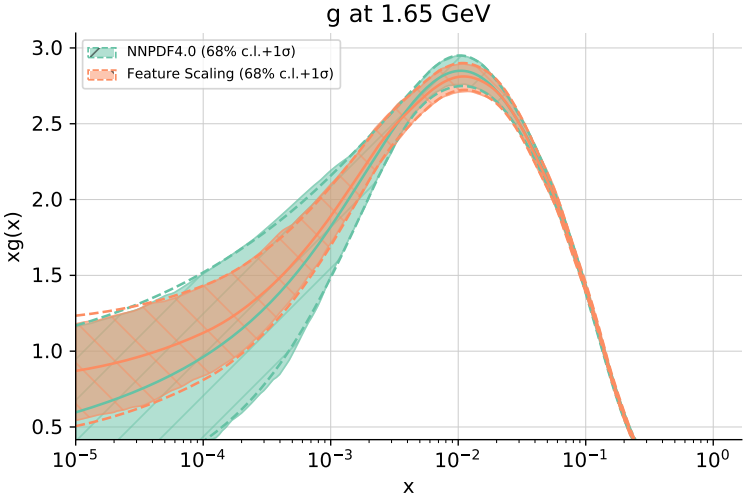

- RESCALE $\Rightarrow$ UNIFORMLY DISTRIBUTED INPUT (ECDF+INTERPOLATION)
  ECDF$(x_0)$: FRACTION OF INPUT POINTS WITH $x \leq x_0$
  - NO NEED FOR PREPROCESSING
  - NO NEED FOR DOUBLE INPUT (LINEAR/LOG)

