2021

Simultaneous Monte Carlo Analysis of Parton Densities and Fragmentation Functions

Eric Moffat
Old Dominion University, emoff003@odu.edu

W. Melnitchouk

Ted C. Rogers
Old Dominion University, trogers@odu.edu

N. Sato

Jefferson Lab Angular Momentum Collaboration

Simultaneous Monte Carlo analysis of parton densities and fragmentation functions

E. Moffat,[1,*] W. Melnitchouk[2,†] T. C. Rogers,[1,2,‡] and N. Sato[2,§]

Jefferson Lab Angular Momentum (JAM) Collaboration

[1]*Department of Physics, Old Dominion University, Norfolk, Virginia 23529, USA*
[2]*Jefferson Lab, Newport News, Virginia 23606, USA*

We perform a comprehensive new Monte Carlo analysis of high-energy lepton-lepton, lepton-hadron and hadron-hadron scattering data to simultaneously determine parton distribution functions (PDFs) in the proton and parton to hadron fragmentation functions (FFs). The analysis includes all available semi-inclusive deep-inelastic scattering and single-inclusive e^+e^- annihilation data for pions, kaons and unidentified charged hadrons, which allows the flavor dependence of the fragmentation functions to be constrained. Employing a new multistep fitting strategy and more flexible parametrizations for both PDFs and FFs, we assess the impact of different datasets on sea quark densities and confirm the previously observed suppression of the strange quark distribution. The new fit, which we refer to as "JAM20-SIDIS," will allow for improved studies of universality of parton correlation functions, including transverse momentum dependent (TMD) distributions, across a wide variety of process, and the matching of collinear to TMD factorization descriptions.

I. INTRODUCTION

The standard parton correlation functions of QCD, such as collinear parton distribution functions (PDFs) and fragmentation functions (FFs), are being utilized in an increasingly diverse range of phenomenological applications. Beyond their traditional role in predicting new high energy phenomena, they also enter frequently into the study of more complex and extended objects like transverse momentum dependent (TMD) PDFs and FFs and generalized parton distributions (GPDs), where they are needed to understand the transition between different factorization regions. Both TMDs and GPDs are central to the study of the nonperturbative parton structure of hadrons, and understanding how they encapsulate their longitudinal and transverse features will be critical to current experimental programs at Jefferson Lab and elsewhere, as well as to the

*emoff003@odu.edu
†wmelnitc@jlab.org
‡trogers@odu.edu
§nsato@jlab.org

future Electron-Ion Collider. These considerations provide one of the main motivations for the study of collinear PDFs and FFs in this paper.

The great value of PDFs and FFs extracted from global QCD data analysis lies with their predictive power, or "universality." However, the translation from experimental data to quark and gluon operator structures is a challenging inverse problem. It is not possible to exactly constrain parton correlations from data alone since this connection involves nontrivial convolution integrals in a factorization formalism (whose accuracy itself is difficult to quantify in any given instance) and because of the limited quantity of available data. The complexity of the inverse problem is also magnified by the number of flavor degrees of freedom involved.

Nevertheless, assessing and maximizing the universality of collinear PDFs and FFs is crucial given the increasingly broad scenarios where they are used. A major focus in the current effort by the Jefferson Lab Angular Momentum (JAM) Collaboration is therefore to both test and broaden the predictive power of parton correlation functions. This is achieved through a Bayesian inference procedure in which PDFs and FFs are extracted simultaneously, and the uncertainty quantification associated with particular parametrizations of parton correlation functions is given in terms of a Bayesian posterior distribution. To test universality, the system of equations relating observables to

parton correlation functions must of course exceed the total number of correlation functions involved—a minimum requirement is that the parton correlation functions be overconstrained by the data in the fit. Of course, realizing this in practical analyses requires that all parton correlation functions be truly fitted simultaneously. This is a major numerical and technological challenge, and traditionally PDFs and FFs have thus been extracted in separate procedures. However, simultaneous fits can be achieved with the Bayesian Monte Carlo approach and have been implemented recently in the JAM17 [1] analysis of helicity PDFs, and in the JAM19 [2] analysis of unpolarized PDFs and FFs. The same basic methodology was also applied in the three-dimensional JAM3D20 [3] study, in the first combined analysis of TMD observables that satisfies the overconstraining criterion.

In this paper, we extend the previous work by performing the first simultaneous and overconstrained fit of unpolarized PDFs and FFs that utilizes both charged hadron production in semi-inclusive deep-inelastic scattering (SIDIS) and single-inclusive e^+e^- annihilation (SIA). This is partly motivated by a number of recent observations associated with the study of TMD PDFs. For example, significant tension has recently been found between fits performed with standard sets of PDFs and FFs and fixed order perturbative QCD calculations in processes including SIDIS [4,5], Drell-Yan (DY) [6], and SIA into wide-angle hadron pairs [7]. A number of suggested solutions and explanations have been proposed to account for this, including a possible need for power suppressed corrections [8] at the moderate scales of most SIDIS experiments. However, more tests of the limits of applicability of standard collinear factorization are needed before it is possible to draw firm conclusions. Given that the majority of data used to constrain collinear correlation functions (both PDFs and FFs) are either highly inclusive or exist are at very high scales, or both, it is perhaps not surprising that tension arises when these are evolved downward and used to make predictions at lower scales and for highly differential observables. Indeed, there have been few tests that Q^2-scaling, a hallmark of the collinear perturbative regime, actually holds to a reasonable approximation in SIDIS measurements at moderate Q^2. Our hope is that the new combined fit, which we refer to as "JAM20-SIDIS," will help to shed light on this and similar issues in the future.

In Sec. II we begin the discussion by summarizing the methodology used in our simultaneous Monte Carlo analysis, including the parametrizations used for the distributions and the multistep Bayesian inference algorithm. Details of the datasets included in the fit are summarized in Sec. III, while in Sec. IV we discuss the criteria for universality and how these are met in this analysis. A detailed discussion of the numerical results is

given in Sec. V, where we present the fitted PDFs and FFs, as well as detailed comparisons of data to theory. Finally, in Sec. VI we summarize our conclusions and discuss the implications of our analysis. Some formulas for SIDIS cross sections and structure functions are collected in the Appendix.

II. THEORETICAL FRAMEWORK

In this section we give an overview of the theoretical framework on which our analysis is based, including the observables to be fitted, the parametrizations used for the PDFs and FFs, details of the perturbative QCD setup, and Bayesian inference strategy employed.

A. Observables and factorization

In this analysis we work in standard collinear factorization [9–11], in which QCD cross sections are separated into perturbatively calculable partonic hard factors convoluted with nonperturbative PDFs and/or FFs. We perform calculations of all observables consistently to order α_s in the QCD coupling. Details of the basic theoretical setups for the inclusive DIS, inclusive Drell-Yan lepton-pair production and SIA reactions are provided in the literature [10,12] and will not be repeated here. However, since SIDIS is a comparatively novel addition to global QCD analyses, we review it in more detail in the Appendix.

The processes considered in the present analysis can be summarized as follows:

$$\ell + N \rightarrow \ell + X, \qquad \text{inclusive DIS},$$

$$\ell + N \rightarrow \ell + h^{\pm} + X, \qquad \text{semi-inclusive DIS},$$

$$N_1 + N_2 \rightarrow \ell^+ + \ell^- + X, \quad \text{Drell-Yan lepton-pair production},$$

$$\ell^+ + \ell^- \rightarrow h^{\pm} + X, \qquad \text{single-inclusive annihilation},$$

where $h^{\pm}$ represent charged pions, kaons, or unidentified hadrons, and the nucleon N (or $N_{1,2}$) in the initial state can be either a proton or a neutron (in practice, deuteron). Within the framework of collinear factorization, the cross sections for each of these processes can be written schematically as convolutions of hard functions and the nonperturbative parton distribution and fragmentation functions,

$$\frac{d\sigma_{\text{DIS}}}{dQ^2 dx_{\text{Bj}}} = \sum_i \mathcal{H}_i^{\text{DIS}} \otimes f_i \tag{1}$$

$$\frac{d\sigma_{\text{SIDIS}}}{dQ^2 dx_{\text{Bj}} dz_h} = \sum_{ij} \mathcal{H}_{ij}^{\text{SIDIS}} \otimes f_i \otimes D_j^h \tag{2}$$

$$\frac{d\sigma_{\text{DY}}}{dQ^2 dx_{\text{F}}} = \sum_{ij} \mathcal{H}_{ij}^{\text{DY}} \otimes f_i \otimes f_j \tag{3}$$

$$\frac{d\sigma_{\text{SIA}}}{dQ^2 dz_h} = \sum_j \mathcal{H}_j^{\text{SIA}} \otimes D_j^h \tag{4}$$

where the symbols $\otimes$ represent the convolution integrals in longitudinal momentum fractions of the hard scattering functions $\mathcal{H}_{ij}$ and the PDFs f_i and FFs D_j^h for parton flavors i, j (see the Appendix). In each process, Q represents the hard scale given by the photon virtuality, $Q \gg$ hadron masses, which allows the observables to be factorized into the short-distance perturbative and long-distance nonperturbative parts.

For the inclusive DIS and SIDIS processes,

$$x_{\text{Bj}} = \frac{Q^2}{2p \cdot q} \tag{5}$$

is the usual Bjorken scaling variable, while for the DY process the analogous scaling variables are defined as

$$x_1 = \frac{Q^2}{2p_1 \cdot q}, \qquad x_2 = \frac{Q^2}{2p_2 \cdot q}, \tag{6}$$

where p_1 and p_2 denote the incoming hadron momenta, with the Feynman scaling variable given by

$$x_{\text{F}} = x_1 - x_2. \tag{7}$$

In the DY center of mass frame, and in the limit of negligible hadron masses ($\ll Q$), the virtual photon rapidity can be written in terms of x_1 and x_2 as

$$y = \frac{1}{2} \ln \frac{x_1}{x_2}. \tag{8}$$

For the processes involving fragmentation to a hadron h in the final state, we have

$$z_h = \frac{p_h \cdot p}{q \cdot p} \qquad [\text{SIDIS}] \tag{9}$$

for SIDIS in Eq. (2), while

$$z_h = \frac{2p_h \cdot q}{Q^2} \qquad [\text{SIA}] \tag{10}$$

for SIA in Eq. (4).

B. Perturbative QCD and numerical setups

For our numerical analysis we make use of Mellin space techniques to enable fast evaluations of observables needed for the Bayesian analysis. In particular, we solve the DGLAP evolution equations analytically in Mellin space [13], which allows one to effectively render high-dimensional momentum space convolutions from process-specific factorization theorems, along with

the integrals in the DGLAP equations, in the form of lower-dimensional inverse Mellin transforms. For example, for the inclusive DIS observables one can write schematically,

$$\frac{d\sigma_{\text{DIS}}}{dQ^2 dx_{\text{Bj}}} = \sum_i \frac{1}{2\pi i} \int dN x_{\text{Bj}}^{-N} \widetilde{\mathcal{H}}_i^{\text{DIS}}(N, \mu) \\ \times U_{ij}^{\text{S}}(N, \mu, \mu_0) \tilde{f}_j(N, \mu_0), \tag{11}$$

where N here is the conjugate variable to x_{Bj}, $\tilde{f}_j(N, \mu_0)$ is the Mellin moment of the PDF $f_j(x, \mu_0)$, defined by

$$\tilde{f}_j(N, \mu_0) = \int_0^1 dx x^{N-1} f_j(x, \mu_0), \tag{12}$$

and $\widetilde{\mathcal{H}}_i^{\text{DIS}}(N, \mu)$ is the corresponding moment of the partonic DIS cross section. The analytic solution for the DGLAP evolution is entirely encoded in the evolution matrix $U_{i,j}^{\text{S}}$ that evolves the moments $\tilde{f}_j(N, \mu_0)$ of the PDFs from a given input scale μ_0 to the relevant DIS hard scale $\mu = Q$. A similar expression can be written for the SIA cross section,

$$\frac{d\sigma_{\text{SIA}}}{dQ^2 dz_h} = \sum_{ij} \frac{1}{2\pi i} \int dM z_h^{-M} \widetilde{\mathcal{H}}_i^{\text{SIA}}(M, \mu) \\ \times U_{ij}^{\text{T}}(M, \mu, \mu_0) \widetilde{D}_j^h(M, \mu_0), \tag{13}$$

where M is the Mellin conjugate variable for z_h, $\widetilde{D}_j^h(M, \mu_0)$ is the moment of the FF, and $\widetilde{\mathcal{H}}_i^{\text{SIA}}$ is the moment of the partonic SIA cross section. The superscripts S and T in the evolution matrix distinguish between the spacelike and timelike evolution for the PDFs and FFs, respectively, which are encoded in the corresponding DGLAP splitting kernels.

