Semi-inclusive deep-inelastic scattering, parton distributions, and fragmentation functions at a future electron-ion collider

Elke C. Aschenauer[*]
Physics Department, Brookhaven National Laboratory, Upton, New York 11973, USA

Ignacio Borsa[†] and Rodolfo Sassot[‡]
Departamento de Física and IFIBA, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Ciudad Universitaria, Pabellón 1 (1428) Buenos Aires, Argentina

Charlotte Van Hulse[§]
University of the Basque Country UPV/EHU, Spain and School of Physics, University College Dublin, Dublin, Ireland

(Received 7 March 2019; published 7 May 2019)

We present a quantitative assessment of the impact a future electron-ion collider would have in the determination of parton distribution functions in the proton and parton-to-hadron fragmentation functions through semi-inclusive deep-inelastic electron-proton scattering data. Specifically, we estimate the kinematic regions for which the forthcoming data are expected to have the most significant impact in the precision of these distributions, computing the respective correlation and sensitivity coefficients. Using a reweighting technique for the sets of simulated data with their realistic uncertainties for two different center-of-mass energies, we analyze the resulting new sets of parton distribution functions and fragmentation functions, which have significantly reduced uncertainties.

DOI: 10.1103/PhysRevD.99.094004

I. INTRODUCTION AND MOTIVATION

The quest for a quantitative picture of lepton-hadron and hadron-hadron interactions in terms of the basic constituents of matter and in the framework of perturbative quantum chromodynamics (pQCD) involves nonperturbative quantities that encode the details about the internal structure of hadrons and the mechanism leading to confinement. Parton distribution functions (PDFs) [1] and fragmentation functions (FFs) [2] stand out among these essential ingredients needed for a theoretical description of hard scattering processes. In the last two decades, remarkable progress has been made to determine these nonperturbative inputs, but the need for calculations of hadronic processes with unprecedented precision, to validate the Standard Model of fundamental interactions and our picture of matter at extreme conditions, gives the improvement of our knowledge of PDFs and FFs a crucial role in the searches for new physical phenomena.

The requirement for increased precision becomes especially relevant in the case of quarks generated through QCD radiation (*sea quarks*), which are typically less constrained than their valence counterparts, due to the comparatively reduced flavor separation power of the data generally included in global analyses [1,3,4]. An appealing solution to this lack of stringent constraints for the sea quark distributions is to take advantage of data from hadron production in semi-inclusive deep-inelastic scattering (SIDIS), which probe different quark flavor combinations depending on the final-state hadron. The idea, originally proposed by Feynman and Field [5,6], has never been exploited in modern global PDF extractions, since on the one hand, it involves the cumbersome task of a simultaneous PDF and FF extraction [7], and on the other hand, it requires access to semi-inclusive data, of the same precision as the inclusive data. While recent semi-inclusive data [8–13] have been important to reduce the uncertainties on the fragmentation functions, the precision of these extractions is still behind compared to that achieved for valence quark PDFs, due to the higher statistical precision and the kinematic coverage of totally inclusive data.

A U.S.-based electron ion collider (EIC) [14,15] with high energy and high luminosity, capable of a versatile

[*]elke@bnl.gov
[†]iborsa@df.uba.ar
[‡]sassot@df.uba.ar
[§]cvhulse@mail.desy.de

range of beam energies, polarizations, and ion species will, for the first time, be able to systematically explore and map out the dynamical system that is the ordinary QCD bound state. The EIC is foreseen to play a transformative role in the understanding of the rich variety of structures at the subatomic scale. It will open up the unique opportunity to go far beyond the present one-dimensional picture of nuclei and nucleons at high energy, where the composite nucleon appears as a bunch of fast-moving (anti)quarks and gluons of which transverse momenta or spatial extent are not resolved. Specifically, by correlating the information of the quark and gluon longitudinal momentum component with their transverse momentum and spatial distribution inside the nucleon, it will enable nuclear femtography. Such femtographic images will provide, for the first time, insight into the QCD dynamics inside hadrons, such as the interplay between sea quarks and gluons. The EIC's landmark in precision and kinematic coverage for SIDIS processes will provide differential and accurate constraints on the distributions that quantify the structure of the proton and of nuclei, and on their counterparts in the final state that describe the fragmentation of quarks and gluons into hadrons [16,17]. In particular, the EIC will allow us to probe unprecedentedly low ranges in the longitudinal parton momentum fraction in SIDIS, over various decades in photon virtuality squared, thereby allowing us to probe sea quarks for the first time with very high precision.

In this paper, we are assessing the impact that future EIC charged pion and kaon SIDIS data would have on PDFs and FFs, with particular focus on sea-quark distributions and the possibility to see charge and flavor symmetry breaking among them. In order to quantify that impact, we follow the strategy discussed in Ref. [7], but now using EIC pseudodata with realistic uncertainties.

The approach relies heavily on the application of the so-called reweighting technique for PDFs and FFs, developed by the NNPDF Collaboration [18,19] and extended to a Hessian uncertainty analysis [20]. This method allows us to modify PDFs or FFs in order to incorporate the information coming from datasets that were not included in their original global extractions, avoiding a full time-consuming refit, but preserving the statistical rigor for the uncertainty estimates. The method has already been successfully demonstrated in different applications [18–21]. Another useful tool to assess the impact of new data in a global fit is to define and calculate correlation and sensitivity coefficients between the experimental data under consideration and PDFs or FFs. These also give a comparative estimate of the impact in different kinematic regions [22,23].

Using the above mentioned tools, we have found that the forthcoming EIC pion and kaon SIDIS data will have a significant impact in the determination of PDFs and FFs not only for sea quarks but also for the up and down quark distributions in the proton and the favored FFs for pions and kaons. The improvement in the parton distributions is most noticeable for the strange quarks, especially for values of the Bjorken variable x_B below 10^{-2}, which are comparatively less determined in modern PDF fits. Our results also highlight the advantage a high center-of-mass system (c.m.s.) energy configuration of the EIC could have in the determination of the PDFs, as well as in constraining charge and flavor symmetry breaking among the proton constituents, due to the extended reach to lower x_B's, which can in leading order (LO) be associated with the momentum fraction of the incoming nucleon taken by the struck quark in the electron rest frame.

The remainder of the paper is organized as follows: in the following sections, we briefly comment on the next-to-leading-order (NLO) theoretical description of SIDIS, and on the generation of the EIC pseudodata and the corresponding uncertainties for the different energy configurations under consideration. Then, we sketch very briefly the main features of the PDF reweighting technique, its extension to FFs evaluated within the Hessian approach, and how it applies to the present study. In Sec. V A, we present the results for the correlation and sensitivity coefficient calculations, assessing the kinematic region where the new data are expected to constrain PDFs and FFs the most. In Sec. V B, we discuss in detail the outcome of our reweighting exercise for the EIC pseudodata, with special interest in the light sea quark distributions and possible flavor and charge symmetry breaking. We briefly summarize the main results and conclusions in Sec. VI.

II. SIDIS CROSS SECTION AT NLO

The cross section for the production of a final-state hadron H in deep-inelastic electron-nucleon scattering, $eN \to e'HX$, in the current-fragmentation region can be written in full analogy to the inclusive deep-inelastic (DIS) case, but in terms of the semi-inclusive structure functions F_1^H and F_L^H [24,25]:

$$\frac{d\sigma^H}{dx_B dy dz} = \frac{2\pi\alpha^2}{Q^2}\left[\frac{(1+(1-y)^2)}{y} 2F_1^H(x_B, z, Q^2) + \frac{2(1-y)}{y} F_L^H(x_B, z, Q^2)\right], \quad (1)$$

where x_B, the inelasticity y, and the virtuality of the exchanged photon Q^2 are the usual DIS variables, defined in terms of the nucleon, the photon, and the incoming electron four-momenta, p_N, q_γ, and k_e, respectively:

$$x_B = \frac{Q^2}{2p_N \cdot q_\gamma}, \qquad y = \frac{q_\gamma \cdot p_N}{k_e \cdot p_N}, \qquad Q^2 = -q_\gamma^2, \quad (2)$$

while z is the analog of x_B for the fragmentation process:

$$z = \frac{p_H \cdot p_N}{p_N \cdot q_\gamma}, \quad (3)$$

which at the lowest order in QCD can be interpreted as the fraction of the fragmenting parton momentum carried by the final-state hadron with momentum p_H. In the collinear leading-twist NLO approximation, factorization allows us to write the structure functions F_1^H and F_L^H in Eq. (1) as convolutions of the quark and gluon distribution functions in the nucleon, denoted as f_q and f_g, respectively, and the FF of D_j^H into the final hadron H:

$$2F_1^H(x,z,Q^2) = \sum_{q,\bar{q}} e_q^2 \Big\{ f_q(x,Q^2) D_q^H(z,Q^2) + \frac{\alpha_s(Q^2)}{2\pi} [f_q \otimes C_{qq}^1 \otimes D_q^H + f_q \otimes C_{gq}^1 \otimes D_g^H + f_g \otimes C_{qg}^1 \otimes D_q^H](x,z,Q^2) \Big\}, \quad (4)$$

$$F_L^H(x,z,Q^2) = \frac{\alpha_s(Q^2)}{2\pi} \sum_{q,\bar{q}} e_q^2 [f_q \otimes C_{qq}^L \otimes D_q^H + f_q \otimes C_{gq}^L \otimes D_g^H + f_g \otimes C_{qg}^L \otimes D_q^H](x,z,Q^2), \quad (5)$$

where $C_{ij}^{1,L}$ are the NLO $\overline{\text{MS}}$ coefficient functions [24–26]. Fragmentation functions are sensitive to the flavor structure of the hadron, and thus the choice of specific hadrons in the final state allows us to disentangle the contributions of the different flavors of quarks.

In recent years, increasingly precise SIDIS measurements have been performed [10,11], which are nicely described by PDFs and current FFs at NLO accuracy. Together with the single-inclusive measurements in e^+e^- annihilation [8,9] and hadron production in proton-proton collisions [12,13], they have allowed the extraction of FFs in global QCD analyses with unprecedented precision [27,28], updating previous, less comprehensive determinations [29], and bringing FF accuracy closer to that of the better determined valence PDFs.

In this paper we restrict ourselves to the case of transverse-momentum-integrated final-state hadrons produced in the current-fragmentation region. The QCD framework to describe transverse-momentum-dependent final-state hadron production is known at NLO accuracy [30], as well as hadron production in the target fragmentation region in terms of fracture functions [25,31,32].

III. SIMULATED DATA FOR SIDIS AT AN EIC

Two preconceptual designs for a future high-energy ($\sqrt{s}$: 20–100 GeV upgradeable to 140 GeV) and high-luminosity (10^{33-34} cm^{-2} s^{-2}) polarized EIC have evolved, using existing infrastructure and facilities [14]. One proposes to add an electron storage ring to the existing Relativistic Heavy-Ion Collider (RHIC) complex at Brookhaven National Laboratory (BNL) to enable electron-ion collisions. The other preconceptual design proposes a new electron and ion collider ring at Jefferson Laboratory (JLab), utilizing the 12 GeV upgraded CEBAF facility as the electron injector.

To span most of the kinematic coverage of an EIC, the studies are performed for lepton beam energies of 5 GeV and 20 GeV in combination with proton beam energies of 100 GeV and 250 GeV, respectively, using the Monte Carlo generator PYTHIA-6 [33,34] to simulate DIS events. The presented results are based on data with a statistical uncertainty corresponding to an integrated luminosity of 10 fb^{-1}. We consider only events with $Q^2 > 1$ GeV^2, a squared invariant mass of the photon-nucleon system $W^2 > 10$ GeV^2, and an inelasticity $0.01 < y < 0.95$. The kinematic range covered in Q^2 and x_B is shown in Fig. 1 for two c.m.s. energies. At the highest c.m. energy, four decades in Q^2 are spanned, while x_B reaches from 10^{-4} to 1.0. At fixed Q^2, higher c.m.s. energies allow us to access the lower region in x_B, while lower c.m.s. energies can give complementary information at higher x_B.

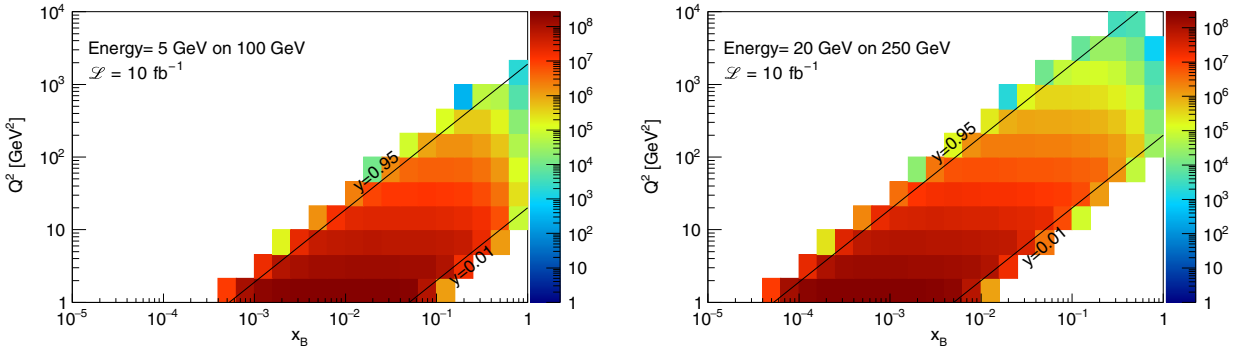

FIG. 1. Expected distribution of DIS events in bins of x_B and Q^2 for two electron-proton beam energy combinations: 5 GeV on 100 GeV (left), and 20 GeV on 250 GeV (right). The two lines indicate the limits on the x-Q^2 plane requiring $0.01 < y < 0.95$. The scattered lepton is required to be between -4 and 4 in rapidity.

TABLE I. Range in hadron (pion, kaon, and proton) momentum (p_H) covered in the various rapidity regions by different particle-identification detectors.

