JHEP

Published for SISSA by Springer

Received: February 1, 2022
Accepted: March 22, 2022
Published: April 14, 2022

New tool for kinematic regime estimation in semi-inclusive deep-inelastic scattering

The Jefferson Lab Angular Momentum (JAM) collaboration

M. Boglione,[a] M. Diefenthaler,[b] S. Dolan,[c] L. Gamberg,[c] W. Melnitchouk,[b] D. Pitonyak,[d] A. Prokudin,[c,b,1] N. Sato[b] and Z. Scalyer[c,2]

[a] *Dipartimento di Fisica, Università di Torino, INFN-Sezione Torino, Torino, Italy*

[b] *Jefferson Lab, Newport News, VA 23606, U.S.A.*

[c] *Science Division, Penn State University Berks, Reading, PA 19610, U.S.A.*

[d] *Department of Physics, Lebanon Valley College, Annville, PA 17003, U.S.A.*

E-mail: elena.boglione@to.infn.it, mdiefent@jlab.org, skd140@psu.edu, lpg10@psu.edu, wmelnitc@jlab.org, pitonyak@lvc.edu, prokudin@jlab.org, nsato@jlab.org, scalyer@p4flight.com

Abstract: We introduce a new phenomenological tool based on momentum region indicators to guide the analysis and interpretation of semi-inclusive deep-inelastic scattering measurements. The new tool, referred to as "affinity", is devised to help visualize and quantify the proximity of any experimental kinematic bin to a particular hadron production region, such as that associated with transverse momentum dependent factorization. We apply the affinity estimator to existing HERMES and COMPASS data and expected data from Jefferson Lab and the future Electron-Ion Collider. We also provide an interactive notebook based on Machine Learning for fast evaluation of affinity.

Keywords: Specific QCD Phenomenology, Deep Inelastic Scattering or Small-X Physics

ArXiv ePrint: 2201.12197

[1] On sabbatical leave at Temple University, Philadelphia, PA, U.S.A.
[2] Now at Prepared for Flight, LLC, York, PA 17402, U.S.A.

https://doi.org/10.1007/JHEP04(2022)084

Contents

1 Introduction

Providing a precise partonic description of hadronic structure from quantum chromodynamics (QCD) factorization theorems has been a topic of great interest for over half a century. From inclusive and semi-inclusive deep-inelastic scattering experiments we know that hadrons have a complex internal structure involving quarks, antiquarks and gluons (generically partons) and their interactions. In addition to the partons' collinear momenta, which is highly correlated with the direction of a fast-moving parent hadron, partons also possess intrinsic transverse motion and structure. Several types of high-energy scattering measurements are known to be sensitive to this intrinsic transverse momentum, including semi-inclusive deep-inelastic leptoproduction of hadrons h [1–3], $\ell N \to \ell' h X$, inclusive electron-positron annihilation to almost back-to-back hadrons [4–6], $e^+e^- \to h_1 h_2 X$, and Drell-Yan lepton-pair or weak gauge boson production in NN scattering [7], $N N \to \{\ell^+\ell^-, Z, W^\pm\} X$, where N represents a proton or neutron (deuteron) in the initial state.

Interpreting these measurements in terms of QCD requires factorization theorems that are valid for the process and the kinematic reach of the measurement. For transverse momentum dependent (TMD) scattering reactions, TMD factorization [8–11] describes these processes in terms of a collinear perturbative (hard) scattering cross section and nonperturbative TMD parton distribution functions (TMD PDFs) and fragmentation functions (TMD FFs) (collectively referred to as "TMDs") [1, 2, 12]. A condition implicit in the proof of TMD factorization in semi-inclusive deep-inelastic scattering (SIDIS), where at

JHEP04(2022)084

leading order the final state hadrons are fragments of the struck quark, is that a clear separation exists between the momentum of the struck quark in the target nucleon and that of partons that are spectators to the hard collision. In this framework the fragmentation of a quark into hadrons is independent of the production mechanism of the quark [13]. Fragmentation is thus described by a function of the momentum fraction of the quark carried by the produced hadron, which is independent of the momentum fraction of the parent nucleon carried by the struck quark. In this scenario the hadron is said to be in the *current* fragmentation region.

By contrast, if the produced hadron moves in nearly the same direction as the target, the hadron is said to be in the *target* fragmentation region, and the relevant factorization theorem is then formulated in terms of fracture functions [14–17]. A clear distinction between the current and target fragmentation regions requires a sufficiently large separation in the momentum of the current and target fragments, and for this purpose it is convenient to use *rapidity* to delineate these regions. Berger [13, 18] provided a specific rapidity gap criterion to study the dynamics of quark fragmentation in the current fragmentation region, although in practice the delineation into distinct current, target, and central fragmentation regions is rarely sharp [13, 18–21].

In addition, partons that populate the rapidity gap between current and target regions also fragment into hadrons, and these form the *central* fragmentation region [21]. This region can be referred to as a *soft-central* region, where soft gluons emitted in the cascade after the hard scattering give important contribution to centrally produced hadrons. By soft, we mean partons with all four components of the momentum being of order $\mathcal{O}(m)$, where m is a typical hadron mass.

Following a careful examination of the approximations involved in QCD factorization [11], recently Boglione et al. [20, 22] introduced new quantitative criteria for classifying fragmentation regions in terms of various ratios, R_i, of partonic and hadronic momenta, which are particularly useful at small and moderate values of the momentum transfer, Q. Traditionally, the applicability of TMD factorization in the current region has been linked solely to the small size of the transverse momentum of the produced hadron P_{hT} and the rapidity region. It was found [20, 22], however, that the applicability can also depend on so-called *region indicators*, characterized by the ratios R_i that reflect the proximity of any given kinematic configuration to a particular partonic region of SIDIS.

Typically, in TMD phenomenology, data are filtered by the value of the hadron transverse momentum P_{hT} in the Breit frame [23, 24], or by the photon transverse momentum in the hadron-hadron frame, $q_T \simeq P_{hT}/z_h$ [25, 26], where $z_h = P \cdot P_h / P \cdot q$, with P the momentum of the initial hadron and q is the momentum transfer from the incident lepton. It was found [20, 22], however, that cuts on P_{hT} or q_T applied in analyses of SIDIS data may not be sufficient to guarantee that the data, at given kinematics, are uniquely within the current fragmentation region. Since the observed hadrons can be produced via different physical mechanisms, identifying SIDIS cross sections in a kinematic region corresponding to TMD factorization requires particular attention. It is crucial, therefore, to analyze the role that data cuts play in discriminating the current region from the target and central fragmentation regions, and assess their impact on the extraction of TMDs from future

JHEP04(2022)084

SIDIS data from Jefferson Lab (JLab), COMPASS at CERN, and the future Electron-Ion Collider (EIC). Indeed, application of the region indicators was already recently discussed by the HERMES Collaboration [27].

In this paper we implement the region indicators introduced in refs. [20, 22] to quantify the confidence of the proximity of SIDIS observables to a particular physical mechanism. The new tool, which we refer to as "affinity", $\mathcal{A}$, combines information from a variety of partonic configurations and the resulting ratios, R_i, into a single estimate of the proximity to a particular hadron production mechanism, which ranges from 0% to 100%. We carry out the affinity analysis for kinematics relevant to existing and future facilities, and provide an affinity profile across the phase space for each kind of physical mechanism for hadron production in SIDIS [28]. Ultimately, these results will provide a well-defined methodology for determining the degree of confidence that a given kinematical configuration may be described in terms of TMDs, given assumptions about the partonic kinematics.

We begin in section 2 by briefly recalling the region indicators. To assess the proximity of the data at given kinematics to a specific physical mechanism, in section 3 we introduce the affinity, $\mathcal{A}$, as a global estimator. In section 4 we apply the new affinity tool to the analysis of existing data from the HERMES and COMPASS experiments, and discuss the analysis of data expected from Jefferson Lab and the future EIC. In section 5 we present the results of training a neural network using the `TensorFlow` package, with various choices for the underlying demarcation of regions, and introduce a `Google Colab` interactive notebook for visualizing the affinity at EIC kinematics. Finally, in section 6 we summarize our findings and discuss possible future applications of this analysis.

2 Region indicators

The methodology of the region indicators R_i was presented in refs. [20, 22] to help delineate different hadron production mechanisms by including some information on the underlying momentum flow in the partonic subprocess with visualization tools that were made available at https://sidis.herokuapp.com. In this section we review the definitions of the region indicators, and discuss how they characterize the current, target and central regions of kinematics.

2.1 Current, target, and central regions

A typical diagram for the momentum flow in the current fragmentation region of the SIDIS process is sketched in figure 1. The figure illustrates the scattering of an incident lepton via the exchange of a virtual photon of momentum q (with $q^2 \equiv -Q^2$) from an initial parton of momentum k_i in a nucleon of momentum P to at least one hadronizing parton k_f, with k_X the total momentum of all other unobserved partons from the partonic subprocess. A detailed discussion of SIDIS kinematics can be found in ref. [22].

The current region is associated with the fragmentation of the parton after it has absorbed the incoming virtual photon. The outgoing parton fragments into the detected hadron of momentum P_h, which moves in approximately the same direction and with similar rapidity as the fragmenting parton if the transverse momentum P_{hT} is small. Consequently,

JHEP04(2022)084

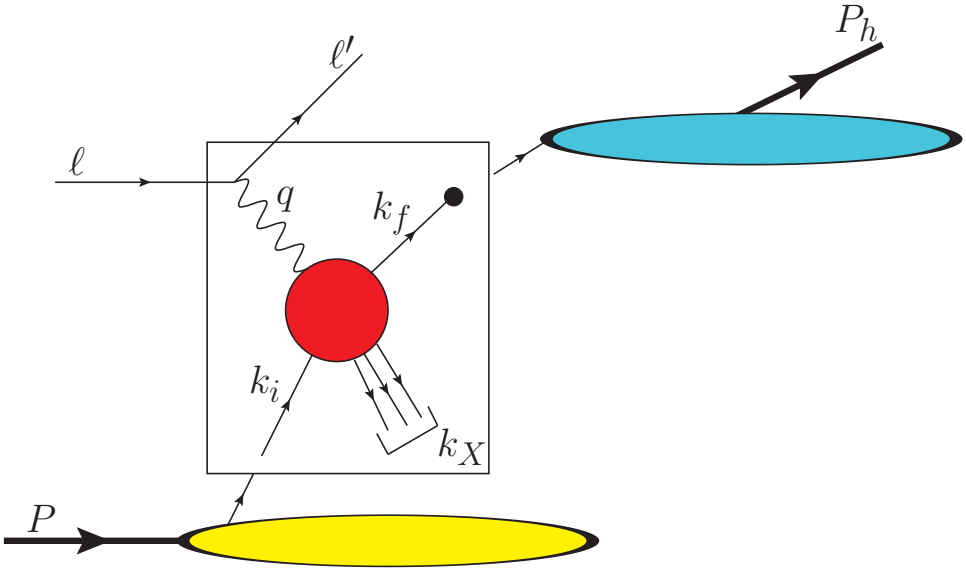

Figure 1. Sketch of the SIDIS process, with scattering of an incident lepton (momentum ℓ) via the exchange of a virtual photon (q) from a parton (k_i) inside a nucleon (P), with the scattered parton (k_f) fragmenting to a hadron (P_h) in the final state and unmeasured hadronic debris (k_X). The lower (yellow) blob represents the residual system after removal of the parton from the incoming nucleon, while the upper right (cyan) blob represents the fragmentation of the outgoing parton into the observed hadron. The rectangle envelopes the parton level subprocess, and the arrows represent momentum flow. The black dot indicates the parton associated with the observed hadron.

in our momentum coordinate conventions the produced hadrons are in the region of negative rapidity. In this case, well-established TMD factorization theorems are valid — see refs. [8–11, 29–35]. Hard QCD radiation may produce a large hadronic transverse momentum P_{hT} in the current region, which would shift the rapidity of the hadron towards central or positive values. In such cases a treatment based on collinear QCD factorization theorems [9, 11, 36] is appropriate.

The target region is associated with the fragmentation of spectator partons, which originate in the target nucleon P but do not experience the hard collision with the virtual photon. These partons continue to move predominantly in the direction of the parent nucleon, and the products of their hadronization are found at positive values of the rapidity. The corresponding momentum flow picture here would be similar to that in figure 1, but with the produced hadron originating from the lower (yellow) blob representing the remnant of the incoming nucleon.

The central region is characterized by the production of hadrons that are neither the products of a hard scattering nor associated in any obvious way with a current quark or target remnant direction. These hadrons are fragmentation products of quarks and gluons that fill the central rapidity region between the struck parton and the target hadron remnants [14, 21, 37]. While the identification of current and target regions is well defined by criteria which establish a clear rapidity separation between the collinear and soft subgraphs in large-Q asymptotics of factorization [11], in reality these rapidity gaps are filled

JHEP04(2022)084

by centrally produced hadrons [37]. These are the phenomena that are approximated in Monte Carlo event generators by string [38] or cluster hadronization [39]. Indeed, proofs of factorization do not yet account for graphical structures of cluster and string hadronization [37] that characterize the production of hadrons in the central region.

On the other hand, the fastest moving hadrons in opposite hemispheres in string-like fragmentation [40], being separated by space-like distances, are in line with the independent hadronization that one obtains in proofs of large-Q asympotics in TMD factorization. Additionally, as pointed out by Collins [37], all intermediate rapidity regions between the current and target contribute at leading power in the hard scale, and it is imperative to classify the momentum regions associated with the central region. Thus, pinning down the central region is considerably more complicated and is the subject of active research [21].

