JHEP

Published for SISSA by Springer

Received: January 26, 2022
Revised: May 25, 2022
Accepted: June 30, 2022
Published: August 5, 2022

Machine learning of log-likelihood functions in global analysis of parton distributions

DianYu Liu,[a] **ChuanLe Sun**[a] **and Jun Gao**[a,b,c]

[a] *INPAC, Shanghai Key Laboratory for Particle Physics and Cosmology, School of Physics and Astronomy, Shanghai Jiao-Tong University, Shanghai 200240, China*

[b] *Key Laboratory for Particle Astrophysics and Cosmology (MOE), Shanghai 200240, China*

[c] *Center for High Energy Physics, Peking University, Beijing 100871, China*

E-mail: dianyu.liu@sjtu.edu.cn, chlsun60@sjtu.edu.cn, jung49@sjtu.edu.cn

Abstract: Modern analysis on parton distribution functions (PDFs) requires calculations of the log-likelihood functions from thousands of experimental data points, and scans of multi-dimensional parameter space with tens of degrees of freedom. In conventional analysis the Hessian approximation has been widely used for the estimation of the PDF uncertainties. The Lagrange Multiplier (LM) scan while being a more faithful method is less used due to computational limitations, and is the main focus of this study. We propose to use Neural Networks (NNs) and machine learning techniques to model the profile of the log-likelihood functions or cross sections for multi-dimensional parameter space in order to overcome those limitations which work beyond the quadratic approximations and meanwhile ensures efficient scans of the full parameter space. We demonstrate the efficiency of the new approach in the framework of the CT18 global analysis of PDFs by constructing NNs for various target functions, and performing LM scans on PDFs and cross sections at hadron colliders. We further study the impact of the NOMAD dimuon data on constraining PDFs with the new approach, and find enhanced strange-quark distributions and reduced PDF uncertainties. Moreover, we show how the approach can be used to constrain new physics beyond the Standard Model (BSM) by a joint fit of both PDFs and Wilson coefficients of operators in the SM effective field theory.

Keywords: Parton Distributions, Deep Inelastic Scattering or Small-X Physics

ArXiv ePrint: 2201.06586

https://doi.org/10.1007/JHEP08(2022)088

JHEP08(2022)088

Contents

JHEP08(2022)088

1 Introduction

Precise understanding of the parton structure of the proton is a central topic of QCD [1, 2]. The parton structure can be described by parton distribution functions (PDFs), which represent distributions of momentum fractions of the proton carried by quarks and gluons, for instance in the case of QCD collinear factorization [3]. They are usually determined by fitting to a variety of experimental data, such as data from proton-proton collision, proton-antiproton collision, electron-proton collision, and neutrino–nucleus scatterings. Besides, there have also been recent developments on calculating PDFs from first principles based on the large momentum effective theory [4] and lattice QCD simulations [5].

Especially, PDFs play important roles in LHC studies. For example, PDF uncertainties represent one of the dominant uncertainties in measurements of the Higgs boson

couplings [6]. Better control of PDF uncertainties are necessary in direct searches for new heavy resonances [7] and indirect searches for new physics beyond the SM [8]. Furthermore, PDF uncertainties also have a large impact on precision measurements of the SM parameters including the strong coupling constant [9], the weak mixing angle and the W boson mass [10, 11].

Modern analysis of PDFs requires calculations of the log-likelihood functions from thousands of experimental data points, and scans of multi-dimensional parameter space with tens of degrees of freedom. There are several groups providing regular updates of PDFs via global fits, see refs. [12–19] for recent results on PDF determinations. The difference between those PDF sets is mainly due to the choice of the experimental data sets, the theoretical calculations used, and the parametrization form of PDFs.

PDF uncertainties can be determined with three methods: the Hessian [20, 21], Monte Carlo (MC) [22], and Lagrange Multiplier (LM) [23, 24] method. There also exist recently developed approaches, meta analysis [25], ePump [26] and L2 sensitivity [27], on accessing impacts of experimental data on PDFs based on the Hessian method. In the Hessian method, the log-likelihood function (χ^2) of a global fit is approximated with a quadratic form of the PDF parameters at the neighborhood of the global minimum. The uncertainties are thus determined through error PDFs along eigenvector directions, constructed by requiring the increase of the total χ^2 of 1 or of a certain tolerance. In the MC method, one can obtain the PDF uncertainties from an ensemble of PDF replicas which are fitted to an ensemble of "pseudo-data". Those pseudo-data are generated from the probability distributions related to the original experimental data sets. On another hand, for the LM method, PDF uncertainties of an observable can be determined from the profiled χ^2 as a function of the observable, without relying on any assumptions about the behavior of the χ^2 at the neighborhood of the global minimum. This means PDF uncertainties estimated from the LM method are more robust than those from the Hessian method. However, the LM method requires a detailed scan of the PDF parameter space for every observable studied, which is usually time consuming.

This drawback can be overcome with the help of machine learning (ML). ML has been widely used in studies of high-energy physics in recent years. In many cases, ML is used for classifications such as particle identification and event selection in experimental data analysis [28]. Neural networks (NNs) are also helpful in regression problems, for example, applications of NNs in the study of PDFs have been pioneered by the NNPDF collaboration [29]. Dependence of PDFs on the momentum fraction are parametrized using NNs, which ensures a great flexibility [30]. On another hand, dependence of the χ^2 or any physics quantity, such as the cross section, on PDFs is complex in general. NNs offer an opportunity to relate physics quantities to PDFs efficiently. One can build NNs with PDFs as input variables to model their PDF dependence. Compared with traditional methods, NNs can greatly improve efficiencies on generating predictions for those physics quantities.

With above motivations, in this paper we propose a new approach with which PDF uncertainties can be calculated efficiently using the LM method with the assistance of NNs. It takes three steps to achieve this goal. First, we construct and train NNs to model the χ^2 of each individual data set used in the global fit with PDFs. Second, we construct and

JHEP08(2022)088

train other NNs to associate the physics quantity to be studied with PDFs. Finally, we can perform LM scans to determine PDF uncertainties in a robust way. The speed of LM scans can be improved by several orders of magnitude due to the introduction of the NNs. We demonstrate above idea in the framework of CT18 NNLO global analysis [12] and beyond. We show how the new approach can help to understand various PDF uncertainties and the interplay between different data sets in the global fit. Moreover, we explore several directions beyond CT18 as will be explained below.

Only a few data sets in the CT18 global fit are sensitive to the strange-quark PDFs. The dimuon production in neutrino scatterings provides an opportunity to directly constrain strange-quark distributions in the nucleon. In recent NOMAD measurements [31], a sample of about 9×10^6 events of inclusive charged-current deep-inelastic scattering (CCDIS), together with about 15344 events of dimuon production, is collected. The large statistics lead to a better control on various systematic errors and also an improvement in statistical uncertainties. We include the NOMAD data in the global fit and evaluate the impact on the PDFs using the aforementioned approach.

The High Luminosity LHC (HL-LHC) is supposed to accumulate an integrated luminosity of $3000\,\mathrm{fb}^{-1}$ for ATLAS and CMS and of $300\,\mathrm{fb}^{-1}$ for LHCb [32]. We take two of those HL-LHC pseudo-data sets constructed in refs. [33, 34], the high-mass Drell-Yan data and the forward W/Z production data, and evaluate their impacts on PDFs. Our projection shows they can largely improve separations of different flavors, especially for sea quarks.

In the searches for new physics beyond the SM from scatterings involving nucleons, for instance at HERA or LHC, one key problem is on the degeneracy of PDF variations and the new physics contributions, especially in cases when similar measurements are used in both the global fit of PDFs and in the searches of new physics. Ideally a joint global fit including both PDFs and model parameters of the new physics should be performed, see refs. [35–39] for examples. We demonstrate successful application of our approach in such scenario by a simultaneous fit of both PDFs and the Wilson coefficient of lepton-quark contact interactions in the SM effective field theory (SMEFT).

The rest of this paper is organized as follows. In section 2, we describe the basic setup of our approach, including architectures of the NNs, PDF parametrizations and experimental data sets considered in the global fit. In section 3, we discuss performances of the approach and show that the accuracy of approximations with NNs are far sufficient for phenomenological studies. In section 4, we explain the method of LM scans and discuss several features of the CT18 analysis based on the new approach. In section 5, we study the impact of the NOMAD measurements and of the two pseudo-data of HL-LHC on PDFs, and show a joint fit with both PDFs and new physics contributions. Finally, we conclude in section 6.

2 Setup of the Neural Network program

In this section, we give a brief introduction to the setup of our NNs, including the architectures, the input variables, and the target functions. We further explain the training processes from the generation of samples to the minimization of the loss function.

JHEP08(2022)088

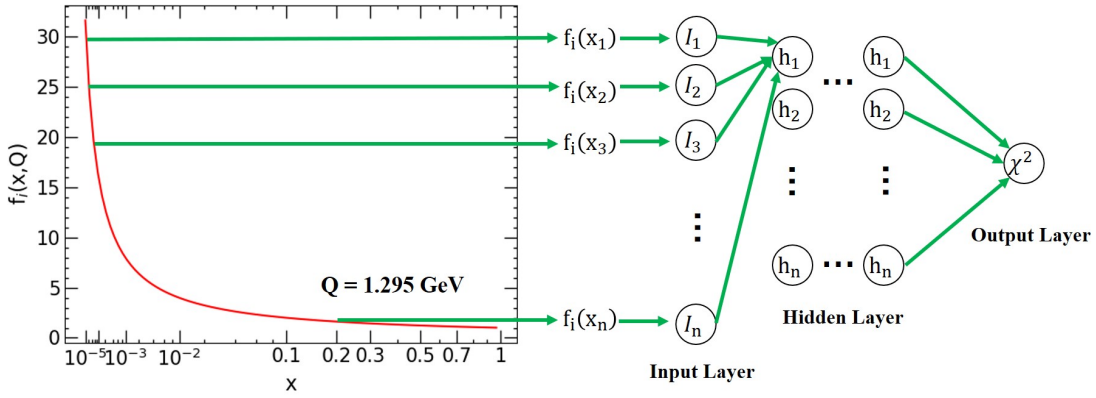


Figure 1. An example of the architecture of NNs in this work, taking χ^2 as the target function.

Target	No. of hidden layers	No. of nodes for each hidden layer	Activation functions for each layer	No. of total params
χ^2	2	60,40	tanh, (x^2+2), linear	7581
σ, $f_i(x,Q)$, $f_i(x,Q)/f_j(x,Q)$	1	40	tanh, linear	3441

Table 1. The architecture of NNs in this paper. Structure is set up for either χ^2 or other quantities.

2.1 Basic setup of NNs

The general structure of NNs includes three parts: the input layer, several hidden layers and the output layer. Each of these layers contains a collection of nodes termed by perceptrons. There exist various implementations of NNs, and we use Keras [40] in this work. From the NNs built by Keras, PDFs as inputs are associated with either χ^2 or physics quantities as outputs. The log-likelihood function χ^2 quantifies agreements between theory predictions and experimental measurements for each data set and is calculated according to [12]. The physics quantities considered include cross sections of several benchmark processes at the LHC, and PDFs or their ratios at different Q values. An example of the architecture of our NNs is shown in figure 1, in which the inputs are PDFs at an initial scale and the outputs are the χ^2 of the fit to an experimental data set. In this figure, the PDFs $f_i(x,Q)$ are evaluated at an initial scale of $Q = 1.295\,\text{GeV}$ with x selected among 14 different values, and $i \in \{g, u, d, \bar{u}, \bar{d}, s\}$ runs over all parton flavors. We always assume $s = \bar{s}$ at the initial scale. They altogether form the input layer with 84 nodes $\{I_1, I_2 \ldots I_{84}\}$. In addition, the differences in setups between NNs for different target functions are shown in table 1. The choice on the architecture is based on the observation that cross sections or evolved PDFs are in general non-linear functions of the PDF parameters. The χ^2 is positive defined and is a sum of various individual terms that depend on cross sections quadratically, and thus can be approximated by a more complicated architecture as prescribed. We include more details on the construction of our NNs in appendix A.

JHEP08(2022)088

To construct the l_{th} layer of a NN, we define

$$b_i^{(l)} = \begin{cases} \sum_j w_{ij}^{(l)} I_j, & (l=1), \\ \sum_j w_{ij}^{(l)} h_j^{(l-1)}, & (l>1), \end{cases} \tag{2.1}$$

where $b_i^{(l)}$ is the value before the activation of the i_{th} node in the l_{th} layer, $w_{ij}^{(l)}$ is the weight matrix connecting the $(l-1)_{th}$ layer to the l_{th} layer, I_j is the value of the j_{th} node in the input layer, and $h_j^{(l-1)}$ is the value of the j_{th} node in the $(l-1)_{th}$ layer. The value of the i_{th} node in the l_{th} layer is then obtained by applying the activation function $t^{(l)}$ on $b_i^{(l)}$:

$$h_i^{(l)} = t^{(l)}\left(b_i^{(l)}\right). \tag{2.2}$$

This procedure iterates over all hidden layers, and in the end we obtain a single value for the output layer. The activation functions used include the conventional choices of linear, and tanh types, as well as a customized one of quadratic form, depending on the target functions and layers. Note we constrain elements of the weight matrix of the output layer to be positive for the NN associated with the χ^2 since it is positive definite. Elements in the weight matrix, $w_{ij}^{(l)}$, are trained to minimize the so-called loss function, which is defined as

$$d_{\text{loss}} = \frac{1}{n}\sum_{k=1}^{n}\left(A_{\text{NN}}^k(w_{ij}) - A_{\text{TR}}^k\right)^2, \tag{2.3}$$

where n is the total number of events in the training sample and A_{TR}^k and A_{NN}^k are the truth of the target function and the prediction from NNs for the k_{th} event.

2.2 PDF parametrization form

The parametrization form of PDFs used at the initial scale Q_0 is

$$f_i\left(x, Q_0\right) = a_0 x^{a_1-1}(1-x)^{a_2} P_i\left(y; a_3, a_4, \ldots\right), \tag{2.4}$$

where $\{a_1, a_2, \ldots\}$ are free parameters, and the behavior of x^{a_1} at $x \to 0$ and $(1-x)^{a_2}$ at $x \to 1$ is guided by Regge theory and spectator counting rules respectively. $P_i\left(y; a_3, a_4, \ldots\right)$ is a polynomial dependent on $y \equiv \sqrt{x}$ ($y \equiv 1-(1-\sqrt{x})^{a_3}$) for valence quark and gluon PDFs (light-quark sea PDFs). Parametrization forms used here are the same as in the CT18 NNLO analysis [12].

For the valence-quark (u_v and d_v) PDF,

$$\begin{aligned} f_v\left(x, Q_0\right) &= a_0 x^{a_1-1}\left(1-x\right)^{a_2} P_v\left(y\right), \\ P_v\left(y\right) &= \sinh\left[a_3\right]\left(1-y\right)^4 + \sinh\left[a_4\right] 4y\left(1-y\right)^3 + \sinh\left[a_5\right] 6y^2\left(1-y\right)^2 \\ &\quad + \left(1+\frac{1}{2}a_1\right) 4y^3(1-y) + y^4. \end{aligned} \tag{2.5}$$

For the gluon PDF,

$$\begin{aligned} f_g\left(x, Q_0\right) &= a_0 x^{a_1-1}(1-x)^{a_2} P_g(y), \\ P_g(y) &= \sinh\left[a_3\right]\left(1-y\right)^3 + \sinh\left[a_4\right] 3y(1-y)^2 + (3+2a_1)y^2(1-y) + y^3. \end{aligned} \tag{2.6}$$

JHEP08(2022)088

For the sea quark ($\bar{u}$, $\bar{d}$ and $s \equiv \bar{s}$) PDF,

$$\begin{aligned} f_{\bar{q}}(x, Q_0) &= a_0 x^{a_1 - 1}(1-x)^{a_2} P_{\bar{q}}(y), \\ P_{\bar{q}}(y) &= (1-y)^5 + a_4 5y(1-y)^4 + a_5 10y^2(1-y)^3 + a_6 10y^3(1-y)^2 \\ &\quad + a_7 5y^4(1-y) + a_8 y^5. \end{aligned} \tag{2.7}$$

In all, we have 8 free parameters for valence quarks after applying the valence sum rules and letting a_1 be equal for u_v and d_v. We have 15 free parameters for sea quarks after fixing some of those a_i or letting them be equal for different flavors [12]. We are left with 5 free parameters for gluon after applying the momentum sum rule. The total number of free PDF parameters is 28.

2.3 Targets and samples

In this paper, we associated PDFs with χ^2 and other physics quantities through our NNs. Details of these target functions are described in the following:

- The individual χ^2 of each data set in an NNLO global analysis of PDFs. We use the same 39 experimental data sets as in CT18 NNLO global analysis. These experimental data sets are summarized in table 2. The theoretical calculations used are explained in the CT18 paper [12]. We take those calculations from CT18 except for minor updates on NNLO K-factors of several data sets. The global χ^2 is simply a sum of the 39 individual χ^2.

- The cross sections of Higgs boson pair (top-quark pair with a Higgs boson) production in proton-proton collisions at center of mass energy $\sqrt{s} = 13\,\text{TeV}$ or $100\,\text{TeV}$. They are computed at leading (next-to-leading) order in QCD using MG5_aMC@NLO [41] and AMCfast [42] to provide an interface with APPLgrid [43]. We choose these two processes for demonstrations, and any scattering cross sections at hadron collisions can be included in a similar way.

- The PDFs and PDF ratios at various x and Q values. They are obtained using HOPPET [44] with DGLAP evolutions at NNLO.

We first generate randomly a training sample consisting of 6000 replicas of PDFs and another test sample of 2000 replicas to prevent from over training. Details about the generation of the replicas of PDFs can be found in appendix A. We compute all the target functions (χ^2 or physics quantities) for each of the replicas, which can be time consuming depending on whether the fast interpolation approaches, like APPLgrid or FastNLO, are used or not. However, we only need to perform these heavy calculations once for all. Afterwards we construct a NN for each of the target function considered with the architectures shown in table 1. We train each NN for about 10 hours, depending slightly on the architecture, on a single CPU-core (2.4 GHz) according to the loss function defined in eq. (2.3). Thus for all χ^2 of the 39 individual data sets that takes about 390 core-hours in total for the training process. We found a very good performance of the

JHEP08(2022)088

ID	Experimental data set	N_{pt}	ID	Experimental data set	N_{pt}
160	HERA I+II 1 fb^{-1}, H1 and ZEUS NC and CC reduced cross section comb. [45]	1120	101	BCDMS F_2^p [46]	337
102	BCDMS F_2^d [47]	250	104	NMC F_2^d/F_2^p [48]	123
108	CDHSW F_2^p [49]	85	109	CDHSW $x_B F_3^p$ [49]	96
110	CCFR F_2^p [50]	69	111	CCFR $x_B F_3^p$ [51]	86
124	NuTeV $\nu\mu\mu$ SIDIS [52]	38	125	NuTeV $\bar{\nu}\mu\mu$ SIDIS [52]	33
126	CCFR $\nu\mu\mu$ SIDIS [53]	40	127	CCFR $\bar{\nu}\mu\mu$ SIDIS [53]	38
145	H1 σ_r^b [54]	10	147	Combined HERA charm production [55]	47
169	H1 F_L [56]	9	201	E605 Drell-Yan process [57]	119
203	E866 Drell-Yan process $\sigma_{pd}/(2\sigma_{pp})$ [58]	15	204	E866 Drell-Yan process $Q^3 d^2\sigma_{pp}/(dQdx_F)$ [59]	184
225	CDF Run-1 lepton A_{ch}, $p_{Tl} > 25$ GeV [60]	11	227	CDF Run-2 electron A_{ch}, $p_{Tl} > 25$ GeV [61]	11
234	D$\varnothing$ Run-2 muon A_{ch}, $p_{Tl} > 20$ GeV [62]	9	260	D$\varnothing$ Z rapidity [63]	28
261	CDF Run-2 Z rapidity [64]	29	266	CMS 7 TeV 4.7 fb^{-1}, moun A_{ch}, $p_{Tl} > 35$ GeV [65]	11
267	CMS 7 TeV 840 fb^{-1}, electron A_{ch}, $p_{Tl} > 35$ GeV [66]	11	268	ATLAS 7 TeV 35 pb^{-1} W/Z cross section, A_{ch} [67]	41
281	D$\varnothing$ Run-2 9.7 fb^{-1} electron A_{ch}, $p_{Tl} > 25$ GeV [68]	13	504	CDF Run-2 inclusive jet production [69]	72
514	D$\varnothing$ Run-2 inclusive jet production [70]	110	245	LHCb 7 TeV 1.0 fb^{-1} W/Z forward rapidity cross section [71]	33
246	LHCb 8 TeV 2.0 fb^{-1} $Z \to e^- e^+$ forward rapidity cross section [72]	17	249	CMS 8 TeV 18.8 fb^{-1} muon charge asymmetry A_{ch} [73]	11
250	LHCb 8 TeV 2.0 fb^{-1} W/Z cross section [74]	34	253	ATLAS 8 TeV 20.3 fb^{-1}, Z p_T cross section [75]	27
542	CMS 7 TeV 5 fb^{-1}, single incl. jet cross section, R = 0.7 (extended in y) [76]	158	544	ATLAS 7 TeV 4.5 fb^{-1}, single incl. jet cross section, R = 0.6 [77]	140
545	CMS 8 TeV 19.7 fb^{-1}, single incl. jet cross section, R = 0.7, (extended in y) [78]	185	573	CMS 8 TeV 19.7 fb^{-1}, $t\bar{t}$ norm. double-diff. top p_T and y cross section [79]	16
580	ATLAS 8 TeV 20.3 fb^{-1}, $t\bar{t}$ p_T^t and $m_{t\bar{t}}$ abs. spectrum [80]	15			

Table 2. Experimental data sets involved in the global fit [12].

resulting NNs without much tuning on the training process for all target functions studied, which will be reported in the next section. In a later stage for the evaluation of the target functions with arbitrary PDF parameters, we can simply use the optimized NNs rather than direct calculations. Comparison between computational cost of the NNs and the direct computations are summarized in appendix A where substantial improvements in the speed from the NNs are observed.

3 Validation of NNs

In this section, we perform several comparisons between the truths and the predictions from our NNs before we apply them to further phenomenological studies. We emphasize that the entire NN approach we discussed so far and in the following is bound to the CT18 parametrization form, especially with the CT18 PDF set. All PDF replicas for training and testing are sampled from the CT18 PDFs. A first attempt of generalization to other parametrization forms or even independent of PDF parametrization shows promising results, and is detailed in appendix A. It should be noted that the NNs should be retrained in general if the underlying PDF parametrization changes.