Rapidity	Pion momentum [GeV]	Kaon momentum [GeV]	Proton momentum [GeV]
$-3.5 <$ rapidity < -1.0 (RICH)	$0.5 < p_H < 5.0$	$1.6 < p_H < 5.0$	$3.0 < p_H < 8.0$
$-1.5 <$ rapidity < -1.0 (dE/dx)	$0.2 < p_H < 0.6$	$0.2 < p_H < 0.6$	$0.2 < p_H < 1.0$
$-1.0 <$ rapidity < 1.0 (DIRC and dE/dx)	$0.2 < p_H < 4.0$	$0.2 < p_H < 0.7$ $0.8 < p_H < 4.0$	$0.2 < p_H < 1.1$ $1.5 < p_H < 4.0$
$1.0 <$ rapidity < 3.5 (RICH)	$0.5 < p_H < 50.0$	$1.6 < p_H < 50.0$	$3.0 < p_H < 50.0$
$1.0 <$ rapidity < 1.5 (dE/dx)	$0.2 < p_H < 0.6$	$0.2 < p_H < 0.6$	$0.2 < p_H < 1.0$

For SIDIS measurements, it is critical to detect the scattered lepton with high precision over a wide range in x and Q^2. For the highest $\sqrt{s}$ and $Q^2 \sim 1$ GeV^2, the scattered lepton is at a rapidity of -4. The scattering angle of the scattered lepton is measured in forward-tracking detectors, and its energy is measured with an electromagnetic calorimeter, covering a rapidity range up to -4 and 4. The different hadrons need to be identified with high efficiency and high purity. To cover the widest range in x_B, Q^2, z, and the hadron transverse momentum with respect to the virtual photon p_T^H, it is crucial to integrate particle-identification detectors into the EIC detector over a wide range in rapidity. We follow in this paper the EIC convention that positive rapidity corresponds to the proton beam direction. Detailed design studies for a general-purpose EIC detector provided the following results important for this study. The magnetic field of the detector is of critical importance for the lowest detectable hadron momentum p_H and the achievable momentum resolution, especially at large rapidities ($\eta \sim |3.5|$). Particle momenta are limited to a minimal value of 0.5 GeV imposed by the presence of a 3 T magnet for momentum reconstruction.

For this study, we assume hadron identification detectors spanning a rapidity range $-3.5 < \eta < 3.5$. We consider identifying pions, kaons, and protons at low hadron momentum p_H by means of the measurement of energy loss per path length (dE/dx), and at medium to high hadron momentum p_H by Cherenkov radiation in a ring imaging Cherenkov (RICH) detector in the backward ($-3.5 < \eta < -1$) and forward ($1 < \eta < 3.5$) rapidity regions, while at midrapidity ($-1 < \eta < 1$), the energy loss in the gas of a time projection chamber (TPC) in combination with a detector of internally reflected Cherenkov (DIRC) light is considered. The restrictions on the range of detectable hadron momentum associated with particle identification capabilities are specified in Table I.

The cross section differential in x_B, Q^2, z, and p_T^H for two c.m.s. energies $\sqrt{s} = 45$ GeV and 140 GeV accounting for the above described detector performance is presented in Fig. 2. In this figure, the differential cross section is shown for positively charged pions as a function of x_B for different ranges in Q^2, z, and p_T^H. Note that a finer binning in Q^2 is possible, but for clarity only a subdivision per decade is presented here. As already discussed, different beam energies allow us to probe complementary regions in x_B and Q^2 independent of z and p_T^H. Measurements of SIDIS at an EIC will give access to extremely low p_T^H and z.

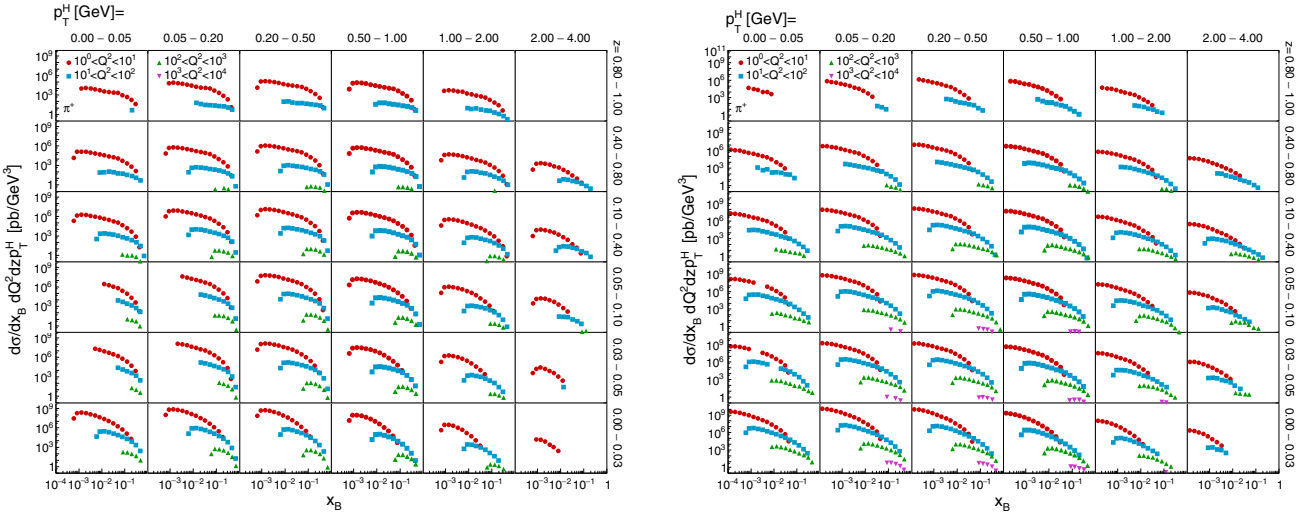

FIG. 2. Differential cross section as a function of x_B for bins in Q^2, z, and p_T^H for two center-of-mass energies 45 GeV (left) and 140 GeV (right).

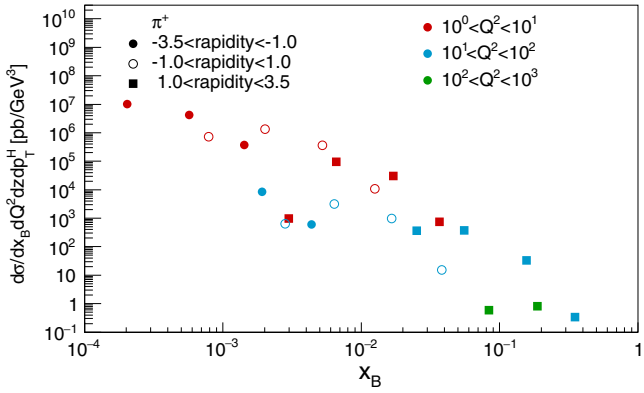

FIG. 3. The simulated differential cross section for three bins in Q^2 (indicated by the different colors) as a function of x_B for $0.4 < z < 0.8$ and $0.2 < p_T^H < 0.5$ with all PID detector requirements (see Table I) applied, separated into three rapidity ranges: backward rapidity, $-3.5 < \eta < -1$ (filled circles); midrapidity, $-1 < \eta < 1$ (open circles); and forward rapidity, $1 < \eta < 3.5$ (filled squares).

The advantage of particle detection and identification over a large range in rapidity is illustrated in Fig. 3, where the four-differential cross section for pion production is shown for the three rapidity ranges $-3.5 < \eta < -1.0$, $-1.0 < \eta < 1.0$, and $1.0 < \eta < 3.5$, and for different ranges in Q^2, at $\sqrt{s} = 140$ GeV. The pion fractional energy and transverse momentum are limited for this figure to $0.4 < z < 0.8$ and $0.2 < p_T^H < 0.5$. All particle identification cuts as listed in Table I are applied. As can be seen, the lower-Q^2 region is accessed at backward rapidity, while higher-Q^2 values are reached at forward rapidity. At fixed Q^2, lower values of x_B are covered at backward rapidity, while the higher-x_B region is probed at forward rapidity. Hence, providing particle identification at backward, mid, and forward rapidity is important in order to cover the widest range in x_B and Q^2 possible.

An important point about probing various regions in rapidity is the enhanced lever arm to separate current and target fragmentation. This is illustrated in Fig. 4. Here, the upper panels show the distribution of pions originating from a struck quark (left) and from the target remnant (right) in the Q^2-rapidity plane for the DIS subprocess $\gamma^* q \to q$ in PYTHIA-6. The bottom panels show the distributions of the struck quark (left) and the target remnant (right) from which the pion originates.[1] While one has to be very careful with the interpretation of the classification of hadrons and their origin in Monte Carlo generators, this plot illustrates clearly that there exists a correlation between the direction of a hadron and its origin. As expected, target remnants are populating regions in rapidity that are much more forward than what is correlated with the struck quark, and its associated pions follow the same trend. While the correlation is not 100% and in reality many more subprocesses than the one exemplified here contribute, the figure illustrates that by covering different regions in rapidity, one can obtain an improved separation of current and target fragmentation. Note that these correlations reveal also a clear W^2 dependence, as shown for two regions in W^2 in Fig. 5.

While particle-identification detectors will most likely not allow for a full coverage in acceptance, they should be chosen to provide a minimal loss in kinematic coverage. Similarly, the choice of the magnet strength is a compromise between the loss of low-momentum, i.e., low-z hadrons, the fraction of which increases with increasing magnetic field, and the degradation in momentum resolution at high momenta, which is inversely proportional to the strength of the magnetic field. The kinematic regions where particles are lost due to particle identification requirements and the presence of a magnetic field are shown in Fig. 6 for positively charged kaons for $\sqrt{s} = 140$ GeV. The open circles correspond to the cross section not requiring a lower-momentum cut due to the magnetic field and no restriction due to particle identification in the rapidity range between -4 and 4. All other symbols represent the situation requiring particle identification, as detailed in Table I, and different lower-momentum cutoffs. As seen from the figure, data at higher x_B values are lost at backward rapidity, because of the particle identification requirements. However, the same kinematic region is accessible at midrapidity, if the minimal momentum cut can be below 0.80 GeV. The complementarity offered by the various rapidity ranges, provided they are equipped with the appropriate detector components, is clearly illustrated in this figure. For the lower center-of-mass energy, the same conclusions hold for pions, kaons, and protons.

In the following, all impact studies for PDFs and FFs are performed based on simulated data that satisfy DIS and particle identification requirements for hadrons from Table I, with a lower-momentum cutoff of 0.5 GeV. The corresponding cross section as a function of x_B binned in Q^2 and z unfolded for detector effects is illustrated in Fig. 7 for pions at a c.m.s. energy of 140 GeV. The uncertainties correspond to a integrated luminosity of 10 fb^{-1}.

Besides the statistical uncertainties, one would need to also consider the systematic uncertainties. They consist of an overall systematic uncertainty of 1.4% on the luminosity determination and a bin-by-bin systematic uncertainty to account for the challenges to identify hadrons over a wide kinematic range and any other detector effects, which cannot be fully unfolded. A current conservative estimate of the bin-by-bin systematic uncertainty is 3.5%. It should

[1]The struck quark is selected using internal PYTHIA-6 information by cutting on the status code KS equal to 11 or 12, and the parent particle with KS = 21. The target remnant was selected requiring either a quark or a diquark through KS = 11 or 12 and the nucleon as the parent particle.

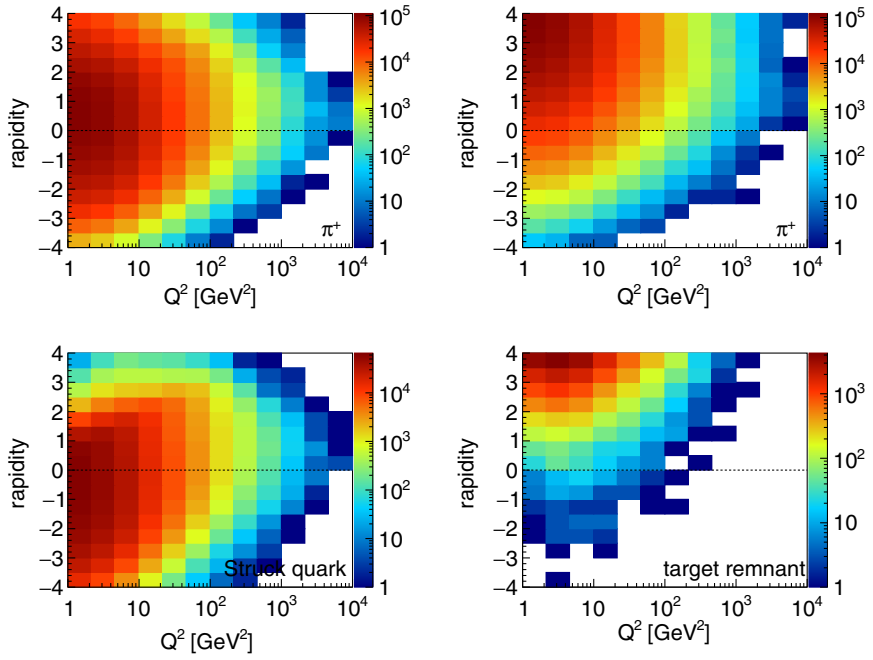

FIG. 4. Kinematic range in Q^2 and rapidity at $\sqrt{s} = 140$ GeV for pions originating from a struck quark (top left) and from the target remnant (top right), as well as for the struck quark (bottom left) and the target remnant (bottom right) themselves for the DIS subprocess $\gamma^* q \to q$ in PYTHIA-6.