THE GLUON

DEFAULT (TWO INPUTS) VS. ONLY LINEAR



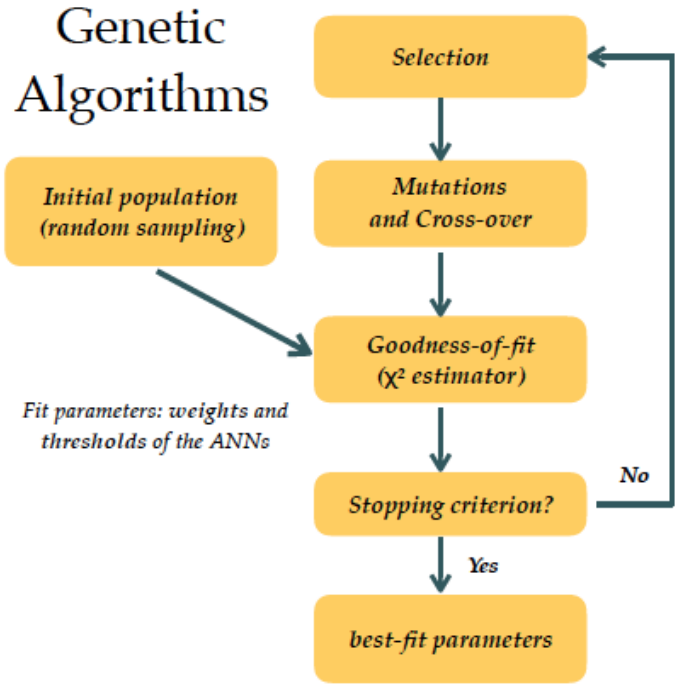

DEFAULT (PREPROCESSED) VS FEATURE SCALING

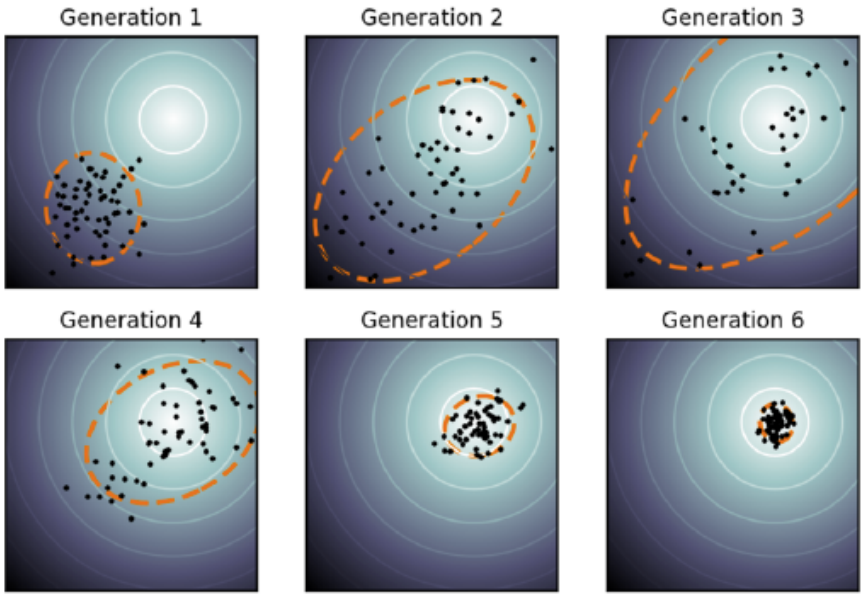


# NEURAL NETWORK TRAINING

# GENETIC ALGORITHMS

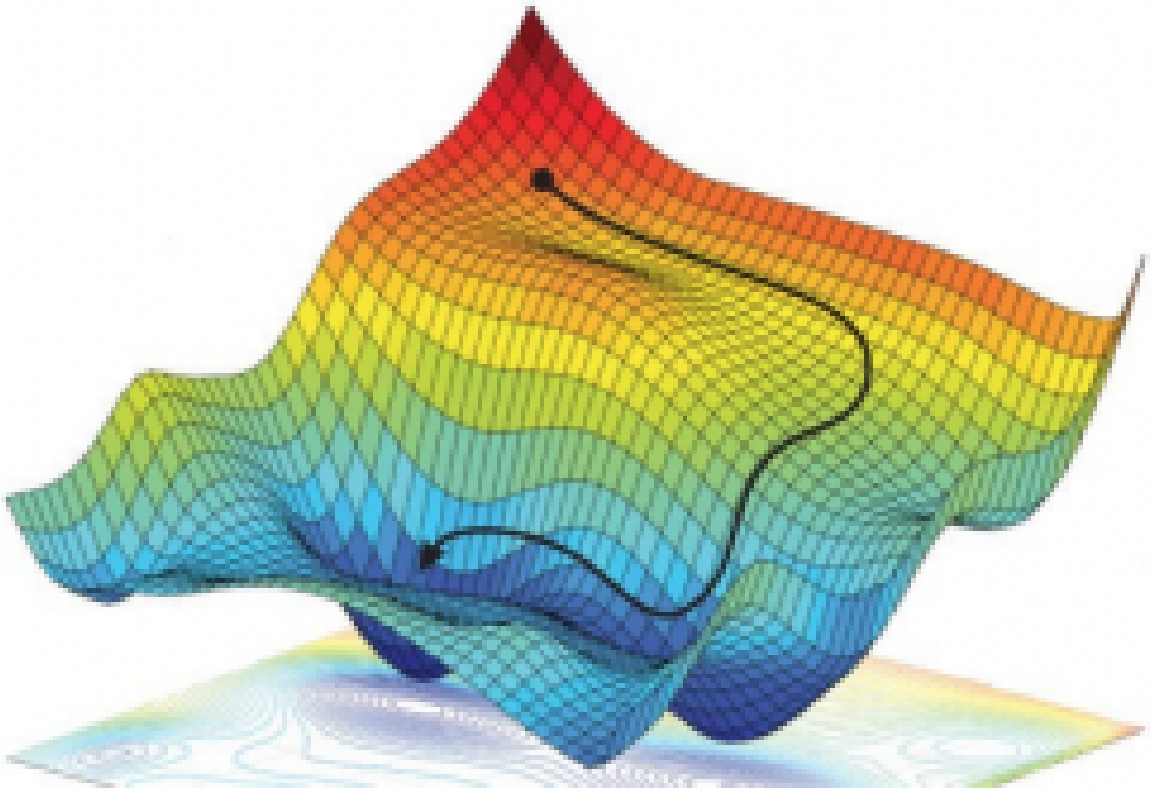
- BASIC IDEA: RANDOM MUTATION OF THE NN PARAMETER
- SELECTION OF THE FITTEST



# GRADIENT DESCENT

- BASIC IDEA: COMPUTE GRADIENT OF LOSS W.R. TO PARAMETERS
- SELECT DIRECTION OF DESCENT



# MINIMIZATION ALGORITHMS: DESIDERATA

- FAST CONVERGENCE
- DO NOT STOP ON LOCAL MINIMA
- EXPLORE SPACE OF MINIMA (DEGENERATE CASE)

