Intrinsic statistical separation of subpopulations in heterogeneous collective motion via dimensionality reduction

Pei Tan[1,*] and Christopher E. Miles[2,†]

[1]*Mathematical, Computational, and Systems Biology Graduate Program, University of California, Irvine 92697, USA*
[2]*Department of Mathematics, University of California, Irvine 92697, USA*

(Received 18 May 2023; revised 3 November 2023; accepted 12 December 2023; published 19 January 2024)

Collective motion of locally interacting agents is found ubiquitously throughout nature. The inability to probe individuals has driven longstanding interest in the development of methods for inferring the underlying interactions. In the context of heterogeneous collectives, where the population consists of individuals driven by different interactions, existing approaches require some knowledge about the heterogeneities or underlying interactions. Here, we investigate the feasibility of identifying the identities in a heterogeneous collective without such prior knowledge. We numerically explore the behavior of a heterogeneous Vicsek model and find sufficiently long trajectories intrinsically cluster in a principal component analysis-based dimensionally reduced model-agnostic description of the data. We identify how heterogeneities in each parameter in the model (interaction radius, noise, population proportions) dictate this clustering. Finally, we show the generality of this phenomenon by finding similar behavior in a heterogeneous D'Orsogna model. Altogether, our results establish and quantify the intrinsic model-agnostic statistical disentanglement of identities in heterogeneous collectives.

DOI: 10.1103/PhysRevE.109.014403

I. INTRODUCTION

Systems of locally interacting agents that display spatiotemporal collective behaviors beyond the capabilities of individuals are found ubiquitously throughout the physical world at a range of scales [1,2]. Notable examples include fish schooling [3,4], birds flocking [5,6], insect [7,8] and bacterial swarming [9,10], human crowds [11], cell migration [12,13], and other subcellular processes [14,15].

Most attention has been paid towards investigating homogeneous collectives, where all agents evolve and interact via the same dynamics. However, real collectives are richly heterogeneous [16,17]. Such heterogeneities arise from bacterial length differences [18], mixed-species collectives [19], leader-follower behaviors in animals [20–23] or cell migration [24–27], and lane formation in human crowds [28]. The collective motion of heterogeneous systems has consequently been investigated extensively and found to be even richer than that of the homogeneous variety [29–34].

Alongside the studies of the emergent behavior of collectives, a parallel thread of investigations has developed and applied methods for the inverse problem of deducing the underlying interactions from trajectories [35–46]. This quest is of natural scientific interest due to the ability to observe only the correlated trajectories of the interactive collective, making disentangling individual interactions inherently challenging, especially with heterogeneities [24]. Recent advances have broken ground on the ability to infer interactions in heterogeneous collectives using clever and sophisticated approaches. However, these approaches, while powerful and elegant, seemingly share a unifying feature of requiring knowledge of the collective or its heterogeneities. For instance, methods that provide flexible nonparametric tests of heterogeneities [47], or the ability to infer the interactions [48] in heterogeneous collectives, both require knowledge of the particle identities *a priori*. The work in Ref. [49] addresses this with a mixture model fit alongside sparse identification of the interactions. While able to identify the identities, the success of this method hinges on the ability to correctly specify a library of underlying interactions. Other methods for detecting heterogeneities work well but are limited to specific contexts such as the detection of dissenting directions among neighbors [50] or only leader-follower interactions [51,52]. In this paper, we seek to address whether particle identities can be detected in heterogeneous collectives with no prior information about the collective or the structure of the heterogeneities.

To study disentangling heterogeneities in collectives, we investigate a heterogeneous variant of the classical Vicsek model [53]. This model is renowned as the textbook minimal example of a collective motion with rich behavior [54,55]. Consequently, many variants have been considered [56], including those with heterogeneities [57,58] such as the ones we propose here. We first consider a setup with two populations of Vicsek particles with different parameters (interaction radii, noise magnitude, velocity), but still interacting as a single collective. After performing dimensionality reduction on the trajectories, we find that in this latent space, the trajectories cluster into their identities for sufficiently long observations. In this paper, we quantify the parameter-dependent timescale required for accurate clustering through numerical simulation. Next, we show that this clustering phenomenon persists in a heterogeneous Vicsek model with more than two species. Lastly, to establish that this is truly a model-free phenomenon,

[*] peit3@uci.edu
[†] chris.miles@uci.edu

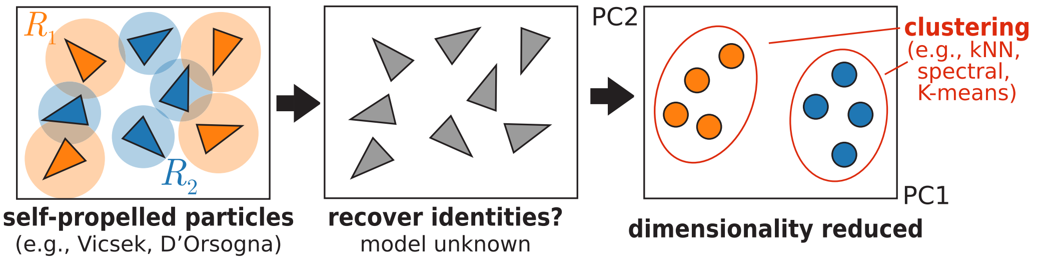

FIG. 1. Schematic of the investigation. The central question of this paper is whether (and how) the identities of particles in heterogeneous collectives (e.g., a Vicsek model with two different interaction radii for each subpopulation R_1, R_2) can be recovered from trajectory data with no model information. We find that dimensionality reduction via PCA (principal component analysis) yields distinct clustering of the subtypes over sufficiently long timescales characterized in our paper.

we consider a heterogeneous D'Orsogna model [59] and find similar clustering behavior. Altogether, our results are summarized in Fig. 1 and establish the ability to cluster heterogeneous collectives in a model-free manner with no prior knowledge of the underlying model or heterogeneities.

II. SETUP

A. Classical Vicsek

The classical Vicsek model describes the evolution of N self-propelled particles moving in two-dimensional space at a constant speed v and with fluctuating direction. The direction of each particle is governed by two factors: noise and local interactions with neighbors. Specifically, each particle averages the orientations over all neighbors within a specified radius, R. In symbols, $\theta_{i,t}$, the orientation of particle i at frame t, evolves as

$$\theta_{i,t+1} = \langle \theta_{j,t}(t) \rangle_{\|\boldsymbol{x}_{i,t} - \boldsymbol{x}_{j,t}\| < R} + \eta. \tag{1}$$

The particle positions are updated with these orientations:

$$\boldsymbol{x}_{i,t+1} = \boldsymbol{x}_{i,t} + v\Delta t \begin{pmatrix} \cos(\theta_{i,t}) \\ \sin(\theta_{i,t}) \end{pmatrix}. \tag{2}$$

The noise η is chosen from a uniform distribution governed by a scalar magnitude $0 \leqslant \sigma \leqslant 1$, such that $\eta \sim \text{unif}(-\sigma\pi, \sigma\pi)$. The particles are constrained to an $L \times L$ periodic box, where distances are computed in a manner that respects the periodicity of the domain. For systems with large N, naive $O(N^2)$ comparisons are prohibitive. We instead employ a standard KD-tree [60] $O(n \log_2 n)$ implementation for computational scalability. Particles are initialized with uniformly random orientation and position within the box. For all simulations, unless noted otherwise, $t = 1000$ steps are taken for equilibration and then discarded for analysis. This choice is discussed further in the text in Sec. III C.

