Prepared for submission to JHEP JLAB-THY-22-xxxx

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

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] **Z. Scalyer**[e]

[a] *Dipartimento di Fisica, Università di Torino, INFN-Sezione Torino, Via P. Giuria 1, 10125 Torino, Italy*

[b] *Jefferson Lab, Newport News, Virginia 23606, USA*

[c] *Science Division, Penn State University Berks, Reading, Pennsylvania 19610, USA*

[d] *Department of Physics, Lebanon Valley College, Annville, Pennsylvania 17003, USA*

[e] *Prepared for Flight, LLC, York, Pennsylvania 17402, USA*

E-mail: elena.boglione@to.infn.it, mdiefent@jlab.org, skd140@psu.edu, lpg10@psu.edu, wmelnitchouk@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: Semi-Inclusive Deep-Inelastic Scattering, fragmentation, Quantum chromodynamics, Electron-Ion Collider

[1] On leave at Temple University, Philadelphia, USA

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 momentum, 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 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 z of the quark carried by the produced hadron, which is independent of the momentum fraction x 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 [13, 18], 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 inside the current fragmentation region. Since the observed hadrons can be produced via different physical mechanisms, identifying SIDIS cross sections at a given kinematic point with TMD factorization formulas 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 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 Sec. 2 by briefly recalling the results of Refs. [20, 22], and introduce the region indicators, R_i, that delineate different hadron production mechanisms in various regions of kinematics. To assess the proximity of the data at given kinematics to a specific physical mechanism, in Sec. 3 we introduce the affinity, $\mathcal{A}$, as a global estimator. In Sec. 4 we apply the new affinity tool to the analysis of existing data from the HERMES and COMPASS experiments, and discuss expectations for the analysis of data expected from Jefferson Lab and the future EIC. In Sec. 5 we present the results of training a neural network using the `TensorFlow` package, with various choices for the underlying demarcation

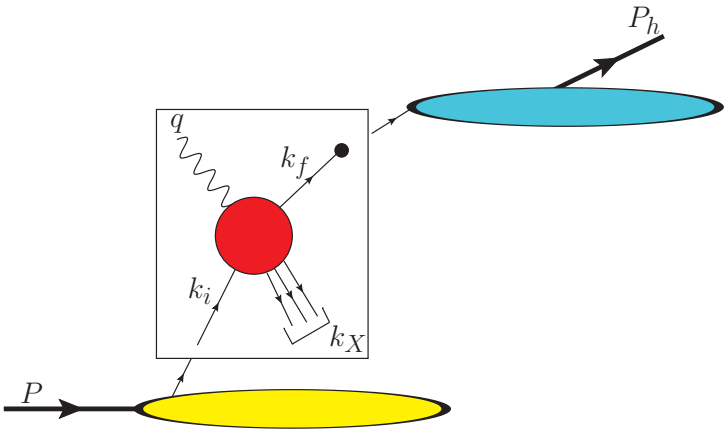

Figure 1. Sketch of the SIDIS process, with scattering of a virtual photon (momentum q) from a parton (k_i) inside a nucleon (P), with the scattered parton (k_f) fragmenting to a hadron (P_h) in the final state and unmeasured hadronic debris (k_X). The lower (yellow) blob represents the residual system 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.

of regions, and introduce a `Google Colab` interactive notebook for visualizing the affinity at EIC kinematics. Finally, in Sec. 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 introduced in Refs. [20, 22] to help delineate different hadron production mechanisms by including some information on the underlying momentum flow in the partonic subprocess. 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 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, 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 case a treatment based on collinear QCD factorization theorems [9, 11, 36] would be 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 Fig. 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]. While the identification of current and target regions is well defined by criteria which establish a clear rapidity separation between the collinear and soft sub-graphs in large-Q asymptotics of factorization [11], in reality these rapidity gaps are filled by centrally produced hadrons [37]. These are the phenomena that are approximated in Monte-Carlo event generators as string [38] and 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. Yet, it is important to note that the fastest moving hadrons in opposite hemispheres in string like fragmentation [40], being space-like separated, are in line with the independent hadronization 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. While there has not been a proof of TMD factorization which incorporates the central fragmentation region, nonetheless it is imperative to classify the momentum regions associated with the central region. Thus, pinning down the central region is considerably more complicated and remains 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].

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 $R_2 \ll 1$. 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 current analysis we use the cuts

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

to select the data. 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}.

The region of large transverse momentum is characterized by a ratio analogous 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,

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

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

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 perturbative QCD 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 perturbative QCD calculations are needed to describe data in this region.

Finally, the intermediate 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 treatment of evolution 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 defined using a Bayesian formulation as

$$\mathcal{A}\big(x_{\rm Bj}, Q^2, z_h, P_{hT}\big|\text{region}\big) = \int \mathrm{d}\{R_i\} \int \mathrm{dPS}\ \mathcal{P}\big(\{R_i\}\big|x_{\rm Bj}, Q^2, z_h, P_{hT}; k_i, k_f, \delta k_T, \varphi, \varphi_i, \varphi_k, \xi, \zeta\big) \times \Theta\big(\{R_i\}\big|\ \text{region}\big)\ \pi\big(k_i, k_f, \delta k, \varphi, \varphi_i, \varphi_k, \xi, \zeta\big|x_{\rm Bj}, Q^2, z_h, P_{hT}\big), \tag{3.1}$$

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.

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 2. Examples of sizes of region indicator ratios corresponding to particular regions of SIDIS. The "×" means "irrelevant or ill-defined." See the text for more details.

Region	R_0	R_1	R_1'	R_2	R_3	R_4
TMD	small	small	×	small	×	×
transition	small	small	×	small	×	×
collinear	small	small	×	large	small (LO pQCD)	small
collinear	small	small	×	large	large (HO pQCD)	small
target	small	large	small	×	×	×
central	small	not small	not small	small	×	×

where $x_{\rm Bj} = Q^2/2P\cdot q$ is the Bjorken scaling variable, Q is the momentum transfer from the incoming lepton, and φ, φ_i, and φ_k are the azimuthal angles of $\boldsymbol{q}_T$, $\boldsymbol{k}_i$, and $\boldsymbol{\delta k}_T$, respectively. The phase space is given by $\mathrm{dPS} \equiv \mathrm{d}k_i\,\mathrm{d}k_f\,\mathrm{d}\delta k_T\,\mathrm{d}\varphi\,\mathrm{d}\varphi_i\,\mathrm{d}\varphi_k\,\mathrm{d}\xi\,\mathrm{d}\zeta$. In essence, the affinity is the integral of the multivariate distribution $\mathcal{P}(\{R_i\}|\ldots)$ with the *proximity function*, chosen to be the Heaviside function, $\Theta(\{R_i\}|\,\text{region})$, that enhances or vetoes the affinity at given SIDIS kinematics as a function of a specific physical region according to Table 2. The integral is also weighted and marginalized over the prior distribution, π, for the internal parton kinematics. In practice, Eq. (3.1) can be implemented by sampling parton momenta from a given choice of prior distribution and the proximity function, and collecting samples for $\{R_i\}$ to estimate the integral using Monte Carlo methods.

To implement the affinity distribution in Eq. (3.1) requires specifying the proximity function, in addition to the Bayesian prior. The ratios R_i depend on the kinematic variables $x_{\rm Bj}$, Q^2, z_h, P_{hT} and the (nonperturbative) partonic parameters k_i, k_f, δk_T, each of order $\mathcal{O}(m)$, where m is a typical hadronic mass scale, and three azimuthal angles φ, φ_i and φ_k, defined such that $\boldsymbol{q}_T\cdot\delta\boldsymbol{k}_T = q_T\,\delta k_T\cos(\varphi-\varphi_k)$, for example. For the prior function we use normal distributed priors for the momenta k_i, k_f, and δk, and flat priors for the angles φ, φ_i, and φ_k. For the partonic momentum fractions ξ and ζ, we sample uniformly from values slightly greater than the kinematic momentum fractions $x_{\rm Bj}$ and z_h, respectively, and independent of external kinematics. The prior function can then be written as

$$\begin{aligned}\pi\big(k_i,k_f,\delta k,\varphi,\varphi_i,\varphi_k,\xi,\zeta\big|x,Q^2,z,P_{hT}\big) \;=\; & k_i(m,\Delta^2)\,k_f(m,\Delta^2)\,\delta k_T(m,\Delta^2)\\ & \times\,\Theta(0<\varphi<2\pi)\,\Theta(0<\varphi_i<2\pi)\,\Theta(0<\varphi_k<2\pi)\\ & \times\,\Theta(x_{\rm Bj}<\xi<x_{\rm Bj}+\epsilon)\,\Theta(z_h<\zeta<z_h+\epsilon)\,, \qquad (3.2)\end{aligned}$$

where the small sampling region is represented by the parameter ϵ, which we set to $\epsilon = 0.1$. To ensure the positivity of k_i, k_f, and δk_T, these are distributed according to the absolute value of the normal distribution $\big|\mathcal{N}(m,\Delta^2)\big|$, with mean $m = 0.5$ GeV and standard deviation $\Delta = 0.5$ GeV. While more sophisticated choices for the these distributions can be made, Eq. (3.2) will be sufficient to illustrate the practical use of affinity. For the proximity function we use Heaviside functions,

$$\Theta\big(\{R_i\}|\text{region}\big) = \prod_i \Theta\big(R_i < R_i^{\max}(\text{region})\big)\,\Theta\big(R_i > R_i^{\min}(\text{region})\big)\,, \qquad (3.3)$$

