Physica A 681 (2026) 131077

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Blind identification of collective motion criticality using sequence model predictive entropy variance

Tianyi Wu, Zhangang Han *

School of Systems Science, Beijing Normal University, Beijing, 100875, People's Republic of China

ARTICLE INFO

Keywords:
Collective motion
Vicsek model
Sequence models
Phase transitions

ABSTRACT

Detecting critical transitions in collective systems from limited data is a central challenge in complex systems science. Traditional methods often require knowledge of control parameters and access to global system states, while typical machine learning approaches rely on multi-agent snapshots. Here, we introduce a novel, blind methodology to identify criticality in collective motion using only single-agent trajectory data. Our approach quantifies the temporal fluctuation in the predictive certainty of a sequence model (RWKV-7) tasked with forecasting an agent's movement. This metric, the predictive entropy variance ($\overline{\mathrm{Var}(H)}$), is hypothesized to peak at the transition. We rigorously test this method on the Vicsek model. Our results show that $\overline{\mathrm{Var}(H)}$ exhibits a sharp peak that accurately identifies the critical region, significantly outperforming a simpler statistical benchmark. Crucially, the method proves to be versatile: it successfully detects the transition in both low-density ($\rho = 0.5$) and high-density ($\rho = 2.0$) regimes. Furthermore, a model trained on data from the smallest system (L=32) demonstrates remarkable generalization, correctly identifying the critical point in larger systems (up to L=256) and even across different physical regimes. This work establishes predictive entropy variance as a robust, data-driven indicator for criticality, offering a powerful tool for analyzing complex systems from purely local observations.

1. Introduction

Collective motion phenomena are ubiquitous in nature and society [1–3]. Examples span biological scales, from bacterial swarms and biofilms where individuals coordinate motility [4] and collective cell migrations crucial for tissue development and wound healing [5], to the canonical examples of bird flocks and fish schools. Such coordination principles are also actively studied and engineered in artificial systems, like autonomous robot swarms designed for exploration or cooperative tasks [6], and provide frameworks for understanding complex human and social dynamics, including pedestrian crowds, vehicular traffic, and the formation and spread of collective opinions [7]. These systems often exhibit complex emergent behaviors, such as spontaneous pattern formation and coordinated large-scale movement, arising from simple local interactions among individual constituents. Understanding the principles governing these non-equilibrium systems is a central challenge in statistical physics and complex systems science.

A particularly fruitful paradigm for studying collective motion has been the investigation of phase transitions between distinct macroscopic states, analogous to equilibrium phase transitions. The Vicsek model (VM), introduced in 1995 [8], has become a cornerstone in this field. In its simplest form, it describes point particles moving at a constant speed that tend to align their direction

* Corresponding author.
E-mail addresses: wu.tianyi@mail.bnu.edu.cn (T. Wu), zhan@bnu.edu.cn (Z. Han).

https://doi.org/10.1016/j.physa.2025.131077
Received 6 May 2025; Received in revised form 23 August 2025
Available online 30 October 2025

of motion with their local neighbors, subject to a certain level of noise. Despite its simplicity, the VM exhibits a rich phenomenology. It displays a kinetic transition from a disordered, gas-like state at high noise to a state of collective motion (flocking) with long-range polar order at low noise, a phenomenon theoretically predicted by hydrodynamic theories [9,10] and extensively studied numerically [11]. The precise nature of this transition, particularly whether it is continuous or discontinuous (first-order), has been a subject of considerable research and debate, often depending on factors like density and specific model variants, with evidence suggesting complex behavior potentially involving inhomogeneous states like bands or phase separation, especially when density fluctuations are coupled [11–13]. Studying this transition provides fundamental insights into the mechanisms of self-organization in active matter systems [14,15].

Traditionally, the phase transition in the VM and related models is characterized by simulating the system across a range of the control parameter, typically the noise amplitude η, and measuring a global order parameter, such as the average particle polarization (or magnetization) Φ [8,11]. The transition point or critical region is often identified by analyzing the behavior of Φ or its susceptibility $\chi = N(\langle \Phi^2 \rangle - \langle \Phi \rangle^2)$, which is expected to peak near criticality. These standard measures have been widely employed in detailed numerical characterizations of the transition [11,12,16]. More recently, information-theoretic quantities like the full system's entropy have also been successfully employed to detect and characterize these transitions [17,18]. While powerful, this standard approach faces significant limitations when considering real-world systems. Firstly, it relies critically on the ability to precisely know and systematically vary the control parameter (η). However, in many biological, social, or ecological systems, the effective "noise" or equivalent governing parameters are often unknown, intrinsically fluctuating, or simply not externally controllable, posing major challenges for model parameterization, validation and dealing with inherent parameter uncertainty [19–21]. Secondly, calculating global order parameters necessitates access to the states (e.g., position and velocity) of a large fraction, if not all, individuals in the system simultaneously. While parameter-independent measures like the Binder cumulant are powerful for locating critical points from global snapshots, they still rely on this very assumption of system-wide data access. This level of global data acquisition is often infeasible in practice due to observational constraints such as occlusion, large group sizes, and limited sensor range [22–24].

The inherent difficulties in applying traditional modeling frameworks to complex real-world systems actively motivate the search for alternative methodologies. Specifically, challenges such as reliably inferring interaction rules or parameters from often noisy or incomplete empirical observations of collective behavior [25,26], and overcoming practical observational constraints that limit access to comprehensive, system-wide data [22,23], underscore the need for new approaches. Consequently, there is a growing emphasis on developing data-driven techniques capable of analyzing system behavior based primarily on available measurements, as seen in fields grappling with complex dynamics like disaster resilience [27], alongside parameter-agnostic frameworks such as equation-free modeling [28]. The goal is to devise methods that can detect critical states or phase transitions using potentially limited observational data (e.g., from a subset of individuals) without relying on prior knowledge or external control of underlying system parameters. Such 'blind identification' capabilities are particularly valuable, as they offer the potential for monitoring complex system states and anticipating critical shifts using only accessible time series data [29].

While machine learning (ML) offers powerful tools for analyzing phases [30,31], prominent approaches often analyze spatial configurations via methods like Convolutional Neural Networks (CNNs) [32–34] or apply unsupervised learning to configuration snapshots [35,36]. Such methods frequently rely on analyzing system-wide or global configuration data, which inherently requires simultaneous state information from many agents across the system. Furthermore, some ML strategies operate within supervised frameworks, for instance, training classifiers to distinguish between phases, which necessitates labeled data (i.e., knowing which phase each configuration belongs to) and thus potentially inherits limitations regarding parameter dependence [37]. In stark contrast, our work explores the potential of using only the information embedded within the temporal dynamics of a *single agent's trajectory*. While the principle of ergodicity suggests that a single long trajectory theoretically contains information about the system's collective states, the practical challenge lies in extracting this subtle information effectively. Simple statistical analyses often fail to reveal the signature of a collective phase transition, a point we will demonstrate explicitly in this work. This motivates a shift in perspective: instead of relying on global snapshots, can we infer collective properties from the much more accessible data of a single trajectory? Encouragingly, recent work has shown that it is indeed possible to infer global properties, such as the size of a collective, purely from the statistical analysis of a single unit's random motion [38]. Building upon this principle, our work explores the potential of using the rich temporal dynamics embedded within a single agent's trajectory not just to infer static properties, but to detect the dynamic signature of a collective critical transition. We propose a fundamentally different approach that moves beyond direct statistical analysis of the trajectory itself. Instead, we focus on probing the *dynamic predictability* of the agent's motion by leveraging a sequence model. We hypothesize that the characteristic fluctuations and complex dynamics near a critical region will cause the predictive certainty of an appropriate sequence model to vary significantly over time, and that this variation, quantified by the *variance of the predictive Shannon entropy*, will exhibit a distinct peak signaling proximity to criticality. This approach quantifies the fluctuations in the learnability of the dynamics, a higher-order feature that we posit is a highly sensitive signature of criticality.

