

# Universality of scaling entropy in charged hadron multiplicity distributions at the LHC

L. S. Moriggi<sup>\*</sup>

*Universidade Estadual do Centro-Oeste (UNICENTRO),  
Campus Cedeteg, Guarapuava 85015-430, Brazil*

F. S. Navarra<sup>†</sup>

*Instituto de Física, Universidade de São Paulo,  
Rua do Matão, 1371, CEP 05508-090, Cidade Universitária, São Paulo, SP, Brazil*

M. V. T. Machado<sup>‡</sup>

*High Energy Physics Phenomenology Group, GFPAE, Institute of Physics,  
Federal University of Rio Grande do Sul (UFRGS),  
Caixa Postal 15051, CEP 91501-970, Porto Alegre, RS, Brazil*



(Received 11 June 2025; accepted 30 September 2025; published 16 October 2025)

In this work, we investigate the scaling behavior of the entropy associated with the charged hadron multiplicity distribution  $P(N)$  in proton-proton collisions at the LHC. We show that the growth of this entropic indicator as a function of the Bjorken- $x$  variable exhibits universal behavior, consistent with observations from deep inelastic scattering. This universality suggests that the entropy scaling is a property of the initial state and reflects the diffusive nature of gluon dynamics at small  $x$ . Furthermore, we demonstrate that high-multiplicity events are not accurately described by traditional Koba-Nielsen-Olesen scaling and require a more precise description based on a diffusion scaling framework. This new scaling emerges naturally from the universal growth of partonic entropy and offers deeper insight into the dynamics of particle production in high-energy hadronic collisions.

DOI: [10.1103/lynp-yzk5](https://doi.org/10.1103/lynp-yzk5)

## I. INTRODUCTION

Understanding the probability distribution  $P(N)$  for producing  $N$  charged hadrons in high-energy collisions is fundamental to unraveling the dynamics of quarks and gluons during the initial and final stages of the collision process. Although  $P(N)$  is one of the most extensively studied observables in high-energy physics, recent LHC data have pushed the limits of our theoretical understanding. In particular, *high-multiplicity events*—in which the number of produced particles greatly exceeds the mean—reveal significant discrepancies between theoretical predictions and experimental observations. These large fluctuations challenge conventional QCD expectations and expose tensions between standard event generators and

data not only in proton-proton ( $pp$ ), but also in heavy-ion ( $AA$ ) collisions.

The spectral shape of charged hadrons exhibits a strong dependence on event multiplicity, in a manner incompatible with existing QCD-inspired models and event generators [1]. Specifically, the variation of transverse momentum ( $p_T$ ) spectra across multiplicity classes leads to behaviors that are difficult to reconcile with current theoretical frameworks.

Without proper theoretical control of multiplicities in  $pp$  collisions, it becomes increasingly unrealistic to interpret final-state observables in heavy-ion collisions with confidence. In fact, phenomena traditionally attributed to quark-gluon plasma (QGP) formation may originate from a lack of understanding of the underlying gluon dynamics in the initial state. A recent ALICE study on multiplicities in  $pPb$  collisions [2] reveals large discrepancies between the data and the predictions from widely used Monte Carlo event generators. It also shows incompatibility with the GLASMA model [3], highlighting serious limitations in current theoretical approaches. Moreover, several QGP-like signals observed in  $AA$  collisions are now being detected in small systems like  $pp$  and  $pPb$  collisions [2], raising fundamental questions about their true origin. This

<sup>\*</sup>Contact author: [lucasmoriggi@unicentro.br](mailto:lucasmoriggi@unicentro.br)

<sup>†</sup>Contact author: [navarra@if.usp.br](mailto:navarra@if.usp.br)

<sup>‡</sup>Contact author: [magnus@if.ufrgs.br](mailto:magnus@if.ufrgs.br)

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI. Funded by SCOAP<sup>3</sup>.

convergence of phenomena across system sizes constitutes one of the most urgent open problems in high-energy hadron physics.

In the study of multiplicity distributions, one important subject of discussion is the famous Koba-Nielsen-Olesen (KNO) scaling [4,5], which is expected to be valid at asymptotically high energies. From the experimental side, KNO scaling is simply the fact that the function  $\langle N \rangle P(N)$  plotted against the variable  $N/\langle N \rangle$  is independent of collision energy  $\sqrt{s}$ . This prediction was made in the 1970s (before QCD) and was based on general assumptions about high-energy processes. From the theoretical side, there were several attempts to understand it in terms of quark and gluon dynamics. This has proven to be a difficult task, because particle production occurs in both soft (small momentum transfer) and hard (large momentum transfer) regimes, and we can perform reliable calculations only in the latter. Calculations with perturbative QCD (pQCD) have shown that KNO is found for particles produced in jets in high-multiplicity events [6,7], where  $N > \langle N \rangle$ . This conclusion depends on the choice of resummation schemes and relies on some assumptions. For minimum bias events, which include a large contribution from soft events, KNO scaling was obtained in the framework of the color glass condensate (CGC) approach [8,9]. This behavior relied on the assumption that the effective theory describing color charge fluctuations at a scale of the order of the saturation momentum is approximately Gaussian. Moreover, both nonlinear saturation effects and running-coupling evolution were required to obtain KNO scaling. More recently, a multiplicity distribution satisfying KNO scaling was derived by solving the Mueller dipole evolution equation in the double logarithm approximation [10]. In spite of all these efforts, a clear connection between KNO scaling and QCD has not yet been established.

The multiplicity distributions measured by ALICE at different center-of-mass energies exhibit significant deviations from KNO scaling, particularly in the tails of the distribution where  $N/\langle N \rangle > 4$  [11]. These deviations grow with increasing collision energy and are most pronounced in the high-multiplicity regime, where rare events dominate. The violation of KNO scaling reflects a breakdown of self-similarity and suggests the emergence of new dynamical scales not accounted for in early scaling hypotheses.

Various theoretical frameworks have been proposed to describe these deviations, including negative binomial distributions (NBDs) with soft and semihard components [12] and multicomponent models [13]. While these approaches offer reasonable phenomenological fits, they typically involve a large number of free parameters that lack a clear interpretation in QCD and often exhibit inconsistent behavior when analyzed across different kinematic regimes.

The CMS Collaboration also investigated KNO scaling at different pseudorapidity windows [14], showing stronger violations in broader rapidity intervals and at higher

multiplicities. A comprehensive ATLAS study over a wide energy range ( $\sqrt{s} = 0.9\text{--}13$  TeV) [15] confirmed that KNO violations become increasingly significant at lower energies and large multiplicities.

These findings raise several open questions:

- (i) Are the apparent collective effects in small systems the result of final-state interactions, or do they reflect intrinsic initial-state dynamics, such as gluon saturation?
- (ii) Can a universal scaling behavior be established across different collision systems, such as  $ep$ ,  $pp$ , and  $AA$ ?
- (iii) What mechanisms govern the rare, large-multiplicity events found in the tails of  $P(N)$ ?
- (iv) Is KNO scaling a fundamental feature, or should it be replaced by a more general scaling law?

While a fully robust theoretical framework is still lacking, the idea of *scaling laws* provides a promising bridge between empirical regularities in the data and theoretical models rooted in partonic dynamics. One such approach is based on *small- $x$  scaling*, motivated by gluon saturation models. In this picture, the gluon distribution reaches a saturation scale  $Q_s$ , which governs the typical transverse momentum and particle multiplicity. Evidence of such scaling behavior has been observed in inclusive particle production at both HERA [16–20] and the LHC [20–25].