One could, for instance, employ a conservative definition of the current and target regions by selecting stringent criteria for the smallness of P_{hT} (or q_T) or the largeness of the rapidity gap. This is not feasible in the central region, however, where hadronization produces approximately uniform distributions in rapidity. To define the central region one could consider identifying the region by exclusion, such that any kinematic configuration that is not strictly in the current or target regions falls within this. Alternatively, a more conservative approach would omit slices of the process phase space by including in the central region only those configurations that are genuinely soft, according to some specific criteria (see ref. [21] for a discussion). In either case, it is clear that in practice the boundaries of the central region remain rather "blurred".

To classify the relevant kinematic regions, refs. [20, 22] proposed several ratios of partonic and hadronic momenta, as we summarize in the following. To begin with, in order to ensure a partonic interpretation of the process, the ratio R_0 of partonic momenta to the hard scale Q^2, referred to as the *general hardness* ratio, was introduced,

$$R_0 \equiv \max\left(\left|\frac{k_i^2}{Q^2}\right|, \left|\frac{k_f^2}{Q^2}\right|, \left|\frac{\delta k_T^2}{Q^2}\right|\right) . \tag{2.1}$$

Here, δk_T^2 is a parameter that characterizes the size of the intrinsic transverse momentum of the parton, which is $\mathcal{O}(m^2)$, where m is a typical hadron mass scale of the reaction. The smallness of R_0, $R_0 \ll 1$, is the minimal requirement needed for the application of a partonic description of the SIDIS process [20, 22].

To isolate the current fragmentation region from the target and central fragmentation regions, we define the *collinearity* ratio [20], R_1, by

$$R_1 \equiv \frac{P_h \cdot k_f}{P_h \cdot k_i} . \tag{2.2}$$

The collinearity must be small for current fragmentation and large for target and central fragmentation. To further distinguish the target region, we also consider the *target proximity* ratio R_1',

$$R_1' \equiv \frac{P_h \cdot P}{Q^2}, \tag{2.3}$$

which is expected to be small for target fragmentation [20, 22].

JHEP04(2022)084

2.2 TMD and collinear current regions

Historically, most phenomenological studies of SIDIS have focused on the current fragmentation region. The analysis of this region can be refined by introducing additional ratios to distinguish the ranges of applicability of descriptions based on QCD collinear and TMD factorization theorems [28].

For this purpose it is useful to introduce the *transverse hardness* ratio, R_2, defined as [22]

$$R_2 \equiv \frac{|k^2|}{Q^2}, \tag{2.4}$$

where $k \equiv k_f - q$. This ratio is relevant because the $2 \to 1$ scattering process $\gamma^* q \to q'$ dominates in the TMD regime, which applies if $|k^2|/Q^2 \simeq 0$. Moreover, as shown in ref. [22], one can write this ratio as

$$R_2 \approx (1 - \hat{z}_N) + \hat{z}_N \frac{q_T^2}{Q^2}\,, \tag{2.5}$$

in terms of the partonic variable $\hat{z}_N$, defined as the ratio of the "$-$" light-front momentum components of k_f and q in the Breit frame,

$$\hat{z}_N \equiv \frac{k_f^-}{q^-} = \frac{z_N}{\zeta}\,. \tag{2.6}$$

The hadronic fragmentation variable z_N here is defined as $z_N = P_h^-/q^-$, with $\zeta = P_h^-/k_f^-$ the momentum fraction of the parton carried by the produced hadron in the Breit frame (see ref. [22] for further details). The smallness of R_2 is needed to establish the existence of the TMD current fragmentation region. Note that if $q_T^2/Q^2 \sim 1$, then $R_2 \sim 1$ for both large and small values of $\hat{z}_N$, while if $q_T^2/Q^2 \ll 1$ and $\zeta \sim z_N$, as in the TMD current fragmentation region, then the transverse hardness ratio becomes $R_2 \ll 1$. On the other hand, a large value for the transverse hardness ratio R_2 would generally indicate the dominance of QCD subprocesses, such as gluon radiation, $\gamma^* q \to g q'$, to generate the observed transverse momentum P_{hT}. In refs. [25, 26] the ratio q_T^2/Q^2 was used to filter data appropriate for a TMD factorization description. Following ref. [25], which performs an N^{3}LO simultaneous fit of Drell-Yan and SIDIS data, in the present analysis we use the cuts

$$Q > 2 \text{ GeV}, \quad \frac{q_T}{Q} < 0.25 \tag{2.7}$$

to select the data.

The region of large transverse momentum is characterized by a ratio similar to those above. In analogy with the general hardness ratio R_0 in eq. (2.1), we introduce the *large transverse momentum* ratio, R_4,

$$R_4 \equiv \max\left(\left|\frac{k_i^2}{k^2}\right|, \left|\frac{k_f^2}{k^2}\right|, \left|\frac{\delta k_T^2}{k^2}\right|, \left|\frac{k_{iT}^2}{k^2}\right|\right) \tag{2.8a}$$

$$= \frac{1}{R_2}\max\left(\left|\frac{k_i^2}{Q^2}\right|, \left|\frac{k_f^2}{Q^2}\right|, \left|\frac{\delta k_T^2}{Q^2}\right|, \left|\frac{k_{iT}^2}{Q^2}\right|\right), \tag{2.8b}$$

JHEP04(2022)084

Ratio		Definition
R_0	general hardness	$\max\left(\left\lvert\frac{k_i^2}{Q^2}\right\rvert, \left\lvert\frac{k_f^2}{Q^2}\right\rvert, \left\lvert\frac{\delta k_T^2}{Q^2}\right\rvert\right)$
R_1	collinearity	$\frac{P_h \cdot k_f}{P_h \cdot k_i}$
R_1'	target proximity	$\frac{P_h \cdot P}{Q^2}$
R_2	transverse hardness	$\frac{\lvert k^2\rvert}{Q^2}$
R_3	spectator virtuality	$\frac{\lvert k_X^2\rvert}{Q^2}$
R_4	large transverse momentum	$\max\left(\left\lvert\frac{k_i^2}{k^2}\right\rvert, \left\lvert\frac{k_f^2}{k^2}\right\rvert, \left\lvert\frac{\delta k_T^2}{k^2}\right\rvert, \left\lvert\frac{k_{iT}^2}{k^2}\right\rvert\right)$

Table 1. Summary of the diagnostic ratios and their definitions used for identifying different fragmentation regions in SIDIS. The particle momenta are defined as in figure 1.

using the definition of R_2 in eq. (2.4). Transverse momentum can be said to be generated by perturbative mechanisms if $R_4 \ll 1$. The smallness of R_4 will be used in this analysis to determine the extent of the collinear QCD current region, instead of the requirement that the transverse hardness ratio R_2 be large.

We can also explore the region associated with gluon radiation in more detail by introducing the *spectator virtuality* ratio, R_3, defined by

$$R_3 \equiv \frac{|k_X^2|}{Q^2}\,, \tag{2.9}$$

where $k_X = k_i + q - k_f$. Small values of R_3 correspond to $2 \to 2$ parton kinematics, and the corresponding region may be explained by low-order (LO) perturbative QCD (pQCD) calculations. On the other hand, large R_2 and R_3 values correspond to $2 \to n$ scattering, where $n \geq 3$, so that higher-order (HO) pQCD calculations are needed to describe data in this region.

Finally, the region of matching of TMD and collinear factorizations is characterized by the presence of intermediate values of R_2, so that both the TMD and collinear pictures are approximately valid, and a smooth transition between these is possible. For completeness, in table 1 we summarize the definitions of all the ratios that act as region indicators in SIDIS analysis.

In addition to the transverse hardness ratio, it is also useful to consider the logarithm measure, $|\ln R_2|$, which is typical of the type of large logarithm that requires the q_T-resummation component from the Collins-Soper-Sterman (CSS) treatment of evolution

JHEP04(2022)084

Region	R_0	R_1	R_1'	R_2	R_3	R_4
TMD	small	small	$\times$	small	$\times$	$\times$
matching	small	small	$\times$	small	$\times$	$\times$
collinear	small	small	$\times$	large	small (LO pQCD) large (HO pQCD)	small
target	small	large	small	$\times$	$\times$	$\times$
central	small	not small	not small	small	$\times$	$\times$

Table 2. Examples of sizes of region indicator ratios corresponding to particular regions of SIDIS. The symbol "$\times$" denotes "irrelevant or ill-defined" (see text for further details).

when $R_2 \to 0$. If the logarithm measure $|\ln R_2|$ becomes larger than $\mathcal{O}(1)$, then q_T-resummation effects may need to be taken into account.

The resulting catalogue of possible values of region indicators is presented in table 2. As shown, the proximity of a given physical mechanism is characterized by the different sizes of the R_i ratios, which in turn depend not only on the external kinematics of the SIDIS reaction, but also on the internal active parton momenta. Since the latter are not physical observables, the use of R_i requires prior knowledge of the parton momenta, which can be inferred from nonperturbative treatments of QCD or from phenomenological analyses where the proximities of regions are estimated on the basis of agreement between data and theory.

3 Affinity

To facilitate the assessment of the proximity of data at a given set of kinematics to a specific physical mechanism, we introduce *affinity*, $\mathcal{A}$, as a global estimator using a Bayesian formulation,

$$\mathcal{A}(x_{\rm Bj}, Q^2, z_h, P_{hT}|\text{region}) = \int \mathrm{d}\{R_i\}\ \ \Theta(\{R_i\}|\ \text{region}) \times \int \mathrm{d}^4k_i\, \mathrm{d}^4k_f\, \mathrm{d}^4\delta k_{\rm T}\ \mathcal{P}(\{R_i\}|x_{\rm Bj}, Q^2, z_h, P_{hT}; k_i, k_f, \delta k_T)\ \pi(k_i, k_f, \delta k_T), \quad (3.1)$$

where $x_{\rm Bj} = Q^2/2P\cdot q$ is the Bjorken scaling variable. The second line of eq. (3.1) is a joint distribution for the SIDIS R_i indicators marginalized over a given choice for the prior distribution π, with $\mathcal{P}$ the conditional probability density for $\{R_i\}$. The latter is a function chosen according to the *prior beliefs* for the possible ranges of intrinsic partonic momenta. Given such a joint distribution, the affinity of a given SIDIS kinematic bin to a given region is defined by marginalizing the joint distribution over all possible values of $\{R_i\}$ with the *proximity function* Θ that selects a given region in R_i-space, according to table 2.

In practice, the implementation of eq. (3.1) requires the use of Monte Carlo methods, in which one must sample the four-vectors of the parton momenta. Using light-front coordinates, we parametrize the initial and final parton four-momenta as $k_i = (k_i^+, k_i^-, \boldsymbol{k}_{iT})$ and $k_f = (k_f^+, k_f^-, \boldsymbol{k}_{fT})$, and the intrinsic transverse momentum of the parton

JHEP04(2022)084

as $\delta k_T = (0, 0, \delta \boldsymbol{k}_T)$, with the components given by

$$k_i^+ = \frac{Q}{\sqrt{2}\,\hat{x}_N}, \quad k_i^- = \frac{\hat{x}_N}{\sqrt{2}\,Q}(k_i^2 + \boldsymbol{k}_{iT}^2), \quad \boldsymbol{k}_{iT} = (|\boldsymbol{k}_{iT}|\cos\varphi_{k_i},\ |\boldsymbol{k}_{iT}|\sin\varphi_{k_i}), \tag{3.2}$$

$$k_f^+ = \frac{1}{\sqrt{2}\,Q\hat{z}_N}(k_f^2 + (q_T\,\hat{z}_N\sin\varphi_h - |\delta\boldsymbol{k}_T|\sin\varphi_{\delta k})^2 + (q_T\,\hat{z}_N\cos\varphi_h - |\delta\boldsymbol{k}_T|\cos\varphi_{\delta k})^2),$$

$$k_f^- = \frac{Q}{\sqrt{2}\,\hat{z}_N}, \quad \boldsymbol{k}_{fT} = (|\delta\boldsymbol{k}_T|\cos\varphi_{\delta k} - q_T\,\hat{z}_N\cos\varphi_h,\ |\delta\boldsymbol{k}_T|\sin\varphi_{\delta k} - q_T\,\hat{z}_N\sin\varphi_h), \tag{3.3}$$

$$\delta\boldsymbol{k}_T = (|\delta\boldsymbol{k}_T|\cos\varphi_{\delta k},\ |\delta\boldsymbol{k}_T|\sin\varphi_{\delta k}), \tag{3.4}$$

where $x_N \equiv -q^+/P^+ = 2x_{\rm Bj}/(1+\sqrt{1+4x_{\rm Bj}^2M^2/Q^2})$ and $\hat{x}_N = x_N/\xi$. In this form, the partonic momenta are parametrized in terms of the two momentum fractions $0 < \xi, \zeta < 1$, two invariant masses $k_i^2 < 0$ and $k_f^2 > 0$, two transverse momenta $|\boldsymbol{k}_{iT}|$, $|\delta\boldsymbol{k}_T| > 0$, and two partonic angular variables φ_{k_i} and $\varphi_{\delta k}$ in the Breit frame, as well as the external kinematic variables x_N, z_N, Q, q_T, and φ_h. In terms of these variables, we can write eq. (3.1) as

$$\begin{aligned}\mathcal{A}(x_{\rm Bj}, Q^2, z_h, P_{hT}|\,\text{region}) = &\int \mathrm{d}\{R_i\}\ \Theta(\{R_i\}|\,\text{region}) \\ &\times \int \mathrm{d}\xi\ \mathrm{d}\zeta\ \mathrm{d}k_i^2\ \mathrm{d}k_f^2\ \mathrm{d}|\boldsymbol{k}_{iT}|\ \mathrm{d}|\delta\boldsymbol{k}_T|\ \mathrm{d}\varphi_{k_i}\ \mathrm{d}\varphi_{\delta k}\ \mathrm{d}\varphi_h \\ &\times \mathcal{P}(\{R_i\}\,|\,\xi,\zeta,k_i^2,k_f^2,|\boldsymbol{k}_{iT}|,|\delta\boldsymbol{k}_T|,\varphi_{k_i},\varphi_{\delta k}) \\ &\times \pi(\xi,\zeta,k_i^2,k_f^2,|\boldsymbol{k}_{iT}|,|\delta\boldsymbol{k}_T|,\varphi_{k_i},\varphi_{\delta k}\,|\,x_{\rm Bj},Q,z_h,q_T,\varphi_h)\ .\end{aligned} \tag{3.5}$$

As given, eq. (3.5) is quite general up to the freedom of choosing the priors and the proximity function. The priors for the parton momenta are selected as follows:

- **Momentum fractions:** since the physical values of ξ and ζ are bounded in the ranges $x_{\rm Bj} < \xi < 1$ and $z_h < \zeta < 1$, and the lowest order diagrams have $\xi = x_{\rm Bj}$ and $\zeta = z_h$, we use flat priors within a window ϵ around $x_{\rm Bj}$ and z_h,

$$\pi(\ldots) \supset \theta(x_{\rm Bj} < \xi < x_{\rm Bj} + \epsilon)\ \theta(z_h < \zeta < z_h + \epsilon), \tag{3.6}$$

  where θ is the Heaviside step function, and we take with $\epsilon = 0.1$.