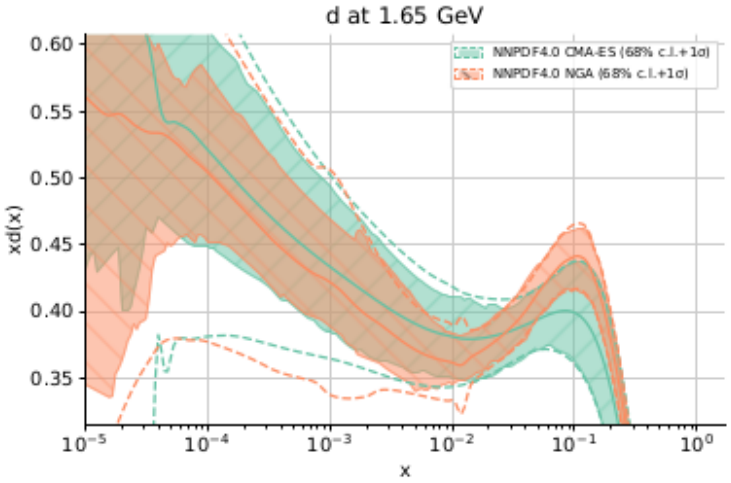
## GENETIC ALGORITHMS

- DIFFERENT EPOCHS; VARIABLE MUTATION RATE
- REWEIGHTING DIFFERENT DATA CONTRIBUTIONS TO LOSS
- NODAL MUTATION
- COVARIANCE MATRIX ADAPTATION (CMA)

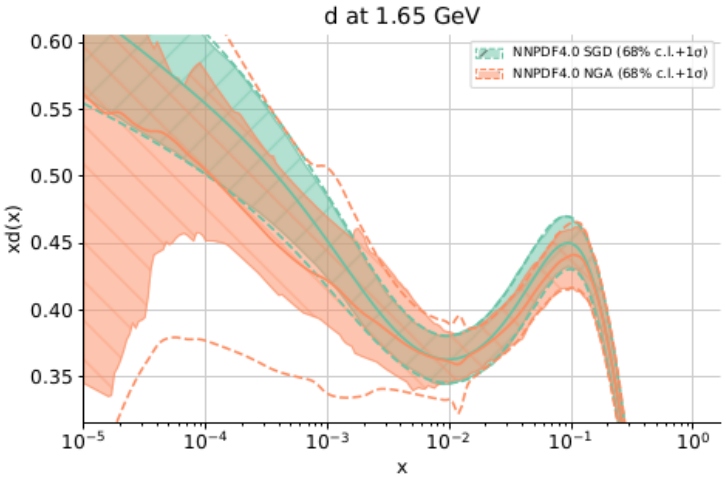
## GRADIENT DESCENT

- GLOROT NORMAL/UNIFORM INITALIZATION
- ADAPTIVE GRADIENT / ADAPTIVE MOMENT
- STOCHASTIC GD
- BATCH GD

NAIVE GA VS. CMA




GA (NAIVE) VS GD (ADADELTA)