In [21,26], a new observable called *scaling entropy* was introduced, proposed as a consequence of the universal growth of gluon distributions at small Bjorken- $x$ , resulting in the functional form:

$$S(x) = \lambda \log(1/x) + \text{const.}$$

Remarkably, the entropy extracted from the charged hadron multiplicity distribution in deep inelastic scattering (DIS) agrees with this prediction, and the growth exponent  $\lambda$  has been measured directly from the data.

In this study, we investigate the implications of scaling entropy in the context of multiplicity distributions at the LHC. We perform a phenomenological analysis of the probability distribution  $P(N)$  from ALICE, CMS, and ATLAS data. We demonstrate that, after subtracting the soft component of the spectrum, the entropy of hadron production in  $pp$  collisions exhibits scaling behavior consistent with that found in HERA  $ep$  data. This suggests the presence of universal initial-state dynamics governing particle production.

As a consequence of this behavior, we argue that *KNO scaling* must be replaced with a more accurate *diffusion scaling*, which naturally emerges from the universal growth of parton distributions and entropy in the initial state. This new perspective offers a path toward a deeper understanding of high-multiplicity events and their connection to fundamental QCD dynamics.

## II. THEORETICAL FRAMEWORK AND MAIN PREDICTIONS

In the dipole picture of DIS [27], the internal structure of the proton is probed via the scattering of a virtual photon. This process can be factorized into the convolution of the photon wave function and a dipole-target scattering amplitude  $\mathcal{N}(x, r)$ , resulting in the total cross section:

$$\sigma_{\gamma^* p}(x, Q^2) = \sigma_0 \int d^2\mathbf{r} dz |\psi_\gamma(r, z)|^2 \mathcal{N}(x, r), \quad (1)$$

where  $\mathbf{r}$  is the transverse size of the dipole,  $z$  is the longitudinal momentum fraction carried by the quark, and  $\psi_\gamma(r, z)$  is the photon light-cone wave function.

The scattering amplitude  $\mathcal{N}(x, r)$  must respect unitarity in position space, i.e.,  $\mathcal{N}(x, r) \leq 1$ , which ensures the saturation of interaction probability at large  $r$ . This leads to the emergence of a dynamical scale known as the *saturation scale*,  $Q_s(x)$ , which marks the transition between dilute and dense partonic systems and is a decreasing function of Bjorken- $x$ .

In momentum space  $k_T$ , the unitarity constraint manifests as the growth of the variance of the transverse momentum distribution with decreasing  $x$ . In a parton cascade picture, emissions follow an evolution where the characteristic timescale is governed by the uncertainty principle,  $t \sim 1/x$ , leading to anomalous diffusion in transverse momentum space:

$$\langle k_T^2(x) \rangle \sim \left( \frac{1}{x} \right)^\lambda, \quad (2)$$

where  $\lambda$  is related to the *hard Pomeron intercept*, characterizing the rate of growth of partonic activity and gluon density at small  $x$ .

We can define the Fourier transform  $\mathcal{P}(x, k_T)$  of the scattering amplitude  $\mathcal{N}(x, r)$ . It is normalized to unity and can be interpreted as a probability distribution containing all information about the interaction process at the partonic level. The form of  $\mathcal{P}(x, k_T)$  can be derived from the maximum entropy principle using Lagrange multipliers to optimize the distribution in a canonical ensemble. The most general form of nonextensive statistical mechanics, the Tsallis entropy [28], is given by

$$S_q = \int d^2k_T \frac{1 - [\mathcal{P}(x, k_T)]^q}{q - 1}, \quad (3)$$

with constraints

$$\langle k_T^2 \rangle_q = \frac{\int d^2k_T k_T^2 [\mathcal{P}(k_T)]^q}{\int d^2k_T [\mathcal{P}(k_T)]^q} = \beta^{-1}, \quad \int d^2k_T \mathcal{P}(k_T) = 1. \quad (4)$$

The Tsallis  $q$  index represents the degree of nonextensivity of the distribution.

This framework leads to a stationary state with a power-law distribution, as proposed in Refs. [20,21]. The Lagrange parameter  $\beta$  can be linked to the scaling hypothesis:

$$\langle k_T^2(x) \rangle_q \sim \beta^{-1} (x_s/x)^\lambda. \quad (5)$$

Since  $1/x$  corresponds to the timescale probed by the soft gluons, the interaction with gluons of varying energies results in an anomalous diffusionlike process. This connects  $\langle k_T^2(x) \rangle_q$  to a generalized form of Einstein's relation for diffusion. Under these conditions, the probability distribution can be expressed in a scaling form:

$$\mathcal{P}(x, k_T^2) \sim \frac{1}{x^{-\lambda}} f(k_T^2/x^{-\lambda}). \quad (6)$$

The resulting partonic entropy in the Tsallis formalism is given by

$$S_q^{\text{parton}}(x) = \frac{1}{q-1} - \left( \frac{2-q}{q-1} \right)^q (\pi Q_s^2(x))^{1-q}. \quad (7)$$

This general form has been used to establish a relationship between the saturation scale and the overlap area in  $pp$  collisions [21]. Scaling exists regardless of whether  $q = 1$  or  $q \neq 1$ ; however, integrated  $k_T$  data cannot distinguish differences associated with high- $k_T$  degrees of freedom. Such effects are observable in  $p_T$ -differential cross sections.

The entropy associated with the initial-state partonic distribution is defined by

$$S^{\text{parton}}(x) = - \int \mathcal{P}(x, k_T) \log[\mathcal{P}(x, k_T)] d^2k_T, \quad (8)$$

which is the Boltzmann-Gibbs form in  $q = 1$  limit. Under the scaling form (6), the entropy exhibits a universal logarithmic behavior:

$$S^{\text{parton}}(x) = C + \lambda \log\left(\frac{1}{x}\right), \quad (9)$$

where the constant  $C$  encodes all the model-dependent details of  $\mathcal{P}$  and the  $\lambda \log(1/x)$  term reflects a universal scaling behavior associated with the growth of gluon densities. This scaling relation underpins the anomalous diffusionlike behavior of gluons and connects the entropy to partonic momentum distributions.

Experimentally, we can define a related observable from the multiplicity distribution of charged hadrons:

$$S^{\text{mult}} = - \sum_N P(N) \log P(N), \quad (10)$$

where  $P(N)$  is the probability of producing  $N$  charged hadrons in a given event class.

In  $pp$  collisions,  $S^{\text{mult}}$  depends on the pseudorapidity interval  $\eta$  in which the measurement is performed and on the center-of-mass energy  $\sqrt{s}$  of the collision. The relevant Bjorken- $x$  value for the partonic subprocess can be estimated from the minimal transverse momentum  $p_T$  and the pseudorapidity of the produced particles:

$$x_{a,b} = \frac{p_T}{\sqrt{s}} e^{\pm\eta}. \quad (11)$$

Since most partonic interactions occur at low  $x$ , the effective value of  $x$  for the subprocess is approximately

$$x \approx \frac{e^{-\eta}}{\sqrt{s}}, \quad (12)$$

assuming  $p_T \sim 1$  GeV for soft-scale interactions.

Substituting this into Eq. (9), we can write a phenomenological scaling form for the experimentally measured entropy:

$$S^{\text{mult}}(\sqrt{s}, \eta) = \lambda \log\left(\frac{1}{x}\right) + f(\eta), \quad (13)$$

where the first term depends only on the initial-state gluon dynamics through  $x$  and the second term  $f(\eta)$  may include contributions from hadronization and final-state effects independent of  $\sqrt{s}$ . This scaling hypothesis provides a quantitative approach to test the principle of *local parton-hadron duality* [29], which posits that final-state hadron distributions closely reflect initial partonic configurations.

