QCD AT THE LHC
III: PDFs

STEFANO FORTE
UNIVERSITÀ DI MILANO & INFN

UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI FISICA

Istituto Nazionale di Fisica Nucleare

XLIX Herbstschule für Hochenergiephysik

Maria Laach, 13 September 2017

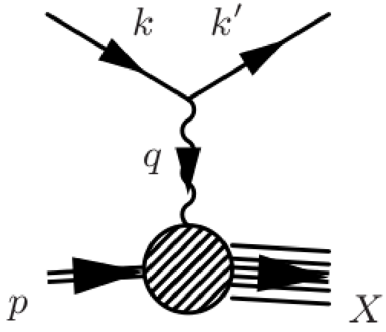
SUMMARY

LECTURE III: PDFS

- FROM DATA TO PDFS: FACTORIZATION
- FROM DATA TO PDFS: PDF DETERMINATION
- DISENTANGLING PDFS: THE ROLE OF THE DATA
- DETERMINING PDFS: HESSIAN VS. MONTE CARLO APPROACH
- PDF UNCERTAINTIES: TOLERANCE AND CROSS-VALIDATION
- CLOSURE TESTING
- THE IMPACT OF LHC DATA

FACTORIZATION REMINDER I: DEEP-INELASTIC SCATTERING

THE STRUCTURE FUNCTIONS



Lepton fractional energy loss: $y = \frac{p\cdot q}{p\cdot k}$;
gauge boson virtuality: $q^2 = -Q^2$
Bjorken x: $x = \frac{Q^2}{2p\cdot q}$
lepton-nucleon CM energy: $s = \frac{Q^2}{xy}$;
virtual boson-nucleon CM energy $W^2 = Q^2 \frac{1-x}{x}$;

$$\frac{d^2\sigma^{\lambda_p\lambda_\ell}(x,y,Q^2)}{dxdy} = \frac{G_F^2}{2\pi(1+Q^2/m_W^2)^2}\frac{Q^2}{xy}\left\{\left[-\lambda_\ell\, y\left(1-\frac{y}{2}\right)xF_3(x,Q^2) + (1-y)F_2(x,Q^2)\right.\right.$$

$$\left.\left. + y^2 x F_1(x,Q^2)\right] - 2\lambda_p\left[-\lambda_\ell\, y(2-y)x g_1(x,Q^2) - (1-y)g_4(x,Q^2) - y^2 x g_5(x,Q^2)\right]\right\}$$

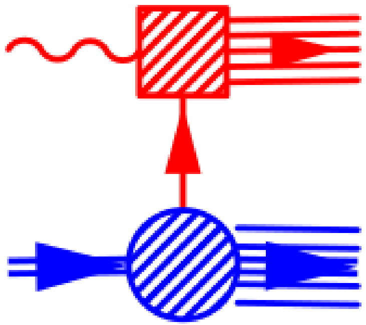
$\lambda_l \;\to$ lepton helicity
$\lambda_p \;\to$ proton helicity

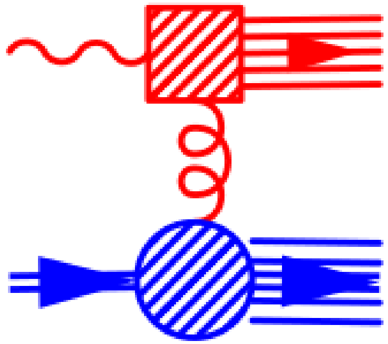
	PARITY CONS.	PARITY VIOL.
UNPOL.	$F_1,\ F_2$	F_3
POL.	g_1	$g_4,\ g_5$

FACTORIZATION REMINDER I

STRUCTURE FUNCTIONS AND PDFS

STRUCTURE FUNCTION=HARD COEFF. (PARTONIC STRUCTURE FUNCTION) $\otimes$PARTON DISTN.





$$F_2(x,Q^2) = x\sum_i \int_1^1 \frac{dy}{y} C_i\left(\alpha_s(Q^2), \frac{x}{y}\right)\left[q_i(y,Q^2) + \bar{q}_i(y,Q^2)\right] + C_g\left(\alpha_s(Q^2), \frac{x}{y}\right) g(y,Q^2)$$

q_i quark, $\bar{q}_i$ antiquark, g gluon

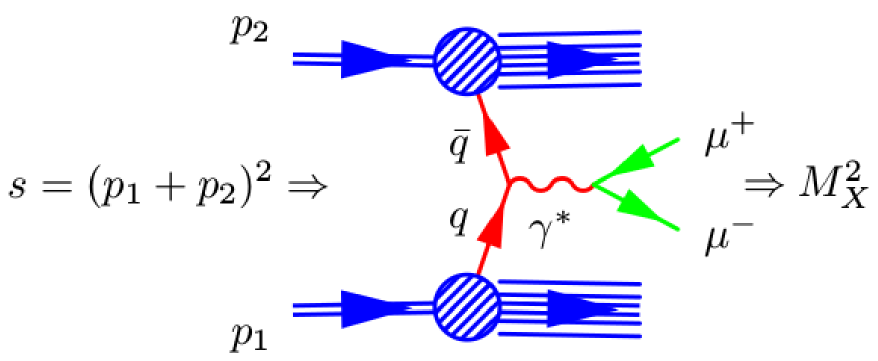
FACTORIZATION REMINDER II
HADRONIC PROCESSES
THE PARTON LUMINOSITY

$$\sigma_X(s, M_X^2) = \sum_{a,b} \int_{x_{\min}}^1 dx_1\, dx_2\, f_{a/h_1}(x_1) f_{b/h_2}(x_2) \hat\sigma_{q_a q_b \to X}\left(x_1 x_2 s, M_X^2\right)$$

$$\sigma_X(s, M^2) = \sigma_0 \sum_{a,b} \int_\tau^1 \frac{dx}{x} \mathcal{L}_{ab}\left(\frac{\tau}{x}\right) C\left(x, \alpha_s(M_H^2)\right)$$

- PARTON LUMINOSITY $\mathcal{L}_{ab}(\tau) = \int_\tau^1 \frac{dx}{x} f_{a/h_1}(x) f_{b/h_2}(\tau/x)$
- COEFFICIENT FUNCTION $\hat\sigma_{q_a q_b \to X}\left(x_1 x_2 s, M_X^2\right) = \sigma_0 C\left(\frac{M_X^2}{x_1 x_2 s}, \alpha_s(M_H^2)\right)$

EXAMPLE: THE DRELL-YAN PROCESS
AT LEADING ORDER



- Hadronic c.m. energy: $s = (p_1 + p_2)^2$
- Momentum fractions $x_{1,2} = \sqrt{\frac{\hat s}{s}} \exp \pm y$;
  Lead. Ord. $\hat s = M^2$
- Partonic c.m. energy: $\hat s = x_1 x_2 s$
- Invariant mass of final state X (dilepton, Higgs,...):
  $M_W^2 \Rightarrow$ scale of process
- Scaling variable $\tau = \frac{M_X^2}{s}$

$$\Rightarrow M^2 \frac{d\sigma}{dM^2} = \sigma_0 \mathcal{L}(\tau); \quad \sigma_0 = \frac{4}{9}\pi\alpha\frac{1}{s};$$

FACTORIZATION REMINDER III
EVOLUTION EQUATIONS...

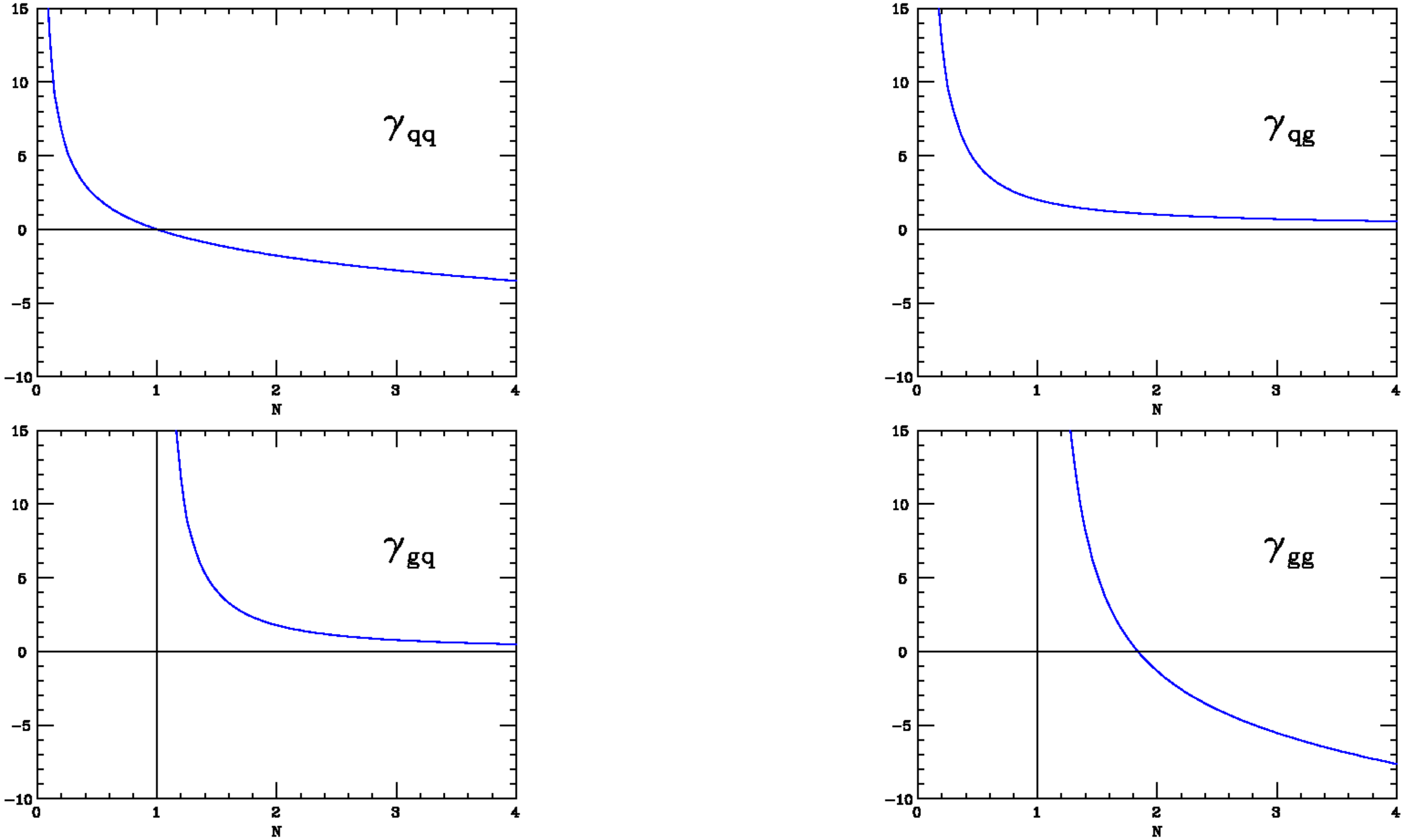
$$\frac{d}{dt} q_{NS}(N,Q^2) = \frac{\alpha_s(t)}{2\pi}\gamma_{qq}^{NS}(N,\alpha_s(t))q_{NS}(N,Q^2),$$

$$\frac{d}{dt}\begin{pmatrix}\Sigma(N,Q^2)\\ g(N,Q^2)\end{pmatrix} = \frac{\alpha_s(t)}{2\pi}\begin{pmatrix}\gamma_{qq}^{S}(N,\alpha_s(t)) & 2n_f\gamma_{qg}^{S}(N,\alpha_s(t))\\ \gamma_{gq}^{S}(N,\alpha_s(t)) & \gamma_{gg}^{S}(N,\alpha_s(t))\end{pmatrix}\begin{pmatrix}\Sigma(N,Q^2)\\ g(N,Q^2)\end{pmatrix}.$$

- LOG SCALE $t = \ln\frac{Q^2}{\Lambda^2}$:
- ANOMALOUS DIMENSIONS VS. SPLITTING FUNCTIONS
  $\gamma(N,\alpha_s(t)) \equiv \int_0^1 dx\, x^{N-1} P(x,\alpha_s(t))$
- SINGLET $\Sigma(x,Q^2) = \sum_{i=1}^{n_f}\left(q_i(x,Q^2)+\bar{q}_i(x,Q^2)\right)$ VS.
  NONSINGLET $q^{NS}(x,Q^2) = q_i(x,Q^2) - q_j(x,Q^2)$
  COMBINATIONS OF QUARK PDFS
- PERTURBATIVE EXPANSION OF ANOMALOUS DIMENSION
  $\gamma_i(N,\alpha_s(t)) = \gamma_i^{(0)}(N) + \alpha_s(t)\gamma_i^{(1)}(N) + \ldots \Rightarrow$
  LOG RESUMMATION: LO $\Leftrightarrow$ LLQ^2; NLO $\Leftrightarrow$ LLQ^2, ...