JHEP08(2022)088

3.1 χ^2 of the global fit

In figure 2, we show the predictions to truths ratios of χ^2 for three experimental data sets: measurements of the proton structure function by BCDMS, measurements of inclusive DIS reduced cross sections at HERA and measurements of the inclusive jet cross sections at $\sqrt{s} = 7\,\text{TeV}$ by CMS. The ratio of total χ^2 for the full data set is also shown in the lower-right panel. The ratios are calculated for the PDFs from the aforementioned training sample and test sample of NNs as well as the CT18 NNLO PDFs. The CT18 NNLO PDFs consist of a central PDF set and 56 error PDFs in a total of 28 Hessian eigenvector directions. The horizontal axis represents the truths of χ^2. Each mark corresponds to a PDF set from these three samples of PDF sets. The green squares and the blue circles represent the ratios corresponding to the PDFs from the training sample and test sample respectively, and the purple triangles represent the ratios corresponding to the PDFs from CT18 NNLO. We find good agreement between the training and test samples, although the NN produces greater deviation than the original CT18 NNLO PDFs. We find that the predictions and the truths in general agree within 1 per mille for each data set. For the total χ^2, the deviation is within 0.6 per mille.

We define $\Delta\chi^2$ as the difference between a certain χ^2 value and its value at the best fit, which is conventionally used in the determination of PDF uncertainties. In figure 3, differences between predicted and true $\Delta\chi^2$ denoted as $\delta(\Delta\chi^2_\alpha) \equiv \Delta\chi^2_{\alpha,pre} - \Delta\chi^2_{\alpha,tru}$ are demonstrated for each data set, where α represents the PDF set used in the comparison. Here we choose a sample of PDF sets consisting of the 56 Hessian error PDFs in CT18 NNLO set, which are represented by the marks distributed along the vertical direction. We also show similar results for the total χ^2. It can be seen that, for each individual data set the $\delta(\Delta\chi^2_\alpha)$ is at most 1 unit, and the $\delta(\Delta\chi^2_\alpha)$ for the full data set are within 2 units. The extent of $\delta(\Delta\chi^2_\alpha)$ for HERA inclusive DIS data set and for total χ^2 is slightly larger than other experimental data sets. It should be noticed that the number of data points of HERA inclusive DIS data set and the full data set is 1120 and 3671 respectively. Besides, the $\Delta\chi^2_\alpha$ of the full data set for most of the 56 Hessian error PDFs is about 100 units. This indicates the relative deviation of NNs predictions is below 2% for $\Delta\chi^2$.

Furthermore, we compare the $\Delta\chi^2$ for the full data set along the 28 eigenvector directions of CT18 NNLO PDFs by scans of PDF parameters. A variable d is introduced to measure the distance that PDF parameters go along the direction of a certain eigenvector. The variation of PDF parameters for the scan along the j_{th} eigenvector direction can be written as:

$$a_i^{j,scan}(d) = \begin{cases} d\left(a_i^{2j-1} - a_i^0\right) + a_i^0, & (d > 0)\,, \\ d\left(a_i^0 - a_i^{2j}\right) + a_i^0, & (d < 0)\,, \end{cases} \tag{3.1}$$

where i represents the index of the PDF parameters, $\{a_i^0\}$ represents PDF parameters for the central PDF of CT18 NNLO, $\{a_i^{2j-1}\}$ and $\{a_i^{2j}\}$ represent PDF parameters of the two error PDFs in the j_{th} eigenvector direction of CT18 NNLO. The total χ^2 are computed using the new set of PDF parameters. We define $\Delta\chi^2 \equiv \chi^2(d) - \chi^2(d=0)$, and compare the

JHEP08(2022)088

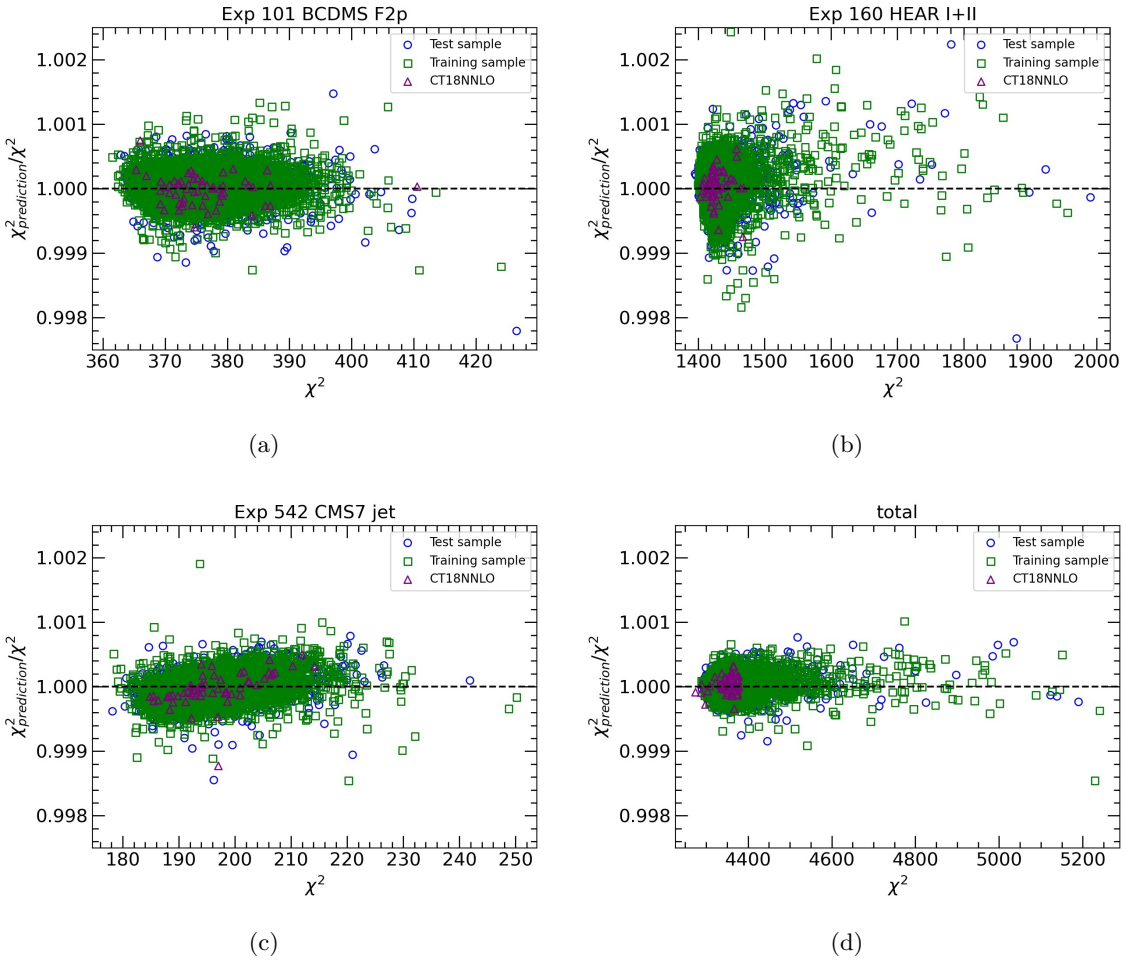


Figure 2. The predictions to truths ratios of χ^2 for experimental data sets for measurements of the proton structure function by BCDMS, measurements of inclusive DIS reduced cross sections at HERA and measurements of the inclusive jet cross sections at $\sqrt{s} = 7\,\text{TeV}$ by CMS as well as for the full data set.

truths of $\Delta\chi^2$ and the predictions from NNs as a function of d for a few selected eigenvector directions in figure 4.

We find that the NNs can describe well the dependence of the $\Delta\chi^2$ on PDF parameters in all Hessian eigenvector directions. The predictions and the truths in general agree within 2% in all directions, which agrees with figure 3. We also notice that $\Delta\chi^2$ has a sizable deviation from quadratic shape in some directions. The NNs can reproduce well the asymmetric and non-quadratic behavior of $\Delta\chi^2$, which is one of the main advantages as comparing to the traditional Hessian method. It is justified to say the deviation of χ^2 due to the NNs approximation is negligible for the PDF parameter space of interest.

3.2 Physics quantities

In figure 5, we show the ratios of the predictions to the truths for the cross section of top-quark pair with a Higgs boson production in proton-proton collisions at a center of

JHEP08(2022)088

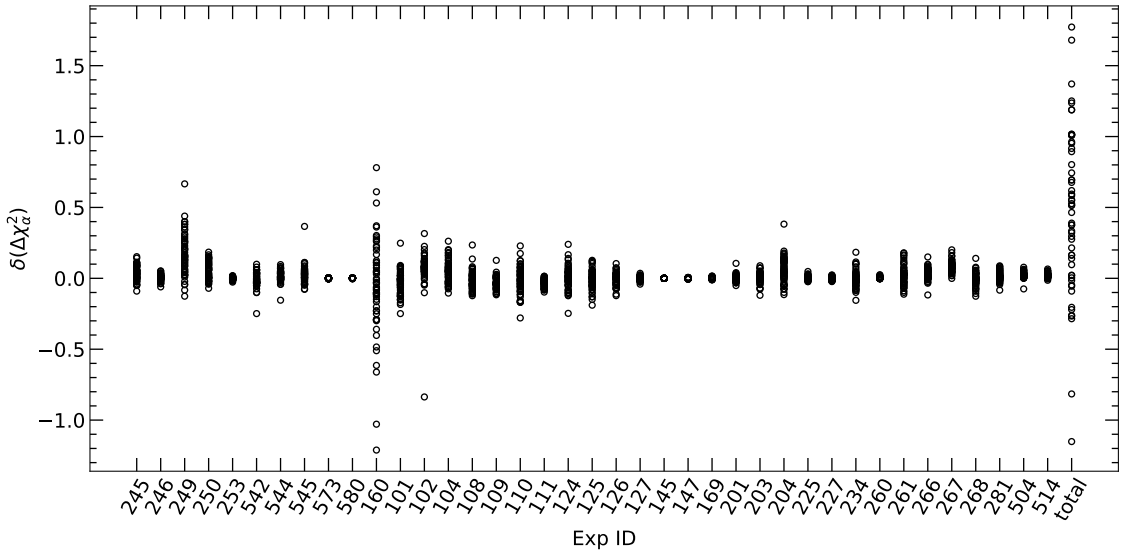


Figure 3. The $\delta(\Delta\chi^2_\alpha)$ corresponding to the 56 error PDFs of CT18 NNLO are represented by the 56 marks distributed along the vertical direction for each individual data set and for the full data set.

mass energy $\sqrt{s} = 13\,\text{TeV}$, and for the PDF ratio $d/u(x = 0.3, Q = 100\,\text{GeV})$. The ratios are calculated for the PDFs from the training sample and test sample of NNs as well as the CT18 NNLO PDFs. We find good agreements between the distributions of marks for training sample and for test sample. Deviations for these two physics quantities are in general within 0.15 per mille and 0.2 per mille, respectively. The performance of the NNs for these two physics quantities is better than that for χ^2, which is because the dependence of χ^2 on PDFs is more complex than the cases for cross sections or PDF ratios. The dependence of these physics quantities on PDFs is even close to linear.

We further summarize the relative difference between the predictions from NNs and the truths for various physics quantities in figure 6. For each physics quantity, 57 marks distributed along the vertical direction correspond to the results from the 57 CT18 NNLO PDFs. The results for the cross sections of Higgs boson pair production and top-quark pair with a Higgs boson production in proton-proton collisions at center of mass energy $\sqrt{s} =$ 13 TeV or 100 TeV are shown in this figure. We find the predictions from NNs and the truths for these cross sections agree within 0.2 per mille. In addition, the results for cross sections of $pp \to hh$ and $pp \to ht\bar{t}$ with high invariant mass $m_{hh} > 2.5\,\text{TeV}$ or $m_{ht\bar{t}} > 2.5\,\text{TeV}$ are also shown in figure 6, and the relative difference between the predictions and truths in general agree within 0.3 per mille. Comparisons for strangeness ratio $R_s \equiv \dfrac{s(x,Q) + \bar{s}(x,Q)}{\bar{u}(x,Q) + \bar{d}(x,Q)}$ and PDF ratios d/u and $\bar{d}/\bar{u}$ at various x and Q are also shown in this figure, and the predictions and the truths agree within 1 per mille. We also show the results for PDF values for g and s-quark at various x and Q, and the deviations between the predictions and the truths are within 0.75 per mille.

JHEP08(2022)088

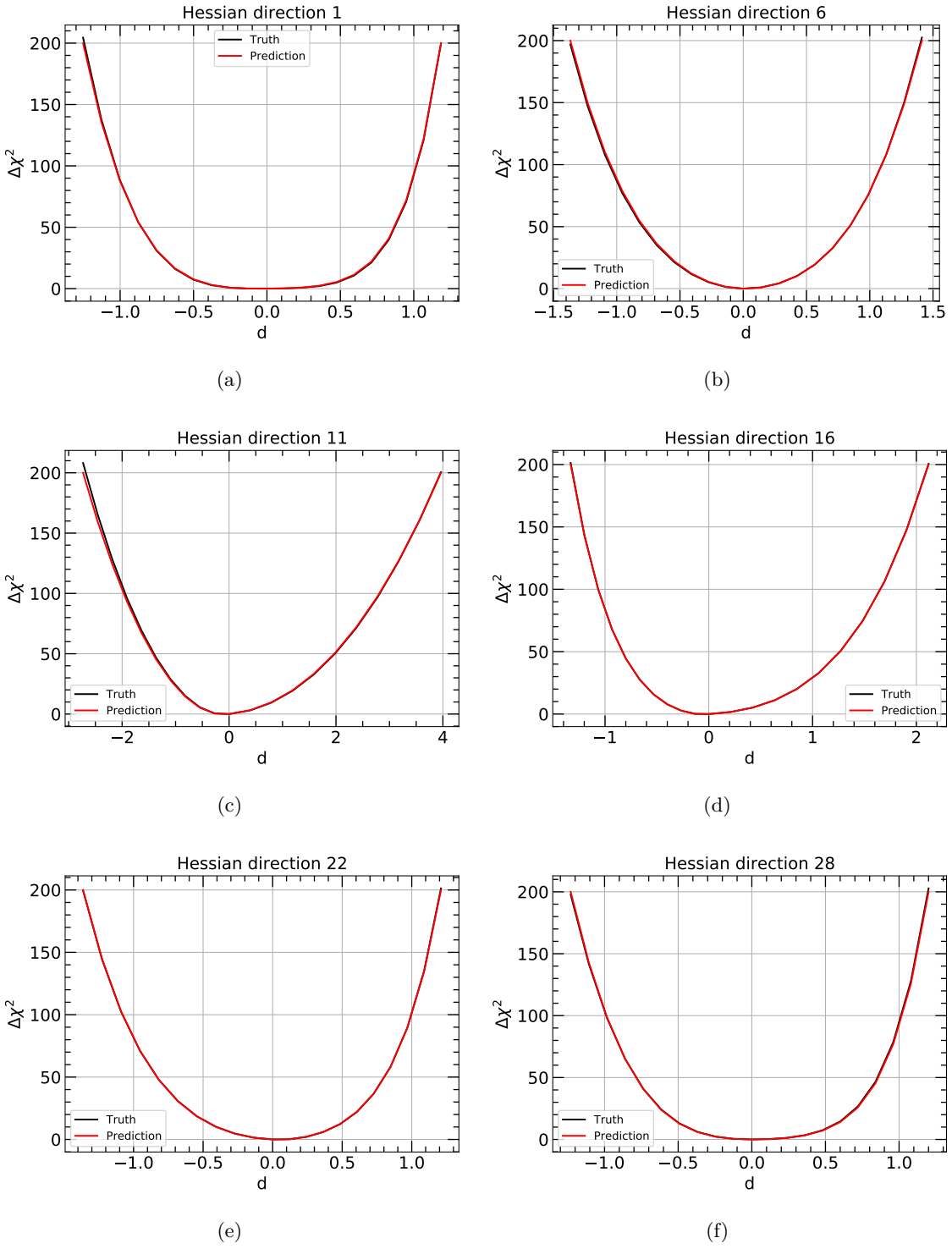
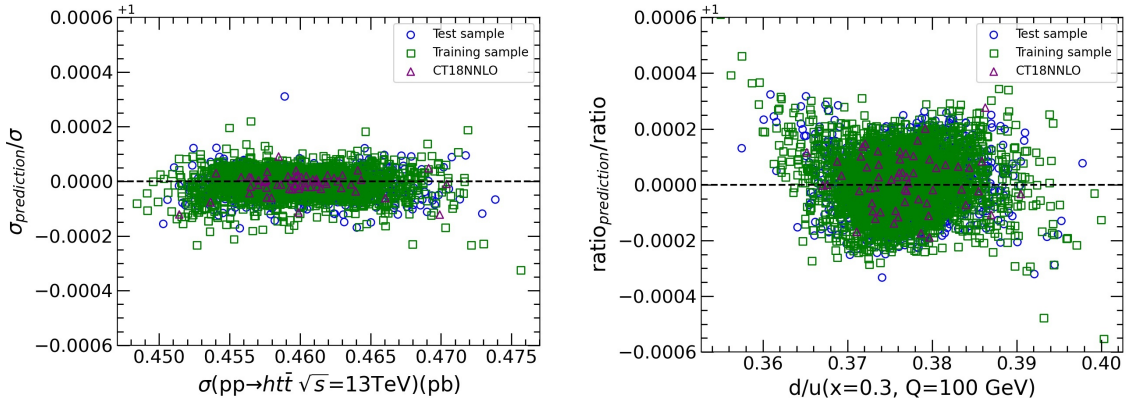


Figure 4. The variation of $\Delta\chi^2$ with d along the 1st, 6th, 11th, 16th, 22th and 28th CT18NNLO Hessian eigenvector directions.

JHEP08(2022)088



(a) (b)

Figure 5. The ratios of the predictions from NNs to the truths for the cross section of top-quark pair with a Higgs boson production in proton-proton collisions at a center of mass energy $\sqrt{s} =$ 13 TeV and the PDF ratio $d/u(x = 0.3, Q = 100\,\text{GeV})$.

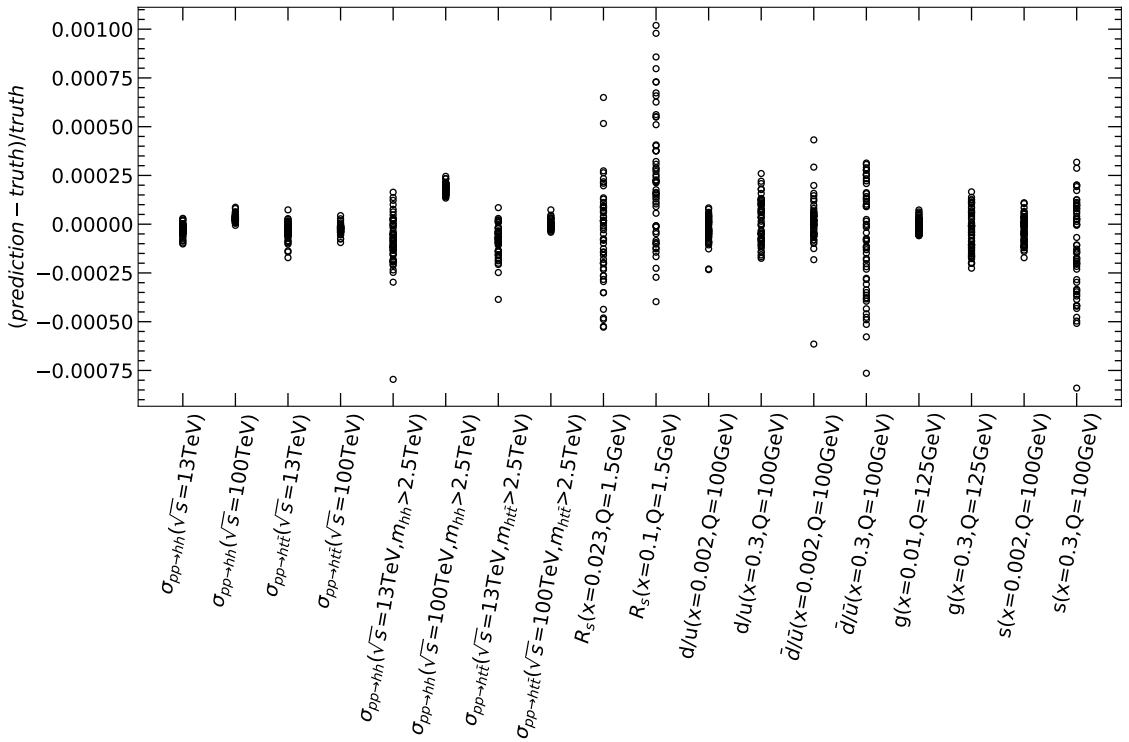


Figure 6. The relative difference between the predictions from NNs and the truths for various physics quantities. For each physics quantity, 57 marks distributed along the vertical direction correspond to the results from 57 PDFs of CT18 NNLO.

JHEP08(2022)088

4 Lagrange Multiplier scans

LM scan is a robust method to estimate PDF uncertainties, which was originally developed in refs. [23, 24]. In this method a physics quantity $X(\{a_i\})$ is introduced to the global fit as a Lagrange multiplier. Then the new function that needs to be minimized in the global fit becomes

$$\Psi(\lambda, \{a_i\}) \equiv \chi^2(\{a_i\}) + \lambda X(\{a_i\}), \tag{4.1}$$

where λ is a specified constant. For each value of λ, one can determine a set of $\{a_i\}$, $X(\{a_i\})$ and χ^2 by minimizing Ψ. Here the χ^2 corresponds to the minimum of a constrained fit with $X\{a_i\}$ fixed to the corresponding value. Specially, the central value of $X(\{a_i\})$ and the global minimum of χ^2, $\chi^2_{\rm min}$, can be determined by setting $\lambda = 0$. A parametrically defined curve (X, χ^2) can be determined by repeating the minimization for many values of λ. This means the χ^2 of the global fit depends on the value of $X(\{a_i\})$ and can be represented as $\chi^2 = \chi^2_{\rm min} + \Delta\chi^2$. The PDF uncertainty of $X(\{a_i\})$ can be determined by requiring $\Delta\chi^2 + P = T$, here T is the so-called "tolerance factor". We assume that 90% CL region corresponds to $T = 100$. The penalty term P, called Tier-2 penalty, is introduced to ensure the tolerance will be reached as soon as any data set shows disagreement at 90% CL. The detailed definition of the penalty term can be found in refs. [1, 81].

