Parton distributions need representative sampling

Aurore Courtoy[,1,*] Joey Huston[,2,†] Pavel Nadolsky[,3,‡] Keping Xie[,4,∥] Mengshi Yan[,5,§] and C.-P. Yuan[,2,¶]

[1]*Instituto de Física, Universidad Nacional Autónoma de México, Apartado Postal 20-364,*
01000 Ciudad de México, Mexico
[2]*Department of Physics and Astronomy, Michigan State University, East Lansing, Michigan 48824, USA*
[3]*Department of Physics, Southern Methodist University, Dallas, Texas 75275-0181, USA*
[4]*Pittsburgh Particle Physics, Astrophysics, and Cosmology Center, Department of Physics and Astronomy,*
University of Pittsburgh, Pittsburgh, Pennsylvania 15260, USA
[5]*School of Physics and State Key Laboratory of Nuclear Physics and Technology,*
Peking University, Beijing 100871, China

In global QCD fits of parton distribution functions (PDFs), a large part of the estimated uncertainty on the PDFs originates from the choices of parametric functional forms and fitting methodology. We argue that these types of uncertainties can be underestimated with common PDF ensembles in high-stake measurements at the Large Hadron Collider and Tevatron. A fruitful approach to quantify these uncertainties is to view them as arising from sampling of allowed PDF solutions in a multidimensional parametric space. This approach applies powerful insights gained in recent statistical studies of large-scale population surveys and quasi-Monte Carlo integration methods. In particular, PDF fits may be affected by the big data paradox, which stipulates that more experimental data do not automatically raise the accuracy of PDFs—close attention to the data quality and sampling of possible PDF solutions is as essential. To test if the sampling of the PDF uncertainty of an experimental observable is truly representative of all acceptable solutions, we introduce a technique ("a hopscotch scan") based on a combination of parameter scans and stochastic sampling. With this technique, we examine the PDF uncertainty on key LHC cross sections at 13 TeV obtained with the public NNPDF4.0 fitting code, while accounting for the likelihood distribution. We show that the uncertainties on the charm distribution at a large momentum fraction x and gluon PDF at small x are enlarged. In PDF ensembles obtained in the analytic minimization (Hessian) formalism, the tolerance on the PDF uncertainty must be based on sufficiently complete sampling of PDF functional forms and choices of the experiments.

I. INTRODUCTION

Precision phenomenology at hadron colliders relies upon accurate predictions in the Standard Model (SM). An overwhelming number of such theoretical predictions require parton distribution functions (PDFs) in a proton, the nonperturbative functions $f_a(x, Q)$ quantifying probabilities for finding quarks and gluons in a proton at an energy scale Q above 1 GeV. Multiple groups [1–9]

provide increasingly sophisticated parametrizations of PDFs by fitting a growing collection of precise experimental datasets to advanced multiloop calculations. High-luminosity (HL) measurements at the Large Hadron Collider (LHC) and planned DIS experiments (Electron-Ion Collider [10], Large Hadron Electron Collider [11], Muon-Ion Collider [12] …), combined with the progress in perturbative QCD calculations, open opportunities both to learn about the PDFs and to find their new applications. The global QCD analysis to determine the PDFs can be now attempted by a broad circle of users thanks to the publicly available xFITTER [13] and NNPDF [8] fitting codes. A recent whitepaper [14] contributed to the Snowmass'2021 Summer Study reviews ongoing progress in the PDF analysis.

In this article, we summarize a study of a rarely discussed source of some observed differences between the published parton distributions. A lot of attention has been dedicated to various factors that determine the accuracy of PDFs, usually associated with a combination of experimental, theoretical, PDF parametrization, and

---

[*]aurore@fisica.unam.mx
[†]huston@msu.edu
[‡]nadolsky@smu.edu
[∥]xiekeping@pitt.edu
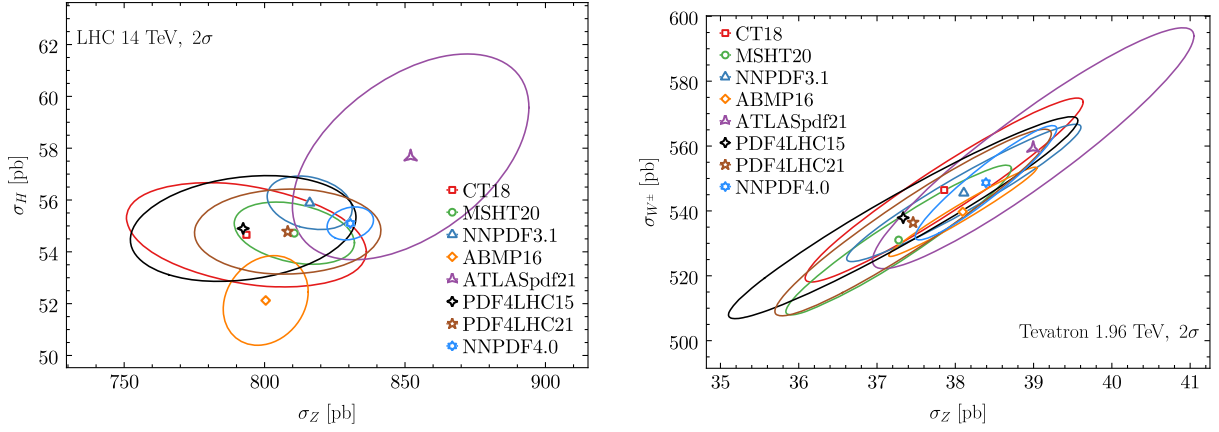[§]msyan@pku.edu.cn
[¶]yuanch@msu.edu

FIG. 1. NNLO theoretical predictions for 95% CL PDF uncertainties for total cross sections of Z and SM Higgs boson production at the LHC 14 TeV (left) and Z and $W^{\pm}$ boson production at the Tevatron 1.96 TeV (right). The Higgs cross sections are obtained at NNLO multiplied by an N3LO/NNLO K factor of 1.097 and by an EW K factor of 1.0514. Predictions are shown for PDF4LHC21 [15], PDF4LHC15 [29], NNPDF4.0 [8], CT18 [5], MSHT20 [7], ABMP16 [3], and ATLASpdf21 [9] NNLO PDFs with $\alpha_s(M_Z) = 0.118$.

methodological sources. Estimates of PDF uncertainties are needed for inference from QCD experiments at the LHC and other facilities [15–17]. However, in addition to the accuracy of individual PDF fits, or "fitting accuracy," another factor in the total uncertainty may be as consequential, reflecting the accuracy of exploration, or *sampling*, of the space of acceptable PDF solutions. This space is truly vast when large data samples are fitted using many parameters. In fact, its exploration can be notoriously difficult, as the sampling of multidimensional spaces is exponentially inefficient [18,19].

In this context, sampled solutions can be obtained by varying the fitted data (e.g., by resampling random fluctuations in the measured values as in the NNPDF analyses [8]), models of theory and experiment (e.g., the PDF parametrization forms [5] or model parameters [20,21] as in the CT and MSHT studies), and in other ways. The pivotal role of adequate sampling in large-scale data analyses has been emphasized in statistics applications across diverse fields, notably in connection with large-scale population surveys [22,23], multidimensional quasi-Monte Carlo (QMC) integration [24], medical research [25], variance-bias separation in machine learning [19,26], and studies of predictivity of complex models [27]. Sampling issues are also pertinent to the multivariate PDF fits. One key observation from the above studies is that large samples do not guarantee convergence to the correct solution, contrary to the common expectation based on the law of large numbers. The reason is that nominally small biases in sampling of possible solutions, such as in the selection of best-fit models of PDFs obtained using various choices of experiments or functional forms of PDFs, may *grow* as the volume of fitted data and complexity of the analysis increase. The growth reflects common difficulties with sampling of parameter spaces of high dimensionality. In an (unfortunately nonrare) situation

when the sampling is unrepresentative of the population of allowed solutions, one may end up with a wrong conclusion described by Xiao-Li Meng [22] as the *big-data paradox*, namely, "the bigger the data, the surer we fool ourselves." Sampling accuracy must be controlled in high-stake phenomenological measurements, such as the recent measurement of W boson mass by the CDF collaboration [28], together with other uncertainties.

Incomplete sampling of PDF solutions may result in unstated sources of the differences among central PDFs or the PDF uncertainties. The possible existence of such differences is suggested by an observation that, while several recent global analyses constrain the PDFs with comparably strong sets of fitted experimental data, in some phenomenologically important cases these analyses arrive at noticeably different estimates of PDF uncertainties. In the CTEQ-TEA analyses, some of the sampling uncertainty is included, as discussed below. On the other hand, when the "experimental" 1σ PDF uncertainty is defined according to the $\Delta\chi^2 = 1$ criterion, as in ABM and HERAPDF studies, this uncertainty does not account for sampling over a sufficient class of PDF parametrizations and other factors, which must be done separately.

As another example, the estimates for the correlated uncertainties on key LHC and Tevatron total cross sections presented in [14,15] vary between the recent PDF fits. In Fig. 1, the 95% confidence level (CL) uncertainty regions on the Z, Higgs, and $W^{\pm}$ total production cross sections at the LHC 14 TeV and Tevatron 1.96 TeV vary in size in a large range. It has been demonstrated that the uncertainties may reflect as much the fitting methodology as the strength of experimental constraints. Indeed, while the differences with the nonglobal (ABMP'16 and ATLAS) and combined (PDF4LHC21) ensembles are reasonably understood, the differences between three global fits—CT18, MSHT'20, and NNPDF3.1—require additional attention. When CT18

[5] and NNPDF3.1.1 [4] NNLO PDFs were compared in Sec. II of the 2021 benchmarking study by the PDF4LHC group [15], the former systematically predicted a larger uncertainty in the moderate x region than the latter. The magnitudes of the MSHT20 NNLO uncertainties [7] in these comparisons tended to lie between the CT18 and NNPDF3.1 ones. More intriguingly, in the course of the PDF4LHC21 exercise, the three global PDF groups conducted fits to a set of common data, using common settings, so as to establish comparisons/benchmarks. The common dataset (termed the "reduced set") was diverse enough to provide constraints on all PDF flavors, but limited enough so that all groups were expected to find similar estimates of PDF uncertainties. In the fits to the same reduced dataset [[15], Sec. III, especially Figs. 3.4 and 3.5], the NNPDF3.1 (reduced) uncertainties came out to be systematically smaller than the CT18 (reduced) and MSHT20 (reduced) uncertainties.

The discrepancies in estimated uncertainties have a variety of implications. For example, they can explain different conclusions about the strength of evidence for the nonperturbative (intrinsic) charm component of the proton obtained by the NNPDF [30] and CTEQ-TEA [31] groups. They also affect projections for sensitivity of planned new experiments, just like related mathematical issues affect planning and policies in other fields [27]. Such differences are often attributed to the tolerance conventions chosen by the global analysis groups. ["Tolerance" refers to the prescription for estimating the PDF uncertainty, see the discussion in Ref. [16].] Are the tolerance conventions mostly subjective, or can some conventions perform better than the others? The question is sharpened by formulating it as a problem about *sampling* of a *specific* PDF-dependent observable that PDFs themselves.

Our article presents an introduction to the sampling of PDF solutions, followed by a presentation of a technique to improve sampling of PDF uncertainty for user-selected QCD observables. Section II reviews mathematical essentials for this discussion. Among these, we first introduce the trio identity, useful for quantifying the convergence of sampling estimates. The trio identity for the sample deviation (Sec. II C) and cornerstone properties of multidimensional (quasi) Monte Carlo integration (Sec. II D) demonstrate that complex, large-scale analyses are at an elevated risk of an unaccounted sampling bias. Global QCD analyses must strive for representative sampling of all acceptable solutions, which may increase the resulting PDF uncertainties or effective tolerance.

Section II also points out fundamental difficulties in performing an all-inclusive test for representative sampling in a multiparametric global fit. Such sweeping test is likely impractical. On the other hand, a practical question "What is the sampling uncertainty on a given observable X?" can be highly tractable using the already available technology for PDF fits. We point out the general

rationale in Sec. II D. We then continue to Sec. III, where we show that the question about PDF uncertainties on specific QCD observables can be explored using the general framework for large-scale surveys and QMC integration presented in Sec. II. Section III A reviews major types of sampling arising in PDF fits, from experimental data to models for systematics. As a specific application, Sec. III B investigates the PDF uncertainty on the LHC benchmark cross sections using the LHAPDF grids of NNPDF4.0 error sets and publicly available MCGEN [21,32,33] and NNPDF [34,35] fitting codes to compute the χ^2 of the included datasets. The hopscotch sampling technique introduced there suggests that the PDF uncertainty on key LHC cross sections at 13 TeV is larger than the nominal uncertainty obtained with the published NNPDF4.0 error sets. Section III C explains the algorithm of the hopscotch scans. Section III D offers a possible interpretation of our findings. The PDF uncertainties must be also enlarged in the case of the strangeness-antistrangeness asymmetry and fitted charm PDF at large momentum fractions, as demonstrated in Sec. III E. Section IV contains conclusions.

II. QCD SAMPLING PROBLEM AND THE TRIO IDENTITY

A. Setup of the problem; the R mechanism

We start by discussing multidimensional sampling in a simplified context, by considering the probability for a QCD observable G dependent on the PDFs, such a collider cross section σ or perhaps the QCD coupling strength α_s determined from hadron scattering measurements. Predictions for observables are the ultimate targets for the propagation of the PDF uncertainties. The goal of the physics endeavor is to estimate the truth value G_{truth} of G that is objectively realized in Nature. Historically, at most we can hope to determine the expectation value $E_p(G) = \frac{1}{N_p}\sum_{i=1}^{N_p} G_i$ on the population of many measurements or other determinations G_i of G, where N_p is a very large number of determinations.

We assume that the determinations G_i are properly designed, so that $E_p(G)$ agrees with G_{truth} (i.e., $E_p(G) - G_{\text{truth}}$ is arbitrarily small) for N_p that is sufficiently large. For example, for $G = \alpha_s$, the population expectation $E_p(\alpha_s)$ could be computed on a future sample of many measurements obtained after several more decades of well-funded research. If G is a cross section σ computed with a multiparameter PDF ensemble, $E_p(\sigma)$ can be the expectation value with the PDF ensemble that densely and representatively samples the whole parameter space.

