# Mapping the sensitivity of hadronic experiments to nucleon structure

Bo-Ting Wang,[1,*] T. J. Hobbs,[1,4,†] Sean Doyle,[1] Jun Gao,[2] Tie-Jiun Hou,[3] Pavel M. Nadolsky,[1,‡] and Fredrick I. Olness[1]

[1] *Department of Physics, Southern Methodist University, Dallas, Texas 75275-0181, USA*
[2] *INPAC, Shanghai Key Laboratory for Particle Physics and Cosmology,*
*School of Physics and Astronomy, Shanghai Jiao-Tong University, Shanghai 200240, China*
[3] *School of Physics Science and Technology, Xinjiang University, Urumqi, Xinjiang 830046 China*
[4] *Jefferson Lab, EIC Center, Newport News, Virginia 23606, USA*

(Received 10 July 2018; revised manuscript received 16 October 2018; published 28 November 2018)

Determinations of the proton's collinear parton distribution functions (PDFs) are emerging with growing precision due to increased experimental activity at facilities like the Large Hadron Collider. While this copious information is valuable, the speed at which it is released makes it difficult to quickly assess its impact on the PDFs, short of performing computationally expensive global fits. As an alternative, we explore new methods for quantifying the potential impact of experimental data on the extraction of proton PDFs. Our approach relies crucially on the Hessian correlation between theory-data residuals and the PDFs themselves, as well as on a newly defined quantity—the *sensitivity*—which represents an extension of the correlation and reflects both PDF-driven and experimental uncertainties. This approach is realized in a new, publicly available analysis package PDFSense, which operates with these statistical measures to identify particularly sensitive experiments, weigh their relative or potential impact on PDFs, and visualize their detailed distributions in a space of the parton momentum fraction $x$ and factorization scale $\mu$. This tool offers a new means of understanding the influence of individual measurements in existing fits as well as a predictive device for directing future fits toward the highest impact data and assumptions. Along the way, many new physics insights can be gained or reinforced. As one of many examples, PDFSense is employed to rank the projected impact of new LHC measurements in jet, vector boson, and $t\bar{t}$ production and leads us to the conclusion that inclusive jet production at the LHC has a potential for playing an indispensable role in future PDF fits. These conclusions are independently verified by preliminarily fitting this experimental information and investigating the constraints they supply using the Lagrange multiplier technique.

DOI: 10.1103/PhysRevD.98.094030

## I. INTRODUCTION

The determination of collinear parton distribution functions (PDFs) of the nucleon is becoming an increasingly precise discipline with the advent of high-luminosity experiments at both colliders and fixed-target facilities. Several research groups are involved in the rich research domain of the modern PDF analysis [1–7]. By quantifying the distribution of a parent hadron's longitudinal momentum among its constituent quarks and gluons, PDFs offer both a description of the hadronic structure and an essential ingredient of perturbative QCD computations. PDFs enjoy a symbiotic relationship with high-energy experimental data, in the sense that they are crucial for understanding hadronic collisions in the Standard Model (SM) and beyond, while reciprocally benefiting from a wealth of high-energy data that constrain the PDFs. In fact, since the start of the Large Hadron Collider Run II (LHC Run II), the volume of experimental data pertinent to the PDFs is growing with such speed that keeping pace with the rapidly expanding data sets and isolating measurements of greatest impact present a significant challenge for PDF fitters. This paper intends to meet this challenge by presenting a method for identifying high-value experiments which constrain the PDFs and the resulting SM predictions that depend on them.

That such expansive data sets can constrain the PDFs is a consequence of the latter's universality—a feature which relies upon QCD factorization theorems to separate the inherently nonperturbative PDFs (at long distances) from process-dependent, short-distance matrix elements.

[*] botingw@mail.smu.edu
[†] tjhobbs@smu.edu
[‡] nadolsky@physics.smu.edu

For instance, the cross section for inclusive single-particle hadroproduction (of, e.g., a weak gauge boson $W/Z$) in proton-proton collisions at the LHC is directly sensitive to the nucleon PDFs via an expression of the form

$$\sigma(AB \to W/Z + X) = \sum_n \alpha_s^n(\mu_R^2) \sum_{a,b} \int dx_a dx_b \times f_{a/A}(x_a,\mu^2)\hat{\sigma}^{(n)}_{ab\to W/Z+X}(\hat{s},\mu^2,\mu_R^2) f_{b/B}(x_b,\mu^2), \qquad (1)$$

in which $f_{a/A}(x_a,\mu^2)$ represents the PDF for a parton of flavor $f_a$ carrying a fraction $x_a$ of the 4-momentum of proton $p_A$ at a factorization scale $\mu$; the $n$th-order hard matrix element is denoted by $\hat{\sigma}^{(n)}_{ab\to W/Z+X}(\hat{s},\mu^2,\mu_R^2)$ and is dependent upon the partonic center-of-mass energy $\hat{s} = x_a x_b s$, in which $s$ is the center-of-mass energy of the initial hadronic system; and $\mu_R$ is the renormalization scale in the QCD coupling strength $\alpha_s(\mu_R)$. In Eq. (1), subleading corrections $\sim\Lambda^2/M^4_{W/Z}$ have been omitted, and we emphasize that factorization theorems like Eq. (1) have been proven to arbitrary order in $\alpha_s$ for essential observables in the global PDF analysis, such as the inclusive cross sections in deep-inelastic scattering (DIS) and Drell-Yan processes. For compactness and generality, we shall refer henceforth to a PDF for the parton of flavor $f$ simply as $f(x,\mu)$.

Given this formalism, one is confronted with the problem of finding those experiments that provide reliable new information about the PDF behavior. With the proliferation of potentially informative new data, incorporating them all into a global QCD fit inevitably incurs significant cost both in terms of computational resources and required fitting time. Indeed, tremendous progress in the precision of PDFs and robustness of SM predictions is driven by the technology for performing global analysis that has vastly grown in complexity and sophistication. Nowadays, the state of the art in perturbative QCD (pQCD) treatments are done at next-to-next-to-leading order (NNLO) [and increasingly even next-to-next-to-next-to-leading order (N$^3$LO)], and advanced statistical techniques are commonly employed in PDF error estimation. The magnitude of this subject is vast, and we refer the interested reader to Refs. [8,9] for comprehensive reviews. The tradeoff of this progress is that the impact of an experiment on the ultimate PDF uncertainty is often hard to foresee without doing a complicated fit. Various publications claim sensitivity of new experiments to the PDFs. In this paper, we look into these claims using statistical techniques that bypass doing the fits and with an eye on theoretical, experimental, and methodological components relevant at the NNLO precision.

The potential cost is steepened by the large size of the global data sets usually involved. This point can be seen in Fig. 1, which plots the default data set considered in the present analysis in a space of partonic momentum fraction $x$ and factorization scale $\mu$. We label these data as the "CTEQ-TEA set," given that it is an extension of the 3287 raw data points (given by the sum over $N_{\rm pt}$ in Tables II and III) treated in the NNLO CT14HERA2 analysis of Ref. [10], now augmented by the inclusion of 734 raw data points (given by the sum over $N_{\rm pt}$ in Table IV) from more recent LHC data. These raw measurements can ultimately be mapped to 5227 typical $\{x,\mu\}$ values in Fig. 1, such that each symbol corresponds to a data point from an experiment shown in the legend, at the approximate $x$ and $\mu$ values characterizing the data point as described in Appendix A. The experiments are labeled by a short-hand name which includes the year of final publication (e.g., "HERAI + II'15"—corresponding to the 2015 combined HERA Run I and Run II data), following the translation key also given in Tables II–IV of Appendix B. The experiments included in the CT14HERA2 analysis are listed in the left column and upper part of the right column of the legend, while the newer LHC data considered for the upcoming CTEQ-TEA analysis are the last 14 entries of the right column.

The growing complexity of PDF fitting stimulates development of less computationally involved approaches to estimate the impact of new experimental data on full global fits, such as Hessian profiling techniques [51] and Bayesian reweighting [52,53] of PDFs. Although these approaches do simulate the expansion of a particular global fit by including (a) theretofore absent data set(s), they are also limited in that the interpretation of their outcomes is married to the specific PDF parametrization and definition of PDF errors. For example, conclusions obtained by PDF reweighting regarding the importance of a given data set strongly depend on the assumed statistical tolerance or the choice of reweighting factors [54,55].

Parallel to these efforts, the notion of using correlations between the PDF uncertainties of two physical observables was proposed in Refs. [56,57] as a means of quantifying the degree to which these quantities were related based upon their underlying PDFs. The PDF-mediated correlation $C_f$ in this case, which we define in Sec. III A, embodies the Pearson correlation coefficient computed by a generalization of the "master formula" [58] for the Hessian PDF uncertainty. The Hessian correlation was deployed extensively in Ref. [59] to explore implications of the CTEQ6.6 PDFs for envisioned LHC observables. It proved to be instrumental for identifying the specific PDF flavors and $x$ ranges most correlated with the PDF uncertainties for $W$, $Z$, $H$, and $t\bar{t}$ production cross sections as well as other processes. The Pearson correlation coefficient has also proven to be informative in the approach based on Monte Carlo PDF replicas; see, e.g., Refs. [60,61]. However, the PDF-mediated correlation with a theoretical cross section is only partly indicative of the sensitivity of the experiment. The constraining power of the experiment also depends on the size of experimental errors that were not normally

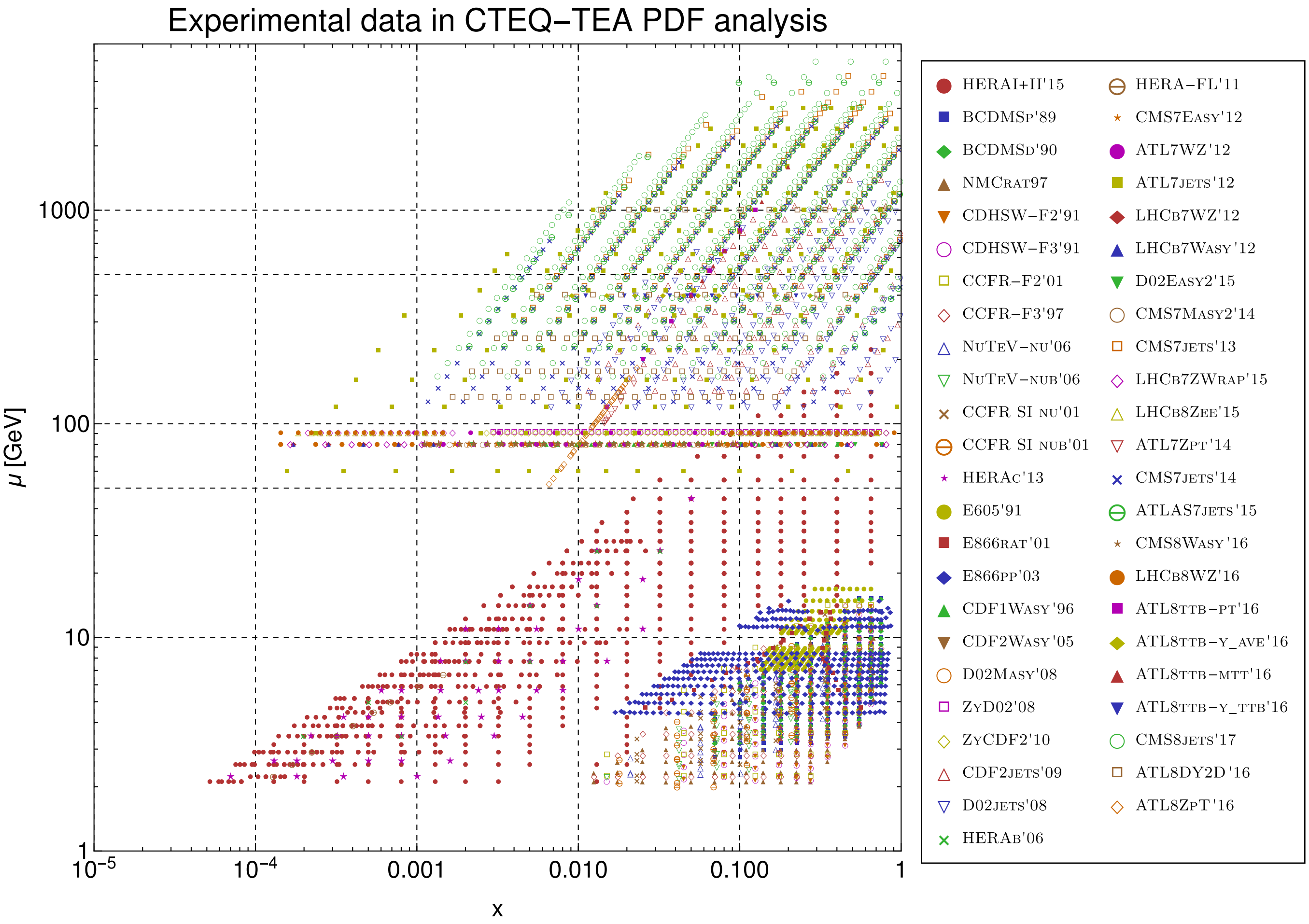

FIG. 1. A graphical representation of the space of $\{x, \mu\}$ points probed by the full data set treated in the present analysis, designated as CTEQ-TEA. It represents an expansion to include newer LHC data of the CT14HERA2 data set [10] fitted in the most recent CT14 framework [1], which involved measurements from Run II of HERA [6]. Details of the data sets corresponding to the short-hand names given in the legend may be found in Tables II–IV. The HERA combined data set HERAI + II'15 consists of both neutral-current (NC) and charge-current (CC) scattering events.

considered in correlation studies, as well as on correlated systematic effects that are increasingly important.

As a remedy to these limitations, we introduce a new format for the output of CTEQ-TEA fits and a natural extension of the correlation technique to quantify the sensitivity of any given experimental data point to a PDF-dependent observable of the user's choice. In this approach, we work with *statistical residuals* quantifying the goodness of fit to individual data points. We demonstrate that the complete set of residuals computed for Hessian PDF sets characterizes the CTEQ-TEA fit well enough to permit a means of gauging the influence of empirical information on PDFs in a fashion that does not require complete refits.

A generalization of the PDF-mediated correlations called the *sensitivity* $S_f$—to be characterized in detail in Sec. III B—better identifies those experimental data points that tightly constrain PDFs by both merit of their inherent precision and their ability to discriminate among PDF error fluctuations. Such an approach aids in identifying regions of $\{x, \mu\}$ for which PDFs are particularly constrained by physical observables.

In fact, in the numerical approach presented in the forthcoming sections, the user can quantify the sensitivity of data not only to individual PDF flavors but even to specific physical observables, including the modifications due to correlated systematic uncertainties in every experiment of the CT14HERA2 analysis. For example, for Higgs boson production via gluon fusion ($gg \to H$) at the LHC 14 TeV, the short-distance cross sections are known up to N$^3$LO with a scale uncertainty of about 3% [62]. It has been suggested that $t\bar{t}$ production and high-$p_T$ $Z$ boson production on their own constrain the gluon PDF in the $x$ region sensitive to the LHC Higgs production and that these are comparable to the constraints from LHC and Tevatron data [63,64]. Verifying the degree to which this hypothesis is true has been difficult without actually including all these data in a fit.

As an alternative to doing a full global fit, we can critically assess this supposition in the context of the entire global data set of Fig. 1 using the Hessian correlations and sensitivities, $|C_f|$ and $|S_f|$. The detailed procedure is explained in Secs. III A and III B. In the example at hand, we could rely on the established wisdom that the theoretical cross sections that have an especially large correlation with $\sigma_{H^0}$ may constrain the PDF dependence of $\sigma_{H^0}$, say, when $|C_f| \gtrsim 0.7$ [59]. Along this reasoning, the left frame in Fig. 2 illustrates 310 experimental data points in $\{x, \mu\}$ space that have the highest absolute correlation, $|C_f|$, between the point's statistical residual defined in Sec. III A and the cross section $\sigma_{H^0}$ at 14 TeV via the CT14HERA2 NNLO PDFs. To locate such points in the figure, we highlighted them with color according to the convention shown on the color scale to the right. The respective $|C_f|$ for the highlighted data points ranges between 0.42 and 1. The rest of the data points have smaller correlations and are shown in gray.

We find that the 310 data points with the highest correlation for $\sigma_{H^0}$ belong to 20 experiments. Nearly all of them are contributed by HERA neutral current (NC) DIS, LHC and Tevatron jet production, and HERA charm production. Some of the data points with highest $|C_f|$ come from high-$p_T$ $Z$ boson and even $t\bar{t}$ production.

The correlations $C_f$, however, do not reflect the experimental uncertainties, which vary widely across the experiments. In the left panel of Fig. 2, fewer than 30 points have a strong correlation of 0.7 or more, but more data points impose relevant constraints in the global fit. To include the information about the uncorrelated and correlated experimental errors, in the right panel of Fig. 2, we plot the distributions of 310 data points with the highest sensitivity parameter $S_f$, which more faithfully reproduces the actual constraints during the fitting. In general, we find substantial differences between the $C_f$ and $S_f$ distributions. Even the most significant correlations, of order $|C_f| \sim 0.7$ and above, do not guarantee a significant contribution of the experimental point to the log-likelihood $\chi^2$ if the errors are large. On the other hand, we argue that $|S_f|$ is closely related to a contribution of the data point to $\chi^2$. According to the distribution in the right figure, the 310 data points with the highest sensitivity $|S_f|$ to $\sigma_{H^0}(14\text{TeV})$ arise from 27 experiments. Among these data points, only some have a

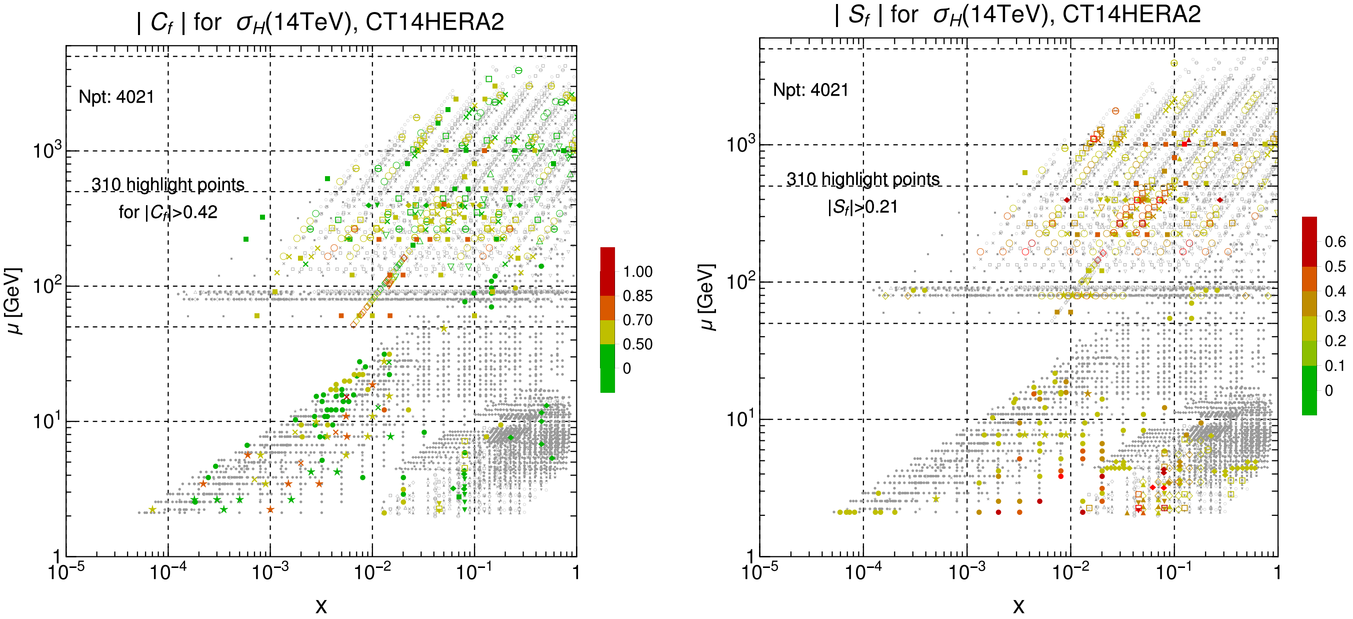

FIG. 2. For the full CTEQ-TEA data set of Fig. 1, we show the absolute correlation $|C_f|$ and sensitivity $|S_f|$ associated with the 14 TeV Higgs production cross section $\sigma_{H^0}(14\text{ TeV})$. 310 input data points with most significant magnitudes of $|C_f|$ and $|S_f|$ are highlighted with color. When only the $|C_f|$ plot is considered, only a very small subpopulation of jet production data (diagonal open circles and closed squares with $\mu \gtrsim 100$ GeV) exhibits significant correlations with $|C_f| > 0.7$ (orange and red colors), as well as some HERA DIS, high-$p_T$ $Z$ boson, and $t\bar{t}$ production data points. Our novel definition for the sensitivity in the right panel, on the other hand, reveals more points that have comparable potency for constraining the Higgs cross section. In this case, a larger fraction of the jet production points is important (especially CMS measurements of CMS8jets'17 and CMS7jets'14), as well as a number of other processes at smaller $\mu$, particularly DIS data from HERA, BCDMS, NMC, CDHSW, and CCFR (experiments HERAI + II'15, BCDMSd'90, NMCrat'97, CDHSW-F2'91, CCFR-F2'01, and CCFR-F3'97). Although its cumulative impact is comparatively modest, ATLAS $t\bar{t}$ production data (ATL8ttb-pt'16, ATL8ttb-y_ave'16, ATL8ttb-mtt'16, and ATL8ttb-y_ttb'16) register significant per-point sensitivities, as do E866 $pp$ Drell-Yan pair production (E866 pp'03), LHCb $W$, $Z$ production (LHCb7ZWrap'15 and LHCb8WZ'16), and charge lepton asymmetries at D0 and CMS (D02Masy'08, CMS7Masy2'14, and CMS7Easy'12). Similarly, some of the high-$p_T$ $Z$ production information (ATL7ZpT'14 and ATL8ZpT'16) from ATLAS provide modest constraints.

large correlation $|C_f|$ with $\sigma_{H^0}(14\text{TeV})$. Nonetheless, they have medium-to-large sensitivity, $|S_f| > 0.21$, according to the criterion developed in Sec. III B. We stress that, while one might suggests plausible dynamical reasons why certain experiments might be particularly sensitive to Higgs production via the gluon PDF, (e.g., via the leading-order $qg$ and $gg$ hard cross sections in jet production and DGLAP scaling violations in inclusive DIS), this reasoning alone does not predict the actual sensitivity revealed by $S_f$ in the presence of multiple experimental constraints.