In DIS, Eq. (9) can be directly tested thanks to the hard scale  $Q^2$ . In contrast, LHC data are often collected with low- $p_T$  cuts, where soft physics contributions may obscure the scaling behavior. To mitigate this, we divide particle production into two components: a soft component that is dominant at low multiplicities ( $N < N_{\text{min}}$ ) and a semihard component that is dominant at high multiplicities ( $N > N_{\text{min}}$ ). Scaling behavior is expected only in the latter.

While there is no universal definition for  $N_{\text{min}}$ , our analysis (to be shown in the next section) reveals that for  $N_{\text{min}} \sim 10$  the extracted growth rate  $\lambda$  stabilizes and becomes consistent with that observed in HERA DIS data, i.e.,  $\lambda_{\text{LHC}} = \lambda_{\text{H1}}$ . This observation supports the universality of the entropy growth law across different collision systems.

To quantify fluctuations, we also consider the variance of the multiplicity distribution:

$$\text{Var}(N) = \sum (N - \langle N \rangle)^2 P(N). \quad (14)$$

The central goal of this work is to confront the theoretical prediction of parton scaling entropy in Eq. (9) with the

experimentally measured multiplicity entropy in Eq. (10) and to investigate whether a universal scaling behavior underlies particle production in high-energy hadron collisions.

Before presenting the main results, it is timely to compare the scaling entropy to recent entropy-based approaches to small- $x$  physics. One of them is the entanglement (von Neumann) entropy evaluated in the CGC approach, which is obtained by taking into account soft gluons in the wave function of a fast-moving hadron. The calculation can be done in the field basis or in the number representation basis [30]. The leading contribution in terms of saturation scaling has the following form [30]:

$$S_{EE}^{\text{CGC}} \propto \frac{1}{2} S_{\perp} \frac{C_F}{4\pi} \tilde{Q}_s^2, \quad (15)$$

where  $\tilde{Q}_s^2(x) = (9/4)(x_0/x)^{\lambda_g}$  and  $S_{\perp}$  are the gluon saturation scale and the transverse size of the hadron or nucleus, respectively. The entropy  $S_{EE}^{\text{CGC}} \sim x^{\lambda_g}$  has a stronger growth on  $x$  compared to the scaling entropy,  $S^{\text{parton}} \sim \ln(1/x)^{\lambda}$ .

Another formalism is the Wehrl entropy in QCD [31], which is the semiclassical analog of the von Neumann entropy. It is obtained in terms of quantum phase space distributions, such as Wigner or Husimi distributions. The QCD Wigner distribution  $W(x, \vec{k}, \vec{b})$  is a generalization of the usual collinear parton distribution functions and depends on parton transverse momentum  $\vec{k}$ , impact parameter  $\vec{b}$ , and longitudinal momentum fraction  $x$ . In case the Wigner distribution is positive definite, then the Wehrl entropy associated with the gluons can be defined in the following way [32]:

$$S_W = - \int d^2b d^2k x W_g(x, k, b) \ln [x W_g(x, k, b)], \quad (16)$$

where  $xg(x)$  is the usual collinear gluon distribution, with  $xg(x) = \int d^2b d^2k x W_g(x, k, b)$ . The Weizsacker-Williams gluon Wigner distribution [33,34] can be written in terms of the forward  $S$  matrix of a QCD color dipole of transverse size  $\vec{r}$ , transverse momentum  $\vec{k}$  at impact parameter  $\vec{b}$  in the adjoint representation,  $\mathcal{S}_A$ :

$$x W_g(x, k, b) = \frac{C_F}{2\pi^4 \alpha_s} \int d^2\vec{r} \frac{e^{i\vec{r} \cdot \vec{k}}}{r^2} (1 - \mathcal{S}_A(x, \vec{r}, \vec{b})). \quad (17)$$

The Weizsacker-Williams Wigner distribution can be analytically evaluated in the case of a Gaussian form for the  $S$  matrix,  $\mathcal{S}_A(x, r, b) = \exp[-\vec{r}^2 \tilde{Q}_s^2(x, b)/4]$ , where  $\tilde{Q}_s^2(x, b) = (N_c/C_F) Q_s^2(x, b)$  is the impact-parameter-dependent gluon saturation scale. Specifically, for the Gaussian  $S$  matrix one obtains the following Wehrl entropy [35]:

$$S_W \approx \frac{2N_c S_{\perp}}{6\pi^2 \alpha_s} Q_s^2(x), \quad (18)$$

where the parametric behavior is  $S_W \propto S_\perp Q_s^2(x)$ . This behavior is quite similar to the von Neumann entropy computed in the CGC formalism.

The parton entanglement entropy proposed in Ref. [36] (Kharzeev-Levin) is constructed based on the color dipole cascade equation in  $(1+1)$  dimensions. At small  $x$  the relation between the entanglement entropy and the gluon distribution,  $xg(x)$ , is simple:

$$S_{EE}^{KL} = \ln[xg(x)] \sim \ln x^{-\Delta}, \quad (19)$$

where  $\Delta = 4\alpha_s N_c \ln(2)/\pi$  is related to the hard Pomeron (Balitsky-Fadin-Kuraev-Lipatov) intercept. The main point is the identification of the probabilities of microstates  $p_n$  with the probabilities of finding  $n$  dipoles inside the hadron  $P_n(Y)$ . The entanglement entropy  $S = -\sum_n P_n \ln(P_n)$  takes a similar form in the  $(3+1)$ -dimensional case. The seminal paper [36] gave rise to a series of phenomenological analyses of this entropy notion in recent years [35,37–54]. This entropy is very close to ours, where the parameter  $\lambda$  is related, basically, to the energy dependence of the saturation scale  $Q_s$ .

Finally, we mention the QCD dynamical entropy [55]  $\Sigma^{Y_0 \rightarrow Y}$ . It provides a microscopic definition of entropy analogous to the Boltzmann approach, based on the QCD dynamics of the CGC, and can be understood as a sort of relative entropy or Kullback-Leibler divergence [56,57]. The microscopical point of view of this entropy is related to the balance between branching and recombination in the saturation regime of dense QCD states. In this regime, gluons organize themselves into cells of typical saturation scale size  $R_s(Y) = 1/Q_s(Y)$ , where  $Q_s \sim e^{\lambda Y}$  is the momentum saturation. From this, the nonlinear energy evolution can be evaluated as a compression  $R_s(Y_0) \rightarrow R_s(Y) < R_s(Y_0)$  for  $Y_0 < Y$ . The entropy is evaluated through the transverse momentum probability distributions  $P(k, Y)$ , where  $k$  is the gluon transverse momentum. These distributions are obtained from the unintegrated gluon distributions  $\varphi(k, Y)$ . In the simplest case of a Gaussian

in  $k$  unintegrated gluon distributions and the geometric scaling property, one obtains [58]

$$\begin{aligned} \Sigma^{Y_0 \rightarrow Y}(\Delta Y) &= \int d^2k P(k, Y) \ln \left[ \frac{P(k, Y)}{P(k, Y_0)} \right] \\ &= 2(e^{\lambda \Delta Y} - 1 - \lambda \Delta Y), \end{aligned} \quad (20)$$

where  $\Delta Y = Y - Y_0$ . It is hard to compare this notion of entropy to ours, since it is a relative entropy. However, the main input is similar: a transverse momentum distribution probability.