The same procedure can be extended for the case of SIDIS, which gives

$$\frac{d\sigma_{\text{SIDIS}}}{dQ^2 dx_{\text{Bj}} dz_h} = \sum_{ijkl} \frac{1}{(2\pi i)^2} \int dN x_{\text{Bj}}^{-N} \\ \times \int dM z_h^{-M} \widetilde{\mathcal{H}}_{ik}^{\text{SIDIS}}(N, M, \mu) \\ \times U_{ij}^{\text{S}}(N, \mu, \mu_0) \tilde{f}_j(\mu_0) U_{kl}^{\text{T}}(M, \mu, \mu_0) \\ \times \tilde{D}_j^h(M, \mu_0). \tag{14}$$

For the case of the Drell-Yan process, a special treatment is required since the Mellin moments for the partonic cross sections are not known. For this we employ the strategy developed by Stratmann and Vogelsang [14], where by the Mellin moments are numerically precalculated and used as

lookup tables during the analysis. The resulting expression can be written schematically as

$$
\frac{d\sigma_{DY}}{dQ^2 dx_F} = \sum_{ijkl} \frac{1}{(2\pi i)^2} \int dN \int dM \widetilde{\mathcal{H}}_{ik}^{DY}(N,M,\mu)
$$
$$
\times U_{ij}^S(N,\mu,\mu_0)\tilde{f}_j(\mu_0) U_{kl}^S(M,\mu,\mu_0)\tilde{f}_l(\mu_0),
$$
(15)

where the relevant inverse Mellin factors x_1^{-N} and x_2^{-M} arising from the scaling variables x_1 and x_2 for the incident nucleons N_1 and N_2, respectively, in Eq. (3) are integrated numerically with the hard scattering cross section and contained inside $\widetilde{\mathcal{H}}_{ik}^{DY}(N,M,\mu)$.

The analytic solutions for the evolution matrices are computed at next-to-leading logarithmic accuracy using splitting kernels up to $\mathcal{O}(\alpha_s^2)$ and the truncated solution for the single evolution operators (see Ref. [13] for details). We employ the zero-mass variable flavor scheme for solving the DGLAP evolution equations, setting the input scale for the PDFs and FFs at $\mu_0 = m_c$. The numerical values for the mass thresholds are taken from the PDG values in the $\overline{\text{MS}}$ scheme [15]: $m_c = 1.28$ GeV and $m_b = 4.18$ GeV. The strong coupling is evolved numerically using the QCD beta function up to $\mathcal{O}(\alpha_s^2)$, using the boundary condition $\alpha_s(M_Z) = 0.118$ at the Z boson mass, $M_Z = 91.18$ GeV. Finally, all the process specific hard coefficients are computed at fixed next-to-leading order in pQCD, which are available in the literature [14,16–18].

C. Parametrization of nonperturbative functions

For the nonperturbative parton distribution and fragmentation functions we use standard parametrizations that have been utilized in the literature. Namely, for the dependence on the parton momentum fraction x of the PDF $f(x)$ we use the template function,

$$
f(x,\mu_0) \to T(x;\boldsymbol{a}) = \mathcal{M} \frac{x^\alpha (1-x)^\beta (1+\gamma\sqrt{x}+\delta x)}{\int_0^1 dx\, x^{\alpha+1}(1-x)^\beta(1+\gamma\sqrt{x}+\delta x)},
$$
(16)

where $\boldsymbol{a} = \{\mathcal{M}, \alpha, \beta, \gamma, \delta\}$ is a vector containing the shape parameters (α, β, γ, and δ) and a normalization coefficient ($\mathcal{M}$) to be fitted. The integral in the denominator ensures that the value of the normalization coefficient $\mathcal{M}$ is equal to the second moment (x-weighted integral) of the function $T(x;\boldsymbol{a})$. For fitting the PDFs, we assume isospin symmetry to relate the PDFs in the neutron, $f_{i/n}(x)$, to those in the proton, $f_{i/p}(x) \equiv f_i(x)$, switching the $u \leftrightarrow d$ and $p \leftrightarrow n$ labels for the light quark flavors, and taking the PDFs for other flavors equal for the proton and neutron.

In practice, we parametrize the valence u and d quark distributions, $u_v \equiv f_u - f_{\bar{u}}$ and $d_v \equiv f_d - f_{\bar{d}}$, directly

using the template function [Eq. (16)]. The gluon distribution, $g \equiv f_g$, is also directly parametrized per Eq. (16). For the sea quark and antiquark distributions, we use five functions parametrized as in Eq. (16). These are a flavor symmetric sea function (S) that dominates at very low x and flavor specific functions [$q_0(\bar{q}_0)$] for the s, $\bar{u}$, $\bar{d}$, and $\bar{s}$ that take into account the possible nonperturbative origin of the sea. The distributions for s, $\bar{u}$, $\bar{d}$, and $\bar{s}$ are constructed from these according to: $q(\bar{q}) \equiv f_{q(\bar{q})} = S + q_0(\bar{q}_0)$. Note that s and $\bar{s}$ are parametrized separately because their contributions to the K^+ and K^- SIDIS cross sections differ. We do not fit the charm and bottom PDFs, and their contributions are generated purely from the DGLAP evolution. In total there are eight parametrized PDF functions being fitted. For the valence quark PDFs u_v and d_v and the nonperturbative sea components $\bar{u}_0$ and $\bar{d}_0$, we use the four shape parameters as in Eq. (16); for all other distributions we set the γ and δ parameters to zero. This gives 24 free shape parameters and 8 free normalization parameters. The number of free parameters is further reduced by valence number sum rules, which constrain the normalization parameters $\mathcal{M}$ for the u_v, d_v, and $s - \bar{s}$ distributions, whose lowest moments are required to be 2, 1, and zero, respectively. The normalization for the gluon PDF is determined using the momentum sum rule. With these constraints, there is a total of 28 free parameters for the PDFs.

For the z dependence of FFs, the functional form follows a similar template,

$$
D(z,\mu_0) \to T(z;\boldsymbol{a}) = \mathcal{M} \frac{z^\alpha(1-z)^\beta(1+\gamma\sqrt{z}+\delta z)}{\int_0^1 dz\, z^{\alpha+1}(1-z)^\beta(1+\gamma\sqrt{z}+\delta z)},
$$
(17)

where again the integral in the denominator ensures that the coefficient $\mathcal{M}$ corresponds to the second moment (z-weighted integral) of the function. In addition to the fragmentation to pions and kaons studied in earlier JAM analyses of SIA and SIDIS data [2,19], here we consider also the inclusive production of unidentified charged hadrons, $h^\pm$. Accounting for unidentified hadrons can be implemented in two ways. First, the hadron FFs can be fit independently from those for pions and kaons, as preferred by the NNPDF Collaboration [20]. Alternatively, one can take advantage of existing knowledge of specified hadron FFs and add a fitted residual correction to their sum. Such an approach was adopted by de Florian et al. (DSS) [21], for example, in which a residual correction was fitted to the sum of previously obtained pion, kaon, and proton fragmentation functions.

In our analysis we follow the latter approach, but include only the pion and kaon FFs, so that the residual term $D_i^{\text{res}^+}$ parametrizes the difference between the total hadron FF $D_i^{h^+}$ and the $D_i^{\pi^+}$ and $D_i^{K^+}$ functions,

$$D_i^{h^+} = D_i^{\pi^+} + D_i^{K^+} + D_i^{\text{res}^+}. \tag{18}$$

To reduce the total number of residual FFs being fit, we assume SU(3) flavor symmetry for light quarks and antiquarks,

$$D_q^{\text{res}^+} = D_u^{\text{res}^+} = D_d^{\text{res}^+} = D_s^{\text{res}^+}, \tag{19a}$$

$$D_{\bar{q}}^{\text{res}^+} = D_{\bar{u}}^{\text{res}^+} = D_{\bar{d}}^{\text{res}^+} = D_{\bar{s}}^{\text{res}^+}, \tag{19b}$$

where $D_q^{\text{res}^+}$ and $D_{\bar{q}}^{\text{res}^+}$ are parametrized per the template [Eq. (17)]. To allow for differentiation between the residual FFs for light quarks and antiquarks, we leave $\mathcal{M}$ and β for $D_{\bar{q}}^{\text{res}^+}$ as free parameters but fix α, γ, and δ to be the same as for $D_q^{\text{res}^+}$. This achieves a similar constraint on the parameters as the condition used by DSS [21], $2D_{\bar{q}}^{\text{res}^+} = (1 - z)^{\beta'} D_{q+\bar{q}}^{\text{res}^+}$.

For the pion FFs, $D_i^{\pi^+}$, we reduce the number of fitted functions by grouping the light quarks into "favored" (valence) and "unfavored" (nonvalence) flavors,

$$D_{\text{fav}}^{\pi^+} = D_u^{\pi^+} = D_{\bar{d}}^{\pi^+}, \tag{20a}$$

$$D_{\text{unf}}^{\pi^+} = D_d^{\pi^+} = D_s^{\pi^+} = D_{\bar{u}}^{\pi^+} = D_{\bar{s}}^{\pi^+}, \tag{20b}$$

where $D_{\text{fav}}^{\pi^+}$ and $D_{\text{unf}}^{\pi^+}$ are parametrized as in Eq. (17). For the parameters of the kaon FFs, $D_i^{K^+}$, we equate the "unfavored" flavors,

$$D_{\text{unf}}^{K^+} = D_d^{\pi^+} = D_s^{\pi^+} = D_{\bar{u}}^{\pi^+} = D_{\bar{d}}^{\pi^+}, \tag{21}$$

but leave the favored FFs $D_u^{K^+}$ and $D_{\bar{s}}^{K^+}$ independent. Here $D_{\text{unf}}^{K^+}$, $D_u^{K^+}$, and $D_{\bar{s}}^{K^+}$ are parametrized per Eq. (17). For the heavier flavors, we assume the charm and bottom quark and antiquark FFs to be equivalent, $D_c^{h^+} = D_{\bar{c}}^{h^+}$ and $D_b^{h^+} = D_{\bar{b}}^{h^+}$ for $h = \pi, K, \text{res}$, with $D_c^{h^+}$ and $D_b^{h^+}$ parametrized per Eq. (17). Finally, the gluon FFs $D_g^{h^+}$ for $h = \pi, K, \text{res}$ are also parametrized according to Eq. (17). We use charge conjugation symmetry to relate FFs for opposite charges by

$$D_q^{h^+} = D_{\bar{q}}^{h^-}, \tag{22}$$

where $h = \pi, K, \text{res}$. This results in five fitted functions for pions and residual hadrons, and six for kaons.

At this point, there are 17 shape parameters and 5 normalization parameters for residual hadrons, 20 shape parameters, and 5 normalization parameters for pions, and 24 shape parameters and 6 normalization parameters for kaons. The number of shape parameters is reduced further because throughout the fitting procedure, the parameters γ and δ for the gluon, charm, and bottom FFs are fixed at zero. In the end there are 16 free parameters to be fitted for residual charged hadron FFs, 19 free FF pion parameters,

and 24 free parameters for the kaon FFs. Together with the 28 PDF parameters, we have a total of 87 free parameters for the fitted functions. In addition, there are also 42 free parameters associated with normalization of various data sets, making for a total of 129 free parameters to be fitted in the analysis.

D. Bayesian inference

Our methodology for extracting nonperturbative PDFs and FFs is based on the general premise of Bayesian inference. Namely, we use Bayes' theorem to define a multivariate probability distribution $\mathcal{P}$ for the shape parameters characterizing the PDFs and FFs (the posterior) at a given input scale μ_0,

$$\mathcal{P}(\boldsymbol{a}|\text{data}) \sim \mathcal{L}(\boldsymbol{a}, \text{data})\pi(\boldsymbol{a}), \tag{23}$$

where $\mathcal{L}$ is a standard Gaussian likelihood function,

$$\mathcal{L}(\boldsymbol{a}, \text{data}) = \exp\left(-\frac{1}{2}\chi^2(\boldsymbol{a}, \text{data})\right), \tag{24}$$

with the χ^2 function defined by

$$\chi^2(\boldsymbol{a}) = \sum_{i,e}\left(\frac{d_{i,e} - \sum_k r_e^k \beta_{i,e}^k - T_{i,e}(\boldsymbol{a})/N_e}{\alpha_{i,e}}\right)^2 + \sum_k (r_e^k)^2 + \left(\frac{1 - N_e}{\delta N_e}\right)^2. \tag{25}$$

Here, $d_{i,e}$ is the value of the ith data point for the experimental dataset e, with $T_{i,e}$ the theoretical prediction for the data point; $\alpha_{i,e}$ is the uncorrelated systematic and statistical uncertainty for each data point added in quadrature; $\beta_{i,e}^k$ is the kth source of point-to-point correlated systematic uncertainties for the ith bin of dataset e, with r_e^k the related weight; and N_e and δN_e are the normalization and normalization uncertainty for each data set, respectively. In Eq. (23), $\pi(\boldsymbol{a})$ is the prior distribution for the set of parameters $\boldsymbol{a}$, which is used as input for a given fit to the data.