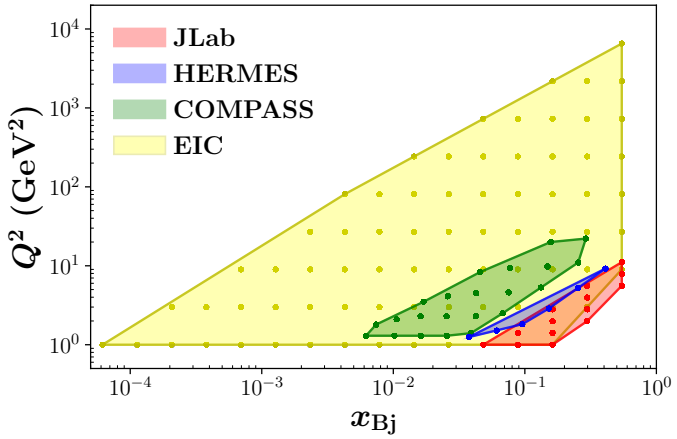

Figure 2. Kinematical reach of Q^2 (GeV^2) versus $x_{\rm Bj}$ for data from existing experiments at Jefferson Lab (red), HERMES (blue), COMPASS (green), and the future EIC (yellow). Bin centers are indicated by filled circles, with each bin representing measurements for different values of z_h and P_{hT}.

where the first Θ 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 $\Theta\big(R_2^{\rm max}({\rm region}) < R_2 < R_2^{\rm min}({\rm region})\big)$ for R_2 to define intermediate values. A special treatment will be used for the central region.

Since for reasonable values for $R_i^{\rm min}$ and $R_i^{\rm max}$ 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 $R_i^{\rm max}({\rm TMD}) = 0.3$, 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 $R_i^{\rm min}({\rm region}) = 0.9$, for $i = 0, 1, 2$. For other ratios we set $R_i^{\rm max}({\rm region}) = 0.3$ and $R_i^{\rm min}({\rm region}) = 3R_i^{\rm max}({\rm region}) = 0.9$. These values in principle depend on kinematics and may vary, so that more fine tuning may be required to delineate the regions with greater accuracy.

4 Applications

Let us now turn to the study of the typical kinematics of the future EIC, which will have a variable *c.m.* energy from $\sqrt{s} = 20$ to 140 GeV. We consider experimental measurements to be performed in the kinematic region shown in Figure 2 where bin centers in $x_{\rm Bj}$ and Q^2 are indicated by dots. We study 7400 bins, as simulated in the Yellow Report on the EIC [41], for π^+ production in $x_{\rm Bj}$, z_h, Q^2, and P_{hT}.

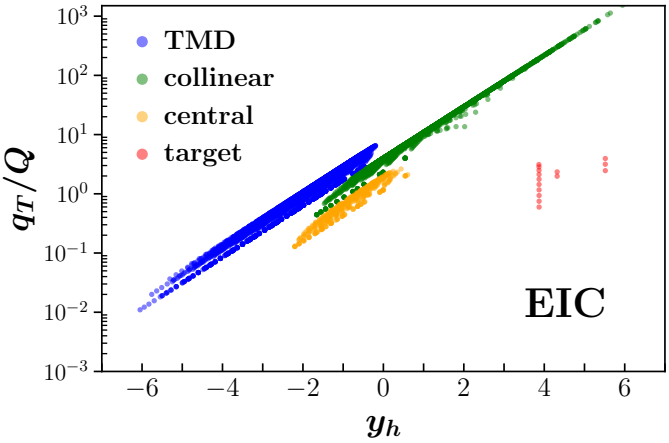

Figure 3. Distribution of bins corresponding to TMD, collinear, central, and target regions as a function of q_T/Q vs y_h. For the visual presentation we multiply q_T/Q of the TMD region by 8, collinear region by 4 and the central region by 2. This makes the groups of points shift vertically, so that they become visible also in the region where they overlap.

In this section we demonstrate the utility of the region indicators introduced in the previous section. After providing a practical definition of affinity that can be used in calculations, we apply this definition to data from existing experiments and to expected data from future facilities. In Figure 2 we illustrate the kinematic reach in $x_{\rm Bj}$ and Q^2 of the experiments that we consider in this paper, namely, Jefferson Lab, HERMES, COMPASS, and the future EIC. Since the EIC covers the largest kinematic range, we expect wider ranges of values for the region indicators compared with the other facilities.

We will plot the kinematical reach of the experiments in terms of the produced hadron rapidity

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

While rapidity is an observable, it can be challenging to measure, in particular at large values where the particle trajectories are close to the beam pipe and neither their energies nor their total momenta can be precisely determined. Thus, pseudorapidity instead of rapidity is often used in experimental work. It is a function of the polar angle between the particle trajectory and the beam axis, and it is ideal for discussions of the 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. Figure 3 shows the kinematics of the EIC projected data, categorized according to affinity values exceeding the threshold of 5% for various fragmentation regions. In order to distinguish between the regions, we use color code for the bins for TMD affinity (blue), collinear affinity (green), central affinity (yellow), and target affinity (red). The initial proton is always in the positive rapidity range, while the produced hadron has either positive or negative rapidity. As discussed in the previous sections, produced hadrons in the negative rapidity range 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

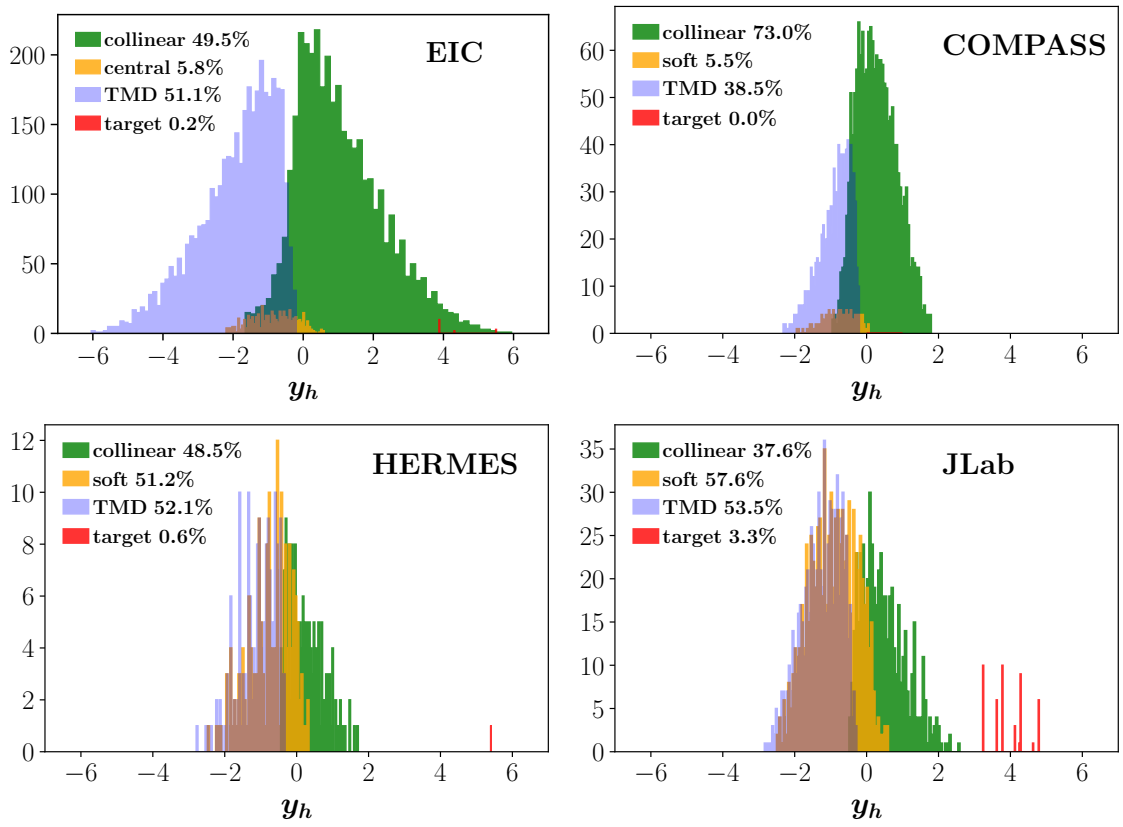

Figure 4. Phase space in rapidity y_h of the produced hadron at the EIC, COMPASS, HERMES, and Jefferson Lab. Histograms corresponding to various kinematic regions are shown. The percentage of bins with affinity of 5% or more for TMD, collinear, central and target regions is indicated in the legend.