### III. RESULTS AND DISCUSSIONS

Different LHC collaborations adopt distinct methodologies for extracting the charged hadron multiplicity distribution  $P(N)$ , often applying different kinematic cuts, pseudorapidity windows, and transverse momentum thresholds. For this analysis, we compiled data from CMS, ALICE, and ATLAS across several center-of-mass energies  $\sqrt{s}$  and pseudorapidity intervals  $\eta$ , when available. We define four datasets for comparison:

- (i) *Set 1.* CMS data [14] with a large rapidity window, divided into five intervals from  $\eta_{\max} = 0.5$  to  $\eta_{\max} = 2.5$ . Non-single-diffractive events were selected using a minimum-bias trigger and a lower  $p_T$  threshold of 0.1 GeV. Each rapidity window is provided for different values of  $\sqrt{s}$ .
- (ii) *Set 2.* ATLAS data [15,59,60] with a large rapidity window of  $\eta_{\max} = 2.5$ . Events were collected using a single-arm minimum-bias trigger, with selections to reduce diffractive contributions. Only charged particles with transverse momentum  $p_T > 0.1$  GeV were included.
- (iii) *Set 3.* ALICE data [61]. Inclusive primary charged particles with  $0.15 \text{ GeV} < p_T < 10 \text{ GeV}$  were measured in a narrow pseudorapidity range of  $\eta_{\max} = 0.8$ .
- (iv) *Set 4.* ALICE data [11] at extended rapidity coverage at different values of pseudorapidity from 0.5 to 1.5.

TABLE I. Extracted  $\lambda$  values from different LHC datasets at various pseudorapidity windows and center-of-mass energies.

| Dataset | $p_T$ threshold (GeV) | $\eta_{\max}$ | $\sqrt{s}$ (GeV)              | $\lambda \pm \Delta\lambda$ | $\chi^2/\text{d.o.f.}$ |
|---------|-----------------------|---------------|-------------------------------|-----------------------------|------------------------|
| 1—CMS   | 0.1                   | 0.5           | 900, 2360, 7000               | $0.32 \pm 0.02$             | 0.0003                 |
| 1—CMS   | 0.1                   | 1.0           | 900, 2360, 7000               | $0.34 \pm 0.02$             | 0.0098                 |
| 1—CMS   | 0.1                   | 1.5           | 900, 2360, 7000               | $0.33 \pm 0.02$             | 0.0179                 |
| 1—CMS   | 0.1                   | 2.0           | 900, 2360, 7000               | $0.32 \pm 0.02$             | 0.0058                 |
| 1—CMS   | 0.1                   | 2.5           | 900, 2360, 7000               | $0.32 \pm 0.02$             | 0.0003                 |
| 2—ATLAS | 0.1                   | 2.5           | 2760, 900, 7000, 8000, 13000  | $0.32 \pm 0.01$             | 1.1045                 |
| 3—ALICE | 0.15                  | 0.8           | 2760, 5020, 7000, 8000, 13000 | $0.32 \pm 0.01$             | 0.9907                 |
| 4—ALICE | 0.05                  | 0.5           | 900, 2760, 7000, 8000         | $0.29 \pm 0.03$             | 0.4514                 |
| 4—ALICE | 0.05                  | 1.0           | 900, 2760, 7000, 8000         | $0.30 \pm 0.02$             | 0.4980                 |
| 4—ALICE | 0.05                  | 1.5           | 900, 2760, 7000, 8000         | $0.29 \pm 0.01$             | 0.5657                 |

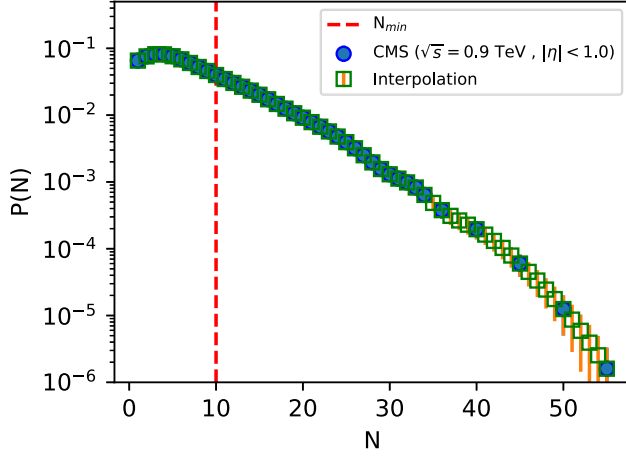


FIG. 1. Multiplicity data (full box) and interpolated values (unfilled) at some  $N$  bins. The vertical line shows the cut imposed of  $N_{\min} = 10$ .

Non-single-diffractive events with a lower  $p_T$  threshold of approximately 0.05 GeV.

Because of differences in analysis techniques and  $p_T$  thresholds, the multiplicity distributions in these sets are not directly comparable. A summary of the experimental characteristics and cuts applied is presented in Table I.

In some cases (notably CMS), the  $P(N)$  values are not reported for all consecutive values of  $N$ , which prevents direct computation of the entropy via Eq. (10). To address this, we perform a smooth interpolation using exponential cubic splines to estimate missing points and associated uncertainties. The procedure is illustrated in Fig. 1, where the interpolated bins are shown unfilled, and a vertical dashed line indicates the cutoff value  $N_{\min}$  applied to exclude soft contributions.

In ALICE set 3 data, the available  $N$  range is limited. To ensure convergence in entropy calculations, we extrapolate the high- $N$  tail of the distribution, which typically contributes about 5% to the total entropy at  $\sqrt{s} = 0.9$  TeV but significantly impacts the extracted value of  $\lambda$ . These preprocessing steps standardize the datasets for meaningful comparison.

We then fit the experimental entropy  $S^{\text{mult}}$  to the scaling form in Eq. (13), assuming a logarithmic dependence on  $x$  for fixed  $\eta$ . The fit extracts the entropy growth rate  $\lambda$  for each pseudorapidity interval. Results for sets 1 and 2 are shown in Fig. 2, where they are compared with the entropy extracted from deep inelastic scattering data [26].

We observe that the extracted values of  $\lambda$  across different LHC datasets and pseudorapidity windows are consistent, within uncertainties, with the DIS reference value [26]:

$$\lambda_{\text{DIS}} = 0.322 \pm 0.007. \quad (21)$$

In particular, the CMS data at  $\sqrt{s} = 0.9$  TeV for the most central pseudorapidity range yield entropy values

nearly identical to those in DIS. At larger rapidity windows, entropy increases with  $\sqrt{s}$ , but the rate  $\partial S / \partial \log(1/x)$  remains consistent with the universal value  $\lambda$ . This suggests that, despite experimental differences in  $P(N)$  extraction, the underlying entropy growth rate is a robust feature of QCD dynamics:

$$\frac{\partial S}{\partial \log(1/x)} = \lambda. \quad (22)$$

Next, we analyze ALICE sets 3 and 4 separately, as presented in Fig. 3. Set 3 (central rapidity) shows good agreement with  $\lambda_{\text{DIS}}$ , while set 4 (extended rapidity) exhibits slightly smaller entropy growth rates, around 7% lower than the other datasets. We believe that the difference is due to lower  $p_T$  thresholds in set 4, which allow for more soft events. Nevertheless, internal consistency is observed within the ALICE measurements at different  $\eta$ .

An important complementary observable is the variance of the multiplicity distribution. For example, in the case of the NBD, the variance is given by

$$\text{Var}(N) = \langle N \rangle^2 + \frac{\langle N \rangle^2}{k}, \quad (23)$$

where  $k$  is the shape parameter and both  $\langle N \rangle$  and  $k$  can vary with  $x$ . Thus, NBD-based models imply that both variance and entropy encode information about partonic dynamics.