In principle, given the Bayesian posterior distribution, one can estimate confidence regions for a generic observable $\mathcal{O}$ (such as a PDF or a function of PDFs or FFs) by integrating over an d-dimensional parameter space,

$$\text{E}[\mathcal{O}] = \int d^d \boldsymbol{a} \mathcal{P}(\boldsymbol{a}|\text{data})\mathcal{O}(\boldsymbol{a}), \tag{26a}$$

$$\text{V}[\mathcal{O}] = \int d^d \boldsymbol{a} \mathcal{P}(\boldsymbol{a}|\text{data})(\mathcal{O}(\boldsymbol{a}) - \text{E}[\mathcal{O}])^2, \tag{26b}$$

where E and V are the expectation value and variance of the observable $\mathcal{O}$, respectively. Due to the significant numerical expense of evaluating the likelihood function, the explicit

usage of Eqs. (26) is often not practical. Instead, a more efficient option is to build Monte Carlo parameter samples $\{a_k; k = 1, ..., n\}$, which contain all parameters, including the N_e from Eq. (25), that are faithfully distributed according to the posterior distribution. These can in turn be used to evaluate the integrals in Eqs. (26) as Monte Carlo sums,

$$E[\mathcal{O}] = \frac{1}{n} \sum_{k=1}^{n} \mathcal{O}(a_k), \tag{27a}$$

$$V[\mathcal{O}] = \frac{1}{n} \sum_{k=1}^{n} (\mathcal{O}(a_k) - E[\mathcal{O}])^2. \tag{27b}$$

Our Monte Carlo sampling strategy is based on data resampling methodology, whereby multiple maximum likelihood optimizations are carried out. Each optimization consists of taking a random point in parameter space and fitting the parameters to data that have been distorted away from the central values by Gaussian shifts within the quoted uncertainties. To build the Monte Carlo samples, we use the multistep strategy developed in Ref. [2], where the PDF and FF parameters are preoptimized to minimize evaluating the likelihood in parameter regions that are strongly disfavored. To that end we start by first considering PDF and FF parameters separately using flat priors, with the resulting samples from each type of hadron structure combined at a later stage to build new prior samples for the final runs. The workflow is illustrated in Fig. 1, where each step is represented as vertical arrows that accumulate additional experimental data from the previous step, with the posterior samples at each step becoming the priors for the subsequent step. This strategy allows the samples to become more

optimized and avoids unnecessary likelihood evaluations in regions of parameters space by disfavoring those regions in earlier stages of the multistep chain.

III. DATA SETS

The data sets used in the present analysis include the primary electromagnetic processes that traditionally have been used in global QCD analyses, namely, inclusive DIS, Drell-Yan lepton-pair production (which constrain PDFs), SIA (which constrains FFs), and SIDIS (which constrains both PDFs and FFs). The inclusive DIS data are measurements of the $F_2(x_{Bj}, Q^2)$ structure function performed by the BCDMS [22,23] and New Muon Collaborations [24,25] at CERN, and from experiments at SLAC [26], as well as from reduced electron and positron cross sections from the H1 and ZEUS Collaborations [27] at DESY. These include both proton [22,24,26] and deuteron [23,25,26] targets, and with both neutral and charged current probes [27]. For the kinematics we implement cuts of $W^2 > 10$ GeV^2 and $Q^2 > m_c^2$, where $W^2 = M^2 + Q^2(1 - x_{Bj})/x_{Bj}$, in order to select DIS data that can be fitted within leading power factorization.

For Drell-Yan lepton-pair production data we use differential cross section measurements $d^2\sigma_{DY}/dQdx_F$ by the E866/NuSea Collaboration [28–30] at Fermilab, which include proton scattering from proton and deuteron targets. We include data in the range $Q^2 > 36$ GeV^2. Excluding lower Q^2 data is recommended by Ref. [31], which demonstrated that inclusion of the lower Q^2 data results in deteriorated prediction quality with no reduction in uncertainty when compared with fits to DIS data alone.

All SIA measurements are of the normalized differential cross sections $(d\sigma_{SIA}/dz_h)/\sigma_{tot}$ for the reaction
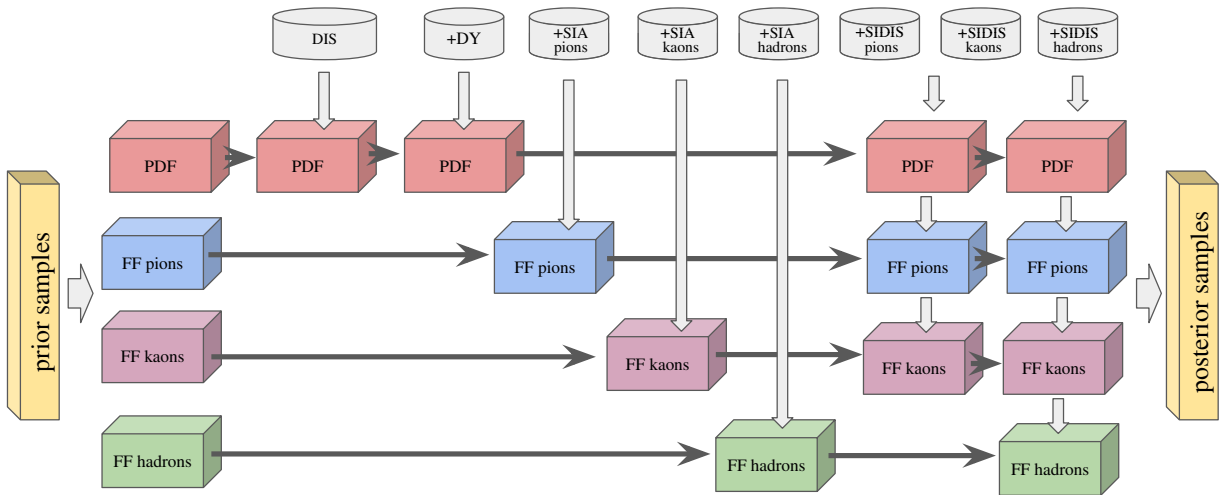

FIG. 1. Schematic illustration of the multistep workflow employed in our simultaneous Monte Carlo analysis. Each box represents a collection of Monte Carlo samples associated with a specific nonperturbative hadronic structure (PDFs, FFs). The vertical arrows indicate the inclusion of additional datasets from which new optimized Monte Carlo samples (posteriors) are generated as input (priors) for the next step.
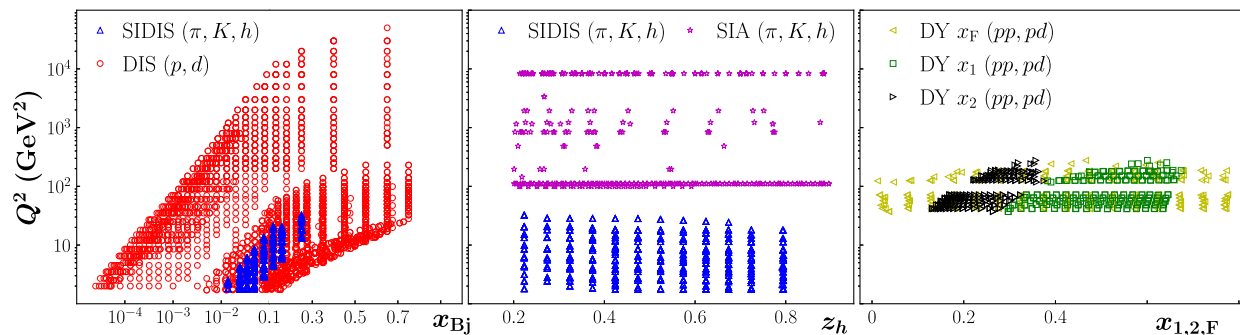
FIG. 2. Kinematic coverage of data used in this analysis, with Q^2 versus the Bjorken scaling variable $x_{\rm Bj}$ for inclusive DIS [22–27] and SIDIS data [50,51] (left panel), fragmentation variable z for SIDIS and SIA data [32–43,45–49] (central panel), and momentum fractions x_1, x_2, x_F for Drell-Yan data [28–30] (right panel).

$e^+e^- \to (\pi^\pm, K^\pm, h^\pm)X$. The data are from experiments performed by the TASSO [32–34] and ARGUS [35] Collaborations at DESY, by the TPC [36–39], HRS [40], SLD [41] and *BABAR* [42] Collaborations at SLAC, by the OPAL [43,44], ALEPH [45] and DELPHI [46] Collaborations at CERN, and by the TOPAZ [47] and Belle [48,49] Collaborations at KEK. As shown in Fig. 2, the SIA data cover the large-Q^2 region where a leading power description in terms of FFs should be accurate. Approximately half of the SIA data points have $Q \approx M_Z$, while the Belle and *BABAR* B factories have lower $Q \approx 10.5$ GeV. To ensure applicability of the leading power formalism, the SIA data in our fits are restricted to the range $0.2 < z_h < 0.9$.

Identification of heavy quark flavors for some of the SIA datasets is achieved through measurement of the total energy and momentum in secondary vertices. The flavor tagged cross sections for a specific flavor $q = c$ or b are particularly sensitive to the D_q^h, $D_{\bar{q}}^h$ and D_g^h fragmentation functions into the observed hadron h. In general, however, care needs to be taken with the precise method for separating primary quark flavors, and there are ongoing discussions regarding the optimal approach to this. For more in-depth discussion see, for example, Ref. [19].

Finally, the critical addition in this work compared with the previous JAM19 analysis [2] is the inclusion of unidentified charged hadron data, along with charged pions and kaons, in the SIDIS off deuterium targets from the COMPASS Collaboration [50,51] at CERN. Since the SIDIS data $d\sigma_{\rm SIDIS}^{h^\pm}/dQ^2 dx_{\rm Bj} dz_h$ are differential in $x_{\rm Bj}$ and z_h, they combine information on both PDFs and FFs, which appear in the description of SIA, Drell-Yan, and DIS data. Furthermore, as illustrated in Fig. 2, the SIDIS data have significant overlap in $x_{\rm Bj}$ and z_h with the $x_{\rm Bj}$ and x_F range of inclusive DIS and Drell-Yan data, respectively, and the z_h range of SIA data, so that the combined analysis constitutes a genuine test of their universality. For the COMPASS SIDIS data we use the same kinematic cuts on W^2 and Q^2 as for inclusive DIS and

restrict the fragmentation variable to $0.2 < z_h < 0.8$ in order to exclude data from the target fragmentation region and avoid large-z threshold corrections.

IV. ASSESSING UNIVERSALITY

Before proceeding to the results of our numerical analysis, we briefly discuss the criteria for universality of the PDFs and FFs and how these are implemented in our analysis. Extracting parton correlation functions, and using the extractions to test models of parton structure, is a nontrivial inverse problem, the detailed examination of which is beyond the scope of the present paper. However, a claim that the success of a fit is a measure of the predictive power of the PDFs and FFs requires a number of basic minimal conditions to be met:

1. The system of unknown correlation functions must be overconstrained, by which we mean that the constraints on unknown correlations imposed by data (or other theoretical constraints such as sum rules) must be greater than the total number of functions involved.

2. Each unknown correlation function must appear at least twice within the set of factorization formulas relating the correlation functions to physical observables.

3. There must be reasonable kinematical overlap between the observables so that correlation functions can be compared within similar ranges of parton momentum fractions.

Using isospin invariance to relate the PDFs in the proton to those in the neutron, we have seven independent PDFs: $f_u, f_d, f_s, f_{\bar{u}}, f_{\bar{d}}, f_{\bar{s}}$ and f_g, with PDFs for heavy flavors generated perturbatively. For the FFs, there are five functions for π^+ production: $D_u^{\pi^+}$, $D_{\bar{u}}^{\pi^+}$, $D_c^{\pi^+}$, $D_b^{\pi^+}$ and $D_g^{\pi^+}$, assuming that for equal u and d quark masses we can equate $D_{\bar{d}}^{\pi^+} = D_u^{\pi^+}$. Charge symmetry allows all the FFs for π^- production to be related to those for π^+ production. For K^+ production, there are six independent FFs: $D_u^{K^+}$,

$D_{\bar{s}}^{K^+}$, $D_{\bar{u}}^{K^+}$, $D_c^{K^+}$, $D_b^{K^+}$ and $D_g^{K^+}$, where we differentiate between the u and $\bar{s}$ functions. Again, using charge symmetry the FFs for K^- can be obtained from these six K^+ FFs. Finally, for the unidentified charged hadrons $h^\pm$ or residual FFs, we identify five independent functions: $D_u^{\mathrm{res}+}$, $D_{\bar{u}}^{\mathrm{res}+}$, $D_c^{\mathrm{res}+}$, $D_b^{\mathrm{res}+}$ and $D_g^{\mathrm{res}+}$. This makes then a total of 23 functions to be determined.