- **Angular variables:** while angular modulations become nontrivial for spin-dependent observables, we restrict our analysis to unpolarized SIDIS reaction and therefore use flat priors,

$$\pi(\ldots) \supset \theta(0 < \varphi_{k_i} < 2\pi)\ \theta(0 < \varphi_{\delta k} < 2\pi)\ \theta(0 < \varphi_h < 2\pi). \tag{3.7}$$

- **Parton virtualities and intrinsic transverse momenta:** we use Gaussian distributions $\mathcal{G}$ to estimate the parton dynamics in the infrared region,

$$\pi(\ldots) \supset\ \mathcal{G}\Big(\sqrt{|k_i^2|}\,\Big|\,m,\Delta\Big)\ \mathcal{G}\Big(\sqrt{|k_f^2|}\,\Big|\,m,\Delta\Big)\ \mathcal{G}(|\delta\boldsymbol{k}_T||m,\Delta)\ \mathcal{G}(|\boldsymbol{k}_{i,T}||m,\Delta), \tag{3.8}$$

  with mean $m = 0.5\,\text{GeV}$ and width $\Delta = 0.5\,\text{GeV}$ guided by nonperturbative QCD fits of TMD widths [23, 41], relevant for the experiments we consider in this paper.

JHEP04(2022)084

While more sophisticated choices for the these distributions can be made, eq. (3.5) will be sufficient to illustrate the practical use of affinity. For our proximity function we use a flat distribution of the form

$$\Theta(\{R_i\}\,|\,\text{region}) = \prod_i \theta(R_i < R_i^{\max}(\text{region}))\,\theta(R_i > R_i^{\min}(\text{region})) \tag{3.9}$$

that isolates the desired region. The first θ function is used to define the region of "small" values and the second θ is used to define the region of "large" values. Note that both θ functions are not always needed for each region. In the case of the TMD-collinear transition region, we will also use $\theta(R_2^{\max}(\text{region}) < R_2 < R_2^{\min}(\text{region}))$ for R_2 to define intermediate values.

Since for reasonable values for $R_i^{\min}$ and $R_i^{\max}$ we have *a priori* no quantitative knowledge of specific region boundaries other than the qualitative estimates in table 2, in practice we need to appeal to existing TMD phenomenology for guidance. Specifically, we tune the allowed ranges of R_i such that for the kinematic bins where ref. [25] found a good agreement between data and phenomenology, the affinity $\mathcal{A} \sim 1$, and for the excluded kinematic bins the affinity $\mathcal{A} \sim 0$.

In terms of the kinematic cuts in eq. (2.7), our implementation translates as "small" R_i values, with $R_i^{\max}(\text{TMD}) = 0.3$, for $i = 0, 1, 2$. As there are no studies of collinear, central or target regions available to quantify "large" values, we will define as "large" any value that is at least 3 times greater than "small", so that $R_i^{\min}(\text{region}) = 0.9$, for $i = 0, 1, 2$. For other ratios we set $R_i^{\max}(\text{region}) = 0.3$ and $R_i^{\min}(\text{region}) = 3R_i^{\max}(\text{region}) = 0.9$. These values in principle may vary depending on kinematics, so that more fine tuning may be required to delineate the regions with greater accuracy.

We stress that *affinity* is a more powerful tool than the region indicators introduced in refs. [20, 22] and numerically available at https://sidis.herokuapp.com.

4 Applications

In this section we demonstrate the practical utility of the region indicators introduced above by applying our operative definition of affinity to data from existing experiments, as well as to expected data from future facilities. In figure 2 we illustrate the kinematic reach in x_{Bj} and Q^2 of the experiments considered in this paper, namely, Jefferson Lab, HERMES, COMPASS, and the future EIC, with measurements to be performed at points with bin centers in x_{Bj} and Q^2 indicated. We will study the kinematic reach of the experiments in terms of the produced hadron rapidity, defined by

$$y_h \equiv \frac{1}{2} \ln \left| \frac{P_h^+}{P_h^-} \right| .$$

While rapidity is an observable, it can be challenging to measure, particularly at large values, where particle trajectories are close to the beam pipe and neither their energies nor their total momenta can be precisely determined. In practice, pseudorapidity is often used instead, which is a function of the polar angle between the particle trajectory and the

JHEP04(2022)084

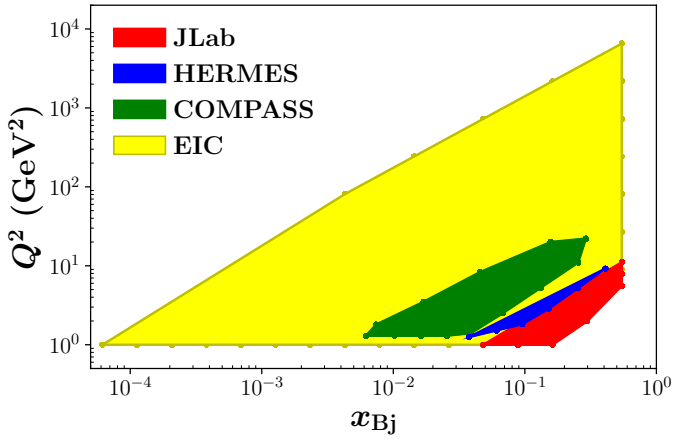

Figure 2. Kinematical reach of Q^2 (GeV^2) versus x_{Bj} for data from existing experiments at Jefferson Lab (red), HERMES (blue), COMPASS (green), and the future EIC (yellow). Bin centers are indicated by filled circles, with each bin representing measurements for different values of z_h and P_{hT}.

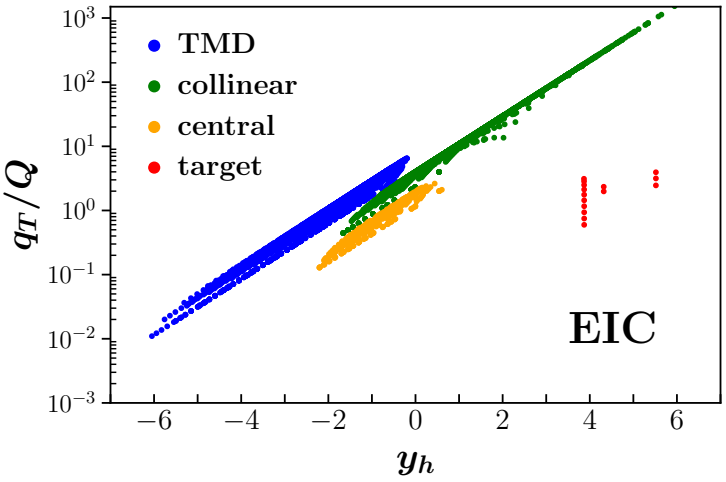

Figure 3. Distribution of bins corresponding to the TMD (blue), collinear (green), central (yellow), and target (red) regions as a function of q_T/Q and rapidity y_h. To make the points distinguishable in regions of overlap, we scale q_T/Q in the TMD region by a factor 8, in the collinear region by 4, and in the central region by 2.

beam axis, and thus ideal for discussions of acceptance coverage of collider detectors and the placement of their various components. For highly relativistic particles, rapidity and pseudorapidity are almost identical and both can be used for physics discussions.

For typical kinematics expected at the future EIC, with variable center of mass energy from $\sqrt{s} = 20\,\mathrm{GeV}$ to $140\,\mathrm{GeV}$, we simulated semi-inclusive deep-inelastic π^+ production for 7400 bins in x_{Bj}, z_h, Q^2, and P_{hT}, as in the EIC Yellow Report EIC [42]. Figure 3 shows the kinematics of the EIC projected data, categorized according to affinity values exceeding the threshold of 5% for various fragmentation regions. The initial proton is always in the positive rapidity range, while the produced hadron has either positive or

JHEP04(2022)084

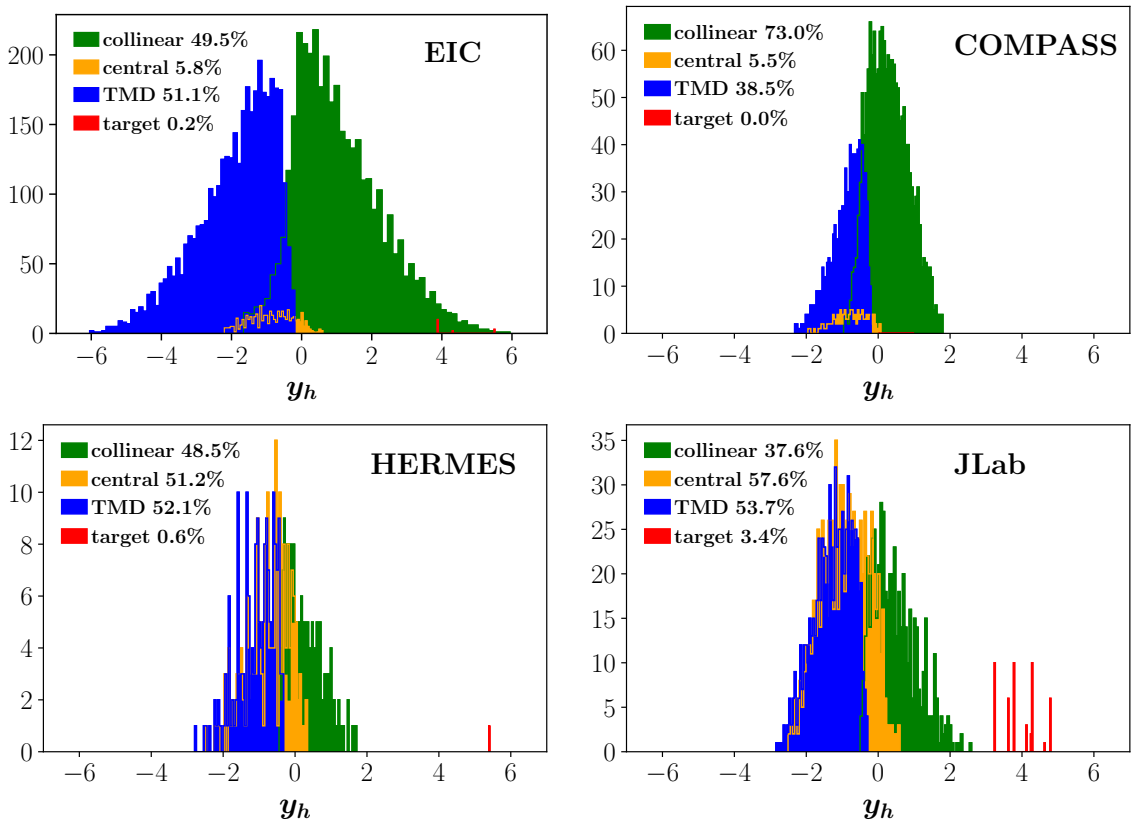

Figure 4. Phase space in rapidity y_h of produced hadrons at EIC, COMPASS, HERMES, and Jefferson Lab kinematics, with for TMD (blue), collinear (green), central (yellow) and target (red) regions indicated. The legends show the percentage of all bins with corresponding affinity above 5%.

negative rapidity. As discussed above, hadrons produced at negative rapidity are likely to be in the current fragmentation region. Hadrons with higher values of q_T migrate into the central and positive rapidities and may originate from hard gluon scattering; therefore, they will be described by collinear QCD. The central region, where low energy partons hadronize, is likely to be in the intermediate region of rapidity. Finally, the target region is typically associated with hadrons with positive rapidities.