As one noticeable difference from the $|C_f|$ figure, while inclusive DIS at HERA continues to contribute a large number of data points (about 80) with a high $|S_f|$, also the fixed-target DIS experiments (BCDMS, NMC, CDHSW, and CCFR) contribute about the same number of sensitive points in the right panel that were not identified by large correlations. Other sensitive points belong to the jet production data sets from ATLAS and CMS and some vector boson production experiments (muon charge asymmetries at D0 and CMS, E866 low-energy Drell-Yan production, and LHCb 7 TeV $W$ and $Z$ cross sections).

On the other hand, HERA charm production and ATLAS 7/8 TeV high-$p_T$ $Z$ production have suppressed sensitivities despite their large correlations, reflecting the larger experimental uncertainties in these measurements. While the LHC $t\bar{t}$ production experiments have large per-point sensitivities, they contribute relatively little to $\chi^2$ because of their small total number of data points. From this comparison, one finds, perhaps somewhat unexpectedly, that fixed-target DIS experiments impose important constraints on $\sigma_{H^0}(14\text{TeV})$, thus complementing the HERA inclusive DIS data. One would conclude that efforts to constrain PDF-based SM predictions for Higgs production by relying only on a few points of $t\bar{t}$ data, but to the neglect of high-energy jet production points, would be significantly handicapped by the absence of the latter. We will return to this example in Sec. IV.

The discriminating power of a sensitivity-based analysis therefore forms the primary motivation for this work, and we present the attendant details below. To assess information about the PDFs encapsulated in the residuals for large collections of hadronic data implemented in the CTEQ-TEA global analysis, we make available a new statistical package PDFSENSE to map the regions of partonic momentum fractions $x$ and QCD factorization scales $\mu$ where the experiments impose strong constraints on the PDFs. In companion studies, we have applied PDFSENSE to select new data sets for the next generation of the CTEQ-TEA global analysis, to quantitatively explore the physics potential for constraining the PDFs at a future Electron-Ion Collider [65–68] and Large Hadron-Electron Collider [69], and to investigate the potential of high-energy data to inform lattice-calculable quantities [70] like the Mellin moments of structure functions [71] and quark quasidistributions [72]. We reserve many instructive results for follow-up publications currently in preparation, while presenting select calculations in this article to demonstrate the power of the method. We find that the sensitivity technique generally agrees with the preliminary CTEQ-TEA fits and Hessian reweighting realized in the EPUMP program [73]. However, assessing the sensitivity is much simpler than doing the global fit. It does not require access to a fitting program or the application of (potentially subtle) PDF reweighting techniques.

The remainder of the article proceeds as follows. Pertinent aspects of the PDFs and their standard determination via QCD global analyses are summarized in Sec. II. Then, we introduce *normalized residual variations* to extract, visualize, and quantify statistical information about the global QCD fit. In Sec. III, we construct a number of statistical quantities that characterize the PDF constraints in the global analysis using the residual variations. In Sec. IV, we apply the thus constructed sensitivity parameter to examine the impact of various CTEQ-TEA data sets on extractions of the gluon PDF $g(x,\mu)$. In this section and in the conclusion contained in Sec. V, we emphasize a number of *physics insights* that we obtained by applying our sensitivity analysis techniques. Additional aspects of the technique and supplementary tables are reserved for Appendixes A and B, respectively.

## II. PDF PRELIMINARIES

### A. Data residuals in a global QCD analysis

While various theoretical models exist for computing nucleon PDFs [74–76], unambiguous evaluation of the PDFs entirely in terms of QCD theory is not yet possible due to the fact that the PDFs can in general receive substantial nonperturbative contributions at infrared momenta. For this reason, precise PDF determination has proceeded mainly through the technique of the QCD global analysis—a method enabled by QCD factorization and PDF universality.

In this approach, a highly flexible parametric form is ascribed for the various flavors in a given analysis at a relatively low scale $Q_0^2$. For example, one might take the input PDF for a given quark flavor $f$ to be a parametric form,

$$f(x,\mu^2 = Q_0^2) = A_{f,0}x^{A_{f,1}}(1-x)^{A_{f,2}}F(x;A_{f,3},\ldots), \quad (2)$$

in which $F(x;A_{f,3},\ldots)$ can be a suitable polynomial function, e.g., a Chebyshev or Bernstein polynomial, or replaced with a feed-forward neural network $\text{NN}_f(x)$ as in the NNPDF approach. While the full statistical theory for PDF determination and error quantification is beyond the intended range of this analysis, roughly speaking, a best fit is found for a vector $\vec{A}$ of $N$ PDF parameters $A_l$ by minimizing a goodness-of-fit function $\chi^2$ describing agreement of the QCD data and physical observables computed in terms of the PDFs. Based on the behavior of $\chi^2$ in the neighborhood of the global minimum, it is then possible to construct an ensemble of error PDFs to quantify uncertainties of PDFs at a predetermined probability level.

There are various ways to evaluate uncertainties on PDFs, e.g., the Hessian [56,58], the Monte Carlo [77,78], and the Lagrange multiplier approaches [79]. In this analysis, our default PDF input set is CT14HERA2, which uses the Hessian method to estimate uncertainties and is therefore based on the quadratic assumption for $\chi^2(\vec{A})$ in the vicinity of the global minimum. In the Hessian method, an orthonormal basis of PDF parameters $\vec{a}$ is derived from the input PDF parameters $\vec{A}$ by the diagonalization of a Hessian matrix $H$, which encodes the second-order derivatives of $\chi^2$ with respect to $A_l$. The eigenvector PDF combinations $\vec{a}_l^{\pm}$ are found for two extreme variations from the best-fit vector $\vec{a}_0$ along the direction of the $l$th eigenvector of $H$ allowed at a given probability level. The uncertainty on a QCD observable $X$ can then be estimated with one of the available master formulas [57,58], the "symmetric" variety of which is

$$\Delta X = \frac{1}{2}\sqrt{\sum_{l=1}^{N}(X_l^+ - X_l^-)^2}. \tag{3}$$

In the CTEQ-TEA global analysis, the $\chi^2$ function accounts for multiple sources of experimental uncertainties, as well as for some prior theoretical constraints on the $a_l$ parameters. Consequently, the global $\chi^2$ function takes the form

$$\chi^2_{\rm global} = \sum_E \chi^2_E + \chi^2_{\rm th}, \tag{4}$$

where the sum runs over all experimental data sets ($E$) and $\chi^2_{\rm th}$ imposes theoretical constraints. The complete formulas for $\chi^2_E$ and $\chi^2_{\rm th}$ can be found in Ref. [80]. For the purposes of this paper, we express $\chi^2_E$ for each experiment $E$ in a compact form as a sum of squared *shifted residuals* $r_i^2(\vec{a})$, which are summed over $N_{\rm pt}$ individual data points $i$ in this experiment, as well as the contributions of $N_\lambda$ best-fit nuisance parameters $\bar{\lambda}_\alpha$ associated with correlated systematic errors:

$$\chi^2_E(\vec{a}) = \sum_{i=1}^{N_{\rm pt}} r_i^2(\vec{a}) + \sum_{\alpha=1}^{N_\lambda} \bar{\lambda}^2_\alpha(\vec{a}). \tag{5}$$

In turn, $r_i(\vec{a})$ for the $i$th data point is constructed from the theoretical prediction $T_i(\vec{a})$ evaluated in terms of PDFs, total uncorrelated uncertainty $s_i$, and the shifted central data value $D_{i,\rm sh}(\vec{a})$:

$$r_i(\vec{a}) = \frac{1}{s_i}(T_i(\vec{a}) - D_{i,\rm sh}(\vec{a})). \tag{6}$$

This representation arises in the Hessian formalism due to the presence of correlated systematic errors in many experimental data sets, which require $\chi^2_E$ to depend on nuisance parameters $\lambda_\alpha$. This is in addition to the dependence of $\chi^2_E$ on the PDF parameters $\vec{a}$ and theoretical parameters such as $\alpha_s(M_Z)$ and particle masses. The $\lambda_\alpha$ parameters are optimized for each $\vec{a}$ according to the analytic solution derived in Appendix B of Ref. [58]. Optimization effectively shifts the central value $D_i$ of the data point by an amount determined by the optimal nuisance parameters $\bar{\lambda}_\alpha(\vec{a})$ and the correlated systematic errors $\beta_{i\alpha}$:

$$D_i \to D_{i,\rm sh}(\vec{a}) = D_i - \sum_{\alpha=1}^{N_\lambda} \beta_{i\alpha}\bar{\lambda}_\alpha(\vec{a}). \tag{7}$$

It should be noted that the contribution of the squared best-fit nuisance parameters to $\chi^2_E$ in Eq. (5) is dominated in general by the first term involving the shifted residuals, which tends to be much larger—especially for more sizable data sets.

We point out also that some alternative representations for $\chi^2$ include the correlated systematic errors via a covariance matrix $(\mathrm{cov})_{ij}$, rather than the above-mentioned CTEQ-preferred form that explicitly operates with $\lambda_\alpha$. Various $\chi^2$ definitions in use are reviewed in Ref. [81], as well as in Ref. [7]. Crucially, however, the representations based upon operating with $\lambda_\alpha$ and $(\mathrm{cov})_{ij}$ are derivable from each other [80]. From an extension of the derivation in Ref. [58], we may relate the shifted residual to the covariance matrix at an $i$th point and optimal nuisance parameters as

$$r_i(\vec{a}) = s_i \sum_{j=1}^{N_{\rm pt}} (\mathrm{cov}^{-1})_{ij}(T_j(\vec{a}) - D_j), \tag{8}$$

$$\bar{\lambda}_\alpha(\vec{a}) = \sum_{i,j=1}^{N_{\rm pt}} (\mathrm{cov}^{-1})_{ij}\frac{\beta_{i\alpha}}{s_i}\frac{(T_j(\vec{a}) - D_j)}{s_j}, \tag{9}$$

where

$$(\mathrm{cov}^{-1})_{ij} = \left[\frac{\delta_{ij}}{s_i^2} - \sum_{\alpha,\beta=1}^{N_\lambda}\frac{\beta_{i\alpha}}{s_i^2}A^{-1}_{\alpha\beta}\frac{\beta_{j\beta}}{s_j^2}\right], \tag{10}$$

and

$$A_{\alpha\beta} = \delta_{\alpha\beta} + \sum_{k=1}^{N_{\rm pt}}\frac{\beta_{k\alpha}\beta_{k\beta}}{s_k^2}. \tag{11}$$

Thus, even for those PDF analyses which operate with the covariance matrix, one is still able to determine the shifted residuals $r_i$ from $(\mathrm{cov}^{-1})_{ij}$ using Eq. (8). In this article, we conveniently follow the CTEQ methodology and obtain

$r_i(\vec{a})$ directly from the CTEQ-TEA fitting program, together with the optimal nuisance parameters $\bar{\lambda}_\alpha(\vec{a})$ and shifted central data values $D_{i,\mathrm{sh}}(\vec{a})$.

### B. Visualization of the global fit with the help of residuals

The shifted residuals $r_i$ draw our interest because, in consequence of the definitions in Eqs. (5) and (6), they contain substantial low-level information about the agreement of PDFs with every data point in the global QCD fit in the presence of systematic shifts. The response of $r_i(\vec{a})$ to the variations in PDFs depends on the experiment type and kinematic range associated with the $i$th data point, and the totality of these responses can be examined with modern data-analytical methods. The sum of squared residuals over all points of the global data set renders the bulk of the log-likelihood, or experimental, component $\chi_E^2$ of the global $\chi^2$. In turn, the root-mean-squared residual $\langle r_0\rangle_E$ for experiment $E$ and the central PDF set $\vec{a}_0$ is tied to $\chi_E^2(\vec{a}_0)/N_{\mathrm{pt}}$, the standard measure of agreement with experiment $E$ at the best fit:

$$\langle r_0\rangle_E \equiv \sqrt{\frac{1}{N_{\mathrm{pt}}}\sum_{i=1}^{N_{\mathrm{pt}}} r_i^2(\vec{a}_0)} = \sqrt{\frac{1}{N_{\mathrm{pt}}}\left(\chi_E^2(\vec{a}_0) - \sum_{\alpha=1}^{N_\lambda}\bar{\lambda}_\alpha^2(\vec{a}_0)\right)} \approx \sqrt{\frac{\chi_E^2(\vec{a}_0)}{N_{\mathrm{pt}}}}. \tag{12}$$

Notice that $\langle r_0\rangle_E \approx 1$ when the fit to the experimental data set $E$ is good.

We will now invoke the Hessian formalism to first organize the analysis of the PDF dependence of individual residuals and then introduce a framework to evaluate sensitivity of individual data points to PDF-dependent physical observables. To test the effectiveness of the proposed method, we study constraints using CT14HERA2 parton distributions [10] fitted to data sets from DIS processes, $Z \to l^+l^-$, $d\sigma/dy_l$, $W \to l\nu$, and jet production ($p_1p_2 \to jjX$) We include both the experiments that were used to construct the CT14HERA2 data set as well as a number of LHC experiments that may be fitted in the future. The experimental data sets are summarized in Tables II–IV.

Given the urgency in improving constraints on the gluon PDF for investigations of the Higgs sector, we focus attention on several candidate experiments that may probe $g(x,\mu)$: high-$p_T$ $Z$-boson production (ATL8ZpT'16 and ATL7ZpT'14), $t\bar{t}$ production (ATL8ttb-pt'16, ATL8ttb-y_ave'16, ATL8ttb-mtt'16, and ATL8ttb-y_ttb'16), as well as high-luminosity or alternative data sets for jet production, such as the high-luminosity ATLAS 7 TeV jet data (ATLAS7jets'15) that are to replace the counterpart low-luminosity set ATL7jets'12, or the CMS 7 TeV jet data set (CMS7jets'14) that extends to lower jet $p_T$ and higher rapidity, $2.5 < |y_j| < 3$, than the previously fitted CMS 7 TeV jet data set (CMS7jets'13).[1] The dependence of such experiments on $g(x,\mu)$ is scrutinized in a number of ways. We examine their statistical properties using both the PDFs from the CT14HERA2 NNLO analysis, which already impose significant constraints on the large-$x$ gluon using the Tevatron inclusive jet data sets, CDF2jets'09 and D02jets'08, and in some comparisons using a special version of the NNLO PDFs that are fitted to the same CT14HERA2 data set, except without including the above jet data sets. For yet another aspect, we investigate a range of measurements of Drell-Yan pair production cross sections and charge lepton asymmetries with the goal of understandomg their sensitivity predominantly to the (anti)quark sector.

To parametrize the response of a residual $\vec{r}_i$, we evaluate it for every eigenvector PDF $\vec{a}_l^\pm$ of the CT14HERA2 PDF set with $N = 28$ PDF parameters. Then, given the normalized residual variations,

$$\delta_{i,l}^\pm \equiv (r_i(\vec{a}_l^\pm) - r_i(\vec{a}_0))/\langle r_0\rangle_E, \tag{13}$$

between the residuals for the PDF eigenvectors $\vec{a}_l^\pm$ and for the CT14HERA2 central PDF $\vec{a}_0$, we construct a $2N$-dimensional vector

$$\vec{\delta}_i = \{\delta_{i,1}^+, \delta_{i,1}^-, \ldots, \delta_{i,N}^+, \delta_{i,N}^-\} \tag{14}$$

for each data point of the global data set.

The components of $\vec{\delta}_i$ parametrize responses of $r_i$ to PDF variations along the independent directions given by $\vec{a}_l^\pm$. The differences are normalized to the central rms residual $\langle r_0\rangle_E$ of experiment $E$ [see Eq. (12)] so that the normalized residual variations do not significantly depend on $\chi^2(\vec{a}_0)/N_{\mathrm{pt}}$, the quality of fit to experiment $E$. Recall that a substantial spread over the fitted experiments is generally obtained for $\chi_E^2/N_{\mathrm{pt}}$. Moreover, it is reasonable to expect significantly larger values for $\chi_E^2/N_{\mathrm{pt}}$ for the experiments that have not been yet fitted but are included in the analysis of the residuals, e.g., the new LHC experiments shown in Fig. 1. With the definitions in Eqs. (13) and (14), however, $\vec{\delta}_i$ is only weakly sensitive to $\chi_E^2/N_{\mathrm{pt}}$.

Thus, we represent the PDF-driven variations of the residuals of a global data set by a bundle of vectors $\vec{\delta}_i$ in a $2N$-dimensional space.[2] This mapping opens the door to applying various data-analytical methods for classification of the data points and identifying the data points of the utmost utility for PDF fits. As the length of $\vec{\delta}_i$ is equal to

[1] As a result, a small number of data points that contributes to both the data sets CMS7jets'14 and CMS7jets'13 is double-counted in the histograms, without affecting the conclusions.

[2] In this section, we consider separate variations along $\vec{a}_l$ in the positive and negative directions. Alternatively, it is possible to work with a vector of $N$ symmetric differences $\delta_{i,l} \equiv (r_i(\vec{a}_l^+) - r(\vec{a}_l^-))/(2\langle r_0\rangle_E)$ and arrive at similar conclusions. Symmetric differences will be employed to construct correlations and sensitivities in Sec. III.

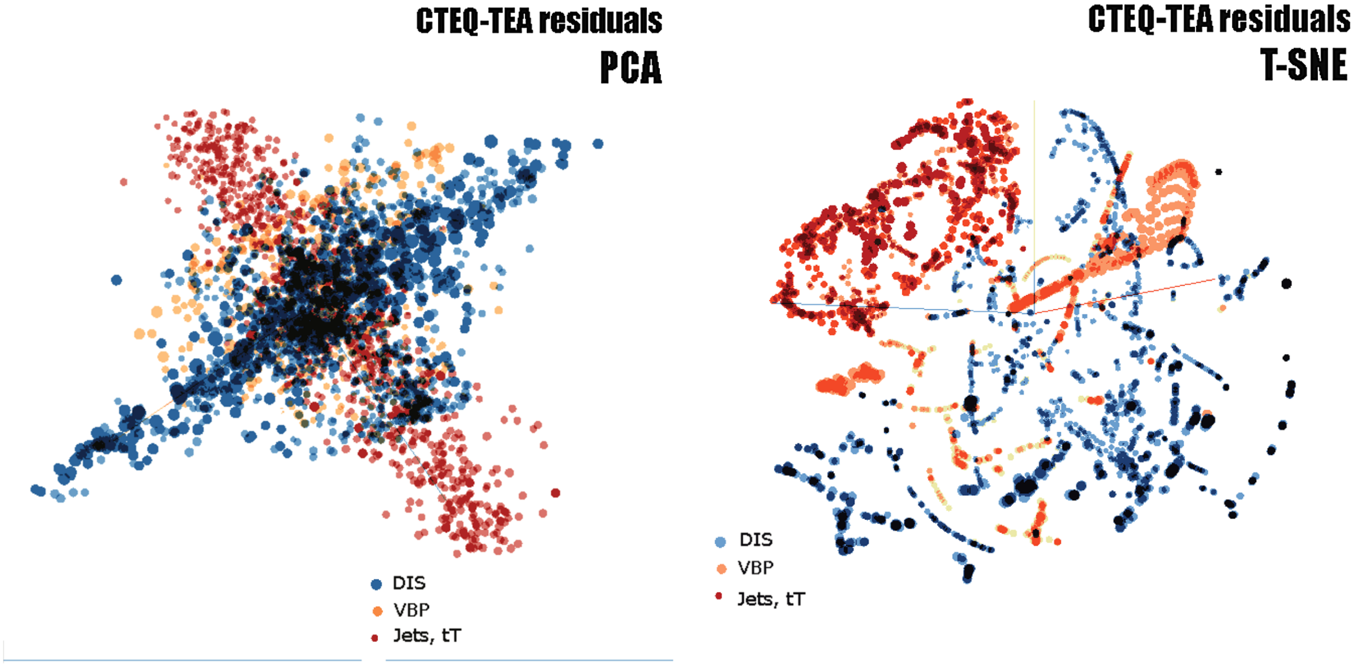

FIG. 3. Distributions of residual variations $\vec{\delta}_i$ from the CTEQ-TEA analysis obtained by dimensionality reduction methods. Left: a three-dimensional projection of a ten-dimensional manifold constructed by PCA. Right: a distribution from the three-dimensional t-SNE clustering method. Blue, orange, and red colors indicate data points from DIS, vector boson production, and jet/$t\bar{t}$ production processes.

the PDF-induced fractional error on $r_i$ as compared to the average residual at the best fit, it can be argued that important PDF constraints arise from new data points that either have a large $|\vec{\delta}_i|$ or are otherwise distinct from the existing data points. Conversely, new data points with a small $|\vec{\delta}_i|$, or the ones that are embedded in the preexisting clusters of points, are not likely to improve constraints on the PDFs.

## C. Manifold learning and dimensionality reduction

### *1. PCA and t-SNE visualizations*

We illustrate a possible analysis technique carried out with the help of the TensorFlow Embedding Projector software for the visualization of high-dimensional data [82]. A table of 4021 vectors $\vec{\delta}_i$ for the CTEQ-TEA data set (corresponding to our total number of raw data points) is generated by our package PDFSense and uploaded to the Embedding Projector website. As variations along many eigenvector directions result only in small changes to the PDFs, the 56-dimensional $\vec{\delta}_i$ vectors can in fact be projected onto an effective manifold spanned by fewer dimensions. Specifically, the Embedding Projector approximates the 56-dimensional manifold by a ten-dimensional manifold using principal component analysis (PCA). In practice, this ten-dimensional manifold is constructed out of the ten components of greatest variance in the effective space, such that the most variable combinations of $\delta_{i,l}$ are retained, while the remaining 46 components needed to fully reconstruct the original 56-dimensional $\vec{\delta}_i$ are discarded. However, because the ten PCA-selected components describe the bulk of the variance of $\delta_{i,l}$, the loss of these 46 components results in only a minimal relinquishment of information and in fact provides a more efficient basis to study $\delta_{i,l}$ variations.

