Title

A dimensionality reduction approach to model-free clustering of trajectories in heterogeneous collectives

Author

Tan, Pei

Publication Date

2023

UNIVERSITY OF CALIFORNIA,
IRVINE

A dimensionality reduction approach to model-free clustering of trajectories in heterogeneous collectives

THESIS

submitted in partial satisfaction of the requirements
for the degree of

MASTER OF SCIENCE

in Mathematical, Computational, and Systems Biology

by

Pei Tan

Thesis Committee:
Christopher E. Miles, Chair
German A. Enciso Ruiz
Matt McHenry

2023

Contents

List of Figures

ACKNOWLEDGMENTS

I would like to express my sincerest appreciation to all those who assisted or contributed to my academic journey.

Thank Dr. German A. Enciso Ruiz and Dr. Matt McHenry for being on the thesis committee. Thank Dr. Axel Almet for providing invaluable suggestions. Thank John James Palacios, Trini Nguyen, Andres Felipe Guerrero Ramirez, and Dr. Benjamin Luke Walker, for giving inspiring talks and discussions in group meetings.

Most of my thanks are for Dr. Christopher E. Miles, my thesis advisor, for his constant support, guidance, and motivation during my master's degree and thesis writing.

Also, many thanks for the love and support of my family and friends.

Special thanks for Dr. Melanie L. Oakes, who kindly returned the mistakenly delivered package containing the most expensive purchase for this project: memory modules.

ABSTRACT OF THE THESIS

A dimensionality reduction approach to model-free clustering of trajectories in heterogeneous collectives

By

Pei Tan

Master of Science in Mathematical, Computational, and Systems Biology

University of California, Irvine, 2023

Christopher E. Miles, Chair

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 naturally cluster with dimensionality reduction computed by PCA. 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'Orsogona model. Altogether, our results quantify the ability to disentangle identities in heterogeneous collectives in a model agnostic manner.

Chapter 1

Introduction

1.1 Collective behaviors in real world

In the natural world, it is common to observe animals and creatures live or migrate in collective, resulting in a synchronized and spatially organized pattern of movement. Notable examples include fish schooling [1,2], birds flocking [3,4], insect [5,6] and bacterial swarming [7,8], human crowds [9], cell migration [10,11], and other subcellular processes [12,13] (Fig. 1.1). Within the system, individuals rely on local cues, such as the positions, motion, or changes in motion of others, to make decisions about their movements [14,15].

Local interactions can vary in complexity based on their environment and proximity. Fish, for example, often swim in groups to reduce the energy cost of locomotion [16]. They also repel one another when they come too close, which helps to prevent collisions and maintain a safe distance between individuals [17]. Recent researchers have gone further to study more intricate factors that influence the movements in dynamics systems, moving beyond the basic binary explanation. Such as, fish can perceive complicated interactions via their lateral line

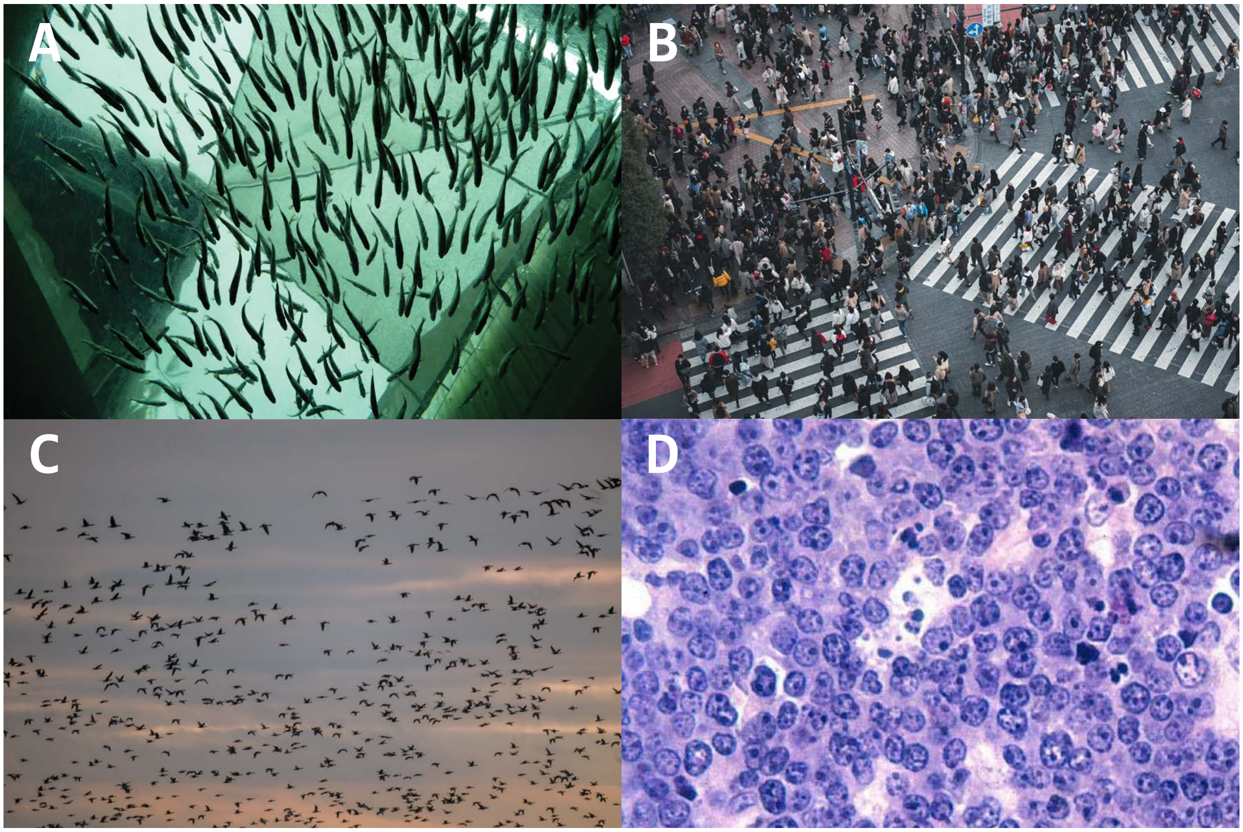

Figure 1.1: **Collective motions in nature.**
A gallery of images related to collective motion. **A**: Fish schooling, photograph by Bruce Warrington. **B**: Human crowd, photograph by Ratapan Anantawat. **C**: Birds flocking, photograph by Jan-Niclas Aberle. **D**: Malignant B-cell lymphocytes seen in Burkitt lymphoma, photograph by Louis M. Staudt.

system, which can respond to stimulation in flow and discern shifts in the environment, such as water temperature and oceanic currents [1, 18].

1.2 Typical collective motion modelings

In the study of collective behavior, it is not always necessary to create complex models that replicate real-life systems with mechanics and dynamics. Instead, simulations can include a basic noise term to take into account various intricate deterministic factors [11,19]. First, the concept of self-propelled particle (SPP) models is introduced [20]. SPP models are made up of particles that interact with each other locally and have an inherent driving force, resulting in a consistent velocity. With this context, Tamás Vicsek proposed the Vicsek model [20], which is a well-known and fundamental model in the field of collective motion. It is regarded as one of the simplest yet effective models for capturing the essential characteristics of collective motion.

The classical Vicsek model describes the evolution of N self-propelled particles moving in 2-dimensional space at a constant speed ν and with fluctuating direction. The direction of each particle is governed by two factors: noise, and local interactions with neighbors (Fig. 1.2). 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

$$\theta_{i,t+1} = \langle \theta_{j,t}(t) \rangle_{\|\boldsymbol{x}_{i,t} - \boldsymbol{x}_{j,t}\| < R} + \eta. \tag{1.1}$$

The particle positions $\boldsymbol{x}$ are updated with these orientations:

$$\boldsymbol{x}_{i,t+1} = \boldsymbol{x}_{i,t} + \nu \Delta t \begin{pmatrix} \cos(\theta_{i,t}) \\ \sin(\theta_{i,t}) \end{pmatrix} \tag{1.2}$$

The noise η is chosen from a uniform distribution governed by a scalar magnitude $0 \leq \sigma \leq 1$, such that $\eta \sim \mathrm{U}(-\sigma\pi, \sigma\pi)$. 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 $\mathcal{O}(N^2)$ comparisons are prohibitive. We instead employ a standard KD-tree [21] $\mathcal{O}(n \log n)$ implementation for computational scalability. Particles are initialized with uniformly random orientation and position within the box.

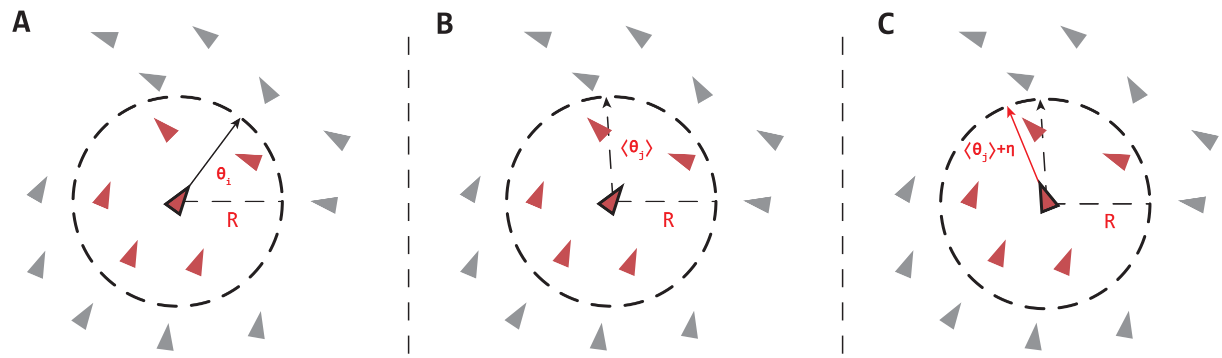

Figure 1.2: **The composition of motion orientation in Vicsek.**
A: In the Vicsek system, particles identify their neighbors using a circular neighborhood with a fixed radius of R (filled red). A particle is oriented towards an angle of θ_i before the interaction. **B**: Particles interact with their neighboring particles by aligning their directions and taking arithmetic mean. **C**: Introduce environmental noise with an intensity of η. This noise and interactions determine the direction of each particle during each Vicsek step.