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, which is why it is often referred to as the "central region". Finally, the target region is likely to be found in the range of positive hadronic rapidities. Figure 3 demonstrates that the region of central rapidities $y_h \sim 0$ will correspond to an admixture of almost all regions. Although the ratio q_T/Q seems 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]. They 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 smaller mass then the proton, and if P_{hT} is small enough, then z_h is 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 later that values of z_h are also small for the target fragmentation region. In Figure 4 we show the distribution of all kinematics bins accessible at current and future experiments as a function of the produced hadron rapidity, y_h, categorized by the affinity as used in Figure 3. One can see that the majority of data at the EIC will correspond to

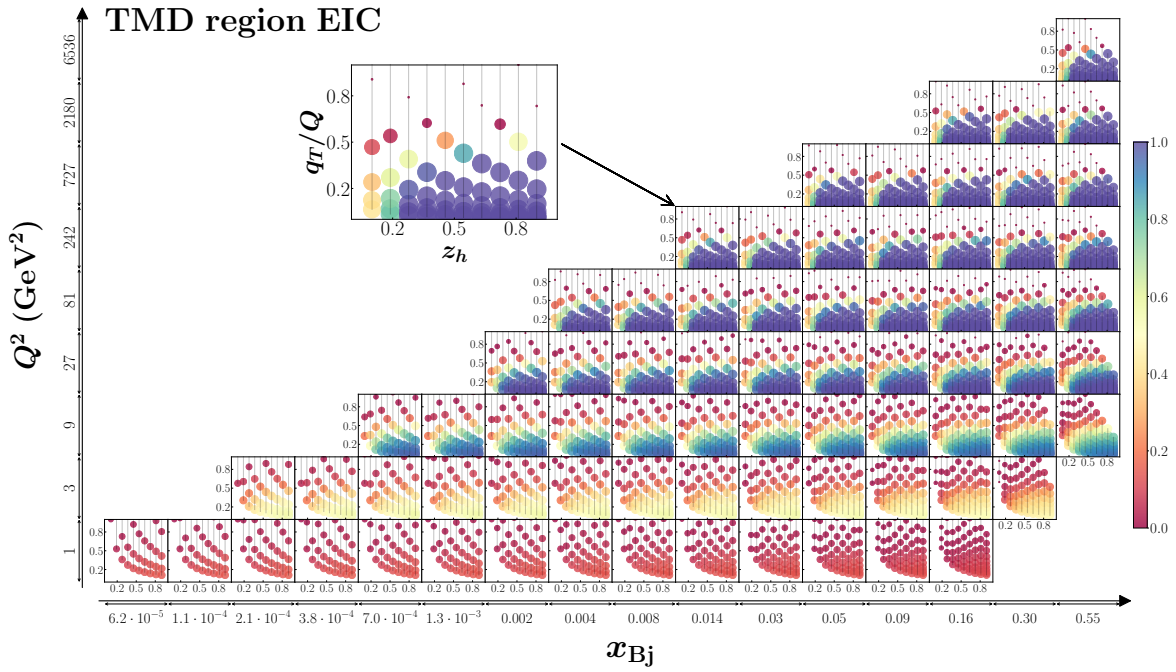

Figure 5. TMD affinity for the EIC. Bin centers are located in the points corresponding to the bin average values of $x_{\rm Bj}$ and Q^2 (GeV^2). In each of these bins, various values of z_h and q_T/Q can be measured. In each bin of fixed z_h and q_T/Q we plot the affinity as a dot with size proportional 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 TMD affinity.

either TMD or collinear QCD fragmentation region. There are also some small fractions of soft or target fragmentation region events.

The relative portions of events from various regions change at other experiments. Figure 4 shows histograms corresponding to COMPASS, HERMES, and Jefferson Lab kinematics. One can see that for lower energy experiments at Jefferson Lab, where the measurements are at $\sqrt{s} = 4.6$ 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 experiments we consider. Notice 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 subsections.

4.1 TMD region

TMD affinity is calculated using Eq. (3.1) by requiring R_0, R_1, and R_2 to be small. The results for the bins of the EIC is shown in Figure 5. One can see that bins with relatively large x_{Bj} and Q^2 (and relatively high z_h and P_{hT}) are particularly important for the TMD

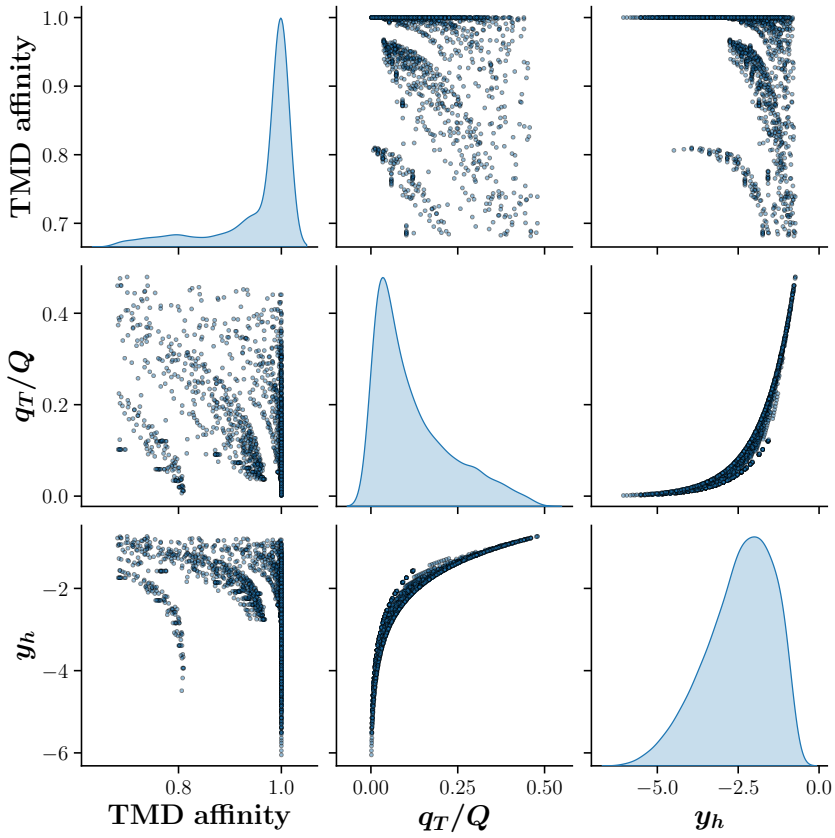

Figure 6. Pair-wise correlations among TMD affinity greater than 70%, q_T/Q, and the produced hadron rapidity y_h for the EIC. The plots on the diagonal represent cumulative distributions corresponding to TMD affinity, q_T/Q, and y_h. The histograms on the diagonal help to visualize the relative abundance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

factorization description. In terms of the applicability of TMD factorization, it means that q_T/Q becomes small enough for this factorization to be valid. We estimate 2325 out of 7400 bins to have TMD affinity of 68% or higher and 1739 bins to have TMD affinity of 95% or higher. As we will discuss later, the rest of the data (or at least part of it) will correspond to different mechanisms, such as that of the collinear factorization scheme.

