

Deep neural network extraction of unpolarized transverse momentum distributions

I. P. Fernando^{*} and D. Keller[†]

Department of Physics, *University of Virginia Charlottesville, Virginia 22904, USA*

 (Received 7 November 2025; accepted 24 April 2026; published 22 May 2026)

Building on the first-ever application of neural networks in transverse momentum distribution (TMD) phenomenology—“Extraction of the Sivers function with deep neural networks”—we now present a momentum-space, physics-informed deep-learning framework for the direct extraction of unpolarized transverse-momentum-dependent parton distribution functions (TMDs) from fixed-target Drell-Yan data (E288, E605). Rather than transforming to impact-parameter space, we remain in $k_\perp$ and embed a normalized integrand $s(x, k_\perp; Q)$ whose autoconvolution produces the observed q_T spectra. We introduce two phenomenological objects in this data-driven approach: the transverse structure kernel $S(q_T, x_1, x_2; Q)$ and the intrinsic-transverse-momentum profile $s(x, k_\perp; Q)$. The extraction proceeds in two steps. Stage I learns the structure kernel $S(q_T, x_1, x_2; Q_M)$ by regressing the cross section with known kinematic prefactors and charge-weighted parton distribution function (PDF) combinations factored out; experimental and PDF uncertainties are propagated with Monte Carlo replicas. Stage II reconstructs $s(x, k_\perp; Q)$ with an end-to-end differentiable $k_\perp$ quadrature layer. Applied to Fermilab cross-section data from experiments E288 and E605, the method reproduces the measured q_T spectra across the Q_M bins and yields x - and Q -dependent TMDs that broaden with Q , with uncertainty bands that consistently propagate experimental, PDF, algorithmic, and methodological components. The approach is minimally biased (no factorized Ansätze and no b_T transform) and provides a transferable template for polarized TMDs and related quantum chromodynamics inverse problems.

DOI: [10.1103/p3sg-k524](https://doi.org/10.1103/p3sg-k524)

I. INTRODUCTION

Transverse-momentum-dependent parton distribution functions (TMDs) provide a fully multidimensional description of hadron structure, encoding the probability densities of partons as functions of both the longitudinal momentum fraction x and the intrinsic transverse momentum $k_\perp$ [1–3]. The TMD factorization formalism [4–14] enables the separation of perturbatively calculable hard coefficients from nonperturbative TMD distributions in processes differential in transverse momentum. Recent comprehensive extractions of unpolarized TMDs can be found in Refs. [15–32], polarized TMDs in Refs. [33–36], and lattice QCD calculations in [37]. The unpolarized TMD, $f_{1,q/N}(x, k_\perp; \mu, \zeta)$, defines the baseline partonic distribution from which all spin-dependent TMDs are normalized and interpreted. Moreover, $f_{1,q/N}$ is essential

for the quantitative description of semi-inclusive deep-inelastic scattering (SIDIS), Drell-Yan (DY), and vector-boson production across both fixed-target and collider kinematics [38–40], where μ denotes the renormalization scale and ζ the rapidity evolution scale.

Historically, extractions of unpolarized TMDs have been carried out in both transverse-momentum space ($k_\perp$) and impact-parameter space (b_T). The modern b_T -space formalism, established within the Collins-Soper-Sterman (CSS) framework [10,41], transforms the transverse convolutions into simple products, thereby rendering TMD evolution more tractable and compact. This has made the b_T representation the current standard approach for contemporary global analyses. High-precision extractions in b_T space have now achieved NNLL to N³LL accuracy, providing excellent agreement with extensive experimental cross-section data sets [16,18,22,26,42,43] as well as transverse single spin asymmetries (Sivers) [33,34,44–46] and (Collins and Transversity) [47–49]. However, many of the competing results for $f_{1,q/N}$ are inconsistent with each other, highlighting the need for a more universal and robust extraction methodology.

Nevertheless, the transverse momentum itself is what mostly closely connects to experimental observables. Performing extractions natively in $k_\perp$ space offers the

^{*}Contact author: ishara@virginia.edu

[†]Contact author: dustin@virginia.edu

most transparent link to measured spectra and can therefore exploit the full information content of the data, especially when combined with deep neural networks (DNNs) to capture nontrivial correlations and nonlinear dependencies. However, such an approach poses significant analytical and computational challenges due to the inherently two-dimensional angular convolutions in transverse momentum [5,50,51].

Early phenomenological studies performed directly in $k_\perp$ space typically adopted simple, nonevolving Gaussian parametrizations for the transverse-momentum distribution, with widths assumed to be independent of x (and, in fragmentation, of z) [52,53]. These models provided an important first baseline for describing low- P_T spectra and spin-averaged observables, but they inherently imposed a factorized form between the longitudinal and transverse degrees of freedom. Incorporating quantum chromodynamic (QCD) evolution consistently within this scheme necessitated reformulating the description in b_T space, where the Collins-Soper-Sterman (CSS) evolution could be implemented systematically [53].

Accumulating experimental evidence has since demonstrated that such a fully factorized picture is inadequate: SIDIS multiplicities reveal a P_T dependence that varies with both x (at fixed z) and z (at fixed x), indicating that $\langle k_T^2 \rangle$ carries a nontrivial x dependence, while $\langle P_T^2 \rangle$ in fragmentation depends explicitly on z [54,55].

Modern analyses have therefore transitioned to b_T -space frameworks with flexible nonperturbative sectors, often incorporating flavor and kinematic dependence in the transverse widths. These fits typically achieve NLL to N^3 LL precision and describe both SIDIS and Drell-Yan data—including fixed-target measurements from E288 and E605 [16,22,26,56–58]—with excellent accuracy.

Following our initial introduction of deep neural-network techniques for extracting TMDs (Sivers Function) [59], and during the ongoing development of our k_T -space analysis of the Unpolarized TMD, Bacchetta *et al.* [28] reported a simple neural network extraction of unpolarized TMDs in impact-parameter (b_T) space. In parallel, the hadron-structure-oriented formulation introduced in Ref. [27] unifies low- and high-energy Drell-Yan and Z^0 production within a consistent theoretical framework, offering new insights into the interplay between perturbative and nonperturbative transverse dynamics. Even more recent results [29] have advanced this frontier with a global extraction at N^4 LL accuracy, incorporating flavor-dependent TMDs and fragmentation functions together with the first application of transverse-momentum moments (TMMs). We note that the level of perturbative accuracy in [29] maybe lower than N^4 LL, due to the requirements of operator product expansion (OPE) matching to collinear fragmentation functions at N^3 LO, which are not yet available.

At the core of the present work, artificial intelligence (AI) provides a uniquely powerful framework for TMD extraction. Flexible neural architectures can capture non-parametric, multiscale correlations across $(x, Q, k_\perp)$ while embedding physics constraints as inductive biases and propagating experimental uncertainties through replica ensembles. Building upon this principle, our previous study performed the first-ever neural network-based extraction of a TMD—the Sivers function—by globally fitting SIDIS data [59]. That work introduced the concept of implicit inclusion and multiplicative compositionality (or functional multiplicativity), demonstrating that even nominally factorized representations can yield minimally biased TMD information when the neural component possesses sufficient expressive capacity. This established a data-driven, physics-informed paradigm for TMD phenomenology and directly motivates the present analysis, which extends the framework by embedding the neural network within the $k_\perp$ convolution integral itself, enabling fully end-to-end differentiable training and direct optimization against experimental observables.

In this work, we pursue a complementary approach to traditional b_T -based extractions by performing the entire analysis directly in momentum space and exploring how modern computational methods and AI can be optimally leveraged for nonperturbative QCD inference. The prevailing preference for the b_T representation in global analyses arises primarily from the mathematical convenience of the inverse problem: the Fourier-Bessel transform converts transverse-momentum convolutions into products, enabling a compact implementation of CSS evolution and resummation [10,41,60]. Moreover, lattice-QCD calculations of the CSS evolution kernel are naturally formulated in impact-parameter space [61], providing valuable first-principles input on the nonperturbative b_T dependence. Nevertheless, analyzing the transverse structure directly in $k_\perp$ space offers distinct advantages for data-driven approaches: it preserves a one-to-one correspondence with the measured spectra, allows experimental uncertainties and kinematic effects to be incorporated without intermediate transformations, and facilitates the inclusion of differentiable, artificial intelligence (AI)-based forward models. Accordingly, both b_T and $k_\perp$ representations remain essential and mutually constraining perspectives on TMD dynamics.

Our strategy is to render the $k_\perp$ inverse problem tractable by numerical means rather than by changing representation. We embed a DNN directly as the integrand of the $k_\perp$ -space convolution and train it on measured cross sections in the small- q_T region where TMD factorization holds, thus eliminating the need to transform to b_T for fitting and then back to $k_\perp$ for interpretation. Concretely, we represent the unpolarized TMD as

$$f_{1,q/N}(x, k_\perp; Q, Q^2) = f_{1,q}(x; Q^2) s(x, k_\perp; Q),$$

where we introduce the *intrinsic-transverse-momentum profile* $s(x, k_\perp; Q)$: a single nonparametric profile learned from data. Unlike factorized ansätze, this joint representation allows the fit to capture correlations among x , Q , and $k_\perp$ supported by the data, while the per-point normalization condition $\int d^2k_\perp s(x, k_\perp; Q) = 1$ for all (x, Q) fixes the overall scaling. This normalization preserves the PDF–TMD matching relation [27],

$$f_{1,q}(x; Q^2) = \int d^2\mathbf{k} f_{1,q}(x, k_\perp; Q^2), \quad (1)$$

ensuring consistency with collinear PDFs.

The forward map from network parameters to observables remains fully differentiable through a quadrature layer that performs the two-dimensional angular convolution *within the training loop* [62]. We do not impose an explicit resummed Sudakov kernel or a fixed-order Y -term (cf. the CSS $W + Y$ decomposition [10,41]); instead, we set $\mu = \sqrt{\zeta} = Q$, take the Drell-Yan hard factor at tree level, and allow the joint profile $s(x, k_\perp; Q)$ to absorb the Q dependence directly from the data in the small- q_T region. Here, Q (or equivalently Q_M) denotes the dilepton invariant mass and serves as the hard scale in Drell-Yan, analogous to the photon virtuality Q^2 in SIDIS. Because no perturbative resummation of cusp or noncusp anomalous dimensions is applied, it is not meaningful to assign an $N^k\text{LL}$ accuracy label to this extraction. Our results should therefore be characterized as a data-driven momentum-space analysis with a tree-level hard factor and no explicit logarithmic resummation, complementary to b_T -space fits where $N^k\text{LL}$ accuracy is defined by the perturbative content of the Sudakov kernel and matching [8,10,16,18,41,63,64]. This distinction is crucial: whereas b_T -space fits infer the Q dependence from perturbative inputs, our momentum-space framework learns it directly from the data, reducing bias not only by avoiding predefined functional ansätze for the $k_\perp$ dependence but also by dispensing with external priors from resummation schemes and the intermediate b_T mapping. The joint profile $s(x, k_\perp; Q)$ is thus determined directly from experiment, with no structural constraints beyond PDF matching, providing a minimally biased, data-driven approach to TMD extraction.

The analysis follows standard fixed-target Drell-Yan practice. We reconstruct (x_1, x_2) from (Q_M, y) at leading order in q_T/Q_M as $x_{a,b} = (Q_M/\sqrt{s})e^{\pm y}$ [10], restrict the fit to the TMD domain $q_T \ll Q_M$ with the condition $q_T < 0.2Q_M$, where the $W + Y$ separation is valid [10,41], veto the Υ resonance region in E288/E605 [56,57], choose $\mu = \sqrt{\zeta} = Q$ as a canonical scale [10], and neglect the Y -term in fits limited to small q_T [41]. Uncertainties are propagated through Monte-Carlo replica ensembles constructed from the reported pointwise errors of E288/E605 (full covariances are not available) [56,57], and

collinear-PDF uncertainties are incorporated by sampling NNPDF4.0 replicas accessed through LHAPDF [65,66]. Within this controlled kinematic window, the method provides a direct, minimally biased extraction in $k_\perp$ space, complementary to global b_T -space analyses that emphasize resummed evolution and matching [10,16,18]. We note that the differentiable-programming framework employed here can, with minimal modification, be adapted to a b_T representation [62].

Our results demonstrate the feasibility of solving the inverse problem entirely in momentum space, without factorizing longitudinal and transverse dynamics or mapping between $k_\perp$ and b_T . In this sense, the present work establishes a data-driven alternative path to unpolarized TMDs that remains directly connected to the measured observables. Methodologically, we employ an end-to-end differentiable inverse-convolution integration in $k_\perp$ that learns correlated x – Q – $k_\perp$ structure directly from the cross sections, aligning with modern learned-operator approaches to inverse problems [67,68].

The remainder of the paper is organized as follows. Section II outlines the DY formalism in $k_\perp$ space and defines the joint, normalized representation $s(x, k_\perp; Q)$ used throughout. Section III details the momentum-space extraction pipeline and describes the two-stage methodology: Stage I, the cross-section-level learning of the transverse structure-kernel $S(q_T; x_1, x_2, Q_M)$, and Stage II, the physics-informed reconstruction of the intrinsic-transverse-momentum profile $s(x, k_\perp; Q)$ with its recursive uncertainty propagation scheme. Section IV presents closure tests with pseudodata, validating the end-to-end inversion and quantifying methodological and algorithmic uncertainties. Section V applies the full procedure to the E288 and E605 fixed-target data. Sections VI and VII describe, respectively, the DNN implementation of the $S(q_T; x_1, x_2, Q_M)$ fit and the differentiable $k_\perp$ inverse convolution that reconstructs $s(x, k_\perp; Q)$. Section VIII presents the extracted TMDs, their scale evolution, and the uncertainty decomposition, while Sec. IX summarizes the conclusions and outlook.

II. FORMALISM

We work in the TMD factorization regime of the DY process, i.e., at small transverse momentum $q_T \equiv |\mathbf{q}_T| \ll Q_M$ (with Q_M being the dilepton invariant mass), where the cross section factorizes into hard, soft, and (TMD) collinear functions [1,2,4,10,69]. The triple-differential cross section can be written as

$$\frac{d\sigma}{dq_T dQ_M dy} = \frac{16\pi^2\alpha^2}{9Q_M^3} q_T F_{UU}^1(x_1, x_2, |q_T|, Q), \quad (2)$$

where the unpolarized structure function F_{UU}^1 is

$$F_{UU}^1 = x_1 x_2 \sum_q e_q^2 \mathcal{H}^{\text{DY}}(Q, \mu) \left(\int d^2 \mathbf{k}_{1T} d^2 \mathbf{k}_{2T} f_{q/1}(x_1, \mathbf{k}_{1T}; \mu, \zeta_1) f_{\bar{q}/2}(x_2, \mathbf{k}_{2T}; \mu, \zeta_2) \right. \\ \left. \times \delta^{(2)}(\mathbf{q}_T - \mathbf{k}_{1T} - \mathbf{k}_{2T}) \right) + (q \leftrightarrow \bar{q}). \quad (3)$$

Here, y is the dilepton rapidity, and (x_1, x_2) are the partonic momentum fractions satisfying, at leading order in q_T/Q_M ,

$$x_1 = \frac{Q_M}{\sqrt{s}} e^{+y}, \quad x_2 = \frac{Q_M}{\sqrt{s}} e^{-y}, \quad x_1 x_2 = \frac{Q_M^2}{s}. \quad (4)$$

The scales $(\mu, \zeta_{1,2})$ denote the renormalization and Collins-Soper (CS) scales associated with TMD evolution [10,18]. The hard factor $\mathcal{H}^{\text{DY}}(Q, \mu) = 1 + \mathcal{O}(\alpha_s)$ contains the perturbative matching coefficients for the partonic subprocess.

For datasets reported in terms of the Feynman variable x_F (notably E605), we convert to rapidity using the leading-order kinematic relations valid in the small- q_T/Q limit,

$$x_F = x_1 - x_2 = \frac{2Q}{\sqrt{s}} \sinh y, \quad y = \sinh^{-1} \left(\frac{x_F \sqrt{s}}{2Q} \right), \quad (5)$$

together with $x_1 x_2 = Q^2/s$. In practice, we use Eq. (5) to populate (y, x_1, x_2) from the published (Q, x_F) bin centers before evaluating the forward map. Thus, for the E605 cross section we used,

$$E \frac{d\sigma^{\text{DY}}}{d^3 q} = \frac{\bar{Q} \cosh \bar{y}}{\pi q_T \sqrt{s}} \frac{d\sigma^{\text{DY}}}{dq_T dy}, \quad (6)$$

where $\bar{Q}, \bar{y}(\bar{x}_F, \bar{Q})$ are bin-averaged values.