Despite the existence of noise and the lack of leader particles or global forces, there is a development of orientational order in the Vicsek model, emerging a transient collective cluster pattern. One intriguing finding in the coupling between density/noise and order [20, 22]. When the noise level or density increases, the system experiences a continuous transition from a disordered state to an ordered and coherent motion. The debate is ongoing about whether the order-disorder phase transition is driven by the level of noise [20, 23–25]. Nonetheless, examining this can provide us with insights into collective motion and how to control and manage it effectively.

Although the Vicsek has historically served as a testbed for investigations of collective motion, one may wonder whether our results are specific to heterogeneities in this model alone. To explore the generality, we also consider a different, historically important alternative: the D'Orsogona model [26, 27]. The D'Orsogona model describes self-propelled particles in 2D, with the position of the ith particle $\boldsymbol{x}_i$ evolving as

$$\frac{\mathrm{d}\boldsymbol{x}_i}{\mathrm{d}t} = \boldsymbol{v}_i, \quad \frac{\mathrm{d}\boldsymbol{v}_i}{\mathrm{d}t} = (\alpha - \beta\|\boldsymbol{v}_i\|^2)\boldsymbol{v}_i - \nabla U(\boldsymbol{x}_i), \tag{1.3}$$

where

$$U(\boldsymbol{x}_i) = \sum_{i \neq j}^{N} \left[C_r e^{-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|/\lambda_r} - C_a e^{-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|/\lambda_a} \right]. \tag{1.4}$$

In the model, the parameter α describes the self-propulsion magnitude and β is the friction magnitude. The potential Eq.(1.4) 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.

According to the model equations, the D'Orsogna model is designed to simulate nonlinear interactions between particles through mutual attractive and repulsive forces. Such distinctions, compared to the alignment-focused Vicsek model, lead the D'Orsogna model to different patterns and dynamics. D'Orsogna model can produce various topological patterns, including mills, rings, collective swarms, and group escape [26,28]. Additionally, these patterns can be further classified as either single or double for both ringing and milling patterns [28].

1.3 Heterogeneous collective identification on trajectories

Most attention has been paid towards investigating homogeneous collectives, where all agents evolve and interact via the same dynamics. However, real animal collectives are richly heterogeneous, which is inevitable [29, 30]. Such heterogeneities are frequently observed in various species: bacterias grown from a single cell but still varied in length differences and swarming speed [31]; mixed-species shoals of fish or insects found in wild fields [32]; leader-follower behaviors in animals [33–36]; cell migration may be attributed to the heterogeneity in border cell population [37–39] and personal velocity and preferences result in lane formation in human crowds [40]. The collective motion of heterogeneous systems has consequently been investigated extensively and found to be even richer than that of the homogeneous variety [41–45].

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 [46–51]. 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 [37]. 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 non-parametric tests of heterogeneities [52], or the ability to infer the interactions [53] in heterogeneous collectives, both require knowledge of the particle identities a priori. The work in [54] 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 [55] or only leader-follower interactions [56, 57]. In this work, 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 [20]. This model is renowned as the textbook minimal example of a collective motion with rich behavior [58, 59]. Consequently, many variants have been considered [60], including those with heterogeneities [61,62] such as the ones we propose here. We first consider a setup with two populations of Vicsek particles with different parameters, including interaction radii, noise magnitude, and particle numbers, but still interacting as a single indistinguishable 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 work, 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, we consider a heterogeneous D'Orsogona model [26] and find similar clustering behavior. Altogether, our results are summarized and establish the ability to cluster heterogeneous collectives in a model-free manner with no prior knowledge of the underlying model or heterogeneities.

Chapter 2

Model Simulation

2.1 Heterogeneous Vicsek and clustering

We consider a variant on the classical Vicsek model with subpopulation amount $M \geq 2$. Specifically, denote $\boldsymbol{\phi} = (\nu, \sigma, R)$ 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 $\boldsymbol{\phi}_j = (\nu_j, \sigma_j, R_j)$. 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. Previous studies have examined this model, with some even exploring more complex scenarios. [62, 63].

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 is more subtle. One framing of the problem is that of time-series clustering, for which there are two standard branches of approaches [64]. One can assign and cluster based on an appropriate metric between trajectories, such as Euclidean distance or dynamic time warping [65]. 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. For the sake of discerning the intrinsic separation of the identities, we opt for the simple but classical approach of performing principal component analysis (PCA). PCA simplifies a dataset by creating new variables called principal components (PC). These components are ordered based on the amount of variance they explain in the data. The first component explains the most variance, while the last one explains the least variance. By selecting only the first few principal components, PCA can approximate the original data table while reducing its dimensionality and retaining critical information [66]. It is worthwhile to note that PCA can outperform nonlinear dimensionality reductions in certain contexts [67].

The last technical complication is to decide on the "data" to be dimensionally reduced. Here, we choose $\theta_i(t)$, 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., $\hat{\theta}_{i,t} = \texttt{atan2}(\boldsymbol{x}^y_{i,t+1} - \boldsymbol{x}^y_{i,t}, \boldsymbol{x}^x_{i,t+1} - \boldsymbol{x}^x_{i,t})$, where $\boldsymbol{x}^{x,y}_{i,t}$ correspond to the x, y component of the positions. Naive PCA does not preserve the structure of angular data [68], 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 perform PCA on the $N \times t$ matrix

$$T_t = \tan(\Theta_t) = \begin{bmatrix} \tan\theta_{1,0} & \tan\theta_{1,1} & \cdots & \tan\theta_{1,t} \\ \tan\theta_{2,0} & \tan\theta_{2,1} & \cdots & \tan\theta_{2,t} \\ \vdots & \vdots & \ddots & \vdots \\ \tan\theta_{n,0} & \tan\theta_{n,1} & \cdots & \tan\theta_{n,t} \end{bmatrix}. \tag{2.1}$$

Using the first two principal components, both of them are normalized to have zero mean and unit variance, we then perform standard K-means clustering on these scores, which is a technique to partition data points into K clusters by iteratively assigning points to the nearest center until convergence [69]. It is important to note that K-mean is not the only option for the heterogeneous Vicsek model. Alternative cluster approaches, including K-nearest neighbors [70] and spectral clustering [71], have been tested and present the ability to produce similar results.

So far, we would like to propose the PCA dimensionality reduction cluster, a classification pipeline for heterogeneous collective motion trajectory data, utilizing PCA dimensionality reduction: First, use PCA to extract the first two PC scores, which represent the most informative and significant features and capture the majority of the variability within the data. Then proceed to the clustering stage, which groups particles into their categories (Fig. 2.1).

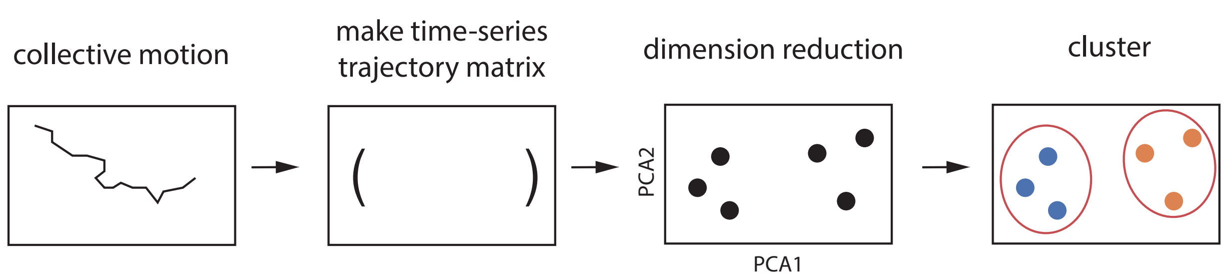

Figure 2.1: **The overview of PCA dimensionality reduction clustering method.** Trajectories can be obtained from motions and then transformed into a time series data matrix. To identify variations in different subpopulations of particles, a dimensionality reduction method is applied on the time series matrix. In this particular study, PCA is utilized as the chosen dimensionality reduction technique.

Chapter 3

Results

3.1 Two subpopulation Vicsek model 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 steps can be seen in Fig. 3.1.

From the snapshots of particle positions (Fig 3.1 panel A, B, C), it can be observed that the collective particles move in cohesive groups, one of the characteristics of the flocking phenomenon. However, the flock is transient and hardly distinguishable as it appears in both subpopulations. In addition, we analyze the orientation distribution within each subpopulation, aiming to understand the spatial alignment or directional tendencies exhibited by