In Figure 6 we show correlation plots of TMD affinity, q_T/Q, and the produced hadron rapidity y_h. It appears clearly that the correlation between y_h and q_T is very strong. In particular, 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 definition of the TMD affinity, 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 the plots one can see 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 y_h appear to be good proxies for TMD affinity, especially when used in combination. In fact, considering only one or the other indicators may considerably limit

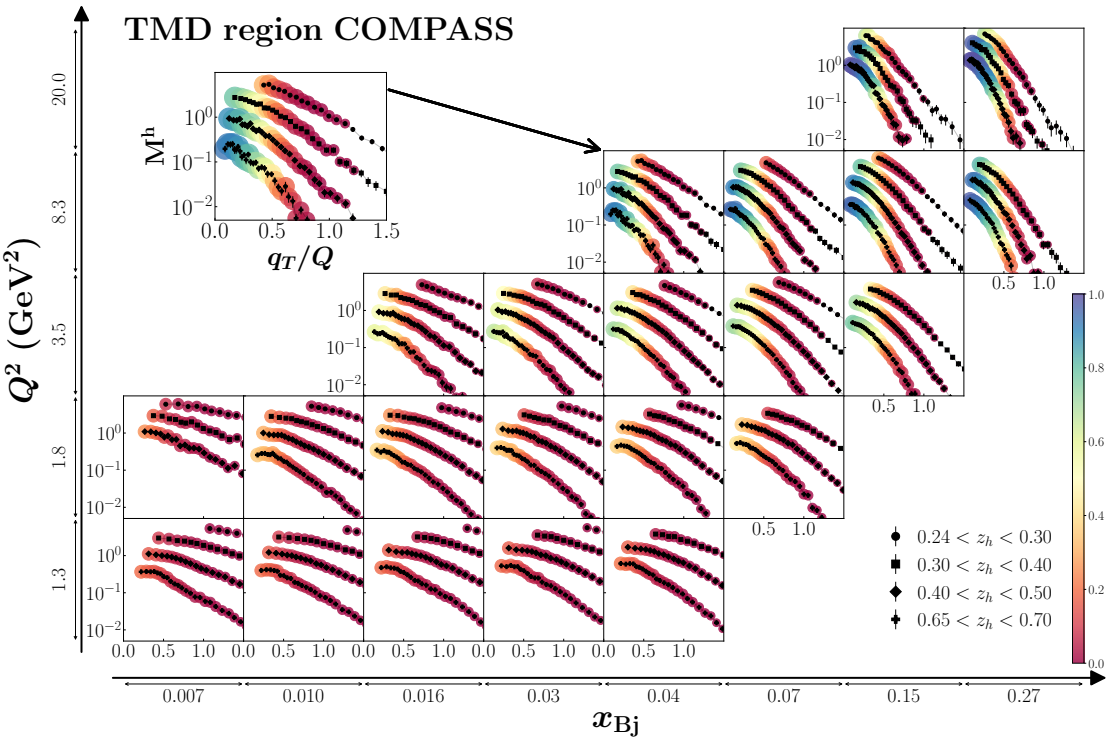

Figure 7. TMD region for COMPASS data [43]. The data (black symbols) on M^h for h^+ obtained in muon-deuteron SIDIS are shown as function of q_T/Q. In each panel there are up to four bins in z_h, the data-set at the top (bottom) corresponds to the lowest (highest) range in z_h and vice versa.

the accuracy in establishing the region boundaries. This was also reported in Ref. [42], where an algorithm based on region indicator ratios involving both q_T/Q and the observed hadron rapidity, y_h, was introduced for the study of $e^+e^- \to hX$ processes. One can see that requiring $y_h < -2.5$ will ensure that the TMD affinity is larger than ∼80%, while the cut $q_T/Q < 0.25$ results in a TMD affinity greater than ∼80%. Moreover, as the correlation between y_h and q_T/Q is very strong, it will be interesting to have information on y_h directly from experimental measurements.

As mentioned previously, we do not attempt a phenomenological description of the experimental data; it is therefore interesting to compare the affinity data selection to those of existing analyses, such as those presented in Refs. [25, 26]. If we apply Eq. (2.7) to the EIC data, then 2116 bins survive out 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.

The next-to-leading-logarithmic (NLL) precision analysis of SIDIS, Drell-Yan, and Z-boson production data in Ref. [24] used the following selection criteria:

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

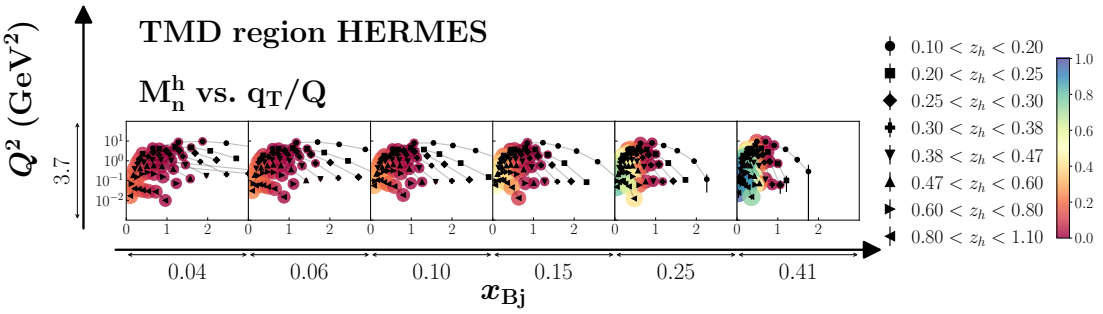

Figure 8. TMD region for HERMES multiplicities [45] as functions of P_{hT} (GeV). Data for M_n^h are shown for π^+ produced off a hydrogen target. In each panel there are up to eight z_h bins with the top (bottom) data-set corresponding to the lowest (highest) values of z_h.

If we apply these cuts, 2148 bins (out of 7400 EIC bins) 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 phenomenology of TMDs are

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

Notice 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 therefore be difficult to describe in a TMD approach.

If we apply these cuts, 671 bins survive and 396 of those belong to TMD affinity of 68% or higher. It is interesting to note that only 50% of the data selected by 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 [43] of charged hadrons produced in collisions of 160 GeV longitudinally polarized muons scattered off a deuteriumn target in the typical SIDIS kinematics $Q^2 > 1\,\mathrm{GeV}^2$, $W^2 > 25\,\mathrm{GeV}^2$, $0.003 < x_{\mathrm{Bj}} < 0.7$, $0.1 < y < 0.9$, $0.2 < z_h < 1$, where $W^2 = (P+q)^2$ and $y = P\cdot q/P\cdot \ell$. The multiplicity [43] is defined as

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

In Figure 7 we present the bins explored by COMPASS and the data corresponding to the positive hadron multiplicity. TMD affinity is shown on top of 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. One can see that, as in the case of the EIC, higher Q bins have higher TMD affinity for low values of P_{hT}. For higher values of z_h and for higher values of x_{Bj}, one expects higher TMD affinity, as seen in Figure 7.

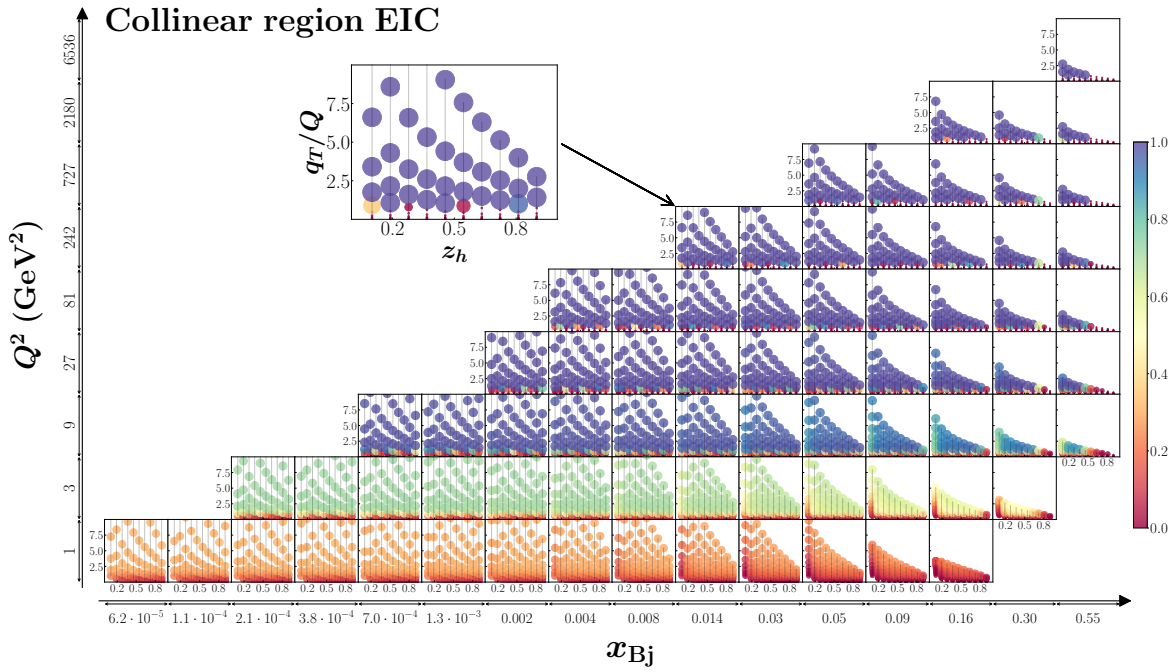

Figure 9. Collinear affinity for the EIC. Bin centers are located in the points corresponding to the bin average values of $x_{\rm Bj}$ and Q^2 (GeV^2). In each of these bins, various values of z_h and q_T/Q can be measured. In each bin of fixed z_h and q_T/Q we plot the affinity as a dot with size proportional 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.