B. Heterogeneous Vicsek and clustering pipeline

We consider a variant on the classical Vicsek model with $M \geqslant 2$ subpopulations. Specifically, denote $\boldsymbol{\phi} = (v, \sigma, R)$ as the parameters governing the motion of a particle in the classical Vicsek model. In the heterogeneous collective, particles belonging to subpopulation j evolve via the parameter set $\boldsymbol{\phi}_j = (v_j, \sigma_j, R_j)$. Particles interact regardless of their membership in a subpopulation. In total, the collective consists of N particles that can be decomposed into their group membership $N = \sum_{j=1}^{M} N_j$, where N_j denotes the number of particles in subpopulation j. This model has been considered in previous studies and is a more general case of some leader-follower models.

The heterogeneous Vicsek model is straightforward to simulate and generate trajectories for testing. However, performing the cluster analysis on the resulting trajectories in an unsupervised model-agnostic manner does not seem to have a clearly outlined path in the existing literature.

The first design decision we must make is the input data to the procedure. We assume that only positional information is available, and the particle identity is known frame-to-frame, allowing for the formation of trajectories. To reduce each trajectory to a scalar quantity, we consider $\theta_i(t)$, the orientations. While it may not be possible to directly access these for experimental observations, the orientations can be estimated by the frame-to-frame displacement, e.g., $\hat{\theta}_{i,t} = \mathtt{atan2}(x^y_{i,t+1} - x^y_{i,t}, x^x_{i,t+1} - x^x_{i,t})$, where $\boldsymbol{x}^{x,y}_{i,t}$ correspond to the x, y component of the positions. Naive dimensionality reduction does not preserve the structure of angular data [61], so we transform $\tau_{i,t} := \tan\theta_{i,t}$. Alternatively, we tested $\tilde{\tau}_{i,t} := [\cos\theta_{i,t}, \sin\theta_{i,t}]$, which doubles the trajectory length but may be more generalizable to 3D data, and found no difference in our results. In summary, for t observations of a collective with N particles, we consider our data to be the $N \times t$ matrix:

$$X_t = \tan(\Theta_t) = \begin{bmatrix} \tan\theta_{1,0} & \tan\theta_{1,1} & \cdots & \tan\theta_{1,t} \\ \tan\theta_{2,0} & \tan\theta_{2,1} & \cdots & \tan\theta_{2,t} \\ \vdots & \vdots & \ddots & \vdots \\ \tan\theta_{n,0} & \tan\theta_{n,1} & \cdots & \tan\theta_{n,t} \end{bmatrix}. \tag{3}$$

Equipped with this data, there are two notable branches of approaches for unsupervised clustering time series [62]. One can assign and cluster based on an appropriate metric between trajectories, such as Euclidean distance or dynamic time warping [63]. However, the choice of such a metric for collective motion data is not obvious to the authors. Therefore, we consider the second main avenue for clustering time series: dimensionality reduction. A zoo of possible linear and nonlinear approaches for dimensionality reduction of time series exists. We opt for a pragmatically simple approach of principal component analysis (PCA). While classical, it is worthwhile to note that PCA can outperform nonlinear dimensionality reductions in certain contexts [64] and has interpretability as linear transformations of the original data. There may be more complex dimensionality reduction procedures that better separate the data, but PCA would nonetheless always be the benchmark to compare the performance with and therefore serve as the basis of the remainder of this paper.

We briefly review PCA for self-containment of our approach's description. Further details can be found in Ref. [65] For the data matrix X_t in (3), PCA corresponds to a $t \times t$ weight matrix W_t, whose columns are the eigenvectors of $X_t^{\mathsf{T}} X_t$, from which a component matrix T_t can be computed by $T_t = X_t W_t$. Each column of T_t is scaled to have unit variance and zero mean. This construction corresponds to a linear change of basis to orthogonal directions that maximize variances within the data. In practice for dimensionality reduction, only the first L columns of T_t are considered, defining a linear transformation of each row of the data X_t (a particle trajectory) into an L dimensional vector of scores. Throughout the remainder of this paper, we consider $L = 2$ due to the ability to visualize the scores. However, we found little performance dropoff or gain for larger or smaller values of L, including $L = 1$ for the heterogeneous Vicsek model.

While PCA is notably useful in transforming the data to a more easily clusterable description, it is not itself, a clustering technique. Therefore, we must finally choose some approach for procedurally identifying clusters. In practice, we considered alternatives (K-nearest neighbors [66], spectral clustering [67]) but find this choice matters very little due to the intrinsic behavior of separation between the two particle populations in PCA space. Unless otherwise noted, all clustering in the remainder of the text is done using K means, which assigns C cluster identities $\mathcal{S} = \{S_1, \ldots, S_C\}$ based on the optimization of the total distance away from centroids within each cluster, $\arg\min_{\mathcal{S}} \sum_{i=1}^{C} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$. Here, $\mathbf{x}$ are L-dimensional vectors of principal component (PC) scores for each trajectory, and $\boldsymbol{\mu}_i$ are the centroids (means) computed from the cluster assignments. This optimization is done using `scikit-learn`'s standard `KMeans` function with the known number of clusters specified.

III. RESULTS

A. Two subpopulation heterogeneous Vicsek models cluster over sufficiently long times

We first demonstrate the dimensionality-reduction-based clustering on a setup with two subpopulations of particles that differ only in one attribute. Specifically, we take two types of particles, $N_1 = 200, N_2 = 200$, with $\boldsymbol{\phi}_1 = (\nu_1, \sigma_1, R_1) = (0.01, 0.1, 0.05)$ and $\boldsymbol{\phi}_2 = (\nu_2, \sigma_2, R_2) = (0.01, 0.3, 0.05)$. That is, the two particles differ only in their magnitude of noise. Other simulation parameters are set to $L = 1$, $\Delta t = 1$. The results of the simulation over increasingly long times can be seen in Fig. 2.