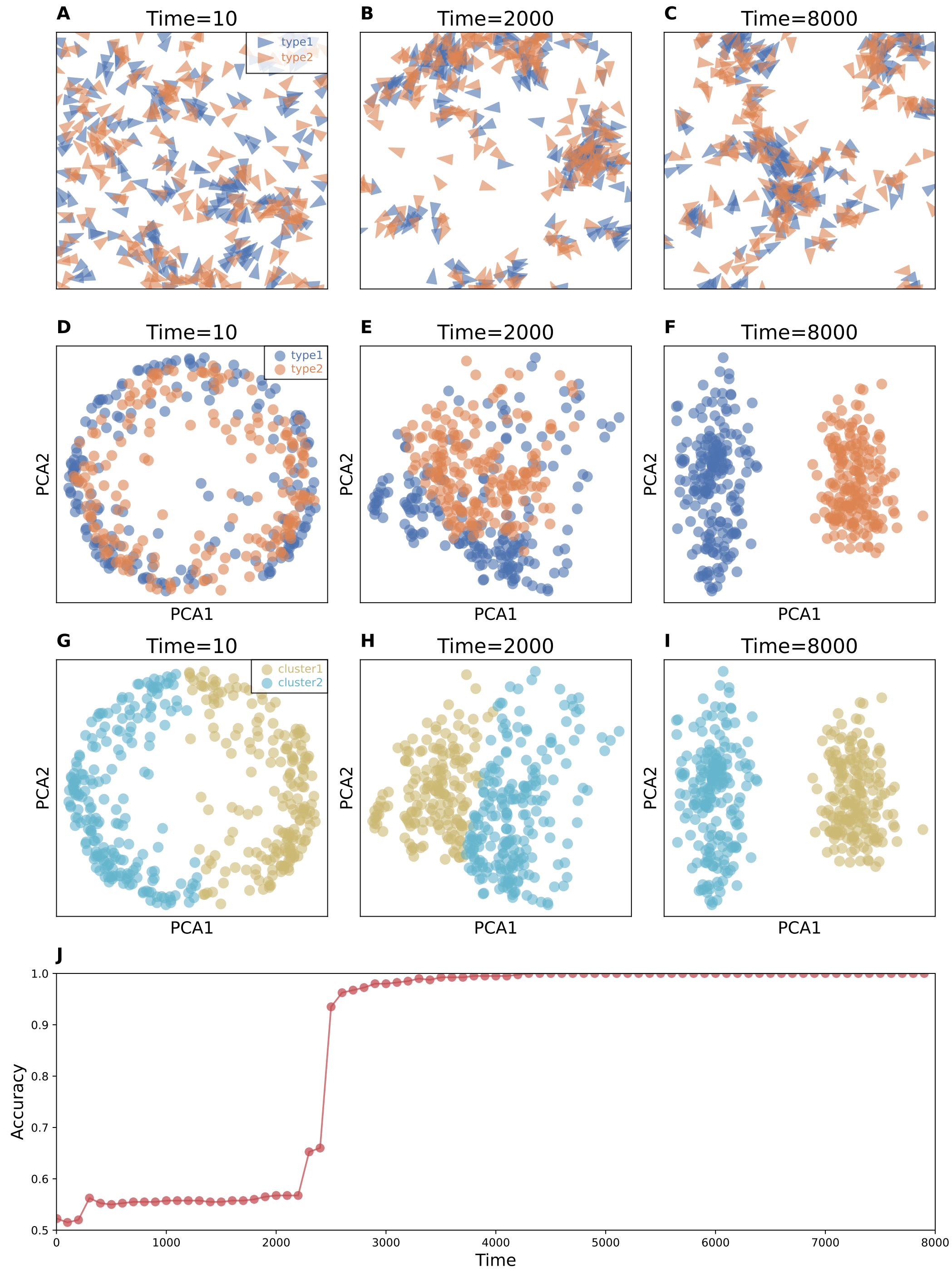

Figure 3.1: **Two subtype Vicsek model simulation and clustering.** **ABC**: snapshots of the particle positions in a heterogeneous Vicsek simulation with two types of particles and display no apparent pattern. **DEF**: The first two principal component scores for each trajectory, colored by particle type. **GHI**: 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 $N_1 = 200, N_2 = 200$.

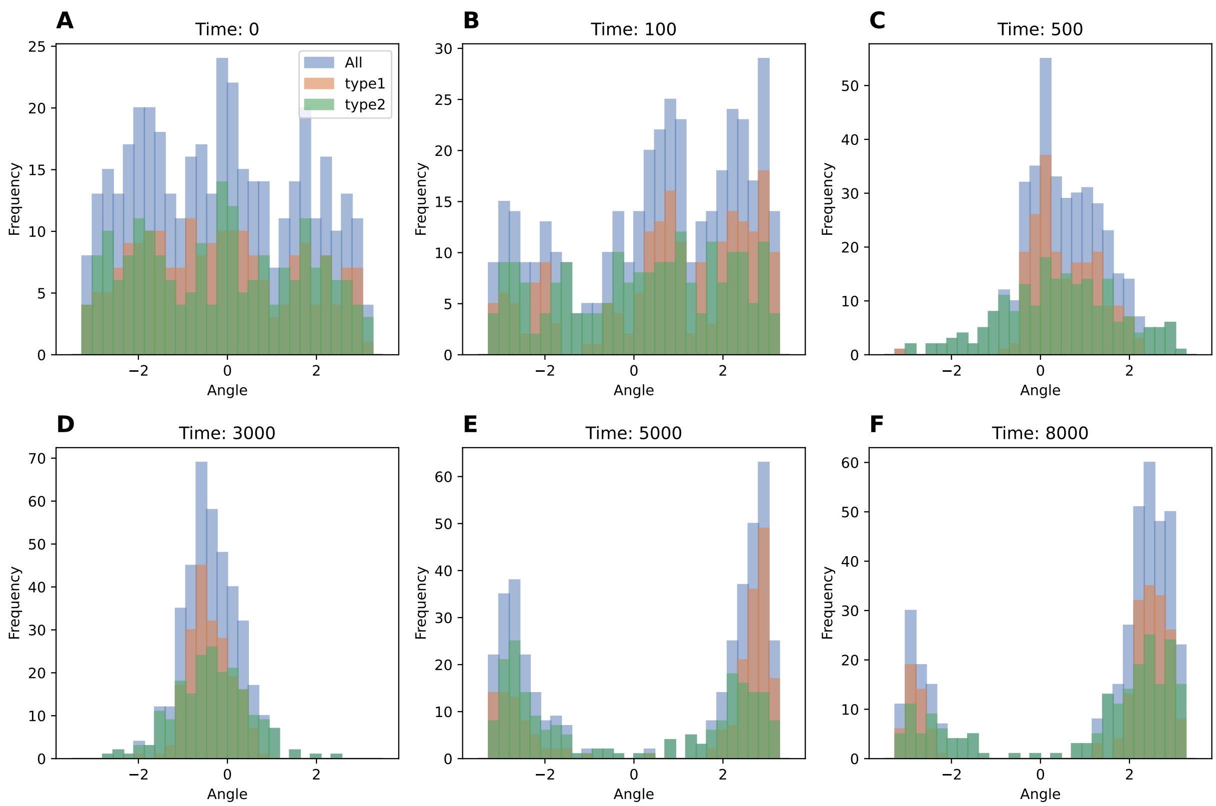

Figure 3.2: **Orientation distribution in two subpopulations Vicsek motion.** The histogram shows the orientation distribution at selected time points: 0 (**A**), 100 (**B**), 500 (**C**), 3000 (**D**), 5000 (**E**), 8000 (**F**). The orientation data are sourced from two subpopulations Vicsek system $N_1 = 200$,$N_2 = 200$ with $\boldsymbol{\phi}_1 = (\nu_1, \sigma_1, R_1)$= $(0.01, 0.1, 0.05)$ and $\boldsymbol{\phi}_2$= $(\nu_2, \sigma_2, R_2) = (0.01, 0.3, 0.05)$. In order to avoid ambiguity, all angles are adjusted to $[-\pi, \pi]$.

individuals within different subpopulations. However, no significant variations are found 3.2. Next, performing the PCA clustering strategy as we stated earlier, the first two PC scores of each trajectory are displayed (Fig 3.1 panel D, E, F). In early times, these scores from various subsets are closely intertwined. However, with the passage of time, they gradually become more distinguishable, although they are still not entirely separate. Eventually, the PC scores for different subpopulations form two distinct clusters. In order to measure the reliability of using PC scores as a resource for the clustering classifier, we conduct K-means clustering on the PC scores and calculate the accuracy. The accuracy is suboptimal initially (around 50%), but it improves as more trajectory data is gathered. Panel J illustrates this improvement, showing that the accuracy finally stabilizes at 100%.

3.2 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 the number of particles $N_1 + N_2$, the ratio of N_1/N_2 in two subpopulations can be seen in Fig. 3.3.

In Fig. 3.3 panel A, we see the effect of differing levels of noise between the two subpopulations of particles, ranging from 2.5 to 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 perfect accuracy.

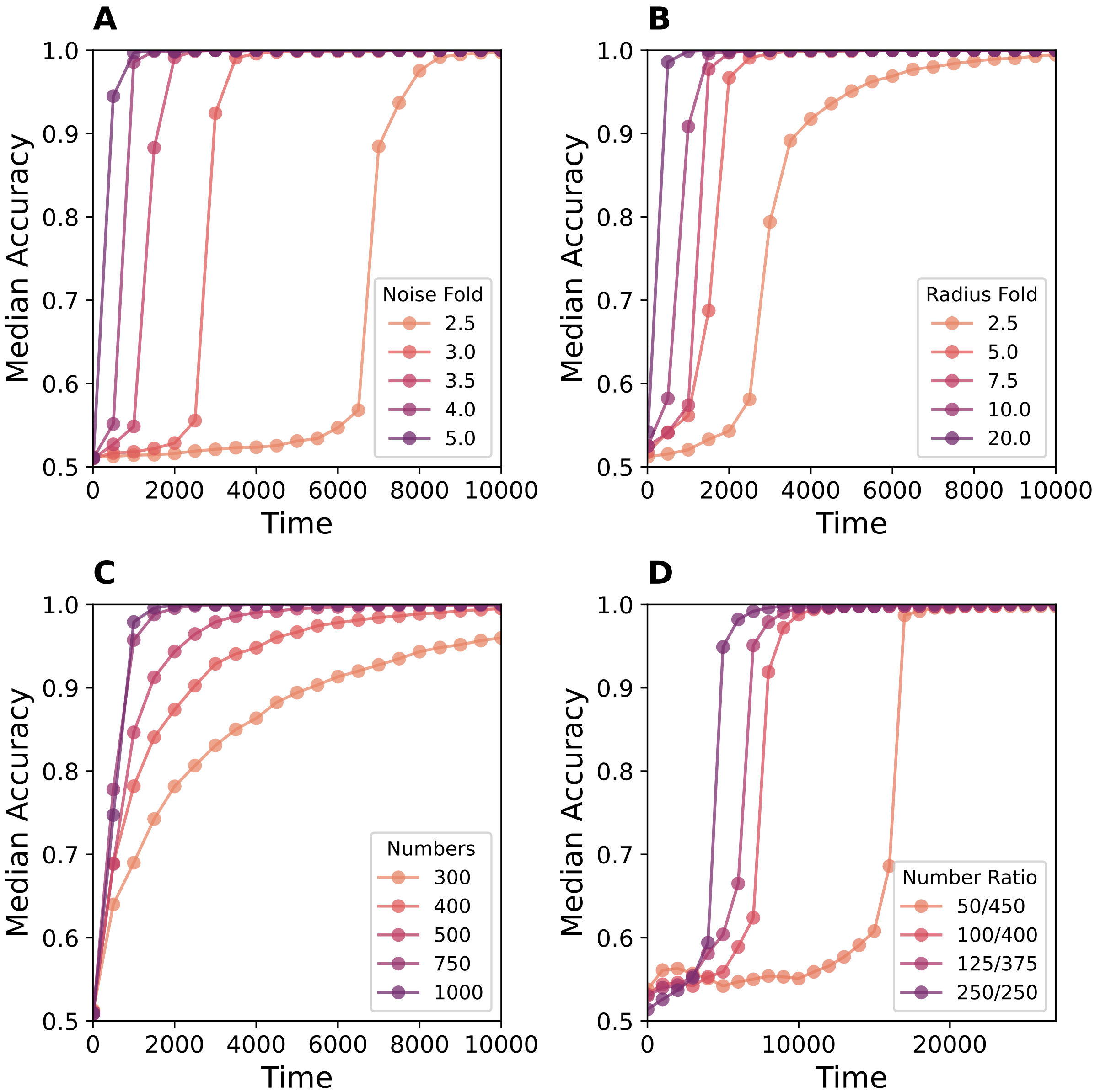

Figure 3.3: **Parameter influence on timescale of accurate clustering.**
A: Median accuracy $N_{\text{sim}} = 100$ of clustering for a two sub-species heterogenous Vicsek model with only noise magnitude different. "Noise fold" refers to the ratio of σ_2/σ_1. **B**: Median accuracy clustering two sub-species 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.