Capturing the subtle temporal patterns underlying this hypothesis necessitates powerful sequence modeling. While architectures like Long Short-Term Memory (LSTM) networks were developed to mitigate challenges such as the vanishing gradient problem inherent in simpler recurrent networks [39,40], effectively capturing dependencies over *very* long sequences can still pose difficulties for these models. The Transformer architecture [41] significantly advanced capabilities for handling long-range dependencies, primarily through its attention mechanism. However, this mechanism typically incurs computational complexity that scales quadratically with the sequence length, potentially limiting its applicability to the extremely long time series that might be required to fully capture phenomena like critical slowing down near phase transitions.

Efficient architectures like the Receptance Weighted Key Value (RWKV) model [42–45] have emerged aiming to bridge this gap, offering linear computational scaling suitable for very long sequences while incorporating sophisticated state-updating mechanisms. We specifically employ the recent RWKV-7 variant [45]. Its potential suitability for modeling critical dynamics stems from several factors: (i) Its linear complexity allows for the efficient processing of the long sequences necessary to potentially capture the extended temporal correlations arising from critical slowing down. (ii) Its enhanced theoretical expressive power and state-tracking capabilities compared to certain classes of recurrent models [46,47], attributed to its non-diagonal, delta-rule-based state updates [45], offer promise for modeling the potentially complex, non-trivial history dependence characteristic of the critical region.

Leveraging this well-suited architecture, our methodology involves training dedicated RWKV-7 models on composite datasets of single-agent trajectories from different physical regimes of the Vicsek model. By calculating the variance of the Shannon entropy of the model's sequential predictions, we derive our proposed indicator, $\overline{\mathrm{Var}(H)}$.

The main contributions of this work are threefold: (1) We introduce the predictive entropy variance as a novel indicator for criticality and demonstrate its effectiveness using only single-agent trajectory data. (2) We validate this indicator against the global order parameter and prove its superiority over simpler statistical benchmarks. (3) We establish the method's universality by showing its success in both low- and high-density regimes, reveal the model's generalization capabilities across different system sizes, and confirm its consistency with the principles of finite-size scaling.

The remainder of this paper is organized as follows: Section 2 details our extensive simulation protocols, the data preprocessing steps, the model training strategy, and the formulation of our indicator and benchmarks. Section 3 presents the core results, systematically demonstrating the indicator's validation, superiority, universality, and finite-size scaling behavior. Section 4 provides a physical interpretation of these findings, discusses their broader implications, and outlines future research directions. Finally, Section 5 provides concluding remarks.

2. Methodology

2.1. The Vicsek model and simulation protocol

We investigate collective motion using the standard two-dimensional Vicsek model with angular noise [8]. The system comprises $N = \rho L^2$ point particles moving within a square domain of linear size L with periodic boundary conditions. To assess the properties of our proposed method, we performed simulations across different system sizes and particle densities.

Specifically, we focused on two representative density regimes. For a low-density regime ($\rho = 0.5$), simulations were conducted for system sizes $L \in \{32, 64, 128, 256\}$ to investigate the finite-size effects on our indicator. For a high-density regime ($\rho = 2.0$), we used system sizes of $L \in \{32, 64\}$ to test the method's applicability in a different dynamical context. In all simulations, particles move with a constant speed $v_0 = 1$, and the interaction radius is set to $R = 1$.

The system evolves in discrete time steps. At each step t, particle i updates its angle of motion θ_i according to:

$$\theta_i(t+1) = \arg\left(\sum_{j \in \mathcal{N}_i(t)} e^{i\theta_j(t)} \right) + \Delta\theta_i(t), \tag{1}$$

where $\mathcal{N}_i(t)$ is the set containing particle i and its neighbors j satisfying $d_{ij}(t) \le R$. The noise term $\Delta\theta_i(t)$ is drawn independently for each particle and time step from a uniform distribution $U[-\eta/2, \eta/2]$. The noise amplitude η serves as the control parameter, varied systematically from 0.1 to 4.0 in steps of $\Delta\eta = 0.1$.

For each parameter set (L, ρ, η), simulations were run for a total of $T_{total} = 100,000$ steps. To ensure measurements were taken in the statistically stationary state, the initial $T_{eq} = 50{,}000$ steps were discarded. A total of $N_{runs} = 20$ independent simulation runs were performed for each parameter set, starting from random initial particle positions and orientations.

2.2. Time series acquisition and preprocessing

For each simulation run conducted in the statistically stationary state ($t > T_{eq}$), the angular trajectory $\theta(t)$ of a single, randomly selected particle was recorded. This trajectory was sampled every $\Delta t_{sample} = 50$ simulation steps, resulting in a time series containing $T_{seq} = 1000$ data points per run. We denote this raw sequence as $\{\theta(t_k)\}_{k=1}^{T_{seq}}$, with angles represented in degrees within the range $[0, 360)$.

The raw angle time series subsequently underwent two essential preprocessing stages. First, to reduce high-frequency noise inherent in the stochastic simulation and to allow the model to focus on more persistent directional changes indicative of the collective state, we employed a vector-averaging Exponential Moving Average (EMA). Each angle $\theta(t_k)$ was initially converted to radians ($\theta_{rad}(t_k) = \theta(t_k) \times \pi/180$) and then represented as a 2D unit vector $(x_k, y_k) = (\cos\theta_{rad}(t_k), \sin\theta_{rad}(t_k))$. An EMA with a span parameter $W_{EMA} = 10$ was applied independently to the sequences of x-components $\{x_k\}$ and y-components $\{y_k\}$, yielding smoothed component sequences $\{\bar{x}_k\}$ and $\{\bar{y}_k\}$. The smoothed angle in radians, $\theta_{smooth,rad}(t_k)$, was then reconstructed using the four-quadrant inverse tangent, $\arctan2(\bar{y}_k, \bar{x}_k)$ (Eq. (2)). Finally, these smoothed radian angles were converted back to degrees and mapped consistently to the $[0, 360)$ range using Eq. (3), yielding the smoothed, continuous angle sequence $\{\theta_{smooth}(t_k)\}_{k=1}^{T_{seq}}$.

$$\theta_{smooth,rad}(t_k) = \arctan2(\bar{y}_k, \bar{x}_k) \tag{2}$$

$$\theta_{smooth}(t_k) = (\theta_{smooth,rad}(t_k) \times 180/\pi + 360) \mod 360 \tag{3}$$

Second, following the smoothing procedure, the continuous angles $\theta_{smooth}(t_k)$ were discretized into $N_{bins} = 360$ integer values, corresponding to degrees $[0, 359]$. This discretization was achieved by rounding the smoothed angle to the nearest integer degree, employing banker's rounding for half-integer cases (consistent with 'numpy.around'), and ensuring the result wraps correctly within the $[0, N_{bins} - 1]$ range using the modulo operation, as shown in Eq. (4).

$$\theta_{discrete}(t_k) = \text{round}(\theta_{smooth}(t_k)) \mod N_{bins}. \tag{4}$$

This sequence of preprocessing steps produced the final integer time series $\{\theta_{discrete}(t_k)\}_{k=1}^{T_{seq}}$ used as input for the sequence model, generated for each of the N_{runs} runs at every noise level η for a given system size L.

2.3. Sequence modeling via RWKV-7

The temporal structure of the discretized angle sequences was modeled using the RWKV-7 architecture [45], a novel sequence modeling architecture that blends the strengths of Transformers and Recurrent Neural Networks (RNNs). As an RNN, RWKV-7 enjoys constant memory and computational cost during inference, allowing it to efficiently process arbitrarily long sequences. However, its design also permits fully parallelizable training, similar to a Transformer. This combination results in a linear time complexity ($O(L)$) with respect to sequence length L, making it an ideal candidate for analyzing the long temporal correlations present in critical dynamics.