FACTORIZATION REMINDER III

...& THE LEADING ORDER ANOMALOUS DIMENSIONS



QUALITATIVE FEATURES

- AS Q^2 INCREASES, PDFS DECREASE AT LARGE x & INCREASE AT SMALL x DUE TO RADIATION
- GLUON SECTOR SINGULAR AT $N = 1 \Rightarrow$ GLUON GROWS MORE AT SMALL x
- $\gamma_{qq}(1) = 0 \Rightarrow$ NUMBER OF QUARKS CONSERVED

FACTORIZATION IV
PARTON KINEMATICS VS. HADRON KINEMATICS

$$\sigma(\tau) = \int_\tau^1 \frac{dy}{y} \sum_{ij} \mathcal{L}_{ij}(y)\hat{\sigma}\left(\frac{\tau}{z}\right); \quad \mathcal{L}_{ij}(y) \equiv \int_y^1 \frac{dx_1}{x_1} q_i(y) q_j\left(\frac{y}{x_1}\right)$$

- q_i QUARKS AND GLUONS
- PARTONIC CHANNEL ij DEPENDS ON PHYSICAL PROCESS (e.g. $W^+ \Rightarrow u\bar{d}$ fusion)
- WHICH PARTON MOMENTUM FRACTIONS CONTRIBUTE TO A GIVEN HADRONIC PROCESS ?

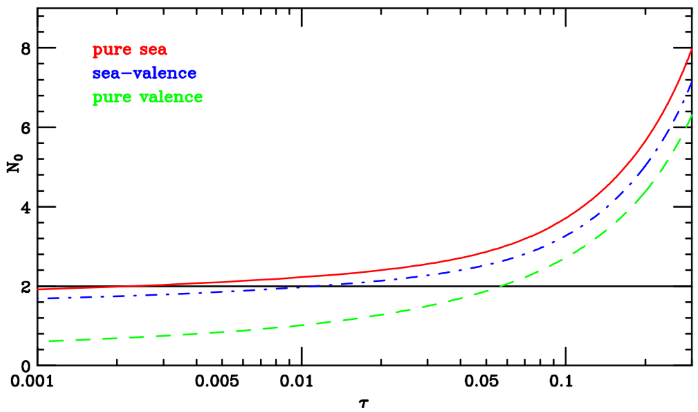
INVERSION OF MELLIN TRANSFORMS

$$f_n = \int_x^1 x^{n-1} f(x) \Leftrightarrow F(x) = \int_{-i\infty}^{+i\infty} x^{-n} f_n$$

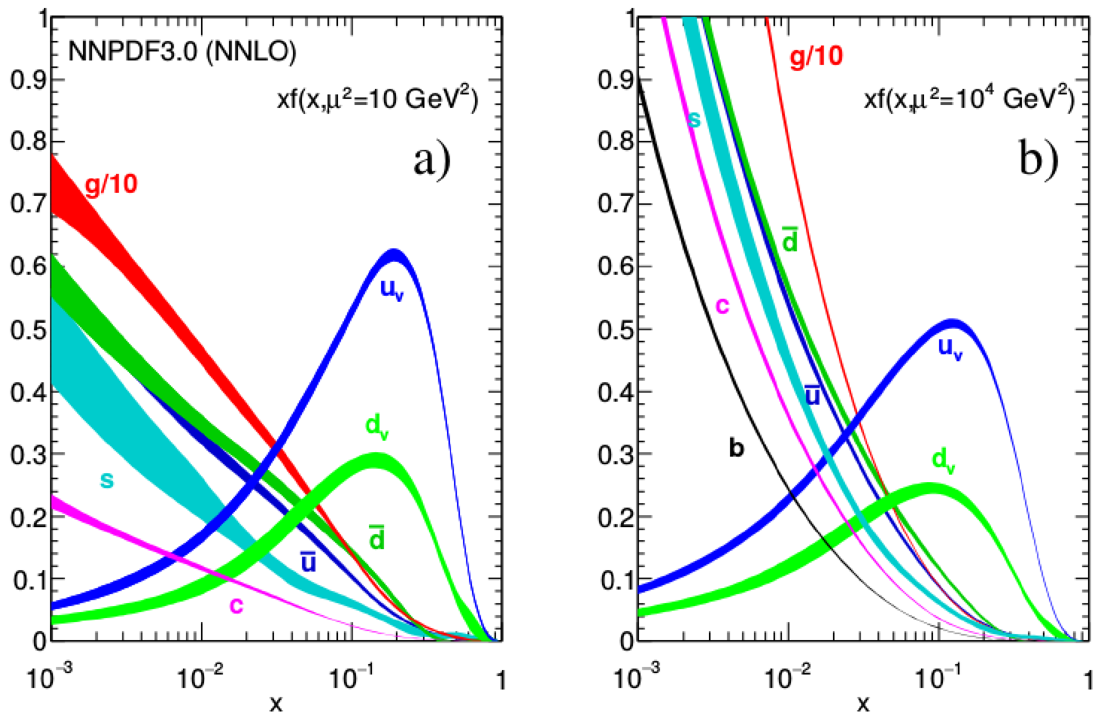
integrate to the right of convergence abscissa

- MELLIN INVERSION DOMINATED BY SADDLE POINT
- POSITION OF SADDLE CONTROLLED BY LUMINOSITY
  DEPENDENCE ON x OF $\mathcal{L}$ POWERLIKE, OF $\hat{\sigma}$ LOGARITHMIC
- PDF PEAKED AT SMALL x ("SEA" $\bar{q}$ VS. "VALENCE" $q-\bar{q}$) $\Rightarrow$ LUMI PEAKS AT SMALL N

SADDLE VS $\tau = Q^2/s$



THE PDFs

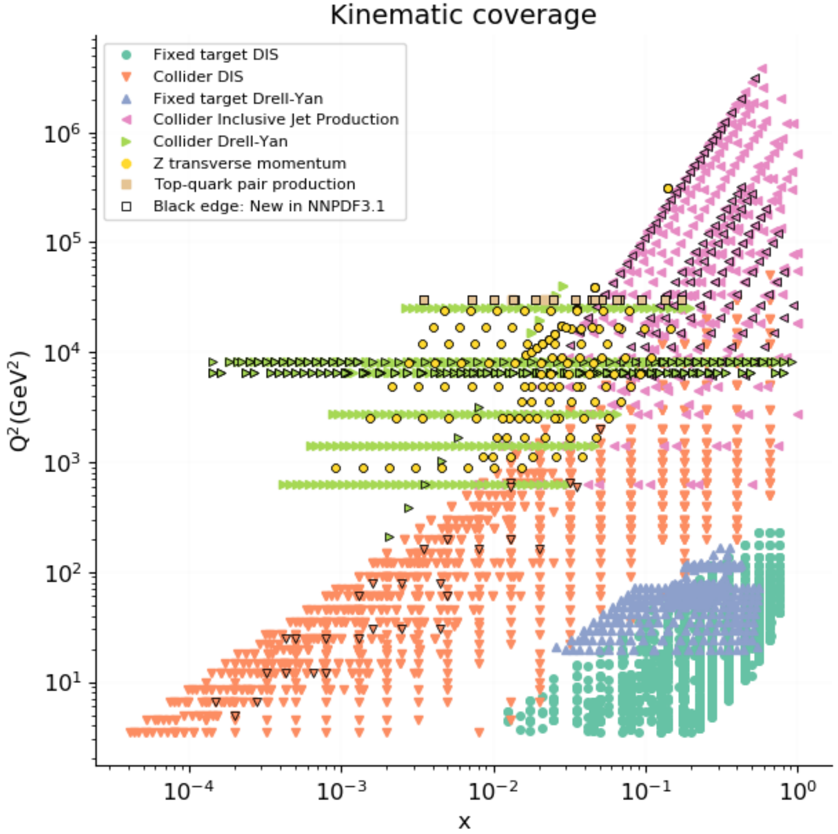


(PDG 2016)

- THE MOMENTUM PROBABILITY DENSITY $xf_i(x)$ IS SHOWN AT TWO DIFFERENT SCALES (LEFT $\Rightarrow$ LOW SCALE; RIGHT $\Rightarrow$ HIGH SCALE)
- PDFs VS x AT ONE SCALE $Q_0^2 \Rightarrow$ DETERMINED FOR ALL SCALES BY EVOLUTION EQUATIONS
- AS $x \geq 1$ KINEMATIC CONSTRAINT $f_i(x) = 0$
- "VALENCE" UP AND DOWN: PEAKED AT $x \sim 0.3$; EXPECT $f_x(x) \underset{x\to 1}{\sim} (1-x)_i^\beta$
- "SEA" ANTIQUARK AND GLUON GROW AT SMALL x
- "SINGLET" AND GLUON MIX $\Rightarrow$ ALL PDFs LOOK THE SAME AS $x \to 0$

PDF DETERMINATION

DATA → PARTON DISTRIBUTIONS



ISSUES AND TASKS:

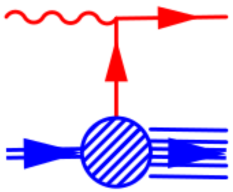
- FROM PHYSICAL OBSERVABLES TO PDFs: SOLVE EVOLUTION EQUATIONS, CONVOLUTE WITH PARTON-LEVEL CROSS-SECTIONS
- DISENTANGLING PDFs: CHOOSE A BASIS OF PDFs ($2N_f$ QUARKS + 1 GLUON) & A SET OF SUITABLE PHYSICAL PROCESSES TO DETERMINE THEM ALL
- PROBABILITY IN THE SPACE OF FUNCTIONS: CHOOSE A STATISTICAL APPROACH (HESSIAN, MONTE CARLO, . . .)
- UNCERTAINTY ON FUNCTIONS: CHOOSE A FUNCTIONAL FORM

DISENTANGLING PDFs

- DEEP-INELASTIC SCATTERING DATA ON PROTON ABUNDANT AND PRECISE
- CC F_1 AND F_3 IN PRINCIPLE PROVIDE FOUR COMBINATIONS, AND NC F_1 TWO MORE $\Rightarrow$ ALL LIGHT FLAVORS
- HERA DATA ONLY DETERMINE FOUR COMBINATIONS OF PDFs:
FIXED COMBINATION OF F_1 F_3, SO NC AND $\pm$CC WITH $e^{\pm}$, PLUS SEPARATE NC γ AND Z FROM SCALE DEPENDENCE
- $W^{\pm}$ AND Z PRODUCTION (INCLUDING DOUBLE DIFFERENTIAL: MASS AND RAPIDITY) PROVIDE A LARGE AMOUNT OF INFORMATION
- WHEN PRODUCING ELECTROWEAK FINAL STATES, THE GLUON CAN ONLY BE ACCESSED FROM SCALE DEPENDENCE OR HIGHER ORDERS
...EXCEPT IN HIGGS PRODUCTION!
- JET PRODUCTION GIVES A DIRECT HANDLE ON THE GLUON

FLAVOR SEPARATION FOM DIS & DY:
LEADING ORDER PARTON CONTENT

DEEP-INELASTIC SCATTERING



NC $F_1^{\gamma} = \sum_i e_i^2\,(q_i + \bar{q}_i)$

NC $F_1^{Z,\,\mathrm{int.}} = \sum_i B_i\,(q_i + \bar{q}_i)$

NC $F_3^{Z,\,\mathrm{int.}} = \sum_i D_i\,(q_i + \bar{q}_i)$

CC $F_1^{W^+} = \bar{u} + d + s + \bar{c}$

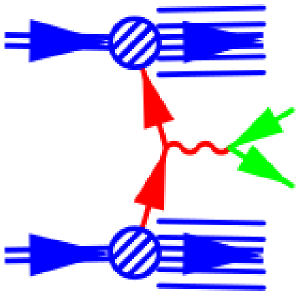
CC $-F_3^{W^+}/2 = \bar{u} - d - s + \bar{c}$

ℓ	e	V	A
u,c,t	$+2/3$	$(+1/2 - 4/3\sin^2\theta_W)$	$+1/2$
d,s,b	-1/3	$(-1/2 + 2/3\sin^2\theta_W)$	-1/2
ν	0	$+1/2$	$+1/2$
e,μ, τ	-1	$(-1/2 + 2\sin^2\theta_W)$	-1/2

$B_q(Q^2) = -2e_q V_\ell V_q P_Z + (V_\ell^2 + A_\ell^2)(V_q^2 + A_q^2)P_Z^2$; $D_q(Q^2) = -2e_q A_\ell A_q P_Z + 4V_\ell A_\ell V_q A_q P_Z^2$; $P_Z = Q^2/(Q^2 + M_Z^2)$

$$W^+ \to W^- \Rightarrow u \leftrightarrow d,\ c \leftrightarrow s;\ p \to n \Rightarrow u \leftrightarrow d$$