Figure 4 shows the variation from the mean, quantified by the square root of the variance and entropy. We observe the linear behavior  $S = \log(\sqrt{\text{Var}(N)}) + \text{cte}$ . This indicates that the entropy is governed not primarily by the average particle number but by the event-by-event fluctuations around that mean. This dispersion originates from quantum fluctuations in the initial state, particularly in the small- $x$  regime. That means that the entropy effectively captures the spread of the parton distribution in momentum space.

Under the diffusion scaling hypothesis, the probability distribution for gluon  $k_T$  diffusion suggests a connection between entropy, variance, and initial-state  $x$  scaling. By analogy, the multiplicity distribution  $P(N)$  is expected to obey

$$P(N) \sim \frac{1}{x^\lambda} F\left(\frac{(N - \langle N \rangle)}{x^\lambda}\right), \quad (24)$$

where  $F$  is a universal scaling function. We do not formally derive a direct link between Eqs. (6) and (24); rather, this connection is proposed as a hypothesis to be tested against experimental data. Equation (6) describes a random walk in transverse momentum space, driven by gluon activity in the initial state. In contrast, Eq. (24) suggests that this underlying partonic behavior manifests experimentally as

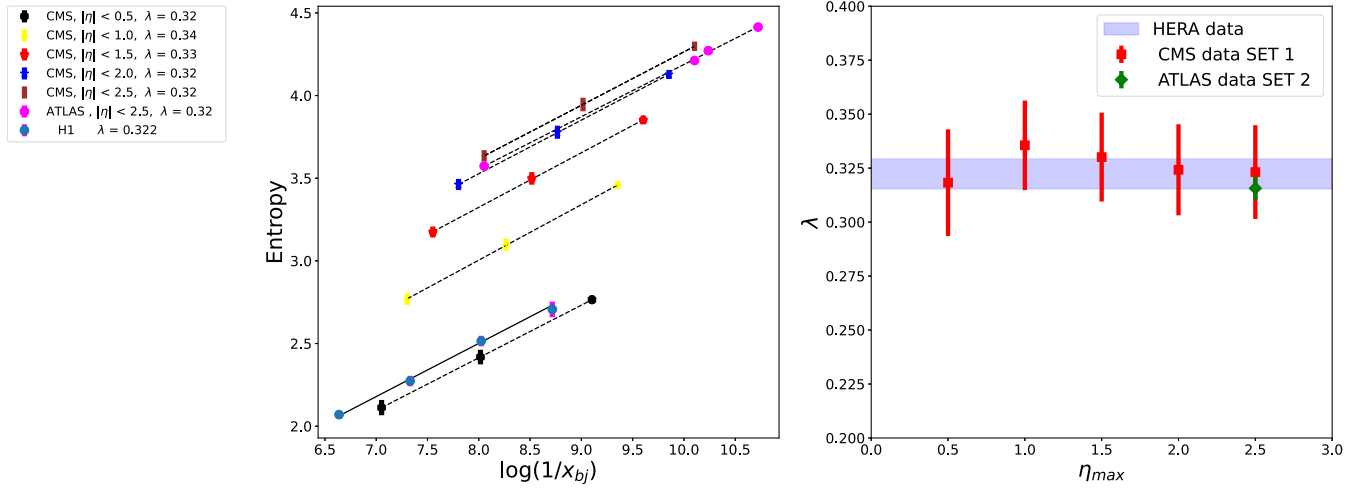


FIG. 2. Left side: experimental entropy (color bars) compared with scaling entropy line as a function of  $\log(1/x)$  in each rapidity bin when available. The full lines correspond to DIS entropy determined from H1 data. Right: resulting  $\lambda$  compared with blue interval corresponding from H1 fit uncertainty.

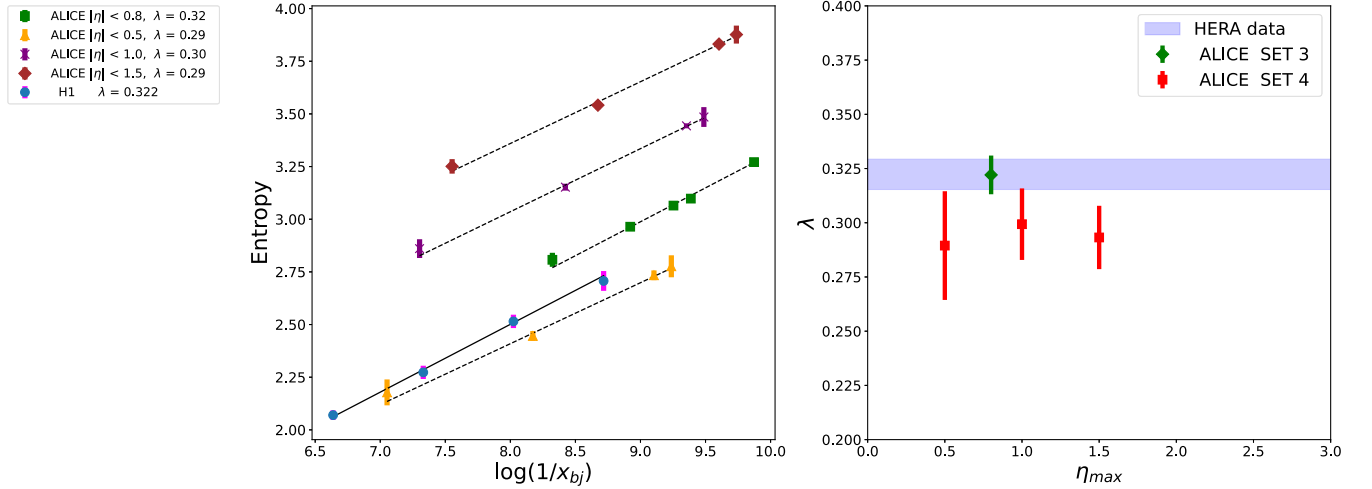


FIG. 3. The same as Fig. 2 for sets 3 and 4.

fluctuations in the number of produced particles, following a similar scaling trend. In essence, the idea is that fluctuations in the transverse momentum of the initial state are mapped onto multiplicity fluctuations in the final state.

Figure 5 shows the diffusion scaling behavior for all datasets analyzed, confirming the remarkable universality of this scaling form across different experiments and kinematic regimes.

Additionally, we compare the traditional KNO scaling with the proposed diffusion scaling using ATLAS data [15,59,60]. Figure 6 demonstrates that KNO scaling fails in the high-multiplicity tail ( $N \gg \langle N \rangle$ ), while diffusion scaling remains valid. It is worth mentioning that the data description is qualitatively good; however, there are still clear discrepancies. The diffusive scaling curve in Fig. 5 is obtained from the  $\lambda$  values extracted from the entropy.

Uncertainties in this measurement will also appear as discrepancies in the scaled probability  $P(N)x^{-\lambda}$ . The comparison is shown for multiple values of  $N_{min}$ , and we find that the diffusive behavior becomes more evident when soft events are excluded. Notably, KNO scaling appears more reliable at lower multiplicities, suggesting that large fluctuations originate from partonic systems exhibiting diffusionlike behavior rather than self-similar scaling.