As figure 3 demonstrates, the region of central rapidities $y_h \sim 0$ corresponds to an admixture of almost all regions. Although the ratio q_T/Q appears to be a good indicator for separating the TMD and collinear regions, a residual overlapping among central, collinear and TMD regions can only be resolved by accounting also for the value of the hadron rapidity. We note that there are two solutions for y_h [see eq. (20) of ref. [20]], which are on opposite sides of the proton rapidity, and the solution that corresponds to the target fragmentation region is severely constrained by kinematics. The final state hadron has a mass smaller than that of the proton, and if P_{hT} is small enough, then z_h will be small. One can see from figure 3 that target region bins are located in the positive range of rapidity, and q_T/Q is small. We will see below that values of z_h are also small for the target fragmentation region.

JHEP04(2022)084

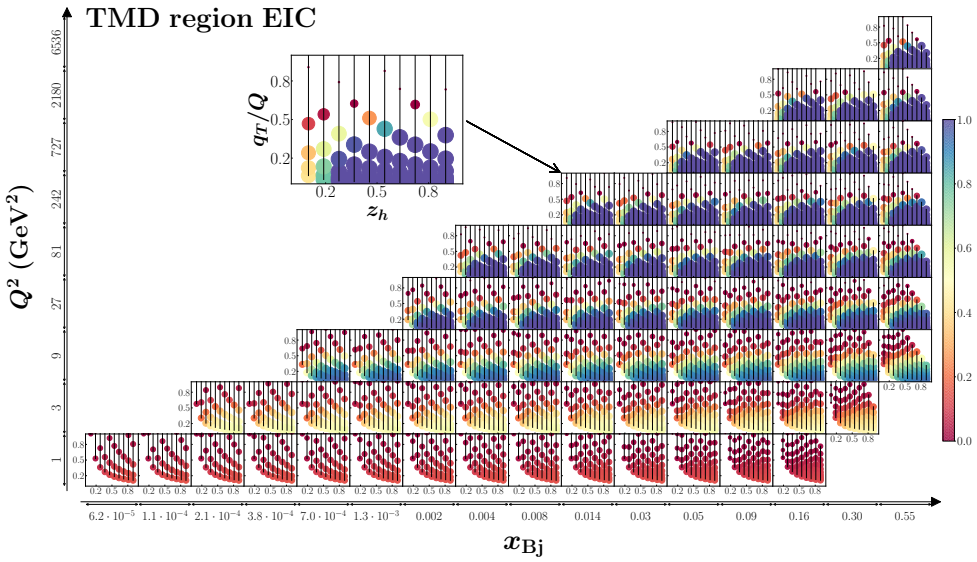

Figure 5. TMD affinity for EIC kinematics. Bin centers are located in the points corresponding to the bin averaged values of x_{Bj} and Q^2, and in each of these bins various values of z_h and q_T/Q can be measured. In each bin of fixed z_h and q_T/Q, the affinity is indicated by a dot with size proportional to the corresponding affinity value. The affinity is color coded according to the scheme on the right of the panels: red (and smaller) symbols correspond to low TMD affinity, while dark blue (and larger) symbols correspond to high TMD affinity.

The distribution of all kinematic bins accessible at current and future facilities is shown in figure 4 as a function of the produced hadron rapidity, y_h, categorized by the affinity as in figure 3. One can see that at the EIC the majority of the data will correspond to either the TMD or collinear QCD fragmentation regions, with small fractions of events in the soft and target fragmentation regions.

At other, lower energy facilities, such as COMPASS, HERMES, and Jefferson Lab, the reach in rapidity is clearly smaller. At Jefferson Lab kinematics, for example, where the measurements are at center of mass energy $\sqrt{s} = 4.6\,\mathrm{GeV}$, one is likely to encounter larger portions of events from central and target fragmentation regions. At the same time, one expects to have large fractions of events that correspond to TMD and collinear factorization for all the experiments discussed. Note also that the regions can overlap; consequently, the sum of percentages for affinities does not equal 100%. We will study each region in more detail in the following.

4.1 TMD region

TMD affinity is calculated using eq. (3.1) by requiring the region indicators R_0, R_1, and R_2 to be small. The results for the bins at the EIC kinematics are shown in figure 5. One can see that bins with relatively large x_{Bj} and Q^2 values (and relatively high z_h and P_{hT}) are particularly important for the TMD factorization description. In terms of

JHEP04(2022)084

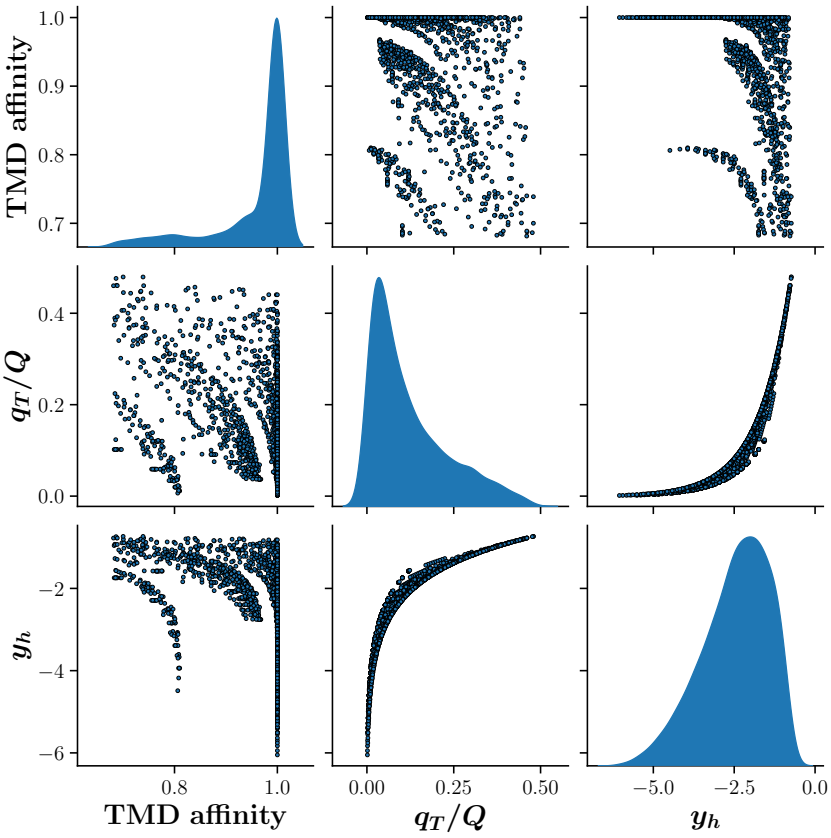

Figure 6. Pair-wise correlations among TMD affinity greater than 70%, q_T/Q, and the produced hadron rapidity y_h for EIC kinematics. The histograms on the diagonal help to visualize the relative abundance of points in other pair-wise plots.

the applicability of TMD factorization, this suggests that q_T/Q becomes sufficiently small for the factorization to be valid. We estimate 2325 out of the 7400 bins to have TMD affinity of 68% or higher and 1739 bins to have TMD affinity of 95% or higher. As discussed below, the remainder of the data (or at least part of it) will correspond to different mechanisms, such as those associated with the collinear factorization scheme.

In figure 6 we show correlation plots of TMD affinity, the ratio q_T/Q, and the rapidity y_h. The correlation between y_h and q_T/Q appears very strong. In particular, we observe that the more negative the produced hadron rapidity y_h, the lower the values of q_T/Q, which is typical of the TMD factorization region. Using our settings for the TMD affinity definition, namely $R_{0,1,2} < 0.3$, we obtain a TMD affinity of 68% or larger for $y_h \in (-6, 0)$ and $q_T/Q \lesssim 0.4$. From our results we find that hadrons in the negative y_h rapidity region are likely to have low values of q_T/Q and large values of TMD affinity.

Both q_T/Q and the rapidity y_h appear to be good proxies for TMD affinity, especially when used in tandem. In fact, taking into account only a single indicator may considerably limit the accuracy in establishing the region boundaries. This was also reported in ref. [43], where an algorithm based on region indicator ratios involving both q_T/Q and y_h was

JHEP04(2022)084

introduced for the study of the $e^+e^- \to hX$ process. Moreover, as the correlation between the rapidity y_h and q_T/Q is very strong, it will be important to obtain information on y_h directly from experimental measurements.

As mentioned above, in this analysis we do not attempt a phenomenological description of the experimental data. However, it is instructive to compare the affinity data selection to those of existing analyses, such as those presented in refs. [25, 26]. If we apply the kinematic cuts as in eq. (2.7) to the EIC data, we find 2116 bins survive from the total of 7400. This subset can be compared to the number of bins that correspond to TMD affinity of 68% or higher, which is 2325. Therefore, only 344 bins do not correspond to cuts from ref. [25], and 94% of data selected by cuts from ref. [25] belong to the TMD region with affinity of 68% or higher.

In the next-to-leading-logarithmic precision analysis of SIDIS, Drell-Yan, and Z-boson production data in ref. [24], the following selection criteria were used:

$$\begin{aligned} &Q^2 > 1.4\,\mathrm{GeV}^2\,, \quad 0.2 < z_h < 0.74\,, \\ &P_{hT} < \min\left[0.2\,Q,\, 0.7\,z_h Q\right] + 0.5\ \mathrm{GeV}\,. \end{aligned} \tag{4.1}$$

Applying these cuts to the sample of projected EIC data in our analysis, we find 2148 bins (from 7400) are selected, and 1504 of those have TMD affinity of 68% or higher. In addition, 838 of these bins do not belong to the bins selected by eq. (2.7) in ref. [25].

Other selection criteria used in refs. [23, 44] for leading order TMD phenomenology are

$$Q^2 > 1.63\,\mathrm{GeV}^2, \quad 0.2 < z_h < 0.6\,, \quad 0.2 < P_{hT} < 0.9\ \mathrm{GeV}. \tag{4.2}$$

Note that from the point of view of factorization proofs the conditions $q_T \ll Q$ and $P_{hT} \ll Q$ are equivalent. However, depending on the numerical value for z_h, data which satisfy $P_{hT} \ll Q$ may not satisfy $q_T \ll Q$, and may therefore be difficult to describe in a TMD approach. Applying the cuts (4.2), we find that 671 bins survive, and 396 of those belong to TMD affinity of 68% or higher. Interestingly, only 50% of the data selected by the cuts from refs. [23, 44] overlaps with the data selected by cuts from refs. [24, 25] in the case of the EIC bins we study.

The COMPASS Collaboration performed measurements [45] of charged hadrons produced in collisions of 160 GeV longitudinally polarized muons scattered off a deuterium target at typical SIDIS kinematics, for $Q^2 > 1\,\mathrm{GeV}^2$, $W^2 \equiv (P+q)^2 > 25\,\mathrm{GeV}^2$, $0.003 < x_{\mathrm{Bj}} < 0.7$, $0.2 < z_h < 1$, and $0.1 < y < 0.9$, where $y = P\cdot q/P\cdot \ell$. The multiplicity in ref. [45] is defined as the ratio of SIDIS to inclusive DIS cross sections,

$$M^h \equiv \frac{\mathrm{d}^4\sigma_{\mathrm{SIDIS}}/\mathrm{d}x_{\mathrm{Bj}}\,\mathrm{d}Q^2\mathrm{d}z_h\,\mathrm{d}P_{hT}^2}{\mathrm{d}^2\sigma_{\mathrm{DIS}}/\mathrm{d}x_{\mathrm{Bj}}\,\mathrm{d}Q^2}\,. \tag{4.3}$$

In figure 7 we present the kinematic bins covered by COMPASS and the data corresponding to the positive hadron multiplicity. The TMD affinity is represented by the colored circles superimposed over each data point. In each bin we plot the data for four z_h bins indicated in the legend as a function of q_T. As in the case of the EIC, higher-Q bins have higher

JHEP04(2022)084

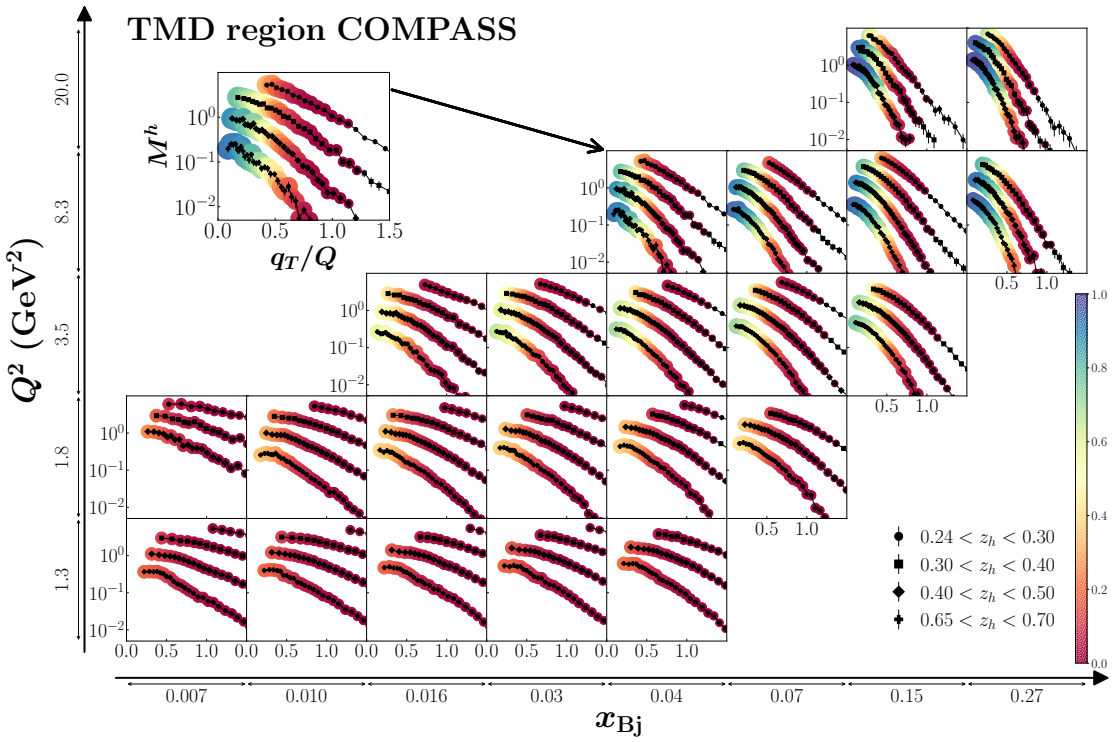

Figure 7. TMD region for COMPASS data [45] on multiplicities M^h for hadrons h^+ (black symbols) produced in muon-deuteron SIDIS, as a function of q_T/Q. In each panel there are up to four bins in z_h, with the dataset at the top (bottom) corresponding to the lowest (highest) range in z_h.