To keep the data-driven inference aligned with the measured q_T spectra and to minimize representation-induced modeling choices associated with transforming to b_T and back, we perform the extraction natively in momentum space and adopt a single, joint profile for the *unpolarized* TMD,

$$f_{q/h}(x, k_\perp; Q, Q^2) = f_{q/h}(x; Q^2) s_h(x, k_\perp; Q), \quad (7)$$

where $h \in \{a, b\}$ labels the incoming hadron, and $s_h(x, k_\perp; Q)$ is a nonparametric profile that captures correlations among $x, k_\perp$, and Q . Equation (7) is applied only in the TMD region; throughout this work, we restrict to $q_T < 0.2Q$ and do not include a Y -term or fixed-order matching by design, so the extracted profiles should be

interpreted as data-driven momentum-space TMDs within this domain. The per- (x, Q) -point normalization

$$\int d^2 k_\perp s_h(x, k_\perp; Q) = 1, \quad \forall (x, Q) \in \mathbb{R} \times \mathbb{R}_{>0}, \quad (8)$$

ensures consistency with the collinear PDFs and fixes the overall normalization.

In Eq. (7) we employ a single transverse-momentum profile $s_h(x, k_\perp; Q)$ that is shared among quark flavors within a given hadron h . This should be understood as a controlled modeling choice motivated by the limited flavor sensitivity of the fixed-target Drell-Yan data considered here (E288 and E605), which predominantly constrain PDF-weighted quark-antiquark combinations rather than independent flavor-resolved intrinsic transverse-momentum distributions. Even with a common profile s_h , the assembled TMDs $f_{q/h}(x, k_\perp; Q, Q^2) = f_{q/h}(x; Q^2) s_h(x, k_\perp; Q)$ retain nontrivial flavor dependence through the collinear PDFs.

Modern global extractions that incorporate broader datasets (e.g., collider Drell-Yan/Z measurements and, in many cases, SIDIS) can resolve additional flavor dependence in the transverse sector and therefore employ flavor-dependent $k_\perp$ profiles. Importantly, the present inverse-problem framework is modular and can be generalized straightforwardly by promoting $s_h \rightarrow s_{q/h}$ (equivalently $\lambda_h \rightarrow \lambda_{q/h}$) and conditioning the integrand network on a discrete flavor label. We defer such extensions to future global analyses where the corresponding degrees of freedom are better constrained by data.

With Eq. (7), the $k_\perp$ -space convolution in Eq. (3) reduces to a two-point angular autoconvolution of the profiles s :

$$S^{(ab)}(q_T; x_1, x_2, Q) \equiv \int_0^\infty dk_\perp k_\perp \int_0^{2\pi} d\phi s_a(x_1, k_\perp, Q) \\ \times s_b \left(x_2, \sqrt{q_T^2 + k_\perp^2 - 2q_T k_\perp \cos \phi}, Q \right). \quad (9)$$

and the structure function becomes

$$F_{UU}^1 = \mathcal{H}^{\text{DY}}(Q, \mu) \sum_q e_q^2 [x_1 f_{q/a}(x_1; Q^2) x_2 f_{\bar{q}/b}(x_2; Q^2) S^{(ab)}(q_T; x_1, x_2, Q) \\ + x_1 f_{\bar{q}/a}(x_1; Q^2) x_2 f_{q/b}(x_2; Q^2) S^{(ab)}(q_T; x_1, x_2, Q)]. \quad (10)$$

For pp scattering with flavor-symmetric transverse profiles, one may have $S^{(ba)} = S^{(ab)}$; for proton–nucleus kinematics, distinct beam/target profiles s_a and s_b can be retained without further assumptions, which we will address in future work.

For comparison with legacy presentations of fixed-target data, we use

$$E \frac{d\sigma^{\text{DY}}}{d^3q} = \frac{1}{2\pi q_T} \frac{d\sigma^{\text{DY}}}{dq_T dy}. \quad (11)$$

We set $\mu = \sqrt{\xi} = Q$ and restrict it to the small- q_T region where TMD factorization applies, neglecting the Y -term that restores the fixed-order tail at $q_T \sim Q_M$ [10,69]. Collinear inputs $f_{q/h}(x; Q^2)$ are taken from modern global fits (e.g. NNPDF4.0) via LHAPDF [65,66].

In Eq. (9) the profiles $s_h(x, k_\perp; Q)$ are represented by a physics-informed deep neural network (DNN) and embedded directly as the *integrand* of the transverse-momentum convolution. A differentiable quadrature layer carries out the radial and angular integrations, so that the forward map from network parameters to the observable cross sections is smooth and trainable by gradient methods. This design allows us to learn nontrivial x - $k_\perp$ - Q correlations without imposing factorized functional forms. When training on experimental data (E288/E605 [56,57]), the measured cross sections are provided in finite (Q_M, y, q_T) bins; we evaluate the theory at the bin centers, include the Jacobian in Eq. (11), and have verified that bin-width effects are negligible at the present level of precision. Although the b_T representation remains mathematically convenient—the analogue of Eq. (9) becomes a product after a Bessel-Fourier transform—our analysis stays entirely in $k_\perp$ space and does not rely on mapping to b_T for fitting. This keeps the extraction close to the measured q_T spectra while retaining the ability to cross-check against b_T -space results when needed.

III. METHODOLOGY

Our extraction proceeds in momentum space and is organized in two stages. In Stage I, we learn the transverse structure kernel $S(q_T; x_1, x_2, Q_M)$ directly from measured DY cross sections after factoring out known kinematic prefactors and collinear PDFs. In Stage II, we solve the inverse problem in $k_\perp$ by training a physics-informed

integrand $s(x, k_\perp, Q)$ whose autoconvolution reproduces the learned S with per- (x, Q) normalization and regularity constraints. The unpolarized TMD is finally assembled as

$$f_{1,q/N}(x, k_\perp; Q, Q^2) = f_{1,q}(x; Q^2) s(x, k_\perp; Q), \quad (12)$$

with $\int d^2k_\perp s(x, k_\perp; Q) = 1$ enforced for every (x, Q) in the fit support, where the flavor-dependence is carried by the collinear PDFs.

We use fixed-target DY cross sections from E288 (200, 300, 400 GeV) [56] and E605 (800 GeV) [57]. We test our output with the standard small- q_T cuts for TMD factorization, reconstruct (x_1, x_2) from (Q_M, y) at leading order in q_T/Q_M , exclude the Υ region by vetoing $9 \text{ GeV} < Q_M < 11 \text{ GeV}$, and treat measured cross sections in finite bins. For each Q_M bin, the width ΔQ_M is taken from the data, and we approximate the bin integral analytically via

$$\int_{Q_M - \frac{\Delta Q_M}{2}}^{Q_M + \frac{\Delta Q_M}{2}} \frac{dQ'}{Q'^3} = \frac{1}{2} \left(\frac{1}{(Q_M - \frac{\Delta Q_M}{2})^2} - \frac{1}{(Q_M + \frac{\Delta Q_M}{2})^2} \right), \quad (13)$$

which enters the rescaling from the triple-differential form in Eq. (2) to our learning target (see Sec. III A). Sample-dependent weights are applied to balance under-represented kinematic slices; these are fixed once at the data-preparation stage and used consistently throughout training. Equation (13) treats the explicit tree-level prefactor $\propto Q^{-3}$ exactly over the finite mass bin, while evaluating the remaining (smooth) factors at the bin center Q_M ; this is a midpoint approximation whose leading correction scales as $\mathcal{O}(\Delta^2)$ for symmetric bins. It is also possible to evaluate the full bin integral numerically by adding a low-order quadrature over Q' inside the differentiable forward map.

A. Stage I: Cross-section-level learning of $S(q_T; x_1, x_2, Q_M)$

We build a supervised model that maps $(q_T, x_1, x_2, Q_M) \mapsto S(q_T; x_1, x_2, Q_M)$ and is trained against the measured cross section after removing known prefactors and isolating the PDF combinations. Concretely, we write Eq. (10) as

$$\begin{aligned} \frac{d\sigma}{dq_T dQ_M dy} = & \frac{(4\pi\alpha)^2}{9 \cdot 2\pi} \left[\int_{Q\text{-bin}} \frac{dQ'}{Q'^3} \right] \sum_q e_q^2 [x_1 f_{q/a}(x_1; Q_M^2) x_2 f_{\bar{q}/b}(x_2; Q_M^2) S(q_T; x_1, x_2, Q_M) \\ & + x_1 f_{\bar{q}/a}(x_1; Q_M^2) x_2 f_{q/b}(x_2; Q_M^2) S(q_T; x_1, x_2, Q_M)], \end{aligned} \quad (14)$$

where the prefactor $\frac{(4\pi\alpha)^2}{18\pi}$ converts to the experimental units used by E288/E605. The per-point training target is then

$$S_{\text{calc}}(q_T; x_1, x_2, Q_M) = \frac{\left(\frac{d\sigma}{dq_T dQ_M dy}\right)}{\frac{(4\pi\alpha)^2}{18\pi}} \left[\int_{Q\text{-bin}} \frac{dQ'}{Q'^3} \right]^{-1}, \quad (15)$$

while the inputs carry two auxiliary channels with pre-computed PDF weights for the $q\bar{q}$ and $\bar{q}q$ orderings. The network outputs the PDF-weighted sum of S contributions and is trained with a weighted mean squared loss using the experimental per-point weights. Progressive training with an early-stopping rule is employed to stabilize optimization. After training a model ensemble over cross-section replicas and PDF replicas, we project the network(s) onto a fine grid in (x_1, x_2, q_T, Q_M) and explicitly evaluate both orderings by swapping (x_1, x_2) in the input. We impose pp beams-targets exchange symmetry such that $Sq\bar{q} = S\bar{q}q$, intentionally neglecting nuclear effects from the target. For proton-proton kinematics the kernel may be symmetric under the exchange $(x_1, x_2) \leftrightarrow (x_2, x_1)$, i.e., $S(q_T; x_1, x_2, Q) = S(q_T; x_2, x_1, Q)$. Accordingly, in Stage I, we impose this symmetry and define the symmetrized predictor,

$$S_{\text{sym}}(q_T; x_1, x_2, Q) = \frac{1}{2} [S_\theta(q_T; x_1, x_2, Q) + S_\theta(q_T; x_2, x_1, Q)], \quad (16)$$

which is used both in the training loss and when projecting to the (x_1, x_2, q_T, Q) grid.¹ For each replica r , we compute $S_{\text{sym}}^{(r)}$ on the grid; the Stage I label is the replica mean and band, $\bar{S}_{\text{sym}} = \langle S_{\text{sym}}^{(r)} \rangle_r$ with 1σ dispersion from the ensemble. We use $\bar{S}$ (for $\bar{S}_{\text{sym}}$) in the rest of the article for convenience. These $\bar{S}$ labels are then used in Stage II to reconstruct $s(x, k_\perp, Q)$. For reference, the Stage I network starts with four input nodes for the kinematics (q_T, x_1, x_2, Q_M) and intermediate layers with widths (128, 128, 128, 64), totaling $N_{\text{par}} = 41,985$ trainable parameters including the output node, with the activation choices: “swish” (for intermediate layers) and “softplus” (for output layers). For proton-nucleus kinematics this exchange symmetry would *not* be imposed as in Eq. (16), and the two orderings would be kept distinct. This type of analysis will be addressed in future work.

B. Stage II: Physics-informed reconstruction of $s(x, k_\perp, Q)$ in $k_\perp$ space

The Stage II network represents the normalized profile $s(x, k, Q)$ and is trained directly against the Stage I labels $\bar{S}$ through the momentum-space autoconvolution

¹Equivalently, one can enforce symmetry by (i) sorting (x_1, x_2) at the input, (ii) using symmetric invariants, e.g., $u = x_1 + x_2$, $v = x_1 x_2$, or (iii) adding a small penalty $\lambda_{\text{sym}} |S_\theta(q_T; x_1, x_2, Q) - S_\theta(q_T; x_2, x_1, Q)|^2$ to the loss.

$$S(q_T; x_1, x_2, Q) = \int_0^{2\pi} d\phi \int_0^\infty dk k s(x_1, k; Q) \times s(x_2, \sqrt{q_T^2 + k^2 - 2q_T k \cos \phi}; Q). \quad (17)$$

To avoid *ad hoc* functional forms, we parametrize the $s(x, k, Q)$ logarithmic slope using a compact feedforward DNN with $\lambda(x, k, Q) \geq 0$ and set

$$s(x, k, Q) = \exp \left[- \int_0^k \lambda(x, \kappa, Q) d\kappa \right], \quad s(x, 0, Q) = 1, \quad (18)$$

so monotonicity and integrability follow by construction [see Sec. VII B for details on the DNN architecture for $\lambda(x, k, Q_M)$]. The per- (x, Q) normalization $\int d^2\mathbf{k} s(x, k_\perp; Q) = 1$ is imposed softly on a high-order Gauss-Legendre grid with a cosine taper that switches on at $k \simeq 0.85k_{\text{max}}$ and a tail mask beyond $0.9k_{\text{max}}$. The forward operator evaluates Eq. (17) with $N_\phi = 64$ angular points and an adaptive radial grid of $R_{\text{nodes}} = 4096$ nodes up to $k_{\text{max}} = 5$, using a trapezoidal rule in k and precomputed targets for physics anchors (see below). The full map from network parameters to S is differentiable, enabling end-to-end gradient training.

The map from the normalized profile $s(x, k; Q)$ to the projected observable $\bar{S}(q_T; x_1, x_2, Q)$ is an integral (auto-convolution) operator and is therefore smoothing. As a result, the inverse reconstruction is not uniformly well conditioned in the radial variable k : variations of s in some k regions can induce only weak changes in the resolved q_T spectra. To ensure numerical stability of the inversion, we retain the mass constraint by definition and, during model development, we introduce auxiliary soft regularizers to suppress unphysical oscillations in weakly constrained directions. These auxiliary terms are used only for prototyping and are removed for the final production extraction once the architecture and hyperparameters have been fixed by recursive closure tests. Further details are provided in Sec. VII.

C. Uncertainties and the recursive extraction scheme

Our uncertainty quantification adheres to a robust recursive framework [59], which is designed to stabilize extraction during error quantification through the process of propagation of experimental errors. This approach systematically distinguishes between precision and accuracy, while also quantifying contributions from algorithmic and methodological sources. The approach iterates between pseudodata closure and real data fitting until the pseudodata reproduce the real cross sections within experimental errors. In each outer iteration $r = 0, 1, 2, \dots$ and we

- (1) Generate pseudodata from a current generator $G^{(r)}$ over the full (x_1, x_2, q_T, Q_M) support and train the Stage I model ensemble to closure, then project and average to produce $\tilde{S}^{(r)}$.
- (2) Reconstruct $s^{(r)}(x, k, Q)$ by Stage II training against $\tilde{S}^{(r)}$, propagate to observables, and compare with the experimental cross sections.
- (3) Update the pseudodata generator and its nuisance settings so that the next round better matches the real data within the quoted uncertainties.

The iteration stops when pseudodata and real data predictions agree (binwise) within the experimental errors. The *accuracy* of the method is then assessed by the kinematically resolved residual between the known generator truth and the final extraction on the matched pseudodata,

$$\Delta_{\text{acc}}(x, k, Q) = s_{\text{extr}}^{(r_\star)}(x, k, Q) - s_{\text{true}}^{(G^{(r_\star)})}(x, k, Q), \quad (19)$$

with $r_\star$ being the last iteration. The *precision* is quantified by the spread of the replica ensemble at the level of S (Stage I) and propagated to s (Stage II), while the *methodological* component collects controlled variations described next. This recursive strategy reduces variance, mitigates initialization sensitivity, and provides a transparent, kinematics-dependent separation of precision and accuracy; it follows the schema outlined in our earlier work [60] and is adapted here to the unpolarized case.

Because full experimental covariance matrices are not available for the E288/E605 sets, we propagate experimental uncertainties by Monte Carlo replicas of the cross sections in Stage I and by PDF replicas in the collinear inputs. The Stage I ensemble induces an ensemble of S grids whose mean and 1σ dispersion define the labels and their “precision” bands for Stage II. On top of this statistical component, the following are also strictly quantified and minimized:

- (i) Algorithmic errors from network initialization, optimizer settings, and training schedules (tracked by repeating trainings with different seeds and schedules, quantified without experimental errors included).
- (ii) Integration-scheme errors from changing the Stage II evaluator (direct angular-radial quadrature vs autoconvolution) and from tightening/loosening N_ϕ , R_{nodes} , and k_{max} . There is a separate component of methodological error found from the residual of $s^{(r)}(x, k, Q)$ and a test function $s_{\text{true}}^{(r_\star)}(x, k, Q)$ (the mean model of the final iteration).
- (iii) Experimental input error from the cross sections (already included via the Stage I replica ensemble).
- (iv) Collinear-PDF error from PDF replicas propagated to final observables.

We analyze these components separately and as a composition through the stages. When quoting a single

band, we combine them, but we also display kinematics-dependent residuals $\Delta_{\text{acc}}(x, k, Q)$ to make the precision component explicit.

With $s(x, k; Q)$ in hand, we form observables by inserting Eq. (17) into Eq. (10) and reconstruct the cross sections in the same binning and with the same Jacobian $E d\sigma/d^3q = (2\pi q_T)^{-1} d\sigma/(dq_T dy)$ used for the experimental presentations. The per- (x, Q) normalization ensures matching onto the collinear PDFs, and the small- q_T working region avoids any explicit modeling of the Y -term tail.