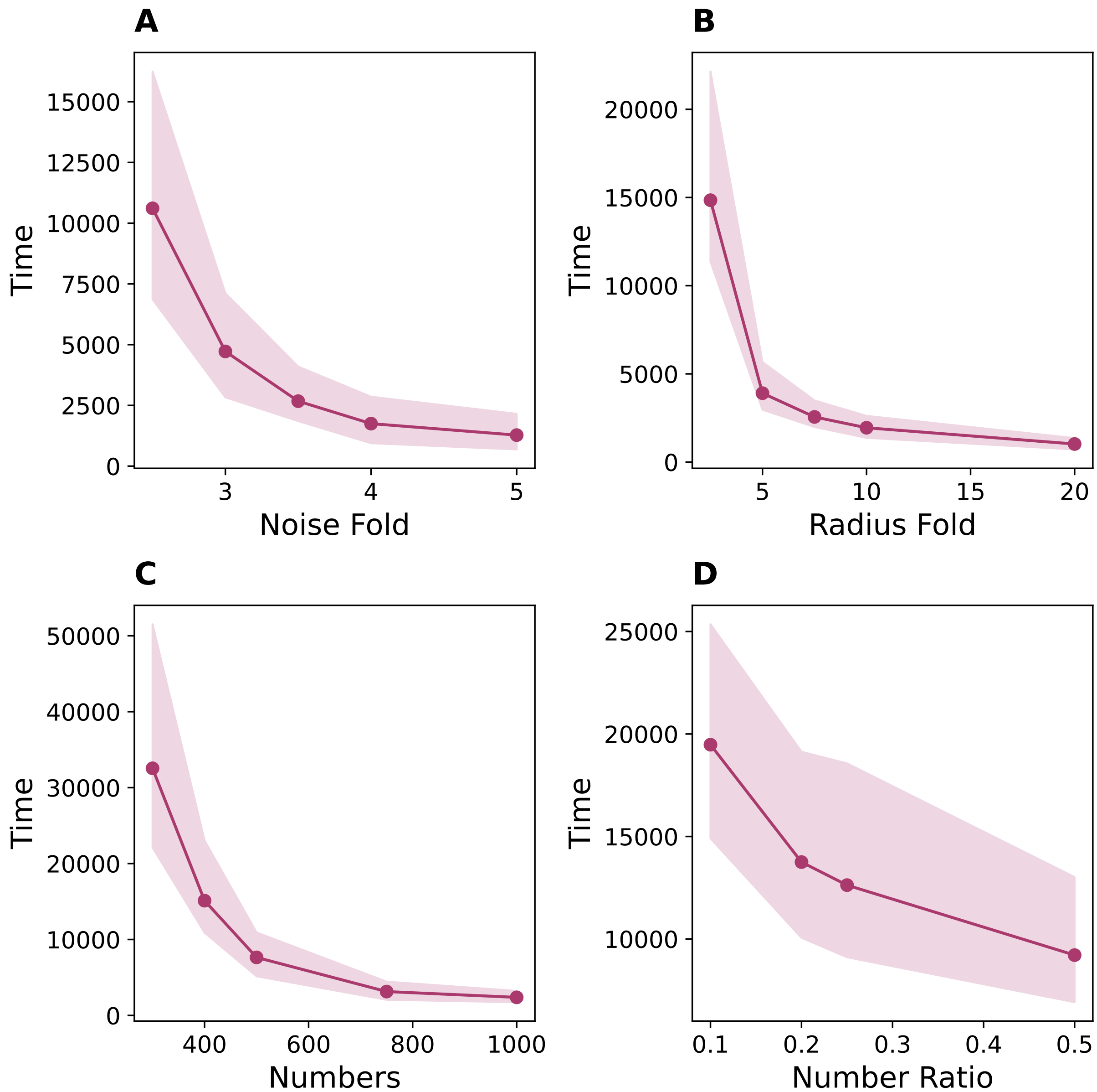

Figure 3.4: **Parameter influence on required time for accurate clustering.** Required time for perfectly accurate clustering from 100 times repeated clustering simulation for a two sub-species heterogenous Vicsek model with various parameters setting. The ”required time” refers to the point in time when accuracy reaches and stabilizes at 100%. It is determined through the method of dichotomy. The scatter plot displays the median required time with bars representing 95% value to 5% value. In **A**, only noise magnitude is different. “Noise fold” refers to the ratio of σ_2/σ_1. In **B**, only interaction radii are different For **C**, the ratio $N_1/N_2 = 1$ is fixed, but the total number of particles $N_1 + N_2 = N$ is increased. For **D**, the ratio $N_1 + N_2 = N$ is fixed, but ratio of $N_1/N_2 = 1$ is varied.

In panel 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 panel C. Lastly, we fix N and vary the ratio of the two subtypes, seen in panel D. Here, we find that greater asymmetry produces longer accurate clustering time. Additionally, we explored the tendency of improving accuracy and identified the point in time when accuracy reaches and stabilizes at 100%, which also supports the finding (Fig. 3.4). 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.

3.3 Cluster relies on simulation time, not initial condition.

In the previous analysis, accurate clustering emerges as the trajectories increase in length. However, it is not immediately clear whether this timescale is shaped by the increased length itself or the equilibration from the initialization of the simulations. To investigate this, we return to the two subpopulation setup with heterogeneous noise and prolong the time window of collective simulation. The point is to maintain the identical dynamic system to guarantee the trajectory only differs in initial positions and initial orientations. We then split the trajectory into two sections, with time window of $[0, t]$ and $[t, 2t]$, to compare their discrepancy in clustering accuracy changes over time. As the Fig. 3.5 demonstrates, the accuracy curve is nearly identical, regardless of the initial configuration of the particles. This suggests the timescales of accuracy are truly representative of the timescale required to observe these populations rather than an artifact of the initialization.

3.4 More than two subpopulations can be clustered.

The previous examples explore a heterogeneous collective with only two subpopulations. However, the dimensionality 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. 3.6. Again, the collective itself does not seem to display any apparent pattern in positions (panels ABC), but the PC scores separate over sufficiently long times (panels DEF). For long trajectories, the accuracy approaches 100% (panel 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 panel J, we plot the silhouette score [72], a metric for choosing the number of clusters. We see that for intermediate times, an incorrect number of clusters may be inferred. When time window is 2000, the maximum silhouette score is at clusters of 5, while the correct cluster is 3. But at sufficiently long times, for instance, time window is 10000, 3 clusters are recovered correctly in the silhouette score.

3.5 Model-free clustering is generalizable to a heterogeneous D'Orsogona model.

The D'Orsogona model can display considerably more complex behavior than the Vicsek counterpart. Depending on the parameter values chosen, D'Orsogona model has various possible behaviors. Here, we investigate a heterogeneous version of the D'Orsogona model

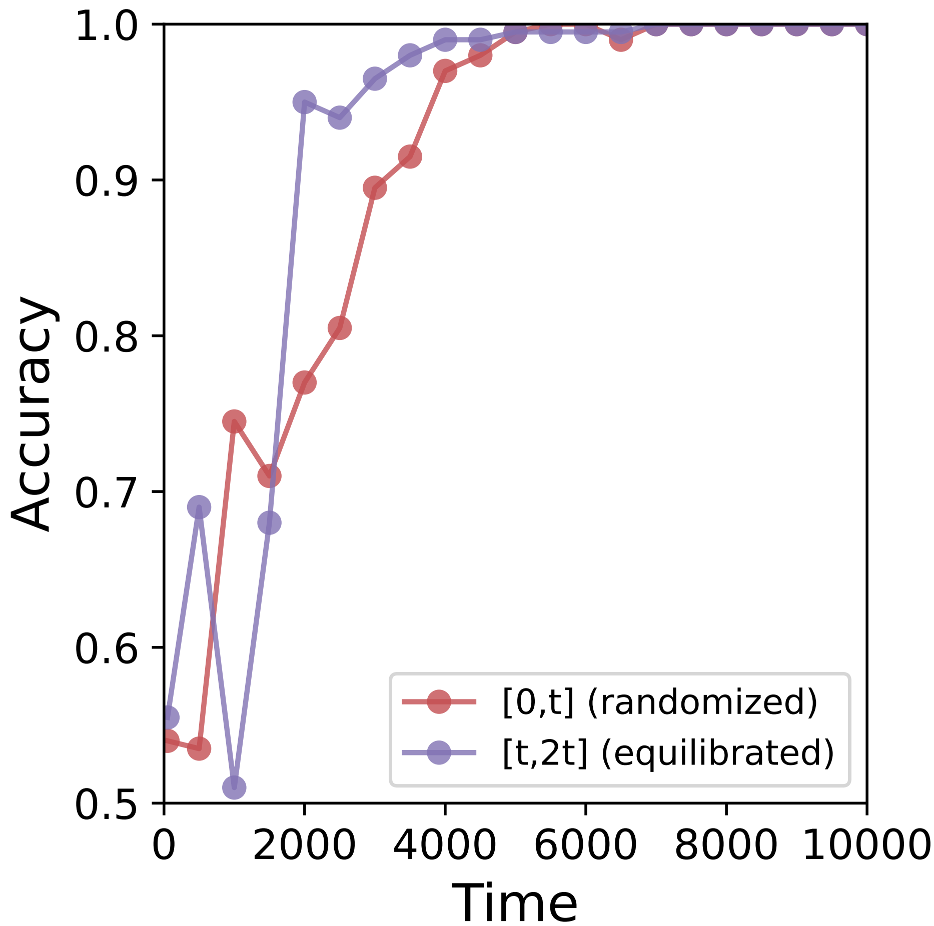

Figure 3.5: **Clustering timescale dependence on initialization.** Clustering accuracy for a single simulation of a heterogeneous Vicsek collective with each subpopulation differing only by noise magnitude, the same as Fig. 3.1. In the red curve, the trajectory window is previously shown $[0, t]$, and in blue, the trajectory from $[t, 2t]$. These two trajectory sections are derived from same Vicsek system.