TMD affinity for low values of P_{hT}. For larger values of z_h and $x_{\rm Bj}$ one expects higher TMD affinity, as seen in figure 7.

The COMPASS data were also used in the phenomenological study in ref. [25]. For the case of h^+ production, 138 bins survive after the cuts defined in eq. (2.7). We find 200 bins actually have TMD affinity of 68% or higher, and 81 of them do not survive after applying eq. (2.7). At the same time, 1165 bins are selected by eq. (4.1), but only 200 of them have TMD affinity of 68% or higher, while 767 bins are selected by eq. (4.2), but only 106 have TMD affinity of 68% or higher. The conclusion drawn from this is that additional phenomenological work is needed to delineate the TMD region more precisely.

The HERMES Collaboration measured [46] the multiplicity of pion and kaon production in the scattering of 27.6 GeV positrons from proton and deuteron targets in the SIDIS kinematics $Q^2 > 1\,\text{GeV}^2$, $W^2 > 10\,\text{GeV}^2$, $0.023 < x_{\rm Bj} < 0.4$, $0.2 < z_h < 0.7$, and $y < 0.85$. The measured multiplicity in ref. [46] was defined as

$$M_n^h \equiv \frac{\mathrm{d}^4\sigma_{\rm SIDIS}/\mathrm{d}x_{\rm Bj}\,\mathrm{d}Q^2\,\mathrm{d}z_h\,\mathrm{d}P_{hT}}{\mathrm{d}^2\sigma_{\rm DIS}/\mathrm{d}x_{\rm Bj}\,\mathrm{d}Q^2}\,. \tag{4.4}$$

In figure 8 we show the bins explored by HERMES and the data corresponding to the positive pion multiplicity. In each bin we plot the data for the z_h values indicated in the legend as a function of P_{hT}. The TMD affinity is represented by the colored circles superimposed over each data point. We observe that the affinity is larger for higher z_h and $x_{\rm Bj}$ values,

JHEP04(2022)084

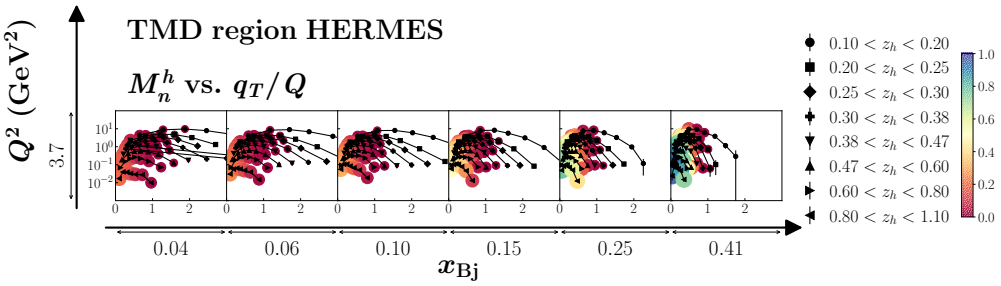

Figure 8. TMD region for HERMES multiplicities M_n^h [46] (black points) as a function of P_{hT} for π^+ produced from a hydrogen target. In each panel there are up to eight z_h bins with the top (bottom) dataset corresponding to the lowest (highest) values of z_h.

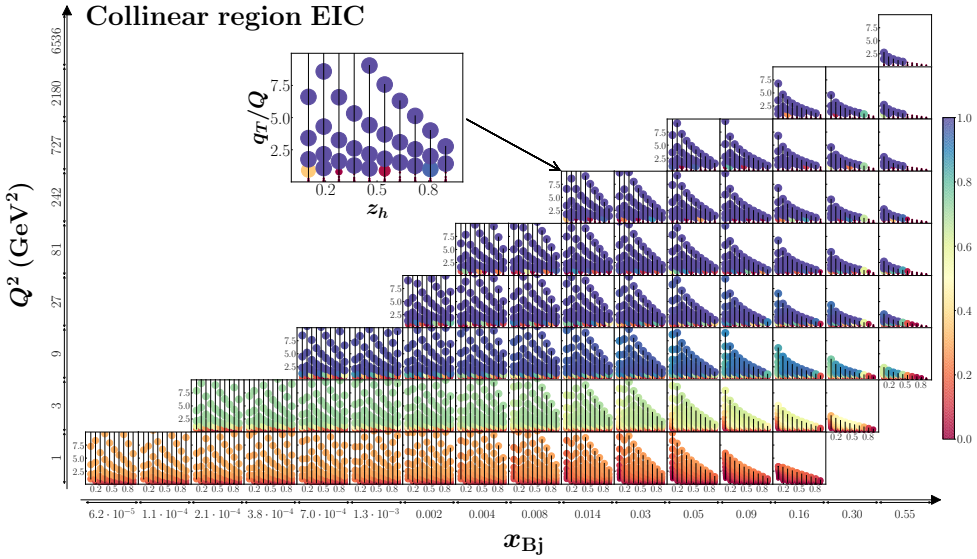

Figure 9. Collinear affinity for EIC kinematics. Bin centers are located in the points corresponding to the bin average values of x_{Bj} and Q^2 (GeV^2). In each of these bins, various values of z_h and q_T/Q can be measured. In each bin of fixed z_h and q_T/Q we plot the affinity as a dot with size proportional to the corresponding affinity value. Affinity is also color coded, according to the scheme on the right of the figure panel: red (and smaller) symbols correspond to low TMD affinity while dark blue (and larger) symbols to high collinear affinity.

and for relatively small P_{hT}. The HERMES data were also used in the phenomenological study of ref. [25]. For the case of π^+ production from a proton target, we find that 34 bins survive after the cuts in eq. (2.7), and 36 bins have TMD affinity of 68% or higher.

4.2 Collinear region

The collinear current fragmentation region is complementary to the TMD region. It covers the region of current fragmentation where hard parton recoil is important, and has negligi-

JHEP04(2022)084

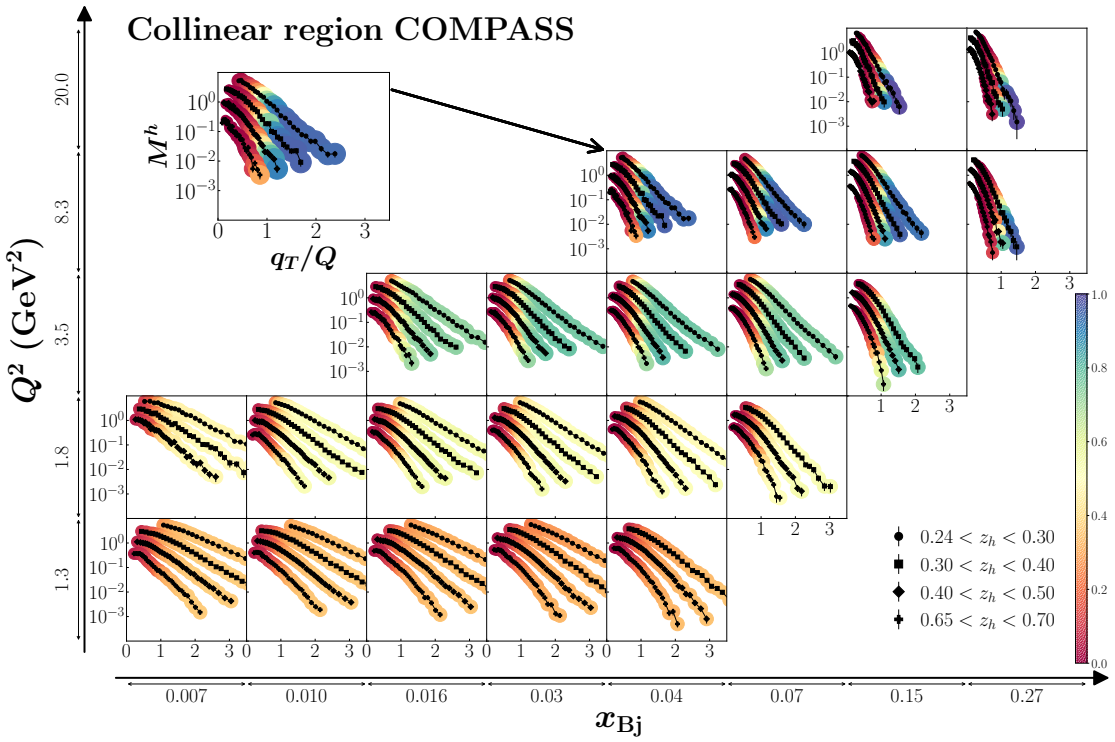

Figure 10. Collinear region for COMPASS multiplicities [45]. The data (black symbols) on M^h for h^+ obtained in muon-deuteron SIDIS are shown as a function of q_T/Q. In each panel there are up to four bins in z_h, with the dataset at the top (bottom) corresponding to the lowest (highest) range in z_h and vice versa.

ble sensitivity to the parton intrinsic transverse momentum. In terms of region indicators, here the ratio R_2 becomes large, while R_4 remains small. This region has been discussed recently in refs. [47, 48], where a significant tension was found between data and theory at COMPASS and HERMES kinematics, with deviations up to an order of magnitude. Wang et al. [47] showed that such deviations are marginally improved by the inclusion of $\mathcal{O}(\alpha_s^2)$ corrections. Similar observations have been made by Bacchetta et al. [49] in the context of the analysis of Drell-Yan cross sections differential in the transverse momentum of the lepton pair.

With the affinity tool in hand, we can now examine the interpretation of existing and future data in the large transverse momentum regime. In figure 9 we present the affinity results at EIC kinematics, showing the ranges for $q_T/Q < 10$ to focus on the larger transverse momentum regime. We estimate that 1750 out of the 7400 bins have a collinear affinity of 68% or higher, while 1170 bins have a collinear affinity of 95% or higher. As expected, the collinear affinity is larger for increasing values of Q and q_T/Q, while becoming smaller for $q_T/Q < 1$. Note that for our chosen values of parton monenta, the collinear affinity values at low Q are not large even when $q_T/Q > 1$, which indicates that in general the $q_T/Q \sim 1$ criterion to estimate the transition from TMD to collinear physics cannot be assumed at low values of Q.

JHEP04(2022)084

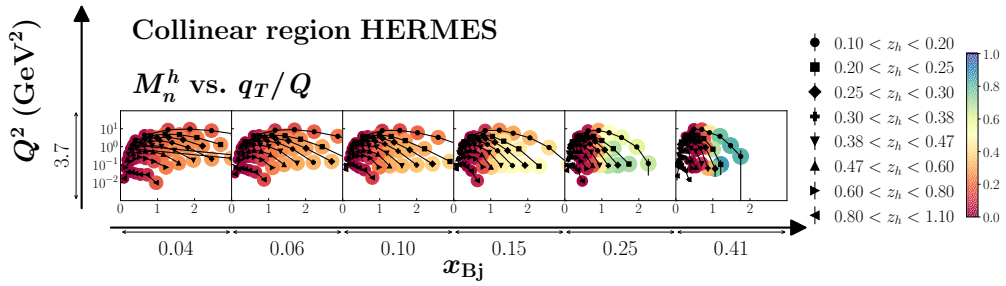

Figure 11. Affinity to the collinear region for the HERMES multiplicity data [46].