The conundrum for many studies is that achieving such large N_p may not be feasible. Often one selects a sample of

N_s replicas from the population, with $N_s < N_p$ or even $N_s \ll N_p$, and estimates the sample expectation value as

$$E_s(G) = \frac{1}{N_s} \sum_{i=1}^{N_s} G_i = E_p(RG)/E_p(R). \qquad (1)$$

In the last step, we expressed the sample expectation $E_s(G)$ as a ratio of population expectations $E_p(RG)$ and $E_p(R)$, where R_i is an array of N_p "sampling indicators" such that for each element G_i of the population

$$R_i = \begin{cases} 1, & \text{if the } i\text{th element is in the sample,} \\ 0, & \text{if it is not in the sample,} \end{cases} \qquad \text{for } i = 1, \dots, N_p. \qquad (2)$$

The sample expectation deviation $\Delta_E \equiv E_s(G) - E_p(G) \approx E_s(G) - G_{\text{truth}}$ is controlled by the accuracy of each determination, or a replica in the case of PDFs, G_i, as well as by the accuracy of *sampling* of N_s replicas from the population. The fitting accuracy/sampling accuracy distinction and the representation using the R indicators ("the R mechanism") are borrowed from the study [22] of large-scale surveys, in which "fits" or "replicas" are equivalent to "responses to the survey." Namely, the accuracy of a single replica G_i can be raised by reducing experimental, theoretical, and computational errors. From here, we will assume that the individual G_i are sufficiently accurate. In contrast, the sampling accuracy reflects how adequately we sample the population of N_p acceptable replicas. If such sampling is biased, the magnitude of the sample deviation can be estimated using the R mechanism, see Eq. (3).

Small biases due to insufficiently representative sampling of large populations may produce large deviations. Surveys of the COVID-19 vaccination rate with very large samples of responses and small statistical uncertainties (e.g., Delphi-Facebook) greatly overestimated the actual vaccination rate published by the Center for Disease Control (CDC) after some time delay [23]. The deviation has been traced to the sampling process. In contrast to the random error, which decreases as $1/\sqrt{N_s}$, the sample expectation deviation can grow with both N_p and N_s.

Concurrently with the formalism for the large population surveys, a related statistical formalism has been developed to understand convergence of quasi-Monte Carlo (QMC) methods for multidimensional integration [24]. Insights from these formalisms help us to elucidate our problem in the context of the PDF analysis, in which it can be posed as follows:

Problem 1. Estimate an expectation value $E_p(G)$ of an observable G on a [possibly unknown] population of N_p replicas, given a sample of N_s values G_i, where $N_s < N_p$.

To get such estimate, it suffices to adopt an R mechanism that renders $\Delta_E = E_s(G) - E_p(G) = 0$ within a prescribed error. In this section, we discuss convergence of such sampling estimates.

B. A toy example

As a toy example, consider a population of NNLO Higgs boson cross sections $G \equiv \sigma_{gg \to H}$ at the LHC c.m. energy 14 TeV. The cross sections are computed with $N_p = 900$ error sets of the baseline PDF4LHC21 PDF ensemble [15] consisting of 300 MSHT20, 300 NNPDF3.1.1, and 300 CT18′ replicas, illustrated on the left panel of Fig. 2. [The replicas are ordered as in the actual 900-replica baseline ensemble. The mean cross section of the CT18′ (NNPDF3.1.1) subset is slightly lower (higher) than the population mean.] We have $E_p(G) = 47.492$ pb and wish to obtain a close estimate by sampling only $N_s = 300$ replicas out of 900.

If we randomly select $N_s = 300$ replicas from the whole population, we obtain $\Delta_E = 0 \pm 0.033$ pb, where the 68% CL uncertainty is computed by repeating the random selection 1000 times. In this case, E_s and E_p are statistically indistinguishable (see the middle panel of Fig. 2). It is known on general grounds that, with the random sampling from the whole distribution, Δ_E decreases as $1/\sqrt{N_s}$, independently of N_p [22].

As an instance of a different sampling, let us select 100 replicas from each of the MSHT, NNPDF, and CT18 subsamples, for a total of $N_s = 300$ replicas. In this case, we still get $\Delta_E = 0 \pm 0.03$ pb, i.e., no deviation. Since the PDF4LHC21 baseline set of 900 replicas is constructed by randomly selecting 300 replicas from each of the MSHT20, NNPDF3.1.1, and CT18′ 1000-replica samples, we conclude that this PDF4LHC21 nonrandom combination prescription introduces no appreciable deviation.

We can generate various sample combinations selecting 100 replicas from one of the three groups, 200 replicas of second group, and none of the third. Two of those are shown on the right panel of Fig. 2, with deviations of $\Delta_E = -0.206 \pm 0.036$ pb and $\Delta_E = 0.138 \pm 0.031$ pb. In this instance, the bias was introduced by hand, but in realistic situations the bias can arise from apparently small departures from the random (probabilistic) sampling at various stages of the analysis.

C. The trio identity

The trio identity [22,24] for the sample expectation deviation Δ_E is a representation introduced to examine convergence of the sampling algorithm. For our problem, the trio identity takes the form
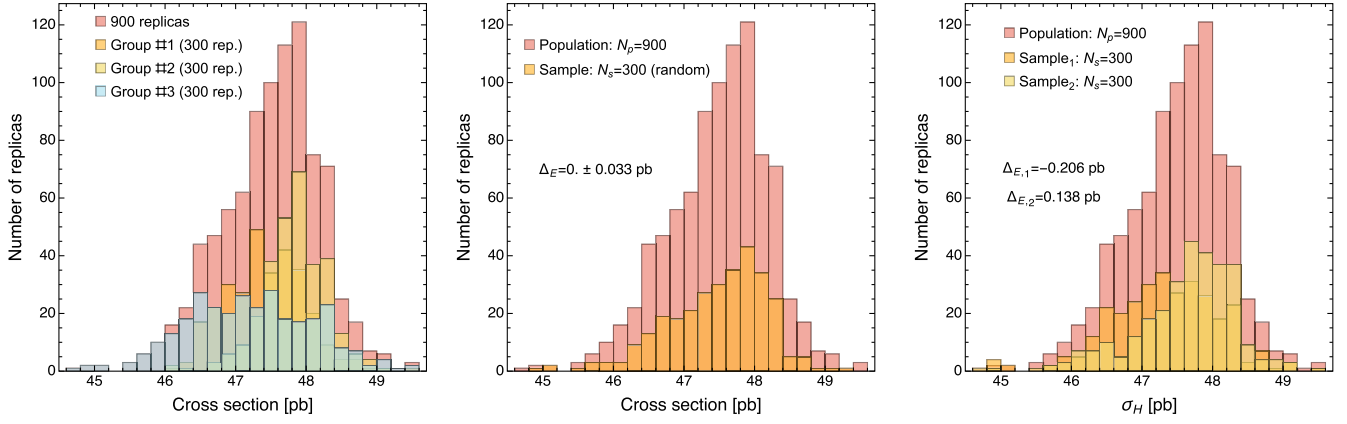
FIG. 2.   The R mechanism illustrated on a toy example. The left panel shows a histogram of predictions for NNLO Higgs cross sections at the LHC based on 900 error PDFs of the PDF4LHC21 PDF ensemble ("the population"). The PDF4LHC21 ensemble is composed of three groups with 300 predictions; the resulting histograms are superimposed. The middle panel shows 300 predictions that are randomly sampled from the 900-member population, producing a sample average $\hat{\mu}$ that is indistinguishable from the population average μ (i.e., $\Delta_E \approx 0$). The right panel depicts two biased samples. Each contains 100 predictions from one group, 200 from a second group, and none from the third. The biased samplings result in nonzero Δ_E values.

$$\Delta_E = E_s(G) - E_p(G) = \frac{\text{Cov}_p[R,G]}{E_p(R)}$$

$$= \underbrace{\text{Corr}_p[R,G]}_{\text{confounding correlation}} \cdot \underbrace{\sqrt{\frac{N_p}{N_s} - 1}}_{\text{measure discrepancy}} \cdot \underbrace{\text{Var}_p[G]}_{\text{population variation}}.$$

(3)

The three factors on the right-hand side are population expectations with different dependencies on N_s and N_p. In Eq. (3), $\text{Cov}_p[A,B] \equiv E_p((A - E_p(A))(B - E_p(B)))$ is the population covariance. *Variation* $\text{Var}_p[G] \equiv \sqrt{\text{Cov}_p[G,G]}$ reflects the complexity of the population distribution.[1] *Measure discrepancy* $\sqrt{N_p/N_s - 1}$ is due to the mismatch between the sizes of the population (N_p) and the sample (N_s). The *confounding correlation* $\text{Corr}_p[R,G]$ lies between -1 and 1. It quantifies efficiency of the sampling algorithm in comparison to simple random sampling. The confounding correlation reflects methodology of the analysis. Methodological correlations play a central role in precise PDF analyses [36], together with data-driven [37] and theory-driven [38–41] correlations.

If the sampling exercise is repeated N_R times while keeping the same N_s, each time choosing a different R array, one can estimate a mean-square error (MSE) of the sample deviation for a given R mechanism:

$$\text{MSE}_R(\Delta_E) \equiv E_R(\Delta_E^2) = E_R(\text{Corr}_p[R,G]^2)$$

$$\cdot \left(\frac{N_p}{N_s} - 1\right) \cdot \text{Var}_p(G)^2.$$

(4)

---

[1]Variation $\text{Var}_p[G]$ and standard deviation $\bar{\sigma}_G$ are related as $\bar{\sigma}_G^2 = N_p/(N_p - 1)\text{Var}_p[G]^2$.

The trio identity establishes dependence of the sample deviation Δ_E on the sampling algorithm [22].

(1) Under simple random sampling (SRS), when replicas are independently selected with identical probability, the sample deviation converges to the truth as $1/\sqrt{N_s}$ in compliance with the law of large numbers:

$$\text{SRS:}\ \Delta_E \to 0 \quad \text{as } N_s \to N_p,$$

with

$$\text{MSE}_{\text{SRS}}(\Delta_E) \equiv V_{\text{SRS}} = \kappa\bar{\sigma}_G^2/N_s,$$
$$\text{where } \kappa \equiv (N_p - N_s)/N_p \sim 1. \quad (5)$$

Comparison of Eqs. (4) and (5) shows that $E_{\text{SRS}}(\text{Corr}_p[R,G]^2) = 1/(N_p - 1)$.

(2) For an arbitrary sampling algorithm, the sample deviation satisfies

$$\Delta_E = \text{Corr}_p[R,G]\sqrt{N_p - 1}\sqrt{V_{\text{SRS}}},$$
$$\text{MSE}_R(\Delta_E) = E_R(\text{Corr}_p[R,G]^2)(N_p - 1)V_{\text{SRS}}. \quad (6)$$

For the sampling deviation to vanish as N_s increases, $\text{Corr}_p[R,G]$ should decrease at least as fast as $o(1/\sqrt{N_p - 1})$. Absent this behavior, unrepresentative sampling may lead to a situation when the sample deviation remains large in spite of misleadingly small standard error estimates. Meng dubbed this situation as "the big-data paradox," which is clearly undesirable and unfortunately can go unnoticed if sampling accuracy is not controlled to a sufficient degree.

D. Quasi-Monte Carlo integration

The trio identity elucidates why quasi-Monte-Carlo (QMC) methods for multidimensional integration may

converge at a faster or slower rate compared to the Monte Carlo integration based on SRS [24]. When integration is performed over a unit hypercube in N_{par} dimensions, the sample deviation Δ_E coincides with the *(hyper)cubature error* and can be decomposed into three factors that play the same roles as in Eq. (3).

Of particular interest to us is the convergence of QMC integration when N_{par} is large. In this limit, the minimal number of MC replicas that guarantees a convergent integral for an arbitrary integrand grows as $2^{N_{\text{par}}}$ [42], reflecting the curse of dimensionality that was pointed out long ago [18,19]. Not only dense sampling of a high-dimensional volume requires an exponentially growing number N_p of replicas, such as $2^{N_{\text{par}}} \sim 10^{30}$ for $N_{\text{par}} = 100$; suppression of the confounding correlation to the adequate degree is likely as a daunting feat.

The sample expectation of a QCD observable $G(\vec{a})$ in PDF fits is merely an integral of the weighted probability function $P(\vec{a})$ over N_{par} PDF parameters $\vec{a}$:

$$E_s(G) = \int G(\vec{a})P(\vec{a})d\vec{a}. \qquad (7)$$

We immediately conclude that convergence of $E_s(G)$ to the truth for an arbitrary $G(\vec{a})$ is not at all guaranteed in a PDF fit that depends on too many parameters and does not control for representative sampling.

In such a complex fit, one practically cannot know if the sample PDF uncertainty covers the truth values for all $G(\vec{a})$. On the positive side, it follows from Eq. (7) that, if $G(\vec{a})$ is known to substantially depend only on a few components of $\vec{a}$, estimation of $E_s(G)$ becomes highly tractable. The reason is that the convergence rate of QMC integration is controlled by *the effective number of components*, i.e. directions in the parameter space, along which the variance of the integrand is significant [43]. If the number of such components is small, integration can be arranged so as to give more weight to the sampling of the manifold spanning the corresponding "large dimensions." For example, the coordinates in the subspace with highest variances of $G(\vec{a})$ can be sampled most densely. The coordinates in the complementary subspace with low variances can be either fixed or sampled with a low density. Techniques exist for ranking the N_{par} coordinates according to the variance of the integrand using the analysis of variance (ANOVA) [44], principal component analysis (PCA), or another dimensionality reduction method. Accuracy of integration can be iteratively improved by adding contributions from the coordinates with lower variances [45]. See discussion in Sec. 8 of Ref. [24].

The role of effective dimensions in accounting for large uncertainties from complex models, beyond Monte-Carlo integration, was recently highlighted for the broader context of applied science. Reference [27] stresses the role of uncertainties in the decision making of new policies in the real world, "where reliance on excessively complex and overconfident models may have deleterious social-environmental consequences."