The violation of the KNO scaling is interesting. In the context of the color glass condensate, the KNO scaling is related to the multiplicity distributions of soft gluons. It was shown in Ref. [9] that an approximately Gaussian effective theory (with running coupling evolution) describing color charge fluctuations at scales  $\sim Q_s$  is able to reproduce the scaling. In particular, the NBD distribution can be derived

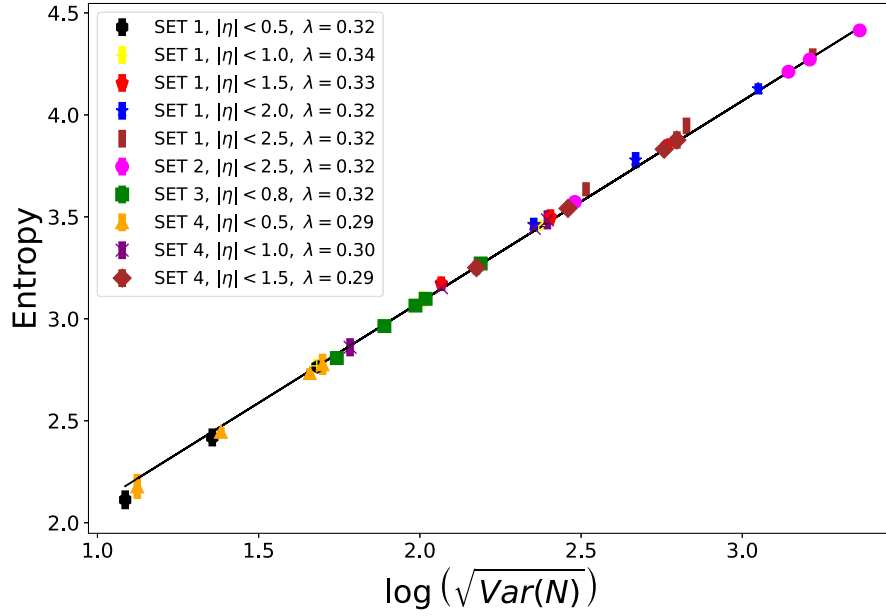


FIG. 4. Entropy as a function of the dispersion to each dataset. The full line is a linear global fit.

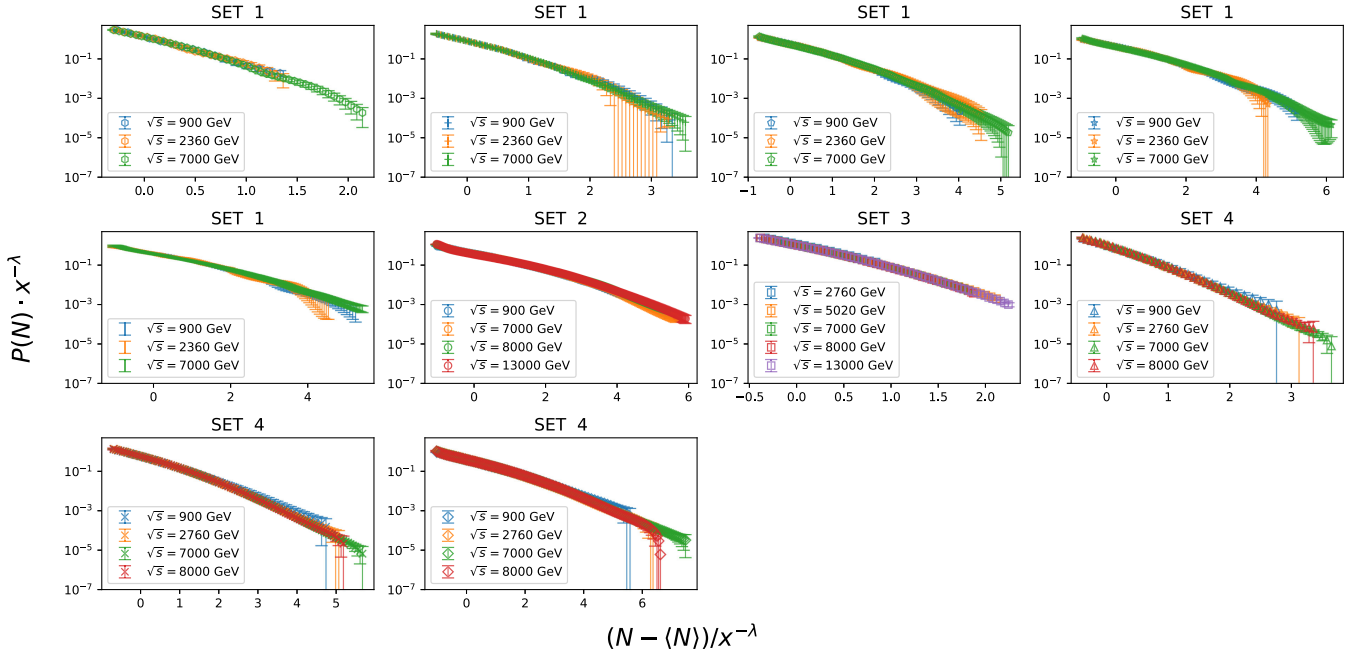


FIG. 5. Each dataset scaling line at different rapidities.

from this small- $x$  framework [9,62]. The role played by the selection of the transverse momentum  $p_T$  of produced charged particles within jets has been investigated recently in Ref. [63]. Based on the ATLAS data, it has been shown that the low- $p_T$  region does not exhibit KNO scaling, and the higher- $p_T$  range gradually moves toward the scaling. On the other hand, all data can be described by a sub-Poissonian distribution, with a distinct  $\delta$  parameter appearing in the birth rate  $\mu_n$  in the cascade equations when the

multiplicity is  $n$ ,  $\mu_n = \lambda(n+1)^{-\delta} + \sigma$ . The different  $\delta$  values from the fits suggest that the various transverse momentum ranges are driven by heterogeneous dynamics. Concerning high- $p_T$  jets, there is a general consensus that the double logarithmic approximation calculations in pQCD are able to describe the observed KNO scaling [6,10,64].

In an approach closest to ours, proposed by Levin and collaborators [38,48,52,65], the multiplicity distribution is