In Figs. 2(a)–2(c), the snapshots of the particle positions show that the collective evolves in a manner that integrates both subpopulations with no apparent pattern. The first two PC scores of each trajectory are shown in Figs. 2(d)–2(f). At early times, the scores are not separable by eye. After some time passes, the scores seem to begin to separate but not to a degree that can be fully disentangled. Finally, at long times, the PC scores of the trajectories corresponding to different types separate into two distinct clusters. Figures 2(g)–2(i) show the result of running K-means clustering on the PC scores. Initially, the clustering is inaccurate (around 50%, as expected, by random assignments of two categories) but progressively gains accuracy until eventually stabilizing at 100% as more data is accumulated on the trajectory [Fig. 2(j)].

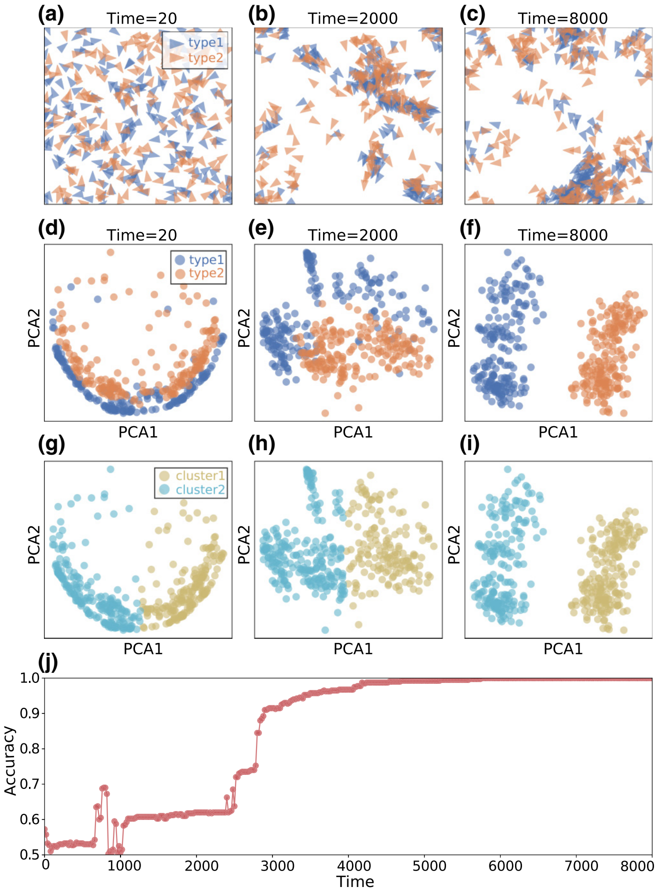

FIG. 2. Two subtype Vicsek model simulation and clustering. (a)–(c) Snapshots of the particle positions in a heterogeneous Vicsek simulation with two types of particles and displaying no apparent pattern. (d)–(f) The first two principal component scores for each trajectory, colored by particle type. (g)–(i) Results of K-means clustering on PC scores. (j) Clustering accuracy approaches 100% as the trajectories become longer. The two populations differ only in their noise magnitude $\sigma_1 = 0.1, \sigma_2 = 0.3$ and otherwise $\nu = 0.01, R = .05$, with $L = 1$, $\Delta t = 1$, and particle counts $N_1 = 200, N_2 = 200$.

B. Time to accurately cluster is dependent on which parameters are heterogeneous

The previous result shows that the PC scores in a single collective with two different noise magnitudes cluster over sufficiently long times. This leaves the natural question of what shapes the timescale for accurate clustering. Due to stochasticity, this timing will differ in each collective. We perform $N_{\text{sim}} = 100$ simulations for each parameter set to evaluate the typical time to cluster accurately for the corresponding scenario. The results of varying the heterogeneity in noise σ, the interaction radius R, and number of particles N_1, N_2, and the ratio of N_1, N_2 can be seen in Fig. 3.

In Fig. 3(a), we see the effect of differing levels of noise between the two subpopulations of particles, ranging from 2.5 to

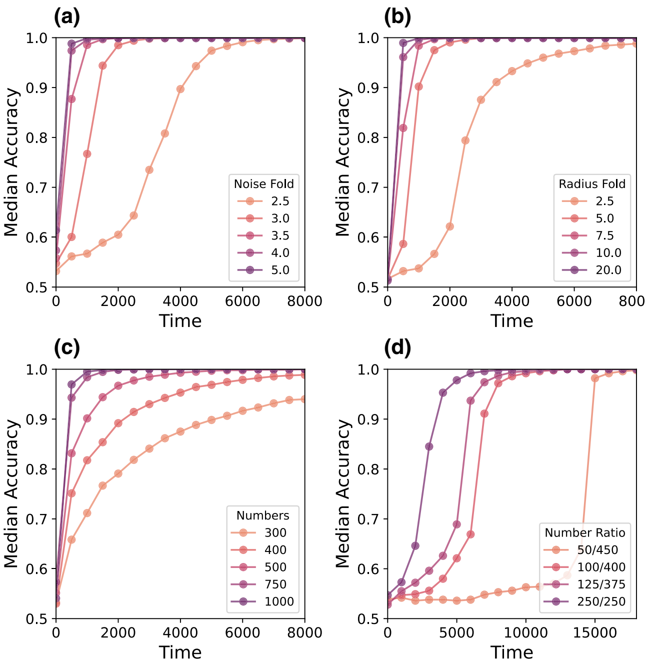

FIG. 3. Parameter influence on timescale of accurate clustering. (a): Median accuracy $N_{\text{sim}} = 100$ of clustering for a two sub-species heterogenous Vicsek model with only noise magnitude different. “Noise fold” refers to the ratio of σ_2/σ_1. (b): Median accuracy clustering two sub-species with only interaction radii different (c): Median accuracy clustering with the ratio $N_1/N_2 = 1$ fixed but the total number of particles $N_1 + N_2 = N$ is increased. (d): Median accuracy clustering with the ratio $N_1 + N_2 = N$ fixed but ratio of two groups is varied.