We encourage the readers to download the table of the normalized residual variations $\vec{\delta}_i$ for CT14HERA2 NNLO from the PDFSense website [83] and explore it for themselves using the Embedding Projector [82] or another program for multidimensional data visualization such as a tour [84]. These tools help one to understand the detailed PDF dependence of individual data sets *without doing the global fit*. Performing such task has been challenging for nonexperts, if not for the PDF fitters themselves. With the proposed method, we can visually examine the PDF dependence of the residuals from the diverse data sets before quantitatively characterizing these distributions using the estimators developed in the next sections. In the future, a computer algorithm can be written to select the experimental data for PDF fits, based on the residual variations, and with minimal involvement from humans.

To offer an illustration, while grasping the full PDF dependence of the data points in the original 56-parameter space is daunting, in the ten-dimensional representation obtained via PCA, some directions result in efficient separation of the data points of different types according to their residual variations. The left panel of Fig. 3 shows one such three-dimensional projection of $\vec{\delta}_i$ that separates clusters of residual variations arising from data for DIS, vector boson production, and jet/$t\bar{t}$ production. In this

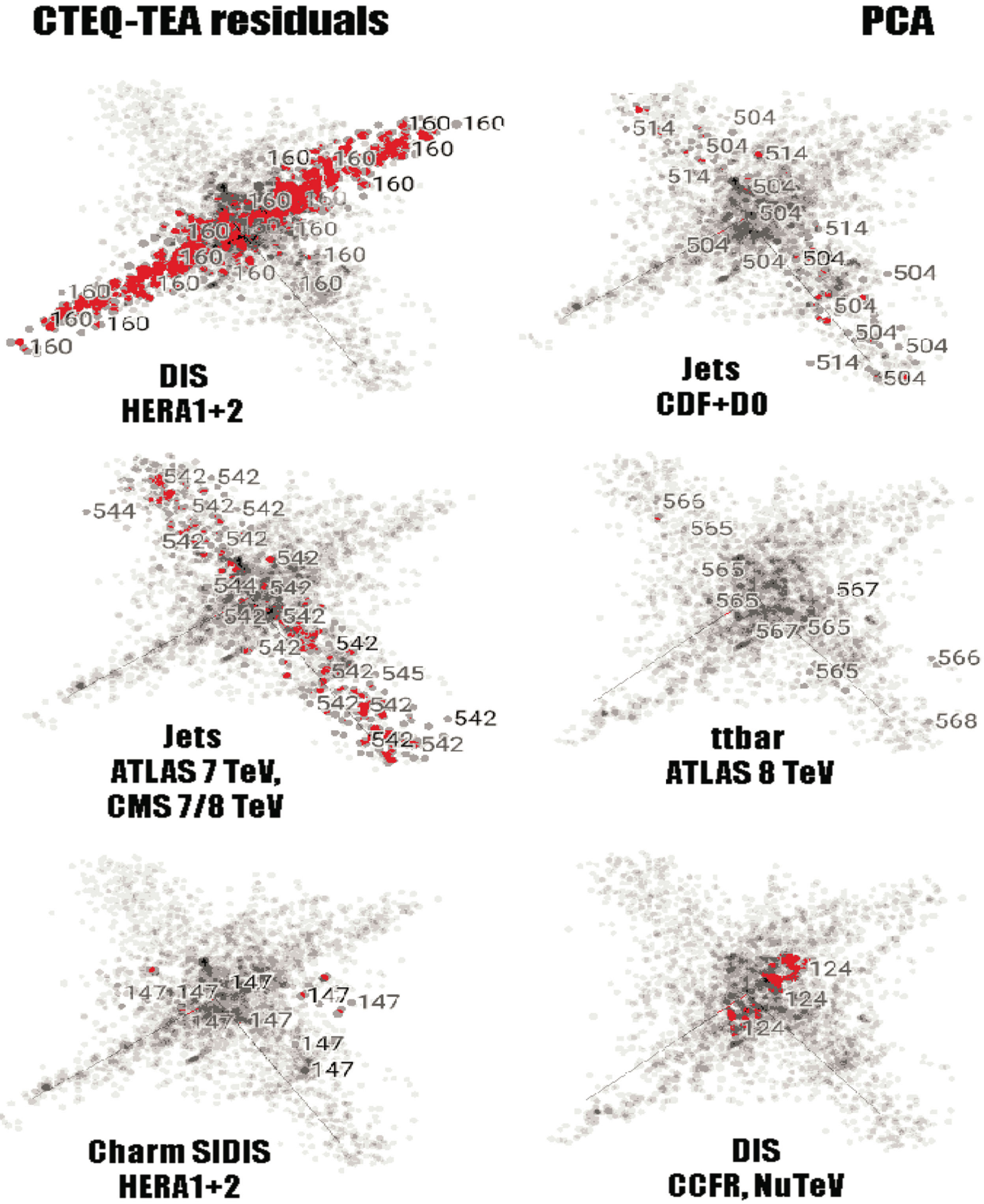

FIG. 4. The PCA distribution from Fig. 3, indicating distributions of points from classes of experiments. In the numbering scheme used here, points labeled 1XX correspond to fixed-target measurements and 5XX correspond to jet and $t\bar{t}$ production as given in Tables II–IV. The specific experiments are noted in the plots.

example, the jet/$t\bar{t}$ cluster, shown in red, is roughly orthogonal to the blue DIS cluster and intersects it. This separation is quite remarkable, as it is based only on numerical properties of the $\vec{\delta}_i$ vectors and not on the metadata about the types of experiments that is entered only after the PCA is completed; in other projections, the data types are not separated. The underlying reasons for this separation, namely, dependence on independent PDF combinations, will be quantified by the sensitivities in the next section.

As an alternative, the Embedding Projector can organize the $\vec{\delta}_i$ vectors into clusters according to their similarity using *t*-distributed stochastic neighbor embedding (t-SNE) [85]. A representative three-dimensional distribution of the vectors obtained by t-SNE is displayed in the right panel of Fig. 3. In the figure, we show that the t-SNE method is able to identify and separate the clusters of data according to the experimental process (DIS, vector production, or jet production). In fact, the readers can perform the t-SNE analysis on the Embedding Projector website themselves and verify that it actually sorts the $\vec{\delta}_i$ vectors into the clusters according to their values of $x$ and $\mu$, and even the experiment itself. This exercise demonstrates, yet again, that the statistical residuals provided in PDFSENSE reflect the key properties of the global fit. Information can be extracted from them and examined in a number of ways.

The breakdown of the vectors over experiments in the PCA representation is illustrated by Fig. 4. Here, we see that the bulk of the DIS cluster from the left Fig. 3 originates with the combined HERA1 + 2 DIS data (HERAI + II'15). The jet cluster in Fig. 3 will be dominated by ATLAS and CMS inclusive jet data sets (CMS7jets'14, ATLAS7jets'15, and CMS8jets'17), which add dramatically more points across a wider kinematical range on top of the CDF Run-2 and D0 Run-2 jet production data sets (CDF2jets'09 and D02jets'08).

In contrast, although the $t\bar{t}$ production experiments (ATL8ttb-pt'16, ATL8ttb-y_ave'16, ATL8ttb-mtt'16, and ATL8ttb-y_ttb'16) are generally characterized by large $\vec{\delta}_i$ vectors, they contribute only a few data points lying within the jet cluster of Fig. 4 and, by themselves, will not make much difference in a global fit. The same conclusion applies to data from high-$p_T$ $Z$ production, which has too few points to stand out in a fit with significant inclusive jet data samples. We return to this point in the discussion of reciprocated distances below.

It is also interesting to note that semi-inclusive charm production at HERA (HERAc'13) lies between, and partly overlaps with, the DIS and jet clusters. Finally, CCFR/NuTeV dimuon semi-inclusive DIS (SIDIS) (CCFR-F2'01, CCFR-F3'97, NuTeV-nu'06, and NuTeV-nub'06) extends in an orthogonal direction, not well separated from the other data sets in the selected three-dimensional projection.

### *2. Reciprocated distances*

As a complement to the visualization methods based on PCA and t-SNE just presented, it is also possible to evaluate another similarity measure based on the distances between the vectors of the residual variations. For example, rather than applying the PCA to an ensemble of $\vec{\delta}_i$ vectors to perform dimensionality reduction, we might instead compute over the vector space a pairwise *reciprocated distance* measure, which we define as

$$\mathcal{D}_i \equiv \left( \sum_{j \neq i}^{N_{\text{all}}} \frac{1}{|\vec{\delta}_j - \vec{\delta}_i|} \right)^{-1}, \qquad (15)$$

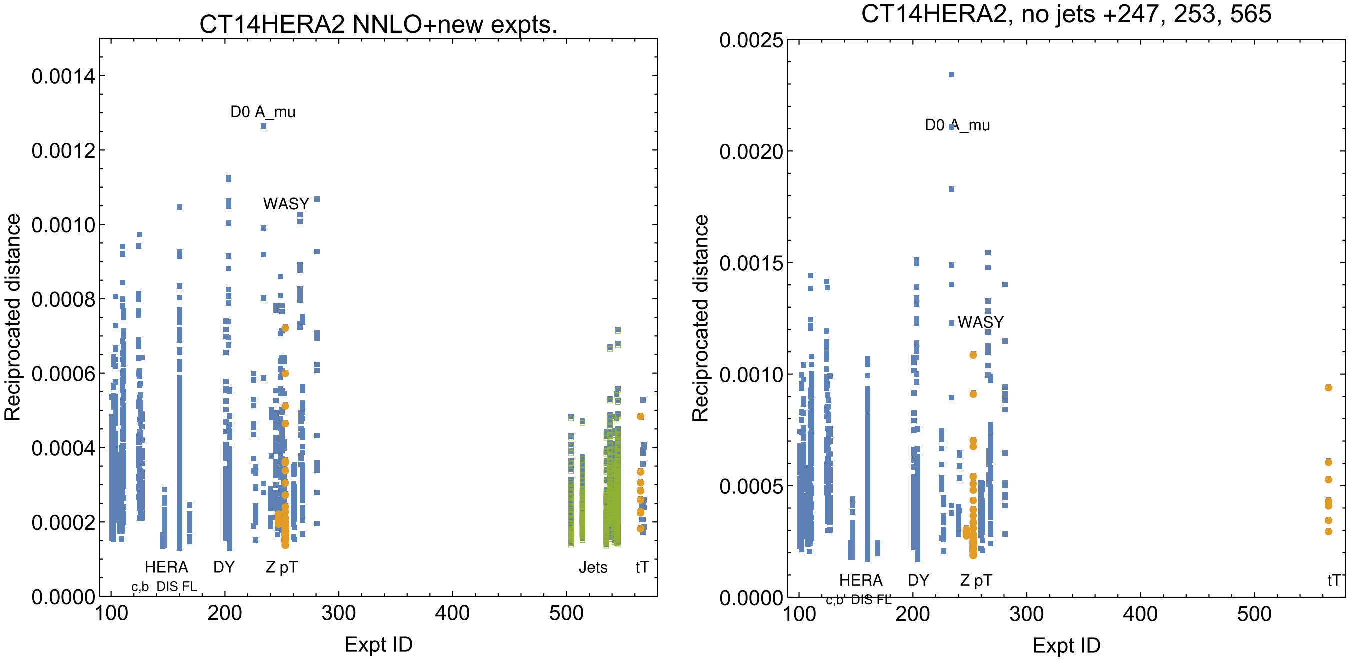

FIG. 5. A plot of the reciprocated distances $\mathcal{D}_i$ obtained from the PDFs fitted to the full CT14HERA2 data set (left) and to the CT14HERA2 data set without jet production experiments (right). The horizontal axis displays numerical experimental CT identifications of the constituent CTEQ-TEA data sets, for each of which is shown a column of values of the reciprocated distance. We highlight columns corresponding to experiment (Expt.) identifications ATL7ZpT'14 [247], ATL8ZpT'16 [253], and ATL8ttb-pt'16 [565] as discussed in the text.

and evaluate for the $i$ points in each experimental data set. We allow the sum over $j$ in Eq. (15) to run over all the data points in the CTEQ-TEA set regardless of experiment (denoted by $N_{\mathrm{all}}$). The distances can be computed either in the 56-dimensional space or in the reduced dimensionality space.[3] We plot the result of applying Eq. (15) to the 56-dimensional residual variations of the full CTEQ-TEA data set computed using two PDF ensembles: CT14HERA2 fitted to all data in the left panel and CT14HERA2 fitted only to the DIS and vector boson production data (excluding jet production data) in the right panel. Figure 5 represents the distribution of the reciprocated distances over individual experiments of the CTEQ-TEA data set. The CT experiment identification number is shown on the abscissa, and the $\mathcal{D}_i$ values for every point of the experiment are indicated by the scatter points.

The advantage of the definition in Eq. (15) is that it enables a quantitative measure of the degree to which separate experiments broadly differ in terms of their residual variations and therefore provides information analogous to that found in Figs. 3 and 4. For example, by inspection of Eq. (15), it can be seen that those experimental measurements which are widely separated from the rest of the CTEQ-TEA data set in space of $\vec{\delta}_i$ vectors will correspond to comparatively large values of $\mathcal{D}_i$, and experiments that systematically differ from the rest of the total data set are thus expected to have especially tall distributions in the panels of Fig. 5. On this basis, it can be seen that information yielded by W asymmetry measurements (D02Masy'08, CMS7Masy2'14, and D02Easy2'15) is particularly distinct, as are the combined HERA DIS data (HERAI + II'15) and fixed-target Drell-Yan measurements, such as E605 (E605'91) and E866 data (E866rat'01 and E866pp'03). Similarly, direct comparison of the $\mathcal{D}_i$ distributions in the panels of Fig. 5 allows one to compare constraints with and without the jet data. We note that the 7 and 8 TeV ATLAS high-$p_T$ $Z$ production (ATL7ZpT'14 and ATL8ZpT'16) and $t\bar{t}$ production (ATL8ttb-pt'16) provide a number of "remote" points and hence are potentially useful in the fits sensitive to the gluon. On the other hand, new jet production experiments (CMS7jets'14, ATLAS7jets'15, and CMS8jets'17) all include large numbers of points characterized by significant reciprocated distances.

## III. QUANTIFYING DISTRIBUTIONS OF RESIDUAL VARIATIONS

We have demonstrated that the multidimensional distribution of the $\vec{\delta}_i$ vectors reflects the PDF dependence of individual data points. In this section, we will focus on numerical metrics to assess the emerging geometrical picture associated with the $\vec{\delta}_i$ distribution and to visualize the regions of partonic momentum fractions $x$ and QCD factorization scales $\mu$ where the experiments impose strong constraints on a given PDF-dependent observable $X$.

Gradients of $r_i$ in a space of Hessian eigenvector PDF parameters $\vec{a}$ are naturally related to the PDF uncertainty. Recall that in the Hessian method the PDF uncertainty on $X(\vec{a})$ is found as

$$\Delta X(\vec{a}) = X(\vec{a}) - X(\vec{a}_0) = \vec{\nabla} X|_{\vec{a}_0} \cdot \Delta\vec{a}, \qquad (16)$$

where $\vec{a}_0$ is the best-fit combination of PDF parameters, and $\Delta\vec{a}$ is the maximal displacement along the gradient that is allowed within the tolerance hypersphere of radius $T$ centered on the best fit [56,58]. The standard master formula

$$\Delta X = |\vec{\nabla} X| = \frac{1}{2}\sqrt{\sum_{l=1}^{N} (X_l^+ - X_i^-)^2} \qquad (17)$$

is obtained by representing the components of $\vec{\nabla} X$ by a finite-difference formula,

$$\frac{\partial X}{\partial a_i} = \frac{1}{2}(X_i^+ - X_i^-), \qquad (18)$$

in terms of the values $X_l^{\pm}$ for extreme displacements of $\vec{a}$ within the tolerance hypersphere along the $l$th direction.

In this setup, a dot product between the gradients provides a convenient measure of the degree of similarity between PDF dependence of two quantities [59]. A dot product $\vec{\nabla} r_i \cdot \vec{\nabla} f$ between the gradients of a shifted residual $r_i$ and another QCD variable $f$, such as the PDF at some $\{x, \mu\}$ or a cross section, can be cast in a number of useful forms.

### A. Correlation cosine

The correlation for the $i$th $\{x, \mu\}$ point, which we define following Refs. [8,56,57,59] as

$$C_f \equiv \mathrm{Corr}[f, r_i] = \frac{\vec{\nabla} f \cdot \vec{\nabla} r_i}{\Delta f \Delta r_i}, \qquad (19)$$

can determine whether there *may* exist a predictive relationship between $f$ and goodness of fit to the $i$th point. The correlation function $\mathrm{Corr}[X, Y]$ for the quantities $X$, $Y$ in Eq. (19) represents the realization in the Hessian formalism of Pearson's correlation coefficient, which we express as

$$\mathrm{Corr}[X, Y] = \frac{1}{4\Delta X \Delta Y}\sum_{j=1}^{N}(X_j^+ - X_j^-)(Y_j^+ - Y_j^-), \qquad (20)$$

with the sum in these expressions being over the $j$ parameters of the full PDF model space. Geometrically,

[3] Alternative definitions for the reciprocated distance can be also used, with qualitatively similar conclusions. For example, we could sum over all experimental data, but excluding those points belonging to the same experiment as point $i$ and normalizing $\mathcal{D}_i$ by $(N_{\mathrm{pt}} - N_{\mathrm{all}})/N_{\mathrm{pt}}$ to compensate for different numbers of points in the experiment.

$\mathrm{Corr}[X,Y]$ represents the cosine of the angle that determines the eccentricity of an ellipse satisfying $\chi^2(\vec{a}) < \chi^2(\vec{a}_0) + T^2$ in the $\{X,Y\}$ plane. This latter point follows from the fact that the mapping of the tolerance hypersphere onto the $\{X,Y\}$ plane is an ellipse with an eccentricity that depends on the correlation of $X$ and $Y$, which is given in turn by Eq. (20) above.

$\mathrm{Corr}[f, r_i]$ does not indicate how constraining the residual is, but it may indicate a predictive relation between $r_i$ and $f$. On the basis of previous work [59], we say that the (anti)correlation between $X$ and $Y$ is significant roughly if $|\mathrm{Corr}[X,Y]| \gtrsim 0.7$, while smaller (anti)correlation values are less robust or predictive. Following this rule of thumb, correlations have been used successfully to identify PDF combinations that dominate PDF uncertainties of complicated observables, for instance to show that the gluon uncertainty dominates the total uncertainty on LHC $W$ and $Z$ production or that the uncertainty on the ratio $\sigma_W/\sigma_Z$ of $W^\pm$ and $Z^0$ boson cross sections at the LHC is dominated by the strangeness PDF, rather than $u$ and $d$ (anti)quark PDFs [59].

### B. Sensitivity in the Hessian method

The correlation $C_f$ alone does not fully encode the potential impact of separate or new measurements on improving PDF determinations in terms of the uncertainty reduction. Rather, we employ $\vec{\nabla} f \cdot \vec{\nabla} r_i$ again to define the sensitivity $S_f$ to $f$ of the $i$th point in experiment $E$,

$$S_f \equiv \frac{\vec{\nabla} f \cdot \vec{\nabla} r_i}{\Delta f \langle r_0 \rangle_E} = \frac{\Delta r_i}{\langle r_0 \rangle_E} C_f, \tag{21}$$

where $\Delta r_i$ and $\langle r_0 \rangle_E$ are computed according to Eqs. (3) and (12), respectively. In other words, $\Delta r_i$ again represents the variation of the residuals across the set of Hessian error PDFs, and we normalize it to the rms residual for the whole data set $E$ to reduce the impact of random fluctuations in the data values $D_{i,\mathrm{sh}}$. This definition has the benefit of encoding not only the correlated relationship of $f$ with $r_i$ but also the comparative size of the experimental uncertainty with respect to the PDF uncertainty. In consequence, for example, if new experimental data have reported uncertainties that are much tighter than the present PDF errors, these data would then register as high-sensitivity points by the definition in Eq. (21).

Geometrically, $S_f$ represents a projection onto the direction of the gradient $\vec{\nabla} f$ of the residual variation $\vec{\delta}_i$, defined in Sec. III using the symmetrized formula for $\delta_{i,l}$ noted in Footnote 2, namely,

$$\delta_{i,l} \equiv (r_i(\vec{a}_l^{\,+}) - r(\vec{a}_l^{\,-}))/(2\langle r_0 \rangle_E). \tag{22}$$

Figure 6 shows a pictorial illustration of this interpretation. This interpretation suggests that the total strength of constraints along the direction of $\vec{\nabla} f$ can be quantified by summing projections $S_f$ onto this direction of all individual vectors $\vec{\delta}_i$.

As with correlations, only a sufficiently large absolute magnitude of $|S_f|$ is indicative of a predictive constraint of the $i$th point on $f$. Recall that $r_i^2$ is the contribution of the $i$th point to $\chi^2$ and that only residuals with a large enough $\Delta r_i$ as compared to the rms residual $\langle r_0 \rangle_E$ are sensitive to PDF variations. The $S_f$ magnitude is of order $\Delta r_i/\langle r_0 \rangle_E$, which suggests an estimate of a minimal value of $S_f$ that would be deemed sensitive according to the respective $\chi^2$ contribution. For the numerical comparisons in this study, we assume that $|S_f|$ must be no less than 0.25 to indicate a predictive constraint, as the PDF uncertainty of the $i$th residual contributes no less than $r_i^2 = 0.0625$ to the variation in the global $\chi^2$. The reader can choose a different minimal value in the PDFSense figures depending on the desired accuracy. The cumulative sensitivities that we obtain in later sections are independent of this choice.

Yet another possible definition, which we list for completeness, is to further normalize the sensitivity as

$$S_f' \equiv \frac{\vec{\nabla} f \cdot \vec{\nabla} r_i}{f_0 \langle r_0 \rangle_E} = \frac{\Delta f}{f_0} S_f. \tag{23}$$

For instance, if $f$ is the PDF $f(x_i, \mu_i)$ or parton luminosity evaluated at the $\{x_i, \mu_i\}$ points extracted according to the data, the definition of $S_f'$ in Eq. (23) deemphasizes those points where the PDF uncertainty $\Delta f(x_i, \mu_i)$ is small compared to the best-fit PDF value $f_0(x_i, \mu_i)$—analogously to how $S_f$ deemphasizes (relative to the correlation $C_f$) those data points of which the normalized residual variations $\Delta r_i/\langle r_0 \rangle_E$ have already been more tightly constrained.