DRELL-YAN



$$L^{ij}\,(x_1, x_2) \equiv q_i(x_1, M^2)\bar{q}_j(x_2, M^2)$$

γ $$\frac{d\sigma}{dM^2 dy}(M^2, y) = \frac{4\pi\alpha^2}{9M^2 s}\sum_i e_i^2 L^{ii}(x_1, x_2)$$

W $$\frac{d\sigma}{dy} = \frac{\pi G_F M_V^2\sqrt{2}}{3s}\sum_{i,j}|V_{ij}^{\mathrm{CKM}}|L^{ij}(x_1, x_2)$$

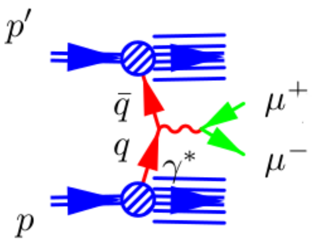
Z $$\frac{d\sigma}{dy} = \frac{\pi G_F M_V^2\sqrt{2}}{3s}\sum_i \left(V_i^2 + A_i^2\right) L^{ij}(x_1, x_2)$$

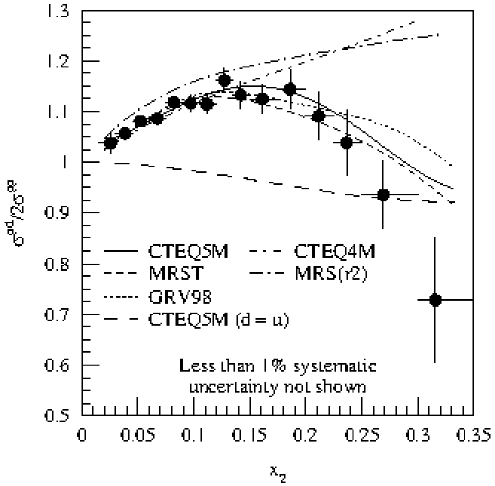
$V_{ij}^{\mathrm{CKM}} \to$ CKM MATRIX ($i = u,\ c\,t,\ j = d,\ s\,b$), $V_{ij}^{\mathrm{CKM}} = 1 + O(\lambda)$; $\lambda = \sin\theta_C \approx 0.22$

IMPACT OF TEVATRON DRELL-YAN DATA: QUARKS AND ANTIQUARKS AT A $p\bar{p}$ COLLIDER (TEVATRON)

BY CHARGE CONJUGATION $\bar{q}_{\bar{P}} = q_p$

DRELL-YAN p/d ASYMMETRY

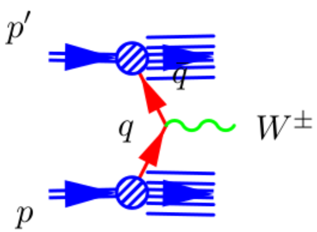


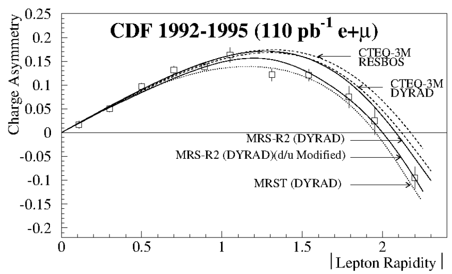


$$\frac{\sigma^{pn}}{\sigma^{pp}} \sim \left. \frac{\frac{4}{9} u^p \bar{d}^p + \frac{1}{9} d^p \bar{u}^p}{\frac{4}{9} u^p \bar{u}^p + \frac{1}{9} d^p \bar{d}^p} \right|_{\text{large } x} \approx \frac{\bar{d}}{\bar{u}}$$

E866 (2001)

$W^{\pm}$ ASYMMETRY





$$\frac{\sigma^{p\bar{p}}_{W+}}{\sigma^{p\bar{p}}_{W-}} = \frac{u^p(x_1)d^p(x_2)+\bar{d}^p(x_1)\bar{u}^p(x_2)}{d^p(x_1)u^p(x_2)+\bar{u}^p(x_1)\bar{d}^p(x_2)} \sim \frac{u^p d^p}{d^p u^p}$$

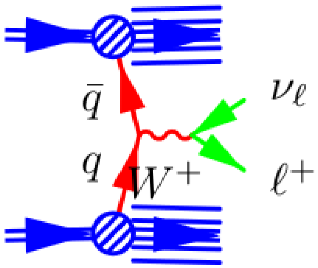
if x_1, x_2 in valence region, neglecting HQ & Cabibbo suppr.

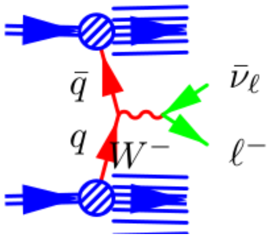
CDF (1998)

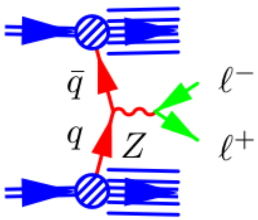
IMPACT OF LHC DRELL-YAN DATA:

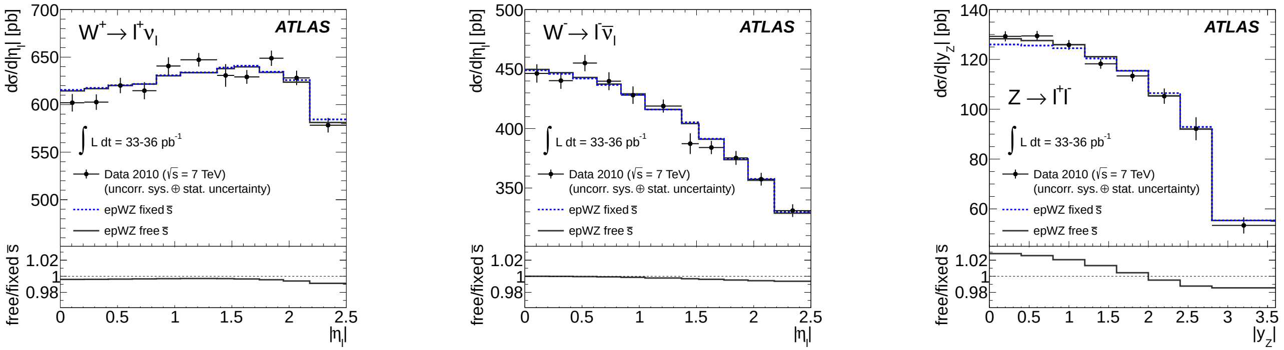
$W^{\pm}$ AND Z PRODUCTION

W AND Z CROSS SECTIONS





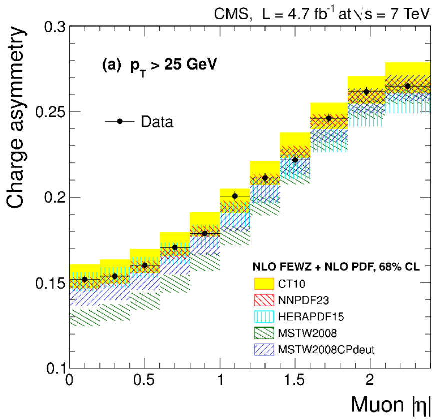




ATLAS (2012)

$$\sigma^{p\bar{p}}_{W+} = u\bar{d} + c\bar{s}; \quad \sigma^{p\bar{p}}_{Z} = u\bar{u} + d\bar{d} + s\bar{s}:$$ STRANGENESS DETERMINED BY COMPARISON

W MUON ASYMMETRY



CMS (2013)

$$\frac{\sigma^{p\bar{p}}_{W+}}{\sigma^{p\bar{p}}_{W-}} = \frac{u(x_1)\bar{d}(x_2)+\bar{d}(x_1)u(x_2)}{d(x_1)\bar{u}(x_2)+\bar{u}(x_1)d(x_2)}$$

"VALENCE" $x \Rightarrow$ NEGLECT STRANGENESS

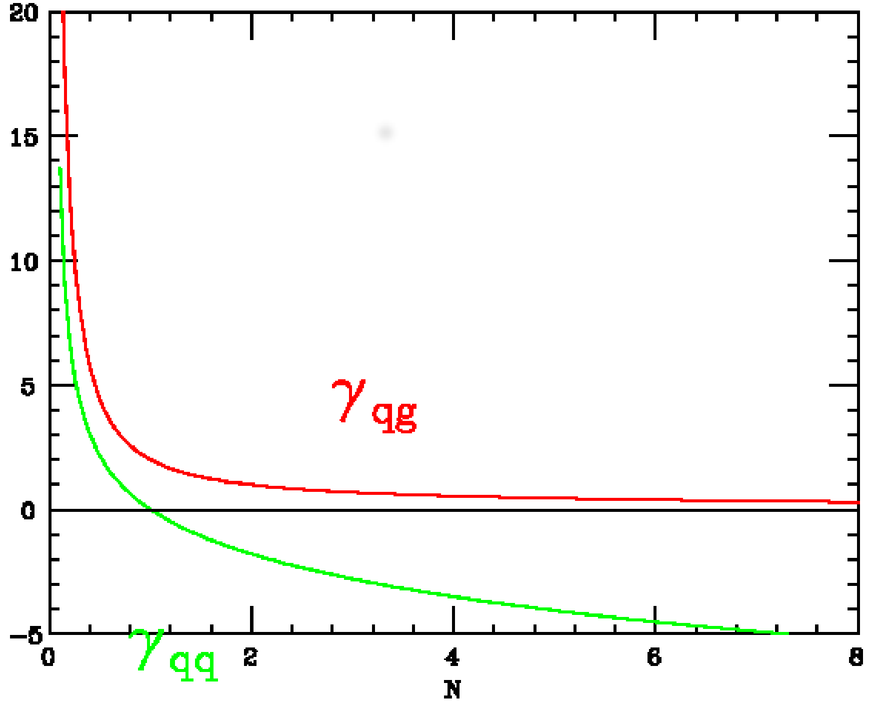
$\Rightarrow$ DETERMINE $\bar{u} - \bar{d}$

THE GLUON FROM DIS

SCALE DEPENDENCE OF FLAVOR SINGLET STRUCTURE FUNCTIONS

$$\frac{d}{dt}F_2^s(N,Q^2)=\frac{\alpha_s(Q^2)}{2\pi}\left[\gamma_{qq}(N)F_2^s+2\,n_f\gamma_{qg}(N)g(N,Q^2)\right]+O(\alpha_s^2)$$

ANOMALOUS DIMENSIONS



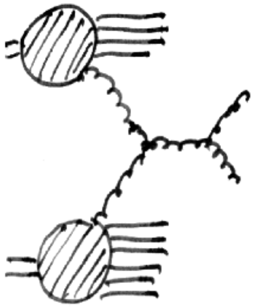
LARGE x GLUON DIFFICULT TO DETERMINE FROM DEEP-INELASTIC SCATTERING

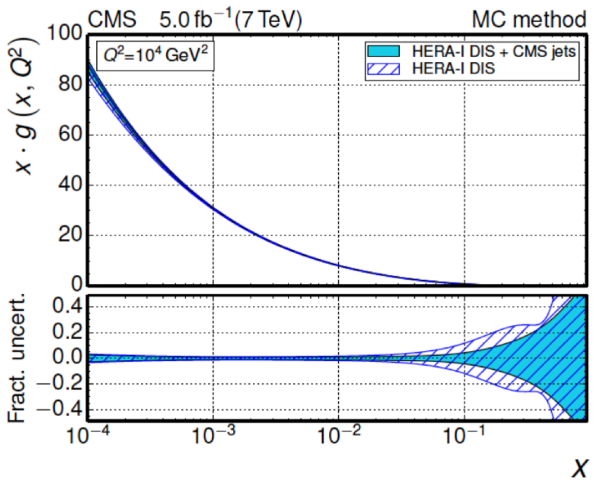
THE GLUON IN HADRONIC COLLISIONS

THE GLUON ONLY INTERACTS THROUGH QCD

JETS

GLUON



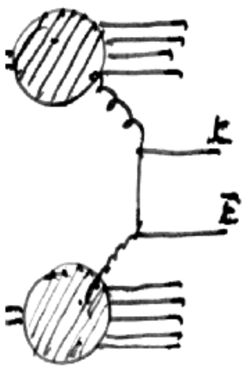


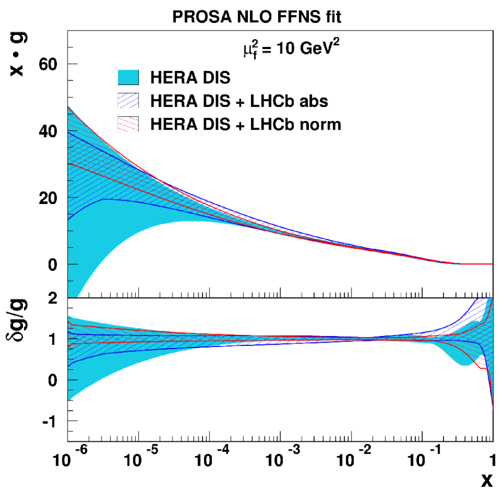
CMS (2014)