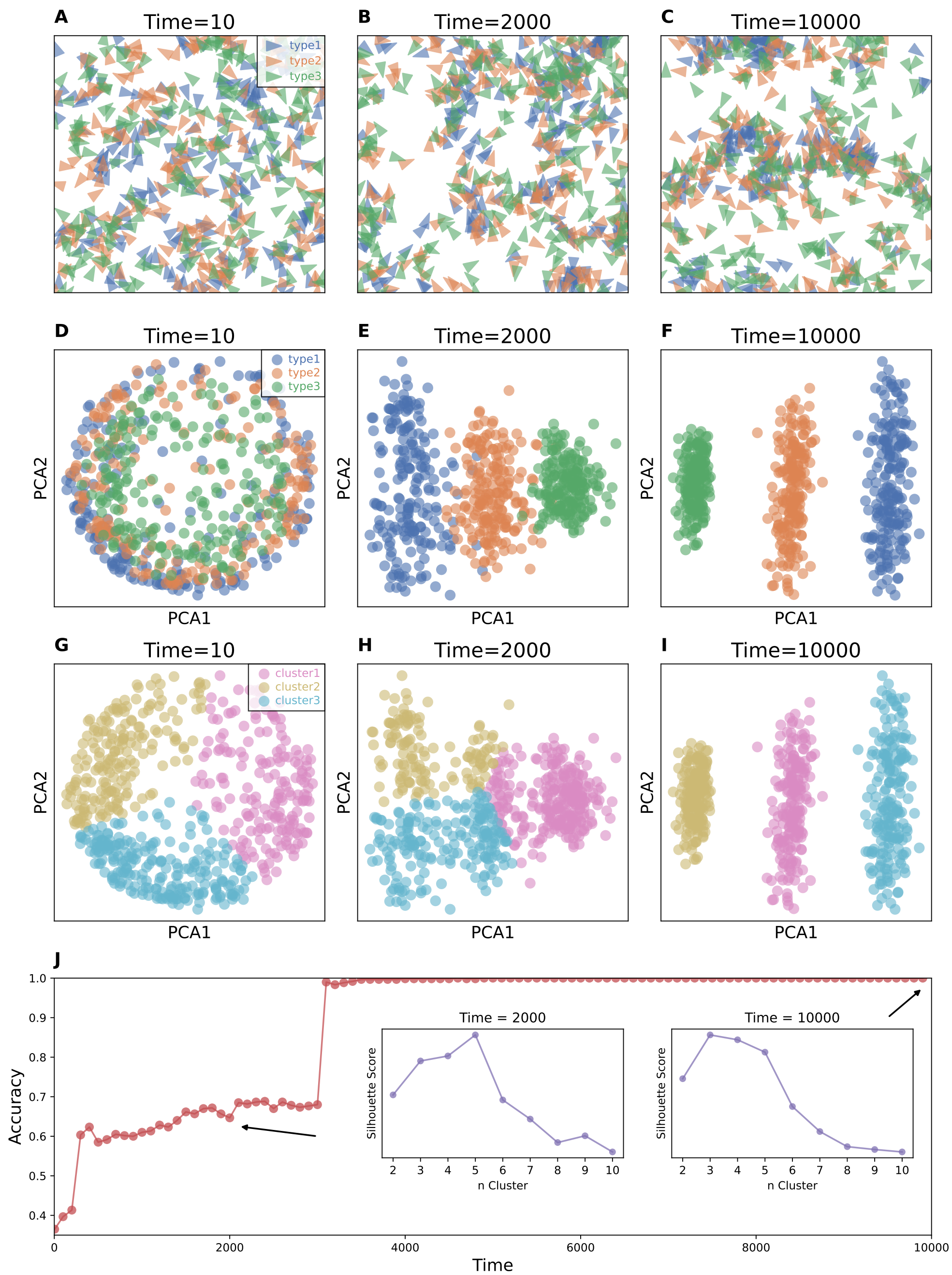

Figure 3.6: **Three subtype Vicsek model simulation and clustering.**
ABC: snapshots of the particle positions in a heterogeneous Vicsek simulation with three types of particles and display no apparent pattern. DEF: The first two principal component scores for each trajectory, colored by particle type. **DEF**: 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$N_1 = 200, N_2 = 200, N_3 = 200$.

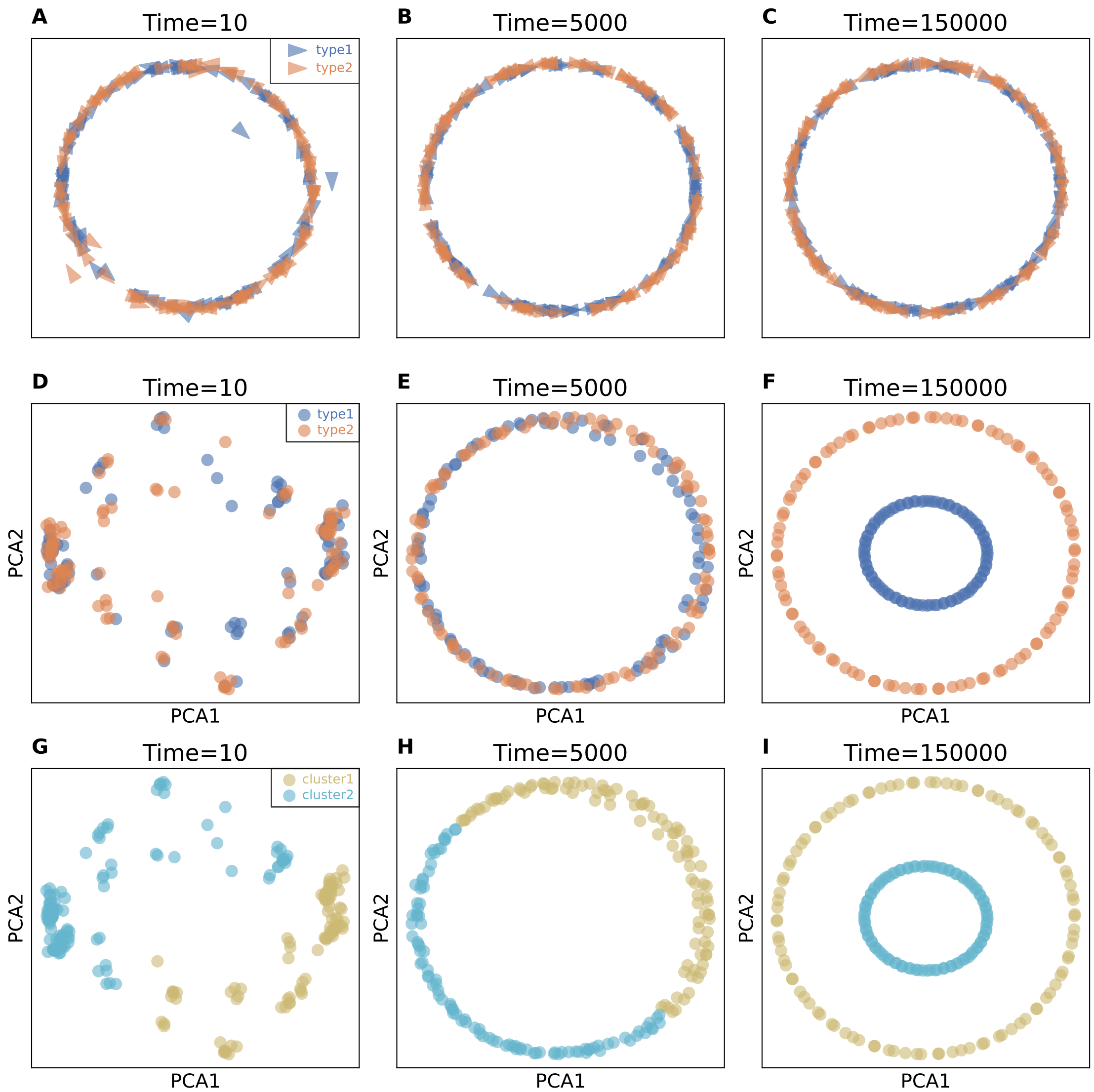

Figure 3.7: **Heterogeneous D'Orsogna model simulation and clustering. ABC**: snapshots of the particle positions in a heterogeneous D'Orsogona model simulation with two types of particles and display no apparent pattern. **DEF**: The first two principal component scores for each trajectory, colored by particle type. **DEF**: Results of spectral clustering on PC scores. Simulation parameter 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 $\beta_2 = 0.775$.

with two subpopulations of particles that each have different parameters. For simplicity, 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'Dorsogona model, whereas in Vicsek it is constant. We again use the orientation alone as the data input to the dimensionality reduction, with $\tau_{i,t} = \mathtt{atan2}(\boldsymbol{v}^y_{i,t}, \boldsymbol{v}^x_{i,t})$ where $\boldsymbol{v}^x_{i,t}$ and $\boldsymbol{v}^y_{i,t}$ 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 re-sampled via linear interpolation to be equally spaced observations by $\Delta t = 1$.

In Fig. 3.7 we see the results of the heterogeneous D'Orsogona 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 (Fig. 3.7 panels ABC) but otherwise the identities of each subpopulation do not seem distinguishable. The PC values shown in DEF 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 [71] recovers the true identities with flawless accuracy.