COMPASS data was used in the phenomenological study of Ref. [25]. For the case of h^+ production, 138 bins survive after Eq. (2.7) cuts. 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. At this point it seems to be clear that additional phenomenological work is needed to delineate the TMD region more precisely.

The HERMES Collaboration measured the multiplicity of pion or kaon production in the scattering of 27.6 GeV positrons off proton and deuteron targets in the SIDIS kinematics $Q^2 > 1\,\text{GeV}^2$, $W^2 > 10\,\text{GeV}^2$, $0.023 < x_{\rm Bj} < 0.4$, $y < 0.85$, $0.2 < z_h < 0.7$. The measured multiplicity [45] is

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

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

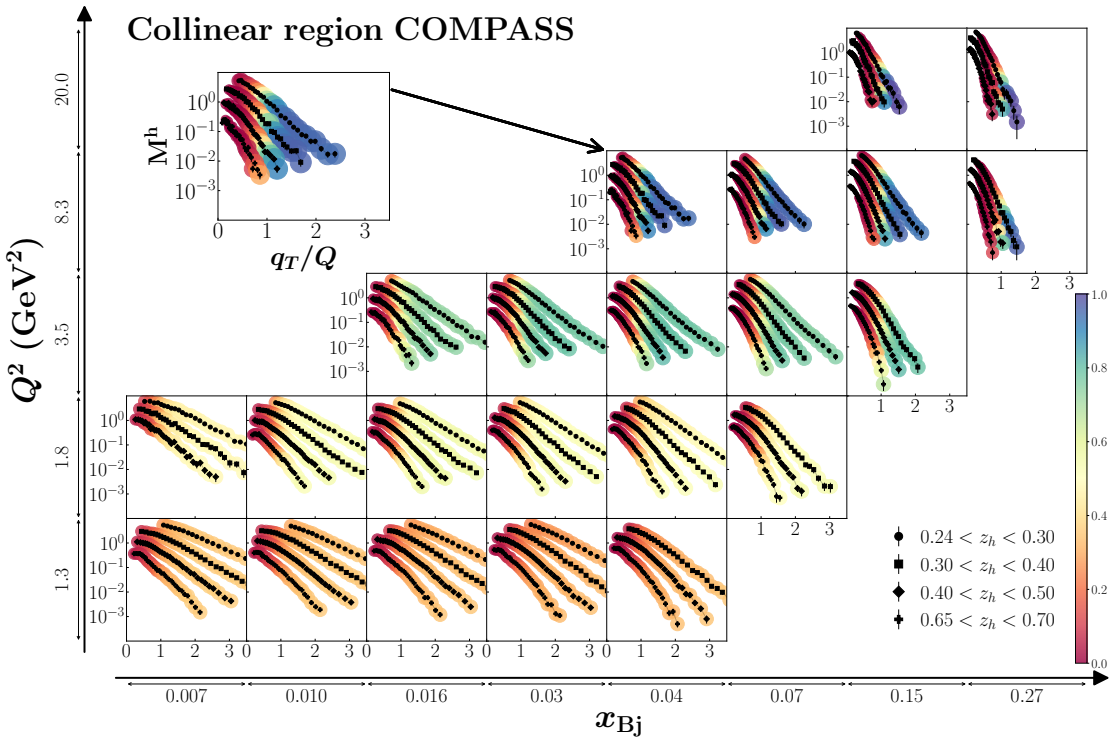

Figure 10. Collinear region for COMPASS multiplicities [43]. The data (black symbols) on M^h for h^+ obtained in muon-deuteron SIDIS are shown as function of q_T/Q. In each panel there are up to four bins in z_h, the data-set at the top (bottom) corresponds to the lowest (highest) range in z_h and vice versa. The yellow band at the bottom of each panel indicate the region where $q_T/Q > 1$.

the legend as a function of P_{hT}. TMD affinity is shown on top of each data point. We see the affinity is larger for higher z_h and $x_{\rm Bj}$ and relatively small P_{hT}. HERMES data was also used in the phenomenological study of Ref. [25]. For the case of π^+ production off a proton target, 34 bins survive after Eq. (2.7) cuts.

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 negligible sensitivity on parton intrinsic transverse momentum. In terms of region indicators, this means that R_2 becomes large while R_4 remains small. This region has been recently discussed in [46, 47] where a significant tension was found between data and theory at COMPASS and HERMES kinematics, with deviations up to one order of magnitude. Ref. [46] showed that such deviations are marginally improved by the inclusion of $\mathcal{O}(\alpha_S^2)$ corrections. Similar observations have been made in Ref. [48], 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 how to interpret existing and future data in the large transverse momentum regime. In Figure 9 we present the affinity results at the EIC kinematics showing the ranges for $q_T/Q < 10$ to focus on the larger transverse

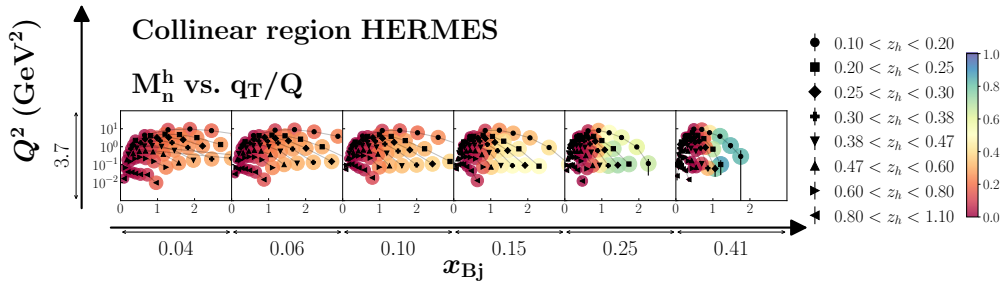

Figure 11. Affinity to the collinear region for the HERMES multiplicity data [45].

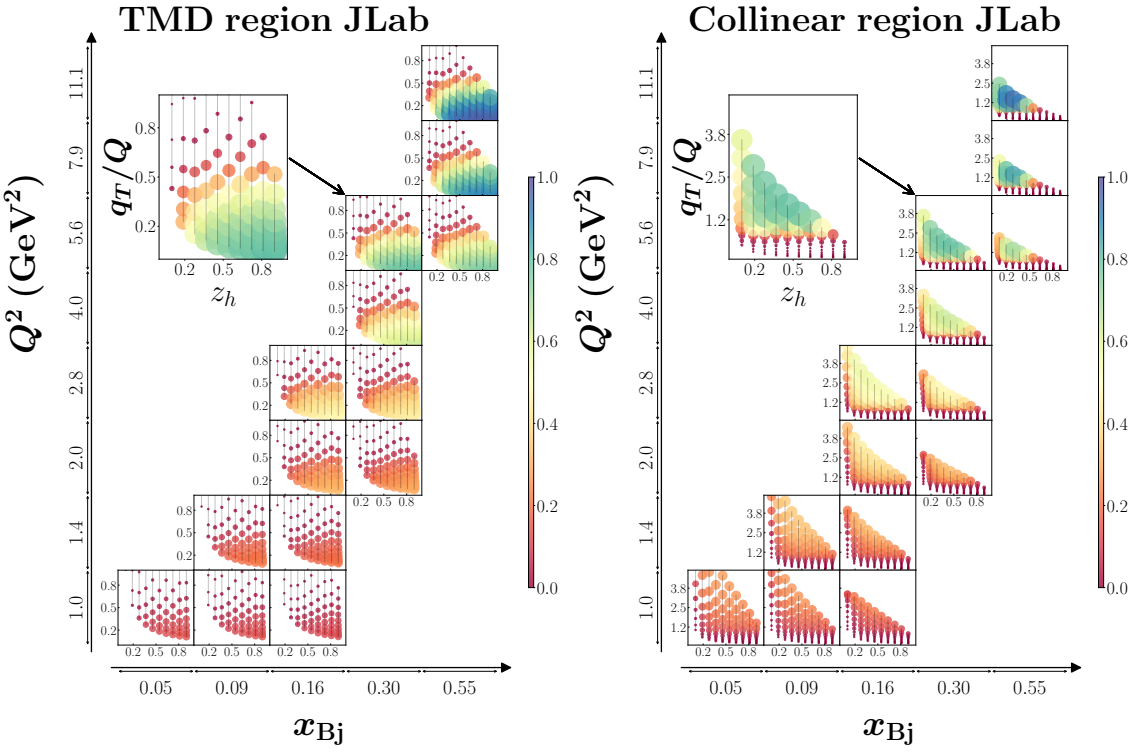

Figure 12. Affinity to the TMD and collinear regions at Jefferson Lab.

momentum regime. We estimate that 1750 out of 7400 bins have a collinear affinity of 68% and higher, while 1170 bins have a collinear affinity of 95% and higher. As expected, the collinear affinity is larger for increasing values of Q and q_T/Q while becoming smaller below $q_T/Q < 1$. Notice 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 taken for granted at low values of Q.