IV. CLOSURE TESTS

We validate the momentum-space extraction with three closure tests that mirror the steps of the pipeline: a cross-section-level closure using pseudodata generated with fixed-target kinematics; a closure on the convolution kernel $S(q_T)$; and a closure on the intrinsic profile $s(k)$. The latter two follow the inverse-integration workflow of Sec. III. Once a DNN is trained, both $S(q_T)$ and $s(k)$ can be sampled on arbitrarily fine meshes in q_T or k ; they are therefore not limited by sparse experimental binning. An example of one of the basic cross-section closure tests is displayed in Fig. 1 using experimental kinematics and errors, while the $S(q_T)$ and $s(k)$ closures appear in Fig. 2. We emphasize that the latter (Stage-II closure test) is a level-0 (noiseless) inverse-operator validation designed to isolate inversion bias; statistical fluctuations are instead propagated through the replica ensemble, which provides the higher-level uncertainty on the reconstructed $s(x, k_\perp, Q)$ which is studied in the end-to-end closure tests.

A. Cross sections

For the initial controlled test we adopt a simple generator with no explicit x - or Q -dependence for simplicity,

$$s_{\text{true}}(k) = \frac{1}{\pi} e^{-k^2}, \quad (20)$$

which implies the analytic autoconvolution

$$S_{\text{true}}(q_T) = \frac{1}{2\pi} \exp\left(-\frac{q_T^2}{2}\right). \quad (21)$$

Pseudodata are produced at E288-like kinematics (with the Υ veto) using Eq. (10) together with modern collinear inputs (e.g., NNPDF4.0). We then carry out the end-to-end fit, training directly on the triple-differential cross sections and reconstructing the theory at bin centers with the invariant Jacobian of Eq. (11). Agreement between prediction and pseudodata across (q_T, Q_M, y) slices demonstrates a consistent forward map; representative comparisons are shown in Fig. 1.

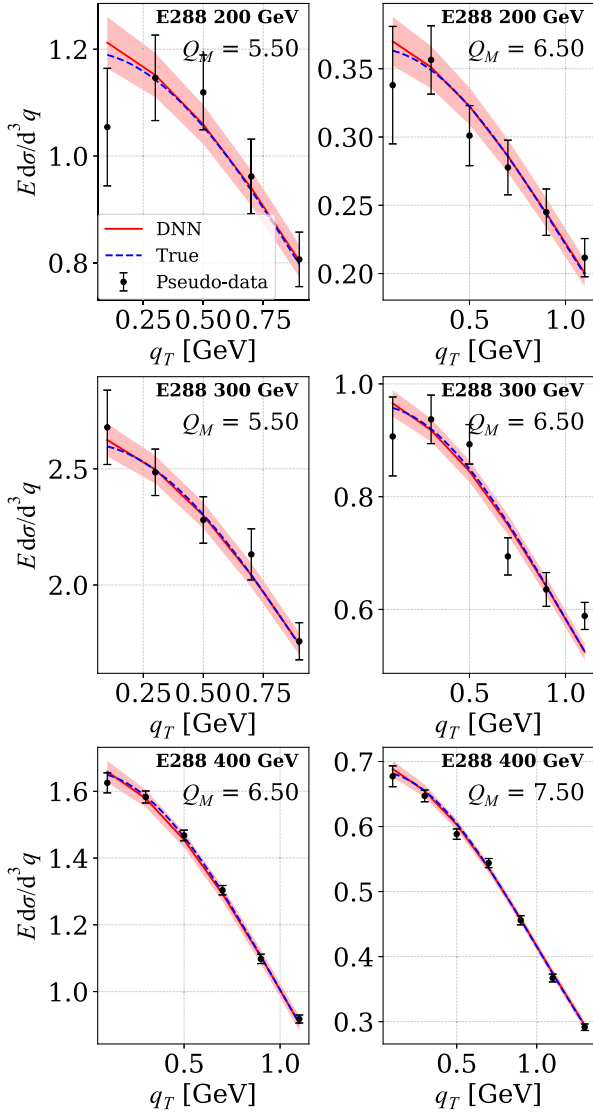


FIG. 1. Cross section closure with pseudodata at E288-like kinematics (in black) with the same scale of error from the experiment. The dashed line (in blue) show the pseudodata (true cross sections), and the solid line and the error band (in red) represent the “mean” and 1σ DNN prediction (with 1000 replicas), respectively, and are constructed at bin centers using Eq. (11).

B. $S(q_T)$ closure (inverse integration)

Given the trained model, we extract $S(q_T; x_1, x_2, Q)$ by factoring out known kinematic prefactors and PDF combinations, and we compare it against the truth computed from the generator via the same momentum-space operator used in the fit,

$$S(q_T; x_1, x_2, Q) = \int_0^{2\pi} d\phi \int_0^{k_{\max}} dk k s(x_1, k; Q) \times s\left(x_2, \sqrt{q_T^2 + k^2 - 2q_T k \cos\phi}; Q\right). \quad (22)$$

The inference in this case is essentially continuous in $S(q_T; x_1, x_2, Q)$, limited only by computational resources. We sample $S(q_T)$ on a dense q_T grid (far denser than the experimental binning since we are in the general problem mapping a DNN model directly). The left panel of Fig. 2 shows that the learned kernel reproduces the generator line shape within the quoted bands over the resolved range.

C. $s(k)$ closure (inverse problem)

We invert for the integrand by training the physics-informed $s(x, k, Q)$ against the $S(q_T)$ labels from the previous step, enforcing the per- (x, Q) normalization $\int d^2k s = 1$ and smoothness priors. For the generator above, the target is known analytically, so we test whether the trained $s(k)$ recovers the original line shape,

$$s_{\text{true}}(k) = \frac{1}{\pi} e^{-k^2} \quad \text{vs} \quad s_{\text{extr}}(k). \quad (23)$$

The right panel of Fig. 2 demonstrates near-perfect recovery across the resolved k window, with a mild systematic deviation at very small k . That low- k feature reflects the interplay of low- q_T leverage, the k -space taper controlling boundary effects, and the finite angular sampling; it is included in the methodological error budget below.

The residual $\Delta_{\text{acc}}(x, k, Q)$ in $s(k)$ is a crucial diagnostic, as it quantifies the error associated with the inverse mapping. The aim is to achieve a near-perfect recovery of the input $s(k)$; any deviation reflects an uncertainty that must be incorporated into the total error budget. This component of the methodological uncertainty is intrinsic to inverse problems of this kind. The width of the algorithmic error distribution among replicas in the reconstructed $s(k)$ should be negligible since the mapping is direct and free from additional stochastic sources of error. The residual term $\Delta_{\text{acc}}(x, k, Q)$ must likewise be minimized by design, with any remaining quantifiable contribution propagated as a kinematically dependent bias in the fidelity of the reconstructed $s(k)$.

D. Recursive precision/accuracy accounting

To fully quantify *precision* from *accuracy* and stabilize the inversion, we follow a recursive protocol: generate pseudodata from a current generator $G^{(r)}$, perform the cross section fit to obtain $S^{(r)}$, reconstruct $s^{(r)}$, compare with real data, update $G^{(r)}$, and repeat. Iterations stop when pseudodata match the real cross sections (binwise) within experimental errors across kinematics. The final, kinematics-dependent *accuracy* residual is given in Eq. (19) and is reported alongside the *precision* band obtained from the replica ensemble (shown as a 1σ algorithmic band on $S(q_T)$ and $s(k)$).

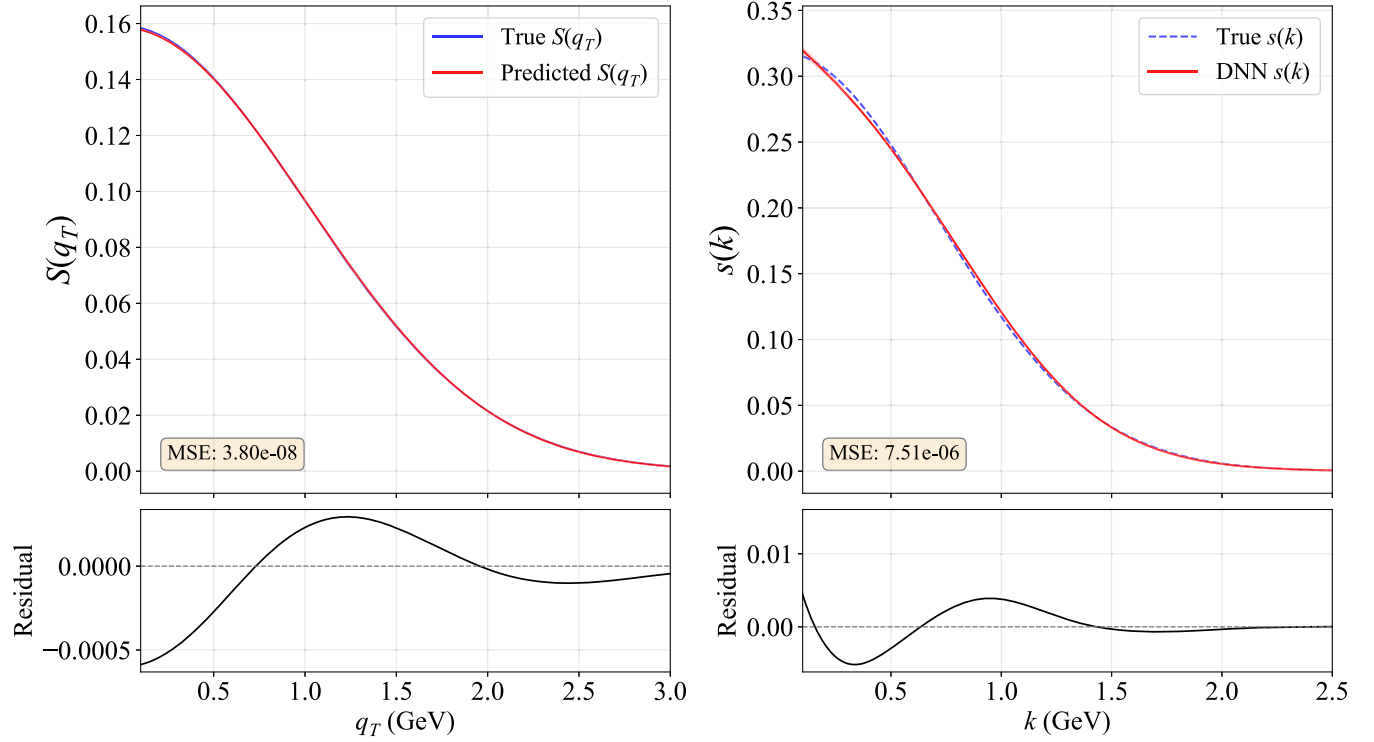


FIG. 2. Closure for $S(q_T)$ and $s(k)$ using the inverse-integration workflow (results with 1000 replicas). Left: $S(q_T)$ comparison (learned vs generator). Right: $s(k)$ comparison (extracted vs generator). Where shown, the shaded band indicates the 1σ algorithmic spread across independent trainings.

E. Methodological error

We monitor four contributions in addition to the statistical replica spread. The *algorithmic* component captures sensitivity to network initialization and optimizer schedule and appears as the 1σ band around the mean in $S(q_T)$ and $s(k)$. The *integration/inversion* component is obtained by switching evaluators (direct angular-radial quadrature versus autoconvolution integration) and tightening/loosening N_ϕ , radial nodes, and $k_{\max}$; the induced shift is quoted as a systematic. The *experimental* component comes from cross-section replicas propagated through the Stage I fit into the S labels and the *collinear PDF* component from PDF replicas used in the final observable reconstruction. To probe sensitivity to spectral content in k , we also vary the generator around the baseline (e.g., width and tail deformations) and take the difference between the true and deduced means as a *generator-modeling* contribution. In the Gaussian test, this analysis yields a small low- k deviation that remains subdominant to the quoted precision band.

F. Remarks on density and kinematic coverage

Because $S(q_T)$ and $s(k)$ are produced from trained DNNs rather than directly from sparse detector binning, they can be evaluated on arbitrarily fine grids. We therefore present closure comparisons on dense (q_T, k) meshes,

while ensuring that anchors (normalization and low- q_T curvature) are computed with the same quadrature used during training. This dense evaluation is crucial for diagnosing localized biases (e.g., the low- k behavior noted above) and for defining kinematics-aware accuracy residuals within the recursive protocol (see Figs. 1 and 2).

V. FITS WITH EXPERIMENTAL DATA

The extraction of unpolarized TMDs from fixed-target DY data proceeds through a momentum-space pipeline that connects the measured cross sections directly to the underlying transverse-momentum distributions using the same two stage and recursive methodology previously outlined and demonstrated with the pseudodata. We continue to work entirely in $k_\perp$ space, using a deep neural network (DNN) as a physics-informed *integrand* for the transverse convolution. The procedure imposes per- (x, Q) normalization and smoothness while minimizing functional bias by avoiding parametric ansätze.

A. Step 1—Cross-section replicas and learning

$$S(q_T; x_1, x_2, Q_M)$$

The starting point is the triple-differential cross-section data from the E288 and E605 experiments, binned in (q_T, Q_M, y) . We reconstruct (x_1, x_2) at leading order in q_T/Q_M and apply the standard cuts ($q_T < 0.2Q_M$ and the Υ

veto, $9 \text{ GeV} < Q_M < 11 \text{ GeV}$). For each dataset we generate ensembles of *experimental replicas* by sampling the reported statistical uncertainties; full covariance matrices are not available and are therefore not used. A supervised model is then trained on the replicas to predict the structure function F_{UU}^1 , from which we isolate the convolution kernel $S(q_T; x_1, x_2, Q_M)$ by dividing out known kinematic prefactors and the collinear PDF combinations (with PDF replicas used for uncertainty propagation). The loss is a weighted mean squared error that respects the experimental errors and kinematic coverage, and progressive training is employed for stability. The trained ensemble is projected onto a fine (x_1, x_2, q_T, Q_M) grid to produce averaged S labels with associated bands, which become the targets for the inversion in Step 2.

B. Step 2—Reconstruction of $s(x, k_\perp; Q)$ in $k_\perp$ space

To solve the inverse problem, we represent the intrinsic profile $s(x, k_\perp; Q)$ by a normalized DNN and embed it as the integrand of the transverse-momentum convolution, as described in Eq. (22). The forward operator is implemented via a differentiable angular-radial quadrature inside the training loop, so the map from network parameters to observables remains smooth and trainable with gradient methods. We enforce the per- (x, Q) normalization $\int d^2k s(x, k; Q) = 1$ and mild curvature regularization, and we include light physics anchors (e.g., low- q_T behavior and an integrated moment of S) to stabilize the inversion. The result is a nonparametric $s(x, k_\perp, Q)$ that reproduces the learned S labels across the full kinematic range of the experimental data.

C. Step 3—Recursion and error extraction

The same recursive refinement previously described in the methodology is applied to stabilize the inversion and to separate precision from accuracy. Each outer iteration consists of: (i) fitting the cross sections to obtain S on a dense grid, (ii) reconstructing $s(x, k, Q)$ against those S labels, (iii) generating pseudodata at higher resolution from the current s and refitting, and (iv) repeating until pseudodata agree with the real cross sections, binwise, within their reported errors. This loop reduces algorithmic variance, mitigates initialization sensitivity, and quantifies accuracy via kinematics-dependent residuals evaluated on matched pseudodata. All fits are performed with the Υ veto and with the conventional $q_T < 0.2Q_M$. No further variation is seen after two iterations within the experimental error.

D. Summary of TMD assembly

After convergence, the unpolarized TMDs are assembled as

$$f_{1,q/N}(x, k_\perp; Q, Q^2) = f_{1,q}(x; Q^2) s(x, k_\perp, Q), \quad (24)$$

with $f_{1,q}(x; Q^2)$ taken from modern global fits of the collinear PDF and $s(x, k_\perp; Q)$ obtained from the inversion above. The ensemble of trained models propagates experimental and PDF uncertainties to S and, subsequently, to s and to the final cross section predictions. Representative cross-section comparisons and TMD slices are shown in Figs. 3–5.

VI. THE DNN $S(q_T; x_1, x_2, Q_M)$ FIT

In Stage I, we learn the momentum-space kernel $S(q_T; x_1, x_2, Q_M)$ directly from the triple-differential DY cross sections by factoring out the known kinematic prefactors and the collinear PDF combinations, and training a compact DNN on the resulting target.² The learned S is then used in Stage II to reconstruct the intrinsic profile $s(x, k_\perp, Q)$ via the differentiable inverse integration described next.