Experience with high-dimensional integration thus raises a warning for the analyses that fit a large number of flexible functions using a modest number of fitted replicas. While these analyses excel at finding acceptable sets of functions describing the data, they are nevertheless prone to the risk of a sampling bias that grows with the dimensionality of the problem. Apparent reduction of the variance does not eliminate this risk because of the big-data paradox quantified by the trio identity. It has been known for a while that precise sampling of χ^2 in the vicinity of the global minimum becomes inefficient with traditional MC replicas: the majority of such replicas have too large $\Delta\chi^2$ because of high dimensionality of the parameter space [[21], Sec. 3. B]. All-inclusive testing for representative sampling thus is difficult with a lot of free parameters. Fortunately, typical QCD cross sections depend on specific combinations of PDFs that can be established using dataset diagonalization [46] (for example, implemented as optimization of Hessian sets for specific experiments in the ePump package [47]) or a related method. Sampling of a known PDF combination can be tested with a greatly reduced cost based on the dimensionality-of-integration argument presented above. Hopscotch scans described in the next section realize such test in practice.

III. SAMPLING TESTS AND HOPSCOTCH SCANS

A. PDF uncertainties on QCD observables as a sampling problem

Section II summarized recent mathematical approaches to statistical surveys of large datasets and QMC integration of functions dependent on many parameters. In this section, we advance a viewpoint that the same approaches can guide estimation of PDF dependence of specific QCD cross sections. In this case, we consider a population of predictions $\{G_i\}$ for an observable G based on a large collection of PDF sets that will be obtained in the future. Without the loss of generality, we assume that the PDF sets are indexed by independent countable parameters and are acceptable according to the goodness-of-fit criteria explained below.

A prediction based on one such PDF set plays the role of an individual response to the survey, given by the numerical value G_i. Predictions based on one published PDF ensemble can then be viewed as a sample with the size N_s that is smaller than the population size N_p. Again without the loss of generality, we can assume that the expectation values can be computed using the unweighted average as in Eq. (1), or, if so necessary, using the weighted average as in Ref. [22]. The formalism from survey studies [22,23] then tells us that, given the complexity of PDF models, confounding correlations may dominate the sampling bias Δ_E even when the sample SRS deviation, proportional to $1/\sqrt{N_s}$, is small.

Validation of representative sampling is thus as essential as the tests of quality of individual fits, such as strong goodness-of-fit tests on resulting PDFs [16] and the closure test [48] of the agreement of a trial fit with a predetermined truth value within the uncertainty. However, for an all-out sampling test, the computation of the confounding correlation in the trio identity, Eq. (3), requires to know the population distribution as an input, which is not known while the fits are performed. The confounding correlation can be predicted to a degree by using a model population distribution based on simulated pseudodata in the same spirit as done in the closure test. On the other hand, tests of representative sampling are simpler for QCD observables with low effective dimensionality.

But what exactly is sampled in the PDF fits? Several types of PDF sampling are performed based either on a known or unknown probability distribution. The uncertainties from each sampling type may or may not be included as a part of the final uncertainty. To illustrate how the groups handle various types of sampling, we will compare two recent NNLO PDF analyses, CT18 [5] based on the analytic χ^2 minimization and NNPDF4.0 based on the MC sampling of neural network parametrizations of PDFs [8]. We outline some common categories of sampling, leaving out technical details of specific realizations.[2]

a. Sampling of experimental datasets occurs when these datasets are selected for the fit. As a variation, only a part of the dataset can be included. Some datasets may be included with χ^2 weights that are different from unity, as has been done in PDF fits circa year 2000. If there are inconsistencies among the datasets, inclusion of a dataset from the global fit may result in a larger-than-nominal shift of the expectation value. The associated variation is latent in the PDF fits with significant tensions among the experiments, including the recent global PDF analyses. The strengths of tensions among the fitted PDF sets are comparable in CT18, MSHT'20, and NNPDF3.1 fits, as reflected by $\chi^2/N_{\rm pt}$ values of experimental datasets in Tables 2.1–2.3 of [15]. Such tensions can be identified with techniques described in [16,49]. Standard techniques for estimation of the corresponding PDF uncertainty, like jackknife cross-validation (computing the expectation value on an ensemble of fits with one experiment left out at a time), are hardly practical in the global fits. Instead, global PDF fits may resort to a remedy of increasing the χ^2 tolerance associated with one standard deviation from $\Delta\chi^2 = 1$ to a larger value. More complex tolerance prescriptions can be alternatively used [16]. In the CT18 family of PDFs, a special CT18Z ensemble is provided to obviate the change in the PDFs upon the inclusion of the ATLAS 7 TeV W/Z production [50] that runs into tension

with dimuon SIDIS experiments. The difference between the CT18 and CT18Z central values can exceed the sum of 90% intervals of two ensembles. The MSHT'20 and NNPDF4.0 analyses publish only the PDF ensembles for the default selection of experiments. Mutual consistency of the datasets is thus a part of the data-quality requirement for the reduction of uncertainties.

b. Sampling of experimental data fluctuations is the most familiar type of sampling. The NNPDF and other (pre-)MC approaches generate PDF sets by resampling and cross-validation of the experimental data. In this paradigm, multiple replicas of the fitted data are constructed by randomly fluctuating the data's central values according to the experimental uncertainties. For each replica of data, the PDFs are found by fitting to the training part of the replica and simultaneously cross-validating against the complementary, control part. The final PDFs optimally agree with both training and control parts based on a criterion that depends on the log-likelihood function χ^2 computed with respect to the fluctuated data. Expectation values are then computed using an unweighted average of predictions on an ensemble of such replicas. The NNPDF group [51] calls this approach "importance sampling." It is called "resampling" or "bootstrap" by other groups.

The CT approach, on the other hand, finds the best-fit PDFs by minimization of the log-likelihood χ^2 computed with respect to unfluctuated data, i.e. the published data with specified statistical and systematic errors (see below). Expectation values in this approach are computed using the best-fit PDF, confidence intervals are estimated using Hessian eigenvector (EV) sets. The CT fit can also produce MC error sets, usually done by the conversion of the final Hessian PDF sets [21]. Reciprocally, the NNPDF4.0 MC replica sets have been also converted into an ensemble of 50 Hessian PDFs, which reproduces the expectations, standard deviations, and correlations of the NNPDF4.0 MC replicas. Resampling of experimental uncertainties and conversions between Hessian and Monte Carlo PDFs are well understood and numerically accurate.

c. Sampling of PDF functional forms and fitting/training methodologies is another common source of an explicit or latent uncertainty. It is independent from the data resampling uncertainty. In the discussion of data fluctuations, we assumed that the MC replicas are generated with the same training methodology, including the same choices for the sizes and contents of training and control partitions, as well as the same condition to finish replica training. We also assumed that the parametrization forms of the PDFs, whether given by an analytic function or by a neural network of a certain architecture, do not change in the course of an individual fit or training cycle. Such settings of the fit of course can also be varied.

In contrast to the experimental uncertainty, these choices are associated with the prior probability. One aspect of this kind is that the final PDFs, whether produced by analytical

---

[2]In addition to the categories discussed here, Monte-Carlo integration uncertainty of theoretical calculations may be important in some cases.

minimization or an AI/ML method, aim to describe the data without underfitting or overfitting parts of data. As a consequence of the fundamental variance-bias dilemma [19,26], overfitting is not sharply defined. Namely, a fit or ML training with an arbitrary functional form can produce multiple solutions that balance between agreeing with the (un)fluctuated data, having the fitted function with high variation, and allowing for random noise. One therefore expects some differences between the overfitting tests adopted by various groups.

Smoothness, such as the absence of sharp features in acceptable PDFs, is a related condition that does not necessarily imply data overfitting. Both CT18 and NNPDF4.0 analyses require the PDFs to satisfy conditions of smoothness, positivity, and integrability, again according to varied prescriptions.

In the CT18 analysis, the candidate fits were repeated with more than 250 alternative functional forms and produced substantial spread of PDF solutions. The tolerance of the published CT18 ensembles, such as those shown in the next subsections, was increased so that their Hessian PDF uncertainty covers the solutions obtained with the alternative parametrizations. CT18 parametrizations utilize Bernstein polynomials, which allow examination of a variety of flexible, yet usually smooth, functional forms.

The NNPDF4.0 analysis adopts a specific optimized algorithm to select the architecture, train neural networks, and impose smoothness and other prior conditions. As a part of the algorithm, the final 100 or 1000 NNPDF4.0 replicas are selected from a larger pool of replicas, many of which exhibit nonsmooth, short-length features. The algorithm is checked for self-consistency in a closure test by fitting idealized pseudodata and verifying quantitative estimators such as the bias-variance ratio and quantiles of $\Delta\chi^2$ for groups of experiments. The closure test demonstrates that the NNPDF4.0 optimized algorithm is *sufficient* for generating well-behaving PDFs that agree with the known "truth" PDFs.[3] Closure tests, however, do not prove that the use of the NNPDF4.0 settings is a *necessary condition* for obtaining well-behaving solutions under acceptable variations in methodology. This especially applies to the case of fitting inconsistent datasets. Other algorithms, which vary in terms of hyperparameters, priors, and similar setting choices, may exist and produce PDF solutions that enlarge the nominal NNPDF ensemble.

The availability of public NNPDF4.0 MC and Hessian PDFs, together with the public NNPDF4.0 code, opens a possibility for a test to evaluate performance of the NNPDF4.0 algorithm in finding the PDF solutions. Consider the Bayesian likelihood-ratio test in the context of PDF comparisons [16,52]. Suppose two PDF solutions,

A and B, have the same likelihood, but solution A is deemed unlikely compared to B based on the ratio of *posterior* Bayesian probabilities. From this, we conclude that solution A is disfavored because of its lower *prior* probability, not because of its likelihood. Generalizing for a collection of QCD observables, we can identify the regions populated by *new* predictions that have the same or higher likelihood as the nominal NNPDF4.0 regions. Differences between these regions arise from the *prior* conditions imposed on the new and nominal solutions. The ML universal approximation theorem [53–55] implies that both groups of solutions can be approximated by neural networks. The proposed test therefore examines sampling over classes of eligible functions or, in the ML language, eligible neural networks. If, in addition to having low χ^2 values, the new solutions pass all other goodness-of-fit criteria, they must be accounted in the final PDF uncertainty e.g. in the form of an enlarged tolerance. The design and implementation of the test are explained in Sec. III C.

d. Sampling of likelihood functions. There is another ambiguity to consider, associated with the approximation of the likelihood in the PDF fits. Since the experiments rarely provide the full likelihood, it is usually expressed as $P(D|T) = \text{const} \cdot \exp\left(-\chi^2(D, T)/2\right)$, where χ^2 is constructed from experimental data values D_i, theoretical predictions T_i, and associated uncertainties. The log-likelihood χ^2 enters the figure-of-merit function in the fit, where it can be combined with prior conditions or computed with respect to fluctuated D_i, as done during the training of MC replicas. The log-likelihood χ^2 is also used, not necessarily in the same form as during the fit, for external comparisons of PDFs like the ones done in our study. Non-Gaussianities of the errors are frequently neglected, and various approximations are made to the correlated systematic errors, which still lack full understanding [Sec. 5 in [14]]. These choices produce nonidentical forms of χ^2 used by ATLAS [9], HERA [1], CT [5], MSHT [7], and NNPDF [8].

In regard to the correlated errors, the PDF analyses address a common ambiguity when converting percentage uncertainties into absolute ones. For an experiment with N_{pt} data points and N_λ systematic errors, the χ^2 functions used by the three global groups can be reduced to

$$\chi^2 = \sum_{i,j}^{N_{\text{pt}}} (T_i - D_i)(\text{cov}^{-1})_{ij}(T_j - D_j), \quad (8)$$

where

$$(\text{cov})_{ij} = s_i^2 \delta_{ij} + \sum_{\alpha=1}^{N_\lambda} \beta_{i,\alpha}\beta_{j,\alpha} \quad (9)$$

depends on uncorrelated uncertainties s_i, and correlated ones $\beta_{i,\alpha}$. In turn,

$$\beta_{i,\alpha} = \sigma_{i,\alpha}X_i \quad (10)$$

---

[3]Success of the closure test depends on the targeted precision and accuracy of the truth-fit comparisons.

are derived from the tables of published $\sigma_{i,\alpha}$ using unspecified normalization cross sections X_i. It has been observed that plausible choices of X_i nontrivially affect the resulting PDFs. Search for the "least biasing" choices prompted scrupulous investigations [56–59]. The Appendix in [56] reviews the rationales for these choices, which depend on the type of the systematic error, while Refs. [1,5,7–9] detail implementations of $\beta_{i,\alpha}$ in the latest PDF fits.

The groups generally avoid fitting the PDFs with the choice of $X_i = D_i$ for multiplicative errors (so called "experimental" scheme, or "exp"), on the count that it was shown to bias the best-fit results with respect to the truth in relatively simple examples examined by D'Agostini [60,61] and NNPDF [58]. Partly for this reason, CTEQ-TEA analyses normalize all $\beta_{i,\alpha}$ by $X_i = T_i$ (the current theory) [59]. The NNPDF group uses a "t_0 scheme," which has been available in two versions: the pre-NNPDF3.0 analyses multiplied only the normalization uncertainties by an iteratively updated theory value, $X_i = T_i^{(0)}$ [56,58], and the NNPDF3.0 and later analyses normalize all $\beta_{i,\alpha}$ with $T_i^{(0)}$ in several groups of experiments [[48], Sec. 2.4.2], while the rest of the errors are normalized by $X_i = D_i$.[4]

These are not the only χ^2 forms in use, however, and in fact the NNPDF4.0 publication quantifies the quality of the fit and agreement with the experiments with χ^2 values in the "exp" scheme [Sec. 5.1 in [8]]. It can be understood that neither of these conventions is safe from biases by recognizing that X_i are values of an initially unknown function X that is fitted or learned together with the PDFs. As such, X is subject to the already mentioned tradeoff between variance, bias, and noise [19,26], with none of the current implementations systematically controlling for this tradeoff. The exp and t_0 schemes correspond to the zero-bias (with respect to D_i) and low-variance options, respectively, and a sequence of other possible schemes lies in-between. [In the exp scheme, the function X goes through the fluctuating data points D_i. Other schemes use a smoother function.] The well-known demonstrations of D'Agostini's bias assumed at most a few multiplicative errors. The PDF fits deal with many multiplicative errors, whose pulls on the PDFs may have opposite signs or be nonlinear, adding up to an unpredictable effect. For high-statistics data samples, it is even possible that random fluctuations in D_i are smaller than uncertainties in choosing $X_i = T_i^{(0)}$; and, finally, the truth X_i for some $\beta_{i,\alpha}$ in the experimental publication may not exactly coincide with T_i or its user-selected analog $T_i^{(0)}$ in the PDF fit at hand.