### C. Sensitivity in the Monte Carlo method

The above statistical measures are general enough and can be extended to other representations for the PDF uncertainties, such as the representation based on Monte Carlo replica PDFs [60,77,78] of the kind employed, e.g., in the NNPDF framework. A family of Monte Carlo PDFs consists of $N_{\mathrm{rep}}$ member PDF sets $q_a^{(k)}(x,\mu) \equiv \{q^{(k)}\}$, with $k = 1, \ldots, N_{\mathrm{rep}}$, and those are used to determine an expectation value $\langle X \rangle$ for a PDF-dependent quantity $X[\{q\}]$ such as a high-energy cross section:

$$\langle X \rangle = \frac{1}{N_{\mathrm{rep}}} \sum_{k=1}^{N_{\mathrm{rep}}} X[\{q^{(k)}\}]. \tag{24}$$

The resulting Monte Carlo uncertainty on $X$ can be extracted from the ensemble as

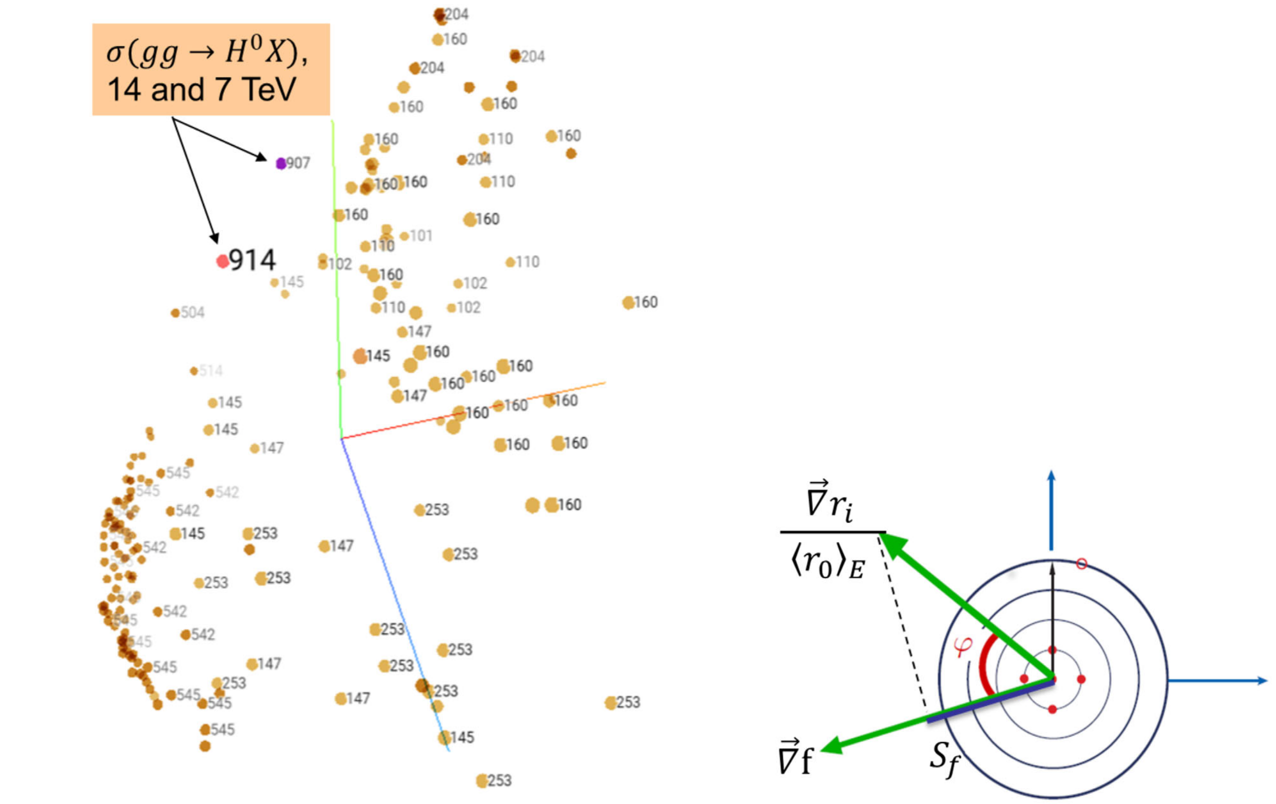

FIG. 6. Left: A PDF-dependent quantity $f$ defines a direction in space of $(2)N$ PDF parameters. The direction is specified by the gradient $\vec{\nabla} f$ in the symmetric convention. Here, the Embedding Projector [82] visualizes the vectors $\vec{\delta}_{907}$ and $\vec{\delta}_{914}$ for NNLO cross sections for Higgs boson production at 7 and 14 TeV and vectors $\vec{\delta}_i$ for CT14HERA2 NNLO data points from Ref. [83] (brown circles), showing only $\vec{\delta}_i$ with the smallest angular distances to $\vec{\delta}_{914}$. These points impose the strongest constraints on the PDF dependence of the Higgs cross sections in the CT14HERA2 analysis, if they have large enough $|\vec{\delta}_i|$. Again, in the numbering scheme used here, points labeled 1XX correspond to fixed-target measurements, 2XX correspond to Drell-Yan processes and boson production, and 5XX correspond to jet and $t\bar{t}$ production as given in Tables II–IV. Right: the sensitivity $S_f$ of the $i$th data residual can be interpreted as the projection of $\vec{\delta}_i \equiv \vec{\nabla} r_i / \langle r_0 \rangle_E$ onto the direction of $\vec{\nabla} f$.

$$\Delta_{\rm MC} X = \left( \frac{1}{N_{\rm rep} - 1} \sum_{k=1}^{N_{\rm rep}} (X[\{q^{(k)}\}] - \langle X \rangle)^2 \right)^{1/2}. \quad (25)$$

In consequence of these definitions, the central value of a particular PDF itself in the NNPDF framework is specified as

$$q_{(0)} \equiv \langle q \rangle = \frac{1}{N_{\rm rep}} \sum_{k=1}^{N_{\rm rep}} q^{(k)}. \quad (26)$$

Akin to the Pearson correlation defined in Eq. (19) of Sec. III A, statistical correlations between two PDF-dependent quantities $X[\{q\}]$ and $Y[\{q\}]$ can be constructed from the PDF replica language above in terms of ensemble averages [60]:

$$\mathrm{Corr}_{\rm MC}[X, Y] = \frac{\langle XY \rangle - \langle X \rangle \langle Y \rangle}{\Delta_{\rm MC} X \Delta_{\rm MC} Y}. \quad (27)$$

Then, using our definitions in Eqs. (19) and (21), we immediately construct the realizations of the correlation and sensitivity for a PDF-dependent quantity $f$ in the Monte Carlo method:

$$C_{f,\rm MC} = \mathrm{Corr}_{\rm MC}[f, r_i], \quad (28)$$

$$S_{f,\rm MC} = \frac{\Delta_{\rm MC} r_i}{\langle r_0 \rangle_E} \mathrm{Corr}_{\rm MC}[f, r_i]. \quad (29)$$

## IV. CASE STUDY: CTEQ-TEA GLOBAL DATA

### A. Maps of correlations and sensitivities

We will now discuss a number of practical examples of using $C_f$ or $S_f$ to quickly evaluate the impact of various hadronic data sets upon the knowledge of the PDFs in a fashion that does not require a full QCD analysis of the type described in Sec. II. For this demonstration, we will continue to study the data set shown in Fig. 1 of the CT14HERA2 analysis [10] augmented by the candidate LHC data.

We have already noted the extent of this data set in the $\{x, \mu\}$ plane in Fig. 1, where it is decomposed into constituent experiments labeled according to the conventions in Tables II–IV. It is instructive to create similar maps in the $\{x, \mu\}$ plane showing the $C_f$ or $S_f$ values for each data point. Such maps are readily produced by the PDFSense program for a variety of PDF flavors and for user-defined observables, such as the Higgs cross section. For demonstration, we have collected a large number of these maps at the companion website [83]. We invite the reader to review these additional figures while reading the paper to validate the conclusions that will be summarized below.

Thus, we obtain scatter plots of $C_f(x_i, \mu_i)$ or $S_f(x_i, \mu_i)$ for a given QCD observable $f = \sigma$, such as the LHC Higgs production cross section shown in Fig. 2, or with a PDF $f$ evaluated at the same $\{x_i, \mu_i\}$ determined by the data points, with examples shown for $g(x_i, \mu_i)$ in Figs. 7 and 8. The typical $\{x_i, \mu_i\}$ values characterizing the data points are found according to Born-level approximations appropriate for each scattering process included in the CTEQ-TEA data set, with the formulas to compute these kinematic matchings summarized in Appendix A. Here and in general, we find it preferable to consider the absolute values $|C_f|$ and $|S_f|$ on the grounds that the signs of $C_f$ and $S_f$ flip when the data points randomly overshoot or undershoot their theory predictions.

Together with the map in the $\{x, \mu\}$ plane, PDFSense also returns a histogram of the values for each quantity it plots. An example is shown for $|C_g|(x_i, \mu_i)$ in the first panel of Fig. 7. One would judge that stronger constraints are in general provided to those PDFs for which the $|C_f|$ histogram has many entries comparatively close to $|C_f| \sim 1$. In the first panel of Fig. 7, we can see that, while the distribution peaks at low correlations, $|C_g| \sim 0$, the distribution has an extended tail in the region $0.7 \lesssim |C_g| \lesssim 1$. This feature shows that, of the 4021 experimental data points within the augmented CT14HERA2 set in Fig. 1, nearly two hundred—specifically, 192—have especially strong ($|C_f| \geq 0.7$) correlations (or anticorrelations) with the gluon PDF. This region of such strong correlations within the histogram is indicated by the horizontal blue bar that runs along the abscissa.

To identify these points, we plot complementary information in the second panel of the same figure—specifically, a map in $\{x, \mu\}$ space of each of the data points shown in Fig. 1. As before, they are colorized according to the magnitude of $|C_g|$ following the color palette in the "rainbow strip" on the right. "Cooler" colors (green/yellow) correspond to weaker correlation strengths, while "hotter" colors (orange/red) represent comparatively stronger correlations, as indicated. To reveal the data points with the highest correlations, we reproduce the same figure in the third panel, but show in color only the data points satisfying $|C_f| > 0.7$. Thus, we obtain two maps in the $\{x, \mu\}$ plane that look similar to the $|C_f|$ map in the left panel of Fig. 2, apart from the differences that (a) Fig. 7 shows the correlation $|C_g|$ for $g(x_i, \mu_i)$ at the same typical values $\{x_i, \mu_i\}$ as in the data, rather than $|C_{\sigma_{H^0}}|$ for the Higgs production cross section in Fig. 2, and (b) Fig. 2 highlights 310 points with the highest $|C_{\sigma_{H^0}}|$.

The correlations for the LHC Higgs production cross section trace those for $g(x_i, \mu_i)$, but not entirely, as we will see in a moment. Large magnitudes of $|C_g|$ in Fig. 7 are found for inclusive jet production measurements, especially those recently obtained by CMS at 8 TeV [49] (experiment CMS8jets'17, inverted triangles) with $|C_g|(x_i, \mu_i)$ as high as 0.85, including at the highest values of $x$ and $\mu$. Beyond these, a sizable cluster of HERA (HERAI + II'15) data points at the lowest values of $x$ is also seen to have large correlations with the gluon PDF, consistent with the common wisdom that HERA DIS constrains the gluon PDF at small $x$ via DGLAP scaling violations. Under the jet production cluster, high-$p_T$ $Z$ production (ATL7ZpT'14 and ATL8ZpT'16) and $t\bar{t}$ production (ATL8ttb-pt'16, ATL8ttb-y_ave'16, ATL8ttb-mtt'16, and ATL8ttb-y_ttb'16) at the LHC show a high $|C_g|(x_i, \mu_i)$ correlation. At the same time, many other measurements, including fixed-target data at large $x$ and $W$ asymmetry data near $\mu \sim 100$ GeV, have feeble correlations with $g(x_i, \mu_i)$ and would therefore be less emphasized by an analysis based solely upon the PDF-residual correlations.

We can also consider the analogous plots for the sensitivity $|S_g|(x_i, \mu_i)$ as defined in Eq. (21), which we plot in Fig. 8. In the first panel, we again consider the histogram, here for the magnitudes of the gluon sensitivity $|S_g|(x_i, \mu_i)$, in which the correlations $|C_g|$ are now weighted by the relative size of the PDF uncertainty $\Delta r_i$ in the residual. As discussed in Sec. III B, this additional weighting emphasizes those data points for which the PDF-driven fluctuations in the residuals are comparatively large relative to experimental uncertainties. This leads to a redistribution of the data points shown in the $|C_g|$ histogram of Fig. 7, with the result being a considerably longer-tailed histogram for $|S_g|$ such that, in this instance, there are 546 raw data points with larger sensitivities, $|S_f| \geq 0.25$, indicated by the horizontal blue bar. Unlike the correlation, $|S_g|$ can be arbitrarily large, depending on the $\Delta r_i$ value. It is suppressed at the data points with large uncertainties or smeared over the regions of data points with correlated systematic uncertainties.

In the second and third panels, we show the respective $\{x, \mu\}$ maps for $|S_g|$, with color highlighting given either for all points or only those with high sensitivities $|S_f| > 0.25$, respectively. $|S_g|$ places additional emphasis on the combined HERA data set (HERAI + II'15) constraining $g(x_i, \mu_i)$ at lowest $x$. In contrast to the $|C_g|$ plot, we observe increased sensitivity in the precise fixed-target DIS data

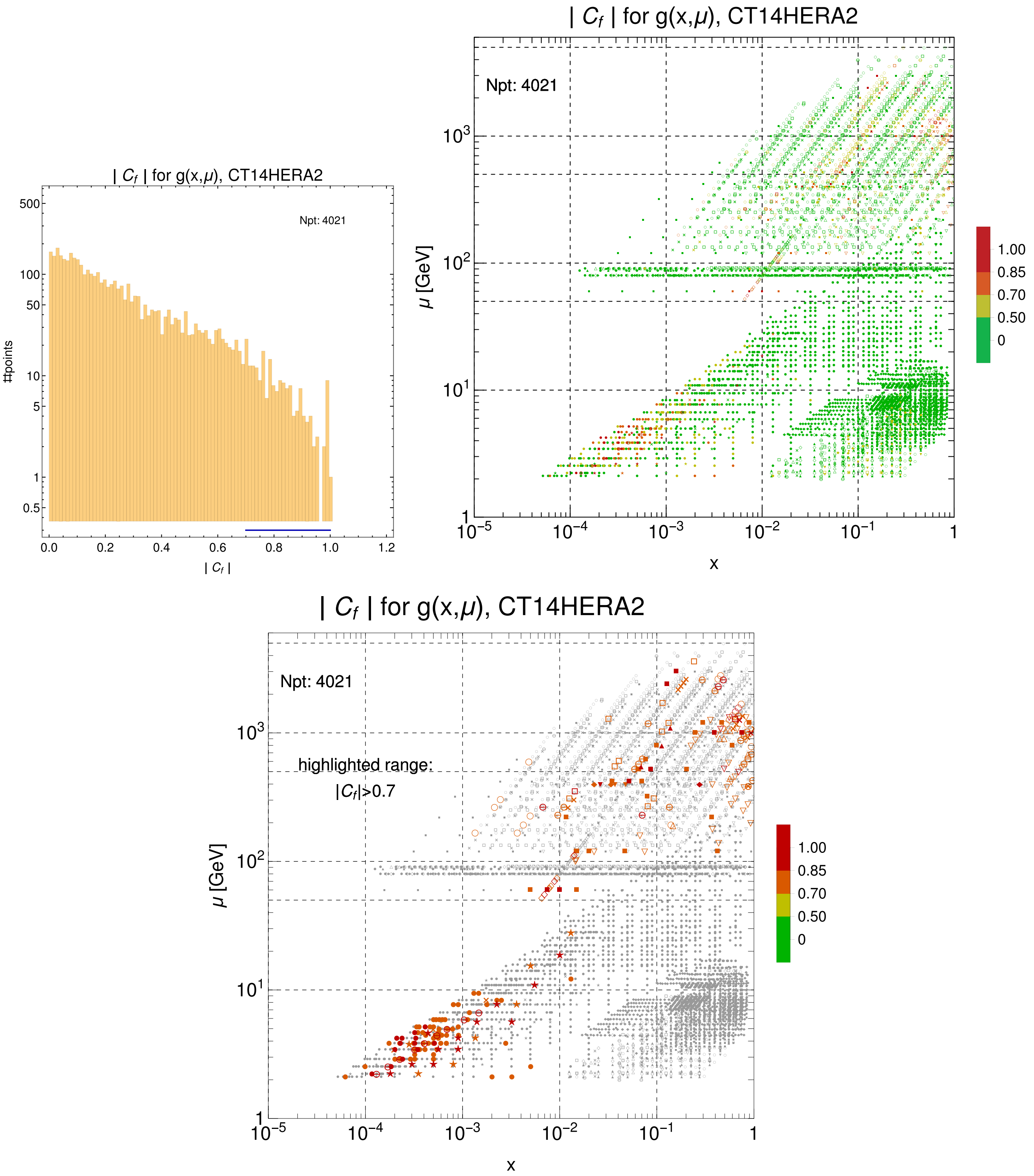

FIG. 7. Representations of the correlation $|C_g|(x_i, \mu_i)$ of the gluon PDF $g(x, \mu)$ with the pointwise residual $r_i$ of the augmented CT14HERA2 analysis. In the first panel, we plot a histogram showing the distribution of correlations for 4021 physical measurements. In the second panel, we show the 5227-point $\{x_i, \mu_i\}$ map corresponding to these data within the full data set, generated as in Appendix A. To adjust for the fact that some measurements of rapidity-dependent quantities match to two distinct points in $\{x_i, \mu_i\}$ space using the rules of Appendix A, we assign weights of 0.5 to these complementary $\{x_i, \mu_i\}$ points in computing the $N_{\mathrm{pt}} = 4021$-count histogram at left. The third figure is the same as the second one, but only the data points satisfying $|C_f| > 0.7$ are highlighted.

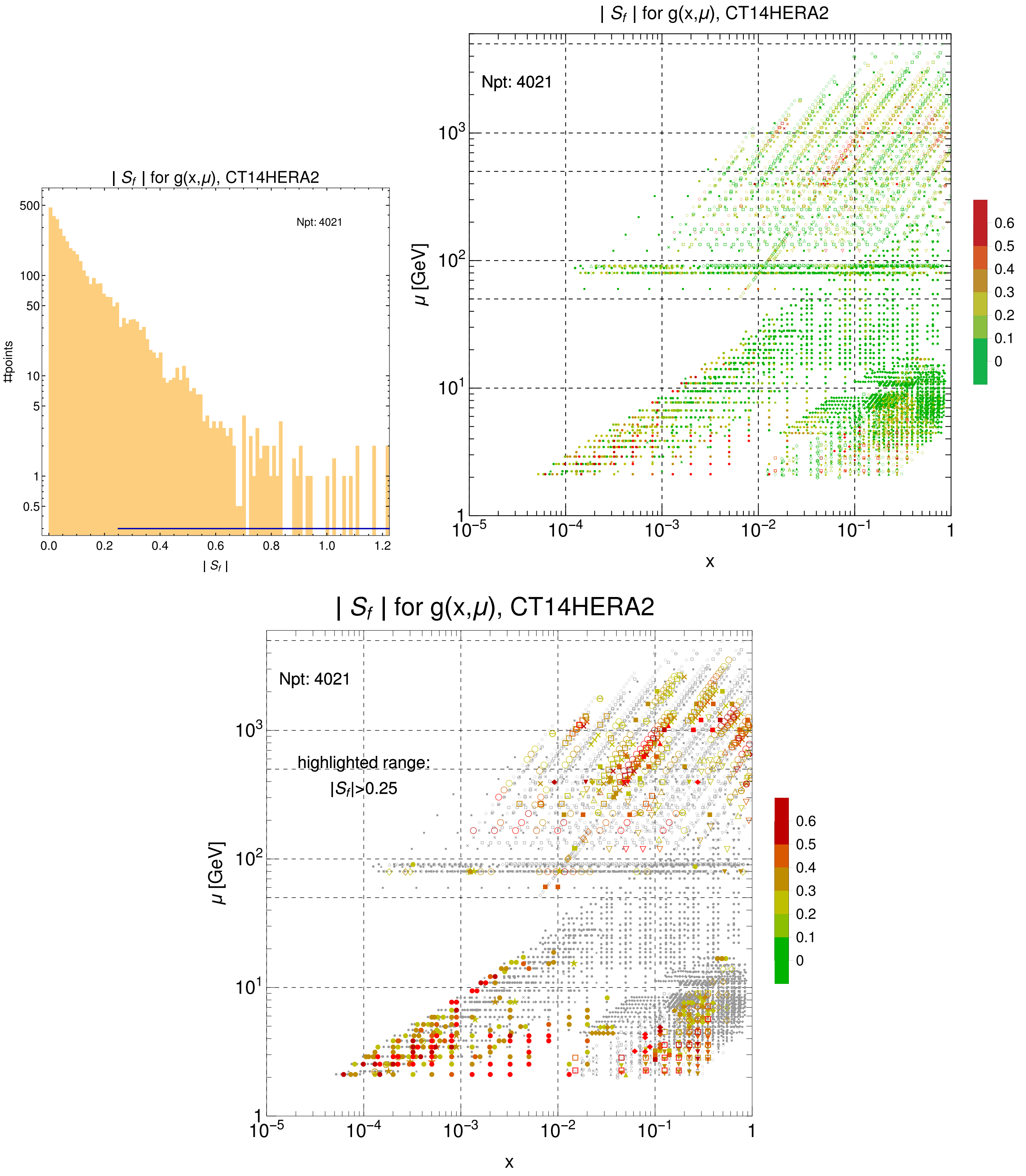

FIG. 8. Like Fig. 7, but for the gluon sensitivity $|S_g|(x_i, \mu_i)$ as defined in Eq. (21). In the third figure, only the data points satisfying $|S_f| > 0.25$ are highlighted.

from BCDMS (BCDMSp'89 and BCDMSd'90) and CCFR (CCFR-F2'01 and CCFR-F3'97), which are sensitive to the gluon via scaling violations despite only moderate correlation values. Similarly, we observe heightened sensitivities at highest $x$ for the LHC (CMS7jets'14, ATLAS7jets'15, and CMS8jets'17) and Tevatron (D02jets'08) jet production data, which have both large correlations with $g(x_i, \mu_i)$ and small experimental uncertainties. Sensitivity $|S_g|$ of

LHC jet experiments (CMS7jets'14, ATLAS7jets'15, and CMS8jets'17) varies in a large range and can significantly improve, depending on the implementation of experimental systematic uncertainties in the analysis, cf. the discussion of the jet data in the next section.