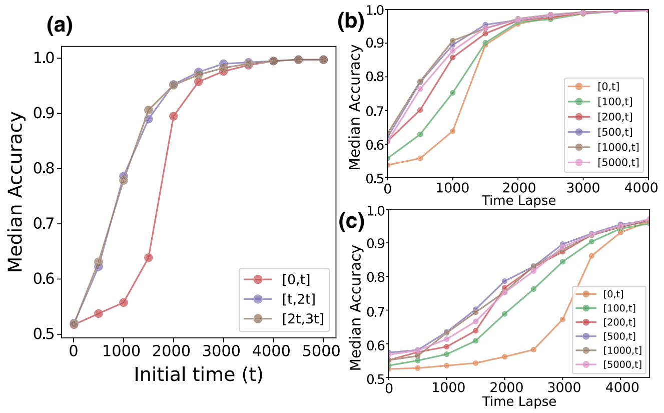

FIG. 4. Clustering timescale dependence transient effects. All simulations have the same parameters as Fig. 2 but retain the initial transient timesteps discarded in all other figures. (a) Median clustering accuracy ($N_{\text{sim}} = 100$) for sliding windows of time $[0, t]$ (red, lowest), $[t, 2t]$ (purple, middle), and $[2t, 3t]$ (purple, top), showing reduced accuracy for trajectories with transient behavior. (b) Cluster accuracy for increasing choices of cutoffs for discarding transient effects. The bottommost curve, labeled $[0, t]$, corresponds to no data discarded. Values around $t > 500$ converge, supporting the choice of $t = 1000$. (c) Same as the previous panel, except with $\sigma_1 = 0.1$, $\sigma_2 = 0.25$, a more challenging clustering [$\sigma_2 = 0.3$ in (b)]. Accurate clustering times are longer, but curves for cutoffs $t > 200$ appear converged.

5.0 “noise fold”, meaning the ratio of σ_2/σ_1. Intuitively, as the populations become more distinct, the ability to distinguish them becomes easier, manifesting as a smaller timescale until all simulations reach 100% accuracy. In Fig. 3(b), a similar effect can be seen for differing only the interaction radii. However, we note that the time for clustering with differing radii takes far longer than clustering noise differences. Next, we investigate the role of particle density by fixing the ratio of N_1 to N_2 in the noise test of the first panels. We then increase the total number of particles $N = N_1 + N_2$ and investigate the time to cluster accurately, finding that the time to cluster decreases with N, as seen in Fig. 3(c). Lastly, we fix N and vary the ratio of the two subtypes, seen in Fig. 3(d). Here, we find that greater asymmetry produces longer accurate clustering time. In sum, we find that (i) the more heterogeneous (in parameter values) the subpopulations, (ii) higher density, and (iii) lower asymmetry in numbers all decrease the critical timescale for clustering accurately.

C. Clustering time is intrinsic and can be disentangled from transient behavior

In the investigation thus far, we have established the intuitive fact that longer trajectories yield higher accuracy in clustering subpopulations. Moreover, the timescale for this accurate clustering is an intrinsic property determined by the parameters of the system. However, it remains unclear whether this emergent timescale is related to transient effects in the system or corresponds to observations equilibrium. For all simulations unless noted otherwise, we discard the first $t = 1000$ time steps in hopes of truly quantifying the equilibrium behavior, but in this section we discuss and investigate this choice. For the heterogeneous two-subpopulation Vicsek model investigated in Fig. 3, we now retain the initial time steps and denote $t = 0$ the initialization with random particle positions and orientations. Then, we investigate sliding windows of time of the trajectories of the same length but at different timepoints. In Fig. 4(a), three curves correspond to clustering accuracy for trajectories limited to $[0, t]$, with no transient effects removed, $[t, 2t]$ and $[2t, 3t]$. The curve with transient effects is notably distinct from those with initial portions discarded and has far slower clustering time, with differences occurring on the timescale of $t = 1000$. To investigate whether $t = 1000$ is an appropriate threshold for cutoff to discard transient time steps, we vary this threshold and compare the median accuracy for each. In Fig. 4(b), we use the same values as previous figures, including $\sigma_1 = 0.1$ and $\sigma_2 = 0.3$, and find that curves with transient times $t > 500$ converge onto each other. In Fig. 4(c), we further investigate the choice of $t = 1000$ cutoff for a harder clustering task $\sigma_1 = 0.1$ and $\sigma_2 = 0.25$. While clustering times are broadly longer, all curves for cutoffs $t > 200$ again converge onto each other. Altogether, these findings suggest that clustering time is indeed a system-specific emergent quantity even at equilibrium rather than an artifact of transient behavior.

D. More than two subpopulations can be clustered

The previous examples explore a heterogeneous collective with only two subpopulations. However, the dimensionality

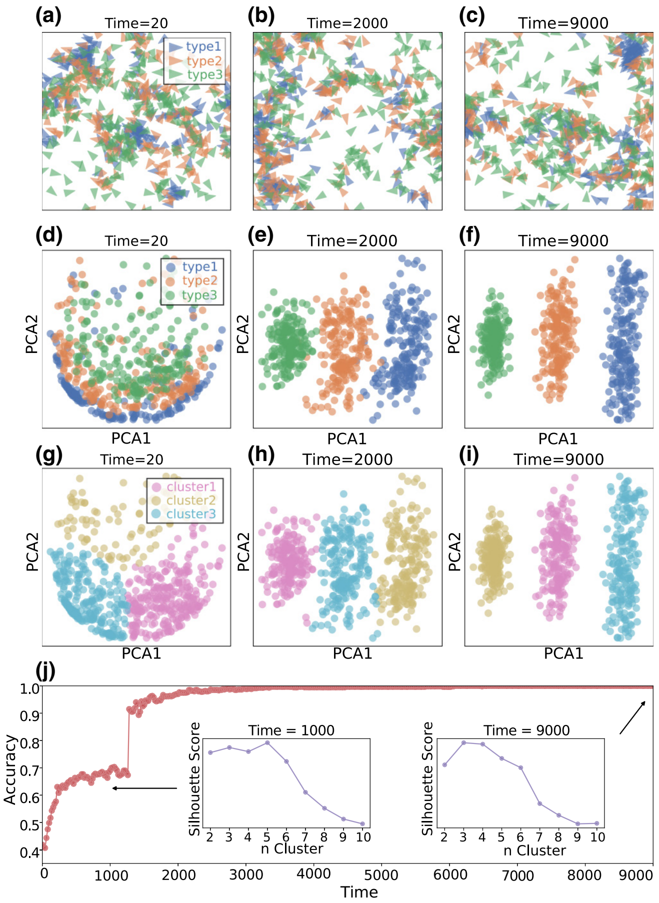

FIG. 5. Three subtype Vicsek model simulation and clustering. (a)–(c) Snapshots of the particle positions in a heterogeneous Vicsek simulation with three types of particles and displaying no apparent pattern. (d)–(f) The first two principal component scores for each trajectory, colored by particle type. (g)–(i) Results of k-means clustering on PC scores. (j) Clustering accuracy approaches 100% as the trajectories become longer. Inset: Silhouette scores at long times correctly identify the number of clusters. The three populations differ only in their noise magnitude $\sigma_1 = 0.1, \sigma_2 = 0.3, \sigma_3 = 0.5$, and otherwise $\nu = 0.01, R = .05$, with $L = 1$, $\Delta t = 1$, and particle counts$N_1 = 200, N_2 = 200$.

reduction and clustering of these latent representations need not be limited to only two populations. We next consider the variation with three subpopulations of Vicsek particles, differing again only by the noise magnitude $\sigma_1 = 0.1, \sigma_2 = 0.3, \sigma_3 = 0.5$. The simulations and clustering procedure can be seen in Fig. 5. Again, the collective itself does not seem to display any apparent pattern in positions [Figs. 5(a)–5(c)], but the PC scores separate over sufficiently long times [Figs. 5(d)–5(f)]. For long trajectories, the accuracy approaches 100% [Fig. 5(j)]. In practice, the number of clusters must be specified for k-means or other clustering algorithms but may be unknown. In the inset of Fig. 5(j), we plot the silhouette score [68], a metric for choosing the number of clusters. We see that for intermediate times, an incorrect number of clusters may be inferred (five clusters is shown as the maximum), but at sufficiently long times, three clusters are recovered in the silhouette score as the correct number.