A. Inputs, targets, and prefactors

For each experimental point, we form two auxiliary PDF channels that collect the charge-weighted $q\bar{q}$ and $\bar{q}q$ combinations at the bin scale Q_M ,

$$f_1(x_1, x_2) \equiv \sum_{q=u,d,s} e_q^2 x_1 x_2 f_{1,q}(x_1; Q_M^2) f_{\bar{q}}(x_2; Q_M^2), \quad (25)$$

$$f_1(x_2, x_1) \equiv \sum_{q=u,d,s} e_q^2 x_1 x_2 f_{1,q}(x_2; Q_M^2) f_{\bar{q}}(x_1; Q_M^2), \quad (26)$$

evaluated with LHAPDF (using NNPDF4.0). The narrow Q_M -bin integral entering the unit conversion and Jacobian is computed analytically,

$$\int_{Q_M-\Delta/2}^{Q_M+\Delta/2} \frac{dQ'}{Q'^3} = \frac{1}{2} \left[\left(Q_M - \frac{\Delta}{2} \right)^{-2} - \left(Q_M + \frac{\Delta}{2} \right)^{-2} \right]. \quad (27)$$

We denote the overall factor

$$K(Q_M) = \frac{(4\pi\alpha)^2}{18\pi} \left[\left(Q_M - \frac{\Delta}{2} \right)^{-2} - \left(Q_M + \frac{\Delta}{2} \right)^{-2} \right] (\hbar c)^2. \quad (28)$$

Given the measured cross section $(\frac{d\sigma}{dq_T dQ_M dy})$, the per-point *regression target* removes $K(Q_M)$ and leaves only S times the PDF combinations,

$$S_{\text{calc}}(q_T; x_1, x_2, Q_M) \equiv \frac{\left(\frac{d\sigma}{dq_T dQ_M dy} \right)}{K(Q_M) [f_1(x_1, x_2) + f_1(x_2, x_1)]} \approx S(q_T; x_1, x_2, Q_M). \quad (29)$$

²We follow the small- q_T setup and notation of Secs. II and III and the Stage I summary there.

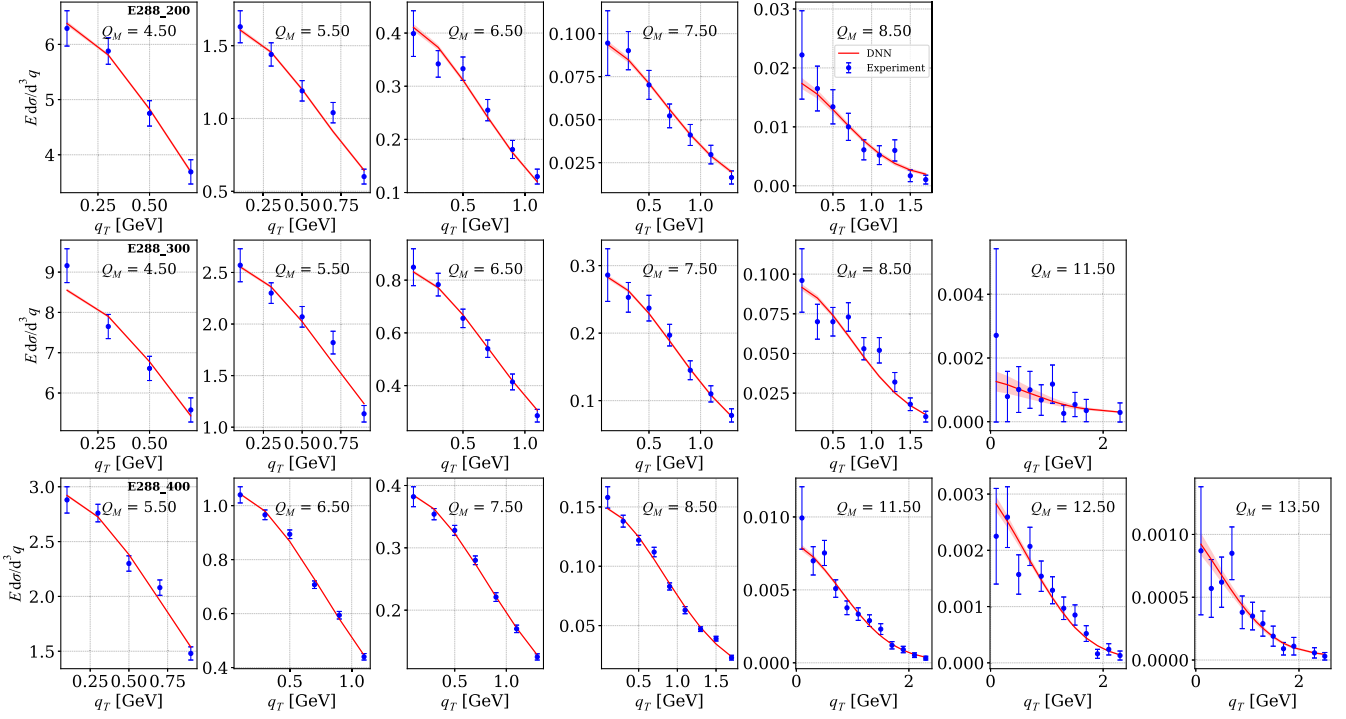


FIG. 3. Representative cross-section comparison after the joint $s(x, k_\perp, Q)$ extraction. Points with error bars are E288 data (with standard cuts), and the curve and band are the DNN prediction and its 1σ uncertainty obtained by propagating the replica ensemble through Eq. (24). As such, only the propagated experimental errors are shown.

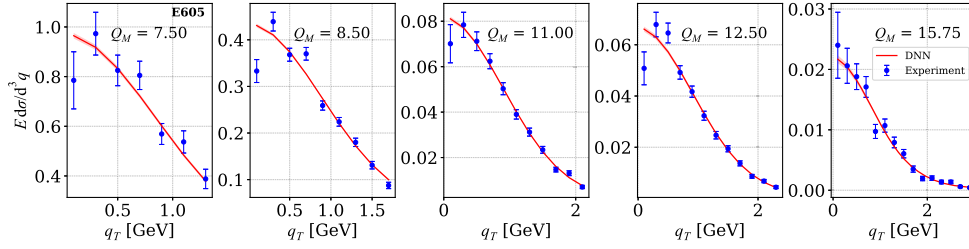


FIG. 4. Representative cross-section comparison after the joint $s(x, k_\perp, Q)$ extraction. Points with error bars are E605 data (with standard cuts); the curve and band are the DNN prediction and its 1σ uncertainty obtained by propagating the replica ensemble through Eq. (24). As such, only the propagated experimental errors are shown.

In practice, we train on Monte Carlo *replicas* of $(\frac{d\sigma}{dq_T dQ_M dy})^{\text{rep}}$ drawn from the reported uncertainties.³

B. Network and forward map

The model is a feed-forward network with inputs (q_T, x_1, x_2, Q_M) and a single output $\hat{S}_{DNN}$ with a `soft-plus` activation. We then multiply $\hat{S}_{DNN}$ by the two PDF channels and sum

$$\hat{S}(q_T; x_1, x_2, Q_M) = \hat{S}_{DNN}(q_T; x_1, x_2, Q_M) \times [f_1(x_1, x_2) + f_1(x_2, x_1)], \quad (30)$$

which establishes sum-aggregation as the canonical permutation-invariant operator [70]. This implements, inside the network graph, the same PDF-weighted structure-function decomposition used in the analytic presentation. For proton-proton (pp) kinematics, the kernel is symmetric under the exchange $(x_1, x_2) \leftrightarrow (x_2, x_1)$. In our implementation, this symmetry is enforced by evaluating both orderings through the explicit PDF channels in Eq. (30). While an explicit symmetrization penalty or input sorting can be applied, it is not required in the default fit. Furthermore, imposing beam-target exchange symmetry,

³Full covariance matrices are unavailable for E288/E605, so we propagate experimental errors via replicas; PDF uncertainties are handled by sampling PDF replicas and repeating the fit.

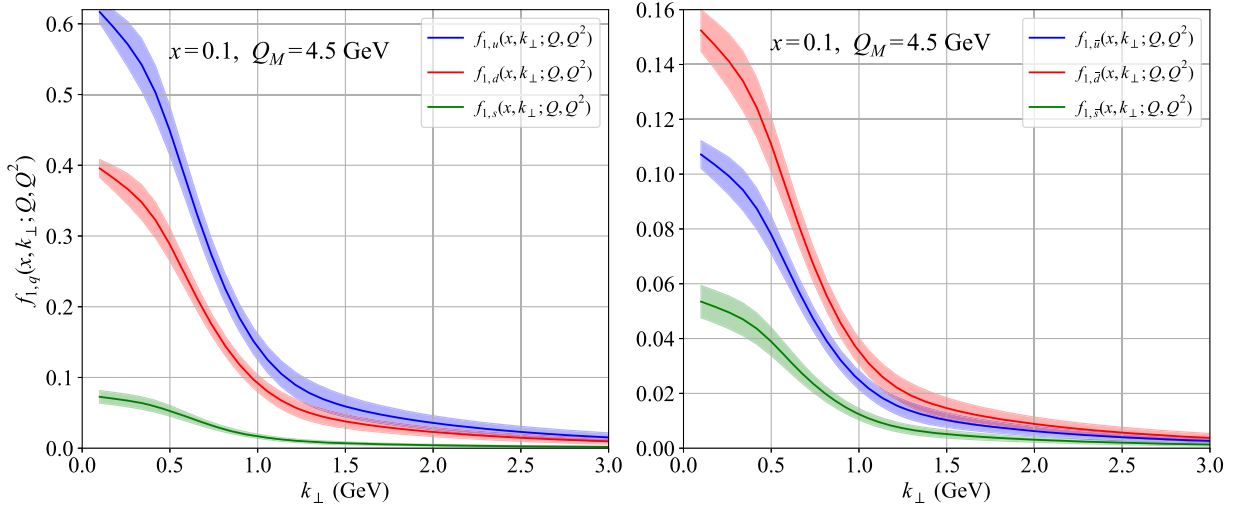


FIG. 5. Unpolarized TMDs reconstructed at $x = 0.1$ and $Q = 4.5$ GeV using Eq. (24). Curves show the central values for selected flavors. Shaded bands indicate the propagated 1σ uncertainties from experimental replicas (Step 1) and collinear PDF replicas. All error contributions are shown.

as done in our fit, introduces a potential bias when using nonhydrogen targets (as in E288 and E605), which is not quantified within our current estimate of systematic uncertainty and will be addressed in future studies.

The DNN architecture for S is already described in Sec. III A. Typical training settings are initial learning rate 2×10^{-3} , batch size 64, and up to 4000 epochs run in two progressive stages (see below).

C. Loss, weights, and progressive training

The objective is a weighted mean squared error (MSE) over training rows i ,

$$\mathcal{L}_S = \frac{1}{N} \sum_{i=1}^N w_i [\hat{S}_i - S_i^{\text{rep}}]^2, \quad (31)$$

with per-row weights w_i used to emphasize under-represented panels in $(\text{dataset}, Q_M)$. Training proceeds progressively in $N_{\text{prog}} = 2$ stages, freezing earlier layers in the first stage and unfreezing in the second. We employ an early-stopping rule keyed to the crossover of validation and training losses after a minimum burn-in (here 100 epochs) with patience and a small tolerance on $(\text{val} - \text{train})$; a ReduceLROnPlateau callback gently lowers the learning rate when the loss stalls.

D. From $\hat{S}$ to cross sections and to the S grid

After training, predictions on any input batch are converted back to cross sections via

$$\left(\frac{\widehat{d\sigma}}{dq_T dQ_M dy} \right) = \hat{S}(q_T; x_1, x_2, Q_M) K(Q_M), \quad (32)$$

using the same $K(Q_M)$ as in Eq. (28). For the hand-off to Stage II we project the trained model onto a fine grid in (x_1, x_2, q_T, Q_M) and *divide out the PDF channels* to recover

$$S(q_T; x_1, x_2, Q_M) = \frac{\hat{S}(q_T; x_1, x_2, Q_M)}{f_1(x_1, x_2) + f_1(x_2, x_1)}, \quad (33)$$

thereby producing the S labels and bands that drive the inverse $k_\perp$ fit in the next section. Uncertainties are propagated by repeating the Stage I training over experimental and PDF replicas and taking the ensemble mean and dispersion on the grid.

E. Position in the pipeline

This Stage I fit isolates the kernel $S(q_T; x_1, x_2, Q_M)$ from data while respecting the same kinematic prefactors and PDF structure used in the analytic formalism, staying natively in momentum space. The resulting S surfaces, with their replica bands, become the targets for Stage II's differentiable inverse convolution, where we reconstruct the normalized intrinsic profile $s(x, k_\perp; Q)$, which reproduces S across the full kinematic phase space learned by the DNN.

VII. DIFFERENTIABLE $k_\perp$ INVERSE CONVOLUTION

Before detailing the forward model used in the inversion, we note that reconstructing a transverse profile from its projected q_T convolution is an ill-conditioned inverse problem. During the initial prototyping stage, we therefore introduce a small set of auxiliary priors—soft pins, moment anchors, curvature and tail stabilizers—to condition the

optimization and to facilitate stable architecture and hyperparameter selection through recursive closure tests. These auxiliary terms are used only to guide model configuration and are subsequently removed, except for the mass normalization which is retained by definition to enforce PDF-TMD consistency along with the cross section components of the loss. The final reported solutions are therefore obtained without removable regularization terms; the temporary priors serve only to stabilize early training and do not determine the physical extraction.

A. Forward model in $k_\perp$ for inversion

In the small- q_T DY regime, the unpolarized structure function is written as Eq. (10) with the beams exchanged for pp kinematics as appropriate. Our inverse problem is, given triple-differential cross sections in (q_T, Q_M, y) (after the standard cuts), to reconstruct a *single* nonparametric profile $s(x, k; Q)$ such that its autoconvolution reproduces the learned S and, through F_{UU}^1 , the measured cross sections.

Two structural constraints are imposed throughout the fit: (i) per-point PDF matching, implemented via the normalization

$$\int d^2k s(x, k, Q) = 2\pi \int_0^\infty dk k s(x, k, Q) = 1, \quad \forall (x, Q). \quad (34)$$

and (ii) monotone falloff in k , enforced by the parametrization below.

B. Monotone integrand via a FiLM-conditioned λ -network

Rather than regressing s directly, we learn its *logarithmic slope* $\lambda(x, k, Q) \geq 0$ with a compact stack of *feature-wise linear modulation* (FiLM)-gated residual blocks [71] and define

$$s(x, k, Q) = \exp \left[- \int_0^k \lambda(x, \kappa, Q) d\kappa \right], \quad (35)$$

$$s(x, 0, Q) = 1. \quad (36)$$

With $\lambda \geq 0$, s is positive and monotonically decreasing in k by construction; the exponential-of-integral map is evaluated *inside* the computation graph and is standard in differentiable-programming settings [72,73]. This parametrization enforces qualitative shape constraints without hand-crafted ansätze [74–76].

The network input is (x, k, Q) . To ease the learning of the radial dependence we expand the scalar k into a fixed feature vector,

$$\psi(k) \equiv (k, k^2, \sqrt{k + \varepsilon}, \log(1 + k)),$$

with a small $\varepsilon > 0$ to avoid the square-root singularity at $k = 0$. The initial hidden state is $h^{(0)} = \phi(W^{(0)}\psi(k))$, where $\phi = \tanh$ and $W^{(0)} \in \mathbb{R}^{W \times 4}$.

Nonseparable (x, Q) dependence is injected through FiLM [77,78]. Let $c(x, Q)$ denote a small conditioning stream; for each block, $\ell = 1, \dots, L$ and it produces a pair of modulation vectors,

$$(\gamma_\ell(x, Q), \beta_\ell(x, Q)) = g_\ell(c(x, Q)) \in \mathbb{R}^W \times \mathbb{R}^W. \quad (37)$$

Given the incoming hidden state $h^{(\ell)} \in \mathbb{R}^W$, the FiLM-gated *residual* block (with width W) computes

$$u^{(\ell)} = \phi(W_1^{(\ell)} h^{(\ell)}) \in \mathbb{R}^W, \quad (38)$$

$$\tilde{u}^{(\ell)} = \gamma_\ell(x, Q) \odot u^{(\ell)} + \beta_\ell(x, Q) \in \mathbb{R}^W, \quad (39)$$

$$h^{(\ell+1)} = \phi(W_2^{(\ell)} \tilde{u}^{(\ell)} + h^{(\ell)}) \in \mathbb{R}^W. \quad (40)$$

Here, $W_1^{(\ell)}, W_2^{(\ell)} \in \mathbb{R}^{W \times W}$ are learnable weight matrices; $\odot$ denotes the Hadamard (elementwise) product; and the additive term $+h^{(\ell)}$ is the residual (skip) connection [71]. For complete clarity, Eqs. (38)–(40) can be written componentwise as

$$u_j^{(\ell)} = \tanh \left(\sum_{i=1}^W (W_1^{(\ell)})_{ji} h_i^{(\ell)} \right), \quad j = 1, \dots, W, \quad (41)$$

$$\tilde{u}_j^{(\ell)} = \gamma_{\ell,j}(x, Q) u_j^{(\ell)} + \beta_{\ell,j}(x, Q), \quad (42)$$

$$h_j^{(\ell+1)} = \tanh \left(\sum_{i=1}^W (W_2^{(\ell)})_{ji} \tilde{u}_i^{(\ell)} + h_j^{(\ell)} \right). \quad (43)$$

In our default implementation, we also consider a scalar-gate variant, where $\gamma_\ell, \beta_\ell \in \mathbb{R}$ are *scalars* per block (and per sample) broadcast across features; Eq. (39) then reduces to $\tilde{u}^{(\ell)} = \gamma_\ell u^{(\ell)} + \beta_\ell \mathbf{1}$ with $\mathbf{1} \in \mathbb{R}^W$ being the all-ones vector. We use $\gamma_\ell = \text{softplus}(\cdot)$ and $\beta_\ell = \tanh(\cdot)$ to ensure a positive gain and bounded shift.