We also observe enhanced sensitivity for *individual points* in a large number of experiments, including CDHSW DIS (CDHSW-F2'91), HERA $F_L$ (HERA-FL'11), the Drell-Yan process (E605'91 and E866pp'03), CDF 8 TeV $W$ charge asymmetry (CMS7Masy2'14), HERA charm SIDIS (HERAc'13), ATLAS high-$p_T$ $Z$ production (ATL7ZpT'14 and ATL8ZpT'16), and especially strongly sensitive points in $t\bar{t}$ production (ATL8ttb-pt'16, ATL8ttb-y_ave'16, ATL8ttb-mtt'16, and ATL8ttb-y_ttb'16). However, since the latter category includes fewer points per each experiment, it constrains the gluon less than the high-statistics DIS and jet production data.

These findings comport with the idea that the gluon PDF remains dominated by substantial uncertainties at both $x \sim 0$ and in the elastic limit $x \to 1$, a fact which has driven an intense focus upon the production of hadronic jets, $t\bar{t}$ pairs, and high-$p_T$ $Z$ bosons, which themselves are measured at large center-of-mass energies $\sqrt{s}$ and are expected to be sensitive to the gluon PDF across a wide interval of $x$, including $x \sim 0.01$ typical for Higgs boson production via gluon fusion at the LHC. Turning back to the distributions of $|C_{\sigma_H}|(x_i, \mu_i)$ and $|S_{\sigma_H}|(x_i, \mu_i)$ for the Higgs cross section $\sigma_H$ at $\sqrt{s} = 14$ TeV in Fig. 2, we notice that they largely reflect the distributions of $|C_g|(x_i, \mu_i)$ and $|S_g|(x_i, \mu_i)$ around $x \sim M_H/\sqrt{s} = 125/14000 = 0.009$ and $\mu = M_H = 125$ GeV. We also see some differences: although the average $x$ and $\mu$ are fixed in $\sigma_H$, it is nonetheless sensitive to some constraints at much lower $x$ values as a result of the momentum sum rule.

The reader is welcome to examine the plots of sensitivities and correlations available on the PDFSense website for a large collection of PDF flavors and PDF ratios, such as $d/u$, $\bar{d}/\bar{u}$, and $(s+\bar{s})/(\bar{u}+\bar{d})$. Sensitivities for other PDF combinations and hadronic cross sections can be computed and plotted in a matter of minutes using the PDFSense program. We will now turn to another aspect of this analysis: summarizing the abundant information contained in the sensitivity plots. For this purpose, we will introduce numerical indicators and propose a practical procedure to rank the experimental data sets according to their sensitivities to the PDFs or PDF-dependent observables of interest.

### B. Experiment rankings according to cumulative sensitivities

Being one-dimensional projections of normalized residual variations $\vec{\delta}_i$ on a given direction in the PDF parameter space, sensitivities can be linearly added to construct a number of useful estimators. By summing absolute sensitivities $|S_f^{(i)}|$ over the data points $i$ of a given data set $E$, we find the maximal cumulative sensitivity of $E$ to the PDF dependence of a QCD observable $f$.

Alternatively, from the examination of multiple $\{x, \mu\}$ maps for $|S_f|$ of various PDF flavors collected on the website [83], we find that the most precise experiments constrain several flavors at the same time, most notably, the combined HERA data. For the purpose of identifying such experiments, we can compute an overall sensitivity statistic for each experiment $E$ to the parton distributions $f_a(x_i, \mu_i)$ evaluated at the same kinematic parameters $\{x_i, \mu_i\}$ as the data. Furthermore, to obtain one overall ranking, we can add up sensitivity measures as an unweighted sum over the "basis PDF" flavors, such as the six light flavors ($\bar{d}$, $\bar{u}$, $g$, $u$, $d$, $s$). To obtain these measures, we say that an experiment $E$ consisting of $N_{\rm pt}$ physical measurements can be characterized by its mean sensitivity per raw data point[4] to a PDF of given flavor $f_a(x, \mu)$: $\langle |S_f^E| \rangle \equiv (N_{\rm pt})^{-1} \sum_{i=1}^{N_{\rm pt}} |S_f|(x_i, \mu_i)$, from which we derive several additional statistical measures of experimental sensitivity. For each experiment and flavor, we then determine a cumulative sensitivity measure, numerically adjusted to the size of each experimental data set $E$, according to $|S_f^E| \equiv N_{\rm pt} \langle |S_f^E| \rangle$. In addition, we also track cumulative, flavor-summed sensitivity measures $\sum_f |S_f^E|$ and $\langle \sum_f |S_f^E| \rangle$, with $f$ running over $\bar{d}$, $\bar{u}$, $g$, $u$, $d$, $s$.

We list the corresponding values of these four types of sensitivities for each experiment of the CTEQ-TEA data set in extensive summary tables provided as Supplementary Material [86]. This is also detailed for categories of experiments from the CTEQ-TEA data set.

With the above estimators, we *quantify* and *compare* the cumulative sensitivities of each experiment to the basic 6 parton flavors. In fact, based on the various trials that we performed, we find that the cumulative sensitivity to the six basic flavors is a good measure of the overall sensitivity to a large range of PDF combinations. Recall that the $N_f = 5$ CT14HERA2 PDFs (with up to 11 independent parton species) are obtained by DGLAP evolution of the six basic parton flavors from the initial scale of order 1 GeV. There exist alternative approaches for measuring the importance of a given experiment in a global fit, for example, by counting the numbers of eigenvector parameters [87] or eigenvector directions [2] that the experiment constrains. Those other methods, however, require access to the full machinery of the global fit, while the sensitivities allow the reader to rank the experiments according to much the same information, for a variety of PDF-dependent observables, with the help of PDFSense, and at a fraction of computational cost.

[4] For those circumstances in which an individual measurement, e.g., obtained via the Drell-Yan process, maps to two sensitivity values in $\{x, \mu\}$ space, we compute the average of these and assign the result to that specific measurement.

In fact, in a companion study, we use the above sensitivity estimators to select the new LHC experiments for the inclusion in the next generation of the CTEQ-TEA PDF analysis. Full tables given in Supplementary Materials [86] provide detailed information about the PDF sensitivities of every experiment of the CTEQ-TEA data set. For a nonexpert reader, along with the full tables, we provide their simplified versions in Tables V–VI, where we rank the experimental sensitivities according to a reward system described in the caption of Table V. In each table, experiments are listed in descending order according to the cumulative sensitivity measure $\sum_f |S_f^E|$ to the six light-parton flavors. For each PDF flavor, the experiments with especially high overall flavor-specific sensitivities receive an "**A**" rating (shown in bold), per the convention in the caption of Table V. Successively weaker overall sensitivities receive marks of "B" and "C," while those falling below a lower limit $|S_f^E| = 20$ are left unscored.

We similarly evaluate each experimental data set based on its point-averaged sensitivity, in this case scoring according to a complementary scheme in which the highest score is "**1**." The short-hand names of the candidate experiments that were *not* included in the CT14HERA2 NNLO fit, that is, the new LHC experiments, are also shown in bold to facilitate their recognition in the tables.

Not only do the sensitivity rankings confirm findings known by applying other methods, but they also provide new insights. According to this ranking system in Tables V–VI, we find that the expanded HERA data set (HERAI + II'15) tallies the highest overall sensitivity to the PDFs, with enhanced sensitivity to the distributions of the $u$- and $\bar{u}$-quarks, as well as that of the gluon. On similar footing, but with slightly weaker overall sensitivities, are a number of other fixed-target measurements, including structure function measurements from BCDMS for $F_2^{p,d}$ (BCDMSp'89 and BCDMSd'90) and CCFR extractions of $xF_3^p$ (CCFR-F3'97)—as well as several other DIS data sets. Among the LHC experiments, the inclusive jet measurements have the highest cumulative sensitivities, with CMS jets at 8 TeV (CMS8jets'17), 7 TeV (CMS7jets'13 and CMS7jets'14), and ATLAS 7 TeV (ATLAS7jets'15) occupying positions 10, 12/13, and 16 in the total sensitivity rankings. They demonstrate the strongest sensitivities among the candidate LHC experiments and at the same time are not precise enough and fall behind the top fixed-target DIS and Drell-Yan experiments: BCDMS, CCFR, E605, E866, and NMC. The two versions CMS7jets'13 and CMS7jets'14 of the CMS 7 TeV jet data that largely overlap have very close sensitivities and rankings in Tables V–VI. The set CMS7jets'13 that extends to higher $p_{Tj}$ has a slightly better overall sensitivity, surpassing the larger data set CMS7jets'14 that includes the extra data points at $p_{Tj} < 100$ GeV or $|y_j| > 2.5$, yet cannot beat CMS7jets'13 except for in the overall sensitivity to the Higgs cross section at 7 TeV.

Going beyond the rankings based upon overall sensitivities, which are more closely tied to the impact of an entire experimental data set in aggregate, it is useful to consider the point-averaged sensitivity as well, which quantifies how sensitive each individual point is. (Some experiments with very high point-averaged sensitivity have a small cumulative sensitivity because of a small number of points.) Based on their high point-averaged sensitivity, CMS $\mu$ asymmetry measurements at 8 and 7 TeV (CMS8Wasy'16 and CMS7Masy2'14) especially stand out, despite their small number of individual points, $N_{\mathrm{pt}} = 11$); this is especially true again for the gluon and $\bar{d}$- and $u$-quark PDFs, for which this set of measurements is particularly highly rated in Table V. Another "small-size" data set with the exceptional point-average sensitivity is the $\sigma_{pd}/(2\sigma_{pp})$ ratio from the E866 lepton pair production experiment (E866rat'01). The average sensitivity of this data set to $\bar{u}$ and $\bar{d}$ PDFs is 0.8, making it extremely valuable for constraining the ratio $\bar{d}/\bar{u}$ at $x \sim 0.1$, in spite of its small size (15 data points).

Aside from the quark- and gluon-specific rankings of specific measurements, we can also assess experiments based upon the constraints they impose on various interesting flavor combinations and observables as presented in Table VI. As was the case with Table V, a considerable amount of information resides in Table VI, of which we only highlight several notable features here. Among these features are the sharp sensitivities to the Higgs cross section (e.g., $|S|_{H7}$, $\langle |S_{H7}| \rangle$, etc.) found for Run I + II HERA data, as well as the tier-C overall sensitivities of the BCDMS $F_2^{p,d}$ and CMS jet production measurements, corresponding to experiments BCDMSd'90, BCDMSp'89, CMS8jets'17, and CMS7jets'14. While their overall sensitivity is small, the corresponding ATLAS $t\bar{t}$ data also possess significant point-averaged sensitivity. On the other hand, measurements of $p_T$-dependent $Z$ production (ATL7ZpT'14 and ATL8ZpT'16) appear to have somewhat less pronounced sensitivity to the gluon and other PDF flavor combinations. The total and mean sensitivities of high-$p_T$ $Z$ boson production experiment ATL8ZpT'16 at 8 TeV are on par with HERA charm SIDIS data (HERAc'13) and provide comparable constraints to charm DIS production, albeit in a different $\{x, \mu\}$ region.

For the light-quark PDF combinations like $u_v$, $d_v$, $d/u$, and $\bar{d}/\bar{u}$, the various DIS data sets—led by Run II of HERA and CCFR measurements of the proton structure function—demonstrate the greatest sensitivity. At the same time, however, Run-2 Tevatron data from D0 on the $\mu$ asymmetry (D02Easy2'15) and Run-1 CDF measurements for the corresponding $A_e(\eta^e)$ asymmetry (CDF1Wasy'96) also exhibit substantial pointwise sensitivity as well. We collect a number of other observations in the conclusion below, Sec. V.

### C. Estimating the impact of LHC data sets on CTEQ-TEA fits

The presented rankings suggest that including the candidate LHC data sets will produce mild improvements

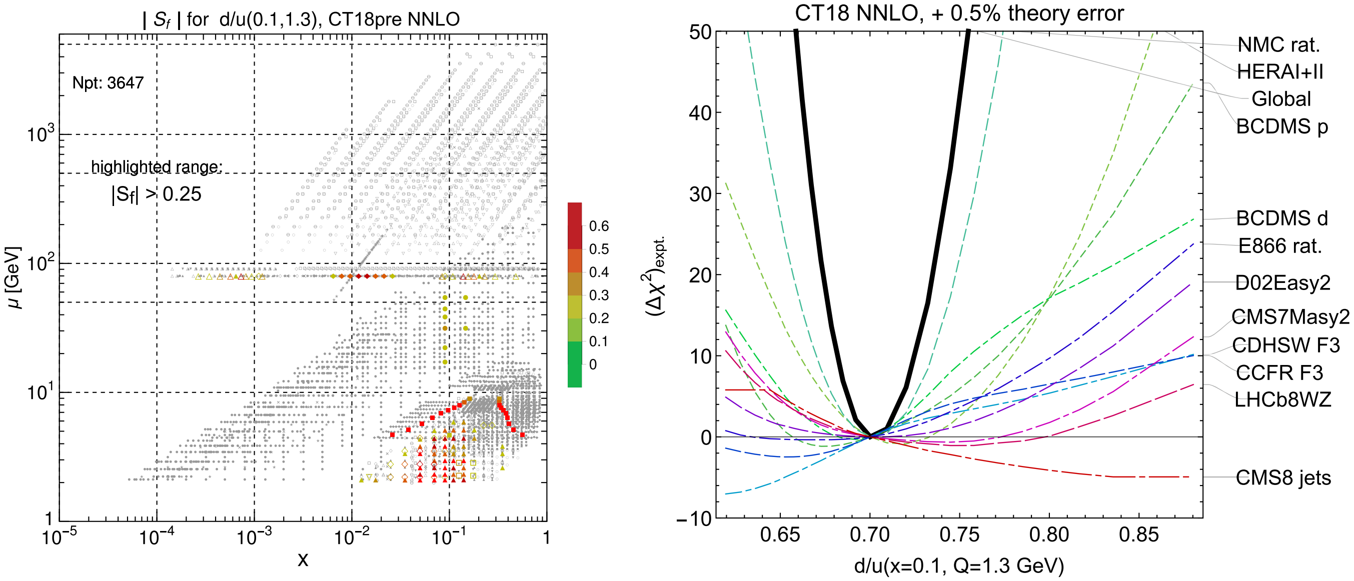

FIG. 9. Left: the PDFSense map for the sensitivity of the fitted data set of the CT18pre NNLO analysis to the $d/u$ PDF ratio, $d/u(x = 0.1, \mu = 1.3\ \mathrm{GeV})$. Right: dependence of $\chi^2$ for the individual and all experiments of the CT18pre data set on the value of $d/u(x = 0.1, \mu = 1.3\ \mathrm{GeV})$ obtained with the LM scan technique. The curves show the deviations $\Delta\chi^2_{\rm expt} \equiv \chi^2_{\rm expt}(\vec{a}) - \chi^2_{\rm expt}(\vec{a}_0)$ from the best-fit values in $\chi^2$ for the indicated experiments, as well as for the totality of all experiments.

in the uncertainties of the CT14 HERA2 PDFs. This projection may appear underwhelming, but keep in mind that the CT14HERA2 NNLO analysis already includes significant experimental constraints, for example, imposed on the gluon PDF at $x > 0.01$ by the Tevatron and LHC jet experiments, CDF2jets'09, D02jets'08, ATL7jets'12, and CMS7jets'13. If all jet experiments are eliminated from the PDF fit, as illustrated in the Supplementary Material [86] tables, the candidate LHC experiments will be promoted to higher rankings, with the CMS 8 and 7 TeV jet experiments (CMS8jets'17 and CMS7jets'13/CMS7jets'14) elevated to positions 4 and 7/8 in the overall sensitivity rankings, respectively.

Our investigations also find that the sensitivities of CMS jet experiments may improve considerably if the current correlated systematic effects are moderately reduced compared to the published values. For instance, by requiring a full correlation of the JEC2 correlation error over all rapidity bins in the CMS 7 TeV jet data set CMS7jets'14, instead of its partial decorrelation implemented according to the CMS recommendation [88], we obtain a very strong sensitivity of the data set CMS7jets'14 to $g$ over the full $\{x, \mu\}$ region but also strong sensitivities to $\bar{u}$, $\bar{d}$, and even $\bar{s}$ PDFs.[5] The overall sensitivity of the data set CMS7jets'14 in this case is elevated to the 4th position from the 13th position in the CT14HERA2 NNLO analysis in Tables V–VI. Similarly, for the CMS 8 TeV jet data set CMS8jets'17, the sensitivity to the above flavors can increase under moderate reduction of systematic uncertainties, easily surpassing the sensitivity of CMS7jets'14 because of the larger number of points in CMS8jets'17.

### D. Comparing PDFSense predictions to postfit constraints from Lagrange multiplier scans

How do the surveys based on PDFSense compare against the actual fits? As we noted, the PDFSense method is designed to provide a fast large-scope estimation of the impact of the existing and future data sets in conjunction with other tools, such as the ePump [73] program for PDF reweighting. It works the best in the quadratic (Hessian) approximation near the best fit and when the new experiments are compatible with the old ones. When detailed understanding of the experimental constraints is necessary, the PDFSense approach must be supplemented by other techniques, such as Lagrange multiplier (LM) scans [79,89,90].

As an illustration of the scope of the differences between the PDFSense predictions before and after the fit, the left panels in Figs. 9 and 10 show the PDFSense maps for $d/u(x=0.1, \mu=1.3\ \mathrm{GeV})$ and $g(x = 0.01, \mu = 125\ \mathrm{GeV})$ evaluated using a preliminary CT18 NNLO fit (designated as "CT18pre") that includes 11 new LHC experimental data sets, namely CMS8jets'17, CMS7jets'14, ATLAS7jets'15, LHCb8WZ'16, CMS8Wasy'16, LHCb8Zee'15, LHCb7ZWrap'15, ATL8ZpT'16, ATL8ttb-pt'16, ATL8ttb-mtt'16, and 8 TeV $t\bar{t}$ production at CMS ("CMS8 ttb pTtyt") [91] in addition to the experiments included in the

[5] With the fully correlated jet energy correction JEC2 source, the data set CMS7jets'14 would provide a strong overall constraint on $s(x, \mu)$ comparable to one of the NuTeV or neutrino CCFR experimental data sets.

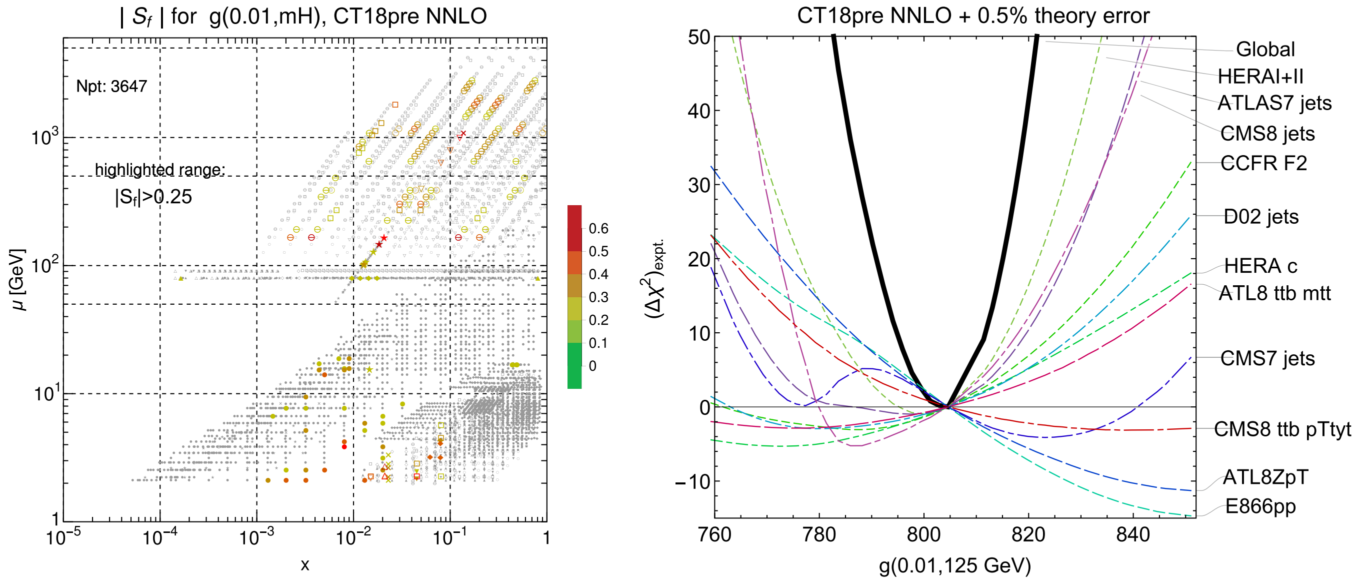

FIG. 10. Like Fig. 9, but comparing the PDFSense map (left) and LM scan (right) for the gluon PDF $g(x = 0.01, \mu = m_H)$ in the Higgs boson production region.

CT14HERA2 fit. The full details of the CT18 fit will be presented in an upcoming publication [92]. Some modifications were made in the methodology adopted in CT18, as compared to CT14HERA2; notably, the PDF parametrization forms and treatment of NNLO radiative contributions have been changed, while some shown curves are also subject to a theoretical uncertainty associated with the QCD scale choices. In accord with the PDFSense predictions based on the CT14HERA2 NNLO PDFs, we find that including the above LHC experiments in the fit produces only mild differences between the CT18pre and CT14HERA2 NNLO PDFs. Consequently, the PDFSense $\{x, \mu\}$ maps based on CT18pre NNLO PDFs are similar to the CT14HERA2 ones [83]. One noticeable difference is that the sensitivity of the new experiments decreases after they are included in the CT18pre fit, because the new information from the newly added experiments suppresses PDF uncertainties of data residuals.