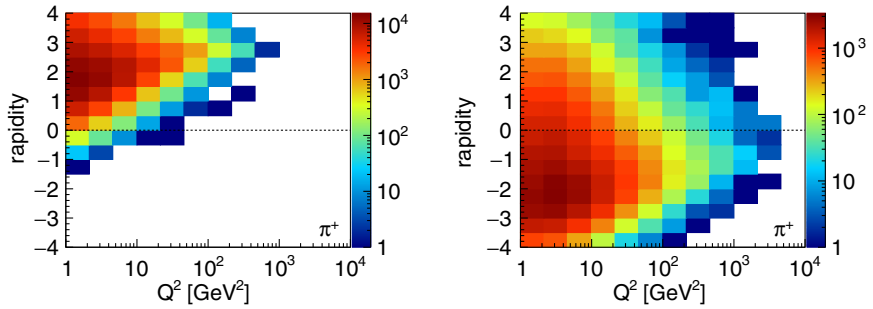

FIG. 5. Kinematic range in Q^2 and rapidity for pions originating from a struck quark for the DIS subprocess $\gamma^* q \to q$ in PYTHIA-6 with 200 GeV^2 $< W^2 <$ 300 GeV^2 (left) and 16000 GeV^2 $< W^2 <$ 18000 GeV^2 (right) for $\sqrt{s} = 140$ GeV.

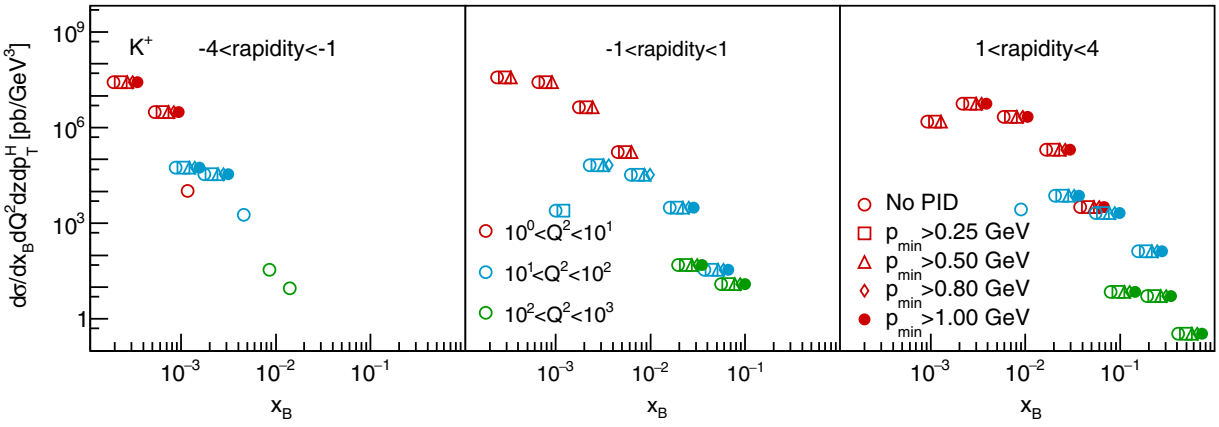

FIG. 6. Differential cross section for kaon production in three regions in rapidity, without particle identification and minimum momentum requirement (open circles), compared to the results with particle identification and for different detectable lowest particle momenta.

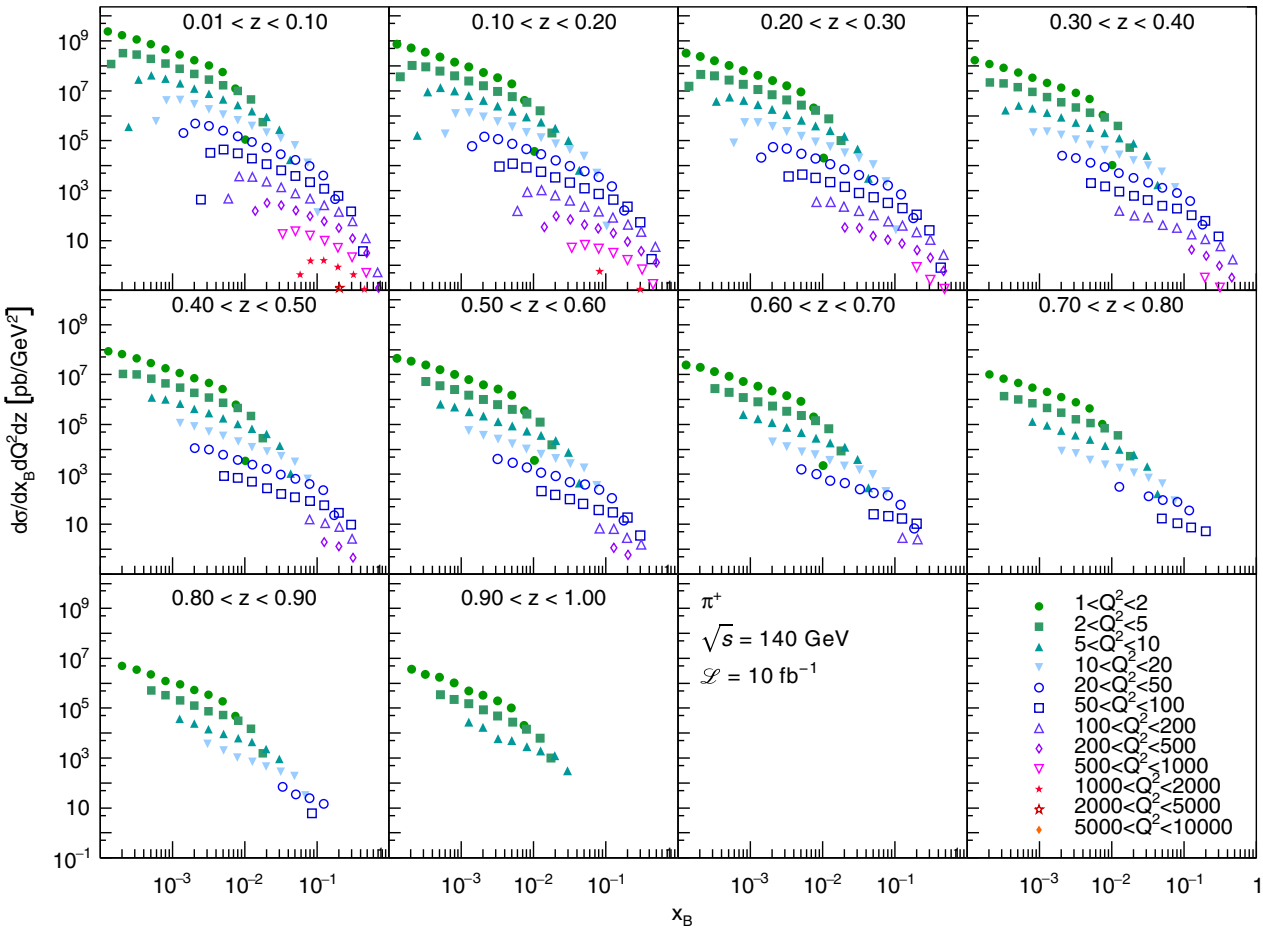

FIG. 7. Differential cross section for pion production at $\sqrt{s} = 140$ GeV as a function of x_B for bins in Q^2 and z measurable at an EIC.

be added in quadrature to the statistical one. As it is difficult without a full detector design to estimate this bin-to-bin uncertainty reliably, we decided to not consider it in our study.

IV. BAYESIAN AND HESSIAN TOOLBOX

A. PDF and FF reweighting with SIDIS data

One of the key ingredients in the strategy pursued in the present analysis is the use of reweighting methods to a set of PDFs or FFs, as a means to incorporate additional information from new data into an existing set, without the need to perform a new global fit [18,19]. Successful demonstrations of the method have been performed in different applications, and more specifically, its usefulness in constraining PDFs with actual SIDIS data has already been shown in Ref. [7]. Here, we briefly recall the main features that are needed for our analysis below.

The method was originally developed based on Bayesian inference and relies on the beforehand generation of a large ensemble of PDF or FF sets $f_i^{(k)}$, by fitting replicas of data obtained by smearing available experimental data according to their experimental and systematic uncertainties and correlations. Here, i is indexing the parton flavor and k the number of the replica. Such an obtained set of PDF or FF replicas forms a precise representation of the underlying probability distribution for the PDFs or FFs. Any observable $\mathcal{O}$ depending on PDFs and FFs can be evaluated by averaging the results for the individual replicas:

$$\langle \mathcal{O} \rangle = \frac{1}{N} \sum_{k=1}^{N} \mathcal{O}[f_i^{(k)}], \tag{6}$$

with N being the number of replicas, and the corresponding variance defined as

$$\Delta \mathcal{O} = \sqrt{\frac{1}{N-1} \sum_{k=1}^{N} \left(\mathcal{O}[f_i^{(k)}] - \langle \mathcal{O} \rangle\right)^2}. \tag{7}$$

Using Bayesian inference, it is possible to assess the effect of a new, independent dataset by updating the probability distribution through the assignment of a new weight $w_k \neq 1$ for each replica. This weight measures the agreement of replica k with the new data. The weighted estimate for any observable then becomes

$$\langle \mathcal{O} \rangle_{\text{new}} = \frac{1}{N} \sum_{k=1}^{N} w_k \mathcal{O}[f_i^{(k)}]. \tag{8}$$

Clearly, replicas with very small weights become irrelevant in the calculation of any observable, thus reducing the spread of the modified probability distribution compared to the original one. As long as the new dataset is not too restrictive, and the number of replicas with non-negligible values of w_k is large enough, reweighted PDFs or FFs will form an accurate representation of the original probability distribution.

The reweighting strategy can also be implemented within the Hessian approach for uncertainties in global PDF or FF extractions [20]. In this case, the large ensemble of replicas needed can be constructed as a Gaussian smearing of the Hessian eigenvector sets:

$$f_k \equiv f_{S_0} + \sum_{i}^{N_{\text{eig}}} \left(\frac{f_{S_i^+} - f_{S_i^-}}{2} \right) R_{ik}. \tag{9}$$

Here, f_{S_0} corresponds to the value of the PDF (FF) obtained with the best-fit parameters, while $f_{S_i^+}$ and $f_{S_i^-}$ are the values of the PDF (FF) evaluated for extreme displacements in the direction of the ith eigenvector. R_{ik} are random numbers with a Gaussian distribution centered at zero and with variance 1. The weights w_k for each replica can be calculated in a completely analogous way as in the case of Monte Carlo–based replicas, and therefore, the reweighted PDF (FF) can be written as

$$f_{\text{new}} \equiv f_{S_0} + \sum_{i}^{N_{\text{eig}}} \left(\frac{f_{S_i^+} - f_{S_i^-}}{2} \right) \left(\frac{1}{N_{\text{rep}}} \sum_{k}^{N_{\text{rep}}} w_k R_{ik} \right). \tag{10}$$

In the following, we use an ensemble of 1000 PDF replicas from Ref. [35] to perform the PDF reweighting, while in the case of the FFs a set of 10^5 replicas is generated from Eq. (9). We compute the weights by comparing to which extent each replica k reproduces the EIC SIDIS pseudodata for charged pions and kaons. The much larger number of starting FF replicas is related to the fact that current sets of FFs are typically much less constrained than PDFs, and the reweighting with very precise data such as that expected from an EIC leaves a comparatively small number of surviving replicas. The SIDIS cross sections are computed at NLO accuracy by convoluting each replica with a variant of the DSS FFs for pions and kaons [27,28], but upgraded so that they use the NNPDF3.0 set of PDFs and the corresponding α_s as input for consistency [7].

Notice that there is a subtlety regarding the inclusion of SIDIS data in the reweighting procedure, since in addition to the experimental uncertainties of the pseudodata, there are also uncertainties associated with the FFs when reweighting PDFs, or conversely with PDFs when reweighting FFs, which are used in the calculation of the observable. These, of course, have to be taken into account when computing the weights. For the FF reweighting, the uncertainties associated with the PDFs are added in quadrature to the experimental uncertainties, and for the PDF reweighting, the FF uncertainties are included in a similar way. The latter case is more involved, since the FF uncertainty estimates already include those of the PDFs used in the original FF extraction, producing a mild double counting that needs to be accounted for. This issue was addressed in Ref. [7], where a criterion on how to include the FF uncertainty consistently was proposed. In the following, we adopt the same procedure.

B. Correlations with Monte Carlo replicas

Another major advantage of Monte Carlo replicas and Hessian eigenvector sets lies in the possibility to use them in order to scan the regions of phase space where the measurements for some observable can potentially constrain the nonperturbative distributions (PDFs and FFs). This can be achieved through the calculation of the correlation coefficients between that observable and the PDF (FF) for a given flavor. The calculation of correlations in both the Hessian and the Monte Carlo formalisms has been discussed in detail in the literature [22,23,36–38].

In the case of a set of replicas for PDFs based on the Monte Carlo method, the correlation coefficient $\rho[f_i, \mathcal{O}]$ between a PDF for a given flavor i and an observable $\mathcal{O}$ (i.e., the cross section for a given process) can be defined as [38]

$$\rho[f_i, \mathcal{O}] = \frac{\langle \mathcal{O} \cdot f_i \rangle - \langle \mathcal{O} \rangle \langle f_i \rangle}{\Delta \mathcal{O} \Delta f_i}, \tag{11}$$

where the mean values are calculated over the ensemble of replicas as in Eq. (6), while the standard deviations for the observable and parton density are given by Eq. (7). Values for $|\rho|$ close to unity indicate that the observable and the PDF are highly correlated, and therefore, including data of that type with competitive experimental uncertainties could in principle further constrain the PDF. Values close to zero are obtained for uncorrelated observables, which would never be able to improve the PDF determination, irrespective of how precise those data are. For simplicity, we omit the dependencies on x_B, Q^2, and z; however, the correlation coefficients are defined for the kinematics of each individual point of the pseudodata, allowing a straightforward comparison between the constraining power of different kinematics.

It is noted that the correlation coefficients can only give insight into the *potential* impact that the new data could have on the PDF or FF determination, but they do not take into account the experimental uncertainties for the observable, which ultimately determine the *actual* constraining power. If, for a given region of phase space, the

experimental uncertainties are large compared to the uncertainty propagated from the PDFs, it is reasonable to expect that these measurements will not constrain the PDFs in this region, regardless of the value of the correlation coefficient.