After L blocks, the slope is produced by a one-neuron head with a `softplus` activation,

$$\lambda(x, k, Q) = \text{softplus}(w_{\text{out}}^\top h^{(L)} + b_{\text{out}}), \quad (44)$$

$$w_{\text{out}} \in \mathbb{R}^W, \quad b_{\text{out}} \in \mathbb{R}, \quad (45)$$

which guarantees $\lambda \geq 0$ and, through Eq. (35), implies $s(x, k; Q) > 0$ and monotone decrease in k . The per- (x, Q) mass constraint $\int d^2k s(x, k; Q) = 1$ is enforced in the physics layer that performs the $k_\perp$ quadrature (Sec. VII C).

The FiLM gate Eq. (39) implements a kinematics-dependent affine transformation of the radial features, enabling (x, Q) -dependent *deformations* of the k profile without factorizing variables or inflating parameters. The residual skip in Eq. (40) provides an identity gradient pathway, which stabilizes deep stacks of blocks. The map $\lambda \mapsto s$ is differentiable, so gradients from both the pointwise S fidelity term and the cross-section-level χ^2 propagate through the FiLM gates and the quadrature layer end-to-end.

The tuned FiLM-conditioned residual DNN architecture is the following: among the three inputs (x, k, Q_M) , the primary variable k undergoes feature expansion into four representations $(k, k^2, \sqrt{k}, \log(1+k))$, which are projected into a 64-dimensional hidden state via a tanh-activated dense layer. In parallel, a conditioning subnetwork processes (x, Q_M) through two ReLU-activated layers of width 32 to produce per-block FiLM parameters—a softplus-activated scale γ and a tanh-activated shift β —one pair per residual block. The trunk consists of 4 residual blocks, each comprising two dense layers of width 64 with tanh activations; between them, the FiLM affine transformation $\gamma \odot \mathbf{h} + \beta$ modulates the intermediate representation, allowing x and Q_M to dynamically gate the network’s response to k . Residual skip connections stabilize training across depth. The final layer is a single softplus unit with a calibrated bias initialization, guaranteeing a positive output $\lambda(x, k, Q_M) \geq 0$ throughout the domain. The maximum network comprises 35,081 trainable parameters.

C. Differentiable $k_\perp$ evaluator and normalization

We embed Eqs. (9) and (35) into a custom *differentiable quadrature layer* that is part of the computational graph [62]:

- (i) Adaptive radial grid: we place $R_{\text{nodes}} \simeq 4096$ nodes and $k_i = k_{\text{max}} \tanh(\beta \xi_i) / \tanh \beta$ with ξ_i uniform in $[0, 1]$ and $\beta \sim 3$, concentrating resolution at small k .
- (ii) Azimuthal integration: the φ integral uses N_φ uniform points ($N_\varphi = 64$ –128; see ablations), and for each (q_T, k_i, φ_j) we evaluate $k' = \sqrt{q_T^2 + k_i^2 - 2q_T k_i \cos \varphi_j}$ and linearly interpolate $s(x_2, k', Q)$ on the $\{k_i\}$ grid.
- (iii) Trapezoid in k and tapering: the k -integral uses a trapezoid rule on the adaptive grid. To suppress endpoint artifacts we multiply the *left* profile by a cosine taper $t(k)$ that turns on smoothly for $k \gtrsim 0.85k_{\text{max}}$ and track a “tail-mass” beyond $0.9k_{\text{max}}$.
- (iv) Normalization: for every (x, Q) we evaluate the mass $m_0[s] = 2\pi \int dk k s$ on a high-order Gauss-Legendre grid and *normalize* $s \mapsto s/m_0[s]$ (“left” normalization). A symmetric option applies the same to the right profile (“both”); in practice we use “left” to reduce redundancy while retaining stability.

Because Eq. (35) is an ordered cumulative integral, the discrete Jacobian is lower triangular: on the grid one has $\partial s(k_j)/\partial \lambda(k_i) = -s(k_j)\Delta k_i$ for $i \leq j$ and 0 otherwise. Backpropagation through the φ -integral and the k -trapezoid therefore yields stable, banded gradients all the way to the FiLM parameters.

Let $m_2[f] \equiv 2\pi \int dk k^3 f(k)$ denote the second radial moment: for a convolution of two *independent, normalized, isotropic* profiles one has $m_2[S] = m_2[s_1] + m_2[s_2]$; for identical beams, this gives $m_2[s] = \frac{1}{2}m_2[S]$. We therefore penalize $(m_2[s] - \frac{1}{2}m_2[S])^2$, with $m_2[S]$ estimated from the learned S via $2\pi \int dq q^3 S(q)$. Two low- q_T anchors stabilize the origin: $S(0)$ and a pin on $S''(0)$ obtained by local quadratic fits at the first q_T nodes. We also include a curvature prior in k and a tail-mass penalty beyond the taper.

D. Direct cross-section supervision and the inverse loss

In addition to the pointwise S loss, the model is supervised *directly* by the experimental triple-differential cross sections. We define an aggregator G_i that maps each training row i [with inputs $X_i = (x_{1,i}, x_{2,i}, q_{T,i}, Q_{M,i})$] to the corresponding experimental kinematic bin $j = (\text{dataset}, Q_M, q_T)$ and forms the cross section⁴

$$A_j^{\text{pred}} = \sum_{i \in B_j} G_i S_\theta(q_{T,i}; x_{1,i}, x_{2,i}, Q_{M,i}), \quad (46)$$

$$G_i = \left[\sum_{q=u,d,s} e_q^2 \left(f_q(x_{1,i}; Q_M^2) f_{\bar{q}}(x_{2,i}; Q_M^2) + f_{\bar{q}}(x_{1,i}; Q_M^2) f_q(x_{2,i}; Q_M^2) \right) \right] \mathcal{K}(Q_M), \quad (47)$$

where the factor $\mathcal{K}(Q_M)$ collects the tree-level hard prefactor and is defined in Eq. (28). With measurements $(A_j^{\text{exp}} \pm dA_j)$, we define the cross-section loss,

$$\mathcal{L}_{\text{CS}} = \frac{1}{J} \sum_{j=1}^J w_j \left(\frac{A_j^{\text{pred}} - A_j^{\text{exp}}}{dA_j} \right)^2, \quad (48)$$

where w_j are optional per-panel weights [e.g., to focus selected $(\text{dataset}, Q_M)$ panels]. The pointwise S loss on training rows $i \in \mathcal{T}$, $\mathcal{L}_S = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} \omega_i [S_\theta(X_i) - S^{\text{targ}}(X_i)]^2$, uses inverse-variance weights $\omega_i \propto 1/\sigma_i^2$. In the nominal optimization, we set $\omega_i = 1$.

The total objective combines the data-consistency terms with the physics-based regularizers described above,

⁴For brevity, we denote the measured (binned) cross section as A_j^{exp} and the prediction as A_j^{pred} .

$$\begin{aligned}
\mathcal{L}_{\text{total}} = & \mathcal{L}_S + \lambda_{\text{CS}} \mathcal{L}_{\text{CS}} + \lambda_{\text{mass}} [m_0[s] - 1]^2 \\
& + \lambda_{m_2} \left[m_2[s] - \frac{1}{2} m_2[S] \right]^2 + \lambda_0 [S(0) - S_0]^2 \\
& + \lambda_2 [S''(0) - S_2]^2 + \lambda_{\text{curv}} \|\partial_k^2 s\|_{w(k)}^2 \\
& + \lambda_{\text{tail}} m_0[s; k > k_{\text{tail}}].
\end{aligned} \tag{49}$$

with $\|\cdot\|_{w(k)}$ a curvature norm reweighted by a learned, non-negative profile $w(k)$ (see below), and $k_{\text{tail}} \simeq 0.9 k_{\text{max}}$. All terms are computed inside the same differentiable layer, so gradients from $\mathcal{L}_{\text{CS}}$ flow through the aggregator Eq. (47) and the $k_{\perp}$ evaluator to the FiLM parameters.

We employ a two-phase schedule matched to the implementation. Phase I (pretrain): minimize the pointwise S data-fidelity loss $\mathcal{L}_S$ (weighted MSE; optionally log-MSE) to stabilize the forward map. Phase II (fine-tune): switch to full-batch Adam on the combined objective $\mathcal{L}_{\text{total}}$ in Eq. (49) with a nonzero cross-section weight $\lambda_{\text{CS}} \in [0.1, 0.3]$, for $\mathcal{O}(10^2\text{--}10^3)$ steps. Unless stated otherwise, we use $\lambda_{\text{CS}} = 0.20$. All symbols λ in (49) are *hyperparameters* that control the relative strength of physics anchors and regularizers with respect to the data-fidelity terms. We treat the anchor weights as fixed across replicas and datasets to avoid hyper-overfitting; only model parameters are optimized. For clarity we list them below together with the quantities they weight:

- (i) λ_{CS} multiplies the cross-section-level χ^2 term $\mathcal{L}_{\text{CS}}$, which compares the *differentiably aggregated* prediction to the binned experimental cross sections.
- (ii) λ_{mass} penalizes deviations from per- (x, Q) unit mass: $m_0[s] \equiv 2\pi \int_0^\infty dk k s(x, k, Q)$ and the penalty is $[m_0[s] - 1]^2$.
- (iii) λ_{m_2} anchors the second moment of s to the small- q_T information in S : $m_2[s] \equiv 2\pi \int_0^\infty dk k^3 s(x, k, Q)$, matched to $\frac{1}{2} m_2[S]$ extracted from the learned $S(q_T)$ (the factor $\frac{1}{2}$ is the Jacobian relating the two moments).
- (iv) λ_0 and λ_2 softly pin the $q_T \rightarrow 0$ behavior of S via $[S(0) - S_0]^2$ and $[S''(0) - S_2]^2$, where (S_0, S_2) are low- q_T anchors inferred from the S fit. (These anchors may be set to zero in ablations.)
- (v) λ_{curv} is a Tikhonov-type smoothness weight on the radial profile $\|\partial_k^2 s\|_{w(k)}^2$, with

$$\|\partial_k^2 s\|_{w(k)}^2 \equiv \int_0^{k_{\text{max}}} dk w(k) [\partial_k^2 s(x, k, Q)]^2, \tag{50}$$

where $w(k) \geq 0$ is a user-specified (or learned) non-negative weight, and k_{max} is the maximal k covered by the evaluator.

- (vi) λ_{tail} penalizes excess probability in the far tail, $m_0[s; k > k_{\text{tail}}] \equiv 2\pi \int_{k_{\text{tail}}}^\infty dk k s(x, k, Q)$, to confine mass within the region constrained by data.

In our default configuration, we found stable performance with $\lambda_{m_2} \sim 5 \times 10^{-2}$, $\lambda_{\text{curv}} \sim 2 \times 10^{-4}$, $\lambda_{\text{tail}} \sim 10^{-3}$, and very small or vanishing (λ_0, λ_2) ; the unit-mass constraint is also enforced by the evaluator's normalization, so λ_{mass} can be kept small.⁵

To make the smoothness term focus where the data are most informative, we optionally adapt $w(k)$ from the gradient of the S fidelity term,

$$\begin{aligned}
w_{\text{new}}(k) \propto & \frac{|\partial \mathcal{L}_S / \partial s(x, k, Q)|^\alpha}{(|\partial_k^2 s(x, k, Q)|^\beta + \epsilon)}, \\
& \text{clipped to } [1, w_{\text{max}}],
\end{aligned} \tag{51}$$

with $(\alpha, \beta, w_{\text{max}}) = (0.9, 1.0, 6.0)$ by default and $\epsilon > 0$ for stability; w_{new} is renormalized to unit mean. In the cross section, a driven setup is used for the main results.

E. Discretization and numerical settings

Unless otherwise noted, we use $k_{\text{max}} = 5 \text{ GeV}$, $R_{\text{nodes}} = 4096$, $N_\varphi = 64\text{--}128$, a taper onset at $0.85 k_{\text{max}}$ with a tail mask beyond $0.9 k_{\text{max}}$, and 512-point Gauss-Legendre quadrature for the normalization and anchor moments. These settings were verified to be numerically converged for the resolved (q_T, k) windows and are consistent with the ablations reported elsewhere in the paper.

F. Target anchors

The moment anchors in the loss function connect the intrinsic transverse-momentum profile $s(x, k; Q)$ to the experimentally accessible structure kernel $S(q_T; x_1, x_2, Q)$ through their statistical moments. To make this connection explicit, let $\mathbf{K}_{1,2}$ denote the two independent, isotropic transverse-momentum vectors carried by the partons in the incoming hadrons, each distributed according to the normalized radial density $s_{1,2}(k)$ with $k = |\mathbf{K}|$. The observable transverse momentum of the lepton pair is their vector sum

$$\mathbf{Q} \equiv \mathbf{K}_1 + \mathbf{K}_2 \tag{52}$$

so that $S(q_T)$ is the radial probability density of $\mathbf{Q}$ with $q_T = |\mathbf{Q}|$.

Because $\mathbf{K}_1$ and $\mathbf{K}_2$ are statistically independent and isotropic, their mean-squared magnitudes add

$$\langle Q^2 \rangle = \langle K_1^2 \rangle + \langle K_2^2 \rangle. \tag{53}$$

In two-dimensional polar coordinates one may write the second radial moment of any normalized isotropic density $f(k)$ as

⁵Exact scalings can differ with grid resolution k_{max} and the choice of $w(k)$; we keep all λ 's fixed across replicas and datasets.

$$m_2[f] \equiv 2\pi \int_0^\infty dk k^3 f(k) = \langle K^2 \rangle_f. \quad (54)$$

Applying this to the convolution relation between $s_{1,2}$ and S gives

$$m_2[S] = m_2[s_1] + m_2[s_2]. \quad (55)$$

For identical beams ($s_1 = s_2 = s$) this reduces to

$$m_2[s] = \frac{1}{2} m_2[S], \quad (56)$$

which is exactly the relation used as the second-moment anchor in (49).

Role in the training schema. Equation (56) supplies a physically interpretable constraint that ties the learned intrinsic width of $s(x, k; Q)$ to the measured width of the observable $S(q_T; x_1, x_2, Q)$. In practice, $m_2[S]$ is computed from the Stage I kernel prediction via

$$m_2[S] = 2\pi \int_0^\infty dq_T q_T^3 S(q_T), \quad (57)$$

and the penalty term $[m_2[s] - \frac{1}{2} m_2[S]]^2$ in Eq. (49) stabilizes the inversion by preventing $s(x, k; Q)$ from drifting to unphysical widths that would otherwise yield degenerate $S(q_T)$ spectra. Together with the low- q_T anchors on $S(0)$ and $S''(0)$, this constraint fixes the overall scale and curvature of the reconstructed distribution, ensuring that the learned $s(x, k; Q)$ reproduces both the integrated and differential behavior of the measured cross sections while remaining consistent with the autoconvolution relation in Eq. (22).

G. End-to-end differentiability

The entire model is trained as a single differentiable computation that links the learnable parameters of the network to the final observable cross sections. Denoting by θ the collective set of *trainable parameters* of all neural components (weights and biases of the FiLM-conditioned λ -network, the curvature weight profile $w(k)$ if enabled, and the output head), the forward map can be written schematically as

$$\theta \mapsto \lambda_\theta(x, k, Q) \mapsto s_\theta(x, k, Q) \xrightarrow{\text{quadrature layer}} S_\theta(q_T; x_1, x_2, Q) \xrightarrow{\text{PDFs \& binning}} A_j^{\text{pred}}(\theta) \quad (58)$$

Each arrow corresponds to a differentiable operator inside the computational graph:

- (i) $\lambda_\theta(x, k, Q)$ is produced by the FiLM-gated residual network described in Sec. VII B; it depends explicitly on (x, k, Q) and implicitly on the learnable parameters θ .
- (ii) $s_\theta(x, k, Q)$ is obtained through the cumulative integral $s = \exp[-\int_0^k \lambda d\kappa]$, implemented differentially on the adaptive k -grid within the same graph. The Jacobian $\partial s(k_j)/\partial \lambda(k_i)$ is lower-triangular, ensuring stable gradient flow during backpropagation.
- (iii) The *quadrature layer* performs the radial and azimuthal integrations that map $(s_1, s_2) \rightarrow S(q_T)$ via the convolution; its discretized form retains exact differentiation with respect to θ .
- (iv) The final stage combines S_θ with charge-weighted parton densities $f_q(x_i; Q^2)$ and experimental bin integrals to yield $A_j^{\text{pred}}(\theta)$, the theoretical cross section in bin j as defined in Eq. (46).

Because every stage is differentiable, gradients of the total loss $\mathcal{L}_{\text{total}}(\theta)$ propagate seamlessly from the experimental χ^2 and pointwise S terms down to the fundamental model parameters,

$$\frac{\partial \mathcal{L}_{\text{total}}}{\partial \theta} = \frac{\partial \mathcal{L}_{\text{total}}}{\partial A^{\text{pred}}} \frac{\partial A^{\text{pred}}}{\partial S_\theta} \frac{\partial S_\theta}{\partial s_\theta} \frac{\partial s_\theta}{\partial \lambda_\theta} \frac{\partial \lambda_\theta}{\partial \theta}.$$

This chain rule—executed automatically by the deep-learning framework—ensures that both pointwise S supervision and the cross-section-level χ^2 drive the FiLM-conditioned λ -network toward an $s(x, k; Q)$ whose autoconvolution reproduces the measured $S(q_T; x_1, x_2, Q)$ while simultaneously satisfying the physics anchors and normalization constraints.