3.6 Other dimension reduction methods are prospective to function as classifiers.

Through above research, we explore the validity of a PCA-based dimensionality-reduction clustering method to accurately identify the heterogeneity in collective motion. The fact is, there are numerous choices available for dimensionality reduction on large data sets [73]. PCA is first proposed because it is both simple and quick, and has the potential for nonlinear transformations [66, 67]. However, it is still unclear whether other dimension reduction methods can function as well as a classifier on motion trajectories.

One such method is Deep Temporal Clustering (DTC), which combines dimensionality reduction and temporal clustering using an unsupervised learning approach [74]. In contrast to PCA, DTC relies on a BI-LSTM neural network to reduce data dimension [75]. In order to compare the effectiveness of these two approaches, we test them on Vicsek model trajectories and observe that both methods show an increase in accuracy with longer time simulations, as depicted in Fig 3.8. We also assess the algorithm performance by calculating loss and accuracy change in train and test data, as depicted in Fig 3.9. It is a preliminary study of dimensionality-reduction clustering methods in heterogeneous identification. Further investigation is needed to fully understand the potential of these methods.

3.7 Limitations on multiple datasets

We have thus far investigated the ability to interrogate a single collective at a time and find 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. 3.1, we now run three separate simulations. 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.3.10. At short times, no apparent pattern is seen. As time progresses (panel B), the PC scores split into 3 groups. This pattern continues at long times (panel C), and each of the 3 groups splits into

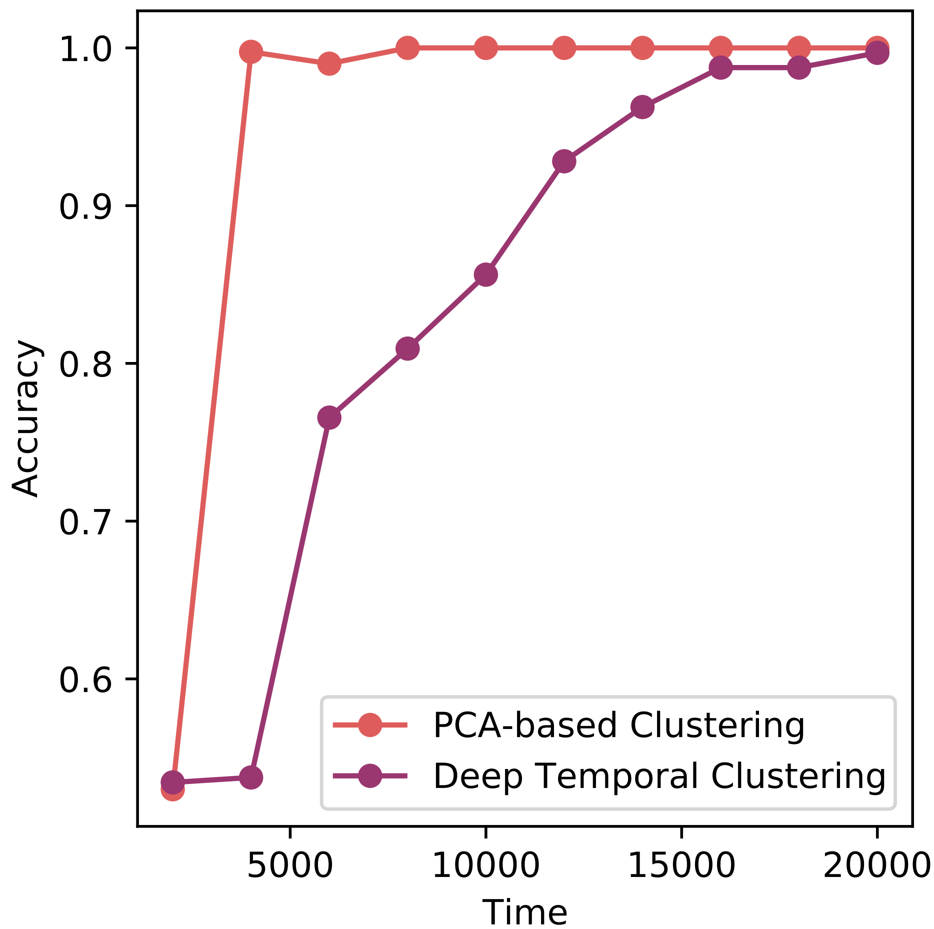

Figure 3.8: **Accuracy comparison over time between clustering methods.**

Both deep temporal clustering and PCA-based clustering method are applied to one two subpopulations Vicsek system with $N_1 = 200$, $\boldsymbol{\phi}_1 = (\nu_1, \sigma_1, R_1) = (0.01, 0.1, 0.05)$ v.s. $N_2 = 200$, $\boldsymbol{\phi}_2 = (\nu_2, \sigma_2, R_2) = (0.01, 0.3, 0.05)$. For deep temporal clustering, training and test set are split with a ratio of 80/20.

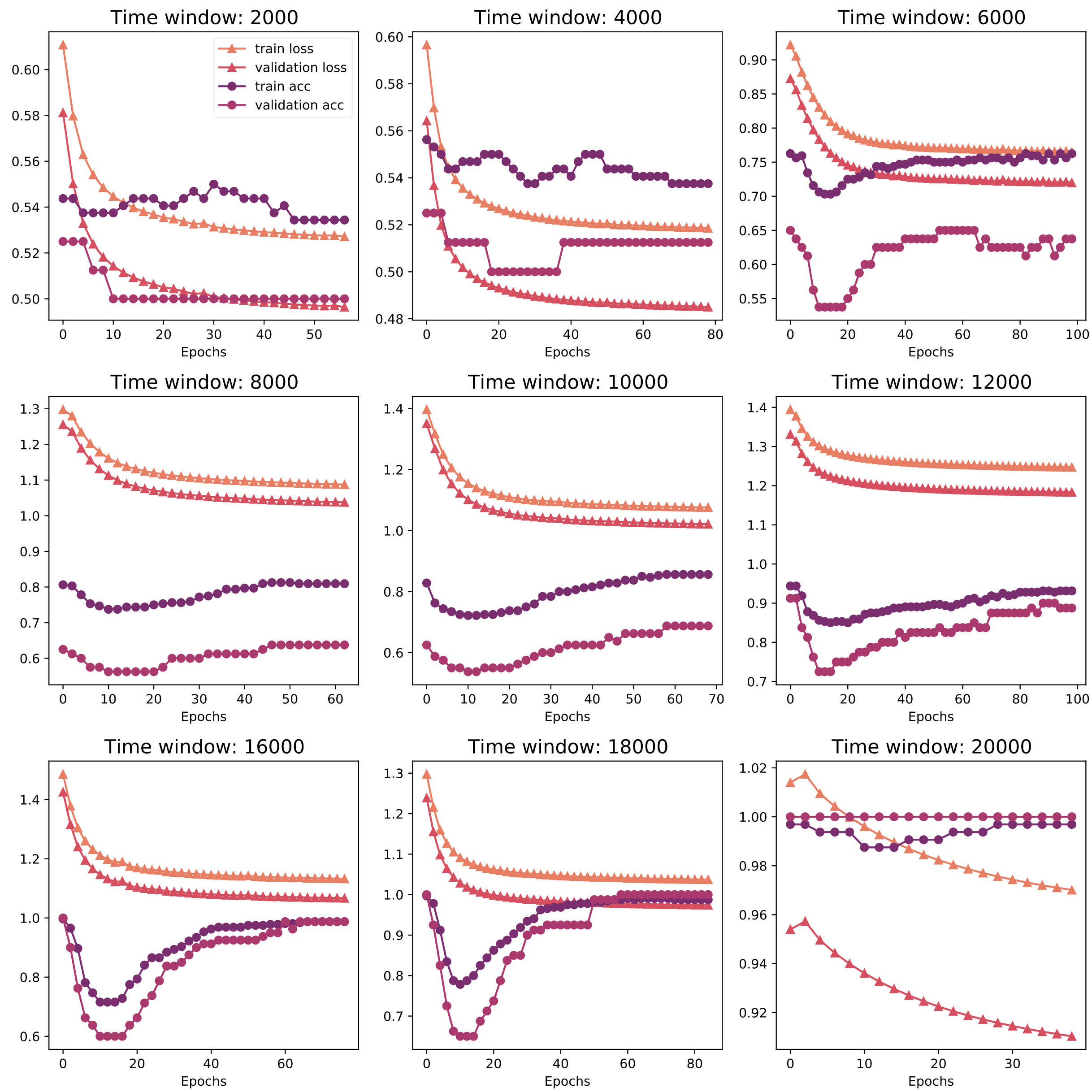

Figure 3.9: **The changes of loss and accuracy for DTC construction in various window lengths.**

To ensure the optimal performance of a model, it is essential to keep a close eye on its convergence and performance. One way to achieve this is by monitoring the loss and accuracy plots, which allow for the identification of potential issues such as overfitting or underfitting. Besides, adjustments can be made to ensure that the model is operating at peak efficiency.

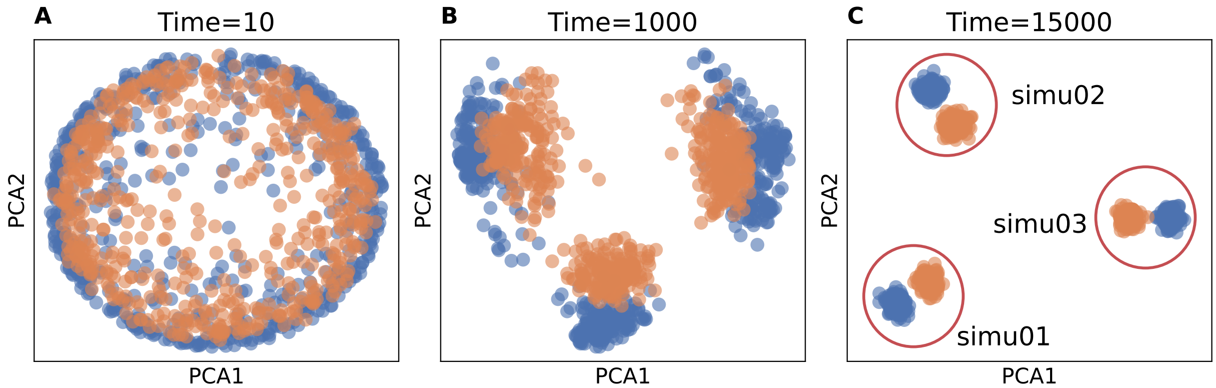

Figure 3.10: **Clustering fails to combine multiple experiments.**
ABC: PC scores over increasingly long trajectories with three separate collectives concatenated into a single dataset. The same setup of two subpopulation Vicsek with different noises as in Fig. 3.1, but initialized with different random value sets.