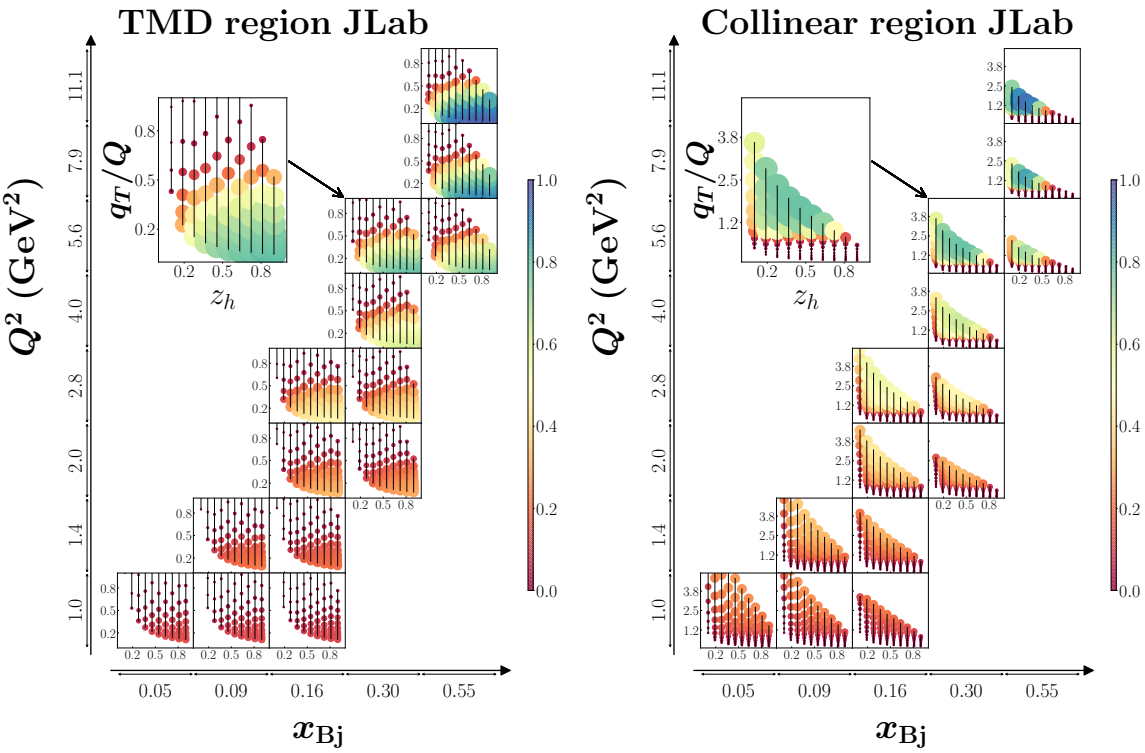

Figure 12. Affinity to the TMD and collinear regions at Jefferson Lab kinematics.

The collinear affinity values at COMPASS kinematics are shown in figure 10, along with the actual experimental multiplicities. Interestingly, the affinity values are in most cases rather small even for $q_T/Q > 1$, which is consistent with the tension between data and theory found in refs. [47, 48]. In contrast, the affinity values become notably larger for high values of Q. However, those regions are close to the edge of the phase space, and it is likely that threshold effects need to be taken into account to achieve a satisfactory description of the cross section in this region [50]. Similar results can be found in figures 11

JHEP04(2022)084

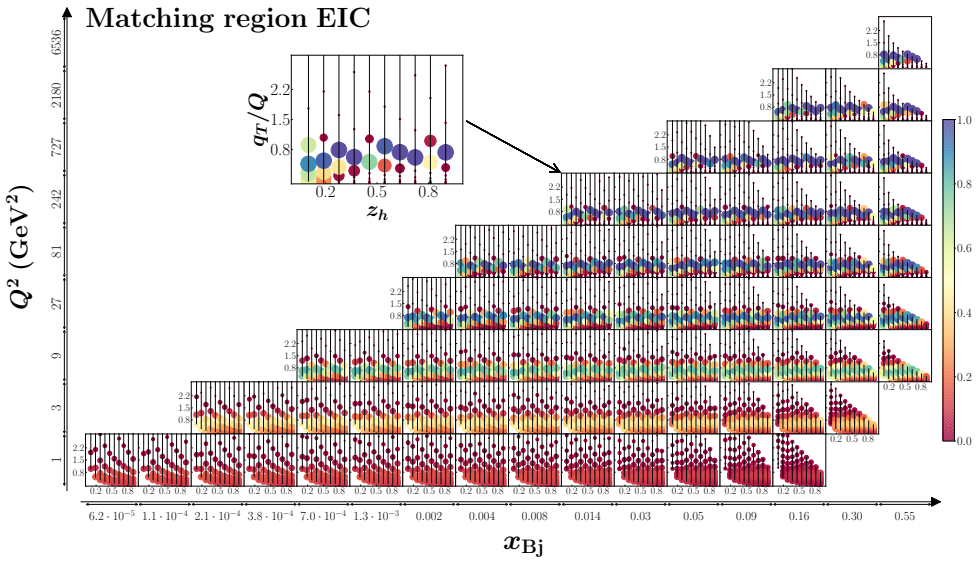

Figure 13. Affinity for the TMD-collinear matching region for EIC kinematics. Bin centers are located in the points corresponding to the bin average values of x_{Bj} and Q^2. In each bin of fixed z_h and q_T/Q we plot the affinity as a dot with size proportional to the corresponding affinity value. Affinity is also color coded, according to the scheme on the right of the figure panel.

and 12 at HERMES and Jefferson Lab kinematics, respectively. It is evident that here no kinematic region shows a strong affinity to collinear factorization. We stress, however, that these observations are based on our specific choices of parton momenta, and in general the results should be viewed only as rough estimates.

4.3 TMD-collinear matching region

The TMD-collinear matching region covers a range of q_T values such that $\Lambda_{\mathrm{QCD}} \ll q_T \ll Q$, and represents the region where one expects a smooth transition between the TMD and collinear regimes. In this intermediate region a description of the data may be possible in either the TMD or collinear schemes. Traditionally, in this region one would implement the $W + Y$ construction by CSS [51], which should ensure a smooth cross section over a wide region of q_T, with controllable error. The existence of such a region is one of the important requirements of the CSS formalism.

In fact, matching the SIDIS cross section across the TMD and collinear regions turns out to be rather challenging, especially with regard to lower energy experiments, as discussed in refs. [20, 28, 52], where different matching procedures have been proposed. While such a discussion goes beyond the scope of this paper, it is important to stress that once a certain definition of this region is chosen, our affinity algorithm can identify and correctly map it, in exactly the same way as for the TMD and collinear regions.

In figure 13 we show the TMD-collinear matching region for EIC kinematics, as determined by the affinity tool. As expected, it correctly covers the range of intermediate values

JHEP04(2022)084

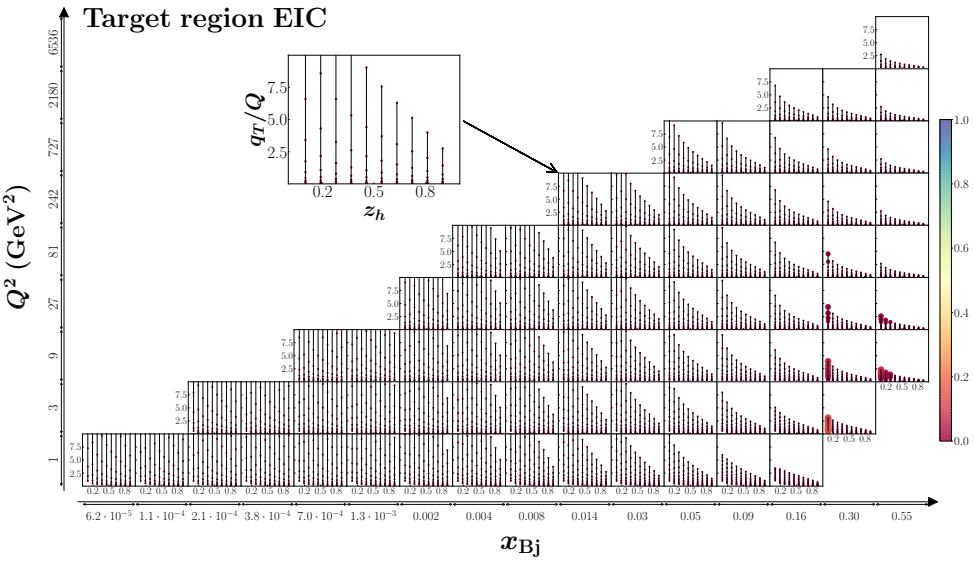

Figure 14. Affinity for the target fragmentation region at EIC kinematics.

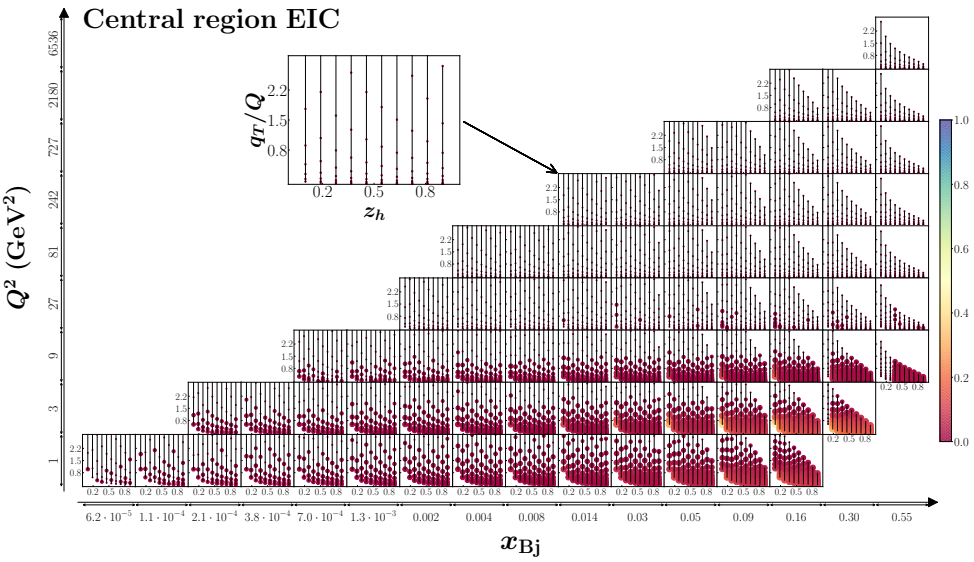

Figure 15. Affinity for the central fragmentation region at EIC kinematics.

of q_T, and turns out to be relevant at rather large values of Q^2 corresponding to moderate and large values of x_{Bj}.

JHEP04(2022)084

4.4 Target and central regions

According to our estimates, at EIC kinematics only a relatively small number of bins is expected to be associated with the target and central fragmentation regions. Indeed, only 15 bins for the target region and 457 bins for the central region exceed an affinity of 5%. The target and central fragmentation regions for the bins of EIC are shown in figures 14 and 15.

As discussed in section 2.1, partons that do not undergo an interaction with the virtual photon hadronize and move predominantly in the direction of the nucleon. These target fragmentation hadrons will be found in the region of positive rapidity, close to the beam. While the experimental measurement of such hadrons is challenging, the study of target fragmentation is important both phenomenologically and theoretically. These processes are usually described in terms of fracture functions [14–17], which are conditional probabilities of producing a hadron the remnant of a nucleon that carries a fraction $\approx 1 - x_{\rm Bj}$ of the nucleon's momentum. For such hadrons the notion of z_h is not well defined, as in this case $z_h \approx 0$. Fracture functions are important ingredients in the description of diffractive hadroproduction, and there are attempts to derive factorization formalisms for those processes [53]. Factorized formulas for TMD fracture functions were conjectured in ref. [16], and the evolution equations along with a more detailed study of factorization was proposed in ref. [17]. Ref. [16] also derives correlations that can be studied experimentally. All this body of work will be important for the planning of EIC detectors [42].

Our estimates of affinity to the target fragmentation region for EIC kinematics are shown in figure 14. One can see that the target region is characterized by relatively large values of $x_{\rm Bj}$ and small values of z_h. Obviously, a more detailed study of the target fragmentation region is needed in order to fully realize the potential of the EIC. Such a study may include a detailed Monte Carlo model of the hadrons produced in the target fragmentation region together with `Geant4` detector simulations.

The last region we will discuss in this section is the central fragmentation region. It is known that the region in rapidity between the struck quark and the nucleon remnants will be filled with radiation that is needed to neutralize the color and make the production of colorless hadrons from colored partons possible. In the event generators in ref. [54] one employs a Lund string model that describes fragmentation as the fracturing of a flux tube created between by the colored quark and the remnant of the nucleon (see refs. [40, 55] and references therein). As a result, the rapidity between the produced hadron and the remnants of the nucleon is filled with hadrons. It is of course interesting to reconcile fragmentation models, especially ones that include spin, such as the Lund string model [56] or the Nambu-Jona-Lasinio model of fragmentation [57, 58], with results of QCD. A recent attempt to use Feynman-graph structures was presented in ref. [21]. In addition, the central region is relevant for factorization, as soft radiation plays an important role for proofs of factorization, including TMD factorization [11, 37].

Even though the central region is incorporated in Monte Carlo generators used in experimental analyses, this region has not yet been explored in detail empirically, and future experimental studies, such as of rapidity distributions of hadrons, will be needed. We estimate that for the central region we have a majority of events in the low-Q^2 and

JHEP04(2022)084

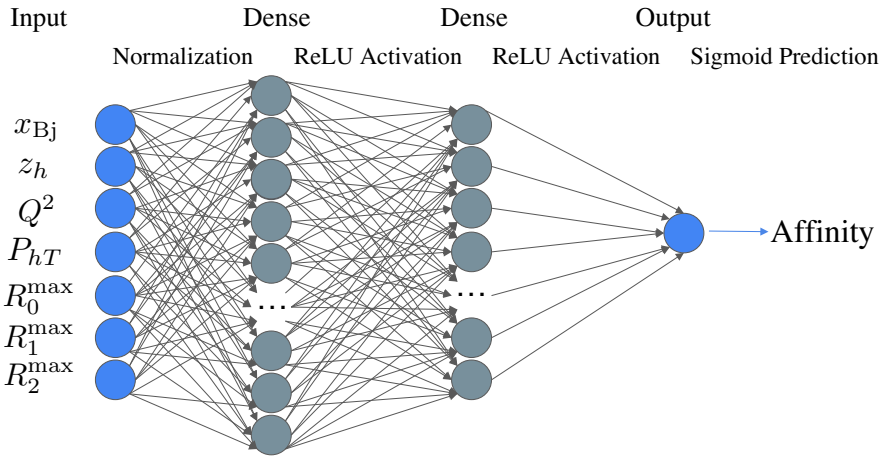

Figure 16. Schematic illustration of the architecture of the neural network devised for calculating affinity. Each arrow represents a trainable parameter that each neural network must learn through training. Input, hidden layers, and output are represented by the (stack of) blobs. The normalization layer and ReLU activation functions are necessary to mitigate the vanishing gradient problem.

low-P_{hT} region, as indicated in figure 15. These hadrons are present in the region of central rapidities, $y_h \sim 0$, and as we explored in figure 4, the contribution to this region of rapidity comes from most of the fragmentation regions investigated in this paper. We expect that in the future more detailed experimental, theoretical, and phenomenological studies will allow more precise delineations of this region [21, 37].