E. Model-free clustering is generalizable to a heterogeneous D'Orsogna model

Although the Vicsek has historically served as a test bed for investigations of collective motion, one may wonder whether our results are specific to heterogeneities in this model alone. To explore the generality, we next consider a different, historically important alternative: the D'Orsogna model [59]. The D'Orsogna model describes self-propelled particles in 2D, with the position of the ith particle $\boldsymbol{x}_i$ evolving as

$$\frac{d\boldsymbol{x}_i}{dt} = \boldsymbol{v}_i, \quad \frac{d\boldsymbol{v}_i}{dt} = (\alpha - \beta\|\boldsymbol{v}_i\|^2)\boldsymbol{v}_i - \nabla U(x_i), \tag{4}$$

where

$$U(\boldsymbol{x}_i) = \sum_{i \neq j}^{N} \left[C_r e^{-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|/\lambda_r} - C_a e^{-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|/\lambda_a} \right]. \tag{5}$$

In the model, the parameter α describes the self-propulsion magnitude and β is the friction magnitude. The potential Eq. (5) is a Morse-like potential between all pairs of particles. The two length scales are l_a and l_r, and represent attraction and repulsion, respectively. Each of those magnitudes is governed by C_a and C_r.

The D'Orsogna model can display considerably more complex behavior than the Vicsek counterpart. Depending on the parameter values chosen, possible behaviors range from single mills and double mills to swarms and escapes [43,59]. Here, we investigate a heterogeneous version of the D'Orsogna model with two subpopulations of particles that each have different parameters but in a parameter regime where each subpopulation displays the same qualitative behavior. This choice was motivated by the intuition that a setup where subtypes display different qualitative behaviors should be easier to cluster and less interesting to investigate. We choose β, the friction, to differ. The interaction potential sums all neighbors, both in and out of the subtype. One key difference is that the magnitude of the velocity may change in the D'Orsogna model, whereas in Vicsek it is constant. We again use the orientation alone as the data input to the dimensionality reduction, with $\tau_{i,t} = \mathtt{atan2}(v^y_{i,t}, v^x_{i,t})$, where $v^x_{i,t}$ and $v^y_{i,t}$ represent the x, y component of the velocity observed spaced time intervals enumerated by t. The ODEs are solved numerically using SciPy's Dormand-Prince dopri5 method and then resampled via linear interpolation to be equally spaced observations by $\Delta t = 1$.

In Eq. (4), we see the results of the heterogeneous D'Orsogna simulation and clustering analysis. For the parameters chosen where attraction is stronger than repulsion, a ring behavior appears with particles moving both clockwise and counterclockwise [Figs. 6(a)–6(c)], but otherwise the identities of each subpopulation do not seem distinguishable by eye. The PC values shown in Figs. 6(d)–6(f) do not initially separate the identities, but as longer trajectories are observed, the PC scores from each subtype separate into two circles: those in type 1 with a smaller radius. Due to the shape of the PC scores, k-means expectedly fails to recover the true identities, but standard spectral clustering [67] shown in Figs. 6(g)–6(i) recovers the true identities with flawless accuracy.

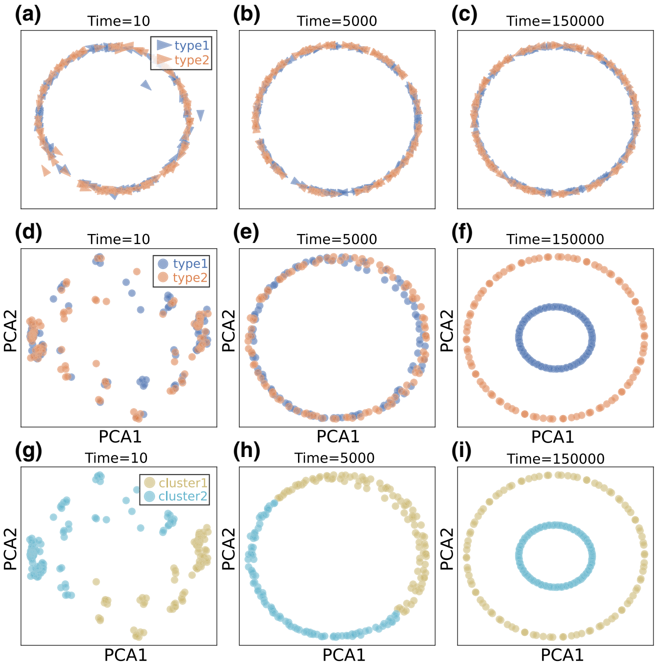

FIG. 6. Heterogeneous D'Orsogna model simulation and clustering. (a)–(c) Snapshots of the particle positions in a heterogeneous D'Orsogna model simulation with two types of particles and displaying no apparent pattern. (d)–(f) The first two principal component scores for each trajectory, colored by particle type. (g)–(i) Results of spectral clustering on PC scores. Simulation parameters are $N_1 = 200$, $N_2 = 200$ with shared parameters: $\alpha = 1.50$, $l_a = 1.0$, $l_r = 0.9$, $C_a = 1.0$, $C_r = 0.9$, but differing $\beta_1 = 0.80$ and $\beta_2 = 0.775$.