2 subgroups, resulting in 6 total clusters. However, the 3 predominant groups correspond to the 3 distinct simulations. Therefore, 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 that the 3 observations were from the same heterogeneous collective.

Chapter 4

Discussion

Time-scale collective motion trajectory analysis has always been in challenges, due to the complexity of time series data with spatiotemporal dependencies [76]. Additionally, the high dimensionality, representing the positions or velocities of multiple particles over time, makes it more difficult to extract meaningful patterns and structures. The increasing of time length also will result in an exponential growth of the feature space. Therefore, dimensionality reduction algorithms are crucial in uncovering latent and relevant information.

Another benefit of dimensionality reduction clustering is the potential to reveal intrinsic heterogeneities within the data. By mapping the data into a lower-dimensional space, this algorithm is capable of highlighting underlying patterns and relationships that may be difficult to detect in original high-dimensional space. The technique provides a more insightful analysis and interpretation of trajectory data.

Besides, it is impressive that the clustering can work with different kinds of models. Even though the Vicsek and D'Orsogna models have unique dynamics and interactions, the generality of analysis can still extract important information from high-dimensional data without any prerequisite. This analysis focuses on finding key features and structures in the data to identify meaningful clusters or groups that exhibit similar motion patterns or behaviors rather than fitting on a specific predetermined model.

One limitation or shortcut of PCA-based dimension reduction cluster analysis is that it may not provide explicit information about the specific group or cluster to which each particle belongs. The reduced-dimensional representation obtained through the algorithm does not directly label or assign data points to particular categories. To overcome this, it may be helpful to combine PCA with algorithms or select discriminative dimension reduction techniques, such as neural networks.

Another challenge within the clustering algorithm is in finding the right balance between revealing heterogeneity and reducing noise. Aggressive noise reduction may inadvertently remove important variance and obscure subtle differences among subgroups, leading to a loss of information and potentially misleading results. Conversely, insufficient noise reduction may result in a reduced ability to differentiate between true patterns and noise, leading to overfitting and decreased classification performance. For this purpose, PC scores are applied to preserve the maximum variance in the data, potentially capturing heterogeneity [66]. And neural network algorithms such as DTC implement the feature selection layers before the classifier encoder.

Chapter 5

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 investigate a heterogeneous Vicsek model. To cluster, the orientations are transformed to non-angular 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'Orsogona model. For this model, we find that spectral clustering was necessary due to the complexity of the PCA scores, but these scores did also separate distinctly over long time scales. 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 [53–55]. 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 that helps steer methods such as [51, 54], which we anticipate can readily handle learning interactions and assigning identities across observations.

There are several avenues of future interest stemming from our work, 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 [56,57,77] or Granger causality [78]. The choice of PCA for dimensionality reduction was for simplicity, but future work could also investigate the use of nonlinear approaches such as autoencoders [79] or LSTM architectures [75]. 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 [80]) can reveal the intrinsic lower dimensional structure of these heterogeneous collectives. Such lower dimensional structures have been analytically derived elsewhere for noisy interacting systems [81], and may reveal further insights about the nature of intrinsic disentanglement of heterogeneities we investigate in this work.

References

[1] Simon Hubbard, Petro Babak, Sven Th. Sigurdsson, and Kjartan G. Magnússon. A model of the formation of fish schools and migrations of fish. *Ecological Modelling*, 174(4):359–374, 2004.

[2] Jitesh Jhawar, Richard G. Morris, U. R. Amith-Kumar, M. Danny Raj, Tim Rogers, Harikrishnan Rajendran, and Vishwesha Guttal. Noise-induced schooling of fish. *Nature Physics*, 16(4):488–493, April 2020.

[3] William Bialek, Andrea Cavagna, Irene Giardina, Thierry Mora, Edmondo Silvestri, Massimiliano Viale, and Aleksandra M. Walczak. Statistical mechanics for natural flocks of birds. *Proceedings of the National Academy of Sciences*, 109(13):4786–4791, 2012.

[4] Hangjian Ling, Guillam E Mclvor, Kasper van der Vaart, Richard T Vaughan, Alex Thornton, and Nicholas T Ouellette. Costs and benefits of social relationships in the collective motion of bird flocks. *Nature ecology & evolution*, 3(6):943–948, 2019.

[5] Andrew J. Bernoff, Michael Culshaw-Maurer, Rebecca A. Everett, Maryann E. Hohn, W. Christopher Strickland, and Jasper Weinburd. Agent-based and continuous models of hopper bands for the Australian plague locust: How resource consumption mediates pulse formation and geometry. *PLOS Computational Biology*, 16(5):e1007820, May 2020.

[6] Jasper Weinburd, Jacob Landsberg, Anna Kravtsova, Shanni Lam, Tarush Sharma, Stephen J Simpson, Gregory A Sword, and Jerome Buhl. Anisotropic Interaction and Motion States of Locusts in a Hopper Band. Preprint, Animal Behavior and Cognition, November 2021.

[7] He-Peng Zhang, Avraham Be'er, E-L Florin, and Harry L Swinney. Collective motion and density fluctuations in bacterial colonies. *Proceedings of the National Academy of Sciences*, 107(31):13626–13630, 2010.

[8] Fernando Peruani, Jörn Starruß, Vladimir Jakovljevic, Lotte Søgaard-Andersen, Andreas Deutsch, and Markus Bär. Collective motion and nonequilibrium cluster formation in colonies of gliding bacteria. *Physical Review Letters*, 108(9):098102, 2012.

[9] Kevin W. Rio, Gregory C. Dachner, and William H. Warren. Local interactions underlying collective motion in human crowds. *Proceedings of the Royal Society B: Biological Sciences*, 285(1878):20180611, May 2018.

[10] Előd Méhes and Tamás Vicsek. Collective motion of cells: From experiments to models. *Integrative Biology*, 6(9):831–854, July 2014.

[11] Ricard Alert and Xavier Trepat. Physical models of collective cell migration. *Annual Review of Condensed Matter Physics*, 11:77–101, 2020.

[12] Volker Schaller, Christoph Weber, Christine Semmrich, Erwin Frey, and Andreas R Bausch. Polar patterns of driven filaments. *Nature*, 467(7311):73–77, 2010.

[13] Christopher E. Miles, Jie Zhu, and Alex Mogilner. Mechanical Torque Promotes Bipolarity of the Mitotic Spindle Through Multi-centrosomal Clustering. *Bulletin of Mathematical Biology*, 84(2):29, February 2022.

[14] Iain D Couzin, Jens Krause, et al. Self-organization and collective behavior in vertebrates. *Advances in the Study of Behavior*, 32(1):10–1016, 2003.

[15] Andreas Deutsch, Peter Friedl, Luigi Preziosi, and Guy Theraulaz. Multi-scale analysis and modelling of collective migration in biological systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 375(1807):20190377, September 2020.

[16] Stefano Marras, Shaun S Killen, Jan Lindström, David J McKenzie, John F Steffensen, and Paolo Domenici. Fish swimming in schools save energy regardless of their spatial position. *Behavioral ecology and sociobiology*, 69:219–226, 2015.

[17] C. M. Breder. Equations descriptive of fish schools and other animal aggregations. *Ecology*, 35(3):361–370, 1954.

[18] Amberle McKee, Alberto P Soto, Phoebe Chen, and Matthew J McHenry. The sensory basis of schooling by intermittent swimming in the rummy-nose tetra (hemigrammus rhodostomus). *Proceedings of the Royal Society B*, 287(1937):20200568, 2020.

[19] Tamás Vicsek and Anna Zafeiris. Collective motion. *Physics Reports*, 517(3-4):71–140, 2012.

[20] Tamás Vicsek, András Czirók, Eshel Ben-Jacob, Inon Cohen, and Ofer Shochet. Novel type of phase transition in a system of self-driven particles. *Physical Review Letters*, 75(6):1226–1229, August 1995.

[21] Joshua M Brown, Terry Bossomaier, and Lionel Barnett. Review of data structures for computationally efficient nearest-neighbour entropy estimators for large systems with periodic boundary conditions. *Journal of Computational Science*, 23:109–117, 2017.

[22] Guillaume Grégoire and Hugues Chaté. Onset of collective and cohesive motion. *Physical Review Letters*, 92(2):025702, 2004.

[23] Hugues Chaté, Francesco Ginelli, Guillaume Grégoire, and Franck Raynaud. Collective motion of self-propelled particles interacting without cohesion. *Phys. Rev. E*, 77:046113, Apr 2008.

[24] Gabriel Baglietto and Ezequiel V. Albano. Nature of the order-disorder transition in the vicsek model for the collective motion of self-propelled particles. *Phys. Rev. E*, 80:050103, Nov 2009.

[25] Gabriel Baglietto, Ezequiel V Albano, and Julián Candia. Criticality and the onset of ordering in the standard vicsek model. *Interface Focus*, 2(6):708–714, 2012.

[26] 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. *Physical Review Letters*, 96(10):104302, March 2006.

[27] Herbert Levine, Wouter-Jan Rappel, and Inon Cohen. Self-organization in systems of self-propelled particles. *Physical Review E*, 63(1):017101, 2000.

[28] Dhananjay Bhaskar, Angelika Manhart, Jesse Milzman, John T. Nardini, Kathleen M. Storey, Chad M. Topaz, and Lori Ziegelmeier. Analyzing collective motion with machine learning and topology. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(12), 12 2019. 123125.

[29] Jolle W Jolles, Andrew J King, and Shaun S Killen. The role of individual heterogeneity in collective animal behaviour. *Trends in ecology & evolution*, 35(3):278–291, 2020.

[30] 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*, pages 1–33. Springer, 2022.

[31] Ashley JW Ward, TM Schaerf, ALJ Burns, JT Lizier, Emanuele Crosato, Mikhail Prokopenko, and Michael M Webster. Cohesion, order and information flow in the collective motion of mixed-species shoals. *Royal Society Open Science*, 5(12):181132, 2018.