In the right panels of Figs. 9 and 10, we illustrate the constraints on the same quantities, $d/u$ $(0.1, 1.3\ \text{GeV})$ and $g(0.01, 125\ \text{GeV})$, in the candidate CT18pre NNLO fit, now obtained with the help of LM scans. A LM scan [79,89,90] is a powerful technique that elicits detailed information about a PDF-dependent quantity $X(\vec{a})$, such as a PDF or cross section, from a constrained global fit in which the value of $X(\vec{a})$ is fixed by an imposed condition. By minimizing a modified goodness-of-fit function $\chi^2_{\rm LM}(\lambda, \vec{a})$ that includes a "generalized-force" term equal to $X(\vec{a})$ with weight $\lambda$, in addition to the global $\chi^2_{\rm global}$ in Eq. (4), a LM scan reveals the parametric relationship between $X(\vec{a})$ and $\chi^2_{\rm global}$ or $\chi^2_{\rm expt}$ contributions from individual experiments, including any non-Gaussian dependence. In the LM scans at hand, the modified fitted function takes the form

$$\chi^2_{\rm LM}(\lambda, \vec{a}) = \chi^2_{\rm global}(\vec{a}) + \lambda X(\vec{a}), \tag{30}$$

and $X(\vec{a})$ are $d/u(x, \mu)$ or $g(x, \mu)$ at a specific location in $\{x, \mu\}$ space. For the optimal parameter combination $\vec{a} \equiv \vec{a}_0$ at which $\chi^2_{\rm global}(\vec{a})$ is minimized, we find in Fig. 9 that $d/u(0.1, 1.3\ \text{GeV}) \approx 0.7$. The LM scan for the $d/u$ then consists of a series of refits of the parameters $\vec{a}_k$, as the multiplier parameter $\lambda$ is dialed along a set of discrete values $\lambda_k$, effectively pulling $d/u$ away from the value ∼0.7 at $\vec{a} = \vec{a}_0$ preferred by the global fit. The right panel of Fig. 9 shows the relationship between $d/u(0.1, 1.3\ \text{GeV})$ and $\chi^2_{\rm global}$ that is quantified this way and similarly for $g(0.01, 125\ \text{GeV})$.

We can also examine how the $\chi^2$ changes for the individual experiments. Figures 9 and 10 show the curves for 11 experiments with the largest variations $\max(\chi^2) - \min(\chi^2)$ in the shown ranges of $d/u$ and $g$, i.e., the most constraining experiments. We notice that, while the $\Delta\chi^2$ dependence is nearly Gaussian for the total $\chi^2$, it is sometimes less so for the individual experiments. Some experiments may be inconsistent when they have a large best-fit $\chi^2(\vec{a}_0)$ or prefer an incompatible $X$ value. Figure 9 is an example of a good agreement between the experiments, when the individual $\Delta\chi^2_{\rm expt}$ curves are approximately quadratic and minimized at about the same location. Figure 10 shows more pronounced inconsistencies, notably in the case of the E866pp and ATL8ZpT curves that prefer a significantly larger $g(0.01, 125\ \text{GeV})$ than in the rest of the experiments.

The LM procedure thus allows a systematic exploration of the exact constraints from the experiments on $X$ without relying on the Gaussian assumption that is inherent to the PDFSense method. Both PDFSense and LM scans

TABLE I. We list the top ten experiments predicted to drive knowledge of the $d/u$ PDF ratio and of the gluon distribution in the Higgs region according to PDFSense and LM scans. For both, we list the PDFSense evaluations based both on the CT14HERA2 fit and on a preliminary CT18pre fit in the first and second columns on either side of the double-line partition.

| $d/u(x = 0.1, \mu = 1.3$ GeV) | | | $g(x = 0.01, \mu = 125$ GeV) | | |
|---|---|---|---|---|---|
| PDFSense | | LM scan | PDFSense | | LM scan |
| CT14HERA2 | CT18pre | CT18pre | CT14HERA2 | CT18pre | CT18pre |
| HERAI + II’15 | NMCrat’97 | NMCrat’97 | HERAI + II’15 | HERAI + II’15 | HERAI + II’15 |
| BCDMSp’89 | HERAI + II’15 | CCFR-F3’97 | CMS8jets’17 | CMS8jets’17 | CMS8jets’17 |
| NMCrat’97 | BCDMSp’89 | HERAI + II’15 | CMS7jets’14 | CMS7jets’14 | ATL8ZpT’16 |
| CCFR-F3’97 | CCFR-F3’97 | BCDMSd’90 | ATLAS7jets’15 | E866pp’03 | E866pp’03 |
| E866pp’03 | BCDMSd’90 | BCDMSp’89 | E866pp’03 | ATLAS7jets’15 | ATLAS7jets’15 |
| BCDMSd’90 | E605’91 | CDHSW-F3’91 | BCDMSd’90 | BCDMSd’90 | CCFR-F2’01 |
| CDHSW-F3’91 | E866pp’03 | E866rat’01 | CCFR-F3’97 | BCDMSp’89 | D02jets’08 |
| CMS8jets’17 | E866rat’01 | CMS7Masy2’14 | D02jets’08 | D02jets’08 | HERAc’13 |
| E866rat’01 | CMS8jets’17 | NuTeV-nu’06 | NMCrat’97 | NMCrat’97 | NuTeV-nub’06 |
| LHCb8WZ’16 | CDHSW-F3’91 | CMS8jets’17 | BCDMSp’89 | CDHSW-F2’91 | CCFR-F3’97 |

successfully identify the experiments with the strongest sensitivity to $X$, while their specific rankings of such experiments are not strictly identical and reflect the chosen ranking prescription and settings of the global fit. We emphasize that, though informative, the LM scans are computationally intensive, with a typical 30-point scan at NNLO requiring ~6500 CPU core hours on a high-performance cluster. This is in contrast to the PDFSense analysis, which can be run for our entire 4021-point data set on a single CPU core of a modern workstation in ~5 min, representing a $\sim 0.8 \times 10^5$ savings in computational cost.

Let us further illustrate these observations by referring again to Figs. 9 and 10, as well as to Table I that displays the top 10 experiments with the largest cumulative sensitivity to $d/u(0.1, 1.3\ \mathrm{GeV})$ and $g(0.01, 125\ \mathrm{GeV})$ according to PDFSense and LM scans, with either CT14HERA2 or CT18pre PDFs used to construct the PDFSense rankings. In the PDFSense columns, the experiments are ranked in order of descending cumulative sensitivities $\sum_{i=1}^{N_{\mathrm{pt}}} |S_f|(x_i, \mu_i)$ according to the same prescription as in Sec. IV B. For the LM scans, the table shows the experiments that have the largest variations $\max(\chi^2) - \min(\chi^2)$ in the range of $X$ corresponding to $\Delta\chi^2_{\mathrm{global}} \leq 100$, that is, within approximately the 90% probability level interval of the CT18pre NNLO PDFs. As the residual uncertainties $\Delta r_i$ in the sensitivities $S_f$ are normalized to the root-mean-squared residuals $\langle r_0 \rangle_E$ at the best fit, cf. Eq. (21), we similarly divide $\max(\chi^2) - \min(\chi^2)$ by the best-fit $\chi^2(\vec{a}_0)/N_{\mathrm{pt}}$ of the experiment in the rankings for the LM scans in Table I.

From the side-by-side examination of the figures and the table, we can draw a broad conclusion that both the prefit PDFSense and postfit LM scan approaches agree in identifying the most constraining experiments, even though they may result in different orderings of these experiments. This agreement is especially impressive in the instance of $d/u(x = 0.1, \mu = 1.3\ \mathrm{GeV})$, when the rankings agree on eight out of ten leading experiments, confirming the dominance of the NMC $p/d$ ratio, HERAI + II, CCFR $F_3$, and BCDMS $p$ and $d$ measurements. For $g(x = 0.01, \mu = m_H)$, for which we see more tension and non-Gaussian behavior in Fig. 10, both PDFSense and LM scans concur on the crucial role played by the top five to six experiments, namely, HERAI + II, E866pp, and inclusive jet production data from CMS, ATLAS, and D0 Run-2. The upward pull on $g$ from the incompatible ATL8ZpT data set seen in Fig. 10 modifies the rankings of the trailing experiments, such as CMS7 jets or BCDMS. Based upon an extended battery of LM scans we have performed, including the two examples presented here, we conclude that the PDFSense surveys perform as intended.

Lastly, we reiterate that a number of subtleties exists in comparing the results of LM scans and PDFSense sensitivity plots. Most importantly, PDFSense is intended by conception as a tool to quantify the anticipated *average* impact of potentially unfitted data based upon their precision in comparison to the PDF uncertainties. We discussed simplifying assumptions made in PDFSense in order to bypass certain complexities of the full fit and obtain quick estimates. LM scans, on the other hand, provide postfit assessments of the contributions of specific data to the global $\chi^2$ function, as specific quantities predicted by the QCD analysis are dialed away from their optimal values. In the comparisons we made, the detailed pictures produced by both PDFSense and the LM scans depend on a variety of theoretical settings like pQCD scale choices, as well as upon the specific implementation of correlated experimental uncertainties (from up to ~100 different sources in some experiments) and the parametric forms chosen for the nonperturbative parametrizations at the starting scale $\mu = Q_0$. The inclusion of additional theory uncertainties and decorrelation of some experimental correlated errors are necessitated in a few experiments by the relatively large $\chi^2$ values that would otherwise be obtained. All these have some peripheral effect on the specific orderings of experiments shown in Table I. Thus, rather than anticipating an exact point-to-point matching between the PDFSense and LM methods, we instead expect, and indeed find, the general congruity between the most important experiments identified by the two approaches illustrated in this section.

## V. CONCLUSIONS

In the foregoing analysis, we have confronted the modern challenge of a rapidly growing set of global QCD data with new statistical methodologies for quantifying and exploring the impact of this information. These novel methodologies are realized in a new analysis tool PDFSense [83], which allows the rapid exploration of the impact of both existing and potential data on PDF determinations, thus providing a means of weighing the impact of measurements of QCD processes in a way that allows meaningful conclusions to be drawn without the cost of a full global analysis. We expect this approach to guide future PDF fitting efforts by allowing fitters to examine the world's data *a priori*, so as to concentrate analyses on the highest impact data sets. In particular, this work builds upon the existing CT framework with its reliance on the Hessian formalism and assumed quasi-Gaussianity, but these features do not impact the validity of our analysis and conclusions. Our approach provides a means to carry out a detailed study of data residuals, for which we explored novel visualizations in several ways, including the PCA, t-SNE, and reciprocated distance approaches discussed in Sec. II C. These techniques show promise for moving forward by providing useful insights into the numerical relationships among data sets and experimental processes.

Crucial to this analysis is the leveraging of both the existing and proposed statistical measures laid out in Secs. III A and III B. Of these, the flavor-specific sensitivity $S_f$ of Eq. (21) for a data point to the PDF serves as a particularly powerful discriminator, and we deployed it and the correlation $C_f$ of Eq. (19) to map PDF constraints provided by data over a wide range in $\{x, \mu\}$. This was facilitated by the fact that the sensitivity and correlation are readily computable over the extent of the global data set. The companion website collects a large number of figures illustrating the sensitivities to various flavors as a function of $x$ and $\mu$.

To quantify the abundant information contained in the maps of sensitivities, in Sec. IV B, we presented statistical estimators to systematically rank and assess subsidiary data sets within the world's data according to their potential to be influential in constraining PDFs. We note that one is allowed some freedom in choosing a specific ranking prescription, but we find our conclusions to be stable against variations among these possible choices. In this context, we reaffirmed the unique advantage of DIS and jet production for determination of the PDFs.

Many intriguing physics results can be established using our sensitivity methods, and the specific results in the previous sections are only illustrative examples. We stress that these results take the complementary form of sensitivity tables (for example, Table V) and $\{x, \mu\}$ plots (such as Fig. 2), which respectively offer global categorizations of the experimental landscape and detailed mappings of the placements of PDF constraints in $\{x, \mu\}$ space. In totality, the full range of physics insights from this method is beyond the scope of the present article, but the interested reader can explore them using our PDFSense package in Ref. [83]. We mention only a representative sample of these to motivate the reader:

(i) A wide range of experimental processes possess sensitivity to the nucleon's quark sea distributions; for example, for the distribution $\bar{d}(x, \mu)$, the $\sigma_{pd}$ DY measurements of E866 (E866rat'01) exhibit strong sensitivity, but so do DY data from E605 (E605'91) as well as (at larger $\mu$) information on the $\mu$-production asymmetry $A_\mu(\eta)$ from CMS at 7 TeV (CMS7Masy2'14); at high $x$ and $\mu$, CMS inclusive jet data (CMS8jets'17 and CMS7jets'14) also acquire some sensitivity to $\bar{u}$ and $\bar{d}$. Still, however, the recent HERA data (HERAI + II'15) registers the greatest overall sensitivity.

(ii) Were they taken cumulatively together as a single data set, CMS jet production at 7 and 8 TeV (CMS7jets'14 and CMS8jets'17) would provide a total sensitivity $|S_s^E| = 11.9 + 8.11$ to $s(x, \mu)$ that is comparable to one of the NuTeV (NuTeV-nu'06) or CCFR (CCFR SI nu'01 and CCFR SI nub'01) dimuon SIDIS experiments, which have very strong average sensitivity to the strange distribution. Still, the strongest constraint is contributed by a mix of the DIS measurements, including $\nu\mu\mu$ data from NuTeV (NuTeV-nu'06), data on $\nu(\bar{\nu})\mu\mu$ processes from SIDIS at CCFR (CCFR SI nu'01 and CCFR SI nub'01), as well as the inclusive DIS data at lower $x$ from HERA1 + 2 (HERAI + II'15) that actually has the strongest cumulative sensitivity. Similarly, various vector boson production data sets have a rank-3 point-averaged sensitivity to the strangeness, including the $A_\mu(\eta^\mu)$ data from D0 (D02Masy'08) and CMS (CMS8Wasy'16 and CMS7Masy2'14), as well ATLAS $W/Z$ production (ATL8DY2D'16 and ATL7WZ'12) and high-$p_T$ $Z$ production (ATL8ZpT'16) cross sections. Although each of the individual vector boson production data sets has a weak cumulative sensitivity to $s(x, \mu)$ because of a small number of data points, in totality, a group of *mutually consistent* LHC experiments on vector boson production can provide a competing constraint on $s(x, \mu)$ that confronts the low-energy CCFR/NuTeV constraints.

(iii) Knowledge of the charm distribution $c(x,\mu)$ is most influenced by a number of data sets, with HERA (HERAI + II'15) at low $x$ especially important. Fixed target measurements, particularly those of CDHSW on the proton's $F_2^p$ structure function (CDHSW-F2'91) have strong sensitivity at slightly higher $x \sim 10^{-1}$, while a wide range of jet measurements, including 7 TeV data from ATLAS (ATLAS7jets'15) and CMS (CMS7jets'14) and 8 TeV CMS (CMS8jets'17) points are also sensitive. This pattern of sensitive measurements broadly follows the corresponding plot for $|S_g|(x_i,\mu_i)$ [as well as $|S_b|(x_i,\mu_i)$] due to the dominance of boson fusion graphs in heavy quark production. The data sets of importance we identify are broadly consistent with the conclusions of the recent CT14 analysis [93] of the nucleon's intrinsic charm [76].

(iv) One can also study the correlations and sensitivities for various derived PDF combinations. For instance, for the $\bar{d}/\bar{u}$ ratio representing deviations from flavor symmetry in the nucleon sea, the E866 experiment (E866rat'01) shows exceptional point-averaged sensitivity, $\langle|S_{\bar{d}/\bar{u}}|\rangle = 1.67$, such that its C ranking for its overall sensitivity to $\bar{d}/\bar{u}$ places it in the company of only a few other DIS and Drell-Yan (DY) experiments, despite their much larger number of measurements, $N_{\mathrm{pt}} = 15$. At somewhat lower $x \gtrsim 0.01$, NMC data on the structure function ratio $F_2^d/F_2^p$ (NMCrat'97) show sensitivity in the range $0.8 < |S_{\bar{d}/\bar{u}}| < 2$. At still lower $x$, the CMS 8 and 7 TeV $A_\mu$ points (CMS8Wasy'16 and CMS7Masy2'14) and $W/Z$ data from LHCb (LHCb8WZ'16) show strong pull, corresponding to point-averaged rankings of 2, **1**, and 2, respectively.

(v) We also consider the PDF ratio $d/u(x,\mu)$, which often serves as a discriminant among various nucleon structure models, especially at high $x$. For $x > 0.1$ an amalgam of fixed-target experiments, including the NMC $F_2^d/F_2^p$ data (NMCrat'97) particularly, but also $F_2^p$ measurements from BCDMS (BCDMSp'89) and CCFR (CCFR-F2'01) as well as $xF_3^p$ data from CCFR drive the current status. At higher $\mu$, however, the LHCb $W/Z$ data (LHCb8WZ'16) and $A_e(\eta)$ measurements from Run-2 of D0 (D02Easy2'15) also constrain the high $x$ behavior of $d/u$ together with $A_\mu(\eta)$ points from CMS at 7 TeV (CMS7Masy2'14).

(vi) More generally, we note that, among the new LHC experiments to be considered for future global fits, the data sets for inclusive jet production are expected to have the greatest impact, followed by a group of vector boson production experiments at ATLAS, CMS, and LHCb. We find that the constraints from jet production at the LHC depend significantly on the treatment of experimental systematic uncertainties—especially the correlated systematic errors. It is conceivable that, with the full implementation of NNLO theoretical cross sections and modest reduction in the experimental systematic uncertainties, the constraints from the LHC jet production will catch up in strength to the effect of adding a large fixed-target DIS data set, such as BCDMS $F_2^p$ (BCDMSp'89). Meanwhile, the magnitude of the constraint on the gluon PDF from high-$p_T$ $Z$ production (ATL8ZpT'16) is comparable to those from the combined HERA SIDIS charm data set (HERAc'13) or inclusive jet production from CDF Run-2 (CDF2jets'09); that is, the high-$p_T$ $Z$ data are significant in the event that the jet data sets are not included, in overall consistency with the findings in Ref. [64]. The smaller ATLAS $t\bar{t}$ production data sets (ATL8ttb-pt'16, ATL8ttb-y_ave'16, ATL8ttb-mtt'16, and ATL8ttb-y_ttb'16) have strong point-by-point sensitivity to the gluon but will have a more diminished role when combined with other, larger data sets. HERA DIS (HERAI + II'15), BCDMS $F_2^d$ (BCDMSd'90), and CMS inclusive jets at 8 TeV (CMS8jets'17) render the strongest overall constraints on the Higgs production cross section at the LHC according to the rankings in Table VI.

Quantifying correlations and sensitivities thus provide a comprehensive means of evaluating the ability of a global data set to constrain our knowledge of nucleon structure. It must be emphasized, however, that this analysis is not a substitute for actually performing a QCD global analysis, which remains the single most robust means of determining the nucleon PDFs themselves. Rather, the method presented in the paper is a guiding tool to both supplement and direct fits by gauging the potential for improving PDFs with the incorporation of new data sets.

The essential ingredients of this study are the PDF-residual correlation and sensitivity $|C_f|$ and $|S_f|$, with the latter representing an extension of the correlation used elsewhere in the modern PDF literature. These definitions are robust enough that we can exhaustively score the data points in an arbitrary global data set to construct and map the resulting distributions, as shown in Figs. 7 and 8. Accordingly, we found it possible to impose cuts on these distributions to identify points of especially strong correlation ($|C_f| > 0.7$) or sensitivity ($|S_f| > 0.25$); we stress that these cuts are chosen as approximate indicators, and any user can adjust them freely. On the other hand, the distributions themselves, as shown in the second panels of Figs. 7 and 8, are not subject to such cut choices. Although the conclusions of this analysis are resistant to alterations in the basic approach, it is worth noting that other formats are possible for evaluating experimental sensitivities and performing the rankings of measurements. For example, one might use somewhat different matchings than those outlined in Appendix A to extract $\{x,\mu\}$ points from the experimental data, but we expect the resulting impact on the overall picture to be minor. Similarly, while the ordering inside ranking tables like Table V was decided according to the total sensitivity to serve our specific goal of identifying the most valuable experiments for the CTEQ-TEA fit, for other purposes, one might produce alternative tables ranked according to point-averaged sensitivities, or sensitivities to specific flavors. Such alternate conventions would also yield important information, and PDFSense allows the user to do this. It should be stressed that these elections for the form of our presentation can always be recovered from the more fundamental information—the numerical values of the sensitivities detailed in the Supplementary Materials [86].

While we have demonstrated these techniques in the context of the CT14 family of global fits, they are of sufficient generality that one could readily repeat our analysis using alternative PDF sets. For the sake of testing this point and validating our predictions for the most decisive experiments in the CTEQ-TEA data set, we performed a preliminary fit including the CT14HERA2 and the candidate LHC experiments (CT18pre) and directly compared PDFSense predictions against Lagrange multiplier scans quantifying the constraints these fitted measurements imposed on select quantities. This provided a demonstration of the robustness of our sensitivity-based analysis, which identified the same sets of high-impact measurements *before fitting*. The results of this study can be expected to vary somewhat depending on the specifics of the PDF sets used to compute $|C_f|$ and $|S_f|$, but we see this as an advantage of PDFSense. One could imagine exploiting them to undertake a systematic analysis of the impact of various theoretical assumptions implemented in competing global fits (e.g., the choice of input PDF parametrization or the status of the perturbative QCD treatment implemented in various processes). The sensitivity $S_f$ can be constructed either from the Hessian or Monte Carlo PDF uncertainties, as prescribed by Eqs. (21) and (29), while the shifted residuals that are crucial to our analysis can be recovered from any type of covariance matrix, as argued in relation to Eq. (8). In the same spirit but on the side of the data, PDFSense empowers the user to evaluate the combined impact of multiple experimental data sets—for example, to evaluate the extent to which the impact of a proposed experiment might be diminished by the constraints already imposed by existing measurements. These various functions collectively suggest a number of possible avenues to use the presented approach and the PDFSense tool to advance PDF knowledge in the coming years.

## ACKNOWLEDGMENTS

We thank our CTEQ-TEA colleagues, Davison Soper, and Madeline Hamilton for support and insightful discussions and appreciate helpful clarifications concerning the LHC experimental data sets from Alexander Glazov, Uta Klein, Bogdan Malescu, and Klaus Rabbertz. We also thank German Valencia, Ursula Laa, and Dianne Cook for helpful discussions related to data visualizations based on the PCA and t-SNE methods. This work was supported in part by the U.S. Department of Energy under Grant No. DE-SC0010129 and by the National Natural Science Foundation of China under Grant No. 11465018. T. J. Hobbs acknowledges support from an EIC Center Fellowship. The work of J. G. is sponsored by Shanghai Pujiang Program.

## APPENDIX A: APPROXIMATE KINEMATICAL VARIABLES

In this section, we describe in detail our method for identifying the values of $\{x_i, \mu_i\}$ that correspond to experimental data.