In order to have a better estimate of the impact of the actual data in a global fit, one can define a scaled correlation or sensitivity coefficient [23] as

$$S[f_i,\mathcal{O}] = \frac{\langle \mathcal{O}\cdot f_i\rangle - \langle\mathcal{O}\rangle\langle f_i\rangle}{\xi\Delta\mathcal{O}\Delta f_i}, \tag{12}$$

where the scaling factor

$$\xi \equiv \frac{\delta\mathcal{O}}{\Delta\mathcal{O}} \tag{13}$$

is defined as the ratio of the experimental uncertainties of the measurement $\delta\mathcal{O}$ and the theoretical uncertainty for that same observable propagated from the PDFs $\Delta\mathcal{O}$. The scaled correlation coefficient suppresses those regions of phase space for which the experimental uncertainties are large compared to the uncertainty associated with the PDFs, while it enhances those regions where the largest impact on the distributions is expected. Of course, the scaled coefficients are no longer constrained to vary within $[-1,1]$.

C. Correlations within the Hessian approach

While several sets of PDF replicas based on the Monte Carlo method are nowadays available, this is not the case for the FFs. In Ref. [22], Monte Carlo–based FFs have been produced; however, they do not include charge separation, nor do they include SIDIS data. On the other hand, extractions like those in Refs. [27,28], that include flavor separation and SIDIS data, estimate uncertainties using the Hessian strategy and therefore the previous method cannot be directly applied. Nevertheless, it is still possible to quantify the correlations within the Hessian formalism. One can define a correlation coefficient analogous to $\rho[f_i,\mathcal{O}]$, in terms of Hessian eigenvector sets following Ref. [23]:

$$\rho[D_q^H,\mathcal{O}] = \frac{\vec{\nabla}D_q^H\cdot\vec{\nabla}\mathcal{O}}{\Delta D_q^H\Delta\mathcal{O}}, \tag{14}$$

where the gradient is taken in the space of Hessian eigenvector FF parameters and can be approximated by this finite difference:

$$\frac{\partial X}{\partial x_i} = \frac{1}{2}(X_i^+ - X_i^-), \tag{15}$$

where $X_i^\pm$ represents the values of X for extreme displacements along the direction of the ith eigenvector, for a given tolerance. Similarly, the uncertainty for any observable can be estimated as

$$\Delta X = |\vec{\nabla}X| = \frac{1}{2}\sqrt{\sum_{i=1}^{N}(X_i^+ - X_i^-)^2}, \tag{16}$$

so that the expression for the correlation in Eq. (14) can be recast as

$$\rho[D_q^H,\mathcal{O}] = \frac{1}{4\Delta D_q^H\Delta\mathcal{O}}\sum_{i=1}^{N}[(D_q^H)_i^+ - (D_q^H)_i^-](\mathcal{O}_i^+ - \mathcal{O}_i^-). \tag{17}$$

As in the case of the PDF correlations, it is worth noting that the correlations defined in Eq. (17) do not account for the experimental uncertainties of the new data or the precision already achieved in the distributions, so it is convenient to define a sensitivity coefficient [23]:

$$S[D_q^H,\mathcal{O}] = \frac{1}{4\xi\Delta D_q^H\Delta\mathcal{O}} \times\sum_{i=1}^{N}[(D_q^H)_i^+ - (D_q^H)_i^-](\mathcal{O}_i^+ - \mathcal{O}_i^-), \tag{18}$$

where again ξ is given by Eq. (13).

V. RESULTS

A. Correlations

In this section, we present the results for the correlation and sensitivity coefficients between pion and kaon pseudodata and the nonperturbative distributions (PDFs and FFs), assessing the regions of phase space where the data have the largest impact on the determination of these distributions. We also assess the impact of two different c.m.s. energies ($\sqrt{s} = 45$ GeV and 140 GeV) of the future EIC.

Figures 8 and 9 show the correlation coefficients between PDFs for different quark flavors and the SIDIS cross sections for charged pions and kaons as a function of x_B for c.m.s. energies $\sqrt{s} = 140$ GeV and $\sqrt{s} = 45$ GeV, respectively. The coefficients for π^+ and π^- are represented by the dotted (blue) and dash-dotted (light-blue) lines, respectively, while those for K^+ and K^- are shown as the dashed (pink) and the long dash-dotted (violet) lines, respectively. The correlation coefficients are calculated for the kinematics $\{x_B, Q^2, z\}$ of each pseudodata point, evolving the PDFs to the adequate $\{x_B, Q^2\}$, while the lines interpolate between data points at the same $\{z, Q^2\}$.

As expected, larger correlations are typically found for quark flavors that are valencelike for the final-state hadron, e.g., $\bar{d}$ in π^+, in the region of x_B where such flavor is most abundant in the proton target. The valence flavors show larger correlations at larger x_B; e.g., at larger x_B, π^+ (π^-) production cross sections show a stronger correlation with

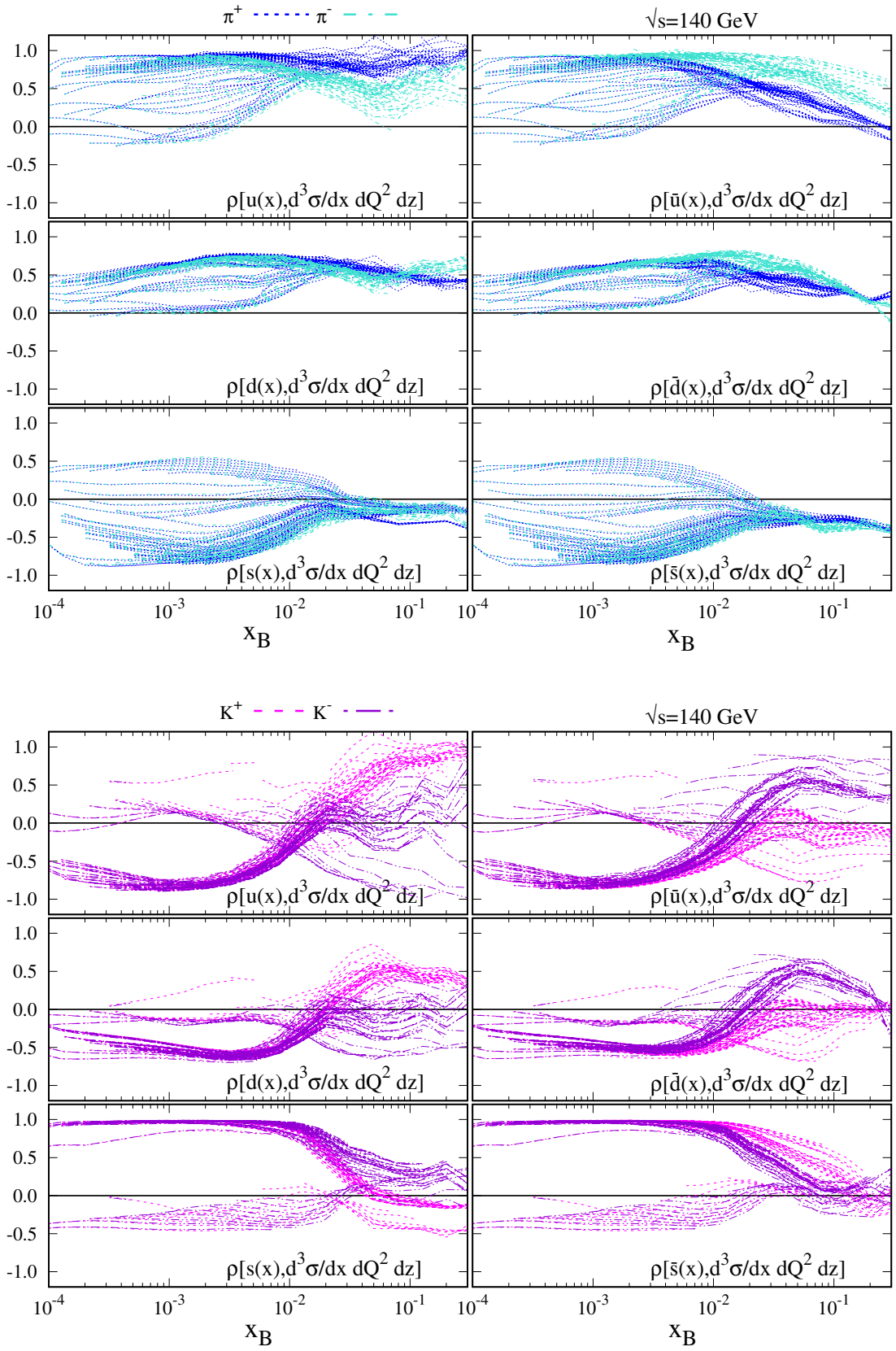

FIG. 8. Correlation coefficient ρ between the charged kaon (magenta and violet) and pion (cyan and blue) production in SIDIS at an EIC, and the light-quark PDFs, as a function of x_B at $\sqrt{s} = 140$ GeV. Each box in the figure represents the correlation with one specific quark flavor. Each line corresponds to a different bin in Q^2 and z.

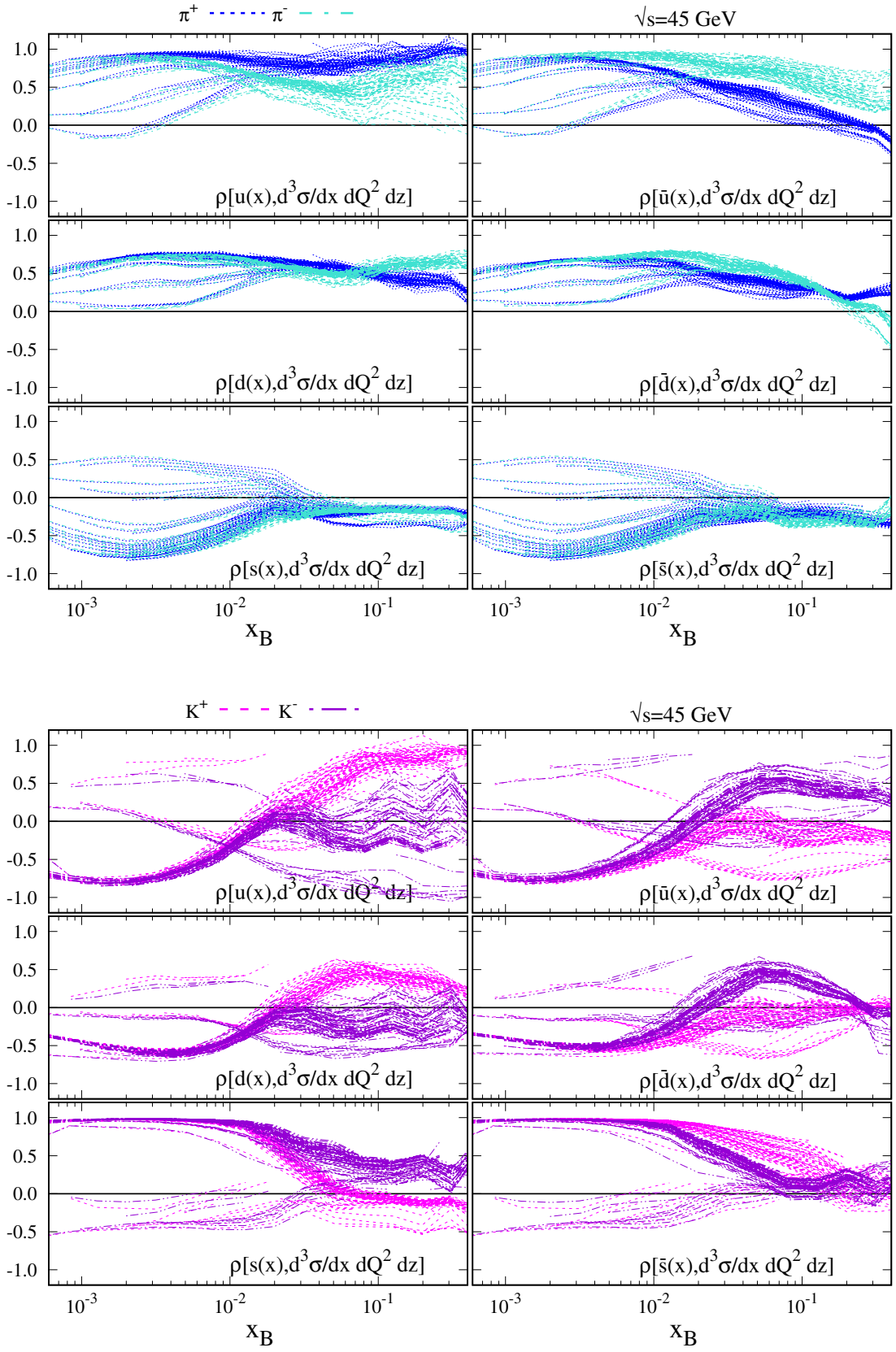

FIG. 9. Same as Fig. 8, but for $\sqrt{s} = 45$ GeV.

$u(\bar{u})$ and $\bar{d}(d)$ quark distributions, while the ones with s and $\bar{s}$ quarks are suppressed. For lower values of x_B, the data probe sea-quark distributions, for which $s \sim u \sim d$ and $q = \bar{q}$, this balances the correlation coefficients of π^+ and π^- and enhances the anticorrelations with strange quarks, which become of the same magnitude as the light quarks.

In the case of u quarks, the correlation coefficient for the simulated cross section for positively charged pions is close to 1 for the full range of x_B probed, while the same holds for $\bar{u}$ and the cross section for negatively charged pions, as is foreseeable, considering that $D_u^{\pi^+} = D_{\bar{u}}^{\pi^-}$. Ultimately, most of the constraints for these distributions will therefore come from the pion production data. It is also worth noticing that due to the electric charge factors, the correlation coefficients for the (anti)up-quark distribution are enhanced compared to those of the (anti)down-quark distribution.