Mechanistically, RWKV-7 features a highly expressive state-updating mechanism based on a generalized "delta rule". A key distinction from Transformers, whose attention mechanism operates on an immutable Key-Value cache of past states, is that RWKV-7's internal RNN state is mutable and completely recomputed at each timestep. This allows the model to actively "edit" its memory by selectively adding new information and removing old information on a per-channel basis. This capability for dynamic state tracking is theoretically powerful — enabling RWKV-7 to solve problems beyond the recognized capabilities of standard Transformers, such as recognizing all regular languages [45,46]. For our task, this mutable state is particularly advantageous, as it allows the model to continuously update its "understanding" of the agent's fluctuating local environment, a crucial feature for capturing the complex, history-dependent dynamics near a phase transition.

In this work, we employed a specific RWKV-7 model with $L_{rwkv} = 3$ layers, each with a hidden state dimensionality of $D_{rwkv} = 128$. For each density regime ($\rho = 0.5$ and $\rho = 2.0$), a dedicated model was trained on a composite dataset constructed by pooling all $N_{runs} \times N_{\eta}$ single-agent sequences generated from the L=32 simulations for that specific density. Importantly, the model training was conducted blindly, without providing the specific noise value η associated with any given training sequence.

The models were trained using an autoregressive objective: at each time step t_k, predicting the probability distribution of the next angle $\theta_{discrete}(t_{k+1})$ given the entire preceding sequence $(\theta_{discrete}(t_1), \ldots, \theta_{discrete}(t_k))$. Training employed the AdamW optimizer [48] with parameters ($\beta_1 = 0.9, \beta_2 = 0.99, \epsilon = 10^{-18}$). A cosine learning rate schedule [49] was used, decaying from an initial rate of 6×10^{-4} to a final rate of 6×10^{-5} over the training duration, preceded by 10 warmup steps. The models were trained for 3 epochs with a batch size of 16.

2.4. Quantifying predictability fluctuation: Predictive entropy variance

Once trained, the unified RWKV-7 model was employed in inference mode to analyze the dynamics encoded in individual sequences. For each preprocessed sequence $\{\theta_{discrete}(t_k)\}_{k=1}^{T_{seq}}$, corresponding to a specific simulation run and noise level η (although η remained unknown to the model during this process), we performed the following analysis. At each time step k from 1 to $T_{seq} - 1$, the model processed the history $(\theta_{discrete}(t_1), \ldots, \theta_{discrete}(t_k))$ to compute the predictive probability distribution $P_k = \{p_{k,i}\}_{i=0}^{N_{bins}-1}$ over the possible subsequent angles $\theta_{discrete}(t_{k+1})$. Subsequently, the Shannon entropy H_k of this distribution was calculated using Eq. (5):

$$H_k = -\sum_{i=0}^{N_{bins}-1} p_{k,i} \log_2 p_{k,i}, \tag{5}$$

where the term is zero if $p_{k,i} = 0$. This value H_k quantifies the instantaneous uncertainty or unpredictability of the model's forecast at step k. To capture the fluctuation in this predictability over the entire trajectory, we computed the variance of the resulting entropy sequence $\{H_k\}_{k=1}^{T_{seq}-1}$ according to Eq. (6):

$$\text{Var}(H) = \frac{1}{T_{seq} - 2} \sum_{k=1}^{T_{seq}-1} (H_k - \bar{H})^2, \quad \text{where } \bar{H} = \frac{1}{T_{seq} - 1} \sum_{k=1}^{T_{seq}-1} H_k. \tag{6}$$

This variance, $\text{Var}(H)$, measures the temporal variability of the predictive uncertainty for a single trajectory.

The primary metric utilized in this study to characterize the system's behavior as a function of noise is the average entropy variance, $\overline{\text{Var}(H)}(\eta)$. This was obtained by averaging the individual $\text{Var}(H)$ values over the $N_{runs} = 20$ independent simulation runs conducted for each noise level η, as defined in Eq. (7):

$$\overline{\text{Var}(H)}(\eta) = \frac{1}{N_{runs}} \sum_{\text{run}=1}^{N_{runs}} \text{Var}(H)_{\text{run},\eta}. \tag{7}$$

We hypothesize that this average entropy variance, $\overline{\text{Var}(H)}(\eta)$, exhibits distinct behavior, particularly a peak, in the vicinity of the critical noise region, thereby serving as an indicator of criticality.

2.5. Benchmark metrics and validation

To validate our findings and evaluate the performance of our method, we compared the $\overline{\mathrm{Var}(H)}$ indicator against two distinct benchmarks.

First, to provide the physical context of the collective phase transition, we computed the global order parameter, the mean polarization $\langle \Phi \rangle$. This allows for a direct comparison between the signal detected by our single-agent indicator and the macroscopic state of the entire system.

Second, to assess the added value of our predictive modeling approach, we introduced a benchmark based on a direct statistical analysis of the single-agent time series. Specifically, we computed the circular variance of the angular increments for each trajectory. This provides a baseline to determine if the complex dynamics captured by the sequence model are necessary to identify the critical region, or if a simpler measure of angular dispersion would suffice.

Circular variance. The circular variance is a measure of the dispersion of angular data, analogous to the linear variance but adapted for periodic quantities. For a set of angular increments $\{\Delta\theta_k\}_{k=1}^{T_{seq}-1}$ from a single trajectory, we first compute the mean resultant vector $\bar{R}$:

$$\bar{R} = \frac{1}{T_{seq}-1} \sum_{k=1}^{T_{seq}-1} e^{i\Delta\theta_k}, \tag{8}$$

where the angles are in radians. The length of this vector, $|\bar{R}|$, quantifies the concentration of the data around the mean angle. The circular variance, V, is then defined as:

$$V = 1 - |\bar{R}|. \tag{9}$$

The value of V ranges from 0 (all data points are identical) to 1 (data are uniformly distributed on the circle). The final benchmark metric for each noise level η was then obtained by averaging the individual V values over the $N_{runs} = 20$ independent simulation runs conducted for that specific η. This allows for a direct comparison of the information captured by our predictive model versus a simpler measure of angular dispersion.

3. Results

In this section, we present a comprehensive evaluation of our proposed indicator, the average predictive entropy variance $\overline{\mathrm{Var}(H)}$, by analyzing its behavior across different system sizes and particle densities. We validate its effectiveness against the global order parameter and demonstrate its superiority over a simpler statistical benchmark.

3.1. Validation and superiority of the indicator at low density ($\rho = 0.5$)

We first assess our method in the low-density regime ($\rho = 0.5$). Fig. 1 provides a multi-faceted analysis for system sizes L=32 and L=64. The left panels (a, c) serve as a primary validation, comparing $\overline{\mathrm{Var}(H)}$ with the macroscopic state of the system, represented by the average polarization $\langle \Phi \rangle$. For both system sizes, a distinct and sharp peak in $\overline{\mathrm{Var}(H)}$ emerges. Crucially, this peak is precisely located within the region where the order parameter $\langle \Phi \rangle$ exhibits its steepest decline, confirming that our single-agent indicator accurately captures the collective phase transition.

To demonstrate the necessity of our predictive modeling approach, we then compare our indicator against a statistically-motivated benchmark: the circular variance of the angular increments, derived from the same single-agent trajectories. The right panels (b, d) of Fig. 1 illustrate this comparison. In stark contrast to $\overline{\mathrm{Var}(H)}$, the circular variance is a featureless, monotonically increasing function of the noise parameter η. It fails to provide any signal of the underlying phase transition. This result strongly suggests that simple statistical measures of trajectory fluctuations are insufficient to detect criticality. Instead, the sophisticated temporal pattern recognition and predictive capabilities of the sequence model are essential to distill the critical signature from a single agent's motion.