[32] Shlomit Peled, Shawn D. Ryan, Sebastian Heidenreich, Markus Bär, Gil Ariel, and Avraham Be'er. Heterogeneous bacterial swarms with mixed lengths. *Physical Review E*, 103(3):032413, March 2021.

[33] 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. *Proceedings of the Royal Society B: Biological Sciences*, 280(1752):20122564, February 2013.

[34] Bertrand Collignon, Axel Séguret, Yohann Chemtob, Leo Cazenille, and José Halloy. Collective departures and leadership in zebrafish. *PloS ONE*, 14(5):e0216798, 2019.

[35] Nobuaki Mizumoto, Sang-Bin Lee, Gabriele Valentini, Thomas Chouvenc, and Stephen C. Pratt. Coordination of movement via complementary interactions of leaders and followers in termite mating pairs. *Proceedings of the Royal Society B: Biological Sciences*, 288(1954):20210998, July 2021.

[36] Luis Gómez-Nava, Richard Bon, and Fernando Peruani. Intermittent collective motion in sheep results from alternating the role of leader and follower. *Nature Physics*, 18(12):1494–1501, December 2022.

[37] Linus J. Schumacher, Philip K. Maini, and Ruth E. Baker. Semblance of Heterogeneity in Collective Cell Migration. *Cell Systems*, 5(2):119–127.e1, August 2017.

[38] Taejin Kwon, Ok-Seon Kwon, Hyuk-Jin Cha, and Bong June Sung. Stochastic and heterogeneous cancer cell migration: experiment and theory. *Scientific reports*, 9(1):1–13, 2019.

[39] Lei Qin, Dazhi Yang, Weihong Yi, Huiling Cao, and Guozhi Xiao. Roles of leader and follower cells in collective cell migration. *Molecular Biology of the Cell*, 32(14):1267–1272, 2021.

[40] Dawei Zhang, Haitao Zhu, Simo Hostikka, and Shi Qiu. Pedestrian dynamics in a heterogeneous bidirectional flow: overtaking behaviour and lane formation. *Physica A: Statistical Mechanics and its Applications*, 525:72–84, 2019.

[41] Gil Ariel, Oren Rimer, and Eshel Ben-Jacob. Order–disorder phase transition in heterogeneous populations of self-propelled particles. *Journal of Statistical Physics*, 158:579–588, 2015.

[42] Katherine Copenhagen, David A. Quint, and Ajay Gopinathan. Self-organized sorting limits behavioral variability in swarms. *Scientific Reports*, 6(1):31808, August 2016.

[43] Maria del Mar Delgado, Maria Miranda, Silvia J Alvarez, Eliezer Gurarie, William F Fagan, Vincenzo Penteriani, Agustina di Virgilio, and Juan Manuel Morales. The importance of individual variation in the dynamics of animal collective movements. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1746):20170008, 2018.

[44] Christian Hoell, Hartmut Löwen, and Andreas M. Menzel. Multi-species dynamical density functional theory for microswimmers: Derivation, orientational ordering, trapping potentials, and shear cells. *The Journal of Chemical Physics*, 151(6):064902, August 2019.

[45] Gal Netzer, Yuval Yarom, and Gil Ariel. Heterogeneous populations in a network model of collective motion. *Physica A: Statistical Mechanics and its Applications*, 530:121550, September 2019.

[46] Warren M. Lord, Jie Sun, Nicholas T. Ouellette, and Erik M. Bollt. Inference of Causal Information Flow in Collective Animal Behavior. *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, 2(1):107–116, June 2016.

[47] Colin J. Torney, Myles Lamont, Leon Debell, Ryan J. Angohiatok, Lisa-Marie Leclerc, and Andrew M. Berdahl. Inferring the rules of social interaction in migrating caribou. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1746):20170385, May 2018.

[48] Fei Lu, Ming Zhong, Sui Tang, and Mauro Maggioni. Nonparametric inference of interaction laws in systems of agents from trajectory data. *Proceedings of the National Academy of Sciences*, 116(29):14424–14433, 2019.

[49] Udoy S. Basak, Sulimon Sattari, Kazuki Horikawa, and Tamiki Komatsuzaki. Inferring domain of interactions among particles from ensemble of trajectories. *Physical Review E*, 102(1):012404, July 2020.

[50] Julienne LaChance, Kevin Suh, Jens Clausen, and Daniel J. Cohen. Learning the rules of collective cell migration using deep attention networks. *PLOS Computational Biology*, 18(4):e1009293, April 2022.

[51] Arshed Nabeel, Vivek Jadhav, Danny Raj M, Clément Sire, Guy Theraulaz, Ramón Escobedo, Srikanth K. Iyer, and Vishwesha Guttal. Data-driven discovery of stochastic dynamical equations of collective motion, 2023.

[52] 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. *Journal of The Royal Society Interface*, 18(176):rsif.2020.0925, 20200925, March 2021.

[53] Fei Lu, Mauro Maggioni, and Sui Tang. Learning interaction kernels in heterogeneous systems of agents from multiple trajectories. *J. Mach. Learn. Res.*, 22(1), jan 2021.

[54] Daniel A. Messenger, Graycen E. Wheeler, Xuedong Liu, and David M. Bortz. Learning anisotropic interaction rules from individual trajectories in a heterogeneous cellular population. *Journal of The Royal Society Interface*, 19(195):20220412, 2022.

[55] Arshed Nabeel and Danny Raj Masila. Disentangling intrinsic motion from neighborhood effects in heterogeneous collective motion. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 32(6):063119, June 2022.

[56] Sachit Butail, Violet Mwaffo, and Maurizio Porfiri. Model-free information-theoretic approach to infer leadership in pairs of zebrafish. *Physical Review E*, 93(4):042411, 2016.

[57] Violet Mwaffo, Sachit Butail, and Maurizio Porfiri. Analysis of pairwise interactions in a maximum likelihood sense to identify leaders in a group. *Frontiers in Robotics and AI*, 4:35, 2017.

[58] Francesco Ginelli. The physics of the Vicsek model. *The European Physical Journal Special Topics*, 225:2099–2117, 2016.

[59] András Czirók and Tamás Vicsek. Collective behavior of interacting self-propelled particles. *Physica A: Statistical Mechanics and its Applications*, 281(1-4):17–29, 2000.

[60] Hugues Chaté, Francesco Ginelli, Guillaume Grégoire, Fernando Peruani, and Franck Raynaud. Modeling collective motion: variations on the vicsek model. *The European Physical Journal B*, 64:451–456, 2008.

[61] M Carmen Miguel, Jack T Parley, and Romualdo Pastor-Satorras. Effects of heterogeneous social interactions on flocking dynamics. *Physical Review Letters*, 120(6):068303, 2018.

[62] Swarnajit Chatterjee, Matthieu Mangeat, Chul-Ung Woo, Heiko Rieger, and Jae Dong Noh. Flocking of two unfriendly species: The two-species vicsek model. *Physical Review E*, 107(2):024607, 2023.

[63] Sagarika Adhikary and SB Santra. Pattern formation and phase transition in the collective dynamics of a binary mixture of polar self-propelled particles. *Physical Review E*, 105(6):064612, 2022.

[64] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. Time-series clustering–a decade review. *Information systems*, 53:16–38, 2015.

[65] Young-Seon Jeong, Myong K Jeong, and Olufemi A Omitaomu. Weighted dynamic time warping for time series classification. *Pattern recognition*, 44(9):2231–2240, 2011.

[66] Michael Greenacre, Patrick JF Groenen, Trevor Hastie, Alfonso Iodice d'Enza, Angelos Markos, and Elena Tuzhilina. Principal component analysis. *Nature Reviews Methods Primers*, 2(1):100, 2022.

[67] Heather J. Zhou, Lei Li, Yumei Li, Wei Li, and Jingyi Jessica Li. PCA outperforms popular hidden variable inference methods for molecular QTL mapping. *Genome Biology*, 23(1):210, October 2022.

[68] Karen Sargsyan, Jon Wright, and Carmay Lim. GeoPCA: A new tool for multivariate analysis of dihedral angles based on principal component geodesics. *Nucleic Acids Research*, 40(3):e25–e25, February 2012.

[69] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[70] Shichao Zhang, Xuelong Li, Ming Zong, Xiaofeng Zhu, and Ruili Wang. Efficient knn classification with different numbers of nearest neighbors. *IEEE Transactions on Neural Networks and Learning Systems*, 29(5):1774–1785, 2017.

[71] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17:395–416, 2007.

[72] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[73] Tak chung Fu. A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1):164–181, 2011.

[74] Naveen S. Madiraju. *Deep Temporal Clustering: Fully Unsupervised Learning of Time-domain Features*. PhD thesis, 2018. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2023-03-03.

[75] Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. A review of recurrent neural networks: Lstm cells and network architectures. *Neural Computation*, 31(7):1235–1270, 2019.

[76] T. Warren Liao. Clustering of time series data—a survey. *Pattern Recognition*, 38(11):1857–1874, 2005.

[77] N Orange and N Abaid. A transfer entropy analysis of leader-follower interactions in flying bats. *The European Physical Journal Special Topics*, 224(17-18):3279–3293, 2015.

[78] Keisuke Fujii, Naoya Takeishi, Kazushi Tsutsui, Emyo Fujioka, Nozomi Nishiumi, Ryoya Tanaka, Mika Fukushiro, Kaoru Ide, Hiroyoshi Kohno, Ken Yoda, Susumu Takahashi, Shizuko Hiryu, and Yoshinobu Kawahara. Learning interaction rules from multi-animal trajectories via augmented behavioral models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 11108–11122. Curran Associates, Inc., 2021.

[79] Yasi Wang, Hongxun Yao, and Sicheng Zhao. Auto-encoder based dimensionality reduction. *Neurocomputing*, 184:232–242, 2016.

[80] John Toner and Yuhai Tu. Flocks, herds, and schools: A quantitative theory of flocking. *Physical Review E*, 58(4):4828, 1998.

[81] Niccolò Zagli, Grigorios A. Pavliotis, Valerio Lucarini, and Alexander Alecio. Dimension reduction of noisy interacting systems. *Physical Review Research*, 5(1):013078, February 2023.