Similar features can be found for the kaon production cross section. In this case, stronger correlations are obtained for the u ($\bar{u}$) and $\bar{s}$ (s) quarks, in agreement with the K^+ (K^-) valence composition. For values $x_B > 10^{-2}$, the correlation with $\bar{s}$ (s) almost vanishes, as the data probe mainly the proton's valence distributions for these values of x_B, while the proton only has sea strange quarks. For lower values of x_B, one can access the (anti)strange-quark distributions. In this x_B range, the correlation coefficients for these distributions get close to 1, while some anticorrelation is obtained with (anti)up and (anti)down quarks. Comparing the correlation coefficients, it can be anticipated that the constraint on the strange content of the proton will essentially come from the kaon data at small x_B. The results seem to indicate that kaon data could also be relevant for the determination of the (anti)up-quark distributions at higher values of x_B. However, as will be discussed later in this section, for higher x_B, these data become less relevant.

Regarding the different energy configurations, it is worthwhile noticing that while the correlation coefficients obtained for the lower-energy configuration span slightly higher values of x_B than those obtained for the higher-energy configuration, the correlations do not show significantly different features.

In order to have a better insight into which datasets best constrain the PDFs, it is illustrative to plot the correlation coefficients as a function of both x_B and Q^2. In Figs. 10 and 11, we show the correlation coefficients for pion and kaon production in SIDIS, and the parton distribution for the different light quarks over the x-Q^2 plane. For each pseudodata point, we plot a circle with a radius proportional to the absolute value of the correlation coefficient. Similar considerations on the correlations to those discussed for Figs. 8 and 9 hold here; however, now the Q^2 dependence of the correlations is made explicit. Notice that larger values of Q^2 are correlated to larger values of x_B, as usual for DIS experiments. For these values, a clear hierarchy emerges with the largest correlation coefficients for quark flavors that are *valence*-like for the final-state hadron, have the largest charge ($e_q = 2/3$) factor, and are *valence*-like also in the proton target. The weakest correlation is, as expected, for the

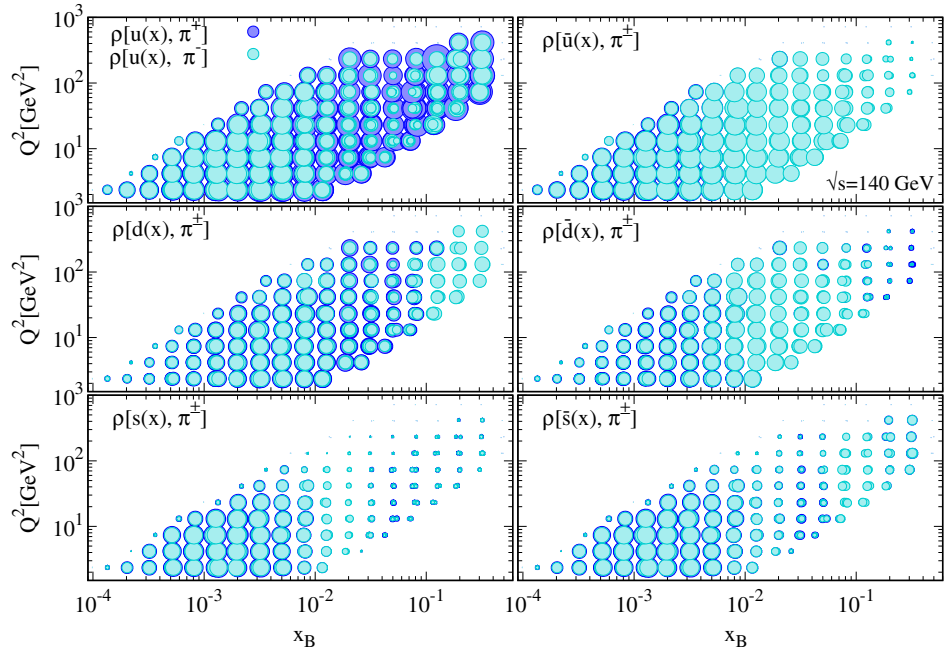

FIG. 10. Correlation coefficient ρ between the cross section for charged pion production and the PDFs of the light quarks, as a function of both the Bjorken variable x_B and the square of the momentum transfer Q^2. For each data point, a circle of which the radius represents the correlation is depicted.

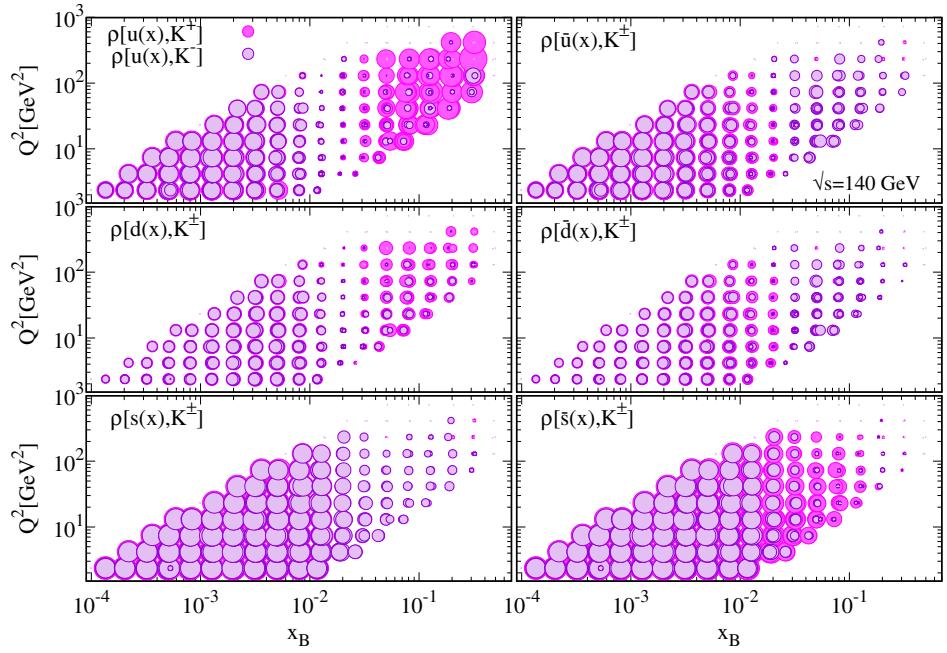

FIG. 11. Same as Fig. 10, for charged kaon production.

strange quarks and pions at larger x_B and Q^2. However, the full strength of this kind of plot will become more apparent when studying the sensitivity coefficients.

As discussed in Sec. IV B, the correlation coefficients only give an estimate of the *potential* impact that a new dataset could have if included in a new global fit, because the experimental precision of the data is not taken into account, and more specifically, because the correlation coefficients do not describe how precise the new data are compared to those used for the PDF determination nor how well the new data are described by the existing PDFs, within their uncertainty. In this respect it is more instructive to examine the sensitivity, or weighted correlations, defined in Sec. IV B. In Figs. 12 and 13, we show the sensitivity coefficient as a function of both x_B and Q^2. As before, the size of the circle for each data point is determined by the absolute value of the sensitivity coefficient. Notice that contrary to the correlation coefficients, the sensitivity coefficients are not normalized to unity, but instead they are proportional to the ratio between the uncertainty in the cross section propagated from the PDFs and that coming from the measurement.

Comparing Fig. 10 with Fig. 12, and Fig. 11 with Fig. 13, it becomes evident that the most significant impact on the PDFs is expected to come from the low-x_B region, which for SIDIS, like for DIS, is associated with the low-Q^2 region. Even though the charged hadron production data have high correlations with different parton distributions throughout the complete kinematic range covered, the most important impact is expected for $x_B < 10^{-2}$ and $Q^2 < 10^2$, since for higher x_B and Q^2, the PDFs are already well constrained.

At this point, it is also enlightening to compare the sensitivity estimates obtained for $\sqrt{s} = 140$ GeV with those for $\sqrt{s} = 45$ GeV. The latter are shown in Figs. 14 and 15 for pions and kaons, respectively. For this lower c.m.s. energy, the impact of the SIDIS data is restricted to the kinematic region given by $10^{-3} < x_B < 10^{-2}$, where the highest values of sensitivity are obtained. However, in spite of the high correlations of the cross sections at higher values of momentum fractions, the expected impact is diluted by the relative error. Notice that for this energy configuration, the most sensitive region explored with the $\sqrt{s} = 140$ GeV c.m.s. configuration—i.e., the one shown to have the greatest constraining power in Figs. 12 and 13, $10^{-4} < x_B < 10^{-3}$—is not probed.

Regarding the impact that EIC SIDIS data could have in the extraction of fragmentation functions, it is worth noting that SIDIS data have a central role in global fits, since they provide almost all the separation between quark and antiquark fragmentation and a good deal of that between flavors. The remarkably precise data from inclusive single-hadron production in electron-positron annihilation (SIA) are mostly sensitive to the singlet combination of fragmentation functions, while hadron production in proton-proton collisions mainly probes gluon fragmentation. As explained in Sec. IV B, the correlation and sensitivity coefficients can also be defined within the improved Hessian approach, considering the variations of the observables over the Hessian eigenvector sets, which is the technique implemented in the charge- and flavor-discriminated DSS extractions of FFs and their updates [27–29].

In order to establish the kinematic regions where the EIC SIDIS data could have the most significant impact for FFs,

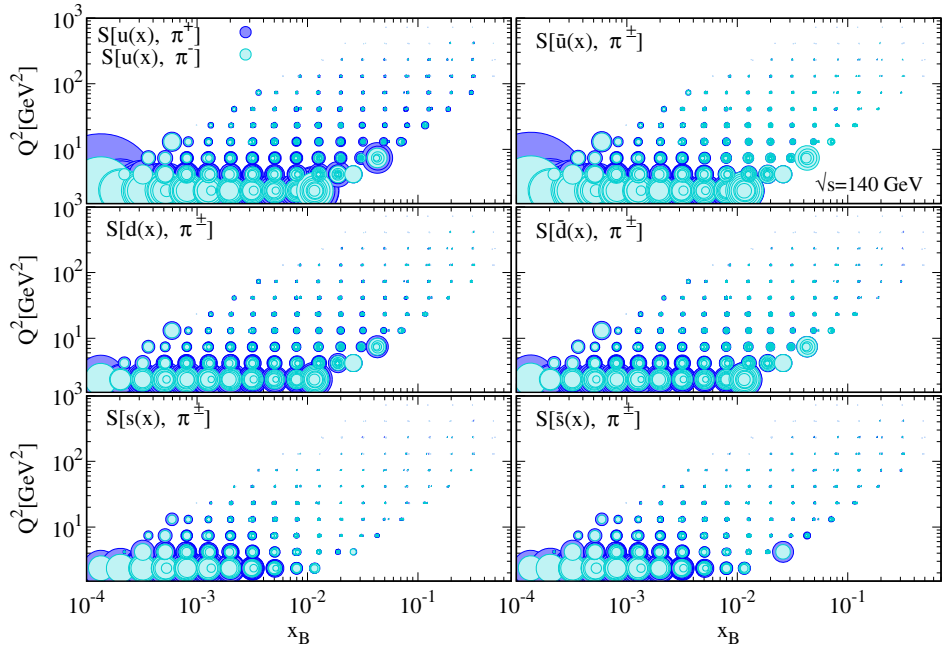

FIG. 12. Sensitivity coefficients S between the cross section for charged pion production and the different light-quark parton distributions, as a function of x_B and the transferred momentum squared Q^2. As in Fig. 10, each circle corresponds to a particular kinematic configuration $\{x_B, Q^2, z\}$ associated with a point from the pseudodata. Its radius corresponds to the value of the sensitivity coefficient for that particular configuration.

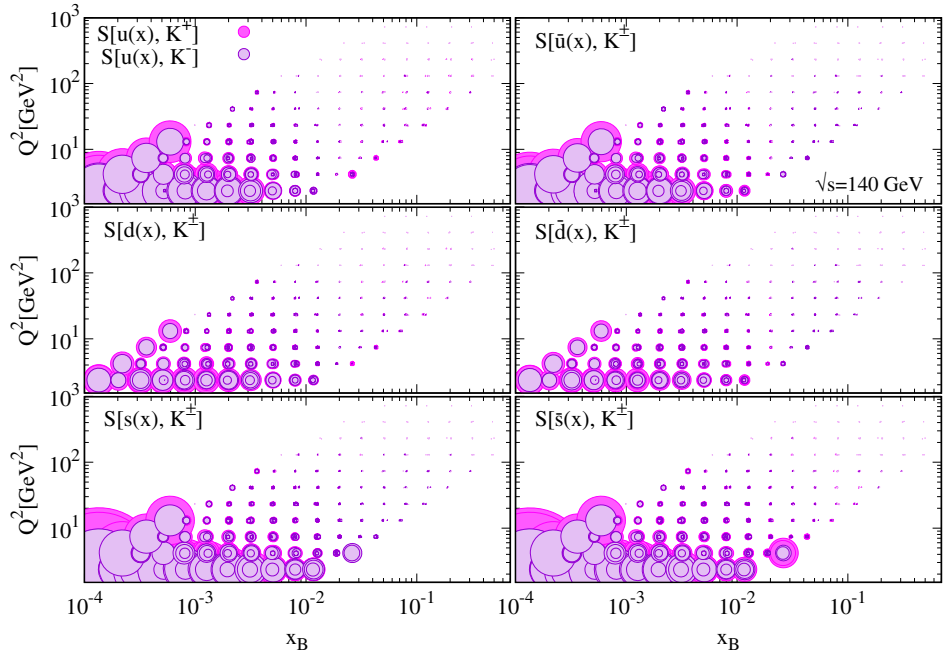

FIG. 13. Same as Fig. 12, for charged kaon production.

we compute the sensitivity coefficients between the cross section for charged pion and kaon production and the *plus* and *minus* combinations $D^{H^\pm}_{q+\bar{q}}$ and $D^{H^\pm}_{q-\bar{q}}$ discriminating for each final-state hadron, and for each of the light-quark flavors. The former are the combinations expected to be constrained by SIA, while the latter are better constrained by SIDIS. In Figs. 16 and 17 we show the sensitivities as a function of z and Q^2. The left panels correspond to the coefficients for the cross sections for kaon production, while the right panels are associated with the cross sections

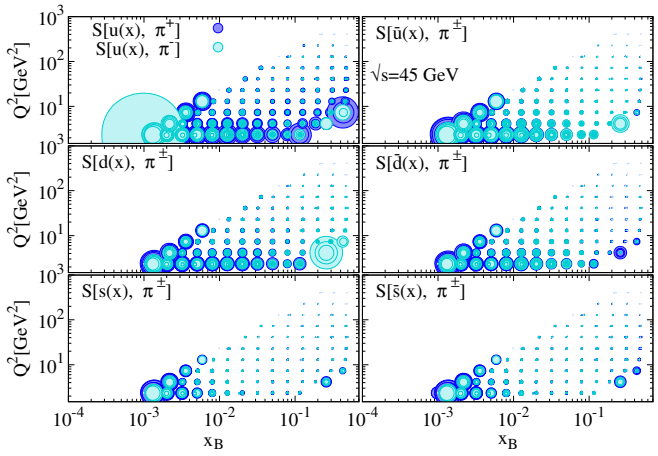

FIG. 14. Same as Fig. 12, for a c.m.s. energy $\sqrt{s} = 45$ GeV. To make the comparison clear, we keep the same scales as in the previous plots.