F. Limitations on multiple data sets

We have thus far investigated the ability to interrogate a single collective at a time and found that we need sufficiently long trajectories for accurate clustering. However, in practice, experimental constraints limit the ability to take long observations. Instead, it may be more practical to obtain replicates of experiments. We therefore investigate the feasibility of combining data from multiple distinct observations of the same heterogeneous collective. Returning to the setup with two subpopulations of Vicsek particles with differing noise magnitude with run $N_1 = 200$, $N_2 = 200$, as in Fig. 2, we now run three separate simulations. Each simulation is initialized with different random configurations, and then run to the steady state with these transient values discarded, the same as previous figures. The three simulations are concatenated into 3×400 trajectories in one data matrix to cluster. The resulting PC values for the concatenated data can be seen in Fig. 7. At short times, no apparent pattern is seen. As time progresses [Fig. 7(b)], the PC scores split into three groups. This pattern continues at long times [Fig. 7(c)], and each of the three groups splits into two subgroups, resulting in six total clusters. However, the three predominant groups correspond to the three distinct simulations. Therefore, the clustering distinguishes different simulations rather than the same groups between simulations. That is, there does not appear to be a way to tell from the PC scores alone that the three observations were from the same heterogeneous collectives. Intuitively, this is because the temporal structures (correlations) that allow for the statistical separation are limited to a single observation. However, this does not mean the task of identification across multiple experiments is impossible but rather that it seems to require different techniques that incorporate model structure, e.g., the mixture modeling of Ref. [49].

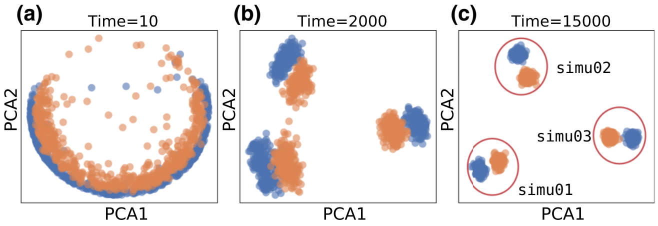

FIG. 7. Clustering fails to combine multiple experiments. (a)–(c) PC scores over increasingly long trajectories with three separate collectives concatenated into a single data set. The same setup of two-subpopulation Vicsek with different noises, as in Fig. 2.

IV. CONCLUSION

In summary, we have investigated the ability to perform clustering to recover the true identities of particles in heterogeneous collectives without prior knowledge of the heterogeneities or underlying model. To do so, we first investigated a heterogeneous Vicsek model. To cluster, the orientations are transformed to nonangular data and then dimensionally reduced via PCA. In these latent dimensions, we find that the trajectories naturally separate over sufficiently long timescales. We find that this timescale is decreased by larger differences in noise magnitudes, larger differences in interaction radii, higher particle densities, and equal subpopulation numbers. The method was readily extended to a heterogeneous Vicsek setup with three types of particles, where the number of clusters was also recovered via a silhouette score. Finally, we show that the premise also extends to other models of collectives by investigating a heterogeneous D'Orsogna model. For this model, we find that spectral clustering was necessary due to the complexity of the PCA scores, but these scores also separate distinctly over long timescales. Ultimately, our results add an important vignette to the growing literature on inferring interactions in collectives, especially those with heterogeneities.

We emphasize that the approach is not intended as an end-all solution to the identification of heterogeneous collectives, but rather complementary to existing approaches. That is, it can be seen as a step of exploratory data analysis to shape the necessary user input to more sophisticated methods such as Refs. [48–50]. One key limitation of our methodology was the inability to identify whether heterogeneities were the same type across different observations. However, the methodology proposed here could be used to identify the existence of heterogeneities and help steer methods such as Refs. [46,49], which we anticipate can readily handle learning interactions and assigning identities across observations.

There are several avenues of future interest stemming from our paper, in both the theory and practice of inferring heterogeneous collectives. It would be interesting to compare the performance of dimensionality reduction approaches to disentangling heterogeneities to those based on information-theoretic quantities like transfer entropy [51,52,69] or Granger causality [70]. The choice of PCA for dimensionality reduction was for simplicity, but future work could also investigate the use of nonlinear approaches such as autoencoders [71] or LSTM architectures [72]. Further, our investigation of heterogeneous collectives was purely numerical. It is therefore of clear interest to explore whether powerful analytical approaches (e.g., Toner-Tu theory [73]) can reveal the intrinsic lower dimensional structure of these heterogeneous collectives. We emphasize the plausibility of future analytical progress by noting the appearance of clusters from a single PC, effectively the covariance between the positions of particles. Such lower dimensional structures have been analytically derived elsewhere for noisy interacting systems [74] and may reveal further insights about the nature of intrinsic disentanglement of heterogeneities we investigate in this paper.

Python code for performing the simulations of the heterogeneous collectives and the clustering analysis therefore can be found at Ref. [75].

ACKNOWLEDGMENT

We thank M. Hua, L. Sun, and M. Wang for working on early preliminaries of this project while undergraduates at New York University.

[1] T. Vicsek and A. Zafeiris, Collective motion, Phys. Rep. **517**, 71 (2012).

[2] A. Deutsch, P. Friedl, L. Preziosi, and G. Theraulaz, Multi-scale analysis and modelling of collective migration in biological systems, Philos. Trans. R. Soc. London, Ser. B **375**, 20190377 (2020).

[3] S. Hubbard, P. Babak, S. T. Sigurdsson, and K. G. Magnússon, A model of the formation of fish schools and migrations of fish, Ecolog. Model. **174**, 359 (2004).

[4] J. Jhawar, R. G. Morris, U. R. Amith-Kumar, M. Danny Raj, T. Rogers, H. Rajendran, and V. Guttal, Noise-induced schooling of fish, Nat. Phys. **16**, 488 (2020).

[5] W. Bialek, A. Cavagna, I. Giardina, T. Mora, E. Silvestri, M. Viale, and A. M. Walczak, Statistical mechanics for natural flocks of birds, Proc. Natl. Acad. Sci. USA **109**, 4786 (2012).

[6] H. Ling, G. E. Mclvor, K. van der Vaart, R. T. Vaughan, A. Thornton, and N. T. Ouellette, Costs and benefits of social relationships in the collective motion of bird flocks, Nat. Ecol. Evol. **3**, 943 (2019).

[7] A. J. Bernoff, M. Culshaw-Maurer, R. A. Everett, M. E. Hohn, W. C. Strickland, and J. Weinburd, Agent-based and continuous models of hopper bands for the Australian plague locust: How resource consumption mediates pulse formation and geometry, PLoS Comput. Biol. **16**, e1007820 (2020).

[8] J. Weinburd, J. Landsberg, A. Kravtsova, S. Lam, T. Sharma, S. J. Simpson, G. A. Sword, and J. Buhl, Anisotropic interaction and motion states of locusts in a hopper band, bioRxiv 2021.10.29.466390 (2023).

[9] H.-P. Zhang, A. Be'er, E.-L. Florin, and H. L. Swinney, Collective motion and density fluctuations in bacterial colonies, Proc. Natl. Acad. Sci. USA **107**, 13626 (2010).

[10] F. Peruani, J. Starruß, V. Jakovljevic, L. Søgaard-Andersen, A. Deutsch, and M. Bär, Collective motion and nonequilibrium cluster formation in colonies of gliding bacteria, Phys. Rev. Lett. **108**, 098102 (2012).