- ONE-JET INCLUSIVE USED TO CONSTRAIN THE LARGE x GLUON SINCE TEVATRON
- WIDE KINEMATIC REGION AT LHC

TOP

GLUON





prosa (LHCb data) 2015)

- WIDE RAPIDITY RANGE: CAN ACCESS WIDE x REGION

DETERMINING PDFS
THE HESSIAN APPROACH

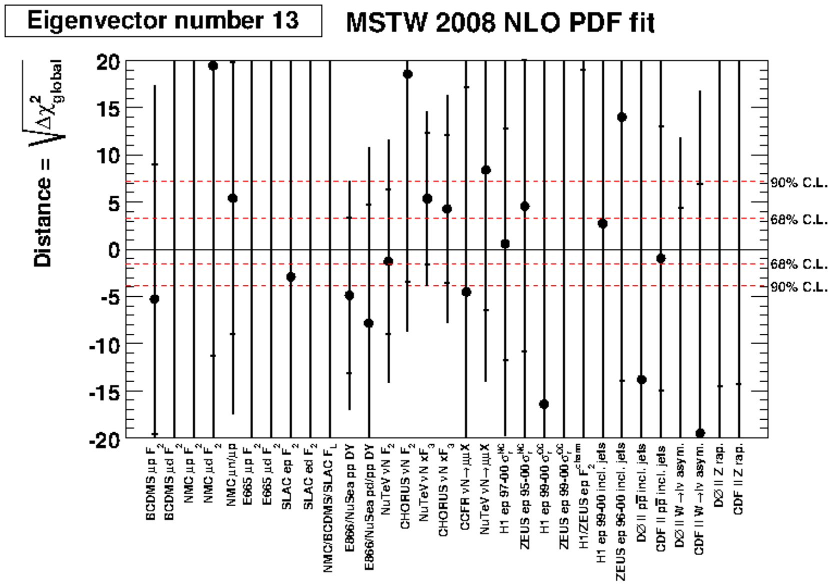
- CHOOSE A FIXED FUNCTIONAL FORM
  - SINCE 1973, PHYSICALLY MOTIVATED ANSATZ $f_i(x, Q_0^2) = x^\alpha (1-x)^\beta g_i(x)$;
    $g_i(x)$ POLYNOMIAL IN x OR $\sqrt{x}$
  - MMHT 2015:
    - BASIS FUNCTIONS g; $u_v = u - \bar{u}$; $d_v = d - \bar{d}$; $S = 2(\bar{u} + \bar{d}) + s + \bar{s}$; $s_+ = s + \bar{s}$; $\Delta = \bar{d} - \bar{u}$; $s_- = s - \bar{s}$.
    - FOR ALL BUT Δ s_-, $g \Rightarrow x f_i(x, Q_0^2) = A x^\alpha (1-x)^\beta \left(1 + \sum_{i=1}^4 a_i T_i(y(x))\right)$;
      T_i CHEBYSHEV POLYNOMIALS, $y = 1 - 2\sqrt{x} \leftrightarrow$ MUST MAP $x = [0,1]$ INTO $y = [-1,1]$;
      $T_i(-1) = T_i(1) = 1$
    - GLUON $x g(x, Q_0^2) = A x^\alpha (1-x)^\beta \left(1 + \sum_{i=1}^2 a_i T_i(y(x))\right) + A' x T \alpha' (1-x)^{\beta'}$
    - SEA ASYMMETRY $x\Delta(x, Q_0^2) = A x^\alpha (1-x)^\beta (1 + \gamma x + \epsilon x^2)$
    - STRANGENESS ASYMMETRY $x\Delta(x, Q_0^2) = A x^\alpha (1-x)^\beta (1 - x/x_0)$
    - 41 PARAMETERS, 4 FIXED BY SUM RULES
    - 12 PARMS FIXED AT BEST FIT, REMAINING 25 USED FOR (HESSIAN) COVARIANCE MATRIX
- EVOLVE TO DESIRED SCALE & COMPUTE PHYSICAL OBSERVABLES
- DETERMINE BEST-FIT VALUES OF PARAMETERS
- DETERMINE ERROR BY PROPAGATION OF ERROR ON PARMS. $\Delta\chi^2 = 1$('HESSIAN METHOD');
  PARM. SCANS ALSO POSSIBLE ('LAGR. MULTIPLIER METHOD')

PROLEMS: TOLERANCE

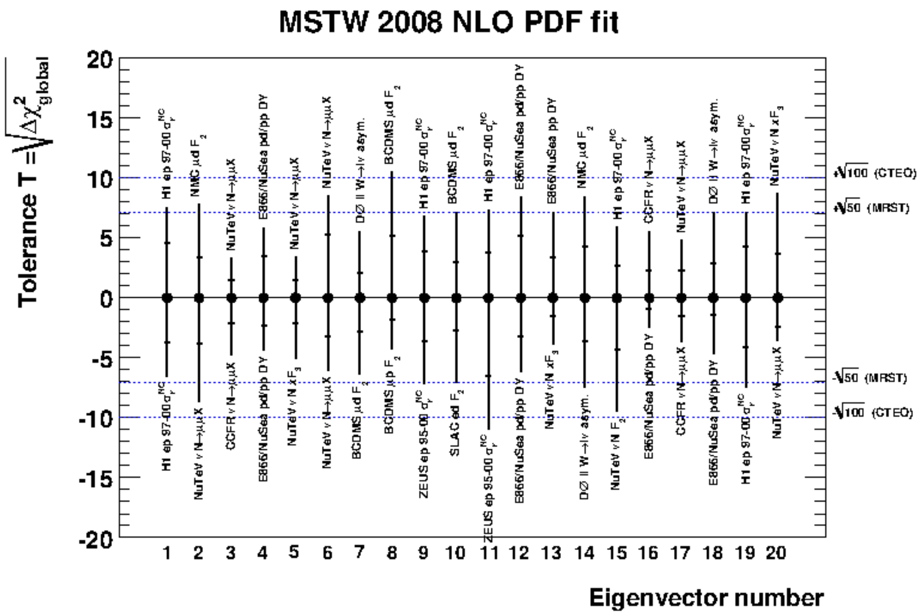
- IN GLOBAL HESSIAN FITS, UNCERTAINTITES OBTAINED BY $\Delta\chi^2 = 1$ UNREALISTICALLY SMALL

- UNCERTAINTIES TUNED TO DISTRIBUTION OF DEVIATIONS FROM BEST-FITS FOR INDIVIDUAL EXPERIMENTS

MSTW TOLERANCE PLOT FOR 13TH EIGENVEC.



GLOBAL MSTW TOLERANCE



- (MSTW/MMHT) FOR EACH EIGENVECTOR IN PARAMETER SPACE DETERMINE CONFIDENCE LIMIT FOR THE DISTRIBUTION OF BEST-FITS OF EACH EXPERIMENT

- RESCALE $\Delta\chi^2 = T$ INTERVAL SUCH THAT CORRECT CONFIDENCE INTERVALS ARE REPRODUCED

DOES THE NEED FOR TOLERANCE REFLECT PARAMETRIZATION BIAS?

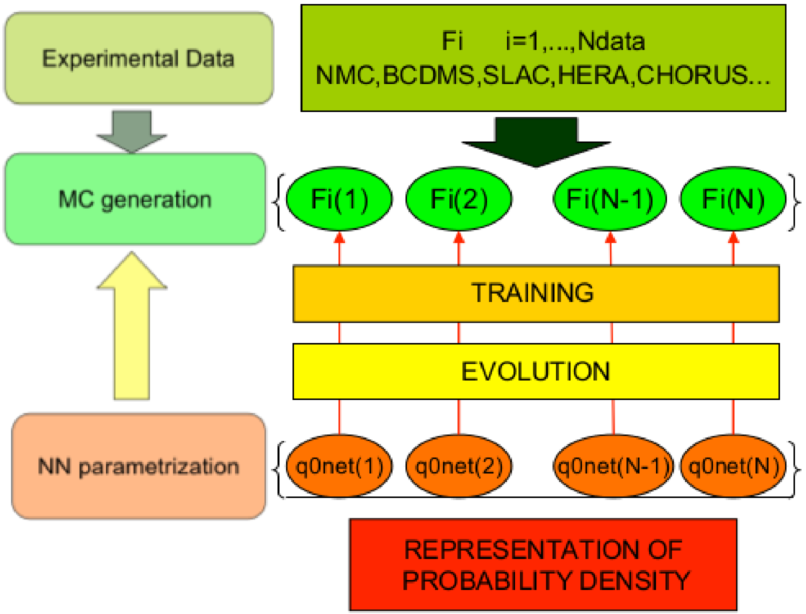
DETERMINING PDFS II

THE MONTE CARLO APPROACH

BASIC IDEA: MONTE CARLO SAMPLING OF THE PROBABILITY MEASURE IN THE (FUNCTION) SPACE OF PDFS

- GENERATE A SET OF MONTE CARLO REPLICAS $\sigma^{(k)}$ OF THE ORIGINAL DATASET $\sigma^{(\mathrm{data})}$
  $\Rightarrow$ REPRESENTATION OF $\mathcal{P}[\sigma]$ AT DISCRETE SET OF POINTS IN DATA SPACE
- FIT A PDF REPLICA TO A DATA REPLICA
  $\Rightarrow$ EACH PDF REPLICA $f_i^{(k)}$ IS A BEST-FIT PDF SET FOR GIVEN DATA REPLICA
- THE SET OF NEURAL NETS IS A REPRESENTATION OF THE PROBABILITY DENSITY:

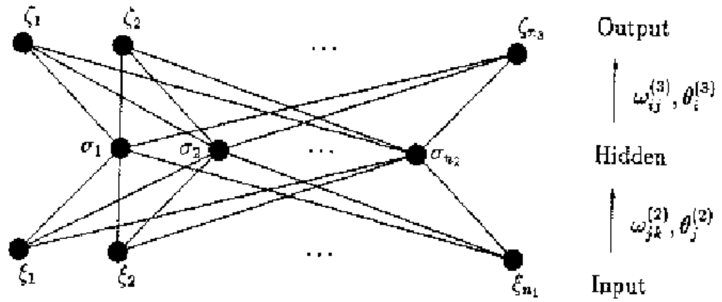
$$\langle f_i \rangle = \frac{1}{N_{rep}} \sum_{k=1}^{N_{rep}} f_i^{(k)}$$



NEURAL NETWORKS & THE MC APPROACH

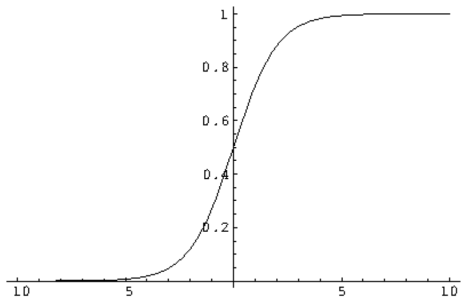
- EACH PDF REPLICA FITTED TO A DATA REPLICA
  $\Rightarrow$ NEED BEST-FIT, COVARIANCE MATRIX IN PARAMETER SPACE NOT NEEDED
- CAN USE VERY LARGE PARAMETRIZATION

NEURAL NETWORKS



MULTILAYER FEED-FORWARD NETWORKS

- Each neuron receives input from neurons in preceding layer and feeds output to neurons in subsequent layer
- Activation determined by weights and thresholds
  $$\xi_i = g\left(\sum_j \omega_{ij}\xi_j - \theta_i\right)$$
- Sigmoid activation function
  $g(x) = \frac{1}{1+e^{-\beta x}}$