3.2. Universality and generalization in the high-density regime ($\rho = 2.0$)

To test the universality of our methodology, we applied it to the high-density regime ($\rho = 2.0$), which exhibits a qualitatively different transition behavior. A new RWKV-7 model was trained specifically on L=32, $\rho = 2.0$ data to specialize in this dynamical context. The results are presented in Fig. 2.

The left panels (a, c) demonstrate the method's universality. The $\overline{\mathrm{Var}(H)}$ indicator, calculated using the new, specialized model, again exhibits a clear and sharp peak. This peak precisely coincides with the abrupt drop in the order parameter $\langle \Phi \rangle$ that characterizes the transition in this regime, for both L=32 and L=64. This confirms that our methodology is robust across different dynamical regimes.

Furthermore, we conducted a stringent test of the model's cross-domain generalization capability. We used the original model, trained *only* on low-density ($\rho = 0.5$) data, to analyze the high-density ($\rho = 2.0$) trajectories. The results, shown in the right panels (b, d), are striking. Even without exposure to high-density dynamics during training, the original model still correctly identifies the critical region, with its $\overline{\mathrm{Var}(H)}$ peak aligning with the order parameter's drop. This remarkable generalization suggests that the sequence model is not merely memorizing patterns specific to one type of transition. Instead, it appears to have learned a more abstract feature of the order–disorder transition, one that is shared between the distinct dynamical regimes of $\rho = 0.5$ and $\rho = 2.0$. The full extent of this transferability remains a subject for future investigation.

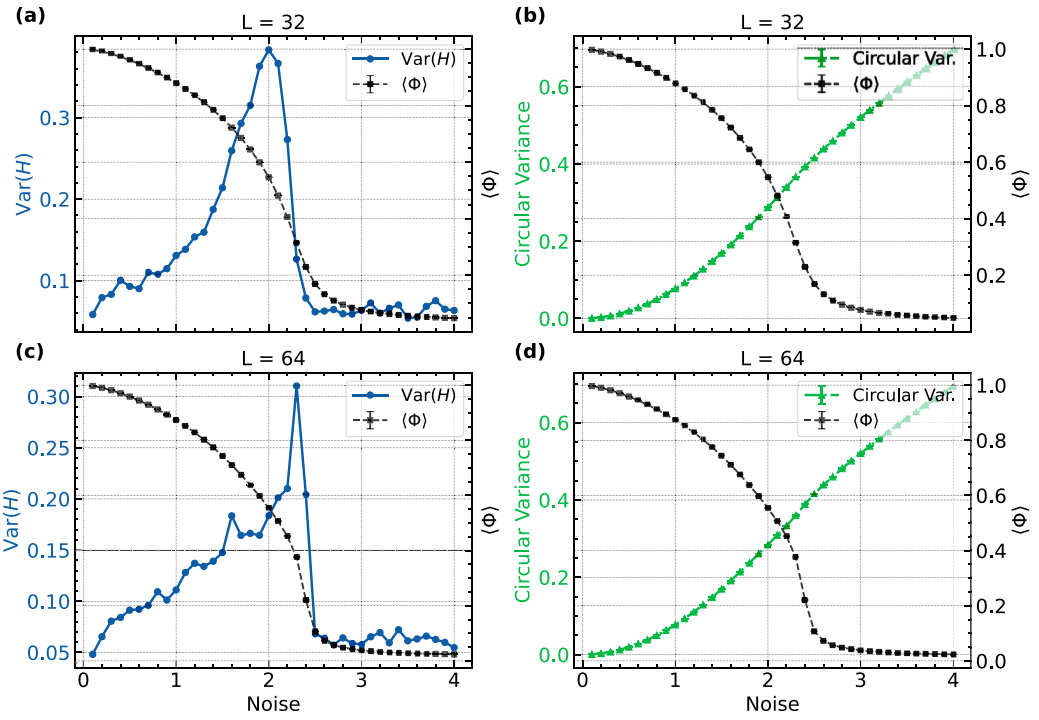

Fig. 1. Validation and superiority of the predictive entropy variance metric ($\rho = 0.5$). Comparison of our proposed indicator, $\overline{\mathrm{Var}(H)}$, with standard metrics for the low-density regime. **(a, c)** Validation against the global order parameter. The peak of $\overline{\mathrm{Var}(H)}$ (blue, left axis) aligns with the descent region of the mean polarization $\langle \Phi \rangle$ (black, right axis) for both L=32 and L=64. **(b, d)** Superiority over a simple statistical benchmark. The circular variance of angular increments (green, left axis) fails to show any distinct feature at the transition, instead increasing monotonically with noise, while the order parameter (black, right axis) shows the transition. This highlights the necessity of our predictive modeling approach.

3.3. Finite-size scaling behavior

Finally, to investigate whether our indicator behaves as a robust physical observable, we examined its finite-size scaling (FSS) properties in the low-density regime ($\rho = 0.5$). Fig. 3 presents the results across four system sizes, from L=32 to L=256.

Panel (a) displays the behavior of $\overline{\mathrm{Var}(H)}$. As the system size L increases, the peak in $\overline{\mathrm{Var}(H)}$ becomes progressively sharper, and its position systematically shifts to higher noise values (from $\eta \approx 2.0$ for L=32 to $\eta \approx 2.5$ for L=256). For comparison, panel (b) shows the behavior of the order parameter $\langle \Phi \rangle$, which exhibits the well-known sharpening of the transition curve with increasing system size. This observed correspondence in scaling behavior between our single-agent indicator and the global order parameter supports the physical relevance of our proposed metric.

4. Discussion

The results presented herein demonstrate the efficacy of utilizing the variance of predictive Shannon entropy, derived from a sequence model processing single-agent trajectories, as a potent indicator for identifying the critical region in the Vicsek model. This methodology achieves a key objective: detecting proximity to a phase transition *without prior knowledge* of the underlying control parameter (η) or the system's global state.

4.1. Entropy variance as a signature of critical dynamics

The central observation is the pronounced peak in the average entropy variance $\overline{\mathrm{Var}(H)}$ near the established critical noise level ($\eta \approx 2.0$ for $L = 32$). We interpret this peak as reflecting the heightened dynamical complexity and predictability fluctuations inherent to critical phenomena. In the ordered (low η) and disordered (high η) phases, the system dynamics, while drastically different, are relatively stable. In the ordered phase, particle movement is largely deterministic (following neighbors), leading to consistently low predictive entropy. In the disordered phase, movement is highly stochastic but statistically uniform, resulting in consistently high predictive entropy. In both stable phases, the *variance* of this entropy along a trajectory remains low.

Conversely, near the critical point, the system exhibits characteristics such as large-scale correlations, critical slowing down, and potentially intermittent switching between more ordered and disordered configurations (especially prominent in finite systems or near coexistence regions, as might be relevant for $\rho = 0.5$ [11,50]). A single agent experiences these fluctuating collective dynamics