The quark and gluon PDFs are constrained by their appearance in several sum rules; in particular, the number sum rules,

$$\int_0^1 \mathrm{d}x(f_q(x) - f_{\bar{q}}(x)) = n_q, \qquad (28)$$

where $n_u = 2$, $n_d = 1$ and $n_s = 0$, and the momentum sum rule,

$$\sum_{i=q,\bar{q},g} \int_0^1 \mathrm{d}x \, x f_i(x) = 1. \qquad (29)$$

Note that in Sec. II C these constraints were specifically used to fix the values of the normalization parameters for several fitted functions. However, for the purpose of assessing universality, they are simply counted as additional independent equations which include and thus constrain the PDFs.

The data sets discussed in Sec. III also constrain the light quark and gluon PDFs since they appear in expressions for multiple independent observables. Counting these and also the four sum rules (28) and (29),

$$f_i(x) \overset{i \neq c,b}{\to} \begin{cases} 6 & \mathrm{DIS} \\ 2 & \mathrm{Drell\text{-}Yan} \\ 6 & \mathrm{SIDIS} \\ 4 & \mathrm{sum\ rules} \end{cases}, \qquad (30)$$

there is a total of 18 relations between the light quark PDFs. The heavy quarks appear in an even greater number of observables. The light quark fragmentation functions appear in at least one SIA observable and, because of charge conjugation invariance, in two SIDIS observables,

$$D_i^{\pi^+}(z) \overset{i \neq c,b,g}{\to} \begin{cases} 1 & \mathrm{SIA} \\ 2 & \mathrm{SIDIS} \end{cases}, \qquad (31)$$

and similarly for the kaon and charged hadron fragmentation functions.

For a robust stress-test of universality, there should be reasonable overlap of the ranges in parton momentum fraction for both the PDFs and the FFs. An indication for how well this is achieved in the current fit can be be gleaned from the kinematical coverage plots shown in Fig. 2. To lowest order in α_s, the kinematical variables x_{Bj}, x_1, x_2 and z_h approximate the parton momentum fractions x and z,

respectively, while QCD evolution relates all values of Q^2. Figure 2 confirms that PDFs and FFs are both constrained by multiple processes in overlapping regions of momentum fractions.

In summary, our analysis does indeed fulfill the basic criterion for qualifying as a test of universality and retaining predictive power for the PDFs and FFs more generally. Note, however, that the momentum sum rule for FFs has not been imposed in the analysis. Instead, this will be used as a consistency check for the final fit in Sec. V.

V. NUMERICAL ANALYSIS

In this section we present the results of our simultaneous Monte Carlo analysis of PDFs and FFs. We begin with a survey of the fitted cross sections for the various global datasets used in this study, focusing especially on the quality of agreement with the SIDIS and SIA data on $\pi^\pm$ and $K^\pm$, as well as unidentified $h^\pm$ production. We then present our final fitted PDFs and FFs, and discuss the vital role played by the SIDIS and SIA datasets in particular in constraining the strange quark distribution in the proton.

A. Data and theory agreement

To assess the agreement of the fitted results with the various datasets, in Fig. 3 we show the reduced χ^2 for each individual experiment, which is defined by

$$\chi_{\mathrm{red}}^2 = \frac{1}{N} \sum_{i,e} \frac{1}{\alpha_{i,e}^2} \left(d_{i,e} - \mathrm{E}\left[\sum_k r_e^k \beta_{i,e}^k + T_{i,e}/N_e \right] \right)^2. \qquad (32)$$

Here, the expectation value $\mathrm{E}[...]$, as defined in Eq. (27a), represents the mean theory, including optimized multiplicative and additive corrections to match the data, with N the total number of data points. In Fig. 3 we show the mean and standard deviation of the Monte Carlo residuals for each experiment e, where the residual per data point is defined as

$$\mathrm{residual}(e,i) = \frac{1}{\alpha_{i,e}} \left(d_{i,e} - \mathrm{E}\left[\sum_k r_e^k \beta_{i,e}^k + T_{i,e}/N_e \right] \right). \qquad (33)$$

For the inclusive DIS, Drell-Yan and SIDIS datasets we find excellent overall agreement between data and theory, with χ_{red}^2 values close to 1. The χ_{red}^2 for the SIA datasets are slightly higher, but nonetheless the overall fit is very good, giving a total reduced $\chi_{\mathrm{red}}^2 = 1.26$ for almost 5000 data points. The values of χ_{red}^2 for each type of dataset and for each specific hadron in the final state are summarized in Table I, along with the number of data points for each dataset.
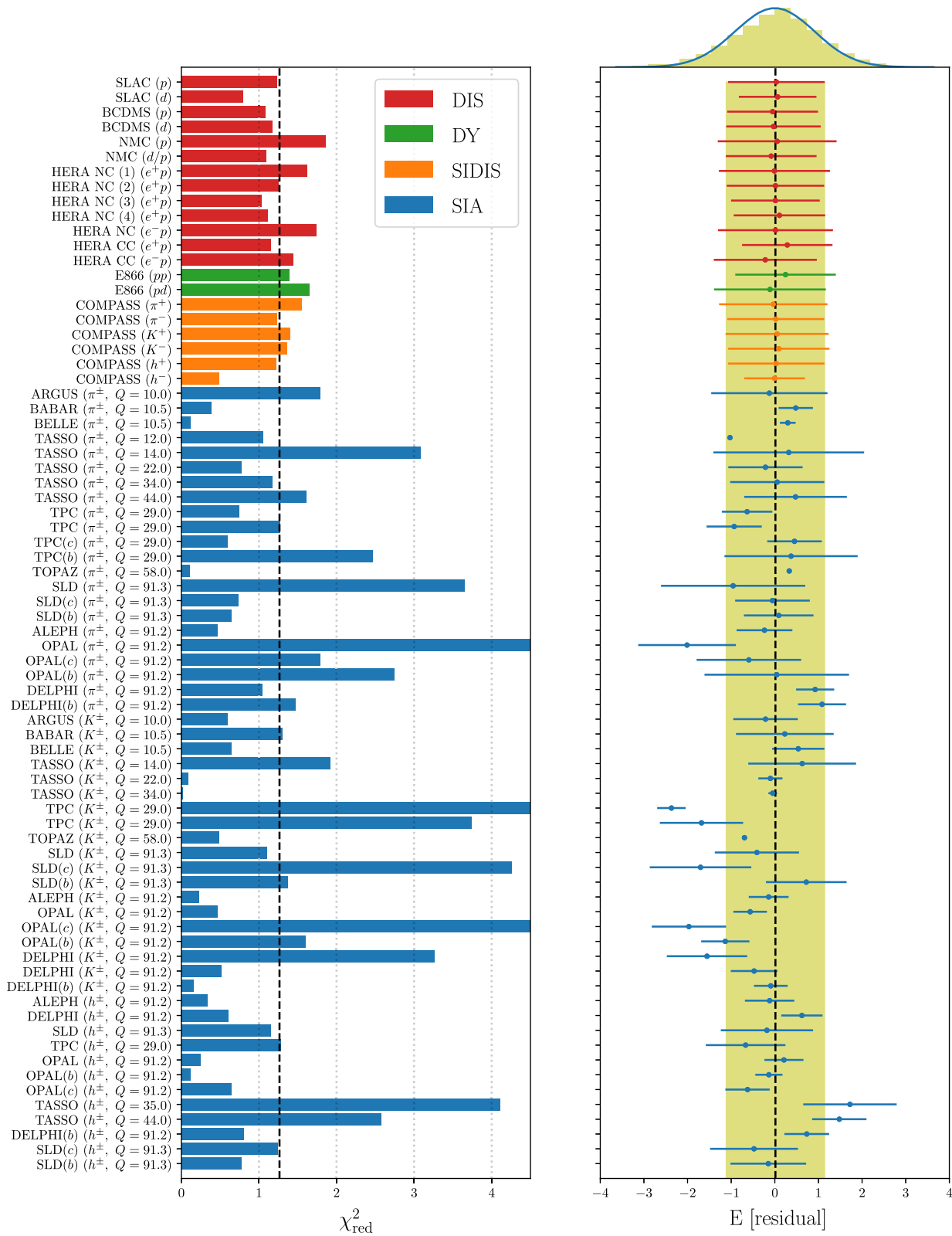
FIG. 3.   Reduced χ^2_{red} values for each DIS (red), DY (green), SIDIS (orange) and SIA (blue) experiment considered in this analysis (left column), along with the corresponding mean and standard deviation of the residuals for each experiment, E (right column).

TABLE I. Reduced χ^2_{red} values for each type of dataset (DIS, Drell-Yan, SIDIS, SIA) considered in this analysis, together with the number of data points N_{dat} for each dataset.

Reaction		χ^2_{red}	N_{dat}
DIS		1.29	2680
DY		1.52	250
SIDIS	$\pi^{\pm}$	1.39	498
	$K^{\pm}$	1.38	494
	$h^{\pm}$	0.85	498
SIA	$\pi^{\pm}$	1.09	231
	$K^{\pm}$	1.37	213
	$h^{\pm}$	1.15	120
Total		1.26	4984

The residuals profile for the DIS, Drell-Yan and SIDIS datasets is well centered around zero, with variances ~1, indicating an average Gaussian behavior of their associated likelihood function. The variance for the SIDIS h^- data from COMPASS, however, is found to be up to $\approx 50\%$ below unity, suggesting a deviation from a Gaussian likelihood. This may be due to the fact that these data are dominated by systematic uncertainties, which is also reflected by the relatively small reduced χ^2_{red} values, especially for the COMPASS h^- data relative to the rest of DIS and SIDIS datasets.

A more detailed comparison with the COMPASS SIDIS is made in Figs. 4–6, where we show the z_h dependence of the $\pi^{\pm}$, $K^{\pm}$ and $h^{\pm}$ multiplicities, respectively, which are defined as ratios of SIDIS to inclusive DIS cross sections at the same x_{Bj} and Q^2,

$$\frac{dM^h}{dz_h} = \frac{d\sigma^h_{\text{SIDIS}}/dQ^2 dx_{\text{Bj}} dz_h}{d\sigma_{\text{DIS}}/dQ^2 dx_{\text{Bj}}}. \quad (34)$$

The agreement between theory and the experimental z_h spectrum is quite remarkable, given that it spans some 2 orders of magnitude, which suggests that at these kinematics a leading power perturbative QCD factorization at next-to-leading order provides sufficient accuracy to describe the data. Interestingly, the differences between the multiplicities for positively and negatively charged hadron species increase with x_{Bj}, especially for kaons, and in the valence region these can differ by an order of magnitude for low values of Q^2. Such differences can enhance our ability to extract flavor dependent effects in nonperturbative PDFs and parton to kaon FFs from the data. The new dataset included for the first time in the present JAM analysis, namely the unidentified charged hadron data shown in Fig. 6, are also well described by our nonperturbative ansatz for the corresponding FFs. In contrast to the excellent agreement with the z_h dependence of the data in Figs. 4–6, we note that analysis of the same data differential in the hadron transverse momentum using existing PDFs and FFs within TMD factorization results in poor agreement between predictions and data [4,5], indicating that further work is needed to understand the SIDIS transverse momentum spectra.