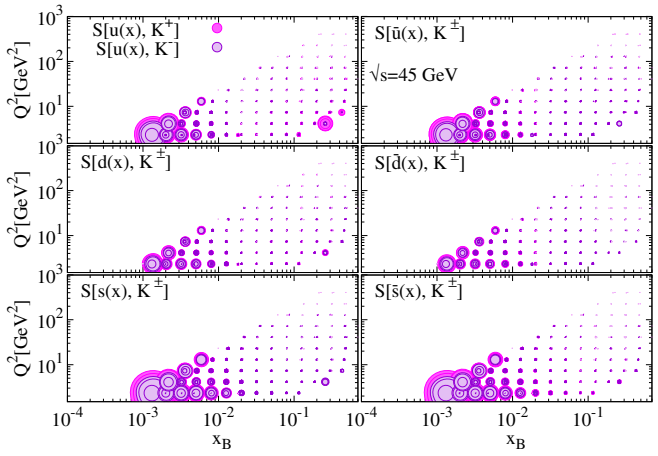

FIG. 15. Same as Fig. 14, for charged kaon production.

for pion production. The coefficients are calculated for the c.m.s. energy of $\sqrt{s} = 140$ GeV.

As can be seen in Figs. 16 and 17, the sensitivities typically grow as Q^2 and z decrease, mainly because of the FF uncertainties, which increase in these limits, and are non-negligible for both the *plus* and *minus* combinations, suggesting a significant constraining power not only for the charge separation, but also competitive for discriminating between quark flavors. Notice that in the DSS FF extractions [27,28], the only data below $z \sim 0.1$ come from LEP experiments, at very high-energy scales, which explains the impressive increase of the sensitivity.

For completeness, we also include in Fig. 17 the results for the sensitivity coefficients between the gluon FF $D_g^{H^\pm}$ and the charged hadron production cross sections. These are found to be marginal, since the constraining power of SIDIS data is not competitive with the RHIC and ALICE proton-proton collision data already included in the global fits, except below $z \simeq 0.2$. We do not show the correlations for $D_{d-\bar{d}}^{K^\pm}$ and $D_{s-\bar{s}}^{\pi^\pm}$, since these combinations are assumed to vanish in the DSS sets because of flavor symmetry considerations.

B. Results for the reweighting using EIC SIDIS pseudodata

While the correlation and sensitivity coefficients are very useful tools to anticipate and identify the kinematic regions where a given dataset can be most relevant for constraining parton densities or fragmentation functions, ultimately the effect of the inclusion of the new data on the distributions needs to be explicitly assessed by performing new global fits or a reweighting of a set of replicas. In this section, we present and discuss the results of the reweighting exercise performed using the pseudodata generated for charged pion and kaon production in SIDIS described in Sec. IV, and we show the resulting set of modified PDFs and FFs, as well as combinations of these distributions that quantify the degree of the charge and flavor symmetry breaking.

We start with the nonstrange light-quark PDFs. In Fig. 18, we show the effect of reweighting a set of 1000 PDF replicas of the NLO NNPDF3.0 set with EIC SIDIS pseudodata. The four panels on the left-hand side correspond to a set of pseudodata at a c.m.s. energy of $\sqrt{s} = 140$ GeV, while those on the right-hand side correspond to $\sqrt{s} = 45$ GeV. In both cases, the number of effective replicas $N_{\rm eff}$ is above 80, ensuring that the modified distributions are an accurate representation of the original probability distribution.

In reweighting, pseudodata with $z < 0.1$ are excluded, since the FFs used to compute the central values have rather large uncertainties that hinder any constraining effect. On the other hand, pseudodata points with $Q^2 < 2$ GeV^2 are also excluded from the reweighting, since their statistical power is so restrictive that the resulting number of effective replicas after the reweighting is extremely low ($N_{\rm rep} \approx 10$). Similarly, it should also be noted that the pseudodata coming from the two alternative c.m.s. configurations are not combined into a single reweighting, given that the constraints imposed by the whole dataset leave a low number of effective replicas. This indicates that if the whole dataset were to be included in a global fit, the impact on the uncertainties would be stronger, but it would require either a new global fit or a reweighting with a much larger number of replicas.

Since the pseudodata are generated around the NNPDF3.0 best-fit result, the main effect on the distributions is expected to be a reduction on the uncertainty bands, with a very minor variation of the central values. Indeed, the new SIDIS information can at most balance small tensions already present between the datasets of the original fit. The distributions and the corresponding uncertainty bands are normalized to the NNPDF3.0 best-fit result,

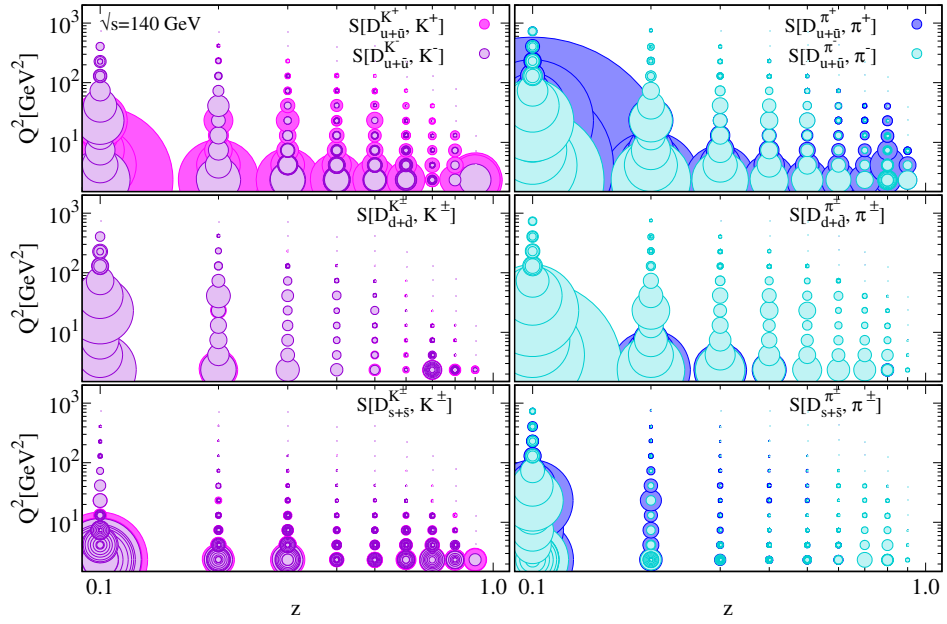

FIG. 16. Sensitivity coefficients between the cross section for charged hadron production at $\sqrt{s} = 140$ GeV, pions (blue and light blue) and kaons (pink and violet), and the singlet FF combination $D^{H^\pm}_{q+\bar{q}}$ for the different light quark flavors. The coefficients are obtained using the Hessian formalism described in Sec. IV B (see text).

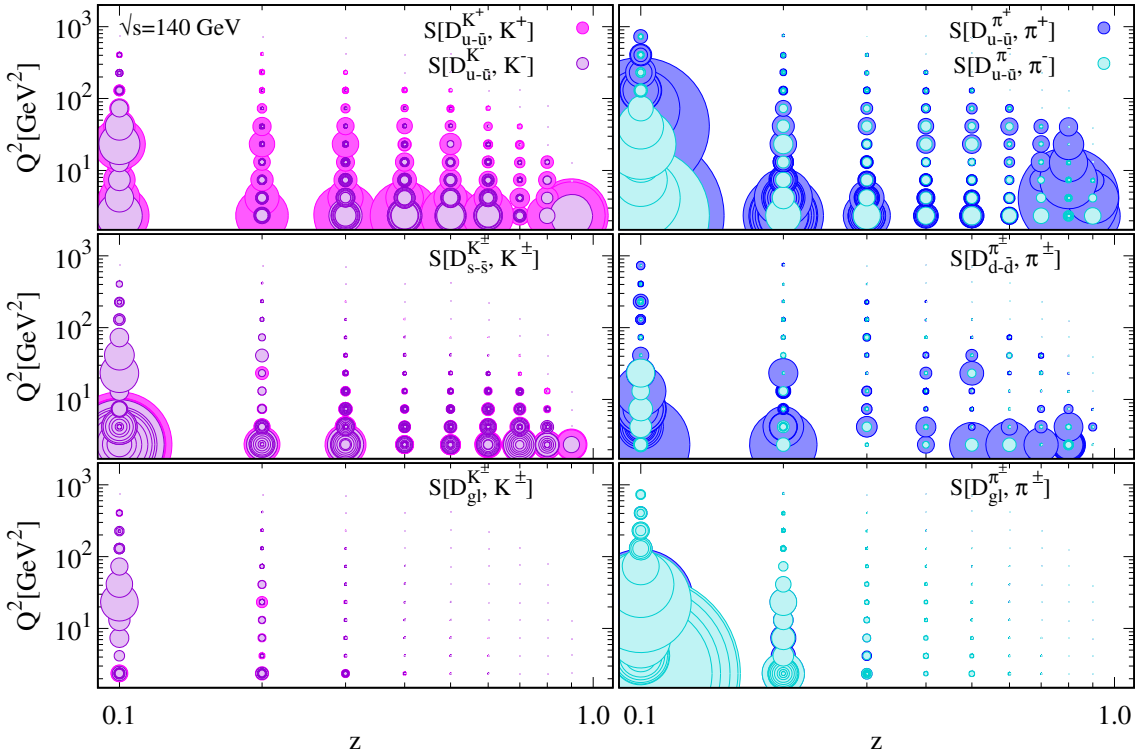

FIG. 17. Same as Fig. 16, for the nonsinglet combination of FFs $D^{H^\pm}_{q-\bar{q}}$.

represented in the plots by the dashed (black) lines with light gray bands. The reweighted results are plotted as solid (green) lines with dark gray uncertainty bands. The upper panels correspond to the u and $\bar{u}$ quark distributions, while the lower panels show the analogous result for the d and $\bar{d}$ quark.

The most noticeable feature in the plots is the significant reduction in the uncertainty bands. The inclusion of the EIC

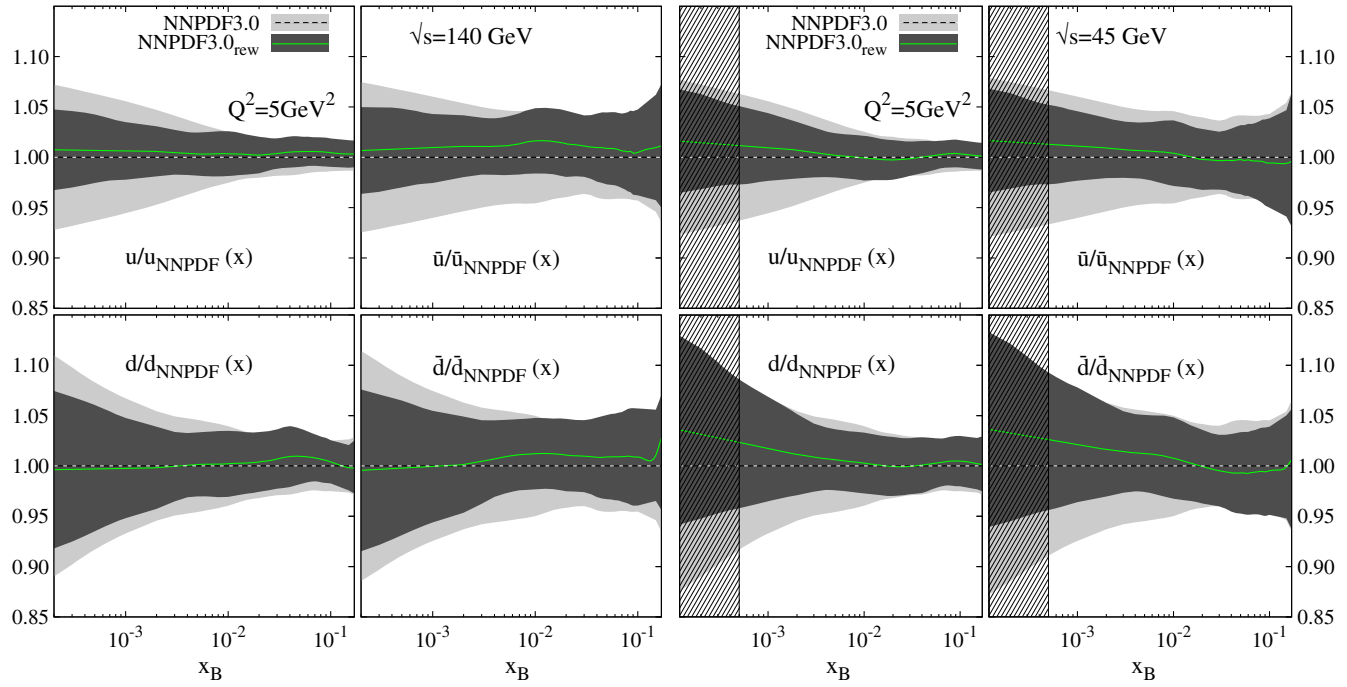

FIG. 18. Reweighting of NNPDF3.0 NLO replicas for the u and $\bar{u}$ quark distribution (upper panels) and d and $\bar{d}$ quark distribution (lower panels) with EIC pseudodata. The four panels on the left-hand side correspond to $\sqrt{s} = 140$ GeV pseudodata, while those on the right-hand side are for $\sqrt{s} = 45$ GeV. The shaded area is the region of x_B not covered by the latter energy configuration. The distributions are normalized to the NNPDF3.0 best fit. The solid (green) lines and dark gray bands represent the results for the distributions after the reweighting procedure and the corresponding uncertainty bands, respectively. All results are shown at a scale of $Q^2 = 5$ GeV^2.