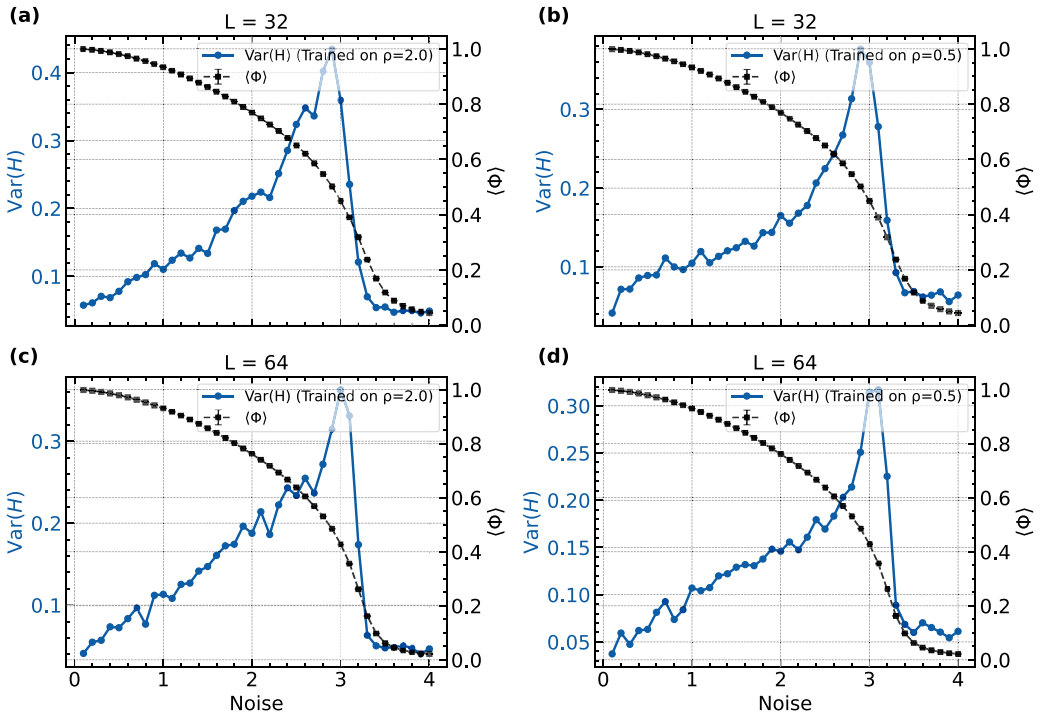

Fig. 2. Performance of the predictive entropy variance in the high-density regime ($\rho = 2.0$). The behavior of the $\overline{\mathrm{Var}(H)}$ indicator is shown for a different physical regime. **(a, c)** Results using a model trained on $\rho = 2.0$ data. For both L=32 and L=64, the peak of $\overline{\mathrm{Var}(H)}$ (blue, left axis) aligns with the region where the order parameter $\langle \Phi \rangle$ (black, right axis) undergoes an abrupt drop. **(b, d)** Results using the original model trained *only* on $\rho = 0.5$ data. The peak of $\overline{\mathrm{Var}(H)}$ produced by this model also aligns with the sharp transition in $\langle \Phi \rangle$ when applied to the $\rho = 2.0$ trajectories.

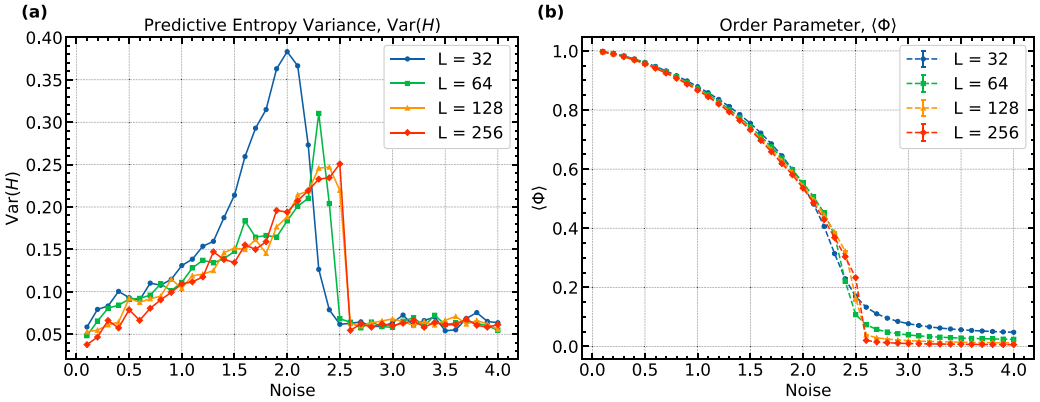

Fig. 3. Finite-size scaling behavior of the proposed indicator ($\rho = 0.5$). Comparison of the scaling properties of our indicator and the order parameter across four system sizes. **(a)** The predictive entropy variance $\overline{\mathrm{Var}(H)}$. As the system size L increases, the peak becomes sharper and shifts to higher η. **(b)** The corresponding order parameter $\langle \Phi \rangle$ curves.

through its local interactions. The RWKV-7 model, attempting to predict the agent's next move based on its history, consequently faces rapidly changing levels of predictability. Its internal "belief state" about the likely future direction fluctuates significantly over time — sometimes the local environment suggests order, sometimes disorder. This fluctuation in the model's predictive certainty manifests directly as a high variance in the output Shannon entropy sequence. Thus, $\mathrm{Var}(H)$ serves as a proxy for the temporal variability of the local dynamical predictability, which is maximized near the transition.

4.2. Inferring global state from local information

A significant implication of our findings is that substantial information about the collective state of the system, particularly its proximity to criticality, is encoded within the long-term trajectory of just a single agent. This contrasts with traditional approaches

that rely on calculating global order parameters (like polarization) requiring information from a large fraction, if not all, agents. Our method demonstrates the potential for inferring macroscopic system properties from microscopic, local observations, which is particularly valuable in scenarios where global data acquisition is infeasible or costly, a common situation in real-world complex systems (e.g., tracking a single animal in a large group, observing one stock in a market). The use of a powerful sequence model like RWKV-7 appears crucial for effectively extracting these subtle temporal patterns related to criticality from the single-agent data.

4.3. Validation against physical observables and benchmarks

A primary contribution of this work is the validation of our proposed indicator, $\overline{\mathrm{Var}(H)}$. As demonstrated in Fig. 1(a, c), for the low-density regime, the peak of our single-agent metric aligns well with the critical region identified by the global order parameter $\langle \Phi \rangle$.

Furthermore, the necessity of a sophisticated predictive model is underscored by the failure of a simpler, statistically-motivated benchmark. Fig. 1(b, d) shows that the circular variance of angular increments, while being a proper measure for angular data, is insensitive to the critical transition and fails to produce any indicative signal. This highlights that the signature of criticality is embedded in higher-order temporal patterns that can only be effectively extracted by a model trained to understand the underlying dynamics.

4.4. Universality and generalization of the method

Our investigation into the high-density regime ($\rho = 2.0$) reveals the universality and generalization capabilities of our approach (Fig. 2). First, when a new model is trained specifically on high-density data, the $\overline{\mathrm{Var}(H)}$ indicator once again successfully identifies the critical region, marked by an abrupt drop in polarization. This demonstrates the universality of our methodology across different dynamical regimes.

Second, we found that the original model trained solely on low-density data could also correctly identify the transition in the high-density system. The full extent of this transferability remains a subject for future investigation.

4.5. Consistency with finite-size scaling theory

Finally, to assess whether our indicator behaves as a robust physical observable, we examined its finite-size scaling (FSS) properties. The results for the low-density regime ($\rho = 0.5$) across four system sizes are presented in Fig. 3.

The behavior of our indicator is broadly consistent with the principles of statistical physics. As shown in panel (a), the most prominent and systematic trend is the shift of the peak location in $\overline{\mathrm{Var}(H)}$ to higher noise values as the system size L increases, moving from $\eta \approx 2.0$ for L=32 to approximately $\eta \approx 2.5$ for L=256. While the shape and height of the peak also vary with system size, the consistent rightward shift of the critical point estimate is a hallmark of finite-size scaling. This trend mirrors the behavior of the global order parameter $\langle \Phi \rangle$ shown in panel (b), where the transition curve also shifts and sharpens with increasing L. This correspondence in scaling behavior supports the physical relevance of our proposed single-agent metric.