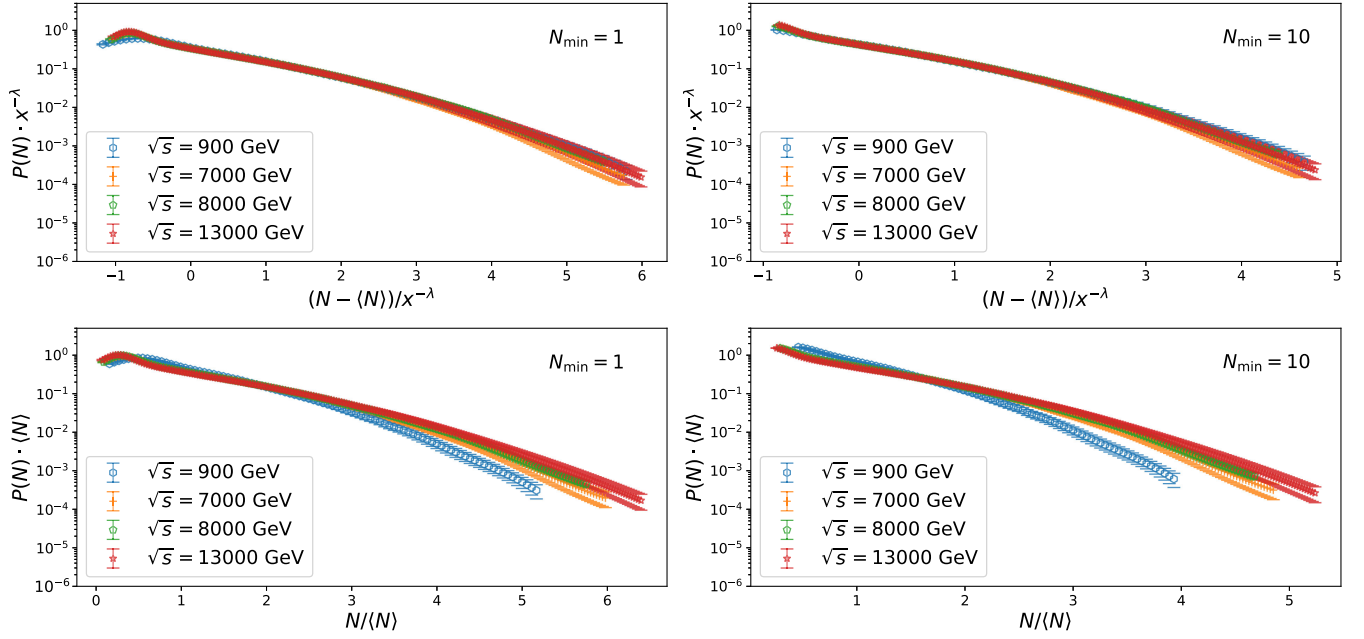


FIG. 6. Comparison between diffusion scaling (top) and KNO scaling (bottom) for two different values of  $N_{\min}$ . Data from ATLAS Collaboration [15,59,60], with  $|\eta| < 2.5$ .

related to the entanglement entropy of gluons in maximally entangled states. In the framework of the parton model, the high-energy multiplicity distribution has the following form [36]:

$$\frac{\sigma_n}{\sigma_{\text{in}}} = \frac{\bar{N}}{\bar{N} + 1} P^{\text{NBD}}(1, \bar{N}, n), \quad (25)$$

where quantities  $N$  (with  $\bar{N} = N - 1$ ),  $\sigma_n$ , and  $\sigma_{\text{in}}$  are the average multiplicity, the cross section for producing  $n$  hadrons, and the inelastic cross section, respectively.

The distribution of Eq. (25) is able to correctly describe the data at multiplicities  $n/\langle n \rangle \lesssim 4$  [38]. On the other hand, it overestimates the data for large multiplicities. In order to mitigate this shortcoming, an approach that sums all Pomeron diagrams (based on  $s$ -channel unitarity of the QCD Reggeon field theory) was considered [66]. The new multiplicity distribution now takes the form

$$P_n = \frac{1}{N} \left[ 1 - \left( \frac{1}{N} \right)^{1-\gamma} \right]^{n-1}, \quad (26)$$

where  $\gamma$  is the dipole-dipole scattering amplitude in the Born approximation of pQCD.

After this correction [Eq. (26)], the system of partons with large multiplicities is suppressed in relation to the distribution in Eq. (25), and LHC data are satisfactorily described [38]. The key point is that the distribution of Eq. (25) is obtained by considering the creation of a dilute system of partons at central rapidities, whereas the distribution in Eq. (26) is related to a dense system of partons.

The final part of our analysis focuses on assessing the suitability of applying a fixed multiplicity cut at  $N_{\min} = 10$ . To evaluate the impact of this choice on the extracted values of  $\lambda$ , we repeated the fitting procedure across a range of  $N_{\min}$  values.

When low-multiplicity events are included (i.e., small  $N_{\min}$ ), the entropy growth rate is reduced, as expected from the dominance of soft processes and the slow increase of the total cross section in the absence of strong kinematic constraints. As we increase  $N_{\min}$ , the contributions of semihard and hard processes become more significant, leading to faster entropy growth. This acceleration continues until a saturation point is reached around  $N_{\min} \sim 10$ , beyond which the growth rate stabilizes. This behavior is illustrated in Fig. 7, indicating that, above this threshold, particle production exhibits a universal scaling with respect to the Bjorken- $x$  variable. This is also the point in which the values of  $\lambda$  extracted from LHC data are equal to the HERA given by (21).

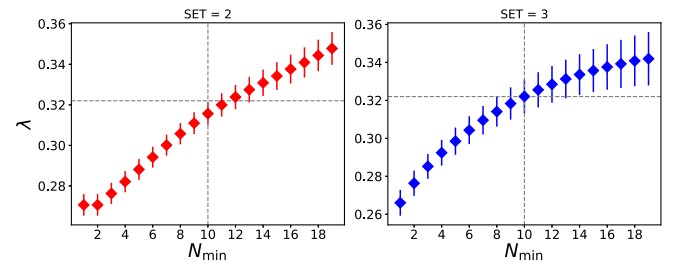


FIG. 7. The resulting  $\lambda$  from set 2 and set 3 considering different cuts in multiplicity  $N_{\min}$ .

#### IV. SUMMARY AND CONCLUSIONS

This work was driven by the central questions mentioned in the Introduction section. Namely, whether a universal behavior governs charged hadron multiplicities across different high-energy collision systems—from HERA to the LHC—and whether the growth of the so-called scaling entropy exhibits universality rooted in initial-state QCD dynamics.

Our analysis demonstrates that, once low-multiplicity events are excluded, the entropy scaling observed in LHC data matches with remarkable precision the trends previously established in deep inelastic scattering at HERA. The extracted entropy growth rate, characterized by the parameter  $\lambda$ , is consistent across experiments and pseudorapidity windows, reinforcing the idea of a common underlying mechanism.

Moreover, we showed that at large multiplicities—where fluctuations are most significant—the multiplicity distributions obey a precise diffusion scaling. This behavior strongly suggests that the observed growth in entropy and variance with energy is not a consequence of final-state effects or hadronization but rather reflects a universal scaling pattern associated with partonic dynamics in the initial state.

In conclusion, our findings support the existence of a universal entropy growth law and diffusion scaling behavior in high-energy hadronic collisions, bridging observations from  $ep$  scattering at HERA to  $pp$  collisions at the LHC. This provides further evidence that the most extreme multiplicity events encode fundamental information about the small- $x$  structure of hadrons and the dynamics of parton cascades.

Beyond its application to proton-proton collisions, the scaling entropy framework offers a valuable tool for disentangling initial- and final-state contributions in more complex systems, such as proton-nucleus and nucleus-nucleus collisions. Because entropy reflects early-stage partonic dynamics, it provides a clean observable largely insensitive to hadronization effects when analyzed above the soft-multiplicity threshold. This makes it a promising diagnostic for isolating initial-state physics in environments where final-state interactions, such as collective flow or quark-gluon plasma formation, are dominant.

Looking ahead, the scaling entropy method may also play an important role in future measurements at the upcoming Electron-Ion Collider, where precise control over the Bjorken- $x$  and  $Q^2$  kinematics will allow for stringent tests of universality and gluon saturation effects. Similarly, applying this framework to high-multiplicity events in heavy-ion collisions could help clarify which observables are rooted in initial-state structure versus those that emerge from final-state dynamics.

#### ACKNOWLEDGMENTS

M. V. T. M. acknowledges funding from the Brazilian agency Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) with Grant No. CNPq/303075/2022-8. F. S. N. gratefully acknowledges the support from the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

#### DATA AVAILABILITY

The data are not publicly available. The data are available from the authors upon reasonable request.