5 Interactive affinity tool

The calculation of affinity is numerically demanding and time consuming. In order to facilitate the computations we use Machine Learning techniques to train a neural network model for fast evaluation of affinity.

Firstly, we generate the affinity data that will be used in training the neural networks by varying the maximum values of three ratios, R_0, R_1, and R_2, and keeping other nonperturbative parameters fixed, as discussed in the paper. The affinity is defined with respect to the values $R_0^{\max}$, $R_1^{\max}$, and $R_2^{\max}$ varying each in the range $(0.05, 1.25)$. We generate ≈ 15000 configurations[1] of those parameters, and for each configuration of these values we produce the corresponding set of affinities for each of the 7400 EIC experimental bins.

We use the `TensorFlow` [59] framework to create and train four neural networks that predict the TMD, collinear, target, and central affinity regions with seven input values of x_{Bj}, z_h, P_{hT}, Q^2, $R_0^{\max}$, $R_1^{\max}$, and $R_2^{\max}$. Choosing the best neural network is not always straightforward or deterministic, as there are many hyperparameters to be adjusted, such

[1]This number gave very good predictive power of the models.

JHEP04(2022)084

as the number and the width of hidden layers, as well as the minimization algorithms to be used. A good hyperparameter combination can significantly improve the performance of the network, and we employ the hyper-band `KerasTuner` [60] for tuning the hyperparameters for our networks. The repository that contains the resulting neural network models can be accessed at a `GitHub` repository [61].

`KerasTuner` hyperparameter search results in each net consisting of four layers: the input layer, two hidden layers, and the output layer. A pictorial representation of the architecture of this network is presented in figure 16, where arrows correspond to the internal parameters that each neural net must learn through training, and the input, output and hidden layers are represented by the (stack of) blobs. Training is an iterative process of updating weights by providing the model input data accompanied by affinity values.

The first hidden layer has 576 neurons for the TMD, 960 for the collinear, 896 for the target, and 576 for the central regions. The second hidden layer has 160 neurons for the TMD, 544 for the collinear, 256 for the target, and 736 for the central regions. For neuron activation we use the rectified linear activation (ReLU) function $f(x) = \max(0, x)$, where x is the input of an activation layer, or the weighted sum of each node in that layer. Here, the weights are trainable parameters and the sum runs over the output of the last layer. The normalization layer and ReLU activation functions are necessary to mitigate the vanishing gradient problem; in fact, when back propagation fails, the weights cannot be updated.

To obtain predictions of affinities bounded on the interval from 0 to 1, we choose the activation function of the last layer (output) as the so-called sigmoid function, which maps any real value to the range $(0, 1)$. This is a cumulative distribution function of a logistic distribution, and transforms the weighted sum of the second dense layer output into a probability or confidence of prediction. Other output functions were also considered in the tuning process.

We used 80% of the data for training and 20% for validation of the network learning, which provides a set of input data to test the network at the end of each training iteration. This procedure allows one to avoid over-fitting and to check the accuracy of the predictions. We use the validation predictions to calculate the mean squared error (MSE), for which back propagation minimizes. Minimization of the validation loss, performed with the Adam minimizer, optimizes the weight parameters. The model architecture that provides the smallest validation MSE (loss function) was considered to be the best. Each model training lasted between 50 and 500 epochs (iterations over the training data). After the training, the resulting neural networks were saved in order to be available for use in further `TensorFlow` applications. For each hyper search, the networks with the five smallest validation MSE were summarized. When the best network was unclear, an independent test set was used to inform the architecture chosen to be further trained and implemented in the affinity tool.

Finally, 100 datasets for each of the 7400 experimental bins and corresponding affinities were independently generated for testing, which allowed for visual and measurable comparisons of networks. The predicted and generated affinity values shown in figure 17 provide a visualization of the error distribution, with a center-line representing perfect predictions and a cloud of points depicting residuals (differences between the values of generated and predicted affinity) for the TMD and collinear network models. The scatter plots in the upper panels show two parallel red lines, equally spaced about the black line of zero

JHEP04(2022)084

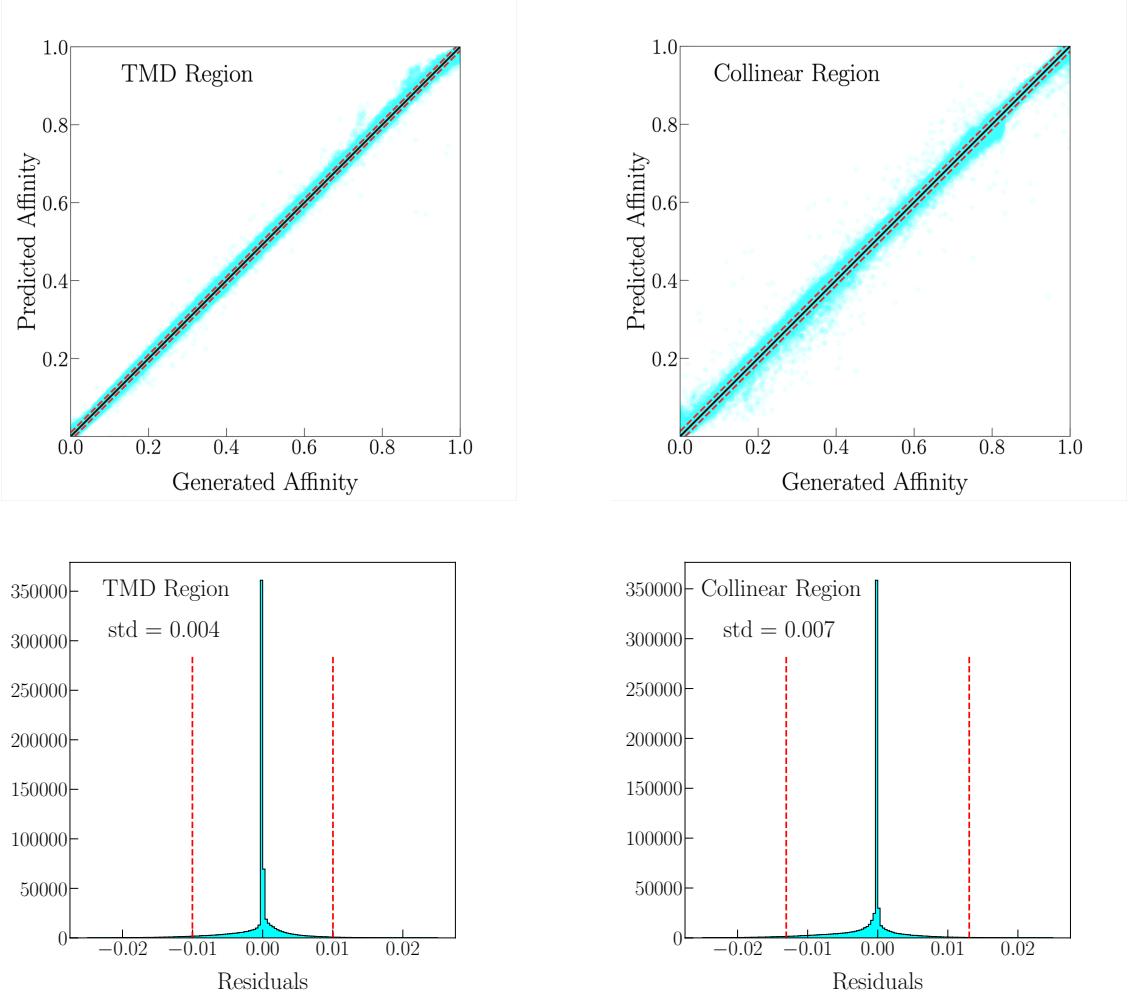

Figure 17. Upper panels: scatter plots of predicted affinity versus the affinity generated with test data. The left (right) panel corresponds to the TMD (collinear) affinity. In each panel the black solid central line represents a perfect prediction, corresponding to zero residuals, while the points located between the two parallel red dashed lines contain 95% of all residuals. Lower panels: histograms of residuals for the TMD (left) and collinear (right) regions. The red dashed lines showing the intervals that contain 95% of all residuals, with the standard deviation "std" indicated. Note that the actual range of the residual values is wider than that shown in the histograms, but the outliers in the plots are omitted for better visual clarity.

residuals, which contains 95% of all residuals. The cloud of points shows higher variation in the TMD network error for larger values of affinity compared to smaller values, while the variation of error for collinear appears to be more uniform over the range of affinity values. Histograms of residuals are shown in the lower panels of figure 17. Around 11% of all predictions have zero residuals, resulting in the histograms being sharply peaked. The standard deviations, 0.004 for TMD affinity and 0.007 for collinear affinity, confirm a very good accuracy of predictions of the trained neural network models.

JHEP04(2022)084

The validated and tested affinity tool has been made publicly available. To minimize the need for installation of additional software for the affinity tool users, we utilize a `Google Colab` notebook that allows users to launch the tool from a web browser without additional installation. The notebook clones the corresponding `GitHub` repository that contains the neural networks and all scripts that produce the interactive plots. The user can choose values for $R_0^{\max}$, $R_1^{\max}$, and $R_2^{\max}$ and generate plots corresponding to the TMD, target, central, and collinear regions at the EIC using neural network models trained for these regions. The interactive notebook with user instructions can be found in ref. [62].

6 Conclusions

SIDIS measurements offer a tremendous opportunity to learn about the partonic structure of nucleons. For a correct phenomenological interpretation of the information they encode, it is vital to develop tools that allow experimental data to be connected to the corresponding theoretical framework. Factorization theorems only apply under specific kinematic conditions, essentially dictated by power counting. It is therefore very important to be able to identify as precisely as possible the sensitivity of each data subset to those kinematic requirements.

In this paper we have implemented the region indicators, R_i first introduced in refs. [20, 22], to quantify our confidence in the proximity of SIDIS observables to a particular physical mechanism. For this purpose we have devised a new tool, "affinity", to facilitate the separation of phase space regions where different factorization formalisms apply. We quantify affinity by combining information from the Monte Carlo generation of partonic configurations and the resulting ratios R_i into a single estimate of proximity to a particular hadron production region, which ranges from 0% to 100%. The affinity to the TMD current fragmentation region is estimated for HERMES and COMPASS datasets for unpolarized multiplicities, and for Jefferson Lab and EIC kinematics. We also quantify the proximity of the current fragmentation region for large transverse momenta described by a collinear QCD treatment, and the transition region from the TMD to collinear factorization descriptions [28]. The central and target regions are also addressed, but these require further phenomenological investigation and dedicated theoretical studies.

Our affinity tool shows that a large portion of experimental bins can be associated with either TMD or collinear physics, for all considered experiments, and especially for the future EIC. Lower energy experiments such as those at Jefferson Lab, however, show a non-negligible admixture of central and target fragmentation events. The affinity tool can be applied in phenomenological analyses to select kinematic bins that are sensitive to the kinematic region of interest. It can also be used to guide the development of new SIDIS experiments, and to incorporate the region estimator into experimental analyses. For this reason we also provide a publicly available interactive tool that allows the study of affinity according to any choice for the separation of the kinematic regions. This tool is based on a neural network model trained with Machine Learning techniques which allows a fast evaluation of the affinity. The affinity interactive tool is available as a `Google Colab` notebook [62], which can easily be accessed and run from any browser without the need of additional software.

JHEP04(2022)084

Acknowledgments

We would like to thank Ted Rogers for helpful discussions and collaboration in the early stages of this research. This work has been supported by the National Science Foundation under Grants No. PHY-2011763 (D.P.), No. PHY-2012002 (A.P., S.D., Z.S.), the U.S. Department of Energy, under contracts No. DE-FG02-07ER41460 (L.G.) and No. DE-AC05-06OR23177 (M.D., A.P., N.S., W.M.) under which Jefferson Science Associates, LLC, manages and operates Jefferson Lab, and within the framework of the TMD Topical Collaboration. A.P. would like to thank Temple University for hospitality and support during his sabbatical leave. The work of N.S. was supported by the DOE, Office of Science, Office of Nuclear Physics in the Early Career Program. M.B. acknowledges funding from the European Union's Horizon 2020 research and innovation programme under grant agreement STRONG — 2020 — No 824093.

References

[1] A. Kotzinian, *New quark distributions and semiinclusive electroproduction on the polarized nucleons*, *Nucl. Phys. B* **441** (1995) 234 [hep-ph/9412283] [INSPIRE].

[2] P.J. Mulders and R.D. Tangerman, *The complete tree level result up to order* $1/Q$ *for polarized deep inelastic leptoproduction*, *Nucl. Phys. B* **461** (1996) 197 [*Erratum ibid.* **484** (1997) 538] [hep-ph/9510301] [INSPIRE].