---

[4]In particular, the gluon at $x > 0.1$ varies depending on the additive or multiplicative treatment of the errors in inclusive jet production. For illustrations, see Fig. 18 in [59] and Figs. 60 and 61 in [48].

The sampling test proposed above can also explore dependence on the form of the likelihood, given that the NNPDF4.0 fitting code can return χ^2 values in the exp and t_0 schemes. As stated in a note of the NNPDF4.0 code manual, "the t_0 method is **not** used* by default *by [applications of the validphys code other than replica training], and instead the default is to compute the experimental χ^2" [62]. Our sampling test is agnostic about the generation of PDFs and just compares available PDF solutions without actually fitting them.

B. NNPDF4.0 probability regions for the LHC benchmark cross sections

The main findings of the test are summarized in Fig. 3, which shows the PDF uncertainties on LHC cross sections at $\sqrt{s} = 13$ TeV computed at NNLO in the QCD coupling strength according to the settings listed in the Appendix. For experimental collaborations, it is important to know which theoretical predictions are acceptable given the latest experimental and theoretical constraints. In the exercise to be presented, all results are obtained using either error PDFs available in the LHAPDF library [63], or linear combinations thereof.

The elliptical regions seen in Fig. 3 are projections of N_{eig}-dimensional volumes populated by PDFs with low χ^2 for the indicated Hessian PDF ensembles, where N_{eig} is the number of EV directions in the Hessian ensemble. For each Hessian PDF ensemble and χ^2 scheme, there is one such approximately ellipsoidal volume. The ellipses of the same style in the six panels of Fig. 3 are the volume's projections onto the two-dimensional planes specified by the pairs of indicated total cross sections. *Here and in the following, the χ^2 is computed with respect to the published (unfluctuated) central data values, except when we explicitly say otherwise.*

Let us now go over the shown probability regions one-by-one.

In the NNPDF4.0 approach, the Monte-Carlo error PDFs are constructed by optimized training of neural networks on replicas of randomly fluctuated experimental data. Each replica fit achieves a good χ^2 with respect to its fluctuated dataset, while practically all MC replicas have a very high χ^2 (by hundreds of units) with respect to the unfluctuated data. The individual MC PDFs are thus poor fits to the published (unfluctuated) dataset—but their average (called the "central replica", or "replica 0") agrees with the unfluctuated data much better [21]. The standard deviation on the ensemble of 1000 NNPDF replicas thus essentially provides a 68% experimental error with fixed methodological settings, estimated around the central replica as the spread of best fits to randomly fluctuated data points for the chosen methodology, including selection (sampling) of experiments, training and cross validation algorithm, and treatment of systematic effects.

To examine the dependence of χ^2 from the NNPDF4.0 code on PDF parameters, as discussed in Sec. III A, we will
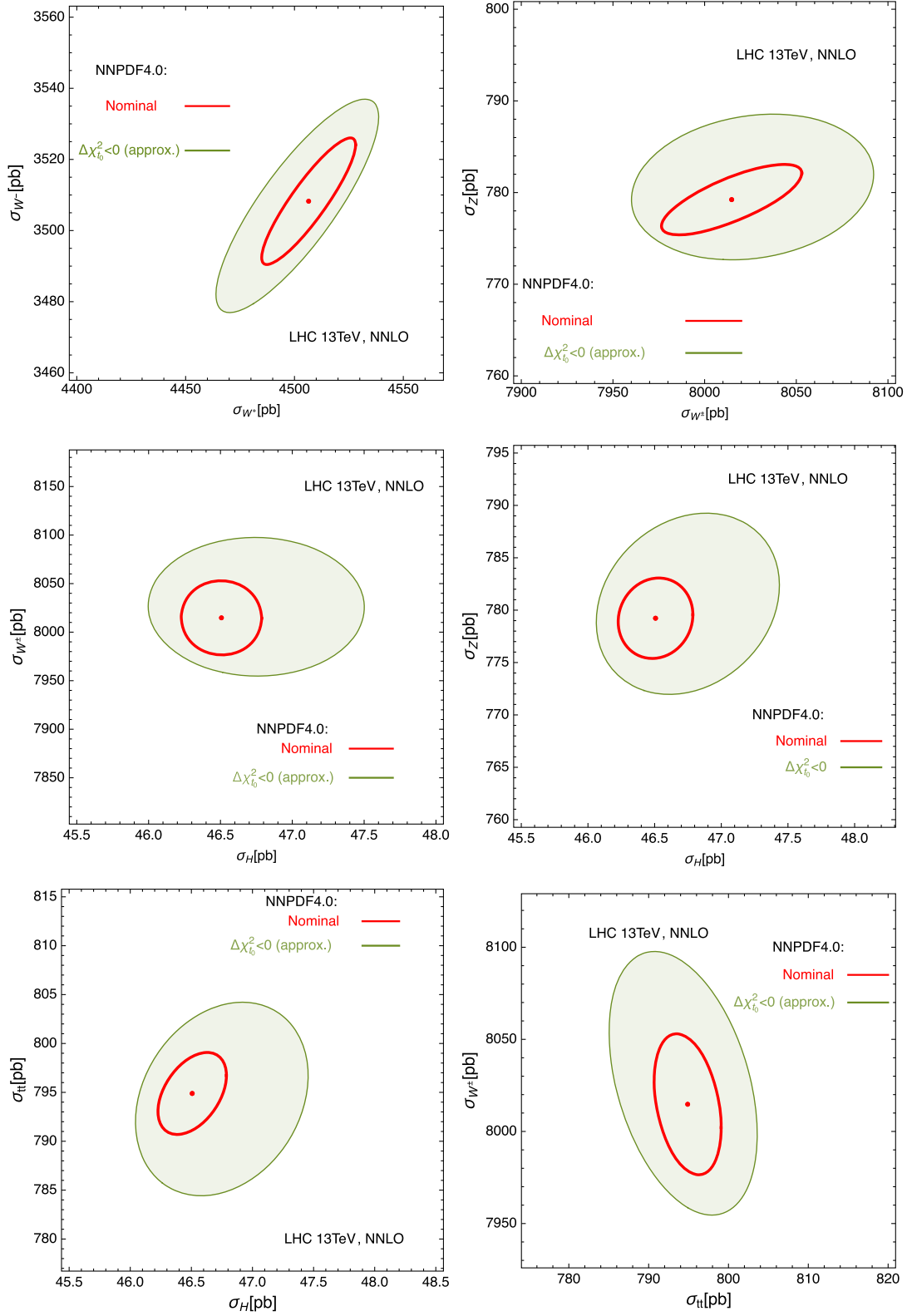
FIG. 3. Red solid ellipses: the nominal 68% probability regions from the NNPDF4.0 NNLO analysis [8]. Green ellipses with a solid contour: approximate regions with acceptable NNPDF4.0 solutions that have the same $\chi^2_{t_0}$ as the central PDF replica in the NNPDF4.0 publication. The reconstruction of the approximate ellipses is described in Sec. III C.

employ the NNPDF4.0 Hessian EV set—50 error PDFs, with 1 error PDF per each of $N_{eig} = 50$ EV directions, obtained by post-fit conversion of NNPDF4.0 MC replicas. In the NNPDF implementation, the Hessian set is centered on the NNPDF4.0 replica 0 and reproduces 1σ symmetric PDF uncertainties of the MC set, i.e., the regions containing 68% of the MC replicas. Section III D illustrates that, indeed, the NNPDF4.0 Hessian ensemble reproduces well these regions.

In the standard interpretation, however, Hessian eigenvectors form a basis in linear space populated by vectors of PDF solutions in the vicinity of the global minimum of χ^2. The position of the global minimum obviously depends on the χ^2 scheme, chosen to be the t_0 scheme in the figures. The approach that provides 1 EV set per EV direction is usually based on the expectation that the global minimum is very close to replica 0, and hence the χ^2 is more or less symmetric with respect to replica 0. We will compare locations of replica 0 and global minima, and then examine the χ^2 dependence in this space.

In Fig. 3, red solid ellipses delineate the 68% probability regions computed with the published NNPDF4.0 error PDFs and centered on the predictions from the NNPDF4.0 central replica.

Overlaid on the nominal uncertainties, the green ellipses indicate *approximate* regions containing PDF solutions that have better χ^2, according to the NNPDF fitting code, than the NNPDF4.0 central replica 0. Section III A pointed out that comparisons in NNPDF publications use two forms of χ^2 as the figure-of-merit, called "t_0" and exp and computed with respect to the unfluctuated data, which differ in their implementation of experimental systematic uncertainties. The green ellipses delineate regions for each pair of cross sections in which our analysis have found regularly behaving PDF solutions that have $\Delta\chi^2 \equiv \chi^2 - \chi_0^2 \lesssim 0$ according to the "t_0" definition, where χ_0^2 is computed for replica 0. Inside the ellipses, χ^2 shows quasi-Gaussian dependence on the PDF parameters with both definitions. The found low-χ^2 PDF solutions are linear superpositions of the NNPDF4.0 Hessian replicas. Among them, we find a few with $\Delta\chi^2$ as low as -37 with the t_0 definition. Their χ^2 values are computed using the published NNPDF4.0 code [34,35].

We construct the alternative NNPDF4.0 solutions using an algorithm that we call a *hopscotch scan*, which performs focused sampling of PDF combinations giving the dominant contribution to the PDF uncertainty of the shown cross sections. Section III C explains this algorithm and construction of the approximate ellipses. The technique combines Lagrange multiplier (LM) scans of PDF parameters [64] along the Hessian EV directions with stochastic sampling of the few "large" dimensions associated with the largest variance of the LHC cross sections under consideration, in accord with the general discussion in Sec. II D.

It is obvious from Fig. 3 that the ellipses containing the PDFs that have $\Delta\chi^2 \leq 0$ are larger than the nominal

68% CL NNPDF4.0 uncertainties. These PDFs have been examined for possible nonsmooth features and other pathologies. We did not observe obvious flaws and found them to be acceptable, according to the CT18 procedure [5], in light that they achieve the same or better χ^2 as the nominal fit, are smooth, and fall within the nominal NNPDF4.0 errors nearly everywhere.

The t_0 definition of the χ^2 reflects a certain implementation of the experimental correlated systematic errors. We pointed out that other definitions exist, reflecting incomplete knowledge of systematic uncertainties provided to the global fits [[14], Sec. 5, and references therein]. Dependence on the χ^2 definition must be scrutinized as a part of a more complete exercise. Even within this limited scope, it is clear that there can be acceptable solutions with $\Delta\chi^2 \leq 0$ outside the nominal NNPDF4.0 uncertainty. Section III D offers a plausible interpretation of such alternative solutions.

C. A hopscotch scan, technical implementation

In this section we describe the construction of the alternative PDF solutions that led to the ellipses of Fig. 3. The procedure realizes the general considerations in Sec. II D, namely:

(1) The NNPDF4.0 Hessian ensemble establishes natural basis coordinates $\vec{a}$ in space of MC replicas.

(2) For a typical QCD observable $G(\vec{a})$, the largest variances are associated with 4-8 "large dimensions" in $\vec{a}$ space.

(3) The PDF uncertainty on $G(\vec{a})$ can be estimated with a moderate number of MC PDF replicas that vary along the large directions.

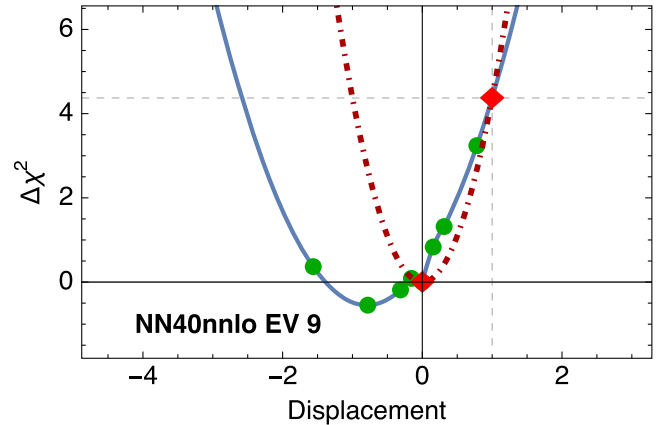We generate LHAPDF6 tables for the sample PDF replicas using the MCGEN program [21,32,33] and the

FIG. 4. Magnification of the χ^2 scan for EV direction 9. The green points and blue curve are the actual $\Delta\chi^2$ values and their interpolation from the scan. The dot-dashed red curve represents a symmetric parabola that would be obtained given only the central replica 0 and the EV set 9 (two red diamonds), and assuming that the third point necessary to build the parabola is the mirror of $(1\sigma, \Delta\chi^2(1\sigma))$.
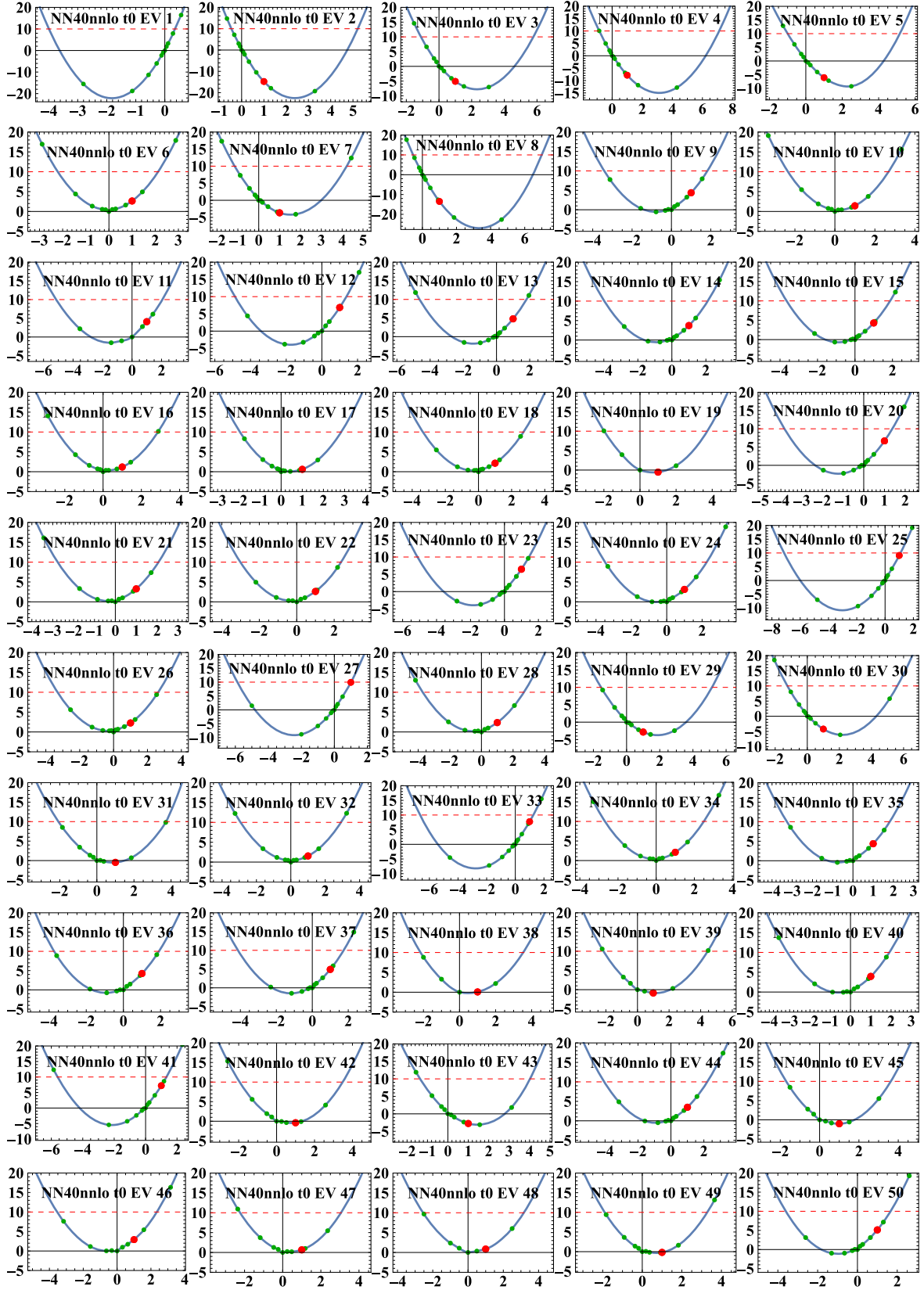
FIG. 5.   Scans of the total χ^2 in the experimental definition along 50 eigenvector directions of the NNPDF4.0 Hessian NNLO PDF ensemble. The red point corresponding to a displacement of $+1$ in the x axis is not shown in the case of EV 1, as it is out of scale. Its value is $\Delta\chi^2 = 32$. See text.