The collinear affinity values at COMPASS kinematics are shown in Figure 10 along

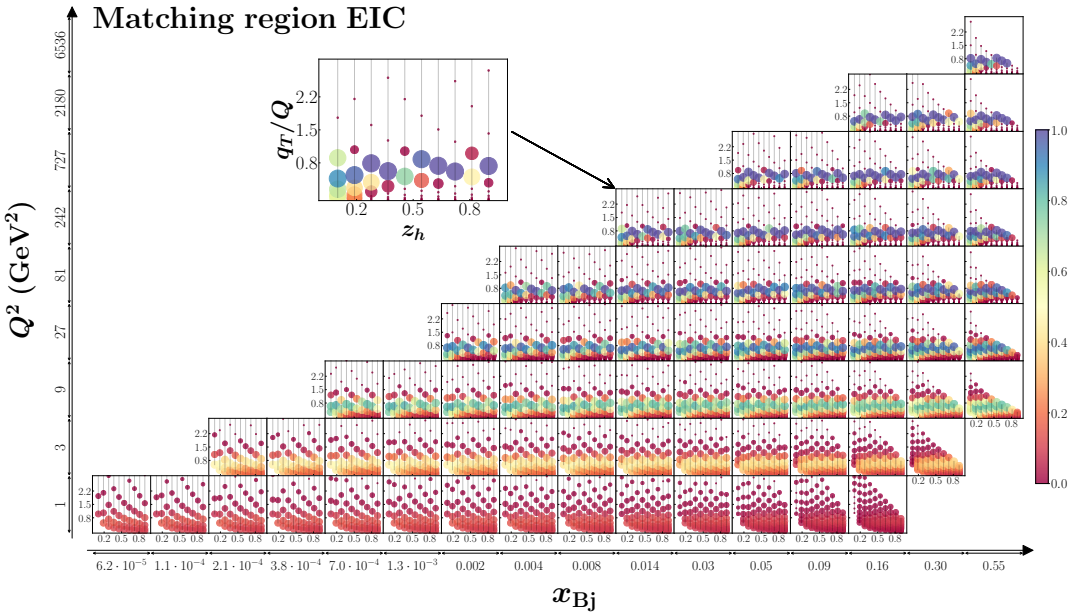

Figure 13. Affinity for the TMD-collinear matching region for the EIC. Bin centers are located in the points corresponding to the bin average values of $x_{\rm Bj}$ and Q^2 (GeV^2). In each bin of fixed z_h and q_T/Q we plot the affinity as a dot with size proportional to the corresponding affinity value.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

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. [46, 47]. In contrast the affinity values become notably larger for large values of Q. However, those regions are close to the edge of the phase space, and it is rather likely that threshold effects need to be taken into account to achieve a satisfactory description of the cross section in this particular region [49]. Similar results can be found in Figure 11 and Figure 12 at HERMES and JLab 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 one should view the results only as rough estimates.

4.3 TMD-collinear matching region

The TMD-collinear matching region covers the range of q_T values such that $\Lambda_{QCD} \ll q_T \ll Q$, and represents the region where one would expect a smooth transition between the TMD and collinear regimes. In this intermediate region the description the data might be possible in either TMD or collinear schemes. Traditionally, this is where one would implement the $W + Y$ construction by Collins-Soper-Sterman (CSS) [50] 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.

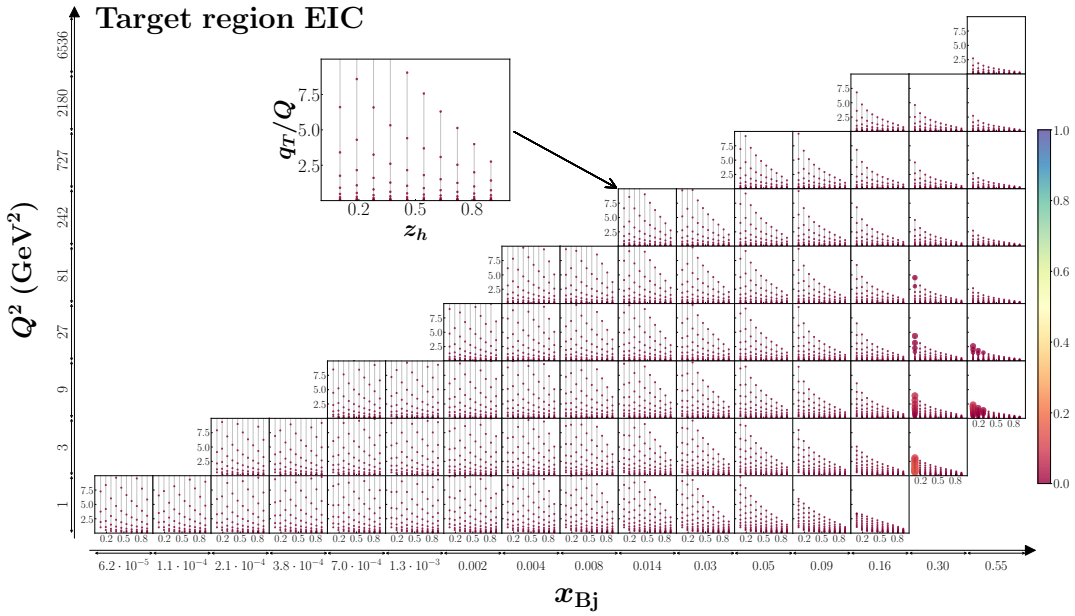

Figure 14. Affinity for target fragmentation region at the EIC.

In fact, matching the SIDIS cross section across TMD and collinear regions turns out to be rather challenging, especially with regard to low energy experiments, as discussed by several authors (see for example Refs. [20, 28, 51]), where different matching procedures have been proposed. We do not enter into the discussion here, which would go well beyond the purposes of this work. Nevertheless, it is important to stress that, once a certain definition of this region is chosen, our affinity algorithm will be able to identify and correctly map it, in exactly the same way as for the TMD and collinear regions.

In Figure 13 we show the matching region for the EIC as determined by the affinity tool. As expected, it correctly covers the range of intermediate values of q_T, and it turns out to be relevant at rather large values of Q^2 corresponding to moderate and large values of x_{Bj}.

4.4 Target and central regions

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

As we 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 rapidities, close to the beam. The experimental measurement of such hadrons is challenging; however, the study

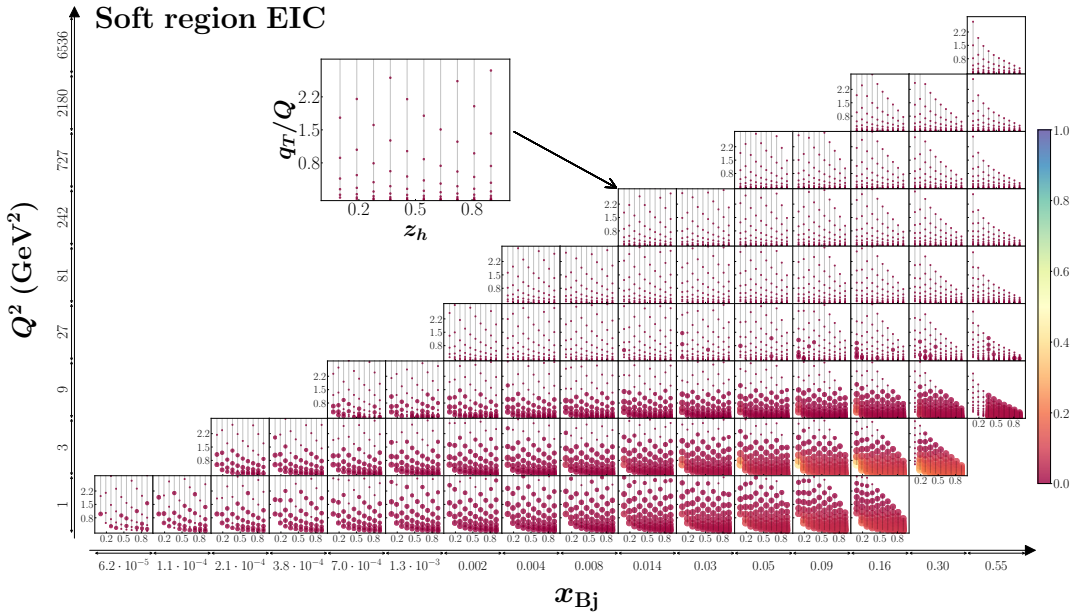

Figure 15. Affinity for central fragmentation region at the EIC.

of the target fragmentation is very interesting from both phenomenological and theoretical points of view. These processes are usually described in terms of fracture functions [14–17] which are conditional probabilities of producing a hadron with energy fraction z off the remnant of a nucleon that carries $1 - x_{\mathrm{Bj}}$ fraction of the nucleon's momentum P. For such hadrons the notion of z_h is not well defined as $z_h \approx 0$ in this case. Fracture functions are important ingredients for the description of diffractive hadro-production and there are attempts of derivation of factorization formalisms for those processes, see Ref. [52]. Factorized formulas for TMD fracture functions were conjectured in Ref. [16] and the evolution equations and 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 deserves attention from the experimental community, in particular for the planning of the EIC detectors [41].