For the SIA data sets, there is a somewhat wider spread in the data versus theory comparisons, as seen in Figs. 7–9 for the $\pi^{\pm}$, $K^{\pm}$ and unidentified charged hadron $h^{\pm}$ final states, respectively. Generally, the $\pi^{\pm}$ data have the best agreement among the SIA datasets, with a reduced $\chi^2_{\text{red}} = 1.09$, followed by the hadron data with $\chi^2_{\text{red}} = 1.15$, and lastly the kaon data, which have an overall reduced
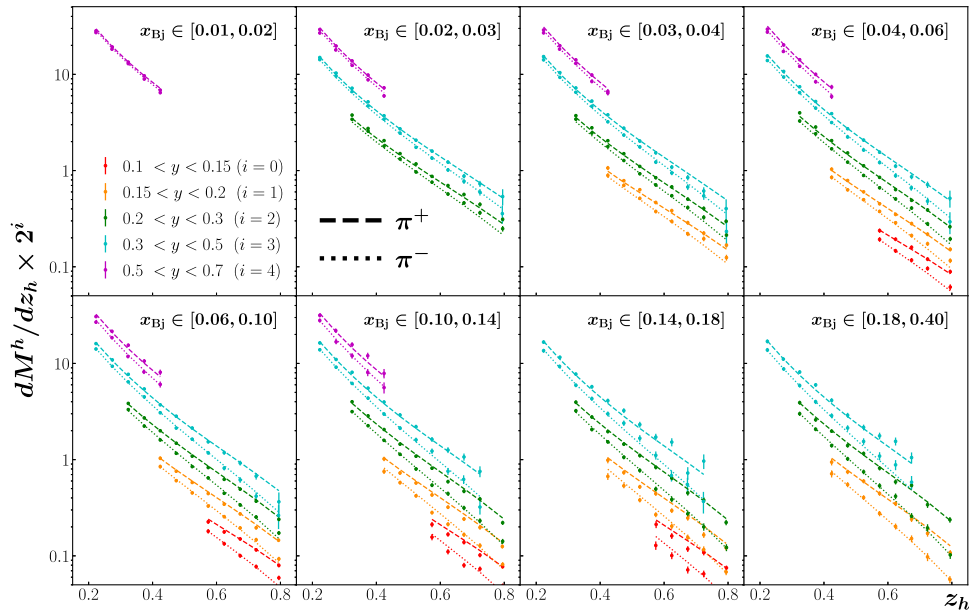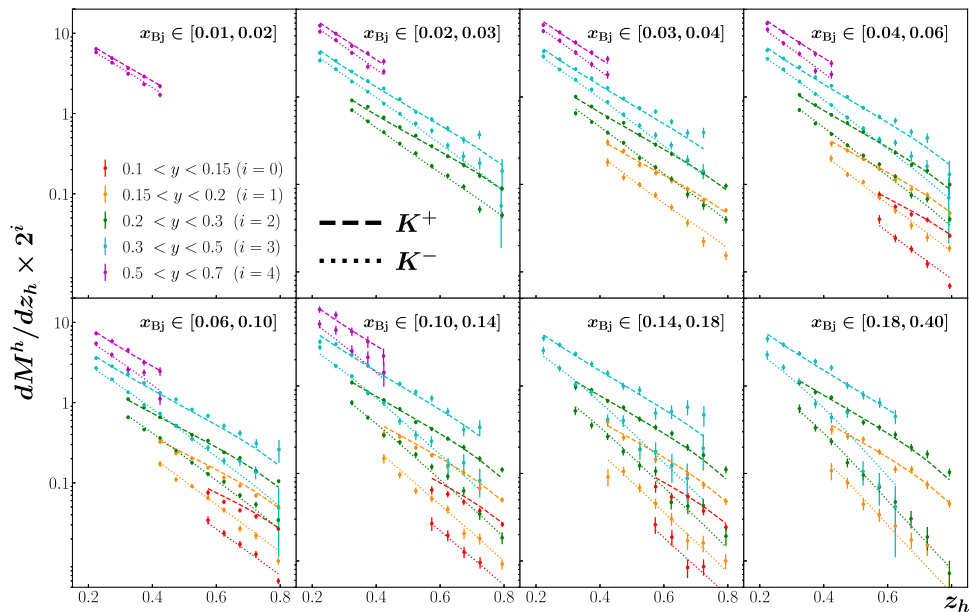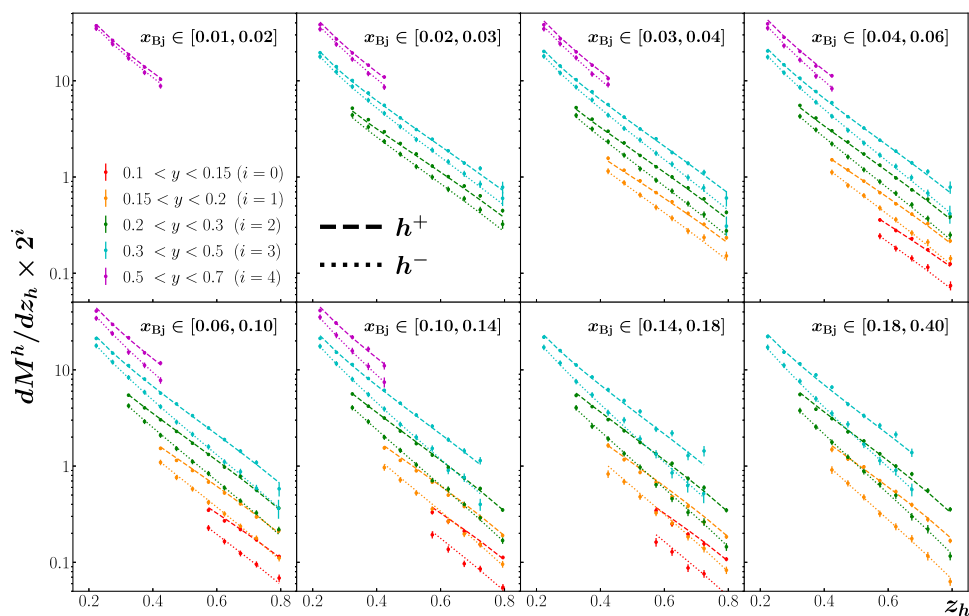

FIG. 4. Comparison of the multiplicities dM^h/dz_h for $h = \pi^+$ (dashed lines) and π^- (dotted lines) production with the COMPASS data [50,51] in various bins of x_{Bj} and y (offset by a factor 2^i).

FIG. 5. As in Fig. 4, but for $K^{\pm}$ COMPASS SIDIS data [50,51].

FIG. 6. As in Fig. 4, but for unidentified hadron $h^{\pm}$ COMPASS SIDIS data [50,51].

$\chi^2_{\rm red} = 1.37$. For about 3/4 of the $\approx$40 SIA datasets, we find very good agreement with the global fit, with $\chi^2_{\rm red} \approx 1$ or below. For the remaining datasets that have larger $\chi^2_{\rm red}$ values, to better understand the reasons for some of the tensions between data and theory we discuss in the following some individual cases ranked by the reduced $\chi^2_{\rm red}$ values.

Starting with the datasets that have the largest $\chi^2_{\rm red}$ values, namely, $\chi^2_{\rm red} \gtrsim 3$, we identify the OPAL ($\pi^{\pm}$ and $c \to K^{\pm}$), TPC ($K^{\pm}$), SLD ($\pi^{\pm}$ and $c \to K^{\pm}$), DELPHI ($K^{\pm}$), and TASSO ($\pi^{\pm}$ and $h^{\pm}$ at 35 GeV) datasets. For the inclusive

OPAL ($\pi^{\pm}$) data, we observe in Fig. 7 that for $z_h < 0.5$ the data are indeed in tension with the corresponding inclusive ALEPH and SLD results, and the overall trend of the data/ theory ratio suggests a possible normalization issue with this dataset. This can also be said for the DELPHI ($K^{\pm}$) which appears to have some tension with the corresponding inclusive OPAL and ALEPH results. Similarly, from Fig. 8 we find that the TPC ($K^{\pm}$) spectrum lies below the theory, suggesting again a normalization problem with these data. The situation for the TASSO ($\pi^{\pm}$) data is less clear, as only the $Q = 14$ GeV dataset seems to give a bad fit, while data
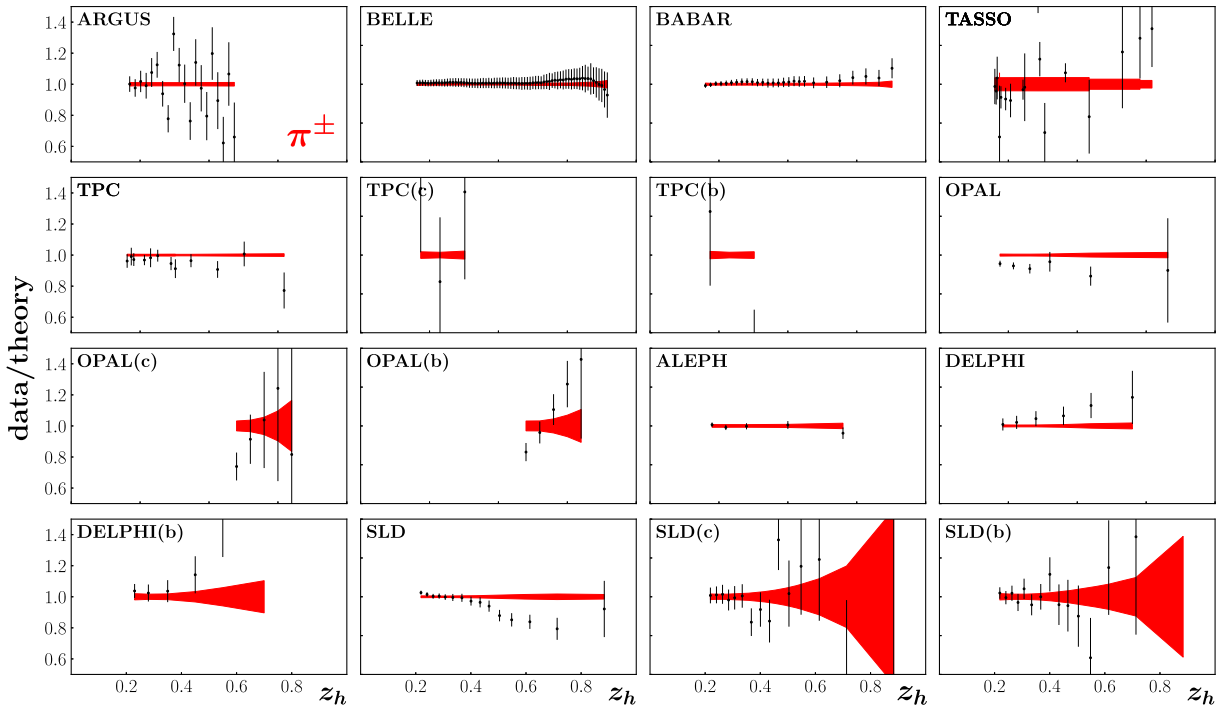
FIG. 7. Data to theory ratios for SIA $\pi^\pm$ production cross sections versus z_h, with the bands indicating the uncertainty on the fitted result.
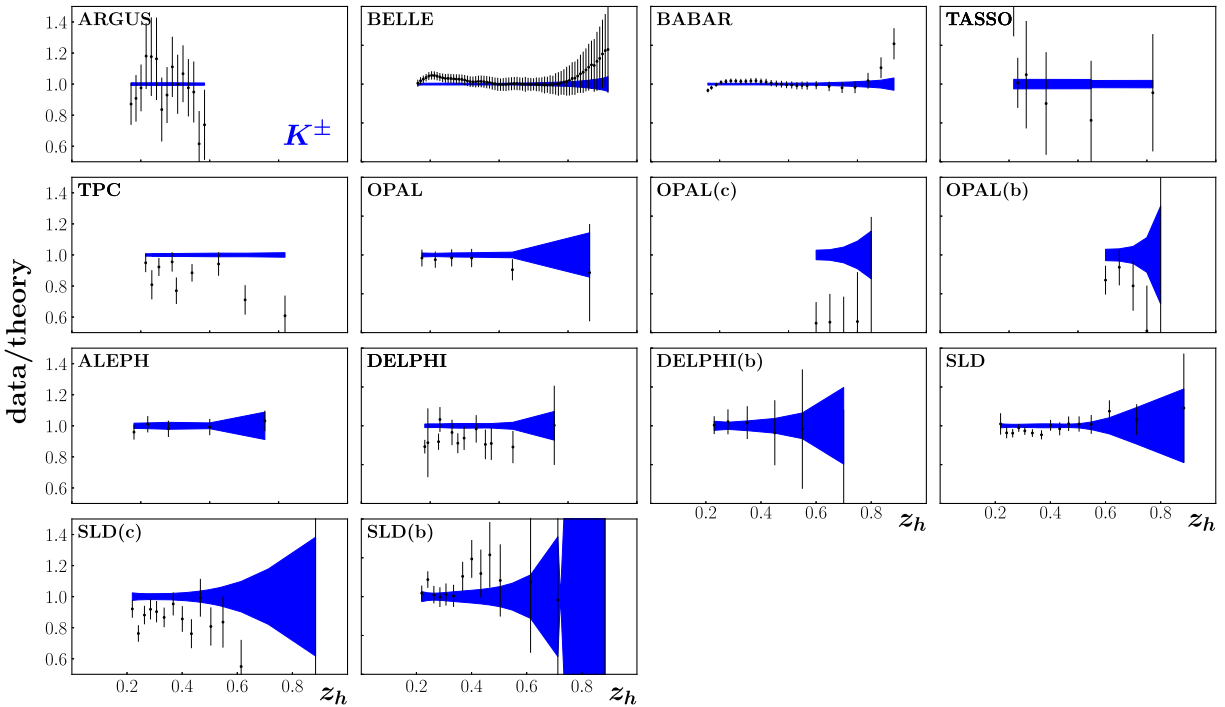

FIG. 8. As in Fig. 7, but for SIA $K^\pm$ production.