Conceptually, Eq. (58) realizes an *end-to-end differentiable inverse problem* in momentum space: the model learns to invert the $k_\perp$ convolution operator directly, without factorized ansätze or a detour through the b_T representation [60]. The differentiable structure allows the optimizer to exploit analytic gradients instead of finite-difference updates, yielding stable and physically interpretable solutions. Such end-to-end differentiable inversion schemes have become canonical in modern scientific machine learning for ill-posed inverse problems [67,68].

As a robustness check, we performed systematic ablation studies to assess the role of all auxiliary constraints introduced for early-stage stabilization. The essential cross section component of the loss and mass normalization constraint are retained by definition since it enforces PDF-TMD consistency; however, all other removable terms—including soft pins, moment anchors, curvature priors, and tail penalties—are used only during the initial prototyping phase to identify stable architectures and hyperparameters through recursive closure validation.

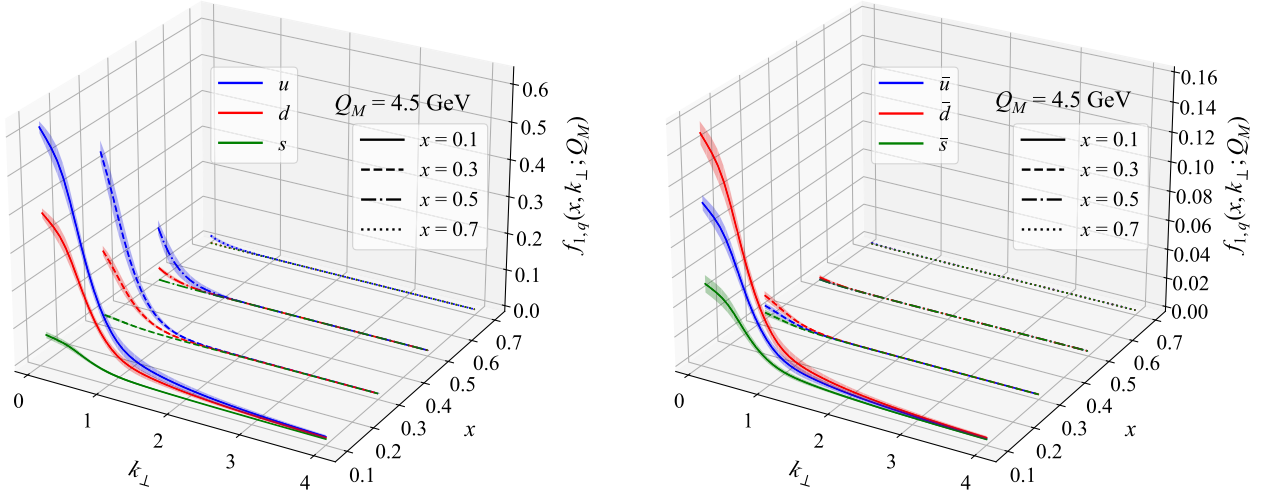


FIG. 6. Transverse-momentum-dependent parton distributions $f_{1,q}(x, k_{\perp}; Q_M)$ at $Q_M = 4.5$ GeV shown for u , d , and s quarks (left) and antiquarks (right) as functions of Bjorken- x and transverse momentum $k_{\perp}$. The distributions are reconstructed from the joint profile $s(x, k_{\perp}; Q)$ and assembled using Eq. (24). Line styles correspond to fixed- x slices, illustrating the evolution of the transverse-momentum width with x . All error contributions are shown.

Once convergence and stability are established, these auxiliary penalties are removed one by one in a controlled sequence of closure tests, and the final model is obtained without them. Across the resolved $(x, k_{\perp}, Q)$ domain, eliminating these removable constraints leads to changes below $\sim 3\%$ in both magnitude and shape of the extracted TMDs, with no statistically significant impact on the cross-section description. This confirms that the final reported solution is determined by the data and the inverse-convolution constraint itself, while the auxiliary terms serve only as numerical stabilizers during model conditioning and do not materially bias the extraction.

VIII. RESULTS

With the momentum-space pipeline of Secs. V–VII and the joint representation $f_{1,q/N}(x, k_{\perp}; Q, Q^2) = f_{1,q}(x; Q^2)s(x, k_{\perp}; Q)$, we extract the unpolarized TMDs for all quark-flavors in $SU(3)_{\text{flavor}}$ ($u, d, s, \bar{u}, \bar{d}, \bar{s}$) across the E288/E605 phase space. All results shown apply the standard kinematic cuts (including the Υ veto) and $q_T < 0.2Q$, where the small- q_T restriction emphasizes the TMD regime. The end-to-end fit reproduces the measured cross sections panel-by-panel and then inverts for the intrinsic profile $s(x, k_{\perp}; Q)$ via the differentiable $k_{\perp}$ convolution described in Sec. VII. Representative cross-section comparisons appear in Figs. 3 (for E288) and 4 (for E605).

A. Flavor dependence at a fixed scale

At fixed (x, Q) the reconstructed TMDs exhibit the expected flavor hierarchy without imposing a factorized ansätze. Figure 5 shows $f_{1,q}(x, k_{\perp}; Q, Q^2)$ (left) at $x = 0.1$ and $Q = 4.5$ GeV. The u distribution is largest, followed

by d and then s , with all curves falling smoothly in $k_{\perp}$. The normalization $\int d^2\mathbf{k}_{\perp} s(x, k_{\perp}; Q) = 1$ enforces PDF matching in Eq. (1); the band reflects the replica ensemble propagated through the two-stage fit. The $f_{\bar{q}}(x, k_{\perp}; Q, Q^2)$ distribution for antiquarks (right panel) shows $f_{\bar{d}} > f_{\bar{u}}$. To visualize the global behavior across $(x, k_{\perp})$ at a fixed scale, we show the u -quark surfaces at $Q = 4.5$ GeV and $Q = 7.5$ GeV in Figs. 6 and 7, respectively, for quarks (left) and antiquark (right) in each case. These plots make explicit the joint $x - k_{\perp}$ structure learned from the data in momentum space.

The extraction captures a mild but systematic Q dependence directly in $k_{\perp}$ space. Because the fixed-target Drell-Yan data that we used in this paper (E288, E605) probe only a limited lever arm in the hard scale, the resulting evolution of the intrinsic transverse-momentum profile is expected to be modest in direct $(x, k_{\perp})$ slices. As the scale increases the TMDs exhibit a redistribution of strength toward larger $k_{\perp}$ and a corresponding suppression of the small- $k_{\perp}$ peak, while the per- (x, Q) normalization of s preserves the overall mass of the profile. This trend can be discerned by comparing the TMD plots at $Q = 4.5$ and 7.5 GeV (Figs. 6 and 7) and is quantified more clearly by the moment-based width $\sigma_k(x, Q^2)$ shown in Fig. 10. The observed broadening is consistent with expectations from transverse-momentum resummation in the small- q_T domain, but here it is inferred empirically from the data within the momentum-space inversion, rather than imposed through a predefined b_T -space evolution kernel. Unlike b_T -space formulations where the Sudakov exponent makes the perturbative/nonperturbative transition explicit, our present implementation does not attempt such a separation; instead, it learns an effective Q dependence from data

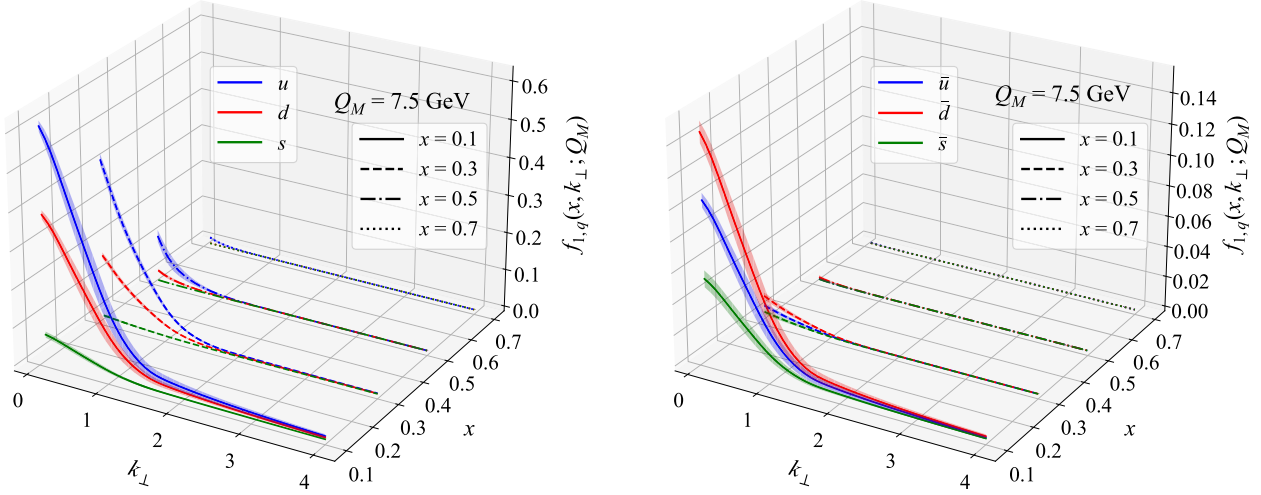


FIG. 7. Transverse-momentum-dependent parton distributions $f_{1,q}(x, k_\perp; Q_M)$ at $Q_M = 7.5$ GeV shown for u , d , and s quarks (left) and antiquarks (right) as functions of Bjorken- x and transverse momentum $k_\perp$. The distributions show mild $k_\perp$ broadening over the fixed-target lever arm Q_M . All error contributions are shown.

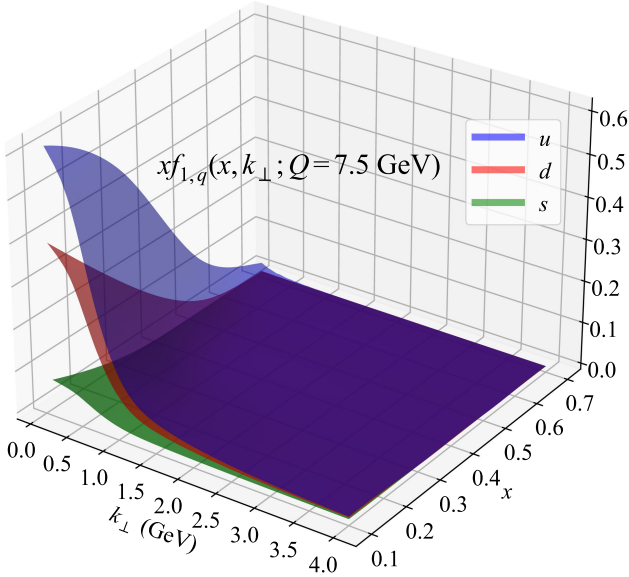


FIG. 8. Reconstructed $(x, k_\perp)$ surfaces of the unpolarized TMDs $f_{1,q}(x, k_\perp; Q, Q^2)$ for $q = u, d, s$ at fixed $Q = 7.5$ GeV shown to illustrate the joint $(x, k_\perp)$ structure and flavor hierarchy at a single scale.

within the fixed-target lever arm by design. The framework can be coupled to a standard Sudakov/matching prior when advantageous.

Figure 8 displays the reconstructed u -quark transverse-momentum-dependent parton distribution $f_{1,u}(x, k_\perp; Q)$ at a fixed scale $Q = 7.5$ GeV across the $(x, k_\perp)$ plane. The surface is obtained directly in momentum space from the joint profile $s(x, k_\perp; Q)$, assembled according to Eq. (24), and represents the replica-averaged mean of the full Monte Carlo ensemble propagated through

the differentiable convolution. At this fixed scale, the distribution exhibits a smooth, monotonic falloff in $k_\perp$, together with a gradual suppression at large x driven by the collinear-PDF normalization. The purpose of this figure is to illustrate the nonfactorized $(x, k_\perp)$ structure learned from the data at a single scale, rather than to demonstrate scale evolution. The Q dependence of the extracted transverse structure is assessed by comparing fixed- Q reconstructions at different scales and is quantified more directly by the RMS width $\sigma_k(x, Q^2)$ shown in Fig. 10. The surface also demonstrates that the neural network-based mapping preserves positivity and per- (x, Q) normalization while retaining the data-driven shape of the extracted TMD. The color map encodes the magnitude of $f_{1,u}(x, k_\perp; Q)$, with lighter tones corresponding to larger probability density at small $k_\perp$.

A complementary view of the evolution is provided at the fixed transverse momentum. Figure 9 shows u, d, s surfaces versus (x, Q) at $k_\perp \approx 0.5$ GeV. We focus on moderate-to-large values of x , where the E288/E605 data have the greatest constraining power; lower- x regions are better addressed in global analyses that include collider data. The pattern across flavors tracks the interplay between Sudakov broadening and the collinear baseline $f_{1,q}(x; Q^2)$, with the net change at fixed $k_\perp$ varying with x (as anticipated from DGLAP-driven x evolution of the collinear inputs) while remaining smooth by construction of $s(x, k_\perp; Q)$.

To compress the $(x, k_\perp)$ information into a single, scale-sensitive observable, we form the root mean square (RMS) transverse-momentum width

$$\sigma_k(x, Q) \equiv \sqrt{\langle k_\perp^2 \rangle - \langle k_\perp \rangle^2}, \quad (59)$$

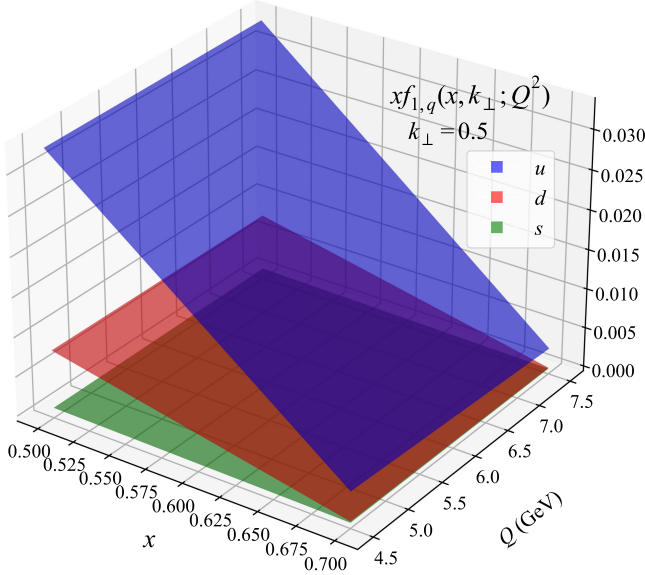


FIG. 9. Reconstructed (x, Q) surfaces of the unpolarized TMD at fixed representative values of $k_\perp$. The figure illustrates the joint dependence on Q and x in the kinematic region where the fixed-target data provide the strongest constraints.

where $\langle k_\perp^2 \rangle$ and $\langle k_\perp \rangle$ are defined as

$$\langle k_\perp^2 \rangle \equiv \frac{\int d^2 \mathbf{k}_\perp (k_\perp^2) w(x, k_\perp; Q)}{\int d^2 \mathbf{k}_\perp w(x, k_\perp; Q)}, \quad (60)$$

and

$$\langle k_\perp \rangle \equiv \frac{\int d^2 \mathbf{k}_\perp (k_\perp) w(x, k_\perp; Q)}{\int d^2 \mathbf{k}_\perp w(x, k_\perp; Q)}. \quad (61)$$

Here, we use as weight either the fully assembled unpolarized TMD $f_u(x, k_\perp; Q, Q^2)$ or, equivalently, the normalized profile $s(x, k_\perp; Q)$ since $\int d^2 \mathbf{k}_\perp s(x, k_\perp; Q) = 1$ by construction. This follows from the joint momentum-space representation $f_{1,q/N}(x, k_\perp; Q, Q^2) = f_{1,q}(x; Q^2) s \times (x, k_\perp; Q)$ and the per- (x, Q) normalization imposed in the inverse-integration stage, see Eqs. (8) and (9) and the assembly step summarized in Eq. (14).