LHAPDF tables of the NNPDF4.0 NNLO Hessian ensemble as the input. The total χ^2 of the NNPDF4.0 analysis was evaluated using the public code released by NNPDF [34,35], without refitting. Specifically, the χ^2 is computed by the `perreplica_chi2_table` function of program `validphys` included in the NNPDF code. We activate the t_0 definition by setting option "`use_t0: True`" in the NNPDF code and using the theory reference values for the `210713-n3fit-001` PDF set provided with the NNPDF code. The kinematics cuts were fixed to be the same as in the NNPDF4.0 global analysis [8]. The minimum values of Q^2 and W^2 for DIS measurements were hence chosen to be 3.49 GeV^2 and 12.5 GeV^2, respectively.

The Hessian representation of the NNPDF4.0 ensemble provides the central replica (f_0) and $N_{\text{eig}} = 50$ error PDF sets f_i corresponding to displacements by a $+1\sigma$ value ($f_0 + \Delta f_i$) along each independent eigenvector (EV) direction. The total $\Delta\chi_i^2$ of each EV set, computed with respect to replica 0, varies among the individual EV sets, with some $\Delta\chi_i^2$ being as large as $+32$ (for EV 1) or low as -15 (for EV 2), and with the majority no more than 2–5 units in magnitude. As only one error set is provided per EV direction, this creates an expectation of an approximately symmetric quadratic behavior of $\Delta\chi^2$ centered on f_0. This expectation is illustrated in Fig. 4 for EV set 9 as a red parabola, in which the red points correspond to replica 0 and EV set 9. The horizontal axis is labeled in units of the 1σ displacement for EV set 9, and the vertical axis shows $\Delta\chi^2$.

As an alternative to the red parabola, the actual $\Delta\chi^2$ behavior might have been very irregular, which may happen if NN fits show large deviations from Gaussianity. To test which of the two hypotheses is correct, we explicitly computed the $\Delta\chi^2$ at green points, for which the LHAPDF tables are constructed as $f_0 + w_i\Delta f_i$, where the real parameter w_i quantifies the displacement on the respective horizontal axis. Figure 5 shows these $\Delta\chi^2$ scans for all N_{eig} EV directions. In each EV direction, we evaluate χ^2 at 16 green points for a total of 800 points, with Fig. 5 showing only the points with $\Delta\chi^2$ below a few tens. We observe that $\Delta\chi^2$ follows regular dependence consistent with a quadratic one along all EV directions. However, the minima of the χ^2 are displaced from the central replica along many EV directions. Blue curves interpolating the green points are consistent with symmetric parabolas whose minima, $f_{i,\text{min}} \neq f_0$, are displaced from replica 0 in many EV directions and render negative $\Delta\chi_{i,\text{min}}^2$ for some EV directions that can be as low as ≈ -15 (for EV 2). The widths of the reconstructed parabolas vary noticeably. These observations strongly suggest the regular, quasiquadratic behavior of χ^2 in the vicinity of the central NNPDF4.0 replica and the existence of a displaced global minimum in parameter space for which the χ^2 is smaller than the value provided by the central replica. EV sets and

replicas with negative $\Delta\chi^2$ were also pointed out in a thesis by the NNPDF collaboration [65]. Yet that study did not provide further details, such as regular, approximately Gaussian dependence of χ^2 revealed by the hopscotch scans.

The hopscotch scan technique explores such low-χ^2 region by focusing on specific QCD cross sections. [Finding the displaced global minimum in the whole 50-dimensional space is more computationally expensive and beyond our study's scope, as complexity of combinatorial and geometrical factors increases quickly.] We draw a low-dimensional "court" based on the χ^2 behavior gleaned from the EV direction scans and then repeatedly "throw a marker" according to one of the strategies to generate the PDF replicas at points inside the court.

Initially, to find a region with replicas satisfying $\Delta\chi^2 \leq T^2$ in the plane of two cross sections, such as $\sigma_{t\bar{t}}$ and σ_Z, we use the interpolated parabolas in Fig. 5 to find up to two "pole" PDF sets corresponding to $\Delta\chi^2 = T^2$ for each of 50 EV directions. We plot the $\{\sigma_{t\bar{t}}, \sigma_Z\}$ pair for each pole set, as is done for $T^2 = 0$ in the upper left panel of Fig. 6. In the N_{eig} dimensional space, the pole sets correspond to the corners of a rectangular block whose projection on the $\{\sigma_{t\bar{t}}, \sigma_Z\}$ plane is a polygon with the corners corresponding to the EV directions with the largest displacements of cross sections from the central predictions. In the upper left Fig. 6, these are EV directions 5, 2, 7, 1, 30, 11, 3, and 12. The other EV directions (examined, but not shown in the figure) generate smaller displacements. For this cross section pair, we initially generated 2×300 replicas in the court consisting of two rectangular blocks spanning complementary groups of EV directions. A replica is generated in an n-dimensional block as $f = f_0 + \sum_{i=1}^{n} w_i\Delta f_i$, where each w_i is a random real number that is uniformly distributed along the ith EV direction between the two corresponding pole sets with $\Delta\chi^2 = T^2$. We also generated the replicas for three more pairs of cross sections: Z vs $W^{\pm}$ (summed over the W boson charges); W^+ vs W^-; $t\bar{t}$ vs H.

In the lower row of Figs. 6, we show the dominant EV directions and replica samples for the σ_Z vs. σ_H pair, which was not included in the generation of replicas. However, since this pair shares the dominant directions with the sampled cross section pairs, we can predict the PDF uncertainties for this pair as well.

Our cumulative set from all scans contains 2329 PDF ensembles.[5] In the right column of Fig. 6, we use varied

---

[5]While we refer to the hopscotch ensembles as "replicas," they are not MC replicas in the sense adopted in the NNPDF formalism. The hopscotch replicas simultaneously have very good χ^2 with respect to the central data values and large displacements for the selected cross sections. The traditional MC replicas are obtained by randomly fluctuating the data or PDF parameters instead of directed search, like the hopscotch scans: the majority of them have a positive $\Delta\chi^2$ in the range of hundreds of units [21].
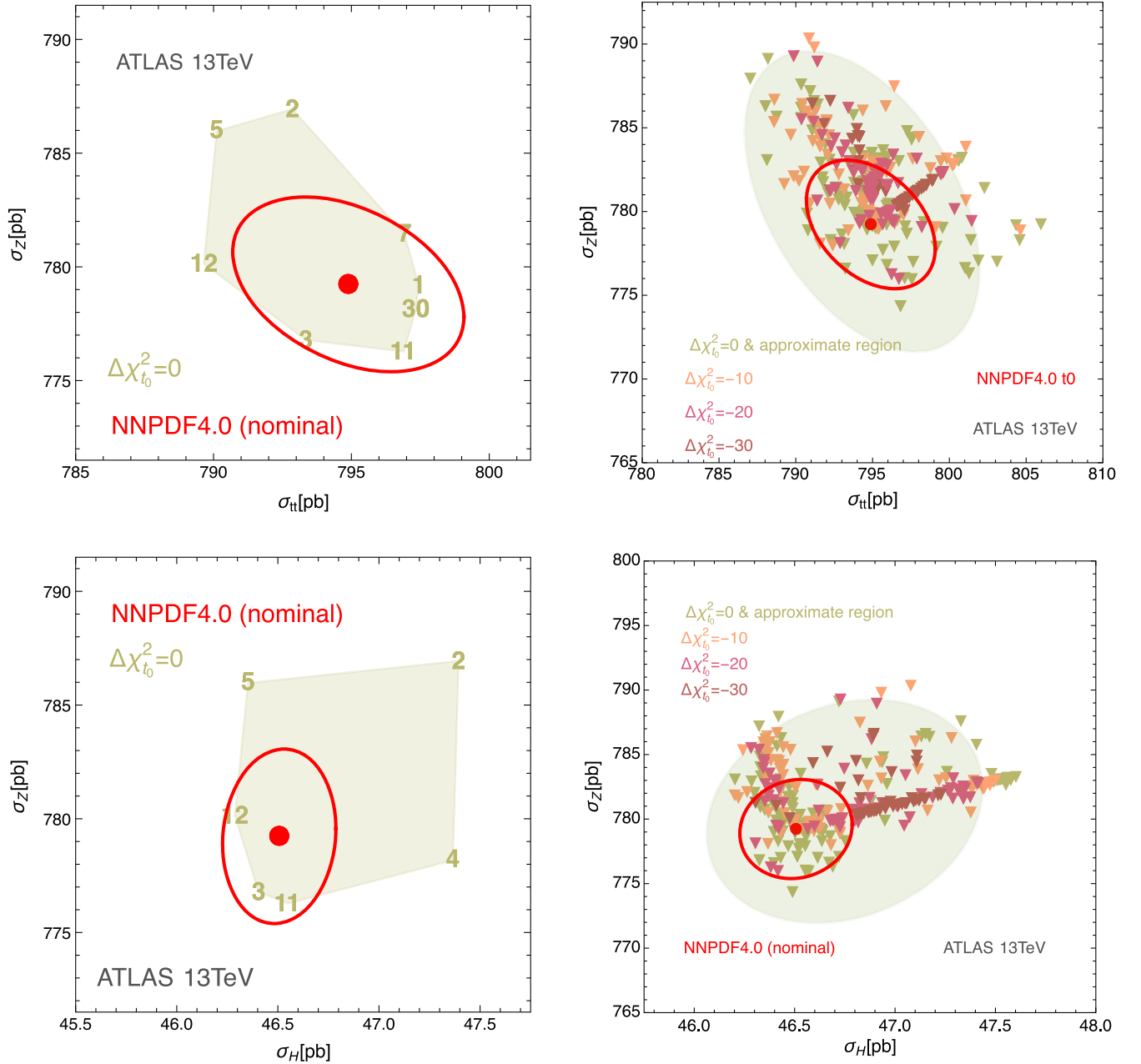
FIG. 6.   Intermediate hopscotch scan results for Z vs. $t\bar{t}$ cross sections (upper row) and Z vs. Higgs boson cross sections (lower row) for ATLAS at 13 TeV. See the Appendix for details of the computation. The left panels shows polygons formed by the pole sets with $\Delta\chi^2 = 0$. In the right panels, the green triangles correspond to $\Delta\chi^2 = 0$, with replicas with lower $\Delta\chi^2$ shown in increasing hue. Green ellipses are approximate regions fitted to the $\Delta\chi^2 < 0$ boundary points. Red ellipses correspond to the 68% probability regions from the published NNPDF4.0 Hessian set.

colors to plot subsamples of replicas that have $\Delta\chi^2 \pm 3$ around the $\Delta\chi^2$ values specified in the figure. The distribution of these replicas is consistent with that of an apparently displaced global minimum, near which some replicas have $\Delta\chi^2$ as low as -37 units. The lowest $\Delta\chi^2$ corresponds to the regions populated by brown markers.

The hopscotch scan is mainly a search algorithm and, in the current realization, does not include any convergence criteria nor the certainty to find the true global minimum.

[These aspects can be further developed along the lines discussed in Sec. II D.] The role of the hopscotch is to reduce the dimensionality of the search for solutions with a lower χ^2 and to identify regions in the cross section space corresponding to such solutions.