In comparison, we briefly describe the calculation of PDF uncertainties in the framework of the Hessian method. Given the physics quantity $X(\{a_i\})$, the asymmetric PDF uncertainties can be calculated as [82]

$$\begin{aligned} \delta^+ X &= \sqrt{\textstyle\sum_{i=1}^{N_d} \left[\max\left(X_{2i-1} - X_0, X_{2i} - X_0, 0\right)\right]^2}, \\ \delta^- X &= \sqrt{\textstyle\sum_{i=1}^{N_d} \left[\max\left(X_0 - X_{2i-1}, X_0 - X_{2i}, 0\right)\right]^2}, \end{aligned} \tag{4.2}$$

where X_0 represents the value of the physics quantity with the central PDF of the Hessian set, X_{2i-1} (X_{2i}) represents the value of the physics quantity with the error PDF of the Hessian set in the positive (negative) direction of the i_{th} eigenvector in the N_d-dimensional PDF parameter space.

4.1 LM scans on PDFs

We first study PDF values and ratios with LM scans based on the aforementioned NNs approximation of χ^2 and physics quantities. The results are shown in figure 7. The black and the red solid lines represent $\Delta\chi^2$ and $\Delta\chi^2 + P$ respectively. The dot and the dash lines indicate the contributions to $\Delta\chi^2$ from individual data sets. The blue and the green vertical dot-dash lines indicate the uncertainties at 90% CL determined with the LM method by requiring $\Delta\chi^2 + P = 100$ and with the Hessian method from the published CT18 NNLO PDFs, respectively. Among the generic features of the scans, it can be seen that the profile of the total $\Delta\chi^2$ and individual $\Delta\chi^2$ show almost a quadratic dependence on the variable at the neighborhood of the global minimum, which is a requirement of the Hessian method. Some individual data sets prefer PDF values or ratios that differ significantly from those at the global minimum. Besides, the HERA inclusive DIS data play important roles in all

JHEP08(2022)088

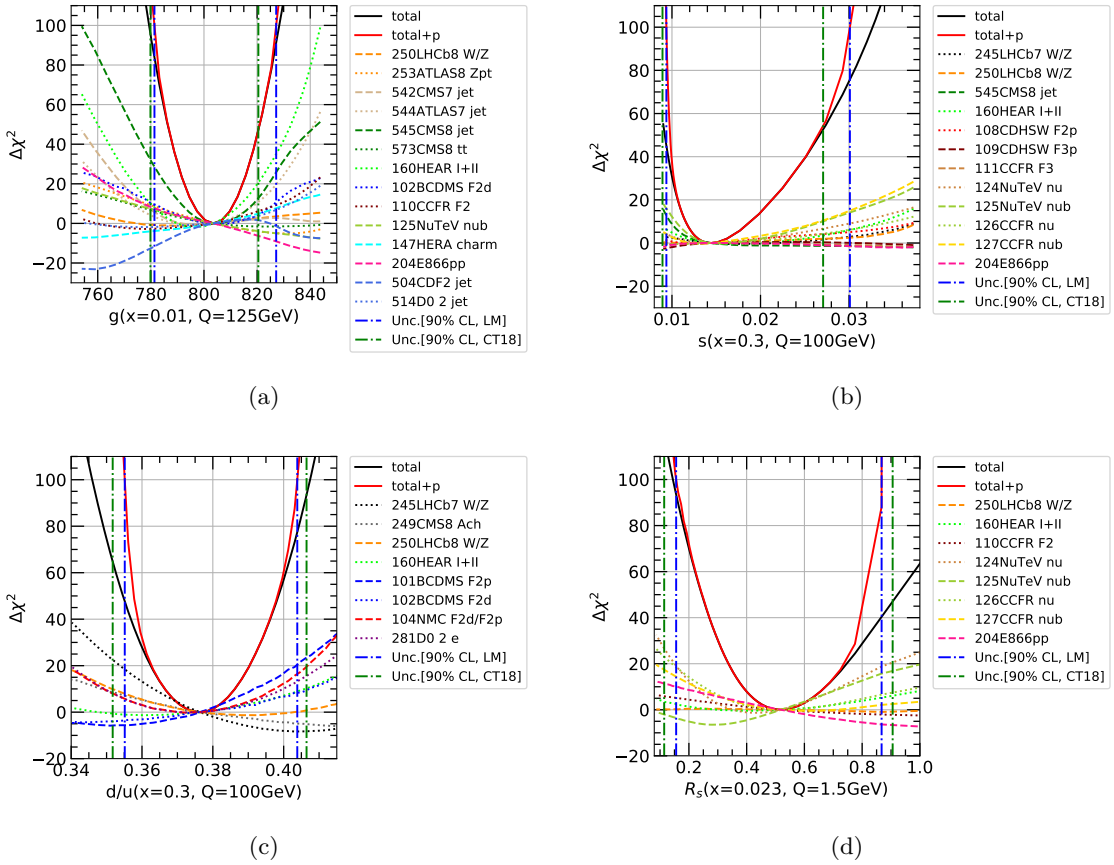


Figure 7. LM scans on the g ($x = 0.01\,\text{GeV}$, $Q = 125\,\text{GeV}$), s ($x = 0.3$, $Q = 100\,\text{GeV}$), d/u ($x = 0.3$, $Q = 100\,\text{GeV}$) and R_s ($x = 0.023$, $Q = 1.5\,\text{GeV}$). The black and the red solid lines represent $\Delta\chi^2$ and $\Delta\chi^2 + P$ respectively. The dot and the dash lines indicate the contributions to $\Delta\chi^2$ from individual data sets. The blue and the green vertical dot-dash lines indicate the uncertainties at 90% CL determined with the LM method by requiring $\Delta\chi^2 + P = 100$ and with the Hessian method from the published CT18 NNLO PDFs, respectively.

cases, which can be understood as due to the high experimental precision and the large number of data points. The penalty term also gives strong constraints on some PDF values or ratios. In addition, there are some slight differences between the uncertainties determined with the Hessian method and the LM scans, which is expected.

In the upper-left panel of figure 7, we show the results of LM scans on the gluon PDF at $Q = 125\,\text{GeV}$ and $x = 0.01$. We find that the HERA inclusive DIS data and the LHC jet data give the leading constraints. In addition, the CDF inclusive jet data (Exp. ID = 504) prefers a smaller value of the gluon PDF. At the global minimum, the χ^2 for the CDF inclusive jet data is elevated by about 20 units. The Hessian method gives a smaller PDF uncertainty than the estimation based on the LM scans.

In the upper-right panel, we show the results of LM scans on the s-quark PDF at $Q = 100\,\text{GeV}$ and $x = 0.3$. In this panel, the NuTeV and the CCFR dimuon data together with the HERA inclusive DIS data give the dominant constraints. These experimental

JHEP08(2022)088

data are consistent with the global fit. A marked deviation from the quadratic shape can be observed in the profile of the $\Delta\chi^2$. In this case, a notable difference in uncertainties manifests between the LM method and the Hessian method, and the LM method should give the more reliable result.

In the bottom-left panel of figure 7, we show the results of LM scans on the PDF ratio d/u at $Q = 100\,\text{GeV}$ and $x = 0.3$. The d/u ratio is dominantly constrained by the LHCb W and Z boson production and the fixed target experiments BCDMS and NMC. Contrasted with previous situations, the LM method gives a smaller PDF uncertainty for the d/u ratio.

In the bottom-right panel, we show the results of LM scans on the strangeness ratio R_s at $Q = 1.5\,\text{GeV}$ and $x = 0.023$. We find that the NuTeV and the CCFR dimuon data and HERA inclusive DIS data give the dominant constraints. It is also worthy noting that the NuTeV dimuon data with anti-neutrinos (Exp. ID = 125) prefers $R_s \approx 0.25$ which is smaller than the best fit value from the global fit $R_s = 0.52$ and results in a large penalty term. At the global minimum, the χ^2 for the NuTeV dimuon data with anti-neutrinos is elevated by about 7 units. The LM method predicts $R_s = 0.52^{+0.35}_{-0.36}$ at 90% CL that has smaller uncertainties than $R_s = 0.52^{+0.39}_{-0.41}$ from the Hessian method.

Above scans have also been performed in the CT18 analysis [12], and our results are consistent with those, which further proves the validity of our approach. After demonstrating the great efficiency and validity of our approach on LM scans, we are now ready to perform a systematic study on the PDF values and ratios for all flavors at a series of x values spreading over a wide range.

In figure 8, we compare the PDF uncertainties at 68% CL at $Q = 1.295\,\text{GeV}$ between the LM method and the Hessian uncertainties from the CT18 NNLO PDFs. The blue and the red solid lines represent the central values of the CT18 NNLO and the PDFs determined with the aforementioned NNs approximation of χ^2 respectively. The blue and the red hatched areas represent the uncertainties determined with the Hessian method and the LM method respectively. The results are normalized to the central value of CT18 NNLO PDFs. We find good agreements of both the uncertainties and the central values between the two methods. A notable difference, however, can be seen for u, d, $\bar{u}$, $\bar{d}$ and s-quark at small-x ($\lesssim 10^{-4}$), as well as for d, $\bar{u}$ and s-quark at large-x ($\gtrsim 0.4$). This indicates a failure of the quadratic approximation in these regions. The uncertainties from the LM method can be either larger or smaller than the uncertainties from the Hessian method depending on the flavor and the x value.

In the lower-right panel, we find that s-quark PDFs have large uncertainties for both $x \lesssim 0.001$ and $x \gtrsim 0.4$. This is because the large-x and small-x behavior of the s-quark are mostly constrained by the extrapolation of the PDF parametrization. The error band of s-quark of CT18 NNLO PDFs covers negative PDF values at $x \gtrsim 0.4$. This unphysical behavior implies a limitation of the Hessian method. On the contrary, the error band determined with the LM method is bounded above zero in all regions.

We also perform the LM scans on the general PDF ratios that is defined as

$$R_f \equiv \frac{f_i(x_1, Q)}{f_j(x_2, Q)}. \tag{4.3}$$

JHEP08(2022)088

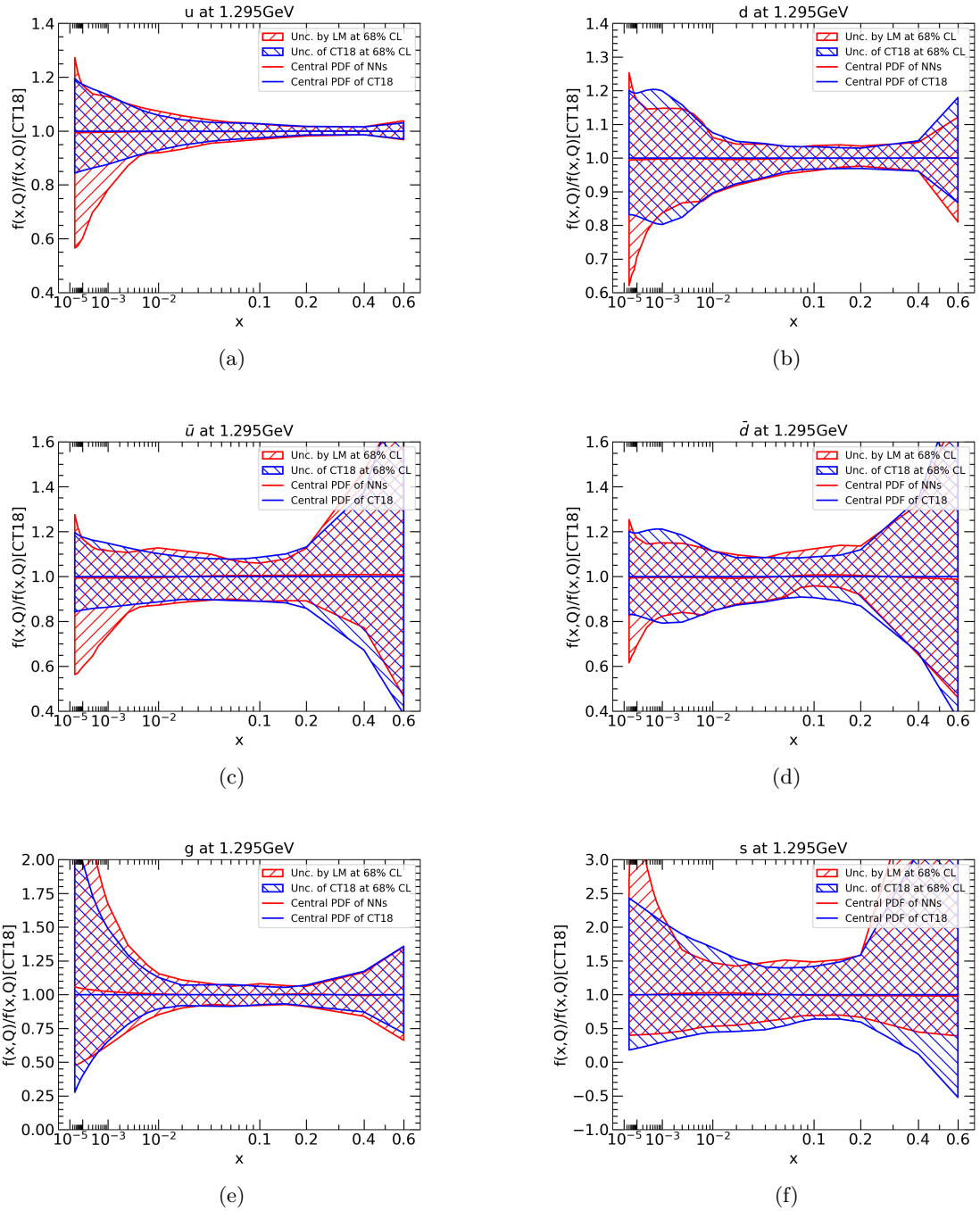
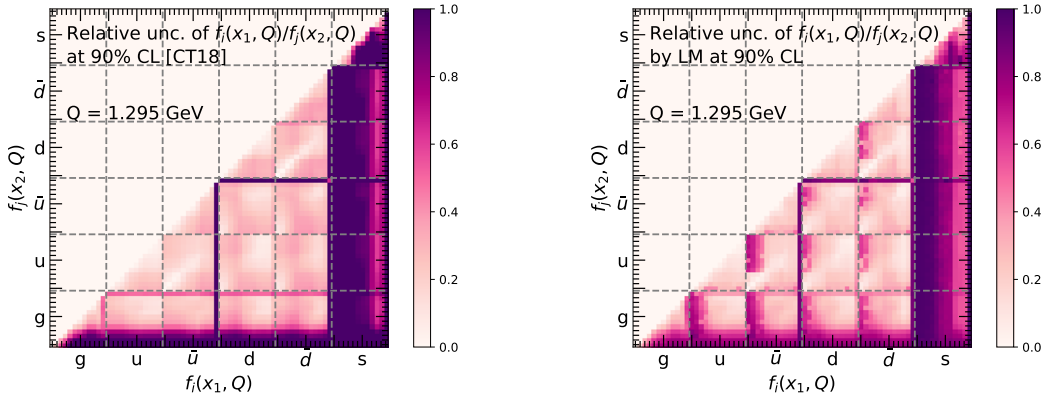


Figure 8. The parton distribution functions at $Q = 1.295\,\text{GeV}$ for u, d, $\bar{u}$, $\bar{d}$, g and s. The blue and the red solid lines represent the central values of the CT18 NNLO and the PDFs determined with the aforementioned NNs approximation of χ^2 respectively. The blue and the red hatched areas represent the uncertainties at 68% CL determined with the Hessian method and the LM scans respectively.

JHEP08(2022)088

	1	2	3	4	5	6
x	3×10^{-5}	7×10^{-5}	3×10^{-4}	7×10^{-4}	3×10^{-3}	7×10^{-3}
	7	8	9	10	11	12
x	0.01	0.02	0.06	0.1	0.2	0.6

Table 3. The x values selected for the calculation of the uncertainties of R_f.



(a) Hessian

(b) LM

Figure 9. The relative uncertainties of $R_f = f_i(x_1, Q)/f_j(x_2, Q)$ determined with the Hessian method and the LM method at 90% CL are shown in panel (a) and (b) respectively. The color code represents the relative uncertainties of the ratio R_f. The relative uncertainties are calculated at $Q = 1.295\,\text{GeV}$ with x_1 and x_2 selected among 12 values from $3{\times}10^{-5}$ to 0.6 listed in table 3, and $i, j \in \{g, u, \bar{u}, d, \bar{d}, s\}$ runs over all parton flavors.

The relative uncertainties of R_f are calculated at $Q = 1.295\,\text{GeV}$ with x_1 and x_2 selected among 12 values from $3{\times}10^{-5}$ to 0.6 listed in table 3, and $i, j \in \{g, u, \bar{u}, d, \bar{d}, s\}$ runs over all parton flavors. The results at 90% CL are shown in figure 9(a) and (b) for the Hessian and LM method respectively. The x and the y axis indicate the numerator and the denominator, and color code represents the relative uncertainties of the ratio R_f. By comparison of the two panels, we find good agreements between the uncertainties determined with the LM method and the Hessian method in most regions. Similar to figure 8, there are notable differences at small-x_1, especially for those with sea quarks in the numerator.

4.2 LM scans on cross sections

Higgs bosons are produced dominantly through gluon fusions at the LHC. The inclusive gluon-fusion cross-section has been calculated to next-to-next-to-next-to-leading order in QCD [83], which further reduces the scale variations and makes the PDF uncertainties even more important. Besides, the production of Higgs boson pair and top-quark pair associated with a Higgs boson are of equal importance for studies of the Higgs boson self-coupling and the top-quark Yukawa coupling.

JHEP08(2022)088

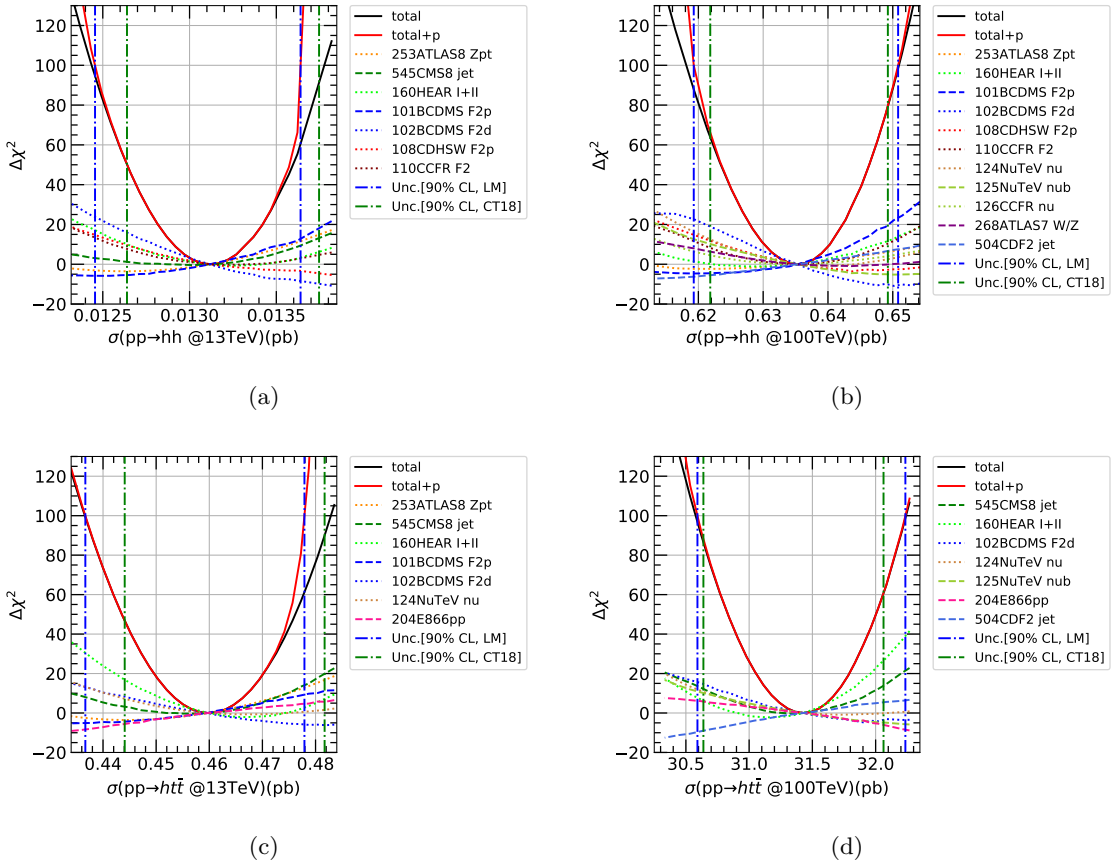


Figure 10. LM scans on the $\sigma_{pp\to hh}$ and $\sigma_{pp\to ht\bar{t}}$ at $\sqrt{s} = 13\,\text{TeV}$ or $100\,\text{TeV}$.

In figure 10 we show the results of LM scans on $\sigma_{pp\to hh}$ and $\sigma_{pp\to ht\bar{t}}$ at $\sqrt{s} = 13\,\text{TeV}$ or $100\,\text{TeV}$. For $\sigma_{pp\to hh}$ at $\sqrt{s} = 13\,\text{TeV}$, in the upper-left panel, the behaviors of χ^2 are very much similar to that shown in the upper-left panel of figure 7 for the gluon PDF. That is because the cross section of $pp \to hh$ at $13\,\text{TeV}$ is strongly correlated with the gluon PDF at $x \sim 0.02$. Constraints from HERA inclusive DIS data, BCDMS proton and deuterium data, CMS $8\,\text{TeV}$ jet data and ATLAS $8\,\text{TeV}$ Z p_T data stand out as expected. In addition, the BCDMS proton data and ATLAS $8\,\text{TeV}$ Z p_T data both prefer a larger cross section contrasted with the BCDMS deuterium data which prefers a smaller value. For $\sigma_{pp\to hh}$ at $\sqrt{s} = 100\,\text{TeV}$, in the upper-right panel, the constraints are distributed among more data sets and are related to PDFs at small-x.