4.6. Relation to other methods and potential advantages

Our approach offers a distinct alternative to traditional methods for detecting phase transitions. Unlike methods requiring parameter scans and global order parameter calculations, it operates “blindly” on time series data. Compared to other time-series analysis techniques sometimes used for criticality detection (e.g., Detrended Fluctuation Analysis (DFA) [51], analysis of autocorrelation functions, or simpler variance measures on position/angle increments), our method leverages the predictive power of a sophisticated sequence model. It quantifies fluctuations in the predictability of the dynamics, potentially capturing more complex temporal structures and non-linear dependencies than simpler statistical measures. The RWKV model, trained autoregressively, implicitly builds a model of the underlying dynamical process conditioned on the observed history. The entropy variance then probes the stability of this learned model’s predictions over time. A key advantage is the potential applicability to systems where the underlying equations or control parameters are unknown, relying solely on observed time series. However, direct quantitative comparisons of sensitivity, data requirements, and robustness against different noise types between our method and established time-series techniques remain an important direction for future work.

Our approach, centered on the variance of predictive entropy derived from a sequence model (RWKV-7), offers a distinct perspective compared to traditional time-series analysis techniques often employed for detecting critical transitions or characterizing complex dynamics. Standard methods frequently rely on calculating statistical properties directly from the observed time series itself. For instance, increases in variance and lag-1 autocorrelation are widely recognized indicators associated with critical slowing down near tipping points, often explored in the context of early warning signals (EWS) [29,52].

Other established techniques delve into different aspects of time series structure. Detrended Fluctuation Analysis (DFA), for example, focuses on quantifying long-range temporal correlations and their scaling behavior, which can change across different dynamical regimes [51,53]. Alternatively, methods rooted in information theory, such as Permutation Entropy [54] or Sample Entropy [55], aim to measure the complexity, regularity, or predictability inherent in the signal’s patterns.

In contrast to these approaches that directly analyze statistical moments, correlation structures, or complexity patterns within the raw data, our methodology quantifies the fluctuations in the *predictability* of the dynamics as perceived by a trained sequence

model. The model is trained autoregressively to learn the underlying process generating the single-agent trajectory across different (unknown) noise levels. The resulting indicator, $\overline{\mathrm{Var}(H)}$, captures the temporal variability of the model's certainty (inverse entropy) about the agent's immediate future state. We hypothesize that this focus on the *dynamics of predictability* makes our method particularly sensitive to the critical region, where an agent experiences rapidly changing local environments fluctuating between more ordered and disordered configurations, leading to significant variations in short-term predictability.

A key potential advantage stems from leveraging a powerful sequence model capable of capturing potentially complex, non-linear, and long-range temporal dependencies that might be missed by simpler statistical measures. The model implicitly builds a representation of the dynamics conditioned on the observed history, and the entropy variance probes the stability and consistency of this learned representation over time along a trajectory. Furthermore, as demonstrated, this approach operates "blindly" using only single-agent data, without requiring knowledge of the control parameter (η) or access to global system information, addressing key limitations of traditional phase transition analysis in potentially data-scarce or parameter-unknown real-world scenarios.

Nevertheless, direct quantitative comparisons concerning the sensitivity, data requirements (e.g., minimum time series length, sampling frequency), computational cost, and robustness against observational noise or different system types between our predictive entropy variance method and established techniques like those cited above [52–54] are warranted. Such comparative studies represent an important avenue for future work to precisely delineate the relative strengths and weaknesses and the optimal application domains for each approach in the analysis of complex systems near criticality.

It is also worth noting a key aspect of our modeling choice: we did not impose any symmetry constraints on the model's predictive distributions. While one might intuitively assume that the prediction for the next angle should be centered around the current angle of motion, the underlying dynamics of the Vicsek model are not necessarily symmetric. The alignment interaction introduces a form of temporal persistence or "momentum", meaning the particle is often more likely to continue turning in the same direction than to reverse its turn. An unconstrained sequence model like RWKV-7 is free to learn these inherent asymmetries from the data. Imposing a symmetry constraint would have forced the model to ignore this crucial information, likely degrading its ability to accurately model the trajectory and, consequently, to detect the subtle fluctuations in predictability that signal criticality.

4.7. Limitations and future directions

Despite the promising results presented, several limitations warrant discussion, alongside fruitful avenues for future research. Firstly, the findings may exhibit some dependence on the specific sequence model employed. While the RWKV-7 architecture proved effective here, further investigation using alternative models, such as LSTMs or Transformers, would be valuable to determine if the observed entropy variance peak is a general property of predictive models applied to critical dynamics or specific to certain architectures. Closely related is the sensitivity to model hyperparameters (e.g., number of layers, hidden dimensions) and the specifics of the training regime, which merit systematic exploration. Secondly, the data requirements for robust detection need further characterization. The optimal time series length (T_{seq}) and sampling frequency (Δt_{sample}) likely depend on the system's intrinsic timescales, particularly near the critical point due to critical slowing down. Similarly, the influence of the chosen preprocessing steps, such as the EMA smoothing window (W_{EMA}) and the angle discretization level (N_{bins}), should be systematically assessed for their impact on the results.

Furthermore, the scope of our method's applicability warrants broader validation. While this work has demonstrated success for the stochastic Vicsek model across two density regimes, its performance on other collective motion paradigms remains to be explored. This includes models with different interaction rules, such as those incorporating explicit attraction and repulsion forces. More fundamentally, an important future direction is to test the method on deterministic chaotic systems, in contrast to the stochastic systems studied here. For instance, agent-based models that are deterministic yet exhibit complex collective behavior and phase transitions (such as certain variants of the Couzin model [56] or the three-zone model) would provide a stringent test case. It is an open question whether the fluctuation in predictability, as captured by $\overline{\mathrm{Var}(H)}$, can serve as a universal indicator of criticality in systems where the unpredictability stems from deterministic chaos rather than explicit stochastic noise. A crucial next step also involves applying the methodology to real-world empirical data from systems like animal groups or pedestrian flows, which present additional challenges such as measurement noise and non-stationarity. The computational cost associated with training deep sequence models, although inference is generally efficient, must also be considered as a practical factor, especially when compared to computationally cheaper traditional statistical measures.

Finally, deeper theoretical understanding and more rigorous quantitative analysis represent important future directions. While we offer a plausible interpretation centered on predictability fluctuations, investigating the internal representations learned by the RWKV model, potentially using techniques from explainable AI [57], could provide more concrete insights into how the model discerns critical dynamics. Additionally, a detailed quantitative study of the scaling properties of the entropy variance peak – examining how its height, width, and location scale with system size L – could establish more rigorous connections to critical exponents and universality classes, further solidifying the link between this data-driven indicator and fundamental concepts in statistical physics. Addressing these points will be crucial for establishing the broader applicability and theoretical underpinnings of using predictive entropy variance for criticality detection.

5. Conclusion

In this study, we have introduced and validated a novel, data-driven methodology for the blind identification of critical transitions in collective motion systems. By applying a sequence model (RWKV-7) to analyze single-agent trajectories, we have shown that the variance of the model's predictive Shannon entropy, $\overline{\mathrm{Var}(H)}$, serves as a robust and effective indicator of criticality. Our approach circumvents the need for prior knowledge of control parameters or access to global system information, two significant challenges in the study of many real-world complex systems.

Our principal findings are threefold. First, we demonstrated the effectiveness of our indicator in the low-density Vicsek model. The peak of $\overline{\mathrm{Var}(H)}$ accurately pinpoints the critical region identified by the global order parameter, whereas a benchmark based on circular variance fails to do so. This result highlights the necessity of a predictive modeling approach to extract the subtle statistical signatures of criticality.

Second, we established the universality and cross-regime generalization of our method. The indicator successfully identifies the critical point not only in the low-density regime but also in the high-density regime. Furthermore, a model trained exclusively on low-density data was able to generalize and correctly locate the transition in the high-density system, suggesting that the model captures a universal feature of the order–disorder transition.