While our set of solutions is not exhaustive, it can be used to estimate the size of the projected area for a given value of $\Delta\chi^2$, say $\Delta\chi^2 < 0$. The sample's convex hull gives a crude boundary of this region. On the other hand, since
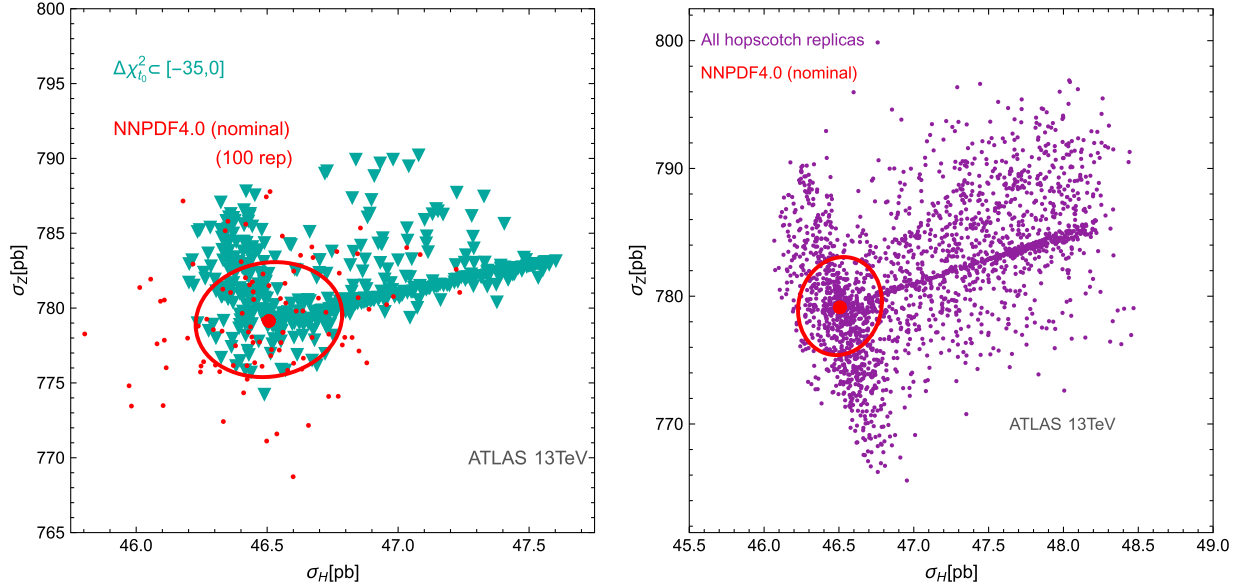
FIG. 7. Left: Hopscotch scan results for the Higgs vs. Z cross section for ATLAS at 13 TeV. Here we show clouds of alternative replicas (cyan) that have $-35 \leq \Delta\chi^2 \leq 0$ with respect to the NNPDF4.0 central replica. The red points indicate predictions with the 100-replica NNPDF4.0 ensemble, whose $\Delta\chi^2$ values are typically at a few positive hundreds of units. See text and footnote 5 for more details. Right: The distribution of 2329 hopscotch replicas.

the EV scans in Fig. 5 are strongly indicative of the approximately Gaussian behavior of χ^2, it seems reasonable to assume that the populated regions in the cross section planes are approximately elliptical. With this information, a highly effective approach to estimate the boundary is to fit an ellipse to the outermost points of the replica subsample in the cross section plot. The quadratic form describing each ellipse can be computed algebraically using a public Mathematica program from [66] for reconstruction of multidimensional ellipsoids from such projections. A 2-dimensional ellipse can be reconstructed by having as few as 6 points on the convex hull of the sample. In our case, we select no less than 15 outermost points per ellipse, so they can be fitted with good certainty.

These approximate elliptical regions for $\Delta\chi^2 < 0$ are shown in Figs. 3 and 6 in light green. These approximate areas covered by all hopscotch replicas with $\Delta\chi^2 < 0$ have been chosen to be centered on the center-of-mass of the hulls drawn on the left-hand side (lhs) of Fig. 6, instead of the minimum χ^2 value found through our scans. Their absolute position is hence indicative of the initial hopscotch "court." *To each of the hopscotch replicas with $\Delta\chi^2 < 0$, represented by triangles in the right-hand side (rhs) of Fig. 6, one could also associate the experimental uncertainty due to data resampling in the same way as the nominal uncertainties (red ellipses) are computed.*

D. The hopscotch scans find the missing good solutions

The hopscotch exercise demonstrates the degree to which predictions for LHC cross sections depend on the

sampling procedures and priors adopted by the groups. To the question: "Which of our generated replicas are acceptable for predicting the LHC cross sections?," the answer accounting only for the likelihoods is "Apparently, all of them that have good χ^2," echoing the likelihood-ratio test described in Sec. III A.

If we also want to explore the priors, seeking acceptable PDF solutions becomes a notorious "needle in a high-dimensional haystack" issue recognized in studies of quasi-MC integration [24,42,45]. To see this, let us take a step back and recall that each NNPDF MC replica is specified by a vector of a large size (of order 800 elements) containing NN latent parameters. The closure test demonstrates that of order 1000 MC replicas reproduce, within some accuracy, expected uncertainties in the PDFs and predictions due to the fluctuations of the pseudodata when training the replicas with a fixed methodology. When predicting a vector of N observables, predictions based on the MC replicas are distributed relatively isotropically. This is illustrated in Fig. 7(left) and Fig. 8, where 2-dimensional projections of the vectors of N LHC cross sections, computed for the 100 nominal replicas (red points) and 1000 replicas (green points), can be converted into approximately spherical distributions by coordinate rotations and rescalings.

Hessian PDFs provide a convenient eigenvector basis that captures PDF variations in "only" 50 dominant dimensions around the NNPDF replica 0. Examinations of χ^2 along the 50 EV directions in Fig. 5 suggest that the global χ^2 minimum is displaced with respect to replica 0 in a direction that does not coincide with any EV direction. And, if we identify a few EV directions that dominate a
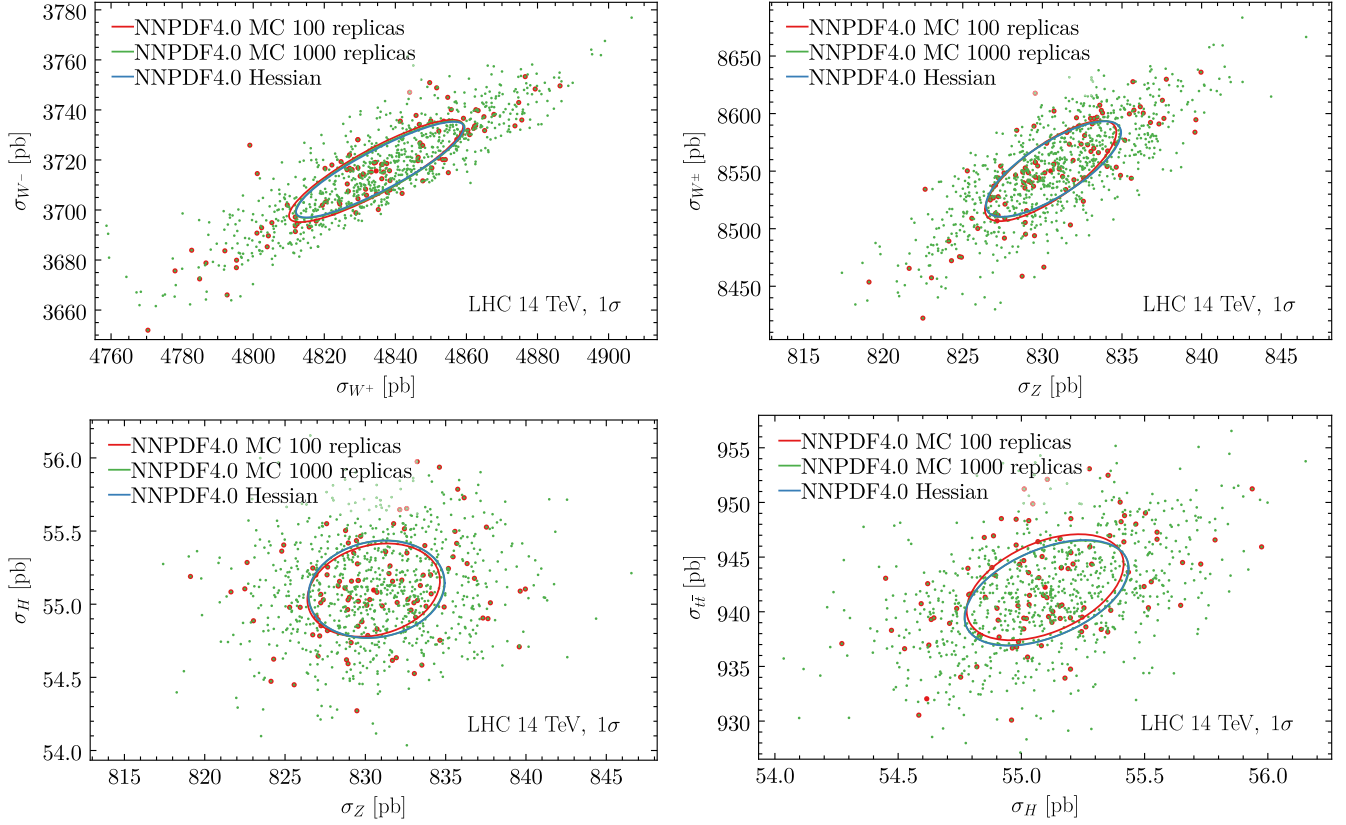
FIG. 8. LHC total cross sections at 14 TeV predicted using the NNPDF4.0 NNLO 1000-replica, 100-replica, and Hessian PDF ensembles. The ellipses indicate 1σ probability regions computed with each ensemble.

given cross section, we can sample these directions more densely than allowed by the isotropic sampling based on 1000 replicas.

Figure 7 illustrates how the hopscotch scans perform targeted sampling of the parameter space based on the guidance from the quasi-parabolic χ^2 distributions in Fig. 5. For each selected pair of cross sections, the hopscotch replicas densely populate a low-dimensional region in the parameter space where χ^2 decreases, while the cross sections show high variability. In Fig. 7(left), we show predictions with the hopscotch replicas that have $-35 < \Delta\chi^2 < 0$ with respect to the χ^2 of the NNPDF4.0 central replica. The low-χ^2 replicas are selected out of 2329 replicas that populate lower-dimensional hyperplanes in which χ^2 decreases or increases slowly as a function of the selected cross sections. These hyperplanes and directions of the scans are identified based on Fig. 5. In Fig. 7(right), we see the projection of the distribution of 2329 replicas (with any χ^2) on the σ_Z vs σ_H plane. The replicas are denser in the regions where the hyperplanes cross the projection plane. We remind the reader that this set of solutions is not exhaustive.

We already noted that both the NNPDF 100-replica ensemble and the NNPDF Hessian ensemble reproduce well the underlying distribution of replicas in their 1000-replica ensemble. An illustration is provided in Fig. 8,

where the LHC cross sections are predicted using the three ensembles. Here the clouds of 100 replicas are consistent with the density distributions of 1000 replicas. The 68% probability regions given by the ellipses are also consistent among the three ensembles.

The NNPDF4.0 Hessian ensemble employed for χ^2 scans in Fig. 5 captures overall properties of the underlying replica distribution. Yet, the low density and distribution of MC replicas does not capture the features of χ^2 revealed by the Hessian scans in Fig. 5 or predict the parametric dependence of the replicas with the negative $\Delta\chi^2$ that have been noticed before [65].

Upon a closer examination of the hopscotch scan, its generated alternative PDFs for $\Delta\chi^2 \leq 0$ appear to pass the standard validation adopted in the CT fits. They are linear combinations of well-behaving Hessian sets that are sufficiently smooth and positive in the x region with the data constraints. At $Q = 2$ GeV, only a few of them are negative in the extrapolation regions, where their behavior can be easily adjusted without changing the agreement with the data. We haven't scrutinized systematically the integrability of T_3 and T_8, as done in the NNPDF4.0 fit, yet we observed no compelling reason to discard these alternative solutions.

If the hopscotch solutions are acceptable, a natural question to raise is why they are not covered by the

nominal NNPDF4.0 ensemble. Since these solutions have a good χ^2, the conclusion from our test is that they are disqualified by the NNPDF prior probability. Indeed, the preliminary studies by NNPDF indicate that some of these replicas (possibly a few dozen out of 2329) fail NNPDF4.0 requirements for smoothness of PDF solutions [67]. If so, the dependence on the priors would be best investigated in collaborative, comprehensive benchmarking exercises among the PDF-fitting groups, using agreed-upon criteria and computational tools.

We also observe that any hopscotch solution can be represented by a neural network in accord with the universal approximation theorem [53–55]. The challenge of representative sampling in a high-dimensional space must therefore be also present in the NN approach. We argued in Sec. III A that the use of data resampling (called "importance sampling" by NNPDF), combined with a fixed methodology that makes specific choices for the NN architecture, the cost function, stopping and smoothness conditions [34,68], does not address samplings over methodology-related settings at the various levels of the global analysis. In the NNPDF4.0 analysis and closure test, the hyperparameters of the methodology were optimized according to a convention, not sampled in the optimum's vicinity. Variations in training methodology are a part of the full uncertainty, together with the theoretical uncertainty and another insufficiently understood source of uncertainty due to the prescription for experimental systematic errors. We have emphasized that the distribution of replicas with good χ^2 depends on the χ^2 definition. This dependence cannot be neglected at the contemporary accuracy level.

E. A case study: Quark sea flavor composition and small-x gluon

Implications of the hopscotch PDF solutions for uncertainties on various QCD observables are of significant practical interest. The χ^2 scans along the NNPDF4.0 Hessian EV directions in Fig. 5 indicate that, for each EV direction, there is a displaced PDF set that has exactly the same χ^2 as the NNPDF central replica 0. Just accounting for these alternative sets can enlarge the nominal PDF uncertainty. On the companion website [69], we provide the LHAPDF grids for two 50-member ensembles of the alternative sets with $\Delta\chi^2 = 0$ for the two χ^2 definitions, as well as figures comparing these PDFs with the nominal NNPDF4.0 NNLO uncertainty bands.