The cross section of $pp \to ht\bar{t}$ mainly depends on the gluon, u-quark and d-quark PDFs. For $\sigma_{pp\to ht\bar{t}}$ at $\sqrt{s} = 13\,\text{TeV}$, in the lower-left panel of figure 10, similar behaviors of the χ^2 as $\sigma_{pp\to hh}$ are observed. At $\sqrt{s} = 100\,\text{TeV}$, the constraints from HERA inclusive DIS data predominate. In addition, constraints from NuTeV dimuon data, CMS jet data and BCDMS proton data also play important roles.

JHEP08(2022)088

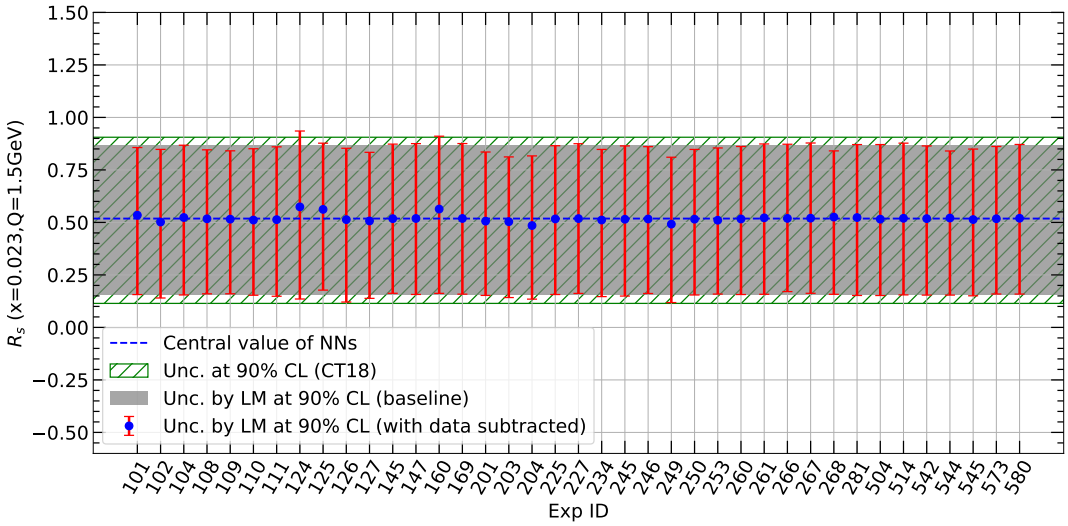


Figure 11. The results of LM scans on the R_s ($x = 0.023$, $Q = 1.5\,\text{GeV}$) with data subtracted. The horizontal axis represents the experimental data set removed from the LM scans. The blue mark and the red error bar respectively indicate the central value and uncertainties at 90% CL determined with the LM method with the rest of the data sets. The green hatched area and the gray band represent the uncertainties at 90% CL determined with the Hessian method and the LM method with the full data set respectively.

4.3 Study on impact of individual data sets

In order to assess the contribution from an individual experimental data set, we remove one data set at a time, and repeat the LM scans on physics quantities with the rest of the data sets. Difference between the fit with and without the data set can be an assessment of its contribution.

The results for R_s at $x = 0.023$ and $Q = 1.5\,\text{GeV}$ are shown in figure 11. After the removal of each data set, we find that R_s value and its uncertainty are only changed slightly, as represented by those error bars comparing to the uncertainty from LM scans with the full data set represented by the gray band. The subtraction of a single data set shows largest effects for NuTeV dimuon production data (Exp. ID = 124, 125), CCFR dimuon production data (Exp. ID = 126, 127), E866 Drell-Yan data (Exp. ID = 204) and HERA inclusive DIS data (Exp. ID = 160). In addition, the NuTeV dimuon production data and HERA inclusive DIS data prefer a smaller R_s contrasted with E866 Drell-Yan data which prefers a larger value, that is consistent with the bottom-right panel of figure 7.

In figure 12 we show the results for $\bar{d}/\bar{u}$ at $x = 0.3$ and $Q = 100\,\text{GeV}$. The E866 Drell-Yan ratio data (Exp. ID = 203) gives the dominant constraints. The fit without E866 Drell-Yan ratio data predicts a result of $\bar{d}/\bar{u} = 1.26^{+0.82}_{-0.59}$, while the fit with the full data set expects $\bar{d}/\bar{u} = 1.28^{+0.20}_{-0.33}$. After the inclusion of E866 Drell-Yan ratio data, the uncertainties of $\bar{d}/\bar{u}$ are reduced by almost 60%. That is because the penalty term of E866 Drell-Yan ratio data provides a strong constraint on $\bar{d}/\bar{u}$. In addition, constraints from NMC deuteron data (Exp. ID = 104) and HERA inclusive DIS data also play important roles.

JHEP08(2022)088

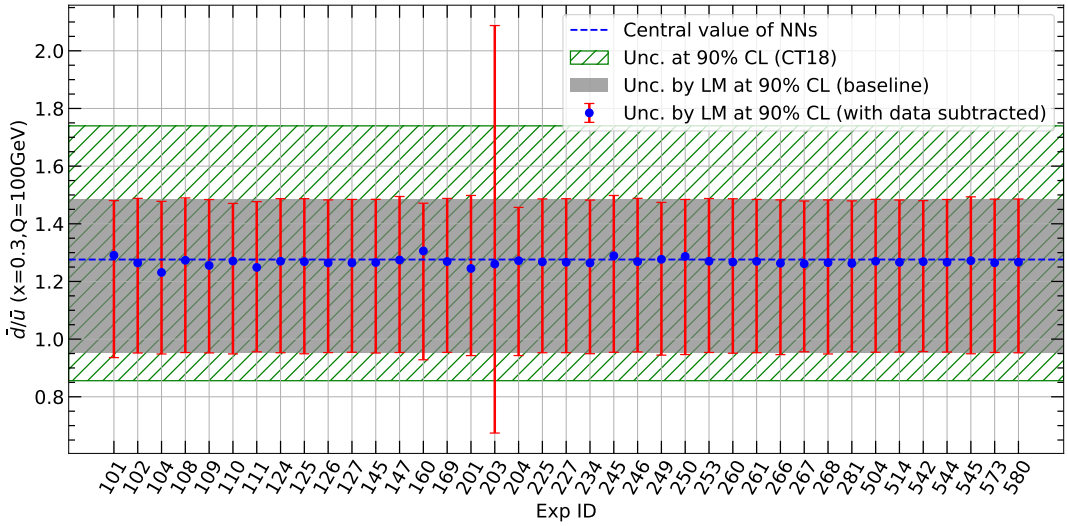


Figure 12. The same as figure 11, but for the results of LM scans on the $\bar{d}/\bar{u}$ ($x = 0.3$ and $Q = 100\,\text{GeV}$).

In figure 13 we show the results for $\sigma_{pp\to hh}$ at $\sqrt{s} = 13\,\text{TeV}$. The constraints from HERA inclusive DIS data predominate as expected. In addition to that, constraints from BCDMS proton and deuterium data (Exp. ID = 101, 102) and ATLAS 8 TeV Z p_T data (Exp. ID = 253) also play important roles. The fit without HERA inclusive DIS data expects $\sigma_{pp\to hh} = 0.0129^{+0.0007}_{-0.0009}$ pb, while the fit with the full data set gives $\sigma_{pp\to hh} = 0.0131^{+0.0005}_{-0.0007}$ pb. An upward shift of about 2×10^{-4} pb is observed when we incorporate HERA inclusive DIS data, and the uncertainties of $\sigma_{pp\to hh}$ are reduced by almost 20%. In addition, the HERA inclusive DIS data and BCDMS deuterium data both prefer a larger $\sigma_{pp\to hh}$ contrasted with BCDMS proton data and ATLAS Z p_T data which prefer a smaller value, that is consistent with the upper-left panel of figure 10.

4.4 Two-dimensional LM scans

Besides the PDF uncertainties, it is possible to quantify other statistical estimators such as the correlation between two physics quantities with two-dimensional LM (2-D LM) scans. That can be achieved by adding a second physics quantity into eq. (4.1). The new function that needs to minimized in the global fit becomes

$$\Psi\left(\lambda_1, \lambda_2, \{a_i\}\right) \equiv \chi^2\left(\{a_i\}\right) + \lambda_1 X_1\left(\{a_i\}\right) + \lambda_2 X_2\left(\{a_i\}\right), \tag{4.4}$$

where λ_1 and λ_2 are specified constants, and $X_1(\{a_i\})$ and $X_2(\{a_i\})$ represent the two physics quantities of interest. Similar to eq. (4.1), the constrained minimum of χ^2 from the global fit depends on X_1 and X_2, and can be written as $\chi^2 = \chi^2_{\text{min}} + \Delta\chi^2$, where $\chi^2_{\text{min}} = \chi^2(\lambda_1 = 0, \lambda_2 = 0)$. The contour of $\Delta\chi^2 + P$ in the plane of X_1 vs. X_2 can be an assessment of the correlation between X_1 and X_2.

As examples in figure 14 we show contours for $\bar{d}/\bar{u}$ ($x = 0.3$, $Q = 100\,\text{GeV}$) vs. R_s ($x = 0.023$, $Q = 1.5\,\text{GeV}$) and $\sigma_{pp\to hh}$ ($\sqrt{s} = 13\,\text{TeV}$) vs. $\sigma_{pp\to hh}$ ($\sqrt{s} = 100\,\text{TeV}$) determined

JHEP08(2022)088

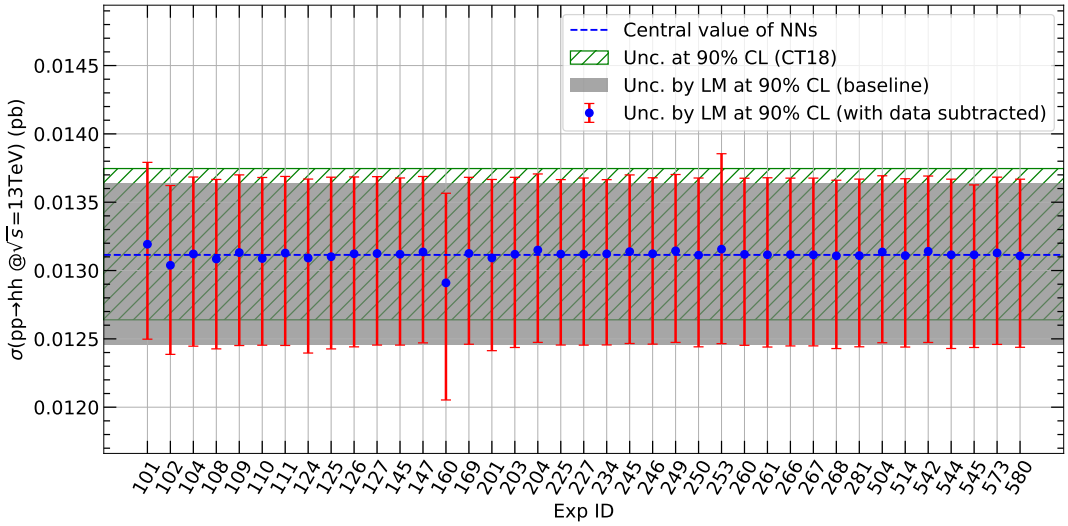
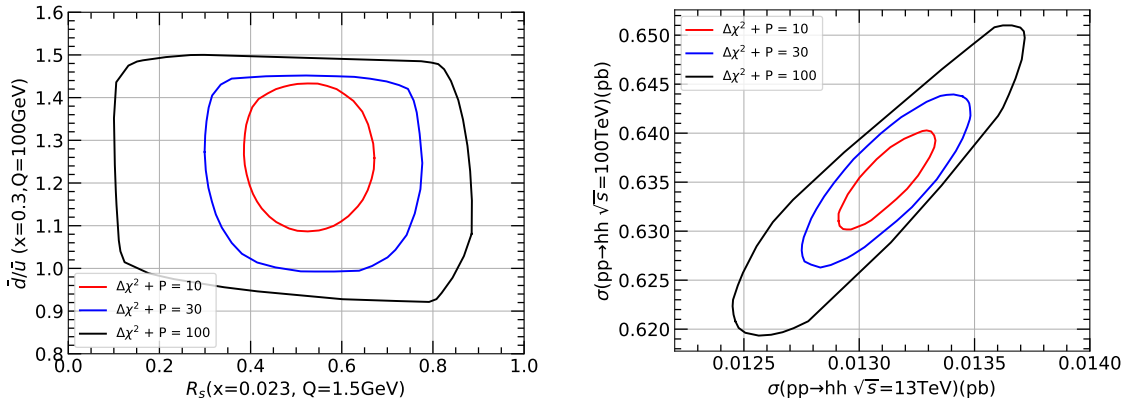


Figure 13. The same as figure 11, but for the results of LM scans on the $\sigma_{pp\to hh}$ at $\sqrt{s} = 13\,\text{TeV}$.



(a) (b)

Figure 14. Contour plot of $\Delta\chi^2$ plus Tier-2 penalty term on the plane of $\bar{d}/\bar{u}$ ($x = 0.3$, $Q = 100\,\text{GeV}$) vs. R_s ($x = 0.023$, $Q = 1.5\,\text{GeV}$) and $\sigma_{pp\to hh}$ ($\sqrt{s} = 13\,\text{TeV}$) vs. $\sigma_{pp\to hh}$ ($\sqrt{s} = 100\,\text{TeV}$).

with the 2-D LM scans. In the left panel, a weak correlation between strangeness ratio R_s and $\bar{u}/\bar{d}$ ratio is observed. That is because the two quantities are dominantly constrained by different experimental data sets. At small $\Delta\chi^2 + P$ the contour shows an elliptic shape. When $\Delta\chi^2 + P$ gets larger, the shape of the contour becomes irregular due to the increase of penalty term contributions. On the contrary, the right panel demonstrates a strong correlation between $\sigma_{pp\to hh}$ ($\sqrt{s} = 13\,\text{TeV}$) and $\sigma_{pp\to hh}$ ($\sqrt{s} = 100\,\text{TeV}$) since both processes are sensitive to gluon PDFs and constrained by the relevant experimental data sets.

JHEP08(2022)088

5 Applications

In this section, we evaluate the impact of the NOMAD measurements and of two pseudo-data sets of HL-LHC on PDFs based on the new approach. In addition, we study constraints on the new physics with a joint fit of both PDFs and the Wilson coefficient of lepton-quark contact interactions in the framework of the SMEFT.

5.1 Constraint from NOMAD data

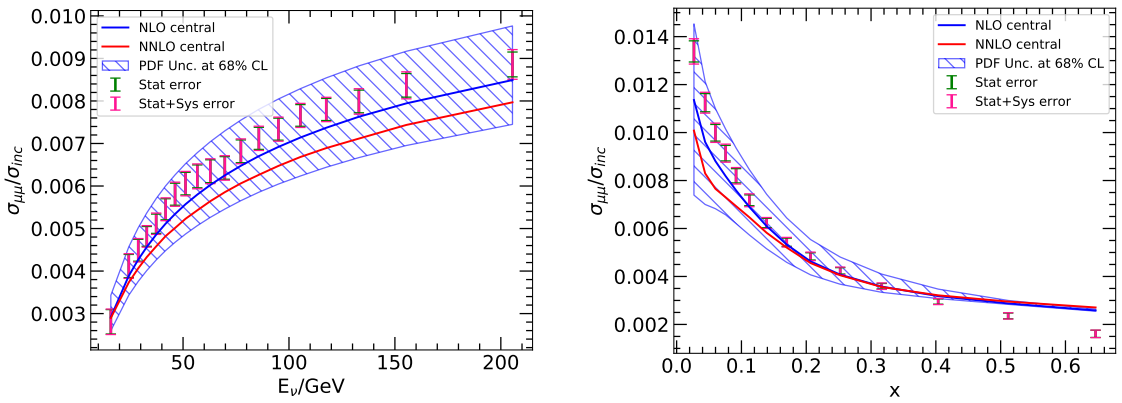
The charm-quark production in CCDIS process provides a unique sensitivity to the strange-quark distribution in the nucleon, with a clean signal of two muons with opposite charges in the final state. Recently, NOMAD collaboration reported a measurement of dimuon production in the neutrino-iron scattering experiment [31]. A sample of about 9×10^6 inclusive CCDIS events, including 15344 dimuon events, is collected, providing a reduced statistical uncertainty. Observables are taken to be the ratios of dimuon to inclusive cross-sections, which provides a large cancellation of the common systematic uncertainties presented in both the numerator and the denominator. Final results are distributed among three differential variables: the reconstructed neutrino energy E_ν, the Bjorken x and the partonic center of mass energy $\sqrt{\hat{s}}$. By the supplement of data from NOMAD, the improvement in the constraint on s-quark PDFs are studied in this section using the same NNs approach on χ^2 mentioned in previous sections.

On the theoretical side, structure functions in S-ACOT-χ general mass scheme up to NNLO are constructed, so that a full consideration of the charm-quark mass is included [84–86]. Predictions of inclusive CCDIS and open charm production cross-sections are made from these constructions, and dimuon cross sections are derived by further applying the inclusive decay branching ratio of charm quark to muon. The significant uncertainties of the decay branching ratio contribute as one of the dominant systematic errors on the dimuon cross sections, which are summarized in appendix C.

In figure 15 we show comparison of NOMAD data and our predictions at both NLO and NNLO, as well as the Hessian PDF uncertainties at 68% CL for distributions over E_ν or x. The PDF uncertainties can be as large as 10% in most regions. This directly comes from the large uncertainties of the predictions of dimuon cross-sections, and can be further traced back to the poor knowledge about s-quark PDFs. In both distributions, most of the data points are consistent with our predictions, while a significant deviation can be found in the last two points of the distribution over Bjorken x. That can be due to the modeling of heavy nuclear corrections used in the experimental analysis. We will discard those two data points when including NOMAD data in our later global fit. It is also noted that the inclusion of NOMAD data to the global fit can improve the consistency with almost no cost of tension with the other data sets [87, 88]. Most of the data lie above our NLO predictions of central values. Given this fact, an increased s-quark PDF is expected after the inclusion of NOMAD data, and this increase gets larger due to the negative corrections from NNLO.

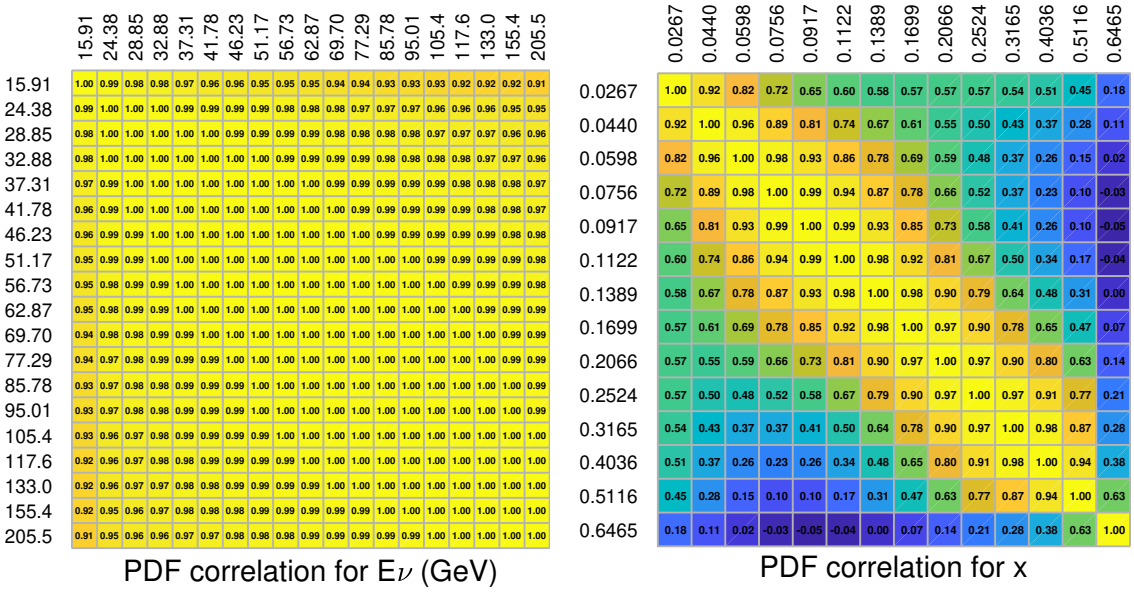
As mentioned earlier, NOMAD presents measurements on three distributions. The different sensitivities of distributions over E_ν or x are illustrated in figure 16. It shows the PDF induced correlations among bins of each distributions calculated with CT18 NNLO

JHEP08(2022)088



(a) (b)

Figure 15. NLO (blue line) and NNLO (red line) predictions for ratios of dimuon to CCDIS inclusive differential cross-sections with respect to neutrino energy (panel a) and Bjorken x (panel. b). The blue hatched areas represent the Hessian PDF uncertainties of the NLO predictions at 68% CL. NOMAD data are also shown with statistical uncertainties and the combination of both statistical and systematic uncertainties.



(a) (b)

Figure 16. PDF induced correlations between theory predictions for different experimental bins, for NOMAD distribution in neutrino energy (a) and in Bjorken-x (b), calculated with CT18 NNLO Hessian PDF set. Numbers in the axis represent the center of each bin, and numbers in the table represent the correlation cosine for each pair of bins.

JHEP08(2022)088

Hessian PDFs. We find that for E_ν distribution, all data points are strongly correlated, and similar results are found for $\sqrt{\hat{s}}$ distribution which is not shown here. Both of them only impose constraints on the overall normalization of the s-quark distribution. Thus their constraints are diluted due to the systematic errors on the inclusive branching ratio of charm quark to muon (0.094±0.01). On the other hand, the correlation pattern is nontrivial for x distribution which imposes further constraints on the shape of s-quark PDFs. We can not simply combine all these distributions from NOMAD data due to the lack of public statistical correlation between these distributions. Hence in the following, only the x distribution is included in our global analysis.