Our estimates of affinity to the target fragmentation region for the EIC kiematics are shown in Figure 14. One can see that the target region is characterized by relatively large values of x_{Bj} and small values of z_h. Obviously a more detailed study of target fragmentation region is needed in order to fully realize the potential of the EIC. Such a study may include a detailed Pythia generation of the hadrons produced in the target fragmentation region together with the Geant detector simulation.

The last region we will discuss in this section is the central 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 off colored partons possible. In the event generators [53] one employs a Lund string model

that describes fragmentation as fracturing of a flux tube that is created between by the colored quark and the remnant of the nucleon, see for instance Refs. [40, 54] 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 very interesting to reconcile fragmentation models, especially the ones that include spin, e.g., the Lund string model [55] or Nambu–Jona-Lasinio model of fragmentation [56, 57], with results of QCD. A recent attempt to use Feynman-graph structures is presented in Ref. [21]. In addition, the central region is interesting from the factorization point of view, as the soft radiation plays an important role for proofs of factorization, for instance of TMD factorization [11, 37].

Even though the central region is incorporated in any Monte Carlo generator used in experimental analyses, this region was not yet explored experimentally in a very detailed fashion, and future experimental studies are very desirable. In order to do it effectively, the studies of the rapidity distribution of hadrons should be performed. We estimate that for the central region we have a majority of events in the low-Q^2 and low-P_{hT} region, see 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 we investigated in this paper. We expect that in the future more detailed experimental, theoretical, and phenomenological studies will allow one to delineate this region more precisely [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.

First 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. The affinity is defined with respect to the values $R_0^{\max}$, $R_1^{\max}$, and $R_2^{\max}$ varying in the range $(0.05, 1.25)$. For each set of these values we generate the corresponding set of affinities for each of the 7400 bins and store them. There are just under 14000 total generated data-sets, which are of the order of several gigabytes; therefore, the direct usage of this information is not feasible.

We use the `TensorFlow` [58] framework to create and train four neural networks that predict TMD, collinear, target, and central affinity regions with seven input values $x_{\rm Bj}$, z_h, P_{hT}, Q^2, $R_0^{\max}$, $R_1^{\max}$, and $R_2^{\max}$. Neural networks in general are difficult to configure as there are many parameters (called hyperparameters) to be adjusted, which comprise the number and the width of hidden layers and even the algorithms of minimization to be used. A good hyperparameter combination can highly improve the performance of the network. We employ the hyper-band `KerasTuner` [59] for tuning the hyperparameters for our networks.

We train four separate neural nets to predict the affinity value to TMD, collinear, target, and central regions. `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 vectors

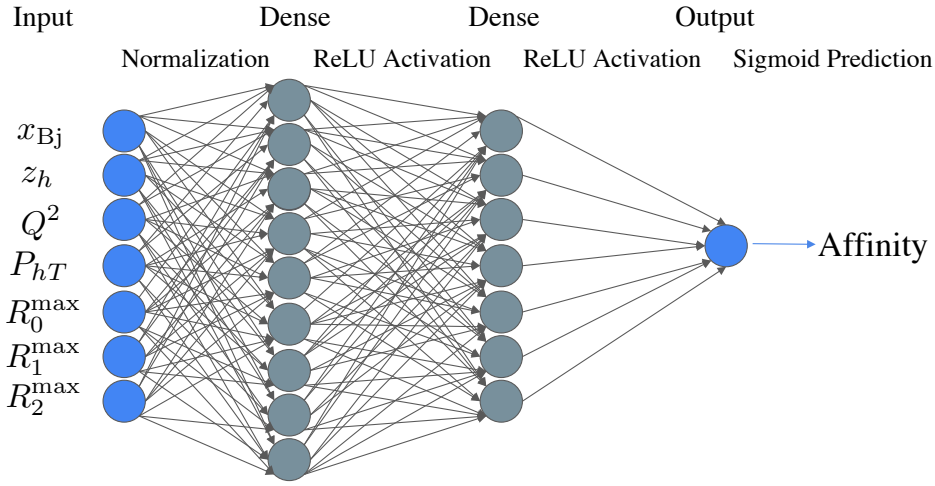

Figure 16. Illustration of the architecture of the neural network devised for calculating affinity. Each vector represents a trainable parameter that each neural network must learn through training. Input, hidden layers, and output are represented by the (stack of) blobs. The normalization layer and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

correspond to the parameters that each neural network must learn through training; input, output and hidden layers are represented by the (stack of) blobs. Training is an iterative process of updating weights through differential tensor calculus by showing the model input examples accompanied with an affinity value.

The first hidden layer has 576 neurons for TMD, 960 for collinear, 896 for target, and 576 for central regions. The second hidden layer has 160 neurons for TMD, 544 for collinear, 256 for target, and 736 for 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, i.e., 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, weights cannot be updated.

In order 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 $(0, 1)$. It is the cumulative distribution function of a logistic distribution, which 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 hide about 20% of the training data from the network learning, which provides a validation set of input examples to test the network at the end of each training iteration. We use the validation predictions to calculate the mean squared error (MSE), for which

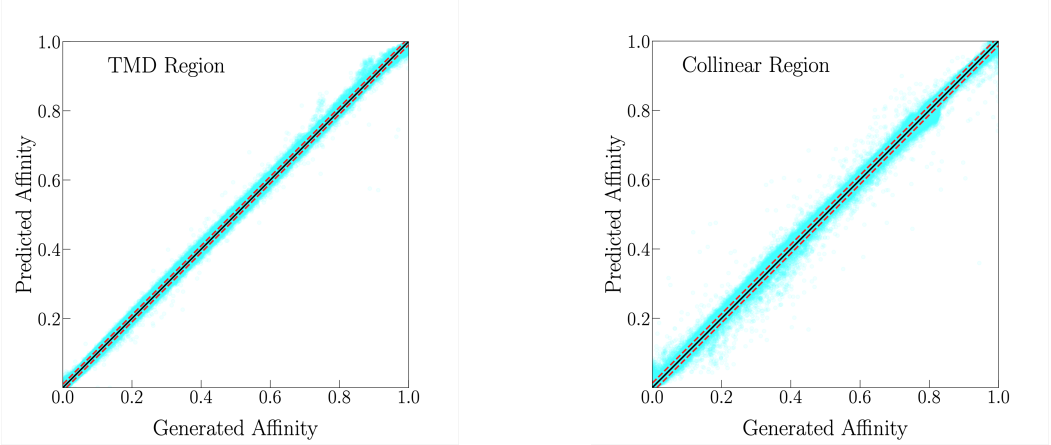

Figure 17. Scatter plot of predictions vs. test data. The left (right) panel corresponds to TMD (collinear) affinity. In each panel the black central line represents the perfect prediction, corresponding to zero residuals, while the points located between the two parallel red lines contain 95% of all residuals.

back propagation minimizes. Minimization of the validation loss optimizes the weight parameters. The model architecture that provides the smallest validation MSE (loss function) was considered best.

We used 80% of the data for training and 20% for validation in order to avoid overfitting and to check the accuracy of our predictions. Each model training lasted between 50 and 500 epochs (i.e., iterations over training data), and we used MSE and Adam minimizer for training. After the training, the resulting neural networks are saved and can be further used in `TensorFlow` applications.