For example, variations along Hessian EV directions 25 and 33 influence strongly the flavor composition of sea quarks and antiquarks at $x > 0.2$, where the relevant experimental constraints remain very weak. Figure 9 presents two illustrations. The left panel shows the nominal NNPDF4.0 uncertainty at the 68% probability for the strange-antistrange asymmetry, $A_{\text{str}}(x, Q) \equiv (s(x, Q) - \bar{s}(x, Q))/(s(x, Q) + \bar{s}(x, Q))$ at $Q = 1.7$ GeV. In the recent NNLO fits that allow strange quark and antiquark PDFs to differ, the CT18As [70,71], MSHT'20, and NNPDF4.0 analyses all prefer a very large positive $A_{\text{str}}(x, Q)$ at $x > 0.3$, which can even exceed 100% by allowing the $\bar{s}$ PDF to go negative [Sec. 4.5 in [15]]. Such behavior may reflect some tensions between the experiments. Among these fits, the positive $A_{\text{str}}(x, Q)$ in NNPDF4.0 may be taken to be most significant at $x \approx 0.2$, given the smallest nominal uncertainty. However, the alternative EV set 33 for $\Delta\chi^2_{t_0} = 0$ in the left panel is consistent with a negative $A_{\text{str}}(x, Q)$. From the plot of parabolas for EV direction 33 in Fig. 5, we see that even deeper negative variations of $A_{\text{str}}(x, Q)$ are allowed if χ^2_{exp} is used, or if simultaneous variations along EV direction 33 and other EV directions are considered. [Note that the EV directions specified by the NNPDF4.0 Hessian set do not change among the χ^2 definitions.]
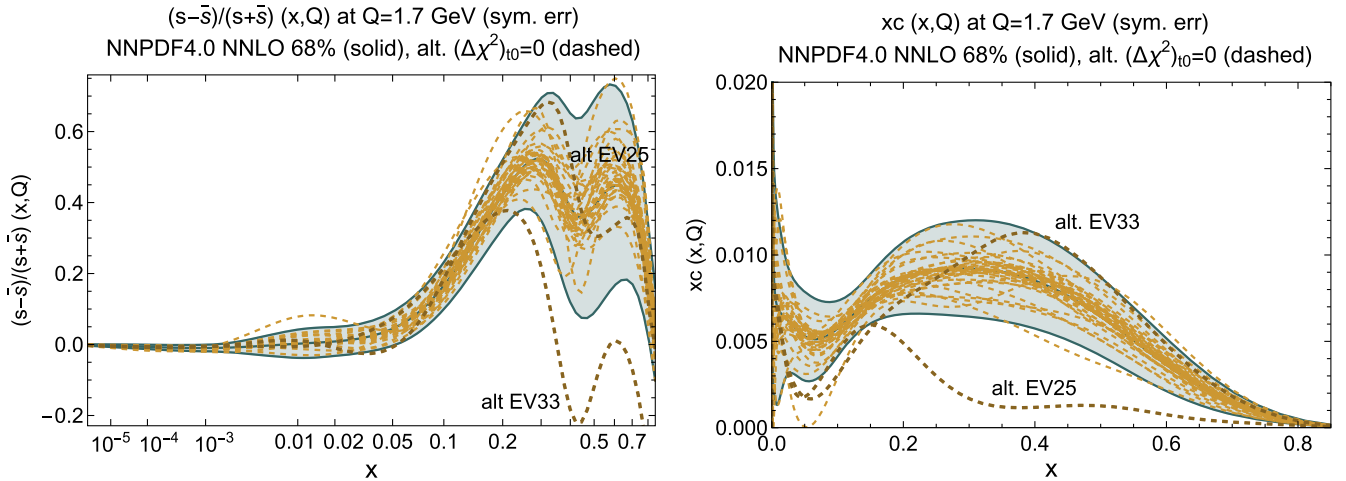

FIG. 9.   Solid bands indicate the nominal 68% NNPDF4.0 uncertainties for strangeness asymmetry (left) and charm PDF (right) at $Q = 1.7$ GeV. The alternative EV sets with $\Delta\chi^2_{t_0} = 0$ are plotted as dashed lines.

The right panel of Fig. 9 shows the counterpart plot for charm PDF $c(x, Q)$ at $Q = 1.7$ GeV. The nominal error band may suggest a significant nominal enhancement of the charm PDF at $x = 0.2$–0.3, approximately in the same x range where a large $A_{str}(x, Q)$ appears in the left panel. The hopscotch analysis shows that the uncertainty on charm PDF is increased by considering the χ^2 variations along the EV directions revealed in Fig. 5. Most notably, the second $\Delta\chi^2_{t_0} = 0$ set for EV direction 25 results in the very small charm at $x > 0.3$ at $Q = 1.7$ GeV. When evolved down to $Q < m_c = 1.51$ GeV, this EV set will result in a vanishing fitted charm at a low scale. After this set is included in the PDF uncertainty, the NNPDF fit does not statistically prefer a nonzero fitted charm at the initial scale, as would be concluded based on the nominal 1σ uncertainty [30]. Including variations along the other EV directions, e.g., 33 that favors a smaller (larger) charm PDF at $x = 0.05$–0.1 ($x > 0.4$), as well as uncertainties in the model for systematic errors, further washes out the preference for the nonzero fitted charm at large x and low Q. Indeed, the recent CT18 FC analysis conveys that there is no evidence for intrinsic charm so far [31].

Similar examinations for other PDF flavors and flavor combinations (collected on the companion website [69]) indicate that the alternative $\Delta\chi^2 = 0$ solutions expand the uncertainty on the gluon PDF at low x and on the T_3 and T_8 combinations of quark and antiquark flavors. One of the unexpected findings of the NNPDF4.0 future test was that the fit without including the HERA DIS data preferred the general growth of the gluon PDF at $Q = 1.65$ GeV and $x < 10^{-3}$, where no constraints were available. See Fig. 29 for $xg(x, Q)$ at $Q = 1.65$ GeV in [8], where the solid green band for NNPDF4.0 without the HERA data does not cover the blue and red bands that include these data. A similar, also less pronounced trend is also seen with the NNPDF3.1 methodology in their Fig. 28. Historically, the solutions with the growing, flat, or even decreasing gluon at $x < 10^{-2}$ were allowed in the CTEQ fits in the early 1990s, before the advent of the HERA data. Indeed, no experimental constraints existed in the pre-HERA data in this region, similarly to the current situation with the nuclear PDFs that have an essentially unconstrained gluon at $x < 10^{-2}$ and may be affected by strong nuclear shadowing. See, for example, Fig. 6 in the 1995 ZEUS publication [72], in which the pre-HERA data at $W^2 < 500$ GeV^2 and $Q^2 \geq 4.5$ GeV^2 do not favor any particular trend of the $\gamma^* p$ total cross section at $W^2 > 500$ GeV^2, and hence they do not constrain the gluon PDF at $x = Q^2/(W^2 + Q^2) \lesssim 0.01$ via scaling violations. Therefore, it is surprising that the NNPDF4.0 pre-HERA future test disfavors the post-HERA small-x gluon behavior.

The $\Delta\chi^2 = 0$ variations with the alternative EV sets 1, 2, and 4 expand the nominal uncertainty in $xg(x, Q)$ of the full NNPDF4.0 set, especially in the downward direction at $x < 10^{-2}$. They modify the NNPDF pre-HERA future tests, too. In the same vein, considering the hopscotch solutions indicates larger uncertainties on the flavor combinations $T_3 \equiv u + \bar{u} - d - \bar{d}$ and $T_8 \equiv u + \bar{u} + d + \bar{d} - 2s - 2\bar{s}$ than seen in Fig. 49 of Ref. [8]. We note that the hopscotch replicas agree with the sum rules and integrability, especially as the PDF behaviors at $x \to 0$ (outside of the data region) can always be adjusted to obtain convergent first moments.

IV. CONCLUSIONS

PDF uncertainties in high-stake measurements (Higgs cross sections, W boson mass…) should be examined for robustness of results to sampling of available experimental datasets and PDF parametrizations. Likewise, tests of manifestations of nonperturbative QCD, such as the asymptotic large-x behavior of intrinsic charm, depend on interpretations of PDF uncertainties [30,31,73]. Sampling biases may arise in PDF fits operating with large populations of possible solutions. Increasing the volumes of the fitted data and parametric space may increase, not reduce, the sample expectation deviation. An undetected deviation may result in a wrong prediction with a low nominal uncertainty. Sampling biases may limit reduction of the PDF uncertainties and explain some differences between the PDF sets.

For these reasons, global fits are potentially vulnerable to unrepresentative sampling when their overall scope (including the number of PDF parameters, size of datasets, range of possible assumptions) grows. As a way to mitigate the risk of underestimation in specific applications, statistical literature suggests to swap democratic sampling in all dimensions for preferential sampling in fewer dimensions that are most relevant to the task at hand.

In the Monte-Carlo (MC) replica method, constructing the Hessian eigenvector (EV) sets from the MC PDF set introduces a convenient coordinate system for such dimensionality reduction. Taking the W boson mass measurements as an example, we could identify the few Hessian sets that give the largest contribution to the M_W PDF error. It is then more effective to sample these EV directions with a higher density of replicas to look for acceptable PDFs that may be outside of the nominal MC uncertainty. We presented a technique of hopscotch scans to perform such estimation.

With this technique that does not require PDF refitting, we have demonstrated that the NNPDF4.0 fitting code allows alternative solutions of their global fit that predict the LHC cross sections outside of the nominal NNPDF4.0 uncertainties, while having the same total χ^2 as the NNPDF4.0 central replica and satisfying typical validation criteria adopted in the CT fit. Literature on ML and analyses of high dimensionality suggests that those solutions may exist and are not necessarily ruled out on the

basis of a low *prior* probability. Instead, for those solutions that display an acceptable value of the likelihood, representative sampling over methodological settings will contribute to the confounding correlation. A related observation is that the dependence of the distribution of acceptable predictions on the prescription for implementation of experimental systematic errors cannot be neglected at the targeted level of accuracy [Sec. 5.1 in [14]].

In the other two examples presented in Fig. 9, we show that including the low-χ^2 solutions from the hopscotch scans relaxes the NNPDF4.0 uncertainty on the flavor composition of sea quarks and antiquarks at $x > 0.2$ and $Q < 2$ GeV. As a result, both a negative strange-antistrange asymmetry and a zero fitted (intrinsic) charm PDF are statistically allowed at the Q scale of order 1.5 GeV.

In either the MC or Hessian methods, a comprehensive range of fits must be explored to understand variations due to the functional forms and other choices. This viewpoint is taken in the CTEQ-TEA family of analyses, in which the tolerance on the fixed PDF functional form of the published set is selected so as to cover candidate best-fit PDFs found with the alternative choices. In other words, one must pay attention both to the quality of accepted fits and their representative sampling. For example, when some experiments disagree, it should be either understood that fitting all experiments at once will either fail the strong goodness-of-fit test [16] or, if such a fit is nevertheless accepted, the tolerance may need to be increased, as the experimental tensions suggest a larger uncertainty on the full population.

Instead of considering a large population of N_p acceptable solutions, for specific predictions, the trio identity equation (3) can help to design a procedure that produces unbiased and reliable estimates using a sample of a smaller size $N_s \ll N_p$. The overall spirit of this approach is similar to dataset diagonalization [46] and replica unweighting [51,74]. The R mechanism realises a generalization of such techniques and can select fits based on the value of χ^2 or other figures of merits.

We make LHAPDF6 grids of the alternative PDF replicas available for the future analyses [69].

ACKNOWLEDGMENTS

APPENDIX: COMPUTATION OF HADRONIC CROSS SECTIONS

In this section, we summarize settings of the computations of LHC cross sections shown in the main part of the article. The cross sections are computed at NNLO in the QCD coupling strength without cuts, unless specified otherwise.

Drell-Yan $W^\pm/Z$ production. For $W^\pm/Z$ boson production at the Tevatron 1.96 TeV, we impose the CDF fiducial cuts [28],

$$W^\pm: \; 30 < p_T^{\ell,\nu} < 55 \text{ GeV}, \qquad |\eta_\ell| < 1,$$
$$u_T < 15 \text{ GeV}, \qquad 60 < m_T < 100 \text{ GeV}; \quad \text{(A1)}$$

$$Z: \; 30 < p_T^{\ell} < 55 \text{ GeV}, \qquad |\eta_\ell| < 1,$$
$$u_T < 15 \text{ GeV}, \qquad 66 < m_{\ell\bar{\ell}} < 116 \text{ GeV}, \quad \text{(A2)}$$

where

$$u_T = |\vec{p}_T^{\ell} + \vec{p}_T^{\nu(\bar{\ell})}|, \qquad m_T = \sqrt{2(p_T^{\ell} p_T^{\nu} - \vec{p}_T^{\ell} \cdot \vec{p}_T^{\nu})}. \tag{A3}$$

For W/Z boson production at the LHC, we adopt the ATLAS 13 TeV fiducial cuts [75],

$$W^\pm: \; p_T^{\ell,\nu} > 25 \text{ GeV}, \qquad |\eta_\ell| < 2.5, \qquad m_T > 50 \text{ GeV}; \tag{A4}$$

$$Z: \; p_T^{\ell} > 25 \text{ GeV}, \qquad |\eta_\ell| < 2.5,$$
$$66 < m_{\ell\bar{\ell}} < 116 \text{ GeV}. \quad \text{(A5)}$$

The theoretical calculation is performed with a fast computation table APPLgrid [76] at NLO, combined with NNLO/NLO point-by-point K-factors calculated with MCFM [77,78]. The renormalization and factorization scales are set equal to the invariant mass of the lepton pair, $m_{\ell\bar{\ell}}$ or $m_{\ell\nu}$.

Top-quark pair production. Top-quark pair production is measured by both ATLAS and CMS groups at 13 TeV [79,80] and presented in the form of total cross sections.

Here we take the public code top++ [81] to compute these cross sections at NNLO, with the threshold logarithms of soft gluons resummed up to the NNLL level. The factorization and renormalization scales are set to the top-quark mass m_t.

Higgs production. The calculation is done with GGHIGGS [82] using the factorization and renormalization scales equal to m_H.

Associated production of Higgs bosons and top-quark pairs. Recently a part of the NNLO calculation for $t\bar{t}H$ production came out [83], while no public code has been released yet. Instead, we make predictions using MADGRAPH_AMC@NLO [84] interfaced with PINEAPPL [85] at NLO, and using NNLO PDFs. The renormalization and factorization scales are set to be equal to the partonic collision energy $\sqrt{\hat{s}}$.

[1] H. Abramowicz *et al.* (H1 and ZEUS Collaborations), Combination of measurements of inclusive deep inelastic $e^{\pm}p$ scattering cross sections and QCD analysis of HERA data, Eur. Phys. J. C **75**, 580 (2015).

[2] A. Accardi, L. T. Brady, W. Melnitchouk, J. F. Owens, and N. Sato, Constraints on large-x parton distributions from new weak boson production and deep-inelastic scattering data, Phys. Rev. D **93**, 114017 (2016).

[3] S. Alekhin, J. Blümlein, S. Moch, and R. Placakyte, Parton distribution functions, α_s, and heavy-quark masses for LHC Run II, Phys. Rev. D **96**, 014011 (2017).

[4] R. D. Ball *et al.* (NNPDF Collaboration), Parton distributions from high-precision collider data, Eur. Phys. J. C **77**, 663 (2017).

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

[6] S. Bailey and L. Harland-Lang, Differential top quark pair production at the LHC: Challenges for PDF fits, Eur. Phys. J. C **80**, 60 (2020).