EXAMPLE: A 1-2-1 NN

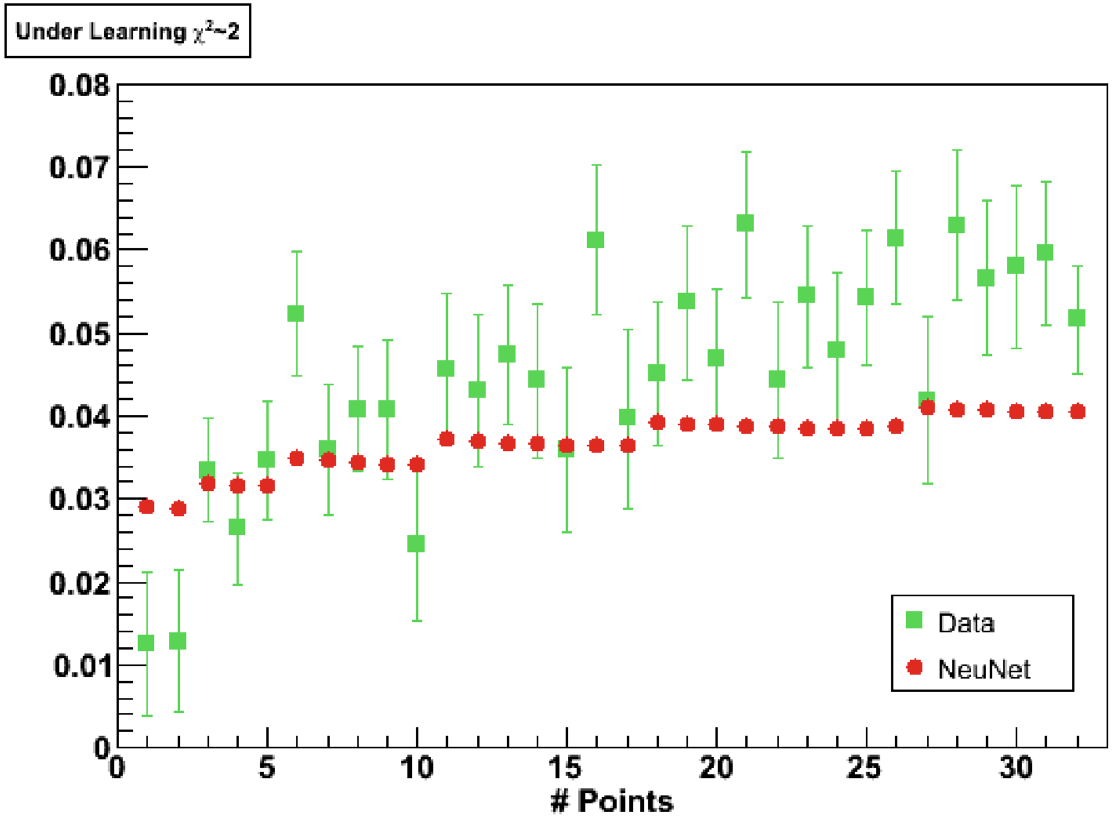
$$f(x) = \frac{1}{1+e^{\theta_1^{(3)} - \frac{\omega_{11}^{(2)}}{1+e^{\theta_1^{(2)} - x\omega_{11}^{(1)}}} - \frac{\omega_{12}^{(2)}}{1+e^{\theta_2^{(2)} - x\omega_{21}^{(1)}}}}}$$

THANKS TO NONLINEAR BEHAVIOUR,
ANY FUNCTION CAN BE REPRESENTED
BY A SUFFICIENTLY BIG NEURAL NETWORK

NEURAL LEARNING

- ONE CAN CHOOSE A HIGHLY REDUNDANT PARAMETRIZATION
  EXAMPLE: NNPDF: $2-5-3-1$ NN FOR EACH PDF: $37 \times 7 = 259$ PARAMETERS
- MINIMIZATION ("LEARNING") CAN BE PERFORMED USING GENETIC ALGORITHMS
- COMPLEXITY INCREASES AS THE FITTING PROCEEDS
- $\Rightarrow$ THE BEST FIT IS NOT THE ABSOLUTE MINIMUM:
  MUST LOOK FOR OPTIMAL LEARNING POINT

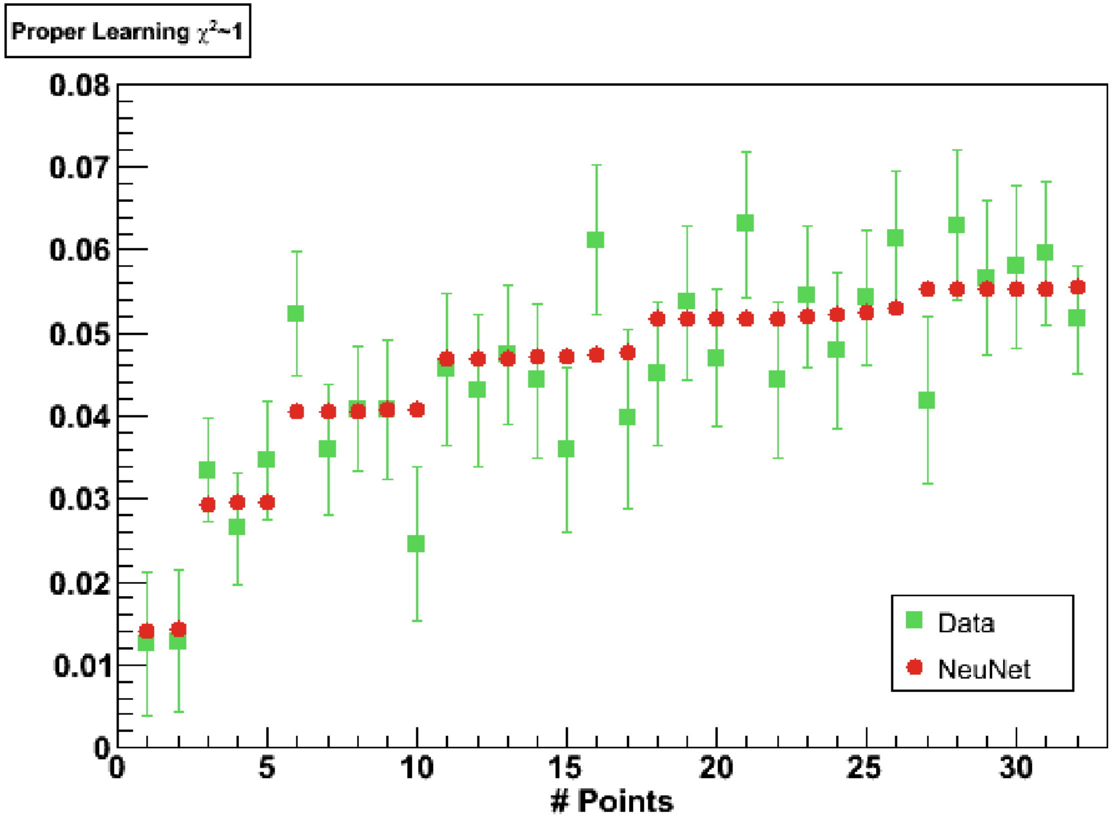
UNDERLEARNING



NEURAL LEARNING

- ONE CAN CHOOSE A HIGHLY REDUNDANT PARAMETRIZATION
  EXAMPLE: NNPDF: $2-5-3-1$ NN FOR EACH PDF: $37 \times 7 = 259$ PARAMETERS
- MINIMIZATION ("LEARNING") CAN BE PERFORMED USING GENETIC ALGORITHMS
- COMPLEXITY INCREASES AS THE FITTING PROCEEDS
- $\Rightarrow$ THE BEST FIT IS NOT THE ABSOLUTE MINIMUM:
  MUST LOOK FOR OPTIMAL LEARNING POINT

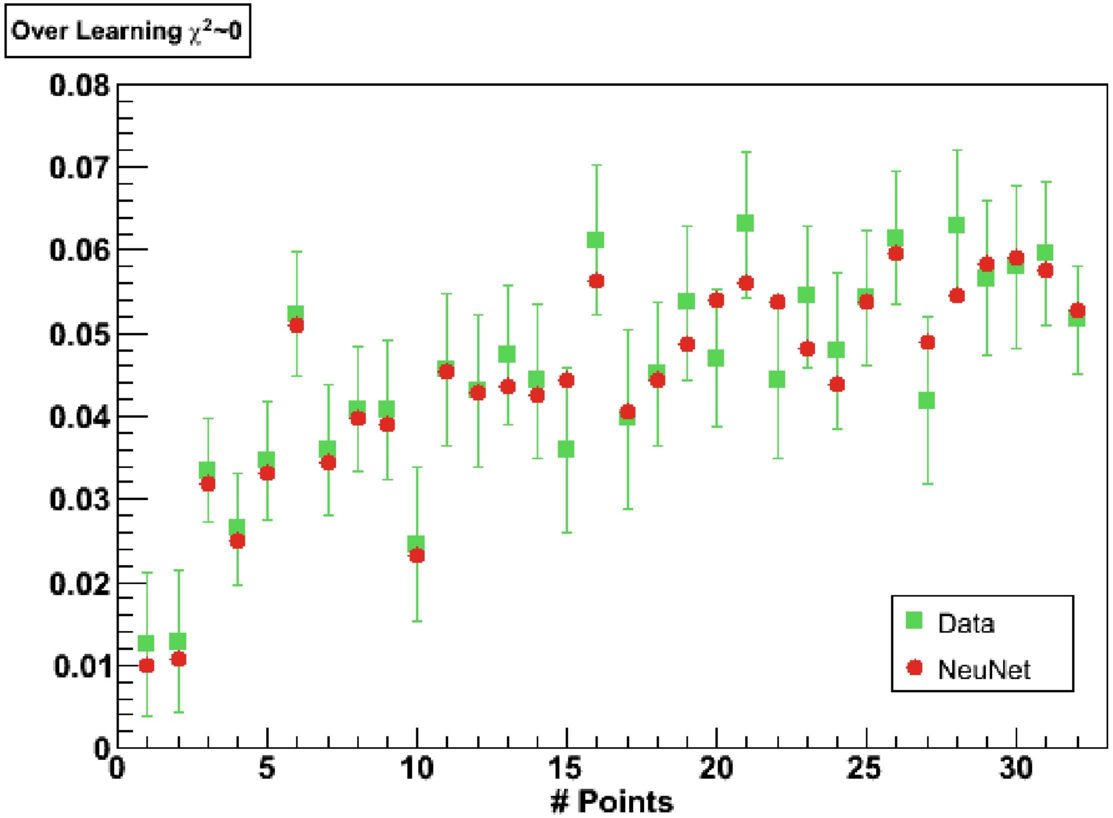
PROPER LEARNING



NEURAL LEARNING

- ONE CAN CHOOSE A HIGHLY REDUNDANT PARAMETRIZATION
  EXAMPLE: NNPDF: $2-5-3-1$ NN FOR EACH PDF: $37 \times 7 = 259$ PARAMETERS
- MINIMIZATION ("LEARNING") CAN BE PERFORMED USING GENETIC ALGORITHMS
- COMPLEXITY INCREASES AS THE FITTING PROCEEDS
- $\Rightarrow$ THE BEST FIT IS NOT THE ABSOLUTE MINIMUM:
  MUST LOOK FOR OPTIMAL LEARNING POINT