Figure 10 shows $\sigma_k(x, Q_M)$ for the u flavor. Over the fixed-target kinematic window probed by E288 and E605, $\sigma_k(x, Q_M)$ exhibits a pronounced dependence on x and only a mild, systematic variation with Q_M mostly visible at large x , consistent with the limited hard-scale lever arm of these data. The dominant x dependence reflects non-trivial correlations between x and the intrinsic transverse-momentum profile learned by the momentum-space inversion, while the weak Q_M dependence encodes a modest redistribution of transverse momentum across scales within the fitted small- q_T domain. Because σ_k is constructed from normalized moments of the reconstructed distribution, it is insensitive to the overall collinear

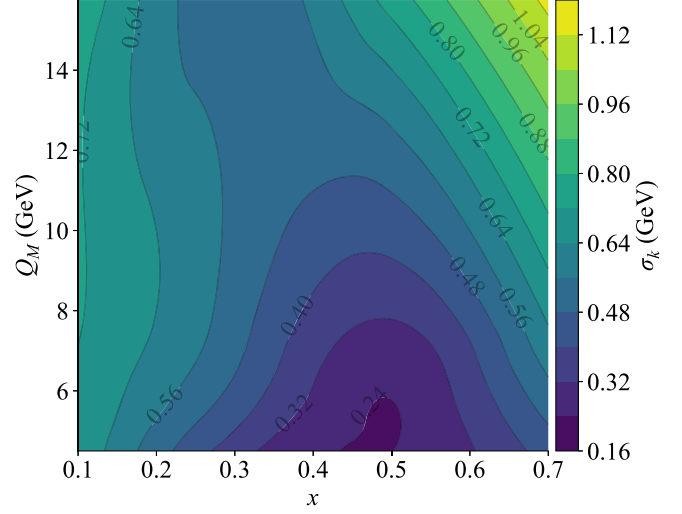


FIG. 10. Root mean square (RMS) transverse-momentum width $\sigma_k(x, Q_M)$ of the extracted unpolarized u -quark TMD computed from replica-averaged normalized moments as $\sigma_k = \sqrt{\langle k_\perp^2 \rangle - \langle k_\perp \rangle^2}$. The field is shown over the full (x, Q_M) domain covered by the E288 and E605 kinematics used in the analysis. So values displayed in that band reflect the smooth interpolation of the learned model. Across the fixed-target lever arm, σ_k exhibits a dominant dependence on x and a modest but systematic increase with Q_M that is most apparent at large x , consistent with the limited range in hard scale available in these data and with the fact that no explicit Sudakov kernel is imposed.

normalization and isolates genuine shape information. In this sense, the $\sigma_k(x, Q_M)$ field provides a compact scalar mapping of the transverse width learned directly from data, rather than a stand-alone determination of perturbative Sudakov evolution.

B. Uncertainty characterization

As detailed in Secs. III C and IV D, uncertainties are propagated through a Monte Carlo replica workflow augmented by a dedicated closure-based estimate of method-induced effects in the inverse step. Here, we summarize the adopted error components and the way they enter the final bands, and we report a compact phase-space-averaged budget in Table I. Terminology and metrics follow the generalized uncertainty framework of Ref. [79].

We organize the total uncertainty into four contributions: (i) experimental (cross-section uncertainties), (ii) collinear inputs [PDF uncertainty entering $f_{1,q}(x; Q^2)$], (iii) algorithmic/numerical (training stochasticity and evaluator settings), and (iv) methodological (bias and spread associated with the inverse reconstruction of the normalized profile $s(x, k_\perp; Q)$ and with other modeling choices not captured by the replica loop).

a. Experimental replicas: because full covariance matrices are not available, experimental uncertainties are propagated by generating Monte Carlo replicas of each reported

TABLE I. Phase-space-averaged uncertainties (relative to the TMD central value). Values are global averages across the resolved $(x, k_\perp, Q)$ region. All sources are propagated through the MC ensemble except the methodological terms, which are determined recursively (closure/tuning) and added *a posteriori*.

Source	Average scale and comment
Experimental	$\approx 12\%$ stat + 25% (15%) norm. E288 (E605); +10% syst. (E605).
Collinear PDF inputs	4–9% (typ. 6–7%); PDF replicas.
Algorithmic/numerical	1–3% (typ. 2%); DNN retraining + evaluator scans.
Methodological	< 1.5% total; inverse reconstruction of $s(x, k_\perp, Q)$.
Accuracy component	$\sim 1.0\%$ phase-space-averaged residual in s .
Precision component	$\sim 0.2\%$ generator-variation width; other method terms < 0.3%.

cross-section point using the quoted per-bin (stat + uncorrelated syst) uncertainties. In addition, where an overall (luminosity/normalization) uncertainty is quoted for a dataset, we treat it as a fully correlated systematic by introducing a replica-dependent nuisance normalization factor $N_d^{(k)}$ common to all bins i in dataset d ,

$$A_{i,d}^{(k)} = N_d^{(k)} A_{i,d}^{\text{data}} + r_{i,d}^{(k)} \sigma_{i,d}^{\text{ptp}}, \quad (62)$$

$$N_d^{(k)} = 1 + r_d^{(k)} \delta_d^{\text{norm}}, \quad r_d^{(k)}, r_{i,d}^{(k)} \sim \mathcal{N}(0, 1). \quad (63)$$

Here, $\sigma_{i,d}^{\text{ptp}}$ denotes the point-to-point (uncorrelated) uncertainty used for bin i , and δ_d^{norm} is the quoted relative normalization uncertainty for dataset d . For E605 we use $\delta^{\text{norm}} = 15\%$ and include the quoted 10% point-to-point systematic in quadrature with the published bin uncertainties [57], while for E288, we use $\delta^{\text{norm}} = 25\%$ [56] as a correlated normalization component (common within each E288 subset used in the fit).

b. Collinear (PDF) inputs: collinear-PDF uncertainty is propagated by sampling a PDF replica in sync with each experimental replica and assembling $f_{1,q/N}(x, k_\perp; Q, Q^2) = f_{1,q}(x; Q^2) s(x, k_\perp; Q)$ for that sample.

c. Algorithmic/numerical: we quantify residual stochastic and numerical scatter by repeating otherwise identical trainings with different random seeds and by coherently varying key evaluator settings (e.g., quadrature orders and grid parameters). This component is assessed *without* experimental smearing and is kept subdominant by construction.

d. Methodological error: to quantify method-induced systematics independently of experimental and PDF uncertainties, we evaluate a two-part budget: a *methodological accuracy* (bias under noiseless closure) and a *methodological precision* (spread under controlled variations of modeling choices). Let Ω denote a kinematic grid in (x, k, Q) with $k \equiv |\mathbf{k}_\perp|$ and empirical sampling weights $\rho(x, Q)$ (proportional to the training density). Define the pointwise residual and an L^2 accuracy metric

$$\Delta_{\text{acc}}(x, k, Q) \equiv s_{\text{extr}}(x, k, Q) - s_{\text{true}}(x, k, Q), \quad (64)$$

$$\mathcal{A}[s] \equiv \left[\sum_{(x,Q) \in \Omega_{x,Q}} \rho(x, Q) 2\pi \sum_{k \in \Omega_k} k \Delta_{\text{acc}}(x, k, Q)^2 \Delta k \right]^{1/2}, \quad (65)$$

which discretizes $\int dx dQ \rho(x, Q) \int dk (2\pi k) \Delta^2$ on the evaluator grid.

Stage I (kernel S): we scan the S -network architecture and training hyperparameters (depth/width, activations, optimizer, learning rate/batch size, and early stopping) to minimize the validation $\mathcal{L}_S$ while monitoring generalization (train vs val). The scan is embedded in a closure loop (noiseless pseudodata) until $\mathcal{A}[s]$ and the binwise χ^2 stabilize across admissible settings. Residual biases at this point are recorded as Stage I methodological *accuracy*; the dispersion across the admissible hyperparameter set defines the Stage I methodological *precision*.

Stage II (inverse integrand s): we perform a stricter recursive calibration of the differentiable evaluator and anchor weights:

- (1) Construct a realistic pseudodata generator with an analytic, normalized ground-truth profile $s_{\text{true}}(x, k, Q)$ and the exact forward map to S .
- (2) Run the full pipeline on noiseless pseudodata (closure) and compute $\mathcal{A}[s]$.
- (3) Fit the real data once to obtain replica means for S and s .
- (4) Use these means as the *truth* for a second pseudodata round to *tune* evaluator settings (radial/azimuthal quadrature orders, adaptive-node parameter, taper onset, tail mask) and anchor weights $(\lambda_{\text{CS}}, \lambda_{m_2}, \lambda_{\text{curv}}, \lambda_{\text{tail}}, \dots)$, reducing residual structure.
- (5) Iterate steps 2–4 until $\mathcal{A}[s]$ and per-panel residuals plateau.

The resulting noiseless bias $\Delta_{\text{acc}}(x, k, Q)$ is assigned to the Stage II methodological *accuracy*. Varying the pseudogenerator (e.g., smooth deformations of widths/tails within a narrow neighborhood of the calibrated truth) yields a spread of $\mathcal{A}[s]$, which we report as the Stage II

methodological *precision*. By construction, these methodological components are *not* produced by the MC replica loop; once their scales are fixed by the recursion above, they are added *a posteriori* as a dedicated systematic.

e. Assumptions not included in the methodological error: certain analysis choices are treated as fixed assumptions and are therefore *excluded* from the methodological variation accounting: (i) beam-target interchange symmetry $S(q_T; x_1, x_2, Q) = S(q_T; x_2, x_1, Q)$; (ii) the positivity constraint $s(x, k; Q) \geq 0$ (hence $f_{1,q/N} = f_q(x; Q^2)s \geq 0$); and (iii) inverse-problem stabilizers (moment anchors, curvature regularization, tail penalties, and tapering), which improve conditioning at the cost of some expressiveness. These ingredients are stress-tested in the methodological study, but any residual loss of information is not fully quantifiable with the present data alone.

f. Combination and intervals: unless stated otherwise, central values are replica means and uncertainty bands are empirical 68% central intervals of the *full* ensemble. Quantile intervals are used for robustness against non-Gaussian tails. Methodological terms are incorporated by augmenting the ensemble with the calibrated systematic described above.

The methodological accuracy is set by the phase-space-averaged residual of s in the recursive closure study; additional methodological contributions are found to be smaller and are included in the quoted precision line. The numerical values in Table I are intended as summaries; per-bin bands always reflect the full MC ensemble (augmented by the methodological systematic) at the quoted 68% confidence level.

C. A comparison with other work

A qualitative comparison of $xf_{1,u}(x, k_\perp; Q, Q^2)$ with the results presented in Ref. [28] and of $f_{1,u}(x, k_\perp; Q, Q^2)$ with those reported in Ref. [27], is shown in Fig. 11. This comparison is intended for orientation only and should not be interpreted as a like-for-like comparison across extractions with different perturbative content and datasets. In particular, the results of Ref. [28] are obtained from a global analysis, including fixed-target data from Fermilab together with hadronic collider data from the Tevatron, RHIC, and the LHC, and incorporate higher-order perturbative ingredients, whereas Ref. [27] is based on a fit to E288 data alone. By contrast, the present extraction is constrained only by fixed-target Drell-Yan data from E288 and E605 and is intentionally performed with a tree-level Drell-Yan hard factor and without imposing an explicit resummed Sudakov kernel or a fixed-order Y -term. In addition, the analyses vary in how comprehensive uncertainty quantification and error propagation is handled. For convenience, both TMD representations, $xf_{1,u}(x, k_\perp; Q, Q^2)$ and $f_{1,u}(x, k_\perp; Q, Q^2)$, are displayed in the same figure.

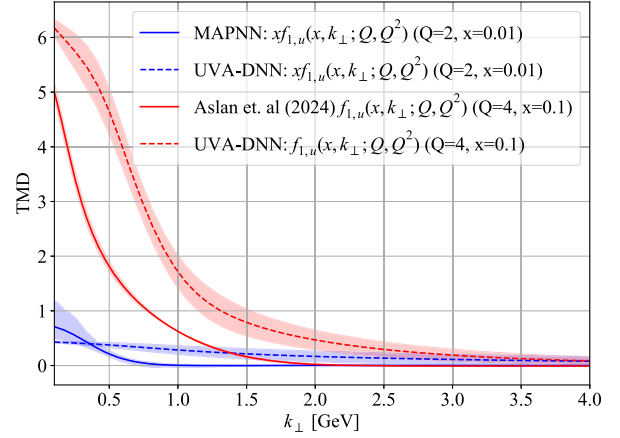


FIG. 11. Comparison of unpolarized u -quark TMD distributions. Results from MAPNN [28] (solid blue curve) obtained from a global analysis, including fixed-target and hadronic collider data (Tevatron, RHIC, and LHC) with higher-order perturbative ingredients; from Aslan *et. al.* [27] (solid-red-curve) based on a fit to E288 fixed-target data; and from this work (dashed curves) constrained by E288 and E605 fixed-target Drell-Yan data and performed with a tree-level Drell-Yan hard factor and no imposed Sudakov kernel or Y -term. The comparison is intended as a qualitative illustration of relative shape and magnitude rather than a matched perturbative-accuracy test.

IX. CONCLUSION

We have presented a momentum-space, physics-informed deep learning framework that extracts purely data-driven unpolarized TMDs directly from fixed-target Drell-Yan data without mapping to impact-parameter space for fitting. The method embeds a normalized, nonparametric profile $s_h(x, k_\perp; Q)$ as the *integrand* of the $k_\perp$ convolution and trains it end-to-end against cross sections through a differentiable quadrature layer. A two-stage pipeline first learns the structure kernel $S(q_T; x_1, x_2, Q_M)$ at the cross-section level and then solves the inverse problem to reconstruct $s_h(x, k_\perp; Q)$, enforcing per- (x, Q) normalization. This construction captures the joint $(x, k_\perp, Q)$ correlations supported by the data while avoiding restrictive factorized functional ansätze in the transverse sector.

Applied to E288/E605, the approach reproduces the measured q_T spectra panel-by-panel in the small- q_T domain, demonstrating that the learned integrand and forward map are well conditioned within the kinematic support of the data (Figs. 3 and 4). The extracted TMDs exhibit the expected flavor hierarchy and smooth profiles in $k_\perp$ that respect per- (x, Q) normalization. Global surfaces in $(x, k_\perp)$ at fixed Q (Figs. 6 and 7) and complementary (x, Q^2) slices at fixed $k_\perp$ (Fig. 9) make the learned multidimensional structure explicit. Over the fixed-target hard-scale window, the extracted scale dependence is mild but systematic, consistent with the limited lever arm in Q_M of these data; in the present implementation

it should be interpreted as an effective, empirically learned Q -dependence within the fitted small- q_T domain.

To characterize the transverse width through its second moment, we introduce the RMS width $\sigma_k(x, Q_M) = \sqrt{\langle k_\perp^2 \rangle - \langle k_\perp \rangle^2}$ computed from normalized moments of the reconstructed distributions. The resulting field $\sigma_k(x, Q_M)$ exhibits a pronounced dependence on x and only a moderate variation with Q_M (Fig. 10), particularly at high x , providing a compact summary of the transverse width that is insensitive to overall collinear normalization and thus isolates shape information learned in momentum space.

Uncertainties are propagated through Monte Carlo replica ensembles (1000 replicas) and reported as 68% central intervals. We distinguish experimental (cross section) and collinear (PDF) inputs from algorithmic components, and we include a separate methodological term associated with the small systematic deviations through end-to-end closure tests. The latter is quantified using a recursive pseudodata/closure schema performed on a tuned pseudodata generator, followed by iterative refits to the experimental cross section in which replica means from previous passes define an effective “truth” until residual deviations stabilize. The resulting phase-space-averaged residual in s_h defines a methodological accuracy component that is added *a posteriori* as a dedicated systematic. The final phase-space-averaged budget (Table I) shows experimental and PDF contributions as leading terms, with subdominant algorithmic uncertainty and a small but explicitly quantified methodological accuracy from the DNN fit to the cross section and the inverse reconstruction.

The present implementation is intentionally designed to be maximally data-driven in the transverse sector. We restrict to the TMD region $q_T < 0.2Q$, set $\mu = \sqrt{\zeta} = Q$, employ a tree-level hard factor, and do not impose an explicit resummed Sudakov kernel, fixed-order matching, or a Y -term; consequently no N^k LL resummation label is assigned. In addition, the fixed-target lever arm in Q_M bounds sensitivity to scale dependence, and the absence of published covariance matrices requires diagonal replica resampling of the reported uncertainties. Finally, in the present DY-only fixed-target study, we adopt a flavor-independent transverse profile within each hadron, with flavor dependence entering through the collinear PDFs;

broader datasets are required to robustly constrain fully flavor-dependent intrinsic $k_\perp$ shapes.

These scope choices do not limit the generality of the inverse-problem framework itself. The central methodological advance is the differentiable inverse-operator strategy—neural integrand, physics constraints, and end-to-end training through the $k_\perp$ convolution—which provides a transferable template for extracting other intrinsic partonic distributions defined implicitly by multidimensional integral transforms. The same machinery can be extended modularly to incorporate higher-order perturbative ingredients (hard factors, explicit evolution kernels, matching, and Y -terms when desired) if one wishes to enforce a chosen perturbative accuracy, and the strategy can be implemented equivalently in $k_\perp$ - or b_T -space. Likewise, the framework generalizes straightforwardly to flavor-dependent profiles $s_{q/h}$, hadron- and nucleus-dependent transverse structure, and global datasets, including collider DY/Z and SIDIS, which enlarge the lever arms in both x and Q . In this way, physics-informed differentiable inversion provides a scalable, low-bias methodology for future extractions that remain close to measured spectra while enabling systematic uncertainty accounting within a unified Monte Carlo workflow.