at other energies can be described fairly well. This again hints at a problem with the overall normalization for this dataset. The same behavior appears also in the TASSO ($h^\pm$) data in Fig. 9, where both the $Q = 35$ and 45 GeV datasets

are above the theoretical cross sections. The case of SLD and OPAL ($c \to K^\pm$) data in Fig. 8 shows a clear overestimation of the z_h spectra. While one can argue that this problem could be a reflection of the need for a more
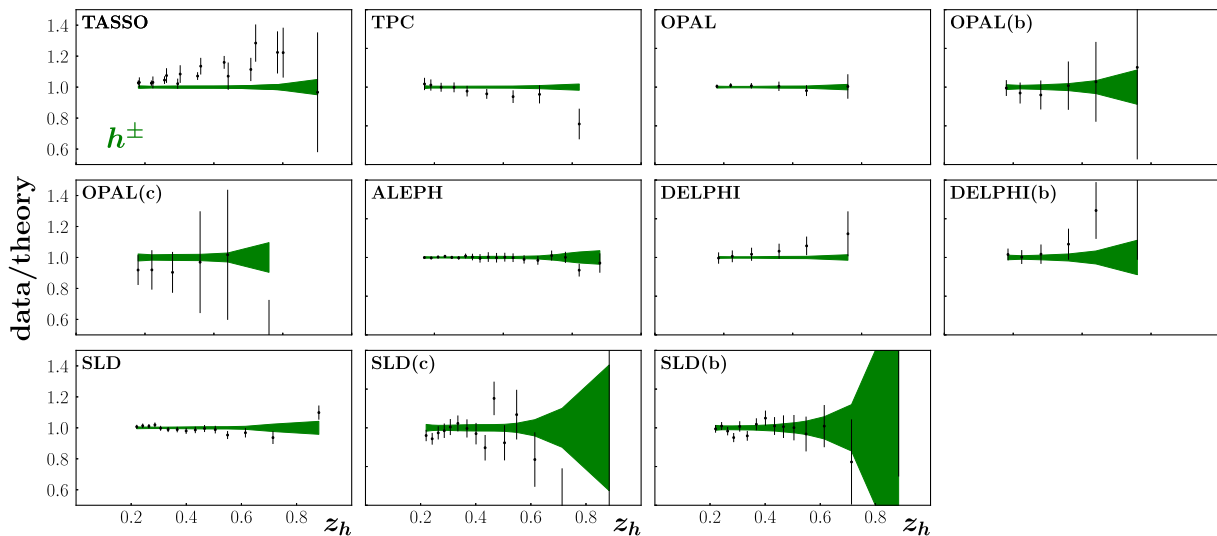
FIG. 9.   As in Fig. 7, but for SIA unidentified charged hadron $h^{\pm}$ production.

sophisticated heavy quark treatment in our theory, the description of b-tagged data from SLD, DELPHI and OPAL is relatively good, so that an explanation in terms of a normalization uncertainty in the SLD and OPAL $(c \rightarrow K^{\pm})$ data may be more relevant.

For SIA datasets that have smaller, but still large, χ^2_{red} values, $2 \lesssim \chi^2_{\mathrm{red}} \lesssim 3$, we identify the b-tagged TPC $(b \rightarrow \pi^{\pm})$, OPAL $(b \rightarrow \pi^{\pm})$, and TASSO ($h^{\pm}$ at 44 GeV). For the case of the TPC $(b \rightarrow \pi^{\pm})$ data, we see from Fig. 7 that for the largest z_h bin the theory overestimates the data. On the other hand, good agreement is found for the SLD $(b \rightarrow \pi^{\pm})$ data at the same kinematics. It is possible that at the smaller Q values of TPC relative to SLD, the range in z_h where leading power factorization is applicable is narrower, in particular for the b-tagged data. The z_h dependence of the OPAL $(b \rightarrow \pi^{\pm})$ data appear to be clearly different from the theory, even within the large uncertainties. We note here that the OPAL data are presented as truncated moments as a function of the lower limit of the integration, z_h^{min}, and the inclusion of the very high z_h bins may be problematic for the validity of factorization theorems at $z_h \rightarrow 1$. Lastly, as with TASSO ($h^{\pm}$ at 35 GeV), the somewhat large χ^2_{red} values for the 44 GeV data is likely attributable to a problem with overall normalization.

For datasets that have $\chi^2_{\mathrm{red}} \lesssim 2$, we consider the agreement to be generally acceptable. Indeed, the vast majority of datasets in this category have $\chi^2_{\mathrm{red}} \approx 1$ or below. These include all of the recent high-statistics B-factory data from *BABAR* ($\pi^{\pm}$, $K^{\pm}$) and Belle ($\pi^{\pm}$, $K^{\pm}$), most of the TASSO ($\pi^{\pm}$, $K^{\pm}$), TPC ($\pi^{\pm}$, $c \rightarrow \pi^{\pm}$) and SLD ($h^{\pm}$, $b \rightarrow \pi^{\pm}$, $b \rightarrow h^{\pm}$) datasets, all of the ALEPH ($\pi^{\pm}$, $K^{\pm}$, $h^{\pm}$) and most of the DELPHI ($\pi^{\pm}$, $K^{\pm}$, $b \rightarrow K^{\pm}$, $h^{\pm}$, $b \rightarrow h^{\pm}$) data, along with the older ARGUS ($K^{\pm}$), TOPAZ ($\pi^{\pm}$, $K^{\pm}$) and OPAL ($K^{\pm}$, $h^{\pm}$, $c \rightarrow h^{\pm}$, $b \rightarrow h^{\pm}$) data. Slightly higher, but

still reasonable, χ^2_{red} values are obtained for the ARGUS ($\pi^{\pm}$), TPC ($h^{\pm}$), DELPHI ($b \rightarrow \pi^{\pm}$), and SLD ($K^{\pm}$, $c \rightarrow \pi^{\pm}$, $c \rightarrow h^{\pm}$) datasets.

Finally, we note that most of the large χ^2_{red} values found in this analysis were absent in the previous JAM Monte Carlo analysis of fragmentation functions [19]. The main reason is the restriction of the SIA datasets here to the range $0.2 < z_h < 0.8$, chosen to coincide with the range over which the SIDIS data in this work are able to be described within collinear factorization. For the LEP data in particular there are many data points at $z_h < 0.2$ which can be well fitted within the current framework, and which would reduce the overall χ^2_{red}. A careful point by point comparison of the individual χ^2_{red} values for the various datasets indeed confirms that similar discrepancies also occurred in Ref. [19]. However, for consistency in our joint analysis of PDFs and FFs, we restrict the kinematic range to the region where both SIA and SIDIS can be simultaneously described. The same choice for the z_h range was made in the recent JAM19 analysis, which required SIDIS data to be restricted to $z_h \gtrsim 0.2$ to ensure separation of the target and current fragmentation regions.

B. Parton distributions and fragmentation functions

The proton PDFs from our simultaneous fit are displayed in Fig. 10 at a scale $\mu^2 = 10~\mathrm{GeV}^2$, where we focus on the kinematic region of parton momentum fractions $x \gtrsim 0.01$ that is constrained by the SIDIS data. For comparison, we contrast our results with other next-to-leading order PDF parametrizations, namely, from the CJ15 [52] and NNPDF3.1 [53] global analyses. Compared with the other fits, our valence u and d quark distributions have slightly larger magnitude in the intermediate-x region, $x \sim 0.1$,
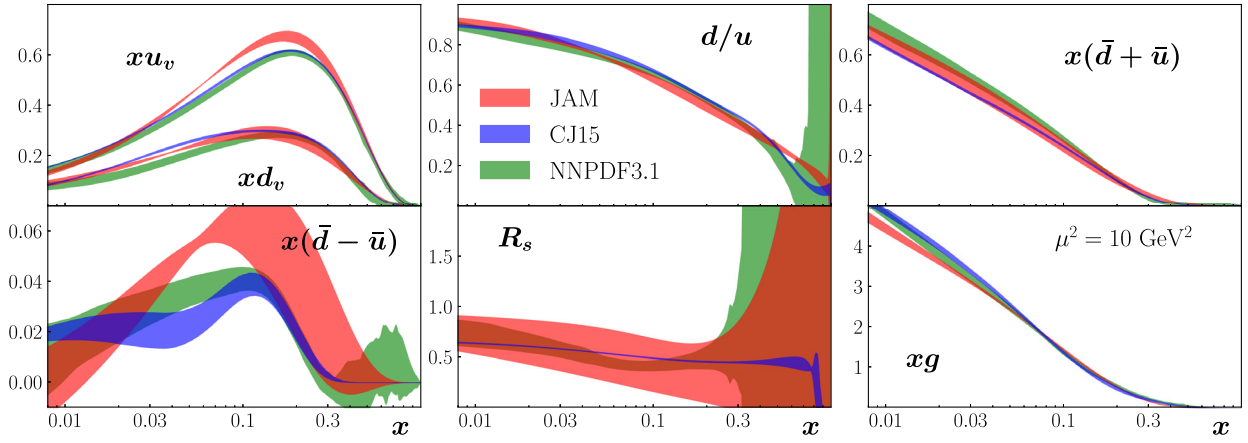
FIG. 10. Proton PDFs from the present JAM20-SIDIS analysis (red bands) versus x at a scale $\mu = 10$ GeV^2, compared with the CJ15 [52] (blue bands) and NNPDF3.1 [53] (green bands) parametrizations. The bands shown represent the mean $\pm 1\sigma$.

with a compensating stronger suppression at small x, needed to ensure that the valence number sum rules are respected. The ratio d/u is quite compatible with the results from the other groups, on the other hand, but has a significantly larger uncertainty at large x compared with the CJ15 result, reflecting the Monte Carlo nature of our analysis.

The intermediate-x enhancement in the valence PDFs in our fit is correlated with the slightly smaller $\bar{d} + \bar{u}$ light antiquark sea compared with the CJ15 and NNPDF3.1 parametrizations. This in turn is correlated with the behavior of the strange quark sea, as seen in the ratio,

$$R_s = \frac{s + \bar{s}}{\bar{d} + \bar{u}} \qquad (35)$$

of the strange to nonstrange sea quark PDFs. In Fig. 10 this ratio is generally larger in our analysis than for the other parametrizations, with a somewhat bigger uncertainty. This is understood from the fact that in the CJ15 fit R_s is fixed to be 0.4 at the input scale, with deviations from the constant value arising only from DGLAP evolution. For the NNPDF3.1 fit the uncertainties are smaller because of their inclusion of the neutrino DIS data, which we do not include in our analysis because of unknown nuclear corrections in neutrino scattering [54–56]. Our light antiquark asymmetry $\bar{d} - \bar{u}$ is also compatible with the other groups, but again with a larger uncertainty, which may be related to the absence of collider W and lepton asymmetry data in our fit. Finally, for the gluon distribution, the magnitude and uncertainties are very similar across all the analyses, even though our fit does not include jet production data from hadron colliders. This reflects the fact that the HERA DIS data, which are included here, provide strong constraints on the shape of the gluon PDF via scaling violations.

For the parton to hadron FFs, we show in Fig. 11 the z dependence of the FFs at a scale $\mu^2 = 100$ GeV^2 for the

positively charged π^+, K^+ and unidentified hadrons h^+, as well as for the residual hadrons δh^+, defined as the difference between h^+ and the sum of π^+ and K^+ (so that the total is given by $h^+ = \pi^+ + K^+ + \delta h^+$). For most of the flavors we find that the quark $\to \pi^+$ fragmentation dominates, as expected from the pion being the lightest hadron in the QCD spectrum. Exceptions to this are for $\bar{s} \to K^+$ and $c \to K^+$ at intermediate z values and for b quark fragmentation into residual hadrons δh^+.

For gluon fragmentation, pion production dominates for z up to $\sim$0.5–0.6, above which kaon fragmentation becomes as sizeable as the pion. This is consistent with the findings of previous FF analyses [19,57], which observed that the production of heavier particles such as kaons requires larger momentum fractions from the fragmenting gluon compared to the production of lighter particles.

The production of hadrons heavier than kaons, as indicated in Fig. 11 by the residual hadrons δh^+, can be sizable and comparable to that of kaons, especially for the d and s quarks and at large values of z. The relatively large $d \to \delta h^+$ FF can be understood in terms of the fragmentation into protons. Note that we have imposed flavor symmetry for the residual hadron fragmentation, so that $D_d^{\delta h^+} = D_s^{\delta h^+}$. In principle, the presence of hyperons such as Σ^+ should brake this relation, but we leave analysis of such effects for future work. As the case for the $g \to K^+$, the fragmentation of gluons into heavier particles peaks at large z, where larger momentum fractions from the fragmenting gluons are needed for the production of heavier particles.