- 
- [1] S. Acharya *et al.* (ALICE Collaboration), *Eur. Phys. J. C* **79**, 857 (2019).
  - [2] S. Acharya *et al.* (ALICE Collaboration), *Eur. Phys. J. C* **85**, 919 (2025).
  - [3] B. Schenke, P. Tribedy, and R. Venugopalan, *Phys. Rev. C* **89**, 024901 (2014).
  - [4] A. M. Polyakov, *Sov. Phys. JETP* **32**, 296 (1971), <http://www.jetp.ras.ru/cgi-bin/e/index/e/32/2/p296?a=list>.
  - [5] Z. Koba, H. B. Nielsen, and P. Olesen, *Nucl. Phys.* **B40**, 317 (1972).
  - [6] Y. L. Dokshitzer and B. R. Webber, *J. High Energy Phys.* **08** (2025) 168.
  - [7] Y. L. Dokshitzer and B. R. Webber, [arXiv:2507.07691](https://arxiv.org/abs/2507.07691).
  - [8] A. Dumitru and Y. Nara, *Phys. Rev. C* **85**, 034907 (2012).
  - [9] A. Dumitru and E. Petreska, [arXiv:1209.4105](https://arxiv.org/abs/1209.4105).
  - [10] Y. Liu, M. A. Nowak, and I. Zahed, [arXiv:2302.01380](https://arxiv.org/abs/2302.01380).
  - [11] J. Adam *et al.* (ALICE Collaboration), *Eur. Phys. J. C* **77**, 33 (2017).
  - [12] A. Giovannini and R. Ugoccioni, *Phys. Rev. D* **59**, 094020 (1999); **69**, 059903(E) (2004).
  - [13] I. Zborovský, *Eur. Phys. J. C* **78**, 816 (2018).
  - [14] V. Khachatryan *et al.* (CMS Collaboration), *J. High Energy Phys.* **01** (2011) 079.
  - [15] G. Aad *et al.* (ATLAS Collaboration), *New J. Phys.* **13**, 053033 (2011).
  - [16] K. J. Golec-Biernat and M. Wusthoff, *Phys. Rev. D* **59**, 014017 (1998).
  - [17] A. M. Stasto, K. J. Golec-Biernat, and J. Kwiecinski, *Phys. Rev. Lett.* **86**, 596 (2001).
  - [18] M. Praszalowicz and T. Stebel, *J. High Energy Phys.* **03** (2013) 090.
  - [19] F. Gelis, R. B. Peschanski, G. Soyez, and L. Schoeffel, *Phys. Lett. B* **647**, 376 (2007).

- [20] L. S. Moriggi, G. M. Peccini, and M. V. T. Machado, *Phys. Rev. D* **102**, 034016 (2020).
- [21] L. S. Moriggi, G. S. Ramos, and M. V. T. Machado, *Phys. Rev. D* **110**, 034005 (2024).
- [22] L. McLerran and M. Praszalowicz, *Phys. Lett. B* **741**, 246 (2015).
- [23] T. Osada and T. Kumaoka, *Phys. Rev. C* **100**, 034906 (2019).
- [24] T. Osada, *Phys. Rev. C* **103**, 024911 (2021).
- [25] M. Praszalowicz, *Phys. Lett. B* **727**, 461 (2013).
- [26] L. S. Moriggi and M. V. T. Machado, *Phys. Rev. D* **111**, 014017 (2025).
- [27] N. N. Nikolaev and B. Zakharov, *Z. Phys. C* **49**, 607 (1991).
- [28] C. Tsallis, *J. Stat. Phys.* **52**, 479 (1988).
- [29] Y. L. Dokshitzer, V. A. Khoze, S. I. Troian, and A. H. Mueller, *Rev. Mod. Phys.* **60**, 373 (1988).
- [30] H. Duan, C. Akkaya, A. Kovner, and V. V. Skokov, *Phys. Rev. D* **101**, 036017 (2020).
- [31] Y. Hagiwara, Y. Hatta, B.-W. Xiao, and F. Yuan, *Phys. Rev. D* **97**, 094029 (2018).
- [32] Y. Hatta, B.-W. Xiao, and F. Yuan, *Phys. Rev. Lett.* **116**, 202301 (2016).
- [33] Y. V. Kovchegov and A. H. Mueller, *Nucl. Phys. B* **529**, 451 (1998).
- [34] F. Dominguez, C. Marquet, B.-W. Xiao, and F. Yuan, *Phys. Rev. D* **83**, 105005 (2011).
- [35] G. S. Ramos and M. V. T. Machado, *Phys. Rev. D* **101**, 074040 (2020).
- [36] D. E. Kharzeev and E. M. Levin, *Phys. Rev. D* **95**, 114008 (2017).
- [37] Z. Tu, D. E. Kharzeev, and T. Ullrich, *Phys. Rev. Lett.* **124**, 062001 (2020).
- [38] E. Gotsman and E. Levin, *Phys. Rev. D* **102**, 074008 (2020).
- [39] V. Andreev *et al.* (H1 Collaboration), *Eur. Phys. J. C* **81**, 212 (2021).
- [40] D. E. Kharzeev and E. Levin, *Phys. Rev. D* **104**, L031503 (2021).
- [41] E. Levin, *Phys. Rev. D* **104**, 056025 (2021).
- [42] K. Zhang, K. Hao, D. Kharzeev, and V. Korepin, *Phys. Rev. D* **105**, 014002 (2022).
- [43] Y. Liu, M. A. Nowak, and I. Zahed, *Phys. Rev. D* **105**, 114027 (2022).
- [44] Y. Liu, M. A. Nowak, and I. Zahed, *Phys. Rev. D* **105**, 114028 (2022).
- [45] M. Hentschinski, K. Kutak, and R. Straka, *Eur. Phys. J. C* **82**, 1147 (2022).
- [46] H. G. Dosch, G. F. de Teramond, and S. J. Brodsky, *Phys. Lett. B* **850**, 138521 (2024).
- [47] M. Hentschinski, D. E. Kharzeev, K. Kutak, and Z. Tu, *Phys. Rev. Lett.* **131**, 241901 (2023).
- [48] E. Levin, *Eur. Phys. J. C* **84**, 662 (2024).
- [49] P. Caputa and K. Kutak, *Phys. Rev. D* **110**, 085011 (2024).
- [50] M. Hentschinski, D. E. Kharzeev, K. Kutak, and Z. Tu, *Rep. Prog. Phys.* **87**, 120501 (2024).
- [51] Y. Hatta and J. Montgomery, *Phys. Rev. D* **111**, 014024 (2025).
- [52] E. Levin, *Phys. Rev. D* **111**, 016019 (2025).
- [53] M. Ouchen and A. Prygarin, arXiv:2508.12102.
- [54] K. Kutak and M. Praszalowicz, arXiv:2508.13781.
- [55] R. Peschanski, *Phys. Rev. D* **87**, 034042 (2013).
- [56] H. Umegaki, *Kodai Math. Sem. Rep.* **14**, 59 (1962).
- [57] S. Floerchinger and T. Haas, *Phys. Rev. E* **102**, 052117 (2020).
- [58] G. S. Ramos and M. V. T. Machado, *Phys. Rev. D* **105**, 094009 (2022).
- [59] G. Aad *et al.* (ATLAS Collaboration), *Eur. Phys. J. C* **76**, 403 (2016).
- [60] M. Aaboud *et al.* (ATLAS Collaboration), *Eur. Phys. J. C* **76**, 502 (2016).
- [61] S. Acharya *et al.* (ALICE Collaboration), *Phys. Lett. B* **845**, 138110 (2023).
- [62] F. Gelis, T. Lappi, and L. McLerran, *Nucl. Phys. A* **828**, 149 (2009).
- [63] G. R. Germano, F. S. Navarra, G. Wilk, and Z. Włodarczyk, *Phys. Rev. D* **110**, 034026 (2024).
- [64] X.-P. Duan, L. Chen, G.-L. Ma, C. A. Salgado, and B. Wu, arXiv:2503.24200.
- [65] E. Gotsman and E. Levin, *Eur. Phys. J. C* **81**, 99 (2021).
- [66] A. Kovner, E. Levin, and M. Lublinsky, *J. High Energy Phys.* **08** (2016) 031.