pseudodata leads to a reduction of the uncertainty of order 30% for the up quark, driven by the new kaon and pion data, and of order 20% for the down quark, led by the pion data. It is also worth noticing that the kinematic region where the impact of the SIDIS pseudodata is most important is precisely the region $x_B < 10^{-2}$, as anticipated from the sensitivity coefficients calculation depicted in Fig. 12. As stated in the previous section, in spite of the high correlation between the pion cross section and the (anti)up-quark distribution for higher values of x_B, the inclusion of the pseudodata through the reweighting procedure hardly modifies the distributions in that kinematic region. Indeed, while a smaller impact for the high-x_B region was expected according to the sensitivity coefficients, the fact that the distributions are hardly modified in that kinematic configuration is the result of the increasing uncertainty associated with the FFs. As mentioned in Sec. IV, the theoretical uncertainty coming from the FFs must be included in the reweighting procedure, thus attenuating the impact of the pseudodata in the regions where these uncertainties become larger than those of the PDFs.

In Fig. 19, we show the pseudodata estimates for the production of positively charged pions as a function of x_B for representative bins of Q^2 and z. The pseudodata are presented in a (Data-Theory)/Theory plot together with the theoretical uncertainties for the cross-section estimate coming from the PDFs (light blue band) and from the FFs (dark blue band). Clearly, while the uncertainties propagated from the FFs are roughly independent of x_B, those coming from the PDFs grow for smaller values of x_B, since at these values the PDFs are considerably less well known than for the valence region. Naturally, the FF uncertainties limit the impact of the reweighting process in the kinematic region where the PDFs uncertainties are comparatively better determined. Iterating the reweighting procedure, as was demonstrated in Ref. [7] with actual SIDIS data, would yield more accurate FFs, which in turn would constrain the PDFs better. In any case, we see from this first step of the iterative procedure that the impact on the distributions is quite significant. Eventually, a combined PDF and FF global fit would yield in a single, albeit more involved step, a similar result.

The results with pseudodata generated for the lower c.m.s. energy of 45 GeV, on the right-hand side, show that the reduction in the uncertainty bands is not as large as in the case of the higher c.m.s. energy. Nevertheless, the pseudodata for this configuration still imposes sizable constraints on the distributions. The reweighting with this pseudodataset leads to a reduction in the uncertainty of the order of 20% in the case of the u and $\bar{u}$ quarks, and around 10% for the d- and $\bar{d}$-quark distributions. At variance with

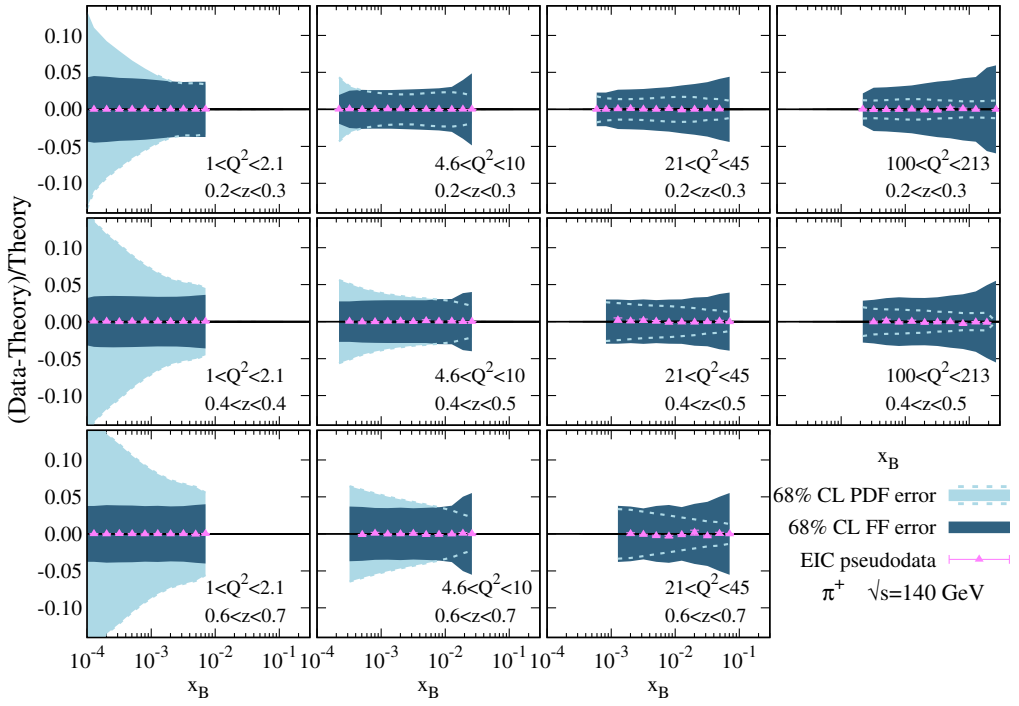

FIG. 19. Pseudodata estimates for the production of π^+ at $\sqrt{s} = 140$ GeV as a (Data-Theory)/Theory plot. The bands represent the uncertainty in the theoretical estimate coming from the PDFs (light blue) and from the FFs (dark blue). The data are plotted for representative bins of Q^2 and z, as a function of x_B. For those regions where the uncertainty coming from the FFs becomes larger than that of the PDFs, the error band of the latter is represented by the light blue dashed lines.

the higher c.m.s. energy, some deviations from the original best fit are produced for $x_B < 10^{-3}$, due to the absence of pseudodata points constraining the behavior of the replicas, which is fixed by the higher-x_B data.

As for the higher c.m.s. energy, the kinematic region constrained by the inclusion of the new data at lower c.m.s. energy coincides with the region of larger values of the sensitivity coefficient, now restricted to $10^{-3} < x_B < 10^{-2}$. Once again, it should be noticed that whereas the sensitivity coefficients suggest a more moderate impact for the higher-x_B region, the completely unmodified distributions are a result of the inclusion of the growing theoretical uncertainties coming from the FFs in the reweighting, which dilute the constraining power of the new dataset.

Similar results are obtained for the (anti)strange-quark distribution, which is depicted in Fig. 20 (upper-left panel), together with the flavor (upper-right and lower-left panels) and charge (lower-right panel) symmetry breaking. Again, the four panels on the left-hand side correspond to a set of pseudodata at a c.m.s. energy of $\sqrt{s} = 140$ GeV, while those on the right-hand side correspond to the $\sqrt{s} =$ 45 GeV set. As could be expected from the relatively poor determination of the strange-quark content of the proton, the most striking feature is an even more noticeable reduction in the uncertainty for the s-quark distribution, which is of the order of 75% for momentum fractions below 10^{-2}, driven by the kaon data through the reweighting.

The reduction in uncertainty of the strange-quark content of the proton has also a very significant impact on the constraints for the so-called *strange ratio*, shown in the upper-right panel, which has been actively discussed in connection to recent LHC measurements. Our result indicates that EIC SIDIS data would be able to further constrain the x_B dependence of the ratio, suggesting a rather asymmetric scenario at high x_B, while favoring SU(3) flavor symmetry between the light quarks for lower values of the momentum fraction.

Regarding the isospin and charge asymmetries, shown in the lower panels, no significant improvements in the uncertainty estimates are found. On the other hand, no important deviations from the original value are observed, which is fully consistent with the fact that the pseudodata were generated from theoretical estimates already containing the same degree of symmetry breaking, and the procedure does not introduce any spurious imbalance between $\bar{u}$ and $\bar{d}$ and between s and $\bar{s}$.

The reweighting of the FF replicas yields comparable results in terms of impact, although with some specific features related to the FF extractions used as a starting point. In Fig. 21, we show the effect of reweighting a set of 10^5 replicas of the variants of the DSS14 and DSS17 sets of FFs (based on NNPDF3.0) for pions and kaons with EIC SIDIS pseudodata for the c.m.s. energy configuration of $\sqrt{s} = 140$ GeV. In both cases, the sets of FF replicas are

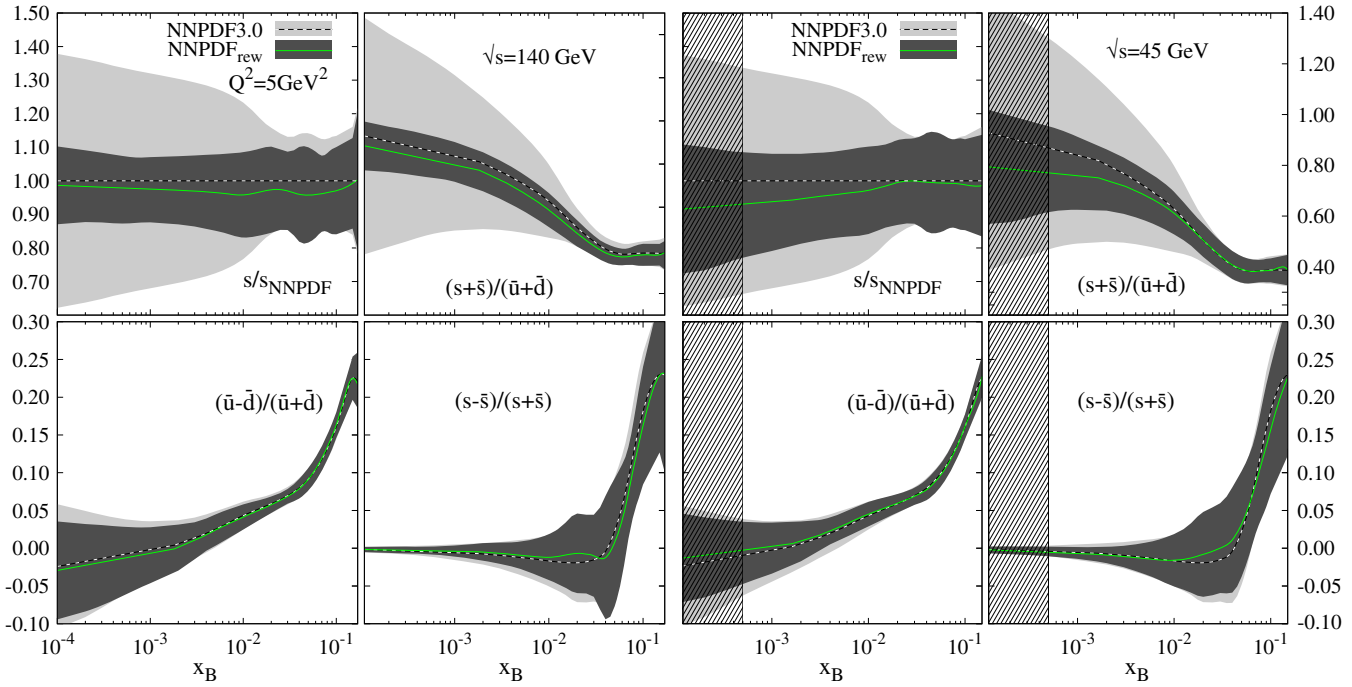

FIG. 20. The same as Fig. 18, but for the strange-quark distribution (upper panels) and for the PDF combinations sensitive to charge and isospin symmetry breaking (lower panels). Again, the results are shown at a scale of $Q^2 = 5$ GeV^2 and are normalized to the NNPDF3.0 best fit.