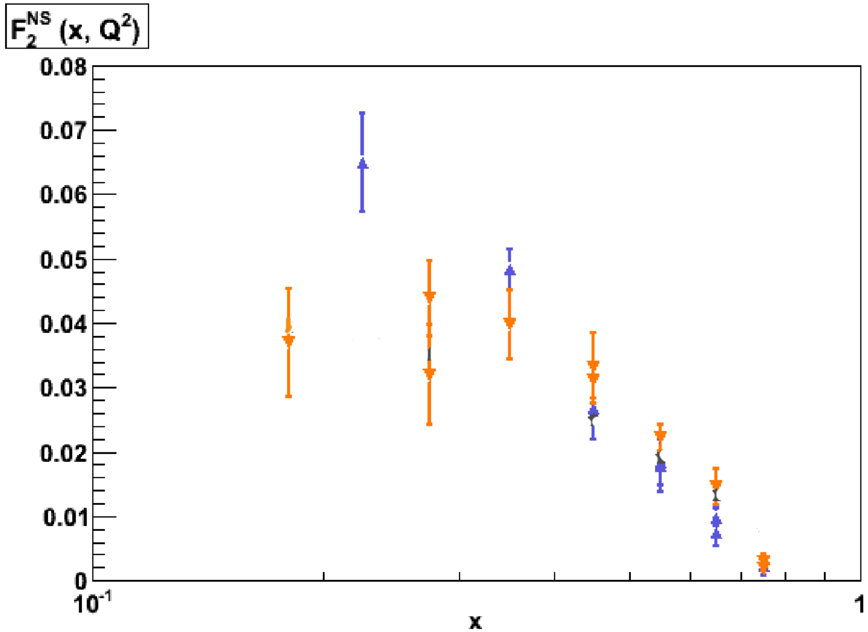
OVERLEARNING



OPTIMAL FIT: CROSS-VALIDATION

GENETIC MINIMIZATION:
AT EACH GENERATION, χ^2 EITHER UNCHANGED OR DECREASING

- 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

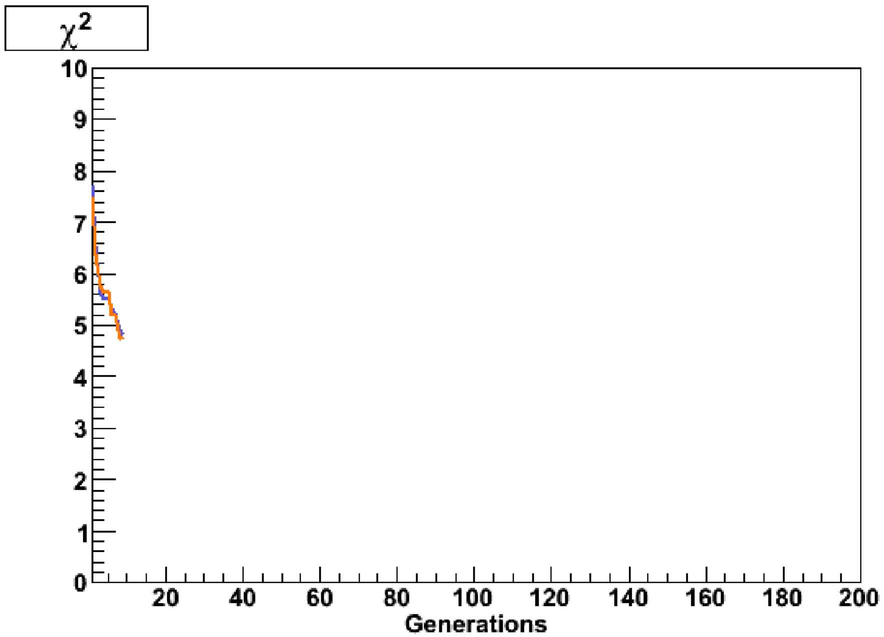


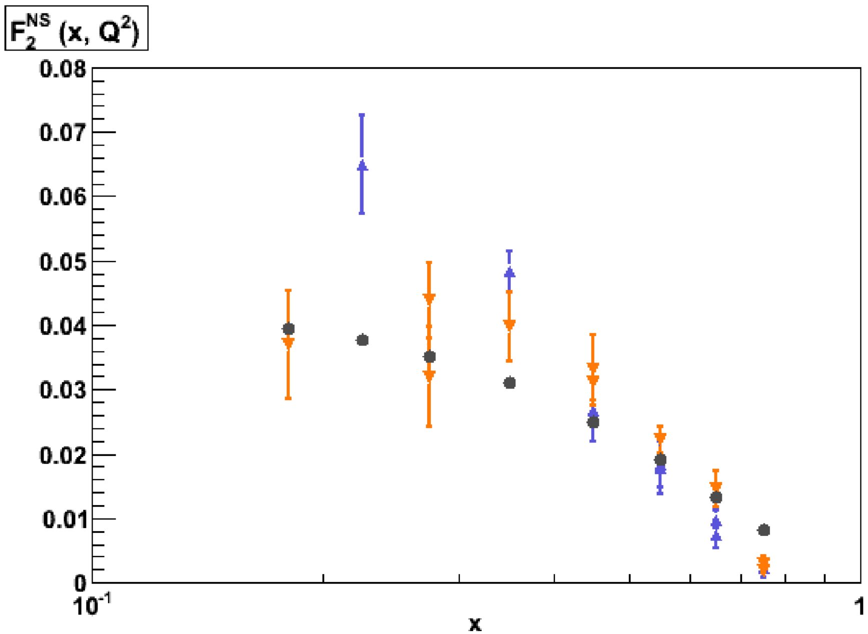
OPTIMAL FIT: CROSS-VALIDATION

GENETIC MINIMIZATION:
AT EACH GENERATION, χ^2 EITHER UNCHANGED OR DECREASING

- 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!



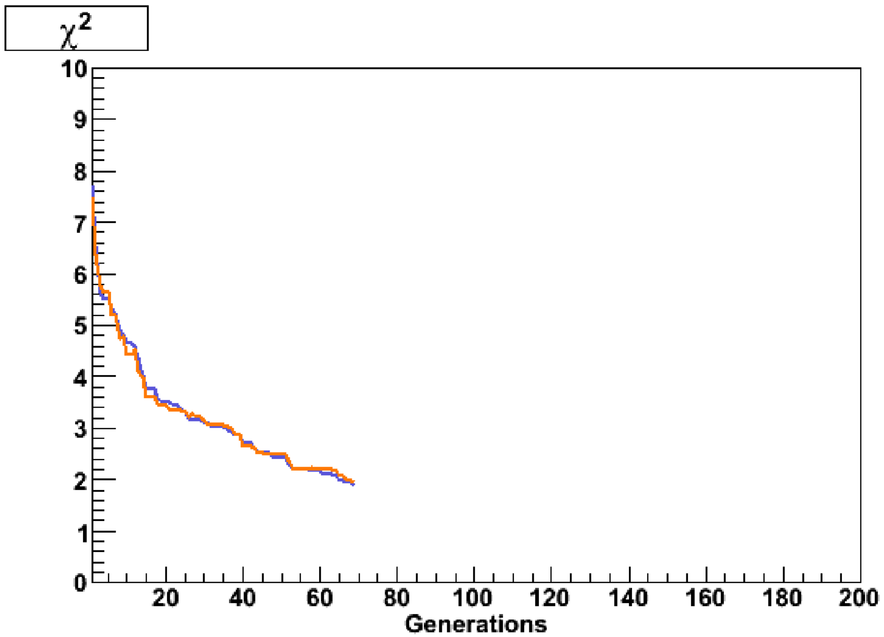


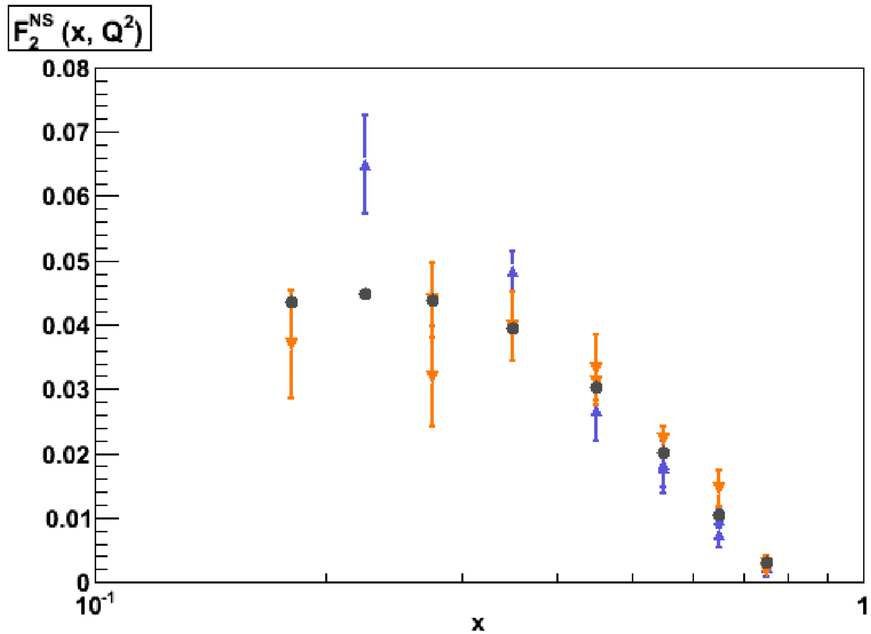
OPTIMAL FIT: CROSS-VALIDATION

GENETIC MINIMIZATION:
AT EACH GENERATION, χ^2 EITHER UNCHANGED OR DECREASING

- 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!



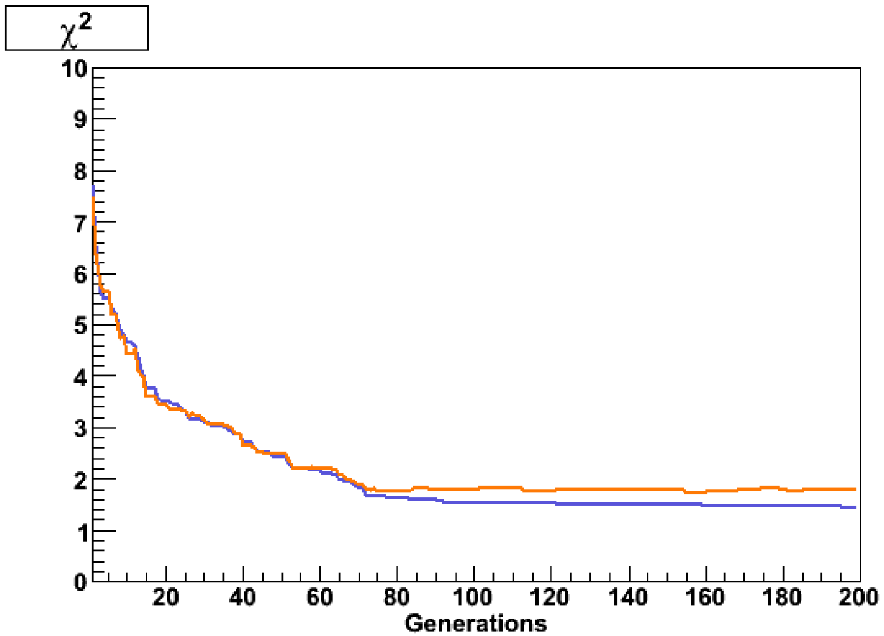


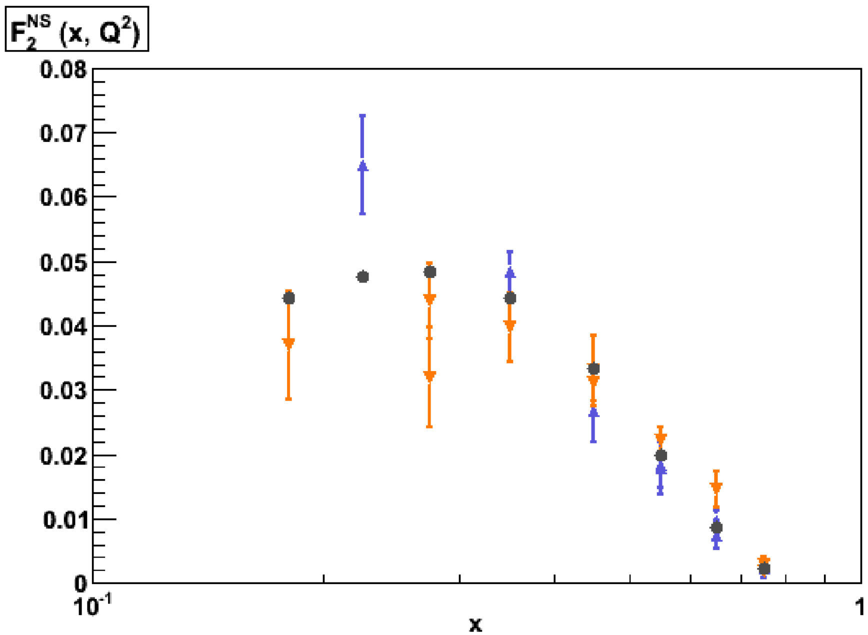
OPTIMAL FIT: CROSS-VALIDATION

MINIMIZE BY GENETIC ALGORITHM:
AT EACH GENERATION, χ^2 EITHER UNCHANGED OR DECREASING

- 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!