For each experimental data point $i$, we can establish an approximate relation between the kinematical quantities for that data point and unobserved quantities specifying the PDFs: the partonic momentum fraction $x$ and QCD factorization scale $\mu$. For example, in DIS, $x$ and $\mu$ are approximately equal to Bjorken $x_B$ and momentum transfer $Q$ according to the Born-level kinematic relation. Although this relation is violated by higher-order radiative contributions, it will approximately hold in most scattering events. The same overall logic can be followed to relate the kinematical quantities in every process of the CTEQ-TEA global set to the *approximate* unobserved quantities $x$ and $\mu$ in the PDFs. These relations vary by process and are used to assign approximate pairs $\{x_i, \mu_i\}$ for each data point.[6]

Specifically, for DIS, which primarily measures the differential cross sections of the form $d^2\sigma/(dx_B dQ^2)$, we simply take

$$\mu_i \approx Q|_i, \qquad x_i \approx x_B|_i \tag{A1}$$

as mentioned above, where the kinematical variables inside $|_i$ are evaluated at their experimentally measured values for the $i$th data point. The above approximate relations hold even when (N)NLO radiative contributions are included.

For one-particle-inclusive particle production in hadron-hadron scattering of the form $AB \to CX$, we plot two $x$ values if the rapidity $y_C$ is known:

[6] It should be pointed out that, while there are 5227 $\{x, \mu\}$ points generated by the 4021 physical measurements in the default CTEQ-TEA data set of this study, occasionally there are instances in which $|C_f|$ and $|S_f|$ cannot be meaningfully computed for select flavors. For example, since the bottom quark PDF $b(x, \mu)$ has no sensible definition below its partonic threshold (i.e., for $\mu < m_b = 4.75$ GeV), it is not possible to evaluate $|S_b|$ for data points extracted at $\mu$ scales below the $b$-quark mass. Similarly, there are situations when the extracted parton fraction $x_i \approx 1$, such that some PDF flavors $f(x_i, \mu_i) \approx 0$, and the Hessian procedures described in this paper do not yield a well-defined correlation or sensitivity. In these cases, we simply redact the associated $\{x_i, \mu_i\}$ points.

$$\mu_i \approx Q|_i, \qquad x_i^{\pm} \approx \frac{Q}{\sqrt{s}}\exp(\pm y_C)|_i. \tag{A2}$$

We set $y_C = 0$ if the rapidity is integrated away. We point out that for processes of this type, Eq. (A2) implies that a measurement in a single rapidity bin can in fact probe two distinct values of $x$; for this and other potential reasons, the number of raw data points in such an experiment ($N_{\rm pt}$) should not be expected to match the number of extracted $\{x, \mu\}$ points in the figures.

In vector boson production, $AB \to (\gamma^*, Z \to \ell\bar{\ell})X$ or $AB \to (W \to \ell\nu_\ell)X$, we set $Q = m_{\ell\bar{\ell}}$ (invariant mass of the lepton pair), and $y_C = y_\ell$ if a single-lepton rapidity is provided or $y_C = y_{\ell\bar{\ell}}$ if the lepton-pair rapidity is provided. If the rapidity $y_\ell$ of the lepton is known, yet $y_{\ell\bar{\ell}}$ of the pair is unknown, we use the fact that $y_\ell \sim y_{\ell\bar{\ell}} \pm 1$ for most events because of the shape of the decay leptonic tensor.

Thus, the momentum fractions $x_i^{\pm}$ can still be estimated as $x_i^{\pm} \approx (Q/\sqrt{s}) \exp(\pm y)|_i$, where $y \sim y_\ell$ (up to an error of less than 1 unit):

(i) In single-inclusive jet production, $AB \to j + X$, we set $Q = 2p_{Tj}$, $y_C = y_j$.

(ii) In single-inclusive $t\bar{t}$ pair production, $AB \to t\bar{t}X$, we set $Q = m_{t\bar{t}}$, $y = y_{t\bar{t}}$ if known, or 0 otherwise.

(iii) In single-inclusive top (anti)quark production, $AB \to (\bar{t})tX$, we take $Q = 2p_{Tt}$, $y = 0$ for $d\sigma/dp_{T_t}$ (as in experiment ATL8ttb-pt'16). On the other hand, for $d\sigma/d\langle y_t\rangle$ or $d\sigma/dy_{t\bar{t}}$, in which the $t\bar{t}$ invariant mass is integrated out (experiments ATL8ttb-y_ave'16 and ATL8ttb-y_ttb'16), we take an average mass scale $\mu_i = 400$ GeV that is slightly above the observed peak of $d\sigma/dm_{t\bar{t}}$ at $m_{t\bar{t}} \approx 2m_t$.

Lastly, for the $d\sigma/dp_T^Z$ measurements from $AB \to (\gamma^*, Z \to \ell\bar{\ell})X$ in experiments ATL7ZpT'14 and ATL8ZpT'16, we take $Q = \sqrt{(p_T^Z)^2 + (M_Z)^2}$, $y_C = y_Z$. (Here, $Q$ denotes the boson's transverse mass, not the invariant mass.)

## APPENDIX B: TABULATED RESULTS

In Tables II–IV, we provide a detailed key for the individual experiments mapped in Fig. 1, including the physical process, number of points, and luminosities, where available. We group these tables broadly according to subprocess—Table II corresponds to DIS experiments, while Tables III and IV collect various measurements for the hadroproduction of, e.g., gauge boson, jet, and $t\bar{t}$ pairs—and thus provide a translation key for the experimental shorthand names given in Fig. 1.

In Tables V and VI, we collect the flavor-specific ($|S_f^E|$) and overall ($\sum_f |S_f^E|$) sensitivities for the experimental data sets contained in this analysis. In Table V, we list the total and point-averaged sensitivities for each main flavor ($\bar{d}$, $\bar{u}$, $g$, $u$, $d$, $s$), while Table VI gives the corresponding information for a number of quantities derived from these, as explained in the associated captions.

In the Supplementary Materials [86], we enclose a series of additional tables that further illustrate the details of our sensitivity analysis. These include a detailed breakdown of the various CTEQ-TEA experiments according to physical process (Supplemental Table I) and associated sensitivity rankings, both for individual PDF flavors (Supplemental Table II) and for various derived quantities (Supplemental Table III). In addition, in Supplemental Tables IV and V, we give numerical values of sensitivities corresponding to the

TABLE II. Experimental data sets considered as part of CT14HERA2 and included in this analysis: deep-inelastic scattering. We point out that the numbering scheme (CT ID#) included in this and subsequent tables follows the standard CTEQ labeling system with, e.g., experiment identifications of the form 1XX representing DIS experiments, etc.

| Experiment name | CT ID# | Data set details | | $N_{\rm pt}$ |
|---|---|---|---|---|
| BCDMSp'89 | 101 | BCDMS $F_2^p$ | [11] | 337 |
| BCDMSd'90 | 102 | BCDMS $F_2^d$ | [12] | 250 |
| NMCrat'97 | 104 | NMC $F_2^d/F_2^p$ | [13] | 123 |
| CDHSW-F2'91 | 108 | CDHSW $F_2^p$ | [14] | 85 |
| CDHSW-F3'91 | 109 | CDHSW $F_3^p$ | [14] | 96 |
| CCFR-F2'01 | 110 | CCFR $F_2^p$ | [15] | 69 |
| CCFR-F3'97 | 111 | CCFR $xF_3^p$ | [16] | 86 |
| NuTeV-nu'06 | 124 | NuTeV $\nu\mu\mu$ SIDIS | [17] | 38 |
| NuTeV-nub'06 | 125 | NuTeV $\bar{\nu}\mu\mu$ SIDIS | [17] | 33 |
| CCFR SI nu'01 | 126 | CCFR $\nu\mu\mu$ SIDIS | [18] | 40 |
| CCFR SI nub'01 | 127 | CCFR $\bar{\nu}\mu\mu$ SIDIS | [18] | 38 |
| HERAb'06 | 145 | H1 $\sigma_r^b$ (57.4 pb$^{-1}$) | [19,20] | 10 |
| HERAc'13 | 147 | Combined HERA charm production (1.504 fb$^{-1}$) | [21] | 47 |
| HERAI + II'15 | 160 | HERA1 + 2 combined NC and CC DIS (1 fb$^{-1}$) | [6] | 1120 |
| HERA-FL'11 | 169 | H1 $F_L$ (121.6 pb$^{-1}$) | [22] | 9 |

TABLE III. Same as Table II, showing experimental data sets for production of vector bosons, single-inclusive jets, and $t\bar{t}$ pairs.

| Experiment name | CT ID# | Data set details | | $N_{\mathrm{pt}}$ |
|---|---|---|---|---|
| E605'91 | 201 | E605 DY | [23] | 119 |
| E866rat'01 | 203 | E866 DY, $\sigma_{pd}/(2\sigma_{pp})$ | [24] | 15 |
| E866pp'03 | 204 | E866 DY, $Q^3 d^2\sigma_{pp}/(dQdx_F)$ | [25] | 184 |
| CDF1Wasy'96 | 225 | CDF Run-1 $A_e(\eta^e)$ (110 $\mathrm{pb}^{-1}$) | [26] | 11 |
| CDF2Wasy'05 | 227 | CDF Run-2 $A_e(\eta^e)$ (170 $\mathrm{pb}^{-1}$) | [27] | 11 |
| D02Masy'08 | 234 | DØ Run-2 $A_\mu(\eta^\mu)$ (0.3 $\mathrm{fb}^{-1}$) | [28] | 9 |
| LHCb7WZ'12 | 240 | LHCb 7 TeV $W/Z$ muon forward-$\eta$ Xsec (35 $\mathrm{pb}^{-1}$) | [29] | 14 |
| LHCb7Wasy'12 | 241 | LHCb 7 TeV $W$ $A_\mu(\eta^\mu)$ (35 $\mathrm{pb}^{-1}$) | [29] | 5 |
| ZyD02'08 | 260 | DØ Run-2 $Z$ $d\sigma/dy_Z$ (0.4 $\mathrm{fb}^{-1}$) | [30] | 28 |
| ZyCDF2'10 | 261 | CDF Run-2 $Z$ $d\sigma/dy_Z$ (2.1 $\mathrm{fb}^{-1}$) | [31] | 29 |
| CMS7Masy2'14 | 266 | CMS 7 TeV $A_\mu(\eta)$ (4.7 $\mathrm{fb}^{-1}$) | [32] | 11 |
| CMS7Easy'12 | 267 | CMS 7 TeV $A_e(\eta)$ (0.840 $\mathrm{fb}^{-1}$) | [33] | 11 |
| ATL7WZ'12 | 268 | ATLAS 7 TeV $W/Z$ Xsec, $A_\mu(\eta)$ (35 $\mathrm{pb}^{-1}$) | [34] | 41 |
| D02Easy2'15 | 281 | DØ Run-2 $A_e(\eta)$ (9.7 $\mathrm{fb}^{-1}$) | [35] | 13 |
| CDF2jets'09 | 504 | CDF Run-2 inclusive jet $(d^2\sigma/dp_T^j dy_j)$ (1.13 $\mathrm{fb}^{-1}$) | [36] | 72 |
| D02jets'08 | 514 | DØ Run-2 inclusive jet $(d^2\sigma/dp_T^j dy_j)$ (0.7 $\mathrm{fb}^{-1}$) | [37] | 110 |
| ATL7jets'12 | 535 | ATLAS 7 TeV inclusive jet $(d^2\sigma/dp_T^j dy_j)$ (35 $\mathrm{pb}^{-1}$) | [38] | 90 |
| CMS7jets'13 | 538 | CMS 7 TeV inclusive jet $(d^2\sigma/dp_T^j dy_j)$ (5 $\mathrm{fb}^{-1}$) | [39] | 133 |

TABLE IV. Same as Table II, showing experimental data sets for production of vector bosons, single-inclusive jets, and $t\bar{t}$ pairs that were not incorporated in the CT14HERA2 fit but included in our augmented CTEQ-TEA set.

| Experiment name | CT ID# | Data set details | | $N_{\mathrm{pt}}$ |
|---|---|---|---|---|
| LHCb7ZWrap'15 | 245 | LHCb 7 TeV Z/W muon forward-$\eta$ Xsec (1.0 $\mathrm{fb}^{-1}$) | [40] | 33 |
| LHCb8Zee'15 | 246 | LHCb 8 TeV Z electron forward-$\eta$ $d\sigma/dy_Z$ (2.0 $\mathrm{fb}^{-1}$) | [41] | 17 |
| ATL7ZpT'14 | 247 | ATLAS 7 TeV $d\sigma/dp_T^Z$ (4.7 $\mathrm{fb}^{-1}$) | [42] | 8 |
| XCMS8Wasy'16 | 249 | CMS 8 TeV W muon, Xsec, $A_\mu(\eta^\mu)$ (18.8 $\mathrm{fb}^{-1}$) | [43] | 33 |
| LHCb8WZ'16 | 250 | LHCb 8 TeV W/Z muon, Xsec, $A_\mu(\eta^\mu)$ (2.0 $\mathrm{fb}^{-1}$) | [44] | 42 |
| ATL8DY2D'16 | 252 | ATLAS 8 TeV Z $(d^2\sigma/d\|y\|_{ll}dm_{ll})$ (20.3 $\mathrm{fb}^{-1}$) | [45] | 48 |
| ATL8ZpT'16 | 253 | ATLAS 8 TeV $(d^2\sigma/dp_T^Z dm_{ll})$ (20.3 $\mathrm{fb}^{-1}$) | [46] | 45 |
| CMS7jets'14 | 542 | CMS 7 TeV inclusive jet, $R = 0.7$, $(d^2\sigma/dp_T^j dy_j)$ (5 $\mathrm{fb}^{-1}$) | [47] | 158 |
| ATLAS7jets'15 | 544 | ATLAS 7 TeV inclusive jet, $R = 0.6$, $(d^2\sigma/dp_T^j dy_j)$ (4.5 $\mathrm{fb}^{-1}$) | [48] | 140 |
| CMS8jets'17 | 545 | CMS 8 TeV inclusive jet, $R = 0.7$, $(d^2\sigma/dp_T^j dy_j)$ (19.7 $\mathrm{fb}^{-1}$) | [49] | 185 |
| ATL8ttb-pt'16 | 565 | ATLAS 8 TeV $t\bar{t}d\sigma/dp_T^t$ (20.3 $\mathrm{fb}^{-1}$) | [50] | 8 |
| ATL8ttb-y_ave'16 | 566 | ATLAS 8 TeV $t\bar{t}d\sigma/dy_{\langle t/\bar{t}\rangle}$ (20.3 $\mathrm{fb}^{-1}$) | [50] | 5 |
| ATL8ttb-mtt'16 | 567 | ATLAS 8 TeV $t\bar{t}d\sigma/dm_{t\bar{t}}$ (20.3 $\mathrm{fb}^{-1}$) | [50] | 7 |
| ATL8ttb-y_ttb'16 | 568 | ATLAS 8 TeV $t\bar{t}d\sigma/dy_{t\bar{t}}$ (20.3 $\mathrm{fb}^{-1}$) | [50] | 5 |

TABLE V. For each experiment $E$, we have defined its flavor-specific sensitivity $|S_f^E|$ and its point-averaged counterpart $\langle|S_f^E|\rangle$ in Sec. IV B. Using these quantities, we tabulate the total overall (i.e., flavor-summed) sensitivity and a flavor-dependent sensitivity for the various experiments in our data set, ordering the table in descending magnitude for the overall sensitivity. Thus, row 1 for the combined HERA Run I + Run 2 data set has the greatest overall sensitivity, while row 47 for the H1 $\sigma_r^b$ reduced cross section has the least overall sensitivity according to that metric. For each flavor, we award particularly sensitive experiments a rank A, B, C or 1∗,1,2,3 based on their total and point-averaged sensitivities, respectively. These ranks are decided using the criteria: $C \Leftrightarrow |S_f^E| \in [20, 50]$, $B \Leftrightarrow |S_f^E| \in [50, 100]$, and $A \Leftrightarrow |S_f^E| > 100$ according to the total sensitivities for each flavor and, analogously, $3 \Leftrightarrow \langle|S_f^E|\rangle \in [0.1, 0.25]$, $2 \Leftrightarrow \langle|S_f^E|\rangle \in [0.25, 0.5]$, $1 \Leftrightarrow \langle|S_f^E|\rangle \in [0.5, 1]$, and $1{*} \Leftrightarrow \langle|S_f^E|\rangle > 1$ according to the point-averaged sensitivities. Experiments with sensitivities falling below the lowest ranks (that is, with $|S_f^E| < 20$ or $\langle|S_f^E|\rangle < 0.1$) are not awarded a rank for that category/flavor. Note that we sum over the light quark + gluon flavors to compute $\langle\sum_f|S_f^E|\rangle$ within this and subsequent tables. Also, new experimental data sets not originally included in CT14HERA2 are indicated by bold experiment names in the second column.

| | | | | | Rankings, CT14 HERA2 NNLO PDFs | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| No. | Experiment | $N_{\rm pt}$ | $\sum_f\lvert S_f^E\rvert$ | $\langle\sum_f\lvert S_f^E\rvert\rangle$ | $\lvert S_{\bar d}^E\rvert$ | $\langle\lvert S_{\bar d}^E\rvert\rangle$ | $\lvert S_{\bar u}^E\rvert$ | $\langle\lvert S_{\bar u}^E\rvert\rangle$ | $\lvert S_g^E\rvert$ | $\langle\lvert S_g^E\rvert\rangle$ | $\lvert S_u^E\rvert$ | $\langle\lvert S_u^E\rvert\rangle$ | $\lvert S_d^E\rvert$ | $\langle\lvert S_d^E\rvert\rangle$ | $\lvert S_s^E\rvert$ | $\langle\lvert S_s^E\rvert\rangle$ |
| 1 | HERAI + II'15 | 1120. | 620. | 0.0922 | B | | A | 3 | A | 3 | A | 3 | B | | C | |
| 2 | CCFR-F3'97 | 86 | 218. | 0.423 | C | 1 | C | 1 | | 3 | B | 1 | C | 2 | | |
| 3 | BCDMSp'89 | 337 | 184. | 0.0908 | | | C | | C | | B | 3 | C | | | |
| 4 | NMCrat'97 | 123 | 169. | 0.229 | C | 2 | | | | | C | 2 | B | 2 | | |
| 5 | BCDMSd'90 | 250 | 141. | 0.0939 | C | | | | C | 3 | C | 3 | C | 3 | | |
| 6 | CDHSW-F3'91 | 96 | 115. | 0.199 | C | 2 | C | 2 | | 3 | C | 2 | C | 3 | | |
| 7 | E605'91 | 119 | 113. | 0.158 | C | 2 | C | 2 | | | | 3 | | | | |
| 8 | E866pp'03 | 184 | 103. | 0.0935 | | 3 | C | 3 | | | C | 3 | | | | |
| 9 | CCFR-F2'01 | 69 | 89.1 | 0.215 | | 3 | | 3 | C | 2 | | 3 | | 2 | | 3 |
| 10 | **CMS8jets'17** | 185 | 87.6 | 0.0789 | | | | | C | 3 | | | | | | |
| 11 | CDHSW-F2'91 | 85 | 82.4 | 0.162 | | 3 | | 3 | | 3 | | 3 | C | 3 | | |
| 12 | CMS7jets'13 | 133 | 63.8 | 0.0799 | | | | | C | 3 | | | | | | |
| 13 | NuTeV-nu'06 | 38 | 58.9 | 0.259 | | 3 | | 3 | | | | 3 | | 3 | C | 1 |
| 14 | **CMS7jets'14** | 158 | 57.5 | 0.0606 | | | | | C | 3 | | | | | | |
| 15 | CCFR SI nub'01 | 38 | 49.4 | 0.217 | | 3 | | 3 | | | | 3 | | 3 | C | 1 |
| 16 | **ATLAS7jets'15** | 140 | 48.2 | 0.0574 | | | | | | 3 | | | | | | |
| 17 | CCFR SI nu'01 | 40 | 48. | 0.2 | | 3 | | 3 | | | | 3 | | 3 | C | 1 |
| 18 | **LHCb8WZ'16** | 42 | 41.4 | 0.164 | | 3 | | 3 | | 3 | | 3 | | 2 | | |
| 19 | ATL7WZ'12 | 41 | 39.6 | 0.161 | | 3 | | 3 | | | | 3 | | 3 | | 3 |
| 20 | **CMS8Wasy'16** | 33 | 39.2 | 0.198 | | 2 | | 3 | | | | 3 | | 2 | | 3 |
| 21 | D02jets'08 | 110 | 37.5 | 0.0568 | | | | | | 3 | | | | | | |
| 22 | NuTeV-nub'06 | 33 | 36.7 | 0.185 | | 3 | | 3 | | | | 3 | | 3 | | 2 |
| 23 | **ATL8DY2D'16** | 48 | 34.7 | 0.121 | | 3 | | 3 | | | | 3 | | | | 3 |
| 24 | E866rat'01 | 15 | 33.3 | 0.37 | | 1 | | 1 | | | | 3 | | 2 | | |
| 25 | ATL7jets'12 | 90 | 30.4 | 0.0563 | | | | | | 3 | | | | | | |
| 26 | **LHCb7ZWrap'15** | 33 | 30.2 | 0.152 | | 3 | | 3 | | 3 | | 3 | | 3 | | |
| 27 | CMS7Masy2'14 | 11 | 29.4 | 0.446 | | 1 | | 2 | | 2 | | 2 | | 1 | | 3 |
| 28 | CDF2jets'09 | 72 | 21.5 | 0.0497 | | | | | | 3 | | | | | | |
| 29 | **ATL8ZpT'16** | 45 | 17.2 | 0.0638 | | | | | | 3 | | | | | | 3 |
| 30 | HERAc'13 | 47 | 15.1 | 0.0537 | | | | | | 3 | | | | | | |
| 31 | D02Masy'08 | 9 | 15. | 0.278 | | 3 | | 3 | | | | 2 | | 2 | | 2 |
| 32 | CMS7Easy'12 | 11 | 14.3 | 0.216 | | 2 | | 3 | | 3 | | 3 | | 2 | | |
| 33 | D02Easy2'15 | 13 | 14. | 0.18 | | 3 | | 3 | | | | 3 | | 2 | | |
| 34 | ZyD02'08 | 28 | 11.6 | 0.0693 | | | | | | | | 3 | | 3 | | |
| 35 | ZyCDF2'10 | 29 | 11.2 | 0.0647 | | | | | | | | 3 | | | | |
| 36 | CDF1Wasy'96 | 11 | 8.83 | 0.134 | | 3 | | 3 | | | | 3 | | 2 | | |