For production of hadrons initiated by heavy quarks, we find similar fragmentation of charm quarks into pions and kaons, but a rather different pattern for the fragmentation of bottom quarks. Some of this difference can be explained by the flavor-changing properties of u-type quarks decaying into d-type quarks. While the charm quark can decay into strange quarks and hence enhance K^+ production, the same does not occur for bottom quarks, which suppresses kaon
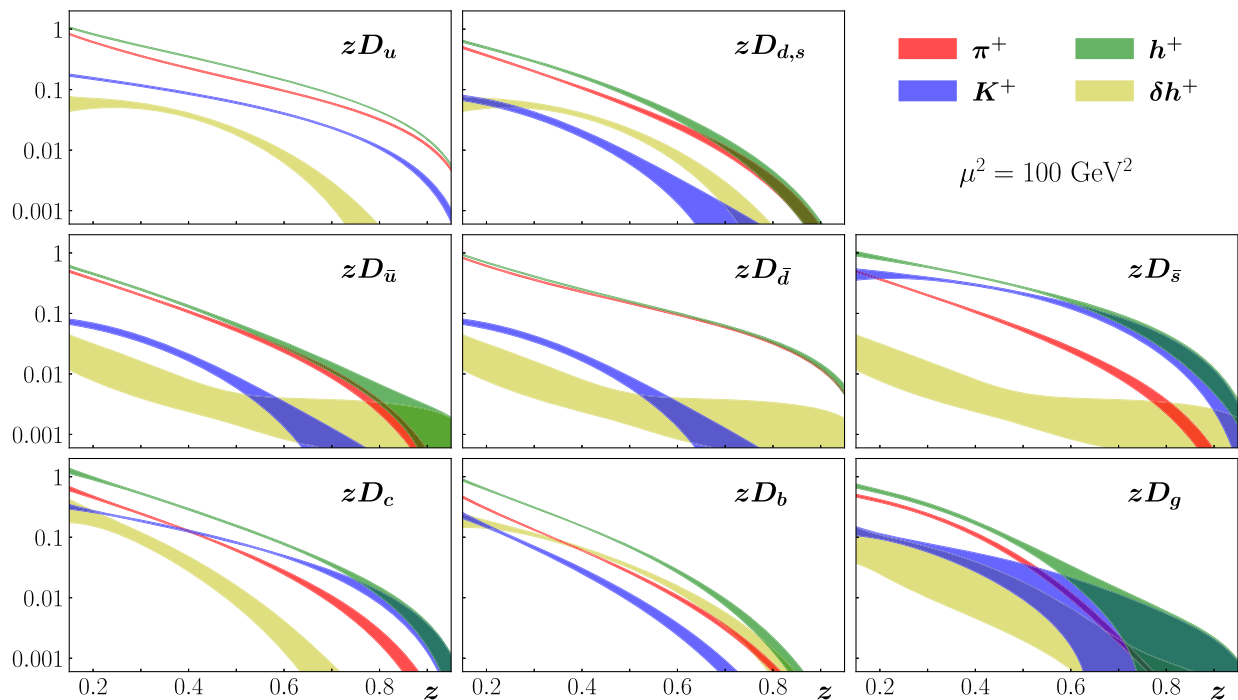
FIG. 11. Parton to hadron FFs versus z at $\mu^2 = 100$ GeV^2 from the JAM20-SIDIS analysis for various parton flavors fragmenting to π^+ (red bands), K^+ (blue bands), unidentified hadrons h^+ (green bands), and residual hadrons δh^+ (yellow bands), defined as the difference between h^+ and the sum of π^+ and K^+. The bands shown are mean $\pm 1\sigma$.

production relative to pion production due to the mass difference. Interestingly, the production of other species of charged hadrons is much larger for b quarks than for c quarks, which may be understood from the greater phase space available for b quarks to decay into heavier hadrons to which charm quarks cannot transition.

In Fig. 12 we present truncated moments,

$$\langle z \rangle_i^h = \int_{z_{\min}}^1 dz\, z D_i^h(z), \qquad (36)$$

for each flavor i and final state hadron h, where we take the lower limits on the z integration $z_{\min} = 0.2$ to restrict the moment to the region of SIDIS kinematics. The truncated moment indicate how energetic is the production different a hadron species h relative to the parent parton i. In general, we find that the production of hadron species heavier than pions and kaons is typically produced with lower energies, which is consistent with the physical picture whereby more energy is required to produce heavier hadrons than lighter hadrons.

As expected, the favored fragmentation of $\bar{d}$ quarks is predominantly into highly energetic pions, while for the antistrange $\bar{s}$ the production rate of energetic kaons is slightly higher than that of pions. The unfavored fragmentation of d, s and $\bar{u}$ quarks follows a similar pattern, with the lightest (pion) state produced at the highest energies

followed by kaons and other heavier charged hadrons. An exception to this behavior is for charm and bottom quark fragmentation: for c quarks kaons are produced with energies comparable to those of pions, while for b quarks kaon production is suppressed with heavier mass hadrons produced at similar energies as pions.

Interestingly, the production of hadrons from gluons follows the same pattern as for u-quark fragmentation. While the latter can be explained in terms of mass differences between the produced hadron species, the fact that u quarks and gluons give a similar average energy profile across hadron species is intriguing. On perturbative grounds one can argue that gluon fragmentation is enhanced because of the $C_A = 3$ factor in the gluon splitting function, P_{gg}, relative to quark splitting functions, P_{qq} and P_{gq}, which are proportional to $C_F = 4/3$. The absence of direct constraints on the gluon FF beyond scaling violations, however, anything drawing more than speculative conclusions at present.

We conclude the discussion of our numerical results by focusing on the correlation between the strange to non-strange PDF ratio R_s and the strange to kaon fragmentation function $D_{\bar{s}}^{K^+}$. In Fig. 13 we show R_s and the $\bar{s} \rightarrow K^+$ FF, with individual Monte Carlo samples color coded by the scaled χ^2_{red} intensity (with darker replicas indicating higher likelihoods) computed for the specific cases of SIA ($K^\pm$) and SIDIS (K^+, K^-) datasets. The SIA datasets have a clear
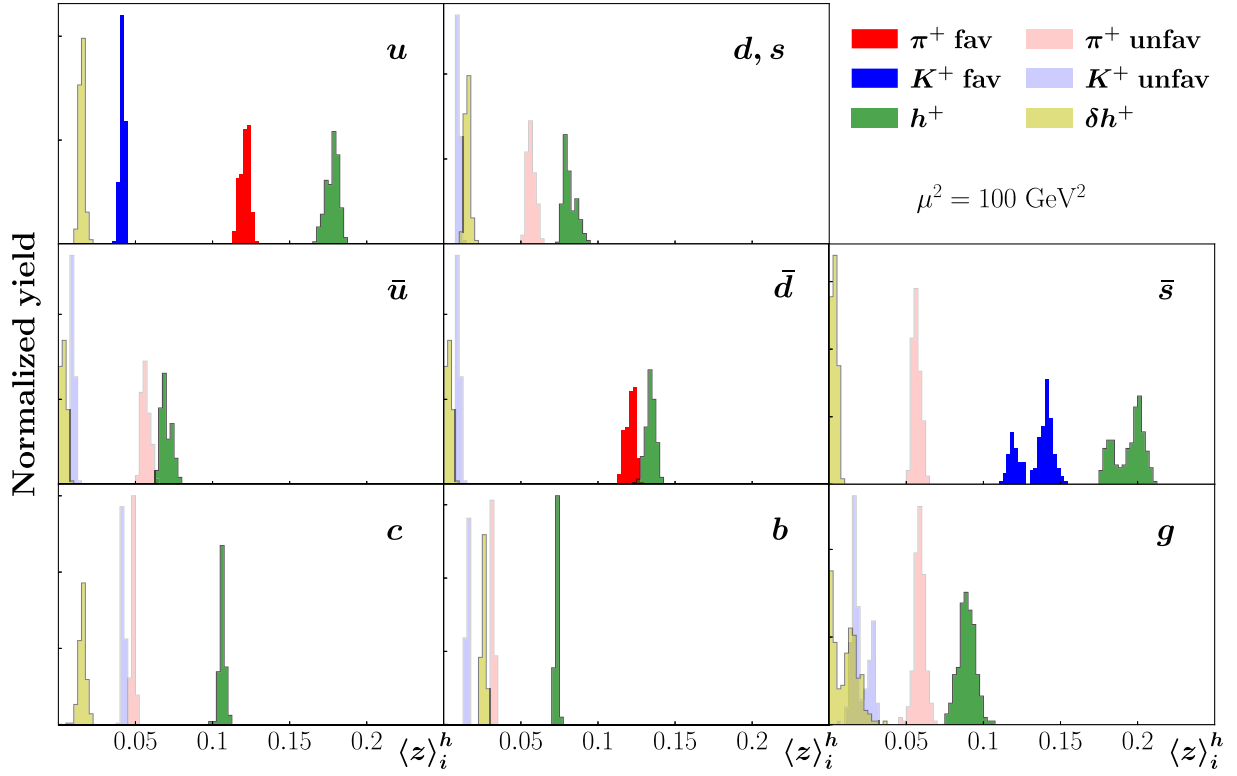
FIG. 12. Normalized yield of truncated moments $\langle z \rangle_i^h$ of the $i \to h$ FFs $z D_i^h$, for the favored π^+ (red) and favored K^+ (blue), unfavored π^+ (light red) and unfavored K^+ (light blue), the total hadron h^+ (green) and residual hadron δh^+ (yellow) FFs, at a scale of $\mu^2 = 100$ GeV^2.
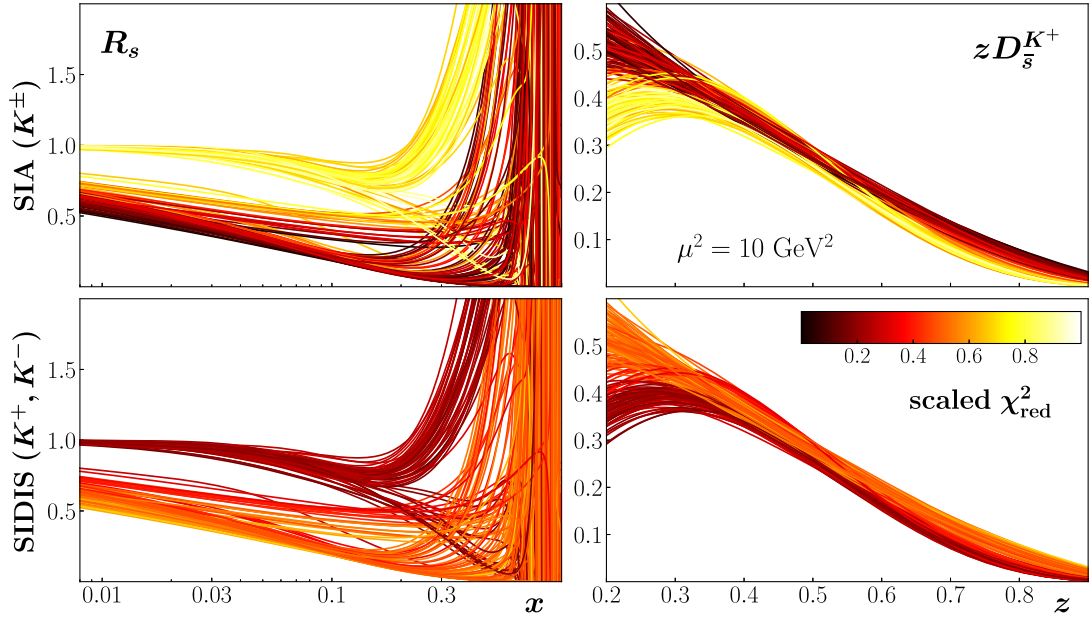

FIG. 13. Monte Carlo samples for the R_s ratio (left) and $z D_{\bar{s}}^{K^+}$ FF (right) at $\mu^2 = 10$ GeV^2, color coded according to the scaled χ^2_{red} for the SIA ($K^\pm$) (top row) and SIDIS (K^+, K^-) (bottom row) datasets.

preference for a smaller R_s and enhanced $D_{\bar{s}}^{K^+}$, as was found in the previous JAM19 analysis [2]. Interestingly, the SIDIS (K^+, K^-) data, which have smaller χ^2_{red}, have a slight tendency to favor solutions with a larger R_s and smaller $D_{\bar{s}}^{K^+}$, however, this preference is much weaker than the preference of the SIA data for smaller R_s values.

We also note that in the current analysis we have extended the flexibility of the PDF and FF parametrizations, which allowed us to obtain a more uniform Monte Carlo distribution of R_s compared JAM19, where a more restricted parametrization gave rise to multiple solutions. Our new analysis confirms that the most probable solutions found in JAM19 did not result from parametrization bias, and corroborates the need for a suppressed strange quark PDF in the proton in order to simultaneously describe both the SIA and SIDIS datasets within leading power QCD factorization.

VI. CONCLUSION

In this paper we have presented the results of a simultaneous Monte Carlo analysis of PDFs and FFs constrained by a diverse array of data from inclusive and semi-inclusive DIS, Drell-Yan lepton-pair production, and SIA in e^+e^- collisions. The analysis extends the previous JAM19 [2] simultaneous fit by including in addition unidentified charged hadrons in the final states of SIDIS and SIA, and increasing the flexibility of the PDF and FF parametrizations.

The analysis—referred to as "JAM20-SIDIS"—represents the most comprehensive determination of parton to hadron ($\pi^\pm$, $K^\pm$, $h^\pm$) FFs fitted concurrently with spin-averaged parton distributions, broadening the test of universality of parton correlation functions to more observables. The more thorough exploration of the parameter space and reduced $\chi^2_{\rm red}$ values for each of the ≈ 70 datasets fitted in this study confirmed the previous finding [2] that the combination of SIA and SIDIS datasets have a strong preference for a smaller strange to nonstrange PDF ratio, R_s, correlated with an enhanced $D_{\bar{s}}^{K^+}$ FF. As further tests of this scenario, we plan in future to extend the experimental datasets to include weak-boson and jet production in hadronic collisions, from both Tevatron and LHC data, as well as to relax the W^2 cuts for inclusive DIS to incorporate more fixed-target DIS data at high $x_{\rm Bj}$ values [58].