In figure 17, we compare u, $\bar{u}$, $\bar{d}$ and s-quark PDFs at $Q = 1.295\,\text{GeV}$ from fits with and without the inclusion of NOMAD data. The PDF uncertainties are shown through hatched areas with relevant colors. NOMAD data are taken from the distribution over Bjorken x excluding the last two points, with predictions calculated up to NNLO in QCD. Predictions for data sets 124-127 (dimuon measurements from NuTeV and CCFR) in the global fit are replaced with their NNLO versions when including NOMAD data, in order to match on the theoretical precision. Note in the fit without NOMAD data the predictions for data sets 124-127 are evaluated at NLO similar to those in CT18. All PDFs are normalized to the central value without NOMAD data in figure 17. In the upper-left panel, almost no change occurs in the region $x \gtrsim 0.1$ of u-quark PDF, and a negligible downward shift smaller than 2% can be seen for $x \lesssim 0.05$. Slight downward shifts on both central value and uncertainty region can also be observed in the $\bar{u}$ (upper-right panel) and the $\bar{d}$-quark (lower-left panel) PDFs. The downward shifts observed are required to stabilize the W and Z production cross sections at collider experiments. The improvement in the constraints on u, $\bar{u}$ and $\bar{d}$-quark PDFs are smaller than about 3%. This insensitivity of u, $\bar{u}$ and d-quark PDFs to NOMAD data is an indication of the CKM suppression in the charm-quark production. The constraint on s-quark PDF is, however, markedly improved around $x = 0.05$. In the region of $x \sim 0.05$, the s-quark PDF achieves a factor of two better precision when NOMAD data are incorporated. This is because NOMAD data peak at neutrino energy $E_\nu \approx 30\,\text{GeV}$, which implies a sensitivity to kinematic region with Bjorken $x \sim 1/(1 + 2M_{\text{nucleon}}E_\nu/Q^2) \sim 0.03$ at $Q = 1.295\,\text{GeV}$. An upward shift of more than 15% is also observed in most regions. It is indeed a manifestation of the trend of prediction-data comparison shown in figure 15.

Both ABM and NNPDF groups considered the impact of NOMAD data [87, 88]. As to the analysis of ABM group, an at most 5% downward shift is reported near region $x \approx 0.05$ at scale $Q = 3\,\text{GeV}$ when NOMAD data are incorporated into the fit with only NuTeV/CCFR data (data sets 124-127 in this paper) [87]. More data sets are considered in the work of NNPDF group [88]. With the analysis performed there, NOMAD data together with ATLAS W/Z data sets [89, 90] contribute to a marked enhancement of s-quark PDF at $Q = 10\,\text{GeV}$ compared with CT18 data sets. It is noted that, between these two kinds of data sets, ATLAS W/Z data sets are already reported to give a larger s-quark PDF compared with CT18 data sets [12], and the work of NNPDF group further demonstrated that ATLAS W/Z data sets prefer a larger s-quark PDF compared with NOMAD data. Finally, both the two groups and our analysis indicate strong constraints on s-quark PDF in the region near $x \approx 0.05$, given the incorporation of NOMAD data.

JHEP08(2022)088

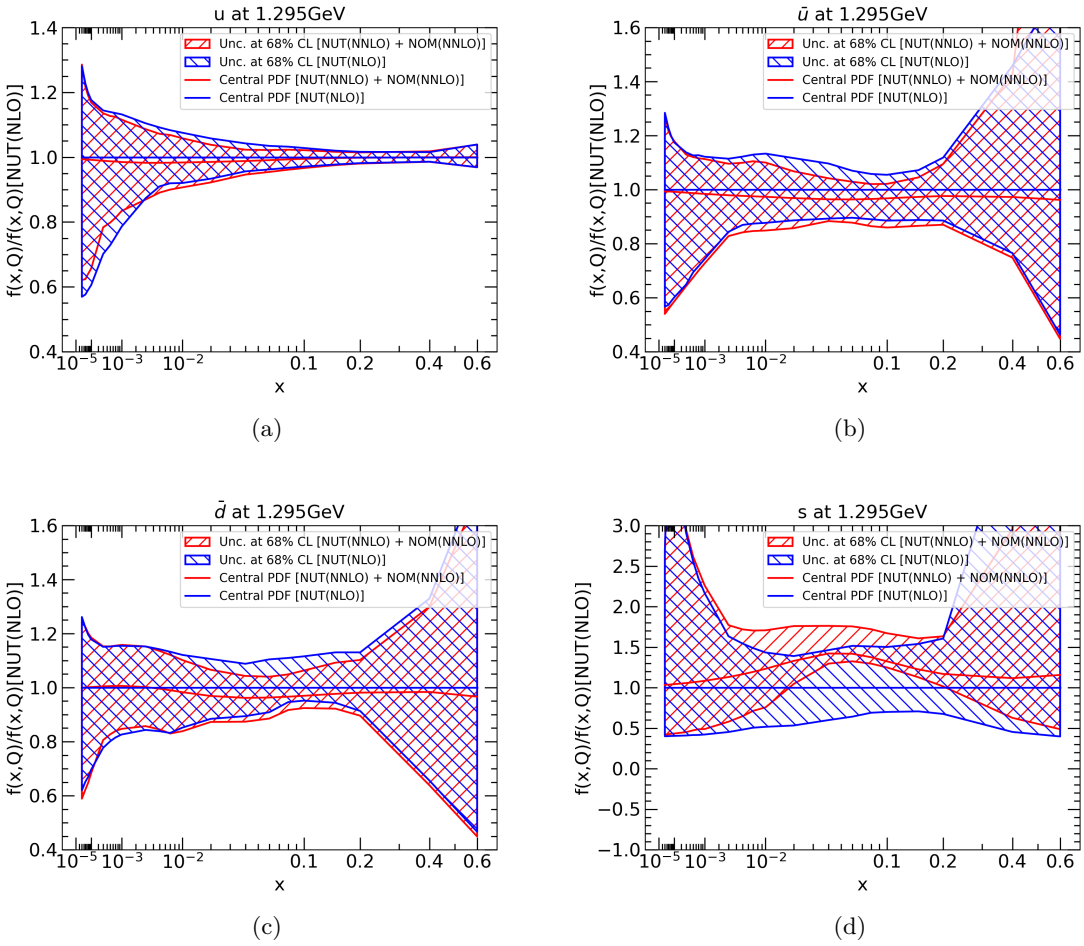


Figure 17. The parton distribution functions at $Q = 1.295\,\text{GeV}$ for u, $\bar{u}$, $\bar{d}$, and s. The red and the blue solid lines represent the central values with and without NNLO NOMAD data respectively, and the red and the blue hatched areas represent the respect uncertainties at 68% CL. When NOMAD (NOM) data are incorporated, 124-127 data sets (NUT) are replaced with their NNLO version. Central values are normalized to the NLO version.

We also compare the sensitivities to s-quark PDF between NOMAD data and the other data sets in figure 18, in which we show LM scans on s-quark PDF and R_s. This comparison is set at a scale of $Q = 1.5\,\text{GeV}$ and $x = 0.1$ in panel (a). In this panel, NOMAD data predominate over the other experimental data sets. When x gets smaller to be 0.023 as in panel (b), NuTeV and CCFR neutrino DIS experiments become more important but still the NOMAD data show the most prominence. In panel (a), the fit without NOMAD data predicts $R_s(x = 0.1, Q = 1.5\text{GeV}) = 0.40^{+0.35}_{-0.20}$ at 90% CL, while fit including NOMAD data expects $R_s(x = 0.1, Q = 1.5\text{GeV}) = 0.54^{+0.24}_{-0.06}$, giving improved constraints by a factor of two. In panel (b), the corresponding values are $R_s(x = 0.023, Q = 1.5\text{GeV}) = 0.53^{+0.33}_{-0.38}$ and $R_s(x = 0.023, Q = 1.5\text{GeV}) = 0.70^{+0.40}_{-0.17}$, respectively. NOMAD data hence give about 20% reduction on PDF uncertainties. It is also noted that a slight tension exists between NOMAD data and data from the other two neutrino DIS experiments, i.e., NuTeV and

JHEP08(2022)088

CCFR, in both panels. The latter two experiments both prefer smaller R_ss contrasted with NOMAD data which prefer a larger one. Further investigations on the interplay of the three experiments and of different theories are included in appendix C. Moreover, the ATLAS W/Z data [89], which prefer an especially larger $R_s(x = 0.023, Q = 1.38\text{GeV}) \sim 1$, show an even stronger tension with these two neutrino DIS experiments. NOMAD data, however, compromise between these two extremes. This conclusion is also observed in the analysis of [88]. Meanwhile, a similar result of $R_s(x = 0.023, Q = 1.6\text{GeV}) = 0.71 \pm 0.1$ is obtained in that work once NOMAD data are included.

On the other hand, we let the scale increase to be $Q = 100\,\text{GeV}$ in panel (c) and panel (d). The case with $x = 0.3$ shows more sensitive than that with $x = 0.002$. In panel (d), it can be seen that NuTeV and CCFR data become comparable with NOMAD data. No significant shift in the central value is found when we incorporate NOMAD data, but an almost 30% better constraints on s-quark PDF is achieved. In panel (c), the sensitivity of NOMAD data becomes worse due to the favor of large-x at this scale, and collider data now play important roles. Only improvement of a few percent in the constraint on s-quark PDF can be obtained.

5.2 Impact of High-luminosity LHC

LHC data play important roles on constraining PDFs as shown in table 2. And the upgrade of the LHC, the HL-LHC, is expected to accumulate a total integrated luminosity of $\mathcal{L} = 3000\,\text{fb}^{-1}$ for ATLAS and CMS and $300\,\text{fb}^{-1}$ for LHCb. In this section, we take two of those HL-LHC pseudo-data sets constructed in ref. [33], and evaluate their impact on PDFs within the framework of CT18 based on our new approach.

The HL-LHC pseudo-data are generated for processes of Drell-Yan production with high dilepton invariant mass and W and Z boson production in the forward region. Details of these pseudo-data are described as follows:

- The distribution of dilepton invariant mass $d\sigma(pp \to l^+l^-)/dm_{ll}$ of high-mass Drell-Yan process at $\sqrt{s}$ =14 TeV, covered by the ATLAS experiment, is generated according to the following requirements: $p_T^{l_{1(2)}} \geq 40$ (30) GeV, $|\eta^l| \leq 2.5$, and $m_{ll} \geq 116\,\text{GeV}$. The total number of data points is 21. The binning and the systematic uncertainties are determined from refs. [33, 91].

- The distributions for W and Z boson production in the forward region at $\sqrt{s}$ =14 TeV, covered by the LHCb experiment, are generated according to the following cuts: $p_T^l \geq 20\,\text{GeV}$, $2.0 \leq \eta^l \leq 4.5$. An additional requirement for Z production is that $60\,\text{GeV} \leq m_{ll} \leq 120\,\text{GeV}$. The total number of data points is 90. The binning and the systematic uncertainties are determined from refs. [33, 74].

We include those pseudo-data in the CT18 global fit and quantify their impact on PDFs. In figure 19 we show a comparison of the PDFs with and without HL-LHC pseudo-data, together with the published Hessian set of CT18. All results are normalized to the central value of CT18. The PDF uncertainties are shown through hatched areas with relevant colors. In figure 19, a significant reduction in PDF uncertainties can be found in all cases

JHEP08(2022)088

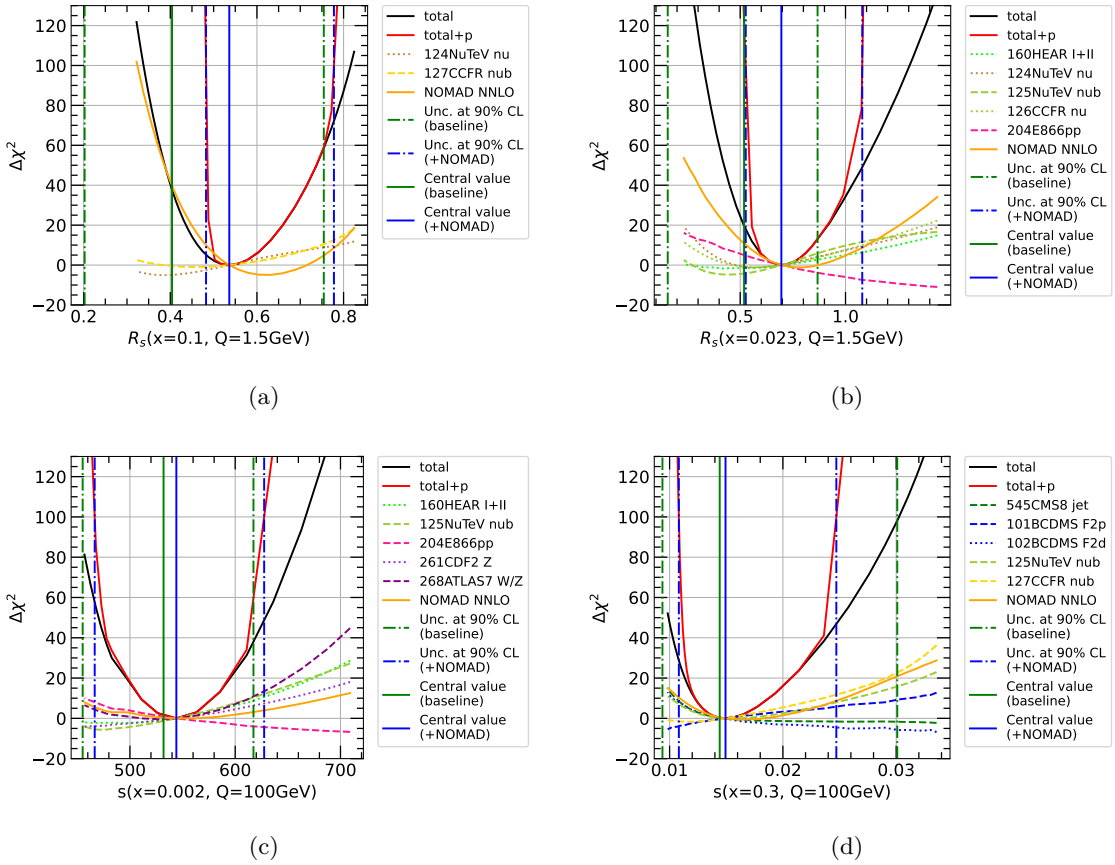


Figure 18. LM scans on the R_s at $Q = 1.5\,\text{GeV}$ and $x = 0.1$ or 0.023 (upper panels), and LM scans on the s-quark at $Q = 100\,\text{GeV}$ and $x = 0.002$ or 0.3 (lower panels). The blue and the green vertical solid (dot-dash) lines represent the central values (uncertainties) with and without NOMAD data respectively.

once including the pseudo-data, especially for sea quarks. In the upper-left panel, the PDF uncertainties are reduced by almost a factor of 2, from about 30% to about 15%, at small-x. Similar improvements can also be observed in d, $\bar{u}$ and s-quark PDFs. That is because HL-LHC pseudo-data contribute a great improvement in statistics, and cover the kinematic regions where PDFs are not determined well. Specifically, the process of high-mass Drell-Yan is directly sensitive to sea quarks at large-x, and the process of forward W/Z production constrains the s-quark PDF at both small-x and large-x. In the lower-right panel, we find that the HL-LHC gives about 30% reduction on PDF uncertainties of s-quark in the regions of $x \sim 0.01$ and $x \sim 0.1$. This result highlights the importance of the process of forward W/Z production.

We show the results of LM scans on R_s and u/d ratio in figure 20. We find measurements of high-mass Drell-Yan process and forward W/Z production process at HL-LHC give strong constraints on R_s and u/d ratio at both small-x and large-x. The PDF uncertainties of R_s and u/d are significantly reduced after the inclusion of pseudo-data.

JHEP08(2022)088

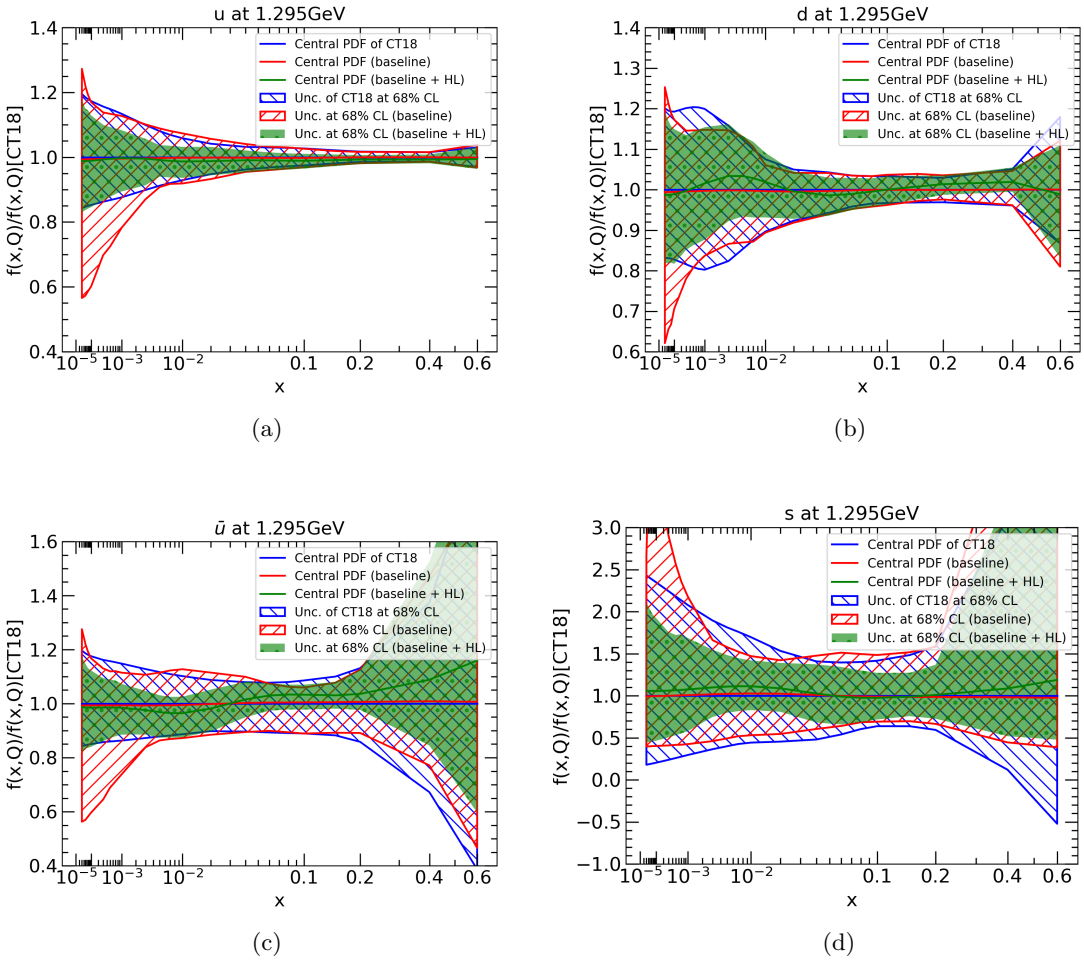


Figure 19. The parton distribution functions at $Q = 1.295\,\text{GeV}$ for u, d, $\bar{u}$ and s. The red and the green solid lines represent the central values without and with HL-LHC pseudo-data respectively, and the red and the green hatched areas represent the respective uncertainties at 68% CL. The results are normalized to the central value of CT18 NNLO (blue solid line).

In the upper-left panel of figure 20 for R_s at $x = 0.023$ and $Q = 1.5\,\text{GeV}$, the constraints from HL-LHC pseudo-data predominate as expected. In addition to that, constraints from NuTeV dimuon, CCFR dimuon and HERA inclusive DIS data also play important roles. The fit without pseudo-data predicts a result of $R_s = 0.53^{+0.33}_{-0.38}$ at 90% CL, while fit including pseudo-data gives $R_s = 0.54^{+0.22}_{-0.19}$. After the inclusion of pseudo-data, the PDF uncertainties are reduced by almost 50%. As x increases to 0.1 in the upper-right panel, HL-LHC forward W/Z data becomes more important. Fit without pseudo-data gives a result of $R_s = 0.40^{+0.35}_{-0.20}$, while fit including pseudo-data gives $R_s = 0.39^{+0.16}_{-0.16}$.

For d/u at $x = 0.002$ and $Q = 100\,\text{GeV}$, in the lower-left panel, the most strong constraints originate from HL-LHC pseudo-data together with LHC W and Z boson data and the fixed target experiments E866 and NMC. The results of d/u are $0.946^{+0.038}_{-0.032}$ and $0.954^{+0.026}_{-0.032}$ corresponding to fit without and with pseudo-data respectively. After

JHEP08(2022)088

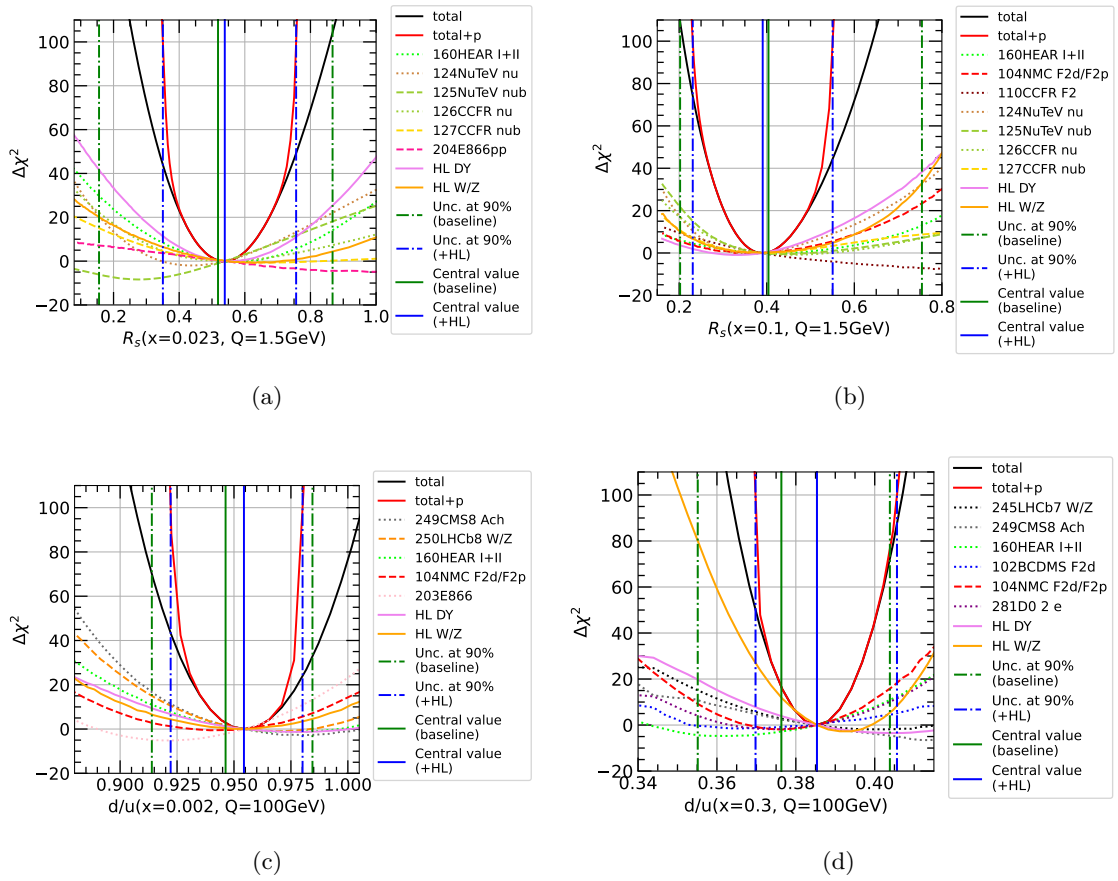


Figure 20. LM scans on the R_s at $Q = 1.5\,\text{GeV}$ and $x = 0.023$ or 0.1 (upper panels), and LM scans on the d/u at $Q = 100\,\text{GeV}$ and $x = 0.002$ or 0.3 (lower panels). The blue and the green vertical solid (dot-dash) lines represent the central values (uncertainties) with and without HL-LHC pseudo-data respectively.