# THEORY

# THEORY PREDICTIONS:

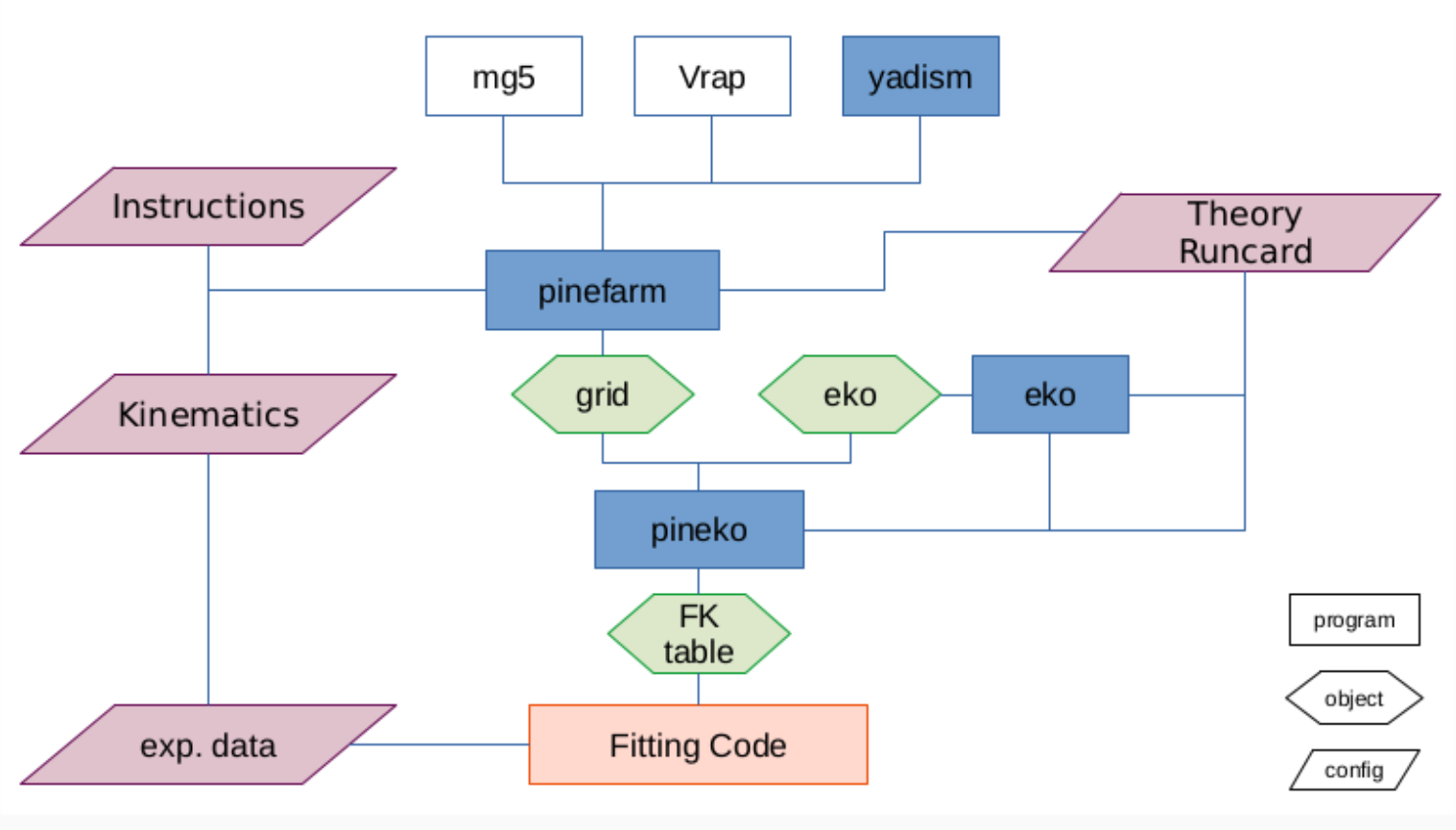
## TASKS

- SOLVE EVOLUTION EQUATIONS
- CONVOLUTE WITH HARD CROSS-SECTIONS
- ENFORCE CONSTRAINTS (SUM RULES, INTEGRABILITY, POSITIVITY)

## DESIDERATA

- FAST EVALUATION
- ANALYTIC GRADIENTS
- $\Rightarrow$ GRID-BASED APPROACH

# THE THEORY PIPELINE




- PINEFARM: GRID INTERFACE TO MATRIX ELEMENT CODES
- EKO: EVOLUTION GRIDS
- PINEKO $\Rightarrow$ FK TABLE: GRID FROM PDF TO OBSERVABLES