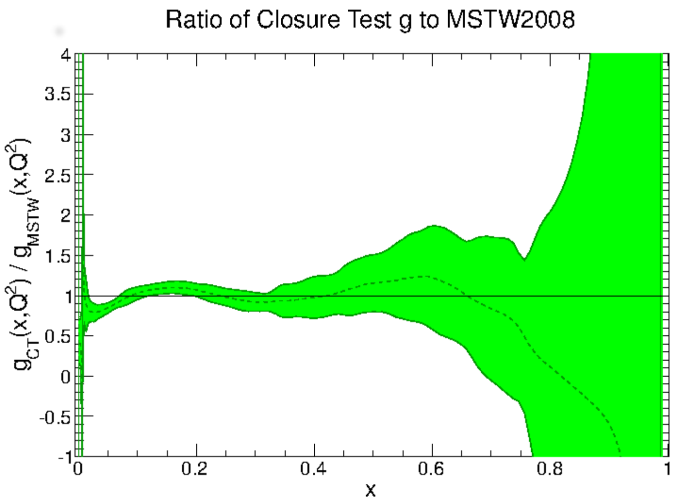
TESTING THE PDF DETERMINATION:

CLOSURE TESTS

- ASSUME PDFS KNOWN: GENERATE FAKE EXPERIMENTAL DATA
- CAN DECIDE DATA UNCERTAINTY (ZERO, OR AS IN REAL DATA, OR . . .)
- FIT PDFS TO FAKE DATA
- CHECK WHETHER FIT REPRODUCES UNDERLYING "TRUTH":
  - CHECK WHETHER TRUE VALUE GAUSSIANLY DISTRIBUTED ABOUT FIT
  - CHECK WHETHER UNCERTAINTIES FAITHFUL
  - TRACE DIFFERENT SOURCES OF UNCERTAINTY

TESTING THE PDF DETERMINATION

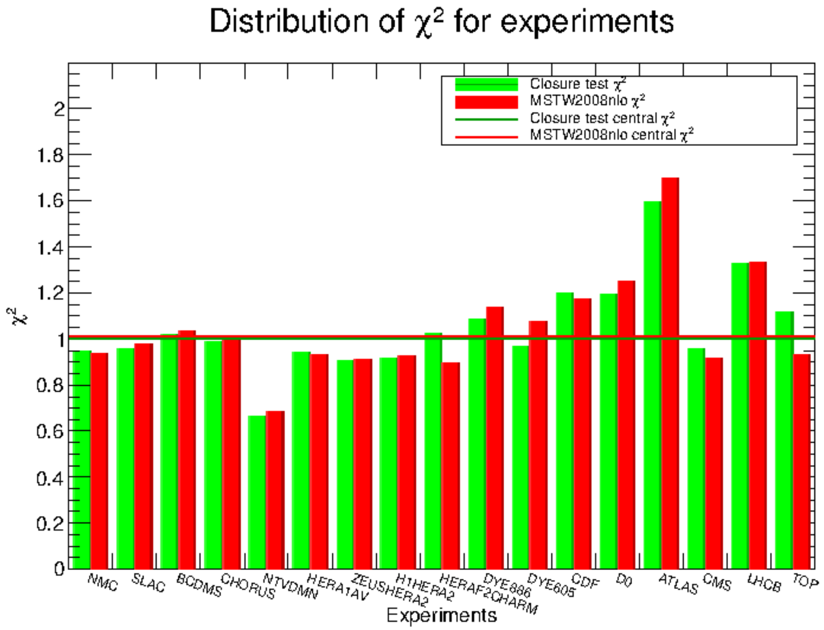
THE GLUON: FITTED/"TRUE"



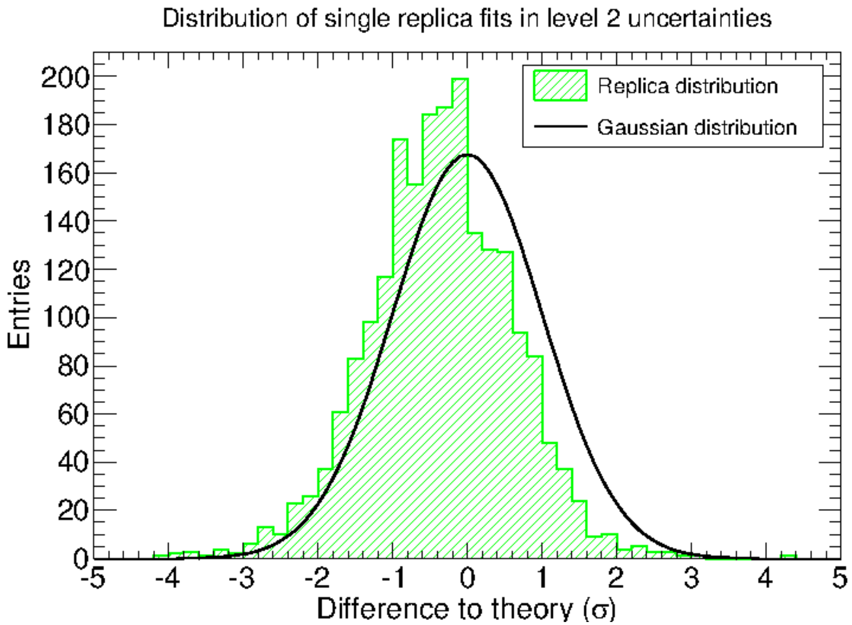
RESULTS

- CENTRAL VALUES:
  COMPARE FITTED VS. "TRUE" χ^2
  BOTH FOR INDIVIDUAL EXPERIMENTS
  & TOTAL DATASET
  FOR TOTAL $\Delta\chi^2 = 0.001 \pm 0.003$
- UNCERTAINTIES: DISTRIBUTION OF DEVIATIONS BETWEEN FITTED AND "TRUE" PDFS
  SAMPLED AT 20 POINTS BETWEEN 10^{-5} AND 1
  FIND 0.699% FOR ONE-SIGMA,
  0.948% FOR TWO-SIGMA C.L.

LEVEL-2 FITTED χ^2 VS "TRUE"

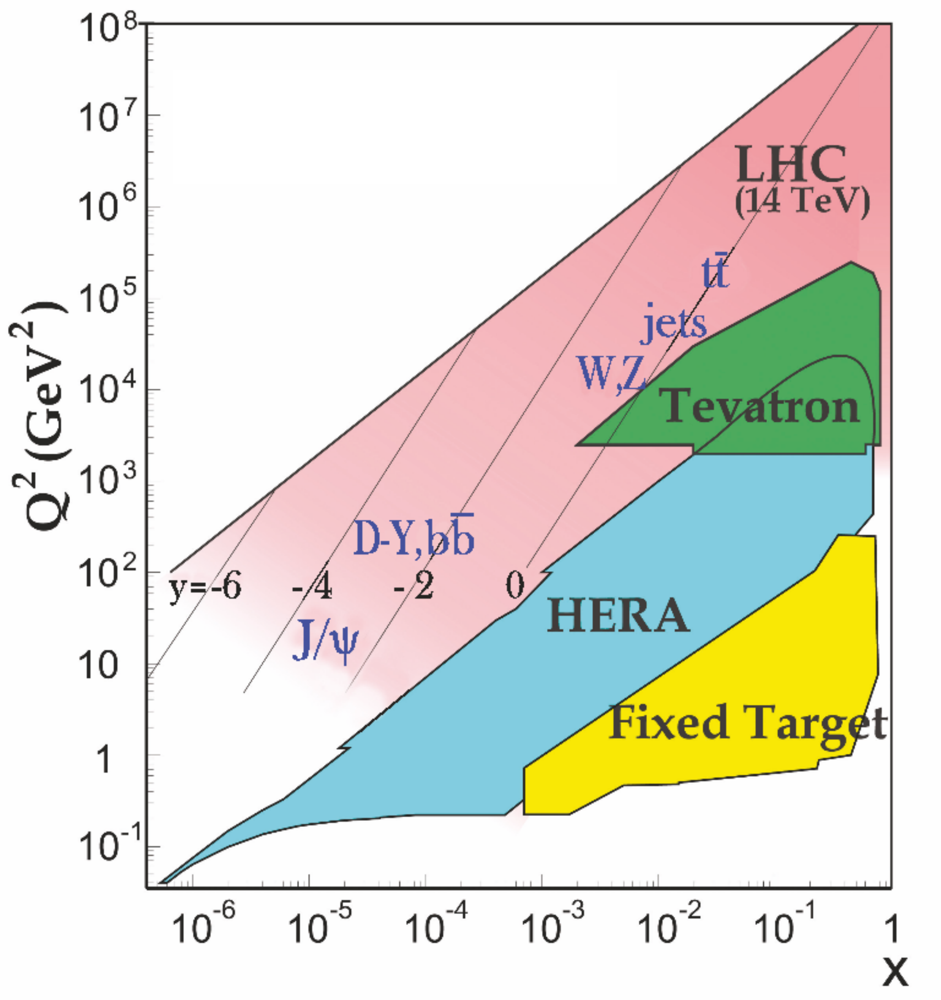


NORM. DISTRIBUTION OF DEVIATIONS

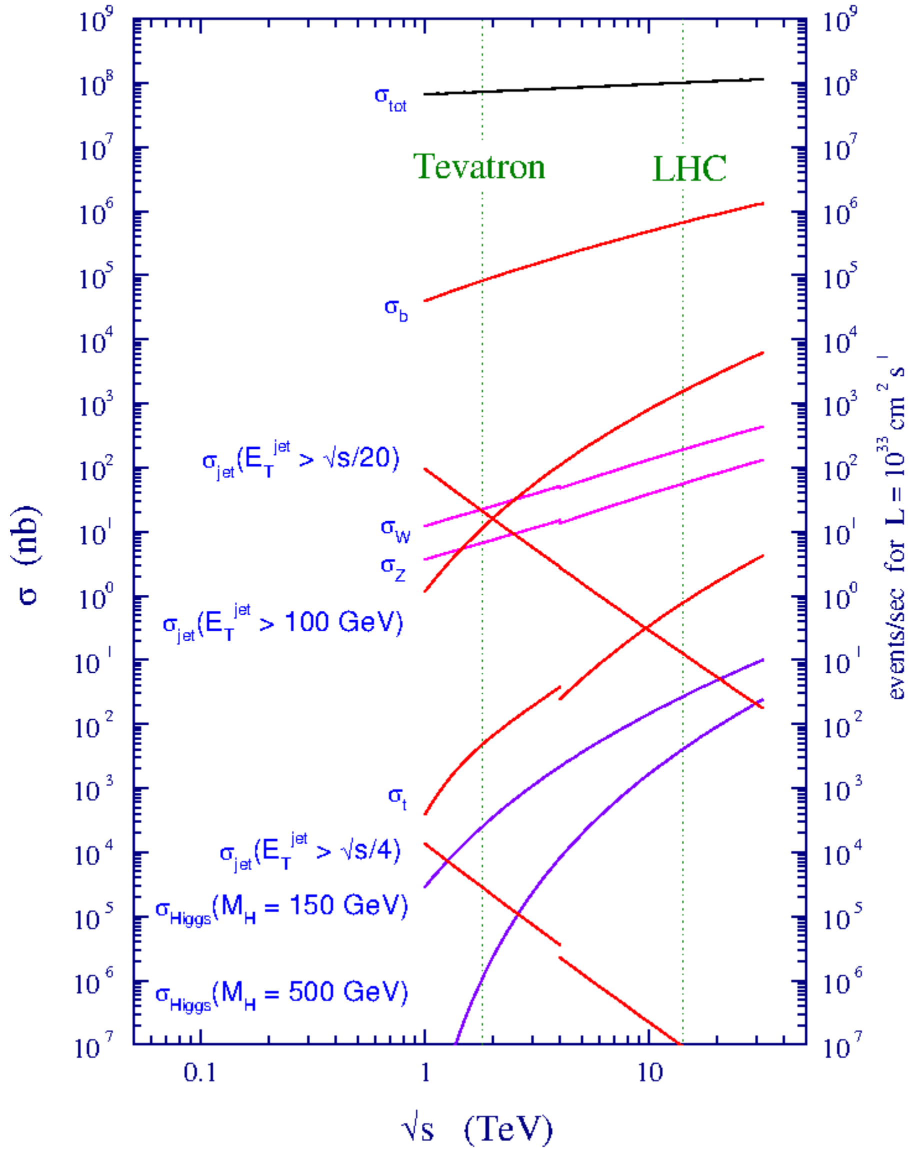


BEFORE AND AFTER THE LHC I

KINEMATIC PLANE



HADRONIC CROSS-SECTIONS

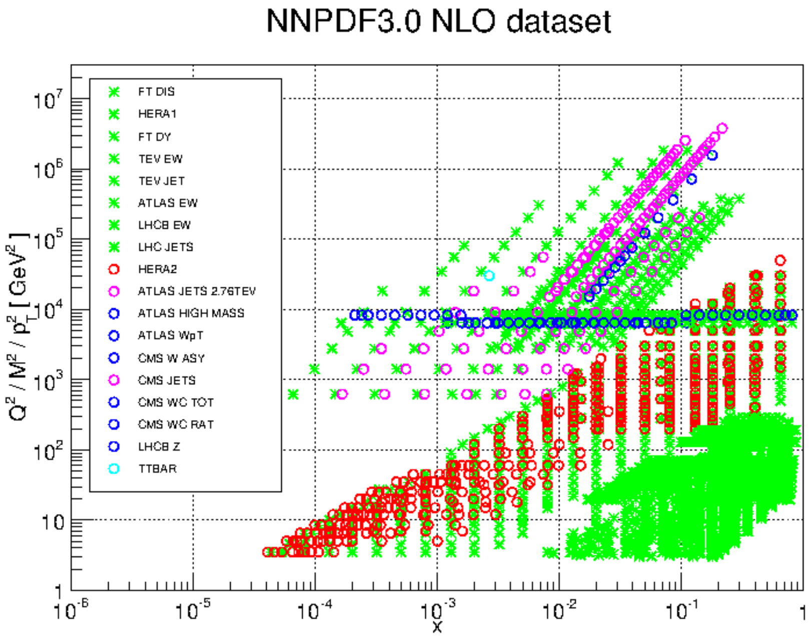


- Q^2 : INVARIANT MASS OF FINAL STATE $\Rightarrow$ WIDENING OF AVAILABLE PROCESSES
- AS ENERGY GROWS, DROP OF CROSS-SECTION MAY BE OFFSET BY GROWTH OF SMALL x PDFs