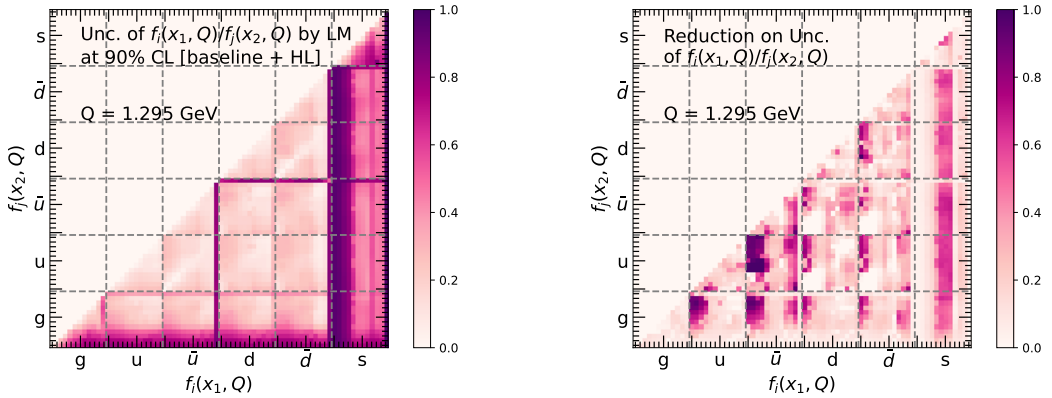
the inclusion of pseudo-data, the PDF uncertainties are reduced by almost 30%. In the lower-right panel for $x = 0.3$ and $Q = 100\,\text{GeV}$, pseudo-data predominate over the other experimental data sets. Fit without and with pseudo-data give $d/u = 0.376^{+0.028}_{-0.021}$, and $d/u = 0.385^{+0.020}_{-0.016}$ respectively. PDF uncertainties are reduced by almost 25% in this case. Both of the two HL-LHC processes prefer a larger d/u, and their inclusion leads to an increase of the central value.

In figure 21, we show the results for the general PDF ratio R_f as defined in eq. (4.3). The uncertainties of R_f are also determined with the LM method. In panel (a), we show the relative uncertainties of R_f at 90% CL, ΔR_f, from the fit with inclusion of HL-LHC pseudo-data. To compare with the results from the fit without HL-LHC pseudo-data, a reduction factor of ΔR_f,

$$y_{\text{red}} = 2\frac{\Delta R_f^{\text{base}} - \Delta R_f^{\text{base}+HL}}{\Delta R_f^{\text{base}} + \Delta R_f^{\text{base}+HL}}, \tag{5.1}$$

is shown in panel (b). We find that the relative uncertainties of R_f have a noticeable

JHEP08(2022)088



(a) CT18NNLO + HL-LHC

(b) Reduction

Figure 21. The relative uncertainties of $R_f = f_i(x_1, Q)/f_j(x_2, Q)$ determined with the LM method at 90% CL are shown in panel (a), where $Q = 1.295\,\text{GeV}$. In panel (b) we show the reduction factors on the relative uncertainties of R_f between with and without HL-LHC pseudo-data.

reduction in general. For the case of the u and $\bar{u}$-quark PDFs as the numerator, we find that the HL-LHC gives about 80% reduction on relative uncertainties in the region of $x_1 \lesssim 0.001$, which is because the HL-LHC pseudo-data give strong constraints on u and $\bar{u}$-quark PDFs in this region as shown in figure 19. In addition, for the case of the s-quark PDFs as the numerator, we find that the HL-LHC gives about 50% reduction on relative uncertainties in the region of $x_1 \sim 0.01$. However, for the $f_d(x_1, Q)/f_u(x_2, Q)$, $f_d(x_1, Q)/f_{\bar{u}}(x_2, Q)$, $f_{\bar{d}}(x_1, Q)/f_u(x_2, Q)$ and $f_{\bar{d}}(x_1, Q)/f_{\bar{u}}(x_2, Q)$, we find that the reduction factors on relative uncertainties are minor in the region of $x_1 \lesssim 1 \times 10^{-3}$ and $x_2 \lesssim 1 \times 10^{-3}$. That is because the correlation between u-quark PDFs and d-quark PDFs at small-x that originates from the parametrization form of PDFs. Besides, for the ratios of gluon PDFs $f_g(x_1, Q)/f_g(x_2, Q)$, the uncertainties are reduced by only a few percent, which is expected due to the weak correlations between the two HL-LHC processes and gluon PDFs.

5.3 Constraint on new physics with the global fit

PDFs and their uncertainties play important roles in the indirect searches for new physics beyond the SM. In this case, the scale of the new physics can be well beyond the typical scale of hard scatterings, and its effects can be formally described in the framework of the SMEFT. PDFs are determined by fitting to a variety of experimental data under the assumption of the SM. This leads to a problem that the degeneracy of PDF variations and the new physics contributions cannot be identified. Therefore, to assess and furthermore constrain the new physics, a joint global fit including both PDFs and model parameters of new physics should be performed. In this paper, we only consider one dimension-six operator, namely the lepton-quark contact interactions, to model the BSM effects in the

JHEP08(2022)088

SMEFT framework,

$$\mathcal{L}_{\text{SMEFT}} = \mathcal{L}_{\text{SM}} + \sum_{i,j} \frac{c_{ij}}{\Lambda^2}(\bar{q}_i\gamma_\mu q_i)(\bar{l}_j\gamma^\mu l_j)$$
$$= \mathcal{L}_{\text{SM}} + \frac{\tilde{c}}{\Lambda^2}\sum_{i,j} e_{q_i}e_{l_j}(\bar{q}_i\gamma_\mu q_i)(\bar{l}_j\gamma^\mu l_j), \tag{5.2}$$

where c_{ij} is the Wilson coefficient, l_j and q_i represent fields of charged leptons and quarks of flavor j and i respectively, and $e_{q_i(l_j)}$ are the corresponding electric charges. We assume the new interactions being vector-current type and have a flavor structure similar to the QED coupling for simplicity. Thus the contributions from the new physics are parametrized by a single variable of the effective Wilson coefficient $\tilde{c}$ that is normalized to the QED coupling.

In the case of data sets of the CT18 global fit, the DIS and the Drell-Yan processes receive contributions from this operator. Processes with relatively large Q^2 are especially sensitive to BSM effects, where Q is the momentum transfer. Most of the data of Drell-Yan process included in the CT18 analysis are collected near the Z-pole region, which is less sensitive to new physics. Hence, we only consider the HERA DIS process due to its large Q^2. The amplitude of SM contributions from QED interactions is proportional to $1/Q^2$, and the amplitude of the BSM contributions is proportional to $\tilde{c}/\Lambda^2$.[1] Hence, the total cross section including the BSM effects can be written as:

$$\sigma_{\text{total}} = (1 + \frac{\tilde{c}}{\Lambda^2}Q^2)^2 \times \sigma_{\text{DIS}}. \tag{5.3}$$

A new NNs is built by adding the parameter $\tilde{c}/\Lambda^2$ into the input layer. An association between the 29 variables $\{a_i, \tilde{c}/\Lambda^2\}$ and χ^2 is constructed. With the new NNs, χ^2 is recalculated and the results of LM scans on $\tilde{c}/\Lambda^2$ are shown in figure 22. HERA inclusive DIS data give the dominant constraints as expected. The LM scans predict a result of $\tilde{c}/\Lambda^2 = 0.56^{+9.16}_{-9.16}$ TeV^{-2} at 90% CL, which is consistent with the SM. The interplay between PDFs and BSM effects in the framework of the SMEFT has been studied in previous works [35, 36, 39]. A simultaneous determination of the PDFs and BSM effects from DIS data based on the NNPDF framework was presented in ref. [35]. The Wilson coefficients of the lepton-quark contact interactions (i.e. l-u, l-d, l-s and l-c contact interactions) are constrained by the HERA inclusive DIS data. The most stringent bounds are obtained for u-quark, followed by d-quark, and then c-quark and s-quark. The constraint on the Wilson coefficients for u-quark converted to $\tilde{c}/\Lambda^2$ is [-6.5 TeV^{-2}, 39.2 TeV^{-2}] at 90% CL. In ref. [36], the BSM effects are constrained by the high-mass Drell-Yan data. The result converted to $\tilde{c}/\Lambda^2$ is [-6.5 TeV^{-2}, 57.8 TeV^{-2}] at 95% CL.

In figure 23, we compare u, d, $\bar{u}$ and g PDFs at $Q = 1.295\,\text{GeV}$ determined by fitting with and without the new physics contributions. The PDF uncertainties are shown through hatched areas with relevant colors. The PDFs from two fits are almost indistinguishable for both central value and the uncertainties. In addition, we find that the central value is slightly changed if the $\tilde{c}/\Lambda^2$ is fixed at -8.60 or 9.72 TeV^{-2}, as represented by the two black

[1] The weak interactions from Z boson induce a different energy dependence of $1/(Q^2 + M_Z^2)$ which we neglect here for simplicity.

JHEP08(2022)088

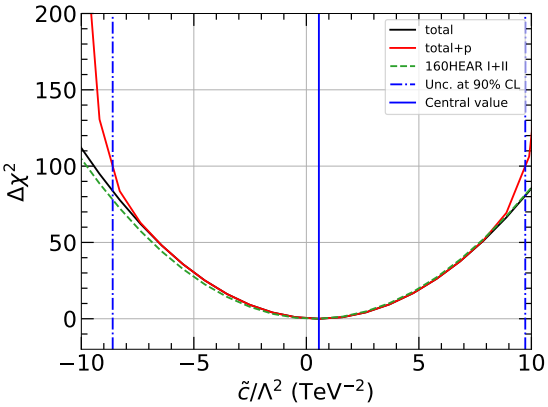


Figure 22. LM scans on $\tilde{c}/\Lambda^2$. The blue vertical solid line represents the central value of $\tilde{c}/\Lambda^2$, and the blue vertical dot-dash lines represent the uncertainties at 90% CL.

solid lines. Specifically, in the upper-left panel, a shift as large as 2% can be observed in the region of $x \sim 0.02$. Similar shifts can also be observed in panel (b) and panel (c). Besides, in the lower-right panel, a shift as large as 10% can be observed at both small-x ($\sim 10^{-4}$) and large-x (~ 0.6). These shifts on PDFs are required to compensate for the contributions from the new physics on DIS cross sections. Our approach can be extended to include more EFT operators from new physics which we leave for future studies.

6 Conclusion

Better understanding on parton distributions is essential for precision physics at hadron colliders, as well as for study of QCD. Nowadays the analysis of PDFs requires calculations of the log-likelihood functions χ^2 from thousands of experimental data points, and scans of multi-dimensional parameter space with tens of degrees of freedom. Such analyses will benefit from development of new methods and improvement of computing efficiencies, for instance by various interpolation approaches. In this paper we propose a new approach of using Neural Networks and machine learning techniques to model the dependence of the χ^2 or any physics quantities on the PDFs. We demonstrate the high accuracy of our approach through detailed comparisons in the PDF parameter space of interest, taking the CT18 NNLO analysis as an example. Importantly, compared with direct calculations the computational cost on calculating χ^2 are reduced by several orders of magnitude. The improvement ensures efficient scans of the full PDF parameter space and is desirable for the determination of PDF uncertainties.

Based on our NNs, we perform a series of LM scans to reevaluate PDF uncertainties in the CT18 NNLO analysis, and to understand the interplay between different data sets. The LM method is generally more reliable through a scan of the χ^2 along the trajectory of constrained minimum of the physics quantity studied. Our new approach renders such extensive scans almost costless and ensures the possibility of detailed comparisons of PDF uncertainties determined from the LM and Hessian method. We first perform LM scans

JHEP08(2022)088

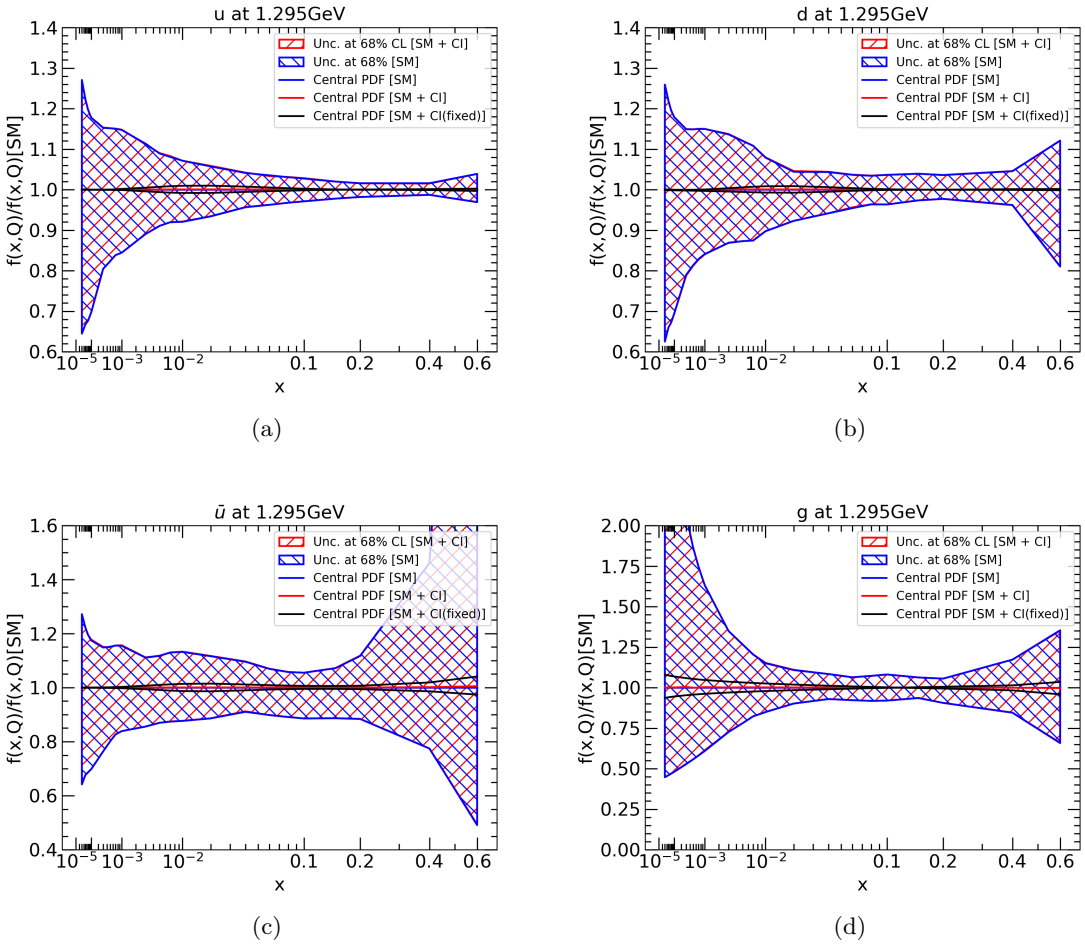


Figure 23. The parton distribution functions at $Q = 1.295\,\text{GeV}$ for u, d, $\bar{u}$ and g. The blue and the red solid lines represent the central values determined by fitting without and with the new physics contributions respectively, and the blue and the red hatched areas represent the respective uncertainties at 68% CL. The black solid lines represent the PDFs when $\tilde{c}/\Lambda^2$ is fixed at -8.60 or $9.72\,\text{TeV}^{-2}$.

on PDF values and ratios at various x and Q values, and find the results from the LM method and the Hessian method agree well in general. However, a notable difference can be observed in the small and the large-x regions. Since the quadratic approximation fails in the region where PDF uncertainties are large, and the results from the LM method are more reliable. Besides, we perform LM scans on the production cross sections of the Higgs boson pair and the top-quark pair in association with a Higgs boson, at the LHC or future colliders, as well as two dimensional scans on a pair of PDFs or cross sections. Furthermore, using LM scans we study the impact of individual data sets in the CT18 NNLO analysis by subtracting and adding back one data set at a time.

We show further applications of our approach on several extensions of the CT18 NNLO analysis. Especially, we study the impact of the NOMAD dimuon data on constraining the strange-quark PDFs. Theoretical predictions are calculated in the S-ACOT-χ general mass

JHEP08(2022)088

scheme up to NNLO, based on which the NNs are constructed and LM scans are performed. We find that the NOMAD data place stringent constraints on the strange-quark PDFs at intermediate and large-x regions. At $x \sim 0.05$, for example, the PDF uncertainties of the strange quark are reduced by almost a factor of 2. An upward shift of more than 15% in the strange-quark PDF as well as slight downward shift in the u and d-quark PDFs are also observed in most regions. We show the interplay of the NOMAD data and other data sets in the CT18 by detailed LM scans on R_s and s-quark PDFs at different scales and x values. The global fit with NOMAD data predicts $R_s(x = 0.023, Q = 1.5\text{GeV}) = 0.70^{+0.40}_{-0.17}$ at 90% CL and a slight tension between NOMAD and NuTeV data is observed. We also present a series of variant fits for clarifications on the impact of different theory predictions and of different choices of the decay branching ratio of the charm quark.

Afterwards, we study the impact of two HL-LHC pseudo-data constructed in ref. [33], including the high-mass Drell-Yan data and the forward W/Z production data. We find potentially large reduction on PDF uncertainties of the sea quarks. These results highlight the importance of HL-LHC measurements. Besides, we performed a joint fit on both PDFs and effects of new physics beyond the SM. We take the lepton-quark contact interactions as an example that are described by high dimensional operators in the SMEFT. We determine the effective Wilson coefficient to be $\tilde{c}/\Lambda^2 = 0.56^{+9.16}_{-9.16}\,\text{TeV}^{-2}$ at 90% CL as mostly constrained by the HERA inclusive DIS data. Foreseen extensions of the study would be to include more SMEFT operators in the joint fit that is under investigation.

Acknowledgments

This work was sponsored by the National Natural Science Foundation of China under the Grant No. 11875189 and No.11835005. JG would like to thank members of CTEQ-TEA collaboration for helpful discussions and proofreading of the manuscript.

A More on the Neural Network approach

In this appendix we collect various details of the NN approach, including on the architectures and parametrization dependence, the generation of training and test samples, and the performances in terms of computational cost. One important feature of our approach is to use directly PDF values as inputs to the NNs rather than the PDF parameter themselves. That ensures a great flexibility of the functional space since we can select PDF values at an arbitrary number of x points. In our current study with CT18 parametrization form, we select the x grid consisting of 14 points for each PDF flavor with their values shown in table 4. They are selected randomly with the only criteria being distributed evenly in $\ln x$.

We explain briefly on the mathematical model behind our NNs. The true dependence of our target function, for instance, the χ^2, on the PDF parameters is uniquely determined by the theory and experimental data used in the global fit, which we denote as A_{TR}. On another hand if we exchange the PDF parameters by the PDF values at discrete x points, the mapping is not unique since we have input PDF values far more than the number of PDF parameters. Thus we arrive at a bunch of possible functions $\{A^*_{\text{TR}}\}$ depending

JHEP08(2022)088

	1	2	3	4	5	6	7
x	3.30×10^{-5}	3.73×10^{-4}	3.50×10^{-3}	1.24×10^{-2}	2.48×10^{-2}	4.34×10^{-2}	8.59×10^{-2}
	8	9	10	11	12	13	14
x	0.118	0.167	0.206	0.302	0.406	0.637	0.831

Table 4. The x values we choose for the training and test samples of NNs.

explicitly on $\{I_k\}$ which is the PDF value at the k_{th} node, satisfying

$$\chi^2_{\rm truth} = A_{\rm TR}(a) = A^*_{\rm TR}(\{I_k(a)\}). \tag{A.1}$$

The purpose of our NNs is to construct an explicit function of $\{I_k\}$ depending on a set of tuneable parameters t_β. By the training procedure we update $A_{\rm NN}$ iteratively until it converges to the neighborhood of one of the truth function $A^*_{\rm TR}$ with a choice on the parameters $\hat{t}_\beta$. Finally we arrive at our approximation to the χ^2 dependence on the PDF parameters as

$$\chi^2_{\rm pred} = A_{\rm NN}\left(\{I_k(a)\}\,;\,\{\hat{t}_\beta\}\right). \tag{A.2}$$

As from above one expects that the outcome NN (or equivalently the solution $\hat{t}_\beta$) in general depends on the parametrization form of PDFs. However, in practice one can approximate either PDFs or cross sections in terms of interpolated functions on a dense x-grid with sufficient accuracy, as implemented successfully in APPLgrid [43], FastNLO [92], and FastKernal [14]. In that sense there may exist an almost universal solution for different parametrization forms if one start with a sufficiently large number of PDF inputs. We leave that for future investigations.

The 8000 PDF replicas used for training and test are generated through a randomly sampling of the PDF parameters defined in eq. (2.4) with the help of CT18 NNLO Hessian PDF set. Each replica or PDF parameters $a_i^{\rm rep}$ is determined by 28 randomly distributed variables r_j, namely

$$a_i^{\rm rep} = a_i^0 + \sum_{j=1}^{28} r_j \left(a_i^{2j-1} - a_i^{2j}\right)/2, \tag{A.3}$$

where $\{a_i^0\}$ represent the i_{th} PDF parameter of the central PDF of CT18 NNLO, $\{a_i^{2j-1}\}$ and $\{a_i^{2j}\}$ represent the i_{th} PDF parameter of the error PDFs in the plus and the minus direction of the i_{th} eigenvector respectively. For each r_j we use a Gaussian sampling with mean value 0 and variance $1/\sqrt{28}$ which ensures coverage of the PDF parameter space with average increase of global χ^2 of a few hundred units comparing to CT18 best fit as shown in figure 2. We note that performances of the trained NNs are not sensitive to the choice of training samples as far as we are within or close to the uncertainty range of CT18.