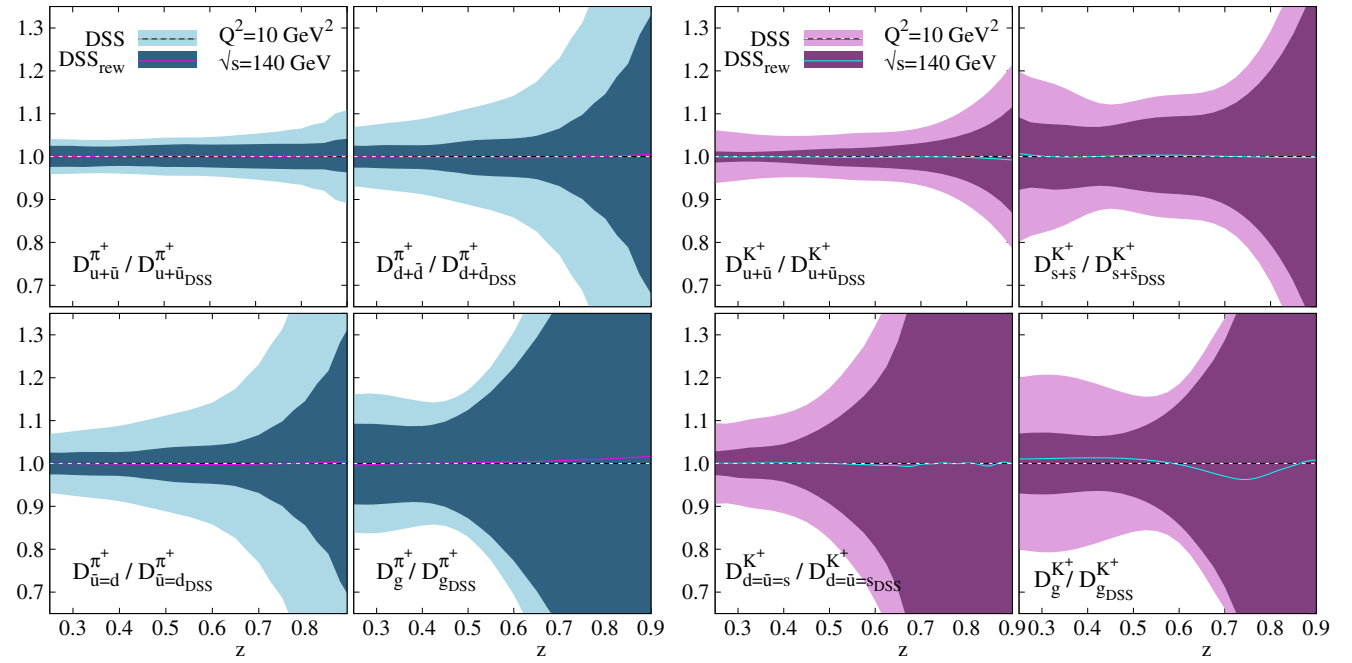

FIG. 21. Reweighting of the DSS NLO parton-to-pion and the parton-to-kaon fragmentation function replicas for the combinations $q + \bar{q}$ associated with the final hadron valence quarks (upper panels), as well as for the unfavored flavors of quarks and gluons (lower panels) with EIC pseudodata of c.m.s. energy $\sqrt{s} = 140$ GeV. As in Figs. 18 and 20, the results are normalized to the DSS best fit. In the case of parton-to-pion FFs, the modified distributions are represented by the pink line, while their modified uncertainties are represented by the dark blue band. Analogously, the original central value and uncertainty are given by the black and white dashed line and the light blue band, respectively. The inverse color scheme is used in the case of parton-to-kaon FFs. All results are shown at a scale of $Q^2 = 10$ GeV^2.

generated according to Eq. (9), from random variations in the parameter space, followed by an analogous application of the Bayesian inference procedure, described in previous sections. In both cases, a sufficiently large number of effective replicas survives after the reweighting exercise, with $N_{\mathrm{eff}}^{(\pi)} \approx 500$ and $N_{\mathrm{eff}}^{(K)} \approx 200$.

As in the case of the PDF reweighting, the plots show the modified distributions and their estimated uncertainties normalized to the reference value of DSS FFs, depicted by the black and white dashed lines. The modified parton-to-pion FFs are represented by the solid (magenta) lines, and their uncertainties by the darker (blue) bands. The inverse color scheme is used with the parton-to-kaon FFs, with light blue lines representing the modified FFs and violet bands representing their uncertainties. In both cases, the upper panels show the FFs of the *plus* combinations $D_{q+\bar{q}}^{H^+}$ associated with the final hadron valence quarks, whereas the lower panels correspond to the FFs for the unfavored light quarks and the gluon. Once again, since the pseudodata used for the reweighting procedure were generated by smearing the NLO estimate with DSS sets of FFs, no important deviation from the original sets is to be expected.

The improvement in the determination of both pion and kaon FFs is remarkable: In the case of parton-to-pion FFs, the reduction in the uncertainty of $D_{u+\bar{u}}^{\pi^+}$ is of the order of 25%, while for $D_{d+\bar{d}}^{\pi^+}$, the reduction is of the order of 30%. Even more impressive is the effect on the FFs associated with unfavored quark flavors, which show a reduction in the uncertainty of approximately 60%. This important improvement is mainly due to the relatively poor constraints for the unfavored flavors in the global fits. Notice that $D_q^{\pi^+}$ is assumed to be the same for $\bar{u}$ and d in the global fit.

It is also worth mentioning the impact of the pseudodata on the gluon-to-pion fragmentation function for low values of z, which shows a reduction of the uncertainty of the order of 40%. In this case, the constraints come not only from the NLO contribution to the cross section associated with the hadronization of gluons, but also through the evolution equations, which depend critically on the gluon FF.

As in the case of the reweighting with PDF replicas, the reweighting of FF replicas necessarily involves the inclusion of the theoretical uncertainties coming from the PDFs. Once again, the relatively smaller impact of the reweighting in the region of high z is associated with the larger PDF uncertainties, which grow with z for a fixed value of $\{x_B, Q^2\}$, as can be seen in Fig. 19. Feeding the reweighting with improved PDFs in subsequent iterations would eventually exploit the full constraining power of the data.

Regarding the parton-to-kaon FFs, the results shown in Fig. 21 should be taken with some caution, since the much more rigid functional form assumed for some of the DSS kaon FFs could be too restrictive for the generation of faithful replicas. In fact, the reweighting results in a significantly lower number of effective replicas compared to the pion reweighting. While the constraints on the FFs for the combinations $u + \bar{u}$, $s + \bar{s}$ are once again impressive, with reductions in the uncertainties of $D_{u+\bar{u}}^{K^+}$ and $D_{s+\bar{s}}^{K^+}$ around 70% and 60%, respectively, the less flexible parametrizations for the unfavored FFs and gluons could translate into an artificial reduction of the uncertainties. The comparison to actual SIDIS data instead of simulated cross sections generated from the DSS sets will eventually indicate the need of a new FF fit with more flexibility or different flavor symmetry assumptions. In any case, the results clearly show that the EIC SIDIS measurements largely exceed in precision the current global analysis and therefore have a significant potential for the improvement of FF extractions.

VI. SUMMARY

The semi-inclusive production of hadrons in deep-inelastic electron-proton scattering offers a remarkably versatile tool to probe both the flavor content of the proton and the way in which the different parton flavors confine into final-state hadrons. QCD factorization allows us to model the corresponding cross sections in terms of nonperturbative parton distribution and fragmentation functions in such a way that precise cross-section measurements impose very stringent constraints on these distributions.

The key advantage of SIDIS data in the determination of the PDFs lies in the fact that the flavor composition of the final-state hadrons probes a specific combination of partonic flavors, giving access to flavor-dependent information that is entangled in more inclusive measurements. Consequently, the unprecedented precision and kinematic coverage of SIDIS measurements at a future EIC will certainly enhance our knowledge on PDFs and FFs, and provide new insights into the inner structure of the nucleon, and the interactions among its most basic constituents. In this paper, we have made quantitative assessment of the improvements.

Despite the technical difficulties involved in a simultaneous extraction of both PDFs and FFs, techniques based on Bayesian inference allow us to refine our knowledge on the nonperturbative distributions, including the critical information coming from SIDIS data. Through the implementation of reweighting techniques, we studied in detail the constraints that measurements at the future EIC would impose on the parton distribution functions of the proton, as well as on the parton-to-hadron fragmentation functions, by using simulated data with realistic uncertainties.

We confirm the remarkable impact that EIC SIDIS data would have on the PDFs, especially on those of light quarks of radiative origin, which are comparatively less constrained than their valence counterparts. Our study suggests that outstanding reductions in the uncertainties of these distributions can be obtained, which we estimate to be of

the order of 75% in the case of the strange-quark content of the proton, 30% for the up quark and 20% for the down quark (for the most energetic configuration of $\sqrt{s} = 140$ GeV). In addition, our results indicate that it will be possible to constrain the strong parton momentum fraction dependence of the *strangeness ratio*, and have complementary estimates of the charge symmetry breaking.

We also find that the most significative effect on the parton distributions will be achieved with the much wider kinematic range covered by the EIC running at a large c.m.s. energy, for which more stringent constraints are found.

Regarding the fragmentation functions, we have also estimated the kinematic configurations where the EIC data could enhance the precision of FFs in future global analyses, as well as the improvement in the precision of these distributions. Our results indicate that EIC SIDIS data would have a significant effect on the determination of the FFs, complementing the present measurements since they span a wider kinematic range than that currently probed.

It is noted that the alternative reweighting of PDFs and of FFs shown here could be seen as the first step in an iterative processes equivalent to a combined global fit, that would eventually exploit the full constraining power of the forthcoming EIC semi-inclusive data. The significant impact already obtained in these first steps highlights the importance that the forthcoming measurements at the EIC will have on the determination of the nonperturbative PDFs and FFs, taking them to a new standard in precision, and therefore refining our picture of the partonic structure of matter.

ACKNOWLEDGMENTS

We warmly acknowledge Pia Zurita and Marco Stratmann for interesting comments and suggestions and their help with the reweighting methodologies. This work was supported in part by CONICET and ANPCyT. C. V. H. acknowledges the support from the Basque government (Grant No. IT956-16) and the Ministry of Economy and Competitiveness (MINECO) (Juan de la Cierva), Spain.

[1] J. Butterworth *et al.*, J. Phys. G **43**, 023001 (2016).
[2] A. Metz and A. Vossen, Prog. Part. Nucl. Phys. **91**, 136 (2016).
[3] J. Rojo *et al.*, J. Phys. G **42**, 103103 (2015).
[4] J. Rojo, Proc. Sci. , **DIS2016**, 018 (2016).
[5] R. P. Feynman, *Photon-Hadron Interactions* (W. A. Benjamin, Reading, MA, 1972).
[6] R. D. Field and R. P. Feynman, Phys. Rev. D **15**, 2590 (1977).
[7] I. Borsa, R. Sassot, and M. Stratmann, Phys. Rev. D **96**, 094020 (2017).
[8] J. P. Lees *et al.* (*BABAR* Collaboration), Phys. Rev. D **88**, 032011 (2013).
[9] M. Leitgab *et al.* (Belle Collaboration), Phys. Rev. Lett. **111**, 062002 (2013).
[10] A. Airapetian *et al.* (HERMES Collaboration), Phys. Rev. D **87**, 074029 (2013).
[11] C. Adolph *et al.* (COMPASS Collaboration), Phys. Lett. B **767**, 133 (2017).
[12] G. Agakishiev *et al.* (STAR Collaboration), Phys. Rev. Lett. **108**, 072302 (2012).
[13] B. B. Abelev *et al.* (ALICE Collaboration), Phys. Lett. B **736**, 196 (2014).
[14] A. Accardi *et al.*, Eur. Phys. J. A **52**, 268 (2016).
[15] http://icfa-bd.kek.jp/Newsletter74.pdf.
[16] E. C. Aschenauer, R. Sassot, and M. Stratmann, Phys. Rev. D **92**, 094030 (2015).
[17] E. C. Aschenauer *et al.*, arXiv:1602.03922.
[18] R. D. Ball, V. Bertone, F. Cerutti, L. Del Debbio, S. Forte, A. Guffanti, J. I. Latorre, J. Rojo, and M. Ubiali (NNPDF Collaboration), Nucl. Phys. **B849**, 112 (2011); **B854**, 926(E) (2012); **B855**, 927(E) (2012).
[19] R. D. Ball, V. Bertone, F. Cerutti, L. Del Debbio, S. Forte, A. Guffanti, N. P. Hartland, J. I. Latorre, J. Rojo, and M. Ubiali, Nucl. Phys. **B855**, 608 (2012).
[20] H. Paukkunen and P. Zurita, J. High Energy Phys. 12 (2014) 100.
[21] N. Armesto, J. Rojo, C. A. Salgado, and P. Zurita, J. High Energy Phys. 11 (2013) 015.
[22] V. Bertone, N. P. Hartland, E. R. Nocera, J. Rojo, and L. Rottoli (NNPDF Collaboration), Eur. Phys. J. C **78**, 651 (2018).
[23] B. T. Wang, T. J. Hobbs, S. Doyle, J. Gao, T. J. Hou, P. M. Nadolsky, and F. I. Olness, Phys. Rev. D **98**, 094030 (2018).
[24] W. Furmanski and R. Petronzio, Z. Phys. C **11**, 293 (1982).
[25] D. Graudenz, Nucl. Phys. **B432**, 351 (1994).
[26] D. de Florian, M. Stratmann, and W. Vogelsang, Phys. Rev. D **57**, 5811 (1998).
[27] D. de Florian, R. Sassot, M. Epele, R. J. Hernandez-Pinto, and M. Stratmann, Phys. Rev. D **91**, 014035 (2015).
[28] D. de Florian, M. Epele, R. J. Hernandez-Pinto, R. Sassot, and M. Stratmann, Phys. Rev. D **95**, 094019 (2017).
[29] D. de Florian, R. Sassot, and M. Stratmann, Phys. Rev. D **75**, 114010 (2007).
[30] A. Daleo, D. de Florian, and R. Sassot, Phys. Rev. D **71**, 034013 (2005).
[31] A. Daleo, C. A. Garcia Canal, and R. Sassot, Nucl. Phys. **B662**, 334 (2003).
[32] A. Daleo and R. Sassot, Nucl. Phys. **B673**, 357 (2003).

[33] T. Sjöstrand, P. Edén, C. Friberg, L. Lönnblad, G. Miu, S. Mrenna, and E. Norrbin, Comput. Phys. Commun. **135**, 238 (2001).
[34] T. Sjöstrand, S. Mrenna, and P. Skands, Comput. Phys. Commun. **178**, 852 (2008).
[35] R. D. Ball *et al.* (NNPDF Collaboration), J. High Energy Phys. 04 (2015) 040.
[36] P. M. Nadolsky, H. L. Lai, Q. H. Cao, J. Huston, J. Pumplin, D. Stump, W. K. Tung, and C.-P. Yuan, Phys. Rev. D **78**, 013004 (2008).
[37] R. D. Ball, L. Del Debbio, S. Forte, A. Guffanti, J. I. Latorre, A. Piccione, J. Rojo, and M. Ubiali (NNPDF Collaboration), Nucl. Phys. **B809**, 1 (2009); **816**, 293(E) (2009).
[38] A. Guffanti and J. Rojo, Nuovo Cimento C **033**, 65 (2010).