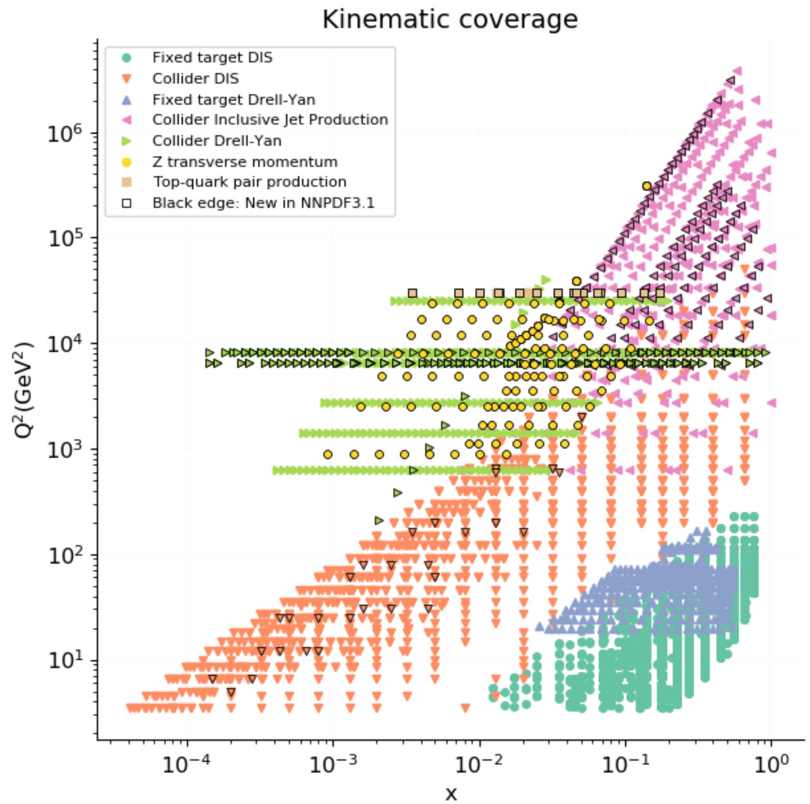
BEFORE AND AFTER THE LHC II

PDFs WITH RUN I DATA

NNPDF3.0 (2014) DATASET



NNPDF3.1 (2017) DATASET

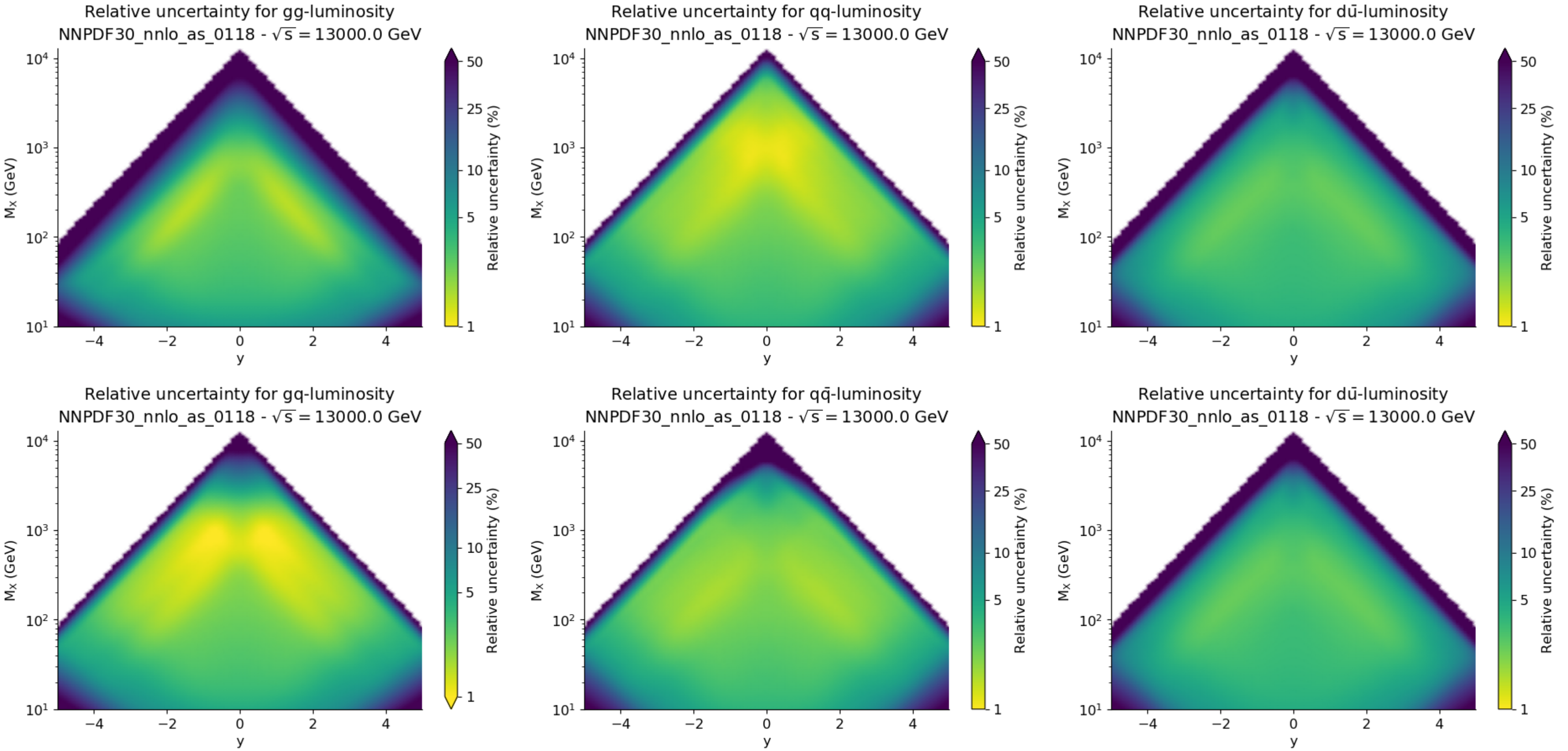


NEW DATA (NNPDF3.1 VS NNPDF3.0):

- TEVATRON LEGACY Z RAPIDITY, W ASYMMETRY & JET DATA
- ATLAS W, Z RAPIDITY, AND TOTAL XSECT (INCL. 13TEV), HIGH AND LOW MASS DY, JETS
- CMS W ASYMMETRY, $W + c$ TOTAL & RATIO, DOUBLE-DIFFERENTIAL DY AND JETS
- LHCB W AND Z RAPIDITY DISTRIBUTIONS
- ATLAS AND CMS Z p_T DISTRIBUTIONS
- ATLAS AND CMS TOP TOTAL CROSS-SECTION & DIFFERENTIAL RAPIDITY DISTRIBUTION

THE IMPACT OF LHC DATA
PDF UNCERTAINTIES: PAST ⇒ PRESENT (NNPDF3.0 NNLO)

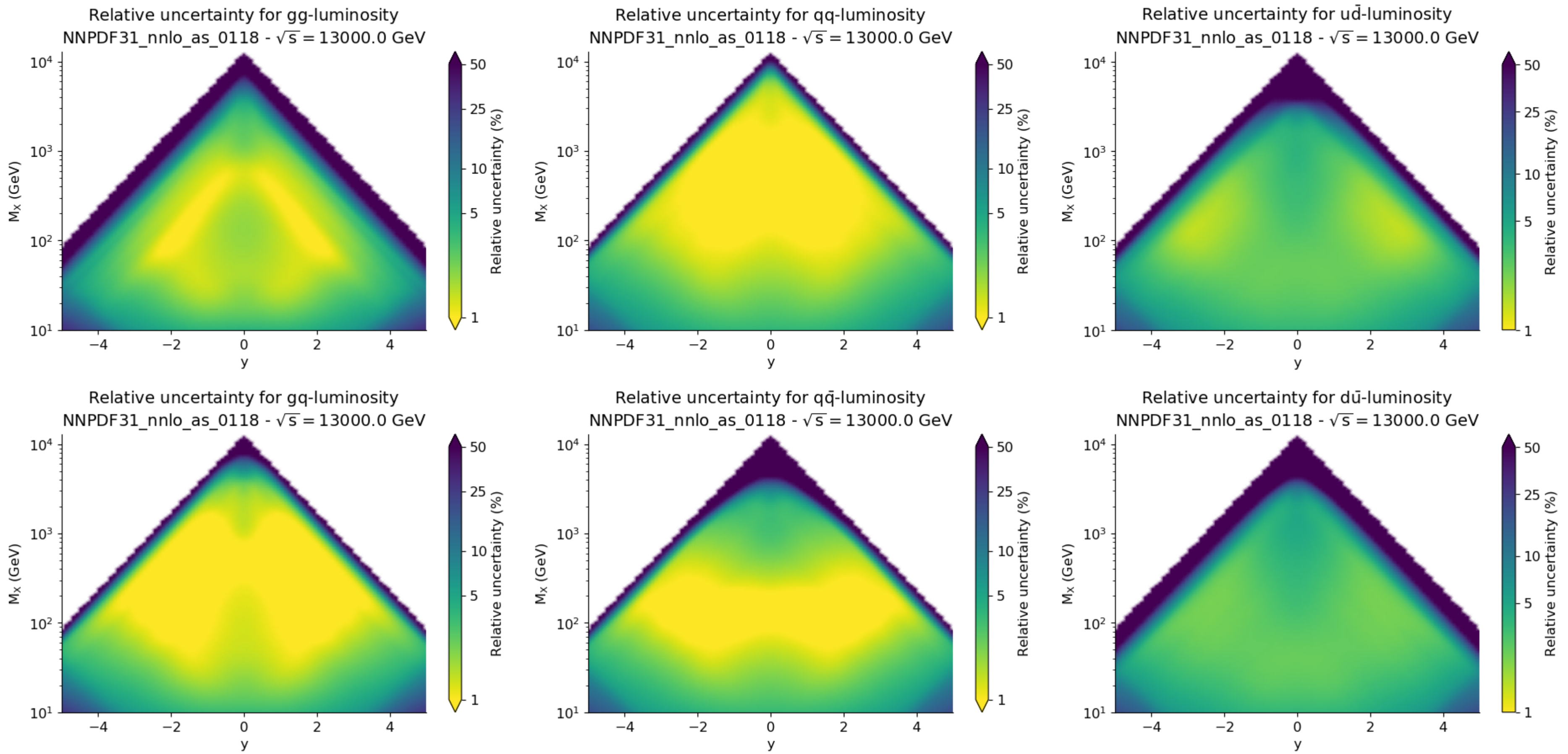
GLUON | SINGLET | FLAVORS



- GLUON BETTER KNOWN AT SMALL x, VALENCE QUARKS AT LARGE x, SEA QUARKS IN BETWEEN
- SWEET SPOT: VALENCE Q - G; UNCERTAINTIES DOWN TO 1%
- UP BETTER KNOWN THAN DOWN; FLAVOR SINGLET BETTER THAN INDIVIDUAL FLAVORS

THE IMPACT OF LHC DATA
PDF UNCERTAINTIES: PRESENT ⇒ FUTURE (NNPDF3.1 NNLO)

GLUON SINGLET FLAVORS



- GLUON BETTER KNOWN AT SMALL x, VALENCE QUARKS AT LARGE x, SEA QUARKS IN BETWEEN
- SWEET SPOT: VALENCE Q - G; UNCERTAINTIES DOWN TO 1%
- UP BETTER KNOWN THAN DOWN; FLAVOR SINGLET BETTER THAN INDIVIDUAL FLAVORS
- NEW LHC DATA ⇒ SIZABLE REDUCTION IN UNCERTAINTIES

SUMMARY

- PDF EXTRACTION NEEDS INPUT FROM A LARGE VARIETY OF PROCESSES $\Rightarrow \sim 5000$ DATAPOINTS&
  HIGHEST ACCURACY CALCULATIONS $\Rightarrow$ NNLO+NNLL
- STATE OF THE ART ANALYSIS TOOLS $\Rightarrow$ NEURAL NETWORKS DEVELOPED OR FORTHCOMING $\Rightarrow$ MACHINE LEARNING
- STATISTICAL VALIDATION TOOLS $\Rightarrow$ CLOSURE TESTS
- ACCURACY GOAL FOR NEW PHYSICS SEARCHES $\Rightarrow$ 1%