We further test performance of our NN approach with alternative PDF parametrization forms taking the target function of χ^2 of the ATLAS 7 TeV jet data as an example. We have chosen MMHT2014 [93], NNPDF3.1 [94] and NNPDF4.0 [14] NNLO PDFs with 4000 MC PDF replicas each generated from the corresponding Hessian PDF sets with

JHEP08(2022)088

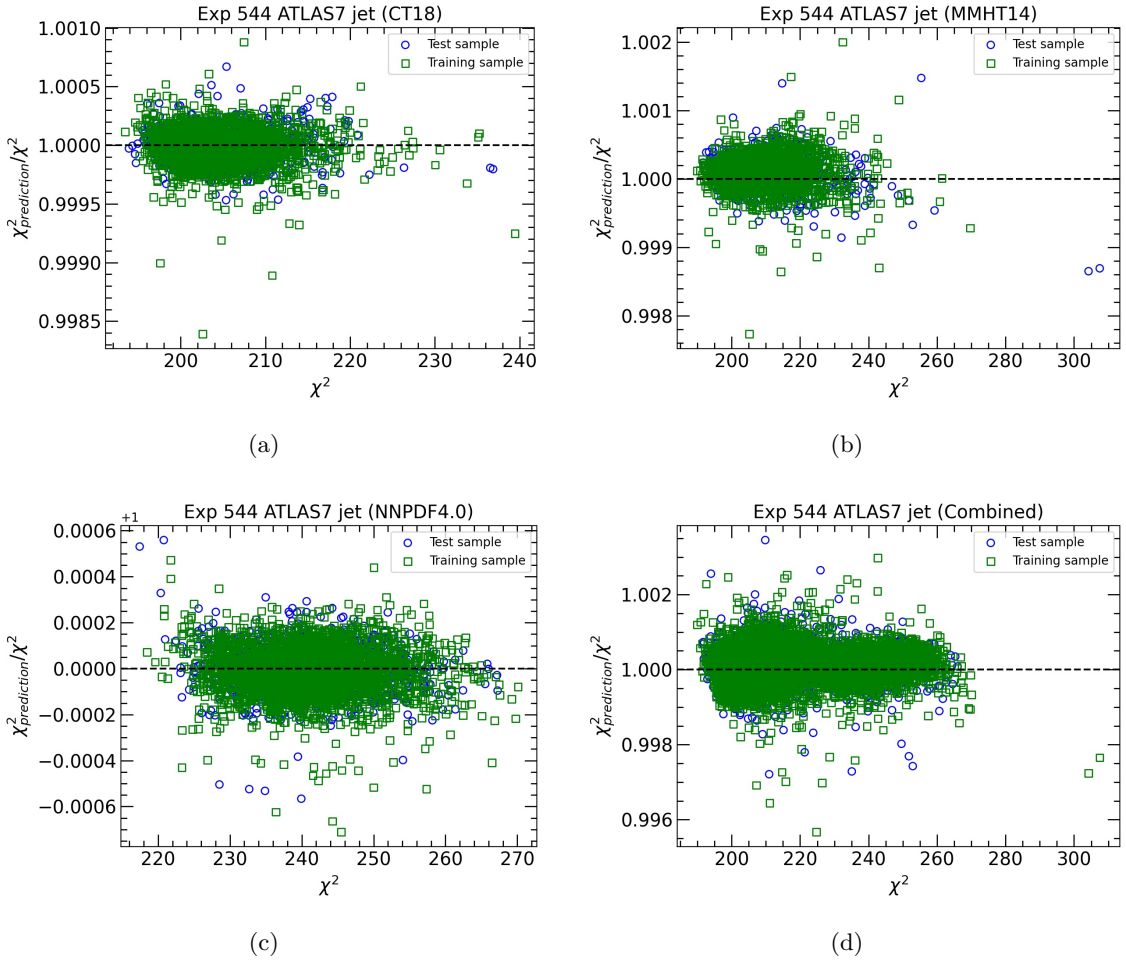


Figure 24. The predictions to truths ratios of χ^2 for measurement of the inclusive jet cross sections at $\sqrt{s} = 7\,\text{TeV}$ by ATLAS when training the NNs to individual PDF parametrization forms including CT18, MSHT14 and NNPDF4.0, or an ensemble of PDF replicas of the three.

LHAPDF6 [95].[2] We use the same architecture as used for CT18 except for extensions to include 9 PDF flavors, namely with $\bar{s}$, c, and b-quark PDFs in addition. The NNs have been trained and tested for each individual parametrization form with the corresponding MC replicas. We find very good performance of the NNs in cases of MMHT2014 and NNPDF4.0 as shown in figure 24, similar to the case of CT18. Interestingly, we find performance of the same architecture is much better for the parametrization form of NNPDF4.0 than NNPDF3.1, possibly due to the smooth conditions applied in NNPDF4.0 [14]. We also try to train the NNs with an ensemble of PDF replicas, 12000 replicas in total, from CT18, MMHT14 and NNPDF4.0. The accuracy of the trained NNs is only marginally worse than the NNs trained to individual parametrization forms. That hints the possibility of a universal NN to accommodate for a variety of smooth PDF parametrizations, as discussed at earlier this section.

[2]We have not used the native MC replicas of NNPDF since the numbers of replicas are limited to be 1000 in that case.

JHEP08(2022)088

cost \ target / method	χ^2	σ	$f(x,Q)$
NNs	0.70 ms	0.41 ms	0.37 ms
traditional	10^7(200) ms	10^6(20) ms	20(2) ms

Table 5. Comparison on computational cost between NNs and traditional methods. Numbers in parenthesis represent cases if fast interpolations on PDFs are used.

Finally we summarize the performances of our NNs in terms of computational cost in table 5 comparing to the traditional approaches. Note that we have not included the time cost for the process of training of the NNs since we do not need to repeat it in later scans of the PDF parameters. In table 5 the numbers indicate the time cost on a single CPU-core (2.4 GHz) of calculating the target functions for a single point in the PDF parameter space. For χ^2 the cost includes those for the calculations of the needed cross sections (taking 10 points per data set as an example) and for the multiplications with covariance matrix. In the conventional approach the computing efficiency varies significantly, e.g., for the χ^2, depending on the number of data points, the perturbative order of the theory calculations, and importantly whether the fast interpolation algorithms are used or not. Thus included numbers only represent typical average cost in the CT18 NNLO analysis for a direct calculation or using fast interpolations (shown in parenthesis). The fast interpolation method for calculations of a single cross section involves more PDF values on a dense grid and thus is slower than the NN approaches. Nevertheless, the NN approaches lead to significant improvement in general and ensure efficient scans of the PDF parameter space with much less cost. The NNs was programmed with PYTHON2.7, and we expect further reduction of the computational cost if transferred into more efficient programming languages like Fortran/C++. We are planning to provide open source access for the NN framework used together with trained NNs for various target functions of CT18 in the near future.

B Hessian PDF set

We further generate a Hessian PDF set based on the χ^2 profile obtained with the NN approaches. The Hessian error matrix on the PDF parameters is calculated using a numeric method of finite difference. We use an iterative algorithm on diagonalization of the Hessian matrix that is developed in refs. [20, 23] and used in later CTEQ analyses. The iterative procedure greatly improves the performance of Hessian approximation in the case of large number of free parameters (28 here) and in the existence of flat directions. Once all orthogonal eigenvectors are determined, two error PDFs are generated for each eigenvector by scanning along the plus and the minus directions and looking for solutions with $\Delta\chi^2 + P = 100$ (for 90% CL).

We compare the PDF uncertainties at 68% CL from the Hessian PDF set to the published CT18 NNLO PDFs in figure 25. We find very good agreements between predictions of

JHEP08(2022)088

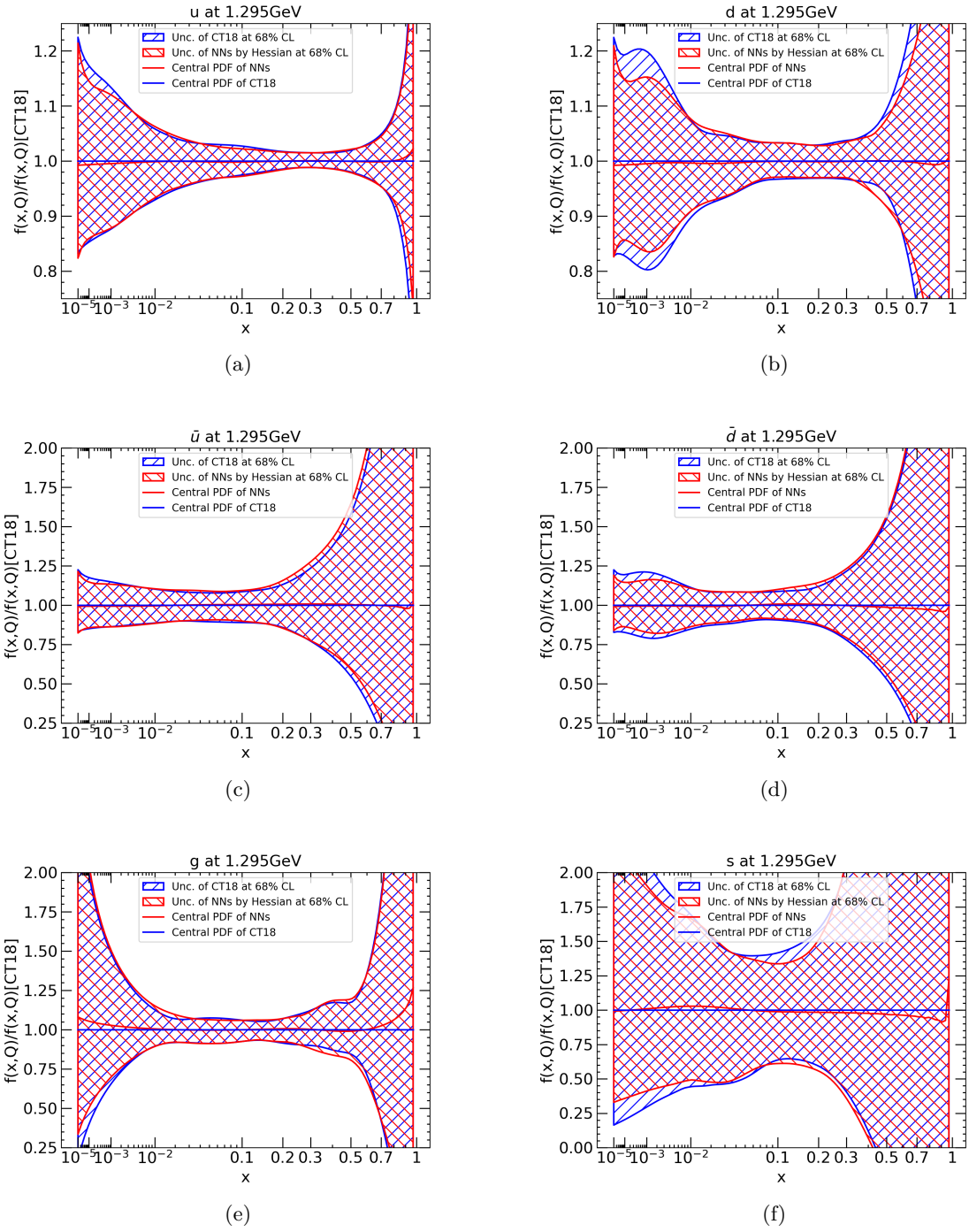


Figure 25. The parton distribution functions at $Q = 1.295\,\text{GeV}$ for u, d, $\bar{u}$, $\bar{d}$, g and s. The blue and the red solid lines represent the central values of CT18NNLO and NNs respectively. The blue and the red hatched areas represent the uncertainties of CT18NNLO and final PDFs in this paper at 68% CL respectively.

JHEP08(2022)088

the two Hessian PDF sets in general. However, some notable differences can be seen for d-valence and gluon PDFs at large-x (~ 0.4) as well as for sea quarks at $x \lesssim 10^{-3}$. There are two reasons that lead to the differences in the new Hessian set and the CT18 set. First as mentioned earlier in the global fit presented in this paper the NNLO K-factors used for predictions of the Drell-Yan data have been updated comparing to those used in the CT18 analysis. Besides, when calculating the Hessian error matrix numerically we use a step size of $\Delta\chi^2 = 10$ on sampling of the PDF parameters while a value of ~ 1 is used in the CT18 analysis. The dependence on choices of this step size reflect one intrinsic uncertainty of the Hessian approaches [23].

C Variant fits with NOMAD data

In this appendix we present a series of global fit with inclusion of the NOMAD data and with different theories or different choices of decay branching ratios of charm quark to muon. In section 5.1 when comparing to the global fit without NOMAD data, we use NLO cross sections (but with NNLO PDFs) for NuTeV and CCFR dimuon data to be consistent with the CT18 analysis. Thus the changes observed after including the NOMAD data can be due to both the NOMAD data or the changes of theories for the other two dimuon data. Now we further consider different choices of the theory predictions, namely either calculated at NLO or NNLO in QCD, to disentangle their effects.

Furthermore, to compare with dimuon data, one has to convert the cross sections of charm-quark production to production of dimuon which relies on the input of inclusive semileptonic branching ratio $\mathrm{Br}(c \to \mu)$. Since NOMAD dimuon data extend down to $E_\nu \sim 6\,\mathrm{GeV}$ the energy dependence of $\mathrm{Br}(c \to \mu)$ is taken into account in NOMAD analysis, with a parametrization form

$$Br(c \to \mu) = \frac{a}{1 + b/E_\nu}, \tag{C.1}$$

where a and b are free parameters. In the NOMAD paper it suggests values of $a = 0.094 \pm 0.010$ and $b = 6.6 \pm 3.9\,\mathrm{GeV}$ as measured by the E531 experiment [96]. The uncertainties on parameters a and b will propagate into the unfolded charm-quark cross sections and are treated as additional correlated systematic errors that are summarized in table 6 for distribution in Bjorken-x. Other correlated systematic uncertainties for NOMAD data can be found in ref. [31]. On the other hand, for NuTeV and CCFR data, since the neutrino energies are sufficiently high, a constant value of $\mathrm{Br}(c \to \mu)$=0.099 ± 0.010 has been suggested [52] and is used in the CT18 analysis. The central value is slightly higher than the parameter a used in our nominal fit of NOMAD data. Thus we perform variant fits using $a = 0.099 \pm 0.010$ for NOMAD data to further investigate the impact of this overall normalization on the outcome PDFs.

In figure 26 we compare the strange-quark PDFs at $1.295\,\mathrm{GeV}$ from all variant fits. We show the PDF uncertainties at 68% CL from LM scans for fits with and without NOMAD data and using NNLO predictions from dimuon production consistently. That can be compared with figure 17 where NLO predictions are used in fit without NOMAD data. We also present central PDFs obtained with NLO predictions or with higher branching ratio for

JHEP08(2022)088

x_{Bj}	Bin center	$\sigma_{\mu\mu}/\sigma_{cc} \pm \delta^{\mathrm{stat}} \pm \delta^{\mathrm{syst}}$ (10^{-3})	δ^a, %	δ^b, %
0.0000 – 0.0336	0.0267	13.383 ± 0.441 ± 0.289	10.6	5.3
0.0336 – 0.0511	0.0440	11.245 ± 0.380 ± 0.210	10.6	6.8
0.0511 – 0.0672	0.0598	9.991 ± 0.347 ± 0.201	10.6	7.7
0.0672 – 0.0836	0.0756	9.141 ± 0.324 ± 0.189	10.6	8.3
0.0836 – 0.1000	0.0917	8.198 ± 0.297 ± 0.169	10.6	8.8
0.1000 – 0.1246	0.1122	7.176 ± 0.225 ± 0.144	10.6	9.0
0.1246 – 0.1535	0.1389	6.229 ± 0.195 ± 0.118	10.6	9.4
0.1535 – 0.1870	0.1699	5.427 ± 0.171 ± 0.106	10.6	9.6
0.1870 – 0.2277	0.2066	4.837 ± 0.151 ± 0.093	10.6	9.9
0.2277 – 0.2800	0.2524	4.235 ± 0.133 ± 0.083	10.6	10.0
0.2800 – 0.3590	0.3165	3.595 ± 0.113 ± 0.072	10.6	10.0
0.3590 – 0.4583	0.4036	2.955 ± 0.111 ± 0.062	10.6	10.1
0.4583 – 0.5838	0.5116	2.355 ± 0.120 ± 0.055	10.6	9.9
0.5838 – 0.7500	0.6465	1.607 ± 0.150 ± 0.047	10.6	9.4

Table 6. NOMAD measurements on the Bjorken-x distribution of dimuon to inclusive CC cross section ratio, including the binning, central values, statistical and total systematic uncertainties. The last two columns show the additional correlated systematic uncertainties in percentages, if converting back to production cross sections of charm-quark, due to input parameter a and b respectively. These additional errors are derived based on theoretical cross sections at NLO with CT18 NNLO PDFs.

NOMAD data. We find including the NNLO corrections leads to a moderate increase of the strange-quark PDF, which is in consistent with the conclusions in ref. [12]. The inclusion of NOMAD data results in about 20% enhancement of the strange-quark PDF at x around 0.05 and a significant reduction of the PDF uncertainties. Changing to $a = 0.099 \pm 0.010$ for NOMAD only induces a minor reduction of the strange-quark PDF.

We further summarize the total or individual χ^2 of all variant fits together with predictions on $R_s(x = 0.023, Q = 1.5\mathrm{GeV})$ with uncertainties at 68% CL in table 7. By comparison with the χ^2 we find the NNLO predictions in general lead to a slightly worse fit with increase on χ^2 of a few units. However, in all cases the global fit can describe well various dimuon data as can be seen from the χ^2 per number of degree of freedoms. When comparing fits in the last two rows we find using a consistent branching ratio in different data sets results in a better fit and reduced PDF uncertainties.

JHEP08(2022)088

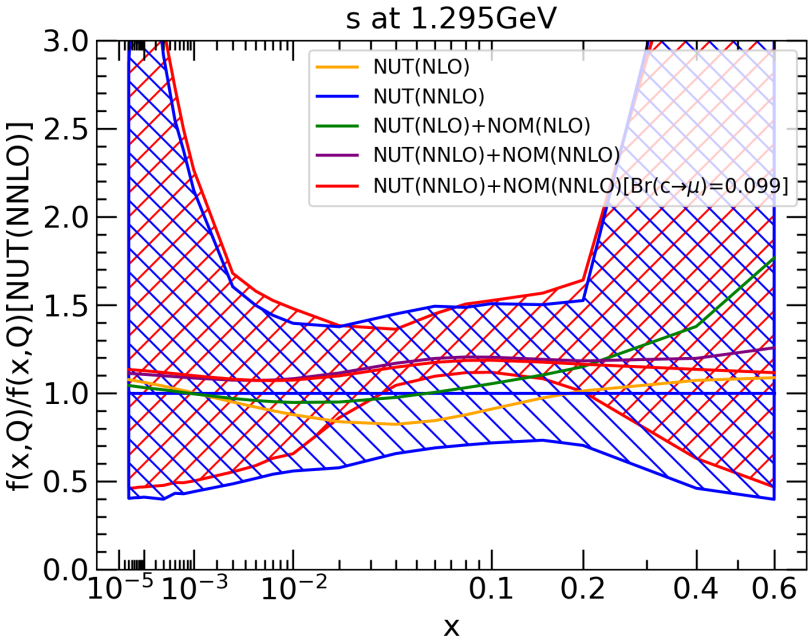


Figure 26. Strange-quark PDF at $Q = 1.295\,\text{GeV}$ from LM scans in global fits with various conditions. The PDF uncertainties are shown for 68% CL.

data sets	χ^2_{total} (3671/3683)	χ^2_{nomad} (12)	χ^2_{124} (38)	χ^2_{125} (33)	χ^2_{126} (40)	χ^2_{127} (38)	$R_s(0.023,1.5\text{GeV})$
NUT (NLO)	4272.64	—	18.83	39.55	29.89	19.42	$0.518^{+0.349}_{-0.363}$
NUT (NNLO)	4268.77	—	21.44	32.84	34.06	22.54	$0.616^{+0.441}_{-0.377}$
NUT (NLO) + NOM (NLO)	4286.28	8.39	24.81	41.95	29.30	18.79	$0.593^{+0.256}_{-0.155}$
NUT (NNLO) + NOM (NNLO)	4291.47	14.47	28.13	34.26	34.21	22.15	$0.695^{+0.384}_{-0.169}$
Br(c→ μ)=0.099	4289.61	13.79	27.38	34.06	34.10	22.13	$0.685^{+0.290}_{-0.174}$

Table 7. Total χ^2 and individual χ^2 of dimuon data for global fits with various conditions. Numbers in the first row indicate the total number of data points. The last column includes predictions on $R_s(x = 0.023, Q = 1.5\text{GeV})$ with uncertainties at 90% CL.

JHEP08(2022)088

References

[1] J. Gao, L. Harland-Lang and J. Rojo, *The Structure of the Proton in the LHC Precision Era*, *Phys. Rept.* **742** (2018) 1 [arXiv:1709.04922] [INSPIRE].

[2] K. Kovařík, P.M. Nadolsky and D.E. Soper, *Hadronic structure in high-energy collisions*, *Rev. Mod. Phys.* **92** (2020) 045003 [arXiv:1905.06957] [INSPIRE].

[3] J.C. Collins, D.E. Soper and G.F. Sterman, *Factorization of Hard Processes in QCD*, *Adv. Ser. Direct. High Energy Phys.* **5** (1989) 1 [hep-ph/0409313] [INSPIRE].

[4] X. Ji, Y.-S. Liu, Y. Liu, J.-H. Zhang and Y. Zhao, *Large-momentum effective theory*, *Rev. Mod. Phys.* **93** (2021) 035005 [arXiv:2004.03543] [INSPIRE].

[5] LATTICE PARTON collaboration, *Unpolarized isovector quark distribution function from lattice QCD: A systematic analysis of renormalization and matching*, *Phys. Rev. D* **101** (2020) 034020 [arXiv:1807.06566] [INSPIRE].

[6] LHC HIGGS CROSS SECTION WORKING GROUP collaboration, *Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector*, arXiv:1610.07922 [INSPIRE].