An important application of the current results will be in benchmark calculations of transverse momentum dependent cross sections and in particular for the small transverse momentum region where the transition from collinear factorization to TMD factorization is expected to set in. One motivation for the present project was to assess the possible role of limitations in collinear PDF and FF fits in explaining discrepancies between theory and data in the range of intermediate and large transverse momentum across a number of transversely differential processes [4–7]. For this, a truly simultaneous analysis of parton distribution and fragmentation functions across the standard set of electromagnetic processes, integrated over all transverse momentum, is necessary. The general success of the collinear fits for transverse momentum integrated cross sections that we have examined here, and their evident predictive power, suggests that factors unique to the transverse momentum differential treatment are responsible for the tension with data. We plan to address this also in future work.

APPENDIX: SEMI-INCLUSIVE DIS

In this Appendix we summarize the basic cross section and structure function formulas relevant for our analysis of the semi-inclusive leptoproduction of a hadron h (with four-momentum p_h) in the deep-inelastic scattering of a lepton ℓ (momentum l) from a nucleon N (momentum p) via the exchange of a virtual photon (momentum q),

$$\ell + N \rightarrow \ell' + h + X, \qquad (A1)$$

where the final state hadron is integrated over all transverse momentum. The formal setup follows standard methods described in Refs. [59–62], for example, and we closely follow the specific techniques in Ref. [63] utilizing the double Mellin moment method from Ref. [14].

The spin-averaged cross section is parametrized in terms of the semi-inclusive structure functions F_1^h and F_L^h,

$$\frac{d\sigma}{dx_{\rm Bj}dy dz_h} = \frac{4\pi\alpha^2}{Q^2}\left[\frac{1+(1-y)^2}{y}F_1^h(x_{\rm Bj}, z_h, Q^2)\right.$$
$$\left. + \frac{(1-y)}{x_{\rm Bj}y}F_L^h(x_{\rm Bj}, z_h, Q^2)\right], \qquad (A2)$$

which are functions of the Bjorken scaling $x_{\rm Bj}$, the hadron fragmentation scaling variable z_h, and the four-momentum transfer squared Q^2, defined in the standard way as

$$x_{\rm Bj} = \frac{Q^2}{2p\cdot q}, \qquad z_h = \frac{p_h\cdot p}{q\cdot p}, \qquad Q^2 = -q^2, \quad (A3)$$

with $y = q\cdot p/l\cdot p$ the inelasticity. Our conventions for the semi-inclusive structure function definitions are directly related to the conventions for the inclusive DIS structure functions, with $F_T = 2xF_1$ and $F_L = F_2 - 2xF_1$ the transverse and longitudinal structure functions, respectively. (Note that other conventions use instead $F_1 \rightarrow 2F_1$ and $F_L \rightarrow F_L/x$ [64].)

In the current fragmentation region, the factorization formulas for the semi-inclusive structure functions, expanded to order α_s in the hard part, are given by

$$2F_1^h(x_{\rm Bj}, z_h, Q^2)$$
$$= \sum_{q,\bar{q}} e_q^2 \Big\{ f_q(x_{\rm Bj}, Q^2) D_q^h(z_h, Q^2) + \frac{\alpha_s(Q^2)}{2\pi} [f_q \otimes C_{qq}^1 \otimes D_q^h$$
$$+ f_q \otimes C_{gq}^1 \otimes D_g^h + f_g \otimes C_{qg}^1 \otimes D_q^h](x_{\rm Bj}, z_h, Q^2) \Big\}, \quad (A4)$$

$$\frac{1}{x} F_L^h(x_{\rm Bj}, z_h, Q^2)$$
$$= \sum_{q,\bar{q}} e_q^2 \frac{\alpha_s(Q^2)}{2\pi} [f_q \otimes C_{qq}^L \otimes D_q^h + f_q \otimes C_{gq}^L \otimes D_g^h$$
$$+ f_g \otimes C_{qg}^L \otimes D_q^h](x_{\rm Bj}, z_h, Q^2), \quad (A5)$$

where f_i and D_j^h label the PDF of flavor i in the proton and parton $j \to$ hadron h FF, respectively. The functions C_{ij}^1 (C_{ij}^L) are the lowest-order hard scattering coefficient functions for the F_1 (F_L) structure functions, and the symbol $\otimes$ denotes the convolution integral over longitudinal momentum fractions, $[A \otimes B](x) \equiv \int_x^1 (\mathrm{d}z/z) A(z) B(x/z)$. Explicit expressions for C_{ij}^1 (C_{ij}^L) are given, for example, in the appendixes of Ref. [63].

In all of the expressions above, kinematical corrections from a nonzero target and final state hadron mass are neglected. To focus attention on the current fragmentation region, in our analysis we impose the kinematic cuts $z_h > 0.2$ and $W^2 > 10\,\mathrm{GeV}^2$, and for the hard scale we choose $Q^2 > m_c^2$. Therefore, the kinematical corrections may not be entirely negligible [65] at the energies of some of the experiments, although for now we set aside a fuller account of their effect to a future dedicated analysis of mass corrections in SIDIS.

[1] J. J. Ethier, N. Sato, and W. Melnitchouk, Phys. Rev. Lett. **119**, 132001 (2017).

[2] N. Sato, C. Andres, J. Ethier, and W. Melnitchouk, Phys. Rev. D **101**, 074020 (2020).

[3] J. Cammarota, L. Gamberg, Z.-B. Kang, J. A. Miller, D. Pitonyak, A. Prokudin, T. C. Rogers, and N. Sato, Phys. Rev. D **102**, 054002 (2020).

[4] J. O. Gonzalez-Hernandez, T. C. Rogers, N. Sato, and B. Wang, Phys. Rev. D **98**, 114005 (2018).

[5] B. Wang, J. Gonzalez-Hernandez, T. Rogers, and N. Sato, Phys. Rev. D **99**, 094029 (2019).

[6] A. Bacchetta, G. Bozzi, M. Lambertsen, F. Piacenza, J. Steiglechner, and W. Vogelsang, Phys. Rev. D **100**, 014018 (2019).

[7] E. Moffat, T. Rogers, N. Sato, and A. Signori, Phys. Rev. D **100**, 094014 (2019).

[8] T. Liu and J.-W. Qiu, Phys. Rev. D **101**, 014008 (2020).

[9] J. C. Collins, D. E. Soper, and G. F. Sterman, Adv. Ser. Dir. High Energy Phys. **5**, 1 (1989).

[10] R. K. Ellis, W. J. Stirling, and B. R. Webber, Cambridge Monogr. Part. Phys., Nucl. Phys., Cosmol. **8**, 1 (1996).

[11] J. C. Collins, *Foundations of Perturbative QCD* (Cambridge University Press, Cambridge, England, 2011).

[12] R. Devenish and A. Cooper-Sarkar, *Deep Inelastic Scattering* (Oxford University Press, Oxford, 2004).

[13] A. Vogt, Comput. Phys. Commun. **170**, 65 (2005).

[14] M. Stratmann and W. Vogelsang, Phys. Rev. D **64**, 114007 (2001).

[15] P. Zyla et al. (Particle Data Group), Prog. Theor. Exp. Phys. **2020**, 083C01 (2020).

[16] E. Floratos, C. Kounnas, and R. Lacaze, Nucl. Phys. **B192**, 417 (1981).

[17] S. Kretzer, Phys. Rev. D **62**, 054001 (2000).

[18] C. Anastasiou, L. J. Dixon, K. Melnikov, and F. Petriello, Phys. Rev. Lett. **91**, 182002 (2003).

[19] N. Sato, J. Ethier, W. Melnitchouk, M. Hirai, S. Kumano, and A. Accardi, Phys. Rev. D **94**, 114004 (2016).

[20] V. Bertone, N. P. Hartland, E. R. Nocera, J. Rojo, and L. Rottoli, Eur. Phys. J. C **78**, 651 (2018).

[21] D. de Florian, R. Sassot, and M. Stratmann, Phys. Rev. D **76**, 074033 (2007).

[22] A. Benvenuti et al., Phys. Lett. B **223**, 485 (1989).

[23] A. Benvenuti et al., Phys. Lett. B **237**, 592 (1990).

[24] M. Arneodo et al., Nucl. Phys. **B483**, 3 (1997).

[25] M. Arneodo et al., Nucl. Phys. **B487**, 3 (1997).

[26] L. W. Whitlow, E. M. Riordan, S. Dasu, S. Rock, and A. Bodek, Phys. Lett. B **282**, 475 (1992).

[27] H. Abramowicz et al., Eur. Phys. J. C **75**, 580 (2015).

[28] E. Hawker et al., Phys. Rev. Lett. **80**, 3715 (1998).

[29] R. Towell et al., Phys. Rev. D **64**, 052002 (2001).

[30] J. C. Webb, Measurement of continuum dimuon production in 800-GeV/c proton nucleon collisions, Ph.D. thesis, New Mexico State University, 2003.

[31] S. Alekhin, K. Melnikov, and F. Petriello, Phys. Rev. D **74**, 054033 (2006).

[32] R. Brandelik et al., Phys. Lett. **94B**, 444 (1980).

[33] M. Althoff et al., Z. Phys. C **17**, 5 (1983).

[34] W. Braunschweig et al., Z. Phys. C **42**, 189 (1989).

[35] H. Albrecht et al., Z. Phys. C **44**, 547 (1989).

[36] X.-Q. Lu, Heavy quark jets from e^+e^- annihilation at 29 GeV, Ph.D. thesis, John Hopkins University, 1986.

[37] H. Aihara et al., Phys. Rev. Lett. **52**, 577 (1984).

[38] H. Aihara et al., Phys. Rev. Lett. **61**, 1263 (1988).

[39] G. D. Cowan, Inclusive $\pi^\pm$, $K^\pm$ and $p\bar{p}$ production in e^+e^- annihilation at $\sqrt{s} = 29$ GeV, Ph.D. thesis, University of California, Berkeley, 1988.
[40] M. Derrick *et al.*, Phys. Rev. D **35**, 2639 (1987).
[41] K. Abe *et al.*, Phys. Rev. D **69**, 072003 (2004).
[42] J. Lees *et al.*, Phys. Rev. D **88**, 032011 (2013).
[43] R. Akers *et al.*, Z. Phys. C **63**, 181 (1994).
[44] G. Abbiendi *et al.*, Eur. Phys. J. C **16**, 407 (2000).
[45] D. Buskulic *et al.*, Z. Phys. C **66**, 355 (1995).
[46] P. Abreu *et al.*, Eur. Phys. J. C **5**, 585 (1998).
[47] R. Itoh *et al.*, Phys. Lett. B **345**, 335 (1995).
[48] M. Leitgab *et al.*, Phys. Rev. Lett. **111**, 062002 (2013).
[49] M. Leitgab, Precision measurement of charged pion and kaon multiplicities in e^+e^- annihilation at Q = 10.52 GeV, Ph.D. thesis, UIUC, 2013.
[50] C. Adolph *et al.*, Phys. Lett. B **764**, 1 (2017).
[51] C. Adolph *et al.*, Phys. Lett. B **767**, 133 (2017).
[52] A. Accardi, L. T. Brady, W. Melnitchouk, J. F. Owens, and N. Sato, Phys. Rev. D **93**, 114017 (2016).
[53] R. D. Ball *et al.*, Eur. Phys. J. C **77**, 663 (2017).
[54] A. Accardi, F. Arleo, W. Brooks, D. D'Enterria, and V. Muccifora, Riv. Nuovo Cimento **32**, 439 (2010).
[55] K. Kovarik, I. Schienbein, F. Olness, J. Yu, C. Keppel, J. Morfin, J. Owens, and T. Stavreva, Phys. Rev. Lett. **106**, 122301 (2011).
[56] N. Kalantarians, C. E. Keppel, and M. E. Christy, Phys. Rev. C **96**, 032201 (2017).
[57] M. Hirai, S. Kumano, T.-H. Nagai, and K. Sudoh, Phys. Rev. D **75**, 094009 (2007).
[58] C. Cocuzza *et al.* (to be published).
[59] G. Altarelli, R. K. Ellis, G. Martinelli, and S.-Y. Pi, Nucl. Phys. **B160**, 301 (1979).
[60] P. Nason and B. R. Webber, Nucl. Phys. **B421**, 473 (1994).
[61] W. Furmanski and R. Petronzio, Z. Phys. C **11**, 293 (1982).
[62] D. Graudenz, Nucl. Phys. **B432**, 351 (1994).
[63] D. de Florian, M. Stratmann, and W. Vogelsang, Phys. Rev. D **57**, 5811 (1998).
[64] D. P. Anderle, F. Ringer, and W. Vogelsang, Phys. Rev. D **87**, 034014 (2013).
[65] M. Boglione, A. Dotson, L. Gamberg, S. Gordon, J. Gonzalez-Hernandez, A. Prokudin, T. Rogers, and N. Sato, J. High Energy Phys. 10 (2019) 122.