Finally, 100 data-sets of 7400 input examples and corresponding affinities were independently generated for testing, which allowed for visual and measurable comparisons of networks. Plots of predicted and generated affinity values provide a visualization of the error distribution, having a center-line representing perfect predictions and a cloud of points depicting variations in the error. Calculating the bin width necessary to contain a center 50% of residuals provides a measure of the error distribution (here"residual" indicates the difference between the values of generated and predicted affinity). In Figure 17 we provide a visualization of predicted (output) affinity values versus test (input) affinity values for TMD and collinear network models. The plots show two parallel red lines, equally spaced about the black line of zero residuals, which contains 95% of all residuals. The width of this center bin is approximately 0.02 (0.026), containing 37000 residuals within ± 0.01 (0.013) for TMD (collinear). 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.

Other visual and measurable comparisons were also considered. Choosing the best neural network is not always straightforward or deterministic. For each hyper search, the networks with the five smallest validation MSE were summarized. When the best network was unclear, the independent test set was used to inform the architecture chosen to be

further trained and implemented in the affinity tool.

With the affinity tool implemented, validated and tested, we are now ready to make it publicly available. In order to ease the burden of installation of any additional software for the users of the tool, we use a `Google Colab` notebook that allows the user to launch the tool from a web browser without any 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 the values for $R_0^{\max}$, $R_1^{\max}$, and $R_2^{\max}$ and generate plots corresponding to EIC TMD, target, central, and collinear regions using neural net models trained for these regions. The interactive notebook with user instructions can be found in [60].

6 Conclusions

SIDIS measurements offer a great opportunity to learn about the partonic structure of nucleons. For a correct phenomenological interpretation of the information they encode, it is fundamental to develop tools that allow us to connect the experimental data 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, $\{R_i\}$, introduced in Refs. [20, 22] to quantify our confidence in the proximity of SIDIS observables to a particular physical mechanism in terms of a new tool called "affinity". This facilitates the separation of the 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 estimates of 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. Little can be said for these kinematic ranges, however, where a well-defined factorization theorem is not available. They will indeed deserve further phenomenological studies as well as to a more theoretical level.

Our affinity tool shows that a large portion of experimental bins can be associated to either TMD or collinear physics, for all considered experiments, and especially for the future EIC. Lower energy experiments such as Jefferson Lab 12 GeV, however, show a non-negligible admixture of central and target fragmentation events.

The publicly available interactive tool developed for this study is based on a neural network model trained with machine learning techniques which allows a fast evaluation of affinity. The architecture of the net consists of four layers: input, output and two hidden intermediate layers.

The affinity tool can be applied in phenomenological analyses to select kinematic bins that are sensitive to the kinematic region of interest. It can be also used to guide the development of new SIDIS experiments and to incorporate the region estimator into experimental analyses. For this reason, we also provide an interactive tool that allows the study of affinity according to any personal choice on how to separate the kinematic regions. The affinity interactive tool is freely available as a `Google Colab` notebook, which can easily be accessed and run from any browser, without the need of additional software.

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.* **B441** (1995) 234 [hep-ph/9412283].

[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.* **B461** (1996) 197 [hep-ph/9510301].

[3] A. Bacchetta, M. Diehl, K. Goeke, A. Metz, P.J. Mulders et al., *Semi-inclusive deep inelastic scattering at small transverse momentum*, *JHEP* **0702** (2007) 093 [hep-ph/0611265].

[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.* **B504** (1997) 345 [hep-ph/9702281].

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

[6] D. Boer, *Angular dependences in inclusive two-hadron production at BELLE*, *Nucl.Phys.* **B806** (2009) 23 [0804.2408].

[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].

[8] J.C. Collins and D.E. Soper, *Parton Distribution and Decay Functions*, *Nucl. Phys.* **B194** (1982) 445.

[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.

[10] X. 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].

[11] J. Collins, *Foundations of perturbative QCD*, Cambridge University Press (11, 2013).

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

[13] E.L. Berger, *SEMI-INCLUSIVE INELASTIC ELECTRON SCATTERING FROM NUCLEI*, in *NPAS Workshop on Electronuclear Physics with Internal Targets, SLAC*, p. 82, 1987.

[14] L. Trentadue and G. Veneziano, *Fracture functions: An improved description of inclusive hard processes in QCD*, *Phys. Lett.* **B323** (1994) 201.

[15] M. Grazzini, L. Trentadue and G. Veneziano, *Fracture functions from cut vertices*, *Nucl. Phys.* **B519** (1998) 394 [hep-ph/9709452].

[16] M. Anselmino, V. Barone and A. Kotzinian, *SIDIS in the target fragmentation region: Polarized and transverse momentum dependent fracture functions*, *Phys. Lett.* **B699** (2011) 108 [1102.4214].

[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 [1903.00809].

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

[19] S.J. Joosten, *Fragmentation and nucleon structure in semi-inclusive deep-inelastic scattering at the HERMES experiment*, Ph.D. thesis, Illinois U., Urbana, 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.* **B766** (2017) 245 [1611.10329].

[21] J. Collins and T.C. Rogers, *Graphical Structure of Hadronization and Factorization in Hard Collisions*, 1801.02704.

[22] M. Boglione, A. Dotson, L. Gamberg, S. Gordon, J.O. Gonzalez-Hernandez, A. Prokudin et al., *Mapping the Kinematical Regimes of Semi-Inclusive Deep Inelastic Scattering*, *JHEP* **10** (2019) 122 [1904.12882].

[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 [1312.6261].

[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 [1703.10157].

[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 [1912.06532].

[26] A. Bacchetta, V. Bertone, C. Bissolotti, G. Bozzi, F. Delcarro, F. Piacenza et al., *Transverse-momentum-dependent parton distributions up to N^3LL from Drell-Yan data*, *JHEP* **07** (2020) 117 [1912.07550].

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

[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.* **D94** (2016) 034014 [1605.00671].

[29] J.C. Collins and D.E. Soper, *Back-To-Back Jets in QCD*, *Nucl. Phys. B* **193** (1981) 381.

[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.

[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].

[32] S. Aybat and T.C. Rogers, *TMD Parton Distribution and Fragmentation Functions with QCD Evolution*, *Phys.Rev.* **D83** (2011) 114042 [1101.5057].

[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 [1412.3820].

[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 [1111.4996].

[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 [1402.0869].

[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.

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

[38] T. Sjostrand, S. Mrenna and P.Z. Skands, *PYTHIA 6.4 Physics and Manual*, *JHEP* **05** (2006) 026 [hep-ph/0603175].

[39] G. Corcella, I.G. Knowles, G. Marchesini, S. Moretti, K. Odagiri, P. Richardson et al., *HERWIG 6: An Event generator for hadron emission reactions with interfering gluons (including supersymmetric processes)*, *JHEP* **01** (2001) 010 [hep-ph/0011363].

[40] B. Andersson, *The Lund model*, vol. 7, Cambridge University Press (7, 2005), 10.1017/CBO9780511524363.

[41] R. Abdul Khalek et al., *Science Requirements and Detector Concepts for the Electron-Ion Collider: EIC Yellow Report*, 2103.05419.

[42] M. Boglione and A. Simonelli, *Kinematic regions in the $e^+e^- \to h\,X$ factorized cross section in a 2-jet topology with thrust*, 2109.11497.

[43] COMPASS collaboration, *Transverse-momentum-dependent Multiplicities of Charged Hadrons in Muon-Deuteron Deep Inelastic Scattering*, *Phys. Rev. D* **97** (2018) 032006 [1709.07374].

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

[45] 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 [1212.5407].

[46] 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 [1903.01529].

[47] 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 [1808.04396].

[48] 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 [1901.06916].

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

[50] J.C. Collins, D.E. Soper and G. Sterman, *Factorization of hard processes in qcd*, *Adv. Ser. Direct. High Energy Phys.* **5** (1988) 1 [hep-ph/0409313].

[51] 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 [1801.01480].

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

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

[54] B. Andersson, G. Gustafson, G. Ingelman and T. Sjostrand, *Parton Fragmentation and String Dynamics*, *Phys. Rept.* **97** (1983) 31.

[55] 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 [1802.00962].

[56] 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 [0906.5362].

[57] H.H. Matevosyan, A. Kotzinian and A.W. Thomas, *Monte Carlo Implementation of Polarized Hadronization*, *Phys. Rev. D* **95** (2017) 014021 [1610.05624].

[58] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro et al., *TensorFlow: Large-scale machine learning on heterogeneous systems*, 2015.

[59] T. O'Malley, E. Bursztein, J. Long, F. Chollet, H. Jin, L. Invernizzi et al., "Keras tuner." https://github.com/keras-team/keras-tuner, 2019.

[60] M. Boglione, M. Diefenthaler, S. Dolan, L. Gamberg, W. Melnitchouk, D. Pitonyak et al., "Affinity tool." https://colab.research.google.com/github/QCDHUB/SIDIS-Affinity/blob/main/interactive_affinity_tool.ipynb, 2021.