[11] K. W. Rio, G. C. Dachner, and W. H. Warren, Local interactions underlying collective motion in human crowds, Proc. Phys. Soc. London, Sec. B **285**, 20180611 (2018).

[12] E. Méhes and T. Vicsek, Collective motion of cells: From experiments to models, Integr. Biol. **6**, 831 (2014).

[13] R. Alert and X. Trepat, Physical models of collective cell migration, Annu. Rev. Condens. Matter Phys. **11**, 77 (2020).

[14] V. Schaller, C. Weber, C. Semmrich, E. Frey, and A. R. Bausch, Polar patterns of driven filaments, Nature (London) **467**, 73 (2010).

[15] C. E. Miles, J. Zhu, and A. Mogilner, Mechanical torque promotes bipolarity of the mitotic spindle through multi-centrosomal clustering, Bull. Math. Biol. **84**, 29 (2022).

[16] J. W. Jolles, A. J. King, and S. S. Killen, The role of individual heterogeneity in collective animal behaviour, Trends Ecol. Evol. **35**, 278 (2020).

[17] G. Ariel, A. Ayali, A. Be'er, and D. Knebel, Variability and heterogeneity in natural swarms: Experiments and modeling, in *Active Particles, Volume 3: Advances in Theory, Models, and Applications* (Springer, Cham, Switzerland, 2022), pp. 1–33.

[18] S. Peled, S. D. Ryan, S. Heidenreich, M. Bär, G. Ariel, and A. Be'er, Heterogeneous bacterial swarms with mixed lengths, Phys. Rev. E **103**, 032413 (2021).

[19] A. J. Ward, T. Schaerf, A. Burns, J. Lizier, E. Crosato, M. Prokopenko, and M. M. Webster, Cohesion, order and information flow in the collective motion of mixed-species shoals, R. Soc. Open Sci. **5**, 181132 (2018).

[20] J. E. Herbert-Read, S. Krause, L. J. Morrell, T. M. Schaerf, J. Krause, and A. J. W. Ward, The role of individuality in collective group movement, Proc. Phys. Soc. London, Sec. B **280**, 20122564 (2013).

[21] B. Collignon, A. Séguret, Y. Chemtob, L. Cazenille, and J. Halloy, Collective departures and leadership in zebrafish, PLoS ONE **14**, e0216798 (2019).

[22] N. Mizumoto, S.-B. Lee, G. Valentini, T. Chouvenc, and S. C. Pratt, Coordination of movement via complementary interactions of leaders and followers in termite mating pairs, Proc. Phys. Soc. London, Sec. B **288**, 20210998 (2021).

[23] L. Gómez-Nava, R. Bon, and F. Peruani, Intermittent collective motion in sheep results from alternating the role of leader and follower, Nat. Phys. **18**, 1494 (2022).

[24] L. J. Schumacher, P. K. Maini, and R. E. Baker, Semblance of heterogeneity in collective cell migration, Cell Syst. **5**, 119 (2017).

[25] X. Fu, S. Kato, J. Long, H. H. Mattingly, C. He, D. C. Vural, S. W. Zucker, and T. Emonet, Spatial self-organization resolves conflicts between individuality and collective migration, Nat. Commun. **9**, 2177 (2018).

[26] T. Kwon, O.-S. Kwon, H.-J. Cha, and B. J. Sung, Stochastic and heterogeneous cancer cell migration: Experiment and theory, Sci. Rep. **9**, 1 (2019).

[27] L. Qin, D. Yang, W. Yi, H. Cao, and G. Xiao, Roles of leader and follower cells in collective cell migration, Mol. Biol. Cell **32**, 1267 (2021).

[28] D. Zhang, H. Zhu, S. Hostikka, and S. Qiu, Pedestrian dynamics in a heterogeneous bidirectional flow: Overtaking behaviour and lane formation, Physica A **525**, 72 (2019).

[29] G. Ariel, O. Rimer, and E. Ben-Jacob, Order–disorder phase transition in heterogeneous populations of self-propelled particles, J. Stat. Phys. **158**, 579 (2015).

[30] K. Copenhagen, D. A. Quint, and A. Gopinathan, Self-organized sorting limits behavioral variability in swarms, Sci. Rep. **6**, 31808 (2016).

[31] M. del Mar Delgado, M. Miranda, S. J. Alvarez, E. Gurarie, W. F. Fagan, V. Penteriani, A. di Virgilio, and J. M. Morales, The importance of individual variation in the dynamics of animal collective movements, Proc. Phys. Soc. London, Sec. B **373**, 20170008 (2018).

[32] C. Hoell, H. Löwen, and A. M. Menzel, Multi-species dynamical density functional theory for microswimmers: Derivation, orientational ordering, trapping potentials, and shear cells, J. Chem. Phys. **151**, 064902 (2019).

[33] G. Netzer, Y. Yarom, and G. Ariel, Heterogeneous populations in a network model of collective motion, Physica A **530**, 121550 (2019).

[34] B. Khelfa, R. Korbmacher, A. Schadschneider, and A. Tordeux, Heterogeneity-induced lane and band formation in self-driven particle systems, Sci. Rep. **12**, 4768 (2022).

[35] R. Lukeman, Y.-X. Li, and L. Edelstein-Keshet, Inferring individual rules from collective behavior, Proc. Natl. Acad. Sci. USA **107**, 12576 (2010).

[36] R. P. Mann, Bayesian inference for identifying interaction rules in moving animal groups, PLoS ONE **6**, e22827 (2011).

[37] J. E. Herbert-Read, A. Perna, R. P. Mann, T. M. Schaerf, D. J. T. Sumpter, and A. J. W. Ward, Inferring the rules of interaction of shoaling fish, Proc. Natl. Acad. Sci. USA **108**, 18726 (2011).

[38] Y. Katz, K. Tunstrøm, C. C. Ioannou, C. Huepe, and I. D. Couzin, Inferring the structure and dynamics of interactions in schooling fish, Proc. Natl. Acad. Sci. USA **108**, 18720 (2011).

[39] J. Gautrais, F. Ginelli, R. Fournier, S. Blanco, M. Soria, H. Chaté, and G. Theraulaz, Deciphering interactions in moving animal groups, PLoS Comput. Biol. **8**, e1002678 (2012).

[40] W. M. Lord, J. Sun, N. T. Ouellette, and E. M. Bollt, Inference of causal information flow in collective animal behavior, IEEE Trans. Mol. Biol. Multi-Scale Commun. **2**, 107 (2016).

[41] C. J. Torney, M. Lamont, L. Debell, R. J. Angohiatok, L.-M. Leclerc, and A. M. Berdahl, Inferring the rules of social interaction in migrating caribou, Proc. Phys. Soc. London, Sec. B **373**, 20170385 (2018).