(Table continued)

TABLE V. *(Continued)*

| | | | | | Rankings, CT14 HERA2 NNLO PDFs | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| No. | Experiment | $N_{\mathrm{pt}}$ | $\sum_f \lvert S_f^E\rvert$ | $\langle\sum_f \lvert S_f^E\rvert\rangle$ | $\lvert S_{\bar d}^E\rvert$ | $\langle\lvert S_{\bar d}^E\rvert\rangle$ | $\lvert S_{\bar u}^E\rvert$ | $\langle\lvert S_{\bar u}^E\rvert\rangle$ | $\lvert S_g^E\rvert$ | $\langle\lvert S_g^E\rvert\rangle$ | $\lvert S_u^E\rvert$ | $\langle\lvert S_u^E\rvert\rangle$ | $\lvert S_d^E\rvert$ | $\langle\lvert S_d^E\rvert\rangle$ | $\lvert S_s^E\rvert$ | $\langle\lvert S_s^E\rvert\rangle$ |
| 37 | LHCb7WZ’12 | 14 | 7.27 | 0.0866 | | 3 | | | | | | | | 3 | | |
| 38 | **LHCb8Zee’15** | 17 | 7.1 | 0.0696 | | | | | | | | 3 | | | | |
| 39 | **ATL8ttb-pt’16** | 8 | 6.2 | 0.129 | | 3 | | 3 | | 2 | | | | | | |
| 40 | LHCb7Wasy’12 | 5 | 6.11 | 0.204 | | 2 | | 3 | | | | 3 | | 2 | | 3 |
| 41 | **ATL7ZpT’14** | 8 | 5.84 | 0.122 | | 3 | | 3 | | 3 | | 3 | | 3 | | |
| 42 | HERA-FL’11 | 9 | 3.99 | 0.0739 | | | | | | 2 | | | | | | |
| 43 | **ATL8ttb-mtt’16** | 7 | 3.81 | 0.0907 | | | | | | 2 | | | | | | |
| 44 | CDF2Wasy’05 | 11 | 3.7 | 0.056 | | | | | | | | | | 3 | | |
| 45 | **ATL8ttb-y_ttb’16** | 5 | 3.37 | 0.112 | | | | | | 2 | | | | | | |
| 46 | **ATL8ttb-y_ave’16** | 5 | 3.2 | 0.107 | | | | | | 2 | | | | | | |
| 47 | HERAb’06 | 10 | 1.14 | 0.0191 | | | | | | | | | | | | |

TABLE VI. A horizontal continuation of the information in Table V, containing the flavor-dependent total and mean sensitivities of a number of derived quantities, as opposed to the individual flavors given in Table V. Going across, the total and mean sensitivities are tabulated for valence distributions of the $u$ and $d$ quarks, the partonic flavor ratios $\bar d/\bar u$ and $d/u$, and the Higgs production cross section $\sigma_{pp\to H^0X}$ at 7, 8, and 14 TeV, respectively. The ranking criteria, ordering, and other conventions are again as described in Table V.

| | | | | | | Rankings, CT14 HERA2 NNLO PDFs | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| No. | Expt. | $\lvert S_{u_v}^E\rvert$ | $\langle\lvert S_{u_v}^E\rvert\rangle$ | $\lvert S_{d_v}^E\rvert$ | $\langle\lvert S_{d_v}^E\rvert\rangle$ | $\lvert S_{\bar d/\bar u}^E\rvert$ | $\langle\lvert S_{\bar d/\bar u}^E\rvert\rangle$ | $\lvert S_{d/u}^E\rvert$ | $\langle\lvert S_{d/u}^E\rvert\rangle$ | $\lvert S_{H7}^E\rvert$ | $\langle\lvert S_{H7}^E\rvert\rangle$ | $\lvert S_{H8}^E\rvert$ | $\langle\lvert S_{H8}^E\rvert\rangle$ | $\lvert S_{H14}^E\rvert$ | $\langle\lvert S_{H14}^E\rvert\rangle$ |
| 1 | HERAI + II’15 | B | | C | | C | | B | | B | | B | | B | |
| 2 | CCFR-F3’97 | B | 1 | B | 1 | | | C | 2 | | 3 | | 3 | | 3 |
| 3 | BCDMSp’89 | B | 3 | C | | C | | C | 3 | C | | | | | |
| 4 | NMCrat’97 | C | 2 | C | 3 | C | 2 | B | 1 | | | | | | |
| 5 | BCDMSd’90 | C | | C | 3 | | | C | | C | | C | | | |
| 6 | CDHSW-F3’91 | C | 2 | C | 2 | | | | 3 | | | | | | |
| 7 | E605’91 | C | 3 | C | 3 | | | | | | | | | | |
| 8 | E866pp’03 | C | 3 | | | | | | | | | | | | |
| 9 | CCFR-F2’01 | | 3 | | 3 | | 3 | | 3 | | 3 | | 3 | | 3 |
| 10 | **CMS8jets’17** | | | | | | | | | | 3 | C | 3 | C | 3 |
| 11 | CDHSW-F2’91 | | 3 | | 3 | | | | 3 | | 3 | | 3 | | |
| 12 | CMS7jets’13 | | | | | | | | | | 3 | | 3 | | 3 |
| 13 | NuTeV-nu’06 | | | | | | | | | | | | | | |
| 14 | **CMS7jets’14** | | | | | | | | | | 3 | | 3 | | 3 |
| 15 | CCFR SI nub’01 | | | | | | | | | | | | | | |
| 16 | **ATLAS7jets’15** | | | | | | | | | | | | | | |
| 17 | CCFR SI nu’01 | | | | | | | | | | | | | | |
| 18 | **LHCb8WZ’16** | | 3 | | 3 | | 2 | | 2 | | 3 | | 3 | | |
| 19 | ATL7WZ’12 | | 3 | | | | 3 | | 3 | | | | | | |
| 20 | **CMS8Wasy’16** | | 3 | | 3 | | 2 | | 2 | | | | | | |
| 21 | D02jets’08 | | | | | | | | | | | | | | |
| 22 | NuTeV-nub’06 | | | | | | | | | | | | | | |
| 23 | **ATL8DY2D’16** | | 3 | | | | 3 | | 3 | | | | | | |
| 24 | E866rat’01 | | 2 | | 2 | C | 1* | | 2 | | 3 | | 3 | | |
| 25 | ATL7jets’12 | | | | | | | | | | 3 | | 3 | | 3 |
| 26 | **LHCb7ZWrap’15** | | 3 | | 3 | | 2 | | 2 | | 3 | | 3 | | |

*(Table continued)*

TABLE VI. *(Continued)*

| | | Rankings, CT14 HERA2 NNLO PDFs | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| No. | Expt. | $\lvert S^E_{u_v}\rvert$ | $\langle\lvert S^E_{u_v}\rvert\rangle$ | $\lvert S^E_{d_v}\rvert$ | $\langle\lvert S^E_{d_v}\rvert\rangle$ | $\lvert S^E_{\bar{d}/\bar{u}}\rvert$ | $\langle\lvert S^E_{\bar{d}/\bar{u}}\rvert\rangle$ | $\lvert S^E_{d/u}\rvert$ | $\langle\lvert S^E_{d/u}\rvert\rangle$ | $\lvert S^E_{H7}\rvert$ | $\langle\lvert S^E_{H7}\rvert\rangle$ | $\lvert S^E_{H8}\rvert$ | $\langle\lvert S^E_{H8}\rvert\rangle$ | $\lvert S^E_{H14}\rvert$ | $\langle\lvert S^E_{H14}\rvert\rangle$ |
| 27 | CMS7Masy2'14 | | 2 | | 2 | | 1 | | 1 | | 3 | | 3 | | 3 |
| 28 | CDF2jets'09 | | | | | | | | | | | | | | |
| 29 | **ATL8ZpT'16** | | | | | | | | | | | | | | 3 |
| 30 | HERAc'13 | | | | | | | | | | 3 | | 3 | | 3 |
| 31 | D02Masy'08 | | 2 | | 2 | | 2 | | 2 | | | | | | 3 |
| 32 | CMS7Easy'12 | | 3 | | 3 | | 2 | | 2 | | | | | | |
| 33 | D02Easy2'15 | | 3 | | 2 | | 3 | | 2 | | | | | | |
| 34 | ZyD02'08 | | 3 | | | | | | | | | | | | |
| 35 | ZyCDF2'10 | | 3 | | | | | | | | | | | | |
| 36 | CDF1Wasy'96 | | 3 | | 2 | | 3 | | 2 | | | | | | |
| 37 | LHCb7WZ'12 | | | | | | 3 | | 3 | | | | | | |
| 38 | **LHCb8Zee'15** | | | | | | | | | | | | | | |
| 39 | **ATL8ttb-pt'16** | | 3 | | | | | | | | 2 | | 2 | | 2 |
| 40 | LHCb7Wasy'12 | | 3 | | 3 | | 2 | | 2 | | 3 | | 3 | | 3 |
| 41 | **ATL7ZpT'14** | | | | | | | | 3 | | 3 | | 3 | | 3 |
| 42 | HERA-FL'11 | | | | | | | | | | 3 | | 3 | | |
| 43 | **ATL8ttb-mtt'16** | | | | | | | | | | 3 | | 3 | | 3 |
| 44 | CDF2Wasy'05 | | | | 3 | | | | 3 | | | | | | |
| 45 | **ATL8ttb-y_ttb'16** | | | | | | | | | | 2 | | 2 | | 3 |
| 46 | **ATL8ttb-y_ave'16** | | | | | | | | | | 2 | | 2 | | 3 |
| 47 | HERAb'06 | | | | | | | | | | | | | | |

rankings shown in Tables V and VI. In Supplemental Tables VI and VII, numerical values of sensitivities corresponding to Supplemental Tables II and III are also given. Lastly, in Supplemental Tables VIII and IX, sensitivity ranking tables of the CTEQ-TEA data set based upon a companion fit that excluded jet data are given, and corresponding numerical values are shown in Supplemental Tables X and XI.

[1] S. Dulat, T.-J. Hou, J. Gao, M. Guzzi, J. Huston, P. Nadolsky, J. Pumplin, C. Schmidt, D. Stump, and C.-P. Yuan, Phys. Rev. D **93**, 033006 (2016).

[2] L. A. Harland-Lang, A. D. Martin, P. Motylinski, and R. S. Thorne, Eur. Phys. J. C **75**, 204 (2015).

[3] R. D. Ball *et al.* (NNPDF Collaboration), Eur. Phys. J. C **77**, 663 (2017).

[4] S. Alekhin, J. Blümlein, S. Moch, and R. Placakyte, Phys. Rev. D **96**, 014011 (2017).

[5] A. Accardi, L. T. Brady, W. Melnitchouk, J. F. Owens, and N. Sato, Phys. Rev. D **93**, 114017 (2016).

[6] H. Abramowicz *et al.* (ZEUS and H1 Collaboration), Eur. Phys. J. C **75**, 580 (2015).

[7] S. Alekhin *et al.*, Eur. Phys. J. C **75**, 304 (2015).

[8] J. Gao, L. Harland-Lang, and J. Rojo, Phys. Rep. **742**, 1 (2018).

[9] J. Butterworth *et al.*, J. Phys. G **43**, 023001 (2016).

[10] T.-J. Hou, S. Dulat, J. Gao, M. Guzzi, J. Huston, P. Nadolsky, J. Pumplin, C. Schmidt, D. Stump, and C.-P. Yuan, Phys. Rev. D **95**, 034003 (2017).

[11] A. C. Benvenuti *et al.* (BCDMS Collaboration), Phys. Lett. B **223**, 485 (1989).

[12] A. C. Benvenuti *et al.* (BCDMS Collaboration), Phys. Lett. B **237**, 592 (1990).

[13] M. Arneodo *et al.* (New Muon Collaboration), Nucl. Phys. **B483**, 3 (1997).

[14] J. P. Berge *et al.*, Z. Phys. C **49**, 187 (1991).

[15] U.-K. Yang *et al.* (CCFR/NuTeV Collaboration), Phys. Rev. Lett. **86**, 2742 (2001).

[16] W. G. Seligman *et al.*, Phys. Rev. Lett. **79**, 1213 (1997).

[17] D. A. Mason, Ph. D. thesis, Oregon University, 2006, http://lss.fnal.gov/archive/thesis/2000/fermilab-thesis-2006-01.pdf.

[18] M. Goncharov *et al.* (NuTeV Collaboration), Phys. Rev. D **64**, 112006 (2001).

[19] A. Aktas *et al.* (H1 Collaboration), Eur. Phys. J. C **40**, 349 (2005).
[20] A. Aktas *et al.* (H1 Collaboration), Eur. Phys. J. C **45**, 23 (2006).
[21] H. Abramowicz *et al.* (ZEUS and H1 Collaboration), Eur. Phys. J. C **73**, 2311 (2013).
[22] F. D. Aaron *et al.* (H1 Collaboration), Eur. Phys. J. C **71**, 1579 (2011).
[23] G. Moreno *et al.*, Phys. Rev. D **43**, 2815 (1991).
[24] R. S. Towell *et al.* (NuSea Collaboration), Phys. Rev. D **64**, 052002 (2001).
[25] J. C. Webb *et al.* (NuSea Collaboration), arXiv:hep-ex/0302019.
[26] F. Abe *et al.* (CDF Collaboration), Phys. Rev. Lett. **77**, 2616 (1996).
[27] D. Acosta *et al.* (CDF Collaboration), Phys. Rev. D **71**, 051104 (2005).
[28] V. M. Abazov *et al.* (D0 Collaboration), Phys. Rev. D **77**, 011106 (2008).
[29] R. Aaij *et al.* (LHCb Collaboration), J. High Energy Phys. 06 (2012) 058.
[30] V. M. Abazov *et al.* (D0 Collaboration), Phys. Lett. B **658**, 112 (2008).
[31] T. A. Aaltonen *et al.* (CDF Collaboration), Phys. Lett. B **692**, 232 (2010).
[32] S. Chatrchyan *et al.* (CMS Collaboration), Phys. Rev. D **90**, 032004 (2014).
[33] S. Chatrchyan *et al.* (CMS Collaboration), Phys. Rev. Lett. **109**, 111806 (2012).
[34] G. Aad *et al.* (ATLAS Collaboration), Phys. Rev. D **85**, 072004 (2012).
[35] V. M. Abazov *et al.* (D0 Collaboration), Phys. Rev. D **91**, 032007 (2015); **91**, 079901(E) (2015).
[36] T. Aaltonen *et al.* (CDF Collaboration), Phys. Rev. D **78**, 052006 (2008); **79**, 119902(E) (2009).
[37] V. M. Abazov *et al.* (D0 Collaboration), Phys. Rev. Lett. **101**, 062001 (2008).
[38] G. Aad *et al.* (ATLAS Collaboration), Phys. Rev. D **86**, 014022 (2012).
[39] S. Chatrchyan *et al.* (CMS Collaboration), Phys. Rev. D **87**, 112002 (2013); **87**, 119902(E) (2013).
[40] R. Aaij *et al.* (LHCb Collaboration), J. High Energy Phys. 08 (2015) 039.
[41] R. Aaij *et al.* (LHCb Collaboration), J. High Energy Phys. 05 (2015) 109.
[42] G. Aad *et al.* (ATLAS Collaboration), J. High Energy Phys. 09 (2014) 145.
[43] V. Khachatryan *et al.* (CMS Collaboration), Eur. Phys. J. C **76**, 469 (2016).
[44] R. Aaij *et al.* (LHCb Collaboration), J. High Energy Phys. 01 (2016) 155.
[45] G. Aad *et al.* (ATLAS Collaboration), J. High Energy Phys. 08 (2016) 009.
[46] G. Aad *et al.* (ATLAS Collaboration), Eur. Phys. J. C **76**, 291 (2016).
[47] S. Chatrchyan *et al.* (CMS Collaboration), Phys. Rev. D **90**, 072006 (2014).
[48] G. Aad *et al.* (ATLAS Collaboration), J. High Energy Phys. 02 (2015) 153; 09 (2015) 141(E).
[49] V. Khachatryan *et al.* (CMS Collaboration), J. High Energy Phys. 03 (2017) 156.
[50] G. Aad *et al.* (ATLAS Collaboration), Eur. Phys. J. C **76**, 538 (2016).
[51] S. Camarda *et al.* (HERAFitter Developers' Team), Eur. Phys. J. C **75**, 458 (2015).
[52] R. D. Ball, V. Bertone, F. Cerutti, L. Del Debbio, S. Forte, A. Guffanti, J. I. Latorre, J. Rojo, and M. Ubiali (NNPDF Collaboration), Nucl. Phys. **B849**, 112 (2011); **B855**, 927 (E) (2012).
[53] R. D. Ball, V. Bertone, F. Cerutti, L. Del Debbio, S. Forte, A. Guffanti, N. P. Hartland, J. I. Latorre, J. Rojo, and M. Ubiali, Nucl. Phys. **B855**, 608 (2012).
[54] N. Sato, J. F. Owens, and H. Prosper, Phys. Rev. D **89**, 114020 (2014).
[55] H. Paukkunen and P. Zurita, J. High Energy Phys. 12 (2014) 100.
[56] J. Pumplin, D. Stump, R. Brock, D. Casey, J. Huston, J. Kalk, H. L. Lai, and W. K. Tung, Phys. Rev. D **65**, 014013 (2001).
[57] P. M. Nadolsky and Z. Sullivan, eConf **C010630**, P510 (2001).
[58] J. Pumplin, D. R. Stump, J. Huston, H. L. Lai, P. M. Nadolsky, and W. K. Tung, J. High Energy Phys. 07 (2002) 012.
[59] P. M. Nadolsky, H.-L. Lai, Q.-H. Cao, J. Huston, J. Pumplin, D. Stump, W.-K. Tung, and C.-P. Yuan, Phys. Rev. D **78**, 013004 (2008).
[60] R. D. Ball, L. Del Debbio, S. Forte, A. Guffanti, J. I. Latorre, A. Piccione, J. Rojo, and M. Ubiali (NNPDF Collaboration), Nucl. Phys. **B809**, 1 (2009); **B816**, 293(E) (2009).
[61] S. Carrazza, S. Forte, Z. Kassabov, and J. Rojo, Eur. Phys. J. C **76**, 205 (2016).
[62] C. Anastasiou, C. Duhr, F. Dulat, E. Furlan, T. Gehrmann, F. Herzog, A. Lazopoulos, and B. Mistlberger, J. High Energy Phys. 05 (2016) 058.
[63] M. Czakon, N. P. Hartland, A. Mitov, E. R. Nocera, and J. Rojo, J. High Energy Phys. 04 (2017) 044.
[64] R. Boughezal, A. Guffanti, F. Petriello, and M. Ubiali, J. High Energy Phys. 07 (2017) 130.
[65] A. Accardi *et al.*, Eur. Phys. J. A **52**, 268 (2016).
[66] D. Boer *et al.*, arXiv:1108.1713.
[67] S. Abeyratne *et al.*, arXiv:1209.0757.
[68] E. C. Aschenauer *et al.*, arXiv:1409.1633.
[69] J. L. Abelleira Fernandez *et al.* (LHeC Study Group), J. Phys. G **39**, 075001 (2012).
[70] H.-W. Lin *et al.*, Prog. Part. Nucl. Phys. **100**, 107 (2018).
[71] M. Gockeler, R. Horsley, E.-M. Ilgenfritz, H. Perlt, P. E. L. Rakow, G. Schierholz, and A. Schiller, Phys. Rev. D **53**, 2317 (1996).
[72] X. Ji, Phys. Rev. Lett. **110**, 262002 (2013).
[73] C. Schmidt, J. Pumplin, and C. P. Yuan, arXiv:1806.07950.
[74] G. R. Farrar and D. R. Jackson, Phys. Rev. Lett. **35**, 1416 (1975).
[75] T. J. Hobbs, M. Alberg, and G. A. Miller, Phys. Rev. C **91**, 035205 (2015).
[76] T. J. Hobbs, J. T. Londergan, and W. Melnitchouk, Phys. Rev. D **89**, 074008 (2014).
[77] W. T. Giele and S. Keller, Phys. Rev. D **58**, 094023 (1998).
[78] W. T. Giele, S. A. Keller, and D. A. Kosower, arXiv:hep-ph/0104052.

[79] D. Stump, J. Pumplin, R. Brock, D. Casey, J. Huston, J. Kalk, H. L. Lai, and W. K. Tung, Phys. Rev. D **65**, 014012 (2001).

[80] J. Gao, M. Guzzi, J. Huston, H.-L. Lai, Z. Li, P. Nadolsky, J. Pumplin, D. Stump, and C.-P. Yuan, Phys. Rev. D **89**, 033009 (2014).

[81] R. D. Ball *et al.*, J. High Energy Phys. 04 (2013) 125.

[82] http://projector.tensorflow.org.

[83] PDFSense, http://metapdf.hepforge.org/PDFSense/.

[84] D. Cook, U. Laa, and G. Valencia, Eur. Phys. J. C **78**, 742 (2018).

[85] L. van der Maaten and G. Hinton, J. Mach. Learn. Res. **9**, 2579 (2008).

[86] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevD.98.094030 for additional tabulated sensitivities.

[87] J. Pumplin, Phys. Rev. D **81**, 074010 (2010).

[88] V. Khachatryan *et al.* (CMS Collaboration), Eur. Phys. J. C **75**, 288 (2015).

[89] J. Pumplin, D. R. Stump, and W. K. Tung, Phys. Rev. D **65**, 014011 (2001).

[90] R. Brock, D. Casey, J. Huston, J. Kalk, J. Pumplin, D. Stump, and W. K. Tung, in *Workshop on B Physics at the Tevatron: Run II and Beyond, Batavia, Illinois, 1999* (2000), p. 159, http://inspirehep.net/record/528731.

[91] A. M. Sirunyan *et al.* (CMS Collaboration), Eur. Phys. J. C **77**, 459 (2017).

[92] T.-J. Hou *et al.* (to be published).

[93] T.-J. Hou, S. Dulat, J. Gao, M. Guzzi, J. Huston, P. Nadolsky, C. Schmidt, J. Winter, K. Xie, and C. P. Yuan, J. High Energy Phys. 02 (2018) 059.