Finally, we confirmed the physical consistency and cross-size generalization of our indicator. A single model, trained only on data from the smallest system (L=32), successfully revealed the finite-size scaling of the critical point across a range of larger system sizes up to L=256. The indicator's peak exhibits systematic shifts and sharpening consistent with the behavior of established physical observables, establishing it as a robust, data-driven physical quantity that generalizes well across scales.

In conclusion, this work demonstrates that significant information about a system's collective state is encoded in the temporal dynamics of a single constituent, and that this information can be effectively decoded using modern sequence models. Our proposed indicator, the predictive entropy variance, offers a practical tool for analyzing complex systems where traditional analysis methods are challenging to apply.

CRediT authorship contribution statement

Tianyi Wu: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Zhangang Han:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Zhangang Han reports financial support was provided by National Natural Science Foundation of China. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62176022).

Data availability

Data will be made available on request.

References

[1] T. Vicsek, A. Zafeiris, Collective motion, Phys. Rep. 517 (3) (2012) 71–140, http://dx.doi.org/10.1016/j.physrep.2012.03.004.
[2] D.J.T. Sumpter, Collective Animal Behavior, Princeton University Press, Princeton, 2011, http://dx.doi.org/10.1515/9781400837106.
[3] C. Bechinger, R. Di Leonardo, H. Löwen, C. Reichhardt, G. Volpe, G. Volpe, Active particles in complex and crowded environments, Rev. Modern Phys. 88 (2016) 045006, http://dx.doi.org/10.1103/RevModPhys.88.045006.
[4] M.E. Cates, J. Tailleur, Motility-induced phase separation, Annu. Rev. Condens. Matter Phys. 6 (Volume 6 2015) (2015) 219–244, http://dx.doi.org/10.1146/annurev-conmatphys-031214-014710.
[5] P. Friedl, D. Gilmour, Collective cell migration in morphogenesis, regeneration and cancer, Nature Rev. Mol. Cell Biol. 10 (7) (2009) 445–457, http://dx.doi.org/10.1038/nrm2720.
[6] M. Brambilla, E. Ferrante, M. Birattari, M. Dorigo, Swarm robotics: A review from the swarm engineering perspective, Swarm Intell. 7 (1) (2013) 1–41, http://dx.doi.org/10.1007/s11721-012-0075-2.
[7] C. Castellano, S. Fortunato, V. Loreto, Statistical physics of social dynamics, Rev. Modern Phys. 81 (2009) 591–646, http://dx.doi.org/10.1103/RevModPhys.81.591.
[8] T. Vicsek, A. Czirók, E. Ben-Jacob, I. Cohen, O. Shochet, Novel type of phase transition in a system of self-driven particles, Phys. Rev. Lett. 75 (1995) 1226–1229, http://dx.doi.org/10.1103/PhysRevLett.75.1226.
[9] J. Toner, Y. Tu, Long-range order in a two-dimensional dynamical XY model: How birds fly together, Phys. Rev. Lett. 75 (1995) 4326–4329, http://dx.doi.org/10.1103/PhysRevLett.75.4326.