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

[8] R. D. Ball *et al.*, The path to proton structure at 1% accuracy: NNPDF Collaboration, Eur. Phys. J. C **82**, 428 (2022).

[9] G. Aad *et al.* (ATLAS Collaboration), Determination of the parton distribution functions of the proton using diverse ATLAS data from pp collisions at $\sqrt{s} = 7$, 8 and 13 TeV, Eur. Phys. J. C **82**, 438 (2022).

[10] R. Abdul Khalek *et al.*, Science requirements and detector concepts for the electron-ion collider: EIC yellow report, Nucl. Phys. **A1026**, 122447 (2022).

[11] P. Agostini *et al.* (LHeC and FCC-he Study Group), The large hadron-electron collider at the HL-LHC, J. Phys. G **48**, 110501 (2021).

[12] D. Acosta and W. Li, A muon–ion collider at BNL: The future QCD frontier and path to a new energy frontier of $\mu + \mu-$ colliders, Nucl. Instrum. Methods Phys. Res., Sect. A **1027**, 166334 (2022).

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

[14] S. Amoroso *et al.*, Snowmass 2021 whitepaper: Proton structure at the precision frontier, arXiv:2203.13923.

[15] R. D. Ball *et al.* (PDF4LHC Working Group), The PDF4LHC21 combination of global PDF fits for the LHC Run III, J. Phys. G **49**, 080501 (2022).

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

[17] J. Gao, L. Harland-Lang, and J. Rojo, The structure of the proton in the LHC precision era, Phys. Rep. **742**, 1 (2018).

[18] R. Bellman, *Adaptive Control Processes. A Guided Tour* (Princeton University Press, Princeton, NJ, 1961).

[19] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)* (Springer-Verlag, Berlin, Heidelberg, 2006).

[20] G. Watt and R. S. Thorne, Study of Monte Carlo approach to experimental uncertainty propagation with MSTW 2008 PDFs, J. High Energy Phys. 08 (2012) 052.

[21] T.-J. Hou *et al.*, Reconstruction of Monte Carlo replicas from Hessian parton distributions, J. High Energy Phys. 03 (2017) 099.

[22] X.-L. Meng, Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election, Ann. Appl. Stat. **12**, 685 (2018).

[23] V. Bradley, S. Kuriwaki, M. Isakov, D. Sejdinovic, X.-L. Meng, and S. Flaxman, Unrepresentative big surveys significantly overestimated US vaccine uptake, Nature (London) **600**, 695 (2021).

[24] F. J. Hickernell, The trio identity for Quasi-Monte Carlo error, in Monte Carlo and Quasi-Monte Carlo Methods (MCQMC 2016), *Springer Proceedings in Mathematics & Statistics*, edited by A. B. Owen and P. W. Glynn (2018), Vol. 241, p. 3, arXiv:1702.01487.

[25] P. Msaouel, The big data paradox in clinical practice, Cancer investigation **40**, 567 (2022).

[26] S. Geman, E. Bienenstock, and R. Doursat, Neural networks and the bias/variance dilemma, Neural Comput. **4**, 1 (1992).

[27] A. Puy, P. Beneventano, S. A. Levin, S. L. Piano, T. Portaluri, and A. Saltelli, Models with higher effective dimensions tend to produce more uncertain estimates, Sci. Adv. **8**, eabn9450 (2022).

[28] T. Aaltonen *et al.* (CDF Collaboration), High-precision measurement of the W boson mass with the CDF II detector, Science **376**, 170 (2022).

[29] J. Butterworth *et al.*, PDF4LHC recommendations for LHC Run II, J. Phys. G **43**, 023001 (2016).

[30] R. D. Ball, A. Candido, J. Cruz-Martinez, S. Forte, T. Giani, F. Hekhorn, K. Kudashkin, G. Magni, and J. Rojo NNPDF Collaboration, Evidence for intrinsic charm quarks in the proton, Nature (London) **608**, 483 (2022).

[31] M. Guzzi, T. J. Hobbs, K. Xie, J. Huston, P. Nadolsky, and C. P. Yuan, The persistent nonperturbative charm enigma, arXiv:2211.01387.

[32] J. Gao and P. Nadolsky, A meta-analysis of parton distribution functions, J. High Energy Phys. 07 (2014) 035.

[33] The mcgen program for generation of Monte-Carlo replicas and algebraic manipulations with LHAPDF6 tables, https://metapdf.hepforge.org/mcgen/.

[34] R. D. Ball *et al.* (NNPDF Collaboration), An open-source machine learning framework for global analyses of parton distributions, Eur. Phys. J. C **81**, 958 (2021).

[35] The NNPDF open source code, http://nnpdf.mi.infn.it/nnpdf-open-source-code/.

[36] R. D. Ball, S. Forte, and R. Stegeman, Correlation and combination of sets of parton distributions, Eur. Phys. J. C **81**, 1046 (2021).

[37] A. Apyan, D. Froidevaux, and B. Malaescu. Presentations on studies of data-driven correlations in PDF fits, PDF4LHC meeting, CERN, 2018, https://tinyurl.com/4wcnd8xn; LHC EW Precision sub-group workshop, IPPP Durham UK, 2019, https://tinyurl.com/2p8d8ba3; LHC Electroweak WG meeting, CERN, 2019, https://tinyurl.com/2p8tcn5b.

[38] R. Abdul Khalek *et al.* (NNPDF Collaboration), A first determination of parton distributions with theoretical uncertainties, Eur. Phys. J. C **79**, 838 (2019).

[39] R. Abdul Khalek *et al.* (NNPDF Collaboration), Parton distributions with theory uncertainties: General formalism and first phenomenological studies, Eur. Phys. J. C **79**, 931 (2019).

[40] R. D. Ball and R. L. Pearson, Correlation of theoretical uncertainties in PDF fits and theoretical uncertainties in predictions, Eur. Phys. J. C **81**, 830 (2021).

[41] S. Forte and Z. Kassabov, Why α_s cannot be determined from hadronic processes without simultaneously determining the parton distributions, Eur. Phys. J. C **80**, 182 (2020).

[42] I. H. Sloan and H. Woźniakowski, An intractability result for multiple integration, Math. Comput. **66**, 1119 (1997).

[43] R. E. Caflisch, W. J. Morokoff, and A. B. Owen, Valuation of mortgage-backed securities using Brownian bridges to reduce effective dimension, J. Comput. Finance **1**, 27 (1997).

[44] R. Liu and A. B. Owen, Estimating mean dimensionality of analysis of variance decompositions, J. Am. Stat. Assoc. **101**, 712 (2006).

[45] I. H. Sloan and H. Woźniakowski, When are Quasi-Monte Carlo algorithms efficient for high dimensional integrals?, J. Complexity **14**, 1 (1998).

[46] J. Pumplin, Data set diagonalization in a global fit, Phys. Rev. D **80**, 034002 (2009).

[47] T.-J. Hou, Z. Yu, S. Dulat, C. Schmidt, and C. P. Yuan, Updating and optimizing error parton distribution function sets in the Hessian approach. II., Phys. Rev. D **100**, 114024 (2019).

[48] R. D. Ball *et al.* (NNPDF Collaboration), Parton distributions for the LHC Run II, J. High Energy Phys. 04 (2015) 040.

[49] J. C. Collins and J. Pumplin, Tests of goodness of fit to multiple data sets, arXiv:hep-ph/0105207.

[50] M. Aaboud *et al.* (ATLAS Collaboration), Precision measurement and interpretation of inclusive W^+, W^- and Z/γ^* production cross sections with the ATLAS detector, Eur. Phys. J. C **77**, 367 (2017).

[51] 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, Reweighting and Unweighting of Parton Distributions and the LHC W lepton asymmetry data, Nucl. Phys. **B855**, 608 (2012).

[52] D. E. Soper and J. C. Collins, Issues in the determination of parton distribution functions, arXiv:hep-ph/9411214.

[53] G. Cybenko, Approximation by superpositions of a sigmoidal function, Math. Control Signals Syst. **2**, 303 (1989).

[54] K. Hornik, M. Stinchcombe, and H. White, Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks, Neural Netw. **3**, 551 (1990).

[55] K. Hornik, Approximation capabilities of multilayer feed-forward networks, Neural Netw. **4**, 251 (1991).

[56] R. D. Ball *et al.*, Parton distribution benchmarking with LHC data, J. High Energy Phys. 04 (2013) 125.

[57] J. Pumplin, D. R. Stump, J. Huston, H. L. Lai, P. M. Nadolsky, and W.-K. Tung, New generation of parton distributions with uncertainties from global QCD analysis, J. High Energy Phys. 07 (2002) 012.

[58] R. D. Ball, L. Del Debbio, S. Forte, A. Guffanti, J. I. Latorre, J. Rojo, and M. Ubiali (NNPDF Collaboration), Fitting parton distribution data with multiplicative normalization uncertainties, J. High Energy Phys. 05 (2010) 075.

[59] J. Gao, M. Guzzi, J. Huston, H.-L. Lai, Z. Li, P. Nadolsky, J. Pumplin, D. Stump, and C. P. Yuan, CT10 next-to-next-to-leading order global analysis of QCD, Phys. Rev. D **89**, 033009 (2014).

[60] G. D'Agostini, On the use of the covariance matrix to fit correlated data, Nucl. Instrum. Methods Phys. Res., Sect. A **346**, 306 (1994).

[61] G. D'Agostini, Bayesian reasoning in high-energy physics: Principles and applications, Report No. CERN-99-03, 1999.

[62] The NNPDF code manual, Chi square figures of merit, https://docs.nnpdf.science/figuresofmerit/index.html.

[63] LHAPDF6 library of parton distribution functions, https://lhapdf.hepforge.org/.

[64] D. Stump, J. Pumplin, R. Brock, D. Casey, J. Huston, J. Kalk, H.-L. Lai, and W.-K. Tung, Uncertainties of predictions from parton distribution functions. 1. The Lagrange multiplier method, Phys. Rev. D **65**, 014012 (2001).

[65] N. Lambri, Optimized regression models for parton distribution functions determination using deep learning models, master thesis, Master's thesis, Università degli Studii di Milano, 2020, http://nnpdf.mi.infn.it/wp-content/uploads/2021/01/thesislambri.pdf.

[66] R. Anwar, M. Hamilton, and P. M. Nadolsky, Direct ellipsoidal fitting of discrete multi-dimensional data, SMU J. Undergrad. Res. **5**, 1 (2019).

[67] M. Ubiali, *Proceedings of the 8th International Workshop on High Precision for Hard Processes (hp2 2022)* (Durham, UK, 2022), https://lhapdf.hepforge.org/.

[68] S. Carrazza and J. Cruz-Martinez, Towards a new generation of parton densities with deep learning models, Eur. Phys. J. C **79**, 676 (2019).

[69] https://ct.hepforge.org/PDFs/2022hopscotch/.

[70] T.-J. Hou, H.-W. Lin, M. Yan, and C. P. Yuan, Impact of lattice $s(x) - \bar{s}(x)$ data in the CTEQ-TEA global analysis, arXiv:2204.07944.

[71] T.-J. Hou, H.-W. Lin, M. Yan, and C. P. Yuan, Impact of lattice strangeness asymmetry data in the CTEQ-TEA global analysis, arXiv:2211.11064.

[72] M. Derrick *et al.* (ZEUS Collaboration), Measurement of the proton structure function F2 at low x and low Q^2 at HERA, Z. Phys. C **69**, 607 (1996).

[73] A. Courtoy and P. M. Nadolsky, Testing momentum dependence of the nonperturbative hadron structure in a global QCD analysis, Phys. Rev. D **103**, 054029 (2021).

[74] R. D. Ball, V. Bertone, F. Cerutti, L. Del Debbio, S. Forte, A. Guffanti, J. I. Latorre, J. Rojo, and M. Ubiali (NNPDF Collaboration), Reweighting NNPDFs: The W lepton asymmetry, Nucl. Phys. **B849**, 112 (2011); **855**, 927(E) (2012).

[75] G. Aad *et al.* (ATLAS Collaboration), Measurement of $W^{\pm}$ and Z-boson production cross sections in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector, Phys. Lett. B **759**, 601 (2016).

[76] T. Carli, D. Clements, A. Cooper-Sarkar, C. Gwenlan, G. P. Salam, F. Siegert, P. Starovoitov, and M. Sutton, *A posteriori* inclusion of parton density functions in NLO QCD final-state calculations at hadron colliders: The APPLGRID Project, Eur. Phys. J. C **66**, 503 (2010).

[77] J. M. Campbell, R. K. Ellis, and W. T. Giele, A Multi-Threaded Version of MCFM, Eur. Phys. J. C **75**, 246 (2015).

[78] J. Campbell and T. Neumann, Precision phenomenology with MCFM, J. High Energy Phys. 12 (2019) 034.

[79] M. Aaboud *et al.* (ATLAS Collaboration), Measurements of top-quark pair to Z-boson cross-section ratios at $\sqrt{s} = 13$, 8, 7 TeV with the ATLAS detector, J. High Energy Phys. 02 (2017) 117.

[80] A. M. Sirunyan *et al.* (CMS Collaboration), Measurement of the $t\bar{t}$ production cross section, the top quark mass, and the strong coupling constant using dilepton events in pp collisions at $\sqrt{s} = 13$ TeV, Eur. Phys. J. C **79**, 368 (2019).

[81] M. Czakon and A. Mitov, Top++: A program for the calculation of the top-pair cross-section at hadron colliders, Comput. Phys. Commun. **185**, 2930 (2014).

[82] M. Bonvini, S. Marzani, C. Muselli, and L. Rottoli, On the Higgs cross section at $N^3LO + N^3LL$ and its uncertainty, J. High Energy Phys. 08 (2016) 105.

[83] S. Catani, I. Fabre, M. Grazzini, and S. Kallweit, $t\bar{t}H$ production at NNLO: the flavour off-diagonal channels, Eur. Phys. J. C **81**, 491 (2021).

[84] R. Frederix, S. Frixione, V. Hirschi, D. Pagani, H. S. Shao, and M. Zaro, The automation of next-to-leading order electroweak calculations, J. High Energy Phys. 07 (2018) 185; 11 (2021) 85.

[85] S. Carrazza, E. R. Nocera, C. Schwan, and M. Zaro, PineAPPL: Combining EW and QCD corrections for fast evaluation of LHC processes, J. High Energy Phys. 12 (2020) 108.