ACKNOWLEDGMENTS

The authors would like to thank Michael Wagman and Ted Rogers for their insightful discussions. We also thank Ted Rogers, J. O. Gonzalez-Hernandez (HSO-Approach), and C. Bissolotti (MAP Collaboration) for providing their TMD data points for the comparison. We acknowledge Research Computing at the University of Virginia for providing computational resources and technical support, which have contributed to the results reported in this publication. For additional information, see Ref. [80]. This work was supported by the U.S. Department of Energy under Contract No. DE-FG02-96ER40950.

DATA AVAILABILITY

The data that support the findings of this article are not publicly available. The data are available from the authors upon reasonable request.

-
- [1] X. Ji, Gauge-invariant decomposition of nucleon spin, *Phys. Rev. Lett.* **78**, 610 (1997).
 - [2] X. Ji, J.-P. Ma, and F. Yuan, Transverse-momentum-dependent gluon distributions and semi-inclusive processes at hadron colliders, *J. High Energy Phys.* **07** (2005) 020.

- [3] R. Boussarie *et al.*, TMD handbook, [arXiv:2304.03302](https://arxiv.org/abs/2304.03302).
- [4] X. Ji, J.-P. Ma, and F. Yuan, QCD factorization for spin-dependent cross sections in DIS and Drell–Yan processes at low transverse momentum, *Phys. Lett. B* **597**, 299 (2004).

- [5] X. Ji, J.-P. Ma, and F. Yuan, QCD factorization for semi-inclusive deep-inelastic scattering at low transverse momentum, *Phys. Rev. D* **71**, 034005 (2005).
- [6] A. Vladimirov, V. Moos, and I. Scimemi, Transverse momentum dependent operator expansion at next-to-leading power, *J. High Energy Phys.* **01** (2022) 110.
- [7] J.-Y. Chiu, A. Jain, D. Neill, and I. Z. Rothstein, A formalism for the systematic treatment of rapidity logarithms in quantum field theory, *J. High Energy Phys.* **05** (2012) 084.
- [8] M. G. Echevarría, A. Idilbi, and I. Scimemi, Factorization theorem for Drell-Yan at low q_T and transverse-momentum distributions on-the-light-cone, *J. High Energy Phys.* **07** (2012) 002.
- [9] T. Becher and M. Neubert, Drell-Yan production at small q_T , transverse parton distributions and the collinear anomaly, *Eur. Phys. J. C* **71**, 1665 (2011).
- [10] J. C. Collins, *Foundations of Perturbative QCD* (Cambridge University Press, Cambridge, England, 2011).
- [11] J. C. Collins, D. E. Soper, and G. Sterman, Factorization of hard processes in QCD, in *Perturbative Quantum Chromodynamics*, edited by A. H. Mueller (World Scientific, Singapore, 1989), pp. 1–91.
- [12] J. C. Collins and D. E. Soper, Back-to-back jets: Fourier transform from b to k_T , *Nucl. Phys.* **B197**, 446 (1982).
- [13] J. C. Collins and D. E. Soper, Back-to-back jets in QCD, *Nucl. Phys.* **B193**, 381 (1981).
- [14] J. C. Collins and D. E. Soper, Erratum: Back-to-back jets in QCD, *Nucl. Phys.* **B213**, 545 (1983).
- [15] M. Anselmino, M. Boglione, J. O. Gonzalez Hernandez, S. Melis, and A. Prokudin, Unpolarized transverse momentum dependent distribution and fragmentation functions from SIDIS multiplicities, *J. High Energy Phys.* **04** (2014) 005.
- [16] 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, *J. High Energy Phys.* **06** (2017) 081.
- [17] V. Bertone, I. Scimemi, and A. Vladimirov, Extraction of unpolarized quark transverse-momentum-dependent parton distributions from Drell-Yan/Z-boson production, *J. High Energy Phys.* **06** (2019) 028.
- [18] I. Scimemi and A. Vladimirov, Nonperturbative structure of semi-inclusive deep-inelastic and Drell-Yan scattering at small transverse momentum, *J. High Energy Phys.* **06** (2019) 137.
- [19] A. Vladimirov, Pion-induced Drell-Yan processes within TMD factorization, *J. High Energy Phys.* **10** (2019) 090.
- [20] M. Bury, F. Hautmann, S. Leal-Gomez, I. Scimemi, A. Vladimirov, and P. Zurita, PDF bias and flavor dependence in TMD distributions, *J. High Energy Phys.* **10** (2022) 118.
- [21] V. Moos, I. Scimemi, A. Vladimirov, and P. Zurita, Extraction of unpolarized transverse-momentum distributions from the fit of Drell-Yan data at N⁴LL, *J. High Energy Phys.* **05** (2024) 036.
- [22] A. Bacchetta, V. Bertone, C. Bissolotti, G. Bozzi, F. Delcarro, F. Piacenza, M. Radici, and A. Signori (MAP Collaboration), Unpolarized transverse-momentum distributions from a global fit of Drell-Yan and semi-inclusive deep-inelastic scattering data, *J. High Energy Phys.* **10** (2022) 127.
- [23] M. Boglione, J. O. Gonzalez-Hernandez, and A. Simonelli, Transverse-momentum-dependent fragmentation functions from recent BELLE data, *Phys. Rev. D* **106**, 074024 (2022).
- [24] P. C. Barry, L. Gamberg, W. Melnitchouk, E. Moffat, D. Pitonyak, A. Prokudin, and N. Sato, Tomography of pions and protons via transverse-momentum-dependent distributions, *Phys. Rev. D* **108**, L091504 (2023).
- [25] M. Boglione and A. Simonelli, Full treatment of the thrust distribution in single inclusive $e^+e^- \rightarrow hX$ processes, *J. High Energy Phys.* **09** (2023) 006.
- [26] A. Bacchetta, V. Bertone, C. Bissolotti, G. Bozzi, M. Cerutti, F. Delcarro, M. Radici, L. Rossi, and A. Signori (MAP Collaboration), Flavor dependence of unpolarized quark transverse-momentum distributions from a global fit, *J. High Energy Phys.* **08** (2024) 232.
- [27] F. Aslan, M. Boglione, J. O. Gonzalez-Hernandez, T. Rainaldi, T. C. Rogers, and A. Simonelli, Phenomenology of TMD parton distributions in Drell-Yan and Z^0 boson production in a hadron-structure-oriented approach, *Phys. Rev. D* **110**, 074016 (2024).
- [28] A. Bacchetta, V. Bertone, C. Bissolotti, M. Cerutti, M. Radici, S. Rodini, and L. Rossi, Neural-network extraction of unpolarized transverse-momentum-dependent distributions, *Phys. Rev. Lett.* **135**, 021904 (2025).
- [29] V. Moos, I. Scimemi, A. Vladimirov, and P. Zurita, Determination of unpolarized TMD distributions from the fit of Drell-Yan and SIDIS data at N⁴LL, *J. High Energy Phys.* **11** (2025) 134.
- [30] P. C. Barry, A. Prokudin, T. Anderson, C. Cocuzza, L. Gamberg, W. Melnitchouk, E. Moffat, D. Pitonyak, J. W. Qiu, N. Sato, A. Vladimirov, and R. M. Whitehill, First simultaneous analysis of transverse-momentum-dependent and collinear parton distributions in the proton, *arXiv:2510.13771*.
- [31] M. Cerutti, L. Rossi, S. Venturini, A. Bacchetta, V. Bertone, C. Bissolotti, and M. Radici (MAP Collaboration), Extraction of pion transverse momentum distributions from Drell-Yan data, *Phys. Rev. D* **107**, 014014 (2023).
- [32] L. Rossi *et al.*, New insights from the flavor dependence of quark transverse momentum distributions in the pion, *arXiv:2509.25098*.
- [33] M. Bury, A. Prokudin, and A. Vladimirov, Extraction of the Sivers function from SIDIS, Drell-Yan, and $W^\pm/Z$ -boson production data with TMD evolution, *J. High Energy Phys.* **05** (2021) 151.
- [34] M. Bury, A. Prokudin, and A. Vladimirov, Extraction of the Sivers function from SIDIS, Drell-Yan, and $W^\pm/Z$ data at next-to-next-to-next-to leading order, *Phys. Rev. Lett.* **126**, 112002 (2021).
- [35] M. Horstmann, A. Schäfer, and A. Vladimirov, Study of the worm-gear- t function g_{1T} with semi-inclusive DIS data, *Phys. Rev. D* **107**, 034016 (2023).
- [36] K. Yang, T. Liu, P. Sun, Y. Zhao, and B.-Q. Ma, First extraction of transverse-momentum dependent helicity distributions, *Phys. Rev. Lett.* **134**, 121902 (2025).
- [37] J.-C. He, M.-H. Chu, J. Hua, X. Ji, A. Schäfer, Y. Su, W. Wang, Y.-B. Yang, J.-H. Zhang, and Q.-A. Zhang (Lattice Parton Collaboration (LPC)), Unpolarized transverse momentum dependent parton distributions of the nucleon from lattice QCD, *Phys. Rev. D* **109**, 114513 (2024).

- [38] D. Boer *et al.*, Gluons and the quark sea at high energies: Distributions, polarization, tomography, <https://arxiv.org/abs/1108.1713>.
- [39] A. Accardi *et al.*, Electron-ion collider: The next QCD frontier, *Eur. Phys. J. A* **52**, 268 (2016).
- [40] R. Abdul Khalek *et al.*, Science requirements and detector concepts for the electron-ion collider, *Nucl. Phys. A* **1026**, 122447 (2022).
- [41] J. C. Collins, D. E. Soper, and G. Sterman, Transverse momentum distribution in Drell–Yan pair and W and Z boson production, *Nucl. Phys. B* **250**, 199 (1985).
- [42] I. Scimemi and A. Vladimirov, Analysis of vector boson production within TMD factorization, *Eur. Phys. J. C* **78**, 89 (2018).
- [43] A. Bacchetta, A. Bongallino, M. Cerutti, M. Radici, and L. Rossi, Exploring the three-dimensional momentum distribution of longitudinally polarized quarks in the proton, *Phys. Rev. Lett.* **134**, 121901 (2025).
- [44] M. G. Echevarría, Z.-B. Kang, A. Prokudin, and I. Vitev, QCD evolution of the Sivers asymmetry, *Phys. Rev. D* **89**, 074013 (2014).
- [45] A. Bacchetta, F. Delcarro, C. Pisano, and M. Radici, The three-dimensional distribution of quarks in momentum space, *Phys. Lett. B* **827**, 136961 (2022).
- [46] Z.-B. Kang, A. Prokudin, P. Sun, and F. Yuan, Extraction of quark Sivers functions from SIDIS data with QCD evolution, *Phys. Rev. D* **93**, 014009 (2016).
- [47] M. Anselmino, M. Boglione, U. D’Alesio, A. Kotzinian, F. Murgia, A. Prokudin, and C. Turk, Transversity and Collins functions from SIDIS and e^+e^- data, *Phys. Rev. D* **75**, 054032 (2007).
- [48] M. Anselmino, M. Boglione, U. D’Alesio, S. Melis, F. Murgia, and A. Prokudin, Simultaneous extraction of transversity and collins functions from SIDIS and e^+e^- data, *Phys. Rev. D* **87**, 094019 (2013).
- [49] Z.-B. Kang, A. Prokudin, P. Sun, and F. Yuan, Nucleon tensor charge from Collins azimuthal asymmetry measurements, *Phys. Rev. D* **91**, 071501 (2015).
- [50] W. Vogelsang and F. Yuan, Single-transverse-spin asymmetries: From deep inelastic scattering to hadronic collisions, *Phys. Rev. D* **72**, 054028 (2005).
- [51] P. Schweitzer, T. Teckentrup, and A. Metz, Intrinsic transverse parton momenta in deeply inelastic reactions, *Phys. Rev. D* **81**, 094019 (2010).
- [52] M. Anselmino, M. Boglione, U. D’Alesio, A. Kotzinian, F. Murgia, and A. Prokudin, Extracting the Sivers function from polarized SIDIS data and making predictions, *Phys. Rev. D* **72**, 094007 (2005).
- [53] M. Anselmino, M. Boglione, S. Melis, and A. Prokudin, Simultaneous extraction of transversity and Collins functions from new semi-inclusive deep inelastic scattering and e^+e^- data, *Phys. Rev. D* **87**, 094019 (2013).
- [54] A. Signori, A. Bacchetta, M. Radici, and G. Schnell, Investigations into the flavor dependence of partonic transverse momentum, *J. High Energy Phys.* **11** (2013) 194.
- [55] M. Aghasyan *et al.* (COMPASS Collaboration), Transverse-momentum-dependent multiplicities of charged hadrons in muon–deuteron deep inelastic scattering, *Phys. Rev. D* **97**, 032006 (2018).
- [56] A. S. Ito, R. J. Fisk, H. Jöstlein, D. M. Kaplan, S. W. Herb, D. C. Hom, L. M. Lederman, H. D. Snyder, J. K. Yoh, B. C. Brown, C. N. Brown, W. R. Innes, R. D. Kephart, K. Ueno, and T. Yamanouchi, Measurement of the continuum of dimuons produced in high-energy proton–nucleus collisions, *Phys. Rev. D* **23**, 604 (1981).
- [57] G. Moreno, C. N. Brown, W. E. Cooper, D. M. Kaplan, T. J. Nash, M. H. Schub, W. J. Spalding, T. A. Trippe, Y. C. Wang, H. J. Frisch *et al.*, Dimuon production in proton–copper collisions at $\sqrt{s} = 38.8$ GeV, *Phys. Rev. D* **43**, 2815 (1991). fermilab E605.
- [58] A. Bacchetta, V. Bertone, C. Bissolotti, G. Bozzi, F. Delcarro, F. Piacenza, and M. Radici, Transverse-momentum-dependent parton distributions up to N³LL from Drell–Yan data, *J. High Energy Phys.* **07** (2020) 117.
- [59] I. P. Fernando and D. Keller, Extraction of the Sivers function with deep neural networks, *Phys. Rev. D* **108**, 054007 (2023).
- [60] R. N. Bracewell, *The Fourier Transform and Its Applications*, 3rd ed. (McGraw–Hill, New York, 1999).
- [61] P. Shanahan, M. Wagman, and Y. Zhao, Collins–Soper kernel for TMD evolution from lattice QCD, *Phys. Rev. D* **102**, 014511 (2020).
- [62] A. G. Baydin, B. A. Pearlmutter, A. A. Radul, and J. Siskind, Automatic differentiation in machine learning: A survey, *J. Mach. Learn. Res.* **18**, 1 (2018), <https://par.nsf.gov/servlets/purl/10066574>.
- [63] F. Landry, R. Brock, P. M. Nadolsky, and C. P. Yuan, Tevatron run-1Z-boson data and Collins–Soper–Sterman resummation formalism, *Phys. Rev. D* **67**, 073016 (2003).
- [64] A. V. Konychev and P. M. Nadolsky, Universality of the Collins–Soper–Sterman nonperturbative function in gauge boson production, *Phys. Lett. B* **633**, 710 (2006).
- [65] R. D. Ball *et al.*, The path to proton structure at 1% accuracy, *Eur. Phys. J. C* **82**, 428 (2022).
- [66] A. Buckley, J. Ferrando, S. Lloyd, K. Nordström, B. Page, M. Rüfenacht, M. Schönherr, and G. Watt, LHAPDF6: Parton density access in the LHC precision era, *Eur. Phys. J. C* **75**, 132 (2015).
- [67] S. Arridge, P. Maass, O. Öktem, and C.-B. Schönlieb, Solving inverse problems using data-driven models, *Acta Numer.* **28**, 1 (2019).
- [68] J. Adler and O. Öktem, Learned primal–dual reconstruction, [arXiv:1707.06474](https://arxiv.org/abs/1707.06474).
- [69] A. Bacchetta, D. Boer, M. Diehl, and P. J. Mulders, Matches and mismatches in the descriptions of semi-inclusive processes at low and high transverse momentum, *J. High Energy Phys.* **08** (2008) 023.
- [70] M. Zaheer, S. Kottur *et al.*, Deep sets, [arXiv:1703.06114](https://arxiv.org/abs/1703.06114).
- [71] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, [arXiv:1512.03385](https://arxiv.org/abs/1512.03385).
- [72] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud, Neural ordinary differential equations, [arXiv:1806.07366](https://arxiv.org/abs/1806.07366).
- [73] B. Mildenhall *et al.*, NeRF: Representing scenes as neural radiance fields for view synthesis, [arXiv:2003.08934](https://arxiv.org/abs/2003.08934).
- [74] B. Amos, L. Xu, and J. Z. Kolter, Input convex neural networks, [arXiv:1609.07152](https://arxiv.org/abs/1609.07152).
- [75] S. You *et al.*, Deep lattice networks and partial monotonic functions, [arXiv:1709.06680](https://arxiv.org/abs/1709.06680).

- [76] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios, Neural spline flows, [arXiv:1906.04032](#).
- [77] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, Film: Visual reasoning with a general conditioning layer, [arXiv:1709.07871](#).
- [78] V. Dumoulin *et al.*, A learned representation for artistic style, [arXiv:1610.07629](#).
- [79] D. Keller, Experimental uncertainty propagation in neural network extraction in hadronic physics, [arXiv:2509.11456](#).
- [80] <https://rc.virginia.edu>.