[3] A. Bacchetta, M. Diehl, K. Goeke, A. Metz, P.J. Mulders and M. Schlegel, *Semi-inclusive deep inelastic scattering at small transverse momentum*, *JHEP* **02** (2007) 093 [hep-ph/0611265] [INSPIRE].

[4] D. Boer, R. Jakob and P.J. Mulders, *Asymmetries in polarized hadron production in* e^+e^- *annihilation up to order* $1/Q$, *Nucl. Phys. B* **504** (1997) 345 [hep-ph/9702281] [INSPIRE].

[5] D. Boer, *Sudakov suppression in azimuthal spin asymmetries*, *Nucl. Phys. B* **603** (2001) 195 [hep-ph/0102071] [INSPIRE].

[6] D. Boer, *Angular dependences in inclusive two-hadron production at BELLE*, *Nucl. Phys. B* **806** (2009) 23 [arXiv:0804.2408] [INSPIRE].

[7] R.D. Tangerman and P.J. Mulders, *Intrinsic transverse momentum and the polarized Drell-Yan process*, *Phys. Rev. D* **51** (1995) 3357 [hep-ph/9403227] [INSPIRE].

[8] J.C. Collins and D.E. Soper, *Parton distribution and decay functions*, *Nucl. Phys. B* **194** (1982) 445 [INSPIRE].

[9] J.C. Collins, D.E. Soper and G.F. Sterman, *Transverse momentum distribution in Drell-Yan pair and W and Z boson production*, *Nucl. Phys. B* **250** (1985) 199 [INSPIRE].

[10] X.-D. Ji, J.-P. Ma and F. Yuan, *QCD factorization for semi-inclusive deep-inelastic scattering at low transverse momentum*, *Phys. Rev. D* **71** (2005) 034005 [hep-ph/0404183] [INSPIRE].

JHEP04(2022)084

[11] J. Collins, *Foundations of perturbative QCD*, Cambridge University Press, Cambridge, U.K. (2009).

[12] D. Boer and P.J. Mulders, *Time reversal odd distribution functions in leptoproduction*, *Phys. Rev. D* **57** (1998) 5780 [hep-ph/9711485] [INSPIRE].

[13] E.L. Berger, *Semi-inclusive inelastic electron scattering from nuclei*, in *NPAS workshop on electronuclear physics with internal targets*, SLAC, U.S.A. (1987), p. 82.

[14] L. Trentadue and G. Veneziano, *Fracture functions: an improved description of inclusive hard processes in QCD*, *Phys. Lett. B* **323** (1994) 201 [INSPIRE].

[15] M. Grazzini, L. Trentadue and G. Veneziano, *Fracture functions from cut vertices*, *Nucl. Phys. B* **519** (1998) 394 [hep-ph/9709452] [INSPIRE].

[16] M. Anselmino, V. Barone and A. Kotzinian, *SIDIS in the target fragmentation region: polarized and transverse momentum dependent fracture functions*, *Phys. Lett. B* **699** (2011) 108 [arXiv:1102.4214] [INSPIRE].

[17] X.P. Chai, K.B. Chen, J.P. Ma and X.B. Tong, *Fracture functions in different kinematic regions and their factorizations*, *JHEP* **10** (2019) 285 [arXiv:1903.00809] [INSPIRE].

[18] P.J. Mulders, *Current fragmentation in semiinclusive leptoproduction*, *AIP Conf. Proc.* **588** (2001) 75 [hep-ph/0010199] [INSPIRE].

[19] S.J. Joosten, *Fragmentation and nucleon structure in semi-inclusive deep-inelastic scattering at the HERMES experiment*, Ph.D. thesis, Illinois U., Urbana, IL, U.S.A. (2013).

[20] M. Boglione, J. Collins, L. Gamberg, J.O. Gonzalez-Hernandez, T.C. Rogers and N. Sato, *Kinematics of current region fragmentation in semi-inclusive deeply inelastic scattering*, *Phys. Lett. B* **766** (2017) 245 [arXiv:1611.10329] [INSPIRE].

[21] J. Collins and T.C. Rogers, *Graphical structure of hadronization and factorization in hard collisions*, arXiv:1801.02704 [INSPIRE].

[22] M. Boglione et al., *Mapping the kinematical regimes of semi-inclusive deep inelastic scattering*, *JHEP* **10** (2019) 122 [arXiv:1904.12882] [INSPIRE].

[23] M. Anselmino, M. Boglione, J.O. Gonzalez Hernandez, S. Melis and A. Prokudin, *Unpolarised transverse momentum dependent distribution and fragmentation functions from SIDIS multiplicities*, *JHEP* **04** (2014) 005 [arXiv:1312.6261] [INSPIRE].

[24] A. Bacchetta, F. Delcarro, C. Pisano, M. Radici and A. Signori, *Extraction of partonic transverse momentum distributions from semi-inclusive deep-inelastic scattering, Drell-Yan and Z-boson production*, *JHEP* **06** (2017) 081 [*Erratum ibid.* **06** (2019) 051] [arXiv:1703.10157] [INSPIRE].

[25] I. Scimemi and A. Vladimirov, *Non-perturbative structure of semi-inclusive deep-inelastic and Drell-Yan scattering at small transverse momentum*, *JHEP* **06** (2020) 137 [arXiv:1912.06532] [INSPIRE].

[26] A. Bacchetta et al., *Transverse-momentum-dependent parton distributions up to N^3LL from Drell-Yan data*, *JHEP* **07** (2020) 117 [arXiv:1912.07550] [INSPIRE].

[27] HERMES collaboration, *Azimuthal single- and double-spin asymmetries in semi-inclusive deep-inelastic lepton scattering by transversely polarized protons*, *JHEP* **12** (2020) 010 [arXiv:2007.07755] [INSPIRE].

JHEP04(2022)084

[28] J. Collins, L. Gamberg, A. Prokudin, T.C. Rogers, N. Sato and B. Wang, *Relating transverse momentum dependent and collinear factorization theorems in a generalized formalism*, *Phys. Rev. D* **94** (2016) 034014 [arXiv:1605.00671] [INSPIRE].

[29] J.C. Collins and D.E. Soper, *Back-to-back jets in QCD*, *Nucl. Phys. B* **193** (1981) 381 [*Erratum ibid.* **213** (1983) 545] [INSPIRE].

[30] J.C. Collins and D.E. Soper, *Back-to-back jets: Fourier transform from B to K-transverse*, *Nucl. Phys. B* **197** (1982) 446 [INSPIRE].

[31] R. Meng, F.I. Olness and D.E. Soper, *Semiinclusive deeply inelastic scattering at small q_T*, *Phys. Rev. D* **54** (1996) 1919 [hep-ph/9511311] [INSPIRE].

[32] S.M. Aybat and T.C. Rogers, *TMD parton distribution and fragmentation functions with QCD evolution*, *Phys. Rev. D* **83** (2011) 114042 [arXiv:1101.5057] [INSPIRE].

[33] J. Collins and T. Rogers, *Understanding the large-distance behavior of transverse-momentum-dependent parton densities and the Collins-Soper evolution kernel*, *Phys. Rev. D* **91** (2015) 074020 [arXiv:1412.3820] [INSPIRE].

[34] M.G. Echevarria, A. Idilbi and I. Scimemi, *Factorization theorem for Drell-Yan at low q_T and transverse momentum distributions on-the-light-cone*, *JHEP* **07** (2012) 002 [arXiv:1111.4996] [INSPIRE].

[35] M.G. Echevarria, A. Idilbi and I. Scimemi, *Unified treatment of the QCD evolution of all (un-)polarized transverse momentum dependent functions: Collins function as a study case*, *Phys. Rev. D* **90** (2014) 014003 [arXiv:1402.0869] [INSPIRE].

[36] G. Altarelli, R.K. Ellis, M. Greco and G. Martinelli, *Vector boson production at colliders: a theoretical reappraisal*, *Nucl. Phys. B* **246** (1984) 12 [INSPIRE].

[37] J. Collins, *Do fragmentation functions in factorization theorems correctly treat non-perturbative effects?*, *PoS* **QCDEV2016** (2017) 003 [arXiv:1610.09994] [INSPIRE].

[38] T. Sjöstrand, S. Mrenna and P.Z. Skands, *PYTHIA 6.4 physics and manual*, *JHEP* **05** (2006) 026 [hep-ph/0603175] [INSPIRE].

[39] G. Corcella et al., *HERWIG 6: an event generator for hadron emission reactions with interfering gluons (including supersymmetric processes)*, *JHEP* **01** (2001) 010 [hep-ph/0011363] [INSPIRE].

[40] B. Andersson, *The Lund model*, volume 7, Cambridge University Press, Cambridge, U.K. (2005) [INSPIRE].

[41] A. Signori, A. Bacchetta, M. Radici and G. Schnell, *Investigations into the flavor dependence of partonic transverse momentum*, *JHEP* **11** (2013) 194 [arXiv:1309.3507] [INSPIRE].

[42] R. Abdul Khalek et al., *Science requirements and detector concepts for the electron-ion collider: EIC yellow report*, arXiv:2103.05419 [INSPIRE].

[43] M. Boglione and A. Simonelli, *Kinematic regions in the $e^+e^- \to hX$ factorized cross section in a 2-jet topology with thrust*, arXiv:2109.11497 [INSPIRE].

[44] Jefferson Lab Angular Momentum collaboration, *Origin of single transverse-spin asymmetries in high-energy collisions*, *Phys. Rev. D* **102** (2020) 054002 [arXiv:2002.08384] [INSPIRE].

[45] COMPASS collaboration, *Transverse-momentum-dependent multiplicities of charged hadrons in muon-deuteron deep inelastic scattering*, *Phys. Rev. D* **97** (2018) 032006 [arXiv:1709.07374] [INSPIRE].

JHEP04(2022)084

[46] HERMES collaboration, *Multiplicities of charged pions and kaons from semi-inclusive deep-inelastic scattering by the proton and the deuteron*, *Phys. Rev. D* **87** (2013) 074029 [arXiv:1212.5407] [INSPIRE].

[47] B. Wang, J.O. Gonzalez-Hernandez, T.C. Rogers and N. Sato, *Large transverse momentum in semi-inclusive deeply inelastic scattering beyond lowest order*, *Phys. Rev. D* **99** (2019) 094029 [arXiv:1903.01529] [INSPIRE].

[48] J.O. Gonzalez-Hernandez, T.C. Rogers, N. Sato and B. Wang, *Challenges with large transverse momentum in semi-inclusive deeply inelastic scattering*, *Phys. Rev. D* **98** (2018) 114005 [arXiv:1808.04396] [INSPIRE].

[49] A. Bacchetta, G. Bozzi, M. Lambertsen, F. Piacenza, J. Steiglechner and W. Vogelsang, *Difficulties in the description of Drell-Yan processes at moderate invariant mass and high transverse momentum*, *Phys. Rev. D* **100** (2019) 014018 [arXiv:1901.06916] [INSPIRE].

[50] A. Kulesza, G.F. Sterman and W. Vogelsang, *Joint resummation for Higgs production*, *Phys. Rev. D* **69** (2004) 014012 [hep-ph/0309264] [INSPIRE].

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

[52] M.G. Echevarria, T. Kasemets, J.-P. Lansberg, C. Pisano and A. Signori, *Matching factorization theorems with an inverse-error weighting*, *Phys. Lett. B* **781** (2018) 161 [arXiv:1801.01480] [INSPIRE].

[53] A. Berera and D.E. Soper, *Behavior of diffractive parton distribution functions*, *Phys. Rev. D* **53** (1996) 6162 [hep-ph/9509239] [INSPIRE].

[54] T. Sjöstrand, *Status and developments of event generators*, *PoS* **LHCP2016** (2016) 007 [arXiv:1608.06425] [INSPIRE].

[55] B. Andersson, G. Gustafson, G. Ingelman and T. Sjöstrand, *Parton fragmentation and string dynamics*, *Phys. Rept.* **97** (1983) 31 [INSPIRE].

[56] A. Kerbizi, X. Artru, Z. Belghobsi, F. Bradamante and A. Martin, *Recursive model for the fragmentation of polarized quarks*, *Phys. Rev. D* **97** (2018) 074010 [arXiv:1802.00962] [INSPIRE].

[57] T. Ito, W. Bentz, I.C. Cloet, A.W. Thomas and K. Yazaki, *The NJL-jet model for quark fragmentation functions*, *Phys. Rev. D* **80** (2009) 074008 [arXiv:0906.5362] [INSPIRE].

[58] H.H. Matevosyan, A. Kotzinian and A.W. Thomas, *Monte Carlo implementation of polarized hadronization*, *Phys. Rev. D* **95** (2017) 014021 [arXiv:1610.05624] [INSPIRE].

[59] M. Abadi et al., *TensorFlow: large-scale machine learning on heterogeneous distributed systems*, arXiv:1603.04467 [INSPIRE].

[60] T. O'Malley et al., *Keras tuner github repository*, https://github.com/keras-team/keras-tuner, (2019).

[61] M. Boglione et al., *Affinity github repository*, https://github.com/QCDHUB/SIDIS-Affinity, (2022).

[62] M. Boglione et al., *Affinity tool*, https://colab.research.google.com/github/QCDHUB/SIDIS-Affinity/blob/main/interactive_affinity_tool.ipynb, (2022).

JHEP04(2022)084