[42] F. Lu, M. Zhong, S. Tang, and M. Maggioni, Nonparametric inference of interaction laws in systems of agents from trajectory data, Proc. Natl. Acad. Sci. USA **116**, 14424 (2019).

[43] D. Bhaskar, A. Manhart, J. Milzman, J. T. Nardini, K. M. Storey, C. M. Topaz, and L. Ziegelmeier, Analyzing collective motion with machine learning and topology, Chaos: Interdiscip. J. Nonlinear Sci. **29**, 123125 (2019).

[44] U. S. Basak, S. Sattari, K. Horikawa, and T. Komatsuzaki, Inferring domain of interactions among particles from ensemble of trajectories, Phys. Rev. E **102**, 012404 (2020).

[45] J. LaChance, K. Suh, J. Clausen, and D. J. Cohen, Learning the rules of collective cell migration using deep attention networks, PLoS Comput. Biol. **18**, e1009293 (2022).

[46] A. Nabeel, V. Jadhav, D. R. M, C. Sire, G. Theraulaz, R. Escobedo, S. K. Iyer, and V. Guttal, Data-driven discovery of stochastic dynamical equations of collective motion, Phys. Biol. **20**, 056003 (2023).

[47] T. M. Schaerf, J. E. Herbert-Read, and A. J. W. Ward, A statistical method for identifying different rules of interaction between individuals in moving animal groups, J. R. Soc. Interface **18**, 2020 (2021).

[48] F. Lu, M. Maggioni, and S. Tang, Learning interaction kernels in heterogeneous systems of agents from multiple trajectories, J. Mach. Learn. Res. **22** (2021).

[49] D. A. Messenger, G. E. Wheeler, X. Liu, and D. M. Bortz, Learning anisotropic interaction rules from individual trajectories in a heterogeneous cellular population, J. R. Soc. Interface **19**, 20220412 (2022).

[50] A. Nabeel and D. R. Masila, Disentangling intrinsic motion from neighborhood effects in heterogeneous collective motion, Chaos **32**, 063119 (2022).

[51] S. Butail, V. Mwaffo, and M. Porfiri, Model-free information-theoretic approach to infer leadership in pairs of zebrafish, Phys. Rev. E **93**, 042411 (2016).

[52] V. Mwaffo, S. Butail, and M. Porfiri, Analysis of pairwise interactions in a maximum likelihood sense to identify leaders in a group, Front. Robot. AI **4**, 35 (2017).

[53] T. Vicsek, A. Czirók, E. Ben-Jacob, I. Cohen, and O. Shochet, Novel type of phase transition in a system of self-driven particles, Phys. Rev. Lett. **75**, 1226 (1995).

[54] F. Ginelli, The physics of the Vicsek model, Eur. Phys. J.: Spec. Top. **225**, 2099 (2016).

[55] A. Czirók and T. Vicsek, Collective behavior of interacting self-propelled particles, Physica A **281**, 17 (2000).

[56] H. Chaté, F. Ginelli, G. Grégoire, F. Peruani, and F. Raynaud, Modeling collective motion: variations on the Vicsek model, Eur. Phys. J. B **64**, 451 (2008).

[57] M. C. Miguel, J. T. Parley, and R. Pastor-Satorras, Effects of heterogeneous social interactions on flocking dynamics, Phys. Rev. Lett. **120**, 068303 (2018).

[58] S. Chatterjee, M. Mangeat, C.-U. Woo, H. Rieger, and J. D. Noh, Flocking of two unfriendly species: The two-species Vicsek model, Phys. Rev. E **107**, 024607 (2023).

[59] M. R. D'Orsogna, Y. L. Chuang, A. L. Bertozzi, and L. S. Chayes, Self-propelled particles with soft-core interactions: patterns, stability, and collapse, Phys. Rev. Lett. **96**, 104302 (2006).

[60] J. M. Brown, T. Bossomaier, and L. Barnett, Review of data structures for computationally efficient nearest-neighbour entropy estimators for large systems with periodic boundary conditions, J. Comput. Sci. **23**, 109 (2017).

[61] K. Sargsyan, J. Wright, and C. Lim, GeoPCA: A new tool for multivariate analysis of dihedral angles based on principal component geodesics, Nucleic Acids Res. **40**, e25 (2012).

[62] S. Aghabozorgi, A. S. Shirkhorshidi, and T. Y. Wah, Time-series clustering—a decade review, Inf. Syst. **53**, 16 (2015).

[63] Y.-S. Jeong, M. K. Jeong, and O. A. Omitaomu, Weighted dynamic time warping for time series classification, Pattern Recognit. **44**, 2231 (2011).

[64] H. J. Zhou, L. Li, Y. Li, W. Li, and J. J. Li, PCA outperforms popular hidden variable inference methods for molecular QTL mapping, Genome Biol. **23**, 210 (2022).

[65] M. Greenacre, P. J. Groenen, T. Hastie, A. I. d'Enza, A. Markos, and E. Tuzhilina, Principal component analysis, Nat. Rev. Methods Primers **2**, 100 (2022).

[66] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, Efficient kNN classification with different numbers of nearest neighbors, IEEE Trans. Neural Networks Learn. Syst. **29**, 1774 (2017).

[67] U. von Luxburg, A tutorial on spectral clustering, Stat. Comput. **17**, 395 (2007).

[68] P. J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. **20**, 53 (1987).

[69] N. Orange and N. Abaid, A transfer entropy analysis of leader-follower interactions in flying bats, Eur. Phys. J.: Spec. Top. **224**, 3279 (2015).

[70] K. Fujii, N. Takeishi, K. Tsutsui, E. Fujioka, N. Nishiumi, R. Tanaka, M. Fukushiro, K. Ide, H. Kohno, K. Yoda, S. Takahashi, S. Hiryu, and Y. Kawahara, Learning interaction rules from multi-animal trajectories via augmented behavioral models, in *Advances in Neural Information Processing Systems*, edited by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (Curran Associates, Red Hook, New York, 2021), Vol. 34, pp. 11108–11122.

[71] Y. Wang, H. Yao, and S. Zhao, Auto-encoder based dimensionality reduction, Neurocomputing **184**, 232 (2016).

[72] Y. Yu, X. Si, C. Hu, and J. Zhang, A review of recurrent neural networks: LSTM cells and network architectures, Neural Comput. **31**, 1235 (2019).

[73] J. Toner and Y. Tu, Flocks, herds, and schools: A quantitative theory of flocking, Phys. Rev. E **58**, 4828 (1998).

[74] N. Zagli, G. A. Pavliotis, V. Lucarini, and A. Alecio, Dimension reduction of noisy interacting systems, Phys. Rev. Res. **5**, 013078 (2023).

[75] https://github.com/tanpei0513/vicsek_trajectory.