[10] J. Toner, Y. Tu, Flocks, herds, and schools: A quantitative theory of flocking, Phys. Rev. E 58 (1998) 4828–4858, http://dx.doi.org/10.1103/PhysRevE.58.4828.
[11] H. Chaté, F. Ginelli, G. Grégoire, F. Raynaud, Collective motion of self-propelled particles interacting without cohesion, Phys. Rev. E 77 (2008) 046113, http://dx.doi.org/10.1103/PhysRevE.77.046113.
[12] G. Grégoire, H. Chaté, Onset of collective and cohesive motion, Phys. Rev. Lett. 92 (2004) 025702, http://dx.doi.org/10.1103/PhysRevLett.92.025702.
[13] A.P. Solon, H. Chaté, J. Tailleur, From phase to microphase separation in flocking models: The essential role of nonequilibrium fluctuations, Phys. Rev. Lett. 114 (2015) 068101, http://dx.doi.org/10.1103/PhysRevLett.114.068101.
[14] M.C. Marchetti, J.F. Joanny, S. Ramaswamy, T.B. Liverpool, J. Prost, M. Rao, R.A. Simha, Hydrodynamics of soft active matter, Rev. Modern Phys. 85 (2013) 1143–1189, http://dx.doi.org/10.1103/RevModPhys.85.1143.
[15] S. Ramaswamy, The mechanics and statistics of active matter, Annu. Rev. Condens. Matter Phys. 1 (Volume 1 2010) (2010) 323–345, http://dx.doi.org/10.1146/annurev-conmatphys-070909-104101.
[16] G. Baglietto, E.V. Albano, Finite-size scaling analysis and dynamic study of the critical behavior of a model for the collective displacement of self-driven individuals, Phys. Rev. E 78 (2008) 021125, http://dx.doi.org/10.1103/PhysRevE.78.021125.
[17] A. Cavagna, P. Chaikin, D. Levine, S. Martiniani, A. Puglisi, M. Viale, Vicsek model by time-interlaced compression: A dynamical computable information density, Phys. Rev. E. 103 (2021) 062141.
[18] B. Sorkin, A. Be'er, H. Diamant, G. Ariel, Detecting and characterizing phase transitions in active matter using entropy, Soft Matter. 19 (2023) 5118–5126.
[19] J. Gunawardena, Models in biology: 'accurate descriptions of our pathetic thinking', BMC Biol. 12 (1) (2014) 29, http://dx.doi.org/10.1186/1741-7007-12-29.
[20] Y. Katz, K. Tunstrøm, C.C. Ioannou, C. Huepe, I.D. Couzin, Inferring the structure and dynamics of interactions in schooling fish, Proc. Natl. Acad. Sci. 108 (46) (2011) 18720–18725, http://dx.doi.org/10.1073/pnas.1107583108.
[21] M. Nagy, Z. Ákos, D. Biro, T. Vicsek, Hierarchical group dynamics in pigeon flocks, Nature 464 (7290) (2010) 890–893, http://dx.doi.org/10.1038/nature08891.
[22] A. Strandburg-Peshkin, D.R. Farine, I.D. Couzin, M.C. Crofoot, Shared decision-making drives collective movement in wild baboons, Science 348 (6241) (2015) 1358–1361, http://dx.doi.org/10.1126/science.aaa5099.
[23] A. Cavagna, I. Giardina, Bird flocks as condensed matter, Annu. Rev. Condens. Matter Phys. 5 (2014) 183–207, http://dx.doi.org/10.1146/annurev-conmatphys-031113-133834.
[24] L.F. Hughey, A.M. Hein, A. Strandburg-Peshkin, F.H. Jensen, Challenges and solutions for studying collective animal behaviour in the wild, Phil. Trans. R. Soc. B 373 (1746) (2018) 20170005, http://dx.doi.org/10.1098/rstb.2017.0005.
[25] J.E. Herbert-Read, A. Perna, R.P. Mann, T.M. Schaerf, D.J.T. Sumpter, A.J.W. Ward, Inferring the rules of interaction of shoaling fish, Proc. Natl. Acad. Sci. 108 (46) (2011) 18726–18731, http://dx.doi.org/10.1073/pnas.1109355108.
[26] R.P. Mann, Bayesian inference for identifying interaction rules in moving animal groups, PLoS One 6 (8) (2011) 1–10, http://dx.doi.org/10.1371/journal.pone.0022827.
[27] T. Yabe, P.S.C. Rao, S.V. Ukkusuri, S.L. Cutter, Toward data-driven, dynamical complex systems approaches to disaster resilience, Proc. Natl. Acad. Sci. 119 (8) (2022) e2111997119, http://dx.doi.org/10.1073/pnas.2111997119.
[28] I.G. Kevrekidis, G. Samaey, Equation-free multiscale computation: Algorithms and applications, Annu. Rev. Phys. Chem. 60 (Volume 60 2009) (2009) 321–344, http://dx.doi.org/10.1146/annurev.physchem.59.032607.093610.
[29] M. Scheffer, J. Bascompte, W.A. Brock, V. Brovkin, S.R. Carpenter, V. Dakos, H. Held, E.H. van Nes, M. Rietkerk, G. Sugihara, Early-warning signals for critical transitions, Nature 461 (7260) (2009) 53–59, http://dx.doi.org/10.1038/nature08227.
[30] J. Carrasquilla, Machine learning for quantum matter, Adv. Phys.: X 5 (1) (2020) 1797528, http://dx.doi.org/10.1080/23746149.2020.1797528.
[31] P. Mehta, M. Bukov, C.-H. Wang, A.G. Day, C. Richardson, C.K. Fisher, D.J. Schwab, A high-bias, low-variance introduction to machine learning for physicists, Phys. Rep. 810 (2019) 1–124, http://dx.doi.org/10.1016/j.physrep.2019.03.001, a high-bias, low-variance introduction to Machine Learning for physicists.
[32] J. Carrasquilla, R.G. Melko, Machine learning phases of matter, Nat. Phys. 13 (5) (2017) 431–434, http://dx.doi.org/10.1038/nphys4035.
[33] W. Rządkowski, N. Defenu, S. Chiacchiera, A. Trombettoni, G. Bighin, Detecting composite orders in layered models via machine learning, New J. Phys. 22 (9) (2020) 093026, http://dx.doi.org/10.1088/1367-2630/abae44.
[34] T. Xue, X. Li, X. Chen, L. Chen, Z. Han, Machine learning phases in swarming systems, Mach. Learn.: Sci. Technol. 4 (1) (2023) 015028, http://dx.doi.org/10.1088/2632-2153/acc007.
[35] L. Wang, Discovering phase transitions with unsupervised learning, Phys. Rev. B 94 (2016) 195105, http://dx.doi.org/10.1103/PhysRevB.94.195105.
[36] S.J. Wetzel, Unsupervised learning of phase transitions: From principal component analysis to variational autoencoders, Phys. Rev. E 96 (2017) 022140, http://dx.doi.org/10.1103/PhysRevE.96.022140.
[37] P. Ponte, R.G. Melko, Kernel methods for interpretable machine learning of order parameters, Phys. Rev. B 96 (2017) 205146, http://dx.doi.org/10.1103/PhysRevB.96.205146.
[38] P. De Lellis, M. Porfiri, Inferring the size of a collective of self-propelled vicsek particles from the random motion of a single unit, Commun. Phys. 5 (2022) 86.
[39] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780, http://dx.doi.org/10.1162/neco.1997.9.8.1735.
[40] Y. Bengio, P. Simard, P. Frasconi, Learning long-term dependencies with gradient descent is difficult, IEEE Trans. Neural Netw. 5 (2) (1994) 157–166, http://dx.doi.org/10.1109/72.279181.
[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc., 2017.
[42] P. Bo, Blinkdl/rwkv-lm: 1.00, 2021, http://dx.doi.org/10.5281/zenodo.5196577.
[43] B. Peng, E. Alcaide, Q. Anthony, A. Albalak, S. Arcadinho, S. Biderman, H. Cao, X. Cheng, M. Chung, L. Derczynski, X. Du, M. Grella, K. Gv, X. He, H. Hou, P. Kazienko, J. Kocon, J. Kong, B. Koptyra, H. Lau, J. Lin, K.S.I. Mantri, F. Mom, A. Saito, G. Song, X. Tang, J. Wind, S. Woźniak, Z. Zhang, Q. Zhou, J. Zhu, R.-J. Zhu, RWKV: Reinventing RNNs for the transformer era, in: K. Bali H. Bouamor (Ed.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 14048–14077, http://dx.doi.org/10.18653/v1/2023.findings-emnlp.936.
[44] B. Peng, D. Goldstein, Q. G. Anthony, A. Albalak, E. Alcaide, S. Biderman, E. Cheah, T. Ferdinan, K. K. GV, H. Hou, S. Krishna, R.M. Jr., N. Muennighoff, F. Obeid, A. Saito, G. Song, H. Tu, R. Zhang, B. Zhao, Q. Zhao, J. Zhu, R.-J. Zhu, Eagle and finch: RWKV with matrix-valued states and dynamic recurrence, in: First Conference on Language Modeling, 2024.
[45] B. Peng, R. Zhang, D. Goldstein, E. Alcaide, X. Du, H. Hou, J. Lin, J. Liu, J. Lu, W. Merrill, G. Song, K. Tan, S. Utpala, N. Wilce, J.S. Wind, T. Wu, D. Wuttke, C. Zhou-Zheng, Rwkv-7 goose with expressive dynamic state evolution, 2025, arXiv:2503.14456.
[46] W. Merrill, A. Sabharwal, The parallelism tradeoff: Limitations of log-precision transformers, Trans. Assoc. Comput. Linguist. 11 (2023) 531–545, http://dx.doi.org/10.1162/tacl_a_00562, https://aclanthology.org/2023.tacl-1.31/.
[47] W. Merrill, J. Petty, A. Sabharwal, The illusion of state in state-space models, in: Proceedings of the 41st International Conference on Machine Learning, ICML'24, JMLR.org, 2024.

[48] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, la, USA, May (2019) 6-9, OpenReview.net, 2019.
[49] I. Loshchilov, F. Hutter, SGDR: stochastic gradient descent with warm restarts, in: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April (2017) 24-26, Conference Track Proceedings, OpenReview.net, 2017.
[50] F. Ginelli, H. Chaté, Relevance of metric-free interactions in flocking phenomena, Phys. Rev. Lett. 105 (2010) 168103, http://dx.doi.org/10.1103/PhysRevLett.105.168103.
[51] C.-K. Peng, S.V. Buldyrev, S. Havlin, M. Simons, H.E. Stanley, A.L. Goldberger, Mosaic organization of dna nucleotides, Phys. Rev. E 49 (1994) 1685–1689, http://dx.doi.org/10.1103/PhysRevE.49.1685.
[52] V. Dakos, S.R. Carpenter, W.A. Brock, A.M. Ellison, V. Guttal, A.R. Ives, S. Kéfi, V. Livina, D.A. Seekell, E.H. van Nes, M. Scheffer, Methods for detecting early warnings of critical transitions in time series illustrated using simulated ecological data, PLoS One 7 (7) (2012) 1–20, http://dx.doi.org/10.1371/journal.pone.0041010.
[53] J.W. Kantelhardt, E. Koscielny-Bunde, H.H. Rego, S. Havlin, A. Bunde, Detecting long-range correlations with detrended fluctuation analysis, Phys. A 295 (3) (2001) 441–454, http://dx.doi.org/10.1016/S0378-4371(01)00144-3.
[54] C. Bandt, B. Pompe, Permutation entropy: A natural complexity measure for time series, Phys. Rev. Lett. 88 (2002) 174102, http://dx.doi.org/10.1103/PhysRevLett.88.174102.
[55] J.S. Richman, J.R. Moorman, Physiological time-series analysis using approximate entropy and sample entropy, Am. J. Physiology-Heart Circ. Physiol. 278 (6) (2000) H2039–H2049, http://dx.doi.org/10.1152/ajpheart.2000.278.6.H2039.
[56] I.D. COUZIN, J. KRAUSE, R. JAMES, G.D. RUXTON, N.R. FRANKS, Collective memory and spatial sorting in animal groups, J. Theoret. Biol. 218 (1) (2002) 1–11, http://dx.doi.org/10.1006/jtbi.2002.3065.
[57] R. Roscher, B. Bohn, M.F. Duarte, J. Garcke, Explainable machine learning for scientific insights and discoveries, IEEE Access 8 (2020) 42200–42216, http://dx.doi.org/10.1109/ACCESS.2020.2976199.