[7] W. Beenakker et al., *NLO+NLL squark and gluino production cross-sections with threshold-improved parton distributions*, *Eur. Phys. J. C* **76** (2016) 53 [arXiv:1510.00375] [INSPIRE].

[8] S. Alioli, M. Farina, D. Pappadopulo and J.T. Ruderman, *Precision Probes of QCD at High Energies*, *JHEP* **07** (2017) 097 [arXiv:1706.03068] [INSPIRE].

[9] ATLAS collaboration, *Determination of the strong coupling constant and test of asymptotic freedom from Transverse Energy-Energy Correlations in multijet events at* $\sqrt{s} = 13$ *TeV with the ATLAS detector*, ATLAS-CONF-2020-025 [INSPIRE].

[10] CMS and ATLAS collaborations, *Standard Model: Electroweak Physics with CMS and ATLAS at* 13 *TeV*, *J. Phys. Conf. Ser.* **1258** (2019) 012015 [INSPIRE].

[11] ATLAS collaboration, *Measurement of the W-boson mass in pp collisions at* $\sqrt{s} = 7$ *TeV with the ATLAS detector*, *Eur. Phys. J. C* **78** (2018) 110 [*Erratum ibid.* **78** (2018) 898] [arXiv:1701.07240] [INSPIRE].

[12] T.-J. Hou et al., *New CTEQ global analysis of quantum chromodynamics with high-precision data from the LHC*, *Phys. Rev. D* **103** (2021) 014013 [arXiv:1912.10053] [INSPIRE].

[13] S. Bailey, T. Cridge, L.A. Harland-Lang, A.D. Martin and R.S. Thorne, *Parton distributions from LHC, HERA, Tevatron and fixed target data: MSHT20 PDFs*, *Eur. Phys. J. C* **81** (2021) 341 [arXiv:2012.04684] [INSPIRE].

[14] NNPDF collaboration, *The path to proton structure at* 1% *accuracy*, *Eur. Phys. J. C* **82** (2022) 428 [arXiv:2109.02653] [INSPIRE].

[15] S. Alekhin, J. Blümlein and S. Moch, *NLO PDFs from the ABMP16 fit*, *Eur. Phys. J. C* **78** (2018) 477 [arXiv:1803.07537] [INSPIRE].

JHEP08(2022)088

[16] H1 collaboration, *Determination of the strong coupling constant $\alpha_s(m_Z)$ in next-to-next-to-leading order QCD using H1 jet cross section measurements*, *Eur. Phys. J. C* **77** (2017) 791 [*Erratum ibid.* **81** (2021) 738] [arXiv:1709.07251] [INSPIRE].

[17] P. Jimenez-Delgado and E. Reya, *Delineating parton distributions and the strong coupling*, *Phys. Rev. D* **89** (2014) 074049 [arXiv:1403.1852] [INSPIRE].

[18] S. Park, A. Accardi, X. Jing and J.F. Owens, *CJ15 global PDF analysis with new electroweak data from the STAR and SeaQuest experiments*, in *28th International Workshop on Deep Inelastic Scattering and Related Subjects*, Online conference U.S.A., 12–16 April 2021 [arXiv:2108.05786] [INSPIRE].

[19] ATLAS collaboration, *Determination of the parton distribution functions of the proton using diverse ATLAS data from pp collisions at $\sqrt{s} = 7$, 8 and 13 TeV*, *Eur. Phys. J. C* **82** (2022) 438 [arXiv:2112.11266] [INSPIRE].

[20] J. Pumplin et al., *Uncertainties of predictions from parton distribution functions. 2. The Hessian method*, *Phys. Rev. D* **65** (2001) 014013 [hep-ph/0101032] [INSPIRE].

[21] A.D. Martin, W.J. Stirling, R.S. Thorne and G. Watt, *Parton distributions for the LHC*, *Eur. Phys. J. C* **63** (2009) 189 [arXiv:0901.0002] [INSPIRE].

[22] S. Forte, L. Garrido, J.I. Latorre and A. Piccione, *Neural network parametrization of deep inelastic structure functions*, *JHEP* **05** (2002) 062 [hep-ph/0204232] [INSPIRE].

[23] J. Pumplin, D.R. Stump and W.K. Tung, *Multivariate fitting and the error matrix in global analysis of data*, *Phys. Rev. D* **65** (2001) 014011 [hep-ph/0008191] [INSPIRE].

[24] D. Stump et al., *Uncertainties of predictions from parton distribution functions. 1. The Lagrange multiplier method*, *Phys. Rev. D* **65** (2001) 014012 [hep-ph/0101051] [INSPIRE].

[25] J. Gao and P. Nadolsky, *A meta-analysis of parton distribution functions*, *JHEP* **07** (2014) 035 [arXiv:1401.0013] [INSPIRE].

[26] C. Schmidt, J. Pumplin, C.P. Yuan and P. Yuan, *Updating and optimizing error parton distribution function sets in the Hessian approach*, *Phys. Rev. D* **98** (2018) 094005 [arXiv:1806.07950] [INSPIRE].

[27] B.-T. Wang et al., *Mapping the sensitivity of hadronic experiments to nucleon structure*, *Phys. Rev. D* **98** (2018) 094030 [arXiv:1803.02777] [INSPIRE].

[28] D. Guest, K. Cranmer and D. Whiteson, *Deep Learning and its Application to LHC Physics*, *Ann. Rev. Nucl. Part. Sci.* **68** (2018) 161 [arXiv:1806.11484] [INSPIRE].

[29] NNPDF collaboration, *Parton distributions for the LHC Run II*, *JHEP* **04** (2015) 040 [arXiv:1410.8849] [INSPIRE].

[30] S. Forte and S. Carrazza, *Parton distribution functions*, arXiv:2008.12305 [INSPIRE].

[31] NOMAD collaboration, *A Precision Measurement of Charm Dimuon Production in Neutrino Interactions from the NOMAD Experiment*, *Nucl. Phys. B* **876** (2013) 339 [arXiv:1308.4750] [INSPIRE].

[32] A. Dainese, M. Mangano, A.B. Meyer, A. Nisati, G. Salam and M.A. Vesterinen, *Report on the Physics at the HL-LHC,and Perspectives for the HE-LHC*, *CERN Yellow Rep. Monogr* **7** 2019 1.

[33] R. Abdul Khalek, S. Bailey, J. Gao, L. Harland-Lang and J. Rojo, *Towards Ultimate Parton Distributions at the High-Luminosity LHC*, *Eur. Phys. J. C* **78** (2018) 962 [arXiv:1810.03639] [INSPIRE].

JHEP08(2022)088

[34] R. Abdul Khalek, S. Bailey, J. Gao, L. Harland-Lang and J. Rojo, *Probing Proton Structure at the Large Hadron electron Collider*, *SciPost Phys.* **7** (2019) 051 [arXiv:1906.10127] [INSPIRE].

[35] S. Carrazza, C. Degrande, S. Iranipour, J. Rojo and M. Ubiali, *Can New Physics hide inside the proton?*, *Phys. Rev. Lett.* **123** (2019) 132001 [arXiv:1905.05215] [INSPIRE].

[36] A. Greljo et al., *Parton distributions in the SMEFT from high-energy Drell-Yan tails*, *JHEP* **07** (2021) 122 [arXiv:2104.02723] [INSPIRE].

[37] M. Madigan and J. Moore, *Parton Distributions in the SMEFT from high-energy Drell-Yan tails*, *PoS* **EPS-HEP2021** (2022) 424 [arXiv:2110.13204] [INSPIRE].

[38] CMS collaboration, *Measurement and QCD analysis of double-differential inclusive jet cross sections in proton-proton collisions at* $\sqrt{s} = 13$ *TeV*, *JHEP* **02** (2022) 142 [arXiv:2111.10431] [INSPIRE].

[39] S. Iranipour and M. Ubiali, *A new generation of simultaneous fits to LHC data using deep learning*, *JHEP* **05** (2022) 032 [arXiv:2201.07240] [INSPIRE].

[40] *Keras*, https://keras.io/.

[41] J. Alwall et al., *The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations*, *JHEP* **07** (2014) 079 [arXiv:1405.0301] [INSPIRE].

[42] V. Bertone, R. Frederix, S. Frixione, J. Rojo and M. Sutton, *aMCfast: automation of fast NLO computations for PDF fits*, *JHEP* **08** (2014) 166 [arXiv:1406.7693] [INSPIRE].

[43] T. Carli et al., *A posteriori inclusion of parton density functions in NLO QCD final-state calculations at hadron colliders: The APPLGRID Project*, *Eur. Phys. J. C* **66** (2010) 503 [arXiv:0911.2985] [INSPIRE].

[44] G.P. Salam and J. Rojo, *A Higher Order Perturbative Parton Evolution Toolkit (HOPPET)*, *Comput. Phys. Commun.* **180** (2009) 120 [arXiv:0804.3755] [INSPIRE].

[45] H1 and ZEUS collaborations, *Combination of measurements of inclusive deep inelastic* $e^{\pm}p$ *scattering cross sections and QCD analysis of HERA data*, *Eur. Phys. J. C* **75** (2015) 580 [arXiv:1506.06042] [INSPIRE].

[46] BCDMS collaboration, *A High Statistics Measurement of the Proton Structure Functions* $F_2(x, Q^2)$ *and R from Deep Inelastic Muon Scattering at High* Q^2, *Phys. Lett. B* **223** (1989) 485 [INSPIRE].

[47] BCDMS collaboration, *A High Statistics Measurement of the Deuteron Structure Functions* $F_2(x, Q^2)$ *and R From Deep Inelastic Muon Scattering at High* Q^2, *Phys. Lett. B* **237** (1990) 592 [INSPIRE].

[48] NEW MUON collaboration, *Measurement of the proton and deuteron structure functions,* F_2^p *and* F_2^d, *and of the ratio* $\sigma_L \sigma_T$, *Nucl. Phys. B* **483** (1997) 3 [hep-ph/9610231] [INSPIRE].

[49] J.P. Berge et al., *A Measurement of Differential Cross-Sections and Nucleon Structure Functions in Charged Current Neutrino Interactions on Iron*, *Z. Phys. C* **49** (1991) 187 [INSPIRE].

[50] CCFR/NUTEV collaboration, *Measurements of* F_2 *and* $xF_3^{\nu} - xF_3^{\bar{\nu}}$ *from CCFR* ν_μ*−Fe and* $\bar{\nu}_\mu$*−Fe data in a physics model independent way*, *Phys. Rev. Lett.* **86** (2001) 2742 [hep-ex/0009041] [INSPIRE].

[51] W.G. Seligman et al., *Improved determination of* α_s *from neutrino nucleon scattering*, *Phys. Rev. Lett.* **79** (1997) 1213 [hep-ex/9701017] [INSPIRE].

JHEP08(2022)088

[52] D.A. Mason, *Measurement of the strange - antistrange asymmetry at nlo in qcd from nutev dimuon data*, Ph.D. Thesis, Department of Physics, University of Oregon, Oregon U.S.A (2006) [DOI] [INSPIRE].

[53] NuTeV collaboration, *Precise Measurement of Dimuon Production Cross-Sections in ν_μ Fe and $\bar{\nu}_\mu$ Fe Deep Inelastic Scattering at the Tevatron*, *Phys. Rev. D* **64** (2001) 112006 [hep-ex/0102049] [INSPIRE].

[54] H1 collaboration, *Measurement of $F_2^{c\bar{c}}$ and $F_2^{b\bar{b}}$ at high Q^2 using the H1 vertex detector at HERA*, *Eur. Phys. J. C* **40** (2005) 349 [hep-ex/0411046] [INSPIRE].

[55] H1 and ZEUS collaborations, *Combination and QCD Analysis of Charm Production Cross Section Measurements in Deep-Inelastic ep Scattering at HERA*, *Eur. Phys. J. C* **73** (2013) 2311 [arXiv:1211.1182] [INSPIRE].

[56] H1 collaboration, *Measurement of the Inclusive $e^\pm p$ Scattering Cross Section at High Inelasticity y and of the Structure Function F_L*, *Eur. Phys. J. C* **71** (2011) 1579 [arXiv:1012.4355] [INSPIRE].

[57] G. Moreno et al., *Dimuon production in proton - copper collisions at $\sqrt{s} = 38.8-GeV$*, *Phys. Rev. D* **43** (1991) 2815 [INSPIRE].

[58] NuSea collaboration, *Improved measurement of the $\bar{d}/\bar{u}$ asymmetry in the nucleon sea*, *Phys. Rev. D* **64** (2001) 052002 [hep-ex/0103030] [INSPIRE].

[59] NuSea collaboration, *Absolute Drell-Yan dimuon cross-sections in 800 GeV/c pp and pd collisions*, hep-ex/0302019 [INSPIRE].

[60] CDF collaboration, *Measurement of the Lepton Charge Asymmetry in W Boson Decays Produced in $p\bar{p}$ Collisions*, *Phys. Rev. Lett.* **81** (1998) 5754 [hep-ex/9809001] [INSPIRE].

[61] CDF collaboration, *Measurement of the forward-backward charge asymmetry from $W \to e\nu$ production in $p\bar{p}$ collisions at $\sqrt{s} = 1.96$ TeV*, *Phys. Rev. D* **71** (2005) 051104 [hep-ex/0501023] [INSPIRE].

[62] D0 collaboration, *Measurement of the muon charge asymmetry from W boson decays*, *Phys. Rev. D* **77** (2008) 011106 [arXiv:0709.4254] [INSPIRE].

[63] D0 collaboration, *Measurement of the Shape of the Boson Rapidity Distribution for $p\bar{p} \to Z/\gamma^* \to e^+e^- + X$ Events Produced at $\sqrt{s}$ of 1.96 TeV*, *Phys. Rev. D* **76** (2007) 012003 [hep-ex/0702025] [INSPIRE].

[64] CDF collaboration, *Measurement of $d\sigma/dy$ of Drell-Yan e^+e^- pairs in the Z Mass Region from $p\bar{p}$ Collisions at $\sqrt{s} = 1.96$ TeV*, *Phys. Lett. B* **692** (2010) 232 [arXiv:0908.3914] [INSPIRE].

[65] CMS collaboration, *Measurement of the Muon Charge Asymmetry in Inclusive $pp \to W + X$ Production at $\sqrt{s} = 7$ TeV and an Improved Determination of Light Parton Distribution Functions*, *Phys. Rev. D* **90** (2014) 032004 [arXiv:1312.6283] [INSPIRE].

[66] CMS collaboration, *Measurement of the Electron Charge Asymmetry in Inclusive W Production in pp Collisions at $\sqrt{s} = 7$ TeV*, *Phys. Rev. Lett.* **109** (2012) 111806 [arXiv:1206.2598] [INSPIRE].

[67] ATLAS collaboration, *Measurement of the inclusive $W^\pm$ and Z/gamma cross sections in the electron and muon decay channels in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector*, *Phys. Rev. D* **85** (2012) 072004 [arXiv:1109.5141] [INSPIRE].

JHEP08(2022)088

[68] D0 collaboration, *Measurement of the electron charge asymmetry in $p\bar{p} \to W + X \to e\nu + X$ decays in $p\bar{p}$ collisions at $\sqrt{s} = 1.96$ TeV*, *Phys. Rev. D* **91** (2015) 032007 [*Erratum ibid.* **91** (2015) 079901] [arXiv:1412.2862] [INSPIRE].

[69] CDF collaboration, *Measurement of the Inclusive Jet Cross Section at the Fermilab Tevatron $p\bar{p}$ Collider Using a Cone-Based Jet Algorithm*, *Phys. Rev. D* **78** (2008) 052006 [*Erratum ibid.* **79** (2009) 119902] [arXiv:0807.2204] [INSPIRE].

[70] D0 collaboration, *Measurement of the inclusive jet cross-section in $p\bar{p}$ collisions at $s^{(1/2)} = 1.96$ TeV*, *Phys. Rev. Lett.* **101** (2008) 062001 [arXiv:0802.2400] [INSPIRE].

[71] LHCb collaboration, *Measurement of the forward Z boson production cross-section in pp collisions at $\sqrt{s} = 7$ TeV*, *JHEP* **08** (2015) 039 [arXiv:1505.07024] [INSPIRE].

[72] LHCb collaboration, *Measurement of forward $\mathrm{Z} \to \mathrm{e}^+\mathrm{e}^-$ production at $\sqrt{s} = 8$ TeV*, *JHEP* **05** (2015) 109 [arXiv:1503.00963] [INSPIRE].

[73] CMS collaboration, *Measurement of the differential cross section and charge asymmetry for inclusive $pp \to W^{\pm} + X$ production at $\sqrt{s} = 8$ TeV*, *Eur. Phys. J. C* **76** (2016) 469 [arXiv:1603.01803] [INSPIRE].

[74] LHCb collaboration, *Measurement of forward W and Z boson production in pp collisions at $\sqrt{s} = 8$ TeV*, *JHEP* **01** (2016) 155 [arXiv:1511.08039] [INSPIRE].

[75] ATLAS collaboration, *Measurement of the transverse momentum and ϕ^*_η distributions of Drell-Yan lepton pairs in proton-proton collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector*, *Eur. Phys. J. C* **76** (2016) 291 [arXiv:1512.02192] [INSPIRE].

[76] CMS collaboration, *Measurement of the Ratio of Inclusive Jet Cross Sections using the Anti-k_T Algorithm with Radius Parameters $R = 0.5$ and 0.7 in pp Collisions at $\sqrt{s} = 7$ TeV*, *Phys. Rev. D* **90** (2014) 072006 [arXiv:1406.0324] [INSPIRE].

[77] ATLAS collaboration, *Measurement of the inclusive jet cross-section in proton-proton collisions at $\sqrt{s} = 7$ TeV using 4.5 fb^{-1} of data with the ATLAS detector*, *JHEP* **02** (2015) 153 [*Erratum ibid.* **09** (2015) 141] [arXiv:1410.8857] [INSPIRE].

[78] CMS collaboration, *Measurement and QCD analysis of double-differential inclusive jet cross sections in pp collisions at $\sqrt{s} = 8$ TeV and cross section ratios to 2.76 and 7 TeV*, *JHEP* **03** (2017) 156 [arXiv:1609.05331] [INSPIRE].

[79] CMS collaboration, *Measurement of double-differential cross sections for top quark pair production in pp collisions at $\sqrt{s} = 8$ TeV and impact on parton distribution functions*, *Eur. Phys. J. C* **77** (2017) 459 [arXiv:1703.01630] [INSPIRE].

[80] ATLAS collaboration, *Measurements of top-quark pair differential cross-sections in the lepton+jets channel in pp collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector*, *Eur. Phys. J. C* **76** (2016) 538 [arXiv:1511.04716] [INSPIRE].

[81] S. Dulat et al., *New parton distribution functions from a global analysis of quantum chromodynamics*, *Phys. Rev. D* **93** (2016) 033006 [arXiv:1506.07443] [INSPIRE].

[82] H.-L. Lai et al., *New parton distributions for collider physics*, *Phys. Rev. D* **82** (2010) 074024 [arXiv:1007.2241] [INSPIRE].

[83] C. Anastasiou, C. Duhr, F. Dulat, F. Herzog and B. Mistlberger, *Higgs Boson Gluon-Fusion Production in QCD at Three Loops*, *Phys. Rev. Lett.* **114** (2015) 212001 [arXiv:1503.06056] [INSPIRE].

JHEP08(2022)088

[84] E.L. Berger, J. Gao, C.S. Li, Z.L. Liu and H.X. Zhu, *Charm-Quark Production in Deep-Inelastic Neutrino Scattering at Next-to-Next-to-Leading Order in QCD*, *Phys. Rev. Lett.* **116** (2016) 212002 [arXiv:1601.05430] [INSPIRE].

[85] J. Gao, *Massive charged-current coefficient functions in deep-inelastic scattering at NNLO and impact on strange-quark distributions*, *JHEP* **02** (2018) 026 [arXiv:1710.04258] [INSPIRE].

[86] J. Gao, T.J. Hobbs, P.M. Nadolsky, C. Sun and C.P. Yuan, *General heavy-flavor mass scheme for charged-current DIS at NNLO and beyond*, *Phys. Rev. D* **105** (2022) L011503 [arXiv:2107.00460] [INSPIRE].

[87] S. Alekhin et al., *Determination of Strange Sea Quark Distributions from Fixed-target and Collider Data*, *Phys. Rev. D* **91** (2015) 094002 [arXiv:1404.6469] [INSPIRE].

[88] F. Faura, S. Iranipour, E.R. Nocera, J. Rojo and M. Ubiali, *The Strangest Proton?*, *Eur. Phys. J. C* **80** (2020) 1168 [arXiv:2009.00014] [INSPIRE].

[89] ATLAS collaboration, *Determination of the strange quark density of the proton from ATLAS measurements of the $W \to \ell\nu$ and $Z \to \ell\ell$ cross sections*, *Phys. Rev. Lett.* **109** (2012) 012001 [arXiv:1203.4051] [INSPIRE].

[90] ATLAS collaboration, *Precision measurement and interpretation of inclusive W^+, W^- and Z/γ^* production cross sections with the ATLAS detector*, *Eur. Phys. J. C* **77** (2017) 367 [arXiv:1612.03016] [INSPIRE].

[91] ATLAS collaboration, *Measurement of the double-differential high-mass Drell-Yan cross section in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector*, *JHEP* **08** (2016) 009 [arXiv:1606.01736] [INSPIRE].

[92] T. Kluge, K. Rabbertz and M. Wobisch, *FastNLO: Fast pQCD calculations for PDF fits*, in *14th International Workshop on Deep Inelastic Scattering*, Tsukuba Japan, April 20–24 2006, pp. 483–486 [DOI] [hep-ph/0609285] [INSPIRE].

[93] L.A. Harland-Lang, A.D. Martin, P. Motylinski and R.S. Thorne, *Parton distributions in the LHC era: MMHT 2014 PDFs*, *Eur. Phys. J. C* **75** (2015) 204 [arXiv:1412.3989] [INSPIRE].

[94] NNPDF collaboration, *Parton distributions from high-precision collider data*, *Eur. Phys. J. C* **77** (2017) 663 [arXiv:1706.00428] [INSPIRE].

[95] A. Buckley et al., *LHAPDF6: parton density access in the LHC precision era*, *Eur. Phys. J. C* **75** (2015) 132 [arXiv:1412.7420] [INSPIRE].

[96] FERMILAB E531 collaboration, *Cross-sections for Neutrino Production of Charmed Particles*, *Phys. Lett. B* **206** (1988) 375 [INSPIRE].

JHEP08(2022)088